Analyzing the NYC Subway Dataset

Do more people ride the subway when it is raining versus when it is not raining?

0. References

Anderson-Darling test: https://en.wikipedia.org/wiki/Anderson%E2%80%93Darling_test

Calculating p-value from Mann-Whitney_U:
http://stats.stackexchange.com/questions/116315/problem-with-mann-whitney-u-test-in-scipy

Dummy variable coding: http://www.ats.ucla.edu/stat/mult_pkg/faq/general/dummy.htm

Heteroscedasticity: http://www.statsmakemecry.com/smmctheblog/confusing-stats-terms-explained-heteroscedasticity-heteroske.html

Probability plot: http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.probplot.html

1. Statistical Test

1.1 Which statistical test did you use to analyze the NYC subway data? Did you use a one-tail or a two-tail P value? What is the null hypothesis? What is your p-critical value?

I used the Mann-Whitney U test to determine if the average turnstile entries on the rainy days tended to be higher than the average turnstile entries on the non-rainy days. The Mann-Whitney U tests if we draw randomly from each distribution (rainy and non-rainy), whether one distribution is more likely to generate a higher value than the other.

Given random draws x from the population of turnstile entries on rainy days and random draws y from the population of turnstile entries on non-rainy days, the two tailed hypotheses are:

Null hypothesis: Probability (x > y) = 0.5

Alternative hypothesis: Probability (x > y) ≠ 0.5

The critical value is the U-value from the Mann-Whitney test, which is: 153635120.5.

1.2 Why is this statistical test applicable to the dataset? In particular, consider the assumptions that the test is making about the distribution of ridership in the two samples.

A histogram of the turnstile entries per hour (Figure 3) indicates that the data is right-skewed. Using the Anderson-Darling test for confirmation, we reject the hypothesis that the data are normally distributed (p-value < 0.0001, alpha level: 0.01). Therefore, the Mann-Whitney test is applicable because as a non-parametric test, it does not make assumptions about the underlying data distributions.

1.3 What results did you get from this statistical test? These should include the following numerical values: p-values, as well as the means for each of the two samples under test.

	Rain	No Rain
Mean	2028.20	1845.54
Median	939	893
Interquartile Range	2129	1928

P-value for two-tailed test: 5.48e-06

1.4 What is the significance and interpretation of these results?

The average NYC subway ridership on rainy and non-rainy days is not the same. The averages are significantly different at an alpha level of 0.01.

2. Linear Regression

2.1 2.1 What approach did you use to compute the coefficients theta and produce a prediction for ENTRIESn_hourly in your regression model?

I used Ordinary Least Squares from Statsmodels.

2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?

My final model included: hour, rain, day_week, pressure and also included station as dummy variables. I used reference coding for the dummy variables with `station_1 Ave` as the reference level.

2.3 Why did you select these features in your model? We are looking for specific reasons that lead you to believe that the selected features will contribute to the predictive power of your model.

Input Variable	Reason it was Included
Rain	An objective was to determine if ridership on the subway increased when it was raining.

Day_of_week	The exploratory analysis as shown in Figure 4 indicates that there is a difference in ridership by day of the week.
Hour	It seemed likely that ridership would depend on the time of day. For example, there may be more riders at 9:00am than at 1:00am.
Pressure	The t-test on pressure's beta value indicated that it was significant
Station	It seemed logical that ridership would vary by station. Without station or its proxy, turnstile unit, in the model, the adjusted R^2 value decreased significantly.

2.4 What are the parameters (also known as "coefficients" or "weights") of the non-dummy features in your linear regression model?

Input Variable	Coefficient
Rain	70.4161
Day_of_week	-143.0418
Hour	122.6634
Pressure	-394.6688

2.5 What is your model's R2 (coefficients of determination) value?

R^2: 0.418

Adjusted R^2: 0.416

2.6 What does this R2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R2 value?

The R^2 value means that 42% of the variability in the data is explained by the model. I wonder if there are other non-linear models that might be a better fit for the data, or if perhaps it would be wise to test for interactions between variables.

There are a several reasons why a linear model may not be the best choice for the data:

- The variable of interest, `ENTRIESn_hourly` is conceptually bounded at 0. Negative predictions do not make sense.
- `ENRIESn_hourly` is a discrete variable. Fractional predictions do not make sense.
- An examination of the residuals from the linear model reveal a heteroscedastic pattern of decreasing variance that is not accounted for by the predictions. Constant variance is an assumption for linear regression, and this assumption appears to be broken in this model.

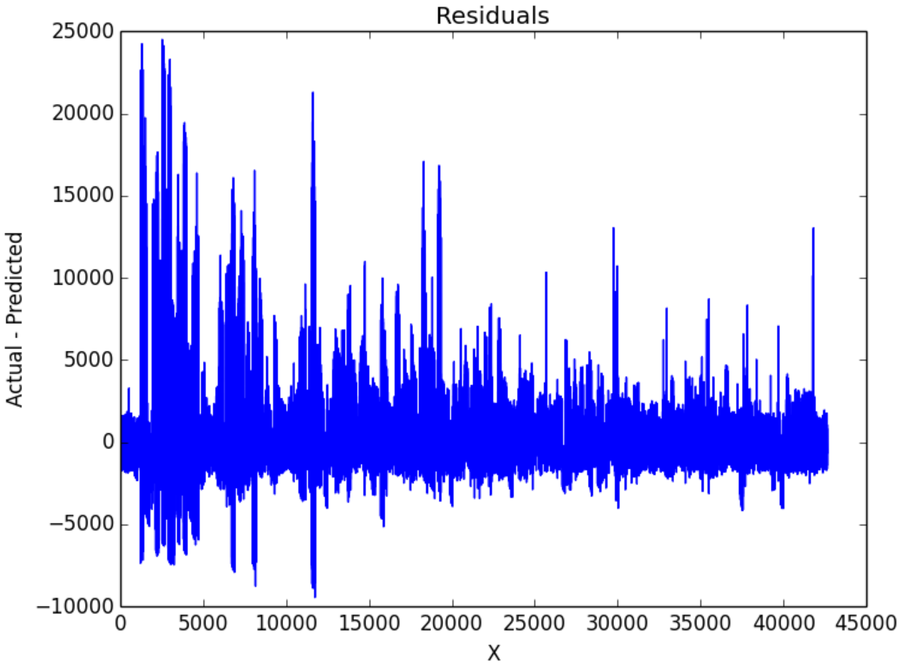

Figure 1: Residual Plot

- This model also violates the linear regression assumption that the residuals are normally distributed. The probability plot in Figure 2 shows that the residuals have a long tail and are not normally distributed.

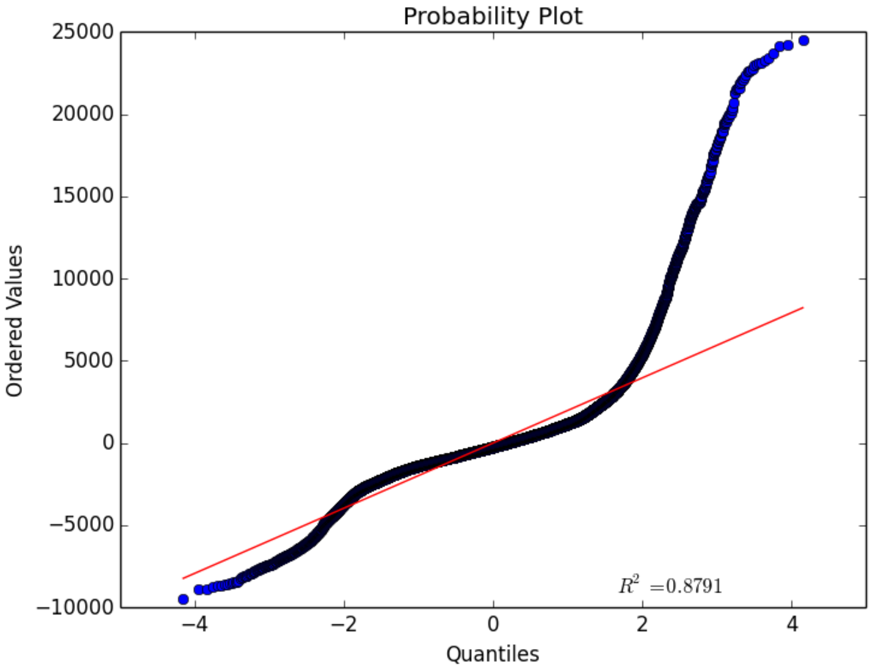

Figure 2: Probability Plot of Residuals

3. Visualization

3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

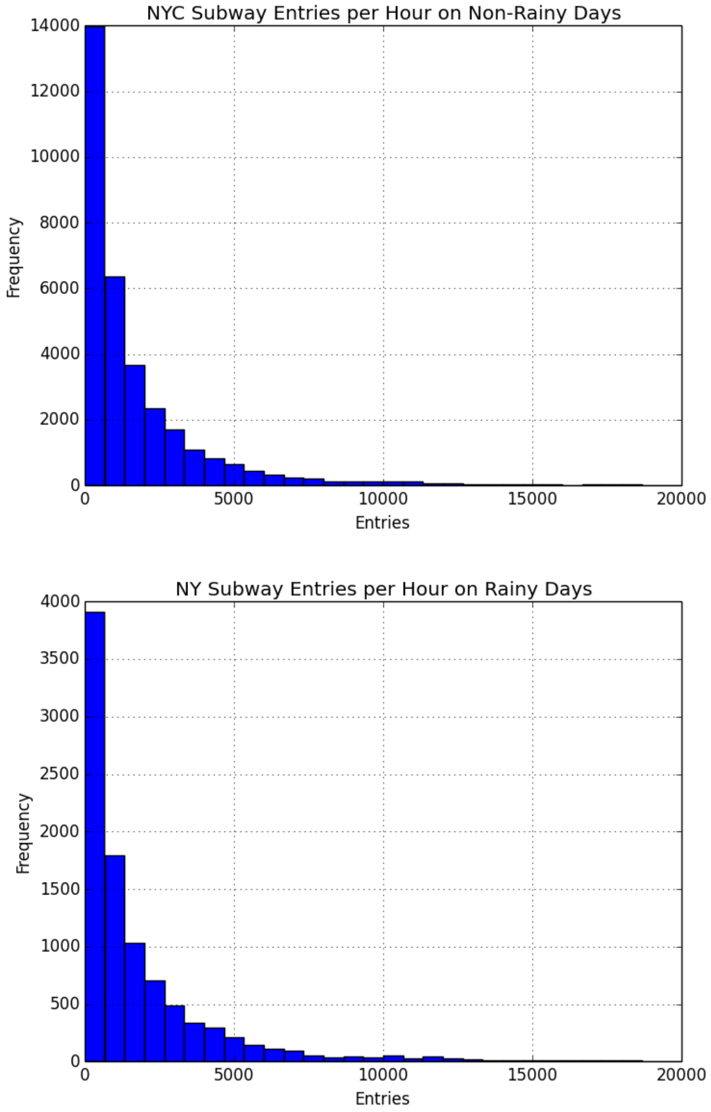

Figure 3: NY Subway Entries per Hour (Rain and No Rain)

Figure 3 shows that both distributions are right skewed. The x-axis in the figure has been truncated, cutting off outliers in the far tail. There are many fewer instances of rainy days than non-rainy days in the sample.

	Rain	**No Rain**
N	9585	33064

About 22% of the days in the sample are rainy.

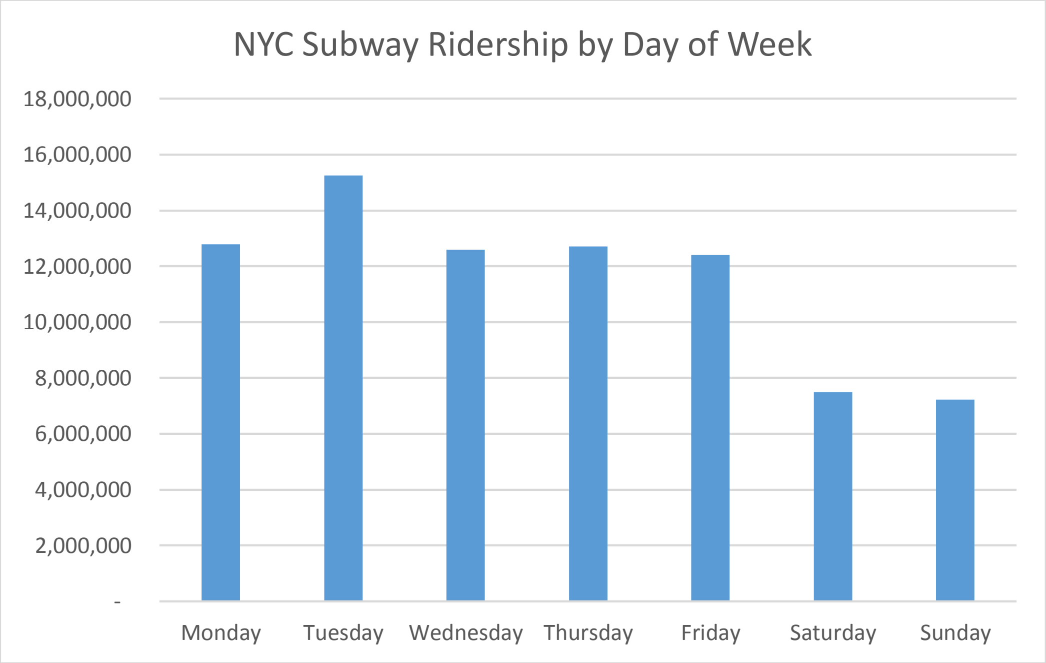

Figure 4: Subway Ridership by Day of Week

The ridership appears to fluctuate by day of the week, with fewer subway riders on Saturday and Sunday.

4. Conclusion

4.1 From your analysis and interpretation of the data, do more people ride the NYC subway when it is raining or when it is not raining?

4.2 What analyses lead you to this conclusion? You should use results from both your statistical tests and your linear regression to support your analysis.

By the analysis presented above, more people ride the NYC subway when it is raining compared to when it is not raining. The Mann Whitney U test leads to the conclusion that there is a difference between the ridership distributions on rainy versus non-rainy days and that ridership tends to be higher on rainy days. Then the model derived from linear regression shows that the rain variable (with values: rain or no rain) is a significant predictor in the model. The coefficient on the rain variable is positive which means that if all other variables are held constant, we'd expect to see an increase of 70 entries in each hour when it is raining versus when it is not raining.

5. Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis, including: Dataset, Analysis, such as the linear regression model or statistical test.

5.2 (Optional) Do you have any other insight about the dataset that you would like to share with us?

The R^2 and adjusted R^2 values seemed low. It is possible that a better result could have been achieved by testing for interactions or testing quadratic terms in the model. I'd like to learn variable selection techniques in python in order to have a better method than trial and error to choose variables for the model.

The dataset is time series data, which implies that an ordinary least squares model is not a good choice. The data represents turnstile entries for the month of May, 2011. Having one month of data coverage may not be sufficient since it might be reasonable to expect a seasonal trend for subway ridership.

The model also had problems with multicollinearity, meaning that there were variables that were contributing similar information. I thought that latitude and longitude could potentially have been good analogs for subway station, which could have benefitted the model by reducing the number of variables, but without subway or turnstile unit in the model, the R^2 and adjusted R^2 values decreased dramatically.