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 the Mann-Whitney U test, as Mann-Whitney not considers the data is in Gaussian distribution or not. And as this is not yet known which data set would be higher or lower, and as the question here is about looking for any significant whether positive or negative, a two tail test is more appropriate here. The null hypothesis is that the two data set/population is almost same, or simply put there is no relation of rainfall with ridership. 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.

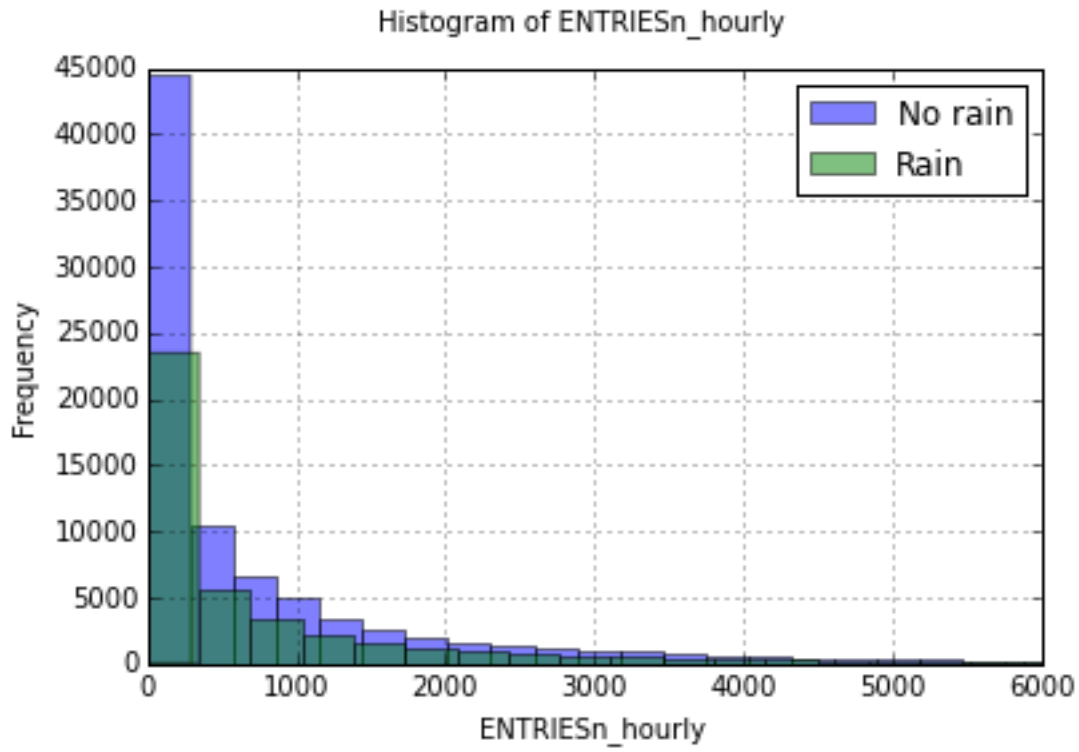


Fig 2.1: Histogram of ENTRIESn_hourly

As shown in fig 2.1, nor the entries during the rain, neither during the No rain the data is normally distributed. In such case Mann Whitney U test is more applicable, while a Welch two sample t-test is not. To confirm the results that neither of the data is normally distributed, Shipro-Wilk test could have been conducted.

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.

(1105.4463767458733, 1090.278780151855, 1924409167.0, 0.019309634413792565)

Mean entries with rain = 1105.4463767458733

Mean entries without rain = 1090.278780151855

U – Statistic = 1924409167.0

p-value = 0.038619268827585131

1.4 What is the significance and interpretation of these results?

The difference between mean entries with rain and without rain is around 1.4%. The Mean value with rain is clearly higher than without rain. The U – Statistic has a very high value, close to the maximum value of 1937202044. A U-stat of half the maximum would indicate that the null hypothesis is true. Further, p-value is around 0.039 satisfy the p-crit value, and the conclusion can be drawn with 95% confidence that the null hypothesis is false. And this conclude that the there is difference in ridership in between rainy days 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?

I used the OLS using statsmodel.

```
>>>
                            OLS Regression Results
==============================================================================
Dep. Variable:        ENTRIESn_hourly   R-squared:                       0.453
Model:                            OLS   Adj. R-squared:                  0.433
Method:                 Least Squares   F-statistic:                     22.32
Date:                Fri, 28 Aug 2015   Prob (F-statistic):               0.00
Time:                        16:14:03   Log-Likelihood:            -1.1488e+05
No. Observations:               13000   AIC:                         2.307e+05
Df Residuals:                   12534   BIC:                         2.342e+05
Df Model:                         465
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
rain        2.417e+13   1.65e+13      1.463      0.144     -8.22e+12  5.66e+13
precipi     3.123e+14   2.14e+14      1.463      0.144     -1.06e+14  7.31e+14
Hour          77.9328      2.187     35.631      0.000        73.645    82.220
meantempi     33.2936      4.601      7.236      0.000        24.275    42.312
```

Code:

```
model = sm.OLS(values, features).fit()

 results_rsquare = model.rsquared

results_predict = model.predict(features)

 results = model.summary()
```

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 are rain, precipitation, hour and mean temperature. UNITS is used as dummy variable.

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.

The hour is obvious choice for this experiment, and affects ridership in office hours and non office hours. The rain, precipitation and temperature can also cause the affect of ridership. For broadening the hypothesis from "people use subway more often in rainy days" to "people use subway more often in bad weather or thunderstorm". I also included thunderstorm, which causes very very slight increase in r-square value.

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

rain 2.416576e+13

precipi 3.122864e+14

Hour 7.793277e+01

meantempi 3.329358e+01

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

0.452983905103

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?

R-square is the percentage of variance, and to measure qualitatively on "goodness of fit" . In our case r-square is 0.4529, that is 45.29% of variation explained by our models.

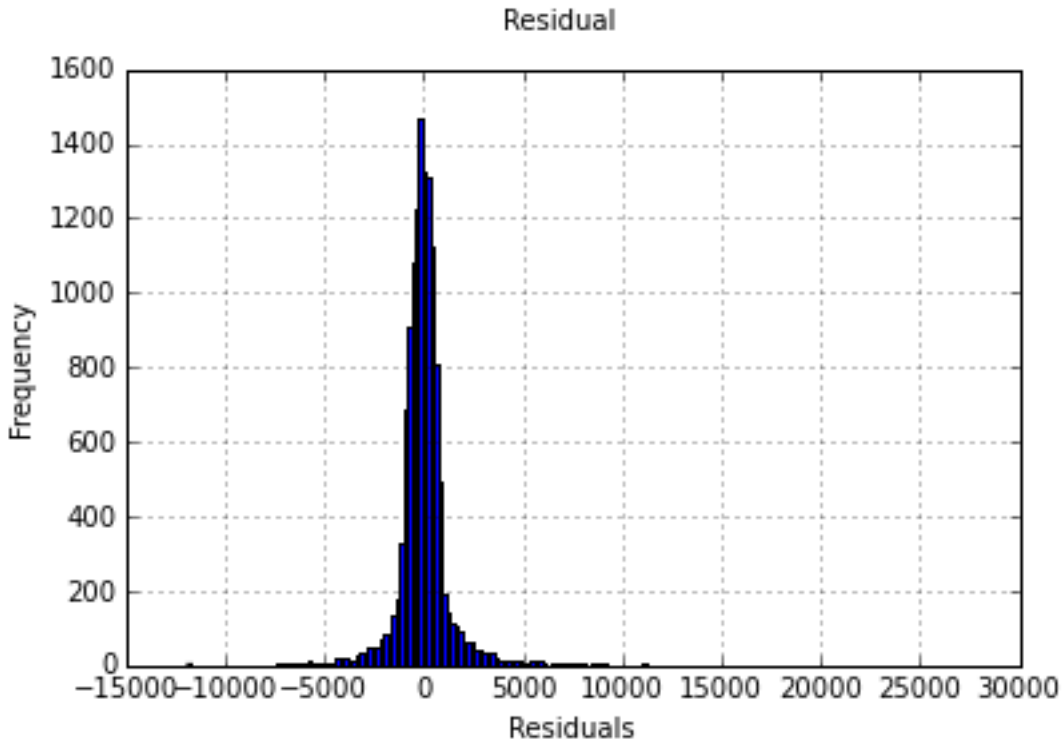


Fig 2.2: Residual plot

To conclude model is a good fit or not, one also have to put for what purpose the model going to use. If the use of the model is for safety, then definitely the model is insufficient. The most of residuals in this plot are in between and close to +/- 5000.

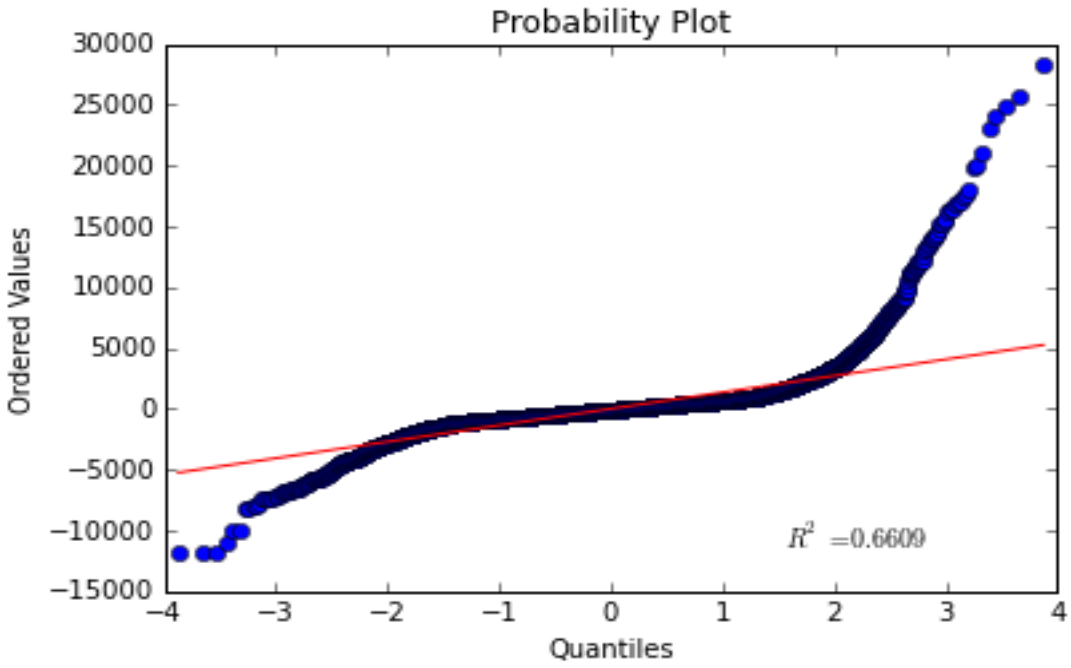


Fig 2.3: Probability plot

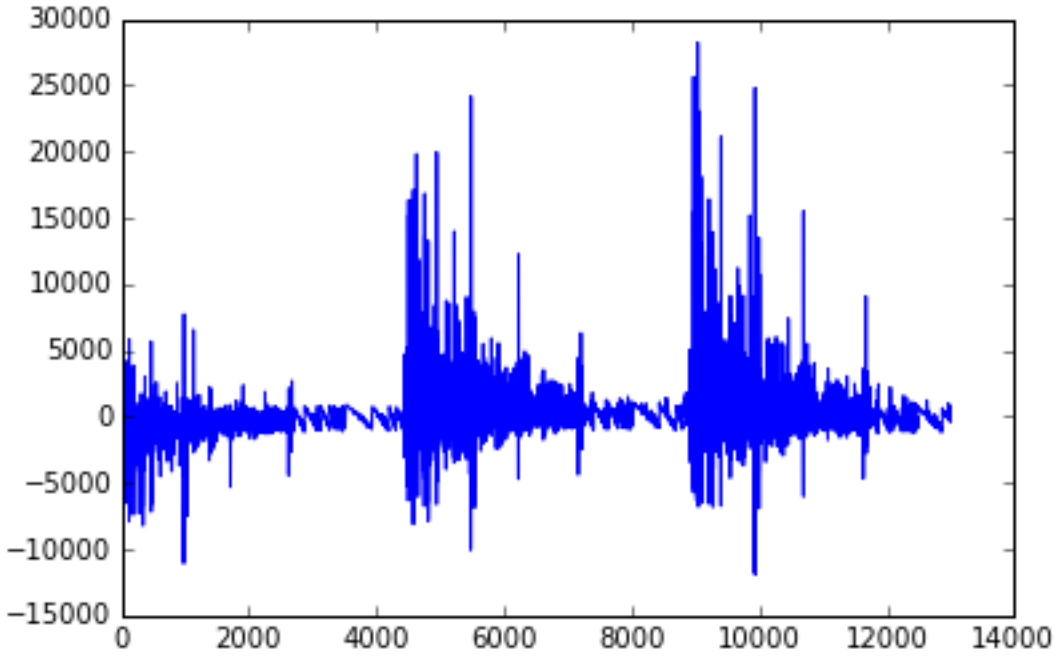


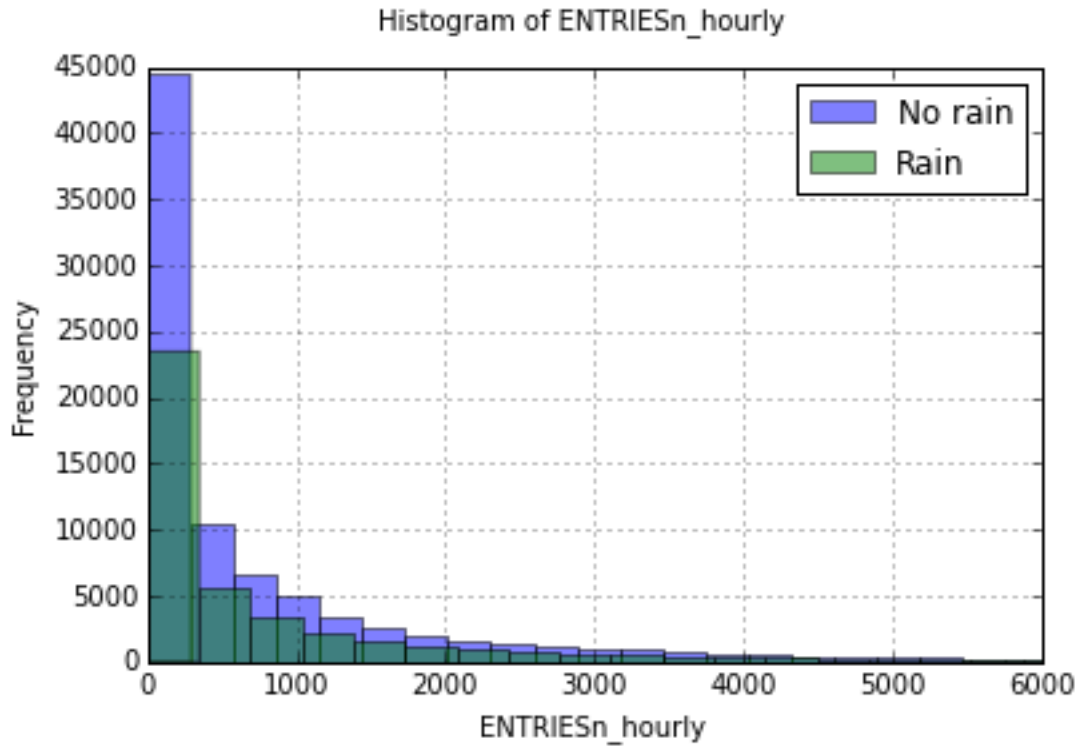
Fig 2.4: Residual per data point

The QQ-plot in fig 2.3 clearly described the heavy tails in the residual distribution. And can conclude some non-linearities may be missed in linear model.

By plotting the residual per data points clearly show the cyclic patterns in the residuals, which suggest a non linear model probably a better fit for dataset.

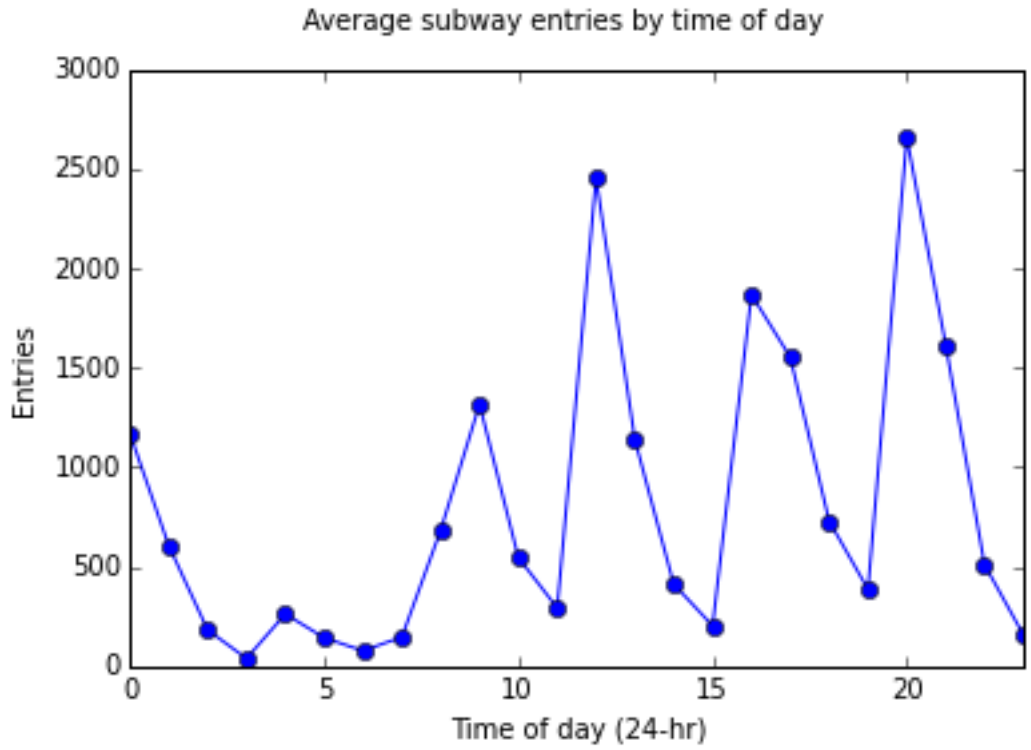
Section 3. Visualization

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.



The plot of Rain and No rain are definitely not normally distributed. This is important to describe that the number of rainy days are much lower than the number of non rainy days. So, this would be incorrect to conclude from this graph that the ridership in rainy days is lower than non-rainy days.

3.2 Include and describe a freeform visualization.



There are several peaks in the plot. Interestingly, most of the entries going on at night an evening instead of office hours from 8 am to 5 pm. And the highest peak at 8 pm is much more shocking. Without any further data, this is impossible to determine what's actually happening.

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?

With results of Mann-Whitney U test (p-value = 0.0386), we can conclude this more people ride the subway when it is raining. The conclusion from only means would be naïve and is insufficient to conclude anything. The Mann Whitney U test needed to confirms that there is a difference.

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.

With the Mann-Whitney U test results, we can reject the null hypothesis of the ridership in rainy days and non-rainy days, with the r-square being at 45.29%. Although the means of both data is not that different and almost similar, but the Mann Whitney U test indicate there is a change in

ridership in rainy vs non rainy days. So, this will be right to say there are significant changes in ridership in rainy and non rainy days.

Section 5. Reflection

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

One of the ambiguities in the data is about total number of entries in comparison to total number of exits. The total number of entries are 13474385, and exists at 10804295. The only possible reasoning of this problem is that there is some problem in counting, or some of the stations are not included in data set. But, as the effect of this is same of rain vs no rain, this had little effect on results.

As we can say by examining the 'UNIT' column, the ridership variation is greatly. This conclude, some stations are severely receiving much more entries than any other stations. The Mann- Whitney U test not consider this, and how the rain and no-rain affect on stations, or how the rain affect the ridership on the same day.

There are many variables that may show colinearity, such as mintempi,maxtempi and meantempi.

Finally, the predictions may have been limited by the use of linear analysis. Some of the variables may have non-linear effects and a non-linear model may predict more accurate results.

REFRENCES

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

http://statsmodels.sourceforge.net/stable/

http://mpastell.com/2013/04/19/python_regression/

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

http://stackoverflow.com/questions/13865596/quantile-quantile-plot-using-scipy

https://www.youtube.com/watch?v=-KXy4i8awOg