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

Ans: As per the analysis on categorical variables from the dataset, I have made following observations

1. Fall has the highest median followed by summer and the weather situation is clear
2. Year 2019 have higher median
3. The month of June and July we observe highest bike rentals and highest spread in the month on September and October
4. Working day and non-working day has similar median
5. Clear weather is most favourable for bike ride
6. Overall median in the weeksdays is similar but Saturday and Wednesday has highest spread
7. People rent more bikes on holiday the n non-holiday, may be they want to have fun with family members

Q2. Why is it important to use *drop_first=True* during dummy variable creation?

Ans: When we create dummy variables , we can represent n variable with n-1.

For example we can represent weathersit variable in 4 category fall,winter,summer,spring

- 000 will correspond to fall
- 001 will correspond to winter
- 010 will correspond to summer
- 100 will correspond to spring

As you can see fall can be as 000, that is when winter, summer, spring are 0 that means it's fall.

We can drop this variable to save some space and ease the calculations

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation the target variable?

Ans: Looking at the pair plot temp has the highest correlation with target variable cnt.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans: There are below assumption on the linear regression

- Linearity : The relationship between X and the mean of Y is linear.
- Homoscedasticity : The variance of residual is the same for any value of X.
- Independence : Observations are independent of each other.
- Normality : For any fixed value of X, Y is normally distributed.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining
the demand of the shared bikes?

Ans: Top three features contributing the demand are as follows

- Yr (.246)
- Light snow (-.090)
- Spring (.198)

Q6. Explain the linear regression algorithm in detail.

Ans: Linear regression is part of supervised learning

Regression analysis is used for three types of applications:

- Finding out the effect of Input variables on Target variable.
- Finding out the change in Target variable with respect to one or more input variable.
- To find out upcoming trends.

Types of regressions are:

- Linear Regression
- Multiple Linear Regression
- Logistic Regression
- Polynomial Regression

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of the data based on some variables. In the case of linear regression the two variables which are on the x-axis and y-axis should be linearly correlated.

Q7. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, but appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties.

Q8. What is Pearson's R?

Ans: Pearson's R is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations,

Its value is between -1 and 1

Q9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and

standardized scaling?

Ans: Scaling in pre-processing on data to convert the data to same normalized form.

We have observed sometimes, collected data set contains features highly varying in range. If we don't do scaling then algorithm only takes higher values in account and not the lower values, this will create the biased model, hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of range.

Minmax scaling:

It brings all of the data in the range of 0 and 1.

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).

Q10. 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. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. We can drop one of the variables from the dataset which is causing this perfect multicollinearity to solve this problem.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables .

Q11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans: Q-Q plot also known as Quantile-Quantile plot, is used to determine if two data sets come from

populations with a common distribution.