Analyzing the New York Subway Dataset

Author : Vijayasarathi Balasubramanian

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?

I used the Mann-Whitney U test to analyze the New York Sub way data and to see the passenger behavior when it rains versus non rainy days. The reason for using Mann-whitney U test is that we had both rainy and non-rainy days, so to start with we had two set of samples to test that's why we went to Man-whitney u test.

As stated above, as we have two different data sets like rainy and sunny data sets we need to use the two tail P value.

Null Hypothesis : NULL hypothesis is a commonly known statistical method used to disprove any hypothesis we have on the test data set. Say if we choose 10 students from a 50 student's class, we might want to test if all of the selected students are of high performers.

p-value : 0.025

1.2 Why is this statistical test appropriate or applicable to the dataset? In particular, consider the assumptions that the test is making about the distribution of ridership in the two samples.

The mann whitney u test is a non-parametric test which does not assume any particular distribution, as opposed to Welch's t-test. Therefore the mann whitney u test is the best fit for the NYC subway data set, simply by appreciating the histogram of the data we can clearly see the data is non-normal though we had two different data sets.

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.

The mann whitney u test returned a p-value of 0.025, so we reject the null hypothesis that both data sets are identical and have the same mean. In other words, both sample means are statistically different.

With rain mean : 1105.4463
Sunny season mean : 1090.2787
p-value: 0.025

Two Tailed Results :
With rain mean : 2210.8926
Sunny season mean : 2180.5574
p-value: 0.05

1.4 What is the significance and interpretation of these results?

These results show that subway usage increases when it rains, in a statistically significant way. On average, it increases by 15 riders per hour.

Analyzing the New York Subway Dataset

Author : Vijayasarathi Balasubramanian

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. Gradient descent (as implemented in exercise 3.5)
2. OLS using Statsmodels
3. Or something different?

Both gradient descent (GD) and OLS models where used to run linear regression on the NYC subway dataset. Both models look for linear relationships between the features and the predicted values or NYC subway rides.

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

In the gradient descent model the features used were: rain, precipitation (precipi), hour of the day (Hour), mean temperature (meantempi) and dummy variables for individual station (UNIT). In the OLS model, the features used where: rain, mean temperature (meantempi) and dummy variables for stations (UNIT) and dummy variables for hours of day (Hour).

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."*

After mixing and matching various features, these were the most relevant and important features based on their explicatory power and statistical significance. I had a bias for choosing the simplest model possible, without losing too much explicatory power or R^2.

I maintained rain, precipitation, hour, wind speed and mean temperature because out of experimentation, I was unable to find R^2 values that were better. I was adding one by one to see the R^2 value changes and how it impacts.

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

The coefficient of 'rain' is -1.20517622e+01
The coefficient of 'mintempi' is -7.17508874e+01
The coefficient of 'Hour' is 4.64126449e+02
The coefficient of 'fog' is 4.30221670e+01

2.5 What is your model's R_2 (coefficients of determination) value?

The R squared for the GD model is 0.461. The R squared for the OLS is 0.525.

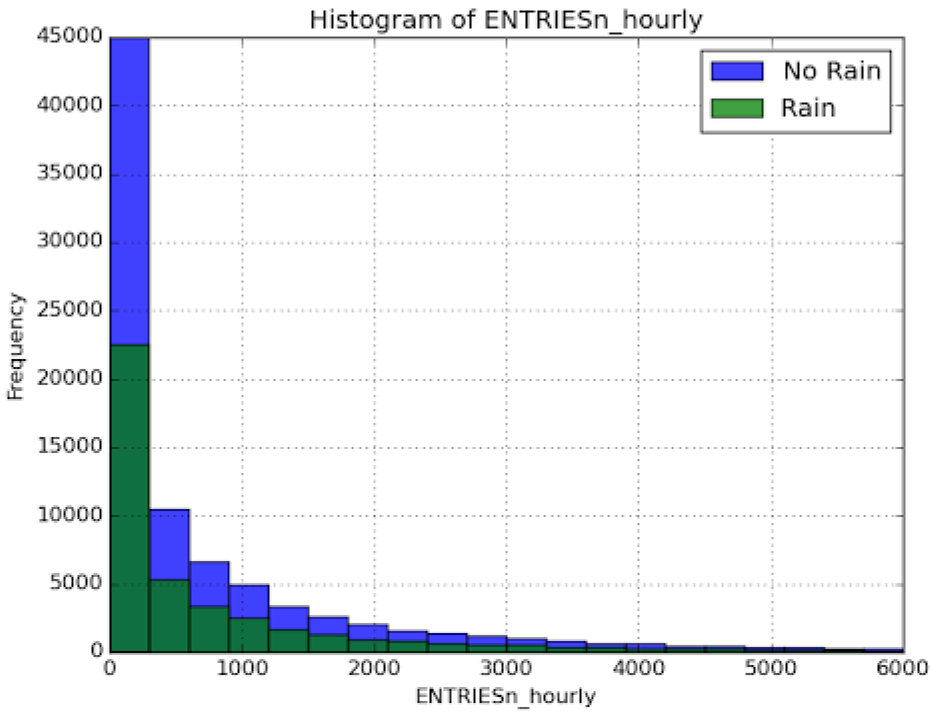
2.6 What does this R_2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R_2 value?

R^2 is essentially the percentage of variance that is explained, and is a quantitative measure of the "goodness of fit." While it only explains 52.5% of variation. To decisively conclude whether or not this model was a good fit certainly depends on the context and use case for the prediction data. The linear model is sufficient for our analysis.

Further and advanced study could include more features or utilize polynomial regressions. However, this might lead to significant over-fitting, and the model may fail on new data sets. In that case, regularization would be a good method to attenuate any over-fitting.

Section 3. Visualization
Please include two visualizations that show the relationships between two or more variables in the NYC subway data. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots, or histograms) or attempt to implement something more advanced if you'd like.
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



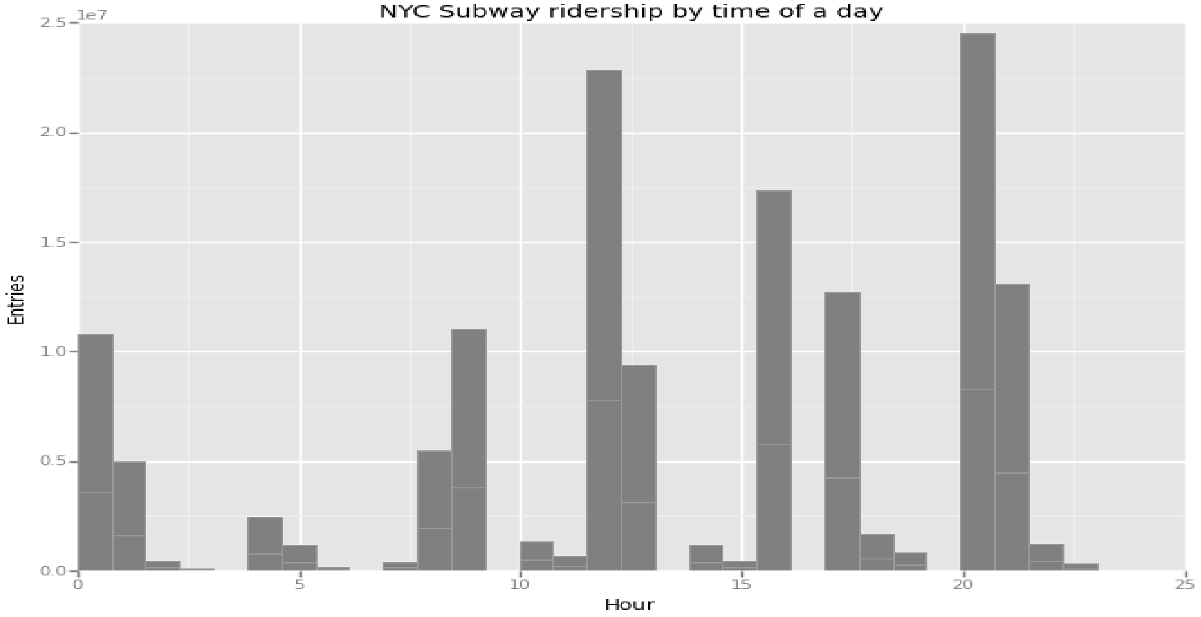
The above histogram shows the NYC subway data for hourly entries for the rain and non-rain season in the same picture.

3.2 One visualization can be more freeform. Some suggestions are:

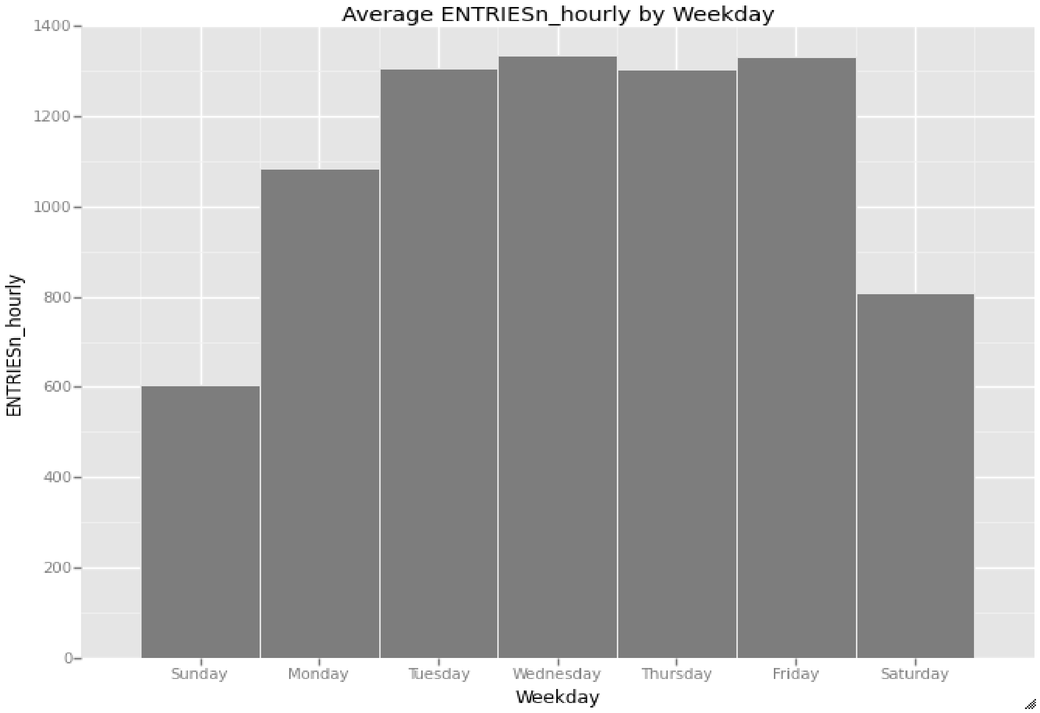
- Ridership by time-of-day
- Ridership by day-of-week

Please include two visualizations that show the relationships between two or more variables in the NYC subway data. You should feel free to implement something that we discussed in class (e.g., scatterplots, line plots, or histograms) or attempt to implement something more advanced if you'd like.
One visualization should be two histograms of ENTRIESn_hourly for rainy days and non-rainy days

> One visualization can be more freeform, some suggestions are: Ridership by time-of-day or day-of-week How ridership varies by subway station Which stations have more exits or entries at different times of day



The above histogram shows NYC rider's entries for the different hours in a day from the whole data set. The graph has been plotted for 24 hours in a day, to cover all the hours. If you interpret the picture closely, you can observe that the people are using the train more in the evening.



The above picture shows the average number of people travel hourly for different days in a week. The graph has been plotted for Sunday to Saturday and we can observe that the people use train more in week days.

Section 4. Conclusion

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?

On average, between 15 and 100 more people ride the NYC subway on a rainy day compared to a non-rainy day.

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 positive coefficient for the rain (0 or 1) parameter indicates that the presence of rain contributes to increased ridership. This may have not been the case for all data points, with the R^2 being approximately 46%; however, the small residuals show relatively high accuracy, given our objectives. Although the means of both data sets are not that different from each other, the Mann-Whitney U test did indicate that there was a statistically significant change in ridership for rain vs. no-rain. It is conscientious to claim that rain increases subway ridership.

Section 5. Reflection
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.

Dataset:
1) There might be omitted variables like festivity or event dates, closed dates for maintenance, etc.

Analysis:
2) In both the linear models a lot of dummy variables were used, which removed a lot of degrees of freedom and increases chances of multicollinearity. Example: we were unable to add three sets of dummy variables for hours of day, day of week and stations.

References:

https://www.python.org/
https://en.wikipedia.org/
http://pandas.pydata.org/
projects in git hub
GGPlot (http://ggplot.yhathq.com/docs/index.html)
GraphPad (http://graphpad.com/guides/prism/6/statistics/index.htm?how_the_mann-whitney_test_works.htm)
statsmodels.regression.linear_model.OLS
(http://statsmodels.sourceforge.net/devel/generated/statsmodels.regression.linear_model.OLS.html)
And all the reference materials in the course