Q1 .From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

ans. Here many insights can be drawn from the plots

1. Season 3,autumn has highest demand for rental bikes
2. Demand for next year has grown
3. Demand is continuously growing each month till June. September month has highest demand. After September, demand is decreasing
4. When there is a holiday, demand has decreased.
5. Weekday and working day is not giving clear picture abount demand.
6. The good weathershit has highest demand
7. During September, bike sharing is high,During the year end and beginning

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

Ans. Techically It is important in order to achieve k-1 dummy variables as it can be used to delete extra column while creating dummy variables

For Example: We have three variables: Furnished, Semi-furnished and un-furnished. We can only take 2 variables as furnished will be 1-0, semi-furnished will be 0-1, so we don't need unfurnished as we know 0-0 will indicate un-furnished. So we can remove it

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

Ans. Atemp and temp has highest correlation equal to 0.63 with target variable.

Q 4. How did you validate the assumptions of Linear Regression after building the model on the training set?
Ans. Distribution of residuals is normal and centred around zero along with mean also zero on plotting a distplot of residuals which must follow normal distribution.

Q 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Ans. As per our final Model, the top 3 predictor variables that contributing significantly towards explaining the demand of the shared bikes are:

1. **Temperature (atemp)** - A coefficient value of '0.4758' indicated that a unit increase in temp variable increases the bike hire numbers by 0.4758 units.
2. **Weather Situation 3 (weathersit_bad)** - A coefficient value of '-0.1728' indicated that, , a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.1728 units.
3. **Year (yr)** - A coefficient value of '0.2462' indicated that a unit increase in yr variable increases the bike hire numbers by 0.2462 units.

General Subjective Questions

Q 1. Explain the linear regression algorithm in detail.

Ans. Linear regression is a statistical practice of calculating a straight line that specifies a mathematical relationship between dependent variable on a number of independent variables also

Linear regression algorithm is defined as an algorithm that provides a linear relationship between an independent variable and a number of dependent variable which may or may not dependent on each other to predict the outcome of future events.

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

Ans Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different. It is often used to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties.

Q 3. What is Pearson's R?

Ans The Pearson correlation method or person correlation cofficient R is the most common method to use for numerical variables; it assigns a value between − 1 and 1, where 0 is no correlation, 1 is total positive correlation, and − 1 is total negative correlation.

This is interpreted as follows: a correlation value of 0.7 between two variables would indicate that a significant and positive relationship exists between the two.

For example A positive correlation signifies that if variable A goes up, then B will also go up, whereas if the value of the correlation is negative, then if A increases, B decreases.

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

Ans 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.

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and

1. **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

Q 5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a high correlation between variables. If the correlation is perfect than R2=1, which in turn makes VIF = 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.
Multicollinearity occurs when independent variables in a regression model are highly correlated with each other, which can cause issues such as unstable parameter estimates and reduced

statistical power. The VIF measures how much the variance of an estimated regression coefficient is increased due to multicollinearity.

Q 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 or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.
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. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.

Use of Q-Q plot in Linear Regression: The Q-Q plot is used to see if the points lie approximately on the line. If they don't, it means, our residuals aren't Gaussian (Normal) and thus, our errors are also not Gaussian.
Importance of Q-Q plot:

I. The sample sizes do not need to be equal.

II. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.

III. The q-q plot can provide more insight into the nature of the difference than analytical methods.