

# IBM Synthetic Data Sets

Erik Altman

Dipali Aphale

Joy Deng

Yadu Nandan B

Saurabh Srivastava

Kelly Xiang



**IBM Z**

**IBM LinuxONE**

**Artificial Intelligence**





IBM Redbooks

**IBM Synthetic Data Sets**

February 2025

**Note:** Before using this information and the product it supports, read the information in “Notices” on page v.

**First Edition (February 2025)**

This edition applies to IBM Synthetic Data Sets.

**© Copyright International Business Machines Corporation 2025. All rights reserved.**

Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

# Contents

<b>Notices</b> .....	v
Trademarks .....	vi
<b>Preface</b> .....	vii
Authors .....	vii
Now you can become a published author, too! .....	viii
Comments welcome .....	ix
Stay connected to IBM Redbooks .....	ix
<b>Introducing IBM Synthetic Data Sets</b> .....	1
Synthetic data in the AI model lifecycle .....	2
<b>Dataset deep dive</b> .....	3
IBM Synthetic Data Sets for Payment Cards .....	4
IBM Synthetic Data Sets for Core Banking and Money Laundering .....	5
IBM Synthetic Data Sets for Homeowners Insurance .....	6
<b>Available editions</b> .....	7
Trial Edition .....	8
Pro Edition .....	8
Enterprise Edition .....	8
<b>Previewing data schemas</b> .....	9
<b>Using real data versus synthetic data</b> .....	10
Speeding up time to value with privacy-compliant data .....	11
Broader and richer data .....	11
Data privacy, security, and compliance .....	12
Saving costs with synthetic training data .....	12
<b>Data generation methodology</b> .....	13
Simulating a realistic world .....	14
Creating regular and varied consumer behavior .....	14
Constructing real assets .....	15
Connecting different parts of a simulated world .....	15
Understanding criminal behavior .....	16
<b>Artificial intelligence ethics</b> .....	17
Fairness .....	18
Robustness .....	18
Value alignment .....	18
Data laws .....	19
Intellectual property .....	19
Transparency .....	19
Privacy .....	20
Conclusion .....	20
<b>Legal usage terms</b> .....	21
<b>Getting started</b> .....	22

Artificial intelligence on IBM Z Solution Templates .....	23
IBM Technology Expert Labs Services .....	23
Starting a proof-of-concept with the AI on IBM Z team .....	23
<b>Frequently asked questions</b> .....	<b>24</b>
<b>Additional resources</b> .....	<b>27</b>
<b>Appendix: Data schemes for IBM Synthetic Data Sets</b> .....	<b>28</b>
Payment cards .....	29
Core banking .....	32
Insurance .....	37

# Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

*IBM Director of Licensing, IBM Corporation, North Castle Drive, MD-NC119, Armonk, NY 10504-1785, US*

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to actual people or business enterprises is entirely coincidental.

## COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

# Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at <https://www.ibm.com/legal/copytrade.shtml>

The following terms are trademarks or registered trademarks of International Business Machines Corporation, and might also be trademarks or registered trademarks in other countries.

IBM®


IBM Cloud®

IBM Security®

IBM Z®

Passport Advantage®

Redbooks®

Redbooks (logo) ®

z/OS®

The following terms are trademarks of other companies:

Other company, product, or service names may be trademarks or service marks of others.



# Preface

IBM Synthetic Data Sets is a family of artificially generated, enterprise-grade datasets that enhance predictive artificial intelligence (AI) model training and large language models (LLMs) to benefit IBM Z® and IBM LinuxONE clients, ecosystems, and independent software vendors. These pre-built datasets are downloadable and packaged as comma-separated values (CSVs) and data definition language (DDL) files, making them familiar to use, and compatible with everything from databases to spreadsheets to hardware platforms to standard AI tools. These datasets also leverage the IBM® industry expertise and domain knowledge of the financial services sector without using any real client seed data, which alleviates security concerns with Personally Identifiable Information (PII). Real data at client sites is often limited in scope to only their own organization's transactions, and clients do not always know which transactions are fraudulent or not. To address this scenario, IBM Synthetic Data Sets were modified for fraud detection use cases so that clients can download and enable development of predictive AI models and LLMs for financial services or optimize existing models for improved accuracy and risk mitigation.

The IBM Synthetic Data Sets family contains the following features:

- ▶ IBM Synthetic Data Sets for Payment Cards
- ▶ IBM Synthetic Data Sets for Core Banking and Money Laundering
- ▶ IBM Synthetic Data Sets for Homeowners Insurance

This IBM Redbooks® publication introduces IBM Synthetic Data Sets and provides information about how IBM Synthetic Data Sets can enhance and optimize your predictive AI model training and LLMs.

## Authors

This publication was produced by a team of specialists from around the world working with the IBM Redbooks team.

**Erik Altman** is a Research Scientist at the IBM T.J. Watson Research Center. He has worked across many technical disciplines, such as computer architecture and artificial intelligence (AI). He has written dozens of scientific papers, and has dozens of issued patents. His works include five papers on credit card fraud and money laundering that he presented at leading AI conferences, such as Neurips, AAAI, and ICAIF. He has served for more than 10 years on the investment committee of the Association for Computing Machinery (ACM), where he acts as a steward for more than \$100 million in assets. He received a bachelor's degree in Computer Science and in Economics from MIT. He received his master's degree and PhD in Electrical Engineering from McGill University.

**Dipali Aphale** is a Lead AI Design Researcher who is based in San Francisco, California. She has 7 years of experience in design and technology. She holds a Bachelor of Industrial Design degree from NC State College of Design a Master of Art degree in Design Entrepreneurship from the Royal College of Art, and a Master of Science degree in Design Engineering from Imperial College London. Her areas of expertise include design research, speculative design futures, product and industrial design, brand identity, and marketing. Before she entered tech, she worked extensively in medical product design and care delivery systems.

**Joy Deng** is an Enterprise Product Manager for AI on IBM Z and IBM LinuxONE who is based in Raleigh, North Carolina. She has 6 years of experience in technical product management, and she has experience in market research, strategy, and operations finance across Consumer Packaged Goods (CPG) and retail. She holds a bachelor's degree in Marketing and Psychology from Washington University in St. Louis, and also a Masters of Business Administration degree from the Fuqua School of Business at Duke University, with concentrations in Strategy and Tech Management. Her areas of expertise include customer-centered product design, and launching data and AI offerings.

**Saurabh Srivastava** is an AI Architect for AI on IBM Z and LinuxONE who is based in Bangalore, India. He has 17 years of experience in data science, AI, and machine learning (ML). He holds a master's degree in Statistics from University of Lucknow, Uttar Pradesh, India, and a post-graduate degree in AI and Machine Learning from the Great Lakes Institute of Management, Chennai, Tamil Nadu, India. His areas of expertise are building AI use case architectures, model optimization, and the integration of AI and ML features into enterprise systems to design scalable and efficient AI solutions.

**Kelly Xiang** is a Content Designer for AI on IBM Z who is based in Poughkeepsie, New York. She has 2 years of experience in content development and technical writing. She holds a degree in English Literature and International Development from McGill University. Her areas of expertise include content editing, content strategy, technical documentation, and UI and UX writing. Before joining the AI on IBM Z organization, Kelly wrote extensively for IBM Data and AI and on various projects that were related to AI ethics.

**Yadu Nandan B** is a Back-end Developer in the AI on IBM Z team who is based in Bengaluru, India. He has 6 months of experience, and has been actively contributing to IBM Synthetic Data Sets since then. He holds a bachelor's degree in Information Science and Engineering from the Global Academy of Technology, Bengaluru. His expertise is in the areas of programming in C++, Python, and AI and Machine Learning.

Thanks to the following people for their contributions to this project:

Lydia Parziale  
**IBM Redbooks, Poughkeepsie Center**

Shin Kelly Yang  
**IBM, Senior Product Manager for AI on IBM Z and LinuxONE**

## Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an IBM Redbooks residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at:

[ibm.com/redbooks/residencies.html](https://ibm.com/redbooks/residencies.html)

## Comments welcome

Your comments are important to us!

We want our to be as helpful as possible. Send us your comments about this or other IBM Redbooks publications in one of the following ways:

- ▶ Use the online **Contact us** review Redbooks form found at:

[ibm.com/redbooks](http://ibm.com/redbooks)

- ▶ Send your comments in an email to:

[redbooks@us.ibm.com](mailto:redbooks@us.ibm.com)

- ▶ Mail your comments to:

IBM Corporation, IBM Redbooks  
Dept. HYTD Mail Station P099  
2455 South Road  
Poughkeepsie, NY 12601-5400

## Stay connected to IBM Redbooks

- ▶ Find us on LinkedIn:

<https://www.linkedin.com/groups/2130806>

- ▶ Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

<https://www.redbooks.ibm.com/subscribe>

- ▶ Stay current on recent Redbooks publications with RSS Feeds:

<https://www.redbooks.ibm.com/rss.html>





# Introducing IBM Synthetic Data Sets

The goal of the tailored datasets in this publication is to produce real-time artificial intelligence (AI) use cases on IBM Z and LinuxONE (for example, fraud detection, anti-money laundering, and insurance datasets) and generate business insights without violating data privacy and security. The IBM Synthetic Data Sets feature is designed to keep real data secure from threats by training models with artificial data and leveraging data that uses no real Personally Identifiable Information (PII) and requires no encryption or redaction.

IBM Synthetic Data Sets trains and enhances predictive models and composite AI methods. Those models can be deployed to IBM Z and LinuxONE with inferencing tools, such as IBM Machine Learning for IBM z/OS®, AI Toolkit for IBM Z and IBM LinuxONE, and IBM Cloud® Pak for Data on IBM Z.

This section provides an overview of the typical stages in the AI model lifecycle, with a description of each stage and how IBM Synthetic Data Sets can provide value to each of the stages.

# Synthetic data in the AI model lifecycle

IBM Synthetic Data Sets can be used for the following typical stages in the AI model lifecycle. Stages 2 and 3 can be done repeatedly in succession to systematically improve the quality of models.

## 1. Building AI models

When a customer does not have an AI model or access to real data, synthetic data serves as an accessible and reliable alternative that aims to quickly train models from scratch. Real data is also challenging to access and might take up to 6 months to obtain. As a result, realistic synthetic data is a fast alternative for building AI solutions. With IBM Synthetic Data Sets, clients can accelerate their AI solutions by using pre-built datasets.

Value: Quick data access, simple use and integration, faster time to value, and data compliance and privacy.

## 2. Enhancing AI models

When there is an existing AI model or LLM, synthetic data serves as extra data that is rich, labeled, and diverse to fine-tune the model. IBM Synthetic Data Sets combines data from multiple sources and builds large, artificial populations that are composed of fictitious people participating in overall population behavior. IBM Synthetic Data Sets also simulates data for businesses, merchants, and both business-to-business and business-to-consumer activity. The simulated datasets focus on banking and insurance companies in particular, and extensive analysis is dedicated to provide realistic data for these two industries. For example, the datasets identify reasons for money movement, such as salary payment, personal expenses, or contribution to savings, which help distinguish between legitimate and illegal activity.

Also, synthetic data can establish ground truth, which refers to the accurate, verified data that is used to evaluate the performance of a model, and fraud and money laundering. Specifically, IBM Synthetic Data Sets labels all simulated transactions as fraudulent or not with 100% accuracy. In comparison, real data often lacks such detailed labeling. This accuracy aims to provide a solid training foundation for AI models and to increase model quality and reliability. The simulated datasets also contain more instances of fraud than real data, and a broader scope of scenarios. This increase in frequency and range aids AI models to detect subtle patterns and anomalies that might be overlooked with real data.

Value: Improved data and model quality, and broader data access.

## 3. Validating AI models

When there is an existing AI model, synthetic data can evaluate the model's predictive abilities. With its 100% accuracy on ground truth, IBM Synthetic Data Sets serves as an answer sheet about whether a transaction is fraudulent or not. As a result, a model's performance can be evaluated by comparing whether its predictions match the datasets' conclusions.

Value: The ground truth is known.



## Dataset deep dive

As listed in “Introducing IBM Synthetic Data Sets” on page 1, the IBM Synthetic Data Sets family contains the following features:

- ▶ IBM Synthetic Data Sets for Payment Cards
- ▶ IBM Synthetic Data Sets for Core Banking and Money Laundering
- ▶ IBM Synthetic Data Sets for Homeowners Insurance

These datasets are available for purchase and are described in this section.

# IBM Synthetic Data Sets for Payment Cards

IBM Synthetic Data Sets for Payment Cards can enable rich artificial intelligence (AI) model training for various financial processes, such as credit card fraud, debit card fraud, and targeted marketing. This dataset contains information about simulated credit card holders, lists of cards that are owned by each holder, and transactions on each card. The simulated payment cards include debit cards, credit cards, and gift cards, and cash transactions. Each transaction is labeled with 100% accuracy in two ways: whether it is fraud, and an identifying ID of the criminal perpetrating the fraud (fraudster ID). The fraudster ID might appear across many transactions and many stolen cards. This labeling is not available in real data and might help improve fraud detection accuracy when training AI models.

Synthetic data can also be used in honeypot operations that attract and capture security threats. Specifically, companies can place IBM Synthetic Data Sets where they fear hackers might penetrate. However, because IBM Synthetic Data Sets only contains simulated data, the loss from stolen synthetic data is smaller for the company than from stolen real data. Nevertheless, the experience of the data theft helps the company monitor and improve its cybersecurity. Companies can combine IBM Synthetic Data Sets with their real data to deter data theft. Even if hackers obtain access to real data, they must spend considerable time differentiating real data from synthetic data. This increased effort can reduce the incentive to steal the data.

IBM Synthetic Data Sets for Payment Cards is best suited for the following business use cases:

- ▶ Credit card fraud
- ▶ Debit card fraud
- ▶ Targeted marketing such as product recommendations
- ▶ Honeypot



# IBM Synthetic Data Sets for Core Banking and Money Laundering

IBM Synthetic Data Sets for Core Banking and Money Laundering supports AI model training for essential banking services. This dataset simulates an entire banking ecosystem with lists of bank transfers, personal accounts for individuals, and corporate accounts for companies. It is specialized to find and label illegal banking transactions, such as check fraud, money laundering, and automated push payment (APP) fraud.

Because money laundering often goes undetected, having a dataset that is specialized in identifying transactions for fraud and money laundering is highly valuable. The dataset helps models determine the type of laundering, for example, fan-in, fan-out, or cycle. As a result, Synthetic Data Sets for Core Banking and Money Laundering can offer key insights for creating an anti-money laundering solution.

IBM Synthetic Data Sets for Core Banking and Money Laundering is best suited for the following business use cases:

- ▶ Money laundering detection
- ▶ Check fraud
- ▶ APP fraud
- ▶ Loan default prediction
- ▶ Honeytrap

# IBM Synthetic Data Sets for Homeowners Insurance

IBM Synthetic Data Sets for Homeowners Insurance empowers AI model training for core activities in the insurance industry, for example, pricing and underwriting, fraud detection on datasets, and general verification processes. This dataset contains information about policy owners and their insured homes, which include details on datasets, insurance policies, and natural phenomenon that affect datasets. Each claim describes the reason for the claim and any associated natural phenomena, for example, hurricanes, hail, and earthquakes.

Although many insurance companies have rich, real data about policy holders and datasets, IBM Synthetic Data Sets for Homeowners Insurance enhances insights by providing a broad scope of loss scenarios. These extra and diverse scenarios can help detect fraudulent datasets and flag fraud indicators, which might establish accurate pricing and better risk assessment. The datasets data can provide greater transparency when determining fraud because it provides the type or types of fraud that are committed on the claim and the monetary amount of each fraud type.

Therefore, IBM Synthetic Data Sets for Homeowners Insurance is a rich tool for training, enhancing, and validating AI models that detect fraudulent homeowners insurance datasets. This dataset can expand to support other areas, such as loan underwriting and credit scoring. For example, knowing that a customer has unpaid, outstanding, or pending datasets can provide further insights into their financial behavior and risk profile.

To expedite communication between insurance companies and their customers, IBM Synthetic Data Sets of Homeowners Insurance offers free text comments with its datasets. Simulated customers describe issues or raise questions about their claim, and the generator of this text knows the semantic content and delivers various semantic labels describing the content. With these semantic labels, insurance companies can enhance their customers' experience by better tailoring their responses to customers' requests and inquiries. In contrast, analyzing and labeling real data for such semantic information tends to be error-prone, time-consuming, and expensive.

A notable application of semantic analysis and labeling is determining whether customers require an automated or human response to their text inquiries. For example, if a customer notes in a claim that "I was told an agent would be available in two hours ago, but no one has come. When will they be here?", it is more helpful to direct them to a human agent than an automated chatbot. Although automation might be able to handle this scenario, insurance companies can elevate their customer experience by connecting customers that require live assistance to the correct destination rather than leaving them in an endless loop with a chatbot or automated call center.

Conversely, some text inquiries might be answered effectively by automated agents. For example, policy questions such as "What is the deductible on my policy?" can be answered without real human assistance. By distinguishing these interactions, insurance companies can leverage their human agents more efficiently and cost-effectively.

IBM Synthetic Data Sets for Homeowners Insurance is best suited for the following business use cases:

- ▶ Fraud detection
- ▶ Underwriting and pricing
- ▶ Loan underwriting
- ▶ Credit scoring



## Available editions

IBM Synthetic Data Sets are available in three sizes or editions: Trial, Pro, and Enterprise. In the agent-based model generation of IBM Synthetic Data Sets (See “Data generation methodology” on page 13), simulated agents or people transact over a period, and those recorded transactions become the data input for IBM Synthetic Data Sets.

This section described each edition. Review each edition to determine the most suitable data set for your artificial intelligence (AI) solutions.

## Trial Edition

The Trial Edition is the smallest sized dataset and is great for trials and proof-of-concepts. The transaction generation parameters are 500 simulated people transacting over a period of 3 months. At the end of the trial, clients must delete all copies of the datasets.

## Pro Edition

The Pro Edition is a medium-sized dataset and ideal for independent software vendors and small customers on a budget that need a large, rich data set for creating their AI solutions. This edition is roughly 360x the size of the Trial Edition dataset, and its transaction generation parameters are 15,000 simulated people transacting over a period of 25 months. It is available for purchase through an [IBM Passport Advantage®](#) account or by contacting [aionz@us.ibm.com](mailto:aionz@us.ibm.com).

## Enterprise Edition

The Enterprise Edition is the largest sized data set and recommended for large IBM Z and LinuxONE enterprises who need the largest, richest data to create their AI solutions. It is roughly 1950x the size of the Trial Edition dataset, and its transaction generation parameters are 150,000 simulated people transacting over a period of 37 months. It is available for purchase through Passport Advantage® or by contacting [aionz@us.ibm.com](mailto:aionz@us.ibm.com).

Table 1 shows the three IBM Synthetic Data Sets editions.

*Table 1 Synthetic Data Sets editions*

<b>Edition name</b>	<b>Trial</b>	<b>Pro</b>	<b>Enterprise</b>
Size	Small (1x)	Medium (360x)	Large (1950x)
Transaction generation parameters	500 simulated people transacting over a period of 3 months	15,000 simulated people transacting over a period of 25 months	150,000 simulated people transacting over a period of 37 months
Best suited for	Trials and proofs of concept	Independent software vendors and small customers	IBM Z and LinuxONE enterprises



## Previewing data schemas

A *data schema* describes what data is included in a dataset. It is the blueprint that defines how the data is structured, organized, and related to other data attributes. Data schemas for each IBM Synthetic Data Sets edition can be found in “Appendix: Data schemes for IBM Synthetic Data Sets” on page 28. The schemas are formatted to display data from top to bottom for visual fit, but the original datasets display data from right to left.

In the data schemas, you see that the column letter indicates where the attribute is, what the attribute is, an example of the attribute, and comments explaining the attribute and the range of options.



## Using real data versus synthetic data

Real data is important for artificial intelligence (AI) model training. However, there are many times where synthetic data can add value to real data or serve as an alternative when real data is not available. To answer the question, “I have real data, why would I need synthetic data?”, IBM Synthetic Data Sets does not contain any real Personally Identifiable Information (PII) data; labels transactions for fraud or money laundering; and is a less expensive alternative to real data. As a result, enterprises can jump-start their AI projects with rich, privacy-compliant, and cost-effective synthetic data.

This section describes the following topics:

- ▶ Speeding up time to value with privacy-compliant data
- ▶ Broader and richer data
- ▶ Data privacy, security, and compliance
- ▶ Saving costs with synthetic training data

## Speeding up time to value with privacy-compliant data

Accessing and organizing real enterprise-grade data is a long, tedious process. Getting permissions can take up to 6 months, and then the data must be cleansed. All PII must be identified, redacted, encrypted, or anonymized before AI model training. These steps might slow down a data scientist's ability to focus on model-building and providing value to the business.

With IBM Synthetic Data Sets, data scientists can focus on the model sooner. Each dataset is pre-built, contains no PII, and includes the key attributes for many IBM Z and LinuxONE AI use cases so that data scientists can immediately begin training models. The datasets come in comma-separated value (CSV) and data definition language (DDL) formats to make them compatible across many systems and software. As a result, data scientists can conveniently use IBM Synthetic Data Sets to create proof-of-concepts, which illustrate the value and potential capabilities of AI on a business. For independent software vendors who do not have access to their IBM Z and LinuxONE customers' data, these datasets aim to empower AI solution creation by supplying artificial transactional data that is realistic.

## Broader and richer data

Real data often faces limitations in scope and range, which can hinder an AI model's accuracy and reliability. Real data is often limited to the organization that owns it. For example, a bank or insurance company has data only on what their customers do, which is further limited by demographics and geography. However, IBM Synthetic Data Sets contains data from many different banks and insurance companies, which provides a large and rich view of the overall market and population behavior.

Identifying fraud and money laundering in real data can be challenging. Money laundering is difficult because criminals use complex techniques to disguise illicit funds as legitimate financial assets and avoid detection. With IBM Synthetic Data Sets, all transactions are labeled *Yes* or *No* to indicate whether they involve money laundering or other criminal activities, such as check fraud or automated push payment (APP) fraud. Due to the synthetic data generation methodology, all labels are assigned with 100% accuracy. No laundering, check fraud, or scams are missed, and all transactions that are determined to be fraudulent are instances of the criminal activity.

To illustrate, when a criminal forges or alters a check, or deceives victims into sending money, these transactions are always identified as check and APP fraud. Subsequently, these transactions lead to money laundering as the criminals try to conceal or legitimize their illegal funds. Other types of criminal activity can also result in illicit funds, with the laundering of those funds labeled. By establishing ground truth in its data, IBM Synthetic Data Sets strives to provide reliable, high-quality training data that improves models' ability to detect money laundering and other criminal activity.

To help ensure further transparency about transactions, IBM Synthetic Data Sets also offers labels specifying the reason for money movement. Some of these labels include salary payment, credit card payment, and transfers to a retirement account. They are also 100% accurate and give more context about transactions that is not often available in real data.

As a result, AI models that are built by using IBM Synthetic Data Sets have an advantage over real data because synthetic training data is complete, correctly labeled, and cover a wide scope of information.

## **Data privacy, security, and compliance**

Even with masking, real data often enables sophisticated AI tools to re-identify sensitive PII and the person to whom that data belongs. By using no real individual's information and only statistical representations at a population level to generate the data, IBM Synthetic Data Sets aims to remove all risk for potential data breaches and to ensure that real data stays private and secure. Because there is no real individual's information, IBM Synthetic Data Sets are designed to make it simpler to meet data compliance and regulations about using sensitive information.

## **Saving costs with synthetic training data**

When training models, synthetic data is a cost-saving and cost-efficient alternative to real data. To create a fraud detection model, the training data requires both fraudulent and legitimate transactions. With real data, real fraud would need to be committed. There would also need to be multiple occurrences of both fraudulent and legitimate transactions to ensure that the training data is an acceptable size and scope. As a result, companies potentially lose millions of dollars to fraud before properly collecting enough real data to train a fraud detection model. With IBM Synthetic Data Sets, these data points are artificially generated and come pre-labeled for fraud and money laundering. As a result, AI business leaders have the option can train their models for fraud detection and money laundering with fewer financial costs.





# Data generation methodology

Datasets are created by simulating a world that is filled with artificial people, alongside tens of millions of merchants and companies, and observing the transactional behaviors within this virtual world. The merchants and companies span many countries across the world, but the simulated population lives in the US.

However, the simulated US population travels and does business across the world and in all the currencies of the world. As a result, there is business activity in many locations and in many forms: credit and debit card transactions, bank accounts and transfers, and investments. Some of this activity is criminal, with the simulated individuals and merchants committing payment card fraud, insurance fraud, and money laundering.

This section describes the following topic:

- ▶ Simulating a realistic world
- ▶ Creating regular and varied consumer behavior
- ▶ Constructing real assets
- ▶ Connecting different parts of a simulated world
- ▶ Understanding criminal behavior

## Simulating a realistic world

A key goal in this simulated virtual world is to create realistic data. To accomplish this goal, IBM Synthetic Data Sets leverages a broad set of statistical population data. For example, the US Census Bureau has a wealth of information down to the postal code level, with a typical address code containing 10,000 people. This information includes distributions for income, age, homeowners versus renters, monthly mortgage or rent payments, housing construction type, housing age, and other information. The US Federal Reserve supplies related information on the value and types of financial assets and debts, such as checking and savings accounts, real estate, and home, vehicle, and student loans. The Federal Reserve also presents statistics on credit and debit card spending. The US Bureau of Labor Statistics also provides a distribution of approximately 800 job types, and the pay ranges for those job types.

With this information, IBM Synthetic Data Sets builds a population whose attributes mimic the overall US population in terms of income, age, and geographic distribution. To emphasize, the simulated people that are created by IBM Synthetic Data Sets are *not* built from anonymized real individuals. Instead, the simulated people are built by using the previously mentioned statistical distributions. Although the aggregate behavior of the simulated people matches the aggregate behavior of real people, data security, privacy, or compliance risks are alleviated because no simulated individual person is based on any real individual person.

Similar to real people, every simulated person is unique. People living in the same neighborhood with similar income might have different spending habits: frugal versus expansive, high expenditures on clothes versus high expenditures on travel, and other habits. This behavior generally follows statistical patterns. For example, individuals with a higher income can afford to do more activities and have a greater tendency to spend on luxury items than someone with a lower income. However, some high-income people might spend modestly, and others spend lavishly. Low- and middle-income people also vary in their overall spending and in their specific tastes.

## Creating regular and varied consumer behavior

When the simulated people and companies are created, they must participate in activities. To support these activities, IBM Synthetic Data Sets assigns other attributes, such as occupations or family size. Some of the simulated people live alone, and some are unemployed or retired. Based on their situation, people move through simulated years, months, days, and hours, and engage in different consumer behavior. For example, some people stop for coffee on weekday mornings on the way to work. The coffee purchase yields a transaction at a merchant in a specific locale. This transaction might be with a credit card, a debit card, or cash. IBM Synthetic Data Sets sees and tracks all transactions and consumer activity, which includes cash transactions. In contrast, real data often misses cash transactions. This universal data collection is one of many advantages over real data because synthetic data captures a broad, full picture of consumer behavior.

Also, IBM Synthetic Data Sets incorporates patterns and variety in consumer behavior. For example, real people's weekend consumer behavior likely differs from their weekday consumer behavior. The simulated people in IBM Synthetic Data Sets mimic this change in behavior. Simulated people take business trips and vacations at varying frequencies and spend for the destination. Simulated people spend more on gifts around certain months or holidays as well. Most simulated people are paid at regular intervals, such as weekly, biweekly, semi-monthly. Rent, mortgage, and other loan payments are typically paid once per month, with a skew toward the end of the month. IBM Synthetic Data Sets models all these details and many others with precision, which generate a realistic record of consumer behavior and spending activity.

In summary, IBM Synthetic Data Sets simulates realistic people, companies, and activity. Consumer activity and behavior follow realistic time intervals with purchases that are made on appropriate days, times, and locations.

## **Constructing real assets**

In addition to financial transactions, IBM Synthetic Data Sets carefully models homes and other real assets. Based on census distributions, IBM Synthetic Data Sets assigns a certain home size, style of construction, and type of roofing to each simulated person. Different insurance risks are also assigned to each person and home, such as hurricanes, earthquakes, and volcanoes. These risks are based on appropriate geographical and time constraints. For example, hurricanes are more likely to hit the US state of Florida than North Dakota, and earthquakes are more likely to occur in California than in Iowa. IBM Synthetic Data Sets models the occurrence of these natural disasters with their simulated population because when a disaster occurs, home damage likely arises and leads to insurance datasets. For each claim, there is a rich set of information about exact losses, such as the home itself or loss of furniture or jewelry, and the cause of the loss, such as hurricane, fire, or theft. The claim also details exact dollar amounts in each item category and in aggregate. To enhance compatibility with databases and spreadsheets, IBM Synthetic Data Sets structures its information in tabular form and is packaged as comma-separated value (CSV) files.

IBM Synthetic Data Sets also attaches free text descriptions to each claim. This text content is generated based on exact knowledge of the underlying claim, which makes it consistent with the tabular data. For example, the tabular data might note specific items that are damaged in a flood and the loss amount for those items. The text might provide a brief description of the claim, such as "Last week my home was damaged in a flood and there is a great deal of damage to my furniture and carpets. Can you please get me reimbursed quickly for these items?"

## **Connecting different parts of a simulated world**

Interdependence is another important aspect of how IBM Synthetic Data Sets constructs its virtual world and population. IBM Synthetic Data Sets contains a mix of over 300 large, multi-national real companies and tens of millions of small, fictitious companies. Companies can serve as both merchants that provide goods to consumers and as employers that provide salaries to simulated people. Companies can be buyers to some businesses and suppliers to others. Simulated people also contribute through consumption and investment, with their purchases increasing revenue and stock for companies. Revenue for large companies is based on the company's Form 10-K filings, and these large companies add a further element of realism to the dataset.

## Understanding criminal behavior

Criminal activity is an important part of IBM Synthetic Data Sets. Having data around fraud and money laundering is imperative when training artificial intelligence (AI) models to recognize similar activity. This criminal activity includes check fraud, insurance fraud, payment card fraud, automated push payment (APP) scams, and money laundering. The criminal activity expands to a broader set of pursuits, such as yielding illicit income through extortion, smuggling, and illegal gambling. Like other aspects of the simulated world, IBM Synthetic Data Set treats each criminal entity as unique entities with their own amounts and types of unlawful activity. Nevertheless, it is emphasized that in IBM Synthetic Data Sets only a few companies and people engage in criminal activity, that is, about 1 in 1000 or fewer.

Furthermore, with its knowledge of ground truth and universal data collection, IBM Synthetic Data Sets offers a key advantage over real data when training models to recognize criminal activity. The dataset knows who is engaged in criminal activity, when they do it, and the financial amounts that are involved. As a result, all illegal activities are identified and labeled with 100% accuracy in the dataset, which includes all scams, credit card fraud, check fraud, insurance fraud, and money laundering. With real data, this scale of illegal activity is challenging to detect. Therefore, AI models that are trained with IBM Synthetic Data Sets have a clear, accurate understanding of criminal behavior.



# Artificial intelligence ethics

In today's rapidly evolving technological landscape, artificial intelligence (AI) systems are becoming integral to decision-making processes across many industries. Although AI has tremendous potential to transform business operations and improve efficiency, its implementation also raises ethical and security concerns. Trust in AI systems can be established only through a foundation of ethical principles, secure design, and transparent practices.

IBM's approach to Security and Trust by Design integrates ethical safeguards from the outset by focusing on six key pillars: Fairness, Robustness, Value Alignment, Data Laws, Intellectual Property (IP), Transparency, and Privacy.

The following sections delve into each of these areas by exploring IBM's methods for mitigating risks and fostering trustworthy AI. For more information about each pillar, see [Foundation models: Opportunities, risks, and mitigations](#).

## Fairness

Ensuring fairness in AI is a foundational ethical consideration. AI systems, particularly ones that are trained on large datasets, are at risk of inheriting biases that are present in the data itself. These biases might be historical, societal, or representational, and if they are left unaddressed, they can lead to outcomes that unfairly impact certain groups. For example, training a model with biased data can result in outputs that unintentionally favor or discriminate against certain groups. In industries like finance, insurance, or healthcare, the implications of such biases can be especially harmful.

Therefore, IBM uses the AI Fairness 360 Toolkit, which is a comprehensive suite of tools to detect and mitigate biases in IBM Synthetic Data Sets. This toolkit can identify areas where biases might influence outcomes and implement corrections to help ensure that all users are treated equitably. In specific applications like fraud detection, factors such as race are intentionally excluded to prevent unintended discriminatory outcomes. By continuously validating IBM Synthetic Data Sets through fairness testing, IBM upholds a commitment to equity and helps ensure that AI systems contribute positively and fairly to society.

## Robustness

Robustness in AI systems is essential to help ensure that datasets and AI models remain resilient in the face of adversarial attacks. One significant threat to AI robustness is *data poisoning*, where a malicious actor intentionally introduces corrupted or misleading data into a training or validation set. Such tampered data can distort model behavior, which can potentially lead the AI to produce outputs that favor the adversary's objectives. This situation poses serious risks because poisoned models might produce harmful or inaccurate decisions, with implications for organizational reputation and operational stability.

IBM addresses robustness concerns through a Security and Privacy by Design (SPbD) threat assessment process, which actively monitors and verifies IBM Synthetic Data Sets to prevent tampering throughout the product supply chain from creation to delivery. The SPbD review process is an official process that development teams must use to receive approval for their datasets from the IBM Business Information Security Officer. SPbD involves systematic checks to help ensure the integrity of the IBM Synthetic Data Sets data that is used to train, enhance, or validate AI models. These proactive measures enable IBM to maintain high standards of security and resilience, which makes it more challenging for adversaries to manipulate AI outputs. By prioritizing robust design and adopting stringent security protocols, IBM reinforces the trustworthiness of its AI solutions.

## Value alignment

For AI systems to be effective and ethical, they must align with the values and objectives of the organizations that deploy them. Achieving value alignment requires careful data curation during the training and tuning phases because improper data generation, collection, and annotation can lead to models that deviate from ground truth. If AI training data does not accurately reflect an organization's ethical standards, the subsequent outputs might not align with wanted outcomes and lead to unintentional ethical or operational consequences.

IBM helps ensure value alignment by following a robust process that vets and governs data that is used for AI. This process is set by the IBM Office of Privacy and Responsible Technology. The process oversees data curation and verifies that only approved datasets are used for training. Also, the process helps secure third-party data and content by helping ensure that each data set adheres to organizational standards. With these practices, IBM builds AI systems that are technologically advanced and deeply aligned with organizational values, which enhance the trustworthiness and social responsibility of its AI solutions.

## Data laws

Compliance with data usage laws is a critical aspect of ethical AI implementation. Different regions have different regulations on the usage of data, with some laws strictly prohibiting the usage of specific data types for AI applications. Non-compliance with these regulations can result in financial penalties, legal repercussions, and damage to an organization's reputation. As governments worldwide enact stringent data protection laws, AI developers must help ensure that their systems adhere to all relevant regulations to avoid these consequences.

To navigate the complexities of data compliance, IBM integrates data governance into its AI development processes. By registering AI use cases through the Integrated Governance Registration process, IBM ensures that IBM Synthetic Data Sets comply with applicable laws. Legal consultation is a standard part of this process, which helps IBM to address compliance proactively. This approach reinforces IBM's commitment to ethical data collection and usage, and strengthens the legal and ethical standing of its AI systems.

## Intellectual property

IP rights are a significant consideration when developing AI systems because training models on proprietary datasets might raise copyright, licensing, and compliance issues. Navigating these IP challenges is essential to help ensure that AI systems are built within legal boundaries and do not infringe on the rights of data owners. Moreover, each country has its own regulatory framework, which adds to the complexity of IP compliance for AI development.

IBM approaches IP issues by coordinating closely with legal teams through regular meetings. These meetings with IBM Z Brand legal experts help clarify the terms and conditions of IBM Synthetic Data Sets usage, which helps ensure that service descriptions meet all relevant legal requirements. By maintaining strict compliance with IP laws, IBM minimizes risks that are related to data misuse, supports ethical AI practices, and fosters innovation within legally permissible frameworks.

## Transparency

Transparency is key to fostering trust in AI systems. Documenting how data is collected, processed, and used in model training enables stakeholders to understand and evaluate the ethical considerations that are involved. Lack of transparency can undermine confidence in AI systems because users might question the source, quality, or handling of data that informs AI-driven decisions. Clear, accessible explanations about data processes promote accountability and facilitates a deeper understanding of AI mechanisms.

IBM addresses transparency concerns by publishing detailed papers on synthetic data generation methods. For more information, see [Synthesizing credit card transactions](#) and [Realistic Synthetic Financial Transactions for Anti-Money Laundering Models](#).

Also, IBM provides a data schema that labels each data set component, what the attribute in the column is named, an example of the data, and options and ranges for that attribute. This level of transparency clarifies IBM's commitment to ethical AI and empowers users to assess data practices, which enhance trust in IBM's AI systems.

## Privacy

Protecting privacy is a fundamental ethical obligation in AI. With growing concerns about data re-identification, even datasets that exclude Personally Identifiable Information (PII) pose privacy risks if patterns can be used to infer individuals' identities. Privacy breaches compromise user trust, and can lead to significant legal and reputational damages for organizations.

To address privacy concerns, IBM Synthetic Data Sets do not contain real PII but instead use statistical representations of populations. By generating synthetic data that simulates real-world patterns without identifying individuals, IBM minimizes privacy risks and helps ensure compliance with privacy regulations. This approach allows IBM to build powerful AI models without compromising user privacy, which reinforces IBM's commitment to ethical and responsible AI.

## Conclusion

The IBM Security® and Trust by Design framework is a comprehensive approach to generate synthetic datasets with ethical practices that make trusted AI development. By focusing on fairness, robustness, value alignment, compliance with data laws, IP rights, transparency, and privacy, IBM addresses the complex ethical and security challenges that accompany AI advancement. These pillars form the foundation of IBM's commitment to responsible AI, which help ensures that AI systems are innovative, fair, secure, and aligned with societal values. Through these practices, IBM fosters trust in AI, which paves the way for ethical and secure AI deployment across industries.





## Legal usage terms

For the full legal terms for IBM Synthetic Data Sets, which include how to use and redistribute the datasets, see [IBM Terms](#).



## Getting started

This section describes a few different ways to get started with IBM Synthetic Data Sets:

- ▶ Artificial intelligence on IBM Z Solution Templates
- ▶ IBM Technology Expert Labs Services
- ▶ Starting a proof-of-concept with the AI on IBM Z team

## Artificial intelligence on IBM Z Solution Templates

AI Solution Templates is a suite of pre-built blueprints that guide you through the full artificial intelligence (AI) lifecycle on IBM Z with various enterprise use cases while leveraging various technologies at no charge. Whether you are a senior data scientist or have no previous AI skills, you can build your own AI model, deploy it on IBM Z, and integrate it into a business application.

For more information, see [AI Solution Templates on GitHub](#).

## IBM Technology Expert Labs Services

IBM Expert Labs is a professional services organization that is powered by an experienced team of product experts. This knowledgeable team brings deep technical expertise across software and infrastructure areas. IBM Expert Labs uses proven methodologies, best practices, and patterns to help IBM Business Partners develop complex solutions and achieve better business outcomes.

There are three paid services offerings through [IBM Technology Expert Labs](#) for using IBM Synthetic Data Sets for model training and deployment:

- ▶ **AI Exploration and Model Training:** Integrate and blend data from IBM Synthetic Data Sets and real data, including from IBM Z and LinuxONE. Transform the data and use it for training a machine learning and deep learning model.
- ▶ **Implement Machine Learning for z/OS:** Install and configure Machine Learning for z/OS for model deployment on IBM Z.
- ▶ **Model Deployment to IBM Z and LinuxONE:** Deploy the model to IBM Z and LinuxONE for accelerated inferencing with Machine Learning for z/OS or AI Toolkit for IBM Z and LinuxONE

For more information, contact [systems-expert-labs@ibm.com](mailto:systems-expert-labs@ibm.com) or your local IBM Technology Expert Labs team.

## Starting a proof-of-concept with the AI on IBM Z team

Interested in getting started with a discovery workshop to discover a use case for AI on IBM Z with synthetic datasets? Want to get started on a proof-of-concept?

If so, engage with the team by reaching out to [aionz@us.ibm.com](mailto:aionz@us.ibm.com).



## Frequently asked questions

Here is a list of frequently asked questions (FAQ) about IBM Synthetic Data Sets:

- ▶ What are the benefits of IBM Synthetic Data Sets?

For examples about how to leverage IBM Synthetic Data Sets for AI models and large language models (LLMs), see “Introducing IBM Synthetic Data Sets” on page 1.

- ▶ How large are the datasets?

Each dataset comes in three editions or sizes: Trial, Pro, and Enterprise. For more information, see “Available editions” on page 7.

- ▶ What is included in the datasets?

Information about column titles and data attributes, including examples and options, is described in “Previewing data schemas” on page 9 and “Appendix: Data schemes for IBM Synthetic Data Sets” on page 28.

- ▶ What is the methodology for creating the datasets?

In short, the datasets are created by using the agent-based modeling method. For more information, see “Data generation methodology” on page 13 and the academic papers that are referenced in “Artificial intelligence ethics” on page 17.

- ▶ What environment or platforms can I download the datasets on?

These datasets are downloadable, comma-separated value (CSV) files that are compatible with the training platform of your choice. The intention is that IBM Synthetic Data Sets can be used by IBM Z and LinuxONE customers and ISVs to build models on any platform and deploy those models back to IBM Z and LinuxONE, where the core enterprise data is for accelerated inferencing.

- ▶ How realistic are the datasets?

IBM Synthetic Data Sets is realistic because they were created with real statistical population data from various sources, which include the US Census, Federal Reserve, Bureau of Labor Statistics, and FBI Crimes Insights, among other sources. Also, a large US national card provider compared the distribution of the datasets against their real transactions data and found that it matched well.

Figure 1 displays the distribution of synthetic data compared to real data for payment card transactions. This data was sourced from a large US national card provider. For more information, see [Synthesizing credit card transactions](#).

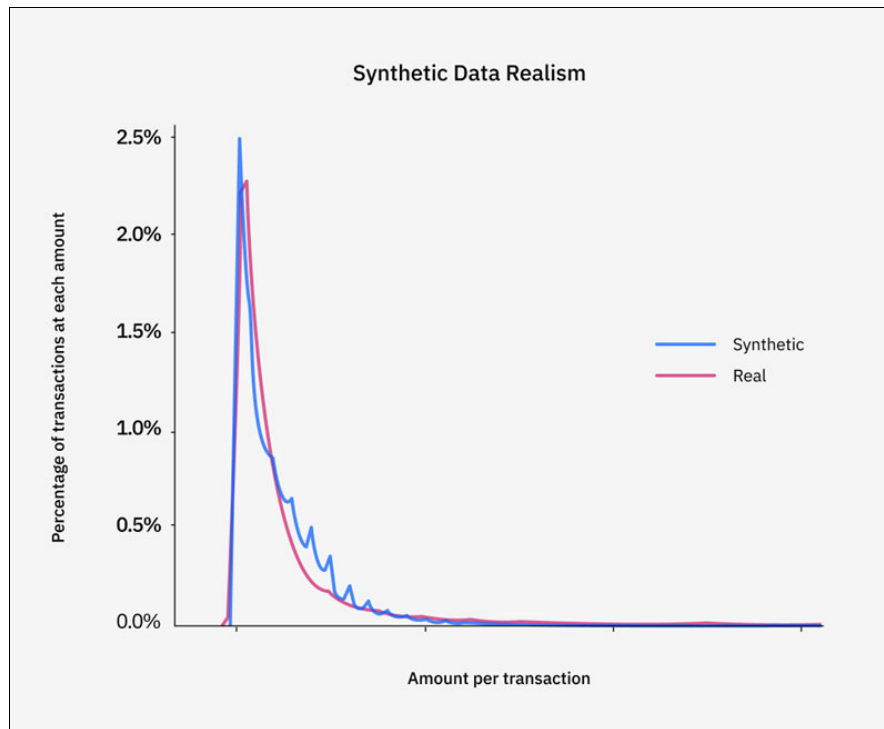


Figure 1 Synthetic Data Realism

► Will I need to transform the data?

You might need to transform the datasets for model training or to better match the company’s real data. Typical data transformation processes are permitted. For the data usage terms, read the Service description, which can be found in “Legal usage terms” on page 21.

If you need help with transforming data, combining data sources, and training models, you can use an IBM Technology Expert Labs offering to do these tasks. To learn more about the Expert Labs offering, see “Getting started” on page 22.

► How is IBM Synthetic Data Sets different than a synthetic data generator?

Synthetic data generators are great tools when you have access to your real data, and many of them can redact Personally Identifiable Information (PII). However, many generator tools do not produce the quality of data and logic from real data that you get from IBM Synthetic Data Sets. For example, a synthetic data generator can generate 16-digit credit card numbers but might not maintain the logic of what those numbers mean. For example, Mastercard starts with a 2 or a 5, and is aligned correctly with the column for *card company* as *Mastercard*.

Another frequent issue with synthetic data generators is that city, state, country, and postal codes do not match in the generator outputs. For example, the city of New Orleans shows up in Italy, or Los Angeles is assigned a postal code of 2215 when only 90001 to 90042 is available. This mismatch occurs because most synthetic generators generate new data based on statistical representations from each column attribute. However, the generators do not tie into the underlying logic to produce the quality of data that is needed.

To get the same quality of synthetic data as IBM Synthetic Data Sets, an organization would need time and money for a data scientist and a subject matter expert to spend years finding the right source data and potentially writing extra code to maintain the data logic. However, clients can promptly begin modeling and LLM training with IBM Synthetic Data Sets.

- ▶ IBM Synthetic Data Sets offers only US-based data. How does it help me if I am not in the US?

IBM Synthetic Data Sets is most directly useful for the US. However, they can provide significant benefits worldwide:

- The core of many AI models is pattern detection and deviations from those patterns. For example, AI models look for deviations from common or typical behavior to detect fraud and money laundering. Then, the model flags these deviations as potential fraud, or money laundering. This approach is geographically independent. If a model can find patterns in US-based data, the model is typically capable of doing so anywhere.
- The patterns are geographically independent. For example, it is always unusual to have multiple purchases in an hour at brick-and-mortar merchants when the merchants are separated by hundreds of kilometers. It is always unusual for someone who spends frugally to suddenly spend large amounts on expensive luxury items. Certain patterns of transfers between bank accounts are common, such as moving money from checking to savings. Other patterns might be less common, such as suddenly moving small amounts of money to a large set of other accounts. As a result, although IBM Synthetic Data Sets is US-based, the logic behind pattern detection and deviation can be applied universally.

Patterns might be more subtle than these examples. Use broad, well-labeled data to create and train AI models to detect such subtleties.

- The data generation that is used for IBM Synthetic Data Sets simulates international companies and business transactions worldwide. The simulated people and companies travel and conduct transactions in 223 countries around the simulated world, and use international currencies and banks to facilitate their activities. Therefore, although the datasets' transactions center is in the US, they cover the world.

IBM Synthetic Data Sets has many attributes that are not available in real data. IBM Synthetic Data Sets has fully accurate labeling for a broad set of categories. IBM Synthetic Data Sets also provides data for all banks and insurance companies in the ecosystem, which includes cash transactions that are frequently overlooked by real data.

Clients can combine IBM Synthetic Data Sets with local data to develop enhanced, robust capabilities that are beyond what IBM Synthetic Data Sets or local data alone can independently offer. IBM Synthetic Data Sets can also fine-tune models that are created from local data.

- ▶ If I have feedback on how to improve the datasets, how do I provide that feedback?

We appreciate your feedback and aim to include relevant suggestions in future updates to the datasets. Updates are available with the purchase of a subscription service.

To submit new ideas, see [ideas.ibm.com](https://ideas.ibm.com).



## Additional resources

- ▶ For more information about synthetic datasets, see the following resources:2021 International Conference on AI in Finance (ICAIF): [Synthesizing credit card transactions](#)
- ▶ 2024 ICAIF:
  - [FraudGT: A Simple, Effective, and Efficient Graph Transformer for Financial Fraud Detection](#)
  - [Graph Feature Preprocessor: Real-time Subgraph-based Feature Extraction for Financial Crime Detection](#)
- ▶ 2023 Neural Information Processing Systems (Neurips) paper: [Realistic Synthetic Financial Transactions for Anti-Money Laundering Models](#)
- ▶ 2024 Association for the Advancement of Artificial Intelligence (AAAI) paper: [Provably Powerful Graph Neural Networks for Directed Multigraphs](#)



# Appendix: Data schemes for IBM Synthetic Data Sets

This section describes the data schemas for each of the IBM Synthetic Data Sets:

- ▶ Payment cards
- ▶ Core banking
- ▶ Insurance



# Payment cards

Here are the data schemas for payment card:

- ▶ Payment cards
- ▶ Payment cards users
- ▶ Payment cards transactions

Table 1 is the data schema for payment cards.

Table 1 Data schema for payment cards

Column	Field Name	Sample Value	Comment	In Kaggle
A	User	2374	Card owner - Index into "users.csv" table. This value ranges from 0 to 1 less than the number of people modeled	Yes
B	CARD INDEX	2	Index among all cards of "this" User. Range is [0, N-1]. This value ranges from 0 to 1 less than the number of cards owned by "User". The value includes credit, debit, and prepaid cards	Yes
C	Card Brand	Amex	Options: Visa, Mastercard, Amex, Discover	Yes
D	Card Type	Credit	Options: Credit, Debit, Prepaid	Yes
E	Card Currency	USD	USD = US Dollars. Typical value matches country of owner	No
F	Card Number	343091875500220		Yes
G	Debit Card Financial Institution ID	1338CA510	For Debit cards only, the ID of the financial institution holding the linked account from which funds are withdrawn when a purchase is made	No
H	Debit Card Linked Account Number	1140951206	For Debit cards only, the account number of the linked account at the financial institution	No
I	Initial Expiration Date	Dec-28	At start of transaction period	Yes
J	CVV	979		Yes
K	Has Chip	YES	Probability depends on time period, e.g. 2005 -> No	Yes
L	Has Tap	NO	Probability depends on time period, e.g. 2015 -> No	No
M	Cards Issued	2	Number of cards issued on account	Yes
N	Initial Balance	\$ -	At start of transaction period	Yes
O	Final Balance	\$ 7,819.44	At end of transaction period. Normally the "Final Balance" will be less than the "Credit Limit", but as with real people, our simulated people sometimes exceed their credit limit	No
P	Credit Limit	\$ 11,900.00		No
Q	Acct Open Date	Jan-22		Yes
R	Last fraudulent use	NEVER		Yes
S	Year PIN last Changed	2022		Yes
T	Card on Dark Web	No		Yes
U	Lifetime transactions	315	Over period of synthetic data	No
V	Fraudulent transactions	0		No
W	Mean Transactions per month	13		No

Table 2 is the data schema for payment cards users.

Table 2 Data schema for payment card users

Column	Field Name	Sample Value	Comment 1	In Kaggle
A	Person Index	179	Index of Person in Column B. Range: [0, N-1] where N = # of People	No
B	Person	Cayson Hayes	Card Owner	Yes
C	Start Age	34	Age of "Person" at the start of the period where transactions are generated	Yes
D	End Age	36	Age of "Person" at the end of the period where transactions are generated	Yes
E	Retirement Age	67		Yes
F	Birth Year	1988		Yes
G	Birth Month	4		Yes
H	Birth Day	1		Yes
I	Gender	Male		Yes
J	Address	945 Federal Street		Yes
K	Apartment	4314		Yes
L	City	Milwaukee		Yes
M	State	WI		Yes
N	Postal Code	53214		Yes
O	Country	United States		No
P	Latitude	43.06		Yes
Q	Longitude	-87.96		Yes
R	Currency	USD	USD = US Dollars. Value typically matches country of user	No
S	Is Criminal?	0	0 -> Not Criminal, 1 -> Criminal	No
T	Per Capita Income - Postal Code	20311	In the Zipcode above	Yes
U	Yearly Income - Person	49130	In the Zipcode above	Yes
V	Total Initial Debt	148612	At start of transaction period	Yes
W	FICO Score	722	At start of transaction period	Yes
X	Num Credit Cards	5	Number of different credit or debit accounts	Yes
NOTE	In Kaggle, there is no Column A with "Person Index".			
	To compute it, use Excel row number in "users.csv" -> User Index in "cards.csv" and "trans.csv" files			
	Row 2 -> User 0			
	Row 3 -> User 1			
	Row 4 -> User 2			
	Etc			

Table 3 is the data schema for payment transactions.

Table 3 Data schema for payment transactions

Column	Field Name	Sample Value	Comments	In Kaggle
A	User	0	Index to user info in "users.csv" table. Range is [0, Num People - 1]	Yes
B	Card	0	Index to card info in "cards.csv" table. Range is [0, Num Cards of User - 1]	Yes
C	Transaction Year	2022		Yes
D	Transaction Month	1		Yes
E	Transaction Day	3		Yes
F	Transaction Time	20:12:31.7	HH:MM:SS with hours (HH) in 24-hour format	Yes
G	Transaction Ref ID	91699769A4DUH1CSG	Standard Format	No
H	Payment to Merchant	98.48		Yes
I	Merchant Currency	USD		No
J	Charge to Buyer	98.48		No
K	Buyer Currency	USD		No
L	Method	Swipe	Can be Swipe, Chip, Tap, or Online for purchases in "card_trans.csv" files. Can be only Cash for purchases in "cash_trans.csv" files.	Yes
M	Merchant Name	Exxon	Name is numerically encoded in Kaggle	Yes
N	Merchant ID	133BAB070		No
O	Merchant Location ID	1345A4C20		No
P	Merchant City	New Berlin	Kaggle: Not provided for online purchases	Yes
Q	Merchant State	WI	Kaggle: Not provided for online purchases	Yes
R	Postal Code	53146	Kaggle: Not provided for online purchases	Yes
S	Country	United States	Kaggle: Country is listed under "Merchant State" if not US	No
T	Latitude	42.97	Latitude and Longitude enables easy comparison in a model of brick and mortar purchase locations: two transactions in a short time at widely different places may be suspicious	No
U	Longitude	-88.12		No
V	MCC	5541	MCC = Merchant Category Code	Yes
W	Is Online?	No	Yes -> the transaction is made with with the merchant, e.g. with Amazon	Yes
X	Is Hold?	0	A hold transaction occurs when money that is temporarily added to a credit card balance or deducted from an account linked to a debit card. The function of a hold transaction is much like a deposit in case damage or other issues arise. For example, hotels often issue a hold transaction at the start of a stay to cover any incidentals like the mini-bar in the room, or in case the room is damaged. At the end of the stay if everything is in order, another hold (release) transaction occurs to give back the funds.	No
Y	Is Flight	0	Flight means that this transaction is for an airline reservation. Transactions that are airline reservations provide additional information beyond "normal" transactions, and that additional information is the data provided in the next 7 fields.	No
Z	Year		Flight1 Start Date. Empty if "Is Flight?" is 0	No
AA	Month		-->	No
AB	Day		-->	No
AC	Flt1 Src Airport		Flight1 Origin Airport -- often near home. This airport marks where travel begins	No
AD	Flt1 Dest Airport		Flight1 Destination Airport. This airport marks where the first leg of travel ends	No
AE	Flt2 Src Airport		Flight2 Origin Airport -- often same as "Flight1 Dest Airport". This airport marks the start of (typically) a return journey	No
AF	Flt2 Dest Airport		Flight2 Destination Airport -- often same as "Flight1 Src Airport". This airport marks the end of (typically) a return journey	No
AG	Is Fraud?	No	No -> This transaction is legitimate. Yes -> This transaction is fraud	Yes
AH	Fraudster ID	0	Lets models track who is doing fraud -> Better accuracy?	No
AI	IBM Internal	No	This value is reserved for IBM use, and its value should not be relied upon.	No
AJ	Errors?		A number of errors can prevent a transaction from being processed, and this field lists any such errors that occurred for the transaction. Most transactions have no errors, in which case this field is blank. Here are errors that can occur: Bad Card Number Bad Expiration Date Bad CVV Bad PIN Bad Zipcode Insufficient Funds Technical Glitch	Yes

# Core banking

Here are the data schemas for core banking:

- ▶ Banks
- ▶ Liquid accounts people
- ▶ Liquid accounts companies
- ▶ Bank transfers
- ▶ Business-to-business (B2B)

Table 4 is the data schema for banks.

Table 4 Data schema for banks

Column	Field Name	Sample Value	Comment	In Kaggle
A	Bank ID	10A3D1540	Unique ID	No
B	Bank Name	ABC Bank	All major US banks supported including JP Morgan Chase, Bank of America, Wells Fargo, Citi, Capital One Financial, U.S. Bank, PNC Financial Services, Bank of New York Mellon, BB&T, SunTrust Bank, Key Bank, Regions Financial, M&T Bank, Citizens Bank, State Street Bank, Ally Financial, Fifth Third	No
C	Num Transactions	29010	Number of bank transfers involving bank (all locations)	No
D	Num Total Locations	2000		No
E	Num non Focus-Country Locations	100	A dataset has one or more focus countries (e.g. the United States) with more detailed simulation than elsewhere. This value lists the number of bank locations not in a focus country.	No
F	Sample City	Moundridge - Kansas - United States	Note: Banks and large banks in particular have many branches in many cities. The "Sample City" is just an example of one of those places, not an exhaustive list. However, the generated data can have any city with a branch	No

Table 5 is the data schema for liquid accounts people.

Table 5 Data schema for liquid accounts people

Column	Field Name	Sample Value	Comment	In Kaggle
A	Financial Institution Name	PQR Bank	All major US banks supported including JP Morgan Chase, Bank of America, Wells Fargo, Citi, Capital One Financial, U.S. Bank, PNC Financial Services, Bank of New York Mellon, BB&T, SunTrust Bank, Key Bank, Regions Financial, M&T Bank, Citizens Bank, State Street Bank, Ally Financial, Fifth Third	No
B	Financial Institution ID	10A3D12D0	Same ID in "bank_xfers.csv"	No
C	Branch	10C061C30		No
D	Account Country	United States	"Account Currency" typically matches this value - but not always	No
E	Account Currency	USD	USD = US Dollars. Value typically matches "Account Country"	No
F	Entity Type	Person	Owner is a person, not a company. For a personal account "Person" is the only option for this "Entity Type" field	No
G	Entity ID	179	Same index as in "users.csv"	No
H	Entity Name	Cayson Hayes	Same name as in "users.csv"	No
I	Account Has Debit Card?	1	1 --> There is a debit card. 0 --> No debit card.	No
J	Index of Debit Card among Cards of Entity	2	The credit and debit cards held by each entity are numbered 0, 1, 2, 3, ... to distinguish among all cards held by the entity. This index indicates the index among all cards of entity where this debit card sits.	No
K	Controlled by Criminal?	0	0 -> Not controlled by Criminal, 1 -> Controlled by Criminal	No
L	MCC	MCC = Merchant Category Code	Field applies only to companies, not people	No
M	Account Type	Checking	Account types include: Cash Checking Account Savings Account Money Market Account Crypto Currency CD: Certificate of Deposit Bond Mutual Fund Stock Stock Options 401k IRA Life Insurance	No
N	Account ID	3020009543		No
O	Max Overdraft	50	In "Account Currency"	No
NOTE: Format is same as in "liquid accts companies.csv"				



Table 6 is the data schema for liquid accounts companies.

Table 6 Data schema for liquid accounts companies

Column	Field Name	Sample Value	Comment	In Kaggle
A	Financial Institution Name	ABC Bank	All major US banks supported including JP Morgan Chase, Bank of America, Wells Fargo, Citi, Capital One Financial, U.S. Bank, PNC Financial Services, Bank of New York Mellon, BB&T, SunTrust Bank, Key Bank, Regions Financial, M&T Bank, Citizens Bank, State Street Bank, Ally Financial, Fifth Third	No
B	Financial Institution ID	10A3D1540	Same ID in "banks.csv"	No
C	Branch	10C13A360		No
D	Account Country	United States	"Account Currency" typically matches this value - but not always	No
E	Account Currency	USD	USD = US Dollars. Value typically matches "Account Country"	No
F	Entity Type	Large_Corporation	Owner of this account is a large corporation. Possible values include: Sole Proprietorship Partnership Small Corporation Large Corporation Local Government State Government National Government"	No
G	Entity ID	10A3DEA60		No
H	Entity Name	Mega Company Ltd		No
I	Account Has Debit Card?	1	1 --> There is a debit card. 0 --> No debit card.	
J	Index of Debit Card among Cards of Entity	2	The credit and debit cards held by each entity are numbered 0, 1, 2, 3, ... to distinguish among all cards held by the entity. This index indicates the index among all cards of entity where this debit card sits.	
K	Controlled by Criminal?	0	0 -> Not controlled by Criminal, 1 -> Controlled by Criminal	No
L	MCC	1520	Merchant Category Code MCC (or closest fit) for company owning account	No
M	Account Type	Checking	Account types include: Cash Checking Account Savings Account Money Market Account Crypto Currency CD: Certificate of Deposit Bond Mutual Fund Stock Stock Options 401k IRA Life Insurance	No
N	Account ID	1342177281		No
O	Max Overdraft	1000	In "Account Currency"	No
NOTE:	Format is same as in "liquid_accts_people.csv"			

Table 7 is the data schema for bank transfers.

Table 7 Data schema for bank transfers

Column	Field Name	Sample Value	Comment	In Kaggle
A	Timestamp	1/1/22 15:29:31	Kaggle to 1 minute. Enterprise to 0.1 second	Yes
B	From Bank	XYZ Bank	Kaggle value is numeric. Enterprise value is text	Yes
C	From Account	3020009543	From Account Number	Yes
D	To Bank	ABC Bank		Yes
E	To Account	1409286471	To Account Number	Yes
F	Amount Paid	1.20	1.2	Yes
G	Payment Currency	USD	Standard 3-letter currency abbreviations. All 143 country currencies are supported, e.g. EUR and CNY. 13 of the leading crypto currencies are also supported, e.g. Bitcoin (BTC)	Yes
H	Amount Received	1.20	1.2	Yes
I	Receiving Currency	USD	As with "Payment Currency" there are 143 country currencies and 13 crypto currencies. Kaggle NOT standard abbreviations	Yes
J	Payment Format	Debit Non-Prepaid	Supported formats: Cash, Credit Card, Debit Card (Non-Prepaid), Debit (Prepaid), Cheque, ACH, Wire, Bitcoin + Interest, Dividend, Reinvest, Bank Fee	Yes
K	From - Initial Balance	1798.00	1798	No
L	From - End Balance	1796.80		No
M	To - Initial Balance	75835545.32	75835545.32	No
N	To - End Balance	75835546.52		No
O	Transaction Type	Personal Expense	Supported types: Higher Interest, Forced Savings, To IRA Account, From IRA Account, From 401k Account, Buy Securities, Sell Securities, Personal Expense, Credit Card Payment, Prepaid Card Refill, Salary Payment, Accounts Receivable, Accounts Payable, Bank Fee, Bank Interest, Dividend Payment, Insurance Payout, Refund, Hold Release	No
P	Laundering Type	None	Supported types: Fan-Out, Fan-In, Cycle, Bipartite, Stack, Random, Scatter-Gather, Gather-Scatter, Other	No
Q	Is Laundering?	0	0 -> Not a laundering transaction, 1 -> Laundering	Yes
R	Is Cheque Fraud?	0	Is transaction a case of cheque fraud?	No
S	Is APP Fraud?	0	Is transaction a case of APP fraud?	No
T	Cheque Fraudster ID	6AA4F900	Unique ID of Fraudster Committing the Cheque Fraud	No
U	APP Fraudster ID	6AA78E20	Unique ID of Fraudster Committing the APP Fraud	No
V	APP Fraud Sequence Number	2	When field is valid, the sequence number is sequential as thefts continue, e.g. 1, 2, 3, .... When field is not valid, i.e. "Is App Fraud?" is 0, then this field is also 0. In APP fraud, scammers and con artists persuade people (1) they have romantic interest; (2) they are relatives in need; (3) they are owed money for a false invoice, e.g. for tech services or for cyber security software; or (4) other tricks. Based on these tricks, the con artist convinces the victim to send them funds, and sometimes to send funds repeatedly, hence the sequence number.	No
W	Sufficient Funds?	1	0 -> From Account has sufficient fund, 1 -> Does not	No
X	Overdraft Okay?	1	0 -> From Account allows overdrafts, 1 -> Does not	No
Y	Is All Cash?	0	1 -> Both accounts are cash, 1 -> Not both cash	No
Z	Is Hold?	0	0 -> Not a hold transaction, 1 -> Hold	No

Table 8 is the data schema for business-to-business (B2B).

Table 8 Data schema for B2B

Column	Field Name	Sample Value	Comment	In Kaggle
A	B2B Company Index	481	Index of company number. Note that B2B = Business-to-Business Relationships. Values go 1, 2, 3, ...	No
B	Company ID	5E2440B0	Unique ID of Company	No
C	Company Name	Stephens Fabricators	Name of Company	No
D	Company MCC (Merchant Category Code)	1520	MCC is an encoding from the credit card industry to label merchant industries. IBM Synthetic Data Sets augments the standard set of MCC values to be able to represent more industries	No
E	Value of B2B Payments Received (USD)	34677	Expected annual US dollar value of B2B payments received from business customers	No
F	Value of B2B Payments Made (USD)	10053	Expected annual US dollar value of B2B payments made to business suppliers	No
G	Number of B2B Payments Received	102	Number of B2B payments received	No
H	Number of B2B Payments Made	30	Number of B2B payments made	No
I	Number of B2B Payments Missed	0	Number of B2B payments missed	No
J	Number of Main Suppliers	4	Number of business suppliers to Company. Each supplier is detailed in L-R	No
K	Number of Main B2B Customers	7	Number of business customers of Company. Each business customer is detailed in S-Y	No
L	Index Number of Supplier	3	Number from 1, 2, 3, 4, ... of supplier among all major suppliers to company	No
M	Supplier Company ID	3C19FA30	Unique ID of Supplier Company	No
N	Supplier Company Name	Athena's Santa Fe Wholesale	Name of Supplier Company	No
O	Supplier Company MCC (Merchant Category Code)	5300	Industry indicator of Supplier Company. See comments above with D for more details	No
P	Supplier Size Category	Large	Is this supplier large, medium, small, or Unspecified	No
Q	Supplier Expected Sales to Company in USD	6505	What were predicted Sales to Company	No
R	Supplier Actual Sales to Company in USD	7444	What were actual Sales to Company	No
S	Index Number of B2B Customer	3	Number from 1, 2, 3, 4, ... of supplier among all major business customers of company	No
T	B2B Customer Company ID	1A9969D0	Unique ID of Business Customer	No
U	B2B Customer Company Name	Sawyer Roofing and Siding	Name of Business Customer	No
V	B2B Customer Company MCC (Merchant Category Code)	1761	Industry indicator of Business Customer. See comments above with D for more details	No
W	B2B Customer Size Category	Unspecified	Is this business customer large, medium, small, or Unspecified	No
X	B2B Customer Expected Purchases from Company in USD	3073	What were predicted Purchases from Company	No
Y	B2B Customer Actual Purchases from Company in USD	3070	What were actual Purchases from Company	No



# Insurance

Here are the data schemas for insurance:

- ▶ Insurance Application
- ▶ Insurance Policy
- ▶ Insurance Claims
- ▶ Insurance Freetext
- ▶ Storms
- ▶ Quakes
- ▶ Volcanoes

Table 9 is the data schema for insurance applications.

Table 9 Data schema for insurance applications

Column	Field Name	Sample Value	Comment	In Kaggle
A	Index to Insurance Policy CSV	235	Claims use this value to refer to the policy / applicant. The value is the same in "policy.csv", that is, rows of "policy.csv" logically continue the corresponding "applic.csv" row.	No
B	Index - Applicant 1	17	Index to applicant in the users.csv file. Information in the two files and others may be cross-linked. If the value is not in the users.csv file, it indicates that the applicant is not a "primary" person in the simulation, but instead, for example, a spouse of a primary person. The primary can be male or female, as can the spouse.	No
C	Name - Applicant 1	Cayson Hayes		No
D	Date of Birth - Applicant 1	04/01/1987	Month / Day / Year format.	No
E	Social Security Number - Applicant 1	786-38-7809		No
F	Drivers License Number - Applicant 1	Z979-3439-5902-75		No
G	Drivers License State - Applicant 1	Wisconsin	At the time of writing, only the 50 US states are supported.	No

Column	Field Name	Sample Value	Comment	In Kaggle
H	Drivers License Country - Applicant 1	United States	At the time of writing, only the United States are supported, but this field facilitates adding other countries in the future.	No
I	Marital Status - Applicant 1	Married	Six values are supported: Married, Separated, Always Single, Divorced, Widowed, and Cohabiting.	No
J	Education Level - Applicant 1	Associates	Nine values are supported: None, High School, Associates, Some College, Bachelors, Masters, PhD, MD, and JD.	No
K	Personal Phone Number - Applicant 1	414-633-6424		No
L	Email Address - Applicant 1	Hayes.7934@google mail.com	Like other aspects of IBM Synthetic Data Sets, the email addresses are fake, but realistic.	No
M	Street Address - To Be Insured	945 Federal Street		No
N	Unit Number - To Be Insured			No
O	City - To Be Insured	Milwaukee		No
P	State - To Be Insured	WI		No
Q	Postal Code - To Be Insured	53214		No
R	Country - To Be Insured	United States		No
S	Months at this Address	61		No

Column	Field Name	Sample Value	Comment	In Kaggle
T	Previous Address - If less than 36 months at Insured Address		When no previous address is needed, for example, there is more than 36 months at the current address, no values are provided for previous address. In the comma-separated value (CSV) file, these fields will be consecutive commas, which indicate empty fields.	No
U	Previous Unit Number - Applicant 1			No
V	Previous City - Applicant 1			No
W	Previous State - Applicant 1			No
X	Previous Postal Code - Applicant 1			No
Y	Previous Country - Applicant 1			No
Z	Current Employer - Applicant 1	Hilton		No
AA	Street Address - Employer of Applicant 1	36532 Eighth Drive		No
AB	Unit Number - Employer of Applicant 1			No
AC	City - Employer of Applicant 1	Milwaukee		No
AD	State - Employer of Applicant 1	WI		No
AE	Postal Code - Employer of Applicant 1	53214		No
AF	Country - Employer of Applicant 1	United States		No
AG	Type of Employer - Applicant 1	Hotels		No
AH	Position at Employer - Applicant 1	Interviewer		No

Column	Field Name	Sample Value	Comment	In Kaggle
AI	Are Self-Employed - Applicant 1?	No		No
AJ	Years on Job - Applicant 1	3		No
AK	Years in this Profession - Applicant 1	18		No
AL	Business Phone - Applicant 1	414-280-7042		No
AM	Index - Applicant 2	17	Index to applicant in the users . csv file. Information in the two files and others may be cross-linked. If the value is not in the users . csv file, it indicates that the applicant is not a "primary" person in the simulation, but instead, for example, a spouse of a primary person. The primary can be male or female, as can the spouse.	No
AN	Name - Applicant 2	Zoey Hayes		No
AO	Date of Birth - Applicant 2	01/07/1976		No
AP	Social Security Number - Applicant 2	392-56-6826		No
AQ	Drivers License Number - Applicant 2	G407-2062-5784-12		No
AR	Drivers License State - Applicant 2	Wisconsin		No
AS	Drivers License Country - Applicant 2	United States		No
AT	Marital Status - Applicant 2	Married		No
AU	Education Level - Applicant 2	Bachelor's		No
AV	Personal Phone Number - Applicant 2	414-312-2984		No
AW	Email Address - Applicant 2	hare3216@icloud.com		No
AX	Current Employer - Applicant 2	Katelyn's Bank		No

Column	Field Name	Sample Value	Comment	In Kaggle
AY	Street Address - Employer of Applicant 2	358 Madison Boulevard		No
AZ	Unit Number - Employer of Applicant 2			No
BA	City - Employer of Applicant 2	Milwaukee		No
BB	State - Employer of Applicant 2	WI		No
BC	Postal Code - Employer of Applicant 2	53215		No
BD	Country - Employer of Applicant 2	United States		No
BE	Type of Employer - Applicant 2	Financial Institution		No
BF	Position at Employer - Applicant 2	Loan Officer		No
BG	Are Self-Employed - Applicant 2?	No		No
BH	Years on Job - Applicant 2	1		No
BI	Years in this Profession - Applicant 2	27		No
BJ	Business Phone - Applicant 2	414-705-2426		No
BK	Any foreclosures; repossessions; or bankruptcies in the last 5 years?	No		No
BL	Any insurance declined; canceled; or non-renewed in the last 3 years?	No		No
BM	Has anyone with a financial interest in the property been convicted of arson; fraud; or other crime related to a loss on a property?	No		No
BN	Residence Type to be Insured	Single Family House		No
BO	Number of Units	1		No

Column	Field Name	Sample Value	Comment	In Kaggle
BP	Year Built	2022	US Dollars	No
BQ	House Value	251573		No
BR	Distance to Fire Hydrant (Feet)	350		No
BS	Distance to Fire Station (Miles)	1		No
BT	Distance to Tidal Water (Miles)	700		No
BU	Angle of Slope with House (Degrees)	0		No
BV	Lot Size (Square Feet)	21903		No
BW	Living Area (Square Feet)	2633		No
BX	Basement Area (Square Feet)	0		No
BY	Garage Area (Square Feet)	420		No
BZ	Garage Capacity (Number of Cars)	2		No
CA	Basement Finished (Percentage)	0		No
CB	Number of Stories	2		No
CC	Construction Style	Wood Frame		No
CD	Number of Bedrooms	4		No
CE	Number of Full Baths	2		No
CF	Number of Half Baths	1		No
CG	Bathroom Quality	High		No
CH	Kitchen Quality	High		No
CI	Fireplace Count	1		No
CJ	Wood Stove Count	0		No
CK	Electrical Service (Amps)	150		No
CL	Roof - Last Update Year	2022		No
CM	Roof - Type of Update (Full / Partial / None)	None		No
CN	Wiring and Electrical - Last Update Year	2022		No

Column	Field Name	Sample Value	Comment	In Kaggle
CO	Wiring and Electrical - Type of Update (Full / Partial / None)	None		No
CP	Heating - Last Update Year	2022		No
CQ	Heating- Type of Update (Full / Partial / None)	None		No
CR	Plumbing - Last Update Year	2022		No
CS	Plumbing - Type of Update (Full / Partial / None)	None	1 = Concrete Slab; 2 = Crawlspace; 3 = Cinderblock Basement; 4 = Poured Concrete Basement; 5 = Stone Basement; 6 = Wood Pilings; 7 = Concrete Pilings	No
CT	Foundation Type (Numeric Code)	1	1 = Brick; 2 = Wood Siding; 3 = Vinyl Siding; 4 = Aluminum Siding; 5 = Stucco; 6 = Concrete Board; 7 = Wood Shingles; 8 = Synthetic Shingles; 9 = Stone; 10 = Poured Concrete; 11 = Logs; 12 = Asbestos Tiles; 13 = EIFSCB: Exterior Insulation Finishing System over Cinder Block; 14 = EIFSS: Exterior Insulation Finishing System over Studs	No
CU	Exterior Wall Type (Numeric Code)	1	1 = A-Frame; 2 = Flat; 3 = Gable with Valler; 4 = Gable with Dormer; 5 = Bonnet; 6 = Butterfly; 7 = Gambrel; 8 = Dome; 9 = Mansard	No
CV	Roof Shape (Numeric Code)	4	1 = Asphalt Shingles; 2 = Shake - Wood; 3 = Shake - Cement; 4 = Aluminum / Metal; 5 = Copper; 6 = Clay Tiles; 7 = Slate Tiles; 8 = Polymer Tiles; 9 = Thatch; 10 = T-Lock; 11 = Asbestos	No

Column	Field Name	Sample Value	Comment	In Kaggle
CW	Roof Material (Numeric Code)	1	1 = Toe Nailing; 2 = Clips; 3 = Single Straps; 4 = Double Straps; 5 = Structural; 6 = Unknown	No
CX	Roof Anchor (Numeric Code)	1	1 = Strong Glass; 2 = Wooden Storm Shutters; 3 = Electric Metal Shutters; 4 = Manual Metal Shutters; 5 = None; 6 = Unknown	No
CY	Wind Protection (Numeric Code)	6	1 = Great to 10 = Horrible	No
CZ	Protection Class (Numeric Code)	1		No
DA	Is Manufactured Home?	No		No
DB	Is Historic?	No		No
DC	Has Historic Tours?	No		No
DD	Is Garage Attached?	Yes		No
DE	Is Garage Heated?	No		No
DF	Has Automated Garage Doors?	Yes		No
DG	Has Carport?	No		No
DH	Has Screen Enclosure?	No		No
DI	Has Walkout Basement?	No		No
DJ	Has Walkup Attic?	Yes		No
DK	Has T-Lock Shingles?	No		No
DL	Has Asbestos Shingles?	No		No
DM	Is Under Construction?	No		No
DN	Is Bolted To Foundation?	No		No
DO	Has Visible Damage?	No		No
DP	Has Deadbolt Locks?	Yes		No
DQ	Has Sprinklers?	No		No



Column	Field Name	Sample Value	Comment	In Kaggle
DR	Has Smoke Detectors?	Yes		No
DS	Has Carbon Monoxide Detectors?	Yes		No
DT	Has Local Theft Alarm?	No		No
DU	Has Central Theft Alarm?	No		No
DV	Has Central Fire Alarm?	No		No
DW	Has Video Surveillance?	No		No
DX	Has Video Monitoring?	No		No
DY	Has Leak Defense System?	Yes		No
DZ	Has Motion Lighting?	Yes		No
EA	Is Teardown?	No		No
EB	Is Gutted and Remodeled?	No		No
EC	Is Visible from Road?	Yes		No
ED	Is Visible to Neighbors?	Yes		No
EE	Occupied Daily?	Yes		No
EF	Has Flood Insurance?	No		No
EG	Has Knob and Tube Wiring?	No		No
EH	Has Fuses?	No		No
EI	Has FPE Electric Panel?	No		No
EJ	Has Lead Pipes?	No		No
EK	Has Iron Pipes?	No		No
EL	Has Polybutylene Pipes?	No		No
EM	Has Lead Paint?	No		No
EN	Has Asbestos?	No		No
EO	Has Fuel Tank Underground?	No		No

Column	Field Name	Sample Value	Comment	In Kaggle
EP	Has Fuel Tank above Ground?	No		No
EQ	Has Fuel Tank in Basement?	No		No
ER	Converted to Private Home from other Use?	No		No

Table 10 is the data schema for insurance policies.

Table 10 Data schema for insurance policies

Column	Field Name	Sample Value	Comment	In Kaggle
A	Index to Insurance Application CSV	235	Claims use this value to refer to the policy / applicant. The value is the same in "policy.csv", that is, rows of "policy.csv" logically continue the corresponding "applic.csv" row.	No
B	Index to Insurance Agency CSV	1	Insurance agency information was not provided in the initial datasets.	No
C	ID for Insurance Company CSV	10A3CE5D0	Insurance agency information was not provided in the initial datasets.	No
D	Coverage Class	HO-4	A Standard Insurance Coverage	No
E	Premium Amount	439.00	In "Monetary Currency"	No
F	Monetary Currency	USD	USD = US Dollars. The value typically matches the country where the home is.	No
G	Months Covered by Premium	3	Quarterly payments	No
H	Start Date	01/15/2024	Month / Day / Year format	No
I	End Date	04/15/2024	Month / Day / Year format	No
J	Theft - Physical Goods: Coverage Limit	40000	In "Monetary Currency" -- as are all monetary values below	No

Column	Field Name	Sample Value	Comment	In Kaggle
K	Theft - Physical Goods: Deductible	200		No
L	Vandalism: Coverage Limit	61000		No
M	Vandalism: Deductible	200		No
N	Riots: Coverage Limit	51000		No
O	Riots: Deductible	200		No
P	Explosion: Coverage Limit	34000		No
Q	Explosion: Deductible	200		No
R	Fire Damage: Coverage Limit	28000		No
S	Fire Damage: Deductible	200		No
T	Hail Damage: Coverage Limit	49000		No
U	Hail Damage: Deductible	200		No
V	Wind Damage: Coverage Limit	45000		No
W	Wind Damage: Deductible	200		No
X	Flood: Coverage Limit	0		No
Y	Flood: Deductible	0		No
Z	Water Damage - Weather: Coverage Limit	61000		No
AA	Water Damage - Weather: Deductible	200		No
AB	Water Damage - Plumbing: Coverage Limit	29000		No
AC	Water Damage - Plumbing: Deductible	200		No
AD	Water Damage - Heating Overflow: Coverage Limit	60000		No
AE	Water Damage - Heating Overflow: Deductible	200		No

Column	Field Name	Sample Value	Comment	In Kaggle
AF	Water Damage - AC Overflow: Coverage Limit	43000		No
AG	Water Damage - AC Overflow: Deductible	200		No
AH	Appliance Flood: Coverage Limit	74000		No
AI	Appliance Flood: Deductible	200		No
AJ	Water Heater: Coverage Limit	59000		No
AK	Water Heater: Deductible	200		No
AL	Frozen Pipes: Coverage Limit	26000		No
AM	Frozen Pipes: Deductible	200		No
AN	Snow / Ice Buildup: Coverage Limit	38000		No
AO	Snow / Ice Buildup: Deductible	200		No
AP	Lightning: Coverage Limit	57000		No
AQ	Lightning: Deductible	200		No
AR	Electrical Current: Coverage Limit	68000		No
AS	Electrical Current: Deductible	200		No
AT	Tree: Coverage Limit	68000		No
AU	Tree: Deductible	200		No
AV	Falling Object: Coverage Limit	39000		No
AW	Falling Object: Deductible	200		No
AX	Aircraft Damage: Coverage Limit	52000		No
AY	Aircraft Damage: Deductible	200		No
AZ	Vehicle caused Damage: Coverage Limit	51000		No

Column	Field Name	Sample Value	Comment	In Kaggle
BA	Vehicle caused Damage: Deductible	200		No
BB	Sinkhole: Coverage Limit	0		No
BC	Sinkhole: Deductible	0		No
BD	Earthquake: Coverage Limit	0		No
BE	Earthquake: Deductible	0		No
BF	Volcano: Coverage Limit	43000		No
BG	Volcano: Deductible	200		No
BH	Mandatory Evacuation: Coverage Limit	0		No
BI	Mandatory Evacuation: Deductible	0		No
BJ	Ordinance Change: Coverage Limit	0		No
BK	Ordinance Change: Deductible	0		No
BL	Building Codes: Coverage Limit	0		No
BM	Building Codes: Deductible	0		No
BN	Eco Upgrade: Coverage Limit	0		No
BO	Eco Upgrade: Deductible	0		No
BP	Identity Theft: Coverage Limit	8000		No
BQ	Identity Theft: Deductible	250		No
BR	Mold: Coverage Limit	0		No
BS	Mold: Deductible	0		No
BT	Termites: Coverage Limit	0		No
BU	Termites: Deductible	0		No
BV	Decayed Foundation: Coverage Limit	68000		No

Column	Field Name	Sample Value	Comment	In Kaggle
BW	Decayed Foundation: Deductible	200		No
BX	Failure to keep safe env: Coverage Limit	25000		No
BY	Failure to keep safe env: Deductible	200		No
BZ	Dwelling: Replacement Cost?	No		No
CA	Dwelling: Coverage Limit	215000		No
CB	Dwelling: Deductible	250		No
CC	Extended Premises: Replacement Cost?	No		No
CD	Extended Premises: Coverage Limit	0		No
CE	Extended Premises: Deductible	0		No
CF	Other Structures: Replacement Cost?	No		No
CG	Other Structures: Coverage Limit	60000		No
CH	Other Structures: Deductible	200		No
CI	Roof Surfaces: Replacement Cost?	No		No
CJ	Roof Surfaces: Coverage Limit	0		No
CK	Roof Surfaces: Deductible	0		No
CL	Yard and Garden: Replacement Cost?	No		No
CM	Yard and Garden: Coverage Limit	0		No
CN	Yard and Garden: Deductible	0		No
CO	Data Recovery: Replacement Cost?	No		No
CP	Data Recovery: Coverage Limit	0		No
CQ	Data Recovery: Deductible	0		No

Column	Field Name	Sample Value	Comment	In Kaggle
CR	Credit Cards: Replacement Cost?	No		No
CS	Credit Cards: Coverage Limit	0		No
CT	Credit Cards: Deductible	0		No
CU	Financial Assets: Replacement Cost?	No		No
CV	Financial Assets: Coverage Limit	0		No
CW	Financial Assets: Deductible	0		No
CX	Rental Income Loss: Replacement Cost?	No		No
CY	Rental Income Loss: Coverage Limit	0		No
CZ	Rental Income Loss: Deductible	0		No
DA	Business Property: Replacement Cost?	No		No
DB	Business Property: Coverage Limit	0		No
DC	Business Property: Deductible	0		No
DD	Home Daycare: Replacement Cost?	No		No
DE	Home Daycare: Coverage Limit	0		No
DF	Home Daycare: Deductible	0		No
DG	Medical Payments: Replacement Cost?	No		No
DH	Medical Payments: Coverage Limit	0		No
DI	Medical Payments: Deductible	0		No
DJ	Liability - Bodily Injury: Replacement Cost?	No		No
DK	Liability - Bodily Injury: Coverage Limit	145000		No

Column	Field Name	Sample Value	Comment	In Kaggle
DL	Liability - Bodily Injury: Deductible	250		No
DM	Liability - Property Damage: Replacement Cost?	No		No
DN	Liability - Property Damage: Coverage Limit	761000		No
DO	Liability - Property Damage: Deductible	250		No
DP	Loss Assessment: Replacement Cost?	No		No
DQ	Loss Assessment: Coverage Limit	0		No
DR	Loss Assessment: Deductible	0		No
DS	Fire Department Charges: Replacement Cost?	No		No
DT	Fire Department Charges: Coverage Limit	0		No
DU	Fire Department Charges: Deductible	0		No
DV	Living Expenses: Replacement Cost?	No		No
DW	Living Expenses: Coverage Limit	51000		No
DX	Living Expenses: Deductible	250		No
DY	Furniture: Replacement Cost?	No		No
DZ	Furniture: Coverage Limit	234000		No
EA	Furniture: Deductible	250		No
EB	Appliances: Replacement Cost?	No		No
EC	Appliances: Coverage Limit	8000		No
ED	Appliances: Deductible	250		No
EE	Electronics: Replacement Cost?	No		No



Column	Field Name	Sample Value	Comment	In Kaggle
EF	Electronics: Coverage Limit	165000		No
EG	Electronics: Deductible	250		No
EH	Beds & Mattresses: Replacement Cost?	No		No
EI	Beds & Mattresses: Coverage Limit	36000		No
EJ	Beds & Mattresses: Deductible	250		No
EK	Apparel: Replacement Cost?	No		No
EL	Apparel: Coverage Limit	50000		No
EM	Apparel: Deductible	250		No
EN	Jewelry: Replacement Cost?	No		No
EO	Jewelry: Coverage Limit	0		No
EP	Jewelry: Deductible	0		No
EQ	Silverware: Replacement Cost?	No		No
ER	Silverware: Coverage Limit	0		No
ES	Silverware: Deductible	0		No
ET	Tools: Replacement Cost?	No		No
EU	Tools: Coverage Limit	28000		No
EV	Tools: Deductible	250		No
EW	Construction Material: Replacement Cost?	No		No
EX	Construction Material: Coverage Limit	0		No
EY	Construction Material: Deductible	0		No
EZ	Books & Magazines: Replacement Cost?	No		No

Column	Field Name	Sample Value	Comment	In Kaggle
FA	Books & Magazines: Coverage Limit	36000		No
FB	Books & Magazines: Deductible	250		No
FC	Sporting Goods: Replacement Cost?	No		No
FD	Sporting Goods: Coverage Limit	170000		No
FE	Sporting Goods: Deductible	250		No
FF	Golf Cart: Replacement Cost?	No		No
FG	Golf Cart: Coverage Limit	0		No
FH	Golf Cart: Deductible	0		No
FI	Cameras: Replacement Cost?	No		No
FJ	Cameras: Coverage Limit	0		No
FK	Cameras: Deductible	0		No
FL	Watches: Replacement Cost?	No		No
FM	Watches: Coverage Limit	0		No
FN	Watches: Deductible	0		No
FO	Furs: Replacement Cost?	No		No
FP	Furs: Coverage Limit	0		No
FQ	Furs: Deductible	0		No
FR	Medical Instruments: Replacement Cost?	No		No
FS	Medical Instruments: Coverage Limit	0		No
FT	Medical Instruments: Deductible	0		No
FU	Musical Instruments: Replacement Cost?	No		No
FV	Musical Instruments: Coverage Limit	0		No
FW	Musical Instruments: Deductible	0		No

Column	Field Name	Sample Value	Comment	In Kaggle
FX	Other Personal Property: Replacement Cost?	No		No
FY	Other Personal Property: Coverage Limit	142000		No
FZ	Other Personal Property: Deductible	250		No
GA	Special Deductibles: Wind - Percentage	0		No
GB	Special Deductibles: Wind - Dollar	0		No
GC	Special Deductibles: Named Storm	0		No
GD	Special Deductibles: Hurricane	0		No
GE	Special Deductibles: Theft	0		No
GF	Special Deductibles: Water	0		No
GG	Special Deductibles: All Other Perils	0		No

Table 11 is the data schema for insurance claims.

*Table 11 Data schema for insurance claims*

Column	Field Name	Sample Value	Comment 1	In Kaggle
A	Index to Insurance Application/Policy CSVs	235	The value in the first column of "applic.csv" or "policy.csv".	No
B	Policy ID	6A9B0C8D0	Alternative value for "index" in previous column.	No
C	Home ID	6A9B05DF0		No
D	Monetary Currency	USD	USD = US Dollars. A typical value matches the country of the owner and home.	
E	Date	07/18/2023	Month / Day / Year Format.	No

Column	Field Name	Sample Value	Comment 1	In Kaggle
F	Cause of Claim	Wind Damage	IBM Synthetic Data Sets supports over 30 causes for claims. Among these causes are Physical Theft, Vandalism, Riots, Explosion, Fire Damage, Hail Damage, Wind Damage, and Flood.	No
G	Assoc w Hurricane	0	Is this claim associated with a hurricane? FALSE -> No	No
H	Assoc w Earthquake	0	Is this claim associated with an earthquake? FALSE -> No	No
I	Assoc w Volcano	0	Is this claim associated with a volcano? FALSE -> No	No
J	Total \$Claimed	12380		No
K	Total \$Paid	1723		No
L	Deductible \$on Claim	5000		No
M	Is Claim Cause Covered	1		No
N	Is Fraud on Claim	0		No
O	Is Detected Fraud on Claim	0		No
P	Item 1 - Dwelling: \$Loss Claimed	6723	US Dollars	No

Column	Field Name	Sample Value	Comment 1	In Kaggle
Q	Item 1: \$Loss Allowed	6723	Detailed breakdowns are provided for 35 types of items: Item 1: House; Item 2: Extended Premises; Item 3: Other Structures; Item 4: Roof Surfaces; Item 5: Yard And Garden; Item 6: Data Recovery; Item 7: Credit Card; Item 8: Financial Assets; Item 9: Rental Income Loss; Item 10: Business Property; Item 11: Home Daycare; Item 12: Medical Payments; Item 13: Liability Bodily Injury; Item 14: Liability Property Damage; Item 15: Loss Assessment; Item 16: Fire Department Charges; Item 17: Living Expenses; Item 18: Furniture; Item 19: Appliances; Item 20: Electronics; Item 21: Beds Mattresses; Item 22: Apparel; Item 23: Jewelry; Item 24: Silverware; Item 25: Tools; Item 26: Construction Material; Item 27: Books Magazines; Item 28: Sporting Goods; Item 29: Golf Cart; Item 30: Cameras; Item 31: Watches; Item 32: Furs; Item 33: Medical Instruments; Item 34: Musical Instruments; Item 35: Other Personal Property	No
R	Item 1 - Fraud: Overstated Value	0		No
S	Item 1 - Fraud: Intentional Damage	0		No

Column	Field Name	Sample Value	Comment 1	In Kaggle
T	Item 1 - Fraud: Fake Theft	0		No
U	Item 1 - Fraud: Fake Repair Bills	0		No
V	Item 1 - Fraud: Inflated Repair Bills	0		No
W	Item 1 - Fraud: Non-Covered Use	0		No
X	Item 1 - Fraud: Non-Covered Damage	0		No
Y	Item 1 - Disallowed: Fraud	0		No
Z	Item 1 - Disallowed: Under Deductible	0		No
AA	Item 1 - Disallowed: Not Covered	0		No
AB	Item 1 - Non-Full: Over Limit	1		No
AC	Item 1 - Non-Full: Depreciation	0		No
AD	Item 1 - Non-Full: Over Market Price	0		No
AE	Item 2 - Extended Premises: \$Loss Claimed	0		No
AF	Item 2: \$Loss Allowed	0		No
AG	Item 2 - Fraud: Overstated Value	0		No
AH	Item 2 - Fraud: Intentional Damage	0		No
AI	Item 2 - Fraud: Fake Theft	0		No
AJ	Item 2 - Fraud: Fake Repair Bills	0		No
AK	Item 2 - Fraud: Inflated Repair Bills	0		No
AL	Item 2 - Fraud: Non-Covered Use	0		No
AM	Item 2 - Fraud: Non-Covered Damage	0		No

Column	Field Name	Sample Value	Comment 1	In Kaggle
AN	Item 2 - Disallowed: Fraud	0		No
AO	Item 2 - Disallowed: Under Deductible	0		No
AP	Item 2 - Disallowed: Not Covered	0		No
AQ	Item 2 - Non-Full: Over Limit	0		No
AR	Item 2 - Non-Full: Depreciation	0		No
AS	Item 2 - Non-Full: Over Market Price	0		No
AT	Item 3 - Other Structures: \$Loss Claimed	0		No
AU	Item 3: \$Loss Allowed	0		No
AV	Item 3 - Fraud: Overstated Value	0		No
AW	Item 3 - Fraud: Intentional Damage	0		No
AX	Item 3 - Fraud: Fake Theft	0		No
AY	Item 3 - Fraud: Fake Repair Bills	0		No
AZ	Item 3 - Fraud: Inflated Repair Bills	0		No
BA	Item 3 - Fraud: Non-Covered Use	0		No
BB	Item 3 - Fraud: Non-Covered Damage	0		No
BC	Item 3 - Disallowed: Fraud	0		No
BD	Item 3 - Disallowed: Under Deductible	0		No
BE	Item 3 - Disallowed: Not Covered	0		No
BF	Item 3 - Non-Full: Over Limit	0		No
BG	Item 3 - Non-Full: Depreciation	0		No

Column	Field Name	Sample Value	Comment 1	In Kaggle
BH	Item 3 - Non-Full: Over Market Price	0		No
BI	Item 4 - Roof Surfaces: \$Loss Claimed	5656		No
BJ	Item 4: \$Loss Allowed	0		No
BK	Item 4 - Fraud: Overstated Value	0		No
BL	Item 4 - Fraud: Intentional Damage	0		No
BM	Item 4 - Fraud: Fake Theft	0		No
BN	Item 4 - Fraud: Fake Repair Bills	0		No
BO	Item 4 - Fraud: Inflated Repair Bills	0		No
BP	Item 4 - Fraud: Non-Covered Use	0		No
BQ	Item 4 - Fraud: Non-Covered Damage	0		No
BR	Item 4 - Disallowed: Fraud	0		No
BS	Item 4 - Disallowed: Under Deductible	0		No
BT	Item 4 - Disallowed: Not Covered	1		No
BU	Item 4 - Non-Full: Over Limit	0		No
BV	Item 4 - Non-Full: Depreciation	0		No
BW	Item 4 - Non-Full: Over Market Price	0		No
BX	Item 5 - Yard and Garden: \$Loss Claimed	0		No
BY	Item 5: \$Loss Allowed	0		No
BZ	Item 5 - Fraud: Overstated Value	0		No
CA	Item 5 - Fraud: Intentional Damage	0		No



Column	Field Name	Sample Value	Comment 1	In Kaggle
CB	Item 5 - Fraud: Fake Theft	0		No
CC	Item 5 - Fraud: Fake Repair Bills	0		No
CD	Item 5 - Fraud: Inflated Repair Bills	0		No
CE	Item 5 - Fraud: Non-Covered Use	0		No
CF	Item 5 - Fraud: Non-Covered Damage	0		No
CG	Item 5 - Disallowed: Fraud	0		No
CH	Item 5 - Disallowed: Under Deductible	0		No
CI	Item 5 - Disallowed: Not Covered	0		No
CJ	Item 5 - Non-Full: Over Limit	0		No
CK	Item 5 - Non-Full: Depreciation	0		No
CL	Item 5 - Non-Full: Over Market Price	0		No
CM	Item 6 - Data Recovery: \$Loss Claimed	0		No
CN	Item 6: \$Loss Allowed	0		No
CO	Item 6 - Fraud: Overstated Value	0		No
CP	Item 6 - Fraud: Intentional Damage	0		No
CQ	Item 6 - Fraud: Fake Theft	0		No
CR	Item 6 - Fraud: Fake Repair Bills	0		No
CS	Item 6 - Fraud: Inflated Repair Bills	0		No
CT	Item 6 - Fraud: Non-Covered Use	0		No
CU	Item 6 - Fraud: Non-Covered Damage	0		No

Column	Field Name	Sample Value	Comment 1	In Kaggle
CV	Item 6 - Disallowed: Fraud	0		No
CW	Item 6 - Disallowed: Under Deductible	0		No
CX	Item 6 - Disallowed: Not Covered	0		No
CY	Item 6 - Non-Full: Over Limit	0		No
CZ	Item 6 - Non-Full: Depreciation	0		No
DA	Item 6 - Non-Full: Over Market Price	0		No
DB	Item 7 - Credit Cards: \$Loss Claimed	0		No
DC	Item 7: \$Loss Allowed	0		No
DD	Item 7 - Fraud: Overstated Value	0		No
DE	Item 7 - Fraud: Intentional Damage	0		No
DF	Item 7 - Fraud: Fake Theft	0		No
DG	Item 7 - Fraud: Fake Repair Bills	0		No
DH	Item 7 - Fraud: Inflated Repair Bills	0		No
DI	Item 7 - Fraud: Non-Covered Use	0		No
DJ	Item 7 - Fraud: Non-Covered Damage	0		No
DK	Item 7 - Disallowed: Fraud	0		No
DL	Item 7 - Disallowed: Under Deductible	0		No
DM	Item 7 - Disallowed: Not Covered	0		No
DN	Item 7 - Non-Full: Over Limit	0		No
DO	Item 7 - Non-Full: Depreciation	0		No
DP	Item 7 - Non-Full: Over Market Price	0		No

Column	Field Name	Sample Value	Comment 1	In Kaggle
DQ	Item 8 - Financial Assets: \$Loss Claimed	0		No
DR	Item 8: \$Loss Allowed	0		No
DS	Item 8 - Fraud: Overstated Value	0		No
DT	Item 8 - Fraud: Intentional Damage	0		No
DU	Item 8 - Fraud: Fake Theft	0		No
DV	Item 8 - Fraud: Fake Repair Bills	0		No
DW	Item 8 - Fraud: Inflated Repair Bills	0		No
DX	Item 8 - Fraud: Non-Covered Use	0		No
DY	Item 8 - Fraud: Non-Covered Damage	0		No
DZ	Item 8 - Disallowed: Fraud	0		No
EA	Item 8 - Disallowed: Under Deductible	0		No
EB	Item 8 - Disallowed: Not Covered	0		No
EC	Item 8 - Non-Full: Over Limit	0		No
ED	Item 8 - Non-Full: Depreciation	0		No
EE	Item 8 - Non-Full: Over Market Price	0		No
EF	Item 9 - Rental Income Loss: \$Loss Claimed	0		No
EG	Item 9: \$Loss Allowed	0		No
EH	Item 9 - Fraud: Overstated Value	0		No
EI	Item 9 - Fraud: Intentional Damage	0		No
EJ	Item 9 - Fraud: Fake Theft	0		No

Column	Field Name	Sample Value	Comment 1	In Kaggle
EK	Item 9 - Fraud: Fake Repair Bills	0		No
EL	Item 9 - Fraud: Inflated Repair Bills	0		No
EM	Item 9 - Fraud: Non-Covered Use	0		No
EN	Item 9 - Fraud: Non-Covered Damage	0		No
EO	Item 9 - Disallowed: Fraud	0		No
EP	Item 9 - Disallowed: Under Deductible	0		No
EQ	Item 9 - Disallowed: Not Covered	0		No
ER	Item 9 - Non-Full: Over Limit	0		No
ES	Item 9 - Non-Full: Depreciation	0		No
ET	Item 9 - Non-Full: Over Market Price	0		No
EU	Item 10 - Business Property: \$Loss Claimed	0		No
EV	Item 10: \$Loss Allowed	0		No
EW	Item 10 - Fraud: Overstated Value	0		No
EX	Item 10 - Fraud: Intentional Damage	0		No
EY	Item 10 - Fraud: Fake Theft	0		No
EZ	Item 10 - Fraud: Fake Repair Bills	0		No
FA	Item 10 - Fraud: Inflated Repair Bills	0		No
FB	Item 10 - Fraud: Non-Covered Use	0		No
FC	Item 10 - Fraud: Non-Covered Damage	0		No
FD	Item 10 - Disallowed: Fraud	0		No

Column	Field Name	Sample Value	Comment 1	In Kaggle
FE	Item 10 - Disallowed: Under Deductible	0		No
FF	Item 10 - Disallowed: Not Covered	0		No
FG	Item 10 - Non-Full: Over Limit	0		No
FH	Item 10 - Non-Full: Depreciation	0		No
FI	Item 10 - Non-Full: Over Market Price	0		No
FJ	Item 11 - Home Daycare: \$Loss Claimed	0		No
FK	Item 11: \$Loss Allowed	0		No
FL	Item 11 - Fraud: Overstated Value	0		No
FM	Item 11 - Fraud: Intentional Damage	0		No
FN	Item 11 - Fraud: Fake Theft	0		No
FO	Item 11 - Fraud: Fake Repair Bills	0		No
FP	Item 11 - Fraud: Inflated Repair Bills	0		No
FQ	Item 11 - Fraud: Non-Covered Use	0		No
FR	Item 11 - Fraud: Non-Covered Damage	0		No
FS	Item 11 - Disallowed: Fraud	0		No
FT	Item 11 - Disallowed: Under Deductible	0		No
FU	Item 11 - Disallowed: Not Covered	0		No
FV	Item 11 - Non-Full: Over Limit	0		No
FW	Item 11 - Non-Full: Depreciation	0		No
FX	Item 11 - Non-Full: Over Market Price	0		No

Column	Field Name	Sample Value	Comment 1	In Kaggle
FY	Item 12 - Medical Payments: \$Loss Claimed	0		No
FZ	Item 12: \$Loss Allowed	0		No
GA	Item 12 - Fraud: Overstated Value	0		No
GB	Item 12 - Fraud: Intentional Damage	0		No
GC	Item 12 - Fraud: Fake Theft	0		No
GD	Item 12 - Fraud: Fake Repair Bills	0		No
GE	Item 12 - Fraud: Inflated Repair Bills	0		No
GF	Item 12 - Fraud: Non-Covered Use	0		No
GG	Item 12 - Fraud: Non-Covered Damage	0		No
GH	Item 12 - Disallowed: Fraud	0		No
GI	Item 12 - Disallowed: Under Deductible	0		No
GJ	Item 12 - Disallowed: Not Covered	0		No
GK	Item 12 - Non-Full: Over Limit	0		No
GL	Item 12 - Non-Full: Depreciation	0		No
GM	Item 12 - Non-Full: Over Market Price	0		No
GN	Item 13 - Liability - Bodily Injury: \$Loss Claimed	0		No
GO	Item 13: \$Loss Allowed	0		No
GP	Item 13 - Fraud: Overstated Value	0		No
GQ	Item 13 - Fraud: Intentional Damage	0		No
GR	Item 13 - Fraud: Fake Theft	0		No

Column	Field Name	Sample Value	Comment 1	In Kaggle
GS	Item 13 - Fraud: Fake Repair Bills	0		No
GT	Item 13 - Fraud: Inflated Repair Bills	0		No
GU	Item 13 - Fraud: Non-Covered Use	0		No
GV	Item 13 - Fraud: Non-Covered Damage	0		No
GW	Item 13 - Disallowed: Fraud	0		No
GX	Item 13 - Disallowed: Under Deductible	0		No
GY	Item 13 - Disallowed: Not Covered	0		No
GZ	Item 13 - Non-Full: Over Limit	0		No
HA	Item 13 - Non-Full: Depreciation	0		No
HB	Item 13 - Non-Full: Over Market Price	0		No
HC	Item 14 - Liability - Property Damage: \$Loss Claimed	0		No
HD	Item 14: \$Loss Allowed	0		No
HE	Item 14 - Fraud: Overstated Value	0		No
HF	Item 14 - Fraud: Intentional Damage	0		No
HG	Item 14 - Fraud: Fake Theft	0		No
HH	Item 14 - Fraud: Fake Repair Bills	0		No
HI	Item 14 - Fraud: Inflated Repair Bills	0		No
HJ	Item 14 - Fraud: Non-Covered Use	0		No
HK	Item 14 - Fraud: Non-Covered Damage	0		No
HL	Item 14 - Disallowed: Fraud	0		No

Column	Field Name	Sample Value	Comment 1	In Kaggle
HM	Item 14 - Disallowed: Under Deductible	0		No
HN	Item 14 - Disallowed: Not Covered	0		No
HO	Item 14 - Non-Full: Over Limit	0		No
HP	Item 14 - Non-Full: Depreciation	0		No
HQ	Item 14 - Non-Full: Over Market Price	0		No
HR	Item 15 - Loss Assessment: \$Loss Claimed	0		No
HS	Item 15: \$Loss Allowed	0		No
HT	Item 15 - Fraud: Overstated Value	0		No
HU	Item 15 - Fraud: Intentional Damage	0		No
HV	Item 15 - Fraud: Fake Theft	0		No
HW	Item 15 - Fraud: Fake Repair Bills	0		No
HX	Item 15 - Fraud: Inflated Repair Bills	0		No
HY	Item 15 - Fraud: Non-Covered Use	0		No
HZ	Item 15 - Fraud: Non-Covered Damage	0		No
IA	Item 15 - Disallowed: Fraud	0		No
IB	Item 15 - Disallowed: Under Deductible	0		No
IC	Item 15 - Disallowed: Not Covered	0		No
ID	Item 15 - Non-Full: Over Limit	0		No
IE	Item 15 - Non-Full: Depreciation	0		No
IF	Item 15 - Non-Full: Over Market Price	0		No



Column	Field Name	Sample Value	Comment 1	In Kaggle
IG	Item 16 - Fire Department Charges: \$Loss Claimed	0		No
IH	Item 16: \$Loss Allowed	0		No
II	Item 16 - Fraud: Overstated Value	0		No
IJ	Item 16 - Fraud: Intentional Damage	0		No
IK	Item 16 - Fraud: Fake Theft	0		No
IL	Item 16 - Fraud: Fake Repair Bills	0		No
IM	Item 16 - Fraud: Inflated Repair Bills	0		No
IN	Item 16 - Fraud: Non-Covered Use	0		No
IO	Item 16 - Fraud: Non-Covered Damage	0		No
IP	Item 16 - Disallowed: Fraud	0		No
IQ	Item 16 - Disallowed: Under Deductible	0		No
IR	Item 16 - Disallowed: Not Covered	0		No
IS	Item 16 - Non-Full: Over Limit	0		No
IT	Item 16 - Non-Full: Depreciation	0		No
IU	Item 16 - Non-Full: Over Market Price	0		No
IV	Item 17 - Living Expenses: \$Loss Claimed	0		No
IW	Item 17: \$Loss Allowed	0		No
IX	Item 17 - Fraud: Overstated Value	0		No
IY	Item 17 - Fraud: Intentional Damage	0		No
IZ	Item 17 - Fraud: Fake Theft	0		No

Column	Field Name	Sample Value	Comment 1	In Kaggle
JA	Item 17 - Fraud: Fake Repair Bills	0		No
JB	Item 17 - Fraud: Inflated Repair Bills	0		No
JC	Item 17 - Fraud: Non-Covered Use	0		No
JD	Item 17 - Fraud: Non-Covered Damage	0		No
JE	Item 17 - Disallowed: Fraud	0		No
JF	Item 17 - Disallowed: Under Deductible	0		No
JG	Item 17 - Disallowed: Not Covered	0		No
JH	Item 17 - Non-Full: Over Limit	0		No
JI	Item 17 - Non-Full: Depreciation	0		No
JJ	Item 17 - Non-Full: Over Market Price	0		No
JK	Item 18 - Furniture: \$Loss Claimed	0		No
JL	Item 18: \$Loss Allowed	0		No
JM	Item 18 - Fraud: Overstated Value	0		No
JN	Item 18 - Fraud: Intentional Damage	0		No
JO	Item 18 - Fraud: Fake Theft	0		No
JP	Item 18 - Fraud: Fake Repair Bills	0		No
JQ	Item 18 - Fraud: Inflated Repair Bills	0		No
JR	Item 18 - Fraud: Non-Covered Use	0		No
JS	Item 18 - Fraud: Non-Covered Damage	0		No
JT	Item 18 - Disallowed: Fraud	0		No

Column	Field Name	Sample Value	Comment 1	In Kaggle
JU	Item 18 - Disallowed: Under Deductible	0		No
JV	Item 18 - Disallowed: Not Covered	0		No
JW	Item 18 - Non-Full: Over Limit	0		No
JX	Item 18 - Non-Full: Depreciation	0		No
JY	Item 18 - Non-Full: Over Market Price	0		No
JZ	Item 19 - Appliances: \$Loss Claimed	0		No
KA	Item 19: \$Loss Allowed	0		No
KB	Item 19 - Fraud: Overstated Value	0		No
KC	Item 19 - Fraud: Intentional Damage	0		No
KD	Item 19 - Fraud: Fake Theft	0		No
KE	Item 19 - Fraud: Fake Repair Bills	0		No
KF	Item 19 - Fraud: Inflated Repair Bills	0		No
KG	Item 19 - Fraud: Non-Covered Use	0		No
KH	Item 19 - Fraud: Non-Covered Damage	0		No
KI	Item 19 - Disallowed: Fraud	0		No
KJ	Item 19 - Disallowed: Under Deductible	0		No
KK	Item 19 - Disallowed: Not Covered	0		No
KL	Item 19 - Non-Full: Over Limit	0		No
KM	Item 19 - Non-Full: Depreciation	0		No
KN	Item 19 - Non-Full: Over Market Price	0		No
KO	Item 20 - Electronics: \$Loss Claimed	0		No

Column	Field Name	Sample Value	Comment 1	In Kaggle
KP	Item 20: \$Loss Allowed	0		No
KQ	Item 20 - Fraud: Overstated Value	0		No
KR	Item 20 - Fraud: Intentional Damage	0		No
KS	Item 20 - Fraud: Fake Theft	0		No
KT	Item 20 - Fraud: Fake Repair Bills	0		No
KU	Item 20 - Fraud: Inflated Repair Bills	0		No
KV	Item 20 - Fraud: Non-Covered Use	0		No
KW	Item 20 - Fraud: Non-Covered Damage	0		No
KX	Item 20 - Disallowed: Fraud	0		No
KY	Item 20 - Disallowed: Under Deductible	0		No
KZ	Item 20 - Disallowed: Not Covered	0		No
LA	Item 20 - Non-Full: Over Limit	0		No
LB	Item 20 - Non-Full: Depreciation	0		No
LC	Item 20 - Non-Full: Over Market Price	0		No
LD	Item 21 - Beds & Mattresses: \$Loss Claimed	0		No
LE	Item 21: \$Loss Allowed	0		No
LF	Item 21 - Fraud: Overstated Value	0		No
LG	Item 21 - Fraud: Intentional Damage	0		No
LH	Item 21 - Fraud: Fake Theft	0		No
LI	Item 21 - Fraud: Fake Repair Bills	0		No

Column	Field Name	Sample Value	Comment 1	In Kaggle
LJ	Item 21 - Fraud: Inflated Repair Bills	0		No
LK	Item 21 - Fraud: Non-Covered Use	0		No
LL	Item 21 - Fraud: Non-Covered Damage	0		No
LM	Item 21 - Disallowed: Fraud	0		No
LN	Item 21 - Disallowed: Under Deductible	0		No
LO	Item 21 - Disallowed: Not Covered	0		No
LP	Item 21 - Non-Full: Over Limit	0		No
LQ	Item 21 - Non-Full: Depreciation	0		No
LR	Item 21 - Non-Full: Over Market Price	0		No
LS	Item 22 - Apparel: \$Loss Claimed	0		No
LT	Item 22: \$Loss Allowed	0		No
LU	Item 22 - Fraud: Overstated Value	0		No
LV	Item 22 - Fraud: Intentional Damage	0		No
LW	Item 22 - Fraud: Fake Theft	0		No
LX	Item 22 - Fraud: Fake Repair Bills	0		No
LY	Item 22 - Fraud: Inflated Repair Bills	0		No
LZ	Item 22 - Fraud: Non-Covered Use	0		No
MA	Item 22 - Fraud: Non-Covered Damage	0		No
MB	Item 22 - Disallowed: Fraud	0		No
MC	Item 22 - Disallowed: Under Deductible	0		No

Column	Field Name	Sample Value	Comment 1	In Kaggle
MD	Item 22 - Disallowed: Not Covered	0		No
ME	Item 22 - Non-Full: Over Limit	0		No
MF	Item 22 - Non-Full: Depreciation	0		No
MG	Item 22 - Non-Full: Over Market Price	0		No
MH	Item 23 - Jewelry: \$Loss Claimed	0		No
MI	Item 23: \$Loss Allowed	0		No
MJ	Item 23 - Fraud: Overstated Value	0		No
MK	Item 23 - Fraud: Intentional Damage	0		No
ML	Item 23 - Fraud: Fake Theft	0		No
MM	Item 23 - Fraud: Fake Repair Bills	0		No
MN	Item 23 - Fraud: Inflated Repair Bills	0		No
MO	Item 23 - Fraud: Non-Covered Use	0		No
MP	Item 23 - Fraud: Non-Covered Damage	0		No
MQ	Item 23 - Disallowed: Fraud	0		No
MR	Item 23 - Disallowed: Under Deductible	0		No
MS	Item 23 - Disallowed: Not Covered	0		No
MT	Item 23 - Non-Full: Over Limit	0		No
MU	Item 23 - Non-Full: Depreciation	0		No
MV	Item 23 - Non-Full: Over Market Price	0		No
MW	Item 24 - Silverware: \$Loss Claimed	0		No
MX	Item 24: \$Loss Allowed	0		No

Column	Field Name	Sample Value	Comment 1	In Kaggle
MY	Item 24 - Fraud: Overstated Value	0		No
MZ	Item 24 - Fraud: Intentional Damage	0		No
NA	Item 24 - Fraud: Fake Theft	0		No
NB	Item 24 - Fraud: Fake Repair Bills	0		No
NC	Item 24 - Fraud: Inflated Repair Bills	0		No
ND	Item 24 - Fraud: Non-Covered Use	0		No
NE	Item 24 - Fraud: Non-Covered Damage	0		No
NF	Item 24 - Disallowed: Fraud	0		No
NG	Item 24 - Disallowed: Under Deductible	0		No
NH	Item 24 - Disallowed: Not Covered	0		No
NI	Item 24 - Non-Full: Over Limit	0		No
NJ	Item 24 - Non-Full: Depreciation	0		No
NK	Item 24 - Non-Full: Over Market Price	0		No
NL	Item 25 - Tools: \$Loss Claimed	0		No
NM	Item 25: \$Loss Allowed	0		No
NN	Item 25 - Fraud: Overstated Value	0		No
NO	Item 25 - Fraud: Intentional Damage	0		No
NP	Item 25 - Fraud: Fake Theft	0		No
NQ	Item 25 - Fraud: Fake Repair Bills	0		No
NR	Item 25 - Fraud: Inflated Repair Bills	0		No
NS	Item 25 - Fraud: Non-Covered Use	0		No

Column	Field Name	Sample Value	Comment 1	In Kaggle
NT	Item 25 - Fraud: Non-Covered Damage	0		No
NU	Item 25 - Disallowed: Fraud	0		No
NV	Item 25 - Disallowed: Under Deductible	0		No
NW	Item 25 - Disallowed: Not Covered	0		No
NX	Item 25 - Non-Full: Over Limit	0		No
NY	Item 25 - Non-Full: Depreciation	0		No
NZ	Item 25 - Non-Full: Over Market Price	0		No
OA	Item 26 - Construction Material: \$Loss Claimed	0		No
OB	Item 26: \$Loss Allowed	0		No
OC	Item 26 - Fraud: Overstated Value	0		No
OD	Item 26 - Fraud: Intentional Damage	0		No
OE	Item 26 - Fraud: Fake Theft	0		No
OF	Item 26 - Fraud: Fake Repair Bills	0		No
OG	Item 26 - Fraud: Inflated Repair Bills	0		No
OH	Item 26 - Fraud: Non-Covered Use	0		No
OI	Item 26 - Fraud: Non-Covered Damage	0		No
OJ	Item 26 - Disallowed: Fraud	0		No
OK	Item 26 - Disallowed: Under Deductible	0		No
OL	Item 26 - Disallowed: Not Covered	0		No
OM	Item 26 - Non-Full: Over Limit	0		No



Column	Field Name	Sample Value	Comment 1	In Kaggle
ON	Item 26 - Non-Full: Depreciation	0		No
OO	Item 26 - Non-Full: Over Market Price	0		No
OP	Item 27 - Books & Magazines: \$Loss Claimed	0		No
OQ	Item 27: \$Loss Allowed	0		No
OR	Item 27 - Fraud: Overstated Value	0		No
OS	Item 27 - Fraud: Intentional Damage	0		No
OT	Item 27 - Fraud: Fake Theft	0		No
OU	Item 27 - Fraud: Fake Repair Bills	0		No
OV	Item 27 - Fraud: Inflated Repair Bills	0		No
OW	Item 27 - Fraud: Non-Covered Use	0		No
OX	Item 27 - Fraud: Non-Covered Damage	0		No
OY	Item 27 - Disallowed: Fraud	0		No
OZ	Item 27 - Disallowed: Under Deductible	0		No
PA	Item 27 - Disallowed: Not Covered	0		No
PB	Item 27 - Non-Full: Over Limit	0		No
PC	Item 27 - Non-Full: Depreciation	0		No
PD	Item 27 - Non-Full: Over Market Price	0		No
PE	Item 28 - Sporting Goods: \$Loss Claimed	0		No
PF	Item 28: \$Loss Allowed	0		No
PG	Item 28 - Fraud: Overstated Value	0		No

Column	Field Name	Sample Value	Comment 1	In Kaggle
PH	Item 28 - Fraud: Intentional Damage	0		No
PI	Item 28 - Fraud: Fake Theft	0		No
PJ	Item 28 - Fraud: Fake Repair Bills	0		No
PK	Item 28 - Fraud: Inflated Repair Bills	0		No
PL	Item 28 - Fraud: Non-Covered Use	0		No
PM	Item 28 - Fraud: Non-Covered Damage	0		No
PN	Item 28 - Disallowed: Fraud	0		No
PO	Item 28 - Disallowed: Under Deductible	0		No
PP	Item 28 - Disallowed: Not Covered	0		No
PQ	Item 28 - Non-Full: Over Limit	0		No
PR	Item 28 - Non-Full: Depreciation	0		No
PS	Item 28 - Non-Full: Over Market Price	0		No
PT	Item 29 - Golf Cart: \$Loss Claimed	0		No
PU	Item 29: \$Loss Allowed	0		No
PV	Item 29 - Fraud: Overstated Value	0		No
PW	Item 29 - Fraud: Intentional Damage	0		No
PX	Item 29 - Fraud: Fake Theft	0		No
PY	Item 29 - Fraud: Fake Repair Bills	0		No
PZ	Item 29 - Fraud: Inflated Repair Bills	0		No
QA	Item 29 - Fraud: Non-Covered Use	0		No

Column	Field Name	Sample Value	Comment 1	In Kaggle
QB	Item 29 - Fraud: Non-Covered Damage	0		No
QC	Item 29 - Disallowed: Fraud	0		No
QD	Item 29 - Disallowed: Under Deductible	0		No
QE	Item 29 - Disallowed: Not Covered	0		No
QF	Item 29 - Non-Full: Over Limit	0		No
QG	Item 29 - Non-Full: Depreciation	0		No
QH	Item 29 - Non-Full: Over Market Price	0		No
QI	Item 30 - Cameras: \$Loss Claimed	0		No
QJ	Item 30: \$Loss Allowed	0		No
QK	Item 30 - Fraud: Overstated Value	0		No
QL	Item 30 - Fraud: Intentional Damage	0		No
QM	Item 30 - Fraud: Fake Theft	0		No
QN	Item 30 - Fraud: Fake Repair Bills	0		No
QO	Item 30 - Fraud: Inflated Repair Bills	0		No
QP	Item 30 - Fraud: Non-Covered Use	0		No
QQ	Item 30 - Fraud: Non-Covered Damage	0		No
QR	Item 30 - Disallowed: Fraud	0		No
QS	Item 30 - Disallowed: Under Deductible	0		No
QT	Item 30 - Disallowed: Not Covered	0		No
QU	Item 30 - Non-Full: Over Limit	0		No

Column	Field Name	Sample Value	Comment 1	In Kaggle
QV	Item 30 - Non-Full: Depreciation	0		No
QW	Item 30 - Non-Full: Over Market Price	0		No
QX	Item 31 - Watches: \$Loss Claimed	0		No
QY	Item 31: \$Loss Allowed	0		No
QZ	Item 31 - Fraud: Overstated Value	0		No
RA	Item 31 - Fraud: Intentional Damage	0		No
RB	Item 31 - Fraud: Fake Theft	0		No
RC	Item 31 - Fraud: Fake Repair Bills	0		No
RD	Item 31 - Fraud: Inflated Repair Bills	0		No
RE	Item 31 - Fraud: Non-Covered Use	0		No
RF	Item 31 - Fraud: Non-Covered Damage	0		No
RG	Item 31 - Disallowed: Fraud	0		No
RH	Item 31 - Disallowed: Under Deductible	0		No
RI	Item 31 - Disallowed: Not Covered	0		No
RJ	Item 31 - Non-Full: Over Limit	0		No
RK	Item 31 - Non-Full: Depreciation	0		No
RL	Item 31 - Non-Full: Over Market Price	0		No
RM	Item 32 - Furs: \$Loss Claimed	0		No
RN	Item 32: \$Loss Allowed	0		No
RO	Item 32 - Fraud: Overstated Value	0		No
RP	Item 32 - Fraud: Intentional Damage	0		No

Column	Field Name	Sample Value	Comment 1	In Kaggle
RQ	Item 32 - Fraud: Fake Theft	0		No
RR	Item 32 - Fraud: Fake Repair Bills	0		No
RS	Item 32 - Fraud: Inflated Repair Bills	0		No
RT	Item 32 - Fraud: Non-Covered Use	0		No
RU	Item 32 - Fraud: Non-Covered Damage	0		No
RV	Item 32 - Disallowed: Fraud	0		No
RW	Item 32 - Disallowed: Under Deductible	0		No
RX	Item 32 - Disallowed: Not Covered	0		No
RY	Item 32 - Non-Full: Over Limit	0		No
RZ	Item 32 - Non-Full: Depreciation	0		No
SA	Item 32 - Non-Full: Over Market Price	0		No
SB	Item 33 - Medical Instruments: \$Loss Claimed	0		No
SC	Item 33: \$Loss Allowed	0		No
SD	Item 33 - Fraud: Overstated Value	0		No
SE	Item 33 - Fraud: Intentional Damage	0		No
SF	Item 33 - Fraud: Fake Theft	0		No
SG	Item 33 - Fraud: Fake Repair Bills	0		No
SH	Item 33 - Fraud: Inflated Repair Bills	0		No
SI	Item 33 - Fraud: Non-Covered Use	0		No
SJ	Item 33 - Fraud: Non-Covered Damage	0		No

Column	Field Name	Sample Value	Comment 1	In Kaggle
SK	Item 33 - Disallowed: Fraud	0		No
SL	Item 33 - Disallowed: Under Deductible	0		No
SM	Item 33 - Disallowed: Not Covered	0		No
SN	Item 33 - Non-Full: Over Limit	0		No
SO	Item 33 - Non-Full: Depreciation	0		No
SP	Item 33 - Non-Full: Over Market Price	0		No
SQ	Item 34 - Musical Instruments: \$Loss Claimed	0		No
SR	Item 34: \$Loss Allowed	0		No
SS	Item 34 - Fraud: Overstated Value	0		No
ST	Item 34 - Fraud: Intentional Damage	0		No
SU	Item 34 - Fraud: Fake Theft	0		No
SV	Item 34 - Fraud: Fake Repair Bills	0		No
SW	Item 34 - Fraud: Inflated Repair Bills	0		No
SX	Item 34 - Fraud: Non-Covered Use	0		No
SY	Item 34 - Fraud: Non-Covered Damage	0		No
SZ	Item 34 - Disallowed: Fraud	0		No
TA	Item 34 - Disallowed: Under Deductible	0		No
TB	Item 34 - Disallowed: Not Covered	0		No
TC	Item 34 - Non-Full: Over Limit	0		No
TD	Item 34 - Non-Full: Depreciation	0		No

Column	Field Name	Sample Value	Comment 1	In Kaggle
TE	Item 34 - Non-Full: Over Market Price	0		No
TF	Item 35 - Other Personal Property: \$Loss Claimed	0		No
TG	Item 35: \$Loss Allowed	0		No
TH	Item 35 - Fraud: Overstated Value	0		No
TI	Item 35 - Fraud: Intentional Damage	0		No
TJ	Item 35 - Fraud: Fake Theft	0		No
TK	Item 35 - Fraud: Fake Repair Bills	0		No
TL	Item 35 - Fraud: Inflated Repair Bills	0		No
TM	Item 35 - Fraud: Non-Covered Use	0		No
TN	Item 35 - Fraud: Non-Covered Damage	0		No
TO	Item 35 - Disallowed: Fraud	0		No
TP	Item 35 - Disallowed: Under Deductible	0		No
TQ	Item 35 - Disallowed: Not Covered	0		No
TR	Item 35 - Non-Full: Over Limit	0		No
TS	Item 35 - Non-Full: Depreciation	0		No
TT	Item 35 - Non-Full: Over Market Price	0		No

Table 12 is the data schema for insurance freetext.

Table 12 Data schema for insurance freetext

Column	Field Name	Sample Value	Comment	In Kaggle
A	Insured Claim Narrative	I hope this damage is covered. Please have somebody come to my house. Suddenly I heard something. There were howling winds on Jul 18. I still have to check things, but here is what I think is lost or damaged: lots of the house and the roof pieces. The losses totaled \$12380. How much is my deductible? My house needs help. My house has 6 bedrooms. My house has 2350 square feet. This is urgent.	These columns in freetext can be viewed as extensions to the columns in "insur_claims.csv", that is, there is a 1:1 mapping of rows here to rows in the claims file.	No
B	Generic Request for a Person	0	Labels about the attributes of the narrative. 1 means that the entry is an instance of the specified type, for example, "Request for a Person". 0 means that the entry is not an instance. More than one field can be 1, and more than one field can be 0. Indeed, the fields can be all 1s or all 0s.	No
C	Request only a Person can Answer	1	Labels about the attributes of the narrative. 1 means that the entry is an instance of the specified type, for example, "Request only a Person can Answer". 0 means that the entry is not an instance. More than one field can be 1, and more than one field can be 0. Indeed, the fields can be all 1s or all 0s.	No



Column	Field Name	Sample Value	Comment	In Kaggle
D	Request for a Fact	1	Labels about the attributes of the narrative. 1 means that the entry is an instance of the specified type, for example, "Request for a Fact". 0 means that the entry is not an instance. More than one field can be 1, and more than one field can be 0. Indeed, the fields can be all 1s or all 0s.	No
E	Request for Advice	0	Labels about the attributes of the narrative. 1 means that the entry is an instance of the specified type, for example, "Request for Advice". 0 means that the entry is not an instance. More than one field can be 1, and more than one field can be 0. Indeed, the fields can be all 1s or all 0s.	No
F	Request for a Prediction	0	Labels about the attributes of the narrative. 1 means that the entry is an instance of the specified type, for example, "Request for a Prediction". 0 means that the entry is not an instance. More than one field can be 1, and more than one field can be 0. Indeed, the fields can be all 1s or all 0s.	No

Table 13 is the data schema for storms.

Table 13 Data schema for storms

Column	Field Name	Sample Value	Comment	In Kaggle
A	Date	44577	Date of Quake	No
B	Magnitude	6.7	On Richter Scale	No
C	Latitude	44.3678	North of Boise, Idaho	No
D	Longitude	-116.2110		No
E	Country	United States		No
NOTE:	File has a list of all significant earthquakes during transaction period where significant insurance losses may be incurred			
All quakes are simulated -- but follow geographic and temporal likelihood of earthquakes				

Table 14 is the data schema for quakes.

Table 14 Data schema for quakes

Column	Field Name	Sample Value	Comment	In Kaggle
A	Date	44577	Date of Quake	No
B	Magnitude	6.7	On Richter Scale	No
C	Latitude	44.3678	North of Boise, Idaho	No
D	Longitude	-116.2110		No
E	Country	United States		No
NOTE:	File has a list of all significant earthquakes during transaction period where significant insurance losses may be incurred			
All quakes are simulated -- but follow geographic and temporal likelihood of earthquakes				

Table 15 is the data schema for volcanoes.

Table 15 Data schema for volcanoes

Excel Column	Field Name	Sample Value	Comment	In Kaggle
A	Date	45123		No
B	Volcano	Three Sisters	Name of Volcano	No
C	VEI	4	Volcanic Explosivity Index	No
D	Latitude	44.1033	West of Bend, Oregon	No
E	Longitude	-121.7692		No
F	Country	United States		No
NOTE:	File has a list of all significant and unexpected volcanic eruptions during transaction period where significant insurance losses may be incurred			
	All eruptions are simulated -- but follow geographic and temporal likelihood of eruptions			
	There are 169 active volcanos in the United States. It is very rare to have an unexpected eruption, and for most data sets there will be no eruptions.			





REDP-5748-00

ISBN 0738461997

Printed in U.S.A.

Get connected

