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?**
    - Month wise demand of share bikes stands minimum in the month of February
    - Median value of demand of shared bikes is maximum in fall season followed by summer, winter and spring.
    - Peak demand of shared bikes is in the month of September.
    - Year 2019 has the maximum amount of demand in shared bikes compared to 2018
    - Median value of demand on a non-holiday day is more compared to a holiday.
    - Average and peak count of demand on clear cloud day is maximum followed by mist cloud and light snow.

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

Using drop_first=True during dummy variable creation is important to avoid multicollinearity in linear regression models. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated with each other, which can lead to several issues:

- Model Interpretation: Multicollinearity makes it challenging to interpret the individual impact of each predictor variable on the target variable. It becomes difficult to discern which predictor is truly contributing to the model's predictions.
- Model Coefficients: In the presence of multicollinearity, the coefficients of the correlated variables can become unstable and sensitive to small changes in the data. This instability can make it challenging to trust the results of the regression analysis.
- Inflated Standard Errors: Multicollinearity tends to inflate the standard errors of the coefficients. Larger standard errors lead to wider confidence intervals, making it harder to determine if a coefficient is statistically significant.
- Reduced Predictive Power: Multicollinearity can reduce the predictive power of the model because it introduces noise and instability into the relationship between predictors and the target variable.
- By setting drop_first=True during dummy variable creation, we are essentially omitting one of the categories (levels) of the categorical variable as a reference category. This reference category is represented by zeros in the dummy variables for the other categories. Omitting one category helps mitigate multicollinearity because it ensures that the dummy variables for the remaining categories are not perfectly correlated.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
    - After dropping variable registered and casual, highest correlation with respect to Y is atemp or temp variable. Temp and atemp is also mutually correlated with each other, hence any one variable either atemp or temp can be considered.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
    - P-value of each variable must be less than 0.05
    - VIF of the independent variable must be less than 5
    - R-Square and adjusted R-square of the model must be as high as possible considering the above criteria.
    - Verifying or variable selection through recursive feature elimination and re-iteration.
    - Error terms must be normally distributed.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
    Below are the top 3 features explaining the demand for shared bikes:-
    i) Fall
    ii) Year
    iii) Winter

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Linear Regression is one of the most fundamental algorithms in the Machine Learning world which comes under supervised learning. Basically it performs a regression task. Regression models predict a dependent (target) value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between the dependent and independent variables, they are considering and the number of independent variables being used.

   Linear regression is a supervised machine learning algorithm that predicts a continuous value (i.e., a real number) based on a set of independent variables. The algorithm fits a linear model to the data, which is a line that minimizes the sum of the squared residuals (errors) between the predicted values and the actual values.

   The linear regression algorithm can be expressed mathematically as follows:

   $y = mx + b$

   Where: y is the dependent variable (the value we want to predict)

   x is the independent variable (the variable we use to predict y)

   m is the slope of the line

   b is the y-intercept

   The goal of the linear regression algorithm is to find the values of m and b that minimize the sum of the squared residuals. This can be done using a variety of methods, such as the least squares method.

   Once the values of m and b have been found, the linear regression model can be used to predict the value of y for any given value of x.

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

   Anscombe's quartet is a set of four data sets that have nearly identical summary statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of 11 (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties.

   The four datasets are:

   Dataset 1: $y = x + 3$

   Dataset 2: $y = 0.5x + 2.5$

   Dataset 3: $y = 3 * sqrt(x) + 1.5$

   Dataset 4: $y = x^2 - 2.5$

   Dataset 1 is a linear relationship, while Dataset 2 is a quadratic relationship. Dataset 3 is a curvilinear relationship, while Dataset 4 is a non-linear relationship

   The Anscombe's quartet is a powerful demonstration of the importance of visualizing data before performing any statistical analysis. It shows that summary statistics alone can be misleading, and that it is important to look at the data visually to get a better understanding of its distribution and relationships.

Here are some of the key takeaways from Anscombe's quartet:
Summary statistics can be misleading. The four datasets in Anscombe's quartet have the same mean, standard deviation, correlation coefficient, and coefficient of determination. However, they look very different when graphed. This shows that summary statistics alone cannot tell the whole story about a dataset.
It is important to visualize data. Visualizing data can help you to identify patterns and relationships that may not be apparent from the summary statistics. This is especially important for datasets that are not normally distributed or that have outliers.
Outliers can have a significant impact on statistical analysis. The outlier in Dataset 3 makes the correlation coefficient and coefficient of determination much higher than they would be if the outlier was removed. This shows that outliers can have a significant impact on statistical analysis, and that it is important to remove them before performing any analysis.
Anscombe's quartet is a valuable tool for data scientists and statisticians. It can help you to avoid making mistakes in your analysis and to get a better understanding of your data.

3. What is Pearson's R?

Pearson's R is a statistical measure that indicates the strength and direction of the linear relationship between two variables. It is a number between -1 and 1, where:
A value of 1 indicates a perfect positive correlation, meaning that the two variables increase or decrease together.
A value of -1 indicates a perfect negative correlation, meaning that the two variables increase and decrease in opposite directions.
A value of 0 indicates no correlation, meaning that there is no relationship between the two variables.
Pearson's R is calculated using the following formula:
$r = (n (\sum xy) - (\sum x)(\sum y))/(\sqrt{} [n \sum x2-(\sum x)2][n \sum y2– (\sum y)2)$

Where:

n is the number of data points
$\sum xy$ is the sum of the products of the x and y values
$\sum x$ is the sum of the x values
$\sum y$ is the sum of the y values
$\sum x2$ is the sum of the squares of the x values
$\sum y2$ is the sum of the squares of the y values

Pearson's R is a widely used measure of correlation, but it is important to note that it has some limitations. For example, it can only be used to measure linear relationships. If the relationship between the two variables is nonlinear, Pearson's R may not be a reliable measure of the strength of the relationship.

Here are some of the strengths of Pearson's R:

- It is a simple and easy-to-calculate measure of correlation.
- It is widely used and understood by statisticians and data scientists.
- It can be used to measure the strength and direction of linear relationships.

Here are some of the limitations of Pearson's R:

- It can only be used to measure linear relationships.

- It is sensitive to outliers.
- It is not a good measure of correlation for small sample sizes.

Overall, Pearson's R is a useful measure of correlation that can be used to understand the relationship between two variables. However, it is important to be aware of its limitations when interpreting the results.

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

   Scaling is the process of transforming the features of a dataset so that they have a similar scale. This is done to improve the performance of machine learning algorithms, as many algorithms are sensitive to the scale of the features.

   There are two main types of scaling:

   Normalization: This involves transforming the features so that they have a mean of 0 and a standard deviation of 1. This is the most common type of scaling.

   Standardization: This involves transforming the features so that they have a mean of 0 and a variance of 1. Standardization is less common than normalization, but it is sometimes preferred because it is more robust to outliers.

   Scaling is performed for several reasons:

   - To improve the accuracy of machine learning models.
   - To make the features more comparable.
   - To reduce the impact of outliers.
   - To improve the convergence of machine learning algorithms.

   The main difference between normalized scaling and standardized scaling is that normalized scaling centers the features around 0, while standardized scaling centers the features around their mean and divides by their standard deviation. This means that standardized scaling is more sensitive to the scale of the features than normalized scaling.

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

   The Variance Inflation Factor (VIF) is a measure used in statistics to assess multicollinearity in a regression analysis. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated with each other. When multicollinearity is severe, it can lead to unstable coefficient estimates and make it difficult to interpret the relationships between independent variables and the dependent variable.

   The formula for calculating the VIF for a particular independent variable in a regression model is as follows:

   VIF = 1 / (1 - R^2)

   Where:

   VIF is the Variance Inflation Factor for the variable in question.

   R^2 is the coefficient of determination obtained by regressing the variable of interest against all the other independent variables in the model.

   Now, when the VIF is calculated, it can become infinite in some cases. This occurs when the R^2 value is equal to 1. The independent variable can be perfectly predicted by a linear combination of the other independent variables in the model.

Reasons for infinite VIF values:

- Perfect Multicollinearity: Infinite VIF occurs when there is perfect multicollinearity in the model. Perfect multicollinearity means that one or more independent variables can be expressed exactly as a linear combination of the other independent variables. In such cases, the regression model becomes unstable, and the VIF becomes infinite because you can't estimate the unique contribution of the perfectly collinear variable.
- Singular or Degenerate Matrix: In a regression analysis, the VIF is calculated using matrix algebra. When there is perfect multicollinearity, the matrix used to calculate VIF becomes singular or degenerate, which means it does not have a full rank. In a singular matrix, you cannot compute the inverse, which is necessary to calculate VIF. Hence, the VIF is undefined, and it is often treated as infinite.

Infinite VIF values are a clear indication that there is a serious problem with multicollinearity in the model, and needs to be addressed.

Methods to deal with multicollinearity:

- Removing one or more of the highly correlated variables from the model.
- Combining correlated variables to create composite variables.
- Using regularization techniques like ridge regression or Lasso regression, which can mitigate the impact of multicollinearity.
- Collecting more data to reduce the correlation between variables.

Infinite VIF values occur when there is perfect multicollinearity in a regression model, making it impossible to estimate the unique effects of the collinear variable(s). Identifying and addressing multicollinearity is important for obtaining reliable and interpretable results in regression analysis.

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 is a graphical tool used in statistics to assess whether a dataset follows a specific theoretical distribution, such as a normal distribution. It is particularly useful for comparing the quantiles of the observed data against the quantiles of the theoretical distribution, typically a normal distribution. The main purpose of a Q-Q plot is to visually inspect the similarity between the observed data and the expected distribution.

i) Assessing Normality:

One common use of Q-Q plots in linear regression is to check the assumption of normality of residuals. In linear regression, it is often assumed that the residuals (the differences between observed and predicted values) are normally distributed with a mean of 0.

To assess this assumption, you create a Q-Q plot of the residuals. If the points on the Q-Q plot closely follow a straight line (usually a 45-degree line), it suggests that the residuals are normally distributed. Deviations from the line indicate departures from normality.

ii) Detecting Outliers:

Q-Q plots can also help identify outliers or extreme values in the dataset. Outliers often appear as data points that deviate significantly from the expected line on the Q-Q plot.

iii) Checking Distributional Assumptions:

Q-Q plots are not limited to normality checks. They can be used to assess whether the data follows other theoretical distributions as well, such as the exponential, logistic, or uniform distributions, depending on the context of your analysis.

iv) Model Validation:

In linear regression, Q-Q plots can be a part of model validation. They can be used to assess the validity of the model assumptions, including linearity, independence of errors, and constant variance (homoscedasticity). Deviations from the expected line in the Q-Q plot can suggest issues with these assumptions.

Importance of Q-Q Plots in Linear Regression:

Assumption Checking: Linear regression models often rely on several assumptions about the data. Checking these assumptions is crucial for the reliability of the regression analysis. A Q-Q plot provides a visual and intuitive way to evaluate the normality assumption, which is fundamental in many regression techniques.

Model Improvement: If a Q-Q plot reveals deviations from normality or other distributional assumptions, it may prompt you to explore alternative modeling approaches or to transform the data to better meet these assumptions.

Quality Control: Q-Q plots are a useful tool for quality control in statistical analysis. They help you identify issues with your data that may affect the validity of your conclusions.

In summary, Q-Q plots are valuable tools in linear regression and statistics for assessing the distributional assumptions of your data, particularly the normality of residuals. They provide a visual way to check the goodness of fit and detect outliers, helping you make informed decisions about your regression model and the validity of its results.