DATA VISUALIZATION USING PYTHON LIBRARIES

-Uttara Ketkar

What is data visualization (DV)?

- Data visualization refers to graphical representation of data.
- It involves gathering insights via visual representation of a dataset.
- This representation makes it easy to understand complex relationships within the data.
- Through the inferences gathered from this form of representation, manipulation and formatting of data becomes simpler.

What are the steps involved in DV?

- Initially, data is **acquired** and then divided into categories.
- The aspects of data that aren't of interest are **filtered** out.
- **Data mining** is done in order to understand patterns or mathematical relations.
- Further, a data representation **visual model**, like bar graph, is decided according to the type of data to be visualized.
- The visualization is **refined** by adding colours and layout formatting for making it user engaging.
- Lastly, **interaction** and manipulation of data is done based on observations of visual data.

(To know more about data mining : click here)

Fig 1: Steps involved in Data Visualization

Python has a number of libraries that can be used for data visualization. It is considered to be the most preferred platform for the complete process data analysis, visualization, study and formatting.

The libraries mainly used for data visualization are – **NumPy, Pandas, Matplotlib** and **Seaborn.**

NumPy:

NumPy is an open source software. This library is generally used while working with multi-dimensional arrays and matrices.

- It provides support of a large number of built-in high-level mathematical operations for array modifications.
- Though working with NumPy arrays might look similar to python lists, the NumPy array takes significantly less memory than lists.
- Implementation of mathematical functions and locating a specific element out of multi-dimensional arrays is quite simple while working with NumPy - It provides useful linear algebra, Fourier transform, and random number capabilities.
- Installation: NumPy can be installed with **conda**, with **pip**, or with a package manager on macOS and Linux:

1. If you use **conda**, you can install it with:
   conda install numpy
2. If you use **pip**, you can install it with:
   pip install numpy

Pandas:

Pandas is a software library used for data analysis and manipulation. It is an open source library. It provides easy to use data structures and analysis tools for python programming language.

- It is a free software being widely used for a number of applications in data science due to the features it offers. These features include the likes of moving window statistics and frequency conversion.
- Pandas supports various file formats - whether it is a JSON or CSV, Pandas can support it all, including Excel and HDF5. Pandas can help to merge various datasets, with extreme efficiency.

- It converts python list, dictionary and NumPy arrays into DataFrames which makes it easy to analyse. It is primarily used for data cleaning - Pandas provides a wide array of built-in tools for the purpose of reading and writing data.
- Installation: Pandas can be installed with **conda**, with **pip**, or with a package manager on macOS and Linux:

✓ If you use **conda**, you can install it with:

conda install pandas

✓ If you use **pip**, you can install it with:

pip install pandas

Cleaning data	**Unique data**	**Merging data sets**	**Reshaping and pivoting**
Alignment	**Data masking**	**Sorting**	**Data filtration**
Indexing	**Visualization**	**Grouping**	**Deletion**
Handling missing data	**Time series analysis**	**Slicing**	**Insertion**

Figure 2: Key features of Pandas

Matplotlib:

Matplotlib is a multi-platform data visualization library. It is the most commonly used library for data visualization.

- It provides an object-oriented API for plotting graphs. Everything in matplotlib is organized in a hierarchy wherein simple built-in functions are used to add and modify plot elements.
- Based on the relation between the variables being plot, matplotlib provides a wide array of plots to choose from starting from simple bar graphs to a complex 3D plot.

- This library allows us to make quality plots with only few lines of code along with additional features like managing titles, axes, margins, sub-plots, colour-palettes and much more!
- Installation: Matplotlib can be installed with **conda**, with **pip**, or with a package manager on macOS and Linux:

✓ If you use **conda**, you can install it with:

conda install matplotlib

✓ If you use **pip**, you can install it with:

pip install matplotlib

Seaborn:

Seaborn is a library for statistical visualizations in python. It provides dataset-oriented API for examining relationships between multiple variables

- It is generally used while working with categorical variables to show observations and aggregate statistics.
- It has control over matplotlib styling and figure plotting. It provides various colour palettes to make the graphs catchy and interesting.
- It provides convenient ways to visualize complex datasets. It can work on data frames and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots.
- Installation: seaborn can be installed with **conda**, with **pip**, or with a package manager on macOS and Linux:

✓ If you use **conda**, you can install it with:

conda install seaborn

✓ If you use **pip**, you can install it with:

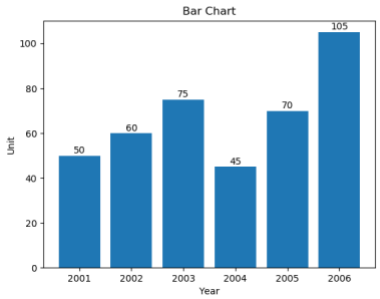
pip install seaborn

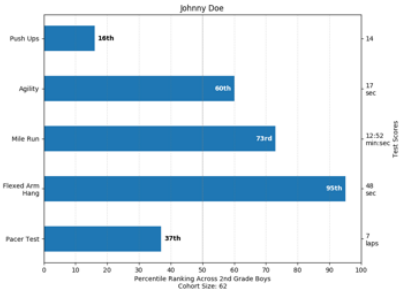
Data visualization graphs and charts that are widely used:

There are a number of plots used for DV. Each graph has certain unique features. They are widely used in various DV models based on the need of visualizations required. Multiple plot arrays can be created (using sub plots) as well as multiple plot lines can be plotted on one graph as well. Graphs titles,

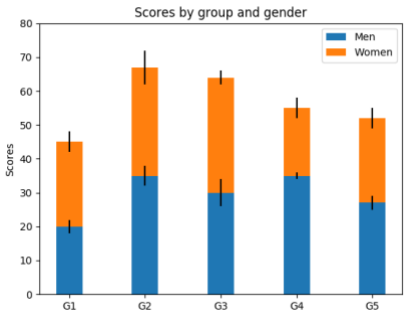
axes titles, colour palettes and shape-size of arrays can be changed to highlight certain data points as and when required.
(for further tips on pyplot refer the following link: pyplot tutorial)
Here are a few plots being used most frequently:

- Bar graph

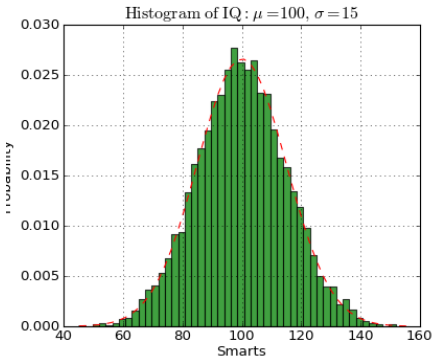




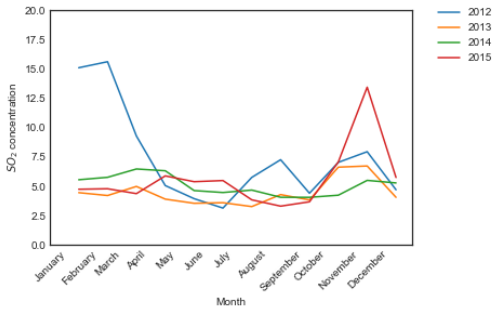
- Stacked bar graph



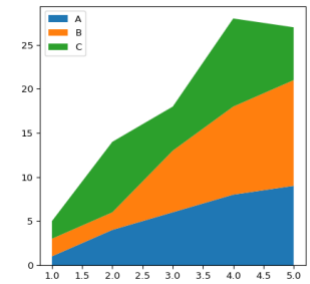
- Histogram



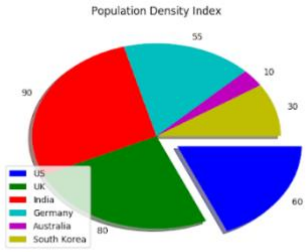
- Line graph



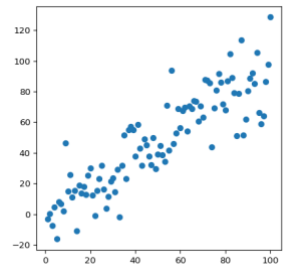
- Area chart



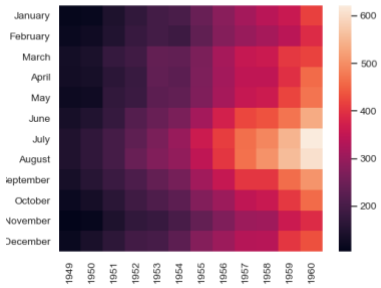
- Pie chart



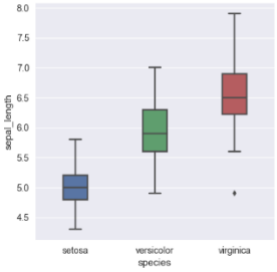
- Scatter plot



- Heat map



- Box plot



Importance of Data Visualization:

- Gives a visual summary of the data which helps to identify patterns and trends.
- It makes it easy to understand the relationship between different data points of our data set.
- The presence of null values can be found out which can be further handled by necessary changes.
- It helps detecting outliers and cleaning data.
- Identifying clusters and evaluate model output becomes easy.
- Visual analysis ensures that we don't get deceptive results and our inferences are perfect.

To Summarize:

Data visualization is an indispensable part of data analysis and machine learning. It is the collection of components that are offered by visualization libraries that make our task of error detection and result deduction simple. A wide array of tools combines to make data procurement, data mining, data cleaning, analysis, model generation and execution easy to handle!