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, 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 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.

- https://www.edx.org/course/introduction-statistics-descriptive-uc-berkeleyx-stat2-1x

I used Stackoverflow on several occasions for learning more about Pandas, Numpy but I didn’t keep track of the pages I visited.

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?
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.
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?

I used the Normal distribution statistical test. The Null Hypothesis Statement that we wanted to prove in the “Intro to Data Science P3” was that the ridership is the same on rainy or non-rainy days. I used the two-tail P value. The p-critical value was 0.024.
I rejected the Null hypothesis as the p-critical value was less than 0.05.

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:
Gradient descent (as implemented in exercise 3.5)
OLS using Statsmodels

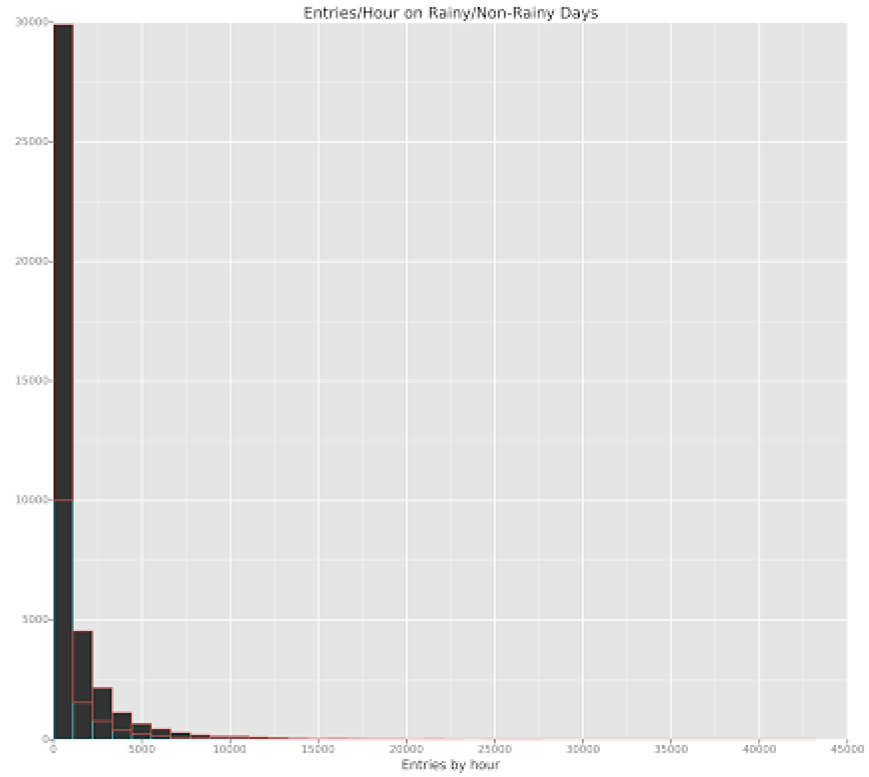
Or something different?
2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?
2.3 Why did you select these features in your model? We are looking for specific reasons that lead you to believe that
the selected features will contribute 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 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?

I used Gradient descent.
I used 'rain', 'precipi', 'Hour', 'meantempi' as features and ‘UNIT’ as a dummy variable.
The experiment was setup to see if ridership increased if it was more rainy outside, as rain causes driving to be more painful.
The coefficients are called Thetas. They are passed as arguments to the Gradient Descent function.
The R-squared value is 0.318137233709. Since the R-squared value is closer to 0, only 31% of the total variation is accounted for by the regression line. This linear model is not a good fit to predict ridership, for this data set.

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,

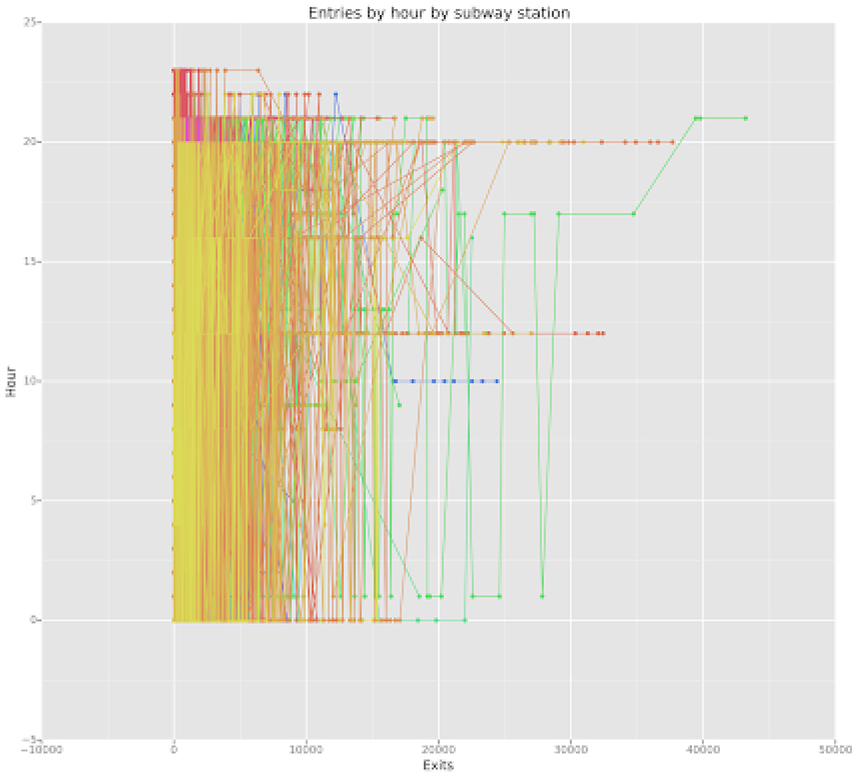
each interval (along the x-axis), the height of the bar for this interval will represent the number of records (rows in our data) that have ENTRIESn_hourly that falls in this interval.
Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.



The red vs blue outline around the bars is for rainy/non-rainy days.

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

The following plot is for Hourly Entries vs Hour and colored based on Subway station entered.



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?

They ride it 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 histogram on hourly entries clearly shows that there were more riders on non-rainy days. Plus when we did our statistical test, our Null Hypothesis (more rainy days meant more ridership) was proven to be inconclusive by the R-squared value.

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.

Definitely there is a fear that the ⅓ dataset that we used has less rainy days by chance! I don't the use of Linear regression using Gradient Analysis is a bad choice yet.

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

It would be great if we could run MapReduce jobs on the full Turnstile data set with the same Null Hypothesis (more rainy days means more ridership) on a cluster of machines! Truly show that Statistics requires large data sets (so that hasty conclusions are not drawn from a smaller sample set). However, the MapReduce programming model + cheap computing power is required for this…