

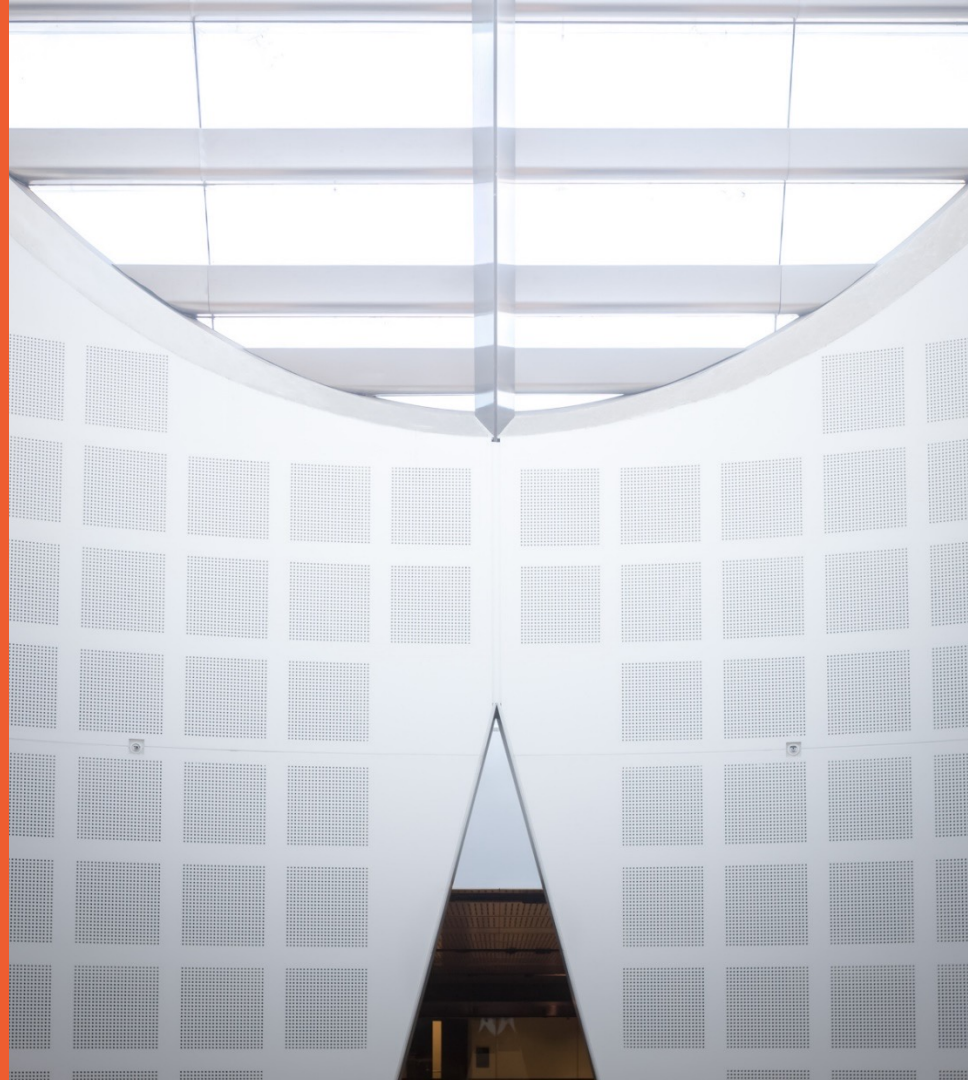
COMP5310: Principles of Data Science

W2: Data Cleaning and Exploration (via Spreadsheet)

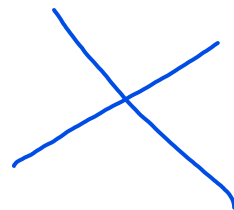
Presented by

Maryam Khanian

Based on slides by previous lecturers of this unit of study



Last week: Introductions and housekeeping



Objective

- Housekeeping; Learn about backgrounds and goals; Define data science.

Lecture

- Welcome, introductions.
- Unit overview, assessment, resources.
- Discuss definitions/scope of data science.

Readings

- Data Science from Scratch: Ch 1.
- Install Anaconda and PostgreSQL.

TO-DO in W1

- Ed Lessons Python modules 1-3.
- Organise into project groups.
- Choose project dataset.

Today: Data cleaning and exploration (via spreadsheet)

Objective

- Use interactive tools to explore a new data set quickly.

Lecture

- Data types, cleaning, preprocessing.
- Descriptive statistics, e.g., mean, stdev, median.
- Descriptive visualisation, e.g., scatterplots, histograms.

Readings

- [Introduction to Data Mining](#): Ch 2.1.1
- Data Science from Scratch: Ch 2-3.

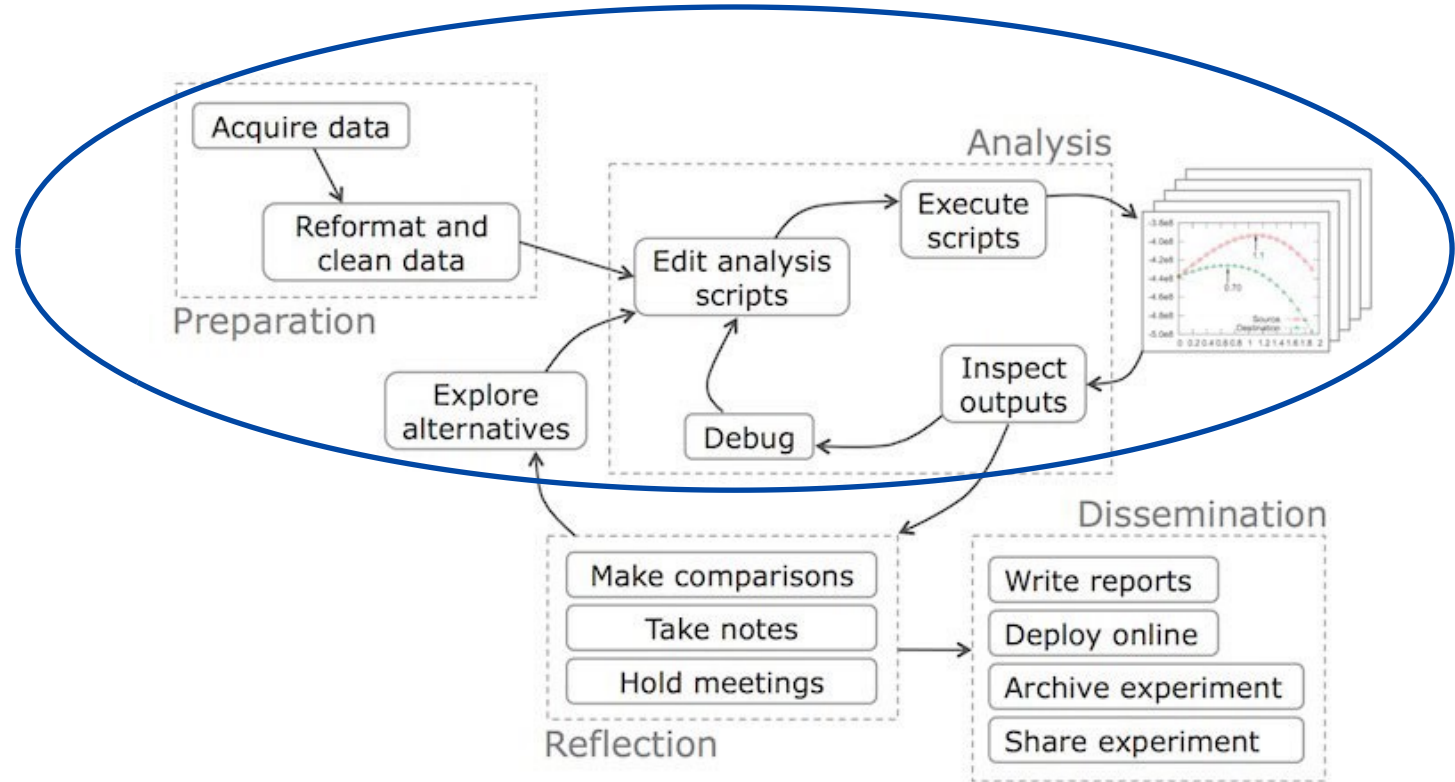
Exercises

- Spreadsheets: Visualisation.
- Spreadsheets: Descriptive statistics.

TO-DO in W2

- Ed Lessons Python modules 4-6.
- Ed Lessons SQL modules 16-17.
- Explore project data.

Exploratory Analysis Workflow



Example dataset

2020 Remote Working Survey Responses:

<https://data.nsw.gov.au/data/dataset/nsw-remote-working-survey>

PRELIMINARIES: TYPES OF DATA

- Data is collection of examples (also called *instances, records, observations, objects*)
- Examples or Objects are described with attributes (*features, variables*)

**Examples
/Objects**

Attributes (features)

Class

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Types of Data

There are different types of data

- **Nominal**
 - Examples: ID numbers, eye color, zip codes
- **Ordinal**
 - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height {tall, medium, short}
- **Interval**
 - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
- **Ratio**
 - Examples: temperature in Kelvin, length, counts, elapsed time (e.g., time to run a race)

Properties of Data Types

- The type of data depends on which of the following properties/operations it possesses:
 - Distinctness: $= \neq$
 - Order: $< >$
 - Differences are meaningful : $+ -$
 - Ratios are meaningful $* /$
- Nominal data: distinctness
- Ordinal data: distinctness & order
- Interval data: distinctness, order & meaningful differences
- Ratio data: all 4 properties/operations

Interval Data

What year were you born?
1972
1972
1982
1987
1991

- Interval scales provide information about order, and also possess equal intervals.
- Values encode differences.
- Equal intervals between values.
- Addition is defined.
- e.g., degrees in Celcius.
- No true Zero and can represent values below zero. (No multiplication or division)

Interval Data (cont')

What year were you born?
1972
1972
1982
1987
1991

- **Central tendency** can be measured by mode, median, or mean.
- **Dispersion** can be estimated by the Inter-Quartile Range (IQR), stdev, variance

Calculating descriptive statistics

- First sort values, then:
 - **Median** is the middle value (or average of two middle values).
 - **Minimum** is the first value.
 - **Maximum** is the last value.
 - **10th percentile** is item at index $0.1 * N$.
 - **90th percentile** is item at index $0.9 * N$.
 - **Range** is Maximum minus Minimum.
 - **IQR** is the difference between the first and third quartile.

$$\text{Ex } 25\% - 75\%$$

Interval Data

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r + 1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

- How to calculate the median for the given output data:

[1,1,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,**3**,**3**,3,3,3,3,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,5,5,5,5,5,5]

- How to calculate the IQR:

$[1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3]$ $[3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3]$ Q_2
 $[3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4]$ $[4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5]$ Q_4
 A_3

- i.e.: $Q3 - Q1 = 4 - 3 = 1$.

Calculating descriptive statistics

- **Mean** is the sum of values divided by the number of values: $\frac{\sum X_i}{N}$
- **Variance:** $\frac{\sum (X_i - \text{mean})^2}{N-1}$
- **Standard deviation:** $\sqrt{\text{variance}}$

Ratio Data

How long have you been in your current job?

(Reponses edited for example: scale in years)

2 years

10 years

8 years

4 years

45 years

- All operations supported by interval data are good here
- Zero defined.
- Multiplication defined.
- Ratio is meaningful.
- e.g., length, weight, income, degree in Kelvin
- Degree in Celcius is not ratio data

Ordinal Data

My organisation encouraged people to work remotely

NA

Strongly disagree

Disagree

Somewhat disagree

Neither agree nor disagree

Somewhat agree

Agree

Strongly agree

- Values are ordered.
- No distance is implied.
- e.g., rank, agreement.
- Central tendency can be measured by mode or median.
- The mean and stdev cannot be defined from an ordinal set.

Ordinal Data

- Countable: can assign a positive integer one-to-one to each response.
- Order defined:
 1. Strongly Disagree
 2. Disagree
 3. Somewhat Disagree
 4. Neither Agree nor Disagree
 5. Somewhat Agree
 6. Agree
 7. Strongly Agree

Nominal Data

Which of the following best describes your industry?

Manufacturing

Wholesale Trade

Electricity, Gas, Water and Waste Services

Professional, Scientific and Technical Services

Transport, Postal and Warehousing

- Values are names.
- No ordering is implied.
- e.g., football jersey numbers.

What about text data?

What do you like about remote work? (Manufactured example)

Avoiding my commute

Going to the gym at lunch time

Staying home with my dog

Spending lunch with my family

Peace and quiet while working

- Not defined as traditional data type in statistics.
- Requires interpretation, coding or conversion.
- More in future lectures...

Self Assessment (1)

- Which one(s) of these data types can be ordered?
 - Nominal
 - Ordinal
 - Interval
 - Ratio
 - None of them
 - All of them

Self Assessment (2)

- Which one(s) of these data types can use multiplication and division operations?
 - Nominal
 - Ordinal
 - Interval
 - Ratio
 - None of them
 - All of them

Levels of Measurement

	Nominal	Ordinal	Interval	Ratio
Countable	✓	✓	✓	✓
Order defined		✓	✓	✓
Difference defined (addition, subtraction)			✓	✓
Zero defined (multiplication, division)				✓

Measures of Central Tendency

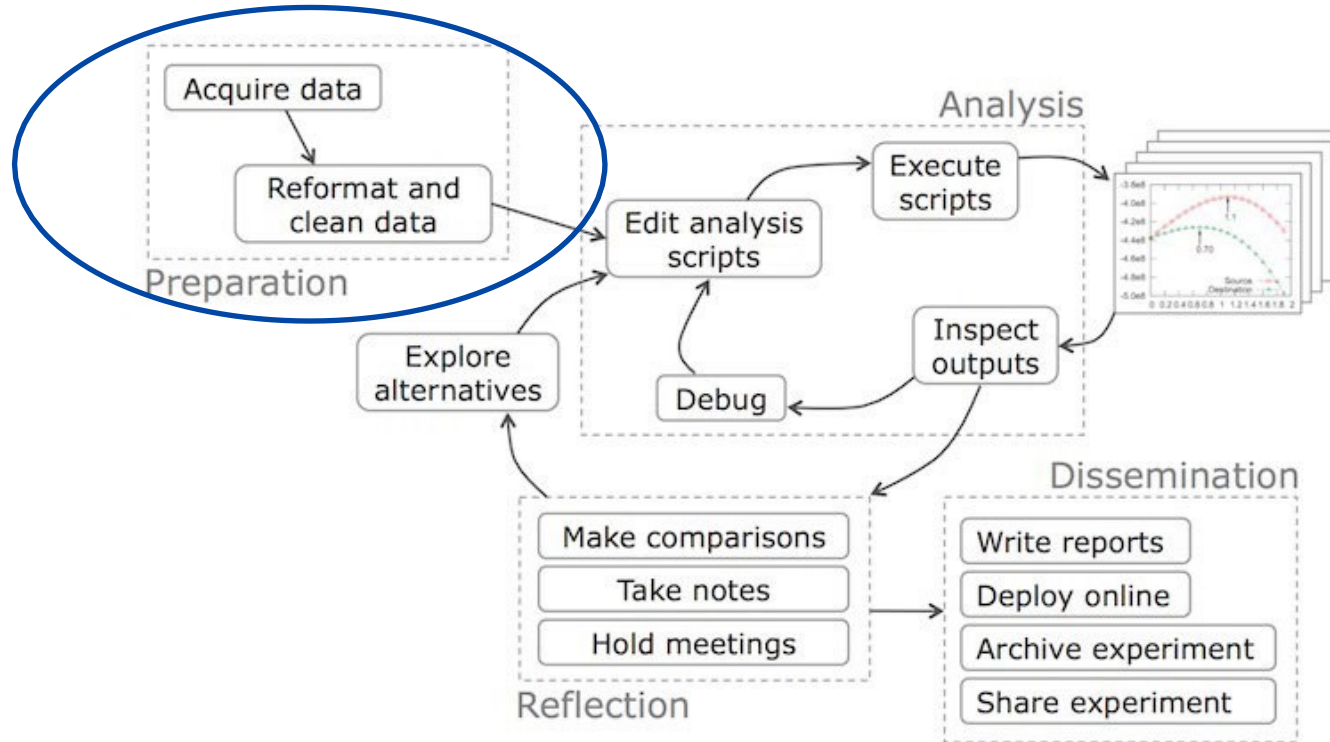
	Nominal	Ordinal	Interval	Ratio
Mode	✓	✓	✓	✓
Median		✓	✓	✓
Mean			✓	✓

Measures of Dispersion

	Nominal	Ordinal	Interval	Ratio
Counts / Distribution	✓	✓	✓	✓
Minimum, Maximum		✓	✓	✓
Range		✓	✓	✓
Percentiles		✓	✓	✓
Standard deviation, Variance			✓	✓

DATA ACQUISITION AND CLEANING


Exploratory Analysis Workflow

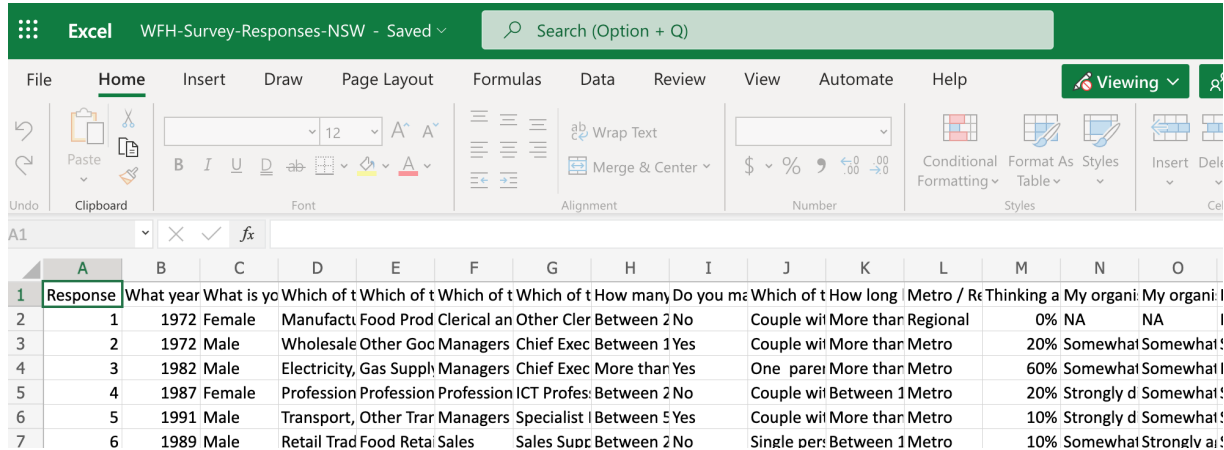


Data Acquisition – Where does data come from?

- File Access
 - You or your organisation might already have a data set, or a colleague provides you access to data.
 - Web download from an online data server.
 - Typical exchange formats: CSV, Excel, sometimes also XML.
- Programmatically (Cf. Data Science from Scratch, Ch 9)
 - Scraping the web (HTML).
 - Using APIs of Web Services (XML/JSON).
- Database Access (Week 4 onwards).
- Collect data yourself, e.g., via a survey.
- **This week:** Using data from the WFH survey.

Acquire data

- Create new Excel spreadsheet
 - Go to your university email.
 - Click the Spreadsheet button. 
 - File > Open > navigate to WFH survey data.



The screenshot shows the Microsoft Excel interface with the following data in the spreadsheet:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Response	What year	What is yo	Which of t	Which of t	Which of t	Which of t	How many	Do you m	Which of t	How long	Metro / R	Thinking a	My organi	My organi
2	1	1972	Female	Manufactu	Food Prod	Clerical an	Other Cler	Between 2	No	Couple wil	More than	Regional	0%	NA	NA
3	2	1972	Male	Wholesale	Other Gro	Managers	Chief Exec	Between 1	Yes	Couple wil	More than	Metro	20%	Somewhat	Somewhat
4	3	1982	Male	Electricity	Gas Supply	Managers	Chief Exec	More than	Yes	One parent	More than	Metro	60%	Somewhat	Somewhat
5	4	1987	Female	Profession	Profession	Profession	ICT Profes	Between 2	No	Couple wil	Between 1	Metro	20%	Strongly d	Somewhat
6	5	1991	Male	Transport	Other Trar	Managers	Specialist	Between 5	Yes	Couple wil	More than	Metro	10%	Strongly d	Somewhat
7	6	1989	Male	Retail Trad	Food Retail	Sales	Sales Supr	Between 2	No	Single per	Between 1	Metro	10%	Somewhat	Strongly d

Cleaning and Transforming Data

- Real data is often '*dirty*'.
- Important to do some data cleaning and transforming first.
- Typical steps involved:
 - Type and name **conversion**.
 - **Filtering** of missing or inconsistent data.
 - **Unifying** semantic data representations.
 - **Matching** entries from different sources.
- Later also:
 - **Rescaling** and optional **dimensionality reduction**.

Exercise: Reformat and clean data

Review and discuss

- Any problems with columns in spreadsheet?
- How should we fix those problems?

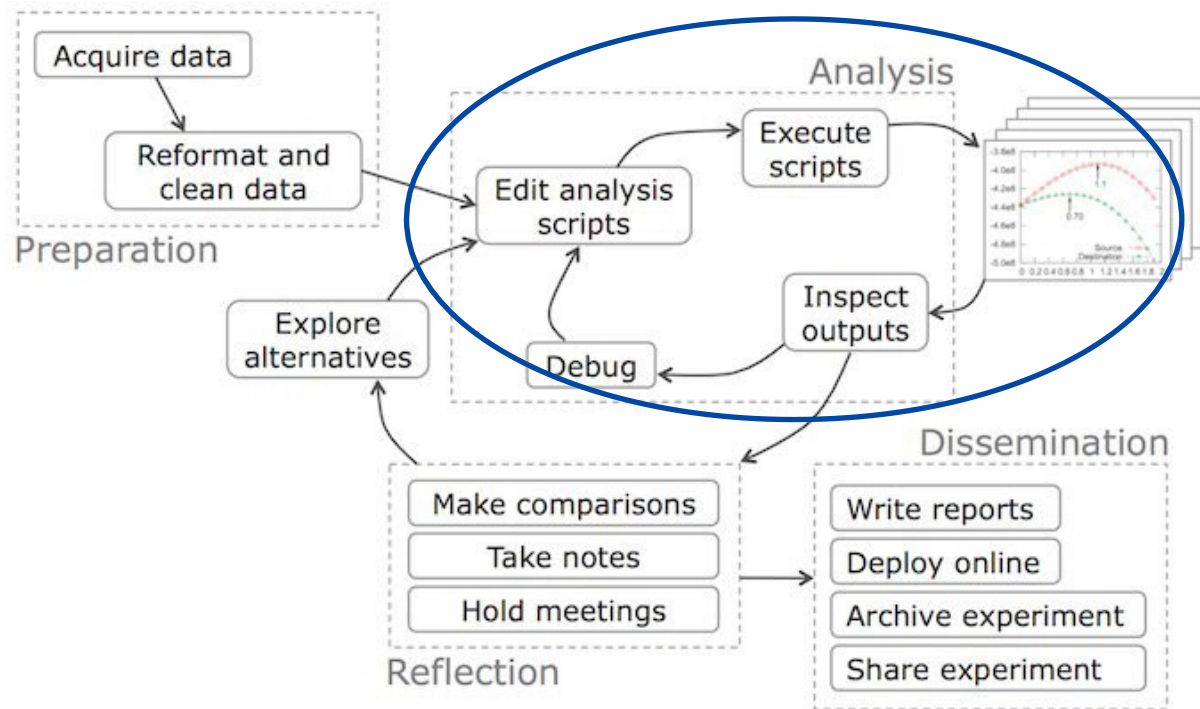


Clean

- Change any text to numeric values in “Number of years...” columns.
- Check format of "Thinking about your current job, how much of your time did you spend remote working last year?" Note that rounding applied on top of underlying data. Is this intentional?

**WHAT QUESTIONS
CAN WE ANSWER?**

Exploratory Analysis Workflow



Exercise: What questions can you ask?

- Review WFH Survey data.
- List 3 questions you can ask.
- Discuss how you would answer each question with this data.

Some descriptive questions

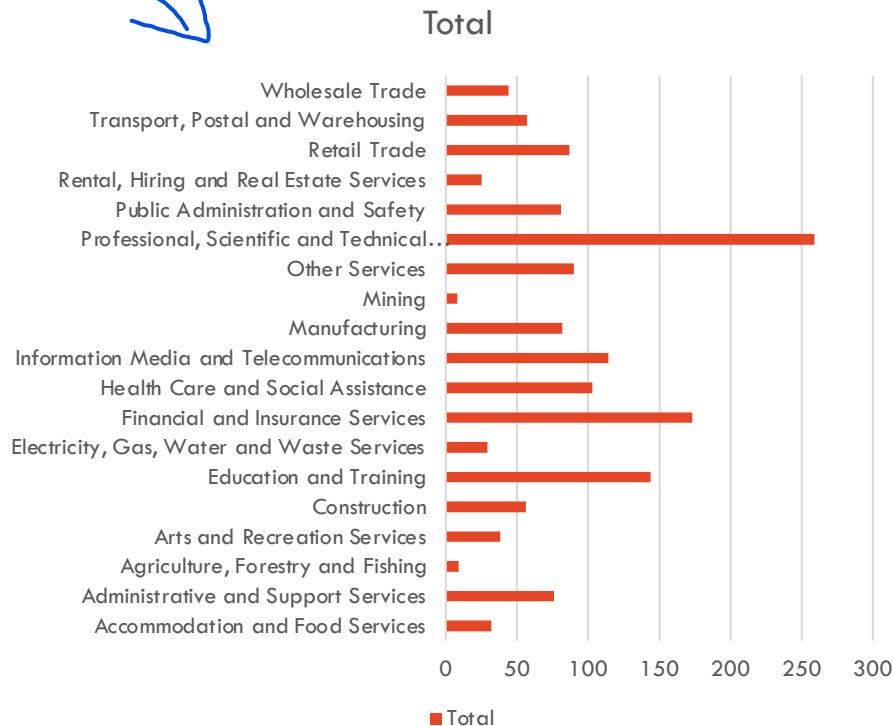
- What industries do people spend more time WFH?
- Do more people who manage than not manage WFH?
- In what industries do people with dependents (e.g., children) WFH the most?
- Do large organizations encourage more people to WFH?

WFH = working from home

PIVOT TABLES

Table and bar chart of industry

Which of the following best describes your industry?	Count of Which of the following best describes your industry?
Accommodation and Food Services	32
Administrative and Support Services	76
Agriculture, Forestry and Fishing	9
Arts and Recreation Services	38
Construction	56



Creating a pivot table

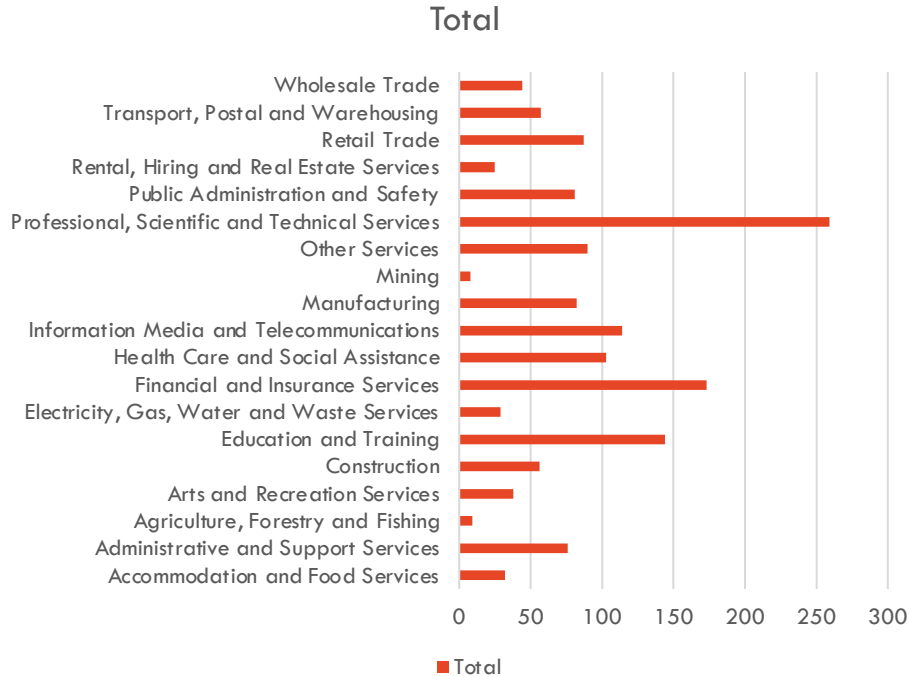
- Summarize data by calculating statistics over sub-populations.
 - e.g., count of industry by name.
- In Excel:
 - Select data range (e.g., C1:En)
 - Go to Insert > Pivot Table (should insert a new sheet).
 - Select industry under row.
 - Select industry under value.
 - Summarize by count.

Exercise: Using a pivot table to summarise data

- **Pivot table**
 - Create a table of average age by industry.
- **Discuss/explore**
 - What other statistics can we calculate?
 - What other variable combinations could we explore?

SUMMARISING NOMINAL DATA

Summarise nominal data with bar charts



- **Measures of central tendency:**
 - Mode.
- **Measures of dispersion:**
 - Counts/distribution

Calculating the Mode

- The most frequent value.
- Defined for nominal data, but spreadsheets might not compute.
- Can be read from a bar chart.

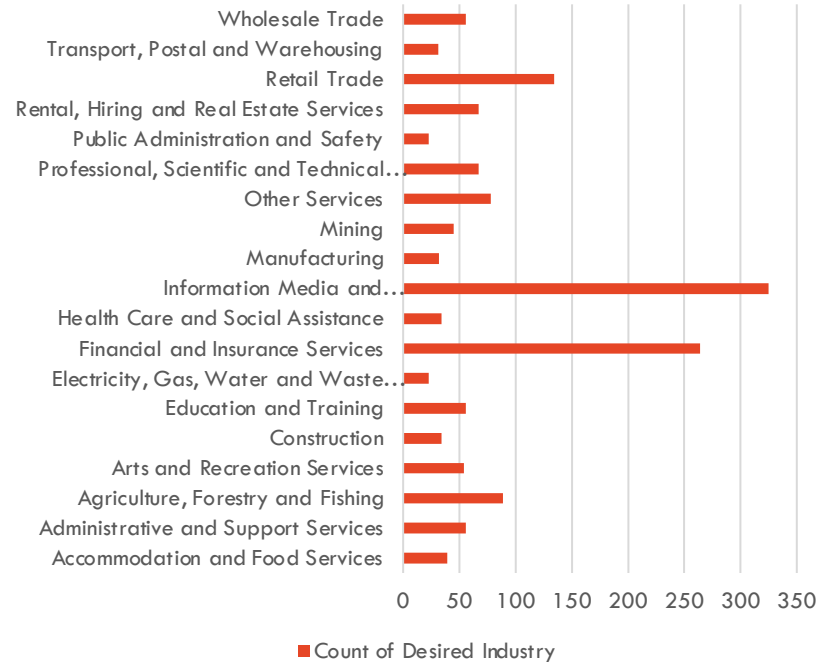
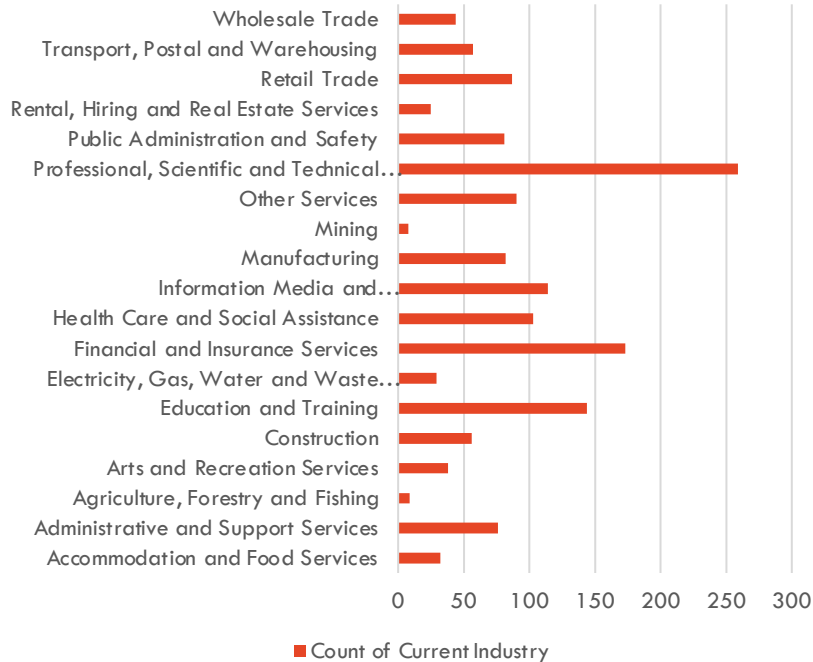
Create Bar Charts

- Count frequency of each category.
- Display on bar chart.
- In Excel
 - Needs a column of responses and a column of counts (can be aggregated in a pivot table).
 - Select data range (e.g., A2:B20).
 - Insert > Bar Chart.

Exercise: Exploring nominal data

- **Visualise**
 - Create histograms of current and desired industries (synthetic data).
- **Discuss**
 - What do we need to do to make these comparable?
 - What is the mode?

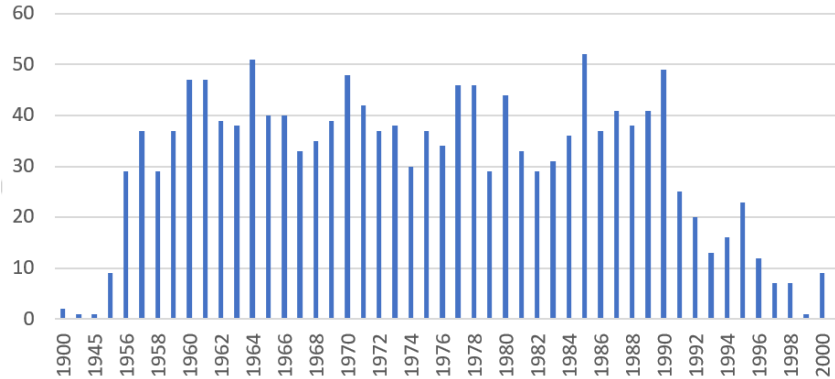
Bar charts comparing known and future industries



Discuss: Do modes differ? Ranges? Number of responses?

SUMMARISING ORDINAL DATA

Summarise ordinal data: histograms, median, percentiles



- **Measures of central tendency:**
 - Median, mode.
- **Measures of dispersion:**
 - Counts/distribution.
 - Min/max/range.
 - Percentiles.

Creating a Histogram chart

- Count frequency, e.g., of ordinal values within each category
- Display on histogram chart with one variable grouped inside
- In Excel:
 - Needs a column of responses and a column of counts (can be aggregated in a pivot table).
 - Insert > Pivot Table > Select full range of data in the spreadsheet > Drag and drop column name that holds response data into the Rows and Values field (check Value is set to Count).
 - Select data range from Pivot Table (e.g., A2:B20).
 - Insert > Column Chart.

Exercise: Exploring ordinal data

- **Visualise**
 - Create a histogram diagram of the question “What year were you born? ”
- **Discuss**
 - What do the responses "1900" mean?
 - Does this reflect underlying working population distribution or are some age groups more well-represented in the survey data?

SUMMARISING RATIO DATA:

How do professional/programming experience
compare?

Ratio (and interval) data



- **Measures of central tendency:**
 - Mean, median, mode
- **Measures of dispersion:**
 - Counts/distribution
 - Min/max/range
 - Percentiles
 - Stdev/variance

Creating a Scatterplot

- Plots relationship between two different variables.
- Display, e.g., professional experience on x-axis vs. programming experience on y-axis for each respondent.
- In Excel:
 - Select data range (e.g., D1:En).
 - Insert > Scatter.

REVIEW

W2 Review: Data cleaning and exploration (via spreadsheet)

Objective

- Use interactive tools to explore a new data set quickly.

Lecture

- Data types, cleaning, preprocessing.
- Descriptive statistics, e.g., mean, stdev, median.
- Descriptive visualisation, e.g., scatterplots, histograms.

Readings

- [Introduction to Data Mining](#): Ch 2.1.1
- Data Science from Scratch: Ch 2-3.

Exercises

- Spreadsheets: Visualisation.
- Spreadsheets: Descriptive stats.

TO-DO in W2

- Ed Lessons Python modules 4-6.
- Ed Lessons SQL modules 16-17.
- Explore project data.