Section 1

Question 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?

I decided to apply the **'Mann-whitney-U test'** as it makes no assumptions about the underlying distribution of the two samples in the test. Since we were required to state the conditions(normal or rainy) under which the subway has increased ridership, I was earlier tempted to take a directional approach and conduct a 'one-tail' test. Later, I learned that while conducting the one tailed test, I was making a strong assumption that ridership will only be increased under rainy conditions. Therefore, to avoid any such unwarranted assumptions, I used a 'two tail' significance test.

Details of the Mann-Whitney-U test:
H_0: P(**x > y**) = **0.5 (Null Hypothesis)** **p-critical value: 0.025 (Alpha = 0.05)**
H_A: P(**x > y**) $\neq$ **0.5 (Alternative Hypothesis)**
where:
x : Random draw from a sample of number of passenger entries as recorded by the turnstile after regular 4-hour intervals (in rain)
y : Random draw from a sample of number of passenger entries as recorded by the turnstile after regular 4-hour intervals (without rain)

Null Hypothesis states that random draws from both samples x and y (defined above) are equally likely to be greater than the other whereas the alternate hypothesis states that random draws from both samples x and y (defined above) are not equally likely to be greater than the other.

Question 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.

Histograms of 'hourly turnstile entries' in 'rain' and 'without rain' conditions reveal the underlying distributions which do not appear normal. This can be further validated by applying shapiro tests for both the samples which returned p-values very close to zero. Hence, 'Mann whitney- U test' is appropriate choice for the case at hand as it makes no assumptions about the distributions of the two samples.

Question 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 numerical results for the 'Mann Whitney U' test are listed below:
U-statistic = 153635120.5
Z-score = - 4.54 (Calculated by using the formula given in lecture notes, scipy function returned a 'nan' as p-value, negative as smaller U statistic is conventionally returned by scipy function)
Critical Z-score = - 1.645
p-value = $5.48 \times 10^{-6} \leq 0.025$

Descriptive Statistics: Mean entries (rain) = 2028.20
Mean entries (without rain) = 1845.54

Question 1.4: What is the significance and interpretation of these results?

Given the numerical results (p-value much lower than the p-critical value), we reject the null hypothesis in favour of the alternative. The low p-value suggests that there is significant evidence (with the given level of confidence i.e 95%) that random draws from the passenger entries in rain and passenger entries without rain are not equally likely to be greater than the other.

Section 2

Question 2.1: What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn_hourly in your regression model:

a) Gradient descent (as implemented in exercise 3.5)
b) OLS using Statsmodels
c) Or something different?

I used the OLS (Ordinary least squares) from the statsmodels package.

Question 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 'weekday', 'hour', 'UNIT' and 'tempi' features to predict ridership in the linear model. Among these features, since 'hour' and 'UNIT' are categorical features holding more than two possible discrete levels, dummy variables representing the levels of these variables are part of the model.

Question 2.3: Why did you select these features in your model?

I started off by making a few plots in order to understand the relationship between ridership and other features in the data set.

1. I added the 'weekday' variable after I observed that the ridership is reduced on weekends by plotting density curves.
2. I also concluded by a bar chart that the average subway entries vary with the changing 'hour' of the day in which the data is collected.
3. From problem set 3, I remembered the 'UNIT' variable as a part of our gradient descent model. When I added it to the model, the R^2 increased from 15.6 % to a 54.2 %.
4. Finally, I added the 'tempi' feature because initially I had observed an increasing trend in ridership with increasing temperature (but this observation was later rejected by the linear model, details in section 5)

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

The coefficient for the non dummy variable, 'tempi' is -12.7504 with a 95 % CI [-15.245, -10.256]

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

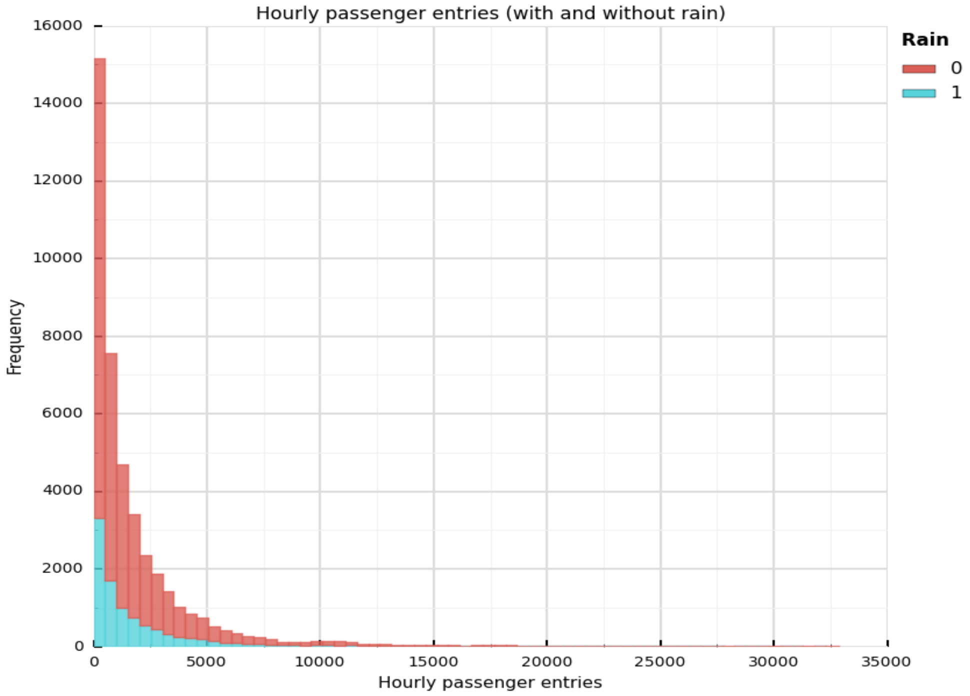
The R^2 value for the fitted linear model is 0.542 or 54.2 %

Question 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?

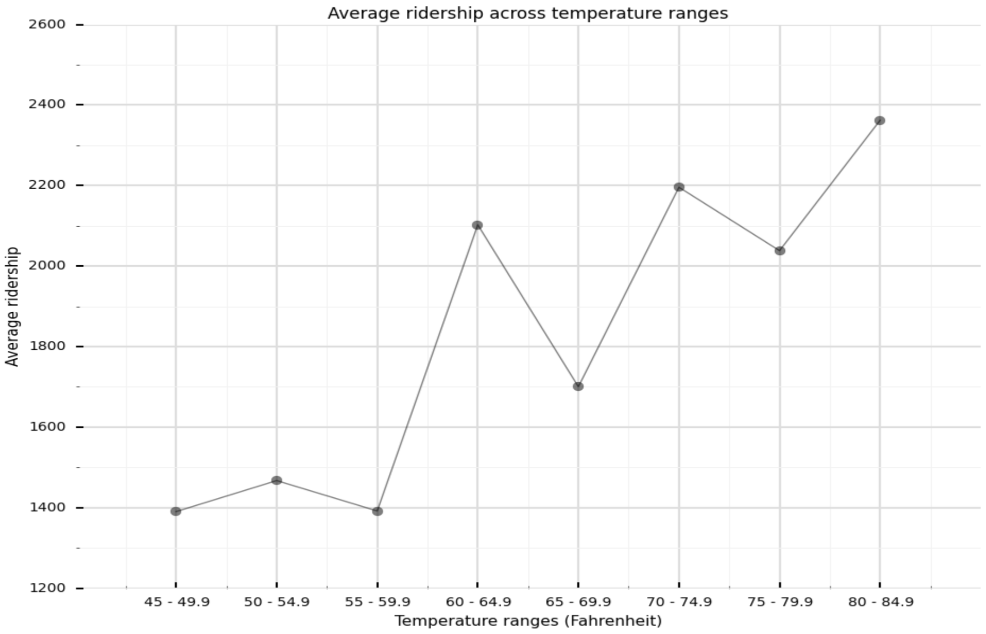
The R^2 value indicates that the fitted model explains 54.2 % of the variability in the given data. Since R^2 is a measure of the goodness of model fit, we can conclude that the model is unable to make very accurate predictions but the predictions will be much improved from the prediction made by mere averaging of the ridership values.

Section 3

The plot gives shows the distribution of the subway ridership (in rain and without rain) which do not appear normal.



The figure below shows the variation of average ridership in different temperature ranges. The visual indicates that although there is no general trend but it seems at higher temperature ranges (60 - 85 Fahrenheit), average ridership is larger as compared to lower temperatures



Section 4

Question 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?

We cannot answer this question solely on the basis of results of a two tail significance test. By rejecting the null hypothesis, we have just inferred that that numbers in one of the two samples (subways entries in rain and without rain) tend to be greater than the numbers in the other. The descriptive statistics associated with the samples on which the significance test was conducted are given in the table below:

	Subway entries in rain	Subway entries without rain
Mean	2048.2	1845.54
Median	939	893
I.Q.R	2129	1928

The above statistics show that mean ridership is greater in rainy conditions as compared to normal conditions Adding to this, a simple regression model with entries as output and rain as the predictor gives us a rain coefficient with the confidence interval [115.549 249.765] and intercept with the confidence interval [1813.726 1877.353]. From this we can conclude that empirically there exists a positive correlation between rain and subway ridership.

Question 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 results from the 'Mann whitney U test' (as stated in the previous answer) led me to conclude that subway ridership in NYC is greater in either rainy or normal conditions. Moreover, by fitting a simple linear regression model with rain as the only predictor, it appeared that rain has a positive correlation with subway ridership.

After adding a few additional features in the linear predictive model, I realized that in presence of variables such as subway station ('UNIT'), hour of the day ('hour'), weekend ('weekday') and recorded temperature ('tempi'), the rain feature is no more significant (critical p-value =0.05). When added to the current model, The 95 % confidence interval for the rain coefficient and its corresponding p-value are [-79.27, 16.96] and 0.204 respectively. Hence as a standalone feature, we do see rain to have a positive correlation with subway ridership and can conclude that under rainy conditions there is increased ridership. But as a feature in the linear model to predict ridership, the rain feature becomes insignificant in the presence of above mentioned variables.

Section 5

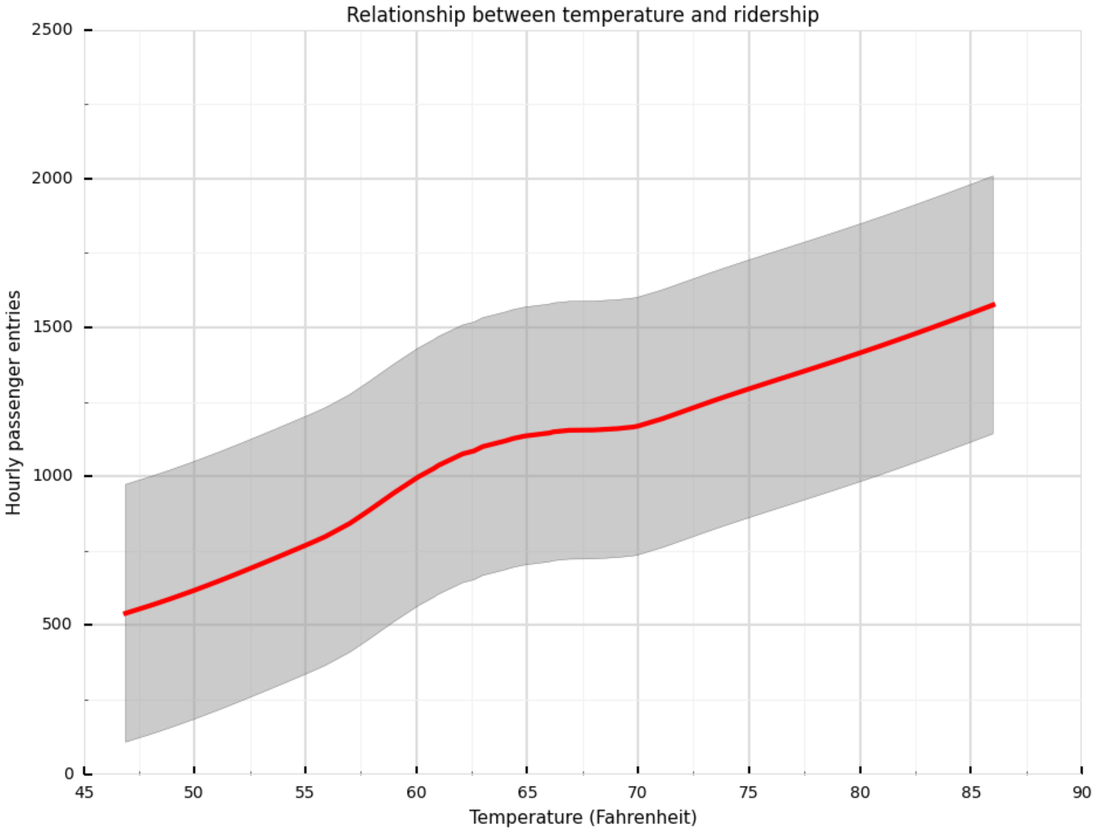
Question 1: Please discuss potential shortcomings of the methods of your analysis

The fitted linear model has an R^2 (coefficient of determination) value of 0.542 which means that 45.8% of variation in the dataset is not explained by the model. Thus we can say that the linear regression predictive model will not make very accurate predictions as compared to some more sophisticated machine learning algorithms. Moreover, the data that we have used to train the linear model is limited to the month of May and due to this limitation, our model may be capturing certain characteristics that are inherent to the conditions in May (unlikely given the selected features in the model but possible).

I chose to use a 'Mann- Whitney U' test to compare the ridership in rainy and non-rainy conditions. Since it is a non-parametric test, it has a lower power as compared to parametric tests that make an assumption about the distribution of the samples in the test. Power for a statistical test is defined as the probability that a test will reject a False NULL hypothesis. Lower power would mean that the "Mann-whitney U" test is less likely to reject a false H_o than its parametric counterparts. Despite being a limitation of the test, it is not relevant to this particular case as were able to reject the H_o.

Question 2: Do you have any other insight about the dataset that you would like to share with us?

Before performing statistical tests or fitting a linear regression model, I made some plots. In one of the plots, I observed an increasing trend of average ridership with increasing temperature (The graph below). But when I fit a linear model, my original observation was challenged as the coefficient of temperature in the model was negative (-13.246) and was significant (95 % CI [-15.856, -10.638]) in presence of variables such as UNIT, hour and weekday. One of the possible explanation is that the temperature variable is related to the hour variable (recorded temperature at busy hours for the subway is higher). Once the hour variable is included in the model, the relationship between temperature and ridership reverses (ridership decreases with temperature when hour is part of the linear model).


Relationship between temperature and ridership
Hourly passenger entries
2500
2000
1500
1000
500
0
45
50
55
60
65
70
75
80
85
90
Temperature (Fahrenheit)