Introduction to Data Science

Analyzing the NYC Subway Dataset

Section 1. Statistical Test

1. The Mann Whitney U test was used to analyze the NYC subway data. A one tailed p-value was used for this test. The null hypothesis was that the ridership during rainy days is not statistically different than the ridership during non-rainy days.
2. This test is significant for this dataset because the dataset does not fit a normal distribution. The Mann Whitney U test is useful as a non-parametric test in which the the dataset can not be assumed to come from any particular distribution, i.e a Normal distribution.
3. The results from this test using the original data set were the following:

   ```
   a.  U statistic:                 1924409167.0
   b.  p-value:                     0.0193096344138
   c.  mean entries with rain:      1090.27878015
   d.  mean entries without rain:   1105.44637675
   ```

4. The significance of these results is that the null-hypothesis is rejected because the p-value is lower than the critical value of 0.05. This means that rainy days have a significant impact on the ridership.

Section 2. Linear Regression

1. I used both Linear Regression and Gradient Descent on the data set. I did this so I could compare resulting R^2 values to determine which would perform better.
2. I used all the available features in my model. For features variables that were dates and time values, I translated those values into absolute integer values based upon a base epoch value. For non-numerical data types I enumerated those values and converted those values to integer values.
3. I made no assumptions about which variables were relevant. I instead relied on Principal Component Analysis (PCA) to determine the predicted values. I imported the `sklearn` Python module and used the included `decomposition.PCA` sub-package. I chose this method because PCA is a well developed method for dimensionality reduction. This means I could include all the variables as a part of the model and the PCA method would:

   > for a set of observed d-dimensional data vectors $\{ t_n \}$, $n \in \{1 . . . N\}$, the q principal axes w_j , $j \in \{1 . . . q\}$, are those orthonormal axes onto which the retained variance under projection is maximal. [1]

4. The resulting R^2 value for Gradient Descent was 0.544397104718 and the resulting value for Least Squares was 0.552180456988. These values were also derived from the use of the original dataset.

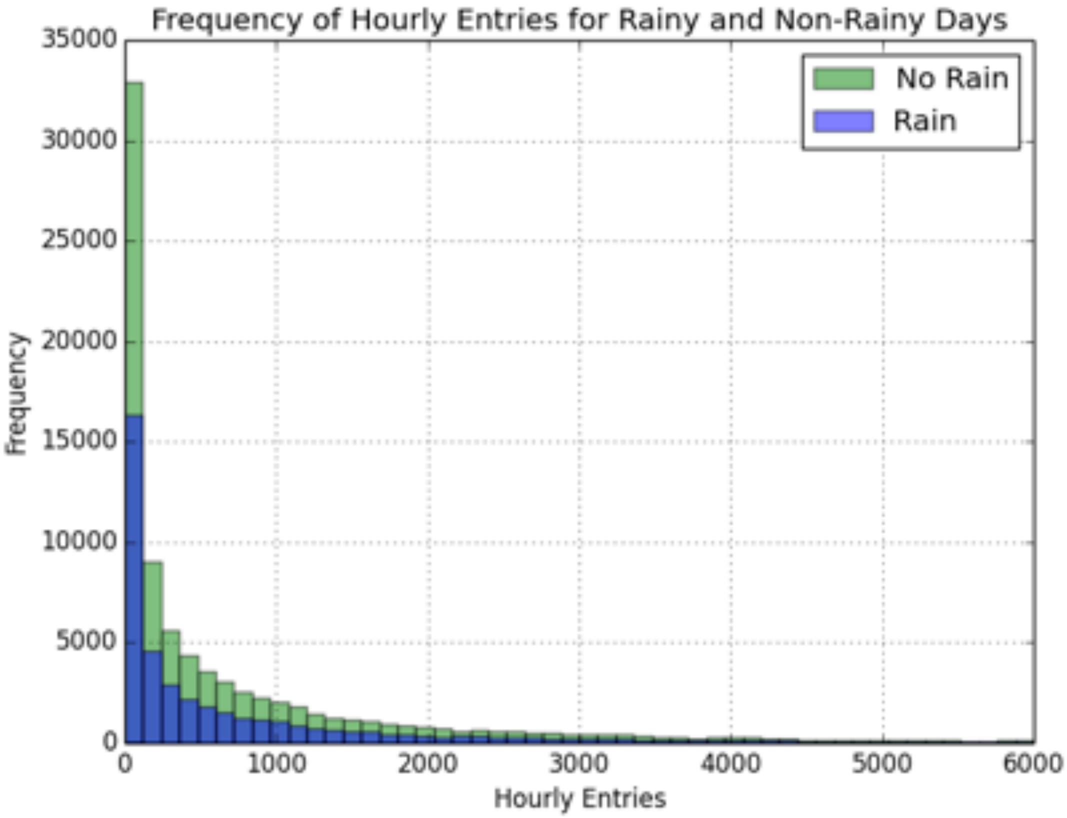
   > The coefficient of determination R^2 is a non dimensional measure of how well a regression model describes a set of data: $R^2 = 0$ when taking the information in the independent variables into account does not improve your ability to predict the dependent variable at all, and $R^2 =1$ when it's

> possible to predict perfectly the dependent variable from the information in the independent variables. [2]

In short, the R^2 value is a valid test for the appropriateness when compared to other linear regression models. In this case, the R^2 values is appropriate, but only when comparing the result of the Gradient Descent method with the result of the Least Squares method.

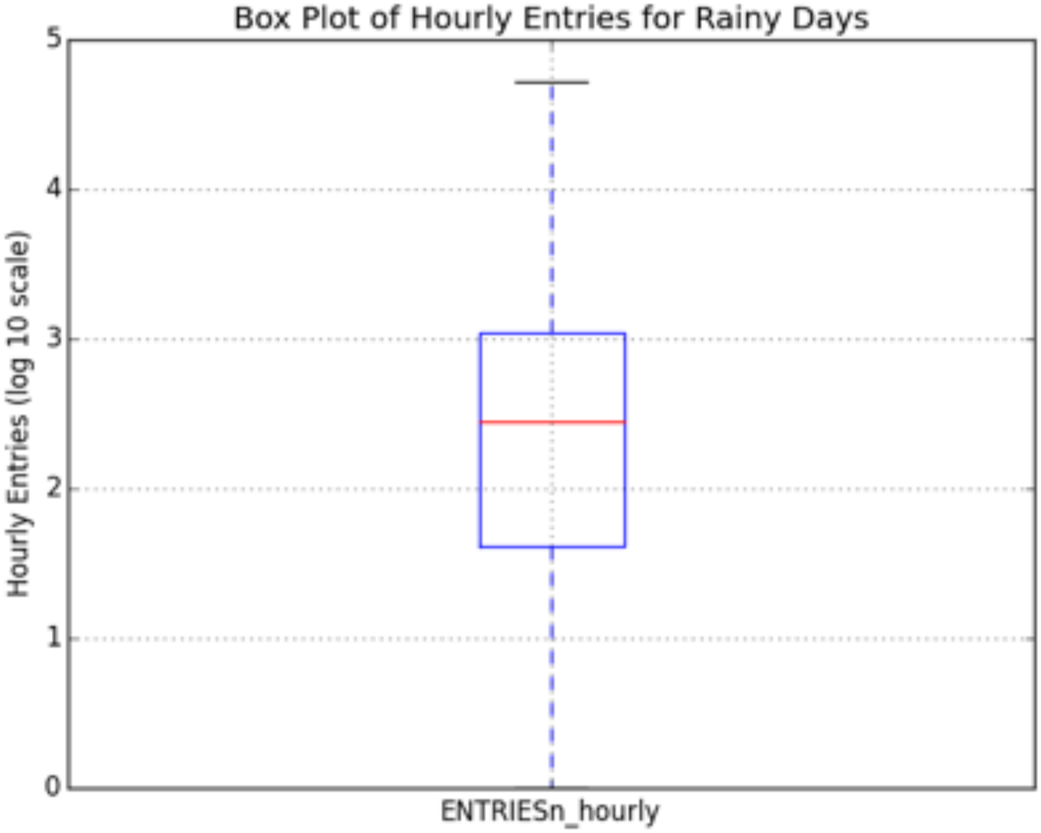
Section 3. Visualization

1. This plot is a histogram of the ridership during rainy days and non-rainy days. This plot suggests that rainy days may have an impact on the subway ridership since the rainy days appear to have fewer entries per hour than non-rainy days.

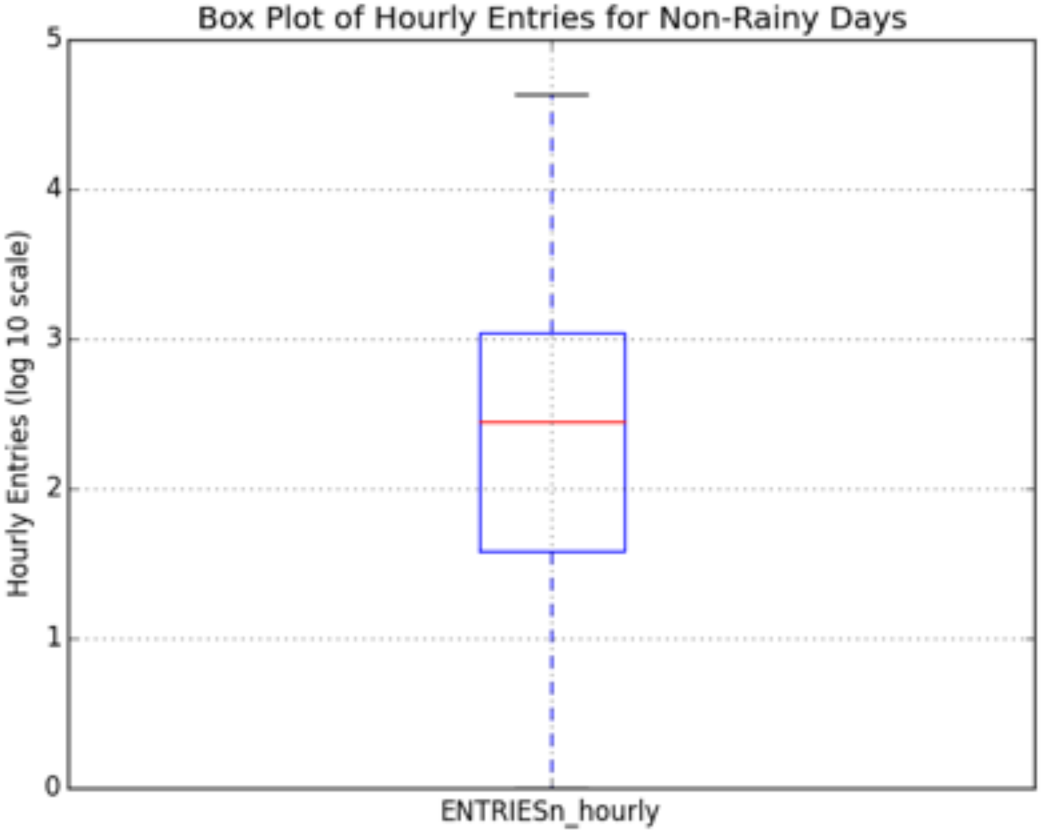


2. These two plots are box plots of the hourly entries for non-rainy days and rainy days. These plots indicate that hourly entries with large numbers of ridership are in the extreme and are therefore outliers. These indicators are true for both rainy and non rainy days.[1]

[1] The graphs for the rainy days and non-rainy days appear to be exactly the same. However, the value of each are different, but the variance between the two is relatively small.



Box Plot of Hourly Entries for Rainy Days



Box Plot of Hourly Entries for Non-Rainy Days

Section 4. Conclusion

1. I would conclude that fewer people ride the subway during rainy days than non-rainy days.
2. The histogram indicates that rainy days and non-rainy days may come from the same distribution based upon their shapes. However, the null hypothesis, that rainy days and non-rainy are derived from the same probability distribution, is rejected based upon The Mann-Whitney U test with a p-value of 0.05.

Section 5. Reflection

1. I investigated selecting certain features, in addition to rain, such as fog, temperature, and weekday, and excluding others such as station. Being selective with these features didn't seem to have a tremendous impact on the accuracy of the predictions. Hence, I used Singular Value Decomposition to include all the values into the prediction model. Interestingly, I tried the same process using the improved data set and found that the results did not differ greatly.
2. My original investigation included the determination of whether foggy days had an impact on ridership in the same way as rainy days. The resulting p-value from the Mann Whitney U Test was 1.95706170955e-06, which would lead to a conclusion of rejecting the null hypothesis since that p-value is well below the 0.05 critical value threshold. I would conclude that fog has a significant impact on the ridership in the same manner as rain. I would further estimate that inclement weather in general has an impact on ridership, suggesting that during these types of days riders would likely stay inside or choose another method of travel.
3. The potential shortcomings of this dataset and the statistical methods used to analysis this dataset include:
   a) The dataset only includes information about the turnstiles. For example, I would hypothesis that if NYC was hosting a large event, like the News Years' Rose Parade or World Series, then this would affect the ridership significantly[3]. I would suggest that examining these events would lead to better predictions, especially given that NYC is known as one the central hubs for commerce, arts and culture, and travel in the world.
   b) Use of the R^2 method to calculate the goodness of fit of the model. The problem is "that R^2 always increases as more variables are added to the regression equation, even if these new variables add little new independent information." [2]

Citations

1. Tipping, Mike E., and Christopher M. Bishop. "Mixtures of Probabilistic Principal Component Analysis." Neural Computation 11.2 (1999): 443-82.
2. Glantz, PhD, Stanton A., and Bryan K. Slinker, DVM, PhD. "Selecting the "Best" Regression Model." Applied Regression and Analysis of Variance. McGraw-Hill, 2013. 248-149.
3. "Subway Series." Wikipedia. Wikimedia Foundation. Web. 3 Apr. 2015. <https://en.wikipedia.org/wiki/Subway_Series>.