Analyzing the NYC Subway Dataset

Section 0.
References

1. In Stackoverflow, slideshare and yhathq.com have gone through various implementations of ggplot. Also gone through the various code snippets posted by fellow Udacians on the forum for better understanding of the concepts and to make solid my foundations. I have also gone through various websites such as Data science central, R-bloggers and Khan academy to clearly understand the fundamentals behind the Linear Regressions and their real time applications as well.
2. http://statsmodels.sourceforge.net/devel/generated/statsmodels.regression.linear_model.OLSResults.conf_int.html - for calculating confidence intervals.

3. https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test – Wiki article on the Mann Whitney U test

Section 1.
Statistical Test

3.1 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?

A: null hypothesis: Rain has no effect on ridership
Alternate hypothesis: subway ridership is different in rainy and non-rainy days.
Statistical test: As the data is not normal, we can't use Welch's test and we used Mann-Whitney U Test.
I used a two-tail test.
P-critical value is 0.05

3.2 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.

A: As we are dealing with a large dataset of more than 30 samples and the data is not normal which we could find from our problem sets. And importantly we are trying to draw comparison between the ridership on rainy and non-rainy days which are two independent groups within a given data set.

3.3 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

A: U value or significance = 1924409167.0
p-value = 0.0386
mean_no_rain = 1090.278
mean_rain = 1105.446

3.4 What is the significance and interpretation of these results?

A: As the p-value is very low we can reject the null hypothesis and can say there is difference in the subway ridership when it rains.

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. *OLS using Statsmodels or Scikit Learn*
2. *Gradient descent using Scikit Learn*
3. *Or something different?*

A: 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?

A: features = dataframe[['rain','fog', 'meanwindspdi','Hour', 'meantempi']]
Dummy variable: UNIT

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.

A: I have used the above mentioned features because when taken all the variables as features and removing the features one by one, I could observe improvement in R^2. Hour is selected as feature because there might be sometime in a day that is very busy and sometime empty. Fog is used because on foggy days people may prefer staying indoors. Rain is selected to see how important it would become in predicting riders, to address the original question.

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

A: rain, fog, meanwindspdi, Hour and meantempi and the weights are -3.30628397e+01, 3.96213350e+01, 4.12074768e+01, 4.57044771e+02, -3.52074425e+01 respectively.

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

A: 0.4581

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?

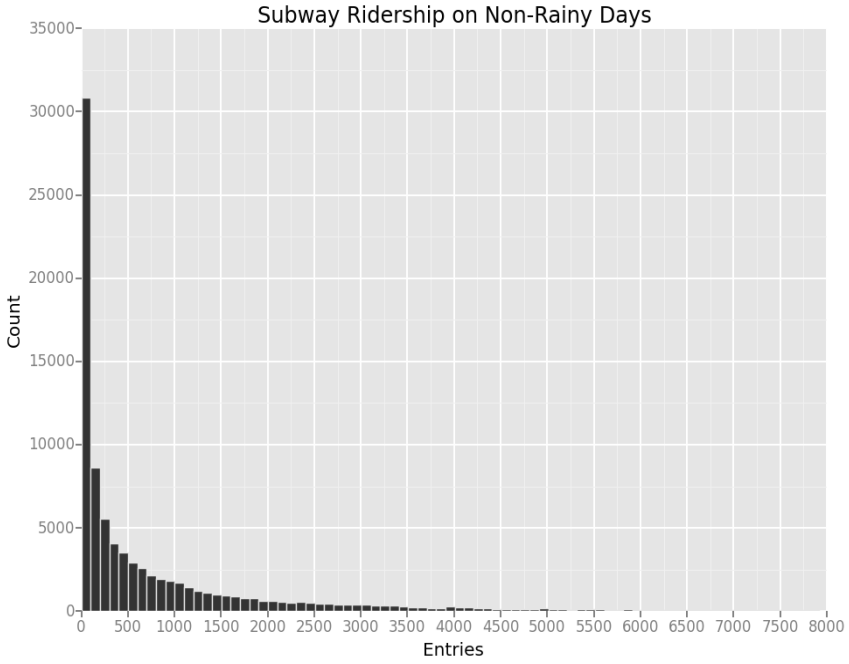
A: R^2 value is 0.4581 and it means that my model can explain 45% of variance in the subway ridership, using features fog, wind, Hour of the day and temperature and UNIT ID which is the location. I think for this dataset linear model is not that appropriate because higher the R^2 value, better the model is and 45% variance is not great when compared to 60% or 90%. We could add all the dummy variables available to get a good R^2 value, but it will increase the complexity of the model. Though it doesn't have the best fit, it is a kind of balance between predictability and complexity at least in my view.

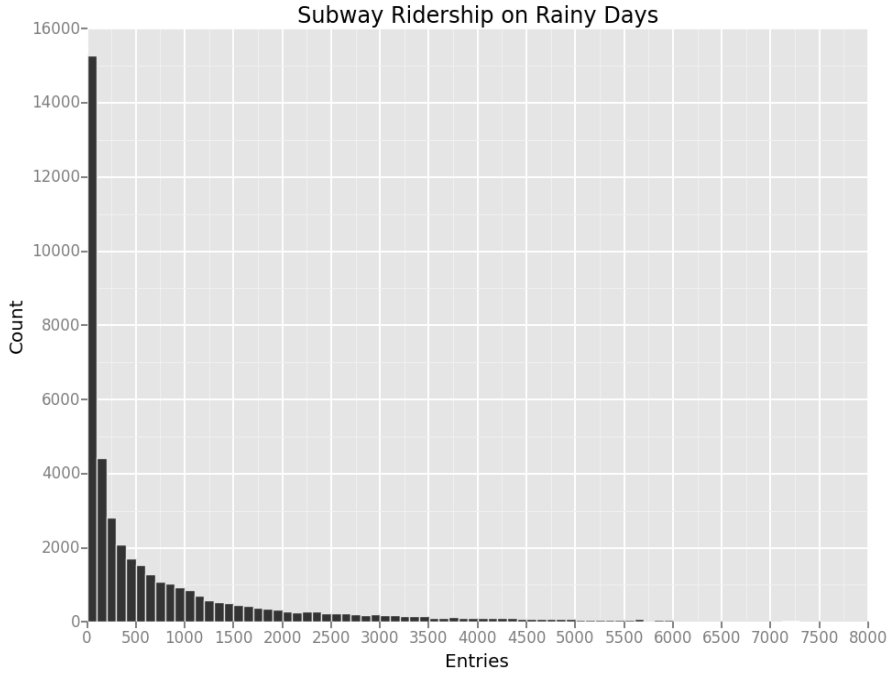
Section 3.
Visualization

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

Description: We could see from the histogram that the ridership is different from rainy to non-rainy days on the basis of data points and both have similar distributions with right-skewed.

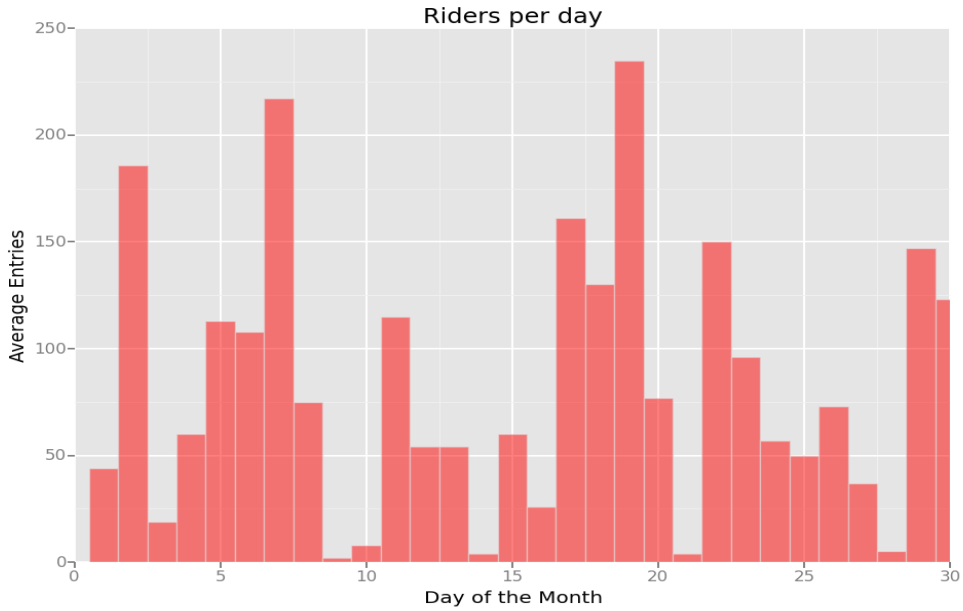
Also it can be observed that the tail on Rainy days histogram is longer than that of non-rainy days.





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.

Description: From the below chart we can interpret that the ridership is more in the 3rd week of the month followed by 2nd week.



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?

A: Yes, from the above statistical test results and the linear regression model we can conclude that the ridership for the subway is different when it rains from non-rain days. And more people are riding on rainy days.

As the p value lies in the p-critical region we can reject the null hypothesis and safely assume the alternate hypothesis that ridership is different from rainy days to non-rainy days. From the linear regression model, the co-efficient of rain -3.30628397e+01is small and negative, which tell us that there is not much effect of rain in subway ridership.

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.

A: From statistical analysis we got the means as mean_no_rain = 1090.278, mean_rain = 1105.446 and U value or significance = 1924409167.0 and p-value = 0.0386. As the mean of rainy days is more than the

non-rainy days and the test support that the difference is significant as the p-value lies in the significant range.
From co-efficient of rain -3.31 is not great when compared to the volume of the footfall, it clearly shows that there are other important features that effect ridership other than rain. However, the co-efficient of rain is not 0 and that indicates that it has surely effect on ridership.
Using statsmodels, I got $-0.097 \leq \theta_{rain} \leq -0.039$. However, the fact that the coefficient is, with 95% confidence, not equal to 0, does indicate that it has an effect. I suppose my original misinterpretation was in the direction of the effect, specifically that the sign of regression coefficients indicate the direction of a relationship. I will still conclude that it has some effect based on this confidence interval. As the reviewer said, for every 1 unit increase in rain, there is the decrease of ridership in terms of the confidence interval mentioned above.

Section 5. Reflection

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

1. Dataset,
2. Analysis, such as the linear regression model or statistical test.

Though for test purpose the data is enough, it would be great if we have a bigger data set. Also it would have been great if we could have discussed and implemented different kind of regression models rather than just confining to linear regression model alone. Ignoring the outliers in the data also has some effect on predictions. At most it all depends on the human behavior of whether using subway or not and it is hard to predict the human behavior perfectly using regression model. Also we have used only a few points and this may lead to wrong results. Adding some variables such as some shopping facilities/eating joints/public toilets also effect the subway ridership. It can also be that it rains on weekends alone, where in which the ridership will be low. We got different results from liner model and statistical tests and this shows that there are more variables affecting the ridership other than attributes given in the dataset. The dataset given is limited i.e., the data here is only for a period of a month and does not account for seasonal changes. Also a good correlation doesn't necessarily prove a good cause and effect.