Mr.Gangadhar Immadi,

Immadi.gangadhar@gmail.com

9986789040

DESCRIPTIVE ANALYTICS

Descriptive Analytics

- Structured / Unstructured data
- Science of describing past data
- Types of Data Measurement Scale
- Population and Sample
- Measures f Central Tendency
- Mesures of Variation
- Data Visualization

Dataset consists of Nominal and Ratio Scale

No.	Gender	Age	Percentage SSC	Board SSC	Percentage HSC	Percentage Degree	Salary
1	M	23	62	Others	88	52	270000
2	M	21	76.33	ICSE	75.33	75.48	220000
3	M	22	72	Others	78	66.63	240000
4	M	22	60	CBSE	63	58	250000
5	M	22	61	CBSE	55	54	180000
6	M	23	55	ICSE	64	50	300000
7	F	24	70	Others	54	65	240000
8	M	22	68	ICSE	77	72.5	235000
9	M	24	82.8	CBSE	70.6	69.3	425000
10	F	23	59	CBSE	74	59	240000

MCT – Mean / Average

- Mathematical average of values and its most frequently used measure
- Population mean μ and Sample mean x

$$\text{Mean}=\bar{x} = \frac{x_1+x_2+\cdots+x_n}{n} = \sum_{i=1}^{n} \frac{x_i}{n}$$

$$\bar{X} = \frac{(270+220+240+250+180+300+240+235+425+240)\times 1000}{10} = 260000$$

$$\sum_{i=1}^{n}\left(X_i - \bar{X}\right) = 0$$

- Suffers from extreme high or low values

Mean of Grouped Data

- Weighted average of class midpoints
- Class frequencies are the weights

$$\mu = \frac{\sum fM}{\sum f} = \frac{\sum fM}{N} = \frac{f_1M_1 + f_2M_2 + f_3M_3 + \cdots + f_iM_i}{f_1 + f_2 + f_3 + \cdots + f_i}$$

Class Interval	Frequency(f)	Class Midpoint(M)	fM
20-under 30	6	25	150
30-under 40	18	35	630
40-under 50	11	45	495
50-under 60	11	55	605
60-under 70	3	65	195
70-under 80	1	75	75
	50		2150

$$\mu = \frac{\sum fM}{\sum f} = \frac{2150}{50} = 43.0$$

Weighted Average

- wish to average numbers
- Assign more importance, or weight, to some of the numbers.

$$Weighted\ \ Average = \frac{\sum xw}{\sum w}$$

Suppose your midterm test score is 83 and your final exam score is 95. Using weights of 40% for the midterm and 60% for the final exam, compute the weighted average of your scores. If the minimum average for an A is 90, will you earn an A?

$$Weighted\ \ Average = \frac{(83)(0.40)+(95)(0.60)}{0.40+0.60}$$

$$= \frac{32+57}{1} = \quad 90.2$$

MCT - Median (or Mid) Value

- Median is the value that **divides** the data into two equal parts
- When n is **odd** value at position **(n + 1)/2** when n is odd
- When n is **even**, the median is the **average** value of **(n/2)th and (n + 2)/2th**
- Number of deposits in a Bank

Day	1	2	3	4	5	6	7
Number of Deposits	245	326	180	226	445	319	260

- 180, 226, 245, 260, 319, 326 ,445,451
- $(n + 1)/2 = (8/2) = 4$

MCT - **Median of Grouped Data**

$$Median = L + \frac{\frac{N}{2} - cf_p}{f_{med}}(W)$$

- ✓ L the lower limit of the median class
- ✓ cf_p = cumulative frequency of class preceding the median class
- ✓ f_{med} = frequency of the median class
- ✓ W = width of the median class
- ✓ N = total of frequencies

Class Interval	Frequency	Cumulative Frequency
20-under 30	6	6
30-under 40	18	24
40-under 50	11	35
50-under 60	11	46
60-under 70	3	49
70-under 80	1	50
	N = 50	

$$= 40 + \frac{\frac{50}{2} - 24}{11}(10)$$

$$= 40.909$$

MCT - Mode

- **Most frequently** occurring **value** in the dataset
- Only measure of central tendency which is valid for qualitative (nominal) data
- Bimodal,Multimodal, No Mode
- For example, (a) Married, (b) Unmarried d (d) Divorced Female.
- Applicable for all types of data scales
- Mode :44

35	41	44	45
37	41	44	46
37	43	44	46
39	43	44	46
40	43	44	46
40	43	45	48

MCT - **Mode of Grouped Data**

- Midpoint of the modal class
- Modal class has the greatest frequency

$$Mode = L_{Mo} + \left(\frac{d_1}{d_1 + d_2}\right)w$$

Class Interval	Frequency
20-under 30	6
30-under 40	18
40-under 50	11
50-under 60	11
60-under 70	3
70-under 80	1

$$30 + \left(\frac{12}{12 + 7}\right)10$$

$$= 36.31$$

MCT – **Percentile,Decile, Quartile**

- Frequently used to identify the **position of the observation in the dataset**(student position)
- P_x, is the value of the data at which ***x* percentage of the data lie below that value**
- Position corresponding to $P_x \approx x\,(n+1)/100$
- P_x is the position in the data calculated , where *n* is the number of observations in the data.
- **Decile** divide the data into 10 equal parts. First decile contains first 10% of the data and **second decile contains first 20%** of the data and so on.

- **Quartile** divides the data into 4 equal parts.
- Example - Time between failures of wire-cut (in hours)

2	**22**	**32**	**39**	**46**	**56**	**76**	**79**	**88**	**93**
3	24	33	44	46	66	77	79	89	99
5	24	34	45	47	67	77	86	89	99
9	26	37	45	55	67	78	86	89	99
21	31	39	46	56	75	78	87	90	102

1. Calculate the mean, median, and mode of time between failures of wire-cuts
2. The company would like to know by what time 10% (ten percentile or P_{10}) and 90% (ninety percentile or P_{90}) of the wire-cuts will fail?
3. Calculate the values of P_{25} and P_{75}.

Solution

- **Mean = 57.64**, **median = 56**, and **mode = 46,89,99**
- The position of $\boldsymbol{P_{10}}$ **= 10 × (51)/100 = 5.1** round off to 5 and value at 5th position is 21
- $P_{10} = 10 \times (51)/100 = 5.1$
  - Approximated as **21 + 0.1 × (value at 6th position – value at 5th position)**

    $= 21 + 0.1(1) = 21.1$
- $P_{90} = 90 \times 51/100 = 45.9$
  - Apprimated as- $90 + 0.9 \times (3) = 92.7$

- $\boldsymbol{P_{25}}$ (1st Quartile or Q_1) = 25 × 51/100 = 12.75 , Value at 12th position is

  = 33

  P_{25} = 33 + 0.75 (value at 13th position – value at 12th position) = 33 + 0.75 (1) = 33.75

- $\boldsymbol{P_{75}}$ (3rd Quartile or Q_3)

  = 75 × 51/100 = 38.25

  Value at 38th position is

  = 86

- P_{75} = 86 + 0.25 (value at 39th position – value at 38th position) = 86 + 0.25 (0) = 86

Measures of Variation / DIspersion

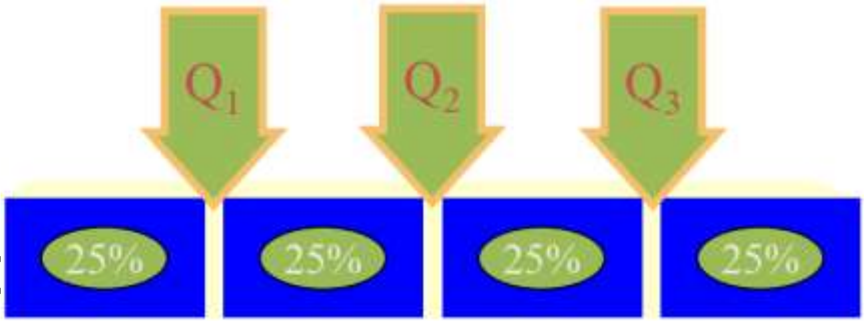
- Describe the spread or the dispersion of a data
- Reliability of measure of central tendency
- To compare dispersion of various samples
- Measures –
  - Range
  - Inter-quartile range
  - Mean Absolute Deviation
  - Variance
  - Standard Deviation
  - Z scores
  - Coefficient of Variation

35	41	44	45
37	41	44	46
37	43	44	46
39	43	44	46
40	43	44	46
40	43	45	48

- **Range** is the difference between maximum and minimum value of the data
  - Ignores all values except extreme values
  - Range = Largest – Smallest = 48 - 35 = 13

MOD – Inter Quartile Range / Distance

- Measure of the distance be and Quartile 3 (Q_3)
- Q1 is equal to the 25th perc



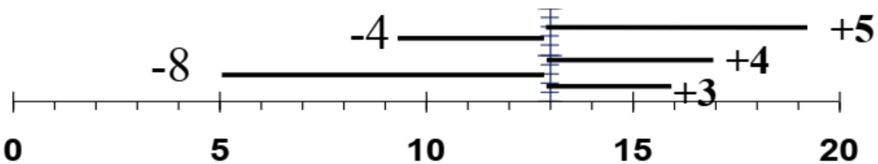
- Quartile values are not necessarily members of the data set
- Ordered array: 106, 109, 114, 116, 121, 122, 125, 129
- Q1 : $i = \frac{25}{100}(8) = 2 \quad Q_1 = \frac{109+114}{2} = 111.5$
- Q2: $i = \frac{50}{100}(8) = 4 \quad Q_2 = \frac{116+121}{2} = 118.5$
- Q3: $i = \frac{75}{100}(8) = 6 \quad Q_3 = \frac{122+125}{2} = 123.5$
- Less influenced by extremes

$$\boxed{Interquartile\ Range = Q_3 - Q_1}$$

MOD - Deviation from mean / Mean Absolute Deviation

- Data set: 5, 9, 16, 17, 18
- Deviations from the mean: -8, -4, 3, 4,

$$\mu = \frac{\sum X}{N} = \frac{65}{5} = 13$$



- Average of the absolute deviations from the mean

X	$X - \mu$	$\lvert X - \mu \rvert$
5	-8	+8
9	-4	+4
16	+3	+3
17	+4	+4
18	+5	+5
	0	24

$$M.A.D. = \frac{\sum \lvert X - \mu \rvert}{N}$$

$$= \frac{24}{5}$$

$$= 4.8$$

MOD - **Population Variance**

- Average of the squared deviations from the arithmetic mean

X	$X-\mu$	$(X-\mu)^2$
5	-8	64
9	-4	16
16	+3	9
17	+4	16
18	+5	25
	0	130

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N}$$

$$= \frac{130}{5}$$

$$= 26.0$$

MOD - **Population Standard Deviation**

- Square root of the variance

X	$X-\mu$	$(X-\mu)^2$
5	-8	64
9	-4	16
16	+3	9
17	+4	16
18	+5	25
	0	130

$$\sigma^2 = \frac{\sum(X-\mu)^2}{N}$$

$$= \frac{130}{5} \quad = 26.0$$

$$\sigma = \sqrt{\sigma^2}$$

$$= \sqrt{26.0}$$

$$= 5.1$$

MOD – Sample Variance / SD

- Average of the squared deviations from the arithmetic mean

$$S^2 = \frac{\sum\left(X - \bar{X}\right)^2}{n-1}$$

X	$X - \bar{X}$	$(X - \bar{X})^2$
2,398	625	390,625
1,844	71	5,041
1,539	-234	54,756
1,311	-462	213,444
7,092	0	663,866

$$= \frac{663,866}{3}$$

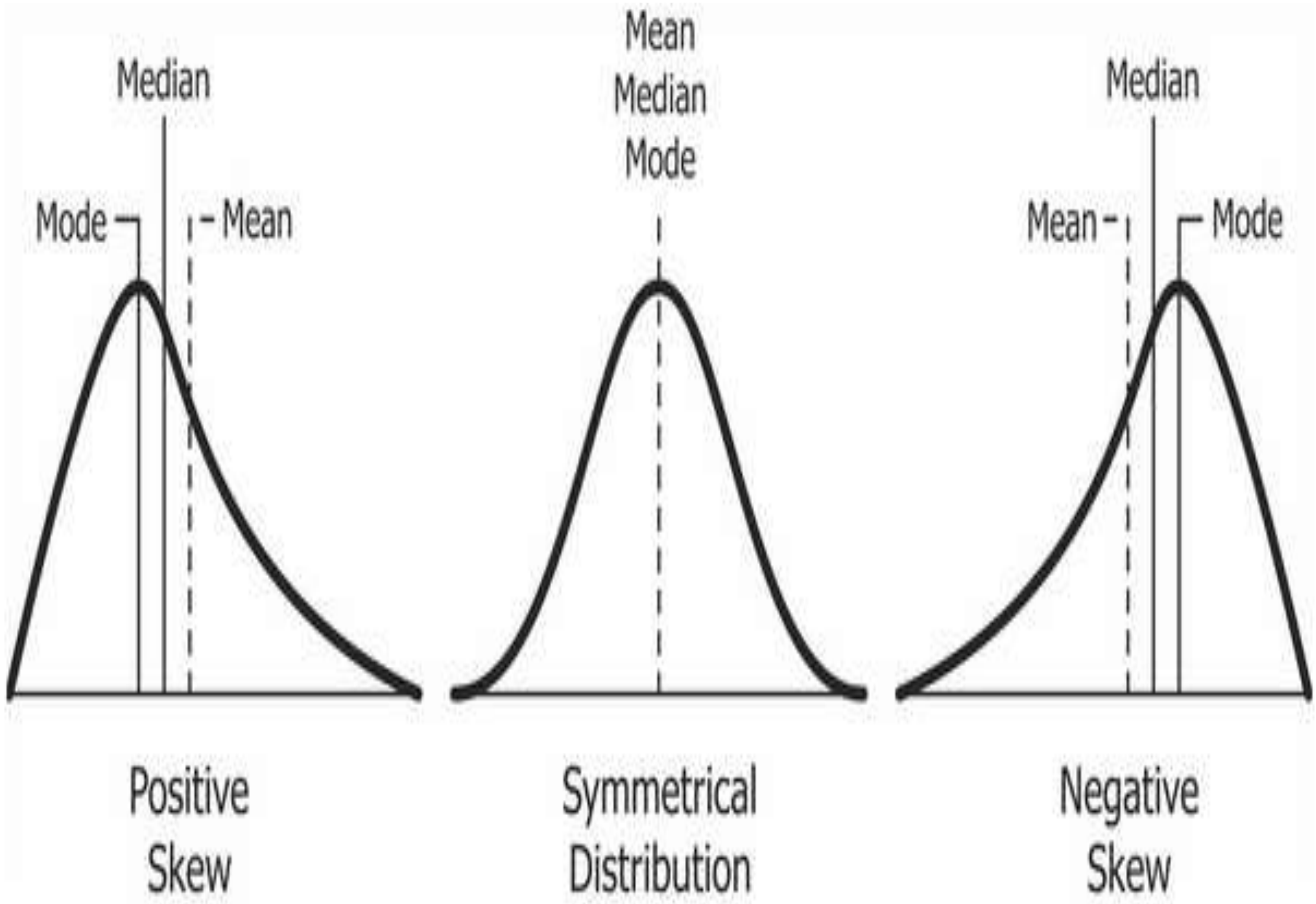
$$= 221,288.67$$

$$S = \sqrt{S^2}$$

$$= \sqrt{221,288.67}$$
$$= 470.41$$

Uses of Standard Deviation

- Indicator of financial risk
- Quality Control
  - construction of quality control charts
  - process capability studies
- Comparing populations
  - household incomes in two cities
  - employee absenteeism at two plants


Median
Mode
Mean
Mean
Median
Mode
Median
Mean
Mode
Positive Skew
Symmetrical Distribution
Negative Skew

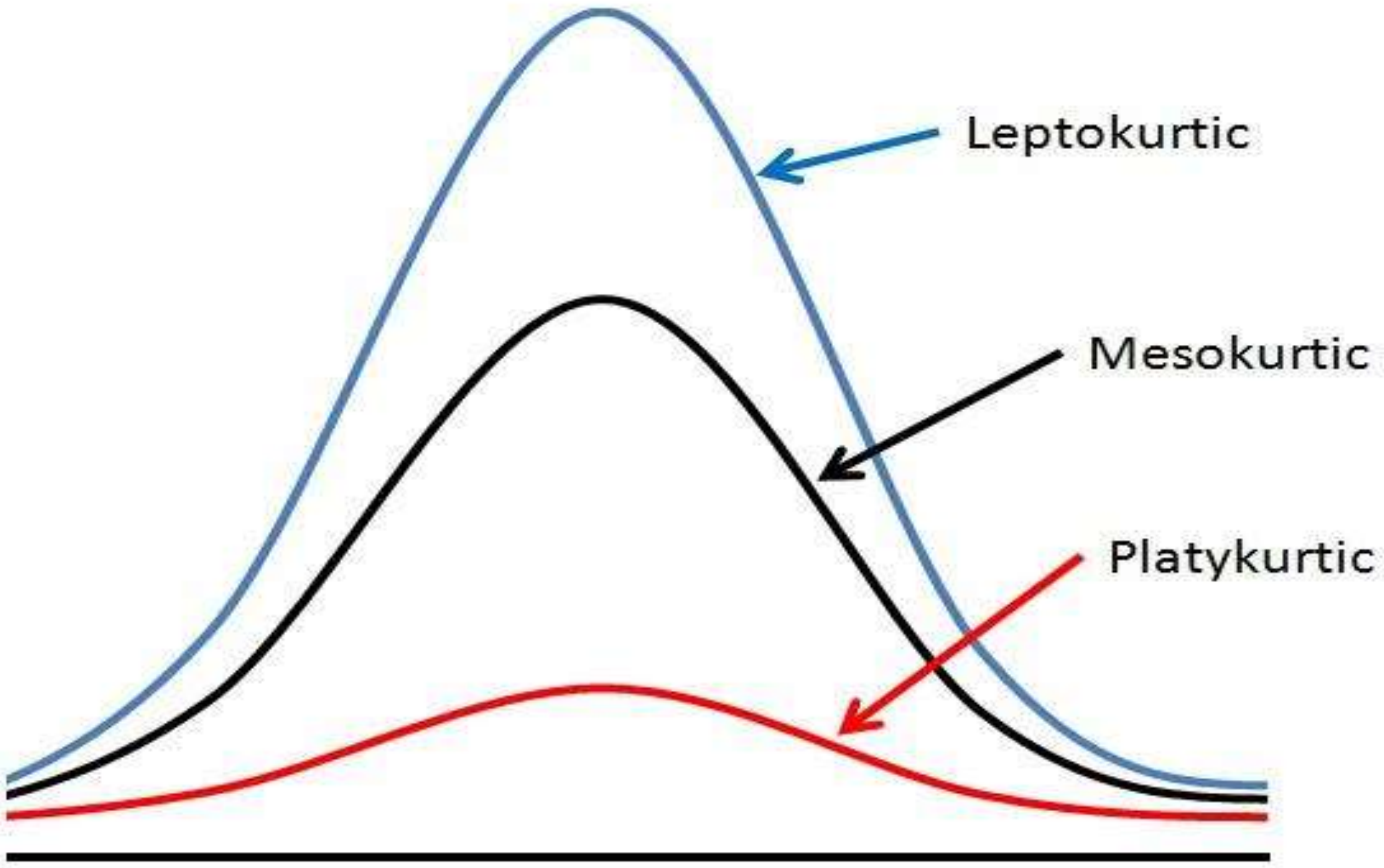
Kurtosis

- **Kurtosis** is another measure of shape, aimed at shape of the tail,
- Checks whther the tail of the data distribution is heavy or light.

$$\text{Kurtosis} = \frac{\sum_{i=1}^{4}\left(X_i - \bar{X}\right)^4 / n}{\sigma^4}$$

- Kurtosis value < 3 --- > **platykurtic distribution**
- Kurtosis value > 3 --- > **leptokurtic distribution.**
- kurtosis value = 3 --- > **standard normal distribution (mesokurtic)**

Leptokurtic, mesokurtic, and platykurtic distributions



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).

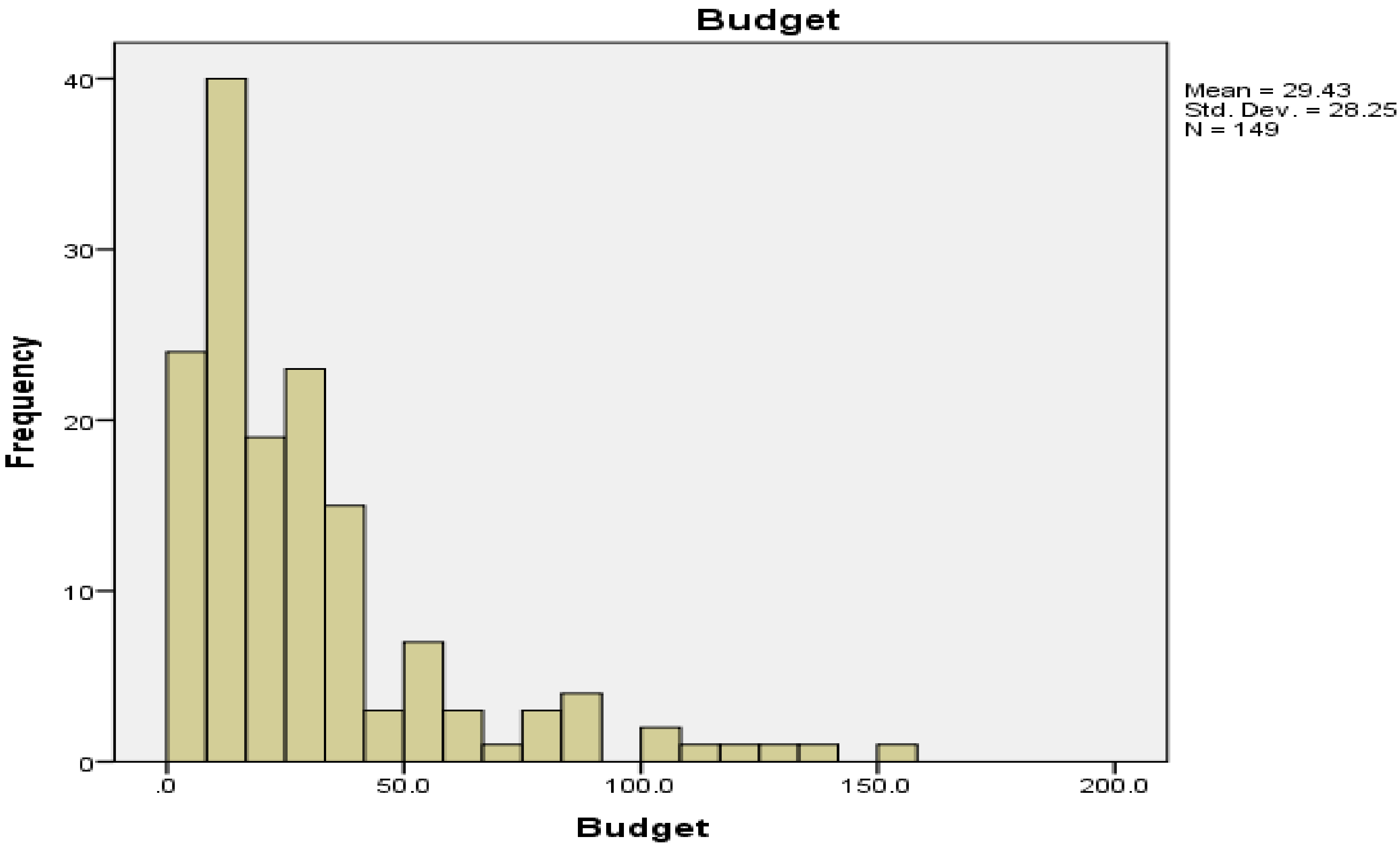
3) Create a frequency distribution (bin in the horizontal axis and frequency in the vertical axis) using the information obtained in steps 1 and 2

Use of Histogram

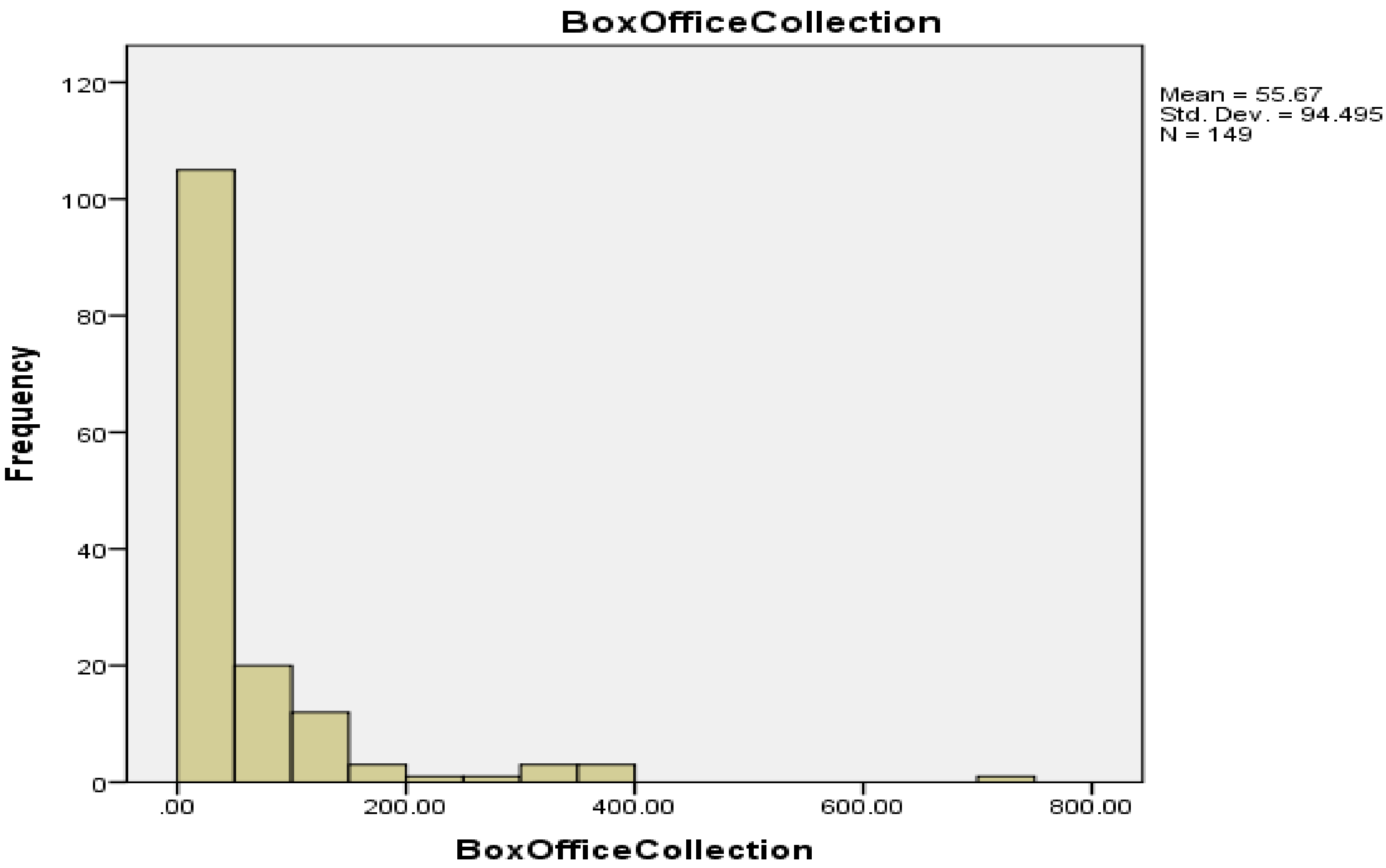
Histogram is very useful since it assists data scientist to identify the following:

- 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

Histogram of Bollywood movie budget

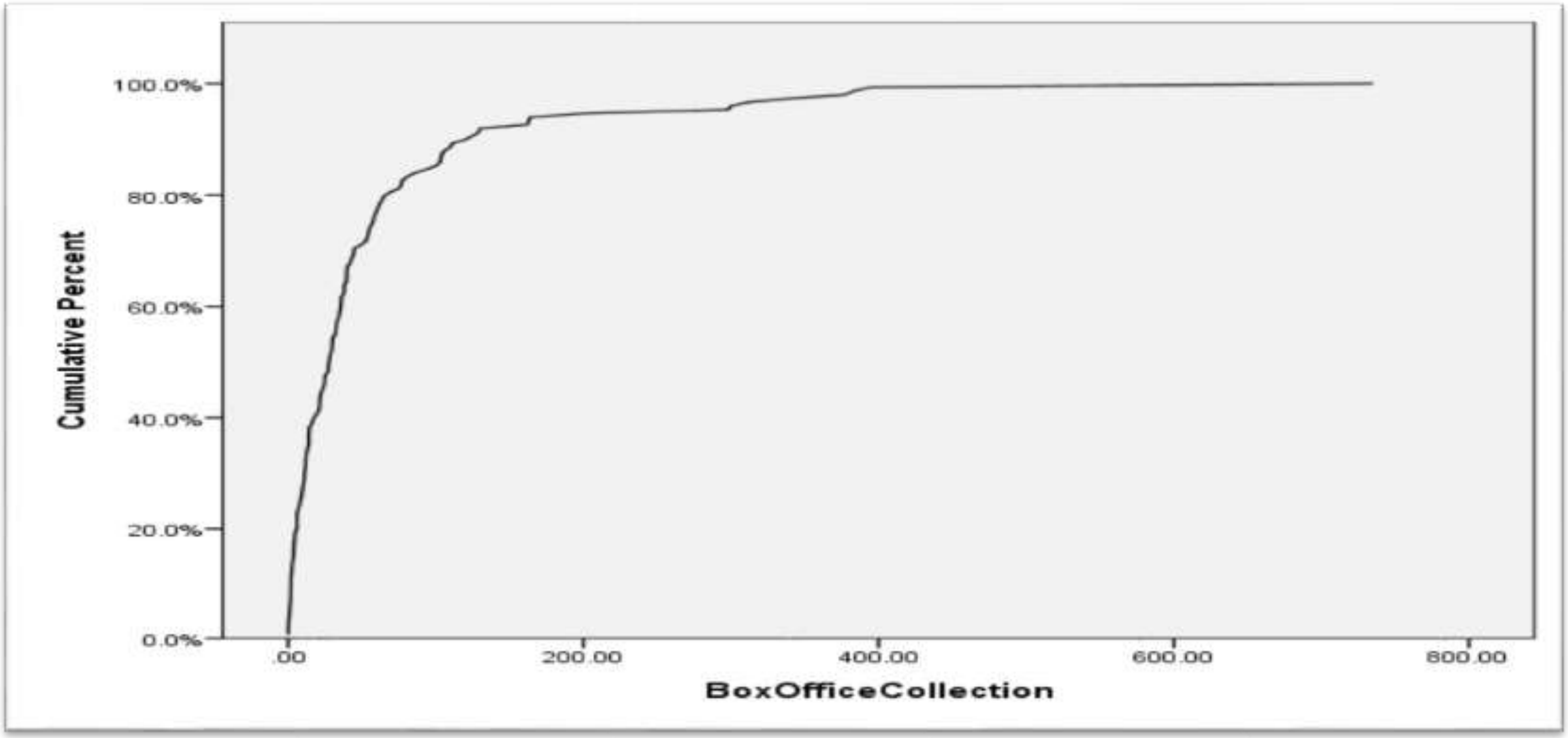


Histogram of Bollywood movie box-office collection

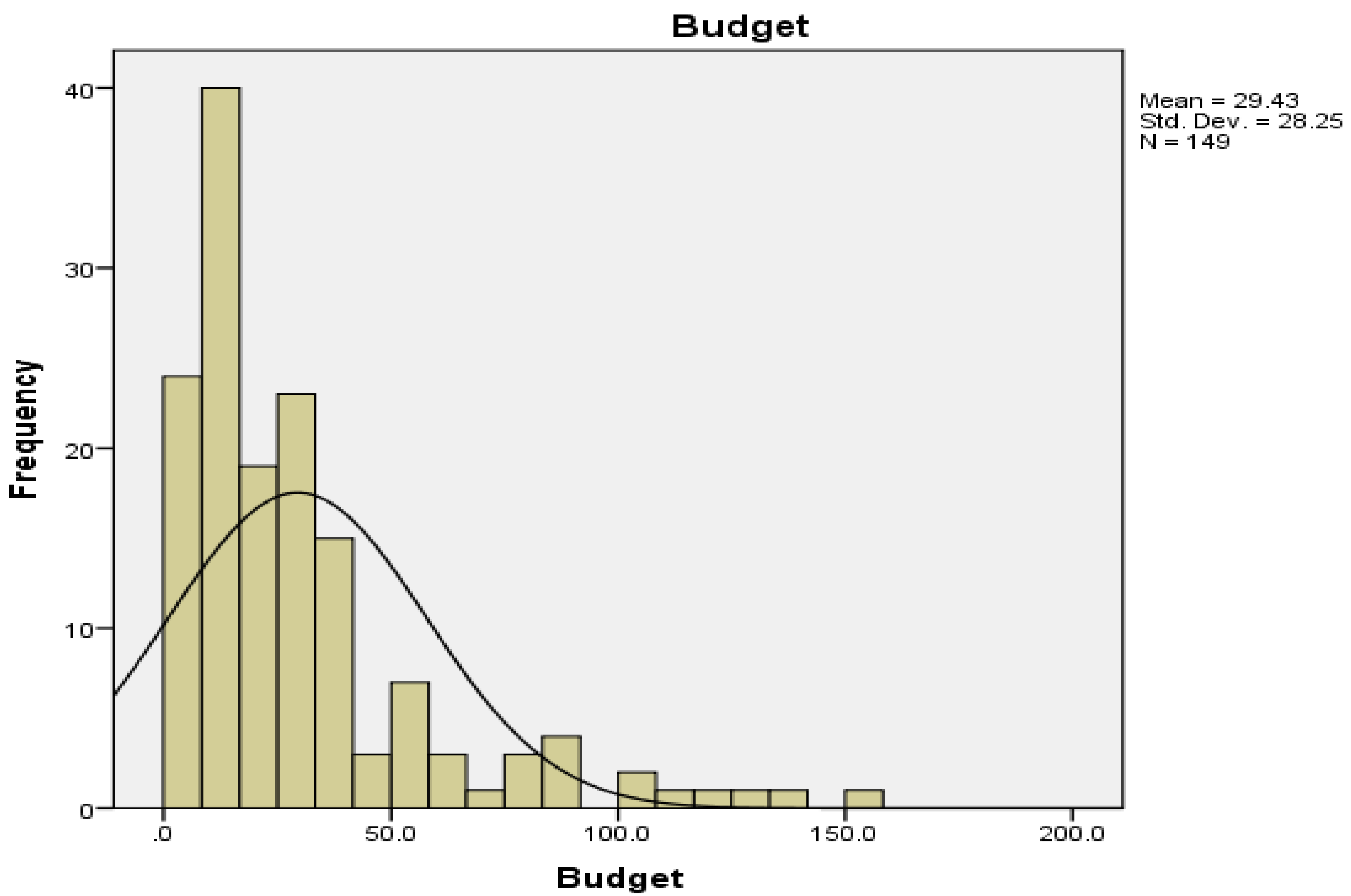


Ogive Curves

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



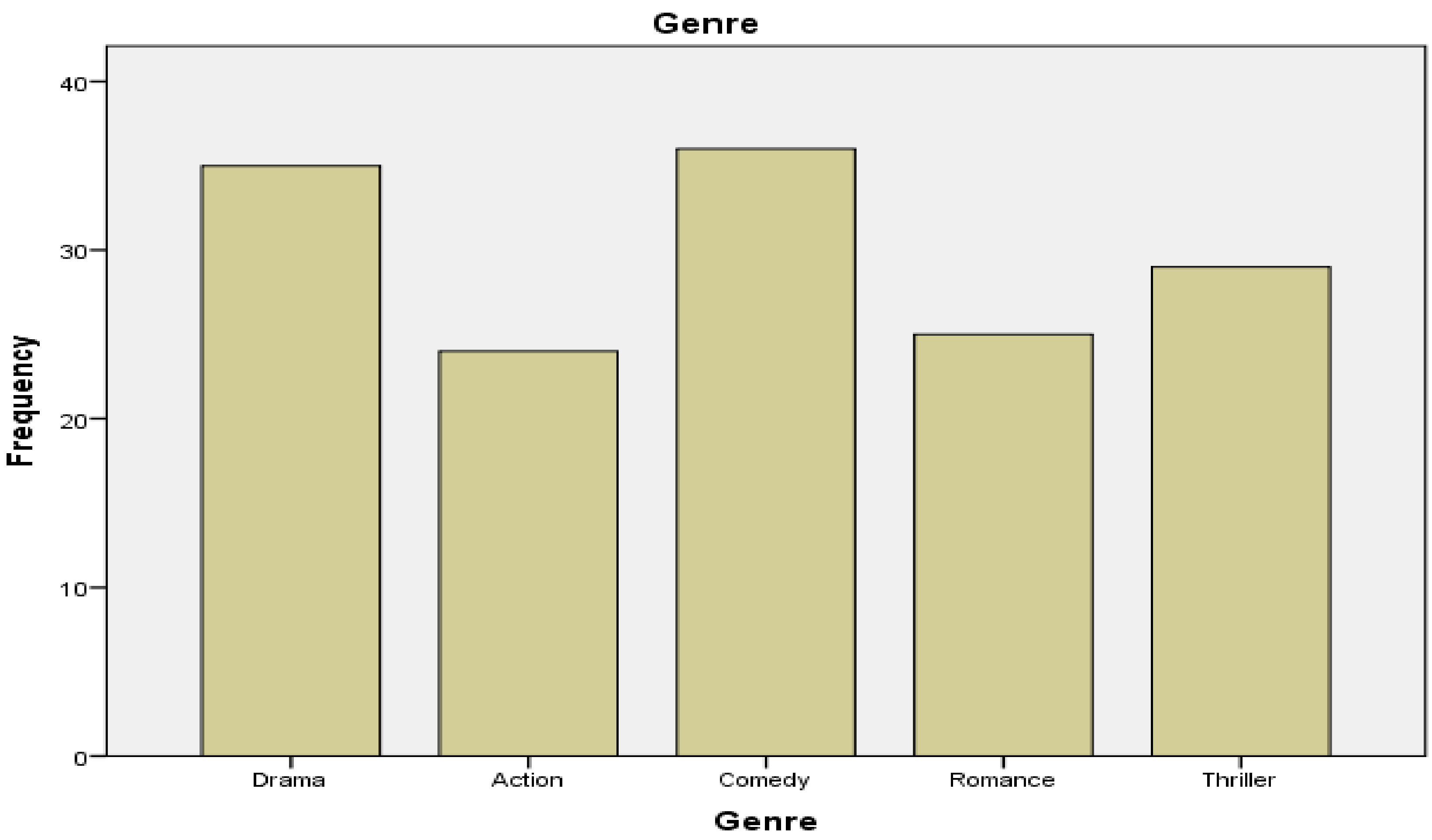
Histogram of Bollywood movie budget along with normal distribution frequency



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

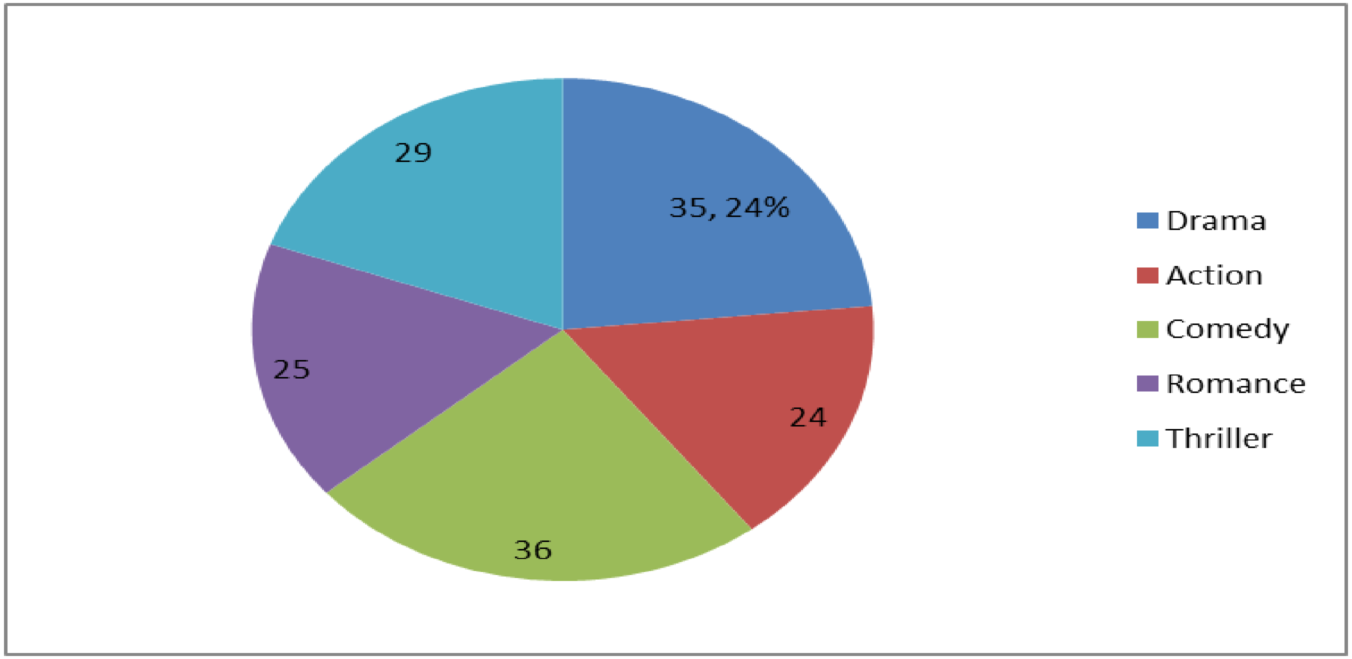
Bar chart for movie genre



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



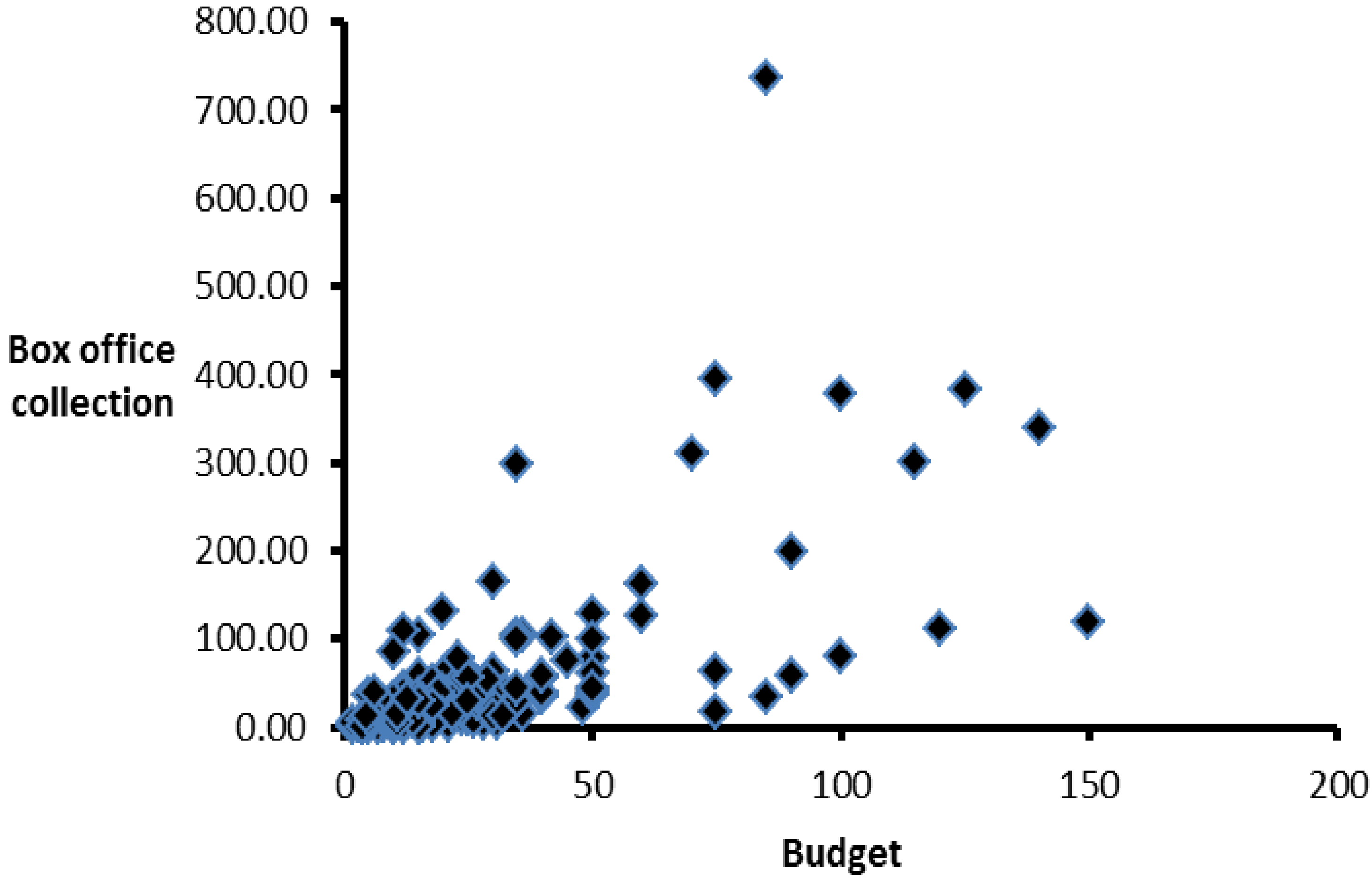
Scatter Plot

- **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 between movie budget and box office collection

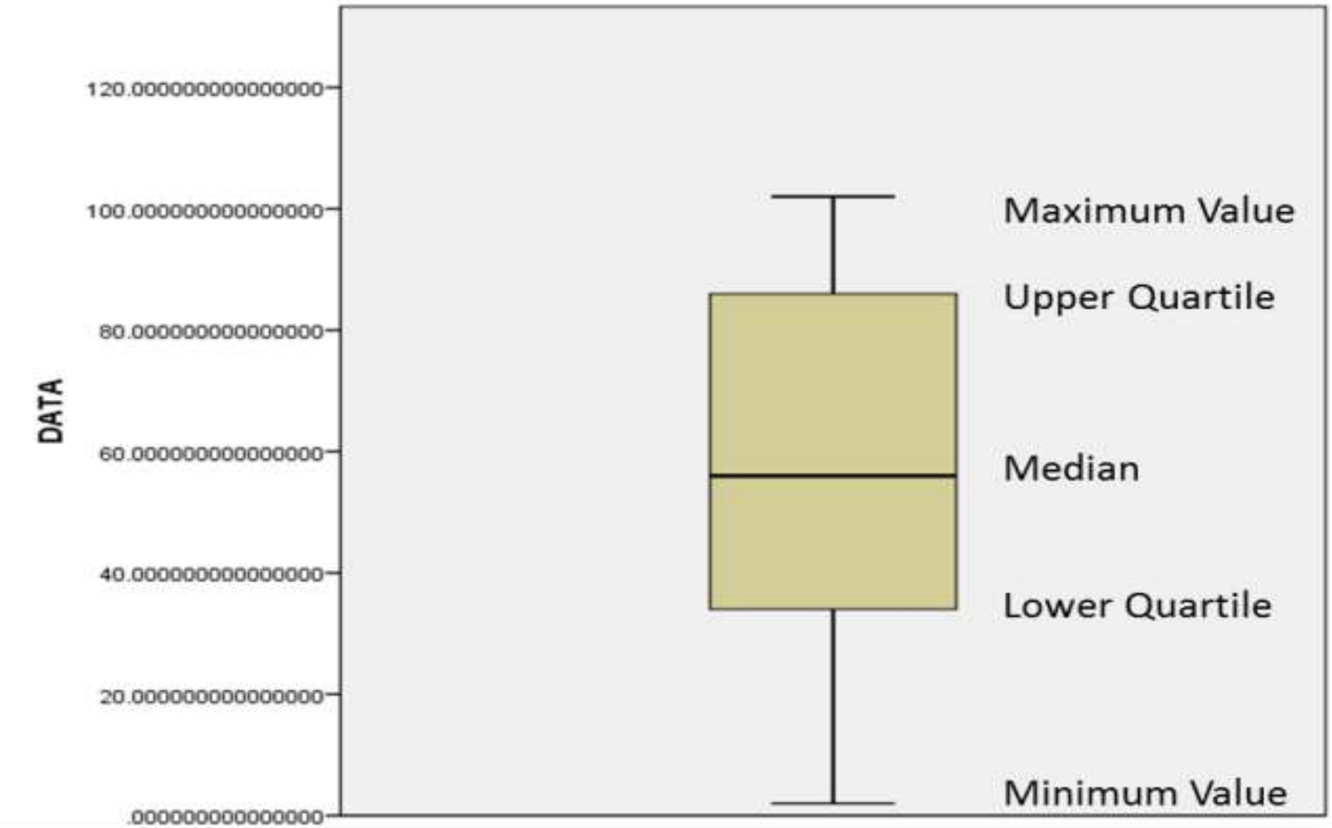


Box Plot (or Box and Whisker Plot)

- **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