Introduction to Data Analysis

UNIVERSIDADE PORTUCALENSE
DEPARTAMENTO CIÊNCIA E TECNOLOGIA

Informatic Engineering | 2024/2025
Data Analysis Lab

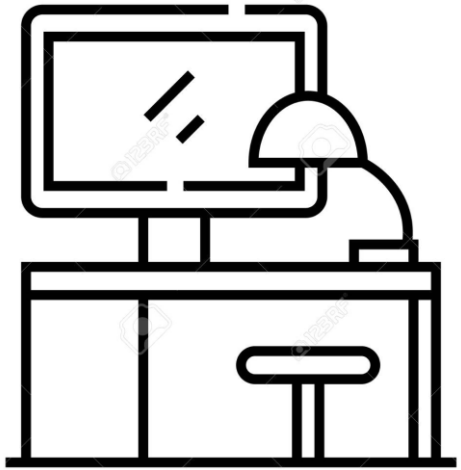
Fátima Leal

Previous Lesson

- Context
- Big Data Definition
- Problems
- Challenges
- Solutions
- Technologies

Outline

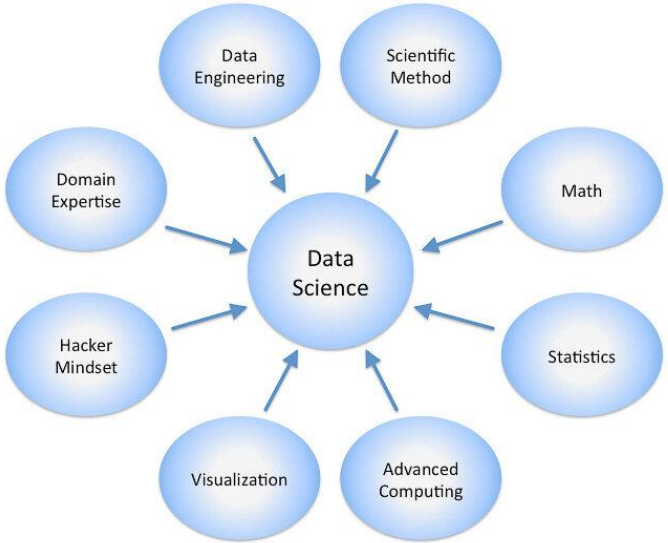
- Introduction to data analysis
- Statistical measures
- Graph Analysis




1
Define why you need data analysis
2
Begin collecting data from sources
3
Clean through unnecessary data
4
Begin analyzing the data
5
Interpret results and apply them

What Is data Science

- Interdisciplinary field that uses scientific methods, processes and algorithms to extract knowledge
- Structured or unstructured data



What is data?

- Data are a large set of bits encoded to represent numbers, texts, images, sounds, videos, etc.
- Without data analysis, data is meaningless.
- When we add information, giving a meaning to them, these data become knowledge.

What is data?

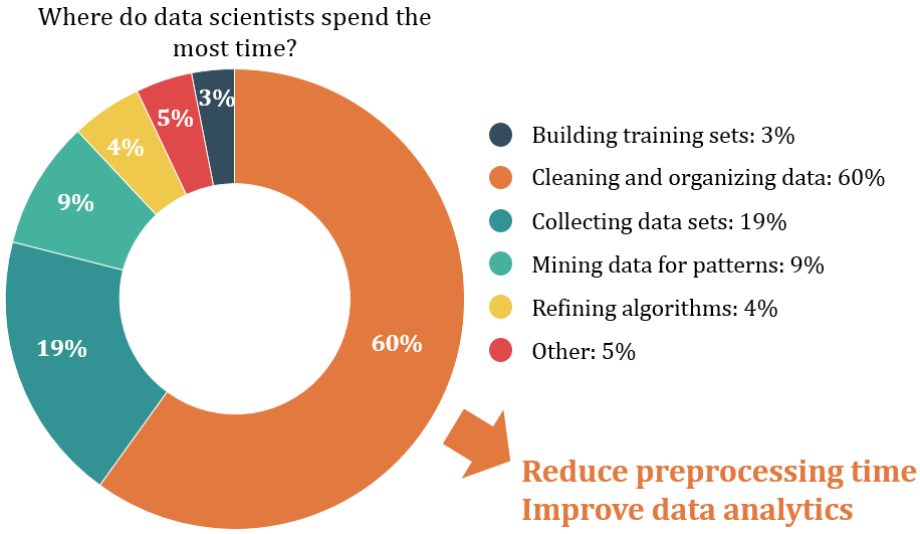
- Attributes
- Features

Contact	Age	Educational level	Company
Andrew	55	1.0	Good
Bernhard	43	2.0	Good
Carolina	37	5.0	Bad
Dennis	82	3.0	Good
Eve	23	3.2	Bad
Fred	46	5.0	Good
Gwyneth	38	4.2	Bad
Hayden	50	4.0	Bad
Irene	29	4.5	Bad
James	42	4.1	Good
Kevin	35	4.5	Bad
Lea	38	2.5	Good
Marcus	31	4.8	Bad
Nigel	71	2.3	Good

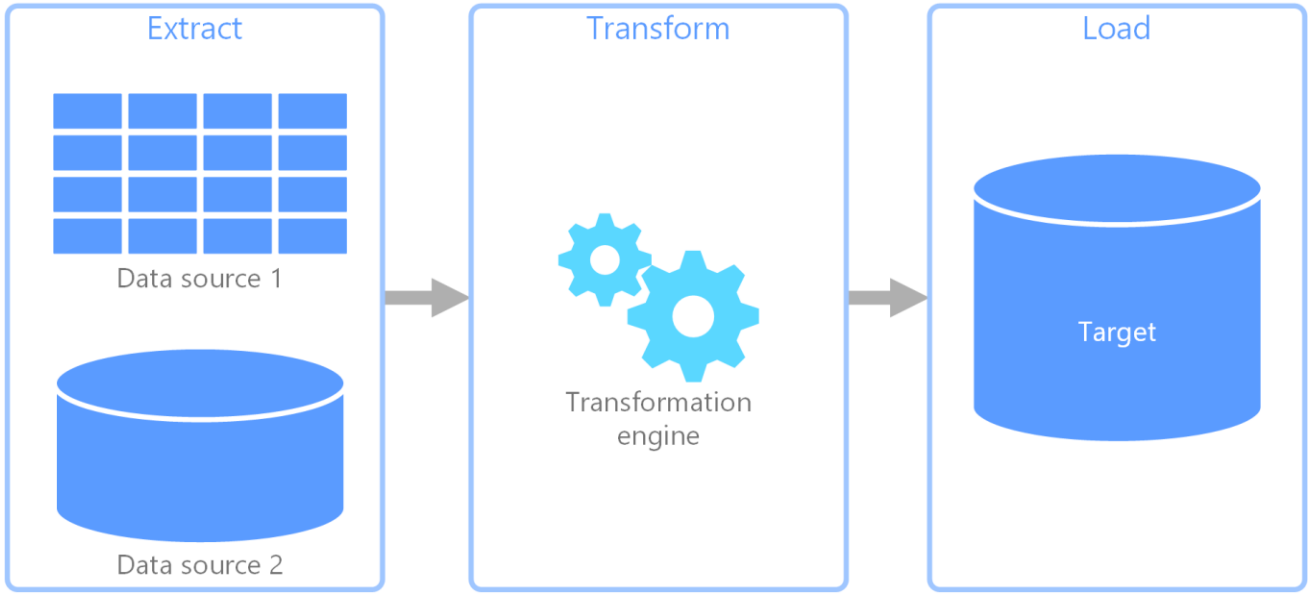
- Sometimes, the data is unstructured and with noise

Introduction to data analysis

- Steps for data analysis:
- Data collection
- Organization
- Pre-processing
- Transformation
- Modelling
- Interpretation



Extract, Transform and Load (ETL)



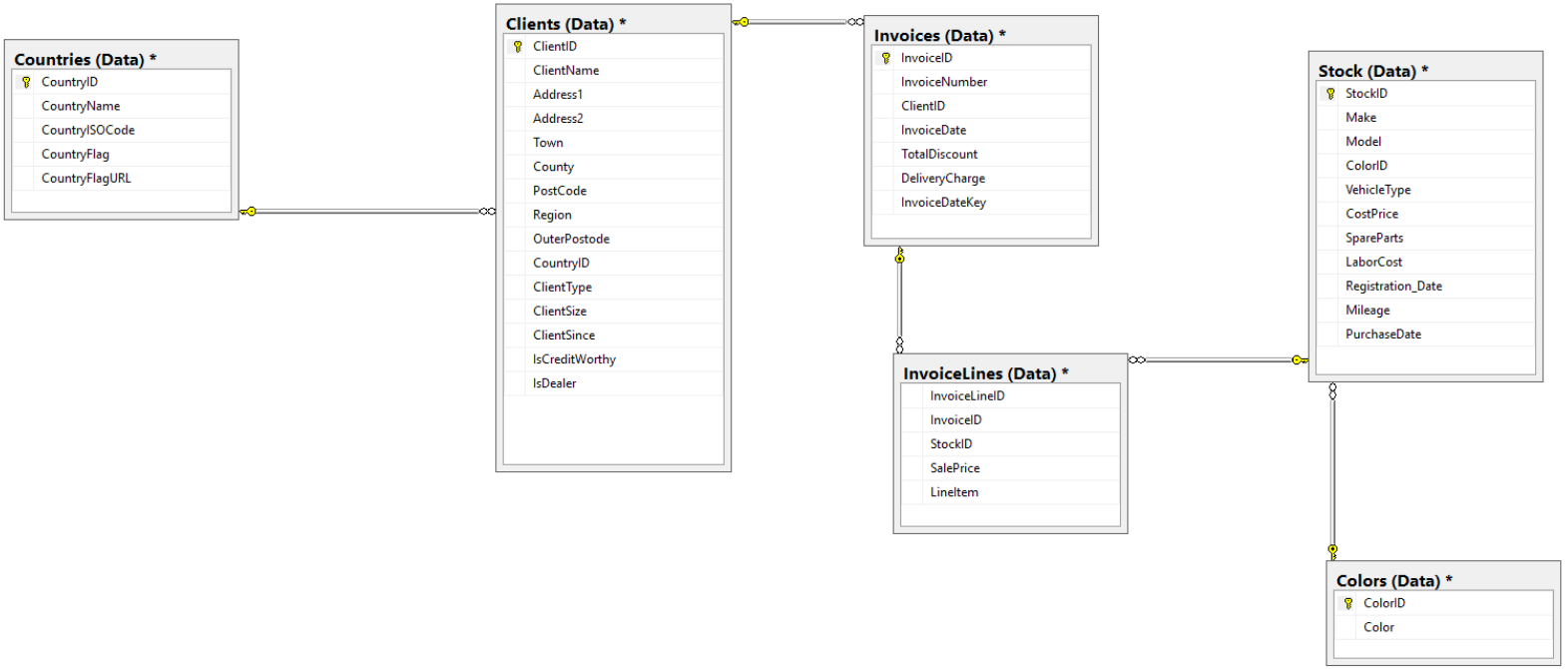
Extract, Transform and Load (ETL)

- **Extract** - ETL identifies the data which can come from structured and unstructured sources, including documents, emails, business applications, databases, equipment, sensors, third parties, and more

- **Transform** - Because the extracted data is raw in its original form, it needs to be mapped and transformed to prepare it for the eventual datastore. In the transformation process, ETL validates, authenticates, deduplicates, and/or aggregates the data in ways that make the resulting data reliable and queryable

- **Load** - This step can entail the initial loading of all the source data, or it can be the loading of incremental changes in the source data. You can load the data in real time or in scheduled batches.

Data Models

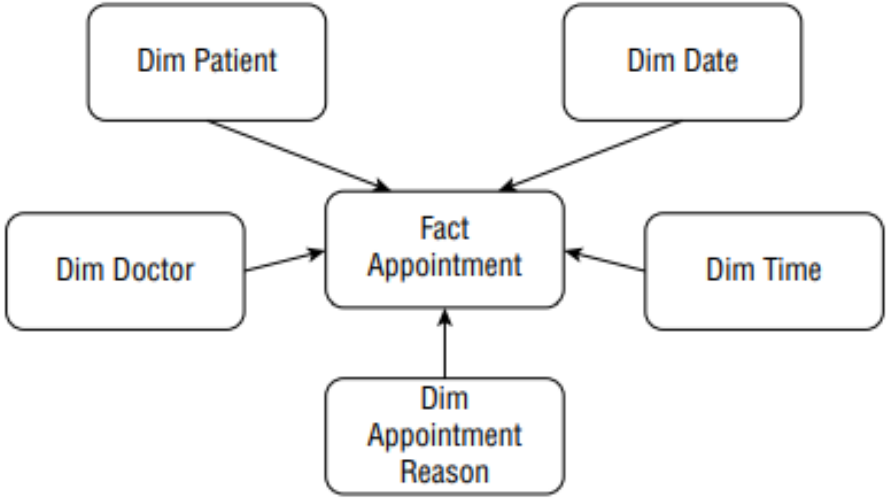

Countries (Data) *
CountryID
CountryName
CountryISOCode
CountryFlag
CountryFlagURL
Clients (Data) *
ClientID
ClientName
Address1
Address2
Town
County
PostCode
Region
OuterPostode
CountryID
ClientType
ClientSize
ClientSince
IsCreditWorthy
IsDealer
Invoices (Data) *
InvoiceID
InvoiceNumber
ClientID
InvoiceDate
TotalDiscount
DeliveryCharge
InvoiceDateKey
InvoiceLines (Data) *
InvoiceLineID
InvoiceID
StockID
SalePrice
LineItem
Stock (Data) *
StockID
Make
Model
ColorID
VehicleType
CostPrice
SpareParts
LaborCost
Registration_Date
Mileage
PurchaseDate
Colors (Data) *
ColorID
Color

Dimensional Data Warehouses

- Fact Tables – contains the “metadata” of an entity, as well as any measures (which are usually numeric values) you want to track and later summarize

- Dimensional Tables - A dimension is property of that entity you can group or “slice and dice” the fact records by, and a dimension table will contain further information of that property.

Dimensional Data Warehouses



Example

- Consider this Fact Table

Order ID	Order Date	Customer	Customer	Country	Book ID	Book Title	Genre	Author	Price per Unit	Quantity	Total Amou	Payment Method	Shipping Type
5001	01/01/2023	C001	John Doe	USA	B101	Data Scien	Technologi	Dr. Smith	30.00	1	30.00	Credit Card	Standard
5002	05/01/2023	C002	Alice Smith	UK	B102	Python Pro	Technologi	Dr. Brown	25.00	2	50.00	PayPal	Express
5003	10/02/2023	C003	Carlos Per	Spain	B103	Artificial In	Technologi	Dr. White	40.00	1	40.00	Credit Card	Standard
5004	15/03/2023	C004	Emily Whit	Canada	B104	History of F	History	Dr. Black	20.00	1	20.00	Debit Card	Standard
5005	20/04/2023	C005	Lucas Silva	Brazil	B105	Philosophy	Philosophy	Dr. Green	15.00	3	45.00	PayPal	Express
5006	25/05/2023	C006	Sophia Lee	Australia	B106	Machine Le	Technologi	Dr. Grey	50.00	2	100.00	Credit Card	Standard
5007	12/06/2023	C007	David Kim	South Kore	B107	World War	History	Dr. Adams	35.00	1	35.00	Bank Transfer	Express
5008	07/07/2023	C008	Emma Bro	Germany	B108	Shakespea	Literature	Dr. Thomp	45.00	1	45.00	PayPal	Standard
5009	18/08/2023	C009	Oliver Jone	France	B109	Cybersecu	Technologi	Dr. Nelson	28.00	3	84.00	Debit Card	Express
5010	23/09/2023	C010	Mia Garcia	Mexico	B110	Modern Ec	Economy	Dr. Carter	33.00	2	66.00	Credit Card	Standard

- Create the data model for this data

Possible Analysis with This Model:

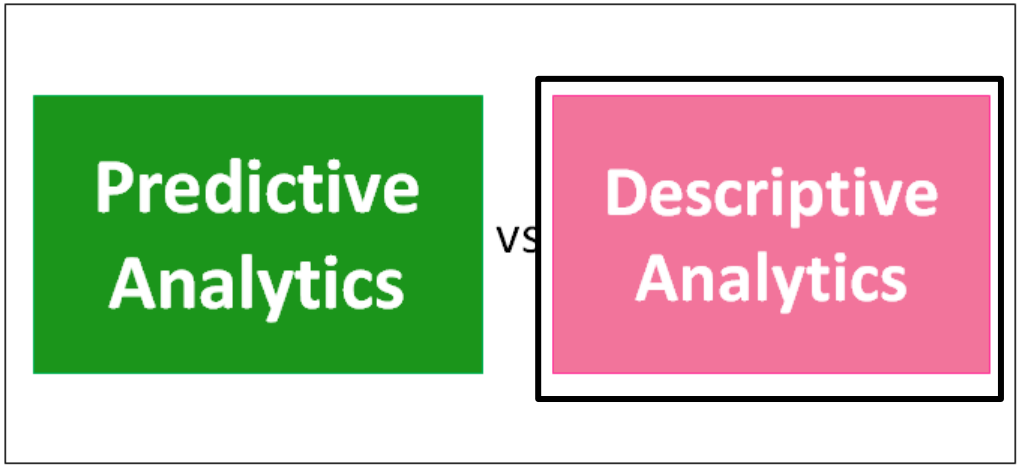
- What is the total sales per country or customer?
- Which book genres are the most sold?
- What is the most commonly used payment method?
- How do sales vary over time?

Introduction to data analysis

- What is the purpose?
    - Answer the research questions and to help determine trends and relationships among variables
- Before data collection, the researcher should accomplish:
    - How to process the data
    - Methods
    - Goals
- After Data collection
    - Process data
    - Prepare tables and graphs
    - Interpret findings

Introduction to data analysis

- **Descriptive analytics**: summarize or condense data to extract patterns
- **Predictive analytics**: extract models from data to be used for future predictions.



Introduction to data analysis

- Descriptive analytics applies algorithms to the data where the result can be statistic
- **Method or technique** is a systematic procedure that allows us to achieve an intended goal.
- **An algorithm** is a self-contained, step-by-step set of instructions easily understandable by humans, allowing the implementation of a given method.
- **A model** in data analytics is a generalisation obtained from data that can be used afterwards to generate predictions.

Descriptive analysis

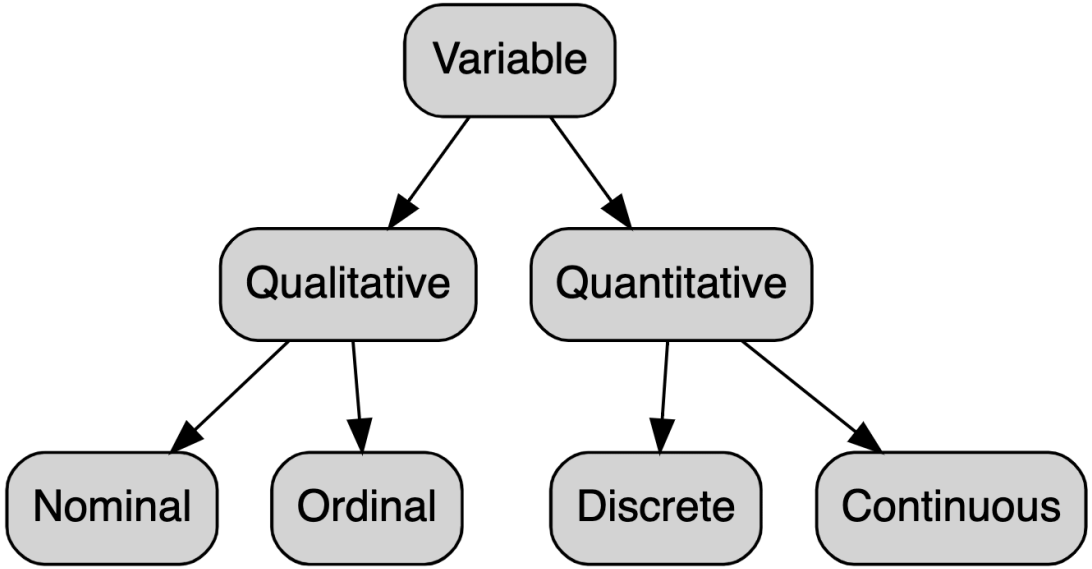
- Descriptive analysis refers to the description of the dataset
- Summarise and describe data:
    - Attribute categorization
    - Data types (numeric, textual, photos, videos, etc.)
    - For numeric properties it is important:
    - Histograms, box plots, and descriptive statistics are useful for understanding characteristics of the individual data attributes.

Descriptive analysis

- Generalised form of a Data Table

		Variables				
Observations		x_1	x_2	x_3	...	x_p
	o_1	x_{11}	x_{12}	x_{13}	...	x_{1p}
	o_2	x_{21}	x_{22}	x_{23}	...	x_{2p}
	o_3	x_{31}	x_{32}	x_{33}	...	x_{3p}
	...	...	...	...	...	...
	o_n	x_{n1}	x_{n2}	x_{n3}	...	x_{np}

Describing Data: Types of variables



Describing Data: Types of variables

- **Discrete variables** - numeric variables that have a countable number of values between any two values

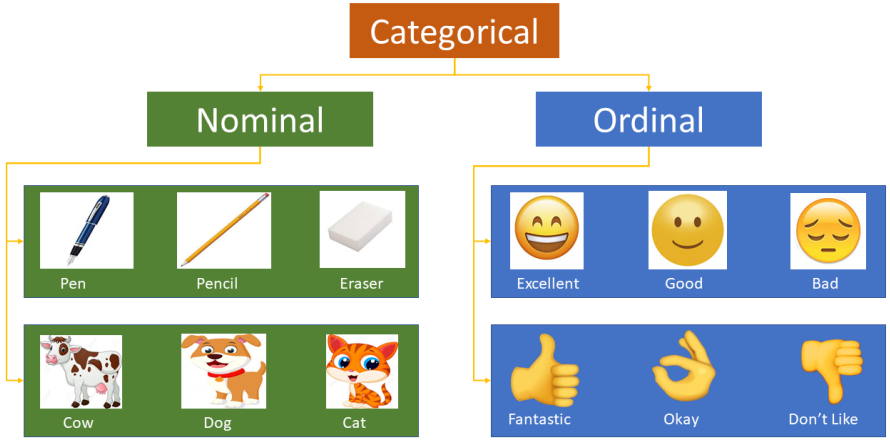
- **Continuous variables** - numeric variables that have an infinite number of values between any two values.

car name	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin
chevrolet chevelle malibu	18	8	307	130	3504	12	70	American
buick skylark 320	15	8	350	165	3693	11.5	70	American
plymouth satellite	18	8	318	150	3436	11	70	American
amc rebel sst	16	8	304	150	3433	12	70	American
ford torino	17	8	302	140	3449	10.5	70	American
ford galaxie 500	15	8	429	198	4341	10	70	American
chevrolet impala	14	8	454	220	4354	9	70	American
plymouth fury iii	14	8	440	215	4312	8.5	70	American
pontiac catalina	14	8	455	225	4425	10	70	American
amc ambassador dpl	15	8	390	190	3850	8.5	70	American
dodge challenger se	15	8	383	170	3563	10	70	American

Describing Data: Types of variables

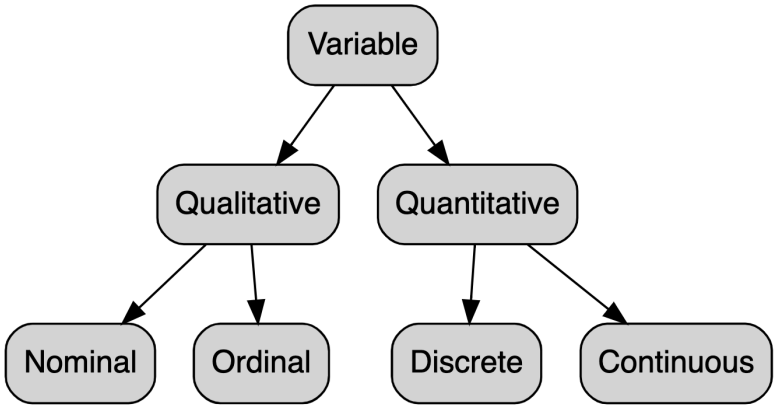
- **Nominal variables** - A qualitative nominal variable is a qualitative variable where no ordering is possible or implied in the levels.

- **Ordinal variables** – a qualitative ordinal variable is a qualitative variable with an order implied in the levels.



Describing Data: Types of variables: exercise

- Using the next hierarchy, classify the variables:



1. Eye color
2. Hair color
3. Temperature in Celsius
4. Types of fruits
5. Time to complete a task
6. Political affiliation
7. Height in feet and inches
8. Brands of cars
9. Shirt Sizes
10. Types of pets

Feature Engineering

- Process of selecting, transforming, extracting, combining, and manipulating raw data to **generate the desired variables**

	text	sentiment
0	RT @NancyLeeGrahn: How did everyone feel about...	Neutral
1	RT @ScottWalker: Didn't catch the full #GOPdeb...	Positive
2	RT @TJMShow: No mention of Tamir Rice and the ...	Neutral
3	RT @RobGeorge: That Carly Fiorina is trending ...	Positive
4	RT @DanScavino: #GOPDebate w/ @realDonaldTrump...	Positive

Feature Engineering

der_Date	Customer_ID	PromoID	Product_ID	Cost	Retail_Price	Shipping_Cost	Category	Genre	Reason	Discount	Age	Gender	Location	Discount_Value
17-11-22	4002	Full Price	376401	11	17	2.5	Fiction	Sci-fi	Not Returned	0.0	60	M	Non-US	0.0
17-01-23	4002	30_OFF	318652	10	17	2.5	Fiction	Thriller	Product Not Wanted	0.3	60	M	Non-US	5.1
17-02-01	1462	30_OFF	376401	11	17	2.5	Fiction	Sci-fi	Not Returned	0.3	40	M	US	5.1
17-08-01	2212	30_OFF	376401	11	17	2.5	Fiction	Sci-fi	Product Not Wanted	0.3	34	F	Non-US	5.1
17-12-14	3643	Full Price	376401	11	17	2.5	Fiction	Sci-fi	Not Returned	0.0	19	F	Non-US	0.0

Descriptive analysis

- Measuring of central tendencies are fundamental. It consists as a statistical index that describes the average of a set of values

- Kinds of Averages:

  - **Mode** – a numeric value in a distribution that occurs most frequently

$$3, 4, 5, 6, 7, 7, 7, 8, 8, 9$$

  - **Median** – an index of average position in a distribution of numbers

$$3, 4, 7, 2, 3, 7, 4, 2, 4, 7, 4$$

Before identifying the median, the values must be sorted:

$$2, 2, 3, 3, 4, 4, 4, 4, 7, 7, 7$$

Descriptive analysis

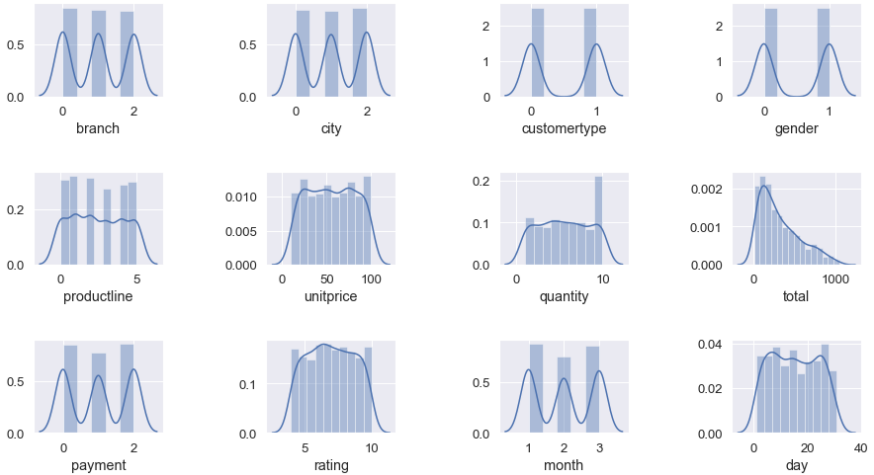
- **Mean** – the point on the score scale that is equal to que sum of the scores divided by the total number of scores

$$3, 4, 5, 7, 7, 8, 9, 9, 9$$

- The sum of all nine values is (3 + 4 + 5 + 7 + 7 + 8 + 9 + 9 + 9) or 61. The sum divided by the number of values is 61 ÷ 9 or 6.78.

Distribution of data

- Provides information in how the different values are distributed. It can be analysed using:
  - Range
  - Quartiles
  - Variance
  - Standard Deviation
  - Shapes

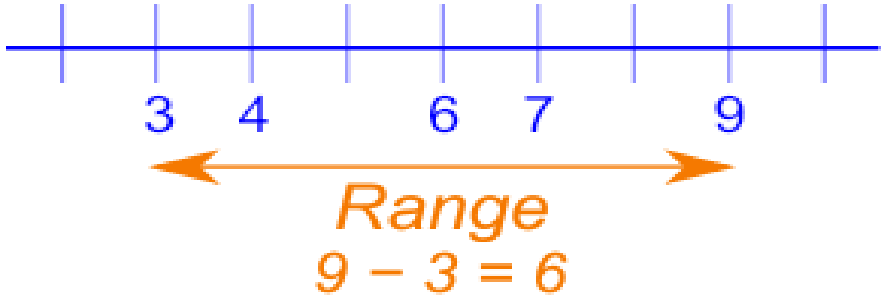


Distribution of data: Range

- **Range** is the variation for a particular variable. It is calculated as the difference between the highest and lowest values.

$$\mathbf{2}, 3, 4, 6, 7, 7, 8, \mathbf{9}$$

- The range is 7 calculated from the highest value (9) minus the lowest value (2).



Distribution of data: Quartiles

- **Quartiles** divide a continuous variable into four even segments based on the number of observations.

- First quartile (Q1) is at the 25%
- Second quartile (Q2) is at the 50
- Third quartile (Q3) is at the 75%

- The calculation for Q2 is the same as the median value (described earlier).

- Example:

Distribution of data: Quartiles

3, 4, 7, 2, 3, 7, 4, 2, 4, 7, 4

The values are initially sorted:

2, 2, 3, 3, 4, 4, 4, 4, 7, 7, 7

Median (Q2) 50 %

Next, the median or Q2 is located in the center:

2, 2, 3, 3, 4, **4**, 4, 4, 7, 7, 7

We now look for the center of the first half (shown underlined) or Q1:

Quartile 1 (25 %)

2, 2, **3**, 3, 4, **4**, 4, 4, 7, 7, 7

The value of Q1 is recorded as 3.

Distribution of data: Quartiles

Finally, we look for the center of the second half (shown underlined) or Q3:

2, 2, 3, 3, 4, **4**, 4, 4, **7**, 7, 7

Q3 (75 %)

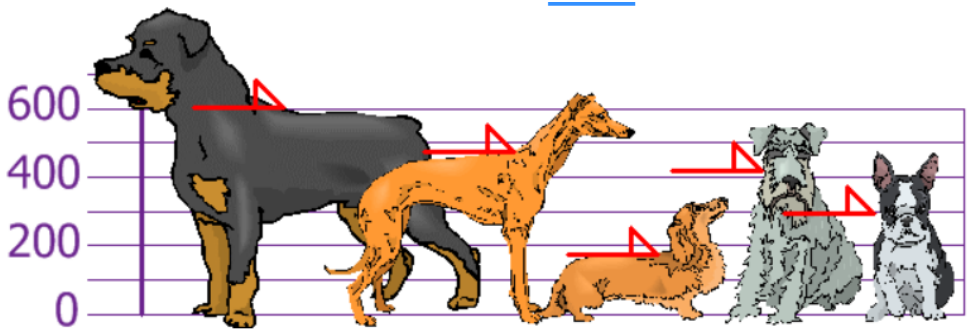
The value of Q3 is identified as 7.

Distribution of data: Variance

- The variance describes the spread of the data and measures how much the values of a variable differ from the mean.

- Let us analyse a funny example

Distribution of data: Variance



The heights (at the shoulders) are: 600mm, 470mm, 170mm, 430mm and 300mm.

Find out the Mean, the Variance, and the Standard Deviation.

Your first step is to find the Mean:

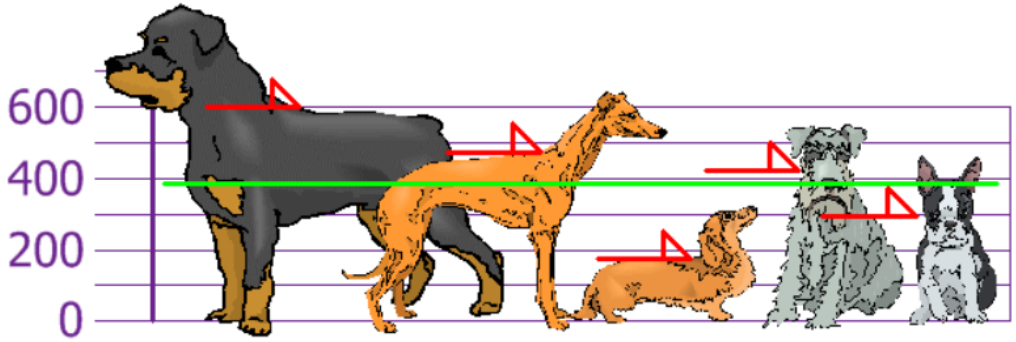
Answer:

$$\text{Mean} = \frac{600 + 470 + 170 + 430 + 300}{5}$$

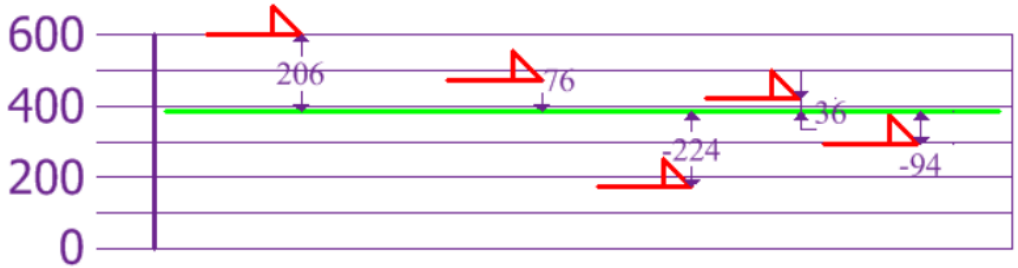
$$= \frac{1970}{5}$$

$$= 394$$

Distribution of data: Variance



Now we calculate each dog's difference from the Mean:



Distribution of data: Variance

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Variance

$$\sigma^2 = \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5}$$

$$= \frac{42436 + 5776 + 50176 + 1296 + 8836}{5}$$

$$= \frac{108520}{5}$$

$$= 21704$$

Distribution of data: Standard Deviation

- The standard deviation is the square root of the variance. For a sample from a population, the formula is

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

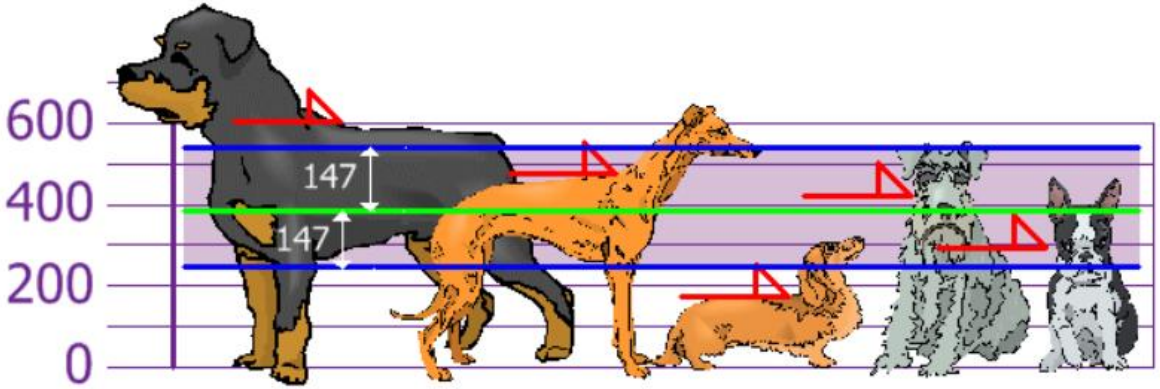
Standard Deviation

$\sigma = \sqrt{21704}$

$= 147{,}32\ldots$

$=$ **147** *(to the nearest mm)*

Distribution of data: Standard Deviation



So, using the Standard Deviation we have a "standard" way of knowing what is normal, and what is extra large or extra small.

Rottweilers **are** tall dogs. And Dachshunds **are** a bit short, right?

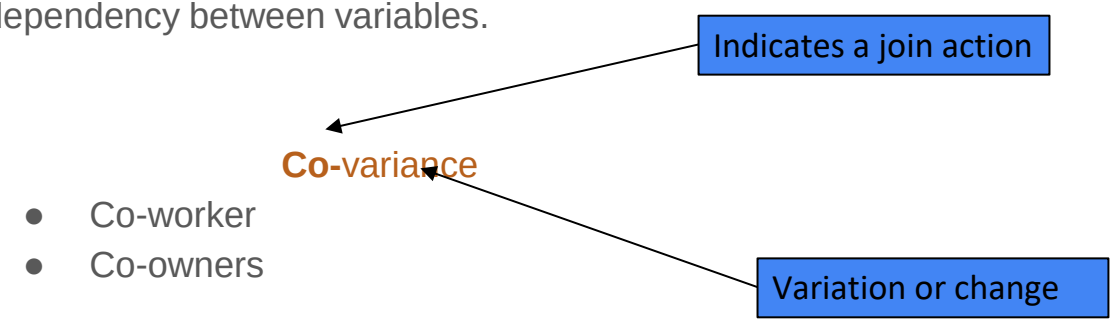
Distribution of data

	phd	service	salary
count	78.000000	78.000000	78.000000
mean	19.705128	15.051282	108023.782051
std	12.498425	12.139768	28293.661022
min	1.000000	0.000000	57800.000000
25%	10.250000	5.250000	88612.500000
50%	18.500000	14.500000	104671.000000
75%	27.750000	20.750000	126774.750000
max	56.000000	51.000000	186960.000000

Distribution of data: Covariance

- **Covariance** is a statistical measure that shows whether two variables are related by measuring how variables change in relation to each other

- It is essentially a measure of the variance between two variables. However, the metric does not assess the dependency between variables.



- Co-worker
- Co-owners

- Covariance measure how two things change together

Distribution of data: Correlation

- **Correlation** is a measure of how two variables change in relation to each other, but it goes one step further than covariance in that correlation tells how strong the correlation is.

Covariance

Correlation

Distribution of data: Covariance

Imagine that you are the owner that a new ice cream shop near the beach!

I sold more when it was hot! Is that true?

Let us analyse data.

Covariance

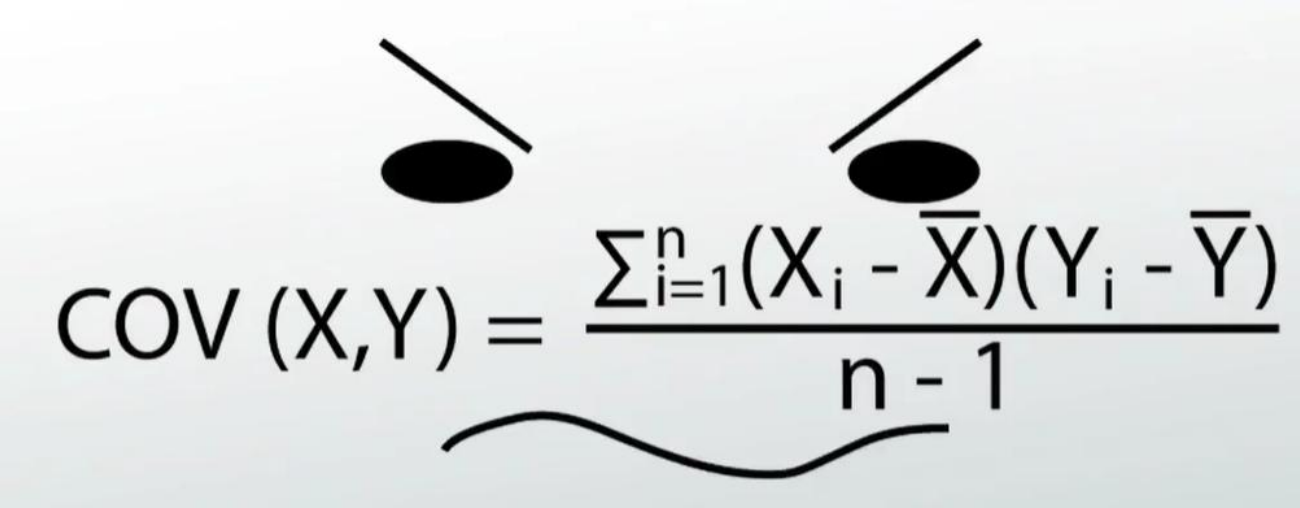
- **Positive covariance** as one increases the other also increases
- **Negative covariance** as one increases the other decreases

Temperature	Number of Customers
98	15
87	12
90	10
85	10
95	16
75	7

$$COV(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n - 1}$$

Covariance

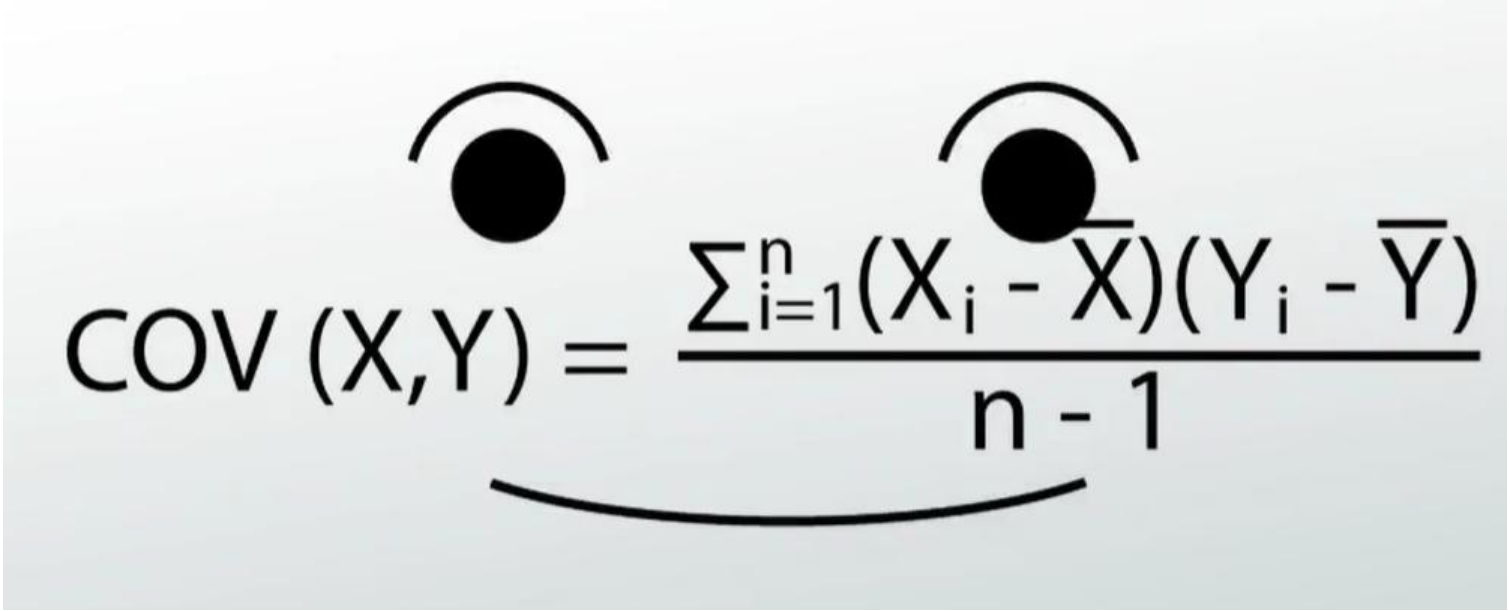
SO scaring!!!!



$$COV(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

Covariance

It is not !!!!



Covariance

Differences between the mean and the variable

$$COV(X,Y) = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{n-1}$$

Total of the sample minus 1

Covariance

Temperature	Number of Customers
98	15
87	12
90	10
85	10
95	16
75	7

Mean = 88.33

Mean = 11.67

Temperature $(x - \bar{x})$	Customers $(y - \bar{y})$	Product $(x - \bar{x})(y - \bar{y})$
98 - 88.33= 9.67	15 - 11.67= 3.33	32.20
87 - 88.33= -1.33	12 - 11.67= 0.33	-0.44
90 - 88.33= 1.67	10 - 11.67= -1.67	-2.79
85 - 88.33= -3.33	10 - 11.67= -1.67	5.56
95 - 88.33= 6.67	16 - 11.67= 4.33	28.88
75 - 88.33= -13.33	7 - 11.67= -4.67	62.25

total = 125.66 /5 = 25.1

Positive relationship

Covariance - Python

- Create a data frame with these values and run the next comands

Temperature	Number of Customers
98	15
87	12
90	10
85	10
95	16
75	7

```
df =df['Temperature'].cov(df['N Customers'])
```

```
df.cov()
```

Correlation

- In the previous example, the number is positive, so we can state that the two variables have a positive relationship; as temperature rises, the number of customers in the store also rises.
- What this doesn't tell us is how strong this relationship is. To find the strength, we need to continue with correlation.
- To determine the strength of a relationship, you must use the formula for correlation coefficient.
- This formula will result in a number between -1 and 1

Correlation

- **Perfect inverse correlation**: the variables move in opposite directions reliably and consistently. In this case the result should be -1
- **Neutral relationship** between the two variables: the result should be near 0
- **Perfect positive correlation**: the variables reliably and consistently move in the same direction as each other. The result should approximate to 1

Correlation - Python

- Create a data frame with these values
- Calculate the correlation between the variables

- Use the folowing comand lines

```
df.corr()
```

Temperature	Number of Customers
98	15
87	12
90	10
85	10
95	16
75	7

Additional examples for correlation and covariance

Distribution of data: Examples

```
import pandas as pd

df = pd.DataFrame({'a': np.random.randint(0, 50, 1000)})
df['b'] = df['a'] + np.random.normal(0, 10, 1000)
df['c'] = 100 - df['a'] + np.random.normal(0, 5, 1000)
df['d'] = np.random.randint(0, 50, 1000)
df.corr()
```

What do you conclude?

Pearson correlation

- There are several types of correlations
- Pearson correlation is the most widely used correlation statistic to measure the degree of the relationship between linearly related variables.

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

r_{xy} = Pearson r correlation coefficient between x and y
n = number of observations
x_i = value of x (for ith observation)
y_i = value of y (for ith observation)

Distribution of data: Examples

```
import pandas as pd

df = pd.DataFrame({'a': np.random.randint(0, 50, 1000)})
df['b'] = df['a'] + np.random.normal(0, 10, 1000)

df['c'] = 100 - df['a'] + np.random.normal(0, 5, 1000)
df['d'] = np.random.randint(0, 50, 1000)
df.corr(method='pearson')
```

What do you conclude?

Pearson is the default correlation method in Python

UNIVERSIDADE
PORTUCALENSE