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://statistics.laerd.com/premium-sample/mwut/mann-whitney-test-in-spss-2.php
http://stackoverflow.com/questions/26886653/pandas-create-new-column-based-on-values-from-other-columns
http://ggplot.yhathq.com/docs/index.html
http://www.statmethods.net/stats/nonparametric.html
http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit
http://docs.scipy.org/doc/numpy/
http://stackoverflow.com/questions/22391433/count-the-frequency-that-a-value-occurs-in-a-dataframe-column
http://stackoverflow.com/questions/9847213/which-day-of-week-given-a-date-python
http://stackoverflow.com/questions/3606697/how-to-set-limits-for-axes-in-ggplot2-r-plots

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?

Test Used: Mann-Whitney U test to analyze the NYC subway data.

P value used: One-tail P value.

Null hypothesis H0: The distribution of number of entries (ie. Ridership on the NYC subway on rainy days and no rainy days) are statistically the same

p-critical value: 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.

Assumption 1: Ridership ("Entriesn_hourly") is measured at the continuous level (ie, it is quantitative)

Assumption 2: Independent variable consist of independent group, rainy day and non-rainy day

Assumption 3: There is no relationship between the observations in each group (the rainy day ridership and non-rainy day ridership)

Assumption 4: The data drawn for the rainy and non-rainy samples are not normally distributed or follow any particular underlying probability distribution

Mann-Whitney U test makes the above assumptions to give a valid result and because our data satisfies these assumptions it is applicable to the dataset that we have.

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.

On implementing the Mann-Whitney U test in Python following are the results that were observed

Mean ridership on rainy day is: 2028.196

Mean ridership on non-rainy day is: 1845.539

p-value is: 5.482e-06 (got a nan in Python , used R to run the test to get the p-value)

W = 153635120.5

1.4 What is the significance and interpretation of these results?

Since observed p-value from the test is significantly less than p-critical value (0.05), we reject the null hypothesis and conclude that the ridership on NYC subway is statistically different between rainy and non-rainy days.

By comparing the means of the two samples, on an average there is an increase of about 7 riders per hour on a rainy day when compared to a non-rainy day.

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 (as implemented in exercise 3.5)

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

Features used were “rain”, “hour”, “weekday”.

Used “conds”, “station” and “units” as dummy variables in addition to the above 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 R^2 value."

Used intuition and initially selected features, Rain, Hour, Weekday, meantempi, day_of_week, wspdi and in addition used "unit" as dummy variable.

After a few runs of the model and looking at the variations in R^2 values narrowed down to the 3 features that were finally chosen, 'rain', 'hour' and 'weekday', along with the 3 dummy variables mentioned above.

There wasn't a significant increase in the R^2 value by including other variables (meantempi & wspdi) and hence were removed from the model to make the model simple with only the required features.

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

The coefficients (or weights) for the 3 non-dummy features are

Rain: 90.53
Hour: 1122.76
Weekday: 554.51

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

R^2 of the model is 0.477

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 a measure of how close the data are to the fitted regression and is the ratio of explained variation to total variation.

So a R^2 value of .477 means that 47.7% of data variability could be explained by the model above, ie we can predict entries with 47.7% accuracy

I think this model is appropriate given this value of R^2, especially to predict human behavior of using the NYC subway based on weather conditions.

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.

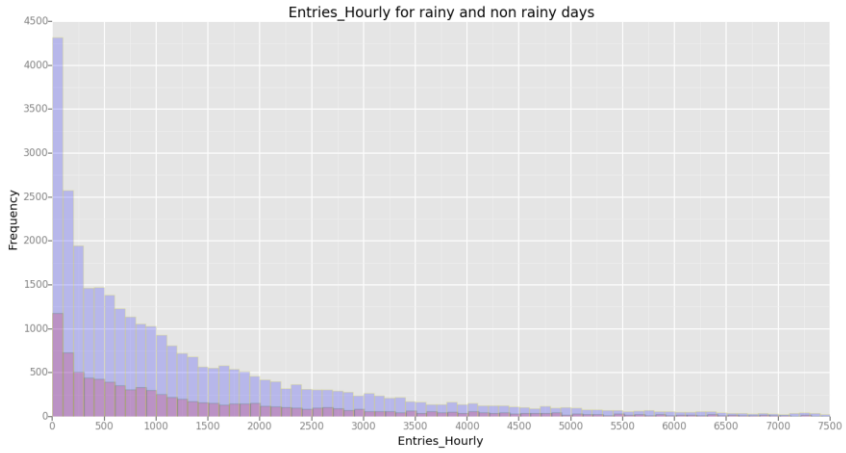
- For the histograms, you should have intervals representing the volume of ridership (value of ENTRIESn_hourly) on the x-axis and the frequency of occurrence on the y-axis. For example, each interval (along the x-axis), the height of the bar for this interval will represent the number of records (rows 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.

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

Histogram that plots ENTRIESn_hourly for rainy and non-rainy days



Legend

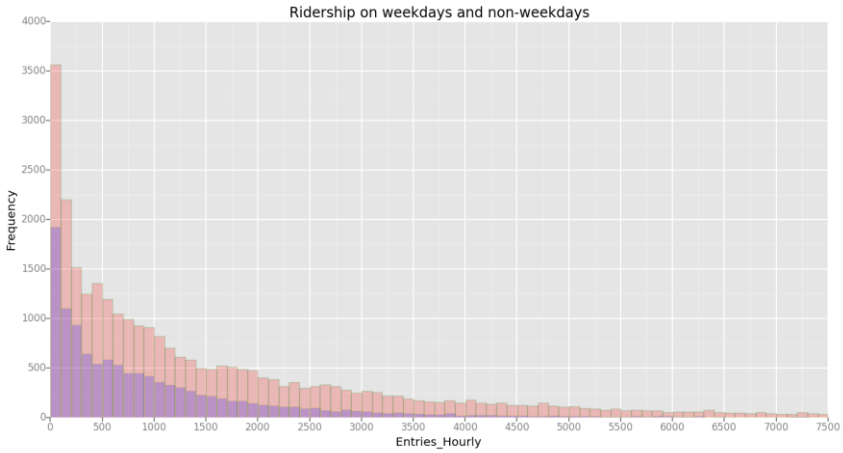
- Red bars plots number of entries for rainy days
- Blue bars plots number of entries non-rainy days

The above histogram plots the count of Entries_hourly for both rainy and non-rainy days in the dataset.

Blue bars represent non-rainy days and red bars the rainy days. Since the number of non-rainy day rows are significantly greater than the rain day rows the blue bars are more pronounced.

As we can see from the chart the distribution of the two samples are very similar and the variances are quite close to each other as well.

Histogram that plots Ridership for weekdays and non-weekdays



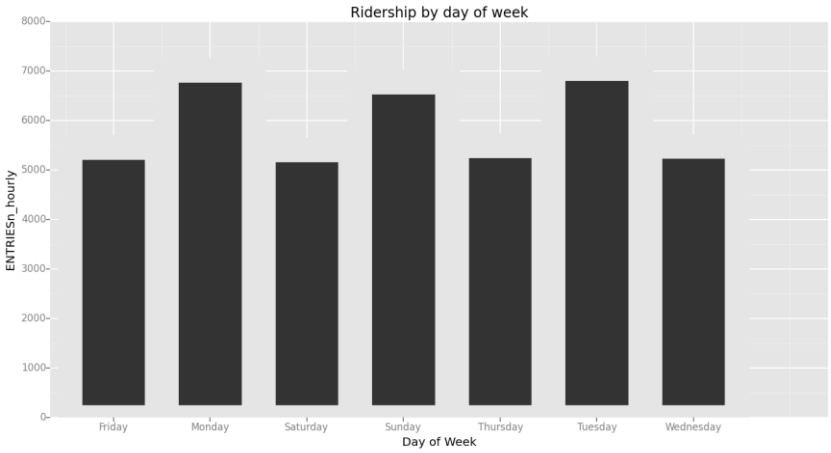
Legend

Red bars plots number of entries for weekdays

Blue bars plots number of entries for non-weekdays

The above histogram plots ENTRIESn_hourly for weekdays and non-weekdays. As can been seen, the number of people using the NYC subway is significantly higher during weekday when compared to a non-weekday

Histogram that plots Ridership by day of week



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.

Conclusion: More people ride the NYC subway when it is raining

From the Mann-Whitney U test we see that the distributions of the two samples (rainy day vs non rainy day) are statistically different from each other and by comparing the means of the two samples we can see that there is an increase of about 7 riders per hour on a rainy day when compared to a non-rainy day.

Further looking at the results from the Gradient Descent model, the coefficient (theta) for the rain variable is 90.53, and since when it rains the value of the rain variable in the dataset is represented by a value 1, the model predicts on an average about 90.53 more people will ride the subway on a rainy day.

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.

The dataset may be missing key data points that may have made some significant contributions to our results, eg. Special occasions, festivals etc…Dataset includes a number of data points that do not contribute much to the results but contributes to the increase in the size of the overall data and makes it difficult to process it, given the limitations of the computer hardware.

Mann-Whitney U is a non-parametric test and is less efficient than other parametric tests. These tests are usually used to get a quick analysis of the data before applying other advanced tests.

The GD model requires us to choose the appropriate alpha term (the learning rate), a large value may overshoot our minimum cost value and we may fail to converge and a small value may require a number of iterations before we converge to the minimum cost value. Choosing the right alpha term may pose a challenge in practical situations, especially when we are dealing with a number of parameters.

Although we have used R2 values to predict the goodness of the test, it may not tell the whole story in some instances. For example there is no guideline to indicate if R2 is adequate for regression model. A good model may have a low R2 value and vice versa. While a high R2 is required for good predictions, it is not sufficient.

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