Exploratory Data Analysis :-

- we deal with two types of statistics:-
  - → Descriptive statistics
  - → Inferential statistics

Descriptive statistics :-

Term given to analysis of the data that helps describe, show or summarize data in a meaningful way.

It is the discipline of quantitatively describing the main features of a collection of information.

Examples of Descriptive statistics :-

→ Frequency distribution which tells how frequencies are distributed over values.

→ Measures of central tendencies (mean, median, mode).

→ Graphs like pie chart, bar chart that describes the data.

Inferential statistics :-

→ Inferential statistics is concerned with making predictions or inferences about a population on the evidence of sample data.

→ To address the issue of generalization, we have tests of significance.

→ A chi-square or T-test, for example, can tell us the probability that the results of our analysis on the sample are representative of the population that the sample represents.

Examples of Inferential statistics :-

→ Linear regression analysis attempts to model the relationship between two variables by fitting a linear equation to the observed data.

→ Analysis of Variance (ANOVA) is a statistical technique that seperates the total variability found within a data set into two components.

→ The correlation analysis is a statistical technique employed to study the measure of relationship between two or more variables.

Population :- It is the set of all objects in a study.

Sample :- is any subset of the population, and a data point is an element of a set of data.

Types of Data Analysis :-

1) Descriptive Analysis
2) Diagnostic Analysis
3) Predictive Analysis
4) Prescriptive Analysis.

Exploratory Data Analysis comes under descriptive analysis either also called as descriptive statistics.

The lifecycle of analytics :-

1) Descriptive Analysis/Analytics →
2) Predictive Analytics → Automation.
(Optimization)

Data :- Information in raw or unorganized form (such as alphabets, numbers, or symbols) that refer to, or represent, conditions, ideas, or objects.

Types of Data :-

→ Cross sectional Data :- Collection of data at same point in time is called cross sectional data.

→ Time Series Data :-

Collection of data over a period of time such as month, quarter or year is called time Series Data.

→ Panel Data :-

Combination of cross sectional data and time series data is called panel data.

Variable :-

- Qualitative Variable
- Quantitative Variable

Qualitative Variable :-

Variable that cannot be counted (no numerical value).

Quantitative Variable :-

Variable that can be counted.

- Variable
  - Numeric
    - Continuous
    - Discrete
  - Categorical
    - Ordinal
    - Nominal

Continuous Variable :-

A variable with infinite number of values.

Example :- "time", "weight", "speed"

Discrete Variable :-

A variable that can be taken certain number of values.

Example :- Number of cars in a parking lot.

Categorical Variables :- Variables that can be put into categories.

Example :- The Category "Toothpaste" might contain Colgate and pepsodent.

Nominal Variable :- Another name for categorical variable.

Ordinal Variable :-
Similar to the categorical variable, but there is a clear order.
For example :- income levels of low, middle and high.

Ratio Variable :- Similar to interval variables, but has a meaningful zero.

Interval Variable :- A meaninful measurement between two variables. Also sometimes used as another name for a continuous variable.

Different file reading methods :-

- Set up your workdirectory in your local machine.

```
import os
os.chdir('C:\\Users\\Public\\Downloads')
```

- Data in 'csv' format.

```
df = pd.read_csv("data.csv")
df = pd.read_excel("data.xlsx", sheetname = "abc")
```

- import zipfile

```
archive = zipfile.Zipfile('T.zip', 'r')
df = archive.read('data.csv')
```

- text_file = open("text.txt", "r")

```
lines = text_file.read().
```

- df = pd.read_json("data.json")
- import xml.etree.ElementTree as ET

```
tree = ET.parse('data.xml')
root = tree.getroot()
print root.tag
```

- pdf2txt.py <pdf-file>.pdf

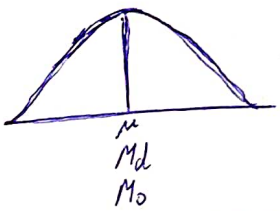
· Measure of Central Tendencies, Measure of Location, Statistical Average (mean, median and mode).

· Let us Consider:-

mean $= \mu$, median $= M_d$, mode $= M_o$.

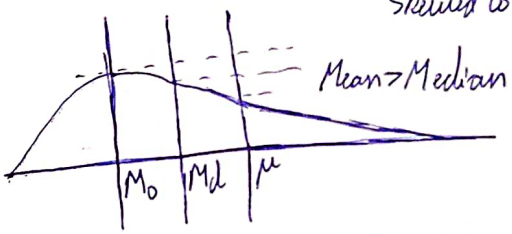
Symmetrical

Mean = Median = Mode



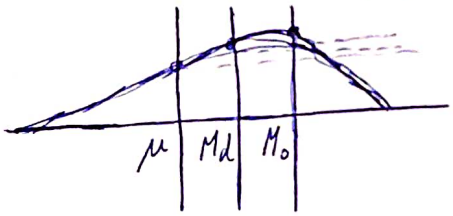
Skewed to right

Mean > Median > Mode



Skewed to left

mean < median < mode



Mean

Arithmetic mean is defined as the sum of all observations in a data set divided by the total number of observations.

$$\text{mean} = \frac{a_1 + a_2 + a_3 + a_4 + \cdots + a_{n-1} + a_n}{n}$$

$$\boxed{\text{mean} = \frac{1}{n} \sum_{i=0}^{n} a_i}$$

· Arithmetic mean is affected by extreme values or fluctuations in sampling.

· It is not the best average to use when the data set contains extreme values or outliers.

Median:- Median is the middle most observation when you arrange data in ascending order of magnitude.
· It is not effected by extreme values. $\boxed{\left(\frac{n+1}{2}\right)}$

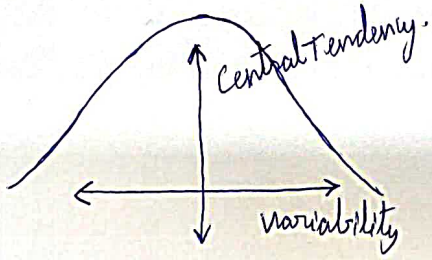
Mode:- Mode is the value which occurs most often. It has the maximum frequency of occurence. Mode also has resistance to outliers.
We can have n number of modes.

MEASURE OF DISPERSION

Central Tendencies describe the central point in a data set. Dispersion or variability describes the spread of the data.

Measure of Central tendency use a single value to describe the center of a data set. (Mean, median and mode).

Variability describes the spread of the data.



Measures of variability or dispersion includes:-
→ Standard deviation
→ Range
→ Kurtosis
→ Skewness

To describe the spread, a number of statistics are available to us, including the range, quartiles, standard deviation.

Range:-

Range is defined as the difference between the highest and lowest value of the variable in the series.

$$\boxed{X_{max} - X_{min}}$$

Quartile:-

Quartiles split the values into four equal parts:-

First Quartile:- divides the smallest 25% of the values from the rest that are larger.

Second Quartile:- (median) divides the 50% of the values from the rest that are larger or equal to median.

Third Quartile :- divides the smallest 75% of the values from the rest that are larger.

First Quartile :- $Q_1 = (n+1)/4$ (n is the sample size)

Third Quartile :- $Q3 = 3(n+1)/4$

Percentiles:-

- Percentiles split a variable into 100 equal parts
- First quartile is equivalent to 25th percentile
- Second quartile is equivalent to 50th percentile.
- Third quartile is equivalent to 75th percentile.

IQR (Inter Quartile Range)

Range computed on middle 50% of the observation after eliminating the highest and the lowest 25% of observations in a data set that is arranged in ascending order. IQR is less effected by outliers.

$IQR = Q3 - Q1$

Quartile Deviation :-

$$QD = (Q3 - Q1)/2$$

Co-efficient of QD $= (Q3 - Q1)/(Q3 + Q1)$

Standard Deviation :-

It is the square root of the variance.

Standard Deviation σ (Sigma) is widely used in business production, quality control techniques etc.

$$\boxed{\sigma = \sqrt{\Sigma (x_i - \bar{x})^2 / n}}$$

Coefficient of Variation :-

$\sigma \rightarrow SD$

$\bar{x} \rightarrow$ mean

$$\boxed{CV = (\sigma / \bar{x}) * 100}$$

Inorder to decide which of the two distributions is more variable, we compare the coefficient of variation.

- The distribution with greater CV is said to be more variable.

(Lesser the CV, more the consistency)

Z-Score :-

- Z score of a value is the difference between that value and the mean, divided by the standard deviation.
- Z score of 0 indicates the value is same as the mean.
- Z score is positive or negative means the value is above or below the mean and by how many standard deviations.

Five Number Summary:-

The five numbers that help describe the center, spread and shape of the data.

They are:-

1) Minimum ($X_{smallest}$)
2) First Quartile (Q_1)
3) Median (Q_2)
4) Third Quartile (Q_3)
5) Maximum ($X_{largest}$)

Shape of the Data:-

(a) Skewness

Skewness measures the extent to which the data values are not symmetrical around the mean.

Possibilities:-

- Mean < Median :- Negative (or) left skewed
- Mean = Median :- Zero skewness or Symmetry.
- Mean > Median :- Positive (or) right skewed

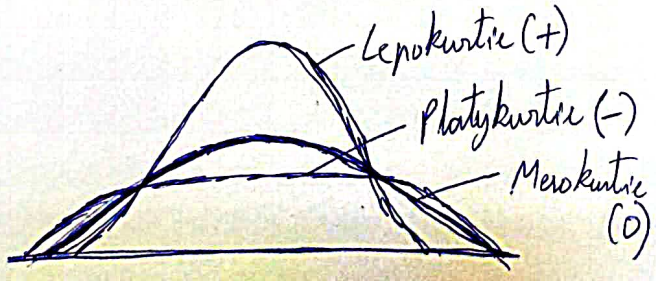
(b) Kurtosis

- Kurtosis = 0 : Mesokurtic
- Kurtosis > 0 : Lepokurtic

A distribution has sharper rising centre peak than the peak of the normal distribution.

- Kurtosis < 0 : Platykurtic

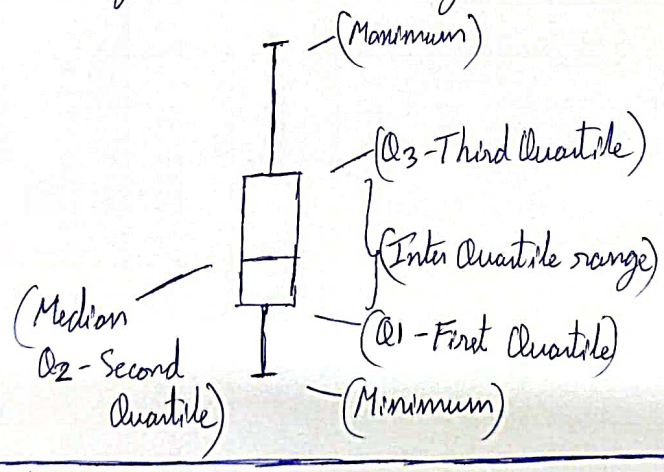
A distribution has slower rising centre peak than the peak of the normal distribution.



Relationships among the five-number summary and distribution shape.

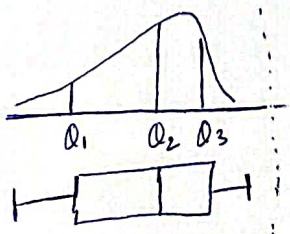
Left Skewed	Symmetric	Right Skewed
Median − $X_{smallest}$ > $X_{largest}$ − Median	Median − $X_{smallest}$ = $X_{largest}$ − Median	M_d − $X_{smallest}$ < $X_{largest}$ − M_d
Q_1 − $X_{smallest}$ > $X_{largest}$ − Q_3	Q_1 − $X_{smallest}$ = $X_{largest}$ − Q_3	Q_1 − $X_{smallest}$ < $X_{largest}$ − Q_3
Median − Q_1 > Q_3 − Median	Median − Q_1 = Q_3 − Median	Median − Q_1 < Q_3 − Median

Boxplot is a Graphical display of the data based on the five-number summary.

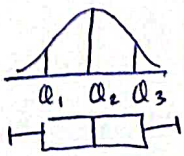


Distribution shape and the Box plot:-

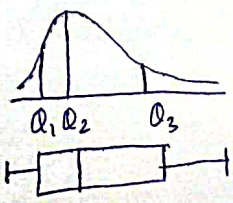
Left skewed



Symmetric



Right skewed



- A value is considered an outlier if it is more than 1.5 times the interquartile range below Q_1 or above Q_3.

Co-Variance :- (How two features of varying w.r.t one another)

- Two measures of the relationship between two numerical variables are covariance and coefficient of correlation.
- Covariance measures the strength of the linear relationships between two numerical variables (X & Y).

$$\boxed{Cov(X,Y) = \Sigma (X - X_{avg})(Y - Y_{avg}) / n-1}$$

where n is the sample size,

X_{avg} is the mean of X

Y_{avg} is the mean of Y.

- Covariance cannot tell whether the value indicates a strong relationship or weak relationship, since the covariance can take any value.

Co-efficient of Correlation :-

- Coefficient of correlation measures the relative strength of a linear relationship between two numerical variables.
- The values of correlation coefficient varies from -1 (perfect negative relationship) to 1 (perfect positive relationship).
- Correlation is a unit measure of change between two variables change with respect to each other.

Univariate Analysis :-

Refers to the analysis of a single variable. The main aim of univariate analysis is to summarize and find patterns in the data. There is only one variable involved in the analysis.

Multi Variate Analysis :-

It is performed to understand interactions between different fields in the data set (or) finding interactions between variable more than 2.

Example :- Pair plot and 3D Scatter plot.

Scatter Plot :-

A scatter plot gives us an idea of the association between two variables. A scatter plot is a visual representation of the degree of correlation between any two columns.

Handling non-numeric data :-

We need to convert categorical text data into numeric data for model building.

One-Hot Encoding :-

One hot encoding is used to create dummy variables to replace the categories in a categorical variable into features of each category and represent it using 1 or 0 based on the presence or absence of the categorical value in the record.

pd.get_dummies

This function does One-hot-Encoding on categorical text.

Convert to number :-

Sometimes if there are too many categories, it is better to go for label encoding rather than using dummy variables.

Sklearn Label Encoder:-

This function converts categorical text into categorical numbers.

Sklearn One Hot Encoder:-

This function does One-Hot-Encoding on categorical numbers.

Normalization and Scaling:-

In this method, we convert variables with different scales of measurements into a single scale.

Standard Scaler normalizes the data using the formula (x-mean) / standard deviation.

Univariate:-
Histogram

Multivariate:-
Heat Map

Bivariate Analysis:-

Numerical Vs Numerical

1) Scatter plot
2) Line plot

Categorical Vs Numerical

1) Bar chart (on summary statistics)
2) Grouped kernel density plots
3) Box plots
4) Violin Plots

Categorical Vs Categorical

1) Stacked bar chart
2) Grouped bar chart
3) Segmented bar chart

If standard deviation and variance value is more, then the data is spread and if the value is less, the data is dense (Concentrated at one point).

- Covariance is also known as Measure of Joint Variability.

This value will be from $-\infty$ to $+\infty$

we will get Magnitude + Direction

Magnitude tells us how large the variation is.

If value close to zero (0 ~) then features are independent.

Direction will tell whether - or +.

If one increases and other increases		then +	If direction is on one side then +.
Decreases	Decreases	+	
Increases	Decreases	-	If direction is opposite then -.
Decreases	Increases	-	

- when we want to know the measure of strength of relationship between two features, we go for correlation.

Correlation consists of Mag and Direction.

Its range is -1 to +1.

Uni-Variate :-

Numerical :-

1. Histogram
2. Distplot
3. KDE Plot
4. Boxplot
5. Violinplot

Categorical :-

1. Bar Graph
2. Pie Chart

Scaling :-

- Used for better interpretation
- To control contribution of features equally

(we need to bring the features in the same scale)

To do this scaling we have two methods,

Z Score [Standardization Method]

$$\boxed{Z = \frac{x-\mu}{\sigma}}$$ where μ is mean and σ is standard deviation.

Example :-

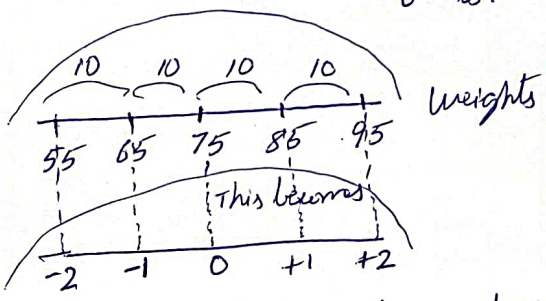
Let us assume, weight of a student $x = 75$ and mean $\mu = 75$, ~~the~~ σ is 10, then Z will be 0

$\boxed{Z = 0}$

If $\boxed{x = 85}$ then $\boxed{Z = 1}$

If $\boxed{x = 65}$ then $\boxed{Z = -1}$

Lets plot these values in a graph :- σ is 10.



Distribution is same, but scaling is different. So, finally, if exam score (for example) is 0, then average score, if +1 or +2 (then good score), if (-1, -2 ...) then low score.

- **Min-Max Scaling :- [Normalization].**

we make the maximum and minimum values of the data to 1 and 0.

$$\boxed{\begin{array}{l} \text{Max} \rightarrow 1 \\ \text{Min} \rightarrow 0 \end{array}}$$

For example :-

Max → 100 Kgs (Suppose the highest weight of
Min → 50 Kgs the guy is 100, $x = 100$)

Max Weight → 100

Min Weight → 50.

· Formula for Min-Max Scaling is,

$$\boxed{\frac{X - X_{min}}{X_{max} - X_{min}}}$$

If we consider, the weight of student as 100 then,

$$\frac{100-50}{100-50} = 1.$$

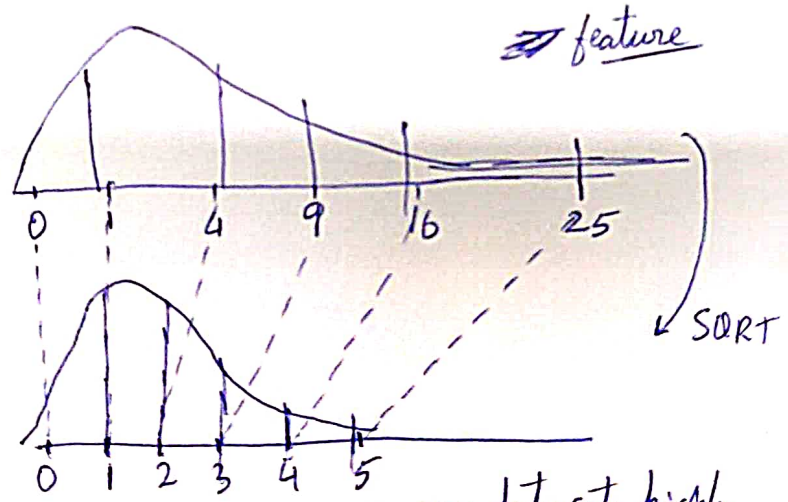
If we consider the weight of student as 50 then,

$$\frac{50-50}{100-50} = 0.$$

· So, the lowest value is becoming 0 and highest is becoming 1.

· Example, if the weight is 75, then, min-max is,

$$\frac{75-50}{100-50} = \frac{25}{50} = 0.5$$



our data gets highly squeezed.

Data times will come closer.

This application is known as TRANSFORMATION.

There are two types of Transformations:-

⇒ SQRT (Light Transformation)

⇒ LOG (Heavy Transformation)

Handling Categorical Features:-

we are converting male into 1 and female into 0.

Gender	
M	1
M	1
F	0
F	0
M	1
F	0
M	1
M	1

Cities	
Nagpur	2
Mumbai	1
Delhi	0
Delhi	0
Mumbai	1
Delhi	0
Nagpur	2
Nagpur	2

Basically this process is known as Label Encoding.

- Using Replace and Map Functions we can change the values.
- In machine learning library, from sklearn.preprocessing import Label Encoder.
- To fit the Label Encoder we can use ~~fit transf~~ fit_transform.

But, there is a limitation in Label Encoding. In the above example, Gender M and F can be given as True or False, and in Cities (where it is nominal feature, using Label Encoding, machine will detect it as ordinal.

ONE HOT ENCODING

~~Example~~:- Let us take Cities,

Cities	N	D	B
N	1	0	0
D	0	1	0
B	0	0	1
B	0	0	1
B	0	0	1
D	0	1	0
N	1	0	0

This is known as One Hot Encoding.

- we use pd.get_dummies for One Hot Encoding.

During ~~to~~ one-hot encoding, we can drop a column, because, we can guess the city easily even in the absence of one column by looking at the 0's and 1's.

· we use pd.get_dummies for One Hot Encoding.

Bi Variate Analysis (2 features)

Num - Num
1. Line Graph
2. Scatter Plot
3. Joint Plot
4. LM Plot

Num - Categorical
1. Bar Graph
2. Box plot
3. Violin plot
4. Strip plot
5. Swarm plot

Category - Category
1. Side-by-Side bar graph
2. Stacked bar graph

· How can we handle null values in a data.
→ Replacing the data
→ Removing the data

· Firstly we will perform boolean on null values.
· Then we can count the number of null values.
· we can either replace them or remove them.

Scenarios
· If column has 90% of null values, then we can drop the column.
· If we have 5-10% of null values in a column, we can drop the rows, but cannot remove the complete column.
· In a column we have 20% - 40% as null, here we cannot drop columns or rows as there is a huge loss. we can go for replace option.

How do we decide by what we want to Replace. (we dont use mode usually)
To replace we look at the data.
If the data is skewed, Mean?

we go with Median.

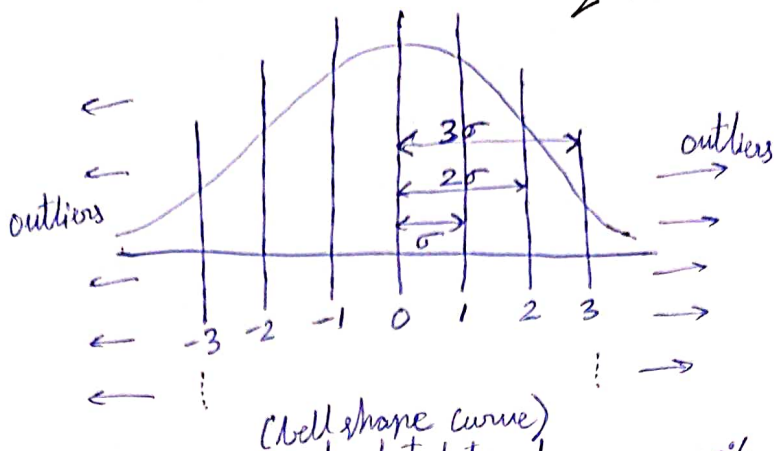
If the data is symmetric

we can go with Mean, but median is not harmful.

Handling the Out liers:-

Firstly let us see w.r.t Z-score.

Z-Score

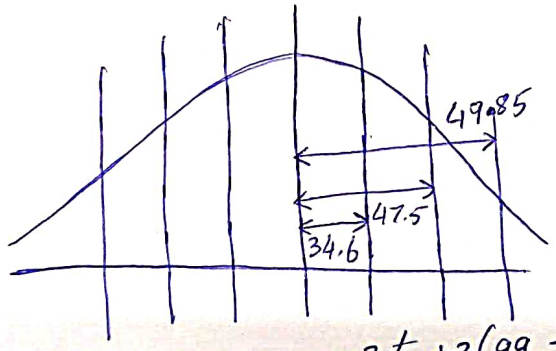


(bell shape curve)

· It seems in a normal distribution diagram, 100% of the data is in between -3 to +3 bandwidth.

(Gaussian or Standard Normal Dis. or Bell shape)

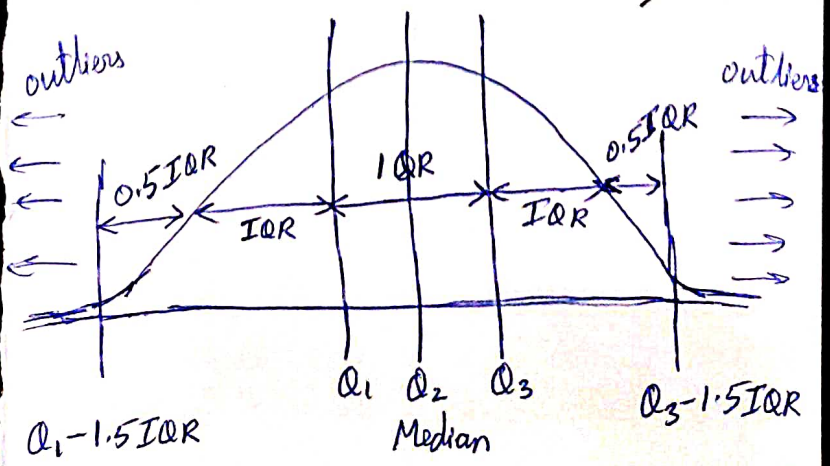
There is a fixed emperical rule:-



-3 to +3 (99.7%)

We can handle the out liers with IQR.

IQR



(Lower Limit)

(Upper Limit)

- value_counts(normalize = True) * 100 gives us the percentage
- print ('Mode of Total Bill column is', mode[0])

  gives us the mode of first column.

- Map and Replace functions can be used during label encoding to change the values.
- we have an option in the ML library for label encoding,

  from sklearn.preprocessing import Label Encoder
- To fit the Label Encoder we can use fit_transform.
- we use pd.get_dummies for one Hot Encoding.
- If it is seasonality / Trend we are going for a line plot.
- Hue is to divide the colors.

  (Example Males and Females)

  (Different Days) etc.

- To import zscore, we use the library scipy.stats.

  from scipy.stats import zscore.

- quantile is used to check Q_1, Q_2, Q_3.

EDA Points

- Example to load a data set,
  df1 = sns.load_dataset('tips')
- For numerical data we use,
  Histogram, KDE plot, Distplot, Box plot, Violin plot
- For categorical data we use,
  Bar graph, Pie chart
- df1.nunique() gives the number of unique values in each column.
- df1['day'].unique()
  gives the days people visited (list of days).
- df1['day'].value_counts()
  (gives the list of each day in the column 'day' with the number of customers visited).
- If we want the percentage of customers visited out of all the customers use df1['day'].value_counts(normalize=True)*100

- To find mean, var, std, we use this:-
  df1['total_bill'].mean(), .median(), .skew(),
  .mode, .var(), .std(), .mad(), .kurt()
- from scipy.stats import zscore
- from sklearn.preprocessing import LabelEncoder
  then use LE.fit_transform
- pd.get_dummies is used for one hot encoding.
- df.dropna() drops all the rows with null values. (axis=0 by default).
- If we want to change the data type:-
  df_adult['sex'] = df_adult['sex'].astype('object')

- <u>Example programs useful</u>:-
  df.groupby(['workclass'])[['age']].median()
- df_adult['hours'].quantile(0.75)

- from sklearn.preprocessing import StandardScaler
  sc = StandardScaler()
  scaled_bmi = sc.fit_transform(dataset[['bmi']])

- from sklearn.preprocessing import MinMaxScaler
  min_max = MinMaxScaler()
  transformed_age = min_max.fit_transform(dataset[['age']])

- Cross sectional data is the current data
- Time series data focus on multiple time intervals while panel data focuses on multiple individuals at multiple time intervals.

- Categorical variable is also called nominal variable
- An interval variable is where the difference b/w two values is meaningful. diff b/w 100°c, 90°c is same as diff b/w 90°c and 80°c.
- A ratio variable, has all the properties of an interval variable, but also has a clear definition of 0.0. when the variable equals 0.0, there is none of that variable. Variables like height, weight, enzyme activity are ratio variables.

- skewness is a measure of symmetry, or more precisely, the lack of symmetry.
- Kurtosis is a measure of whether the data are heavy-tailed (more outliers) (or) light-tailed (less outliers).

Examples for skewness
- wealth, length of comment (+ve skew)
- Age, weight, Height (normal skew)
- Life span of human being (-ve skew)

- Quantiles are just the lines that divide data into equally sized groups.
- Percentiles are just quantiles that divide the data into 100 equally sized groups.

- Univariate analysis (Numerical) Histogram
  KDE plot (or) density plot
  Dist plot (Histogram + Density)

- Univariate analysis (Categorical) → Bar graph
  Pie chart

- Bivariate
  numerical vs numerical
  Scatter plot
  Joint plot (histogram + scatter plot)
  lm plot (line of best fit on the scatter plot)

  numerical vs categorical :- Bar graph
  Box plot
  Violin plot
  strip plot
  swarm plot

  Categorical vs Categorical :-
  Side by side Bar graph
  stacked Bar graph

Pair plot :- pairwise relationship b/w the numeric variables.

- A variable has magnitude and unit.
  Example :- 25 years.
  magnitude :- 25, units :- years

- Normalization helps us to scale down our feature b/w 0 to 1.
- standardization helps us to scale down our feature based on standard normal distribution.
- Normalization → min max scalar
- Standardization → standardscalar

Transformations helps us to transform the data in a better way.

- log transformation is used to remove the skewness of the original data.
- A square root transformation can be used to normalize a skewed distribution.

- abs() function is used to return the absolute value of a number.
Positive num is unaffected and -ve num is converted into positive.

- df.select_dtypes(include = np.number) means we are including all the columns which have numbers.

- Quartile Deviation:-

$QD = (Q_3 - Q_1)/2$

- Coefficient of Quartile Deviation:-

$(Q_3 - Q_1)/(Q_3 + Q_1)$

- Coefficient of Variation:-

$$\boxed{CV = (\sigma/\bar{x}) * 100}$$

- Covariance

$$\boxed{cov(x,y) = \frac{\Sigma(x - x_{avg})(y - y_{avg})}{n-1}}$$

- we need to find like, how far apart the ~~average~~ individual data points are from the average. (or) how spread out, the data points are w.r.t average.

- Mean Absolute deviation is the average of the sum of Abs (Value − Average).

For example :-

Avg is 70.

$\frac{1}{n} \Sigma (x_i - \bar{x})$

Name	Score	Abs(score − Avg)
a	75	5
b	72	2
c	68	2
d	65	5
e	67	3
f	73	3
	MAD	3.16

- Till here we are done with the mean absolute deviation.

- Now lets add one more column and take the square of abs(score − Avg). (variance)

$(\text{Score} - \text{Avg})^2$ $(x_i - \bar{x})^2$
25
4
4
25
9
9

$\Sigma (x_i - \bar{x})^2 = 76$

$\frac{1}{n} \Sigma (x_i - \bar{x})^2 = 12.66$

Average $= 12.66$

$\sqrt{\text{average}} = \sqrt{12.66}$

$= 3.55$

Therefore, standard deviation σ is

$$\sqrt{\frac{\Sigma (x_i - \mu)^2}{N}}$$

Coefficient of Variation (CV) :-

$\frac{\text{standard deviation}}{\text{mean}}$ also known as relative standard deviation.

Standard deviation is the most common measure of variability for a single data set.

Sometimes the standard deviation cannot be used for 2 data sets.
That is when we use, Coefficient of variation.

Example for coefficient of variation :-

First data →
13, 17, 1, 13, 13, 20, 18, 15, 10, 3

Second data → (~~double the of~~ multiply by 10)
130, 170, 10, 130, 130, 200, 180, 150, 100, 30

	First data	Second data
mean	12.3	123
variance	38.01	3801
std	6.16	61.65

· Even though the same points of the first data were used to create second data, we can find exactly the variation of the data points from the mean.
First data has 6.16 and second data has 61.65.
To correct this, we use Coefficient of variation :-
Group1 → CoV is 0.501245
Group2 → CoV is 0.501245

Covariance
· Measures the ~~strength~~ variation between x and y.

$$\boxed{Cov(x,y) = \frac{\Sigma (x - x_{avg})(y - y_{avg})}{n-1}}$$

Covariance cannot tell whether the value indicates a strong relationship or weak relationship, since the covariance can take any value.
Value is from $-\infty$ to $+\infty$
We get magnitude and direction.
Magnitude tells us how large the variation is.

· Covariance captures the variance b/w x and y.

Pearson Correlation Coefficient :-

· In covariance,
$x\uparrow\ y\uparrow \rightarrow$ (+ve) ⇒ we cannot determine how much +ve (or) how much (-ve).
$x\uparrow\ y\downarrow \rightarrow$ (-ve) ⇒ Direction

That is when we go for correlation.
Coefficient of correlation measures the relative strength of the linear relationship between two numerical variables.

$$\boxed{\text{Formula :- } \frac{cov(x,y)}{\sigma_x \sigma_y}} \quad \rho(x,y)$$

· In covariance, we know the direction but we cannot determine the strength of the variance.
· We can determine the strength using correlation.
We can tell the direction of the relationship using correlation.
· value ranges from -1 to +1.