Analysing the NYC Subway Dataset (intro to Data Science final project)

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?

Shapiro-Wilk test was ised to find out if the data is normally distributed. And Mann Whitney test was used to test if ridership on rainy days is the same as 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 results of Shapiro-Wilk show that Mann-Whitney is applicable

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.

rain	**ENTRIESn_hourly (mean)**
0	**1845.539439**
1	**2028.196035**

Shapiro-Wilk results: (0.5943876504898071, 0.0)

Mann-Whitney results: (153635120.5, 2.74106957124e-06)

1.4 What is the significance and interpretation of these results?

Given the results we can reject the null-hypothesis and assume that ridership on rainy days is not the same as is on 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:

both (although the results were quite close to each other):
OLS using Statsmodels or Scikit Learn
Gradient descent using Scikit Learn

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

Here’s the features list I’ve used:
['rain', 'meanprecipi', 'hour', 'meantempi', 'meanwspdi', 'weekday', ‘day_week', 'pressurei', ‘weather_lat','tempi','precipi','wspdi']

Plus two dummy variables: UNIT, conds

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 R^2 value.”

Mostly used those features which significantly improved R^2 value.
Some features like longitude and meanpressure were eliminated as their impact on R^2 value was too litttle.

2.4 What are the parameters (also known as "coefficients" or "weights") of the non-dummy features in your linear regression model?
{**'rain': -57.118964122519799, 'meanprecipi': 145.08826083597626, 'hour': 737.64620742529576, 'meantempi': -409.8639355026196, 'meanwspdi': -96.308550734833773, 'weekday': 509.87956098422029, 'day_week': 96.749824633464428, 'pressurei': -84.432729768667599, 'weather_lat': 140.37641404504953, 'tempi': 340.2303882731585, 'precipi': -69.554341659073842, 'wspdi': 66.670690540261205}**

2.5 What is your model’s R^2 (coefficients of determination) value?

R_ols = 0.49289282801 #OLS Predictions
R_gd = 0.492794955901 #Gradient descent

2.6 What does this R^2 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 R^2 value?

R2 equal 0.49 was the best I could get from the data. But still this is not a good results considering that maximum is 1. So I cannot say that this is a strong model.

Section 3. Visualization

Please include two visualizations that show the relationships between two or more variables in the NYC subway data.

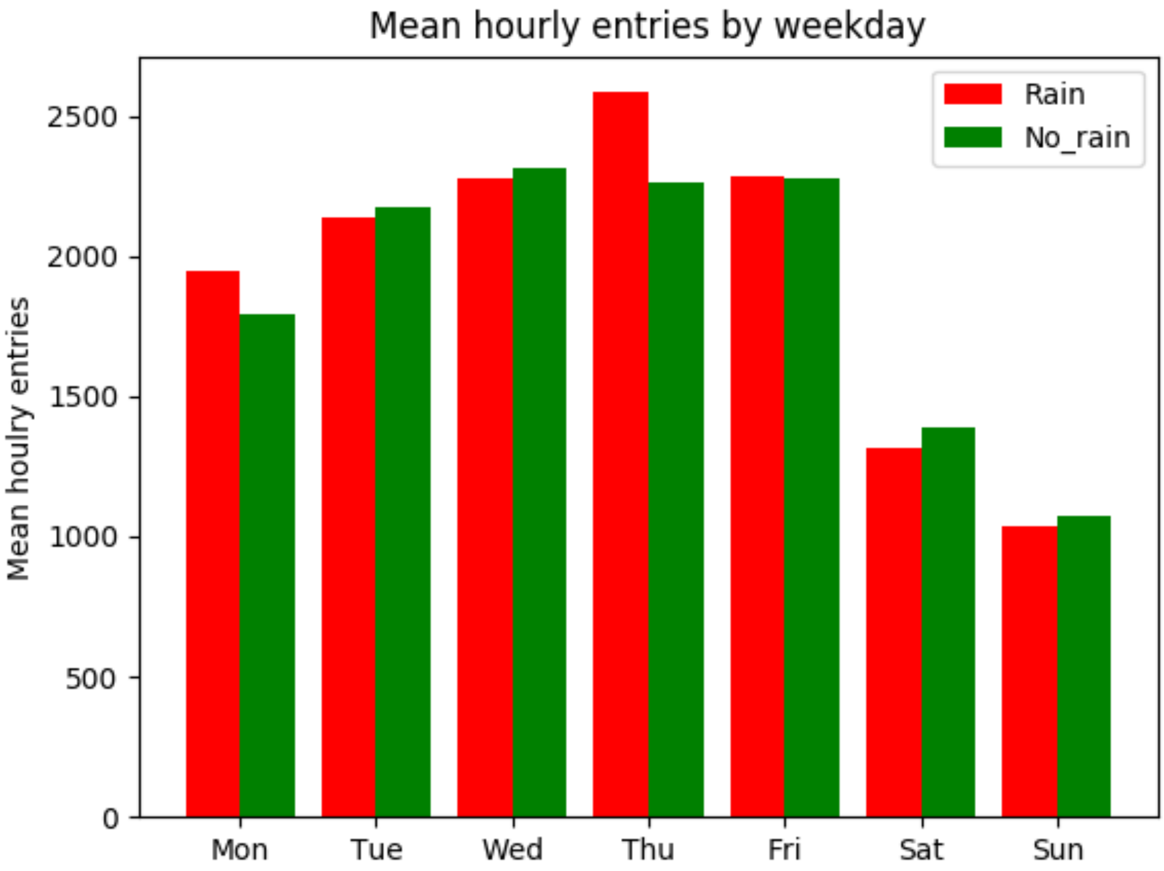
Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

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

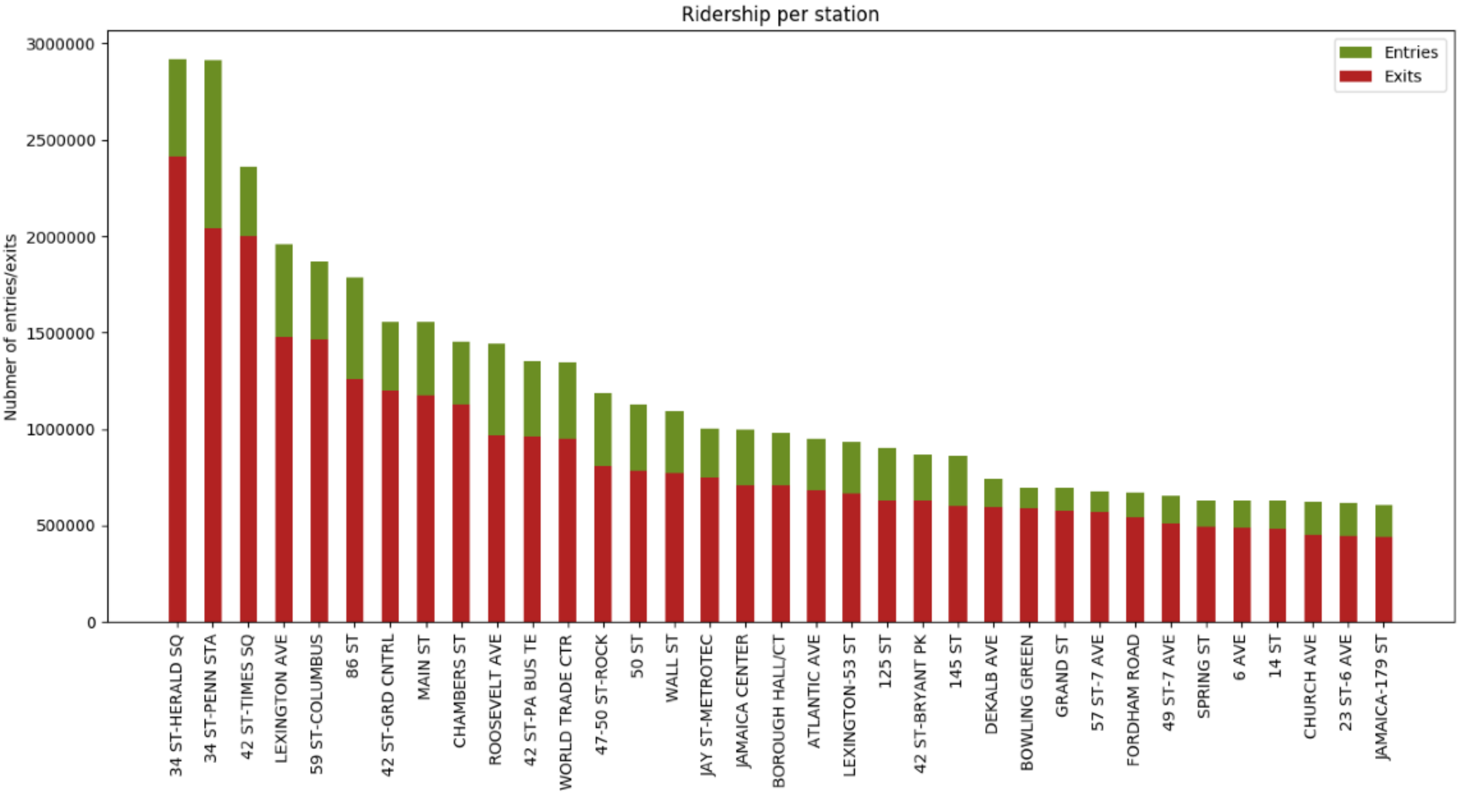
You can combine the two histograms in a single plot or you can use two separate plots.
If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.
For the histograms, you should have intervals representing the volume of ridership (value of ENTRIESn_hourly) on the x-axis and the frequency of occurrence on the y-axis. For example, each interval (along the x-axis), the height of the bar for this interval will represent the number of records (rows in our data) that have ENTRIESn_hourly that falls in this interval.
Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.
Not literally what was asked but this may be more informative. Instead ggplot matplotlib was used because of some problems with compatibility with latest version of python:

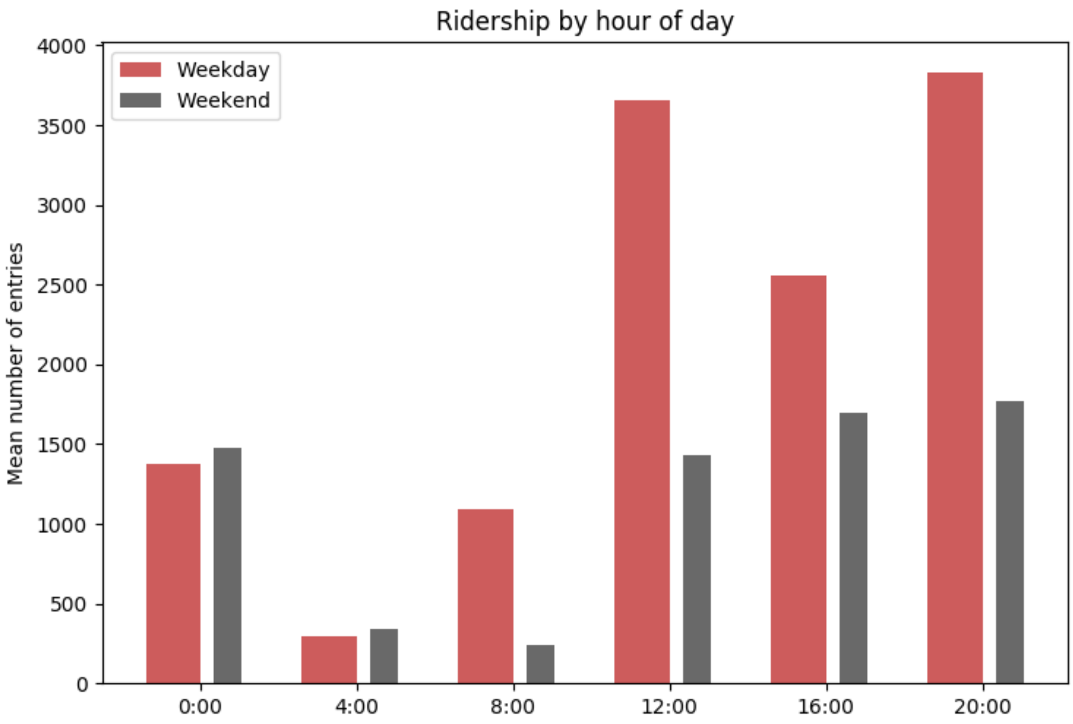
3.2 One visualization can be more freeform. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots) or attempt to implement something more advanced if you'd like. Some suggestions are:

Ridership by time-of-day
Ridership by day-of-week



Two more visualisations: 35 most popular stations (most entries and exits) and ridership by time of day





Section 4. Conclusion

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

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?

According to the data mean hourly ridership when it's raining is higher than when it's not (2028.196035 vs 1845.539439).

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.

Mann-Whitney test + means of 2 samples
Theta for rain was = -57.118964122519799

From visualization by day of week it's hard answer the question so we can just calculate mean riders in rainy and dry days.

Section 5. Reflection

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

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

Dataset,
Analysis, such as the linear regression model or statistical test.

I guess you cannot say that the dataset is big enough. There is not much data to really see correlations between riders and weather (esp fog or wind). Although it's good to see if ridership is connected to the day of week and the hour of the day. Plus we can see the busiest stations.

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