DATA ANALYTICS

Unit 1:Data Visualization

Mamatha.H.R
Department of Computer Science and Engineering

DATA ANALYTICS

Unit 1: Data Visualization

Mamatha H R
Department of Computer Science and Engineering

How can we convey data to users effectively?

- Data visualization aims to communicate data clearly and effectively through graphical representation.
- Data visualization has been used extensively in many applications—for example, at work for reporting, managing business operations, and tracking progress of tasks.
- More popularly, we can take advantage of visualization techniques to discover data relationships that are otherwise not easily observable by looking at the raw data.
- Nowadays, people also use data visualization to create fun and interesting graphics.

What is Data Visualization?

- Data visualization is an integral part of descriptive analytics and it assists decision maker with useful insights

- There are many useful charts such as histogram, bar chart, pie-chart, box-plot that would assist data scientist with visualization of the data

Histogram

- Histogram is the visual representation of the data which can be used to assess the probability distribution (frequency distribution) of the data

- Histograms are created for continuous (numerical) data.

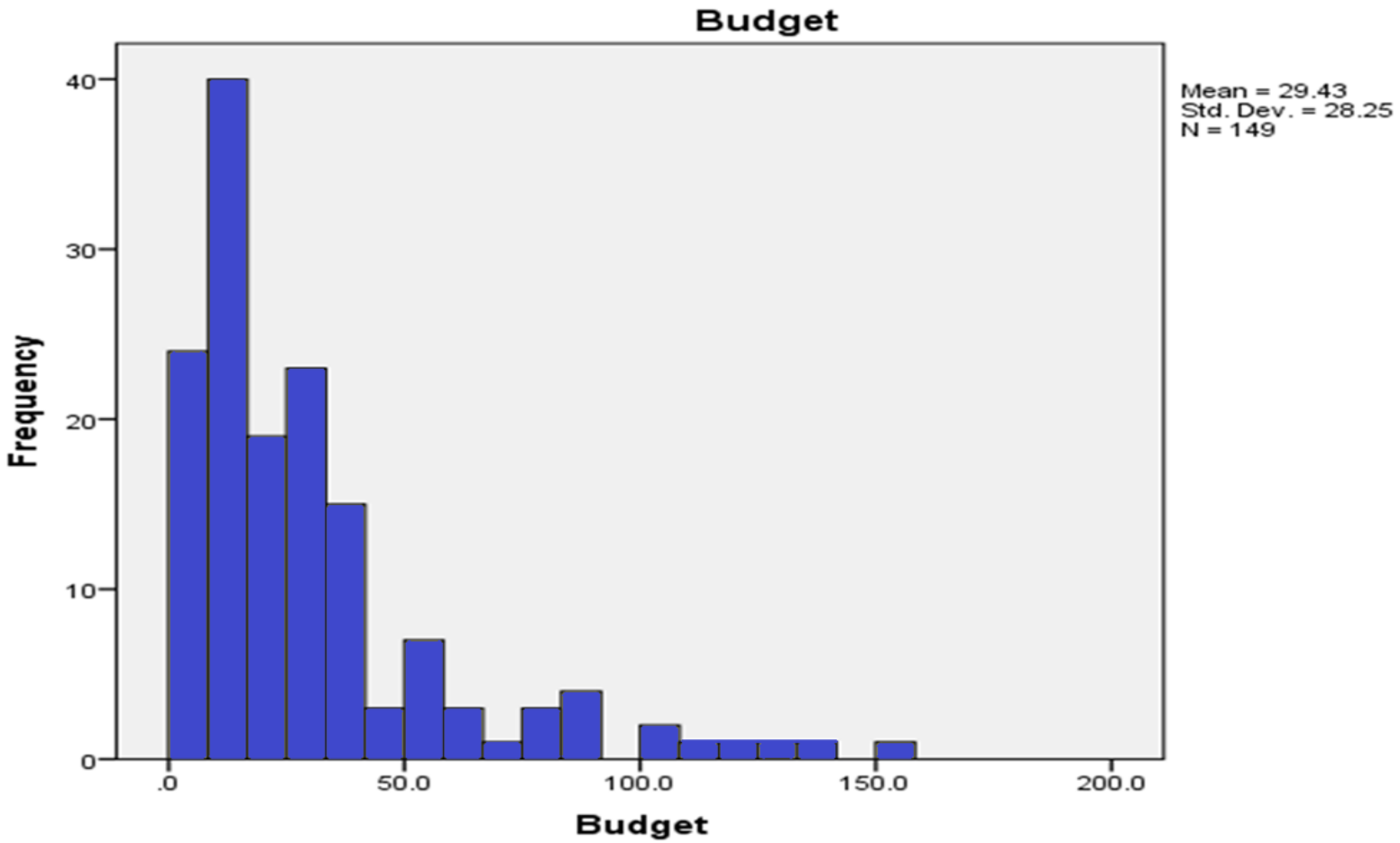
- It is a frequency distribution of data arranged in consecutive and non-overlapping intervals

Use of Histogram

Histogram is very useful since it assists data scientist to identify the following:

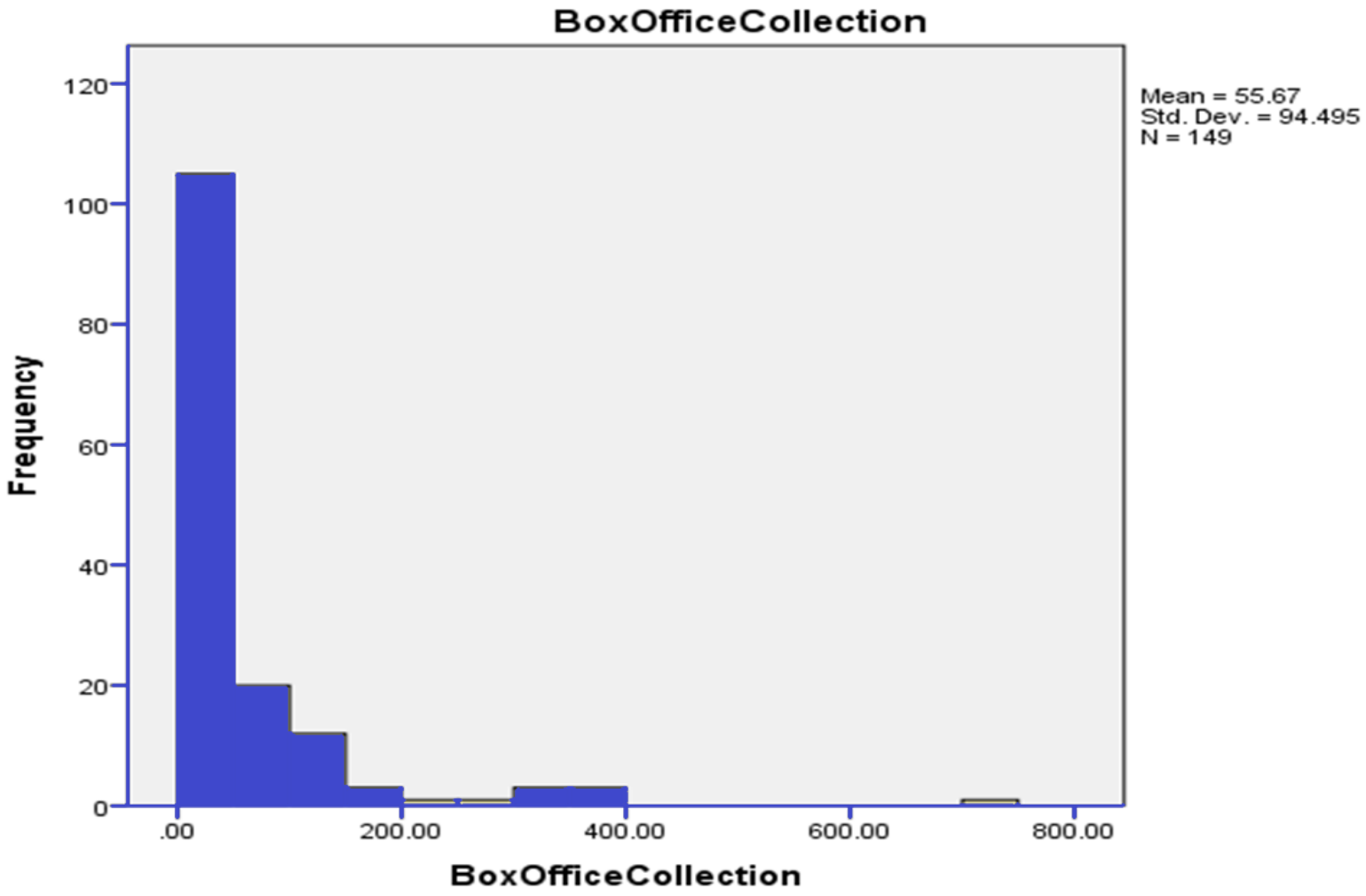
- The shape of the distribution and to assess the probability distribution of the data.
- Measures of central tendency such as median and mode.
- Measures of variability such as spread.
- Measure of shape such as skewness

Histogram of Bollywood movie budget



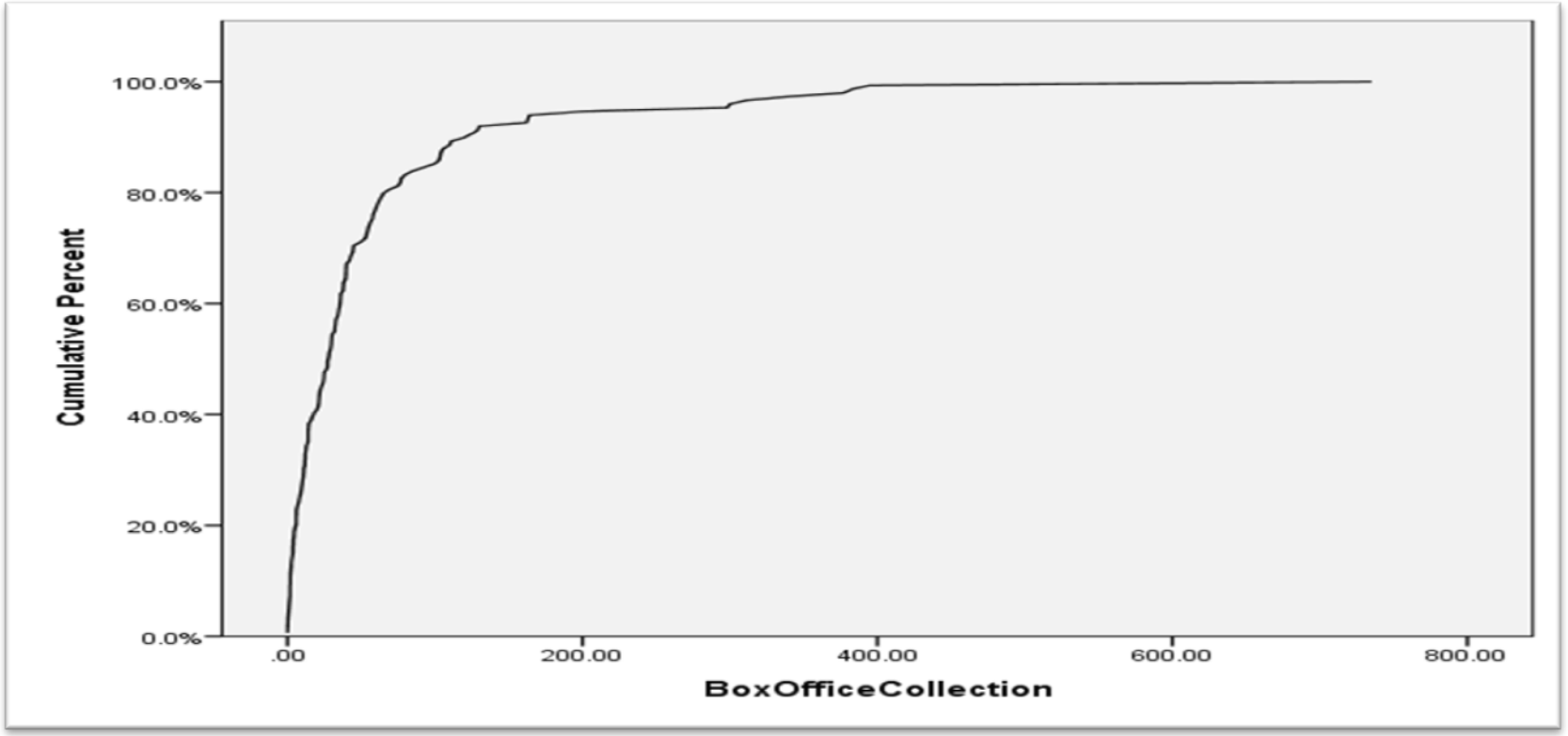
DATA ANALYTICS

Histogram of Bollywood movie box-office collection

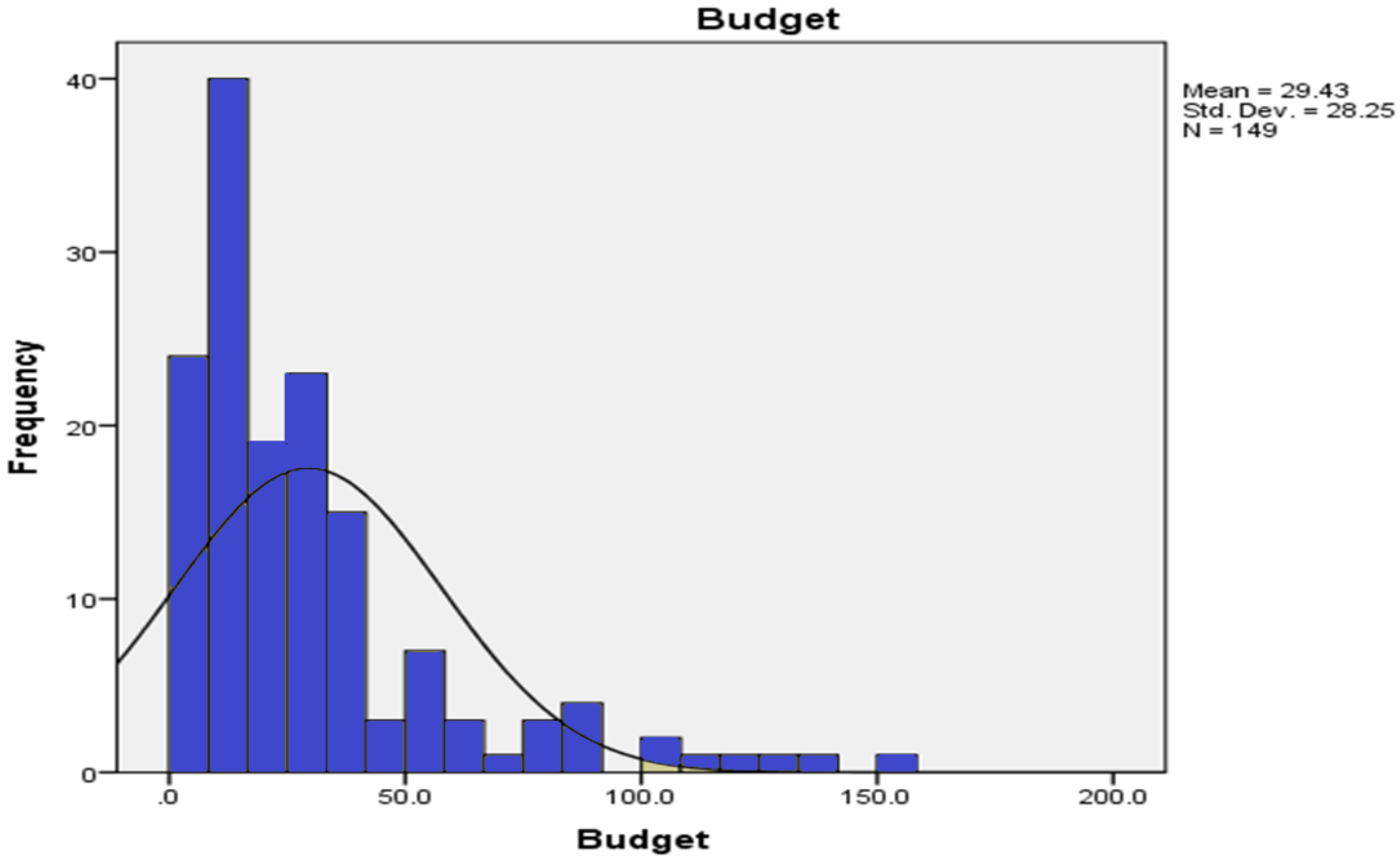


Ogive Curves

- The cumulative histograms are called Ogive curves. The Ogive curve for Bollywood box-office collection is shown below:



Histogram of Bollywood movie budget along with normal distribution frequency



Bar Chart

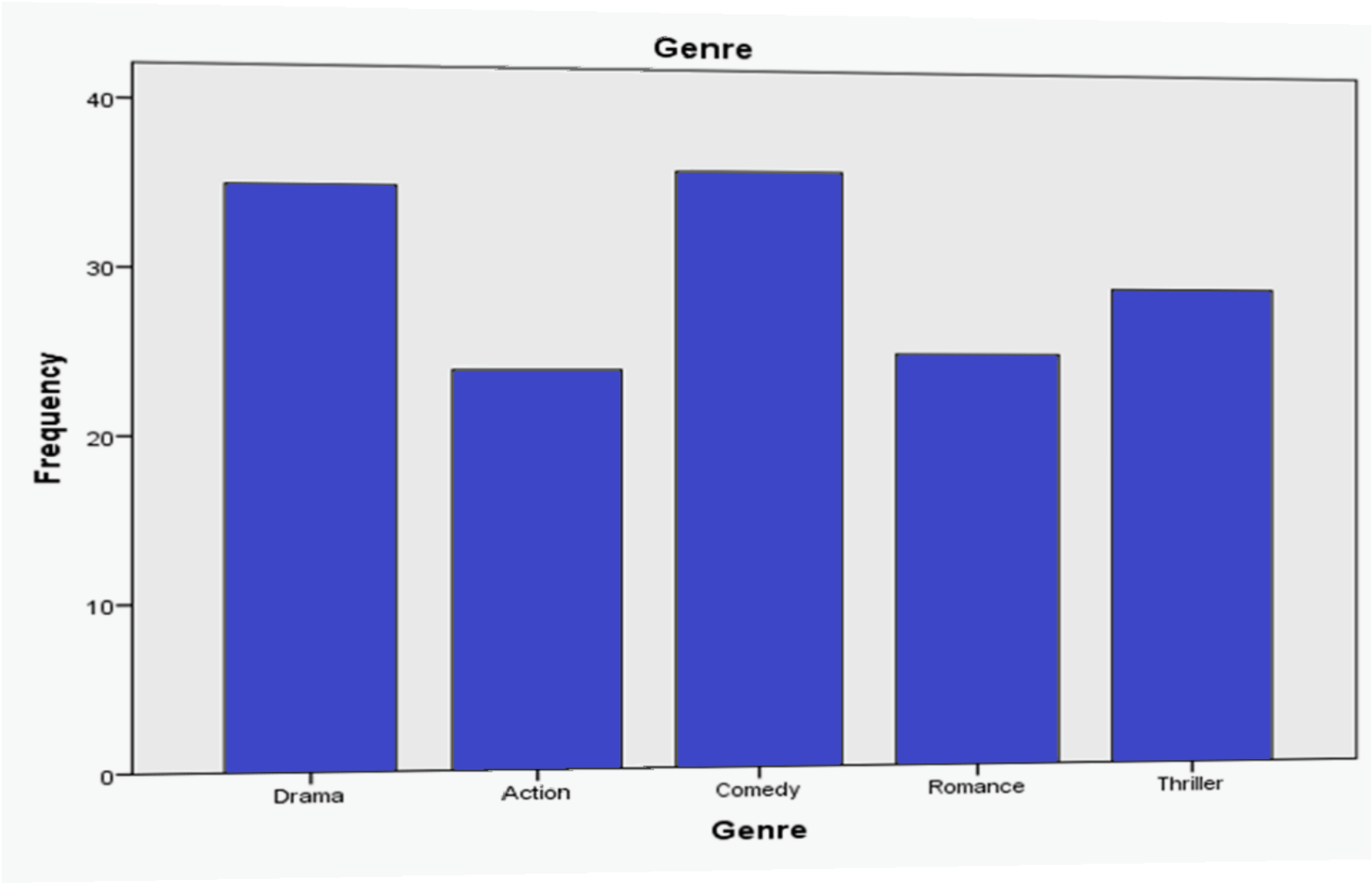
- Bar chart is a frequency chart for qualitative variable (or categorical variable)

- Bar chart can be used to assess the most-occurring and least-occurring categories within a dataset

- Histograms cannot be used when the variable is qualitative

DATA ANALYTICS

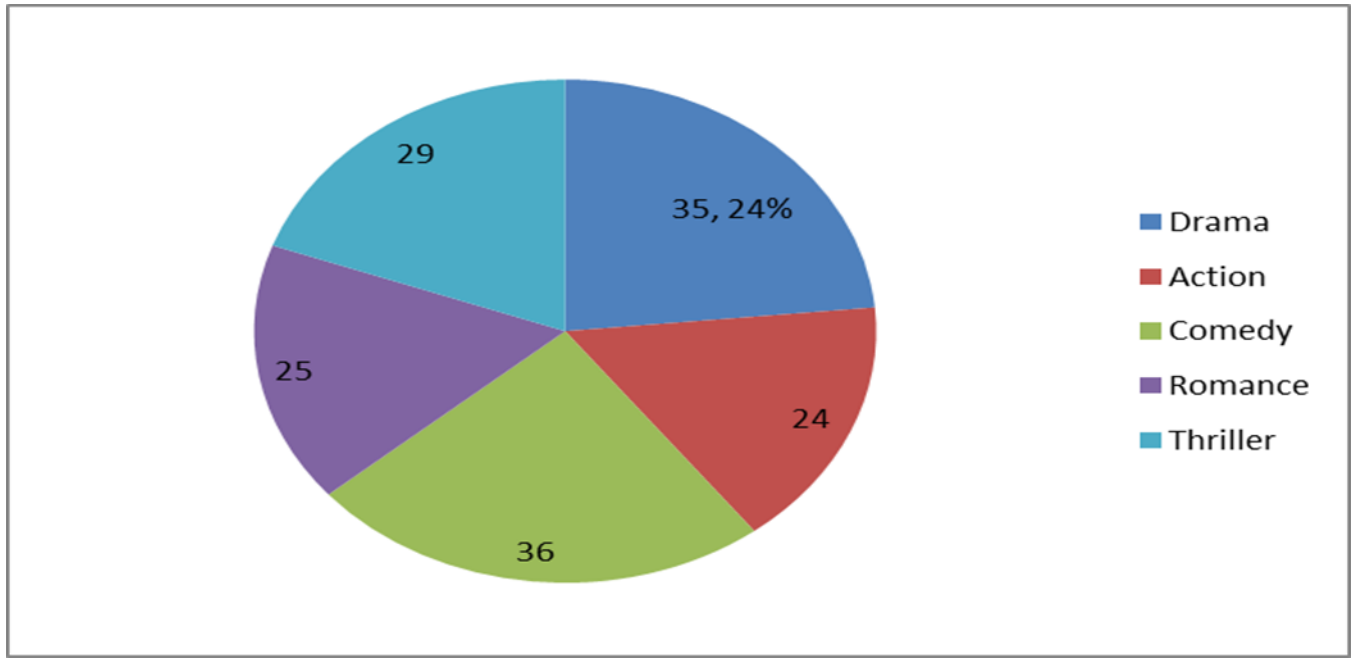
Bar chart for movie genre



Pie Chart

- Pie chart is mainly used for categorical data and is a circular chart that displays the proportion of each category in the dataset

Pie chart for movie genre



Scatter Plot

- Scatter plot is a plot of two variables that will assist data scientists to understand if there is any relationship between two variables

- The relationship could be linear or non-linear

- scatter plot is also useful for assessing the strength of the relationship and to find if there are any outliers in the data

Scatter Plot

- There are many types of coefficients of correlation in scatter points, most popular one is

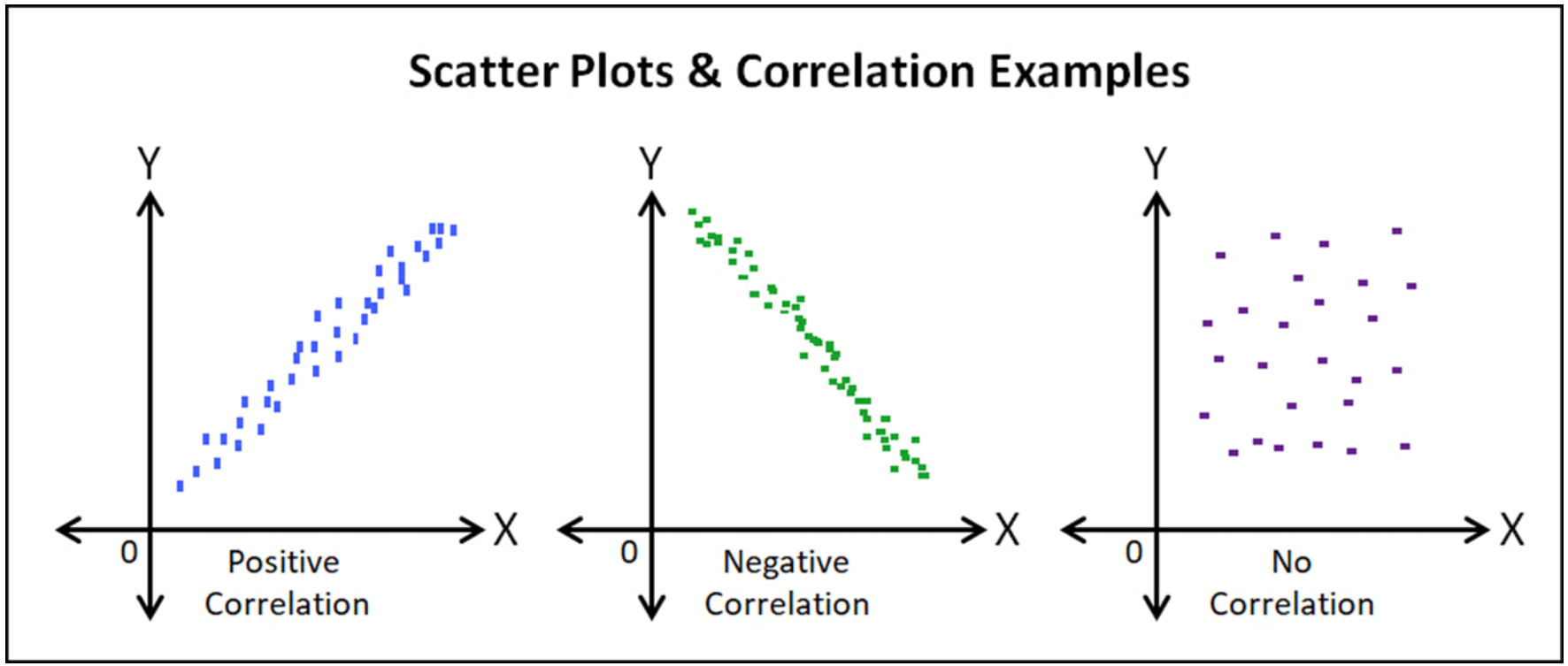
$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

- Pearson's Co-efficient of correlation
- x – value of data point on x-axis
- y – value of data point on y-axis
- n – no of datapoints

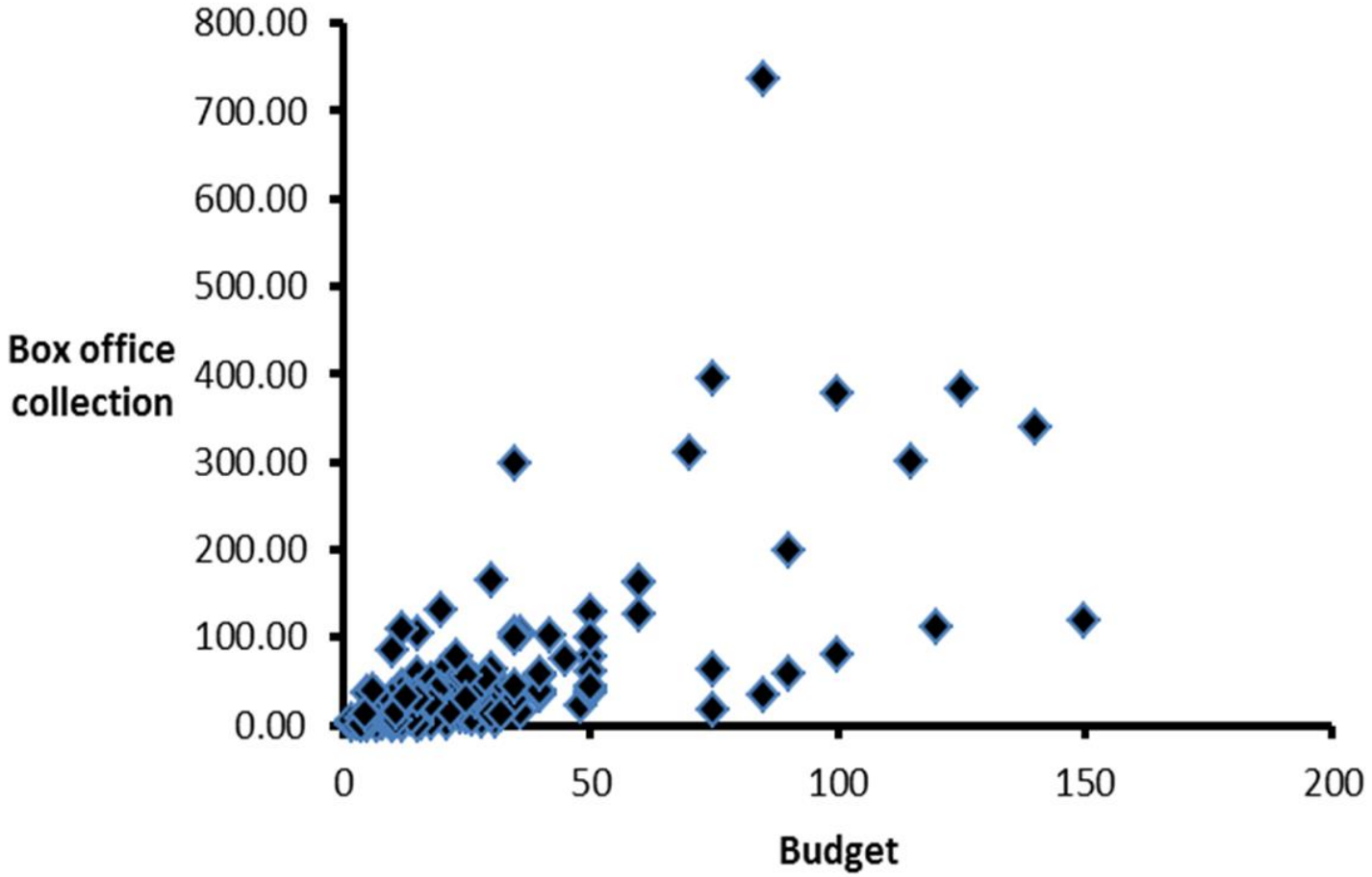
Scatter Plot

- Pearson's co-efficient of correlation:
- co-eff > 0 : positively correlated
- co-eff < 0 : negatively correlated
- co-eff = 0 : no correlation
- +1 or -1, mean perfect correlation between the data points

Scatter Plot



Scatter plot between movie budget and box office collection

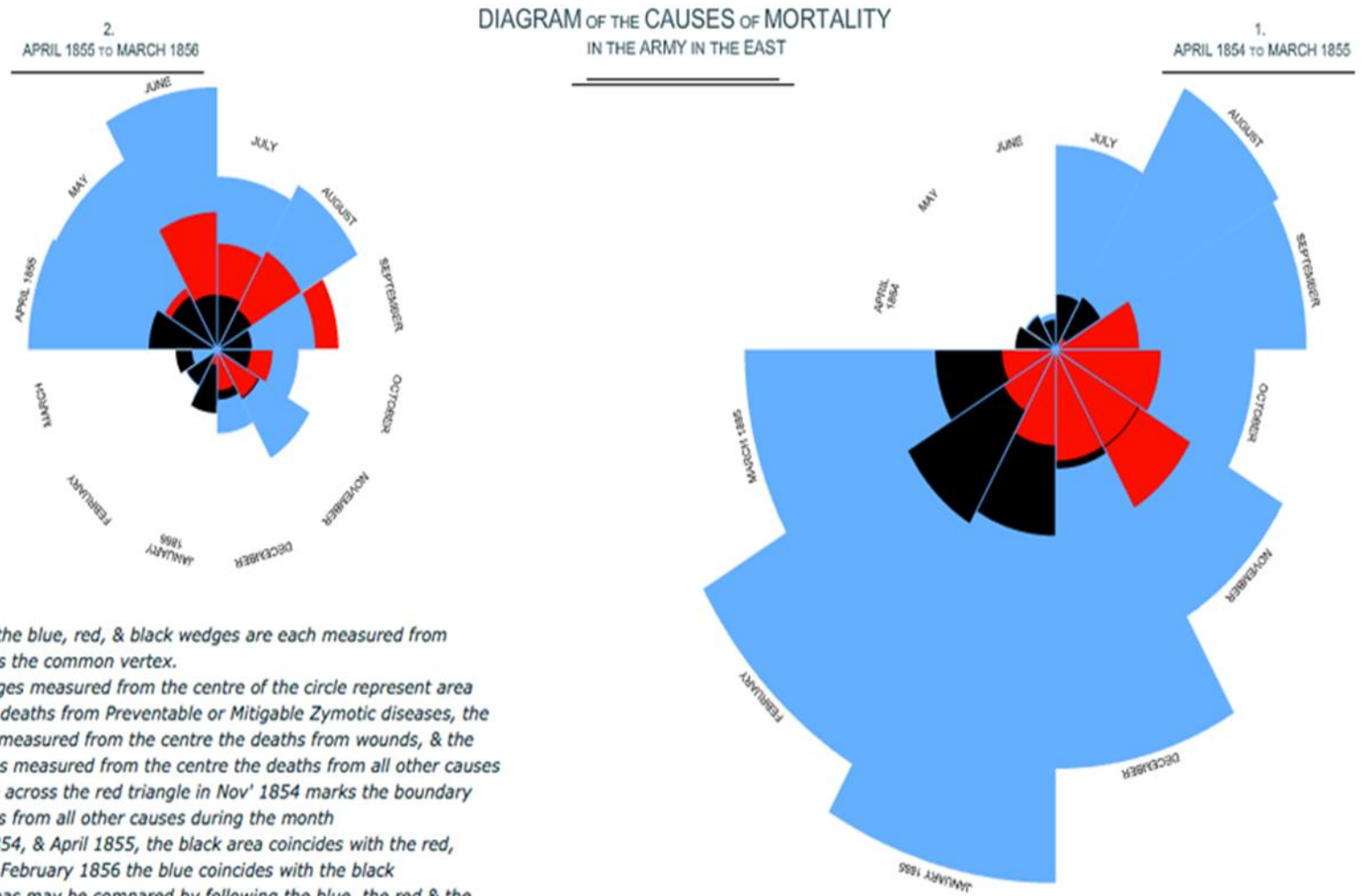


Coxcomb Chart

- Coxcomb chart (also known as polar area chart or roses) is an extension of pie chart made popular by Florence Nightingale (Lewi, 2006)
- In a Coxcomb chart, each area represents a magnitude of the category
- The main difference between the regular pie chart and coxcomb chart is that in the case of pie chart the radius of each sector is same, whereas, in coxcomb chart the radius of the sector is adjusted to create the magnitude of the area

DATA ANALYTICS

Coxcomb chart on causes of mortality in the army prepared by Florence Nightingale



The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.
The blue wedges measured from the centre of the circle represent area for area the deaths from Preventable or Mitigable Zymotic diseases, the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes
The black line across the red triangle in Nov' 1854 marks the boundary of the deaths from all other causes during the month
In October 1854, & April 1855, the black area coincides with the red, in January & February 1856 the blue coincides with the black
The entire areas may be compared by following the blue, the red & the black enclosing lines.

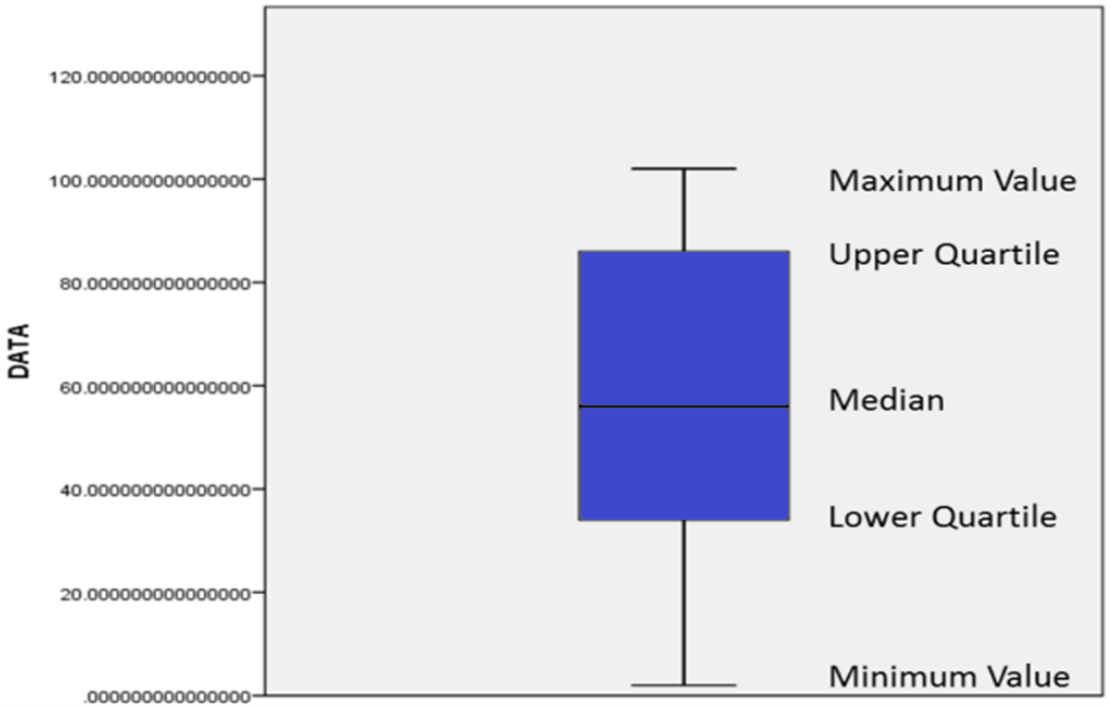
Box Plot (or Box and Whisker Plot)

- Box plot (aka Box and Whisker plot) is a graphical representation of numerical data that can be used to understand the variability of the data and the existence of outliers
- Box plot is designed by identifying the following descriptive statistics:
  - Lower quartile (1st Quartile), median and upper quartile (3rd Quartile).
  - Lowest and highest value
  - Inter-quartile range (IQR).

DATA ANALYTICS

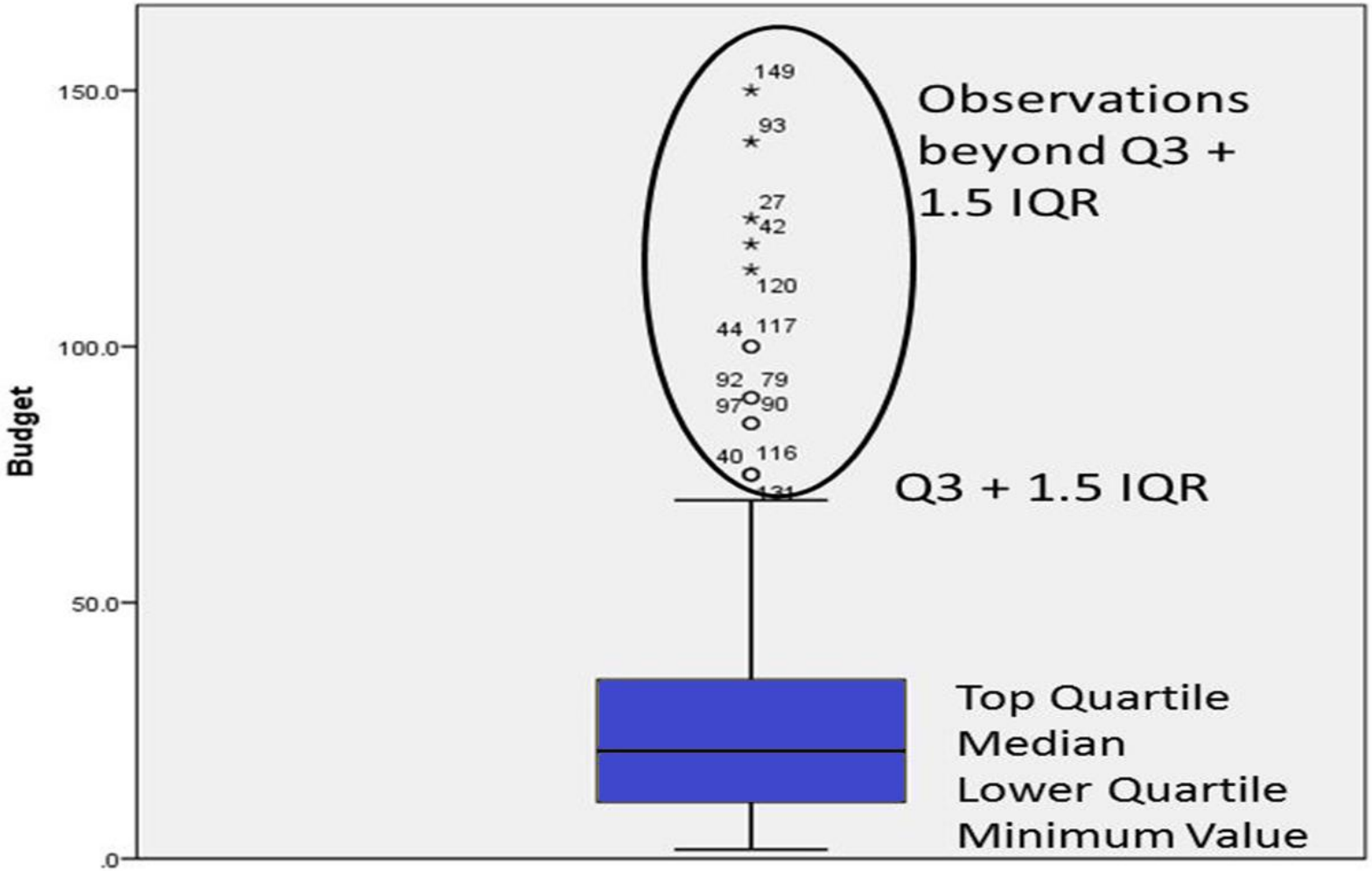
IQR Box Plot

- The box plot is constructed using IQR, minimum and maximum values



Bollywood movie Budget Boxplot

- The box plot for the Bollywood movie budget



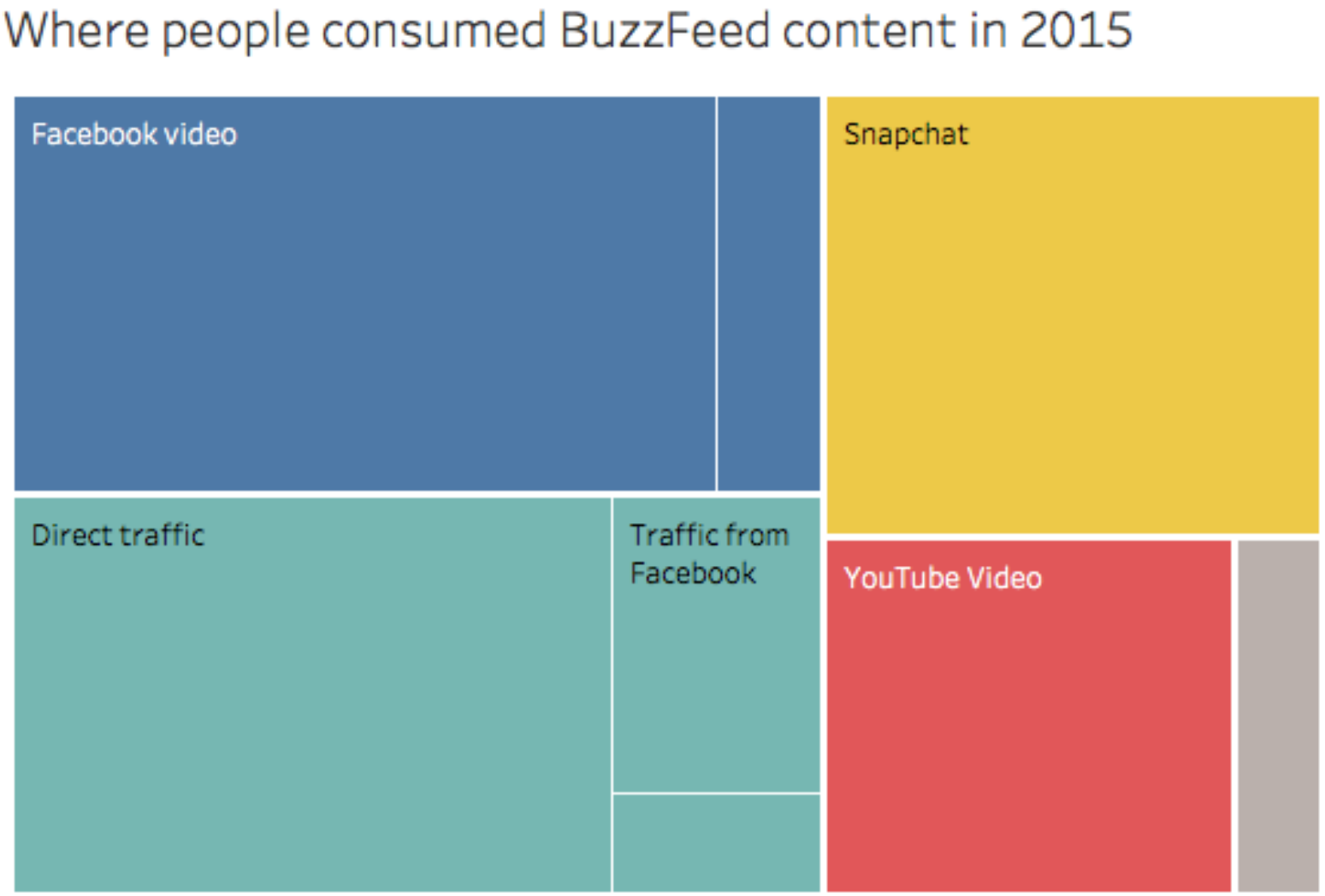
Treemap

- Treemap is a hierarchical map made up of nested rectangles frequently used as part of business intelligence reports which helps organizations to understand the data hierarchically

- The size of rectangle and colour are used for describing/differentiating the characteristics of the data.

DATA ANALYTICS

Treemap



Summary

- Descriptive analytics is beginning of any analytics project that uses data summarization, descriptive statistics, visualization and queries to gain insights about what happened in the past

- Measures of central tendency, measures of variation and measures of shape assist data scientists to understand the data for characteristics such as variability and skewness.

- Descriptive analytics can help data scientists with further analysis of the data by identifying relationships that may exist in the data

Summary

- Data visualization is an integral part of descriptive analytics and plays a major role in business intelligence (BI) by displaying data using innovative graphs and dashboards for easy comprehension of data to top management.

- Descriptive analytics will provide hints for developing predictive analytics models.

Exercises

What are the ideal use cases that warrant the use of a Treemap chart and Coxcomb chart?

References

Text Book:

- "Business Analytics, The Science of Data-Driven Decision Making", U. Dinesh Kumar, Wiley 2017
- Data Mining: Concepts and Techniques by Jiawei Han, Micheline Kamber and Jian Pei, The Morgan Kaufmann Series in Data Management Systems, 3rd Edition.
- Introduction to Data Mining , Tan, Steinbach, Kumar, 2nd Edition

THANK YOU

Dr.Mamatha H R

Professor,Department of Computer Science

mamathahr@pes.edu

+91 80 2672 1983 Extn 834