Assignment-based Subjective Questions:

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

Ans: Analysis of the categorical variables from the dataset are:

- The bike demand is high in mnth_sep i.e., mnth_9 compared to other months.
- The bike demand is high in fall i.e., season_3 compared to other seasons compared to other seasons like season_2.
- The bike demand is high in clear_fewclouds weathersit
- The bike demand is high in weekdays compared to holidays
- The bike demand is high in yr. with index1 i.e., 2019 compared to 2018.

Q2) Why is it important to use drop first=True during dummy variable creation?

Ans: It is important because it prevents from creating the extra column when dummy variables for columns are created that prevents from the correlations among variables.

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

Ans: Among the numerical variables, cnt and temp show equally high correlation with target variable.

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

Ans: Validate the assumptions of linear regression by considering:

- Comparing R2, adjusted r2 of both train and test data.

- Also considered p value and vif value. if the r2 is greater than or equal to 75 to till 85 then linear regression is valid.

- Above 85% it means the model might be over fitting may be due to outliers below 75% may be the model might be underfitting. Also, the ytest and ypred must be linear.

- The mean value of residuals will be nearly or always be equal to 0.

- Plotting the graph between test prediction and residual Based on all this factors we are validating the assumptions of linear regressions.

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

Ans: The top 3 features are:

- Temp
- Yr
- mnth_sep or mnth_9

General Subjective Questions

Q1) Explain the linear regression algorithm in detail?

Ans: Linear Regression is an algorithm based on supervised learning means the labels of data is given based on that training of model is done in Machine Leaning. We split the dataset into test and train and find out the best fit line. Basically, in linear regression algorithm we can analyse the relationship between the dependent and independent variables. Even if there is multicollinearity issue, we can still efficiently find and understand the relationship by the best fit line. The accuracy of the linear regression algorithm can be found out by vif's of its features less than 10, p value less than critical value 0.05, and high f-statistic.

Y=mx+c is the equation for linear regression and best fit line equation.We do it to get best fit line which can be obtained by reducing mean squared error i.e.

difference between actual and predicted values. We can handle over fitting and under fitting scenarios in linear regression efficiently and also sensitive to outliers.

Q2) Explain the Anscombe's quartet in detail?

Ans: Anscombe's quartet is the actually quart let graphs means four graphs that appear to be same but they are different when plotted. These are plotted to show the importance of the factors that might affect the analysis like outliers, etc.

Q3) What is Pearson's R?

Ans: Pearson's R is one of the common ways for measuring the relationship between variables are linear or not. It lies between 1 to -1. If it is 1 then if one variable change then dependent variable also changes. If it is -1 if one variable increase then its dependent variable decreases and vice versa. If it is 0 then no relationship or correlation between independent and dependent variables.

Q4) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans: Scaling means standardizing or normalizing the data in a specific range. It is performed on independent variables and numerical variables. Normalized learning means normalizing or standardising the data from 0 to1 and which is also known as min-max scaling. Standardized learning is the normalizing or standardizing the data based on z scores like standardising the data based on mean and standard deviation.

Q5) You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans: The places where I observed vif infinity are:

- When dropped outliers for some of the variables it is resulting in r2 =1 and then resulting in vif =infinity which is indicating that the variable has high correlation.

- vif is also infinity for some variables which has high correlation with multiple variables.

Q6) What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression?

Ans: Quantile-Quantile called as Q-Q plot is a tool that assess if a data came from some distribution such as logical, normal, etc in graphical way. It also determines if the data sets come from a common distribution of dataset or not. IN linear regression we analyse the training and testing data by using Q-Q plot that the data came with same distribution from dataset or not and also dataset have residuals with normal distribution and also skewness.

Advantages:

1. It can be used with sample sizes also.

2. 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.

Uses:

1. Check if two datasets come from populations with a common distribution.

2. Check if two datasets have common location and scale

3. Check if two datasets have similar distributional shapes

4. Check if two datasets have similar tail behaviour