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?

Answer:

Analysis on the categorical variables is done using the sweetviz tool, scatter, box and bar 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?

Answer:

drop_first = True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. Syntax - drop_first: bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level. Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not A and B, then It is obvious C. So we do not need 3rd variable to identify the C

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

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?

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?

Answer:

Top 3 features variables are :

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

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

Linear regression is a supervised learning algorithm used to predict the relationship between a dependent variable and one or more independent variables. In simple linear regression, there is only one independent variable, while in multiple linear regression, there are two or more independent variables.

The goal of linear regression is to find the best-fit line that can describe the relationship between the dependent variable and independent variable(s). The best-fit line is a line that minimizes the distance between the predicted values and actual values. This line is also known as the regression line or the line of best fit.

The equation of a straight line is given by:

$$y = mx + c$$

where y is the dependent variable, x is the independent variable, m is the slope of the line, and c is the intercept.

To find the slope and intercept of the best-fit line in linear regression, we use the method of least squares. This method calculates the sum of the squared differences between the predicted and actual values. The line with the minimum sum of squared differences is the best-fit line.

In simple linear regression, the slope m and intercept c can be calculated as follows:

$$m = (n\sum xy - \sum x\sum y) / (n\sum x^2 - (\sum x)^2)$$

$$c = (\sum y - m\sum x) / n$$

where n is the number of data points, x and y are the independent and dependent variables, and ∑ represents the sum of all the data points.

In multiple linear regression, the equation of the best-fit line is given by:

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_r x_r$$

where y is the dependent variable, x_1, x_2, ..., x_r are the independent variables, and β_o, β_1, β_2, ..., β_r are the coefficients of the best-fit line.

To calculate the coefficients in multiple linear regression, we use the method of least squares. The coefficients can be calculated using matrix operations or through gradient descent.

Once the coefficients are calculated, the model can be used to make predictions on new data points.

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet is a set of four datasets that have the same statistical properties but are visually very different. It was created by the statistician Francis Anscombe in 1973 to illustrate the importance of visualizing data before making any conclusions based on statistical analysis.

Each dataset in Anscombe's quartet consists of 11 x,y pairs of values. The mean, variance, and correlation coefficients for each dataset are identical, but the scatterplots and regression lines are dramatically different, illustrating the importance of visualizing data.

The four datasets are as follows:

1. Dataset I - This dataset shows a simple linear relationship between x and y. The relationship is strong, with a correlation coefficient of 0.816. The regression line is a good fit for the data.
2. Dataset II - This dataset has a similar linear relationship as dataset I, but with one outlier that dramatically affects the regression line. The outlier has an extreme x value, but a y value that is similar to the other data points. The regression line is no longer a good fit for the data.
3. Dataset III - This dataset has a non-linear relationship between x and y. The relationship is still strong, with a correlation coefficient of 0.816, but a linear regression line is not appropriate for the data.
4. Dataset IV - This dataset has a strong relationship between x and y, but it is driven by one outlier. The relationship between the other data points is weak, with a correlation coefficient of 0.317. The regression line is not a good fit for the data.

The main lesson from Anscombe's quartet is that it is important to visualize data before making any conclusions based on statistical analysis. Even if the statistical properties of two datasets are identical, the visual properties may be very different, leading to different conclusions about the relationship between the variables. Therefore, it is important to plot data and examine the relationships between variables to gain a deeper understanding of the data.

3. What is Pearson's R?

Answer:

Pearson's R, also known as Pearson correlation coefficient, is a statistical measure that represents the strength and direction of the linear relationship between two variables. It is denoted by the symbol 'r' and ranges from -1 to +1.

A value of +1 indicates a perfect positive linear relationship, where an increase in one variable corresponds to an increase in the other variable. A value of -1 indicates a perfect negative linear relationship, where an increase in one variable corresponds to a decrease in the other variable. A value of 0 indicates no linear relationship between the variables.

Pearson's R is calculated by dividing the covariance of the two variables by the product of their standard deviations. The formula for calculating Pearson's R is as follows:

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 sqrt represents the square root function.

Pearson's R is commonly used in various fields such as psychology, economics, and social sciences to analyze the relationship between variables. However, it should be noted that Pearson's R only measures linear relationships between variables and is not suitable for identifying non-linear relationships.

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

Answer:

Scaling is a data preprocessing technique used to transform the numerical features of a dataset to a specific range or distribution. The main objective of scaling is to improve the performance and accuracy of machine learning algorithms by ensuring that the features have similar scales and are not biased towards any particular feature.

There are various scaling techniques, but two of the most common techniques are normalized scaling and standardized scaling.

1. Normalized Scaling: Normalization is a technique that scales the features to a range of 0 to 1. It is achieved by subtracting the minimum value of the feature and then dividing it by the range of the feature. The formula for normalized scaling is as follows:

x_normalized = (x - min(x)) / (max(x) - min(x))

Normalization ensures that all the features have the same scale and distribution, making them equally important to the model. Normalization is suitable for algorithms that assume a Gaussian distribution in the input features, such as linear regression, logistic regression, and neural networks.

2. Standardized Scaling: Standardization is a technique that scales the features to have a mean of 0 and a standard deviation of 1. It is achieved by subtracting the mean of the feature and then dividing it by the standard deviation of the feature. The formula for standardized scaling is as follows:

x_standardized = (x - mean(x)) / std(x)

Standardization transforms the features to have the same scale and distribution, making them equally important to the model. Standardization is suitable for algorithms that do not assume a Gaussian distribution in the input features, such as K-Nearest Neighbors, Support Vector Machines, and Gradient Boosting Machines.

In summary, scaling is performed to ensure that the numerical features in a dataset are transformed to a specific range or distribution. Normalization scales the features to a range of 0 to 1, while standardization scales the features to have a mean of 0 and a standard deviation of 1. The choice of scaling technique depends on the requirements of the machine learning algorithm being used.

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

Answer:

VIF, or variance inflation factor, is a measure of multicollinearity in a linear regression model. It quantifies the extent to which the variance of the estimated regression coefficients is increased due to collinearity among the predictor variables.

The formula for VIF is:

VIF = 1 / (1 - R^2)

where R^2 is the coefficient of determination of a regression model with the predictor variable of interest as the response variable, and all the other predictor variables as predictors.

When the value of VIF is infinite, it means that the coefficient of determination, R^2, is equal to 1. This indicates that the predictor variable of interest can be perfectly predicted by a linear combination of the other predictor variables in the model, which is known as perfect multicollinearity.

Perfect multicollinearity can occur when there is a linear relationship between two or more predictor variables in the model, or when one of the predictor variables is a linear combination of other predictor variables in the model. In either case, it results in a situation where the model cannot distinguish the individual effects of the predictor variables on the response variable, and the regression coefficients become unstable and unreliable.

In practical terms, when VIF is infinite, it is an indication that one or more of the predictor variables in the model should be removed or combined with other predictor variables to reduce multicollinearity and improve the stability and accuracy of the regression model.

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

Answer:

A Q-Q plot, short for quantile-quantile plot, is a graphical method used to assess whether the distribution of a dataset is approximately normal or not. It compares the quantiles of the data to the quantiles of a theoretical distribution, typically the standard normal distribution.

The Q-Q plot is created by plotting the sorted data against the expected values from the theoretical distribution. If the data follows a normal distribution, the points in the Q-Q plot should form a straight line. However, if the data deviates from a normal distribution, the points will deviate from the straight line, indicating that the distribution is not normal.

In linear regression, Q-Q plots are used to assess the assumption of normality for the residuals of a regression model. Residuals are the differences between the observed values and the predicted values from the regression model. The assumption of normality for the residuals is important because it ensures that the statistical inference from the model, such as confidence intervals and hypothesis tests, are valid.

To create a Q-Q plot for the residuals of a linear regression model, the residuals are sorted in ascending order and plotted against the expected values from the standard normal distribution. If the points in the Q-Q plot follow a straight line, it indicates that the residuals are approximately normally distributed. However, if the points deviate from the straight line, it indicates that the residuals are not normally distributed, and the regression model assumptions may be violated.

In summary, a Q-Q plot is a graphical method used to assess the normality of a dataset. In linear regression, Q-Q plots are used to assess the normality assumption of the residuals, which is an important assumption for the validity of the statistical inference from the model.