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?

The demand of bike is less in the month of spring when compared with other seasons.

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

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.

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

From Pair-Plot tells us that there is a LINEAR RELATION between atemp and temp both have same correlation with target variable of 0.63 which is the highest among all numerical variables.

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

According to this assumption there is linear relationship between the features and target. Linear regression captures only linear relationship. This can be validated by plotting a scatter plot between the features and the target.

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

Based on final model top three features contributing significantly towards explaining the demand are:

- Temperature (0.552)
- weathersit: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (-0.264)
- year (0.256)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The goal of linear regression is to find the line (or hyperplane in higher dimensions) that best describes the relationship between the variables.

The algorithm for linear regression involves several steps:

1. Data Preparation: The first step is to collect and prepare the data. This involves gathering data for the dependent and independent variables, cleaning the data by removing any missing values or outliers, and splitting the data into training and testing sets.

2. Model Training: The next step is to train the linear regression model on the training set. This involves finding the line (or hyperplane) that best fits the training data. The line is determined by minimizing the sum of the squared errors between the predicted values and the actual values of the dependent variable.

3. Model Evaluation: After training the model, we evaluate its performance on the testing set. This involves calculating metrics such as the mean squared error or the R-squared value to determine how well the model fits the testing data.

4. Model Optimization: If the model's performance is not satisfactory, we can adjust the model to improve its performance. This may involve adding or removing independent variables, transforming the data, or using regularization techniques such as ridge regression or Lasso regression to reduce overfitting.

5. Model Deployment: Once the model's performance is satisfactory, we can deploy it to make predictions on new data.

2. Explain the Anscombe’s quartet in detail.

The datasets are fundamentally different from each other in terms of their properties and relationships between the variables. Here are the characteristics of each dataset in Anscombe's quartet:

1. Dataset I: This dataset has a linear relationship between x and y, with a strong positive correlation. It is an example of a simple linear regression.

2. Dataset II: This dataset has a non-linear relationship between x and y, with a clear outlier. It is an example of a case where a single outlier can have a significant effect on the correlation coefficient and regression line.

3. Dataset III: This dataset has a linear relationship between x and y, but it is driven by a single outlier. The outlier has a large influence on the correlation coefficient and regression line, demonstrating the importance of identifying and handling outliers.

4. Dataset IV: This dataset consists of two distinct clusters of data, each with its own linear relationship between x and y. It is an example of a situation where the overall relationship between x and y can be misleading if the data are not properly analysed.

Anscombe's quartet highlights the importance of visualizing data and conducting exploratory data analysis to gain a deeper understanding of the underlying relationships and properties of the data. Simply relying on summary statistics such as the correlation coefficient or regression line can be misleading and may result in incorrect conclusions or decisions.

In summary, Anscombe's quartet is a set of four datasets that appear to have similar linear relationships between x and y but are fundamentally different from each other in terms of their properties and relationships between the variables. It serves as a reminder of the importance of visualizing data and conducting exploratory data analysis to gain a deeper understanding of the data.

3. What is Pearson’s R?

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the linear relationship between two variables. It is a number between -1 and 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation.

The formula for Pearson's R is:

r = (Σ[(x - mean(x)) * (y - mean(y))]) / (sqrt(Σ(x - mean(x))^2) * sqrt(Σ(y - mean(y))^2))

where x and y are the two variables, mean(x) and mean(y) are their respective means, and Σ denotes the sum of the values.

Pearson's R is commonly used in data analysis and machine learning to quantify the strength and direction of the relationship between two variables. It can help identify patterns and trends in the data, as well as provide insights into cause-and-effect relationships.

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

Scaling is a data pre-processing technique in which the values of numerical variables are transformed to fit within a specific range or distribution. The goal of scaling is to improve the performance of machine learning algorithms by making the data more suitable for analysis.

Scaling is performed for several reasons:

1. To improve the accuracy of machine learning models: Some machine learning algorithms are sensitive to the scale of the input variables. Scaling the data can help ensure that all variables contribute equally to the model.

2. To speed up the training of machine learning models: Scaling can help improve the convergence rate of some machine learning algorithms, which can reduce the time needed to train the models

3. To make the data more interpretable: Scaling can help standardize the units of measurement used for different variables, which can make the data more easily interpretable.

There are two common types of scaling: normalized scaling and standardized scaling.

Normalized scaling involves transforming the values of a variable to fit within a specified range, typically between 0 and 1.

The formula for normalized scaling is: x_norm = (x - min(x)) / (max(x) - min(x))

Standardized scaling involves transforming the values of a variable to have a mean of 0 and a standard deviation of 1.

The formula for standardized scaling is: x_stand = (x - mean(x)) / std(x)

scaling is a data pre-processing technique used to transform the values of numerical variables to fit within a specific range or distribution. Normalized scaling and standardized scaling are two common types of scaling, each with their own benefits and use cases.

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 of multicollinearity, which is the degree to which independent variables in a multiple regression model are correlated with each other. VIF is calculated for each independent variable, and a high VIF value indicates that the variable may be affected by multicollinearity, which can lead to unstable and unreliable regression coefficients.

Sometimes, the value of VIF can be infinite. This occurs when there is perfect multicollinearity between an independent variable and the other variables in the model. Perfect multicollinearity means that one or more independent variables in the model can be perfectly predicted by a linear combination of the other variables. In other words, there is a perfect correlation between two or more variables, which makes it impossible to estimate the regression coefficients accurately.

Perfect multicollinearity can occur due to several reasons, such as:

1. Including a variable that is a linear combination of other variables in the model.

2. Using dummy variables to represent categorical variables without leaving out a reference category.

3. Using variables that are almost identical or measuring the same thing.

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 to check if a set of data follows a particular distribution, such as the normal distribution. It compares the distribution of the data to the expected distribution by plotting the quantiles of the data against the quantiles of the theoretical distribution.

The use of Q-Q plots in linear regression is to check the normality assumption of the residuals. In linear regression, the residuals are the difference between the actual values of the dependent variable and the predicted values based on the regression equation. The normality assumption of the residuals is important because if the residuals are not normally distributed, it can affect the validity of the regression coefficients and lead to biased or inefficient estimates.

To create a Q-Q plot in linear regression, the residuals are first calculated for each observation. Then, the ordered residuals are plotted against the quantiles of the standard normal distribution. If the residuals are normally distributed, the points on the Q-Q plot will fall on a straight line. If the residuals deviate from a straight line, it indicates that the normality assumption may not hold, and further investigation may be needed.