Analyzing the NYC Subway Dataset

Questions

Xinqi You

May 18, 2015

Section 0. References

[1] http://www.dotnetperls.com/dictionary-python

[2] http://stackoverflow.com/questions/1801668/convert-a-python-list-with-strings-all-to-lowercase-or-uppercase

[3] http://stackoverflow.com/questions/16050952/how-to-remove-all-the-punctuation-in-a-string-python

[4] http://ggplot.yhathq.com/docs/geom_histogram.html

[5] http://stackoverflow.com/questions/19482970/python-get-list-from-pandas-dataframe-column-headers

[6] https://docs.python.org/2/library/datetime.html#strftime-strptime-behavior

[7] http://www.tutorialspoint.com/python/string_join.htm

[8] http://stackoverflow.com/questions/12555323/adding-new-column-to-existing-dataframe-in-python-pandas

[9] http://www.tutorialspoint.com/python/python_basic_operators.htm

[10] https://github.com/upjohnc/udacity

[11] https://github.com/lanttern/Data_Aanlyst_Nanodegree_projects/blob/master/Project1_Analyzing_the_NYC_Subway_Dataset/problem_sets2to5/problem_set4_visualizing_sunway_data/4_1_visualization_1_ridership_by_unit.py

Section 1. Statistical Test

1.1

I use Mann-Whitney U-Test and one-sided P value. The null hypothesis is that two samples (with rain and without rain) come from the same population. My critical p-value is 0.025.

1.2

Our data is not normally distributed so we cannot use t-test, which assumes normal distribution. U-Test doesn't assume the distribution of the data.

1.3

$\mu_{rain} = 1105.45, \mu_{notrain} = 1090.28, p - value = 0.025$

1.4

If we use significance level $\alpha = 0.05$, our p-value < 0.05. Therefore we reject null hypothesis and conclude that the entries on rainy and non-rainy days come from different populations.

Section 2. Linear Regression

2.1

I use both gradient descent and OLS to compute the coefficients theta.

2.2

In the gradient descent model, I select rain, precipi, Hour, meantempi and UNIT as dummy variable.

In the OLS model, I select 'EXITSn_hourly', 'Hour', 'maxpressurei', 'maxdewpti', 'mindewpti', 'minpressurei', 'meandewpti', 'meanpressurei', 'fog', 'rain', 'meanwindspdi', 'mintempi','meantempi', 'maxtempi', 'precipi', 'thunder'.

2.3

In the gradient descent model, I choose these features based on intuition. The entries have large correlation with the weather so I pick several representative variables to indicate the weather condition.

In the OLS model, I just select as many variables as possible to increase R-squared.

2.4

weights = [-9.09928037, 18.31999979, 464.12341413, -47.77325488]

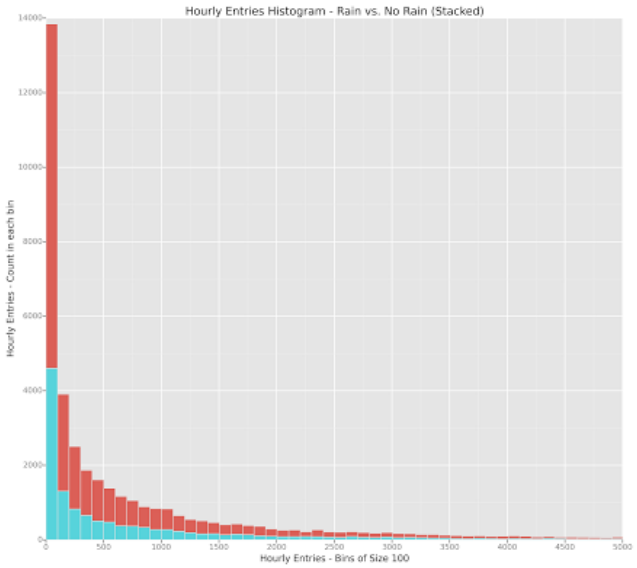
2.5 $R^2 = 0.4640$

2.6

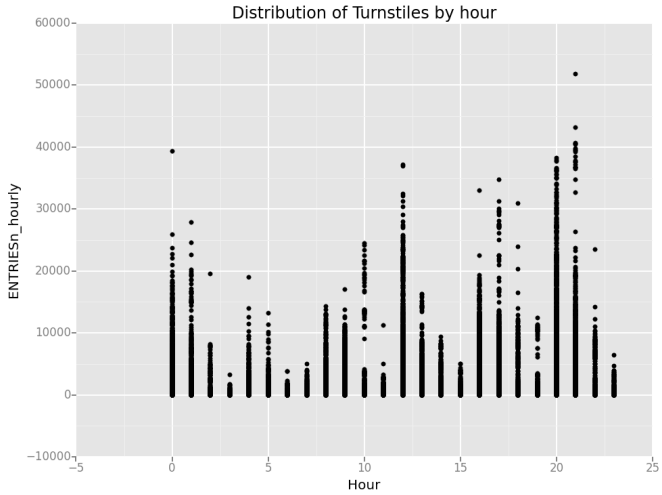
R-squared means how much variance is described in the linear regression model. The R-squared is rather low therefore it's not a good model to predict ridership.

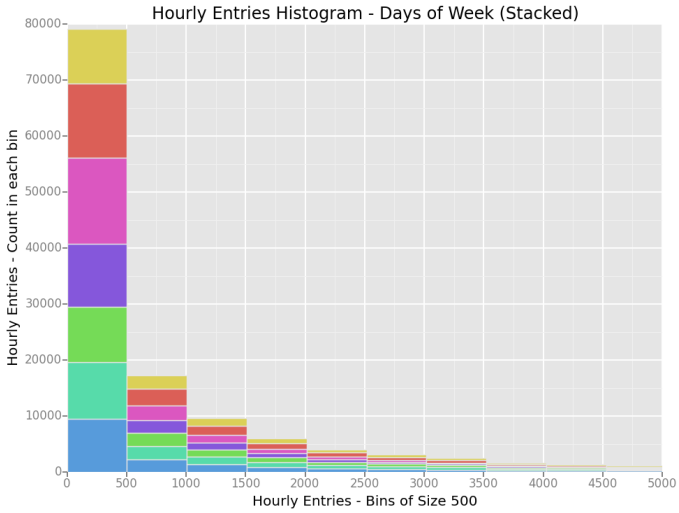
Section 3. Visualization

3.1



3.2





Section 4. Conclusion

4.1

Much fewer people ride the NYC subway when it is raining. From the graph obtained in section 3.1, we can see that on non-rainy days the number of entries almost doubled. The difference is so significant that we cannot ignore.

4.2

Also from our Mann-Whitney U test result, the means of rainy and non-rainy samples are statistically different. The coefficient of rain is -9.09928037 in our gradient descent model indicates that when rain=1, prediction of hourly entries will

be 9.09928037 smaller than the hourly entries when rain =0. In other words, rainy days have smaller ridership than non-rainy days in general.

Section 5. Reflection

5.1

The dataset contains many unnecessary variables about the weather such as max, min and mean of weather condition. Usually we only need the mean. Also the dataset only includes weather as one general factor. There are other factors that affect ridership such as whether there is a big event in the city, etc.

From the scatter plot of our data, we can see that our data doesn't satisfy the normal distribution hypothesis and it does not have strong linear trend. To improve the performance of linear model, we can do some transformations on input variables. Or use clustering method to analyze.