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?

Based on the feedback and in the SciPy documentation I used the two-tail. According to the SciPy doc which states: "*This test corrects for ties and by default uses a continuity correction. The reported p-value is for a one-sided hypothesis, to get the two-sided p-value multiply the returned p-value by 2.*"[1] Also the mean it's pre calculated so it's considered that the two population are the same size.

1.1 What is the null hypothesis?

The null hypothesis help us state if there is no relationship between two measured phenomena. Based in that definition I state the null hypothesis is: ***the rain has no correlation with the ridership***.

1.1 What is your p-critical value?

The p-critical value is `0.05` and my p-value is `0.049998`, so I can reject the null hypothesis and state that the difference is due to random sampling, and conclude instead that the populations are distinct.

[1] scipy.stats.mannwhitneyu: http://docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.stats.mannwhitneyu.html

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. [2] So this help us to understand the distribution of the data in contrast with the Wekch's test.

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 = `0.049998`

1.4 What is the significance and interpretation of these results?

The difference between the rain and no rain it's 1.372%, showing us that the users use more the subway in rainy days. Thanks to the Mann Whitney U-test, we know that this difference its statically relevant. So we can say with great confidence that the null hypothesis is false, letting us to conclude that the rain increments the ridership in the subway.

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 where the chosen features. dummy variables were introduced in each data point, noted 'UNIT', they were initialized with boolean.

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.

[2] Graphpad mann whitney test:
http://www.graphpad.com/guides/prism/6/statistics/index.htm?how_the_mann-whitney_test_works.htm

This were the variables that give the best result using the linear regression model, also I noted that the temperature play also a role in the increment of the ridership. Also the hour variable gave a better picture of the variation in each turnstiles, tied with the rush hours. I was unable to find R^2 values that were better.

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

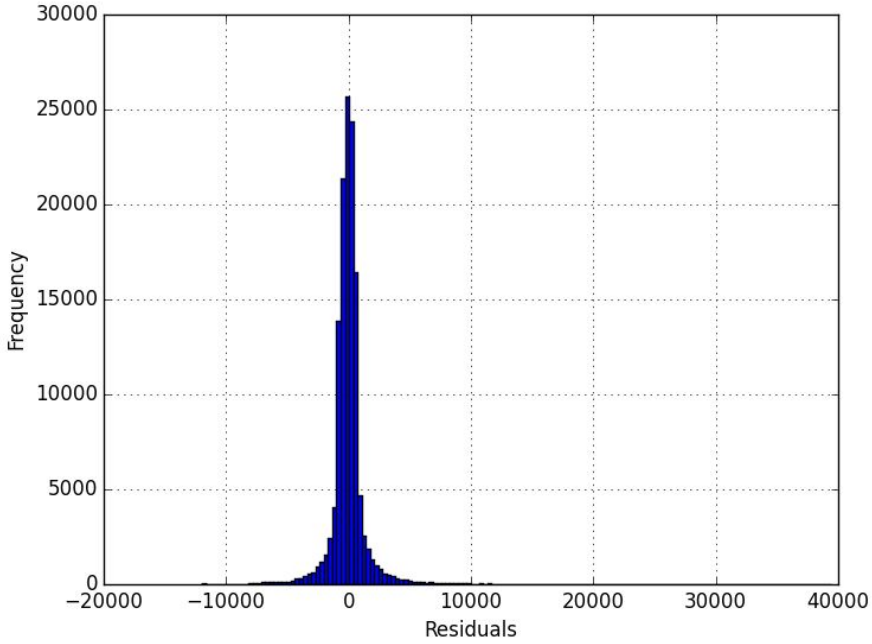
rain = 3.01708749e+01, precipi = 2.06990813e+01, meantempi = -1.04394319e+01

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

```
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.

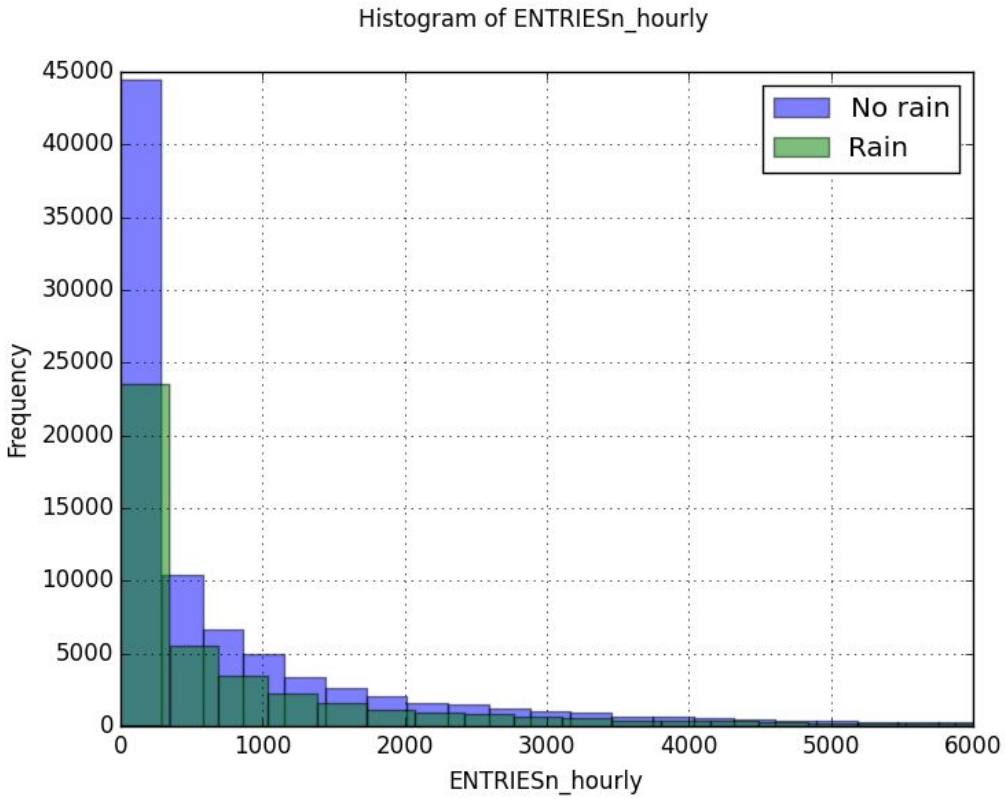


The histogram of the residuals has long tails, which suggests that there are some very large residuals or difference between the predicted data and the actual data. Indicating that this is not a very good fit for the model. This is also reinforced by the low value of the R^2

Section 3. Visualization

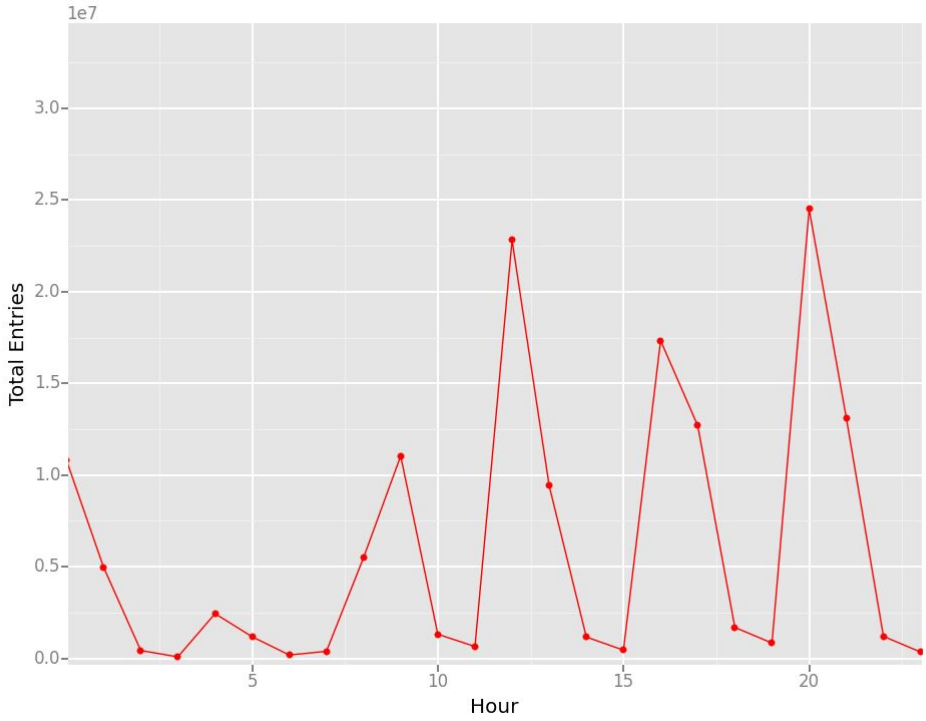
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

Histogram of ENTRIESn_hourly No rain vs Rain



Histograms of subway entries for rainy and no rainy hours.There were less rainy days than there were not rainy days, it would be incorrect to think that subway ridership is less when it rains.

Ridership by time-of-day



Getting the ridership by hour we can get a very good idea of which are the most congested hours in the entire subway system. The bigger ones are at noon and 8pm, contrary to the common belief that it's during rush hours (8-9am and 5-6pm)

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 and big value, if we compared them with the rest values/data. The R^2 (44%) give us a good idea of the certainty to trust on 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, using this column make the results of the linear regression model varied greatly. The linear regression model 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 in the linear regression model. For example based in the ridership by time-of-day the more congested hours are not the rush hours, so I’m curious about what other factors have impact in the ridership.

Even than the linear regression model was a very simple approach, it was adequate for the purpose of this example, but doesn't give us a very good confident value. Including more features or polynomial combinations could have increased the accuracy of the model, but also would make this model only fit with this data set.