Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Holidays - As people tend to take more bike rides during holidays.

Temperature - Higher temperature can lead to an increase in the demand for bikes.

Humidity - Humidity can affect the demand for bikes, as high humidity levels can make cycling uncomfortable.

Wind speed - Higher wind speed can deter people from using bikes, so it is important to consider this variable.

Season - Different seasons can affect the demand for bikes in various ways, so it is crucial to include this variable in the analysis.

Months - Specific months such as January, July, September, November, and December can impact bike usage.

Year - The year 2019 could have a bearing on the demand for bikes.

Day of the week - Sunday could have a different demand pattern compared to other days of the week.

Weather situation - Weather conditions such as light snow or mist and cloudy skies can affect the demand for bikes, and therefore it is important to factor this into the analysis.

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

It is important in order to achieve k-1 dummy variables as it can be used to delete extra column while creating dummy variables.

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

atemp and temp

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

We validate the assumptions of Linear Regression by Residual Analysis by plotting the displot to see if it is normal distribution or not. If it has mean=0 then it is 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 contributing significantly towards explaining the demand of the shared bikes

1. Temperature - Higher temperature can lead to an increase in the demand for bikes.

2. Weather situation - Weather conditions such as light snow or mist and cloudy skies can affect the demand for bikes, and therefore it is important to factor this into the analysis.

3. Year - The year 2019 could have a bearing on the demand for bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical algorithm used to model the relationship between a dependent variable and one or more independent variables. The goal of linear regression is to find the linear equation that best predicts the value of the dependent variable based on the values of the independent variables.

Here are the steps involved in the linear regression algorithm:

Data preparation: The first step in the linear regression algorithm is to collect and prepare the data. This includes identifying the dependent variable and one or more independent variables, and cleaning the data to remove any missing values or outliers.

Model training: The algorithm trains a linear regression model on the data. This involves finding the values of m and b that minimize the difference between the predicted values of y and the actual values of y in the training data. The algorithm uses a cost function such as mean squared error to measure the difference between the predicted and actual values.

Model evaluation: After the model is trained, the algorithm evaluates its performance on a separate set of test data. This involves calculating metrics such as mean squared error, R-squared, and adjusted R-squared to determine how well the model fits the data and how well it generalizes to new data.

Model prediction: Once the model is evaluated, it can be used to make predictions on new data. This involves plugging in new values of x into the equation y = mx + b to predict the corresponding values of y.

2. Explain the Anscombe’s quartet in detail.

Anscombe's quartet is a set of four datasets that have the same statistical properties, yet have very different distributions and patterns when graphed.

Each of the four datasets consists of 11 (x,y) pairs. When plotted, all four datasets have the same mean, variance, correlation coefficient, and regression line. However, they have very different distributions and patterns.

The four datasets:

The first dataset is a simple linear relationship between x and y, where y = 3 + 0.5x. The plot of this dataset shows a clear linear relationship between x and y.

The second dataset is also a linear relationship between x and y, but with one outlier. The plot of this dataset shows the importance of identifying and handling outliers in data analysis.

The third dataset is a non-linear relationship between x and y, where y = 1 + 0.5x + 0.1x^2. The plot of this dataset shows that even non-linear relationships can have a strong correlation coefficient and regression line.

The fourth dataset is a complete outlier to the other three datasets. It consists of two clusters of data, where one cluster has a linear relationship between x and y, and the other cluster has no relationship between x and y. The plot of this dataset shows that relying solely on summary statistics can be misleading when analysing data.

3. What is Pearson’s R?

Pearson's R, also known as Pearson's correlation coefficient or Pearson's product-moment correlation coefficient, is a statistical measure that measures the linear relationship between two continuous variables. It is denoted by the symbol "r" and takes on values between -1 and 1, where -1 represents a perfect negative correlation, 0 represents no correlation, and 1 represents a perfect positive correlation.

Pearson's R is commonly used in many fields, including statistics, social sciences, and natural sciences, to determine whether and to what degree two variables are related to each other. It is calculated by dividing the covariance of the two variables by the product of their standard deviations, as shown in the following formula:

r = Cov(X,Y) / (SD(X) * SD(Y))

where Cov(X,Y) is the covariance of X and Y, and SD(X) and SD(Y) are the standard deviations of X and Y, respectively.

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

   Scaling is a data pre-processing technique in which the values of the features or variables in a dataset are transformed to a specific range or distribution.
   There are several reasons why scaling is performed:

- To improve model performance: Scaling can improve the performance of many machine learning algorithms, such as distance-based algorithms like k-nearest neighbors and clustering algorithms.
- To speed up convergence: Some optimization algorithms, such as gradient descent, converge faster when the features are on a similar scale.
- To improve interpretability: Scaling can make it easier to compare the relative importance of different features in the dataset.

The main difference between normalized scaling and standardized scaling is that normalized scaling preserves the shape of the distribution of the features, while standardized scaling transforms the features to have a normal distribution.

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

   The VIF (Variance Inflation Factor) is a statistical measure used to detect the presence of multicollinearity in a regression analysis.

   In some cases, the VIF can be infinite. This can happen when the coefficient of determination (R^2) is equal to 1, which means that the predictor variable can be perfectly predicted from the other predictor variables. When R^2 is equal to 1, the denominator in the VIF formula becomes 0, resulting in an infinite VIF value.

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 assess the normality of a distribution by comparing the observed distribution of a variable to a theoretical normal distribution. In a Q-Q plot, the observed quantiles of the variable are plotted against the expected quantiles of a normal distribution, and if the points on the plot fall on a straight line, it indicates that the variable is approximately normally distributed.

   Q-Q plot is an important tool in linear regression analysis as it can help to assess the validity of the normality assumption, which is a crucial assumption in many statistical tests and analysis.

If the residuals are not normally distributed, appropriate transformations or alternative regression models may need to be considered to improve the accuracy and reliability of the results.