Analyzing the NYC Subway Dataset
Short Questions
Overview

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?

I used Mann-Whitney U-Test to analyze the NYC subway data. It is appropriate to use 2-tailed P value because we have to compare subway ridership on rainy vs non-rainy days by comparing the means of 2 groups, hence the null hypothesis is "There is no statistical difference between the subway ridership on rainy day vs. non-rainy day". The p-critical value obtained was 0.02499.

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.

Mann-Whitney U-test is applicable to the dataset because the both samples are not normally distributed and this test can be used in case both distributions are same or different in shape whether normal or non normal.
This dataset also fits with the following man-whitney u-test assumptions.

1. We have 1 dependent variable i.e. ridership
2. Independent variable can be grouped into 2 or more categories – in this case rainy column can be grouped into rainy or non-rainy

3. We have independence of observations for rainy vs. non-rainy days i.e. they don't overlap.

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.

Following results were obtained:

Rainy mean: 1105.4463767458733
Non-rainy mean: 1090.278780151855
U: 1924409167.0
P value: 0.024999912793489721

1.4 What is the significance and interpretation of these results?

At significance level of 5%, the null hypothesis stated above can be rejected because the p-value obtained (0.0249) is less than 0.05. Hence alternative hypothesis is accepted which is "There is a statistically significant difference between ridership on NYC subway on rainy vs. non-rainy day".

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:

1. **Gradient descent (as implemented in exercise 3.5)**
2. **OLS using Statsmodels**
3. **Or something different?**

Used Gradient descent.

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

Used following features in the model:

'rain', 'precipi', 'Hour',
'meantempi','fog','meandewpti','meanpressurei','meanwindspdi','mintempi','maxtempi','mindewpti','maxdewpti','minpressurei','maxpressurei'

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.

- **Your reasons might be based on intuition. For example, response for fog might be: "I decided to use fog because I thought that when it is very foggy outside people might decide to use the subway more often."**
- **Your reasons might also be based on data exploration and experimentation, for example: "I used feature X because as soon as I included it in my model, it drastically improved my R2 value."**

Chose some of these by intuition. NYC subway is used by most people for commuting to work hence feature 'Hour' is important. Other features are chosen by trial and error, by seeing the effect of including them on R2 value.

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

rain	-46.58395441
precipi	-11.99832405
Hour	464.7142207
meantempi	-51.5739601
fog	24.66848212
meandewpti	24.42178947
meanpressurei	135.4017576
meanwindspdi	48.68748315

mintempi	-221.074481
maxtempi	105.5967244
mindewpti	-49.30858398
maxdewpti	176.9030947
minpressurei	-50.23206768
maxpressurei	-124.5605923

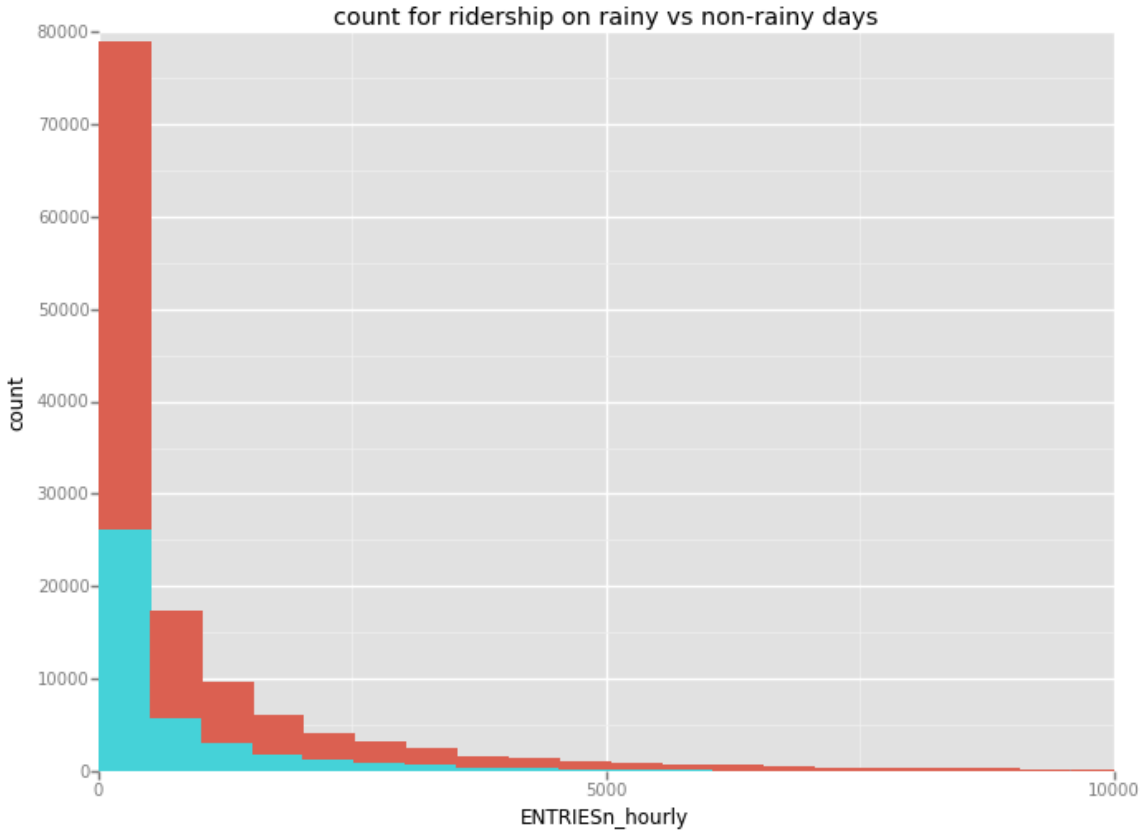
2.5 What is your model's R2 (coefficients of determination) value?

0.460896712543478

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?

Closer the R2 value to 1, better is the goodness of fit. In this case value of 0.46 means 46% of the variation in value of dependent variable i.e. ridership can be attributed to the features used in the model while remaining 54% is by chance.

3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for nonrainy days.

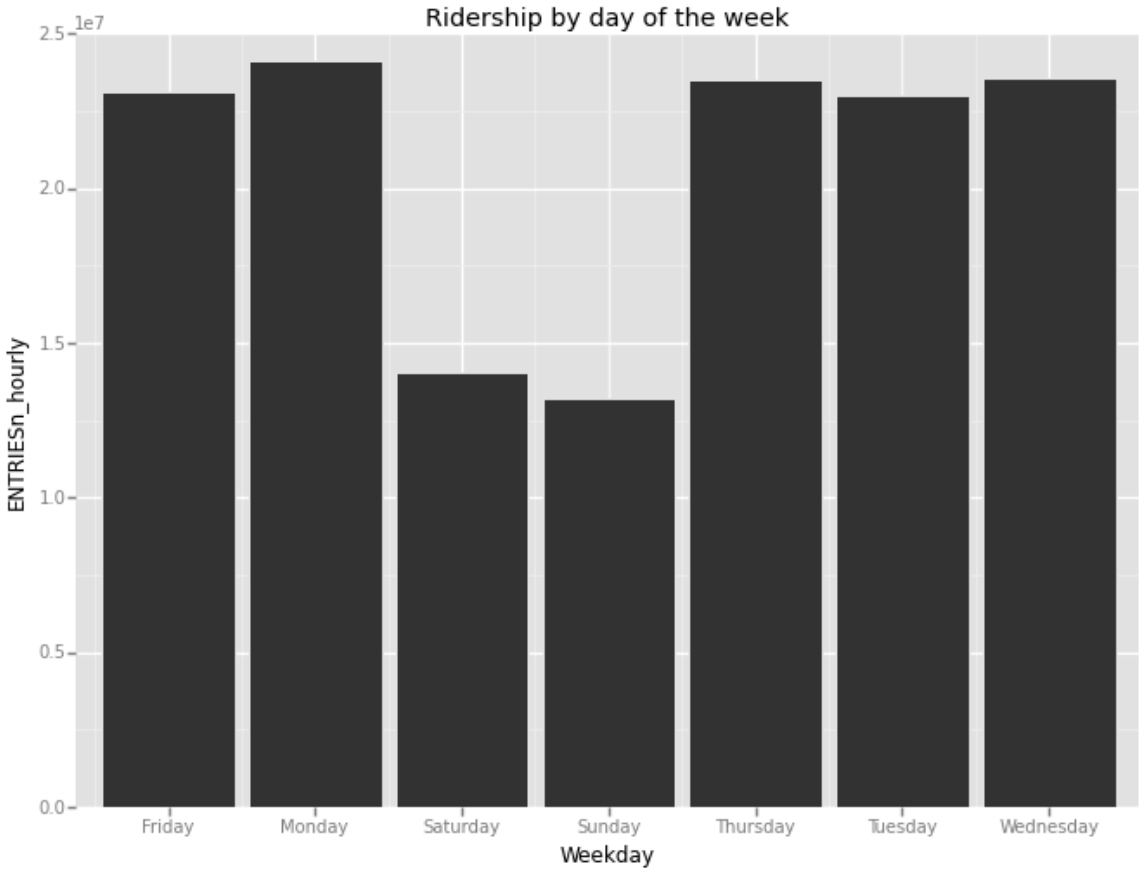


couldn't get the legend working , here blue is rainy and red is non-rainy

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



4.1 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?

From the tests and visualizations it does look like more people ride subway on non-rainy day than rainy day.

4.2 What analyses lead you to this conclusion?
Mann-Whitney U test showed that the two samples are different statistically as alternate hypothesis was accepted. But it doesn't say which group is larger than the other. Linear regression model does offer some insight on the effect of parameter 'rain' on the ridership. The coefficient for rain

comes to -46, which says it has negative effect on the ridership which means rainy day would reduce the ridership. Hence we can conclude that non-rainy days see more ridership than rainy days. But this predictor variable is not most the significant as there are other parameters, which show greater effect on the ridership, based on the coefficient values.
Lastly, from the visualization, the histogram showing rainy vs non–rainy ridership shows more values of Entries on non-rainy days vs rainy days.

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

There are some shortcomings of the dataset used to analyze the ridership.

1. There are more observations on non-rainy days vs rainy days, which creates bias.
2. Some non weather related variables might affect the ridership on some stations e.g. closure of station due to construction activity will reduce the activity, stations close to downtown may not have too much ridership on the weekend.
3. Data could be missing for some stations.
4. Data is not complete for each unit, some Units have captured data only for certain hour of the day.

There are also some shortcomings related to the analysis.

1. In regression model, although rain is considered Independent variable, other variables are actually related to rain e.g. temperature, dewpoint etc. are related to rain so they are not completely independent as is the assumption in regression.

2. The goodness of fit of regression model is 0.46, which is far from perfect value of 1, hence only 46% of the value of ridership can be predicted by independent variables.