Adam Veraszto, Benedek Kornyei,  
Zoltan Bereczki

JunctionXBudapest Hackathon  
Antavo Challange

Project Documentation

10/23/2022  
Budapest

# Introduction:

Our project was to analyse and visualise a big dataset provided by Antavo based on real data. According to this data we had to make some assumptions, predictions and improvements about customer life cycle, customer loyalty and customer experience. We made some pre-processing on the received data and made multiple observations on it. We were working with the checkouts and wanted to determine the goodness of coupons. We examined many aspects of the dataset and visualized the results of these. We also wanted to see if there is a trend in the amount of checkouts and if we can predict it. In the following chapters we will discuss the used methods and the final results with our thoughts on them.

# Dataset pre-processing

At first, we reduced the size of the dataset to be comfortable to work with it. We removed some features that we found unnecessary for our goal. The data had more than half a million customers with nearly 8 million records of events. Our idea was to use smaller group of customers therefore we kept only data of 100 000 customers. We found distinct currencies so we exchanged them to be comparable. We made a new dataset which was covering the checkout events ordered by the customers.

From the remaining values we created a table that contains the data of each remaining user. We then summarized the events for the customers, the records of the table contain:

* the customer ID
* the number of checkouts
* the number of coupon usages
* the customer’s first checkout
* the customer’s last checkout
* the amount spent by the customer
* the amount spent using coupons by the customer
* the average amount of money spent per checkout
* the average amount of money spent per checkout using coupons
* average number of days between checkouts

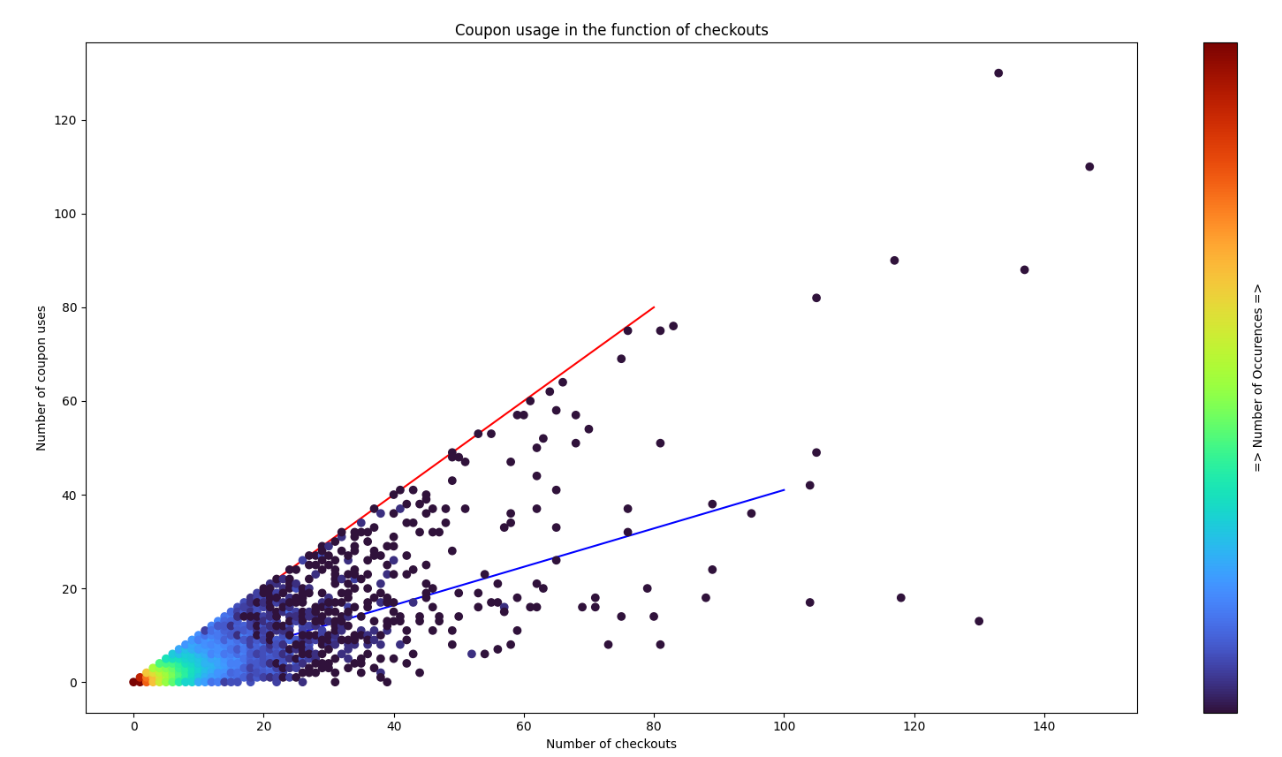
We also created some tables that contain:

* the number of checkouts
* the number of coupon usages
* the amount spent by the customer
* the amount spent using coupons by the customer

These tables contain the above-mentioned features in daily and monthly breakdown.

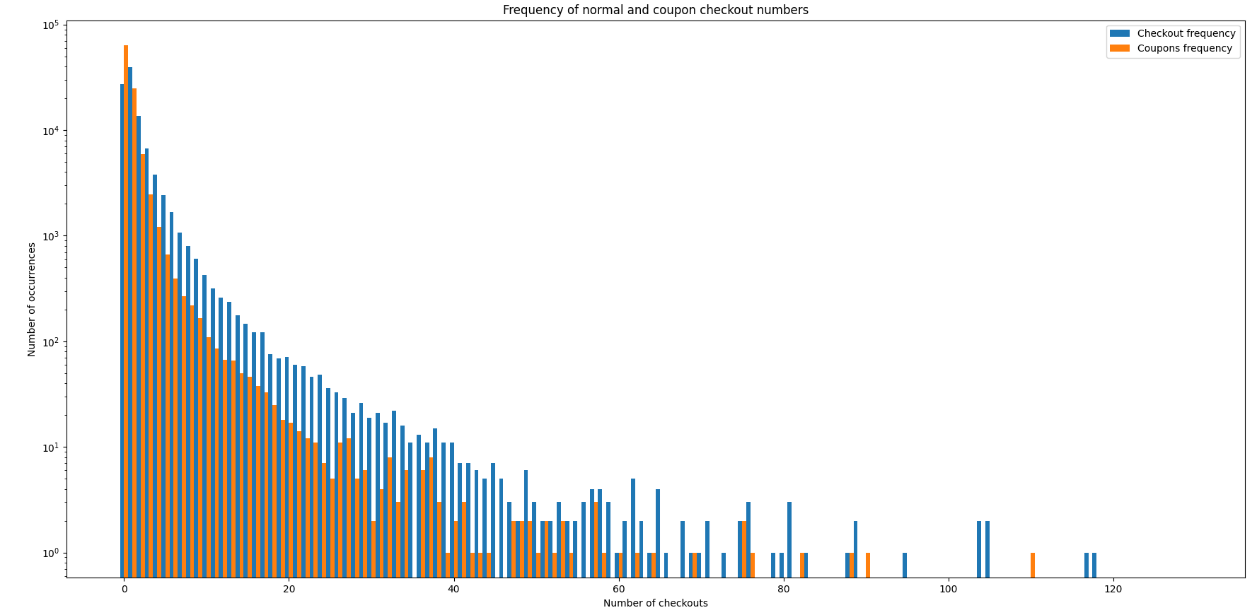
# Customer statistics and results

First, we wanted to map the amount of coupon usages by the amount of checkouts, where the colour shows the frequency of the occurrence. Our hypothesis was that there are many customers, that use coupons at almost every purchase even after a large amount of checkouts. The resulted to be correct as it can be on the 1. Figure. Many data points are close to the 45 degrees line. It also can be seen that most of the purchases can be found above the 22.5 degrees line, which means that more than half of the purchases are made using coupons.



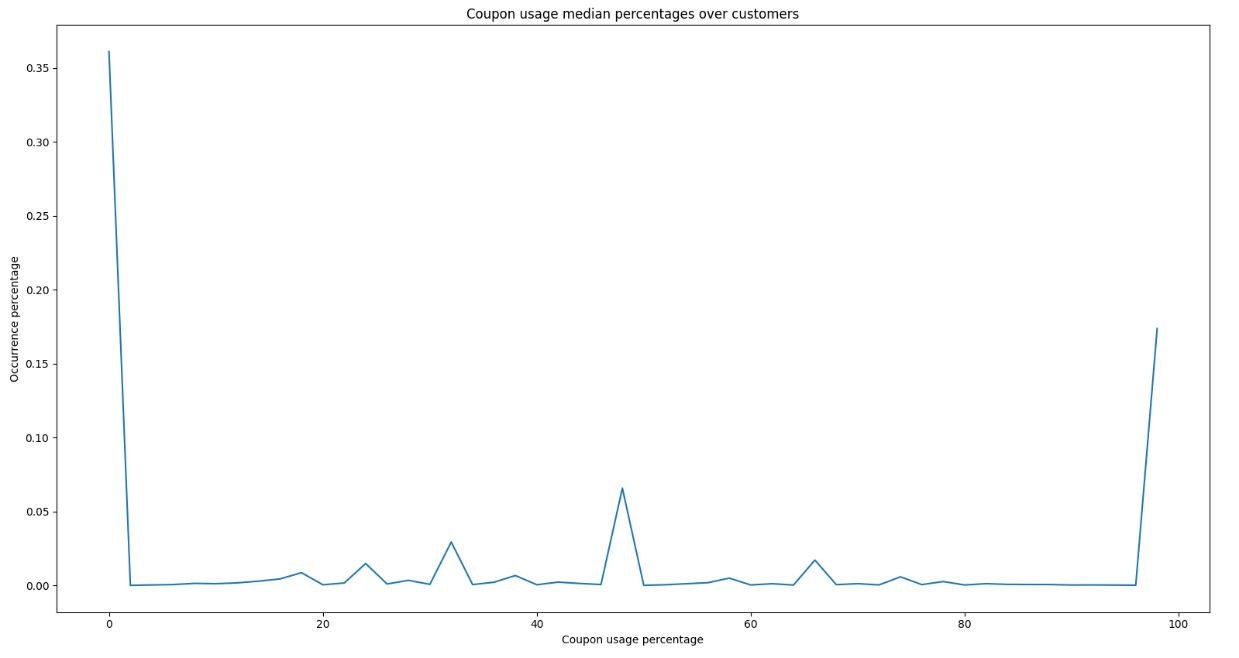
1. Figure Coupon usage in the function of checkouts

The 2. Figure shows the number of customers according to their number of checkouts. It shows that the coupon-system makes a great job getting new customers into the pool, but we can also see that it fails at converting them into non-coupon using customers.

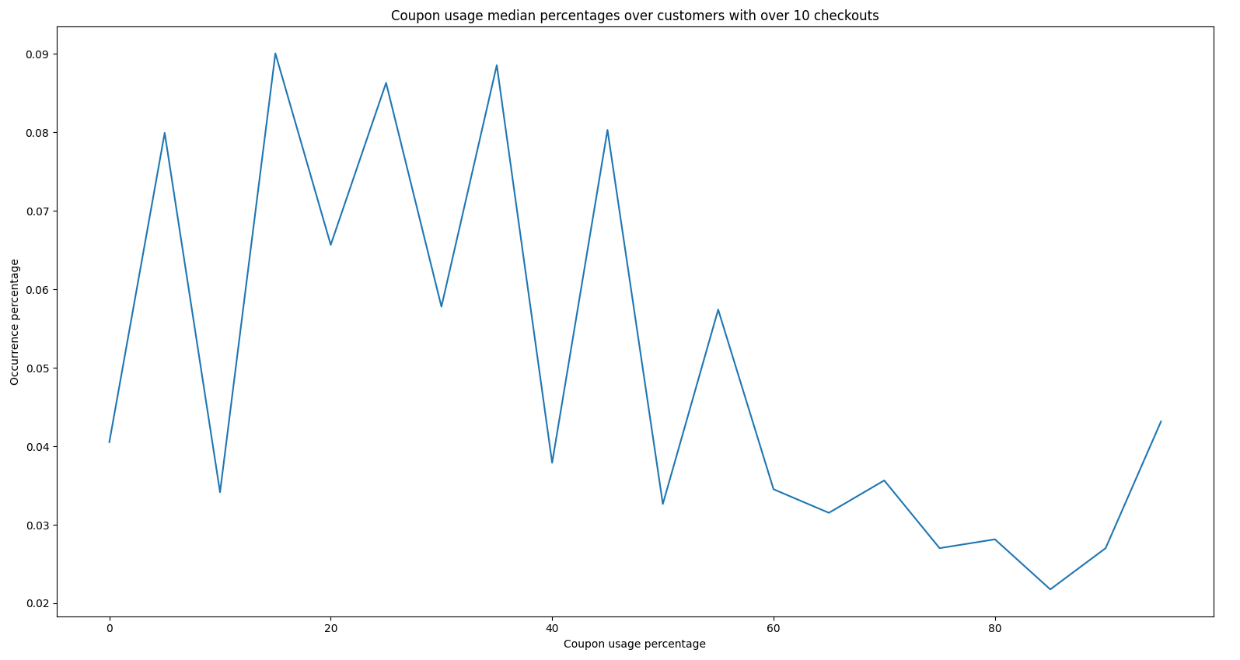


2. Figure Frequency of normal and coupon checkout numbers

The 3. Figure shows the percentage of customers over the percentage of coupon usage of the customer in checkouts. We wanted to predict if the coupon-system is also used after a bigger number of checkouts. We kind of failed, because the data turned out to be really biased towards customers with low number of purchases. That is why small fractions are appearing on the diagram. Therefore, we cut the data for customers with over 10 checkouts, which you can see on the 4. Figure. We realized, that now the data is not biased anymore and it is clearly noticeable that long lasting customer are tend to use less coupons. Involving this part of the pool might increase overall customer satisfaction and lifetime.

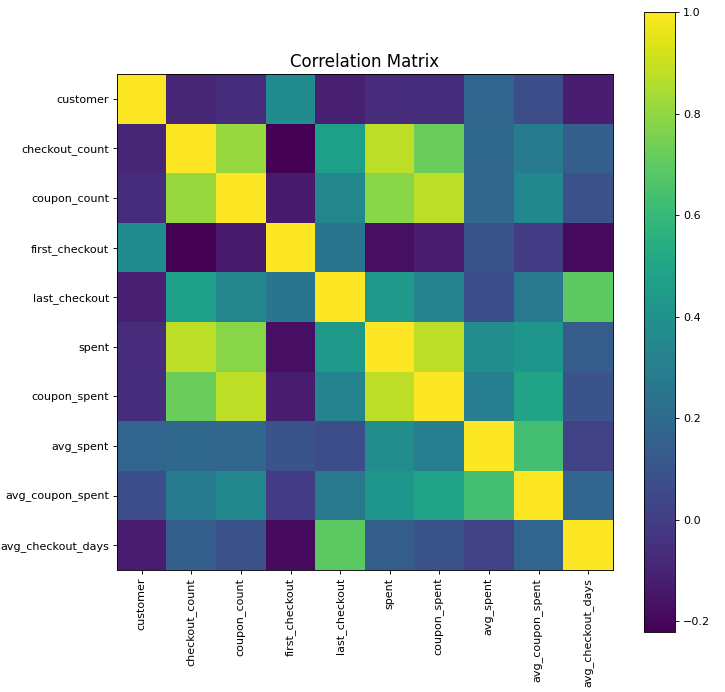


3. Figure Coupon usage median percentages over customers



4. Figure Coupon usage median percentages over customers with over 10 purchases

The 5. Figure shows the correlation matrix between the features. The first that we noticed that there is no correlation between the number checkouts and the first checkout, and between the number of coupon usages and the first checkout. Thus, we assume the coupon-system has no advantage joining any time over another. On the other hand, there is a small correlation between the last checkout and the number of purchases. We assume that staying in the pool is beneficial due to the membership-system.



5. Figure Correlation matrix of the created customer data