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

   Fig1:- Boxplots of all the categorical variables.

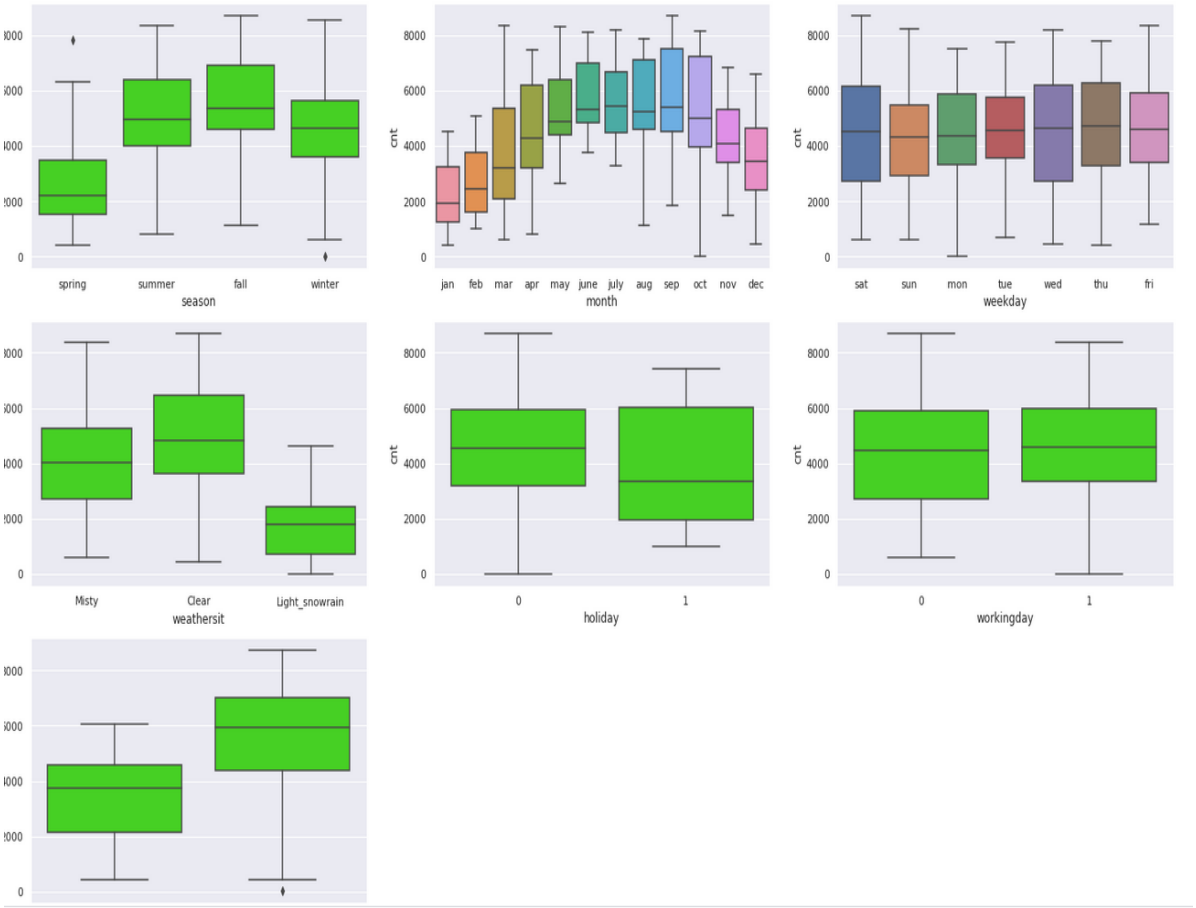
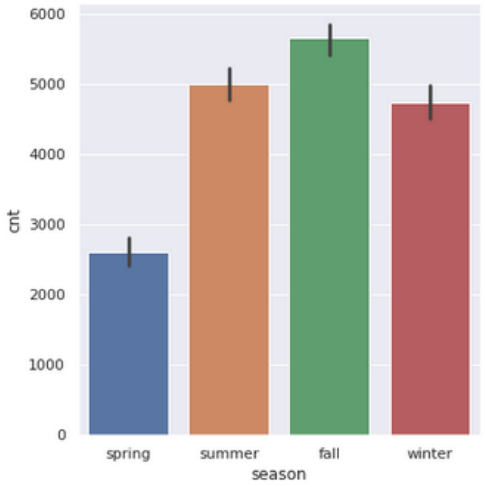
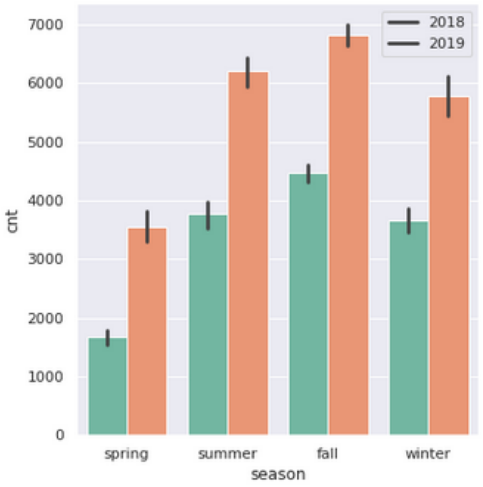
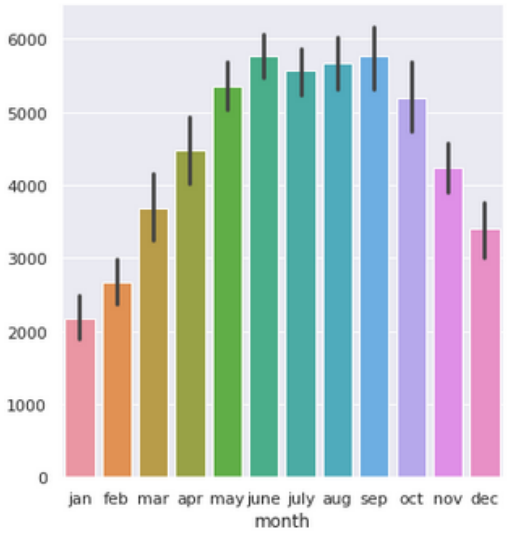
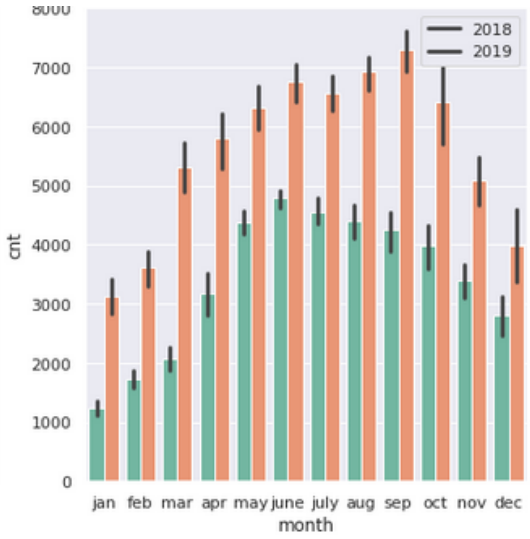


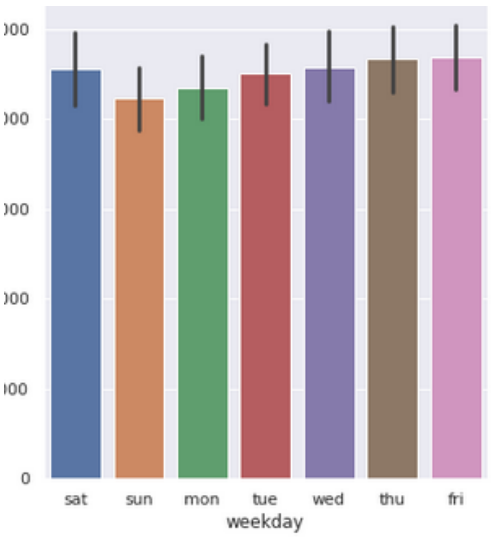
   Fig 2 – 5 Barplots of categorical variables with year comparisons

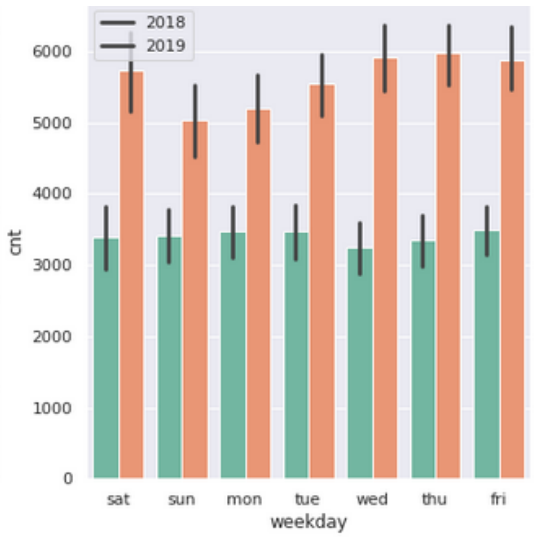

cnt
season
spring
summer
fall
winter

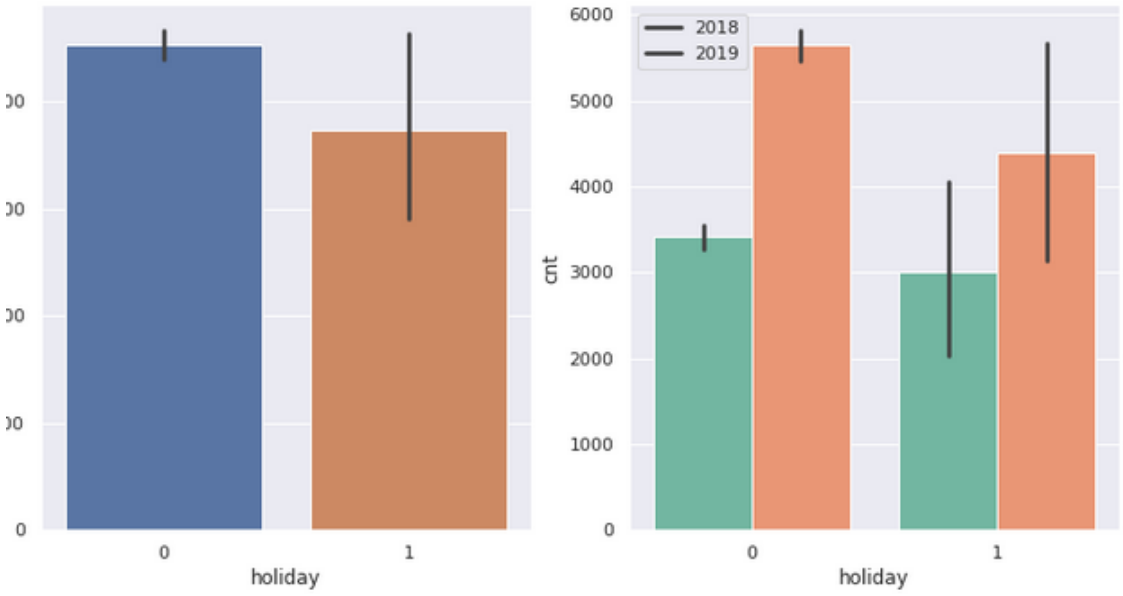

2018
2019
cnt
season
spring
summer
fall
winter


month
jan
feb
mar
apr
may
june
july
aug
sep
oct
nov
dec


2018
2019
cnt
month
jan
feb
mar
apr
may
june
july
aug
sep
oct
nov
dec


weekday
sat
sun
mon
tue
wed
thu
fri


2018
2019
cnt
weekday
sat
sun
mon
tue
wed
thu
fri



The categorical variable used in the dataset are season , year , holiday, weekday ,workingday, and weathersit(weather) and month . These were visualized using a boxplot.These variables had the following effect on our dependant variable: -

1. Season - For the variable season, we can clearly see that the category Fall, has the highest median, which shows that the demand was high during this season. It is least for 1: spring.

2.Year - The year 2019 had a higher count of users as compared to the year 2018.

3. Holiday - rentals reduced during holiday.

4. Weekday - The bike demand is almost constant throughout the week.

5. Workingday – From the "Workingday" boxplot we can see those maximum bookings happening between 4000 and 6000, that is the median count of users is constant almost throughout the week. There is not much of difference in booking whether its working day or not.

6. Weathersit - There are no users when there is heavy rain/ snow indicating that this weather is quite adverse. Highest count was seen when the weather situation was Clear, Partly Cloudy.

7. Month - The number of rentals peaked in September, whereas they peaked in December. This observation is consistent with the observations made regarding the weather. As a result of the typical substantial snowfall in December, rentals may have declined

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

Answer

There are mainly 2 reasons to use drop_first:-

- 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.
- If we don't drop the first column then the dummy variables will be correlated (redundant). This may affect some models adversely and the effect is stronger when the cardinality is smaller. This will increase the multicollinearity. For instance lets look at 2 equations :-

  D1 + D2 + D3 = 1
  D3 = 1-(D1+D2) – ii

  Here D1, D2 and D3 are dummy variables.
  The last equation indicates D3 is perfectly explained by the other two dummy variables D1 and D2.

  If we don’t eliminated highly correlated dummy variables , our model creation will be trapped in Dummy Variable Trap.

Q3 Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer:-

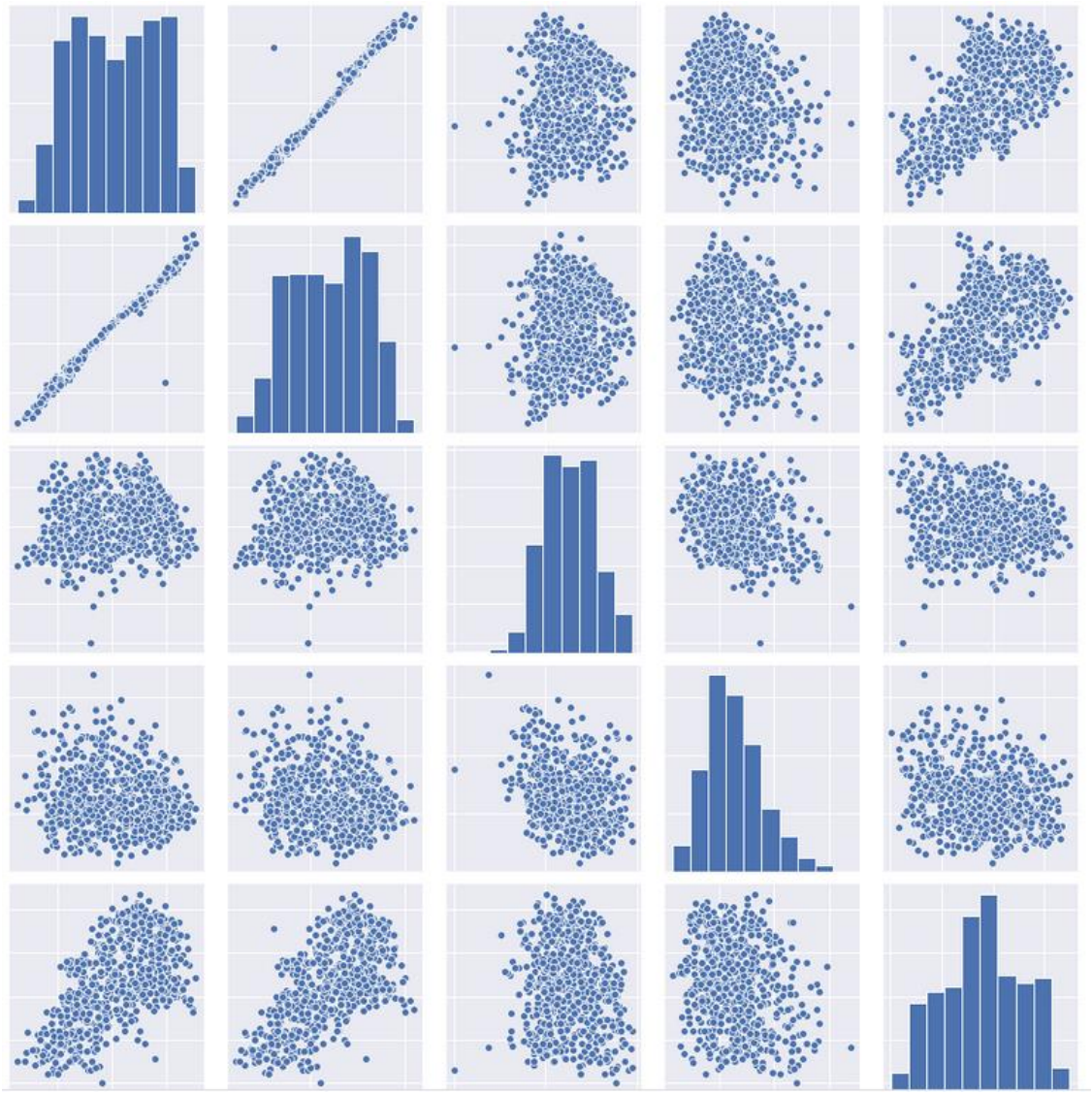


Fig 6:- Pair plots of all numerical variables.

As we can see from above ‘temp’ variable has the highest correlation with the target variable as well as with atemp.

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

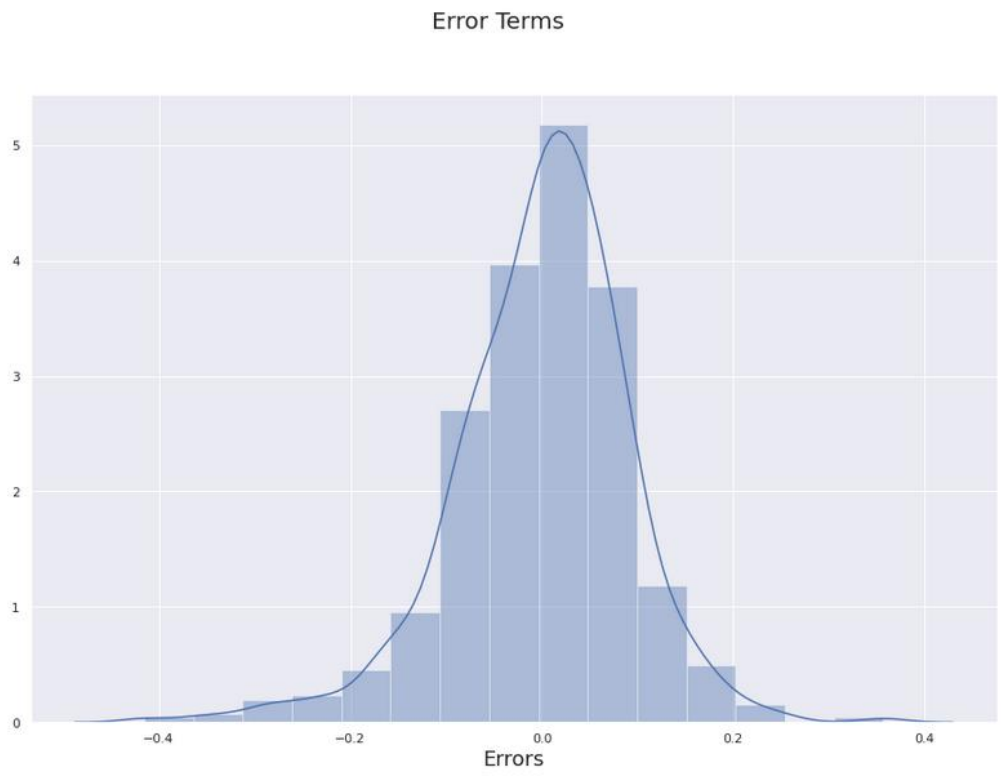


Fig 7 :- Residual analysis of training data

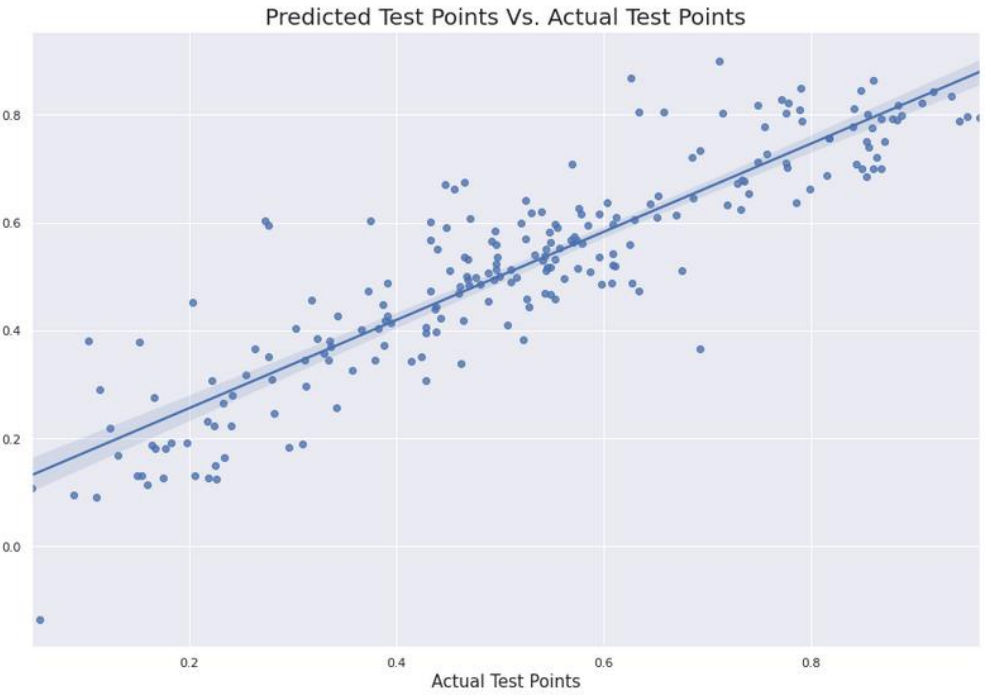


Fig 8:- Predicting Test points vs Actual test points

The assumptions of Linear Regression can be validated as below :-

1. Error terms should be normally distributed as can be seen in fig 7.

2. There should be insignificant multicollinearity between variables.
3. The model must be a linear equation which on test and prediction data , be able to show linearity between variables .
4. There should be no visible pattern in residual values. The model should be a good fit.
5. Independence of residuals by having no auto correlation . The errors and independent variables are uncorrelated.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer :-

1. temp :- It has a positive coefficient of 0.49088
2. year :- It has a positive coefficient of 0.233570
3. weathersit_Light_snowrain :- it has a negative coefficient of 0.284199

General Subjective Questions

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

Answer:-

Linear regression is a statistical technique used for finding the existence of an association relationship between a dependent variable (aka response variable or outcome variable) and an independent variable(predictor variable).

Linear regression is a machine learning algorithm based on supervised learning.

We can only establish the change in the value of the outcome variable(Y) is associated with change in the value of the feature X, i.e , regression technique cannot be used for establishing causal relationship between two variables.

Regression is one of the most popular supervised learning algorithms in predictive analysis.

For instance :- A hospital may be interested in finding the how the total cost of a patient for a treatment varies with the body weight of the patient.

Simple linear regression has only one feature variable and the relationship between the outcome variable and the regression coefficient is linear .

$$Y = \beta 0 + \beta 1X + \varepsilon$$

Here, Y= outcome/dependent variable

β0 = Y intercept

β1 = slope of the regression line, this reflects the effect of X on Y

X = feature variable used to make predictions

To find the best values of β0 and β1, we need to define a cost function that measures how well the line fits the data. A common choice is the mean

squared error (MSE), which is the average of the squared differences between the actual y values and the predicted y values:

$$MSE = \left(\frac{1}{n}\right) * \sum (y - y`)\text{^}2$$

Where , n = number of data points

Y = actual value

Y` = predicated value

The goal is to minimize the MSE by adjustinng β0 and β1. There are different methods to do this, such as gradient descent, normal equation, or using libraries like scikit-learn.

Linear regression can also be extended to multiple input variables (x1, x2, …, xn), in which case the equation becomes:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 \ldots\ldots \beta kXk + \varepsilon$$

Limitations :

It assumes a linear relationship between the input variables and the output variable, which may not always be the case.

It may be sensitive to outliers or multicollinearity

Assumptions :-

1. The errors are assumed to follow a normal distribution.
2. The variance of error is constant for various values of independent variables X.
3. The errors and independent variables are uncorrelated .

2.Explain the Anscombe's quartet in detail.

Answer:-

Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four datasets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to emphasize both the importance of graphing data before analysing it and the effect of outliers and other influential observations on statistical properties.

Let us assume that for 4 datasets the below holds true :-

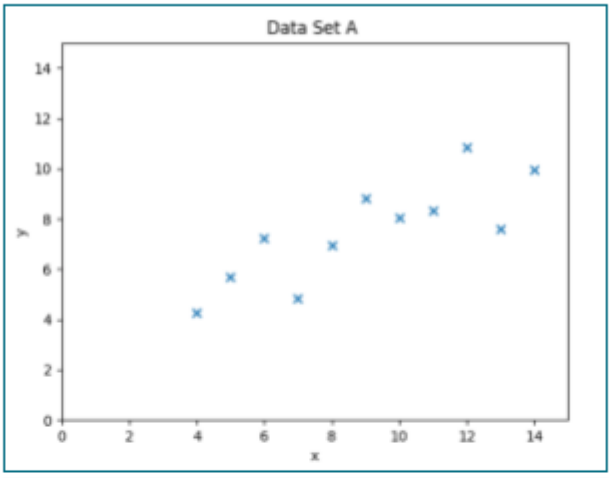
1) Mean of x values in each data set = 9.00
2) Standard deviation of x values in each data set = 3.32
3) Mean of y values in each data set = 7.50
4) Standard deviation of x values in each data set = 2.03
5) Pearson's Correlation coefficient for each paired data set = 0.82
6) Linear regression line for each paired data set: y = 0.500x + 3.00

When looking at this data we would be forgiven for concluding that these data sets must be very similar – but really they are quite different.

Data Set A:

x = [10,8,13,9,11,14,6,4,12,7,5]

y = [8.04, 6.95,7.58,8.81,8.33, 9.96,7.24,4.26,10.84,4.82,5.68]

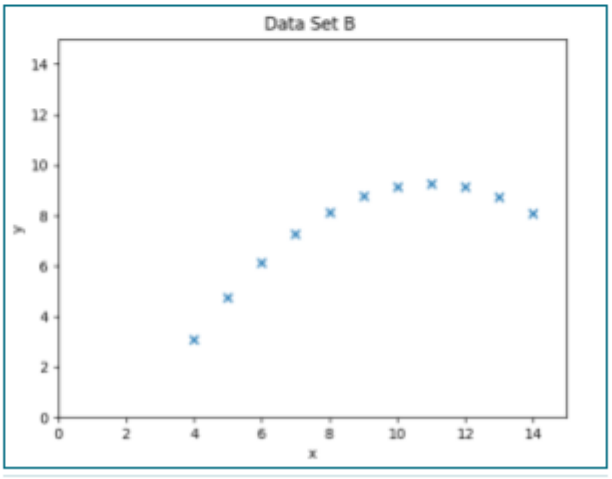


Data Set A does indeed fit a linear regression – and so this would be appropriate to use the line of best fit for predictive purposes.

Data Set B:

x = [10,8,13,9,11,14,6,4,12,7,5]

y = [9.14,8.14,8.74,8.77,9.26,8.1,6.13,3.1,9.13,7.26,4.74]

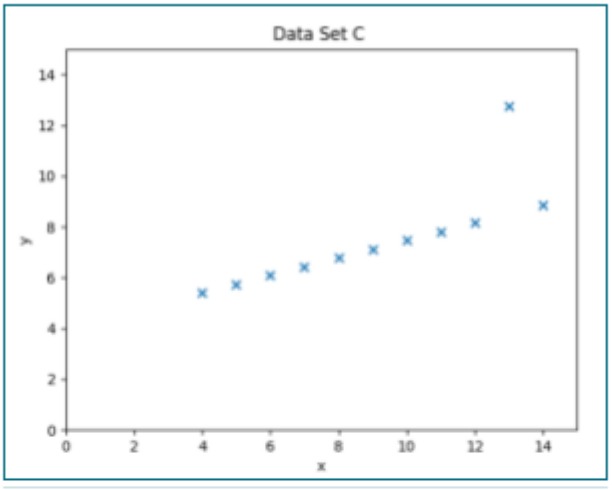


You could fit a linear regression to Data Set B – but this is clearly not the most appropriate regression line for this data. Some quadratic or higher power polynomial would be better for predicting data here.

Data Set C:

x = [10,8,13,9,11,14,6,4,12,7,5]

y = [7.46,6.77,12.74,7.11,7.81,8.84,6.08,5.39,8.15,6.42,5.73]



In Data set C we can see the effect of a single outlier – we have 11 points in pretty much a perfect linear correlation, and then a single outlier. For predictive purposes we would be best investigating this outlier (checking that it does conform to the mathematical definition of an outlier), and then potentially doing our regression with this removed.

Data Set D:

x = [8,8,8,8,8,8,8,19,8,8,8]

y = [6.58,5.76,7.71,8.84,8.47,7.04,5.25,12.50,5.56,7.91,6.89]

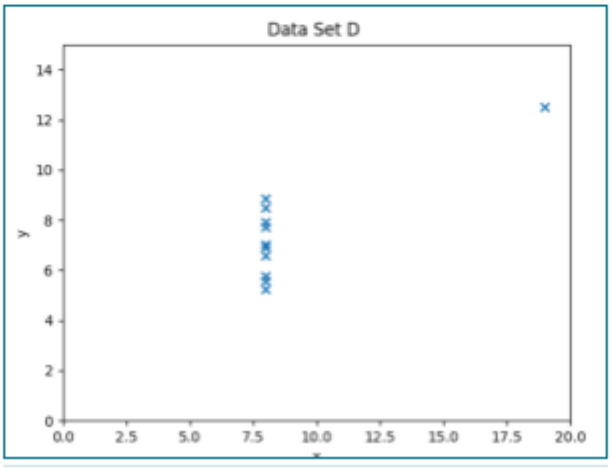


Fig 9:- Each graph plot shows the different behavior irrespective of statistical analysis.

This tells us about the importance of visualising 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 you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.

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. It value ranges between -1 to +1. It shows the linear relationship between two sets of data. In simple terms, it tells us "can we draw a line graph to represent the data?"

English mathematician and statistician Karl Pearson is credited for developing many statistical techniques including Pearson Coefficient

To find the Pearson coefficient, the two variables are placed on a scatter plot. The variables are denoted as X and Y. There must be some linearity for the coefficient to be calculated; a scatter plot not depicting any resemblance to a linear relationship will be useless.

The closer the resemblance to a straight line of the scatter plot, the higher the strength of association. Numerically, the Pearson coefficient is represented the same way as a correlation coefficient that is used in linear regression, ranging from -1 to +1. A value of +1 is the result of a perfect positive relationship between two or more variables. Positive correlations indicate that both variables move in the same direction. Conversely, a value of -1 represents a perfect negative relationship. Negative correlations indicate that as one variable increases, the other decreases; they are inversely related. A zero indicates no correlation.

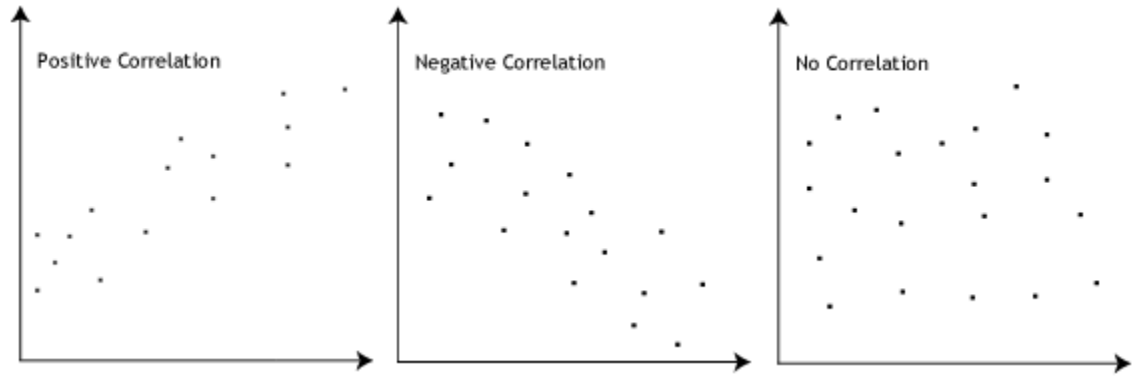


Fig 10 :- Shows all kinds of correlation.

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

Answer:-

Feature scaling is a method used to normalize or standardize the range of independent variables of data. It is performed during the data preprocessing stage to deal with varying values in the dataset. 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, irrespective of the units of the values.

There are generally 2 kinds of scaling

1. Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.

2. Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization,standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus,tackle this issue.

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

Answer**:-**

VIF - Variance Inflation Factor

The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity. If there is perfect correlation, then VIF = infinity. It gives a basic quantitative idea about how much the feature variables are correlated with each other. It is an extremely important parameter to test our linear model.

$$VIF = \frac{1}{1 - R\text{^}2}$$

Where R-1 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables. If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and

it's R-squared value will be equal to 1. So, VIF = 1/(1-1) which gives VIF = 1/0 which results in"infinity".

Therefore, an infinite value of VIF for a given independent variable indicates that it can be perfectly predicted by other variables in the model.

A rule of thumb for interpreting the variance inflation factor:

1 = not correlated.

Between 1 and 5 = moderately correlated.

Greater than 5 = highly correlated.

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

Answer:-

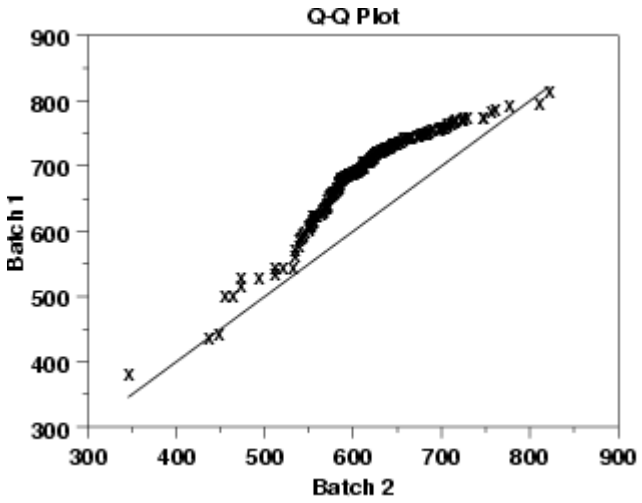
The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

The advantages of the q-q plot are:

1. The sample sizes do not need to be equal.
2. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.



1. These 2 batches do not appear to have come from populations with a common distribution.
2. The batch 1 values are significantly higher than the corresponding batch 2 values.
3. The differences are increasing from values 525 to 625. Then the values for the 2 batches get closer again.