Data visualization

- Data visualization is the practice of translating information into a visual context, such as a map or graph, to make data easier for the human brain to understand and pull insights from.
- The main goal of data visualization is to make it easier to identify patterns, trends and outliers in large data sets.
- The term is often used interchangeably with others, including information graphics, information visualization and statistical graphics.

Data visualization

- Data visualization is one of the steps of the data science process, which states that after data has been collected, processed and modeled, it must be visualized for conclusions to be made.
- Data visualization is also an element of the broader data presentation architecture (DPA) discipline, which aims to identify, locate, manipulate, format and deliver data in the most efficient way possible.

Benefits of Data Visualization

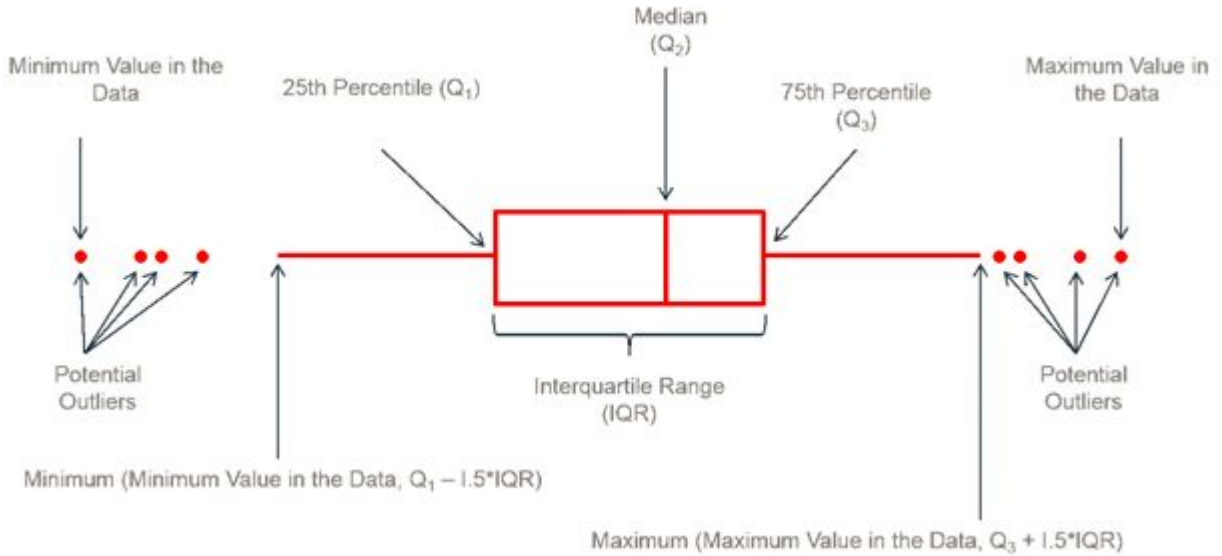
1. Data visualization helps data scientists communicate insights to stakeholders in an easy-to-understand format
2. A visual overview of information is more manageable to identify. It makes it easier to see the trends and patterns of quantitative data.
3. It is definitely faster to gather some insights from the data using a data visualization rather than just studying a chart.
4. The ability to absorb information quickly, improve insights and make faster decisions;
5. An increased understanding of the next steps that must be taken to improve the organization;
6. An easy distribution of information that increases the opportunity to share insights with everyone involved;

Box Plot

A box plot also known as Five Number Summary, summarizes data using the median, upper quartile, lower quartile, and the minimum and maximum values. It allows you to see important characteristics of the data at a glance(visually). This also help us to visualize outliers in the data set.

BOX PLOTS/ box-and-whisker plots:

A box-plot is a very useful and standardized way of displaying the distribution of data based on a five-number summary (minimum, first quartile, second quartile(median), third quartile, maximum). It helps in understanding these parameters of the distribution of data and is extremely helpful in detecting outliers.



A drawback of the box plot is that it is not effective at identifying distributions that have multiple peaks or modes.

Working Example of Box

Let's understand Box plot with this an example.

Step 1 — take the set of numbers given

14, 19, 100, 27, 54, 52, 93, 50, 61, 87,68, 85, 75, 82, 95

Arrange the data in increasing(ascending) order

14, 19, 27, 50, 52, 54, 61, 68, 75, 82, 85, 87, 93, 95, 100

Step 2 — Find the median of this data set. Median is mid value in this ordered data set.

14, 19, 27, 50, 52, 54, 61, **68**, 75, 82, 85, 87, 93, 95, 100

Here it is 68.

Step 3 — Lets find the Lower Quartile.

Lower Quartile is the median from the left of the, medium found in the Step 2(ie. 68)

(14, 19, 27, **50**, 52, 54, 61), 68, 75, 82, 85, 87, 93, 95, 100

Lower Quartile is 50

Step 4 — Lets find the Upper Quartile.

Upper Quartile is the median from the Right of the medium found in the Step 2(ie. 68)

14, 19, 27, 50, 52, 54, 61, 68,(75, 82, 85, **87**, 93, 95, 100)

Upper Quartile is 87

Step 5 — Lets find the Minimum Value

It is value lies in the extreme left from this data set or first value in the data set after ordering.

14, 19, 27, 50, 52, 54, 61, 68, 75, 82, 85, 87, 93, 95, 100

Minimum Value is 14

Step 6 — Lets find the Maximum Value

It is value lies in the extreme Right from this data set or last value in the data set after ordering.

14, 19, 27, 50, 52, 54, 61, 68, 75, 82, 85, 87, 93, 95, **100**

Maximum Value is 100

Median	=> 68
Lower Quartile	=> 50
Upper Quartile	=> 87
Minimum Value	=>14
Maximum Value	=>100

Range :

Range is basically spread of our data set.Range can be found as difference between Maximum Value and Minimum Values.

Range=Maximum Value-Minimum Value
Range= 100-14 = 86

Step 1 : We have 14 records below.

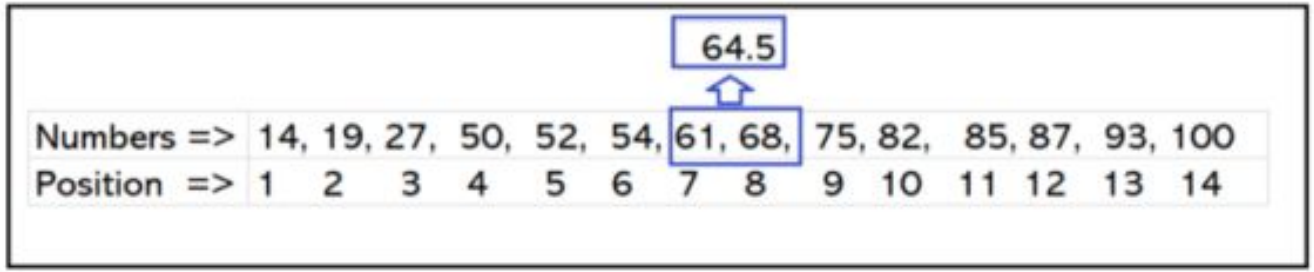
14, 19, 100, 27, 54, 52, 93, 50, 61, 87,68, 85, 75, 82

Arrange the data in increasing(ascending) order

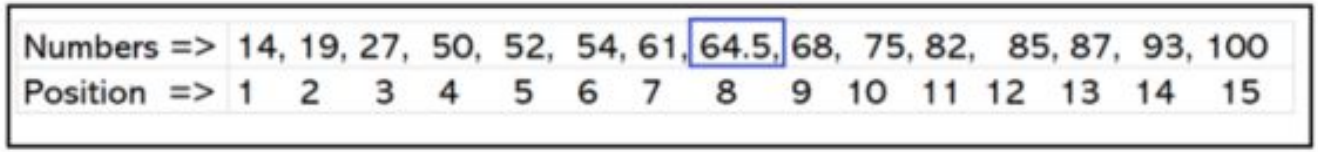
14, 19, 27, 50, 52, 54, 61, 68, 75, 82, 85, 87, 93, 100

Step 2:Since we the even number take the middle two values add them and divide them by 2.

Here Values at position 7 & 8 are middle values



So our new median value is 64.5



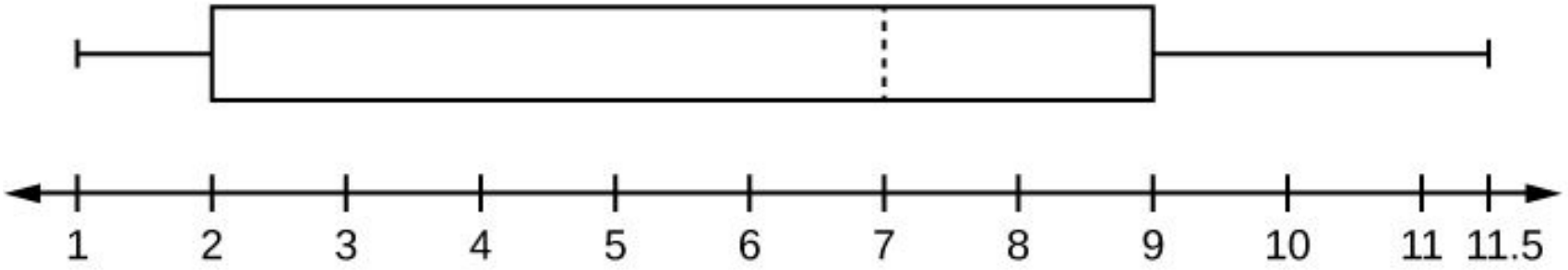
Continue Step 3 to Step 6 to get the values mention in **Working Example of Box Plot** section.Final Result as below

Median	=> 64.5
Lower Quartile	=> 50
Upper Quartile	=> 85
Minimum Value	=>14
Maximum Value	=>100

Consider, again, this dataset.

1, 1, 2, 2, 4, 6, 6.8, 7.2, 8, 8.3, 9, 10, 10, 11.5

The first quartile is two, the median is seven, and the third quartile is nine. The smallest value is one, and the largest value is 11.5. The following image shows the constructed box plot.



The two whiskers extend from the first quartile to the smallest value and from the third quartile to the largest value. The median is shown with a dashed line.

The following data are the heights of 40 students in a statistics class.

59; 60; 61; 62; 62; 63; 63; 64; 64; 64; 65; 65; 65; 65; 65; 65; 65; 65; 65; 66; 66; 67; 67; 68; 68; 69; 70; 70; 70; 70; 70; 71; 71; 73; 74; 74; 75; 77

Construct a box plot with the following properties; the calculator instructions for the minimum and maximum values as w the quartiles follow the example.

- Minimum value = 59
- Maximum value = 77
- Q1: First quartile = 64.5
- Q2: Second quartile or median= 66
- Q3: Third quartile = 70

a. Each quarter has approximately 25% of the data.

b. The spreads of the four quarters are $64.5 - 59 = 5.5$ (first quarter), $66 - 64.5 = 1.5$ (second quarter), $70 - 66 = 4$ (third quarter), and $77 - 70 = 7$ (fourth quarter). So, the second quarter has the smallest spread and the fourth quarter has the largest spread.

c. Range = maximum value – the minimum value = 77 – 59 = 18

d. Interquartile Range: $IQR = Q_3 - Q_1 = 70 - 64.5 = 5.5$.

e. The interval $59 - 65$ has more than 25% of the data so it has more data in it than the interval 66 through 70 which has 25 % of the data.

f. The middle 50% (middle half) of the data has a range of 5.5 inches.

Histograms

- They show the numerical variable's distribution with bars.
- The horizontal axis shows the range, while the vertical axis represents the frequency or percentage of occurrences of a range.
- To construct a histogram, the data is grouped into specific class intervals, or “bins” and plotted along the x-axis. These represent the range of the data. Then, the rectangles are constructed with their bases along the intervals for each class. The height of these rectangles is measured along the y-axis representing the frequency for each class interval.

Where is the peak of the distribution, and is the mode near the mean? Is the distribution symmetric or skewed? Where are the tails?

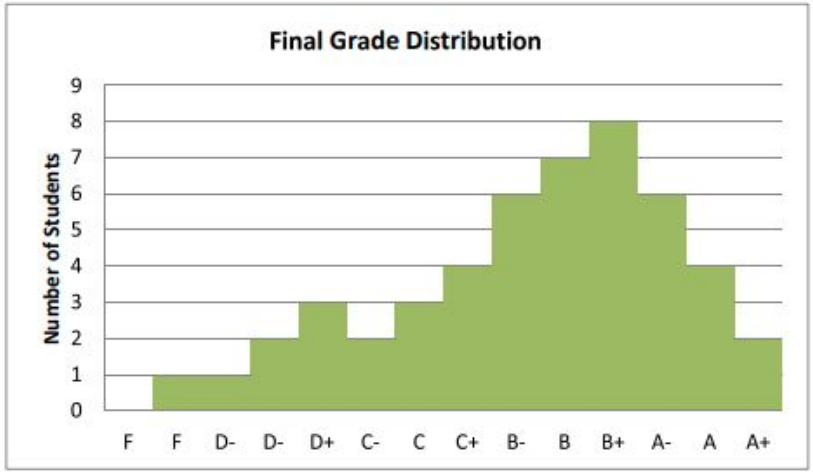


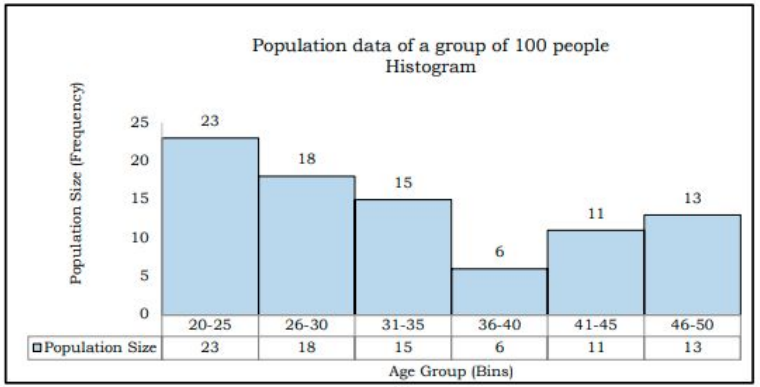
Figure 6: Histogram of Final Exam Grades

Example

Example

Histogram for the population data of a group of 86 people:

Age Group (in years)	Population Size
20-25	23
26-30	18
31-35	15
36-40	6
41-45	11
46-50	13
TOTAL	**86**



Scatter Plots

Scatter Plot :

A Scatter Plot is a graph in which the values of two variables are plotted along two axes, the pattern of the resulting points revealing any correlation present.

With scatter plots we can explain how the variables relate to each other. Which is defined as correlation. Positive, Negative, and None (no correlation) are the three types of correlation.

Limitations of a Scatter Diagram

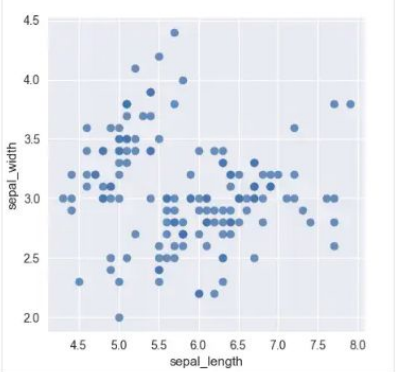
Below are the few limitations of a scatter diagram:

- With Scatter diagrams we cannot get the exact extent of correlation.
- Quantitative measure of the relationship between the variable cannot be viewed. Only shows the quantitative expression.
- The relationship can only show for two variables.

Advantages of a Scatter Diagram

Below are the few advantages of a scatter diagram:

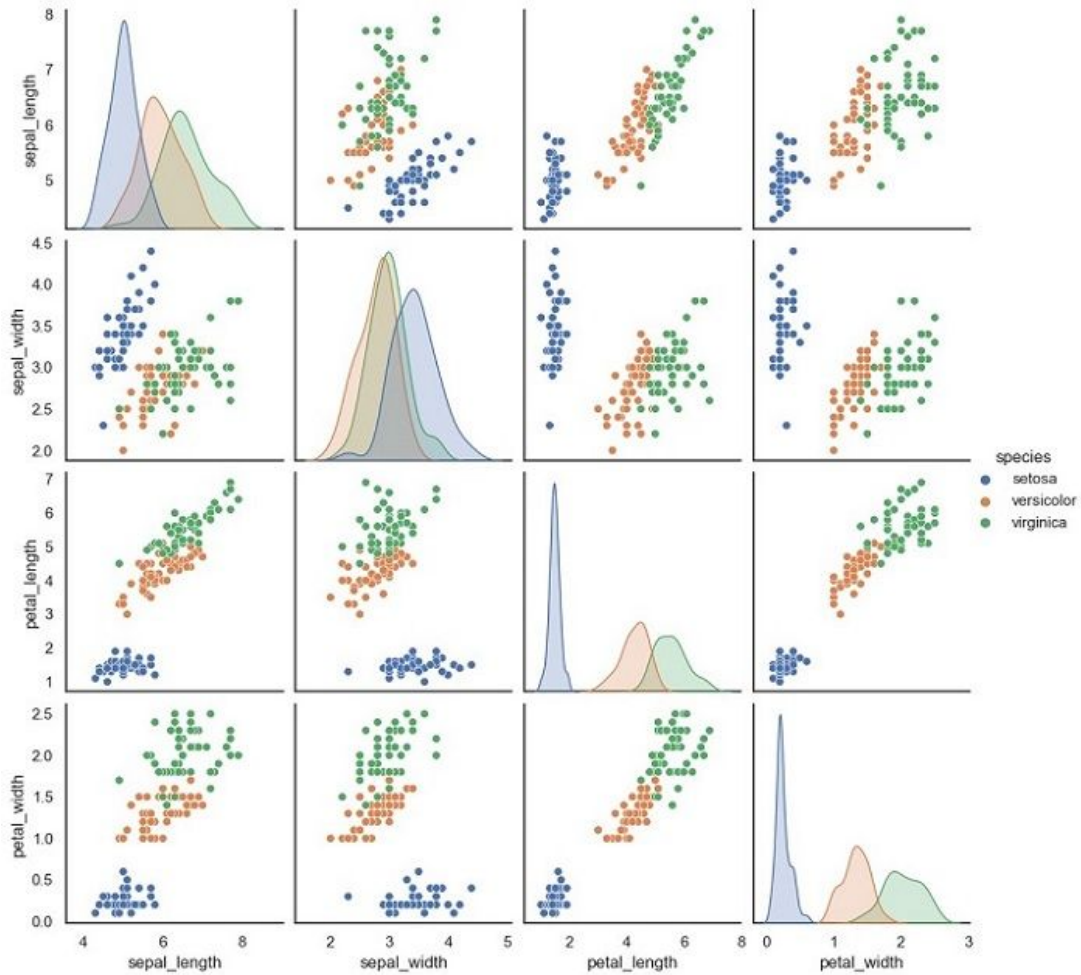
- Relationship between two variables can be viewed.
- For non-linear pattern, this is the best method.
- Maximum and minimum value, can be easily determined.
- Observation and reading is easy to understand
- Plotting the diagram is very simple.



scatter matrix

- A scatter plot matrix is a grid (or matrix) of scatter plots used to visualize bivariate relationships between combinations of variables. Each scatter plot in the matrix visualizes the relationship between a pair of variables, allowing many relationships to be explored in one chart.

- A scatter matrix consists of several pairwise scatter plots of variables presented in a matrix format. It can be used to determine whether the variables are correlated and whether the correlation is positive or negative


sepal_length
sepal_width
petal_length
petal_width
species
setosa
versicolor
virginica