QUESTIONS FOR ANALYZING THE NYC SUBWAY DATASET

Section 0. References

1. Mann–Whitney U test, Wikipedia page: http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test
2. scipy.stats.mannwhiteyu (http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html#scipy.stats.mannwhitneyu)
3. One-tail vs. two-tail P values (http://graphpad.com/guides/prism/6/statistics/index.htm?one-tail_vs__two-tail_p_values.htm)

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?

For analysis of means for ridership ('ENTRIESn_hourly' data) on Rainy vs. Nonrainy days, the Mann Whitney U-test was used. The two-tail P-values was used due to the following null hypothesis: The means of 'ENTRIESn_hourly' data values are the same on rainy and nonrainy days.

p-critical value (U) is equal to 1924409167.0; P =0.0193 (< 0.05 two-tailed)

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.

U-test was chosen due to the following considerations:

- it is nonparametric and more suitable for non-normal distribution of 'ENTRIESn_hourly' data (Welch's T-test was not suitable, because it proposes that two samples follow normal distribution)
- it's null hypothesis assume that two samples are from the same populations and follow the same 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.

with_rain_mean = 1105.4463767458733
without_rain_mean = 1090.278780151855
U = 1924409167.0
p = 0.0193 (< 0.05 two-tailed)

1.4 What is the significance and interpretation of these results?

p = 0.0193 (< 0.05 two-tailed) is significant and can be interpreted that means of two samples are statistically 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:

I've used Gradient descent method 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?

Features list: ['rain', 'meanwindspdi', 'Hour', 'maxdewpti']

Also, the UNIT was added to features using dummy variables

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 decided to use 'rain' because I thought that when it is very rainy outside people might decide to use the subway more often.

I decided to use 'meanwindspdi' because I thought that when it is very windy outside people might decide to use the subway more often.

I decided to use 'Hour' because I thought people might use the subway more often at specific hours: in the morning they go to work, afternoon the might go to some business meetings or events, and in the evening they go back their home.

The 'maxdewpti' feature was used just because of curiosity.

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

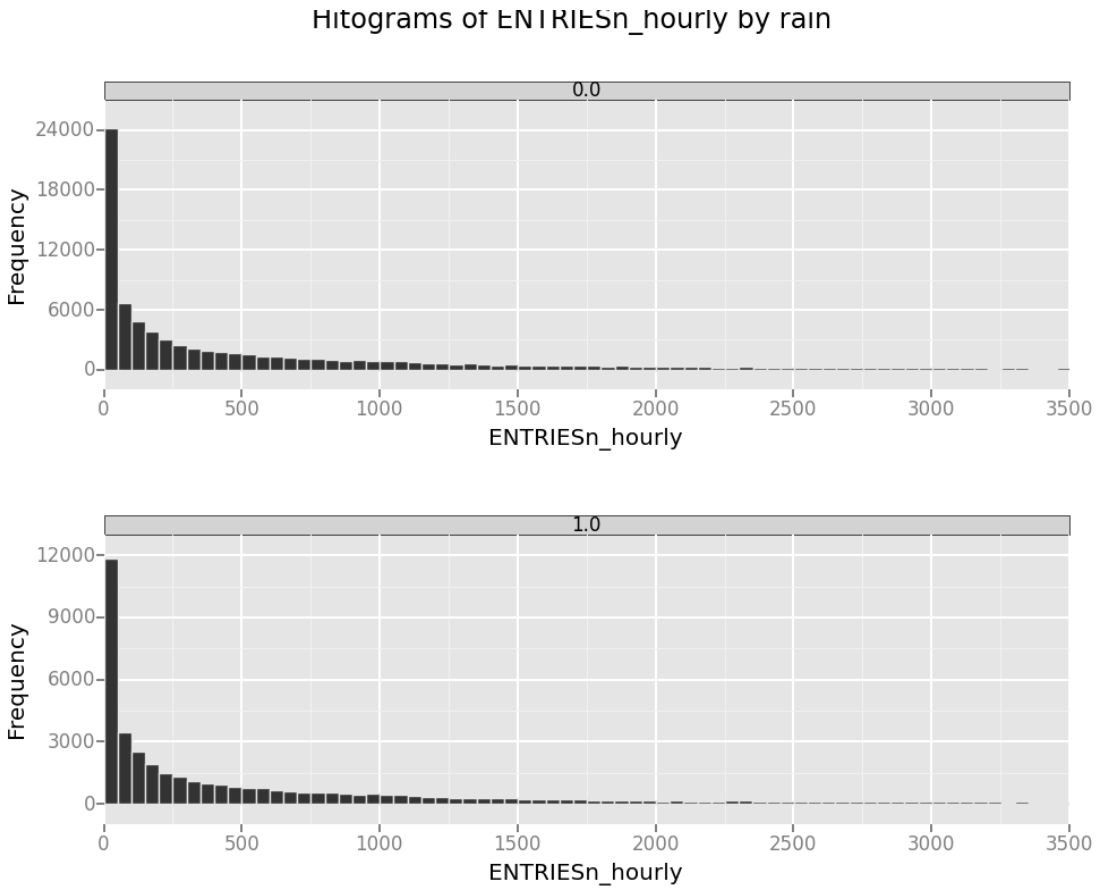
```
R^2 value is 0.464195606053
```

2.5 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?

This R^2 indicate that the model explains about 46 % of variability of the ridership data around its mean. This linear model will produce weak predictions for this dataset, given this R^2 value.

Section 3. Visualization

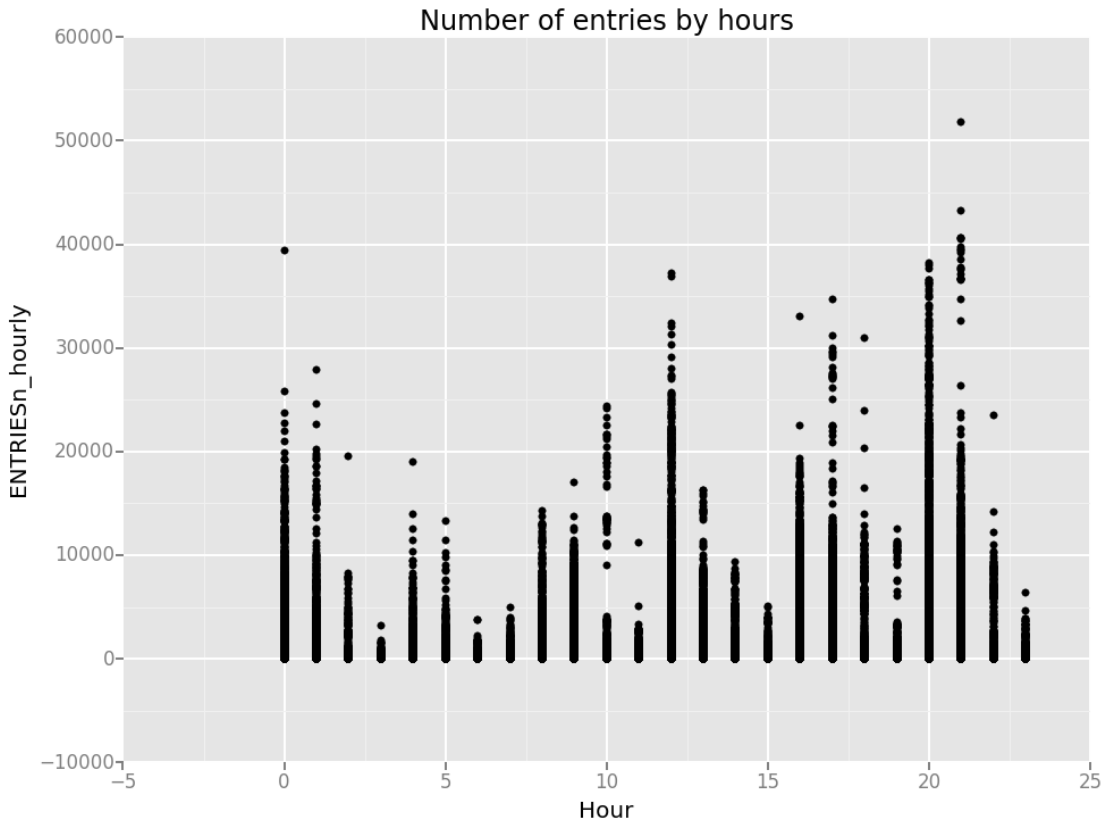
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.



The histograms above show the volume of ridership (value of ENTRIESn_hourly) on the x-axis and the frequency of occurrence on rainy and non-rainy days.

Frequency of using subway during the rainy days almost twice lower than in non-rainy days.

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:



The diagram about shows that number of entries depends hourly. The most frequently people use subway around the noon (12 pm) and between 7 and 22 pm.

Section 4. Conclusion

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?

More people ride the NYC subway when it is not raining. The Mann Whitney U-test of means between two datasets shows statistically significant difference (< 0.05). The difference is also shown on "Histogram of ENTRIESn_hourly by rain" above.

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 Mann Whitney U-test of means between two datasets shows statistically significant difference (< 0.05). Calculated statistics :

- with_rain_mean = 1105.4463767458733
- without_rain_mean = 1090.278780151855
- U = 1924409167.0
- p = 0.0193 (< 0.05 two-tailed)

R^2 `value is 0.464195606053 ,` indicates that the model explains about 46 % of variability of the rider ship data around its mean.

Section 5. Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis, including:

1. Dataset:
    - Have no idea about shortcomings regarding the dataset
2. Analysis, such as the linear regression model or statistical test.
    - The build linear regression model has weak prediction power. The possible problem is due to assumed linearity of relationships it tries to define and predict.

5.2 (Optional) Do you have any other insight about the dataset that you would like to share with us?

Have no insights about the dataset.