Analyzing the NYC Subway Dataset

Section 0. References

References used are as below:

1. http://ggplot.yhathq.com/docs/geom_histogram.html
2. http://www.graphpad.com/guides/prism/6/statistics/index.htm?how_the_mann-whitney_test_works.htm
3. http://spin.atomicobject.com/2014/06/24/gradient-descent-linear-regression/
4. http://pandas.pydata.org/pandas-docs/stable/basics.html

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?

Mann-Whitney U-statistic was used to analyse the NYC subway data.
Used a two tail P Value.
Null Hypothesis The mean of the two samples on an rainy day and on a non-rainy days are the same.
Alternate hypothesis : The mean of the ridership on an non-rainy day is different compared to the rainy day.
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.

This statistical test is applicable to the dataset since the distribution is non-parametric or the dataset is not from any probability distribution .

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.

P-Value =0.0249 for one-tail and 0.0499 for two tailed P-value. Mean riders on an rainy day =1105 ,and mean riders on a non-rainy day=1090. U Value = 1924409167.0

1.4 What is the significance and interpretation of these results?

The ridership on a rainy or a non-rainy days are not significantly different.The difference between the two groups are not different

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?

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?

Minimum Temperature, Precipitation, time or Hour and Fog are the feature variables to predict the entries hourly.

Yes dummy variables was part of the features. The dummy variables were the subway route # OR the route#.The dummy variables stands in for the qualitative variable which will effect the model.The route number in this example acts as the dummy variable is not quantified ,but by making it numeric value it will have the influence on the outcome.This dummy variable is an independent variable has no role in influencing the input 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.

- Your reasons might be based on intuition. For example, response for fog might be: "I decided to use fog because I thought that when it is very foggy outside people might decide to use the subway more often."
- Your reasons might also be based on data exploration and experimentation, for example: "I used feature X because as soon as I included it in my model, it drastically improved my R^2 value."

The minimum temperature = When the temperature is low people may use the subway.

The precipitation = When there is more precipitation,people will use the subway more .

Time or Hour = This is a important factor as the during the morning and during the evening the number of riders will be high.

Fog= When the visibility is low,people will use the subway.

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

-1.25457098e+01 4.68023897e+02 6.83767042e+01 1.50166400e+02

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

R^2 = 0.46490084168

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?

Higher the value of the R^2 ,more accurate is the value of the predicted value to the actual value.

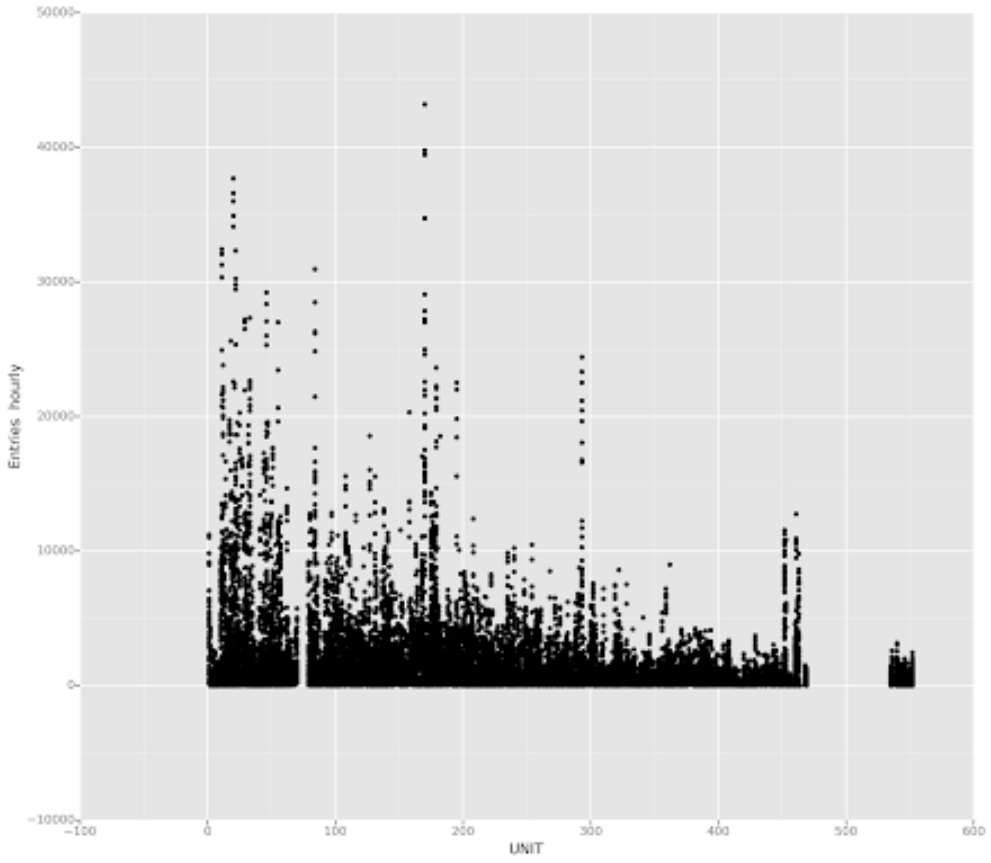
Yes this model is appropriate as we are predicting the linear model.This number gives the approximation on how well the predicted value fits the observed value.

This number is the ratio between the explained variation to total variation.Given the fact that the features used are unbiased this is a good model for the prediction of the outcome.

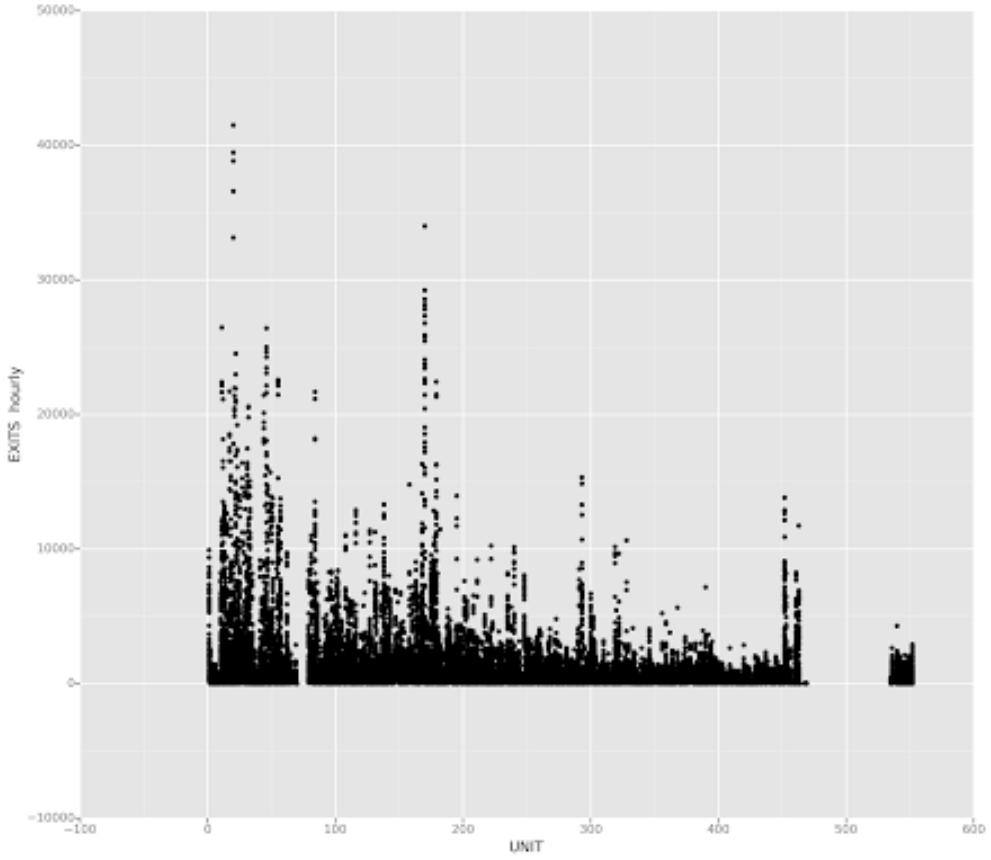
Section 3. Visualization

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

Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.



The above figure shows the relationship between the entries hourly and the subway unit number.The X- axis is the subway unit number.The Y-axis shows the entries hourly for that unit.

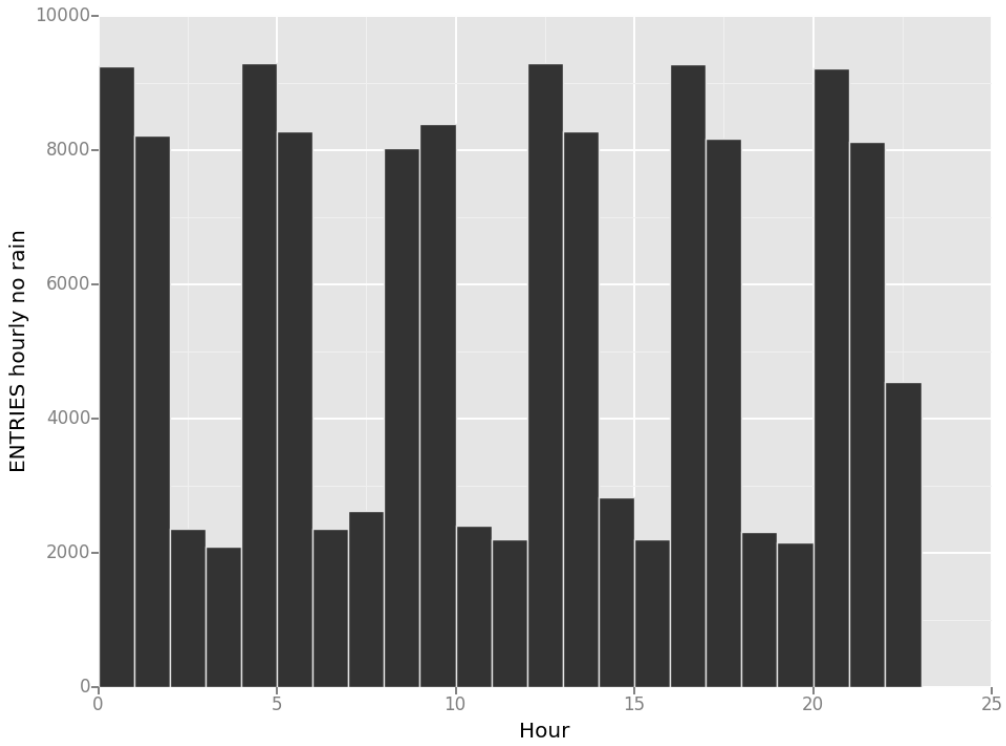


The above figure shows the relationship between the exits hourly and the subway unit number.The X- axis is the subway unit number.The Y-axis shows the exits hourly for that unit.

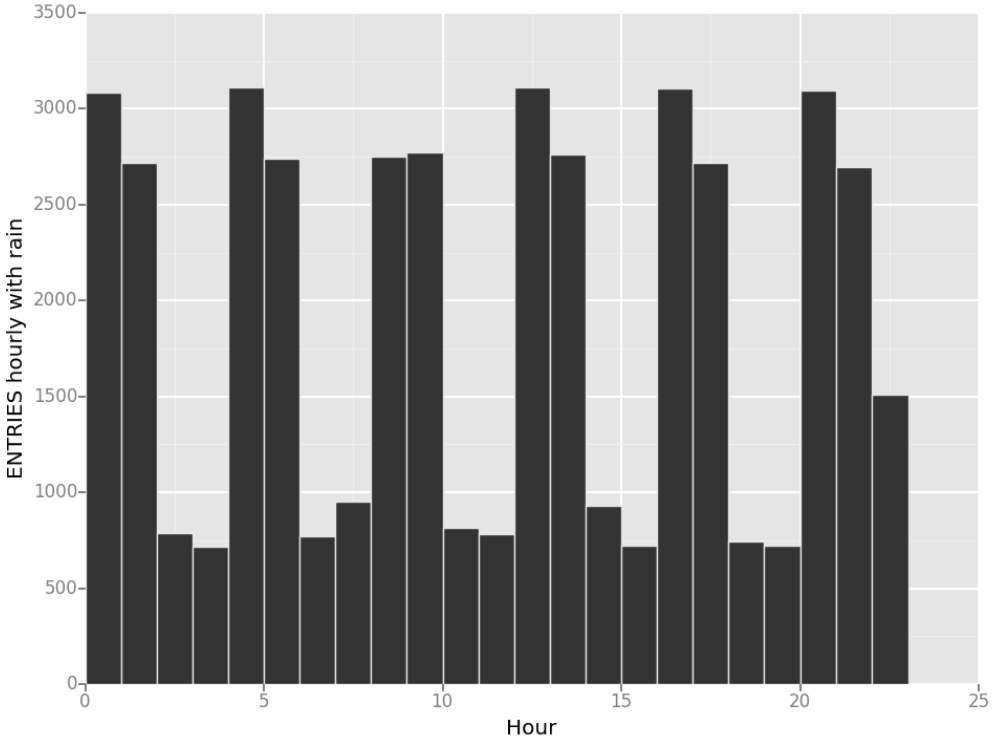
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

- You can combine the two histograms in a single plot or you can use two separate plots.
- If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.
- For the histograms, you should have intervals representing the volume of ridership (value of ENTRIESn_hourly) on the x-axis and the frequency of occurrence on the y-axis. For example, each interval (along the x-axis), the height of the bar for this interval will represent the number of records (rows in our data) that have ENTRIESn_hourly that falls in this interval.

- Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.



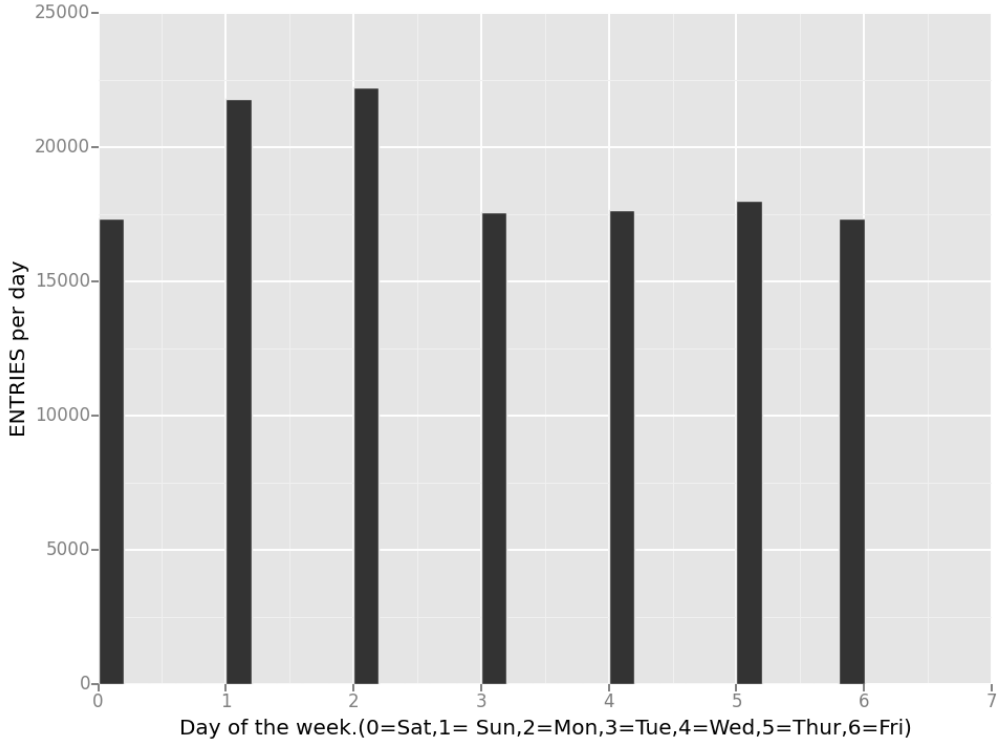
The above bar plot shows the number of hourly entries in subway when it is not raining.



The above bar plot shows the relationship of the hourly number of entries in subway when it is raining.

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 time-of-day
- Ridership by day-of-week



The above relationship shows the ridership for day of the week.The x axis shows the day of the week and the Y axis shows the entries per day.

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 data visualization it looks as if the people ride more when it is not raining and the result is statistically significant to reject the null hypothesis when we use the statistical tools to conclude the result.

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.

The reason for this conclusion is that the when we are using the Mann Whitney test and check for the P-value for two-tailed test turns out to be 0.0498 < 0.05 hence we reject the null hypothesis and accept the alternate hypothesis that the ridership in subway is different between rainy and non-rainy days.The P-value is very marginally less than the P-critical value,but it is statistically significant to reject the null hypothesis.The P-Value indicates the assumption of the ridership between

rainy and non-rainy days is same is not true at 95% Confidence Interval and hence we go with the alternate hypothesis.

In the linear regression we use the various input factor which influence the output to predict the output which is the number of entries in the subway rider. Among the variour factors used the linear regression analysis shows that when using the rain factor to predict the outcome of the number of entries in the subway the outcome didn't alter the R-factor much.R-factor is used to measure the actual output versus the predicted output.This gives the impression that the subway ridership between the rainy and non-rainy days is marginally different.

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,
   The dataset used for the analysis makes an difference in the prediction of the output result.The number of datapoints could be less to make an impact in the prediction of the output.For eg when it is raining if it can be quantified on how much inches it rained then this could play an factor in improving the R-factor.
2. Analysis, such as the linear regression model or statistical test.
   In the linear regression model the following factors influence the results
   a. Outliers in the data points.The outliers in the feature inputs can effect the output significantly. These outliers can be from the environment factors which is external.For eg Hurricanes.
   b. When doing the linear regression the assumption is that feature variables considered should be independent of each other.However in the data what we get this may not hold true for all the feature inputs.For eg the precipitation and rain are dependent of each other.