Project-1

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?

*For the purpose of studying the data and exploring the data set I have used the **Mann-Whitney U test**, as this statistical test doesn't assume that the data is normalized. Also, as part of this test I used the 2 tailed P value. The null hypothesis is that the subway ridership doesn't wary based on weather conditions i.e. the subway ridership is consistent across rainy and non-rainy days*.

Null & Alternate Hypothesis:

H0 – Data is drawn from a normally distributed population
Ha – Data is not drawn from a normally distributed population

Two Tailed hypothesis test was done for this analysis with p-critical value of 0.025.

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.

We can determine the likelihood that our samples are drawn from normally distributed population by following method

1. Shape of the distribution

Below two histograms depict the distribution of hourly entries in our NYC subway data when it's raining and not raining. Shape of the both distributions is positively skewed. Hence we can prove that **data is not normally distributed**.

The below histograms tests prove that the data is drawn from the non-normal distribution. Hence we have to use a non-parametric test to compare our data. One such test is the **Mann-Whitney U test** which was used for the data analysis.

Legends:

Not Raining

Raining

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.

- *Mean Ridership on rainy days - 1105.4463767458733*
- *Mean Ridership on non-rainy days - 1090.278780151855*
- *One sided p-value - 0.024999912793489721*
- *Two sided p-value – 2*0.024 = 0.048*

1.4 What is the significance and interpretation of these results?

The P value of two sided is less than 0.05 which signifies that the 95% confidence level the null hypothesis that is the ridership is same for both rainy days and non-rainy days can be rejected. Meaning there is a significance difference in ridership between rainy 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:

2 Gradient descent (as implemented in exercise 3.5)
3 OLS using Statsmodels
4 Or something different?

OLS using statasmodel produced a higher coefficient value than gradient descent, which implies that the OLS model is more predictable than the gradient descent.

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 in this model are: rain, precipi, hour, fog, meantempi

Dummy variables used in the model: UNIT, day_week, conds, TIMEn

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."

Features:

- **Hour -** By plotting the total number of hourly entries and exits grouped by hour I found that there were spikes during several hours in a day
- **Meantempi –** If the temperature is high I assumed that more people will use subway
- **Rain –** I assumed that when it is raining more people will use subways
- **Fog -** I assumed that when it is foggy more people will use subways
- **Precipi –** With higher precipitation the chances of rain is higher therefore precipication and rain have a direct relationship and therefore I used it as one of my variables

Dummy Variables:

- **UNIT** – Unit is improving the R^2 value drastically. Hence included it in my model.
- **day_week** – Normally traffic is more during weekdays compared to weekends.
- **Conds**: Normally looking at the conditions of the whether people will decide whether to use subways or not.
- **TIMEn**: I used this dummy variable which helps to understand how people will use subway at different times.

2.4 The weights for the different features are as below

Hour	-1.43733045e+02
Meantempi	-7.17010557e+01
Rain	-1.06792970e+02
Fog	3.90683925e+02
Precipi	9.75138102e+00

2.5 What is your model's R2 (coefficients of determination) value?

R^2 value is 0.547721190841

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?

The R2 value explains the variability of the response data around its mean i.e. the R2 value of 48.7% means that the model explains 54.7% variability of the outcome data. Or 54.7% of variations in 'ENTRIESn_hourly' can be explained by the features rain, precipi, fog, hour, meantempi, UNIT, conds, TIMEn and day_week. With the dataset that has been provided I would assume that the model is appropriate. The only way to test it out further would be to train the model with more data or make the model more complex. The current model is complex enough as it has multiple features, the other thing that I have not tried but I would assume can be tried is to use some sort of polynomial regression instead of a linear regression.

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. Ridership by time-of-day or day-of-week

Legends:

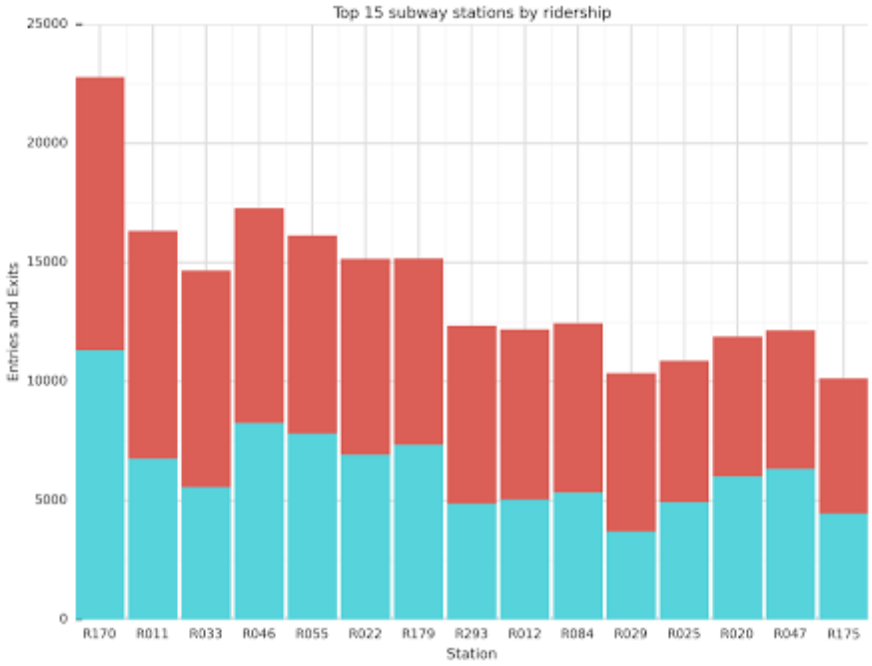
Raining

Not Raining

The above graph shows different ranges in the x-axis which constitute a range to represent volume of ridership and y-axis represents the frequency of people using the subway for these given ranges. The blue colored histograms are the frequency of ridership when it is raining and the orange are the ones when it is not raining.

3.2. One visualization can be more freeform, some suggestions are:

2. Which stations have more exits or entries at different times of day



Legends:

Mean Entries

Mean Exits

The above graph shows the top 15 subway stations based on the hourly entries and exits from each of the subways stations. The above graph represents the mean value of ridership per subway station and not the actual count of entries and exits. The red coloured bars represent mean exits per subway station and the blue coloured bars represent mean Entries per subway stations.

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?

By analyzing the data visually and by performing statistical tests it was observed that more people used the subway when it was raining compared to when it was 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.

Looking at the below table by doing some simple data exploration it can be observed that the median and mean number of entries and exits per hour are higher when it is raining than when it is not raining.

			ENTRIESn_hourly	EXITSn_hourly
not raining	count	87847	87847	87847
	mean	0	1090.27878	883.25961
	std	0	2320.004938	1998.516762
	min	0	0	0
	25%	0	38	31
	50%	0	278	231
	75%	0	1111	846
	max	0	43199	45249
raining	count	44104	44104	44104
	mean	1	1105.446377	894.123572
	std	0	2370.527674	2028.552487
	min	1	0	0
	25%	1	41	33
	50%	1	282	235
	75%	1	1103.25	849
	max	1	51839	41503

Also, on performing the Mann Whitney U-Test on the sample dataset it was observed that the null hypothesis i.e. the ridership remains the same for both rainy and non-rainy days can be rejected as the p-value was less than 0.05. The OLS model that was built predicted the same outcome with a 46.11% stability that more people use the subway when it raining compared to when it is not raining.

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:

Some of the potential shortcomings of the dataset are:-

- The data collected for different subways are at different times of the day and they are not uniform.
- There should be data present per subway station for the same time but collected for both rainy and non-rainy days which are not present in the current data set. This is with the assumption that the same set of people travel on the same routes
- The subways stations can have an additional geographic dimension to show whether the subway station is near a residential neighborhood or near to workplace.

Some of the potential shortcomings of the analysis method are:-

- My analysis is purely based on a liner model. I think applying a nonlinear model with composite features would lead to a better fit over the sample data.

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

References:

Udacity website: https://www.udacity.com/nanodegrees

Blogs: https://piazza.com/class

Nanodegree Phython course : https://www.udacity.com/course/ud036

Sample Project:
http://nbviewer.ipython.org/url/www.asimihsan.com/articles/Intro%20to%20Data%20Science%20-%20Final%20Project.ipynb

Python Course: https://developers.google.com/edu/python/introduction

Pandas: http://www.gregreda.com/2013/10/26/intro-to-pandas-data-structures/

ggplot : https://pypi.python.org/pypi/ggplot/

Statistical Learning http://www-bcf.usc.edu/~gareth/ISL/book.html

Pandas tutorial : http://pandas.pydata.org/pandas-docs/stable/tutorials.html

machine learning:
http://openclassroom.stanford.edu/MainFolder/CoursePage.php?course=MachineLearning

(http://openclassroom.stanford.edu/MainFolder/CoursePage.php?course=MachineLearning)

Statistics
https://www.youtube.com/watch?v=JvS2triCgOY&src_vid=zPG4NjIkCjc&feature=iv&annotation_id=annotation_742878

(https://www.youtube.com/watch?v=JvS2triCgOY&src_vid=zPG4NjIkCjc&feature=iv&annotation_id=annotation_742878)

Learn Python hard way: http://learnpythonthehardway.org/book/

Internet searching through google.

Data Analysis in google.

topic: http://www.statsoft.com/Textbook/Multiple-Regression#cresidual

Here is another helpful blog post on R squared

values: http://blog.minitab.com/blog/adventures-in-statistics/how-high-should-r-squared-be-in-regression-analysis