**A Statistical Analysis of Bike Rentals in Seoul**

**MSBA 620: Data Analytics 1**

**Nicholas Michal**

**December 1, 2020**

**Summary**

Ever since Seoul’s bike sharing system Ddareungi, also known as Seoul Bike, was introduced in October of 2015, there has been an increasing demand for additional bikes and rental stations. This paper will investigate the statistical significance of weather on the number of bikes rented from a public bike renting program. This data is publicly available online in the UCI Machine Learning Repository. Specifically, the team used the Seoul, South Korea bike rental data. The team was able to run a variety of tests to determine the correlation and statistical significance of the different aspects of weather and what role they play in the number of bikes rented.

**Introduction**

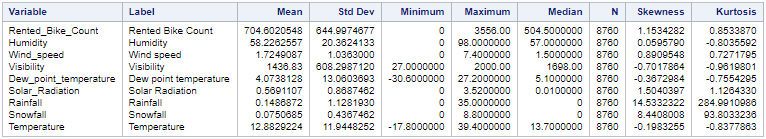
Seoul, South Korea has a bike rental program where people can rent bikes to ride around the city whether it be for commuting, tourists, exercise, recreation activities, or any other reason. In 2017, the city’s bike rental data was tracked with various weather conditions recorded: temperature, humidity, wind speed, visibility, dew point, solar radiation, rainfall, snowfall, and the season of the year. It is suspected that weather conditions affected the number of bikes rented, so the following null and alternative hypothesis were formed:

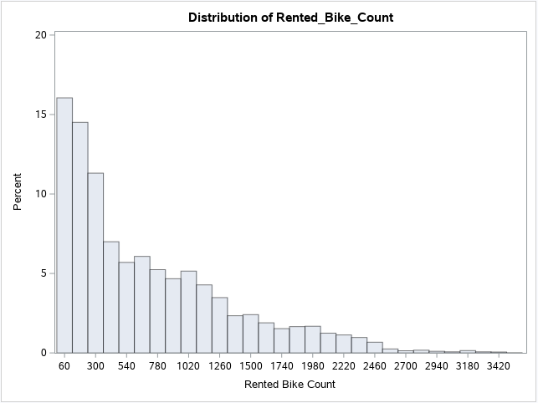
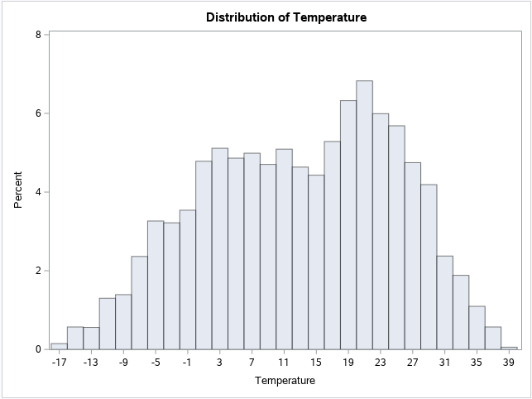
H0: Weather has no effect on bike rentals

H1: Weather effects bike rentals

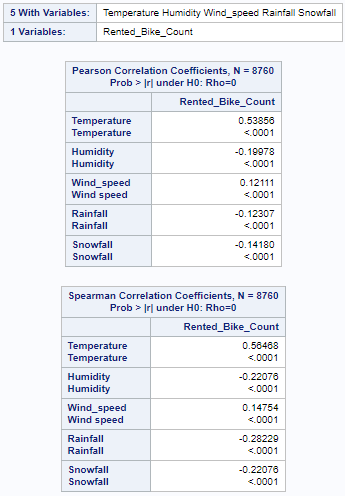
In the following sections, various statistical analyses will be conducted to determine if the null hypothesis should be accepted or rejected.

**Methodology**

The Seoul bike rental dataset was explored by first evaluating the descriptive statistics output in SAS Studio. The rented bike count is around 705 bikes per hour (on average), while the median value is around 505 bikes. Since the rented bike count is skewed, the median provides a better understanding for the ‘typical’ number of bikes rented at a given time.

 Furthermore, the team evaluated distributions for measures used to describe weather conditions. Rainfall and snowfall have skewed distributions since it does not precipitate every day. Also, upon visual inspection of the temperature variable, it seems to have an approximately normal distribution.

In addition to descriptive statistics, the team evaluated the Pearson and Spearman correlations between variables. There are weak correlations between bike rentals and the following measures: humidity, wind speed, rainfall, and snowfall. However, there is a 0.56468 Spearman correlation between rented bike count and temperature.

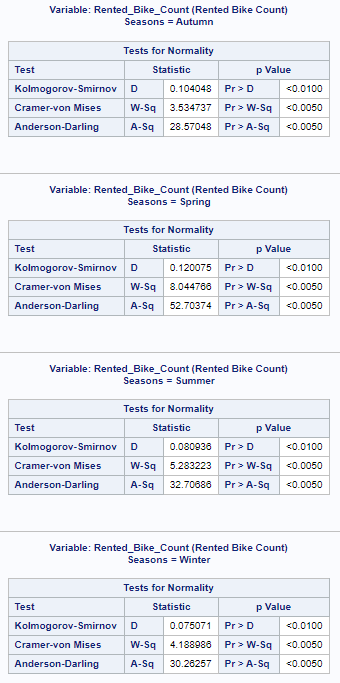


To build upon the statistical analysis of the Seoul bike dataset, a t-test was run , ANOVA, and regression analysis in SAS Studio. The main goal was to see if there is a difference in rented bikes given changing weather conditions or season of the year.

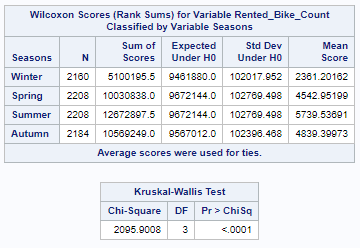
**Results**

*ANOVA*

The first question to answer was whether bike rentals varied by season. A normality test was run to determine whether it should use parametric or nonparametric tests.

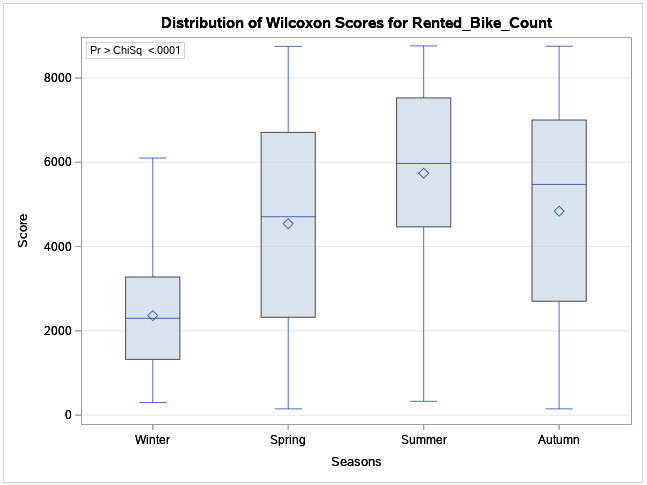


As displayed in the tables above, for every season variable (Autumn, Spring, Summer and Winter), the p-values indicated statistical significance, meaning a nonparametric test would be most appropriate to analyze the data. As a result, the Wilcoxon score and Kruskal-Wallis Test were used.



The Kruskal-Wallis test above is statistically significant with *p <* 0.0001, demonstrating that at least one of the pairs of seasons is different. In the Wilcoxon Scores table, we can see a couple of these differences play out, especially in summer, in which the expected sum is about 24 percent lower than the actual sum of scores, and winter, which has an expected sum almost twice the actual sum of scores.

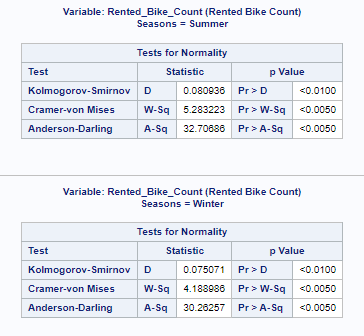
This also plays out in the box plot displaying the distribution of Wilcoxon Scores (below). While the spring and autumn are relatively similar, winter lags far below in the rest while summer is significantly higher and with a smaller window, indicating consistently higher rentals than the other three seasons.



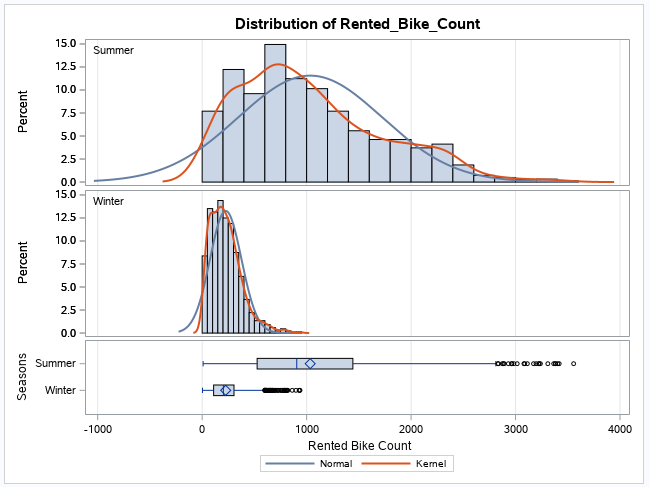
Something is leading to a significant difference in bike rentals between seasons, though whether a correlation exists with weather conditions is yet to be seen, so more tests were conducted to delve deeper into the hypothesis.

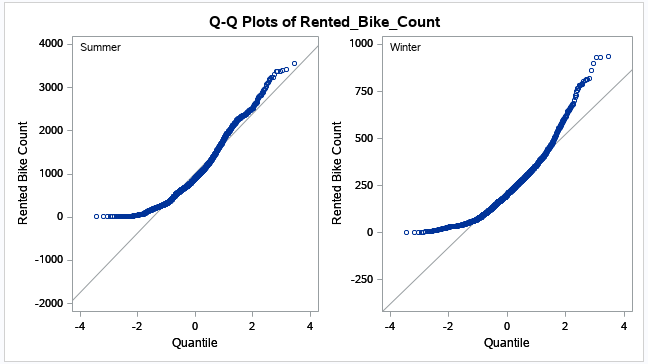
*T-Test*

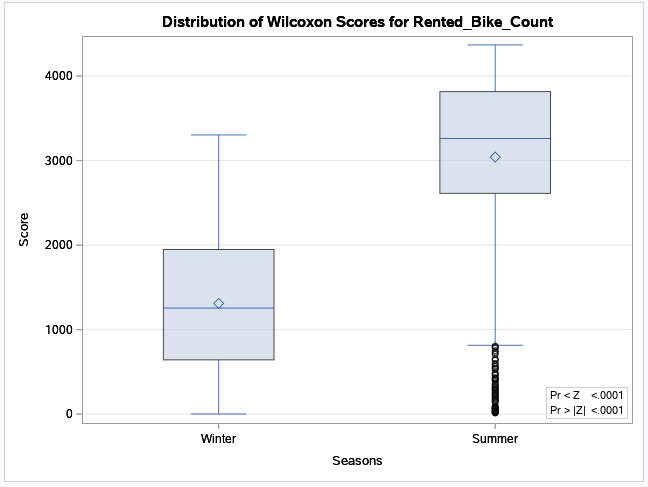
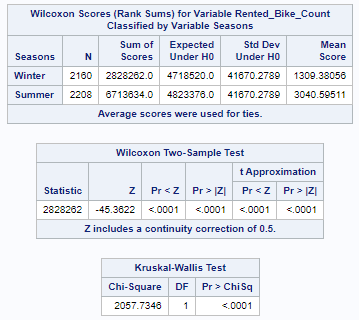
After the ANOVA test showed that there was a statistical difference for bike rentals in at least one of the seasons, a deeper dive was taken into only two of the seasons using a two sample T-test. The decision was made to choose Summer and Winter bike rentals due to their visual and statistical difference in the ANOVA test and plots. They also tend to have different conditions in a typical year.



As you can see from the charts, the p-values for both Summer and Winter were statistically significant which led to running the Wilcoxon two sample test.



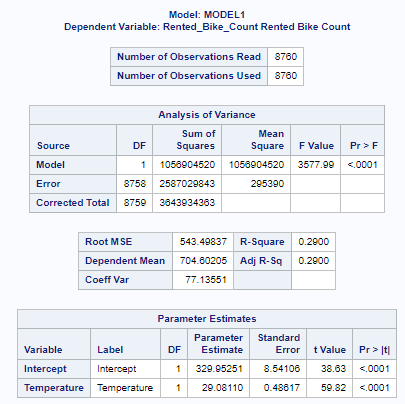




After performing the two-sample t-test, the results show that Summer’s bike rental mean is more than double that of Winter’s bike rental mean. Also, both the Wilcoxon two sample t-test and the Kruskal-Wallis Test both provide scores that are statistically significant (<.0001). For this reason, the conclusion was that there is a significant difference between bike rentals in Summer and Winter.

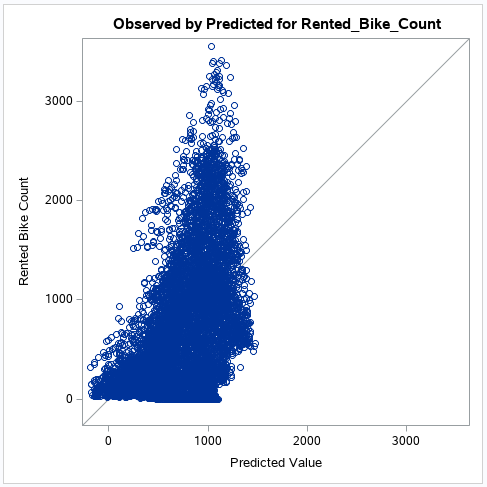
*Regression*

A simple linear regression was created for the number of bikes rented for each temperature. This was determined after a correlation analysis between the two variables showed a correlation of 0.54. The results of the linear regression analysis is shown below.

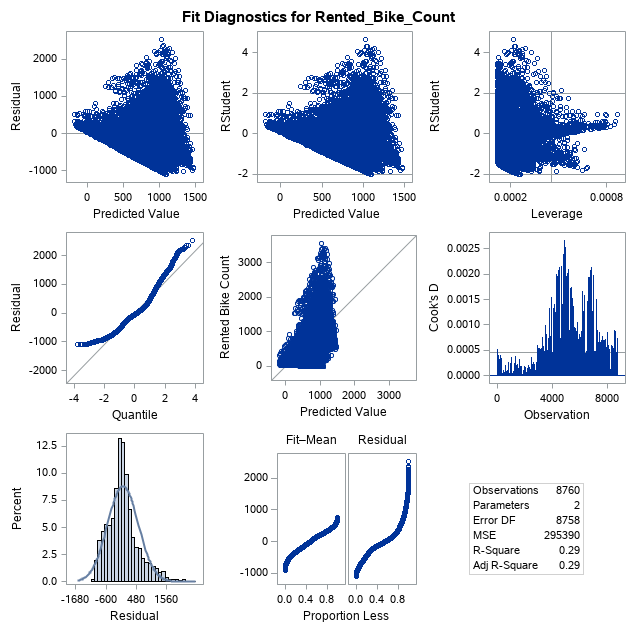


According to the ANOVA output, this data is shown to be statistically significant with a P value less than .0001. These results show, with 95% confidence, that there will be a difference between the number of bikes rented and temperature. Continuing, the R-Square and Adjusted R-Square both show a value of .29 so 29% of the bikes rented can be explained by temperature.

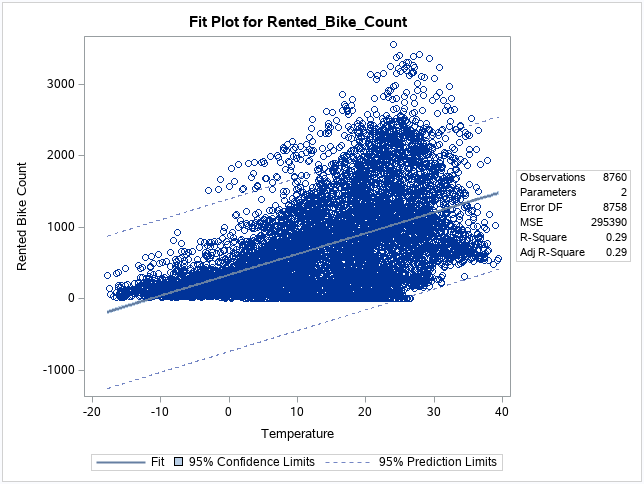
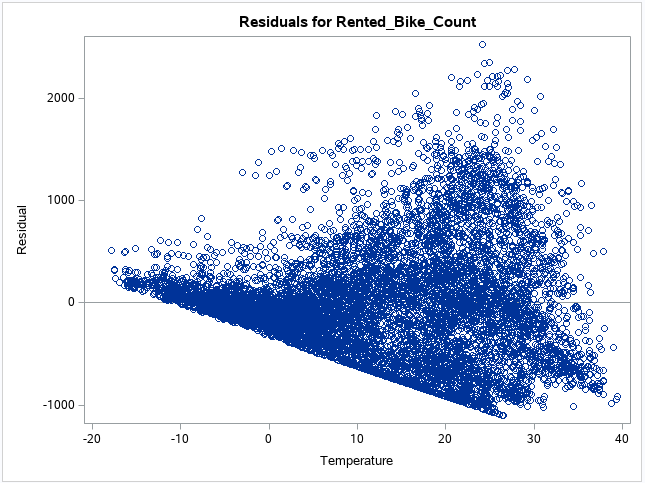
According the Parameter Estimates table, for each rise on temperature by 1 degree Celsius, roughly 29 more bikes will be rented. The intercept also tells us that at the baseline of 0 degrees, roughly 330 bikes are still rented by determined cyclists.



This first plot from the regression analysis shows the observed by predicted number of bikes rented with the predicted value running along the x-axis and the observed value along the y-axis. This particular plot suggests some homoskedasticity with its large deviation towards the top of the plot. While this likely violates an assumption of simple linear regression, it’s still useful for its predictive value.



The goodness-to-fit plots show a decent scattering in the residuals with the line of regression running though many of the points in left-middle Residual/Quantile plot. Furthermore, the histogram in the bottom-left shows a curve that is mostly normal. The determination from this plot matrix is that temperature is a good variable to use to predict bike rentals.



Finally, the Residual and Fit plots reveal more about the model. The Residual plot shows evidence for homoskedasticity with high number of points that do not fall close to the line. The Fit plot shows how many points fall within the 95% confidence interval and how many are within the 95% prediction limits. The area for confidence interval is small, but most of the points still fall within the 95% prediction limits.

The regression analysis shows that while temperature isn’t a perfect predictor of bike rentals, it’s still very useful in predicting number of bikes rented and suggests a correlation where the warmer the weather, the more bikes get rented.

**Conclusion**

Weather conditions are correlated with the number of bikes rented, thus rejecting our null hypothesis. More specifically, the ANOVA test showed there is a statistically significant difference in the number of bike rentals based on the season. In the Wilcoxon tests, you can clearly see the difference in bike rentals between the winter and summer seasons. Lastly, using a regression test, it was determined that even though temperature isn’t a perfect predictor of bike rentals, the difference is still quantifiable and significant.

**References**

* Dataset: https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand
* https://en.wikipedia.org/wiki/Ddareungi