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)**

Answer –

Analysis on the categorical variables is done using the bar and box plot. Observations are below:

- Season - Fall season has more booking for rental bikes
- Year - Booking has increased in year 2019 as compared to 2018. Shows good trend in general for bike sharing company.
- Month - With in an year , July month has attracted more rentals. But in general most bookings are done from May- Oct.
- Holiday - Attracted more bookings even on non-holiday.
- Weekday - Almost same distribution of rentals across whole week.
- Workingday - Rental bookings are almost same on working and non-working day.
- Weathersit - Clearly more bookings on Clear/FewClouds and Mist Cloud weather days. Lowest booking on Light Snow/Thunderstorm days which is obvious.

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

Answer – It helps in reducing the extra column created 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.

Basic idea is if one variable is not A and not B, then it is C for sure. So, we do not need 3rd variable to identify C.

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

Answer – 'temp' variable 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)**

Answer – I have validated the assumption of Linear Regression as mentioned below:

- Normality of error terms: Check if the error terms/residuals are also normally distribu ted. Verified the same using distplot of (y_train - y_train_pred).

- Multicollinearity check: By checking the VIF values and pair plot of numerical variables.
- Linear relationship validation: Linearity should be visible among variables.
- Homoscedasticity for error terms: Verified by plotting residuals against predicted values and verifying there should not be any visible pattern in residual values.

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

Answer – Top 3 features variables are :

1. 'temp' : coefficient as 0.5499
2. weathersit : Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattere d clouds (coefficient as -0.2871)
3. 'year' (coefficient as 0.2331)

General Subjective Questions:

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

Answer:

Linear regression is a statistical method used to find the linear relationship between a dependent variable (also called the response variable or outcome) and one or more independent variables (also called predictors or features). The goal of linear regression is to create a linear equation that can predict the value of the dependent variable based on the values of the independent variables.

There are two types of linear regression:

1. Simple linear regression: In simple linear regression, there is only one independent variable and one dependent variable. The linear equation is of the form y = mx + b, where y is the dependent variable, x is the independent variable, m is the slope of the line, and b is the intercept.
2. Multiple linear regression: In multiple linear regression, there are multiple independent variables and one dependent variable. The linear equation is of the form y = b0 + b1x1 + b2x2 + ... + bnxn, where y is the dependent variable, x1, x2, ..., xn are the independent variables, and b0, b1, b2, ..., bn are the coefficients.

The following are the steps involved in the linear regression algorithm:

1. Data collection: Collect the data for both the dependent and independent variables.
2. Data preprocessing: Clean the data by removing any missing values, outliers, or errors. Also, scale the data if necessary.
3. Data splitting: Split the data into training and testing sets.
4. Model training: Train the linear regression model on the training data using the least squares method. The least squares method minimizes the sum of the squared differences between the predicted and actual values.
5. Model evaluation: Evaluate the model's performance on the testing data using metrics such as mean squared error, root mean squared error, or R-squared.
6. Prediction: Use the trained model to make predictions on new data.
7. Model interpretation: Interpret the coefficients of the model to understand the relationship between the independent and dependent variables.

Linear regression is a widely used algorithm in machine learning, data analysis, and statistics. It is simple, fast, and provides useful insights into the relationship between variables. However, it assumes that the relationship between the variables is linear and does not account for non-linear relationships or interactions between variables.

Assumptions -

The following are some assumptions about dataset that is made by Linear Regression model –

- Multi-collinearity – Linear regression model assumes that there is very little or no multi-collinearity in the data.
- Auto-correlation – Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data.
- Relationship between variables – Linear regression model assumes that the relationship between target and feature variables must be linear.

- Normality of error terms – Error terms should be normally distributed
- Homoscedasticity – There should be no visible pattern in residual values.

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

Answer:

Anscombe's quartet is a set of four datasets that have identical statistical properties, but very different visual representations. The quartet was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data.

Each of the four datasets has 11 pairs of x and y values. The datasets have the same mean, variance, and correlation, but they differ in their distribution, skewness, and outliers.

Here are the details of each dataset in Anscombe's quartet:

1. Dataset I: This dataset is a simple linear relationship between x and y. The x values range from 4 to 21, and the y values range from 4.26 to 10.84. The correlation between x and y is 0.816.
2. Dataset II: This dataset is a non-linear relationship between x and y. The x values range from 4 to 21, and the y values range from 3.10 to 12.74. The correlation between x and y is 0.816.
3. Dataset III: This dataset is a linear relationship between x and y, but with an outlier. The x values range from 4 to 21, and the y values range from 5.39 to 12.50. The correlation between x and y is 0.816.
4. Dataset IV: This dataset is a perfect example of how misleading summary statistics can be. It consists of three clusters of data points, where two clusters have a linear relationship and one has no relationship. The x values range from 8 to 19, and the y values range from 5.25 to 12.5. The correlation between x and y is 0.816, the same as the other three datasets.

The importance of Anscombe's quartet lies in its ability to show that visualizing data is essential to understanding its true nature. Even though the four datasets have identical statistical properties, their visual representations are very different. Dataset IV, for example, demonstrates how summary statistics can be misleading when they are used without visual analysis. Therefore, Anscombe's quartet is often used as a teaching tool to emphasize the importance of data visualization in statistical analysis.

3. What is Pearson's R?(3 marks)

Answer:

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that describes the strength and direction of the linear relationship between two continuous variables. It is denoted by the symbol "r".

The value of r ranges between -1 and 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation between the two variables.

Pearson's R is calculated by dividing the covariance of the two variables by the product of their standard deviations. Mathematically, the formula for Pearson's R is:

$r = Cov(x, y) / (SD(x) * SD(y))$

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.

Pearson's R is commonly used in data analysis to determine the degree of association between two variables. It is used to test hypotheses, evaluate the strength of the relationship, and to identify outliers or influential observations.
Pearson's R assumes that the relationship between the variables is linear and that the variables are normally distributed. It may not be an appropriate measure of correlation for variables that do not have a linear relationship or have extreme values. In such cases, other correlation measures such as Spearman's rank correlation or Kendall's tau may be more appropriate.

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

Answer:

Scaling is the process of transforming the values of variables to a common scale. It is performed to bring all the variables to a similar range so that they can be compared and analyzed effectively.
Scaling is performed for various reasons, such as:

1. To avoid bias due to differences in units of measurement: When different variables are measured in different units, the magnitude of the values may differ, leading to bias in the analysis. Scaling the variables to a common scale eliminates this bias.
2. To improve the performance of machine learning models: Many machine learning algorithms use distance-based measures, such as Euclidean distance, to compare variables. Scaling the variables to a common range can help improve the performance of these algorithms.
3. To make the variables comparable: Scaling can make variables with different ranges and units directly comparable, enabling meaningful analysis and interpretation.

Normalized scaling and standardized scaling are two common types of scaling methods:

1. Normalized scaling: In normalized scaling, the values of the variables are transformed so that they fall between 0 and 1. This scaling method is useful when the range of values for a variable is not known or when the range is very large. The formula for normalized scaling is:

   x_normalized = (x - min(x)) / (max(x) - min(x))
   where x is the original value of the variable, min(x) and max(x) are the minimum and maximum values of the variable, respectively, and x_normalized is the scaled value of the variable.
2. Standardized scaling: In standardized scaling, the values of the variables are transformed so that they have a mean of 0 and a standard deviation of 1. This scaling method is useful when the mean and standard deviation of the variable are important. The formula for standardized scaling is:

   x_standardized = (x - mean(x)) / sd(x)
   where x is the original value of the variable, mean(x) is the mean of the variable, sd(x) is the standard deviation of the variable, and x_standardized is the scaled value of the variable.

The main difference between normalized scaling and standardized scaling is that normalized scaling preserves the original distribution of the variable, while standardized scaling transforms the variable to have a mean of 0 and a standard deviation of 1.

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

Answer:
The Variance Inflation Factor (VIF) is a measure of multicollinearity, which is the extent to which the predictor variables in a regression model are correlated with each other. A high VIF value indicates a high degree of multicollinearity, which can lead to problems in the regression model such as unreliable coefficient estimates, reduced statistical power, and difficulties in interpreting the results.

Sometimes, the value of VIF can be infinite. This happens when one of the predictor variables in the model is a perfect linear combination of one or more of the other predictor variables. In other words, when there is perfect multicollinearity in the model.
Perfect multicollinearity means that one or more predictor variables can be expressed as a linear combination of the other predictor variables with a coefficient of 1 or -1. This can happen when the variables are not properly scaled or when there are errors in the data.

When there is perfect multicollinearity, the VIF for the affected predictor variable becomes infinite because the formula for VIF involves dividing the variance of the predictor variable by the residual variance, which becomes zero in the presence of perfect multicollinearity. This results in a division by zero error and an infinite VIF value.

In such cases, it is necessary to identify and remove the variable causing the multicollinearity, or to use other techniques such as principal component analysis or ridge regression to address the issue of multicollinearity in the model.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**
Answer:
A Q-Q plot (quantile-quantile plot) is a graphical tool used to compare the distribution of a sample of data to a theoretical distribution. The Q-Q plot plots the quantiles of the sample data against the quantiles of the theoretical distribution on a scatter plot.

In linear regression, Q-Q plots are used to check the assumption of normality of the residuals, which is one of the key assumptions of linear regression. The residuals are the differences between the actual values and the predicted values of the response variable.

The Q-Q plot helps to visually assess whether the distribution of the residuals is approximately normal or not. If the residuals are normally distributed, the points in the Q-Q plot should fall approximately along a straight line. Any deviation from a straight line indicates that the residuals are not normally distributed.

The importance of a Q-Q plot in linear regression lies in its ability to help identify potential problems with the assumptions of the regression model. If the Q-Q plot shows significant deviations from a straight line, it suggests that the residuals are not normally distributed and may violate the assumption of normality. This could lead to biased estimates of the regression coefficients, incorrect p-values, and incorrect confidence intervals.
In such cases, the regression model may need to be modified to address the issue of non-normality. For example, the transformation of the response variable or the inclusion of additional predictor variables may help improve the normality of the residuals.

Overall, Q-Q plots are an important tool in linear regression for assessing the normality of residuals and checking the assumptions of the regression model. They provide a quick and visual

way to identify potential problems and help ensure the validity and reliability of the regression analysis.