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?

Answer: A non-parametric test (Mann-Whiteny U test) was used because the distribution is not normal.
Null hypothesis: No significant difference in subway ridership between rainy and non-rainy days.
I used two-sided test with p critical value =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.

Answer: Because the distribution of ridership is not normally distributed, I need to use non-parametric test instead of t-test.

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.

Answer:
Mean of the samples with rain = 1105.44, mean of the samples without rain =1090.28. Two-sided p-value = 0.05.
Based on this p-value, I can reject the null hypothesis with alpha = 0.05

1.4 What is the significance and interpretation of these results?
Answer: This result means that the number of entries are significantly 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:

a. OLS using Statsmodels or Scikit Learn
b. Gradient descent using Scikit Learn

c. Or something different?

Answer: I used OLS using Statsmodels to compute the theta and produce prediction.

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

Answer: I used 'rain', 'precipi', 'hour', 'meantempi','day_week' as input variables. UNIT and conds were used as a dummy variable.

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.

Answer: I used rain/precipi because people may use subway more often during rainy days. I also used day_week, hour, UNIT because the number of people using subways fluctuates over time, day and locations. I also used conds (weather condition) as it may affect people's behavior

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

Answer:

rain	201.284320
precipi	776.804619
hour	122.889759
meantempi	-25.427899
day_week	-146.973990

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

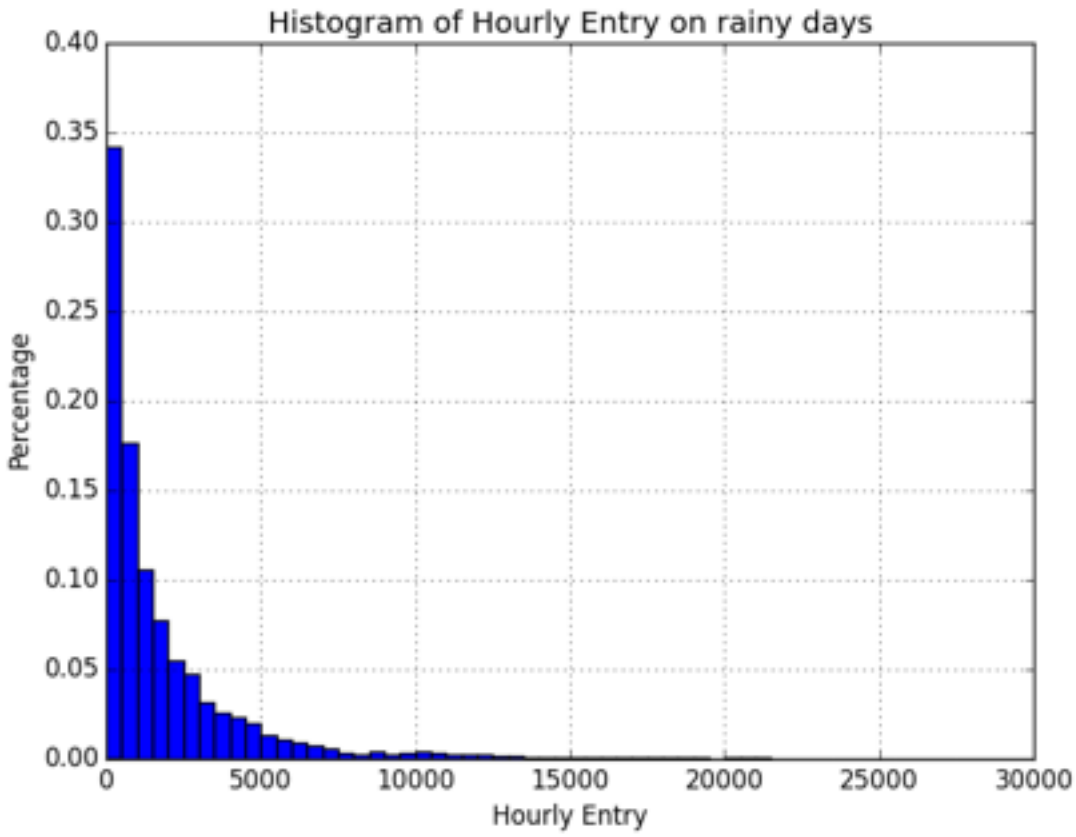
Answer: 0.477

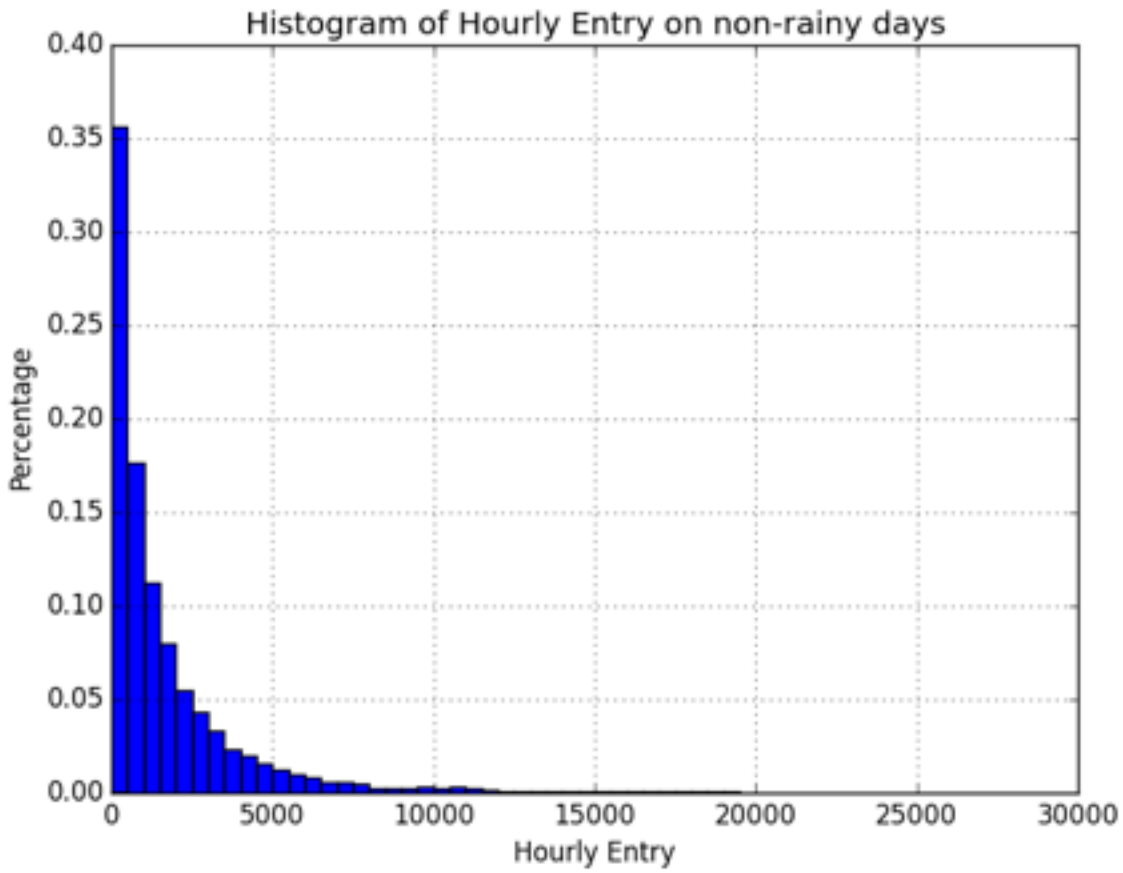
2.6 What does this R^2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?

Answer: R^2 indicates how close the data are to the regression line, and how much the model explains the variability of the data. Considering that this models predicts human behavior, we expect lower level of R_square in general, and 47.7% R^2 is considered to be high. Therefore, I think this linear model is appropriate for this dataset even though there should be better models that can have higher R_square. At the same time, R_square doesn't tell whether this model will predicts well for the unknown data set because we didn't try to divide samples and do a cross validation.

Section 3. Visualization

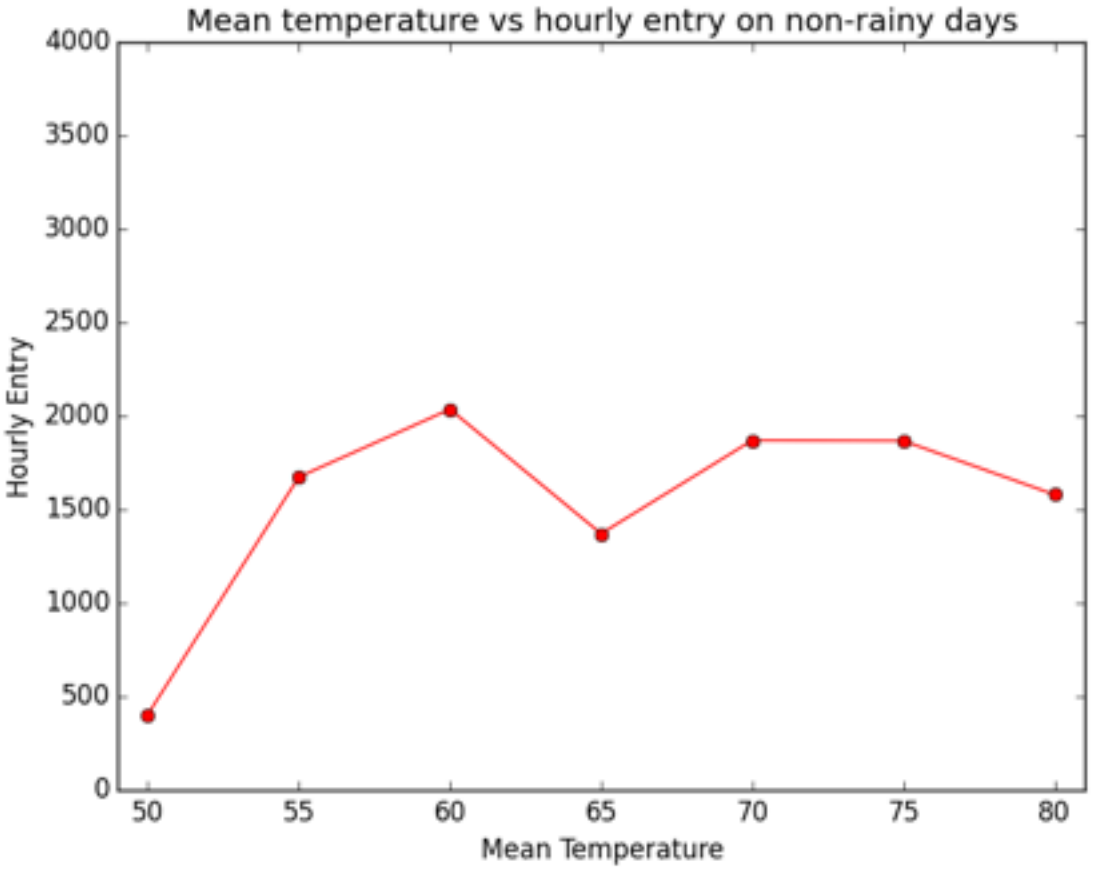
Please include two visualizations that show the relationships between two or more variables in the NYC subway data.
Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

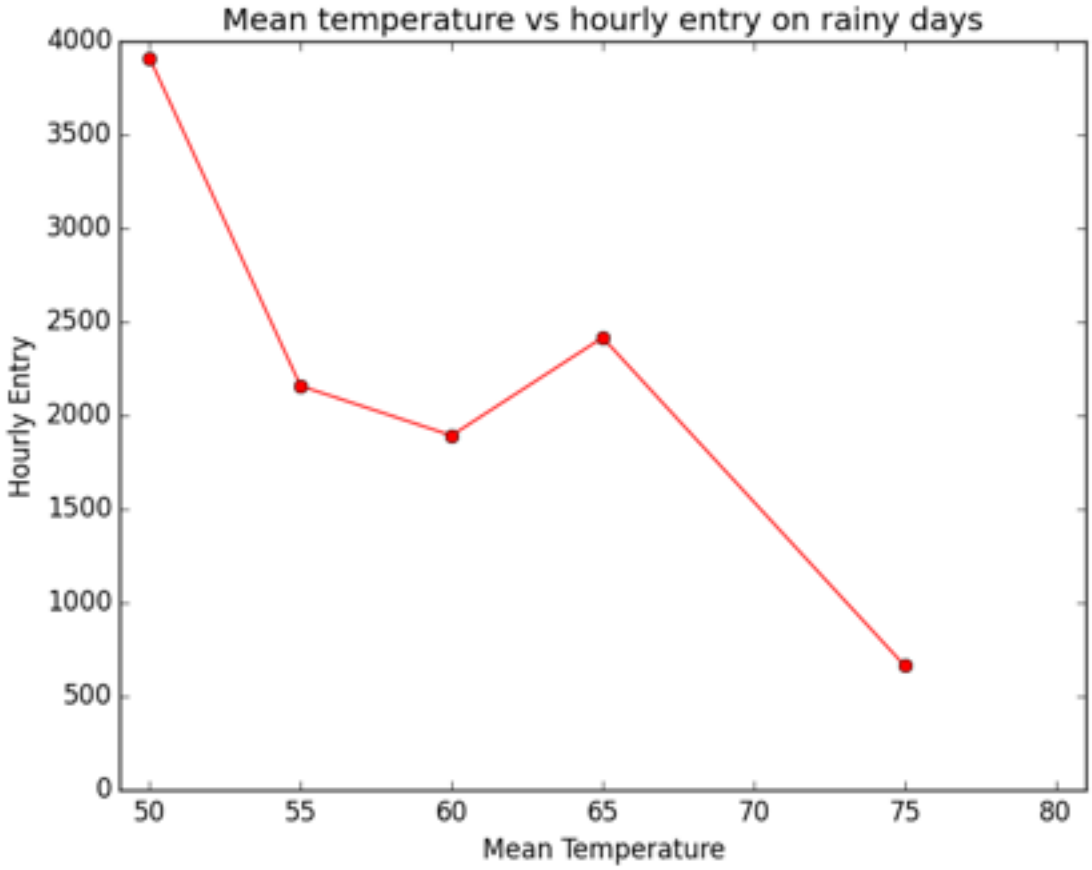




The above figures are the histogram of hourly entry on rainy and non-rainy days. Bin size of the plot is 500, and the histogram is normalized to show the percentage of each bin. From this figure we can observe the followings:

- Percentage of low hourly entries (less than 500) is higher on non-rainy days (0.36 vs 0.34). It means that people tend to ride less on non-rainy days.
- Both histograms show a skewed distribution to the right, but the histogram of non-rainy days has slightly more occurrences in the lower values.


Mean temperature vs hourly entry on non-rainy days
Hourly Entry
Mean Temperature
0
500
1000
1500
2000
2500
3000
3500
4000
50
55
60
65
70
75
80


Mean temperature vs hourly entry on rainy days
Hourly Entry
Mean Temperature
0
500
1000
1500
2000
2500
3000
3500
4000
50
55
60
65
70
75
80

These two figures are show the mean temperature vs mean of hourly entry on rainy days and non-rainy days. X-axis is a 5-degree interval temperature, and Y-axis is the mean of hourly entry of the samples in each temperature interval.

On non-rainy days, mean of hourly entry doesn't depend on mean temperature very much except for the very low mean temperature value. On rainy days, hourly entry is inversely related to the mean temperature, i.e., people ride more when the temperature is lower. From this plot, and the first histogram, we can say that that people tend to ride more when the weather is rainy and cold, which is consistent with the coefficients of linear regression model (positive coefficients for rain, and negative coefficients for mean temperature).

Section 4. Conclusion

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

Answer: Based on Mann-Whiteny U test result, we could reject null hypothesis, and concluded that people's ridership is significantly different depending on rains. The mean of hourly entry on rainy days is higher than the mean on non-rainy days by 25.
From the linear regression model, both rain and precipitation have positive coefficients (201 for rain, 776 for precipi), which means that this model predicts that ridership increases if rains, and it increases more if it rains more.

Section 4. Reflection

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

People's ridership can be affected so many other factors that are not included in this dataset. For example, it may go up due to traffic conditions, or some other activities to reduce air pollution, or some important events. From the test, we could conclude that people ride more when it rains, but it will be hard to predict the ridership based on this data set. In addition, the relationship between some parameters and ridership may not be linear, and it can't be explained by linear

models. In this test, our p-value is almost same as p-critical value, and we could barely reject the null hypothesis.