John Paukovits

Intro to Data Science Project 1 DAND

Analyzing the NYC Subway Dataset

References

Stats Models - Ordinary Least Squares
Python.org - Datetime module documentation
GGPlot - Using all of their documentation as examples
Matplotlib - Axes Limits
Scipy.org - Scipy.stats.mannwhitneyu
Youtube - Great Github Video - https://youtu.be/0fKg7e37bQE

Section 1. Statistical Test

1.1

The Mann-Whitney U Test was used to determine with there was a statistical difference between Hourly Entries into the NYC Subway stations when it was raining and when it was not. This led me to use a two tailed test. The Null Hypothesis states that there is no significant difference between the two samples. Meaning the average hourly entries when it is raining is equal to the average hourly entries when it is not raining. I used an alpha value of .05 or 95% confidence. This means, if the Mann-Whitney U test returns a p value higher than 0.05 we would fail to reject the null.

1.2

The Mann-Whitney U Test is useful for this dataset because it does not assume the dataset is normally distributed. When looking at the histogram in section 3.1, it is clear the distribution of hourly entries in not normally distributed. Using a test, such as the Welch's T test would produce less reliable results.

1.3

The mean hourly entries on rainy days was 1105.45. The average for days it did not rain was 1090.28. The p-value returned by the Mann-Whitney U Test is a one tailed value of .0249.

1.4

The results of this test led me to reject the null. Even though the p-value is just below 0.05, we can say that there is a significant difference in hourly entries when it rains in NYC.

Section 2. Linear Regression

*This sections results are based on the improved data set and run on my local machine, the files are compiled in the Git Repository.

2.1

While completing problem set 3 I used both OLS in Statsmodels and Gradient Descent from Scikit Learn to produce predictions of Hourly Entries. I also mirror both problems on my local machine so I could run a few more variables and iterations

2.2

The features I used for both approaches were ‘rain’, ‘Hour’, and ‘weekday’. I left ‘UNIT’ as the dummy unit and attempted to add ‘stations’ as a second dummy unit when using the improved dataset. I wasn’t satisfied that ‘stations’ was making a large enough improvement in my R^2 as a dummy variable and with the risk of multicollinearity I chose to leave it out.

2.3

The variables I chose were based on trial and error but with some intuition mixed in. I did not attempt every combination of variables because I would guess some are just not as strong as others or too similar. For example I tried to substitute Meantempi with Maxtempi but didn’t see an improvement in my R^2. In the end, running both the OLS and Gradient Descent on my machine over and over again led me to the 4 variable and one dummy variable I chose.

*Sections 2.4 and 2.5 refer to Gradient Descent run on my local machine.

2.4

‘rain’ = -59.377
‘Hour’= 56.461
‘weekday’= 528.287

2.5

The R^2 produced by this approach is .4465

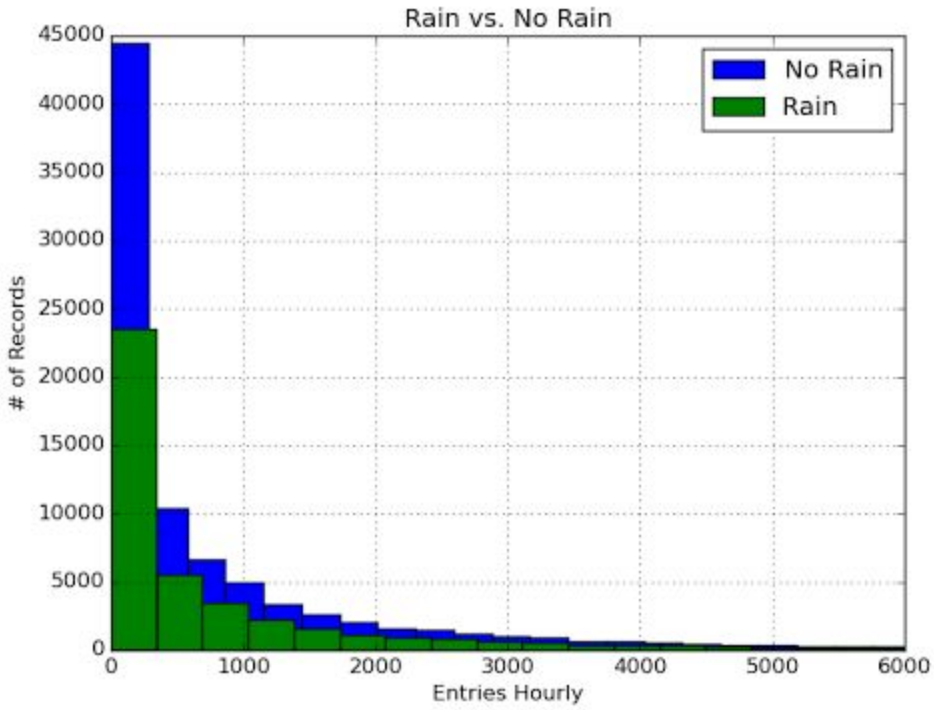
2.6

Although the Gradient Descent approached produced a lower R^2 than when I used OLS in the browser, It can still be used as a model to predict ridership in the NYC subway system. With a R^2 higher than point .4 we can say the model explains a good portion of ridership. The closer the R^2 is to 1, the more variation in the dependant variable can be explained by the independent variable. Specifically in this model, the independent variables chosen explain 44.65% of the variation in the dependent variable fo Hourly Entries.

However, a plot of the residuals derived from the difference between the final predictions and the actual hourly entries shows some very long tales. This means that occasionally throughout the dataset my linear model is not accurate, which is consistent with an R^2 of 44.65%. This should be taken into consideration when applying the predictions found with the model.

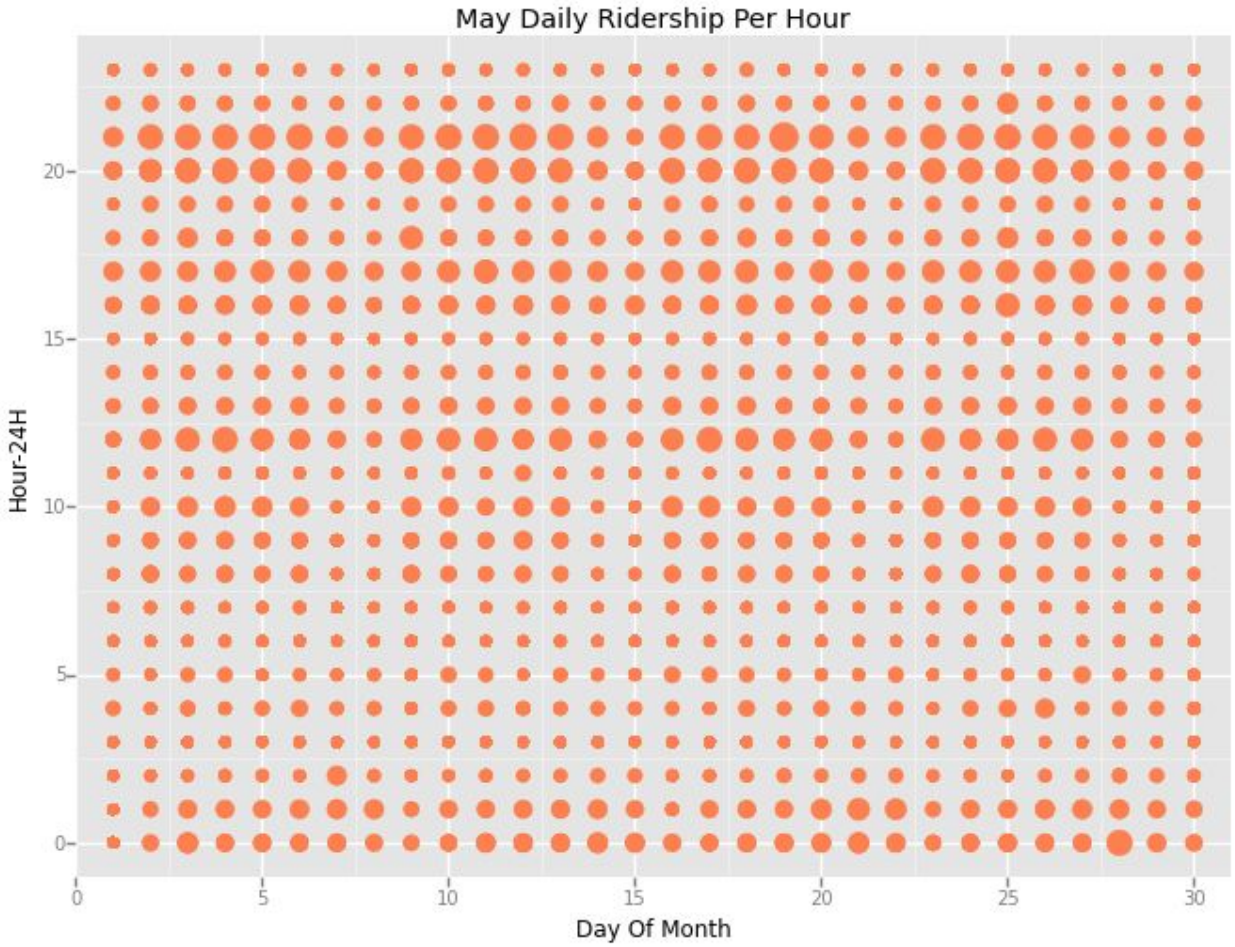
Section 3. Visualization

3.1



This is the histogram taken directly from problem set 3

3.2



This plot shows the Hourly Entries on each day of May by the hour. The size of the points indicate a greater amount of entries as size increases. From this plot we can deduce ridership may be greater during the week and the peak of most days is around 8:00pm. My guess is that post happy hour, the subway seems like a great idea.

Section 4. Conclusion

4.1

Based on the statistical tests done during lesson 3, I am confident that there is a difference between Entries into the NYC subways station when it is raining vs. not raining. When looking at the one tailed alpha value found in the Mann-Whitney U test we can see there is a statistical significance of more hourly entries when it is raining.

4.2

There are three main pieces of information that led me to this conclusion during lesson 3. The first is the histogram plotting hourly entries for both cases. This is a good starting spot to suspect there may be a difference. The second was being able to reject the null hypothesis for the Mann-Whitney U Test. Although the alpha value was very close to .05, the rules of the test can be strictly followed as established from the beginning. The Mann-Whitney U results can also be used to test which direction rain pushes ridership. When comparing the two means we can see the average ridership when it is raining is higher. At that point we can reject the null hypothesis if the p value returned is less than .025 and accept an alternate hypothesis that average hourly entries when it is raining is statistically greater. The third piece of evidence is the coefficient of rain found in the Gradient Descent approach talked about above. The rain variable has a similar effect to hour on the predictions.

Section 5. Reflection

5.1

1. The dataset itself is clearly limited by the timeframe. While completing the NYC Subway analysis I tried to think of myself as a manager trying to anticipate ridership, maybe for scheduling employees or directing train traffic. In that situation I would like data over a longer time frame. For example a variable like average temp may have more effect if we are able to sample a little more variation. Perhaps in May the average temp is fairly moderate. If we were able to include the more extreme average temps, there would be more potential to capture those effects.
2. The quality of any linear regression is always correlated to the quality of the variables you are able to choose from. When you look at the data of the original dataset, you can see that many of the variables will be strongly correlated to themselves. This will cause the problem of multicollinearity and hence limit the variable you are able to pick from.