Everything below is based on the dataset the problem sets provide, which means they should be mostly a subset of the dataset.

Section 1. Statistical Test

1.1

I used Mann-Whitney Test with a two-tail P value. The null hypothesis is: The distributions of the number of entries are statistically the same for rainy days and non rainy days. My P critical value is 0.05.

1.2

Because Mann-Whitney test does NOT assume our data is drawn from any particular underlying probability distribution.

1.3

P value is 0.0249999. The mean hourly entries for rainy days is 1105.4464, the mean hourly entries for non rainy days is 1090.2788.

1.4

The Mann-whitney test tells us that if there's no difference between the two distributions, the probability for their difference to be equal to or greater than the observed value by pure chance is: 0.024999912793489721 * 2 which is just under the commonly used P critical value 0.05.
Thus we can say the hourly entries in rainy days is different than that in non rainy days.

The mean hourly entries with rain is 1105, the mean hourly entries without rain is 1090.

Because the mean is higher in rainy days, we can conclude there're more people using subway when it's rainy.

Section 2. Linear Regression

2.1

Gradient descent and OLS using *statsmodels* in the optional part of the problem set.

2.2

The features I used are:
'rain', 'fog', 'precipi', 'Hour', 'meantempi', 'weekend', and dummy variables created from 'UNIT' feature using *pandas.get_dummies* method. The column "weekend" is created from "DATEn". Its value is 1 if it's a weekend, 0 otherwise.

2.3

I use 'rain' because I think that people are more likely to use subway when it's rainy outside.

I use 'fog' because of the same reason as 'rain'.

I use ‘precipi’ because I think the amount of precipitation has an effect on how much more likely people will use subway when it’s raining outside.

I use ‘Hour’ because I believe how many people are using subway depends on what time it is. For example, during rush hours when people are going to work or leaving from work, there will be a lot people using subway.

I use ‘meantempi’ because I think high and low temperature will make it uncomfortable to travel outside and more people will use subway.

I use ‘weekend’ because I think if it’s not a weekend, a lot of people will need to go to work or go to school resulting in more people using subway. I also noticed R^2 increased after I included the variable “weekend”.

2.4

Result from running my gradient descent code:

Rain	Fog	Precipi	Hour	Meantempi	weekend
-46.64	51.08	-8.29	468.30	-59.01	-251.66

Results from running OLS code:

Rain	Fog	Precipi	Hour	Meantempi	weekend
-46.92	75.57	-13.91	430.16	-75.39	-239.36

The two-tailed p values for the weights calculated using OLS code are:

Rain	Fog	Precipi	Hour	Meantempi	weekend
2.84e-002	2.08e-004	5.27e-001	7.90e-137	2.03e-005	8.64e-044

As we can see, if our pick 0.05 as the P critical value, every feature above is statistically significant except “Precipi” which has a P value = 0.527.
So we can say with more than 95% confidence that we have the correct weights of the features except “Precipi”.

2.5

My R^2 is 0.4752 with gradient descent, 0.4947 with OLS.

2.6

It means 47.52%(47.52% for gradient descent, 49.47% for OLS) of the variation in ridership can be explained with the 6 features I selected.

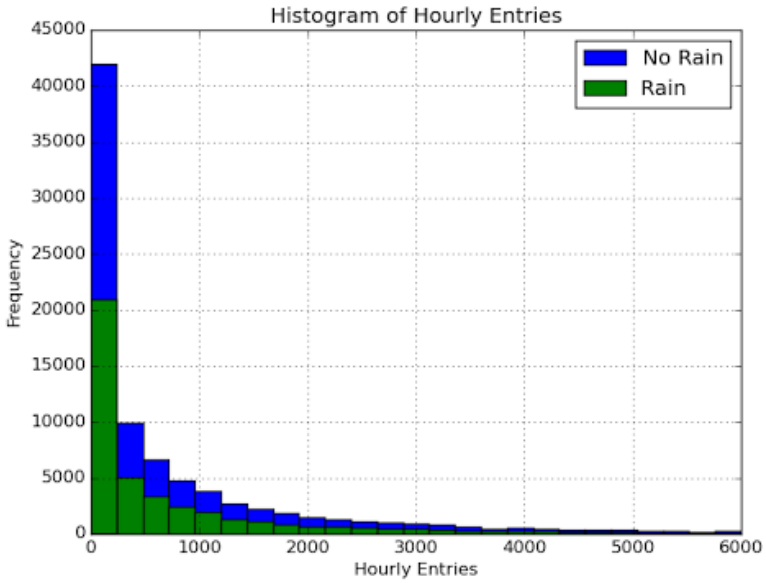
Typically, any attempts to predict human behavior has a relative low R^2, it’s very common to be lower than 50%, so our R^2 is acceptable. But the linear model is not the most accurate model for this problem as some of the features shouldn’t be considered as linear. For example, temperature and hour.

A higher(or lower) temperature doesn't necessarily mean a certain trend of ridership. It depends on whether it's a relatively comfortable temperature for this season. Hour is not a linear feature either. One reason would be that there are a lot of people using subway early in the morning to go to work/school and around 5-9pm to leave from work/school. Thus a non linear model could yield a better fit, especially if we were to look at dataset for more than one month.

Section 3. Visualization

3.1

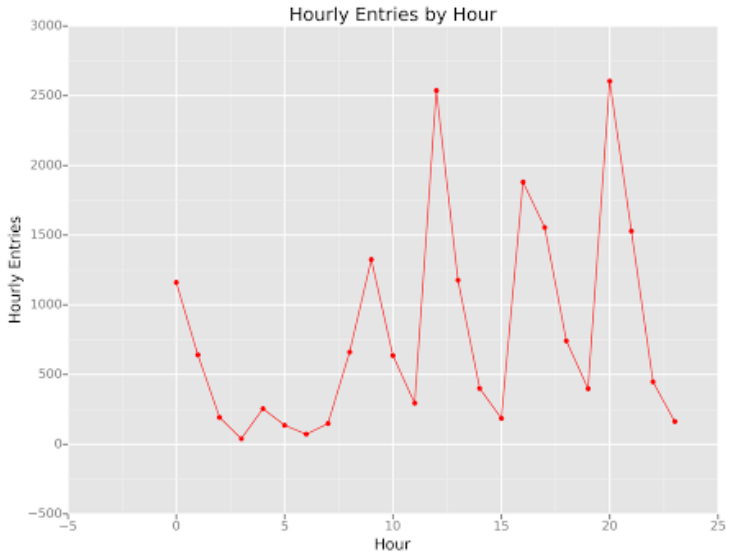
Histogram of ENTRIESn_hourly for rainy days and ENTRIESn_hourly for non rainy days:



The blue bars show the distribution of ENTRIESn_hourly for NON rainy days.
The green bars show the distribution of ENTRIESn_hourly for rainy days.
** This is the same histogram I made in Problem Set 3.1 **
The distribution clearly does NOT resemble a normal distribution. There are a of time low hourly entries, the frequency decreases as hourly entries increases. Very high hourly entries are relatively rare.

From the histogram, we can easily tell there're more non rainy hours than rainy hours. Each blue bar is also roughly twice as high as the green bar with the same hourly entries, so the average hourly entries are probably not very different for rainy and non rainy times based on this histogram. The total number of non rainy hours is approximately twice as many as that of rainy hours.

3.2



Above is a plot of ridership by hour(time of day). I calculated the mean hourly entries at each different time of day.
** This is the same plot I made in Problem Set 4.1 **

From this plot we can see there're a 5 spikes in ridership:
1: Just after midnight, maybe people going home after hanging out at bars and other places.
2: From 7 to 10am, a lot of people going to work/school using subway.
3: Between 11am and 2pm, people going out for lunch and going back to work/school.
4: At roughly 4pm-6pm, people leaving work/school at normal time.
5: Peaks at 8pm, could be people who're working late or people eating dinner outside.
The above analysis mainly focuses on weekdays rather than weekends.

Section 4. Conclusion
4.1
More people ride the NYC subway when it is raining .

4.2
The Mann-Whitney test shows that the two tailed P value is 0.024999912793489721 * 2, which is just under 0.05. So we can say the two samples are different with 95% confidence. And the mean hourly entries is higher when it's raining, so the conclusion is:
More people ride the NYC subway when it is raining.

However, when performing gradient descent or OLS, I found the weight of the "rain" feature is negative. This means if every other feature in the model takes the exact same value, then the predicted hourly entries when it's raining will come out lower than when it's not. This is slightly counter intuitive, given our conclusion is that there're more people using the subway when it's raining. It does still make some sense because when rain is the only variable and every other feature takes the same value, some people may perfer to stay inside rather than going out somewhere using subway when raining.

Correlation does not imply causation. The fact that there're more people using the subway when it's raining does NOT imply raining itself is the cause(as the feature has a negative value). It could be that the more important features(with higher weight) that contribute to more ridership happens to occur very often when it's raining, thus counters the negative weight of the "rain" feature, which in the end results in more people using the subway.

To further explain this idea using statistical results from our sample, I've run some *pandasql* code on the sample.

Here're my findings:

39.8926654741% rainy hours are also foggy.

`Average temperature in May, 2011 when it's not raining: 64.9327414986.`

Average temperature in May, 2011 when it's raining: 62.7960644007.

80.7394156231% rainy hours did not happen in weekends.

Average hour when it's raining: 10.7918902803

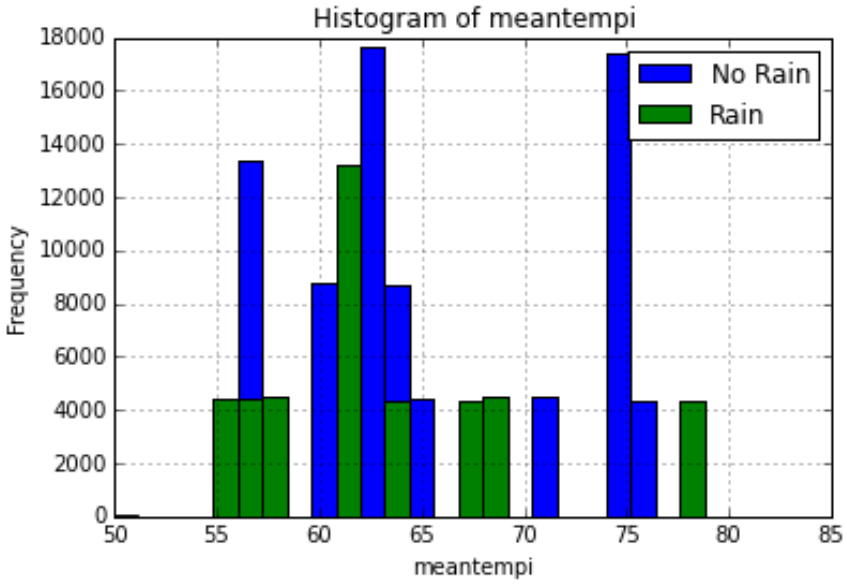
The weights of my features from running OLS are:

Rain	Fog	Precipi	Hour	Meantempi	weekend
-46.92	75.57	-13.91	430.16	-75.39	-239.36

From these weights, there will likely be more people using the subway when it's foggy, later in the day(higer "Hour" value), lower temperature, not weekend.

As we can see, almost 40% of the time when it's raining, it's also foggy, "Fog" feature has a positive weight, its absolute value is larger than that of the "Rain" feature, thus partially counters the negative effect on ridership due to rain.

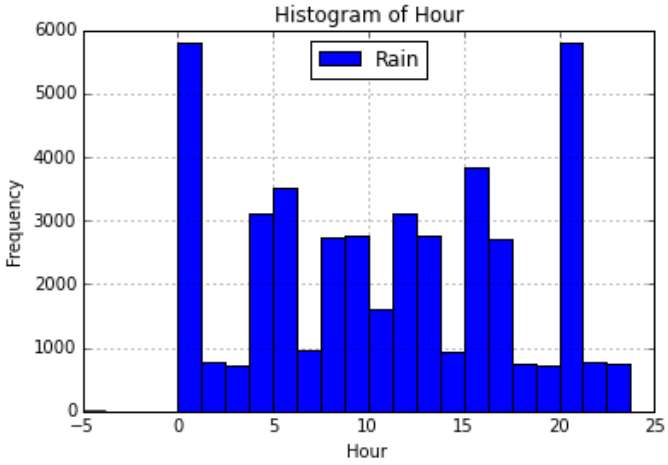
Average temperature when it's raining is slightly lower than that when it's not raining. A two degree difference in mean value is not a huge difference, so I also made a histogram using *matplotlib* to look at the distribution:

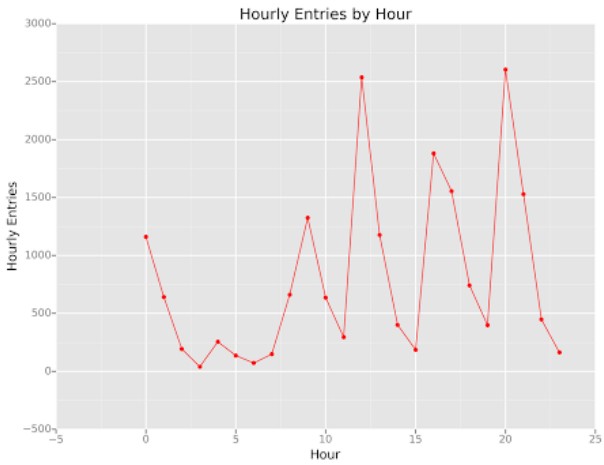


I also ran Mann-Whitney test on them, the result is:
U=1.59e+09, p=0.00e+00
The difference between the two distributions is apparently statistically significant since the test says the chance for the difference to be purely due to chance is lower than the smallest value Python can display.

Since the coefficient of "meantempi" feature is negative and has larger absolute value than that of "rain", the mean difference in temperature is about 2 degrees which is larger than the values of "rain" (either 1 or 0), this will also contribute to resulting in more ridership when it's raining.

On average, the probability that it rains in a workday is: 5/7 = 0.7143. However 80.74% of rains happened during workdays, considering the larger absolute value of the weight of feature "weekend"(which takes the value 1 when it's a weekend, 0 otherwise), this can potentially result in more ridership when it's rainy as well. I made a histogram to demonstrate the hour distribution when it is raining(left):





If we put it side by side with the plot "Hourly Entries by Hour"(right), it's easy to see that the peaks on the two plots somewhat coincide. The peaks on the second plot show that people tend to use subway at certain times, but in May, 2011, it also rains a lot at these times. This is interesting, because these high peaks in rideship are likely due to the fact that people NEED to use the subway to go to work/school/home/lunch/dinner or other places at those times, but the rain also happens to occur often at the same time. This will contribute to resulting in more people using subway when it's raining, but the reason for the increase of ridership is NOT the rain itself.

To sum up, the Mann-Whitney test tells me there are more people using the subway when it's raining. The coefficients calculated using linear regression supports this conclusion. But the rain itself is not necessarily the cause, other features that have larger weights are likely the real reason more people are using subway when it's raining.

Section 5. Reflection

5.1

The dataset has limited data, a larger sample size can definitely help. I also noticed the dataset only has data for any unit every 4 hour, missing a lot of potentially important data. This can potentially affect our accuracy in predicting ridership and drawing conclusion.

To find out if there're more people using subway when it's raining we should also include more months. Or if we're only interested in learning about May in particular, we can include data of May from different years.

More features in the dataset can also help. For example, a feature that indicates whether it's a holiday can help improve our predictions using the model.

The shortcomings mentioned above are all shortcomings of the dataset. As for our analysis tool, a linear model gives a reasonable prediction, but due to the complicated nature of human behavior and the non linear relation between the features and ridership, a linear model is probably not the most accurate choice. For example, the "Hour" feature is definitely much more complicated than a linear model can predict accurately.