Evaluating the Effect of Weather Conditions on Bike Sharing Patterns in Washington D.C.

INF2190H – Data Analytics: Introduction, Methods, and Practical Approaches

December 06, 2023

Professor Tegan Rajkumar-Maharaj
Dawson Butler
Ruiying Wang
Amal Mohamed

Abstract

This research study investigates how meteorological parameters, specifically temperature, feeling temperature, humidity, and windspeed affect the use of Washington D.C.'s renowned public bike-sharing system, Capital Bikeshare. The purpose of this research is to assess whether there is a statistically significant relationship between each meteorological factor and the type of bike rental. Specifically, we analyze how casual and registered bike renters respond to changes in weather using a Kruskal-Wallis H test, along with regression methods to both model the effect of weather has on the number of bike rentals and to predict future bike-share trends. We find that bike rentals are most popular during the summer and fall months. Additionally, registered bikers use Capital Bikeshare as a transit alternative, primarily to commute to and from work, as opposed to casual bike renters that mainly use the service recreationally.

1. Introduction

Bike share programs, also referred to as bicycle-sharing systems, are innovative transit services exclusive to large metropolitan cities, offering low-cost, short-term bike rentals to citizens around the world (Kabra et al., 2018). These programs have not only revolutionized urban transportation by increasing transit accessibility but have also contributed to better public health outcomes and the development of more eco-friendly neighbourhoods (Gebhart & Noland). Prior to the development of Citi Bike, New York's public bicycle sharing system, Washington D.C. was home to the largest bike sharing service in the United States. Capital Bikeshare, commonly known as CaBi, started with over 300 docking stations and 2,500 bicycles (Buehler and Hamre, 2014), and has now expanded to more than 700 stations and 6,000 bicycles (Capital Bikeshare, 2023).

Despite its popularity, public bike system usage experiences considerable fluctuations hourly, daily, monthly and yearly due to several different factors. Among these factors, weather conditions make up "eighty percent of the daily demand fluctuations" (Kumar, 2021). Essentially, people are more hesitant to ride bicycles in adverse weather conditions like extreme cold, humidity, rain, and snowfall (Hanson and Hanson, 1977).

Similar to pre-existing studies on bike-sharing trends in Washington D.C., specifically, Juran Zhang, Qianfan Luo, and Yuntian Shen's comparative analysis "Bike Sharing in Seoul and D.C.", this research paper also aims to investigate how variations in weather conditions, particularly changes in meteorological parameters such as temperature, feeling temperature, windspeed and humidity, affect bike share usage in Washington D.C. To accomplish this, our project uses the Bikeshare dataset sourced from 'An Introduction to Statistical Learning', by Gareth James, Daniela Witten, Trevor Hastie, & Robert Tibshirani'. The dataset contains an hourly and daily count of rental bikes from Capital Bikeshare in Washington, D.C, between 2011 and 2012. The dataset categorizes bikers into two groups: registered bikers and casual bikers. Furthermore, the dataset contains several variables, including weather variables, date and time variables, and data related to seasons and holidays. Summary statistics for our independent and dependent variables are shown in Table 1 and will be further explored in section 3.1.

Although the primary focus of this research paper is to better understand the relationship between bike share patterns and variations in weather, it is imperative to also consider the role of different bike users. As mentioned earlier, Capital Bikeshare categories bikers into two groups: registered and casual bikers. It is likely that this distinction is indicative of differences in bike usage patterns. For example, registered bikers who may heavily rely on bike-share as their main mode of transportation, might show adaptability in the face of adverse weather conditions. However, casual bikers who use bike share systems recreationally might be more affected by bad weather. To help visualize these behavioral patterns, Figure 1 shows the differences in overall bike trends in 2011 and 2012 between registered and casual bikers; a point of interest this research aims to further investigate. As such, our research is guided by the following questions:

Research Question 1: What is the relationship between weather variables and bike usage patterns?
Research Question 2: How do variations in weather affect bike usage patterns between registered and casual bikers?

By addressing these questions, our research aims to gain insight into the various factors causing fluctuations in public bike share systems.

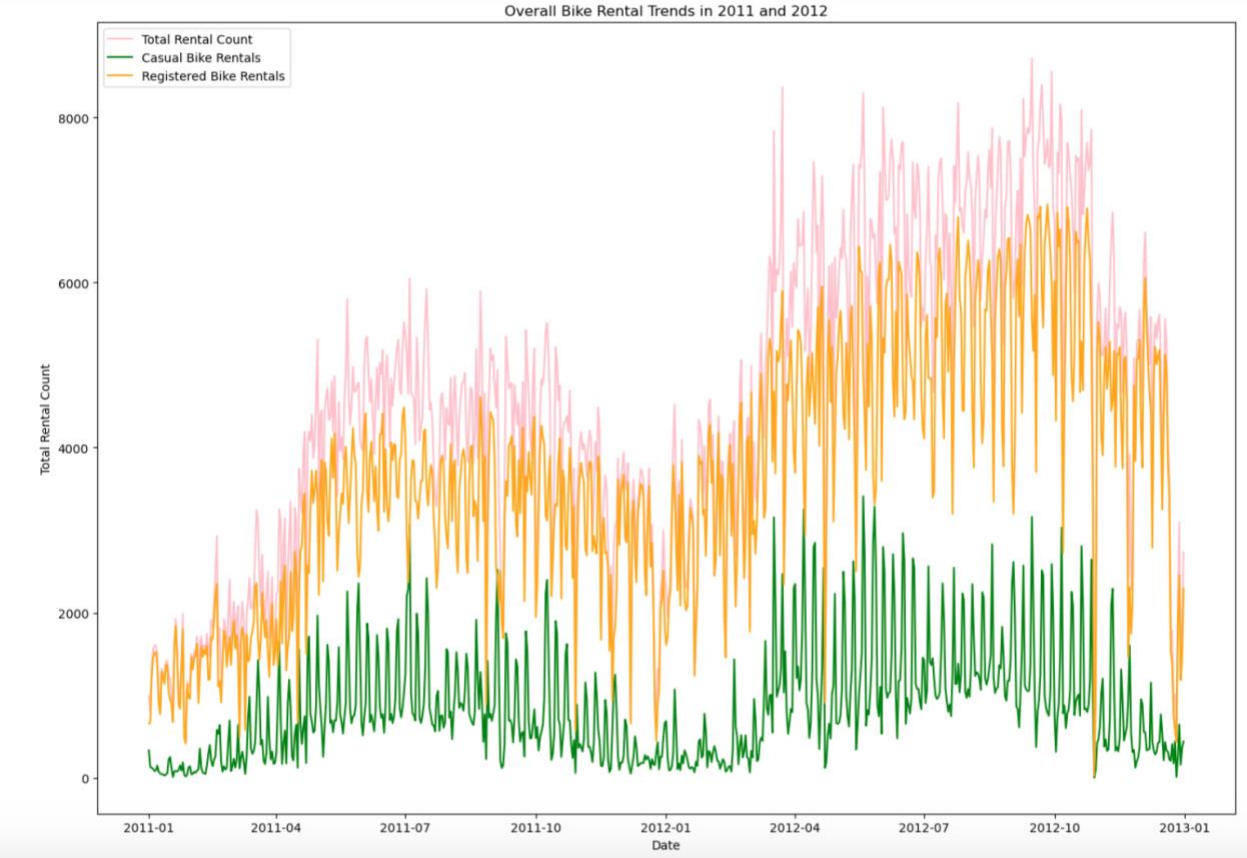


Figure 1: Overall Bike Trends in 2011 and 2012

Table 1: Descriptive statistics for bike rental categories and weather variables

Variable	Mean	Standard Deviation	Minimum	Maximum
Temperature	15.28°C	8.60	-5.22°C	32.50°C
Feeling Temperature	15.31°C	10.76	-10.78°C	39.50°C
Humidity	62.79%	14.24	0.00%	97.25%
Windspeed	12.76 km/h	5.19	1.50 km/h	34.00 km/h
Casual Bike Rentals	848.18	686.62	2	3410
Registered Bike Rentals	3656.17	1560.26	20	6946
Total Bike Rentals	4504.35	1937.21	22	8714

2. Methods

2.1 Dataset Description

The Bikeshare dataset is sourced from 'An Introduction to Statistical Learning', by Gareth James, Daniela Witten, Trevor Hastie, & Robert Tibshirani'. The data originates from Capital Bikeshare's historical records on bike usage patterns in Washington D.C. from 2011 to 2012. It is divided into two separate datasets: one accounting for the total number of bike rentals per day and the other for the total number of bike rentals per hour.

There are a total of 14 variables (Table 2), but for the purpose of our research, we focus on total, casual and registered bike rental patterns, in relation to meteorological parameters like temperature, 'feels like' temperature, humidity, and wind speed. As shown in Table 2, the dataset includes an ordinal variable named 'weather conditions', which is ranked 1 to 4 (clear skies, misty, light rain/snow, and heavy rain/snow). In this study, when we refer to 'weather', we are referencing the measurable meteorological parameters mentioned above as opposed to weather conditions. This is because the aim of our research is to understand the quantifiable effect of weather on bike patterns.

Lastly, it is important to note that while the remaining variables, such as holidays, seasons, and working days provide relevant context to our research, overall, they play a subordinate role in our analysis.

Table 2: 'Bikeshare' dataset variable information

Variable	Type	Description
dteday	Date	Date
season	Categorical	1: Spring, 2: Summer, 3: Autumn, 4: Winter
yr	Categorical	0: 2011, 1: 2012
mnth	Numerical	Month (1 to 12)
wkday	Categorical	Day of the Week
workingday	Binary	Working day = 1 Weekend/Holiday = 0
temp	Continuous	Normalized temperature °C
atemp	Continuous	Normalized feeling temperature °C
hum	Continuous	Normalized humidity
weathersit	Categorical	1: Clear, 2: Misty/Cloudy, 3: Light Rain
windspeed	Continuous	Normalized wind speed
casual	Numerical	Count of casual users
registered	Numerical	Count of registered users
count	Numerical	Count of total rental bikes including both casual and registered

2.2 Data Cleaning

Prior to conducting exploratory data analysis and feature engineering, we pre-processed and cleaned both datasets using several packages, most notably, pandas and numpy in Python and dplyr in R (See Appendix A). This process involved checking for missing values, converting variables to categorical, changing the date variable from an object to a floating-point number, renaming variables, assigning labels to ranges (instead of 1 putting 'winter' for season variable), and de-normalizing the meteorological parameters.

2.3 Statistical Modelling

In addition to our research questions, we hypothesize the following:

Hypothesis 1: As weather conditions become more unfavorable, the total count of bikeshare rentals will decrease.

Hypothesis 2: Adverse weather conditions have more of an impact on casual bikeshare rentals, as opposed to registered bikeshare rentals.

The statistical modelling process started with determining whether our datasets met the assumptions required for parametric statistical testing. Given our objective to evaluate whether there is a statistically significant difference in the means between multiple variables, simultaneously, our primary goal was to check ANOVA assumptions. This includes checking for normality, homogeneity of variances, and independence.

Firstly, we checked for normality by plotting histograms to illustrate the distribution of total rental bikes relative to the following categorical variables: seasons and meteorological parameters. As displayed in Figure 2, there are no "bell shape curves", meaning the data is not normally distributed. The skewed data informs us that the first assumption of ANOVA is not met. Secondly, we checked for homogeneity of variances by creating box plots. Figure 3 shows a series of box plots for the same variables displayed In Figure 2. The box plots show differences in variance across the variables, particularly how different seasons and meteorological parameters affect the total number of bike rentals. Given that both the normality and equal variance assumptions were violated, this research performs the Kruskal-Wallis H-test instead: a non-parametric rank-based test used to check whether statistically significant differences exist among multiple groups of an independent variable against a continuous/ordinal dependent variable. The result of this test is further discussed in the following section.

Lastly, we perform separate linear regression models to learn more about the relationship between each meteorological parameter and each bike rental category: total bike rentals, registered bike rentals, and casual bike rentals. This allows us to better understand how each category reacts to different weather parameters, and ultimately which weather variable most significantly impacts bike rental patterns within the different groups. Due to the non-normal distribution of our datasets, regression analysis proved to be the most complementary model. Next, we perform multiple linear regression to evaluate if all four weather measurements have a relative effect on bike rentals.

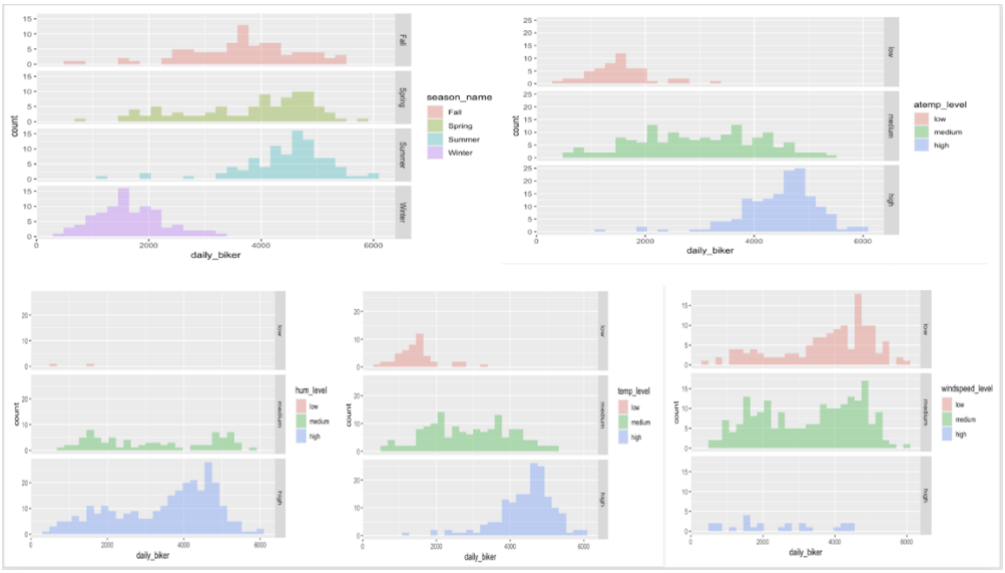


Figure 2: Histograms showing the distribution of total bike rentals across different seasons and meteorological parameters

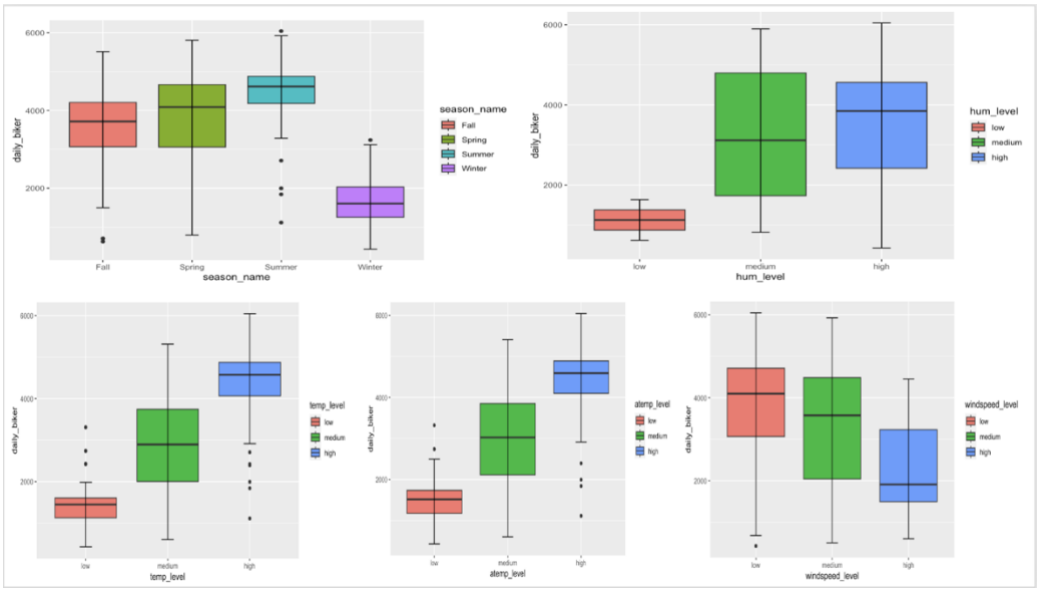


Figure 3: Boxplots showing how different seasons and meteorological parameters affect the total number of bike rentals

3. Results and Discussion

3.1 Exploratory Data Analysis

To gain a holistic overview of our datasets, we perform exploratory data analysis on the following sets of variables: time variables (month, days, hours) and weather variables (seasons and meteorological parameters). The plots below provide insight into how the different variables behave and interact.

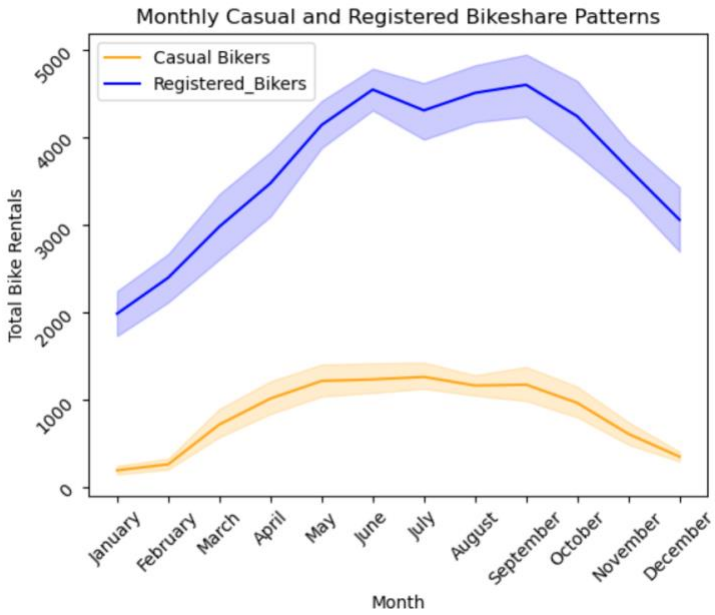


Figure 4: Line graph showing monthly bike rental use for registered and casual bikers

The comparison of casual and registered monthly bike rental patterns in Figure 4 shows that during the warmer months, between July and September, there is a peak bike share usage. Despite more registered bikers participating in the bike-sharing program, this trend still provides similar for casual bikers. However, casual bike usage remains relatively low throughout the year and does not have any major peaks, but rather a gradual increase and decrease. Figure 4 suggests that major peaks and pits in bike usage for registered bikers can be an indication of a strong influence by seasonal changes.

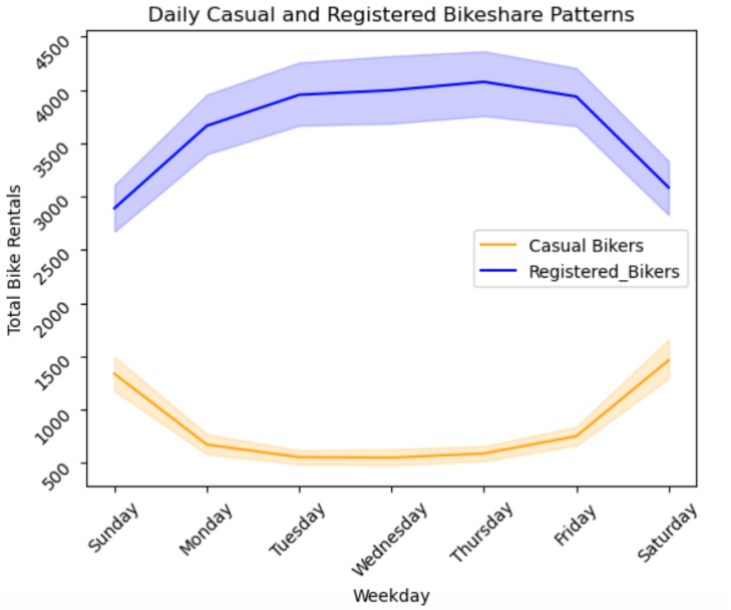


Figure 5: Line graph showing daily bike rental use for registered and casual bikers

Figure 5 shows an interesting relationship between casual and registered bike rentals across the week. It is clear that registered bikers make the most use of Capital Bikeshare during workdays, with a gradual increase in usage between Monday and Friday. In contrast, casual bikers rent bikes mostly on Saturday and Sunday (See Appendix B.1 for a detailed breakdown). This pattern provides some evidence that casual bikers tend to use Capital Bikeshare strictly recreationally.

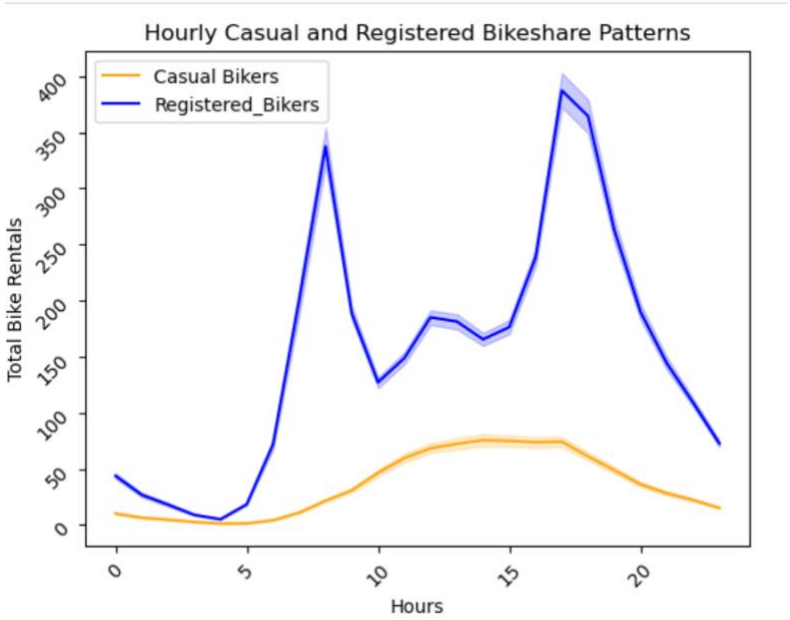


Figure 6: Line graph showing hourly bike rental use for registered and casual bikers

In Figure 6, we see a bimodal distribution on the registered bike rentals line plot, with a peak around 7am and a second peak around 5-6pm. Registered bike rental patterns follow the same commute pattern trends. In contrast, casual bike rentals occur during work and school hours which supports our initial assumption that casual bikers use the service for leisure.

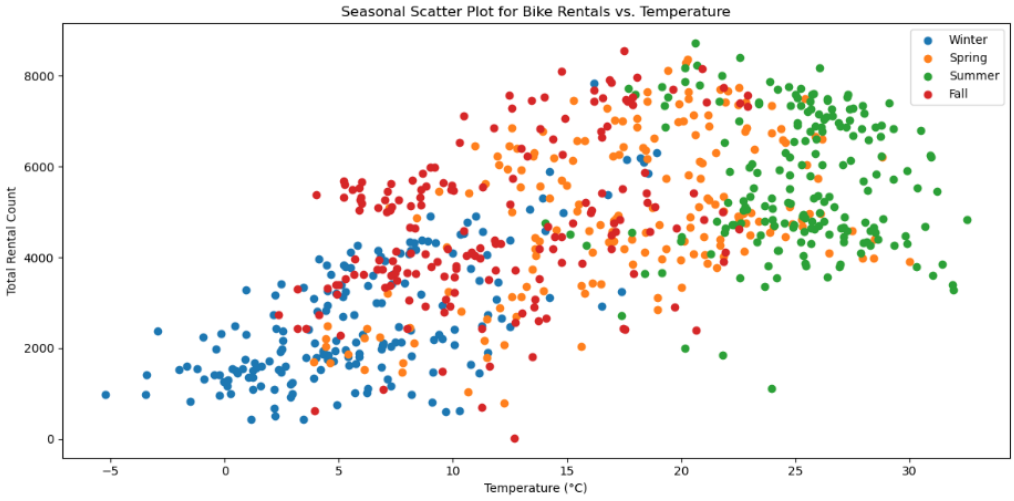


Figure 7: Seasonal scatter plot for bike rentals vs. temperature

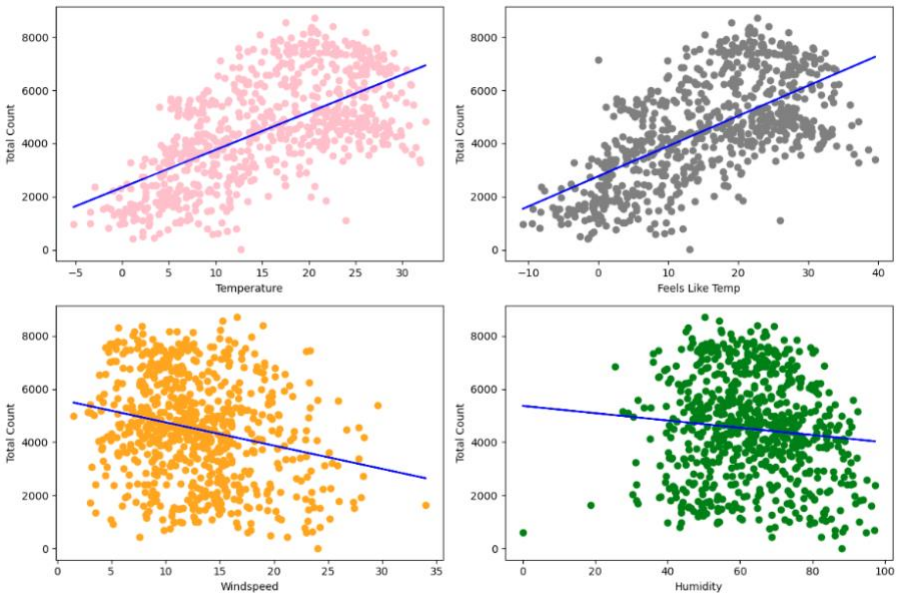


Figure 8: Scatter plot and linear regression line for meteorological parameters

Figures 7 and 8 provide more insight into the behavior of bike rental usage in relation to different seasons and meteorological parameters. In Figure 7 we can see that as we approach the summer months and higher temperatures, as a result, total bike rentals increase. This positive and direct relationship can also be seen with feeling temperature. In contrast, windspeed and humidity have an indirectly proportional relationship with total bike rentals. As windspeed and humidity increase, the total number of bike rentals decreases.

3.2 Kruskal-Wallis H Test

This study uses the Kruskal-Wallis H test due to the non-normally distributed nature of our dataset. Our dataset passes the four fundamental assumptions that underpin this test, which include: dependent variables are ordinal/continuous, independent variables consists of at least two more categorical groups, independence of observation, data distribution must have some type of shape. Consequently, we proceeded to perform the Kruskal-Wallis H test of mean rank, employing a significance level of 0.05. Following the Kruskal-Wallis H test, we conduct a post hoc test, the Dunn's test, to perform pairwise comparisons between each independent level (See Appendix C). To meet the test assumption, we convert the continuous variable including temperature, 'feels' like temperature, and humidity into categorical variable by grouping them into 3 levels, low, medium and high by equal level interval of each group.

3.2.1 Season

The results from the Kruskal-Wallis H test for the seasons variable yielded a significantly small p-value of less than 2.2e-16. This is less than the significance level of 0.05 and suggests that at least one season's mean rank is significantly different from the others. To understand which pairs of groups differed, we conduct a Dunn's test between group levels. The results show a very small, adjusted p-value for all pairs, with the smallest p-value for the Spring and Winter and Summer and Winter pairs (Table 3).

Table 3: R Result for Dunn's Test for Bike Rentals and Seasons

Comparison <chr>	Z <dbl>	P.unadj <dbl>	P.adj <dbl>
Fall - Spring	-1.036189	3.001138e-01	3.001138e-01
Fall - Summer	-4.607618	4.073084e-06	6.109625e-06
Spring - Summer	-3.596210	3.228868e-04	3.874641e-04
Fall - Winter	8.604066	7.694045e-18	1.538809e-17
Spring - Winter	9.714607	2.612541e-22	7.837622e-22
Summer - Winter	13.342305	1.313352e-40	7.880112e-40

3.2.2 Temperature and Feeling Temperature

Results for both the temperature and feeling temperature are quite similar, with both p-values at 2.2e-16, falling beneath the 0.05 significance level. This means that the total number of bike rentals was significantly impacted by both temperature and normal temperature levels. The Dunn's test also revealed a low adjusted p-value, validating that there is a strong relationship between higher temperatures and increased bike rental activity. (See Appendix C)

3.2.4 Windspeed

The Kruskal-Wallis H test for windspeed revealed the lowest p-value yet, 9.649e-06 suggests that at least one wind speed level's mean rank is significantly different from the others, and the negative Z-value and adj. p value from Dunn's test result validating the negative impact of higher windspeed level in the number of bikers. (See Appendix C)

3.2.5 Humidity

In contrast with the other weather variables, humidity levels returned a p-value of 0.0619, which is greater than 0.05, indicating that that the test did not find statistically significant evidence to suggest that the distribution of daily biker counts differ between different humidity levels. While it is worth noting that the p-value is relatively close to 0.05, we use linear regression to further assess this relationship. (See Appendix C)

3.3 Linear Regression

We chose linear and polynomic regression to better understand the relationship between the weather variables and the three different bike rider categories. In addition, using regression analysis benefited our research because of our non-normally distributed data. Finally, through regression plots we can provide a visual representation of the data which is something that can provide an even better understanding of the data.

Figure 9 shows the weather regression analysis for casual bikers. Here we can see that there is a strong direct relationship between both types of temperature and casual bike rentals. There is a positive correlation between the two variables, with the strongest relationship found at medium temperature values. Wind Speed is also highly correlated, but not linearly, and has a negative relationship. Humidity has a parabolic relationship, with the highest quantities being found at medium points.

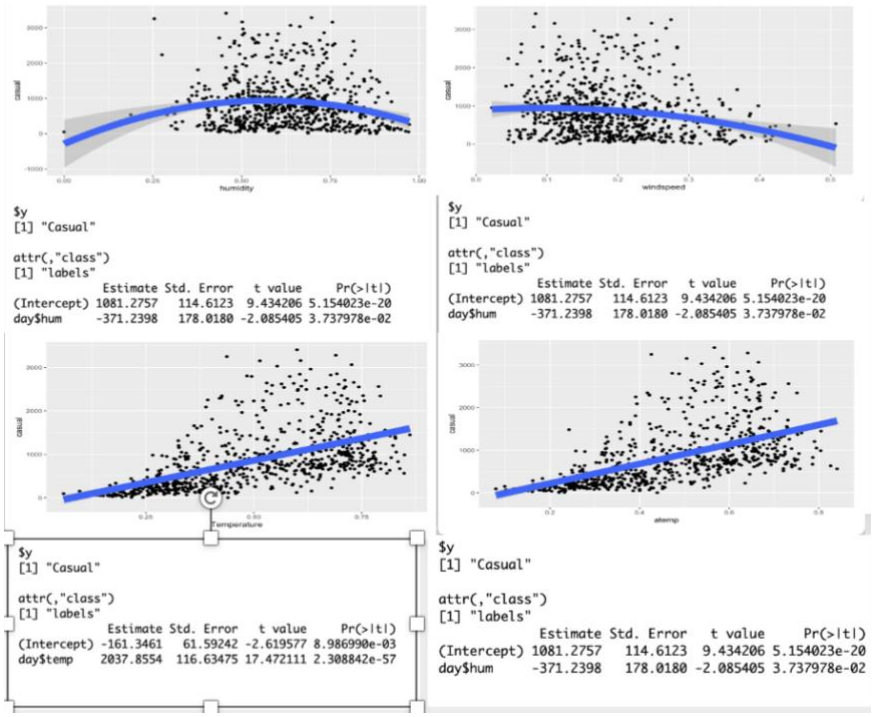


Figure 9: Linear and polynomic regression results and plots showing the effect of the four main weather variables on casual bike rentals

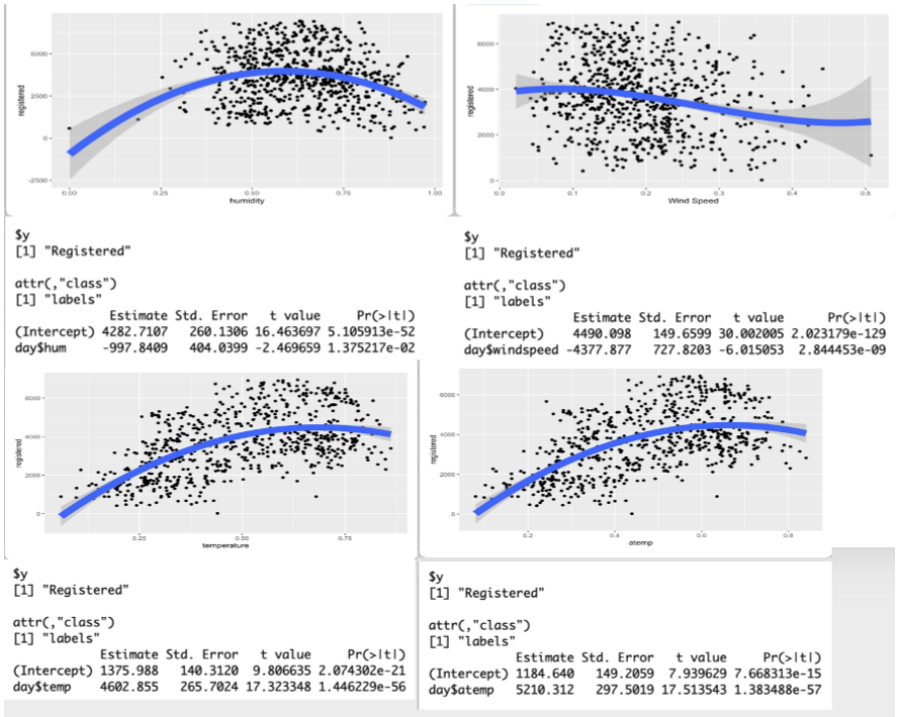
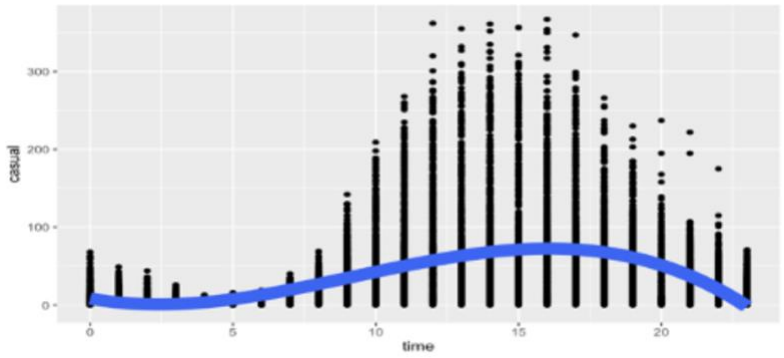


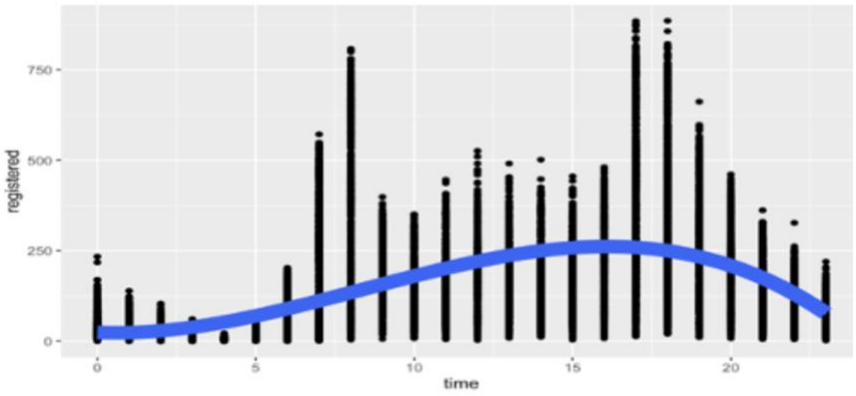
Figure 10: Linear and polynomic regression results and plots showing the effect of the four main weather variables on registered bike rentals

In Figure 10 we can see a polynomic and parabolic relationship with humidity, feeling temperature, and temperature. Similarly, there is also a positively correlated relationship with these variables and registered bike rentals. The three plots show that the highest values are found in the median points at the plots, and the lowest are found at the extremes. In contrast, windspeed is somewhat negatively correlated and has a strong polyrhythmic shape. There is a negative relationship between higher windspeeds and registered bike rentals.




```
$y
[1] "Count"

attr(,"class")
[1] "labels"
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept)  10.8761 0.69421964 15.66666 6.075926e-55
hour$hr       2.1478 0.05158189 41.63865 0.000000e+00
```

Figure 11: Time regression analysis plot for casual bikers




```
$y
[1] "registered"

attr(,"class")
[1] "labels"
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 59.219064   2.072596 28.57241 1.653843e-175
hour$hr      8.189992   0.153998 53.18246  0.000000e+00
```

Figure 12: Time regression analysis plot for registered bikers

The time regression analysis results show that bike rentals could be found at all hours but were most common in the middle of the day and rarest in the early morning hours. Casual bike rentals are mostly found in the middle of the day. In contrast, registered bike rentals had the most variation throughout the day, with most rentals occurring in early morning and early evening, most commonly at 6 am and 5-6 pm.

3.3 Multiple Linear Regression

The final test of this research is multiple linear regression, which provides a way to determine which weather variable had the greatest overall effect on the different bike rental categories. Prior to performing regression on our actual model, we developed a prediction model through additional data from Washington D.C.'s bike-sharing system to predict the overall expected counts of bike rental usage. The data starts on January 1st, 2012, and lasts until January 1st, 2014. In Figure 13 we can see a parabolic shape form as the years progress, showing higher bike rental counts occurring in the spring and summer months. As we can see in Figure 14, the actual model plotted matches yearly parabolic shape of the previous plot, supporting the model's accuracy.

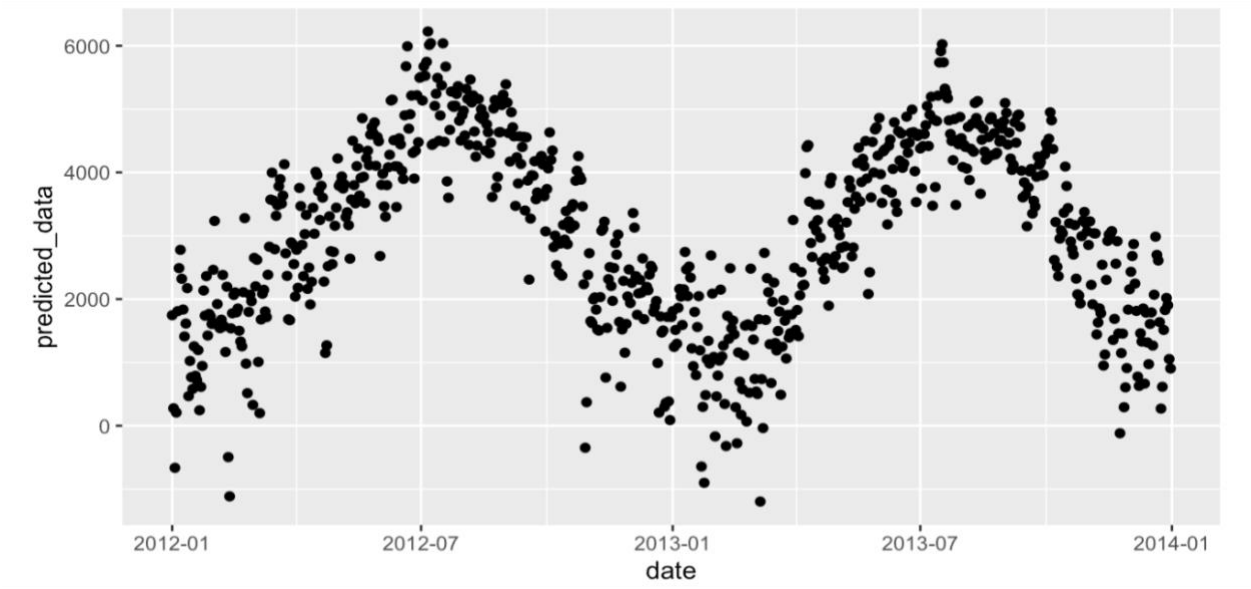


Figure 13: Plot predicting the counts of bike share riders during different dates from different months from 2012 to 2014.

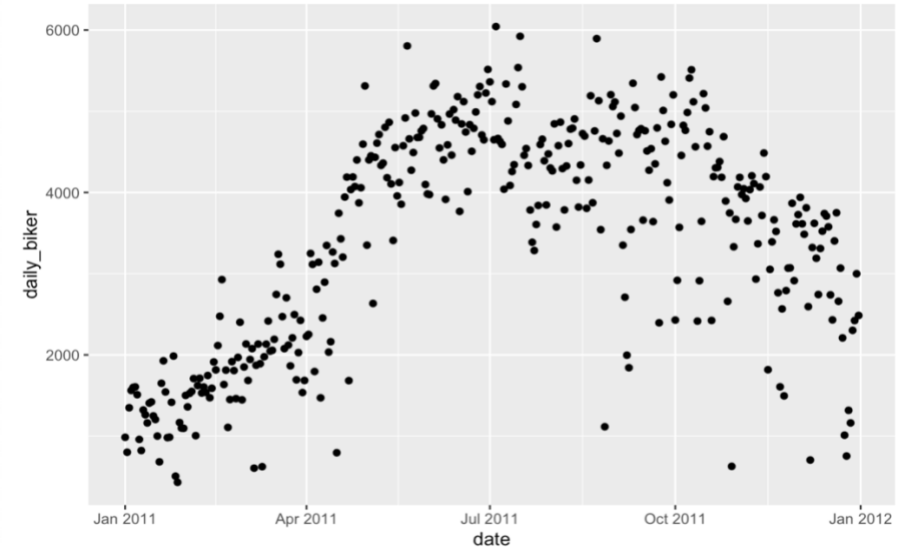


Figure 14: Plot showcasing the actual count of bikers from the Washington bikeshare system in 2011

Next, we perform assumption checks to examine the normality of the weather variable residuals. In Figure 15 we can see how the different residuals compare. We also see that the Q-Q plot shows a near-linear relationship, suggesting collinearity among the data. Finally, we see homoscedasticity in the last plot. Therefore, the conditions are met, ensuring that a multiple linear regression model is possible for the data.

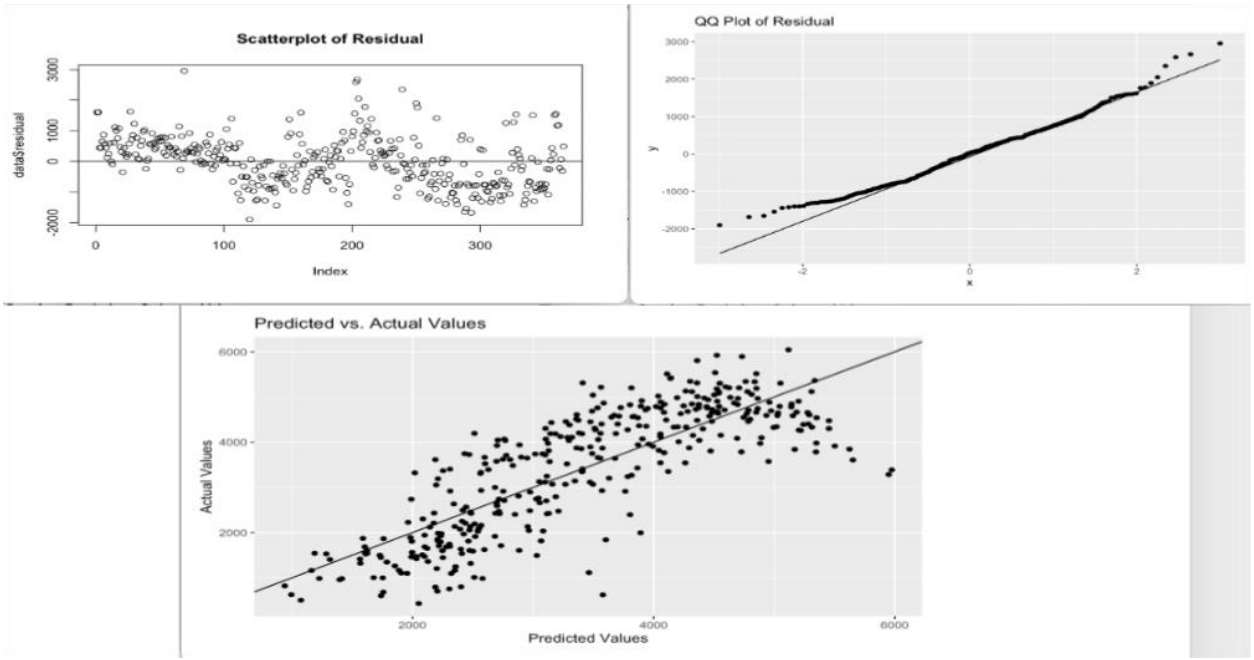


Figure 15: Plots examining the normality of weather variable residuals

The results of our multiple linear regression analysis for casual bikers showed that all four weather variables strongly affect the overall number of casual bikers (See Appendix D). Casual bike rentals were more sensitive to temperature, 'feels' like temperature, and windspeed. In addition, there is a negative relationship between wind speed and humidity, which suggests that the number of casual bikers decreases with elevated wind speed and humidity. The results for registered bikers were a high sensitivity towards temperature, relative to all other variables, with a positive correlation. Some sensitivity towards feeling temperature and wind speed and low sensitivity towards humidity, which suggests a limited effect. Humidity and wind speed have a negative correlation, suggesting that as they increase, registered bikers decrease. A parabolic relationship exists for the temperature and feeling temperature, so quantification may not be perfect in this model. Overall, the indications of the regression analysts suggest that the fundamental use of the bikes from registered versus casual users was fundamentally different between the registered and casual bikers. The regression analysis suggested that registered bikers used bikes at times and dates similar to how people use cars, suggesting that it was a mode of transportation used to go to and from work. In contrast, the casual bikers were greatly affected by weather and were only found during specific dates and times; usually the middle of a weekend or holiday, suggesting that the fundamental use was for leisure.

4. Limitations and Conclusion

The main limitations of the dataset were primarily how the data was measured and its distribution. The dataset was not normally distributed, specifically for the following variables: count, registered, casual, hour, and day. As a result, the data failed the ANOVA test resulting in the use of a Kruskal-Wallis H test. This became commonplace within the data, as tests had to be carefully selected to ensure they could be used. Furthermore, tests had to be canceled due to being impossible to produce, specifically with algorithms and logistic regression algorithms, as the categories of interest came from count measurements rather than categorical or outcome-based measurements, which was what principally those types of tests were used for in the past. Finally, the dataset had several different measurements of possible variables affecting the amount of bike-share riders per day and hour, which all had some influence on the results, making it difficult to quantify in a single test and difficult to quantify the exact relationship that existed between the number of bikers and the variables of interest. In addition, the two types of potential variables they measured, weather and date, were limited in that the dataset was incomplete when measuring the selected variables for those two categories, and to get a complete and accurate picture of the dataset, more variables would be needed to be measured including factors like precipitation. One final issue was an error in the weather situation dataset. It was missing a category in the CSV conversion, making it impossible to work with and forcing that category to be excluded.

Overall, the results from our research regarding the Washington D. C. bike-share system show that weather plays an integral role in bike usage, with warmer temperatures yielding higher bike rentals and adverse weather conditions such as higher windspeed and humidity impacting bike rental usage. Registered bikers predominantly use Capital Bikeshare, particularly during commuting hours within the work week. While registered bikers use the service as a means of transportation, casual bikers appear to use it recreationally. This is provided when assessing casual bike rental counts with the days of the week and seeing the highest peaks on Saturdays and Sundays. Results from this research are important when addressing how to integrate successful

bike-sharing programs within metropolitan cities. Bike-share programs are becoming increasingly popular in the mix of a climate crisis and governments looking for ways to reduce the fossil fuel crisis around the globe (Ma et al., 2018). Bikes represent an environmentally friendly, non-fossil, fuel-based mode of transportation that almost anyone can use for a low price. Bike-share programs are a means to facilitate the distribution of these bikes to people who may not bike often enough to buy their own, for people who cannot afford their own bikes, or for tourists or guests within a village (Kabra et al., 2018). Due to the success of Washington D. C.'s Capital Bikeshare, we conclude that bike-share systems represent a way for cities to reduce fossil fuels by offering an environmentally friendly, cheap mode of transportation.

References

Kabra, Ashish, Belavina, Elena, Girotra, Karan. (2015). Bike-Share Systems: Accessibility and Availability. *Management Science* 66(9): 3803-3824. https://doi.org/10.1287/mnsc.2019.3407

Ma, Xinwei, Ji, Yanjie, Yang, Mingyuan, Jin, Yuchuan, Tan, Xu. (2018). Understanding bikeshare mode as a feeder to metro by isolating metro-bikeshare transfers from smart card data. *Transport Policy* 71(2018): 57-69. https://doi.org/10.1016/j.tranpol.2018.07.008

Scott, Darren M. & Ciuro, Celenna. (2019). What factors influence bike share ridership? An investigation of Hamilton, Ontario's bike share hubs. *Travel Behaviour and Society* 16(2019): 50-58. https://doi.org/10.1016/j.tbs.2019.04.003

Zhang, Juran, Luo, Qianfan, Shen, Yuntian. (2022). Bike Sharing in Seoul and D.C. *CMU education.* https://www.stat.cmu.edu/capstoneresearch/fall2022/315files_f22/team12.html

Buehler, R., & Hamre, A. (2014). Economic benefits of capital bikeshare: A focus on users and businesses (No. VT-2013-06). Mid-Atlantic Universities Transportation Center. https://capitalbikeshare.com/how-it-works

Gebhart, K., & Noland, R. B. (2014). The impact of weather conditions on bikeshare trips in Washington, DC. Transportation, 41, 1205-1225.

Kumar, D. (2021). Meteorological barriers to bike rental demands: a case of Washington DC using NCA approach. Case Studies on Transport Policy, 9(2), 830-841.

Hanson, S., & Hanson, P. (1977). Evaluating the impact of weather on bicycle use. Transportation Research Record, 629, 43-48.

Visual Crossing. (2023). Washington, DC, USA, 2012-01-01-2013-12-31. *Visual Crossing Weather*. Retrieved December 3rd, 2023. https://www.visualcrossing.com/weather/weather-data-services/Washington,DC,USA/metric/2012-01-01/2023-12-02

Appendix

Appendix A: Data Pre-processing and Cleaning

In [5]:
```
#check/handling missing values for day

day.isnull().sum()
```

Out[5]:
```
instant       0
dteday        0
season        0
yr            0
mnth          0
holiday       0
weekday       0
workingday    0
weathersit    0
temp          0
atemp         0
hum           0
windspeed     0
casual        0
registered    0
cnt           0
dtype: int64
```

In [6]:
```
#check/handling missing values for hour

hour.isnull().sum()
```

Out[6]:
```
instant       0
dteday        0
season        0
yr            0
mnth          0
hr            0
holiday       0
weekday       0
workingday    0
weathersit    0
temp          0
atemp         0
hum           0
windspeed     0
casual        0
registered    0
cnt           0
dtype: int64
```

Data Cleaning and Pre-processing

In [2]:
```
#read csv files for day and hour bikeshare data

day = pd.read_csv("day.csv")
hour = pd.read_csv("hour.csv")
```

In [3]:
```
#view both dataframes and get a sense of where tidy data principles need to be applied

day.head()
```

Out[3]:

	instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
0	1	2011-01-01	1	0	1	0	6	0	2	0.344167	0.363625	0.805833	0.160446	331	654	985
1	2	2011-01-02	1	0	1	0	0	0	2	0.363478	0.353739	0.696087	0.248539	131	670	801
2	3	2011-01-03	1	0	1	0	1	1	1	0.196364	0.189405	0.437273	0.248309	120	1229	1349
3	4	2011-01-04	1	0	1	0	2	1	1	0.200000	0.212122	0.590435	0.160296	108	1454	1562
4	5	2011-01-05	1	0	1	0	3	1	1	0.226957	0.229270	0.436957	0.186900	82	1518	1600

In [4]:
```
hour.head()
```

Out[4]:

	instant	dteday	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
0	1	2011-01-01	1	0	1	0	0	6	0	1	0.24	0.2879	0.81	0.0	3	13	16
1	2	2011-01-01	1	0	1	1	0	6	0	1	0.22	0.2727	0.80	0.0	8	32	40
2	3	2011-01-01	1	0	1	2	0	6	0	1	0.22	0.2727	0.80	0.0	5	27	32
3	4	2011-01-01	1	0	1	3	0	6	0	1	0.24	0.2879	0.75	0.0	3	10	13
4	5	2011-01-01	1	0	1	4	0	6	0	1	0.24	0.2879	0.75	0.0	0	1	1

In [8]:
```
#noticing that the dteday variable, which stands for daytime
#is an object which is not helpful for this analysis

#converting dteday using datetime() function for both day and hour dataframes

day['dteday'] = pd.to_datetime(day['dteday'])
hour['dteday'] = pd.to_datetime(hour['dteday'])

#view change
print(day[['dteday']].head())
print(hour[['dteday']].head())
```

```
      dteday
0 2011-01-01
1 2011-01-02
2 2011-01-03
3 2011-01-04
4 2011-01-05
      dteday
0 2011-01-01
1 2011-01-01
2 2011-01-01
3 2011-01-01
4 2011-01-01
```

In [9]:
```
#correcting variables for day --> converting to categorical

day["season"] = day["season"].astype('category')
day["yr"] = day["yr"].astype('category')
day["mnth"] = day["mnth"].astype('category')
day["holiday"] = day["holiday"].astype('category')
day["workingday"] = day["workingday"].astype('category')
day["weathersit"] = day["weathersit"].astype('category')
day["weekday"] = day["weekday"].astype('category')
```

In [10]:
```
#correcting variables for hour --> converting to categorical

hour["season"] = hour["season"].astype('category')
hour["yr"] = hour["yr"].astype('category')
hour["mnth"] = hour["mnth"].astype('category')
hour["holiday"] = hour["holiday"].astype('category')
hour["workingday"] = hour["workingday"].astype('category')
hour["weathersit"] = hour["weathersit"].astype('category')
hour["weekday"] = hour["weekday"].astype('category')
```

In [8]:
```
#noticing that the dteday variable, which stands for daytime
#is an object which is not helpful for this analysis

#converting dteday using datetime() function for both day and hour dataframes

day['dteday'] = pd.to_datetime(day['dteday'])
hour['dteday'] = pd.to_datetime(hour['dteday'])

#view change
print(day[['dteday']].head())
print(hour[['dteday']].head())
```

```
      dteday
0 2011-01-01
1 2011-01-02
2 2011-01-03
3 2011-01-04
4 2011-01-05
      dteday
0 2011-01-01
1 2011-01-01
2 2011-01-01
3 2011-01-01
4 2011-01-01
```

In [9]:
```
#correcting variables for day --> converting to categorical

day["season"] = day["season"].astype('category')
day["yr"] = day["yr"].astype('category')
day["mnth"] = day["mnth"].astype('category')
day["holiday"] = day["holiday"].astype('category')
day["workingday"] = day["workingday"].astype('category')
day["weathersit"] = day["weathersit"].astype('category')
day["weekday"] = day["weekday"].astype('category')
```

In [10]:
```
#correcting variables for hour --> converting to categorical

hour["season"] = hour["season"].astype('category')
hour["yr"] = hour["yr"].astype('category')
hour["mnth"] = hour["mnth"].astype('category')
hour["holiday"] = hour["holiday"].astype('category')
hour["workingday"] = hour["workingday"].astype('category')
hour["weathersit"] = hour["weathersit"].astype('category')
hour["weekday"] = hour["weekday"].astype('category')
```

In [16]:
```
#renaming hour dataframe columns for clarity

hour = hour.rename(columns={
    'dteday': 'Date',
    'season': 'Season',
    'yr': 'Year',
    'mnth': 'Month',
    'hr': 'Hour',
    'holiday': 'Holiday',
    'weekday': 'Weekday',
    'workingday': 'Working_Day',
    'weathersit': 'Weather_Conditions',
    'temp': 'Temp',
    'atemp': 'Feeling_Temp',
    'hum': 'Humidity',
    'windspeed': 'Windspeed',
    'casual': 'Casual_Bikers',
    'registered': 'Registered_Bikers',
    'cnt': 'Total_Bikers'
})
```

In [17]:
```
#replacing numbers in categorical variables with the actual category name

#day
day['Season'] = day['Season'].replace([1, 2, 3, 4], ["Winter", "Spring", "Summer", "Fall"])
day['Month'] = day['Month'].replace([1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12], ["January","February", "March", "April","May","June", "July", "August","September","October", "November", "Decem
day['Holiday'] = day['Holiday'].replace([0, 1], ["no", "yes"])
day['Weekday'] = day['Weekday'].replace([1, 2, 3, 4, 5, 6, 0], ["Monday","Tuesday", "Wednesday", "Thursday", "Friday", "Saturday","Sunday"])
day['Working_Day'] = day['Working_Day'].replace([0, 1], ["no", "yes"])
day['Weather_Conditions'] = day['Weather_Conditions'].replace([1, 2, 3], ["Clear to Partly Cloudy", "Misty to Cloudy", "Light Rain/Snow"])
```

In [18]:
```
#hour
hour['Season'] = hour['Season'].replace([1, 2, 3, 4], ["Winter", "Spring", "Summer", "Fall"])
hour['Month'] = hour['Month'].replace([1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12], ["January","February", "March", "April","May","June", "July", "August","September","October", "November", "Dec
hour['Holiday'] = hour['Holiday'].replace([0, 1], ["no", "yes"])
hour['Weekday'] = hour['Weekday'].replace([1, 2, 3, 4, 5, 6, 0], ["Monday","Tuesday", "Wednesday", "Thursday", "Friday", "Saturday","Sunday"])
hour['Working_Day'] = hour['Working_Day'].replace([0, 1], ["no", "yes"])
hour['Weather_Conditions'] = hour['Weather_Conditions'].replace([1, 2, 3], ["Clear to Partly Cloudy", "Misty to Cloudy", "Light Rain/Snow"])
```

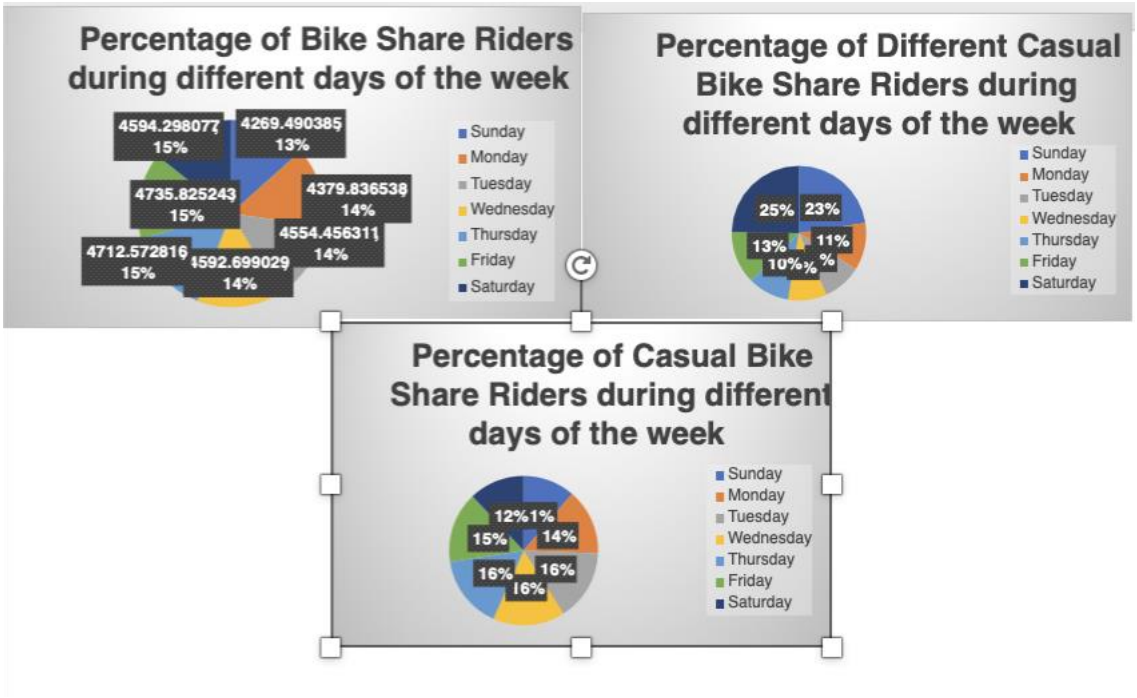
In [21]:

```
#convert temp, feeling temp, humidity, windspeed into actual

#denormalize

#temperature
hour['Temp'] = hour['Temp'] * (39 - (-8)) + (-8)
day['Temp'] = day['Temp'] * (39 - (-8)) + (-8)

#atemp
hour['Feeling_Temp'] = hour['Feeling_Temp'] * (50 - (-16)) + (-16)
day['Feeling_Temp'] = day['Feeling_Temp'] * (50 - (-16)) + (-16)

#humidity
hour['Humidity'] = hour['Humidity'] * 100
day['Humidity'] = day['Humidity'] * 100

#windspeed
hour['Windspeed'] = hour['Windspeed'] * 67
day['Windspeed'] = day['Windspeed'] * 67
```

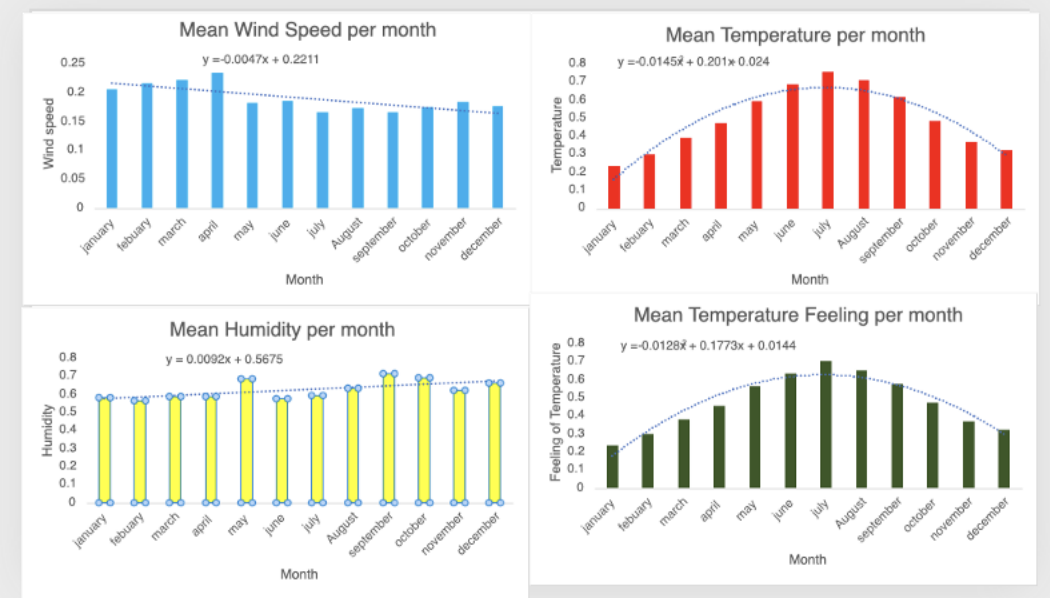
```r
#aggregate bike rental counts and average weather variables by day; add new columns for labeling holiday/workingday/season
daily_data <- bikeshare %>%
  group_by(day) %>%
  summarise(daily_biker = sum(bikers),
            daily_casual=sum(casual),
            daily_registered=sum(registered),
            avg_temp = mean(temp),
            avg_atemp=mean(atemp),
            avg_hum = mean(hum),
            avg_windspeed = mean(windspeed),
            season=max(season),
            month=first(mnth),
            holiday=first(holiday),
            weekday=first(weekday),
            workingday=first(workingday))

#level interval is determined by (max-min)/3
#temp_level: low <0.26, medium 0.26~0.52,high >0.52
daily_data$temp_level<-cut(daily_data$avg_temp,breaks=c(-Inf,0.26,0.52,Inf),labels=c("low","medium","high"))

#atemp_level: low <0.26, medium 0.26~0.52,high >0.52
daily_data$atemp_level<-cut(daily_data$avg_atemp,breaks=c(-Inf,0.26,0.52,Inf),labels=c("low","medium","high"))

#hum_level: low <0.26, medium 0.26~0.52,high >0.52
daily_data$hum_level<-cut(daily_data$avg_hum,breaks=c(-Inf,0.26,0.52,Inf),labels=c("low","medium","high"))

#windspeed_level: low <0.16, medium 0.16~0.32,high >0.32
daily_data$windspeed_level<-cut(daily_data$avg_windspeed,breaks=c(-Inf,0.16,0.32,Inf),labels=c("low","medium","high"))

#label the categorical value
daily_data$season_name <-with(daily_data,ifelse(season==1,'Winter',ifelse(season==2,'Spring',ifelse(season==3,'Summer','Fall'))))
daily_data$holiday_name <-with(daily_data,ifelse(holiday==1,'Holiday',"Non-holiday"))
daily_data$workingday_name <-with(daily_data,ifelse(workingday==1,'Working day',"Non-WorkingDay"))
```

Appendix B: Additional Visualizations from Exploratory Data Analysis

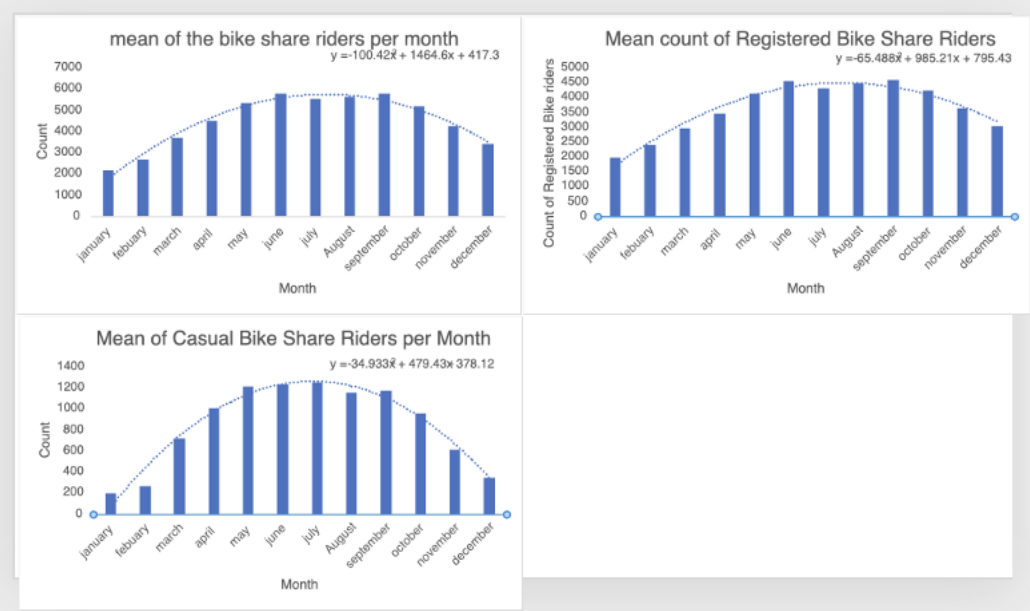
Appendix B.1: Pie chart showing the percentage of each type of biker on different days of the week



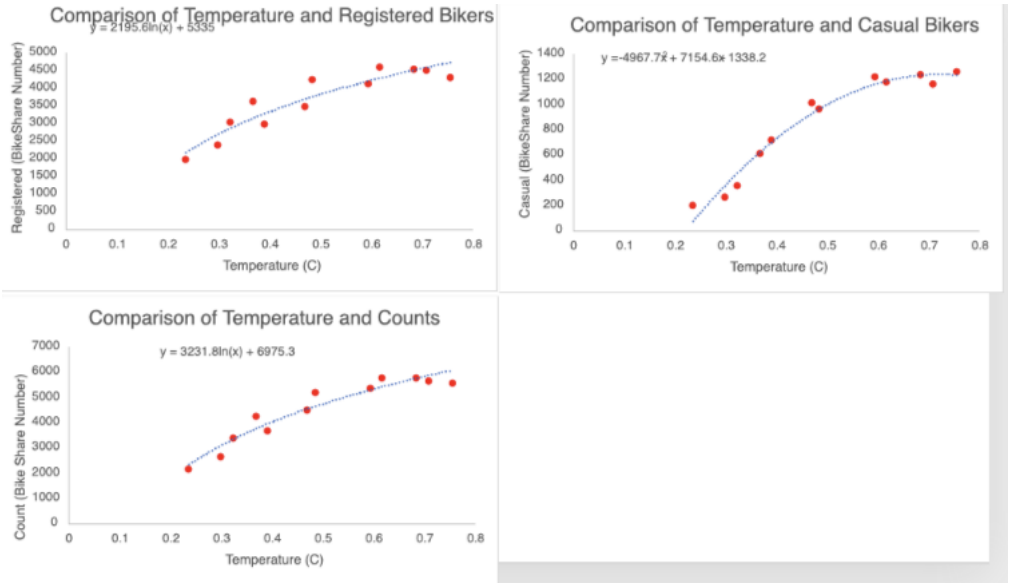
Appendix B.2: Bar chart showing the percentage of each type of biker on different days of the week



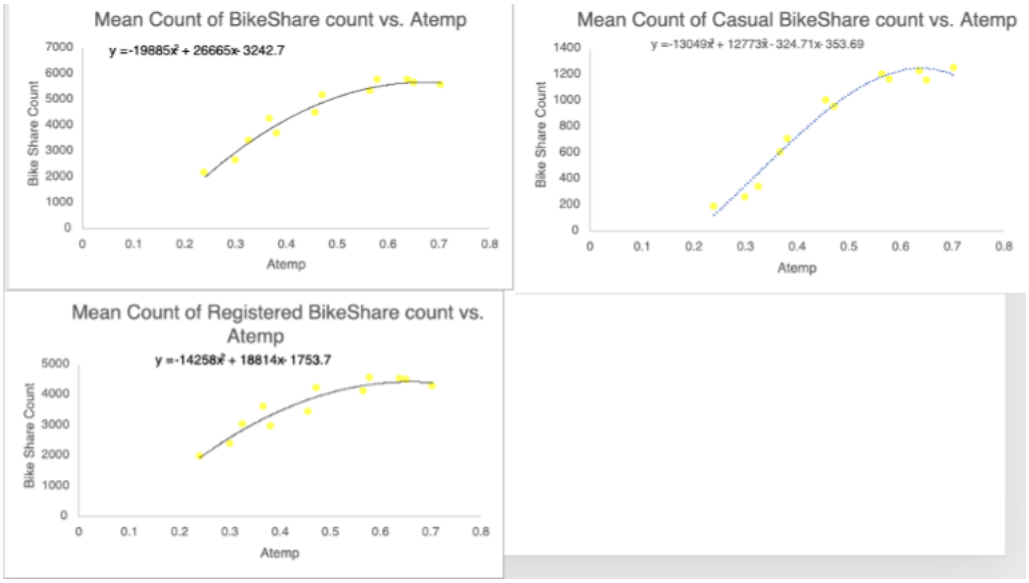
Appendix B.3: Bar chart showing the relationship between four weather parameters and the different months of the year



Appendix B.4: Line plot showing the relationship between the type of bike rental per month and temperature

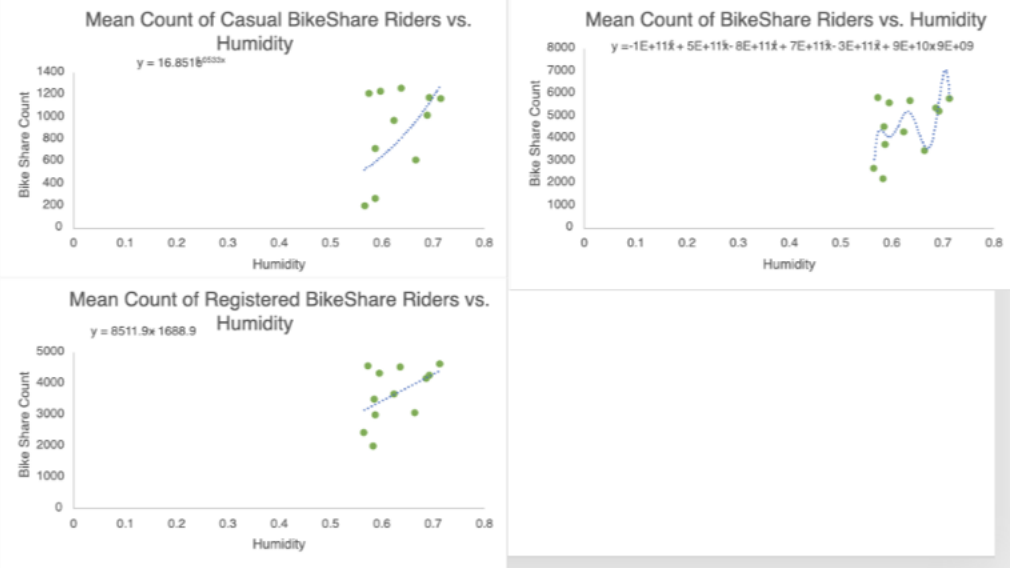


Appendix B.5: Line plot showing the relationship between the type of bike rental per month and feeling temperature



Appendix B.5: Line plot showing the relationship between the type of bike rental per month and windspeed

Appendix A: Data Pre-processing and Cleaning



Appendix C: Kruskal Wallis H Test Source Code

```r
#season
kruskal.test(daily_biker~season_name,data=daily_data)

#P-value <2.2e-16, the test yield a significantly small p value which is less than significance level of 0.05, it suggests that at least one season's mean rank is significantly different from the others. To understand which pairs of groups are different, we will conduct a posthoc test(Dunn's test) between group levels.
```

```
	Kruskal-Wallis rank sum test

data:  daily_biker by season_name
Kruskal-Wallis chi-squared = 190.87, df = 3, p-value < 2.2e-16
```

```r
dunnTest(daily_biker~season_name,data=daily_data,method="bh")
#The test returns a very small adj. p-value of pairs Fall vs. Summer; Spring vs.Summer; Winter vs. other seasons (<0.05), especially winter vs. other seasons. The low p-value indicate a high significant difference in daily count of bikers between seasons, with the strongest difference involving Winter. Yet it is worth noting that the p-value of Fall vs. Spring >0.05, which suggests that the difference between Fall and Spring is not statistically significant.

```

R Console | data.frame 6 x 4

Description: df [6 × 4]

Comparison <chr>	Z <dbl>	P.unadj <dbl>	P.adj <dbl>
Fall - Spring	-1.036189	3.001138e-01	3.001138e-01
Fall - Summer	-4.607618	4.073084e-06	6.109625e-06
Spring - Summer	-3.596210	3.228868e-04	3.874641e-04
Fall - Winter	8.604066	7.694045e-18	1.538809e-17
Spring - Winter	9.714607	2.612541e-22	7.837622e-22
Summer - Winter	13.342305	1.313352e-40	7.880112e-40

6 rows

```r
#temp_level
kruskal.test(daily_biker~temp_level,data=daily_data)
#P-value <2.2e-16, which measures the probability of observing the test statistic as extreme as the observed value under the null hypothesis (mean rank of each group are same from the other), as the test yield a significantly small p value, it suggests that at least one temp level's mean rank is significantly different from the others.  To understand which pairs of groups are different, we will conduct a pairwise comparisons between group levels with corrections for multiple testing.
```

```
	Kruskal-Wallis rank sum test

data:  daily_biker by temp_level
Kruskal-Wallis chi-squared = 202.87, df = 2, p-value < 2.2e-16
```

```r
dunnTest(daily_biker~temp_level,data=daily_data,method="bh")

#The test returns a very small adj. p-value of all pairs, which indicates that the number of daily bikers is significantly affected by the temperature level.
```

R Console | data.frame 3 x 4

Description: df [3 × 4]

Comparison <chr>	Z <dbl>	P.unadj <dbl>	P.adj <dbl>
high - low	12.513818	6.273492e-36	1.882048e-35
high - medium	10.479422	1.073979e-25	1.610969e-25
low - medium	-5.325248	1.008153e-07	1.008153e-07

3 rows

```{r}
#atemp_level
kruskal.test(daily_biker~atemp_level,data=daily_data)
#P-value <2.2e-16, which measures the probability of observing the test statistic as extreme as the observed value under the null hypothesis (mean rank of each group are same from the other), as the test yield a significantly small p value, it suggests that at least one atemp level's mean rank is significantly different from the others.
```

```
	Kruskal-Wallis rank sum test

data:  daily_biker by atemp_level
Kruskal-Wallis chi-squared = 205.33, df = 2, p-value < 2.2e-16
```

```{r}
dunnTest(daily_biker~atemp_level,data=daily_data,method="bh")

#The test returns a very small adj. p-value of all pairs, which indicates that the number of daily bikers is significantly affected by the feeling temperature level.
```

Description: df [3 × 4]

Comparison <chr>	Z <dbl>	P.unadj <dbl>	P.adj <dbl>
high – low	13.067355	5.059419e-39	1.517826e-38
high – medium	10.146467	3.435760e-24	5.153640e-24
low – medium	-5.758605	8.481180e-09	8.481180e-09

3 rows

```{r}
#wind speed level
kruskal.test(daily_biker~windspeed_level,data=daily_data)

#P-value= 9.649e-06, Final-project.pdf significantly small p value, indicating that the difference between the distribution of daily count of bikers or different wind speed levels are statistical significant, suggests that at least one wind speed level level's mean rank is significantly different from the others.
```

```
	Kruskal-Wallis rank sum test

data:  daily_biker by windspeed_level
Kruskal-Wallis chi-squared = 23.097, df = 2, p-value = 9.649e-06
```

```{r}
dunnTest(daily_biker~windspeed_level,data=daily_data,method="bh")

#The test returns a  small adj. p-value of all pairs (<0.05), which indicates that the number of daily bikers is significantly affected by the windspeed level,with statistically significant differences observed between all pairwise comparisons of different windspeed levels.
```

Description: df [3 × 4]

Comparison <chr>	Z <dbl>	P.unadj <dbl>	P.adj <dbl>
high – low	-4.406926	1.048481e-05	3.145442e-05
high – medium	-2.942959	3.250920e-03	3.250920e-03
low – medium	3.139060	1.694908e-03	2.542361e-03

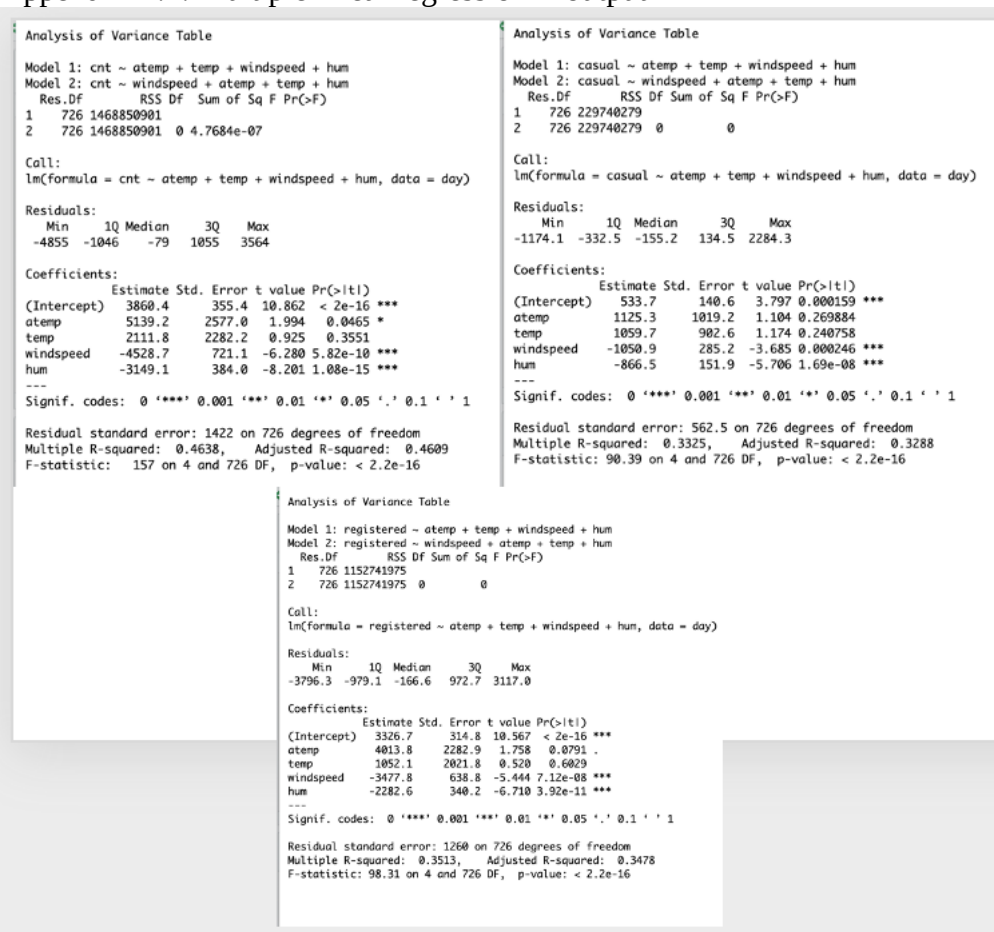
3 rows

```{r}
#hum_level
kruskal.test(daily_biker~hum_level,data=daily_data)

#p-value = 0.06917, is greater than 0.05, indicating that the test did not find statistically significant evidence to suggest that the distribution of daily biker counts differ between different humidity levels, thus failed to reject the null hypothesis. while it is worth noting that the p-value is relatively close to 0.05, which could suggest a trend or practical difference that is not statistically significant, we may need to look further in the impact of combination of factors in the distribution of daily biker counts.
```

```
	Kruskal-Wallis rank sum test

data:  daily_biker by hum_level
Kruskal-Wallis chi-squared = 5.3425, df = 2, p-value = 0.06917
```

Appendix D: Regression Source Code

Appendix D.1: Multiple linear regression R output



```
Analysis of Variance Table

Model 1: cnt ~ atemp + temp + windspeed + hum
Model 2: cnt ~ windspeed + atemp + temp + hum
  Res.Df        RSS Df  Sum of Sq F Pr(>F)
1    726 1468850901
2    726 1468850901  0 4.7684e-07

Call:
lm(formula = cnt ~ atemp + temp + windspeed + hum, data = day)

Residuals:
   Min     1Q Median     3Q    Max
 -4855  -1046    -79   1055   3564

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3860.4      355.4  10.862  < 2e-16 ***
atemp         5139.2     2577.0   1.994   0.0465 *
temp          2111.8     2282.2   0.925   0.3551
windspeed    -4528.7      721.1  -6.280 5.82e-10 ***
hum          -3149.1      384.0  -8.201 1.08e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1422 on 726 degrees of freedom
Multiple R-squared:  0.4638,    Adjusted R-squared:  0.4609
F-statistic:   157 on 4 and 726 DF,  p-value: < 2.2e-16
```

```
Analysis of Variance Table

Model 1: casual ~ atemp + temp + windspeed + hum
Model 2: casual ~ windspeed + atemp + temp + hum
  Res.Df       RSS Df Sum of Sq F Pr(>F)
1    726 229740279
2    726 229740279  0         0

Call:
lm(formula = casual ~ atemp + temp + windspeed + hum, data = day)

Residuals:
     Min       1Q   Median       3Q      Max
-1174.1  -332.5  -155.2   134.5  2284.3

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    533.7      140.6   3.797 0.000159 ***
atemp         1125.3     1019.2   1.104 0.269884
temp          1059.7      902.6   1.174 0.240758
windspeed    -1050.9      285.2  -3.685 0.000246 ***
hum           -866.5      151.9  -5.706 1.69e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 562.5 on 726 degrees of freedom
Multiple R-squared:  0.3325,    Adjusted R-squared:  0.3288
F-statistic: 90.39 on 4 and 726 DF,  p-value: < 2.2e-16
```

```
Analysis of Variance Table

Model 1: registered ~ atemp + temp + windspeed + hum
Model 2: registered ~ windspeed + atemp + temp + hum
  Res.Df        RSS Df Sum of Sq F Pr(>F)
1    726 1152741975
2    726 1152741975  0         0

Call:
lm(formula = registered ~ atemp + temp + windspeed + hum, data = day)

Residuals:
    Min      1Q  Median      3Q     Max
-3796.3  -979.1  -166.6   972.7  3117.0

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3326.7      314.8  10.567  < 2e-16 ***
atemp         4013.8     2282.9   1.758   0.0791 .
temp          1052.1     2021.8   0.520   0.6029
windspeed    -3477.8      638.8  -5.444 7.12e-08 ***
hum          -2282.6      340.2  -6.710 3.92e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1260 on 726 degrees of freedom
Multiple R-squared:  0.3513,    Adjusted R-squared:  0.3478
F-statistic: 98.31 on 4 and 726 DF,  p-value: < 2.2e-16
```