PROJECT REPORT

UDACITY

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. Which statistical test did you use to analyze the NYC subway data?

A: Mann-Whitney U test was used to analyze the data. Following procedure/steps and considerations were taken into account while applying the Mann- Whitney U test to the data set:

1.	Hypothesis Statement	H_0 = No. of Entries for rainy and non-rainy days are same H_A = No. of Entries for rainy and non-rainy days are significantly different
2.	Significance Level	$\alpha = 0.05$
3.	Direction of the test	Two-tailed Test

2. Why is this statistical test appropriate or applicable to the dataset?

A: A histogram representation was used for the initial explanatory data analysis of the NYC subway ridership with respect to rain. It showed (problem 3.1 or visualization 1) that data for entries with rain and entries without rain was skewed (non-normal) so a non-parametric test such as Mann-Whitney U test was used in this case where data was not assumed to be drawn from any particular probability distribution(eg: Normal distribution).

3. What results did you get from this statistical test?

A: The statistical test produced a low p-value of 0.024(one-tailed test). As the direction of the test is two-tailed test, p-value applicable to our test is 2 x 0.0249 = 0.049 (http://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.mannwhitneyu.html)

4. What is the significance of these results?

A: The p-value obtained $p<0.05$ from the test indicates greater statistical significance and stronger evidence for rejecting the null hypothesis which in turns suggests that the distribution of no. of entries between rainy and non-rainy days is statistically different.

Section 2. Linear Regression

1. What approach did you use to compute the coefficients theta and produce prediction in your regression model:

   a. Gradient descent (as implemented in exercise 3.5)

   b. OLS using Stats models

   c. Or something different?

A: For Problem 3.5, Gradient descent as implemented was used.

For Problem 3.8, OLS was implemented via Stats models using an R type formula

(http://statsmodels.sourceforge.net/devel/example_formulas.html)

2. What features did you use in your model? Did you use any dummy variables as part of your features?

A: For Problem 3.5, features used were rain, UNIT, precipi, Hour, mintempi & minpressurei

For Problem 3.8, features used were rain, UNIT, fog, precipi, mintempi & minpressurei

Dummy *Variable: For Problem 3.5, Gradient descent was implemented using the feature 'UNIT' as dummy variable. For Problem 3.8, OLS model was implemented using R type formula and 'UNIT' feature was treated as dummy variable using C() wrapper.*

3. Why are these features appropriate?

A: The features/variables I have selected are keeping in mind the effect they could have on the subway ridership in my opinion.

Response Variable: ENTRIESn_hourly

Predictor Variables: rain, UNIT, fog, precipi, mintempi & minpressurei

Following is a brief description on the basis of which I think these features are an appropriate choice:

Rain: was selected because people would avoid to get wet due to rain and use the subways instead.

Subway Station (UNIT): this feature was included as it drastically improved R^2. As 'UNIT' data is categorical, dummy variable implementation for UNIT was used using C(UNIT) (http://statsmodels.sourceforge.net/devel/example_formulas.html)

Fog: although, not directly but I think fog is related to rain and somewhat has the same effect on people. They would think of the weather getting worse and would prefer subways to get to their respective destinations.

Precipitation (precipi): Rain is liquid form of precipitation and I found it appropriate to be included as feature as it encompasses drizzle, freezing rain along with the regular rain we have.

(http://www.metoffice.gov.uk/learning/rain/what-is-precipitation)

Temperature (mintempi): I chose this feature because mostly temperature tends to fall when it rains and will have a similar effect on ridership as rain.

Pressure (minpressurei): I have included this feature as low pressure areas tend to have bad weather conditions such as rains, clouds and strong winds. (https://answers.yahoo.com/question/index?qid=20080817010439AA75zwb)

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

A: For Problem 3.5, $R^2 = 0.473$

For Problem 3.8, $R^2 = 0.451$

5. What does this R^2 value mean for the goodness of fit for your regression model?

A: R^2 is a statistical measure of how well the regression line approximates the actual data points. R^2 value closer to 1 is considered good so the values obtained in problems 3.5 and 3.8 suggest that the regression model will reasonably fit the data set.

6. Do you think this linear model is appropriate for this dataset, given this R^2 value?

A: Given the obtained R^2 value, I think the features that I chose for the linear model are appropriate as they show relation with rain and will have similar effect on the subway ridership, i.e. people would tend to use subways more often under the features described in section2 question 3. e.g: decrease in temperature will tempt people to take subways as they might think falling temperature will bring in rain.

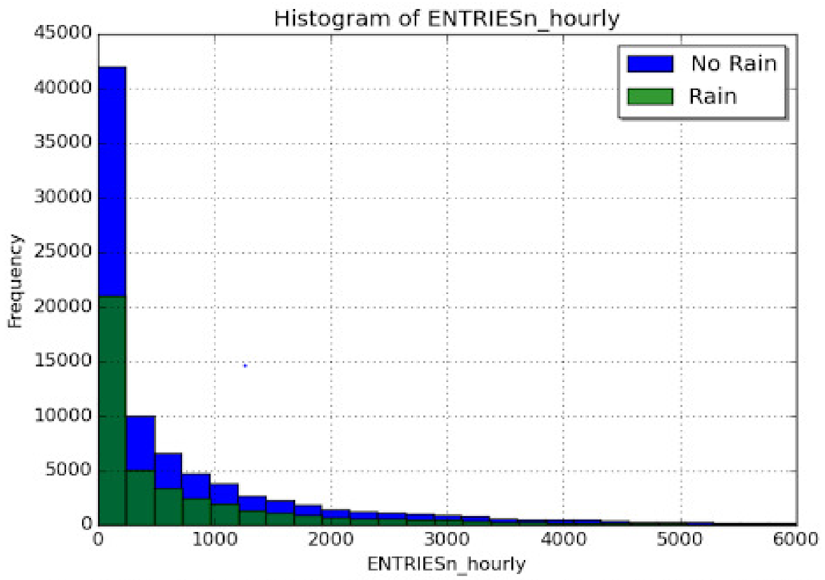
Section 3. Visualization

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., scatterplots, 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.

1. One visualization should be two histograms of ENTRIESn_hourly for rainy days and non-rainy days

A: Visualization 1 shows the 2 histograms comparing ENTRIESn_hourly for rainy days and non-rainy days. Visualization 1 is same as graph for the solution of problem 3.1

Visualization 1 (using Matplotlib)



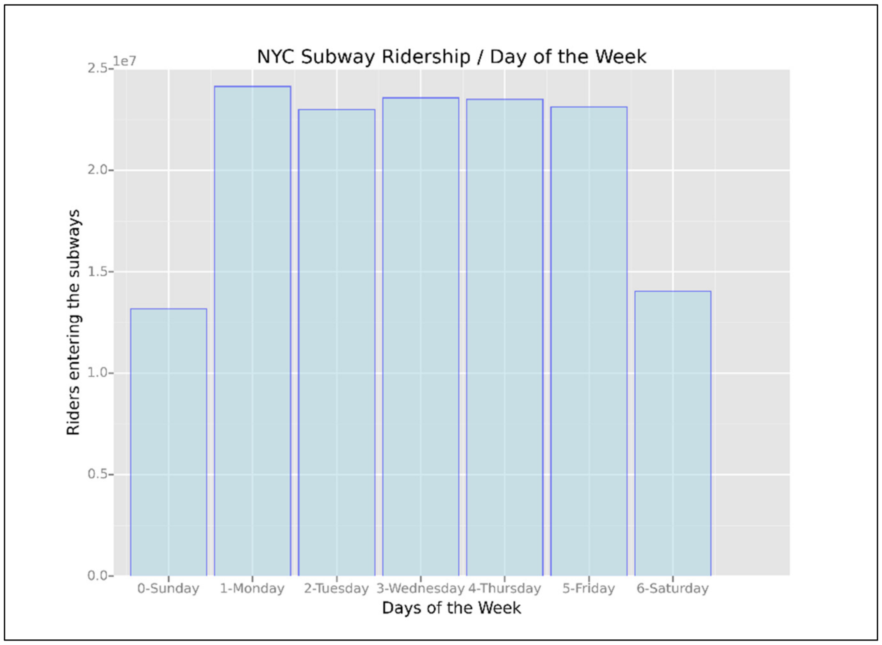
In the above figure, the two histograms compare hourly entries (ENTRIESn_hourly) into NYC subways for rainy days and non-rainy days.

2. One visualization can be more freeform, some suggestions are:
    1. Ridership by time-of-day or day-of-week
    2. How ridership varies by subway station
    3. Which stations have more exits or entries at different times of day

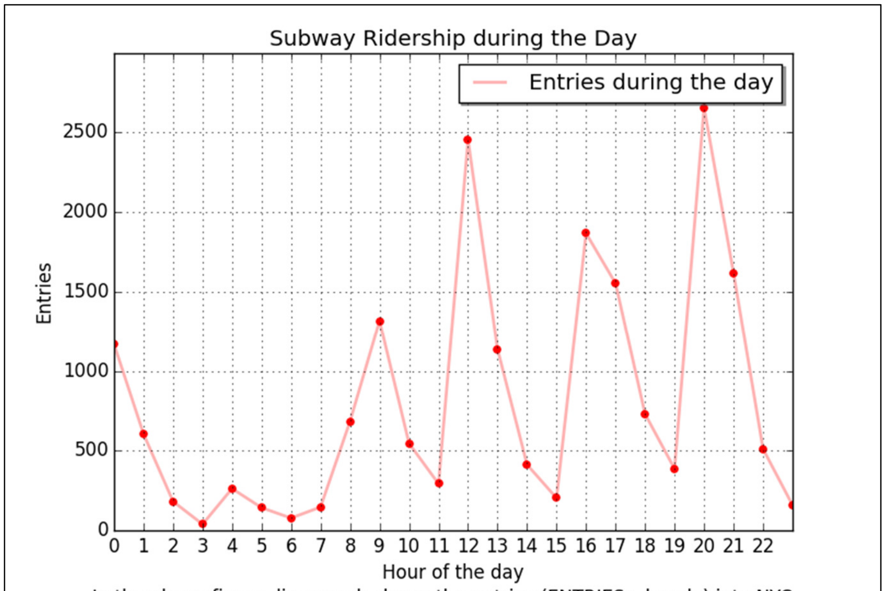
A: Visualization 2 shows Entries of riders into NYC subways for each day of the week. Visualization 2 is same as graph for the solution of exercise 4.1

A: Visualization 3 (Optional): I have add on my own, shows riders into NYC subways during the day.

Visualization 2 (using ggplot)



Visualization 3 (Optional) (Matplotlib)



In the above figure, line graph shows the entries (ENTRIESn_hourly) into NYC subways during 24 hours. It is clearly evident that ridership increases during the day time, i.e. from 9 (9:00 am) onwards.

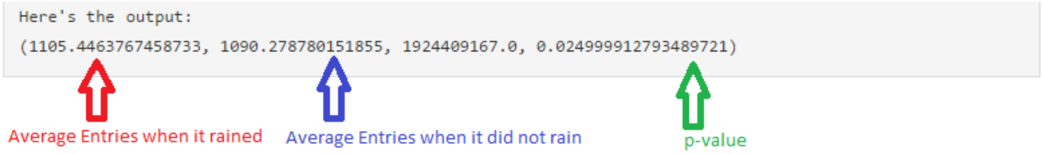
Section 4. Conclusion

Please address the following questions in details, and your answers should be 1-2 paragraphs long.

1. From your analysis and interpretation of the data, do more people ride the NYC subway when it is raining versus when it is not raining?

A: Based on the analysis and the statistical test performed on the data set, there was considerable evidence suggesting that the number of entries into NYC subways statistically differ for rainy and non-rainy days.

The implementation of the Mann-Whitney U test for this purpose was done (using the python scipy library http://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.mannwhitneyu.html) and the results are shown in figure 4.1 below:



```
Here's the output:
(1105.4463767458733, 1090.278780151855, 1924409167.0, 0.024999912793489721)
```

Average Entries when it rained
Average Entries when it did not rain
p-value

Figure 4.1

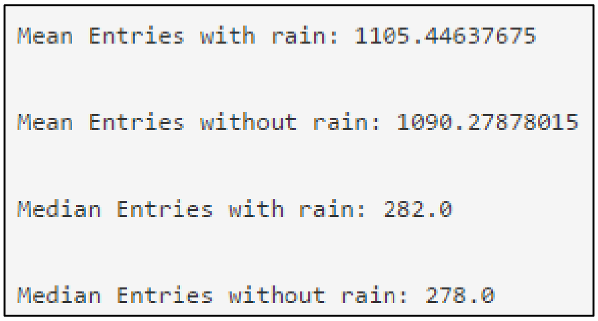
The p-value for the test, p = 0.0249 (for one-tailed test according to the scipy link above) needs to be doubled for the two-tailed test as per the direction of the test chosen in the section 1. Although marginally, p-value of 0.049 for the two-tailed test is less than the significance level set before the start of the test, i.e. p < 0.05, suggesting that the distribution of no. of entries between rainy and non-rainy days is different i.e. Ridership for rainy days != Ridership for non-rainy days.

In addition to this, the mean values for entries into subways for rainy and non-rainy days in figure 4.1 are 1105.5 and 1090.3 respectively. These values suggest that for the data set provided, on average people use subways slightly more when it is raining i.e. Ridership for rainy days > Ridership for non-rainy days.

2. What analyses lead you to this conclusion?

A: Analysis started by choosing Mann-Whitney U test as it is specifically performed for non-normal distributions instead of Welch's t-test. The p-value ($p < 0.05$) indicated greater statistical significance for rejecting the null hypothesis set in section 1 before conducting the test.

From our analysis and answer to section 4 question 1 we know that rain has an effect on the subway ridership. Furthermore, the results in figure 4.2 showed that people use subway more often when it rains i.e. no. of entries for rainy days are more than no. of entries for non-rainy days. As it can be seen in the figure below that mean and median for entries with rainy days are more than for non-rainy days (for the data set provided) suggesting that NYC ridership increases when it rains.



```
Mean Entries with rain: 1105.44637675


Mean Entries without rain: 1090.27878015


Median Entries with rain: 282.0


Median Entries without rain: 278.0
```

Figure 4.2

Linear regression was also performed in addition to the statistical test to further identify the contributing features towards increasing ridership in the NYC subways due to rain. As explained in section 2 question 3, the features for the linear regression were chosen appropriately thinking their relation with rain and effect they would have on the ridership. The R^2 value obtained from linear regression, for the features selected, suggests that people would tend to prefer subways more often in overcast conditions such as rain, fog/mist, cold temperatures, etc.

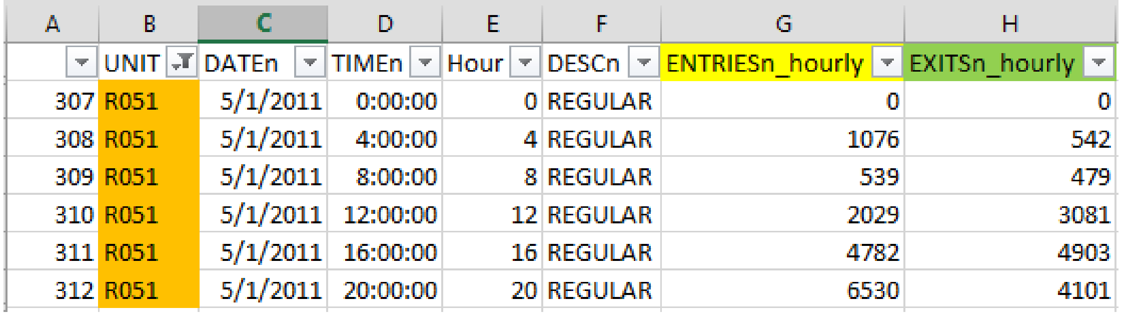
Section 5. Reflection

Please address the following questions in details, and your answers should be 1-2 paragraphs long.

1. Please discuss potential shortcomings of the data set and the methods of your analysis.

A: A few shortcomings which were observed are as follows (just observations):

- The data for no. of entries and exits originally was cumulative and logged at certain times and executing the piece of code in exercise 2.8 and 2.9, I think, has misinterpreted the data in a way. As shown in figure 5.1 taken from the 'turnstile_data_master_with_weather.csv' there are 0 entries at Hour = 0 and 1076 entries at Hour = 4 with all entries missing between the Hour 0 and 4. A better approach could have been to take average of the 2 values and assign them to the time stamps in between 0 and 4 (i.e. Hour 1, 2 and 3).



A	B	C	D	E	F	G	H
	UNIT	DATEn	TIMEn	Hour	DESCn	ENTRIESn_hourly	EXITSn_hourly
307	R051	5/1/2011	0:00:00	0	REGULAR	0	0
308	R051	5/1/2011	4:00:00	4	REGULAR	1076	542
309	R051	5/1/2011	8:00:00	8	REGULAR	539	479
310	R051	5/1/2011	12:00:00	12	REGULAR	2029	3081
311	R051	5/1/2011	16:00:00	16	REGULAR	4782	4903
312	R051	5/1/2011	20:00:00	20	REGULAR	6530	4101

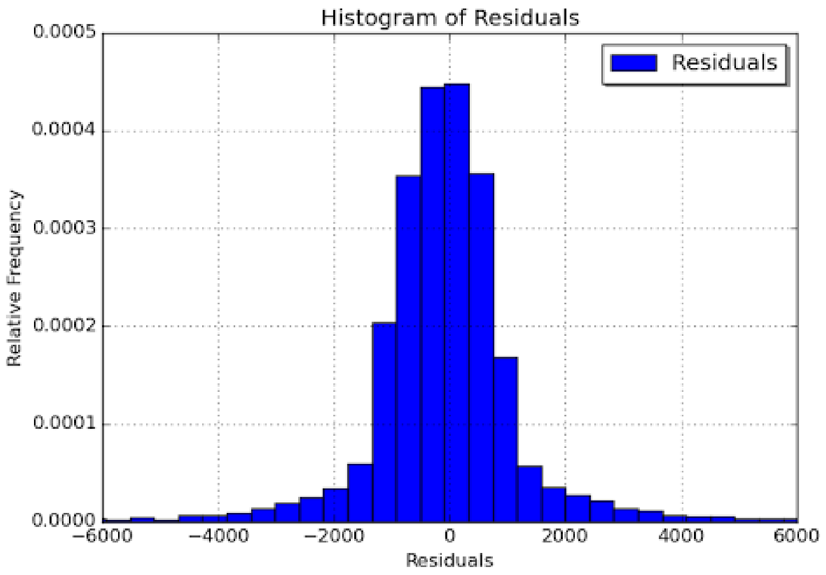
Figure 5.1

- We could include more data (collected over a period of several months) for better use of Mann-Whitney U test as it greatly depends on the sample size. This would finally help us get to a much better conclusion.
- During the wrangling phase in problem set 2.10 we converted time entries to the nearest hour. Although, this was done for the ease of data representation and interpretation but, I think, it overall sacrificed the time stamp details.

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

A: Few of the quick insights about the dataset are as follows (just observations):

- Although, the features chosen could be varied for better correlation but we can fairly suggest that the linear regression performed well as for the plot of residuals (problem set 3.6) showing that error in variance followed a normal distribution. Figure shown below:



In the above figure, the histogram is showing the difference between the actual and the predicted hourly entry data.

Figure 5.2

- Using suggested contributing factors such as vacations, seasons, weekends & holidays on a much larger data set (collected for several months) combined with output of reducer for problem 5.2 (Fog-Rain pair) could help get to a precise result regarding the ridership dependency on rain or weather in general.
- Subway station R549 was seen to have way more entries in the dataset than any other subway station. Having relatively comparable data entries would result in a better and accurate analysis.