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)

Significant variables to predict the demand for shared bikes

- holiday
- temp
- hum
- windspeed
- Season (Summer, Spring)
- Months (January, February, May, June, September, November, December)
- Year (2019)
- Sunday
- Weathersit (Light Snow, Mist + Cloudy)

- Hence when the situation comes back to normal, the company should come up with new offers during summer and spring when the weather is pleasant and also advertise a little for September as this is when business would be at its best.

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

During dummy variable creation, drop_first=True is important because it helps to avoid the "dummy variable trap," which is a situation where one or more of the independent variables in a regression model can be perfectly predicted from the other variables. This is also known as multicollinearity, which can lead to unstable and unreliable regression models.

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

‘Temperature’ has the highest correlation with the target variable

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

1)Normality of error terms
2)Multicollinearity check
3)Linear relationship validation
4)Homoscedasticity
5)Independence of residuals

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

The top 3 features contributing significantly towards explaining the demand of shared bikes are
Spring
Summer
January

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression is a supervised learning algorithm used to predict the value of a continuous output variable based on one or more input variables. It assumes a linear relationship between the input and output variables, and seeks to identify the best linear relationship between them. In other words, it tries to fit a straight line that best describes the relationship between the input and output variables. The general formula for a linear regression model with a single input variable is:

y = b0 + b1 * x

Where:

- y is the output variable (the dependent variable)
- x is the input variable (the independent variable)
- b0 is the intercept, which represents the value of y when x=0

- b1 is the slope of the line, which represents the change in y for a one-unit change in x.

To estimate the values of b0 and b1, the linear regression algorithm uses a technique called ordinary least squares (OLS), which minimizes the sum of the squared differences between the predicted values and the actual values.

The steps involved in performing linear regression are:

1. Data collection: Collect data for the input and output variables. This data is typically in the form of a dataset with several observations or samples.
2. Data preprocessing: Clean and preprocess the data to remove any inconsistencies, missing values, or outliers that may affect the model's performance.
3. Model training: Split the dataset into two parts - a training set and a testing set. Use the training set to fit the linear regression model by estimating the values of b0 and b1 using OLS.
4. Model evaluation: Use the testing set to evaluate the performance of the model. This is done by comparing the predicted values with the actual values of the output variable. Common evaluation metrics for linear regression include the mean squared error (MSE), root mean squared error (RMSE), and R-squared.
5. Model deployment: Once the model has been trained and evaluated, it can be used to make predictions on new data.

Linear regression is a simple yet powerful algorithm that can be used for a wide range of applications, such as predicting housing prices, stock prices, and customer demand. It is also a popular algorithm for feature selection and can be extended to handle multiple input variables, non-linear relationships, and more complex models.

2. Explain the Anscombe’s quartet in detail. (3 marks)

Anscombe's quartet is a set of four datasets that have the same statistical properties when analyzed with basic descriptive statistics but have very different patterns when graphed. These datasets were created by the British statistician Francis Anscombe in 1973 to demonstrate the importance of data visualization in statistical analysis.

The four datasets in Anscombe's quartet are as follows:

1. Dataset I: This dataset consists of a set of (x,y) coordinates that form a straight line with a positive slope. It represents a simple linear relationship between the two variables.

2. Dataset II: This dataset consists of a set of (x,y) coordinates that form a curved pattern that roughly resembles a parabola. It represents a non-linear relationship between the two variables.
3. Dataset III: This dataset consists of a set of (x,y) coordinates that form a straight line with a negative slope, except for one outlier. It represents the impact of an outlier on a linear relationship.
4. Dataset IV: This dataset consists of a set of (x,y) coordinates that form a horizontal line, except for one outlier. It represents the impact of an outlier on a constant relationship.

In summary, Anscombe's quartet emphasizes the importance of data visualization in statistical analysis and demonstrates that relying solely on descriptive statistics can be misleading. By supplementing descriptive statistics with data visualization, one can gain a deeper understanding of the relationships between variables and make more informed decisions.

3. What is Pearson’s R? (3 marks)

4. Pearson's R, also known as Pearson's correlation coefficient, is a statistical measure that quantifies the degree of linear correlation between two continuous variables. It is denoted by the symbol 'r' and can range from -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation.

5. The formula for calculating Pearson's R is:

6. $r = (\Sigma(x - \bar{x})(y - \bar{y})) / \text{sqrt}(\Sigma(x - \bar{x})^2 * \Sigma(y - \bar{y})^2)$

7. where x and y are the two variables being correlated, $\bar{x}$ and $\bar{y}$ are their respective means, and Σ represents the sum of the values.

8. Pearson's R is widely used in statistics, data analysis, and machine learning to measure the strength and direction of the relationship between two variables. It is particularly useful for identifying trends and patterns in data, and for making predictions based on historical data.

4.What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is a preprocessing step in data analysis and machine learning that involves transforming the numerical features of a dataset to a consistent range or distribution. The goal of scaling is to make the features comparable and prevent features with large numerical ranges from dominating the analysis or model fitting process.

Scaling is performed to standardize the data and make it easier to analyze or model. It can help improve the performance of machine learning algorithms and reduce the impact of differences in feature scales on the final results.

Normalized scaling and standardized scaling are two common methods of scaling data. Normalization involves scaling the data so that all values fall within a specified range, usually between 0 and 1. Normalization is performed by subtracting the minimum value from each feature and dividing by the range. Normalized data is useful when the absolute values of the features are not important, but their relative values are.

Standardized scaling involves scaling the data so that it has a mean of 0 and a standard deviation of 1. This is achieved by subtracting the mean from each feature and dividing by the standard deviation. Standardized data is useful when the absolute values of the features are important and the data has a Gaussian distribution.

In summary, scaling is performed to standardize the numerical features of a dataset and prevent differences in feature scales from dominating the analysis or model fitting process. Normalized scaling and standardized scaling are two common methods of scaling data, with each being useful for different types of data and analysis tasks

5)You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

The Variance Inflation Factor (VIF) is a measure of multicollinearity in regression analysis. It measures the degree to which the variance of the estimated regression coefficient is increased due to collinearity among the predictor variables.

In some cases, the VIF value can be infinite. This happens when one of the predictor variables can be perfectly predicted by a linear combination of the other predictor variables. This situation is known as perfect multicollinearity and can occur when two or more predictor variables are highly correlated or when a variable is a linear combination of other variables in the model.

When perfect multicollinearity exists, the regression model cannot be estimated because the matrix of predictor variables is singular, and the inverse of this matrix does not exist. As a result, the regression coefficients cannot be estimated, and the VIF value for the affected variable is infinite.

In summary, an infinite VIF value indicates that there is perfect multicollinearity among the predictor variables, and the regression model cannot be estimated. In such cases, it is necessary to identify and remove the variables causing multicollinearity or to re-specify the model to avoid the issue.

6)What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

A Q-Q plot, short for quantile-quantile plot, is a graphical technique used to assess the distributional similarity between two datasets, typically a sample dataset and a theoretical distribution, such as the normal distribution. The Q-Q plot plots the quantiles of the two datasets against each other on a scatterplot.

In linear regression, Q-Q plots are often used to check the assumption of normality of residuals, which is one of the key assumptions of linear regression. Residuals are the differences between the observed values and the predicted values of the response variable. The assumption of normality of residuals implies that the distribution of the residuals is approximately normal with a mean of 0 and constant variance.

To create a Q-Q plot for residuals, the residuals are first standardized to have mean 0 and standard deviation 1. The standardized residuals are then plotted on the y-axis, while the expected values from a normal distribution with mean 0 and standard deviation 1 are plotted on the x-axis. If the residuals are normally distributed, the Q-Q plot should show a straight line, indicating that the observed values are close to what would be expected under a normal distribution.

If the Q-Q plot shows deviations from a straight line, it may suggest non-normality of the residuals. This can indicate issues such as outliers, heteroscedasticity, or other violations of the assumptions of linear regression. In this case, additional diagnostic tests may be needed to identify and address the underlying issues. Therefore, Q-Q plots are an important tool for assessing the assumptions and checking the validity of linear regression models.