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)

Ans 1 :
The effect on the dependent variable are as follows :

- Season - Bike Rentals are more during the Fall season and then in summer
- Year - Bike Rentals are more in the year 2019 compared to 2018
- Weathersit - Bike Rentals are more in partly cloudy weather
- Weekday - Bike Rentals are more on Saturday,wednesday and thursday

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

Ans 2 :
It is important to use it, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables

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

Ans 3 :
The highest correlation with the target variable is for column 'temp' and 'atemp'. However during RFE process 'atemp' is removed because it is highly correlated to 'temp' column and can solve the purpose.

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

Ans 4 :
- Linearity : All the other variables have a linear relationship with target variables, was inferred by the graphs plotted
- Independence : A general plot of the data points conveyed that there is independence
- Homoscedasticity : Plotted a graph between residual and predicted values. And the variance of the error is constant as per the graph obtained
- Normality : By performing residual analysis, we could confirm that the error terms are normally distributed
- No multicollinearity : By plotting correlation matrix, we could infer that there is minimum to no multicollinearity

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans 5 :
The top 3 features that contribute significantly towards explaining the demand of the shared bikes are
'temp' - Temperature
'Year' - Year
'season winter' - bikes shared in winter season

General Subjective Questions :-

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

Ans 1 :

Linear regression is a type of supervised machine learning algorithm that computes a linear relationship between a dependent variable and one or more independent features. When the number of independent feature is one then it is called as univariate linear regression. In case of more than one feature, it is known as multivariate linear regression.

General equation of linear regression : $y = mx+c$ OR $y = \beta_1.X_1 + \beta_0$

For multivariate linear regression equation : $y = \beta_0 + \beta_1.X_1 + \beta_2.X_2 + \ldots\ldots + \beta_n.X_n$

Assumptions of linear regression :

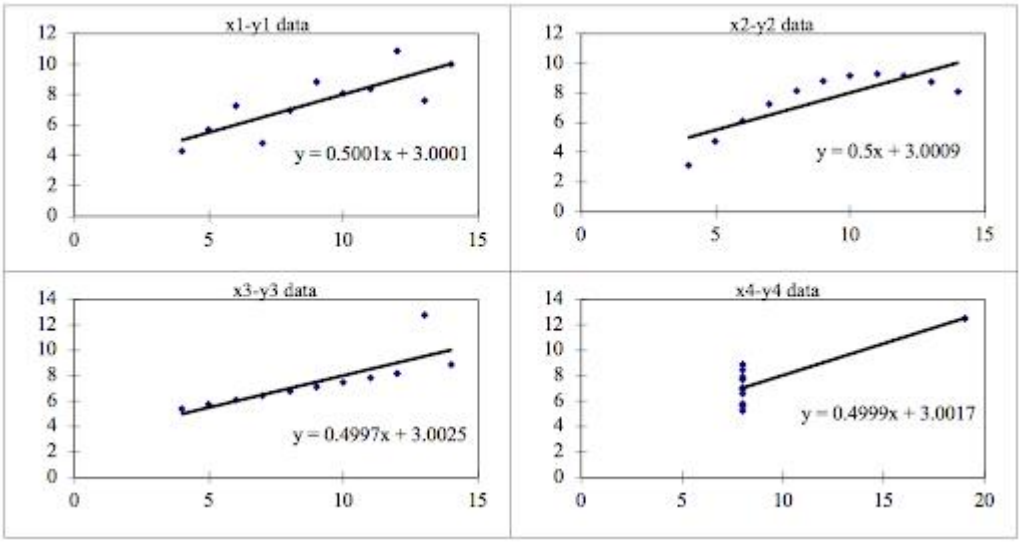
- Linearity : The independent and dependent variables have a linear relationship with one another.
- Independence : The observations in the dataset are independent of each other. This certainly means that the value of the dependent variable for one observation does not depend on the value of the dependent variable for another observation
- Homoscedasticity : Across all levels of the independent variable, the variance of the errors is constant
- Normality : The errors in the model are normally distributed
- No multicollinearity : There is no high correlation between the independent variables.

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

Ans 2 :
Anscombe's quartet was constructed by statistician Francis Anscombe to illustrate the importance of plotting data before you analyze it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another.

Purpose :
It tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.).

Sharing the snip here which shows that after plotting how the graphs turn out to be different from one another :



Interpretation of the above graph :

Graph 1 (x1-y1 data): Fits the linear regression model pretty well.
Graph 2 (x2-y2 data): Cannot fit the linear regression model because the data is non-linear.
Graph 3 (x3-y3 data): Shows the outliers involved in the data set, which cannot be handled by the linear regression model.
Graph 4 (x4-y4 data): Shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

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

Ans 3 :
The Pearson's correlation coefficient (r) is a way of measuring linear correlation. It is a number between -1 and 1 that measures the strength and direction of the relationship between two variables.

Pearson correlation coefficient (*r*)	Correlation type	Interpretation	Example
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the **same direction**.	Baby length & weight: The longer the baby, the heavier their weight.
0	No correlation	There is **no relationship** between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between 0 and −1	Negative correlation	When one variable changes, the other variable changes in the **opposite direction**.	Elevation & air pressure: The higher the elevation, the lower the air pressure.

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

Ans 4 :
Scaling - It is a data pre-processing technique that involves transforming the values of features or variables in a dataset to a similar scale. This is done to ensure that all features contribute equally to the model and to prevent features with larger values from dominating the model.

Why is scaling done - It is needed when working with datasets where the features have different ranges, units of measurement, or orders of magnitude. And we need to bring it at the same level for further modelling.

Difference between normalized scaling and standardized scaling -

Normalization	Standardization
Minumum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling
It is used when features are on different scale	It is used when we want to ensure zero mean and unit standard deviation
Scales values between [0,1] or[-1,1]	It is not bounded to a certain range

It is affected by outliers	It is not much affected by outliers

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

Ans 5 :
VIF becomes infinite when there is a perfect correlation between the variables. This means that one independent variable can be perfectly predicted by other variables in the model.

Lets say that a VIF of 3, this means that the variance of the model coefficient is inflated by a factor 3 due to the presence of multicollinearity.

The approach to be taken for variables having infinite VIF is to remove them from the model as it would degrade the evaluation metrics like r-square and adjusted r-square.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans 6 :
The quantile - quantile (Q-Q) plot is a graphical method for determining whether two samples of data came from the same population or not.

Use and Importance :

- Q-Q plots are useful for checking whether a dataset follows a certain theoretical distribution, such as a normal distribution or a log-normal distribution. If the points on the Q-Q plot fall on a straight line, it indicates that the two datasets have the same distribution. If the points deviate from the straight line, it suggests that the two datasets do not have the same distribution. The degree and direction of deviation from the straight line can provide insights into the nature of the difference between the two datasets.

- They are important and useful for identifying outliers or extreme values in a dataset