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?

- Categorical predictor variables ‘season’ and ‘weathersit’ shows a correlation with the dependent variable ‘cnt’ . The relationship is not always clear positive or negative , but the pearson correlation shows significant relationship with target variables.

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

- Using ‘drop_first’ as 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.

Dummy variables creates ‘n’ columns for ‘n’ unique values of a categorical value, although ‘n-1’ variables are enough to represent all unique values of the categorical column. Here using ‘drop_first’ as true drops a column to help get rid of collinearity.

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

- After removing the insignificant variables , ‘temp’ shows the highest correlation with the target variable ‘cnt’. Showing pearson correlation value of 0.63 with target variable.

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

- First did a residual analysis on the error terms, to see if the error terms are normally distributed and centered around 0.
- R2_score on the testing data is also close to the training accuracy.
- Got a r2_score little more than the training accuracy , which was indicating some overfitting, so added some more randomness on the training and testing dataset which made the model more generalized.

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

- Based on the final model ‘temp’ , ‘yr’ and ‘season’ show high correlation with the target variable , hence they can best describe the demand of shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

- Linear regression algorithm is a supervised algorithm, which tries to predict the value of a target continuous variable based on ‘n’ independent continuous variables.
  The algorithm uses a dataset to find the relationship between the independent variables to the target (or predicted) variable. The algo tries to create a generic model to best forecast the target value based on the input values fed.

2. Explain the Anscombe’s quartet in detail.

- Anscombe quartet is essentially a set of four datasets which are statistically identical but show significant difference when plotted on a graph.
  These types of datasets show why eyeballing and visualizing data is important rather than just looking at statistical values.

3. What is Pearson’s R?

- Pearson’s R or Pearson Correlation Coefficient is the measure of relationship between 2 sets of data. It provides a linear measure of relationship between -1 and 1 between the 2 variables.
  The coefficient value, if less than 0 , signifies a negative relationship between variables whereas a value greater than 0 signifies a positive relationship between the 2 variables.

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

- Data scaling is a technique to transform data in a particular range, while maintaining the statistical relationship and distance ratio between data points.
  Scaling is performed to scale down different values under a standard scale so the large magnitude values or outliers do not impact the model. Also it also helps significantly in training the model faster by reducing the computations.

  Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1 like Min-Max scaling whereas standardized scaling technique the values are centered around the mean with a unit standard deviation.

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

- This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 as 1, which leads to 1/(1-R2) to become infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this multicollinearity.

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

- A Q-Q plot or a Quantile-Quantile plot is graphical method of determining if 2 sample sets belongs to the same population.
  This plots a probability distribution of quantiles of the dataset.

QQ Plot helps in determining if the sample sets follow the same distribution behavior and statistical values . As QQ is a probability plot so we do need to worry about the scaling the datasets.