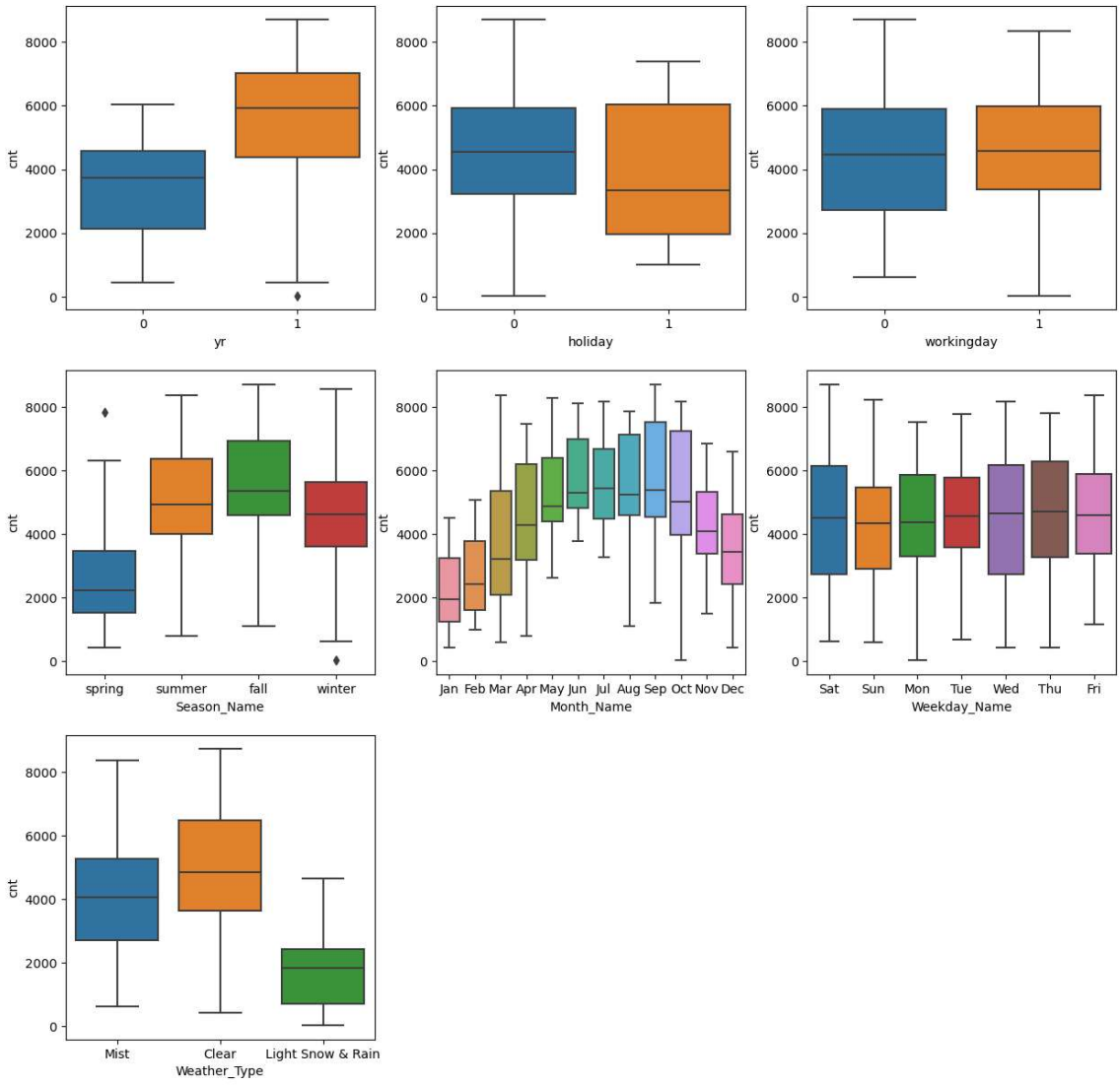
Assignment-based Subjective Questions

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:

Categorical variables were analysed as shown below:

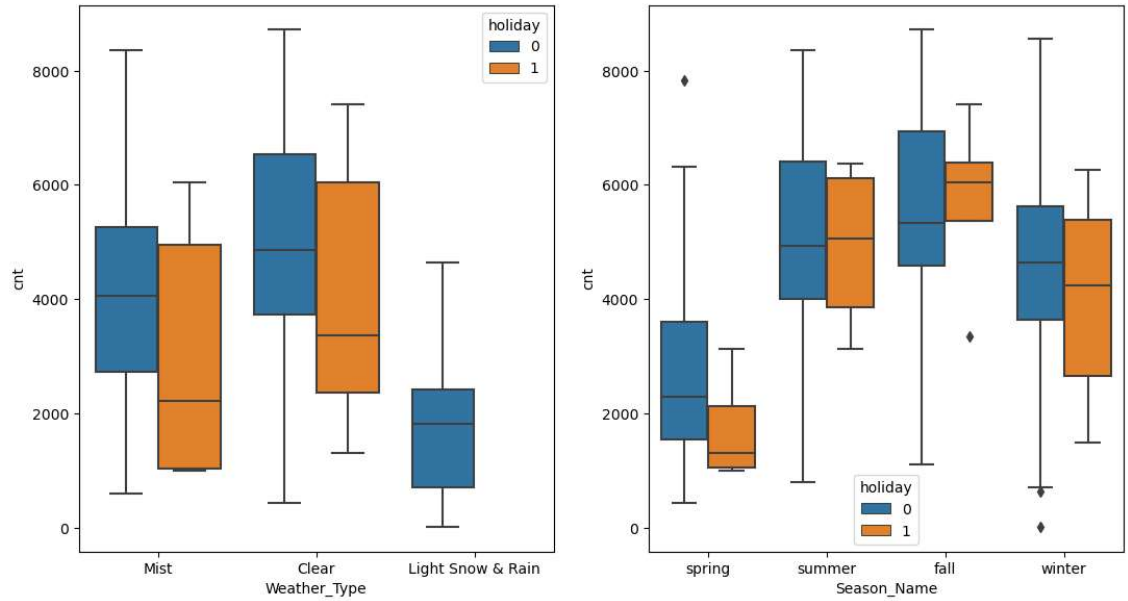


Below are the observations from boxplots:

1. When weather is clear, we see there are more users using bike sharing.
2. When we have snow or rain, the count of users reduces drastically which is also very natural.
3. From May to October, median of no of bike users are higher.
4. In month of July, we see highest median no of users.
5. Median for no of bike users is highest in season of Fall and lowest in the season of spring.
6. No of bike users has increased in year of 2019 compared to 2018.
7. Median of no of users is almost normally distributed across the days of the week.

8. Highest no of users seen on non-holidays.

Multiple variables were also analysed as shown below.



Below are the observations:

1. When weather is clear, median of bike users is high on non-holiday day compared to holiday.
2. When season is fall, median of bike users is high on holiday compared to non-holiday day.
3. When season is winter, median of bike users is high on non-holiday compared to holiday.

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

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. Dummy variables are represented by binary numbers (0,1).

For example, we have created dummy variables for '**Furnishing Status**' which have 3 different entries/categories. However, only 2 are required in the dummy variables in order to explain all 3 entries/categories. As shown below, (1,0) points to furnished, (0,1) points to semi-furnished and (0,0) points to unfurnished. Hence, we can say that only n-1 dummy variables can represent n values.

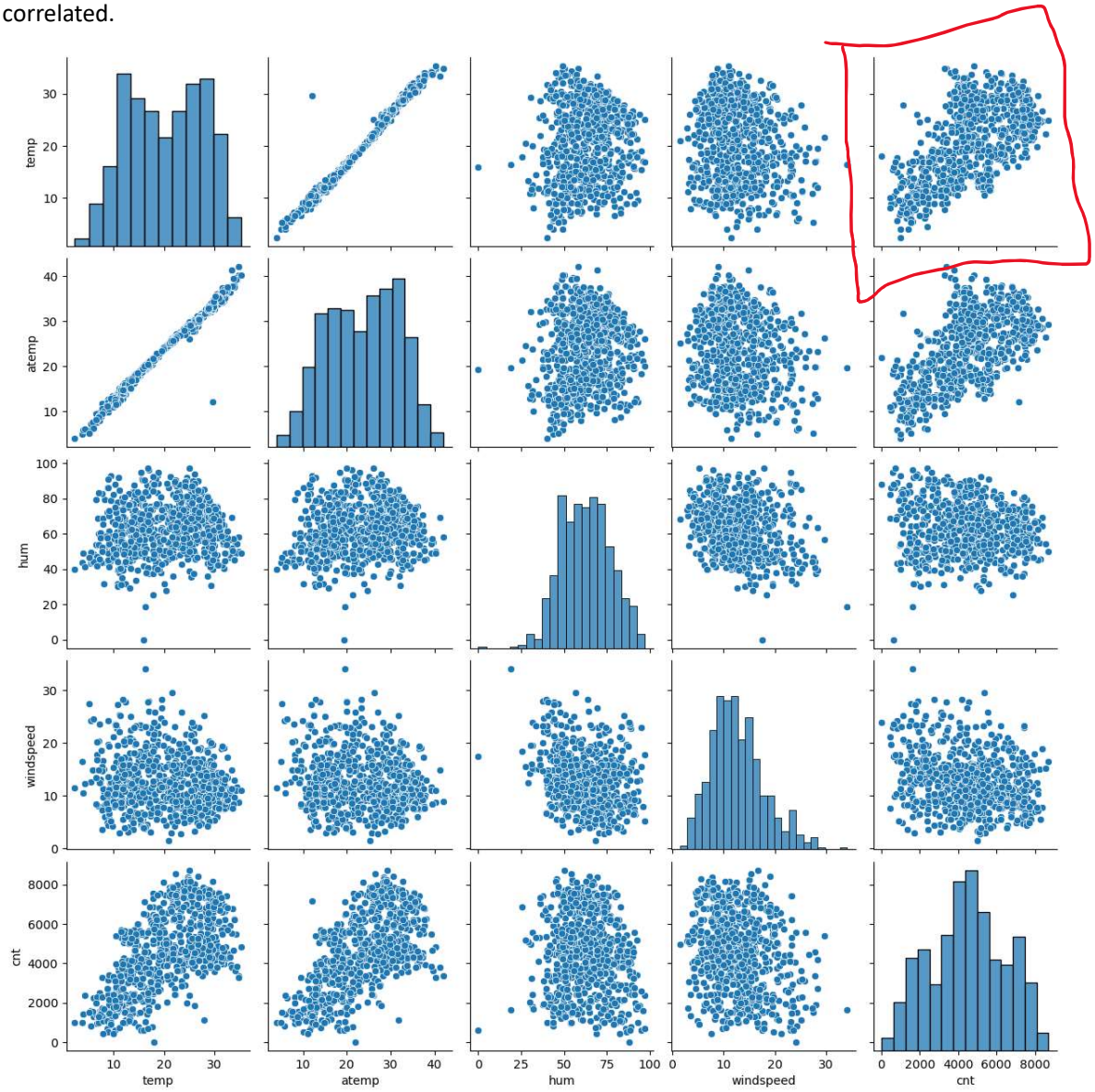
Value	Indicator Variable	
Furnishing Status	furnished	semi-furnished
furnished	1	0
semi-furnished	0	1
unfurnished	0	0

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

Answer:

And '**temp' Vs 'cnt' showed the strongest / highest correlation with the target variable** which is considered for model building. It is marked in red in the below pair-plot screenshot.

We also see that **atemp** has similar correlation with **cnt**. It is because **temp and atemp** are also highly correlated.

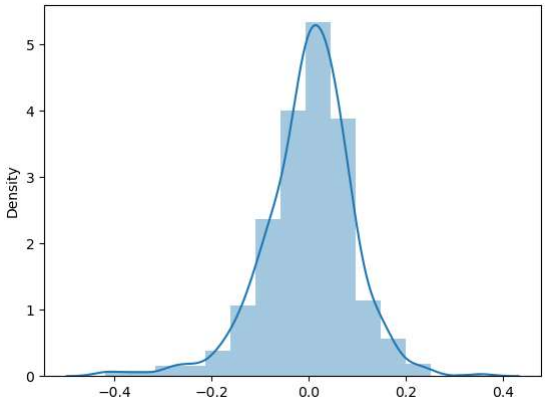


4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

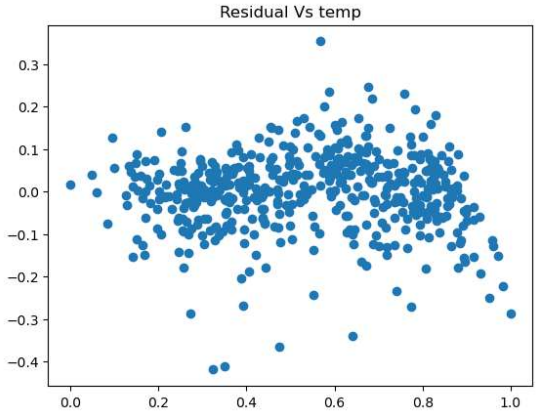
Answer:

I have validated assumptions of Linear Regression after building the model on the training dataset as follows:

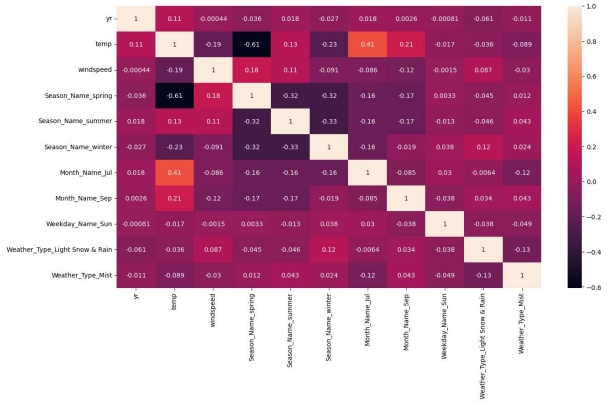
1. **Distribution plot for error terms/residuals:** We can see that it is normally distributed and mean is zero.



2. **There should be no visible pattern in the residual values**: We do not see any clear pattern



3. **Multicollinearity:** There is not significant multicollinearity between independent variables.



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:

The top 3 features contributing significantly towards explaining the demand of the shared bikes are

1. **Temperature (positively)**
2. **Year (positively)**
3. **Weather_Type_Light Snow & Rain (negatively)**

Above list is obtained from the equation for the best fit line which is as mentioned below:

cnt = 0.2034 + (**0.2339** * yr) + (**0.4917** * temp) - (0.1497 * windspeed) - (0.0682 * Season_Name_spring) + (0.0479 * Season_Name_summer) + (0.0818 * Season_Name_winter) - (0.0483 * Month_Name_Jul) + (0.0723 * Month_Name_Sept) - (0.0450 * Weekday_Name_Sun) - (**0.2847** * Weather_Type_Light Snow & Rain) - (0.0802 * Weather_Type_Mist)

General Subjective Questions

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

Answer:

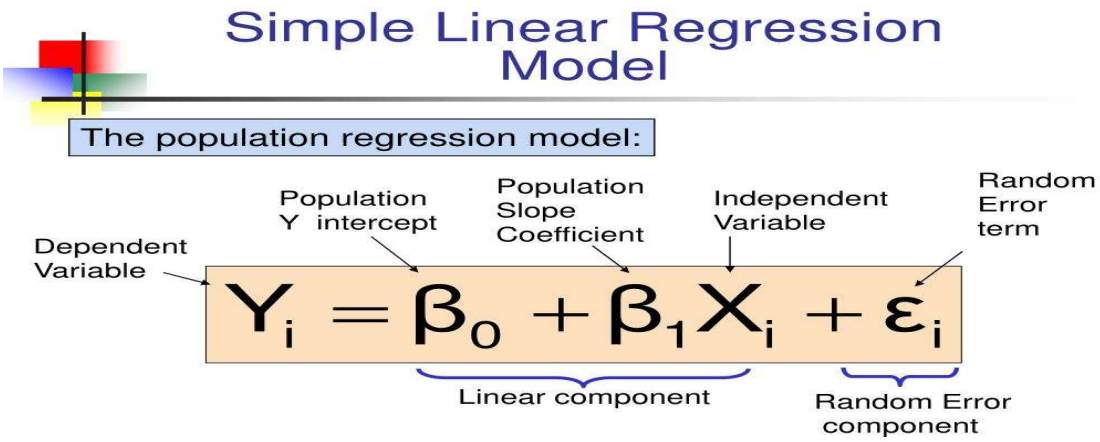
Linear regression is a **machine learning algorithm** used for predicting a continuous numeric output (also called the dependent variable) based on one or more input features (independent variables). It models the relationship between the input features and the output by fitting a linear equation to the observed data points. The primary goal is to find the **best-fitting line that minimizes the difference between the predicted and actual values**.

Linear regression relies on certain assumptions:
Linearity: The relationship between the features and the output is linear.
Independence: The residuals (differences between predicted and actual values) are independent and do not show patterns.
Homoscedasticity: The variance of the residuals is constant across all levels of the predictor variables.
Normality: The residuals are normally distributed.

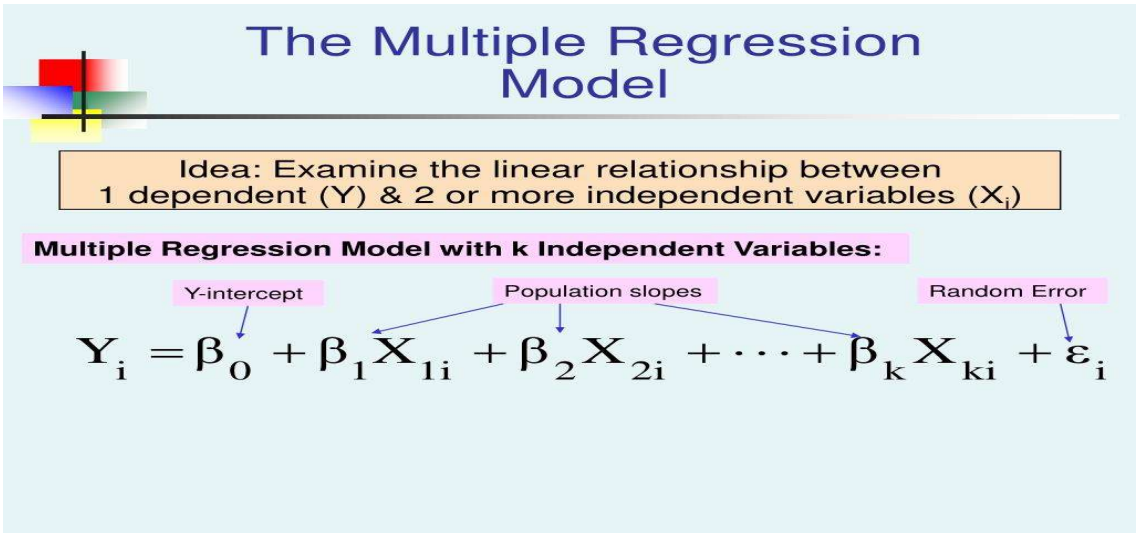
Linear regression is of the following two types –

- **Simple Linear Regression**
- **Multiple Linear Regression**

Simple Linear Regression: In simple linear regression, there's only one input feature (independent variable) and one output (dependent variable). The goal is to fit a line (equation) that best describes the linear relationship between the input and output. Below image shows an example of linear regression equation.



Multiple Linear Regression: When there are multiple input features, it's called multiple linear regression. The goal remains the same: find the best-fitting hyperplane that represents the linear relationship between multiple input features and the output. Below image is an example of multi linear regression equation.



The strength of the linear regression model can be assessed using 2 metrics:

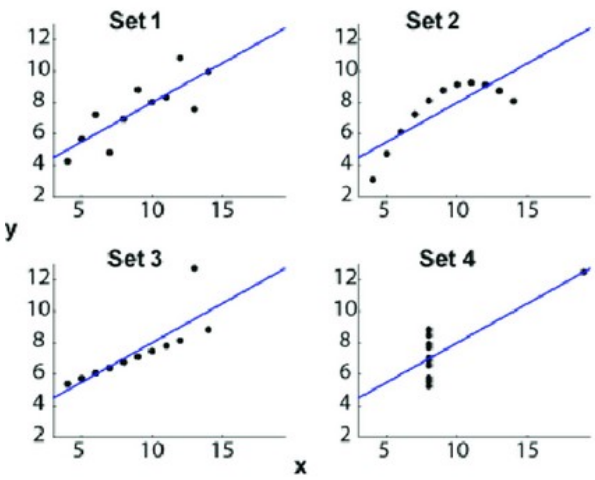
1. R2 or Coefficient of Determination
2. Residual Standard Error (RSE)

In summary, linear regression helps us find the best-fitting line to model the relationship between input and output variables. It's widely used for tasks like predicting house prices, stock prices, and more. Broadly speaking, it is a form of predictive modelling technique which tells us the relationship between the dependent (target variable) and independent variables (predictors). Remember that linear regression assumes a linear relationship, so it's essential to validate this assumption for the specific problem.

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

Answer:

Anscombe's quartet is a perfect example to understand the importance of Data visualization. This was developed 1st by a statistician named Francis Anscombe in 1973. He had taken 4 data sets of X & Y with 11 datapoints of X,Y in each set. All the datasets had almost same basic statistical parameters (Mean, Median, Standard deviation etc.), although when we plot the graphs of X &Y, for the 4 data sets, they look completely different from each other as shown below:-



Mean of Y	**7.50 in all 4 XY plots**
Sample variance of Y	**4.122 or 4.127 in all 4 XY plots**
Correlation (r)	**0.816 in all 4 XY plots**
Linear regression	**y = 3.00 + (0.500 x) in all 4 XY plots**

Data sets for the 4 XY plots

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Hence, these plots tell completely a different story about the relationship of X with Y. Dataset 1 is somewhat linear, Dataset 2 is curved, Dataset 3 is highly linear except a few outliers. The huge effect of outliers is also very much evident in 4th Dataset, where one value of x4 is more than, whereas all other values are same.

By highlighting the differences between visually similar datasets, Anscombe's quartet demonstrates that visualization is a critical step in understanding data and making informed decisions based on it. It serves as a reminder for analysts, researchers, and data scientists to always visualize data before drawing conclusions or making predictions.

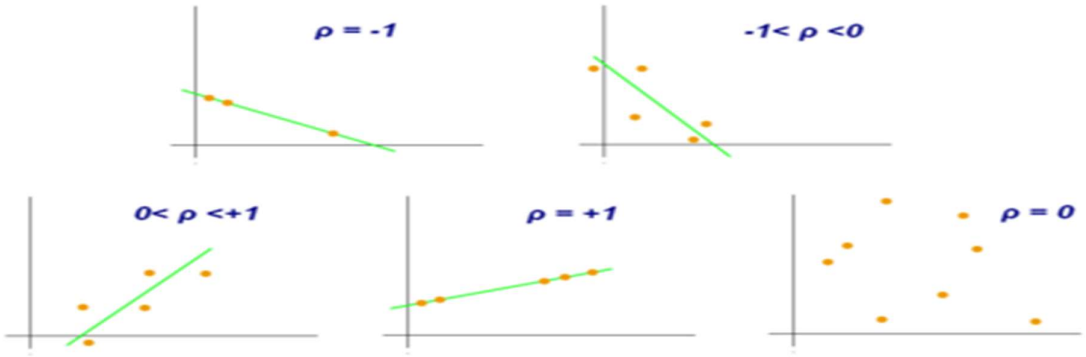
3. What is Pearson's R? (3 marks)

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. 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:



Pearson's correlation coefficient is calculated using the following formula:

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

Where,

r = Pearson Correlation Coefficient

x_i = x variable samples

y_i = y variable sample

$\bar{x}$ = mean of values in x variable

$\bar{y}$ =mean of values in y variable

The absolute value of r indicates the strength of the relationship:

Close to +1 or -1: Strong linear relationship.
Close to 0: Weak or no linear relationship.
Direction of Relationship: The sign of r (+ or -) indicates the direction of the linear relationship:

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

Answer:

Scaling is a process to standardize the numerical data present in the dataset in a fixed range. The purpose of scaling is to standardize the range of features or variables in a dataset, so that they can be compared and analysed more effectively.

Why is scaling important?

Consider a scenario where your dataset contains features with vastly different scales.
For example:

- Age (ranging from 10 to 60)
- Salary (ranging from 1 Lac to 40 Lacs)
- Number of bedrooms in an apartment (ranging from 1 to 5)

If we don't scale these features, the model might give undue importance to the feature with the largest scale (in this case, Salary).
Feature scaling ensures that each feature contributes proportionally to the model's decision-making process. It can reduce impact of extreme values, making the analysis more robust and accurate.

The difference between normalized and standard scaling is as follows:

SL. NO.	Normalized Scaling	Standard Scaling
1	It brings the data in the range of 0 and 1.	It brings all the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).
2.	Sensitive to outliers, as they can significantly affect the scale.	Less sensitive to outliers due to the use of mean and standard deviation.
3.	Preserves the original distribution, but doesn't handle outliers well.	Makes the distribution around 0 and is often preferred for algorithms that assume normal distribution.
4.	MinMax Scaling: $x = \frac{x - min(x)}{max(x) - min(x)}$	Standardisation: $x = \frac{x - mean(x)}{sd(x)}$

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

Answer:

The formula for VIF is: VIF = 1/ (1-R^2)
And VIF will be infinite only when denominator is zero and for denominator to be zero, R^2 has to be one. i.e. R^2 = 1.

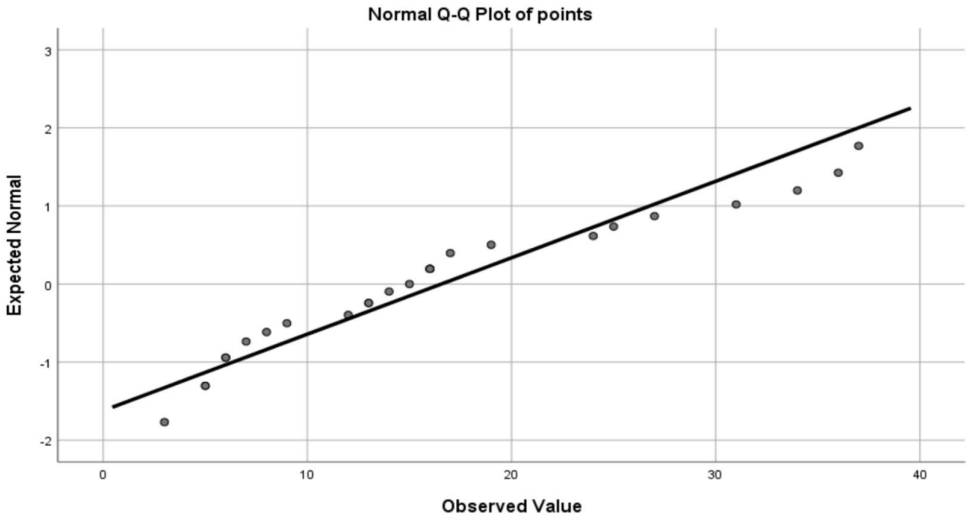
R^2=1 means that there is a perfect correlation between 2 independent variables. The infinite VIF value of a variable means that all the effect of this variable can be expressed by other variable (as they are having perfect correlation), which also would be having infinite VIF value. To rectify this problem, we should remove one of the variables having perfect correlation with other.

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

Answer:

The quantile-quantile (q-q plot) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not. Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution.

It is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight as shown below.



In linear regression, Q-Q plots are used to check the **normality assumption of residuals**. The use and importance of Q-Q plots in linear regression is that it helps us to check whether the **residuals are normally distributed or not**. If the residuals are normally distributed, then it means that the model is correctly specified and we can trust the results of our linear regression model. If the residuals are not normally distributed, then it means that there is something wrong with our model and we need to investigate further.