# Analyzing the NYC Subway Dataset

*Submitted by Mital Shah for Udacity's Data Analyst Nanodegree, Project 1*

## Section 1. Statistical Test

**1.1 Which statistical test did you use to analyse 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?**

*Answer: The Mann Whitney U test [wikiMann]_ is chosen to assess the statistical significance of this result.*

*I have used a two tail p-value because of the null hypothesis.*

*The hypothesis that there is no significant difference between specified populations, any observed difference being due to sampling or experimental error*.

*In using the Mann-Whitney U test, the null hypothesis is that the two populations are the same, or simply put, that rain has no correlation with ridership.*

*We have used a p-critical equal to 0.05, meaning that in case the null hypothesis is false we will require a 95% of confidence.*

*The p-critical value used was 0.05, or 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.**

*Answer:*

*The Mann Whitney U test, or Wilcoxon rank-sum test, is chosen because of characteristics of our samples: we can't use a parametric test because the distributions do not seem to follow any particular and well known probability distribution which we could use to make inferences that could directly report the significance of any difference between both populations.*

*The U test is particularly powerful to assess the significance of the difference between the median of two samples that have similar distributions. The assumptions that our data samples must comply with are basically:*

- *All observations of both groups are independent*
- *The responses are ordinal (so we can use the ranking algorithm of the U test).*

**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.**

*Answer: Mean entries with rain: 1105.446*
*Mean entries without rain: 1090.279*
*U-statistic: 1924409167.0*
*p-value: 0.025*

**1.4 What is the significance and interpretation of these results?**

*Answer: Comparing the means yields 1.4% more subway entries when it rains. This statistic alone is insufficient in drawing conclusions or correlation. The U-statistic has a high value, very close to the maximum value of 1937202044.0, or half the product of the number of values in each data set. A U-statistic of half the maximum would indicate that the null hypothesis is true. Of note, the p-value 0.025 satisfies the p-critical value, and the conclusion can be drawn with 95% confidence that the null hypothesis is false and that ridership is different with vs. without rain.*

## 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?**

*Answer:*

*A machine learning algorithm, batch gradient descent, was used to train the linear regression coefficients. I used the default values of learning rate (alpha) 0.1 and 75 iterations, and also kept the mean normalization feature scaling. The given values were sufficient in converging on a local minimum, as confirmed by plotting the cost history vs. number of iterations.*

*Notes*

*[1] The linear correlation coefficient (r) can take on the following values: −1≤r≤1. If r=+1, then a perfect positive linear relation exists between the explanatory and response variables. If r=−1, then a perfect negative linear relation exists between the explanatory and response variables.*

*[2] The coefficient of determination (R2) can take on the following values: 0≤R2≤1. If R2=0, the least-squares regression line has no explanatory value; if R2=1, the least-squares regression line explains 100% of the variation in the response variable.*

Gradient descent

*The code used to implement the gradient descent algorithm to find the linear regression coefficient can be checked on the python file available at the GitHub repository associated to this work (projectone.py), and on the submissions to the* Introduction to Data Science *class (problem set 3).*

Ordinary Least Squares (with stats models)

*Selecting the same features as in the Gradient Descent exercise, we calculated the coefficients of the linear model by using the OLS implementation of the stats models python library [stats models] .*

Polynomial features with Ridge linear regression

*After analyzing the results from the previous regressions, and for reasons that will become clear after the description of their results, we went a little further and we used a polynomial transformation of the selected features, and another linear regression algorithm, the Ridge regression, was used to find the coefficients and predict ridership. The model used and results will be shown in the interpretation section.*

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

*Answer: Features used included rain (0 or 1), precipitation, mean wind speed, hour, and mean temperature. Per the default configuration, dummy variables were introduced for features 'UNIT' (the turnstile location/identification number), which were categorical in nature. They were initialized with Boolean (0 or 1) features with prefix 'unit,' and each data point would have a '1' in the feature that it "belonged" to. It*

*did not make sense to apply linear regression to the raw 'UNIT' parameters quantitatively; however, it was important to keep track of it as there was a wide variation between different subway stops and account for it first. If this was not done, the differences between different turnstiles would mask the markedly smaller changes due to rain, precipitation, hour, or temperature.*

*Quantitative features used: 'hour','day_week','rain','tempi'.*

*Categorical features used: 'UNIT'. As a categorical feature, this variable required the use of so-called dummy variables.*

**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.**

*Answer:*

*The features were selected based partially on intuition and partially by exploratory analysis. First, it was clear that the behavior for each individual turnstile was mainly a function of the hour of the day and the day of the week, depending on the time of the day, and also a dependence on the day of the week. However the relation is clearly non-linear. We kept the hour as a predictor because is an important predictor, an in a very rough approximation one can see that ridership is lower in the beginning of the day while reaching a peak on the evenings.*

*Ridership vs date for turnstile R084.*

*The figure clearly shows a periodic behavior for the ridership behavior for a particular turnstile, which is a function mainly of the hour of the day and day of the week. Ridership peaks are usually seen at 20 hours, while weekends and holidays (May 30th) being less busy than weekdays.*
*However, we decided to use weekday instead of day_week (the second being the day of the week, i.e, a number between 0 and 6, where 0 is Monday and 6 Sunday), because the major change on ridership behavior is seen between work days and off days (weekends), and weekday can be better modeled by a linear model than day week as it can be checked on Ridership vs day of the week.*

*This plot show the ridership distribution as boxplots for the 7 days of the week (0 is Monday, 6 is Sunday). We can see that even when a relation exist between day of the week and ridership, this relation doesn't look linear, and thus we decided to use weekday instead.*
*Even when UNIT was not a numerical variable, we decided to use it given the different ridership patterns for each turnstile location. When using it as a dummy variable what we will be doing is adding or subtracting a constant offset which is particular for each location. Will this be enough to model the behaviors of different stations?*
*No further experiments where done to try other weather variables, since we were mainly interested in the behavior of the system as a function of precipitations; also, no other linear relationships were apparent from these variables, or there was not enough data to sample the ridership under some conditions (e.g, only 1 or 2 foggy days, no snow, etc.)*

*Finally, out from intuition we left out the variable EXITSn: besides having a highly linearly correlated relation with the ridership variable, it is clear that this variable is not completely independent from the number of entries to subway. Furthermore, it won't be a nice predicting feature, since its value will depend on the number of entries, and it should be treated as a observable or predictable variable on itself.*

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

*Answer: [-4.24736210e+00, 8.60518981e+00, 4.64832083e+01, 4.64163466e+02, -3.19114921e+01, 1.08898857e+02]*

```
print best_results.params
```

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

*Answer: r^2 value is 0.463968815042*

*For n=500, the best R2 value witnessed was 0.85 (with the best r value seen at 0.92).*

*The coefficients found with the gradient descent and OLS algorithms were the same in both cases, which was expected for a successful execution of the gradient descent algorithm. The selected features were enough to obtain a R^2 = 0.481 .*

**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?**

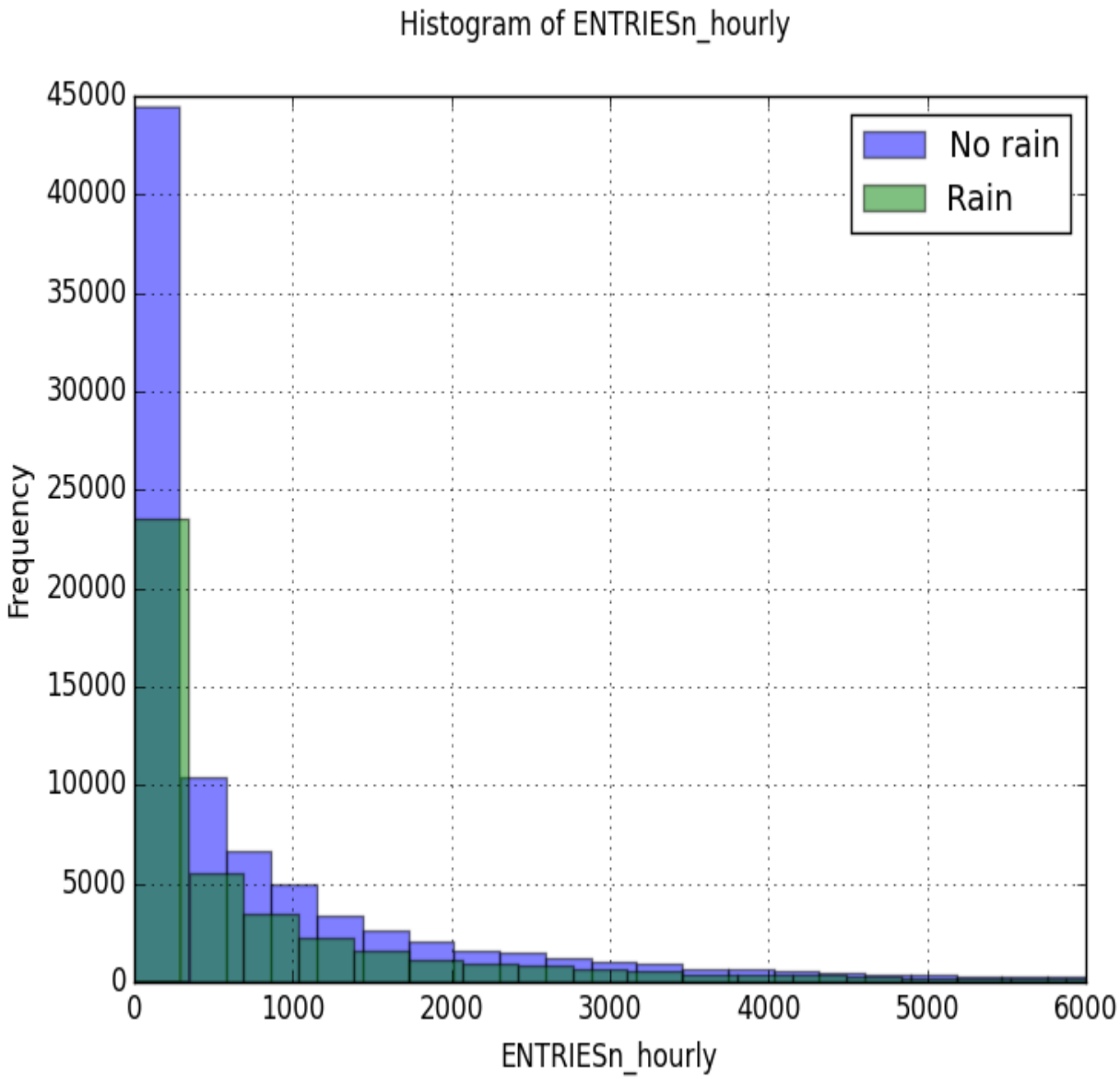
*Answer: R^2 is essentially the percentage of variance that is explained, and is a quantitative measure of the "goodness of fit." While it only explains 45.8% of variation, I think a better metric is to plot the residuals.*

*To decisively conclude whether or not this model was a good fit certainly depends on the context and use case for the prediction data. If this were a use case that had safety and security concerns, it would certainly be insufficient! The residual plot shows that most of the residuals were close to 0 +/- 5,000. Qualitatively, and for the objective of being able to "ballpark" ridership, the linear model is sufficient.*
*Further and advanced study could include more features or utilize polynomial regressions. However, this might lead to significant over-fitting, and the model may fail on new data sets. In that case, regularization would be a good method to attenuate any over-fitting.*

## Section 3. Visualization

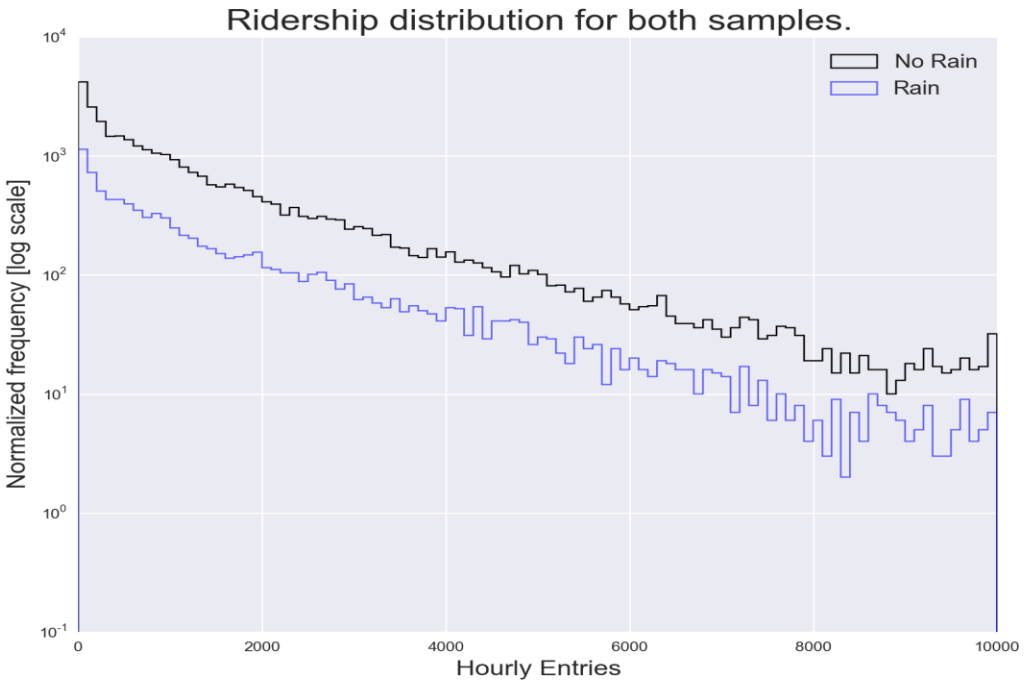
**3.1 Include and describe a visualization containing two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.**

***Answer:***




***Plotting overlaid histograms of subway entries for both rainy and dry hours shows that both distributions are not normally-distributed. Of note, it's important to clarify that these are aggregate values and that there were less rainy days than there were not rainy days; it would be grossly incorrect to draw from this graph that subway ridership is less when it rains.***

***The figure comparing the ridership distributions for rainy and non-rainy days has been already presented in the Chapter 2. We show the same figure, but this time we want to show the different samples sizes by not normalizing the data.***

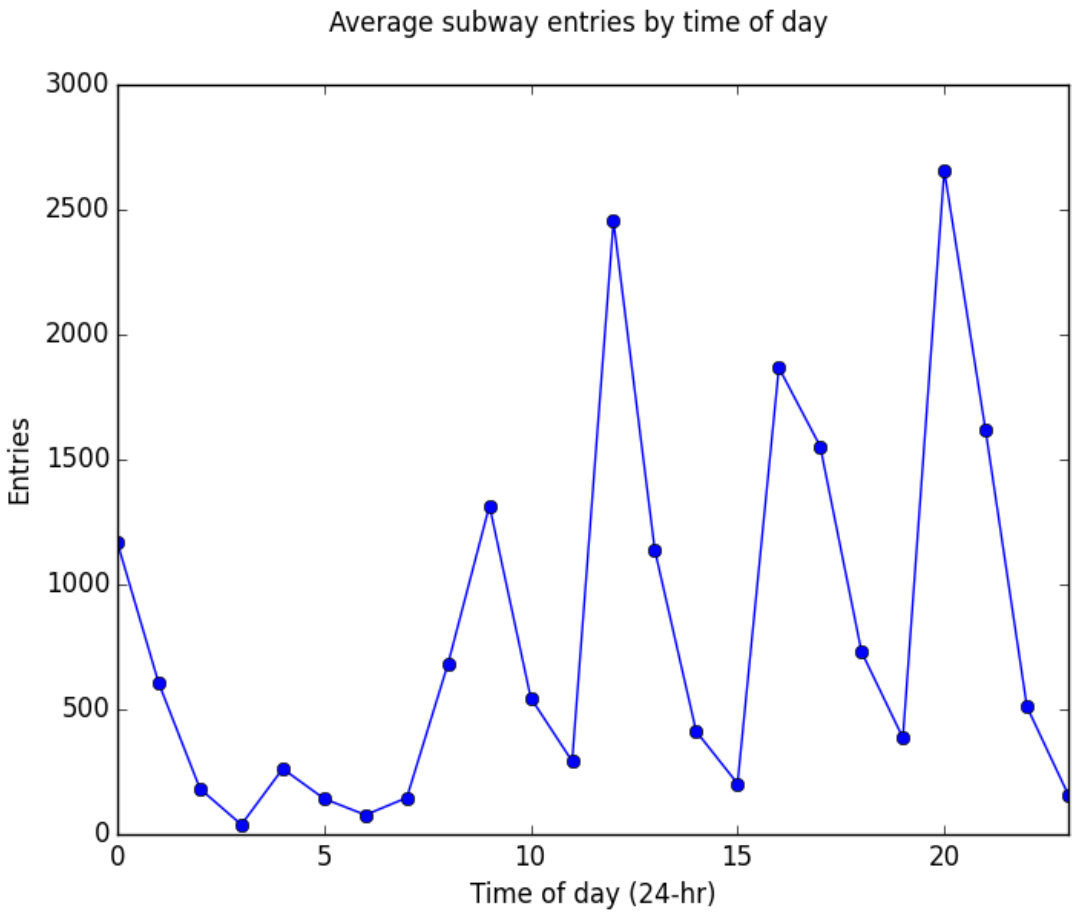



***Ridership distribution comparison between rainy and dry days.***

***Please note the logarithmic scale on axis Y. It was used to allows us to study the visualization with more detail. Both distributions are similar in shape, but the rainy sample is smaller than the rain sample (there was precipitation reported for only 7 days of May 2011), and thus the counts by bin are smaller.***

### 3.2 Include and describe a freeform visualization.

***Answer:***




*By plotting the average number of subway entries at each hour, it's clear that there are several peaks throughout the day, with the most prominent ones being at noon and 8pm. interestingly, these peaks are larger than those during rush hours (8-9am and 5-6pm). It raises some intriguing questions about the demographics and characteristics of NYC subway riders: assuming a 9-5pm workday, why are more subway entries occurring at 5pm vs. 9am? Are more people going out to lunch (12pm) and dinner (8pm), or is that there work schedule? Without any demographic data, it would be impossible to determine these questions from the current data set.*

## 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?**

*Answer: Particularly given the results from the Mann-Whitney U test (p-value: 0.025), we can say with a high level of certainty that more people ride the NYC subway when it*

*is raining. It is important to note that simply looking at the means of both data sets is insufficient, due to variance. The Mann-Whitney U test is needed to quantitatively confirm that the two data sets are statistically different.*

*From the current data set and the analyses performed, it remains inconclusive whether rain has any impact on the number of NYC subway entries. However, based on this data set alone, rain seemed to be an insignificant factor as it related to subway ridership. Thus, further analysis is necessary.*

*On the other hand, based on the data exploration, it seems quite clear that the number of entries is highly dependent on physical location, particularly station position, with specific units having the most importance.*

**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.**

*Answer: Two analyses were performed that backup this conclusion:*

1. *A non-parametric statistical test between two samples, being one sample the data observed in rainy days and the other the data observed in non-rainy days. Even when in the beginning the small difference in the ridership volume reported for each sample seemed to be significant by the statistics reported by the test, it was not clear that both samples were really independent. After aggregating the data from all individual turnstiles to smooth possible selection effects and outliers, the test was run again and no significant nor meaningful difference was found between both samples.*
2. *A second analysis was done by means of the use of a machine learning technique. We tried to fit the data to multiple regression model, were we aimed to find predicting features from with our data. Even when some predicting features were found, as the hour of the day, day of the week, holiday or workday, the rain indicators didn't have either a significant weight or a high p-value, thus confirming by an independent method that precipitations didn't seem to play a roll on the ridership behavior of the NYC subway. or from the fitting to a multiple regression model, support*
3. *We believe that we did not have enough data to answer the question. Even studying the system as whole, removing the complications associated to the different behavior between different turnstiles and locations, the number of rainy points was small: only at 2 times heavy rain was reported, and just for short periods of a day; and only at 8 hours the conditions were "rain". All the other precipitations reported lasted only for*

*short periods of time, affected only specific stations withing the whole NYC area, or where only relate to light rain or drizzle.*

*After obtaining this result, can we discard some of our previous preconceptions or intuitions? One prejudgment was to believe that people would prefer to use the subway on rainy days, instead of using other transportation means as bus or taxis, since the later would mean to be more exposed to the rain. Another preconception, in a opposite direction, was to think that people would prefer instead to remain at home if there is not need to go out, thus only people that need to commute to work would be riding on those conditions.*

*The positive coefficient for the rain (0 or 1) parameter indicates that the presence of rain contributes to increased ridership. This may have not been the case for all data points, with the R^2 being approximately 46%; however, the small residuals show relatively high accuracy, given our objectives. Although the means of both data sets are not that different from each other, the Mann-Whitney U test did indicate that there was a statistically significant change in ridership for rain vs. no-rain. It is conscientious to claim that rain increases subway ridership.*

## Section 5. Reflection

**5.1 Please discuss potential shortcomings of the methods of your analysis, including: data set, linear regression model, and statistical tests.**

*Answer:*

*Several shortcoming have been raised along the pages of this work. Many of them are not only related to the analysis methods but also to the dataset. Here we will summarize the most relevant shortcomings according to our criteria. The order of the list doesn't necessarily reflect the importance.*

1. *The statistical test must be used with caution: besides checking that some basic assumptions about the shape of the samples distribution should be met, the test itself does not indicate any about the actual independence of both samples. One should be careful when comparing the two samples on assessing if any other feature, or selection problem, can be the responsible of a significant difference detected by the test.*

2. *The use of a linear regression, even with multiple features, was shown to be not enough to model the ridership behavior of the NYC subway. The assumption of linearity between the predicting features and the entries by hour was not met for most of the variables, and the residual analysis confirmed the poor fit.*
3. *We believe that we did not have enough data to answer the question. Even studying the system as whole, removing the complications associated to the different behavior between different turnstiles and locations, the number of rainy points was small: only at 2 times heavy rain was reported, and just for short periods of a day; and only at 8 hours the conditions were "rain". All the other precipitations reported lasted only for short periods of time, affected only specific stations withing the whole NYC area, or where only relate to light rain or drizzle.*
4. *One immediate red flag that was presented while exploring the data was that there were markedly more entries than there were exits. The only logical explanations could be that there were miscounts, or some turnstiles/stations were not included in the data set. Presumably, this would have had an equivalent effect on both rain and no-rain data sets, so for the purposes of this study, it likely had little to no effect.*
5. *A combination of increased sample size (larger data set) and normalization by location/turnstile ID could have potentially increased the confidence of both the Mann-Whitney U test and the linear regression model. As we saw from examining the 'UNIT' column, ridership varied greatly. Simply put, some stations and turnstiles were naturally more active than others. The Mann-Whitney U test did not take this into account, and only looked at the subway entry distributions for rain and no-rain. Examining how the same stations at the same day and time varied by rain could have increased the fidelity of the test.*

*The data set under consideration was limited to a single month in the late spring / early summer of a particular year. As a result, among countless other possible factors for which the available data did not account, precipitation may in fact have an increased impact on the number of entries at other times of the year (e.g., during the winter months). Thus, the data set was limited by its temporal locale.*

*The linear regression model that was created, while having very high r and R2 values did not, based on residual analysis, adequately model the data*
*There is in fact not a linear relationship between the explanatory and response variables under consideration; thus, a non-linear model would likely be more appropriate for the current data set.*

*The statistical tests that were employed seemed effective (as long as sample sizes were kept small enough). However, it's unclear how traditional statistical tests relate to massive data sets (esp. since many statistical tests need to be used with relatively small sample sizes; on this point, see 5.2 below).*

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

*Answer: I think that an interesting investigation would be to use gradient descent with logistic regression to see if one might be able to predict if it rained or not given various parameters, to include turnstile location/ID, time of day, and subway entries. Intuitively, this might produce false positives or negatives on special days (e.g. sports game, holidays, etc.).*

*Assuming all statistical tests and learning models were implemented and interpreted correctly, it became clear that computational power was very important in data science, not due to the ability merely to apply methods to data, but in the ability to repeat numerous tests on random samples of data, which, at least in the case of this analysis, encouraged more confidence in test/model results.*

*As the shortcoming, some other insights have been shown within this work. Many of the visualizations presented in the different figures wanted to inform the reader about these findings:*

- *How the behavior changes for holidays, even when only one holiday was present in our data (May 30th, 2011).*
- *The ridership behavior and features changes between locations, specially when comparing the busier downtown stations with the periphery locations.*
- *The precipitations are different within the NYC area for the month of May, with the southern stations reporting a higher precipitation.*

## References


- http://docs.scipy.org/doc/numpy/reference/generated/numpy.dot.html
- http://www.graphpad.com/guides/prism/6/curve-fitting/index.htm?reg_analysischeck_linearreg.htm
- http://pandas.pydata.org/pandas-docs/stable/
- https://bitbucket.org/hrojas/learn-pandas
- https://pypi.python.org/pypi/pandasql
- http://www.sqlite.org/lang.html
- https://docs.python.org/2/library/pprint.html
- Intro to Descriptive statistics course that explains it very well. Link and Programming Foundations with Python course by Udacity
- Documents from Instructor notes: Gradient Descent - Problem of Hiking Down a Mountain, Linear Regression and Understanding the MannWhitney U Test – docs provided by Udacity
- Stanford Machine Learning Class
- SciPy Documentation
- Mann-Whitney U Test Wiki
- GraphPad
- Piazza posts