Analyzing the NYC Subway Dataset

Intro to Data Science – April ’15 Cohort
Cameron Mcleod

Overview

This project consists of two parts. In Part 1 of the project, I completed the questions in Problem Sets 2, 3, 4, and 5 in the Introduction to Data Science course.

This document addresses part 2 of the project. In this project, I looked at the NYC Subway data and figure out if more people ride the subway when it is raining versus when it is not raining. The data set I used was the improved dataset “turnstile_weather_v2.csv”.

In order to figure out if more people ride the subway when it is raining versus not raining, I wrangled the NYC subway data, and used statistical methods and data visualization to draw an interesting conclusion about the subway dataset.

Section 0. References

http://stackoverflow.com/questions/944700/how-to-check-for-nan-in-python

http://stats.stackexchange.com/questions/116315/problem-with-mann-whitney-u-test-in-scipy

http://discussions.udacity.com/t/problem-set-3-8-more-linear-regression-optional/15337

http://statsmodels.sourceforge.net/0.5.0/generated/statsmodels.regression.linear_model.OLS.html

http://pandas.pydata.org/pandas-docs/stable/generated/pandas.get_dummies.html

http://www.itl.nist.gov/div898/handbook/pri/section2/pri24.htm

http://stackoverflow.com/questions/15160123/adding-a-background-image-to-a-plot-with-known-corner-coordinates

http://discussions.udacity.com/t/problem-set-3-item-1-how-to-plot-histogram-just-like-in-the-instructor-note/19480

Section 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?

The statistical test I used to analyze the NYC subway data was the Mann Whitney U-test. I used a two-tailed P value, as the null hypothesis was "there is no difference between subway ridership when it is raining vs. not raining". My p-critical value is -0.01 for a two tailed test, therefore for rainy day ridership to be statistically greater or less than non-rainy day ridership (the null is rejected), the p-value must be < 0.005.

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.

The Mann-Whitney U-test was applicable because there is no underlying assumption that the data sets fall under a normal distribution. We can see from the histogram of entries for rainy and non-rainy days, that the data definitely does not follow a normal distribution (using a Shapiro-Wilk test would also be a good method to show the data is not drawn from a normal distribution). Additionally, the assumption is made that the two samples are independent, and the data is ordinal.

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.

Using the Mann-Whitney U-test on the improved dataset, I found the mean entries to be 2028.20 entries per hour for rainy days, and 1845.54 entries per hour for non-rainy days. When I used the function:

scipy.stats.mannwhitneyu(Entries_rain, Entries_no_rain)

a value "nan" was returned for the p-value. However we can calculate p using a normal approximation of U, due to our large sample size. The code for which is attached. With the normal approximation of U, I determined the p-value to be 0.0000055 (5.5 x 10^-6)

1.4 What is the significance and interpretation of these results?

With our chosen 99% confidence interval, we require a p-critical value < 0.005 for our two-tailed Mann-Whitney U-test. Our returned p-value was 0.0000055, much less than our p-critical value. Therefore we can reject the null hypothesis, and confirm that there is a statistically significant difference between ridership on rainy vs. non-rainy days.

Section 2. Linear Regression

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

B. OLS using statsmodel – I used this method because it is a relatively simply implementation and I am making the assumption that the variables are linearly correlated.

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

I used 'rain', 'hour', 'meantempi' and 'weekday' in my model. For dummy variables I used 'UNIT' and 'conds'.

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.

I used 'rain' because that is the main variable we are testing, from our statistical test there is a high probability of rain determining ridership.

I used 'hour' because the time of day will have a large effect on the ridership for example there should be fewer riders between 2-5 am because most people are sleeping.

I used 'meantempi' because I believe when it is either very hot or very cold outside, subway ridership will be influenced

I used 'weekday' in my model because I assumed that most commuting via subway happens during weekdays, and people are more likely to use vehicles on weekends (going out of town).

I used 'UNIT' because there are probably stations which are vastly more busy than others.

I used 'conds' because people may choose to use the subway if it is overcast or "looks like rain" out, even if it in fact does not rain. Also on a nice day people may be more inclined to walk, and be outside.

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

Feature	Weight
Rain	40.42
hour	856.92
meantempi	-141.96
weekday	424.68

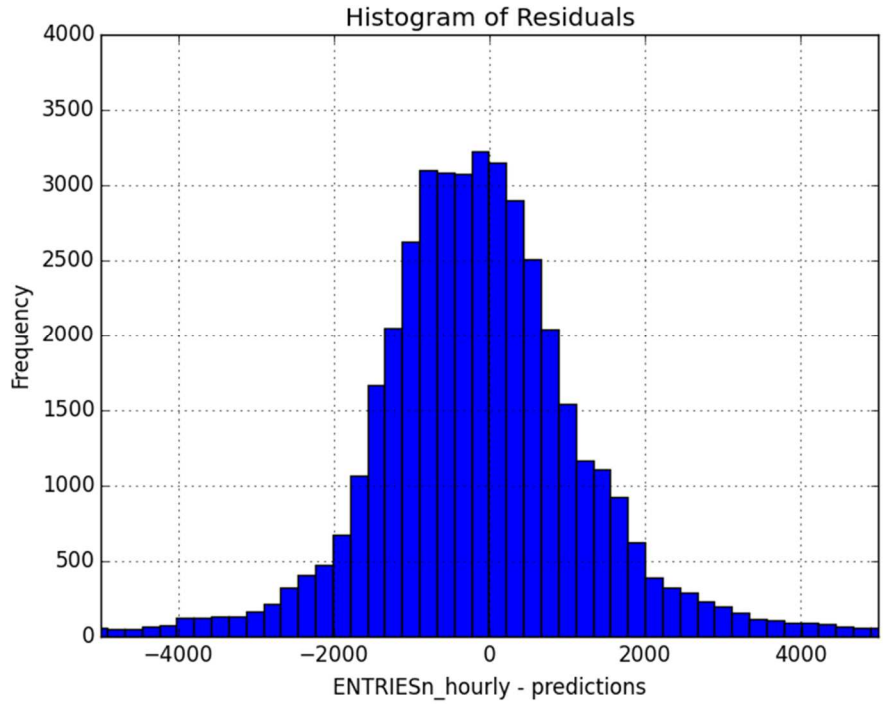
Additionally I added a constant to the fit, which came out to const = 1881.77

2.5 What is your model's R^2 (coefficients of determination) value?

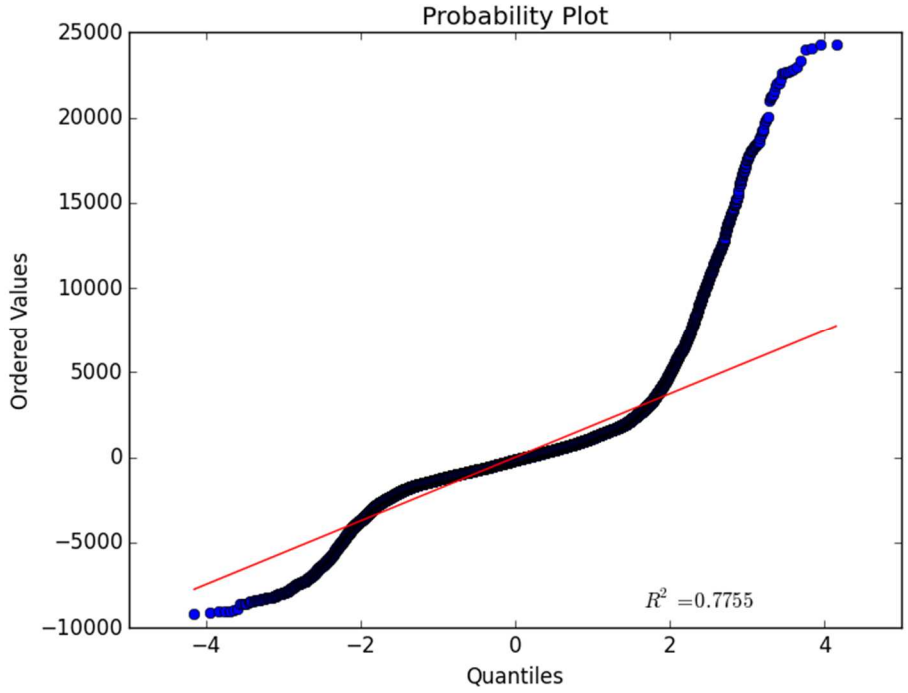
My model's R^2 value was 0.4875

2.6 What does this R^2 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 R^2 value?

To evaluate the goodness of fit I have plotted the residuals below:



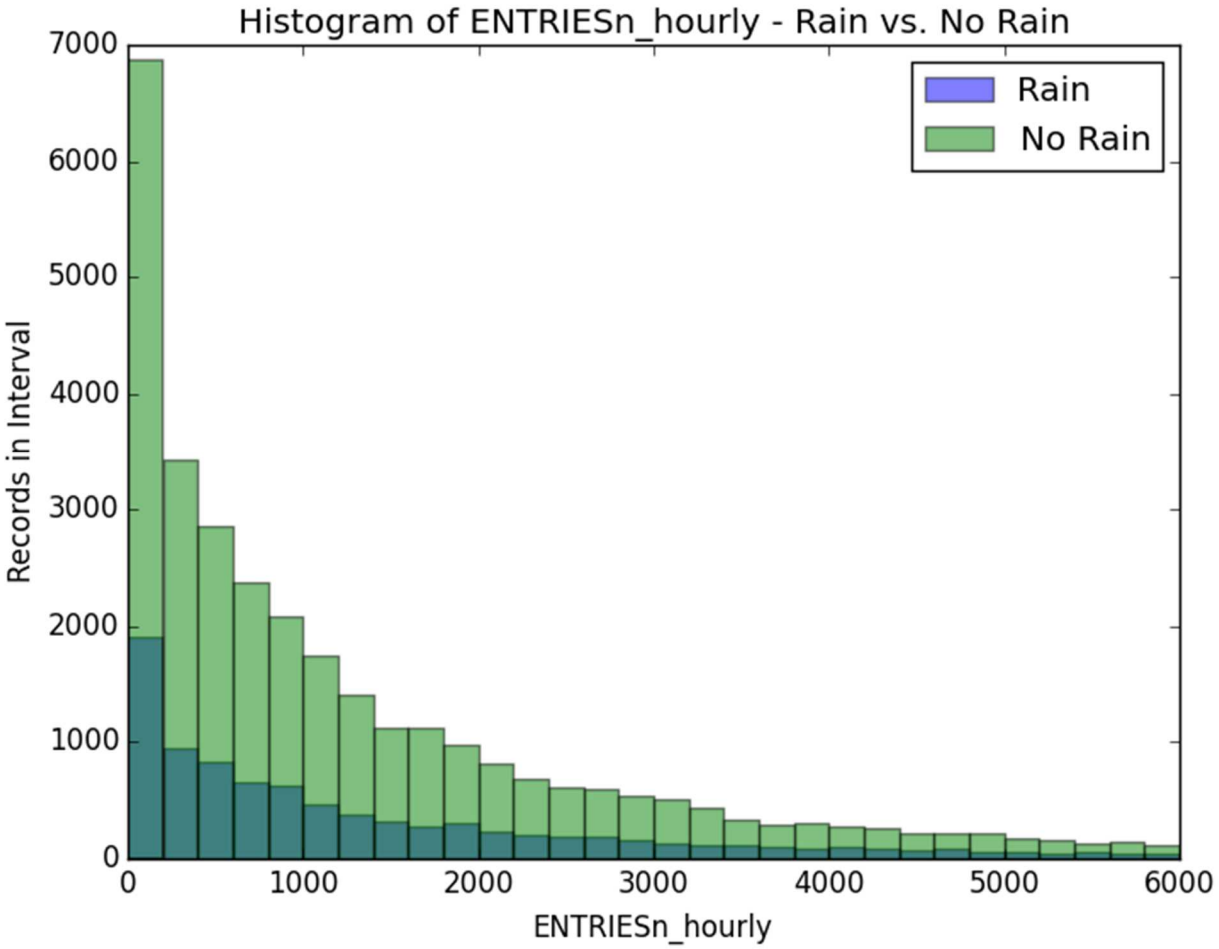
From the plot above, we see our residuals histogram follows a near-normal distribution, possibly demonstrating the appropriateness of our model. An important note is that the mean of our distribution is not at 0, but around -400 entries per hour. This means on average we under predict the hourly entries by approximately 400 per hour. Since our R^2 value is 0.4875, which is not very close to a perfect model with a score of 1, we can assume there will be some inaccuracies in our model. In terms of our baseline R^2 value of 0.4, this R^2 value of 0.4875 is an improvement over the in-class model. On the following page I have a Q-Q Plot which better shows the inaccuracy of our model.



From this plot, we can see the non-normality and non-linearity of our dataset come through. The best fit line fails to accomodate the curve and we see a heavy departure of the predictions from the data, showing the failure of the model for the lower and upper quantiles . From the curvilinear nature of the Q-Q plot, the dataset is clearly non-linear and we will need an improved model if we are to support more meaningful conclusions from this analysis.

Section 3. Visualization

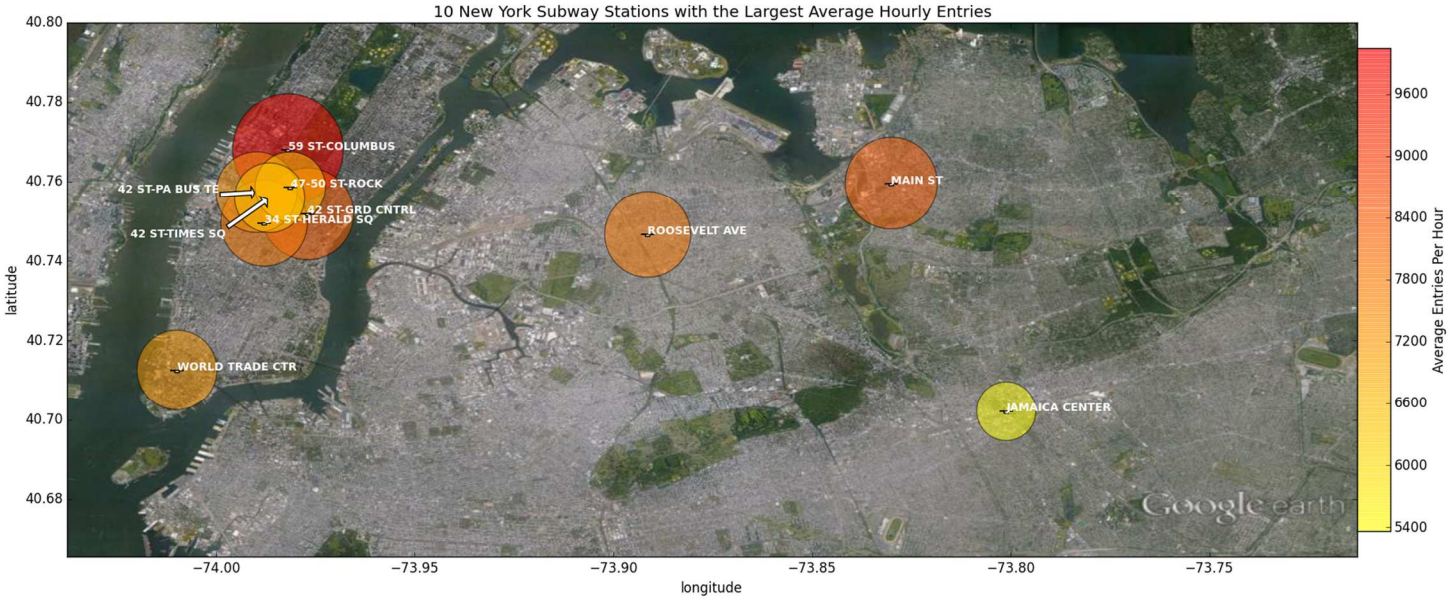
3.1 Two histograms: ENTRIESn_hourly for rainy days and ENTRIESn_hourly for non-rainy days.



In the above histogram ENTRIESn_hourly is depicted for rainy and non-rainy days. Key insights from this diagram include:

- There are considerably less rainy observations than non-rainy observations.
- Both histograms show a similar pattern that has the appearance of exponential decay
- For both histograms the mean > median > mode. And the mode is between 0 and 200 entries hourly – meaning during most hours observed, there were very few riders.
- The graph does cut off the tail of the histogram for readability – however it should be noted there are entries exceeding 30,000 / hour, so although the histogram shows the relationship between the two datasets very well, it does not show the whole picture.

3.2 Freeform Visualization: Top 10 Stations for Average Hourly Entries – Map overlay



In the above visualization the top 10 stations for average “ENTRIESn_hourly” are depicted. These are plotted in a scatterplot with point size and color correlating to average entries per hour. The plot is overlaid on a map using the latitudes and longitudes of each station (“new York top 10.png” attached). Key insights from this visualization include:

- 6 of the 10 busiest stations are in the grand central station / south central park area.
- The top average station has ~10,000 entries per hour.
- Even in the top 10 stations, there is a large range in the average hourly entries (between 10,000 and 5,400)
- There is an idea of scale to what area the New York subway system covers.

Section 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?

From the statistical tests I can say with 99% confidence that more people ride the NYC subway when it is raining. Therefore I was able to reject the null hypothesis, and confirm that there is a statistically significant difference between ridership on rainy vs. non-rainy days. Additionally, using linear regression I found a correlation between subway ridership and whether it was raining or not, which was positively correlated with rain.

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.

Using the means of 2028.20 entries per hour for rainy days, and 1845.54 entries per hour for non-rainy days, my Mann Whitney U test provided a p-value of 5.5x10^-6. Therefore the difference in means is statistically significant ($p < 0.005$ at 99% confidence).

The Ordinary Least Squares method for linear regression from the statsmodel module allowed me to find the correlation coefficient for rain. Rain and ridership were found to be positively correlated with a coefficient of 40.42, giving evidence that is rain increases, ridership increases. This is reinforced by our residuals histogram, which follows a near-normal distribution, demonstrating the appropriateness of our model.

Section 5. Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis, including:

Potential shortcoming of the dataset include the date range over which the data was collected. Since we only have data from the month of May, there is a possibility that in other months ridership is inversely correlated with rain, or there is no statistical significance in the difference of means. An additional shortcoming of the dataset could be that it doesn't include for when stations may have been shut down for maintenance.

Shortcomings in the analysis could include the lack of confidence intervals on my parameters for the linear regression, additionally I believe the data may have been underfit for some features – for example correlation to outdoor temperature may be bimodal and not linear – people could avoid extreme heat or cold by taking the subway. Additionally, my linear regression residuals histogram shows a slight skew for predicted ridership, of around -400 riders per hour. Insight from the Q-Q plot demonstrates that the dataset is clearly non-linear and we will need an improved model if we are to support more meaningful conclusions from this analysis. An improved model accounting for non-linearity would hopefully resolve that inaccuracy.

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

An interesting thing I found from the linear regression model were the coefficents of my features:

Feature	Weight
Rain	40.42
hour	856.92
meantempi	-141.96
weekday	424.68

Weekday was highly correlated to subway ridership, leading me to believe that most people use the subway for commuting to work. A way to test that hypothesis would be to find the hourly entries and exits for subway stations that are in major business centers in the city. Additionally the mean temperature was negatively correlated, meaning as the temperature goes down, ridership goes up. If we had some winter data this hypothesis could be tested.

In my freeform visualization, the top 10 stations for average hourly entries, it was interesting to see some of the busiest stations outside of Manhattan, from further research I found that these stations are connected to either bus depots or ports. As such, they become bottlenecks for subway traffic which could explain their high levels of ridership.