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? (3 marks)

The analysis of the categorial variable yields the following inferences:

- The boxplot between 'weathersit' and 'cnt' infers that most bikes are rented when the weather is Clear, Few clouds, Partly cloudy, Partly cloudy.
- The boxplot between 'season and 'cnt infers that most bikes are rented during the fall season.
- The boxplot between 'holiday' and 'cnt' infers that most bikes were rented when it was not a holiday.
- The boxplot between 'workingday' and 'cnt' infers that most bikes were rented when it was working day.
- The boxplot between 'mnth and 'cnt' infers that most bikes were rented during the month of September.
- The boxplot between 'weekday and 'cnt' infers that most bikes were rented on Monday.

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

The dummy variable created using **drop_first=True** is to reduce the dimension of the categorial variables which can be explained by n-1 levels where n is the level of the categorial variable. The reduction in the dimension of the dummy variable makes the learning of coefficient of the beta-n variables much more efficient as the dimension of the variables is reduces significantly.

Also **drop_first** significantly reduces the possibility of Multicollinearity among categorical variables that can improve the performance of the model.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Among the pair-plots of the numerical variables, the variable 'temp' and 'cnt' have the highest correlation with a value of 0.63.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

After building the Linear Regression Model the validate the assumption:

- Checked the presence of Multicollinearity by using Variance Inflation Factor against the final features which were less than 5.
- Checked the error terms are normally distributed and centered around zero (0)
- Checked if the error terms are independent of each other by plotting a scatter plot of **error_terms** vs **X_train_rfe** variables and they were mostly centered around 0 with no unusual patterns/variance and normally distributed i.e there is **Homoscedasticity**.

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

From the final model, it is found that the top 3 features that contribute significantly towards explaining the demand of the shared bikes are '**year 2019'**, '**temperature'**, and '**spring**'.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a widely used in the field of machine learning and statistics for modeling the relationship between a dependent variable and one or more independent variables. It aims to find the best-fitting linear relationship between the input variables and the output variable, allowing us to make predictions and understand the underlying patterns in the data.

Here's a detailed explanation of the linear regression algorithm:

- Linear regression is used to predict a continuous numeric value (the dependent variable) based on one or more input features (independent variables). The goal is to find a linear equation that best describes the relationship between the input features and the target variable.
- The linear regression model can be represented by the following equation:
  y = b0 + b1x1 + b2x2+ … bnxn
  Where,
  y = the predicted output
  b0 is the intercept
  b1,b2,b3..bn are the coefficient for each independent variable x1,x2,x3,xn
- The Cost Funtion associated with the linear regression should be minimize the difference between the actual dependent values and the predicted dependent value. The most common loss function used in linear regression is the Mean Squared Error.
- The model is trained by repeatedly applying gradient descent to update the coefficients until the cost function converges to a minimum or reaches a stopping criterion.
- Once the model is trained and the coefficients are optimized, you can use the linear regression equation to make predictions for new input data.
- The performance of the linear regression model is often evaluated using metrics like Mean Squared Error, Root Mean Squared Error, and R-squared (coefficient of determination) to assess how well the model fits the data.
- Linear regression assumes that the relationship between the variables is linear, that the errors are normally distributed, and that the variance of the errors is constant (homoscedasticity). It may not perform well when these assumptions are violated or when dealing with complex relationships between variables.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a set of four datasets that were introduced by the statistician Francis Anscombe in 1973 that emphasizes visualizing the data set and not just relying on the summary statistics. Anscombe's quartet explains this by an example that the four data sets have near identical summary statistics such as Mean, Variance and Correlation. This summary statistics information is not sufficient, when a graph is plotted such as a scatter plot then is becoming evident that the four data sets have a different pattern.

- The first data set shows a linear scatter plot.
- The second data set shows a linear scatter plot with variance having outlier.
- The third data set shows weaker linear relationship with higher variance having outliers.
- The fourth data set showcases a curved linear relationship.

So, it can be concluded from Anscombe's quartet that one cannot rely on summary statistics alone to draw a concrete conclusion. It is imperative that the data should be visualized using any graphical representations to view patterns in the data set.

3. What is Pearson's R?

The Pearson's Correlation Coefficient which is denoted by as Pearson's R is a quantifying coefficient that measures the correlation between two continuous variables. The value of the Coefficient r can be between 1 and -1. This coefficient measures the strength of linear correlations and the direction of the linear correlation.

- R = 1, which means that the two continuous variables are highly positively correlated with each other (One variable increases as the other variable increases)
- R = 0, which means that the two continuous variables are not correlated with each other (the change in one variable has no effect on the other variable)
- R = -1, which means that the two continuous variables are highly negatively correlated with each other (One variable increases as the other variable decreases)

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

Scaling is technique where all the values of a variable are transformed into a standardized range. This is to ensure the values have a comparable scale and vectors.

Scaling is performed because it makes the model building during linear regression easier where the machine learns the values of the independent variable's coefficients using gradient decent and uses less computation power as the values are scaled.

Scaling can make the coefficients or weights obtained from linear models more interpretable and comparable.

Normalized scaling, also known as Min-Max scaling transforms the magnitude of variables within the range of 0 and 1. This has a benefit as the outliers too are compressed withing this range. The formula for normalized scaling is :

x (normalized) = x−min(x)/(max(x)−min(x))

here min(x) is the minimum value of the variable and max(x) is the maximum value of the variable and (x) is the value to be scaled.

Standardized transforms the features to have a mean of 0 and a standard deviation of 1. The formula to standardize a feature. x is:

X(standardized) = (x −mean(x))/std(x)

Where mean(x) is the mean value of the feature, and std(x) is the standard deviation of the variable.

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

The Variance Inflation Factor measures multicollinearity between the independent variables (predictor variables) in a linear regression model. An infinite value of VIF indicates that the feature/variable has a perfect linear relationship between the predictor variables. This happens when one variable can be perfectly predicted using a linear combination of other variables, leading to an infinite VIF value. It's a sign of severe multicollinearity, which can distort regression results and affect the interpretability of coefficients.

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

A Q-Q (Quantile-Quantile) plot helps us check if our data fits a certain expected pattern, like a line. In linear regression, it's crucial because it checks if the prediction errors behave normally. If the plot shows a straight line, we're good. If not, something might be off in our analysis, and we need to investigate. It's like a truth checker for our predictions.

In linear regression, we make predictions using data, and we want to be sure they're accurate. We check if the "errors" we make (differences between our guesses and real data) are behaving properly. A Q-Q plot is like a test for these errors. It shows if they're normal, like they should be. If the plot has a straight line, our errors are fine and our predictions are good. It's like a helpful check for our prediction mistakes.