Data Analytics Foundations

Online Workshop 3
Know Your Data

Questions from Module 2?

Attribute and data types

What attribute or data types are the following (you may need to look some of them up):

- colour: values {‘red’, ‘green’, ‘blue’, ‘yellow’}
- greyscale: values {‘white’, ‘light grey’, ‘medium grey’, ‘dark grey’, ‘black}
- grade: ‘h’, ‘d’, ‘c’, ‘p’, ‘z’
- FoR (Field of Research) codes a1 s2
- phone number
- Postcode
- age as a number
- age: ‘child’, ‘teenager’, ‘adult’, ‘middle-aged’, ‘elderly’
- location (longitude, latitude)
- monthly house sale prices in Sydney 100K

Attribute and data types

What attribute or data types are the following (you may need to look some of them up):

- colour: values {‘red’, ‘green’, ‘blue’, ‘yellow’} nominal
- greyscale: values {‘white’, ‘light grey’, ‘medium grey’, ‘dark grey’, ‘black} ordinal
- grade: ‘h’, ‘d’, ‘c’, ‘p’, ‘z’
- grade: ‘h’, ‘d’, ‘c’, ‘p’, ‘z’, ‘w’, ‘y’
- FoR (Field of Research) codes nominal
- phone number (nominal)
- Postcode (nominal)
- age as a number (ratio)
- age: ‘child’, ‘teenager’, ‘adult’, ‘middle-aged’, ‘elderly’ (ordinal)
- location (longitude, latitude) (nominal)
- monthly house sale prices in Sydney (ratio)

KNIME – attribute types

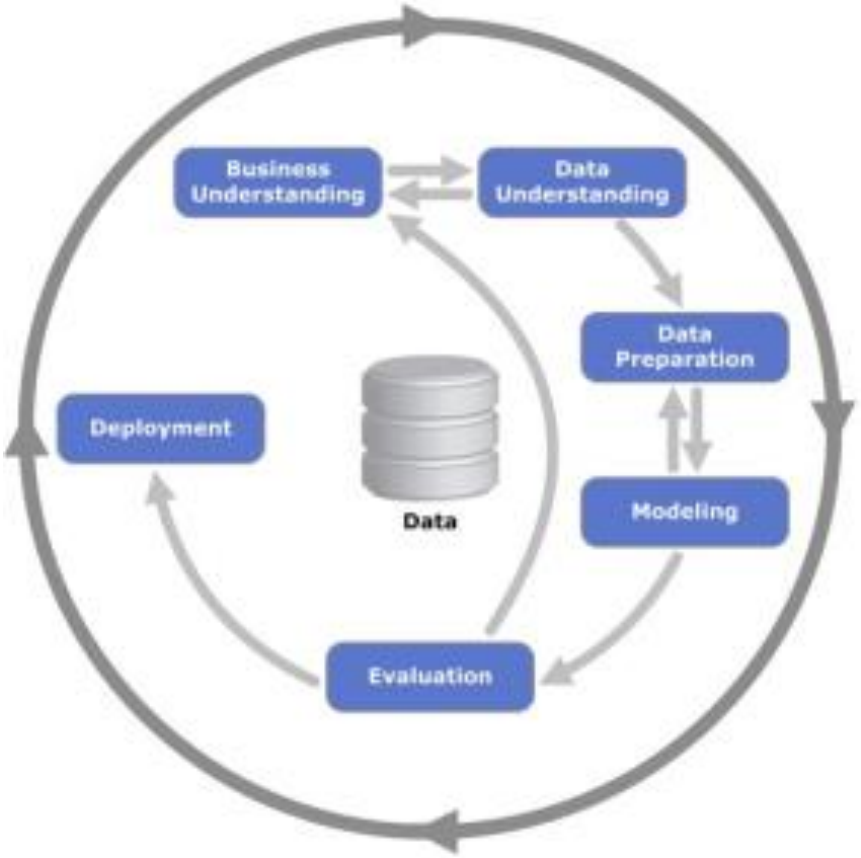
- Download imports-mod.csv from the Workshop 3 page in Canvas.
- Import it into KNIME using CSV Reader or File Reader and look at the file table.
- What attribute type is symboling? Use a Statistics node to look at its distribution.
- Use Number To String to convert symboling to a string. Use another Statistics node to look at its distribution. What has changed?
- Use String To Number to convert price to a number?
- Use String To Date to convert date to a Date format (yyyy-MM-dd)?

KNIME – attribute types

- Download Telco-Customer-Churn.csv from the Workshop 3 page in Canvas.
- Import it into KNIME using CSV Reader and look at the file table.
- Use a Statistics node to look at its distribution.
- What value can be taken from analyzing this dataset?

CRISP-DM (Q&A)

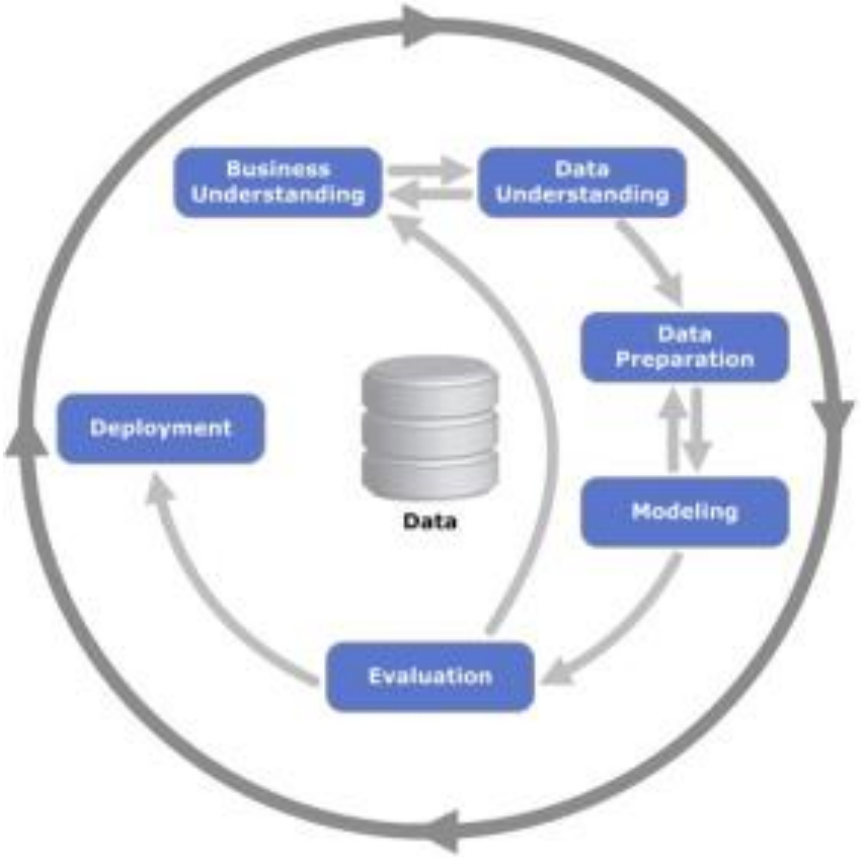
- Client: Me ☺
- Data source: student data (student ID, name, course, current and completed subjects, marks and grades), course data (subjects, prerequisites, undergrad/postgrad/other), …
- Problem: Client wants to better support the students in the class.
- Using the CRISP-DM framework, work out tasks for each of the six phases.



Source: Kenneth Jensen / Wikimedia Commons / Public Domain

CRISP-DM

- Phase1: setup the data mining goal “to predict the students on risk”
- Phase 2: describe data, explore and check quality
- P3: Clean (remove noise, missing values, outliers), integrate, normalize
- P4: Create prediction model
- P5: Test
- P6: Deploy



Source: Kenneth Jensen / Wikimedia Commons / Public Domain