Analyzing the NYC Subway Dataset

January 29, 2016

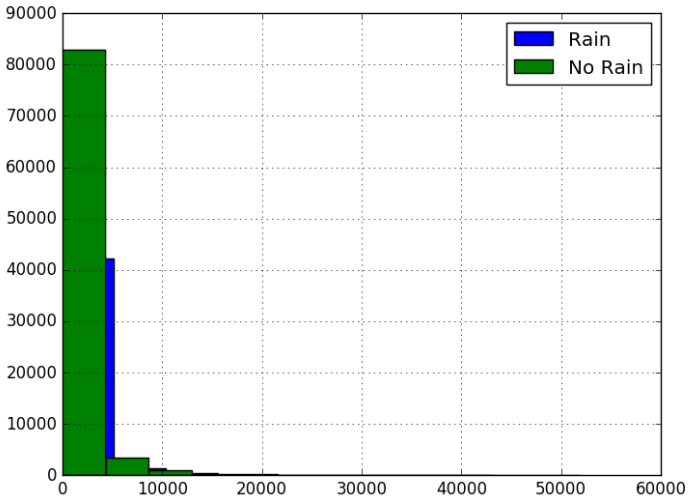
Project Overview

In this project, you look at the NYC Subway data and figure out if more people ride the subway when it is raining versus when it is not raining. You will wrangle the NYC subway data, use statistical methods and data visualization to draw an interesting conclusion about the subway dataset that you've analyzed.

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?

Using turnstile_data_master_with_weather.csv dataset, the histogram of hourly ridership below split by times it was raining versus when it wasn't indicates that this data doesn't seem to be normally distributed-



So we will use the Mann Whitney U Test to examine if more people ride the subway when it is raining versus when it is not raining. Let x be random draws from population X which is ridership on NYC subway per hour per turnstile when it's raining and y be random draws from population Y which is ridership on NYC subway when it isn't raning -

H_0: $P(x > y) = 0.5$

H_a: $P(x > y) \neq 0.5$

We use scipy's MWU two-tailed p-value test and the p-critical value is 0.05.

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 Mann Whitney U test is applicable to the dataset since the NYC subway ridership data is not normal. This statistical test is a non-parametric test since it does not assume the data is drawn from any particular underlying probability 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.

Based on the sample of NYC ridership collected for May 2011, the U statistic we get from the Mann Whitney U test is 1924409167.0 and the p-value is 0.0193. Two times p-value is 0.0386 which is less than the p-critical value of 0.05, so we can reject the null hypothesis at the 5% significance level. On average, there are 1,105 riders per hour per turnstile when it rains and 1,090 riders per hour per turnstile when it doesn't rain on the NYC subway. Hence, on average, there are 15 more riders per hour per turnstile when it rains compared to when it doesn't.

1.4 What is the significance and interpretation of these results?

These results indicate that the ridership in NYC subway is higher when it's raining than when it isn't and this difference is statistically significant. However, we can't infer from these results that rain is the only predictor for ridership in NYC subway. We now examine other factors that may affect how many riders ride in the NYC 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?

I used OLS using Statsmodels to compute the coefficients theta and predictions for ENTRIESn_hourly in my regression model.

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

I used day of the week, hour of the day, rain indicator, UNIT, weather conditions, daily average temperature, daily average wind speed in mph, daily average precipitation, daily average pressure for the location, temperature, wind speed, precipitation and pressure at that time and location. I used dummy variables for UNIT, day of week, hour of day and weather conditions.

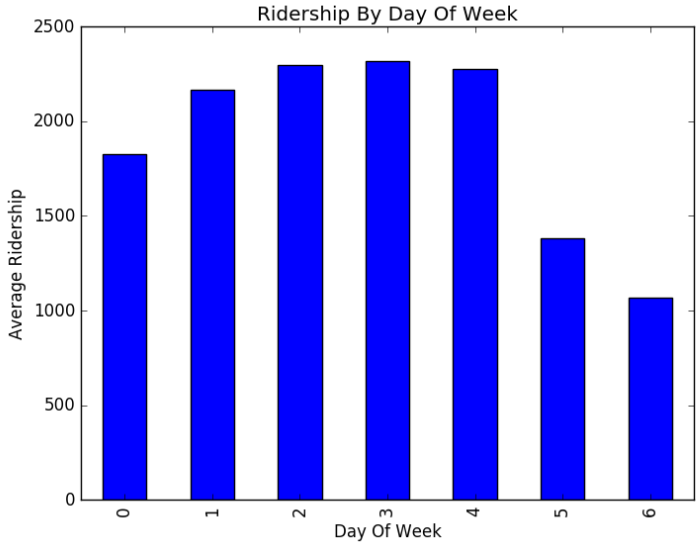
2.3 Why did you select these features in your model? We are looking for specific reasons that lead you to believe that selected features will contribute to the predictive power of your model.

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 R2 value."

Day of Week:

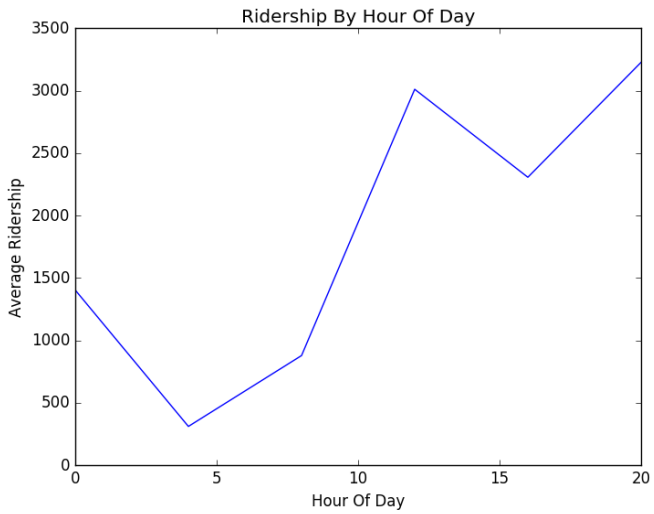
I selected day of the week since the bar plot below shows that ridership is lower on Saturdays and Sundays relative to week days-



I ended up using weekday dummy in my model instead which indicates whether the day of ridership is weekday (1) or weekend (0) because it improved the R-squared of the model.

Hour of Day:

I used hours because ridership is higher between hours 8 am in the morning to 8pm with a slight dip at 4pm as shown below -



I used hours as a dummy variable since it improved the R-squared of the model.

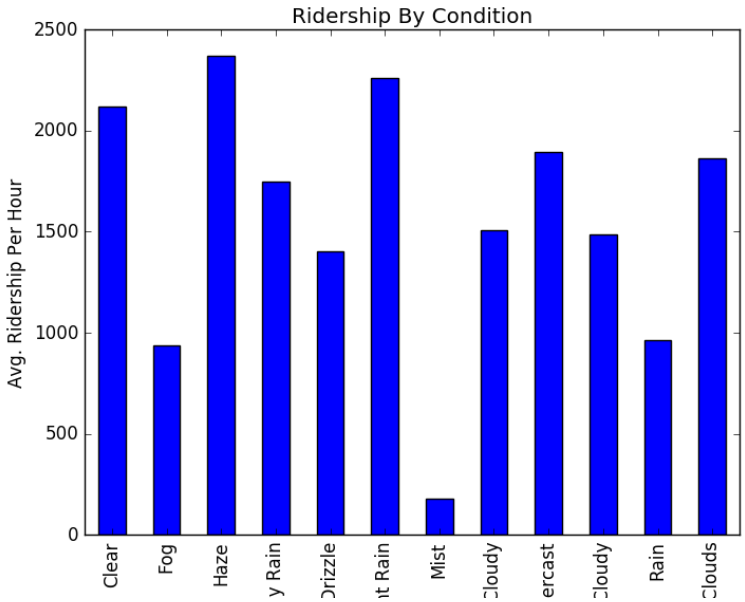
Rain:

I included the rain dummy variable since in section 1, using the Mann-Whitney test, in Section 1, I find that on average, ridership per hour in rainy days is higher than that in days when it doesn't rain and this difference is statistically significant.

UNIT:

The ridership varies across UNITS. The mean ridership across all UNITS in May 2011 was 1,841 riders per hour with a standard deviation of 1,788 riders per hour. The maximum ridership was 10,046 riders per hour. Hence, we use different UNITS as dummy variables in the OLS regression model.

Weather Conditions:



As the above graph shows, ridership per hour varies by weather conditions – according to this sample, the number of riders likely to ride the subway are maximum when the weather is hazy and minimum when it's misty. Hence, we use weather conditions as dummy variables in the model.

Temperature, Wind Speed, Precipitation & Pressure:

I used the daily average of temperature, wind speed in miles per hour, precipitation, pressure for the location as well as the temperature, wind speed, precipitation and pressure at the time and location in the model since each of these variables had a Pearson correlation coefficient with ENTIRES_hourly that was statistically significant. However, the magnitude of each of these coefficients wasn't very large. I also added these variables one by one to the model only if they improved the R-squared of the model from the previous version of the model.

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

The coefficient of weekday in the model is 1,024 which suggests there are likely to be approximately 1,000 riders per hour per UNIT more on weekdays than on weekends. The coefficient on hr_20 (8pm) is 3,931 riders per hour which is higher on average compared to the coefficient on hr_4 (4am) which is 987 riders per hour. The coefficient on daily average temperature and temperature at a specific time during the day is -5.18 and -17.62 respectively which suggests that as temperature increases the average ridership per hour decreases. The coefficient on daily average precipitation for a location is 5,744 which indicates that the average ridership per hour increases by 94 riders when the average precipitation increases by 1 SD (0.016). The coefficient on daily average pressure for the location is 1526 and that on the daily average wind speed is -26.17. The coefficient on pressure for a specific time and location is -2068 and that on wind speed for a specific time and location is 6.85.

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

The R2 for the regression model used is 54.63%

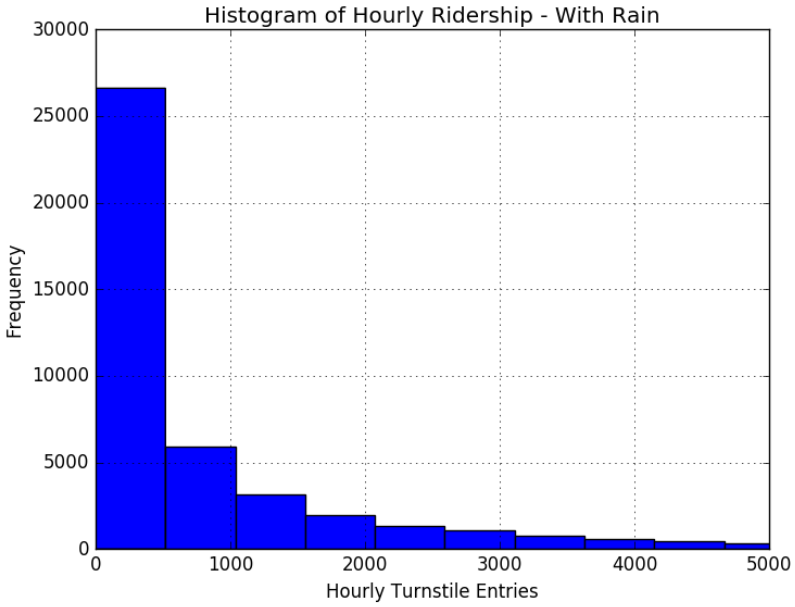
2.6 What does this R2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R2 value?

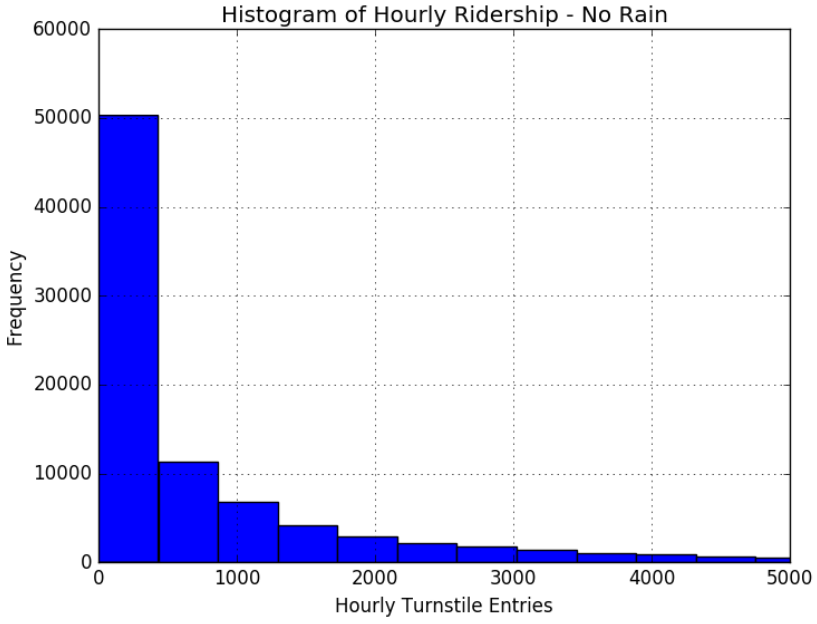
This R2 indicates that the goodness of fit for my regression model is 54.63% which means that 54.63% of the variation in the dependent variable – ENTRIES_hourly is explained by the features used in the model specification. Since only 55% of the variability in the dependent variable is explained by the linear model, I think this model is appropriate to predict ridership using this dataset but needs to be improved so the predictions of hourly ridership can be close to the actual values. The residual plot in Section 4 indicates only 82% of predictions are close to the actual values. This may be improved with additional data and other model specifications as indicated in Section 5.

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.

Below are two plots for histograms: one of ENTRIESn_hourly for rainy days and another for ENTRIESn_hourly for non-rainy days –

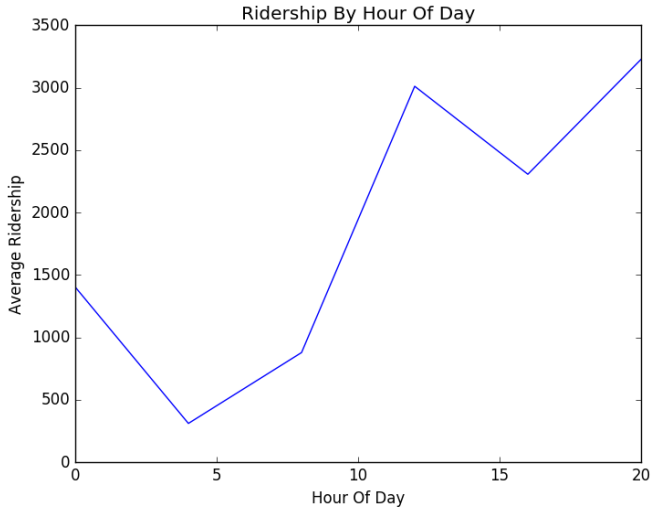




The mean ridership per hour when it rains is 1105 with standard deviation of 2370. The mean ridership per hour when it doesn't rain is 1090 with standard deviation of 2320. The median ridership per hour with rain is 282 and without rain is 278. The maximum ridership per hour with rain is 51,839 and that without rain is 43,199, hence, distribution of ridership with rain is more positively skewed than the distribution of ridership without rain. There are 44,104 observations with rain and 87,847 observations 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.

Below is the visual for average ridership per hour by time-of-day –



As we can observe in the line plot above, average ridership per hour varies by time of day. The maximum ridership is 3200 riders per hour at 8pm and the minimum ridership of 300 riders per hour is at 4am in the morning.

Section 4. Conclusion

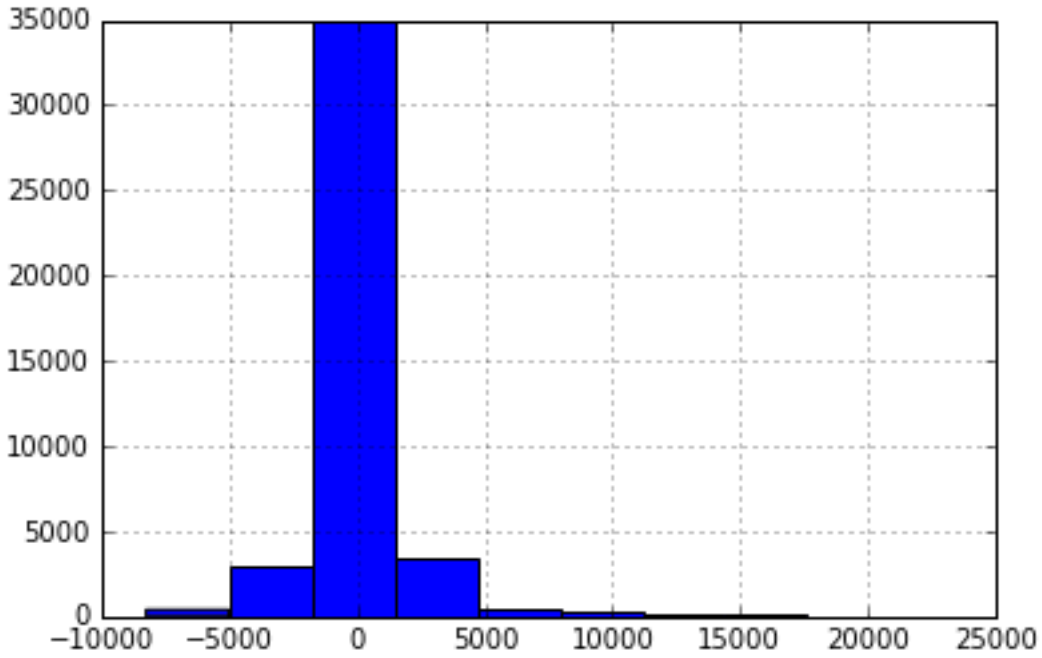
4.1 From your analysis and interpretation of the data, do more people ride the NYC subway when it is raining or when it is not raining?

From my analysis and interpretation of the data, more people ride the NYC subway when it is raining that when it isn't.

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.

Firstly, the dataset used for Mann-Whitney test and linear regression were different - turnstile_data_master_with_weather.csv and turnstile_weather_v2.csv respectively. Using the Mann-Whitney U test in section 1, we found that on average, ridership per hour on NYC subway is 15 riders more when it is raining than when it isn't. However, when I use the same dataset that I used for linear regression (turnstile_weather_v2.csv) for the Mann-Whitney test, I find that the average ridership per hour on NYC subway increases by 182 when it rains than when it doesn't. However, this statistical test doesn't control for other variables that may affect ridership on NYC subway.

Using linear regression we find, all else equal, when it rains, on average, 1121 additional riders ride the NYC subway per hour compared to when it doesn't rain. This correlation is even more pronounced when it rains heavily - all else equals when it rains heavily, on average, 1222 additional riders ride the NYC subway per hour compared to when it doesn't rain. When the weather condition is light rain, then there are 966 riders per hour in the NYC subway. I think the results from the linear regression are likely more robust since the model controls for weekday, hour of day, temperature, precipitation, wind speed in miles per hour and pressure, weather conditions and variation in traffic of riders by UNIT. In addition, the weather condition variable indicates clearly for the time and location whether there was light rain, rain or heavy rain which is more explicit that the rain dummy variable. Since the R-squared of the linear regression model is 54.63, this model also explains 54.63% of the variability in the ridership per hour on the NYC subway. Finally, the residual plot below indicates that most errors, that is, for most of the observations, predictions minus actual values of ENTRIESn_hourly below is zero-



82% of the observations in the dataset have residuals that are either close to or zero. The mean value of residuals is -1.152469e-09 with standard deviation of 1988 riders per hour.

Section 5. Reflection

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

This dataset turnstile_weather_v2.csv has 42,649 observations for the month of May in 2011. Using this dataset, we can control for the weekday and weekend effect but we can't control for holiday effect. We also can't control for the monthly seasonal patterns on the NYC ridership – for example, all else equal, the ridership per hour on NYC maybe affected in the summer months due to tourists visiting New York versus winter months. If we need to build a model that can predict the ridership per hour on NYC for the entire year, then we would need data for other months of the year for a couple of years at least to incorporate monthly seasonal patterns. This dataset also only has hourly ridership data for

every 4 hour interval – 0, 4, 8, 12, 16, 20. Hence, the linear regression model using this dataset may not capture the ridership patterns for other hours of the day.

The dependent variable in the dataset 'ENTRIESn_hourly' has outliers. It has a mean of 1,887 riders per hour and standard deviation of 2,952 riders per hour. When I group by Station, the average minimum ridership per hour is zero for the entire month of May which likely is incorrect data since every station must have a quantifiable amount of riders per hour. The maximum value for 'ENTRIESn_hourly' is 32,184 riders per hour per UNIT. Since the IQR for ENTRIESn_hourly is 1,981, the extreme outlier value for ridership per hour is 8,198. So we would need to examine the values that above the extreme outlier. If these are actual ridership values, then maybe we need to model the prediction of ridership by hour by stratifying this dataset. One way to stratify the sample could be to model small subway stations and large subway stations separately.

The Mann-Whitney U test doesn't control for any variables other than whether it rains or not. The linear regression model assumes that temperature, precipitation, pressure and wind speed at a particular time and location as well daily averages of temperature, precipitation, pressure and wind speed have a linear relationship with ridership per hour on the NYC subway. This is a huge assumption to make since temperature at a particular time and location as well as the daily average temperature for the location may not have a linear relationship with hourly ridership. So we could predict hourly ridership on NYC by using this dataset with other models such as random forest which is an ensemble of decision trees where for each decision tree, the cut-offs for continuous variables are decided based on the measures of node impurity such as Gini index.

References:

1. Introduction to Data Science @ Udacity
2. Mann-Whitney U Test