Final term-project

CS555- Foundation of Machine Learning

Shawana Adbiah Raka

Research Question:

For the analysis, I have selected a dataset from the MBTA website. The dataset contains information about rail ridership for the Massachusetts Bay Transportation Authority (MBTA) categorized by season, time period, route line, and stop.

The question I want to answer through this analysis is-
"What factors significantly influence the total ridership (total_ons) at different stops within the MBTA transit system?"

This question aims to identify and quantify the impact of various factors such as time of day (time_period_name), route characteristics (route_id), and possibly other variables like average_flow on the number of passengers boarding (total ons) at various MBTA stops. The goal is to determine which variables are significant predictors of ridership, which can help in planning, optimizing routes, and potentially improving service efficiency.

The Dataset:

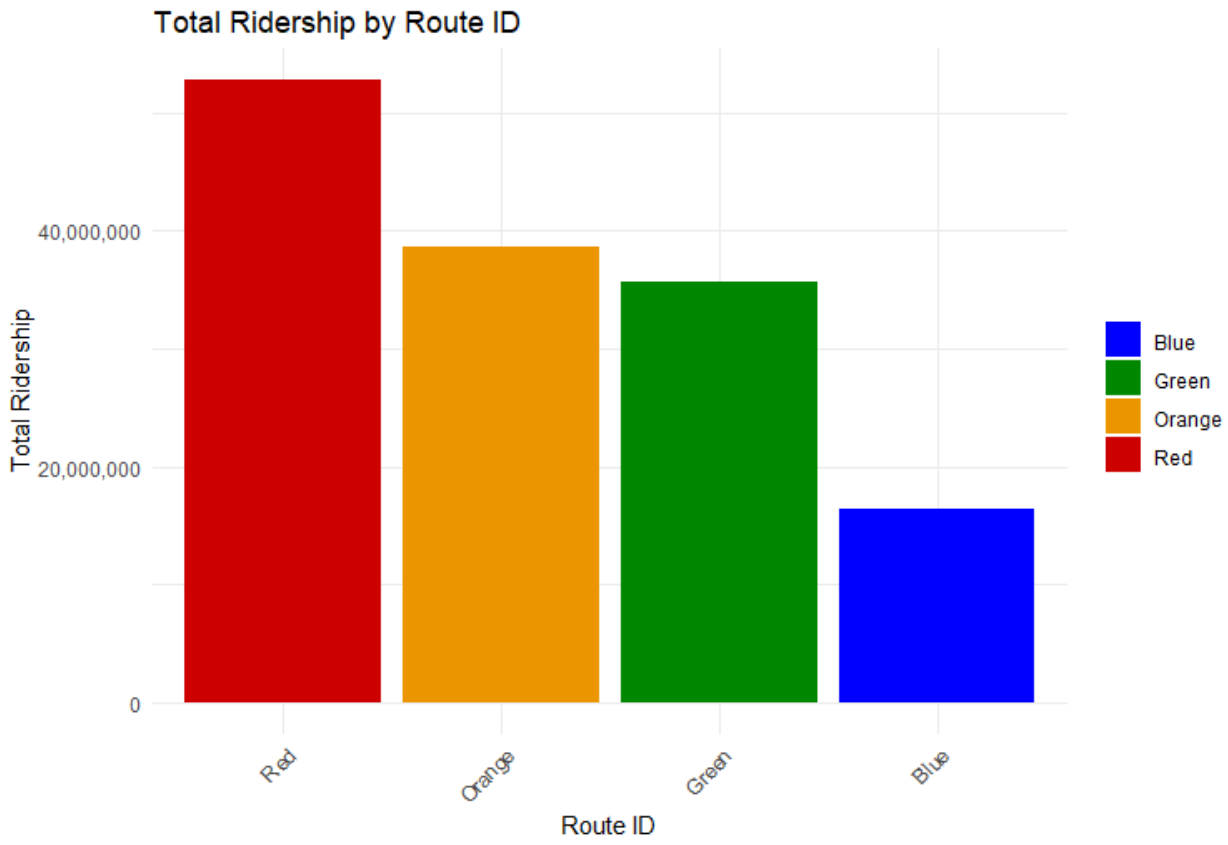
It has 7920 rows and 18 variables and has data from 2017 to 2019. It only contains data for the fall season. It does not have any missing values. Here are the columns included in the dataset:

mode: Numerical identifier for the mode of transportation
season: Season and year during which the data was collected (e.g., Fall 2019).
route_id, route_name: Identifier and name of the route.
direction_id: Numerical identifier for the direction of travel.
day_type_id, day_type_name: Identifier and name for the type of day (e.g., weekday, weekend).
time_period_id, time_period_name: Identifier and descriptive name for the time of day (e.g., VERY_EARLY_MORNING).
stop_name, stop_id: Name and identifier of the stop.
total_ons, total_offs: Total number of passengers getting on and off at each stop.
number_service_days: Number of days the service was operational during the data collection period.
average_ons, average_offs: Average number of ons and offs at each stop per day.
average_flow: Possibly the average total flow of passengers (both ons and offs) at each stop.
ObjectId: Unique identifier for each record.

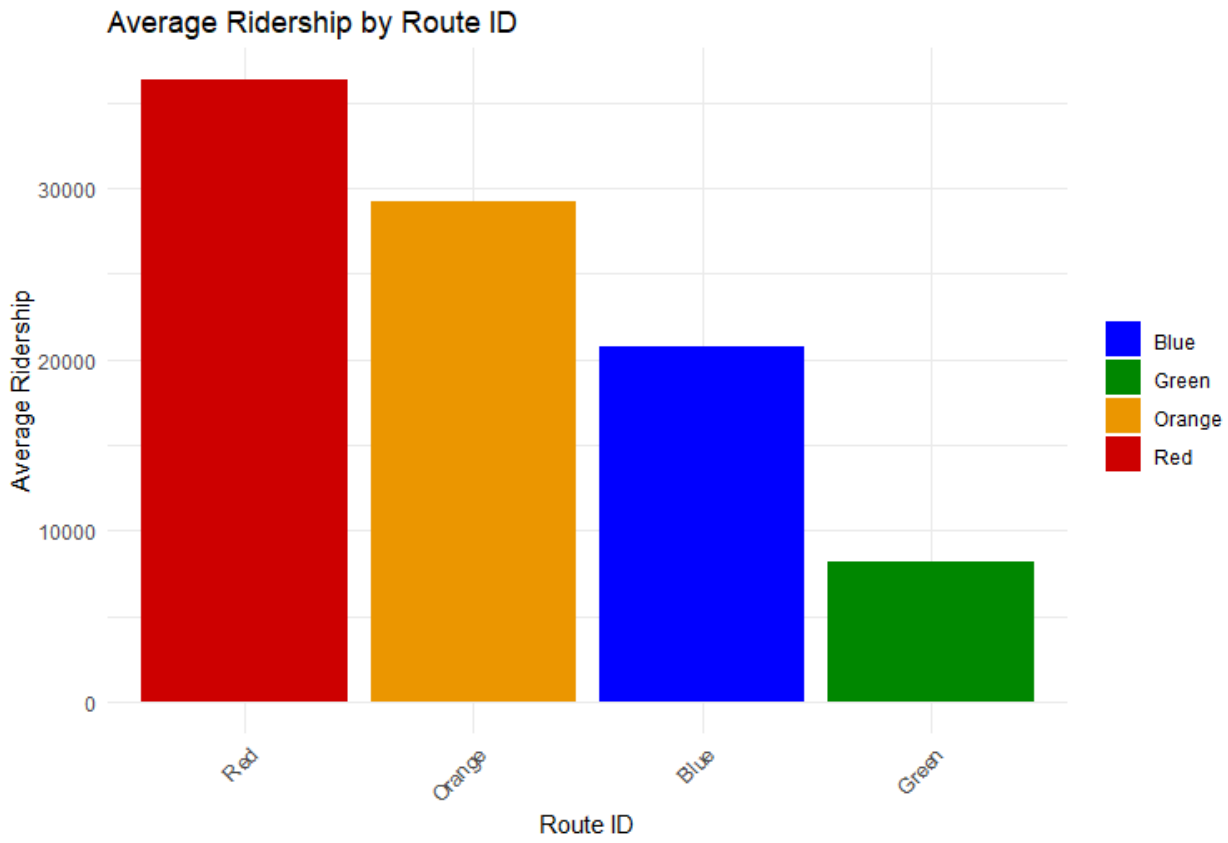
After looking at the variables and the dataset in R, I have decided to remove some variables that do not add much to the analysis.
The variables I am removing are -
Mode, route_name, day_type_id, time_period_id, stop_id, and ObjectId. These variables are mostly id columns.

Initial Look at the Data:

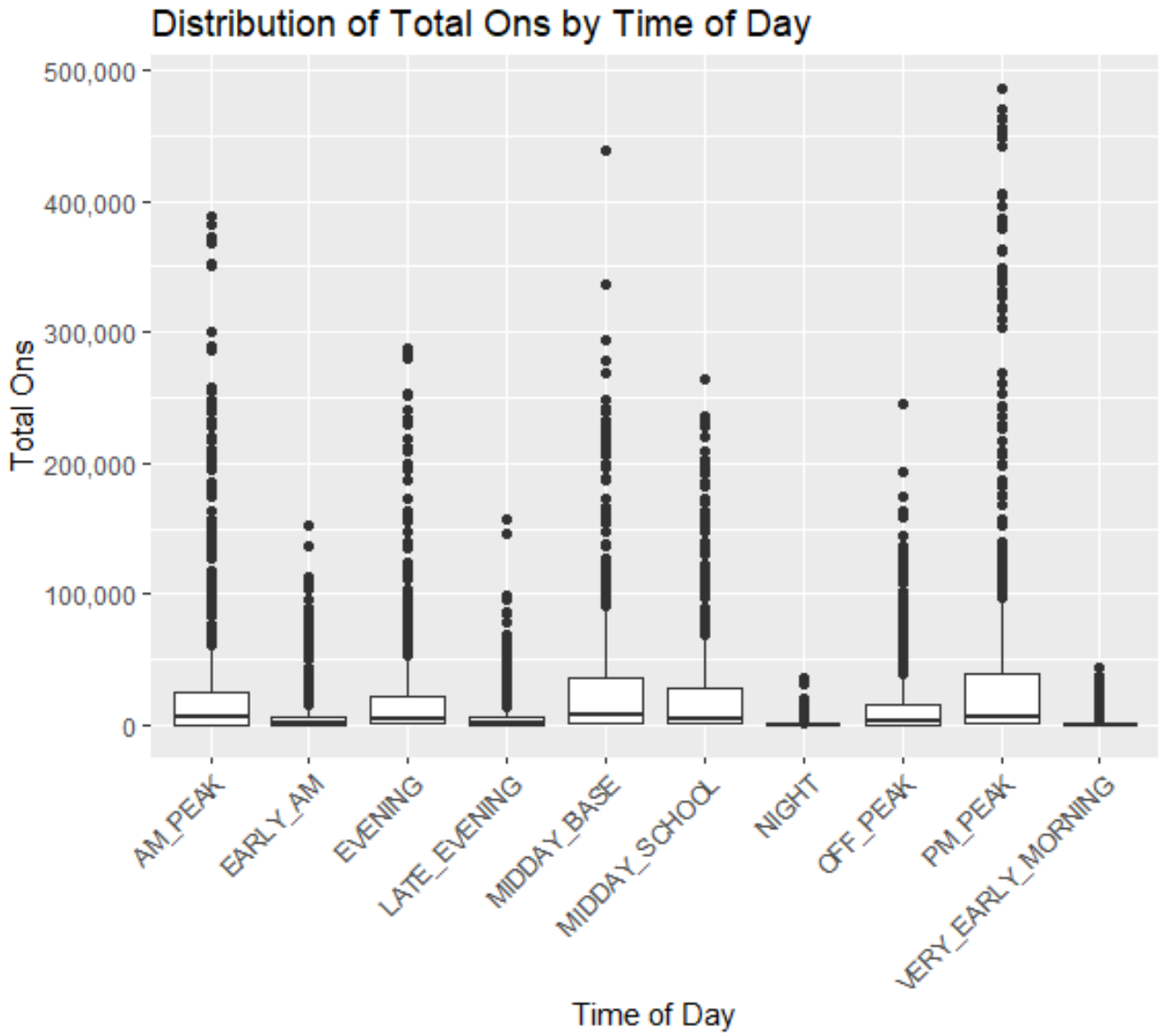
I tried to figure out if based on the routes, the ridership is different or not. The below plot shows that based on the different routes, the total number of passengers is different.



If I take the average number of riders, then we can also see that the total number of riders per route or line is different. However, the average number of riders on the green line is less than the blue line where the total number of riders is larger for the green line. This might be because of the difference in service days or the number of days the route was active. The below plot shows the average number of riders per route.


Average Ridership by Route ID
Average Ridership
30000
20000
10000
0
Red
Orange
Blue
Green
Route ID
Blue
Green
Orange
Red

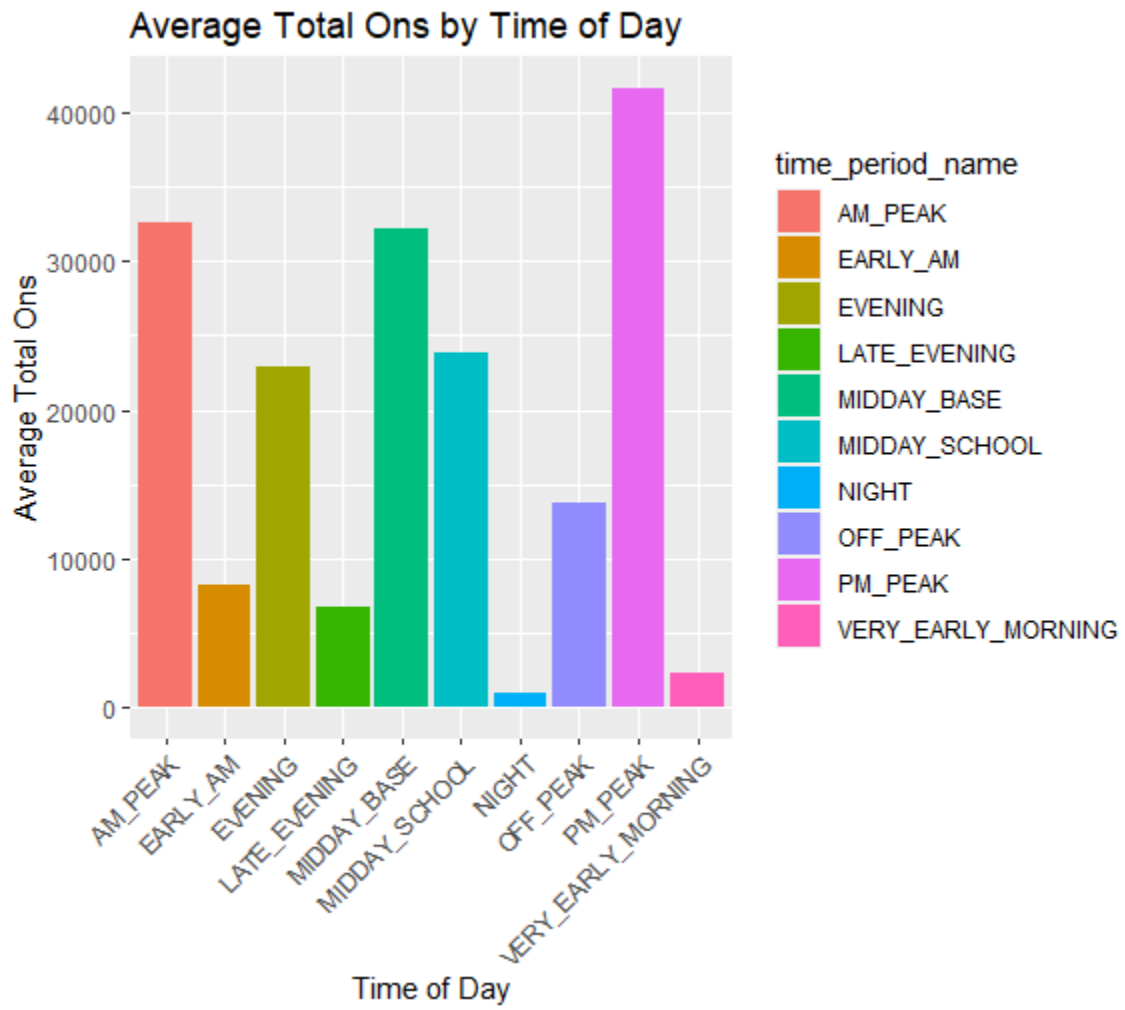
Box Plot



The box plot indicates the distribution of total boardings (total_ons) for different times of the day. The central line in each box represents the median of the distribution, the box itself covers the interquartile range (IQR, from the 25th to the 75th percentile), and the "whiskers" extend to cover the bulk of the data, typically 1.5 times the IQR. Points beyond the whiskers are considered outliers.

From this plot, we can see:
- There's a wide range of total boardings in the dataset, indicated by the spread of points and the length of the whiskers.
- Some time periods have a higher median boarding (e.g., `AM_PEAK`), suggesting these are busier times for the MBTA.
- There are many outliers, particularly during the `AM_PEAK` and `PM_PEAK` periods, indicating extreme values that are significantly higher than the typical range.

Bar Plot



The bar plot shows the average number of boardings (total_ons) for each time of day. This is a more direct visualization of the central tendency (mean) without considering the distribution's spread or outliers.

Observations from the bar plot include:
- AM_PEAK and PM_PEAK have the highest average boardings, which is expected as these are rush hour periods.
- The MIDDAY_BASE and NIGHT have the lowest average boardings, suggesting these are off-peak hours with less frequent ridership.

ANOVA Analysis

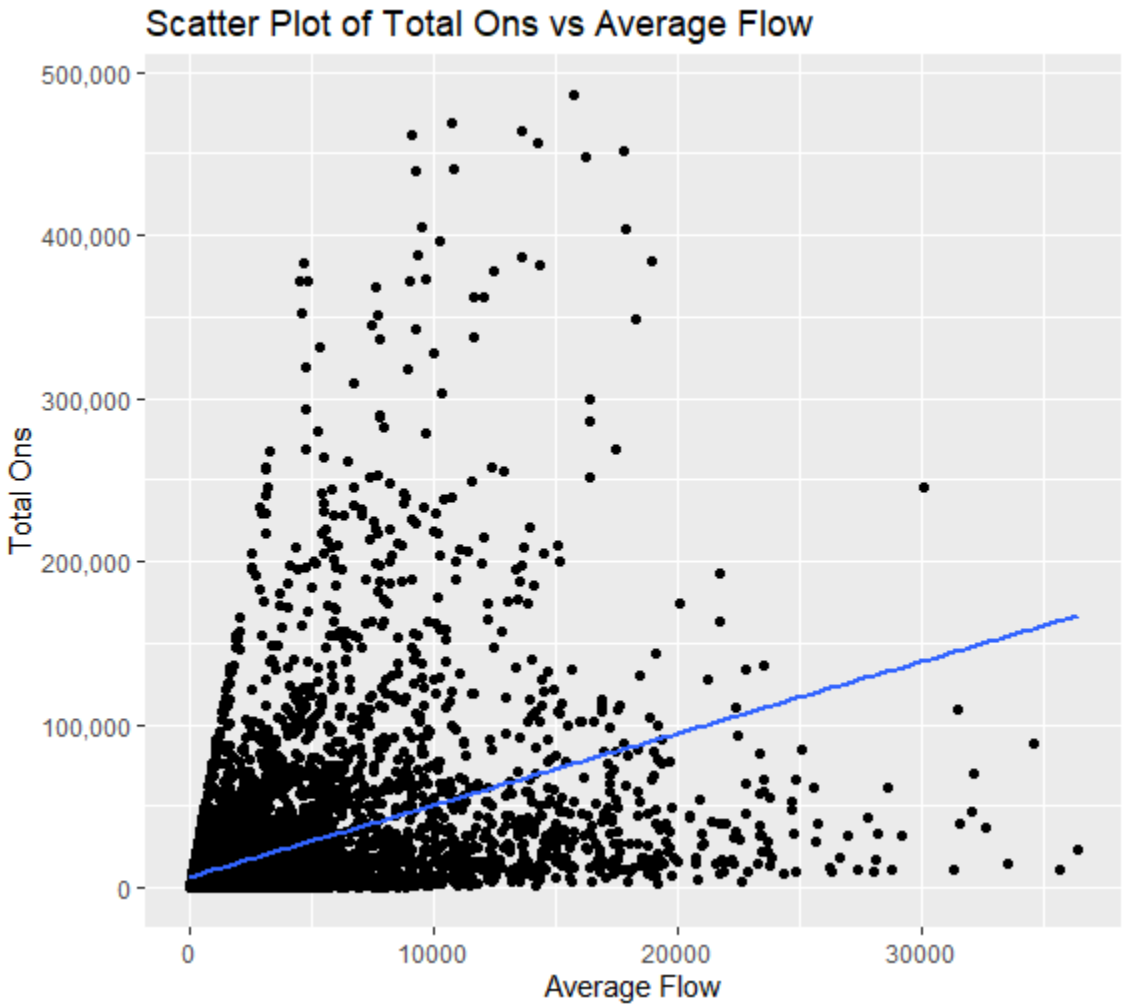
I have conducted an ANOVA analysis to statistically test whether the mean number of total_ons differs significantly between different times of the day.

The ANOVA results (figure 1) indicate that there are statistically significant differences in the number of total boardings among different times of the day (p-value < 2.2e-16). The time_period_name factor's significance means that the time of day has a strong effect on the number of people boarding MBTA trains.

Splitting The Data into Training and Test sets

For the next steps, I have divided the data into training and test sets randomly with a seed value of 555. I took 70% of the data for the training set and 30% of the data for the test set.

SLR Model



I have chosen the average flow to predict total ons in my slr model. I used the training set to build the model.

From the model summary we can see that -
Residual Standard Error (RSE): The RSE of 39,480 on 5542 degrees of freedom indicates that, on average, the actual number of total_ons can differ from the predicted number by 39,480. This is a measure of the model's prediction error.

R-squared: The R-squared value of 0.1825 indicates that about 18.25% of the variability in total_ons is explained by average_flow. This isn't a particularly high value, suggesting that while average_flow does have a significant relationship with total_ons, there's a lot of variation in total_ons that isn't explained by average_flow alone.

F-statistic: The F-statistic of 1237 and the p-value < 2e-16 indicate that the model is statistically significant. The model with average_flow as a predictor is better at explaining the variation in total_ons than a model with no predictors at all.

I used the test set to test the model-
The Root Mean Squared Error (RMSE) of 40,785.26 on the test data indicates the average deviation of the predicted total_ons from the actual total_ons. Given the magnitude of RMSE relative to the typical values of total_ons, it suggests that while my model has statistically significant predictors, the prediction error is quite substantial.

The RMSE is close to the Residual Standard Error (RSE) from the training model summary (39,480), which suggests consistency in model performance from training to testing. However, both values are high, indicating a large average prediction error.

Hypothesis Testing
I want to figure out if the model is statistically significant or not.
Step 1: The Hypotheses
Null Hypothesis (H0): There is no relationship between average_flow and total_ons.
Alternative Hypothesis (H1): There is a relationship between average_flow and total_ons.

Step 2: Significance Level (α)
The significance level is the probability of rejecting the null hypothesis when it is actually true. A common choice is $\alpha = 0.05$, which has a 5% chance of a Type I error (false positive).

Step 3: Compute the Test Statistic
The t-test statistic for the slope is computed as the estimated coefficient divided by its standard error. According to the summary, the t-value for average_flow is 35.18.

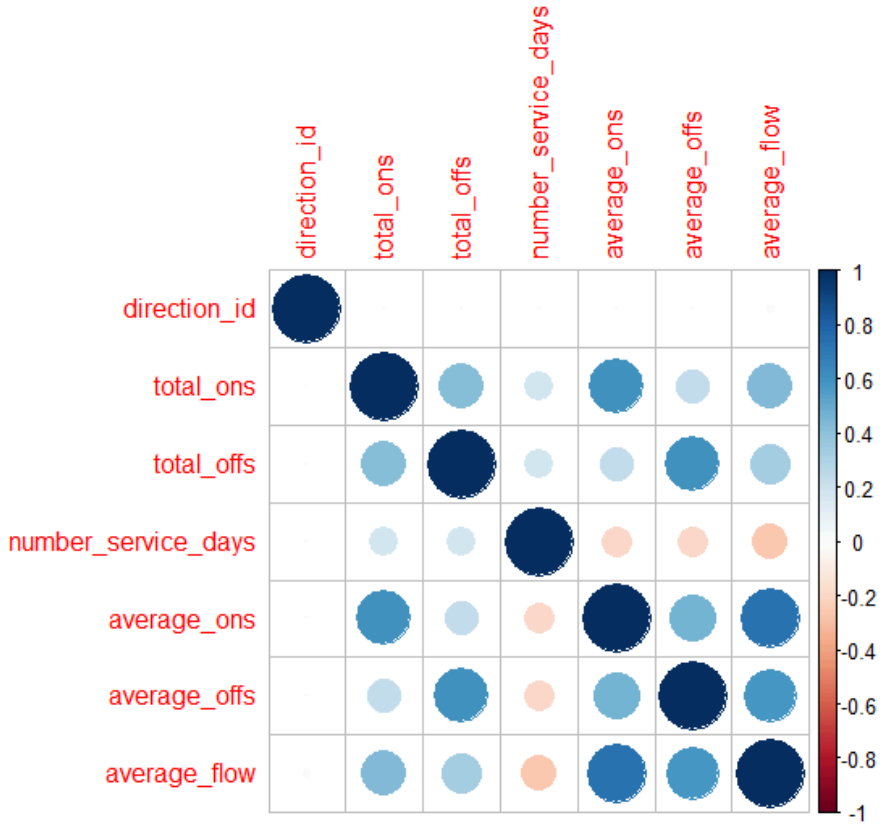
Step 4: p-value
The p-value tells us the probability of observing a test statistic as extreme as the t-test statistic under the null hypothesis. In this summary, the p-value for average_flow is < 2.2e-16, which is extremely low.

Step 5: Decision
If the p-value is less than α, we reject the null hypothesis. Given that the p-value is far below 0.05, we reject the null hypothesis and conclude that there is a statistically significant relationship between average_flow and total_ons.

Finally, the overall F-statistic tests whether at least one predictor variable in the model has a non-zero coefficient. Since the p-value for the F-statistic is also less than 0.05, we can say that the model as a whole is statistically significant.

Correlation Table

The dataset has a few numeric variables. I am building a correlation matrix to see if any of these variables are highly correlated.



From the correlation matrix we can see that none of the variables are very highly correlated. However, average_flow has a strong positive correlation with average_ons (0.7351), indicating that the average flow

of passengers is heavily influenced by the average number of boardings. Thus I am dropping the variable average_ons as it can lead to issues with multicollinearity.

MLR Model

As the SLR model was not sufficiently predicting the average_ons, I have decided to use an MLR model to predict the average_ons. I have used the backward elimination method for this model to get the best results in terms of accuracy. R has dummyfied the categorical variables by itself under the hood.

From the model summary (Figure 9) we can see that the Residual Standard Error is 27,580, suggesting that the predictions of the model typically deviate from the actual values by this amount. The R-squared value of 0.6103 means that about 61.03% of the variability in total_ons is explained by the model. This is a decent fit for real-world data and a lot better than the SLR model. The adjusted R-squared for the number of predictors in the model is 0.601, which is close to the R-squared, indicating a good model fit without being overly inflated by excessive predictors.

Interpretation of Significant Predictors:

Time Period and Day Type: Certain time periods like EARLY_AM, LATE_EVENING, and NIGHT have significant negative coefficients, suggesting lower ridership during these times. PM_PEAK has a positive coefficient, indicating higher ridership.

Seasonal Effects: The negative coefficient for seasonFall 2018 suggests a decrease in ridership during this period compared to the baseline (reference category Fall 2017).

Route and Stop Specifics: Different routes and stops, like route_idOrange and Park Street, show varied effects on ridership, indicating that some routes and stations are more popular than others.

The model's performance summary table:

	Data	RMSE	MAE	R2
1	**Training**	**27255.02**	**14970.55**	**0.6103227**
2	**Testing**	**30407.46**	**16180.72**	**0.5532455**

The performance summary table gives a clear view of how the multiple linear regression model performs on both the training and testing datasets. Here's a breakdown of what these metrics suggest about the model's performance:

Performance Metrics:
The RMSE is higher on the test set compared to the training set, indicating that the model performs slightly worse on unseen data. This could be due to model overfitting, or it might reflect the variability in the new data that wasn't captured during training.

The MAE (Mean Absolute Error) measures the average absolute difference between predicted and actual values, giving a straightforward interpretation of error magnitude without direction. Similar to RMSE, the MAE is higher on the testing data, further suggesting that predictions are less accurate when the model is applied to new data.

The R^2 (Coefficient of Determination) indicates how well the data fit the statistical model (higher values indicate a better fit). The R^2 value for the testing data is lower than for the training data, which is typical in modeling scenarios as the training data is what the model was built on. The reduction in R^2 on the test set shows that the model explains approximately 55.32% of the variability in the testing data, compared to 61.03% in the training data, suggesting some loss of explanatory power when generalized to new data.

5 Step Hypothesis Testing:

I am conducting hypothesis testing to see if the model's predictors are significant or not.

To perform a hypothesis test for the multiple linear regression (MLR) model to determine if at least one predictor is useful in predicting total_ons, I used the F-test for overall model significance. This test evaluates whether the model provides a better fit to the data than a model with no predictors other than the intercept. Below are the five steps of the hypothesis test:

Step 1: Null and Alternative Hypotheses
Null Hypothesis (H0): All regression coefficients in the model are equal to zero (except for the intercept). This suggests that none of the predictor variables have a significant effect on the response variable, total_ons.

Alternative Hypothesis (Ha): At least one regression coefficient is not zero. This implies that at least one predictor has a significant effect on total_ons.

Step 2: Choose the Significance Level (α)
The significance level (α) is commonly set at 0.05 for many tests. Thus I went with it as well. This threshold determines the probability of rejecting the null hypothesis when it is actually true (Type I error).

Step 3: Compute the Test Statistic
The F-statistic is used for this test. It compares the variance explained by the model with the unexplained variance:
This statistic is automatically calculated in the summary output of the linear model in R.

Step 4: Determine the p-value

The p-value for the F-statistic can be found in the summary of the regression model. It indicates the probability of observing the data or something more extreme if the null hypothesis is true.

Step 5: Make the Decision
Comparing the p-value to the α level:
If p-value ≤ α, reject the null hypothesis. There is sufficient evidence to suggest that at least one predictor variable significantly affects total_ons.
If p-value > α, fail to reject the null hypothesis. There is not enough evidence to suggest that any predictor significantly affects total_ons.

Based on the summary :
F-statistic: 65.22
Degrees of Freedom: Regression df = 130, Residual df = 5413
p-value: < 2.2e-16

Decision:
Given the p-value is much less than 0.05, we reject the null hypothesis. This suggests that there is very strong evidence that at least one of the variables in the model significantly affects total_ons.

This method provides a formal statistical basis for assessing the significance of the entire model and thus informs MBTA decision-making regarding factors impacting ridership.

Logistic Regression

I have also built a logistic regression model to figure out based on route_id, time_period_name, and average_flow if the total ridership is high or not.

The summary of the logistic regression model (Figure 11) and the subsequent prediction on the test data provide valuable insights into the model's predictive power and its applicability to making decisions based on ridership patterns.

From the model summary:
Coefficients: The variables like route_idOrange, route_idRed, time_period_nameNIGHT, and time_period_nameOFF_PEAK have significant coefficients with strong effects. Notably, time_period_nameOFF_PEAK has a very large negative coefficient, indicating a strong decrease in the likelihood of high ridership during off-peak times compared to the baseline period.
average_flow: The positive coefficient for average_flow indicates that as the average flow increases, so does the probability of observing high ridership. This is statistically significant and suggests that average flow is a strong predictor of high ridership.

Predictions and Model Evaluation

I calculated the accuracy using the diag() function. The confusion matrix and the calculated accuracy give a clearer picture of model performance:
Confusion Matrix:
True Negatives (0,0): 1034
False Positives (0,1): 154
False Negatives (1,0): 345
True Positives (1,1): 843
The model has an accuracy of approximately 79%, which is quite good for a logistic regression model in a practical application like this.

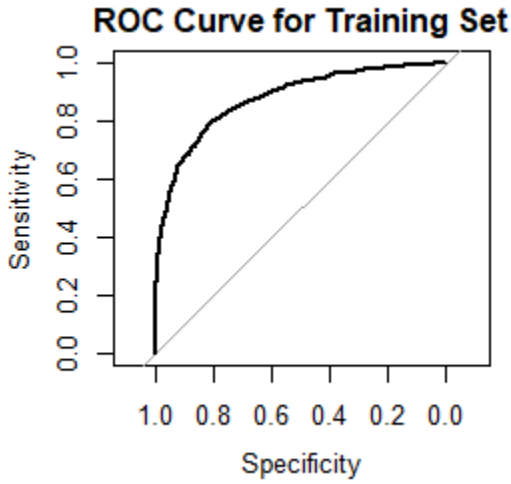
AUC and ROC Curve

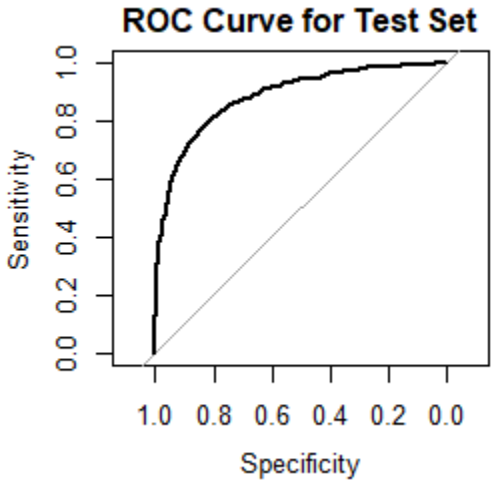
According to the code output (figure 13):
AUC for Training Set: 0.8747
AUC for Test Set: 0.8834

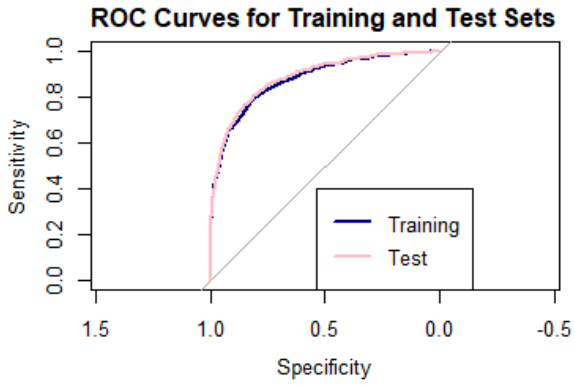
Both AUC values are well above 0.5 and close to 1.0, which suggests that the model does an excellent job at distinguishing between the two classes (high vs. low ridership). Higher AUC values indicate that the model is better at correctly classifying the binary outcome across all possible threshold values.

The AUC of 0.8747 for the training set shows strong predictive accuracy within the data used to train the model. This means the model fits well to the training data, capturing the necessary patterns to distinguish between the outcomes effectively.

An AUC of 0.8834 on the test set is even slightly higher than the training set AUC. This is a positive sign that the model not only generalizes well but perhaps even performs slightly better on unseen data, suggesting that it hasn’t overfitted to the training data.







Observations from the ROC Curves:

Training Set (Blue Curve): The curve is close to the left-hand border and the top border of the ROC space, which indicates a high true positive rate and a low false positive rate for most thresholds.

Test Set (pink Curve): The test ROC curve closely follows the training curve, suggesting that the model's predictive performance is consistent when applied to new data.

The curves suggest that for most thresholds, the model has a good balance of sensitivity and specificity. The curves do not intersect, which often suggests that there is not a point where the model suddenly performs poorly as the threshold is adjusted.

The similarity between the training and test ROC curves suggests that the model has generalized well from the training data to the test data. There is no indication of overfitting, as the test curve would typically fall significantly below the training curve in such cases.

The fact that both curves are close and both AUC values are high suggests that the model should perform well in practice, making reliable predictions about ridership. This consistency is what we would hope to see when deploying a model in a real-world setting. The MBTA can use these insights to refine strategies around ticket pricing, service frequency, and capacity planning based on the predicted ridership patterns.

Insights and Recommendations for MBTA:

1. Off-Peak Times: Given the significant negative coefficients for off-peak times, MBTA might consider strategies to encourage ridership during these times, such as promotional fares or targeted marketing campaigns.

2. Route Optimization: Routes like Orange and Red showing positive effects on high ridership could be optimized to handle more traffic, perhaps by increasing service frequency during peak times to manage load better.

3. Service Enhancements: For routes and times showing a tendency towards lower ridership, consider service enhancements or adjustments to attract more passengers.

This logistic regression analysis provides a solid foundation for understanding factors influencing high ridership and can help MBTA make data-driven decisions to enhance service and operational efficiency.

K-Nearest Neighbors (KNN)

Bonus, I also used knn to figure out the nearest neighbors to Babcock Street. They are -

	stop_name	distance
1	Park Street	17.03575
2	Park Street	17.11082
3	Downtown Crossing	17.21247
4	Science Park	17.26254
5	JFK/Umass	17.30139
6	North Station	17.34427

7	Government Center	17.35605

Conclusion

In this study, I examined the Massachusetts Bay Transportation Authority (MBTA) dataset to identify key factors that influence total onboard ridership (`total_ons`) across various stops in the MBTA network. Through comprehensive data analysis, which included data cleaning, exploratory data analysis (EDA), and statistical modeling, several insights were derived that could aid in better understanding and forecasting ridership patterns.

Key Findings:
Statistical Analysis and Modeling: A linear regression model indicated that average_flow is a significant predictor of total_ons, suggesting that higher passenger flow correlates with increased boarding numbers. This model provided a solid foundation for understanding direct relationships within the data.
ANOVA results further supported the significant effects of different time periods (time_period_name) on ridership, highlighting how ridership patterns vary significantly across different parts of the day.

Practical Implications:
The findings suggest that MBTA could potentially optimize service schedules, focus on high-ridership routes for resource allocation, and tailor services to meet peak demand times more effectively. Understanding factors that significantly impact ridership can also aid in strategic planning, especially in anticipating changes in ridership patterns and preparing for future transit needs.

Recommendations:
Service Adjustment: Based on the analysis, targeted adjustments in service during off-peak times could be beneficial, as these periods showed distinct patterns that might be optimized to improve efficiency.
Further Research: Additional variables, such as weather conditions, special events, and demographic data of the ridership areas, could be incorporated into future models to enhance the accuracy and comprehensiveness of the predictions.
Continued Monitoring: Ongoing analysis is recommended to adapt to trends over time, especially in response to external factors like urban development and economic shifts that could impact ridership.

In conclusion, this study provides a robust analytical foundation for understanding the dynamics of ridership within the MBTA system. By leveraging statistical tools and data visualization, we can significantly improve operational strategies and meet the evolving demands of public transit users effectively.

Appendix

```
>
> # Perform ANOVA
> anova_results <- anova(anova_model)
> anova_results
Analysis of Variance Table

Response: total_ons
                    Df     Sum Sq    Mean Sq F value    Pr(>F)
time_period_name     9 1.3138e+12 1.4598e+11  81.469 < 2.2e-16 ***
Residuals         7910 1.4173e+13 1.7918e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Figure 1

```
>
> # Fit the linear model for ANOVA
> anova_model <- lm(total_ons ~ time_period_name, data = MBTA)
> summary(anova_model)

Call:
lm(formula = total_ons ~ time_period_name, data = MBTA)

Residuals:
   Min     1Q Median     3Q    Max
-41644 -19877  -6708   -363 444193

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                       32598.8     1577.6  20.664  < 2e-16 ***
time_period_nameEARLY_AM         -24377.1     2231.0 -10.927  < 2e-16 ***
time_period_nameEVENING           -9658.6     2231.0  -4.329 1.51e-05 ***
time_period_nameLATE_EVENING     -25830.9     2231.0 -11.578  < 2e-16 ***
time_period_nameMIDDAY_BASE        -390.2     2231.0  -0.175    0.861
time_period_nameMIDDAY_SCHOOL      -8689.7     2231.0  -3.895 9.90e-05 ***
time_period_nameNIGHT             -31581.3     2231.0 -14.156  < 2e-16 ***
time_period_nameOFF_PEAK          -18785.4     1932.1  -9.723  < 2e-16 ***
time_period_namePM_PEAK             9045.0     2231.0   4.054 5.08e-05 ***
time_period_nameVERY_EARLY_MORNING -30292.4     2231.0 -13.578  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 42330 on 7910 degrees of freedom
Multiple R-squared:  0.08483,   Adjusted R-squared:  0.08379
F-statistic: 81.47 on 9 and 7910 DF,  p-value: < 2.2e-16
```

Figure 2

```
> slr_model <- lm(total_ons ~ average_flow, data = train_data)
> summary(slr_model)

Call:
lm(formula = total_ons ~ average_flow, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-146360   -8972   -6955   -3684  417018

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6954.8468   613.9178   11.33   <2e-16 ***
average_flow    4.2373     0.1205   35.18   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39480 on 5542 degrees of freedom
Multiple R-squared:  0.1825,    Adjusted R-squared:  0.1824
F-statistic:  1237 on 1 and 5542 DF,  p-value: < 2.2e-16
```

Figure 3

```
print(paste("RMSE on test data:", RMSE))
[1] "RMSE on test data: 40785.2556944918"
```

Figure 4

```
> print(cor_matrix)
                  direction_id    total_ons   total_offs number_service_days   average_ons average_offs
direction_id         1.0000000000 -0.001284794 0.0002717115          0.0000000 -0.0005864897  0.001381189
total_ons           -0.0012847938  1.000000000 0.4245982447          0.1806413  0.6084513315  0.232682715
total_offs           0.0002717115  0.424598245 1.0000000000          0.1818872  0.2325669820  0.608273208
number_service_days  0.0000000000  0.180641321 0.1818871822          1.0000000 -0.1975623135 -0.199286285
average_ons         -0.0005864897  0.608451331 0.2325669820         -0.1975623  1.0000000000  0.463649397
average_offs         0.0013811892  0.232682715 0.6082732084         -0.1992863  0.4636493970  1.000000000
average_flow         0.0125209949  0.432739026 0.3287810466         -0.2576684  0.7350506692  0.582644786
                    average_flow
direction_id          0.01252099
total_ons             0.43273903
total_offs            0.32878105
number_service_days  -0.25766836
average_ons           0.73505067
average_offs          0.58264479
average_flow          1.00000000
```

Figure 5

```
> # Perform backward elimination
> reduced_model <- step(full_model, direction = "backward")
Start:  AIC=113507.9
total_ons ~ season + route_id + direction_id + day_type_name +
    time_period_name + stop_name + total_offs + number_service_days +
    average_ons + average_offs + average_flow

                       Df  Sum of Sq        RSS    AIC
- day_type_name         1 3.1716e+08 4.1173e+12 113506
- direction_id          1 4.9559e+08 4.1175e+12 113507
- average_flow          1 4.9942e+08 4.1175e+12 113507
<none>                               4.1170e+12 113508
- number_service_days   1 1.9958e+09 4.1190e+12 113509
- route_id              3 5.9818e+09 4.1230e+12 113510
- season                2 1.3590e+10 4.1306e+12 113522
- stop_name           112 2.3115e+11 4.3481e+12 113587
- time_period_name      8 1.9130e+11 4.3083e+12 113744
- total_offs            1 4.4299e+11 4.5600e+12 114072
- average_offs          1 4.4904e+11 4.5660e+12 114080
- average_ons           1 1.0703e+12 5.1873e+12 114787

Step:  AIC=113506.4
total_ons ~ season + route_id + direction_id + time_period_name +
    stop_name + total_offs + number_service_days + average_ons +
    average_offs + average_flow
```

Figure 6

```
                        Df  Sum of Sq        RSS    AIC
- average_flow           1 4.5669e+08 4.1178e+12 113505
- direction_id           1 4.9663e+08 4.1178e+12 113505
<none>                                4.1173e+12 113506
- number_service_days    1 1.9753e+09 4.1193e+12 113507
- route_id               3 6.0015e+09 4.1233e+12 113508
- season                 2 1.3638e+10 4.1309e+12 113521
- stop_name            112 2.3270e+11 4.3500e+12 113587
- time_period_name       9 2.2123e+11 4.3385e+12 113779
- total_offs             1 4.4315e+11 4.5605e+12 114071
- average_offs           1 4.5082e+11 4.5681e+12 114080
- average_ons            1 1.0700e+12 5.1873e+12 114785

Step:  AIC=113505
total_ons ~ season + route_id + direction_id + time_period_name +
    stop_name + total_offs + number_service_days + average_ons +
    average_offs

                        Df  Sum of Sq        RSS    AIC
- direction_id           1 5.2460e+08 4.1183e+12 113504
<none>                                4.1178e+12 113505
- number_service_days    1 2.0179e+09 4.1198e+12 113506
- route_id               3 6.9581e+09 4.1247e+12 113508
- season                 2 1.3513e+10 4.1313e+12 113519
- stop_name            112 2.3226e+11 4.3500e+12 113585
- time_period_name       9 2.2789e+11 4.3457e+12 113786
- total_offs             1 4.4574e+11 4.5635e+12 114073
- average_offs           1 4.8997e+11 4.6077e+12 114126
- average_ons            1 1.9864e+12 6.1042e+12 115685

Step:  AIC=113503.7
total_ons ~ season + route_id + time_period_name + stop_name +
    total_offs + number_service_days + average_ons + average_offs
```

Figure 7

```
                        Df  Sum of Sq        RSS    AIC
<none>                                4.1183e+12 113504
- number_service_days    1 2.0099e+09 4.1203e+12 113504
- route_id               3 6.9645e+09 4.1252e+12 113507
- season                 2 1.3519e+10 4.1318e+12 113518
- stop_name            112 2.3221e+11 4.3505e+12 113584
- time_period_name       9 2.2781e+11 4.3461e+12 113784
- total_offs             1 4.4635e+11 4.5646e+12 114072
- average_offs           1 4.9014e+11 4.6084e+12 114125
- average_ons            1 1.9872e+12 6.1054e+12 115685
>
> # Display the summary of the reduced model
> summary(reduced_model)

Call:
lm(formula = total_ons ~ season + route_id + time_period_name +
    stop_name + total_offs + number_service_days + average_ons +
    average_offs, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-237845  -10326   -1197    7452  265430
```

Figure 8

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27580 on 5413 degrees of freedom
Multiple R-squared:  0.6103,    Adjusted R-squared:  0.601
F-statistic: 65.22 on 130 and 5413 DF,  p-value: < 2.2e-16

>
```

Figure 9

```
> print(performance_summary)
      Data      RMSE       MAE        R2
1 Training 27255.02 14970.55 0.6103227
2  Testing 30407.46 16180.72 0.5532455
>
```

Figure 10

```
> summary(logistic_model)

Call:
glm(formula = high_ridership ~ route_id + time_period_name +
    average_flow, family = binomial(link = "logit"), data = train_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.9760  -0.7330  -0.1416   0.7057   2.4709

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -7.809e-01  1.630e-01  -4.790 1.67e-06 ***
route_idGreen                   -2.073e-01  1.233e-01  -1.681 0.092808 .
route_idOrange                   5.860e-01  1.528e-01   3.835 0.000126 ***
route_idRed                      5.461e-01  1.444e-01   3.781 0.000156 ***
time_period_nameEARLY_AM        -4.767e-01  1.554e-01  -3.067 0.002165 **
time_period_nameEVENING          2.533e-02  1.575e-01   0.161 0.872255
time_period_nameLATE_EVENING    -3.785e-01  1.521e-01  -2.489 0.012807 *
time_period_nameMIDDAY_BASE      3.717e-01  1.609e-01   2.310 0.020879 *
time_period_nameMIDDAY_SCHOOL    2.480e-01  1.586e-01   1.563 0.117983
time_period_nameNIGHT           -1.247e+00  1.766e-01  -7.060 1.66e-12 ***
time_period_nameOFF_PEAK        -2.207e+00  1.741e-01 -12.682  < 2e-16 ***
time_period_namePM_PEAK          9.942e-02  1.618e-01   0.614 0.538951
time_period_nameVERY_EARLY_MORNING -9.694e-01  1.660e-01  -5.840 5.22e-09 ***
average_flow                     8.985e-04  3.971e-05  22.629  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7685.6  on 5543  degrees of freedom
Residual deviance: 4903.1  on 5530  degrees of freedom
AIC: 4931.1

Number of Fisher Scoring iterations: 7
```

Figure 11

```
> print(confusion_matrix)
             TestPredicted
TestObserved    0    1
           0 1034  154
           1  345  843
> # Calculate accuracy
> accuracy <- sum(diagonal(confusion_matrix)) / sum(confusion_matrix)
Error in diagonal(confusion_matrix) : could not find function "diagonal"
> accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
> print(paste("Accuracy:", accuracy))
[1] "Accuracy: 0.789983164983165"
>
```

Figure 12

```
> library(pROC)
> # For the training set
> train_probs <- predict(logistic_model, newdata = train_set, type = "response")
> roc_train <- roc(train_set$high_ridership, train_probs)
Setting levels: control = 0, case = 1
Setting direction: controls < cases
> plot(roc_train, main="ROC Curve for Training Set")
> auc_train <- auc(roc_train)
> print(paste("AUC for Training Set:", auc_train))
[1] "AUC for Training Set: 0.874723776103646"
> # For the test set
> test_probs <- predict(logistic_model, newdata = test_set, type = "response")
> roc_test <- roc(test_set$high_ridership, test_probs)
Setting levels: control = 0, case = 1
Setting direction: controls < cases
> plot(roc_test, main="ROC Curve for Test Set")
> auc_test <- auc(roc_test)
> print(paste("AUC for Test Set:", auc_test))
[1] "AUC for Test Set: 0.883393417905202"
```

Figure 13

https://mbta-massdot.opendata.arcgis.com/search?tags=ridership