Analyze the NYC subway data

Section-1. Statistical Test

1.1
Q. Which statistical test did you use to analyze the NYC subway data?
Answer: I use Mann-Whitney U-Test to analyze the MTA subway data.

Q. Did you use a one-tail or a two-tail P value?
Answer: A two-tail p-value was selected since an appropriate initial question, given the results of the Weather-Related Data section of the *Data Exploration* supplement, is simply whether or not there is a statistically significant difference between the populations

Q. What is the null hypothesis?
Answer: The Mann-Whitney U test is a nonparametric test of the null hypothesis that the distributions of two populations are the same.

Q. What is your p-critical value?
Answer: p=0.0249999127935

1.2
Q. 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:
The data I analyzed include two populations: ridership on rainy days and ridership on non-rainy days. These two populations have possibly unequal variances and sample size. Based on the characteristics of the data, Welch's t-Test or Mann-Whitney U-Test may be used to check the null hypothesis that the mean of two populations is the same against an alternative hypothesis. The Welch's t-Test should meet the following assumptions:

a. Both samples are drawn from normal population
b. The two samples are independent.

Therefore, I examined if the data I used for analysis were normally distributed. First, the histograms for the number of entries per hour for days on rainy days and non-rainy days showed they were not normal distribution. Second, the Shapiro-Wilks test, which is a test to check if a sample came from a normally distributed population, was against the null hypothesis that the populations were normally distributed ($p < 0.05$, Table 1). Taken together, the data did not meet the assumptions for Welch's t-Test. Thus, I chose Mann-Whitney U-Test, which can be used for data with both normal and non-normal distribution.

1.3
Q. 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: To do the test for an alpha level of 0.05, I used the null hypothesis: the ridership on rainy days and non-rainy days are the same. The alternative hypothesis is: the ridership on rainy

days and non-rainy days are not the same. The mean of ENTRIESn_hourly on rainy days was 1105.45 and the mean of ENTRIESn_hourly on non-rainy days was 1090.28. The Mann Whitney U-Test results showed that the p_value for the test was ~0.025

1.4
Q. What is the significance and interpretation of these results?
Answer: Because the p_value (0.025) is less than 0.05, I concluded that the null hypothesis was rejected and the ridership on rainy days and non-rainy days were significantly different.

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?

Answer: Both gradient descent (GD) and OLS models where used to run linear regression on the NYC subway dataset. Both models look for linear relationships between the features and the predicted values or NYC subway rides.

2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?
Answer: In the GD model the features used were: rain, precipitation (precipi), hour of the day (Hour), mean temperature (meantempi) and dummy variables for individual station (UNIT). In the OLS model, the features are used where: rain, mean temperature (meantempi) and dummy variables for stations (UNIT) and dummy variables for hours of day (Hour).

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."

Answer: After mixing and matching various features, these were the most relevant and important features based on their explicatory power and statistical significance. I had a bias for choosing the simplest model possible, without losing too much explicatory power or R^2.

2.5 What is your model's R^2 (coefficients of determination) value?
Answer: The R squared for the GD model is 0.461. The R squared for the OLS is 0.525

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: The R squared for the OLS is 0.525 which means we can explain about 52.5% of the data variability with the model. In other words, our model lets us predict NYC subway entries with 52% accuracy.

Section-3. Visualization

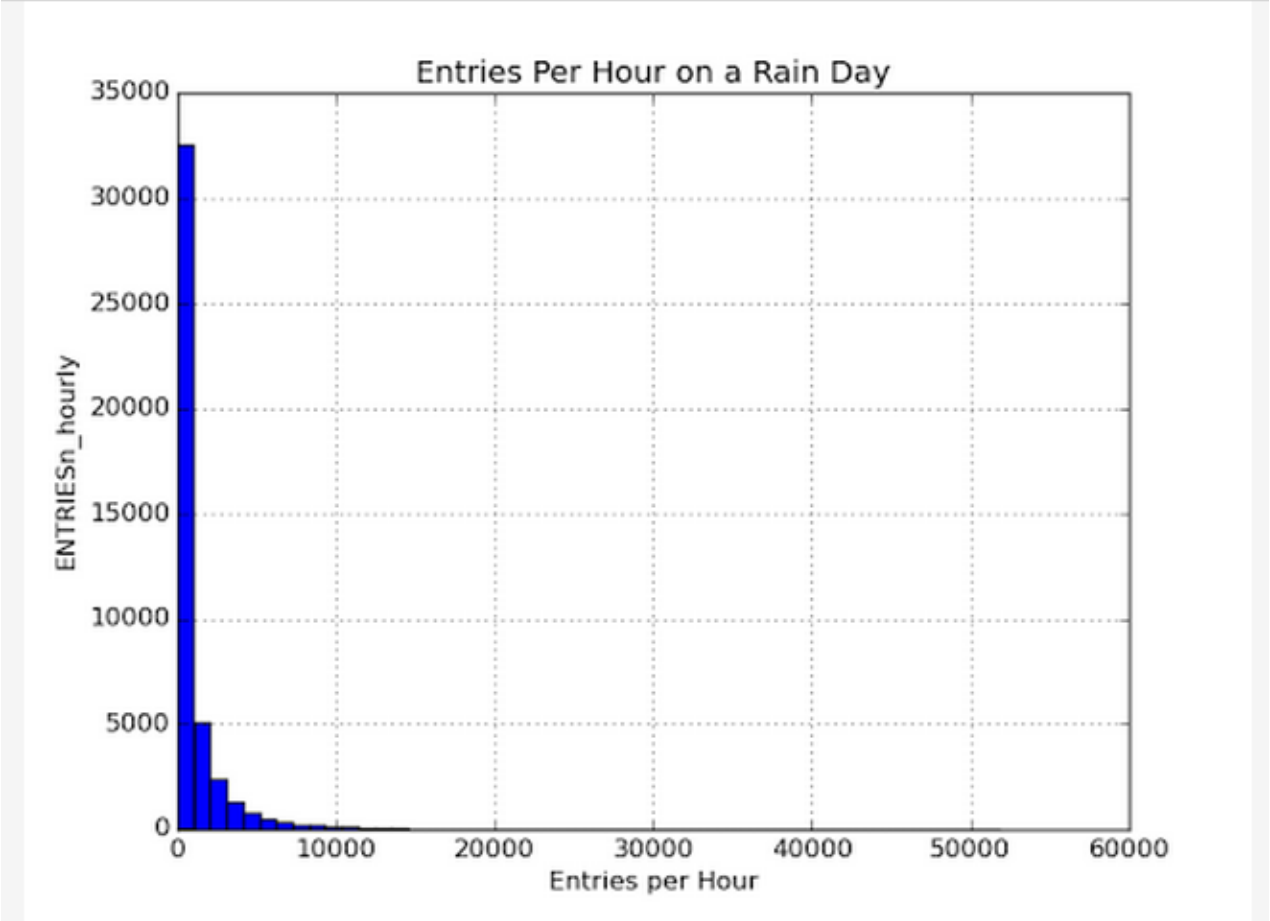
Please include two visualizations that show the relationships between two or more variables in the NYC subway data.Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

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

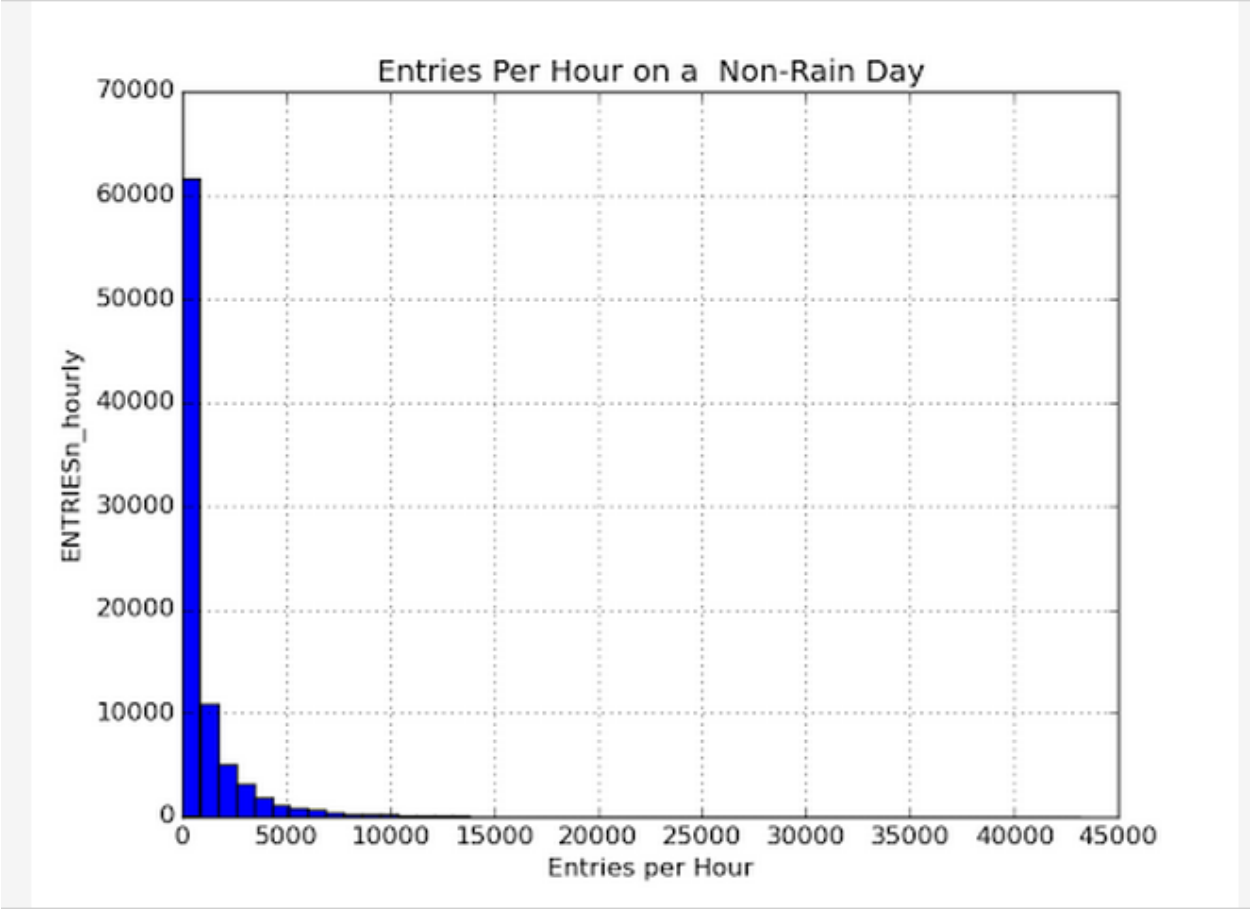
- You can combine the two histograms in a single plot or you can use two separate plots.
- If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.
- For the histograms, 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.

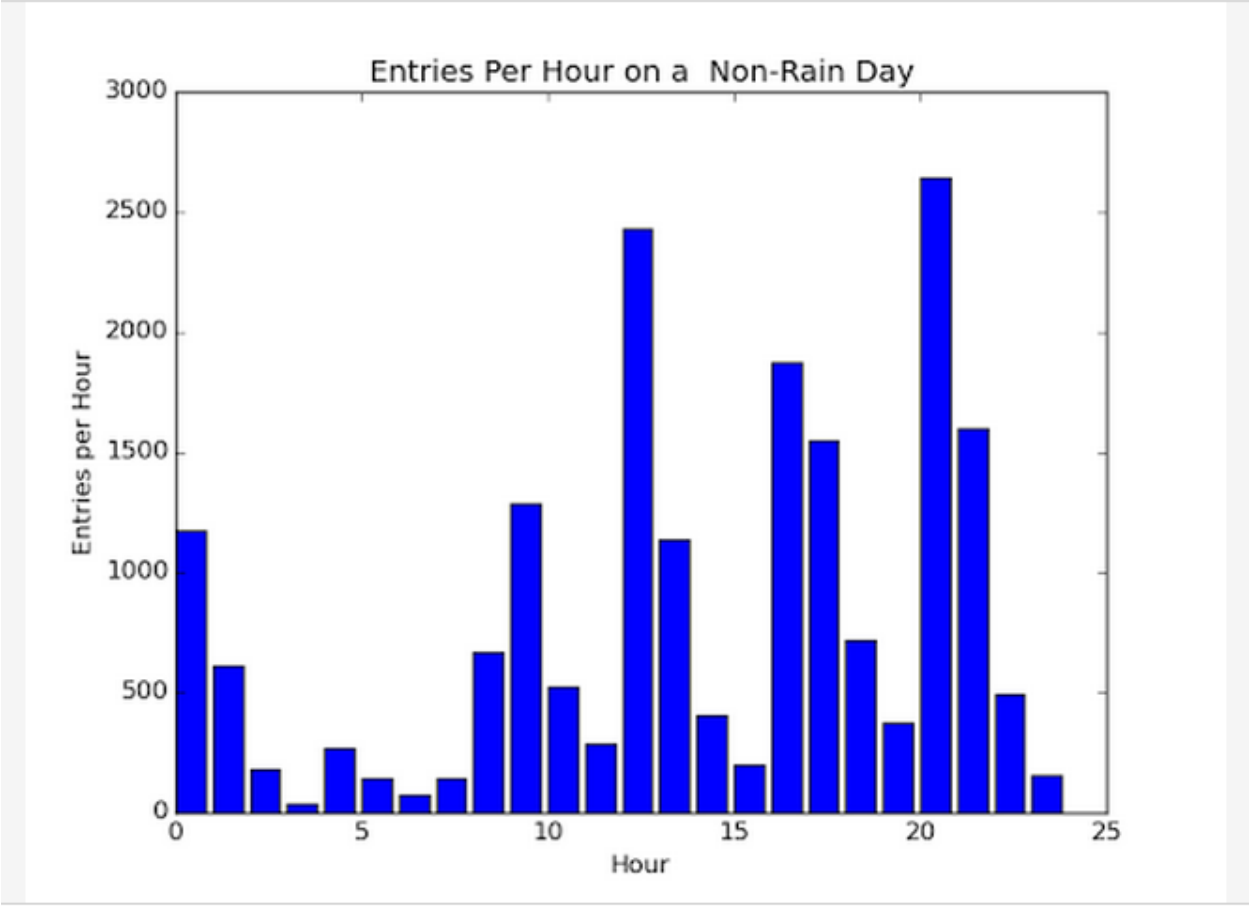
```
 turnstile_weather['ENTRIESn_hourly'][turnstile_weather.rain==1].hist(bins=50
plt.xlabel('Entries per Hour')
plt.ylabel('ENTRIESn_hourly')
plt.title('Entries Per Hour on a Rain Day')
```



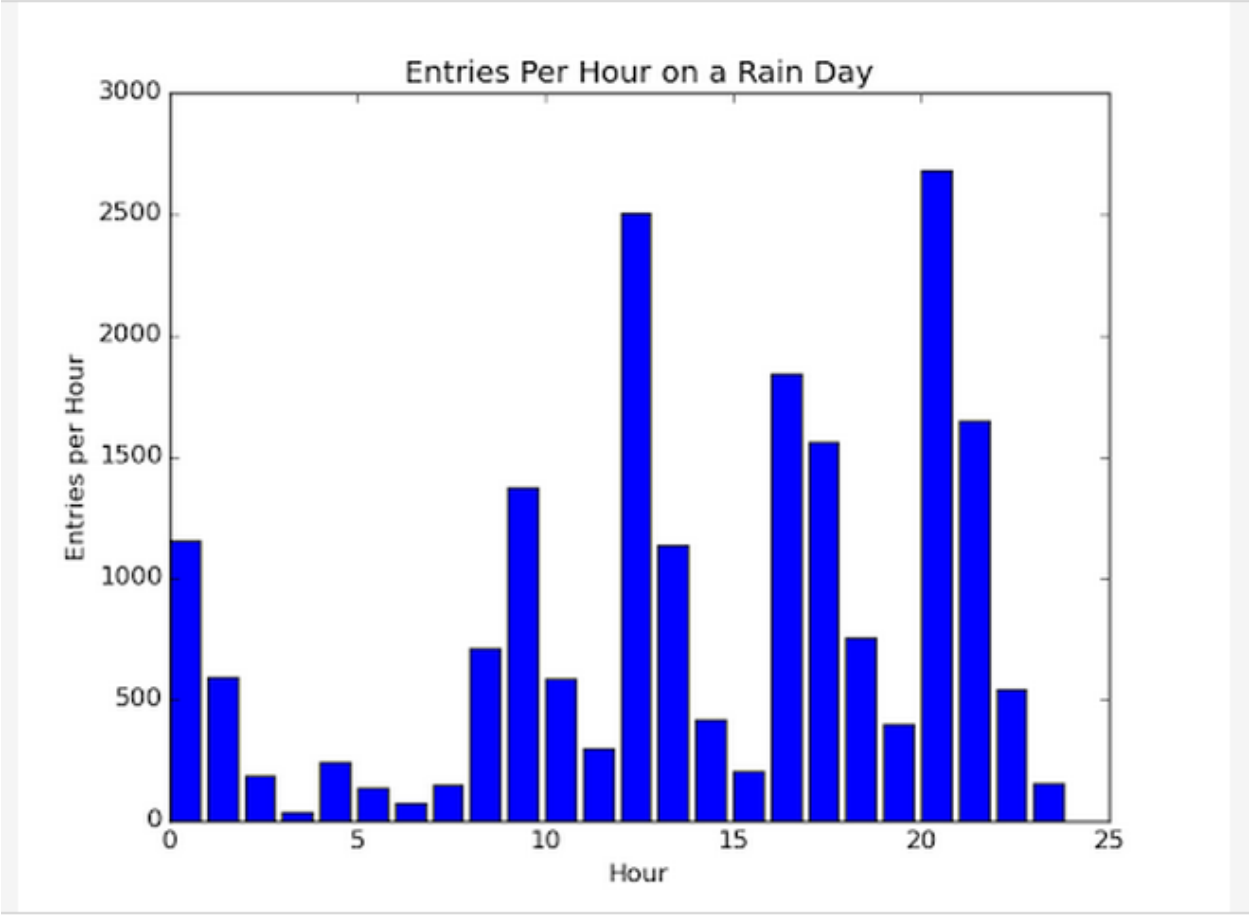
turnstile_weather['ENTRIESn_hourly'][turnstile_weather.rain==0].hist(bins=50
plt.xlabel('Entries per Hour')
plt.ylabel('ENTRIESn_hourly')
plt.title('Entries Per Hour on a Non- Rain Day')



```
Non_rain=turnstile_weather['ENTRIESn_hourly'][turnstile_weather.rain==0]
Hour_non_rain=turnstile_weather['Hour'][turnstile_weather.rain==0]
 p=[]
 for i in range (24):
    hourly=Non_rain[Hour_non_rain==i]
    total=sum(hourly)
    l=len(hourly)
    x=total/l
    p.append(x)
 plt.bar(range(0,24), p)
 plt.ylabel('Entries per Hour')
 plt.xlabel('Hour')
 plt.title('Entries Per Hour on a  Non-Rain Day')
 plt.show()
```



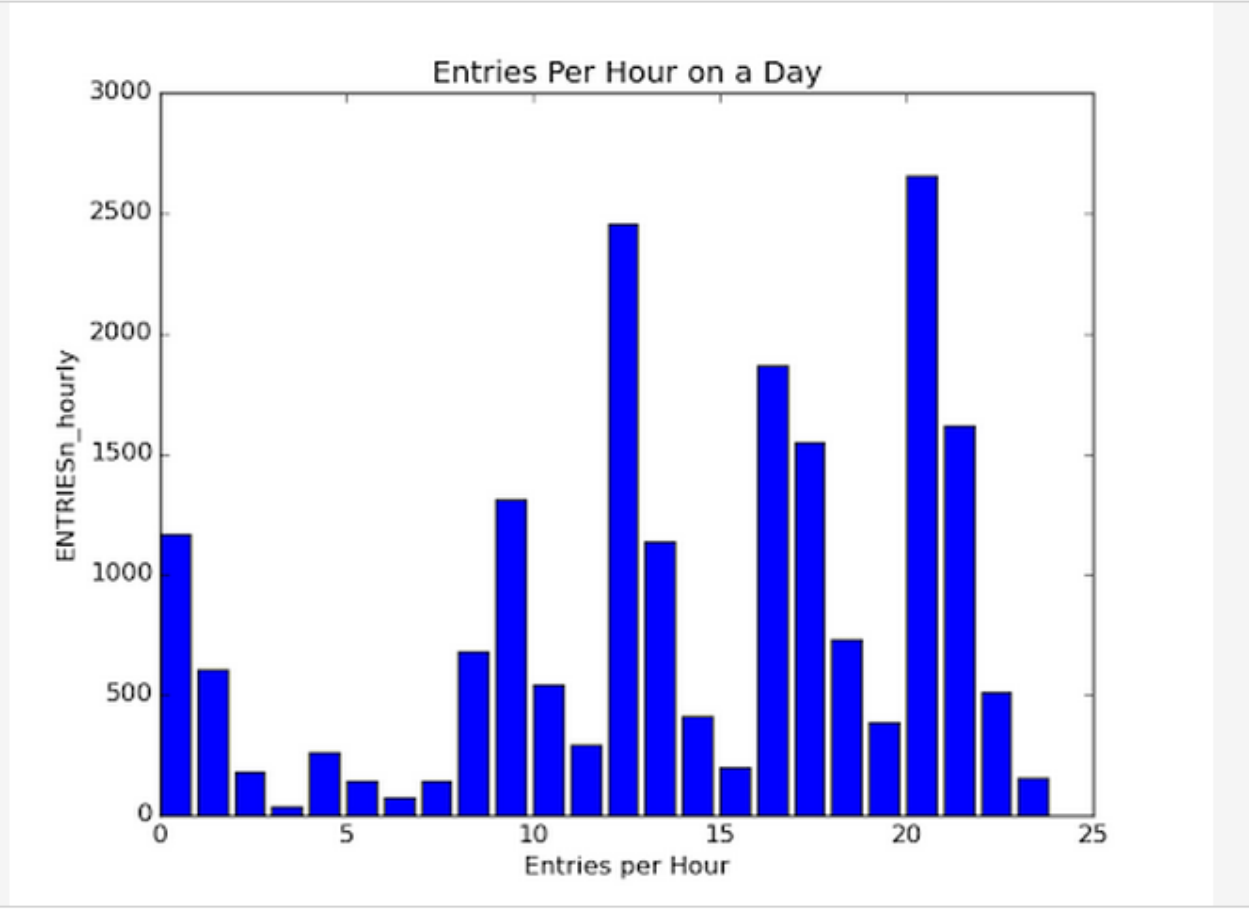
```
rain=turnstile_weather['ENTRIESn_hourly'][turnstile_weather.rain==1]
Hour_rain=turnstile_weather['Hour'][turnstile_weather.rain==1]
p=[]
for i in range (24):
    hourly=rain[Hour_rain==i]
    total=sum(hourly)
    l=len(hourly)
    x=total/l
    p.append(x)
plt.bar(range(0,24), p)
plt.ylabel('Entries per Hour')
plt.xlabel('Hour')
plt.title('Entries Per Hour on a Rain Day')
plt.show()
```



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

```
p=[]
 for i in range (24):
    hourly=turnstile_weather['ENTRIESn_hourly'][turnstile_weather.Hour==i]
    total=sum(hourly)
    l=len(hourly)
    x=total/l
    p.append(x)
 plt.bar(range(0,24), p)
 plt.xlabel('Entries per Hour')
 plt.ylabel('ENTRIESn_hourly')
 plt.title('Entries Per Hour on a Day')
 plt.show()
```



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?

Answer: From the current data set and the analyses performed, it remains inconclusive whether rain has any impact on the number of NYC subway entries. However, based on this data set alone, rain seemed to be an insignificant factor as it related to subway ridership. Thus, further analysis is necessary. On the other hand, based on the data exploration, it seems quite clear that the number of entries is highly dependent on physical location, particularly station position, with specific units having the most importance.

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.
Answer: In this analysis I tested the null hypothesis- the ridership on rainy days and non-rainy days are the same, and an alternative hypothesis- the ridership on rainy days and non-rainy days are not the same at an alpha level of 0.05. I showed that the mean of ENTRIESn_hourly on rainy days (1105.45) was slightly higher than the mean of ENTRIESn_hourly on non-rainy days (1090.28). The Mann-Whitney U-Test results showed a p_value of 0.025 (table 2), which was less than 0.05. Because the p_value (0.025) was less than 0.05, I concluded that the null hypothesis was rejected and the ridership on rainy days and non-rainy days were significantly different. Considering the mean of ENTRIESn_hourly on rainy days was greater than the mean of ENTRIESn_hourly on non-rainy days, the results supported that there were more people ride the NYC subway when it was raining versus when it was not raining on May 2011.

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:

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

Sample size- the dataset only included data from May 2011, increasing sample size may change the results of analysis and conclusions.

Other weather conditions- this analysis only considers rainy and non-rainy condition, but other weather may also impact ridership. For instance, the feature “fog” shows a similar R^2 value as feature “rain” (table 3). It cannot be excluded the possibility that those “fog” or “non-fog” days biased the results of analysis based on rainy days and non-rainy days. To overcome this shortcoming, more detailed comparisons should be used, e.g. comparisons

between “rain and non-fog” and “non-rain and non-fog”, comparisons between “rain and fog” and “non-rain and fog”.

Different months may have different number of rainy days, but this dataset only included data on May. Thus, the number of rainy days or non-rainy days may be biased. To improve this, the data from the whole year should be included for analysis.