Joel Smith
4/6/2015
Intro to Data Science

Analyzing the NYC Subway Dataset

References

https://www.khanacademy.org/math/probability
https://developers.google.com/edu/python/
http://synesthesiam.com/posts/an-introduction-to-pandas.html

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?

Used a two tailed p-critical value of 0.05 with a Mann-Whitney U-test. My null hypothesis was that there is no statistical difference between subway ridership on rainy and non-rainy days.

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

The sample data had a non-parametric distribution. Welch's t test would of only been applicable if the data had a normal distribution.

.

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.

Mean with rain = 1105.45
Mean without rain = 1090.28
U-statistic = 1924409167
p value = 0.02

Note: P - value was generated locally from the "turnstile_data_master_with_weather.csv" dataset. Problem set 3 from the exercises on the Udacity site generated a p value of 0.025.

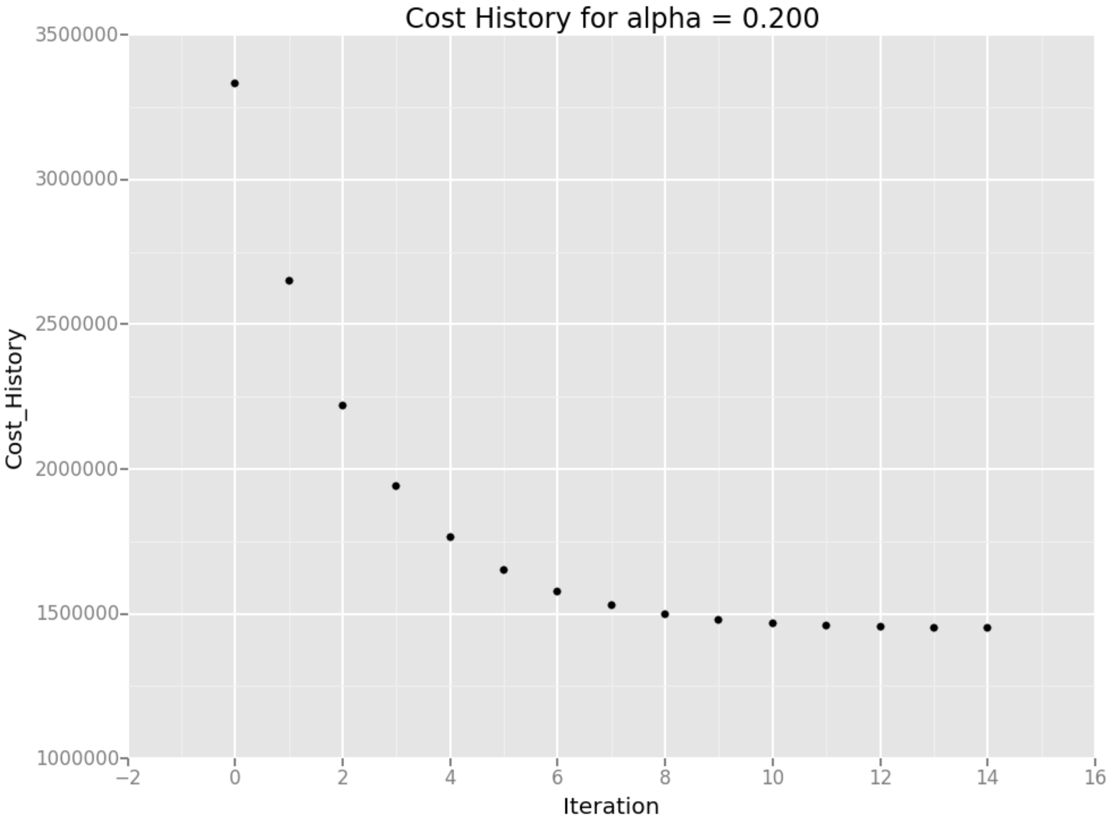
1.4

What is the significance and interpretation of these results?

The Mann-Whitney U-test indicates with a confidence level of 95% that there is a statistical difference between ridership on rainy and non-rainy days. The sample data in particular has a higher average amount of riders on rainy days (1.4%). This is a good indication that people prefer to ride the subway when it rains.

2.1
What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn_hourly in your regression model:
Gradient descent (as implemented in exercise 3.5)
OLS using Statsmodels
Or something different?

Used a gradient descent approach with 15 iterations and an alpha value of 0.2 to train the theta coefficients. Any further iterations had substantial diminishing returns.



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

The features used as input for my model were the hour, mean temperature, mean dew point, and mean pressure along with the values indicating the presence of rain, precipitation, and fog. For dummy variables, I used the “unit” data to distinguish between turnstiles and I also generated and utilized a value that represents the day of week for each data point

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. Your reasons might be based on intuition. For example, response for fog might be: “I decided to use fog because I thought that when it is very foggy outside people might decide to use the subway more often.” Your reasons might also be based on data exploration and experimentation, for example: “I used feature X because as soon as I included it in my model, it drastically improved my R^2 value.”

I selected all the features I believe to be critical for a person in assessing the weather’s state. The humidity, temperature, fog, barometric pressure, and type of precipitation are all factors someone might examine before choosing a method of travel. I also included the hour and day of week since some commuters may prefer different methods of travel at different times or dates. The turnstile unit identifier data was also included to help account for differences in the average amount of traffic each turnstile may receive on average at their locations.

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

```
['rain', 'precipi', 'Hour', 'meantempi', 'fog', 'meandewpti', 'meanpressurei']
[-23.5362258,  -12.5131733,   440.231292,  -38.8345505,
 24.2074176,  -9.13145147,  -25.2129327]
```

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

r^2 = 0.469591287708

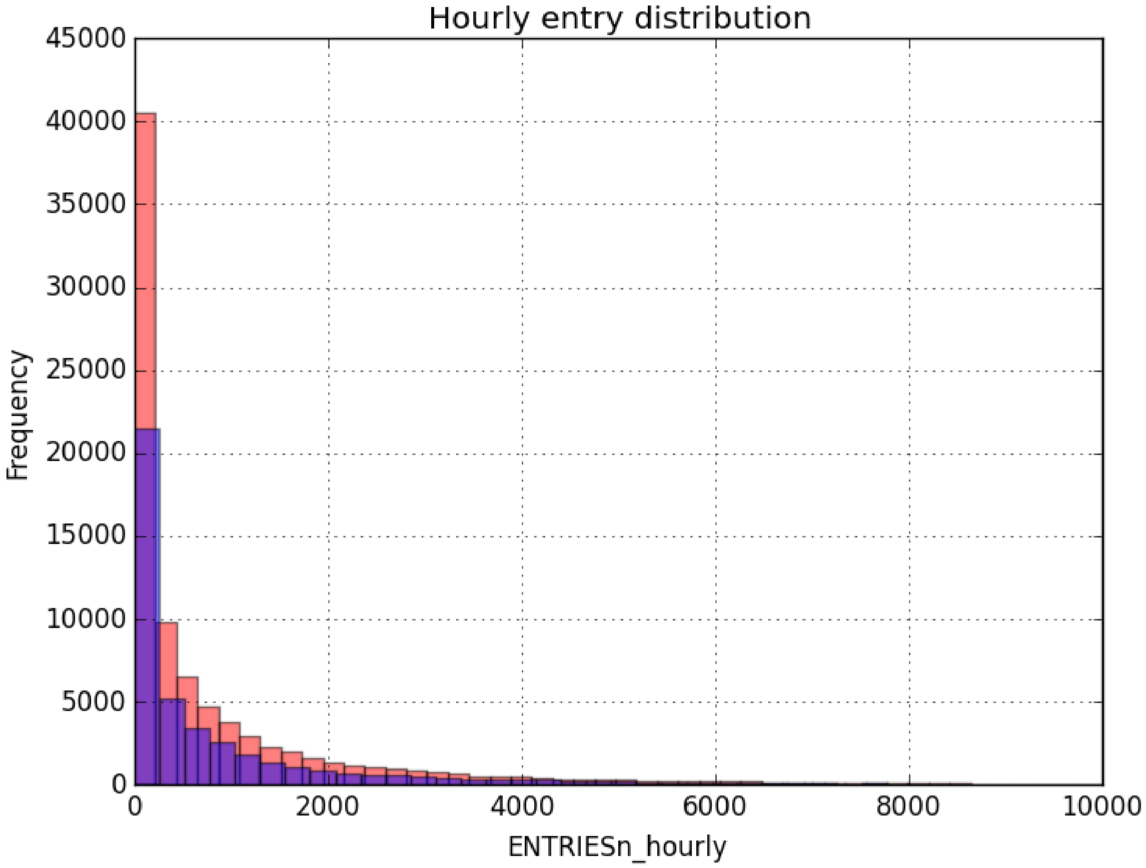
2.6
What does this R^2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?

The value of R^2 indicates that my regression model is able to explain about 47% of the variance seen between subway ridership at different times with different weather conditions. I believe this is a good fit and is appropriate for estimating subway traffic for any non-safety critical purposes.

3.1

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

> **You can combine the two histograms in a single plot or you can use two separate plots.**
>
> **If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.**
>
> **For the histograms, you should have intervals representing the volume of ridership (value of ENTRIESn_hourly) on the x-axis and the frequency of occurrence on the y-axis. For example, each interval (along the x-axis), the height of the bar for this interval will represent the number of records (rows in our data) that have ENTRIESn_hourly that falls in this interval.**
>
> **Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.**

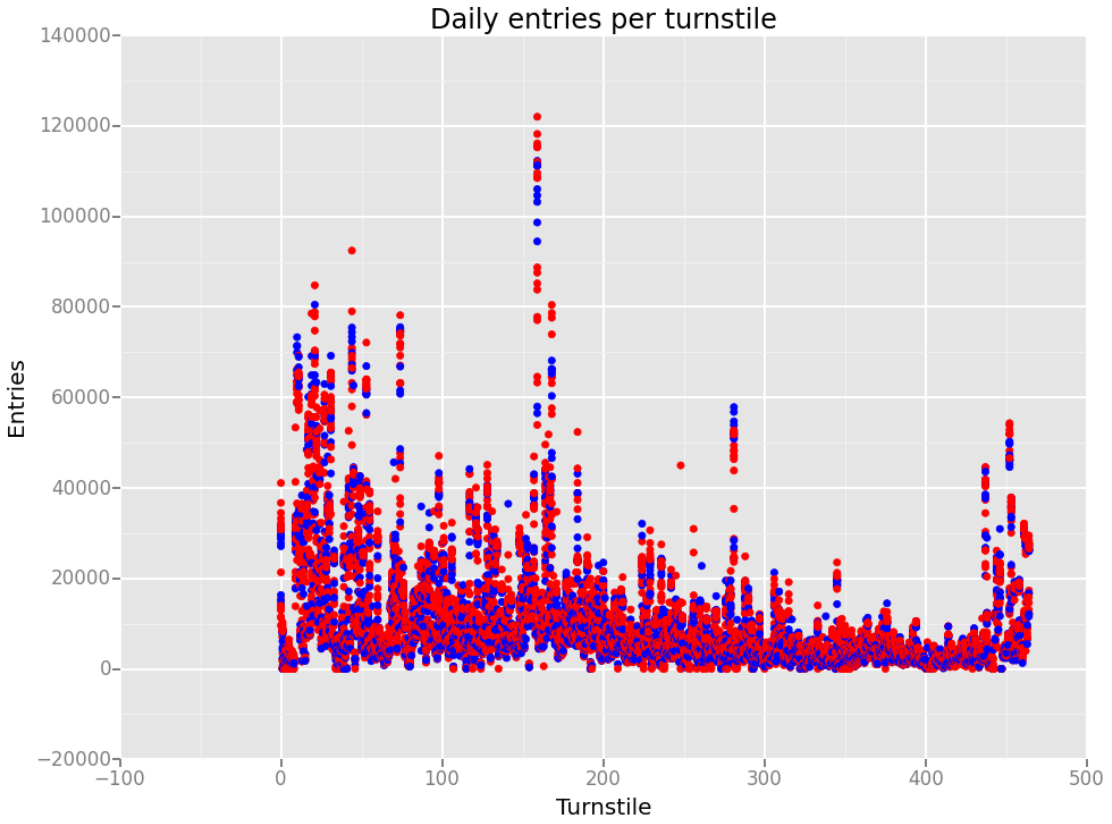


Blue represents entries with rain and red without rain.

3.2

One visualization can be more freeform. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots) or attempt to implement something more advanced if you'd like. Some suggestions are:

Ridership by time-of-day
Ridership by day-of-week



Each data point represents the amount of riders passing through a turnstile per day with the blue points indicating days where there was rain. The x axis is an identifier for the turnstile and the y axis is the amount of riders.

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?

The Mann-Whitney U-test performed indicated with a 95% level of confidence that there was a statistical difference in the average amount of riders on the subway for rainy and non-rainy days. The sample data in particular displayed a slightly higher average amount of riders on rainy days. As a result, it would be statistically sound to conclude that more people prefer to ride the subway when it is raining. However, I feel it may be more accurate to conclude that more people ride the subway when weather conditions are poor.

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

The significance test illustrated with some certainty that the differences in the average amount of riders on rainy and non-rainy days from the sample data could be reflective of the entire New York City subway. The sample data itself has a slightly higher average amount of riders on rainy days.

However, in my linear regression model, the theta coefficient mapped to rain had a value of -23.5362258. When computing the estimated amount of riders per a turnstile data point with normalized values, this has an effect of lowering the rider count projection by 33.2 riders when raining and increasing the projection by 16.7 when it isn't! However, the theta coefficient mapped to fog has a magnitude of 24.2074176. As a result, data points that have fog were projected to have 54 additional riders while those that didn't were projected to have 10.8 less riders. Also lower temperatures, barometric pressure, and dew points tended to increase ridership projections. This is why I find it more appropriate to conclude that poor weather conditions in general tend to increase subway ridership.

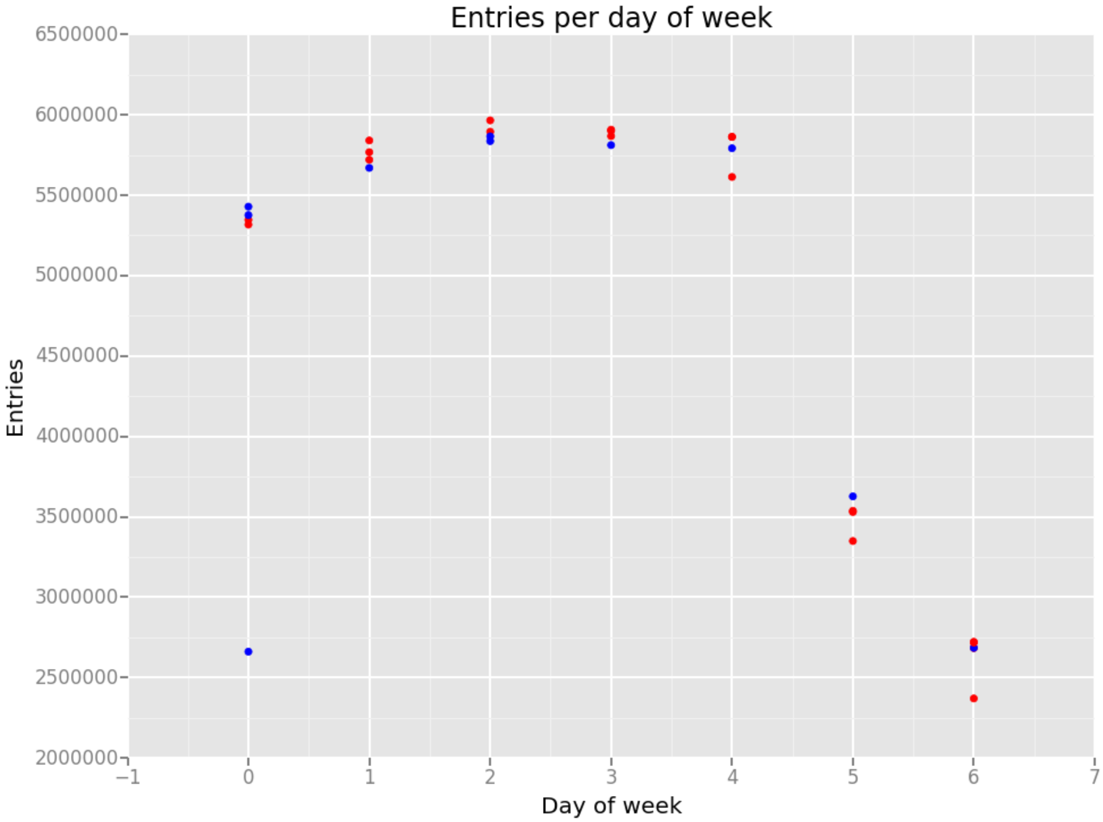
Normalized value for rain = 1.411311
Normalized value for no rain = -0.708556
Normalized value for fog = 2.232578
Normalized value for no fog = -0.447909

5.1
Please discuss potential shortcomings of the methods of your analysis, including:
Dataset,
Analysis, such as the linear regression model or statistical test.

The data set could be expanded to include more dates and/or turnstiles for a bigger sample. Also since my linear regression model had a R^2 value of about 0.47, there is significant room for improvement and there are obviously other factors that this data doesn't encapsulate that may affect subway ridership. Also the significance testing performed only proved that there was difference in the average amount of riders on the subway during rainy and non-rainy days. It doesn't actually conclude that there are always more riders on rainy days.

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

When you look at the total number of people riding the subway for each day of the month, you can see an increase of riders at the beginning of the week when it rains and a decrease at the end. Would need an expanded dataset and to perform some significance testing, but it appears that more people are likely to ride the subway when it rains at the beginning of the week than at the end.



Blue indicates a day where there was any rain and red indicates no rain at all.