ANSWERS TO THE SHORT QUESTIONS OF UDACITY'S NANODEGREE IN BIG DATA FOUNDATION : -

PROJECT TOPIC : - ANALYSIS OF NYC – SUBWAY – DATASET

E – Mail :- sonamlpu657@gmail.com

❖ STASTICAL TEST : -

1. TYPE OF TEST STATISTICAL TEST USED : -

I have used Mann- Whitney U Test for analysing the NYC Subway Dataset. In order to Analyse the effect of independent variables such as "fog" ,"rain" ,"time_of_day", "day_of_time" and "weekdays" on dependent variables such as "ENTRIESn_hourly", "EXITSn_hourly" .

2. USE OF ONE TAIL OR TWO TAIL : -

Since, we are dealing with the two different cases : use of subway in rainy day as well as use of subway in non- rainy day/ clear day. Therefore, it is considered as two sided problem because we cannot say which one sample is higher than the other as both are are not equal.

3. NULL HYPOTHESIS : -

NULL hypothesis means that the no of people entering the subway in raining day is equal to the no of people entering the subway in non- raining day/ clear day.

4. WHY THIS TEST ?

I have used this test due to the following reasons : -

1. Since, the dataset given is not normally distributed and many other tests assumes Normal distribution and Mann-Whitney Test does not assume normal distribution.
2. Observations are independent from each other.
3. Given sample size is more than 20.
4. There are two groups to compare i.e., two tailed test (Raining V/s Not Raining).
5. Response is continunous.

6. RESULTS FROM THE STATISTICAL TEST : -

With Rain : - 1105.446377
Without Rain : - 1090.278780
U:1924409167.0,

One – Tailed p: 0.02499991279
Average mean value of ENTRIESn_hourly in rainy days is slightly more than Without rain.

5. SIGNIFICANCE AND INTERPRETATION OF RESULT : -

According to the following tests, we conclude that the impact of "fog" ,"rain", "time_of_day" , "day_of_time" is very less as compared to the impact of weekdays on the ridership of subway for the given data set. As the no of people entering the subway in holidays i.e ., Saturday and Sunday is less than the regular days.

❖ REFERENCES : -

1. The main source code of the file can be accessed through link: http://localhost:8888/notebooks/Anaconda3/Untitled.ipynb.
2. It can also be access on github through the given link : - https://github.com/Sonam06/NYC-SUBWAY-DATA-ANALYSIS
3. I have used Udacity's Nanodegree course material for completing this project and most of the help was provided by udacity to complete different problems and problem sets of Big Data Foundation course.

❖ LINEAR REGRESSION : -

1. APPROACH USED TO COMPUTE THE COEFFICIENTS OF REGRESSIO -N MODEL : -

I have used plain linear regression model with default parameters value.

2. FEATURES USED IN MODEL : -

I have performed plain regression on the given dataset i.e ., turnstile_data_ Master_with_weather.csv. Rain, tempi, windspdi, meandewpti, meanpressurei Fog, thunder, minpressurei, maxpressurei are the features included in the model. The dummy feature is ENTRIESn_hourly and EXITSn_hourly.

3. COEFFICIENTS OF NON- DUMMY FEATURES : -

RAIN : 115.68829
TEMPI : -112.69670
FOG : - -74.116274
PRESSUREI : - -377.723
WINDSPDI : - 10.8248
MEANTEMPI : - -54.1723

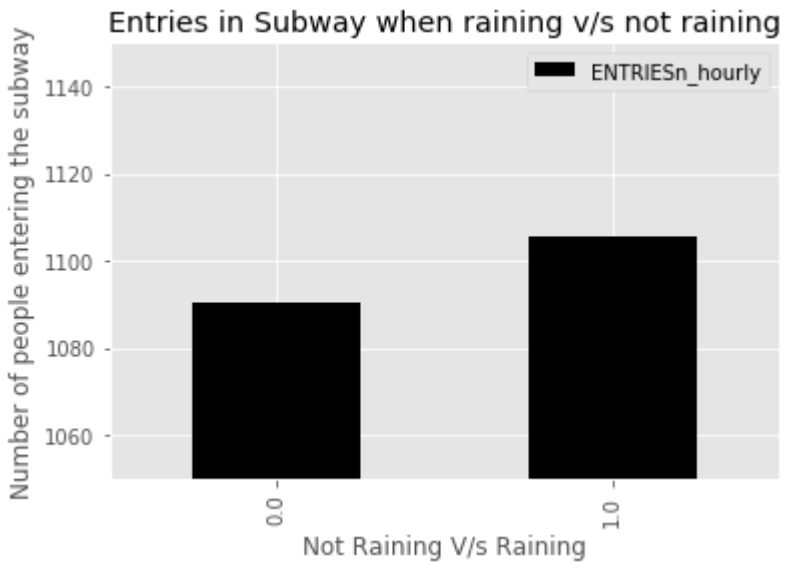
MEANPRESSUREI : - 334.79693
MEANWINDSPDI : - -101.67422

4. R2 VALUE : -

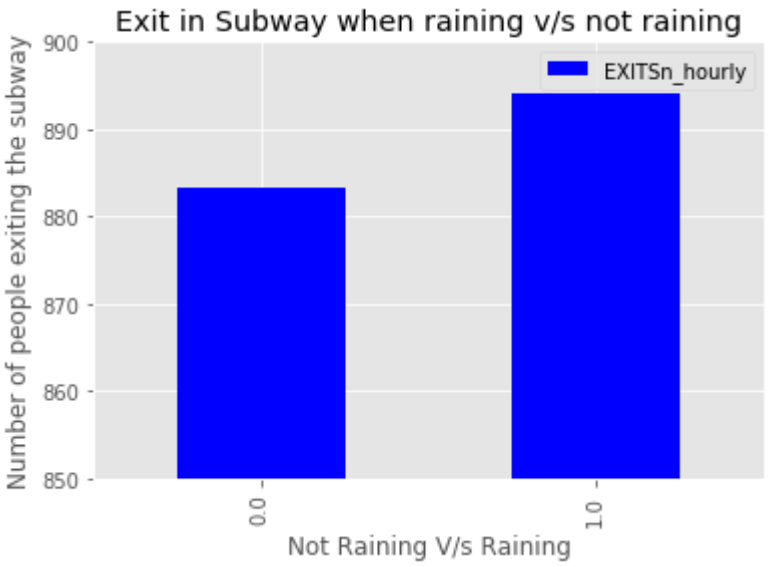
0.148483 is the R2 value for this model.

❖ VISUALISATION : -

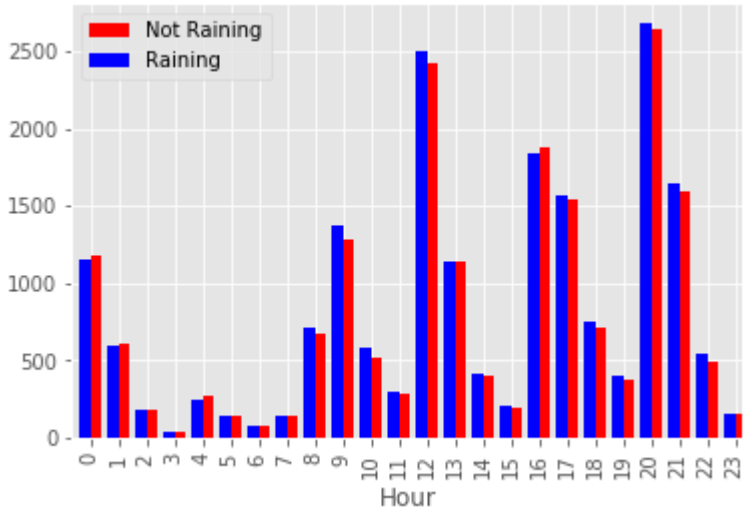
1. ENTRIESn_hourly for both rainy as well as non- rainy days : -



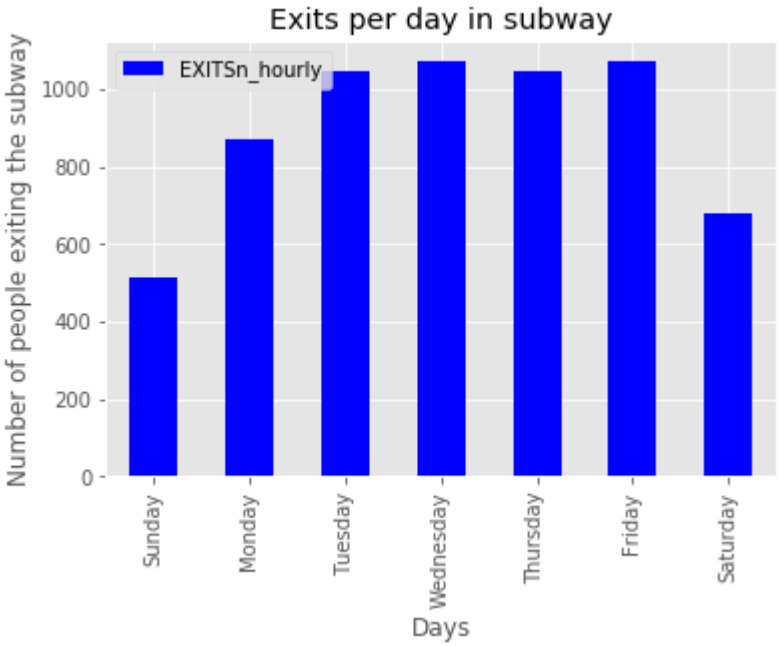
2. EXITS_n hourly (exit in subway) for both rainy day as well as non- rainy day : -



3. TIME OF DAY : -



4. DAY OF WEEK : -



❖ CONCLUSION : -

1. After analysing the given data set, we can conclude that the no of people entering the subway while not raining is more as compared to the no of people entering the subway while not raining. And also, the no of people entering the subway in weekdays is more than the holidays i. e., Saturday and Sunday.

2. I have used Mann-Whitney U Test and linear regression for finding the relation-Ship between rain and riding the subway. The p value is 2*0.02499991279 Which is nearly equal to 0.05 i.e., 95 % and also the coefficient of linear regression Model is positive. Hence, the analysis is correct.

❖ REFLECTION : -

POTENTIAL SHORTCOMINGS OF THE METHODS : -

1. Potential shortcomings can be applied for both dataset and analysis.

1.1 As, the given dataset covers single month i.e., month of june which is too short. There might be the possibility of certain coincidence or some event which is Affecting the ridership of subway. For ex : - We use umbrella only when there is a possibility of rain. As we know that weekdays and rain are correlated, which affect the accuracy and result of test.

1.2 Also, the regression model which we have used i.e., linear regression model is only Applicable when there is a linear relationship with the features used in model. And If it failed to meet the linear relationship, the result may violate for ex., when there Is a extreme low or high temperature, people tend to be in their home which linear Regression cannot map.