BIBDA: Semester Project

Athens University of Economics and Business

Course: Business Intelligence and Big Data Analysis
Professor: Chatziantoniou Damianos
Academic year: Winter Semester 2023-2024

Students:
Alexios Mandelias (ID: 3190106)
Anna-Lefkothea Papavasileiou (ID: 6190101)

Warehouse

Introduction

The dataset comprises a bank's data for all the details for the debit and credit card transactions that were carried out by the bank, for the time period between 2015 and 2020.

In this report we aim to analyze the number and total sum of the transactions, across a number of parameters, namely the gender and age of the consumers, the brand and type of cards, the type of transactions, the city in which they took place, as well as any combination of the above features.

From this analysis we will produce statistical reports and extract useful information about electronic transactions carried out by cards.

Data preprocessing

This dataset was used in another university course. It was provided by the course professor.

The data resides in a single 500MB `.csv` file. We examine the structure of the `.csv` file, find the attributes of each data point and create a single `CardTransactions` table into which we bulk insert the data from that `.csv` file.

We look for any missing or otherwise malformed data, of which we find none, so we continue with importing the data into the warehouse.

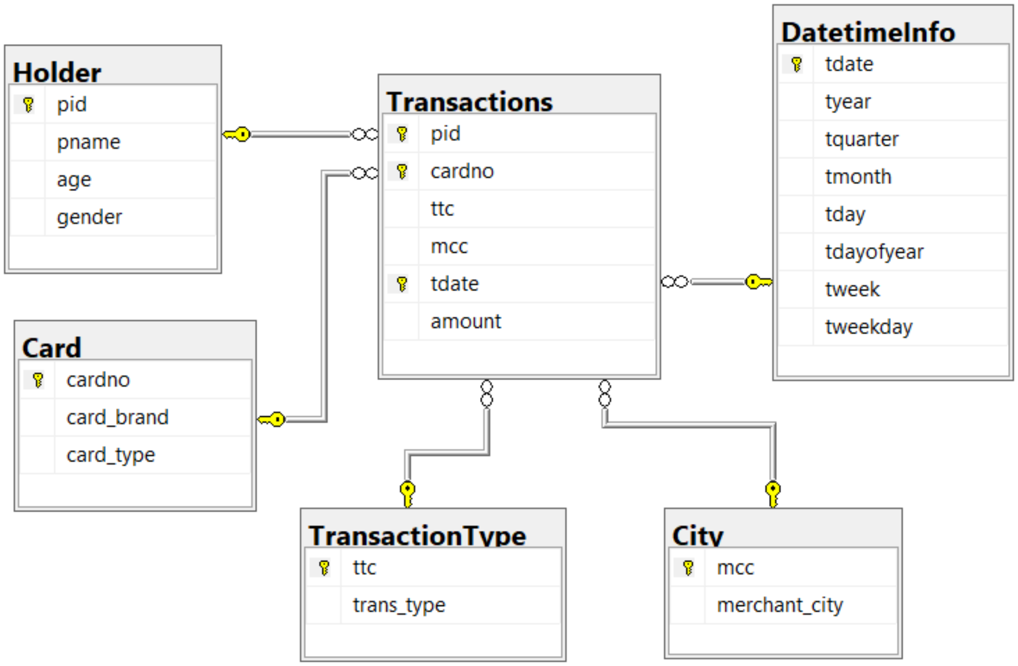
After taking a closer look at our data, we find that the transaction data can be split into five dimension tables and one fact table, as explained in the star schema section below. We split the datetime field into year, quarter, month, week, day, etc. to facilitate further exploration of the data and produce more interesting results.

We observe that data for the year 2020 are limited to the first two months of that year, therefore we will take care to exclude the year 2020 in any visualizations which rely on yearly averages or sums, since the incomplete data of 2020 will produce results which can be very easily misinterpreted.

We attempted to include state data[1], to allow us to perform state-wide analyses and reports, however we quickly noticed that one a city may exist in multiple states (city "Franklin" was present in 29 (!) states), therefore a one-to-one mapping from city to state was not possible to be performed starting from our dataset, which contained only city data.

Star Schema

We split our data in fact and dimension tables, thus creating the following star schema:



[1] https://simplemaps.com/data/us-cities

It contains the fact table "`Transactions`" with the single metric `amount`.

In addition, there are five dimension tables each containing the "who", "how", "where" and "when" of the transactions:

- `Holder` (card owner details)
- `Card` (card information)
- `TransactionType` (chip, online or swipe)
- `City` (where the transaction took place)
- `DatetimeInfo` (year, month, day, etc. of the transaction)

Data Cube

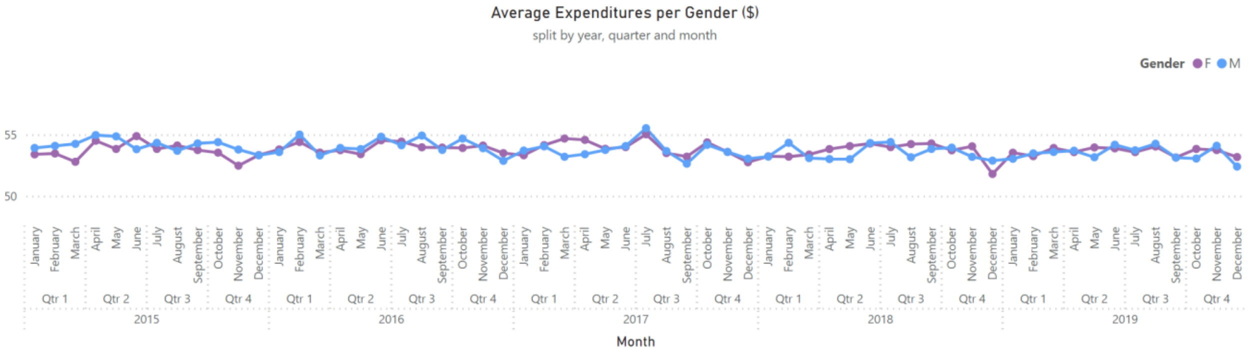
We are interested in examining spending differences across genders, age groups and time periods within a year. We create the following data cube to help us extract relevant information from our data across these dimensions:

```sql
SELECT gender, age, tmonth, SUM(amount) as transaction_amount
FROM Transactions
JOIN Holder ON Transactions.pid = Holder.pid
JOIN DatetimeInfo ON Transactions.tdate = DatetimeInfo.tdate
GROUP BY CUBE (gender, age, tmonth);
```

Visualizations

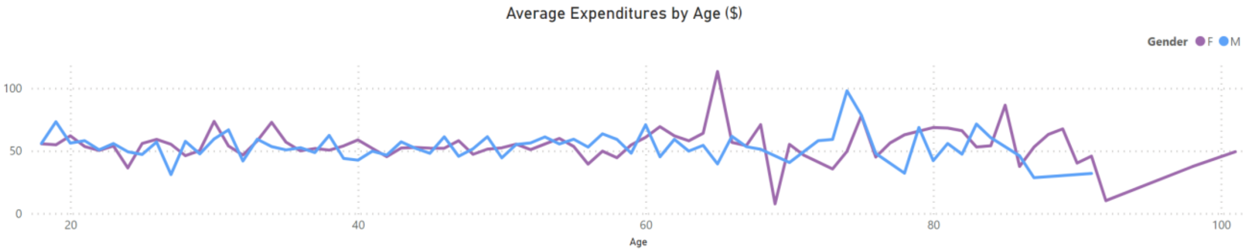
We visualize our data to facilitate our analysis. Power BI is used for all of them.

The first graph shows the average expenditures per gender, split by the year, quarter and month of the transaction:



Both genders appear to spend the same amount of money on average, as the two lines largely overlap.

When we add age to the mix, however, we can observe some interesting results:

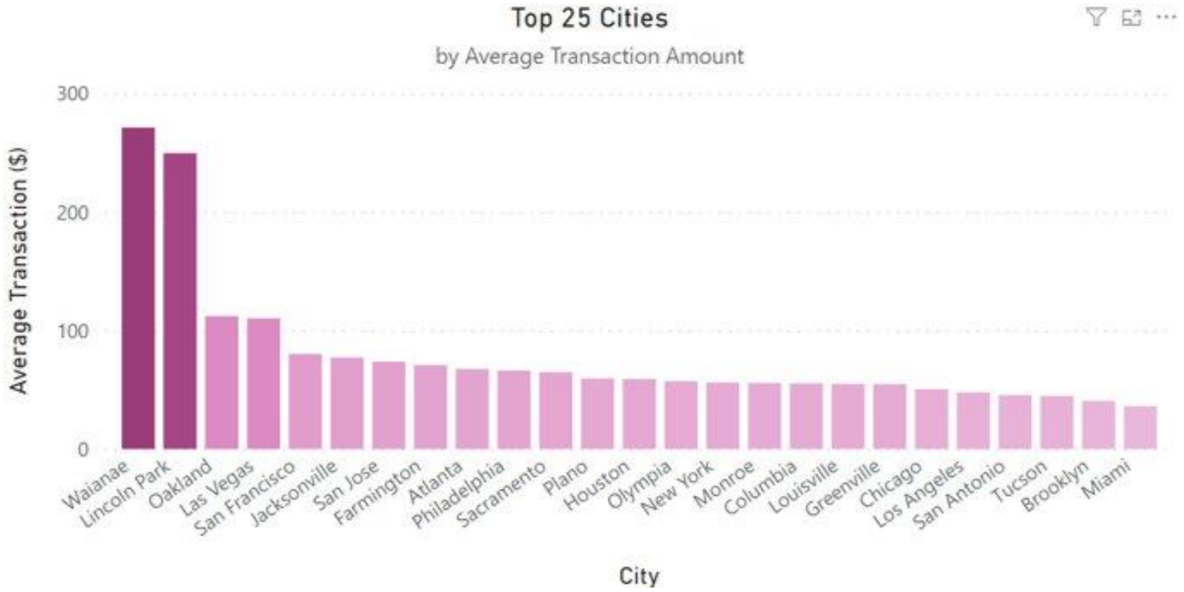


After the age of 60, women's expenses spike and drop significantly a couple of times. The largest spike occurs at age 65, and only a few years later, at age 69, there is a sudden drop.

Similar fluctuations apply to men after age 70, especially at age 74, where we see a sudden spike in the graph.

It is clear that both genders spend more as they reach retirement age, possibly due to the increased expenses of their medical care or the birth of their grandkids. To conclude more about the differences in the spikes and drops between the genders, we would need to conduct separate research.

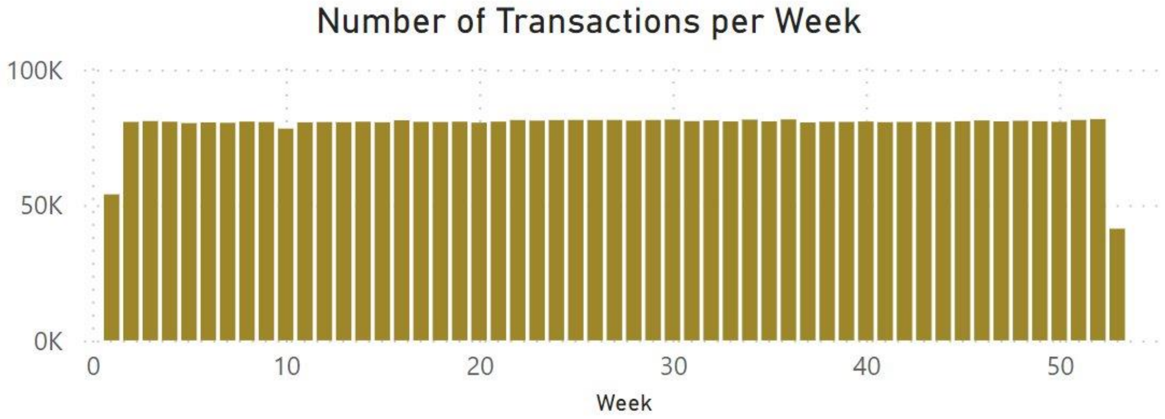
Leaving the gender aspect aside, let's look at the following graph to examine the expenditure distribution across the U.S.:



This graph shows the 25 U.S. cities with the highest average transaction amounts. The cities have been placed in decreasing order, based on the average expenditure amount.

Waianae and Lincoln Park top the graph, as the average transaction amount differs by a staggering 160 USD (Waianae) and 140 USD (Lincoln Park) from the third and fourth cities in the graph, which are Las Vegas and San Francisco. The cities that follow are mostly spread across different states, but there are quite a few Texan cities in the list.

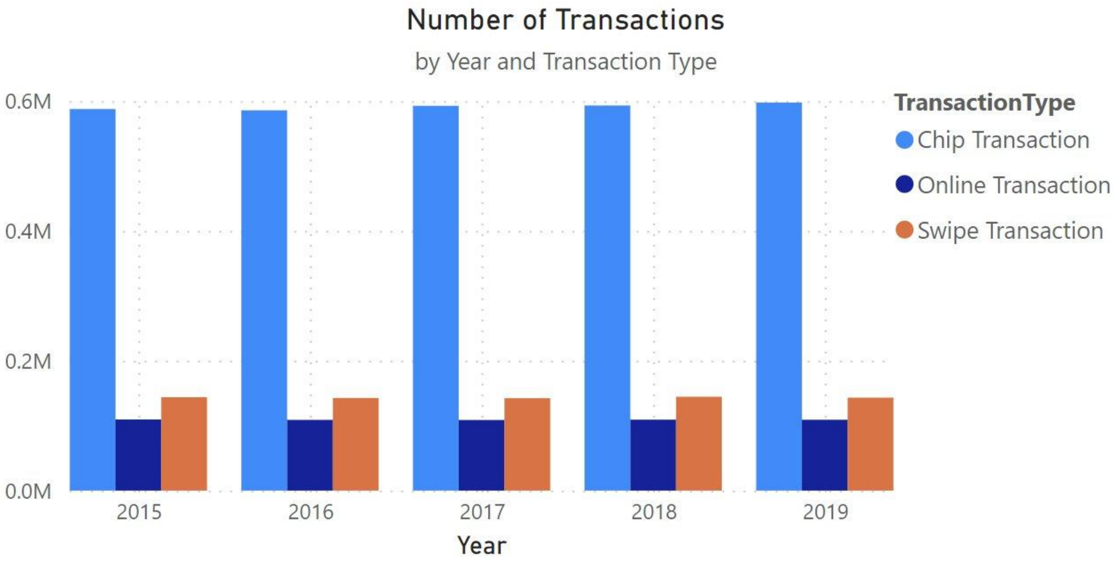
We are also interested in checking if there are specific times during the year, when people tend to spend more money.



We see that people tend to spend way less on the weeks before and after the celebration of the New Year, but otherwise, the expenditure amount remains remarkably consistent.

Lastly, we examine the details regarding the transaction type and the card type people prefer.

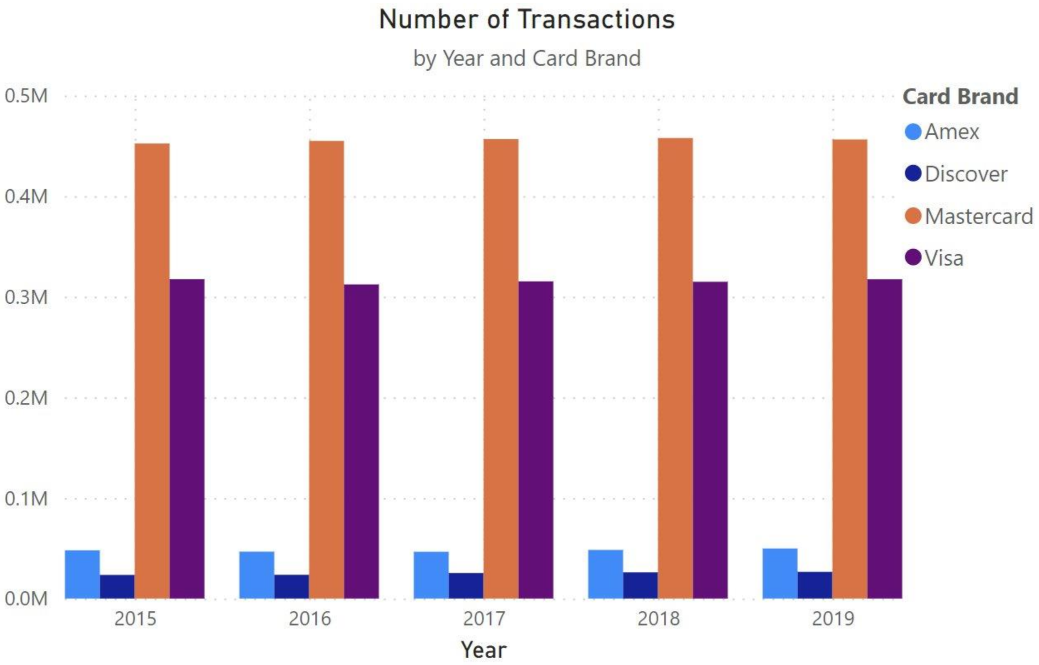
Throughout the years, chip transactions remain the most common transaction type:



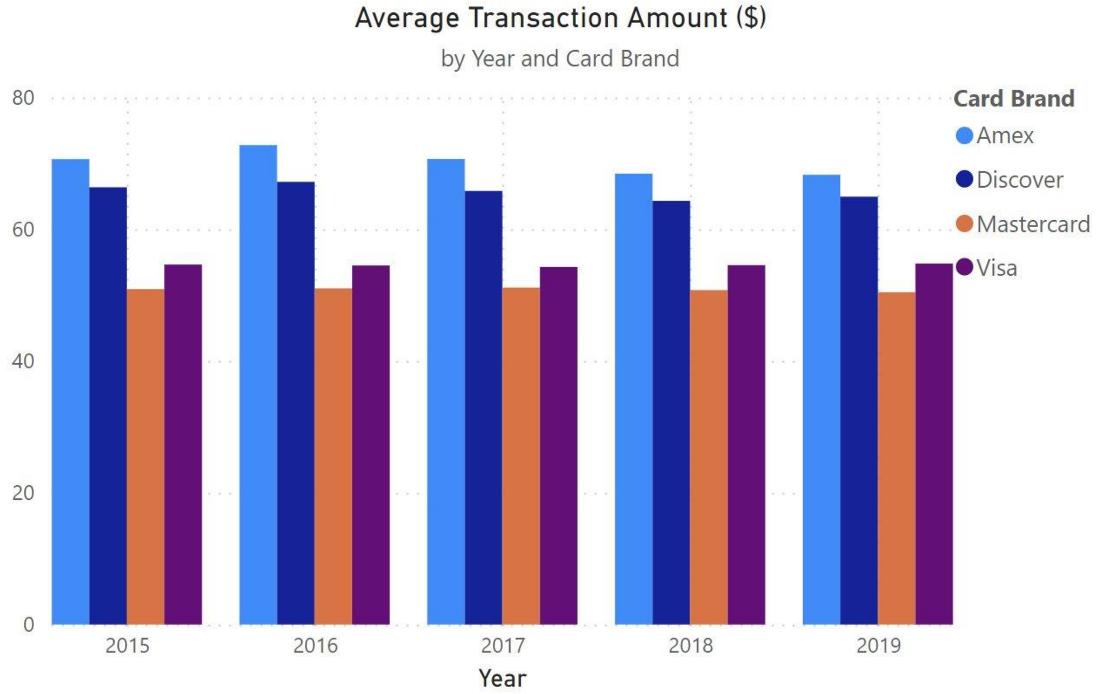
Chip cards have become the go-to payment method for maintaining maximum security while using credit or debit cards.

They are also known as EMV cards, named after Europay, Mastercard and Visa, the three companies that originally created the technical standard.

Therefore, it is not surprising that the most popular card types are Mastercard and Visa:



Despite this, expensive transactions are mostly made using Amex and Discover cards:



Data Mining

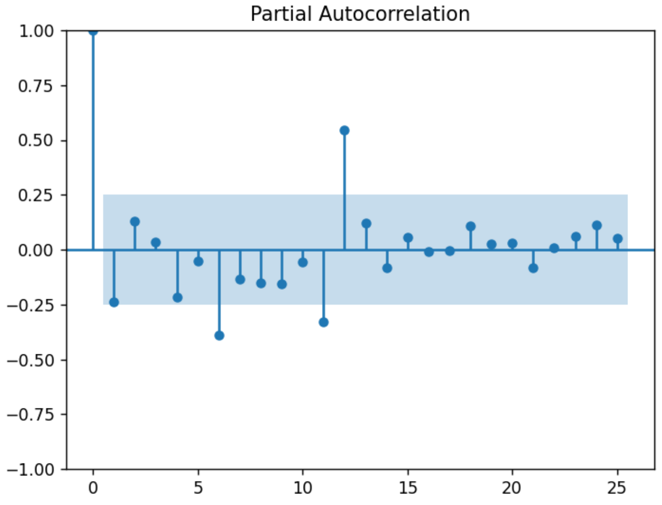
We create two models to extract useful information from our dataset.

Model 1: Predict sum of transactions using an Auto-Regressive model

We wish to use the transaction data from the years 2015-2019 to predict the total amount of transactions per month for the years 2020-2021. We hypothesize that the data (the total sum of transactions) for each month depends heavily on that of the previous months. Therefore, we will be using an auto-regressive model which, based on the data of the past months, will predict the data in the future.

We need to determine the importance of each past month in predicting the next one. We hypothesize that people's spending habits follow a period of 12 months, that is to say that the amount they spend each January is approximately the same, and different than what they spend each February. We further hypothesize that the amount they spend each month may depend on that of the previous month. We test these hypotheses using the PACF plot[2] to determine which past months are statistically significant in predicting future months.

The plot shows that month 12 (exactly one year ago) is a great predictor for the current month, which is to be expected. Furthermore, somewhat unintuitively, month 6 (half a year ago) and month 11 are also statistically significant predictors of the current month. Month 1 (the previous month) is surprisingly not statistically significant, meaning that the total sum of transactions for any month does not strongly correlate to that of the previous month. Months 13-24 (more than one year ago) are very much not statistically significant, which is to be expected since in a period of 12 months, months 1-12 contain most of the information.



[2] https://en.wikipedia.org/wiki/Partial_autocorrelation_function

To build the model we will use months 6, 11, 12 as lag values in the time series of the sum of transactions. The equation of an auto-regressive model for lag values {6, 11, 12} is as follows:

$$y_t = c + \varphi_6 \cdot y_{t-6} + \varphi_{11} \cdot y_{t-11} + \varphi_{12} \cdot y_{t-12} + \varepsilon_t$$

Where y_t the t-th value of the time series, c the constant term, φ_{t-i} the coefficient of the $(\mathrm{t} - \mathrm{l})$-th value of the time series (where l the lag value), and ε_t the error term for y_t.

The parameters of our fitted model are the following:

Parameter	Value	P-Value
constant	1,050,000	0.032
sum_amount.L6	-0.0843	0.180
sum_amount.L11	-0.0659	0.254
sum_amount.L12	0.8738	0.000

We see that the terms y_{t-6} and y_{t-11} have $p_{value} > 0.05$, and are thus not statistically significant at confidence level $a = 0.05$. We fit the model again with the updated formula, without the aforementioned terms:

$$y_t = c + \varphi_{12} \cdot y_{t-12} + \varepsilon_t$$

Parameter	Value	P-Value
constant	290,300	0.189
sum_amount.L12	0.9243	0.000

Again, the constant term has $p_{value} > 0.05$, and is thus not statistically significant at confidence level $a = 0.05$. We fit the model again with the updated formula, without the constant term:

$$y_t = \varphi_{12} \cdot y_{t-12} + \varepsilon_t$$

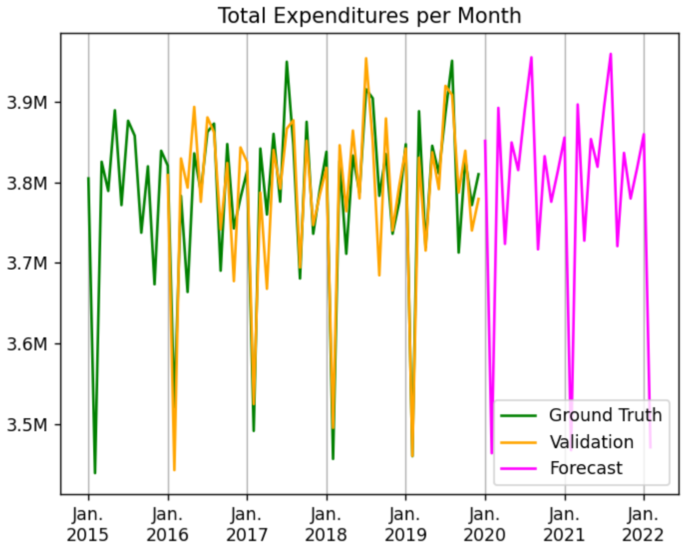
Parameter	Value	P-Value
sum_amount.L12	1.0011	0.000

The formula thus becomes:

$$y_t = 1.0011 \cdot y_{t-12} + \varepsilon_t$$

After discarding the statistically insignificant variables, we conclude that the total transactions for any given month depends solely on that of the same month of the previous year, with a very small upward trend of 0.1% each year.

We plot our data in green, the model's prediction for the train data in orange, and the model's prediction for the future data in magenta. We observe that the model's prediction for the train set very closely follows the ground truth, indicating that the model is working properly.



Furthermore, the model's predictions for the years 2020-2021 are, somewhat expectedly, very similar to the time series of the previous years, following the pattern established by our training data, and giving us further confidence that the model adequately explains the data.

Using this data, any interested party may adjust their policies based on the spending habits of people for each month.

While the model does fit the data very nicely, it is worth noting that we are only using a single variable to predict the response variable: the value of the month one year prior. This is an oversimplification; the economy and spending habits of the citizens of the United States are much more complex than this model would suggest. However, given the limited observations, the small number of explanatory variables, and the very good fit of this model, we have decided to keep this result.

Model 2: Cluster People by Age, total Transactions and City percentile

We wish to use transaction data from 2015-2020 to split people into different groups based on their spending characteristics. Specifically, we wonder whether people with different spending habits who live in areas with different city-wide average spending amounts per capita can be meaningfully distinguished. We hypothesize that people in low-income areas tend to spend similar amounts of money, and that age also plays an important role in that.

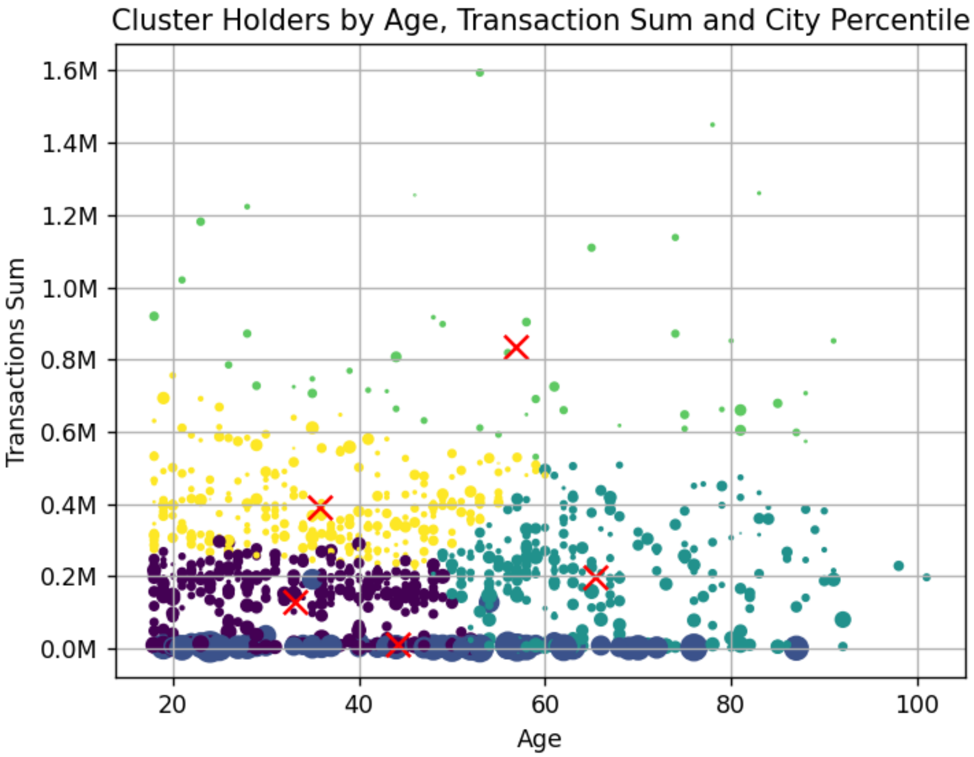
We perform clustering using the K-Means algorithm, splitting people into five clusters.

We observe that people can be meaningfully split in five clusters, each with the following characteristics:

Cluster ID	Age	Transaction Sum	City Class	Support	Support %
3	All	Very Large	High	47	5%
4	Young	Large	Mid-High	257	26%
0	Young	Medium	Mid	342	53%
2	Old	Medium	Mid	261	27%
1	All	Very Small	Low	77	8%

The results of the table are summarized below:

- Cluster ID 3: People with a very large transaction sum, living in very high-class cities.
- Cluster ID 4: Young people with a large transaction sum, living in high-class cities.
- Cluster ID 0: Young people with a medium transaction sum, living in mid-class cities.
- Cluster ID 2: Old people with a medium transaction sum, living in mid -class cities.
- Cluster ID 1: People with a very low transaction sum, living in low-class cities.


Cluster Holders by Age, Transaction Sum and City Percentile
Transactions Sum
1.6M
1.4M
1.2M
1.0M
0.8M
0.6M
0.4M
0.2M
0.0M
Age