Transaction Time Optimization

Tanmayee Anguloori
Submitted on May 1st, 2023

Introduction

This analysis will explore the relationship between various variables at a and the time of a transaction at a supermarket. In their article, "Point of Sale (POS) Data from a Supermarket: Transactions and Cashier Operations', Antczak and Weron describe a dataset containing transaction and operational data of a supermarket in Poland. The purpose of this analysis is to find which variables have the most significant impact on transaction time. This analysis will attempt to pinpoint the most important variables that influence transaction times. The chosen model will be a regression model to help predict the transaction times based on basket size (number of articles in a transaction), payment method (cash vs. card), transaction value (amount), break time, and checkout type (self service vs. general checkout). The hypothesis is that all variables will have a significant relationship on transaction times based on analysis shoen in the Exploratory Data Analysis section of this report. The results of this analysis will help supermarket managers not just localized to Poland but worldwide to understand any areas of improvement within each variable to reduce transaction times.

Note: This proposal is submitted along with an .ipynb file named "tanguloo_Transaction Time Optimization" which contains all R code used to create visualizations and the final model mentioned in this report.

Exploratory Data Analysis- Variable descriptions and Data Wrangling

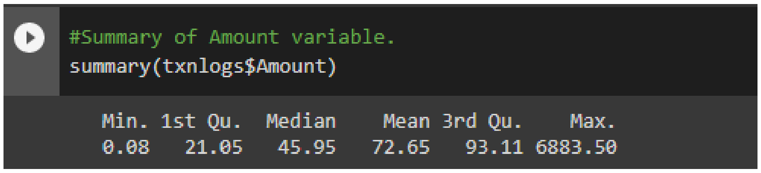
1. *BeginDateTime and EndDate Time*

The day attributes such as dayofweek and timeofDay were generated using the BeginDateTime variable. In addition, time attributes such as hours, minutes and seconds were derived from and for the BeginDateTime and EndDateTime attributes respectively. The code for this is provided in the first few lines in the Google Colab notebook.

2. *Amount*

Figure 1 shows the descriptive statistics of the Amount variable. As with basket size, we see high values running up to a transaction value of $6883.5. Figure 3 shows that these high values may make sense given that a high percentage of higher amounts were observed in December

2017. This makes sense as it is the holiday season and shoppers tend to make high value purchases during this time. As a result, these high values were not dropped from the dataset. The histogram also shows that a higher amount of transactions fall towards the lower end of the histograms confirming that there is nothing unexpected in the transaction values outside of the high values observed in the December time frame. It is interesting to observe however that there may also have been an event in February 2019 which resulted in 25% of high value transactions to be processed (Figure 3).



```
#Summary of Amount variable.
summary(txnlogs$Amount)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.08   21.05   45.95   72.65   93.11 6883.50
```

Fig 1. Descriptive statistics of Amount variable

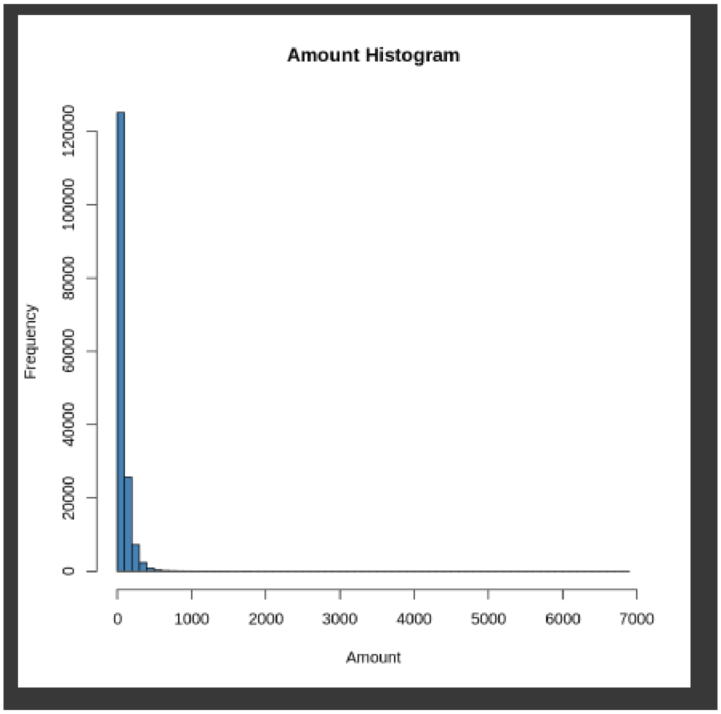
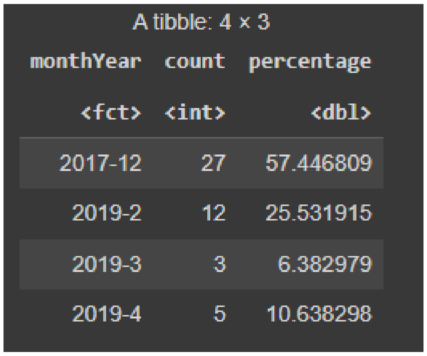


Fig 2. Amount variable’s distribution



A tibble: 4 × 3

monthYear	count	percentage
<fct>	<int>	<dbl>
2017-12	27	57.446809
2019-2	12	25.531915
2019-3	3	6.382979
2019-4	5	10.638298

Fig 3. 57% of transactions more than $1000 prevalent in December 2017.

3. *Transaction time*

The descriptive statistics for transaction time variables before any cleaning followed by its histogram are displayed in Fig 4. One can observe that the distribution is skewed to the right and since the linear regression model will be using transaction time as a dependent variable a log transformation was chosen to normalize the relationship between the dependent and independent variables for the regression model. It is important to note that after the log transformation was performed, the histogram still showed a slightly rightly skewed distribution with the mean and median having a very slight difference which was accepted and interpreted as the distribution being approximately normal (Fig 5).

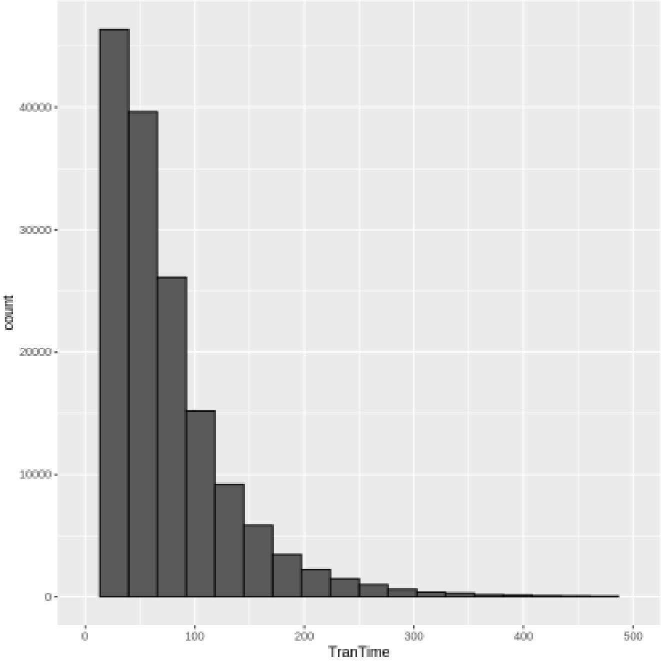


Fig 4. Distribution of the Transaction time variable before log transformation

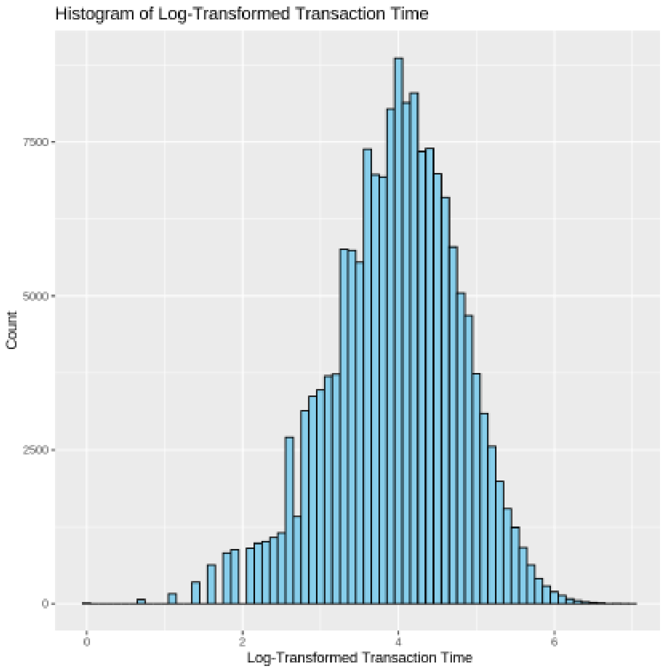
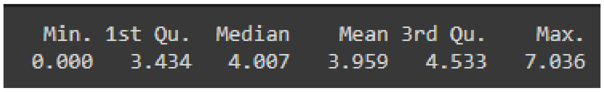


Fig 5. Distribution of Transaction time variable after the log transformation.

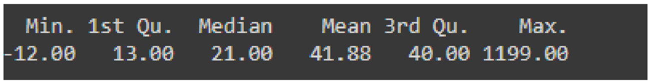


```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   3.434   4.007   3.959   4.533   7.036 
```

Fig 6. Descriptive statistics of Transaction time variable after the log transformation. Though Fig 5 shows a slight skew, this transformation was accepted as being normal due to mean and median being approximately equal.

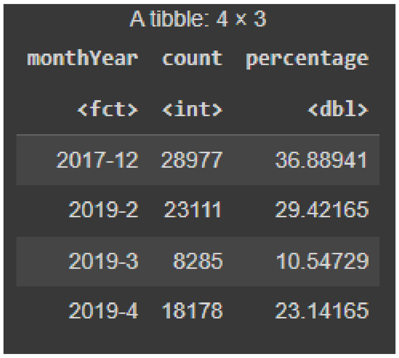
4. Break time

There were some negative breaktime values in the dataset which were removed as can be observed in Figure 7. Also, the max values of 1199 seconds (~19 mins) was further analyzed to find months where breaktimes were high. December of 2017 was found to have highest break times which could be caused by high customer traffic due to it being the holiday season followed by February 2019 (Figure 8). The frequency chart was filtered to show break times above the median value of 21 secs. Transactions can be more complex as more items are discounted or special items are offered during the holiday season. Also, staffing levels may have been inadequate to support high foot traffic. A line graph in Figure 9 shows a day-by-day breakdown of break times in December 2017. One can observe higher break times are seen in the latter part of December coinciding with Christmas and New Years Eve. In February 2019, higher break times are seen increasing towards the end of the week.



```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-12.00   13.00   21.00   41.88   40.00 1199.00
```

Fig 7. Descriptive statistics of Break Time variable



A tibble: 4 × 3

monthYear <fct>	count <int>	percentage <dbl>
2017-12	28977	36.88941
2019-2	23111	29.42165
2019-3	8285	10.54729
2019-4	18178	23.14165

Fig 8. High break times found to be more prevalent in December 2017 and Feb 2019.

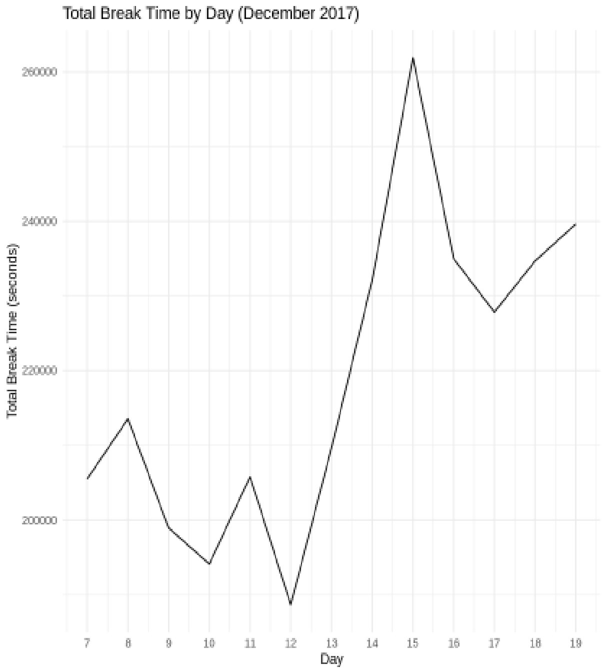


Fig 9. Total break time broken up by day view of Dec 2017

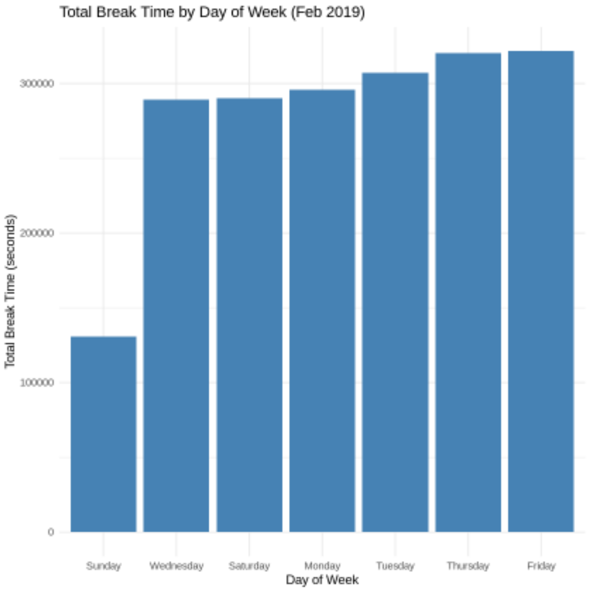
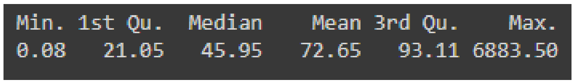


Fig 10. Total break time broken up by day of week for Feb 2019

5. *Amount*

Figure shows the descriptive statistics of the Amount variable. As with basket size, we see high values running up to a transaction value of $6883.5. As observed in the break time variable, the frequency chart in Figure 13 shows high amount values may be correlated with holiday season and an important event in Poland. As a result, these high values were not removed from the analysis. Note that the values for the frequency chart were filtered to all values above 45.95 which is the median value.



```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.08   21.05   45.95   72.65   93.11 6883.50
```

Fig 11. Descriptive statistics of Amount variable

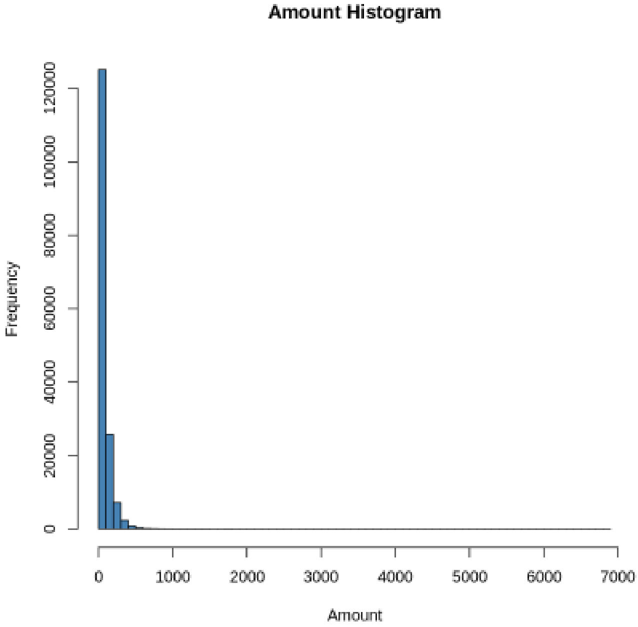
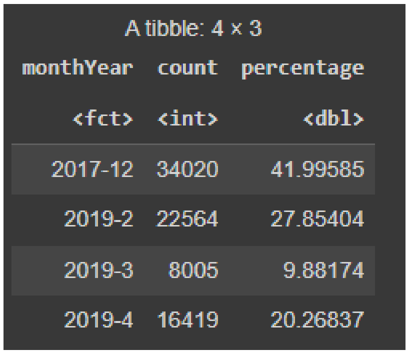


Fig 12. Distribution of the amount variable



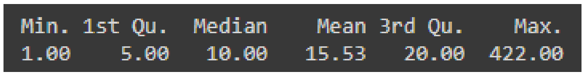
A tibble: 4 × 3

monthYear	count	percentage
<fct>	<int>	<dbl>
2017-12	34020	41.99585
2019-2	22564	27.85404
2019-3	8005	9.88174
2019-4	16419	20.26837

Fig 13. Frequency chart of amounts larger than zł45.95.

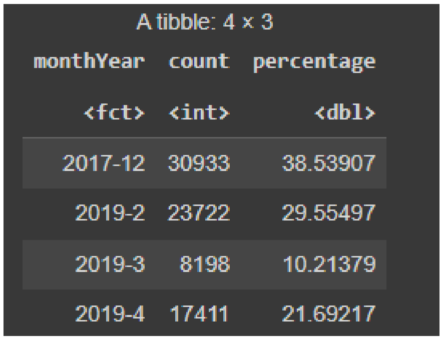
6. *ArtNum (number of articles per transaction or article size)*

Figure 14 shows the descriptive statistics of the basket size variable. Due to the maximum value of 500 articles being high especially for a supermarket, a frequency chart choosing all values over the median of 10 items was created against the month variable. As with both break time and amount variables, December 2017 and February 2019 show especially basket sizes.



```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.00    5.00   10.00   15.53   20.00  422.00
```

Fig 14. Descriptive statistics of the ArtNum variable



A tibble: 4 × 3

monthYear <fct>	count <int>	percentage <dbl>
2017-12	30933	38.53907
2019-2	23722	29.55497
2019-3	8198	10.21379
2019-4	17411	21.69217

Fig 15. Frequency chart of high basket sizes by months.

7. *Break Times* x *Transaction Times* x *Amount* x *Basket Size*

There were similar trends observed for Break Time, Transaction Time, and Amount variables. All three variables show higher values in December of 2017, February of 2019, and April of 2019. This warranted further analysis in the following figures to see if any of the continuous variables showed any visual relationship with each other. Plotting transaction time against break time, one can see a positive visual correlation (Fig 16). Plotting amount against transaction time did not show any strong visual correlation (Fig 18) whereas basket size and transaction time showed a strong positive visual correlation (Fig 17). Plotting break time against basket size showed a moderate negative visual correlation whereas plotted against amount it showed a weak negative correlation(Fig 19 – 20). As a result, it was decided that it was worthwhile to bring in break time and amount into the linear regression model to understand the impact they have on transaction times.

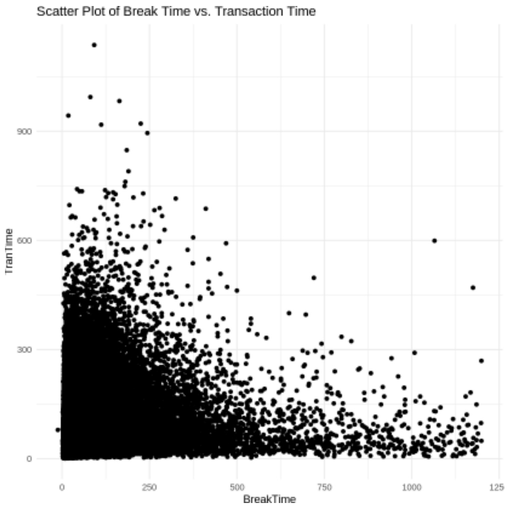


Fig 16: Break time vs Transaction Time

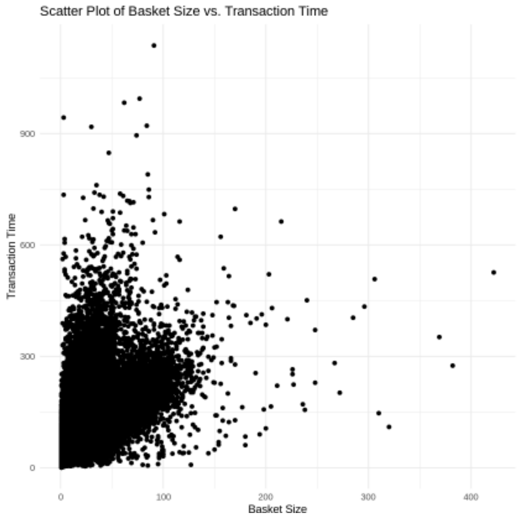


Fig 17: Basket Size vs Transaction Time

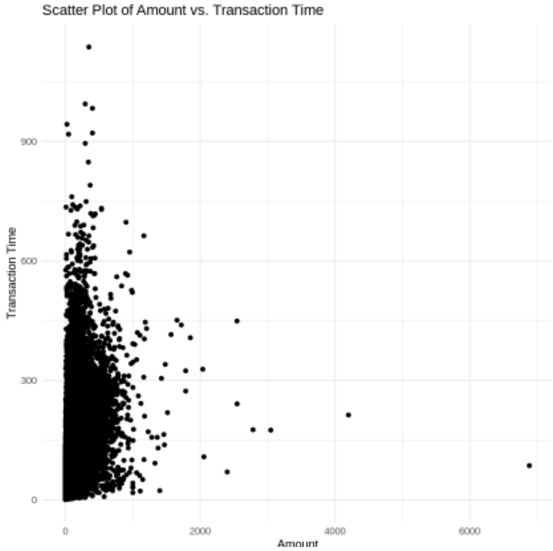


Fig 18: Amount vs Transaction Time

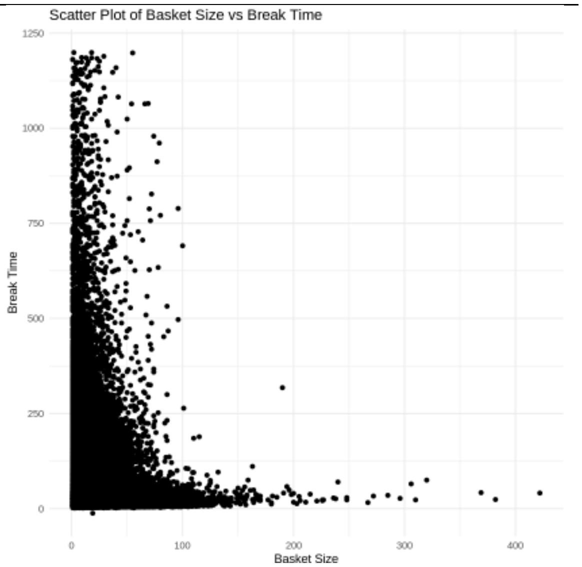


Fig 19: Basket Size vs Break Time

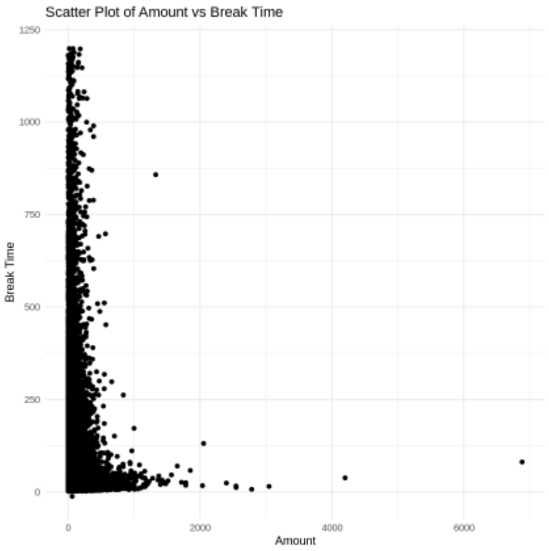


Fig 20: Amount vs. Break Time

8. *Payment Type and Checkout type*

A new payment type variable was created which was derived from and factorized the TNcard and TNcash variables into their respective payment type labels of “Cash”, “Card” and “Both”. There were values where the payment type which were neither cash nor card, and those were removed from the Payment Type variable and thus the dataset. checkoutType was derived from the WorkstationGroupID variable and was relabeled and factorized to be more visualization friendly.

Looking at these two variables against transaction count tells us a lot of about customer preferences at least for the periods analyzed in this report. Full service and cash were preferred by most customers (Figure 21 and 22). Though cash was the most popular payment type, in December 2017 card was more popular (Figure 22 and 23). Perhaps it could be theorized that during the holiday season, customers are more inclined to use cards because they make higher value and more frequent purchases. Though transaction counts were high in December 2017, most customers still preferred full-service stations (Figure 23). Afternoons had the greatest number of transactions and again, cash and full-service stations were preferred by customers (Figure 24). Card was the most popular option at self service except for the mornings where cash was preferred by customers to process transactions at self service stations (Figure 23 and 24).

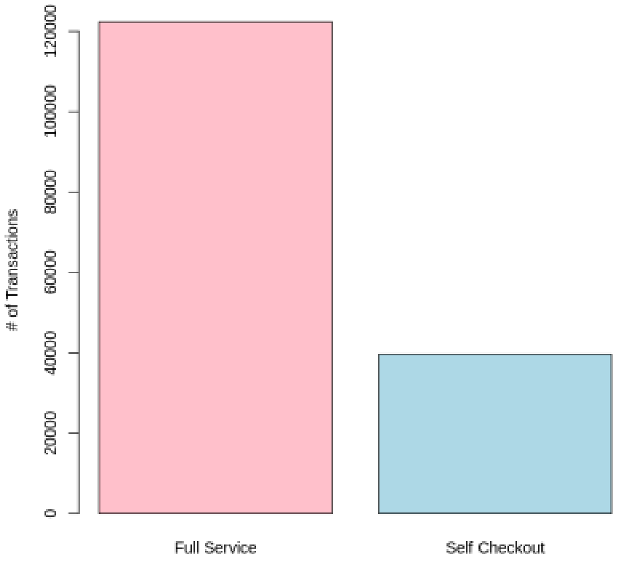


Fig 21: Transaction count by full service vs self checkout

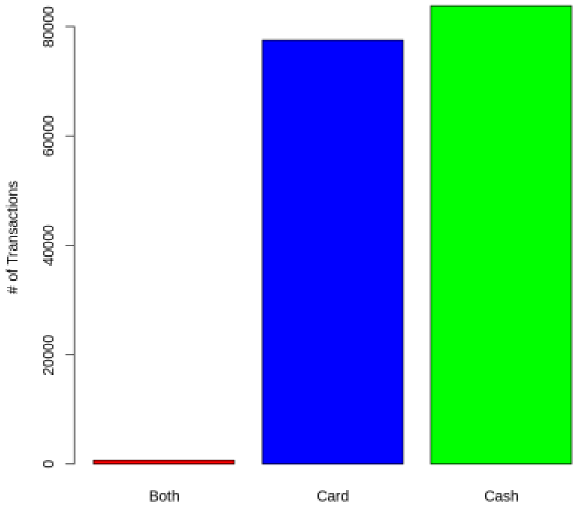


Fig 22: Transaction count by payment type

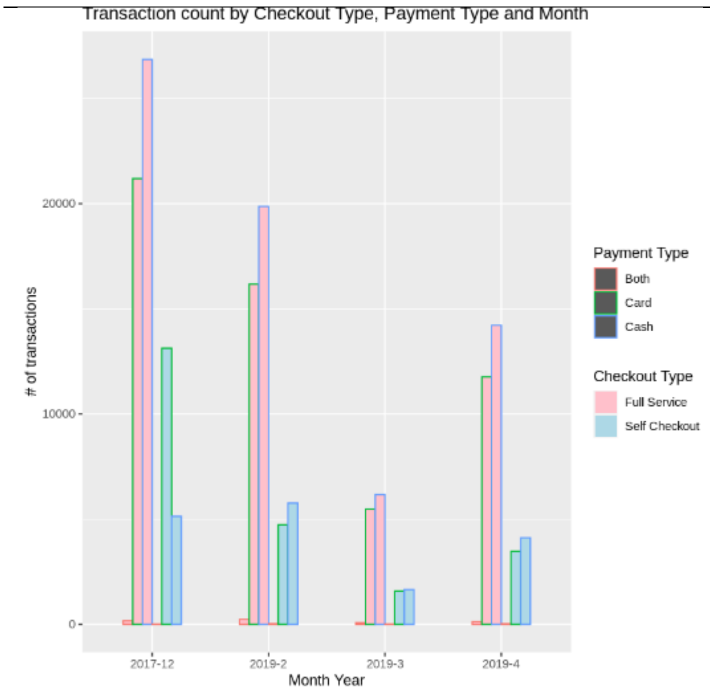


Fig 23: Transaction count by payment type, checkout type categorized by month

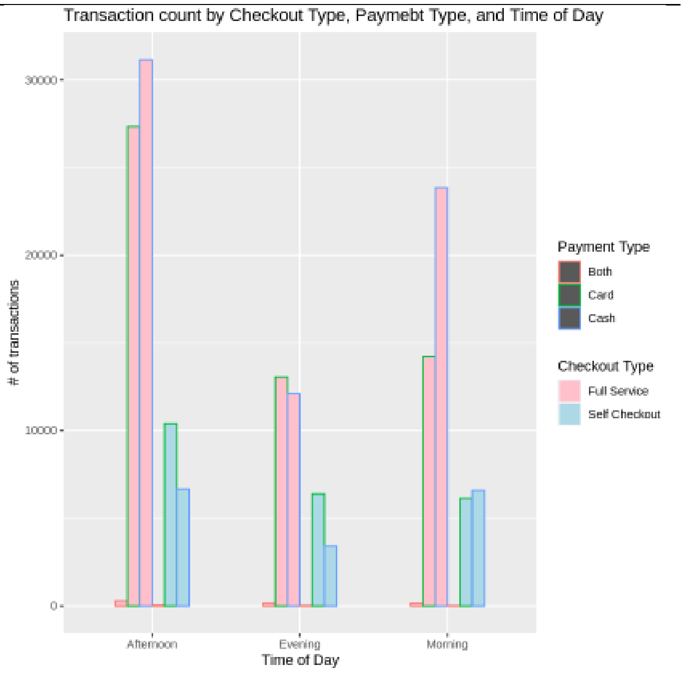


Fig 24: Transaction count by payment type, checkout type categorized by time of day.

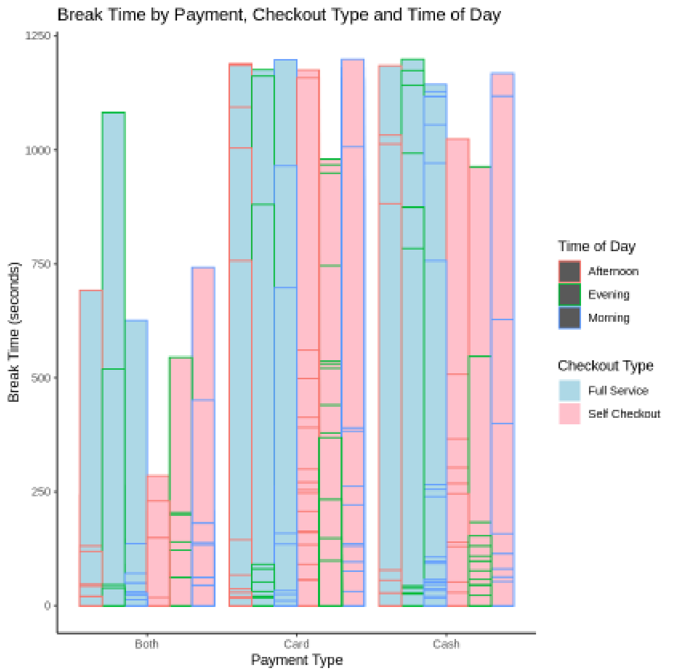


Fig 25: Break times by payment type, checkout type, and time of day

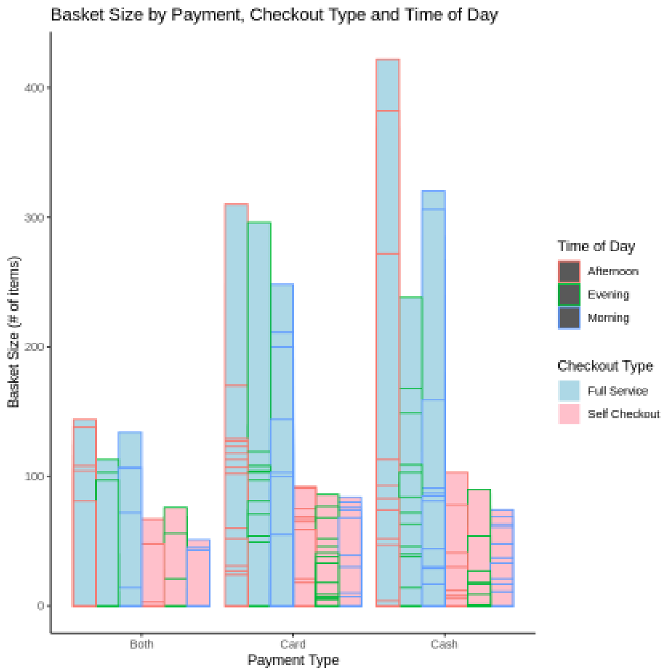


Fig 26: Basket size by payment type, checkout type, and time of day

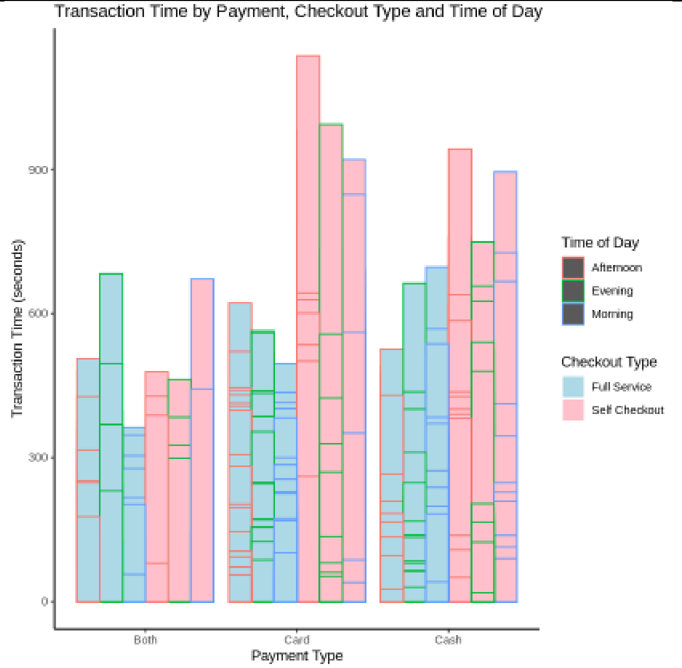


Fig 27: Transaction times by payment type, checkout type, and time of day

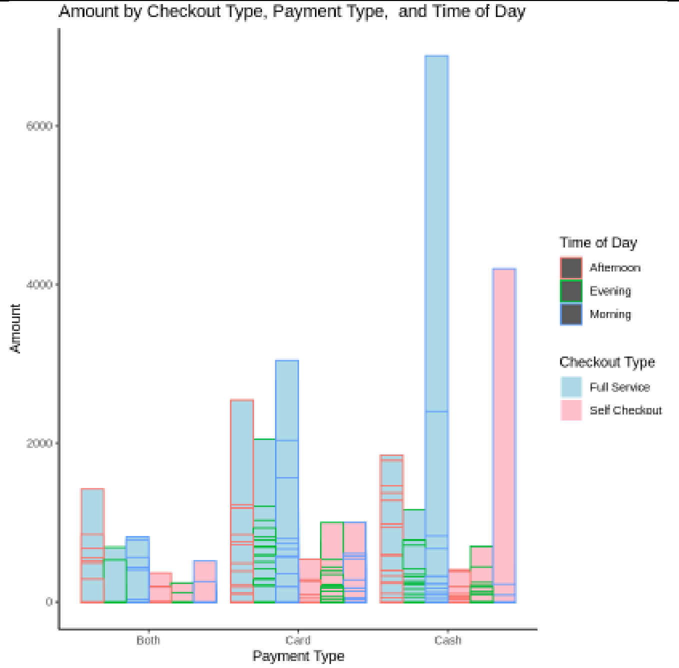


Fig 28: Amount by payment type, checkout type, and time of day

Break times were highest for full service across all payment types with high break times in morning and afternoons (Fig 25). Transaction times were highest at self service stations where cards were used and mostly during afternoons (Fig 27). Full service stations processed higher cash transaction amounts especially seen during mornings (Fig 28). Similarly larger basket sizes were processed at full service stations using cash but moreso in afternoons (Fig 26).

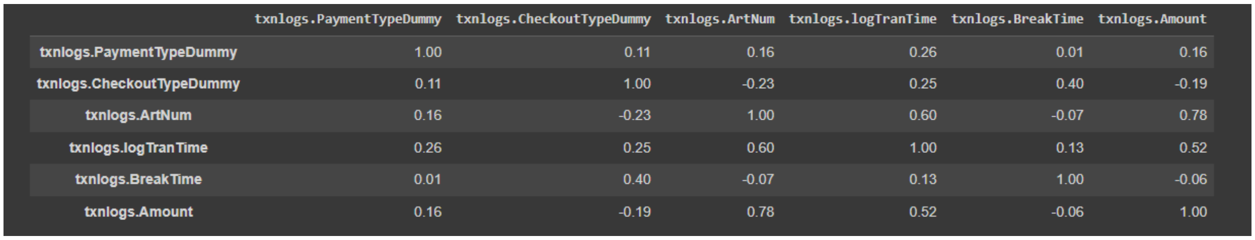
Regression analysis

Dummy variables were created for checkout type and payment type. A correlation matrix shown below identifies a strong relationship between:

- Amount and transaction time (positive)
- Amount and basket size (positive)
- Basket size and transaction time (positive)
- Break time and checkout type (positive)

The correlation matrix shows a moderate relationship between:

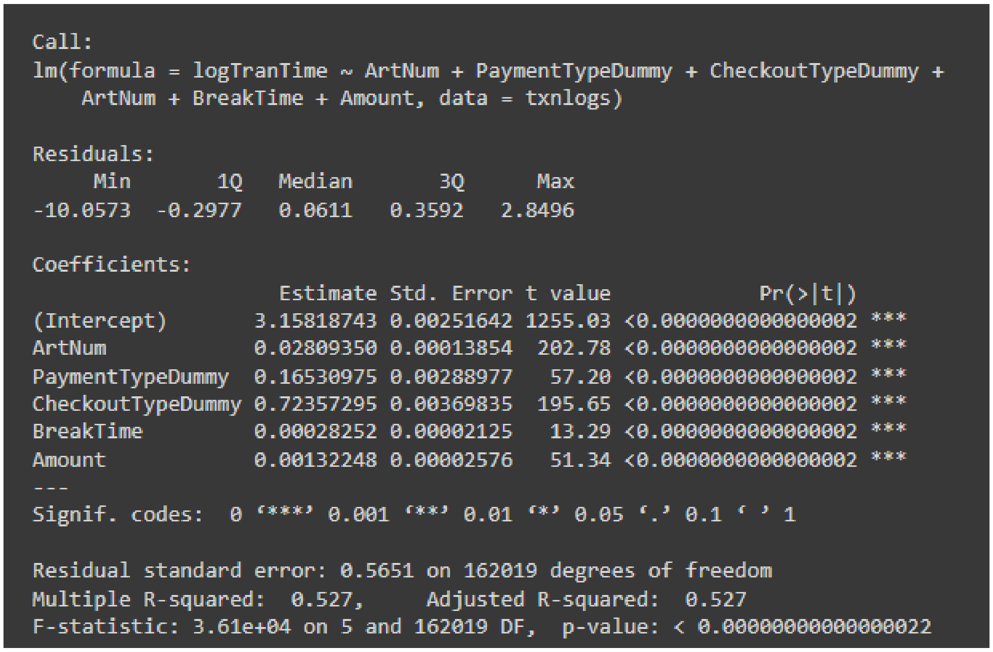
- Transaction time and payment type (positive)
- Transaction time and checkout type (positive)
- Checkout type and basket size (negative)



	txnlogs.PaymentTypeDummy	txnlogs.CheckoutTypeDummy	txnlogs.ArtNum	txnlogs.logTranTime	txnlogs.BreakTime	txnlogs.Amount
txnlogs.PaymentTypeDummy	1.00	0.11	0.16	0.26	0.01	0.16
txnlogs.CheckoutTypeDummy	0.11	1.00	-0.23	0.25	0.40	-0.19
txnlogs.ArtNum	0.16	-0.23	1.00	0.60	-0.07	0.78
txnlogs.logTranTime	0.26	0.25	0.60	1.00	0.13	0.52
txnlogs.BreakTime	0.01	0.40	-0.07	0.13	1.00	-0.06
txnlogs.Amount	0.16	-0.19	0.78	0.52	-0.06	1.00

Fig 29: Correlation matrix with variables used in linear regression model

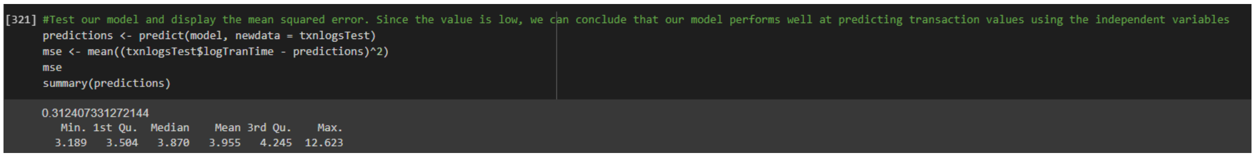
Next, before splitting the dataset into train and test rows, a linear regression was performed on the original dataset with the below results. Residuals still support a skew to the data but each variable is highly significant. Additionally, a score of ~53% tells us that a good chunk of variability can be explained with these variables.



```
Call:
lm(formula = logTranTime ~ ArtNum + PaymentTypeDummy + CheckoutTypeDummy +
    ArtNum + BreakTime + Amount, data = txnlogs)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.0573  -0.2977   0.0611   0.3592   2.8496 

Coefficients:
                    Estimate Std. Error t value            Pr(>|t|)    
(Intercept)       3.15818743 0.00251642 1255.03 <0.0000000000000002 ***
ArtNum            0.02809350 0.00013854  202.78 <0.0000000000000002 ***
PaymentTypeDummy  0.16530975 0.00288977   57.20 <0.0000000000000002 ***
CheckoutTypeDummy 0.72357295 0.00369835  195.65 <0.0000000000000002 ***
BreakTime         0.00028252 0.00002125   13.29 <0.0000000000000002 ***
Amount            0.00132248 0.00002576   51.34 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5651 on 162019 degrees of freedom
Multiple R-squared:  0.527,	Adjusted R-squared:  0.527 
F-statistic: 3.61e+04 on 5 and 162019 DF,  p-value: < 0.00000000000000022
```

Fig 30: Linear regression summary output without training the linear model just yet.

Finally, the dataset was split into train and test rows. The training model was used to produce prediction values using the test dataset. Below figure shows that the MSE of the trained model which predicts transaction times based on the chosen variables is low which means that majority of predicted values are very close to the actual transaction times. The following scatter plot also shows that most points crowd around the fitted line showing a good fit with stray points mostly observed when higher transaction times are trying to be predicted.



```
[321] #Test our model and display the mean squared error. Since the value is low, we can conclude that our model performs well at predicting transaction values using the independent variables
      predictions <- predict(model, newdata = txnlogsTest)
      mse <- mean((txnlogsTest$logTranTime - predictions)^2)
      mse
      summary(predictions)

      0.312407331272144
         Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
        3.189   3.504   3.870   3.955   4.245  12.623 
```

Fig 31: MSE of test model

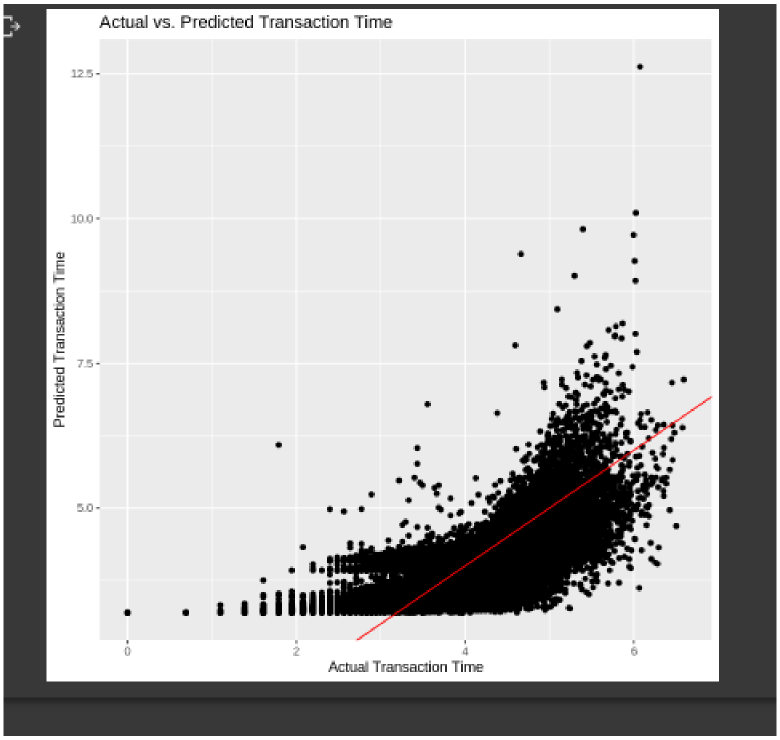


Fig 32: Scatter plot of actual vs predicted transaction times using test data

The choice to not exclude higher transaction time values from the dataset may have contributed to the outliers seen in the scatter plot. As a future analysis, it may be worthwhile to examine if excluding transaction times beyond a certain threshold can bring down the MSE even further and decrease the number of outliers seen on the scatter plot.

Discussion & Limitations

It was rewarding to see that all variables chosen for the analysis were significant in the model. However, for a future iteration of this analysis it would be worth using a backward stepwise regression to see if limiting the number of variables would produce a better MSE and R-squared value. Another limitation found was the outliers in the actual vs predicted scatter plot (Fig 32). Though the transaction time variable was log transformed, it may be worthwhile to remove outliers before transforming the variable to see if the plot produces less stray points. The same would apply with all other high values found in the basket size and amount variables. Though these high values appeared as a pattern in certain time periods, removing unreasonably high values may improve model performance. Lastly, the exploratory analysis section of this paper showed some interesting insights into the relationship between time of day, checkout, and payment type variables. A future iteration of this analysis could include a logistic regression model to see if time of day and day of week influenced a customer's probability around preferring certain checkout and payment types.

Conclusion

This analysis is helpful to supermarket managers as there was found to be a significant relationship between basket size, amount, break time, payment type and checkout type variables and transaction times. In addition to the model showing that checkout type and payment type variables were significant predictors of transaction times, the correlation matrix spoke to a moderate relationship existing between checkout type and transaction time and between payment type and transaction time. Furthermore, looking at Figures 25 and 27, one can see that break times are highest for full service and transaction times are highest for self service. Perhaps, managers should consider bulking up their staffing levels during times of high anticipated store traffic (such as holiday seasons and certain times of week such as afternoons and mornings reported in this analysis) and also run frequent performance tests on the self checkout machines or even increase the number of self checkout machines to meet demand. Lastly, there was also a strong positive correlation observed between amount and transaction time with customers choosing to process high value transactions through full service. An investigation can be conducted to understand why high amount transactions take longer to process. The exploratory analysis coupled with the model and correlation matrix provides many insights and opportunities for managers to drill into the categorical variables to see where they can improve transaction times.

CITATIONS

Antczak, T., & Weron, R. (2019). Point of Sale (POS) data from a supermarket: Transactions and cashier operations. *Data*, *4*(2), 67. https://doi.org/10.3390/data4020067