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?

Answer 1.1

Initially I used Mann-Whitney U test to Analyze New York city subway data set.

We used a Two-tail P value; C=counter validate our null hypothesis.

Null hypothesis in this case was that rain and ridership has no co-relation.

The p-critical values used would be 0.5 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.

I used this test as the data for rain and no-rain is 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.

with_rain_mean = 1105.4463767458733,
without_rain_mean = 1090.278780151855,
U = 1924409167.0,
P = 0.024999912793489721

1.4 What is the significance and interpretation of these results?

Increase of about 1.4% in ridership in New-York subway; But it's not sufficient to prove our null-hypothesis.
Very High value of U statistics 1924409167.0 and very low value for P 0.024; helps us draw conclusion that the proposed null hypothesis is false; and ridership is different for rainey 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:

1. **OLS using Statsmodels or Scikit Learn**
2. **Gradient descent using Scikit Learn**
3. **Or something different?**

1. Initially I use OLS using Statsmodels; without forgetting to add a constant in the features
2. Gradient descent using Scikit Learn - experimented with this as well

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

Answer 2.2

I used features mentioned below; from subway data set:
'rain', 'precipi', 'Hour', 'meandewpti', 'meanwindspdi'
No additional dummy features were used; except default dummy_units which added 'Units' features later to the data frame.

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 went with the features:
'rain', 'precipi', 'Hour', 'meantempi' - as they indicated rain and precip at certain duration of time at a certain temp; as well as they are provided default parameters with R2 0.42
'meandewpti', 'meanwindspdi' – Both of these features dew and wind speed are essential in predicting weather conditions; and improvement in value of R2 indicates the same.

‘fog’ - Just went with a hunch that it might be also useful to predict; increase in value of R2 proves the same,

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

OLS or Ordinary least square is a unweighted linear regression analysis.

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

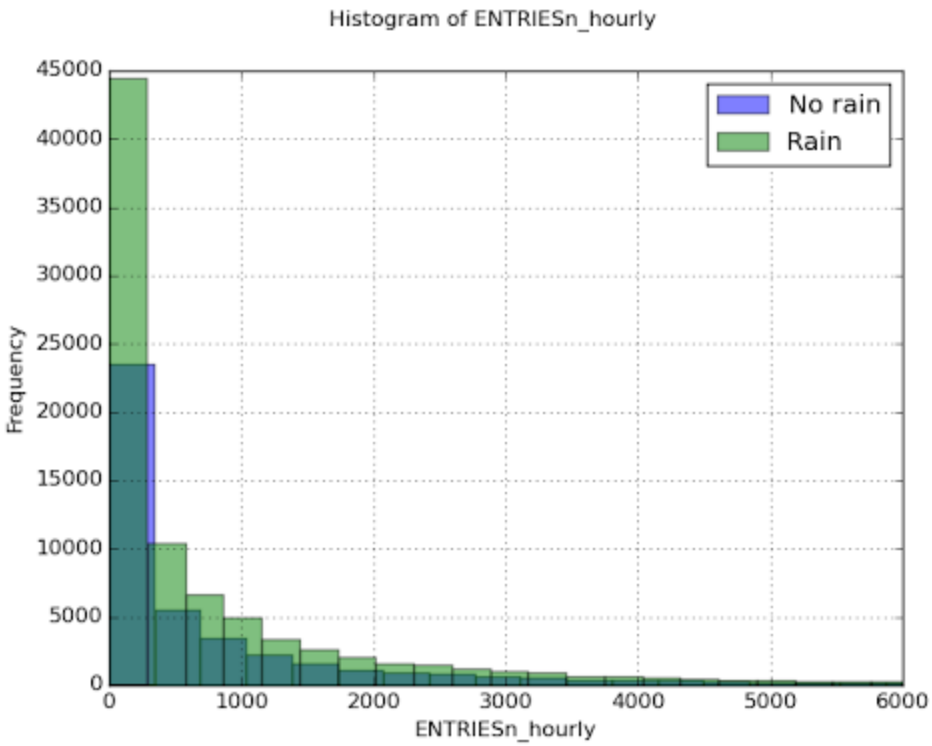
My model's R2 (coefficients of determination) value = 0.480502743303

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 value is suppose to as close to 1 as possible for better fit of the regression model.
I think my liner model ridership is appropriate but I think it can be improved; still.

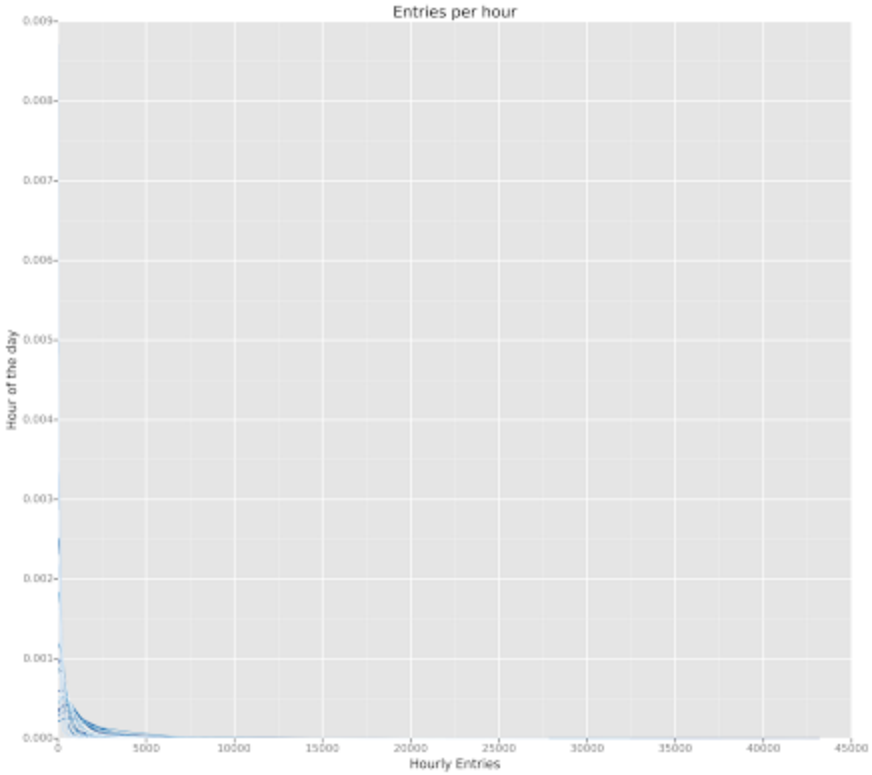
Section 3. Visualization

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



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:

The plot is my attempt to use geom_density to plot Entries per hour data which is colored per 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?

As per my analysis and interpretation of data, people do ride more on NYC subway when it's raining compared to when it's not; and the first and more than 95% conclusive test I did was Mann-Whitney U test; which had critical P-value of 0.25.Though the test was not helpful for predictions.

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.

First what I did was i came to conclusion that people do ride more on NYC subway while it's raining; for that I calculated critical P-value (=0.25 for this case) and U (1924409167.0) giving

us sufficient doubt to believe the null hypothesis; and used mean values of riders on rainy and non-rainy days. Concluding that more people do ride on subway of NYC on rainy days.

Now using this using linear regression I tried to build a basic prediction system based on the data; and was able to understand that rain (0 or 1) is one of most important feature in this but there are other feature that would allowed me to increase accuracy of my model R2 to 0.48.

Section 5. Reflection

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.**

Based on the shortcomings would reflect that if the size of the dataset can be increased it would have helped more in linear regression; which compared to Mann-Whitney's U test does not just based on riders per rainy day for all subway stations of NYC.

But also bases on other features and if the size of the dataset could be increases we can divide it in training-set (60%), cross validation set (20%) and test-set(20%).

Bibliography:

http://en.wikipedia.org/wiki/Ordinary_least_squares
http://en.wikipedia.org/wiki/Gradient_descent
http://www.ats.ucla.edu/stat/mult_pkg/faq/general/tail_tests.htm
http://en.wikipedia.org/wiki/Null_hypothesis
sklearn.linear_model.LinearRegression — scikit-learn 0.14.1 documentation
yhat/ggplot
http://pandas.pydata.org/pandas-docs/stable/10min.html