Predictive Model for Bike Usage in City of Seoul

STAC67 Case Study

Group 24

Shawn Santhoshgeorge - Data Analyst (1006094673)
Shashwat Piyush Doshi - Data Analyst (1005716940)
Rohita Nalluri - Data Analyst (1006154037)
James Chen - Data Analyst (1006166220)

Due Date

December 06, 2021

Introduction

In the past decade, the usage of public rental bikes has increased in various urban cities. In cities, rental bikes have become a crucial part of transportation because of many reasons. They provide free or affordable access to transportation for short-distance trips instead of requiring a private vehicle and help reduce congestion, noise and air pollution in a city (Winters, 2020). Bike Sharing systems are now rapidly growing across the world at nearly 2000 various operating programs in total as of May 2021 (Yu et al., 2021). The demand for these systems is growing at 14.3% compounded annually from 2017 to 2025 (Cooper, 2019).

There are various public bike rental systems across the world, like in the city of Toronto where they have approximately 6,850 bikes and its users have completed 2.9 million trips as of 2020 (Bike Share Toronto, 2021). In New York City there are "... nearly eight hundred thousand (773,000) ride a bicycle regularly" and have completed over 7 million rides as of September 2021 with a 26% growth in daily cycling between 2014 to 2019 (New York City DOT, 2021). One of the biggest bike-sharing systems is found in China at Hangzhou and Wuhan where their residents have access to 90,000 and 70,000 bicycles respectively (Borowska-Stefańska et al., 2021).

Like many other cities within the world, the city of Seoul in South Korea has also founded a bike-sharing system called *Ttareungyi* or Seoul Public Bike started in 2015. It was started to " resolve issues of traffic congestion, air pollution, and high oil prices in Seoul, and to build a healthier society while enhancing the quality of life for Seoul citizens" (Seoul Metropolitan Government, 2015). As of March 20, 2018, it has exceeded 620,000 memberships with 38% of the users using the bikes during rush hour (Seoul Metropolitan Government, 2018). Due to the high usage of bikes in cities, effective management of bike-sharing systems must be developed to ensure the general public can access the service when required.

The following data set from the Seoul Public Bike initiative provides various information like the number of bikes rented, temperature, humidity, wind speed, visibility, solar radiation, snowfall and more. This can be used to see how weather affects the number of bikes rented per hour and makes a predictive model for future uses from the given data from December 1st, 2017 to December 1st, 2018. This is a very important question to consider since an online survey conducted within the geographical region of Asia in 2016 identified, climate as an "... important physical barrier with 25% of respondents strongly agreeing that it constrains them physically, particularly warm or warmth and high precipitation levels."(Mateo-Babiano et al., 2017) among other reasons like infrastructure issues and others.

We will do this by first cleaning the data, making visualizations to see trends between variables and then generating and inspecting the model and summary statistics and using diagnostics to improve the model and deal with outliers and such. Some variables will be left out due to high multicollinearity issues or not contributing much to the model which was found using leaps and AIC criterion.

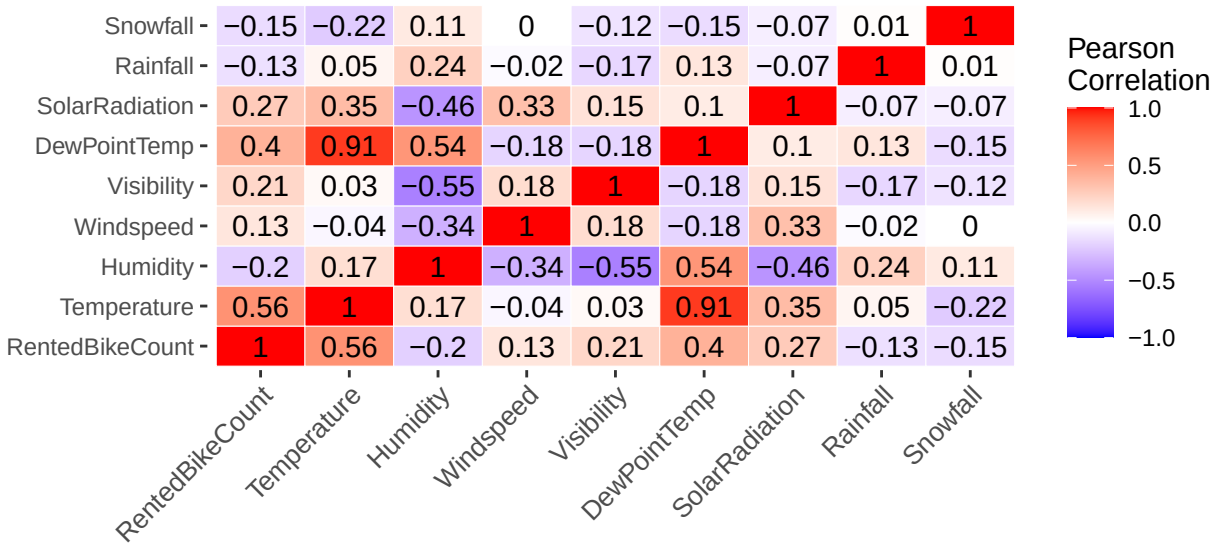
We start first with analyzing the data to verify that there are no missing values in any of the columns, change the format of certain variables to make it easier to make visualization from the data. After taking a look at the column names of the data set we decide to change the names to be it easier to use. Then we ran a check on each column of the data set to look for missing values and found none. We also decided to drop the variable *FunctioningDay* which is there to indicate if the bike-sharing system is available or not and if we observe the number of rented bikes we see that it is 0 for 0 observations where *FunctioningDay* is "No." Since here we are trying to predict how many bikes will be rented when the system is open. For this reason, we can remove these observations. We also formatted the *Date* column to by using a built-in date function to help when we need to make graphs and such.

	Is Null
Date	FALSE
RentedBikeCount	FALSE
Hour	FALSE
Temperature	FALSE
Humidity	FALSE
Windspeed	FALSE
Visibility	FALSE
DewPointTemp	FALSE
SolarRadiation	FALSE
Rainfall	FALSE
Snowfall	FALSE
Seasons	FALSE
Holiday	FALSE
FunctioningDay	FALSE

Exploratory Data Analysis

After cleaning the data set we have 8465 observations. In this section, we will be taking a look at some of the 13 variables and consider their effect on our response variable *RentedBikeCount* and the relationship between explanatory variables themselves. We will also make a decision with regard to which variables to keep for the model building.

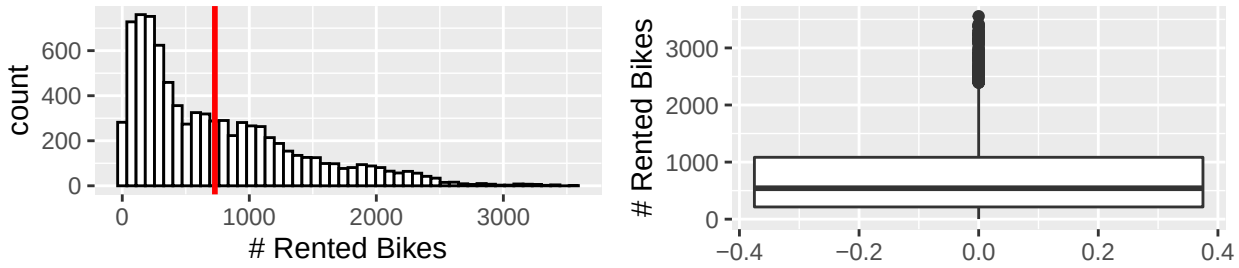
Correlation Heat Map



From Figure 1 which is the Correlation Heat Map we can observe that the correlation between the explanatory variables DewPointTemp and Temperature is 0.91 which is quite high and so one of them will be dropped to reduce multicollinearity. We decided to drop DewPointTemp and kept Temperature as a possible variable to be used since it correlated much strongly to RentedBikeCount. We can also notice that Visibility, SolarRadiation and Snowfall correlate well with RentedBikeCount.

RentedBikeCount

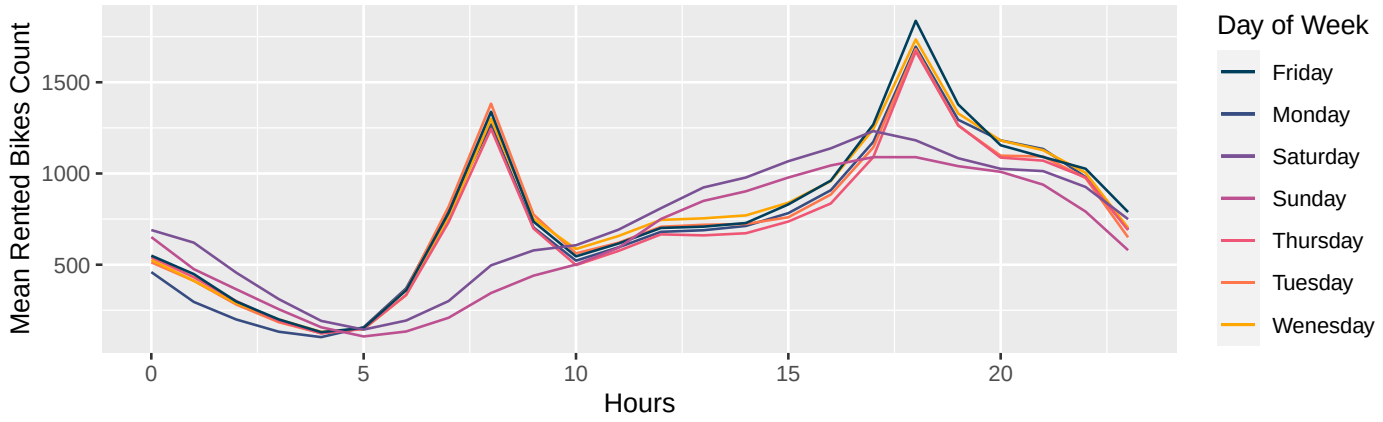
The *RentedBikeCount* is the response variable that we are trying to predict from our model. This variable will be represented as the variable *Y*. From the summary of the variable, we see that at minimum there are 2 bikes rented with a maximum of 3556 bikes rented at once and on average 729 bikes. From the histogram, we see that many of the observations are less than the average number of bikes represented by the blue line rented and from the boxplot we see that number of bikes rent is pretty similar between each observation for the most part with some observations as outliers.



Date

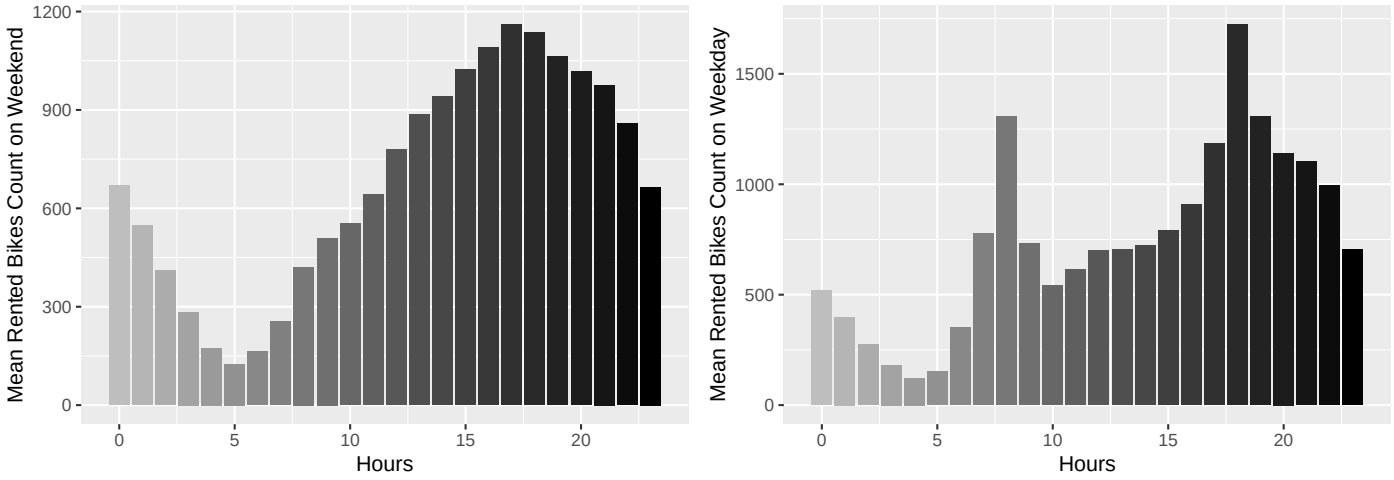
The *Date* column of the data set stores the date during which this observation was taken and is in the format of (year-month-day). This in its self will not be providing much information but breaking the date into specific days of the week can help to see the general trend of usage throughout a week and specifically the demand weekdays and weekends. So after making a graph that plots the average number of bikes used per hour for each day of the week we can see that the demand

change significantly based on if it is a weekday or weekend. The two peaks in the weekday can be explained by the start of workday and school which is approximately 8 AM and the end of the workday and school which is approximately 5 PM on average. Since there is a significant change in demand based on if it is a weekend or not we will make a categorical variable where 1 represents weekend and 0 represents not weekend with a base category that will be set to not weekend.



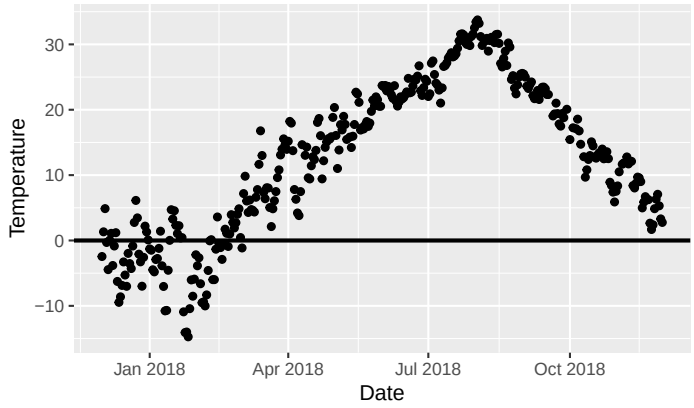
Hours

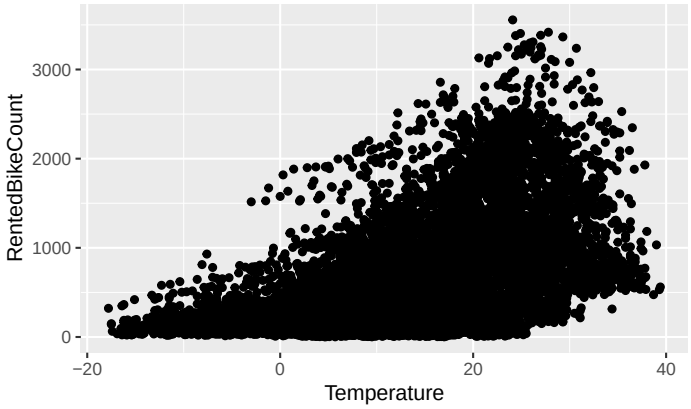
The *Hour* column of the data set records the hour of the day this observation was taken and is in the format of 24 hrs ranging from 0 to 23. As mentioned when we discussed the *Date* column, we saw that the hour played a huge role in the demand for bikes throughout the day. From the graphs below we also see that combining the *Weekend* and *Hour* variables shows more information about the fluctuating demaand of bike usage. So, we make Hour into a categorical variable corresponding to a specific hour of the day with the base category being Hour 23.



Temperature

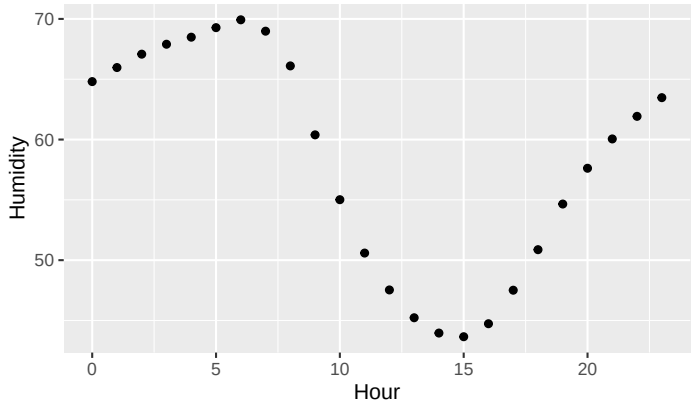
Temperature was measured in degree Celsius (℃) at the hour of observation. The lowest temperature reached was -17.80 ℃ and the highest it went was 39.40 ℃ and on average the temperature was at 12.77 ℃. From the graph on the left, we can see the temperature has a daily trend and with the graph on the right we can see that there is an association between Temperature and RentedBikeCount whereas generally an increase in Temperature increases the number of bikes rented.

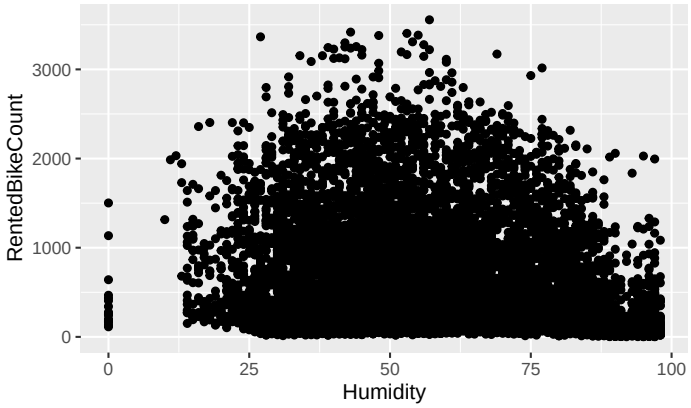




Humidity

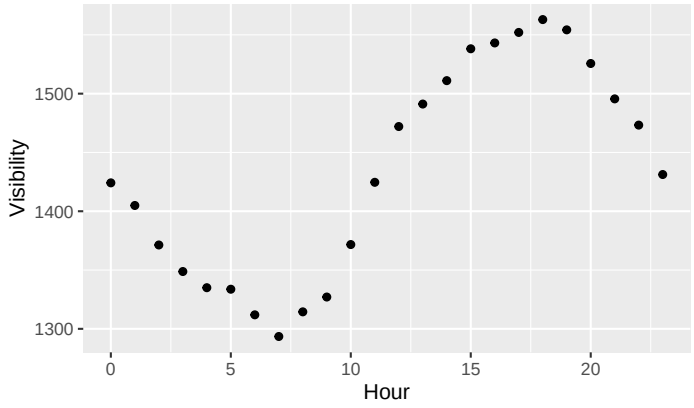
Humidity was measured in % at the hour of observation. In regards to the values, we have 0% as the lowest and 98% as the highest for Humidity with an average of 57 %. From the graph on the left, we can see that Humidity has an hourly trend and there was no noticeable Daily and Monthly Humidity Trend. In the graph on the right, we can see that the number of bikes rented varies slightly based on the graph.

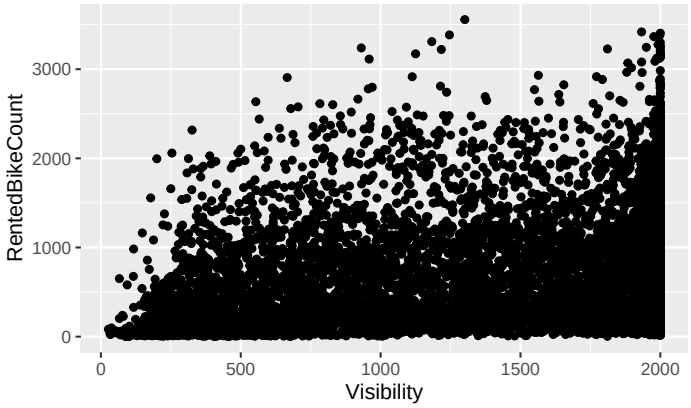




Visibility

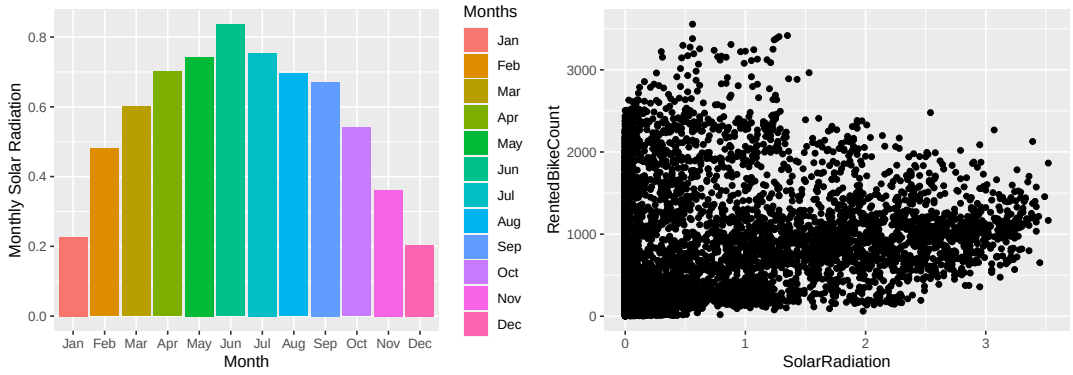
The distance till which an object or light can be clearly seen is measured in meters (m) at the hour of observation. The average distance visible for a biker is 1690 m and it can go down as low as 27 m and as high as 2000 m. From the graph on the left, we can see that Visibility has an hourly trend and that there was no noticeable Daily and/or Monthly Visibility Trend. In the graph on the right, we can see that the number of bikes rented varies based on the Visibility variable and if we go left to right the number of bikes rented increases as Visibility increases.





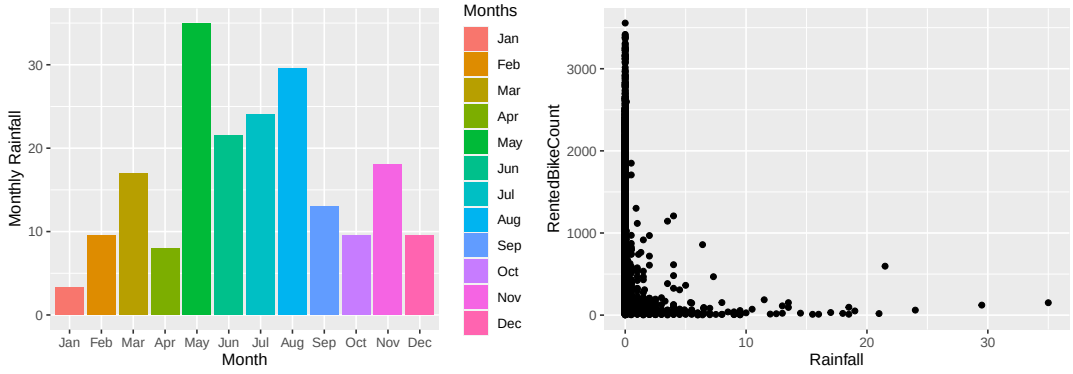
Solar Radiation

The *SolarRadiation* column in the data set is the solar radiation experienced, measured in $\frac{Mj}{m^2}$, at the hour of observation. On average according to this data set Seoul gets 0.5679 $\frac{Mj}{m^2}$ with a maximum of 3.52 $\frac{Mj}{m^2}$ of solar radiation for one observation. From the graph on the left, we can see that solar radiation throughout the year and is quite consistent. Looking at its relationship with RentedBikeCount we see that



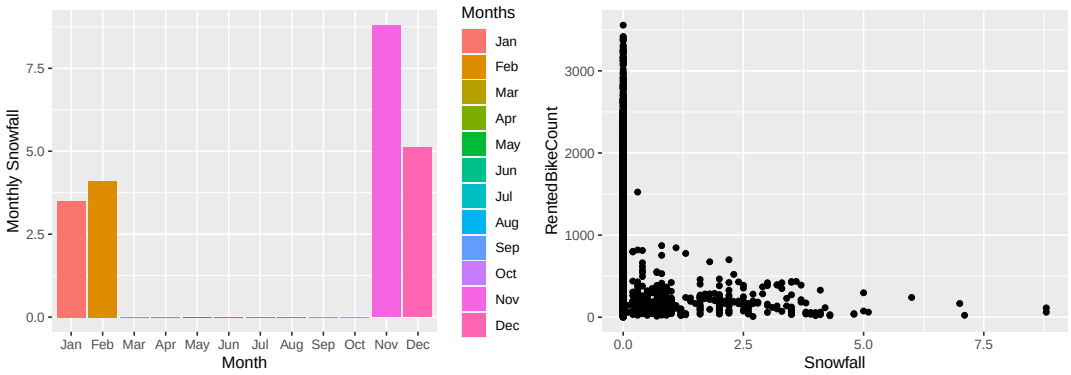
Rainfall

The *Rainfall* column in the data set is the rainfall experienced, measured in millimetres (mm), at the hour of observation. On average according to this data set Seoul gets 0.15 mm with a maximum of 35 mm of rainfall for one observation. From the graph on the left, we can see that rainfall throughout the year with the most amount falling in the month of May. Looking at its relationship with RentedBikeCount we see that it has quite the effect by decreasing the number of bikes rented on those days.



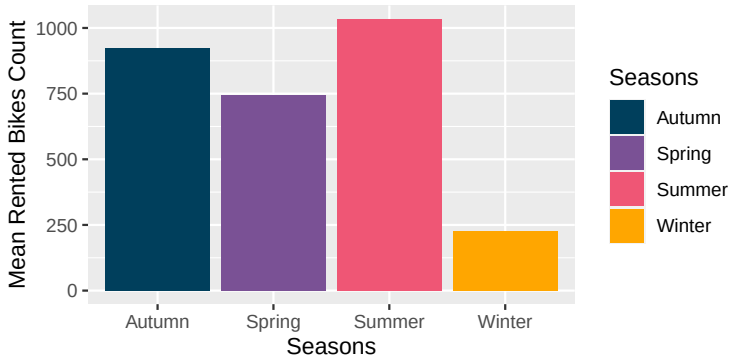
Snowfall

The *Snowfall* column in the data set is the snowfall experienced, measured in centimetres (cm), at the hour of observation. On average according to this data set Seoul gets 0.08 cm or 0.8 mm with a maximum of 8.8 cm of snowfall for one observation. From the graph on the left, we can see that Snowfall only falls in the first 2 and last 2 months of the year and mostly comes down during November. Looking at its relationship with RentedBikeCount we see that it has quite the effect of decreasing the number of bikes rented on those days.



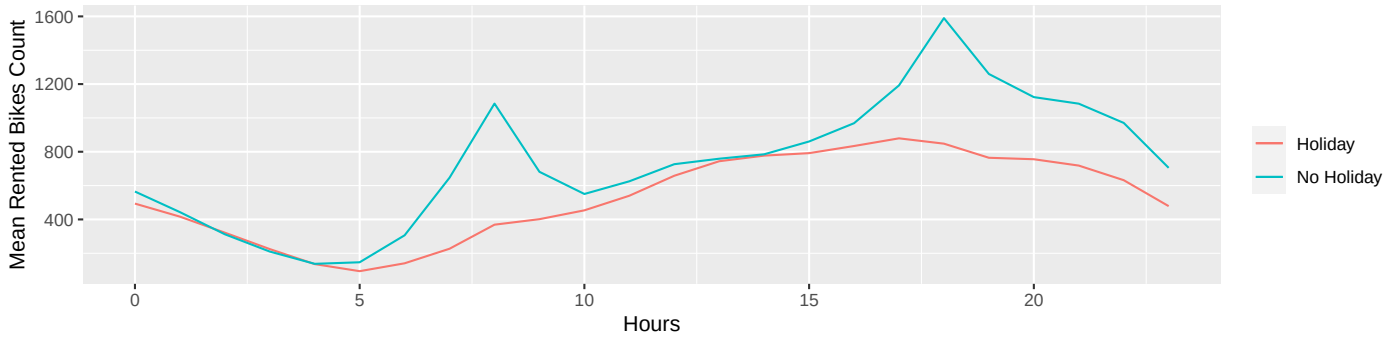
Seasons

Season is a column in the data set that states what Season was this observation taken in. From the graph below we can see that the Season most definitely affects the number of bikes rented on average. In the Winter season, you see the lowest number of bikes rented and in the Summer season, you see the highest number of bikes rented. Autumn and Spring seasons on the other hand are relatively close to each other. Adding the *Seasons* to the model as a categorical variable for the current season in Seoul (Winter, Spring, Summer, Autumn) with the base category being set to Autumn will make an effect on the model.



Holiday

Holiday is a column in the data set that states if that *Date* is a "Holiday" or "No Holiday." Looking at the graph below we see that is quite similar to the day of the week above in the *Date* section. Therefore, adding the *Holiday* variable to the model will most likely add no extra information and so it will be excludedhe model will most likely add no extra information and so it will be excluded



Model Building & Validation

We start with building the model with all the given variables: X0 (Date), X1(Hour), X2(Temperature), X3(Humidity), X4(Windspeed), X5(Visibility), X7(Solar Radiation), X8(Rainfall), X9(Snowfall), X10(Seasons), X11(Holiday), and the weekend variable.

After inspecting the cleaned data, our initial assumption was to classify X10 (Seasons), and the weekend variable as categorical(binary) variables, while the rest should be classified as continuous variables.

Then, by using the package leaps() we find the best subset of continuous variables.

After running it we found that the best subset of continuous variables is all of them except X4 which is windspeed since it had the lowest AIC of -6061.978.

We now decide to include the categorical variables with the model, which are and Weekends. Then, based on several iterations of the ANOVA Table outputs, we tried to fit a model that gave us a considerably good Adjusted R-squared value as well as a good VIF value, among other considerations for fitting the best model.

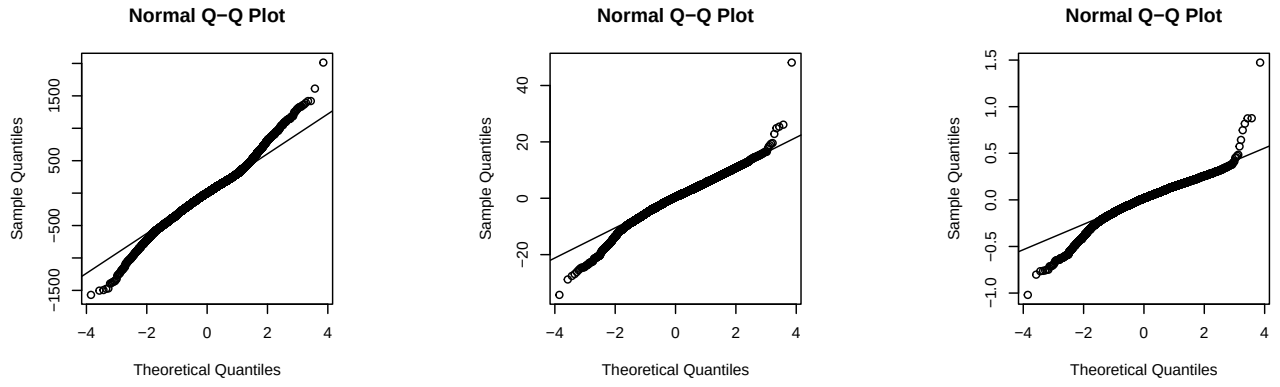
After several iterations and inspections, we came up with the following assumptions:

1. The Seasons and Weekend variables should be included in the model.
2. Convert the Hour variable into a categorical variable and use it with the Seasons categorical variable.

Even after editing our model, we encountered an issue with *EXTREMELY HIGH* multicollinearity. (Did this by calling VIF() function)

So, to solve this, we took out the NA rows in the summary statistics of the model, as well as inspected rows with high P-value. However, instead of removing specific interactions of X7 with hours, we decided to remove interactions of Hours with any variable whenever we encountered a significant number of NA values or high P-values in the summary statistics output. We iteratively conducted this process and checked the VIF output. We needed to find a model with VIF values less than 10.

But, now the Q-Q residual plot does not align with the line y=x, and is *heavy-tailed*, as indicated by the left Q-Q plot. Therefore, we conducted a Square-Root Transformation and obtain the resulting Q-Q Plot, indicated by the middle graph. Now, we tried to align the Normal Q-Q plot with the line y = x. To do this, we applied the Box-Cox Transformation. We investigated this further and found that there is no significant effect on the distribution after applying a Box-Cox transformation, indicated by the right graph. Therefore, we only transform the response variable by taking Yˆ0.5 (i.e., by only performing the square-root transformation).



Hence, we conclude that we have our final model, which is:

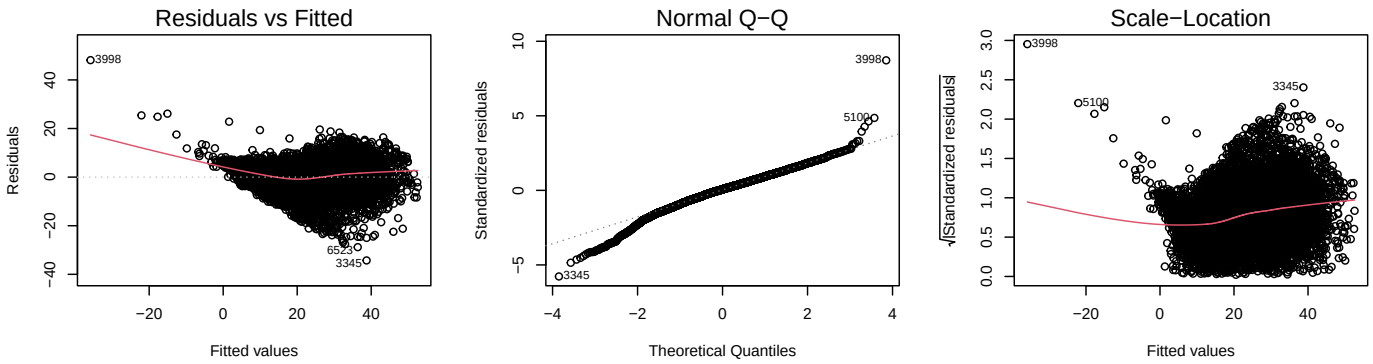
RentedBikeCount = (Rainfall:Weekends + Visibility + Snowfall + Rainfall + Humidity + Temperature + SolarRadiation + SolarRadiation:Weekend + SolarRadiation:Seasons + Temperature:Weekend + Weekend:Hour + Hour)^2

Cross Validation

We randomly choose 70% of our dataset to be the training set, and the remaining 30% to be the validation set. Then we compare our MSPE and MSE to see if they were similar. Since the MSPR computed from the validation set was 35.493. MSE computed from the training set was 36.020. Due to MSPR and MSE being fairly close, it can be concluded that our model is valid.

Diagnostics

Regression Assumptions



Residual Vs. Fitted plot shows residuals are randomly and evenly distributed along a mainly horizontal line, therefore the linear assumption is true. The normal Q-Q plot is mostly linear with a slight left skew. Because the majority of data points are on the line, we conclude the data is normal. Scale - Location plot shows randomly distributed points on a nearly horizontal line, therefore variance is homogeneous.

Outlying Y Observations

Comparing studentized deleted residuals of each observation against our threshold (t.crit = 4.533), 7 observations stood out as outlying Y observations.

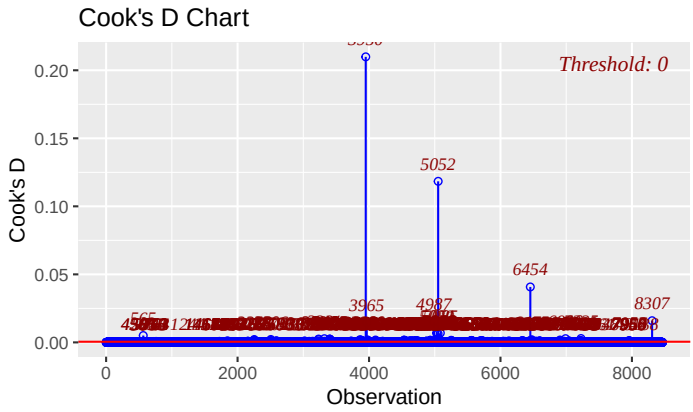
High leverage X Observations

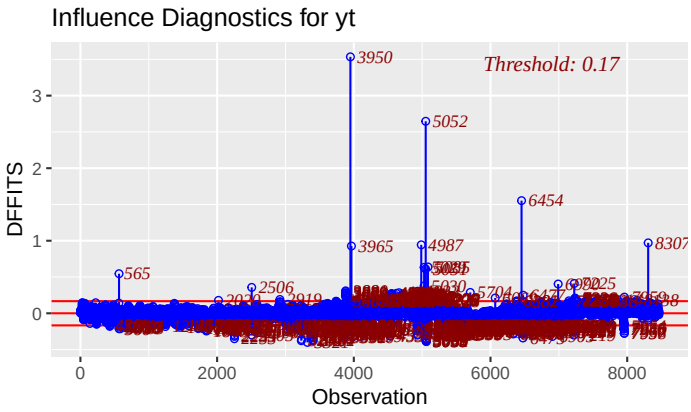
Comparing the leverage of each observation against two times the mean leverage ($\frac{2p'}{n} = 0.013$), 86 observations stood out as high leverage X observations. Comparing leverage of each observation against .5, 0 observations stood out as high leverage x observations.

Influencial Observations

Comparing Cook's Distance and DFBETAS against their respective thresholds, 0 influential observations were found. Comparing DFFITS against its threshold (2*sqrt(p'/n) = 0.161), 129 influential observations were found.

Overall, 3 outlying Y observations were influential, 22 high leverage X observations were influential. After looking closely at each influential outlying observation, we determined that all points were naturally generated and will be kept in our dataset. As for the remaining non-influential outliers, there is no real difference whether they remain in our dataset or not.





Looking at the cook's distance graph, 5 points are clearly larger than the rest (3950, 5052, 6454, 8307, 3965) and looking at DFFITs graph, a multitude of points lay above and below the threshold, confirming our calculations.

Multicollinearity

```
##           X5           X9           X8           X3           X2           X7
##     1.613196     1.106860     1.255359     2.739509     2.400330     9.890820
##           h0           h1           h2           h3           h4           h5
##     2.319102     2.319765     2.320368     2.321563     2.322528     2.325860
##           h6           h7           h8           h9          h10          h11
##     2.326202     2.334889     2.362790     2.475024     2.721096     3.002367
##          h12          h13          h14          h15          h16          h17
##     3.224625     3.294957     3.163660     2.956680     2.699695     2.486660
##          h18          h19          h20          h21          h22 X8:weekends
##     2.386334     2.350171     2.344352     2.335653     2.328209     1.280933
##  weekends:X7       X7:D1        X7:D2        X7:D3  weekends:X2  weekends:h0
##     6.794372     1.626546     2.624579     2.872842     3.187702     1.450655
##  weekends:h1  weekends:h2  weekends:h3  weekends:h4  weekends:h5  weekends:h6
##     1.447096     1.443316     1.439648     1.437170     1.434598     1.431454
##  weekends:h7  weekends:h8  weekends:h9 weekends:h10 weekends:h11 weekends:h12
##     1.431728     1.447201     1.560153     1.751047     1.988405     2.173666
## weekends:h13 weekends:h14 weekends:h15 weekends:h16 weekends:h17 weekends:h18
##     2.222691     2.143231     1.972328     1.734498     1.575764     1.496232
## weekends:h19 weekends:h20 weekends:h21 weekends:h22
##     1.480874     1.479604     1.472783     1.465503
```

The largest VIF was 9.196, for solar radiation, which is less than 10. The mean VIF was 2.303, which is not considerably larger than 1, therefore there is no indication of serious multicollinearity.

Conclusion

When we first started working on the Seoul Bike Sharing dataset, our goal was to figure out how the weather affected the number of bikes rented on an hourly basis.

Our findings showed that weather did not have as large of an effect as we anticipated, with the major effects being from the hour itself, i.e. eight in the morning versus eight at night, whether it was a weekend or not as well as if it was raining. To be more conclusive, we found in order of most to least impact: hour, weekend, temperature, humidity and visibility. As well, we also found that windspeed had no impact.

However, we also need to account for the limitations of our study:

1. The regression model we have developed is based on the models and techniques we had been exposed to during class. As such, there may be a better model and transformation that would fit our purposes better.
2. The predictive model we have designed is based on a fixed data set and as such may not be as accurate when dealing with real-time data points.
3. After applying our transformations, we found that our final models' residuals appear left-skewed albeit trying boxcox which made it worse.
4. Our model is also focused on Seoul as a whole, and the number of bikes rented on an hourly basis may vary amongst different parts of Seoul.

By creating a predictive model that can accurately estimate how many bikes will be needed in regards to a variety of areas, we can ensure that we meet the demand for rental bikes. This is especially important due to the high annual growth rate of rental bike use which can partly be attributed to the increasing awareness of climate change which has led people to prioritize more environmentally modes of transportation among others.

As such, potential areas for future research would be to go further into our data and see how the number of bikes rented on an hourly basis varies amongst different parts of Seoul to optimize the allocation of rental bikes in the city. As well, similar bike rental systems could be popularized and adapted across the world, such as Toronto which could help lower carbon emissions.

Bibliography

(2021) (pp. 1–10). New York City Department of Transportation. Retrieved from https://www1.nyc.gov/html/dot/downloads/pdf/cycling-in-the-city-2021.pdf

A look back at 2020 with bike share toronto: The silver lining. (2021, January 2). Bike Share Toronto. Retrieved from https://bikesharetoronto.com/news/2020-year-in-review/

Borowska-Stefanska, M., Mikusova, M., Kowalski, M., Kurzyk, P., & Wisniewski, S. (2021). Changes in urban mobility related to the public bike system with regard to weather conditions and statutory retail restrictions. *Remote Sensing*, *13*(18), 3597. doi:10.3390/rs13183597

Cooper, S. (2021). Bike rental market size, share, trend, 2019-2025 | industry report. *Web.archive.org*. Retrieved from https://web.archive.org/web/20201022130750/https://www.hexaresearch.com/research-report/bike-rental-market

Mateo-Babiano, I., Kumar, S., & Mejia, A. (2017). Bicycle sharing in asia: A stakeholder perception and possible futures. *Transportation Research Procedia*, *25*, 4975. doi:10.1016/j.trpro.2017.05.375

Seoul public bike. (2015, October). Seoul Metropolitan Goverment. Retrieved from http://english.seoul.go.kr/service/movement/seoul-public-bike/

Seoul public bike. (2018, March 21). Seoul Metropolitan Goverment. Retrieved from http://english.seoul.go.kr/members-seouls-public-bicycle-reach-620000-38-percent-users-use-bike-rush-hours/?keyword=Bike&cat=46

Winters, M. (2020, February 20). Bike share | cycling in cities. *Bike Share*. Retrieved from https://cyclingincities.spph.ubc.ca/motivating-cycling/bikeshare-systems/

Yu, C., O'Brien, O., DeMaio, P., Rabello, R., Chou, S., & Benicchio, T. (2021, October). The meddin bike-sharing world map mid-2021 report. PBSC Urban Solutions. Retrieved from https://bikesharingworldmap.com/reports/bswm_mid2021report.pdf