ANALYZING THE NYC SUBWAY DATASET

Udacity Project 2

JANUARY 24, 2016

AKSHITH R KANDAKATLA

Table of Contents

Questions

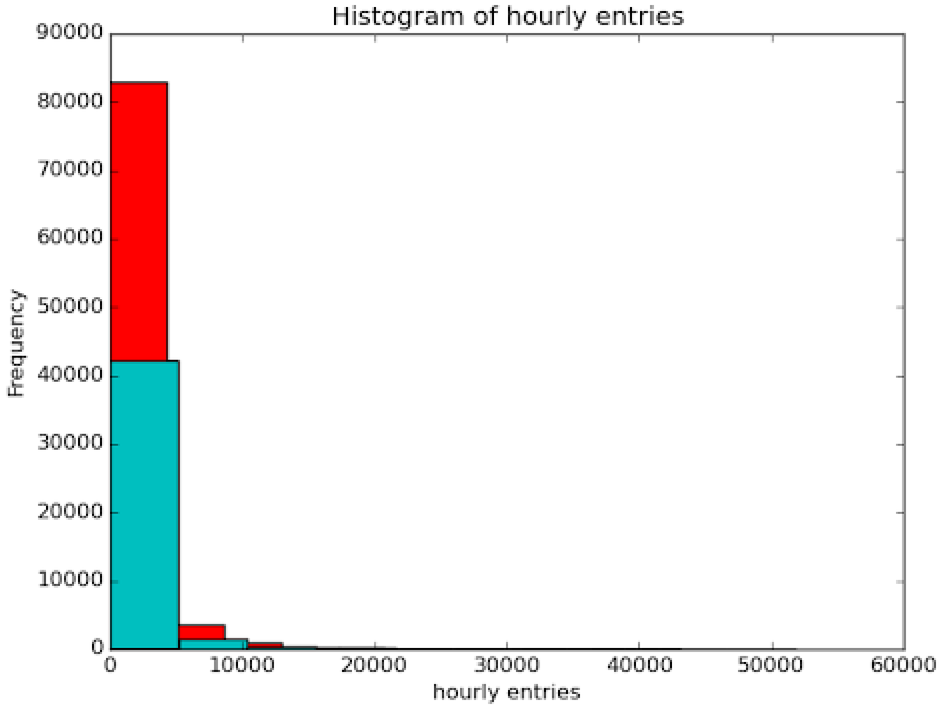
Section 1. Statistical Test

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?

Ans: I used the Mann-Whitney U test to analyze the NYC subway data to determine the significance of rain on ridership. I used a two-tailed test here to signify the difference between rainy & non-rainy circumstances. The null hypothesis is that there is no relation between rain and ridership. The p-critical value is 0.05 which indicates that the null-hypothesis will be rejected 5% of time when true.

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.

Ans[i]: We can observe that the data is not normally distributed from the following plot.



Thus in this case, a non-parametric test such as the Mann-Whitney U test is applicable. Another reason for chosing a non-parametric test is because the feature being analyzed(ENTRIESn_hourly) has only two groups and is ordinal in nature as it has a ranking but no clear interpretation can be made based on it. Some of the other assumptions that the test makes about the distribution are:

- All the entries in the data are independent
- Each data entry is distinct & continuous

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.

Ans: The results from the Mann Whitney U test read as follows:

```
(1105.4463767458733, 1090.278780151855, 1924409167.0, 0.024999912793489721)
```

This indicates that:

- Mean entries with rain: 1105.446
- Mean entries without rain: 1090.279
- U-statistic: 1924409167.0
- p-value: 0.025 (especially since it is a two-tailed test)

1.4 What is the significance and interpretation of these results?

Ans: From the above results we can clearly observe that the mean entries during rain is more than that of the entries without rain. However, this statistic (mean value) alone would not be enough to make a conclusion that the rain causes more ridership.

The Mann Whitney U test tells us that the p-value of 0.025 (two-tailed) satisfies the p-critical value of 0.05. Hence, we can reject the null hypothesis with a 95% confidence level and go with the alternative hypothesis that ridership is affected by rain.

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?

Ans: I have used a linear regression model to compute the coefficients theta. In order to produce a prediction for the chosen features, I used the ordinary least squares method imported from the statsmodel api

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

Ans: The selected features were, rain, precipi, Hour, meantempi, meandewpti, fog, meanpressurei & meanwindspdi. The dummy variables used were UNIT as it was categorical in nature.

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.”

Ans: Initially I had selected all the features assuming that selecting all features would lead to a better model but, as a models efficiency is based on its R^2 value, by trial & error method I kept interchanging the features until the R^2 value improved. While working on this I came across a thought that the ridership might have been affected by the rain but it would have been affected by other features also and hence I added more of the weather related features such as mean pressure & mean wind to conclude that not just rain but bad weather might have been the reason for an increase in the ridership

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

Ans: The coefficients of the non-dummy features in the linear regression model are as follows:

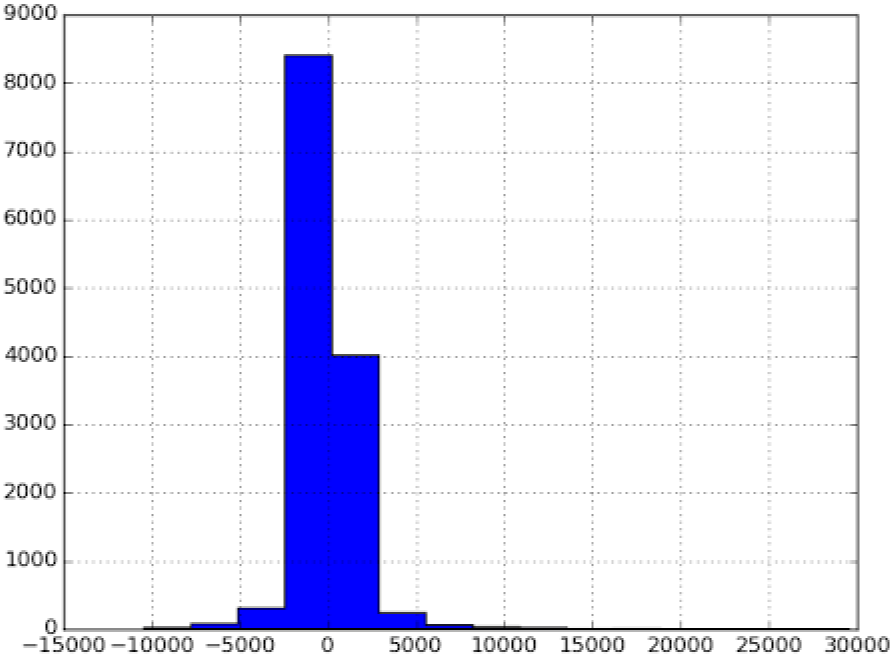
```
Feature           Coefficients
rain                16.186629
precipi            -66.773911
Hour                65.372310
meantempi           -7.016610
meandewpti          -2.719888
fog                202.102007
meanpressurei     -180.820099
meanwindspdi        29.990626
```

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

Ans: R2 value of this model is 0.480589270457.

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?

Ans: The coefficient of determination value explains the goodness of fit for 48.05% of the variation i.e. it is the prediction error is 1-0.4805=51.95% of the variance of the data. Plotting the residuals will give further insights into whether about the metrics as follows:



As seen above we can notice that a majority of the ENTRIESn_hourly values are in the range -5000 to 5000. We can also infer that the model predicts the ENTRIESn_hourly better when the values are small. Thus based on the coefficient of determination value and the residual plot we can conclude that though the linear model can be used to estimate the ridership it would not be suitable to make predictions about the ridership for this dataset given the R^2 value.

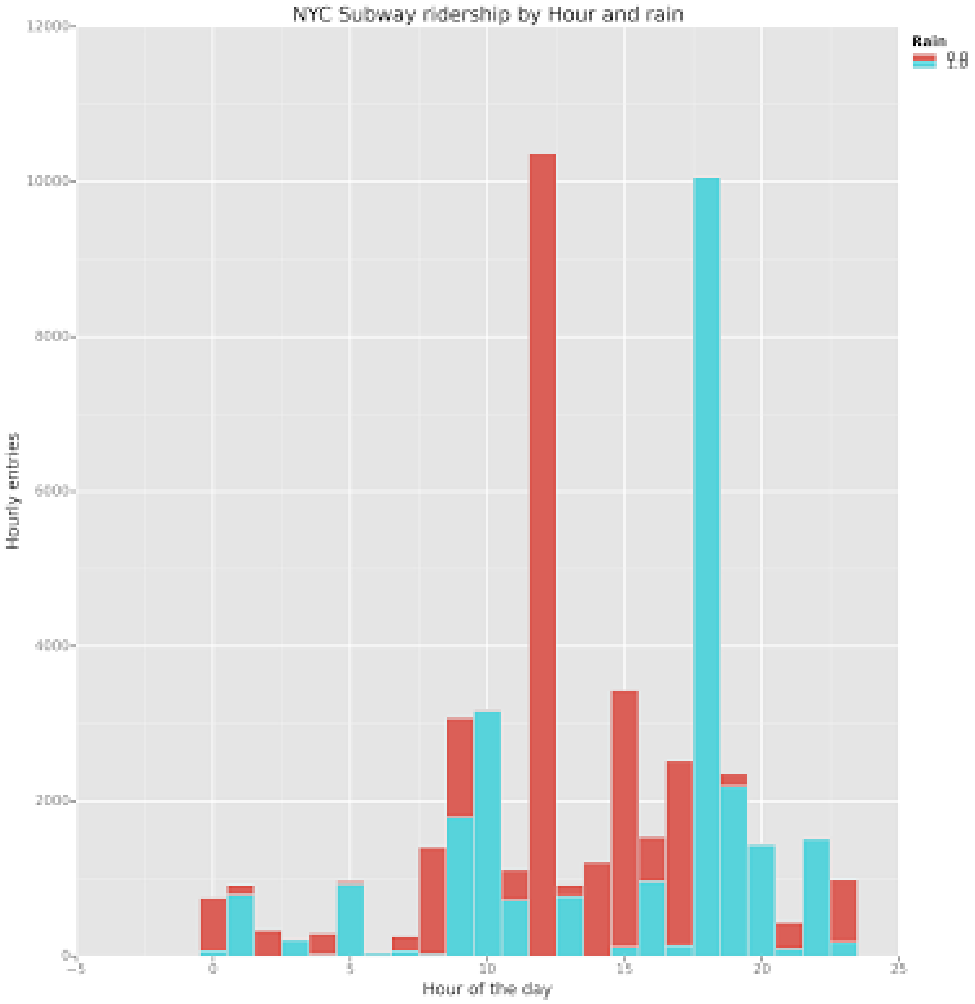
Section 3. Visualization[ii]

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.

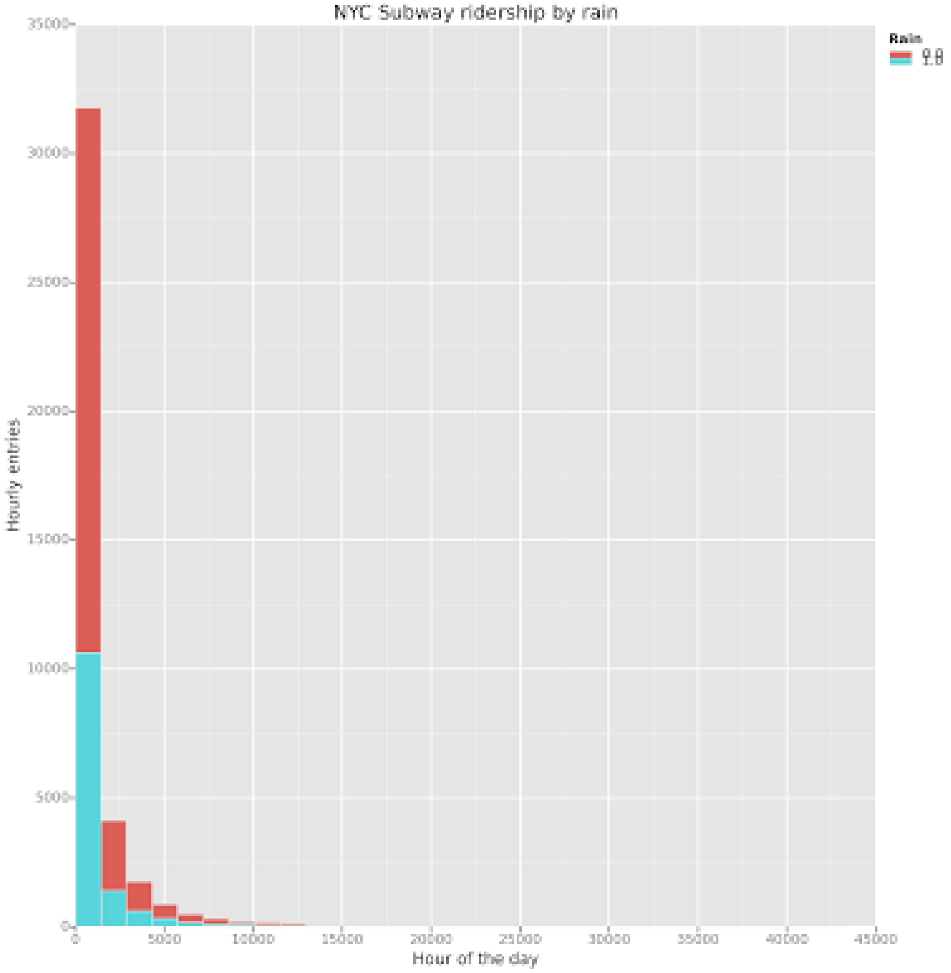
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

Ans: The following is the ggplot with 2 different histograms & the third with a scatterplot respectively. All these plots signify the hourly ridership during rain and non-rainy days.

From the following plot we can infer that there is not much of a similarity in the hourly ridership pattern depending on the rain condition.



From the following plot of the ridership plotted as the dependent variable on the y-axis we can observe



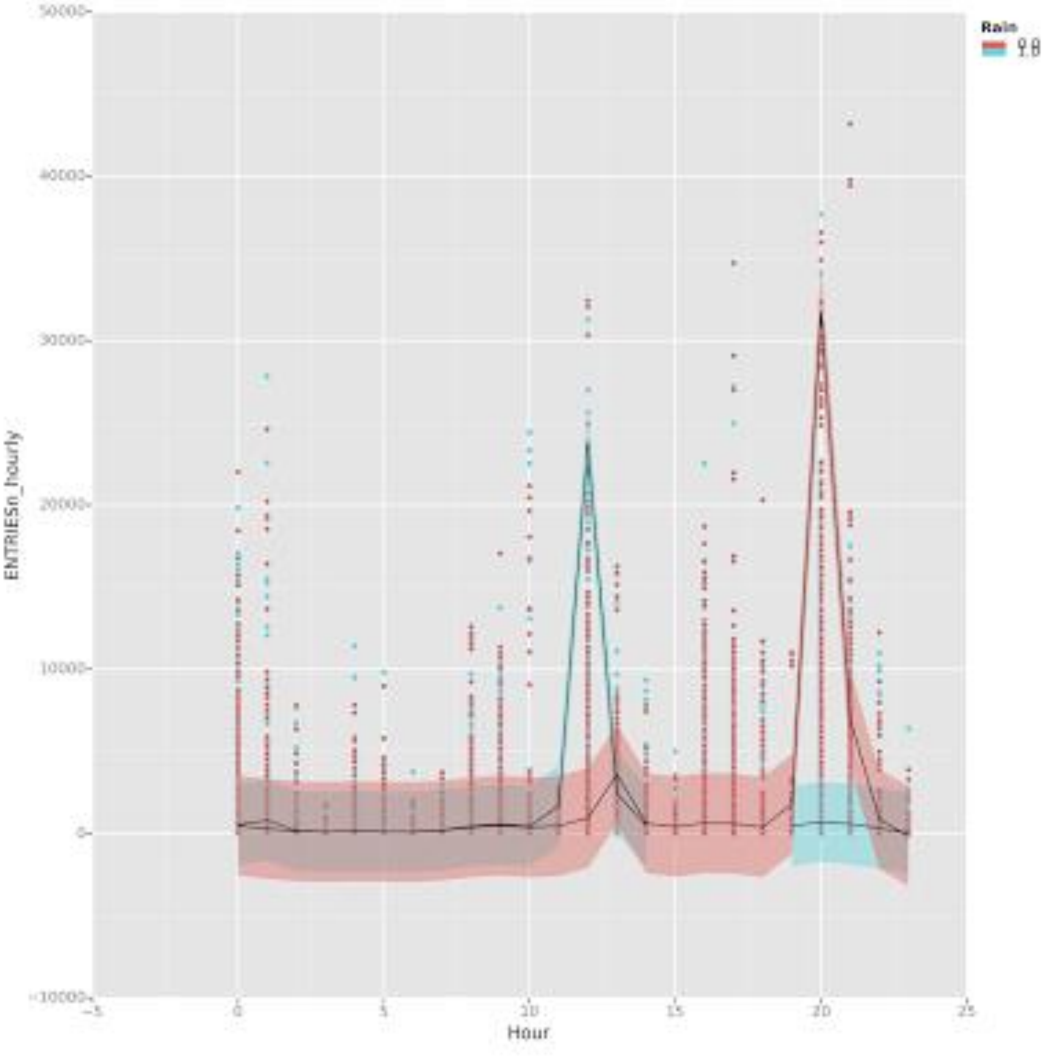
both the distributions are not normally distributed. Though this seems like the ridership during rain is less than that of the ridership when there is no rain, but it is not consistent as it could also be possible that the number of days considered for both the variables are different.

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

Ans: The following is a plot which provides information about the ridership by hour and the effect of rain on ridership along with the most times when it rains. For example we can notice that the maximum ridership is around 12:00 pm with a max ridership because of rain and at 08:00PM during non-rainy days respectively. This plot also gives a sense of the timings during which it rained most for the data set considered for instance, it rained from 12:00 am to 10:00 am generally but it didn’t from 11:00 am to 01:00pm and then from 2:00pm to 07:00pm.



iii

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?

Ans: Yes, based on the p-value (0.025) from the Mann-Whitney U test and from the linear regression, we can be conclusive that more people ride the subway when it is raining than 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.

Ans: The Mann-Whitney U test led us to conclude that there is a relation between the weather (rain) and the subway ridership. As per our calculations, we have obtained a two tailed p-value of 0.025. This means that we can be 95% confident that the null hypothesis (H_0: there is no difference between the ridership based on the rain condition) is false. This was confirmed by the positive co-efficient value which indicates like-wise too. However, with a significantly high correlation of 0.48 we can assume that this might not be applicable to all the data points. Thus based on the Mann-Whitney U test and the visualizations we can conclude that the ridership is affected by the rain condition.

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.

Ans:

The data set appears to be big enough but is within a small range of duration. This ties-up to one of the messages that was given earlier in the course that correlation doesn't imply causation. This is also significant because, we are analyzing a small chunk of data and applying its results to the entire population which can be misleading.

Another possibility that was not addressed was errors. We are not sure about the errors in the data as the presence of errors in the data can steer the conclusions another way altogether. Additionally, I believe that performing a linear regression was not suitable for this data set as there appears to be a non-linear relationship between the features. For instance, there could have been many other reasons that might have affected the ridership as in a public holiday or a special occasion. Allthough linear regression is not very effective against large data, the model could have been improved by adding additional features which might take into consideration the overall population rather than just focusing on a specific data set.

References:

[i] https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test#Assumptions_and_formal_statement_of_hypotheses

[ii] https://pypi.python.org/pypi/ggplot
[iii] https://github.com/yhat/ggplot