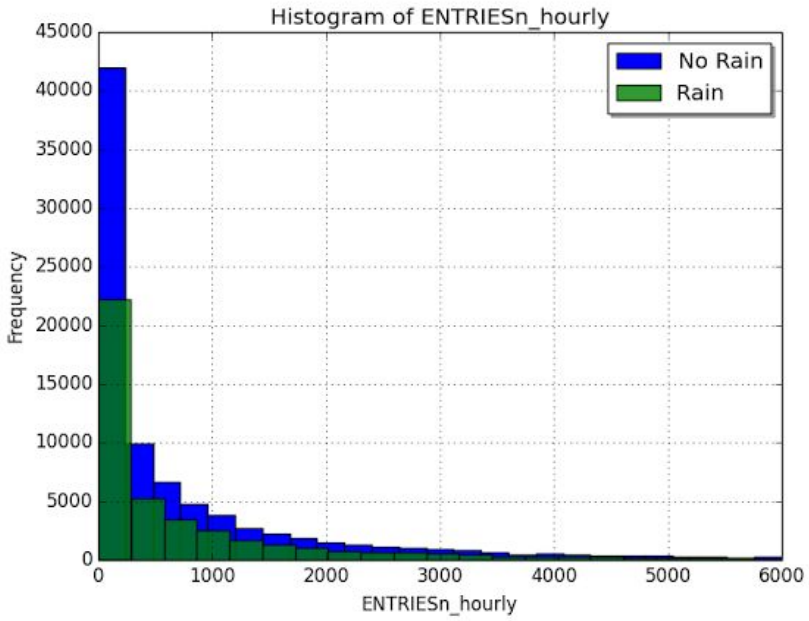
Analyzing the NYC Subway Dataset

What effect does weather have on ridership volume on the NYC subway systems for rainy vs non-rainy days?

Section 1. Statistical Test

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

Before I performed any analysis, I decided to take a look at the NYC subway data. More specifically, I examined the hourly entries in the dataset to determine what distribution the data follows for both "Rain" and "No Rain" variables. This can be achieved with a histogram plot in Python using the pandas and matlibplot libraries. From the figure below, the distribution clearly follows a non-symmetric shape. So, I began my analysis of the NYC subway dataset by performing the Mann-Whitney U-Test since it is a good starting point for non-normal distributions.



Did you use a one-tail or a two-tail P value? What is your p-critical value?

This uses a two-tailed test with a p-critical value of 0.05.

What is the null hypothesis?

I am interested in analyzing whether the null hypothesis should be rejected. The null hypothesis is the mean ridership on rainy days is the same as the mean ridership on non-rainy days.

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 Mann-Whitney U-Test is applicable to the dataset because the histogram plot follows a non-symmetric distribution. Through additional research, I learned that the Mann-Whitney U-Test is frequently used to compare behaviors in people. The objective is to understand what effect does weather have on human behavior for rainy and non-rainy days related to the volume of riders on the NYC subway based on real world data.

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 reported p-value is 0.049999824 and the mean of entries with rain is 1105.45 riders and without rain is 1090.28 riders.

1.4 What is the significance and interpretation of these results?

The outcome of the Mann-Whitney U-test for the “Rain” dataset and “No Rain” dataset has a reported p-value of 0.049999824. With a significance level of 95%, since p is less than 0.05, there is a difference between the two cohorts. Furthermore, the mean of entries with rain (1105 riders) is greater than the mean of entries without rain (1090 riders). As a result, more people ride the subway when it is raining as oppose to when it is not raining. The null hypothesis was proven false. In summary, the distribution of the number of subway entries is statistically different between rainy and 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?

I decided to use Linear Regression with 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?

I used the following features in the pandas dataframe:

- Hour
- Fog
- Rain

My model used 465 dummy variables. Each dummy variable represented a specific turnstile location labeled “UNIT”. Since the dummy variables are categorical in nature, there is not an explicit direction or ordering of values.

2.3 Why did you select these features in your model?

After looking at the raw data, I noticed that all dates were from May. Since the weather in NYC during the month of May is one of the best times of the year, I decided to use input variables that would cause poor weather conditions. For example, I used features Fog and Rain. I also used Hour because I wanted to add structure to the model for when people decide to ride the subway.

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

After 15 iterations with a 0.5 alpha learning rate, my final set of theta updates converged to 468.2 weight for Hour, 49.3 for Fog, and -1.28 for Rain. The strongest non-dummy feature of my model is Hour. I am a bit surprised that the Rain input variable was in the noise, essentially having zero effect on the predicted output.

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

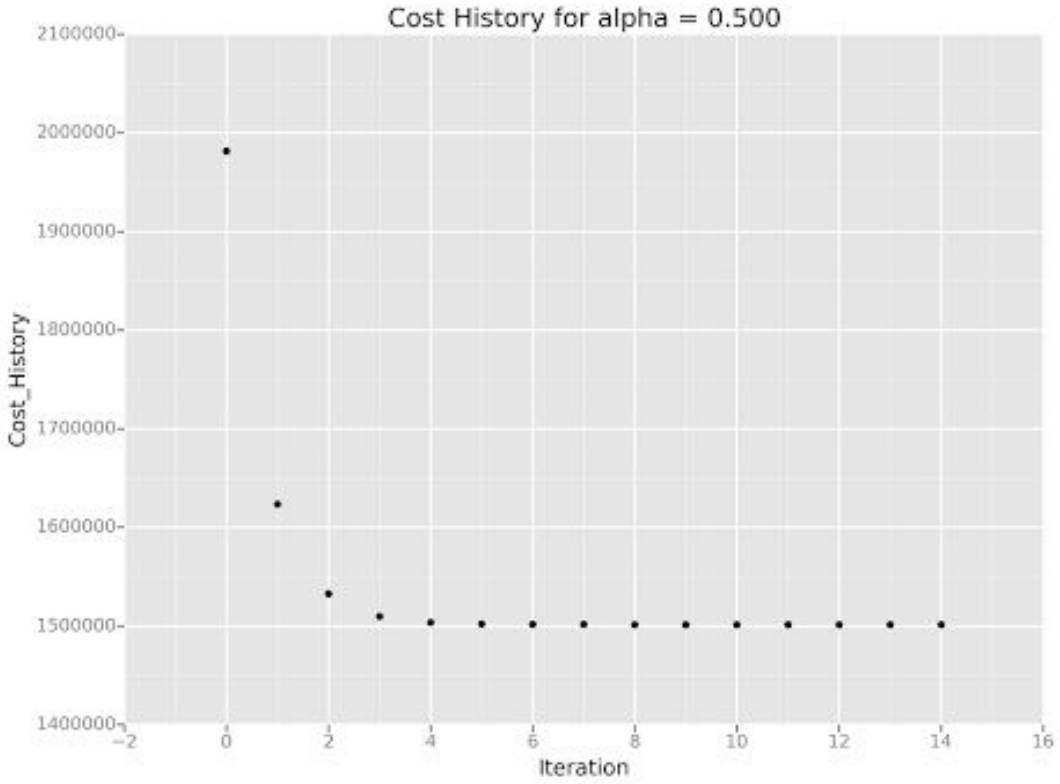
My coefficient of determination is 0.46.

2.6 What does this R2 value mean for the goodness of fit for your regression model?

The R2 value is an evaluation of the model’s performance that ranges from 0 (poor) to 1 (better). Based on my R2 value, I think the data does not perfectly fit-in with my linear model. It is weaker than I expected but fair enough for this project. However, it was better than the targeted 0.2 threshold, which was a requirement for problem set 3.5.

Do you think this linear model to predict ridership is appropriate for this dataset, given this R2 value?

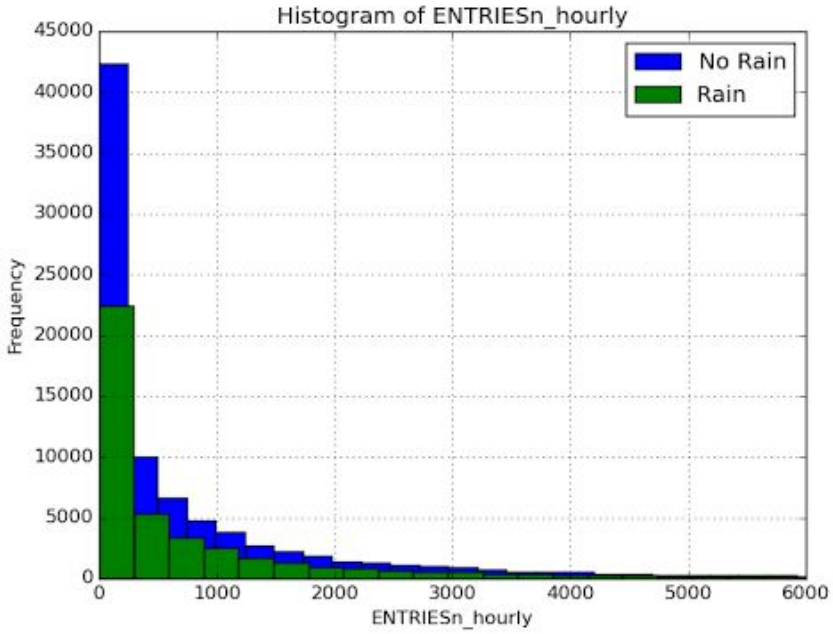
The perfect model would show the cost history converging to a target value of 0. As you can see in the figure below, my cost function decreases by a fair amount from 2M to 1.5M, ultimately converging to the minimum cost. It decreased as we trained the model with the selected features, but it doesn't do super well. However, I believe It has enough data points to make a solid conclusion.



Section 3. Visualization

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

From the figure (histogram plot) below, both distributions clearly follow a non-symmetric shape. The “No Rain” distribution is larger because there were more days (larger sample size) without rain in the dataset.



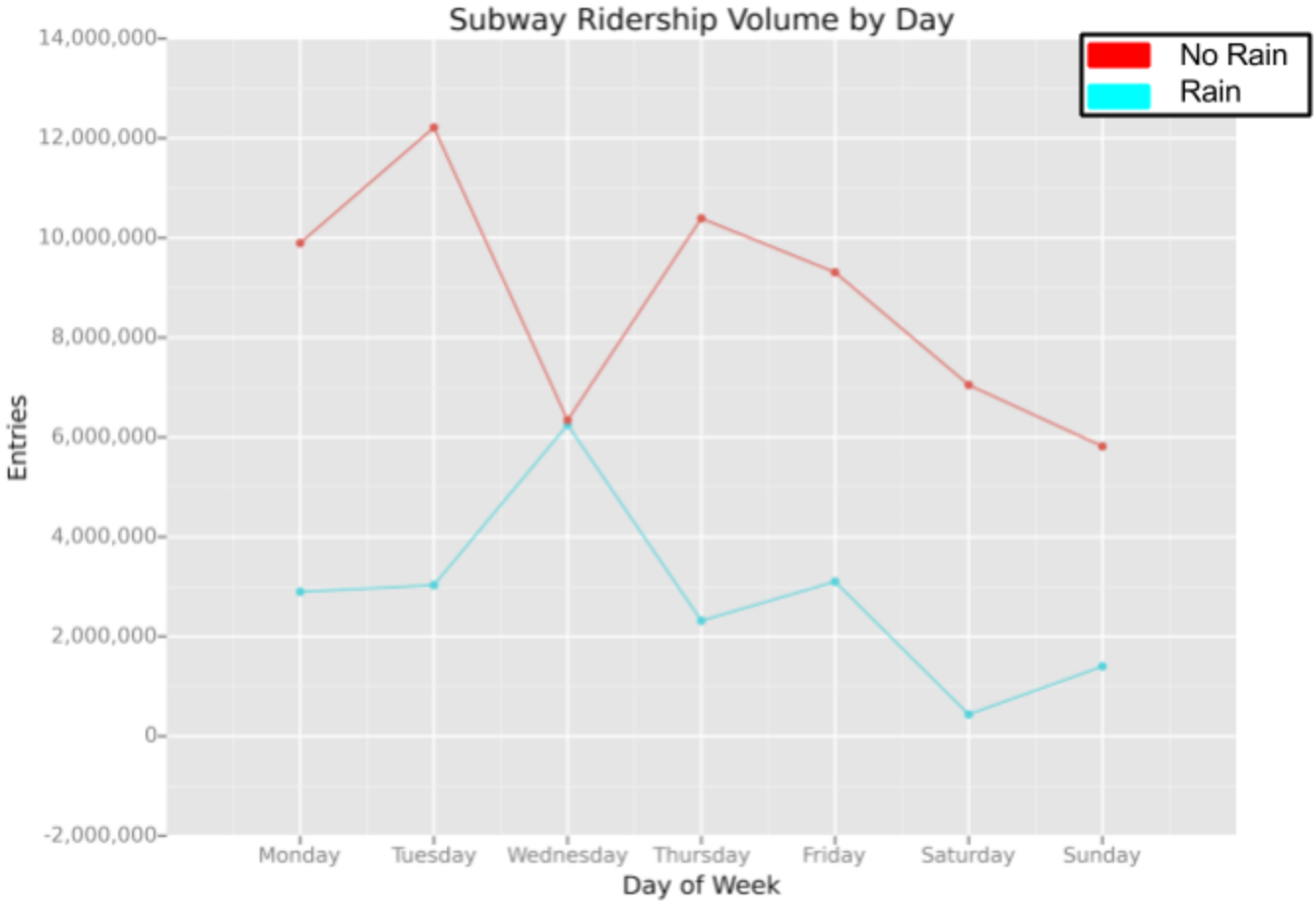
Below is a code snippet of my work to avoid any doubts that I copied/pasted the above figure from the instructor’s notes.

```
import numpy as np
import pandas
import matplotlib.pyplot as plt

def entries_histogram(turnstile_weather):
    '''
    You can read a bit about using matplotlib and pandas to plot histograms here:
    http://pandas.pydata.org/pandas-docs/stable/visualization.html#histograms

    You can see the information contained within the turnstile weather data here:
    https://www.dropbox.com/s/meyki2wl9xfa7yk/turnstile_data_master_with_weather.csv
    '''
    rain = turnstile_weather[turnstile_weather.rain == 1]
    norain = turnstile_weather[turnstile_weather.rain == 0]

    plt.figure()
    binwidth = 175
    norain['ENTRIESn_hourly'].hist(bins=binwidth, label="No Rain")
    rain['ENTRIESn_hourly'].hist(bins=binwidth, label="Rain")
    plt.title('Histogram of ENTRIESn_hourly')
    plt.xlabel('ENTRIESn_hourly')
    plt.ylabel('Frequency')
    plt.legend(loc='best')
    plt.xlim(0,6000)
    return plt
```

3.2 One visualization can be more freeform. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots) or attempt to implement something more advanced if you'd like.

Regardless of weather conditions (wet or dry), the ridership spikes occur during the same peak times (midnight, 9am, noon, 4pm, and 8pm). The valleys between the spikes are very comparable for both plots in terms of total number of entries.

This plot shows the relationship between ridership volume on rainy and non-rainy days for a given day of the week. Each day of the week had some rain. The takeaway from this plot is Wednesday experienced more rain than any other day of the week for the provided dataset. Apparently, Wednesday was bad luck that month. Tuesday was the busiest month in terms of total riders with or without rain.



Section 4. Conclusion

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?

Based on my analysis, more people ride the subway when it is raining as opposed to when it is not raining. I reached this conclusion by gleaning valuable information from the Mann-Whitney U-test. Since the majority of residents in NYC do not own a car, it makes a lot of sense that the weather outside tends to influence how people get around in a major city.

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.

The Mann-Whitney U-test results for the “Rain” dataset and “No Rain” dataset has a reported p-value of 0.049999824, which is right on the borderline of the p-critical value. With a significance level of 95%, since p is less than 0.05, there is a difference between the two populations. To strengthen my argument, supplemental statistics such as the mean is necessary to support the reported p-value result. The mean of entries with rain (1105 riders) exceeded the mean of entries without rain (1090 riders). It was a lot closer than I originally expected, but nonetheless a subtle difference as the mean for the ridership with rain is greater.

As I trained my linear regression model, I was able to reduce the cost function by a decent amount (2M to 1.5M) with an R2 value of 0.46. As a starting point, I built one model per non-dummy feature and ranked the non-dummy features according to coefficient weights. In addition, I decided to build the model gradually by starting with the strongest weight (Hour) and then added the next strongest weights step by step. It turns out that the predicted output from the Hour + Fog model was nearly identical to the predicted output from the Hour + Fog + Rain model. The rain feature essentially had no impact in my model. With an R2 value of 0.46, the model is middle of the road in terms of generalizing the data.

Section 5. Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis, including:

1. **Dataset**

   One potential shortcoming is there could be a disconnect between the model and the data. The project required the use of a linear regression model. What if the data is nonlinear? Trying to fit a linear model on nonlinear data sometimes causes under-fitting. Another potential issue is data corruption and one technique to resolve that is to clean the data.

2. **Analysis**

   One potential shortcoming of linear regression with gradient descent is the cost function could have multiple local minimas. I performed gradient descent just once and started with an initial theta value of 0. Perhaps I could experiment with different initial values of theta and perform gradient descent numerous times to see if I could improve the performance of the model. What if my analysis found the local minimum that isn't actually the global minimum of the cost function?

References:

All websites and materials provided in the classroom.