IMS Health & Quintiles are now IQVIA™

Day - 1
Basic Concepts

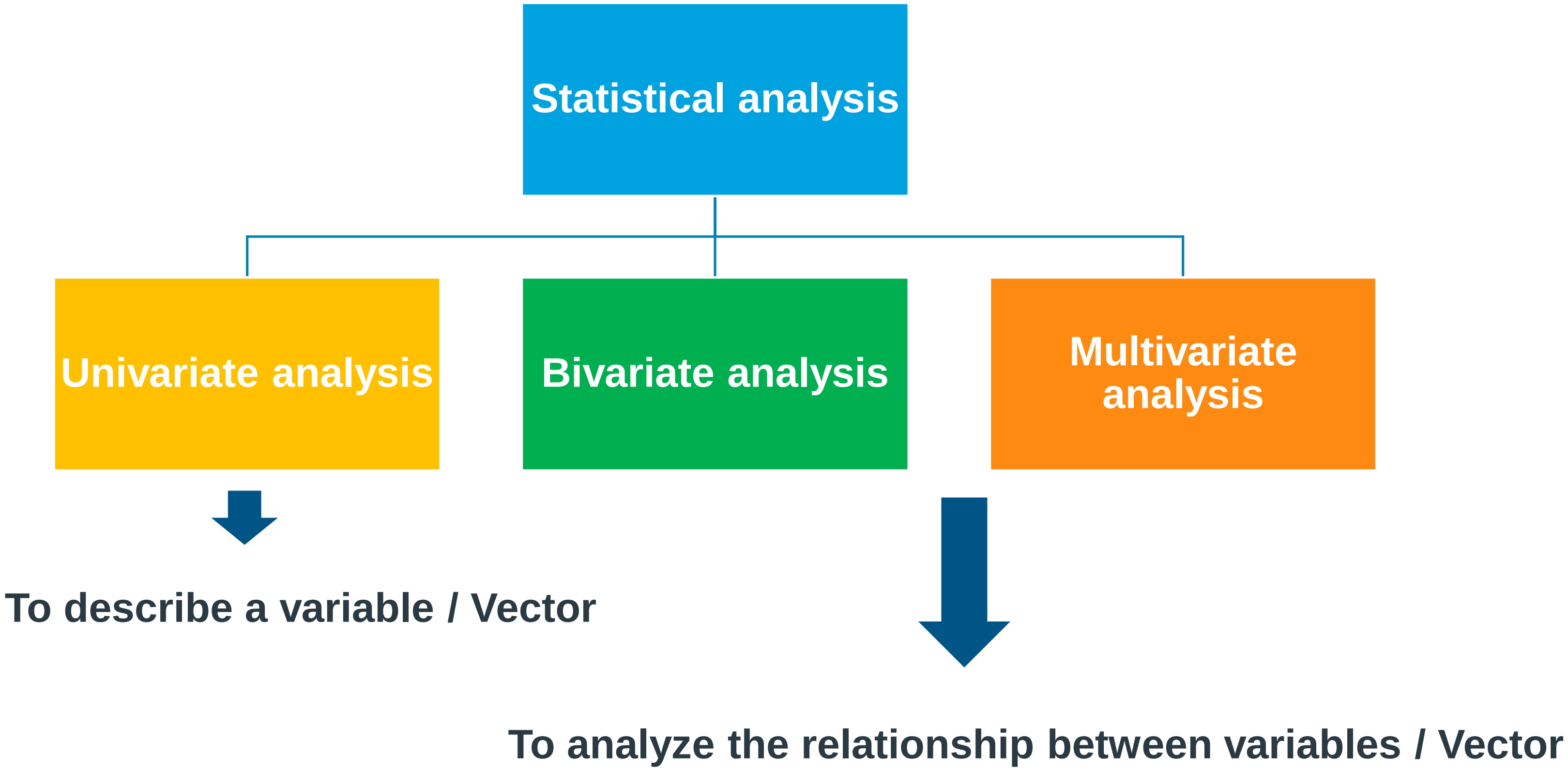
January 8, 2018

What Will I Learn?

- Fundamental of statistics with supporting case studies
- How to run basic statistical analysis in R

What is Statistics?

Statistics is a branch of mathematics dealing with the collection, **analysis**, interpretation, presentation, and organization of data



Statistical analysis

Univariate analysis

Bivariate analysis

Multivariate analysis

To describe a variable / Vector

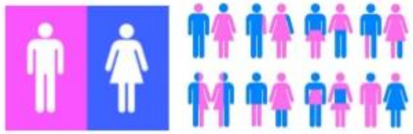
To analyze the relationship between variables / Vector

Types of variable

Variables can be broadly classified into two types

1). **Categorical** : Qualitative Variable. It can be further categorized as nominal, dichotomous and ordinal

- **Nominal Variable** - Have two or more categories, but no intrinsic order (e.g.) Race, Gender



- **Ordinal Variable** - Like nominal variable, but with intrinsic order (e.g.) Performance, Blood group

2). **Continuous** : Quantitative Variable (e.g.) Age, Income, years of experience etc.

Cause effect relationship

To understand cause – affect relationship let us start with an example

Example:

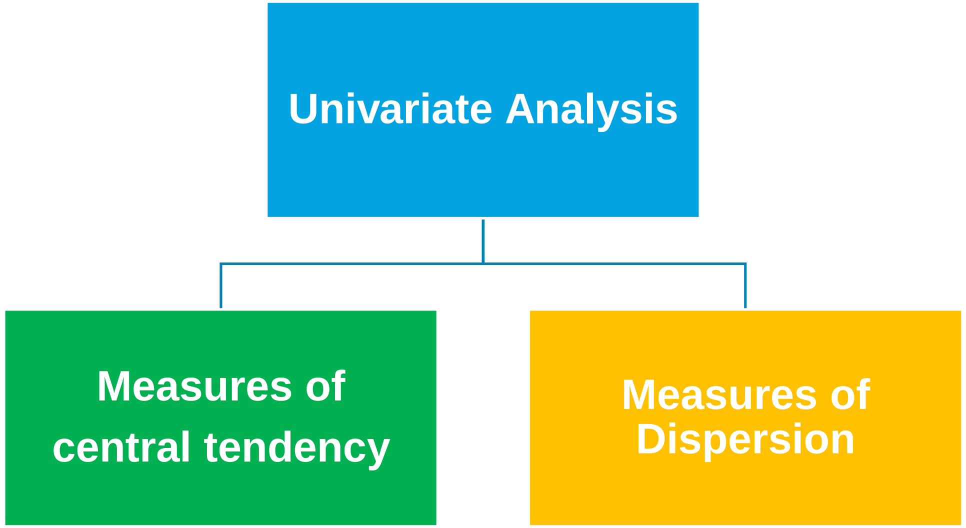
Assume that we are having a supermarket and we are offering 15% discount on the price of products (for particular month). What do we think is going to happen?

Cause : Reduction in price which – **Independent variable**

Effect : Increase in sales – **Dependent Variable**

Univariate analysis

Univariate analysis is the simplest way of analyzing data and it looks at one variable at a time



Univariate Analysis

Measures of central tendency

Measures of Dispersion

Measures of central tendency : (e.g.) sum, mean, median, mode, standard deviation

Measures of dispersion : (e.g.) variance, standard deviation range

Calculation - Measure of central tendency

Data Series : 17, 4, 2, 35, 36, 36, 4, 18, 2, 4, 2, 38

Sum = 17 + 4 + 2 + 35 + 36 + 36 + 4 + 18 + 2 + 4 + 2 + 38 = 198

Mean = (17 + 4 + 2 + 35 + 36 + 4 + 18 + 2 + 4 + 2 + 38) /11 = 16.5

Median = 38, 36, 36, 35, 18, 17, 4, 4, 4, 2, 2, 2 = (17+4) / 2 = 10.5

Mode = 2 , 4

Max = 38

Min = 2

Calculation – Measures of dispersion

S.No	Data Series	$X - \mu$	$(x - \mu)^2$
1	38	21.5	462.25
2	36	19.5	380.25
3	36	19.5	380.25
4	35	18.5	342.25
5	18	1.5	2.25
6	17	0.5	0.25
7	4	-12.5	156.25
8	4	-12.5	156.25
9	4	-12.5	156.25
10	2	-14.5	210.25
11	2	-14.5	210.25
12	2	-14.5	210.25
		Sum	**2,667.00**

Variance = σ^2 = $\sum(x-\mu)^2/N$

Standard Deviation = σ = $\sqrt{\ } \sigma^2$

x= *observation*

μ= *population mean*

N= *number of observations in the population*

Variance = (2,667.00 / 12) = 222.25

Standard deviation = $\sqrt{222.25}$ = 14.91

Case study - 1

We are trying to invest in the equity market. Shown below are the annual return summary for stocks A, B and C over period of 10 years.

Year	Stock A - Return	Stock B - Return	Stock C - Return
2007	55.0%	30.5%	35.0%
2008	18.3%	15.0%	18.3%
2009	2.1%	15.1%	2.1%
2010	47.0%	15.0%	47.0%
2011	34.4%	37.6%	34.4%
2012	23.9%	23.1%	25.0%
2013	3.2%	33.4%	3.2%
2014	34.0%	28.6%	32.0%
2015	1.3%	21.0%	23.0%
2016	27.5%	26.5%	27.5%

Mean	**24.7%**	**24.6%**	**24.8%**
Standard Deviation	18.8%	8.1%	14.0%

Case study - 1

We are trying to invest in the equity market. Shown below are the annual return summary for stocks A, B and C over period of 10 years.

Year	Stock A - Return	Stock B - Return	Stock C - Return
2007	55.0%	30.5%	35.0%
2008	18.3%	15.0%	18.3%
2009	**2.1%**	15.1%	**2.1%**
2010	47.0%	15.0%	47.0%
2011	34.4%	37.6%	34.4%
2012	23.9%	23.1%	25.0%
2013	**3.2%**	33.4%	**3.2%**
2014	34.0%	28.6%	32.0%
2015	**1.3%**	21.0%	23.0%
2016	27.5%	26.5%	27.5%
Mean	**24.7%**	**24.6%**	**24.8%**
Standard Deviation	18.8%	8.1%	14.0%

Case study - 1

We are trying to invest in the equity market. Shown below are the annual return summary for stocks A, B and C over period of 10 years.

Year	Stock A - Return	Stock B - Return	Stock C - Return
2007	55.0%	**30.5%**	35.0%
2008	18.3%	**15.0%**	18.3%
2009	2.1%	**15.1%**	2.1%
2010	47.0%	**15.0%**	47.0%
2011	34.4%	**37.6%**	34.4%
2012	23.9%	**23.1%**	25.0%
2013	3.2%	**33.4%**	3.2%
2014	34.0%	**28.6%**	32.0%
2015	1.3%	**21.0%**	23.0%
2016	27.5%	**26.5%**	27.5%
Mean	24.7%	24.6%	24.8%
Standard Deviation	**18.8%**	8.1%	**14.0%**

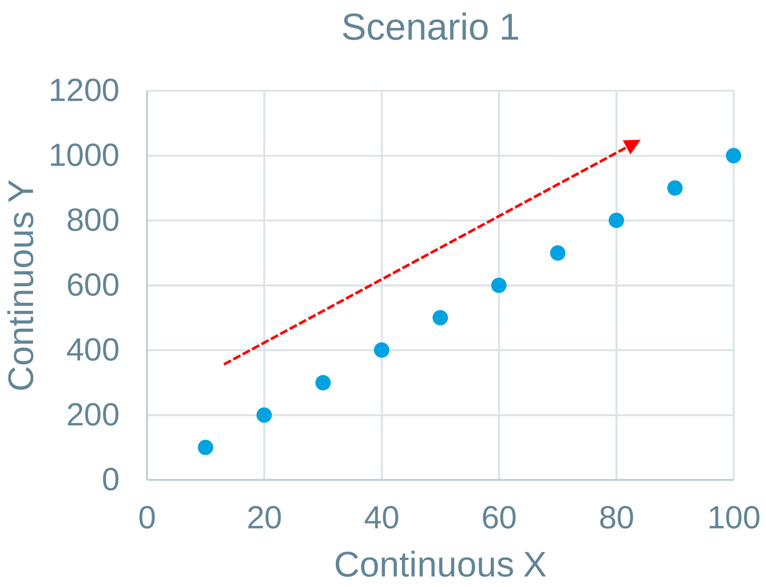
Stock B is my Choice. Yours?

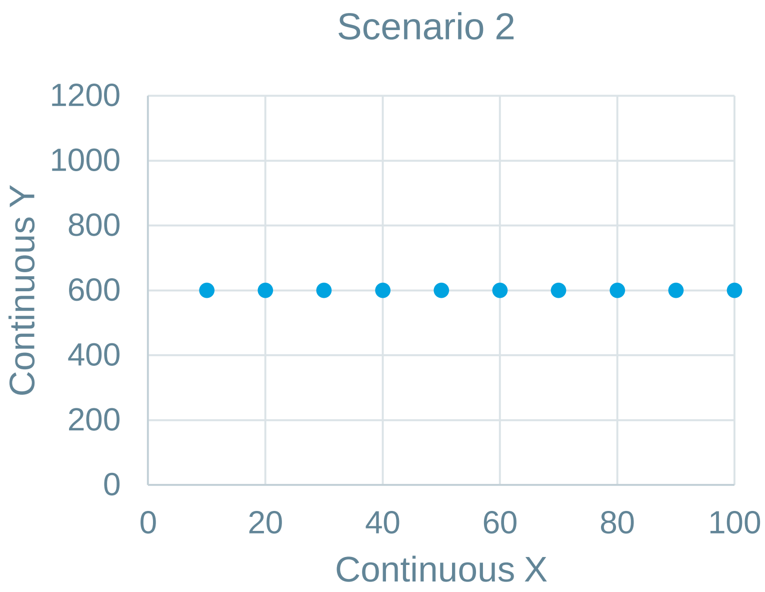
Bivariate analysis

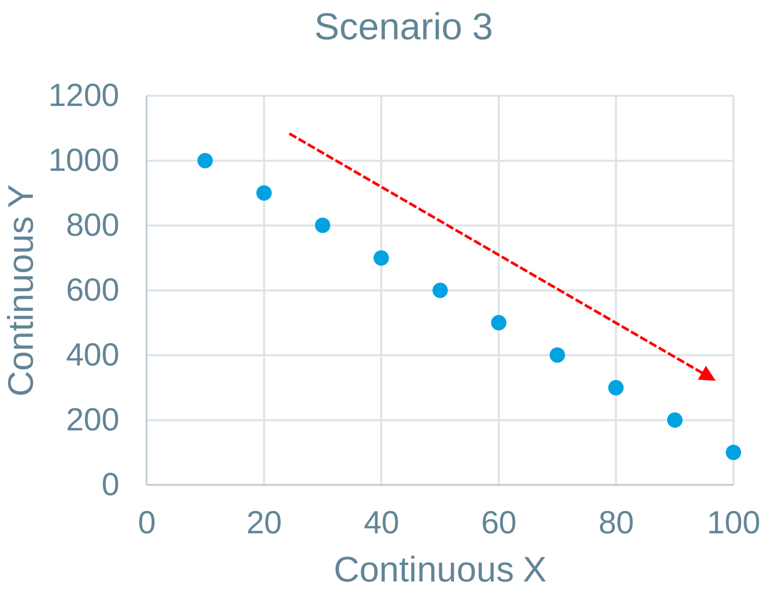
Bivariate analysis is one of the simplest forms of quantitative analysis. It involves the analysis of two variables (often denoted as *X*, *Y*) for the purpose of determining the empirical relationship between them

Variable X	Variable Y	Technique (graphs)
Continuous	Continuous	**Scatter plot**
Categorical	Categorical	**Multiple bar diagram**
Categorical	Continuous	**Box - Plot**
Continuous	Categorical	**Box - Plot**

Scatterplot





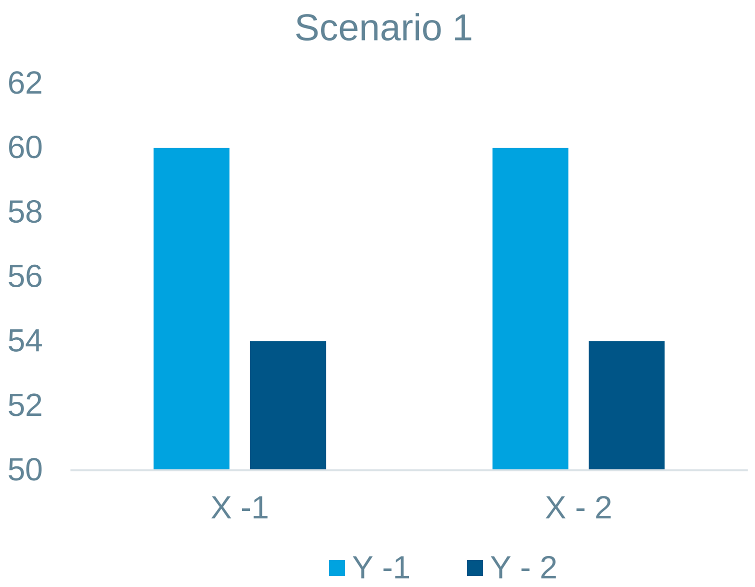


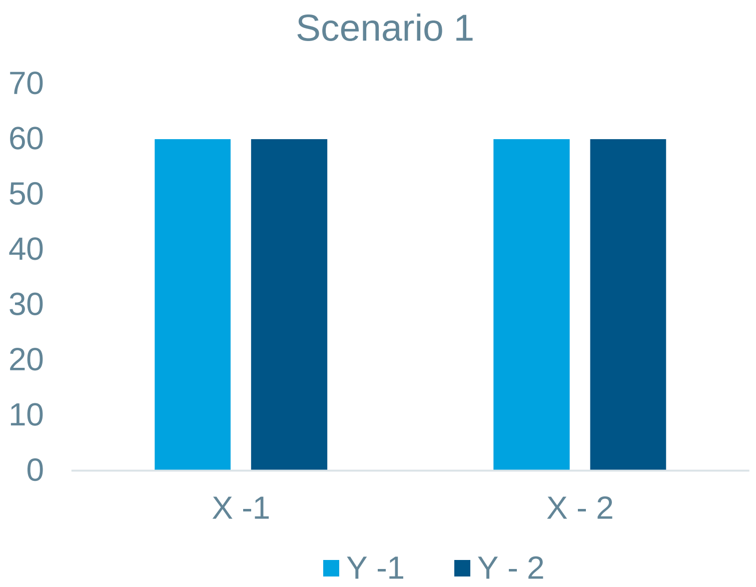
Scenario 1 : The relationship between the variable X any Y is linear and positive

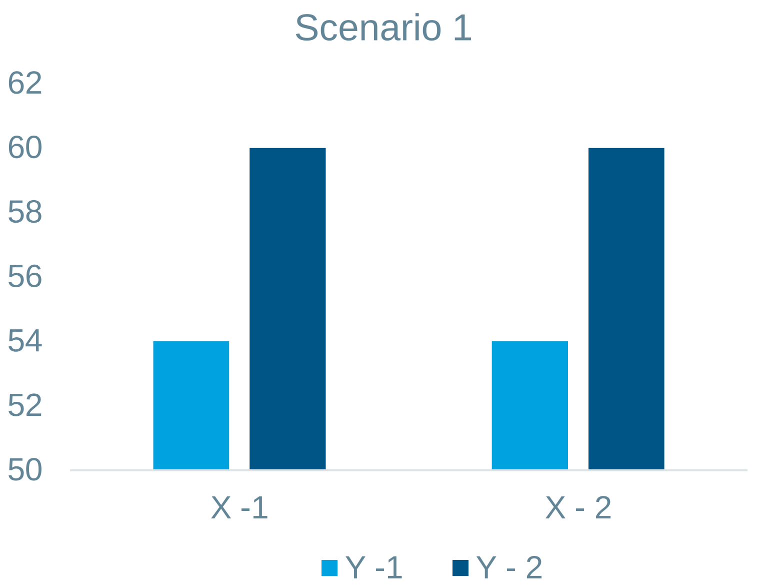
Scenario 2 : There is no relation ship between the variable X and Y

Scenario 1 : The relationship between the variable X any Y is linear and negative

Multiple bar diagram





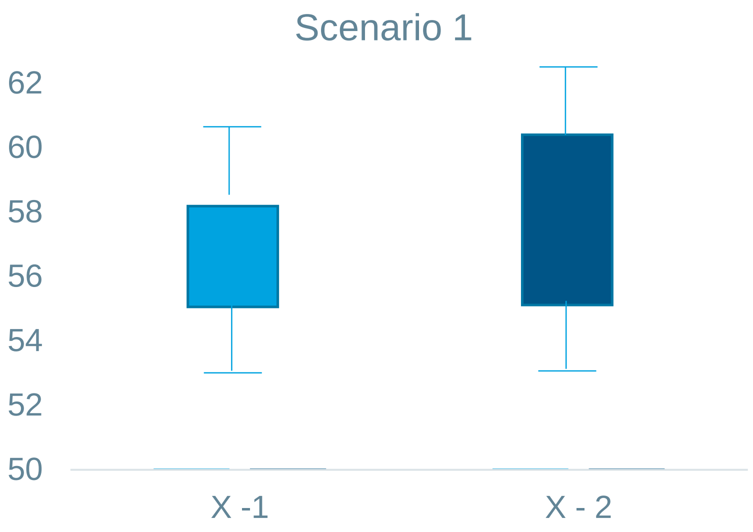


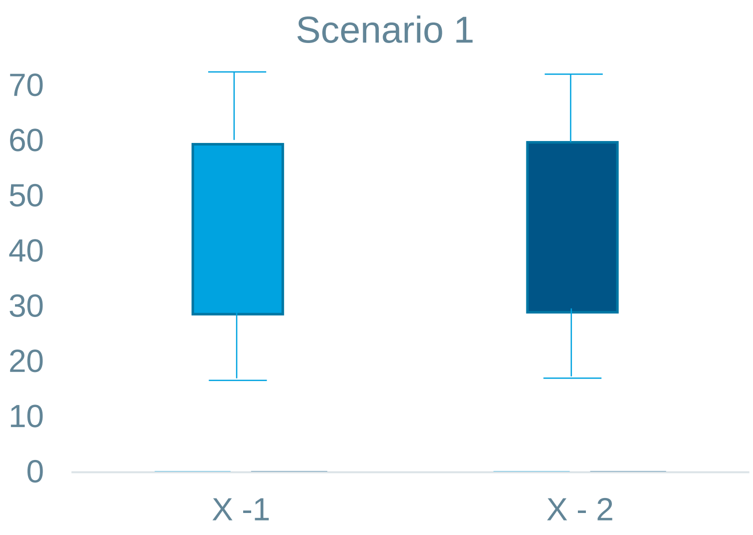
Scenario 1 : There is a relationship between two categories (Y-1 outperforms Y-2)

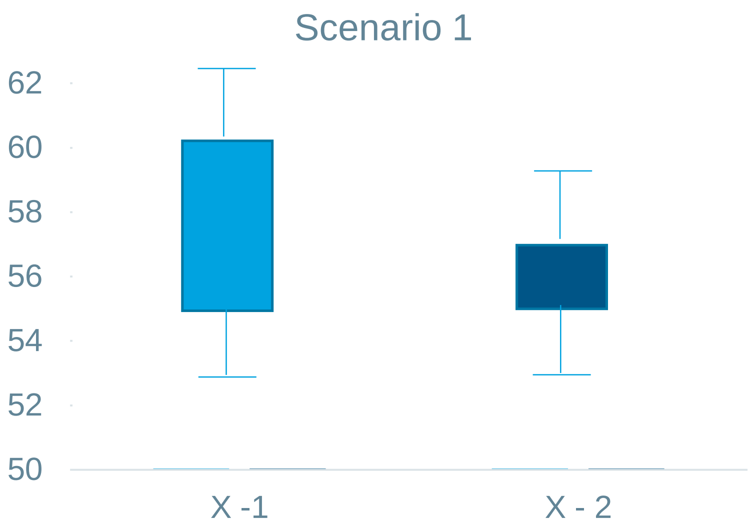
Scenario 2 : No relation ship between two categories

Scenario 1 : There is a relationship between two categories (Y-2 outperforms Y-1)

Box - Plot







Scenario 1 : X – 2 category is outperforming X – 1 category

Scenario 2 : No significant difference

Scenario 1 : X – 1 category is outperforming X – 2 category

Case study - 2

Using direct marketing dataset from a direct marketer, we have to identify and tell the factors that are influencing some customers to spend more than other customers

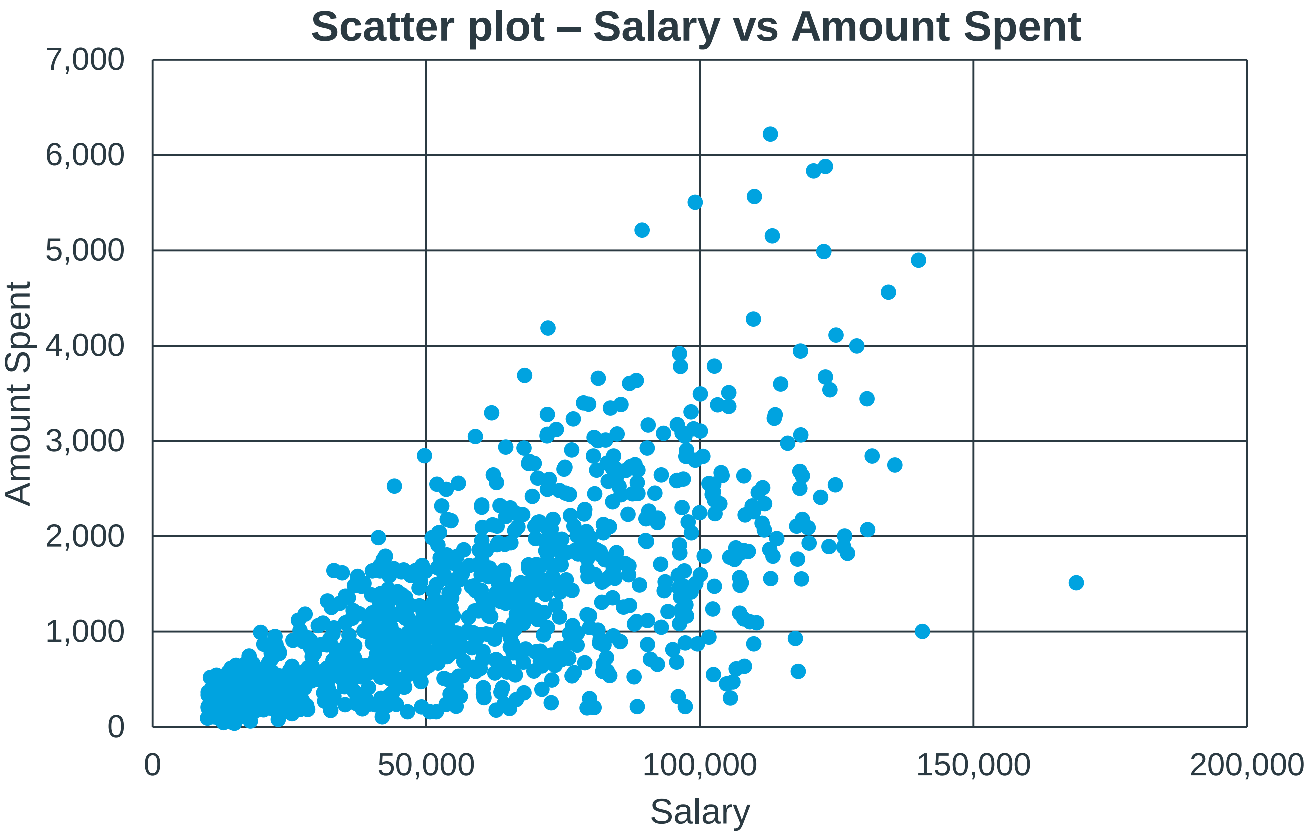
We have access to

- **Age**
- Gender
- **Own hose / rental house**
- **Marital Status**
- **Salary**
- How far the customer live away from the store that sells similar kind of products
- No of children they have
- Volume purchased
- **Amount Spent**

Dataset: Copyright © by Jigsaw academy

Is salary influencing customers to spent more?

- **Salary – Continuous Variable**
- **Amount Spent – Continuous Variable**

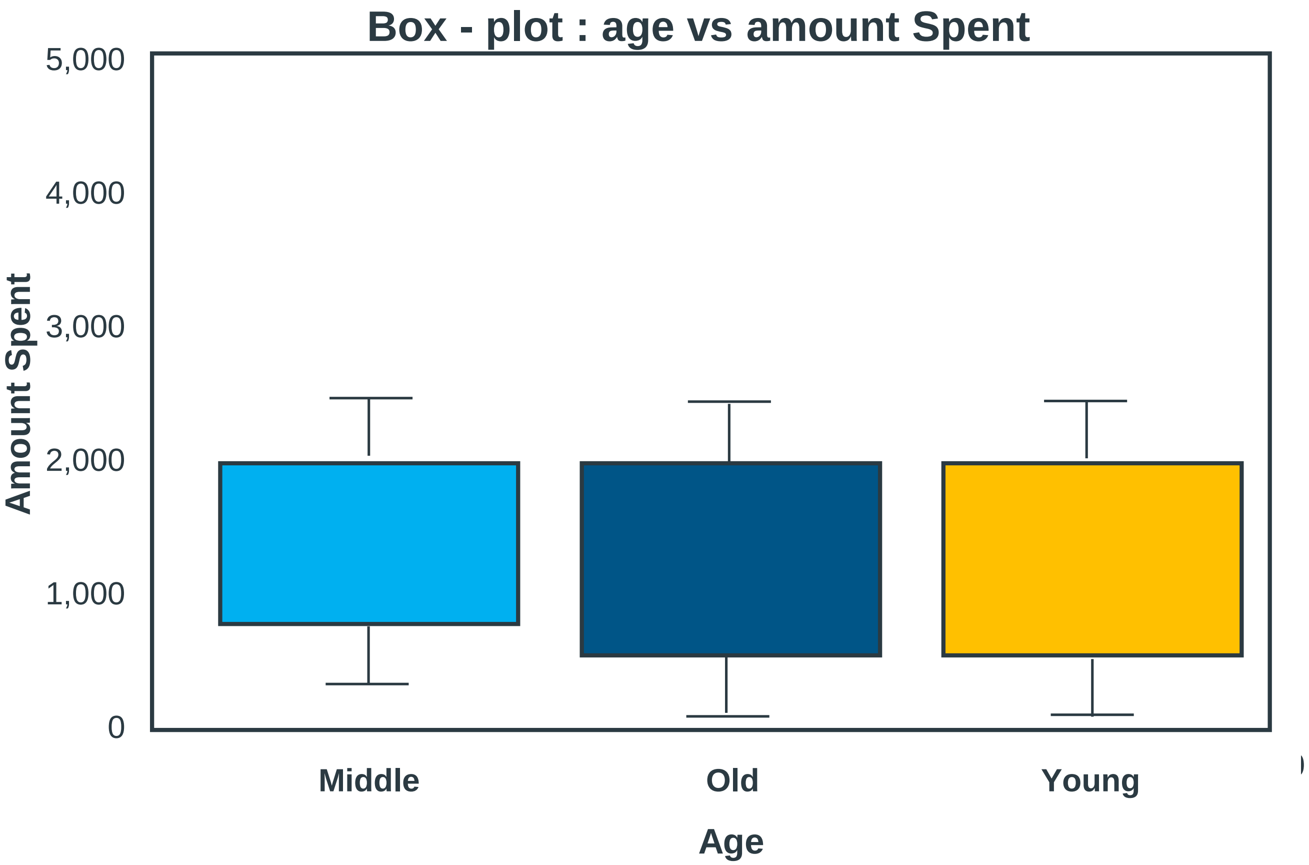


Inference :

- We could see that, as salary increases amount spent is also increasing.
- Which reveals that there is a positive between between two variables Salary and amount spent

Is there any association between age and amount spent?

- **Age – Categorical Variable**
- **Amount Spent – Continuous Variable**



Inference :

- We could see that, middle and old age customers are spending more
- Which reveals that there is a strong relationship between age and amount spent

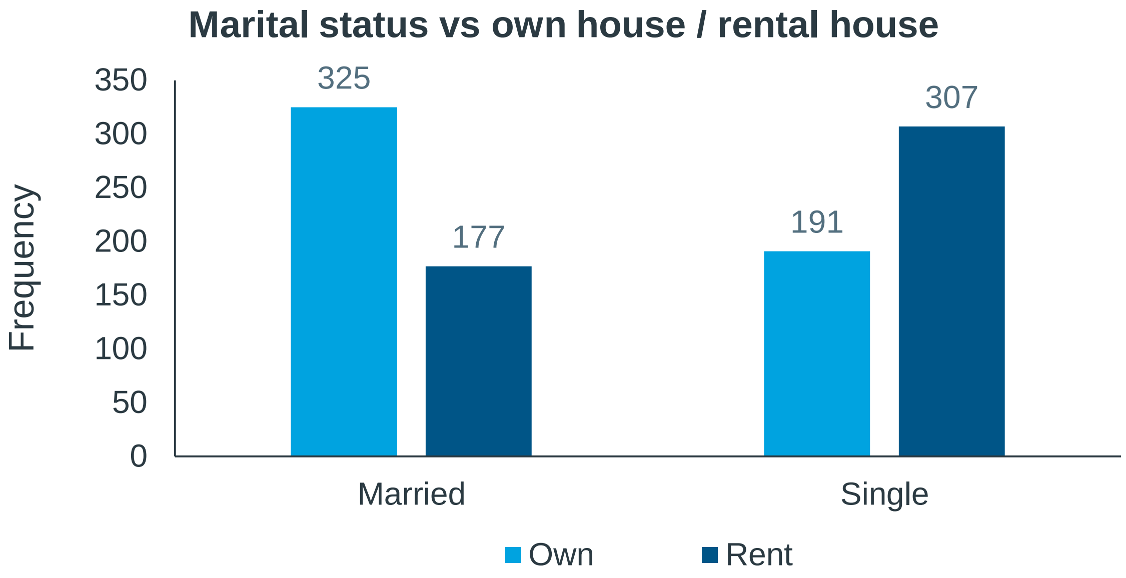
Is there any association between marital status and own hose / rental house?

- Marital Status – Categorical Variable
- Own house / rental house – Categorical Variable

Step 1 : Cross table / contingency table

	Own house	Rental house
Married	325	177
Single	191	307

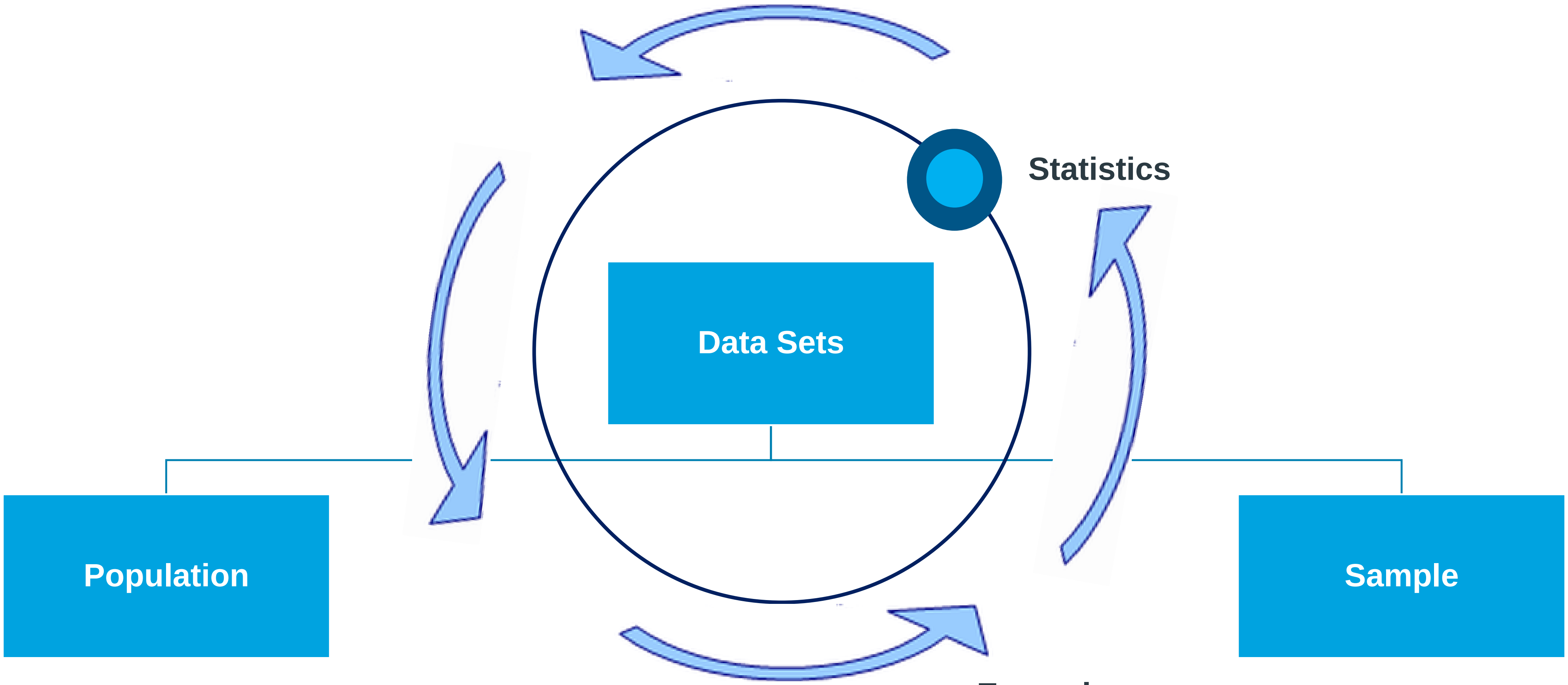
Step 2 : Multiple bar diagram



Inference :

- Most of the married customers are owning house when compared to single
- Therefore there is a strong association between marital status and own house / rental house

Populations vs samples



Examples :

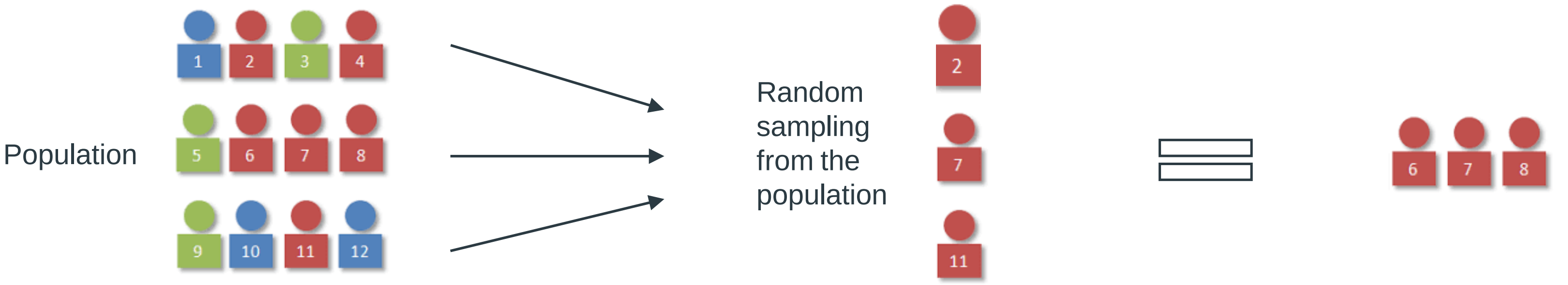
- All the students of a school
- All the customers of a bank

Examples :

- All the class toppers of a school
- All the customers of a bank (with credit card)

How to select a sample?

Simple random sampling :

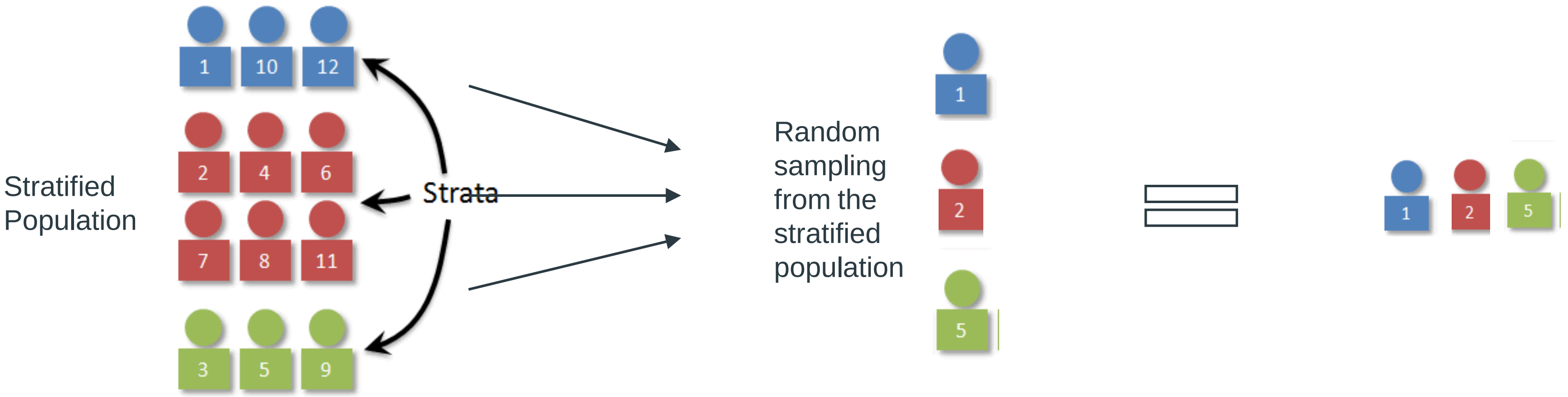


Population

Random sampling from the population

Possibility for biasness

Stratified random sampling :



Stratified Population

Strata

Random sampling from the stratified population

Less possibility for biasness

In analytics most of the time we use sample to make inference about the population

Practical Session