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?

Ans : * The count of rental bikes is low during spring season and high during fall season.
* 2019 has more count of rental bikes than 2018.
* The count of rental bikes is high on clear days and low on rainy and snowy days.
* Average count of rental bikes is high on non-holiday and low on holidays.
* The count of rental bikes increases from January to September and then decreases from October to December which is in relation to the season column.

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

Ans : If we do not use drop_first = True, then n dummy variables will be created, and these predictors(n dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap. This leads to perfect multicollinearity, which causes problems in statistical inference and model fitting. The drop_first=True parameter tells the get_dummies() function to create n-1 dummy variables for a categorical variable with n categories. This means that one category is dropped from the set of dummy variables and used as the reference category.

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

Ans : From the pair plot of numerical variables, it is observed that '**atemp**' has highest correlation with target variable '**cnt**' whereas '**temp'** has the second highest correlation with target variable '**cnt**'.

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

Ans: Assumption 1: There is a linear relationship between target and predictor variable.
We have plotted the scatterplot of target and predictor variables to check the linear relationship.
Assumption 2: Error terms are normally distributed with mean zero.
Residual is calculated by subtracting y_pred from y_test and then plot the histogram of the residuals
From the histogram, we found that the residuals are normally distributed with mean value of zero.
Assumption 3: Residuals are independent of each other.
Durbin-Watson test is used to check the autocorrelation of residuals. The value of Durbin-Watson test is 2.027 for the final model which is close to 2, thus we can conclude that the residuals are independent.

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

Ans : Based on the final model, the top 3 factors affecting the count of rental bikes are:
1. Temperature (temp) - The coefficient of temp is 0.5636, which means that a unit increase in temp will increase the count of rental bikes by 0.5636 units.
2. weathersit_3 - The weathersit_3 indicates ` Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds`. The coefficient of weathersit_3 is -0.3070, which means that a unit increase in weathersit_3 will decrease the count of rental bikes by 0.3070 units.
3. yr - The coefficient of yr is 0.2308, which means that a unit increase in yr will increase the count of rental bikes by 0.2308 units. Thus we can interpret that the count of rental bikes increases year on year by 0.2308 units.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Ans: Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. When the number of the independent feature, is 1 then it is known as Univariate Linear regression, and in the case of more than one feature, it is known as multivariate linear regression. The goal of the algorithm is to find the best linear equation that can predict the value of the dependent variable based on the independent variables. The equation provides a straight line that represents the relationship between the dependent and independent variables.

The mathematical equation can be given as:
Y=mX+b
Here, Y is the dependent variable we are trying to predict.
X is the independent variable we are using to make predictions.
m is the slope of the regression line which represents the effect X has on Y
b is a constant, known as the Y-intercept. If X = 0, Y would be equal to b.

Assumption for Linear Regression Model

Linear regression is a powerful tool for understanding and predicting the behavior of a variable, however, it needs to meet a few conditions in order to be accurate and dependable solutions.

Linearity: The independent and dependent variables have a linear relationship with one another. This implies that changes in the dependent variable follow those in the independent variable(s) in a linear fashion.
Independence: The observations in the dataset are independent of each other. This means that the value of the dependent variable for one observation does not depend on the value of the dependent variable for another observation.
Homoscedasticity: Across all levels of the independent variable(s), the variance of the errors is constant. This indicates that the amount of the independent variable(s) has no impact on the variance of the errors.
Normality: The errors in the model are normally distributed.
No multicollinearity: There is no high correlation between the independent variables. This indicates that there is little or no correlation between the independent variables.

2. **Explain the Anscombe's quartet in detail.**

Ans : 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 analysing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x, y points in all four datasets.
This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the

data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets

The four datasets can be described as:
Dataset 1: this fits the linear regression model pretty well.
Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model.
Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model.

3. What is Pearson's R?
Ans : Pearson's R, also known as Pearson correlation coefficient, is a statistical measure that represents the strength and direction of the linear relationship between two continuous variables. It is named after Karl Pearson, who developed the concept in the late 19th century.
Pearson's R takes values between -1 and +1, where -1 indicates a perfectly negative linear relationship between the variables, +1 indicates a perfectly positive linear relationship, and 0 indicates no linear relationship between the variables.
The coefficient not only states the presence or absence of the correlation between the two variables but also determines the exact extent to which those variables are correlated. It is independent of the unit of measurement of the variables where the values of the correlation coefficient can range from the value +1 to the value -1. However, it is insufficient to tell the difference between the dependent and independent variables
It is independent of the unit of measurement of the variables. For example, suppose the unit of measurement of one variable is in years while the unit of measurement of the second variable is in kilograms. In that case, even then, the value of this coefficient does not change.
The correlation coefficient between the variables is symmetric, which means that the value of the correlation coefficient between Y and X or X and Y will remain the same.

The formula for Pearson's R is:
$r = (n * \Sigma xy - \Sigma x * \Sigma y) / (\sqrt{(n * \Sigma x^2 - (\Sigma x)^2)} * \sqrt{(n * \Sigma y^2 - (\Sigma y)^2)})$
where:
n is the number of observations
Σxy is the sum of the product of the x and y deviations from their respective means
Σx and Σy are the sums of the x and y deviations from their respective means
Σx^2 and Σy^2 are the sums of the squared x and y deviations from their respective means

Pearson's R is commonly used in regression analysis and other statistical applications to determine the strength and direction of the relationship between two variables, and to evaluate the effectiveness of predictive models. However, it only measures the strength and direction of a linear relationship, and may not be appropriate for non-linear relationships or for variables that are not normally distributed.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Ans : Scaling 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. It is is

performed to ensure that each feature contributes equally to the analysis and to eliminate any bias that may be introduced due to the different scales of the features.
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 scaling is a process of scaling the features so that they have a range of values between 0 and 1. This process involves subtracting the minimum value of the feature and dividing by the range of the feature.
The new point is calculated by the below formula for normalization scaling:

```
X_new = (X - X_min)/(X_max - X_min)
```

Standardization scaling is a process of scaling the features so that they have a mean of zero and a standard deviation of one. This process involves subtracting the mean value of the feature and dividing by the standard deviation of the feature. This is often called as Z-score.

```
X_new = (X - mean)/Std
```

The key difference between normalization scaling and standardization scaling is the range of the scaled values. Normalization scaling produces values between 0 and 1, while standardization scaling produces values with a mean of zero and a standard deviation of one. The choice of scaling method depends on the nature of the data and the requirements of the analysis. Normalization scaling is often used for algorithms that require the features to have a similar range of values, such as neural networks. Standardization scaling is often used for algorithms that assume that the data is normally distributed, such as linear regression or logistic regression.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
Ans: The VIF (Variance Inflation Factor) measures the degree of multicollinearity among the predictor variables in a linear regression model. When the value of VIF is infinite, it indicates that there is perfect multicollinearity between the predictor variables, which means that one or more predictor variables are perfectly predicted by the other predictor variables in the model.

Perfect multicollinearity occurs when there is a linear relationship among the predictor variables in the model. For example, if we have two predictor variables that are perfectly correlated, then we can express one of them as a linear combination of the other. In this case, the VIF for the predictor variable that can be expressed as a linear combination of the others will be infinite because it is perfectly predictable from the other predictor variables.

In practice, infinite VIF values can occur due to a few reasons, such as:

- One of the predictor variables is a linear combination of the other predictor variables in the model.
- The data is insufficient to estimate the model parameters, which can happen if there are too few observations or too many predictor variables relative to the sample size.
- The model is over-specified, which means that it has too many predictor variables relative to the sample size, making it difficult to estimate the model parameters.

Infinite VIF values can pose a problem in a linear regression model because it violates the assumption of non-multicollinearity among the predictor variables. Therefore, it is important to identify and address the problem of multicollinearity in the model, either by removing the problematic predictor variable or by transforming the data.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
Ans: A Q-Q plot (Quantile-Quantile plot) is a graphical technique used to compare the distribution of a sample data set to a theoretical distribution, such as the normal distribution. The Q-Q plot plots the quantiles of the sample data against the quantiles of the theoretical distribution. If the data are normally distributed, the points on the Q-Q plot should fall along a straight line. If the data are not normally distributed, the points will deviate from the straight line, indicating that the data are skewed or have heavy tails.

In linear regression, a Q-Q plot is used to assess the normality assumption of the error terms or residuals. The error terms or residuals should be normally distributed with a mean of zero and a constant variance for the linear regression model to be valid. A Q-Q plot of the residuals can help us to visually inspect whether the residuals follow a normal distribution or not. If the residuals follow a normal distribution, the points on the Q-Q plot should fall along a straight line. If the residuals are not normally distributed, the points will deviate from the straight line, indicating that the normality assumption of the error terms may not hold.

The use and importance of a Q-Q plot in linear regression are:

It helps us to assess the normality assumption of the error terms or residuals in a linear regression model.
It provides a visual check of the normality assumption, which is important because formal statistical tests for normality may not be sensitive enough to detect violations of normality.
It can reveal patterns of non-normality that may not be apparent from summary statistics or other diagnostic plots.
It can help us to identify outliers and other unusual observations that may have a disproportionate effect on the linear regression model.