0. References

http://docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.stats.mannwhitneyu.html

http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test

https://statistics.laerd.com/spss-tutorials/mann-whitney-u-test-using-spss-statistics.php

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

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-statistic (python implementation) to analyze data. I used two-tail P Value. Null hypothesis is that two samples (rainy/non-rainy entries) really are the same and that an observed discrepancy between sample means is due to chance. P-critical is 0.05

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.

Mann-Whitney U test can be used to compare differences between two independent groups when the dependent variable is either ordinal or continuous, but not normally distributed. In our case:

1. Our dependent value (entries) is continuous.

2. Our independent value consists of two categorical, independent groups -
"rainy" and "non rainy". Obviously it could not be "rainy" and "non rainy" at one time.

3. We have independency of observations, "entry" occurs independently from all other future/past "entries", it can occur only when it is "rainy" or "non rainy".

4. Our samples are not normally distributed.

All things considered we can use Mann-Whitney 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.

Mean value for "rainy" sample : 1105.4463767458733

Mean value for "non rainy" sample" : 1090.278780151855

U: 1924409167.0

p: 0.049999825587

1.4 What is the significance and interpretation of these results?

$p < 0.05$ therefore we can say that there is a statistical difference between two samples and I can continue with further investigations.

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.

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

I didn't use dummy variables, I used weather (rainy, non rainy), hour of a day, mean temperature and mean pressure.

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.

I used weather, mean temperature and mean pressure because I thought that "good" weather could provoke people to go for a walk instead using subway. Also, while creating plots I discovered a strong dependency between time of day and entries. That's why I also used hour of a day.

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

weather (rainy, non rainy): 1.92440555

hour: 467.82139579

mean temperature: -67.61747639

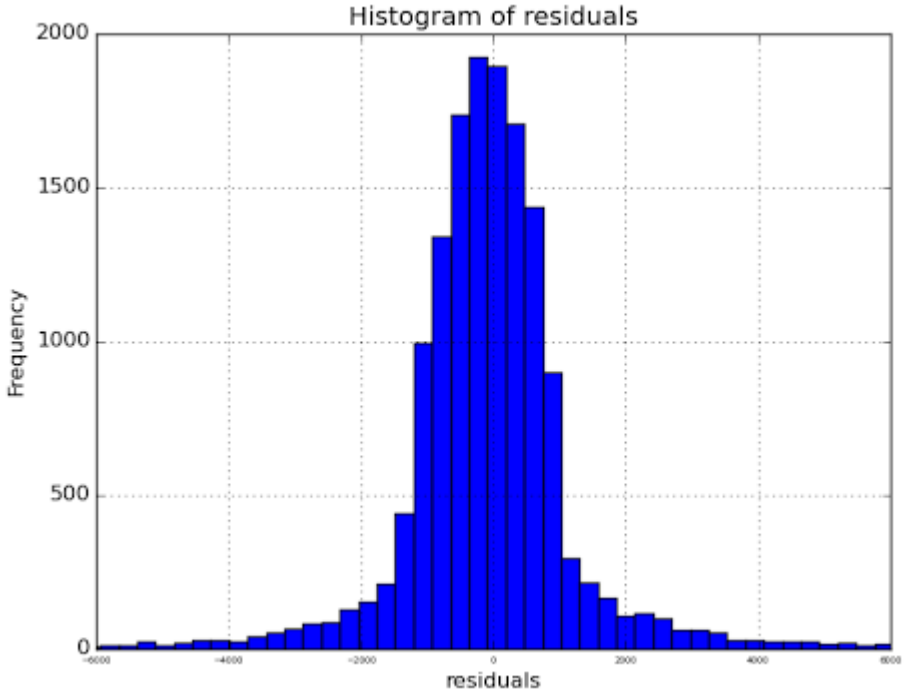
mean pressure: -48.94658952

2.5 What is your model's R^2 (coefficients of determination) value?

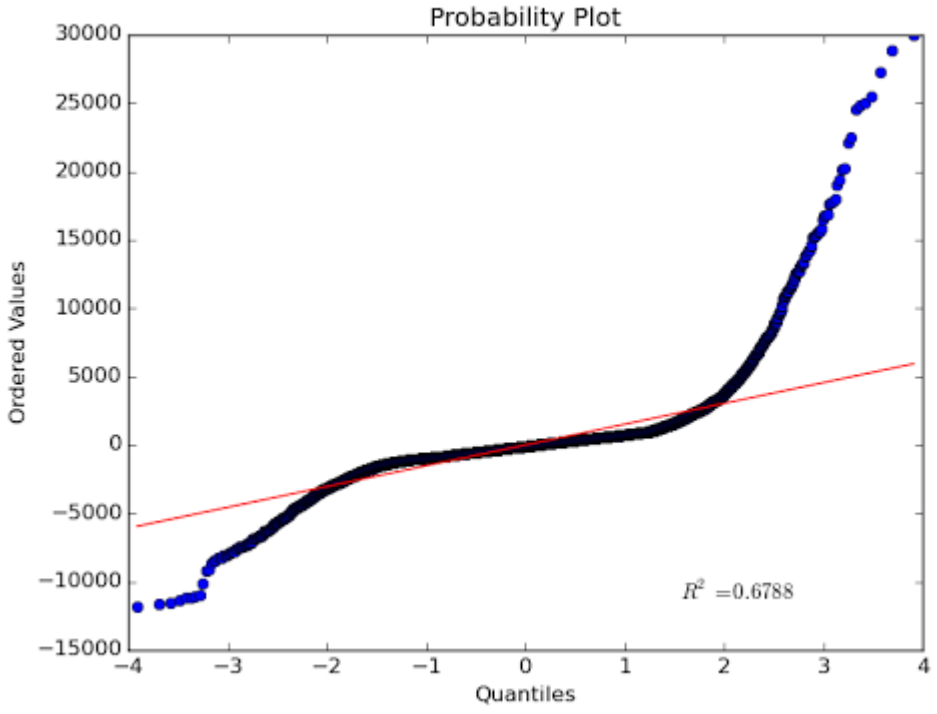
0.464334450963

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?

0.46 mean that we 46% of the variance in the entries can be explained by the explanatory variables (weather, mean temperature, mean pressure and hour of day). The remaining 54% can be attributed to unknown, lurking variables or inherent variability. In order to get better understanding of our model's performance I plotted histogram of residuals



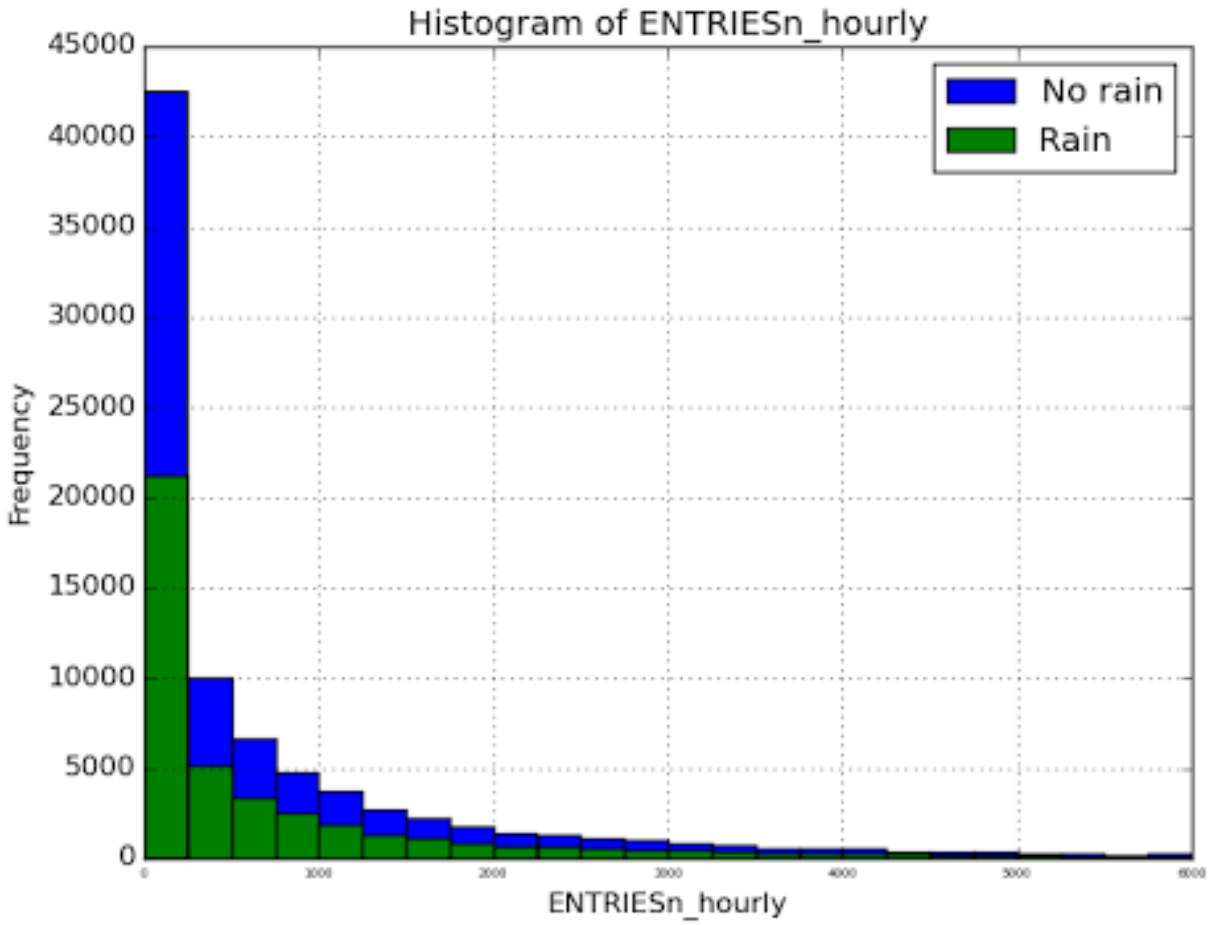
which has bell shape and distribution is approximately close to normal. However as it has long tails then further investigation is required: probability plot may give better answer on this question.



"S" shaped curve on this QQ plot suggests a bimodal distribution of residuals, therefore our model didn't catch some dependencies in input data.

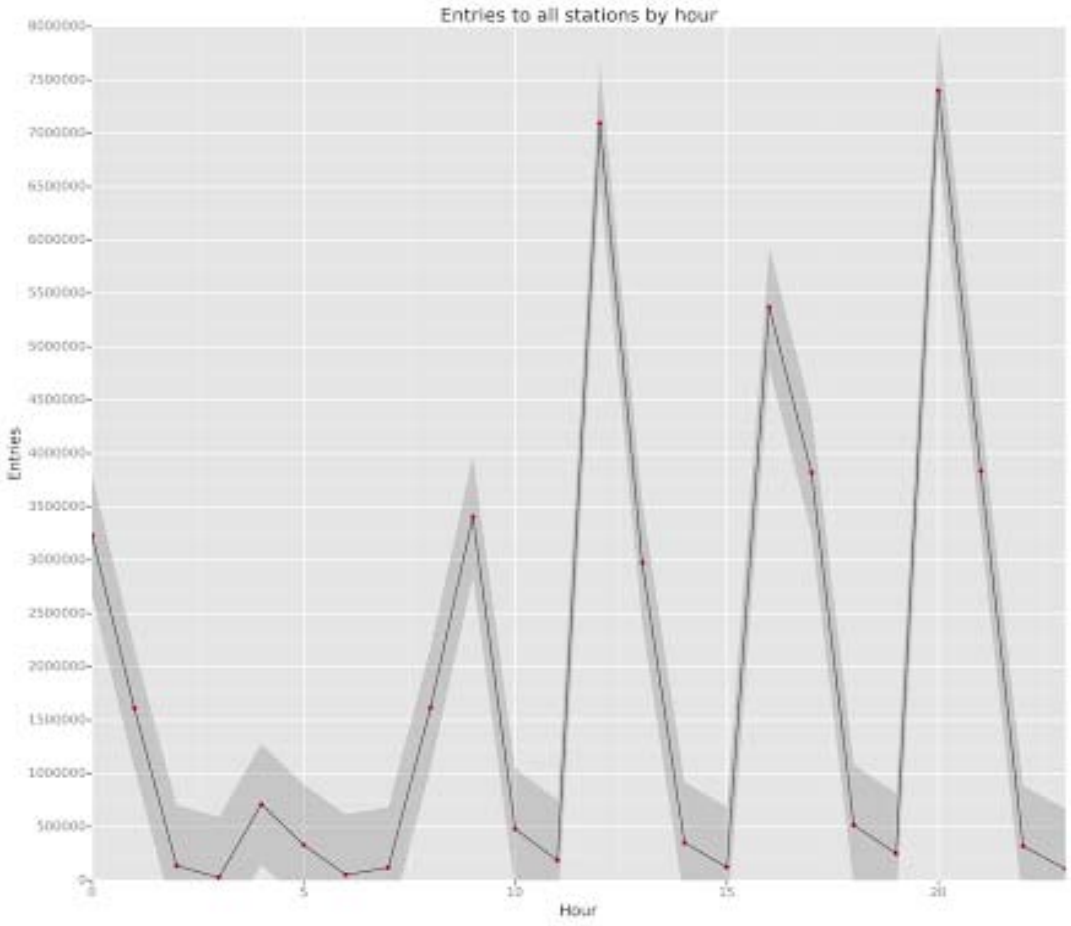
The answer to question "Does this linear model to predict ridership is appropriate for this dataset?" is "it depends on how we are going to use it". If we want to know if there are more ridership when it is rainy, then yes, it is appropriate. But if we are going to use this model to somehow optimize NYC subway (i.e. change the train schedule depending on ridership), then no, this is not going to work.

3.1

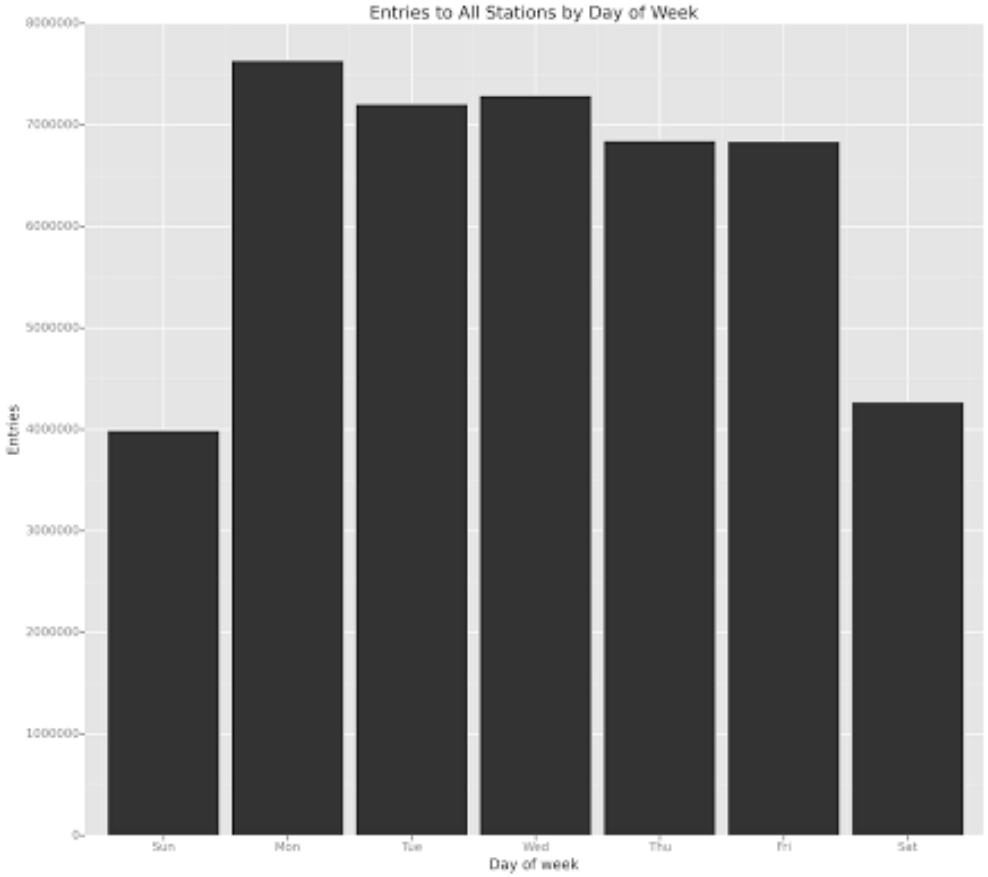


Two histograms combined: hourly entries when it is rainy and when it is not. Both samples are not normally distributed.

3.2



Entries to all NYC Subway stations by hour. There are peaks of ridership in the morning when people ride to their offices, then near 13:00 when people use subway to visit meetings or because of their personal needs. Pike neat 17:00 is the end of the operational hours, people return to home and near 20:00 when they probably attending different events in their free time.



Entries to all stations by day of week. As expected there are much less entries during weekend and there is a high splash of business activity on Monday which slightly decreases until Friday. I cannot explain Wednesday's 2nd highest activity, probably because it is a middle of a work week and people activity raises.

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?

Yes, more people ride the NYC subway when it is raining, linear regression and statistical test proved it. However it is not true that the difference is very big, rain doesn't strongly affect ridership, time of day has much more influence on it.

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.

Mann Whitney U discovered a statistical difference between two samples and comparing means of these two samples helped me to find out that there are more entries on rainy days rather than on non-rainy days. Also I created a linear regression model and obtained coefficient using gradient descent. The coefficient for 'weather' input variable (which can be 1 when it is rainy and

0 when it is not) is 1.92440555 which obviously means that when the weather is rainy then there are more subway entries.

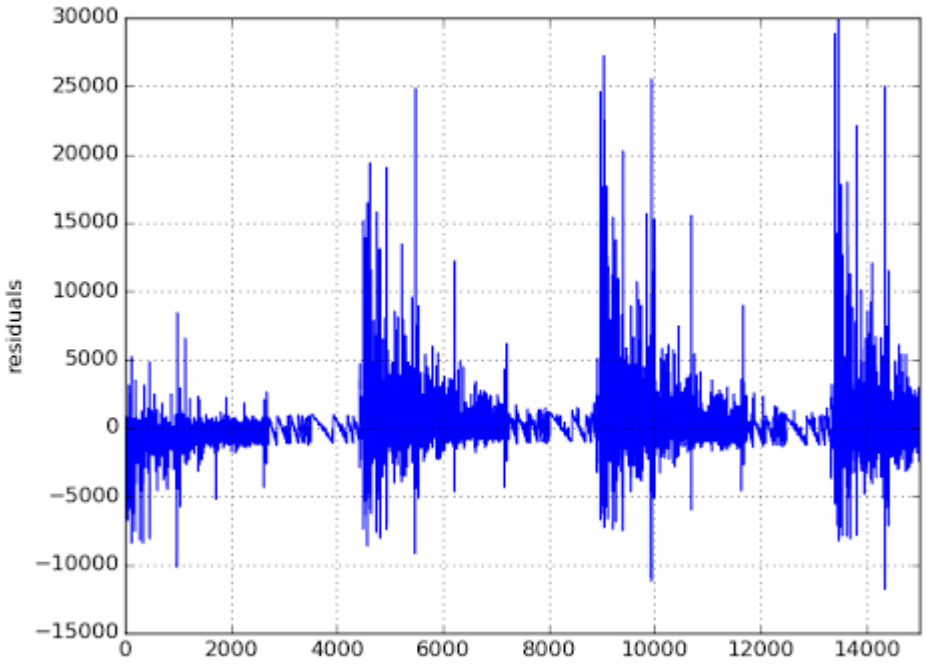
5.1 Please discuss potential shortcomings of the methods of your analysis, including:

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

Dataset provided could be better, I had a list of entries like 'station': 'entries to this station during a period of time', 'weather on this station during a period of time'. If I have a dataset with exact time and weather of each entry then my analysis could be give more precise results.

As statistical test only can provide a "yes"-"no" answer without any good quantitative evaluation therefore it can not be treated as deep data analysis tool, I used it in order to make sure that there is a sense in further investigations.

The analysis of R^2, residuals histogram and QQ probability plot allowed me to infer that our linear model is not a good fit for our data and probably there are some non-linear dependant input/output variables. Plotting the difference between predictions and actual values



confirmed my suggestion: residuals follow cyclical pattern. Therefore in order to get better results we should use non-linear regression.