Question 1:-
From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer:-
In the given assignment I first identified categorical variables like ‘season’, ‘year’, ‘month’, ‘holiday’, ‘weekday’, ‘workingday’, ‘weathersit’ against the target variable count, and EDA visualization was made. As from stats this was clear that weather sit median was around 50,000 approximately and similar we could see for ‘year’ and ‘season’. And the final model building shows a significant growth of R^2 and Adjusted R^2 for year, season, etc.

Question 2:-
Why is it important to use drop_first=True during dummy variable creation?

Answer:-
It is important to use drop_first=True during dummy variable creation as It helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

Answer:-
Temp and atemp has the highest correlation with the count variable.

Question 4:-
How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer:-
We tested normal distribution of error terms which we call as residuals by visualizing plot of the error terms.

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

Answer:-
Top 3 features contributing significantly towards explaining the demand of the shared bikes are :

1. Temperature variable - which means if the temperature increases by one unit the number of bike rentals increases
2. Summer & Winter season.
3. Year - 2019 is a higher demanding year than 2018.

General Subjective Questions:-

Question 1:-
Explain the linear regression algorithm in detail
Answer:-
Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.
Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

Question 2:-
Explain the Anscombe's quartet in detail.
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 fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.

Question 3:-
What is Pearson's R?
Answer:-
Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

Question 4:-
What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Answer:-
What-
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.

Why-
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.

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).

sklearn.preprocessing.scale helps to implement standardization in python.
One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

Question5:-

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

Answer:-

If there is perfect correlation, then VIF = infinity. 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.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

Question 6:-

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

Answer:-

In statistics, a Q–Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

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

Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. 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