Teja Tummalapalli

Project 2:

Section 0 : References

Books:
Python for data analysis , ISBN-13: 978-1449319793
Statistics in a Nutshell ,ISBN-13: 978-1449316822

Websites:

Plotting

http://matplotlib.org/users/legend_guide.html
http://blog.yhathq.com/posts/ggplot-for-python.html

Regression

http://docs.statwing.com/interpreting-residual-plots-to-improve-your-regression
http://www.originlab.com/doc/Origin-Help/Residual-Plot-Analysis
https://www.khanacademy.org/math/probability/regression/regression-correlation/v/r-squared-or-coefficient-of-determination
http://stats.stackexchange.com/questions/164542/is-time-in-linear-regression-a-categorical-or-continuous-variable

Pandas

http://pandas.pydata.org/pandas-docs/version/0.15.2/groupby.html
http://pandas.pydata.org/pandas-docs/stable/indexing.html

Section 1 : Statistical Test

Questions

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?
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.
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.
1.4 What is the significance and interpretation of these results?

Answers

1.1 I have used Mann-Whitney U test to analyze the NYC subway data with two-tailed P-Value.

Null Hypothesis

H_0: The probability of an observation from the population X(With rain) exceeding an observation from the second population *Y* equals the probability of an observation from *Y(Without Rain)* exceeding an observation from *X* : P(*X*>*Y*)=P(*Y*>*X*)

P-Critical Value : 0.05

1.2 The subway data is not normally distributed for ENTRIESn_hourly.As the parametric test like Welch T-Test is not suitable for non-normal distributions ,Mann-Whitney U test is used.

1.3
U-Statistic : 1924409147
P-value : 2 * 0.0249 = 0.049
Mean of sample with rain : 1105
Mean of sample without rain : 1090

1.4 P-Value is less than the P-Critical values, so the null hypothesis is rejected. ie there is a difference in distribution of two samples(with and without rain)

Section 2 : Linear Regression

Questions

2.1 What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn_hourly in your regression model:
2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?
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.
2.4 What are the parameters (also known as "coefficients" or "weights") of the non-dummy features in your linear regression model?
2.5 What is your model's R^2 (coefficients of determination) value?
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?

Answers

2.1 OLS using Statsmod

2.2 Rain,meantempi,day_week,Hour,Unit and conds are used as independent variables in regression. day_week,Hour,Unit,conds are dummy coded.

2.3

Rain : I used the feature rain because, people may tend to used subway less/more when it is raining.
meantempi : People may prefert to use subway less/more based on the temperature.
Day_week: People generally use subway more on weekdays for going to work.
Hour: I used hour because,people may use the subway more in peak hours when compare to normal hours
UNIT: I have used UNIT because,there may be some units which are busier than the others. As an example,the UNIT on a busiest route may have more ridership
conds: The usage of subway may depend up on the different weather conditions.

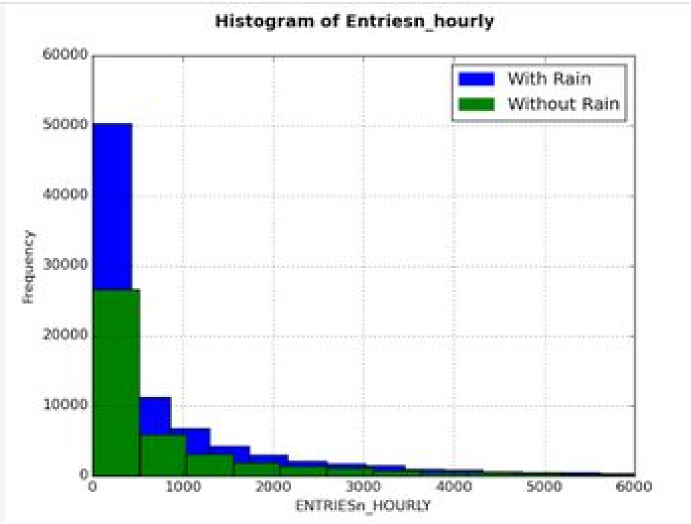
2.4
Params of rain is 11.3
Params of meantempi is -20.5

2.5
The R-Square value is 0.547

2.6
R-Square is used to measure how close the data can be fitted to regression line. It is calculated by taking the the percentage of the response variable variation that is explained by a linear model.
The R-square values is greater than 0.4. So the regression model can be used to predict the ridership.

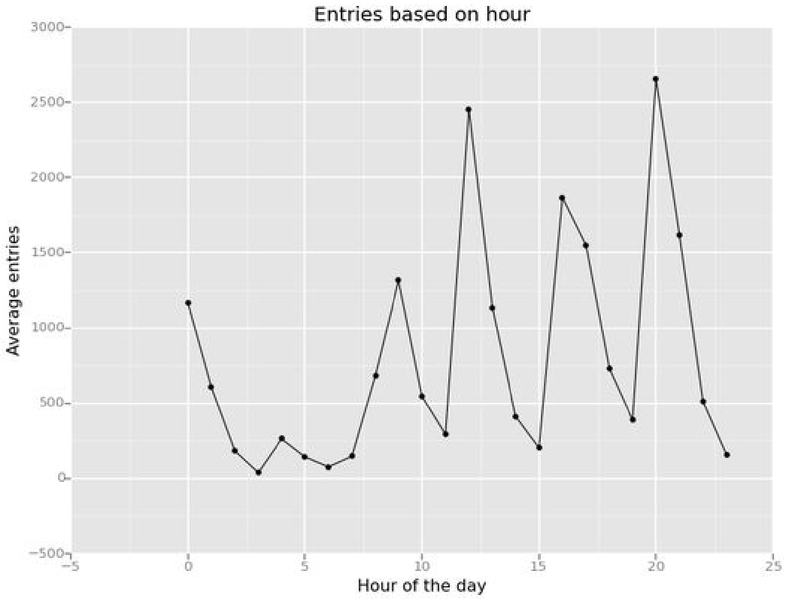
Section 3 : Visualization

3.1



The histogram is used to plot the distribution of ENTRIESn_HOURLY for the sample with and without rain. The plot is not bell curve shaped. We can infer that bot the samples are not normally distributed.

3.2



In this line graph,X-Axis depicts the hour of the day and Y-Axis depicts the average number of entries. Each point represent a hour.From the graph we can infer that there is heavy ridership at 12PM and 8PM. We can also get the overview of ridership based on the time of the day.

Section 4: Conclusion

Questions

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.

Answers

4.1. More people ride the NYC subway when it is raining when compared to not raining.

4.2

- Mann-Whitney U test p value is less than p-critical value. This suggests that there is the difference of distribution of data between rain and no rain. As the test is two sided, we need to explore more stats to find out which sample is having higher ridership. The means of both the sample are calculated and the mean with rain is higher than the mean. This gives an hint.
- Linear regression can be used for further analysis. The co-efficient of Rain variable is 11.3, which implies the ridership increases with rain.

Based on both Mann-Whitney,linear regression we can conclude that the ridership is more when it is raining.

Section 5 : Reflections

5.1

- Dataset of the regression contains data for only one month, the prediction/analysis can be accurate if the data is more.
- While performing regression multicollinearity can be checked to avoid wrong coefficients.
- Hours is dummy coded in the regression, but to preserve the continuity of the time a periodic function can be used.
- Machine learning algorithms like ‘select k best’ can be used to select the important features for regression.