Brendan Schell 14Oct2015

Analyzing the NYC Subway Dataset

Questions

Overview

This project consists of two parts. In Part 1 of the project, you should have completed the questions in Problem Sets 2, 3, and 4 in the Introduction to Data Science course. This document addresses part 2 of the project. Please use this document as a template and answer the following questions to explain your reasoning and conclusion behind your work in the problem sets. You will attach a document with your answers to these questions as part of your final project submission.

Section 0. References

N/A

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?

To analyze the NYC subway data, a two-tailed Mann Whitney U test was performed.

The null hypothesis is that if given a random draw from the population of entries on a rainy day, it is equally likely that the distribution will be greater than or less than that on a non-rainy day. A p-critical value of 0.05 was used (0.025 for each tail).

$$H_0: P(x_{rain} > x_{no\ rain}) = 0.5$$

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.

This statistical test is applicable to the dataset since it is not known whether the data follows a normal distribution or not. Due to this, normality is not assumed and the Mann Whitney U test is used since it is non-parametric. As well, this test does not assume that the samples are of a similar sample size which is ideal since the ratio of rainy to non-rainy days are possibly dissimilar which could lead to different sample sizes for each. A two-tailed test was performed since the null hypothesis seeks to prove that there is no difference in values (either positive or negative) between the two populations.

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.

$$\mu_{with\_rain} = 1105.446, \mu_{without\_rain} = 1090.279, p = 0.024999\ (one - sided)$$

1.4 What is the significance and interpretation of these results?

The significance of these results is that the two-sided p value (0.024999*2 = 0.049998) is below the p-critical value of 0.05 and therefore, the null hypothesis is not retained for a 95%

confidence level. These results can be interpreted as there being less than a 5% probability that the difference in the two distribution samples are explained by the variation in the population distribution alone.

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:

1. **OLS using Statsmodels or Scikit Learn**
2. Gradient descent using Scikit Learn
3. 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?

The input variables that were used in my model were 'rain', 'precipi', 'Hour', 'meantempi',’fog’, ‘meanwindspdi’, and a dummy variable for each unit value in the ‘Unit’ column.

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.”**

Rain : I decided to use rain since I figured that there would be more people using the subway when it rains.

Precipi: I used precipi since it made sense to me that more people would use the subway if it rained more substantially (i.e less ridership if only light rain).

Hour: I used the hour of the day since logically there would be more ridership during rush-hours than non-peak times.

Meantempi: I used the mean temperature since it made sense to me that there would be more ridership during colder temperatures than warmer temperatures when more people might walk or bike instead.

Fog: I used fog since I thought that more people might opt to use the subway instead of driving if it were foggy.

Meanwindspdi: I used the mean wind speed since I figured that fewer people might use the subway if it were very windy.

Unit (dummy variables): I used the units in my model since it makes sense that there would be different ridership numbers based on which unit was being looked at. As well, including it in the model significantly improved the R^2 value.

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

The parameters of the non-dummy features in the linear regression model were as follows:

```
rain              0.040560
precipi         -73.976935
Hour             65.364525
meantempi        -9.491420
fog             214.086883
meanwindspdi     32.568316
```

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

The model's R^2 value was equal to 0.480456675828.

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?

This R^2 value means that 48% of the variation can be explained in terms of the parameters that were modeled. Since an R^2 value of 1 would mean a direct correlation and 0 would mean no correlation, I think that this linear model has some significance, but that there must be other factors that were not considered in this model. It is also possible that this model does not follow a linear relationship and that there could be a different type of relationship between the data (non-1 exponents on the features). Since the R^2 value is not close to unity, I do not think this linear model is appropriate to predict ridership.

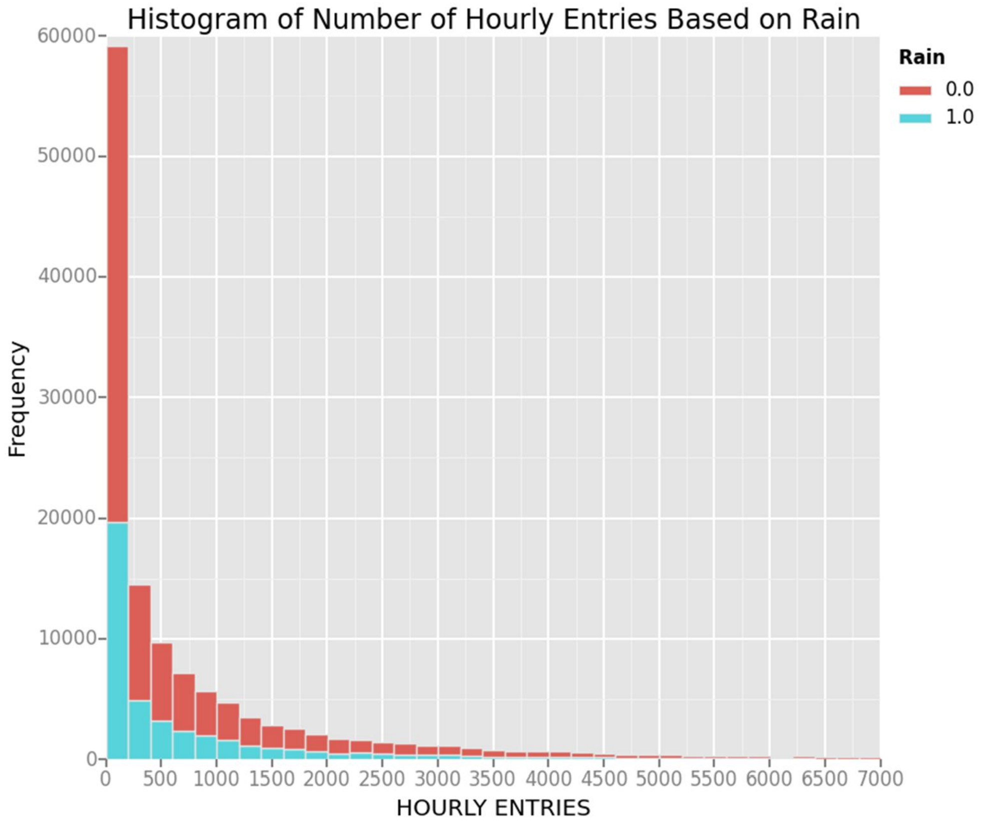
Section 3. Visualization

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.

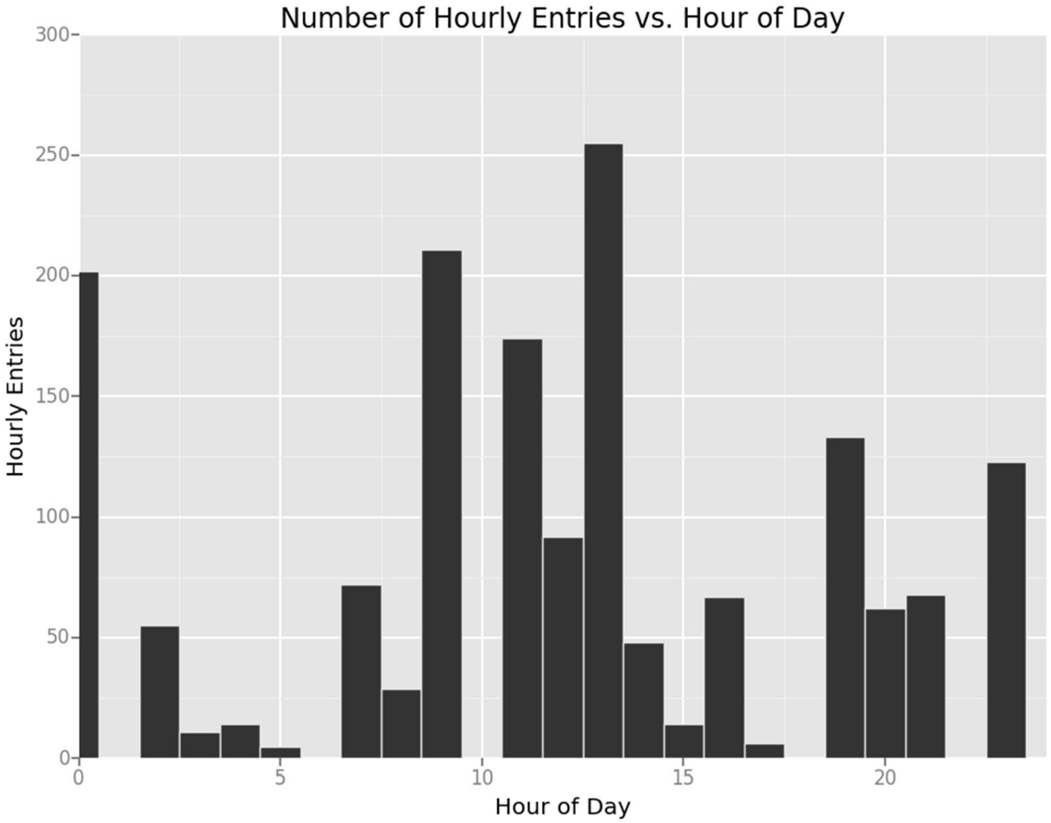
- **Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.**



In this visualization, the frequency of the Number of subway entries per hour is binned by the number of hourly entries when it is raining (blue) and when it is not raining (red). These entries are stacked on top of eachother. It can be seen that there is a much larger frequency of events where it is not raining and it is therefore difficult to compare the data in this form. As the number of hourly entries increases, the proportion of rain to non-rain appears to increase, demonstrating that a larger proportion of the rain events result in high traffic.

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**



In this visualization, the number of hourly entries is displayed against the hour of the day from hour 0 to 24. It can be seen that ridership is at a maximum during the middle of the day and decreases throughout the evening reaching a minimum at hour 5 at which point it increases again. This is in line with what I would expect given that rush hours would tend to be in the middle of the day.

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?

From my analysis, it seems likely that more people ride the subway when it is raining than when it is not. In section 1.3, the population mean ridership was determined to be larger when it is raining compared to when it is not with a 95% level of confidence (mu_rain > mu_no rain, p < p-crit, p > 0). Though this does not guarantee that the increase in values is not explained by the natural variation in the distribution, it means that there is less than a 5% probability that they are equal.

As well, this is supported by visual 3.1 that shows proportional increases in rain to non-rain frequencies as the number of high entries increases. Based on the previously completed analyses, it seems likely that more people ride the subway when it is raining than 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.

The analysis that lead me to this conclusion was the Mann Whitney U test that was performed in section 1.3.

Results:

$$\mu_{with\_rain} = 1105.446, \mu_{without\_rain} = 1090.279, p = 0.024999\ (one - sided)$$

Since the population mean when raining was larger than the population mean when not raining and the p-value was positive and less than the p-critical value of 0.05 (when multiplied by two to account for a two tailed test), this test demonstrated that there was less than a 5% probability of the difference being accounted for by the distribution variation.

As well, in section 2.1 it was determined that rain had a positive coefficient for the linear regression model, meaning that is positively correlated to subway ridership:

Coefficient for rain: 0.040560, R^2 = 0.480456675828.

One interesting thing to note is that there is a large negative coefficient for precipi in the linear regression that was performed. This is possibly due to other coefficients such as rain, fog, and hour that are largely positive accounting for some of the effect that would otherwise be seen in precipi.

As the R^2 value was equal to 0.48, it shows that the linear regression did have some correlation with the subway ridership, even if it did not account for the majority of the variation (significantly less than 1). From these results, it can be shown that there is a high likelihood of a positive correlation with rain and subway ridership.

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:

1. **Dataset,**
2. **Analysis, such as the linear regression model or statistical test.**

One potential shortcoming of the dataset is the possibility of erroneous data (we did not do a very elaborate imputation process or sanity check on data values). As well, weather data was generalized to all of New York even though New York is very large and likely did not always receive uniform rain dispersion. This could be alleviated through better weather data based on the station of entry instead of generalized New York data.

As far as the linear regression is concerned, the coefficient of determination was only 0.48 and was not very close to 1. As a result, the model is not necessarily the best predictor of subway ridership since it does not account for 52% of the variation of the data. As well, the statistical test resulted in a p-value that was smaller than the p-critical value, but not by a significant margin. When considering potential sources of error such as those mentioned previously, it is very possible that slightly different source data could have produced a different outcome in the statistical test. This could be evaluated using a sensitivity analysis.

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