Utrecht University

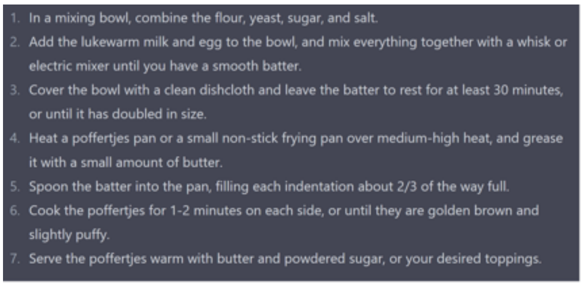
Applied Data Analysis & Visualisation I

Introduction

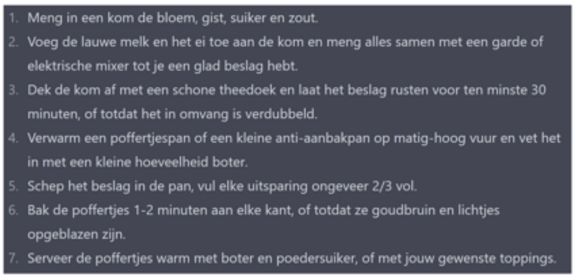
Department of Methodology and Statistics
Anastasia Giachanou

ChatGPT

Could you write two sets of instructions for making Poffertjes, one in English and one in Dutch?

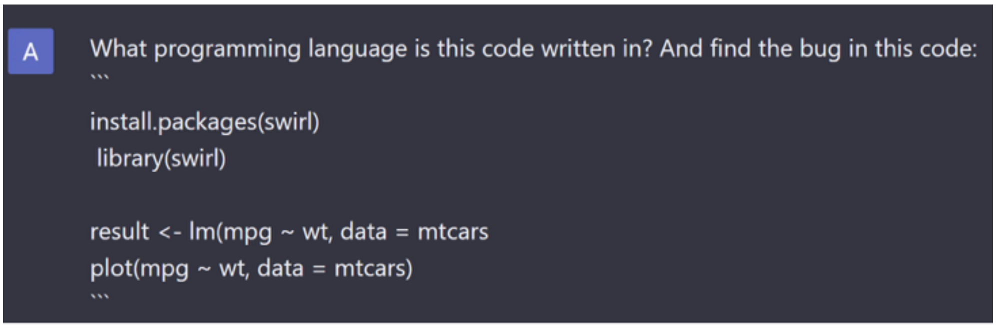


1. In a mixing bowl, combine the flour, yeast, sugar, and salt.
2. Add the lukewarm milk and egg to the bowl, and mix everything together with a whisk or electric mixer until you have a smooth batter.
3. Cover the bowl with a clean dishcloth and leave the batter to rest for at least 30 minutes, or until it has doubled in size.
4. Heat a poffertjes pan or a small non-stick frying pan over medium-high heat, and grease it with a small amount of butter.
5. Spoon the batter into the pan, filling each indentation about 2/3 of the way full.
6. Cook the poffertjes for 1-2 minutes on each side, or until they are golden brown and slightly puffy.
7. Serve the poffertjes warm with butter and powdered sugar, or your desired toppings.



1. Meng in een kom de bloem, gist, suiker en zout.
2. Voeg de lauwe melk en het ei toe aan de kom en meng alles samen met een garde of elektrische mixer tot je een glad beslag hebt.
3. Dek de kom af met een schone theedoek en laat het beslag rusten voor ten minste 30 minuten, of totdat het in omvang is verdubbeld.
4. Verwarm een poffertjespan of een kleine anti-aanbakpan op matig-hoog vuur en vet het in met een kleine hoeveelheid boter.
5. Schep het beslag in de pan, vul elke uitsparing ongeveer 2/3 vol.
6. Bak de poffertjes 1-2 minuten aan elke kant, of totdat ze goudbruin en lichtjes opgeblazen zijn.
7. Serveer de poffertjes warm met boter en poedersuiker, of met jouw gewenste toppings.

ChatGPT

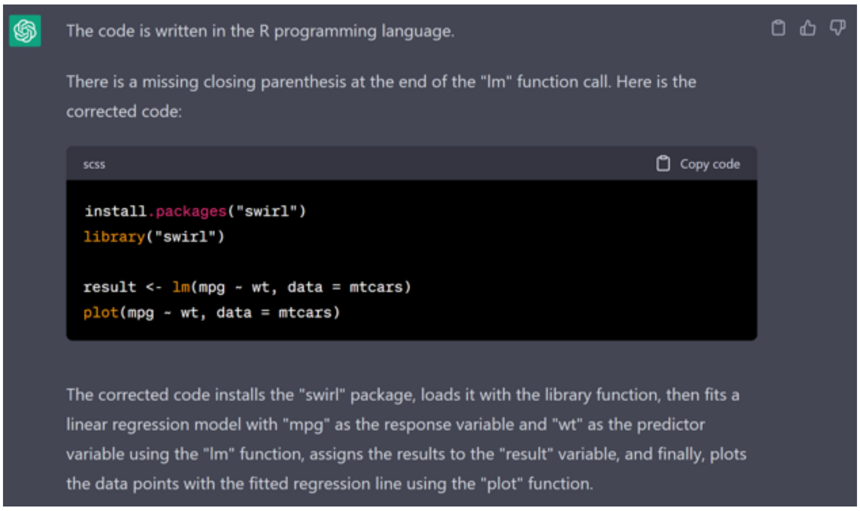


What programming language is this code written in? And find the bug in this code:

```
install.packages(swirl)
 library(swirl)

result <- lm(mpg ~ wt, data = mtcars
plot(mpg ~ wt, data = mtcars)
```

ChatGPT



The code is written in the R programming language.

There is a missing closing parenthesis at the end of the "lm" function call. Here is the corrected code:

```
install.packages("swirl")
library("swirl")

result <- lm(mpg ~ wt, data = mtcars)
plot(mpg ~ wt, data = mtcars)
```

The corrected code installs the "swirl" package, loads it with the library function, then fits a linear regression model with "mpg" as the response variable and "wt" as the predictor variable using the "lm" function, assigns the results to the "result" variable, and finally, plots the data points with the fitted regression line using the "plot" function.

Course setup

- **Course website:** https://adav-course-2024.netlify.app/
- Each week: 1 lecture, 1 practical
  - Lectures: Tuesdays 9.00 - 10.45 OR 11:00 - 12:45
  - Practicals on Thursdays, group-dependent times
- Readings: https://adav-course-2024.netlify.app/course_info/reading/
- This is a **parallel course** for UU wide and Economics ADS minor students
  - Lectures are together
  - Practicals are UU / Economics-specific
  - Exam is together

Lab sessions

- Practical application of the theory learned during the week
- We will use R and R studio
- For the lab sessions, you have to hand in homework which will be graded pass/fail. Deadline two hours before the labs (at 11am).
  - Lab 3 is a ‘do-it-yourself’ lab due to holiday, the full lab needs to be handed in and will account for weeks’ one pass/fail grade
  - Nothing to submit after the labs.
- Answers for each lab can be viewed on the course website on Friday.

Teaching team

Course grading

- Individual homeworks: 20% of your grade
  - A composite score of all pass/fails of lab homework + extra question (week 2)
- Group Assignment (in 3 or 4): 20% of your grade
  - Analyzing data and interactive visualization using data science methods R
  - Part 1 of group assignment is optional (but gives you +0.5 points bonus towards your group assignment grade)
- Exam: 60% of your grade
  - Understanding and comprehension of data analysis methods, theory and skills
  - Digital exam in Remindo, exam is on location

Locations and Blackboard

- Lectures (apart from week 8) and labs are on location
- Lecture on week 8 will be ONLINE
- Attendance is mandatory
- Locations may be differ over the weeks, please consult your time table at https://mytimetable.uu.nl
- Assignment/homeworks submissions via Blackboard

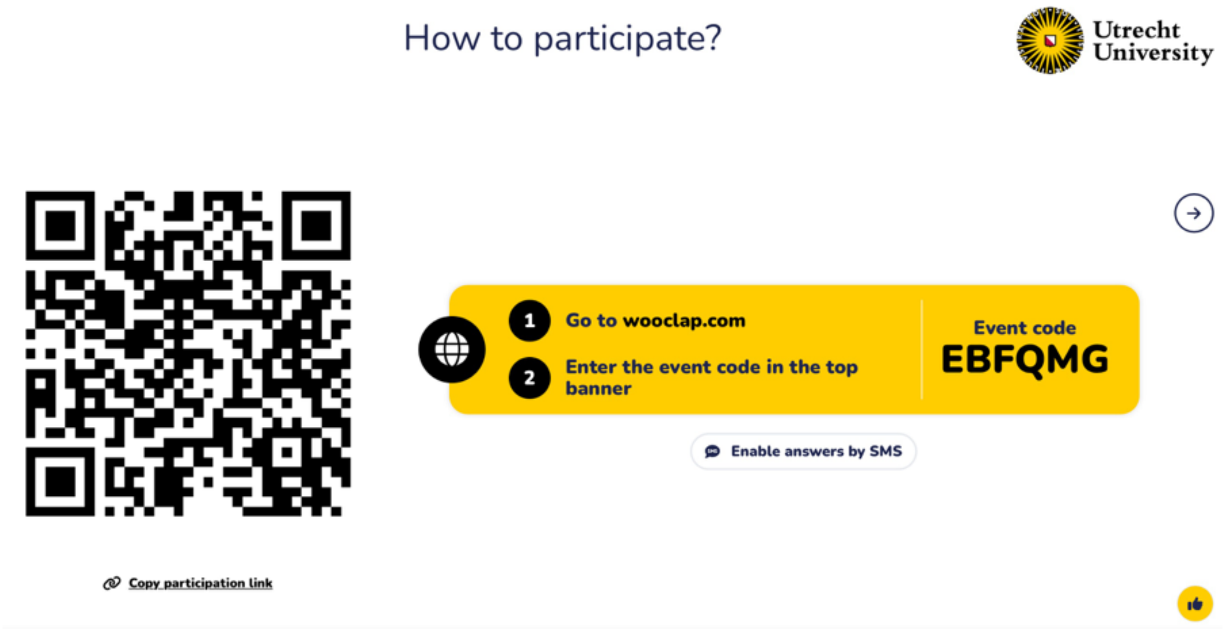
AI Policy

- You can use AI tools to ask about concepts you didn’t fully understand
- You are NOT allowed to copy paste questions to AI tools. Having AI write your homework/assignment constitutes plagiarism.
- In your assignment, you have to acknowledge and specify how you used any AI tool
- It is important to remember that chatGPT and other AI tools are *not a replacement* for your own critical thinking and original ideas. The ultimate goal of this course and any tool used to submit work is to enhance your own learning and understanding, not to undermine it.
- If the teachers suspects any violation of this policy, he/she can ask further questions to you on the assignment/homework during the lab session.

Examples

- Allowed: What is the difference of logistic regression and linear regression?
- Allowed: What is the function to create a histogram in R?
- Not allowed: Generate all the code for the following assignment.

What do you think aDAV is all about? What do you expect to learn in this course?



How to participate?

1. Go to wooclap.com
2. Enter the event code in the top banner

Event code
EBFQMG

Course overview

What	When
Introduction	Week 1
Data visualisation	Week 2
Model fit and cross validation	Week 3
Linear regression for data science	Week 4
Classification	Week 5
Interactive visualisations	Week 6
Tree-based methods	Week 7
Text mining	Week 8
Network Analysis	Week 9
Exam	Week 10

Acknowledgements

Part of the course material for aDAV is based on the courses in Applied Data Science master profile:

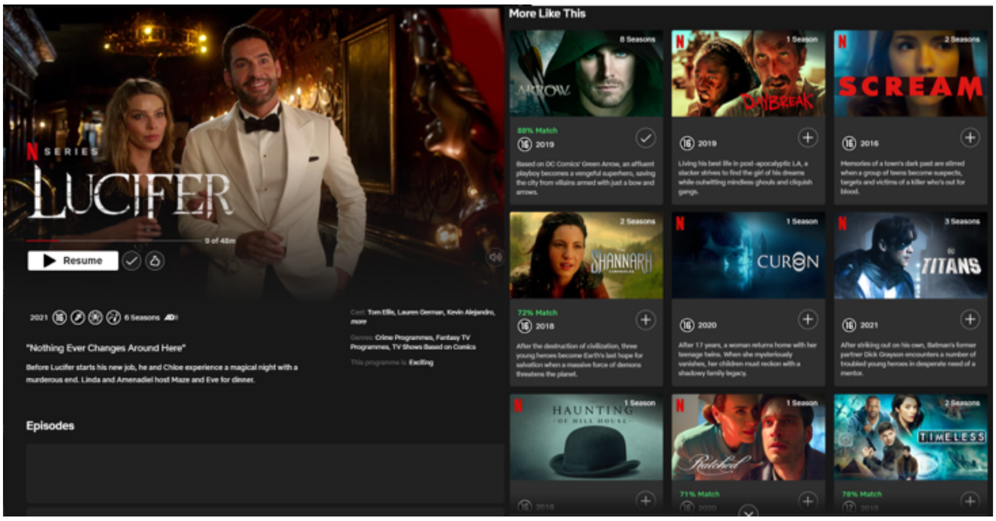
- GSLS ADS profile: https://studyguidelifesciences.nl/profiles/applied-data-science
- GSNS ADS profile: https://students.uu.nl/en/science/applied-data-science/profile-gsns

Points for today

- What are statistical learning / data analysis / data visualisation
- Why do we need them
- Different types of data analysis
- Key data science concepts
  - Supervised vs unsupervised learning
  - Trade-off between model accuracy and interpretability

Learning how to entertain the world

- Machine Learning and Recommendations in **Netflix**
- Netflix Movies and TV Shows dataset from **Kaggle**



"Nothing Ever Changes Around Here"

Before Lucifer starts his new job, he and Chloe experience a magical night with a murderous end. Linda and Amenadiel host Maze and Eve for dinner.

Cast: Tom Ellis, Lauren German, Kevin Alejandro, more

Genres: Crime Programmes, Fantasy TV Programmes, TV Shows Based on Comics

This programme is: Exciting

Based on DC Comics' Green Arrow, an affluent playboy becomes a vengeful superhero, saving the city from villains armed with just a bow and arrows.

Living his best life in post-apocalyptic LA, a slacker strives to find the girl of his dreams while outwitting mindless ghouls and cliquish gangs.

Memories of a town's dark past are stirred when a group of teens become suspects, targets and victims of a killer who's out for blood.

After the destruction of civilization, three young heroes become Earth's last hope for salvation when a massive force of demons threatens the planet.

After 17 years, a woman returns home with her teenage twins. When she mysteriously vanishes, her children must reckon with a shadowy family legacy.

After striking out on his own, Batman's former partner Dick Grayson encounters a number of troubled young heroes in desperate need of a mentor.

How is “data analysis” related to…

- “Data modeling”?
- “Machine learning”?
- “Statistical learning”?
- “Statistics”?
- “Data science”?
- “Data mining”?

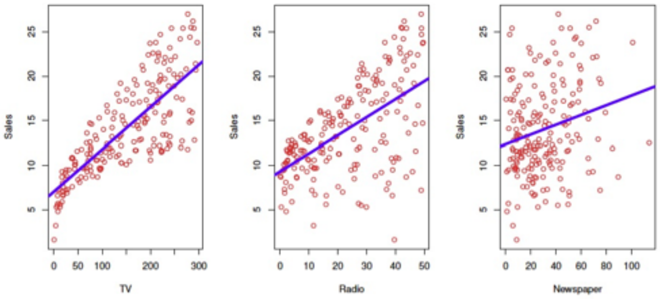
How is “data analysis” related to…

- People from different fields (statistics, computer science, information science, industry) have different goals and standard approaches.
- But they often use the same techniques.
- So each field came up with words that emphasized a different aspect of “data analysis”.
- Because of this history, the terms are not *exact* synonyms.
- But most people can’t tell the difference!

Our emphasis is going to be on **creating results that help you (the analyst) understand the data and make predictions**.

Example

- You are statistical consultant hired by a client to investigate the association between advertising and sales
- The asked you to predict sales increase given an increase in advertising budget
- Advertising budget: Input variables (predictors, features, independent variables), denoted as *X*, *X1* the TV, *X2* the radio, *X3* the newspaper budget
- Sales: Output variable (response, dependent variable) denoted as *Y*



Data science vs empirical research / economics

Oversimplification:

- *Data analysis*: estimating a model *f* to summarize our data, composed of an outcome *Y* and a set of predictors $X : \hat{Y} = \hat{f(X)}$.
- *Data science*: mostly interested in prediction, so in $\hat{Y}$
- *Empirical research / economics*: mostly interested in inference (i.e., explanation): understanding the relationship between **X** and **Y**
  - (so we are interested in the model that generates the predictions)
- Example: use linear regression to summarize our data $\hat{y} = \alpha + \beta 1 X1 + \beta 2 X2 + ... + \epsilon$
  - Data science interested in $\hat{y}$, while in Empirical research / economics interested in β's (e.g., the magnitude, statistical significance)

Some example questions

- Who will win the election?
- Is the climate changing?
- Why are women underrepresented in STEM degrees?
- What is the best way to prevent heart failure?
- What kind of topics are popular on Twitter?
- How many people visited in a given time period your website?
- …

Data analysis: goals

- **Description**
  What happened?
- **Explanation/Diagnostic**
  Why did / does something happen?
- **Prediction**
  What is likely to happen in the future?
- **Prescription**
  What shall we do?

Some example questions

- Who will win the election? **Prediction**
- Is the climate changing? **Description / Prediction**
- Why are women underrepresented in STEM degrees? **Explanation/Diagnostic**
- What is the best way to prevent heart failure? **Prescription**
- What kind of topics are popular on Twitter? **Description**
- How many people visited in a given time period your website? **Description**
- …

Data analysis: modes

- **Exploratory**
  - Following your gut (or other criteria) to some interesting results
  - Investigate datasets open-mindedly to uncover patterns, and understand data
  - Generate new insights
  - Create new hypotheses
- **Confirmatory**
  - Confirmatory data analysis tests specific hypotheses and confirms relationships with targeted analytical techniques
  - Validate or refute hypotheses
  - Analysis predefined

Examples of different kinds of data analysis

Exploratory	Confirmatory
EDA	Hypothesis Testing
Unsupervised learning	Supervised learning
Correlation analysis	Causal modeling

Examples of different kinds of data analysis

Exploratory	Confirmatory
EDA	Hypothesis Testing
Unsupervised learning	Supervised learning
Correlation analysis	Causal modeling

Explorative data analysis

- Describing interesting patterns: use graphs, summaries, to understand subgroups, detect anomalies (“outliers”), understand the data
- Examples: boxplot, barplots, histograms, scatterplots…

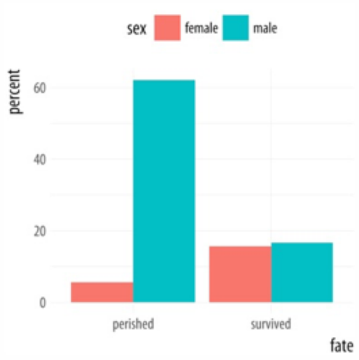


Figure 4.20: Survival on the Titanic, by Sex.

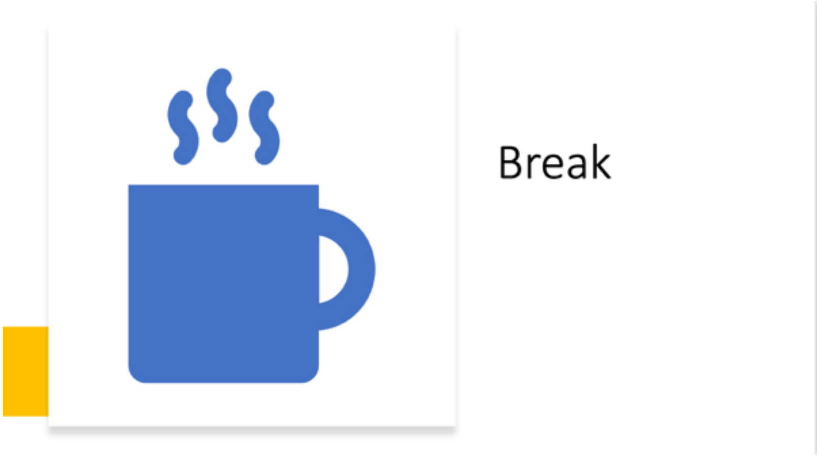
Examples of different kinds of data analysis

Exploratory	Confirmatory
EDA	Hypothesis Testing
Unsupervised learning	Supervised learning
Correlation analysis	Causal modeling

Examples of different kinds of data analysis

- Confirmatory analysis
- You are a scientist and you want to determine if a new drug treatment is effective in reducing blood pressure compared to a placebo.
- Theory testing, hypothesis: new treatment better than placebo
- Analysis can be defined in advance: which outcome variables, how to sample from the population, which method?

10-minute break

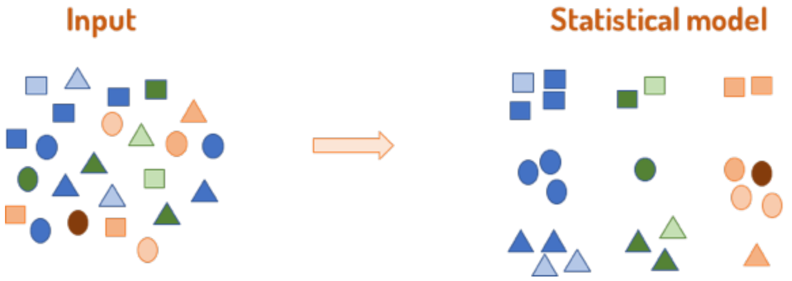


Break

Examples of different kinds of data analysis

Exploratory	Confirmatory
EDA	Hypothesis Testing
Unsupervised learning.	Supervised learning
Correlation analysis	Causal modeling

Unsupervised learning

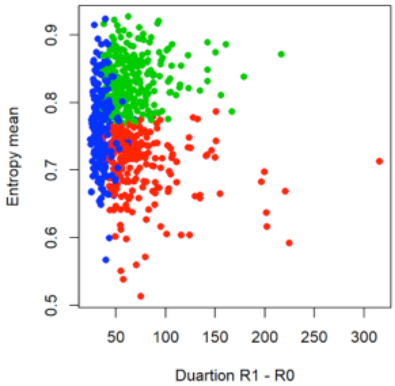


- *Inputs*, but *no outputs*. Try to learn structure and relationships from these data, like detecting unobserved groups (clusters) from the data.
- A retail store has customer data such as age, income, number of purchases, and average purchase amount.
- By applying unsupervised learning techniques like clustering to the data, the retail store can gain valuable insights into customer segmentation, behavior patterns, and preferences.

Unsupervised learning

Clustering:

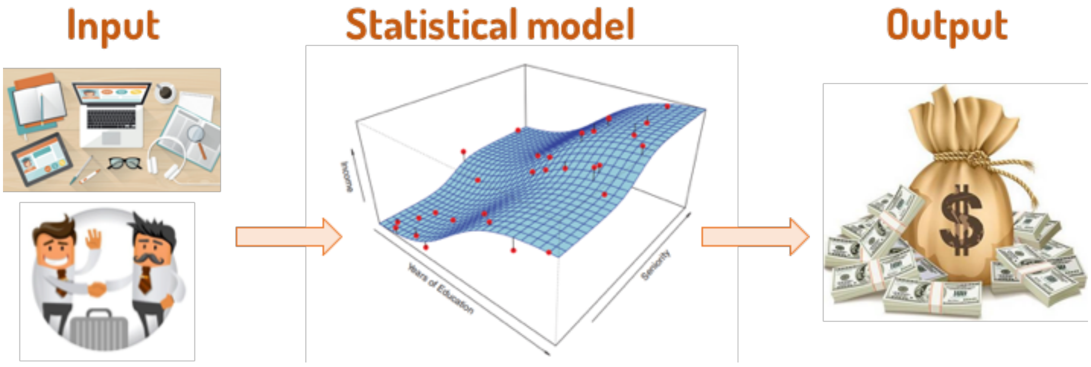
- K-means clustering
- Hierarchical clustering



Examples of different kinds of data analysis

Exploratory	Confirmatory
EDA	Hypothesis Testing.
Unsupervised learning	Supervised learning
Correlation analysis	Causal modeling

Supervised learning

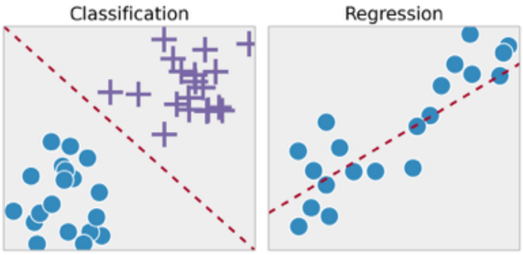


Building a statistical model for *predicting* / estimating an *output* based on one or more *inputs*.

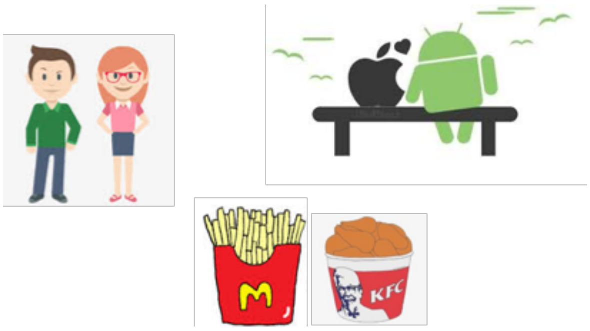
Supervised learning

Most widely used machine learning methods are supervised

- Spam classifiers of e-mail
- Face recognizers over images
- Medical diagnosis systems for patients

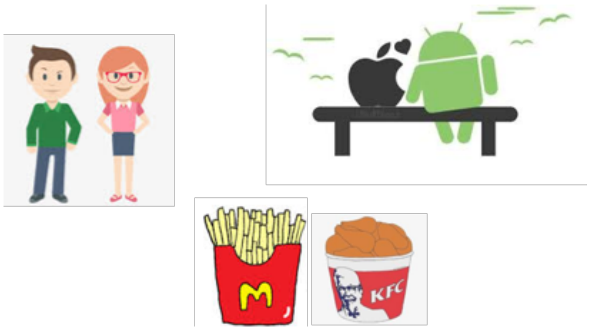


Supervised learning: classification vs regression



Classification: predict to which category an observation belongs (qualitative outcomes)

Supervised learning: classification vs regression



Classification: predict to which category an observation belongs (qualitative outcomes)

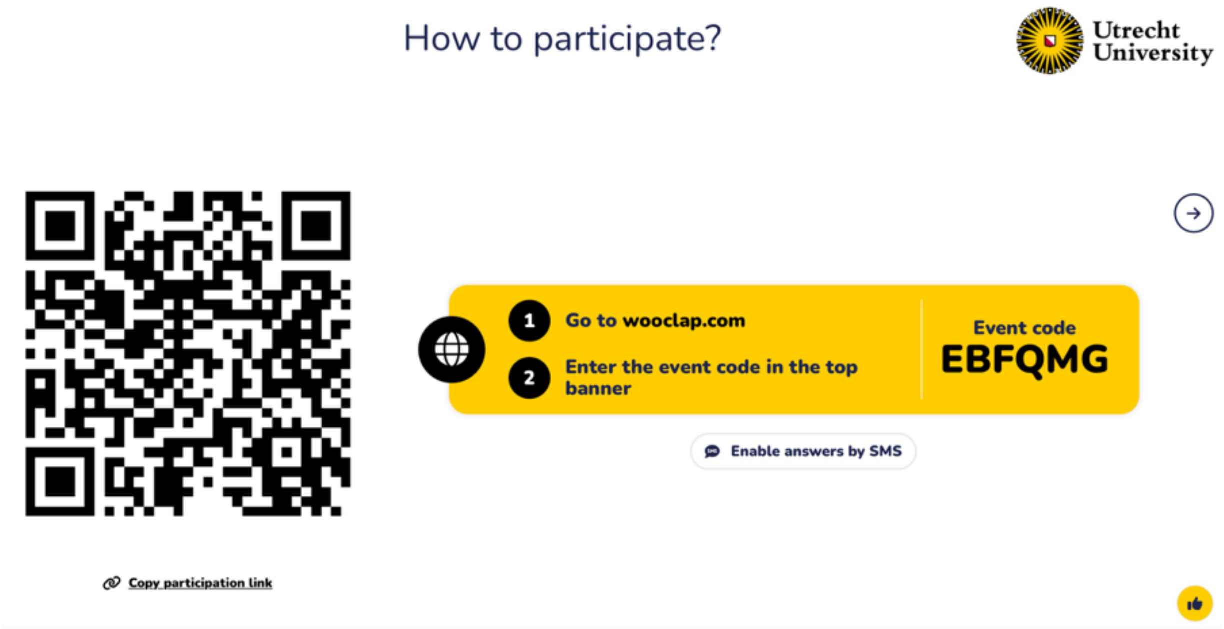
Regression: predict a quantitative outcome

Supervised learning

Methods such as:

- Linear Regression -> week 4
- Logistic Regression -> week 5
- Decision trees / random forests -> week 7
- Support vector machines -> beyond the scope of this course
- Neural networks -> beyond the scope of this course
- and much more

Team up and write an example of supervised or unsupervised learning from your field of study or interest.



How to participate?

Utrecht University

1 Go to wooclap.com

2 Enter the event code in the top banner

Event code

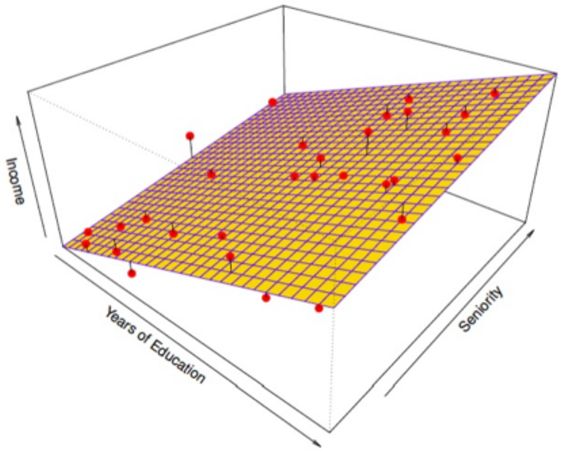
EBFQMG

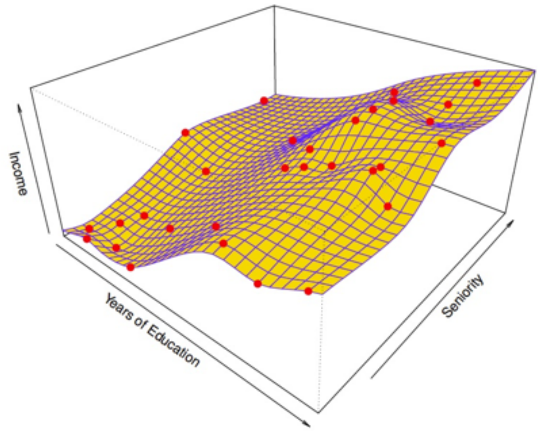
Model accuracy vs interpretability

Image from: https://blog.siftscience.com/2017/machine-learning-black-box/

Model accuracy vs interpretability

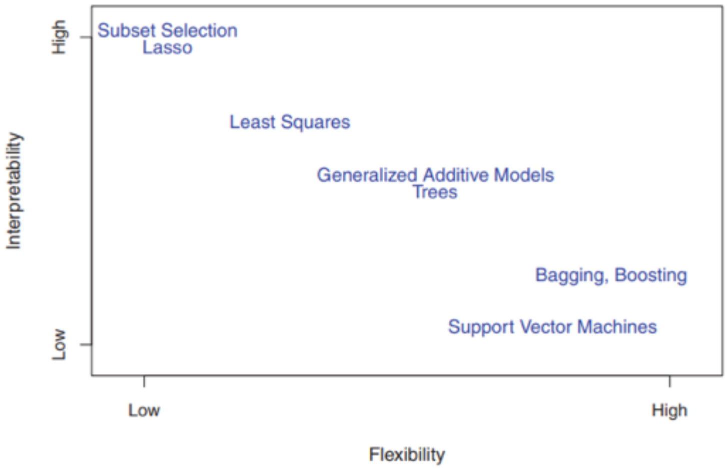
- Some models are less flexible, they can produce a small range of shapes to estimate *f*
- Other models are more flexible and allow curve relationships





Model accuracy vs interpretability

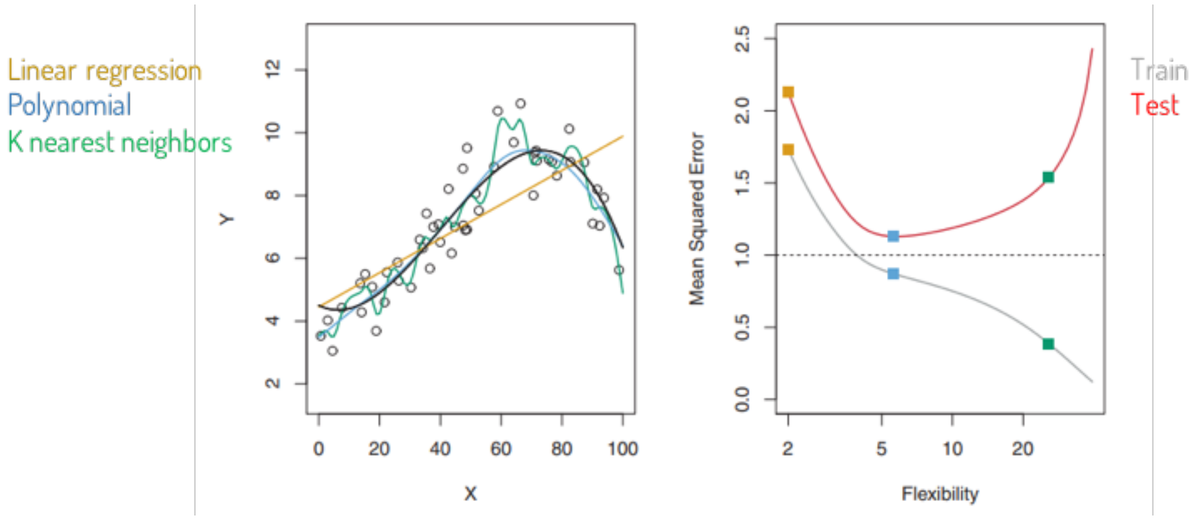
- Why would we ever choose to use a more restrictive method over a flexible approach?



Zoom in: model accuracy

- Measure of model accuracy: mean squared error:
  - $MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2$
  - $MSE = \frac{1}{n} \sum_{i=1}^{n} (outcome_i - predicted_i)^2$
- Hence, MSE represents how wrong our prediction on average is.
- In machine / statistical learning, we obtain the MSE for a *training set* and a *test set*.

Zoom in: model accuracy



Visualisation

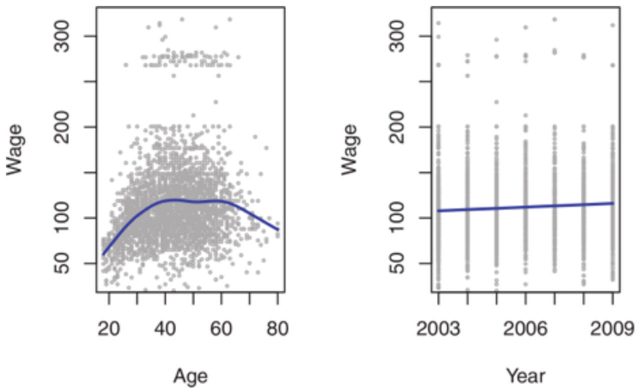
“The simple graph has brought more information to the data analyst’s mind than any other device.” — John Tukey

- Exploratory Data Visualization (week 2)
- Interactive visualization (week 6)

Some examples with visualisations

1. How wages vary over different groups *(James et al. p. 2)*

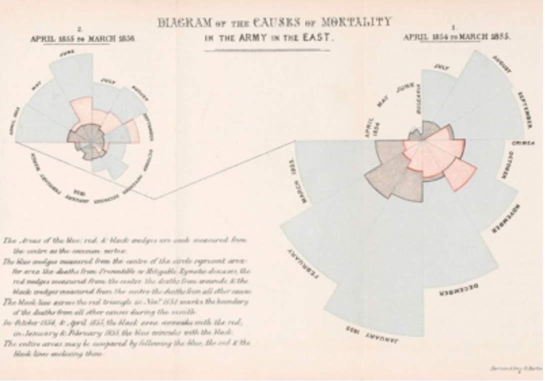
When thinking about important topics, such as whether to stay in school, it helps to know that more highly educated people tend to earn more, but also that there is no difference for top earners.



Some examples with visualisations

2. Healthcare during the Crimean War

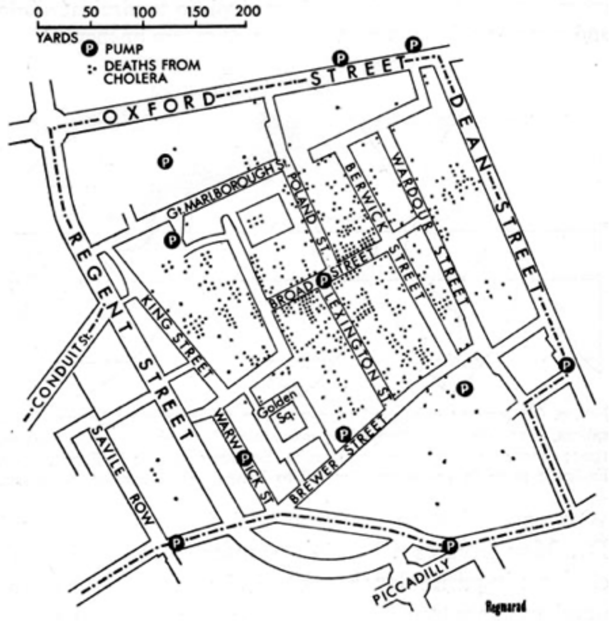
Florence Nightingale’s Rose Diagram, created in 1858, is a an example of data visualization that played a significant role for improvements in healthcare during the Crimean War.



Some examples with visualisations

3. London Cholera outbreak of 1854

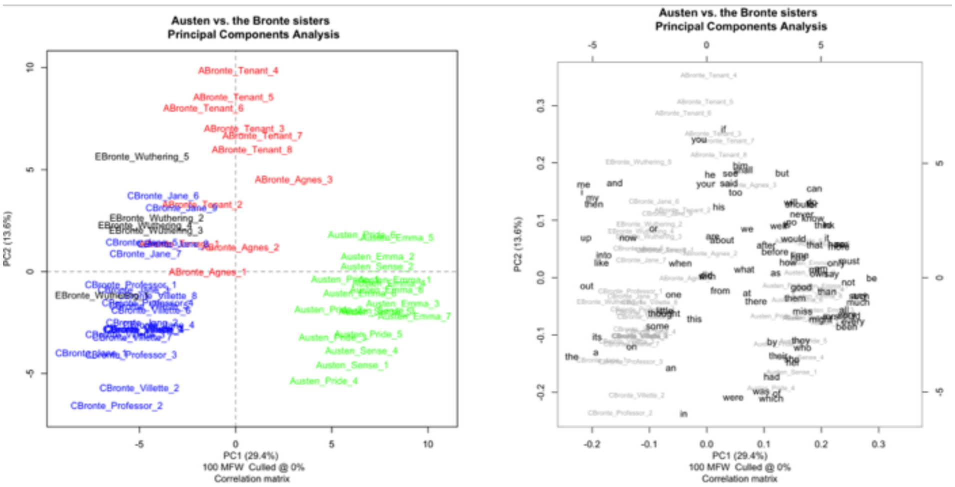
Before John Snow, people thought “miasma” caused cholera and they fought it by airing out the house. It was not clear whether this helped or not, but people thought it must because “miasma” theory said so.



Some examples with visualisations

4. Telling Bronte sisters from Jane Austen

(Eder, M. et al. (2016). Stylometry with R: A Package for Computational Text Analysis. The R Journal 8:1)



Some examples with visualisations

4. Telling Bronte sisters from Jane Austen

Scholars fight over who wrote various songs (Wilhelmus), treatises (Caesar), plays (Shakespeare), etc., with shifting arguments. By counting words, we can sometimes identify the most likely author of a text, and we can explain exactly why we think that is the right answer.

Why we need data analysis and visualisation

- Data analysis and the accompanying visualisations yield insights and solves problems that could not be solved without them.
- On some level, humans do nothing but analyze data;
- They may not do it consistently, understandably, transparently, or correctly, however;
- aDAV help us process more data, and can keep us honest;
- aDAV can also exacerbate our biases when we are not careful;
- Data visualisations make people understand the data and remember the results.

Conclusion

Data analysis..

- …is usually part of a broader “pipeline”, may involve question formulation, budgeting, data collection, communication / dissemination, archiving, engineering, etc. – not covered here!
- …is often aided by visualisation;
- …can have different goals;
- …can have different modes;
- …can be useful, fun, but also dangerous;

In short, it makes sense to learn more about the background and ideas behind aDAV techniques!

Next class

- Lab 1 is on Thursday
- Let your teacher know if you can’t attend
- Submit the homework before 11:00am which will account for weeks’ one pass/fail grade
- Next lecture: next Tuesday about Data visualisation

Read the **Course information** and **Preparation** on the course website.

Have a nice day!