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. Below are my inferences of the categorical variables on the dependent variable cnt:

1) In the fall season, the bikes are rented the most

2) People tends to rent more bike in clear weather

3) Mostly people will rent bikes when there is no holidays

4) Almost everyday the count of bike renting is above 400000

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

Ans. During dummy value creation,to avoid redundant features, it is advisable to use drop_first=True. Dummy variables might be correlated because the first column becomes a reference group during dummy encoding.

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

Ans. The numerical variables 'temp' and 'atemp' both are highly correlated with the target variable 'cnt'.

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

Ans. After building the model on the training set, I carried out the following analysis: -

a) During the spring season, there will be more demand for renting bikes.
b) When there will be heavy rain, snow fall and Thunderstorm, customer will be lesser in number to rent a bike.
c) Month september would be a good time to start a new business strategy

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

Ans. The top 3 features that significantly explain the demand of the shared bikes are:

a) 'atemp'- feeling temperature in Celsius
b) 'mnth' - A subcategory of 'september'
c) 'fall'- A subcategory of 'season'

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

   Ans: Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables . Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output).

   The formula for linear regression is y=mx+c, where y is the dependent variable, x is the independent variable, m is the slope and c is the intercept.

2. **Explain the Anscombe's quartet in detail.**

   Ans: Anscombe's Quartet shows that multiple data sets with many similar statistical properties can still be vastly different from one another when graphed. Additionally, Anscombe's Quartet warns of the dangers of outliers in data sets.

3. **What is Pearson's R?**

   Ans: Pearson's R is a numerical summary of the strength of the linear association between the variables. It varies between -1 and +1. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative. r = 1 means the data is perfectly linear with a positive slope, r = -1 means the data is perfectly linear with a negative slope ,r = 0 means there is no linear association.

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

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

   Scaling is performed during the data pre-processing to handle highly varying values. It is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients.

   Normalized scaling means to scale a variable to have values between 0 and 1, while standardized scaling refers to transforming the data to have a mean of zero and a standard deviation of 1.

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

   Ans: It directly indicates that the particulate variable has severe collinearity. Also, the corresponding variable can be expressed as a linear combination of other variables. In other words, squared multiple correlation of any predictor variable with the other predictors approaches unity.

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

   Ans: The Q-Q plot which is also called as quantile-quantile plot is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential.

   For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption. It’s just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.

   A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that’s roughly straight.