.Analyzing the NYC Subway Dataset
Michal Nadolny

Overview

In this project, you look at the NYC Subway data and figure out if more people ride the subway when it is raining versus when it is not raining.

This project consists of two parts. In Part 1 of the project, you should have 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. Please use this document as a template and answer the following questions to explain your reasoning and conclusion behind your work in the problem sets. You will attach a document with your answers to these questions as part of your final project submission.

I appreciate all the great feedback from the Udacity coaches have provided me to continue improving this project. My references are listed at the bottom of this document. I have used the Improved data set- this version contains extra data points and variables outlined in this document. This discussion post stated this data set can be used.

To start the project I first took a look at some sample data so I can understand the data structure through reading the csv into a dataframe, then using .head() to take a peek at the structure of the data as well as explored the weather variables.

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?

I ran the Mann-Whitney U two-tail test with a null hypothesis (H0) of: There is no difference between the ridership during days it rained and days it did not rain meaning the two populations are the same. My p-critical value is .05

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 two tail test is applicable to the data set because the test does not assume the data is drawn from any particular probability distribution. “The Mann-Whitney U test (also known as Wilcoxon rank-sum test) is a non-parametric test that can be used to test, for two populations with unknown distributions, if we draw randomly from each distribution, whether one distribution is more likely to generate a higher value than the other.” There is also a large set of values and the data is non-normal, making this test applicable.

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.

MannWhitneyU test U= 153635120.5 , p= nan

(See code needed to calculate p with more expanded code because it is not calculating correctly)

P value= 2.74134693571e-06 = 0.000002

P times two because two tailed= 5.48269387142e-06 = 0.000005

No Rain Entries Mean: 1845.53943866

Rain Entries Mean: 2028.19603547

1.4 What is the significance and interpretation of these results?

Since the resulted p-value 0.000005<0.05(p-critical) this means the null hypothesis is rejected. This indicates that there is a difference in ridership during rainy days and 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:

1. **Gradient descent (as implemented in exercise 3.5)**
2. **OLS using Statsmodels**
3. **Or something different?**

To compute the coefficients theta and produce prediction for ENTRIESn_hourly in regression model I applied OLS using Statsmodels.

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, tempi, and pressurei as input variables. And UNIT was my dummy variable.

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.

Here are the reasons for each:

Rain and **hour** were the main features needed to compare ridership.

Tempi because my intuition told me temperature would be a factor in how many people ride.

Pressurei was also based on intuition because I know may people that are affected with strong headaches when the pressure changes, makes them want to call out of work.

The dummy variable of **UNITS** was used as it improved my R2 value from 0.349 to 0.460. However it also added the warning of there might be an indication that there are strong multicollinearity or other numerical problems in my summary.

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

rain	183.8770
hour	121.3171
tempi	4.6235
pressurei	-141.1016

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

My R2 value came out to be 0.460

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?

I found the R2 value was low, meaning the goodness of fit of my regression model was not that good and is low to produce accurate predictions. More details of my OLS regression results are shown in my ipython notebook. No, I don't think this linear model is appropriate for this dataset, plotting residuals which is the differences between predicted and actual values might give us some more insights into this data.

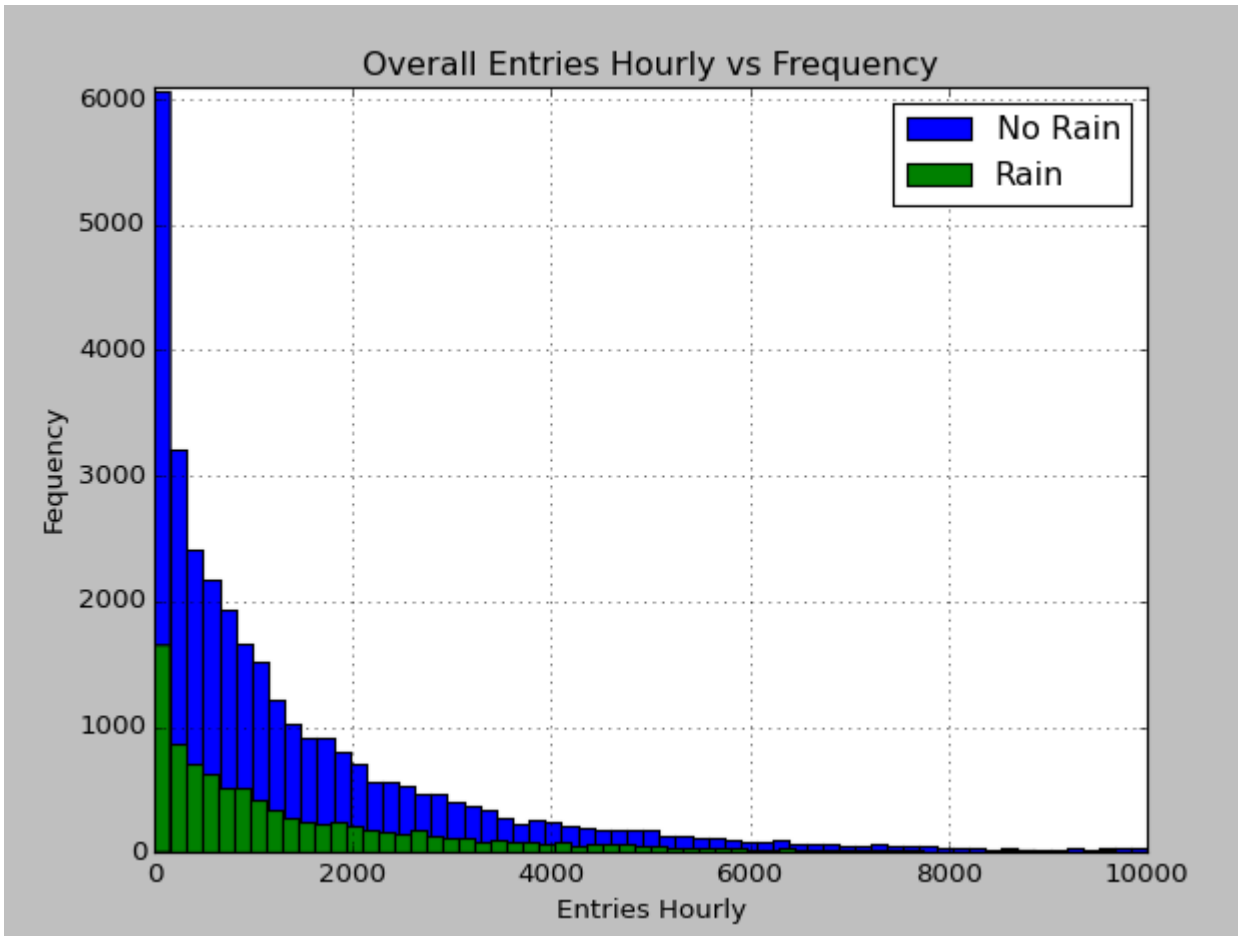
Section 3. Visualization

Please include two visualizations that show the relationships between two or more variables in the NYC subway data. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots, or histograms) or attempt to implement something more advanced if you'd like.

Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

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

- **You can combine the two histograms in a single plot or you can use two separate plots.**
- **If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.**
- **For the histograms, you should have intervals representing the volume of ridership (value of ENTRIESn_hourly) on the x-axis and the frequency of occurrence on the y-axis. For example, each interval (along the x-axis), the height of the bar for this interval will represent the number of records (rows in our data) that have ENTRIESn_hourly that falls in this interval.**
- **Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.**

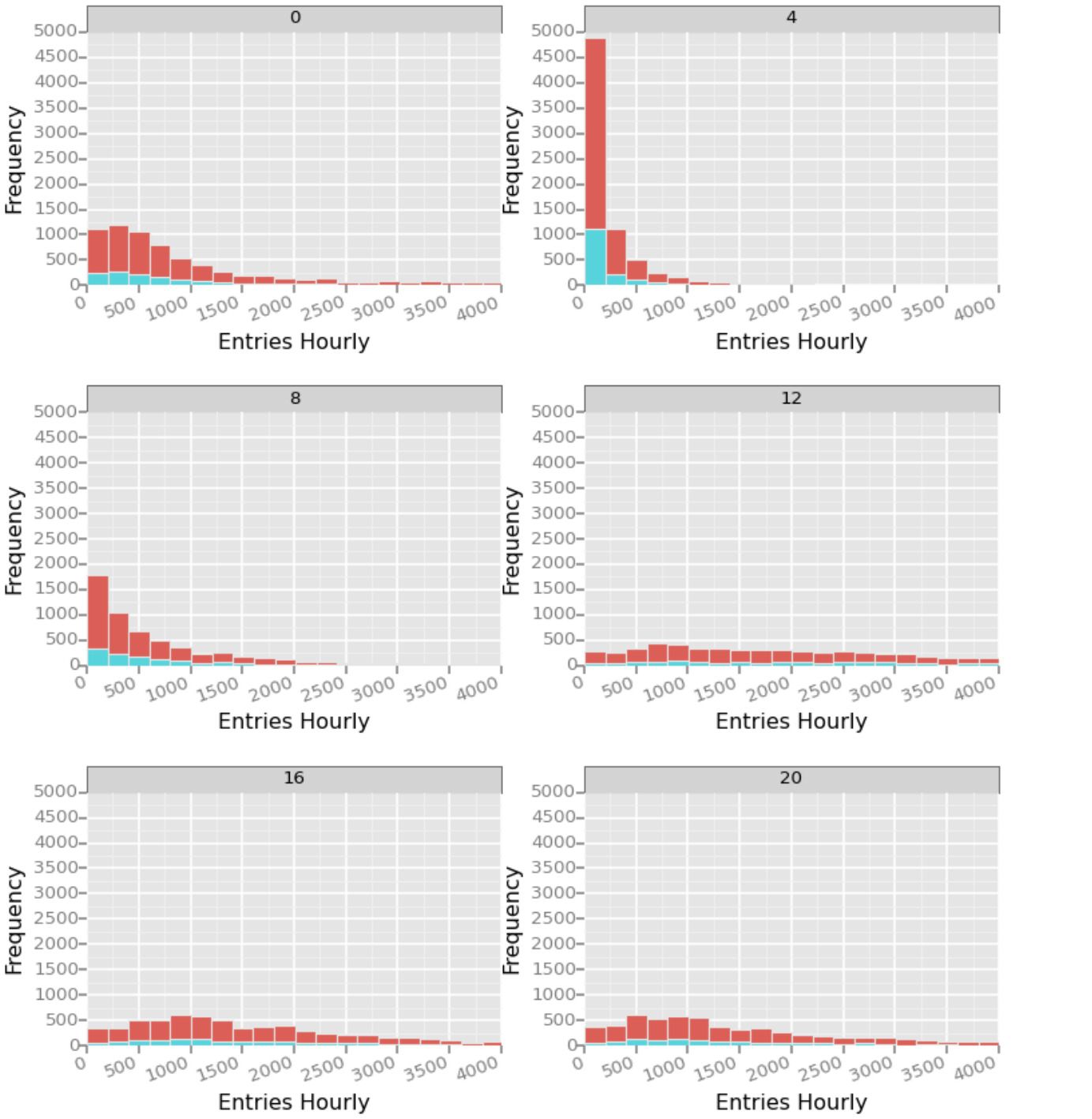


Here is an overall view of the Data. From this visualization we can see that the data is not normally distributed so we cannot run Welch's t test. The chart also shows there are a lot more rows of no rain days than rain days.

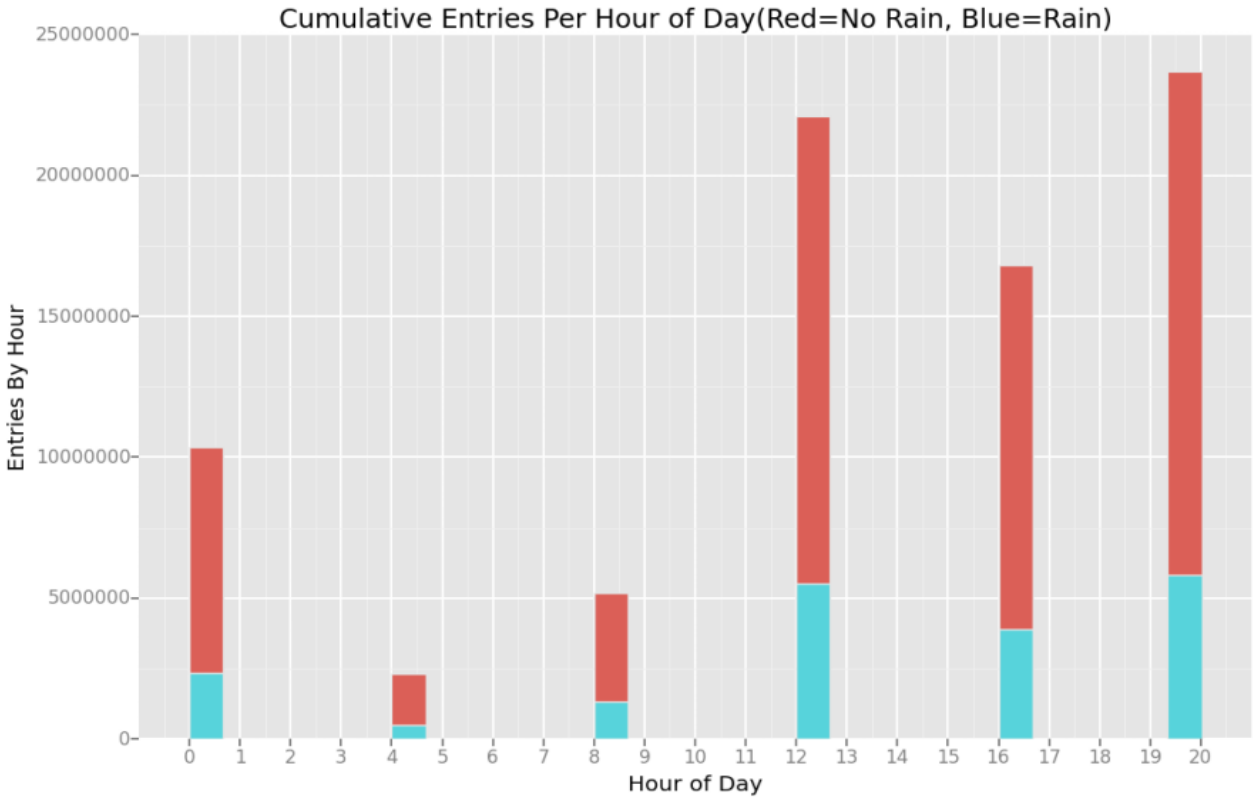
3.2 One visualization can be more freeform. Some suggestions are:

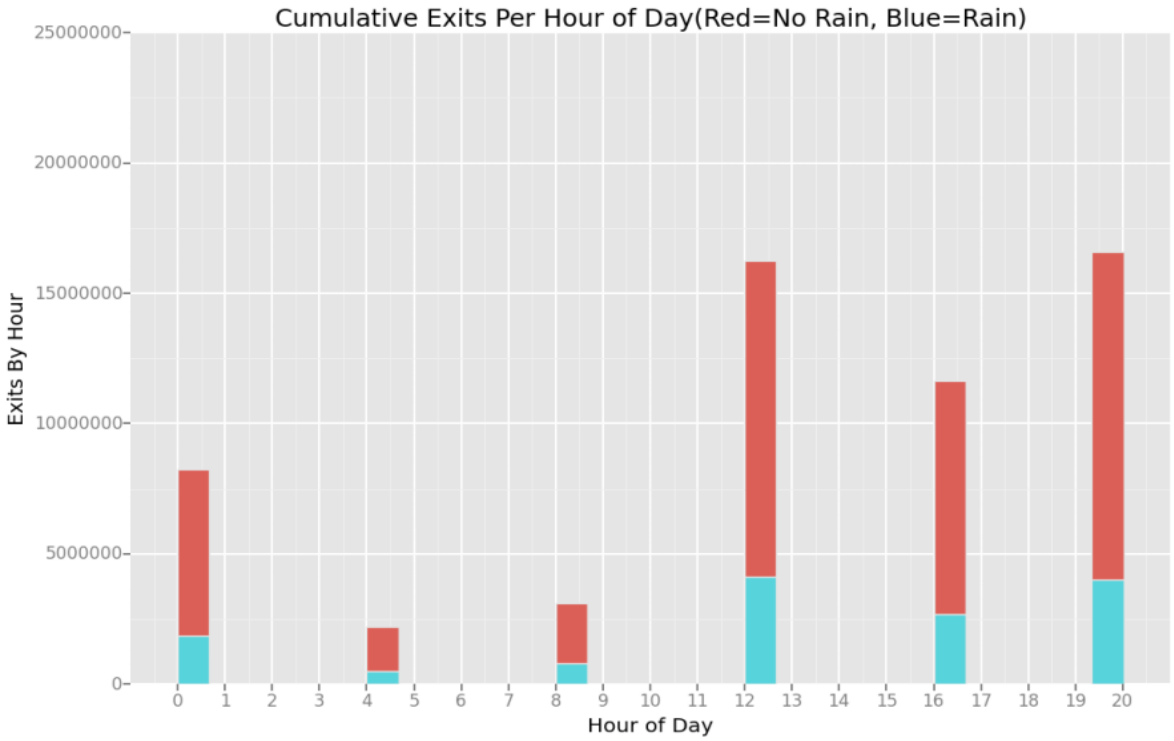
- **Ridership by time-of-day or day-of-week**
- **Which stations have more exits or entries at different times of day**

Here is another view more detailed view of how many people are entering on rainy and non rainy days for each of the sampled hours. We can observe at 4am there is a high frequency of data rows with small amounts of entries, most on non rainy days.



Here are some other interesting visualizations of data which shows the total volumes per hour on Entries and Exits. Here we can observe the highest cumulative ridership at the 20th hour, mostly with no rain. It also shows Entries and Exit volumes are distributed in similar hour patterns.





Section 4. Conclusion

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

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 my analysis and interpretation of the subway turnstiles data it is leads me to think more more people ride the NYC subway when it is 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.

First the Mann-Whitney U two tail test concluded that there was a statistical difference between the two populations and a difference between ridership. Next, the coefficient in the regression model on rain turned out to be 183.8770 which is a positive number so it turns out that would give an effect on ridership. Finally, the No Rain Entries Mean was 1845.53943866 and Rain Entries Mean was 2028.19603547 which lead me to believe ridership was higher when it rained.

Section 5. Reflection

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

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

1. **Dataset,**
2. **Linear regression model,**
3. **Statistical test.**

In my analysis of the dataset I found that the precipi with it's relationship to time and station variables might have made more sense to analyze than the rain variable because it showed the exact precipitation inches, while the rain variable only indicated 1 if it rained at any point in the day. The data rows accounted for 5 different time intervals throughout the day. Ridership can vary based on if it is raining at the time in the day, not just that day overall. Some other variables that could have effect ridership could be events such as basketball, baseball games, and/or power outages due to rain.

As for the statistical test MannWhitneyU two tailed, running the scipy stats one line code for the mann-whitney u test (U,p=scipy.stats.mannwhitneyu(rain,norain)) produced a nan p value. It seemed however if I ran the same one line code on the Original Data set file, it did calculate correctly. The workaround was to calculate the new p value using the more lengthy code which is in my ipython notebook. Here is the link to the discussion board where other users ran into the same issue.

References

http://piazza.com

http://blog.yhathq.com/posts/ggplot-for-python.html

https://github.com/upjohnc/udacity

https://github.com/remondo/Udacity-DS101-IntroToDataScience

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

http://docs.scipy.org/doc/numpy/reference/generated/numpy.mean.html

http://docs.scipy.org/doc/numpy/reference/generated/numpy.sum.html

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

http://en.wikipedia.org/wiki/Ordinary_least_squares

http://en.wikipedia.org/w/index.php?title=Linear_least_squares_(mathematics)

http://en.wikipedia.org/wiki/Polynomial_regression

http://blog.minitab.com/blog/adventures-in-statistics/how-high-should-r-squared-be-in-regression-analysis

http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.describe.html

http://stackoverflow.com/questions/23964236/python-ggplot-rotate-axis-labels