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: As per model, Categorical variables like "Season_spring", "weathersit_Snow", "yr_1" and "weekday_Mon" have strong relationships with Target variable i.e. "Cnt". Rental bikes demand significantly increase in spring season and snow weather days.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
   Answer: This option helps in creating n-1 dummy variables against category variable which has distinct n values. 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)
   Answer: "temp" has the highest correlation with the target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   Answer: I have performed "Residual Analysis" on predicated target values on train data. I have built Histogram of error terms to check normality. Plot of the error terms with X or y to check independence.

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: As per final model, "temp", "season (spring)" & "hum" are top 3 variables impacting shared bikes demand. When "temp" increases then demand for "Shared Bike" increases. Shared bike demand increases in Spring season. People like use Shared bike in more humid environment.

General Subjective Questions

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

   Answer: Linear regression is predictive modelling technique which is used to investigate relationships between Dependent (Target) and Independent (Predictor) variables. There are 2 types of linear regressions as follows:

   Simple Linear Regression – Where dependent variable is associated with single independent variable.

   Multiple Linear Regression - Where dependent variable is associated with multiple independent variables.

   Best fitted regression line $Y = \beta_o + \beta_1 X$ can be found by minimizing cost function (like Ordinary Least Square function for RSS) through any of approaches like "Differentiation" or "Gradient descent method".

   The strength of linear regression model is explained by R2 where R2 = 1 – (RSS/TSS) where RSS is residual sum of squares and TSS is total sum of squares

   In simple linear regression, we assume that X & Y has linear relation. Errors terms are normal distributed, independent, and constant variance. Hypothesis testing in Linear regression is to determine significance of beta coefficients. i.e. H0: $\beta 1=0$; HA: $\beta 1 \neq 0$.

   In multiple linear regression, we need to take care of "Overfitting", "Multicollinearity", and "Feature selection". We need to convert category variables in dummy variables for better prediction. Model assessment and comparison can be done using "Adjusted R-squared" values which increases if new feature improve the model more than would be expected by chance.

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

   Answer: It is set of four dataset which have 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 four datasets that make up 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. It 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.

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

   Answer: This is most common used correlation coefficient to measure strength of relationship between two variables. It is denoted with symbol "R". This formal return values between 1 and -1 where -1 indicate strong negative relationship, 1 indicate strong positive relationship and 0 indicated no relationship. Its formula is represented as follows –

   $$R= n(\sum xy) – (\sum x)(\sum y) / \sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2}$$

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

   Answer: Scaling is process of normalizing values of independent variables in linear regression process. If it is not done, then a regression algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values. There are two major methods to scale the variables, i.e. standardisation and MinMax scaling. Standardisation basically brings all the data into a standard normal distribution with mean zero and standard deviation one. MinMax scaling, on the other hand, brings all the data in the range of 0 and 1. The formulae in the background used for each of these methods are as given below:

   - Standardisation: $x=x-mean(x)/sd(x)$
   - MinMax Scaling: $x=x-min(x)/max(x)-min(x)$

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

    Answer: 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. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

    Answer: It is a scatter plot that shows the relationship between the quantiles of two datasets. The x-axis of a Q-Q plot represents the quantiles of one dataset, while the y-axis represents the quantiles of another dataset. If the two datasets are drawn from the same distribution, the Q-Q plot will be a straight line. Q-Q plots are useful for checking whether a dataset follows a certain theoretical distribution, such as a normal distribution or a log-normal distribution. If the points on the Q-Q plot fall on a straight line, it indicates that the two datasets have the same distribution. If the points deviate from the straight line, it suggests that the two datasets do not have the same distribution. The degree and direction of deviation from the straight line

can provide insights into the nature of the difference between the two datasets. These plots have several advantages over other graphical techniques for comparing distributions. For one, they are not affected by differences in sample size or scale, if the datasets have the same number of observations. They are also useful for identifying outliers or extreme values in a dataset. Finally, They provide a clear and intuitive visual representation of how two datasets compare in terms of their distributions.