Project 1

Analyzing the NYC Subway Dataset

Aishwarya Venketeswaran

Udacity's Data Analyst Nanodegree

Table of Contents

Section 0. References

1. https://www.udacity.com/course/ud359-nd

2. https://docs.python.org/2/

3. http://stackoverflow.com/questions/21098350/python-iterate-over-two-lists-simultaneously

4. http://stackoverflow.com/questions/4576115/python-list-to-dictionary

5. http://stackoverflow.com/questions/1841565/valueerror-invalid-literal-for-int-with-base-10

6. https://www.khanacademy.org/math/probability/statistics-inferential/hypothesis-testing/v/one-tailed-and-two-tailed-tests

7. http://www.graphpad.com/guides/prism/6/statistics/index.htm?how_the_mann-whitney_test_works.htm

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

Section 1. Statistical Test

1.1 The statistical test that I used to analyze the NYC subway data is the Mann-Whitney U-test. I used the two-tailed p value.

-Null hypothesis in this case is that the two distributions that are being compared (ridership on rainy days and ridership on non-rainy days) are statistically the identical.

-Alternative hypothesis is that the ridership on rainy days is different from that on non-rainy days [two-tailed].

$$H_0 : P(x > y) = 0.5$$

$$H_1 : P(x > y) \neq 0.5$$

-P-critical value is 5% or 0.05 (chances of getting a value at least as extreme as the observed)

1.2 Mann-Whitney U test is applicable to this data set because:

- The curves (ridership vs. rainy days and ridership vs. non-rainy days) are not normally distributed. Because the sample is not normal, we cannot apply the t-test in this case.

- Mann Whiney- U test is a non-parametric test (does not assume data to be from any particular probability distribution).

- In this case, both the samples (i.e. the sample for the null hypothesis as well the sample for the alternative hypothesis) are from the same population.

1.3 Results from this statistical test are:

Mean of Entries per hour with rain: 1105.4463767458733

Mean of Entries per hour with no rain: 1090.278780151855

U- value: 1924409167.0

p-value: 0.024999912793489721 *2 = 0.0499

1.4 Significance of these results is that it tells us that the two samples under consideration (entries per hour with rain and entries per hour with no rain) are statistically different. Because the p-value (0.0499) is less than the p-critical value (0.05), we can reject the null hypothesis at p-critical value of 0.05 and conclude in favor of the alternative hypothesis. Hence, we can conclude that there is a difference between the ridership on rainy and the ridership on non-rainy days.

Section 2. Linear Regression

2.1 The approach I used to compute the coefficients theta and produce prediction for ENTRIESn_hourly in my regression model is gradient descent (implemented in exercise 3.5).

2.2 The features (input variables) in my model are: 'rain', 'fog', 'Hour', 'meantempi', 'meanwindspdi', 'precipi'. Only one dummy variable was used – Unit.

2.3 I selected these features in my model because of the following reasons:

- 'Rain' is included because I thought (logically) if it is raining people might decide to use the subway.
- 'Hour' is included in my model because through experimentation I realized that using this feature increased my R^2 value drastically. In fact, using only ‘Hour’ in my list of features gave me a R^2 value of 0.465.
- 'Fog' is included in my model because I thought that when it is foggy, people will use the subway more often.
- 'Meantempi' signifies the temperature and I thought that if it is too cold, people will decide to use the subway due to difficulty in driving.
- 'Meanwindspdi' is included in the model because when it is windy, it is difficult to drive. Hence, I thought that people will choose to use the subway on such days.
- 'Precipi' is included because I thought that depending on the amount of rain (precipitation), people may decide to use the subway.

2.4 The coefficients (or weights i.e. thetas) of the non-dummy features in my linear regression model is :

```
rain               0.040560

fog              214.086883

precipi          -73.976935

Hour              65.364525

meantempi         -9.491420

meanwindspdi      32.568316
```

Intercept: 1282.97263113

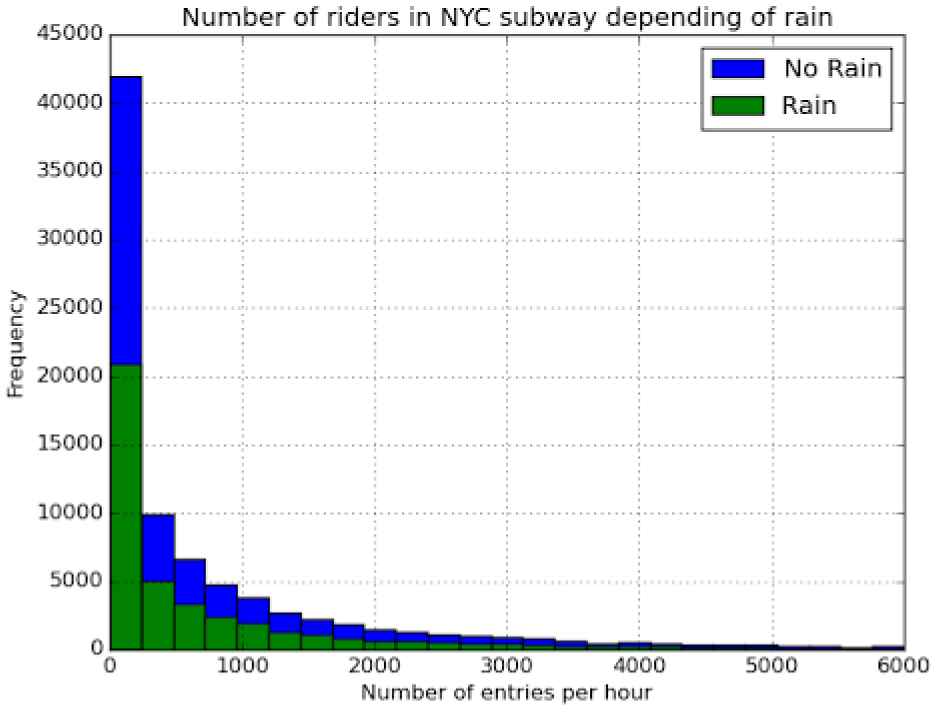
2.5 My model’s R^2 value is 0.480456675828

2.6 This R^2 value means that my linear regression model is a decent way to predict the ridership. The closer the R^2 value is to 1, the better our model is. R^2 value is measure of how well the model fits the data. Hence, I think this linear model to predict ridership is appropriate for this dataset. However, we could have made the R^2 value even better by taking many more features into account such as – may be the variations in cost of gasoline over time and the amount of traffic could affect ridership. Hence, considering the features we were given to predict ridership, I think my model does a good job.

R^2 measures how well the model fits the data. It is 1-ratio of residual variability. In this case R^2 value is 0.48. In this case, because the predictors (: 'rain', 'fog', 'Hour', 'meantempi', 'meanwindspdi', 'precipi') and the independent variable (ENTRIESn_hourly) are related to a certain extent as depicted by our model, the ratio of residual variability is equal to 0.52. Therefore, the R^2 value is 0.48.

Section 3. Visualization

3.1 This visualization (Exercise 3-1) shows the relationship of ENTRIESn_hourly for rainy days and non-rainy days:

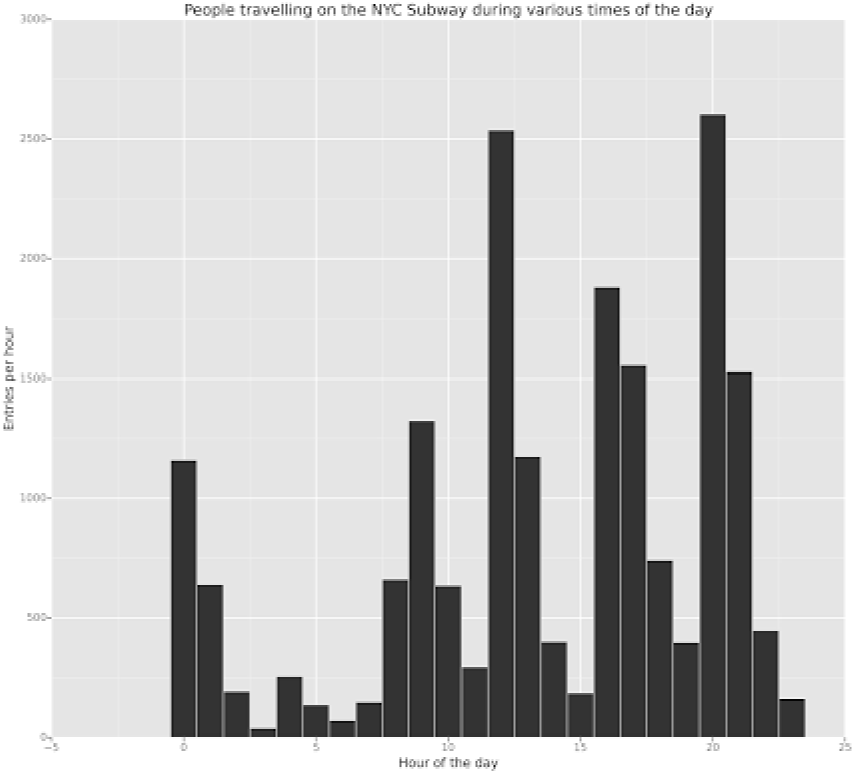


Insights from this visualization are:

- It seems like the number of entries per hour on non-rainy days are higher than that of rainy days. But, this is not true because the number of data points for the “non-rain” are significantly higher than that of “rain”. Hence, this conclusion may not be correct.
- The higher the number of entries per hour, the lower its frequency. (This makes sense because fewer hours have higher usage).
- There are several outliers in the data. Hence, the graph has been cut at about 6000.

Code used to produce this visualization is there in Exercise 3 – 1 [Visualization was created using Matplotlib]

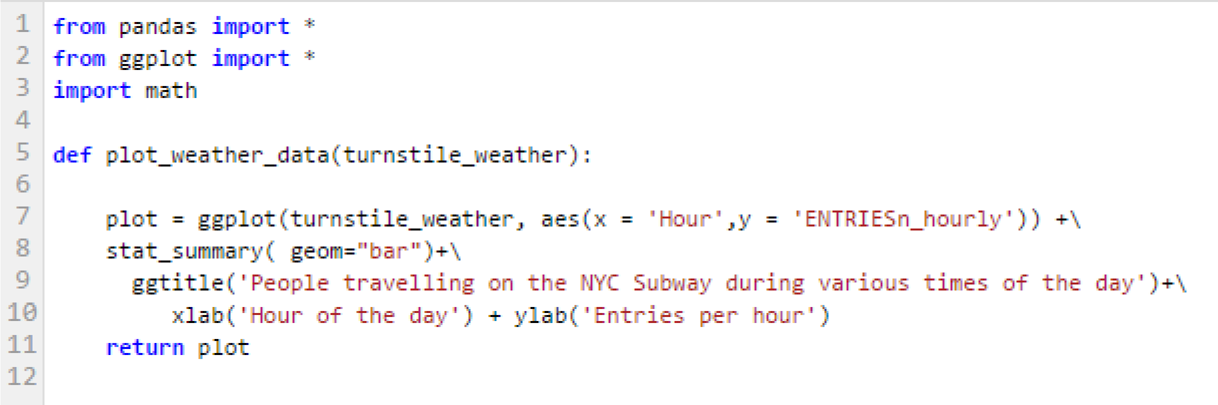
3.2 Visualization (Exercise 4-1) showing the relationship between ridership and hour of the day:



This visualization provides us with a couple of insights:

- The number of entries per hour almost in the same range most of the time (0-1200) throughout the day except at these times of the day:
    - 7.0hrs – 9hrs i.e 7:00AM to 9:00AM (probably because people are travelling to their workplaces, schools, universities, etc.)
    - 12hrs to 13hrs i.e 12:00PM to 1:00PM (could be because people are going to have lunch)
    - 16hrs to 18hrs i.e. 4:00P.M. to 6:00P.M. (could be that people are going back home)
    - 20hrs to 22hrs i.e. 8:00P.M. to 10:00P.M. (could be that people are going back home)
- We must note that amongst the busiest times (5 categories noted above), the number of entries (per hour) is highest in the last category: 8:00P.M. to 10:00P.M.. (This could be that people are travelling together (in cars/buses) during the morning to go to their destinations but travelling back home using the subway)

Code used to produce this visualization is there in Exercise 4-1:



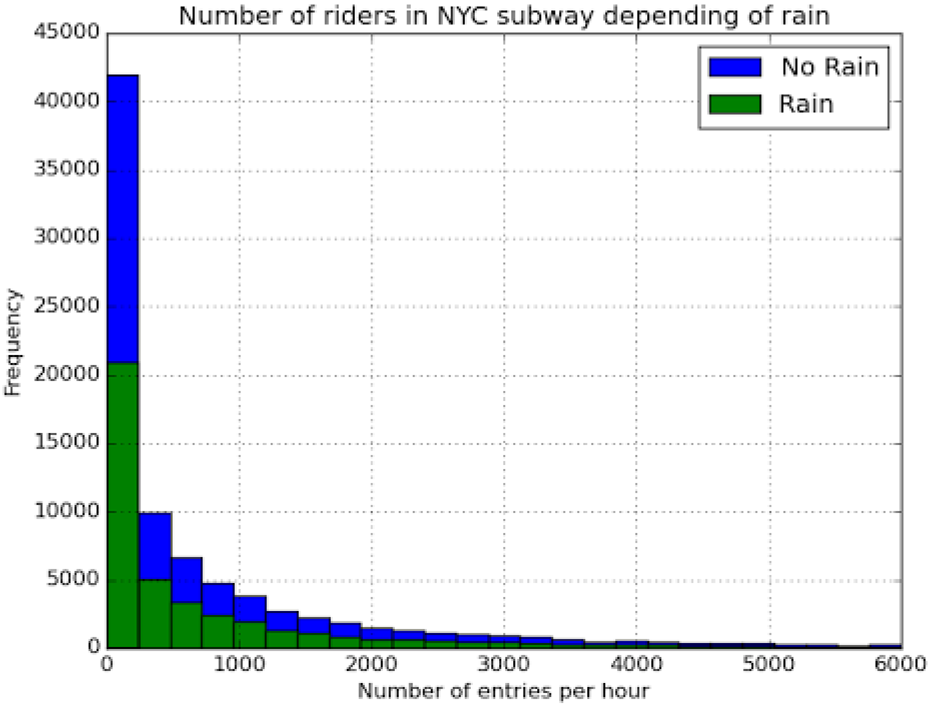
```python
from pandas import *
from ggplot import *
import math

def plot_weather_data(turnstile_weather):

    plot = ggplot(turnstile_weather, aes(x = 'Hour',y = 'ENTRIESn_hourly')) +\
    stat_summary( geom="bar")+\
      ggtitle('People travelling on the NYC Subway during various times of the day')+\
         xlab('Hour of the day') + ylab('Entries per hour')
    return plot

```

Section 4. Conclusion

4.1 From my analysis and interpretation of the data, there is a difference in ridership of the NYC subway when it is not raining vs. when it is raining. Because the p-value (0.0499) is less than the p-critical value (0.05), we can reject the null hypothesis at p-critical value of 0.05 and conclude in favor of the alternative hypothesis – there is difference between ridership depending on rain.

4.2 Here are the analyses that led to this conclusion:

This is the visualization from Problem Set 3-1 again-



- This visualization allows us to conclude that at all times, the number of riders when it is not raining is different from the number of riders when it is raining.
- The mean of the entries with rain is equal to 1105.4463767458733, but the mean of the entries without rain is 1090.278780151855. After doing the Mann-Whitney U test on the data, I found that the U-statistic is 1924409167.0 and the p value is 0.0499. This U-statistic (which is large) along with the p value tells us that the sample from rainy and not rainy days is different. Since the p-value (0.0499) is less than the p-critical value (0.05), we can conclude that there is a statistical difference in the ridership on rainy days vs. that on non-rainy days.
- After using the linear regression model (Problem Set 3- 5), the r^2 value that I computed is 0.463968815042 (close to 1). Hence, our prediction performs pretty well.
- It can be noted that in this sample dataset, the number of rainy days is significantly fewer than the number of non-rainy days. Due to this fact, the numbers [i.e. entries in this case] for the non-rainy days seems exaggerated. Due to the small sample size of the rainy days

[in comparison to the non-rainy days], the conclusions drawn from this small dataset might not be completely accurate.

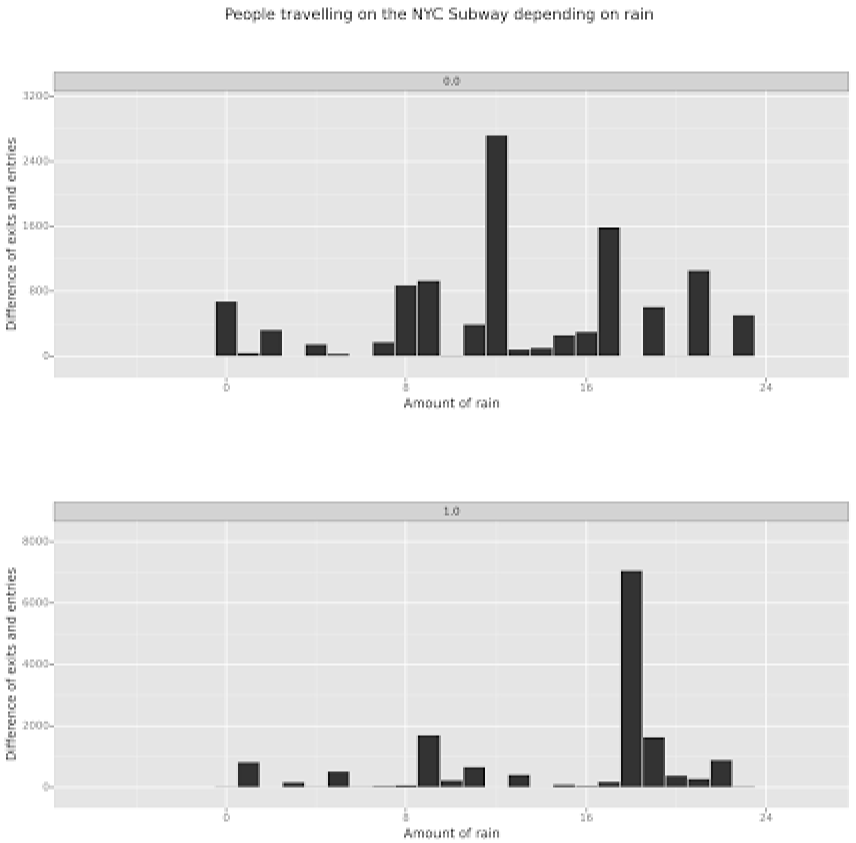
- Since the mean of the entries with rain is equal to 1105.4463767458733, and the mean of the entries without rain is 1090.278780151855, we can conclude that the ridership during rainy days is higher than that of non-rainy days.

Section 5. Reflection

5.1 There are some potential shortcomings of the methods of analysis, including:

1. Dataset: Due to the limitations of the server, only 1/3 of the set is randomly being given for analysis (Problem Set 4). Hence, due to random selection of a subset, the selection might be skewed. Hence, the analysis that is performed on this set could have errors. We must note that we are trying to use the model (by looking the random selection) to decide whether it can be applied to the whole dataset. The sample might contain outliers or it could not be a true representation of the whole dataset. We must consider these factors.
2. Analysis: The use of tests such as the linear regression model or statistical test definitely influences the analysis and the conclusion. The model that we are making does not take into consideration the several factors that affect ridership. We are looking at only some of the factors such as rain, fog, etc. Also, the entries per hour and exits per hour could have some kind of relationship with each other that we are not considering in our model.

5.2 Visualization showing how the number of people using the NYC Subway varies with rain:

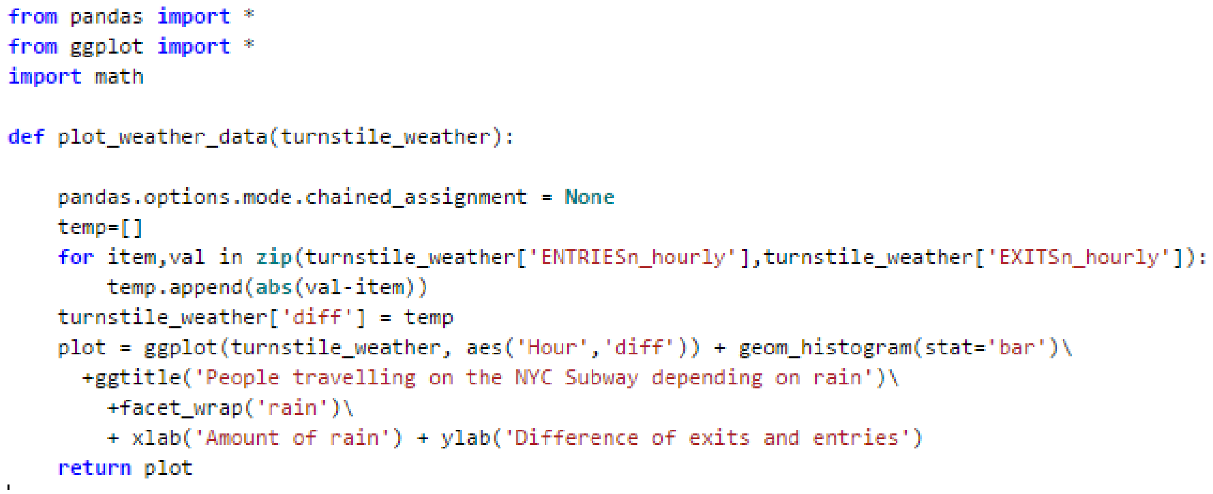

People travelling on the NYC Subway depending on rain

This visualization shows us a couple of interesting details:

- When there is rain, the difference between the entries per hour and the exits per hour is lower. When there is rain people might be using the subway for two-way trips and properly recording all the entries and exits. The difference is highest at 5:00PM probably because people are mostly entering the subway at this time [to go back home].

- On the other hand, when there is no rain, the difference between exits and entries is higher. The difference is highest at about 10:00AM. This might be because people are mostly entering the subway at this time to go their workplace.

Code used to produce visualization is there in Exercise 4-2:



```
from pandas import *
from ggplot import *
import math

def plot_weather_data(turnstile_weather):

    pandas.options.mode.chained_assignment = None
    temp=[]
    for item,val in zip(turnstile_weather['ENTRIESn_hourly'],turnstile_weather['EXITSn_hourly']):
        temp.append(abs(val-item))
    turnstile_weather['diff'] = temp
    plot = ggplot(turnstile_weather, aes('Hour','diff')) + geom_histogram(stat='bar')\
      +ggtitle('People travelling on the NYC Subway depending on rain')\
        +facet_wrap('rain')\
        + xlab('Amount of rain') + ylab('Difference of exits and entries')
    return plot
```