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)

A : Observations from above boxplots for categorical variables:
The month box plots indicate that more bikes are rented during the September month.
The year box plots indicate that more bikes are rented during 2019.
The season box plots indicate that more bikes are rented during the fall season.
The working day and holiday box plots indicate that more bikes are rented during normal working days than on weekends or holidays.
The weekday box plots indicate that more bikes are rented during saturday.
The weathersit box plots indicate that more bikes are rented during Clear, Few clouds, Partly cloudy weather.

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

A: to avoid multicollinearity, if we don't drop ,dummy variables will be correlated. and affects the model adversely.
Also to avoid redundant features.

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

A: count (target variable) has significantly high correlation with temperature (temp)

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

A: residual errors follow normal distribution maintains linear relation between dependant variable (test and predicted)

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

A: **Temperature (temp)** - A coefficient value of '0.5636' indicates that a unit increase in temp variable increases the bike hire numbers by 0.5636 units.

Weather Situation 3 (weathersit_3) - A coefficient value of '-0.3070' indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.3070 units.

Year (yr) - A coefficient value of '0.2308' indicated that a unit increase in yr variable increases the bike hire numbers by 0.2308 units.

General Subjective Questions

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

A: It is a statistical method that is used for predictive analysis. Linear regression is one of the easiest and most popular Machine Learning algorithms. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.
Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

Mathematically, we can represent a linear regression as:
$y= a_0+a_1x+ \varepsilon$,
where,
Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
a0= intercept of the line (Gives an additional degree of freedom)
a1 = Linear regression coefficient (scale factor to each input value).
ε = random error
The values for x and y variables are training datasets for Linear Regression model representation.
Linear regression can be further divided into two types of the algorithm:

- Simple Linear Regression:
  If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.
- Multiple Linear regression:
  If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

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

A: Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.
It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly the same statistical

observations, which provide the same statistical information that involves variance, and mean of all x,y points in all four datasets.
This tells us about the importance of visualizing the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can only be considered a fit for the **data with linear relationships** and is incapable of handling any other kind of datasets.

3. What is Pearson's R? (3 marks)
A: In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient.

However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

Using the formula proposed by Karl Pearson, we can calculate a linear relationship between the two given variables. For example, a child's height increases with his increasing age (different factors affect this biological change). So, we can calculate the relationship between these two variables by obtaining the value of Pearson's Correlation Coefficient r. There are certain requirements for Pearson's Correlation Coefficient:

- Scale of measurement should be interval or ratio
- Variables should be approximately normally distributed
- The association should be linear
- There should be no outliers in the data

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
A: It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

The collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then the algorithm only takes magnitude in account and not units hence incorrect modeling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.
Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

- sklearn.preprocessing.scale helps to implement standardization in python.
- One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
A: If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which leads to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well)

Consider the following linear regression model:
Y = β0 + β1 × X1 + β2 × X2 + β3 × X3 + ε
For each of the independent variables X1, X2 and X3 we can calculate the variance inflation factor (VIF) in order to determine if we have a multicollinearity problem.
Here's the formula for calculating the VIF for X1:

$$VIF_1 = \frac{1}{1 - R^2}$$

R2 in this formula is the coefficient of determination from the linear regression model which has:

- X1 as dependent variable
- X2 and X3 as independent variables

In other words, R2 comes from the following linear regression model:

$X1 = \beta0 + \beta1 \times X2 + \beta2 \times X3 + \varepsilon$
And because R2 is a number between 0 and 1:

- When R2 is close to 1 (i.e. X2 and X3 are highly predictive of X1): the VIF will be very large
- When R2 is close to 0 (i.e. X2 and X3 are not related to X1): the VIF will be close to 1

Therefore the range of VIF is between 1 and infinity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
A: Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution. This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with the same distributions.

Few advantages:
a) It can be used with sample sizes also
b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.
It is used to check following scenarios:
If two data sets —
i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behavior