Assignment-based Subjective Questions: -

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

Ans - I think Numerical variables are just a numeric summary until unless you combine them with categorical variables. i.e- 'temp' is just a numerical column until unless you combine it with any categorical variables such as 'week' of 'day', which could lead us to let us know the average temperature for month or week or which could help us to form some hypothesis.

Ques 2:- Why is it important to use drop _first = True during dummy variable creation?

Ans: - it helps us to reduce the extra columns while creating dummy variables which also helps to reduce to correlations among dummy variables.

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

Ans: - 'atemp' has the highest correlation with target variable 'cnt'.

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

Ans: - We validate the assumptions of Linear Regressions model training set by r-squared Values.

Ques 5: - Based on the final model, which are the top3 features contribution significantly towards explaining the demand of the shared bikes?

Ans: - Based on the final model, Temp, year 2019 and season 'winter' has the highest coeffects values. So we can say these variables decides the demand of bikes.

General Subjective Questions: -

Ques 1:- Explain the linear regression algorithm in detail?

Ans:- Linear regression algorithm is a supervised machine learning algorithms. Where we train a model to predict the dependent variables based on one or more independent variables. Linear models show the relations in the form of linear line between target and predictor.

Ques 2:- Explain the Anscombe's quartet in detail?

Ans :- Anscombe's quartet is made of four datasets which have identical simple statistical properties i.e- mean and standard deviations same but when you plot them on the graph they appear completely different from each other.

Ques 3:- What is Parsons's R?

Ans:- 'Pearson R' is also known as "Pearson Correlations Coefficient". And it measures the linearity between two variables. And the range of correlation is between 1 to -1. There are three types of Correlations: -

1- Positive Correlation: - When both the variables goes in the same direction that's called 'Positive Correlation". When the both the variables are either positive or negative that's called positive correlation.

2- Negative Correlation:- When both the variables have the changes in their course in opposite direction , that is called negative correlation. It means when one variable is positive and other one is negative .
3- Zero Correlation:-When there is no change in any variable that's called 'Zero Correlation.

Ques 4:- What is scaling? Why is scaling performed? What is the difference between normalized scaling and standard scaling?
Ans:- Scaling is a step of pre-processing the data which is applied to independent variables to normalize the data within a particular range. Another befit of scaling is speeding up the calculation in the algorithm.
So often, the data contains different features and different ranges such as numerical values could be in integers or float and other categorical values could be in binary or 1 to 9 range. So in that case if you apply the different algorithms on data so there would be different values according to their previous values, So Scaling helps use to bring the Variables' values in the same range. As we know there are two main approaches to scale the data.

1- Normalization/ Min-Max Scaling: - It brings all data in the range of 0 and 1.
2- Standardization Scaling: - It replaces the values by their Z Scores. It brings all of the data into a standard normal distribution which has mean zero and standard deviation sigma.

Ques 5:- You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans:- It the value of VIF is infinite it means there is perfect correlations between two variables which is 1. In that if we have the correlation value 1, then we need to drop one of the variable from the data set which is causing multicollinearity.

Ques 6:- What is Q-Q? Explain the use and importance of Q-Q plot in linear regression?
Ans:-The Q-Q plots is also knows as Quantile-Quantile plot which helps us to know if two. Samples of data come from the same distribution. And if the normal distribution so when we plot the histogram then mean value is always zero with the normal distribution. That is used for the "Residual Analysis" to see the error distribution in the linear model.