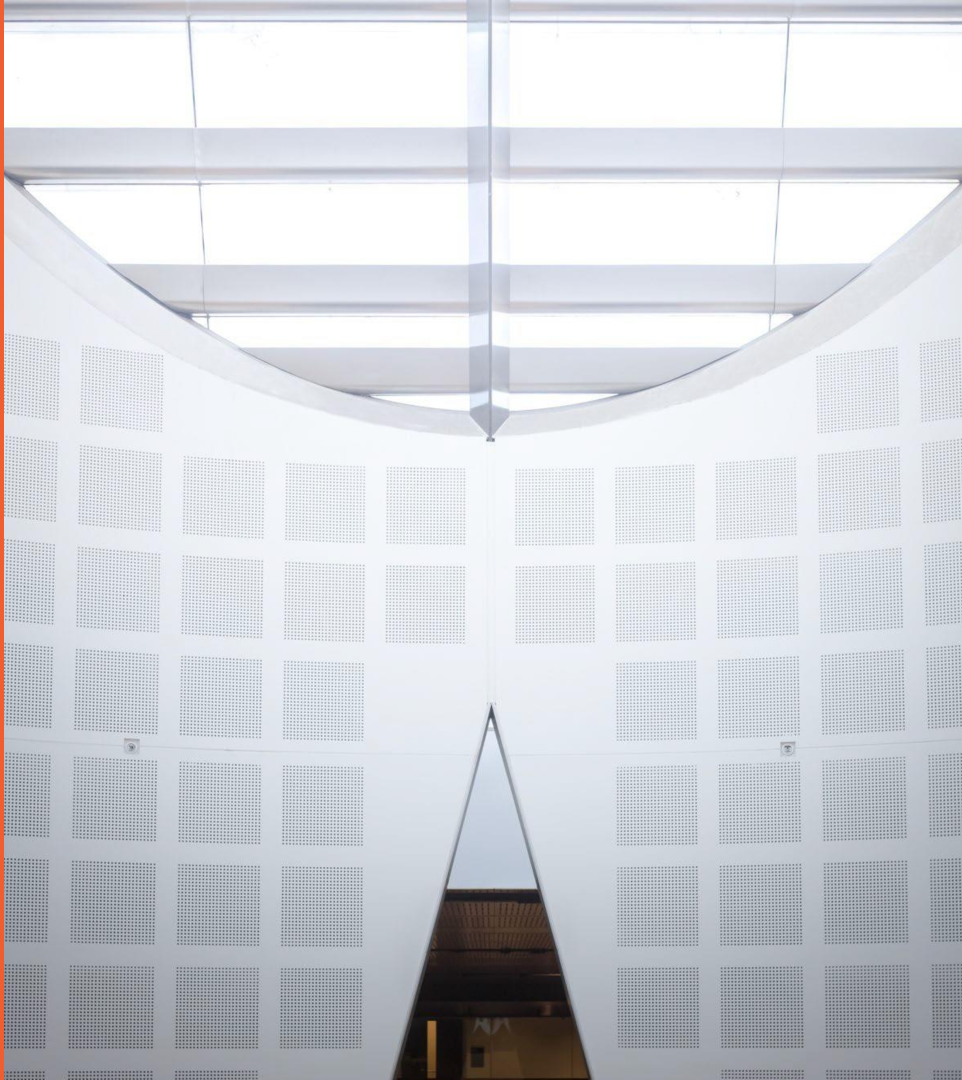
COMP5310: Principles of Data Science

W2: Data Acquisition and Exploration

Presented by
Claire Hardgrove
School of Computer Science
Modified from slides by Dr Ali Anaissi



Overview of Week 2

THE UNIVERSITY OF SYDNEY

Last time: Introductions and Housekeeping

Objective

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

Lecture

- Welcome, introductions
- Unit overview, assessment, resources
- Learning Python with Grok
- Discuss definitions/scope of data science

Readings

- Data Science from Scratch: Ch 1
- Is being a data scientist really the best job in America?
- 8 skills you need to be a data scientist

Exercises

- Introductions / interviews
- Interests / definitions

TODO in W1

- Grok Python modules 1-3
- Choose possible project data

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

- Data Science from Scratch: Ch 2-3

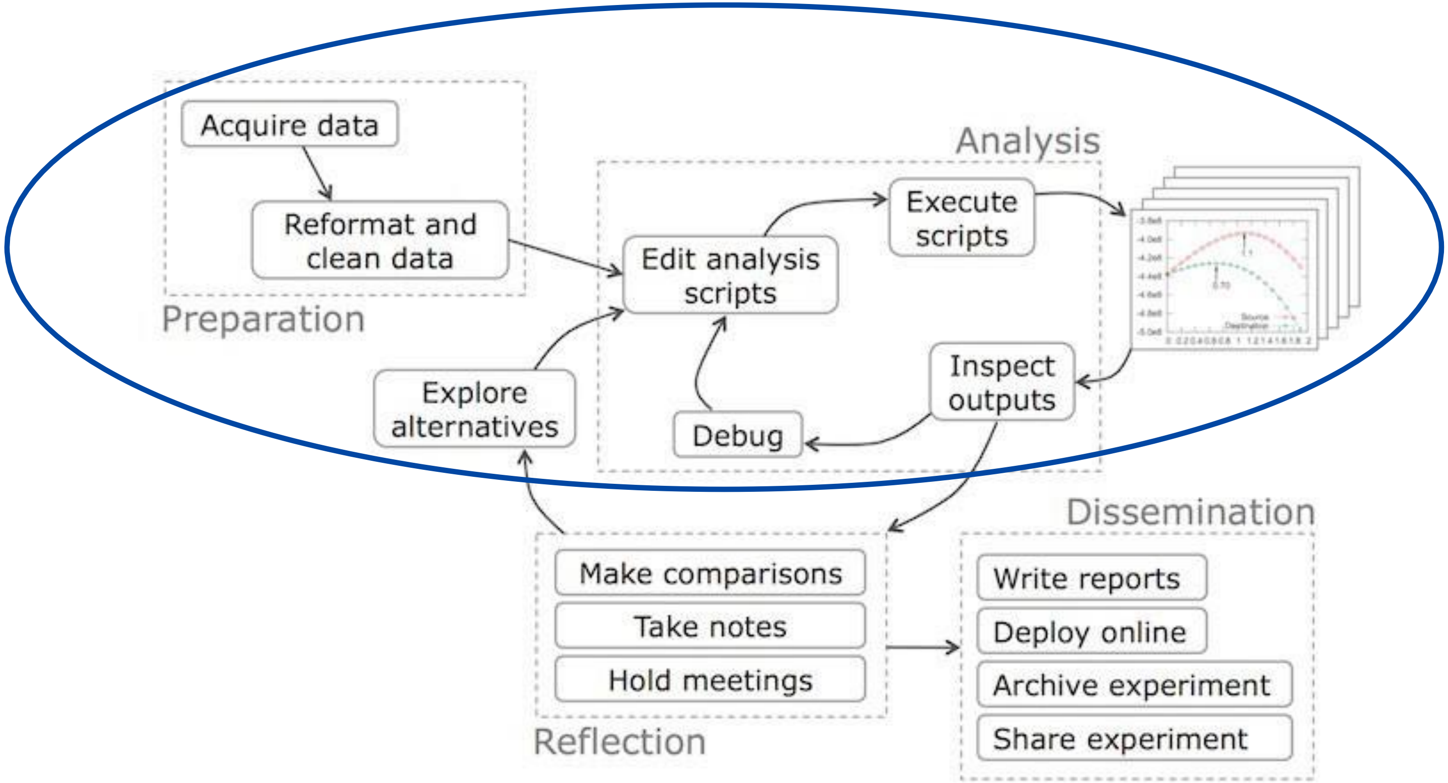
Exercises

- Google Sheets: Visualisation
- Google Sheets: Descriptive stats

TODO in W2

- Grok Python modules 4-6
- Grok SQL modules 16 and 17
- Explore project data

Exploratory Analysis Workflow



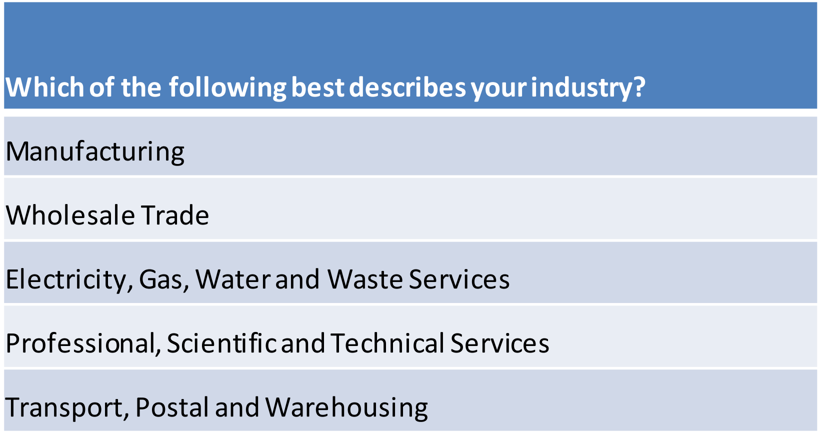
Example dataset

2021 Remote Working Survey Responses (downloaded 4 August 2022):

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

Preliminaries: Types of Data

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
- Eg football jersey numbers

Ordinal Data

My organisation encouraged people to work remotely
NA
Somewhat agree
Somewhat agree
Strongly disagree
Strongly disagree
Somewhat agree
NA
Strongly agree
Strongly agree

- Values are ordered
- No distance is implied
- Eg rank, agreement
- *central tendency* can be measured by mode[1] or median
- the mean cannot be defined from an ordinal set
- dispersion can be estimated by the Inter-Quartile Range (IQR)

[1] The mode is the number that is repeated more often than any other

Ordinal Data

- Countable: can assign a positive integer one-to-one to each response
- Order defined

1. Strongly Disagree
2. Disagree
3. Neither Agree nor Disagree
4. Agree
5. Strongly Agree

Ordinal Data

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

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

the 'cut-off' points
are called **quartiles**

– How to calculate the IQR:

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

– The IQR is the difference between the first and third quartile. i.e:

Q3 – Q1 = 4 – 3 = 1.

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
- No true Zero
- Addition is defined
- Eg. degrees Celcius (not Kelvin)

- *central tendency* can be measured by mode, median, or mean

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

- Values encode differences
- Zero defined
- Multiplication defined
- Ratio is meaningful
- Eg length, weight, income

Levels of Measurement

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

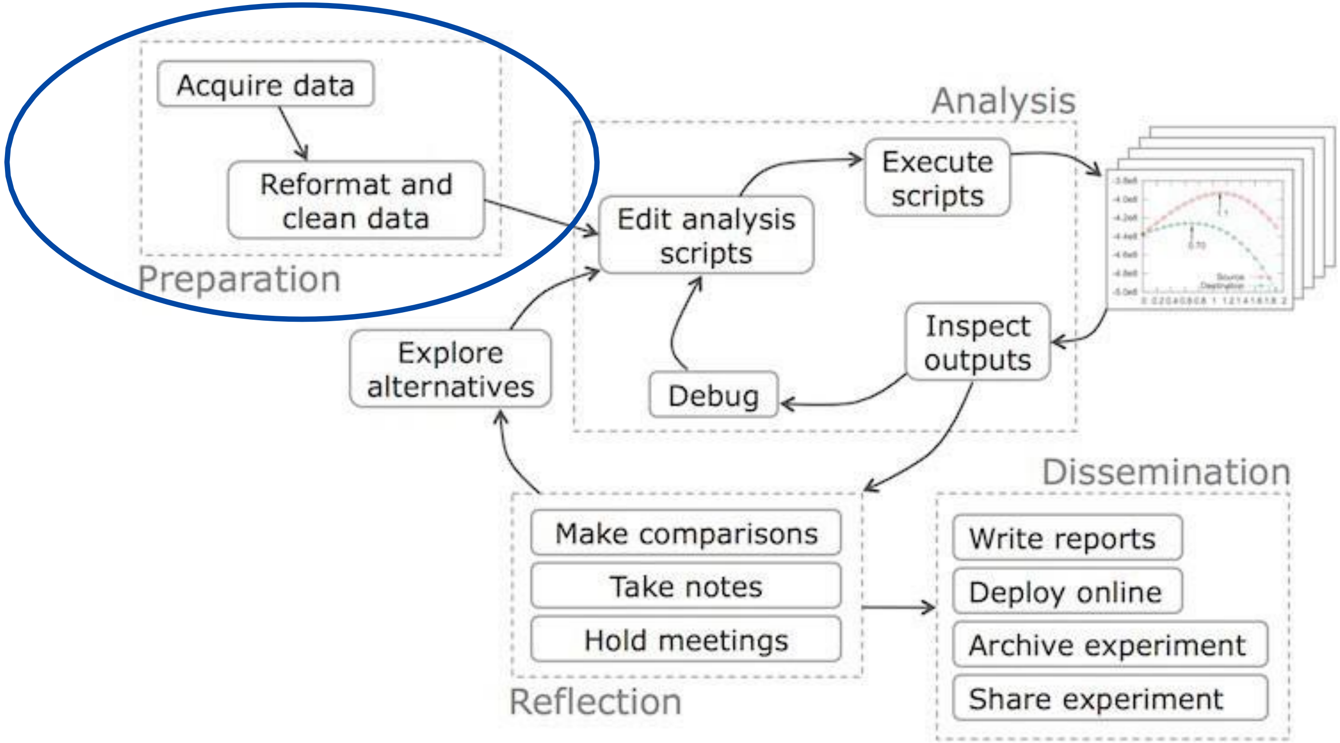
What about text data?

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…

Data Acquisition and Cleaning

Exploratory Analysis Workflow



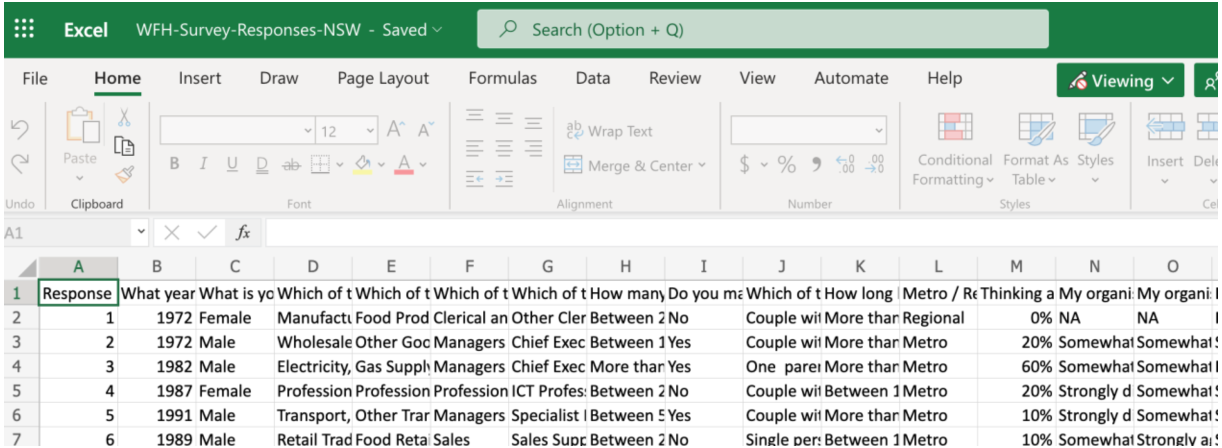
Data Acquisition – Where does data come from?

- File Access
  - You or your organisation might already have a data set, or a colleagues provides you access to data.
  - Or: Web Download from an online data server
  - Typical exchange formats: CSV, Excel, sometimes also XML
- Programmatically
  - Scraping the web (HTML)
  - or using APIs of Web Services (XML/JSON)
    -> Cf. textbook, Ch 9
- Database Access -> Week 4 onwards
- Collect data yourself, eg. 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



	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 ma	Which of t	How long	Metro / Re	Thinking a	My organi:	My organi:
2	1	1972	Female	Manufactu	Food Prod	Clerical an	Other Cler	Between 2	No	Couple wit	More than	Regional	0%	NA	NA
3	2	1972	Male	Wholesale	Other Goo	Managers	Chief Exec	Between 1	Yes	Couple wit	More than	Metro	20%	Somewhat	Somewhat
4	3	1982	Male	Electricity,	Gas Supply	Managers	Chief Exec	More than	Yes	One parer	More than	Metro	60%	Somewhat	Somewhat
5	4	1987	Female	Profession	Profession	Profession	ICT Profes:	Between 2	No	Couple wit	Between 1	Metro	20%	Strongly d	Somewhat
6	5	1991	Male	Transport,	Other Trar	Managers	Specialist I	Between 5	Yes	Couple wit	More than	Metro	10%	Strongly d	Somewhat
7	6	1989	Male	Retail Trad	Food Retai	Sales	Sales Supp	Between 2	No	Single pers	Between 1	Metro	10%	Somewhat	Strongly a

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 of entries from different sources
- Later also:
  - **Rescaling** and optional dimensionality reduction

Exercise: Reformat and clean data

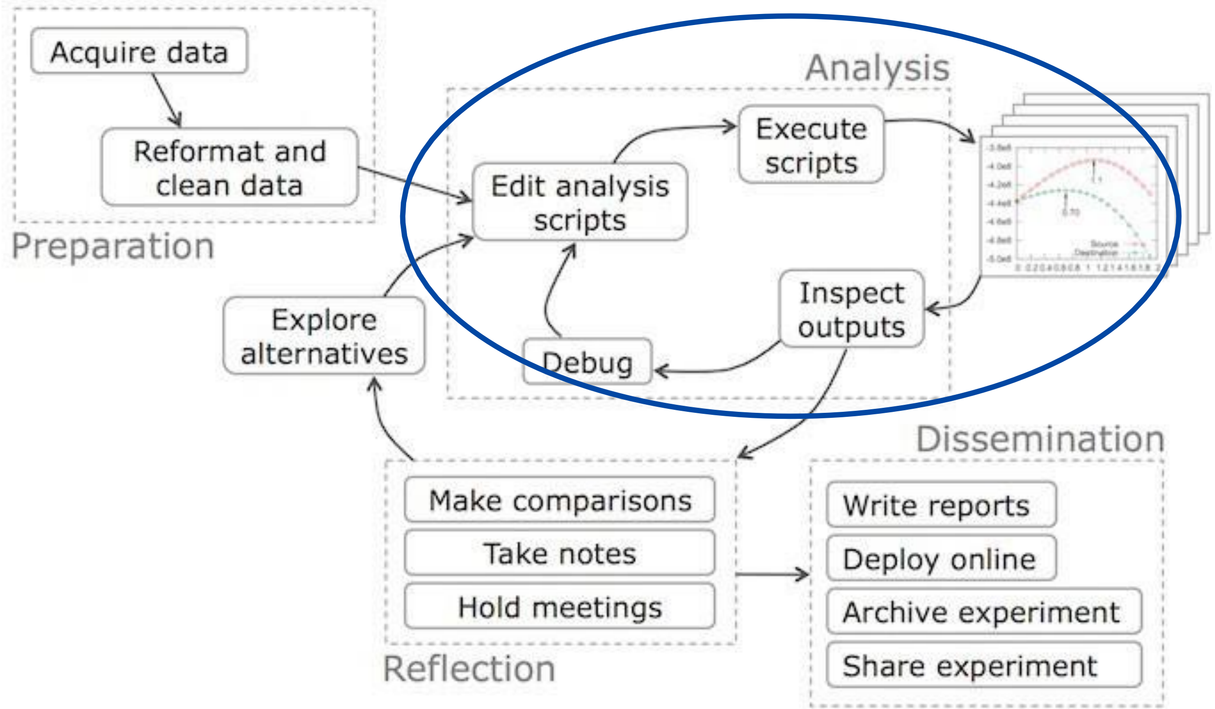
- Review and discuss:
  - Any problems with columns in spreadsheet?
  - How should we fix those problems?
- Clean:
  - Change any text to numeric values in “Number of years…” columns
  - Check format of "Thinking about your current job, how much of your time did you spend remote working last year?" Note that rounding applied on top of underlying data. Is this intentional?

Exercise: What questions can we ask?

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

What Questions Can We Answer?

Exploratory Analysis Workflow



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 = work from home

Pivot Tables

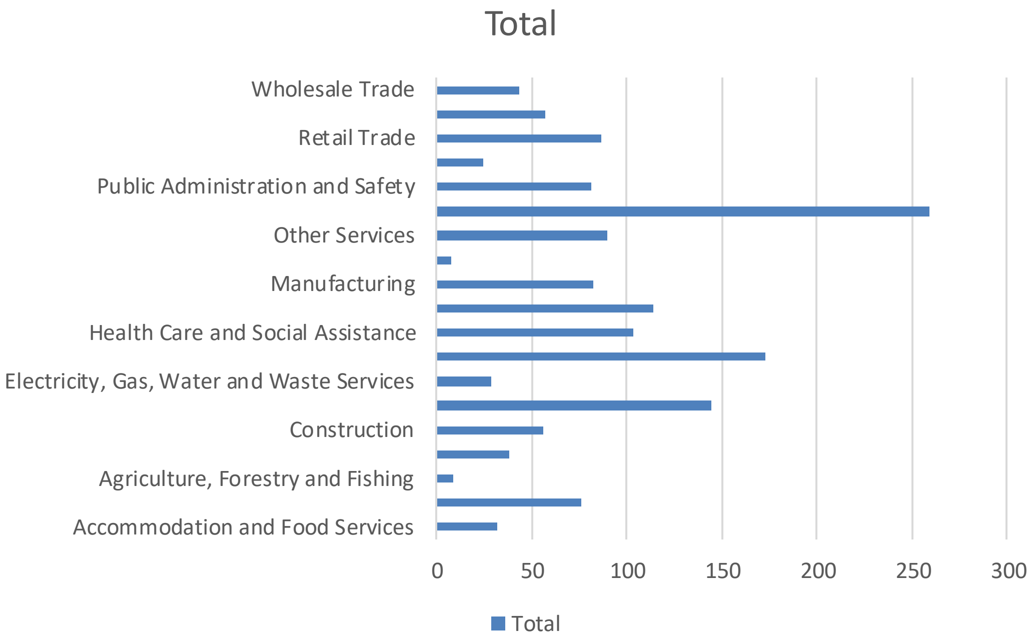
THE UNIVERSITY OF SYDNEY

Creating a pivot table

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

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

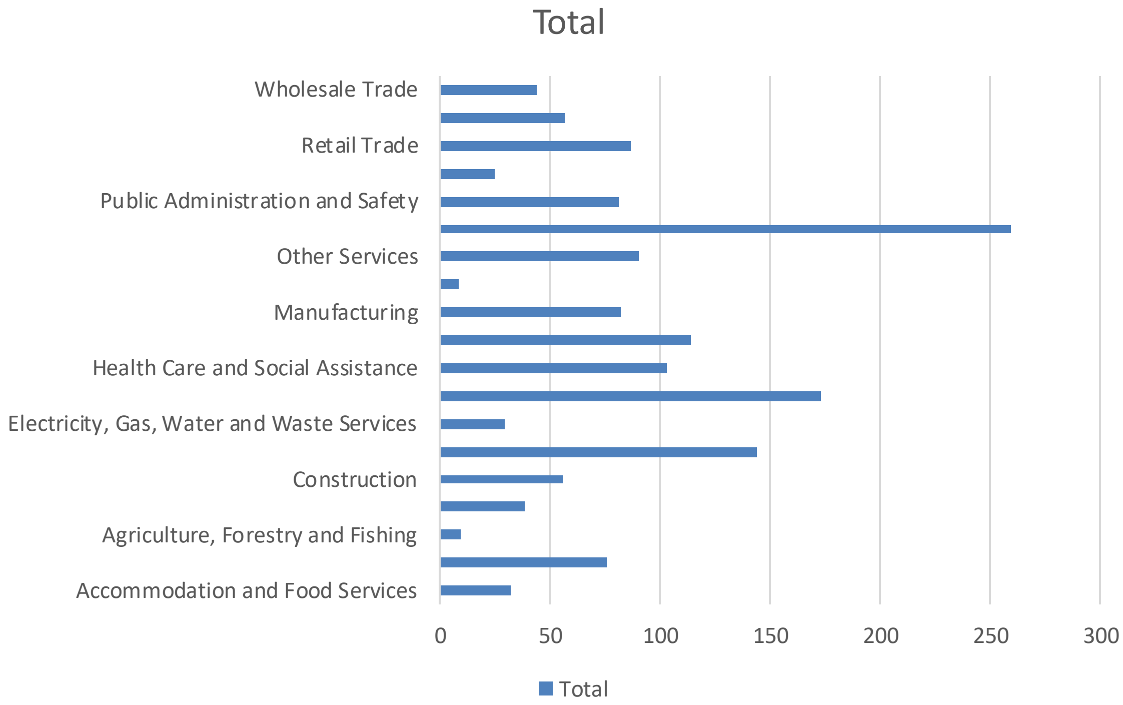


Exercise: Using a pivot table to summarise data

- 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 read from a bar chart

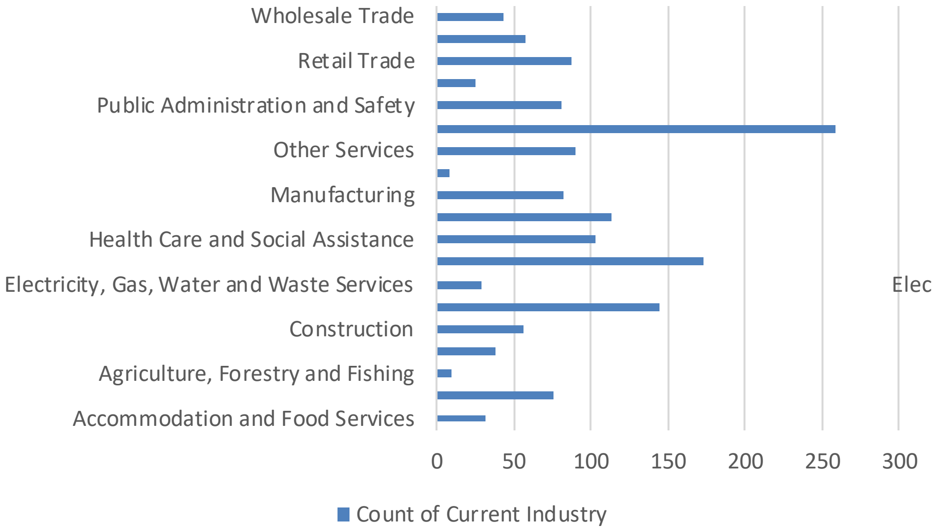
Creating Bar Charts

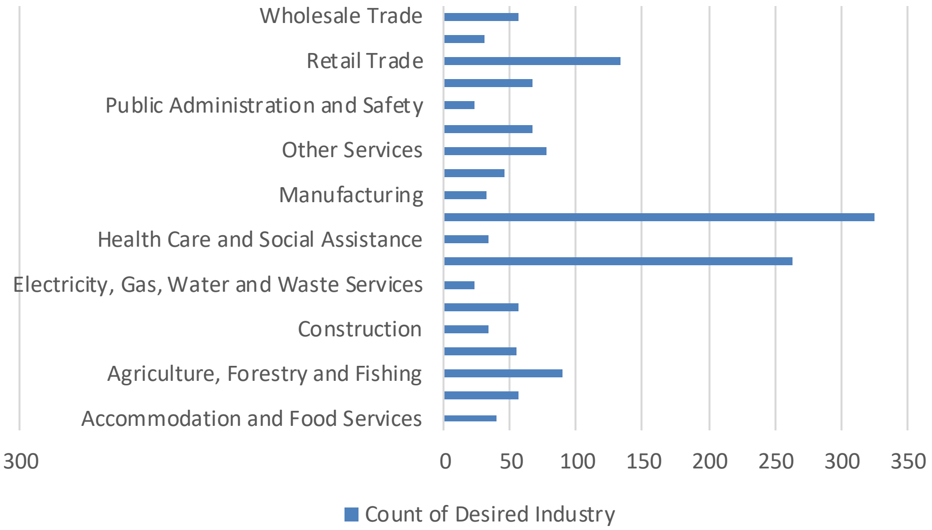
- Count frequency of each category
- Display on bar chart
- In Excel
  - Needs a column of responses and a column of counts (can be aggregated in a pivot table)
  - Select data range (e.g., A2:B20)
  - Insert > Bar Chart

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





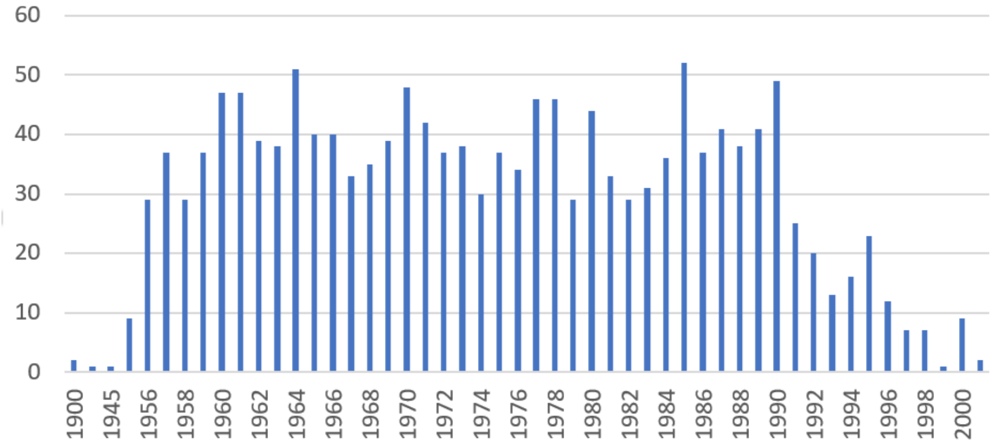
Discuss

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

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

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

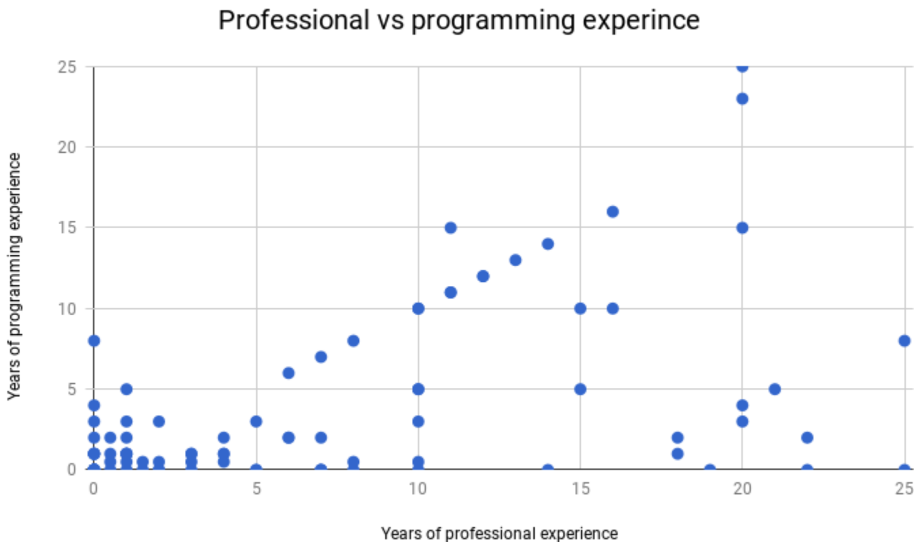
Exercise: Exploring ordinal data

- Visualise:
  - Create a histogram diagram of "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

Calculating descriptive statistics

- ***Median*** and ***percentiles*** good here too
- ***Mean*** is the sum of values divided by the number of values:

$$\frac{\sum X_i}{N}$$

- ***Variance:***

$$\frac{\sum (X_i - mean)^2}{N-1}$$

- ***Standard deviation:***

$$\sqrt{variance}$$

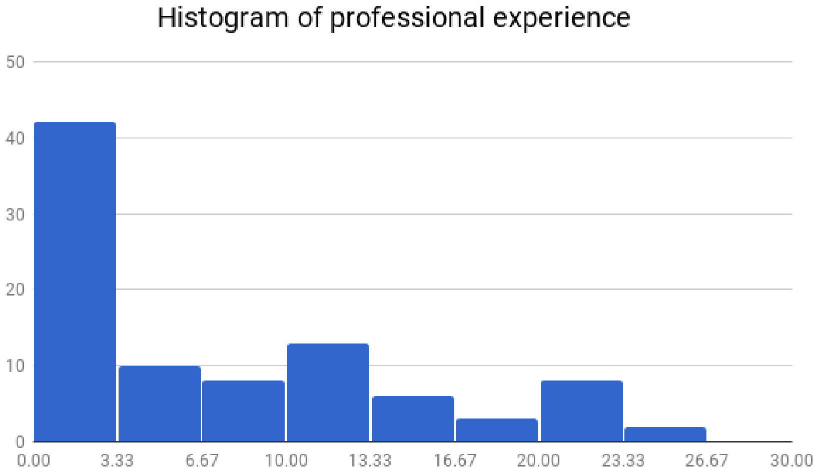
Creating a Scatterplot

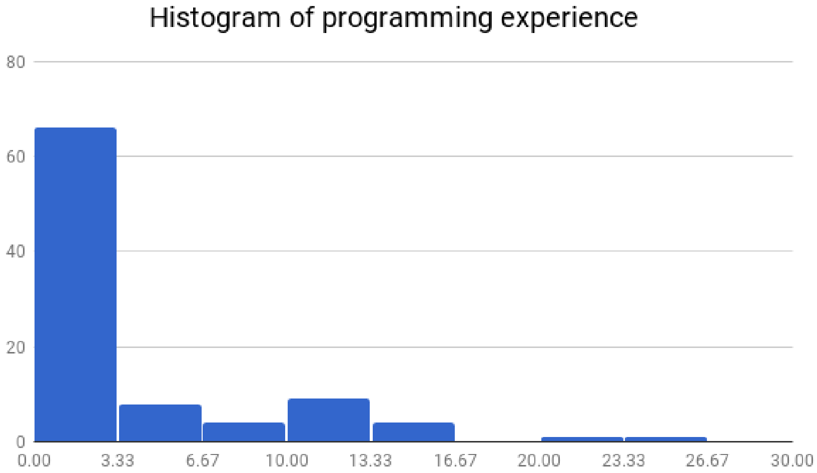
- Plots relationship between two different variables
- Display, e.g., professional experience on x-axis vs. programming experience on y-axis for each respondent
- In Excel
  - Select data range (e.g., D1:En)
  - Insert > Scatter

Exercise: Exploring ratio data

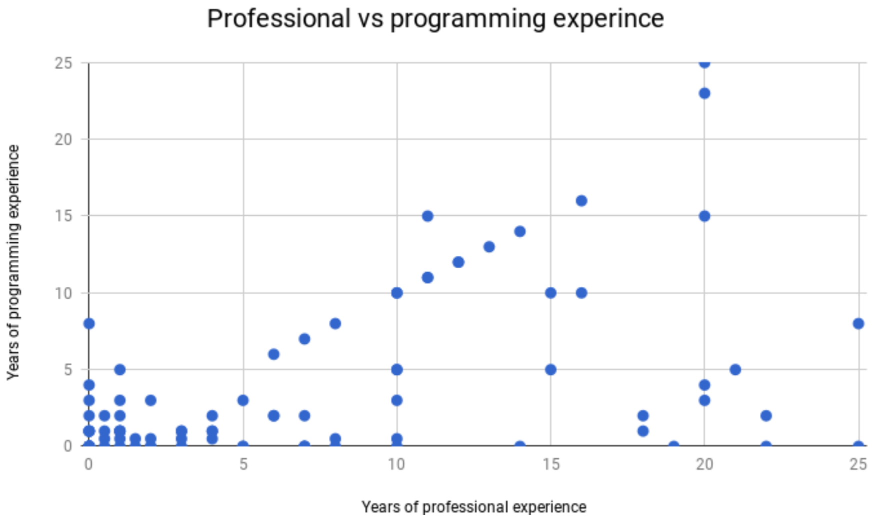
- Visualise:
  - Create a scatter plot of professional vs. programming experience
- Discuss/explore:
  - Is default bin size reasonable?
  - What other kinds of plots can we use to compare experience?
  - How useful are mean and standard deviation numbers?

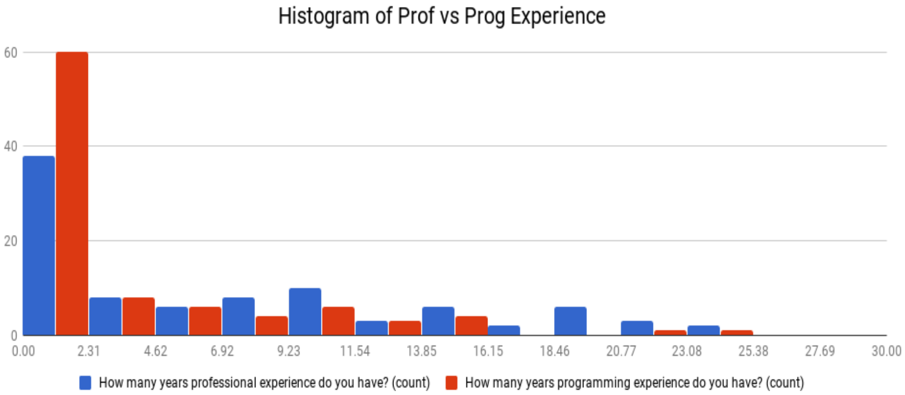
Binned histograms for experience





Comparison with scatterplot and histogram overlays





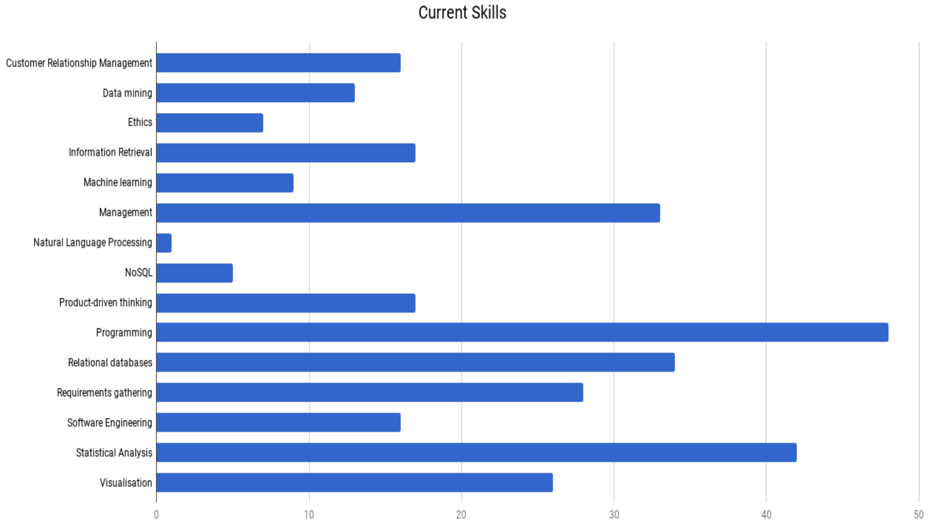
Complex Counting:

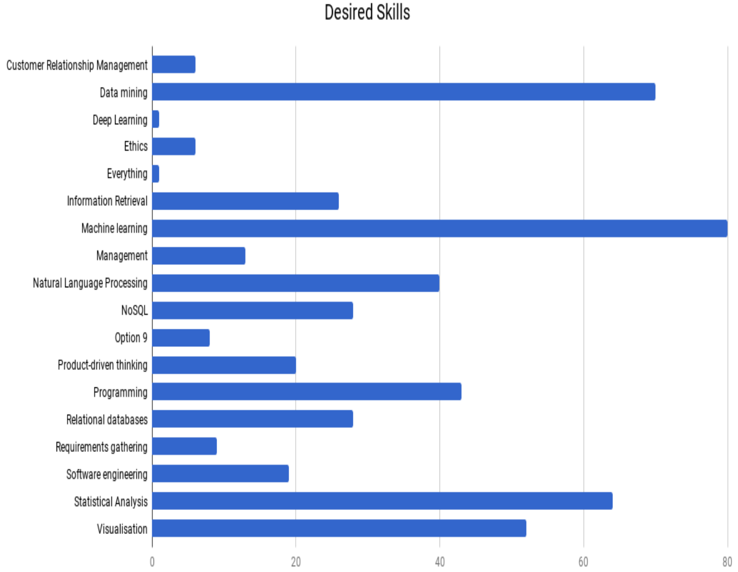
How create a histogram of skills?

- Multiple values in cells within the skills column, e.g.:
  “Software engineering, Requirements gathering, Product-driven thinking”
- Need to split possible values (Google sheets):
  ```
  =sort(unique(transpose(split(join(", ", N2:Nn), ", ", False))))
  ```
- Then count (Google sheets):
  ```
  =countif(N$2:N$n, concat(concat("*", T1), "*"))
  ```

- Could use similar to get word counts
- Better to use programming language (clarity, reusability, etc)

Histograms of current and desired skills (as of 2018...)





Review

Participation

Objective

Ensure everybody is keeping up.

Requirements

- Submit code at end of each exercise

Marked:

- Code/spreadsheets from exercises
- Each week's participation assessed as: all done, partially done, no participation

Marking

- 10% of overall mark

From Week 2 to Week 11: PDF of your lab exercises workbook due Sunday after lab at 23:59pm

W2 Review: Data cleaning and exploration

Objective

Use interactive tools to clean and 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

- Data Science from Scratch: Ch 2-3

Exercises

- Google Sheets: Visualisation
- Google Sheets: Descriptive stats

TODO in W2

- Grok Python modules 4-6 + First SQL module
- Explore project data

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			✓	✓

Next Time

Next week: Data Exploration with Python

Objective

Learn Python tools for exploring a new data set programmatically.

Lecture

- Data types, cleaning, preprocessing
- Descriptive statistics, e.g., median, quartiles, IQR, outliers
- Descriptive visualisation, e.g., boxplots, confidence intervals

Readings

- Data Science from Scratch: Ch 4-5

Exercises

- matplotlib: Visualisation
- numpy/scipy: Descriptive stats

TODO in W2

- Grok Python modules 4-6
- Grok SQL modules 1-2
- Explore and select project data

Project Stage 1

THE UNIVERSITY OF SYDNEY

Project stage 1: Explore, Clean, Load

Objective

Explore a data set and define a research question based on research/business requirement.

Activities

- Individually propose a topic
- Individually choose a dataset
- Individually load and clean the data
- As a group, discuss and recommend topic and dataset for Stage 2A onwards

Output

- See Project Stage 1 specification on Canvas

Marking

- 5% of overall mark

Suggested timeline for Assignment 1 (Project Stage 1)

- W1: Identify possible topics
- W2: Obtain datasets and metadata
- W3: Load data with Python
- W4: Clean and prepare data
- W5: Assess strengths and limitations of each topic/dataset
- W6: Submit 2-page report

Types of projects to consider

- Discover clusters in data
- Learn association rules
- Train a classifier and evaluate prediction accuracy
- Train a regression model and evaluate prediction accuracy

Questions?

THE UNIVERSITY OF
SYDNEY