David Pankiewicz
dwpankiewicz@gmail.com
Data Analyst Nanodegree
5/20/15

Analyzing the NYC Subway Dataset

Section 0: References

Mann Whitney U Function Docs
Pandas – Melt
Numpy – Histograms
Plot Two Histograms in R
Yhat - ggplot
Stack Overflow – Bar Chart in ggplot

Section 1: Statistical Test

1.1 The Mann-Whitney U Test was used to compare ridership in rainy weather vs. ridership in non-rainy weather. A two tailed p-value was used to decide whether to reject the null hypothesis that these populations are the same. A p-critical value of 0.05 was used.

1.2 The test is applicable because the Mann-Whitney U test is a non-parametric test and thus makes no assumptions about the underlying distribution of the data. A Welch's T-test was not applicable because the data is not normally distributed.

1.3 The results of the test:

- With rain mean = 1105.4463767458733
- Without rain mean = 1090.278780151855
- 1 sided p value = 0.0249999127934
- 2 sided p value = 0.049999825587

David Pankiewicz
dwpankiewicz@gmail.com
Data Analyst Nanodegree
5/20/15

1.4 Because the 2 sided p-value is less than 0.05, we reject the null hypothesis that these populations are the same (the value is just barely below 0.05). This leads to the conclusion that are differences between the ridership populations with rain and without.

Section 2. Linear Regression

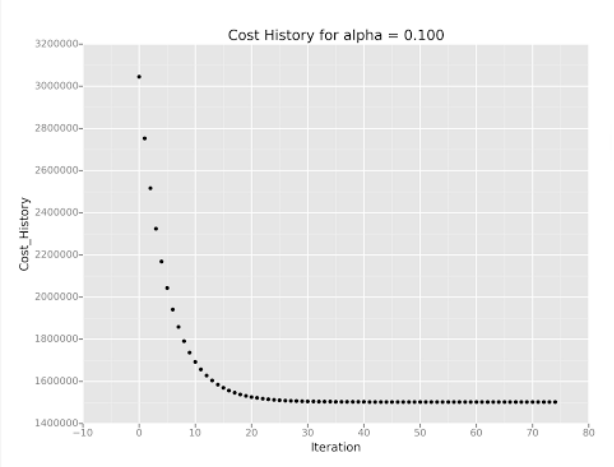
2.1 I used gradient descent to make my predictions.
2.2 I used 'precipi', and 'rain', and the dummy variables 'UNIT' and 'DATEn'.

2.3 I first added variables based on intuition and then tracked their effects on R^2. I then tested deleting variables and noticed if this had any effects on R^2. (See next page for a table).

None of the weather variables added beyond the base model increased R^2 significantly. In fact, the rain and precipitation values had very little predictive power as measured by R^2 changes relative to other non-weather factors. To ensure thoroughness, I tested the effect of removing rain and precipi separately, and then together. Neither produced strong effects.

Upon experimentation, UNIT has by far the largest impact on R^2. Hour and Date each have a measurable impact, but the changes are not near UNIT. Upon consideration, this is logical: UNIT is a proxy for location, and differences in location are intuitively a causal reason for differences in subway use.

I also tested variations in alpha and iterations, but found the effects minimal. This is explained by the cost history chart (below). An alpha value of 0.10 and 75 iterations captures nearly all improvements in minimizing cost.



David Pankiewicz
dwpankiewicz@gmail.com
Data Analyst Nanodegree
5/20/15

R^2 Values for Different Linear Regression Models Predicating NYC Subway Ridership

R^2	Comment	Change in R^2 vs. Baseline	Alpha	Iterations	Variables	Dummy Variables
0.463969	Baseline		0.1	75	rain, precipi, meantempi	UNIT, Hour
0.463968	Remove rain	-0.000001	0.1	75	precipi, meantempi	UNIT, Hour
0.463943	Remove precipi	-0.000025	0.1	75	rain, meantempi	UNIT, Hour
0.463923	Remove rain, precipi	-0.000046	0.1	75	meantempi	UNIT, Hour
0.425845	Remove hour	-0.038123	0.1	75	rain, precipi, meantempi	UNIT
0.463314	Remove meantempi	-0.000655	0.1	75	rain, precipi	UNIT, Hour
0.032375	Remove UNIT	-0.431594	0.1	75	rain, precipi, meantempi	Hour
0.478300	Add DATEn	0.014331	0.1	75	rain, precipi, meantempi	UNIT, DATEn
		Change in R^2 vs. DATEn				
0.478300	Simplified	0.000000	0.1	75	rain, precipi	UNIT, DATEn, Hour
		Change in R^2 vs. Simplified				
0.478300	Higher Alpha	0.000000	0.2	75	rain, precipi	UNIT, DATEn, Hour
0.478300	Higher still Alpha	0.000000	0.5	75	rain, precipi	UNIT, DATEn, Hour
0.477953	Less Alpha	-0.000347	0.05	75	rain, precipi	UNIT, DATEn, Hour
Timed out	Higher iterations	na	0.05	150	rain, precipi	UNIT, DATEn, Hour
Timed out	Higher iterations	na	0.1	150	rain, precipi	UNIT, DATEn, Hour
0.474580	Less iterations	0.003720	0.1	25	rain, precipi	UNIT, DATEn, Hour

2.4 The coefficient for 'rain' is 6.787. The coefficient for 'precipi' is 11.426.

2.5 The R^2 is 0.4783.

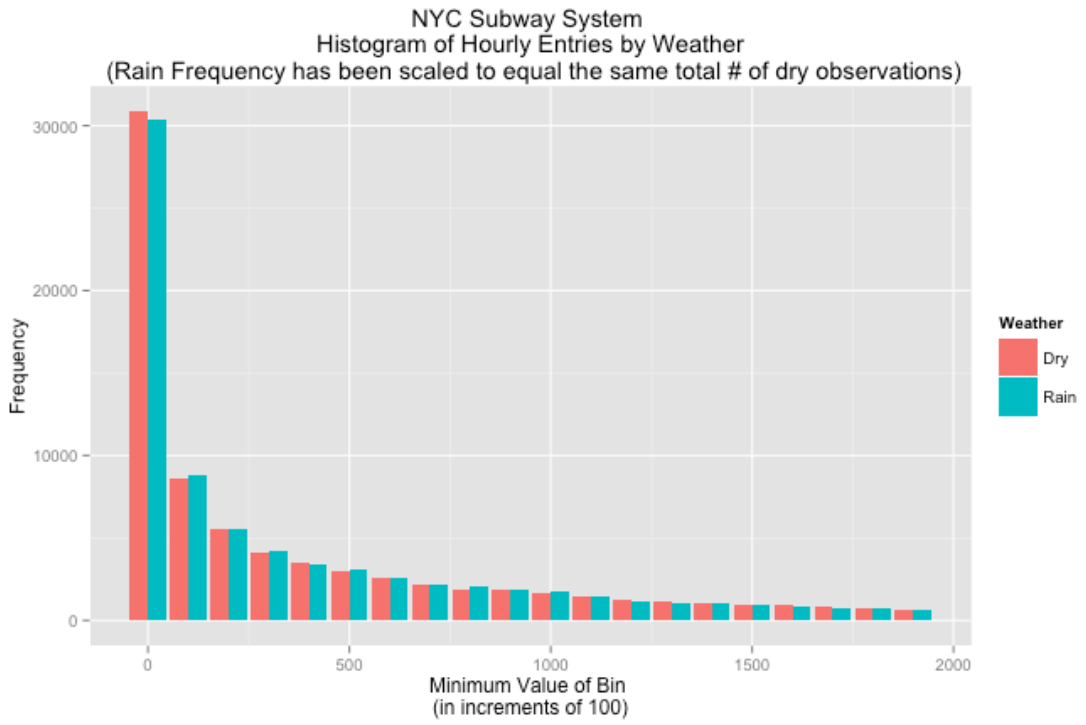
2.6 The R^2 means this model explains roughly 48% of the variability of the response data around its mean. It's difficult to say how good this is without a benchmark value. Given the variability in the data (see section 5. Reflection for more details), this is a good starting point for prediction making. That said, other models (e.g., logistic regression), should be considered and compared based on prediction ability and ease of understanding model itself.

It is worth noting that most of this dataset is weather, and using only weather data produces a tepid R^2 of 0.03. Given that 'UNIT', a categorical variable with over 400 different entries, is driving most of the R^2, linear regression on the raw dataset may not be the most appropriate method to predict ridership.

David Pankiewicz
dwpankiewicz@gmail.com
Data Analyst Nanodegree
5/20/15

Section 3: Visualization

Chart 1



Preparation

Note in the dataset, there are nearly twice as many observations in dry conditions than there are in wet conditions. Thus, a raw histogram (depicted on the next page in in Chart 1B) would give an incomparable histogram, as most of the visible difference in the counts are caused by the size of the samples, and the tiny differences in the distribution are not visible to the eye. Thus, I have scaled the counts for rainy weather to in total equal the number of observations we have for dry conditions, allowing distribution differences to be visible. (This factor was 1.99, the ratio of dry observations to rain observations).

Furthermore, I used a small bin size (100) and limited the graph to bin sizes less than 2000. This captures roughly 90% of our dataset. Entries beyond this point do not have differences that are graphically visible.

Analysis

This chart agrees with the Mann-Whitney U test that the distributions are statistically different, albeit only slightly.

David Pankiewicz
dwpankiewicz@gmail.com
Data Analyst Nanodegree
5/20/15

The graph implies that for observations with less than 1000 entries per hour, rain increases ridership slightly. Although the first bin (0 – 100 entries) lessens, this may still support this hypotheses. Because more people are riding, some counts that would fall into this smallest bin now fall into larger bins (e.g., the 100 – 200 bin). For observations with more than 1000 entries, rain doesn't seem to make a difference. This is likely due other factors (like location) driving the bulk of ridership, regardless of the weather.

Chart 1B: Unscaled Ridership Histogram

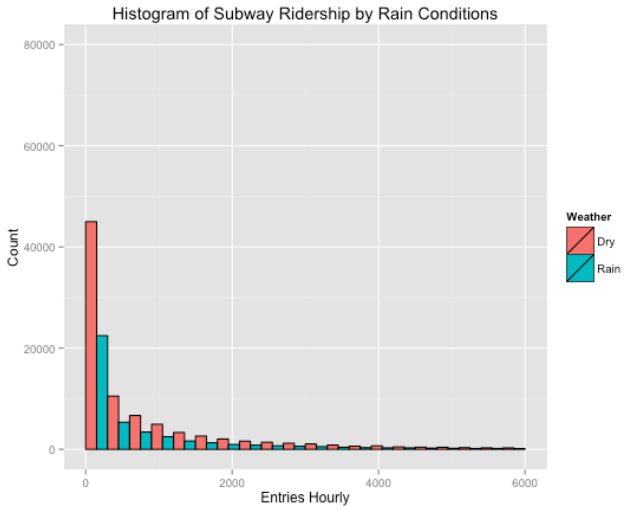
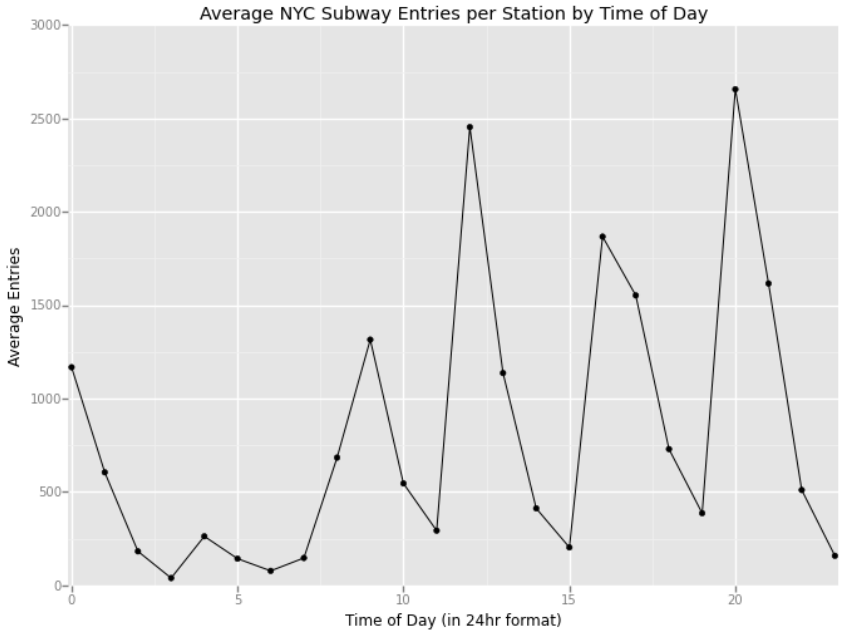


Chart 2:



David Pankiewicz
dwpankiewicz@gmail.com
Data Analyst Nanodegree
5/20/15

Chart 2 represents average NYC subway entries per station by time of day. The troughs in the graph may be explained by the times that most people are either sleeping or working, while the peaks may represent work commutes, lunchtime, and evening activities.

Section 4. Conclusion

4.1 Based on my analysis, slightly more people ride the NYC subway when it is raining. It most be noted, however, that the differences are barely measurable in the given data set.

4.2 My conclusion is based on evidence gathered throughout the preparation and analysis of this data set. First, the Mann Whitney U test results reject the hypotheses that the data comes from similar distributions, thus implying that there are differences in ridership when there is rain. Secondly, building a linear regression model with rain and precipitation as inputs produces coefficients that are positive. To corroborate this, I also ran an OLS model. This also returns positive rain and precipitation coefficients, with 95% confidence intervals that contain only positive numbers (i.e., rejecting the hypotheses that the coefficients are equal to 0). This would mean that holding all other factors constant, rain increases ridership. Finally, based on the histogram created with the data (see Chart 1), rain data shifts towards higher ridership compared to during dry conditions.

Section 5: Reflection

5.1.1 The data is noisy in regard to answering questions of ridership given weather. Most of the variability in entries are related to subway location, which crowds out the possible effects that weather may have on ridership. Furthermore, other factors like the day (using 'DATEn' as a proxy) and the time also significantly influence ridership but have nothing to do with the weather. The data is useful for the purpose of predicting subway ridership given subway station, day, and time, but gives us little predictive ability or ability to answer questions in regard ridership given weather because of the wide variability in the non-weather factors.

5.1.2 Linear regression is often a good starting point for analysis, but will be suboptimal unless the data itself is driven by an underlying linear relationship. Other types of models (e.g. logistic regression) should be applied and analyzed for appropriateness to the data and fit.

David Pankiewicz
dwpankiewicz@gmail.com
Data Analyst Nanodegree
5/20/15

Another shortcoming is that these methods treat all data points equally, and thus make it difficult to tease out differences in results. For example, subway stations that have tiny ridership are likely to have tiny ridership regardless of weather, and weather conditions at 3am are unlikely to affect many riders regardless of subway location. However, in performing a linear regression model, a small subway stop in the wee hours of the morning factors into the model just as equally as a major hub during rush hour (where differences may actually be noteworthy and likely to show in the data). This is not necessarily a shortcoming of the analysis methods used, but instead requires that additional consideration go into data preparation prior to analysis.