Short Questions to Analyzing the NYC Subway Dataset

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 analyse 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?

Mann-Whitney U-test is used to measure the means for non-normal distribution. Two-tailed P value is used to test the null-hypothesis - rain and non-rainy days are not statistically significantly different. P critical value I choose P=0.05 where 'not raining' size of 87847 and 'raining' 44104.

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 MWUt is a non-parametric test which does not assume any particular distribution, as opposed to Welch's t-test. Therefore the MWUt is the best fit for the NYC subway data set. With plotted histogram of the data we see the data is clearly not normal.

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.

We reject null hypothesis since MWUt with a p-value of test statistic: 0.0386 < 0.05(critical value).

- mean entries per hour not raining: 1090.27878015
- mean entries per hour raining: 1105.44637675

1.4 What is the significance and interpretation of these results?

P is small enough to reject null-hypothesis, which in turn states rainy days' ridership is different from that of non-rainy days'. This result shows that subway usage increases when it rains, in a statistical significantly way. On average there are 15 more riders per hour when it rains.

- median entries per hour not raining: 278.0
- median entries per hour raining: 282.0
- mean entries per hour not raining: 1090.27878015
- mean entries per hour raining: 1105.44637675

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?

Gradient Descent and OLS were used to run linear regression on the NYC subway data. Both models look for linear relationships between the features and the predicted values (ENTRIESn_hourly) of 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?

GD used rain, precipi, Hour, meantempi and dummy UNIT. OLS used rain, mintempi and dummies are Hour, UNIT.

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.

If I need to go to somewhere in a rainy day, massive transit helps to avoid finding parking, potentially prolonged trip time, and eliminate danger driving in inclement weather. Temperature is considered for comfort and safety.

'rain', 'precipi', 'Hour', 'meantempi' are the features chosen in GD, and UNIT as dummy variable. This selection is due to experimenting with different combinations. Moving Hour to dummy variable was a last minute move in order to improve R^2. I can't seems to find any other combination have significant improvements over what I found.
feat=weather_turnstile[['rain','mintempi']].join(du).join(hr)
Your R^2 value is: 0.525548088253

- 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."

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

In OLS, coef for rain=73.9803; mintempi=-12.3611. In terms of coefficient as slope, we can interpret for any one more rainy days, there's an increase of 73.98 ENTRIESn_hourly. This is consistent with our earlier postulation. On the other hand, for each degree of increase, we see a decrease of 12.36 ENTRIESn_hourly. Although it is interesting to consider some of the unit_rxxx also increase significantly, but we simply couldn't adjust unit_rxxx as they are stationary. It could be helpful to know their location and have an understanding of why some of the station have more impact than the others.

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 0.526.

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 squared for the OLS is 0.526.

Given $R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$ with $SS_{\text{res}} = \sum_i (y_i - f_i)^2$ and $SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2$,

R^2 is the calculation of error from known measurements. There are many possibilities that OLS would not produce a workable model with a legitimate R^2 value. By reviewing 1 and 2, there are a lot more analysis can be used to explore more details of data distribution. Although in this exercise, I am not going to pursue those possibilities at present time.

Based on its own, our R^2 is good enough for estimating the trend of raining effect on Subway traffic except we have not done things such as residual analysis. If residual is as random as it should, this regression model should serve the purpose.

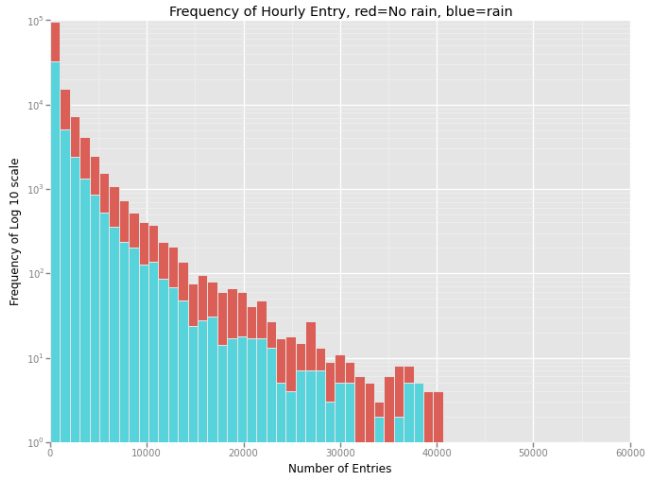
Section 3. Visualization

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.

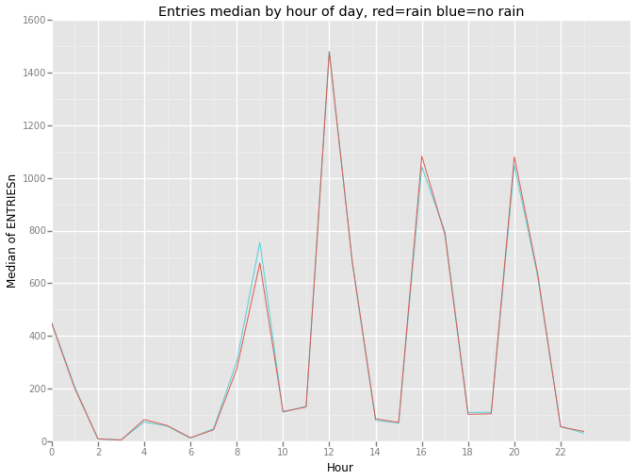
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

- 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.

It seems ggplot has changed to using 'tight' as layout and I can't get legend to plot within the diagram. All diagrams have no legend attached.



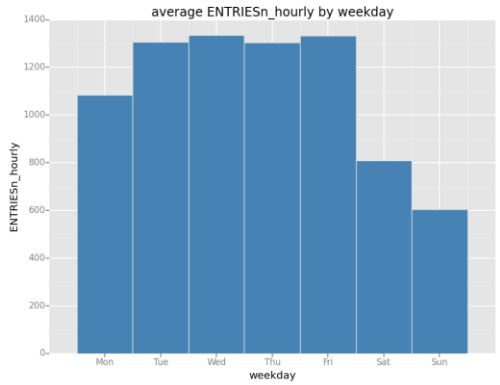
This shows count of ENTRIESn_hour in bin size of 1000. X is the magnitude of recorded ENTRIESn_hour. Y is the frequency within the bin's range.



This draws median of ENTRIESn_hour with group by hour and rain. Color is designated to variable 'rain'.

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

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



Average(mean) plot of ENTRIESn_hour for each weekday.

Weekday used by python indicates Sunday = 6.

The following diagram is essentially the same as number 2, just in stacked bar format. It is removed.

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?

On average, 15 more people per hour ride the NYC subway on rainy days than non-rainy days. Linear regressions with OLS indicate there's a difference between the rain/norain groups.

- median entries per hour not raining: 278.0
- median entries per hour raining: 282.0
- mean entries per hour not raining: 1090.27878015
- mean entries per hour raining: 1105.44637675

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.

Combining the fact there is a difference between rainy days and non-rainy days traffic pattern (per Mann-Whitney test rejected original null hypothesis) and from OLS modeling (with R^2 close to 1 enough), we can conclude (disclaiming some exceptions may occur per section 2.6) there are more people ride the NYC subway then it rains.

Further examination of mean/medium values for each group give us more confidence such assertion is consistent to our view of data.

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. Linear regression model,
3. Statistical test.

1. There could have other reasons affect the outcome such as events or holidays. If riders are more weather sensitive, why we don't see more significant number reflected in numbers. Would be interesting to add other factors into the analysis.
2. The data fitted may not be a linear pattern as the model we used. A more complex model may fit better.
3. Majority of records are made at the top of an hour. Only 20197 of 131951 records (15.3%) are not. Does this imply the decision made by MTA riders can't be examined more closely?
4. Comparing GD vs OLS, it seems the larger number of variables will eventually

lead more complexity for OLS, as the dimensional cubical cost of computation of the matrix. On the other hand, GD seems to be a straightforward solution and should be scaling well.

5. I also noticed if I throw more features at the model, the R squared marginally rises. In effort keeping things simple, I decide such additions are not contributing in the larger pictures. Others may disagree.

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

References:

- http://rpubs.com/nikedenise/3256
- https://github.com/yhat/ggplot
- http://ggplot.yhathq.com/docs/index.html
- http://graphpad.com/guides/prism/6/statistics/index.htm?how_the_mann-whitney_test_works.htm
- http://statsmodels.sourceforge.net/0.5.0/generated/statsmodels.regression.linear_model.OLS.html
- http://en.wikipedia.org/wiki/Coefficient_of_determination