

Linear Regression Assignment

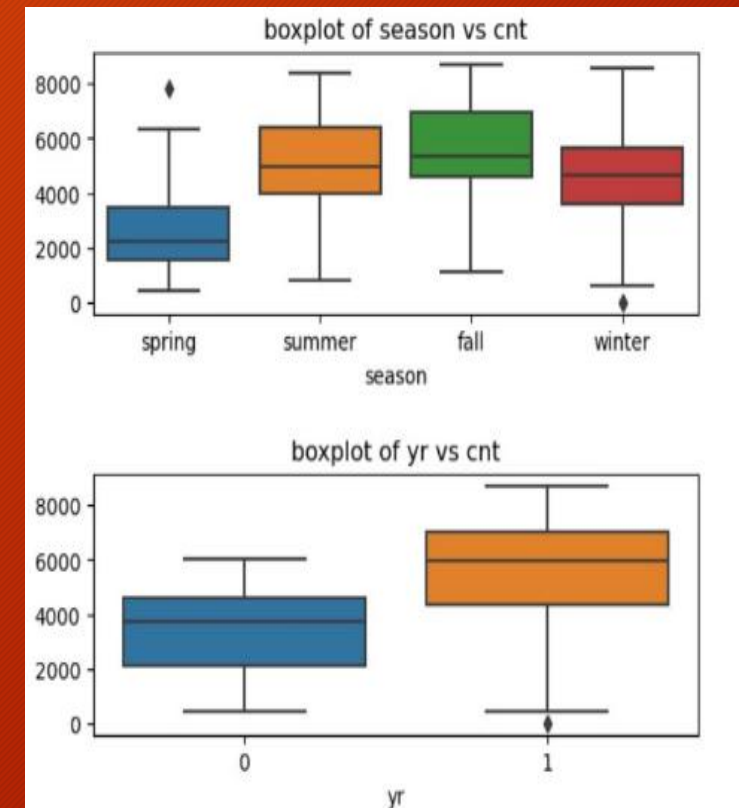
By: Garima rai

Assignment-based Subjective Questions

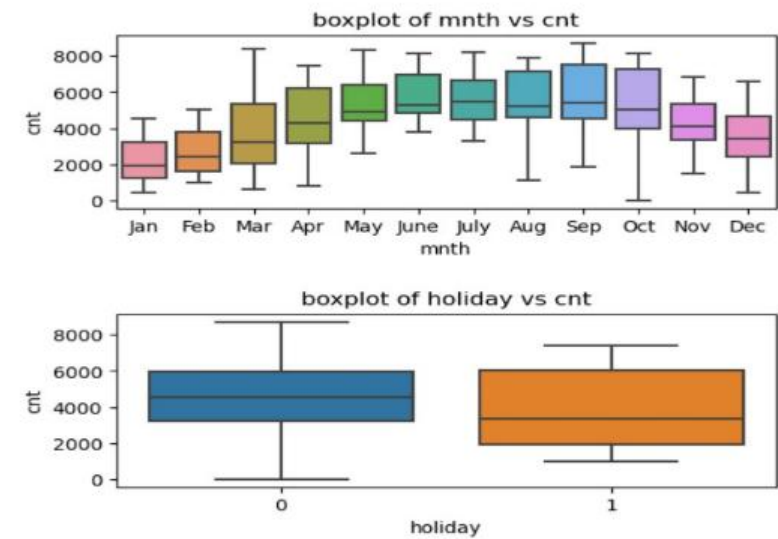
1 From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Ans) Boxplots provide a visual interpretation of the central tendency and distribution and they can also show the presence of outliers. By examining the categorical variables with the target variable through boxplot we can draw following inferences:

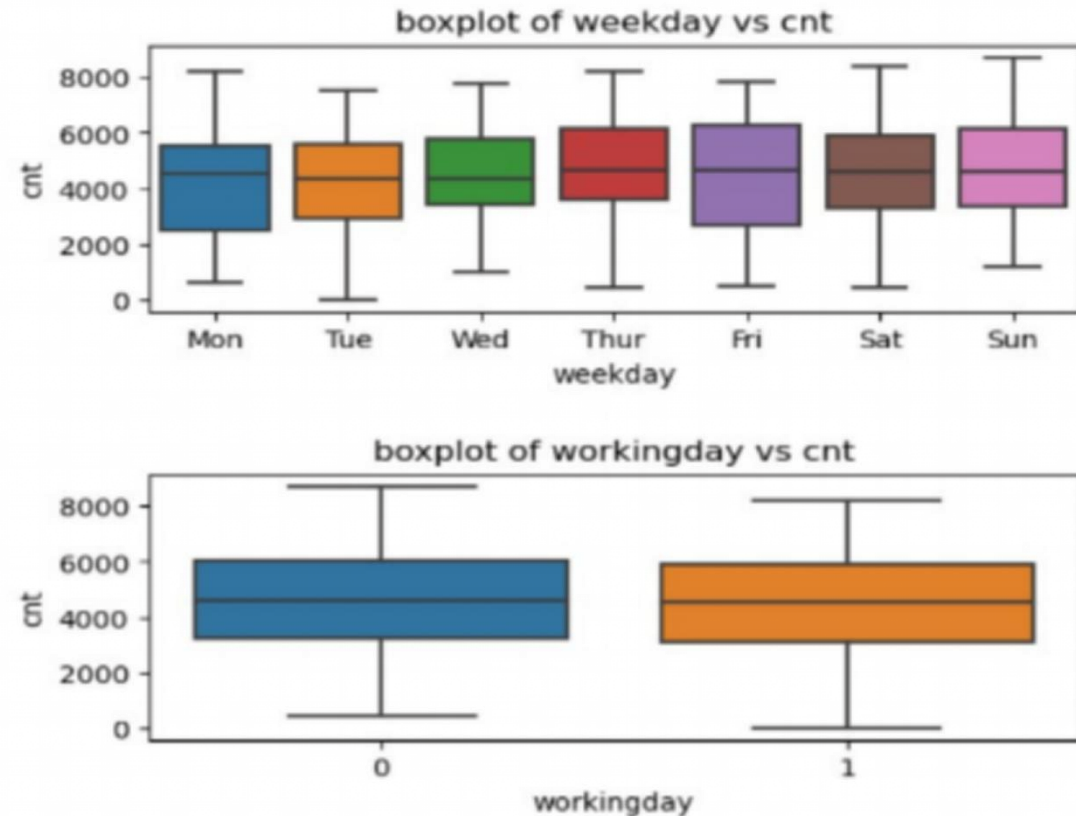
- **Season:** it shows the bike rentals for each season. The median of fall, winter and summer is the highest which indicates a greater number of bikes put on rental during these seasons. Whereas it is lowest in the spring.
- **Year:** this compares the bikes rentals between two years present in the data. Year 1 (2019) shows a hike in rental bikes as compared to year 0 (2018).



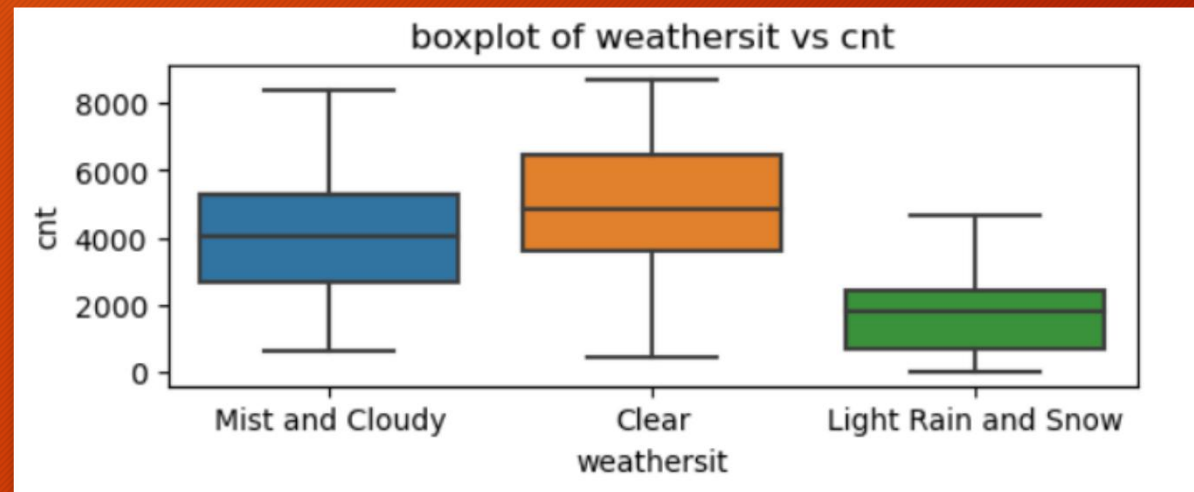
- **Month:** it shows distribution of bike rentals for each month. We can see a considerable increase in the bike rentals from May to October which can be traced from the demands based on season.
- **Holiday:** compares between holidays and non- holidays. The median of rental bikes on holidays is slightly lesser as compared to non-holidays means people don't use it for leisure.



- **Weekday** - the median demand is slightly higher in the middle of the week. Means more bikes were rented in these days.
- **Working day** - There isn't a significant difference in bike demand between working days and non-working days.



- **Weather:** this shows the distribution of bike rentals under different weather conditions. Harsh weather conditions like snow and rain have lower medians whereas clear it is higher on clear days and cloudy days.

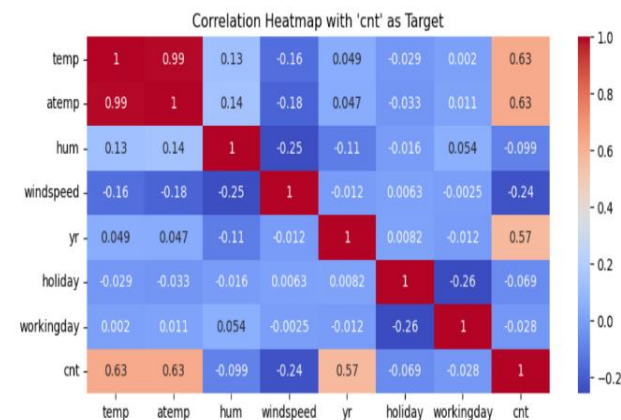


2 Why is it important to use drop first=True during dummy variable creation? (2 mark)

- When we create dummy variables we convert categorical columns into separate binary columns, if we don't drop the first column, the dummy variables will be highly correlated, because it knows the values of all, but one dummy variable tells us the value of the last one too. This will make it tough for the model to work properly and can lead to confusing results. By removing first dummy variable it will remove the correlation and make the data easier to handle.
- It also helps in avoiding redundancy. If we do not drop the first column, we will have an extra column that don't add any value or information but may increase the complexity. By removing one category we make it simple without losing any meaningful insights.

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

- Numerical variable that has highest correlation with the target variable is temperature and atemp it is because as the temperature increases (to a certain extent – comfortable point), people prefer to be outdoors. If the weather is warmer and comfortable people would like to rent a bike and go for a bike ride. People are more inclined to rent bikes for leisure rides, commutes, or exercise



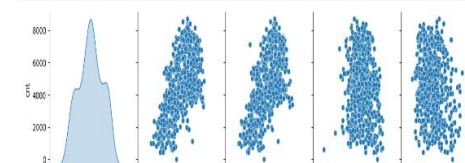
```
In [149]: # Subset the data to include only the specified columns and the target variable 'cnt'
```

```
subset_data = df[['cnt'] + cont_columns]
```

```
# Create the pair plot
```

```
sns.pairplot(subset_data, kind='scatter', diag_kind='kde')  
plt.show()
```

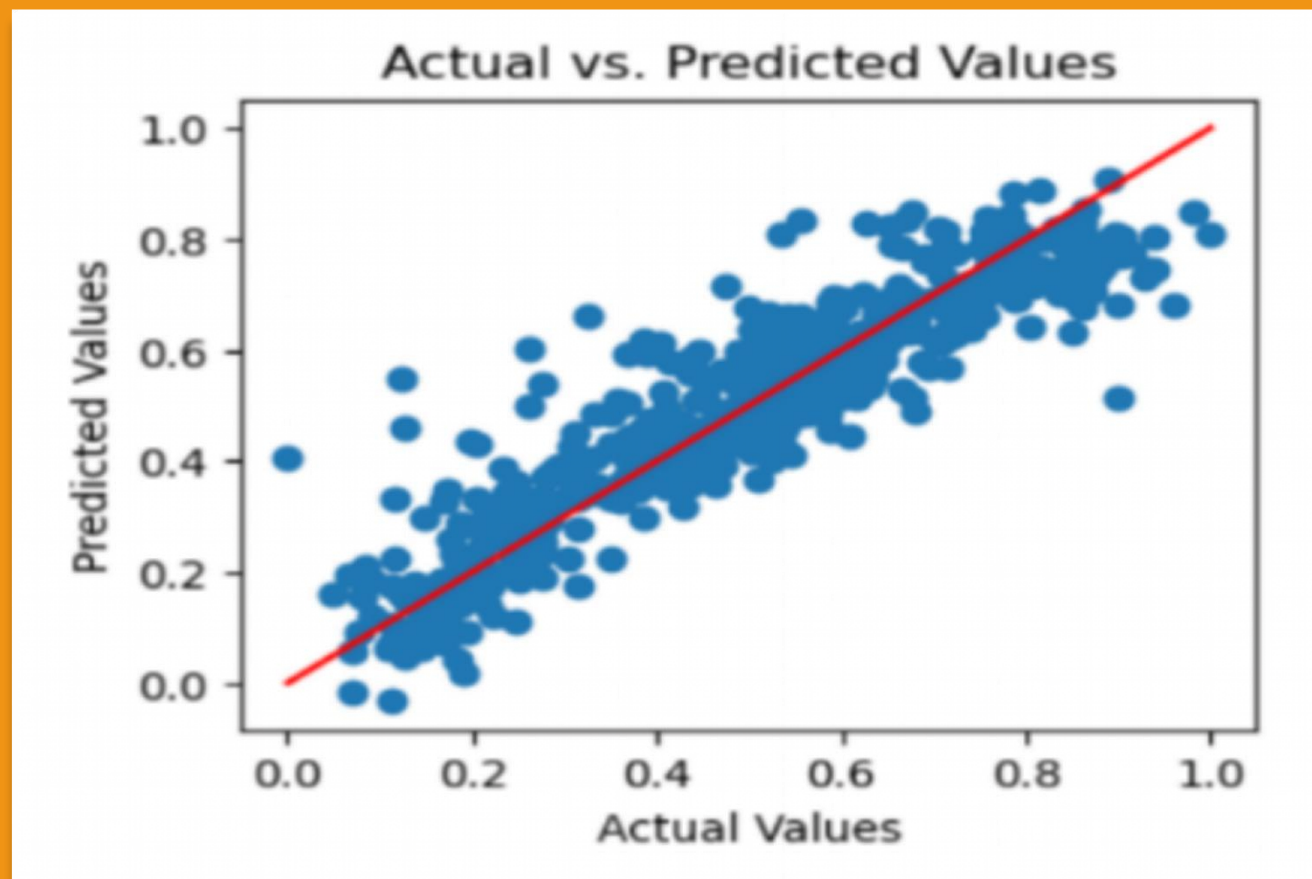
```
# temp and atemp shows a linear relation with the target variable
```



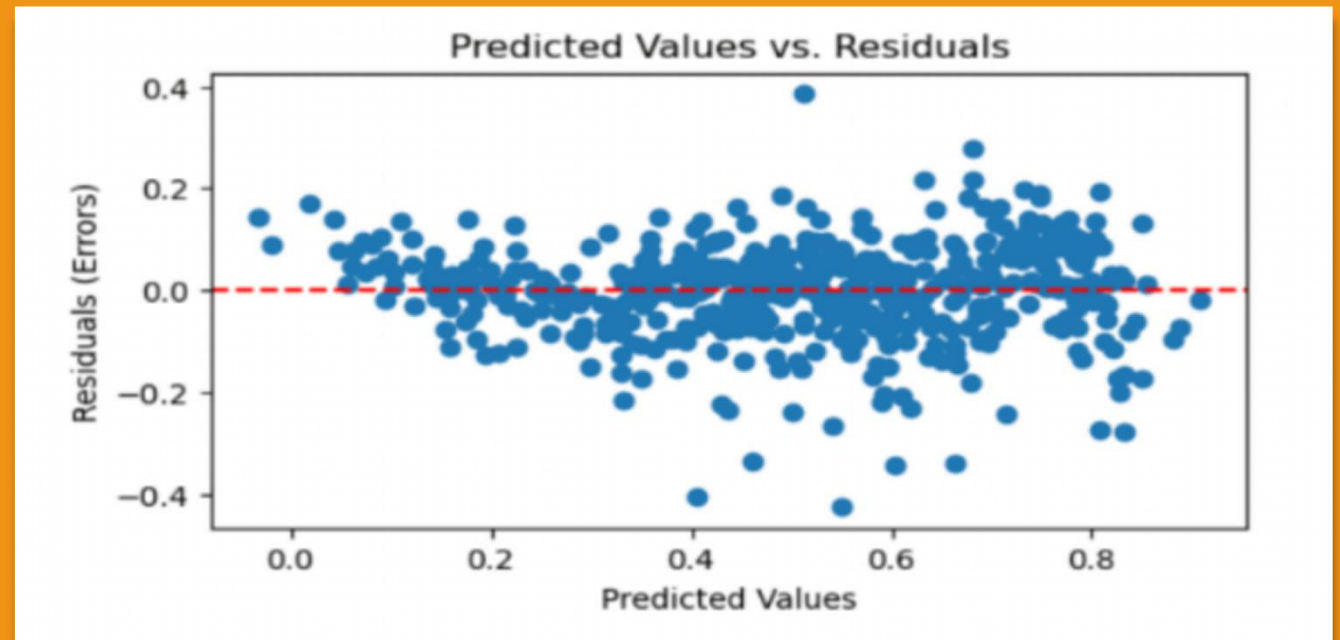
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

- Validating the assumptions of linear regression after building the model on training set is a crucial step to ensure model's accuracy. After building the model on the training set, the assumptions of linear regression can be validated as follows:

1. **Linearity:** by plotting a scatter plot between the predicted and the actual values. A diagonal line will confirm that the relationship is linear. The relationship between the independent variables and the dependent variables should be linear for the model to make accurate predictions.



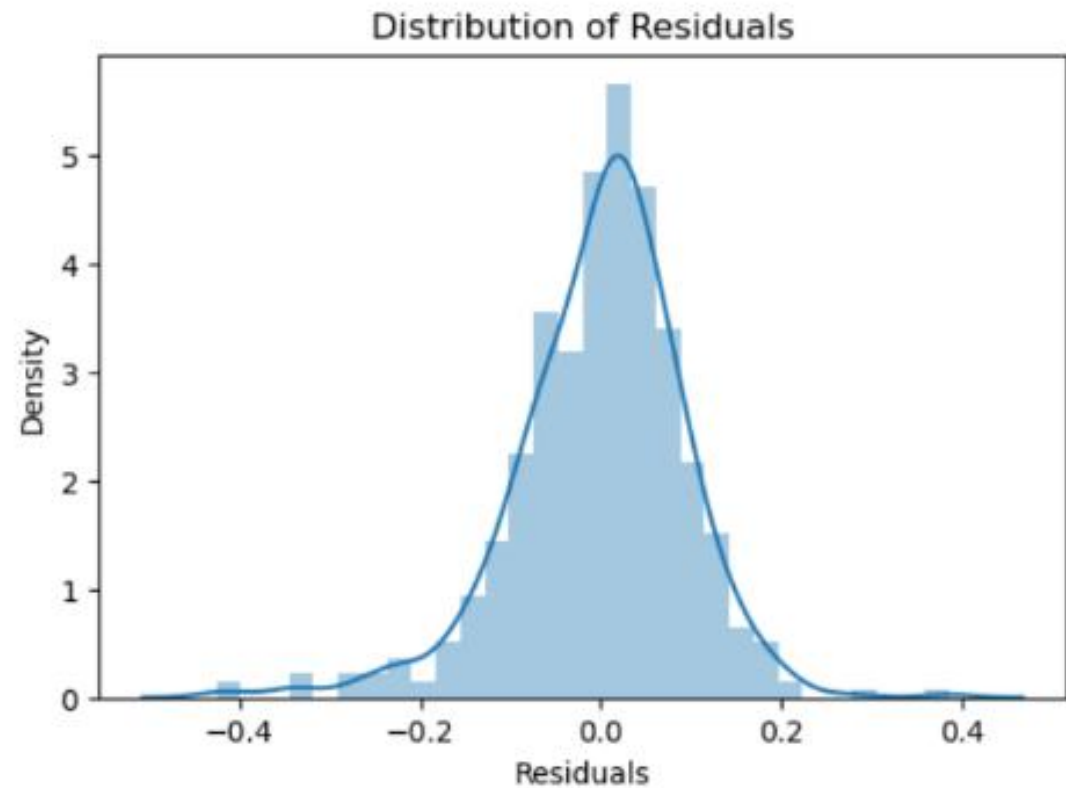
2. Independence of error:
by plotting the residuals(errors) against the predicted values. It should not show any evident pattern or trend in the plot. the residuals should be independent of each other



3. No multicollinearity: by checking the Variance Inflation Factor (VIF) for each predictor variable. The values should be less than 5, if the VIF values are greater than 5-10 it might indicate multicollinearity.

	features	vif
2	temp	4.22
0	yr	2.06
4	summer	1.94
6	July	1.58
5	winter	1.57
9	Mist and Cloudy	1.55
3	spring	1.40
7	Sep	1.34
8	Light Rain and Snow	1.07
1	holiday	1.04


4. Normality of errors:
using a histogram to check
the distribution of residuals.
The residuals should follow
a normal distribution, as
linear regression assumes
that the errors are normally
distributed.



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans) Based on the final model the top 3 features contributing significantly explaining the demand of the shared bikes are:

- 1. Summer:** summer is usually a warmer and more comfortable weather, which can lead to increased outdoor activities. People would like to rent bikes for leisure, commute to office or exercise during summer month compared to colder months, thus we can say that summer season can see a spike in bike rentals.



2. Holiday: people take a break from their daily routine on holidays, they may rent bikes for leisure activities, or the demand may decrease if the bikes are primarily rented for commuting to work or school on regular days. The holiday can capture these variations in the bike rental patterns.

3. Weather (Mist and cloudy): misty and cloudy weather conditions not severe as heavy rains or snow, so some people might enjoy going for a ride under cloudy sky as it can provide a cooler environment, especially during warmer seasons. The atmosphere of a misty morning or evening may appeal to certain individuals who appreciate the cooler temperature it offers.

General Subjective Questions

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

Ans) Linear regression is a supervised machine learning algorithm which is used for predicting the outcome of the variable i.e., dependent variable based on one or more independent variables.

It is called “linear” because it builds a relationship between dependent and the independent variable, means it gives a fit straight line to the data.

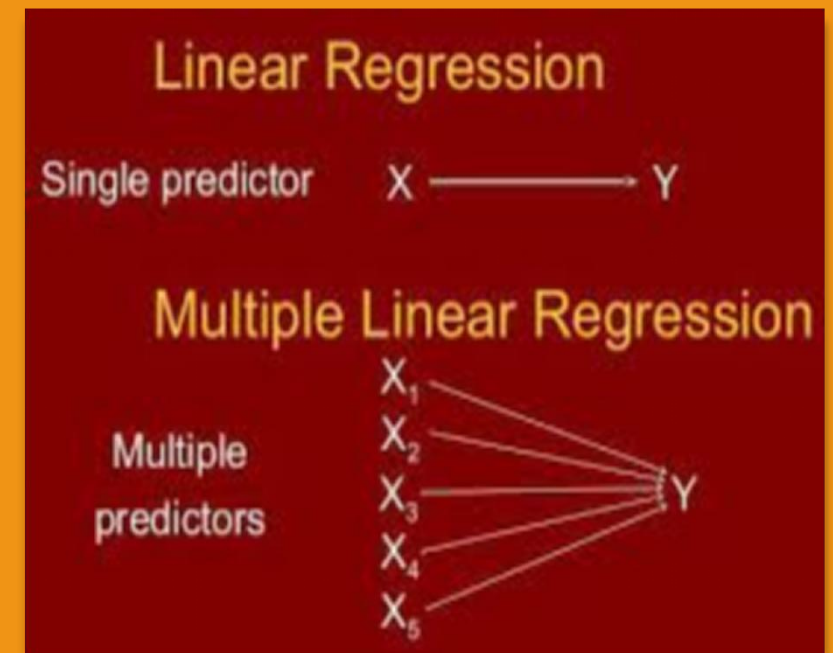
Example: In a company if the sales have increased every month from last few years, by conducting linear analysis on the sales along with the monthly sales we can predict the sales in the coming future.

SIMPLE LINEAR REGRESSION

- In simple linear regression, there is one predictor variable(X) and the outcome variable (Y). the relationship between X and Y is represented as a straight line = $Y = mX + c$.
- The goal is to find the values of 'm' and 'c' that fits the data in the best way. 'm' represents the slope of the line and 'c' represents the intercept value (the value of Y when X is 0).

MULTIPLE LINEAR REGRESSION

- Multiple linear regression is used to estimate the relationships between two or more independent variables and one dependent variable.
- $Y = C + m_1 * X_1 + m_2 * X_2 + \dots + m_n * X_n + E$
- Y is dependent variable that we want to predict.
- C is the intercept (the value of Y when all independent variables are 0).
- m_1, m_2, \dots, m_n are the coefficients of the independent variables.
- X_1, X_2, \dots, X_n are the independent variables.
- E is the error term, that accounts for the unexplained variability in dependent variable.

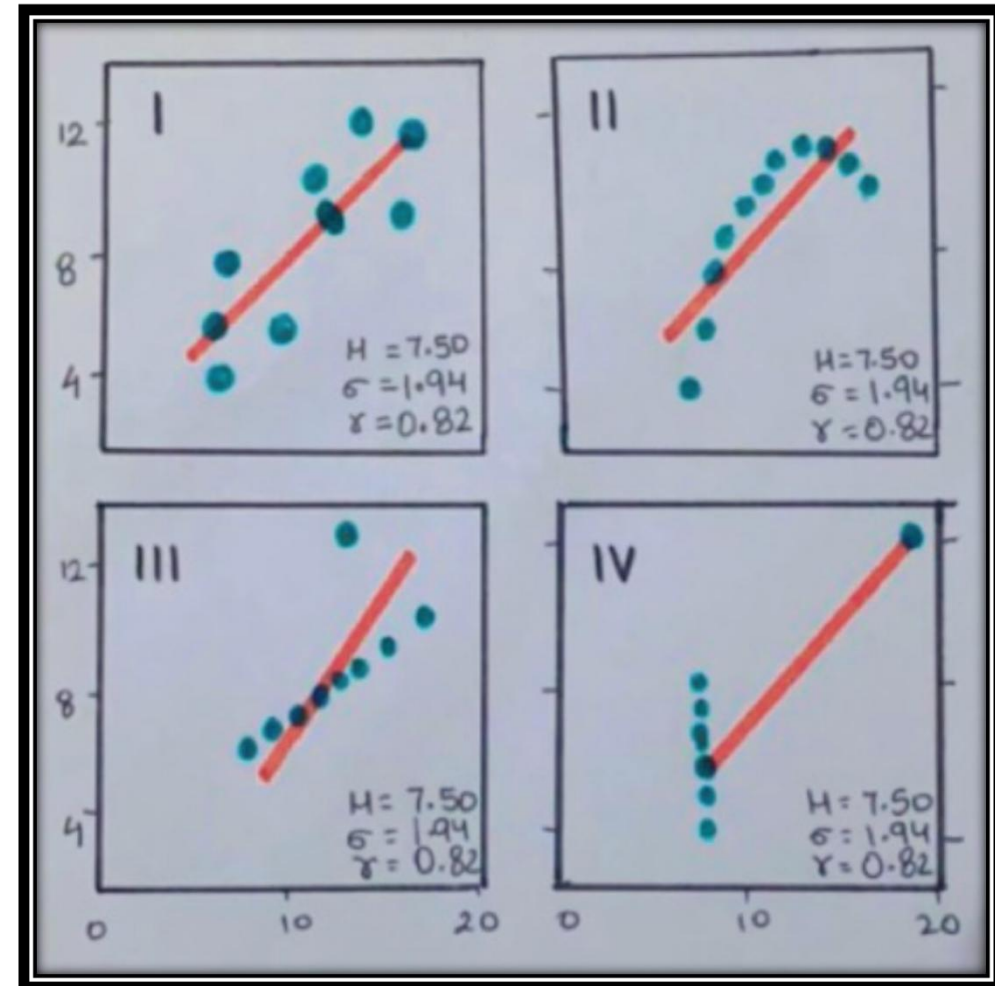


2. Explain the Anscombe's quartet in detail. (3 marks)

- **Ans)** The quartet was created by a statistician Francis Anscombe in 1973 to show the importance of graphing the data before analysing it and to emphasize how statistics alone may not explain the complete story of the data.
- Anscombe's quartet is a set of four small datasets that are very much similar to each other, yet they exhibit very different distributions when they are graphed.
- Anscombe's Quartet consists of four data sets that contains two variables X (independent variable) and Y (dependent variable).
- The summary statistics (mean, variance, correlation, etc.) for each of the four data sets is very similar. This can lead us to believe that the data sets are alike.

Anscombe's Quartet becomes evident when it is graphed, which tell a distinct and often illogical pattern:

- Dataset I - A simple linear relationship
- Dataset II - A curved, non- linear relationship.
- Dataset III - A strong linear relationship but with one outlier.
- Dataset IV: - No linear relationship, but one extreme outlier that influences the regression line.



IMPLICATIONS

Anscombe's Quartet shows that relying only on summary statistics like mean and correlation can be misleading when trying to understand the data.

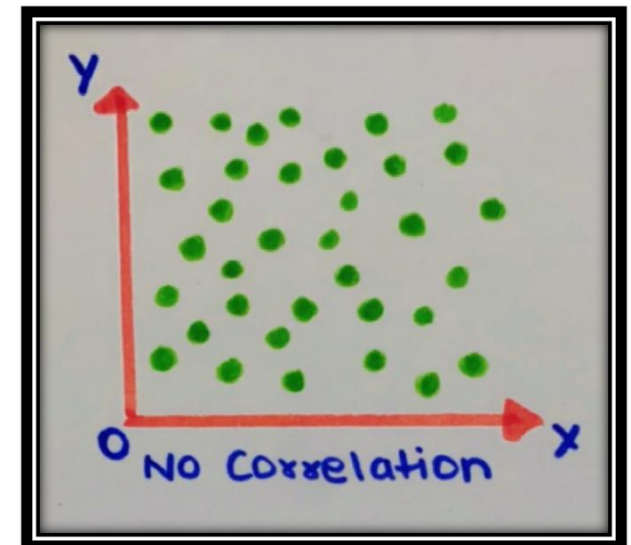
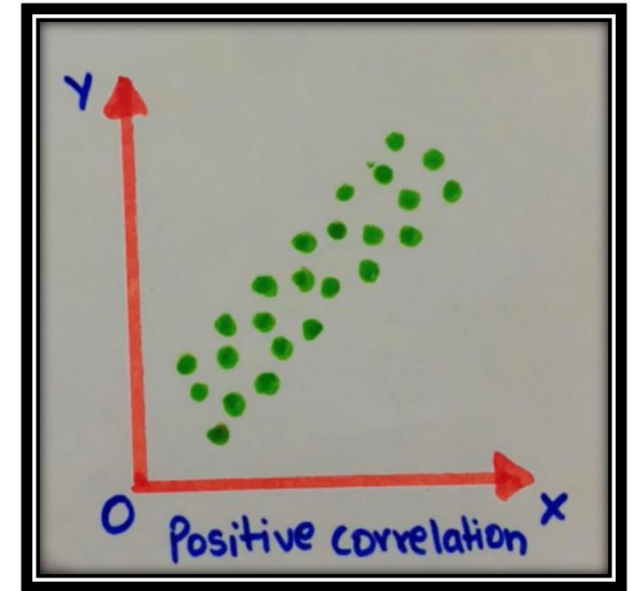
- It emphasises the importance of data visualization to observe the patterns, outliers that may be hidden in the numbers alone.
- Anscombe's Quartet is a powerful reminder that summary statistics alone do not capture the full story of the dataset

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

Ans) Pearson's R is also known as the Pearson correlation coefficient, it is a statistical measure that calculate the strength and direction of the linear relationship between two continuous variables. It is widely used in statistics to evaluate how well two variables are related with each other.

Pearson's R is a number between -1 and +1 that indicates the relationship between two variables.

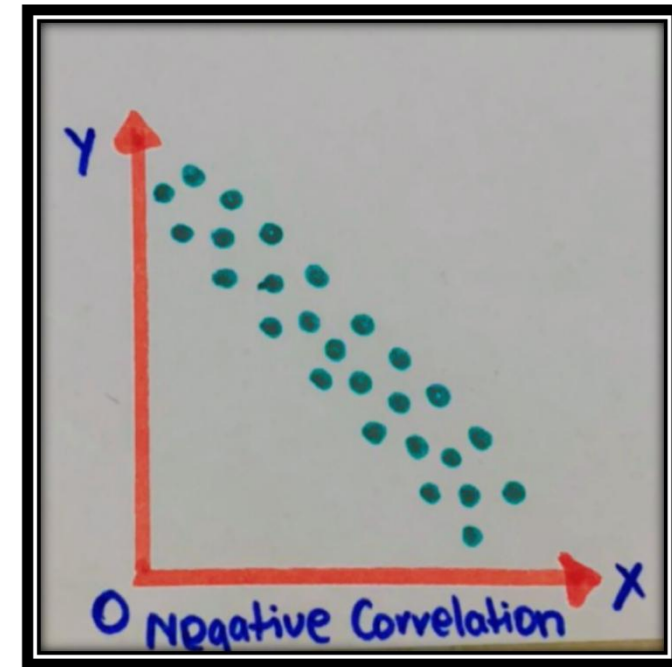
- A value of 0 indicates there is no relation between two variables.
- A value greater than 0 shows a positive relation, means the value of one variable increase so does the value of the other variable.



- A value less than 0 indicates a negative relation, means if the value of one variable increases, the value of the other variable decreases.

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

Achieving the value of +1 or -1 indicates all the data points are passing through the line of best fit, no data points show variation away from the line.



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

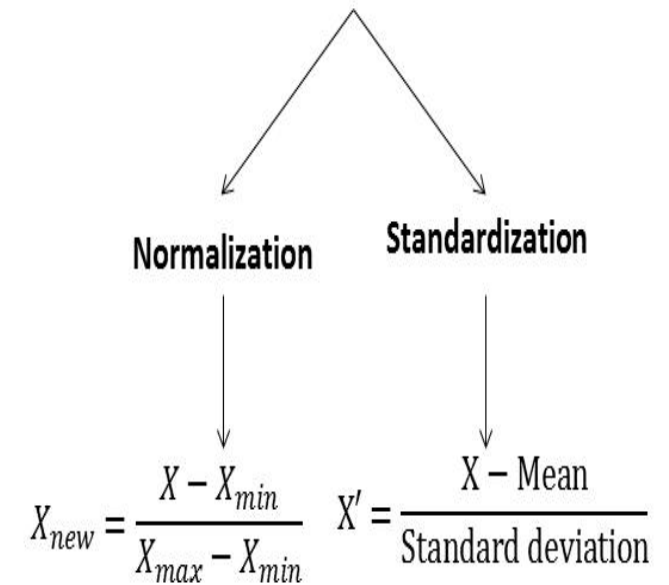
Ans) Scaling refers to the process of converting the values of the variables into specific range or distribution.

Scaling is done to ensure that different variables are on the similar scale or level, which is important for machine learning algorithms. It helps in avoiding the issues related to the stretch of the variables and can lead to better model performances.

Two commonly used scaling techniques are normalizing and standardizing:

- **Normalized Scaling (Min-Max Scaling):** this makes all the values to fit between 0 and 1.
- **Standardized Scaling (Z-Score standardization):** this makes the values have an average (mean) of 0 and a standard deviation of 1.

Feature scaling



DIFFERENCES BETWEEN NORMALIZED AND STANDARDIZED SCALING:

RANGE:

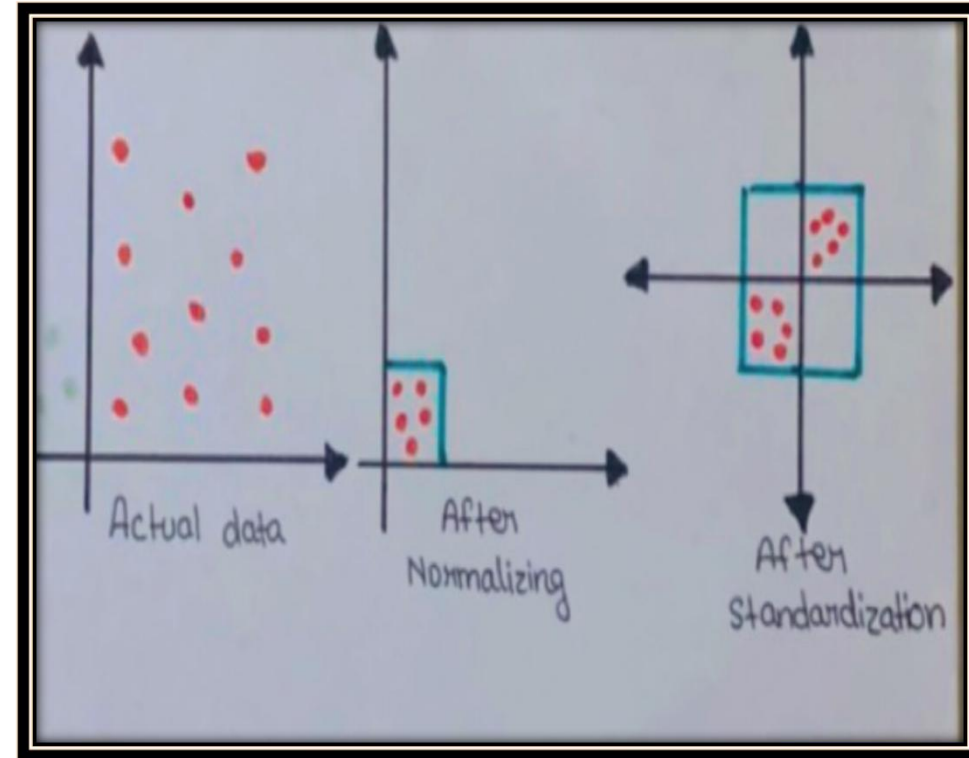
- Normalized scale: transforms data to a fixed range, usually between 0 and 1.
- Standardized scale: it transforms the data to have a mean 0 and standard deviation 1, with no particular range.

EFFECTS ON OUTLIERS:

- Normalized scale: sensitive to outliers because it uses the minimum and maximum values.
- Standardized scale: more robust towards the outliers as it depends on the mean and standard deviation

INTERPRETABILITY:

- Normalized scale: it maintains the original scale of the data which is important for some interpretations.
- Standardized scale: it shifts the data to have a mean of 0, which can make it easier to compare variables on a common scale.



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

Ans) Variance inflation factor (VIF) is used to find out the presence of multicollinearity. Sometimes, in multiple regression (predicting something using multiple variables), one or more of these variables can be predicted perfectly from the others. This perfect prediction creates the problem when we calculate VIF.

VIF is used to check if the factors are too similar and can lead to unreliable results.

The occurrence of an infinite value for the Variance Inflation Factor (VIF) is typically due to a specific relationship between the predictor variables in multiple regression model. This is called “perfect multicollinearity”, and it could lead to unstable or declined regression models.

Perfect multicollinearity occurs when one or more predictor variables in the multiple regression model can be exactly predicted from the combination of other predictor variables.

The VIF is calculated for each predictor variable by regressing it against all the other predictor variables.

In case of multicollinearity, this regression results in an extremely high R-squared value (close to 1) means that a large part of the variation in the variable is explained by some other variables.

- The formula for calculating the VIF involves dividing 1 by (1- R -squared). When R- squared is very close to 1 (due to perfect multicollinearity), the denominator becomes extremely small, approaching zero.
- Dividing 1 by an extremely small number is mathematically equivalent to dividing by zero, which results in an infinite VIF.

This indicates that some variables are not required, making it impossible to obtain reliable coefficient estimates for the affected variables in the model. To fix this issue, it is important to remove one or more unnecessary variables to find alternative ways to model the data without perfect multicollinearity.

$$VIF_i = \frac{1}{1 - R_i^2}$$

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

Ans) A Q-Q plot, short for quantile – quantile plot. It is like a graph that helps us see if our data is normal or follows a certain pattern. In linear regression it is important to know if our data behaves like we expect, (normal distribution). The Q-Q plot shows our actual data on one side and expected normal data on the other. If they make straight line, means our data is close to normal. This helps us make sure our linear regression analysis is logical, and if there are problems with our data, we can spot them using Q-Q plot.

The use and importance of a Q-Q plot in linear regression are:

- In linear regression and many other statistical analyses, it's important to make certain assumptions about the data, including assumptions that the residuals (differences between observed and predicted values) are normally distributed. Q-Q plots are a valuable tool for assessing this assumption.
- Linear regression models assume that the residuals are normally distributed. If this assumption is violated, it can affect the stability of the statistical inferences made from the model.
- Q-Q plots can be helpful in finding out the outliers in a dataset. Outliers are data points that deviate from the pattern and can affect the analysis. Q-Q plots provide a visual way to identify these outliers.
- It allows the comparison of the distribution of two datasets. This is important when assessing whether two sets of data, such as test scores from different schools before and after treatment, have significantly different distributions.

