Analyzing the NYC Subway Dataset

Section 0. References

http://ggplot.yhathq.com/docs/geom_histogram.html
http://www.ats.ucla.edu/stat/mult_pkg/faq/general/tail_tests.htm
http://nbviewer.ipython.org/url/www.alma.cl/~itoledo/Presentation1.ipynb
http://es.wikipedia.org/wiki/Histograma
http://bryansmithphd.com/
https://www.jetbrains.com/pycharm/
http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html
http://stackoverflow.com/questions/17784587/gradient-descent-using-python-and-numpy
http://www.bogotobogo.com/python/python_numpy_batch_gradient_descent_algorithm.php

Section 1. Statistical Test

1.1 Which statistical test did you use to analyze the NYC subway data?

Mann Whitney U-test.

1.1 Did you use a one-tail or a two-tail P value?

One-tail, but if I need the two-tail I can multiply the p value by 2 according with the scipy documentation.

1.1 What is the null hypothesis?

In inferential statistics on observational data, the null hypothesis refers to a general statement or default position that there is no relationship between two measured phenomena.

1.1 What is your p-critical value?

The P value answers this question:

If the groups are sampled from populations with identical distributions, what is the chance that random sampling would result in the mean ranks being as far apart (or more so) as observed in this experiment?

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.

Because the data is not distributed normally. The Mann-Whitney test, is a nonparametric test that compares two unpaired groups. To perform the Mann-Whitney test, Prism first ranks all the values from low to high, paying no attention to which group each value belongs. The smallest number gets a rank of 1. The largest number gets a rank of n, where n is the total number of values in the two groups. Prism then averages the ranks in each group, and reports the two averages. If the means of the ranks in the two groups are very different, the P value will be small.

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 = `0.024999912793489721`

1.4 What is the significance and interpretation of these results?

This means that you can reject the null hypothesis and that the difference is due to random sampling, and conclude instead that the populations are distinct. This means that we can said that the difference in ridership in rain and in non rainy days it's statically 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. OLS using Statsmodels or Scikit Learn
B. Gradient descent using Scikit Learn
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?

Rain, precipi, Hour, meantempi. Yes I use the UNIT dummy variable.

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.

Because I was trying to find a relation of readership and rain. I think that rain was an important part and also the amount of rain. If it's only a light rain the people wouldn't go to the subway. As well, if the temperature it is really cold outside, people wouldn't be likely to be walking.

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

rain = 3.01708749e+01, precipi = 2.06990813e+01, meantempi = -1.04394319e+01

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

```
Your r^2 value is 0.443285548603
```

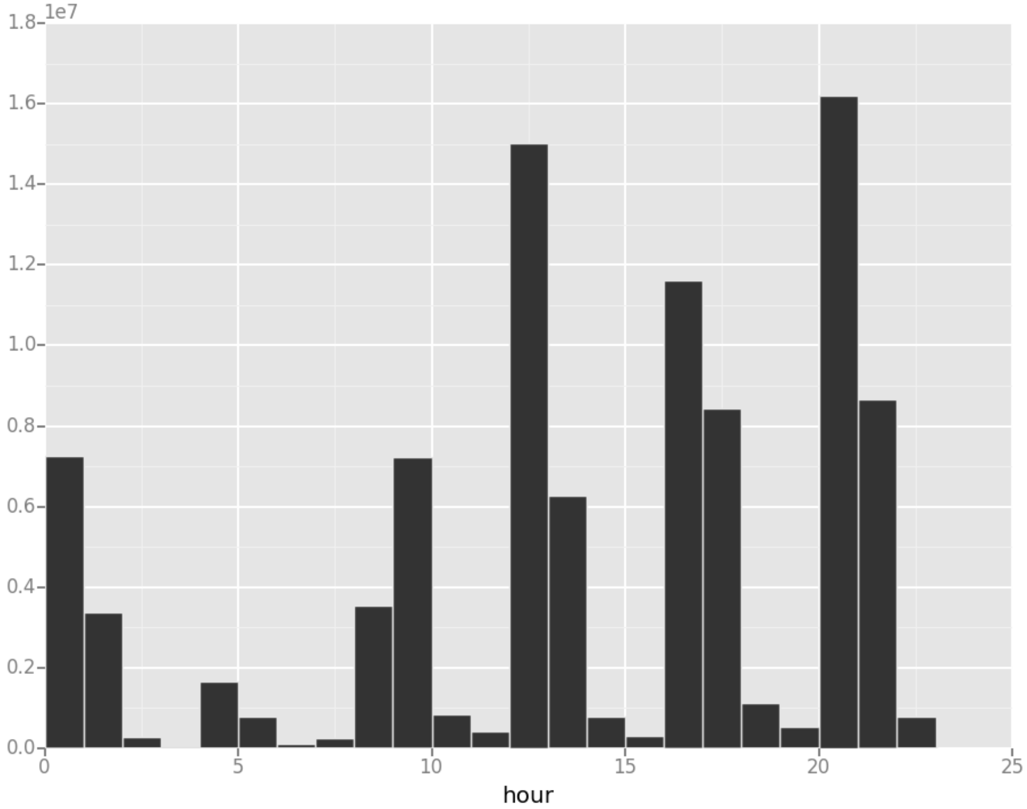
2.6 What does this R^2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?

R^2 is the percentage of variance that is explained, and represents how much certainty we have in order to predict data. In this case a 44.34% percent of certainty it's not a very good percentage if we need to make a critical decision. But it's a good percentage if we are using this feature, like in a app or a non critical way.

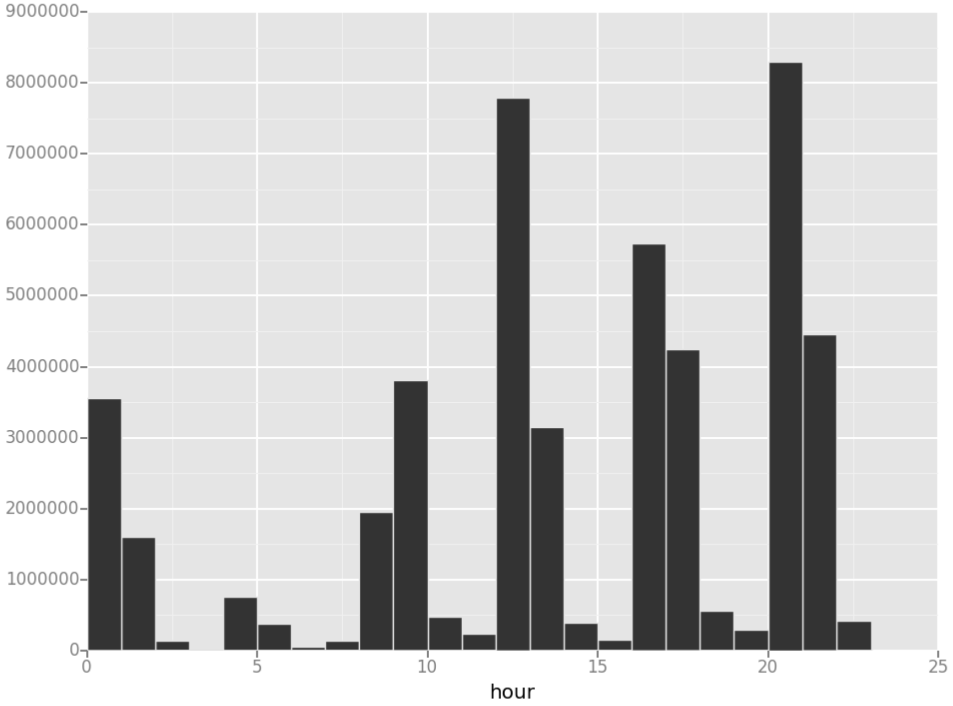
Section 3. Visualization

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

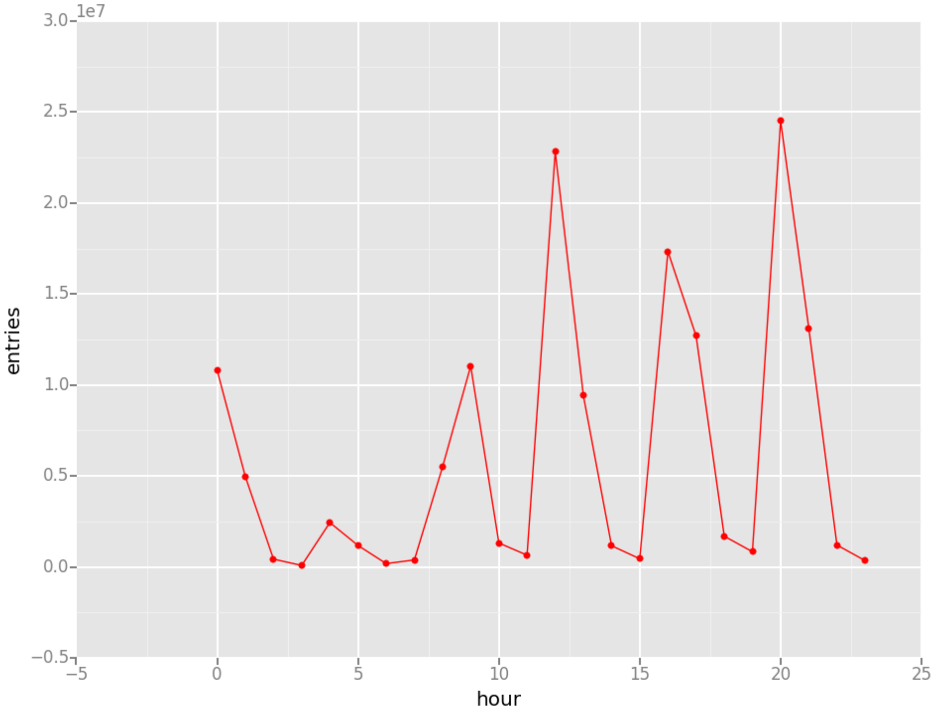
No rain by hour



Rain by hour



Ridership by time-of-day



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?

Yes, more people ride the NYC subway in rainy days. The results from the Mann-Whitney U test, give us a good confidence in this statement. Because of the variance of the means the Mann-Whitney U test is a quantitatively method to confirm that the two data sets are statistically different.

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.

In the linear regression, the coefficient for the rain parameter indicates that the presence of rain contributes to increased ridership, that's because it's a positive value, also a big one compared with others. The R^2 (44%) give us a good idea of the certainty we can trust the predictive model.

Section 5. Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis

One thing that caught my attention was the ‘UNIT’ column, ridership varied greatly. The Mann-Whitney U test did not take the location of the turnstiles into account, and only looked at the subway entry distributions for rain and no-rain. Considering how the same stations at the same day and time varied by rain, could have increased the fidelity of the test.

I would like to make a cross reference with other data sets like an example of taxi ridership, traffic and rush hours. Also more info in the data set like the direction of the trains and the average ride time. As well, a sample size (larger data set) and normalization by location/turnstile ID could increased the confidence of the Mann-Whitney U test and the linear regression model.

Even that the linear regression model was a very simple approach, it was adequate for the purpose of this example. As mentioned in Section 2.6, the inclusion of more features or polynomial combinations could have increased the accuracy of the model.