Section 0.References

- http://graphpad.com/quickcalcs/distMenu/
- https://s3.amazonaws.com/udacity-hosted-downloads/t-table.jpg
- http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html
- Intro to inferential statistics at Udacity

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 one-tailed t-test for independent variables.
Null-hypothesis is: ridership in rainy days is the same or less than on not rainy days: $\mu_r <= \mu_{nr}$
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.

The assumption is that the distribution of ridership in both samples is normal.
t-Test is most applicable because we don't know anything about population, sample's sizes are larger than 30 elements, have different size and variance.

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 of samples with rainy days: μ_r=2028.196
Mean of samples with not rainy days: μ_{nr}=1845.539
T test value: 5.042
The one-tailed P value is less than 0.0001

1.4 What is the significance and interpretation of these results?

Found p value is way below p critical value. Thus we have to reject hypothesis zero. μ_r is statistically significantly higher than μ_{nr}.

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 gradient descent using Scikit Learn.

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

I used following input variables: rain, meantempi, hour, weekday. Also dummy variables were included: UNIT, conds.

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 was selecting variables based on intuition and then I was experimenting with them to confirm gains in R^2 value.

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

rain	*meantempi*	*hour*	*weekday*
5.22423987e+01	-2.06333212e+01	1.19462828e+02	8.94726262e+02

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

$R^2 = 0.473658678629$

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?

R^2 value is far from ideal value – 1, which means that used model might not be good enough for making predictions based on given data set.

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.
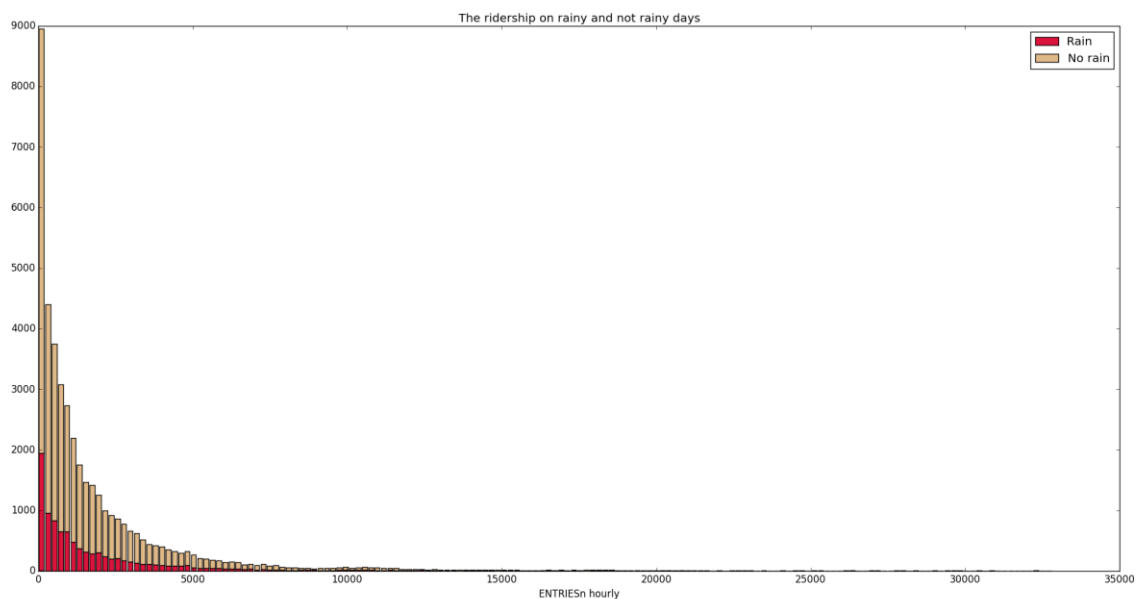

Picture 1 shows that there is much less occurrences and frequencies for not rainy days.
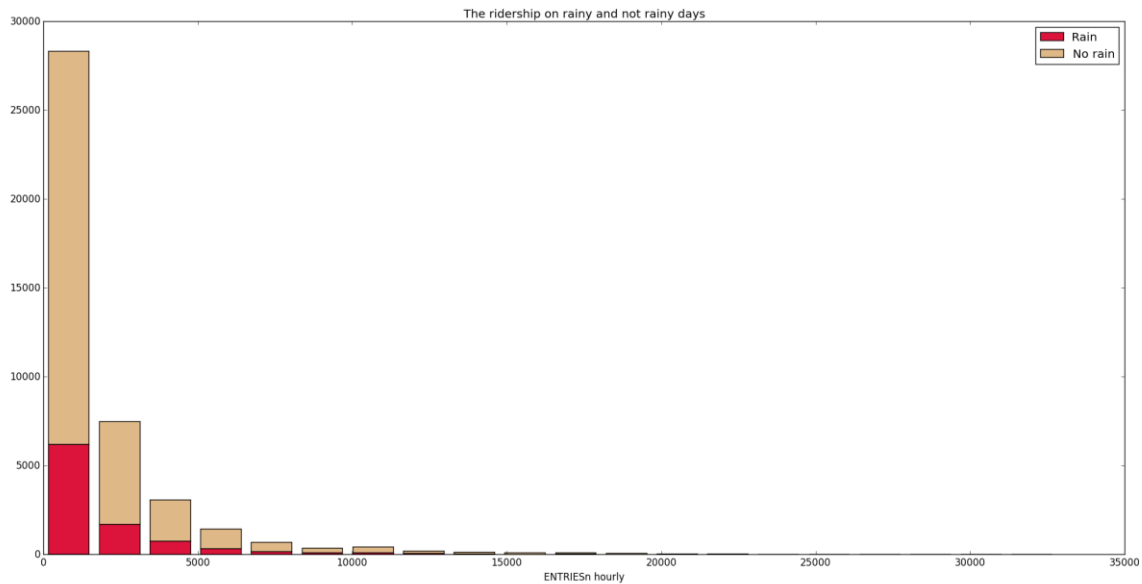
Picture 2 has smaller number of bins comparing to picture 1. Thanks to that we can see higher frequencies for lower values of ENTRIESn_hourly.

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.
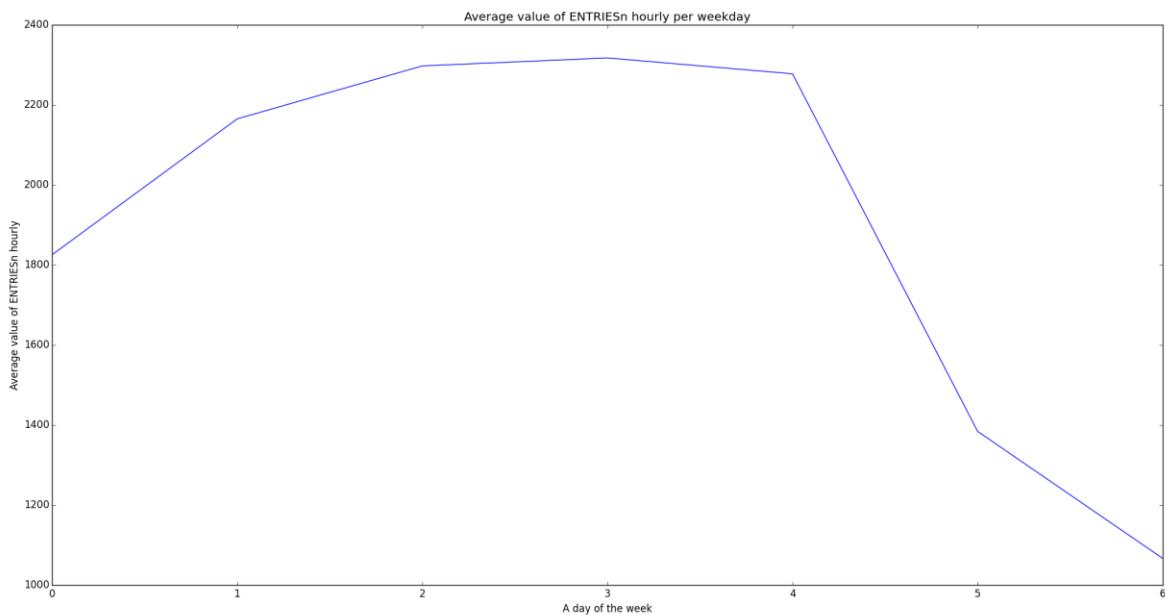
Picture 3 Ridership is almost twice higher on weekday than on weekend days.

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?**

Based on statistical test and linear regression can be conclude that more people ride the NYC subway when it is raining. Visualizations made based on provided data set can be misleading.

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.

From statistical test can be concluded that more people ride the NYC subway when it is raining. The mean of ridership for rainy day is higher than for not rainy days. T test value is positive and much higher than t-critical for one-tile test.

From linear regression results - rain has weight $5.22423987e+01$, which means that the rain has meaningful and positive influence whether people ride subway or not.

Histograms presenting ridership on rainy and not rainy days can be misleading, because sample size with not rainy days is much bigger than sample size with rainy days. This can cause wrong impression and lead to false conclusions.

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.

For statistical test we did an assumption that distribution is normal. We should check if distribution is normal, if not then the conclusion made based on statistical test can be wrong.

Calculated model based on linear regression didn't provide good results although the choose of variables seems to be correct. This leads to conclusion that linear regression might not be enough to make predictions based on provided data set.

I don't see any shortcomings regarding dataset. Data are clean and the sample's sizes are big enough to make fruitful analysis.