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: .

The categorical variables in our dataset has a good effect on the target variable.
1. In weekdays we have seen that there is a good count on Saturday and Sunday. Targeting customers on weekends can increase the count of rentals.
2. In Months, we have seen the rapid increase in rentals count from April to August. We can say that Summer is a good time to target more customers.
3. In Holiday variable, we can see that the holidays are very less compared to not holidays. But still holidays hold the good count of rentals. Holiday is a good day to target customers.
4. For the weather sit variable we can say that the Clear weather is a good condition to improve sales.

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

It is important to drop the first column while dummy variable creation as it helps to reduce one variable. This can reduce correlations among the dummy variables which can be helpful to our model.

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

"**temp**" is the variable that has 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: .

First, the predictor variables are linearly dependent to the target variable.the residuals are independent.The residuals have constant variance and normally distributed.

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

From my final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are Year Variable, temp and Season_spring.

```
from sklearn.feature_selection import SelectKBest,f_classif

X = X3  #independent columns
y = y_train


#apply SelectKBest class to extract top 3 best features
bestfeatures = SelectKBest(score_func=f_classif, k=3)
fit = bestfeatures.fit(X,y)

dfscores = pd.DataFrame(fit.scores_)
dfcolumns = pd.DataFrame(X.columns)

#concat two dataframes for better visualization
featureScores = pd.concat([dfcolumns,dfscores],axis=1)
featureScores.columns = ['Specs','Score']  #naming the dataframe columns


print(featureScores.nlargest(3,'Score'))  #print 3 best features
```

```
           Specs     Score
4  season_spring  1.379902
1             yr  1.291453
3           temp  1.266239
```

Best 3 features for model are temp,yr and season spring

GENERAL SUBJECTIVE QUESTIONS:

1.Explain the linear regression algorithm in detail.

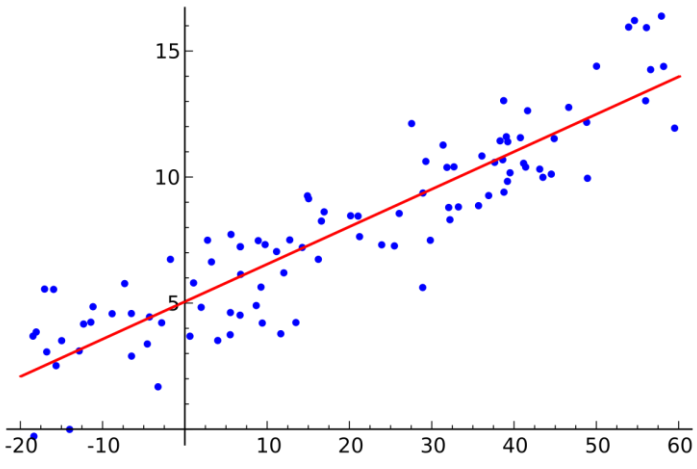
Answer:

Regression is a method of modelling a target value based on independent predictors. This method is mostly used for forecasting and finding out cause and effect relationship between variables. Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.

Linear Regression

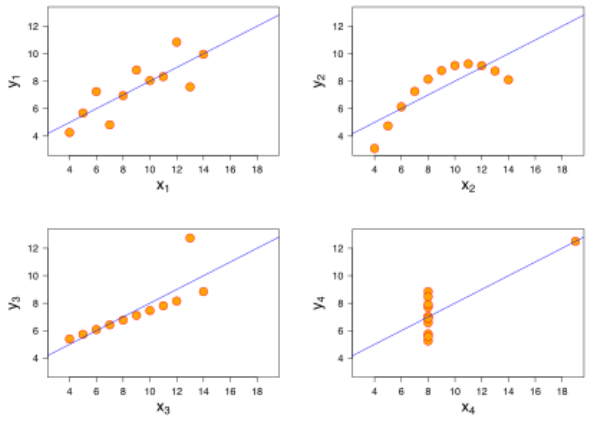
Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable. The red line in the above graph is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best. The line can be modelled based on the linear equation shown below.

y = a_0 + a_1 * x ## Linear Equation



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

Answer:



Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. This was constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties. This tells us about the importance of visualizing the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.statistical properties

3. What is Pearson's R?

Answer:

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. The nearer the scatter of points is to a straight line, the higher the strength of association between the variables. Also, it does not matter what measurement units are used. The relationship between two variables is generally considered strong when their r value is larger than 0.7. The correlation r measures the strength of the linear relationship between two quantitative variables. Pearson r. The strength of the linear relationship increases as r moves away from 0 toward - 1 or 1. It is usually used to express the correlation between two quantities. You could calculate Pearson's r to evaluate whether the two quantities are correlated. R^2 is usually used to evaluate the quality of fit of a model on data. It is used when you have two quantitative variables, and you wish to see if there is a linear relationship between those variables. Your research hypothesis would represent that by stating that one score affects the other in a certain way. The correlation is affected by the size and sign of the r.

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

Answer:

Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step. So, it is extremely important to rescale the variables so that they have a comparable scale. If we do not have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might become very annoying at the time of model evaluation. So, it is advised to use standardization or normalization so that the units of the coefficients obtained are all on the same scale. As you know, there are two common ways of rescaling. It is performed during the data pre - processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values. This is essential for machine

learning algorithms that calculate distances between data. ... Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions do not work correctly without normalization. The terms normalization and standardization are sometimes used interchangeably, but they usually refer to different things. Normalization usually means to scale a variable to have a values between 0 and 1, while standardization transforms data to have a mean of zero and a standard deviation of 1.

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

Infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. In general with the selection of all variables, and proceeds by repeatedly deselecting variables showing a high VIF.

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

The Q - Q plot, or quantile - quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. If both sets of quantiles came from the same distribution, we should see the points forming a line that is roughly straight. It is a graphical technique for determining if two data sets come from populations with a common distribution. A q - q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. The purpose of Q - Q plots is to find out if two sets of data come from the same distribution. A 45 - degree angle is plotted on the Q - Q plot; if the two data sets come from a common distribution, the points will fall on that reference line. It is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential, or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution. If the bottom end of the Q - Q plot deviates from the straight line but the upper end is not, then we can clearly say that the distribution has a longer tail to its left or simply it is left - skewed (or negatively skewed) but when we see the upper end of the Q - Q plot to deviate from the straight line and the lower end.