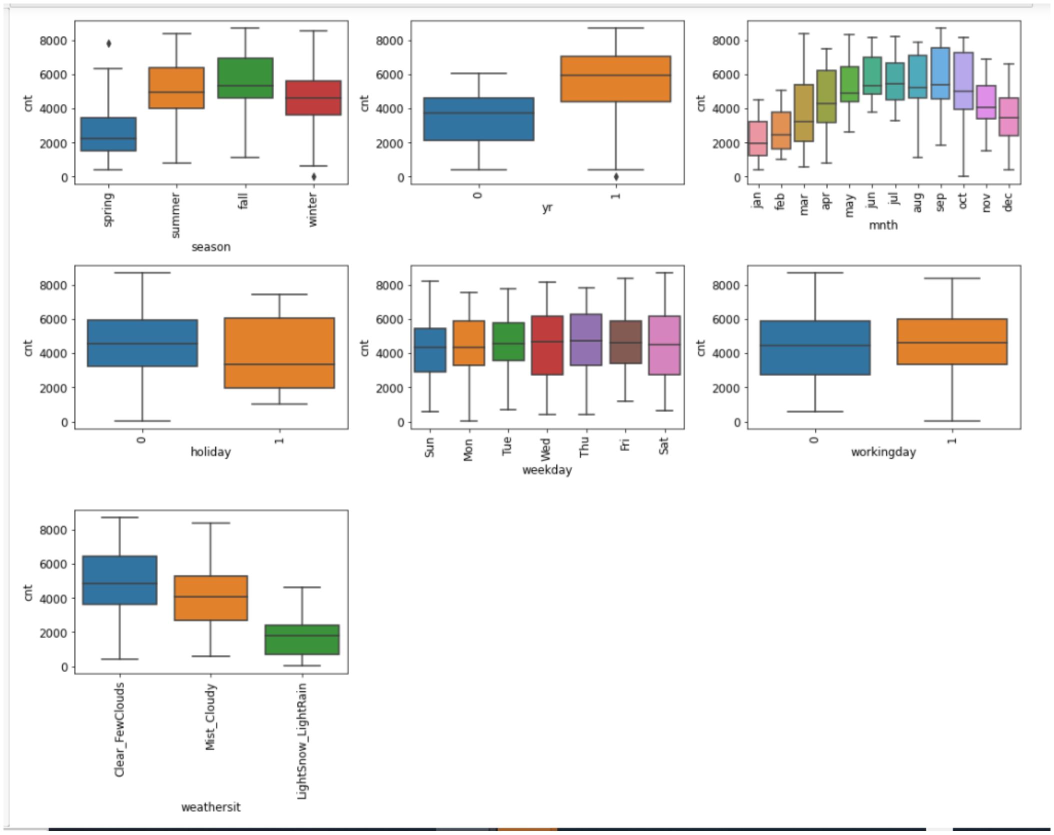
Assignment-based Subjective Questions

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:

   As a part of data visualization, we have analyzed the categorical columns [‘season’,’yr’,’mnth’,’holiday’,’weekday’,’workingday’,’weathersit’] using the boxplot. Below are the few points we inferred from the visualization



- Season: The demand of bikes is less in spring when compared with other seasons. Also fall has highest demand for rental bikes
- The demand of bikes increased in the year(yr) 2019 when compared with year 2018
- When there is a holiday, demand of bikes decreases
- The demand of bikes is almost similar throughout the weekdays (weekday).
- The bike demand is high when weather (weathershit) is clear and Few clouds. However, demand is less in case of Lightsnow and light rainfall. We do not have any data for Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog . May be the company is not operating on those days due to harsh weather conditions or there is no demand of bikes
- During Jun to Sep (mnth) the demand of bikes is high. The month of Jan has the lowest demand of bikes. The demand starts falling from November (probably due to winter and harsh weather condition)

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

Answer:

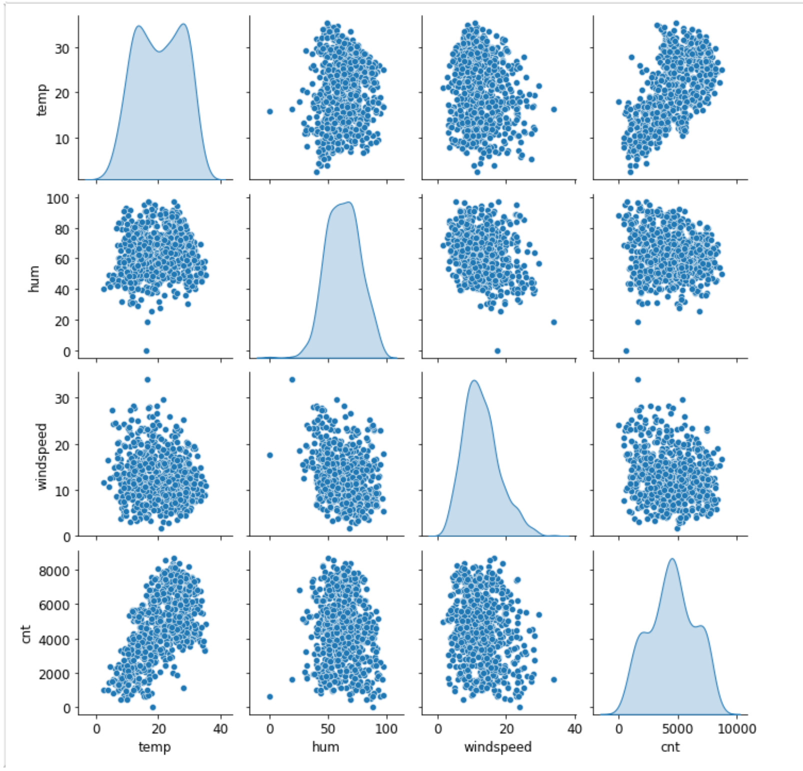
The intention behind the dummy variable is that for a categorical variable with ‘n’ levels, you create ‘n-1’ new columns each indicating whether that level exists or not using a zero or one. drop_first=True is used so that the resultant can match up n-1 levels. Hence it reduces the extra column created during dummy variable creation thereby reducing the correlation among the dummy variables. That is why it is important to use drop_first=True during dummy variable creation.

Example:

Let’s say we have 3 types of values (X,Y,Z) in a Categorical column and we want to create dummy variable for that column. If one variable is not X and Y, then It is obviously Z. So, we do not need the 3rd variable to identify Z.

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

Answer:



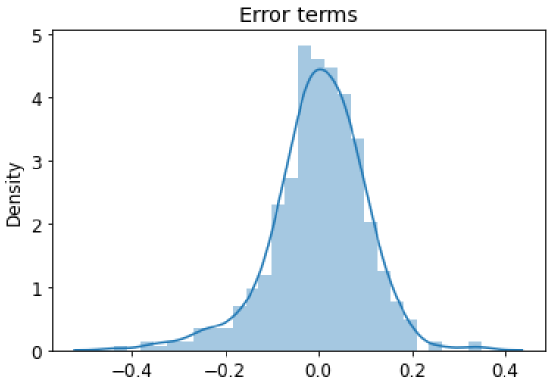
From the pair-plot we observed that ‘temp’ variable has the highest positive correlation with the target variable cnt.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

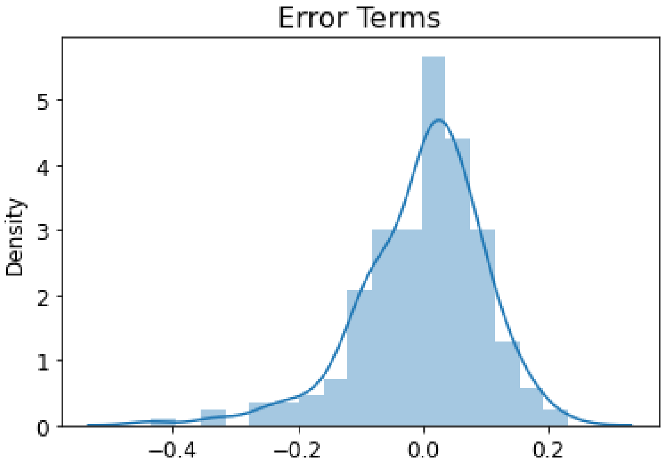
Answer:

The key assumptions of Linear Regression are as follows

1. Normality of error terms- Error terms should be normally distributed.
   We performed Residual Analysis and analyzed the distribution plot of Residue for train data set

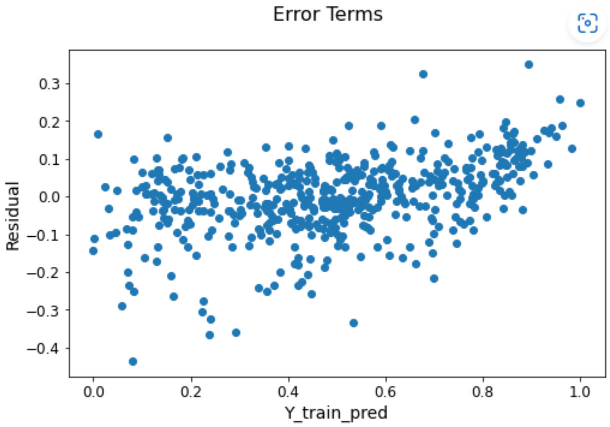


The distribution plot of error term showed the normal distribution with mean at Zero.
We performed Residual Analysis and analyzed the distribution plot of Residue for test data set

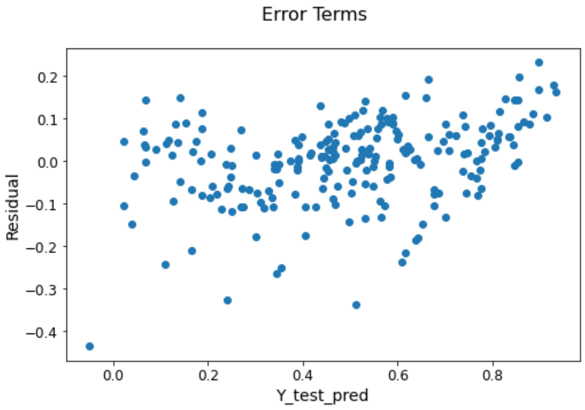


The distribution plot of error term showed the normal distribution with mean at Zero.

2. Homoscedasticity-(There should be no visible pattern in residual values)
We observed the residual plot for train data set

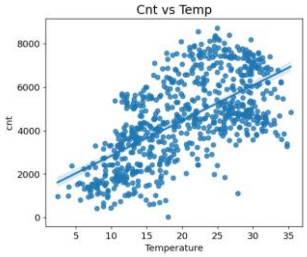


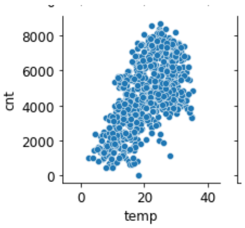
We also observed the residual plot for test data set



In both the cases we found that residual plot is reasonably random. Also the error terms satisfy the rule of having reasonably constant variance (Homoscedasticity)

3. Multi-collinearity- (Linear regression model assumes that there is very little or no multi-collinearity in the data.).
   During Model building we used the technique of detecting Multi-collinearity using VIF values. For our final model all P values are almost 0 and all VIF < 5 thereby ensuring that there is very little or no multi-collinearity in the data.
4. Auto-correlation- There is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors
   We analyzed the Durbin-Watson value for our finally arrived model (2.078). This value is in the middle range (0-4) thereby suggesting less auto correlation.
5. Linear regression model assumes that the relationship between response and feature variables must be linear
   During model building process we analyzed the pair plots to check the relationship between independent and dependent variables thereby ensuring that independent variables exhibit linear relationship with dependent variable





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:

Based on the final arrived model, the top three features that significantly contribute towards explaining the bike sharing demand are as follows

1. Temperature (temp)
2. weathersit : Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
3. year (yr)

General Subjective Questions

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

Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis. The independent variable is also the predictor or explanatory variable that remains unchanged due to the change in other variables. However, the dependent variable changes with fluctuations in the independent variable. The regression model predicts the value of the dependent variable, which is the response or outcome variable being analyzed or studied.

Mathematically the relationship can be represented with the help of following equation −

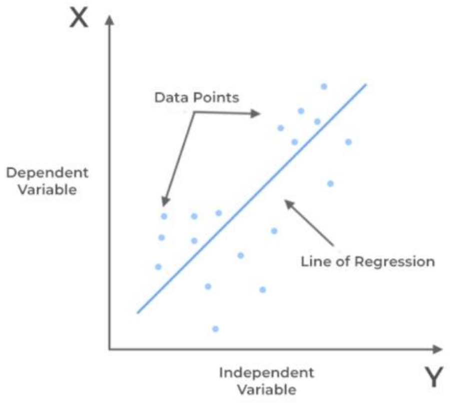
Y = mX + c

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

c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.



Best Fit Line for a Linear Regression Model

The “Line if Regression” is the best fit line for the model. Here the line is plotted for the given data points that suitably fit all the issues. The goal of linear regression is to find this best fit line seen in the above diagram.

There are two types of linear regression

- Simple Linear Regression
- Multiple Linear Regression

Simple linear regression reveals the correlation between the dependent variable and independent variable. This regression type describes the relationship strength between the variables. It is mathematically represented as shown below
Y(x)= p0 + p1*x
Here Y is the continuous value that model tries to predict
x is the independent input variable.
p0 is the y-axis intercept (bias term)
p1 is regression coefficient or the slope of the best-fit straight line of linear regression model. Regression modeling is all about finding the values of unknown parameters of the equation.

Multiple linear regression establishes the relationship between independent variables (2 or more) and the corresponding dependent variable. It is mathematically represented as follows
Y(x)= p0 + p1x1 + p2x2+ … + p(n)x(n)
The machine learning model uses the above formula and different weight values to draw lines to fit. Also, to determine the best fit line, the model evaluates different weight combinations that best fits the data and establishes strong relationship between the variables.

Key Assumptions concerning the data-

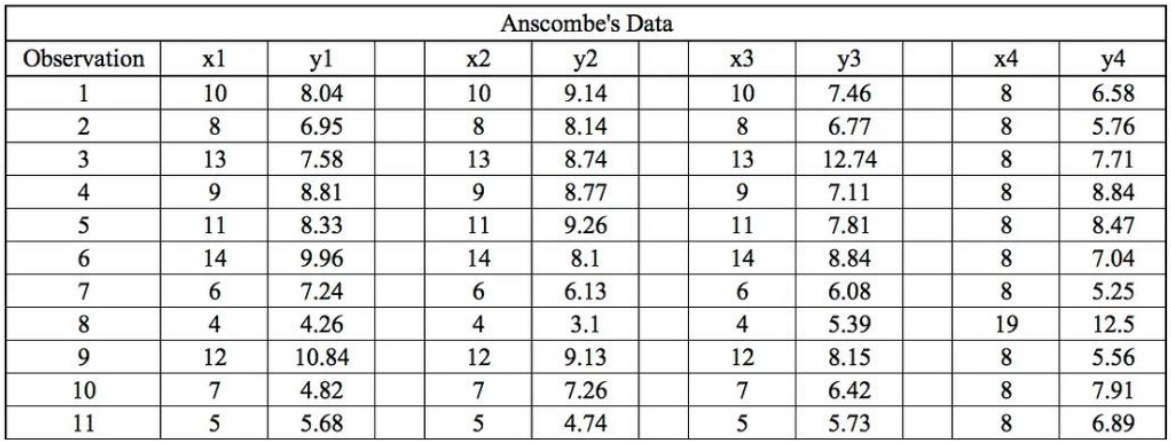
- ✓ Multi-collinearity –
  - Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.
- ✓ Auto-correlation –
  - There is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.
- ✓ Relationship between variables –
  - Linear regression model assumes that the relationship between response 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 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.

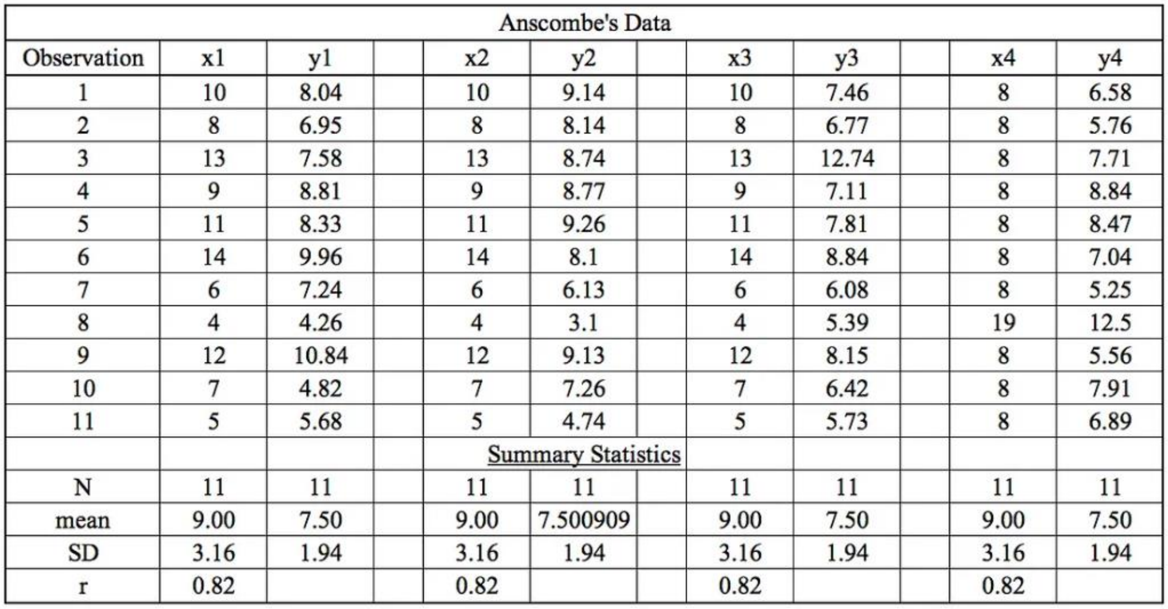
It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

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 us 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 only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets. These four plots can be defined as follows



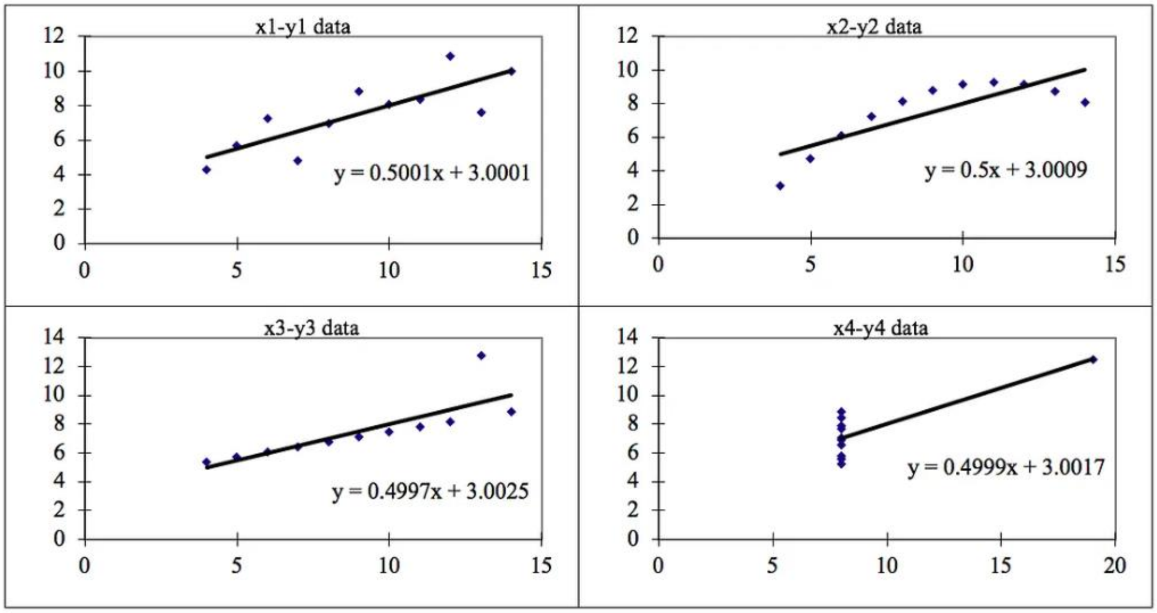
Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89

The statistical information for all these four datasets are approximately similar and can be computed as follows



Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
					Summary Statistics						
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

When these models are plotted on a scatter plot, all datasets generate a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows



The four datasets can be described as:

Dataset 1: this fits the linear regression model pretty well.
Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear

regression model
Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

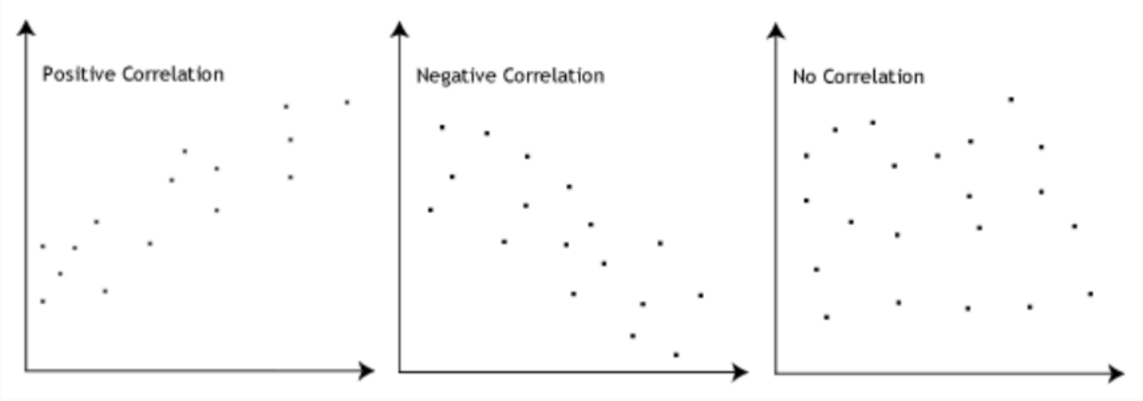
Hence, all the important features in the dataset must be visualized before implementing any machine learning algorithm on them which will help to make a good fit model.

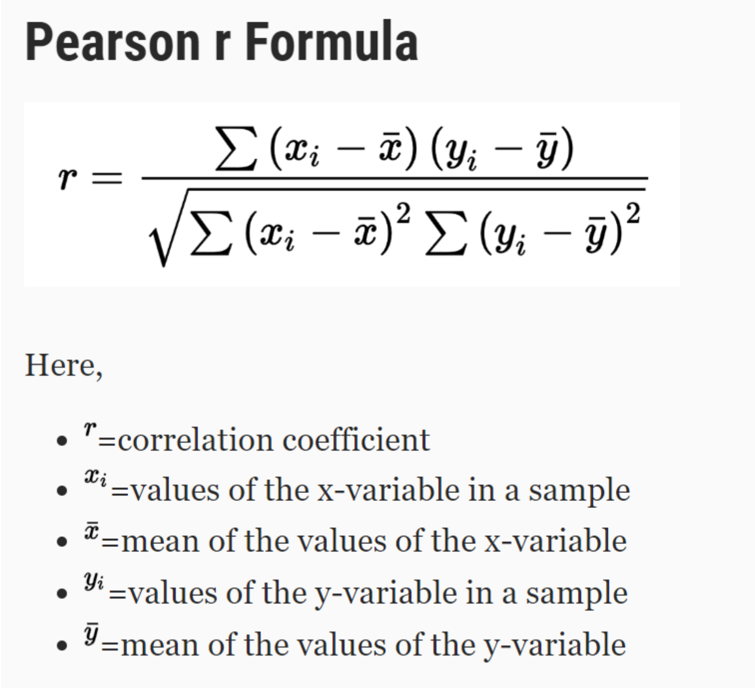
3. What is Pearson's R? (3 marks)
Answer:

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association





Pearson r Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Here,

- r=correlation coefficient
- x_i=values of the x-variable in a sample
- $\bar{x}$=mean of the values of the x-variable
- y_i=values of the y-variable in a sample
- $\bar{y}$=mean of the values of the y-variable

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

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.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done, then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we must do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization typically means rescaling the values into a range of [0,1]. Standardization typically means rescaling data to have a mean of 0 and a standard deviation of 1 (unit variance). The difference between normalized scaling and standardized scaling is provided below

S.NO.	Normalized scaling	Standardized scaling
1.	Minimum and maximum value offeatures are used for scaling	Mean and standard deviation is used for scaling.

2.	It is used when features are of differentscales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.

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

Answer:

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. Hence, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

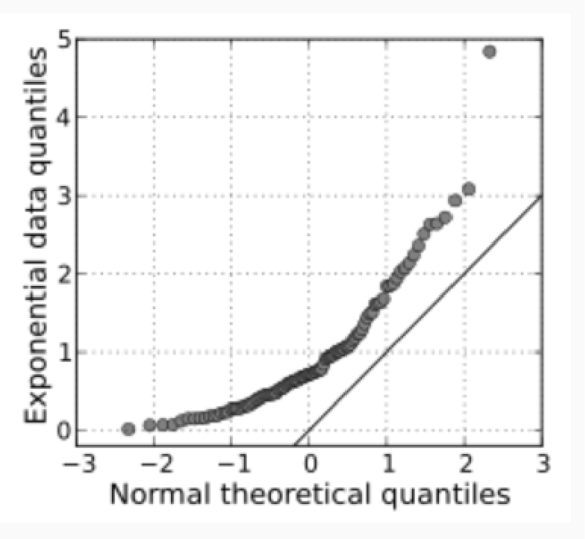
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer:

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. 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.

A Q Q plot showing the 45 degree reference line is depicted below



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.

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.