Capstone Project # 2

Recommendation System - Milestone 1

PowerPoint Slides - Recommendation_System_Milestone1
Code - EDA.ipynb

Problem Statement

Create a recommendation system using credit card transactions records. The data contains transactions from multiple people and contains GPS records. The aim is to provide the customer following

a. The recommendation system should recommend similar merchants to the customer based on similarities between the merchants
b. The recommendation system should recommend similar merchants to the customer based on similarities between users. The recommendation system should consider current customer's location as well in order to recommend the nearest similar merchants.

Data Mining/Data Wrangling

The data for this project came from FinGoal. FinGoal analyzes consumer spending to give radically insightful recommendations built on the experience of other savvy consumers (source: FinGoal.com). The data was anonymized and converted to json format by the company. The file size was 48.2MB. I used pd.read_json to read the file into a data frame. The file had 29 columns and 90086 rows of transactions.

1. **Dropping unnecessary rows and columns**:
   There were a lot of Null values and hidden duplicates in the data. For starters I deleted columns that will not be helpful in creating a recommendation system or were mostly empty. The columns deleted/dropped were

   'detail_category', 'high_level_category', 'transactionid', 'client_id', 'currency_code', 'country', 'pending', 'original_description' ,'category0', 'finsight_api_uid', 'finsight_image', 'insight_ctaurl', 'insight_text', 'original_uid', 'tenant_id', 'accountId'

   The merchant information was in the text of the column "simple_description". There were multiple rows that had exactly the same data (including the date, uid etc) except for transactionid. By deleting transactionid column I was able to delete those duplicates. Other columns were dropped because those were very thinly populated. Column "accountId" is a duplicate of "accountid" hence it was deleted.

2. **Dealing with NaN(Not a Number) values**

   From df.info() output I can see that there are several columns that could be used to fill NAN values of their "like" or “buddy” columns. For example: zip_code and city columns can be used to fill each other's NaN.

   Process: Find a row with NaN in zip_code (NaN) and record this row’s city (city1). Find another row where the city is set to city1 with corresponding zip_code(zip1). Replace NaN of the row with zip_code(zip1) as they both have the same city.
   I am assuming that each city has only one zip_code, which is not a reality, however for the purposes of the recommendation system it was a good option.

   The column ‘accountid’ and ‘uid’ could be considered as “like/buddy” columns but there were several rows that had ‘uid’ but no ‘accountid”. Therefore the logic presented above couldn’t completely fill the NaNs of ‘accountid’. The logic to fill NaN of accountid was to generate random numbers between 10449313 and 10478014 and apply to the rows with unique ‘uid’ that has no ‘accountid’. After this process column ‘accountid’ had no NaN in any of the rows.

3. **Category column**

   In order to go further in the EDA the category column needed to be understood. There were 66 unique values in the category column. The list of values dropped from the category column was:

   'Other Bills', 'Other Expenses', 'Paychecks/Salary', 'Tax' 'Refunds/Adjustments', Transfer', 'Transfers', 'Interest', 'Credit Card Payments', 'Uncategorized', 'Business Miscellaneous', 'Paychecks/Salary', 'Insurance', 'Savings', 'Other Income', 'Securities Trades', 'Deposits', 'Services', 'Tax', 'Taxes', 'Rent', 'Mortgages', 'Loans','Investment Income', 'ATM/Cash Withdrawals','Investment Income', 'Loans', 'Retirement Contributions','Expense Reimbursement', 'Retirement Income', 'Charitable Giving', 'Bank Fees', 'Wages Paid', 'Checks' ,'Rewards','Printing','Payment'

   Majority of the categories in the above list were related to incoming money(salaries, refunds, retirement income etc) or categories that cannot be used to predict better merchants for the user(loans, withdrawals, transfer, interest, charitable giving etc).

   There were 5 categories that had no categoryid associated with them. Since categoryid will be used in the recommendation system, I assigned five different values to the categories that were missing categoryid.

4. **Amountnum Column**

   The columns with less than or equal to zero amount were deleted. 75% of the transactions were under $52 and the number of rows with amounts of $6000 or more looked like outliers. Therefore those rows were dropped. There were several transactions that had $6000 charges however upon investigations it was found that those were duplicates with different dates and accountid.

Data Visualization

1.Number of Transactions per State

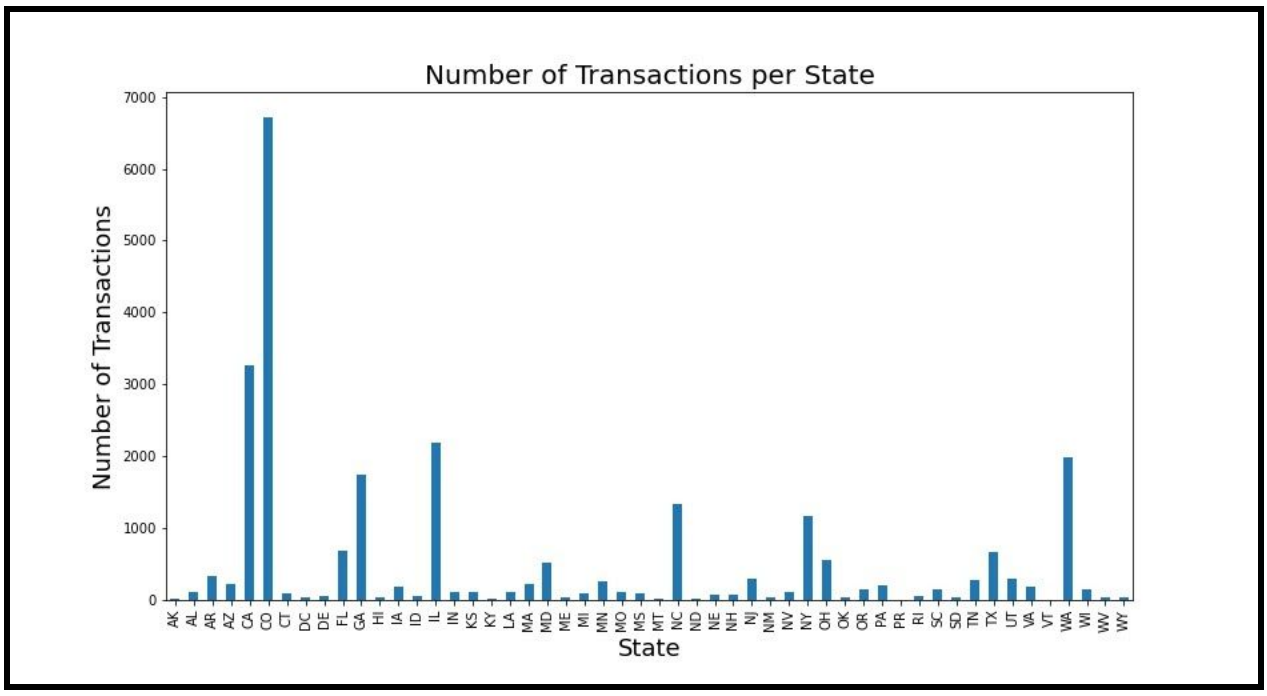

Figure 1: Number of Transactions per State

The most transactions are from Colorado, followed by California and Illinois. Colorado has around 7K transactions in the data which is almost 50% more than California. Since the FinGoal is in Colorado, the highest transactions in that state makes sense.

2. Number of Transactions per Category

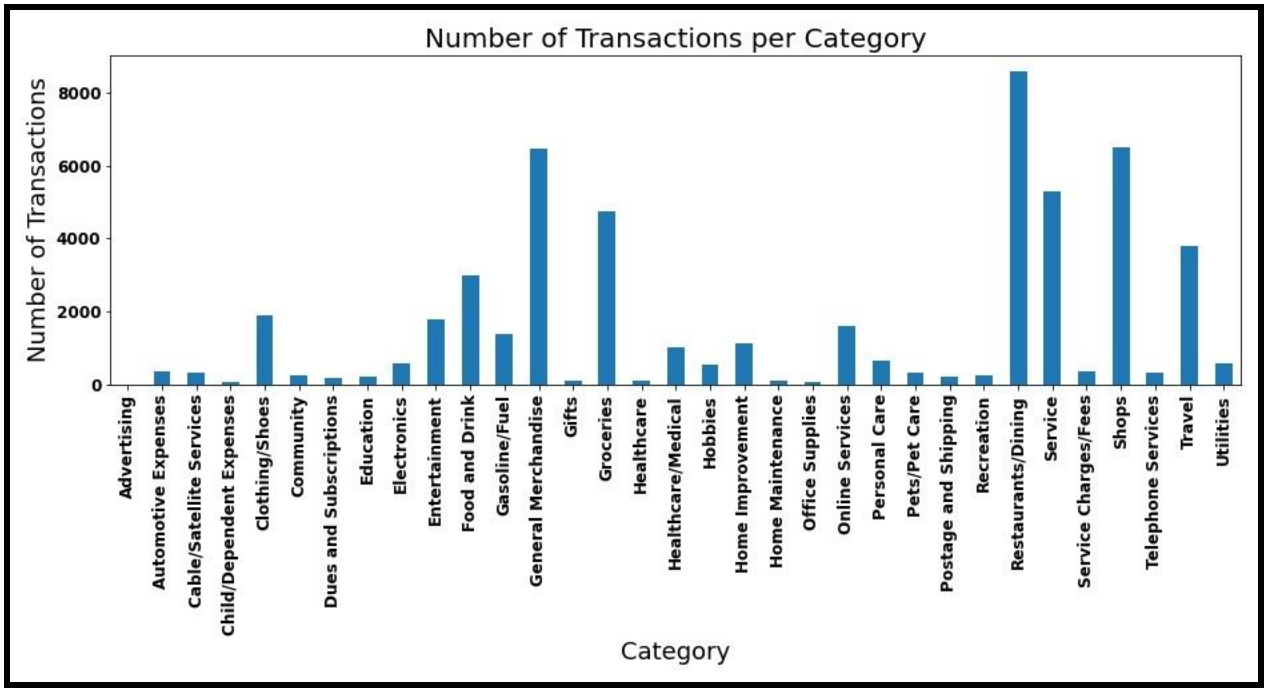

Figure 2: Number of Transactions per Category

From Figure 2 it can be seen that the majority of records are coming from the "Restaurants/Dining" category. Followed by "General Merchandise", "Shops", "Groceries", "Service", and "Travel".

3. Total Amount($) Spent Per Category

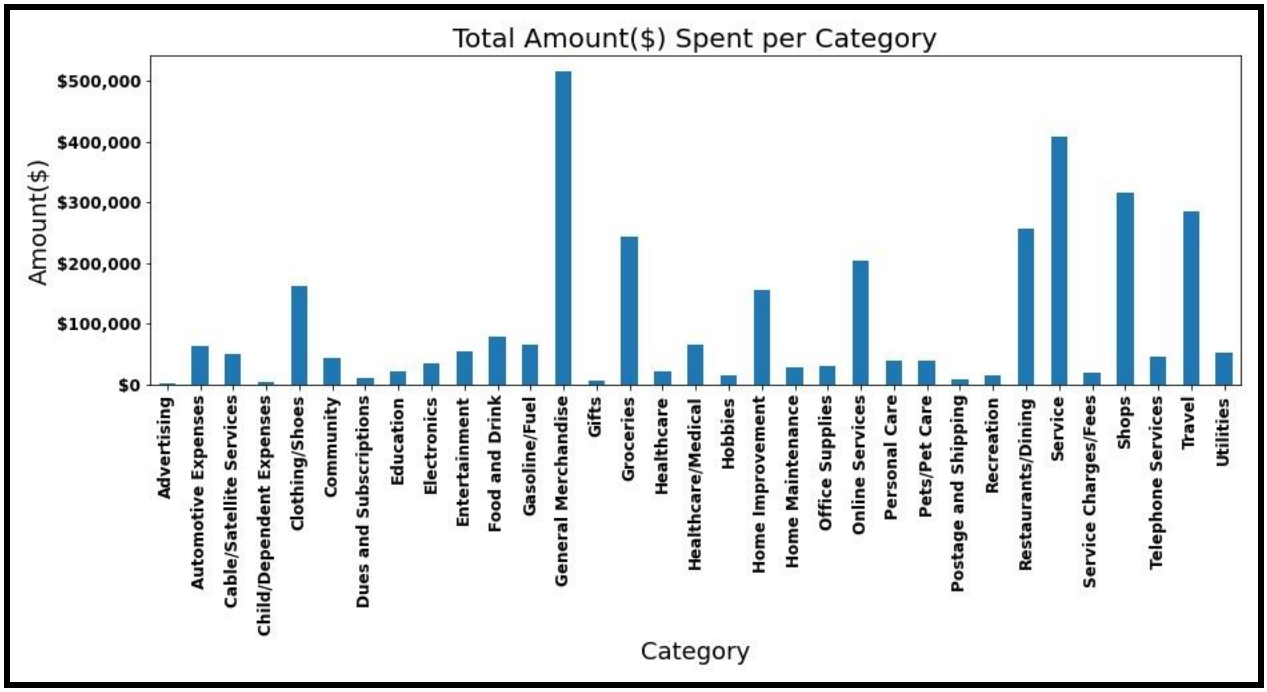

Figure 3: Total Amount($) Spent per Category

The category where most dollars were spent is "General Merchandise". "Service", "Travel", "Restaurants/Dining", "Groceries" come next. Since “General Merchandise” brought in the total of around $500K, looking further into the category and finding the highest earning merchants.

4. Zooming into General Merchandise Category

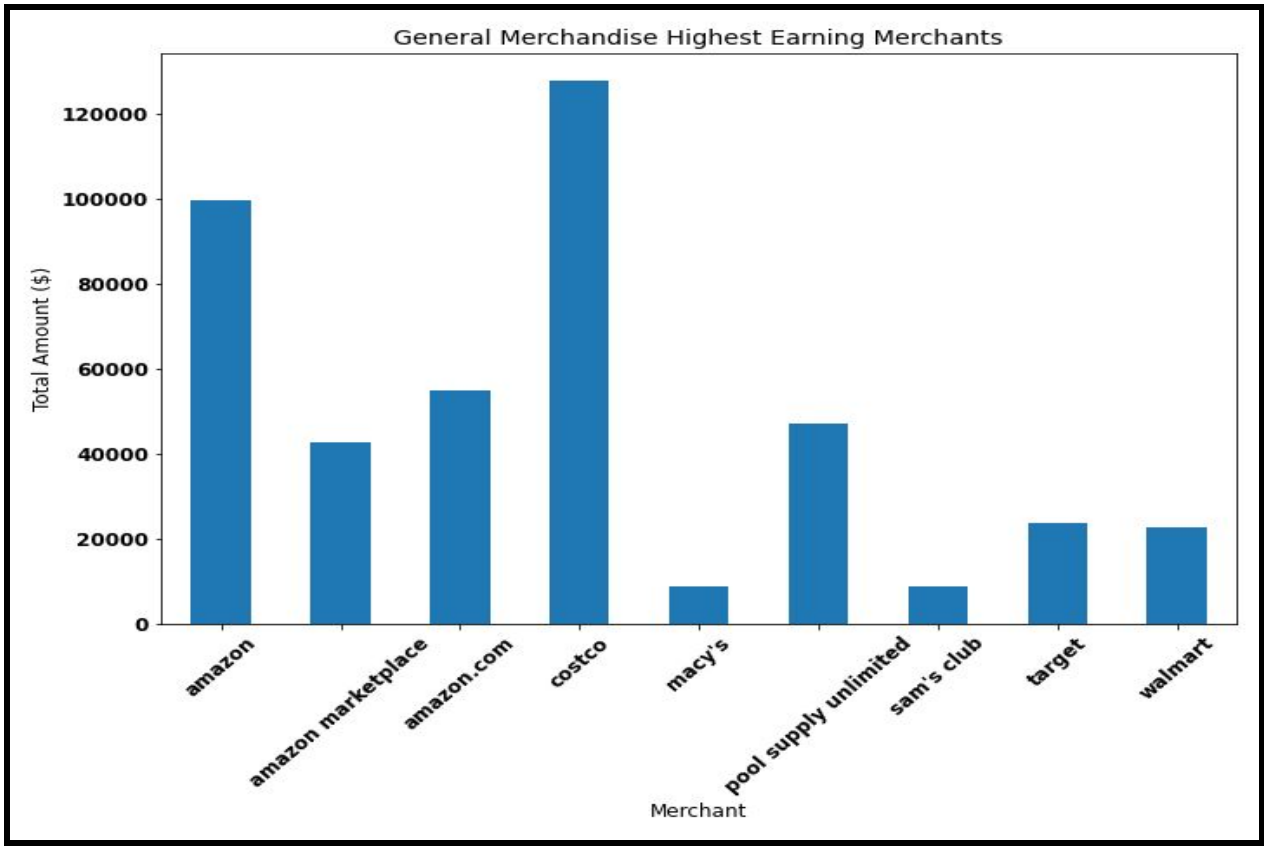

Figure 4: General Merchandise - Highest Earning Merchants

Users spent $128K at Costco which is the highest amount in the General Merchandise category. That's more than a quarter of the amount spent elsewhere. All the high earning merchants are well known names except for "pool supply unlimited".
Upon investigation, the transactions associated with this merchant were added to the data while FinGoal was testing their software. Multiple developers added the same transaction with different uids. Since it actually did occur, I kept on transaction per uid per amount and dropped others that were duplicates.

DateTime Analysis

1.Mean Amount Spent Per Month

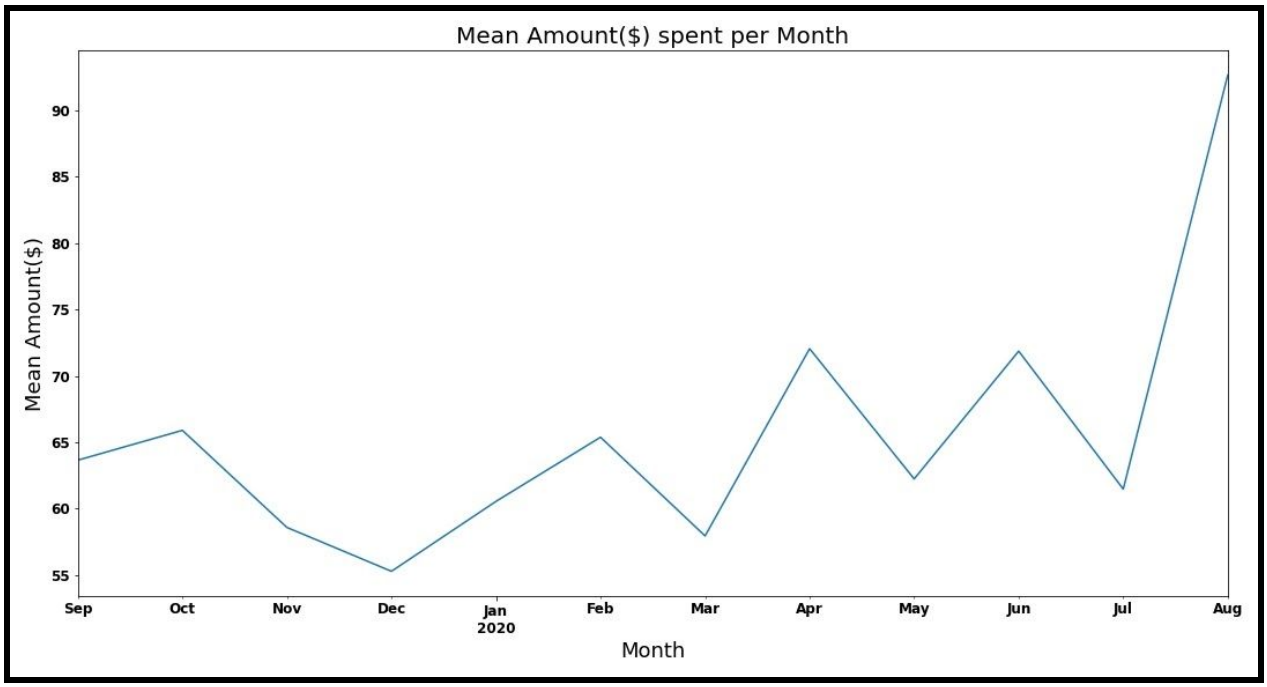

Figure 5: Mean Amount ($) Spent Per Month

The data was collected from Sep 2019 to August of 2020. Clearly there are spikes in the data where the average amount went up in a few months and sharply went down in other months. In April the mean amount went up to 73K which could be a result of all the schools and workplaces being shut down due to CoronaVirus. People were forced to spend money on office and school supplies that they normally didn't. The weirdest spike is in August 2020 where the mean amount went above $90 which is 38% jump from July mean amount.

2. Average Amount Per Month Per Category

In Figure 6 below, the average amount per month per category is shown for the year 2020. The data in the monthly graph below explain Figure 5 very well. For example before CoronaVirus hit, in February, people spent money on everything and when the shut down started to occur majority of spending was limited to "office supplies". The same trend continued in April. In May the effects of Corona continue. In June the spending increased only to be followed by a huge decrease in July. In August, even though the mean maximum amount was around $200 in category General Merchandise, the amount of transactions increased and the spending was up in all categories therefore the mean spiked compared to other months. Figure 7 shows August's graph in detail

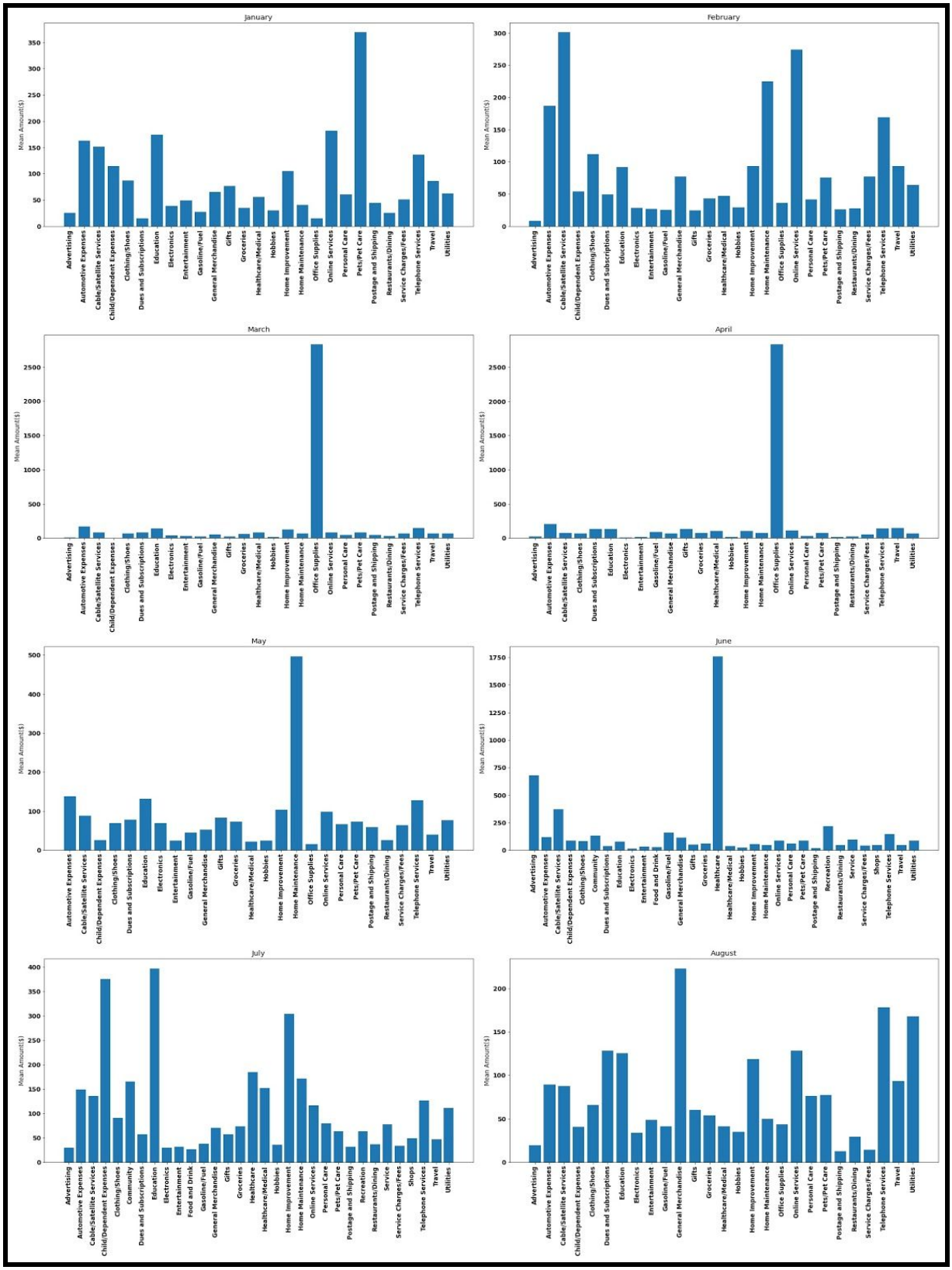

Figure 6: Mean Amount Per Category Per Month

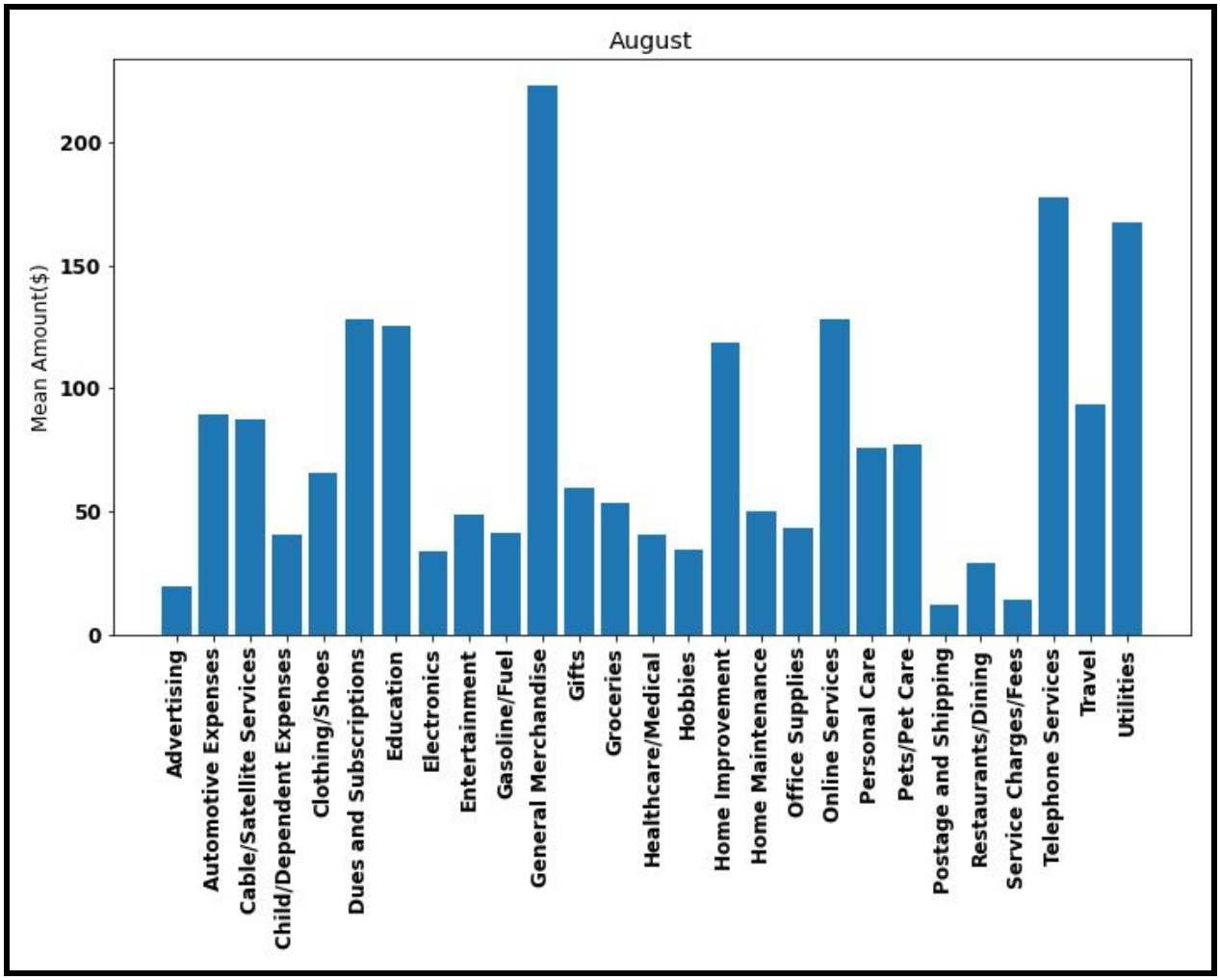

Figure 7: August 2020 - Mean Amount Per Category