Analysis of New York Subway and Weather Data

Overview

The New York subway system is the seventh largest in the world in terms of ridership, carrying more than 1.5 billion passengers per year. It is comprised of 21 interconnected routes and serves the boroughs of Manhattan, Queens, Brooklyn, the Bronx, and Staten Island[1].

This article will analyze a subset of the ridership data, specifically the data from May 2011.

Basic Analysis

By examining some basic information obtained from the data, it can be seen that in the month of May, 2011, the New York subway carried **144,532,327** riders and the mean riders per day was **4,817,744.23**. FIgure 1 shows how the ridership was distributed over the month and is displayed alongside precipitation data to see if rain makes a difference in ridership.

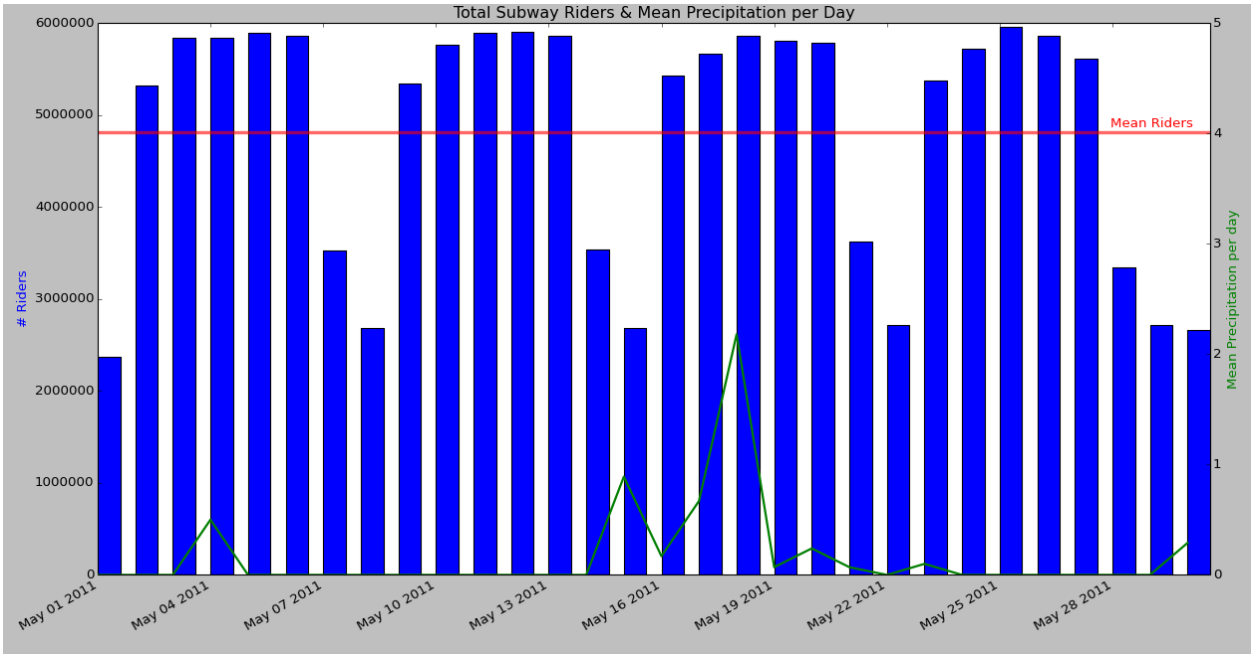

Figure 1 - Subway ridership & mean precipitation per day during May 2011.

From this chart it is not clear whether rain has an appreciable impact on the ridership or not. However, inspecting the histogram data for days when it was raining versus days when it was not raining, a difference is observed (figure 2).

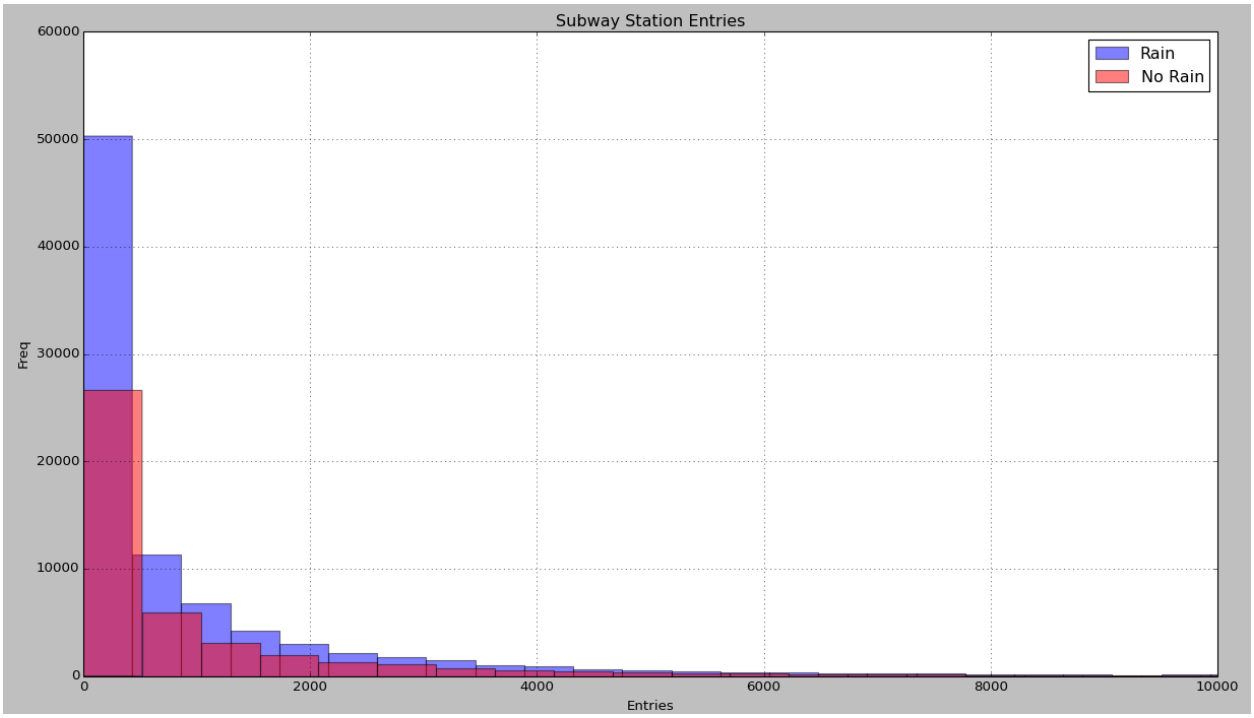

Figure 2 - Histogram of Subway Entries when raining versus when not raining.

This histogram indicates that the data is non-normal so analyzing further using a T-Test may yield inaccurate results. Hence, a **Mann Whitney U** test, a non-parametric test, will be used. The Mann Whitney U test is a test of the hypothesis that two populations are the same, the populations in this case being ridership when not raining and ridership when it is raining.

The test yields a U value of **1924409167.0** and a p value of **0.0193**. The mean values are **1105.45** for when there is no rain and **1090.28** when there is rain. Since the p value is < 0.05, the hypothesis that the two populations are the same can be rejected. That is, ridership when raining and ridership when not raining are different.

Further Analysis

While the data in this case is manageable, the MapReduce programming model can be used to distribute the large quantities of data across multiple computers and process that data in parallel. There are many frameworks and tools that can help implement MapReduce. For example, Hadoop is a framework for processing data that is distributed across multiple computers and Pig is one platform for creating MapReduce programs that can run on Hadoop.

Using a MapReduce algorithm on the subway data, the data can be distilled to display the number of turnstile entries per station (figure 3)

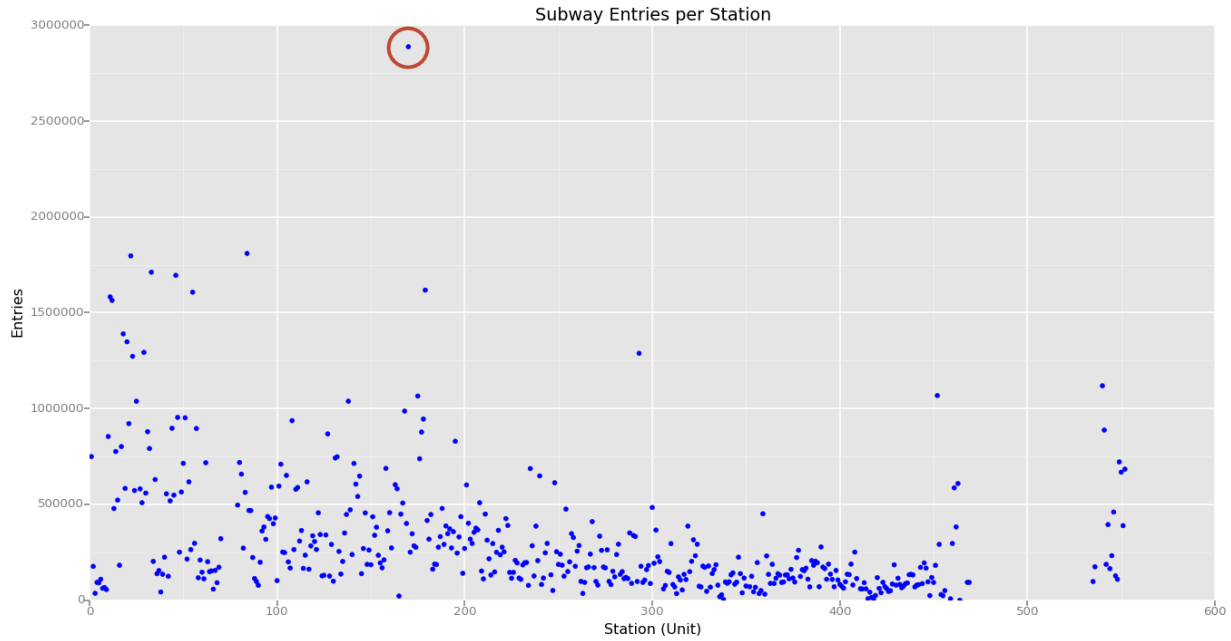

Figure 3 - Subway Entries per Station with station R170 circled

Interestingly, station R170 (circled) is an outlier with a total of 2,887,918. Station R170 is the 14th Street Union Square station in Manhattan[2] so it is understandable that such a centrally located station would have such a high number of riders.

Predicting Ridership

Utilizing machine learning techniques, a model for predicting the subway ridership can be created. In this case a Linear Regression with Gradient Descent technique was used with a features set comprising of **precipitation**, the **min**, **max** and **mean temperature**, **the hour of entry**, the **mean wind speed**, and whether it was **raining** or not.

The R^2 value is a measure of how of good the model is with values closers to 1 being better than those at the 0 end of the scale and in this case the resulting predictions had an R^2 value of **0.459**. Adding the number of **Exits per Hour** to the feature set increases the R^2 value to **0.6216**.

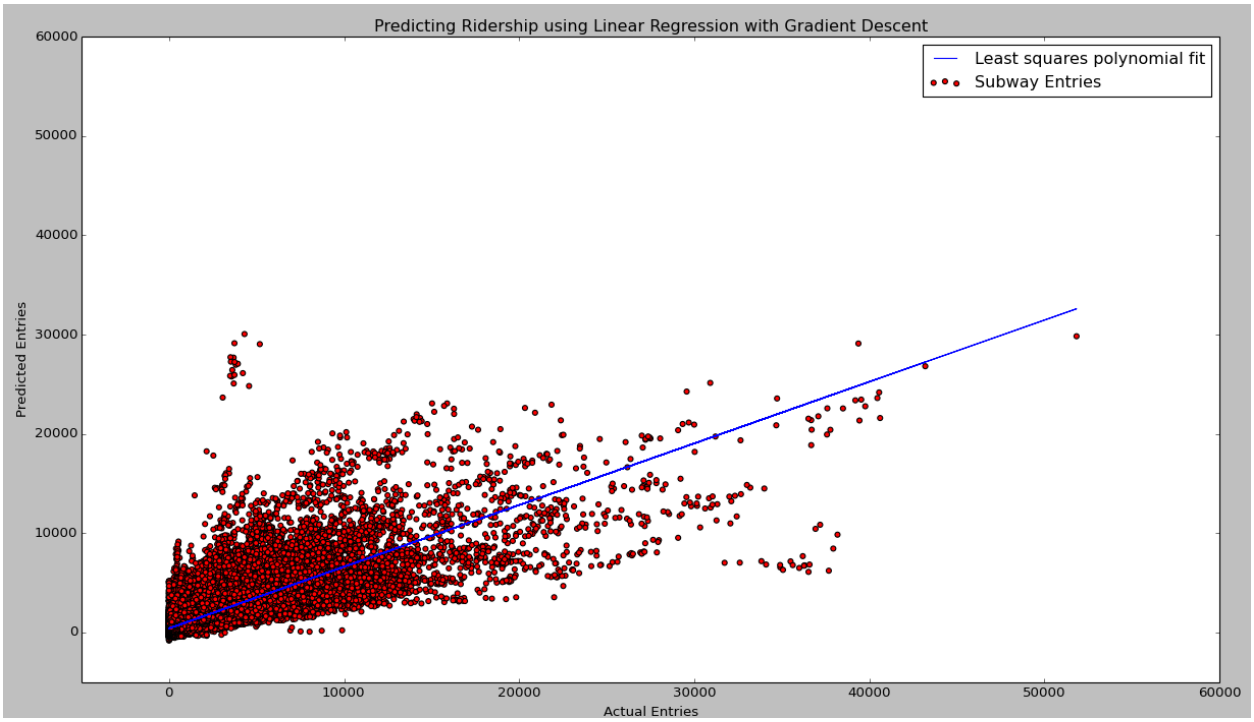

Figure 5 - Scatter plot of actual entries versus predicted entries when also using Exits in the features.

Analysis of the difference between original and predicted values indicates that the residuals mean is **0.029** with a standard deviation of **1437.93**.

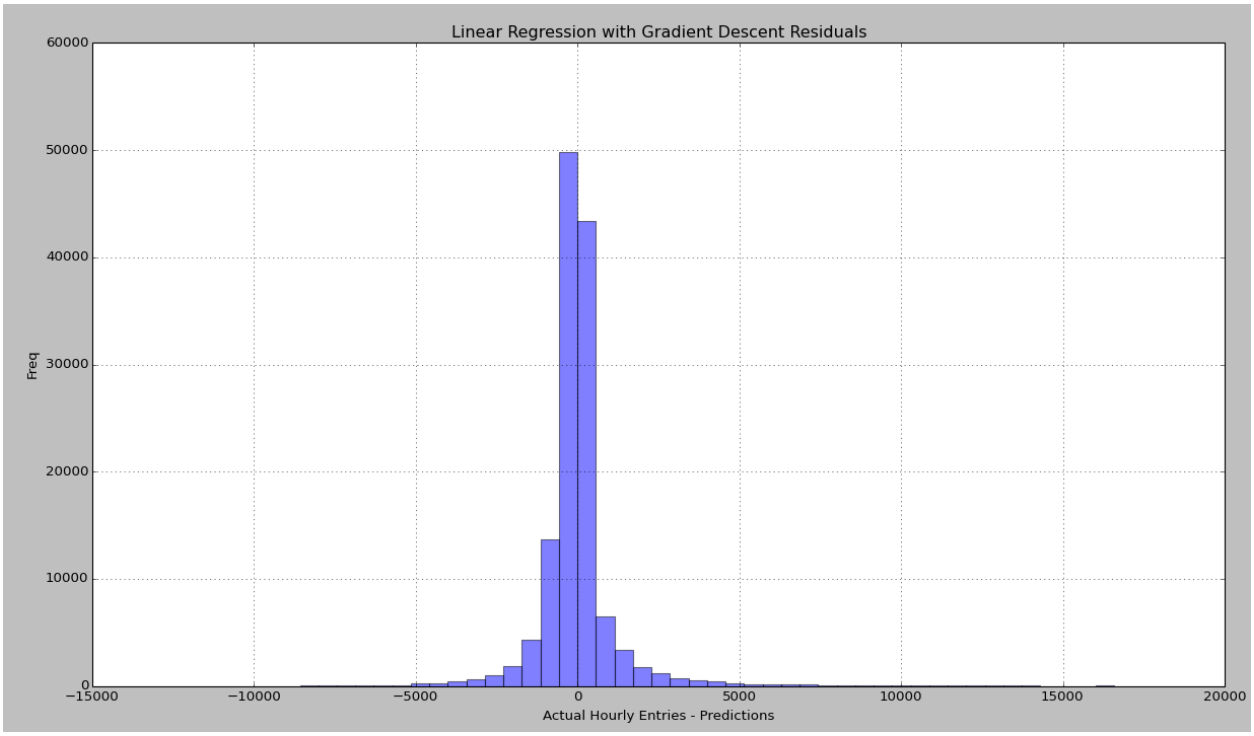

Figure 6 - Linear Regression with Gradient Descent Residuals Histogram

Using only data for station R170, the linear regression with gradient descent yields a model with

an R^2 value of **0.6811** when using the same features as was used for all stations (including exit data).

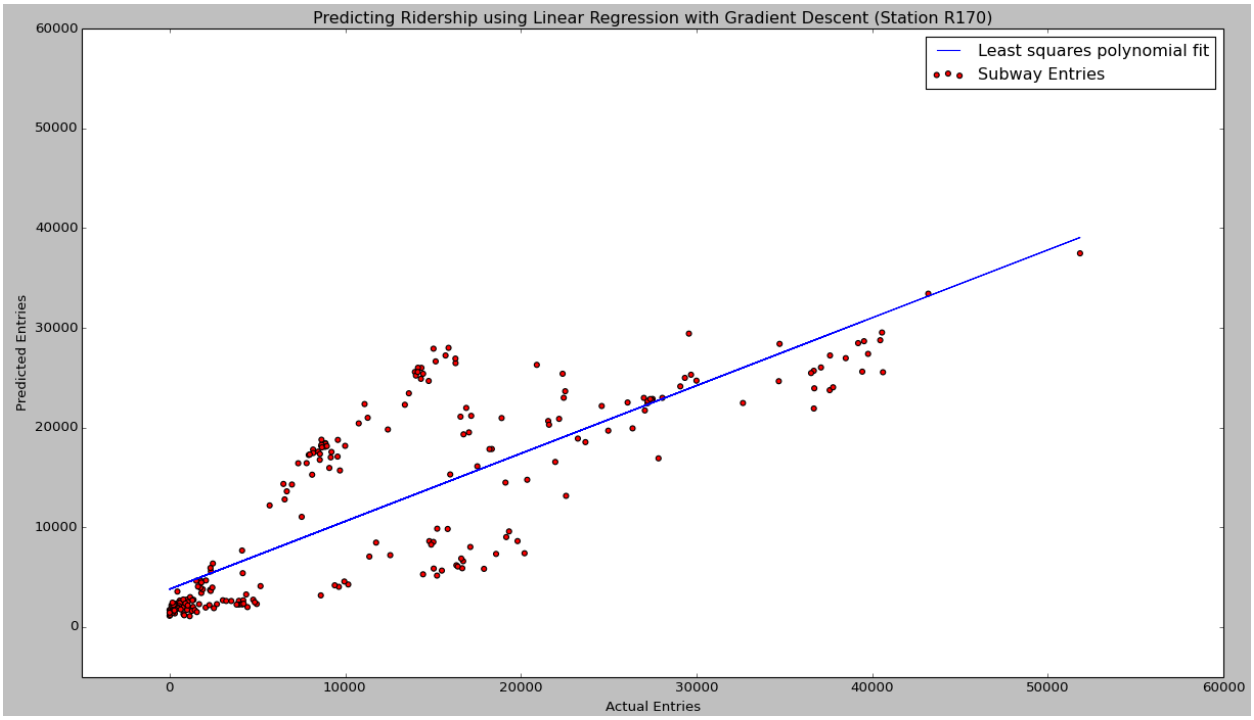

Figure 7 - Scatter plot of actual entries versus predicted entries for station R170

However, examining the residuals shows that this model may be more prone to errors with a residuals mean of **0.3157** and a standard deviation of **6655.7151**. This could perhaps be attributed to the smaller data sample being used to generate this model.

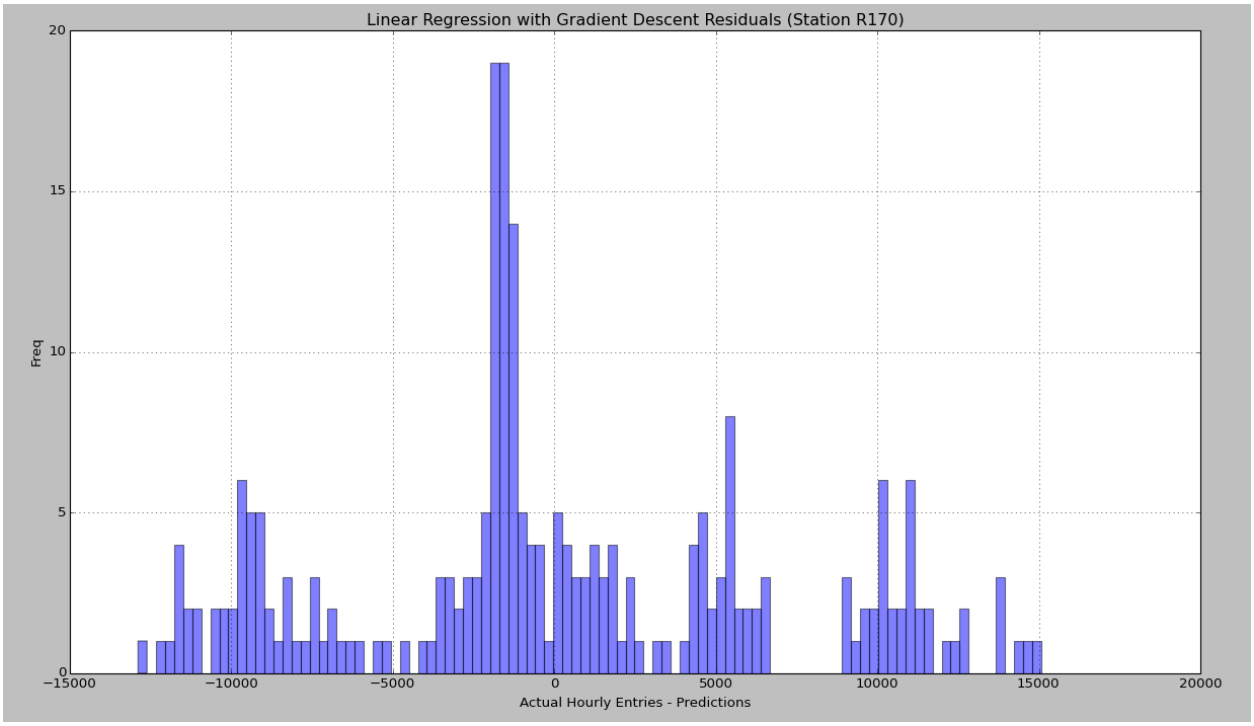

Figure 7 - Linear Regression with Gradient Descent Residuals Histogram for station R170

Improvements

There are a number of areas that could be explored to improve the accuracy of the analysis:-

1. Figure 1 indicates that ridership is different on weekdays than it is on weekends (or Memorial day). Taking this into account in the features used for the linear regression may yield a more accurate model.
2. Another option related to weekend data is to model it separate from the weekday data to arrive at two prediction models.
3. Station R170 is an outlier so removing this from the features may also lead to a better linear regression model.
4. Linear Regression with Gradient Descent is prone to errors if there are multiple minima so restarting with different parameters can help reduce the effect.
5. Different linear regression models, such as Ordinary Least Squares may also yield a better model.

References

[1] MTA Facts and Figures - http://www.mta.info/nyct/facts/ffsubway.htm
[2] MTA Booth Info - http://www.mta.info/developers/resources/nyct/turnstile/Remote-Booth-Station.xls