CS2D7 Data Visualisation
Dr Claire Rocks & Dr Richard Kirk

July 2022
Lecture 2

Acquire, Parse, Filter, Mine
Today

• **Data acquisition:** sourcing and gathering the raw material

• **Data examination:** getting familiar with the key properties and condition of the data

• **Data transformation:** refining the data through modification and consolidation

• **Data exploration:** Using exploratory analysis and research methods to discover insights
Today

• Data acquisition: sourcing and gathering the raw material

• Data examination: getting familiar with the key properties and condition of the data

• Data transformation: refining the data through modification and consolidation

• Data exploration: Using exploratory analysis and research methods to discover insights
Data Acquisition

- Supplied – getting it from someone else
- Downloaded – your workplace might generate data, there are other organizations offering publicly accessible data
- Web scraping – extracting data from webpages
- APIs – allows you to programmatically extract data from sources
- Gathering primary data e.g. through participant questionnaires or measurement devices
- Data foraging – combining multiple sources of disparate data
Ethical Principles

- In groups of 2/3 – draw up your own Ethical Principles based on your reading of *Aspects of Data Ethics in a Changing World: Where Are We Now?*
Possibility driven or purpose driven?

Should data collection be possibility driven or purpose driven? Why?
Aspects of Data Ethics in a Changing World: Where are we now?

• Ethical principles are not necessarily universal. We might note, for example, that the United Kingdom is more bound by ethical considerations of data collection and use than some other countries. This could put us at a technological disadvantage. “Ethical drag” might mean a lack of agility in the face of evolving data technologies.

• To what extent do you agree with this statement?
Today

• Data acquisition: sourcing and gathering the raw material

• **Data examination: getting familiar with the key properties and condition of the data**

• Data transformation: refining the data through modification and consolidation

• Data exploration: Using exploratory analysis and research methods to discover insights
Data Examination

• “Before you know what meal to cook, you need to know what ingredients you have, how much and in what condition” Kirk

• This is all about getting acquainted with the physical properties of the data
  • Types of data
  • Size
  • Range of values
Normalised tabular data offers the most detailed granular form of the data organised by variables and items.

Cross-tabulated data presents a *summarized form* of the data, displaying values that are the result of operations such as group totals, max and min.

Normalised data is sometimes called ‘tall and thin’, cross tabulated data ‘short and fat’.

<table>
<thead>
<tr>
<th>First Name</th>
<th>Last Name</th>
<th>House</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harry</td>
<td>Potter</td>
<td>Gryffindor</td>
<td>1998</td>
</tr>
<tr>
<td>Ron</td>
<td>Weasley</td>
<td>Gryffindor</td>
<td>1998</td>
</tr>
<tr>
<td>Hermione</td>
<td>Granger</td>
<td>Gryffindor</td>
<td>1998</td>
</tr>
<tr>
<td>Albus</td>
<td>Dumbledore</td>
<td>Gryffindor</td>
<td>1989</td>
</tr>
<tr>
<td>Rubeus</td>
<td>Hagrid</td>
<td>Gryffindor</td>
<td>1997</td>
</tr>
<tr>
<td>Severus</td>
<td>Snape</td>
<td>Slytherin</td>
<td>1978</td>
</tr>
<tr>
<td>Tom</td>
<td>Riddle</td>
<td>Slytherin</td>
<td>1945</td>
</tr>
<tr>
<td>Sirius</td>
<td>Black</td>
<td>Pure-blood</td>
<td>1978</td>
</tr>
<tr>
<td>Draco</td>
<td>Malfoy</td>
<td>Pure-blood</td>
<td>1998</td>
</tr>
<tr>
<td>Fred</td>
<td>Weasley</td>
<td>Gryffindor</td>
<td>1996</td>
</tr>
<tr>
<td>Remus</td>
<td>Lupin</td>
<td>Pure-blood</td>
<td>1978</td>
</tr>
<tr>
<td>George</td>
<td>Weasley</td>
<td>Pure-blood</td>
<td>1916</td>
</tr>
<tr>
<td>Neville</td>
<td>Longbottom</td>
<td>Pure-blood</td>
<td>1998</td>
</tr>
<tr>
<td>Arthur</td>
<td>Weasley</td>
<td>Pure-blood</td>
<td>1968</td>
</tr>
<tr>
<td>Ginny</td>
<td>Weasley</td>
<td>Pure-blood</td>
<td>1997</td>
</tr>
<tr>
<td>Minerva</td>
<td>McGonagall</td>
<td>Pure-blood</td>
<td>1954</td>
</tr>
<tr>
<td>Molly</td>
<td>Weasley</td>
<td>Pure-blood</td>
<td>1996</td>
</tr>
<tr>
<td>Dolores</td>
<td>Umbridge</td>
<td>Slytherin</td>
<td>1945</td>
</tr>
<tr>
<td>Alastor</td>
<td>Moody</td>
<td>Pure-blood</td>
<td>1978</td>
</tr>
<tr>
<td>Vernon</td>
<td>Dursley</td>
<td>Pure-blood</td>
<td>1916</td>
</tr>
<tr>
<td>Cornelius</td>
<td>Fudge</td>
<td>Pure-blood</td>
<td>1998</td>
</tr>
<tr>
<td>Peter</td>
<td>Pettigrew</td>
<td>Gryffindor</td>
<td>1978</td>
</tr>
<tr>
<td>Dudley</td>
<td>Dursley</td>
<td>Pure-blood</td>
<td>1945</td>
</tr>
<tr>
<td>Horace</td>
<td>Slughorn</td>
<td>Slytherin</td>
<td>1945</td>
</tr>
<tr>
<td>Percy</td>
<td>Weasley</td>
<td>Pure-blood</td>
<td>1994</td>
</tr>
<tr>
<td>Luna</td>
<td>Lovegood</td>
<td>Ravenclaw</td>
<td>1919</td>
</tr>
<tr>
<td>Cedric</td>
<td>DIGGORY</td>
<td>Pure-blood</td>
<td>1945</td>
</tr>
<tr>
<td>Petunia</td>
<td>Dursley</td>
<td>Pure-blood</td>
<td>1916</td>
</tr>
<tr>
<td>Kreacher</td>
<td>Books</td>
<td>Pure-blood</td>
<td>1972</td>
</tr>
<tr>
<td>Bill</td>
<td>Weasley</td>
<td>Pure-blood</td>
<td>1989</td>
</tr>
<tr>
<td>Bertha    Sr.</td>
<td>Granger</td>
<td>Pure-blood</td>
<td>1945</td>
</tr>
<tr>
<td>Argus</td>
<td>Filch</td>
<td>Pure-blood</td>
<td>1900</td>
</tr>
<tr>
<td>Viktor</td>
<td>Krum</td>
<td>Durmstrang</td>
<td>1945</td>
</tr>
<tr>
<td>Gilderoy</td>
<td>Lockhart</td>
<td>Ravenclaw</td>
<td>1982</td>
</tr>
<tr>
<td>Sybill</td>
<td>Treanomy</td>
<td>Ravenclaw</td>
<td>1982</td>
</tr>
<tr>
<td>Fleur</td>
<td>Dalcor</td>
<td>Beauxbatons</td>
<td>1945</td>
</tr>
<tr>
<td>Lucius</td>
<td>Malfoy</td>
<td>Pure-blood</td>
<td>1972</td>
</tr>
<tr>
<td>Ludo</td>
<td>Bagman</td>
<td>Gryffindor</td>
<td>1998</td>
</tr>
<tr>
<td>Nymphadora</td>
<td>Tonks</td>
<td>Pure-blood</td>
<td>1991</td>
</tr>
<tr>
<td>Gregory</td>
<td>Goyle</td>
<td>Pure-blood</td>
<td>1998</td>
</tr>
<tr>
<td>Vincent</td>
<td>Crabbe</td>
<td>Pure-blood</td>
<td>1998</td>
</tr>
<tr>
<td>Bellatrix</td>
<td>Lestrange</td>
<td>Pure-blood</td>
<td>1998</td>
</tr>
<tr>
<td>Cho</td>
<td>Chang</td>
<td>Ravenclaw</td>
<td>1997</td>
</tr>
<tr>
<td>Dean</td>
<td>Thomas</td>
<td>Pure-blood</td>
<td>1989</td>
</tr>
<tr>
<td>Oliver</td>
<td>Wood</td>
<td>Pure-blood</td>
<td>1994</td>
</tr>
<tr>
<td>Hedwig</td>
<td>Books</td>
<td>Pure-blood</td>
<td>1982</td>
</tr>
<tr>
<td>Rita</td>
<td>skeeter</td>
<td>Pure-blood</td>
<td>1916</td>
</tr>
<tr>
<td>Salazar</td>
<td>Finnigan</td>
<td>Pure-blood</td>
<td>1972</td>
</tr>
<tr>
<td>James</td>
<td>Potter</td>
<td>Gryffindor</td>
<td>1998</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Status</th>
<th>All</th>
<th>Book canon</th>
<th>Harry’s era</th>
<th>Gryffindor</th>
<th>Hufflepuff</th>
<th>Ravenclaw</th>
<th>Slytherin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pure-blood</td>
<td>97</td>
<td>56</td>
<td>20</td>
<td>14</td>
<td>2</td>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td>Half-blood</td>
<td>35</td>
<td>31</td>
<td>10</td>
<td>4</td>
<td>1</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>Muggle-born</td>
<td>11</td>
<td>10</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Squib</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>146</td>
<td>99</td>
<td>34</td>
<td>30</td>
<td>7</td>
<td>9</td>
<td>24</td>
</tr>
</tbody>
</table>
‘Tall and thin’ or ‘Short and fat’

• Working with data in cross tabulated form, without access to the normalised data, reduces the possibilities in terms of analysis
• Certain tools need data to be shaped in a certain way
Levels of data or Scales of measurement

• Scales of measurement is how variables are defined and categorised.
• Developing a thorough understanding of the data types
• The different types of data will have a major influence on a number of aspects of your visualization e.g.
  • The statistical analysis methods you can use
  • The editorial perspectives you can pursue
  • The chart types you can use
  • The appropriateness of colour associations
Types of Data

Quantitative:
- Discrete: Whole numbers that can’t be broken down, such as a number of items
- Continuous: Numbers that can be broken down, such as height or weight
  - Interval: Numbers with known differences between variables, such as time
  - Ratio: Numbers that have measurable intervals where difference can be determined, such as height or weight

Qualitative:
- Nominal: Data used for naming variables, such as hair colour
- Ordinal: Data used to describe the order of values, such as 1 = happy, 2 = neutral, 3 = unhappy

https://studyonline.unsw.edu.au/blog/types-of-data
Discrete or Continuous

- Discrete data is data that has clear spaces between values. Classifying measurements with no in-between e.g. number of seats in a cinema
- Continuous data is data that falls in a constant sequence e.g. height, weight, temperature.
- Discrete data is countable while continuous data is measurable.
T-NOIR classification (Stanley Stevens)

• Textual
• Nominal
• Ordinal
• Interval
• Ratio

Precision goes up
Textual data

- Qualitative data
- Characteristically human
- Usually unstructured
- Typically requires transformation to extract properties and relational information e.g. frequency of key words or NLP to derive sentiment classifications
Nominal Data

- Qualitative data
- Divides data into mutually exclusive labelled classes
- Nominal data can be numeric but is often text-based
- There is potential for hierarchical relationships which will be important when deciding how to represent the data
Ordinal Data

- Qualitative data
- Classifies data into categories which have a natural order or rank
**Interval Data**

- Quantitative data
- Measured along a numerical scale that has equal intervals along adjacent values - introduces precise and continuous intervals, e.g. temperature measurements or the pH scale.
- Lacks a true zero – this is just an established scale position – a temp of 0deg does not mean no temperature – it is a quantitative scale for measuring relative temperature

<table>
<thead>
<tr>
<th>IQ Test Score Bell Curve</th>
</tr>
</thead>
<tbody>
<tr>
<td>By: <a href="http://www.free-iqtest.net">http://www.free-iqtest.net</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HIGH IQ PEOPLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blaise Pascal - 195</td>
</tr>
<tr>
<td>Bobby Fischer - 187</td>
</tr>
<tr>
<td>René Descartes - 180</td>
</tr>
<tr>
<td>Charles Darwin - 165</td>
</tr>
<tr>
<td>Wolfgang Amadeus Mozart - 165</td>
</tr>
<tr>
<td>Albert Einstein - 161</td>
</tr>
<tr>
<td>Nicolaus Copernicus - 160</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>BMI</th>
<th>55</th>
<th>70</th>
<th>85</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mentally inadequate</td>
<td>2.3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low intelligence</td>
<td>13.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Below average</td>
<td>34.1%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

![Image showing IQ test score bell curve with high IQ people and intervals for BMI]
Ratio data

- Quantitative data
- Categorises and ranks data
- Uses continuous intervals (like nominal data) but has a true zero (when variable value = 0, there is none of this variable)
- This is the one you are most likely to encounter
Measurement Scales

• Most data will exist on a linear scale

• Some are based on logarithmic scales e.g. measuring sound in dB or earthquake magnitudes in the Richter Scale.

• A logarithmic scale is a way of displaying numerical data over a very wide range of values in a compact way - the units on an axis are powers, or logarithms, of a base number, usually 10
Shape and Size of the data

• For quantitative data, what are the max and min values?
• In what format are the values presented? How many decimal places?
• For a categorical variable, how many different values are there?
• For textual data, what is the max and min character length or word count?
• You might also want to do some statistical analysis
• The size and shape of your data will have a strong bearing on the design of your visualisation
Some Useful Statistical Methods

- **Frequency distribution**: applied to quantitative values provides information about the shape of the distribution of the values
- **Finding the mean**: the average value
- **Finding the median**: the middle value if all quantities are arranged smallest to largest
- **Finding the mode**: the most common value
- **Frequency counts**: applied to categorical values to understand the frequency of different instances
- **Measurements of spread** e.g. maximum, minimum, range, percentiles and standard deviation
Quality of the data

• Missing values? Empty or null?
• Erroneous values?
• Inconsistencies e.g. capitalization, units of measurement, value formatting, date formatting
• Leading or trailing spaces
• Duplicates
• Expired values
Representative?

- What is missing?
- Does it represent genuine observations or is it influenced by the collection method?
- Does it represent the entirety? Or a sample?
Today

• Data acquisition: sourcing and gathering the raw material
• Data examination: getting familiar with the key properties and condition of the data
• **Data transformation: refining the data through modification and consolidation**
• Data exploration: Using exploratory analysis and research methods to discover insights
Data Transformation

• How to get the data into shape and fit for purpose?
• Likely an iterative process
• Any modifications should be noted/explained
• Three types of transformation
  • Cleaning
  • Creating
  • Consolidating
Cleaning

• Find and replace/remove
• Remove what you don’t need
• Pandas offers a number of methods to help
• You might also want to open data in Excel and clean it up there
• Convert the layout and format for visualization/ importing the data into DataFrame
• Keep copies and backups. Preserve unaltered data so that you can easily return to it if needed.
Creating (1)

• Creating new data from existing values e.g.
  • Creating percentage calculations based on existing quantities
  • Creating a rolling 12 monthly total
  • Using start and end dates to work out duration
  • Converting absolute values into per capita values
  • Using logic based formula to create new categories e.g. Adult or Child
Creating (2)

• For text based data
  • Identify key words or themes and convert into categories
  • Calculate the frequency of certain words
  • Analyse the attributes of text e.g. word count, length of text, duration of read, number of sentences, number of paragraphs
  • Flag up instances of cases (X mentioned) and the existence of relationships (A and B mentioned, C always mentioned before D)
  • Use NLP techniques to determine sentiments, classify words (verb, noun etc) or sentence structures
[Tue 1 March 2005, 10:22 AM] Luigi "add post"
[Tue 1 March 2005, 10:26 AM] Orazio "view discussion"
[Tue 1 March 2005, 11:02 AM] Luigi "add post"
[Tue 1 March 2005, 02:02 PM] Enzo "view discussion"
[Tue 1 March 2005, 02:04 PM] Enzo "view discuss

<table>
<thead>
<tr>
<th>User</th>
<th>Read</th>
<th>Posted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enzo</td>
<td>90</td>
<td>10</td>
</tr>
<tr>
<td>Giorgio</td>
<td>134</td>
<td>20</td>
</tr>
<tr>
<td>Luigi</td>
<td>89</td>
<td>3</td>
</tr>
<tr>
<td>Michele</td>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td>Orazio</td>
<td>117</td>
<td>13</td>
</tr>
</tbody>
</table>
Consolidating

- Source and introduce additional data
- Expand: broaden the values of data we are working with e.g. by adding additional data e.g. Country codes from country names or local authority information from Post Code
- Append e.g. by adding the most up to date data
Today

• Data acquisition: sourcing and gathering the raw material
• Data examination: getting familiar with the key properties and condition of the data
• Data transformation: refining the data through modification and consolidation
• Data exploration: Using exploratory analysis and research methods to discover insights
Data Exploration

• This is all about interrogating the data to see what insights and understanding you get
• Broadening your awareness of the things you could show
• Uses both statistical and visual techniques
Exploratory Data Analysis (EDA)

- Exploratory data analysis has been promoted by John Tukey since 1970 to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments.
- ‘Exploratory data analysis is an attitude, a flexibility, and a reliance on display, not a bundle of techniques’ (John Tukey)
- Not always needed

https://en.wikipedia.org/wiki/Exploratory_data_analysis
Technical, Practical and Conceptual capabilities

• Your instinct
  • Technology plays a part
  • Recognise and pursue potential threads

• Reasoning
  • You cannot do everything and reasoning can help limit the exploration
  • Deductive (follow a curiosity) and Inductive (play around)

• Research
  • Learn as much as you can about the subject and the data

• Discovering nothing still means something
  • This can prove to be the main insight
Lab

- Data Acquisition and Pre-Processing

This afternoon

- Visual Encoding and Different Types of Chart