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?

   Answer: After plotting the categorical variables with the target variables on boxplot and I have inferred following effect on target:

   - Demand has grown from 2018 to 2019
   - Demand is continuously growing each month till June. September has the highest demand. After September, demand is decreasing
   - Fall has highest demand for rental bikes
   - The clear weathersit has the highest demand.
   - When there is a holiday, demand has decreased.

   (3 marks)

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

   Answer: 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. If we do not drop one of the dummy variables created from a categorical variable then it becomes redundant with the dataset as we will have constant variable(intercept) which will create multicollinearity issue.
   (2 mark)

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

   Answer: The feature “temp” has the highest correlation. It is very well linearly related with target “cnt”

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

   Answer: I have checked the following assumptions:
   - Error Terms do not follow any pattern.
   - Multicollinearity check using VIF(s).
   - Ensured the overfitting by looking the R2 value and Adjusted R2
   - Linearity Check.

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

   Answer: Features “holiday”, “temp” and season “hum” are highly related to the target column, so these are top contributing features in model building. (2 marks)

General Subjective Questions

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

   Answer: Linear regression may be defined as the statistical model that analyzes the linear relationship between a dependent variable with a given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

   Mathematically the relationship can be represented with the help of the following equation − Y = mX + c Here, Y is the dependent variable we are trying to predict. X is the independent variable we are using to make predictions. m is the slope of the regression line which represents the effect X has on Y c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

   Furthermore, the linear relationship can be positive or negative in nature.

   Positive Linear Relationship: A linear relationship will be called positive if both independent and dependent variables increase.

   Negative Linear relationship: A linear relationship will be called positive if independent increases and dependent variable decreases

   .

2. Explain the Anscombe's quartet in detail. (3 marks)
   Answer: Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fool the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. Each dataset consists of eleven (x,y) points.

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

   Answer: Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. Pearson's r measures the strength of the linear relationship between two variables. Pearson's r is always between -1 and 1. If data lies on a perfect straight line with negative slope, then r = -1.

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

   Answer: Scaling is a method to normalize the range of independent variables. It is performed to bring all the independent variables on the same scale in regression. If Scaling is not done, then the regression algorithm will consider greater values as higher and smaller values as lower values.

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. Example Weight of a device = 500 grams, and weight of another device is 5 kg.
In this example a machine learning algorithm will consider 500 as greater value which is not the case. And it will make a wrong prediction.

Machine Learning algorithm works on numbers not units. So, before regression on a dataset it is a necessary step to perform.

Scaling can be performed in two ways: Normalization: It scales a variable in range 0 and 1. Standardization: It transforms data to have a mean of 0 and standard deviation of 1

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

(3 marks)

Answer: When there is a perfect relationship then VIF = Infinity whereas if all the independent variables are orthogonal then to each other then VIF = 1.0. Means if a variable is expressed exactly by a linear combination of other variables then it is said that VIF is infinite.

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

Answer: Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution It is used for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. Whether the Distributions is Gaussian, Uniform, Exponential or even Pareto distribution, it can be found out.
Few advantages:
a) It can be used with sample sizes also
b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios: If two data sets:
i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behavior

Below are the possible interpretations for two data sets.
a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.
c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.
d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis

(3 marks)