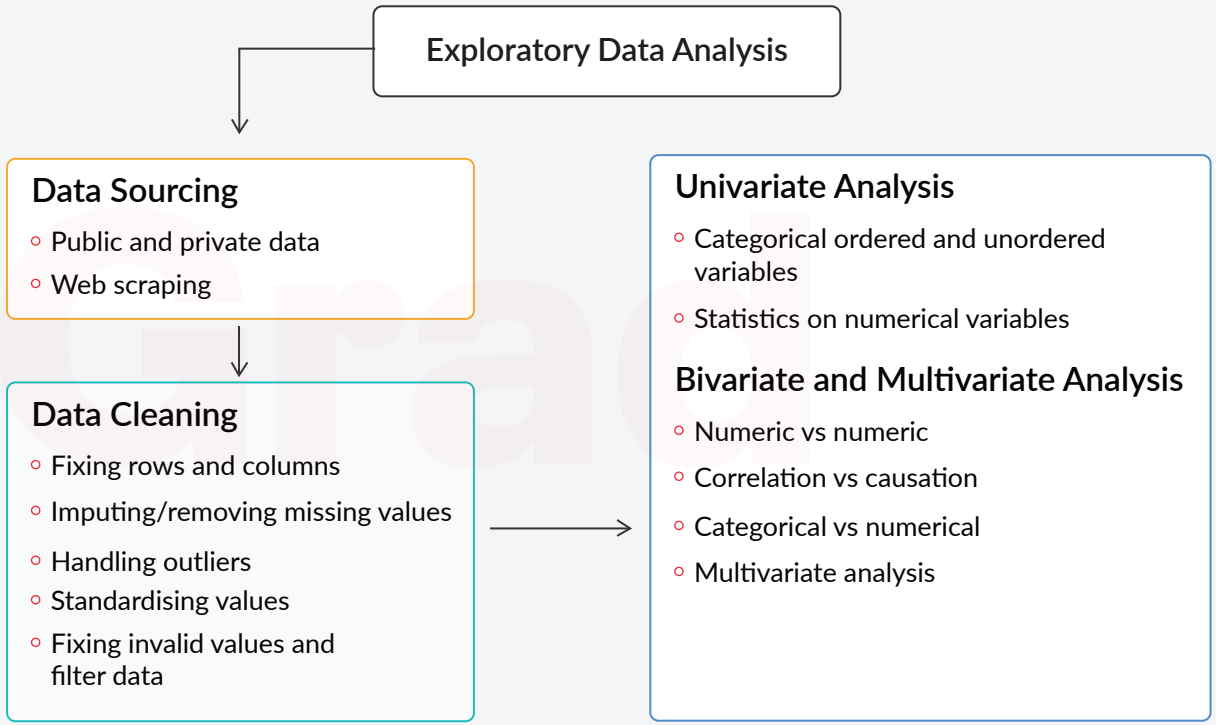
EXPLORATORY DATA ANALYSIS

Exploratory data analysis is essential for any business. It allows data scientists to analyse data before making any assumptions. It also ensures that the results produced are valid and applicable to business outcomes and goals

Common Interview Questions

1. Why is exploratory data analysis significant?
2. What is data wrangling?
3. How can you handle missing values?
4. What is univariate analysis?
5. What is a spurious correlation?
6. Does causation imply correlation?
7. How do you treat the outliers in a dataset?
8. What are the benefits of data cleaning?
9. What is multivariate analysis?
10. What is an interquartile range?



DATA SOURCING

Data Sourcing:

- **Private data:** Any personal, personally identifiable, financial, sensitive or regulated information, such as bank data, credit card data and login credentials
- **Public data:** Available on government websites or open source, for example, population, rainfall and marketing materials

#Storing the data in a CSV file

```
#Storing the data in a CSV file
filename= "imdb_m.csv"
f= open(filename, "w")

headers= "Name, Year, Runtime \n"
f.write(headers)

for container in containers:

  name= container.img["alt"]
  year_mov= container.findAll("span", {"class": "lister-item-year"})
  year=year_mov[0].text
  runtime_mov= container.findAll("span", {"class": "runtime"})
  runtime=runtime_mov[0].text

  print(name + "," + year + "," + runtime + \n")
  f.write(name + "," + year + "," + runtime + \n")

f.close()
```

Web Scraping (From IMDb Website)

```
# Import useful libraries and classes.
from urllib.request import urlopen as uReq
from bs4 import BeautifulSoup as soup


#html upload
my_url= "http://www.imdb.com/search/title?sort=num_votes
desc&start=1&title_type=feature&year=1950,2012"
uclient= ureg(my_url)
page_html= uclient.read()
uclient.close()

#html parser
page_soup=Soup(page_html,"html. parser")
page_soup

#read class from web page.
containers= page_soup findAll("div", {"Class":
"lister-item mode -advanced"}}
print(len(containers))
```

DATA WRANGLING

Various data types with examples

Example	Variable Type	Data Type
Height, weight, age, temperature	Numerical variables	Int, float
Size of clothes, months, types of jobs, blood group	Categorical variable	Object
Grades in exams, education level, months, integer ratings	Ordinal categorical type	Object, int, float
Date, time, timestamp	Date and time variable	Date and time

Strategies to clean a dataset

- Fixing rows and columns
- Imputing/removing missing values
- Handling outliers
- Standardising values
- Fixing invalid values and filtering data

Fixing Rows and Columns

Checklist for fixing rows:

- **Delete summary rows:** Total and Subtotal rows
- **Delete incorrect rows:** Header row and footer row
- **Delete extra rows:** Column numbers, indicators, blank rows, page numbers

Checklist for fixing columns:

- If needed, merge columns for creating unique identifiers
- Split columns to get more data
- **Add column names:** Add column names if missingnumbers
- **Rename columns consistently:** Abbreviations, encoded columns
- **Delete columns:** Delete unnecessary columns
- **Align misaligned columns:** A data set may have shifted columns, which you need to align correctly

Types of Missing Values

MCAR (missing completely at random), MAR (missing at random) and MNAR (missing not at random)

Dealing with missing values

- Impute or delete missing values based on their importance

Imputation on Categorical Columns

1. Categorical columns:

- Impute the most popular category using logistic regression techniques

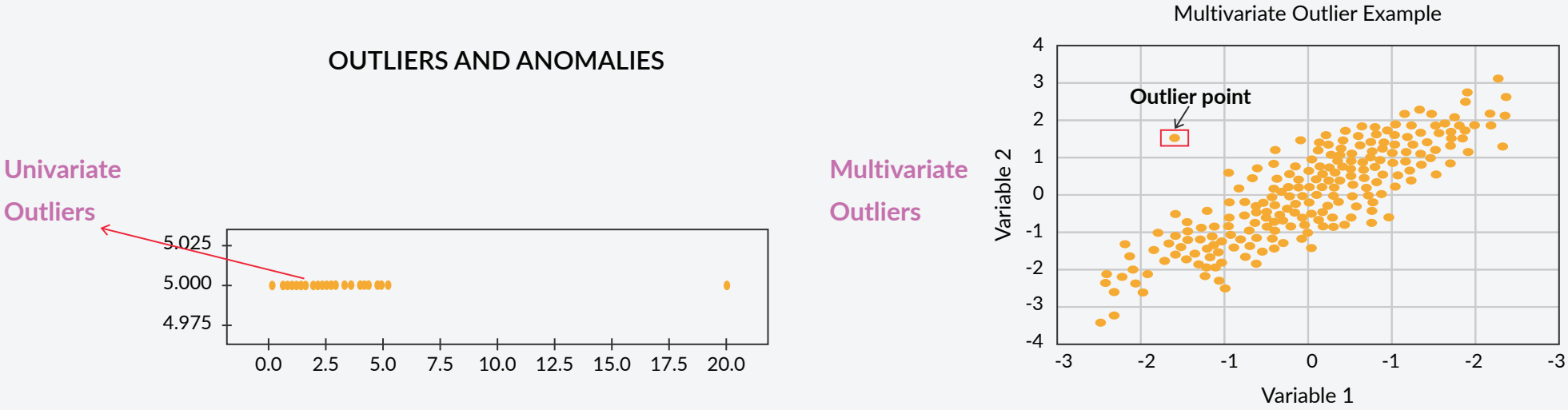
2 . Numerical column:

- Impute the missing value with the mean/median/mode
- The other methods to impute missing values involve the use of interpolation, linear regression

DATA CLEANING

Outliers: Outliers are the values that are much beyond or far from the next nearest datapoint

Types of Outliers



Handling Outliers: Techniques

- **Imputation:** Replace outliers with the median or with any other estimated value
- **Deletion of outliers:** Remove outliers when you have a valid reason
- **Binning of values:** Data distributed in several of bins or buckets
- **Capping outliers:** Set a limit beyond which a value will be considered as an outlier Standardising values
- **Standardise units:** Ensure all observations under one variable are expressed in a common and consistent unit; for example, convert lbs to kg and miles/h to km/h
  - **Standardise units:** Ensure all observations under one variable are expressed in a common and consistent unit; for example, convert lbs to kg and miles/h to km/h
  - **Scale values if required:** Ensure all observations under one variable have a common scale
  - **Standardise precision:** for a better presentation of data, e.g., change 4.5312341 kg to 4.53 kg

Fixing Invalid Values

- Encode Unicode properly
- Convert incorrect data types
- Correct the values that lie beyond a range
- Correct the values not belonging to a list
- Fix incorrect structure

Filter Values

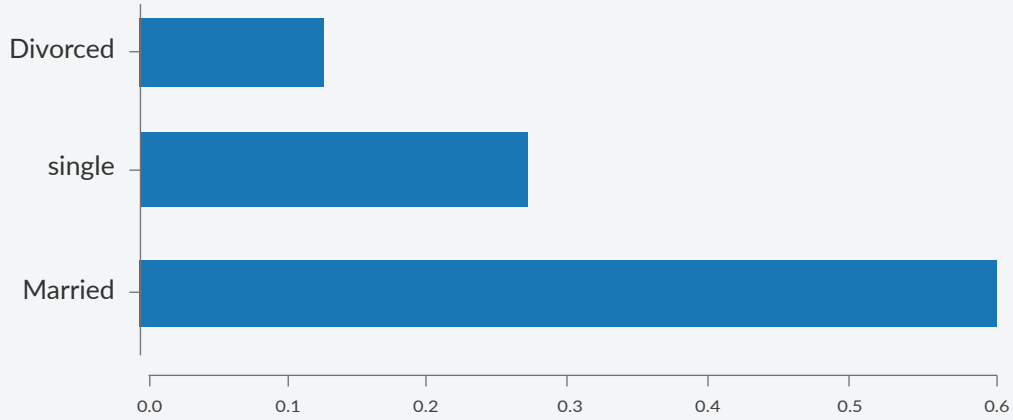
- **Deduplicate data:** Remove identical rows and rows with identical columns
- **Filter rows:** Filter rows by segment and date period
- **Filter columns:** Filter columns relevant to the analysis
- **Aggregate data:** Group by the required keys and aggregate the rest

UNIVARIATE ANALYSIS

Categorical unordered univariate analysis

Unordered data is the type of data that does not have any measurable terms (measurable terms can be high–low, more–less, fail–pass, etc.)
Example: Departments in an organisation – sales, HR, marketing

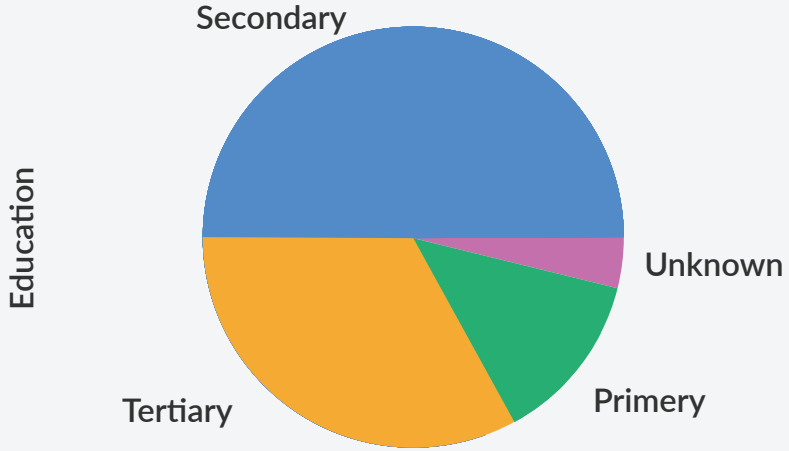
```
1  inp.marital.value_counts(normalize= True).plot.barh ()
2  plt.show()
```



Categorical Ordered Univariate Analysis

Ordered variables are those variables that follow a natural rank of order

Examples: Months – January, February, March.
Education – Primary, secondary, etc.



Statistics on numerical variable

Mean: Average/weighted average of a numerical feature

Median: Midpoint of ordered data

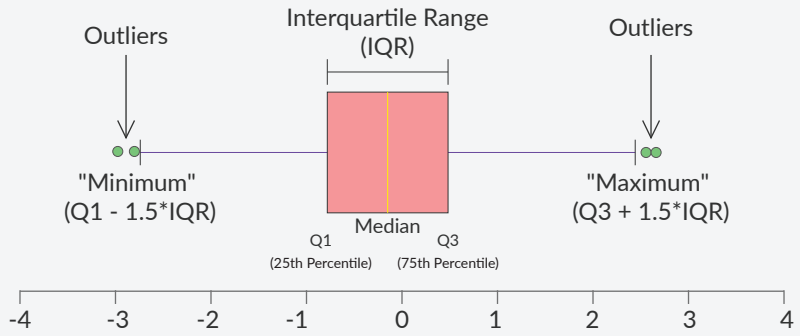
Mode: Highest-occurrence datapoint in a dataset

Standard deviation: How dispersed data is with respect to the mean

Variation: Square root of the standard deviation

Outliers: An observation that lies at an abnormal distance from the other values in a random sample

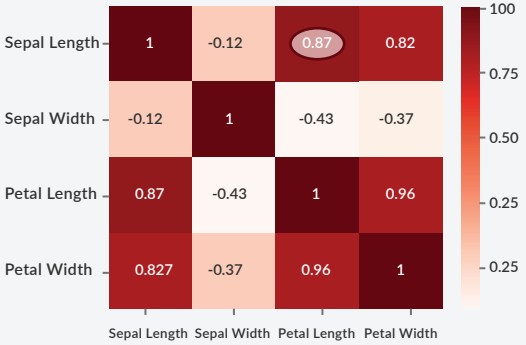
Statistics on Numerical Variables



BIVARIATE AND MULTIVARIATE ANALYSES

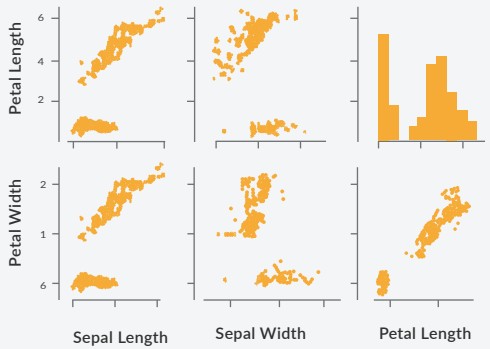
Analysis between two numerical features

IRIS DATABASE: CORRELATION IN FLOWER PARAMETERS



A correlation matrix cannot show the exact distribution, but a pair plot can

IRIS DATABASE: PAIR PLOT OF FLOWER PARAMETERS: SEPALAND PETAL MEASURMENT



Correlation vs Causation

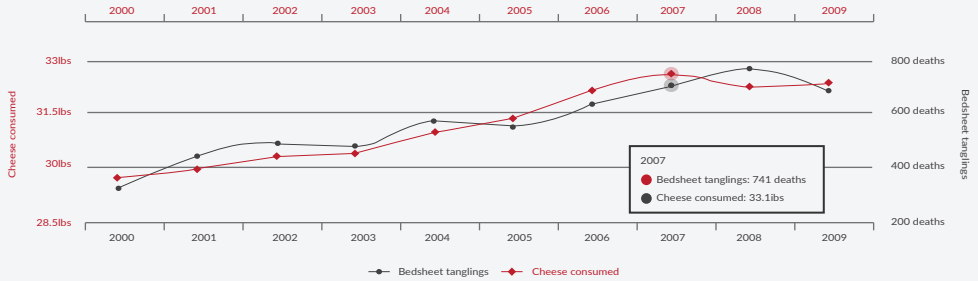
Some numerical variables can be highly correlated, although there may not be any cause for any relationship between them. Such correlations are called spurious correlations

Per capita cheese consumption

Correlates with

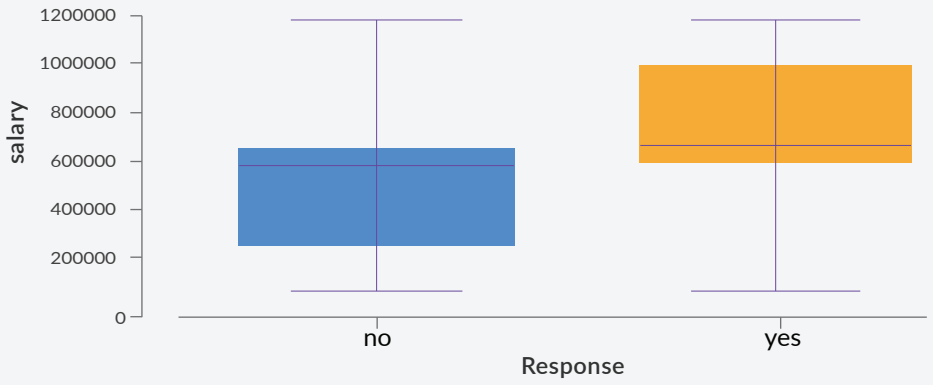
Number of people who died by becoming tangled in their bedsheets

Correlation: 94.71% (r=0.947091)



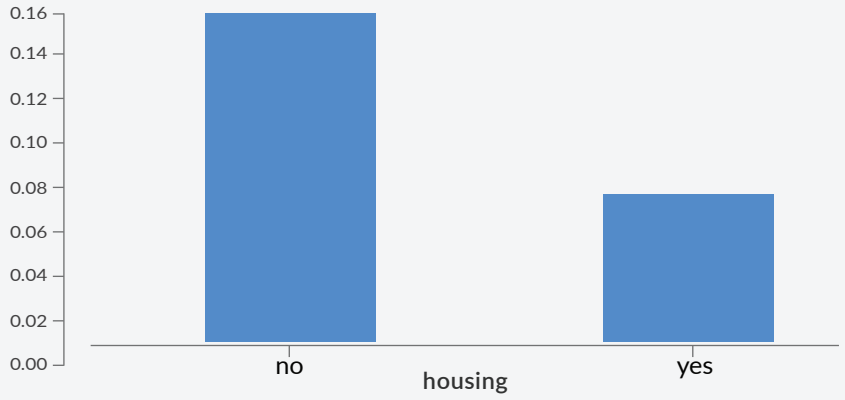
Numerical vs Categorical

In a banking dataset, it is obvious that people with higher salary positively respond to invest in term deposit



Categorical vs Categorical

In a banking dataset, customer without any housing or personal loan are inclined to invest in term deposit



MULTIVARIATE ANALYSIS

In a banking dataset, people who are married and who have completed just their primary education are least likely to give a positive response on term deposits. This can be explained by the fact that people educated only up to the primary level are not aware of the benefits of term investments. Also, married individuals need money to fulfil their daily needs, and they require cash-on-hand to buy the daily essentials; hence, they won't prefer investing in term deposits

Here is a correlation matrix for reference