Analyzing the NYC Subway Dataset

Muhammad Khan.

Section 1. Statistical Test

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

The Mann-Whitney U test was used to find if there was a statistically significant difference between the number of reported entries on rainy and non-rainy days. It is a nonparametric test.

1.2 Did you use a one-tail or a two-tail P value?

A two-tail p-value was used. Simply whether or not there is a statistically significant difference

between the data set.

1.3What is the null hypothesis?

The Null hypothesis is two population are same, but there is a variation in the rain.

Null Hypothesis

H_0: M_1=M_2 or H_0: $M_1-M_2=0$

Alternate Hypothesis

H_1: $M_1 \neq M_2$ or H_1: $M_1-M_2 \neq 0$

--1.4 What is your p-critical value?

```
P critical value of .05
```

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.

The rain or no-rain histograms are not normally-distributed. As such, a non-parametric test Mann-Whitney U is a good fit.

Resubmit the Answer

The data I have include has two populations: ridership on Rainy days and ridership on Non-rainy days. I consider Welch's t-Test, Mann-Whitney Test to check the null hypothesis that the mean of two populations is the same against an alternative hypothesis.
The Welch's t-Test should meet the following assumptions: Both samples are drawn from normal population.
Therefore, I analyzed if the data I have for analysis were normally distributed. The histograms for the number of entries per hour for days on Rainy days and Non-rainy days showed they were not normal distribution.

I chose Mann-Whitney U-Test, it can be used for data with both normal and non-normal distribution.

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.

The Mann-Whitney U-Test tests the null hypothesis that the two samples being compared are derived from the same population. This null hypothesis allows us to test whether there is a statistically significant difference in ridership on rainy and non-rainy days.

```
Here's the correct output:

(Rainy=1105.4463767458733, Non_Rainy=1090.278780151855, U=1924409167.0, P= 0.0499998

)
```

1.4 What is the significance and interpretation of these results?

Resubmit the Answer

In statistics, the *p*-value **is a function of the observed sample results (a statistic) that is used for testing a statistical hypothesis. Before performing the test a threshold value is chosen, called the significance level of the test, traditionally 5% or 1% [1] and denoted as *α*. If the *p*-value is equal to or smaller than the significance level (*α*), it suggests that the observed data are inconsistent with the assumption that the null hypothesis is true, and thus that hypothesis must be rejected.**

Significance level a,0.05

P*2 =0.049998

The numerical results of the Mann-Whitney U-Test show us that p_ value is less than 0.05 this small p-value, we reject the null hypothesis of the Mann-Whitney U-Test.The distribution of the number of entries is statistically different between rainy and non-rainy days in the data set.

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 (GD) and OLS models where used to run linear regression on the NYC subway dataset.

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

In the models the features used were: Rain, Precipi, Hour, meantempi and dummy variables UNIT.

--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.

Resubmit the Answer

These input variables have strong significance in the data set. It strongly impact on the ridership.

Meantempi select as a feature that is a key part of the weather that affects people’s decision making.

Hour features were taken as it is easily observed how ridership changes with time of day and day of week. The train schedule varies between weekdays and weekends.

Precipi and rain, both significantly impact ridership of subway.

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

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

```
Your calculated R^2 value is: 0.318137233709
```

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?

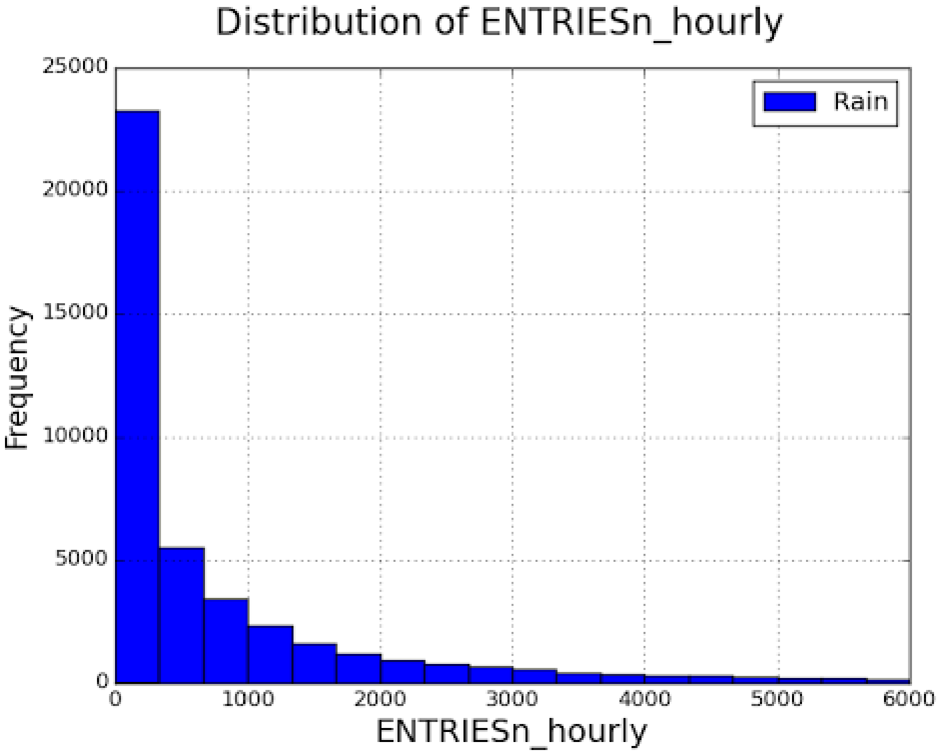
The R squared for the OLS is 0.318 which means we can explain about 32% of the data variability with the model. In other words, our model lets us predict NYC subway entries with 32% accuracy.

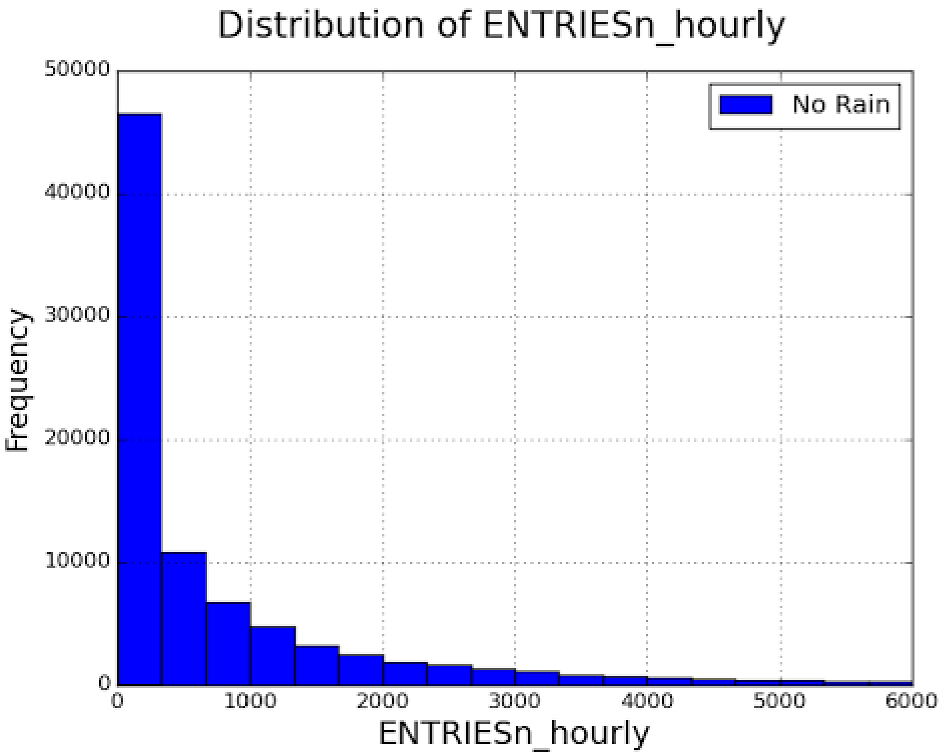
Resubmit the Answer

R squared ranges from 0 to 1. R^2 is essentially the percentage of variance that is explained, and is a quantitative measure of the “goodness of fit.” While it only explains 0.318 of variation.
The R2 value which is 0.318, which show a correlation for the ridership in NYC subway. In this test, the R2 value is 0.318, which suggests a relative low correlation between "Hour" and ridership. Based on this R2 value, this linear model is not very good model for this dataset.

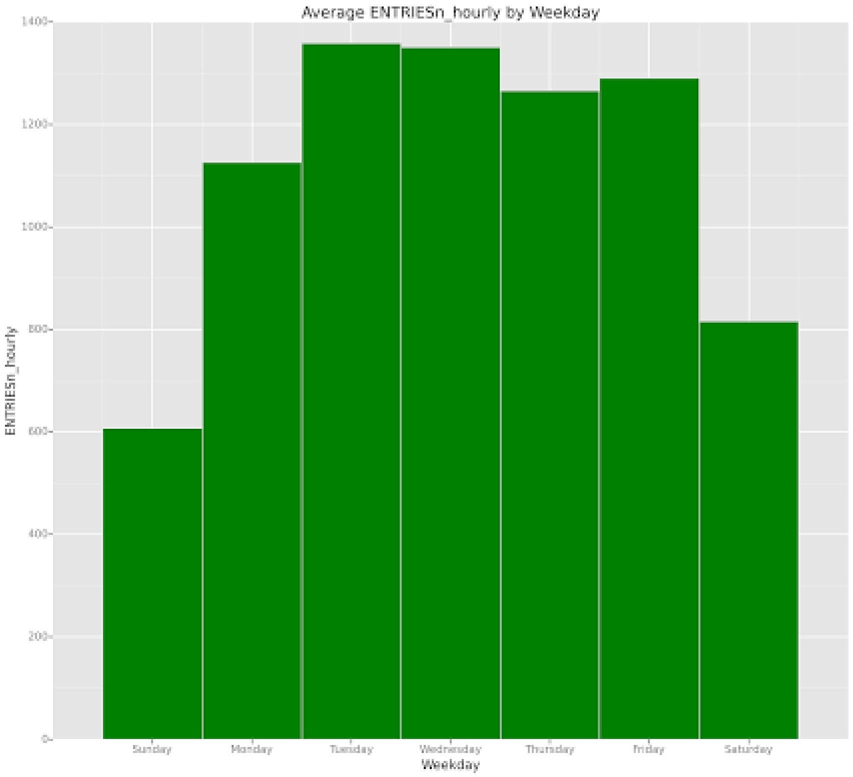
Section 3. Visualization

Resubmit the Answer





Distribution of ENTRIESn_hourly
No Rain
Frequency
50000
40000
30000
20000
10000
0
0
1000
2000
3000
4000
5000
6000
ENTRIESn_hourly



Average ENTRIESn_hourly by Weekday
ENTRIESn_hourly
1400
1200
1000
800
600
400
200
0
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Weekday

The bar chart shows that the average hourly ridership is greater on weekdays than weekends, Saturday showing importantly rising ridership than Sunday. It appears that the average hourly ridership on Monday is significantly different than the rest of the weekdays.

Section 4. Conclusion

1. & 2. 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? What analyses lead you to this conclusion?

The regression result does not show 100% if rain and ridership have a linear relationship. The model that we have created in the data set, with a coefficient of determination of 35%, sound rather deficient to prove that weather has any correlation with ridership. We can watch that there is a slight correlation in the first graph.
If we used the Mann-Whitney test, we watch that there's an important difference in ridership between raining versus not raining hours with only a residual chance of a sampling error. It does not prove that rain causes more entries, it proves that on rainy hours, there is a greater chance of entries.

Resubmit the Answer

The ridership on Rainy days and Non-rainy days are not the same at Significance level alpha , 0.05.The mean of "ENTRIESn_hourly" on Rainy days is =1105.45 which is slightly higher than the mean of "ENTRIESn_hourly" on Non-rainy days is=1090.28.

The Mann-Whitney U-Test results showed a value of 0.049, which is less than 0.05. Thus that the null hypothesis is rejected and the ridership on Rainy days and Non-rainy days are significantly different.

If we see mean of "ENTRIESn_hourly" on Rainy days was greater than the mean of "ENTRIESn_hourly" on non-rainy days, the results supported that there were more people
Ride the NYC subway when it was raining versus when it was not raining in a given month.

Section 5. Reflection

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

The Statistical Analysis of rainy vs non rainy condition that impact on the public transportation system.
We used linear model for the data set we have to predict data base on R square value.
Increasing sample data, would change analysis and trend.

Resubmit the Answer

1- Sample size- data set is limited, if we increase the data set size, it may impact on the result of analysis.

I focus on Analysis of Rainy and Non Rainy days. Furthermore comparison required like fog vs non fog days. It may also impact ridership.
Data from the whole year should be included for analysis.

Sources

http://mathbits.com/MathBits/TISection/Statistics2/correlation.htm
http://www.six-sigma-material.com/Histograms.html