General Subjective Questions

1) Explain the linear regression algorithm in detail

Answer:

Linear regression is a supervised learning algorithm used to predict a continuous output variable (also known as a dependent variable) based on one or more input variables (also known as independent variables or predictors). The algorithm works by finding the best linear relationship between the input variables and the output variable, represented by a straight line equation of the form Y = mX + c, where Y is the output variable, X is the input variable, m is the slope of the line, and b is the intercept.

The goal of linear regression is to find the values of m and b that minimize the sum of the squared errors between the predicted values and the actual values of the output variable. The squared error is calculated as the square of the difference between the predicted value and the actual value.

There are two types of linear regression: simple linear regression and multiple linear regression.

1)Simple linear regression

2)Multiple linear regression

Linear regression is a simple but powerful algorithm that can be used for a variety of tasks, such as predicting the price of a house based on its features, predicting the sales of a product based on advertising expenditure, and predicting the weather based on historical data.

2) Explain the Anscombe's quartet in detail

Answer:

Anscombe's quartet is a collection of four datasets that have the same statistical properties, but look very different when visualized graphically. These datasets were created by the statistician Francis Anscombe in 1973 to illustrate the importance of data visualization in statistical analysis.

Each of the four datasets has 11 observations and two variables: x and y. The summary statistics for the four datasets are as follows:

Despite the similarities in summary statistics, the four datasets look very different when plotted on a graph. Dataset I has a linear relationship between x and y, and the regression line fits the data well. Dataset II has a non-linear relationship between x and y, and the regression line does not fit the data well. Dataset III has an outlier that affects the regression line. Dataset IV has a perfect linear relationship between x and y, except for one outlier that affects the regression line.

The importance of Anscombe's quartet lies in its demonstration that summary statistics alone are not enough to fully understand a dataset. Graphical visualization is essential for identifying patterns, trends, and outliers that may not be apparent from summary statistics alone.

In summary, Anscombe's quartet is a set of four datasets that have the same summary statistics but exhibit very different patterns when visualized graphically. The quartet serves as a reminder of the importance of data visualization in statistical analysis.

3) What is Pearson's R?

Answer:

Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two variables, X and Y. It is a numerical value 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 calculated as the covariance between X and Y divided by the product of their standard deviations. The formula for Pearson's R is as follows:

r = (Σ[(Xi - X_mean)(Yi - Y_mean)]) / (√Σ(Xi - X_mean)^2 * √Σ(Yi - Y_mean)^2)

where Xi is the ith value of X, X_mean is the mean of X, Yi is the ith value of Y, and Y_mean is the mean of Y.

The Pearson correlation coefficient measures the degree to which the relationship between X and Y can be described by a straight line. If the relationship is not linear, Pearson's R may not accurately capture the relationship between the variables. It is also important to note that correlation does not imply causation, and a high correlation between two variables does not necessarily mean that one variable causes the other.

Pearson's R is commonly used in fields such as statistics, finance, and social sciences to analyze the relationship between two variables. It can be used to identify trends and patterns in data, and to make predictions based on the relationship between variables.

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

Answer:

Scaling is the process of transforming data so that it fits within a specific range or distribution. It is often performed on variables in a dataset to bring them to a similar scale or distribution to facilitate analysis and modeling.

Scaling is performed for several reasons:

1. To improve the performance of machine learning algorithms: Many machine learning algorithms are sensitive to the scale of the input features. Scaling can improve the performance of these algorithms by bringing all features to the same scale.

2. To compare variables: When variables in a dataset have different scales or units, it can be difficult to compare them. Scaling can help to make comparisons between variables more meaningful.

3. To reduce the impact of outliers: Scaling can help to reduce the impact of outliers in a dataset by bringing extreme values within the range of other values.

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

Normalized scaling, also known as min-max scaling, transforms the data so that it falls within a specific range, typically between 0 and 1. The formula for normalized scaling is:

X_scaled = (X - X_min) / (X_max - X_min)

where X is the original value, X_min is the minimum value in the dataset, and X_max is the maximum value in the dataset.

Standardized scaling transforms the data so that it has a mean of 0 and a standard deviation of 1. The formula for standardized scaling is:

X_scaled = (X - X_mean) / X_std

where X is the original value, X_mean is the mean of the dataset, and X_std is the standard deviation of the dataset.

The main difference between normalized scaling and standardized scaling is that normalized scaling preserves the original range of the data, while standardized scaling transforms the data to have a mean of 0 and a standard deviation of 1. Normalized scaling is typically used when the original range of the data is important, while standardized scaling is used when it is important to compare the relative position of data points within the distribution.

5) You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   Answer:
   If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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 tool used to assess the normality of a distribution by comparing the quantiles of the data with the quantiles of a theoretical normal distribution. In a Q-Q plot, the observed data is plotted on the x-axis, and the expected values of a normal distribution are plotted on the y-axis.

   The use and importance of a Q-Q plot in linear regression are as follows:

   1. Checking normality assumption: One of the important assumptions of linear regression is that the residuals should be normally distributed. A Q-Q plot can be used to check whether the residuals follow a normal distribution. If the data points on the Q-Q plot fall close to the diagonal line, it suggests that the residuals are normally distributed. If the data points deviate significantly from the line, it suggests that the residuals may not be normally distributed.

   2. Detecting outliers: Outliers in the data can have a significant impact on the regression model. A Q-Q plot can be used to identify outliers by checking for data points that deviate significantly from the diagonal line. Outliers can be a sign of data quality issues or a signal that the model needs to be adjusted.

   3. Assessing model fit: A Q-Q plot can be used to assess the goodness of fit of the regression model. If the residuals follow a normal distribution, it suggests that the model is a good fit for

the data. If the residuals deviate significantly from a normal distribution, it suggests that the model may not be an appropriate fit for the data.

In summary, a Q-Q plot is a useful tool for assessing the normality assumption of linear regression, detecting outliers, and assessing the goodness of fit of the regression model. It helps to ensure that the regression analysis is based on appropriate assumptions and that the results are reliable and meaningful.

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:
I have done analysis on categorical columns using the boxplot and bar plot. Below are the few points we can infer from the visualization –

 Fall season seems to have attracted more booking. And, in each season the booking count has increased drastically from 2018 to 2019.

 Most of the bookings has been done during the month of may, june, july, aug, sep and oct. Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year.

 Clear weather attracted more booking which seems obvious.

 Thu, Fir, Sat and Sun have more number of bookings as compared to the start of the week

When it’s not holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family.

 Booking seemed to be almost equal either on working day or non-working day.

 2019 attracted more number of booking from the previous year, which shows good progress in terms of business.

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:
Normality of error terms
o Error terms should be normally distributed
Multicollinearity check
o There should be insignificant multicollinearity among variables.
Linear relationship validation
o Linearity should be visible among variables

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Answer:
Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes –
1) temp
2)winter
3) sep