BEHIND THE CARDS

An Exploratory Banking Data Analysis

Andromeda Canete-Borys
Data Analyst

MAY 2025

Executive Summary

This exploratory data analysis (EDA) investigates almost 10 years of transaction (13 million+), card, and user data from a banking institution. Using **Python**, **SQL**, and **Excel**, I have uncovered key patterns in user behavior, card ownership, and potential fraud.

The dataset was provided by Caixabank Tech for the 2024 AI Hackathon.

1 **CARD OWNERSHIP IS CONCENTRATED**

Over 80% of users hold 1–4 cards, and only 1% have 8 or more. More than 70% of users actively use all issued cards, indicating high engagement.

2 **More Cards ≠ Substantially More Spend**

There's a statistically significant but weak positive correlation between card count and both monthly transaction volume and spending. Card count alone is not a strong predictor of financial behavior.

3 **High Income Doesn't Always Mean More Cards**

Users earning above $100K do not consistently hold more cards. Additionally, female high earners outperform males by 29% in this bracket.

4 **Red Flags in High-Card Users**

Users with 8–9 cards show disproportionate discretionary spending, particularly in Betting & Casinos (MCC 7995). Some users exhibit suspicious patterns like geographically implausible transactions, warranting fraud investigation.

Steps in Data Analysis

1 Data Collection

The data combines transaction records, customer information, and card data from a banking institution, spanning 2010-2019.

Source: Kaggle

- transactions_data.csv (13,305,916 rows)
- cards_data.csv (6,146 rows)
- users_data.csv (2,000 rows)
- mcc_codes.json (109 rows)

2 Cleaning

Performed preprocessing using Python in Jupyter Notebook and MS Excel.

Cleaned data was then loaded to MySQL Workbench using command line script.

3 Exploration

Designed schemas, joined tables, and executed queries in MySQL Workbench; exported results to Excel for additional wrangling (as needed) or analysis using pivot tables.

Statistical analysis (Pearson correlation) was conducted in Python to examine the relationship between card count and other variables.

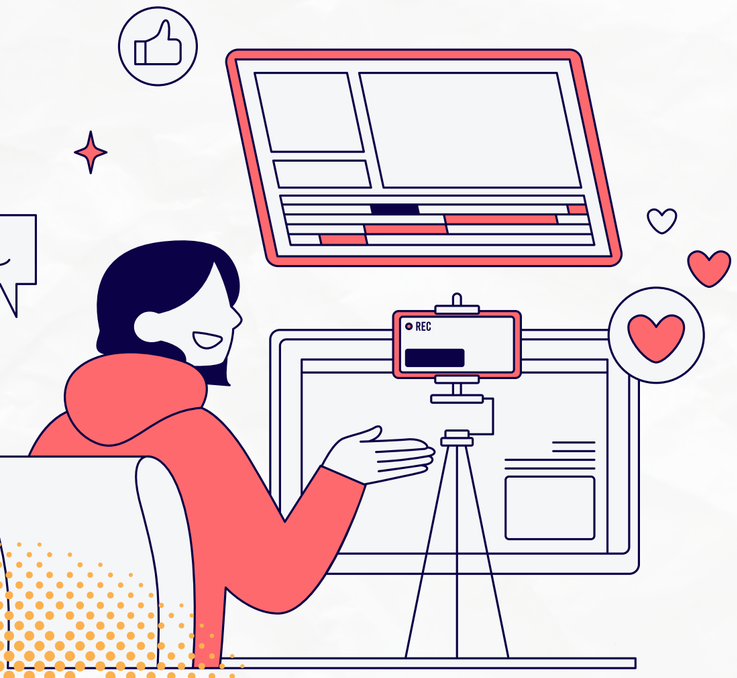
4 Data Visualization

Preliminary visuals were drafted in MS Excel for personal reference and assessment.

However, charts in this deck are from combination of tools - Python and Canva/Flourish.

SOME CODE EXAMPLES ARE SHOWN IN THE NEXT SLIDES

Data Cleaning using Python



jupyter CleaningTransactions Last Checkpoint: 14 days ago

File Edit View Run Kernel Settings Help

Code

[1]:
```
import pandas as pd

# Loading Transactions Table
df = pd.read_csv('transactions_data.csv')

# Clean the column: remove $
df['amount'] = df['amount'].replace(r'\$', '', regex=True)
```

[2]:
```
# Converting zip codes to String type
df['zip'] = df['zip'].astype(str)
```

[3]:
```
# Remove the '.0' in the zip code
df['zip'] = df['zip'].replace(r'\.0', '', regex=True)
```

[4]:
```
top_10 = df.head(10)
top_10
```

[4]:

	id	date	client_id	card_id	amount	use_chip
0	7475327	2010-01-01 00:01:00	1556	2972	-77.00	Swipe Transaction
1	7475328	2010-01-01 00:02:00	561	4575	14.57	Swipe Transaction
2	7475329	2010-01-01 00:02:00	1129	102	80.00	Swipe Transaction
3	7475331	2010-01-01 00:05:00	430	2860	200.00	Swipe Transaction
4	7475332	2010-01-01 00:06:00	848	3915	46.41	Swipe Transaction
5	7475333	2010-01-01 00:07:00	1807	165	4.81	Swipe Transaction

Transaction data has 13 million+ rows.

Therefore, data cleaning in Excel won't be possible.

Cleaned Transaction data using Python in Jupyter Notebook.

Pre-Processing Using SQL

Goal: Examine the Top 10 MCCs by card count bins

1. Create a temporary table with a new column to bin users into three groups depending on their number of cards.

```sql
CREATE TEMPORARY TABLE
bins AS
SELECT
    u.id AS user_id,
    u.yearly_income,
    COUNT(c.id) AS card_count,
    CASE
        WHEN COUNT(c.id) < 5 THEN "1-4 cards"
        WHEN COUNT(c.id) < 8 THEN "5-7 cards"
        ELSE "8-9 cards"
    END AS cardcount_group
FROM users u
LEFT JOIN cards c
    ON u.id = c.client_id
GROUP BY u.id,  u.yearly_income;
```

Pre-Processing Using SQL

Goal: Examine the Top 10 MCCs by card count bins

2 Getting the Top 10 MCCs per cardcount_group to analyze average spend

```sql
WITH ranked_spend AS (
    SELECT
        t.mcc,
        m.merchant,
        b.cardcount_group AS bin,
        AVG(t.amount) AS avg_spend,
        COUNT(DISTINCT t.client_id) as user_count,
        ROW_NUMBER() OVER (PARTITION BY b.cardcount_group ORDER BY SUM(t.amount) DESC) AS rn
    FROM transactions t
    LEFT JOIN mcc m ON t.mcc = m.mcc_code
    LEFT JOIN bins b ON t.client_id = b.user_id
    GROUP BY t.mcc, m.merchant, b.cardcount_group
)
SELECT
    mcc,
    merchant,
    bin,
    avg_spend,
    user_count
FROM ranked_spend
WHERE rn <= 10 -- Getting the Top 10 MCCs
ORDER BY bin, avg_spend DESC;
```

Visualization using Python

Goal: Plot the Top 10 Average Spend by MCC Per Card Bin

```python
import pandas as pd
import matplotlib.pyplot as plt

# Load your data
df = pd.read_csv('/Users/ams/Downloads/mcc-top10-avg.csv')
df['merchant'] = df['merchant'].str.strip()

# Pivot the data: rows are merchants, columns are bins, values are avg_spend
pivot_df = df.pivot(index='merchant', columns='bin', values='avg_spend')

# Sort by average spend across all bins to keep order consistent
pivot_df = pivot_df.loc[pivot_df.mean(axis=1).sort_values().index]

# Define a custom color palette including orange
colors = ['#ADEBB3', '#90D5FF', '#FF7F0E', '#D62728', '#9467BD', '#8C564B', '#E377C2', '#7F7F7F', '#BCBD22', '#17BECF']

# Plot the grouped horizontal bar chart with the custom color palette
pivot_df.plot(kind='barh', figsize=(8, 5), color=colors)
plt.xlabel('Average Spend ($)')
plt.title('Average Spend by Merchant Category and Card Bin')
plt.legend(title='Card Bin')
plt.tight_layout()
plt.show()
```

DATA EXPLORATION

Questions

1. Do users with more cards spend or transact more?
2. Do high-income earners tend to own more cards
3. Are there behavioral differences in users with 8 or more cards
4. Are there any suspicious or anomalous transaction patterns?

GENERAL INSIGHTS

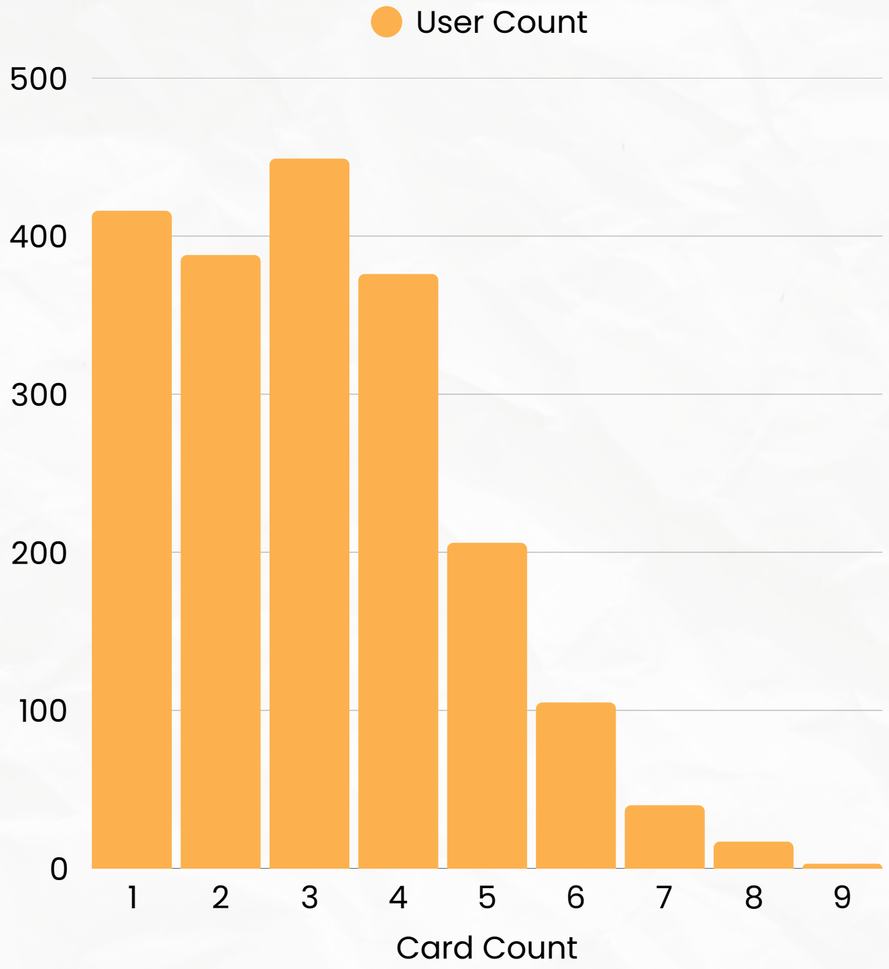
81% of users own 1-4 cards

18% of users own 5-7 cards

1% of users own 8+ cards

71% of users use all cards issued to them

NOTE: All card types are included - Debit, Credit, and Debit (Prepaid)

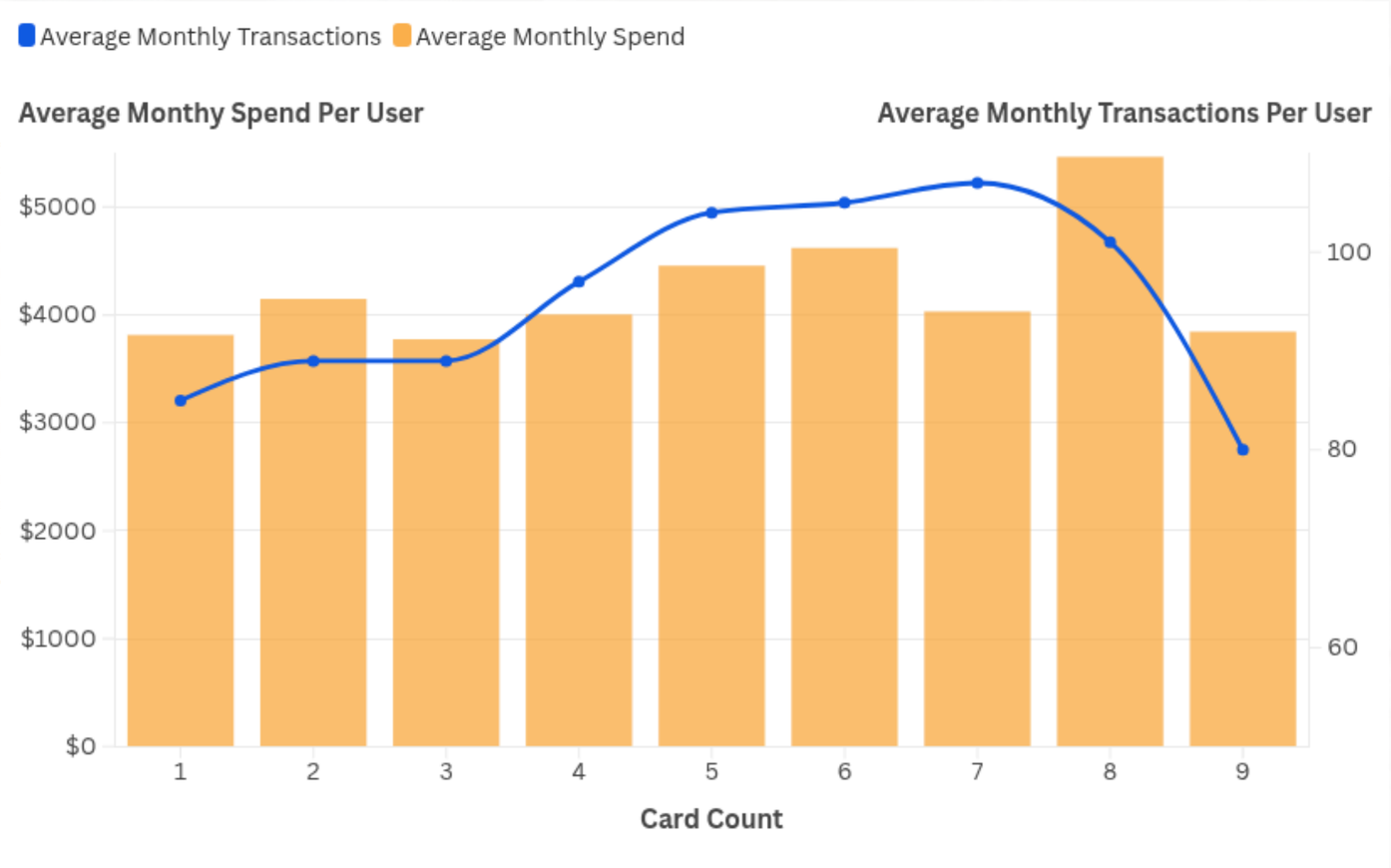


Highest annual income recorded is $307K, earned by a female user who has only two cards issued

Among users earning over $100K, females earn 29% more than males

TRANSACTIONS PER CARD COUNT

Transaction volume and spending tend to slightly rise with more cards, peaking somewhere between 5-6 cards.



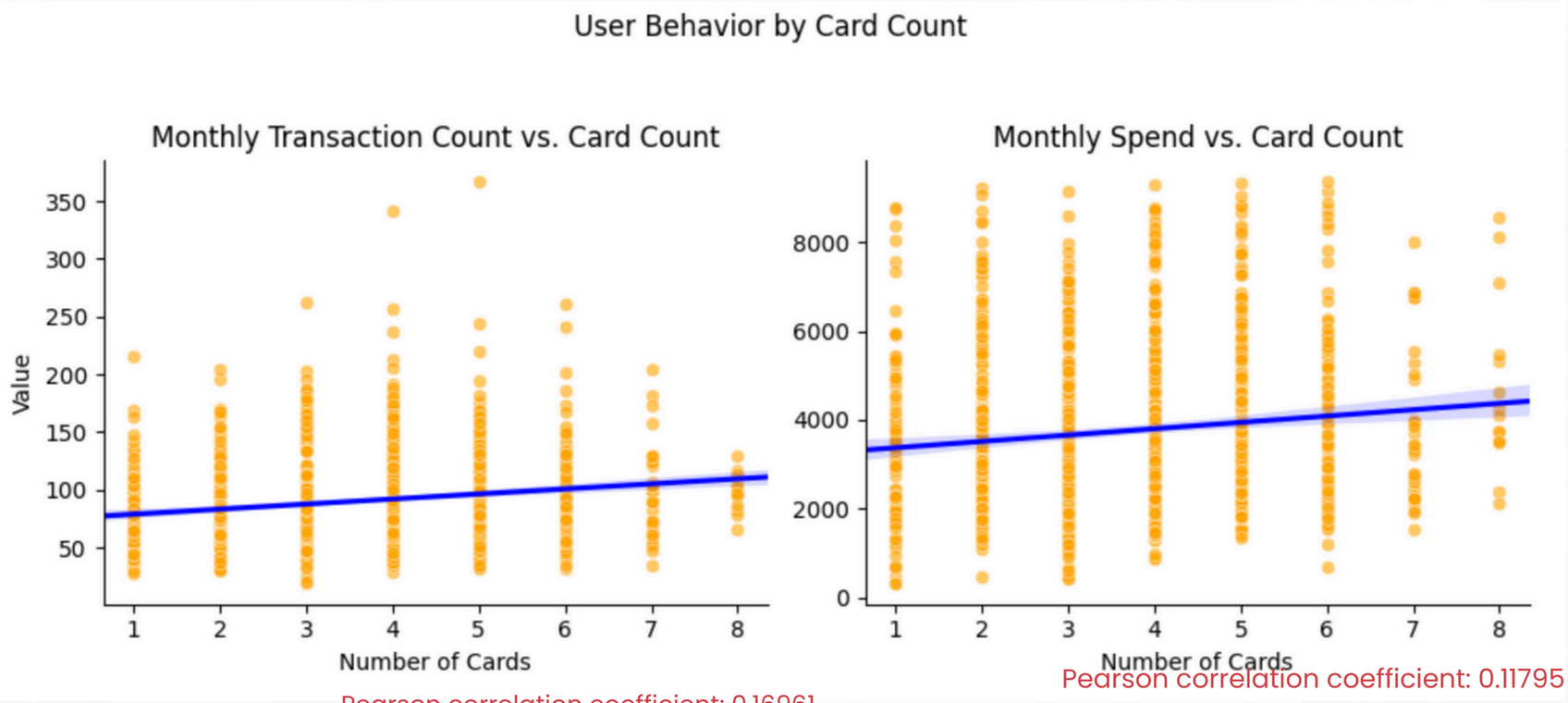
A typical user makes around 90 transactions a month and spends about $3,756 (outliers removed)

Users with 8 cards show unusually high spending, suggesting a need for further investigation.

See slide 13.

STATISTICAL TESTING

There is a statistically significant but **weak positive correlation** between card count and montly transaction volumne / monthly spend. Card count alone is **not a strong predictor** of these two variables.



Pearson correlation coefficient: 0.16961
P-value: 0.0000000056

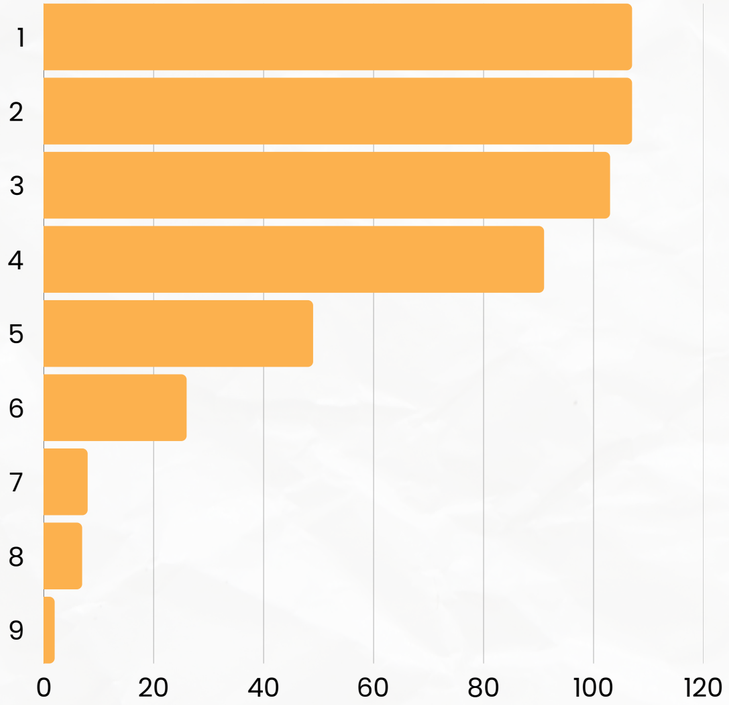
Pearson correlation coefficient: 0.11795
P-Value: 0.00005

** NOTE: Outliers were removed from the above charts. View full Python codes on Github.

HIGHER INCOME ≠ MORE CARDS

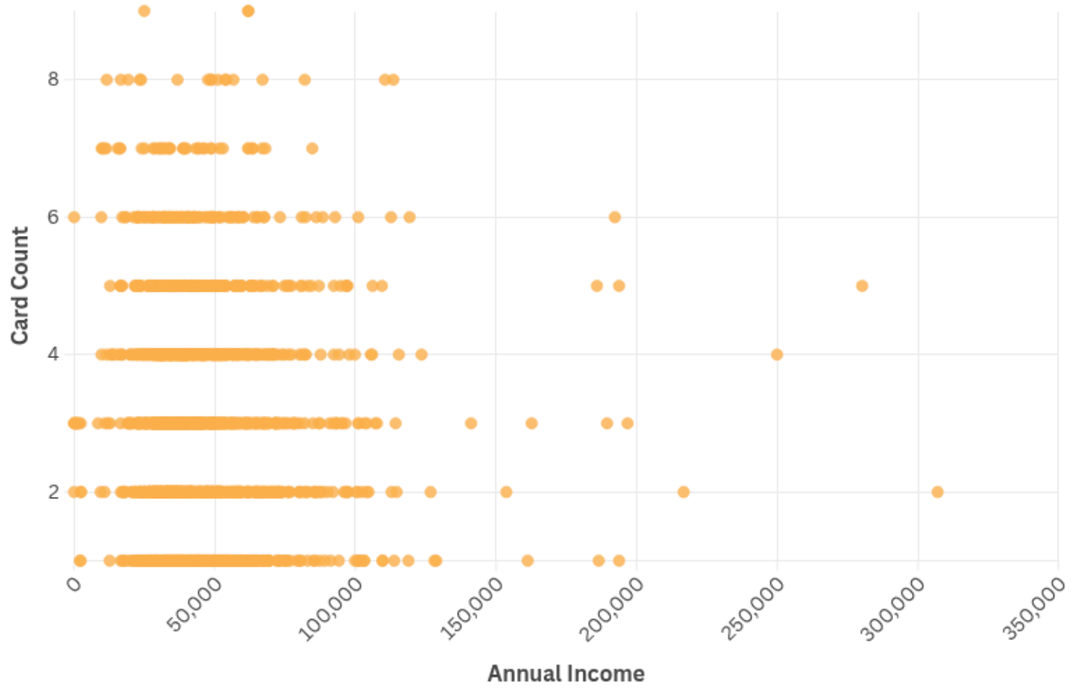
The number of cards per user do not necessarily increase as annual income increases.

Top 25% Income Tier: User Distribution Overview



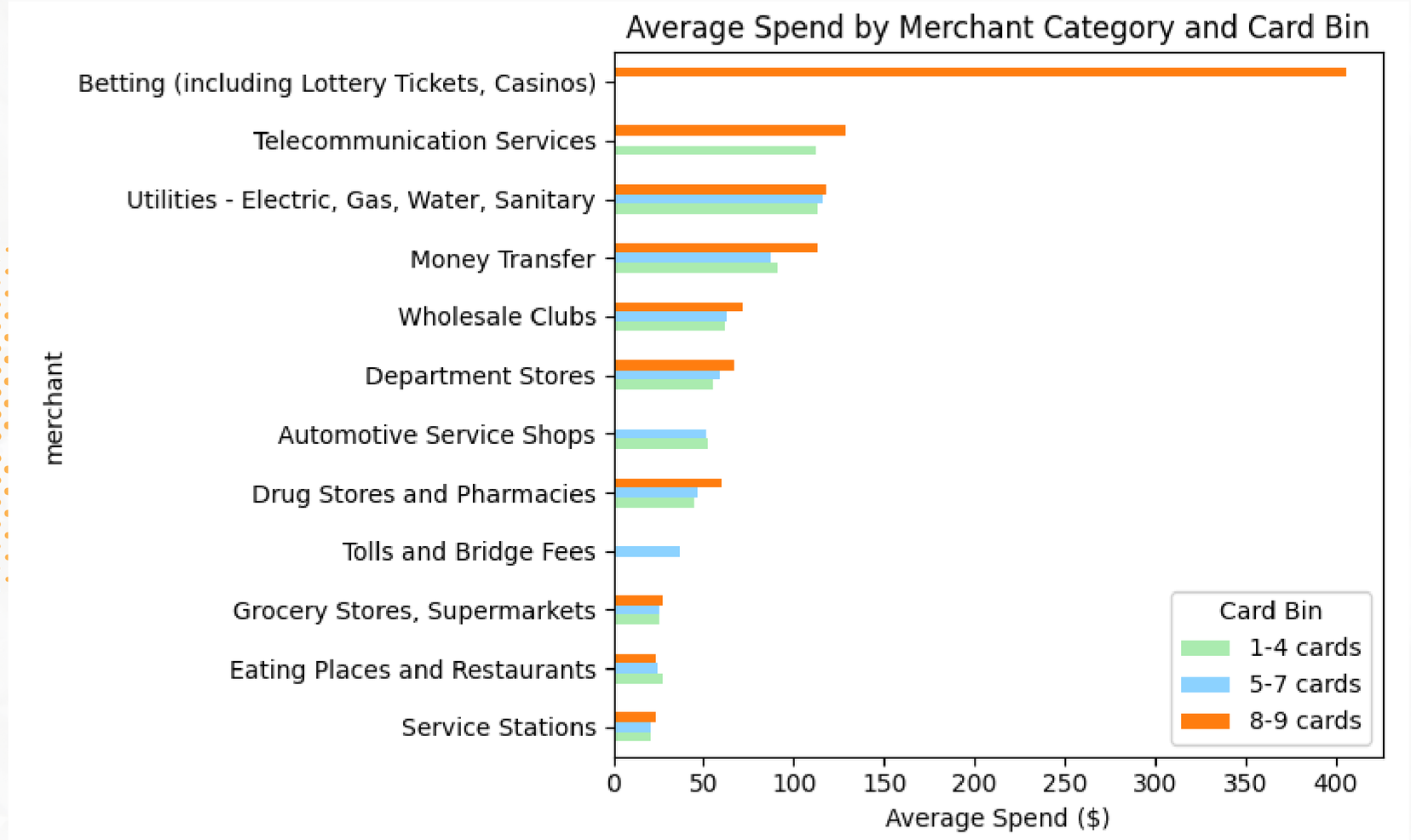
Users earning above $52,698 (75th percentile)

Annual Income vs. Card count (All Users)



** NOTE: View full Python codes on Github.

TOP 10 MERCHANT CATEGORY



High discretionary spend appears exclusively in the Top 10 for users with 8–9 cards, raising flags.

Next, review daily transactions across all bins under 7995: Betting (Lottery tickets and Casinos).

Look at top spenders as sample data.

** NOTE: View full Python codes on Github.

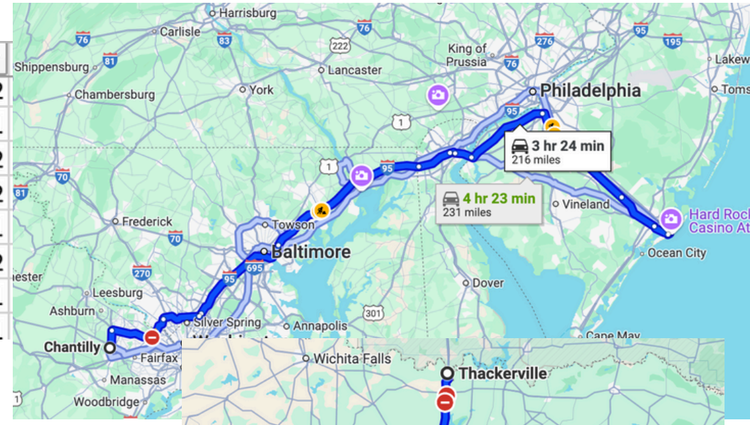
EXAMPLES OF SUSPICIOUS ACTIVITIES

Highlighted transactions occurred in far-apart locations within an hour for both top spenders—raising red flags for potential fraud or unauthorized access.

Recommending a full review by the fraud risk assessment team. There are likely more suspicious transactions under this MCC.

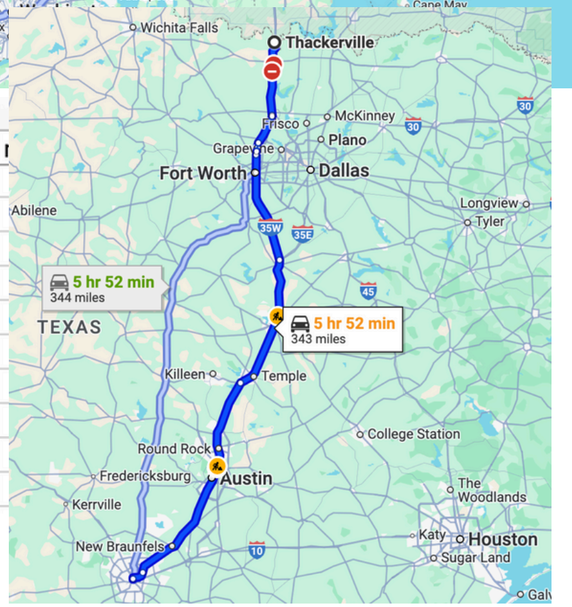
User 989, with 8 cards and a $113K income, spends ~$45K annually on this MCC — his top spending category (30% of annual spend)

date	client_id	card_id	amount	use_chip	merchant	merchant	merchant	zip
2015-03-27 1:21	989	2938	297.71	Chip Transaction	6699	Chantilly	VA	20152
2015-04-02 1:02	989	3039	452.67	Swipe Transaction	31883	Atlantic City	NJ	8401
2015-04-03 1:23	989	2511	388.87	Chip Transaction	6699	Chantilly	VA	20152
2015-04-03 1:40	989	4124	427.66	Chip Transaction	6699	Chantilly	VA	20152
2015-04-09 0:58	989	2511	404.46	Swipe Transaction	31883	Atlantic City	NJ	8401
2015-04-09 1:14	989	2938	393.59	Chip Transaction	6699	Chantilly	VA	20152
2015-04-10 1:15	989	4124	560.36	Swipe Transaction	31883	Atlantic City	NJ	8401
2015-04-16 1:01	989	4613	399.97	Swipe Transaction	31883	Atlantic City	NJ	8401



User 490, owning 4 cards, spends an average of $40K yearly on this MCC alone, exceeding annual income of $33,426

	B	C	D	E	F	G	H	I	J
	date	client_id	card_id	amount	use_chip	merchant	merchant	merchant	zip
	2010-01-09 14:29	490	3769	484.7	Swipe Transa	73661	Thackerville	OK	73459
	2010-01-15 14:00	490	5133	414.59	Swipe Transa	73661	Thackerville	OK	73459
	2010-01-16 14:25	490	176	333.2	Swipe Transa	14306	San Antonio	TX	78230
	2010-01-17 13:54	490	176	445.9	Swipe Transa	73661	Thackerville	OK	73459
	2010-01-17 14:59	490	5133	326.91	Swipe Transa	14306	San Antonio	TX	78230
	2010-01-23 14:15	490	3769	350.31	Swipe Transa	14306	San Antonio	TX	78230
	2010-02-03 13:42	490	5133	321.92	Swipe Transa	73661	Thackerville	OK	73459
	2010-02-03 14:45	490	176	328.94	Swipe Transa	14306	San Antonio	TX	78230
	2010-02-05 14:04	490	5963	356.28	Swipe Transa	14306	San Antonio	TX	78230
	2010-02-08 14:07	490	176	313.31	Swipe Transa	73661	Thackerville	OK	73459



Thank You

- 236-979-3140
- Burnaby, BC
- GitHub
- andromeda@canete.me