TRANSFORM 4EUROPE

Co-funded by the Erasmus+ Programme of the European Union

Data mining and visualisation

David Tomás Díaz
@d_tomas

Universidad de Alicante

lsi Departamento de Lenguajes y Sistemas Informáticos

Contents

- The importance of data
- From data to knowledge
- Exploratory data analysis
  - Descriptive statistics
  - Visualisation
- …and to know more

Contents

- **The importance of data**
- From data to knowledge
- Exploratory data analysis
  - Descriptive statistics
  - Visualisation
- …and to know more

The importance of data

- We live in a world flooded with data
  - Websites that monitor every click of their users
  - Cell phones accumulating location records
  - Smart vehicles collecting driving habits
  - Smart homes collecting lifestyle habits
  - Online stores collecting shopping habits
  - All kinds of government statistics
  - Internet of Things
  - The quantified self
  - ...

The importance of data

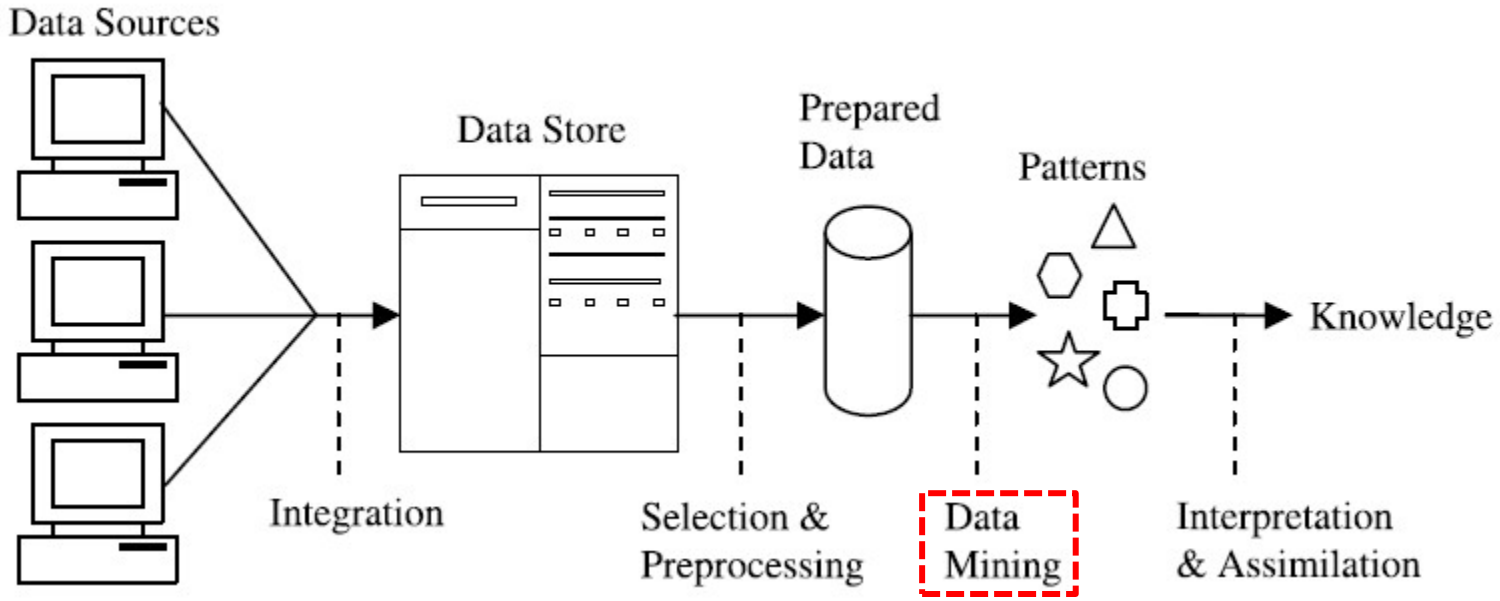
Buried in this data are the answers to countless questions that no one has even thought to ask

Contents

- The importance of data
- **From data to knowledge**
- Exploratory data analysis
  - Descriptive statistics
  - Visualisation
- …and to know more

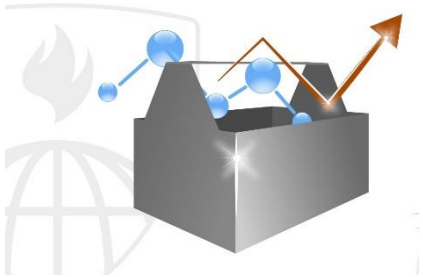
From data to knowledge

- Extracting knowledge from data
  - *Knowledge Discovery in Databases* (KDD)
  - "Non-trivial extraction of implicit, previously unknown and potentially useful information from data



From data to knowledge

- The Data Scientist's Toolbox
  - *Python* (your favourite programming language)
  - *Google Colab* (your favourite development platform)
  - *Pandas* (your favourite library for manipulating tables)
  - *NumPy* (your favourite library for vector manipulation)
  - *Scikit-learn* (your favourite library for *machine learning*)
  - *Seaborn* (your favourite library for visualisation)
  - *Kaggle* (your favorite website to meet the beautiful people)



From data to knowledge

- Data mining
  - Process of discovering patterns in large volumes of data
  - Analogy with a real mining process
    - We start from a mineral (data)
    - We arrive at the refined final product (knowledge)
    - Employs machine learning, statistics, and databases methods

From data to knowledge

- Data mining
  - Knowledge extraction is a challenge due to the wide disparity of existing problems and data types
  - Product recommendation is very different from an intrusion detection application
    - Format of input data
    - Problem definition
  - Even within similar types of problems the differences are quite significant
    - Recommending a product in a database
    - Recommendation of contacts in a social network

From data to knowledge

- Applications
  - Analyze satellite image
  - Analysis of organic compounds
  - Credit card fraud detection
  - Electricity consumption prediction
  - Medical diagnoses
  - House market prediction
  - Targeted marketing
  - Weather forecast
  - Predict TV audience
  - …

From data to knowledge

- Examples
  - Facebook
    - Ask users for their hometown and current location
    - Apparently, the goal is to make it easier for your friends to find you and connect
    - It also analyzes these locations to identify global migration patterns and where fans of different football teams live

From data to knowledge

- Examples
  - 2012 Obama Campaign
    - He employed dozens of data scientists to identify voters who needed extra attention, whose vote was more likely to be useful
    - Identified optimal donor-specific fundraising programs

From data to knowledge

- Examples
  - OkCupid
    - Asks thousands of questions to its members (from climate change to cilantro) to find the most appropriate partners
    - Analyze these results to identify "innocuous" questions to ask to find out how likely someone is to sleep with you on the first date

An aside

Google Colaboratory

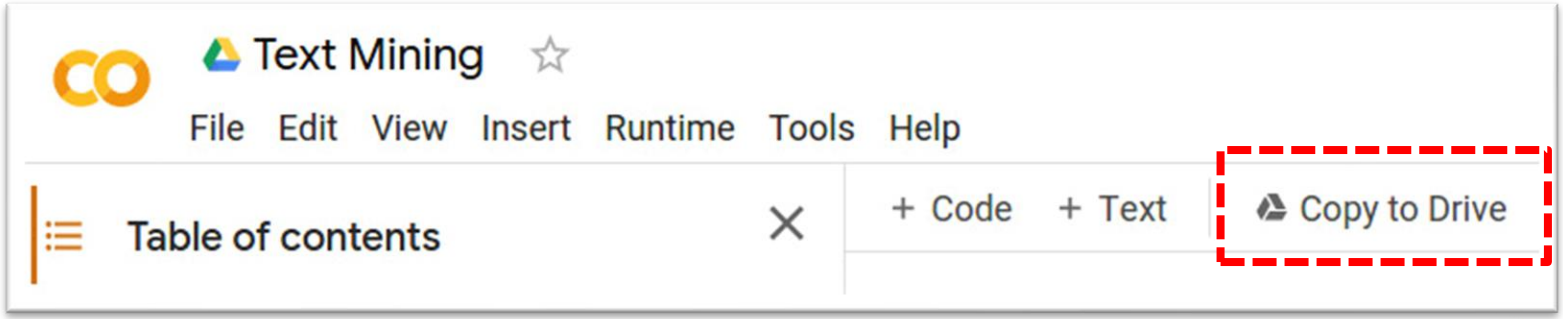
- Free *Jupyter* notebook environment running on Google cloud (Gmail account access)
  - https://colab.research.google.com
- *Jupyter* is an interactive web environment that allows editing and running Python code
- We can run our code on CPU and GPU in the cloud
- *Jupyter* notebooks are saved in Google Drive and can be shared like any other document
- There are some limitations:
  - Initial machine of 12GB RAM and 100GB hard drive
  - Maximum running time: 12 hours
  - If we are more than 90 minutes without using the notebook it disconnects

Google Colaboratory

- Access

https://bit.ly/3Ma4DDA

- Make you own copy in Drive



An aside

Contents

- The importance of data
- From data to knowledge
- **Exploratory data analysis**
  - Descriptive statistics
  - Visualisation
- …and to know more

Exploratory Data Analysis

- Definition
  - Techniques for analyzing data through statistical processing
    - Import, clean and validate
    - View distributions
    - Explore relationships between variables
    - Feature selection
    - Identification of outliers
    - …
  - Used in the initial phase of any data science project
  - The goal is to have a solid knowledge of the data

Exploratory Data Analysis

- Types of analysis
  - Descriptive statistics
    - Mean
    - Median
    - Mode
    - Variance
    - ...
  - Visualisation
    - Histogram
    - Scatter plot
    - Box diagram
    - Word clouds
    - …

Contents

- The importance of data
- From data to knowledge
- Exploratory data analysis
  - **Descriptive statistics**
  - Visualisation
- …and to know more

Descriptive statistics

- Definition
  - Mathematical techniques for summarising or describing datasets quantitatively
  - Identify data properties, noise, and outlieres
  - Common measures used to describe data
    - Central tendency
    - Dispersion

Descriptive statistics

- Central tendency
  - Columns can have thousands of different values
  - A basic step when exploring the data is to get a typical value for each column
  - Central tendency: estimation of where most of the data is located
  - Common measures
    - Mean
    - Median
    - Mode

Descriptive statistics

- Central tendency
  - Mean (also *average*)
    - Most common measure, although sensitive to extreme values

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - Median (also *50th percentile*)
    - Value that occupies the center position in an ordered set
    - If the dataset is even, it will be the average of the two occupying that position
    - Less sensitive to extreme values
      - E.g. household income in the area where Bill Gates lives

Descriptive statistics

- Central tendency
  - Mode
    - The most repeated value or category in a dataset
    - Mainly used for categorial data
    - There can be several modes in a set
      - Unimodal
      - Bimodal
      - Multimodal

Descriptive statistics

- Dispersion
  - The central tendency is a way of summarising a variable
  - Another way to do this is by dispersion (variability), measuring whether the values are clustered or scattered
  - Useful for identifying extreme values (*outliers*)
  - Common measures
    - Range
    - Quantile
    - Interquartile range
    - Standard deviation

Descriptive statistics

- Dispersion
  - Range
    - Difference between the largest and smallest value
    - It is the most basic measure of dispersion
    - It is very sensitive to extreme values
  - Quantile
    - Points taken at regular intervals of a distribution that divide it into sets of equal size
    - Quartiles: divide the distribution into 4 parts (quantile 0.25, 0.50 and 0.75)
    - Percentiles: divide the distribution into one hundred parts
      - The P percentile is a value so that at least P percent of the values have this value or less, and at most 100-P percent take this value or more
      - The median is the same as the 50th percentile

Descriptive statistics

- Dispersion
  - Interquartile range (*IQR*)
    - Usual measure of variability
    - Difference between the 75th percentile (Q3) and the 25th percentile (Q1)

$$IQR = Q_3 - Q_1$$

    - Shows the range covered by the central half of the data
    - Less sensitive to extreme values than standard deviation

Descriptive statistics

- Dispersion
  - Standard deviation
    - Square root of variance
    - Easier to interpret than variance, as it is on the same scale
    - Especially sensitive to extreme values (same as variance)

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

Descriptive statistics

Let's practice!

https://bit.ly/3w9cWKq

Contents

- The importance of data
- From data to knowledge
- Exploratory data analysis
  - Descriptive statistics
  - **Visualisation**
- …and to know more

Visualisation

- Definition
  - Visualising the data allows to highlight its main characteristics
  - Ways to present the data
    - Textual
    - Tabular
    - Graphic
  - The graphic representation is attractive and easy to understand
    - To explore the data
    - To communicate the data (not just to experts)

Visualisation

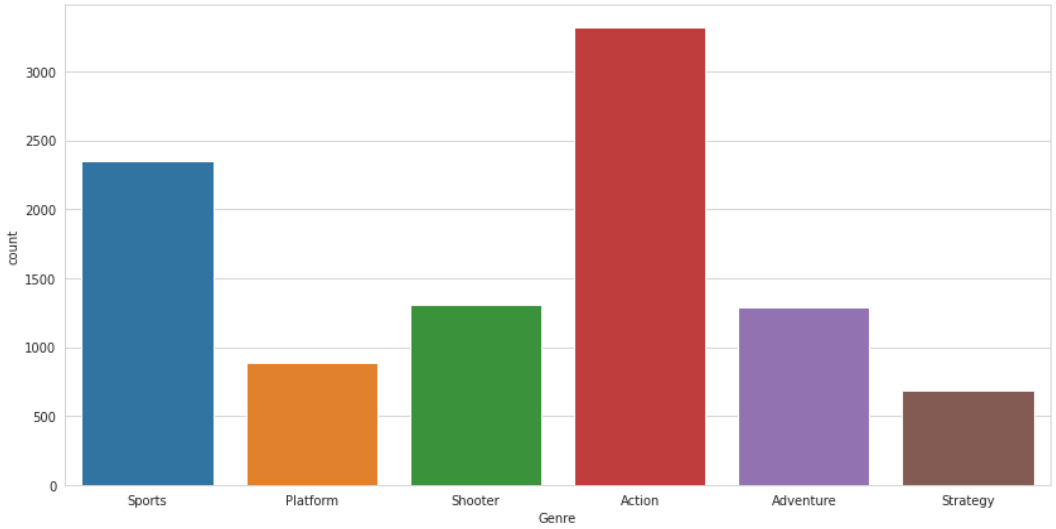
- Visualisation techniques can provide a quick answer to many important questions
  - What range do the observations cover?
  - What is the central tendency?
  - Is the distribution symmetric or is there asymmetry in some direction?
  - Is there evidence of bimodality?
  - Are there significant extreme values?
  - ...

Visualisation

- There are multiple types of diagrams for different purposes:
  - Bar chart (ranking)
  - Histogram (distribution)
  - Density plot (distribution)
  - Line chart (evolution)
  - Scatter plot (correlation)
  - Heat map (correlation)
  - Box plot (distribution and ranking)
  - Violin plot (distribution)
  - Word cloud (ranking)
  - …

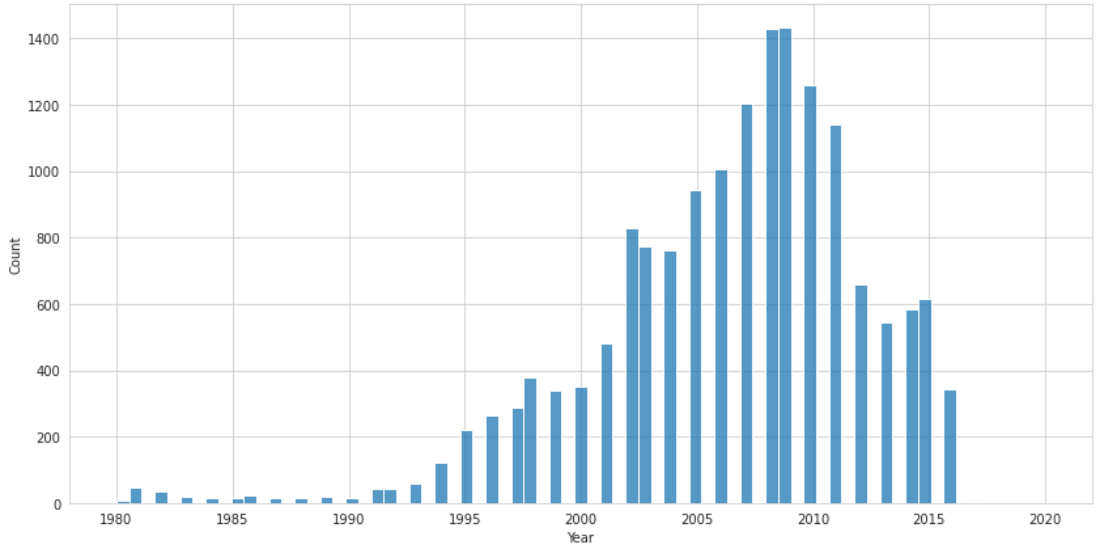
Visualisation

- Bar chart
  - Useful for displaying discrete or categorical value distributions
  - Graphically represents the comparison between data categories



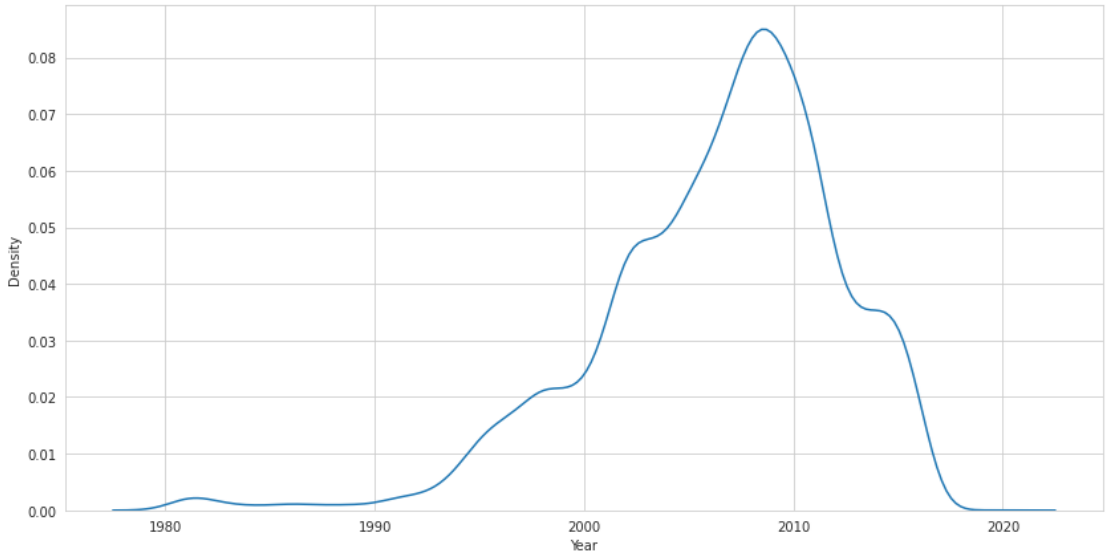
Visualisation

- Histogram
  - Bar chart type to display numeric value distributions
  - Indicates the number of observations that fall within a range of values (*bin*)



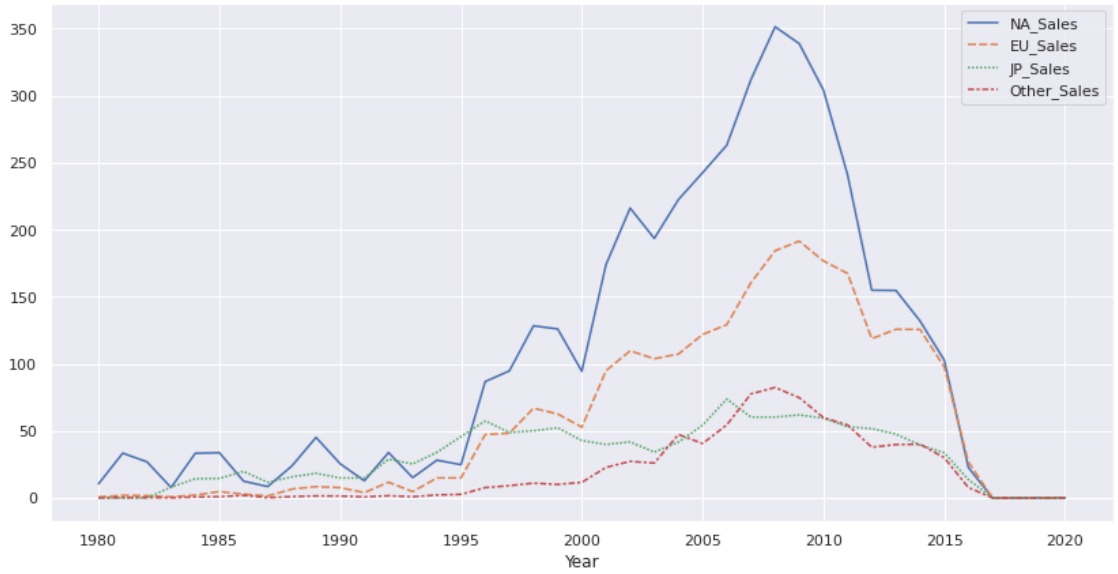
Visualisation

- Density plot
  - It is a variant of the histogram that uses a Gaussian kernel to smooth out values
  - Offer a better view of the shape of the distribution



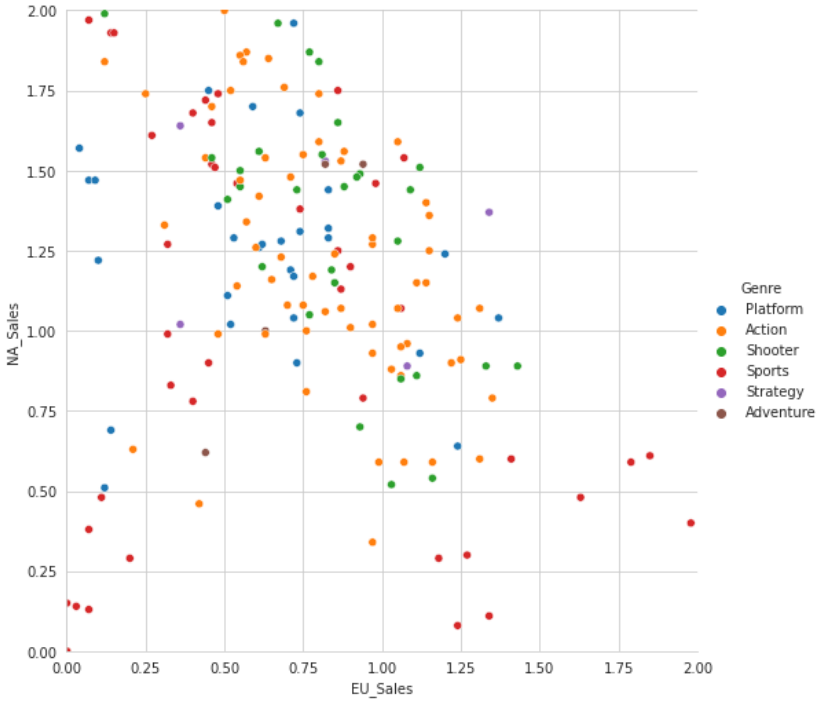
Visualisation

- Line chart
  - Useful for viewing tendencies (very common in time series)
  - Displays data as points joined by straight lines



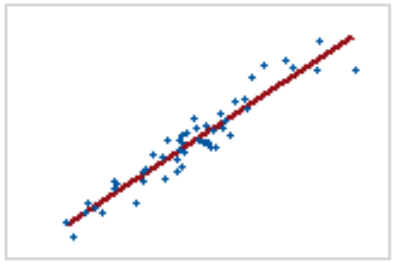
Visualisation

- Scatter plot
  - Useful for identifying relationships, patterns, or trends between two numeric values (bivariate)
  - Visualise data clusters
  - Identify outliers
  - Explore correlations

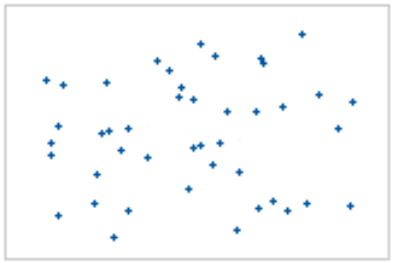


Visualisation

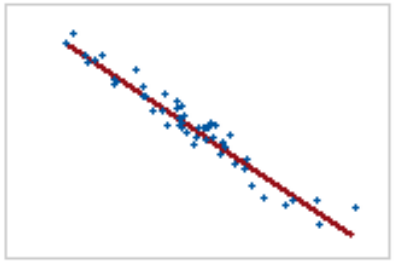
- Scatter plot
  - Two attributes are correlated if one implies the other
    - Positive: when one increases the other too
    - Negative: when one increases the other decreases
    - Neutral: no correlation



Positive correlation



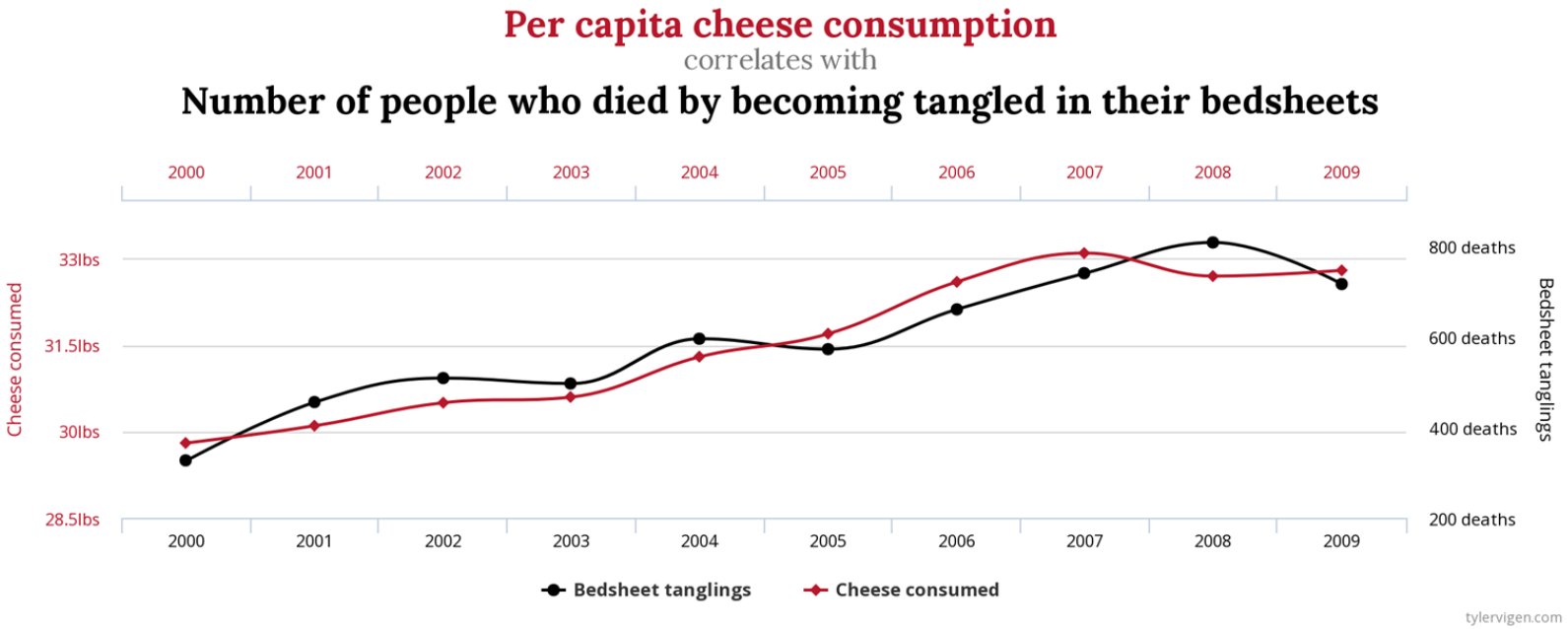
no correlation



Negative correlation

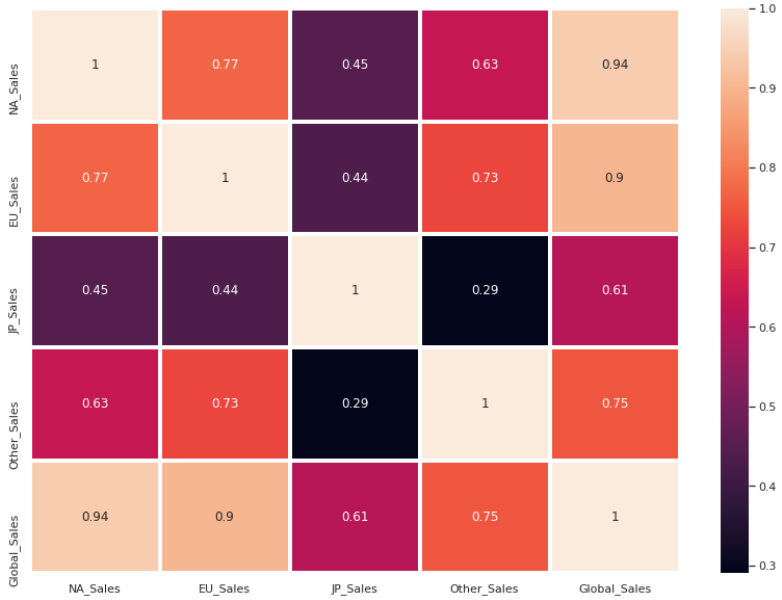
Visualisation

- Scatter plot
  - Correlation does not imply causation



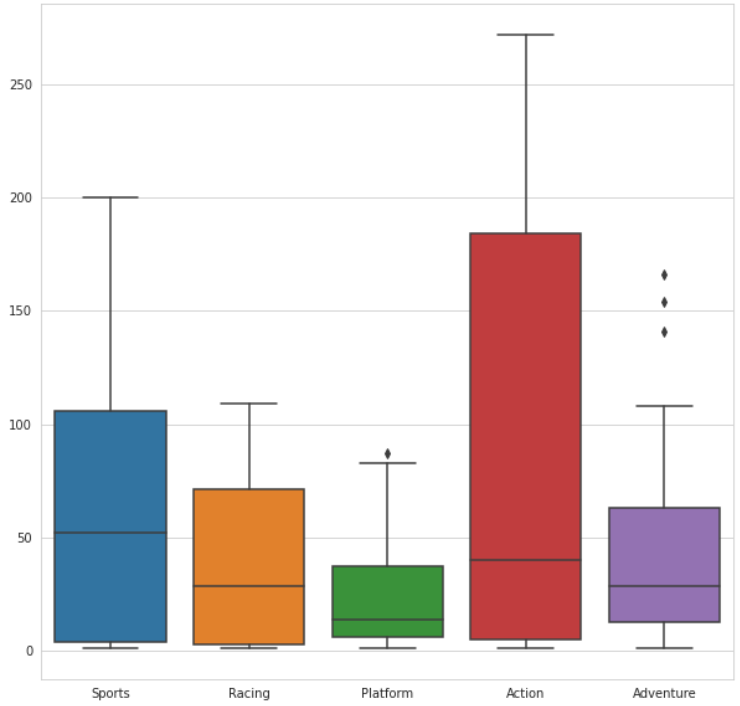
Visualisation

- Heatmap
  - Visualise data using two-dimensional color codes (useful for correlations)
  - Pitch and/or intensity indicates how data varies in space



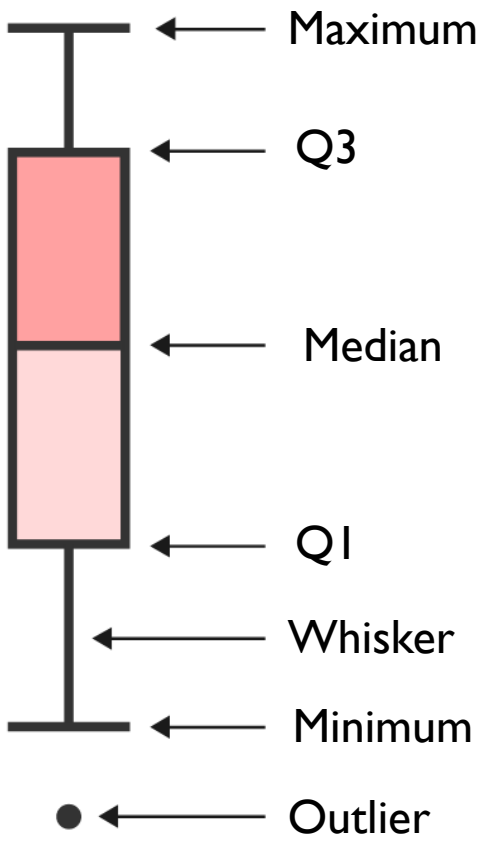
Visualisation

- Box plot
  - Describes numeric data groups using quartiles
  - Useful for data that does not follow a normal distribution



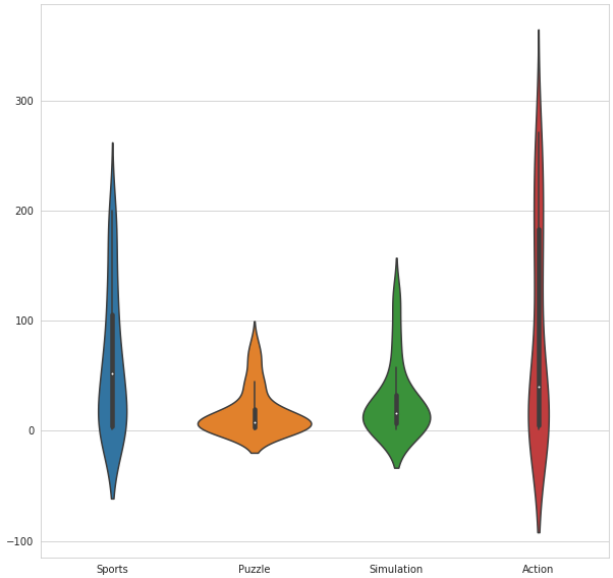
Visualisation

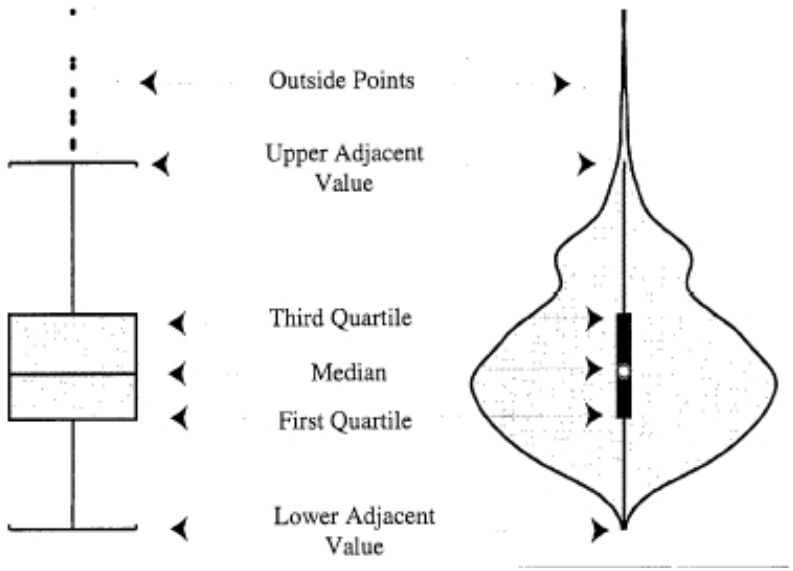
- Box plot
  - Five-number summary
    - Minimum
    - First quartile (Q1)
    - Median
    - Third quartile (Q3)
    - Maximum
  - *Outliers*
    - Below Q1 - 1.5 * IQR
    - Above Q3 + 1.5 * IQR



Visualisation

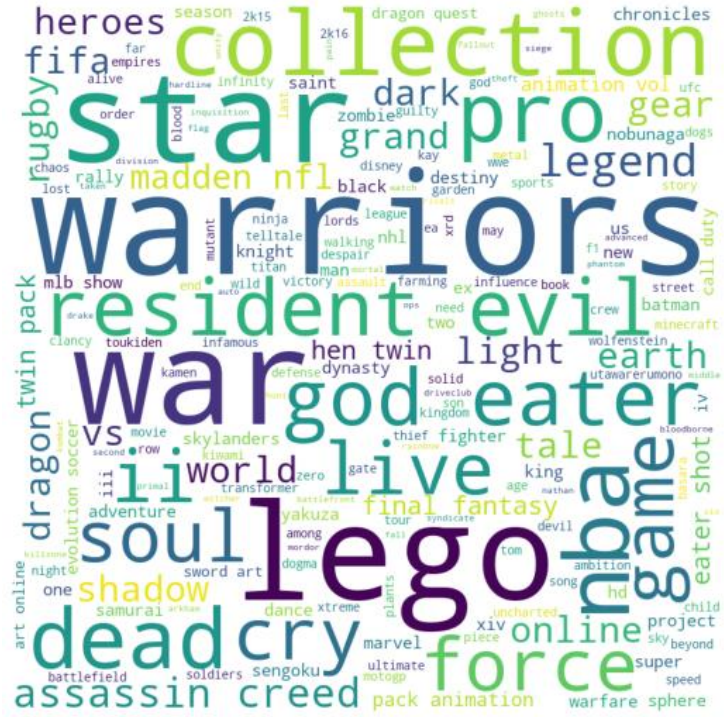
- Violin plot
  - Provides the information of a box plot and also the distribution of values
  - Suitable when there are many values and cannot be visualised individually





Visualisation

- Word cloud
  - Visual representation of the words that appear in a text
  - The size is larger for the most frequent words



Let's practice!

https://bit.ly/3svvwKx

Contents

- The importance of data
- From data to knowledge
- Exploratory data analysis
  - Descriptive statistics
  - Visualisation
- **…and to know more**

… and to know more

- Webs
  - Towards Data Science
  - KDnuggets
  - Kaggle competitions
  - Coursera
- YouTube
  - Dot CSV
- Courses (Warning! Concealed advertising)
  - "Master's Degree in Data Cience", título de la UA

@d_tomas

David Tomás Díaz