Steven Yeh

Project 2: Analyzing the NYC Subway Dataset

Section 0: Sources

OLS Critique: http://www.clockbackward.com/2009/06/18/ordinary-least-squares-linear-regression-flaws-problems-and-pitfalls/

https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test

MTA Subway Ridership: http://web.mta.info/nyct/facts/ridership/#chart_s

Section 1: Statistical Test

1.1

I used the Mann-Whitney U-test. This test gives a one-sided P value by default, which is also what I decided to use because we can assume that the population of people riding the subway on a rainy day should tend to contain larger values than the population on a sunny day.

The null hypothesis in relation to the Mann-Whitney U-test is that the two populations are the same. The populations of people riding the subway on non-rainy and rainy days are the same. The alternative hypothesis would be: "The probability that an observation from the rainy day population is greater than the non-rainy day population is greater than the probability of the opposite."

Upon running the test, the calculated one-tail P value was ~0.025.

1.2

The Mann-Whitney U-test was appropriate since I don't want to assume that this data comes from any specific probability distribution. There is not much justification to assume that the two samples derive from a normal distribution or that the two samples have equal variance. Therefore, a t-test would not be appropriate, whether it be the Student's t-test or Welch's t-test.

1.3

Even though the Mann-Whitney U-test by itself only tests whether or not the two samples comes from the same population, a comparison of the sample means of the two populations shows that on average, rainy days resulted in more subway ridership: Rainy days had a mean of ~1105 hourly entries, non-rainy days had a mean of ~1090 hourly entries. The Mann-Whitney test statistic was 1924409167.0, which corresponded with a p-value of ~0.025.

1.4

A p-value of 0.025 indicates that the probability of obtaining a test statistic at least as extreme as ours (1924409167.0) if the null hypothesis was true is 0.025. If we are using an $\alpha = 0.05$ significance level, then we can state that this result is statistically significant since $0.025 < 0.05$. We can reject the null hypothesis at this level $\alpha = 0.05$, and accept the alternative hypothesis: "The probability that an observation from the rainy day population is greater than the non-rainy day population is greater than the probability of the opposite." The alternative hypothesis reflects the decision to use a one-sided test.

Section 2: Linear Regression

2.1

I used OLS using Statsmodels to perform a standard linear regression.

2.2

I used the features 'rain', 'precipi', 'Hour,' and 'fog'. The dummy variable 'UNIT' was used to separate the subgroups (each individual remote unit).

2.3

'rain' was selected based on the intuition that individuals choose the subway more often on a rainy day. 'precipi' was selected because I believed individuals are more inclined to ride the subway as the amount of precipitation increases. 'Hour' was selected because more people are expected to ride the subway during peak hours (when people go to and from work). 'fog' was chosen as individuals are more likely to ride the subway than walk or drive when visibility is low. 'meantempi' marginally increased by R^2 value (+ <.001), and there is not a clear relationship that can be easily stated, so I did not include this variable in my model. Perhaps on a warmer day, people might decide to walk on a pleasant day, but it is also entirely possible that individuals might enjoy the air-conditioned subway car to escape the heat, even if it might be exacerbated on the platform.

2.4

The parameters for the non-dummy features in my model are as follows:

```
rain           12.688650
precipi        12.247303
Hour           65.304836
fog           133.131960
```

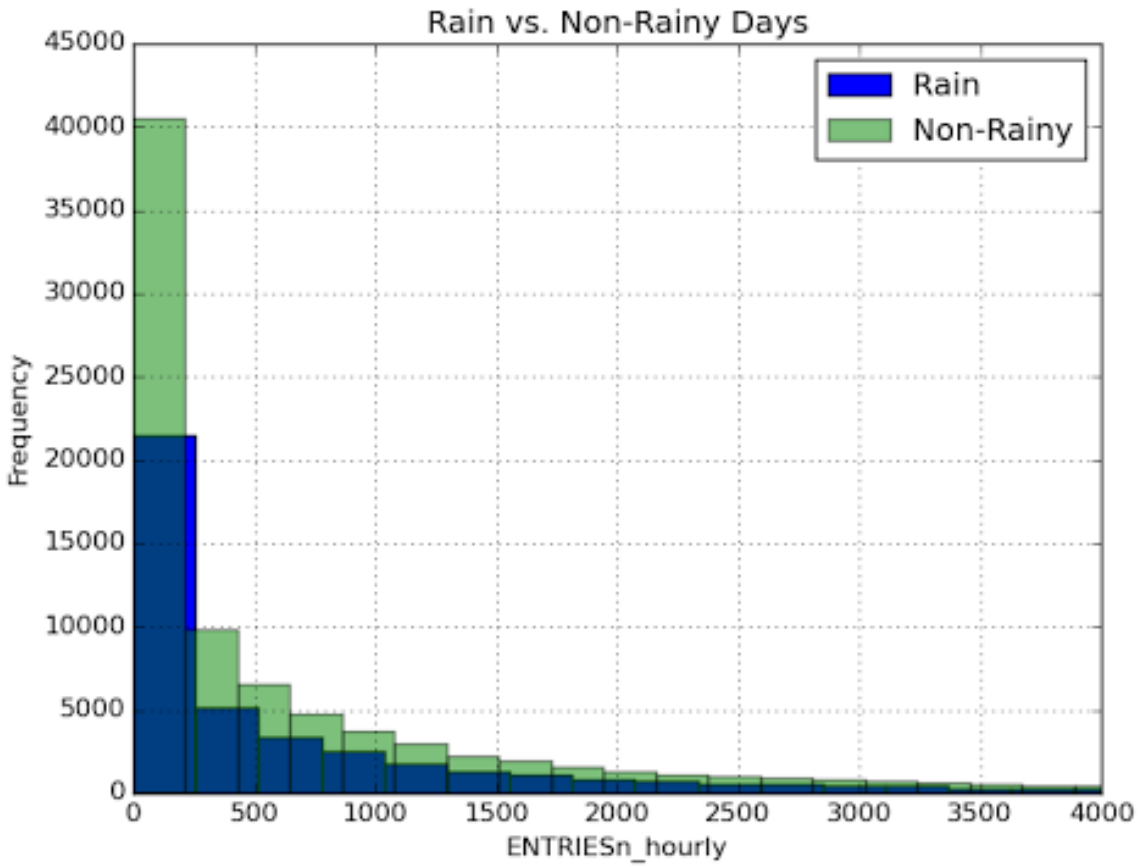
2.5

The model's calculated R^2 value is ~0.4787

2.6

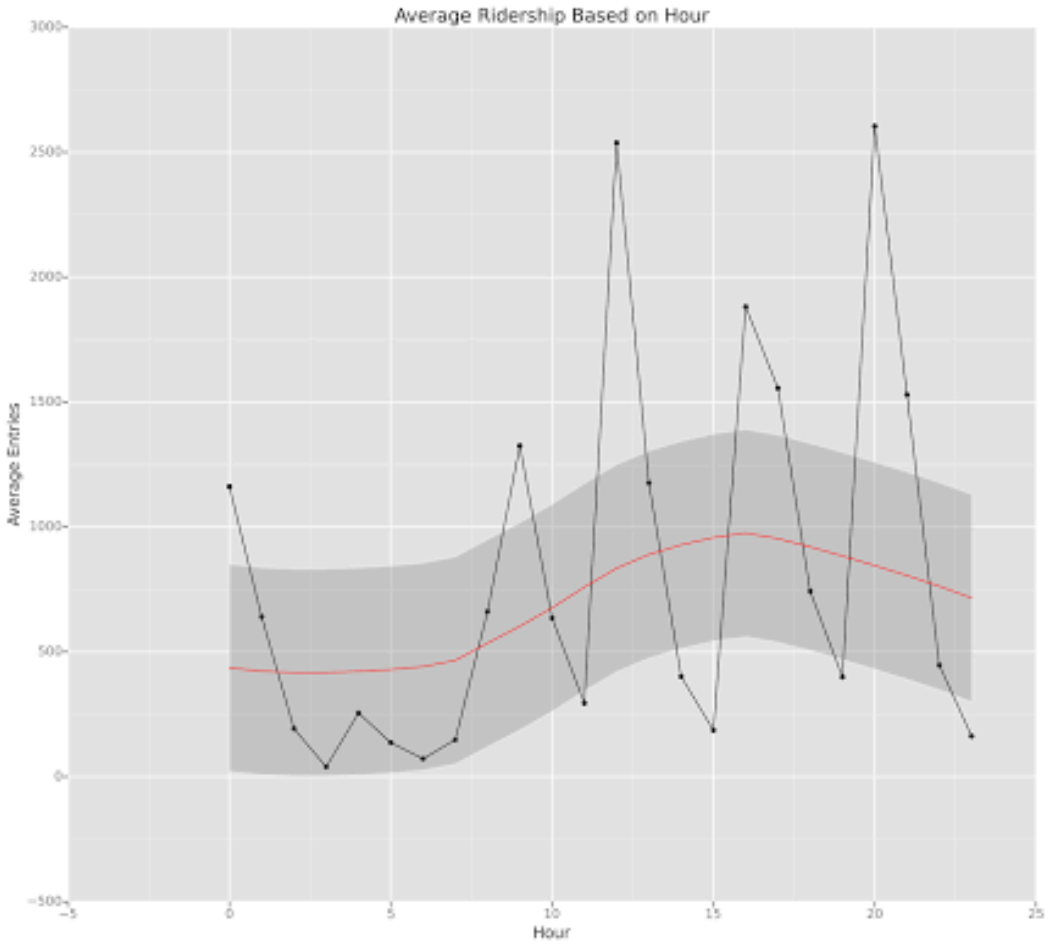
This R^2 value means that this model explains ~47.87% of the variability produced. Of course, an R^2 much closer to 1 would be preferred, as this indicates a perfect fit. I believe that this linear model is not the best given its low R^2 value; there are other variables that exist but for which the data is not present in the table or not collected.

Section 3: Visualization
3.1



A histogram exhibiting frequency of ridership shows that the distributions of riders on rainy and non-rainy days are not from a normal distribution. Nothing can be stated by the relative height of the histograms since there are significantly less data points for "Rainy."

3.2



A line graph with the average number of entries across all units for each hour gives insight into the peak and valley hours where ridership is varied. In addition, a smoothing line was added to provide a clearer look at the pattern. One can see that ridership begins to peak around 8-9 am, with a maximum peaks in the early afternoon to the evening. People traveling to and from work would explain the peaks in the morning and evening. However, since some hours are collected more frequently than others, it is difficult to explain some peaks, such as the one around noon-1pm. Hourly ridership data for each subway station would strengthen this plot.

Section 4: Conclusion

From my analysis of the data, I can conclude via the Mann-Whitney U-test that more people do take the subway on rainy days as opposed to non-rainy days. As evidenced by the histogram and common sense, the Mann-Whitney U-test was appropriate since the data did not appear to come from a normal distribution. The p-value corresponding to the test-statistic was ~0.025, which rejects the null that the ridership counts are the same for both groups and accepts the one-sided alternative hypothesis. The linear regression of the data returned positive parameters for 'rain', and 'precipi', which also indicates that ridership increases on rainy days.

Section 5: Reflection

Possible shortcomings of my analysis include flaws within the dataset and within the linear regression model. The dataset fails to account for extra weather conditions such as whether the sky is overcast or not, or the probability of rain on a given day. It would be reasonable for people to check the weather before choosing a transportation method. Conceivably, someone could check the weather, notice a 75% chance of rain, and elect to take the subway regardless of whether it actually rained that day or not. Some columns were also not very meaningful, such as the measured barometric pressure or dew point, and were thus superfluous to the analysis. It would also be beneficial if the hour intervals were less. Hourly entry and exit data would result in a higher accuracy when conducting analysis.

The linear regression model could have a much better R^2 value, whether it is adding more pertinent variables in the dataset, collecting data for new variables, or manipulating the data itself. There is also the matter of dependence on variables on each other; precipitation and rain are closely related, and these two could also be related to fog.

In general, to predict subway ridership, maybe weather is not the best direction due to the interconnectedness of all the variables. The location of the subway station, or even better, the relative population density surrounding the station could potentially result in more accurate predictions, even if it might be difficult determining the area or range around each station to calculate this density (some stations are close apart; others are a few minutes' ride).