STA141A Final Report: London Bike Sharing Analysis

Written by: Lauren Cordano, Julia Webb, Jesus Leon, and Nilay Varshney

Contributions

Lauren	Background, Methods, Association between season and total number of rides in a day, Association between the day of the week (weekday/weekend) and total rides, Time of day that the highest number of rides were initiated, Questions of Interest and Assumptions, Conclusion
Julia	Methods, Association between the day of the week (weekday/weekend) and total rides, Association between season and total number of rides in a day, Exploring the relationship between the total number of rides in a day and the continuous weather variables
Jesus	Time Series Analysis
Nilay	Association between holiday and number of rides Association between weather code and number of rides

Background

In the last decade, bike-sharing and car-sharing platforms have become more popular due to the high cost of purchasing a car and the high traffic congestion rates. In the UK, transportation accounted for 32% of total greenhouse gas emissions between 2017 and 2018. (UK National Statistics Report 2019, pg 14) Successful bike and car-sharing programs can help mitigate transportation emissions and maximize collective transportation efficiency. Additionally, convenient app-based technologies are making transportation innovation more possible than ever before.

We analyzed Santander Cycle Hire rides, a bike-sharing service based in London. The London Transport Authority collected the data over a period of two years. We were interested in how ridership varied based on time of day, and on weekdays versus weekends. Observations were taken every hour from midnight on January 4, 2015 to January 3, 2017. Our variables include the count of new trips started every hour, the real temperature, the ‘feels like’ temperature, the humidity (in percentage), the wind speed (in kilometers per hour), a numeric weather code (ranging from clear to snowfall), a boolean variable for if it is a holiday, a boolean variable for if it is a weekday, and a numeric season code (0-Spring ; 1-Summer; 2-Fall; 3-Winter).

Understanding how ridership varies based on time of day and type of day can help Santander Cycle Hire optimize how many bikes they offer to customers. It also allows them to effectively market their services to prospective riders.This data set did not offer

map coordinates of where rides began and ended, but that would be valuable information for Santander Cycle Hire to utilize in the future.

Questions of Interest and Assumptions

1. What is the relationship between season and number of rides?
2. What is the relationship between the day of the week (weekend or weekday) and number of rides?
3. What is the relationship between ride counts and weather (in terms of the weather code variable, and the continuous weather variables)? Note that this corresponds with Key Questions 3 and 6 from our project proposal.
4. During what time of day are the most number of rides initiated?
5. Is there an association between holiday and number of rides?

Initially, we were also interested in using logistic regression to predict the type of day (weekend or weekday) given the number of rides, but the professor advised us to forgo this question due to its lack of functionality. Instead we made confidence intervals for mean rides in a day, categorized by type of day (weekday or weekend).

Also, regarding question 3 from above, the original key question in our proposal stated that we planned to use a linear regression model predict total rides in a day. Since we have a time series dataset in which data is collected every hour, we decided it would be more appropriate to use a SARIMA model to forecast rides initiated per hour. Then, we will model our entire dataset using linear regression, in order to gain an understanding of the relationship between between ridership and our continuous weather variables.

Methods

For the forecasting section of our analysis, a SARIMA model is used to forecast hourly bike rides. The last 120 hours of our data set will be used as testing data, and the training data will be composed of the remaining observations.

To explore the linear relationship between rides per day and our continuous weather predictor variables (temperature, 'feels like' temperature, windspeed, humidity), we ran a multiple linear regression using the lm() function in R. However, a possible issue with this method may be dependence between the predictors, given that we are modeling a time series dataset. In an effort to reduce this dependence, we modeled total rides per day as a function of average temperature per day, average 'feels like' temperature per day, wind speed per day, and humidity per day. In an effort to see the effect of dependence on our model, we made a second identical model that was only based on a random sample of 300 observations from the dataset, thinking the magnitude of dependence would be less since the observations were randomly selected. The change in the value of the coefficients, and the Adjusted-R^2 value were relatively small, so we decided to proceed with the analysis.

For the seasonal, type of day (weekday vs. weekend), and the time of day analyses, the Welch t-test will be used, which is the t.test() function in R. Welch's t-test assumes normality, which is satisfied due to our large amount of observations, but the test does not assume equal variance between samples. These t-tests compare the mean total rides per day between the respective groups.

The holiday analysis uses a Shapiro-Wilkes test to determine if the 16 holiday data points meet the assumption of normality. Then, a t-test is used to determine if there is a difference between the mean total rides per day between holidays and non-holidays.

The weather code analysis uses a Welch's ANOVA F-test to compare differences in ridership based on weather code. It also uses a Games-Howell post hoc test to compare number of rides between five pairs of weather code. These t-tests compare the mean total rides per hour. Even though the weather code is essentially ordered by the clearness of the sky (see weather code next page), it would be inappropriate to take a weather code index, like an average or sum, since the numeric values are not consecutive and the index would be highly inflated by numeric values for poorer weather.

Weather Code:

1 = Clear
2 = scattered clouds / few clouds
3 = Broken clouds
4 = Cloudy
**no observations had a weather code of 94 in our dataset.*

7 = Rain/ light Rain shower/ Light rain
10 = rain with thunderstorm
26 = snowfall
*94 = Freezing Fog**

Results

Time Series Analysis and Forecasting:

One of the objectives of this project was to predict the number of rides booked over a 24 hour period. The data is a time series which is indexed hourly. There is data for 2015, 2016, and the first three days of 2017. The data is non stationary. By plotting the time series there is evident seasonality and trend present. Due to the dataset being so large, a zoomed out portion of the initial times is included for reference of time index relative to the plots.

Figure 1: Hourly Number of Rides Booked

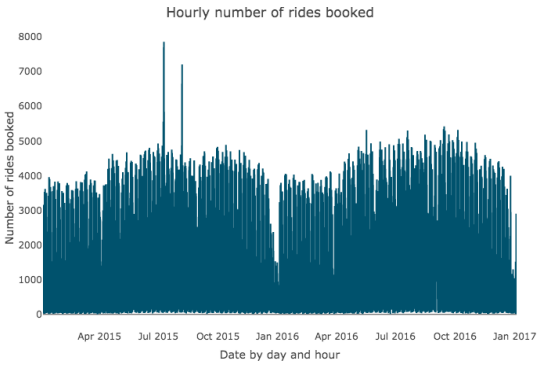
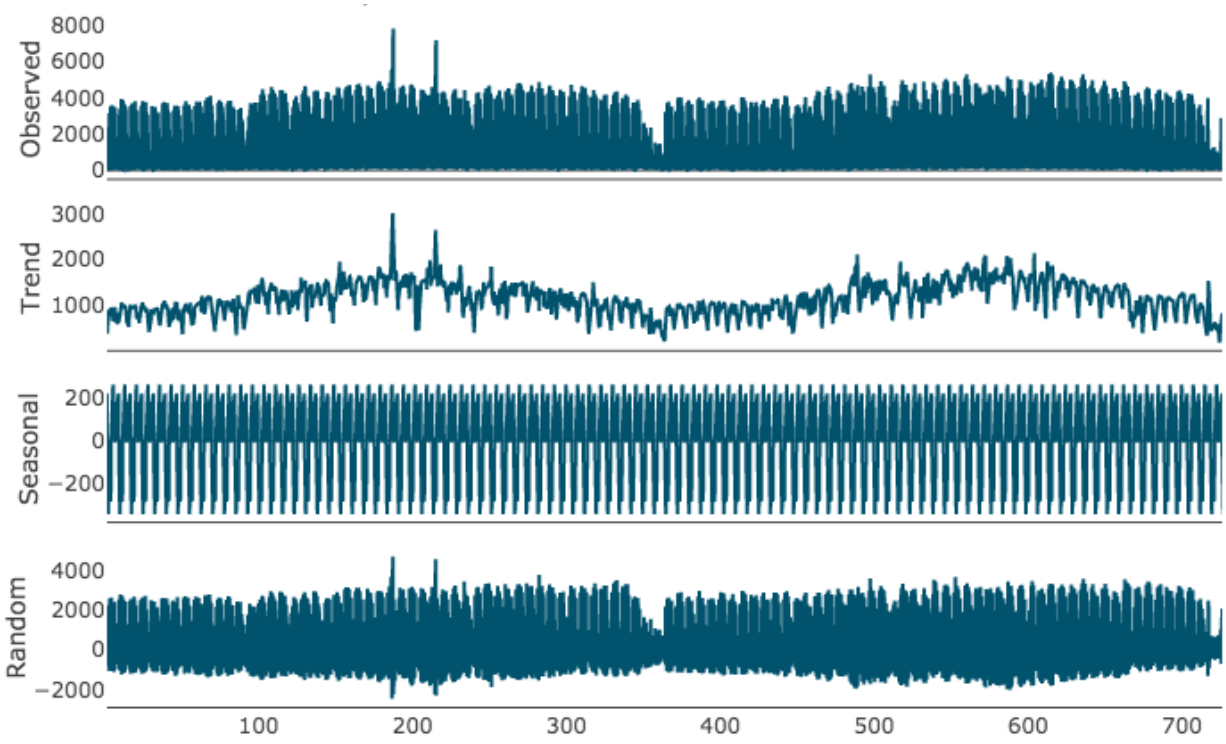


Figure 2: Hourly Number of Rides Booked

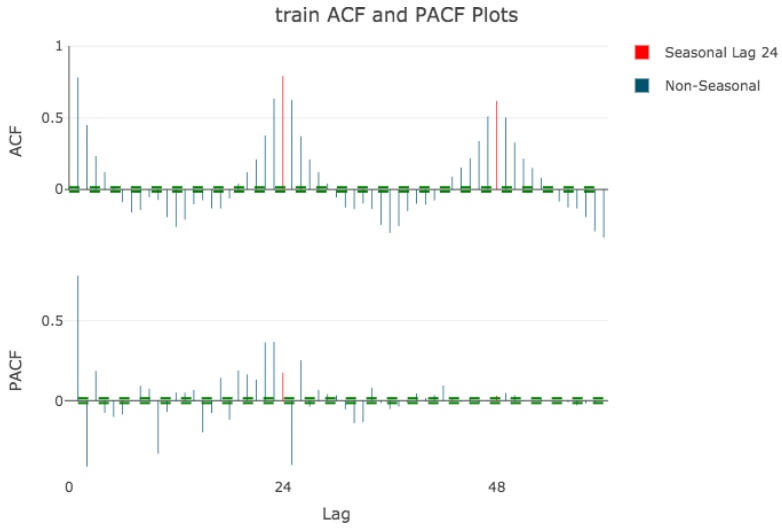
This is an additive time series and it is not stationary. The series was decomposed by it's trend, seasonality, and white noise component. These were then plotted against the original series.

Figure 3: Decomposition of Additive Time Series - DataTS



By this visual analysis of the series there is clear indication of seasonality and trend. We confirm this by analyzing the Autocorrelation function and Partial Autocorrelation function of the series. The PACF function has a peak at lag 1, this confirms that a trend term is present in the data. Furthermore, we can see that the ACF function is correlated and has seasonal peaks at time lags multiple of 24.

Figure 4: Train ACF and PACF Plots



Now that there is confirmation of seasonality and trend we must detrend and get rid of seasonality in order for the data to be stationary. This is accomplished by differencing the time series. Order one differencing detrends the series and lag 24 differencing takes care of the seasonal component. The data now displays time independent variance.

Figure 5: First Seasonal Difference

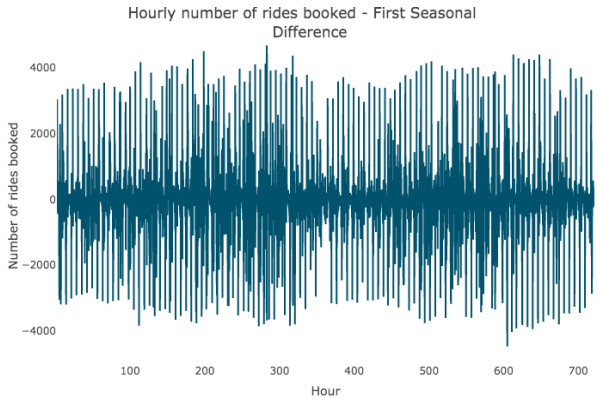
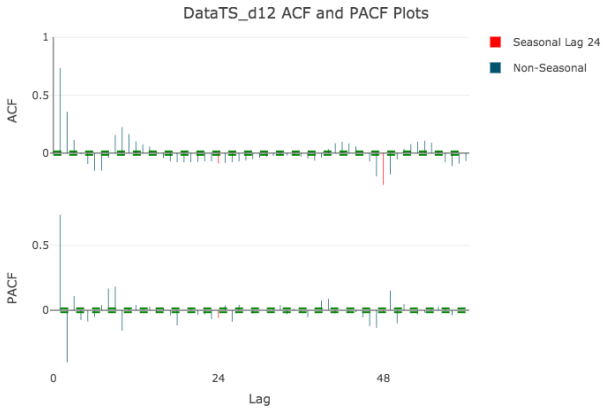


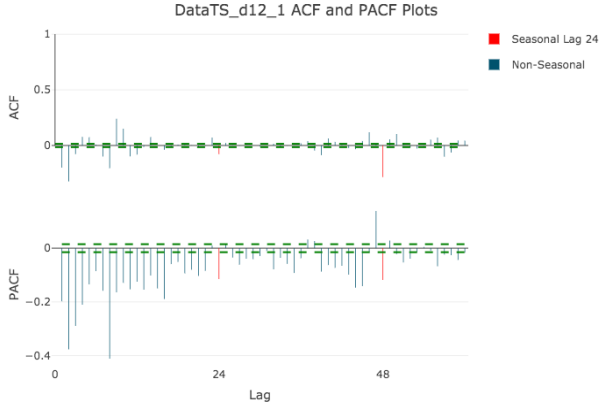
Figure 6: ACF & PACF of First Seasonal Difference



The ACF plot shows no correlation after lag 24 differencing, the seasonality component is resolved, the differenced series is also displayed. However, there is still trend present, so after differencing with order 1 the data is finally stationary. The stationary time series is displayed below in Figure 7.

Figure 7: First Trend Difference

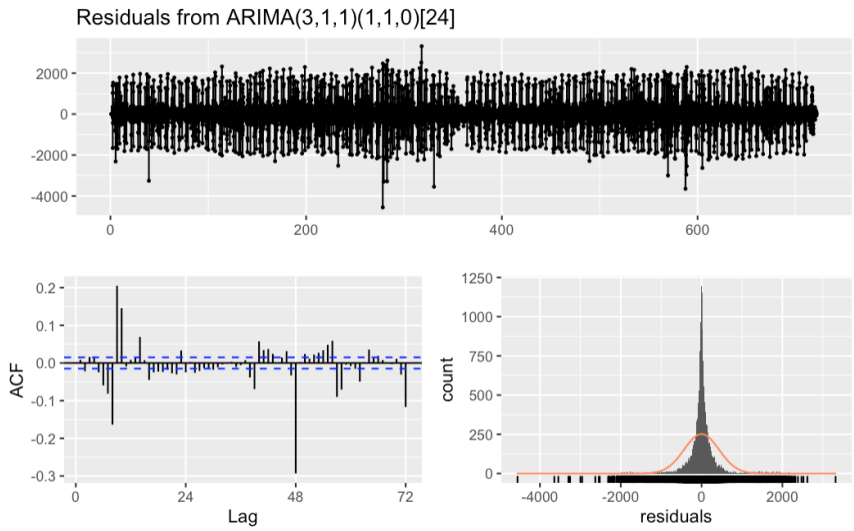
Figure 8: ACF & PACF of First Trend Difference



From the information that the ACF and PACF plots provided, it would be appropriate to fit a SARIMA model, seasonality being 24 time periods. By examining the stationary ACF and PACF we can see that for the non seasonal component the PACF cuts off at lag 2 and the ACF seems to tail off indicating an AR(3) process. The ACF appears to cut off at lag 2 as well. The same deduction process would be applied to the seasonality part of the model, However auto.arima is implemented to choose the best parameters for the SARIMA model. We will use the model SARIMA(3,1,1)(1,1,0)[24] to fit the data

as well as to forecast future values. For performance comparison we will use the model we derived SARIMA(2,1,2)(0,1,0)[24] based on the ACF and PACF plots.
To measure the performance of the models we will compare the AIC score of the two since it is a large dataset. SARIMA(3,1,1)(1,1,0)[24] had an AIC score of 14.95 and SARIMA(2,1,2)(0,1,0)[24] had an AIC score of 14.96. Finally we inspect the residuals, which show no deviation from white noise, no correlation and a normal distribution.

Figure 9: Residuals



The model SARIMA(3,1,1)(1,1,0)[24] performed best, so it is used to fit and forecast our data. The model fit is shown below in red, fitted on the training data, the forecasted data is displayed in green, our model was applied to the test data.

Figure 10a: Forecasted and Fitted Rides

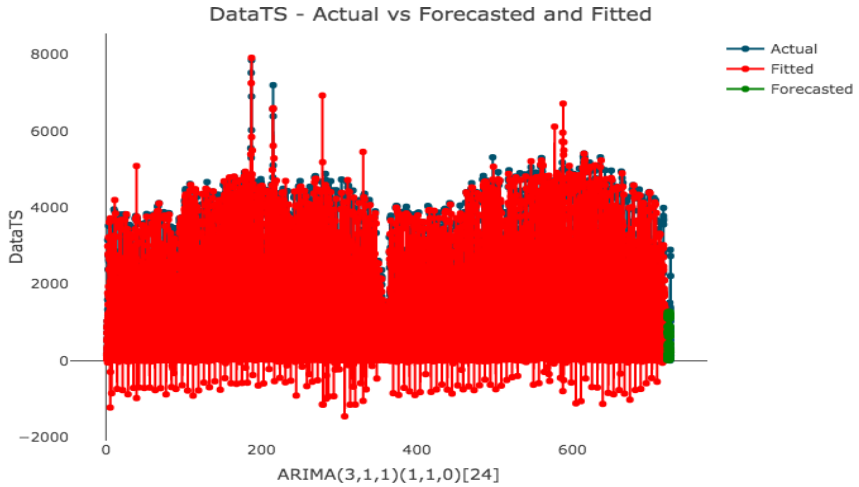
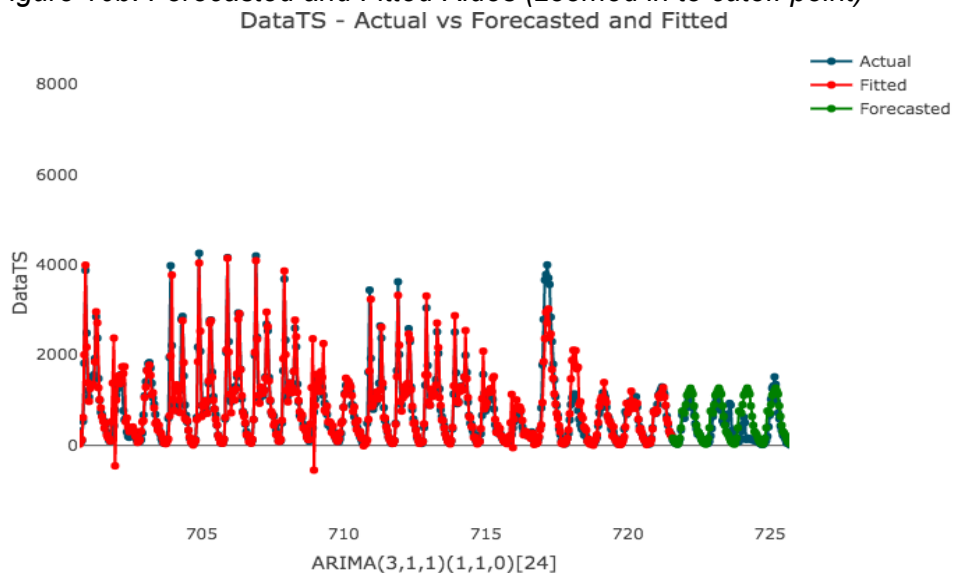
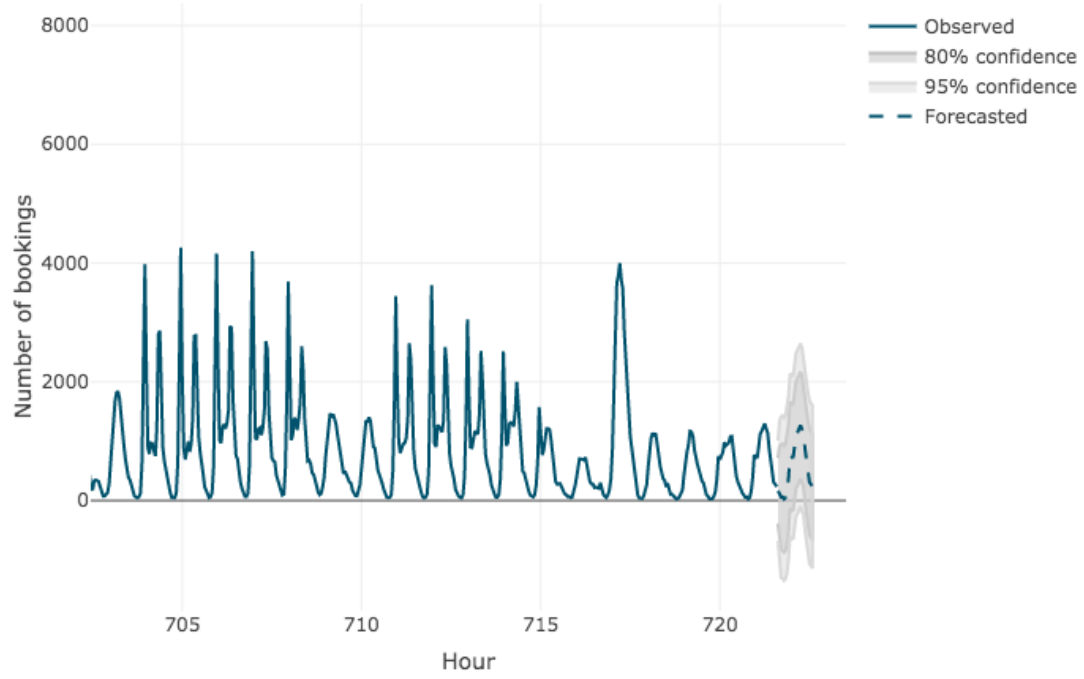


Figure 10b: Forecasted and Fitted Rides (zoomed in to cutoff point)



The 24 hour ahead forecast was applied to the training data. In Figure 11, shown next, there is an 80% CI and a 95% CI threshold on the forecast.

Figure 11: Forecasted Number of Rides Booked



To summarize, as we can see from Figure 3., the data has seasonality and trend present. These components must be removed from the time series to make it stationary. Stationary time series have a constant mean and time independent variance. Trend is removed by differencing once and seasonality is removed by differencing with order of

24, since it is hourly data. Stationarity is important so we can derive the appropriate SARIMA model to fit our train data. Then the model is used to forecast test data.

Investigation to Understand our Dataset:

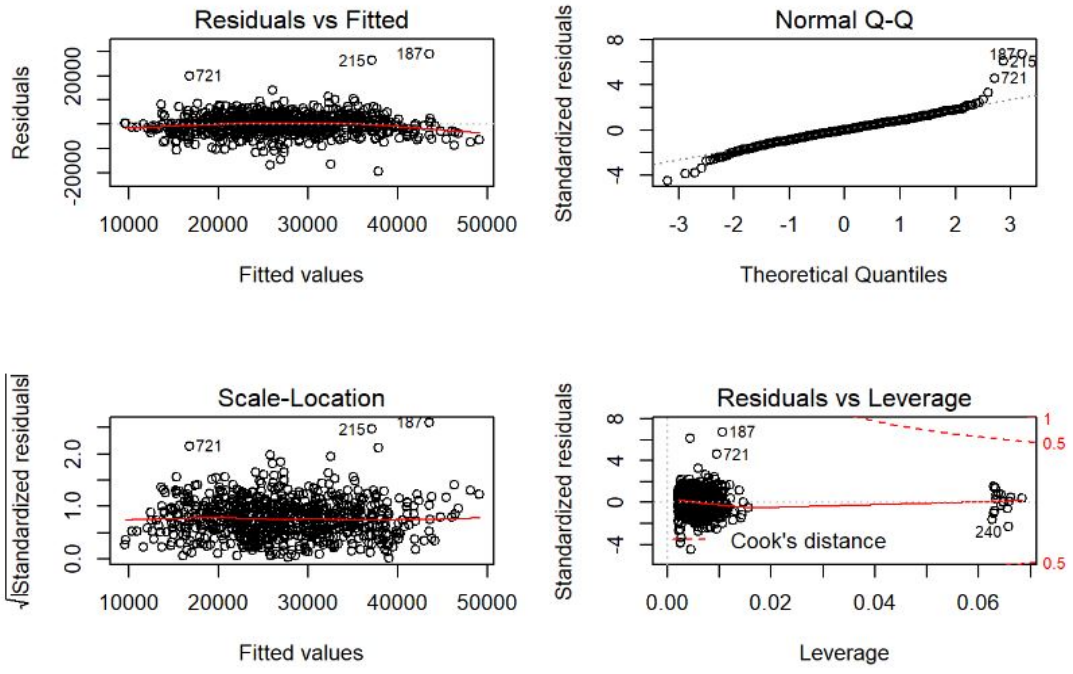
Exploring the relationship between the total number of rides in a day, and the continuous weather variables: temperature, feels like temperature, wind, and humidity.

Initial Model: $\widehat{Y} = \beta_0 + \beta_{temp} + \beta_{temp\,feel} + \beta_{wind} + \beta_{humidity}$

As would be expected, there was a high level of collinearity between the average temperature variable and the average 'feels like' temperature variable, as well as collinearity between the average wind and average humidity. So, in an effort to reduce collinearity while still including all of our data, we created a temperature index variable, which was the sum of our temperature variables, and a wind and humidity variable, which was the sum of the wind and humidity variables. These will be referred to as the temperature index, and the wind/humidity index for the rest of the report. We also decided to include two binary predictors, for weekend and holiday. These predictors showed significance in our results, so their inclusion provides a more accurate estimation of our temperature and wind/humidity index.

Updated Model : $\widehat{Y} = \beta_0 + \beta_{temp\_index} + \beta_{wind\_humidity\_index} + \beta_{weekend} + \beta_{holiday}$

Figure 12: Residuals of Rides ~ Weather



The residuals are evenly scattered about zero in the Figure 12: Residuals vs. Fitted plot. The standardized residuals are also evenly scattered about zero in the Figure 12: Scale-Location, and Residuals vs Leverage plot, and they display a close to linear relationship with the theoretical residual quantiles in Figure 12: Normal QQ plot. These diagnostic plots imply that our residuals are normally distributed about zero, and have equal variance. There are also some notably high leverage points, but their residuals are close to zero, so they are not influential outliers that would negatively affect our model. Additionally, the residual plots indicate that observations 187, 215, and 721 were outliers, however from the Residuals vs Leverage plots indicate that these points do not have high leverage, thus they are not highly influential outliers. It is relevant to note that observation 215 corresponds to a workers strike of the London Underground, and observation 721 corresponds to Christmas day. These events may have influenced the increase in ridership.

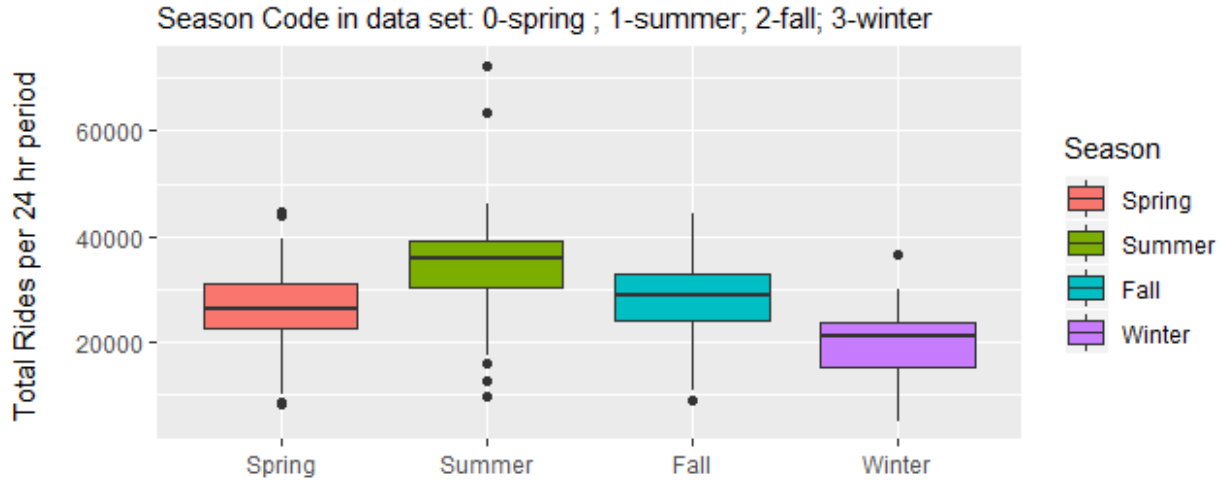
A correlation matrix was also created in R to check between covariance of our predictors. The only notable correlation was -0.3176 between the temperature and wind/humidity index. However, none of the variance inflation factors for the predictors was above 1.2, thus the correlation between the two indexes is not problematic.

The p-values provided from the summary output was less than 2.2e-16 for the temperature index, wind/humidity index, and the weekend binary. The p-value of the holiday binary was 1.12e-12. Thus, the probability of observing a t-statistic more extreme than any of our observed t-statistics is very close to zero, which gives us evidence to reject the null hypotheses that the true coefficient for the predictor variables is zero. The coefficient estimates were 417.10 for the temperature index, indicating that for every single unit increase in temperature index, the expected number of rides in a day increases by about 417. The coefficient estimate for the wind/humidity index was -374.93, indicating that for every single unit increase in wind/humidity, the expected number of rides in a day decreases by about 374. The coefficient for the binary weekend predictor was -5387.92, indicating that on weekend days the total number of rides is about 5387 rides lower than total rides on weekdays. Lastly, the binary holiday predictor had a coefficient of -8006.31, indicating that holidays have an average of about 8006 fewer rides than non-holidays. In terms of the original hourly temperature variables, if enough hourly temperature or ‘feels like’ temperature observations increase such that it raises their respective averages, then this would cause an increase in the temperature index, thus the associated increase in expected rides. On the other hand, in terms of the original hourly wind and hourly humidity variables, if enough hourly wind or hourly humidity observations increase such that it raises their respective averages, then this would cause an increase in the wind/humidity index, thus the associated

decrease in expected rides. The R^2 value for this model is 0.7466, which implies that 74.66% of variation in rides per day can be explained by the model. The R^2 value is also close to the Adjusted R^2 value of 0.7452, which indicates that the value of R^2 is not inflated by the predictors. Finally, the p-value of the F-test was less than 2.2e-16, indicating a strong relationship between the response and predictors.

Association between season and total number of rides in a day

Figure 13: Median Rides per Day Based on Season



Judging from Figure 16 above, it appears that ridership increases in the Summer. Welch t-tests were performed between pairs of means in descending order of highest mean ride count.

Table 1: Mean Number of Rides by Season

Season	**Mean Number of Rides**
Spring	1103.832
Summer	1464.465
Fall	1178.954
Winter	821.7291

Mean of total rides per day in Summer is greater than mean rides per day in Fall:

H_o: $\mu_{Summer} = \mu_{Fall}$ H_a: $\mu_{Summer} > \mu_{Fall}$

The resulting p-value was 2.2e-16, thus there is sufficient evidence to reject the null hypothesis and conclude that the number of rides on an average Summer day is significantly higher than the number of rides on an average Fall day. Because, the Fall has the second highest ridership compared to Summer, we can extrapolate that the

number of rides on an average Summer day are also significantly higher than the number of rides on average Winter and Spring days.

Mean of total rides per day in Fall is greater than mean rides per day in Spring:

H_o: $\mu_{Fall} = \mu_{Spring}$ H_a: $\mu_{Fall} > \mu_{Spring}$

The resulting p-value was 0.01007, thus there is sufficient evidence to reject the null hypothesis and conclude that the number of rides on an average Fall day is significantly higher than the number of rides on an average Spring day. From this result, we also conclude that an average Fall day has significantly more rides than the average Winter day.

Mean of total rides per day in Spring is greater than mean rides per day in Winter:

H_o: $\mu_{Spring} = \mu_{Winter}$ H_a: $\mu_{Spring} > \mu_{Winter}$

The resulting p-value was 2.2e-16, thus there is sufficient evidence to reject the null hypothesis and conclude that an average Spring day has significantly more rides than an average Winter day.

Association between the day of the week (week day or weekend) and total rides

Figure 14 : Hourly Plot of Rides on Weekdays vs. Weekends

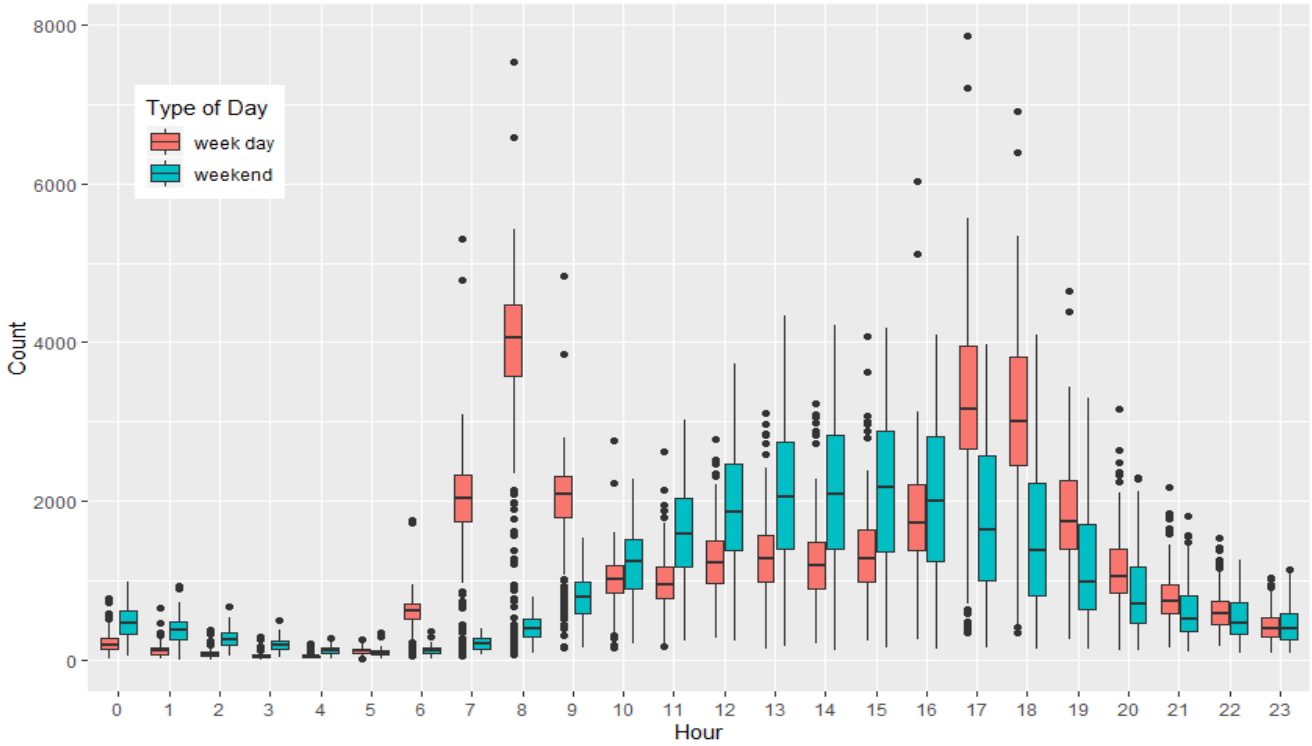
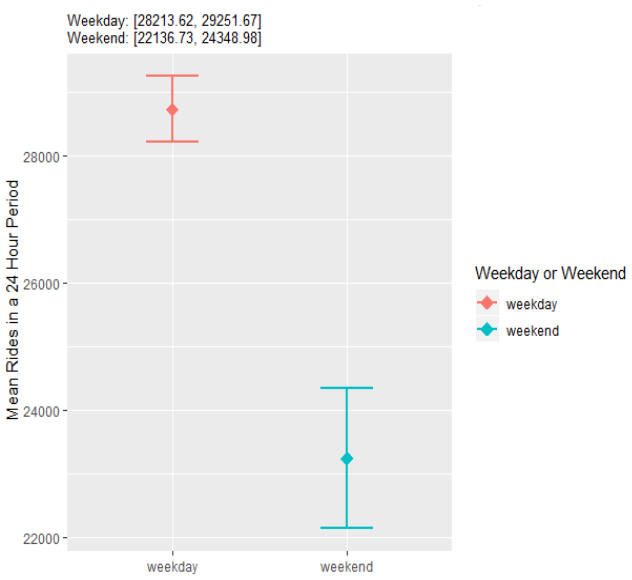


Figure 15: Mean Rides on Weekday vs. Weekend

Figure 16: 95% CI for Rides on Weekdays vs. Weekends



Based on the data shown in Figure 14, we ran a Welch's t-test to see if the mean for total rides in a 24 hour period on a weekday was significantly greater than the mean for total rides in a 24 hour period for a weekend.

H_o: $\mu_{weekday} = \mu_{weekend}$ H_a: $\mu_{weekday} - \mu_{weekend} > 0$

The resulting p-value was 2.369e-13, thus there is sufficient evidence to reject the null hypothesis and conclude that there are more rides initiated on an average weekday than there are on an average weekend day. Figure 13 suggests that the high volume of rides on weekdays may be due to commuters, since there is a spike in ridership between the hours of 7:00am and 9:00am, as well as between 5:00pm and 6:00pm. Based on Figure 15, one can state with 95% confidence that the mean total rides per day on a weekday is between 28,213 and 29,251, while the mean total rides per day on a weekend is between 22,136 and 24,348.

Time of day that the highest number of rides were initiated:

After examining Figure 14, we decided to split up a day into four categories: morning, afternoon, evening, and night to make time of day comparison simpler. The four categories cover the hours between 7am and 10 pm. We decided not to include rides between 10pm and 7am in our analysis because ridership during that time is usually very low.

We also decided to split up weekdays and weekends to maximize the granularity of our analysis. We did this based on a reasonable assumption that riding behavior would differ greatly on weekdays compared to weekends. Specifically, we reasoned that

people likely commute on weekdays and sleep in on weekends. Figure 14 depicts this behavior variance.

A Welch t-test was performed for the total number of rides per 4-hour window to compare the weekday ridership and the weekend ridership by time of day.

Table 2: Mean Number of Rides by Time of Day and Type of Day (Weekday vs. Weekend)

	Weekday	Weekend
Morning Hrs [7:00-10:59]	2681.094	472.2113
Afternoon Hrs [11:00-14:59]	1223.931	2041.266
Evening Hrs [15:00-18:59]	2734.37	1816.274
Night Hrs [19:00-22:59]	857.166	677.9071

Number of rides in the morning (7:00-10:59) is higher on weekdays than on weekends:

H_o: $\mu_{\text{weekday morning}} = \mu_{\text{weekend morning}}$ H_a: $\mu_{\text{weekday morning}} > \mu_{\text{weekend morning}}$

The resulting p-value was less than 2.2e-16, signifying that we can reject the null hypothesis and conclude that the number of rides on an average weekday morning is significantly more than the number of rides on an average weekend morning.

Number of rides in the afternoon (11:-14:59) is higher on weekends than on weekdays:

H_o: $\mu_{\text{weekday afternoon}} = \mu_{\text{weekend afternoon}}$ H_a: $\mu_{\text{weekday afternoon}} > \mu_{\text{weekend afternoon}}$

The resulting p-value was less than 2.2e-16, signifying that we can reject the null hypothesis and conclude that the number of rides on an average weekend afternoon is significantly higher than the number of rides on an average weekday afternoon.

Number of rides in the evening (15:00-18:59) is higher on weekdays than weekends:

H_o: $\mu_{\text{weekday evening}} = \mu_{\text{weekend evening}}$ H_a: $\mu_{\text{weekday evening}} > \mu_{\text{weekend evening}}$

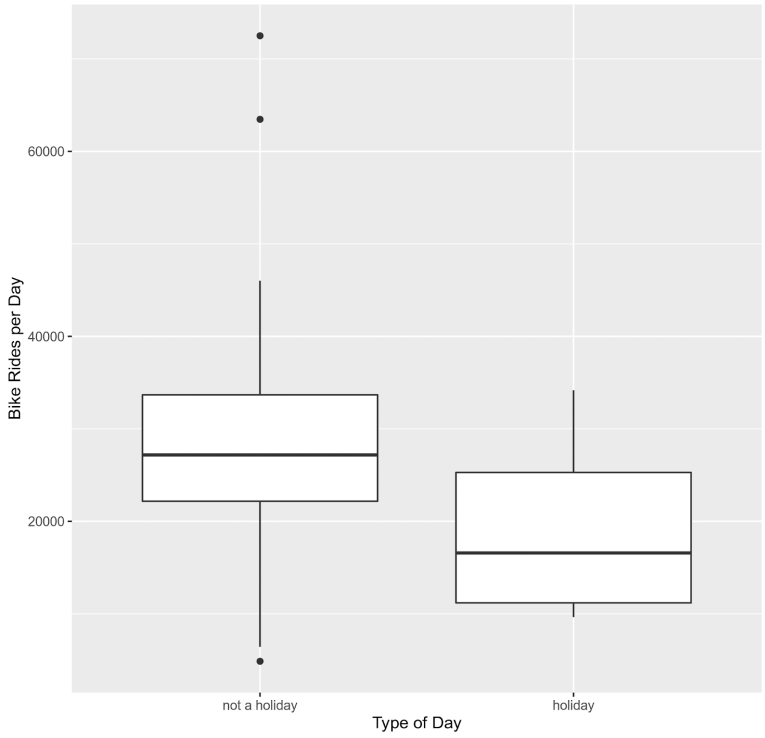
The resulting p-value was less than 2.2e-16, signifying that we can reject the null hypothesis and conclude that the number of rides on an average weekday evening is significantly higher than the number of rides on an average weekend afternoon.

Number of rides at night (19:00-22:59) is higher on weekdays than weekends:

H_o: $\mu_{\text{weekday night}} = \mu_{\text{weekend night}}$ H_a: $\mu_{\text{weekday night}} > \mu_{\text{weekend night}}$

The resulting p-value was less than 2.2e-16, signifying that we can reject the null hypothesis and conclude that the number of rides on an average weekday night is significantly higher than the number of rides on a weekend night.

Association between holiday and number of rides



This section explores the effect of type of day on the daily number of bike rides for two types of days: holidays and non-holidays. This analysis uses the number of rides per day instead of the number of rides per hour because the two-sample test used in this section works better with smaller sample sizes. The first step in the analysis was generating boxplots for the total rides per day for holidays and non-holidays.

Figure 17: Rides on Holidays vs. Non-Holidays

In the dataset, there are 16 holidays and 714 non-holidays. The sample of daily bike ride count for non-holidays satisfies the normality assumption of a Student's two-sample t-test due to its large size, since it can be assumed, using the Central Limit Theorem, that the sample mean for non-holidays is approximately normally distributed. A Shapiro-Wilk test was conducted on the sample of daily bike ride count for holidays to assess the sample's normality. The test statistic was 0.87353, corresponding to a p-value of 0.03081, so at a significance level of 0.01, the aforementioned sample can be considered to have been drawn from a normally distributed population, which means that the assumption of normality had been satisfied.

An F-test for equality of variances was conducted on the two samples in order to assess whether or not the assumption of equal variances was met for the Student's two-sample t-test. The test statistic was 1.064 with 713 and 15 degrees of freedom, corresponding to a p-value of 0.9621, so at a significance level of 0.01, the two samples can be considered to have come from populations with equal variance, which means that the assumption of equal variance had been satisfied.

The Student's two-sample t-test for difference in means was conducted at a significance level of 0.01. The test statistic was 4.1813 with 728 degrees of freedom, corresponding to a p-level of 0.00003251. Since the p-value was much lower than the significance level, the null hypothesis of both samples coming from populations with equal means was rejected. In other words, it was determined that in general, the average number of bike rides in a holiday will be different from the average number of bike rides in a

non-holiday. Therefore, it was confirmed that there is an effect of the type of day (holiday vs. non-holiday) on the number of bike rides that day.

Association between weather code and number of rides

This section explores the effect of the type of weather on the hourly number of bike rides. This analysis uses the number of rides per hour because the weather code does not stay fixed for an entire 24-hour period. The first step in the analysis was generating boxplots for the number of rides per hour for each type of weather.

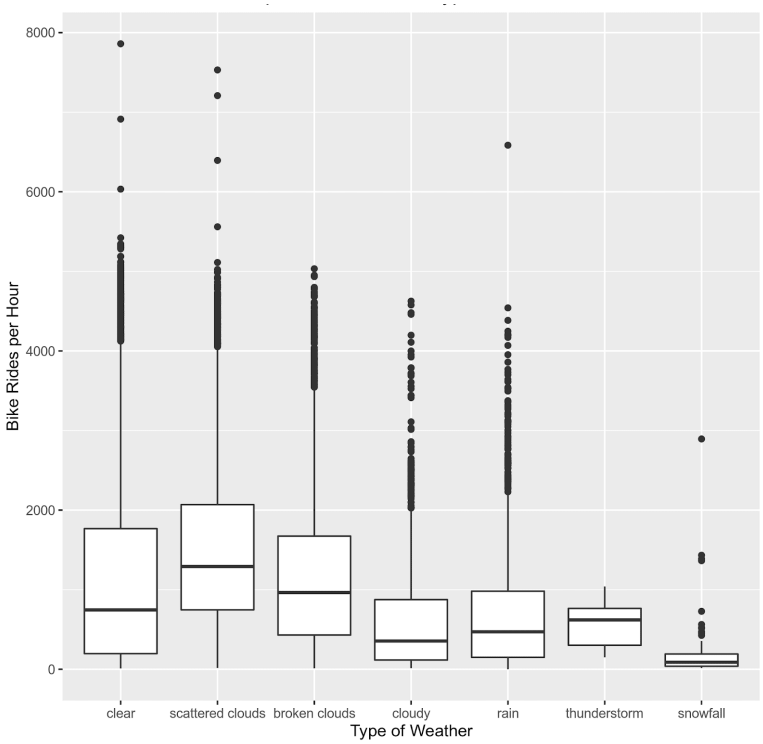


Figure 18: Rides by Weather Code

Based on Figure 18, it seems that the distribution of number of bike rides per hour varies by type of weather. Nonetheless, a one-way ANOVA test was conducted to affirm or dispute the significance of this difference.

In the dataset, there are 6150 hours of clear weather, 4034 hours of scattered clouds, 3551 hours of broken clouds, 1464 hours of cloudy weather, 2141 hours of rainy weather, 14 hours of thunderstorms, and 60 hours of snowfall. The samples of hourly bike ride counts for all types of weather other than thunderstorms satisfy the normality assumption of a Fisher's ANOVA F-test due to their large sizes. A Shapiro-Wilk test was conducted on the sample of hourly bike ride counts for thunderstorms to assess the sample's normality. The test statistic was 0.93877, corresponding to a p-value of 0.4028, so at a significance level of 0.01, the aforementioned sample can be considered to have been drawn from a normally distributed population, which means that the assumption of normality had been satisfied.

A Levene's test for equality of variances was conducted on the samples in order to assess whether or not the assumption of equal variances was met for the Fisher's ANOVA F-test. The test statistic was 98.271 with 6 degrees of freedom, corresponding to a p-value of 0, so at a significance level of 0.01, the two samples cannot be considered to have come from populations with equal variances, which means that the assumption of equal variance had been violated. As a result, the Fisher's ANOVA F-test could not be used. Instead, the Welch's ANOVA F-test was used, as this test requires

samples to come from normal distributions, but not from distributions with equal variances.

The Welch's ANOVA F-test for difference in means was conducted at a significance level of 0.01. The test statistic was 307.7 with 6 and 171.84 degrees of freedom, corresponding to a p-level of 0. Since the p-value was much lower than the significance level, the null hypothesis of all samples coming from populations with equal means was rejected. In other words, it was determined that in general, the average number of bike rides in an hour will depend on the type of weather.

To examine the differences in means of the hourly bike ride counts for each pair of weather types, a Games-Howell post hoc test was conducted at a significance level of 0.01 as an alternative to the Tukey's HSD test, as the former does not assume equal population variances and equal sample sizes, whereas the latter does.

Table 3: Mean Rides based on Weather Code and Game-Howell post hoc Test Results

	Mean Rides	**Comparing to...**	**P-value**	**Result**
Clear	1162.089	Broken clouds	0.774	Not significant
Scattered Clouds	1496.177	N/A	N/A	N/A
Broken Clouds	1195.124	N/A	N/A	N/A
Cloudy	635.2309	Rain	0.04	Not significant
Rain	712.9664	Thunderstorms	0.642	Not significant
Thunderstorms	583.4286	Cloudy	0.663	Not significant
Snowfall	250.85	Thunderstorms	0.025	Not significant

Based on the Game-Howell post hoc tests, there is no significant difference in the average number of bike rides per hour for broken clouds and clear weather, nor for cloudy weather and rainy weather, nor for cloudy weather and thunderstorms, nor for rainy weather and thunderstorms, nor for snowfall and thunderstorms.

Conclusion

Overall, the Santander Cycle Hire bike-sharing program is at least one viable alternative to private vehicle transportation in London. Many days had over 25,000 rides initiated, which is quite substantial.

Our forecasting analysis proves that ridership is predictable over time, which is valuable information for a bike-sharing program that is trying to determine how many bikes a metropolitan area will need, and at what time of day customers will most often be riding. The greatest indicator for volume of rides (based on time of day) was whether it was a weekday or a weekend. This makes sense, as people tend to commute early in the morning on weekdays and sleep in on the weekends. People also tend to run errands or go to various outings on weekend afternoons, whereas on weekday afternoons many people are at work. That being said, other factors are also important to ride volume, including season and weather. Unsurprisingly, Summer had the most rides, while Winter had the least. Additionally, ridership dropped slightly when the weather was poor (cold temperatures or high winds), but the difference between ridership on cloudy hours versus thunderstorm hours was insignificant. This suggests that many people may still prefer to ride bicycles in poor weather rather than purchase an uber, walk, or take the London Tube.

One interesting aspect of bike-sharing that this dataset did not include is the location(s) of where rides begin and end. This is a complex problem for bike-sharing platforms to solve because they want to maximize bike availability for their members, while also minimizing the cost of maintaining their fleet of bikes. In future research, this area would be worth looking into, especially if forecasting ridership is of interest.