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)

   Answer: Below are a few inferences on the anlaysis of categorical variables and their effect on dependant variable

   - **It is clearly visible that Fall has the highest numbers of bikes rented**
   - **On holidays the demand for the rented bike is increased**
   - **On weekdays it is almost the same demand**
   - **Next year demand is going to be high**
   - **September and October is pretty high comapred to the rest of the months in a year. Also, it is very less in start of the year and end of the year.**
   - **Good Weathershit has high demand**

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

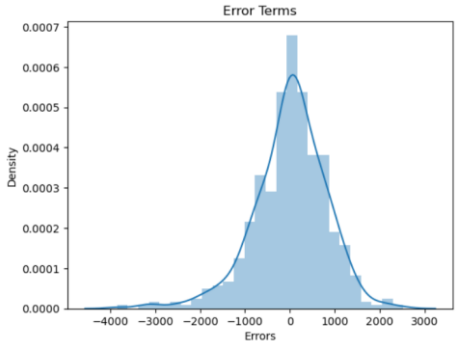
   Answer: It is really important to use drop_first=True as it helps in reducing the extra column created druing dummy variable creation. It also helps to reduce the correlations created among the dummy variables

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

   Answer: *The numerical variable 'atemp' has the highest correlation with target variable 'cnt' with a value of '0.65' followed by 'temp' with a value of '0.64'*

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

   Answer: I have validated the assumptions of Linear Regression by plotting a displot of the residulas and anlysed to see if it is a normal distribution or not. Also, if it has a mean =0. If you look at the below diagram it is distributed normally.



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

   Answer: The following are the top 3 features contributing significantly towards explaining the demands of the shared bikes:

   1. **Temp(temparatues)**
   2. **Light snow**
   3. **Yr(year)**

General Subjective Questions

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

   Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. When the number of the independent feature, is 1 then it is known as Univariate Linear regression, and in the case of more than one feature, it is known as multivariate linear regression. The goal of the algorithm is to find the best linear equation that can predict the value of the dependent variable based on the independent variables. The equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable.

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

   Anscombe's quartet will 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. Anscombe's quartet 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.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

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

   The Pearson coefficient is a type of correlation coefficient that represents the relationship between two variables that are measured on the same interval or ratio scale. The Pearson coefficient is a measure of the strength of the association between two continuous variables.

   - **The Pearson coefficient is a mathematical correlation coefficient representing the relationship between two variables, denoted as X and Y.**
   - **Pearson coefficients range from +1 to -1, with +1 representing a positive correlation, -1 representing a negative correlation, and 0 representing no relationship.**
   - **The Pearson coefficient shows correlation, not causation.**
   - **English mathematician and statistician Karl Pearson is credited for developing many statistical techniques, including the Pearson coefficient, the chi-squared test, p-value, and linear regression.**

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

   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.

   Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. 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.

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

This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables

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

Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against. For example, if you are testing if the distribution of age of employees in your team is normally distributed, you are comparing the quantiles of your team members' age vs quantile from a normally distributed curve.

Advantages:

- **Since Q-Q plot is like probability plot. So, while comparing two datasets the sample size need not to be equal.**
- **Since we need to normalize the dataset, so we don't need to care about the dimensions of values.**