Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about
their effect on the dependent variable? (3 marks)

1. Fall has the highest bike usage followed by summer while spring has the least usage
2. Clear weather has the highest usage while light snow has negligible usage and zero usage in heavy rain
3. There is a significant increase in usage in year 1 [2019] compared to year 0 [2018]
4. Bike sharing usage is low in winter months and consistently increases till June. After that usage remains high till Sept and begins to decline from Oct.
5. Holidays have lower usage though third Quartile has a wider range
6. Across the week, median usage remain consistent, though

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Drop_first=True removes the first dummy variable. This ensures we are able to get the same level of details with an optimised number of columns. For instance, if Spring, Summer and Fall are all False, then we know that the season is Winter. There is no need to have an additional columns for Winter and thus we save on memory usage. On large scale with hundreds of dummy variable, this would be very impactful in optimising memory and computation.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation
with the target variable? (1 mark)

Temp has a high correlation with target variable ie, cnt. We can clearly see the plot showing a trend of increasing cnt with increasing temp. This is also observed in the heat-map plot. Atemp variable also has a high correlation but it is itself dependent on temp as the ‘feeling temperature’.

4. How did you validate the assumptions of Linear Regression after building the model on the
training set? (3 marks)

1. Linear relationship between X and Y - F-Statistic of 263 indicates that independent variables [features selected in the model] have significant effect on dependent variable [in this case ‘cnt’]
2. Error terms are normally distributed with mean at 0 - by plotting the probability distribution of the difference between y values of the train

set to the predicted y values on the same set, we can see that the error terms are normally distributed. We also notice that mean is at 0.
3. Error terms are independent of each other - Durbin-Watson for the final model is 2.02. This means at there is no significant autocorrelation in the residuals.

5. Based on the final model, which are the top 3 features contributing significantly towards
explaining the demand of the shared bikes?

1. Temp: has the highest positive coef of 0.4205 [this also relates to high usage during months from apr to oct]
2. Yr: has the second highest positive coef of 0.2363 [the business is certainly growing from year 2018 to 2019
3. Weather with 'light snow': has the most negative coef of -0.2848 [there is low usage in light snow weather . Please note there is zero usage during heavy rain weather]

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Imagine we have revenue data of a company along with spend on advertisements. Now we have to predict the revenue given the advertisement spend.
We identify that the Revenue is the Target variable as we need to predict the revenue based on a given advertisement spend.

1. We plot the data with Revenue on your axis and ad spend on x axis.
2. We see of there is a general linear trend between the two . If not, then this might no be a good use case for a liner regression
3. Now we need to put line that best fits the provided data. This line is of the from y = mx + b where m is the slope and b is the intercept on y axis
4. The approach is to minimise the sum of squared of the differences between actual and predicted values . This can be achieved using Python libraries of sklearn and statsmodel
5. Once we have a model with coefficients for all the independent variables we need to eliminate features that are of low significance or have a high correlation with other features.
6. We use p-values and VIF scores to remove features from the model one by one and re running the model.
7. We first remove feature with high p-value [greater than 0.05] and again run the model to get new p-values
8. Once we have all features with low p-values we generate the VIF

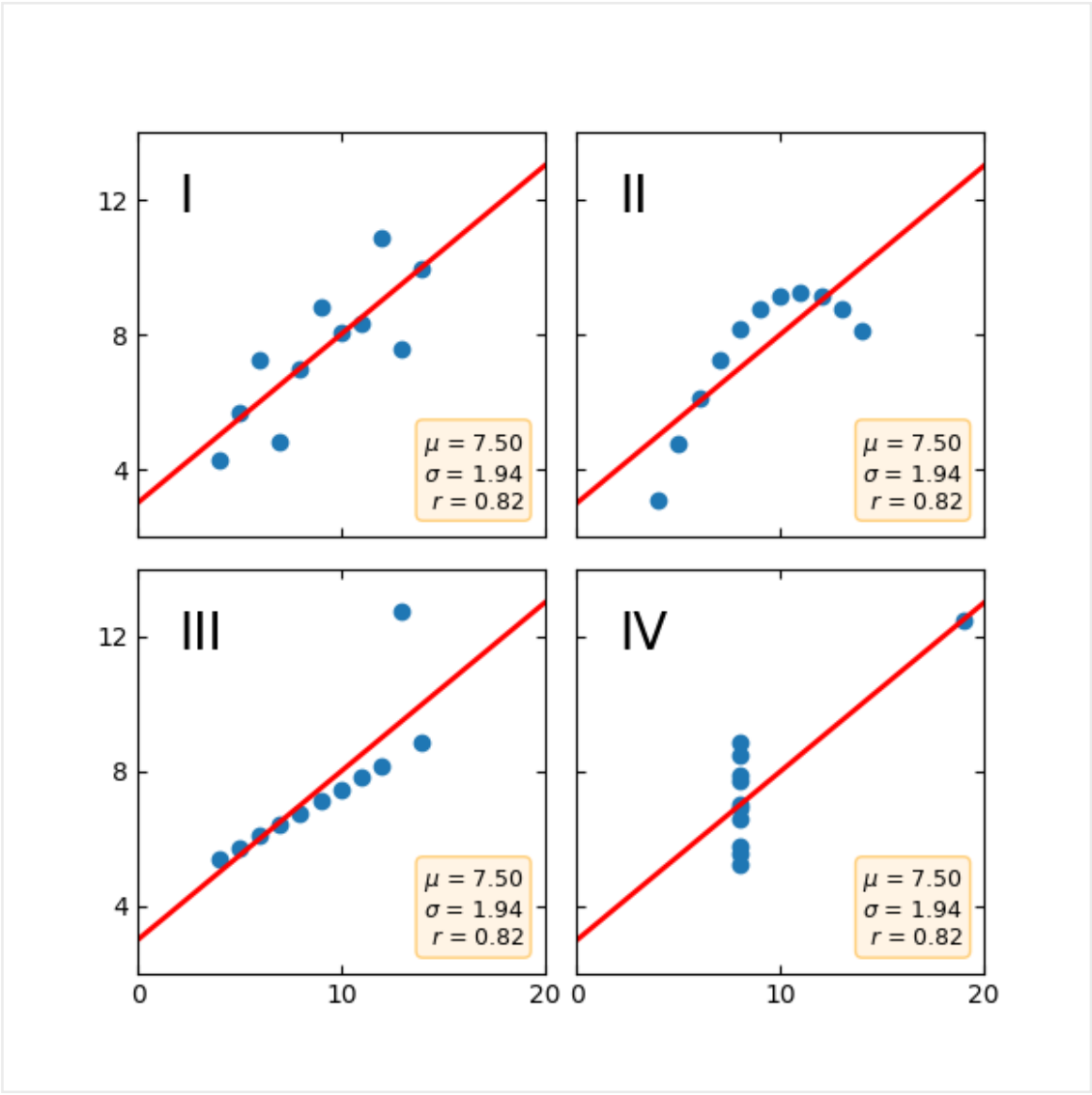
scores and see if there are features with VIF < 5

9. We remove those and run the model again.
10. Repeat 7 to 9 till we have all p-values < 0.05 and al all VIFs < 5
11. Note the coefficients of the features and see which ones have the highest impact on target variable.. Check R2, adjusted R2 and F-Static.
12. Create a distribution plot of residuals [difference of predicted target variable vs actual] to see if it is normal distributed with mean at zero.
13. Apply scaling on Test data - only transform do not fit
14. Get the target and independent features from the Test data set. For the independent features use the sea columns that we got from the modeling exercise
15. Predict the target variable based on the model
16. Evaluate the model by plotting the predicted target variable vs the actual values form the test set. See if the plot looks linear
17. Get the R-squared score on the test data set for the target variable, and the variable containing the predicted values of the target variable on the test set

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet is a group of datasets that have the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation. Each graph plot shows the different behaviour irrespective of statistical analysis. It is used to emphasise the importance of looking at data graphically and not just rely on basic statistic properties.

x = [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]
y1 = [8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68]
y2 = [9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74]
y3 = [7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73]
x4 = [8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8]
y4 = [6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89]



As can be observed, only after plotting on graphs the the datasets show completely different reactions between x and y though they all have the same mean, standard deviation and correlation coefficients

3. What is Pearson's R? (3 marks)

Pearson's R is a statistical formula that measures the strength and direction of the linear relationship between two variable . The value of the Pearson's R is between -1 to +1. When the correlation coefficient comes down to zero, then the data is said to be not related. While, if we are getting the value of +1, then the data are positively correlated and -1 has a negative correlation.

The formula for Pearson correlation coefficient r is given by:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

Where,
r = Pearson correlation coefficient
x = Values in the first set of data
y = Values in the second set of data
n = Total number of values.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is a process of getting all the feature values in the same range like between 0 and 1.

This performed for 1. Ease of interpretation 2. Faster convergence for gradient descent methods.

In Standard scaling the variables are scaled in such a way that their mean is zero and standard deviation is one.

In Normalised or min-max scaling the variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
(3 marks)

VIF is inverse of 1-R^2 so it becomes infinite when R^2 = 1. This means that the variable has perfect multicollinearity with other variables. This indicates that one or more variable in the model are redundant and should be removed.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
(3 marks)
A QQ plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Linear regression models

often assume that the residuals (the differences between observed and predicted values) are normally distributed. Q-Q plots provide a visual tool to assess this assumption. If the residuals are normally distributed, the points on the Q-Q plot should fall approximately along a straight line. Deviations from this line indicate departures from normality, suggesting potential issues with the model's assumptions.