Section 0. References

1.Mann–Whitney *U* test:

http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test

2. Coefficient of determination:

http://en.wikipedia.org/wiki/Coefficient_of_determination

3. Mann–Whitney *U* test and null hypothesis

http://www.graphpad.com/guides/prism/6/statistics/index.htm?how_the_mann-whitney_test_works.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?

ANWSER:

I used Mann-Whitney U-test to analyze the NYC subway data and used a one-tail P-value.

The null hypothesis is that more people will take the subway on rainy days than on not rain days.

The P-critical value is 5%.

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.

ANWSER:

The Mann-Whitney test is a nonparametric test that compares two unpaired groups. As the data size of rainy days and not rain days are different, Mann-Whitney U –test is a good choice.

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.

ANWSER:

The P-value for this U-test is 0.025. Mean of entries on rainy days is 1105.45, mean of entries on not rain days is 1090.28.

1.4 What is the significance and interpretation of these results?

ANWSER:

The P-value is less than the P-critical value, so I should reject this null hypothesis.

Mean value of entries on rainy days is slightly larger than the mean value of entries on not rain 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:

1. OLS using Statsmodels or Scikit Learn
2. Gradient descent using Scikit Learn
3. Or something different?

ANWSER:

I used OLS 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?

ANWSER:

Features used in this model are: 'rain', 'precipi', 'Hour', 'meantempi'. And I added UNIT 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.

ANWSER:

I decided to use rain because I thought when it is rain, more people would like to take the subway instead of walk to their destination.

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

ANWSER:

R-square

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

ANWSER:

The R-square value is 0. 47924770782.

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?

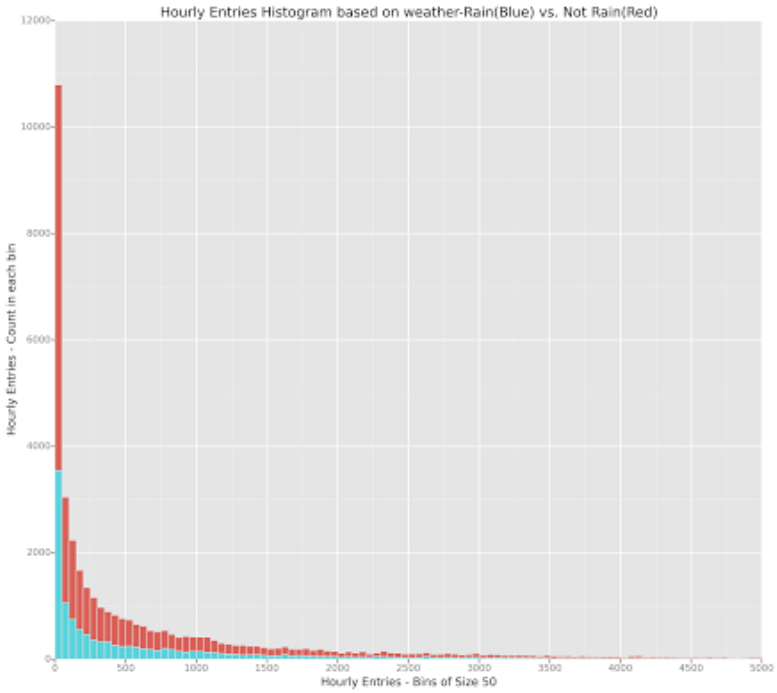
ANWSER:

R^2 is a statistic used in the context of statistical models whose main purpose is either the prediction of future outcomes or the testing of hypotheses, on the basis of other related information. It provides a measure of how well observed outcomes are replicated by the model, as the proportion of total variation of outcomes explained by the model
It is a fraction between 0.0 and 1.0, and has no units. Higher values indicate that the model fits the data better. As the R^2 value is 0. 47924770782, I think this linear model to predict ridership is appropriate for this dataset.

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.

ANWSER:

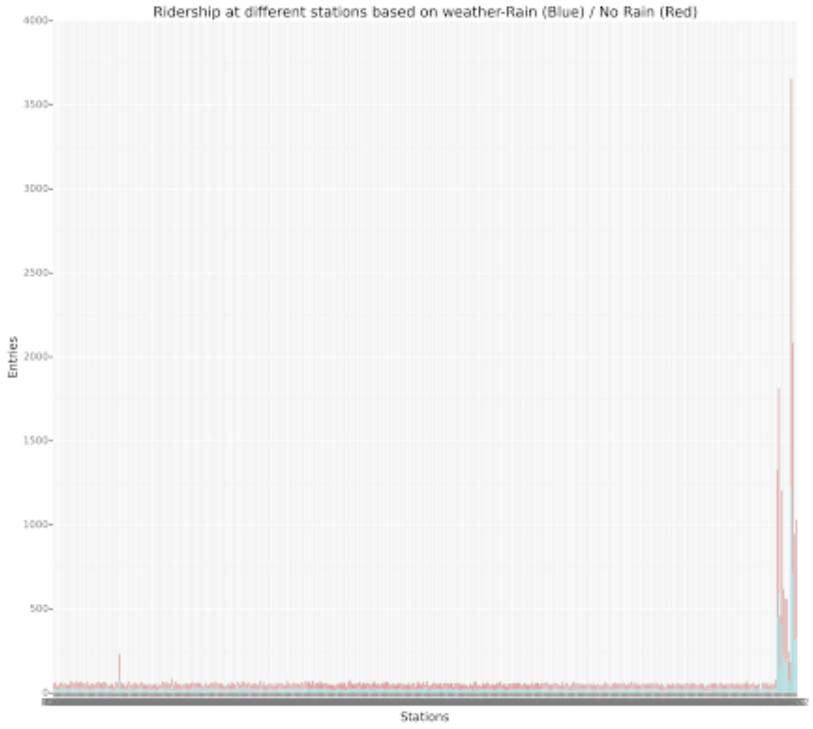


For this plot, x-axis represents the "Hourly Entries - Bins of Size 50" and y-axis represents the "Hourly Entries - Count in each bin" (the frequency of occurrence). The height of the bar for this interval will represent the number of records (rows in our data) that have ENTRIESn_hourly that falls in this interval. So we can see that, when it is not rain, more people take the subway.

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

ANWSER:



For this plot, x-axis represents different stations and y-axis represents the entries at each station. The red bars represent the data of not rain days and the blue bars represent the ridership on rainy days. So we can see that, more people take the subway on not rain days.

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?

ANWSER:

From the analysis above we can see that, more people will ride the NYC subway on not rain days.

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.

ANWSER:

From the U-test we can see that, the P-value we got is 0.025, is less than the P-critical value 0.05. So we can reject the null hypothesis (more people will like to ride the NYC subway on rainy days), which means more people will take the NYC subway on not rain days. Also, form the visualization figures we can see that more people ride the NYC subway on not rain days, the same as the result from the statistical test.

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

ANWSER:

First of all, the dataset is small. It only includes the data of one month, it may not represent the real situations. Second, for the statistical test, the null hypothesis of U-test is that the distributions of both groups are identical, so that there is a 50% probability that an observation from a value randomly selected from one population exceeds an observation randomly selected from the other population. But we are talking about the overall observation over a period of time, not a single observation value. So this U-test may not be the best test for this project.