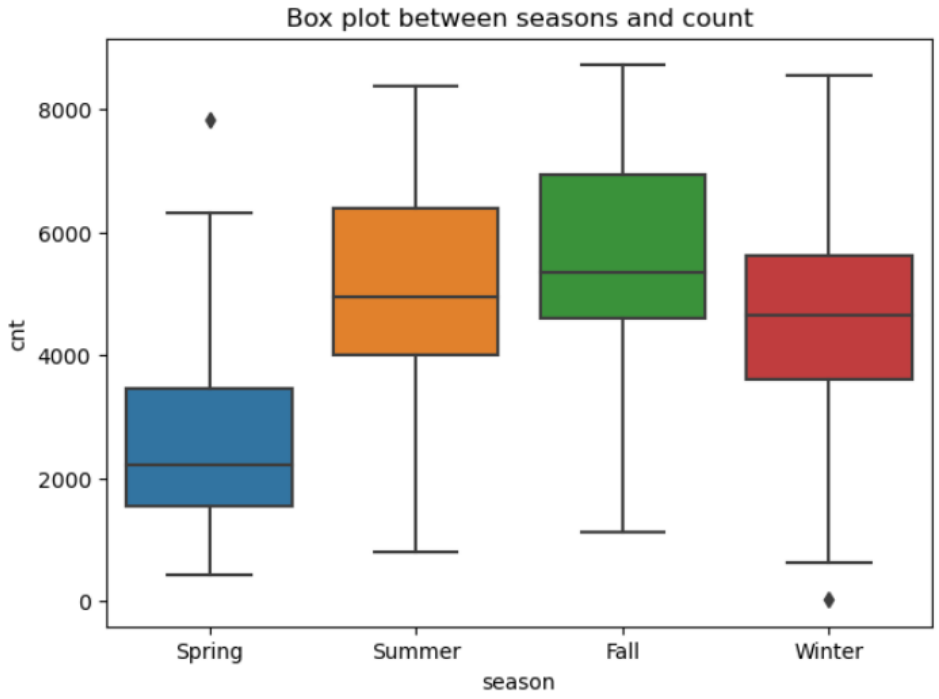
Assignment-based Subjective Questions

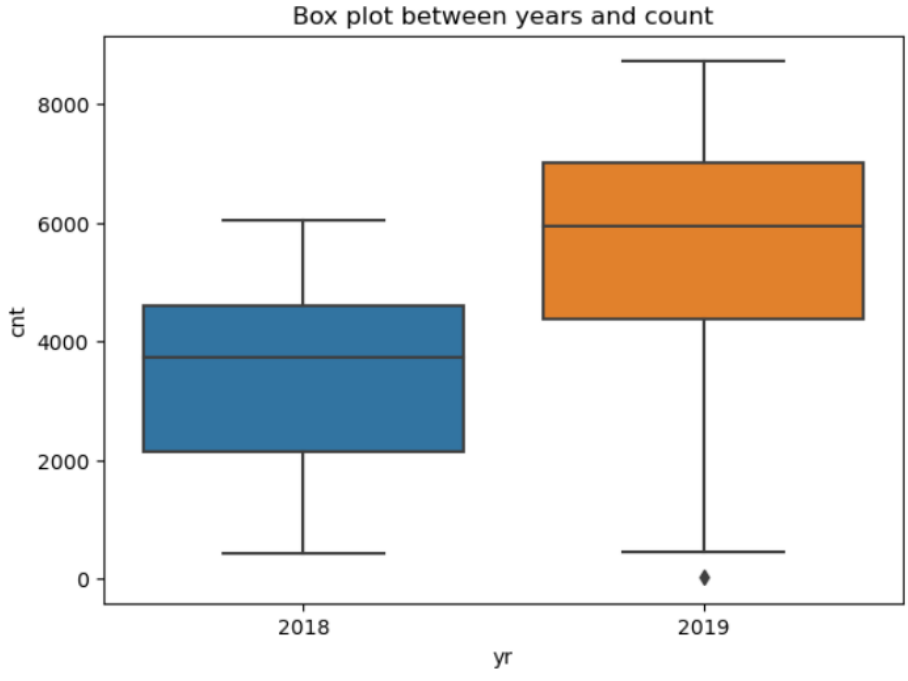
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer:

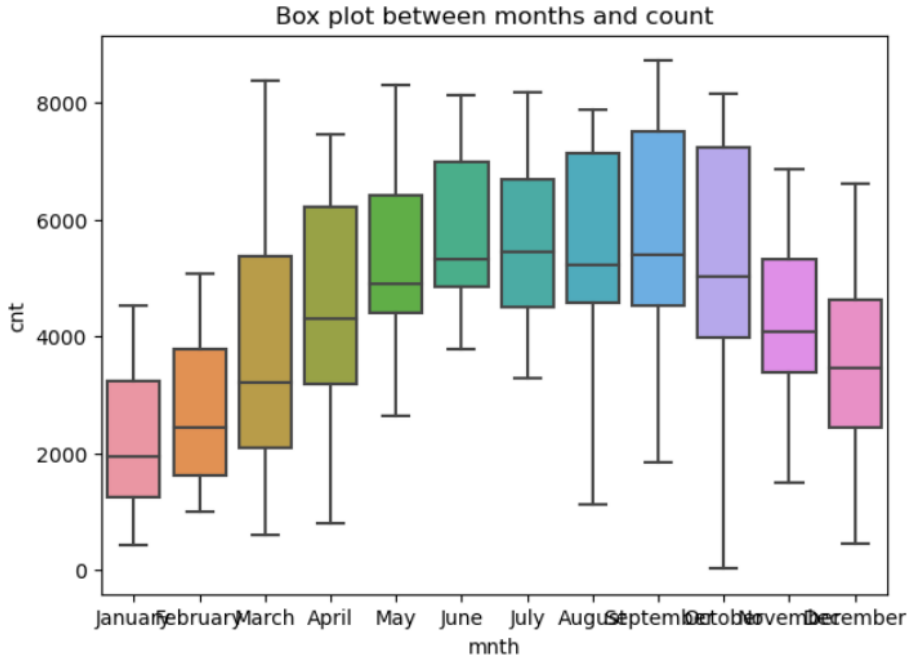
- During Fall, the usage of bikes is comparatively more than the other seasons and Spring at last.



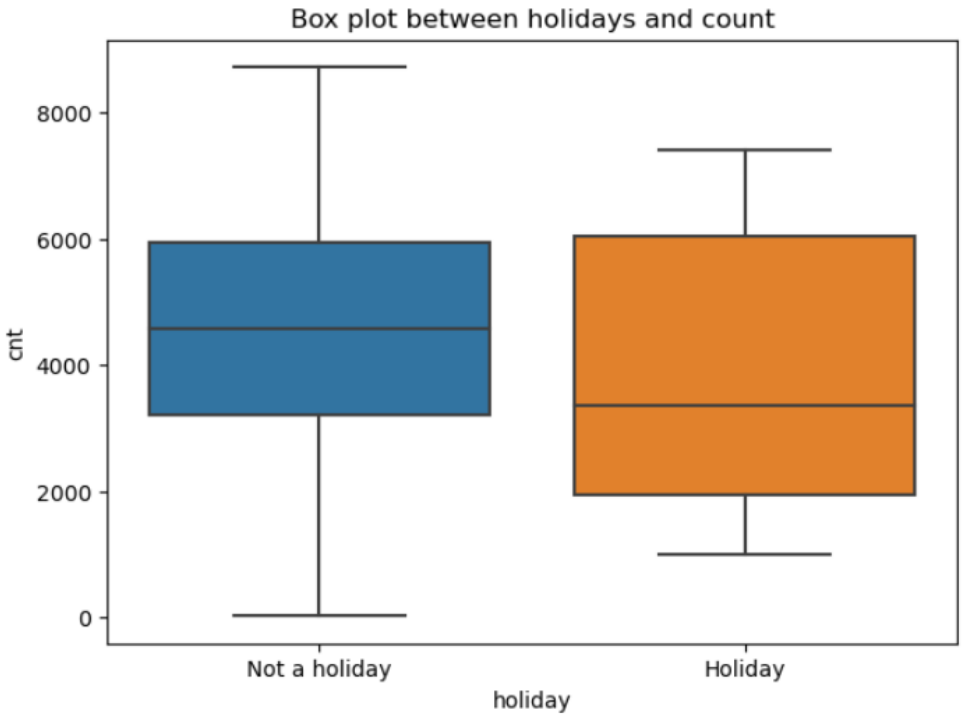
- 2019 has the most usage of bikes than 2018.



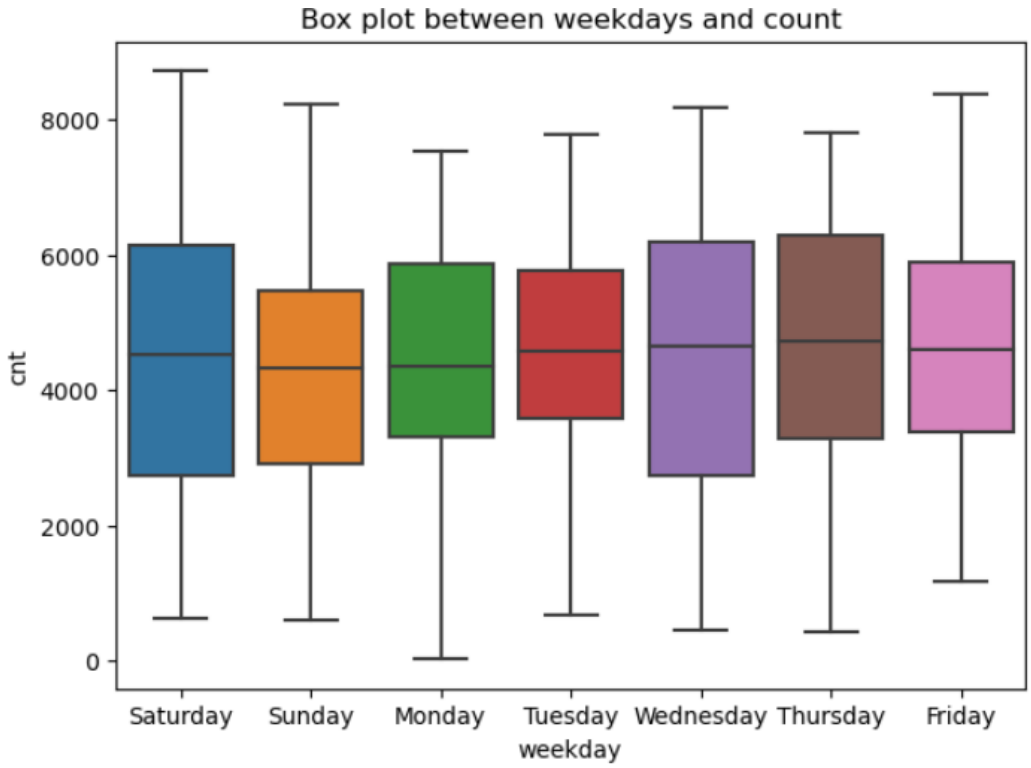
- Between August and October, the usage of bikes is more which implies the Fall season in America.



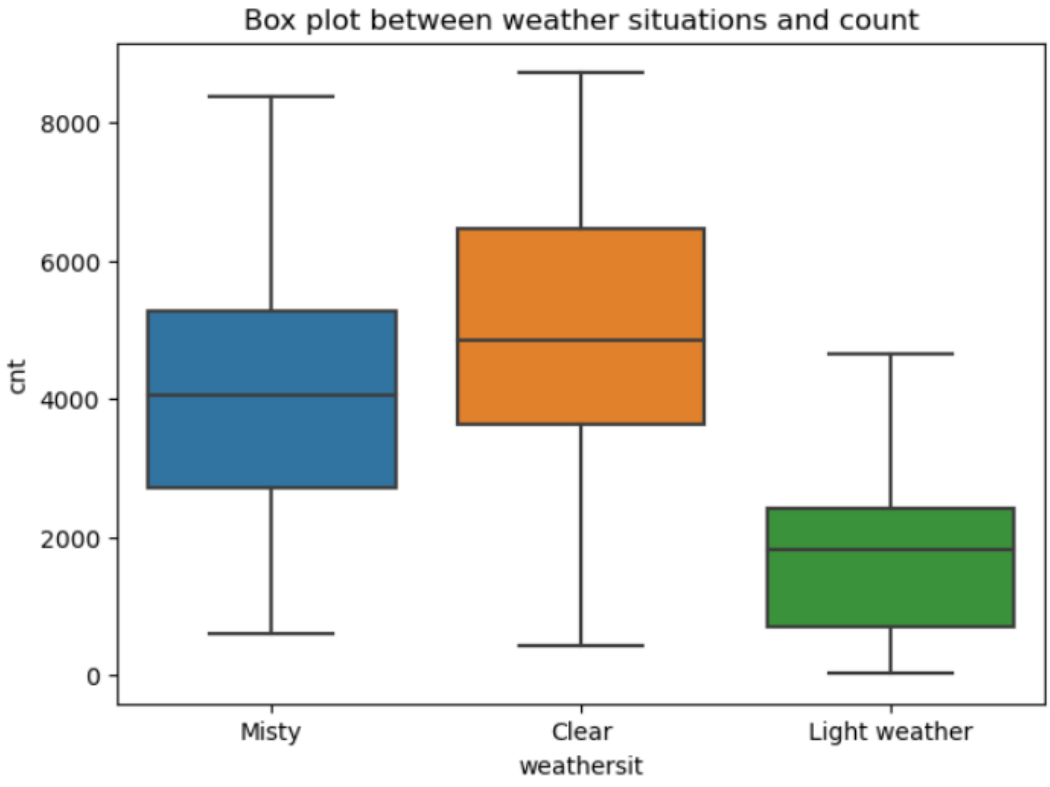
- People use bikes more in working days than holidays.



- People use bikes comparatively same all the days.



- People use bikes more when the weather is clear and don't use bikes in heavy weather conditions.



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

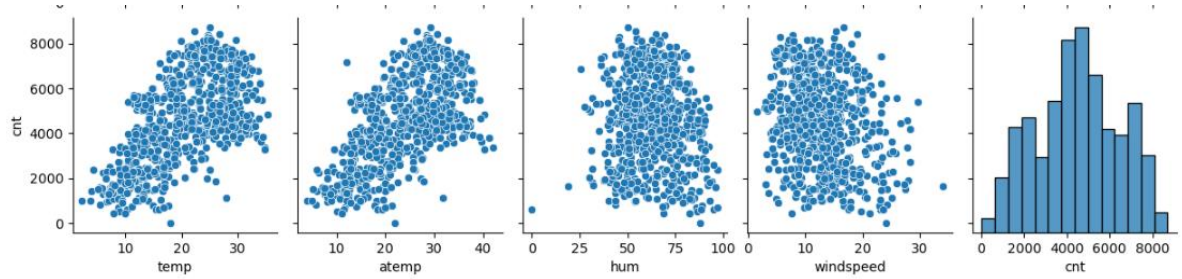
Answer:

This parameter is used in get_dummies( ) function. Two main purpose to use drop_first = True during dummy variable creation are:

- Multicollinearity - It avoids multicollinearity i.e.; it avoids other variables to predict one or more independent variables which leads to the model to be less accurate.
- Interpretability – It makes the interpretation of the variables more straightforward i.e.; it represents the coefficients relative to the dropped variable to predict the dependent variable.

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

Answer:



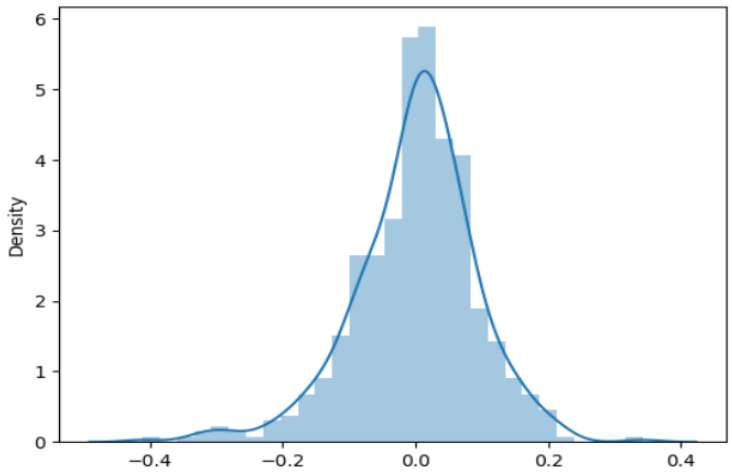
- Target variable – cnt
- Independent variable – temp and atemp (0.63 from correlation heatmap)

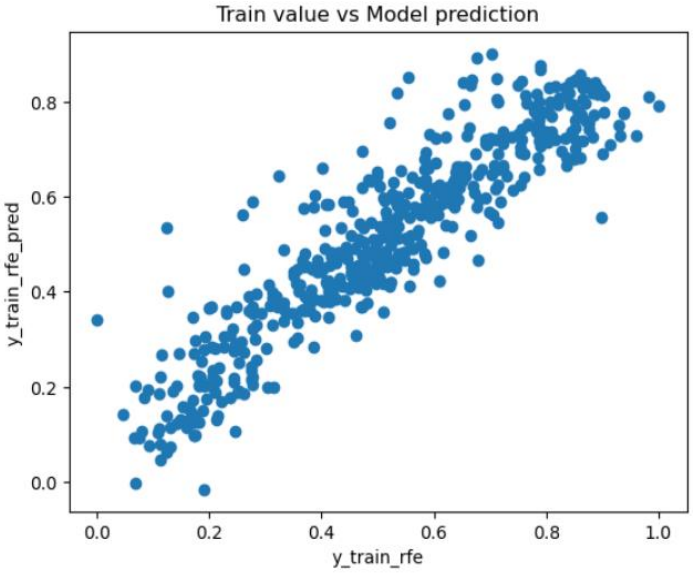
But I considered temp has the high correlation and dropped atemp during manual feature selection.

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

Answer:

- Residual Analysis – It follows a normal distribution centered at zero and shows a linear model.





- R2_score > 80

```
#R Squared Calculation on training dataset
r2_train_rfe = r2_score( y_true = y_train_rfe , y_pred = y_train_rfe_pred)
print(r2_train_rfe)
```

```
0.8295534066833782
```

- Adjusted R Squared > 80
- P values <= 0.05 and VIF <=5

OLS Regression Results

Dep. Variable:	cnt	**R-squared:**	0.830
Model:	OLS	**Adj. R-squared:**	0.826
Method:	Least Squares	**F-statistic:**	242.9
Date:	Wed, 13 Mar 2024	**Prob (F-statistic):**	1.53e-184
Time:	17:12:18	**Log-Likelihood:**	489.69
No. Observations:	510	**AIC:**	-957.4
Df Residuals:	499	**BIC:**	-910.8
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2849	0.040	7.056	0.000	0.206	0.364
temp	0.5147	0.031	16.542	0.000	0.454	0.576
hum	-0.2728	0.032	-8.542	0.000	-0.336	-0.210
windspeed	-0.1887	0.026	-7.194	0.000	-0.240	-0.137
season_Spring	-0.1005	0.015	-6.536	0.000	-0.131	-0.070
season_Winter	0.0653	0.013	5.153	0.000	0.040	0.090
yr_2019	0.2281	0.008	27.102	0.000	0.212	0.245
mnth_July	-0.0805	0.018	-4.538	0.000	-0.115	-0.046
mnth_September	0.0598	0.016	3.723	0.000	0.028	0.091
holiday_Not a holiday	0.0920	0.026	3.472	0.001	0.040	0.144
weathersit_Light weather	-0.1980	0.026	-7.659	0.000	-0.249	-0.147

Omnibus:	51.023	**Durbin-Watson:**	1.968
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	100.492
Skew:	-0.593	**Prob(JB):**	1.51e-22
Kurtosis:	4.823	**Cond. No.**	21.8

	Features	VIF
0	const	94.85
1	temp	2.87
4	season_Spring	2.53
5	season_Winter	1.75
7	mnth_July	1.29
2	hum	1.26
3	windspeed	1.15
8	mnth_September	1.11
10	weathersit_Light weather	1.11
6	yr_2019	1.03
9	holiday_Not a holiday	1.01

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

Answer: By seeing the coefficients, it explains

- temp – 1st
- hum – 2nd
- yr_2019 – 3rd
- weathersit_Light weather – 4th
- windspeed – 5th

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2849	0.040	7.056	0.000	0.206	0.364
temp	0.5147	0.031	16.542	0.000	0.454	0.576
hum	-0.2728	0.032	-8.542	0.000	-0.336	-0.210
windspeed	-0.1887	0.026	-7.194	0.000	-0.240	-0.137
season_Spring	-0.1005	0.015	-6.536	0.000	-0.131	-0.070
season_Winter	0.0653	0.013	5.153	0.000	0.040	0.090
yr_2019	0.2281	0.008	27.102	0.000	0.212	0.245
mnth_July	-0.0805	0.018	-4.538	0.000	-0.115	-0.046
mnth_September	0.0598	0.016	3.723	0.000	0.028	0.091
holiday_Not a holiday	0.0920	0.026	3.472	0.001	0.040	0.144
weathersit_Light weather	-0.1980	0.026	-7.659	0.000	-0.249	-0.147

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

Linear regression is a fundamental statistical and machine learning technique used to model the relationship between a dependent variable (denoted as y) and one or more independent variables (denoted as X). It assumes a linear relationship between the independent variables and the dependent variable. The objective is to create a best fit straight line model that predicts the target variable efficiently. There are two types of linear regression. They are

- Simple Linear Regression (SLR) – one independent variable and one target variable.
- Multiple Linear Regression (MLR) – one or more independent variables and one target variable.

- A simple representation of linear regression with one independent variable is

$$y = mx + c + e$$

where,

- y is the target variable
- x is the independent variable
- m is the slope
- c is the intercept
- e is the error term

  - which can be extended for multiple independent variables.

- In order to find the best fit line, A measure called sum of squared errors is commonly used which is a cost function. To minimize or maximize this cost function, Gradient descent optimization is used.
- The strength of the linear regression model can be assessed using 2 metrics:
  1. R^2 or Coefficient of Determination
  2. Residual Standard Error (RSE)

Steps involving in creating a linear regression model:

- Data Collection
- Data Preprocessing
- EDA
- Train Test Split

- Training the model
- Evaluate the model
- Interpret the model

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet comprises four distinct datasets created by Francis Anscombe in 1973. The purpose of these datasets is to underscore the significance of visually inspecting data and to challenge the reliance only on summary statistics. Each dataset in Anscombe's quartet contains pairs of variables that exhibit similar summary statistics, including means, variances, correlations, and regression lines. But each dataset reveals unique patterns and characteristics when graphed or analyzed further.

1. **Dataset I**: This dataset demonstrates a straightforward linear relationship between the variables, with minimal variability around the regression line.
2. **Dataset II**: Similar to Dataset I in terms of summary statistics, Dataset II also displays a linear relationship. But it is heavily influenced by a single outlier, leading to notable differences in the regression line and correlation coefficient.
3. **Dataset III**: Unlike the first two datasets, Dataset III shows a non-linear relationship between the variables, following a clear quadratic pattern.
4. **Dataset IV**: Dataset IV illustrates the impact of outliers on summary statistics. While most data points cluster around two distinct x-values, one outlier significantly skews the regression line and correlation coefficient.

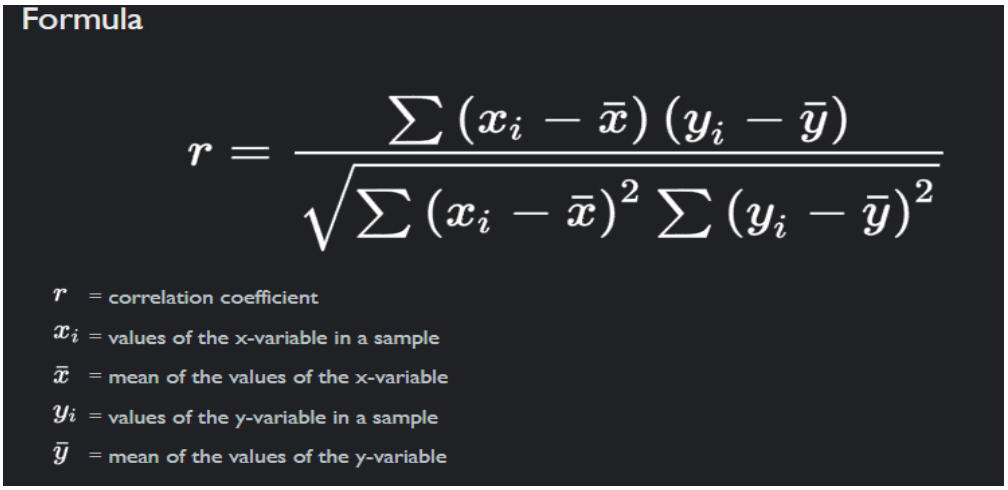
   It concludes that the necessity to combine numerical analysis with visual inspection when interpreting data. It underscores the importance of exploring data graphically to gain a comprehensive understanding of its underlying patterns and relationships.

3. What is Pearson's R?

Answer:

The Pearson correlation coefficient (*r*) is the most common way of measuring a linear correlation only. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. It summarizes the characteristics of a dataset. It is widely used in various fields, including statistics, psychology, economics, and biology, to analyze the strength and direction of relationships between variables.

- 0 < r <= 1 indicates a positive linear relationship
- r = 0 indicates no correlation
- -1 <= r < 0 indicates a negative linear relationship



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

Answer:

Scaling is the process to normalise the data within the particular range. There are two ways to scale the data. They are

- Normalisation – Min Max Scaling – It scales the values between 0 and 1
  - X = (x – xmin) / (xmax – xmin)
- Standardisation – It scales the value that it has a mean of 0 and standard deviation of 1.
  - X = (x – mean) / standard deviation

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

Answer:

VIF(i) = 1 / (1 – Ri^2)

This is the formula to calculate VIF. The reason for the value of VIF to be infinite is that the R Squared values is equal to 1 and the denominator equals to 0. This denotes perfect correlation between variables.

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

Answer:

Q-Q plot is also called as Quantile – Quantile plot is a graph to determine whether 2 datasets come from populations with common distribution. It looks like a scatter plot. It gives a rough straight line when the data came from same distribution. It is used to see if the points form roughly a straight line. If it doesn't, then the residuals aren't Gaussian and thus the errors too.

Importance of Q-Q plot:

- Sample sizes need not to be equal
- Many distribution aspects can be tested simultaneously
- It gives the nature of difference than other analytical methods