Data Visualization - Matplotlib

Agenda

Key Takeaways-

- What is Data Visualization?
- Benefits of Data Visualization
- Matplotlib library and its features
- Box Plot
- Histogram
- Bar Chart
- Scatter Plot
- Line Chart

Data Visualization

Data Visualization is the technique to represent the data/information in a pictorial or graphical format. It enables the stakeholders and decision makers to analyse and explore the data visually and uncover deep insights.

“A Picture is worth a Thousand words.“

Benefits of Data Visualization

- It helps in data analysis, data exploration and makes the data more understandable.
- It identifies the relationships/correlations between the variables.
- It helps in discovering latest trends, hidden patterns in the data.
- Summarises the complex quantitative information in a small space.
- It helps in examining the areas that need attention or improvement.

Python libraries for Data Visualization

- Matplotlib
- Seaborn
- Plotly
- Bokeh
- Altair

Matplotlib

The concept of Matplotlib came from MATLAB(another programming language).It replicates MATLAB's plotting capabilities in Python.

Matplotlib is the most popular data visualization library in Python. It is used to generate simple yet powerful visualizations.

Matplotlib Features

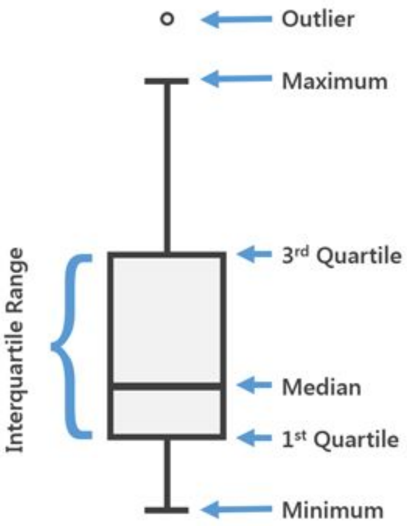
- It supports all popular graphical representations like Bar charts, Histograms, Line charts, Scatter plots, Box plots, etc.

- It is built on top of NumPy so it is fast and efficient.
- Very customizable in general(Support for custom labels and texts).
- It provides high-quality graphics output in many formats.
- It provides full control on every element in a figure.
- It is an open source tool having large community support and cross-platform support.

Box Plot

A box plot (or box-and-whisker plot) is a standardized way to display the distribution of quantitative data based on Five-Point summary (minimum, first quartile(Q1), median(Q2), third quartile(Q3) and maximum).

The box extends from the Q1 to Q3 quartile values, whereas the whiskers extend from the edges of box to the 1.5*IQR. IQR = (Q3 - Q1)



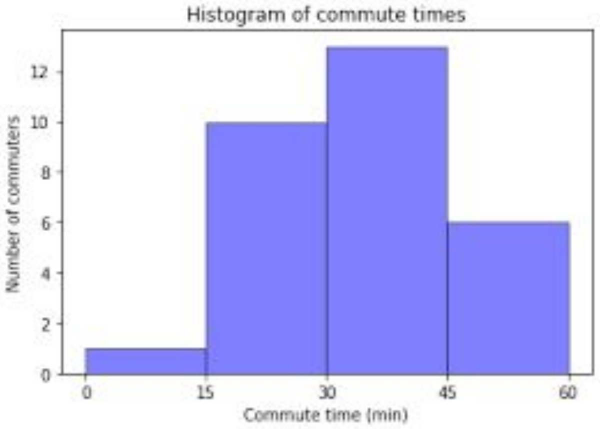
Histogram

A histogram is an accurate representation of the distribution of numerical data.

To construct a histogram, follow these steps −

- Bin (or bucket) the range of values - Divide the entire range of values into a series of intervals.
- Count how many values fall into each interval.

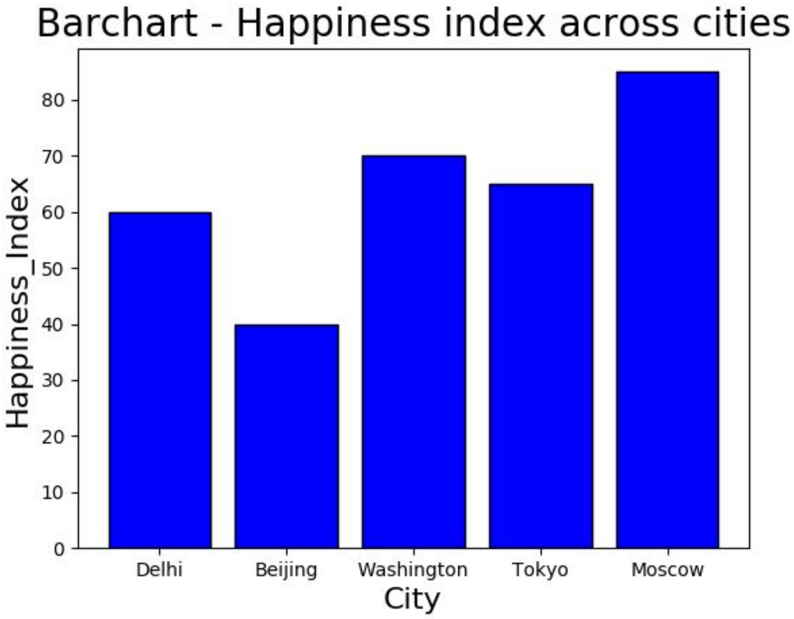
Here, X-axis is about bin ranges where Y-axis talks about frequency.



Bar Chart

A bar chart represents categorical data with rectangular bars with heights proportional to the values that they represent.

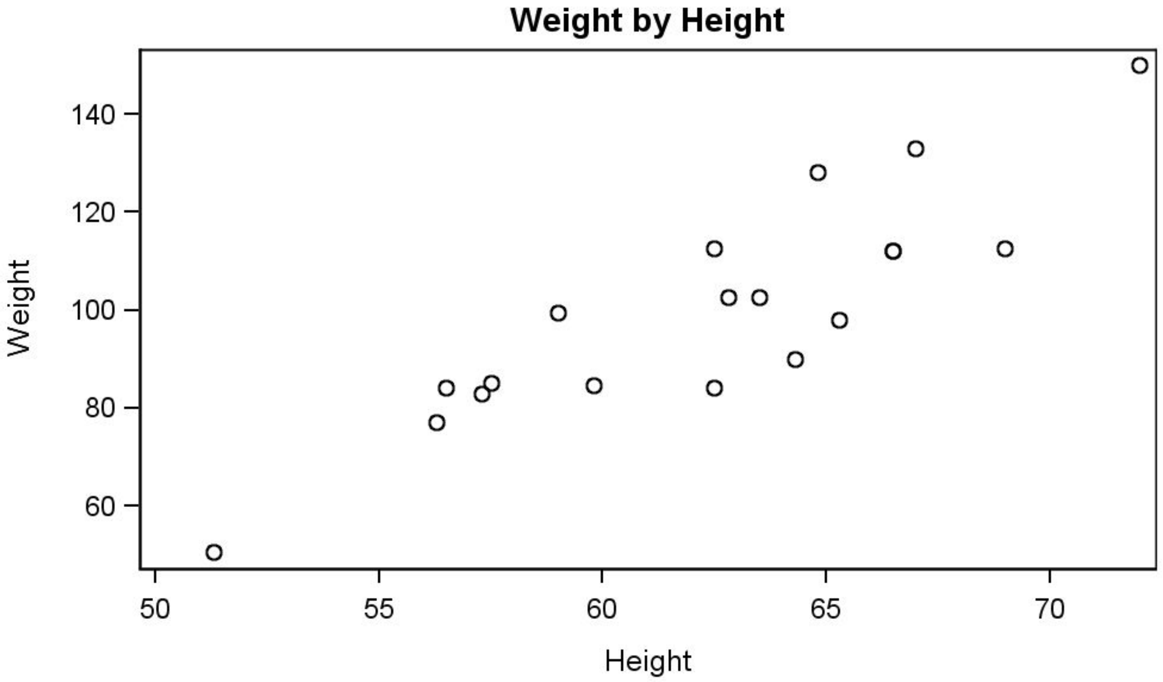
A bar plot shows comparisons among discrete categories.



Scatter Plot

A scatter plot uses dots to represent values for two different numeric variables.

It is really helpful in observing the relationship between two numeric variables.

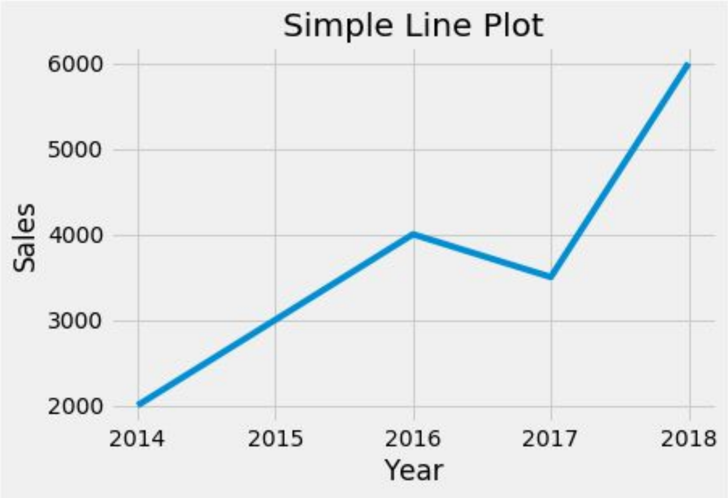


Line Chart

A line chart is a type of chart which displays information as a series of data points called ‘markers’ connected by straight line segments.

It is similar to a scatter plot except that the measurement points are ordered (typically by their x-axis value) and joined with straight line segments.

Line graphs are usually used to find relationship between two numeric variables or to visualize a trend in time series data.



Quiz 1

Choose the correct statement(s).

A. Line charts can be used to assess relationship b/w two numeric variables

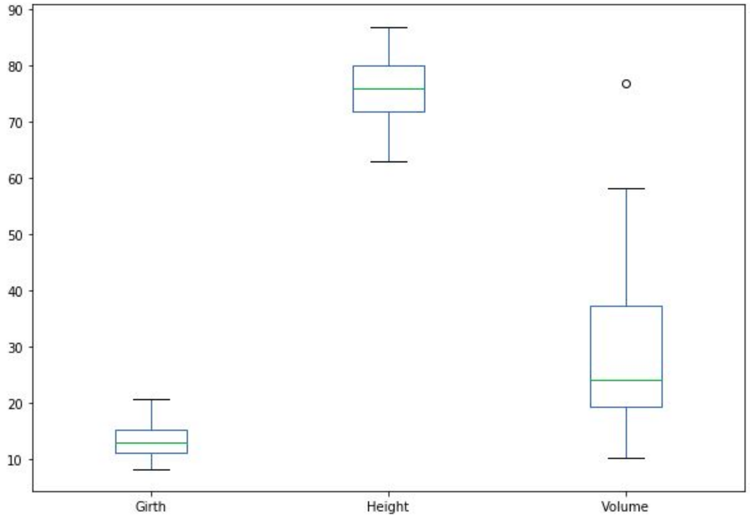
B. Histograms are best suited to non-numeric data.

C. Box Plot indicates the mean also.

D. None of the above

Quiz 2

Choose the correct statement(s).



A. Height has the highest value of 3rd quartile compared to Girth and Volume.

B. Outlier exists in the Height column

C. There is some overlap between the values of Girth and Volume.

D. None of the above