Project: Analyzing the NYC Subway Dataset

Overview

1. Part 1 of the project has been completed. All the questions in Problem Sets 2, 3, 4 and 5 in the Introduction to Data Science course are answered, except the optional ones.
2. I have enclosed folder Code_Used_For_Few_Problems, this has the some of the code I ran in my local system for solutioning and answering short questions of the project.
3. References I have included in the separate reference section below.

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?

Answer:

- **Statistical test**
  We used Mann-Whitney U-Test to analyse the NYC subway data. Analyzing subway ridership on Rainy Vs. Non rainy days.
- We used two-tail P value.
- **Null hypothesis**
  There is no difference in NYC subway ridership between rainy days and non-rainy days.
- p-critical 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.

Answer: Following are the reasons why Mann-Whitney U-Test is applicable for the dataset.

- As observed using histograms of entries for rainy days and non rainy days, both the datasets are identical distributions.
- Both the dataset does not have same number of observations.

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.
Answer: Analyzing subway ridership on Rainy Vs. Non rainy days, based on number of hours, two samples of rainy and non rainy ridership significantly differ (Mean Rainy days=1105.4463767458733, Mean non-rainy days=1090.278780151855 Mann-Whitney U=1924409167.0, p=0.019309634413792565, sig≤.05, 2-tailed).

1.4 What is the significance and interpretation of these results?
Answer: Observing the Mann-Whitney U test results.
Mean Rainy days=1105.4463767458733, Mean non-rainy days=1090.278780151855 Mann-Whitney U=1924409167.0, p=0.038619268827585131, sig≤.05, 2-tailed.

p=0.038619268827585131
By interpretation of p-value, it is observed that p=0.038619268827585131, which does not exceed the null hypothesis declaration that p<=0.05.

There is certainly sufficient information to reject the null hypothesis and to declare that there is a significant difference in NYC subway ridership during rainy days and non rainy days.

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?

Answer: 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?

Answer: Following are the features used in the model.

- 'rain', 'fog', 'precipi', 'Hour'

Yes, used 'unit' as a dummy variable as part of 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."

Answer: For variables 'rain' & 'fog', I thought that during rainy / foggy days people prefer to use subway, as I do myself.

For variables 'precipi' & 'Hour', I saw increased R^2 value, when I included these variables in the model.

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

Answer: Coefficients (or weights) for non-dummy features 'fog', 'rain', 'precipi' & 'Hour' are 2.23056489e+01 -1.15438893e+01 1.75979191e+01 and 4.64517260e+02 respectively.

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

Answer: R^2 value is: 0.457707640824

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?

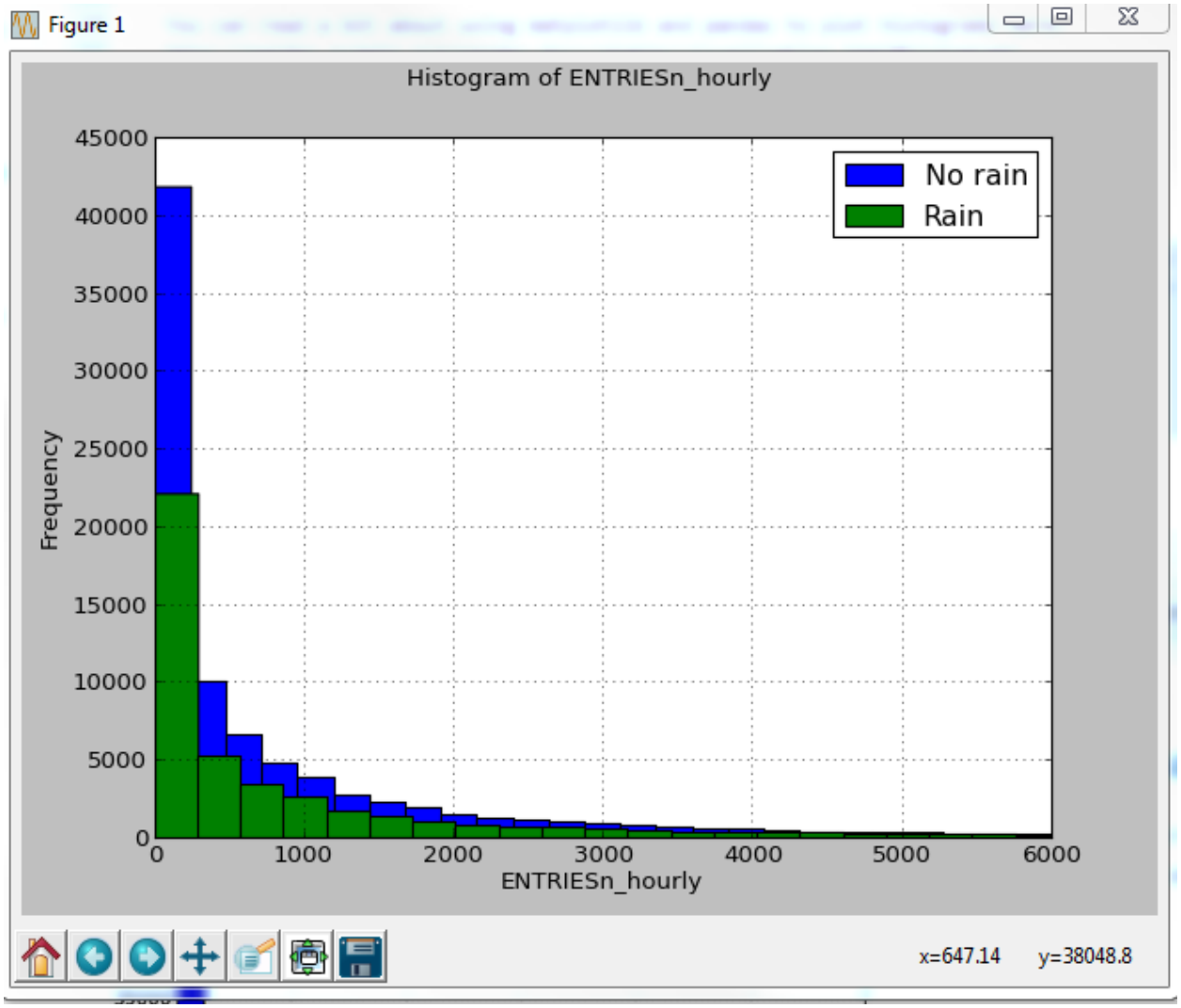
Answer: R^2 (coefficients of determination) value, measure of how much variability in the data, our model is able to capture. Any correctly fitted linear regression model will have $0 \leq R^2 \leq 1$, R^2 bigger the better. As the R^2 value is 0.457707640824, our linear model to predict ridership is appropriate for the dataset.

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.

Answer: Exercise set 3. 1 has the related code for below histogram. This histogram depicts that, both rainy and non-rainy ridership is not normally distributed. We can see that subway ridership is more during no rain days than rainy days.

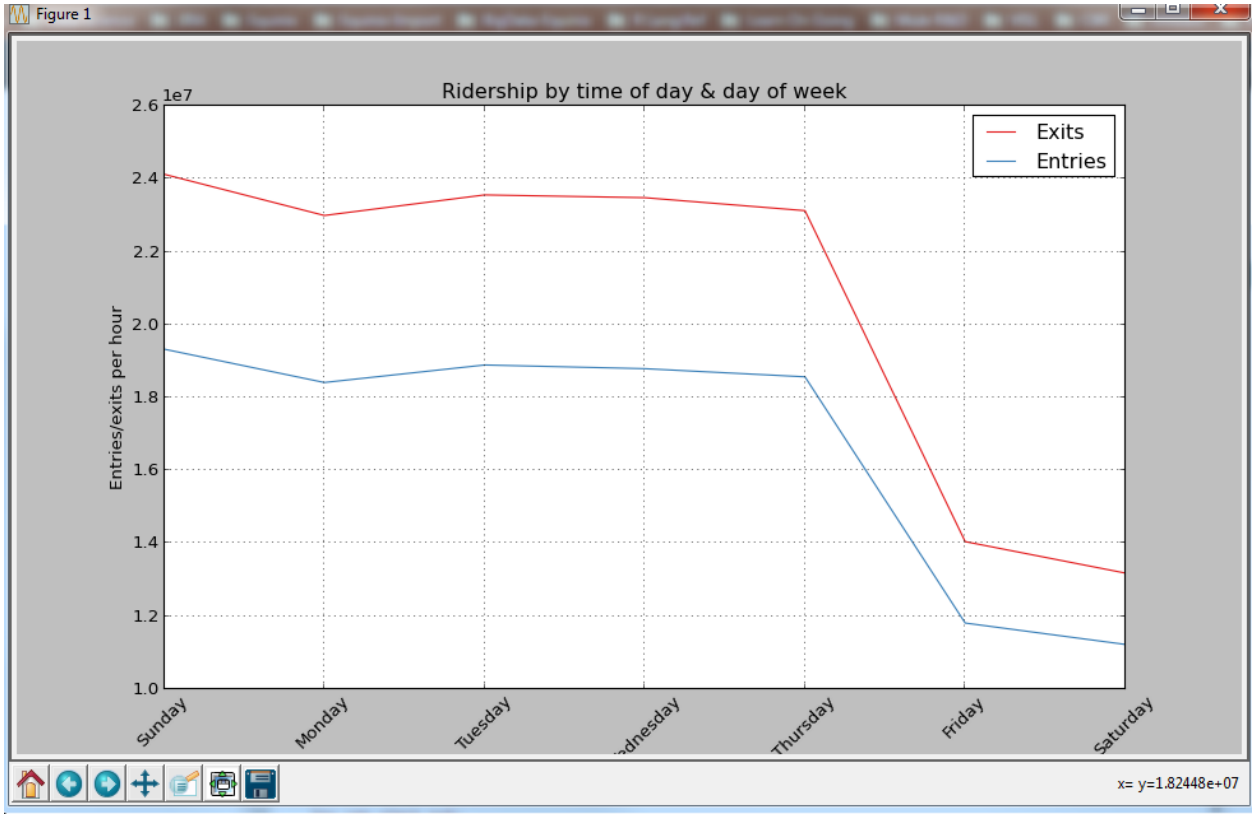


3.2 One visualization can be more freeform. Some suggestions are:

Ridership by time-of-day

Ridership by day-of-week

Answer: Exercise set 4.1 has related code for below line chart. This chart depicts ridership data of entries / exits per hour on a daily basis. Friday and saturdays are least busy days.



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?

Answer:

Histogram:

By plotting the histogram (refer answer to 3.1) of NYC subway ridership data. x-axis represents entries_hourly and y-axis represents the frequency, of rainy and non rainy days, we can observe that non-rainy days there are more entries_hourly. With that observation, numeric estimates indicate that more people travel when it is no rain than number of people travel when it is raining.

Mann-Whitney U test:

Performing Mann-Whitney U test, which does not assume the data is normally distributed, it will lets the answer the question, by comparing 2 samples, 1 sample data of entries_hourly during rain, 1 sample data of entries_hourly during no rain. Result of Mann-Whitney U test

median entries per hour when not raining: 278.0

median entries per hour when raining: 282.0

p-value of test statistic: 0.038619268827585131

Using this it allows us to say that, with a 3.8% degree of certainty, that there is difference in the average number of entries per hour depending on whether its raining or not, and with histogram results, we can accept the hypothesis that ridership per hourly during no rain is greater than ridership per hourly during rain.

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.

Answer: Analyses performed using statistical test, where we accept the hypothesis that there is difference in ridership data during rain or no rain. Linear regression used to predict the entries_hourly for ridership, we can notice that one of the factors which influence the prediction is rain. In both the approaches and the supporting histogram data we can conclude that rain influences the subway ridership and there are more people riding subway during non rainy days than rainy days.

Linear regression:

We performed linear regression to predict entries_hourly, to see how many people ride subway using gradient descent method. Observing th results of the linear regression where we found that R2 value is 0.457707640824, so loosely speaking our model identifies 45% of the variation present in the data it was trained on. Coefficients (or weights) for non-dummy features 'fog', 'rain', 'precipi' & 'Hour' are 2.23056489e+01 -1.15438893e+01 1.75979191e+01 and 4.64517260e+02 respectively. Which reiterates the point that rain influences by magnitude -1.15438893e+01 for subway ridership. More people ride subway during non-rainy days than rainy days.

Mann-Whitney U test:

Performing Mann-Whitney U test, which does not assume the data is normally distributed, it will lets the answer the question, by comparing 2 samples, 1 sample data of entries_hourly during rain, 1 sample data of entries_hourly during no rain. Result of Mann-Whitney U test

median entries per hour when not raining: 278.0

median entries per hour when raining: 282.0

p-value of test statistic: 0.038619268827585131

Using this it allows us to say that, with a 3.8% degree of certainty, that there is difference in the average number of entries per hour depending on whether its raining or not, and with histogram results, we can accept the hypothesis that ridership per hourly during no rain is greater than ridership per hourly during rain.

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.

Answer: Summary of shortcomings of the methods used to analyse subway ridership data.

1. Analysis, such as the linear regression model or statistical test

In our Linear regression we used only four features (‘fog’, ‘rain’, ‘precipi’ & ‘Hour’) to predict the value of entries_hourly, we can improve our model and its fitting to data, by including more features like rain', 'precipi', 'Hour', 'meantempi', 'mintempi', 'maxtempi', 'mindewpti', 'meandewpti', 'maxdewpti', 'minpressurei', 'meanpressurei', 'maxpressurei', 'meanwindspdi'. Alternatively we can use polynomial combinations of features to give our linear regression model a better chance of fitting the data.
Linear regression with four features, R2 : 0.457707640824
Linear regression with additional features, R2 : 0.460322229345
Linear regression with polynomial combination of features, R2 : 0.464004378132
We can also use more advanced regression techniques such as neural networks or support vector regression (SVR).

2. Dataset:

The features available to us, relating to the weather, do not contain enough information to better the ridership of subway. Also we need to feed more data to the model for better predictions. In order to gather and process larger amounts of data we may need to have concurrency model like MapReduce, where massive amounts of data are processed by many computers.

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

References

[1] Introduction to Statastics (https://www.udacity.com/course/viewer#!/c-st101)
[2] Discovering Statistics Using R (http://www.amazon.com/Discovering-Statistics-Using-Andy-Field/dp/1446200469)
[3] How the Mann-Whitney Test Works (http://graphpad.com/guides/prism/6/statistics/index.htm?how_the_mann-whitney_test_works.htm)
[4] One tail vs Two tail p-values (http://graphpad.com/guides/prism/6/statistics/index.htm?one-tail_vs__two-tail_p_values.htm)
[5] Histogram (http://www.itl.nist.gov/div898/handbook/eda/section3/histogra.htm)
[6] Linear Least Squares Regression (http://www.itl.nist.gov/div898/handbook/pmd/section1/pmd141.htm)
[7] ipython notebook, intro to data science project (http://nbviewer.ipython.org/url/www.asimihsan.com/articles/Intro%20to%20Data%20Science%20-%20Final%20Project.ipynb)
[8] Multiple Piazza forum sections (https://piazza.com/class/i4ltdrdhqak4r7?cid=102)