BIA-652 Final Project Report

Point of Sale (POS) Data from Supermarket:

Transactions and Cashier Operations

By: Akhil Chippalthurthy

CWID: 20012188

Introduction:

Checkout operations are an important part of the retail experience. They can have a significant impact on customer satisfaction and loyalty. In recent years, there has been a growing trend towards self-service checkouts. Self-service checkouts offer several advantages, including faster service and lower costs. However, they also have some disadvantages, such as the potential for errors and customer frustration.

This report analyzes checkout operations data from a grocery supermarket in Southern Poland. The data covers three nearly two-week periods: December 7-19, 2017; February 13-26, 2019; and March 28-April 10, 2019. The goal of the analysis is to identify trends and patterns in checkout operations and to develop models that can be used to improve checkout efficiency.

Data Merging:

There are six separate datasets. Three of them are transactions and cashier operations datasets. I have combined the transactions and cashier operations datasets into one dataset which will make it easier to answer the questions below.

Questions:

1. **What is the average transaction time for each checkout type: service (WorkstationGroupID = 1) vs. self-service (WorkstationGroupID = 8)?**

A. The output shows that the average transaction time for service checkouts is **62.037** seconds, while the average transaction time for self-service checkouts is **99.512** seconds. This means that service checkouts are on average **37.474** seconds faster than self-service checkouts.
There are a few possible reasons for this difference in speed. One reason is that service checkouts have cashiers who can help customers with their transactions. Self-service checkouts do not have cashiers, so customers must be able to scan their own items and complete their transactions on

their own. This can be a challenge for some customers, especially those who are not familiar with self-service checkouts.

Another reason why service checkouts may be faster is because they have more experienced cashiers. Cashiers who have been working for a longer period are typically more efficient at scanning items and completing transactions.

Overall, the data suggests that service checkouts are faster than self-service checkouts. This is important information for supermarkets to consider when making decisions about how to configure their checkout areas. If a supermarket wants to improve checkout efficiency, it should consider increasing the number of service checkouts.

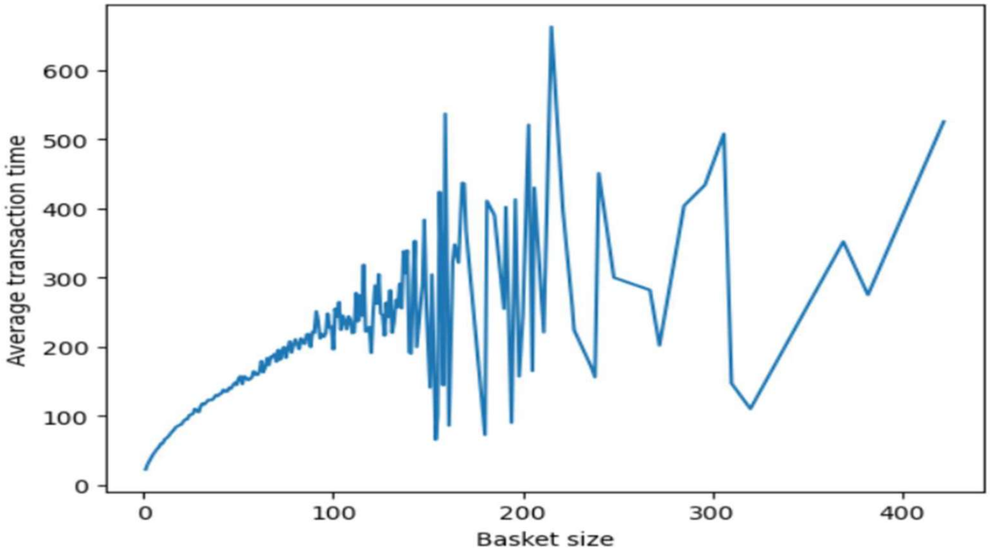
2. **How does the payment method (cash vs. card) impact transaction time?**

A. I found an outlier in the data where both the cash and card transactions were marked as 'False', but there was still a transaction time. Therefore, I removed the rows where both cash and card were marked as 'False'. This is done to ensure that the data only includes rows where a payment method was used. The code then filters the data by payment method type (card vs cash). This is done to calculate the average transaction time for each payment method type. Finally, the code prints the results.

From the results, we can see that the average transaction time for card payments is higher than that for cash payments. This may be due to the additional time required for card authorization and processing. However, we cannot conclude this with certainty without further analysis. Additionally, we can also see that the difference in transaction time between the two payment methods is quite significant.

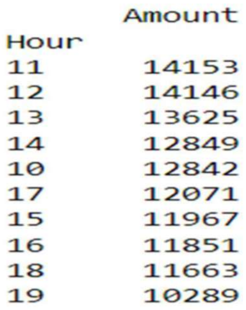
3. **How does the average transaction time change with the basket size? Is there a non-linear relationship between these two variables?**

A. The output graph shows that the average transaction time increases with the basket size. This is because larger baskets require more time to scan and process. The relationship between average transaction time and basket size is non-linear. This means that the increase in average transaction time is not constant as the basket size increases. Overall, the data suggests that the average transaction time increases with the basket size. This is important information for supermarkets to consider when making decisions about how to configure their checkout areas. If a supermarket wants to improve checkout efficiency, it may want to consider offering express lanes for customers with small basket sizes. Here is the output graph:

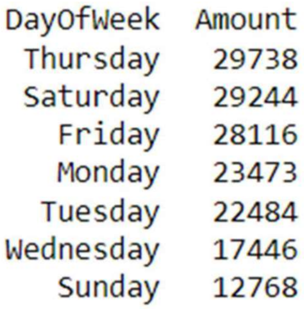


4. **What are the peak hours and days for transactions at the supermarket? Are there any patterns or trends?**

A. According to the output, the peak hours for transactions are around 11:00 AM to 2:00 PM, with the highest number of transactions occurring at 11:00 AM and 12:00 PM. The peak days for transactions are Thursday, Saturday, and Friday, with Thursday having the highest number of transactions and Sunday having the lowest number of transactions. These findings suggest that customers tend to visit the supermarket during the weekdays more often than on the weekends, and that they tend to do so during lunch hours. Below shows the amount of transactions by hour and day:



```
      Amount
Hour
11     14153
12     14146
13     13625
14     12849
10     12842
17     12071
15     11967
16     11851
18     11663
19     10289
```

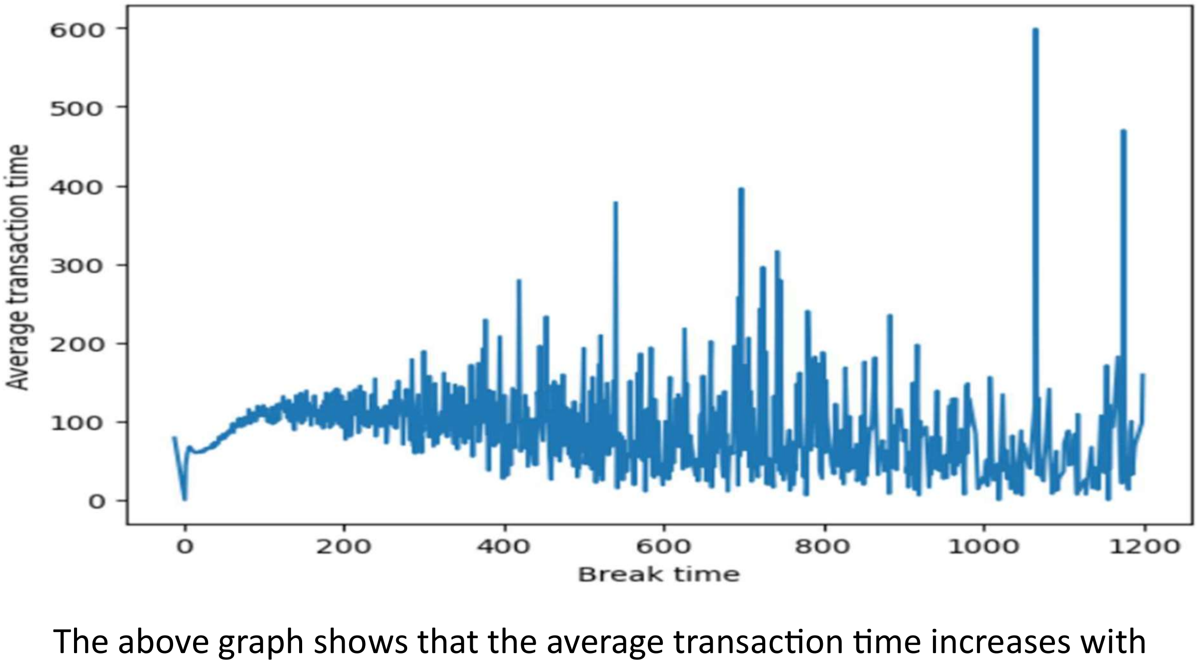


```
DayOfWeek  Amount
 Thursday   29738
 Saturday   29244
   Friday   28116
   Monday   23473
  Tuesday   22484
Wednesday   17446
   Sunday   12768
```

5. **Develop a regression model to predict transaction time with at least the following variables: basket size (ArtNum), payment method, and checkout type. Use the model to answer the following question. How do break times and their durations affect the transaction time of the following transactions?**

A. The model used is a random forest regression model to fit the data and make a prediction for transaction time. The data is cleaned by dropping all rows with missing values and converting the 'ArtNum' and 'BreakTime' columns to integer data type.

After the data is prepared, the model is fitted to the data using the 'fit' method from the 'RandomForestRegressor' class. The model is then used to make a prediction for transaction time using the 'predict' method. The accuracy of the model is evaluated using the 'score' method, which returns the coefficient of determination (R^2) of the prediction. The coefficient of determination is a measure of how well the model fits the data, with higher values indicating a better fit. In this case, the accuracy of the model is 0.659, which is a moderate fit. This means that the model can predict the transaction time with **65.96%** accuracy.



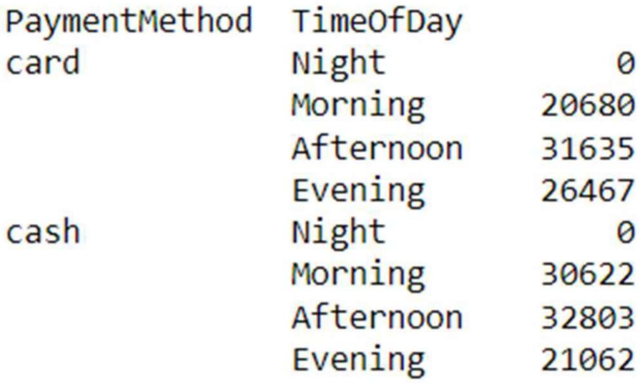
The above graph shows that the average transaction time increases with the break time. This is because customers who take breaks are more likely to have larger baskets and to use cash payments. Both factors can lead to longer transaction times.

6. **Create a new variable representing the time of day (morning, afternoon, evening, and night) based on the BeginDateTime. How do payment methods (cash vs. card) vary across different times of the day?**

A. A new variable has been created, called "TimeOfDay" by mapping the hour of the transaction to a time of day using a dictionary that maps ranges of hours to specific time of day categories. The payment method variable has also been transformed to "cash" or "card" using a lambda function based on the "TNcash" variable. The code then calculates the count of transactions for each payment method and time of day by grouping the data by the "PaymentMethod" and "TimeOfDay" variables, and using the "size()" method to count the number of occurrences in each group.

The output shows that there are zero card transactions during the night, while most transactions for both cash and card occur in the morning and afternoon. There are slightly fewer card transactions in the evening compared to the morning and afternoon, while cash transactions are slightly lower in the evening compared to the morning and afternoon.

Below is the output to show how the payment methods vary across various times of the day:

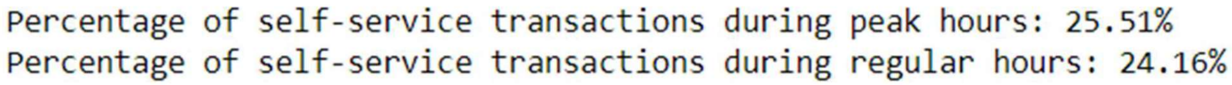


```
PaymentMethod  TimeOfDay
card           Night            0
               Morning      20680
               Afternoon    31635
               Evening      26467
cash           Night            0
               Morning      30622
               Afternoon    32803
               Evening      21062
```

7. **Build a logistic regression model to predict the probability of a customer choosing self-service based on factors such as time of day, day of the week, basket size, and transaction value. Which factors 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?**

A. I have used a logistic regression model to predict the probability of a customer choosing self-service based on the time of day, day of the week,

basket size, and transaction value. Firstly, the data is loaded into the code as 'data'. Then, a new target variable 'SelfService' is created which is set to 1 if the WorkstationGroupID is 8 (which represents self-service checkout), and 0 otherwise. The features are created by extracting the hour of the day, day of the week, and transaction values and are stored in the columns 'TimeOfDay', 'DayOfWeek', 'BasketSize', and 'TransactionValue' respectively. The dataset is then split into training and testing sets using the train_test_split method from scikit-learn. The training data is used to fit the logistic regression model and the model is evaluated on the test data.
The accuracy is **0.8727**, which indicates that the model is performing well. This means that the model can predict the probability of a customer choosing self-service with **87.28%** accuracy.
Next, a new column 'Hour' is created in the 'data' dataframe which contains the hour of the day when the transaction was made. The 'peak_hours' dataframe is created as a subset of the 'data' dataframe which contains only transactions made during peak hours (4 PM to 7 PM) while the 'regular_hours' dataframe contains transactions made during all other times. The percentage of transactions made using self-service checkouts during peak hours and regular hours are then calculated by counting the number of transactions made using self-service checkouts and dividing it by the total number of transactions made during that time.



```
Percentage of self-service transactions during peak hours: 25.51%
Percentage of self-service transactions during regular hours: 24.16%
```

The output also shows that consumers prefer using self-service checkouts during peak hours compared to regular hours. During peak hours, **25.51%** of transactions were made using self-service checkouts, while during regular hours, only **24.16%** of transactions were made using self-service checkouts.