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

Solution :

A. From 'year' variable it is observed that in the year 2019 count of bike rentals increased and became popular than 2018.

B. From 'season' variable it is observed that fall and summer are more favourable for bike rentals than spring.

C. From 'weathersit' variable it is observed that count of bike rentals is more during clear weather.

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

Solution :

A. To reduce the effect of redundant features.

B. To avoid multicollinearity and reduce the effect on the model adversely.

C. It helps in reducing the extra column created during dummy variable creation.

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

Solution:

variable 'temp' 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?

Solution:

Following are my assumptions used to validate the linear regression after building the model on the training set

A. Error terms should be normally distributed

B. There should be insignificant multicollinearity among variables.

C. Linearity should be visible among variables.

D. There should be no visible pattern in residual values.

E. No auto-correlation

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

Solution:

Top 3 features contributing significantly towards explaining the demand of the

shared bikes –

A. temp

B. winter

C. sep

General Subjective Questions:

1. Explain the linear regression algorithm in detail?

Solution:

Linear regression is defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables.

Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

$Y = mX + c$

Here, Y is the dependent variable we are trying to predict.

X is the independent variable we are using to make predictions.

m is the slope of the regression line which represents the effect X has on Y

c is a constant, known as the Y-intercept.

If X = 0, Y would be equal to c.

The linear relationship can be positive or negative−

A. Positive Linear Relationship:

 A linear relationship will be called positive if both independent and dependent variable increases.

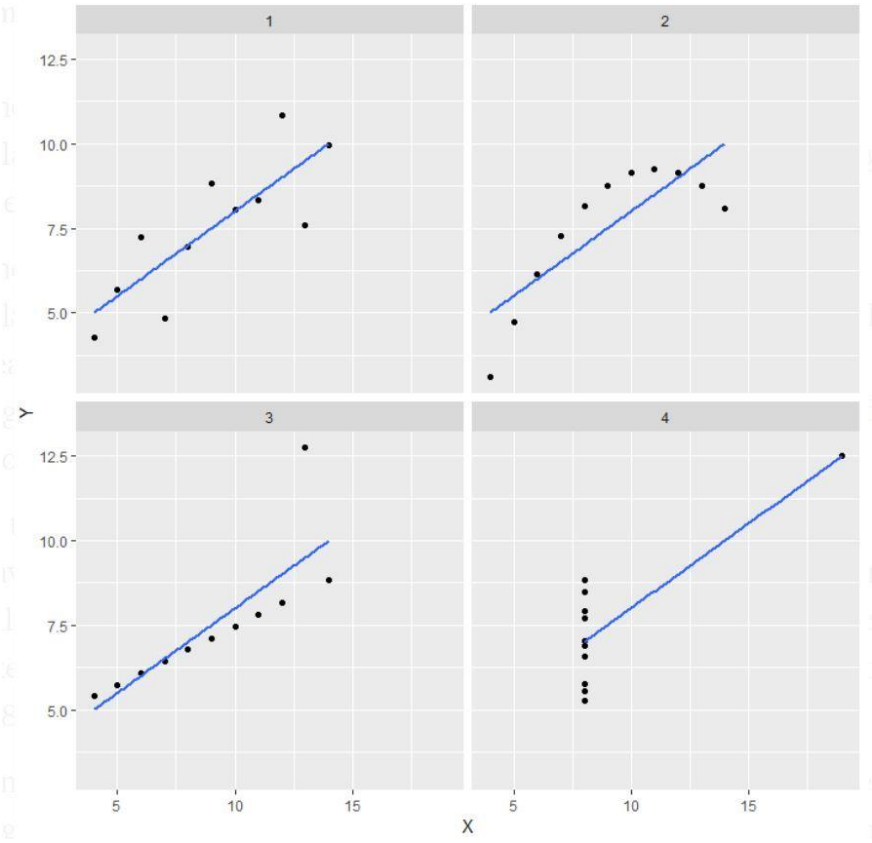
B. Negative Linear relationship:

 A linear relationship will be called positive if independent increases and dependent variable decreases.

2. Explain the Anscombe's quartet in detail?

Solution:

Anscombe's Quartet was developed by statistician Francis Anscombe. It has four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. If things change completely, and we must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.



• First Plot appears to have clean and well-fitting linear models.

• Second Plot is not distributed normally.

• Third Plot is linear, but the calculated regression is thrown off by an outlier.

• Fourth Plot shows that one outlier is enough to produce a high correlation coefficient

Application:

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3. What is Pearson's R?

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.

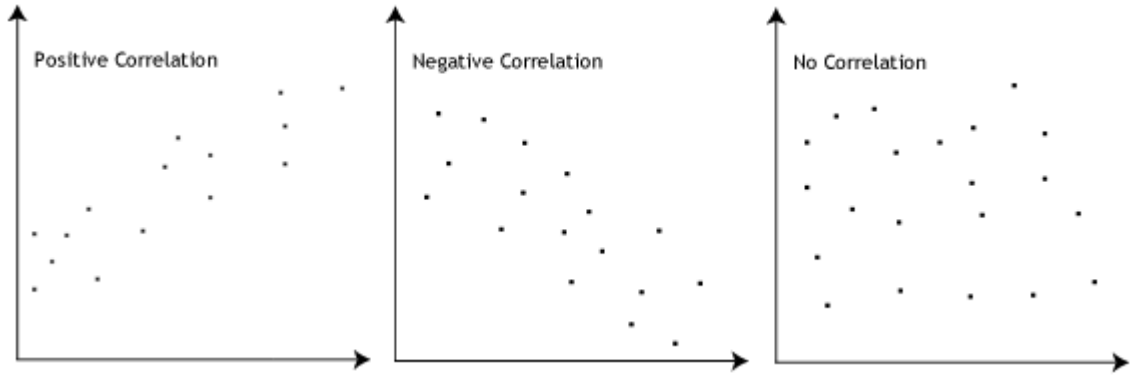
If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson correlation coefficient, r, can take a range of values from +1 to -1.

A value of 0 indicates that there is no association between the two variables.

A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable.

A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



The stronger the association of the two variables, the closer the Pearson correlation coefficient, *r*, will be to either +1 or -1 depending on whether the relationship is positive or negative, respectively.

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

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

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.

So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

Normalized scaling	Standardized scaling
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
Scikit-Learn provides a transformer called MinMaxScaler for Normalization	Scikit-Learn provides a transformer called StandardScaler for standardization
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is really affected by outliers.	It is much less affected by outliers.

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

Solution :

A. VIF = infinity for perfect correlation.

B. 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.

C. 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.

Solution:

A.This plots play a very vital role to graphically analyze and compare two probability distributions by

plotting their quantiles against each other.

B. If the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight-line y = x**.**

C. If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

D. A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.