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.

1.3 What 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?

$P<0.05$

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.

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.

We find that the computed U value gives significant evidence that the two distributions differ ($p < 0.05$, two tailed), therefore there is a significant difference between the number of hourly entries on rainy hours and non-rainy hours.

```
Here's the correct output:

(Rainy=1105.4463767458733, Non_Rainy=1090.278780151855, U=1924409167.0, P= 0.02499991
2793489721)
```

1.4 What is the significance and interpretation of these results?

The numerical results of the Mann-Whitney U-Test show us that 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.
We can say that the probability of rejecting H0 when this hypothesis is true is smaller than the significance level (0.0245 < 0.05) hence we reject this hypothesis.

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.

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

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.

Section 3. Visualization

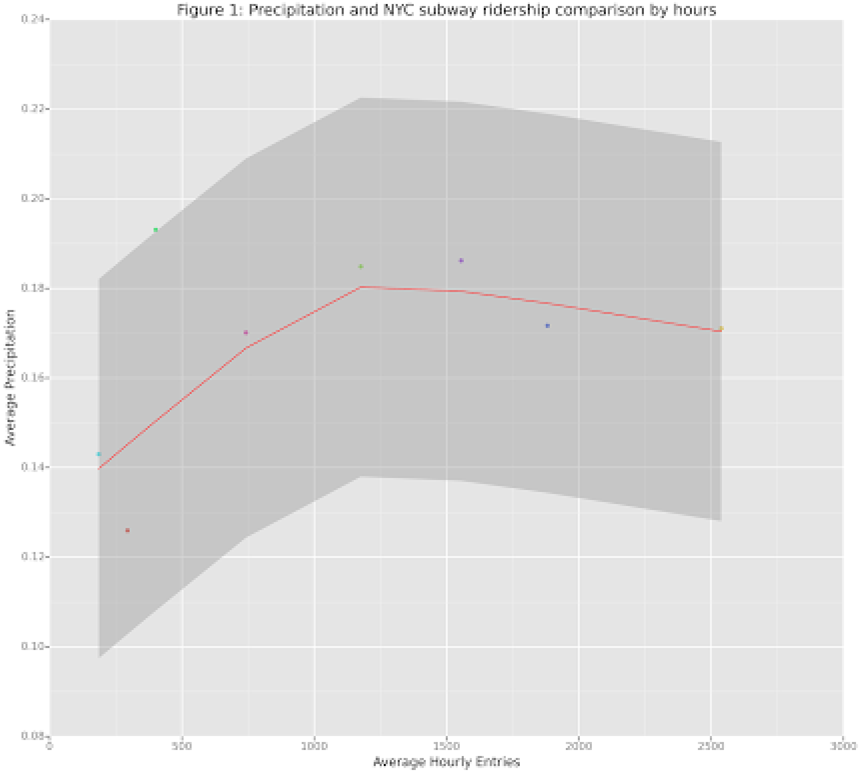


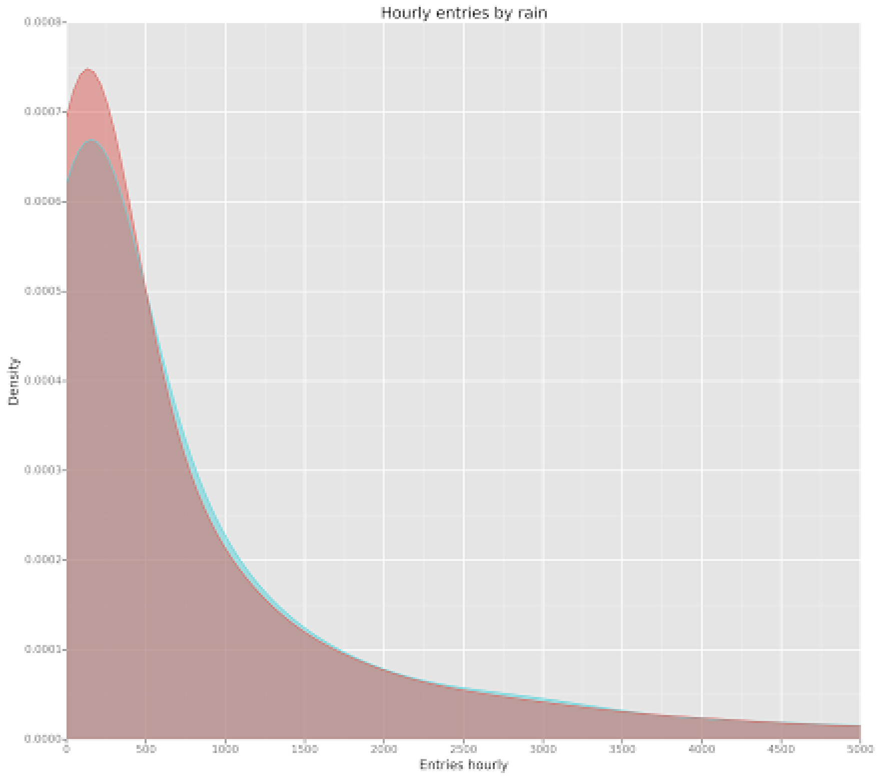
Figure 1: Precipitation and NYC subway ridership comparison by hours

This graph we are calculating ridership and precipitation by hours. There is a little correlation between precipitation and subway ridership.

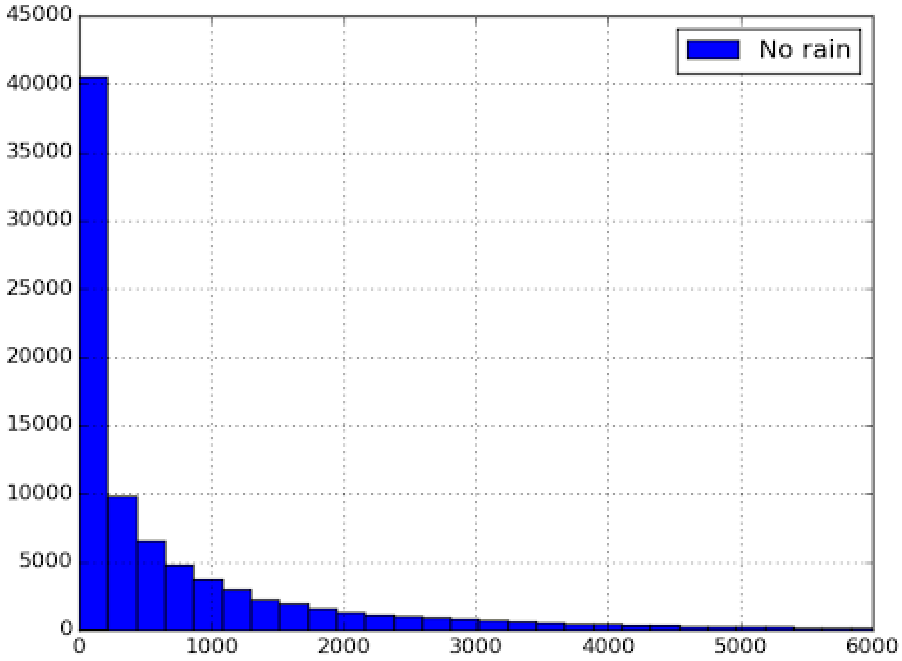
Rain

0

1



This graph is hourly entries with rain and hourly entries without rain because they have significantly different sample sizes to compare. This graph, we can observe that the probability of the ridership variable is smaller on rainy hours compared to non-rainy hours.



The histograms for the number of entries per hour for days on rainy days and non-rainy days showed they were not normal. I select Mann-Whitney U-Test, which can be used for data with both normal and non-normal distribution.

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.

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.

Sources

http://mathbits.com/MathBits/TISection/Statistics2/correlation.htm

http://www.six-sigma-material.com/Histograms.html