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?

Ans : Temp and year has high positive impact on dependent variable count as well as Light_snowrain, holiday and windspeed has high negative impact on dependent variable count

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

Ans : It helps in reducing the extra column created during dummy variable creation. One column can be derived using all other combinations hence its better to remove it. It also helps to reduce correlation due to above-mentioned reason.

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

Ans: temp

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

Ans :

1. Normal distribution of error terms
2. Low VIF
3. High R2 and adjusted R2
4. Low P values

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

Ans :

1. Year
2. Temp
3. Light_snowrain

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans :

Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with a given set of independent variables.

It is represented by following equation −

$Y = mX + c$

Here, Y is the dependent variable we are trying to predict.
X is the independent variable we are using to make predictions.
m is the slope of the regression line which represents the effect X has on Y
c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

Furthermore, the linear relationship can be positive or negative in nature as explained below−

- Positive Linear Relationship: A linear relationship will be called positive if both independent and dependent variable increases.
- Negative Linear Relationship: A linear relationship will be called positive if independent increases and dependent variable decreases.

Linear regression is of the following two types
- Simple Linear Regression
- Multiple Linear Regression

2. Explain the Anscombe's quartet in detail.

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.

I		**II**		**III**		**IV**	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47

I		II		III		IV	
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

For all 4 datasets following is true

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s^2 x	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s^2 y	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places

Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression :	0.67	to 2 decimal places

3. What is Pearson's R?

Ans : It is a statistic that measures the linear correlation between two variables. 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.

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans : Feature Scaling is a technique to standardise the independent features present in the data in a fixed range.

It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Standardization or Z-Score Normalization is the transformation of features by subtracting from mean and dividing by standard deviation

Normalization or Min-Max Scaling is used to transform features to be on a similar scale.
X_new = (X - X_min)/(X_max - X_min)

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

Ans : When the value of VIF is infinite it shows a perfect correlation between two independent variables.

In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Ans : it 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

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

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 same distributions.