Exploratory Data Analysis:

Fundamentals of Statistics

Module 1: Exploratory Data Analysis

Data Management

- Import, sort, merge, aggregate, subset, derive
- Handling of missing data

Descriptive Statistics

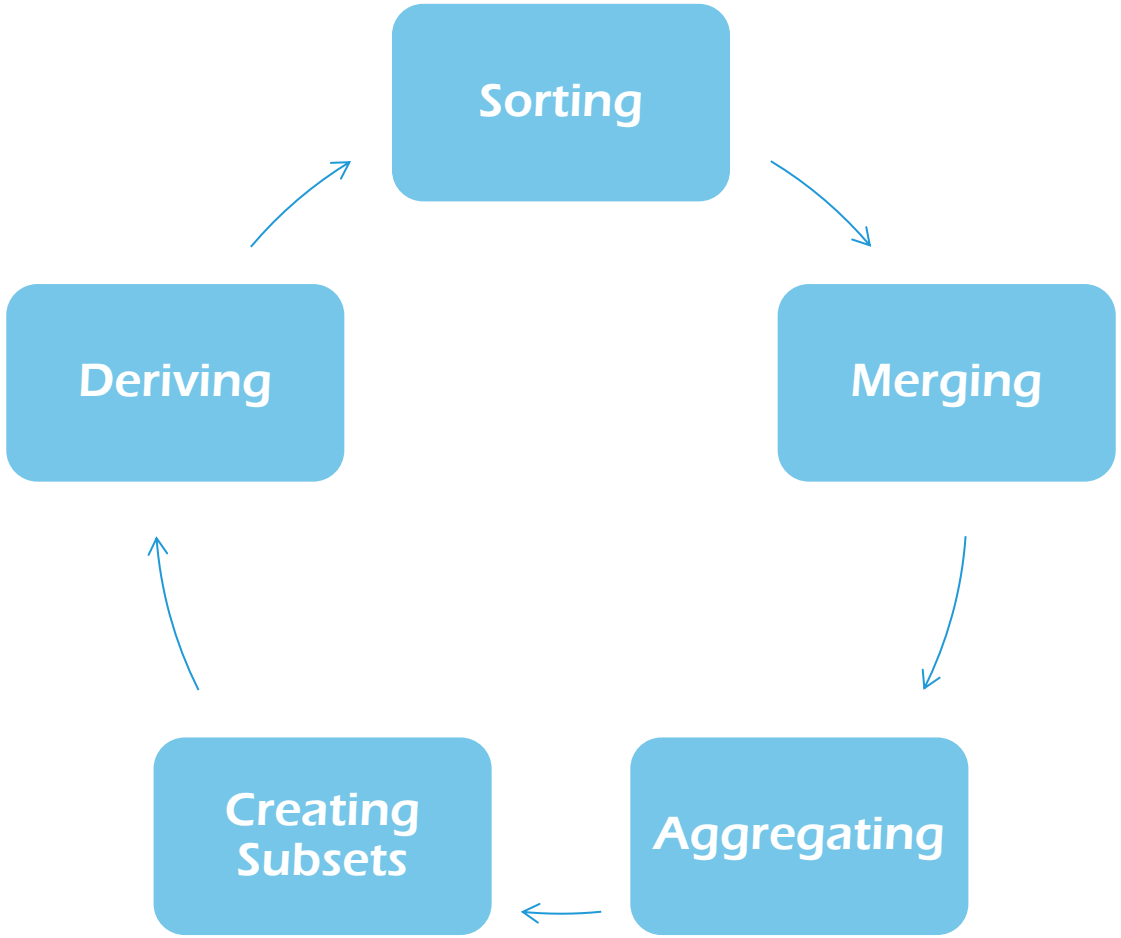
- Central tendency
- Variation
- Shape

Visualization

- Bar charts/histograms
- Box-Whiskers plot
- Contour plot
- Motion chart

- ✓ Critical for successful analytics implementation
- ✓ Good data management helps to
  - assess quality of data
  - improve the quality of data
  - make data analysis ready
- ✓ Provides data insights
- ✓ Guides towards business research problem solution using advanced analytics

Data Management Tasks



Sources of Data

1. **Primary data**

- Data collected by the investigator himself/herself for a specific purpose.
- Direct method of data collection.
- Eg. Data collected for research through questionnaires, interviews.

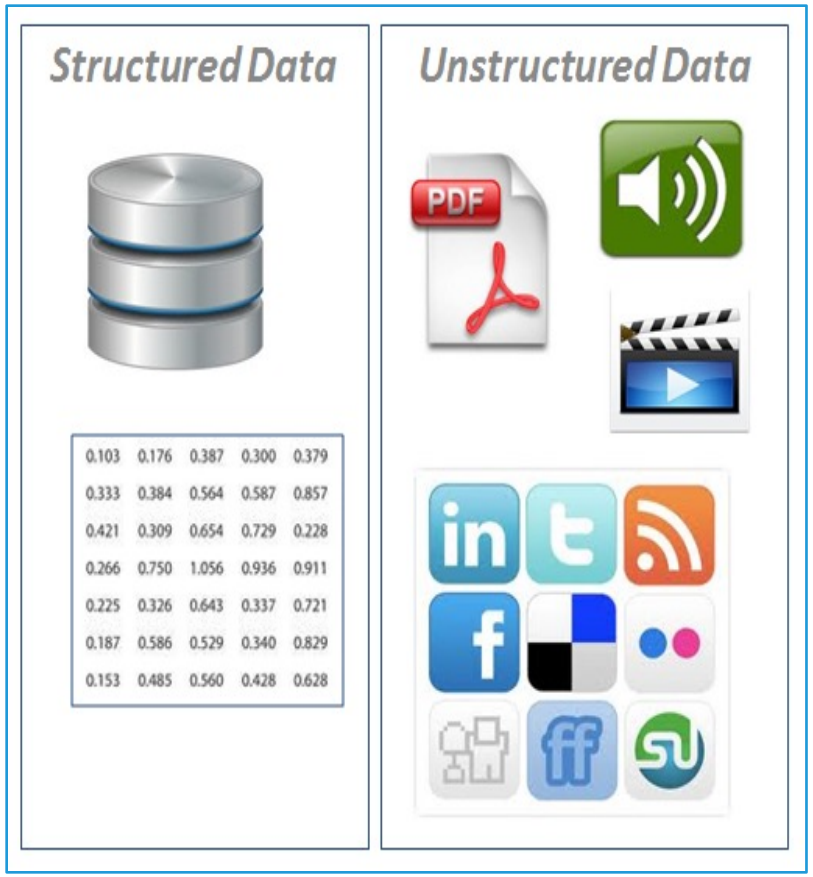
2. **Secondary data**

- Data collected by someone else for some other purpose (but being used by the investigator for another purpose).
- Indirect method of data collection.
- Eg. Census data being used to study the impact of education on income.

Types of Data

1. **Structured data**
- Information is stored with high degree of organisation.
- Contains qualitative data, quantitative data or a mixture of both.
- Eg. Data arranged in Excel file in rows & columns

2. **Unstructured data**
- Information that either does not have a pre-defined data model and/or is not organized in a pre-defined manner.
- Eg. E-mails, tweets, blogs etc.



Measurement Scales

Nominal scale Ordinal scale Numeric(Interval/Ratio Scale)

1. **Nominal scale**

- Placing of data into categories without any order or structure.
- No numerical relationship between categories even if numbers are used for representation.
- eg. Gender, nationality, language, region etc.

2. **Ordinal scale**

- Placing of data into categories such that order of values is meaningful but relative degree of difference is not known.
- eg. Ranking the features of a product on a scale of 1 to 5.
- Likert scale: Psychometric scale commonly used in questionnaires.

Highly Dissatisfied	Dissatisfied	Neutral	Satisfied	Highly Satisfied
1	2	3	4	5

Measurement Scales

Data

Respondent	Gender	Region	Age	Satisfaction Level
1	M	1	23	3
2	M	2	45	4
3	M	2	33	3
4	F	2	25	4
5	F	3	37	2
6	M	1	35	1
7	M	2	41	5
8	F	3	27	2

Description

Region	1	Mumbai
	2	Delhi
	3	Kolkata
Satisfaction Level	1	Highly dissatisfied
	2	dissatisfied
	3	Neutral
	4	Satisfied
	5	Highly satisfied

Gender: Nominal
Region: Nominal
Age: Numeric
Satisfaction Level: Ordinal

Measures of Central Tendency

Measure of Central Tendency:

It is a single value that describe a set of data by identifying the central position within that set of data.

Most commonly used measures of central tendency are :

Mean	**Arithmetic Mean. Commonly known as Average.**
	It is the sum of all values of the variable divided by the total number of values.
Median	**Arrange the data in ascending order, Median is the middle value, if N is odd.**
	If N is even, it is the average of two middle values.
Mode	**It is the most frequently occurring observation in a set of data.**

Calculating Mean, Median, Mode

Consider marks of 12 students in an examination
13, 20, 16, 17, 09, 18, 17, 11, 08, 17, 12, 20

Now, Mean is the sum of all values of the variable divided by the total number of values

$$\frac{\sum_{i=1}^{n} x_i}{n} = \frac{13+20+16+17+09+18+17+11+08+17+12+20}{12} = \frac{178}{12} = 14.83$$

Here, N is even, Median is average of two middle values after arranging the data in ascending order

Data in Ascending order : 08, 09, 11,12, 13, 16, 17, 17, 17, 18, 20, 20

Average of middle two values : $\frac{16+17}{2} = 16.5$

Mode is the most frequently occurring observation in a set of data. Therefore, here, Mode is 17

MEAN	14.83	Average marks scored by the students
MEDIAN	16.5	Half of the students have scored above and half below this
MODE	17	Most frequent observation

Is "Mean" always best measure?

Appropriate Measure of Central Tendency

Type of Variable	Best Measure
Nominal	Mode
Ordinal	Median
Numeric(Interval/Ratio) (Symmetric)	Mean
Numeric(Interval/Ratio) (Not Symmetric)	Median

- Mean is appropriate when the distribution is symmetric. For symmetric distribution, the mean is at the centre.
- For a skewed (not symmetric) distribution, mean is generally not at the centre. Median is better measure of central tendency for a skewed distribution.

Measures of Variation

Measure of Dispersion : In addition to a measure of central tendency, it is desirable to have a measure of dispersion (variation) of data.

A measure of dispersion is an indication of the spread of measurements around the center of the distribution.

Two data sets can have equal mean (measure of central tendency) but vastly different variability.

Eg. Score of Batsman A = (78,62,73,54,76,77) & Score of Batsman B = (92,8,78,34,109,99)

So Average scores of two batsmen in 6 innings is equal(=70) whereas Spread around mean is not identical.

Most commonly used measures of variation are :

Range (calculated as Maximum value – Minimum Value)

Inter-Quartile Range IQR (calculated using lower and upper quartile)

Standard Deviation (calculated using sum of squared deviations about mean)

Coefficient of Variation (CV)

As variance has same units as that of the variable, it is inappropriate to use variance to compare two data sets having different units. Hence, there is a need of a quantity without unit like Coefficient of Variation (CV) for effective comparison.

CV is a relative measure of variation and is used to compare variability in two data sets.

The CV is defined as "Standard Deviation divided by Mean" and is generally expressed as a percentage.

Higher the value of CV, more is the variability.
CV is sometimes referred to as "Relative Standard Deviation".

Case Study - 1

Objective

- To compare the performance of two batsmen using the measures of central tendency and measure of variation

Available Information

- Runs scored by two batsman A and B in **6** matches

Runs Scored

Batsman A	Batsman B
78	92
62	8
73	78
54	34
76	109
77	99

DATA SCIENCE INSTITUTE

Observation and Conclusion

Batsman A	Batsman B
78	92
62	8
73	78
54	34
76	109
77	99
MEAN = 70	MEAN = 70
CV = 13.97%	CV = 57.32%

- Average scores of two batsmen in 6 innings is equal(=70) but the spread around mean is not identical.
- We can see that variability in performance of Batsman B is more than that of Batsman A. Hence, we can infer that Batsman A is a more consistent performer than Batsman B.

Case Study - 2

To learn Descriptive Statistics in R, we shall consider the below case as an example.

Background

Data of 100 retailers in platinum segment of the FMCG company.

Objective

To describe the variables present in the data

Sample Size

Sample size: 100
Variables: Retailer, Zone, Retailer_Age, Perindex, Growth, NPS_Category

Data Snapshot

Retail Data

Variables

Retailer	Zone	Retailer_Age	Perindex	Growth	NPS_Category
1	North	<=2	81.84	3.04	Promoter

Observations

Columns	Description	Type	Measurement	Possible values
Retailer	Retailer ID	numeric	-	-
Zone	Location of the retailer	character	East, West, North, South	4
Retailer_Age	Number of years doing business with the company	character	<=2, 2 to 5, >5	3
Perindex	Index of performance based on sales, buying frequency and buying recency	numeric	-	positive values
Growth	Annual sales growth	numeric	-	positive values
NPS_Category	Category indicating loyalty with the company	character	Detractor, Passive, Promoter	3

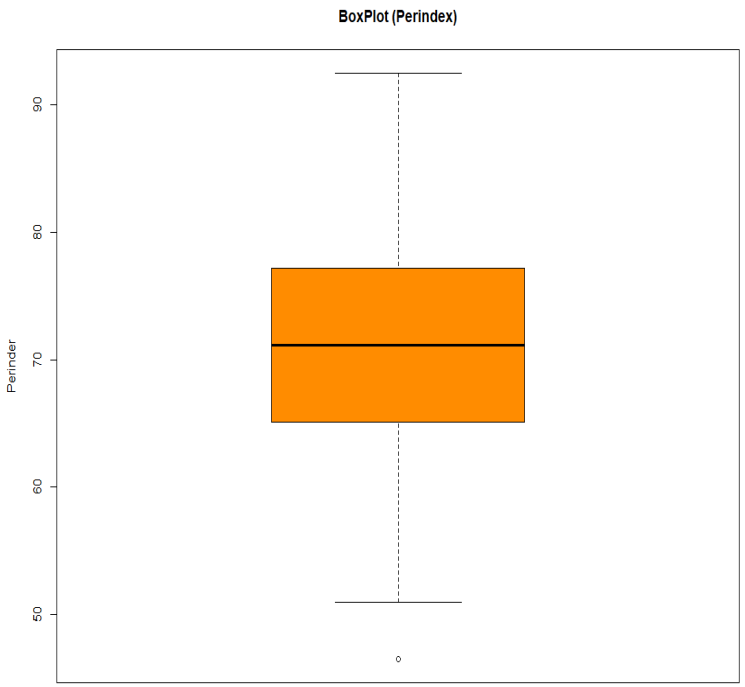
Describing Variables in R

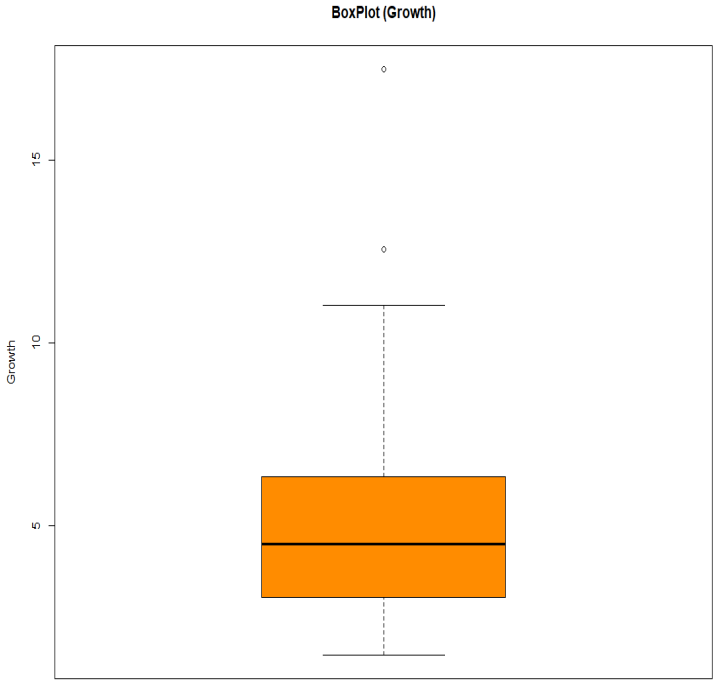
```
#Importing Data
retail_data <-read.csv("Retail_Data.csv" header=TRUE)

#Checking the variable features using summary function
summary(retail_data)

# Output
```

```
   Retailer          Zone     Retailer_Age     Perindex         Growth          NPS_Category
Min.   :  1.00   East :15   <=2   :17   Min.   :46.53   Min.   : 1.470   Detractor:23
1st Qu.: 25.75   North:25   >5    :56   1st Qu.:65.08   1st Qu.: 3.058   Passive  :41
Median : 50.50   South:32   2 to 5:27   Median :71.15   Median : 4.495   Promoter :36
Mean   : 50.50   West :28               Mean   :70.50   Mean   : 5.153
3rd Qu.: 75.25                          3rd Qu.:77.17   3rd Qu.: 6.340
Max.   :100.00                          Max.   :92.49   Max.   :17.500
                                        NA's   :1
```

Understanding Data through Visualisation

```
boxplot(retail_data$Perindex, data= retail_data, main =  "BoxPlot
(Perindex)",ylab = "Perindex",col = "darkorange")

boxplot(retail_data$Growth, data= retail_data, main =  "BoxPlot
(Growth)",ylab = "Growth",col = "darkorange")
```





Here we can see that Perindex variable is distributed symmetrically whereas Growth variable is Positively Skewed.

Measures of Central Tendency in R

```
# Mean for Perindex & Growth Variables
mean(retail_data$Perindex)
[1] NA
```

mean() *in R, gives mean of the variable.*

```
mean(retail_data$Perindex,na.rm = T)
[1] 70.49697
```

Using ***na.rm=T*** *excludes the missing values from the mean*

```
mean(retail_data$Growth,na.rm = T)
[1] 5.1528
```

```
# Median for Perindex & Growth Variables
median(retail_data$Perindex,na.rm = T)
[1] 71.15
```

median() *in R, gives median of the variable.*

```
median(retail_data$Growth,na.rm = T)
[1] 4.495
```

Measures of Central Tendency in R

```
# Summarizing Categorical Variable

freq <- table(retail_data$Zone)
freq

East North South  West
   15    25    32    28
```

table() *in R, gives the frequency of counts of the variable mentioned.*

Measures of Dispersion in R

```
# Standard Deviation/ Variance
sd(retail_data$Perindex,na.rm = T)
```

```
[1] 9.569232
```

sd() in R, gives standard deviation of the variable

```
sd(retail_data$Growth)
```

```
[1] 2.620525
```

```
# Coefficient of Variation
```

```
cv_PI <- sd(retail_data$Perindex,na.rm = T)/
mean(retail_data$Perindex,na.rm = T)*100
cv_PI
```

```
[1]13.57396
```

There is no standard function for CV in R. Hence we calculate it by definition.

```
cv_G <- sd(retail_data$Growth)/mean(retail_data$Growth) *100
cv_G
```

```
[1]50.85633
```

Descriptive Statistics

Bivariate Relationships

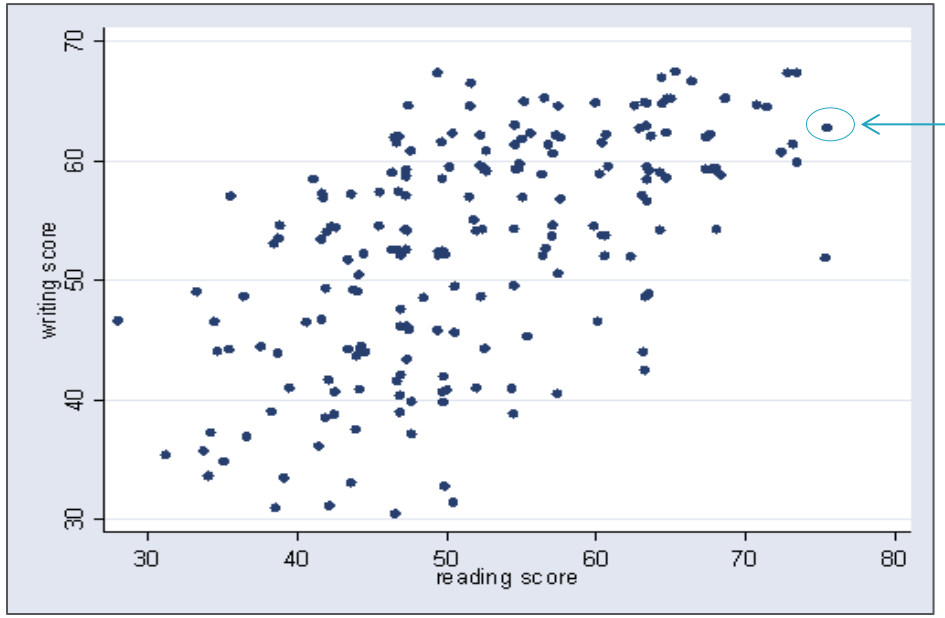
Describing a Bivariate Relationship

The **relationship between two numeric variables** can be described using :

- Scatter Plot : Scatter plot provides nature of relationship graphically
- Correlation Coefficient : Correlation coefficient measures degree of linear relationship
- Simple Linear Regression : Simple Linear Regression gives equation of the type Y=a +bX which can be used to estimate the value Y for any given value of X.

What is a Scatterplot?

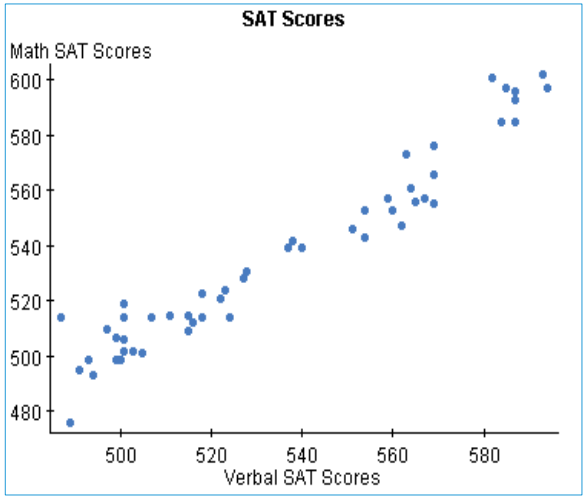
- A scatter plot consists of a X axis (the horizontal axis), a Y axis (the vertical axis), and a series of dots.
- The X-axis and Y-axis represent the values of one variable each.
- Each dot on the scatterplot is one observation from a data set representing the corresponding variable value on X and Y axis respectively
- This plot can be used only for two numeric continuous variables



This dot represents a person having reading score of approx. 74 and writing score of 62.

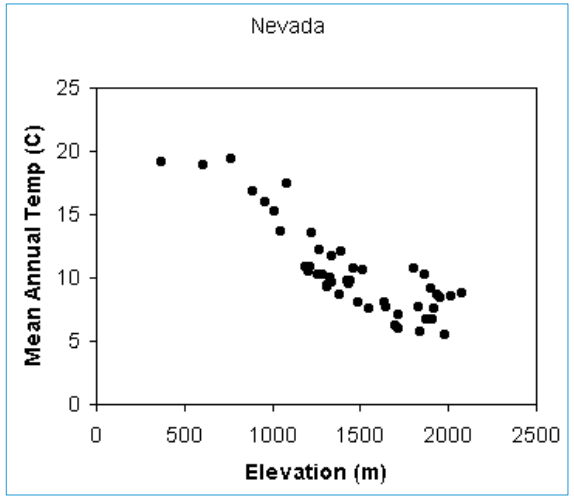
Interpreting a Scatterplot

Positive Correlation



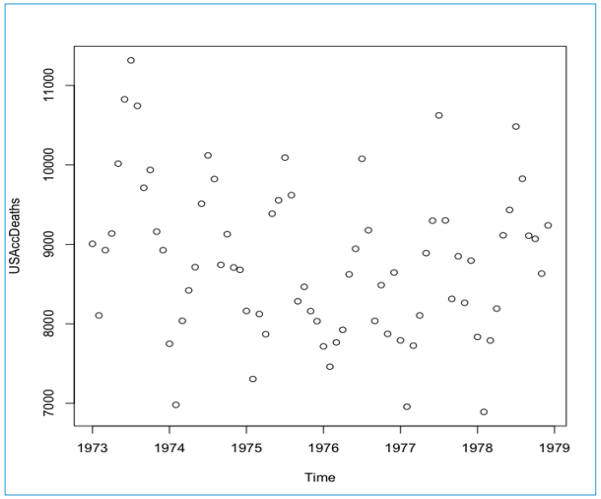
This is a positive sloping (upward) graph.
As the value of one variable increases, the value of other variable also increases.

Negative Correlation



This is a negative sloping (downward) graph.
As the value of one variable increases, the value of other variable tends to decrease.

No Correlation



This is a graph with random pattern.
There is no connection between the two variables.
If value of one variable increases, other might increase/decrease.

Pearson's Coefficient of Correlation

The Pearson's correlation coefficient numerically measures the strength of a linear relation between two variables

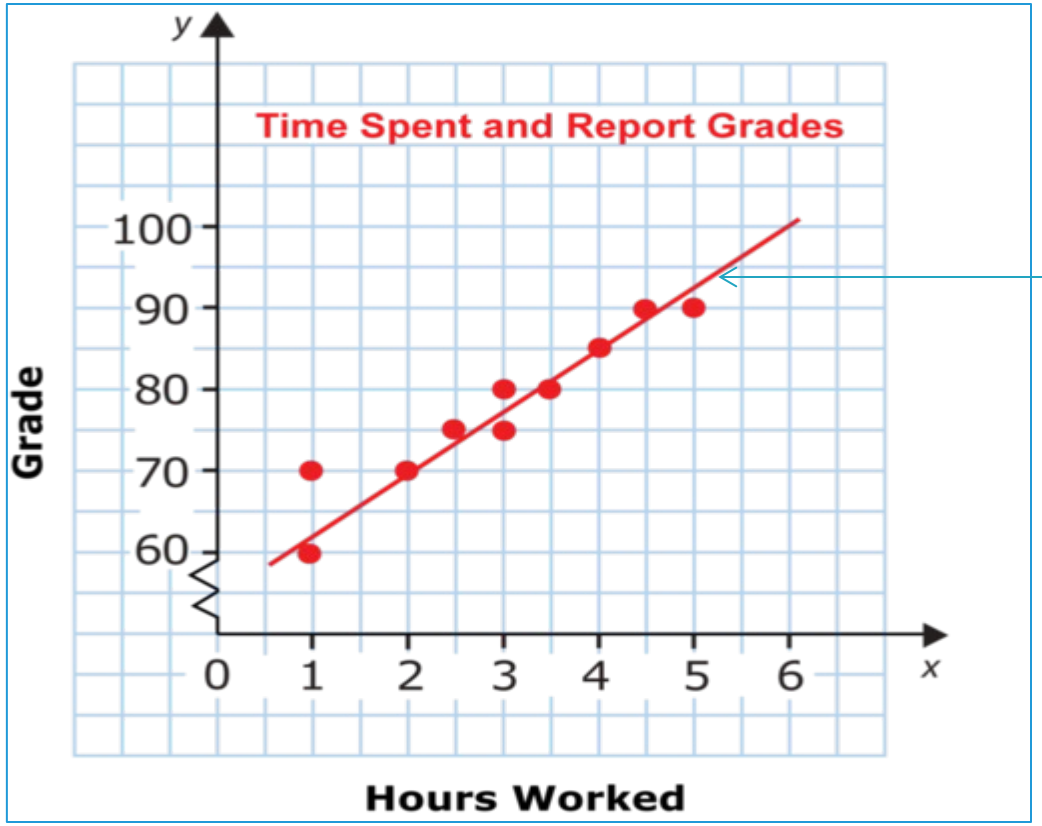
$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}} = \frac{cov\,(X,Y)}{sd(x)sd(y)}$$

RANGE -1 <= r < =1	
Positive Correlation	r > 0
Negative Correlation	r < 0
No Correlation	r = 0

- The two variables can be measured in entirely different units.
- Example, you could correlate a person's age with their blood sugar levels. Here, the units are completely different.
- It is not affected by change of Origin and Scale

Line of Best Fit : Regression Line

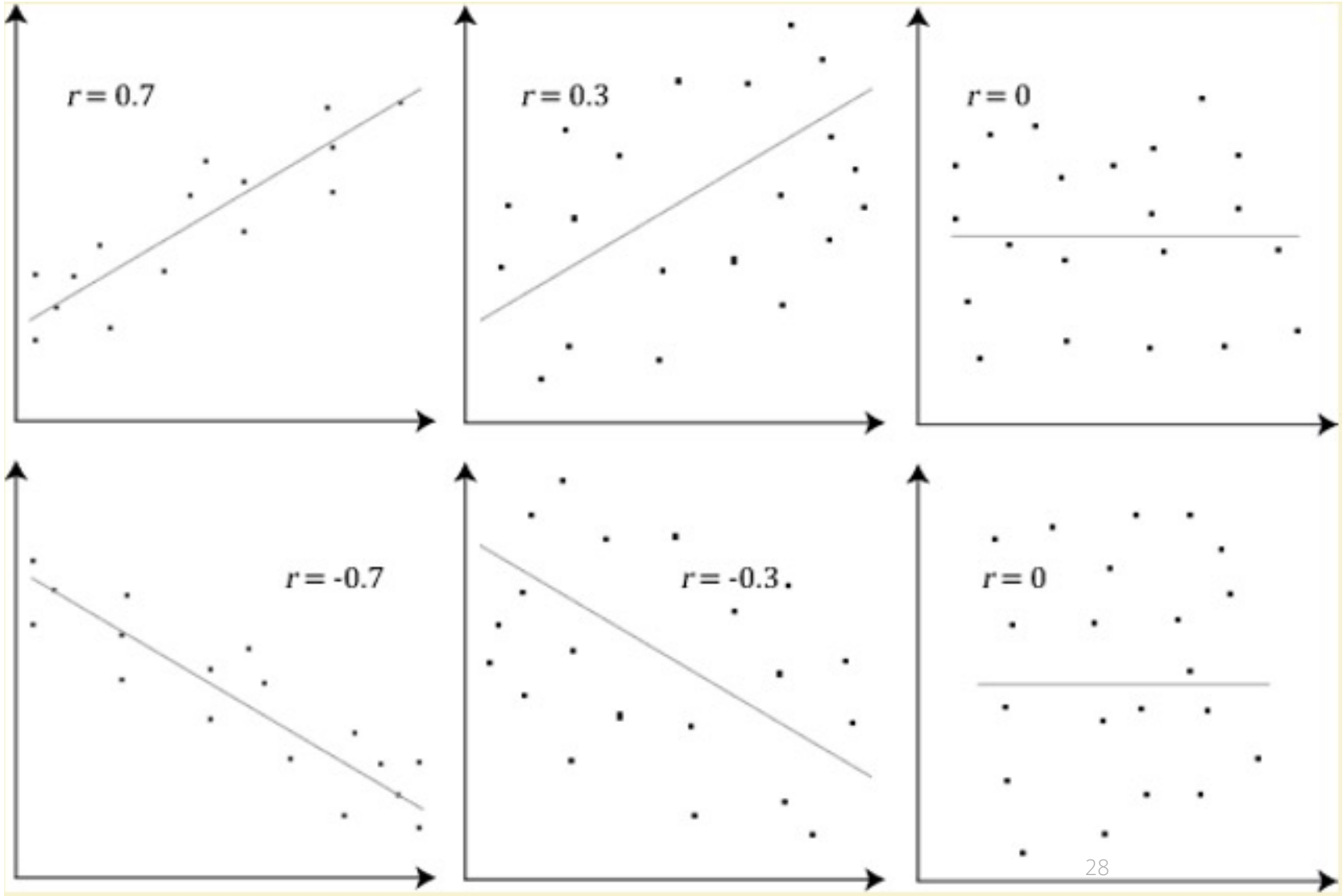
A line of best fit (or "trend" line) is a straight line that best represents the data on a scatter plot.



Line of Best Fit :

This line may pass through some of the points, none of the points, or all of the points.

Relationships and r



DATA SCIENCE INSTITUTE

Simple Linear Regression

The equation of line of best fit is used to describe relationship between two variables

Mathematical form of simple linear regression : $Y = a + bX$ + e

Where,

a : Intercept (The value at which the fitted line crosses the y-axis i.e. X=0)
b : Slope of the Line
e : error which is assumed to be a random variable

NOTE : a and b are population parameters which are estimated using sample

Here, variable Y is known as a 'Dependent' variable, that 'depends on' X which is known as the 'Independent' variable.

Case Study

Background

- A company conducts different written tests before recruiting employees. The company wishes to see if the scores of these tests have any relation with post-recruitment performance of those employees.

Objective

- To study the correlation between Aptitude and Job Proficiency.
- Predict the Job proficiency for a given Aptitude score.

Available Information

- Sample size is 33
- Independent Variables: Scores of tests conducted before recruitment on the basis of four criteria – **Aptitude, Test of English, Technical Knowledge, General Knowledge**
- Dependent Variable **job_prof**: **Job Performance Index** calculated after an employee finishes probationary period (6 months)

Data Snapshot

Job_Proficiency

Variables

Observations

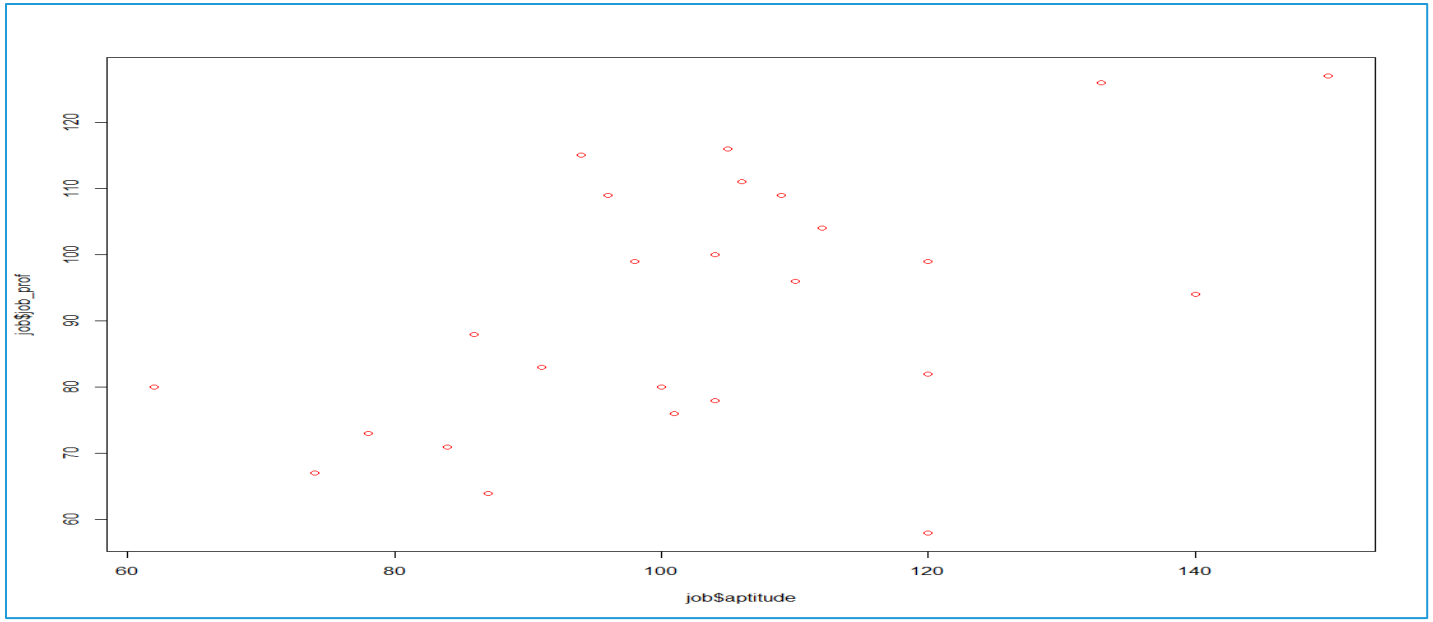
empno	aptitude	testofen	tech_	g_k_	job_prof
1	86	110	100	87	88
2	62	62	99	100	80
3	110	107	103	103	96
4	101	117	93	95	76
5	100	101	95	88	80
6	78	85	95	84	73
7	120	77	80	74	58
8	105	122	116	102	116

Columns	Description	Type	Measurement	Possible values
Empno	Employee Number	numeric	-	positive values
aptitude	Aptitude Score of the Employee	numeric	-	positive values
Testofen	Test of English	numeric	-	positive values
tech_	Technical Score	numeric	-	positive values
g_k	General Knowledge Score	numeric	-	positive values
Job_prof	Job Proficiency Score	numeric	-	positive values

Scatter Plot in R

```
# Importing Data
job<-read.csv("Job_Proficiency.csv",header=T)

# Scatterplot
plot(job$aptitude,job$job_prof,col="red")

# Output
```

- ***plot()*** *gives a scatterplot of the two variables mentioned.*
- ***col=*** *provides color to the points.*



DATA SCIENCE INSTITUTE

Pearson Correlation Coefficient in R

```
# Correlation
cor(job$aptitude,job$job_prof)
[1] 0.5144107
```

cor() *calculates Pearson Correlation Coefficient for the two variables mentioned.*

Pearson Correlation Coefficient	0.5144

There is positive relation between aptitude and job proficiency but the relation is of moderate degree.

Simple Linear Regression

```
# Simple Linear Regression
model1<-lm(job_prof~aptitude, data=job)
model1
```

lm() gives the linear regression model

```
# Output
Call:
lm(formula = job_prof ~ aptitude, data = job)

Coefficients:
(Intercept)     aptitude
    41.3216       0.4922
```

Inferences : Simple Linear Regression

Dependent Variable : **Job Proficiency**

Independent Variable : **Aptitude**

Intercept	Aptitude
41. 3216	0.4922

Equation : **Job Proficiency =** 41. 3216 + 0.4922 * **Aptitude**

Here Job Proficiency changes by 0.4992 units with a unit change in aptitude.