Section 0.References

- https://s3.amazonaws.com/udacity-hosted-downloads/t-table.jpg
- http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html
- http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit
- Intro to inferential statistics at Udacity
- http://en.wikipedia.org/wiki/Multicollinearity

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 two-tailed non-parametric test: Mann-Whitney U test.
Null-hypothesis is: ridership in rainy days is the same as 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.

This statistical test is applicable to the dataset because it doesn't assume any particular distribution.

Executed Shapiro-Wilk test returned values w=0.5938820838928223, p=0.0 for rainy days and w=0.5956180691719055, p=0.0 for not rainy days what allows in both cases to reject hypothesis zero and state that data are not normally distributed.

T-test assumes normal distribution so it can't be used.

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
P value for one-tailed test: 0.0249
P value for two-tailed test: 0.0499

1.4 What is the significance and interpretation of these results?

Found p value is below p critical value. Thus we have to reject hypothesis zero. μ_r is statistically significantly different than μ_{nr}. Mean of samples with rainy days is larger than mean of samples with not rainy days what allows to conclude that ridership is higher on 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 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. Table below shows the results of a selection process.

A list of variables	**R^2 value**
'rain','UNIT'	0.360
'rain','meantempi','UNIT'	0.360
'rain','meantempi','hour','UNIT'	0.449
'rain','meantempi','hour','weekday'	0.104
'rain','meantempi','hour','weekday','UNIT'	0.464
'rain','hour','weekday','UNIT','conds'	0.470
'meantempi','hour','weekday','UNIT','conds'	0.470
'rain','meantempi','hour','weekday','UNIT','conds'	0.474

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=0.474658678629

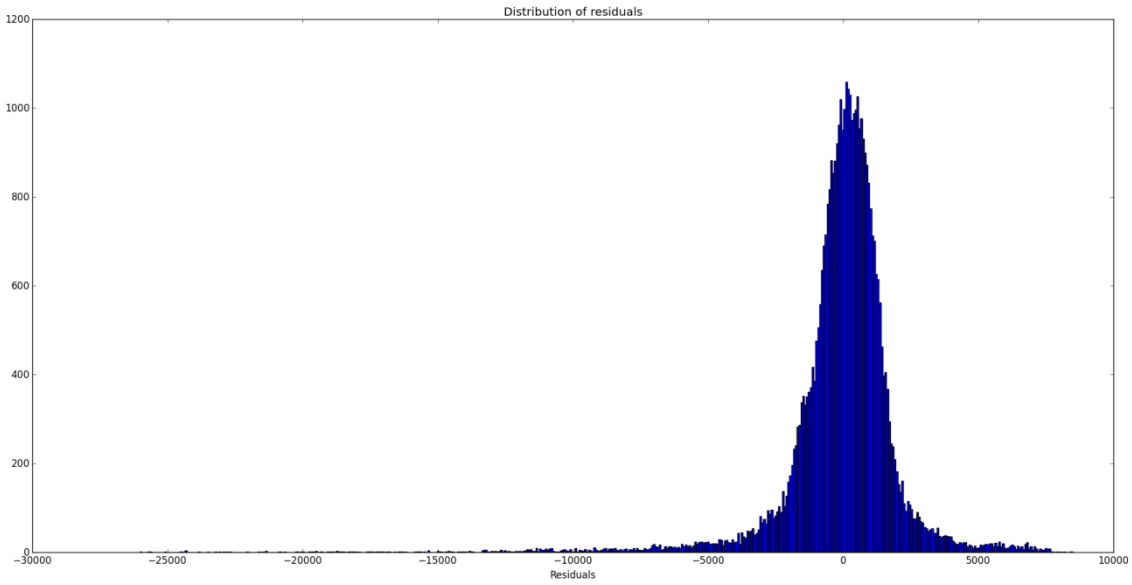
2.6 What does this R^2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?

R^2 is the percentage of the response variable variation that is explained by a linear model. 0% indicates that the model explains none of the variability of the response data around its mean. 100% indicates that the model explains all the variability of the response data around its mean. R-squared does not indicate whether a regression model is adequate.

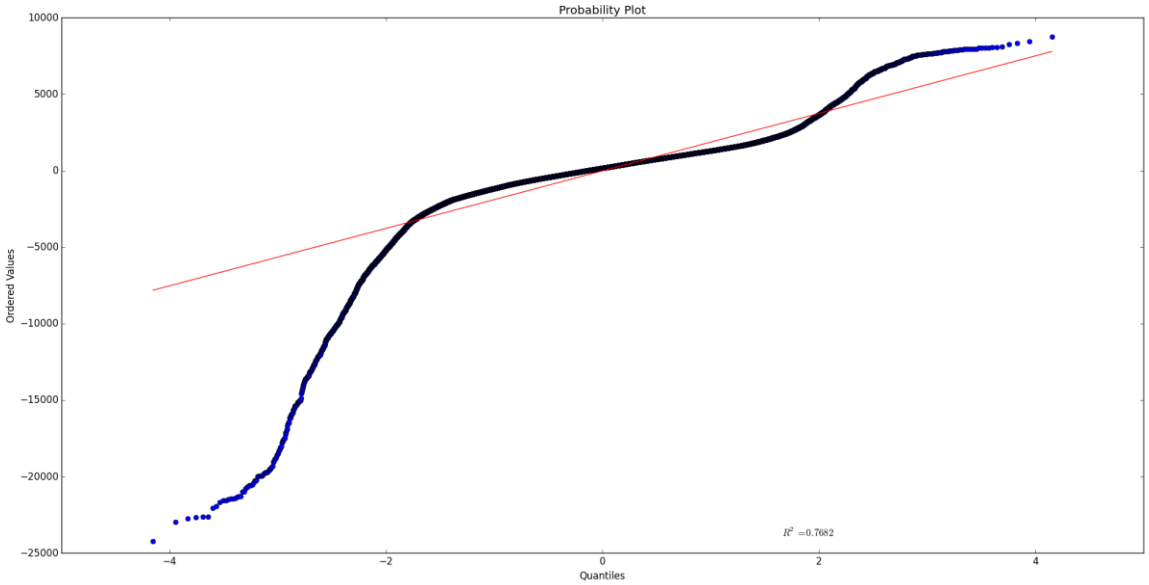
There can be a low R-squared value for a good model, or a high R-squared value for a model that does not fit the data.

R-squared cannot determine whether the coefficient estimates and predictions are biased, which is why it is required to assess the residual plots.

Picture 1 shows a histogram with distribution of residuals. We can observe that the tails of the histogram are long which suggests that some residuals are very high and distribution is not normal. Additionally picture 2 with Q-Q plot proves the same conclusion. Some points follow a strongly nonlinear pattern, suggesting that the residuals are not normally distributed. This leads to conclusions that our model might not be useful and appropriate for given dataset.



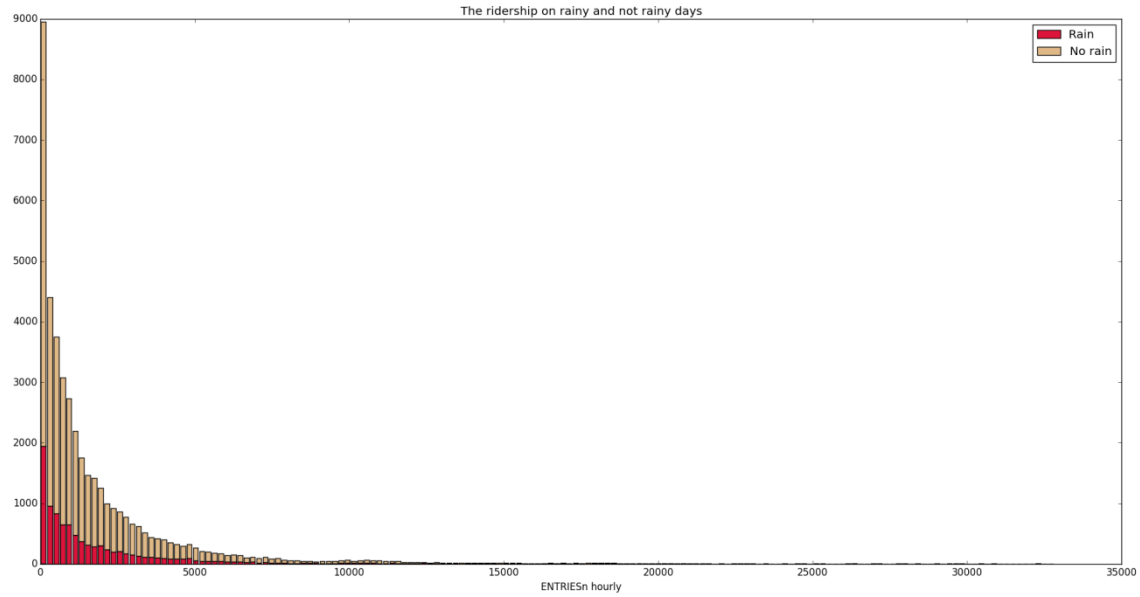
Picture 1 Distribution of residuals



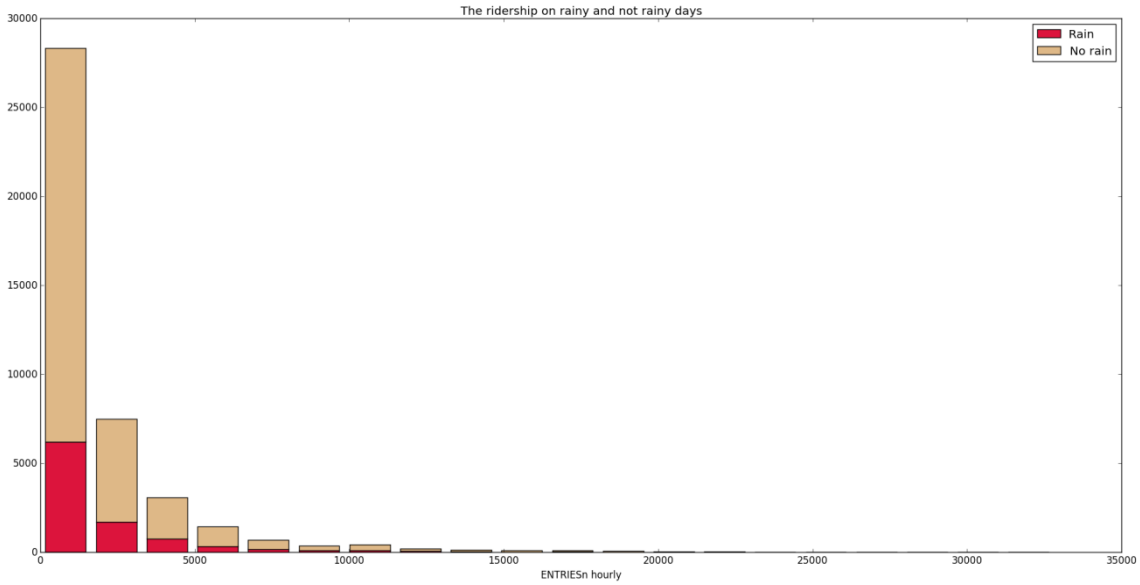
Picture 2 Q-Q plot of residuals

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.

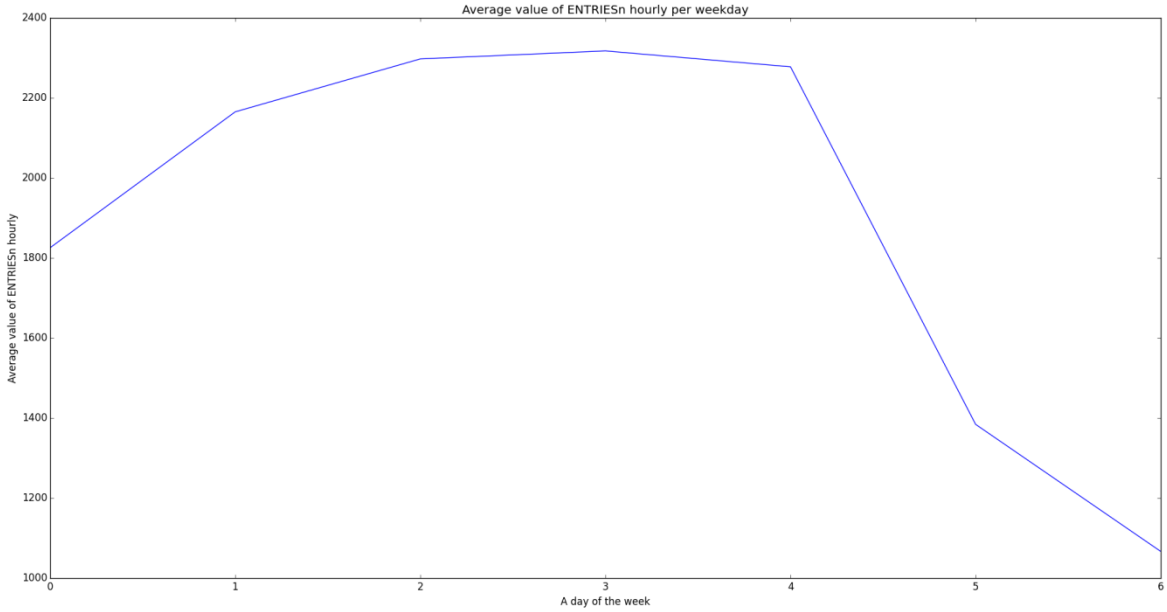


Picture 3 shows that there is much less occurrences and frequencies for not rainy days.



Picture 4 has smaller number of bins comparing to picture 3. 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 5 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.

Calculated model based on linear regression has not very high coefficient of determination and a distribution of residuals is not normal which leads us to conclusion that linear regression might not be a good method for making predictions in our case.

Given dataset was limited to one month – May. For other months or season collected data might look different and the conclusions might be different. On the other hand may is be most representative regarding weather conditions – it is not too warm or too cold.

Additionally the dataset has the closely related variables around temperature and pressure. Including those variables into analysis can lead to collinearity, which can cause some linear regression algorithms to give incorrect results.