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 : **The bike sharing dataset contains six variables related with time, four continuous variables, and the response variable counts of bicycles 'cnt'. The categorical variables in the dataset are season, holiday, working day, and weather. The effect of these categorical variables on the dependent variable (count) can be inferred by analyzing the data using exploratory data analysis techniques such as visualization**

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

Ans : **drop_first=True is important to use, 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 : **'temp','atemp'**

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

Ans : **One way to validate the assumptions of Linear Regression after building the model on the training set is by plotting the residuals distribution. It should come out to be a normal distribution with a mean value of 01.**

**Another way is to detect heteroscedasticity by creating a fitted value vs. residual plot. Once you fit a regression line to a set of data, you can then create a scatterplot that shows the fitted values of the model vs. the residuals of those fitted values**

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

**1)Temperature (temp) - A coefficient value of '0.5173' indicated that a unit increase in temp variable increases the bike hire numbers by 0.5173 units.**

**2)Weather Situation 3 (weathersit_3) - A coefficient value of '-0.2819' indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.2819 units.**

**3)Year (yr) - A coefficient value of '0.2326' indicated that a unit increase in yr variable increases the bike hire numbers by 0.2326 units**.

General Subjective Questions

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

   **Linear Regression is a machine learning algorithm which is based on supervised learning category. It finds a best linear-fit relationship on any given data, between independent**

**(Target) and dependent (Predictor) variables. In other words, it creates the best straight-line fitting to the provided data to find the best linear relationship between the independent and dependent variables. Mostly it uses Sum of Squared Residuals Method.**

**Linear regression is of the 2 types:**

**i. Simple Linear Regression: It explains the relationship between a dependent variable and only one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.**

**Formula for the Simple Linear Regression:**

$$Y=\beta 0+\beta 1X1+\epsilon$$

**ii. Multiple Linear Regression: It shows the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X. It fits a 'hyperplane' instead of a straight line.**

**Formula for the Multiple Linear Regression:**

$$Y=\beta 0+\beta 1X1+\beta 2X2+\ldots+\beta pXp+\epsilon$$

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

**Anscombe's quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph. The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.**

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

**Pearson's correlation coefficient, also called correlation coefficient, a measurement quantifying the strength of the association between two variables. Pearson's correlation coefficient r takes on the values of −1 through +1. Values of −1 or +1 indicate a perfect linear relationship between the two variables, whereas a value of 0 indicates no linear relationship. (Negative values simply indicate the direction of the association, whereby as one variable increases, the other decreases.) Correlation coefficients that differ from 0 but are not −1 or +1 indicate a linear relationship**

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

**Scaling is performed to put different variables on a common scale**

**1. The two most discussed scaling methods are Normalization and Standardization2.Normalization scales the values into a range of 0.1**

**2.Standardization scales data to have a mean of 0 and a standard deviation of 1 (unit variance)2. Standardizing is an example of putting different variables on a common scale1.**

**The difference between normalized scaling and standardized scaling is that the values of a normalized dataset will always fall between 0 and 1, while a standardized dataset will have a mean of 0 and a standard deviation of 1 but the maximum and minimum values are not constrained by any specified upper or lower bounds**

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

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

**A Q-Q plot is a graphical tool for determining if two data sets come from populations with a common distribution such as a Normal, Exponential, or Uniform distribution. In linear regression, Q-Q plot is used to confirm that both the training and test data sets are from populations with the same distributions. Most people use Q-Q plots to fit a linear regression model and check if the points lie approximately on the line. If they don't, the residuals aren't Gaussian and thus the errors aren't either1.**

**The importance of Q-Q plot in linear regression is that it helps us to ensure that our model is based on the right distribution. Having a normal error term is an assumption in regression and we can verify if it's met using this**