Assignment-based Subjective Questions 1.

1.From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

The final Multiple Linear Regression model contains many predictor variables that are categorical in nature and some of them have been encoded to dummy variables.

spring, winter falls under season category and have been dummy encoded. weathersit_2 and weathersit_3 falls under weathersit category and have been dummy encoded. Similarly, month variables fall under mnth category and have been dummy encoded.

From the predictions, it is clear that these variables are statistically significant and explain the variance in model very well.

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

If you set drop_first = True , then it will drop the first category. So if you have K categories, it will only produce K – 1 dummy variables.

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

atemp is having highest correlation with target variable cnt

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

To validate assumptions of the model, and hence the reliability for inference, we go with the following procedures:

Residual Analysis

Linear relationship between predictor variables and target variable

Error terms are independent of each other

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

Below are the 3 features majorly contributing

```
temp             0.4313
yr               0.2347
month_9           0.0842
```

General Subjective Questions

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

Linear Regression finds the best linear relationship between the independent and dependent variables. It is a method of finding the best straight-line fitting to the given data. In technical terms, linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables. It is mostly done by the Sum of Squared Residuals Method. The assumptions of linear regression are:

a. The assumption about the form of the model: It is assumed that there is a linear relationship between the dependent and independent variable

b. Assumptions about the residuals:
1) Normality assumption: It is assumed that the error terms, ε(i), are normally distributed.
2) Zero mean assumption: It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.
3) Constant variance assumption: It is assumed that the residual terms have the same (but unknown) variance, sigma square. This assumption is also known as the assumption of homogeneity or homoscedasticity.
4) Independent error assumption: It is assumed that the residual terms are independent of each other, i.e., their pair-wise covariance is zero.

c. Assumptions about the estimators:
1) The independent variables are measured without error.
2) The independent variables are linearly independent of each other,
i.e., there is no multicollinearity in the data

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

Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Purpose of Anscombe's Quartet

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

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

Pearson correlation coefficient or Pearson's correlation coefficient or Pearson's r is defined in statistics as the measurement of the strength of the relationship between two variables and their association with each other. In simple words, Pearson's correlation coefficient calculates the effect of change in one variable when the other variable changes.

For example: Up till a certain age (in most cases), a child's height will keep increasing as his/her age increases. Of course, his/her growth depends upon various factors like genes, location, diet, lifestyle, etc.

This approach is based on covariance and, thus, is the best method to measure the relationship between two variables**.**