Analyzing the NYC Subway Dataset

Adam Wright

November 2, 2015

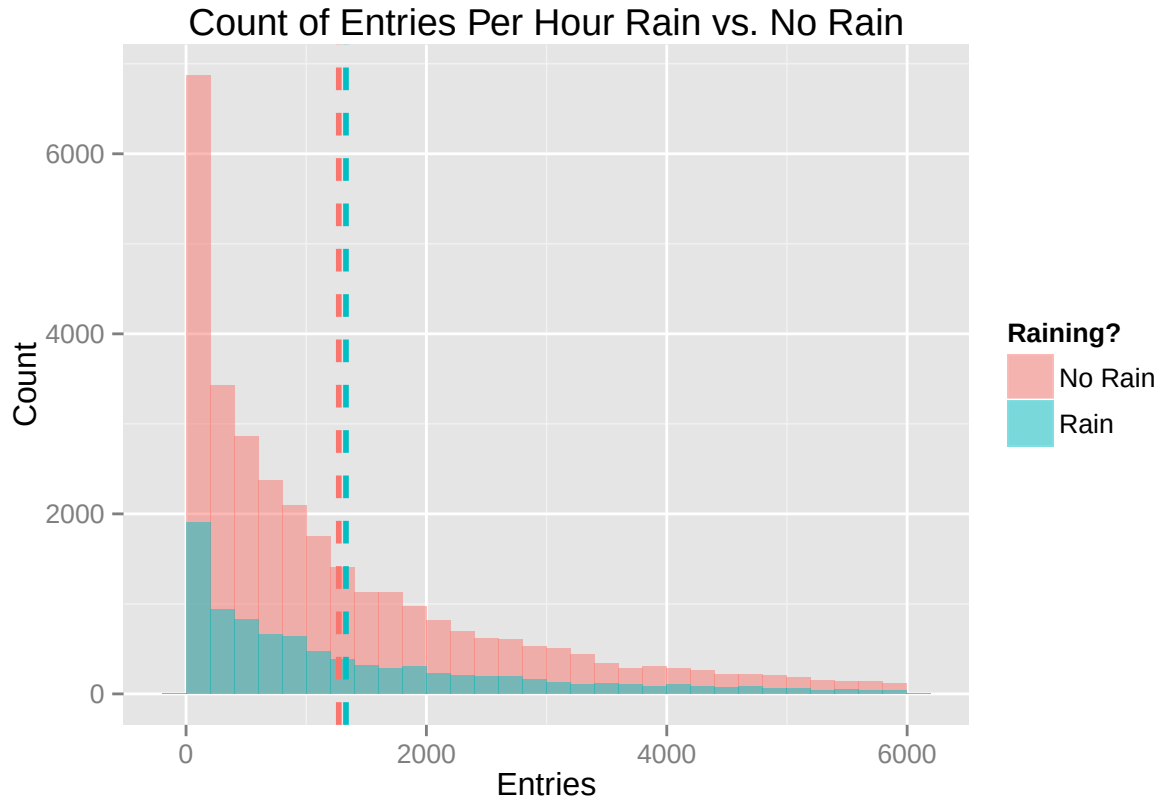
0. References

Please include a list of references you have used for this project. Please be specific - for example, instead of including a general website such as stackoverflow.com, try to include a specific topic from Stackoverflow that you have found useful.

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?

In order to determine what kind of statistical tests (parametric vs. non-parametric) can be applied to a given dataset it is first neccessary to determine whether that data is normally distributed. As can be seen in the figure, below, hourly entries into the NYC subway are definitely not normally distributed, displaying significant positive skew on both rainy ($skew = 1.33$) and non-rainy ($skew = 1.40$) days:



Because the NYC subway entry data is not normally distributed, it is neccessary to use a non-parametric test to test for significant differences between entry conditions (e.g. rainy vs. non-rainy days). In particular, to determine whether hourly entries were significantly different on rainy and non-rainy days **I choose to use the Wilcoxon rank-sum test**, assuming independent samples (i.e. population entering subway on rainy and non-rainy days were not identical).

The null hypothesis for my test is that there will be no significant difference between the number of hourly entries on rainy and non-rainy days. Because my intuition is that people will use the subway more on rainy days in order to avoid getting wet, I used the the **one-tailed p-value for my test, that is that the number of entries on rainy days will be significantly higher compared to non-rainy days**. Finally, I chose to use a **standard** $\alpha = 0.05$.

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.

As described above, the Wilcoxon rank-sum test is applicable to this dataset because the data is not normally distributed. The Wilcoxon rank-sum test also depends on an assumption of homogeneity of variance which can be quantified using Levene's test:

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.

1.4 What is the significance and interpretation of these results? 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: OLS using Statsmodels or Scikit Learn Gradient descent using Scikit Learn Or something different? 2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features? 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 R2 value."
2.4 What are the parameters (also known as "coefficients" or "weights") of the non-dummy features in your linear regression model? 2.5 What is your model's R2 (coefficients of determination) value? 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? Section 3. Visualization

Please include two visualizations that show the relationships between two or more variables in the NYC subway data. Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure. 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. 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 Section 4. Conclusion

Please address the following questions in detail. Your answers should be 1-2 paragraphs long. 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?
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.

Section 5. Reflection

Please address the following questions in detail. Your answers should be 1-2 paragraphs long. 5.1 Please discuss potential shortcomings of the methods of your analysis, including: Dataset, Analysis, such as the linear regression model or statistical test. 5.2 (Optional) Do you have any other insight about the dataset that you would like to share with us?