Analyzing the NYC Subway Dataset

Overview

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.

Section 0. References

Please include a list of references you have used for this project. Please be specific - for example, instead of including a general website such as stackoverflow.com, try to include a specific topic from Stackoverflow that you have found useful.

```
http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.shapiro.html#scipy.stats.shapiro
http://www.statisticshowto.com/what-is-an-alternate-hypothesis/
http://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test
http://en.wikipedia.org/wiki/Coefficient_of_determination
http://stackoverflow.com/questions/9847213/which-day-of-week-given-a-date-python
http://scikit-learn.org/0.14/modules/generated/sklearn.linear_model.SGDRegressor.html
http://en.wikipedia.org/wiki/Welch%27s_t_test
http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html
http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test
http://stackoverflow.com/questions/12190874/pandas-sampling-a-dataframe
http://pandas.pydata.org/pandas-docs/version/0.15.2/cookbook.html#cookbook-plotting
```

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 used Mann-Whitney U test, mainly due to the fact the distribution is not normal. A one sided p value tail test is used normally, when we want to capture which data set is higher or lower with respect to the other. Since, no details are given I used a two tailed test.
- Since, I am using Mann-Whitney U test - the null hypothesise is that the two populations are same, i.e to say that rain has no effect on the ridership.
- My p-critical value is 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.

- Mann-Whitney U test is applicable due to the fact that it does not assume the population to be a part of any distribution and is suitable for non-normal distributions.
- On performing an initial exploratory data analysis on the data set, we found that the distribution of rain vs no rain on ridership was not normally distributed.

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.

- statistics: 1924409167.0
- probability: 0.0249999127935
- means with rain:1105.446377
- meanswithout rain: 1090.278780

1.4 What is the significance and interpretation of these results?

- Simply comparing the means from the statistical test tells that the number of entries is almost 1% more when it rains. However, this cannot be used to draw any kind of conclusions as of yet. As we observe that the p value is below the p critical - we can state that the null hypothesis is false with 95% confidence. This states that the ridership is different with rains and different without 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:

- I used Gradient Descent. The learning rate for the algorithm was kept to the deafult which is equal to 0.1. I also used normalization of features to compute the gradient descent.

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

- features used:
  - weekday
  - rain
  - Peak Hour
- dummy variables used:
  - UNIT: The turnstile location/identification number), which were categorical in nature

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.???
- Rain can effect the number of people getting in to the subway. Deuring rains, people might find it difficult to get into subway stations and hence I used this feature.
- Weekdays should be more crowded than Weekends when people usually take rest. A common logic is weekends people prefer to stay at home
- Peak Hours can effect the people getting in due to work oriented reasons.

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

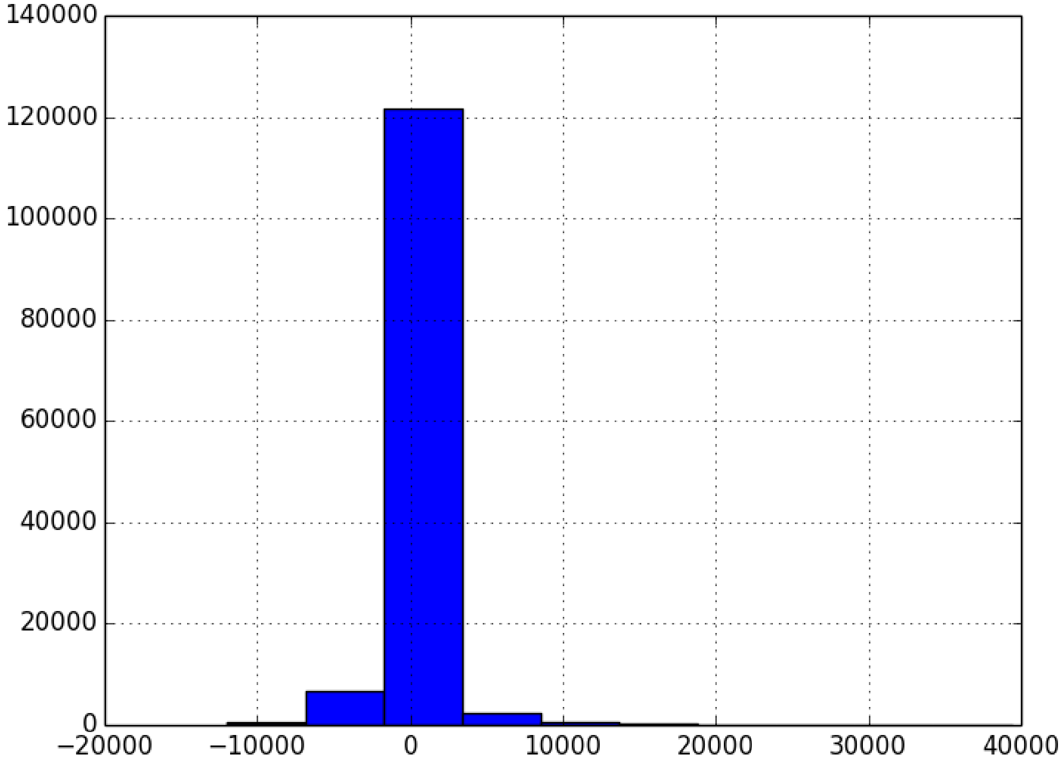
```
Rain:+5.40768773e+01
Hour:2.94382581e+02
Weekday:2.38104670e+02
Peak:3.02560506e+02
```

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

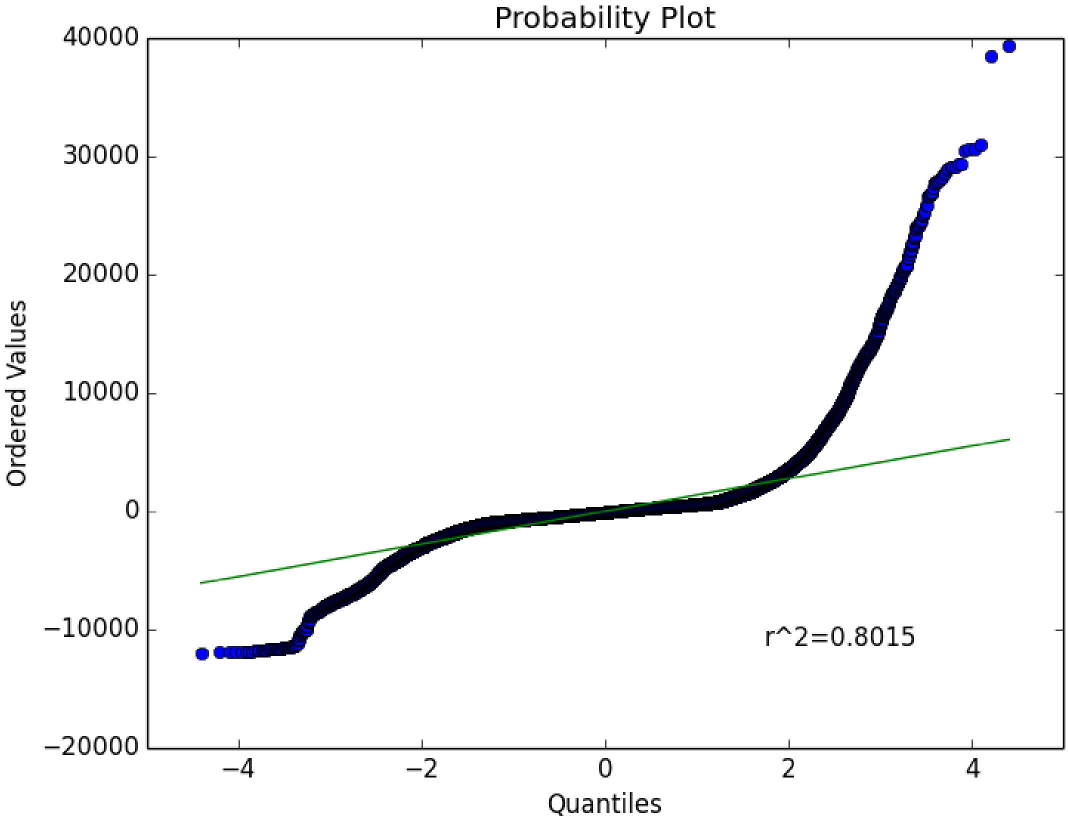
```
0.409970830518
```

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?

- Rˆ2 is essentially the squares of residuals and it signifies the variation in the data set. In our case, it explains 41% of variations and it used for goodness of fit test. a high Rˆ2 value close to 1 explains that we have explained all the variabilty in the dataset. In our case, we have explained almost 41% of variability.



- Figure above shows the residual plot showing that most of them have been capture but with a long tail, we can say that probably the linear model falls short in terms of the level of detail the results need to adhere to. For, a strict restriction i.e to say that +-1000 is not allowed, we could say that the linear mode is insufficient. But, since we are looking to jus ballpark, we could say that the linear mode is sufficient.



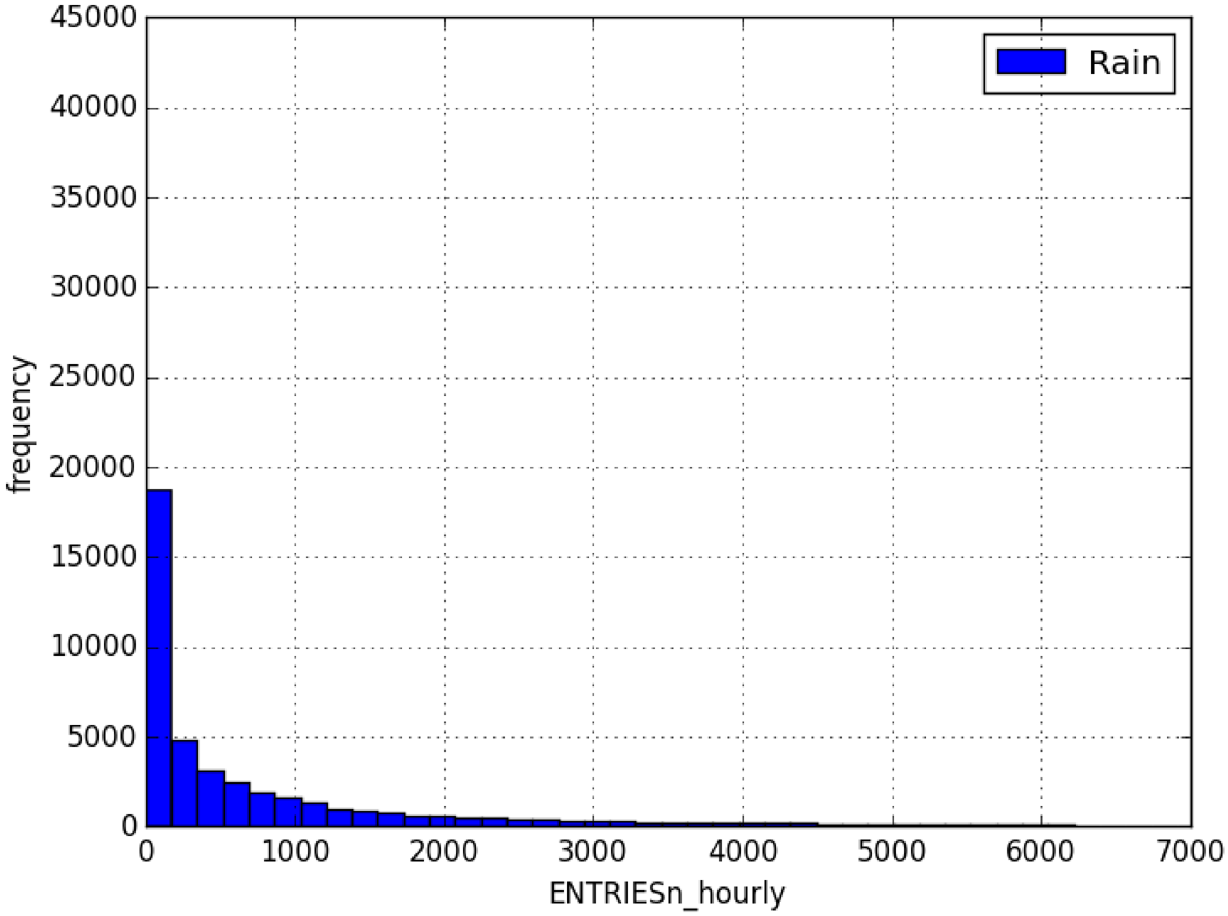
- The probability plot above shows the tail end of the values do not conform to the normal distribution. Once again the context matters and for ballparking the results, the results from linear model are sufficient.

Section 3. Visualization

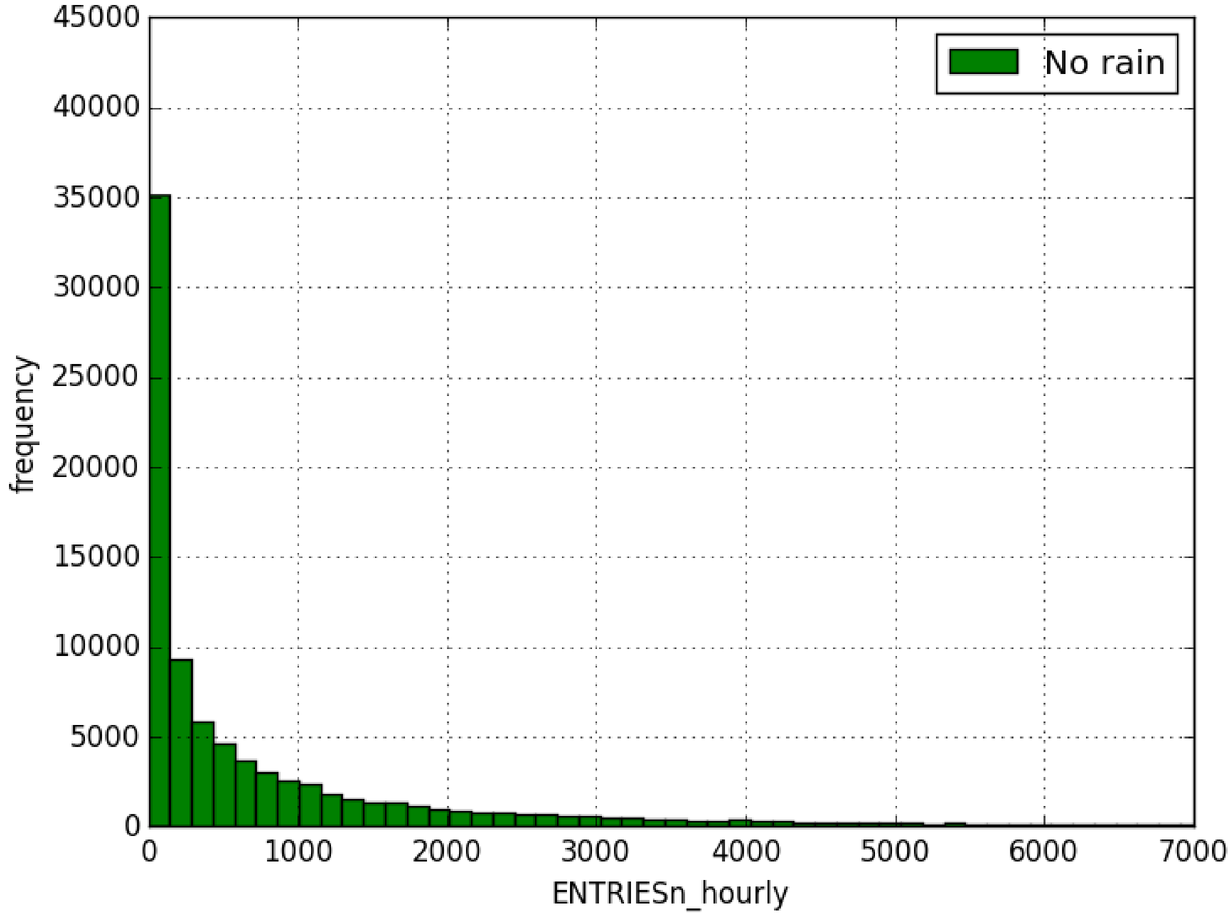
Please include two visualizations that show the relationships between two or more variables in the NYC subway data. Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

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.



- Plot above shows the distribution of subway ridership when there is rain.



- Plot above shows the distribution of subway ridership when there is no rain.

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

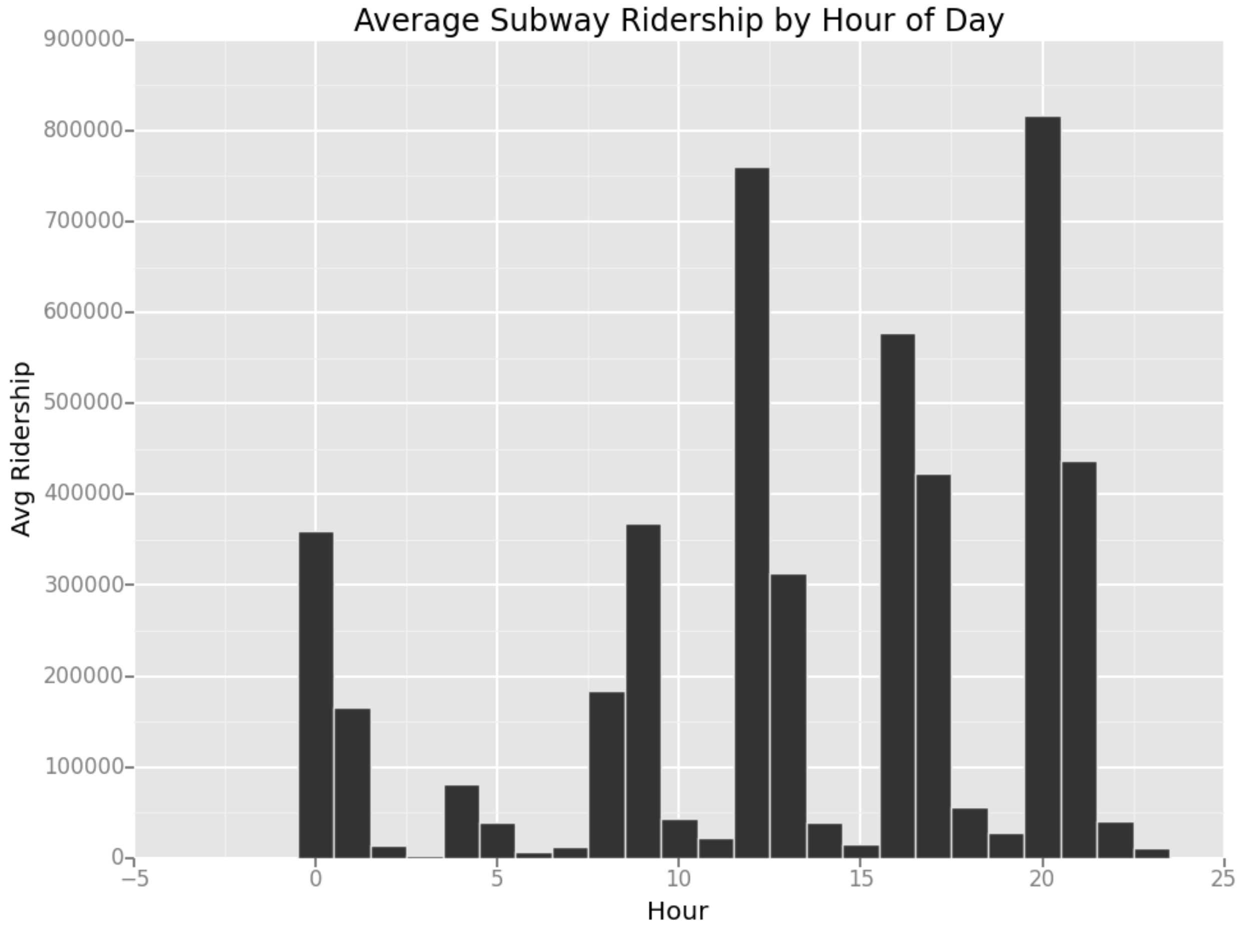
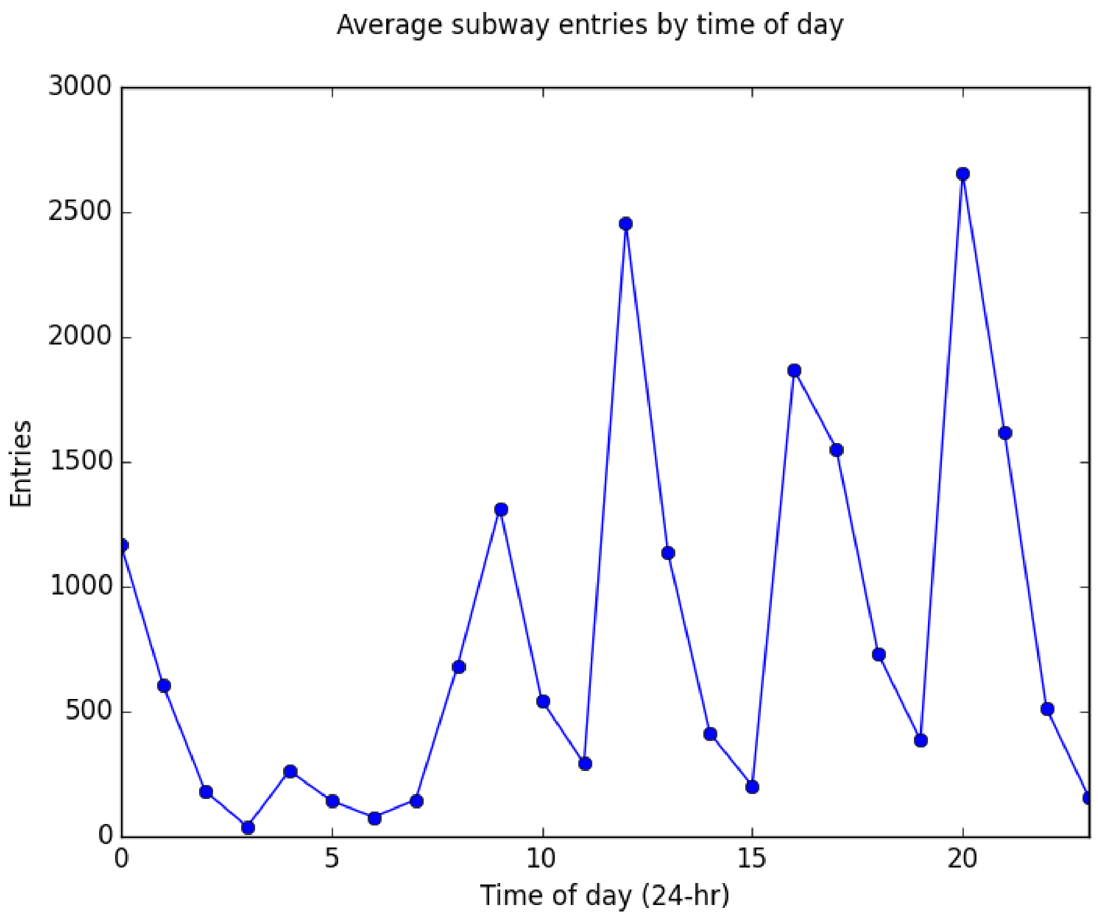


Figure above shows the Entries on an hourly basis.We find that during certain time the subway ridership is at its peak and is much above than the rest of the time. By, further examining the plot in detail i.e plotting the average entries on a daily bases we get the graph below:



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).

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?

- Let’s first observe the Mann-Whitney U test (p-value: 0.025), we can see that more ridership is present in the presence of rain. Given, the fact that the p value falls below the critical value, we can assume that the null hypotheis is false and both the distribution i.e the ridership with rain and without rain are indeed different.

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.

- Statistical test: Let’s first observe the Mann-Whitney U test (p-value: 0.025), we can see that more ridership is present in the presence of rain. Given, the fact that the p value falls below the critical value, we can assume that the null hypotheis is false and both the distribution i.e the ridership with rain and without rain are indeed different.

- Linear regression: Further, the positive co-efficient for rain indicates that there is increased ridership in the case when there is rain thus in unison with the results from the Mann-Whitney U test.

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:

- Number of entries do not match number of exits, probably there could've been some kind of miscount resulting in such a caluclation and this could've easily affected the effect of rain on riderhsip.
- One important factor would've been the inclusion of tursntile lcoation to provide a better insight into the ridership. It could've increased the accuracy and a larger data set could have helped validate this.
- The linear regression model was adequate and could answer the questions presented in this data set. However, from the probability graph, we can see that non linearity is something that could have been modeled to improve the accuracy of the system. Rather, than using the entire data set, the data set could've been split to give a better performance.