Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about
their effect on the dependent variable?
Answer : 1) I could see that in "Season", Summer and Fall the sales are preety good.
2) IN the Year from April to OCtober the sales are preety good
3) Holiday, Weekday and working day, disn't make much of the impact.
4) In weathersit, if the weather is clear then the sales have gone up.

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

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation
with the target variable?
Answer : Looking at the correlation, the Target variable "cnt" has the highest correlation with 'temp' and 'atem' which is 0.63 and 0.63 respectively

4. How did you validate the assumptions of Linear Regression after building the model on the
training set.
Answer : I validated if their is linear assumption between Target and Indepenent Variable using pairplot and Heatmap.
Also I have vrified the residuals distribution if there are following a normal distribution.

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

Answer : Below are the top three feature contributing significantly towards explaining the demand of the shared bikes
1) Temp
2) windspeed
3) mnth_July

General Subjective Questions

1. Explain the linear regression algorithm in detail
Answer : Linear Regression is a machine learning algorithm which is based on supervised learning category. It finds a best linear-fit relationship on any given data, between independent (Target) and dependent (Predictor) variables.
At the time of building a linear model, we assume that the target variable and predictor variables are linearly dependent. But, apart from these, below are few assumptions in linear regression model:
1) Linear relationship between X and y
2) Normal distribution of error terms.
3) Independence of error terms
4)Constant variance of error terms

2. Explain the Anscombe’s quartet in detail.
Answer : Anscombe’s Quartet was developed by statistician Francis Anscombe. This is a method which keeps four datasets, each containing eleven (x, y) pairs. The
important thing to note about these datasets is that they share the same descriptive statistics. Each graph tells a different story irrespective of their

similar summary statistics

3. What is Pearson’s R?
Answer : In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations, thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling
and standardized scaling?
Answer : Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.

It is a technique to standardise the independent variables present to a fixed range in order to bring all values to same magnitudes.Generally performed during the data pre- processing step and also helps in speeding up the calculations in an algorithm.

For normalization, the maximum value you can get after applying the formula is 1, and the minimum value is 0. So all the values will be between 0 and 1. In scaling, you’re changing the range of your data while in normalization you’re mostly changing the shape of the distribution of your data.

5. 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. A large value of VIF indicates that there is a correlation between the variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Answer : The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential.

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