FINANCIAL BEHAVIOR AND RISK ASSESSMENT

Credit Card Data Analysis

TABLE OF CONTENTS

Sourcing Open Data - Summary

Data Source: External Kaggle /IBM synthetic credit card dataset

Credit Card Transactions Dataset

Summary of Data Source

The dataset is an external dataset downloaded from Kaggle, originally generated by IBM Research as part of a multi-agent simulation designed to mimic real world credit card transactions and fraud patterns. Since it is synthetic, it does not come from actual banks or card providers, but it was designed to closely replicate realistic patterns. The dataset can be considered trustworthy for fraud detection modeling and conducting transaction analysis.

Data Collection Method

The data was not collected from real world subjects but instead generated automatically through simulation. The simulation includes 2,000 synthetic consumers making purchases worldwide across multiple decades, logging details such as card type, amount, merchant info, demographic, and fraud label. Since it is synthetic, there are no missing coverage issues.

Variables in the Dataset

The dataset is spread across three CSV files:

- Transaction Data: Includes user, card, year, month, day, time, amount, use chip, merchant name, merchant city, merchant state, zip, transaction amount, MCC (merchant category code), errors label and fraud label.
- Card Info: Includes user, card index, card brand, card type, card number, expires, CVV, has chip, cards issued, credit limit, account open date, year PIN last changed, and card on dark web.
- User Info: Includes person, current age, retirement age, birth year, birth month, gender, address, apartment, city, state, zip code, latitude, longitude, per capita income – zip code, yearly income – person, total debt, FICO score, and num credit cards.

Limitations of the Dataset

The data does not perfectly represent real world fraud behaviors, as consumer behavior may appear overly clean due to its synthetic nature. Fraud labels are determined by simulation model rather than actual criminal activity, while amounts, fraud rates, and merchant category codes (MCC) are realistic, there could still be biases in how fraud patterns were modeled. In addition, some fields are anonymized, meaning the data set cannot be used to infer sensitive insights about real people since no personally identifiable information is included.

Data Relevancy

The dataset is relevant for exploring fraud detection modeling and consumer spending behavior analysis. It allows for practice in joining multiple datasets and working with fraud detection. However, findings should not be based as real world insights but rather simulation-based insights.

Reason for Choosing the Dataset

I chose this dataset because it provides a rich combination of demographic, financial, and transaction-level data that makes it possible to analyze consumer spending behavior and assess user risk profiles. This project supports my goal of pursuing an analyst role in a financial institution, providing a portfolio piece that demonstrates skills in financial data analysis, geospatial analytics, and risk profiling.

Data Profile

Basic Structure

Legend

QN = Quantitative	TV = Time variant	TI = Time invariant	C = Continuous	D = Discrete
QL = Qualitative	O = Ordinal	N = Nominal	B = Binary	

Transaction Data Table

No.	Column	Description	Dtype	Data Type
0	User	Unique numerical value given to each user.	Int64	QL / N / TI
1	Card	Number representing card used by user.	Int64	QL / N / TI
2	Year	A period of the duration of a calendar year.	Int64	QN / D / TV
3	Month	A period of the 12 divisions of a calendar year.	Int64	QN / D / TV
4	Day	A period of division within months of calendar year.	Int64	QN / D / TV
5	Time	A period with the division of a day.	Object	QN / D / TV
6	Amount	Quantity Spent.	Object	QN / C / TI
7	Use Chip	Type of transaction.	Object	QL / B / TV
8	Merchant Name	Value represents the name of the seller.	Int64	QL / N / TV
9	Merchant City	City in which the seller is located.	Object	QL / N / TV
10	Merchant State	State in which the seller is located.	Object	QL / N / TV
11	Zip	Zip code in which the seller is located.	Float64	QL / N / TV
12	MCC	4-digit code assigned to merchant classifying goods and services provided.	Int64	QL / N / TV
13	Errors?	Was there an error at the time of transacting?	Object	QL / B / TV
14	Is Fraud?	Is the transaction deemed fraudulent?	Object	QL /B /TV

- Number of Rows: 24,386,900
- Number of Columns: 15

Card Data Table

No.	Column	Description	Dtype	Data Type
0	User	Unique numerical value given to each user.	Int64	QL / N / TI
1	CARD INDEX	Number representing cards owned by user.	Int64	QN / D / TI
2	Card Brand	Payment card network that facilitates transactions.	Object	QL / N / TI
3	Card Type	Referring to the specific category the card belongs to.	Object	QL / N / TI
4	Card Number	Primary account number that is a unique card identifier	Int64	QL / N / TI
5	Expires	Expiration date on card	Object	QN / D / TI
6	CVV	Security code printed on cards to help prevent fraud during card not present transactions	Int64	QN / D / TI
7	Has Chip	Does the card have an integrated circuit?	Object	QL /B / TI
8	Cards Issued	Number of identical cards issued.	Int64	QN / D / TI
9	Credit Limit	Maximum limit of spending on card.	Object	QN / C / TI
10	Acct Open Date	Date the account was established.	Object	QN / D / TI
11	Year PIN last Changed	Year in which the cards security pin was last changed.	Int64	QN / D / TV

12	Card on Dark Web	Was there credit card information on the dark web?	Object	QL / B / TV

- Number of Rows: 6146
- Number of Columns: 13

User Data Table

No.	Column	Description	Dtype	Data Type
0	Person	Name of card user.	Object	QL / N / TI
1	Current Age	Age of person at time of collecting data.	Int64	QN / D / TV
2	Retirement Age	Expected persons retirement age.	Int64	QN / D / TI
3	Birth Year	Year in which the user was born.	Int64	QN / D / TI
4	Birth Month	Month in which the user was born.	Int64	QN /D / TI
5	Gender	Gender of the user.	Object	QL / N / TI
6	Address	Street number and name of user's address.	Object	QL / N / TI
7	Apartment	Apartment number if available.	Float64	QL / N / TI
8	City	City in which the user's address is located.	Object	QL / N / TI
9	State	State in which the user's address is located.	Object	QL / N /TI
10	Zipcode	Zip code in which the users address is located.	Int64	QL /N / TI
11	Latitude	A geographic coordinate that specifies the north-south position of a point from the equator.	Float64	QN / C / TI
12	Longitude	A geographic coordinate that specifies the east-west position of a point on the Earth's surface.	Float64	QN / C / TI
13	Per Capita Income – Zipcode	Measure of the avg. income earned per person in a specific zip code.	Object	QN / C / TV
14	Yearly Income – Person	User's annual income.	Object	QN / C / TV
15	Total Debt	User's total debt.	Object	QN / C / TV
16	FICO Score	User's credit score as per FICO Company.	Int64	QN / O / TV
17	Num Credit Cards	Number of credit cards user has issued.	Int64	QN / D / TV

- Number of Rows: 2000
- Number of Columns: 18

Descriptive Statistics on Raw Data

Transaction Data Table – Data Accuracy (Numerical Columns)

	User	Amount	Card	Year	Month	Day	Zip code
mean	1001.019	43.63401	1.351366	2012	6.525064	15.71812	50956.44
min	0	-500	0	1991	1	1	501
max	1999	12390.50	8	2016	12	31	99928
std	569.4612	82.02239	1.407154	5.105921	3.472355	8.794073	29397.07

Transaction Data Table – Data Consistency

Columns	Count	Notes:
User	24386900	No missing data
Card	24386900	No missing data
Year	24386900	No missing data
Month	24386900	No missing data
Day	24386900	No missing data
Time	24386900	No missing data
Amount	24386900	No missing data
Use Chip	24386900	No missing data
Merchant Name	24386900	Represented as a Numerical value, No missing data
Merchant City	24386900	Online Purchase appears as city. Will make new column.
Merchant State	21666079	Online orders not reflected.
Zip	21508765	Many online and foreign transactions don't have a zip code.
MCC	24386900	No missing data
Errors?	388431	This count shouldn't be high as most transactions should be successful.
Is Fraud?	24386900	No missing data

Card Data Table – Data Accuracy (Numerical Columns)

	User	CARD INDEX	Card Number	CVV	Cards Issued	Year PIN last Changed
mean	1003.477058	1.472502	4820426000000000	506	1.5	2013
min	0	0	3001055000000000	0	1	2002
max	1999	8	6997197000000000	999	3	2020
std	571.724745	1.463294	1328582000000000	289	0.519191	4.270699

Card Data Table – Data Consistency

Columns	Count	Notes:
User	6146	No missing data
CARD INDEX	6146	No missing data
Card Brand	6146	No missing data
Card Type	6146	No missing data
Card Number	6146	No missing data
Expires	6146	No missing data
CVV	6146	No missing data
Has Chip	6146	No missing data
Cards Issued	6146	No missing data
Credit Limit	6146	No missing data
Acct Open Date	6146	No missing data
Year PIN last Changed	6146	No missing data
Card on Dark Web	6146	No missing data

User Data Table – Data Accuracy (Numerical Columns)

	Current Age	Retirement Age	Birth Year	Birth Month	Apartment
mean	45.3915	66.2375	1974	6.439	693.547348
min	18	50	1918	1	1
max	101	79	2002	12	9940
std	18.414092	3.628867	18.421234	3.565338	1897.157861

	Zipcode	Latitude	Longitude	FICO Score	Num Credit Cards
mean	50535.412	37.389225	-91.554765	709.7345	3.073
min	1060	20.88	-159.41	480	1
max	99508	61.2	-68.67	850	9
std	29359.754742	5.114324	16.283293	67.221949	1.637379

User Data Table – Data Consistency

Columns	Count	Notes.
Person	2000	No missing data. 7 names repeated.
Current Age	2000	No missing data
Retirement Age	2000	No missing data
Birth Year	2000	No missing data
Birth Month	2000	No missing data
Gender	2000	No missing data
Address	2000	No missing data. 2 people might live in one address because 1999 is unique.
Apartment	528	No values missing. Not every address has an apartment number.
City	2000	No missing data
State	2000	No missing data
Zipcode	2000	No missing data
Latitude	2000	No missing data
Longitude	2000	No missing data
Per Capita Income - Zipcode	2000	No missing data
Yearly Income - Person	2000	No missing data
Total Debt	2000	No missing data
FICO Score	2000	No missing data
Num Credit Cards	2000	No missing data

Descriptive Statistics After 1st Cleaning

Transaction Data Table – Data Accuracy (Numerical Columns)

	txn_amount	card_index	txn_year	txn_month	txn_day	txn_zip_code
mean	43.63401	1.351366	2012	6.525064	15.71812	50956.44
min	-500	0	1991	1	1	501
max	12390.50	8	2016	12	31	99928
std	82.02239	1.407154	5.105921	3.472355	8.794073	29397.07

Transaction Data Table – Data Consistency

Columns	Count	Notes:
user_id	24386900	No missing data
card_index	24386900	No missing data
txn_year	24386900	No missing data
txn_month	24386900	No missing data
txn_day	24386900	No missing data
txn_date	24386900	No missing data
txn_time	24386900	No missing data
txn_amount	24386900	No missing data
txn_flag	24386900	No missing data
txn_type	24386900	No missing data. Some online purchases were labeled as Chip Transactions.
merchant_name	24386900	Represented as a Numerical value, No missing data
merchant_city	21666079	Removed all online purchases.
merchant_state	21508753	Online orders and foreign countries are not reflected.
merchant_country	21666079	Removed all online purchases.
txn_zip_code	21508765	Many online and foreign transactions don't have a zip code.
MCC	24386900	No missing data.
errors?	24386900	No missing data.
Is_fraud?	24386900	No missing data

Card Data Table – Data Accuracy (Numerical Columns)

	user_id	card_index	expires	cards_issued	card_limit	acct_open_date	year_PIN_last_changed
mean	1003.477058	1.472502	2020-10-08	1.5	14347	2011-01-15	2013
min	0	0	1997-07-01	1	0	1991-01-01	2002
max	1999	8	2024-12-01	3	151223	2020-02-01	2020
std	571.724745	1.463294	NaN	0.519191	12014.463884	NaN	4.270699

Card Data Table – Data Consistency

Columns	Count	Notes:
user_id	6146	No missing data
card_index	6146	No missing data
card_brand	6146	No missing data
card_type	6146	No missing data
card_number	6146	No missing data, must have 16 digits
expires	6146	No missing data, now datetime64
CVV	6146	No missing data, 3 or 4 digits no 2 digits
has_chip	6146	No missing data
cards_issued	6146	No missing data
credit_limit	6146	No missing data, now as float and no dollar sign.
acct_open_date	6146	No missing data, must have 16 digits
year_PIN_last_changed	6146	No missing data
card_on_dark_web	6146	No missing data

User Data Table – Data Accuracy (Numerical Columns)

	user_id	current_age	retirement_age	birth_year	birth_month	apartment
mean	999.5	45.3915	66.2375	1974	6.439	693.547348
min	0	18	50	1918	1	1
max	1999	101	79	2002	12	9940
std	577.494589	18.414092	3.628867	18.421234	3.565338	1897.157861

	user_zipcode	latitude	longitude	per_capita_income - zip_code	yearly_income - person
mean	50535.412	37.389225	-91.554765	23141.928	45715.882
min	1060	20.88	-159.41	0	1
max	99508	61.2	-68.67	163145	307018
std	29359.754742	5.114324	16.283293	11324.137358	22992.615456

	total_debt	FICO_score	num_credit_cards
mean	63709.694	709.7345	3.073
min	0	480	1
max	516263	850	9
std	52254.453421	67.221949	1.637379

User Data Table – Data Consistency

Columns	Count	Notes.
user_id	2000	Turned the index into the user id.
user_name	2000	No missing data. 7 names repeated.
current_age	2000	No missing data
retirement_age	2000	No missing data
birth_year	2000	No missing data
birth_month	2000	No missing data
gender	2000	No missing data
address	2000	No missing data. 2 people have same address but live in different states.
apartment	528	No values missing. Not every address has an apartment number.
city	2000	No missing data
state	2000	No missing data
user_zip_code	2000	No missing data
latitude	2000	No missing data
longitude	2000	No missing data
per_capita_income – zip-code	2000	No missing data
yearly_income - person	2000	No missing data
total_debt	2000	No missing data
FICO_score	2000	No missing data
num_credit_cards	2000	No missing data

Limitations and Ethical Considerations

Limitations

- Because the data is simulated, spending distributions, transaction frequencies, and demographic relationships may not reflect real-world credit card activity.
- The dataset is suitable for practicing analysis techniques but not for generating insights that can be applied to real-world business or policy decisions.

Privacy

- While credit card data is highly sensitive and in real cases could be re-identified, this dataset is simulated and contains no personal information.
- This dataset is simulated and contains no personal information, so privacy concerns are minimal.

Bias and Fairness

- Real-world credit card data may over- or under-represent certain demographics, but since this dataset is simulated, demographic patterns are not representative.
- Findings should be communicated responsibly, emphasizing that insights come from simulated data and should not be applied in decision-making for real customers.
- Insights should not be used to unfairly target or disadvantage vulnerable groups.

Transparency

- Provide clarity on how the data was cleaned and processed.
- Cleaning and preprocessing choices may oversimplify reality, but this is acceptable for simulated data exploration.

Defining Questions for Analysis Exploration

Spending Behavior Questions

Which merchant category (MCC) has the highest total spend?

How does spending varies across categories by month? Year?

What portions of transactions are in person vs. online?

Which states show the highest spending activities? Which countries?

How does spending differ across income brackets and credit score ranges?

What is the avg. transaction size, and how consistent is it across time?

Card Usage Questions

What portion of transactions are chip, swipe, or online?

Does the avg. transaction amount differ by transaction method?

Which card brands are used most frequently?

Do spending amounts vary across brands?

What is the average credit utilization rate (balance/credit limit) across users?

How many cards does the average customer hold?

Do customers with multiple cards distribute their spending across them?

What transaction type experiences more error? More Fraud?

User Risk Profiling Questions

Do high utilization users also have lower credit scores?

Are certain users engaging in unusually high number of small transactions?

Fraud Detection Questions

What percentage of transactions are labeled as fraudulent vs legitimate?

Are fraudulent transactions more likely to be online or in person?

Do fraud cases cluster by geography or are randomly distributed?

Does transaction type correlate with fraud likelihood?