DATA ANALYTICS

Unit 1:Data Visualization

Mamatha.H.R, Bharathi R
Department of Computer Science and Engineering

DATA ANALYTICS

Unit 1: Data Visualization

Mamatha H R

Department of Computer Science and Engineering

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

Steps to construct histograms

Step 1: Divide the data into finite number of non-overlapping and consecutive bins (interval)

Number of bins, N = $\frac{X_{\max} - X_{\min}}{W}$

Here X_{max} and X_{min} are the maximum and minimum values of the data and W is desired the width of the bin (interval). Intervals in histograms are usually of equal size

Sturges (1926) proposed the following formula for calculating the number of bins

Number of bins, N = $1 + 3.322 \log_{10}(n)$

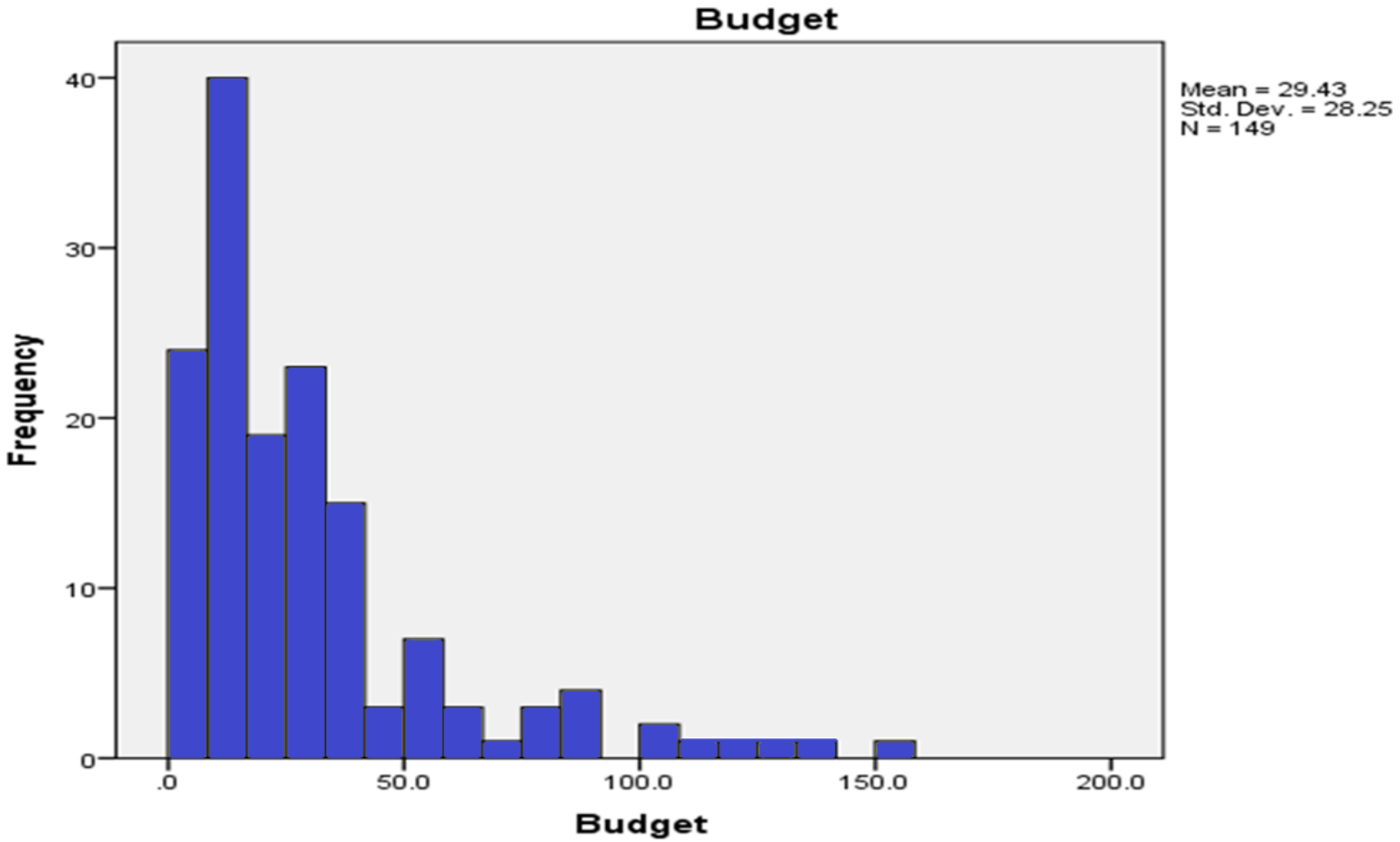
Steps to construct histograms

2) Count the number of observations from the data that fall under each bin (interval).

3) Create a frequency distribution (bin in the horizontal axis and frequency in the vertical axis) using the information obtained in steps 1 and 2

DATA ANALYTICS

Histogram of Bollywood movie budget in crores of rupees



Data file : BollywoodData.xls

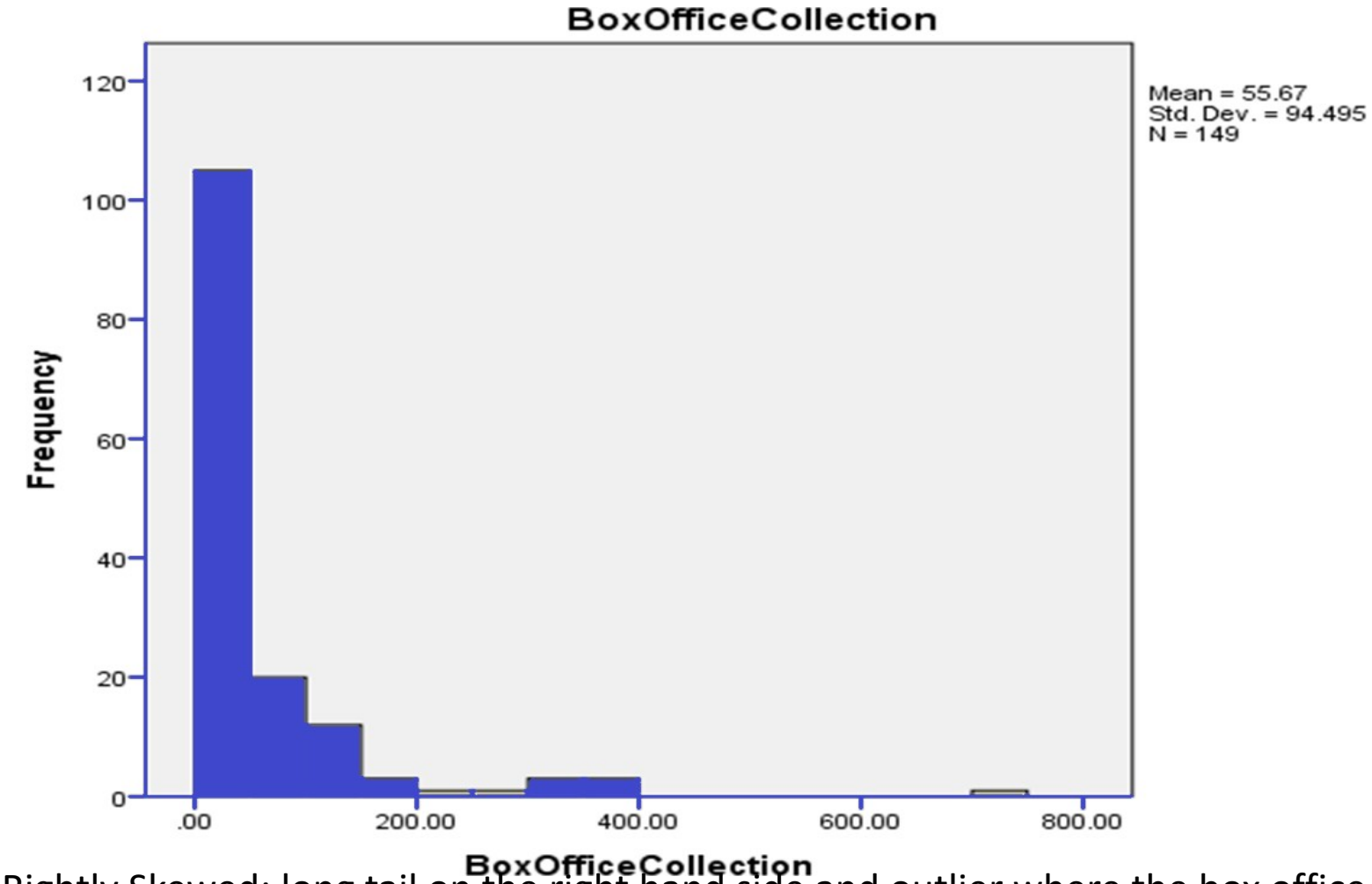
Use of Histogram

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 median and mode.
- Measures of variability such as spread.
- Measure of shape such as skewness

DATA ANALYTICS

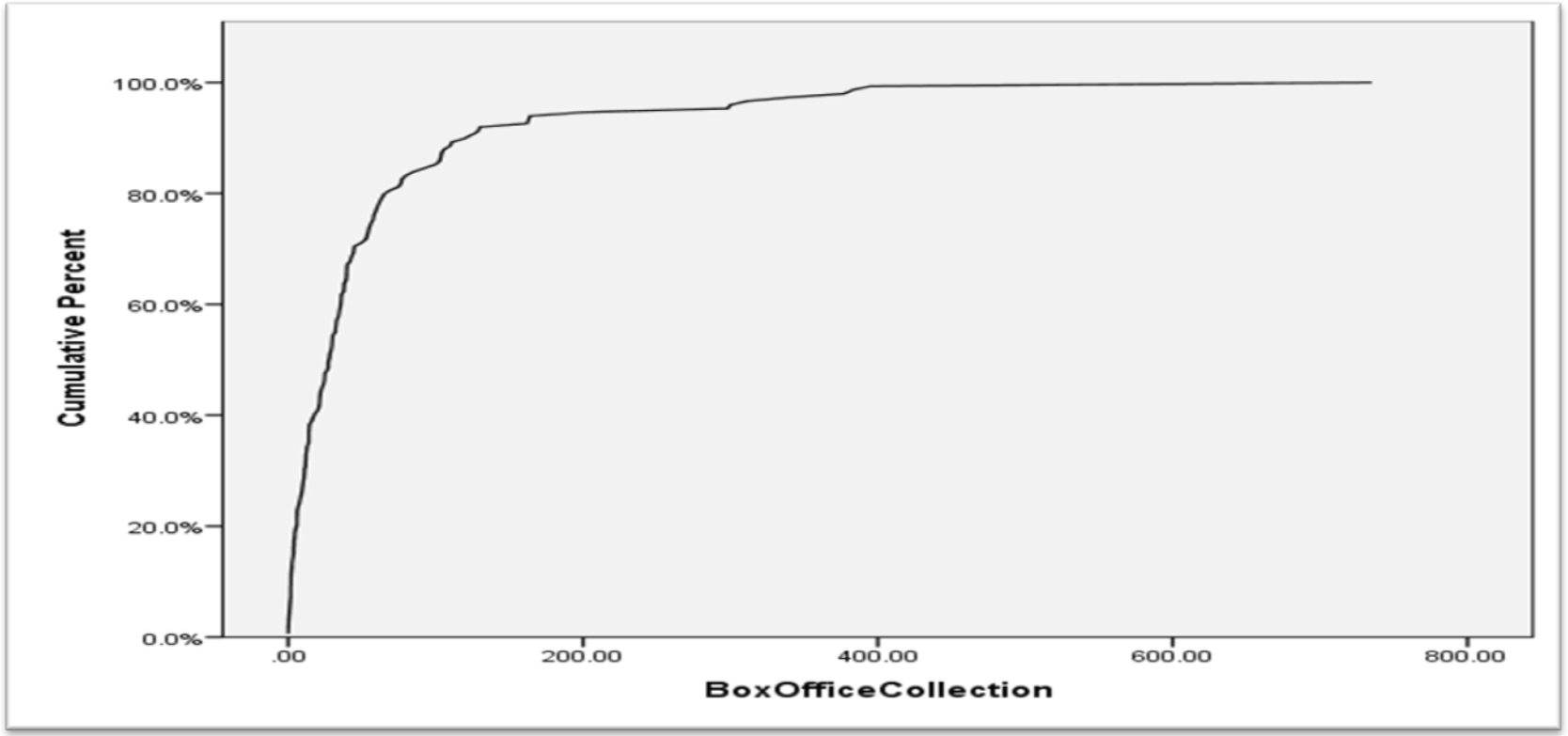
Histogram of Bollywood movie box-office collection



Rightly Skewed; long tail on the right hand side and outlier where the box office collection is more than 700 crores.

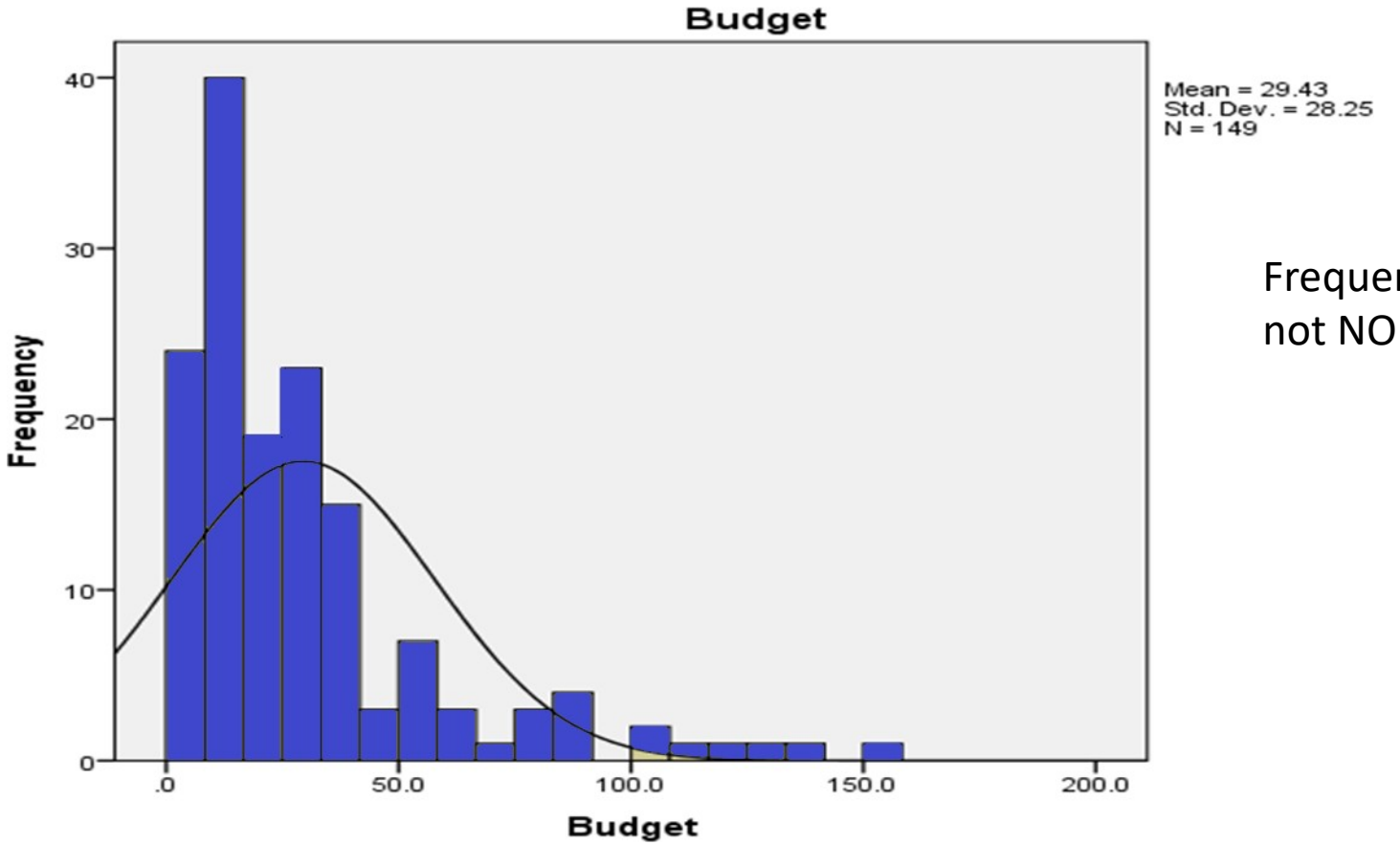
Ogive Curves

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



DATA ANALYTICS

Histogram of Bollywood movie budget along with normal distribution frequency



Frequency distribution of the budget is not NORMAL Distribution

ND superimposed on the histogram

Ogive Curves

- The **cumulative histograms** are called Ogive curves. It can be used to determine how many data values lie above or below a particular value in a data set. The cumulative frequency is calculated from a frequency table

EXAMPLE:

Interval	Frequency	Cumulative frequency
$10 < n \leq 20$	5	
$20 < n \leq 30$	7	
$30 < n \leq 40$	12	
$40 < n \leq 50$	10	
$50 < n \leq 60$	6	

Ogive Curves

- The **cumulative histograms** are called Ogive curves. It can be used to determine how many data values lie above or below a particular value in a data set. The cumulative frequency is calculated from a frequency table

STEP 1 :Compute cumulative frequencies

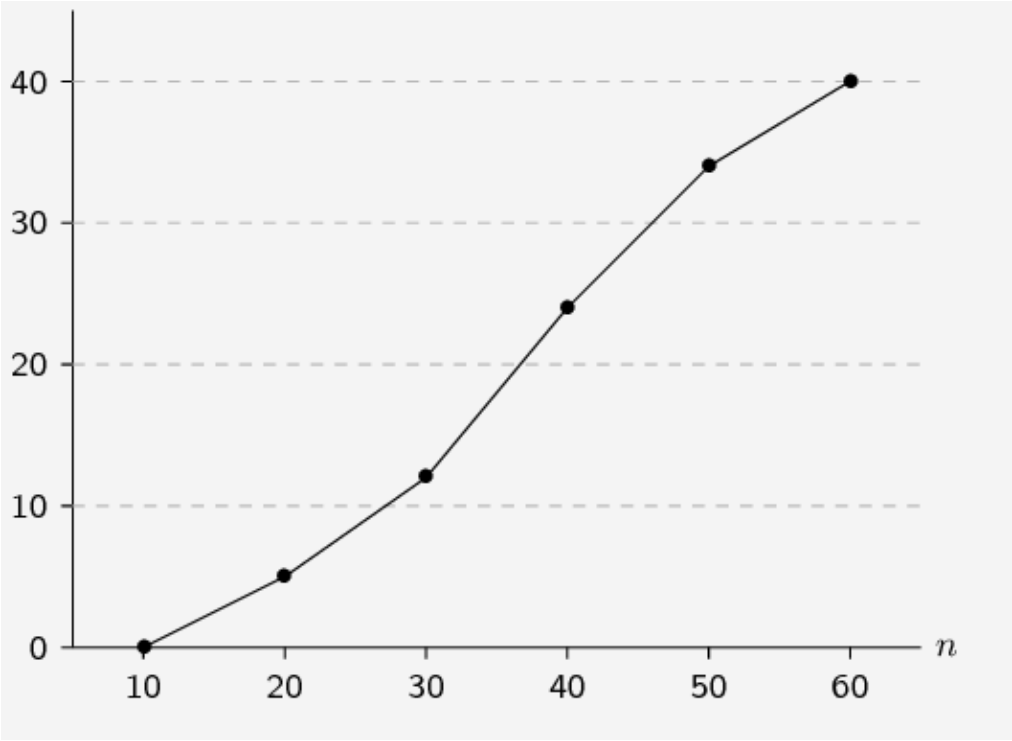
Interval	Frequency	Cumulative frequency
$10 < n \leq 20$	5	5
$20 < n \leq 30$	7	12
$30 < n \leq 40$	12	24
$40 < n \leq 50$	10	34
$50 < n \leq 60$	6	40

Ogive Curves

- The **cumulative histograms** are called Ogive curves. It can be used to determine how many data values lie above or below a particular value in a data set. The cumulative frequency is calculated from a frequency table

STEP 2 :Plot the ogive

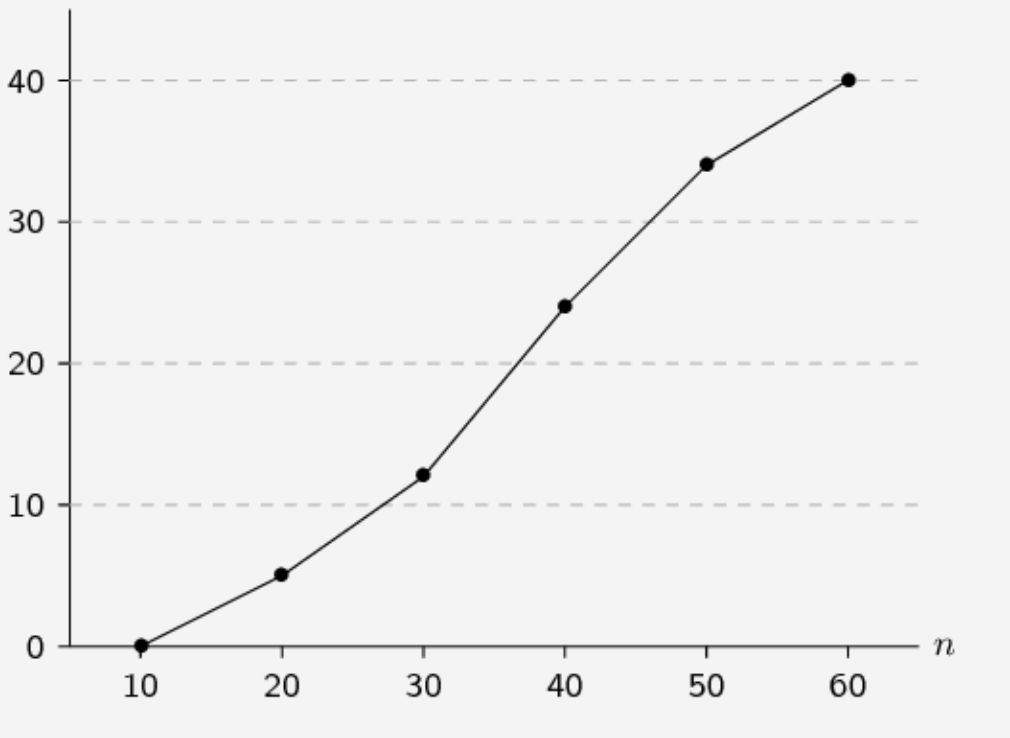
Interval	Frequency	Cumulative frequency
$10 < n \leq 20$	5	5
$20 < n \leq 30$	7	12
$30 < n \leq 40$	12	24
$40 < n \leq 50$	10	34
$50 < n \leq 60$	6	40



DATA ANALYTICS

Ogive CURVES

- Ogives are useful for determining the median, percentiles and five number summary of data.
- Remember that the median is simply the value in the middle when we order the data.
- A quartile is simply a quarter of the way from the beginning or the end of an ordered data set.
- With an ogive we already know how many data values are above or below a certain point, so it is easy to find the middle or a quarter of the data set.



DATA ANALYTICS

OGIVES AND THE FIVE NUMBER SUMMARY

Find the minimum and maximum

The minimum value in the data set is 1, since this is where the ogive starts on the horizontal axis.

The maximum value in the data set is 10 since this is where the ogive stops on the horizontal axis.

Find the quartiles

The quartiles are the values that are 1/4, 1/2 and 3/4 of the way into the ordered data set.

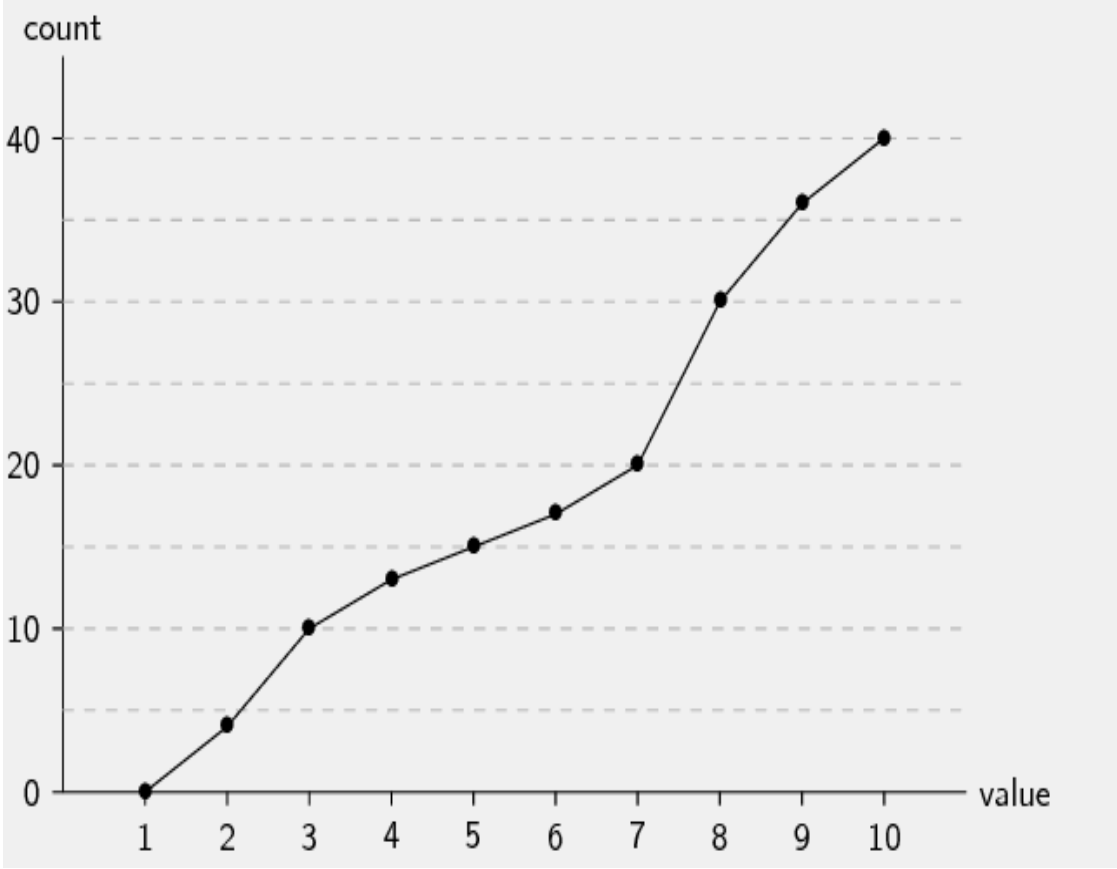
Here the counts go up to 40, so we can find the quartiles by looking at the values corresponding to counts of 10, 20 and 30.

On the ogive a count of

10 corresponds to a value of 3 (first quartile);

20 corresponds to a value of 7 (second quartile); and

30 corresponds to a value of 8 (third quartile).

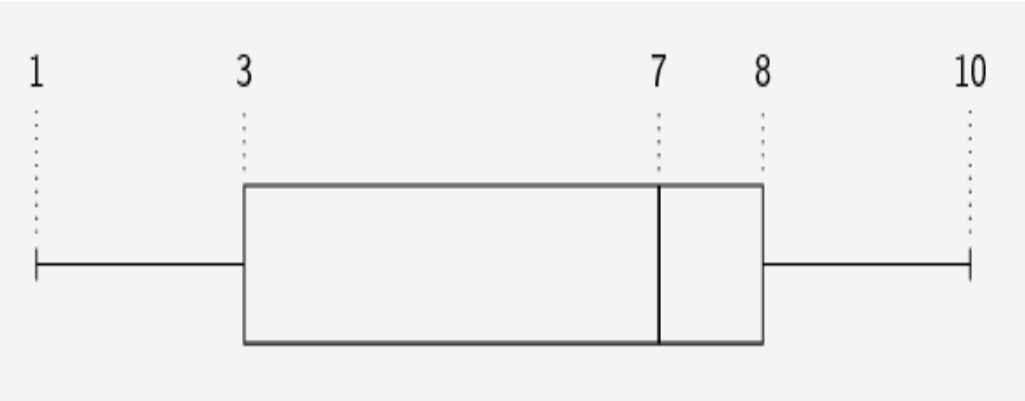


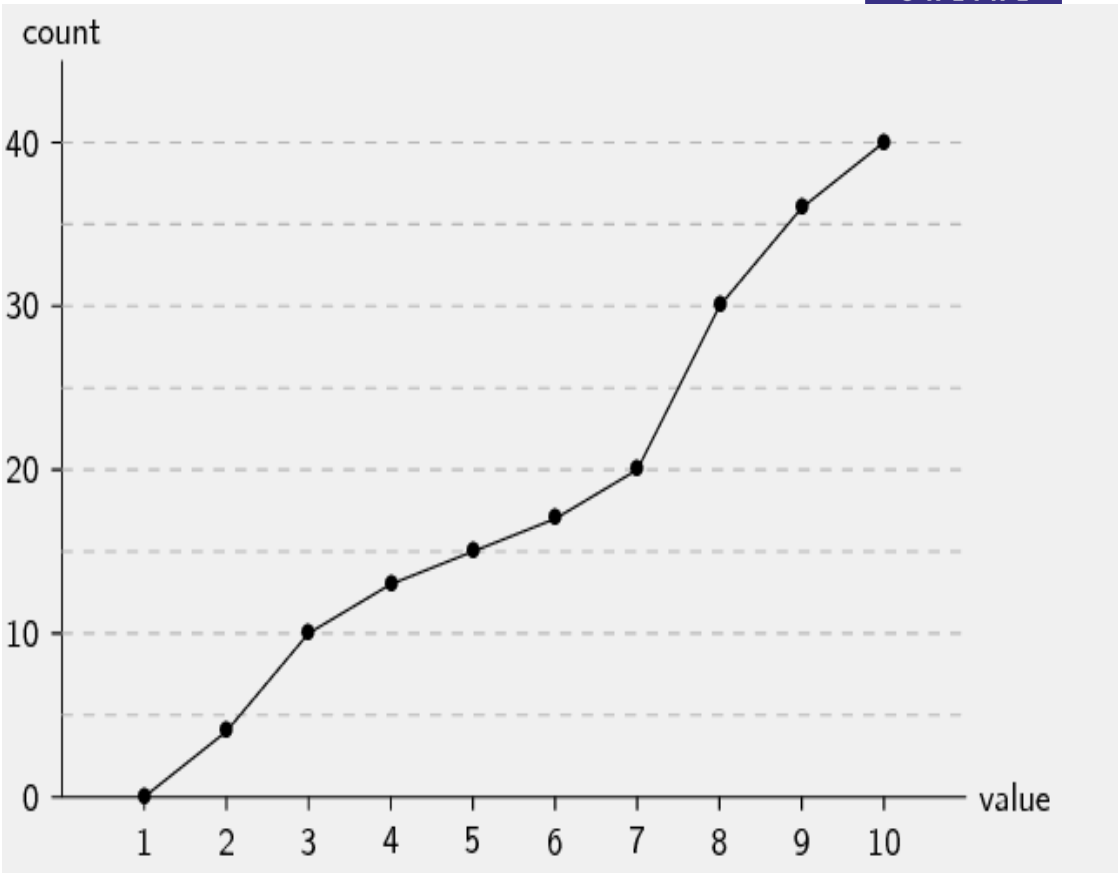
DATA ANALYTICS

OGIVES AND THE FIVE NUMBER SUMMARY

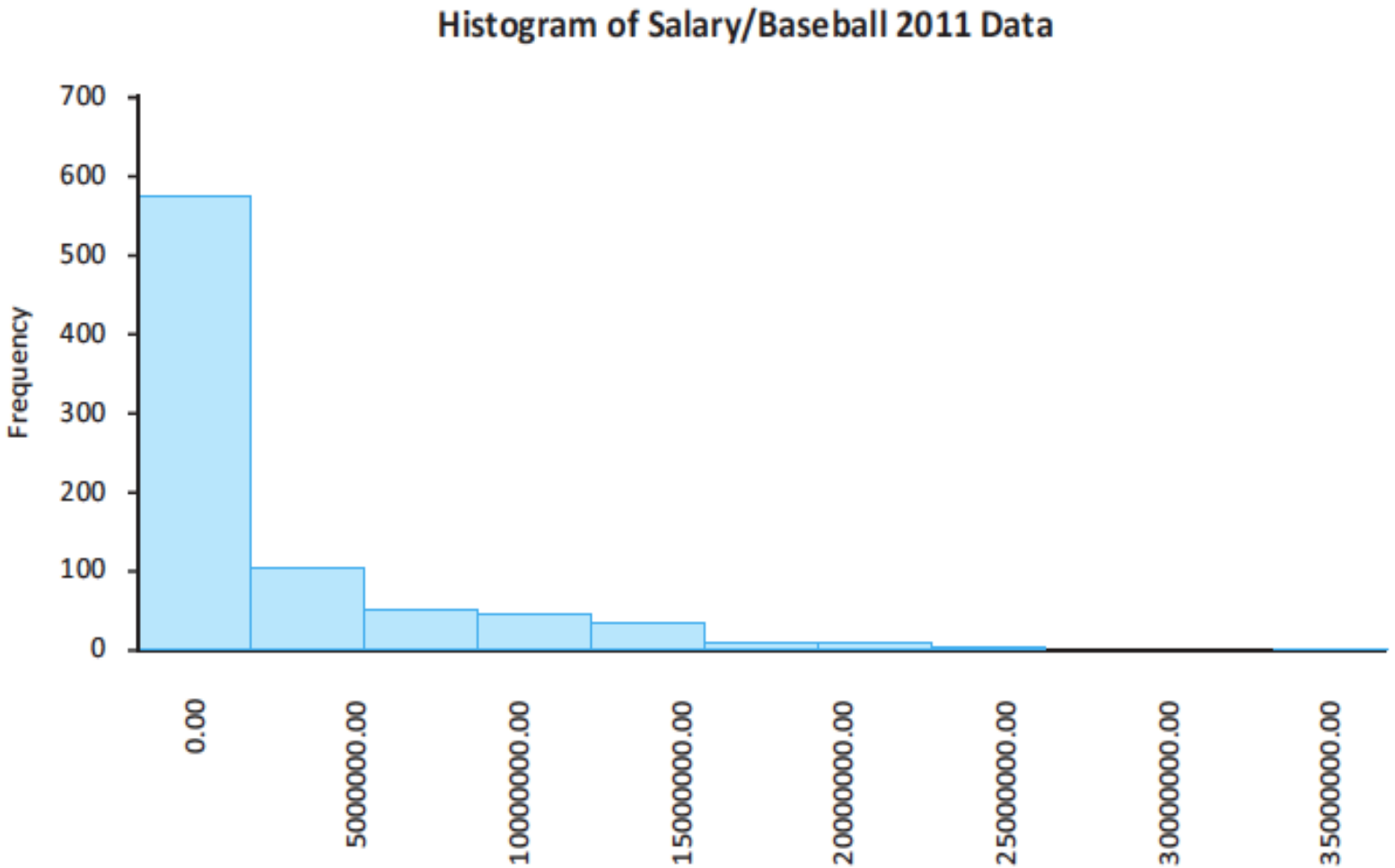
Write down the five number summary

The five number summary is (1;3;7;8;10)The box-and-whisker plot of this data set is given below.





The shape of the salary distribution through a histogram

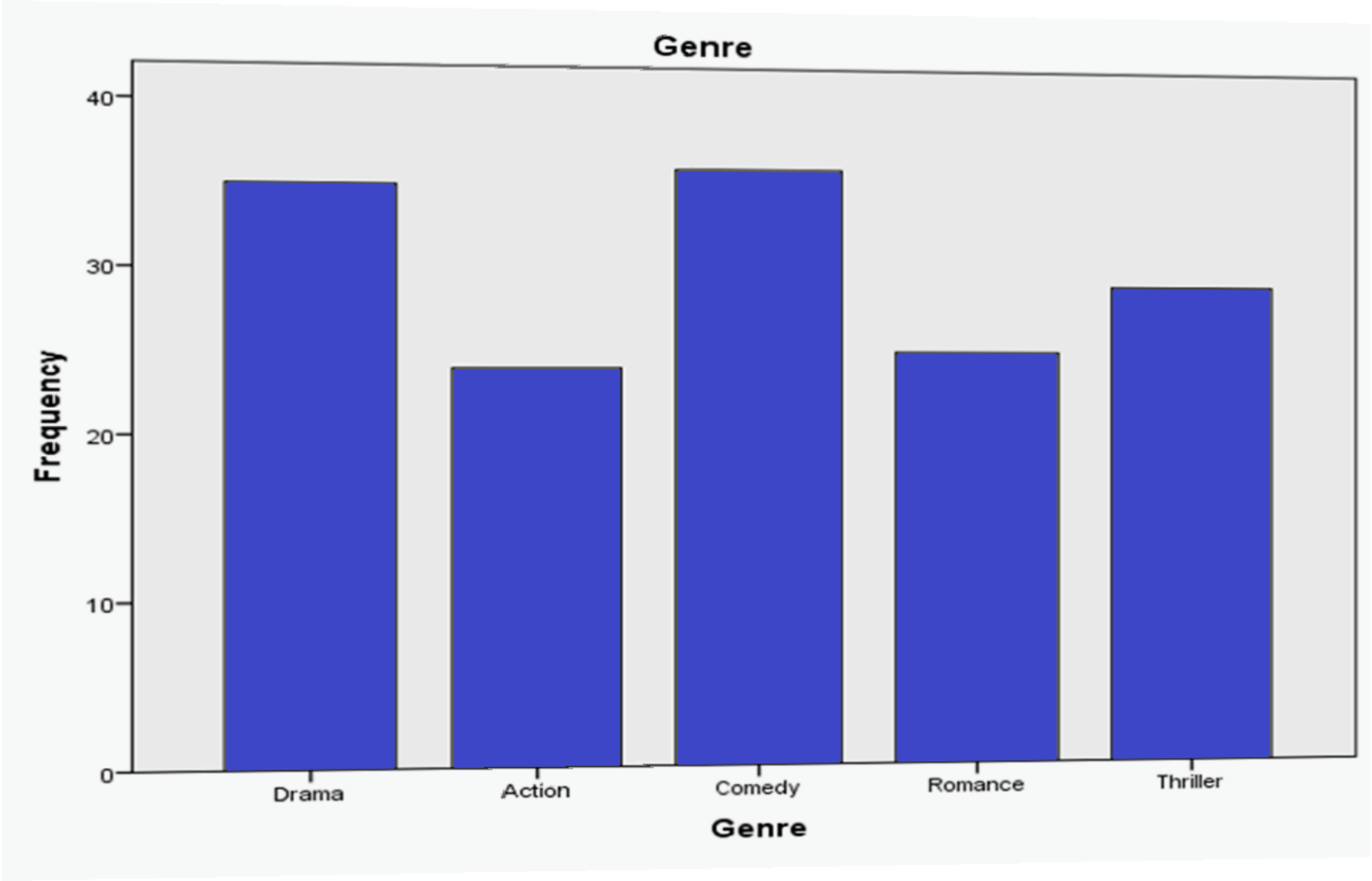


Bar Chart

- 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

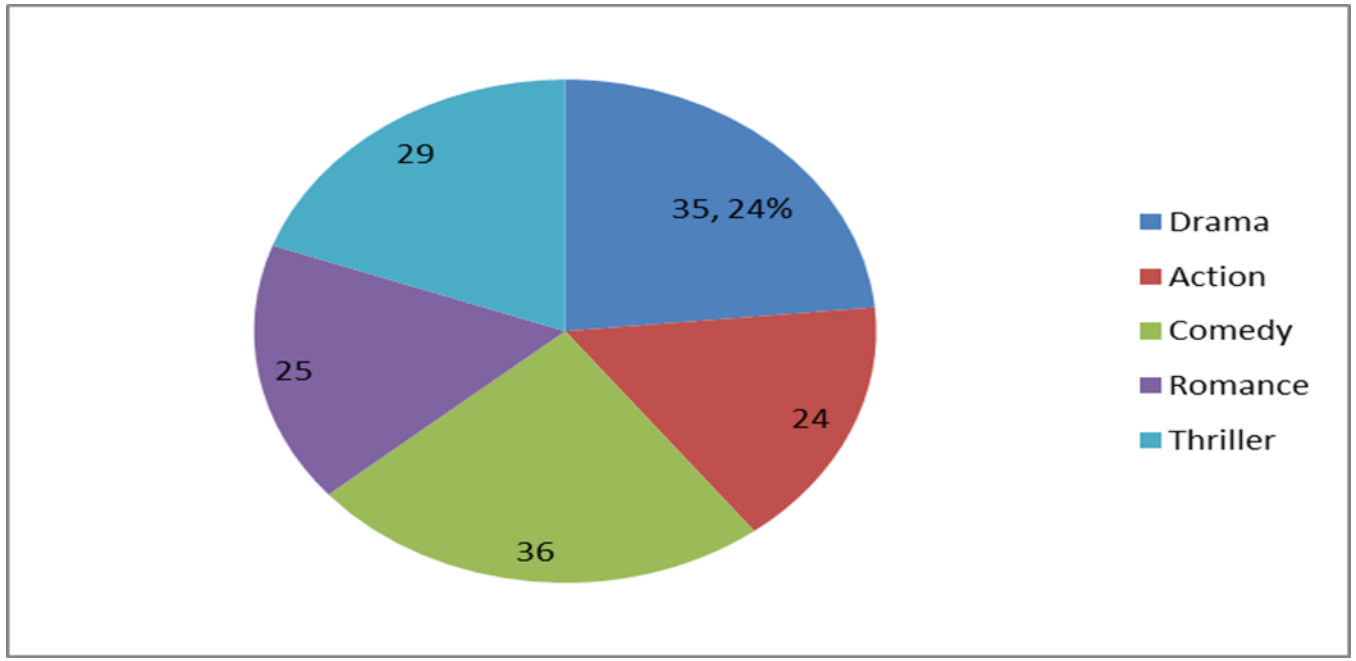
Bar chart for movie genre



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

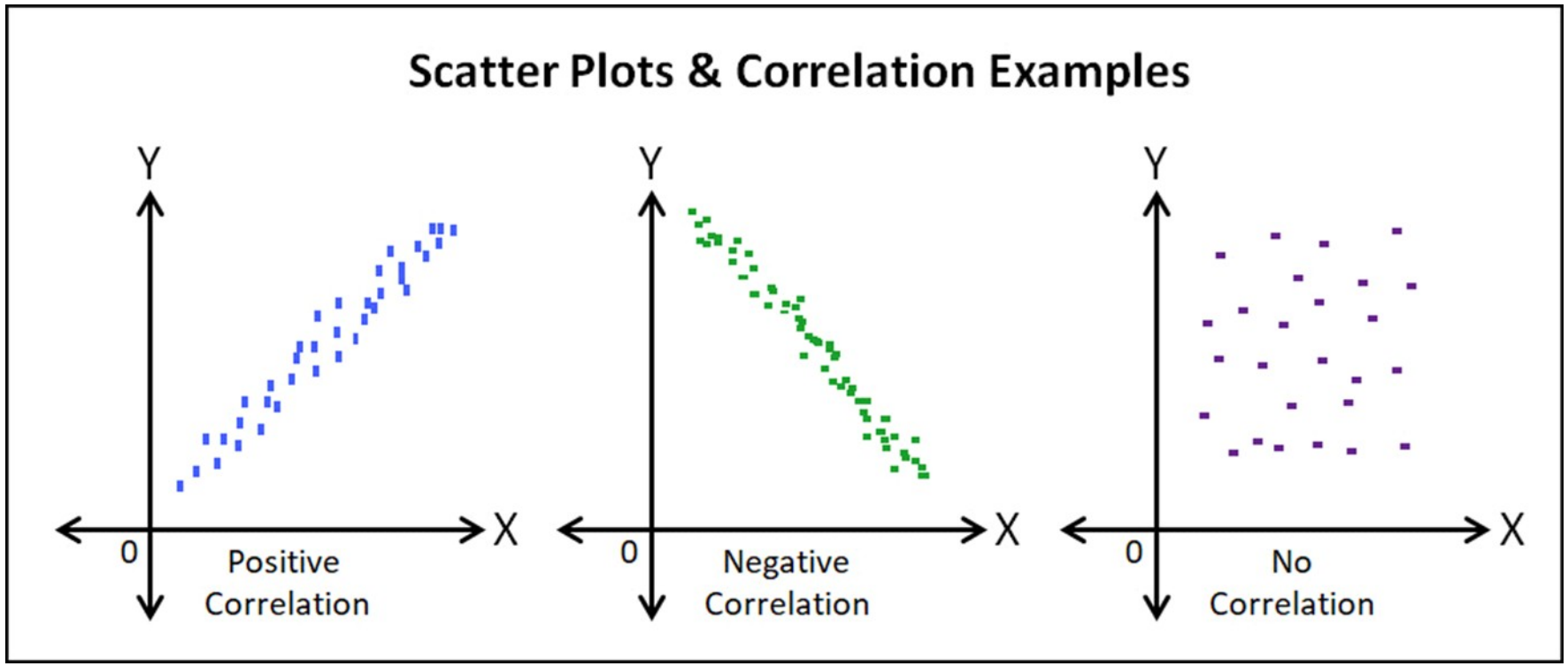
$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

- 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

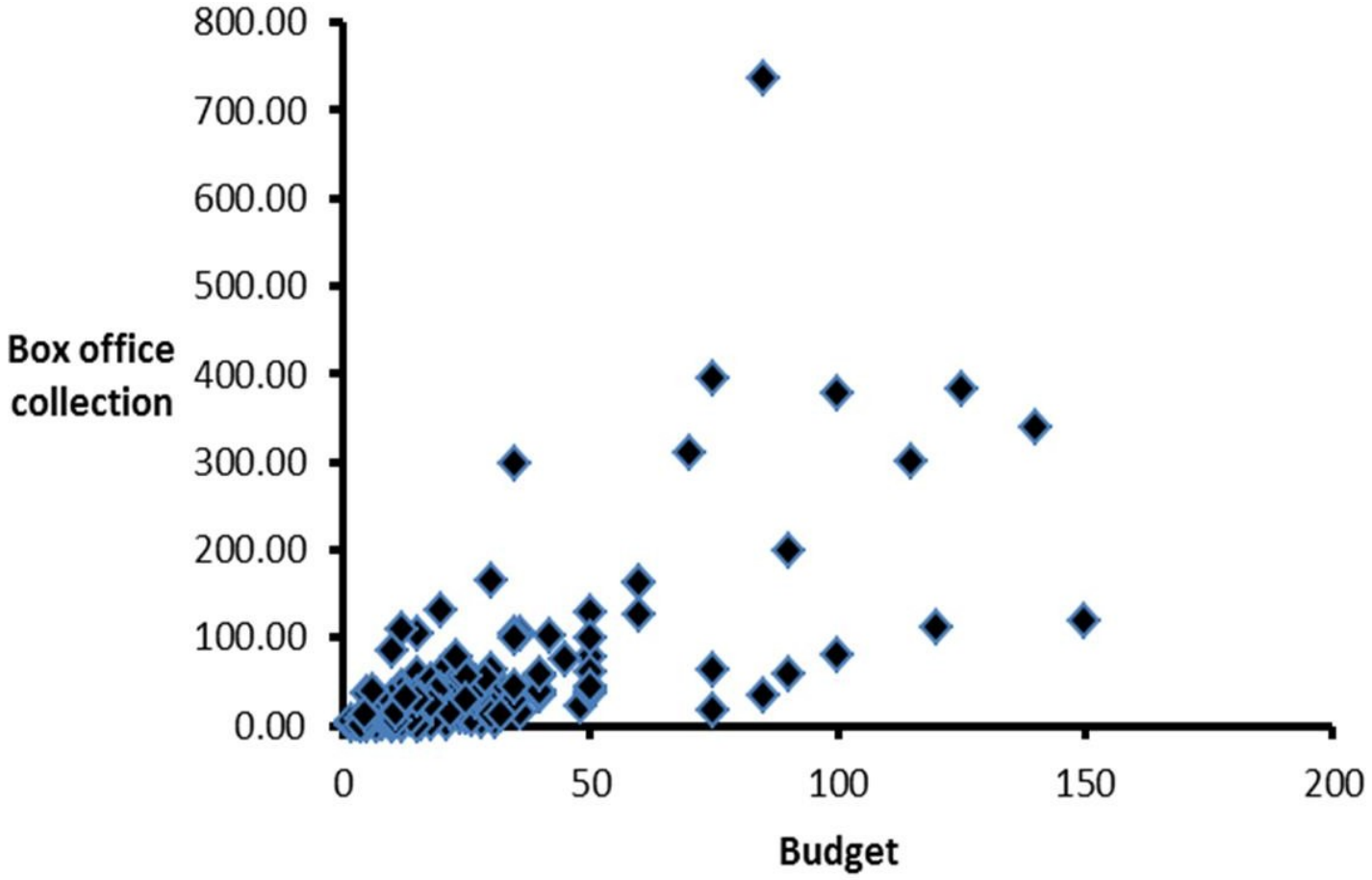
Scatter Plot

- 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



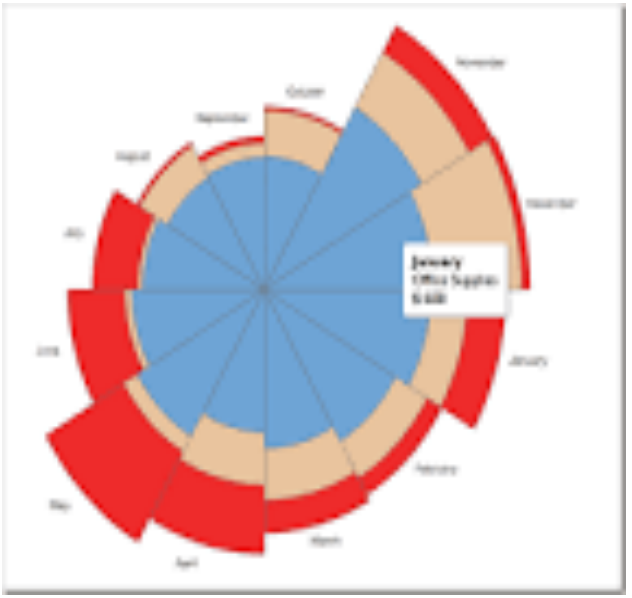
Scatter plot between movie budget and box office collection



DATA ANALYTICS

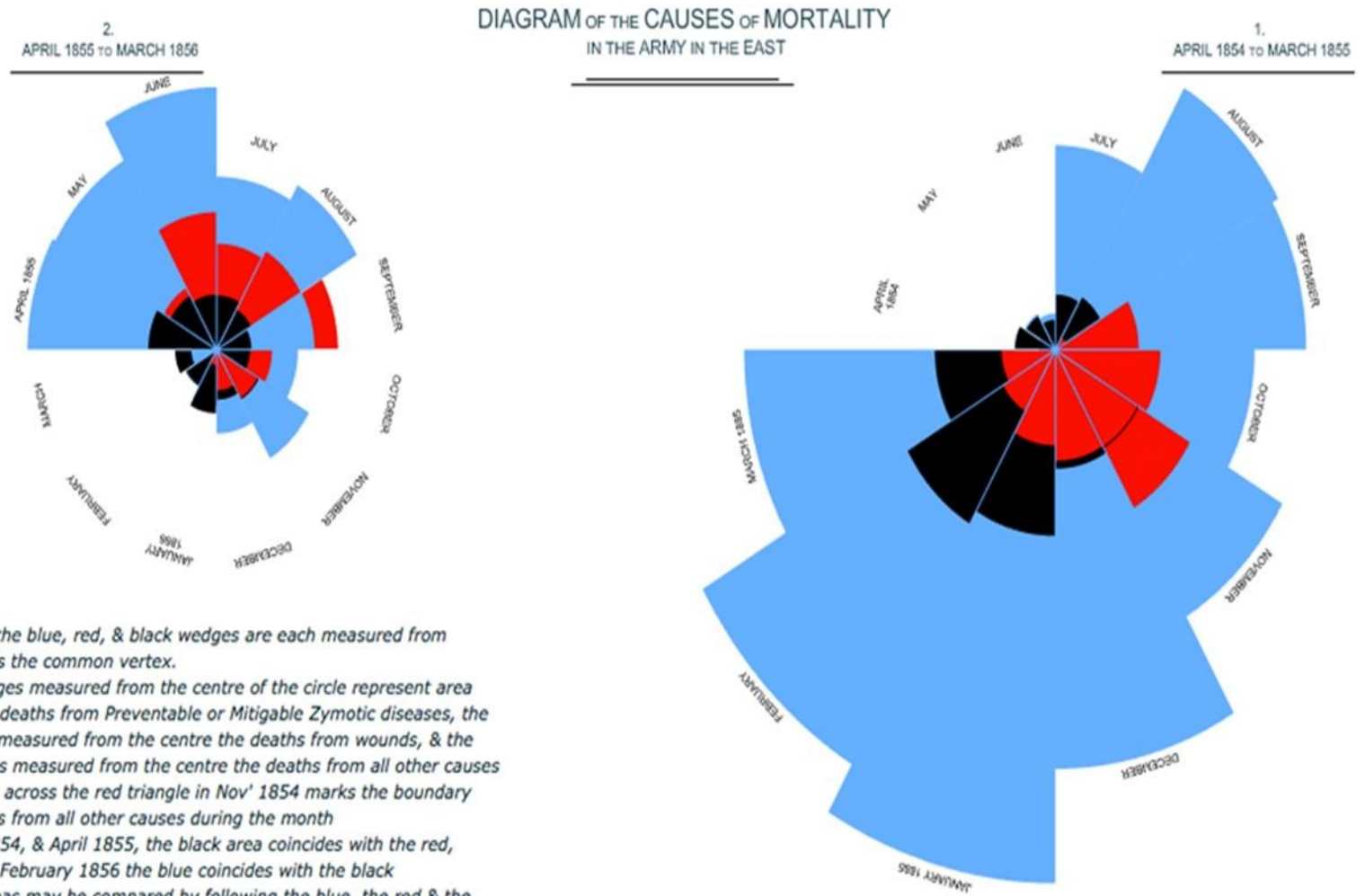
Coxcomb Chart

- 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

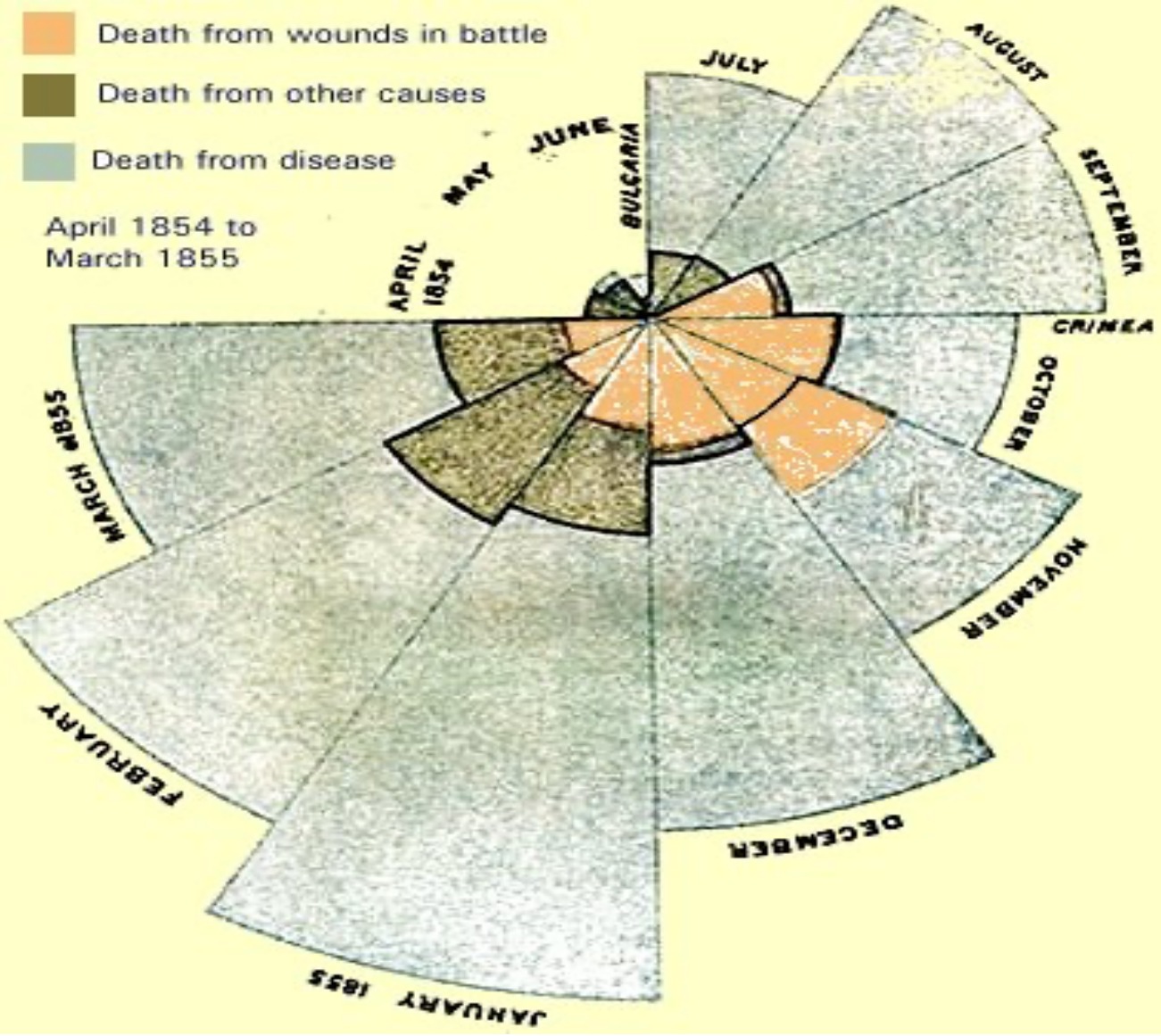
Coxcomb chart on causes of mortality in the army prepared by Florence Nightingale



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.

DATA ANALYTICS

Modified image of the first of the two *Coxcomb Charts* provided by Florence Nightingale in *Notes on Matters Affecting ...* 1858. Crimean War



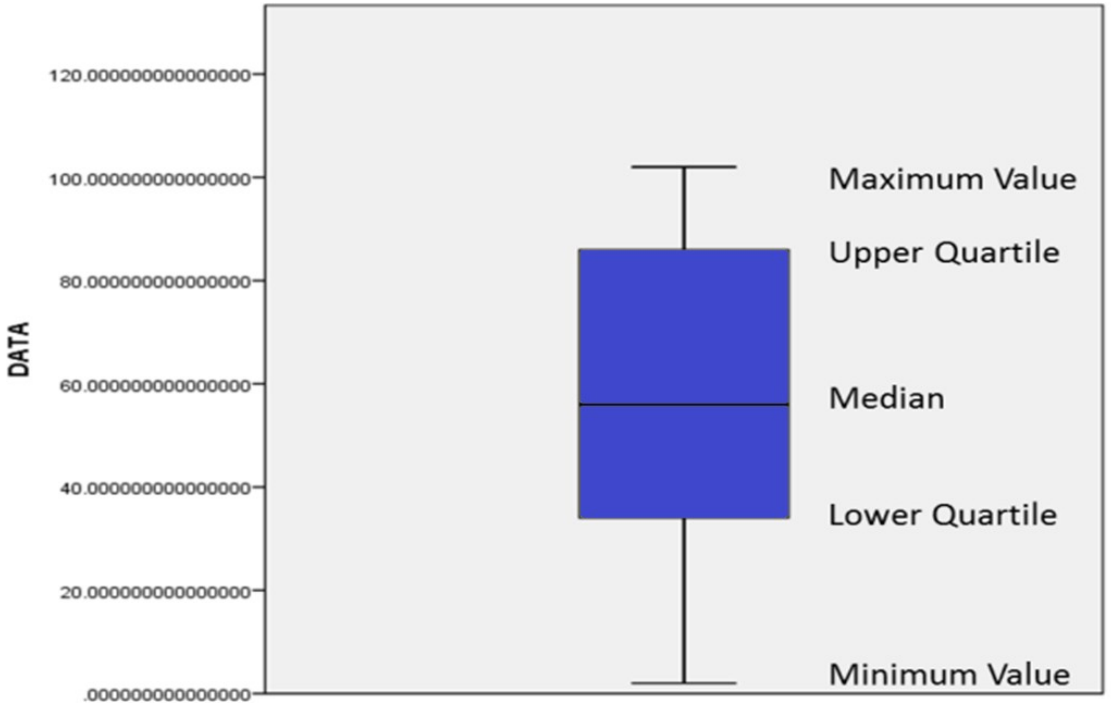
[Florence Nightingale 1820 - 1910]

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).

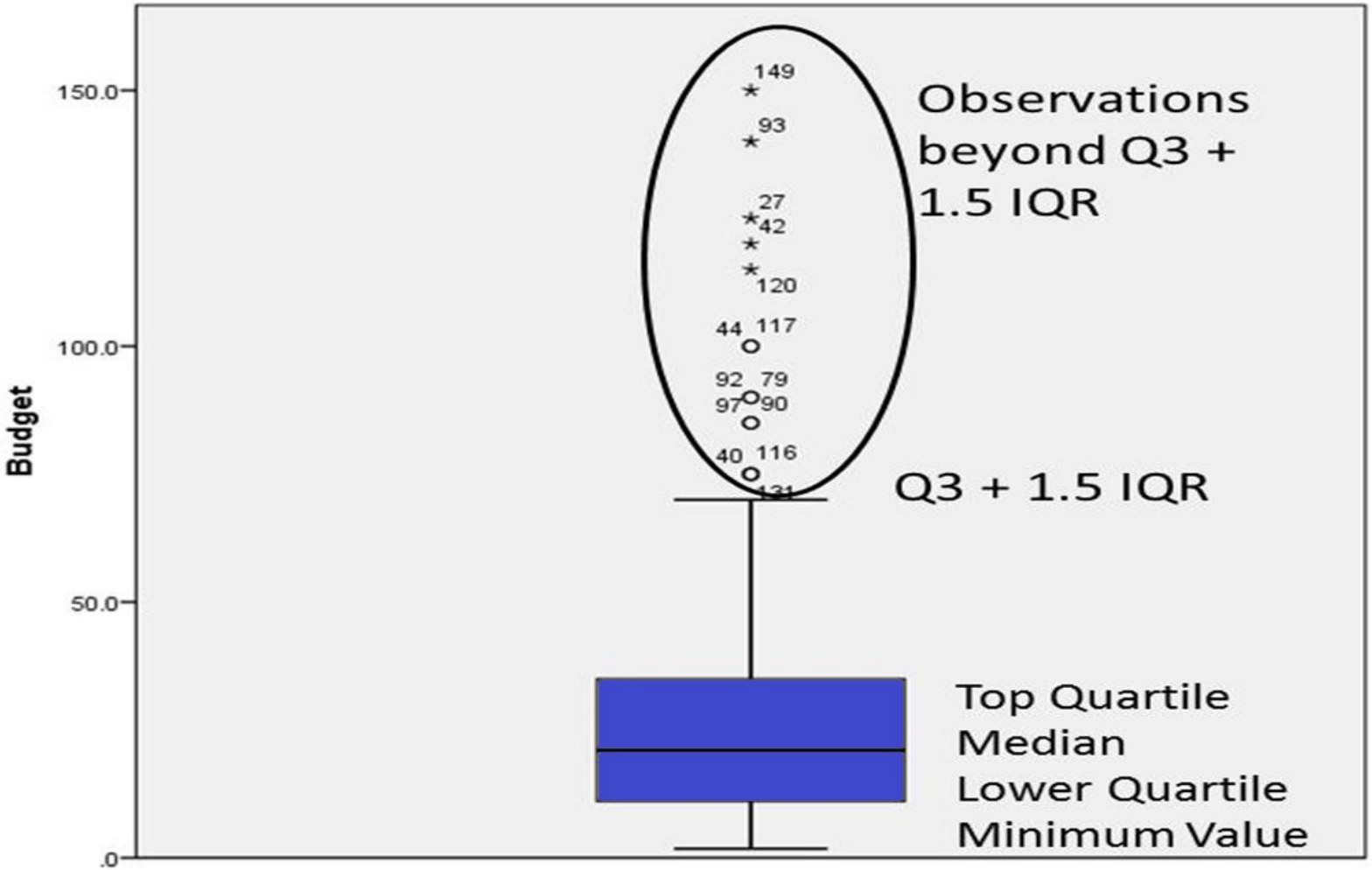
IQR Box Plot

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



Bollywood movie Budget Boxplot

- The box plot for the Bollywood movie budget



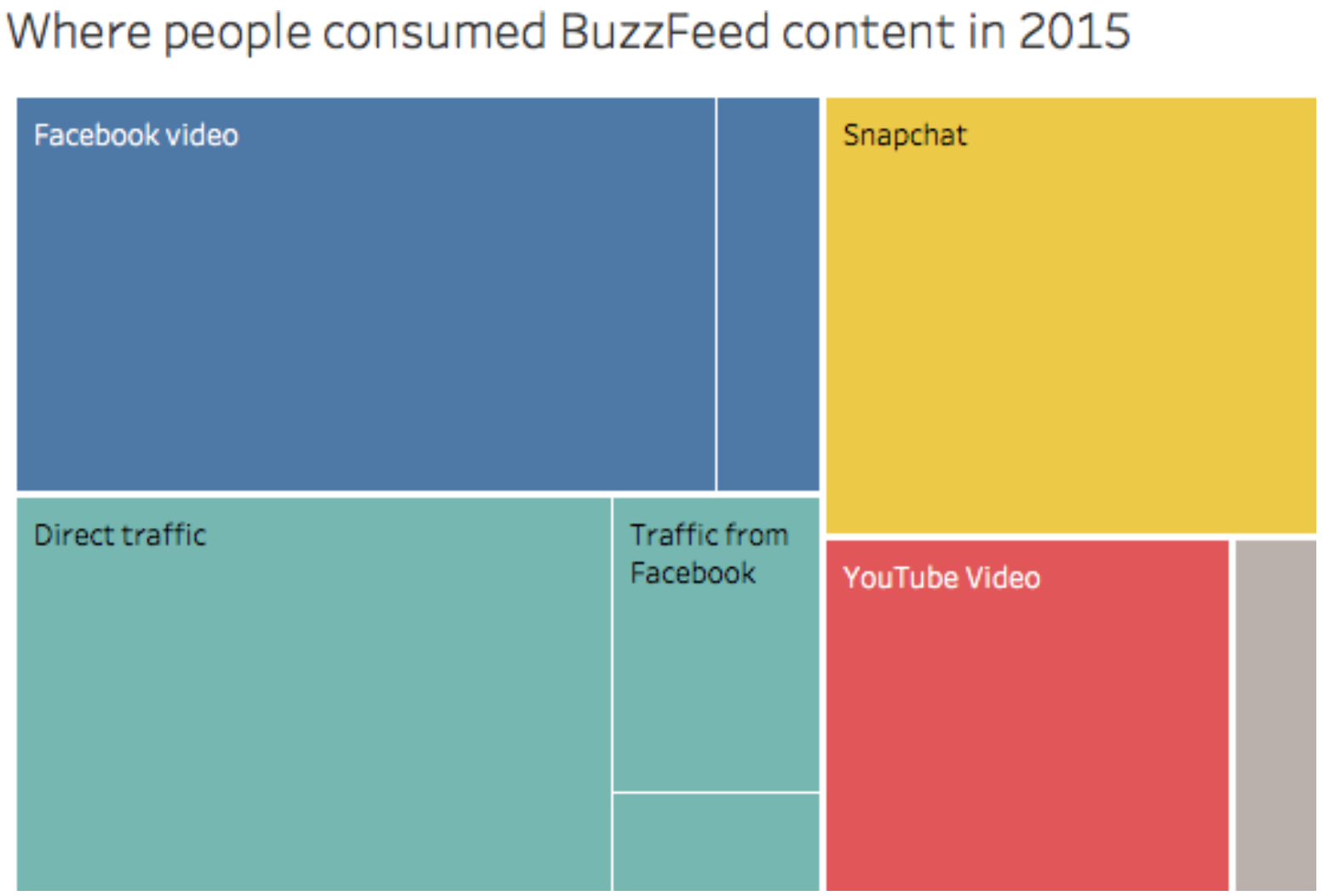
Treemap

- 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



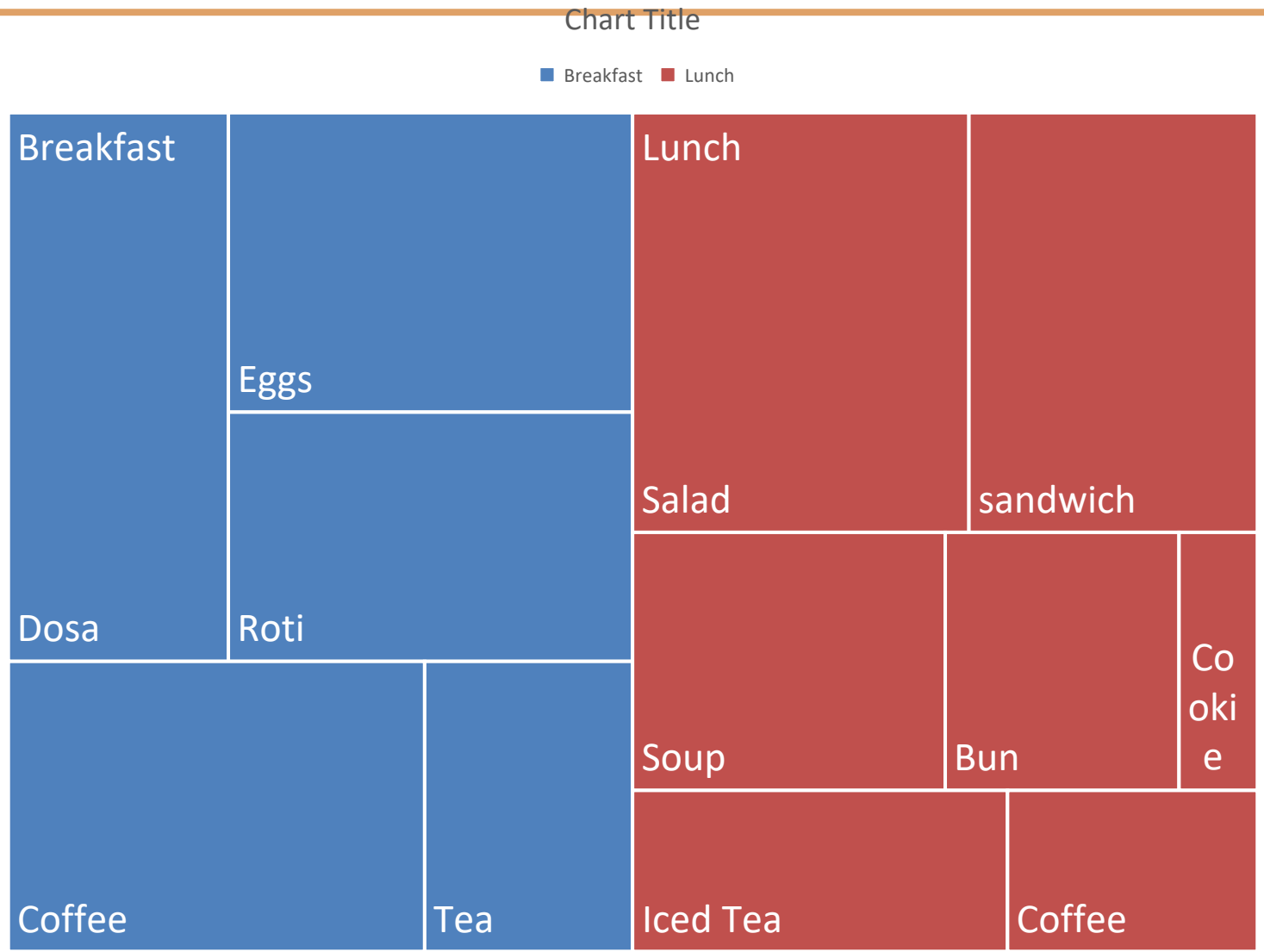
DATA ANALYTICS

Treemap : Example Dataset

Meal	Type	Variety	Price	Quantity	Profit
Breakfast	Beverage	Coffee	20	10	200
Breakfast	Beverage	Tea	30	3	90
Breakfast	Food	Dosa	60	2	120
Breakfast	Food	Roti	50	7	350
Breakfast	Food	Eggs	60	4	240
Lunch	Beverage	Coffee	20	5	100
Lunch	Beverage	Iced Tea	30	15	450
Lunch	Food	Soup	40	4	160
Lunch	Food	sandwich	60	6	360
Lunch	Food	Salad	70	10	700
Lunch	Food	Bun	30	15	450
Lunch	Food	Cookie	10	25	250

DATA ANALYTICS

Treemap : Example Dataset



Summary

1. Histogram and Ogive curve
2. Bar chart
3. Pie chart
4. Scatter plot
5. Coxcomb chart
6. Box plot
7. 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