**Data Science CA2 – Daroogheh**

**Overview**

In this project, we implement a real time data pipeline to process and store data as it gets to the server. We can divide this project into 4 phases. First, we need to get the historical data and store them properly. Then we need to perform some basic analysis on them to gain some general insight about our domain. Next, we need to get the real time data and perform various analysis on them including fraud detection. Lastly, we need to update our database with the new data and then analyze them.

For the first part, we will use spark batch sessions to get the data from the synthetic transaction generator. We then use MongoDB as our database and store the historical data. In the second part we use the spark batch sessions to perform various analysis on the data. In the third section, we use spark stream sessions to get the real time data from the synthetic transaction generator and mark the fraudulent transactions. We then update our database with the incoming data. In all of the stages, we use Kafka in order to mimic a real word scenario in which data comes to our server from different sources and we have to process it.

**Store the Historical Data**

In this section we Store the historical data. The historical data is generated by a synthetic transaction generator. It consists of 20,000 different transactions. We have to store these transactions in our database. For this project, we use MongoDB which is a NoSQL database which means the underlying structure of the database is not tables but collections instead.

We designed a database with 6 different collections. The first collection is transaction summary which stores all of the information about the transactions. Each datapoint here represents one transaction with the transaction id being the key. The second collection is the daily collection where we group the transactions by the date. This table could be used to do temporal analysis and find trends in the data. The third one is the merchant data. This table stores the information about each merchant with the merchant id being the key. We have a similar collection for the customers as well. we store basic data like the transaction count, transaction value, total commission value, successful transactions and failed ones and … in this collection. The last two collections store the information of the customer types and merchant categories. These two collections come really handy when performing future analysis.

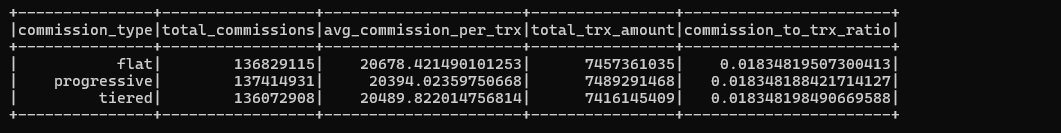
To store the data in our database, we need to perform various aggregations so we use the spark batch sessions to make the processing faster. We first read the data from Kafka and save them in a pandas dataframe. Then we open the dataframe and convert it into a spark dataframe. After that, we start aggregating the data and store it into our collections. After this step our database is filled with the historical data.

**Historical Data Analysis**

Now that we have our historical data saved into the database, we can move on with analyzing this data. After some trials we understood that loading the data again from the database could be time-consuming and unnecessary as we already have the data we want as a spark dataframe ready to be analyzed. So, this part is somehow integrated with the previous part. Right after saving the data, in the same session, we start analyzing the data. We did many different and interesting analysis. However, as the synthetic transaction generator is not really sophisticated and generates the transaction information from a uniform random distribution the reports currently are nor really meaningful. We hope as we replace the synthetic generator with the real-world data our reports will be more meaningful, pinpointing interesting facts about the business.

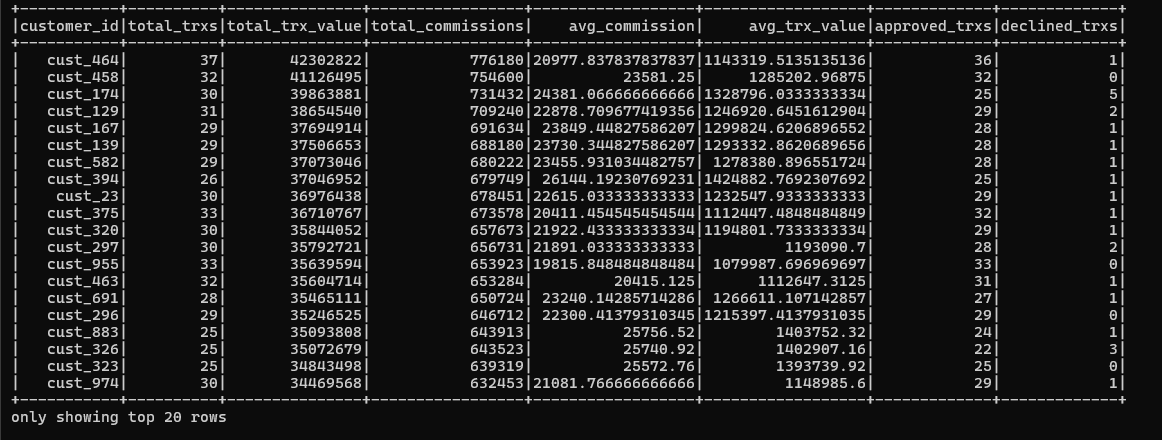
1. **Commission Type**

The first thing we analyzed was the commission strategies as they play a key role in our business. Finding the optimal strategy can significantly boost our income. We grouped the transactions by the commission type and then calculated the commission to transaction value ration. The higher this ratio the higher our income will be for the same amount of transaction value. We can see the output below. However, as stated before, this report is currently meaning less as the ratio for all different strategies are roughly the same.



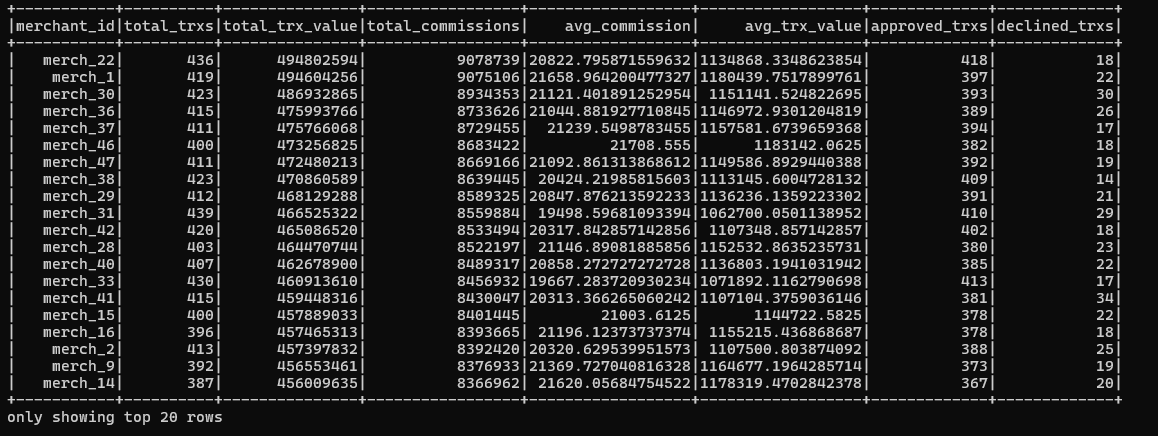
1. **Customer Summary**

The next step is to identify the key customers. Knowing our customers’ behavior enables us to make better business decisions. Therefore, knowing our top customers play a huge role in moving our business forward. Below we can see the basic information about each of our customers.



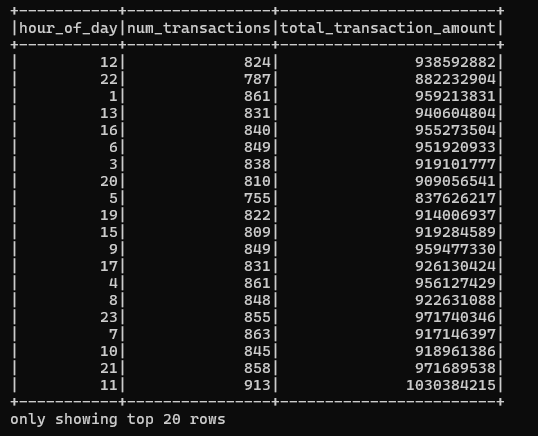
1. Merchant Summary

Just like the consumers, knowing the merchants is very important to us as we can tailor our business plan to best fit their needs. Below we can see the basic information about each of the merchants.



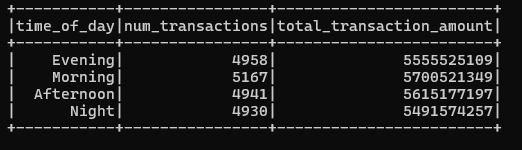
1. Hourly Summary

Now, we can focus on some temporal analysis. First, we can check when our servers are the most crowded and when the load is lighter. This will enable us to better plan our militance phases when some sections of the server might not be available and therefore lose less money.



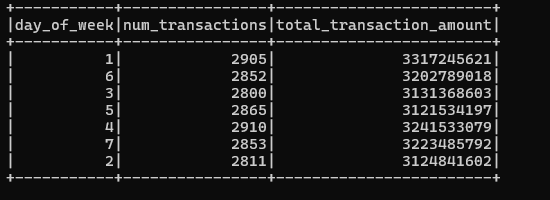
1. Time of the Day Summary

Another table which could be useful for less formal meetings and reports is the table in which we show the number of transactions in each time of the day. For example, knowing that there are more transactions near noon compared to early morning could be insightful.



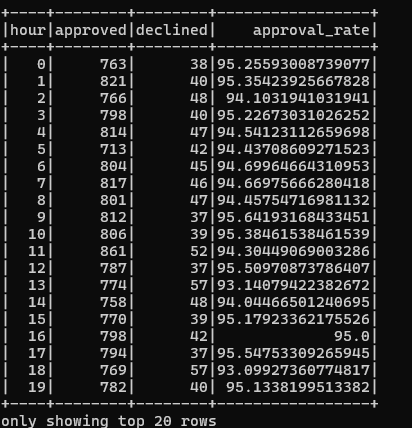
1. Day of the Week Summary

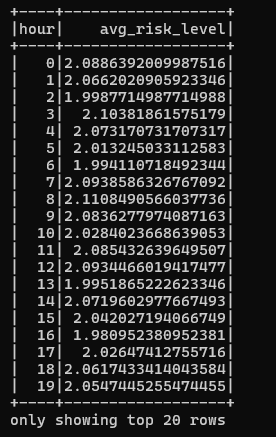
The final temporal analysis on transaction values which could further help us determine the best maintenance time is to understand which days of the week the value of transactions is lower.



1. Risk assessment Summary

The final report that could benefit us is related to risk management. The better we could identify transactions with high risk and high probability of being declined the better we can mange them. Understand when these types of transactions are more likely to happen will let us prepare for them.



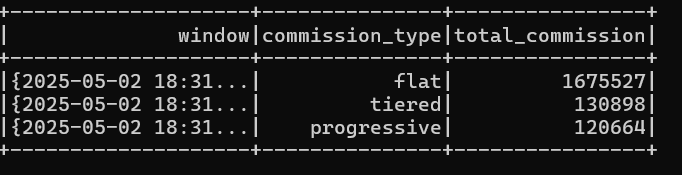


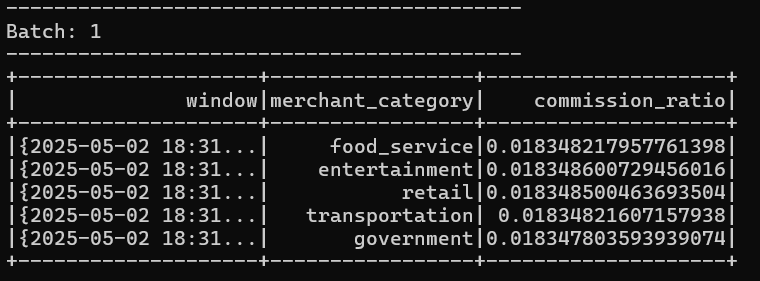
**Real-Time Data Processing and Fraud Detection**

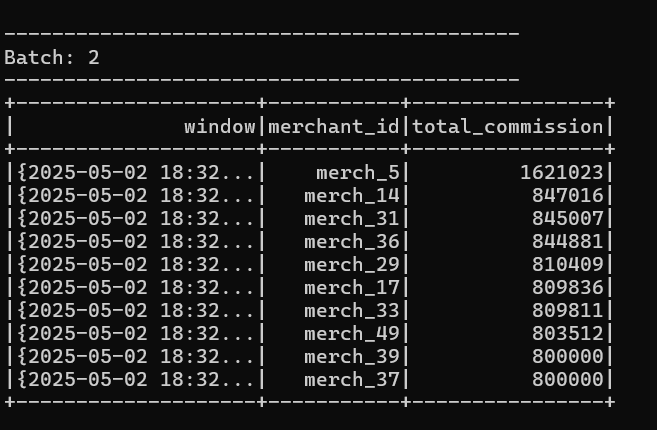
Now that we studied and store the historical data, we can proceed to get the real-time data and store it. We use spark stream sessions for this task which enables us to create minibatches from our data and process each batch on a regular basis. This is essential for processing real-time data effectively. After getting the data, we first summarize it and print some basic reports from it. Then filter the fraudulent transactions and output them to a separate Kafka topic. Lastly, we update the database with the new data.

1. **Basic Reports**

In this section we create very simple reports just to get a sense of the incoming data. Below we can see the results of some of these reports.





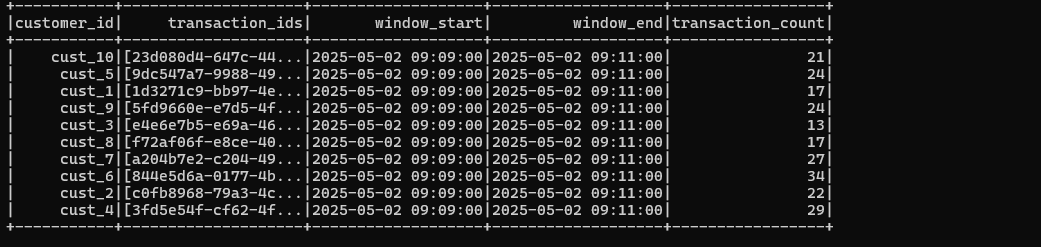


1. **Filter Fraudulent transactions**

The most important step in the real-time data analysis is finding and filtering fraudulent transactions. There are different methods to filter such transactions. One could even use machine learning methods to detect them. In this project, we are going to follow a rule-based method. We have three rules that classify one transaction as fraudulent.

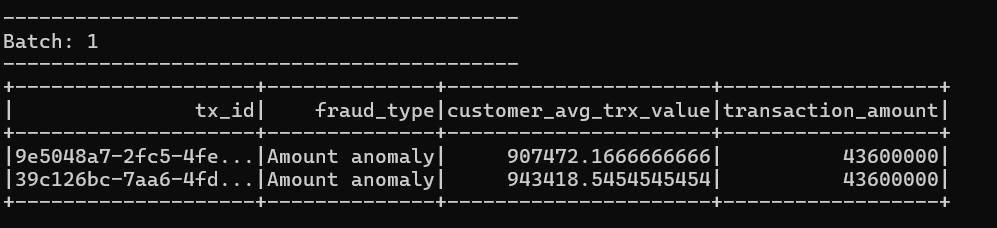
1. **Velocity Check**

If a customer makes more than 5 transactions per 2 minutes, we classify them as potentially fraudulent. To check this, we use a 2-minute window with a 2-minute watermark as well. If we detect such a case, we print the customer id and all the transaction ids that contributed to the velocity error. With the original synthetic generator, no velocity error occurred so just to check the program we limited the number of costumers to 50 while keeping the generation rate the same. Below is a sample output of this report.



1. **Amount Anomaly**

Another rule we have for determining fraudulent activity is comparing the transaction amount with the average transaction of the customer. To get this information, we access our database and join the customer summary topic with the Realtime data. This way, we have access to the average transaction value of the customer initializing each transaction. Again, in this section the original synthetic generator did not produce any errors so we tweaked it manually to test the program. Below we can see a sample report.



1. Geographical impossibility

The last rule for filtering fraudulent transactions is the geographical impossibility. We know that a human cannot move more than 50Km in 5 minutes therefore getting two transactions in 5 minutes from the same customer with over 50Km distance is fishy. We determine the distance using the longitude and the latitude. In this part the synthetic generator didn’t produce any errors and we could not change it in a way that it does so.

1. **Update the Database**

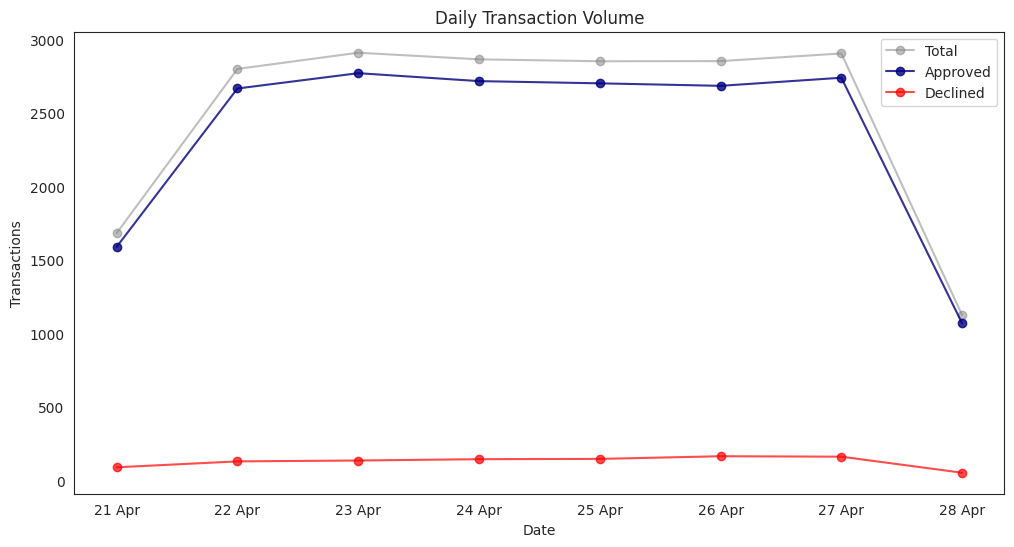
In this section we add all of the new transactions to the database. As we had different collections in our dataset, we should first aggerate the new data and then merge it with the current data. This process is not very straight forward as updating some values of the collections can be challenging. For example, in order to update the average transaction value for each customer we cannot simply take the average of the current and new value but instead we should first calculate the new total transaction value and the total transaction count and then divide them to get the final result. In order to make sure everything is working as expected we created a simple test program for the database in which we print the last data for each collection. By comparing the results of different executions, we can make sure that the database is updated correctly

**Final Data Analysis and Visualizations**

Now that we have our pipeline and the updated database ready, we can proceed to analyze the data and show our collected insight through visualizations. In this section, we first get the data we want from the updated database and then create various visualizations. The goal of these plots is to help us understand our business better. We focus on three aspects of our business. First, we focus on our customers and how we categorize them. Then we focus on our merchants. And lastly, we focus on abnormal and fraudulent behaviors.

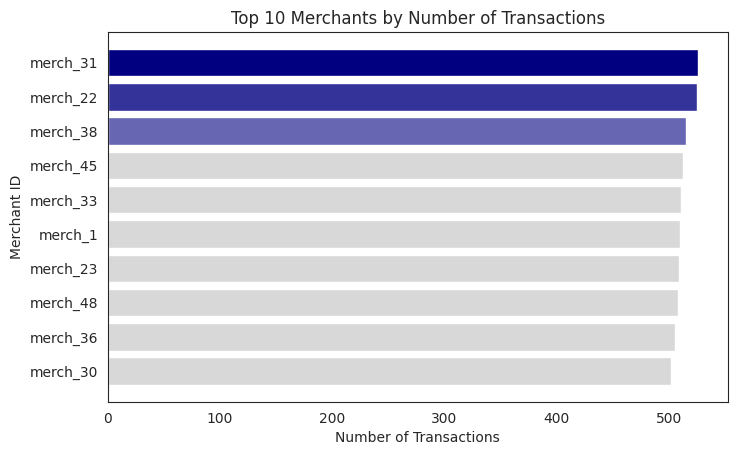
1. **Daily Volume**

We start our analysis with a big picture of the data. We plot the trend of total, approved and decline number of transactions. We can see that in this example data which was generated by the synthetic generator the transaction numbers remain roughly the same for each day. The only datapoint which is different is April 28th which is the current day and as it has not finished yet the number transactions so far is lower than the previous days.

****

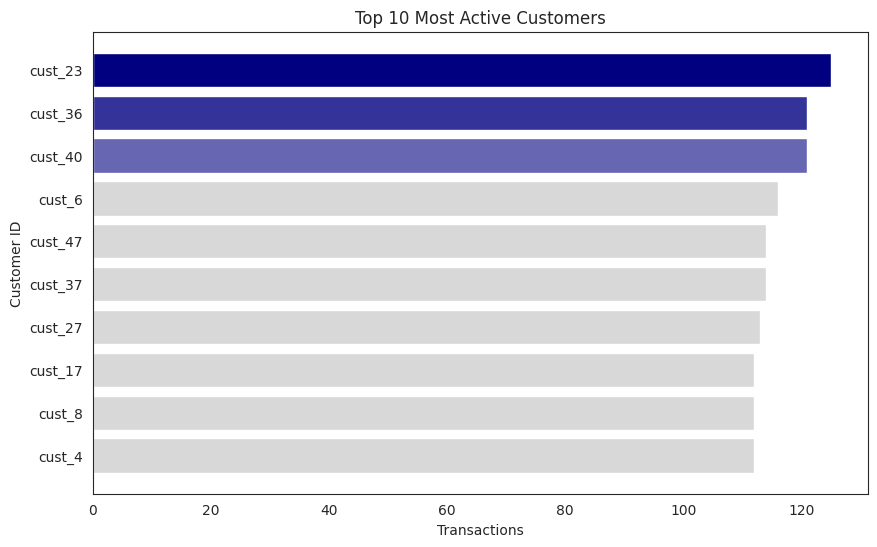
1. **Top Merchants**

Now, we shift our focus to the merchants. We want to find the top merchants by the number of transactions. These are the merchants who are really important to our business. Currently the difference between the merchants is negligible. However, we believe by replacing the synthetic data with real-word data this plot will be much more insightful.

****

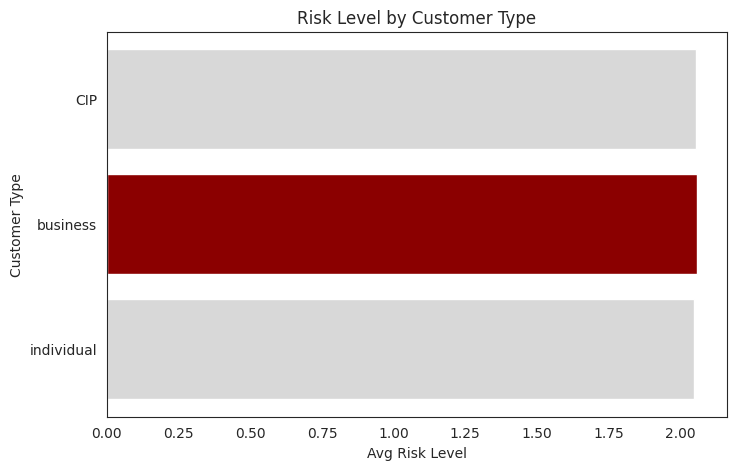
1. **Top Customers**

Just like what we did with the merchants, in this plot we rank our customers by the number of transactions. Again, we cannot really infer anything meaningful from this plot because of the synthetic generator. But in general, by identifying our most active customers, we can analyze their specific behavior and figure out why they use our business and how we can make the most profit from them.

****

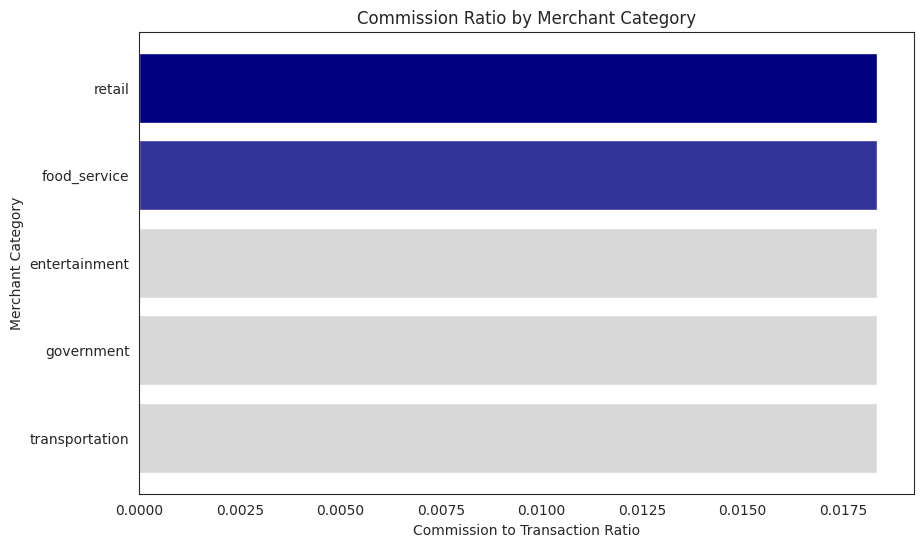
1. **Most Risky Customer Category**

Identifying our most risky costumers is for sure as important as finding our most active customers. These customers can be very harmful for our business. If we do not manage their behavior accordingly, they can leave us with huge losses. In this plot we plot the average risk level of each of our costumer categories. Currently we can see that all of them have the same average risk. However, in real world this plot would give us insight on which category is the riskiest one helping us to better plan to minimize our risk.

****

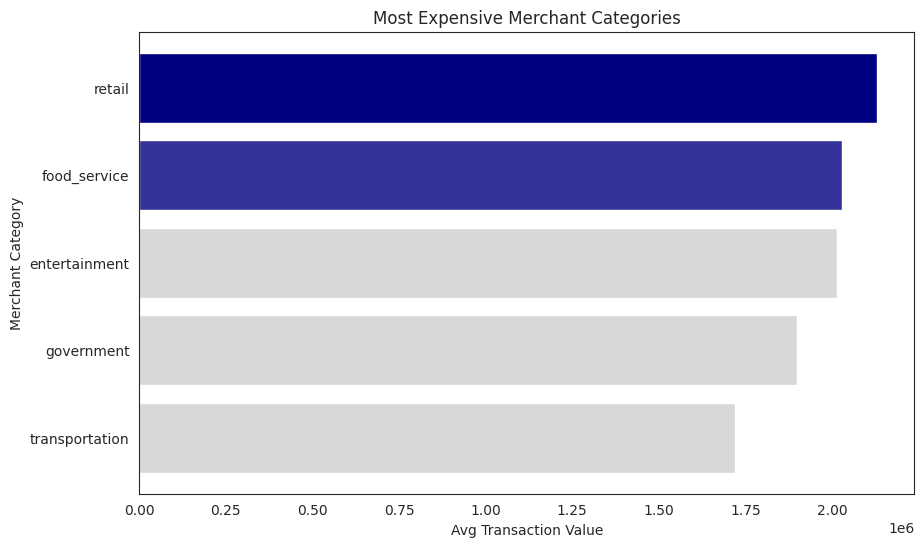
1. **Merchants with Highest Commission to Transaction Ration**

If we wanted to advertise what would be the best merchant category to focus for? This plot is answering this question. These merchants have the highest commission to transaction ration. So, if we increased the transaction volume of these category, they will have the highest increase in commissions. So, our goal with these merchants is to increase their transaction volume.

****

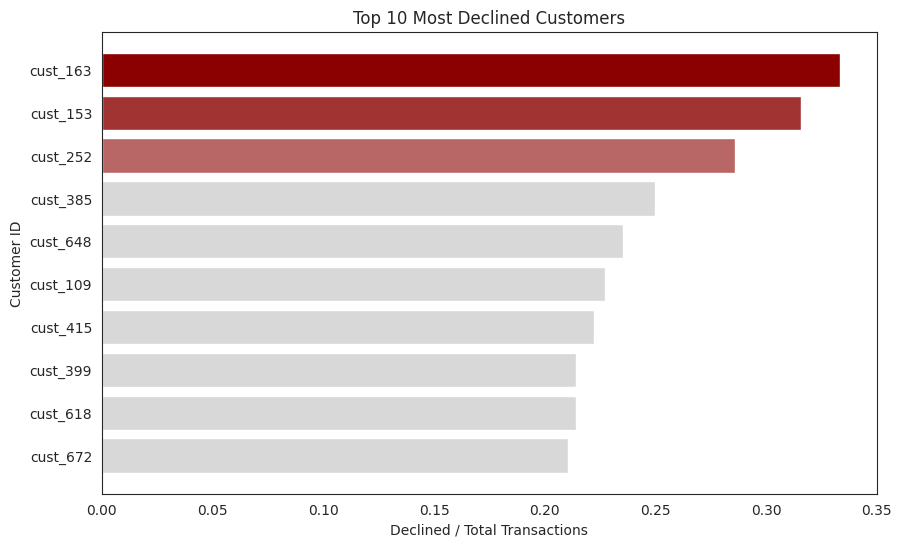
1. **Most Expensive Merchant Category**

In this section we have our merchants ranked by the average transaction value. Merchants with the highest average transaction value are the ones we should really tailor our commission strategy for. Getting the right strategy for calculating the commissions of these merchants can have significant impact on the profitability of our business as the transaction value itself is significant.

****

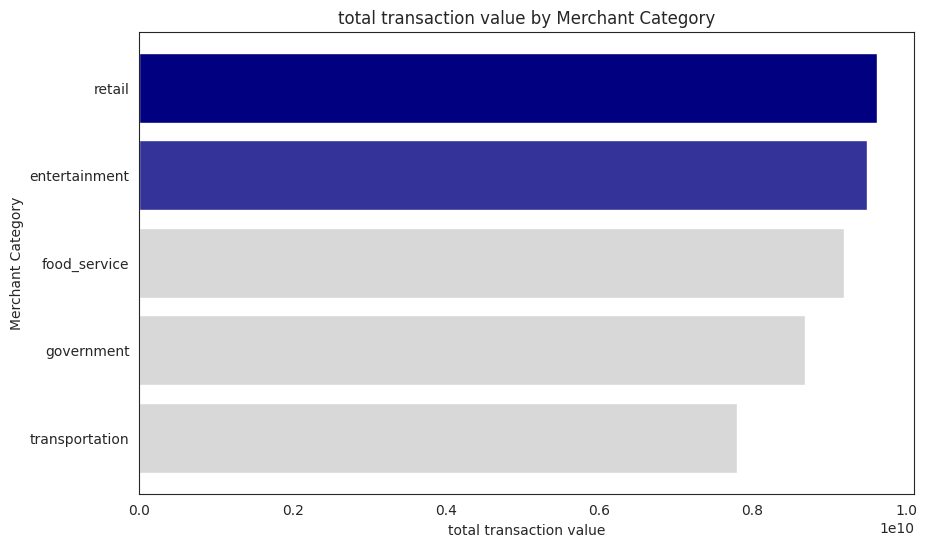
1. **Customers with Highest Declined transactions rate**

**Another method to detect risky costumers is by soring them by the frequency of their declined transactions. A high ration of declined transactions can be due to many factors. We should analyze the details to see if something is wrong with the customer. This high decline frequency could be a consequence of fraudulent activity. Therefore, we need to further analyze each of these customers. So, this plot helps us find potential risky customers.**

****

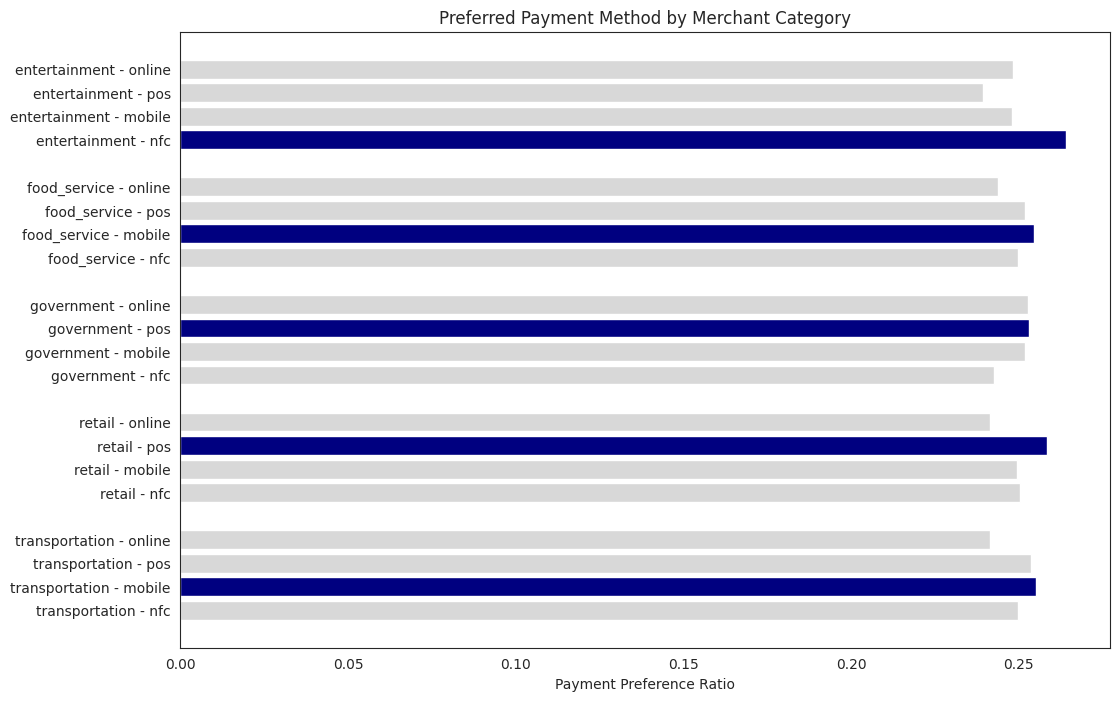
1. **Merchants with Highest Transaction Value**

This plot combines two of the plots we had giving us a more general view of the state of our merchant categories. Here we are not talking about the number of transactions or the average value of each transaction. But instead, we examine the total transaction value. Just like the total number of transactions and the average transaction value, this plot helps us identify our most important merchant category.

****

1. **Merchant Categories’ Preferred Payment Methods**

We previously saw the merchant categories with the highest commission to transaction volume and we discussed why we should focus on them and increase their transaction volume. In this section, we can see their preferences when it comes to the payment method. We can use the insight from this plot to facilitate the use of each method for the main customers of each merchant category. This may lead to an increase in the total transaction volume and total commission paid.

****

**AmirArsalan Shahbazi 81010151**

**Farnoush Fallah 810101484**

**Farjad Fallah 810101483**