Section 0: References

http://stackoverflow.com/questions/16689514/how-to-get-the-average-of-dataframe-column-values

http://stackoverflow.com/questions/7781798/seeing-if-data-is-normally-distributed-in-r

http://pandas.pydata.org/pandas-docs/stable/install.html

http://stackoverflow.com/questions/30406564/python-ggplot-how-do-i-layer-histograms

http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

Section 1:

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 because it has greater efficiency t-test on non-normal distribution.

- Shapiro-welk test was used to determine if data was normal or not.

- Two-tail P value

- My Null hypothesis is: the two populations are the same, rain has no effect on ridership

- p-critical value 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 was used because it has greater efficiency t-test on non-normal distribution.

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 entries with rain = **1105.44637675**
Mean entries without rain = **1090.27878015**
p-value = **0.024999 (1 tailed) -> 2-tailed = 2*0.025 = 0.05**
U-statistic = **1924409167.0**

1.4 What is the significance and interpretation of these results?

When it’s raining, there’s 1.39% more in ridership. Also, the value of U-statistic is very high. U-statistic close to 0.5 of max would mean that the Null Hypothesis is true. Since our U-statistic is very high, this indicates that the null hypothesis is false.

Also, our p-critical value is 0.05 and p-value is 0.0249, which is below 0.05. Therefore, we can conclude with 95% confidence that the null hypothesis is false.

Ridership is different when it’s raining and not raining.

Section 2:

2.1 What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn_hourly in your regression model:

OLS using stats model was used to make predictions

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

Rain, precipitation, hour, mean temp, and also fog, meandewpti, meanpressurei

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 reason I chose these features is that it increase the R^2 value. Here’s the breakdown of the features included, and resulting R^2 value

Features included	**R^2**
precipi, Hour, meantempi	0.458033
rain, precipi, Hour, meantempi	0.458044
Fog, rain, precipi, Hour, meantempi	0.458213

2.4 What are the parameters (also known as "coefficients" or "weights") of the non-dummy features in your linear regression model? (problem set 3.5)

From the gradient_descent code:

```
Rain: 2.92398062e+00
```

```
Precipi: 1.46526720e+0
```

```
Hour: 4.67708502e+02
```

```
Meantempi: -6.22179395e+01
```

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

0.47924770782

After adding **fog** as an input variable, R2 becomes 0.47985

Using "meanpressurei" would yield 0.4795 for R2

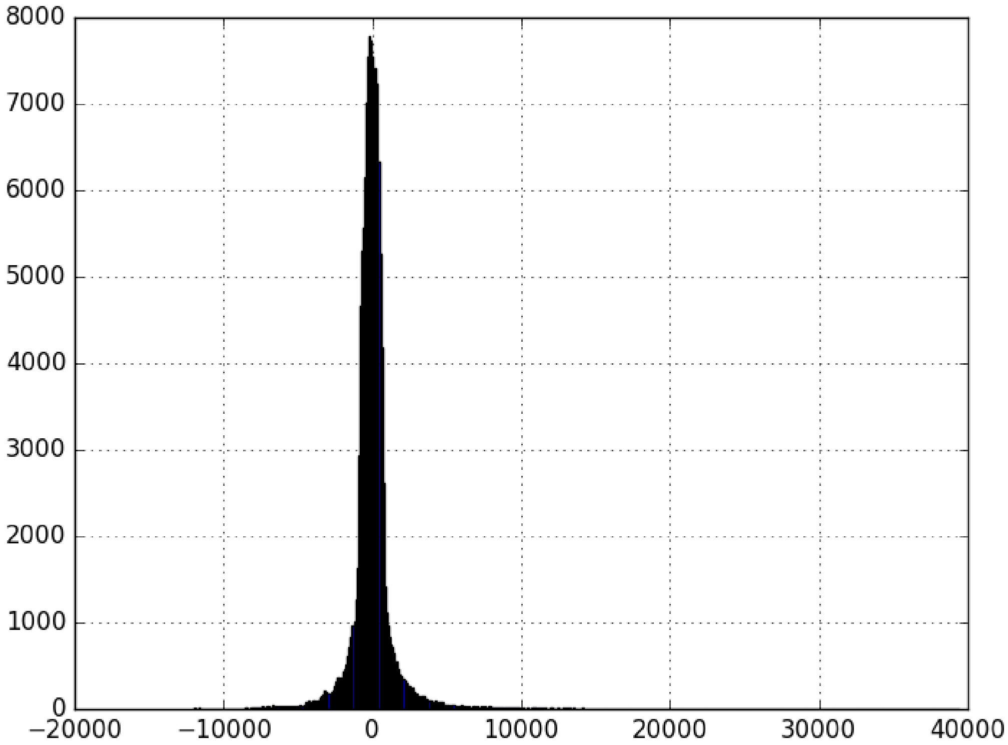
Using "meandewpti" would yield 0.4793 for R2

Other "bad weather" elements did NOT significantly increase the R2 value, only with fog there was a very slight increase in R2

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?

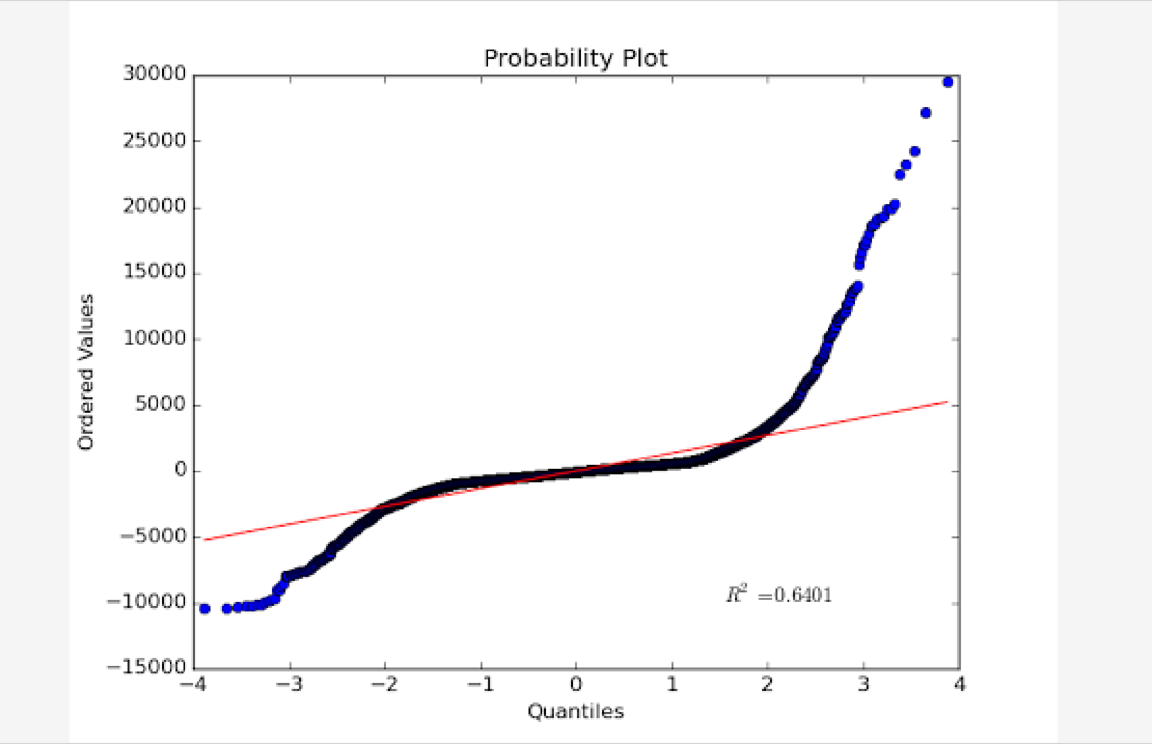
R2 is a statistical measure of how much of the variation can be explained by the model. However, R^2 alone cannot tell us if our model, linear regression, is the right one. We need to look at the tails of the residual plot to conclude if linear regression is a good fit

state whether a linear model is appropriate for this dataset:



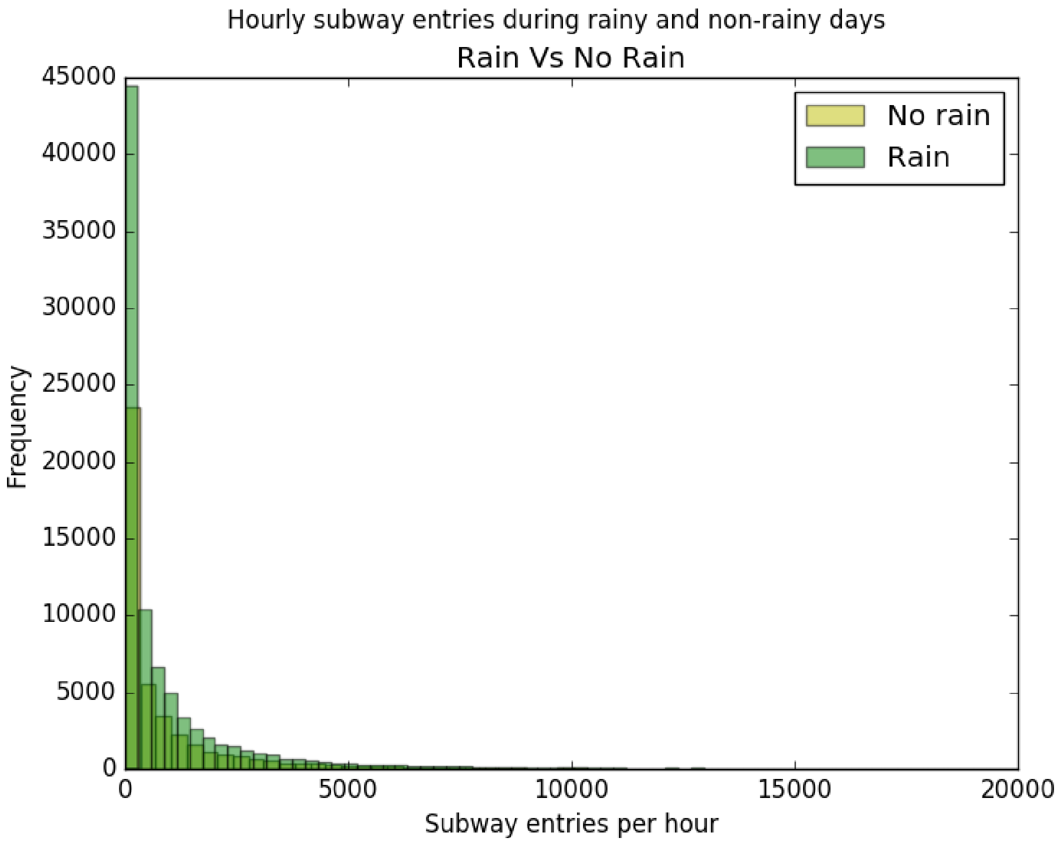
***** residual chart**

A probplot is also created to calculate the quantiles for a probability plot. Here, you can see that it shows our model, linear regression, is not too good of a fit



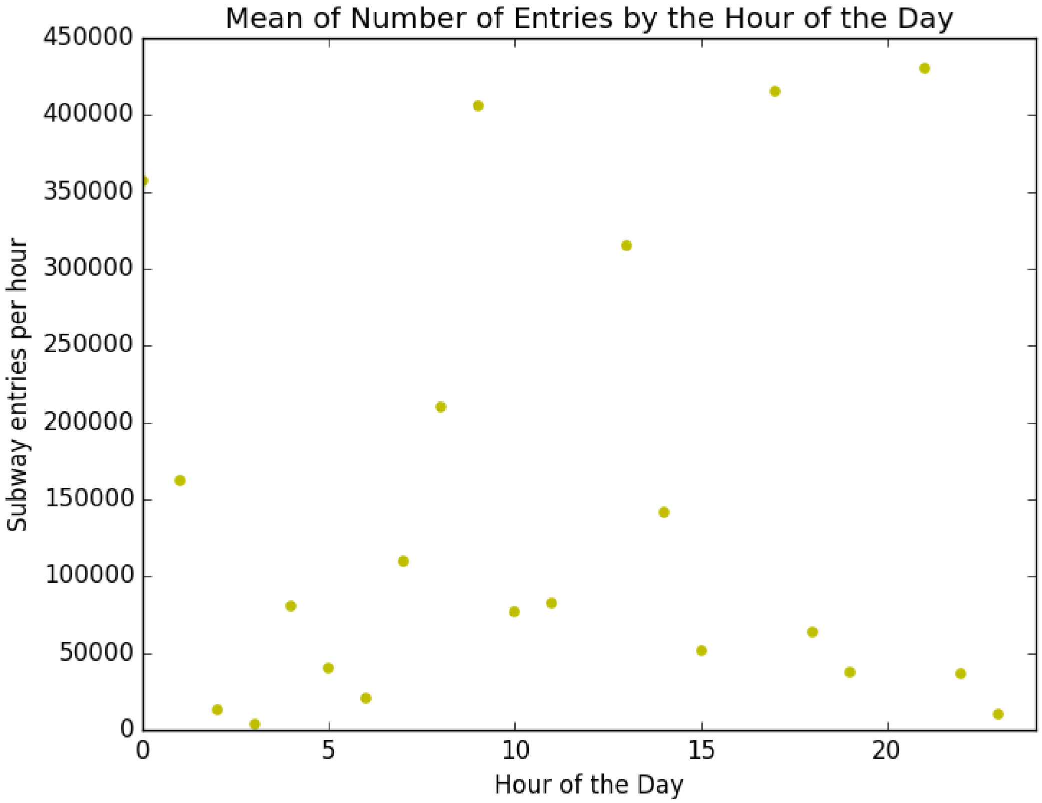
Section 3: Include 2 visualizations

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



* this visualization is about subway entries per hour on rainy and non-rainy days

3.2 One visualization can be more freeform.



*** This visualization is on the aggregate value of mean of number of entries per UNIT and dates, by the time (hour) of the Day**

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?

The results of the Mann-Whitney U test (p-value = 0.05 for two tailed test). Since p-value is small, we can confidently reject the null hypothesis, and conclude that more people ride the NYC subway when it is raining

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.

"Therefore, the appropriate quantity to be analyzed here is the regression coefficient of the rain variable."
A positive regression coefficient of rain (2.92398062e+00) indicates that it has positive increase in ridership, which is consistent with result from the Mann-Whitney test

Section 5: Reflection

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

1. Dataset,

   The dataset covers from 5/1/11 to 5/30/11, only 1 month of data. This seems to be fairly limited and does not have good representation of other months in the year.

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

   On section 2.6, the probability plot shows the our model, linear regression, might not be the best fit. The R2 value, 0.478, indicates that only ~48% of the data can be explained by the model, not the other 52%. With more than half of the data cannot be explained by the model, linear regression is not a good model for this data set.

   Feature Selection – Currently we're only using "trial and error" approach for feature selection. Formal method such as SelectKBest was not used. Therefore, feature selection can be improved.

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