Adam Veraszto, Benedek Kornyei,
Zoltan Bereczki

JunctionXBudapest Hackathon
Antavo Challange

Project Documentation

10/23/2022

Budapest

1. 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.

2. 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

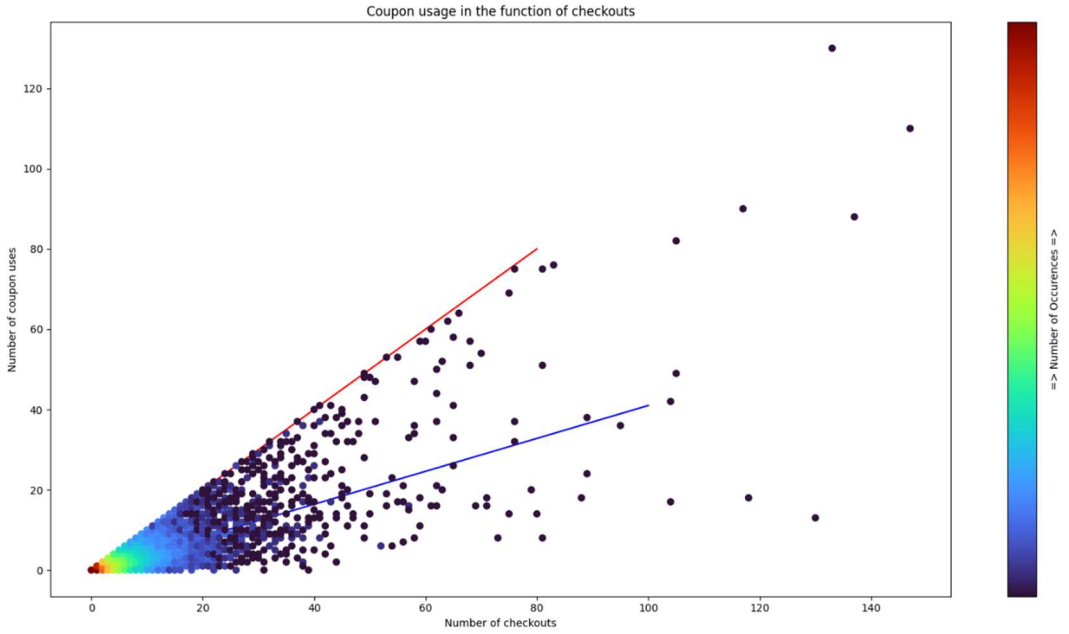
We also created some tables that contain:

- 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.

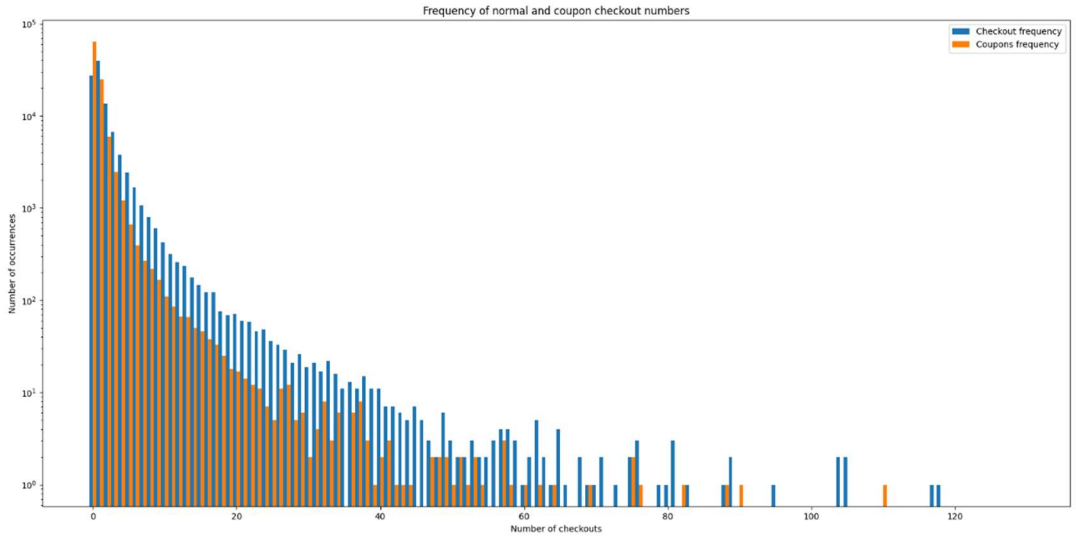
3. Customer statistics and results

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.



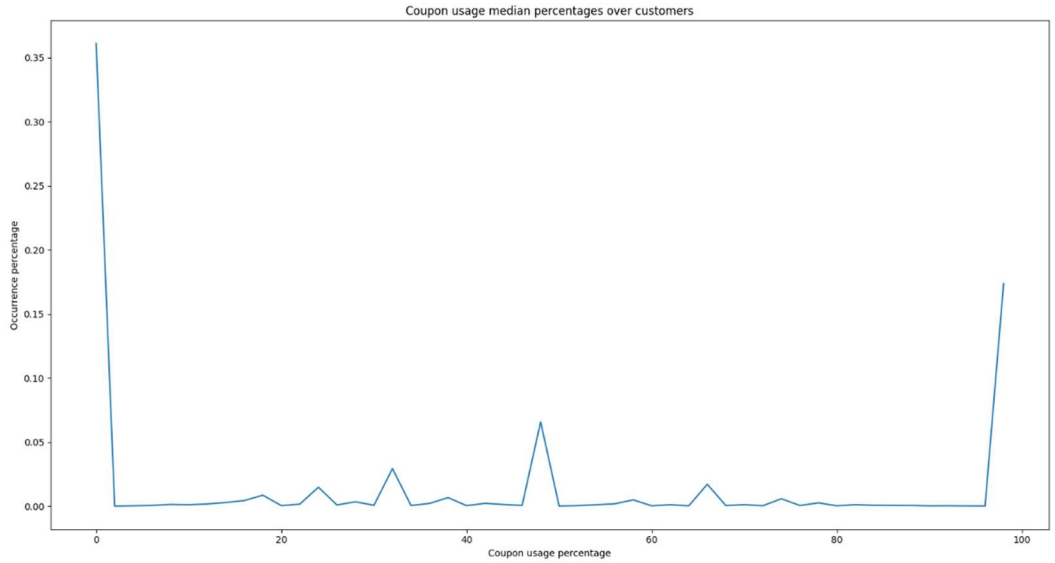
1. Figure Coupon usage in the function of checkouts

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.

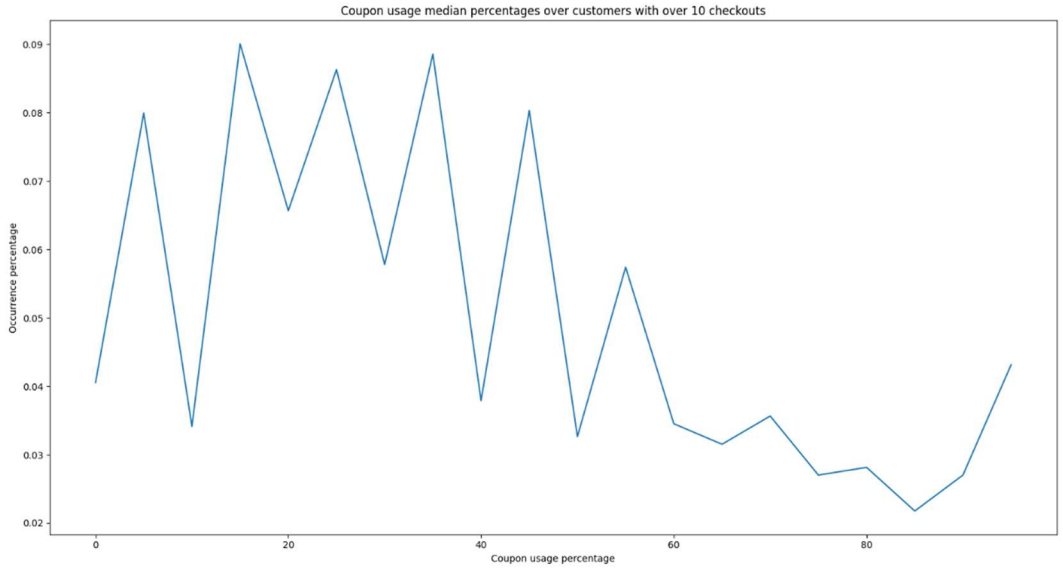


2. Figure Frequency of normal and coupon checkout numbers

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.



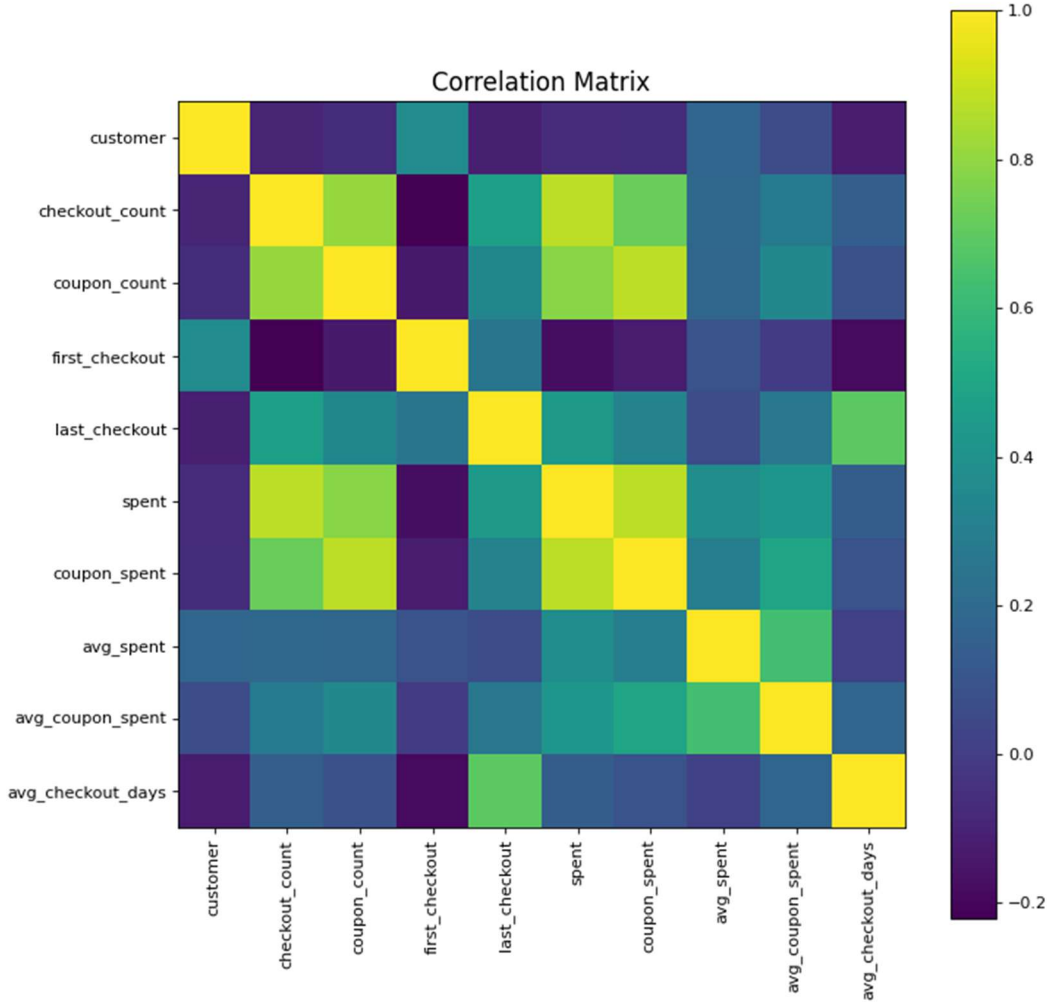
3. Figure Coupon usage median percentages over customers



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

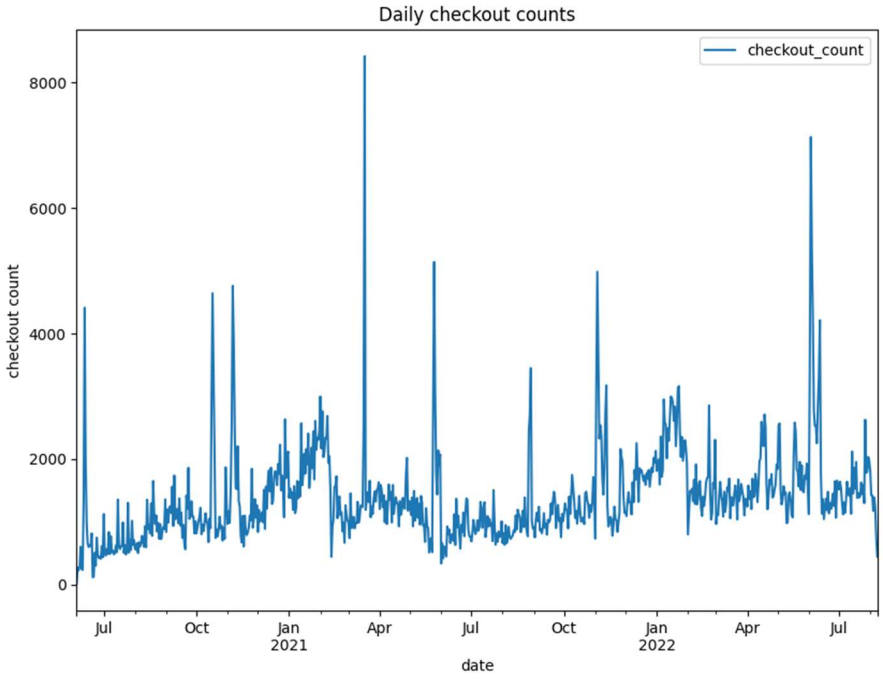
between the last checkout and the number of purchases. We assume that staying in the pool is beneficial due to the membership-system.



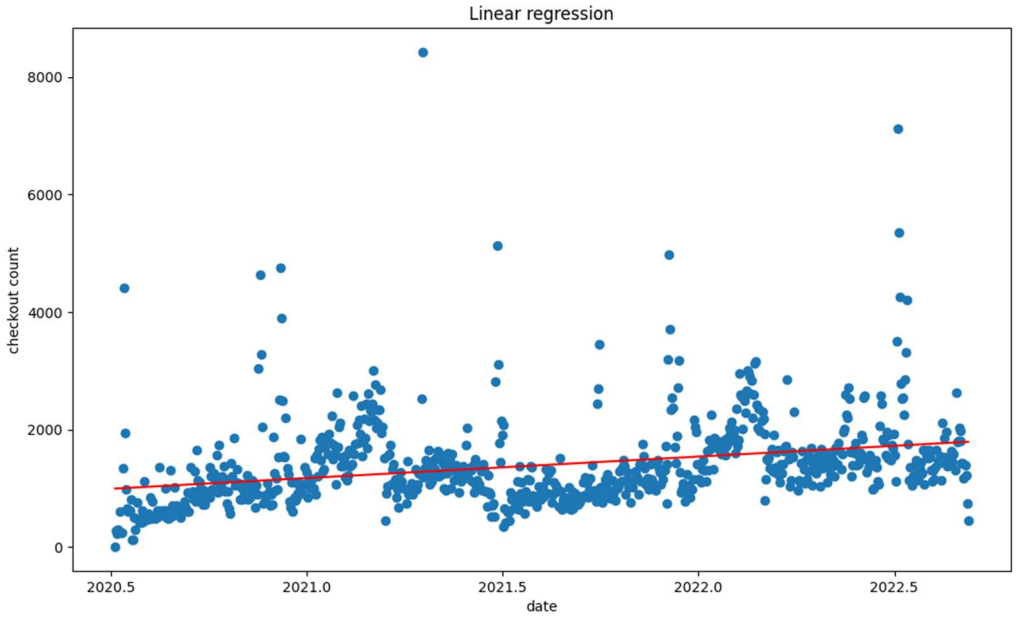
5. Figure Correlation matrix of the created customer data

4. Analytics of daily purchases

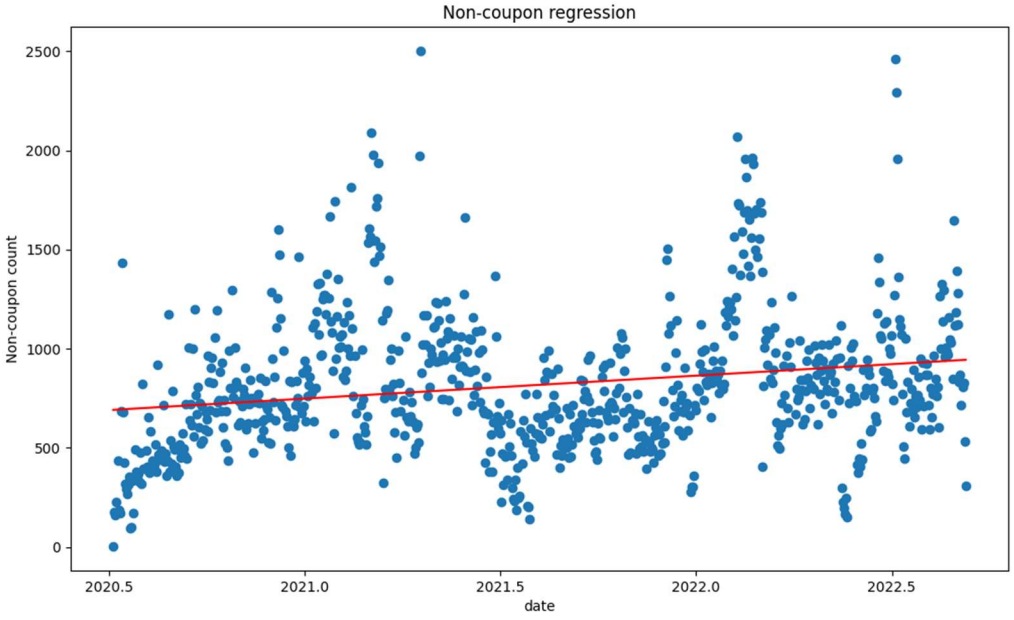
We first simply just plotted the number of checkouts, then we noticed, there might be a small linear increasement over timer. So, we added linear regression on it (7. Figure). We also see that the outlier of this dataset happens to be on the same days, where the outliers are in the daily number of coupon usages. Therefore, we subtract the number of daily coupon usages from the total daily checkouts. This is how we got the daily number of non-coupon checkouts (8. Figure). Interestingly now there was only one outlier and a small linear increment over time. Unfortunately, these charts do not follow any trends, but looking deeper into a chart we might see some regularity Marches. Therefore, we made monthly data tables too.



6. Figure: Daily number of checkouts



7. Figure: Linear regression on daily checkouts

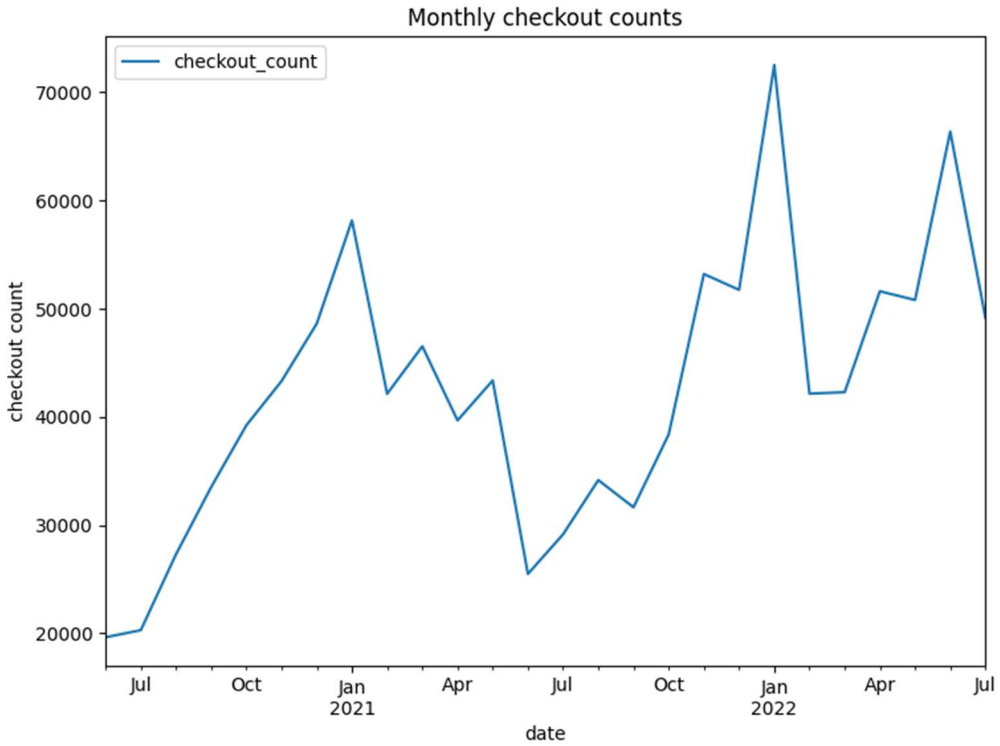


8. Figure: Linear regression on daily non-coupon checkouts

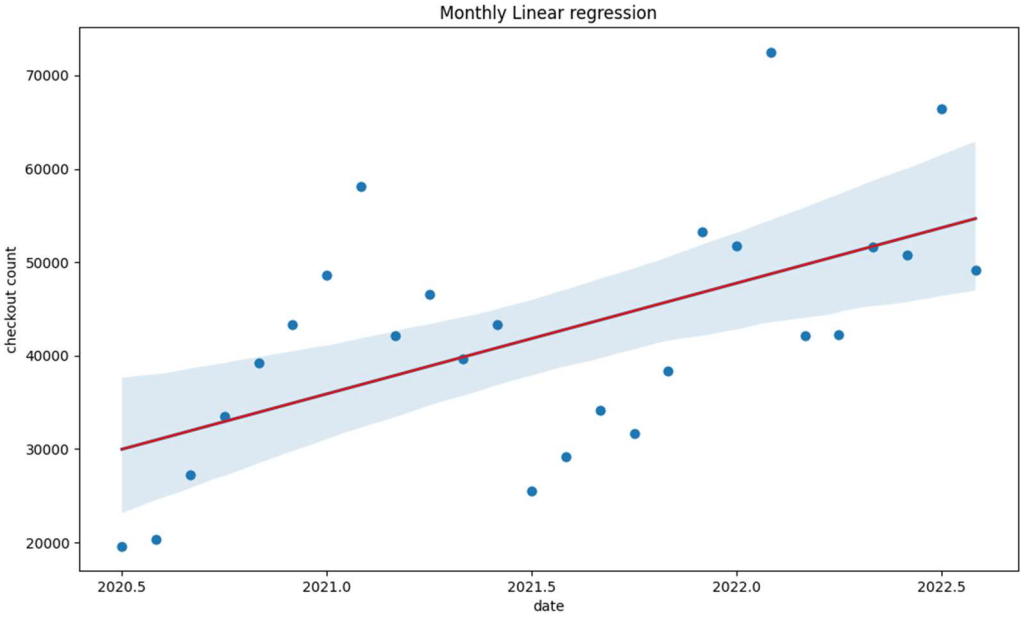
5. Analytics of monthly purchases

After plotting the number of monthly checkouts over time, we got excited as the diagram seemingly showed patterns (9. Figure). So, we put linear regression on this data too (10. Figure), which was not fitting quite well, but we decided to decompose the dataset. We used multiplicative module, and finally we saw clear trend (11. Figure), seasonality (12. Figure). Then we plotted a histogram of the distribution of the series (13. Figure), which was looking a little similar to gaussian curve. Fortunately, the mean and deviation were not the same, there was a clear trend. Therefore, the dataset was not stationary. Then we plotted the first logs and found out the first two have bigger positive correlation (14. Figure), so we set the p parameter of the AR model to be 2. After grid search (15. Figure) we found out that the a (p, q, n) = (1,0,0) the most optimal ARIMA model. Thus, we fitted ARMA model, the parameters (1,0) and we forecasted the dataset (16. Figure). Finally, we used this to predict the amount of checkouts in January of 2023.

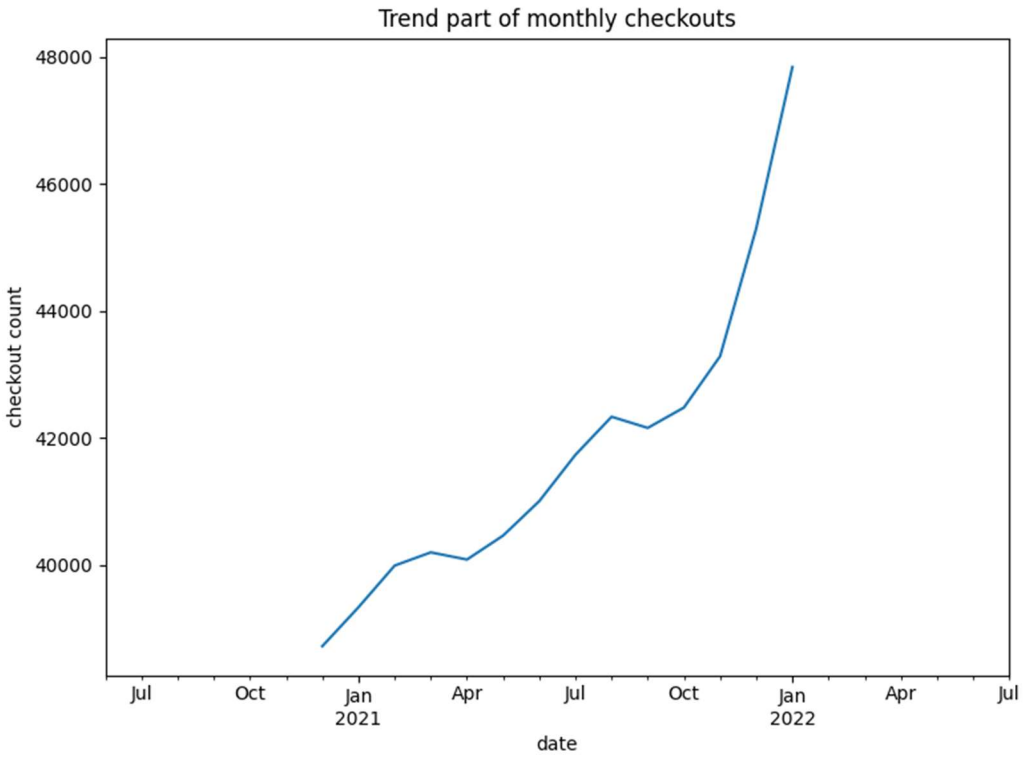
We also did the same thing as before with the spending of non-coupon checkouts. We could also find a trend and seasonality on that dataset. Therefore, we could fit ARIMA model and forecast on that dataset too. We predicted the amount of non-coupon spending in January of 2023. We found out there was no significant difference in percentage increase until that date. Therefore, we assumed that the coupon system can be improved by the earlier mentioned ways.



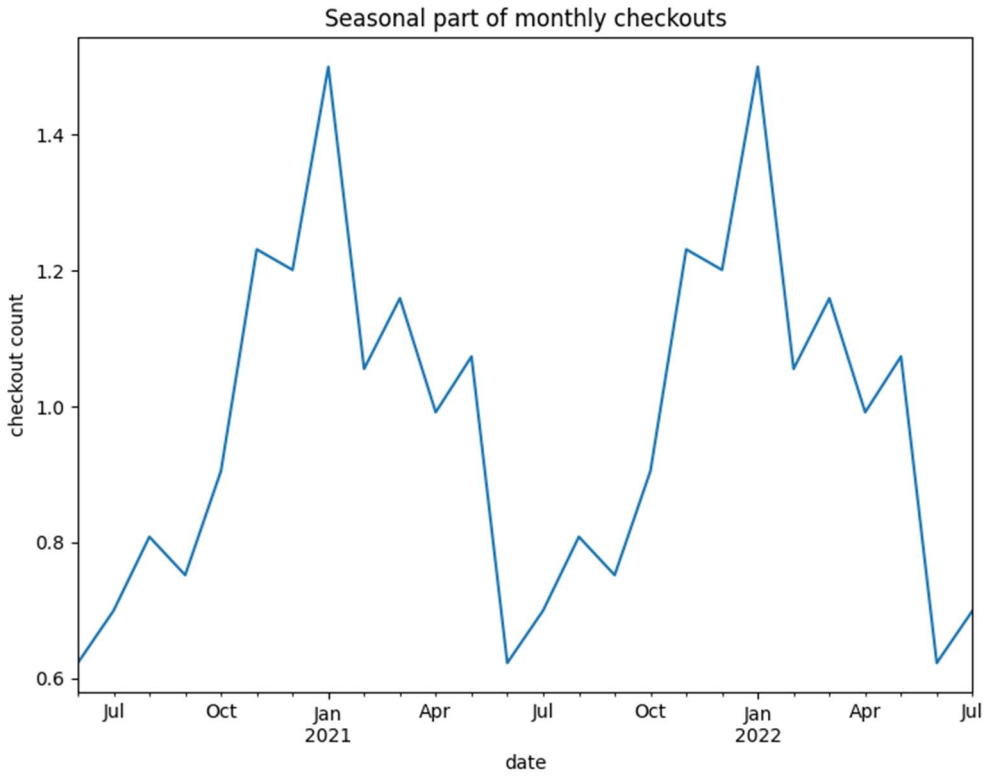
9. Figure: Monthly number of checkouts



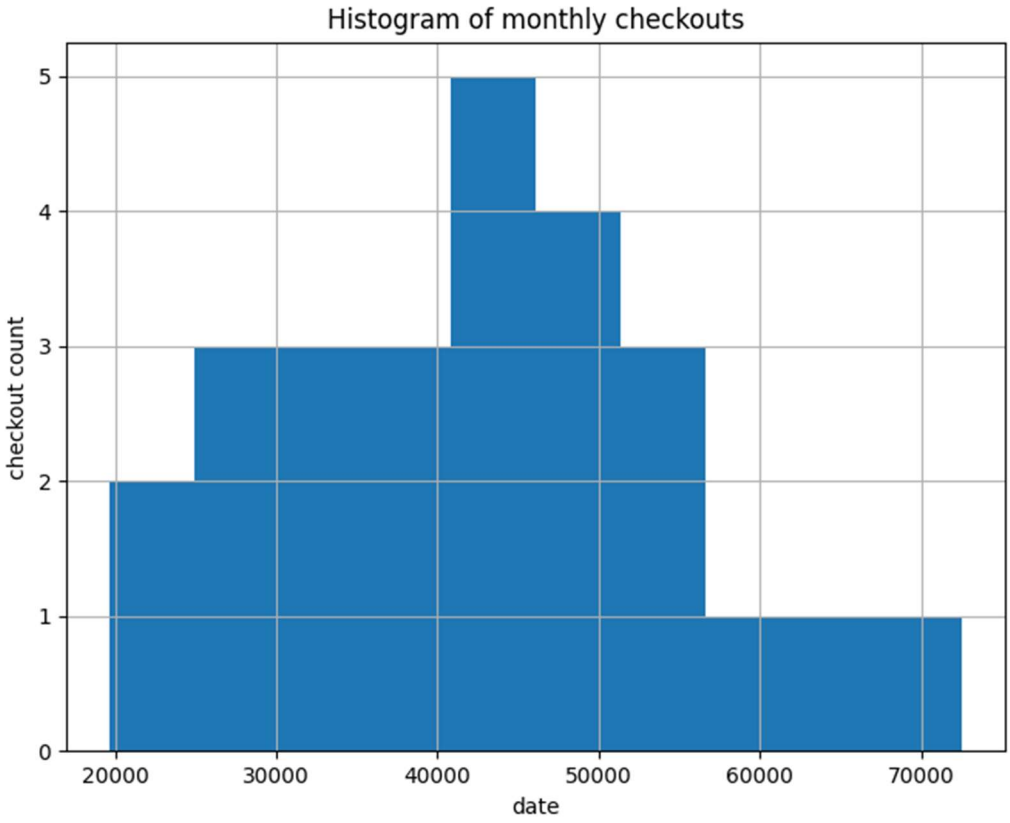
10. Figure: Linear regression on monthly checkouts



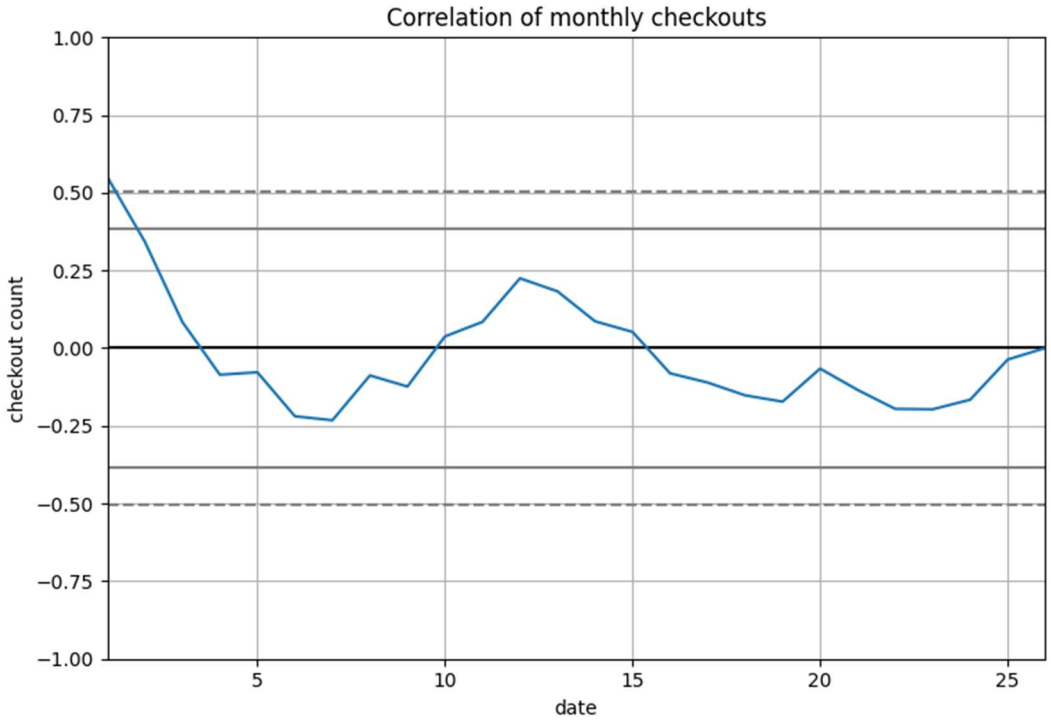
11. Figure: Trend part of monthly checkouts



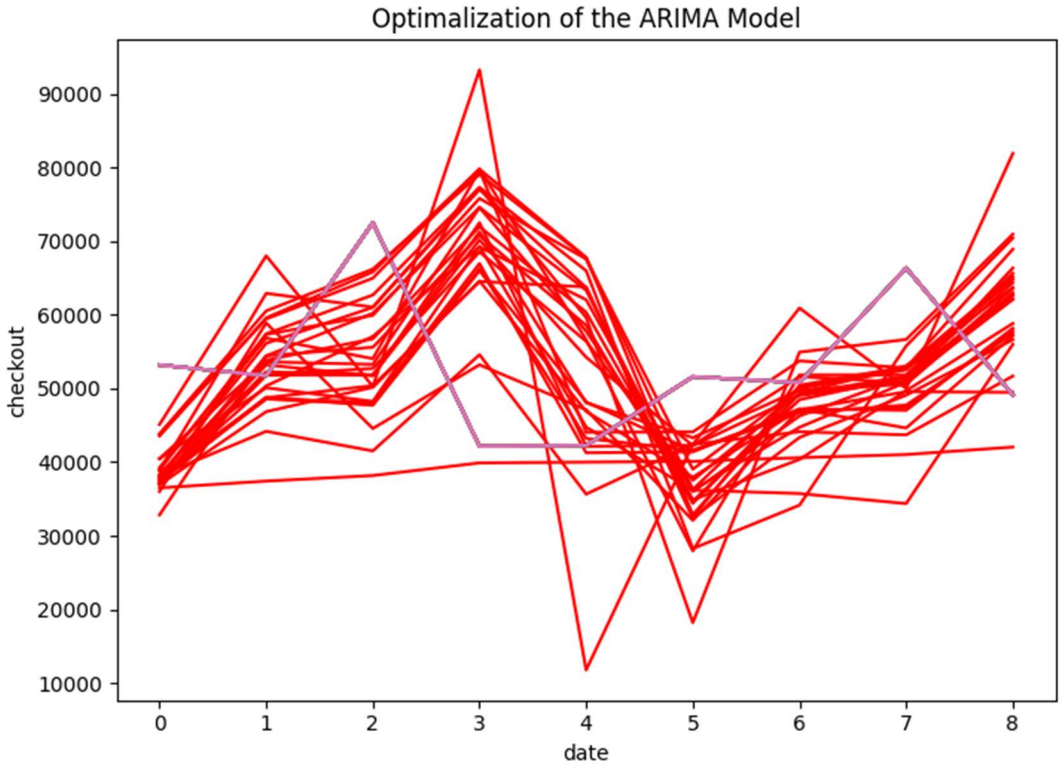
12. Figure: Seasonal part of monthly checkouts



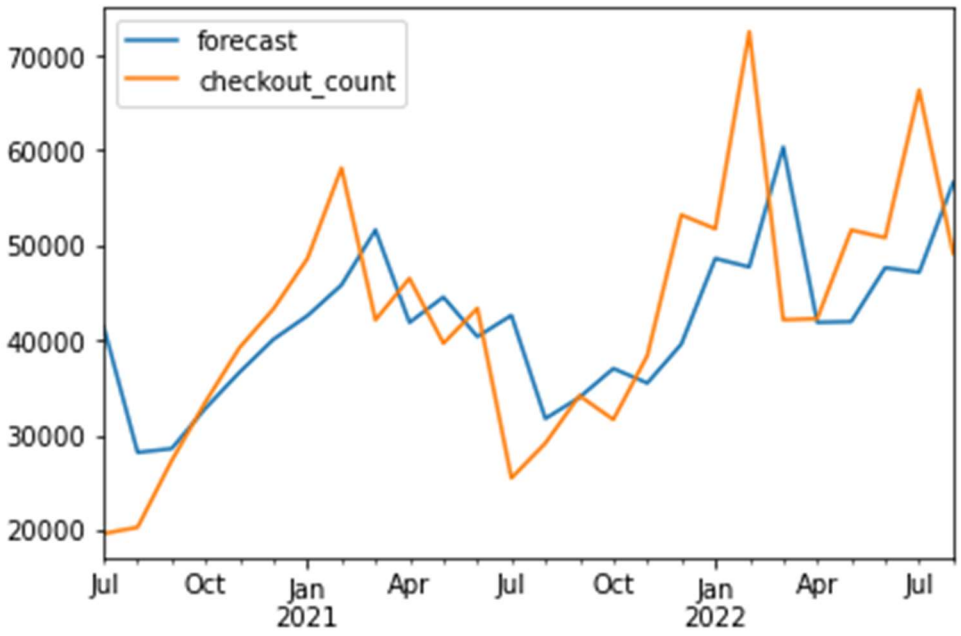
13. Figure: Histogram of the monthly series



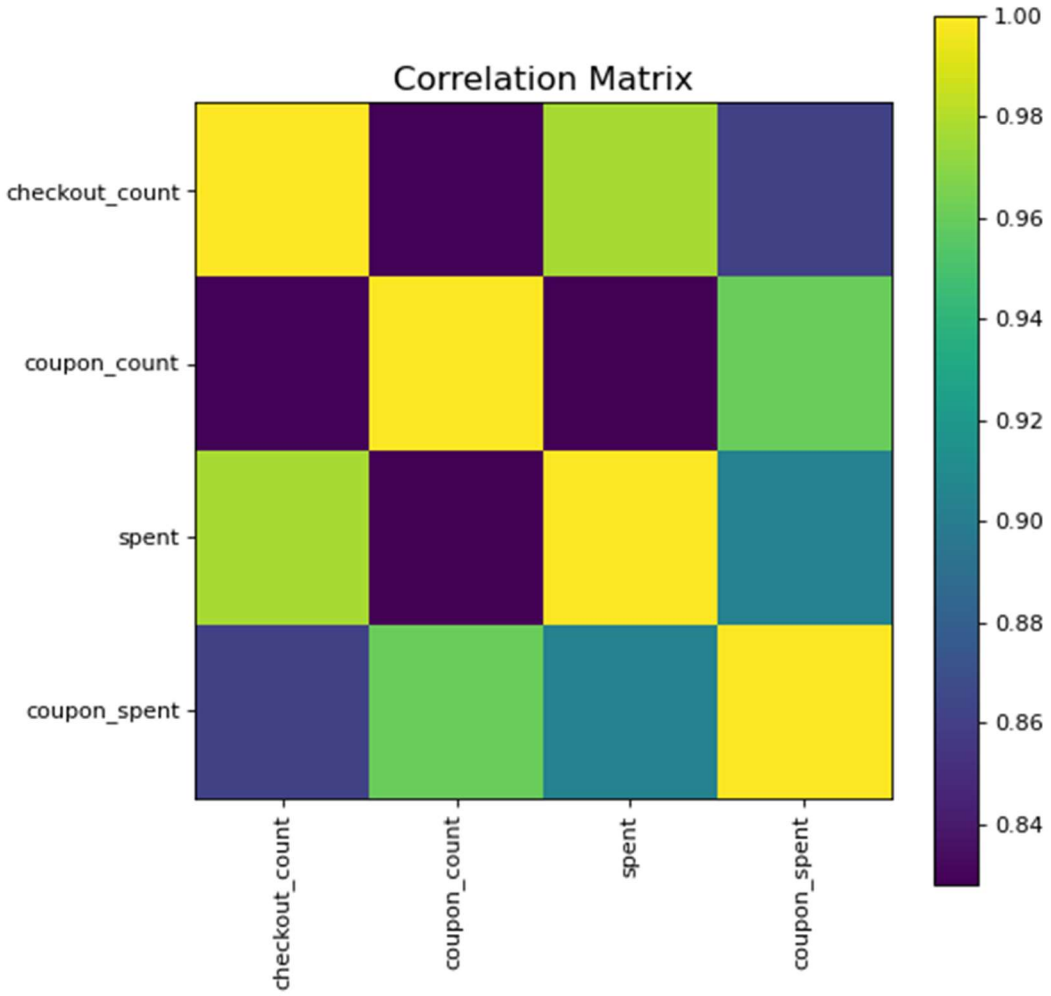
14. Figure: Correlation of monthly checkouts



15. Figure: Grid search of ARIMA

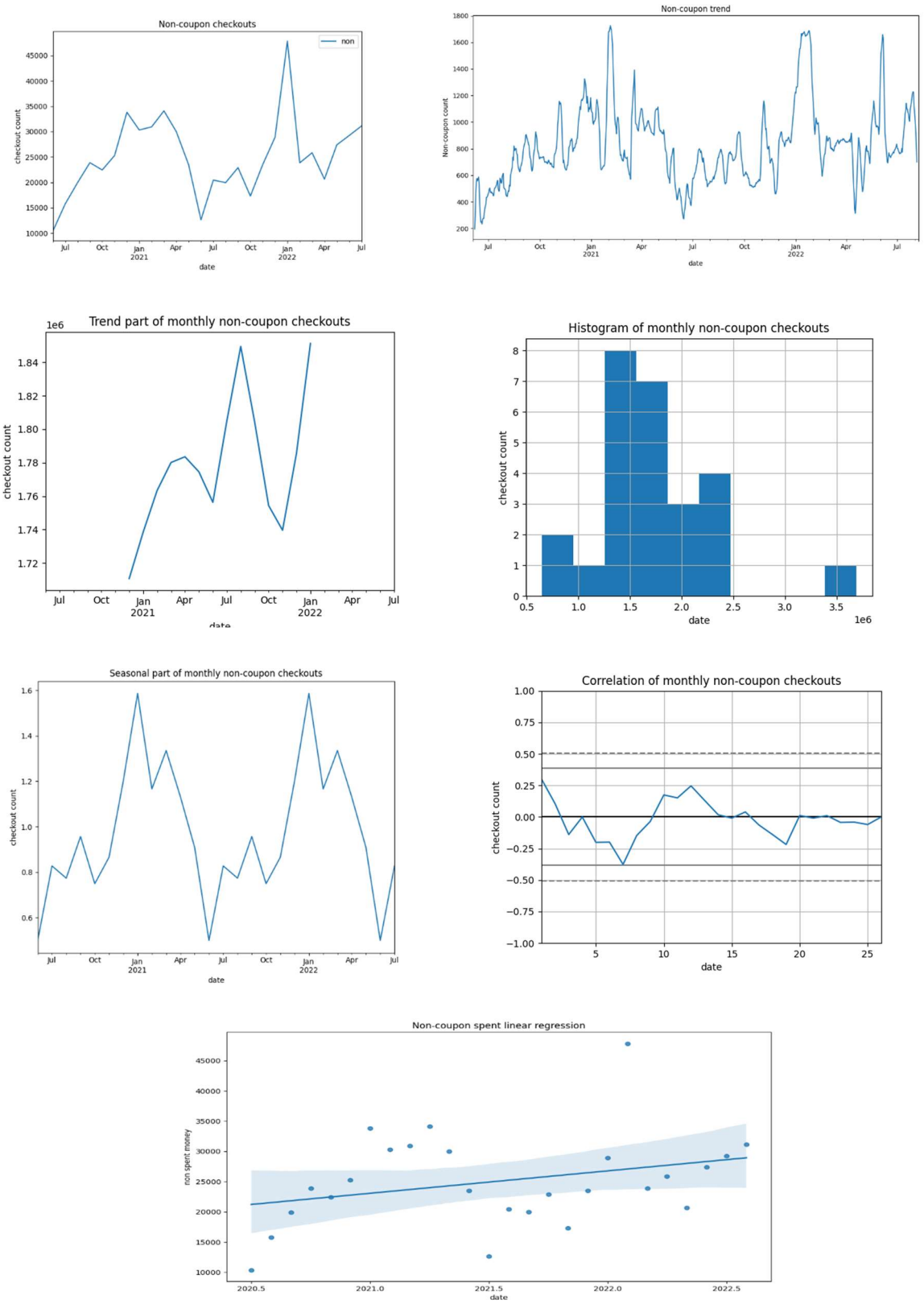


16. Figure: Forecast



17. Figure: Correlation matrix

We made a correlation matrix on the number of checkouts, the amount of coupon usages, the amount of spending on checkouts and the amount of spending on coupon usages. We found out after all, that the correlation was the lowest on checkout count and coupon usage count in the whole matrix. Well it is still a high number but this also suggested us, that the coupon-system is improvable.



18. Figure: Monthly charts of non-coupon data

6. Reconstructing project datasets and visualization

The provided datasets should be placed inside the ‘data’ folder in our project folder. The data folder should contain a ‘gen’ folder, that will contain our generated tables. To be able to run all the scripts of the project, they should be executed in the following order:

- gen_events.py
- gen_reduced_events.py
- gen_checkouts.py
- gen_customers.py
- gen_daily_checkouts.py
- gen_monthly_checkouts.py
- customer_statistics.py
- daily_statistics.py
- monthly_statistics.py

7. Summary

We used the .csv dataset. We created python scripts to generate our specific data and the data processing too. Then we used these datasets to visualize, predict and try to improve the coupon-system and overall experience for customers. We found out that the long-term users are not using many coupons, so you can try to improve that aspect, provide more possibilities and advertisement for these parts of the costumer-pool. We also found out that the current coupon-system could be improved on keeping the new customers, integrate them to not to be just coupon user clients. We think that reducing the number of discounts for newcomers might help, if (for example) the membership status could be upgraded faster.