Analyzing the NYC Subway Dataset

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:

To analyse the NYC subway data the Mann-Whitney U-Test was used. A two-tail P value was used. The null hypothesis used is that the mean number of entries to subway on rainy days (μ_1) is equal to the mean number of entries to subway on days without rain (μ_2)

H_0: $\mu_1 = \mu_2$

H_a: $\mu_1 \neq \mu_2$

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 the Mann-Whitney U-Test is applicable

a) The ridership is being compared between 2 sets of sample data
b) The sample data of the ridership on rainy days and days without rain does not follow a Normal distribution
c) All the Observations in each of the group are independent of the other group i.e. the ridership on the rainy days is independent of the ridership on the days without rain
d) The ridership (rainy / non-rainy days) is Ordinal and can be easily ranked
e) Mann-Whitney U-Test is a non-parametric test and does not assume any particular underlying distribution. Hence, it can be applied in this case

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:

p-value = 0.024999912793489721 ≈ 0.025

Mean for the sample for ridership on Rainy Days = 1105.4463767458733 ≈ 1105.45

Mean for the sample for ridership on Days without rain = 1090.278780151855 ≈ 1090.28

1.4 What is the significance and interpretation of these results?

Answer:

The p-value ≈ 0.025

This is the p-value for one tail

For 2-tails this value will be 2* 0.025 = 0.05

Considering, $\alpha = 0.05$, the p-value is in the critical region signifying that our test result is significant

Hence, we reject the null hypothesis and go with the alternative hypothesis i.e. there is a difference in the ridership during the rainy days and the 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:

The coefficients theta and the prediction for ENTRIESn_hourly was computed using Gradient descent and the OLS method

2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?

Answer:

The features used in the model are UNIT, Hour, meantempi, meanwindspdi

For UNIT, dummy variables are being used

2.3 Why did you select these features in your model?

Answer:

The features which have been selected in the model when compared individually to the response variable ENTRIESn_hourly have higher values for the co-relation coefficient (r) as compared to other features. Also, when these features were used in the model the value of R^2 got improved. Hence, they have been selected in the model

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

Answer:

Using the Gradient Descent algorithm, the coefficients of the non-dummy features is as below:

Hour = 561.28444129

meantempi = -770.94677669

meanwindspdi = 817.96385325

y-Intercept = 1201.58263793

Using the OLS algorithm, the coefficients of the non-dummy features is as below:

Hour = 385.19148093

meantempi = -35.02903267

meanwindspdi = 84.94463678

y-Intercept = 1065.2421

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

Answer:

The value of R^2 using the Gradient Descent algorithm is 0.4721 (Number of iterations = 90, alpha i.e. step size = 0.1)

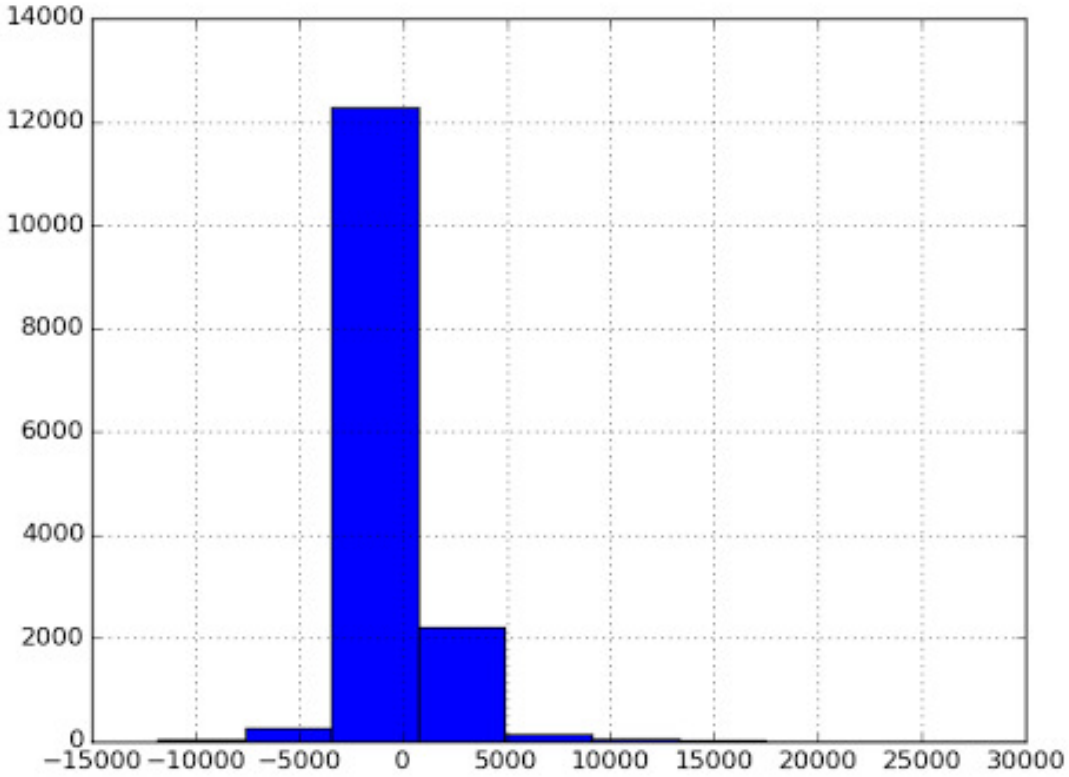
The value of R^2 using the OLS algorithm is 0.4483

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:

Given the value of R^2, 45 - 47% of the variation in the response variable (ridership) is explained by the Predictor variables. The goodness of fit of the regression model is moderate.

The linear model used to predict ridership is appropriate for this dataset as the distribution of the residuals obtained by subtracting the Observed responses from the Predicted responses is approximately a normal distribution with a mean around zero and a constant variance as can be observed from the figure below



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.

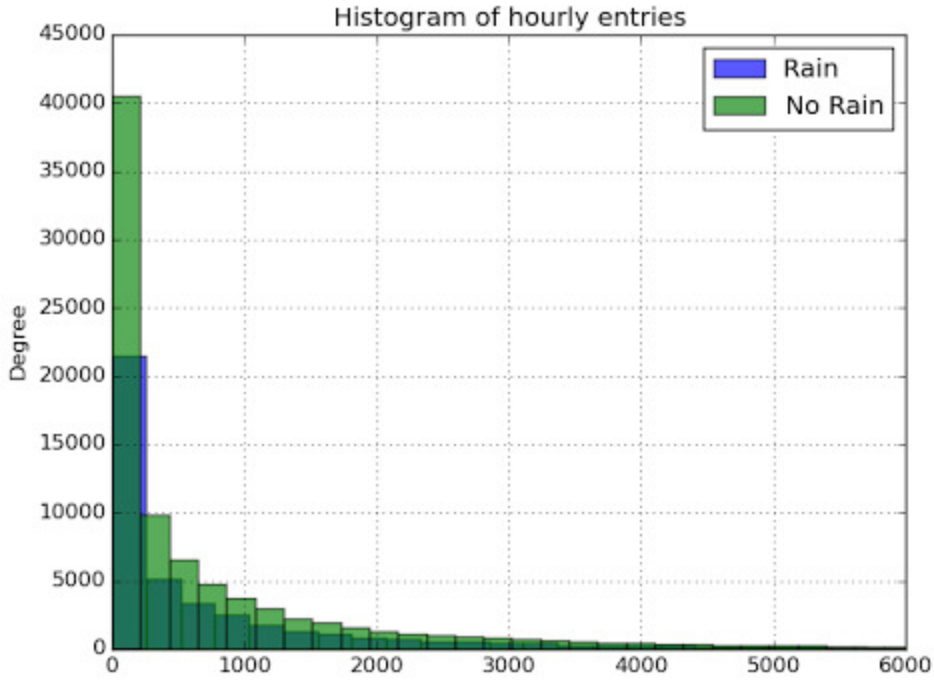
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days

Answer:

Python Code:

```
import numpy as np

import pandas

import matplotlib.pyplot as plt

def entries_histogram(turnstile_weather):

    plt.figure()

    # plot a historgram for hourly entries when it is raining

    t1 = turnstile_weather[turnstile_weather['rain'] == 1]

    t1['ENTRIESn_hourly'].plot(kind='hist',label='Rain',title='Histogram of hourly entries',legend=True,bins=200,alpha=0.65)

    # plot a historgram for hourly entries when it is not raining

    t2 = turnstile_weather[turnstile_weather['rain'] == 0]

    t2['ENTRIESn_hourly'].plot(kind='hist',label='No Rain',title='Histogram of hourly entries',legend=True,bins=200,alpha=0.65)

    plt.xlim(0,6000)

    return plt
```

Graph:



Comments:

This is a Histogram of number of passengers entering the subway hourly showing comparison between when it is raining v/s when it is not raining. It can be seen that the frequency is almost double near zero when it is not raining as compared to when it is raining and this ratio is decreasing as we move away from zero although the frequency remains high for the “no rain” cases even away from zero. This indicates that in the dataset the number of records with “no rain” is more than the “rain” cases. As the ratio of “no rain” to “rain” is higher near zero and is decreasing away from zero we can say that more passengers take the subway when it is raining than when it is not raining but as we move away from zero there is no significant difference in the ridership which shows that there are other factors which have an effect on the ridership of the subway

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

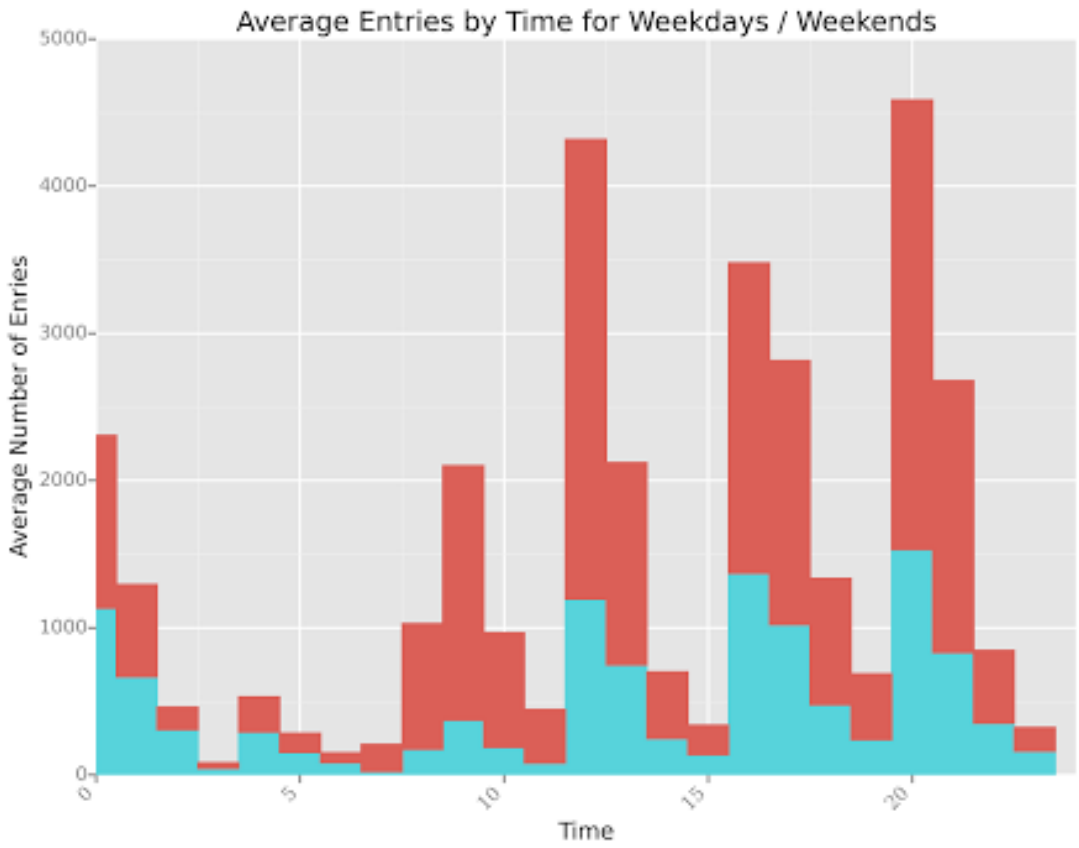
Ridership by time-of-day or day-of-week

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

Answer:

Python Code:

```
from pandas import *

from ggplot import *

import datetime

import pandasql

def plot_weather_data(turnstile_weather):

    df = turnstile_weather.reindex(columns=['Hour','UNIT','ENTRIESn_hourly','DATEn'])

    df['DATEn'] = df['DATEn'].map(lambda x: datetime.datetime.strptime(x, '%Y-%m-%d').strftime('%w'))

    df['DATEn']= df['DATEn'].replace(to_replace=['0', '6'],value='Weekend')

    df['DATEn']= df['DATEn'].replace(to_replace=['1', '2','3','4','5'],value='Weekday')

    # Query to retrieve the data to base the graph on

    q = """

    select DATEn,Hour,  avg(ENTRIESn_hourly) entriesn_hourly

    from df

    group by DATEn,Hour

    """

    df = pandasql.sqldf(q.lower(), locals())

    plot = ggplot(df, aes('Hour','entriesn_hourly',color='DATEn',fill='DATEn',label='DATEn',legend=True)) +
geom_bar(aes(weight = 'entriesn_hourly'),stat='bar') + \

    ggtitle('Average Entries by Time for Weekdays / Weekends') + xlab('Time')  + ylab('Average Number of
Enries') + xlim(0,24) + \

    theme(axis_text_x=element_text(angle=45,  hjust=1,  family="serif",   vjust=1))

    return plot
```



Legend:

Weekend

Weekday

Comments:

This is a Bar Plot of Average number of Passengers entering the subway at different times of the day. It has been segregated into 2 sections. The lower section in color blue represents the passengers entering the subway during weekend. The upper section in color red represents the passengers entering the subway during weekday. We can observe that people entering the subway during most times of the day is higher on weekdays than on weekends. Hence, we can conclude that the subway is busier during the weekdays than on weekends. Also, we can see that on weekdays it is busier at around 9:00 am, around 12:00 pm, between 4:00 and 5:00 pm and between 8:00 and 9:00 pm than the other times of the day

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:

From the Mann-Whitney U-Test it can be seen that there is a difference between the ridership when it is raining versus the ridership when it is not raining and the mean of the ridership when it is raining is greater than the mean of the ridership when it is not raining. Hence, it seems that more people ride the NYC subway when it is raining versus when it is not raining

Also, in the Histogram plotted it can be seen that the ridership on the rainy days is greater than the non-rainy days, though as we move away from zero this difference gets lesser

The co-relation of the ridership to rain is weak as it contributes very less to the R^2 value when taken as one of the input features in the regression model to predict ridership, hence though it seems that more people are riding the subway when it is raining as compared to when it is not raining it is not as significant a factor contributing to the overall ridership as compared to the other factors like UNIT, Hour, meantempi, meanwindspdi

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:

As described in section 4.1, the Mann-Whitney U-Test, the Histogram plotted for ridership on rainy days v/s non-rainy days lead to the conclusion that the ridership on the rainy days is more than the non-rainy days.

The effect of rain on the overall ridership is very less as compared to the other factors which influence the ridership of the NYC subway. This is because; it contributes very less to the R^2 value when chosen as an input feature to the linear regression model for predicting the ridership of the NYC subway

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,

Linear regression model,

Statistical test

Answer:

Most of the Predictor variables used in the model were binary in nature. i.e. it could take only 2 values hence the response variable would be scattered around these 2 values and the probability of response variable getting scattered around 0 or 1 is 50% for these predictor variables of binary nature

One of the Predictor variable used in the Model is the dummy variable for UNIT and the Ridership is very highly co-related to this variable and most of the variance in response variable is explained by this variable alone

The Linear Regression Model has the following limitations:

a) It only looks for the Linear relationships between the Predictor and the response variables and hence is not a good fit if the relationship of the Predictor variable is nonlinear with the response variable
b) It is very sensitive to outliers as it only looks at the mean of the dependent variable and the independent variable

Mann-Whitney U-Test has the following limitations:

a) Even if we don't find a significant difference between the samples still we cannot say that the samples are from the same population
b) This test compares the medians and therefore we cannot refer to the means in the conclusion

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

Answer:

The most significant predictor variable contributing to the Ridership is the UNIT. If we remove this feature and make the prediction for the response variable then the value of R^2 drops by almost 36% using the Gradient Descent and by almost 40% using the OLS algorithm which signifies that ridership is very highly co-related to UNIT as compared to the other features

From the Graph of ridership by time of the day for weekday / weekends we see that the ridership is more during the weekdays as compared to weekends. Also, we can see that on weekdays it is busier at around 9:00 am, around 12:00 pm, between 4:00 and 5:00 pm and between 8:00 and 9:00 pm than the other times of the day