Analyzing the NYC Subway Dataset

Short 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, 4, and 5 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 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 Mann Whitney U - test
Used a Two tailed P Value
Null hypothesis - The two samples under test are same
P Critical value is 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.

The data is not normally distributed hence used Mann Whitney U test. The two samples (rainy and non-rainy) are independent of each other. The responses are ordinal (the ridership on the rainy days are more than non-rainy)

Refer to the histogram in section 3.1

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.

rainy day mean 1105.4463767458733,
Non rainy day mean 1090.278780151855,
U - 1924409167.0,
p value - 0.049999826 (two tailed value, doubled the single tiled value that I got)

1.4 What is the significance and interpretation of these results?
The P value is less than 0.05, hence the null hypothesis is rejected.
The results show the subway ridership (entries) increases when it rains

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?

Used the gradient descent

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

Used following vairables 'rain', 'precipi', 'Hour '
Yes, used a dummy variable based on units

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."

Through the course of exercise (3.1 and 3.3) sensed that 'rain' play critical features/input variables. And by looking in to data understood 'hour' have a relationship with ENTRIESn_hourly. I was experimenting with other input variables and found the 'precipi' improves the R square value. Hence used these features in my model

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

1.62096990e+01 2.52235268e+01 4.30344784e+02

2.5 What is your model's R^2 (coefficients of determination) value?
My R^2 (coefficients of determination) value is 0.482478577098

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?

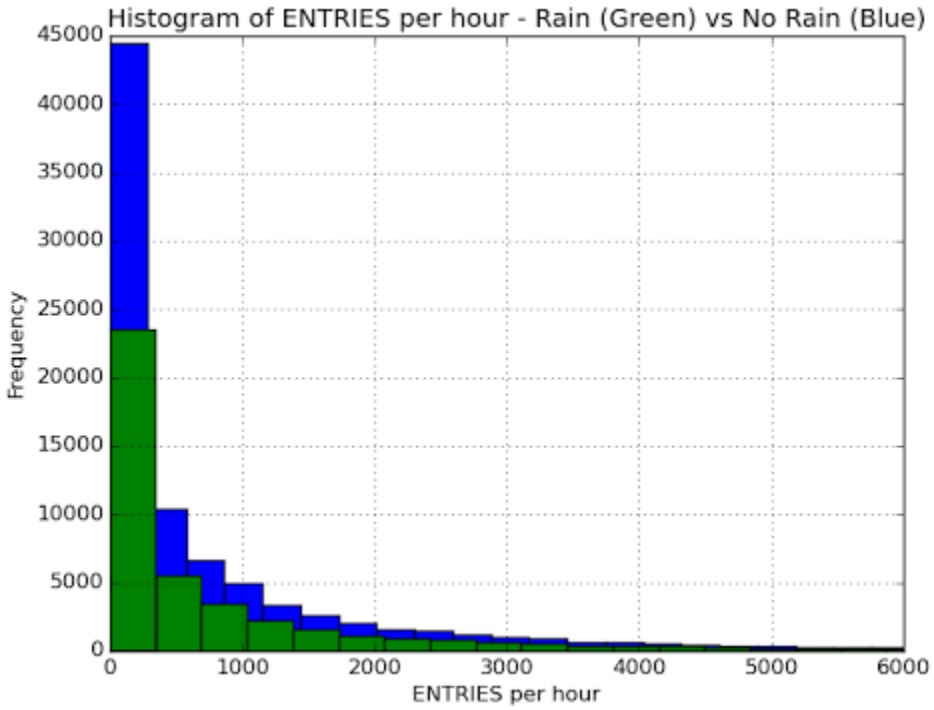
The R-square value is an indicator of how well the model fits the data. Even though my R square is not close to 1.0, but for this dataset the R square value is good enough in terms of fit and to predict ridership

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.



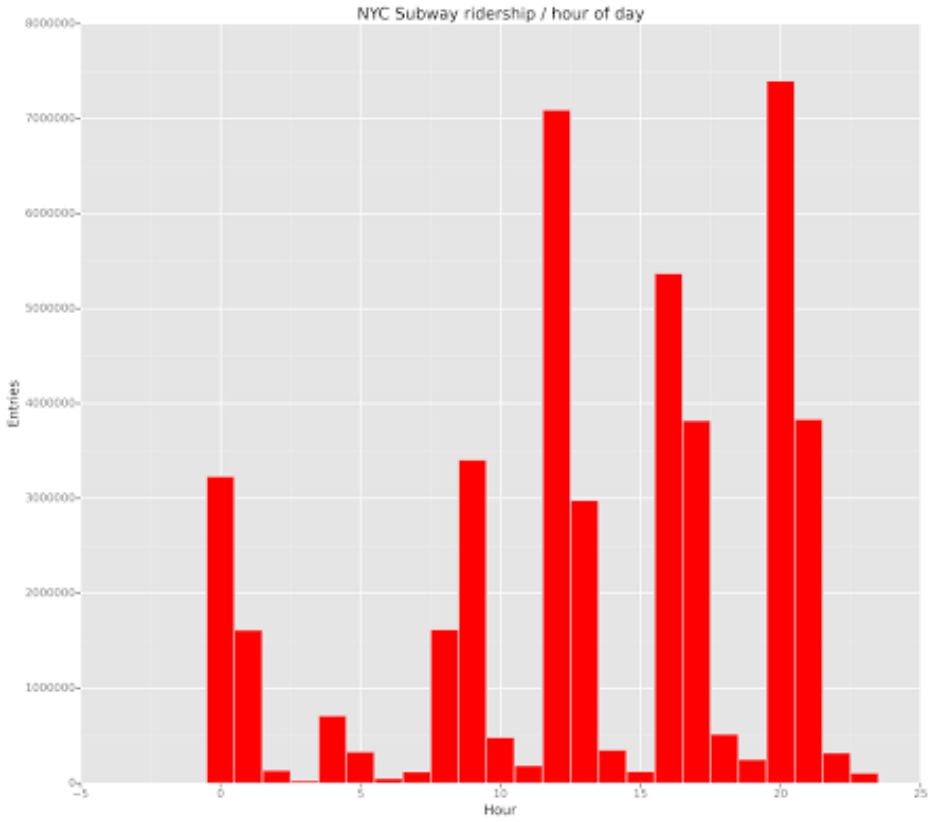
The graph shows the subway entries per hour on a rainy day (green bar) vs non rainy day (blue bar). Histogram shows the data is not normally distributed. Also in the given data set there are more data points for a non-rainy day (which is logical). And the size of bins shows that there are more entries per hour on a rainy day compared to a non-rainy day

Note : It is not the entire data set , maximum range for entries per hour is 6000 and frequency is 45000

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

Ridership by hour of day : Did a rollup/sum of the entries according to the hour of the day and plotted it. It gives an idea on when the peak hours are for the subway



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.

From my analysis more people ride NYC subway when it is raining. I drew this conclusion based on

- Hourly entries mean - It showed that mean is higher on a rainy day (1105) compared to a on rainy day (1090)

- Null Hypothesis & Mann Whitney U test - I started with the null hypothesis that both the samples - Rainy day and non-rainy day are same. And by running the Mann Whitney U test I was able to reject it. The p value less than 0.05 indicated the alternate hypothesis, the sub way ridership increases when it rains.

  Also when I did the linear regression exercise I was able to see the prediction of the ridership had a good variability with the rain as a feature (when I changed the input variable rain I could see R square value increasing). All these led to my conclusion.

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.

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

Few of the shortcomings of the dataset - I find the more number of records for non-rainy days compared to rainy days. And also the mean difference between two samples was less which made it difficult to conclude analysis from logical sense. In such cases the statistical tools came in handy.

Also when doing the linear regression model, i dint see the R square value improving much when trying with multiple features/variables. May be the dataset is designed in such a way not provide a clear indicator of what the key variables are to determine the increase in ridership. Having data for more months would help us to understand patterns - trends, seasonal weather coloration.

There are other analyses that we could do more than rainy day influences on ridership. Things like the best unit (station) to get in which will have least crowd (or hourly entry). And best time to commute to avoid rush.

References

1. Conceptual references Wikipedia and answers.com
2. Syntax reference and examples http://stackoverflow.com/
3. R2 value goodness of fit : http://www.statsoft.com/Textbook/Multiple-Regression#cresidual