Point of Sale (POS) Data from a Supermarket: Transactions and Cashier Operations

Final Project Report

MAY 13, 2023

TARAN KUMAR
10470614

Contents

Introduction

As data analysts, In this project, we are asked to apply the skills that we learned throughout the semester in the class and on the data camp to analyze real-world data from a supermarket, focusing on checkout operations and customer behavior. The project is based on two different datasets that are transaction data and cashier operations data and each of these data is spilt into multiple periods. In total, there are six different Excel sheets that are provided to work on for this project.

Before starting the analysis part, it is necessary to understand what the actual problem is and why the analysis is necessary. The data consists of thousands of entries and shows the point of sale (POS) from a supermarket: Transaction and Cashier Operations for a retail store operation. Retail Operations is basically a term that describes the daily task of employees that they do within a retail store or chain. Retail store operations are increasing day by day and are a relatively wide area of research. As mentioned, it is a wide area of research, the work of retail operations is divided into 7 different parts as per research and those are (1) demand forecasting; (2) in-store logistics; (3) inventory management; (4) assortment and display; (5) product promotion; (6) checkout operations and (7) employee management. Every part has its own importance and contributes a lot to the operations of any store. It is said that checkout operations will attract more customers soon. We will investigate this in more depth as we move forward in the report. Now, the manager has provided us with 7 different questions that can help them and their company to see the customer behavior over self-checkout and supermarket service. Moreover, the manager wants to see what the peak hours and the day are. All of this is discussed in detail below.

Dataset Information

The original dataset for the project was obtained from a checkout/POS Data system from a supermarket. It was stored in six different Excel CSV files with the most important information about transactions and cashier operations. The data consist of retail operations of a grocery supermarket located in the city of south Poland. Moreover, the data covers three nearly two-week periods and due to some restrictions in Poland in 2018, the data has some extended hours according to the timings they were allowed to work. Transaction Dataset has 11 variables with around 160k plus values and the other dataset of operations has 6 variables with around 14k values. The operations file has almost all the variables the same as in the transaction because they both are related. The variables included in the dataset are….

- Work Station Group ID: The workstation group ID is basically the type of checkout that whether it was a self-service checkout or a service provided a staff. Self-service is where the customer checks their items themselves and in service they go to the counter where staff can help them to checkout.

- Transaction ID: Transaction ID is a unique number and contains, data, store ID, checkout ID, and the sequence number.
- Begin Date Time: This basically is the time and date when the transaction is started.
- End Date Time: This is the time and date when the transaction is ended
- Operator ID: This is the cashier ID and it is unique for all cashiers
- Transaction Time: The time taken to do the transaction.
- Break Time: It includes the break time and the idle time in seconds
- Art Num: The Art Number tells the size of the basket and the number of items
- TN cash: This shows that the transaction is paid in cash
- TN card: This shows the transaction is paid by card
- Amount: The value of the transaction, basically the amount spent
- Items: This shows what times were bought and the operation identifier

All of above shows the brief information on what the variables mean. It is important to understand all the variables in detail because every variable has its own importance, and it will help us in data visualization and analysis.

Data Cleaning

Data cleaning is an important part of research and should be done before starting any project. It is a process of removing the data that has errors and data that is incorrect, corrupted, or incomplete. In this research, the cleaning process was carried out to make sure the data is error-free. The data cleaning was done using Python code. At first, I imported the data into Python using importing pandas and datetime modules. All six Excel files are imported into the phyton. Three files are of transaction and the rest three files consist of operators. Then, these 6 files are merged into 2 categories. Once, the data is merged, after this, using Python code all the invalid data rows and missing values are dropped off. Other than this, I asked Python to remove the value that duration is less than 0 or more than 1 hr. The reason is, there is no transaction that takes more than 1 hr. even if you have a large basket size. Not only this, but some of the formatting is also done in the cleaning part and the data is grouped by cashier and date and then summary statistics are calculated such as sum, count, and mean. Finally, once the data is cleaned, it is saved and used as a new file so that other codes can be printed on this.

As mentioned above, it is very important to clean the data because it makes it easy for everyone to interpret and visualize the data. Not only this, but it also makes the data and interpretation more accurate. Luckily, that was good, and didn’t have that much of an issue while cleaning. Lastly, the result of the data is shown in the Python code file.

Data Preview

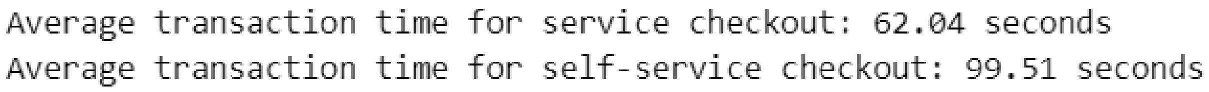
As mentioned in the above paragraph at the initial stage before starting the exploratory analysis, I cleaned the data, and then that data is used to do the analysis with the help of visualization in some of the questions. The data is going to help us to see customer behavior and is focused on the checkout operations that is the person using self-checkout or using the staff counter for checkout. Not only this but I have tried my best to answer all the questions provided to us by the manager in this project.

Exploratory Data Analysis

Now, we know what the data is about and we have got the data cleaned. It is time to do the analysis of the data and get answers to the manager's questions. Here, the main purpose of the manager is to see the behavior of customers, peak times, the average size of basket that customers go for, and what time of check-out they prefer. We will discuss all this in detail below with seven questions that the manager asked….

1. The average transaction time for each checkout type: service

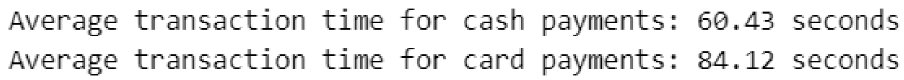
The average transaction time for each checkout type means how much time it requires on average to check out in the supermarket. Here, we have used Python code to do the analysis. First, the code filters the merged dataset and creates new data frames: one for self-service checkout and one for service checkout. The self-service checkout is basically the machine that provides a customer with a mechanism to process the transaction without the help of any person. On the other hand, service checkout is where a human being/cashier is available for you to give you the service and help you to complete your transaction. Now, as two new data frames are created, we then calculate the average transaction time of each type of checkout using the mean function on the transaction time in Python code. And finally, we printed the code. According to the code, we can see that the average transaction time for service checkout is **62.04** seconds and the average transaction time for self-service checkout is **99.51** seconds. I formatted the result to 2 decimal places so that it is easy to understand. As we can see from the result, we can analyze that it takes lesser time for normal service checkout than the self-service. The reason might be cashiers are trained and it is their daily task, compared to the self-service checkout. In self-service checkout, it takes a bit more time to scan the items and sometimes the machine is new to the new customers.



```
Average transaction time for service checkout: 62.04 seconds
Average transaction time for self-service checkout: 99.51 seconds
```

2. Payment method (cash vs. card) impacts transaction time

This is basically to see how the payment method (cash vs card) impacts the transaction time. The reason is the less the transaction time, the more transactions can be done. The cash payment method is where payment is done by cash and the card payment method is where the customer uses their debit or credit card to pay the amount. Here, I have used Python to see the average time that a transaction takes place through cash and card. The first step to do this was to filter the data for the cash and card payments by removing any non-numeric values and the values with errors. This was done to avoid any errors in the calculation. The result is then stored in the cash data variable. Then the next step was to get the average transaction time by taking out the mean of transaction time. This process was followed for both card and cash payment methods. The result that I got for the average transaction time for cash payment is 60.43 seconds and the average time for card payment is 84.12. This means according to the data card payment takes more time for a transaction to complete compared to a cash payment. One of the reasons is that it takes time to read the card by machine and sometimes the card requires you to input the zip code and the pin code. Moreover, sometimes the chip doesn't work and then the card needs to be swiped which takes more time. In the case of cash, it is easy to take and return the change. According to one paper written by Michal Polasik on the efficiency of the POS payment method, his results confirm that cash is significantly a faster way of payment however in today's time most people don't like to carry cash and use their card/phones instead to pay the amount which is competitive to cash in term of the efficiency. I would say that both methods are good and depend on what customers prefer more as per their feasibility.

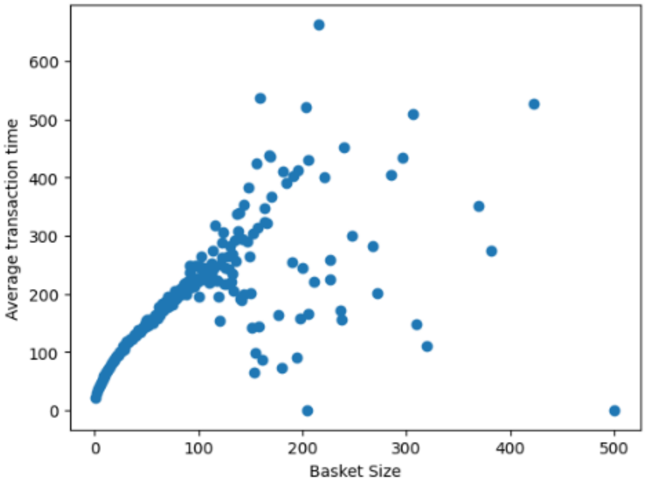


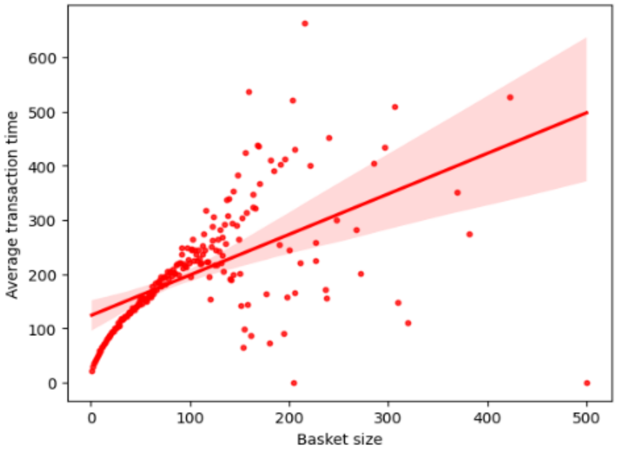
```
Average transaction time for cash payments: 60.43 seconds
Average transaction time for card payments: 84.12 seconds
```

3. Does the average transaction time change with the basket size? Is there a non-linear relationship between these two variables?

This question talks about the basket size and the average transaction time. The manager wants to see if the average transaction time changes with the basket size or not. To do this, we used a Python code where codes read the merged dataset file and group it by the basket size("ArtNum") using group by method, and the mean function is used to average out the transaction time. Both the data are then stored in a new data frame called "basket data". To analyze the result, I have used the regression model here and plotted a scatter plot to see if the relationship is linear or non-linear and if it is significant or not. According to the analysis done using Python, we can see in the graph mentioned below and can analyze that there is a significant relationship between the average transaction time and the basket size. This means that if the basket size is high, it will take more time to complete the transaction. The reason is if you have more items in the basket, it will, of course, take more time to process the transaction as you need to scan it every time and then need pay for it, no matter If it is by card or cash. So, we can say that there is a positive relationship between the two variables. Now, to determine if the relationship is linear or not, we will use the regression line here which is mentioned in the scatter plot below. From the plot, we can analyze that the line is linear however, it has a slight curve at the end so it is very difficult to make an exact decision that the two variables are exactly non-linear or linear. This is also because of the limited data provided to us.

The first scatter plot shows only the scatter plot and in the second one I have added the regression line using Python. I believe using this regression model can help managers a lot to see the changes in the transaction time taken with the different sizes of baskets. This way managers can use this information and can increase the number of cashiers when the basket size is more, or they can also open more self-checkout machines.

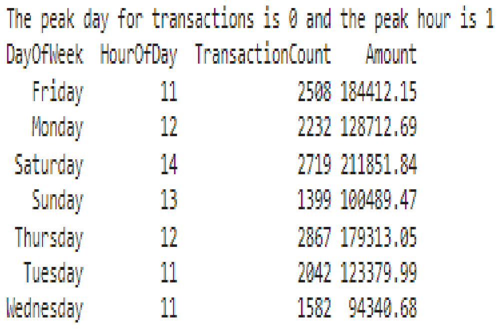




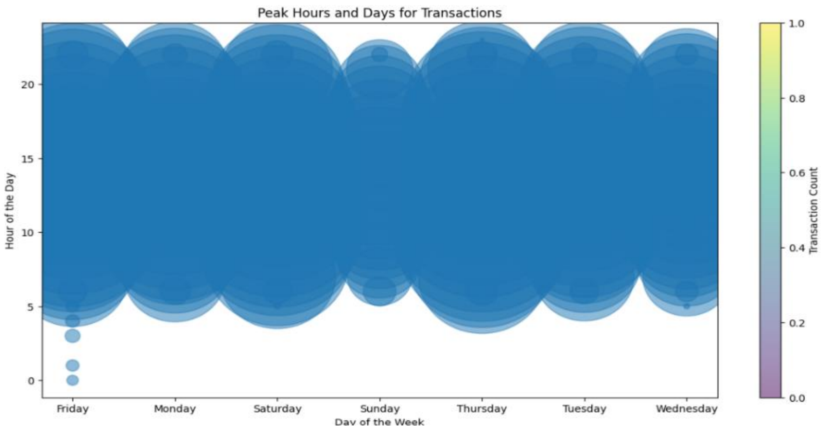
4. Peak hours and days for transactions at the supermarket? Are there any patterns or trends?

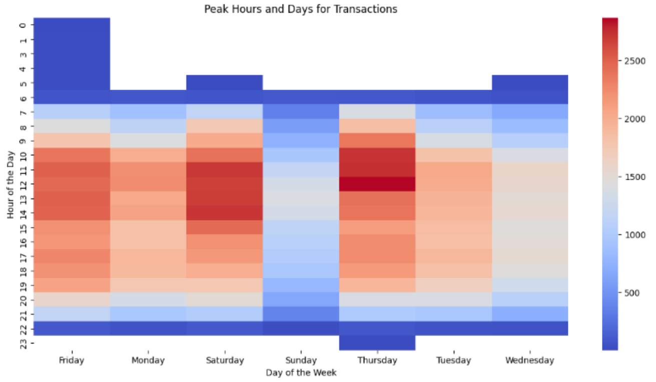
Now, as we move forward, we can see that the manager is trying to gather information regarding the customers that when they are coming the most to buy the items, and at what time. This is going to help them to make adjustments accordingly so that the best service is provided to the customer with fewer resources being wasted. In this question, I am using the code again to see what the peak hours and the days for transactions at the supermarket are and to see if there is any pattern/trend or not.

Here I have first converted the Begin Date Time column to create a new variable Hour of day and day of the week to see the peak and the pattern between them if any. Then the transactions are grouped, and the visualization is done using the seaborn heatmap and the scatter plot. The reason for using two different kinds of visualization was to check whether both visualizations give the same analysis or not. Both the scatter plot and heatmap show the same. They both show the peak day and hours of the transaction. Not only this but I have created 4 different graphs and a table to visualize this in more detail so that it is easy to make analysis and it makes it more accurate.



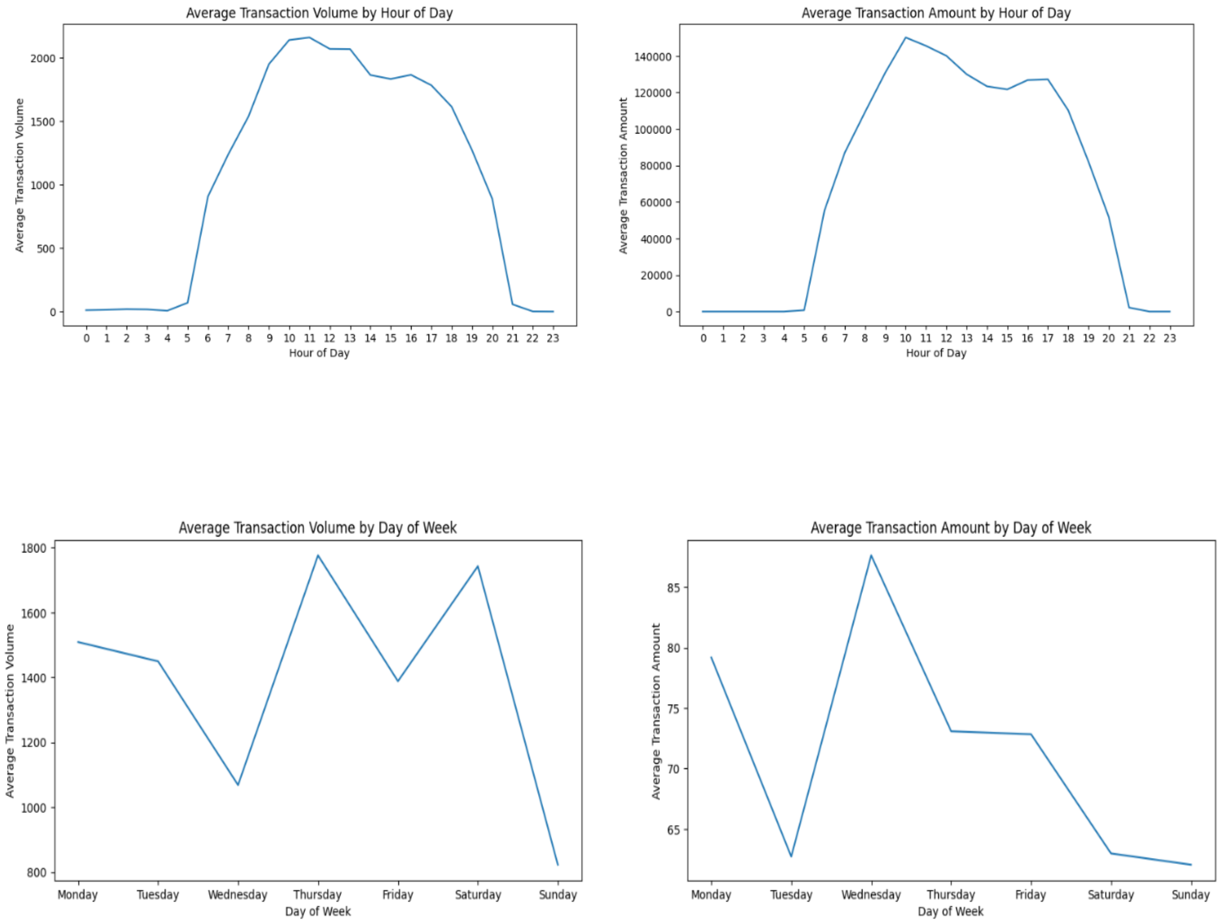
```
The peak day for transactions is 0 and the peak hour is 1
DayOfWeek  HourOfDay  TransactionCount     Amount
   Friday         11              2508 184412.15
   Monday         12              2232 128712.69
 Saturday         14              2719 211851.84
   Sunday         13              1399 100489.47
 Thursday         12              2867 179313.05
  Tuesday         11              2042 123379.99
Wednesday         11              1582  94340.68
```





As we can see above, I have plotted the heatmap, scatter plot and the table that shows the pattern and the trend. We can analyze that there is an increase in the number of transactions on weekends compared to weekdays. The reason is that people are free on weekends, so they prefer buying groceries and other items on weekends compared to weekdays. Secondly, we can see that there is constant growth in the daytime and evening, from morning till evening but after evening there are less no of transactions like from 10 pm onwards. We can also see that the afternoon time like around 1 pm to 2 pm has the highest trend and most people shop at this time. This analysis can help the manager to manage the staff, inventory, stocking, and management of the supermarket. As an add-on, I have also made 4 graphs using Python which can help managers to get the analysis in depth.

Patterns in the peak hours and days for transactions



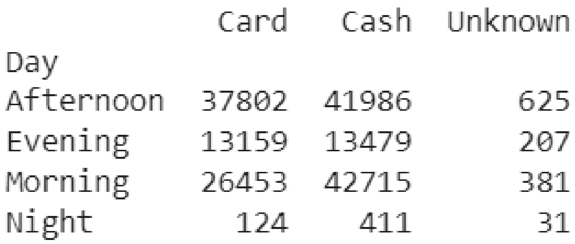
5. Break times and their durations affect the transaction time of the following transactions.

In this, the manager wanted us to use the regression model, to predict transaction time by using three different variables that are basket size, payment method, and checkout type. Here, I have used Python code to run the regression model and to predict the transaction time using different variables. The data was read into the pandas first and then null values were neglected. Then using dummy variables, new columns were created and then the code was run using different steps. Here, I have used an assumption that anyone who has taken a break of more than 60mins means their shift is over. This is because, if the staff is going out for more than 60mins that means the shift might have ended. After this assumption, the model was used to see the effect of break duration on the transaction time.

Looking at the code, we can analyze that the break duration and transaction time are statically significant. If the cashier takes more breaks, people might like it and they might not come next time. However, this model analysis can work both ways. Maybe if the cashier is fresh, they might work more efficiently and result in less transaction time. We can also see that in this model, the mean squared error is 1685.48 and the r-square is 0.541, the r-square evaluates the performance of the model and the mean squared error shows the error overall. Looking at the r square I can say that this analysis needs more information so that we can see and confirm that how much break does the cashier take and what impact they have on the transaction time then.

6. New variable representing the time of day (morning, afternoon, evening, and night) based on the Begin Date Time. How do payment methods (cash vs. card) vary across different times of the day?

Here, I have created a new variable representing the time of the day and have tried to show how the payment method varies across different times of the day. I started by creating a new variable Time of the day that is based on Begin Date Time. Then with the help of code, I grouped the data by different times of the day like morning, afternoon, and night and it then shows the different payment methods used.
We can see in the below tables that were created using Python and can analyze that most transactions are done in the morning time using the cash method of payment. On the second number, it’s afternoon time and it has mostly the card payment. This can help managers to make decisions such as which shifts are the most important and they can appoint more cashiers at that time. Moreover, by looking at the payment method they can take out an estimate that how much time overall is going to spend on the transactions. So, this can help managers a lot in making decisions.



```
            Card   Cash  Unknown
Day
Afternoon  37802  41986      625
Evening    13159  13479      207
Morning    26453  42715      381
Night        124    411       31
```

7. Factors that are the most significant predictors of choosing self-service over cashier service? Do consumers prefer using self-service checkouts during peak hours compared to regular hours?

Lastly, the manager wants to see what some of the most important and significant factors are that people choose self-service over cashier and do they prefer using self-service checkouts during peak hours compared to regular hours. To do this, we used the same files that we used in the above questions and predicted the probability using an accuracy score. We can see in the score that the accuracy is 0.77 which is a good number, and the regular hour probability is 22.83 means around 10 Pm. Moreover, we can see that some of the factors that are most significant for choosing self-service over cashier service are that it is easier and there is no need to stand in line for the cashier. Other than this, customers like to use self-service during peak hours rather than regular hours. Self-service is a good option but sometimes it has difficulties and then you need to call the assistant which takes time. We can say that people prefer using self-service checkout during peak hours compared to regular hours so that their time can be saved and they don’t need to wait longer ques.

Conclusion & Recommendation

As we have seen in the above questions asked the manager. I have tried to answer all the questions and we can see that managers can use this information to make many management decisions. By looking at the peak times, transaction methods, and breaks, the manager can manage the store timing, staff, and inventory level. The information can be used in many other ways also, but I would recommend that some more aspects should be taken into consideration other than these questions. Such as the competition. Like who are the competitors, what pieces they charge, and what items they sell. Lastly, I would say that it was great working on this project, and I look forward to continuing this in the future if I have an opportunity.