Analyzing the NYC Subway Dataset

by Murat Yerlikaya in fulfillment of Udacity's Data Analyst Nanodegree, Project 1

Section-0 References

http://wiki.scipy.org/Tentative_NumPy_Tutorial
http://pandas.pydata.org/pandas-docs/stable/10min.html
http://en.wikipedia.org/wiki/Wikipedia:Statistics

Section-1 Statistical Test

1.1 Which statistical test did you use to analyze 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 the Mann-Whitney U test to analyze the NYC subway data.

The null hypothesis in our case is that both populations are equal, or that there is no significant deviation on both populations medians.
Because of the null hypothesis I will use a two-tail p-value.
I will use a p-critical equal to 0.05, meaning that in case the null hypothesis is false we will require a 95% of confidence.
In using the Mann-Whitney U test, the null hypothesis is that the two populations are the same, or simply put, that rain has no correlation with ridership. The p-critical value used was 0.05, or 5%.

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.

> As shown in Section 3.1, neither the rain or no-rain histograms are normally-distributed. As such, a non-parametric test such as Mann-Whitney U is a good fit, while a test such as Welch's two-sample t-test is not. To understand whether our data is normally distributed or not a Shapiro-Wilk test could have been conducted.

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.

> I used the scipy implementation of the Mann Whitney U test (scipy.stats.mannwhitneyu)

The results from the test are :

```
U-statistic: 1924409167.0

  p-value: 0.025
```

But the user should be aware that scipy reports a p-value for a one-tailed hypothesis, so we multiply by 2 to get the significance for our hypothesis:

```
p-value : 0.05
```

The average from two data samples are follow:

```
  Mean entries with rain: 1105.446
  Mean entries without rain: 1090.279
```

1.4 What is the significance and interpretation of these results?

The distribution of the number of entries is statistically different between rainy and non-rainy days. Comparing the means yields 1.4% more subway entries when it rains. This statistic alone is insufficient in drawing conclusions or correlation. The U-statistic has a high value, very close to the maximum value of 1937202044.0, or half the product of the number of values in each data set. A U-statistic of half the maximum would indicate that the null hypothesis is true. Of note, the p-value 0.05 satisfies the p-critical value, and the conclusion can be drawn with 95% confidence that the null hypothesis is false and that ridership is different with vs. without rain.

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 machine learning algorithm is a branch of artificial intelligence focused on constructing systems that learn from large amount of data to make prediction. So we've determined that we want to figure out which set of parameters provide best predictions for output variable. To define this I used Gradient descent to train the linear regression coefficients. The given values were sufficient in converging on a local minimum, as confirmed by plotting the cost history vs. number of iterations.

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

Here are the included features used in my model. Rain (0 or 1), precipitation, hour and mean temperature. Per the default configuration, dummy variables were introduced for features 'UNIT' (the turnstile location/identification number), which were categorical in nature. They were initialized with boolean (0 or 1) features with prefix 'unit,' and each data point would have a '1' in the feature that it "belonged" to.

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.

I maintained rain, precipitation, hour, and mean temperature because out of experimentation, I was unable to find R^2 values that were better. When I added mean-pressure into my model, I saw slight increase in R^2.

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

	coef
const	1539.1268
rain	29.4645
precipi	28.7264
Hour	65.3346
meantempi	-10.5318

2.5 What is your model's R2 (coefficients of determination) value?

```
0.47924770782
```

2.6 What does this R2 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 R2 value?

R^2 is essentially the percentage of variance that is explained, and is a quantitative measure of the model. For a successful model, we should check residuals as well. The residuals analysis are very good indicators of the behaviors of the ridership that the model can't explain, mainly because it was a very rough assumption to use hour as it is clearly not well modeled by the linear regression.
Finally it is interesting to independently check that even when the rain variable can be fit by a linear model, it significance is very low as can be seen by the low p-value of the coefficient. In fact, removing rain as predictor feature only reduce the R2 by less than 0.0001.

Residuals are the differences between the predicted and the actual values. If you have a good model, you expect to see the residuals a normal distribution.

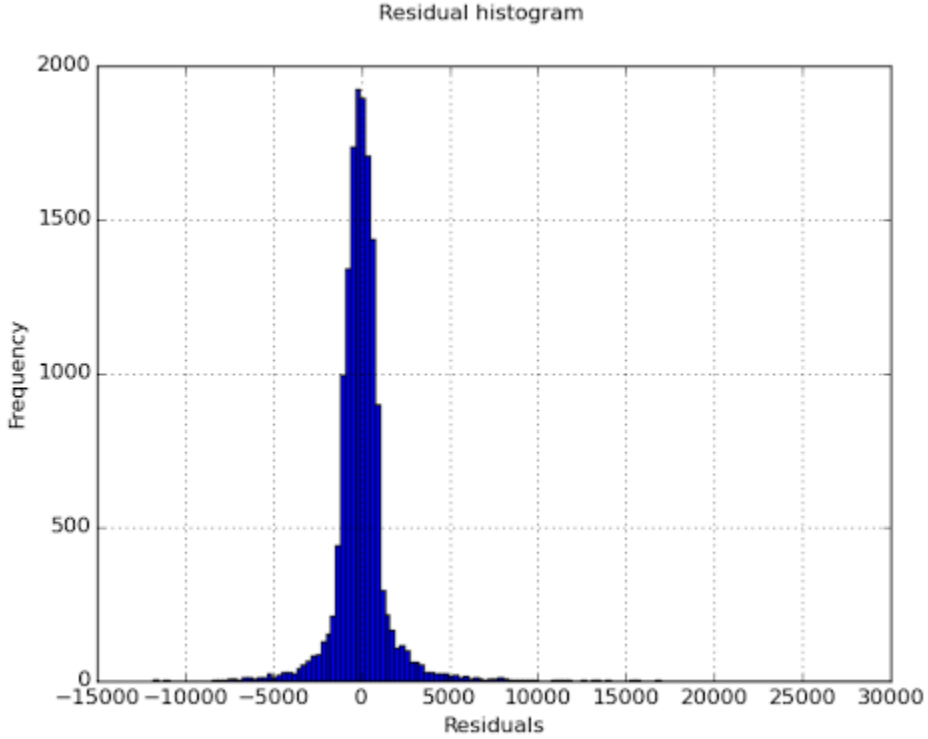

Figure 1: Residuals histogram for linear regression
Randomness and unpredictability are important parts of a regression model; we expect random errors to produce residuals that are normally distributed.

Section-3 Visualization

3.1 Include and describe a visualization containing two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

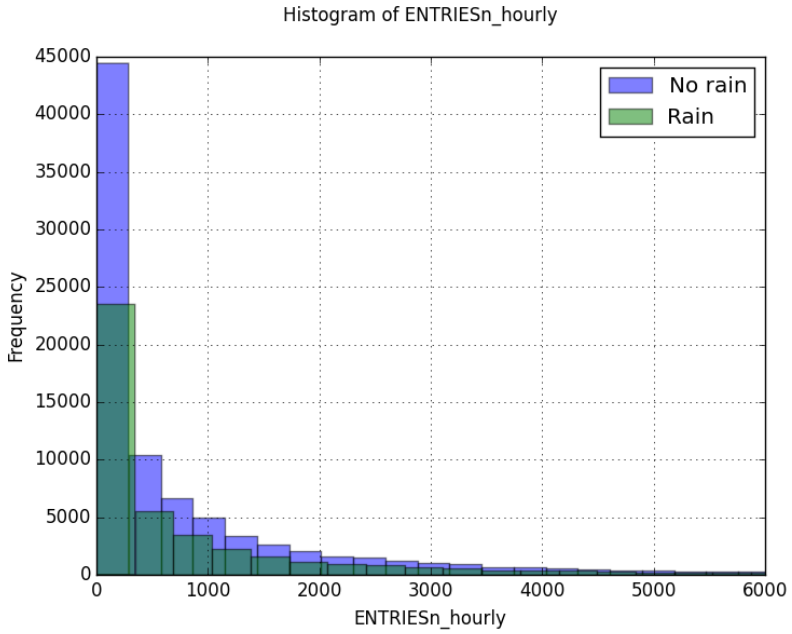

Figure 2: Ridership distribution comparison between rainy and dry days.

Plotting overlaid histograms of subway entries for both rainy and dry hours shows that both distributions are not normally-distributed. Of note, it's important to clarify that these are aggregate values and that there were less rainy days than there were not rainy days; it would be grossly incorrect to draw from this graph that subway ridership is less when it rains.

3.2. Some other visualizations

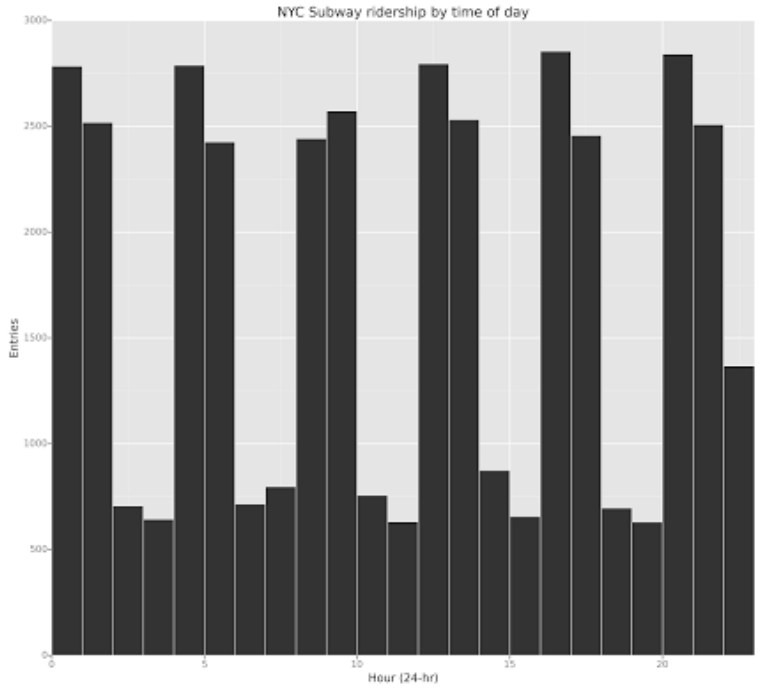

Figure 3: Ridership distribution by time of day.
Graph shows that ridership by time of day. Entries increase early morning, by noon and end of day. It can conclude that this shows traveling from home to business and business to home.

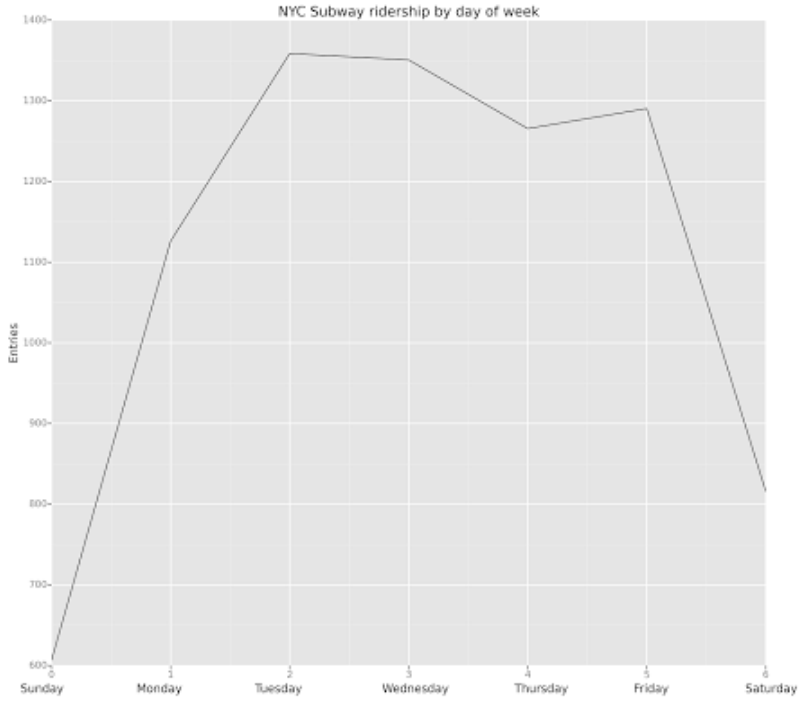

Figure 4: Ridership distribution by day week.
Similar to Figure 3,Ridership increases on work days and decreases on weekend.

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?

It is important to note that simply looking at the means of both data sets is insufficient, due to variance. The Mann-Whitney U test is needed to quantitatively confirm that the two data sets are statistically different.

Particularly given the results from the Mann-Whitney U test (p-value: 0.05), we can say with a high level of certainty that 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.

A non-parametric statistical test between two samples, being one sample the data observed in rainy days and the other the data observed in non-rainy days. Even when in the beginning the small difference in the ridership volume reported for each sample seemed to be significant by the statistics reported by the test, it was not clear that both samples were really independent.

A second analysis was done by means of the use of a machine learning technique. Even when some predicting features were found, as the hour of the day, the day of the week, holiday or workday, the rain indicators didn't have either a significant weight or a high p-value, thus confirming by an independent method that precipitations didn't seem to play a roll on the ridership behavior of the NYC subway, or from the fitting to a multiple regression model.

Section-5 Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis, including: data set, linear regression model, and statistical tests.

A combination of increased sample size (larger data set) and normalization by location/turnstile ID could have potentially increased the confidence of both the Mann-Whitney U test and the linear regression model. As we saw from examining the 'UNIT' column, ridership varied greatly. Simply put, some stations and turnstiles were naturally more active than others. The Mann-Whitney U test did not take this into account, and only looked at the subway entry distributions for rain and no-rain. Examining how the same stations at the same day and time varied by rain could have increased the fidelity of the test.

The use of a linear regression, even with multiple features, was shown to be not enough to model the ridership behavior of the NYC subway. The assumption of linearity between the predicting features and the entries by hour was not met for most of the variables, and the residual analysis confirmed the poor fit.

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

I think that an interesting investigation would be to use gradient descent with logistic regression to see if one might be able to predict if it rained or not given various parameters, to include turnstile location/ID, time of day, and subway entries. Intuitively, this might produce false positives or negatives on special days (e.g. sports game, holidays, etc.).