Section 0. References

Please include a list of references you have used for this project.

Discussion forums
Piazza
https://piazza.com/class/i4ltdrdhqak4r7
Udacity discussion forums-
http://discussions.udacity.com/c/nd002-2015-02-04

Panda
Learning
https://bitbucket.org/hrojas/learn-pandas
Working with Missing Data
http://pandas-docs.github.io/pandas-docs-travis/missing_data.html
Indexing & selecting Data
http://pandas.pydata.org/pandas-docs/stable/indexing.html#boolean-indexing

Numpy
http://docs.scipy.org/doc/numpy/reference/generated/numpy.mean.html
http://docs.scipy.org/doc/numpy/reference/generated/numpy.sum.html

Python DateTime conversion
http://strftime.org/
http://docs.python.org/2/library/datetime.html#datetime.datetime.strptime

CSV Reader/Writer Tutorial
http://goo.gl/HBbvyy

Welch's t- Test
https://www.youtube.com/watch?v=2-ecXltt2vI
Understanding p-value
https://www.youtube.com/watch?v=eyknGvncKLw
Mann Whitney U Test
http://www.statisticssolutions.com/mann-whitney-u-test/

GGPlot
http://blog.yhathq.com/posts/ggplot-for-python.html

Residual Plots
http://www.itl.nist.gov/div898/handbook/pri/section2/pri24.htm

Map-reduce
http://discussions.udacity.com/t/string-maketrans-what-is-this/4676

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?

A two-tailed Mann-Whitney U test was used to analyze the NYC subway data.

Null hypothesis- No. of riders using the subway when it is raining is equal to the no. of riders when it is not raining. Rain has no effect on ridership.

P- Critical value= 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.

Histograms for ridership on rainy vs. non rainy days showed non-normal distributions. See Fig. 1.2. Hence a two-tailed Mann-Whitney U test was chosen for analyzing these two populations. One of the assumptions of the test is that the data is not normally distributed.

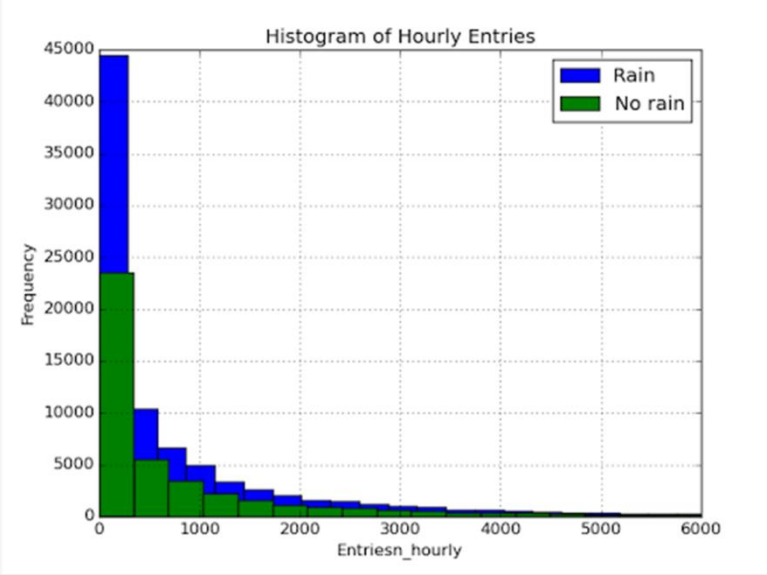

Fig. 1.2: Histogram of Entries/ hourly for rain vs. no rain ridership.

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.

- Average Ridership on a rainy day = 1105.45
- Average Ridership on a non- rainy day = 1090.28
- U= 1924409167.0
- p-value for a 1-tailed test = 0.024999912793489721
- p-value for a 2–tailed test= 0.0499998255869794

(Scipy.stats.mannwhitneyu program was used to analyze the data. The program returned a one-tail p-value. P-value was multiplied by 2 for two-tailed test.)

1.4 What is the significance and interpretation of these results?

Since the p-value is less than p-critical, the null hypothesis can be rejected. Even though the means for the two populations seem similar, we can state that there is a significant difference in ridership on a rainy vs. non rainy day. Rain does have an effect on ridership.

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?

Gradient descent method was used to compute coefficients theta & produce predictions for ENTRIESn_hourly in the regression model.

2.2 What features (input variables) did you use in your model?

The following features were used in the model-

- Rain- Indicator (0 or 1) if rain occurred within the calendar day at the location.
- Precipi- Precipitation in inches at the time and location
- Hour- Hour of the timestamp from TIMEn. Truncated rather than rounded.
- Meantempi- Daily average of tempi for the location
- Meanpressurei- Daily average of pressurei for the location
- Meanwindspdi-Daily average of wspdi for the location.

Did you use any dummy variables as part of your features?

No dummy variables were used as part of features for the model.

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.

Ridership was assumed to be influenced by the following factors-

- Occurrence of rain on that day
- Hour of the day &
- Weather related features such as precipitation, pressure, wind, fog & temperature

Each feature was then added sequentially & R^2 value was monitored for addition of each new feature. Refer to table 2.3 for list of all R^2 values. Upon adding 'meantempi', 'meanpressurei' and 'meanwindspdi' features, no significant increase was observed in the value of R^2 .

Feature	R^2 value
rain, precipi	0.425193646
rain, precipi, Hour	0.463313633
rain, precipi, Hour, meantempi	0.463968815
rain, precipi, Hour, meantempi,meanpressurei	0.464351542
rain, precipi, Hour, meantempi,meanpressurei, meanwindspdi	0.464705994

Table 2.3 Table showing features & corresponding R^2 Value

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

-1.56118306e+00 1.69918111e+00 4.68232699e+02 -4.90397017e+01
-4.22164686e+01 4.97215872e+01 1.49629348e+02

Coefficients	Value
THETA0	-1.5611831E+00
THETArain	1.6991811E+00
THETAprecipi	4.6823270E+02
THETAHour	-4.9039702E+01
THETAmeantempi	-4.2216469E+01
THETAmeanpressurei	4.9721587E+01
THETAmeanwindspdi	1.4962935E+02

Table 2.4 Table showing features & corresponding weights

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

R^2 value extracted was 0.46470599401

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 of 0.46 shows that the model can account for upto 46% of the variability observed in the dataset, using selected features. The R^2 value can certainly be improved upon. For a broad approximate prediction this model is a decent fit. The key is to check for a residuals (difference between expected & observed values) plot. A good residual plot is centered on zero & is normally distributed.

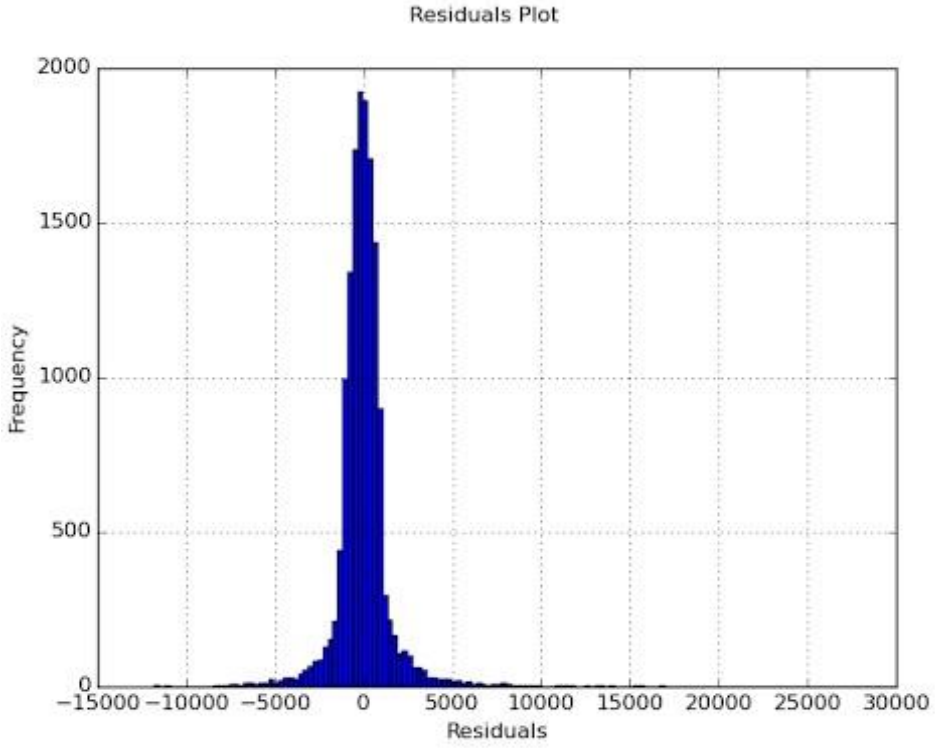

Fig 2.6 Residual histogram plot

Fig.2.6 above shows a histogram plot of all residual values having a normal distribution pattern with a mean of 0 and constant variance. Therefore model equally predicts values that are higher & lower than observed values.

Section 3. Visualization

Please include two visualizations that show the relationships between two or more variables in the NYC subway data.

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

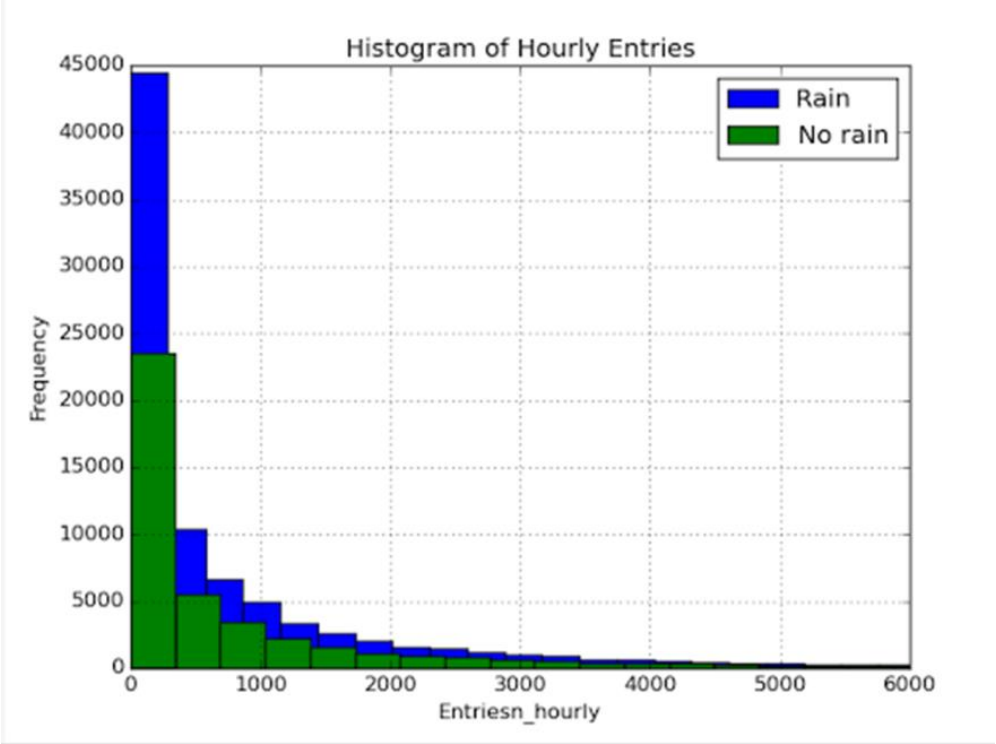

Fig. 3.1: Histogram of Entries/ hourly for rain vs. no rain ridership.

Histogram shows volume of hourly ridership on NYC subway. Plot shows overlay of two histograms- Ridership on rainy days vs. non-rainy days. It can be inferred from the plot that both samples show non- normal distribution. The plot is used to determine the type of statistical test that can be performed on such data sets.

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 day-of-week (Ordered Starting from Monday to Sunday)**

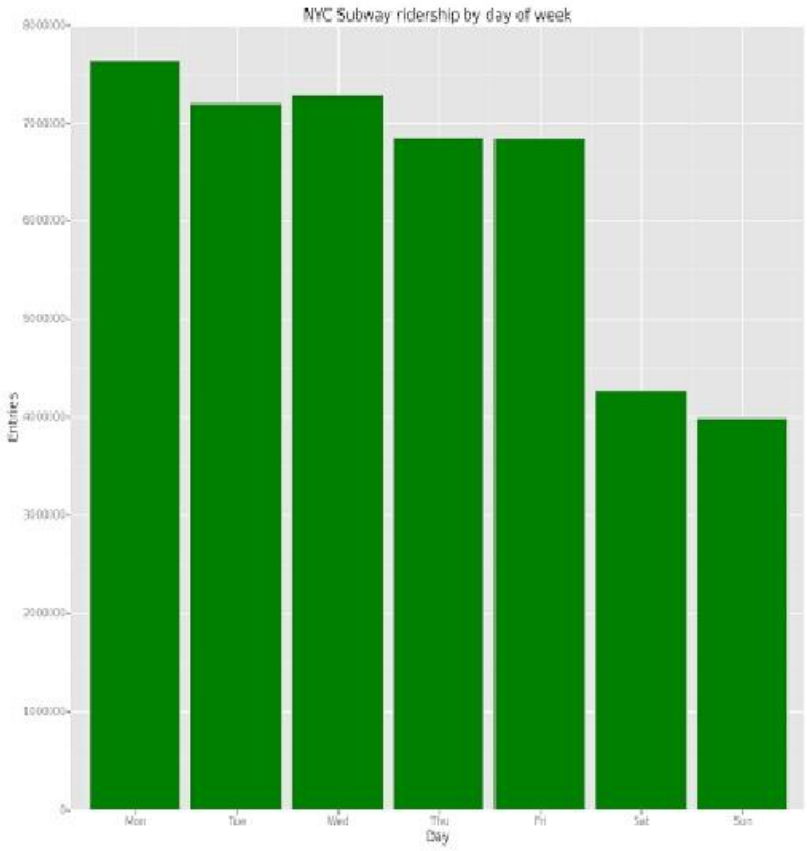

Fig. 3.2: Ridership by day of week

Bar Chart shows aggregate of ridership by day of the week for the month of May 2011. From the plot, the following observations can be made. Ridership is highest on Mondays & slightly decreases over successive week days. There is significant decrease in ridership on the weekend.

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?

From the interpretation of the statistical analyses performed, the conclusion is that people ride the subway more on a rainy day than on a non- rainy day.

Based on the means & Mann-Whitney U test applied on the data-sets show significant difference in ridership on rainy days. The linear regression model having features of Rain, precipitation and the Hour provide a better predictive model for ridership, thereby showing that these features have an effect on ridership.

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.

Means extracted from both sample sizes showed average ridership on a rainy day being more than ridership on a non-rainy day. Based on just the mean, it was insufficient to conclude that more people ride the subway on a rainy day. Sample sizes of each ridership type (rain, no rain) were plotted to choose the type of statistical analysis. Due to the non-normal distribution of the data-set, Mann-Whitney U- test was chosen. The test yielded a significant two tailed p-value of 0.0499998 (< p-critical value of 0.05). This proves that there is a high level of certainty of ridership being different between rainy & non-rainy days. Furthermore, it can be inferred from a one-tailed p-value of 0.02499991, which shows a directionality, that ridership on rainy days is more than on non-rainy days.

Linear regression with gradient descent procedure was implemented to build a model that predicts ridership. A plot of cost_history vs. iteration showed that gradient descent converged to a minimum cost. Coefficient of determination (R^2 value) of 0.46 shows that upto 46% of variability in ridership can be modelled using the selected features. A good model's residual plot should have a normal distribution with a mean of 0 & a finite variance. The residuals were normally distributed, indicating a good model fit. The number of features was adjusted to improve R^2 value.

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:

1. **Dataset**

   Larger sample size for more months in the data-set may aid in extracting statistics that represent actual ridership. Features that may identify different types of riders (for e.g. regular commuters, airport transit, tourists) on the subway may help in building a better model. Larger data set may yield more normally distributed data.

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

   The linear regression analysis applied on the data set does not produce a good predictive model. Adding dummy variables and more features could have enhanced the model. As explained in the lectures, data can be split into training & test set. Doing a cross-validation of the model (obtained from training set) on the test set can help refine the model better.