Module 02

Data Visualization

Data Science Developer

Purwadhika
Startup and Coding School

Outline

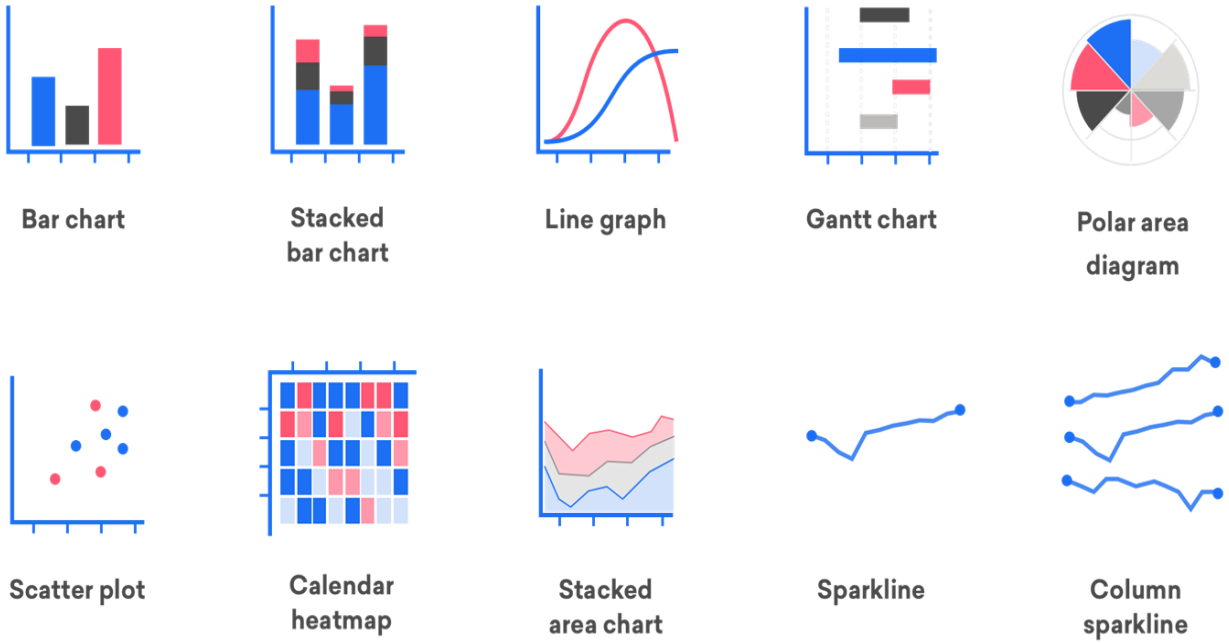
- Basic Concept of Visualization
- Type of Visualization
- Tools for visualization
- Commonly used chart

Objective

- Understand The Concept of Creating data visualization
- Answering analytics question using data visualization

Purwadhika
Startup and Coding School

What is Data Visualization ?



- Data Visualization is the presentation of data in pictorial or graphical format
- Data Visualization is a discipline to understand data by serving it visually so any **pattern or trends**, **composition**, **comparison** and **relationship** can be exposed

Purwadhika
Startup and Coding School

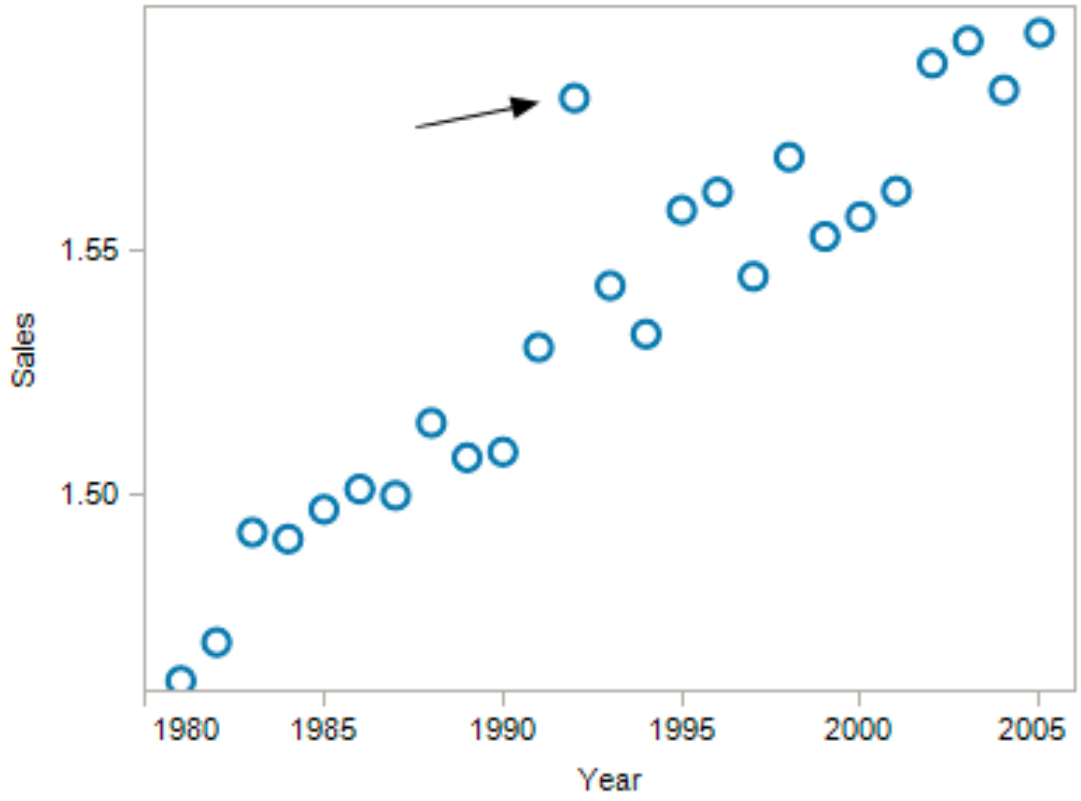
Why is Data Visualization very important ?

- You can utilize data visualization to **explore your data**. Exploring your data can improve your chance to find insight.
- You can also **share your findings** with other by using data visualization. Human brain is easier to process any information using visualization than spreadsheet or report
- Quick and easy way to convey concept in a universal manner.

Purwadhika
Startup and Coding School

Which method is easier to identify uncommon event (anomaly) between year vs sales ?

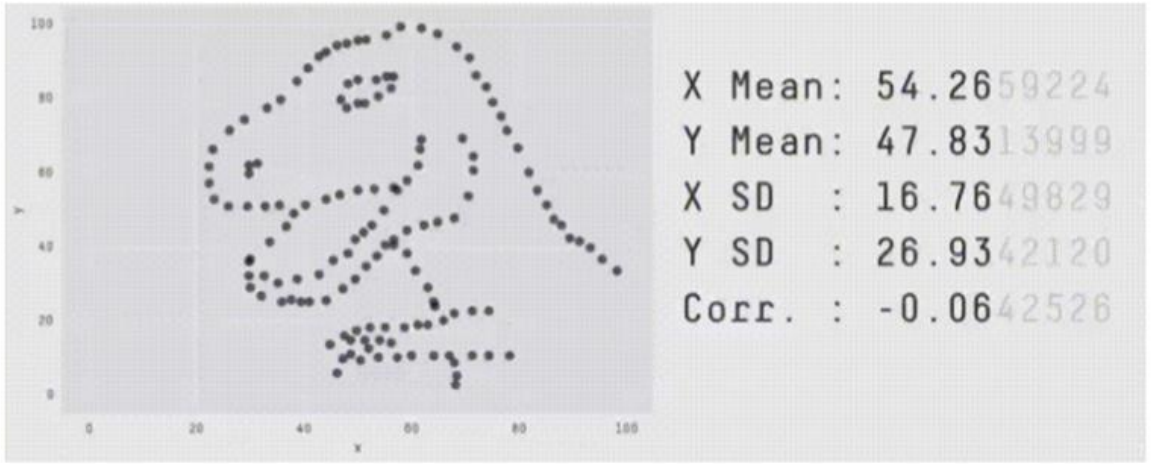
	A	B
1	Year	Sales
2	1981	1.4622
3	1982	1.47004
4	1983	1.49253
5	1984	1.49118
6	1985	1.49722
7	1986	1.50138
8	1987	1.50008
9	1988	1.51493
10	1989	1.50781
11	1990	1.50899
12	1991	1.53037
13	1992	1.58137
14	1993	1.54299
15	1994	1.53307
16	1995	1.55845
17	1996	1.56213
18	1997	1.54488
19	1998	1.56927
20	1999	1.55305
21	2000	1.5571
22	2001	1.56235
23	2002	1.58847
24	2003	1.59309
25	2004	1.58303
26	2005	1.5947



Purwadhika
Startup and Coding School

Visualization can be more meaningfull

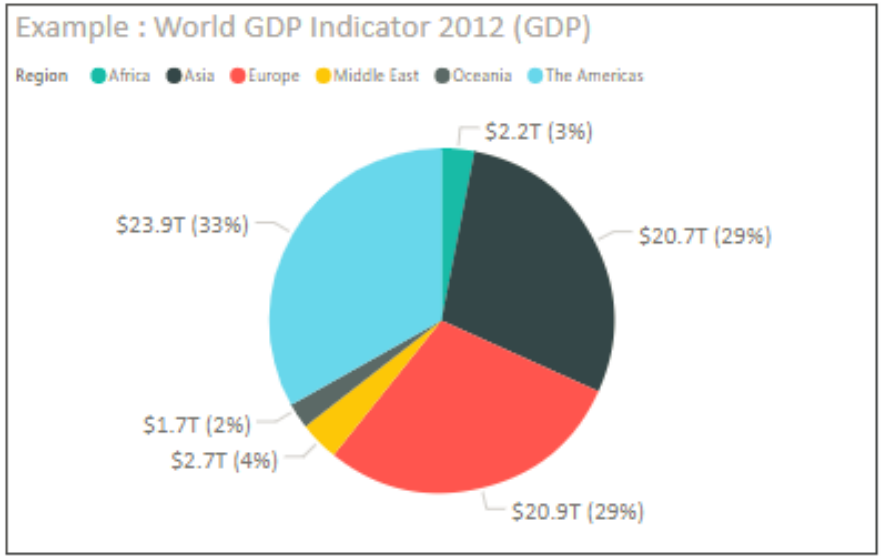
- It's hard to find patterns & derive insights from raw data
- Statistics can summarize data, but may hide patterns in how the data is spread
- We use visual encoding techniques to map data to visual attributes

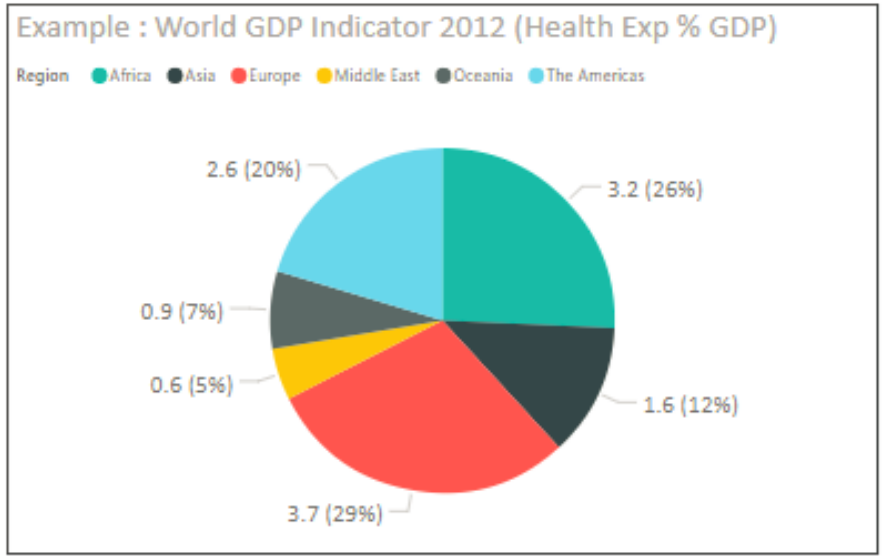


Purwadhika
Startup and Coding School

Example : World GDP Indicator 2012

Region	Population 0-14	Population 15-64	Population 65+	Population Urban	GDP	Health Exp % GDP	Health Exp/Capita
Africa	21.30	29.84	1.86	22.04	$2,232,736,028,370	3.24	$7,941
Asia	9.01	22.66	2.33	17.10	$20,689,754,411,401	1.57	$14,214
Europe	6.72	27.84	6.44	32.79	$20,930,307,966,112	3.69	$120,452
Middle East	3.50	9.01	0.49	10.22	$2,681,972,907,100	0.61	$11,424
Oceania	4.00	8.13	0.87	7.75	$1,727,175,612,930	0.92	$11,652
The Americas	9.88	24.11	3.01	29.48	$23,926,023,856,769	2.57	$33,654
Total	**54.41**	**121.58**	**15.00**	**119.38**	**$72,187,970,782,682**	**12.59**	**$199,337**





QUESTION:
What region has the highest and lowest GDP and Health Exp % GDP?

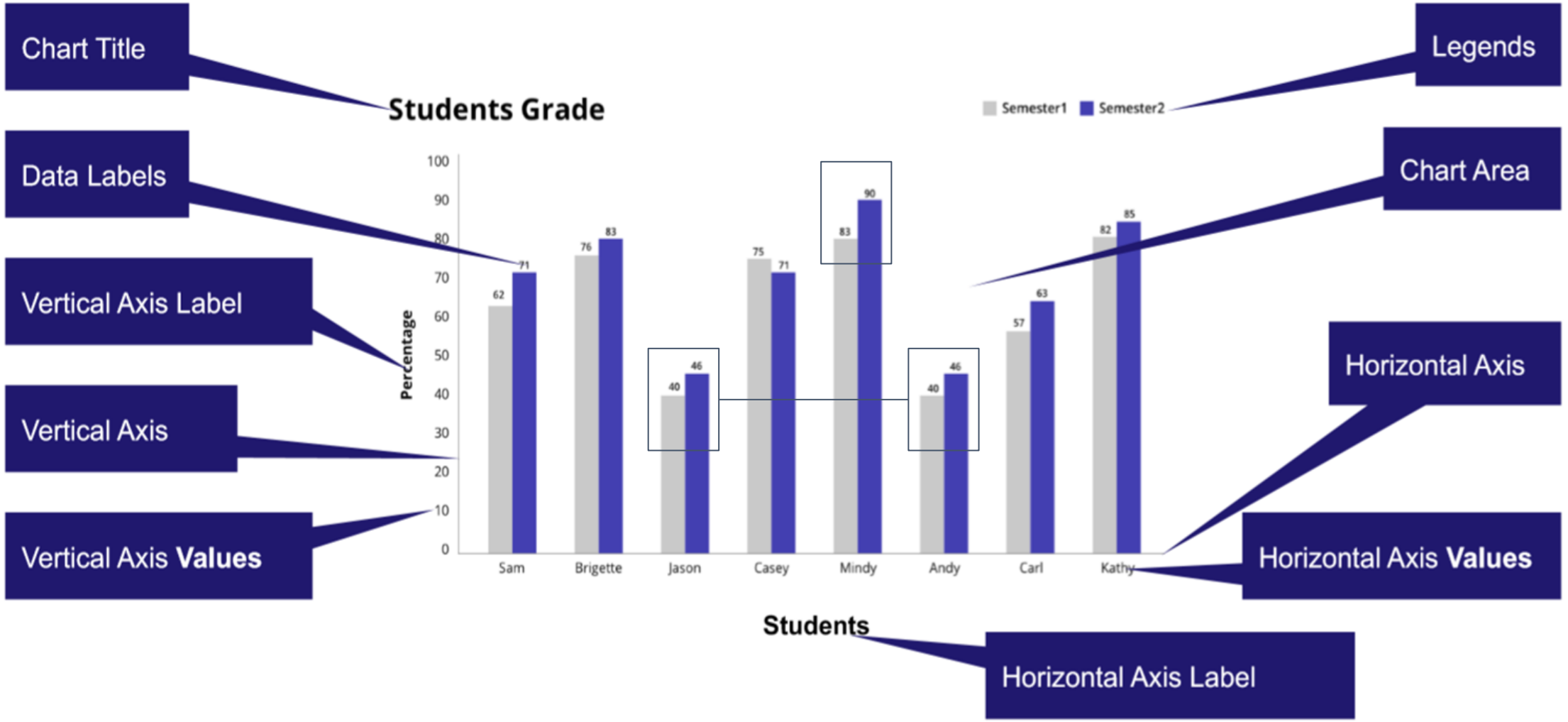
Source data : World Indicator (Power BI)

Purwadhika
Startup and Coding School

How to do Data Visualization ?

- Understand the context of your data
- Making some questions for you data
- Choose appropriate type of visualization and identify the message of each visualization you made
- From technical perspective:
  - Add title to your data visualization to make it more informative
  - Label your axis (x-axis and w-axis)
  - Label your graph (legend)
  - Mark interesting data points
  - Play with color and size
- Get conclusion

Purwadhika
Startup and Coding School



Purwadhika
Startup and Coding School

Some questions that can be answered through data visualization

How is the trend of the sales this past 3 months ?

Which Factor influence customer behavior ?

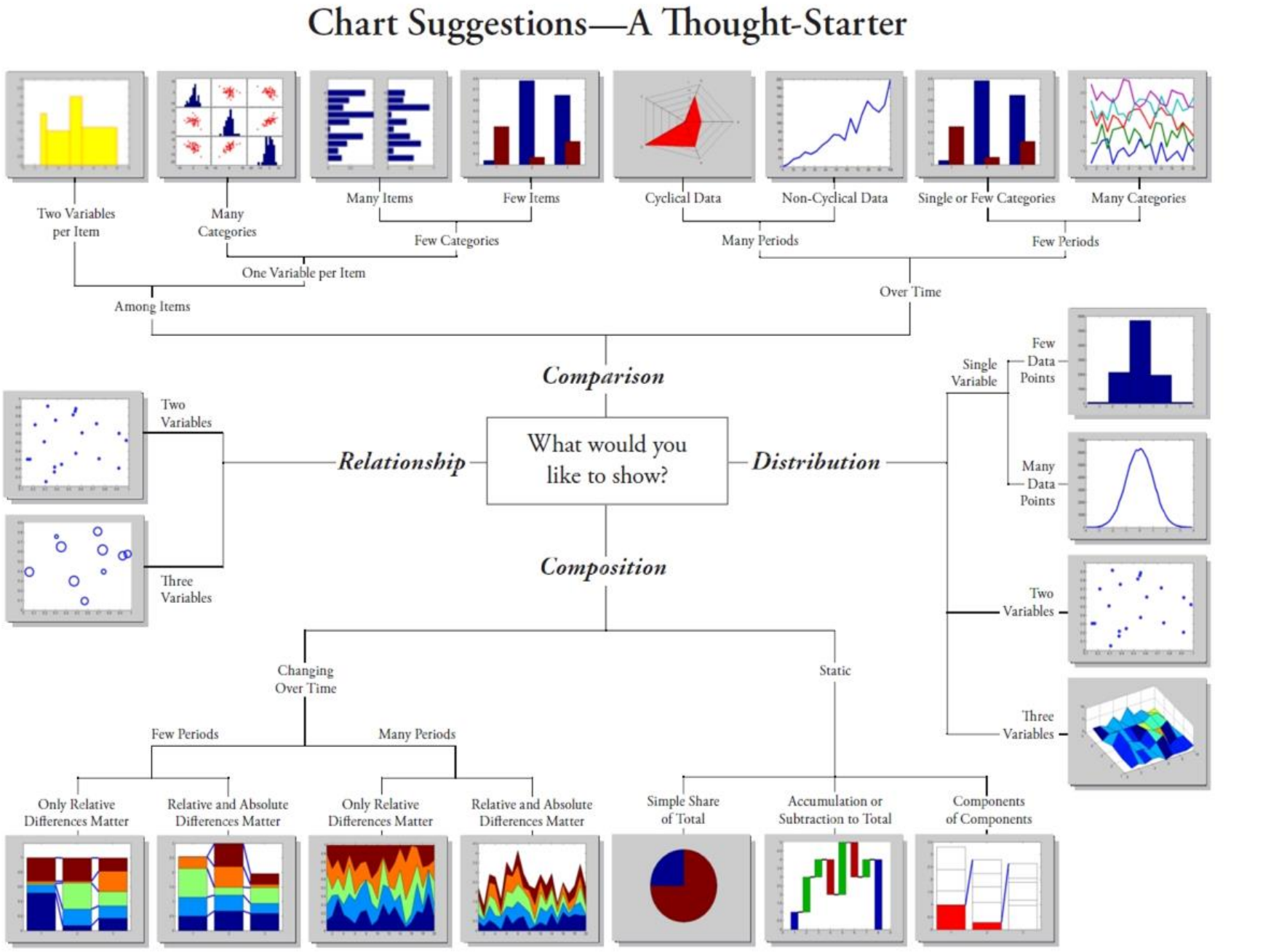
How to place your product to improve sales ?

Which products has the highest sales ?

Purwadhika
Startup and Coding School

Data Visualization Types:

1. Comparison
2. Relationship
3. Distribution
4. Composition



Modified with permision -Doug Hull
blogs.mathworks.com/videos
hull@mathworks.com 2009

www.ExtremePresentation.com

Tool for Visualization

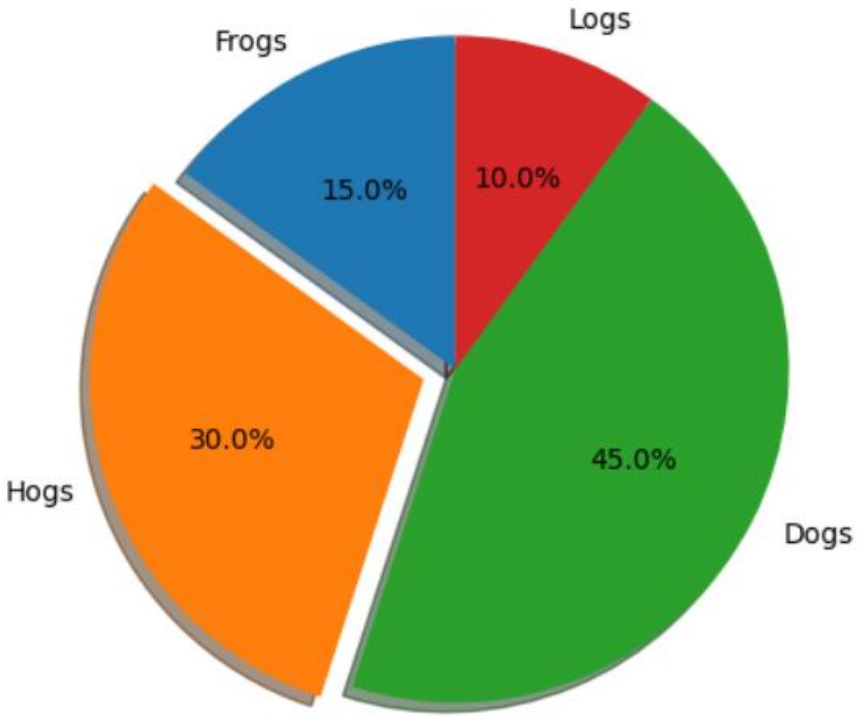
Python:

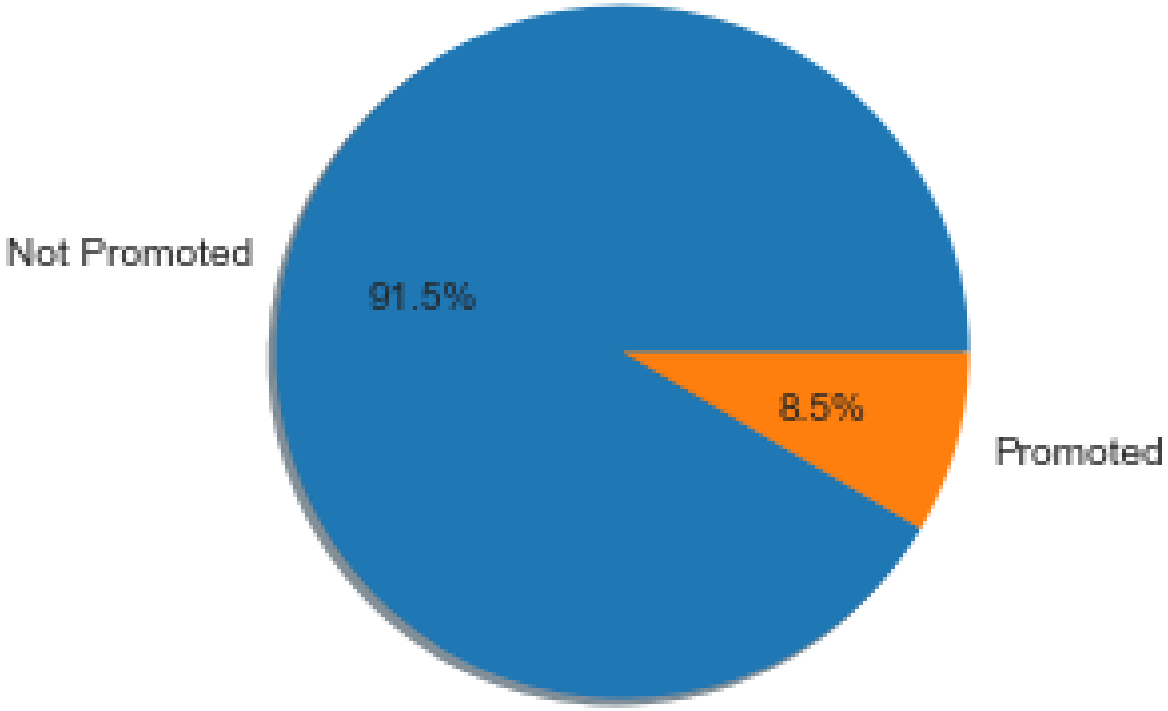
- Pandas
- Matplotlib
- Seaborn
- Plotly
- Bokeh
- Etc

Commonly used:

- *Power BI*
- Tableau
- Microstrategy
- Qlik
- Etc

Purwadhika
Startup and Coding School

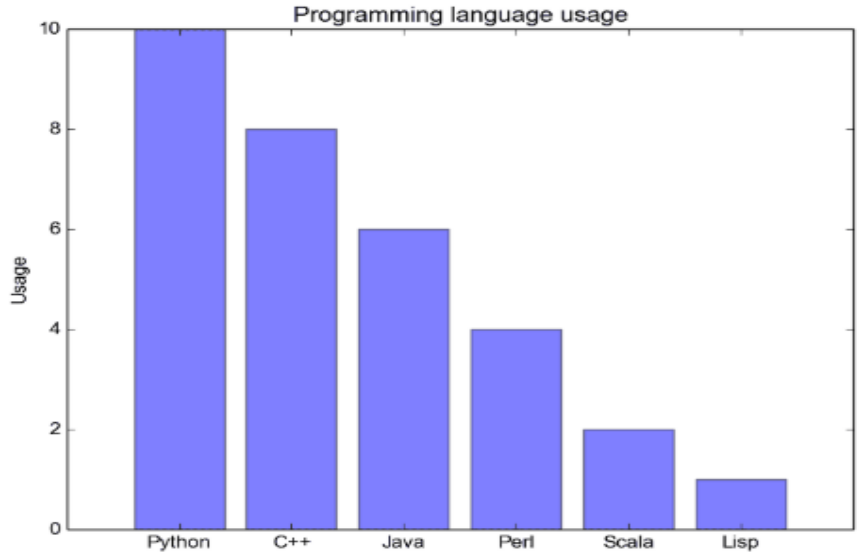




Pie chart

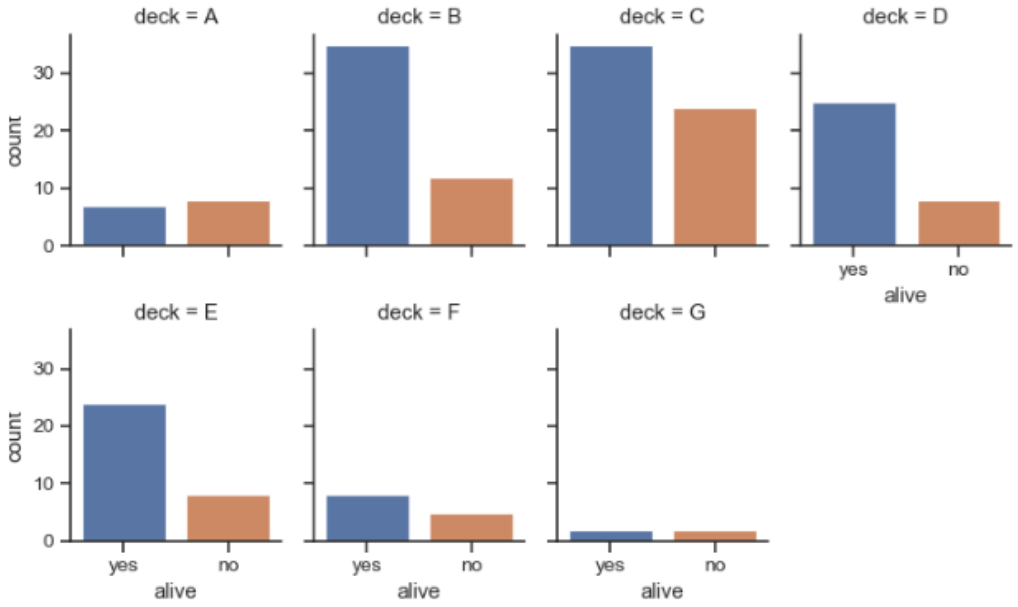
- A circular plot, divided into slices to show numerical proportion of **the categorical data**. They are widely used in the business world.
- Each category are consecutive and non-overlapping
- Main purpose is composition
- Not recommended if there are to many categories

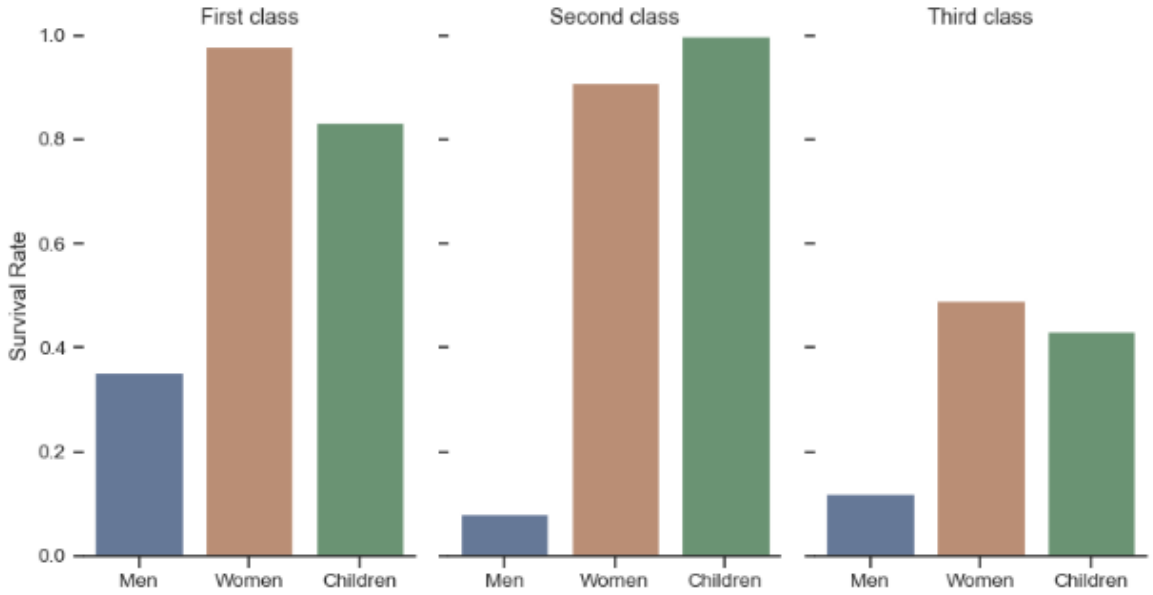
Purwadhika
Startup and Coding School



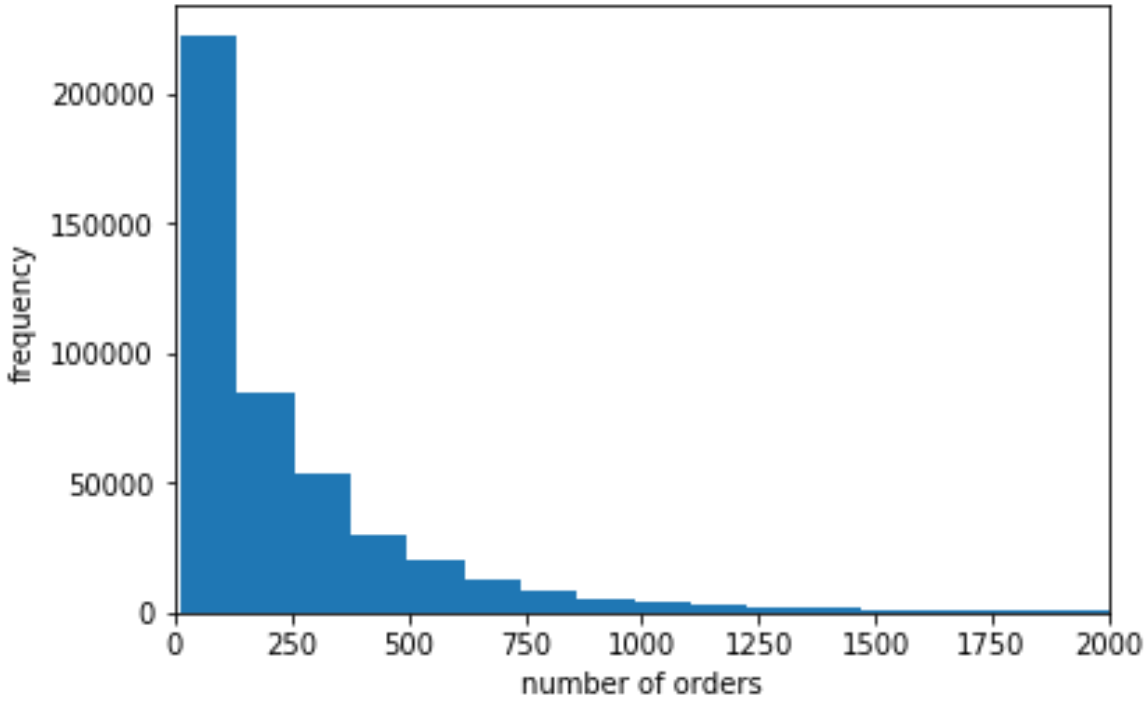
Bar chart

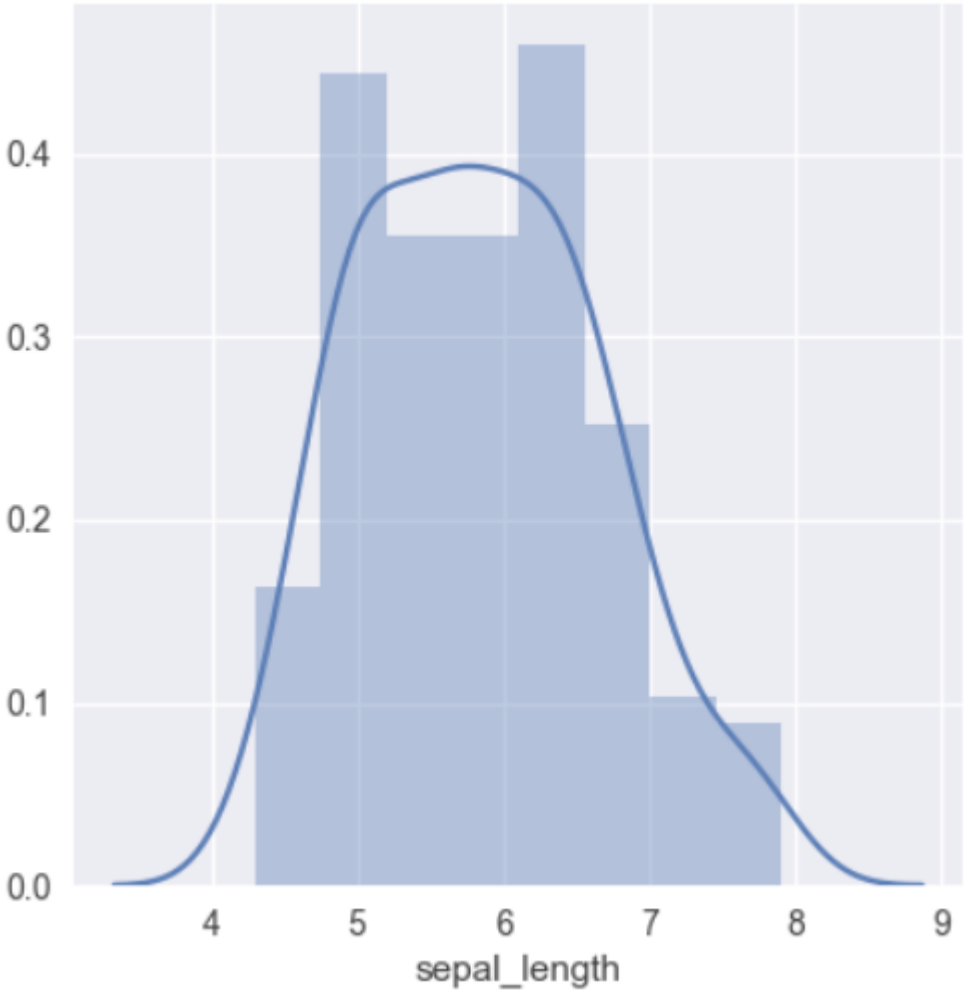
- Represents **categorical data** with rectangular bars. Each bar has a height corresponds to the value it represents. It's useful when we want to **compare** a given numeric value on different **categories**.
- Each category can be consecutive and overlapping
- Can be used to see composition or comparison





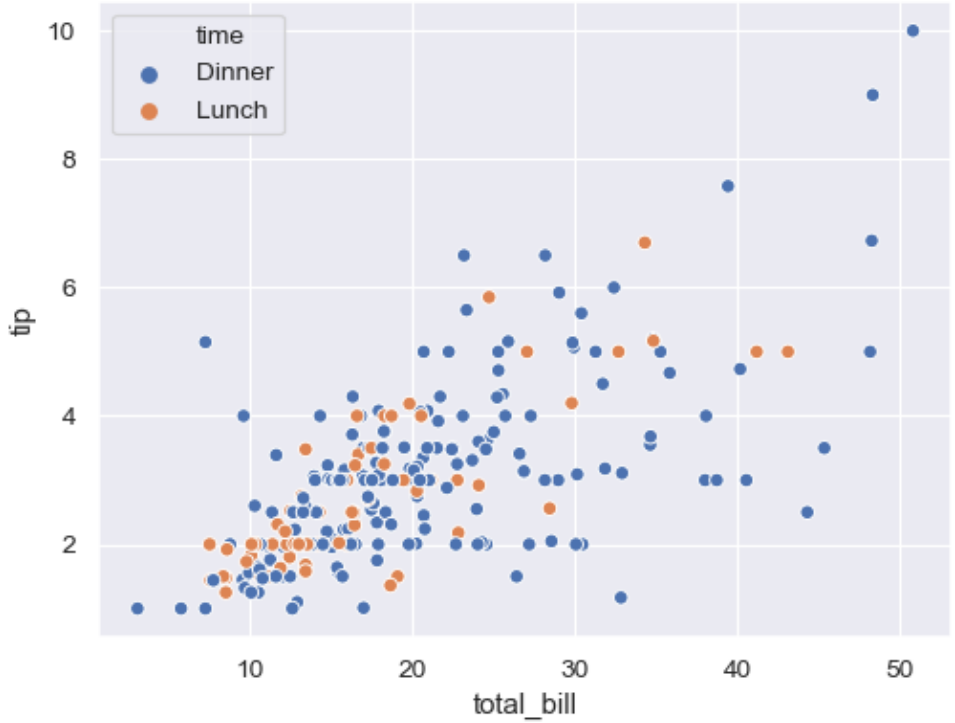
Purwadhika
Startup and Coding School

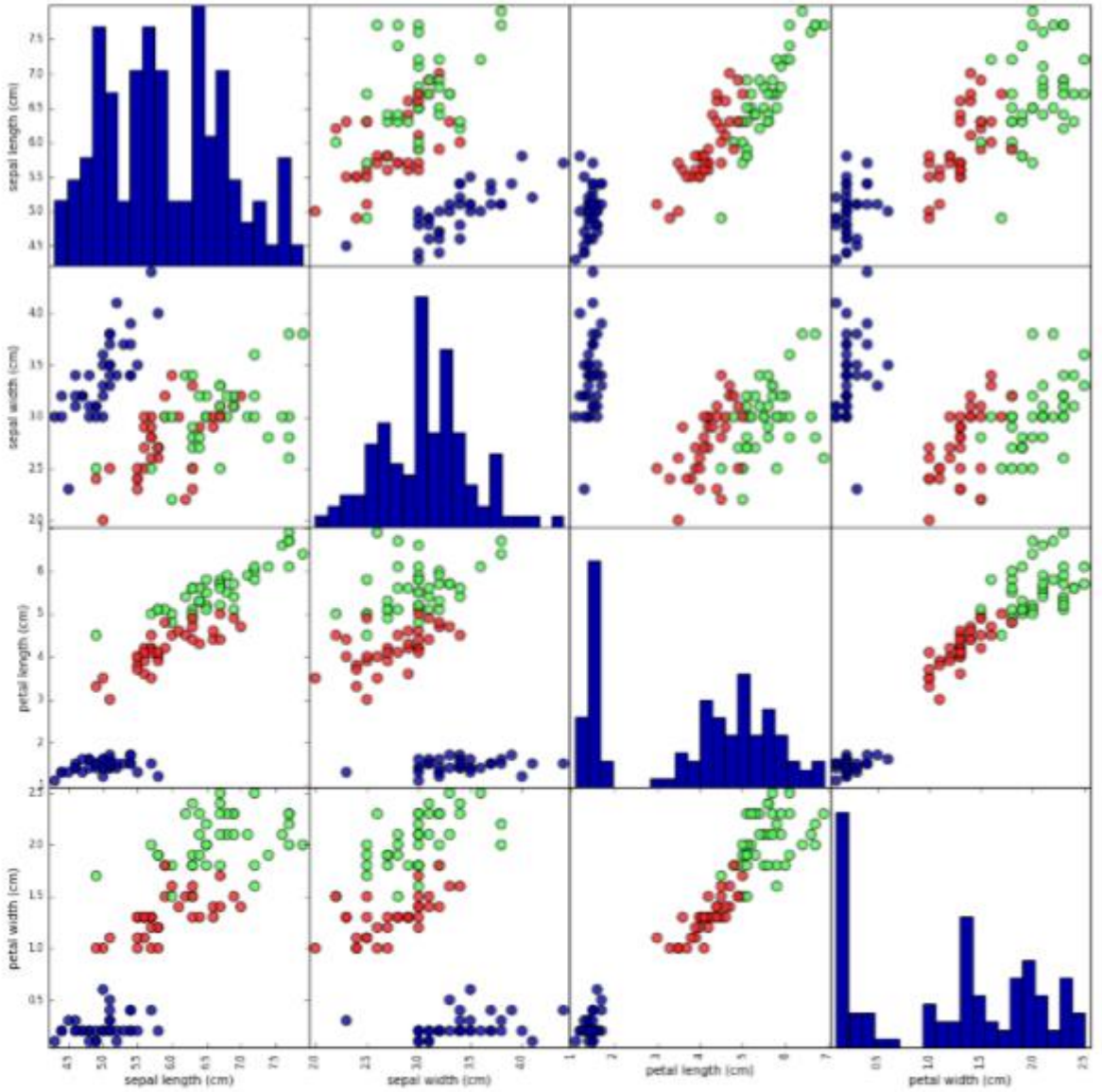




- **Histogram** is an accurate representation of the **distribution of numeric data**.
- A histogram by divide the entire range of values into a series of intervals.
- Then, we count how many values fall into each interval. The intervals are also called **bins**.
- The bins are consecutive and non-overlapping intervals of a variable. They must be adjacent and are often of equal size.

Purwadhika
Startup and Coding School

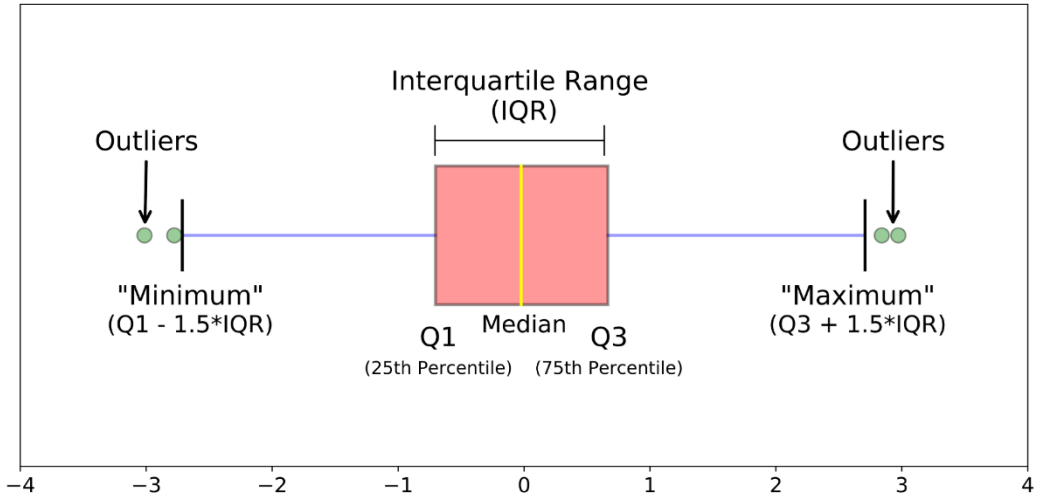


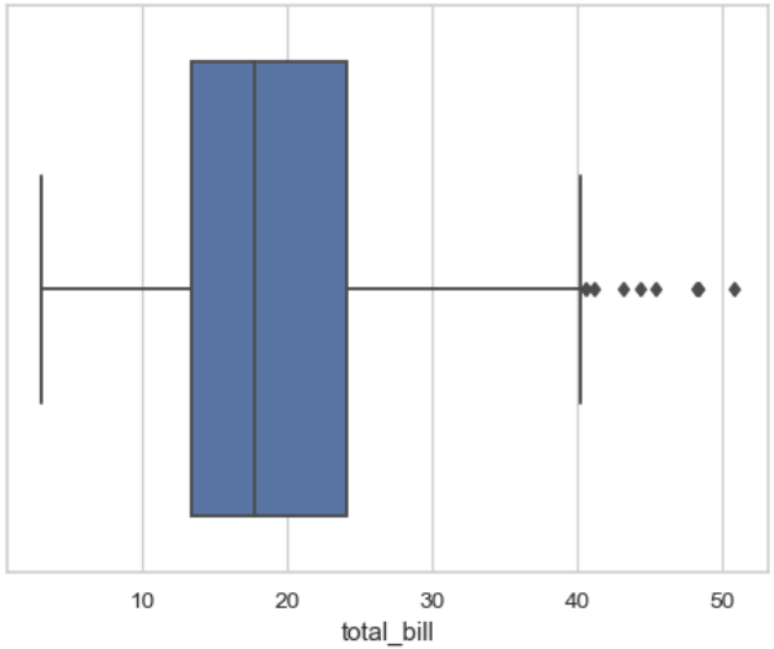


Scatter Plot

- this type of plot shows **all individual data points**. Here, they aren't connected with lines.
- Each data point has the value of the x-axis value and the value from the y-axis values.
- This type of plot can be used to display **trends or correlations**.
- In data science, it shows relationship between two variables.

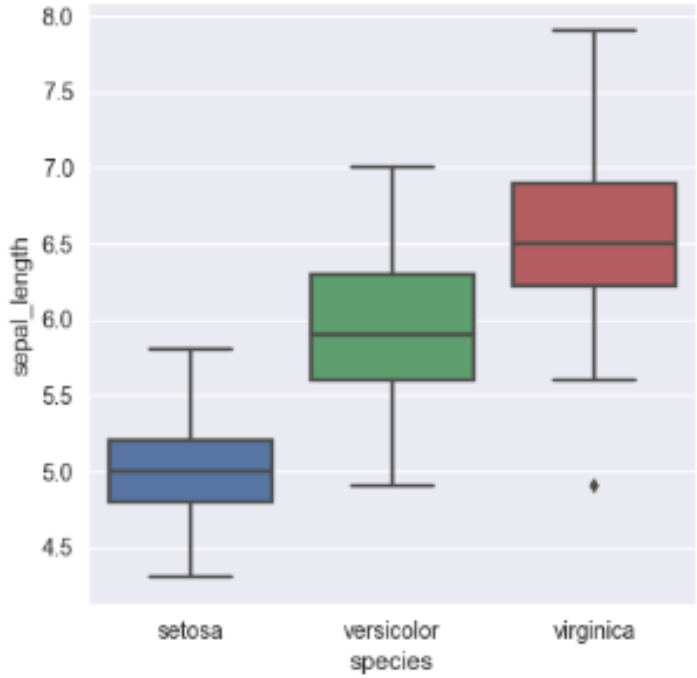
Purwadhika
Startup and Coding School





Box Plot

- Box plot, also called the box-and-whisker plot: a way to show the **distribution of values based on the five-number summary**: minimum, first quartile, median, third quartile, and maximum.
- Can be used for comparison
- Can be used to detect anomaly data/outliers

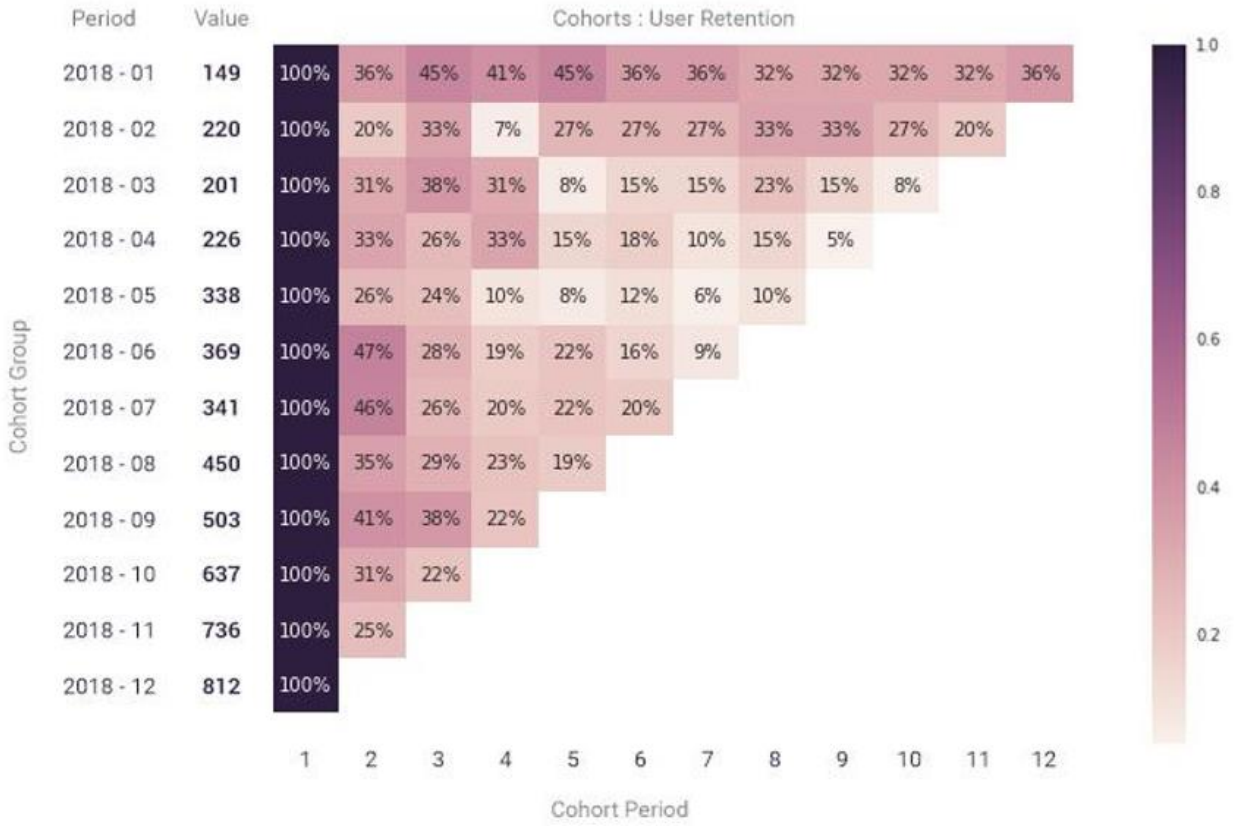


Purwadhika
Startup and Coding School

Line Plot

- A type of plot which displays information as a **series** of **numerical data points** called “markers” connected by **straight lines**.
- In this type of plot, we need the measurement points to be ordered (typically by their x-axis values). This type of plot is often used to visualize a trend in data over intervals of time - a **time series.**
- In this plot we can see trends or any pattern time to time.

Purwadhika
Startup and Coding School



Matrix Plot

Matrix plots allow you to plot data as color-encoded matrices and can also be used to indicate clusters within the data

Purwadhika
Startup and Coding School