SHORT QUESTIONS

Section 1. Statistical Test

1.1 Which statistical test did you use to analyse 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?

I used Mann Whitney U-test to analyse the NYC subway data;

tow-tail;

Null hypothesis: The distributions of both groups are equal under the null hypothesis

Two-sided 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.

In this case, we want to know whether rain or no-rain factor affects the ENTRIESn_hourly data. According to histogram of Question 3.1, the samples data does not seem normally distributed, and it can not be assumed to be drawn from any particular probability distribution. Therefore Mann–Whitney U test, the nonparametric test, is applicable to the dataset in this case.

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.

one-sided P-value from the statistical test: 0.024999

With_rain_mean: 1105.4464

Without_rain_mean: 1090.2788

1.4 What is the significance and interpretation of these results?

Cut-off point of two-sided P-value is set at 0.05. The calculated one-sided P-value in the distribution is 0.0249999, and the two-sided P-value will be 0.0249999*2=0.049999, which is smaller than the P critical value. Therefore the distribution of two samples is 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:

√ a. Gradient descent (as implemented in exercise 3.5)
b. OLS using Statsmodels
c. Or something different?

2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?

I used five features: 'rain', 'precipi', 'Hour', 'meantempi', 'fog'.

Yes. UNIT.

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.”

Firstly I chose features based on my intuition. I thought that when it is foggy or rainy or stormy outside, people might use the subway more often. I also maybe different temperature, precipitation, pressure and different time of a day might affect ridership. Therefore I took 'rain', 'precipi', 'Hour', 'meantempi', 'fog', ‘thunder’ and ‘meanpressurei’, those seven features into consideration.

Then I try different combination of the those features to apply to the test. Base on the experimentation, I finally decided to use the five features, 'rain', 'precipi', 'Hour', 'meantempi', 'fog', which can drastically improved my R^2 value to 0.462.

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

'rain': 25.60;
'precipi': 34.92;
'Hour': 421.14;
'meantempi': -57.88;
'fog': 52.90

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

0.4623

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?

In regression, the R2 coefficient of determination is a statistical measure of how well the regression line approximates the real data points. R-square is 1 minus the ratio of residual variability. When the variability of the residual values around the regression line relative to the overall variability is small, the predictions from the regression equation are good. The value of R^2 shows the proportion of the variance in the test data that is explained by the model.
R-squared = Explained variation / Total variation

In most cases, R-squared is between 0 and 100%:

- 0% indicates that the model explains none of the variability of the response data around its mean.
- 100% indicates that the model explains all the variability of the response data around its mean.

R2 is close to but less than 0.5, so I think this linear model can be used to predict ridership, but not good enough.

Section 3. Visualization

Please include two visualizations that show the relationships between two or more variables in the NYC subway data. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots, or histograms) or attempt to implement something more advanced if you'd like.

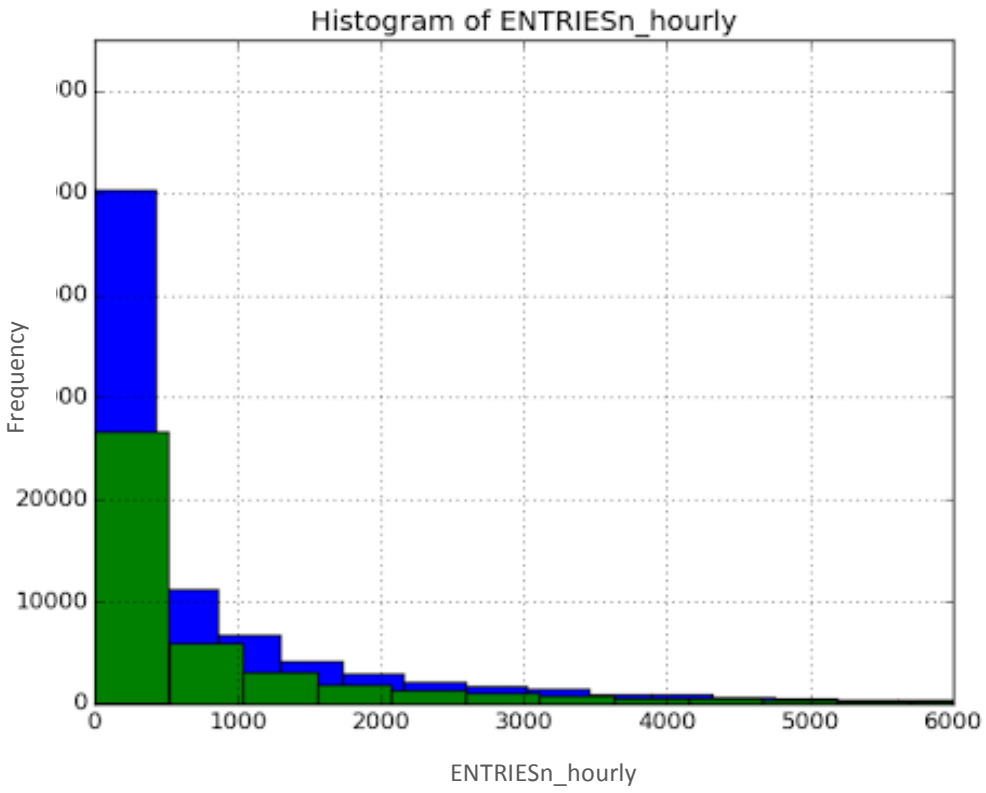
Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

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.

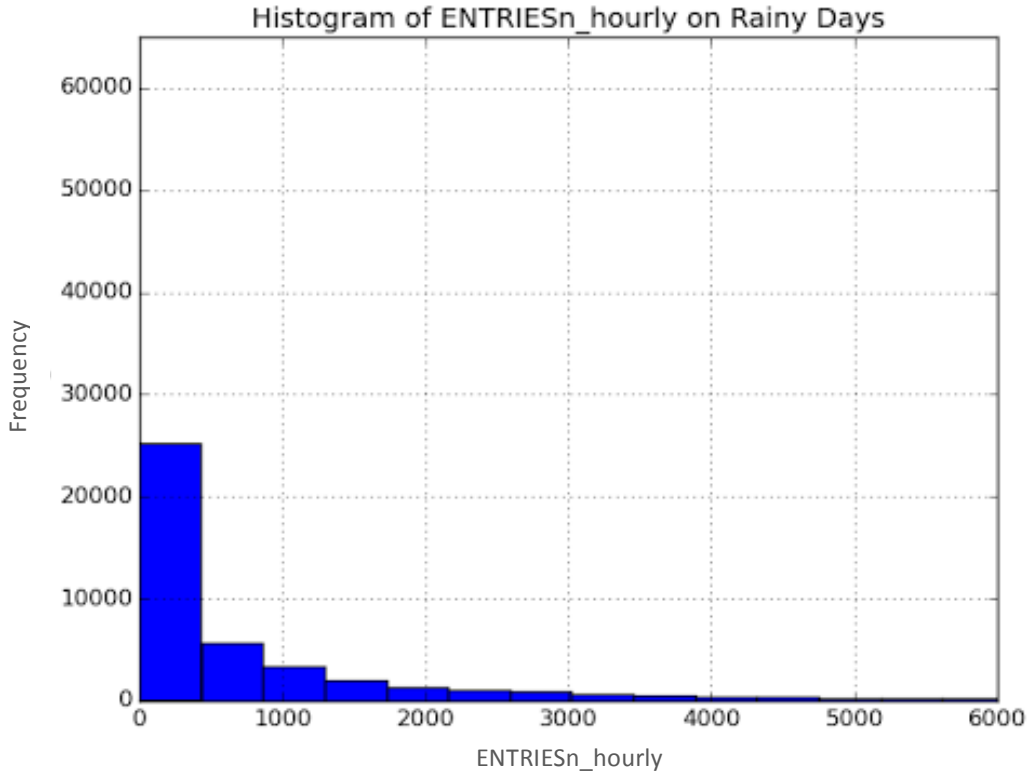


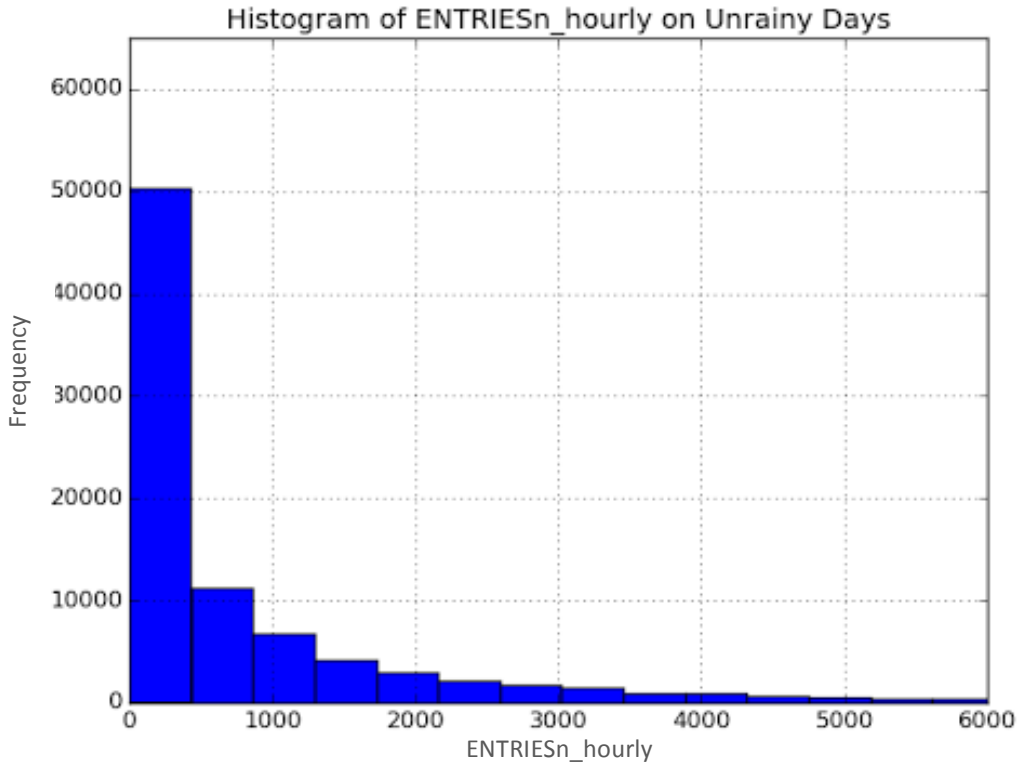
Comment:

Blue: Histogram of ENTRIESn_hourly in unrainy days

Green: Histogram of ENTRIESn_hourly in rainy days

Separated plots:



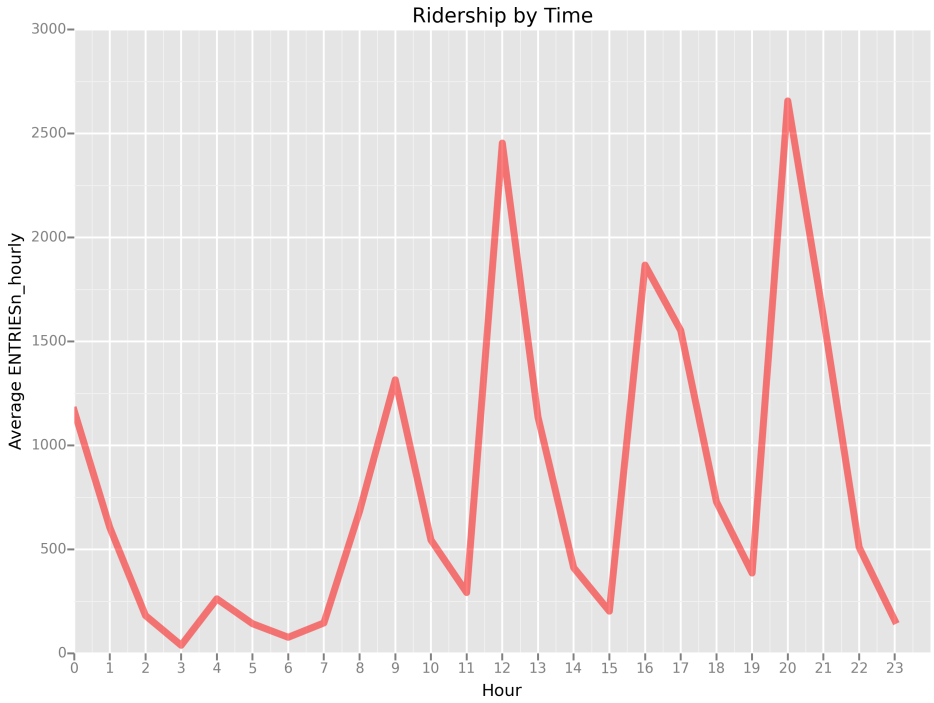


3.2 One visualization can be more freeform. Some suggestions are:

Ridership by time-of-day

Ridership by day-of-week

This is the plot of Ridership by time-of-day. X-axis is the hour of one day; y-axis is the mean of Entriesn_hourly value in different units for every particular time.



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?

More people ride the NYC subway when it is 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.

(1) From Linear Regression

In linear regression, I used the method of gradient descent to compute the coefficient theta and produce prediction for ENTRIESn_hourly in my regression model, in which 'rain' is one of the important features. The

computed coefficient of 'rain' in the model is 2.56002622e+01, which is positive, suggesting more people ride the NYC subway when it is rainy.

(2) From Statistical Tests

According to my test results, the average ridership in rainy days is about 1105.46, while in unrainy days is 1090.28, which is smaller than rainy days. As the calculated p value of my test is smaller than the critical p value, the distribution of two samples in rainy days and unrainy days is statistically different. Therefore I can tell more people ride the NYC subway when it is rainy.

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:

- Dataset,
  (1) The dataset may be lack of meaningful features that have significant effect on ridership.
  (2) According to the note of the course in Udacity, only a subset (~10%) of the data is used in every run in order to make sure the experiment can be finished by us in acceptable time. Using more data should bring better results.
  (3) Some subway units do not have ridership at any time. Maybe the data of those units is missing or those units are abandoned. In either case those units may bring noise, and the dataset including them may not be identically distributed. Therefore the test result base on the dataset is not reliable.
- Analysis, such as the linear regression model or statistical test.
  (1) For the training process, in this case, the training set is also used as the test set, which may cause over fitting problem. Good performance on training set does not mean good result for the prediction of new data. Maybe we should use 90% of entire data for training and 10% for test. Or we can use cross-validation to do the linear regression.

(2) For the model, hypothesis of Linear regression maybe too simple for the dataset. It is possible that the relationship between ridership and features are not simply linear. Therefore R square is not very high.

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

Bug report

in Lesson 5: MapReduce Quiz No.2. Here two bugs in the code in your answer:

(1) key = i.translate(string.maketrans("","", string.punctuation).lower()

Correction: “)” is missed. This line should be:

key = i.translate(string.maketrans("","")**)**, string.punctuation).lower()

(2) if key in word_counts.key():

Correction: “.keys” instead of “.key”. This line should be:

if key in word_counts.key**s**():