Section 0. References

- http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test
- http://en.wikipedia.org/wiki/Welch%27s_t_test
- http://docs.scipy.org/doc/scipy-0.16.0/reference/generated/scipy.stats.mannwhitneyu.html

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 to compare the means of ridership during rainy and non-rainy days. I used 2-tailed test because the aim was to find if there was any significant difference positive **or** negative.

Null Hypothesis: The ridership is same for both rainy and non-rainy days and that there is no statistically significant difference between the two days.

The alpha used to reject/accept the hypothesis was 0.05 or 5%. This means if there is a difference of 5% or more, we can accept the null hypothesis that the ridership is same.

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 assumes that the samples don't have an underlying probability distribution (like normal distribution). Both the rainy and non-rainy distributions for our data have long-tailed shape hence a non-parametric test like Mann-Whitney U test is appropriate for the dataset.

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.

Results

Rain Mean ridership	1105.446
Non-Rain Mean ridership	1090.278
Difference in mean	15.168
U	1924409167.0
P value	0.0249 (2%)

The conclusion is that there is a statistically significant difference in the mean ridership on rainy vs non-rainy days. The p-value is 0.02 < the alpha of 0.05 hence we can safely reject the null hypothesis that there is no significant difference between the rainy and non-rainy day ridership.

1.4 What is the significance and interpretation of these results?
Based on the results, we can conclude that on an average, more people ride the subway on rainy days compared to 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:

OLS using Statsmodels or Scikit Learn
Gradient descent using Scikit Learn
Or something different?

I used the Ordinary Least Squares from Statsmodels to compute predictions for ***ENTRIESn_hourly***.

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

The features used in the prediction were UNIT the dummy variable, hour, rain and fog.

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 R^2 value."*

I tried combinations of different features.
UNIT: was a dummy variable because units in areas of high population are expected to have more riders.
Hour: was used because it has a strong correlation with ENTRIESn_Hourly.
Rain and Fog: I initially used these on a hunch that if it is raining or foggy, people would prefer to ride the subway than drive or walk. I later realized it did not have as much effect as I had hoped.

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

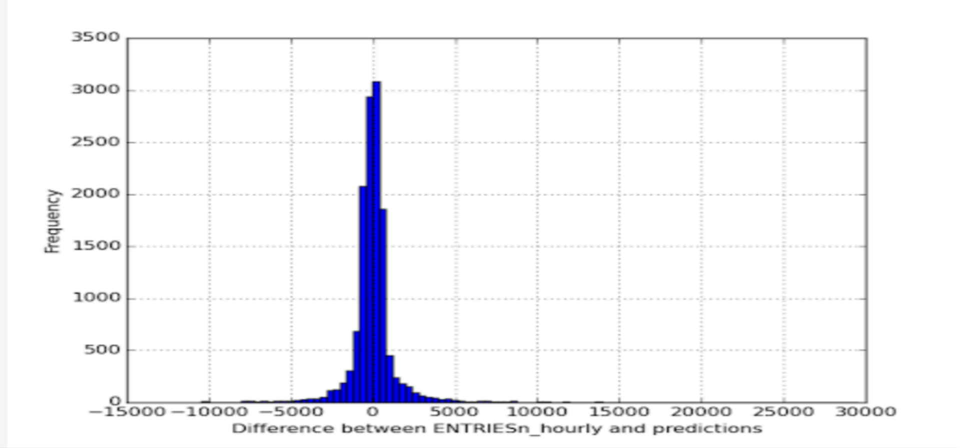
```
                            OLS Regression Results
==============================================================================
Dep. Variable:        ENTRIESn_hourly   R-squared:                       0.479
Model:                            OLS   Adj. R-squared:                  0.460
Method:                 Least Squares   F-statistic:                     25.03
Date:                Sun, 15 Nov 2015   Prob (F-statistic):               0.00
Time:                        22:49:28   Log-Likelihood:            -1.1674e+05
No. Observations:               13195   AIC:                         2.344e+05
Df Residuals:                   12727   BIC:                         2.379e+05
Df Model:                         467
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const        849.1264     32.029     26.511      0.000       786.345   911.908
Hour          65.3009      2.205     29.617      0.000        60.979    69.623
rain          17.4508     35.780      0.488      0.626       -52.684    87.585
fog          137.0755     45.411      3.019      0.003        48.062   226.089
```

2.5 What is your model's R² (coefficients of determination) value?
With the above features and the partial dataset provided on the udacity site, I could get a value of **0.478727011371 which is just under 48%**

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?

R² is the measure of how good the fit is. The R² from the experiment only explains about 48% of the variance. The plot below indicates the residuals. It is clear that the residuals were close to 0+-5000. This prediction only gives an approximation and is nowhere close to being accurate.

Weather this prediction is appropriate or not, entirely depends on the use case and what we need to accomplish using the predictions. If precision is not crucial, then we can use this model to estimate the ridership but if we needed accurate numbers for things like capacity planning of other resources like tickets, security, utilities at the station etc, then this would not give an accurate result.



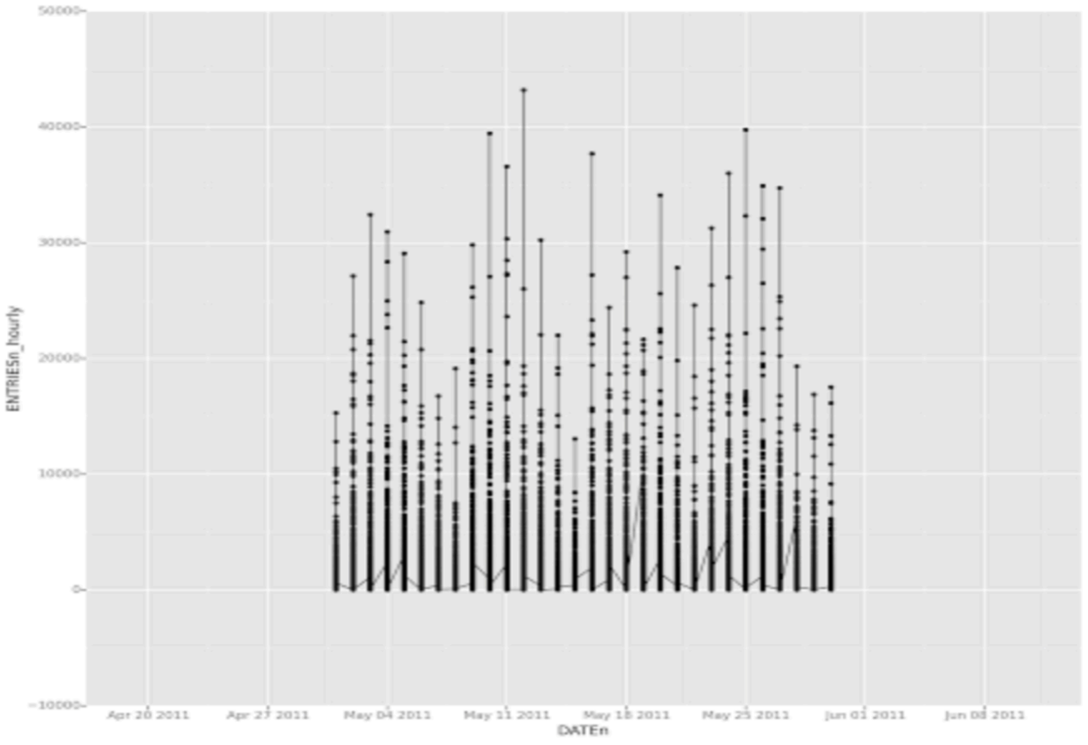
Section 3. Visualization

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.

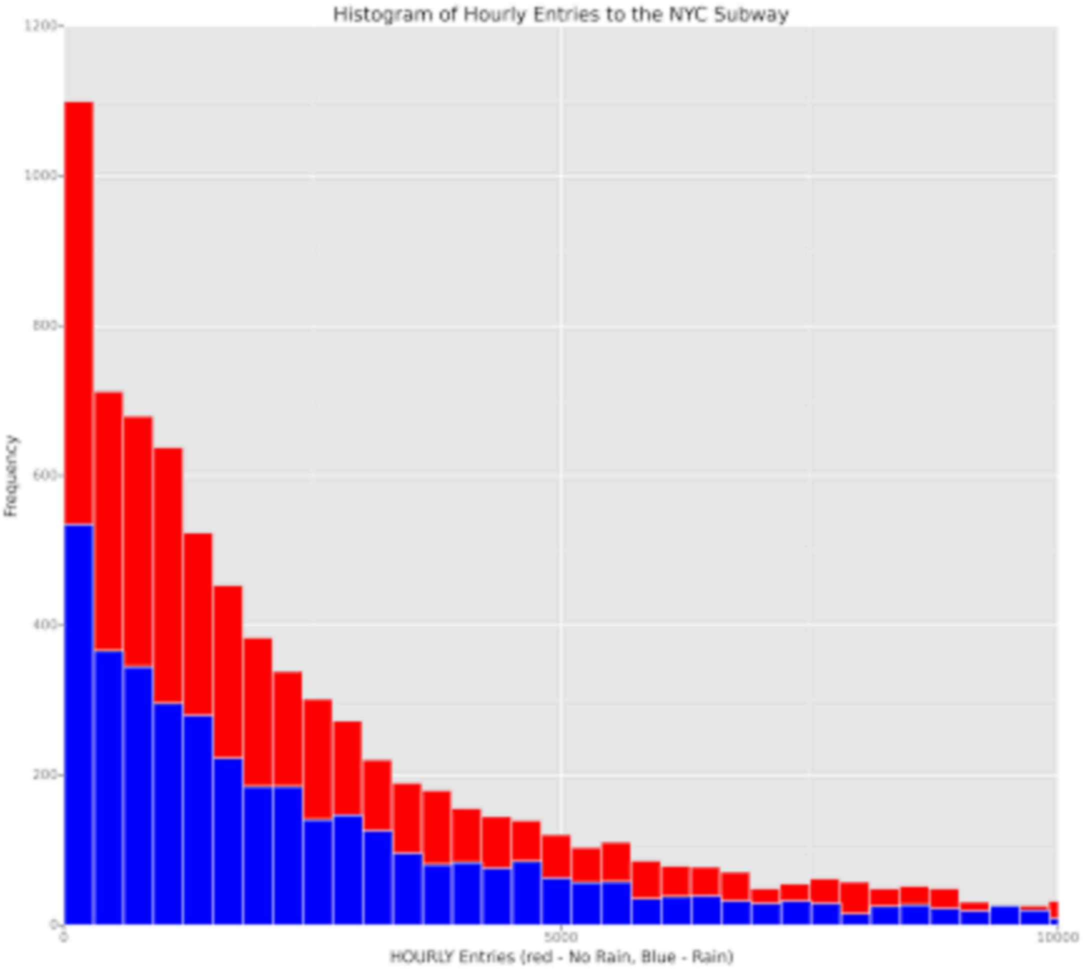
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

> *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.*



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
Ridership by day-of-week



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?
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.

Based on my interpretation of data, more people ride the subway during rains. I say this based on the difference of ~15 between the means in both cases. Although this is a difference, it is not high enough to definitively say that the difference is because of the rains.

This is also further supported by the computation of R^2 in the predictive model. When we add/remove rain from the predictive model, there isn't any significant change in the value.

The final histogram also indicates that the distribution is almost identical.

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:

Dataset,

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

The major shortcoming of the analysis was that the data was limited to a month. We know that season varies throughout the year and if rain were a reason for change in ridership, then we should use data from other months also.

It is not possible to guess weather may is a typical month or not in comparison to the rest of the year. For all we know, May might see higher ridership due to the very fact that it is likely to rain or it might be a month where the ridership is lower than the rest of the year.

Also, if we enhance the model to fit perfectly to the test data we have, it might not work with the actual complete data due to overfitting. We would be training and testing the data based on a single month's data, which might not be a typical month, hence resulting in the model not working as expected.