greatlearning
Power Ahead

Data Visualization and EDA

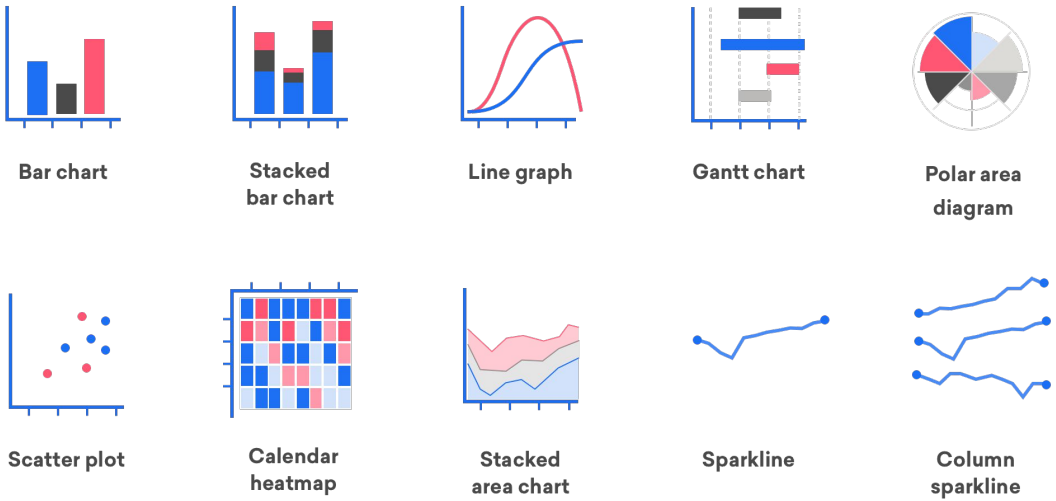
greatlearning
Power Ahead

Agenda

1. Introduction to Visualization
2. Common libraries for Visualization
3. Univariate and Multivariate Analysis
4. Pandas profiling

Introduction to Visualization

- Data visualization is the representation of data or information in a graph, chart, or other visual format to communicate relationships of the data with images.
- We need data visualization because a visual summary of information makes it easier to identify patterns and trends than looking through thousands of rows on a spreadsheet.



Source: https://morphocode.com/location-time-urban-data-visualization/

Common libraries for Visualization

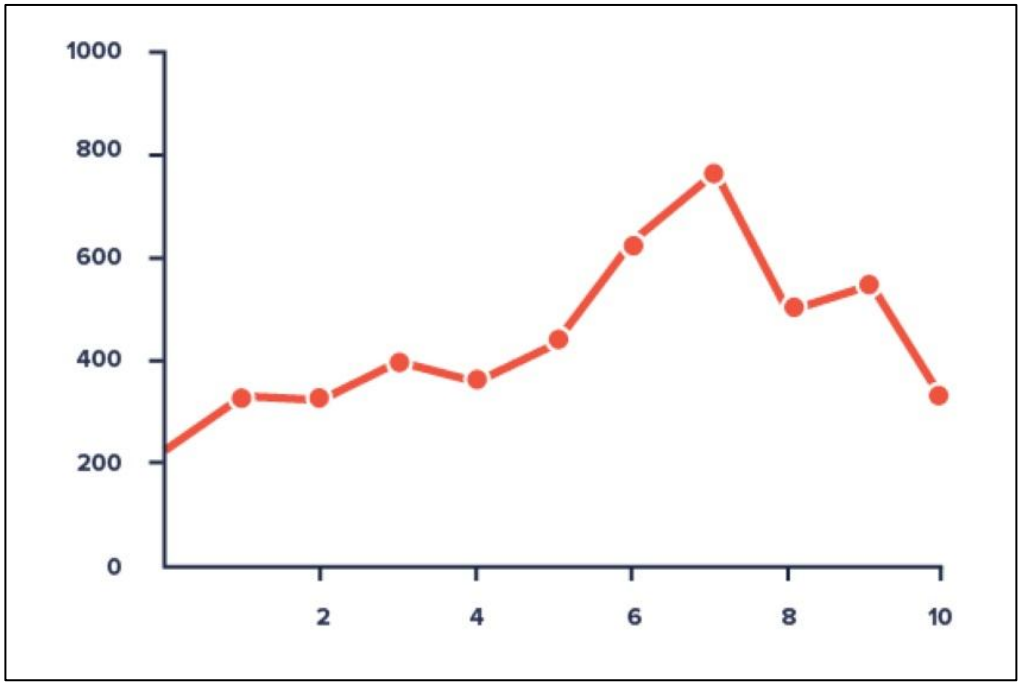
1. **Matplotlib:**
   - Matplotlib is a popular graphical subroutine and is used widely for data visualization applications.
   - It provides a context, one in which one or more plots can be drawn before the image is shown or saved to file. The context can be accessed via functions on *pyplot*.

1. **Seaborn:**
   - Seaborn is complementary to Matplotlib and it specifically targets statistical data visualization.
   - A saying around matplotlib and seaborn is, "*matplotlib tries to make easy things easy and hard things possible, seaborn tries to make a well-defined set of hard things easy too.*"

1. **Plotly:**
   - Plotly provides a web-service for hosting graphs.
   - It is mainly used for interactive visualization, dashboards etc.

Line Chart

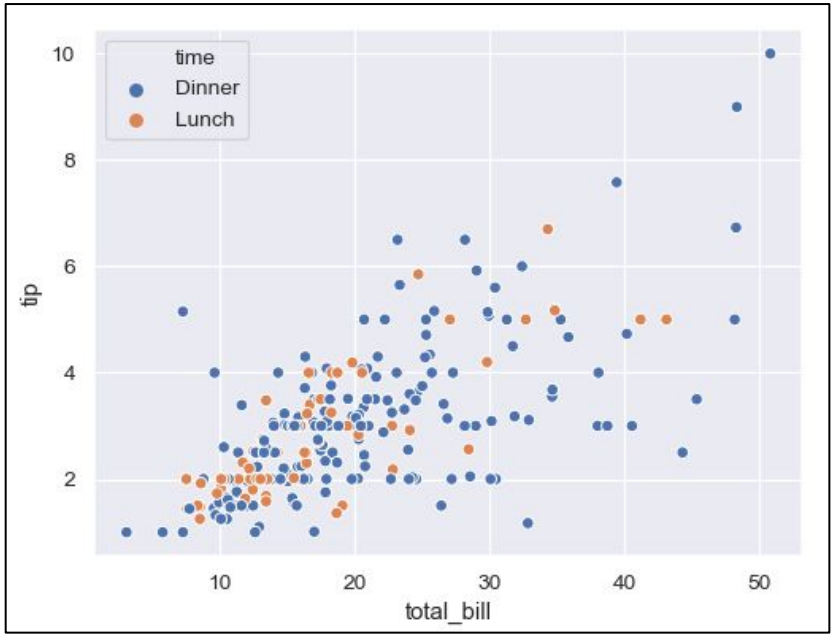
A **line graph** is a graphical display of information that changes continuously over time.



- This plot shows the relationship between the sales and the no. of days
- We can say that sales has been the highest on day 7

Scatter Plot

- A **scatter plot** uses dots to represent values for two different numeric variables.
- The position of each dot on the horizontal and vertical axis indicates values for an individual data point.
- Scatter plots are used to observe relationships between variables.



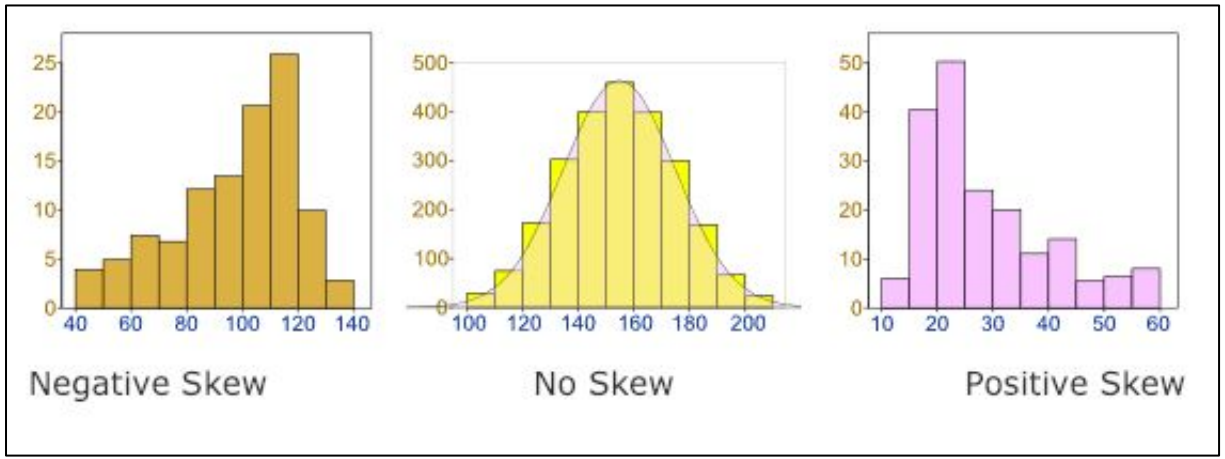
- This plot shows the relationship between the tip and the total bill at the time of lunch and dinner.
- We can say if the total bill is large, the tip can also be large

Histogram and skewness in data

- A **histogram** is a graphical display of data using bars of different heights.
- In a histogram, each bar groups numbers into ranges

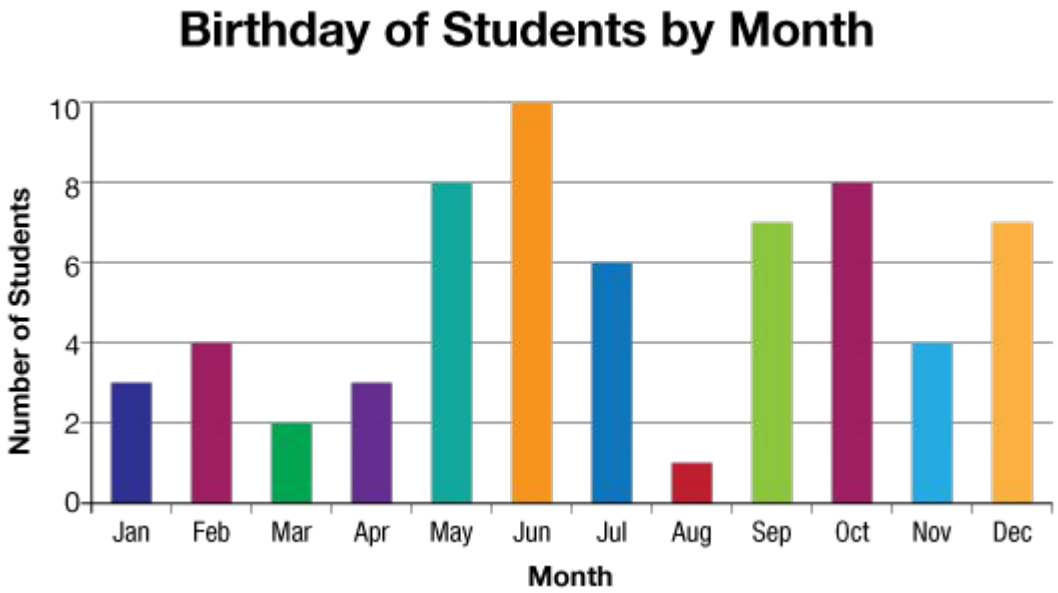
Skewness refers to distortion or asymmetry in a symmetrical bell curve in a set of data

- If the curve is shifted to the left, it is called left skewed. (leftmost curve in the below fig.)
- If the curve is shifted to the right, it is called right skewed. (rightmost curve in the below fig.)



Bar Plot

- A bar chart is a chart that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent.
- The bars can be plotted vertically or horizontally.



- Most of the students celebrated their birthday in June.
- In August, very less students celebrated their birthdays.

Chart Selection

X Variable	Y Variable	Purpose of analysis	Type of chart	Example
Continuous (numerical)	Continuous (numerical)	How Y changes with X	Scatter plot	How cholesterol varies with Age?
Continuous (numerical)	Categorical	How range of X varies for various category levels	Box plot	Cholesterol variation with Men and Women
Categorical	Categorical	What is the number or % of records of X which falls under each category	Stacked bar	How many men have heart disease compared to women?
Continuous	-	Look at the distribution of the values of the X variable	Histogram, boxplot	Distribution of cholesterol ranges
Impact of 2 X variables on a Y variable			Facet_grid()	Distribution of chol across mean and women – compared for people who have and don't have heart disease

Univariate and Multivariate Analysis

Univariate Analysis: Univariate analysis refers to the analysis of a single variable. It is a simplest form of analysis that summarizes and find patterns in the data. **Examples:** Frequency distribution, averages, measure of dispersion etc.

You have several options for describing data with univariate data:

- Frequency distribution tables
- Bar charts
- Histograms
- Frequency polygons

Multivariate Analysis: Multivariate analysis is used to study the interaction between more than one variable. Examples: Correlation, Regression analysis etc.

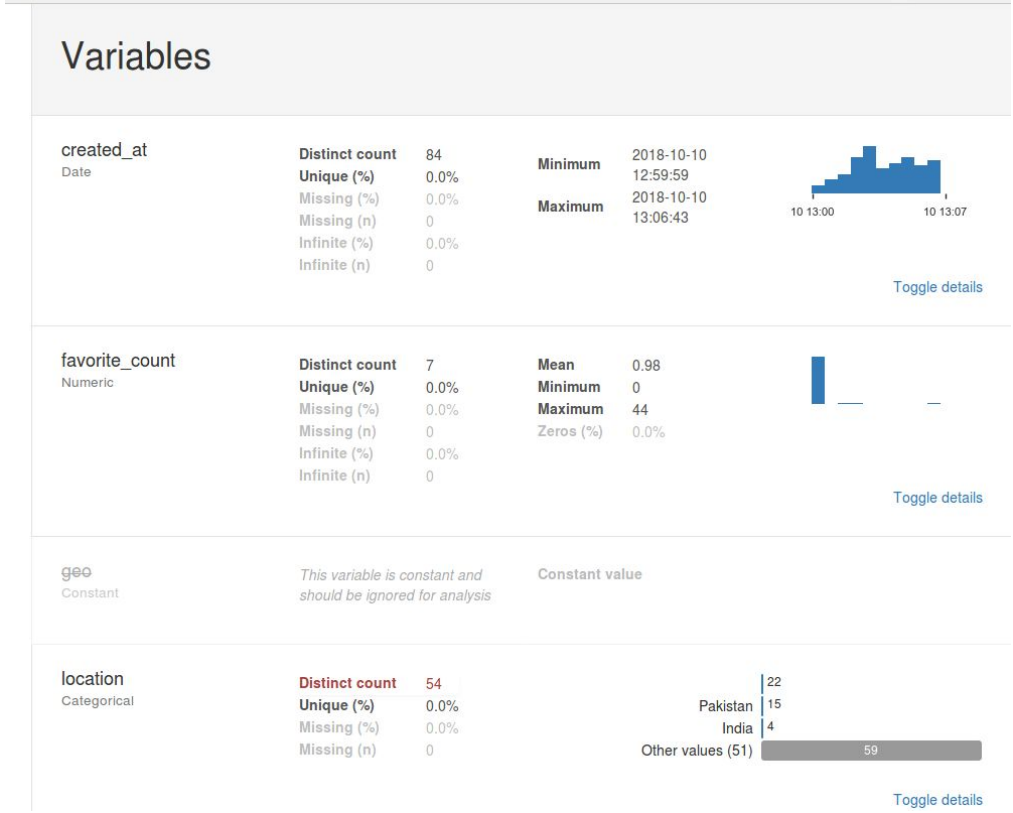
You have several options for describing data with multivariate data:

- Scatter plot
- Pair plot
- Heatmap

Pandas Profiling

- Pandas profiling is an open source python module which helps us do quick exploratory data analysis with a few lines of code.
- It saves all the work of visualizing and checking distribution of each variable.
- It generates a report with all the information available.
- The only problem with pandas profiling is working with large datasets, it takes a lot of time to generate the report.

Sample Report



greatlearning
Power Ahead

Happy Learning !