Introduction to Data Science

Dr. Irfan Yousuf
Department of Computer Science (New Campus)
UET, Lahore

(Week 3; January 29 – February 02, 2024)

Outline

- Linear Regression Analysis
- Exploratory Data Analysis

Relations Between Variables

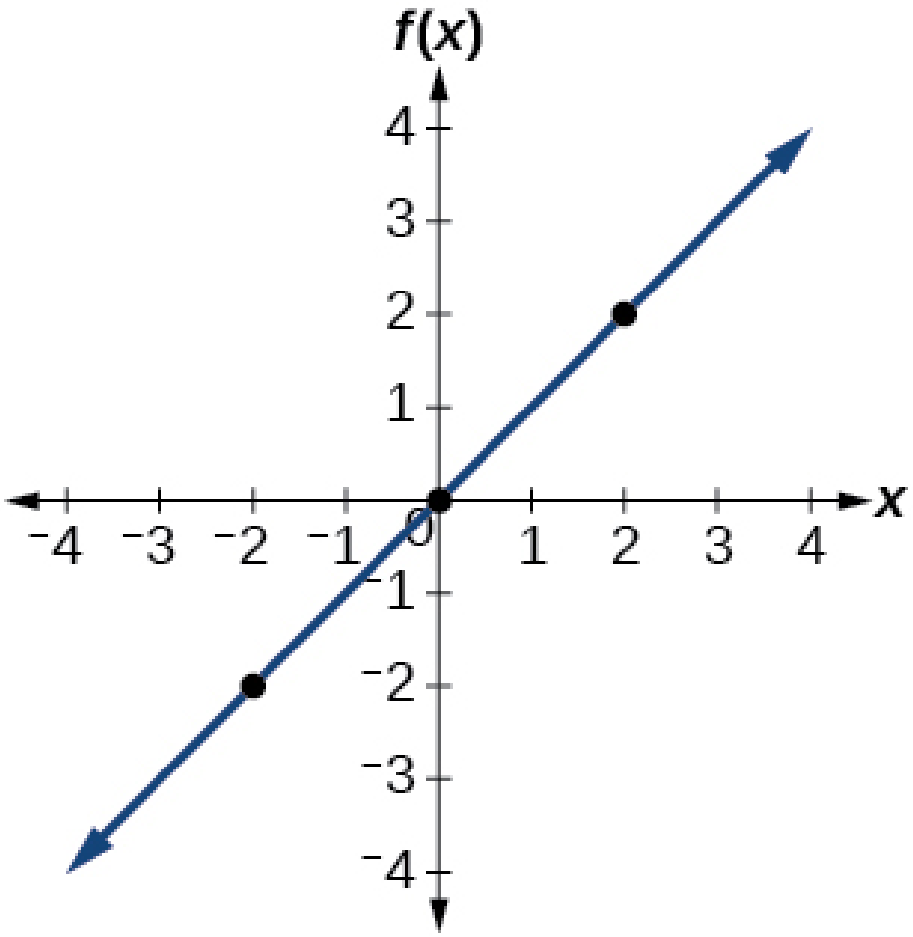
- Functional Relation
- Statistical Relation

Functional Relation

- A functional relation between two variables is a perfect relation where the value of the **dependent variable is uniquely determined by the value of the independent variable**.
- It is expressed by a mathematical formula.

$$y=f(x)$$

Functional Relation

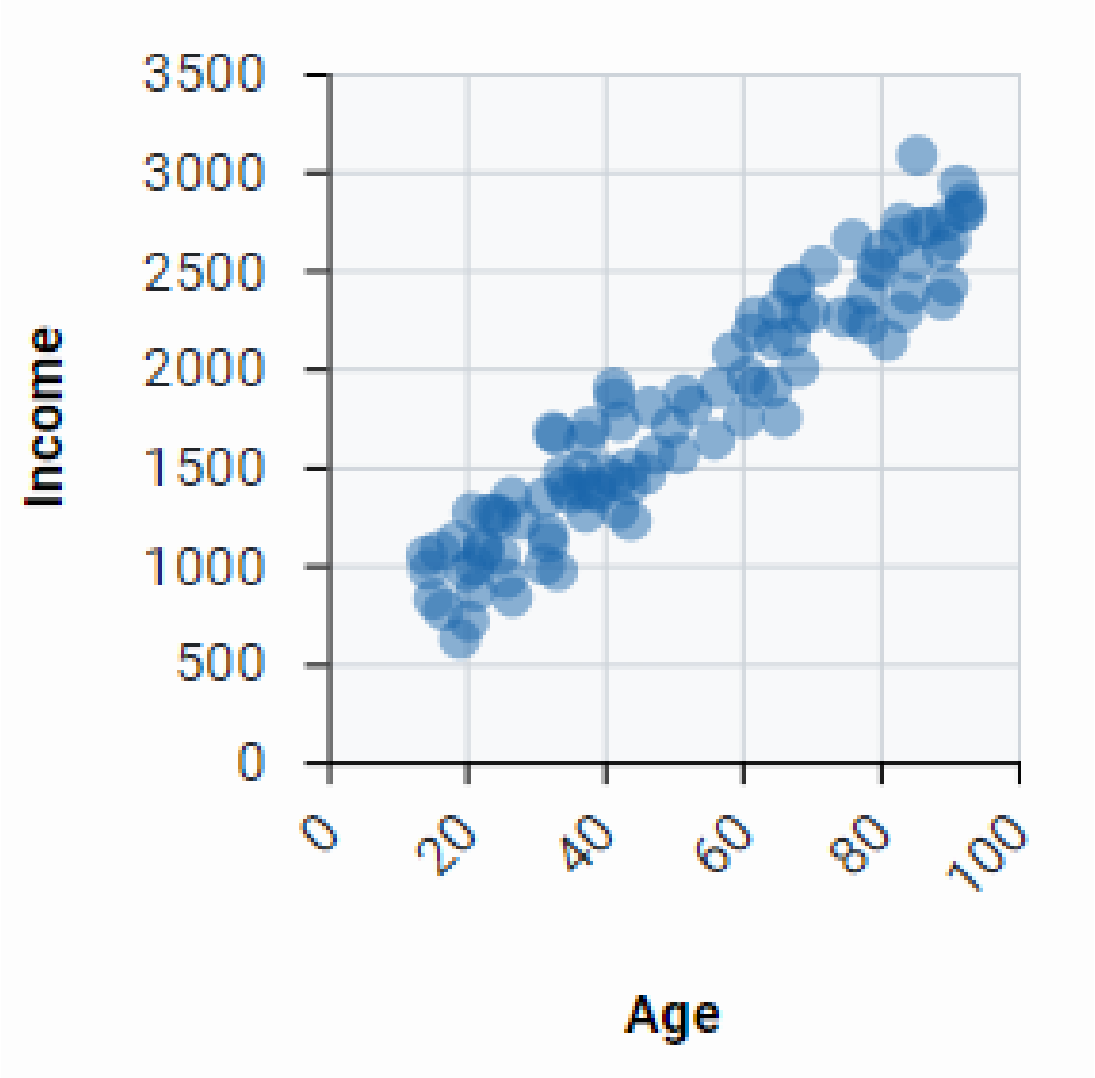


x	$f(x)$
-2	-2
0	0
2	2

Statistical Relation

- A statistical relation between two variables is a relation where **the value of the dependent variable is NOT uniquely determined when the level of the independent variable is specified**.
- If the values of a variable Y increase or decrease when the values of a variable X change, there is a statistical relationship between Variable Y and Variable X.
- It is not an exact relation.
- Examples:
  - The relation between age and income
  - The relation between income and expenditures

Statistical Relation



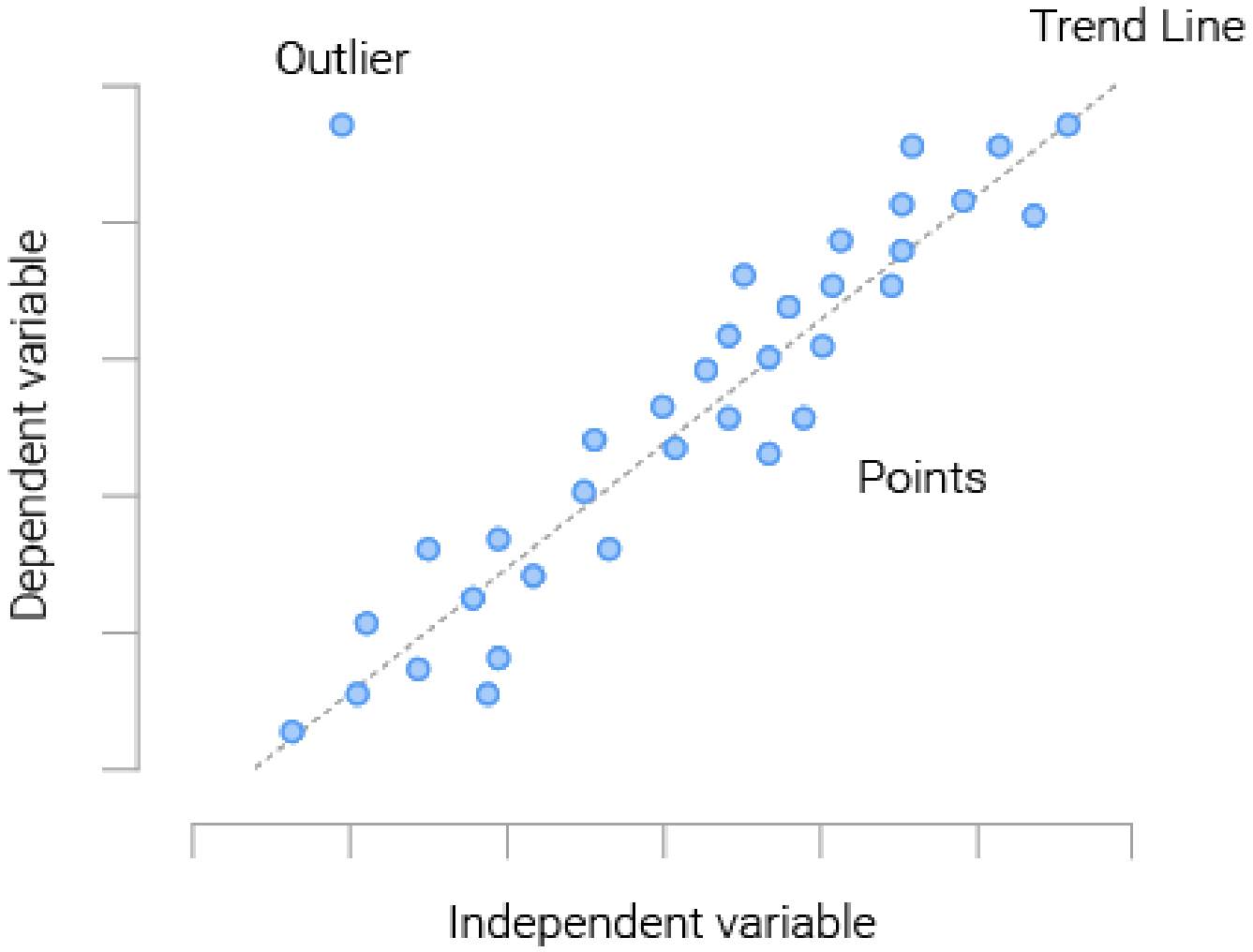
Scatter Plot

- A scatter plot (aka scatter chart, scatter graph) uses dots to represent values for two different numeric variables.

- The position of each dot on the horizontal and vertical axis indicates values for an individual data point.

- Scatter plots are used to observe relationships between variables.

Scatter Plot



Regression Analysis

- Regression analysis provides a method of estimating an average relation (often linear) between two or more variables.

- **Regressor:** The variable that forms the basis of estimation or prediction (aka predictor variable or **independent variable**).

- **Regressand:** The variable whose value depend on the independent variable is called a regressand (aka response variable, predictand variable or **dependent variable**).

Regression Models

- Regression models describe the relationship between variables by fitting a line to the observed data.

- Linear regression models use a straight line, while logistic and nonlinear regression models use a curved line.

- Regression allows you to estimate how a dependent variable changes as the independent variable(s) change.

Simple Linear Regression

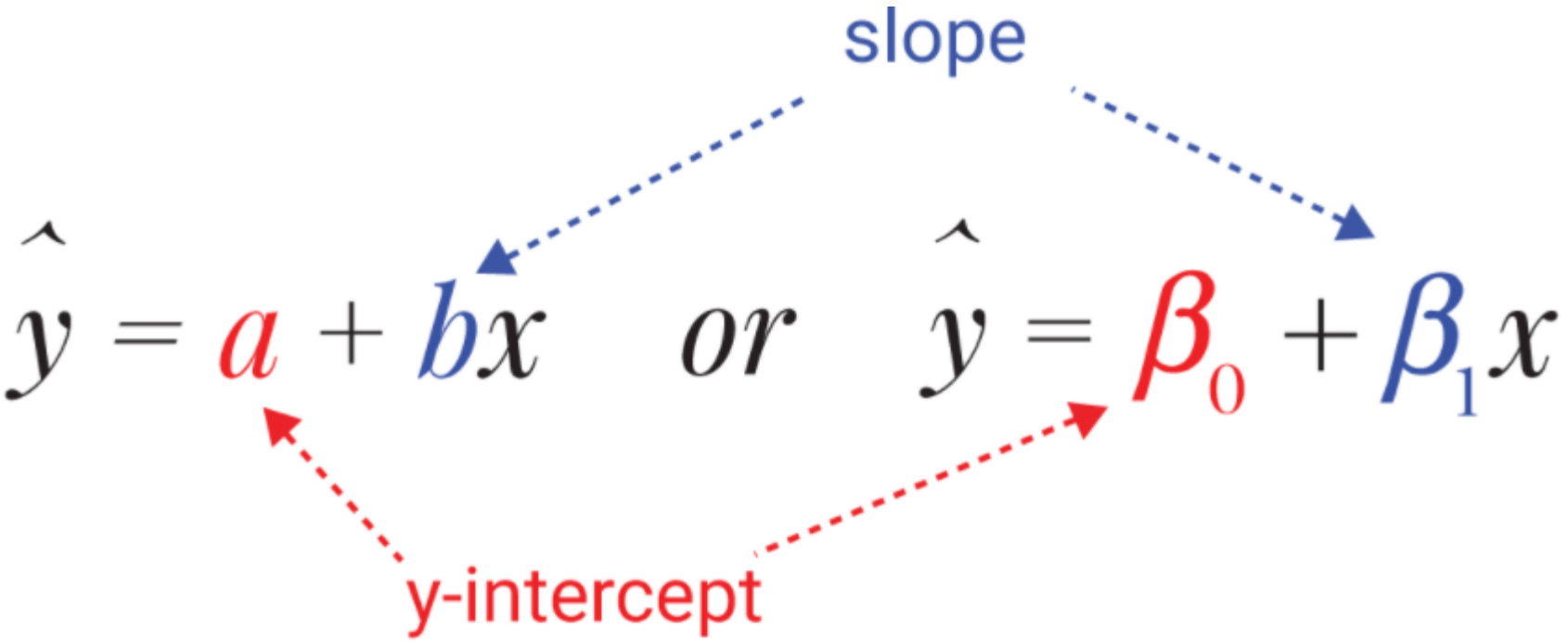
The formula for a simple linear regression is:

$$y = \beta_0 + \beta_1 X + \varepsilon$$

- **y** is the predicted value of the dependent variable (**y**) for any given value of the independent variable (**x**).
- $\mathbf{B_0}$ is the **intercept**, the predicted value of **y** when the **x** is 0.
- $\mathbf{B_1}$ is the regression coefficient – how much we expect **y** to change as **x** increases.
- **x** is the independent variable (the variable we expect is influencing **y**).
- **e** is the **error** of the estimate, or how much variation there is in our estimate of the regression coefficient.

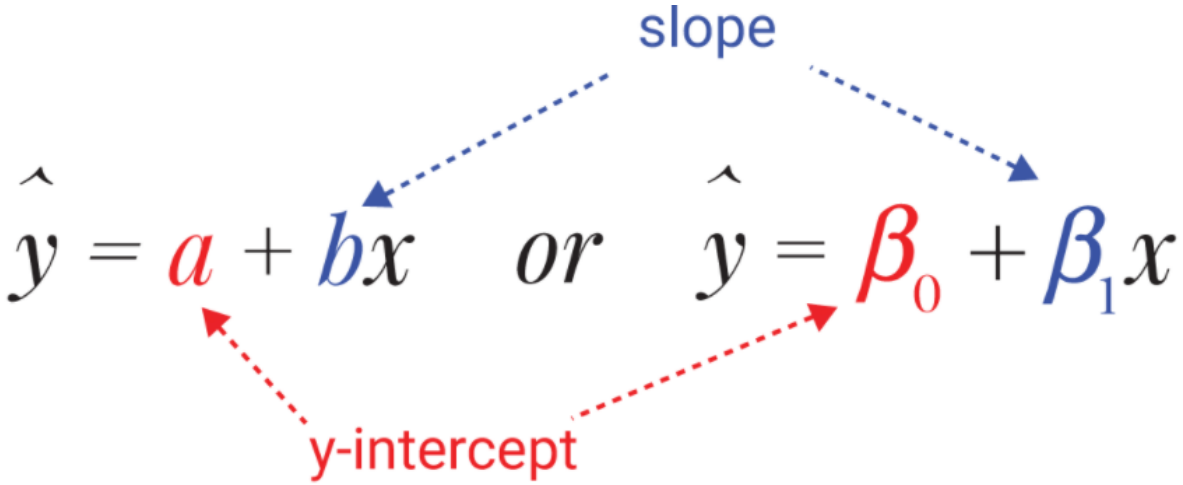
Simple Linear Regression

- Best-fit Line
- Least Squares Method or Least Squares Regression



Simple Linear Regression

- Best-fit Line
- Least Squares Method or Least Squares Regression



$$a = \frac{\sum y - b \sum x}{n}$$

$$b = \frac{n \sum (xy) - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$

Simple Linear Regression: Example

x	y
14.2	215
16.4	325
11.9	185
15.2	332
18.5	406
22.1	522
19.4	412
25.1	614
23.4	544
18.1	421
22.6	445
17.2	408

x	y	xy	x^2
14.2	215	3053.00	201.64
16.4	325	5330.00	268.96
11.9	185	2201.50	141.61
15.2	332	5046.40	231.04
18.5	406	7511.00	342.25
22.1	522	11536.20	488.41
19.4	412	7992.80	376.36
25.1	614	15411.40	630.01
23.4	544	12729.60	547.56
18.1	421	7620.10	327.61
22.6	445	10057.00	510.76
17.2	408	7017.60	295.84
224.1	4829	95506.60	4362.05

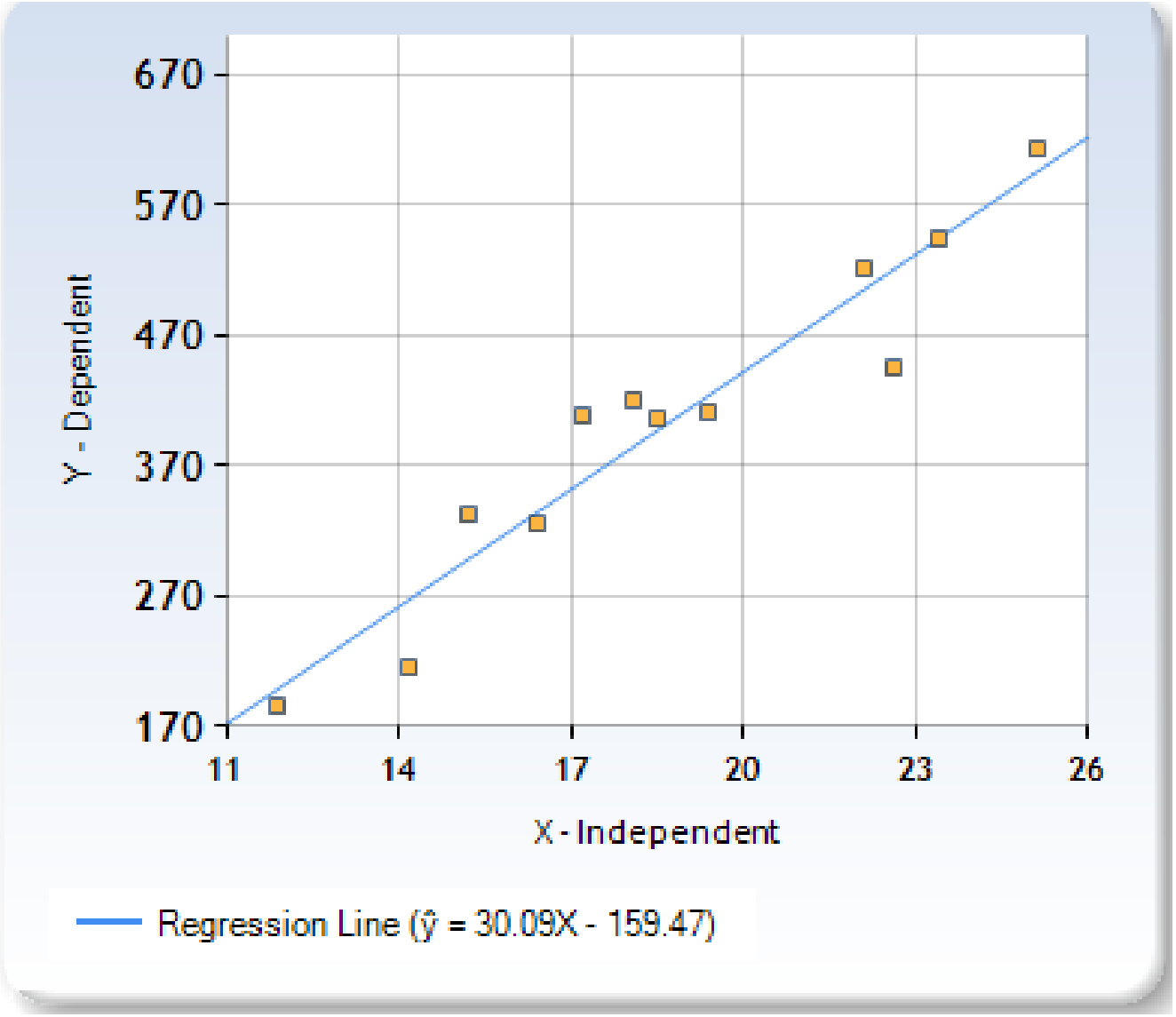
$$a = \frac{\sum y - b\sum x}{n}$$

$$b = \frac{n\sum(xy) - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

b = 30.08

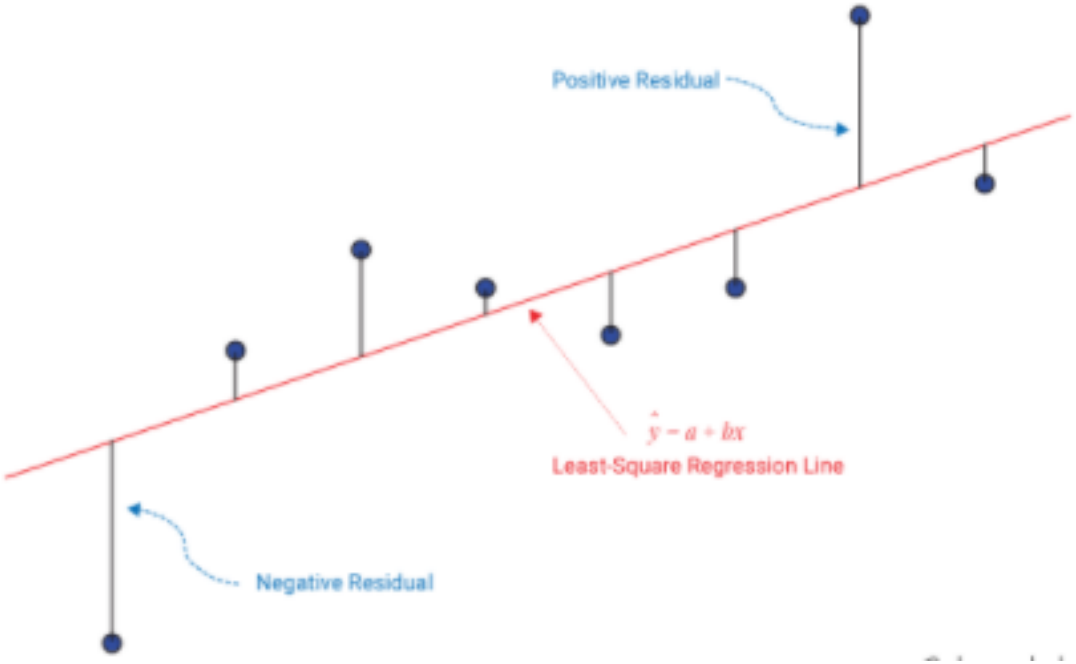
a = -159.4

Simple Linear Regression: Example

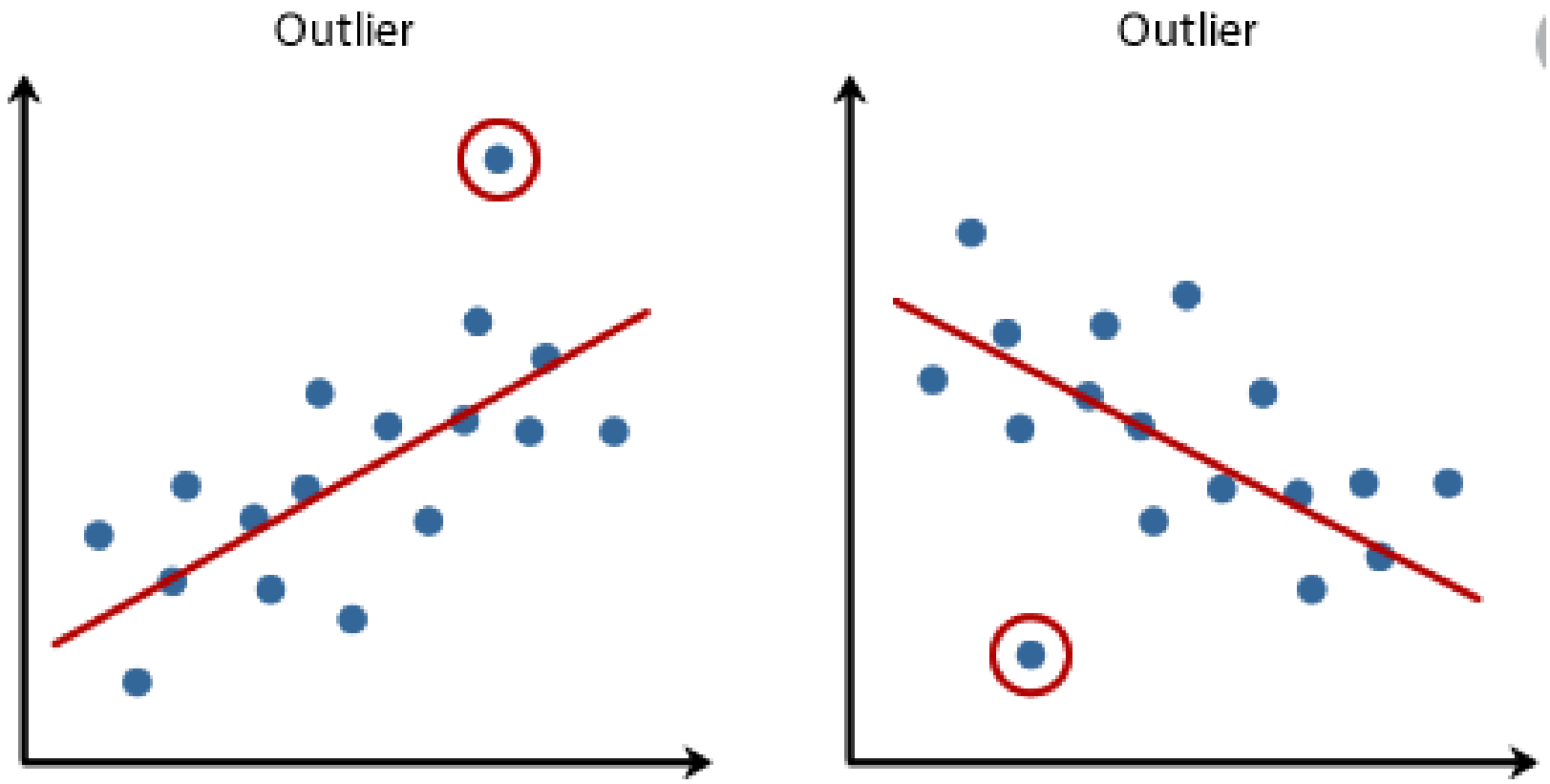


Residuals

- The residuals are the differences between the observed and predicted values.
- It measures the distance from the regression line (predicted value) and the actual observed value. In other words, it helps us to measure error, or how well our regression line "fits" our data.

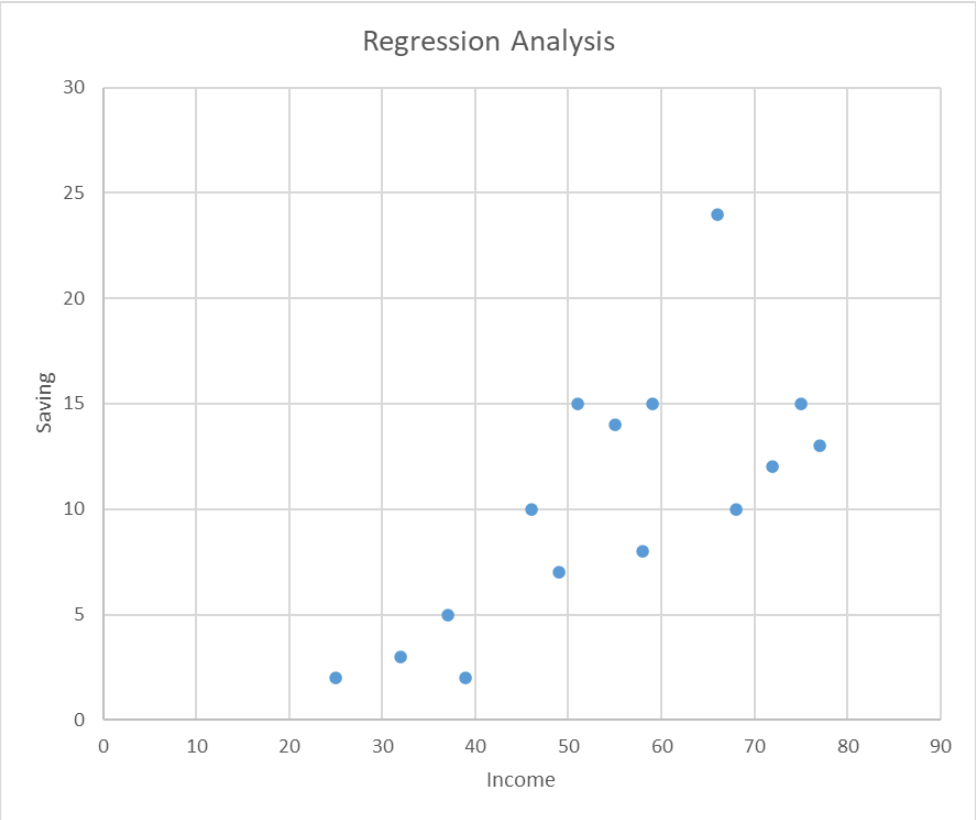


Outliers



Simple Linear Regression: Example

Income	Savings
25	2
32	3
37	5
39	2
46	10
49	7
51	15
55	14
58	8
59	15
66	24
68	10
72	12
75	15
77	13



Income = 85, Saving =?

Income = 42, Saving=?

Simple Linear Regression: Example

Income	Savings
25	2
32	3
37	5
39	2
46	10
49	7
51	15
55	14
58	8
59	15
66	24
68	10
72	12
75	15
77	13

X	Y	X*Y	X^2
25	2	50	625
32	3	96	1024
37	5	185	1369
39	2	78	1521
46	10	460	2116
49	7	343	2401
51	15	765	2601
55	14	770	3025
58	8	464	3364
59	15	885	3481
66	24	1584	4356
68	10	680	4624
72	12	864	5184
75	15	1125	5625
77	13	1001	5929
809	**155**	**9350**	**47245**

$$a = \frac{\sum y - b \sum x}{n}$$

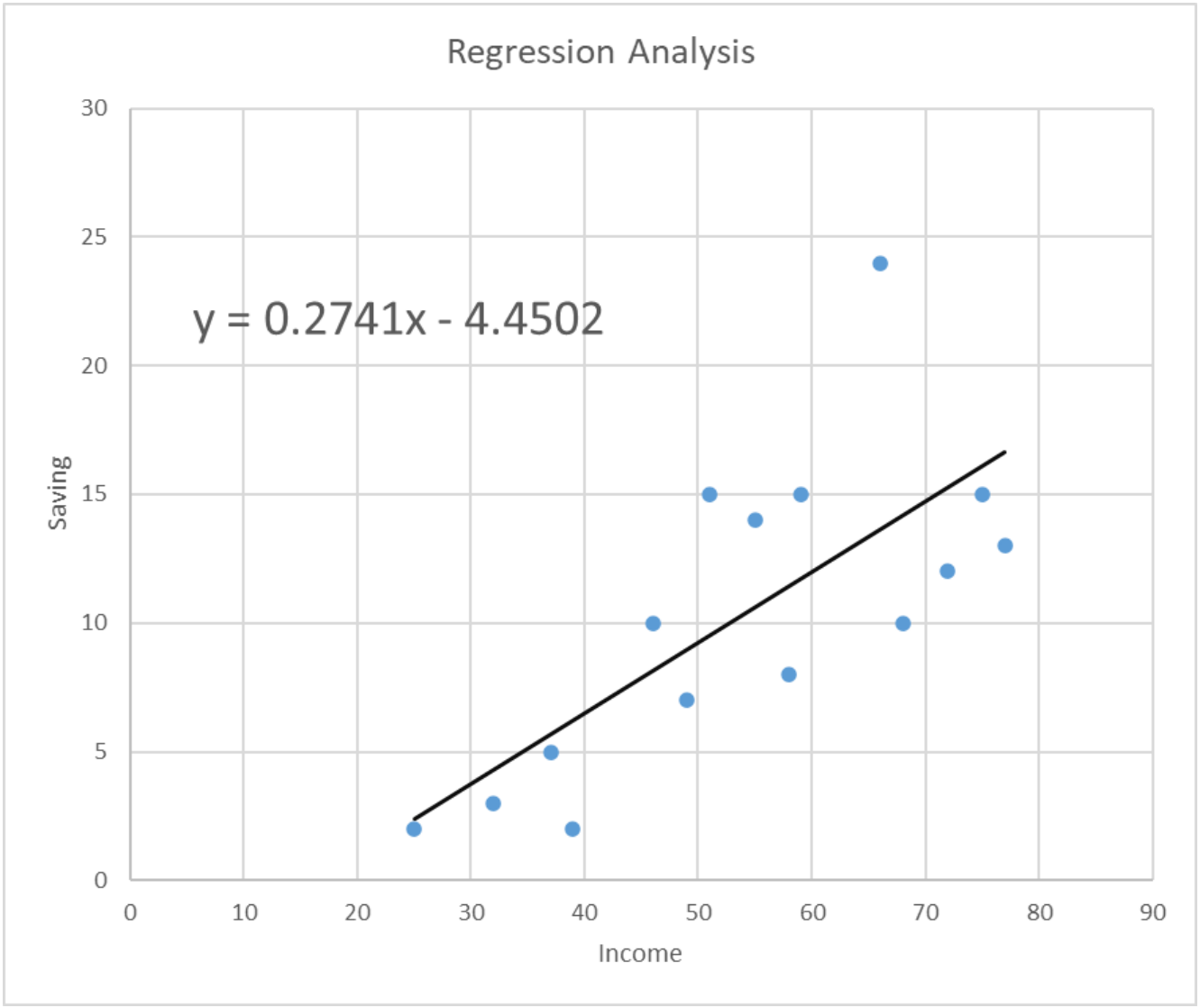
$$b = \frac{n \sum (xy) - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$

a = -4.4502

b = 0.2741

Simple Linear Regression: Example

Income	Savings
25	2
32	3
37	5
39	2
46	10
49	7
51	15
55	14
58	8
59	15
66	24
68	10
72	12
75	15
77	13



Simple Linear Regression: Example

Income	Savings
25	2
32	3
37	5
39	2
46	10
49	7
51	15
55	14
58	8
59	15
66	24
68	10
72	12
75	15
77	13

Income = 85, Saving =?

Income = 42, Saving=?

a = -4.4502

b = 0.2741

Y = -4.4502 + 0.2741(X)

Income = 85, Saving = 18.84

Income = 42, Saving= 7.062

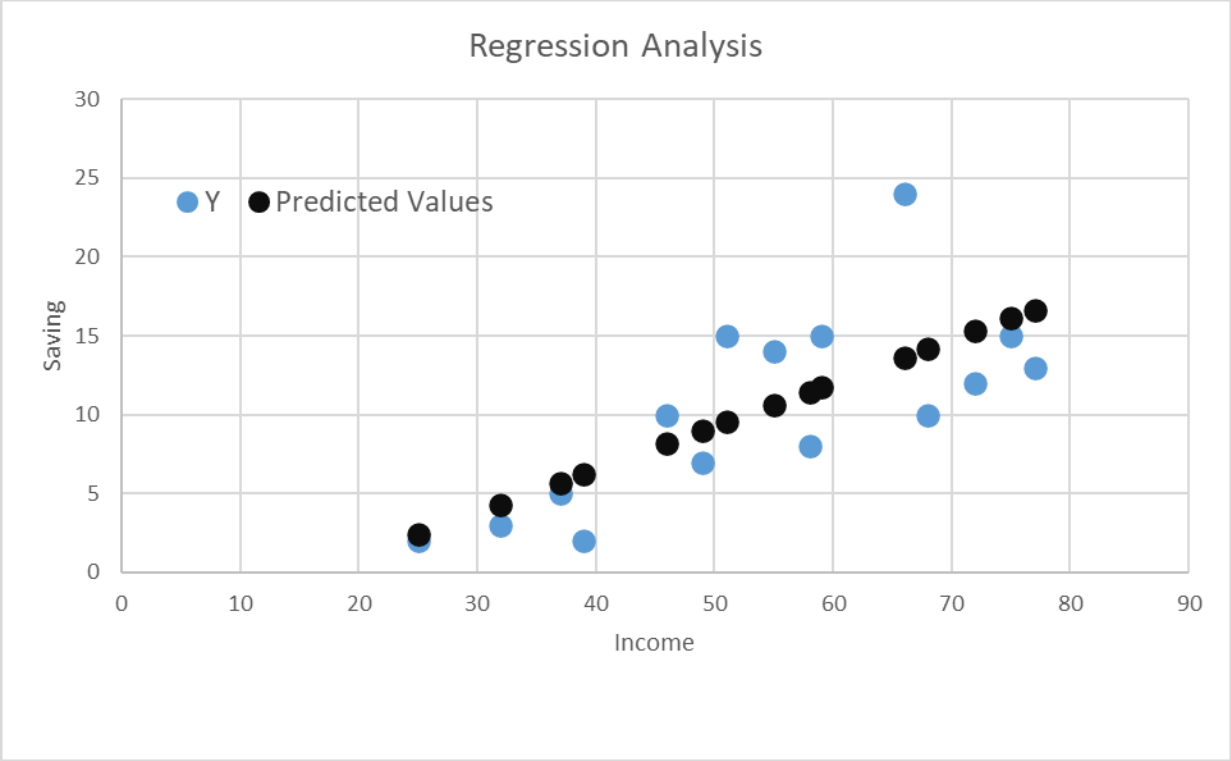
Simple Linear Regression: Example

Income	Savings
25	2
32	3
37	5
39	2
46	10
49	7
51	15
55	14
58	8
59	15
66	24
68	10
72	12
75	15
77	13

X	Y	Predicted Values
25	2	2.4023
32	3	4.321
37	5	5.6915
39	2	6.2397
46	10	8.1584
49	7	8.9807
51	15	9.5289
55	14	10.6253
58	8	11.4476
59	15	11.7217
66	24	13.6404
68	10	14.1886
72	12	15.285
75	15	16.1073
77	13	16.6555

Simple Linear Regression: Example

X	Y	Predicted Values
25	2	2.4023
32	3	4.321
37	5	5.6915
39	2	6.2397
46	10	8.1584
49	7	8.9807
51	15	9.5289
55	14	10.6253
58	8	11.4476
59	15	11.7217
66	24	13.6404
68	10	14.1886
72	12	15.285
75	15	16.1073
77	13	16.6555



R-squared (Goodness of Fit)

- After fitting a linear regression model, you need to determine how well the model fits the data.
- **R-squared is a goodness-of-fit measure for linear regression models**.
- This statistic indicates **the percentage of the variance in the dependent variable that the independent variable can explain** (or that is predictable from the independent variable).
- R-squared measures the **strength of the relationship between the two variables** on a convenient 0 – 100% scale.
- It is also called Coefficient of Determination.

R-squared (Goodness of Fit)

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

y_i = Actual Value
y-hat = Predicted Value
y-bar = Actual Mean

x	y	Predicted (Y)
14.2	215	267.78
16.4	325	333.97
11.9	185	198.58
15.2	332	297.87
18.5	406	397.16
22.1	522	505.47
19.4	412	424.24
25.1	614	595.74
23.4	544	544.59
18.1	421	385.12
22.6	445	520.52
17.2	408	358.04

$$R^2 = 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}},$$
$$= 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \overline{y})^2}.$$

Y= 30.08X – 159.4

$R^2 = 0.9168$

R-squared (Goodness of Fit)

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

y_i = Actual Value
y-hat = Predicted Value
y-bar = Actual Mean

X	Y	Y_{hat}	$(Y - Y_{hat})^2$	$(Y - Y_{avg})^2$
25	2	2.4023	0.1618453	69.3889
32	3	4.321	1.745041	53.7289
37	5	5.6915	0.4781723	28.4089
39	2	6.2397	17.975056	69.3889
46	10	8.1584	3.3914906	0.1089
49	7	8.9807	3.9231725	11.0889
51	15	9.5289	29.932935	21.8089
55	14	10.6253	11.3886	13.4689
58	8	11.4476	11.885946	5.4289
59	15	11.7217	10.747251	21.8089
66	24	13.6404	107.32131	186.869
68	10	14.1886	17.54437	0.1089
72	12	15.285	10.791225	2.7889
75	15	16.1073	1.2261133	21.8089
77	13	16.6555	13.36268	7.1289
			241.87521	513.334

$$R^2 = 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}},$$

$$= 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}.$$

$SS_{RES} = 241.87$

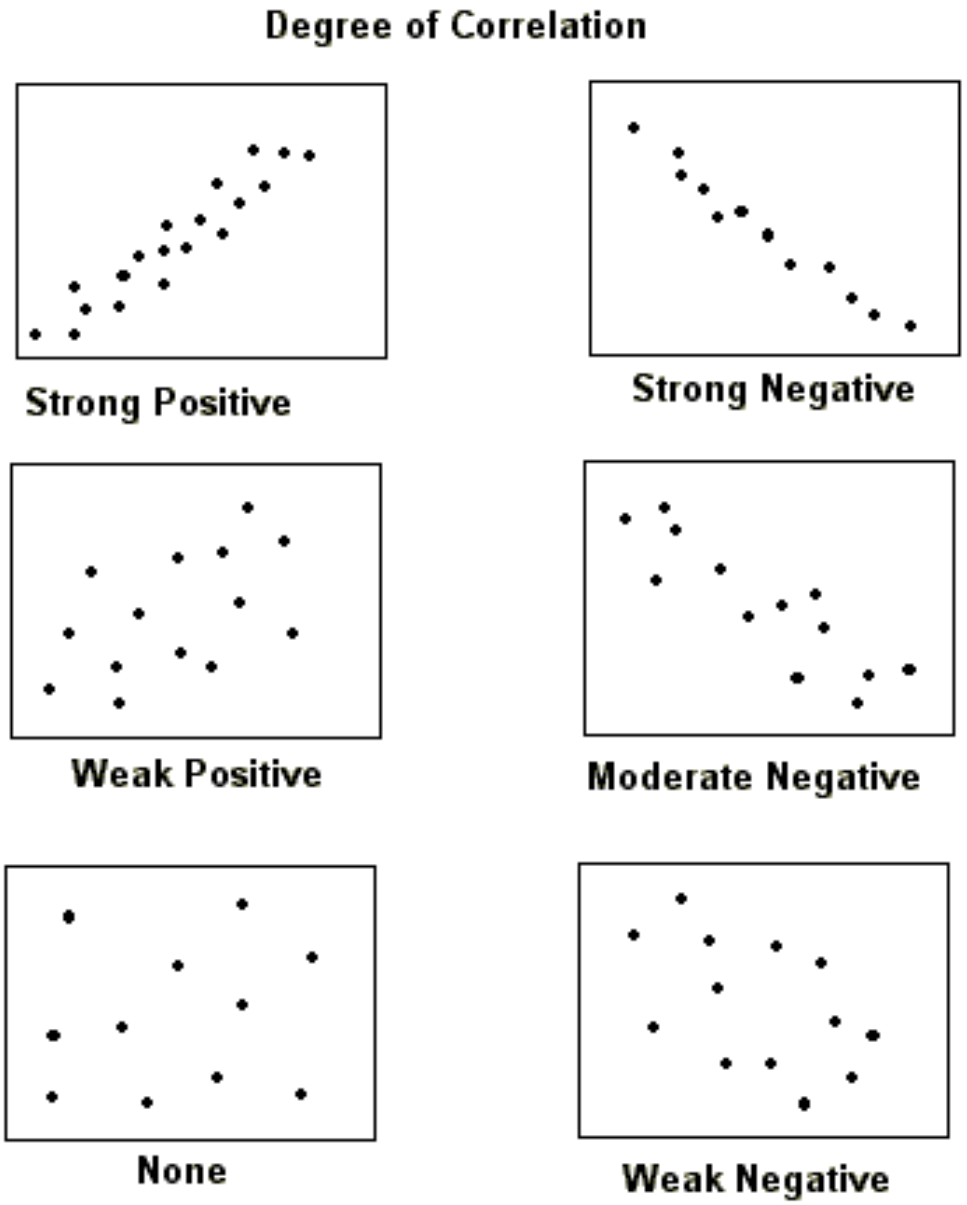
$SS_{TOT} = 513.33$

$R^2 = 0.5288$

Correlation

- The direction and strength of pairwise relationships between two or more numeric variables.
- **-1: Perfect negative correlation**. The variables tend to move in opposite directions (i.e., when one variable increases, the other variable decreases).
- **0: No correlation.** The variables do not have a relationship with each other.
- **1: Perfect positive correlation.** The variables tend to move in the same direction (i.e., when one variable increases, the other variable also increases).

Correlation



Correlation

Correlation Coefficient Formula

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

r = 0.9575

How to Interpret Correlation?

- **Strength:** The greater the absolute value of the correlation coefficient, the stronger the relationship.

- **Direction:** The sign of the correlation coefficient represents the direction of the relationship.

How to Interpret Correlation?

- The correlation, denoted by r, measures the **amount of linear association between two variables**.
- r is always between -1 and 1 inclusive.

- The R-squared value, denoted by R^2, is the **square of the correlation**. It measures the **proportion of variation in the dependent variable that can be attributed to the independent variable**.
- The R-squared value is always between 0 and 1 inclusive.

Correlation vs. Regression

- Correlation is a statistical **measure** that quantifies the direction and strength of the relationship between two numeric variables.

- Regression is a statistical **technique** that predicts the value of the dependent variable Y based on the known value of the independent variable X through an equation.

Drawbacks of Linear Regression

- Linear regression only looks at linear relationships between dependent and independent variables.
- Linear regression looks at a relationship between the mean of the dependent variable and the independent variables.
- Linear regression is sensitive to outliers

Exploratory Data Analysis

Data

- Data is a collection of facts, such as numbers, words, measurements, observations or just descriptions of things.
- **Qualitative data** is descriptive information (it describes something).
- **Quantitative data** is numerical information (numbers).
- Discrete data can only take certain values (like whole numbers)
- Continuous data can take any value (within a range)
- Discrete data is counted, Continuous data is measured

Data Analysis

- Data analysis is the process of collecting, cleaning, analyzing, interpreting, and visualizing data to discover valuable insights or useful information.

Data Science

- Data science is the domain of study that deals with large volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions.

- Data science uses complex (e.g., machine learning) algorithms to build predictive models.

Data Analyst

- A **data analyst** makes sense out of existing data.
- Data analysts typically work with structured data to solve business problems using tools like SQL, R or Python programming languages, data visualization software, and statistical analysis.
  - Collaborating with organizational leaders to identify informational needs
  - Acquiring data from primary and secondary sources
  - Cleaning and reorganizing data for analysis
  - Analyzing data sets to spot trends and patterns that can be translated into actionable insights
  - Presenting findings in an easy-to-understand way to inform data-driven decisions

Data Scientist

- A **data scientist** works on new ways of capturing and analyzing data to be used by the analysts. A data scientist is someone who creates programming code and combines it with statistical knowledge to create insights from data.

  - Gathering, cleaning, and processing raw data
  - Designing predictive models and machine learning algorithms to mine big data sets
  - Developing tools and processes to monitor and analyze data accuracy
  - Building data visualization tools, dashboards, and reports
  - Writing programs to automate data collection and processing

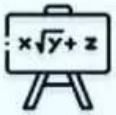
Data Engineer

- A **data engineer** specializes in preparing data for analytical usage. Data Engineering involves the development of platforms and architectures for data processing.
- Data Engineer is responsible for designing the format for data scientists and analysts to work on.
  - Build, test, and maintain dataset pipeline architectures
  - Create new data validation methods and data analysis tools
  - Combine raw information from different sources
  - Explore ways to enhance data quality and reliability

Data Scientist

uses statistics and machine learning to make predictions and answer key business questions

Skills - Math, Programming, Statistics

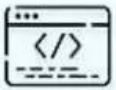


Tech - SQL, Python, R, Cloud

Data Engineer

build and optimize the systems that allow data scientists and analysts to perform their work

Skills - Programming, BigData & Cloud



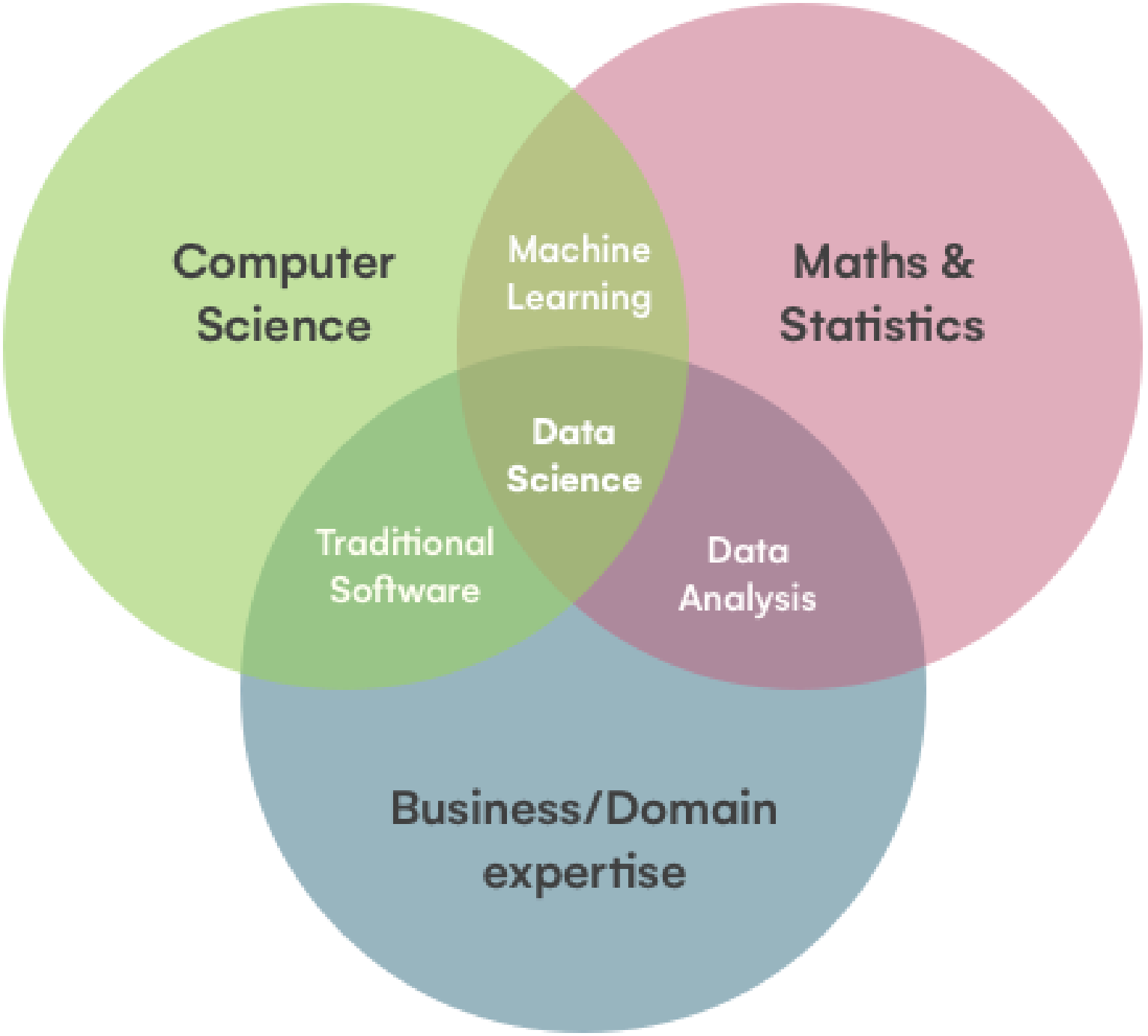
Tech - SQL, Python, Cloud, Distributed Computing

Data Analyst

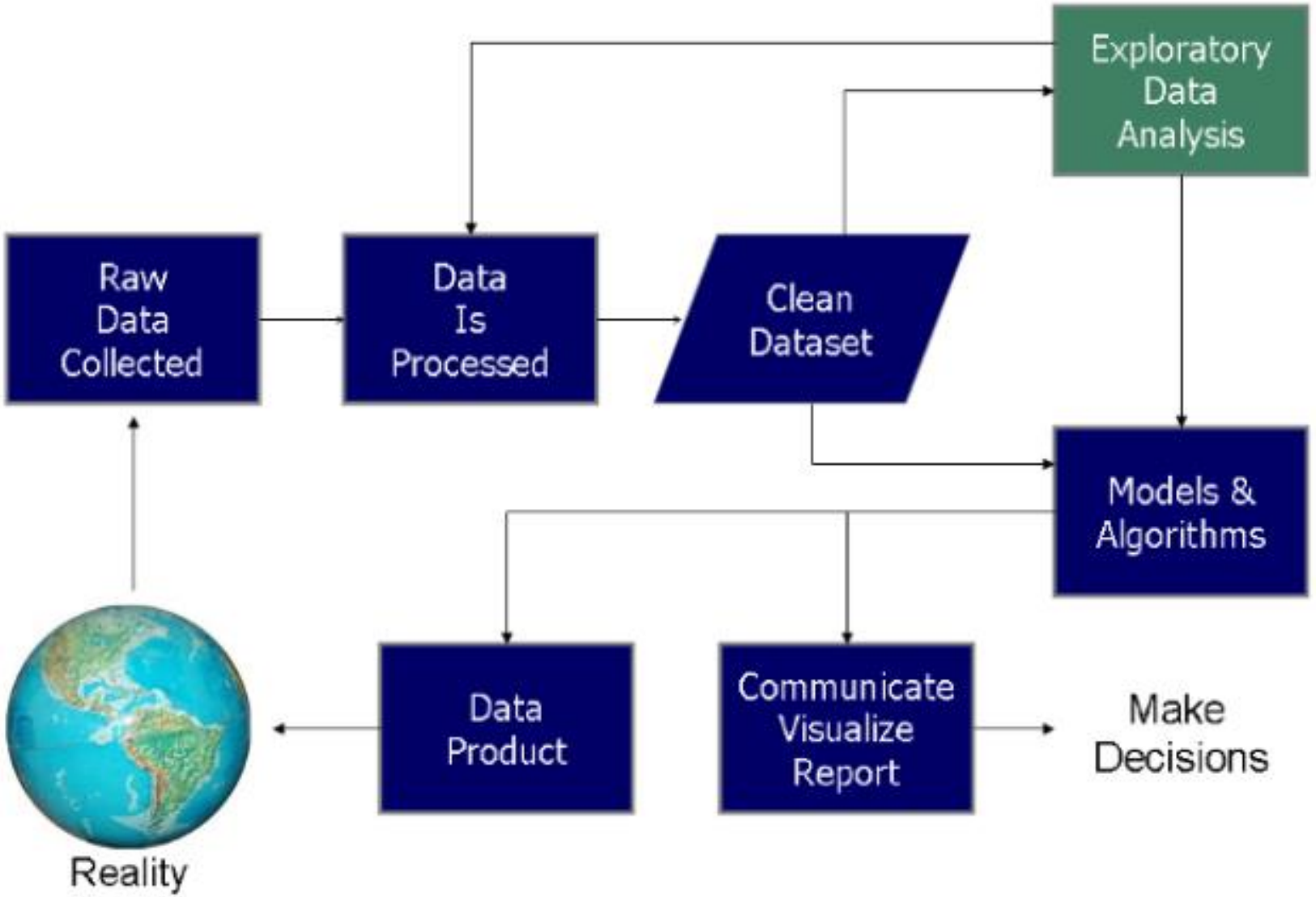
deliver value by taking data, communicating the results to help make business decisions

Skills - Communication, Business Knowledge

Tech - SQL, Excel, Tableau


Computer Science
Machine Learning
Maths & Statistics
Data Science
Traditional Software
Data Analysis
Business/Domain expertise

Data Analysis



Types of Data Analysis

- **Descriptive analysis** looks at past data and tells what happened. This is often used when tracking Key Performance Indicators (KPIs), revenue, sales leads, and more.
- **Exploratory analysis** is an approach of analyzing data to summarize their main characteristics, often using statistical graphics and other data visualization methods.
- **Predictive analysis** predicts what is likely to happen in the future. In this type of research, trends are derived from past data which are then used to form predictions about the future.
- **Prescriptive analysis** is the most advanced form of analysis, as it combines all your data and analytics, then outputs a model prescription: What action to take.

Exploratory Data Analysis

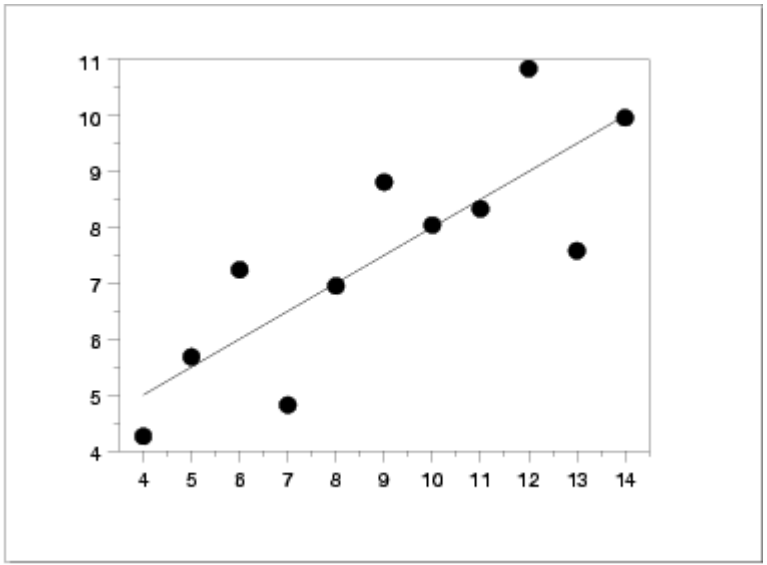
- Exploratory Data Analysis (EDA) is an approach for data analysis that employs a variety of techniques (mostly graphical) to
  - maximize insight into a data set
  - uncover underlying structure
  - extract important variables
  - detect outliers and anomalies
  - test underlying assumptions

Graphics and Exploratory Data Analysis

- Quantitative (summary)
- Graphical (plots)

X	Y
10.00	8.04
8.00	6.95
13.00	7.58
9.00	8.81
11.00	8.33
14.00	9.96
6.00	7.24
4.00	4.26
12.00	10.84
7.00	4.82
5.00	5.68

$N = 11$
Mean of $X = 9.0$
Mean of $Y = 7.5$
Intercept = 3
Slope = 0.5
Residual standard deviation = 1.237
Correlation = 0.816

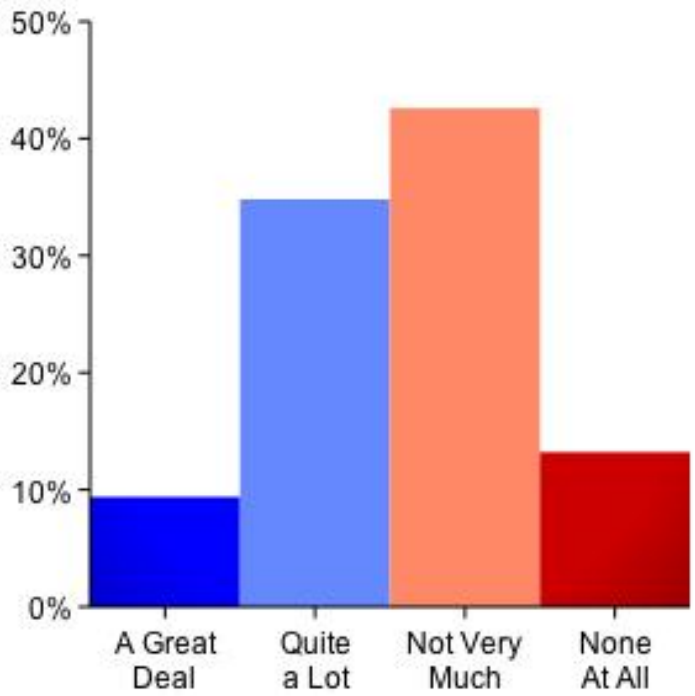


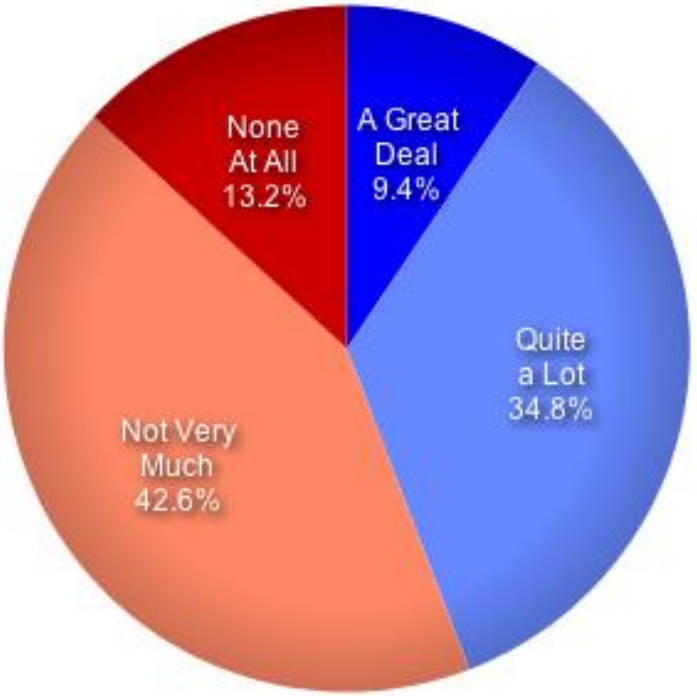
Univariate Data

- Univariate data are the ones which consist of **only one variable**.
- Univariate data analysis are straightforward as we are dealing with only one variable.
- The analysis that we do in case of univariate data analysis doesn't have to do anything with the relationships between variables.
- The main purpose is to describe the data and find patterns that are present.

Univariate Data

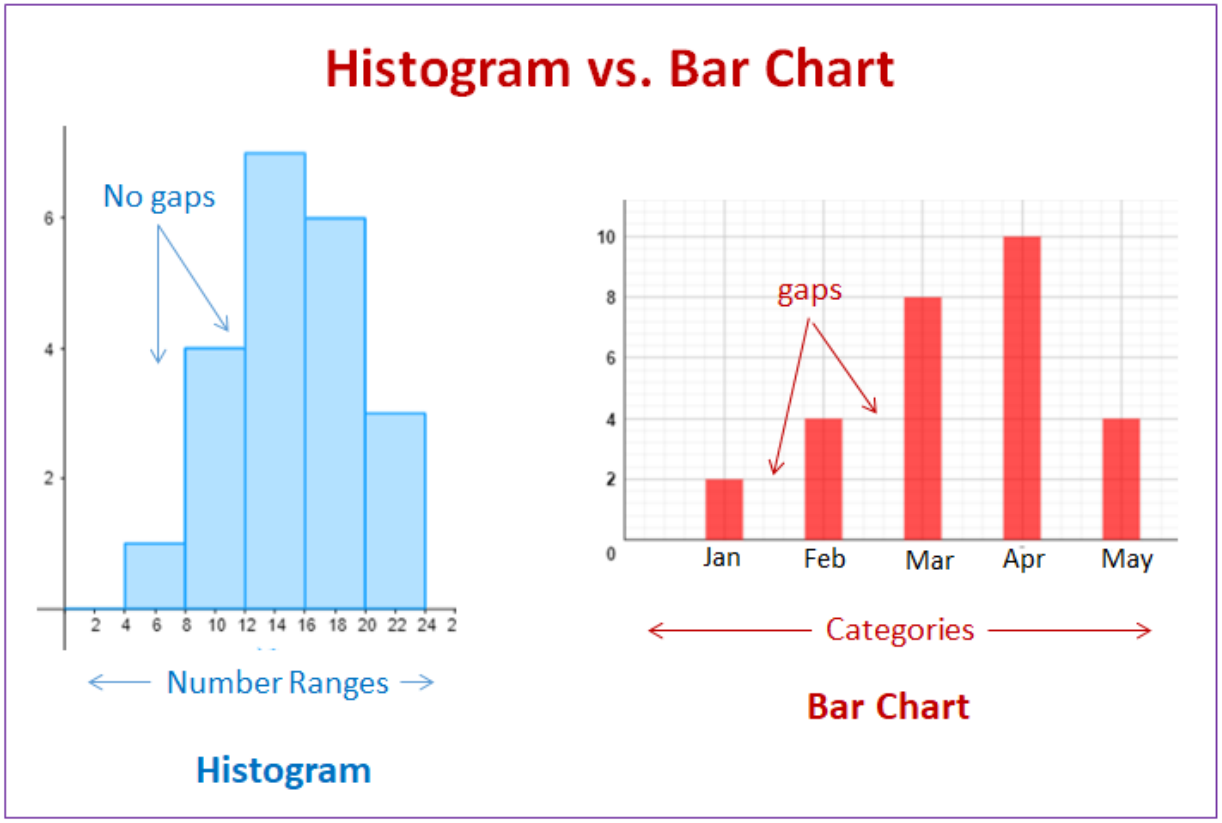
- Bar Chart
- Pie Chart
- Histogram / Frequency Distribution





Univariate Data

- Bar Chart
- Pie Chart
- Histogram / Frequency Distribution

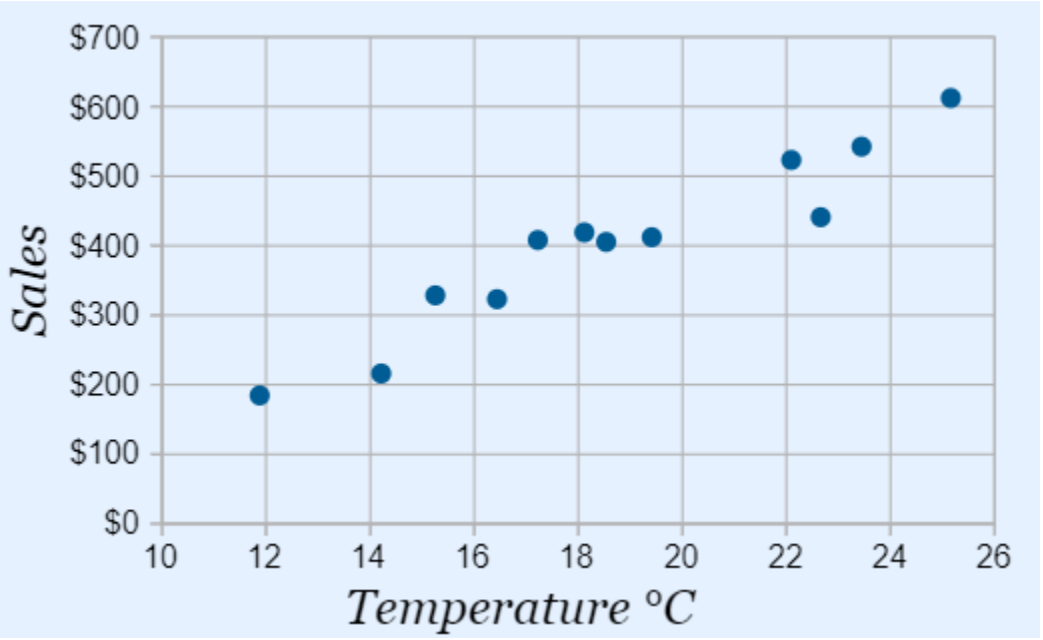


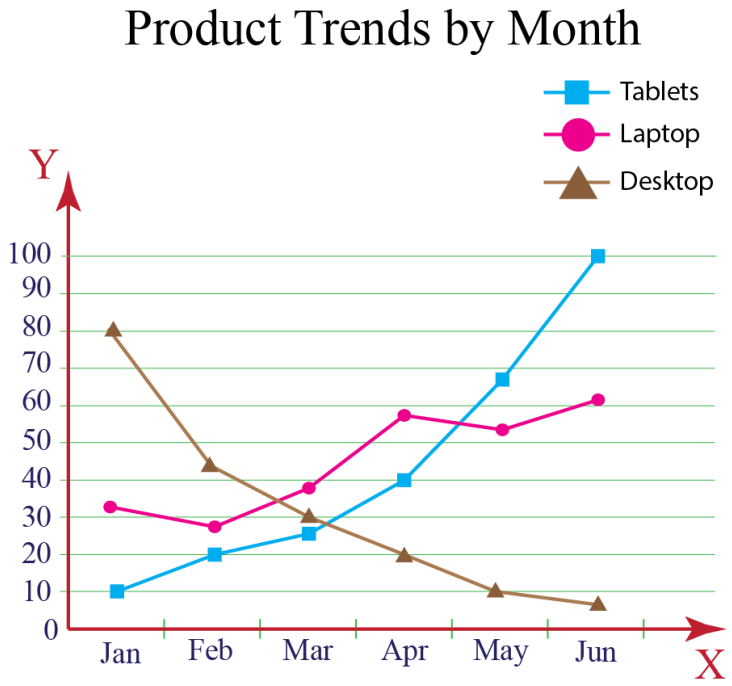
Bivariate Data

- Bivariate data involves **two different variables** where we are concerned about investigating the causes and relationship between those 2 variables.

Bivariate Data

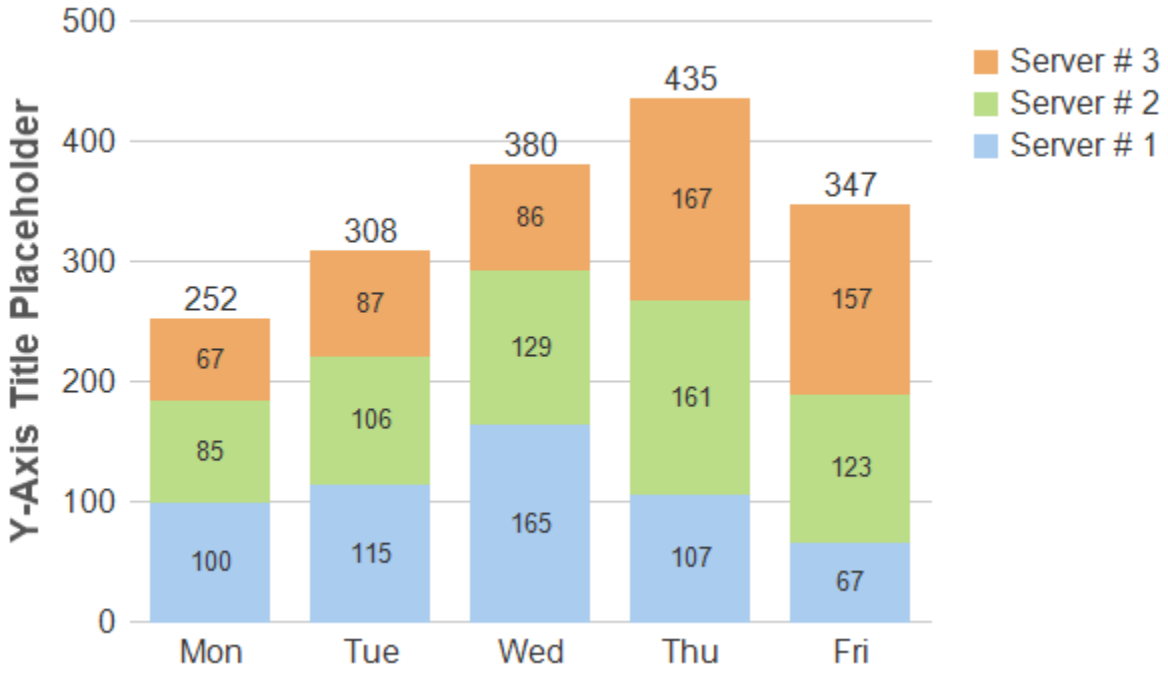
- Scatter Plot
- Line Plot
- Stacked Bar chart





Bivariate Data

- Scatter Plot
- Line Plot
- Stacked Bar chart

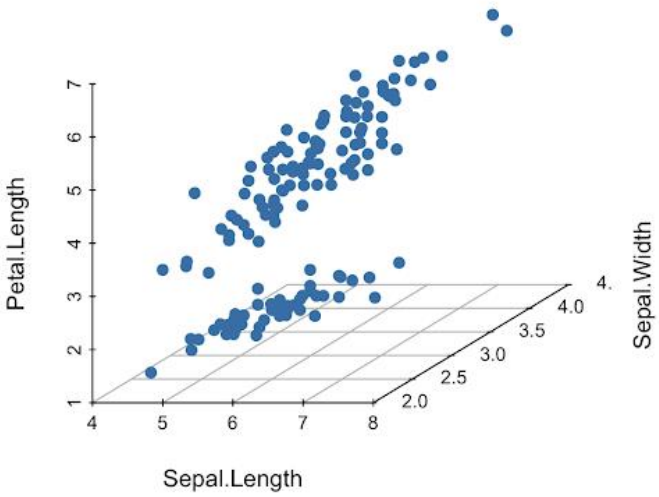


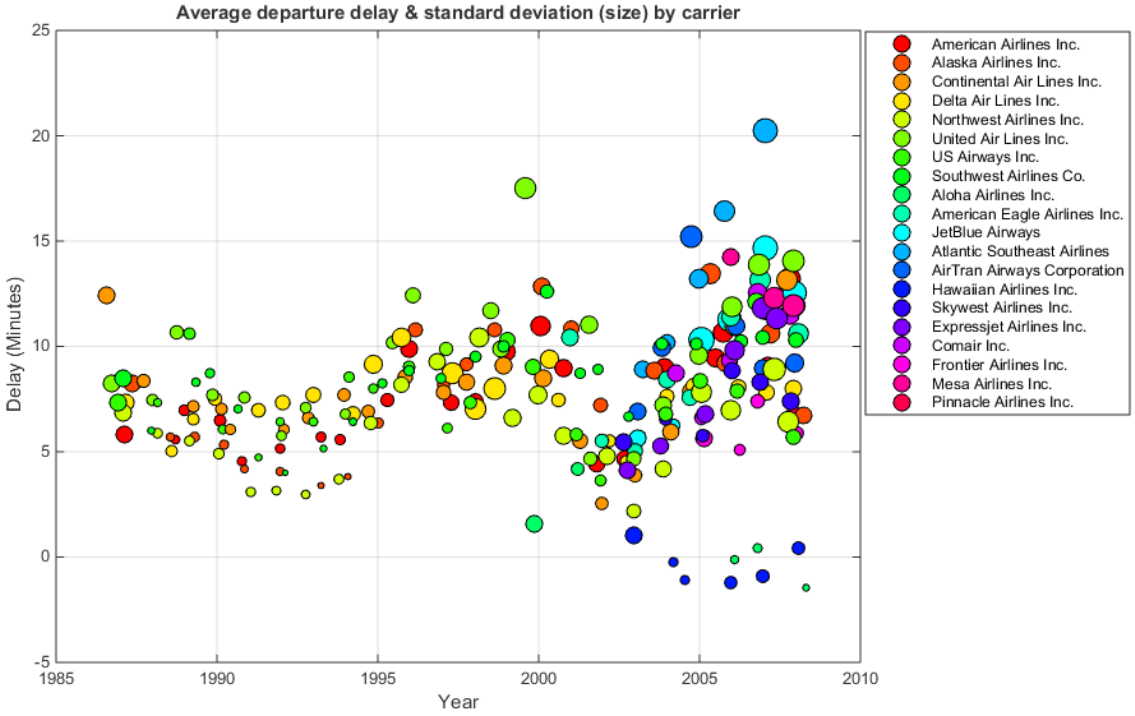
Multivariate data

- Data which **involves 3 or more variables** are termed as Multivariate data. These are similar to bivariate but contains more than one dependent variable.
- “Curse of dimension” is a trouble issue in information visualization.
- The effectiveness of retinal visual elements (e.g. color, shape, size) deteriorates when the number of variables increases

Multivariate data

- Enhanced Basic Plots





EDA using Python

- Matplotlib
- Seaborn

Summary

- Linear Regression
- Exploratory Data Analysis