Statistics for Data Science

Statistics

- It is science of getting information from the data.
- The result of statistics is always approximate and not very accurate.
- If we are working on any particular use case and when we have entire data available, we call it as population (Universal Set)
- If are working on any particular use case and when we have partial data available , we call it as sample. (Subset)

Population and Sample

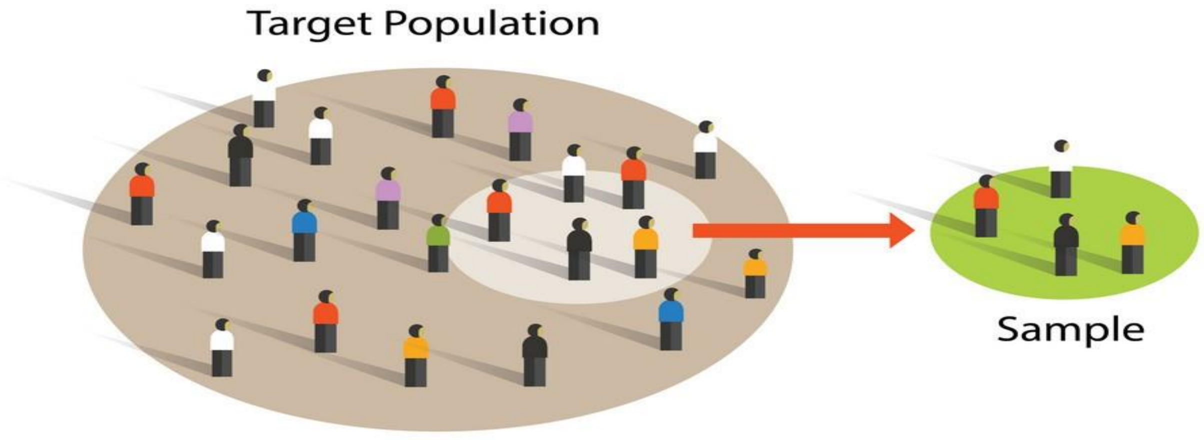
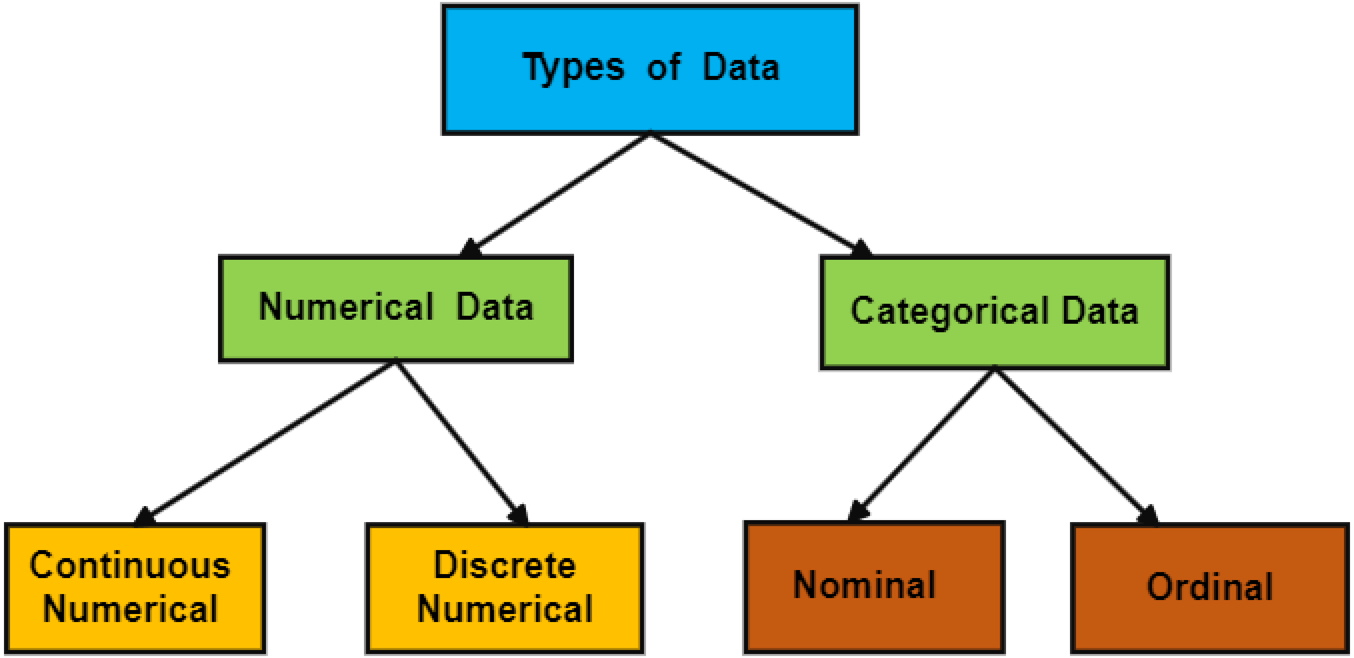


Image Source : VectorStock.com

Types of Data



Types of Data (Numerical)

Numerical Data is further classified as Continuous Numerical and Discrete Numerical

Continuous Numerical

It changes wrt to time. It has infinite values.

For Ex. Age, Height, Weight etc.

Time

Types of Data (Numerical)

Discrete Numerical

Entity that doesn’t change wrt to time.

It has finite range.

For Ex. Grades, No of Students etc.

No of Strawberries

Grades

Types of Data (Categorical Data)

Categorical data is further classified into Nominal and Ordinal. It represents different categories such as sports car brands.

Questions whose answer is in the form of “Yes” or “No” is another instance of categorical data.

Ferrari

Bugati

Are you enrolling for data science program ?

Sports Car brand

YES or NO

Ford GT

Lamborghini

Types of Data (Categorical Data)

Nominal Data

It is a type of categorical data and it doesn't follow any specific order.

For Ex. Gender (Male/Female), City, Season (winter/spring/summer)

Types of Data (Categorical Data)

Ordinal Data

It is a type of categorical data and it strictly follows order.

For Ex. Ranking, Grades

Movie Rating

*

**

***

****

*****

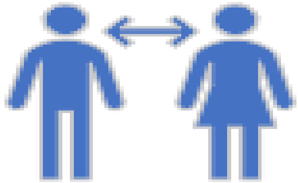
Check Your Understanding

Temperature

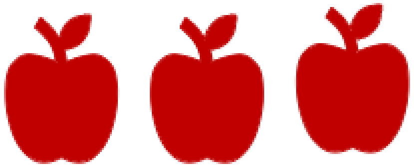
Gender

Speed of Vehicle

No of Stars



Social Distancing



No of Apples

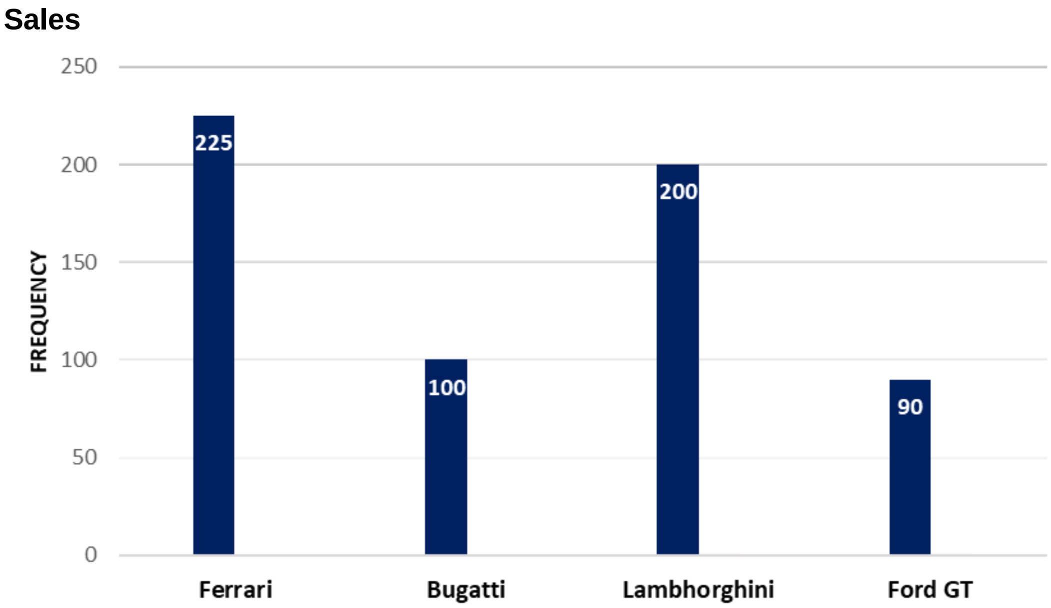
Visualization Techniques for Categorical data

- Frequency Distribution Tables
- Bar Charts
- Pie Charts
- Pareto Diagrams

Visualization Techniques for Categorical data

Sports Car Brand	Frequency
Ferrari	225
Lamborghini	200
Bugatti	100
Ford GT	90
Total	**615**

Frequency Distribution Table



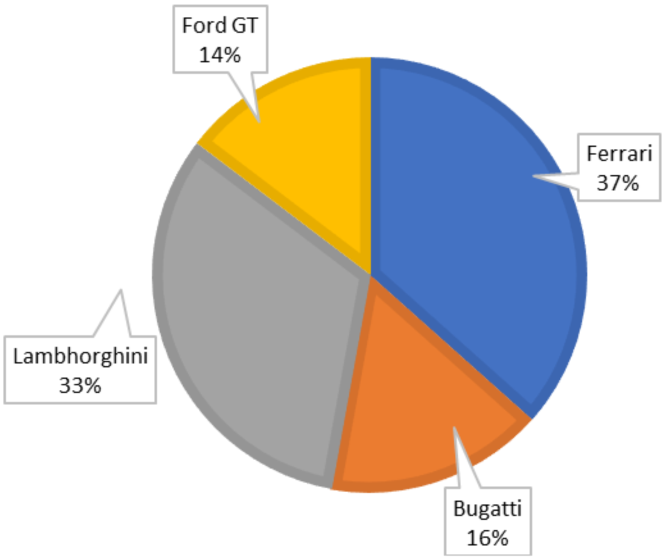
Bar Chart

Visualization Techniques for Categorical data

Sports Car Brand	Frequency	Relative Frequency
Ferrari	225	37 %
Bugatti	100	16 %
Lamborghini	200	33 %
Ford GT	90	14 %
Total	615	100 %

Frequency distribution table

Market Share

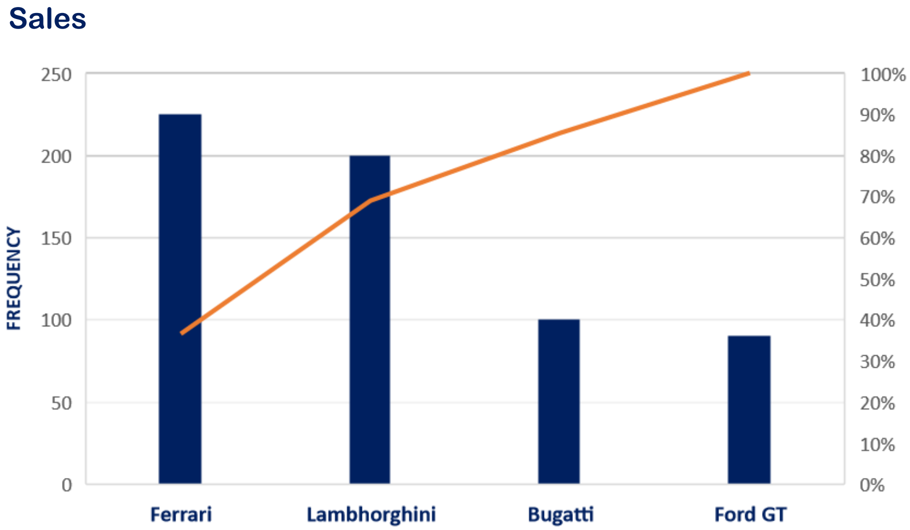


Pie Chart

Visualization Techniques for Categorical data

Sports Car Brand	Frequency	Relative Frequency
Ferrari	225	37 %
Bugatti	100	16 %
Lamborghini	200	33 %
Ford GT	90	14 %
Total	615	100 %

Sports Car Brand	Frequency	Relative Frequency	Cummulative Frequency
Ferrari	225	37 %	37 %
Lamborghini	200	33 %	70 %
Bugatti	100	16 %	86 %
Ford GT	90	14 %	100 %

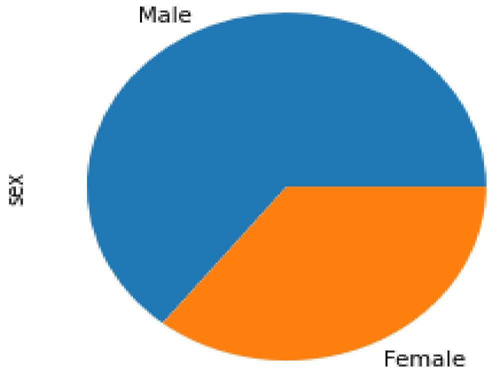


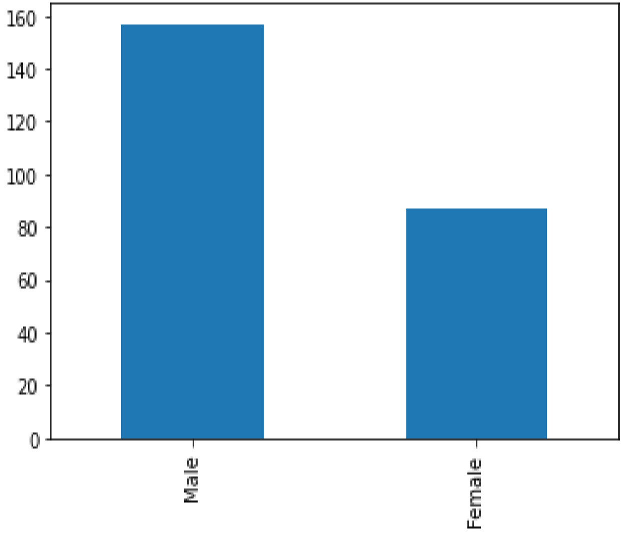
Pareto Diagram

Visualization Techniques for Categorical data

Tips Dataset

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4

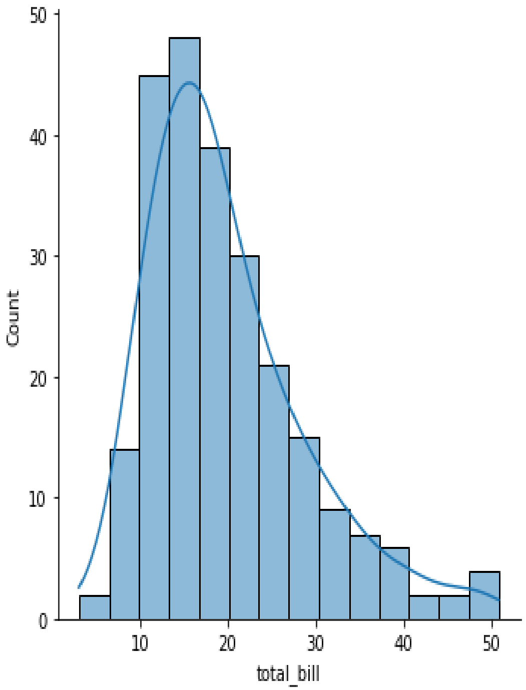


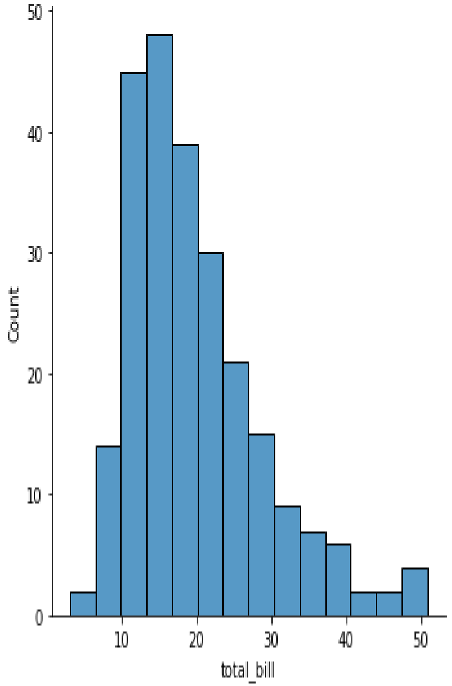


Visualization Techniques for Numerical data

Tips Dataset

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4

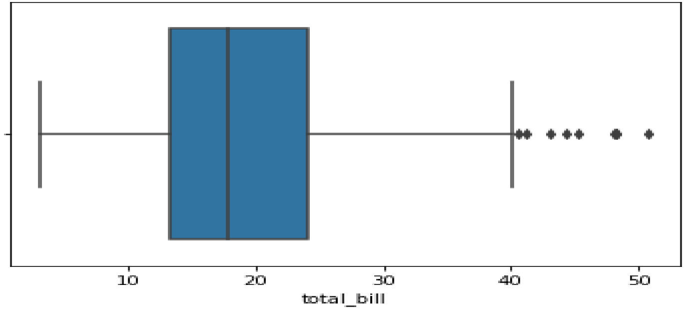


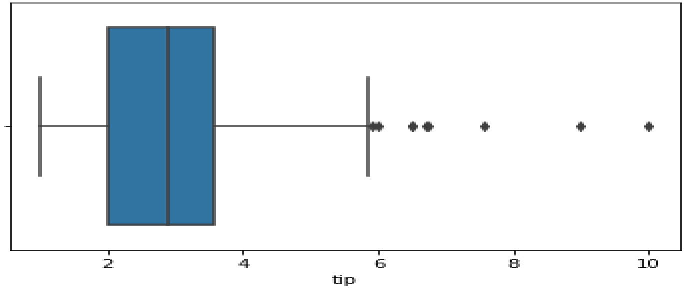


Visualization Techniques for Numerical data

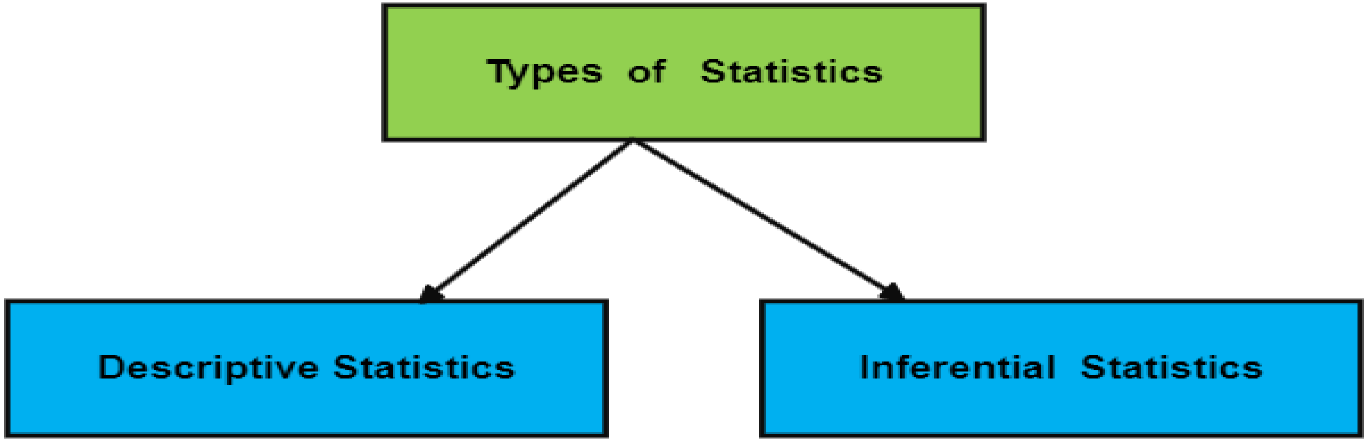
Tips Dataset

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4





Types of Statistics



- It is used to organize and summarize the data.
- It is used to check the quality of data.

- Techniques to draw out inferences about population with sample data.
- To create the predictive models
- To check the quality of model

Descriptive Statistics

Descriptive Statistics

Measures of Central Tendency (Mean, Median, Mode)

Mean :- It is calculated by performing addition of all data points and then dividing by number of data points in the dataset.
For Ex. 1, 5, 4 . The mean is 3.3

Median : It is calculated by performing ordering from lowest to highest and finding the middle no. (In case of two middle numbers, we take the mean of two middle numbers.

Example1. 1, 5, 4
The ordering of number is 1, 4, 5 . The median of 1, 5, 4 is 4

Example2. 1, 5, 5, 4
1, 4, 5, 5
The median of the 1, 5, 5, 4 is 4.5

Mode : It is calculated by finding the number which is occurring most frequently.
For Ex. { 2, 4, 2, 3, 3, 2 } The mode for { 2, 4, 2, 3, 3, 2 } is 2

Measures of Central Tendency (Mean, Median, Mode)

Measures of Central Tendency (Mean, Mode, Median) finds the mid value of the data which will help to understand the quality of the data.

USB Drive Prices

Sr. No	Florida	Washington
1	$10	$10
2	$20	$20
3	$40	$30
4	$40	$40
5	$50	$50
6	$60	$60
7	$70	$70
8	$80	$80
9	$1000	

Mean, Median and Mode
USB Drive Prices

	Florida	Washington
Mean	$152	$45
Median	$50	$45
Mode	$40	

Descriptive Statistics

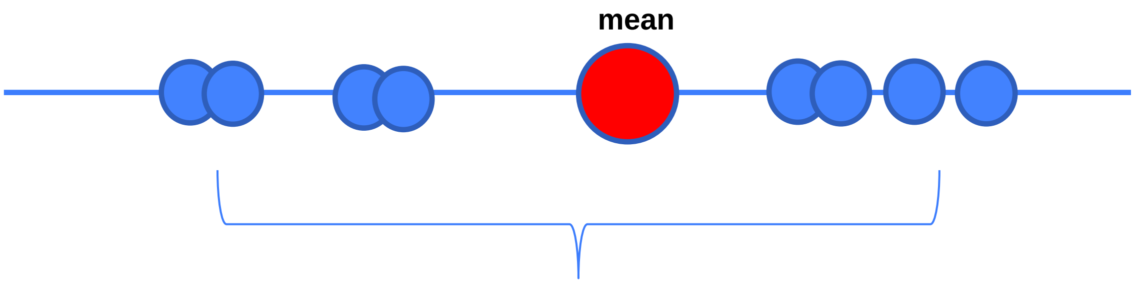
Measures of Dispersion

Variance

It measures the spread of data points around the mean.

$$S^2 = \sum_{1}^{n} (Xi - \bar{X})^2 \div n - 1$$

Sample Variance



Dispersion of data points around the mean.

$$\sigma^2 = \sum_{i=1}^{N} (x_i - \mu)^2 / N$$

Population Variance

Descriptive Statistics

Measures of Dispersion

Xi
10
20
30
40
50
60

Mean : 35
Population Variance : 291.66

$$\sigma^2 = \sum_{i=1}^{N} (x_i - \mu)^2 / N$$

$$= (10-35)^2 + (20-35)^2 + (30-35)^2 + (40-35)^2 + (50-35)^2 + (60-35)^2 / 6$$

= 291.66

Descriptive Statistics

Measures of Dispersion

Xi
10
20
30
40
50
60

Mean : 35
Sample Variance : 350

$$S^2 = \sum_{1}^{n} (Xi - \bar{X})^2 \div n - 1$$

$$= (10-35)^2 + (20-35)^2 + (30-35)^2 + (40-35)^2 + (50-35)^2 + (60-35)^2/5$$

= 350

Measures of Dispersion

Standard Deviation

It also measures the spread of data points around the mean

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

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

Sample standard deviation formula

Population standard deviation formula

Measures of Dispersion

Co-efficient of Variation

It is measure of relative variability. It is used to compare two different datasets.

$$CV = s/\bar{x}$$

$$CV = \sigma/\mu$$

Sample Variance

Population Variance

Standard Deviation and Co-efficient of Variation

USB Drive Prices

Dollars	Rupees
$10	Rs. 810
$20	Rs. 1620
$30	Rs. 2430
$40	Rs. 3240
$50	Rs. 4050
$60	Rs. 4860
$70	Rs. 5670
$80	Rs. 6480
$90	Rs. 7290

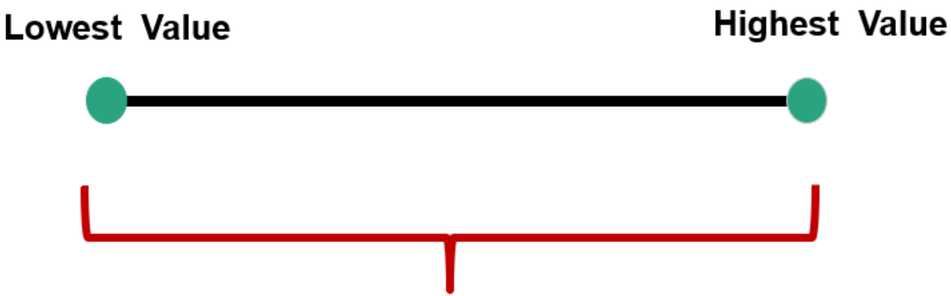
Mean	$50	Rs. 4050
Variance of Sample	$\$^2$ 750	Rs.^{2} 4920750
Standard Deviation of Sample	$ 27.386128	Rs. 2218.2764
Co-efficient of Variance of Sample	0.54	0.54

Variance gives results in squared units. So std deviation is most commonly used

Standard Deviation is the most common measure of variability for a single dataset

Range

Range is difference between maximum data point and minimum data point in the dataset.



Range

Given the numbers as : 4, 7, 8, 10, 15, 17, 20, 25, 35 , 50

Range = highest value - lowest value

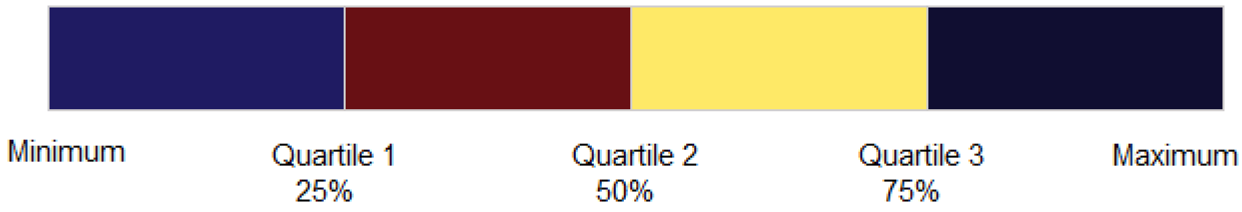
Range = 50 – 4

Range = 46

The drawback with range is it does not compute the spread of most of the data points in the dataset. Range only considers the spread between maximum and minimum data points.

Inter-Quartile Range (IQR)

When the entire dataset which is ordered , splited into four parts known as a quartile. When the dataset is splited into 100 parts then it is known as percentile.



Inter-Quartile Range (IQR)

How to perform outlier detection with IQR ?

- Sort the data in ascending order.
- Compute the value for Q1 and Q3
- Calculate IQR as IQR = Q3 – Q1
- Calculate lower range (LR) = Q1 – 1.5 * IQR
- Calculate upper range = Q3 + 1.5 * IQR

As shown in figure the data points which are falling under LR and greater than UR are the outliers.

Inter-Quartile Range (IQR)

Suppose we have the following dataset as

3, 4, 6, 5, 5, 10, 11, 4, 7, 8,12. (Odd no's) Find the Interquartile range.

Sorted Order

3, 4, 4, 5, 5, 6, 7, 8, 10, 11, 12

Q1 Q2 Q3

Median

IQR = Q3 – Q1

IQR = 10 – 4

IQR = 6

Inter-Quartile Range (IQR)

Suppose we have the following dataset as

4, 4, 6, 10, 10, 10, 11, 12, 16, 18. (Even no's) Find the Interquartile range.

Sorted Order

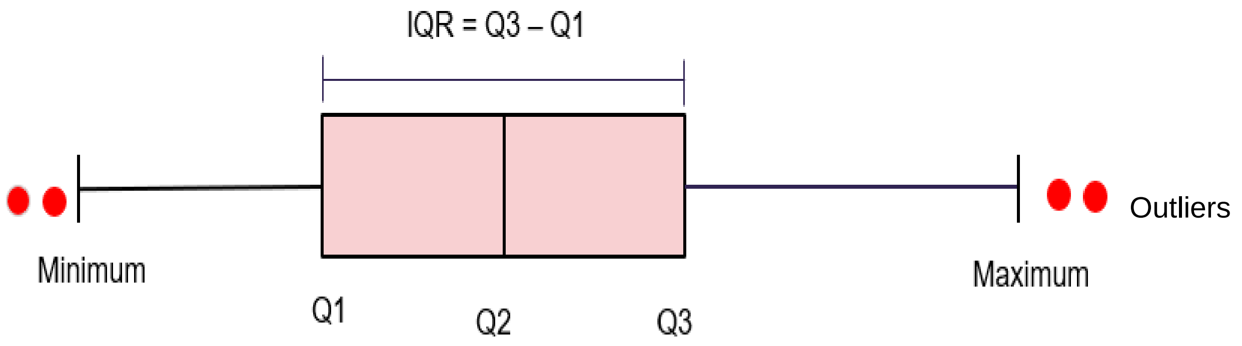
10

4, 4, 6, 10, 10, 10, 11, 12, 16, 18

Q1 Q2 Q3

IQR = Q3 - Q1
IQR = 12 – 6
IQR = 6

Box Plot



Five Number Summary

- Minimum
- Q1
- Q2
- Q3
- Maximum

Box Plot

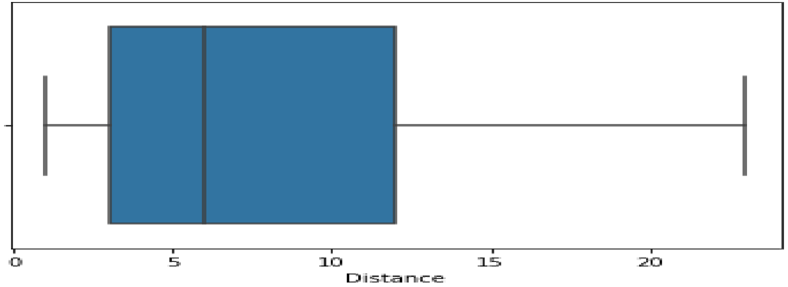
Given dataset as 12, 5, 2, 2, 3, 4, 16, 10, 9, 1, 6, 3, 4, 11, 18, 20, 23

Sorted Order 1, 2, 2, 3, 3, 4, 4, 5, 6, 9, 10, 11, 12, 16, 18, 20, 23

Q1 Q2 Q3

1, 2, 2, 3, 4, 4, 5, 6, 9, 10, 11, 14, 18, 20, 23

Q1 Q2 Q3



Five Number Summary

Minimum : 1
Q1 : 3
Q2 : 6
Q3 : 14
Maximum : 23

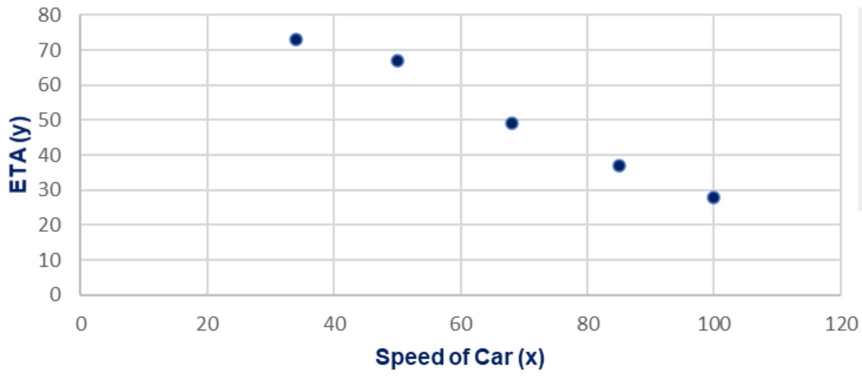
Co-variance

It is used to find the relationship between two variables. It is measure of variability between two variables

Co-variance

Covariance

	Speed of Car (km/h)	ETA (mins.)		$(x-\bar{x})*(y-\bar{y})$
	100	28		- 743
	85	37		- 243
	68	49		- 1
	50	67		- 282
	34	73		- 741
Mean	67	51	**Sum**	- 2,011
Std dev	26	19	**Sample size**	5
			Cov. Sample	- 503



$$COV(x,y) = \sum(x_i - \bar{x}) * (y_i - \bar{y})/n - 1$$

Problems with Co-variance

Covariance could be any number like 4 or 40. It can be number like 0.0022454 or 10 million,

- 4,
- 40,
- 0.0022454

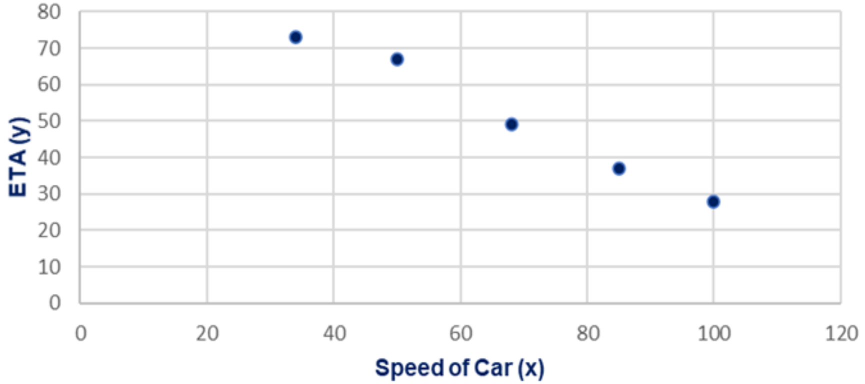
These values are of different scale and interpreting those numbers are difficult.

Correlation Coefficient

It is used to measure strength of relationship between the two variables. It is also measure of variability between two variables.

Covariance

	Speed of Car (km/h)	ETA (mins.)		$(x-\bar{x})*(y-\bar{y})$
	100	28		- 743
	85	37		- 243
	68	49		- 1
	50	67		- 282
	34	73		- 741
Mean	67	51	Sum	- 2,011
Std dev	26.42	19.16	Sample size	5
			Cov. Sample	- 503
			Correlation	(0.99)



Correlation Co-efficient =

$$COV(x,y)/Stdev(x)*Stdev(y)$$

Exploratory Data Analysis

Univariate Analysis

Categorical Data

- Tables
- Bar Chart
- Pie Chart
- Pareto diagram

Quantitative Data

- Histograms
- Numerical Summaries
- Box Plots

Bivariate Analysis

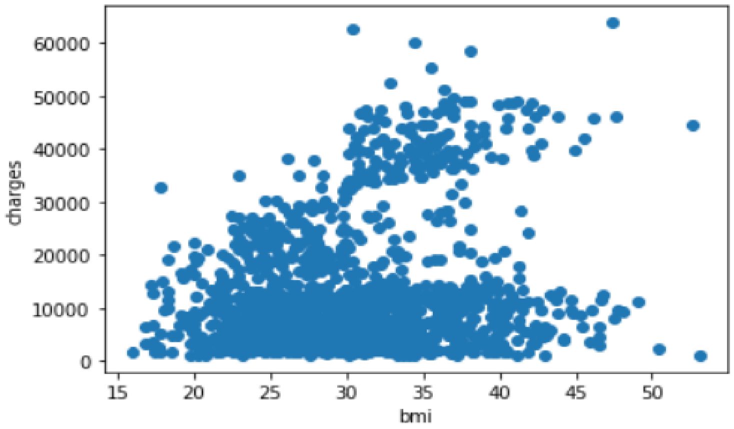
Variables	Type of Plot
C -> Q	Box Plot
Q -> C	Box Plot
C -> C	Contigency Table
Q -> Q	Scatter Plot

C - Categorical Data
Q - Quantitative Data

Bivariate Analysis

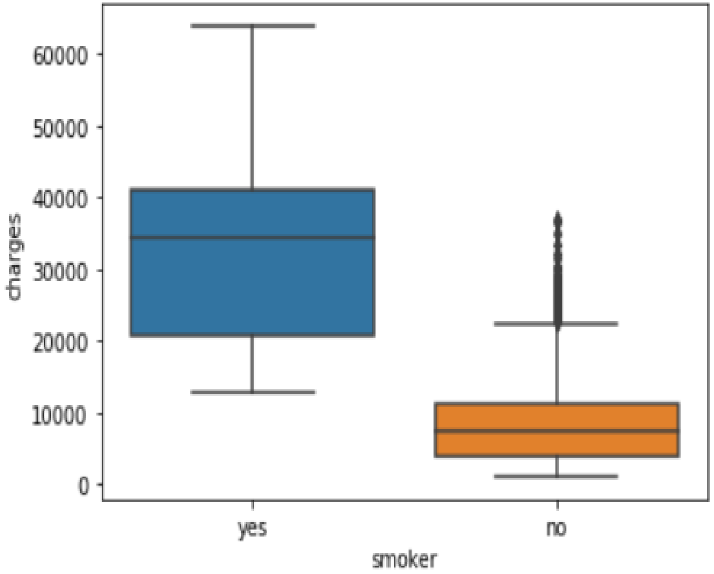
	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

```
plt.scatter(x = "bmi", y = "charges", data=A)
plt.xlabel("bmi")
plt.ylabel("charges")
plt.show()
```




```
sns.boxplot(x = 'smoker', y = 'charges', data = A)
```

<matplotlib.axes._subplots.AxesSubplot at 0x1fdf847cdc8>



Bivariate Analysis

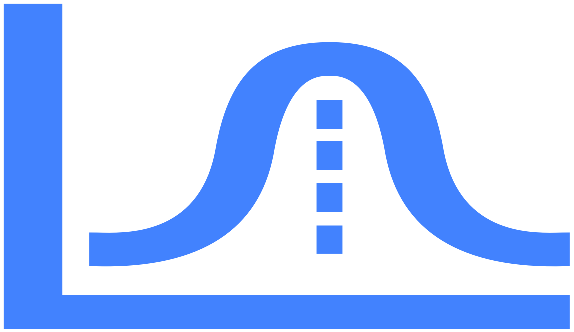
satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	quit	promotion_last_5years	department	salary
0.38	0.53	2	157	3	0	1	0	sales	low
0.80	0.86	5	262	6	0	1	0	sales	medium
0.11	0.88	7	272	4	0	1	0	sales	medium
0.72	0.87	5	223	5	0	1	0	sales	low
0.37	0.52	2	159	3	0	1	0	sales	low

```
pd.crosstab(A.department,A.salary)
```

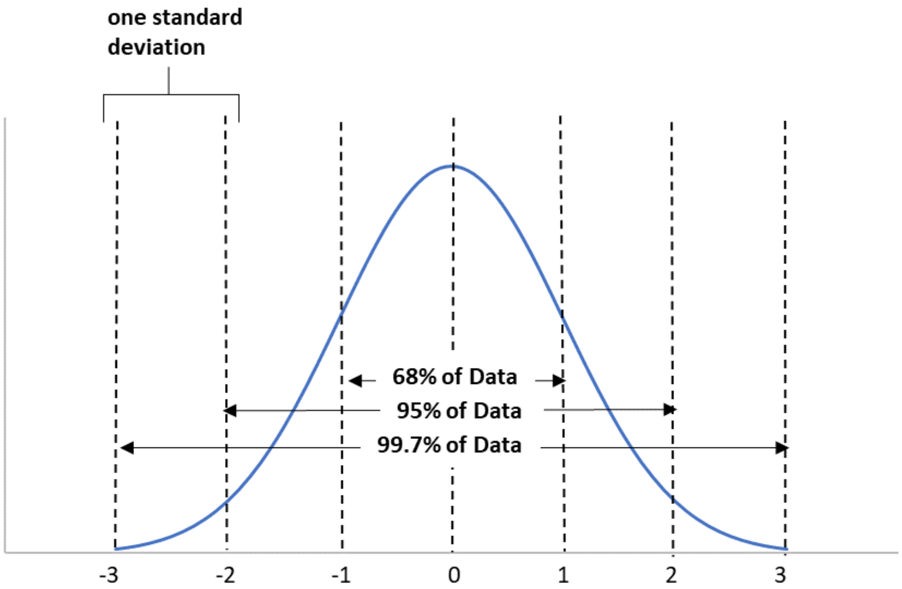
salary	high	low	medium
department			
IT	83	609	535
RandD	51	364	372
accounting	74	358	335
hr	45	335	359
management	225	180	225
marketing	80	402	376
product_mng	68	451	383
sales	269	2099	1772
support	141	1146	942
technical	201	1372	1147

Contingency Table

Inferential Statistics



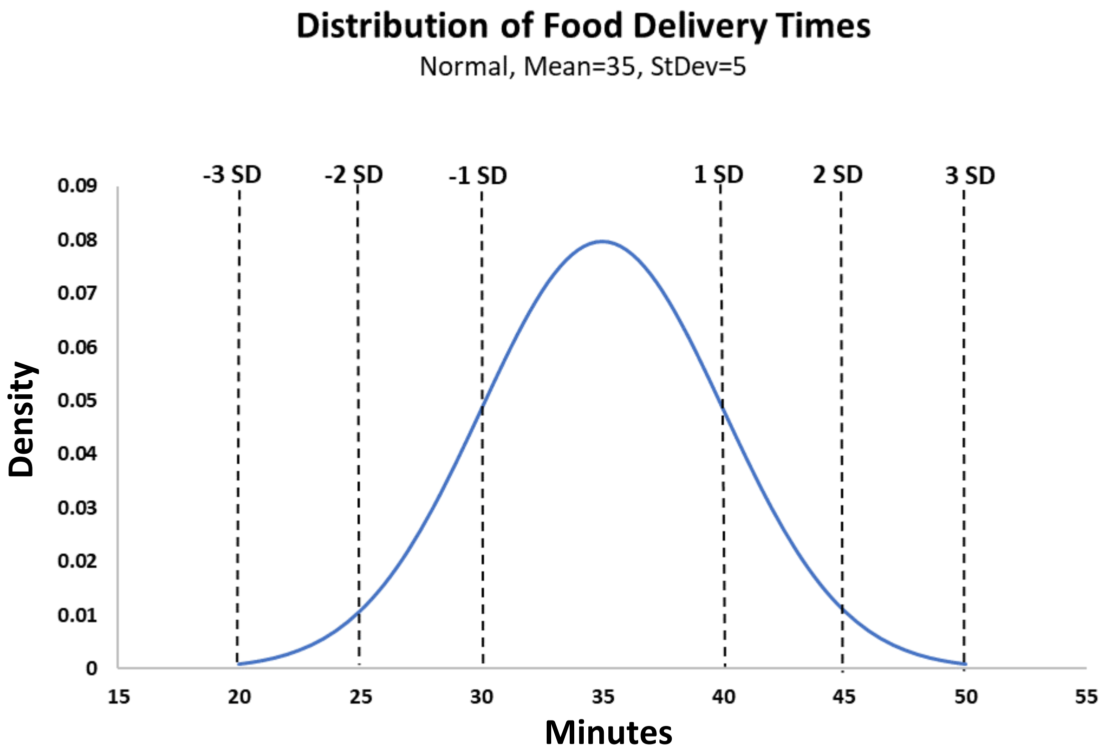
Normal Distribution



Normal Distribution is also referred as bell curve, gaussian distribution.

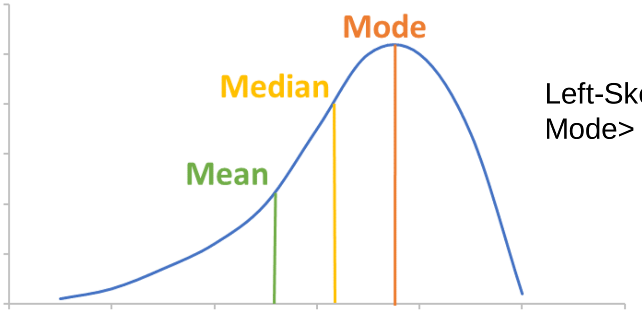
It is symmetrical distribution, where the left hand side is exact mirror image of right hand side.

Normal Distribution



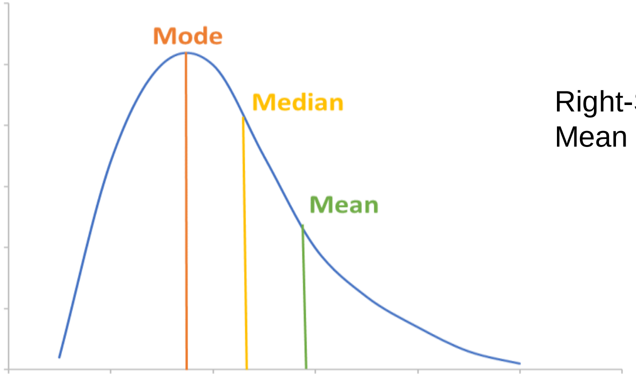
Skewed Distribution

Left-skewed distribution has long tail on the left hand side. It is also referred as negative skewed distribution. The mean here is located on the left of the peak.



Left-Skewed
Mode> Median > Mean

Right-skewed distribution has long tail on the right hand side. It is also referred as positive skewed distribution. The mean here is located on the right of the peak.



Right-Skewed
Mean > Median > Mode

Thank You !!!!!!