Analyzing the NYC Subway Dataset

Questions

Xinqi You

May 22, 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 two-sided P value. The null hypothesis is that two samples (with rain and without rain) come from the same population, i.e. $H_0: \mu_{rain} = \mu_{notrain}, H_a: \mu_{rain} \neq \mu_{notrain}$. The default p-value returned by Scipy Mann Whitney U test is one-sided, which is 0.025 in this case. Since our test is two-sided, the critical p-value is 0.05.

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

$\mu_{rain} = 1105.45, \mu_{notrain} = 1090.28, p - value = 0.05$

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. Means of rainy days and non-rainy days are different at the 95% confidence level.

Section 2. Linear Regression

2.1

I use OLS of Statsmodels to compute the coefficients theta.

2.2

I select rain, meanprecipi, meantempi, weekday, day_week, hour and UNIT. UNIT is the dummy variable.

2.3

'rain, meanpreci, meantempi' will give information about the weather. 'weekday, day_week, hour' provide information about the time that people ride. During rush hours there are more people riding subways. 'UNIT' provides information about locations of riders. There are other variables also include these three information. To simplify the model, I select these representative variables. Note that adding 'UNIT' will greatly improves our model (i.e. R^2) since location of stations matters a lot in the prediction. Some units tend to have more rider and some unites have fewer due to the demographic distribution.

2.4

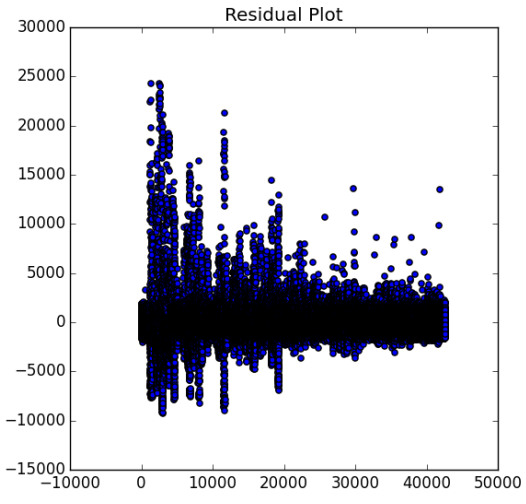
Of the non-dummy features:

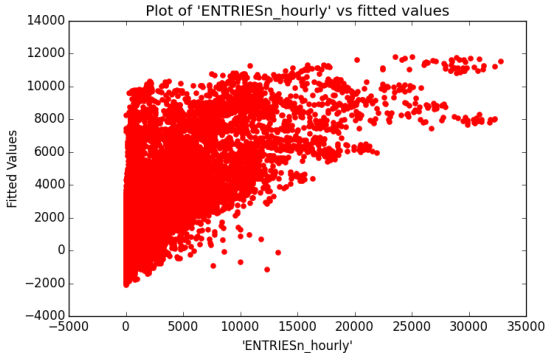
rain	meanprecipi	meantempi	weekday	Day_week	hour
-56.6601	791.1276	-12.6914	1264.9150	76.2369	123.3637

2.5 $R^2 = 0.483$

2.6

R-squared means how much variance is described in the linear regression model. In my model, the OLS model explains 48.3% of total variance of our data. The R-squared is relatively low.





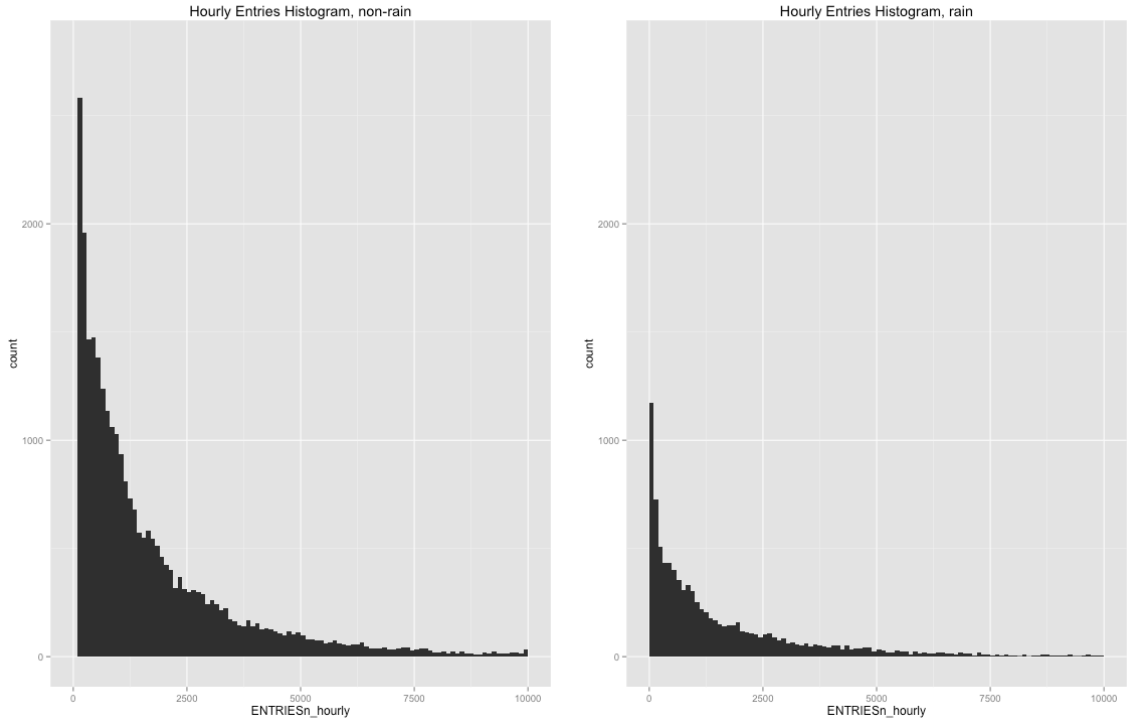
From the residual plot and plot of data vs. fitted value we can also see that linear

model does not have a good fit. Residuals are not scattered equally around 0 and the variation is big. Meanwhile from the relational plot of the data and fitted values, the scatterplot looks quite far away from the ideal straight line. Therefore, it's not a good model to predict ridership.

Section 3. Visualization

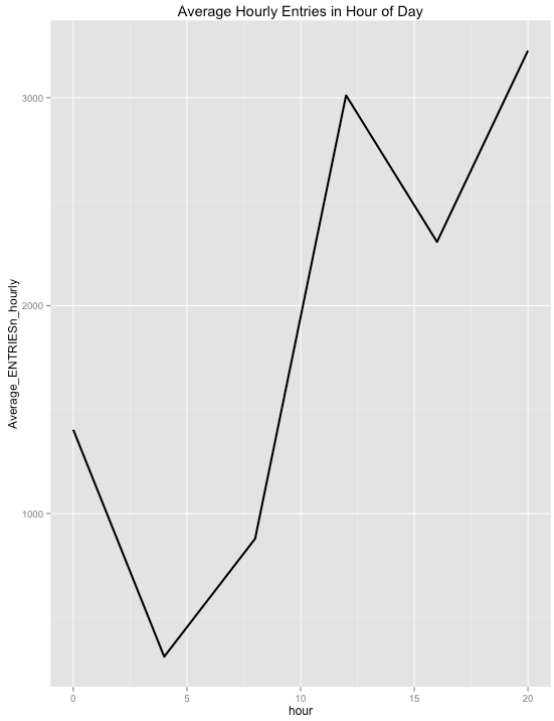
3.1

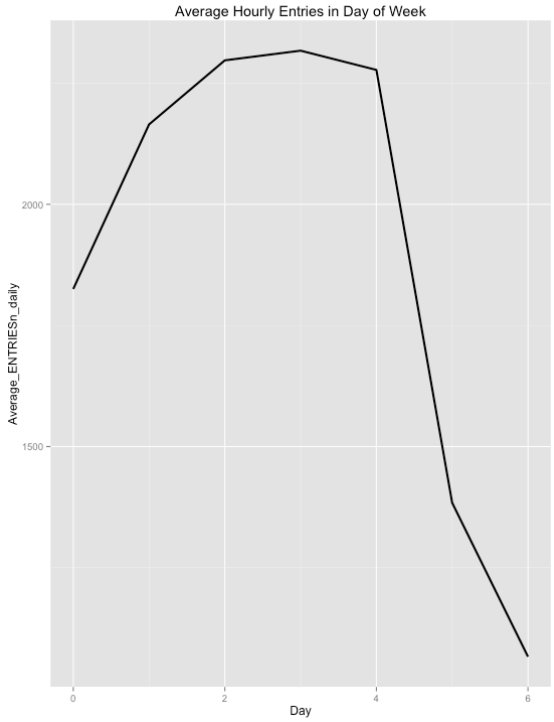
I plot the rainy vs. non-rainy histograms on two separate graphs. Note that non-rainy days have more riders than rainy days.



3.2

To access the ridership by time, I first calculate the average values of hourly entries at each time period. And then plot the line-graph of the average values. The left graph shows that there are more riders in the afternoon and night. The right graphs show that more people take subway on weekdays. Tuesdays, Wednesdays and Thursdays are peak days in the week.





Section 4. Conclusion

4.1

Much fewer people ride the NYC subway when it is raining.

4.2

There are three evidences support his conclusion. First of all, 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. Secondly, from our Mann-Whitney U test result, the means of rainy and non-rainy samples are statistically different. Thirdly, the coefficient of rain is -56.6601 in the OLS model. The negativity of coefficient indicates a negative correlation between rain and hourly entries. In other words, rainy days have fewer riders 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 information about weather and location. There are other factors that affect ridership such as whether there is a big event in the city, etc. Having data of only one month could give us less information about subway ridership in general. It might be the case that people tend to take subway more often or less often in May than other times of year. The analysis and conclusion might be biased in this case.

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.