The purpose of this project is to explore whether weather conditions make a sizable impact on the ridership of NYC Subway System.

SECTION 1. STATISTICAL TEST

1.1 Which statistical test did you use to analyze 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?

Mann-Whitney U Test was used to determine whether group's medians are statistically significantly different. The null hypothesis is that both populations are equal, ea. there is no significant difference in populations' medians. The two-tail P value was selected in accordance with the null hypothesis. In this test we use the significance level of 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 choice of statistical test type is based on the characteristics of the data. The datasets for both rainy and non-rainy days are similarly, but not normally distributed. The data in both groups are independent of each other, and the responses are ordinal.

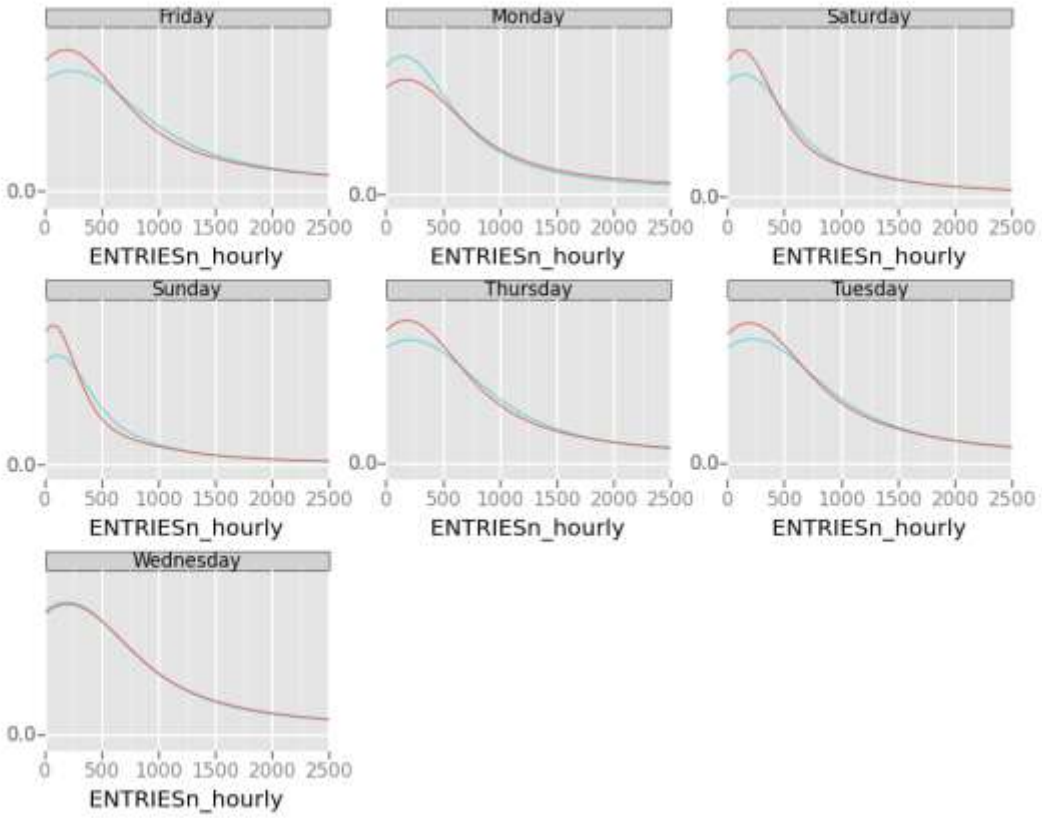


Figure 1 Entry per hour during rainy and non-rainy days

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.

Mean of rainy days	Mean of non-rainy days	U Value	P for two-tail test
1105.45	1090.28	1924409167.0	0.039

1.4 What is the significance and interpretation of these results?

There was a statistically significant difference between the two data sets. Because of that we have to reject the null hypothesis that there is no impact on the ridership.

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:

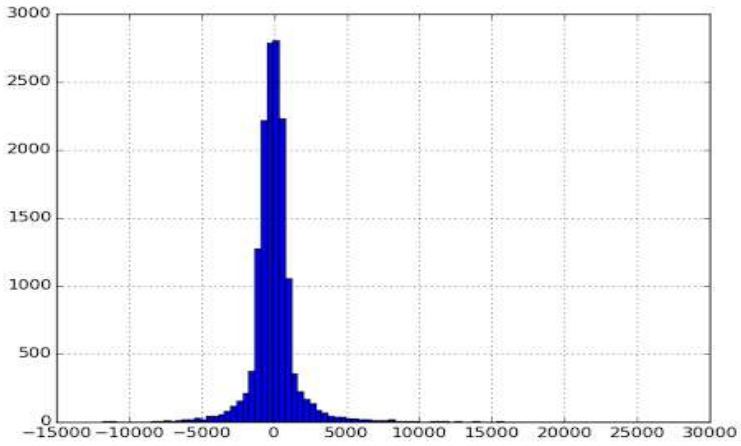
The method of Gradient descent was chosen to model our data.

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

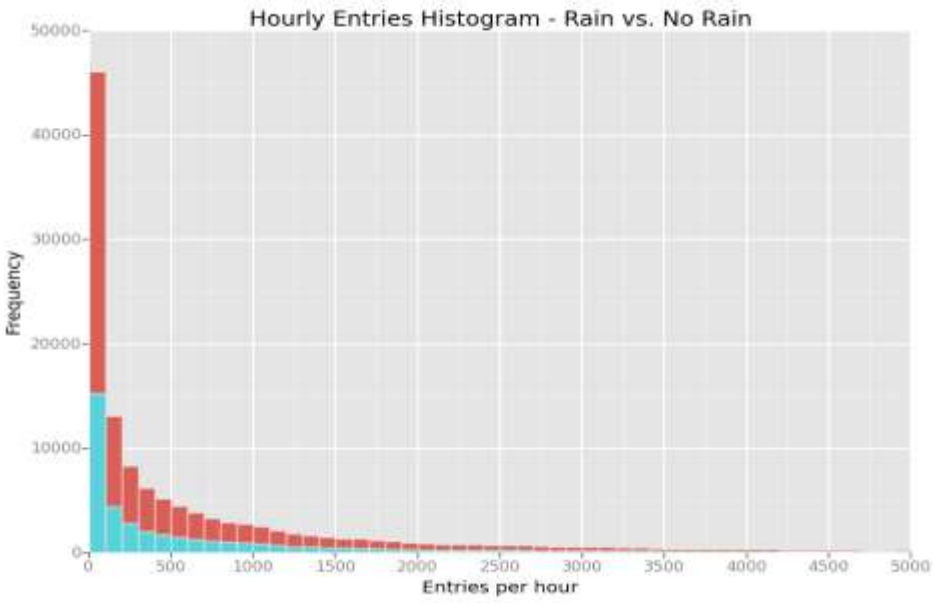
We selected to use *Hour, weekday, rain, Unit*, and *meantempi* variables. Our selection is based on everyday life observations. As everyone knows, there are different patterns of ridership on weekdays and weekends, the time of day makes a big difference too. Stations vary on volume of passengers as well.

2.5 What is your model's R2 (coefficients of determination) value? What does this R2 value mean for the goodness of fit for your regression model?

R^2 value is 0.465. This value is relatively low, that means that model does not predict well.



SECTION 3. VISUALIZATION



SECTION 4. CONCLUSION

The analysis of provided dataset shows that more people ride the NYC subway when it is raining.

SECTION 5. REFLECTION

A non-parametric statistical test between two samples gives us a good reason to believe that rain makes an impact on ridership of NYC subway.

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

https://statistics.laerd.com/premium-sample/mwut/mann-whitney-test-in-spss.php

https://github.com/glamp/ggplot-tutorial