Analyzing the NYC Subway Dataset

Contents

[**Section 0. References** 3](#_Toc417203581)

[**Section 1. Statistical Test** 4](#_Toc417203582)

[**Section 2. Linear Regression** 6](#_Toc417203583)

[**Section 3. Visualization** 8](#_Toc417203584)

[**Section 4. Conclusion** 10](#_Toc417203585)

[**Section 5. Reflection** 11](#_Toc417203586)

# **Section 0. References**

As per my knowledge to strengthen my statistics and python knowledge, I used the following references:

* <http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test>
* <http://en.wikipedia.org/wiki/Student%27s_t-test>
* <http://stackoverflow.com/questions/14941366/pandas-sort-by-group-aggregate-and-column>
* <https://docs.python.org/2/library/time.html>
* <http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html#scipy.stats.mannwhitneyu>
* <https://www.udacity.com/course/ud827>
* <https://www.udacity.com/course/ud201>
* <http://www.codecademy.com/en/tracks/python>
* <http://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.hist.html>
* <http://strftime.org/>

# **Section 1. Statistical Test**

* 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 Mann–Whitney U test to analyze the NYC subway data since it is an appropriate choice for nonparametric non-normally distributed datasets.

I used two-tailed P value we want to assess if there is any difference in distribution of number of entries between rainy and non-rainy days.

Null hypothesis was distribution of number of entries between rainy and non-rainy day is same i.e. the distribution of number of entries between rainy and non-rainy day is not statistically different.

I chose alpha = 0.05 as my P-critical value. So, for p value less than 0.05 we would reject null hypothesis and that would mean that distribution of number of entries between rainy and non-rainy day is statistically different.

* 1. 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.

Since Mann-Whitney U test is non-parametric test which means that it doesn’t assume our data is drawn from any particular underlying probability distribution. So, it became applicable to the dataset because our data is non-normal and using student’s t test or Welch t test would have meant that our data is normal.

* 1. 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.
* Mean entries with rain: 1105.45
* Mean entries without rain: 1090.28
* U-Statistics: 1924409167.0
* p-value: 0.049999825586979442 (two-sided p value)
  1. What is the significance and interpretation of these results?

If we simply compare the means of entries with rain and without rain then it is evident that there is an increase of 1.4% in the subway entries when it rains. However, this statistic alone is not sufficient enough to draw a conclusion. That’s where results from Mann-Whitney U test helps. U statistics for our sample is 1924409167 which is very close to half of the product of number of samples (i.e. 87847 \* 44104/2 = 1937202044) and p value is less than alpha =0.05. So we can say with 95% confidence that distribution of number of subway riders is statistically different between rainy 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?

First, in exercise 3.5 I used Gradient descent to implement linear regression and computed the theta coefficients and produced predictions. Later on, in exercise 3.8 I used OLS using Statsmodels to implement linear regression.

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

In both Gradient descent and OLS using Statsmodel implementation I used the following features:

* Rain
* Precipitation (precipi)
* Fog
* Hour of the day (hour)
* Temperature mean (meantempi)
* Dummy variables for subway stations (UNIT)

The original UNIT feature is categorical. So, we couldn’t use it in our linear regression models as it is. To convert it into dummy/indicator variables I used pandas.get\_dummies.

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.

* Your reasons might be based on intuition. For example, response for fog might be: “I decided to use fog because I thought that when it is very foggy outside people might decide to use the subway more often.”
* Your reasons might also be based on data exploration and experimentation, for example: “I used feature X because as soon as I included it in my model, it drastically improved my R2 value.”

I included rain in my features list because the results from Mann-Whitney U test showed that there was change in number of riders whether it rained that day or not.

I included hour because it made sense that number of riders would vary based on the time of the day

I included precipi because if it is snowing then people would prefer to use the subway than driving in the city themselves. On similar notion, I included meantempi

I included fog because once I included it, it drastically improved my R^2 value.

UNIT was one of the important feature to be considered because there would be few prime subway stations which would be used more often than a few low profile subway stations. However, I couldn’t include the original UNIT feature since it was categorical. So using pandas.get\_dummies I converted it in a dummy variable.

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

From Gradient Descent implementation with alpha = 0.1 and num\_iterations =100 these were the coefficients of the non-dummy features.

Rain: -1.42128836e+01

Precipi: -9.18209020e+00

Fog: 6.53282087e+01

Hour: 4.68337403e+02

Meantempi: -7.43477729e+01

From OLS using Statsmodels, I found the following coefficients for the non-dummy features:

rain -24.525274

precipi -40.228339

fog 232.484066

meantempi -13.472989

Hour 62.241386

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

For Gradient Descent implementation R^2 was 0.464492894286

For OLS using Statsmodels implementation R^2 was 0.484371123678

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 R2value?

This R^2 value means that for Gradient Descent implementation 46.45% (or for OLS using Statsmodels implementation 48.44%) of the proportion of total variance in number of subway riders can be explained by the model.

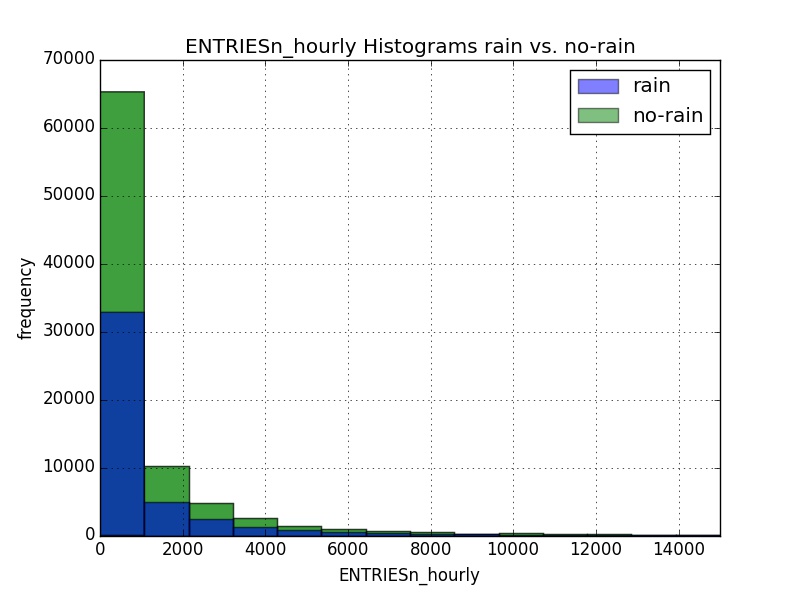
Well, I don’t think this linear model is appropriate to predict ridership since we are able to explain less than 50% variance in the number of subway riders. Another factor here would be to have a look at the residuals. We plotted residuals in Problem Set 3.6 where it was pretty clear that residual for Gradient Descent model were in (-8000, 8000) range so this linear model will not be a good choice if we are looking for very accurate numbers. However, for a ballpark numbers this model might just work.

To improve the accuracy of our model, we either might include more features or combination of features or we might want to explore polynomial regressions.

# **Section 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.

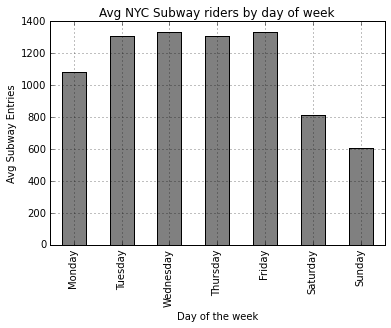
I created this histogram on my local machine. Here is the code image for it. I have attached the code in a separate file.

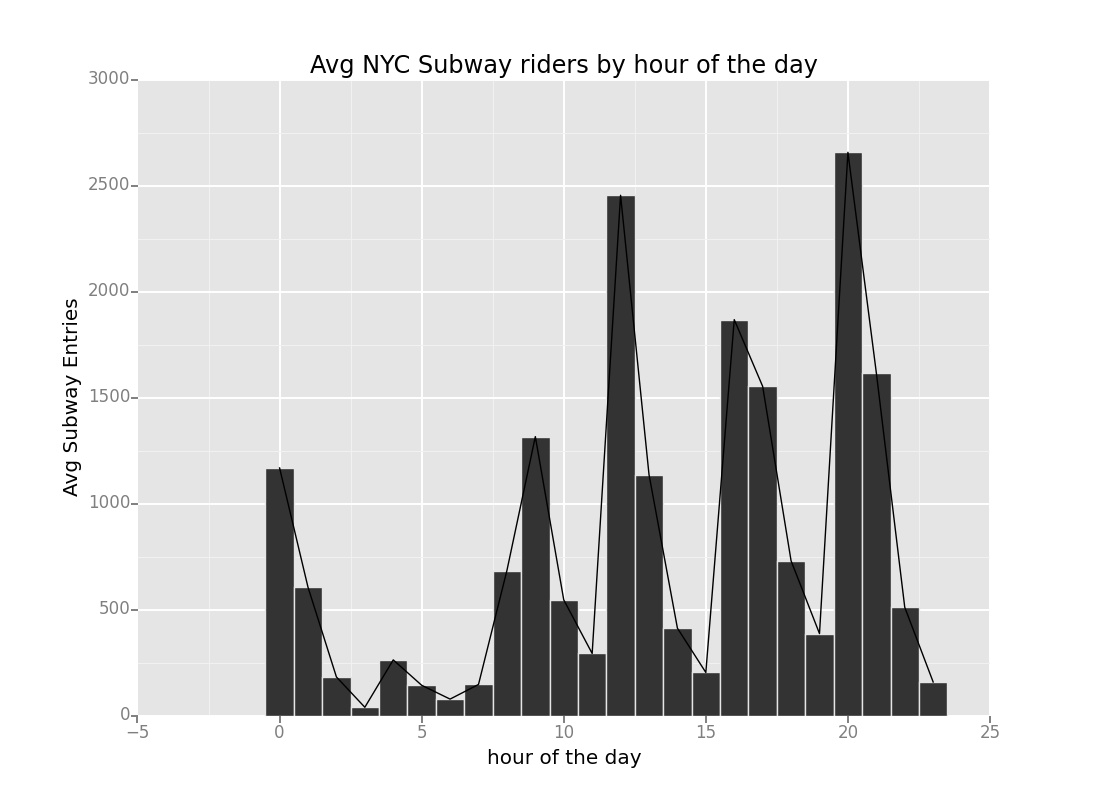


3.2 One visualization can be more freeform. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots) or attempt to implement something more advanced if you'd like. Some suggestions are:

* Ridership by time-of-day
* Ridership by day-of-week

I created this plot on my machine for average number of riders by day-of-week and average number of riders by time-of-day. I have attached the code in a separate file.





# **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?

Even before analyzing the data we would think that it makes more sense to prefer the use of NYC subway more when it is raining vs it is not raining. And, the results from Mann-Whitney test confirm our assumption. Using the p-value from Mann-Whitney test was 0.049999825586979442 we can say with 95% confidence that 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.

We want to assess the effect of rain on the ridership, then we should consider only rain and UNITs. So, if we create a linear model with just rain and dummy variables for UNITs then coefficient for rain comes out to be as follows (I have included the code which I used in Problem Set 3.8 for this problem)

Rain 58.487178

This indicates if keeping all UNITs constant rain will have a positive effect on ridership i.e. rainy days contribute to increased number of riders. Though the fact that average number of riders for rainy days is slightly higher than the average number of riders for non-rainy days is not alt sufficient to draw any conclusions however the results from Mann-Whitney confirm that there is a statistically significant change in number of riders for rainy vs. non-rainy days with 95% confidence level and average number of subway riders for rainy days is higher vs for non-rainy days.

# **Section 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.

First limitation I noticed that the provided subway data just for May. This is a huge limitation because May will have similar type of weather for the entire month. So, this would mean the data provided may be biased. To overcome this limitation we should include data from at least one whole year.

Second thing I would like to add here is that using dummy variables we included all the subway stations in our linear model. Instead of this, we should have focused on one subway station at a time and then create a different model for each subway station. Because, different subway stations have different geographic locations, connectivity to other subway stations and other factors which affect the number of riders. Further, we should also focus on hour of the day since hour of the day also affects the number of riders. So, by selecting one subway station and one hour of the day we will truly be focusing on effects of weather on number of subway riders.

Third thing would be so far we considered implementing linear regression models. But, we should consider implementing polynomial regression models as well because it is not necessary that linear models will fit for all kinds of data sets.

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

Well, I believe that number of subway riders may also depend on holidays, social events, sports matches etc. So, we might see a sudden drop or hike in the number of subway riders which could construe as outliers for the data set and impact the overall model.

Another interesting factor would be to include number of lines/number of other subway stations that each subway station connects to and compensate for such and rest of the features and then which subway stations are used more frequently. Based on this, we can find out which geographical factors such as restaurants, bars, sports stadiums, offices, and colonies make those subway stations more popular than others.