COMP5310: Principles of Data Science

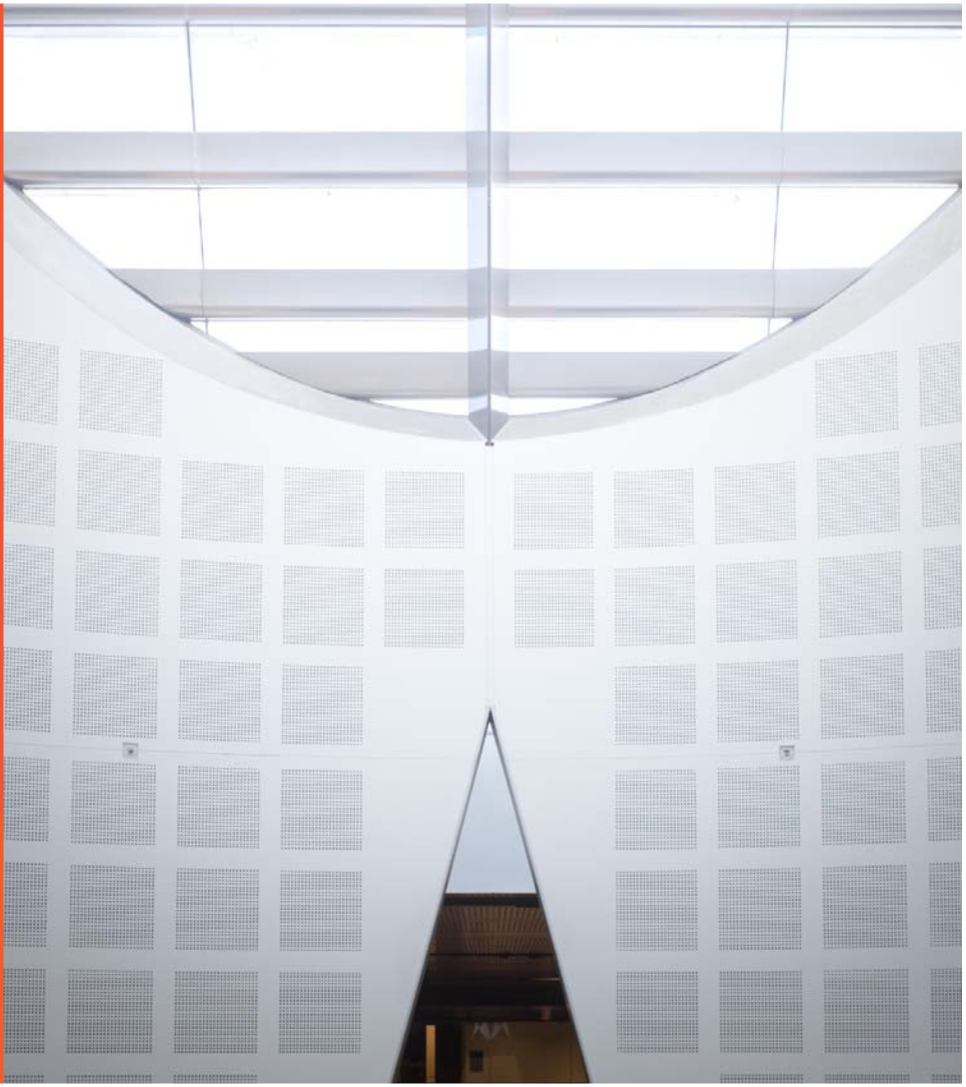
W2: Data Acquisition and Exploration

Presented by

Dr Matloob Khushi

School of IT

THE UNIVERSITY OF SYDNEY



Overview of Week 2

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

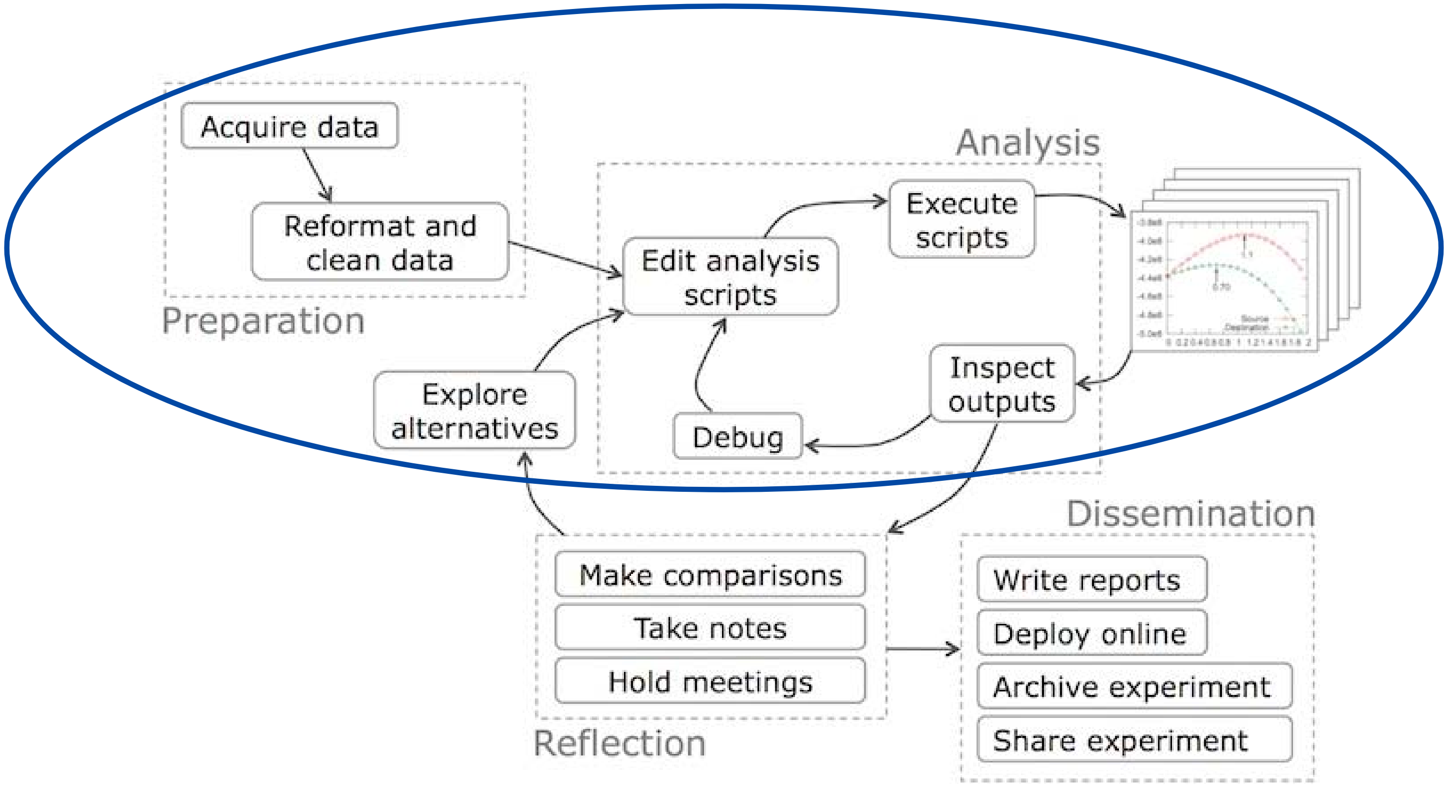
Exercises

- 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



Preparation

Acquire data

Reformat and clean data

Analysis

Edit analysis scripts

Execute scripts

Inspect outputs

Debug

Explore alternatives

Reflection

Make comparisons

Take notes

Hold meetings

Dissemination

Write reports

Deploy online

Archive experiment

Share experiment

Preliminaries: Types of Data

Data Types

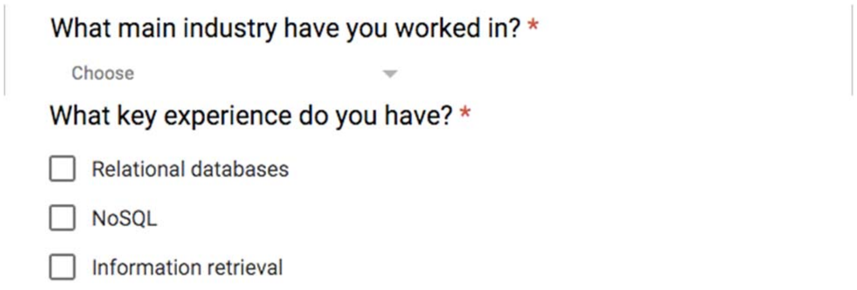
- Text
- Images
- Videos
- Categorical
  - Nominal
    - Dichotomous
  - Ordinal
- Quantitative
  - Interval
  - Ratio

Categorical Data

- A categorical variable is also known as a discrete or qualitative variable and can have two or more categories. It is further divided into two variants, nominal and ordinal. These variables are typically coded as numerical values.

Nominal Data

- This is an unordered category data. This type of variable may be coded in numeric form but these numerical values have no mathematical interpretation and are just labeling to denote categories. For example, colours: black, red and white can be coded as 1, 2 and 3.



What main industry have you worked in? *

Choose

What key experience do you have? *

□ Relational databases

□ NoSQL

□ Information retrieval

- Values are names
- No ordering is implied
- Eg jersey numbers

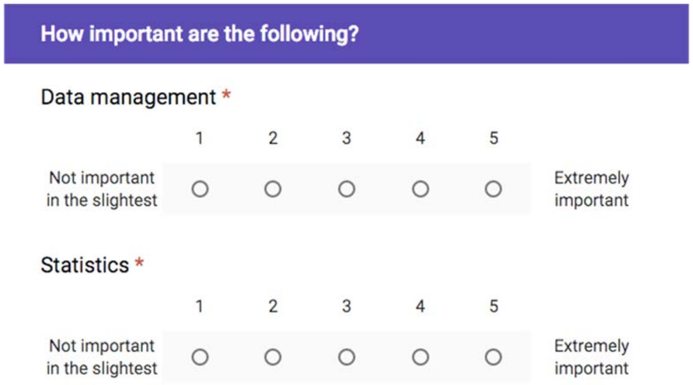
Dichotomous Data

- A dichotomous is a type of nominal data that can only have two possibile values, e.g. true or false, or presence or absence. These are also sometimes referred as binary or Boolean variables.

- True (1) or false (0)
- Gender: male / female

Ordinal Data

- This is ordered categorical data in which there is strict monotonic order. For example, human height (small, medium and high) can be coded into numbers small = 1, medium = 2, high = 3.



How important are the following?

Data management *

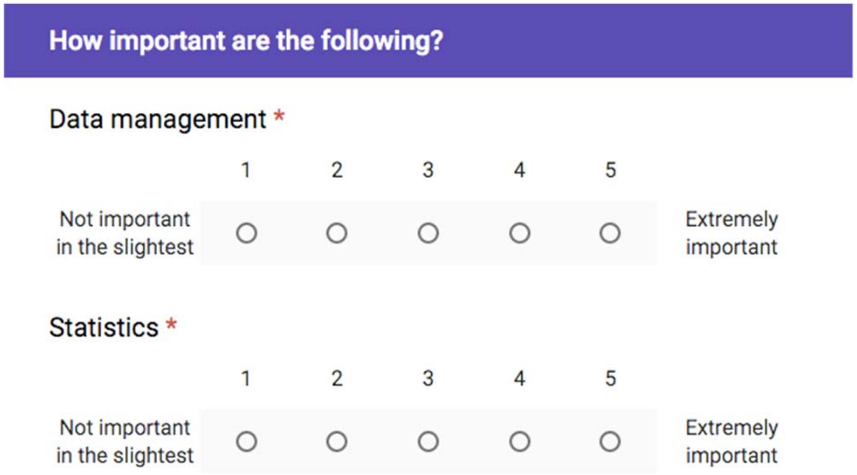
	1	2	3	4	5	
Not important in the slightest	○	○	○	○	○	Extremely important

Statistics *

	1	2	3	4	5	
Not important in the slightest	○	○	○	○	○	Extremely important

- Values are ordered
- No distance is implied
- Eg rank, agreement

Ordinal Data



How important are the following?

Data management *

	1	2	3	4	5	
Not important in the slightest	○	○	○	○	○	Extremely important

Statistics *

	1	2	3	4	5	
Not important in the slightest	○	○	○	○	○	Extremely important

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

We can’t say that the difference between “OK” and “Unhappy” is the same as the difference between “Very Happy” and “Happy?”

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

Ordinal Data

– How to calculate the median:

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

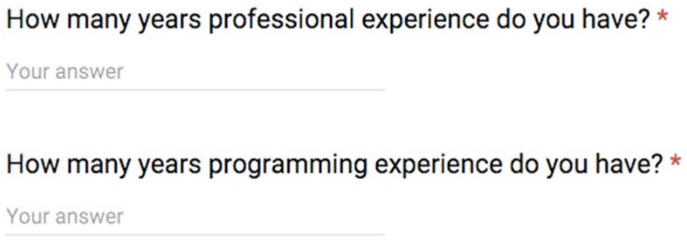
“Thermometer” by Christer Edvartsen is licenced under CC BY 2.0

- 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 Celsius temperature

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

Ratio Data

It is variable that has a true value of zero and represents the total absence of the variable being measured. For example, it makes sense to say a Kelvin temperature of 100 is twice as hot as a Kelvin temperature of 50 because it represents twice as much the thermal energy (unlike Fahrenheit temperatures of 100 and 50).



How many years professional experience do you have? *

Your answer

How many years programming experience do you have? *

Your answer

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

Calculating descriptive statistics

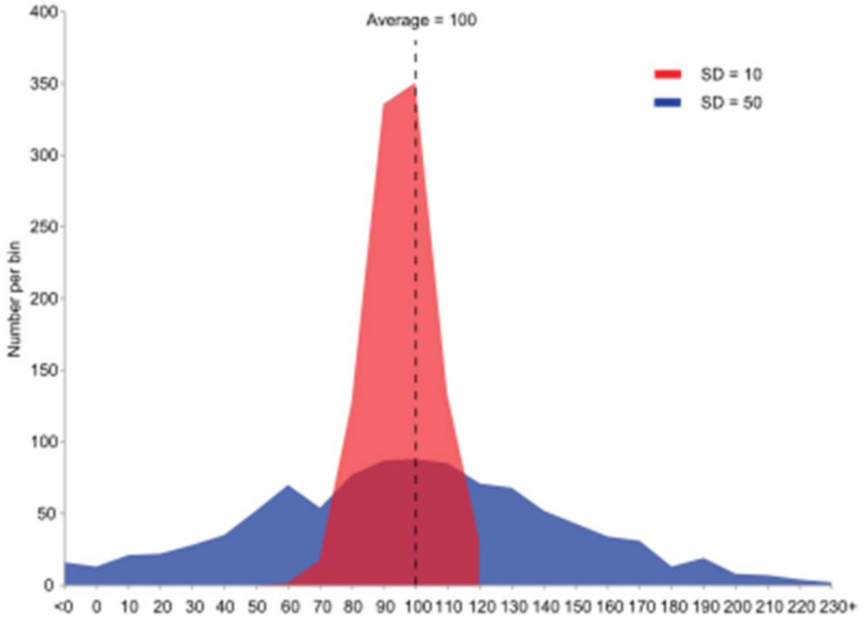
- ***Median*** and ***percentiles*** good here too
- ***Mean*** is the sum of values divided by the number of values:

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

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

- ***Standard deviation:*** $$\sqrt{\mathrm{var}\,iance}$$

What does variance and standard deviation tell us?



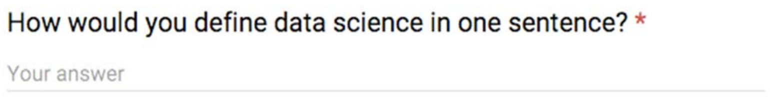
Samples from two populations with the same mean but different variances. The red population has mean 100 and variance 100 (SD=10) while the blue population has mean 100 and variance 2500 (SD=50).

https://en.wikipedia.org/wiki/Variance

Levels of Measurement

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

What about text data?



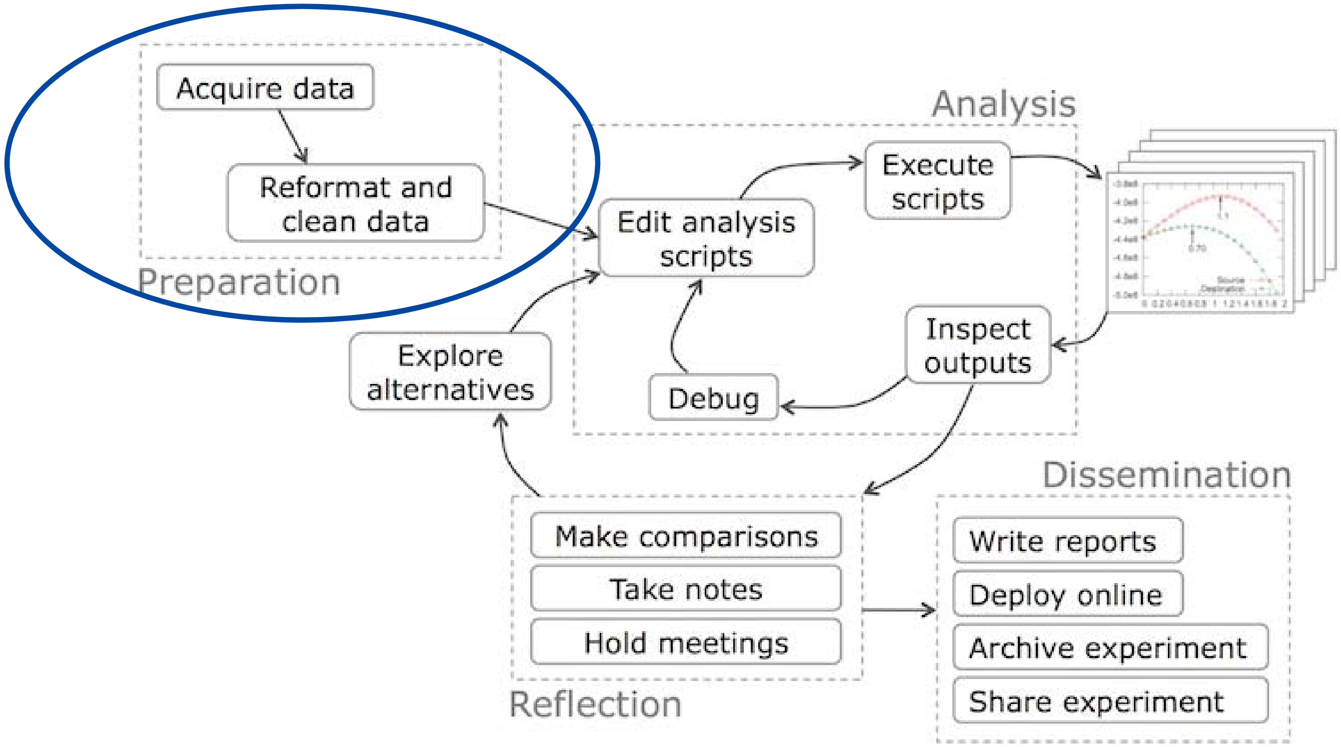
How would you define data science in one sentence? *

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

Data Acquisition and Cleaning

THE UNIVERSITY OF SYDNEY

Exploratory Analysis Workflow

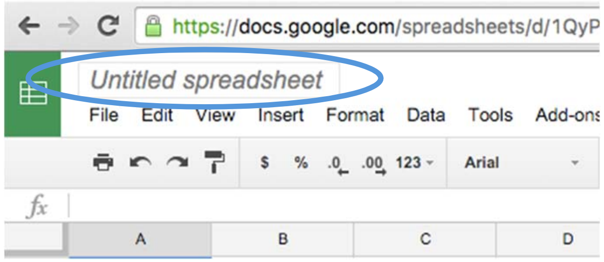


Data Acquisition – Where does data come from?

- 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
  - Scrapping 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 our online survey from Week 1

Exercise: Acquire data



- Create new Google spreadsheet
  - Go to https://docs.google.com/spreadsheets
  - File > New > Spreadsheet
  - Rename “COMP5310 `Survey analysis` **`(NAME, UNIKEY)`**”
- Download response data: https://goo.gl/5YK2g6 (link on Canvas)
- Google Sheets File > Make a copy
- If you need to import data from a CSV file:
  - Google Sheets File > Import
  - Click on Upload
  - Select and load: `Survey_COMP5310_2018s2 - Form Responses 1.csv`

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

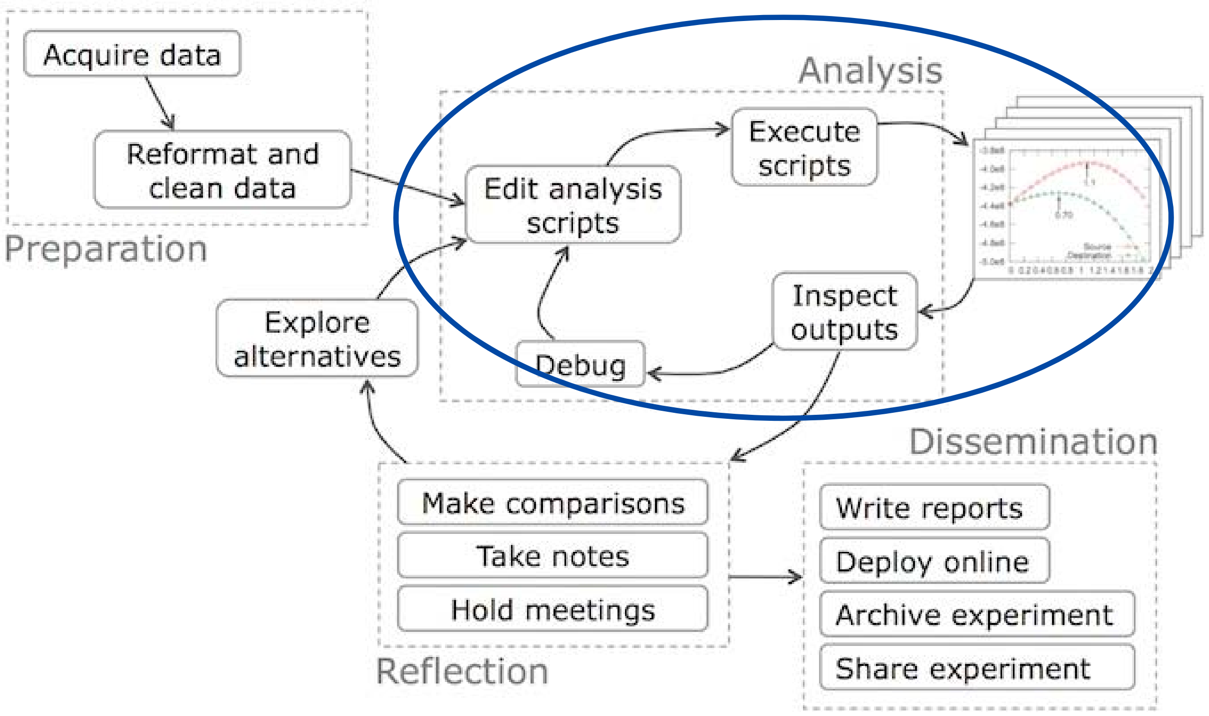
- Groups of 6

- 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

What Questions Can We Answer?

Exploratory Analysis Workflow



Exercise: What questions can we ask?

- In your group, discuss:
  - Review survey questions from W1: https://goo.gl/EZ2QS3
  - List 3 questions we can ask
  - Discuss how you would answer each question with this data

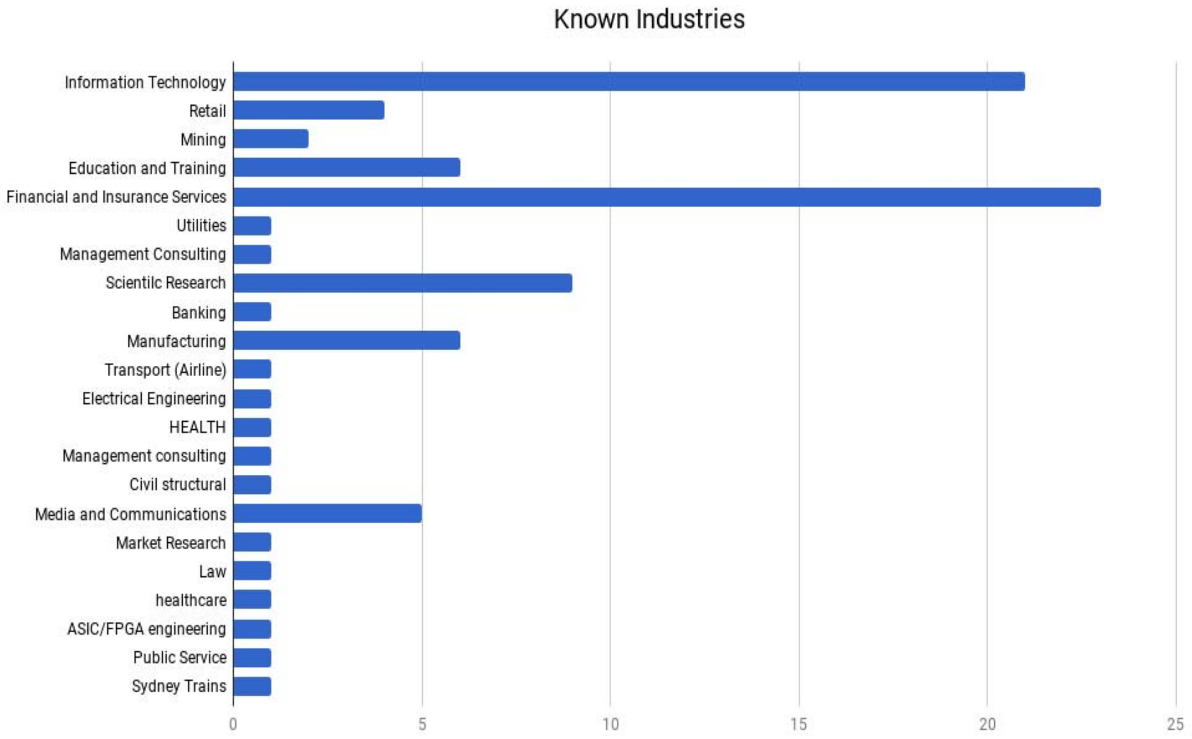
Some descriptive questions

- What areas of data science are considered important?
- How do professional/programming experience compare?
- How does programming experience differ across industries?
- What skills do we know? What would we like to learn?
- Which industries are most desirable? Do past/future differ?
- What skills co-occur most? How strong is the association?

Summarising Nominal Data:

What industries do we know? What would we like to go into?

Summarise nominal data with histograms



Measures of central tendency:

- mode

Measures of dispersion:

- counts/distribution%

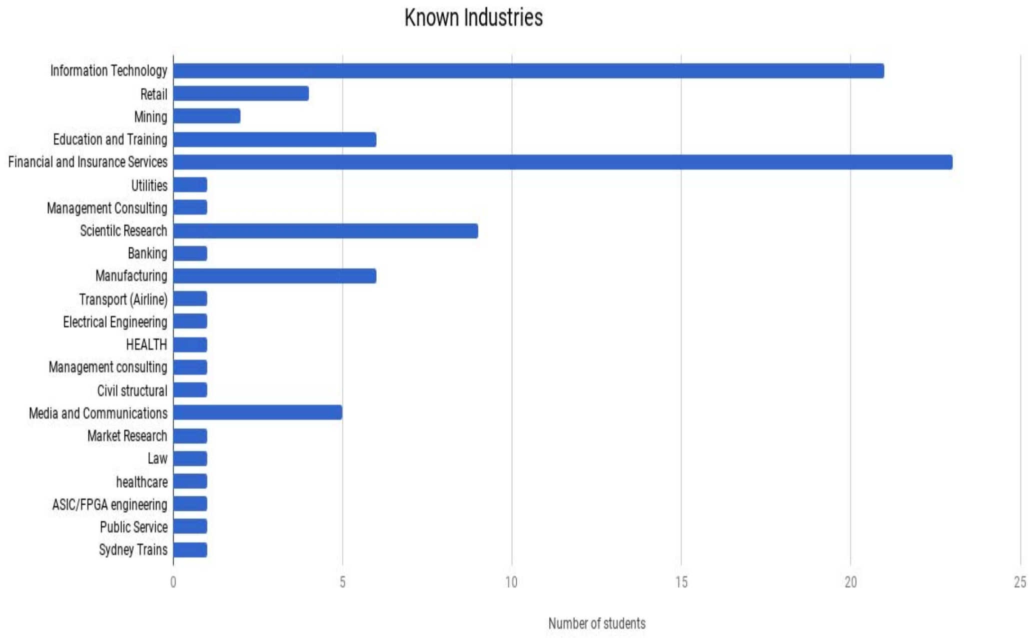
Calculating the Mode

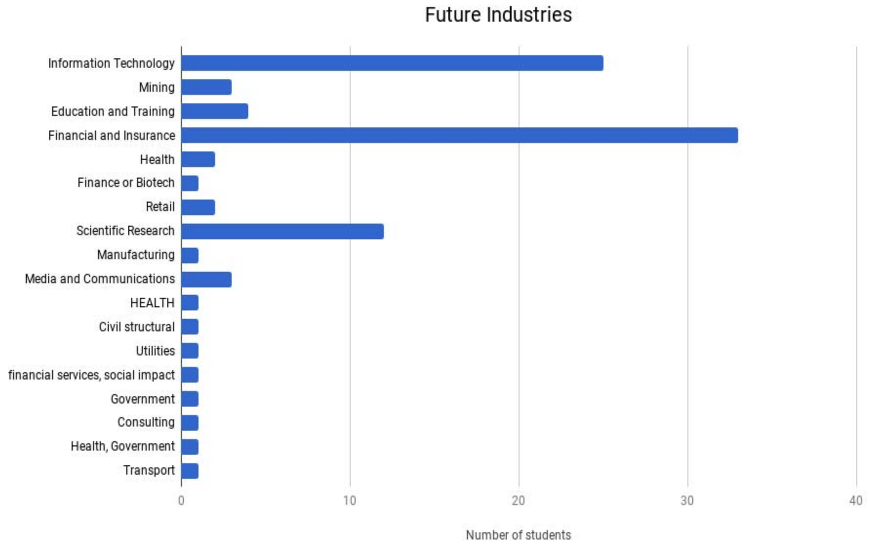
- The most frequent value
- Defined for nominal data, but spreadsheets might not compute
- Can read from a histogram

Creating Histograms (bar charts)

- Count frequency of each category
- Display on bar chart
- In Google Sheets
  - Select data range (e.g., C1:C96)
  - Click “insert chart” icon 📊
  - Ensure on “DATA” tab under, "Use row 1 as headers“ and “Aggregate column C" are selected.
  - Change the chart type, select “bar chart”

Histograms comparing known and future industries





Exercise: Exploring nominal data

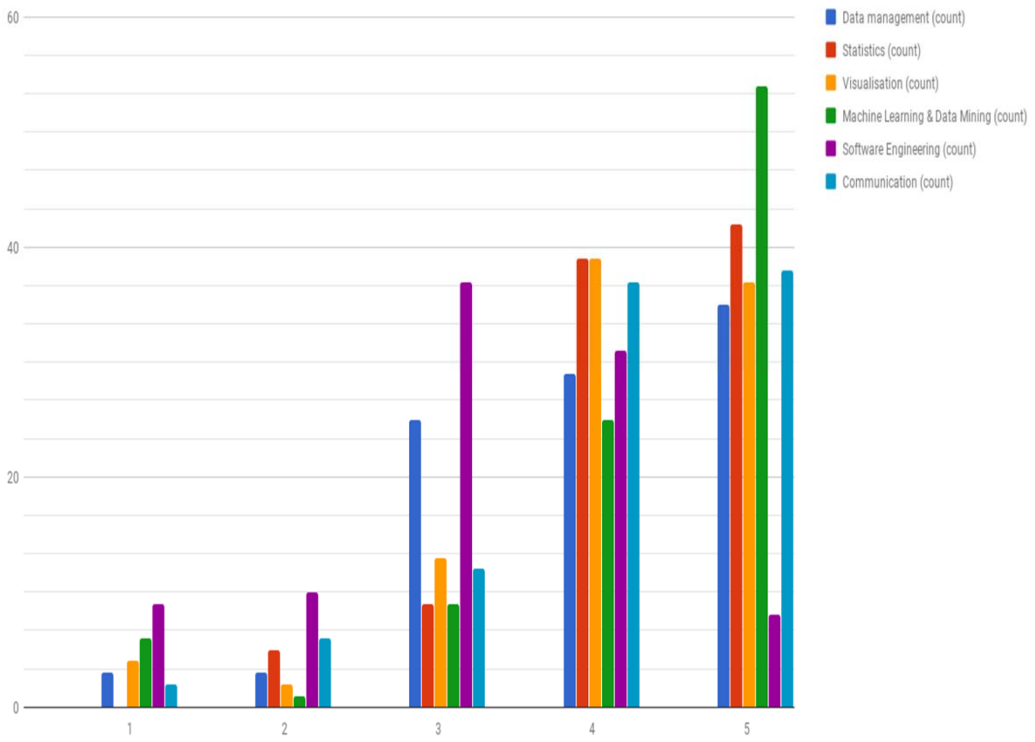
- Visualise:
  - Create histograms of known and future industries
- Discuss:
  - What do we need to do to make these comparable?
  - What is the mode?

Summarising Ordinal Data:

What areas of data science are considered important?

THE UNIVERSITY OF SYDNEY

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

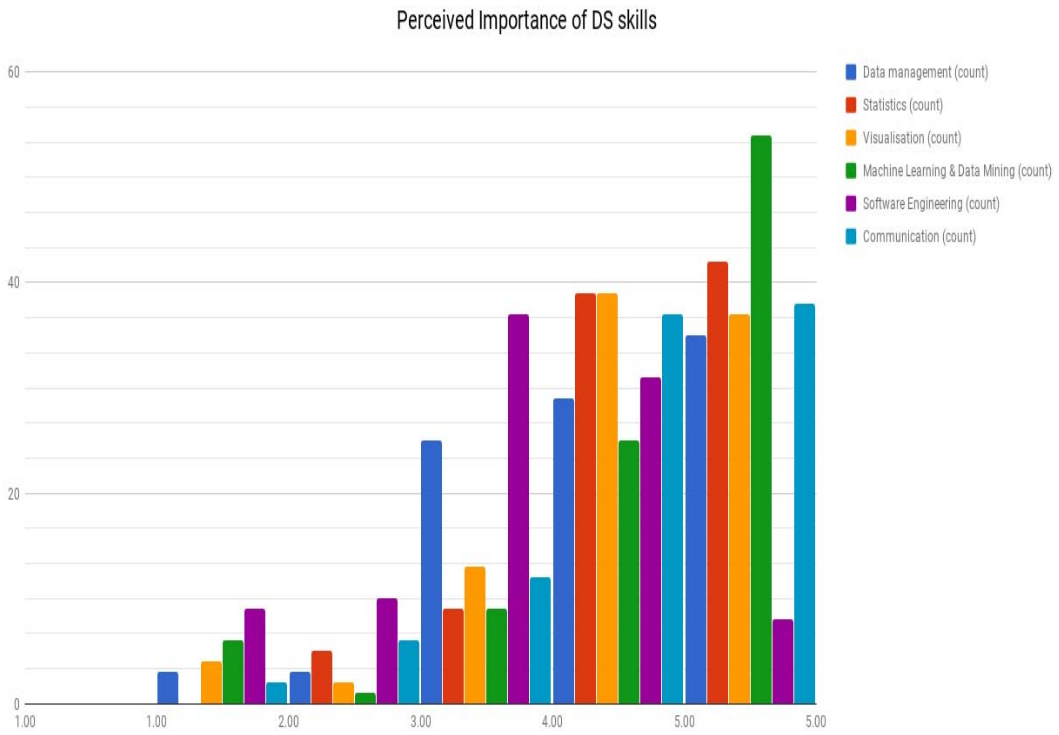
Creating a Histogram chart

- Count frequency, e.g., of ordinal values within each category
- Display on histogram chart with one variable grouped inside
- In Google Sheets
  - Select data range (e.g., G1:L96)
  - Click “insert chart” icon 📊
  - For chart type, select “Histogram Chart”
  - On “DATA” tab under "Series", select “Use row 1 as headers”
  - Configure rest to your liking

Exercise: Exploring ordinal data

- Visualise:
  - Create a histogram diagram of area importance ratings
- Discuss:
  - Which area gets the most high rankings?
  - Are there interesting differences between areas?
  - Do medians differ? Ranges?

Histogram chart comparing areas of data science



Good:

- Illustrates tendency
- Areas differentiated

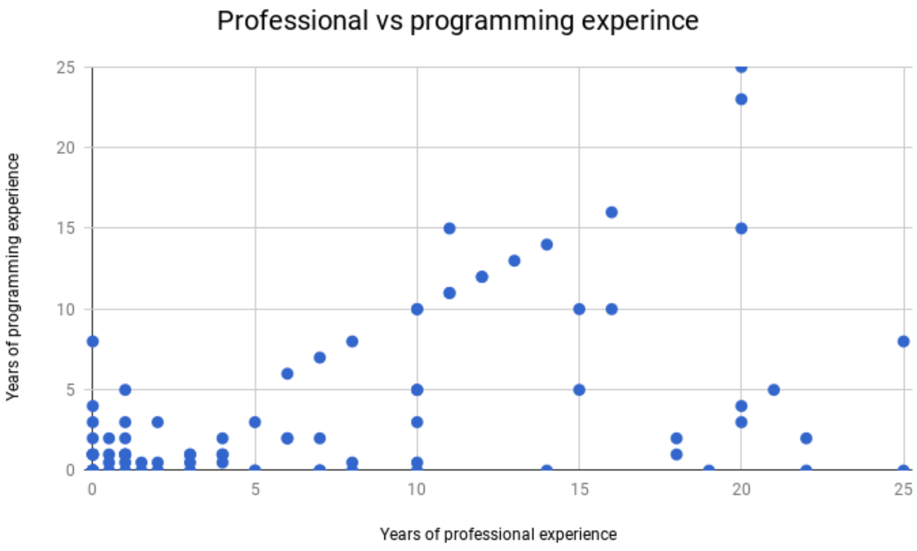
Bad:

- buckets on x-axis not clear and no clear separation
- no axis titles (add manually)

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:***

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

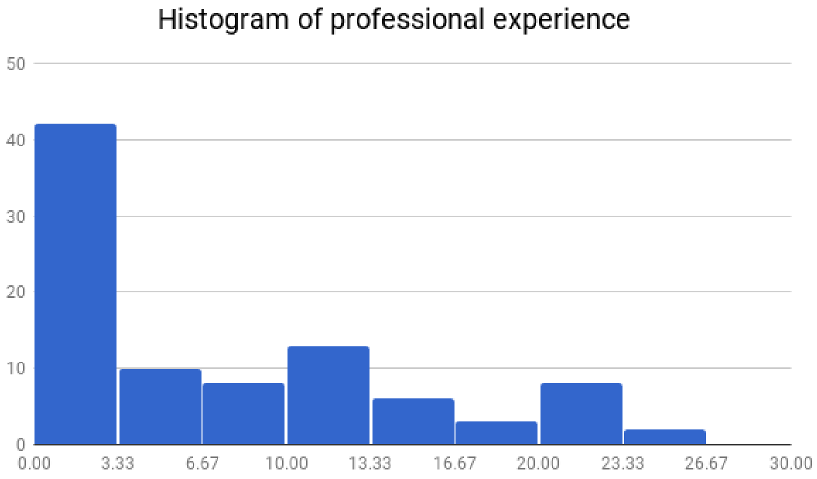
- ***Standard deviation:***

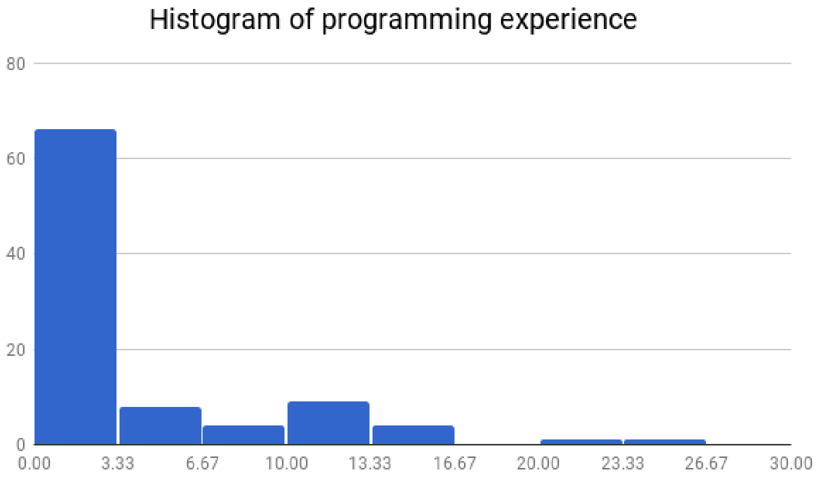
$$\sqrt{variance}$$

Creating a Scatterplot

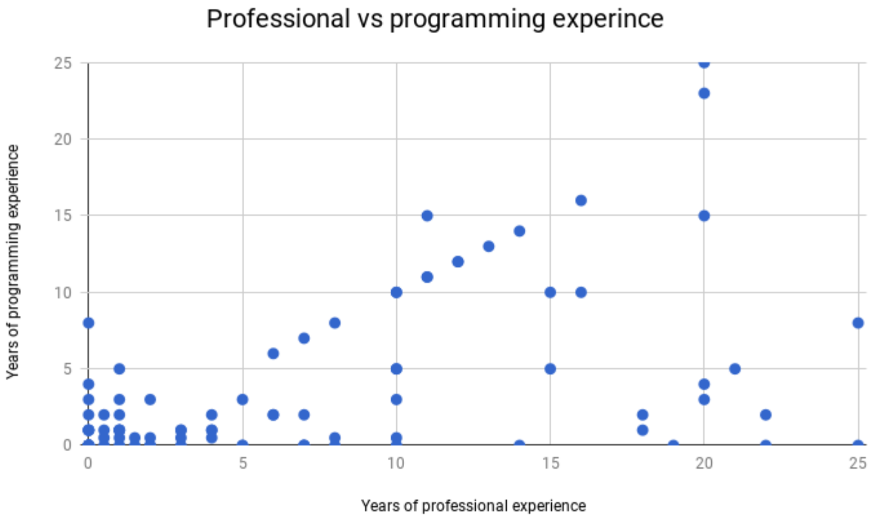
- Plots relationship between two different variables
- Display, e.g., professional experience on x-axis vs. programming experience on y-axis for each respondent
- In Google Sheets
  - Select data range (e.g., D1:E96)
  - Click “insert chart” icon 📊
  - Select chart type “Scatter Chart”

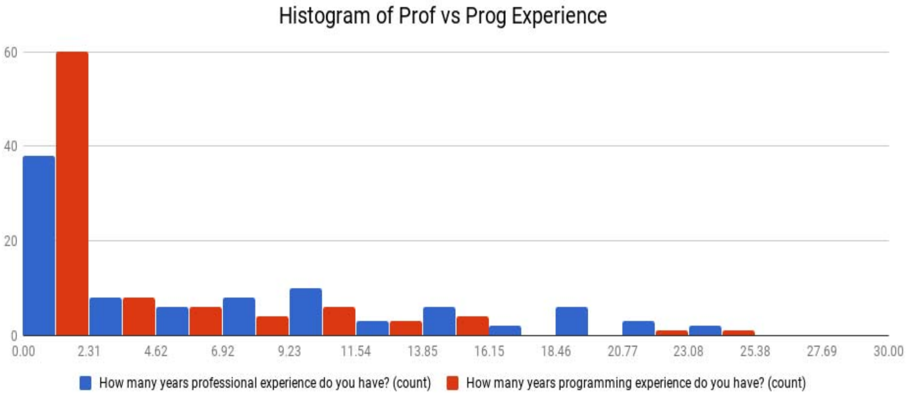
Binned histograms for experience





Comparison with scatterplot and histogram overlays





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?

Pivot Tables:

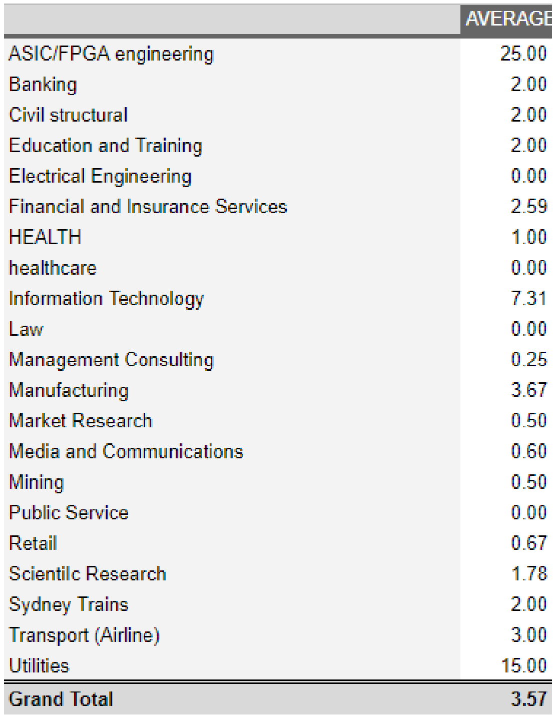
How does programming experience differ across industries?

THE UNIVERSITY OF SYDNEY

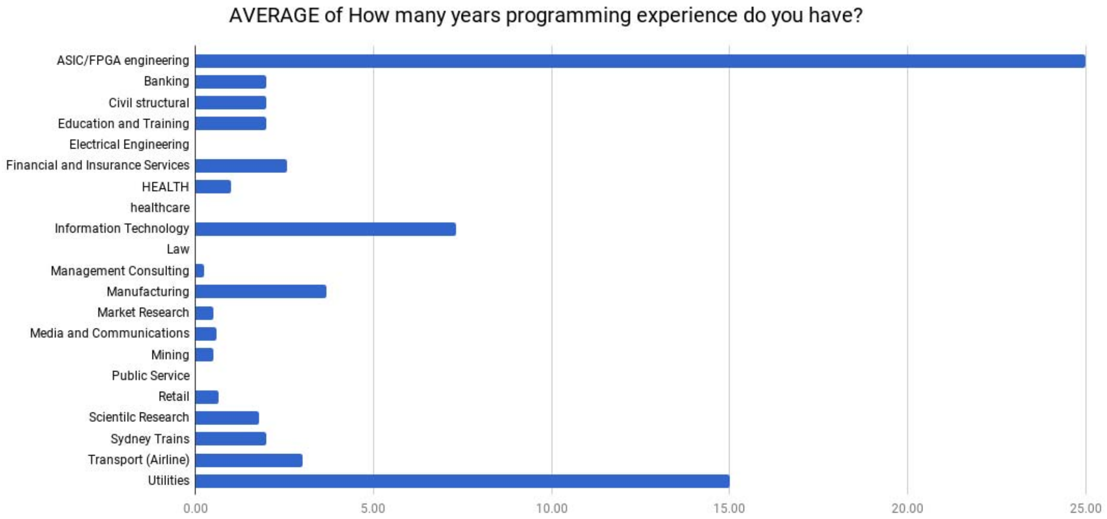
Creating a pivot table

- Summarise data by calculating statistics over sub-populations
- E.g., average programming experience by industry
- In Google Sheets
  - Select data range (e.g., C1:E86)
  - Go to Data > Pivot Table (should insert a new sheet)
  - Select industry under row
  - Select professional experience under value
  - Summarise by average

Table and Histogram of programming by industry



	AVERAGE
ASIC/FPGA engineering	25.00
Banking	2.00
Civil structural	2.00
Education and Training	2.00
Electrical Engineering	0.00
Financial and Insurance Services	2.59
HEALTH	1.00
healthcare	0.00
Information Technology	7.31
Law	0.00
Management Consulting	0.25
Manufacturing	3.67
Market Research	0.50
Media and Communications	0.60
Mining	0.50
Public Service	0.00
Retail	0.67
Scientilc Research	1.78
Sydney Trains	2.00
Transport (Airline)	3.00
Utilities	15.00
Grand Total	**3.57**



Exercise: Using a pivot table to summarise data

- Pivot table:
  - Create a table of average programming experience by industry
- Discuss/explore:
  - What other statistics can we calculate?
  - What other variable combinations could we explore?

Complex Counting:

What skills do we know? What would we like to learn?

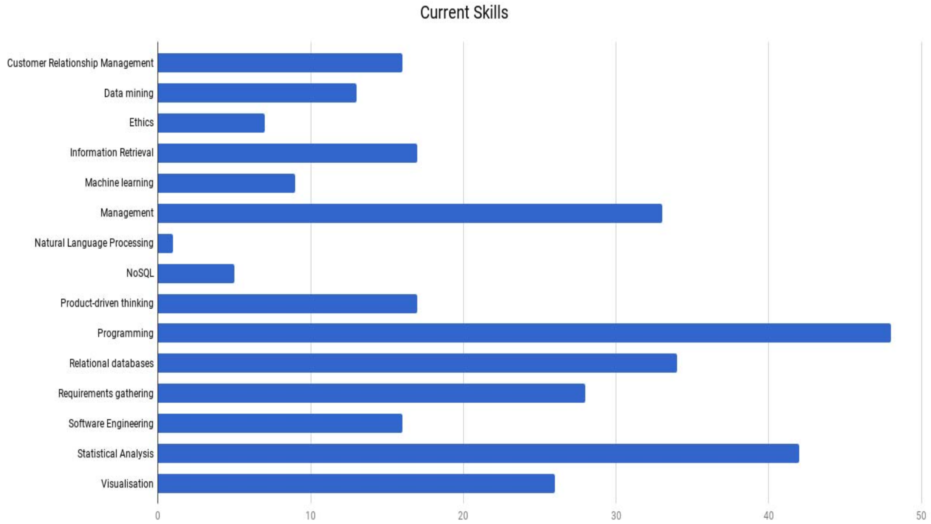
THE UNIVERSITY OF SYDNEY

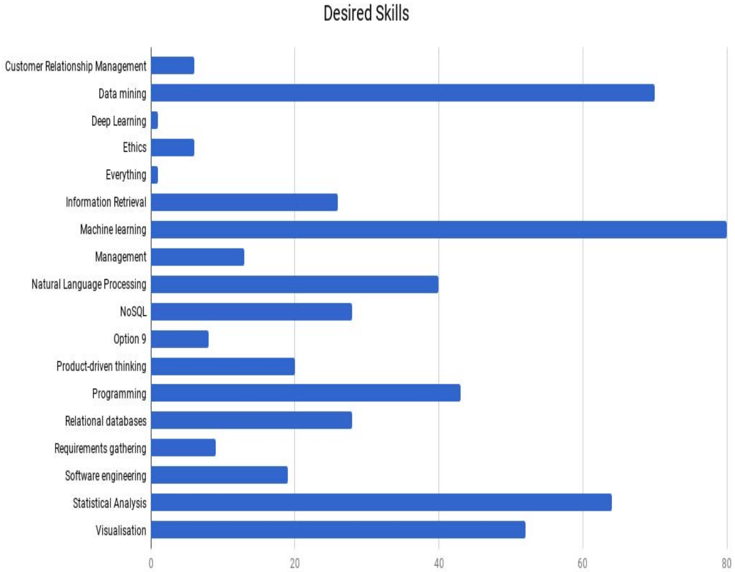
How to create a histogram of skills?

- Multiple values in cells within the skills column, e.g.:
  `“Software engineering, Requirements gathering, Product-driven thinking”`
- Need to split possible values:
  `=sort(unique(transpose(split(join(", ", N2:N96), ", ", False))))`
- Then count:
  `=countif(N$2:N$96, 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 2016…)





Review

THE UNIVERSITY OF SYDNEY

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

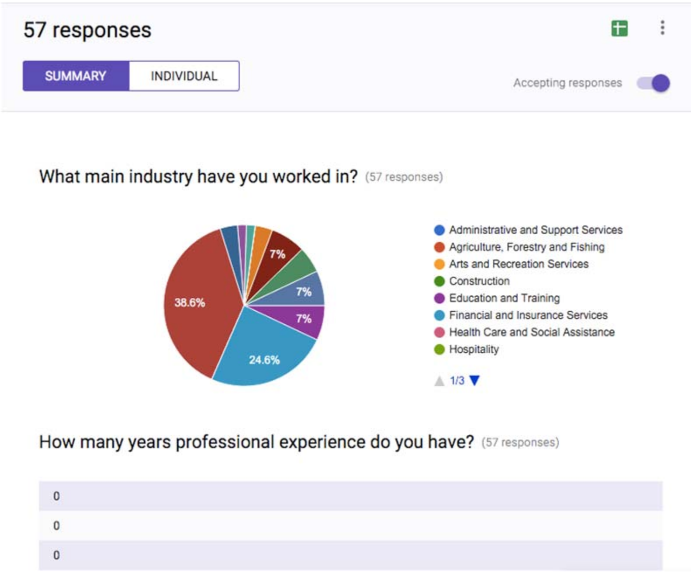
	Nominal	Ordinal	Interval	Ratio
Mode	✓	✓	✓	✓
Median		✓	✓	✓
Mean			✓	✓

Measures of Dispersion

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

Google's answer to exercises…

- Google Forms provides a summary
- Useful but not perfect
  - Legend includes 0 labels
  - Pie chart doesn't show N
  - Does not clean noisy input
  - Labels incomplete
  - Does not handle text



57 responses

SUMMARY
INDIVIDUAL
Accepting responses

What main industry have you worked in? (57 responses)

7%
7%
38.6%
7%
24.6%
● Administrative and Support Services
● Agriculture, Forestry and Fishing
● Arts and Recreation Services
● Construction
● Education and Training
● Financial and Insurance Services
● Health Care and Social Assistance
● Hospitality
1/3

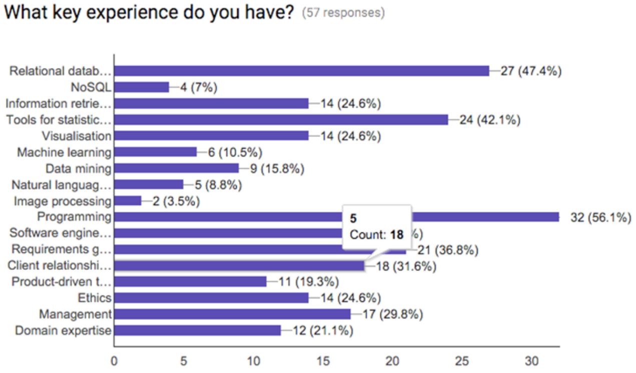
How many years professional experience do you have? (57 responses)

0

0

0

⋮



What key experience do you have? (57 responses)

Tips and Tricks

- Data cleaning important for any meaningful analysis
- Spreadsheet software is good for quick interactive analysis
  Need programmatic analysis for bigger/complex data
- Careful about which types of data allow what kind of measures & viz.
- Measures of central tendancy (e.g., mean) are not sufficient
  Always explore and communicate spread as well (e.g., stdev)
- Good visualisations help convey distributions and relationships
  - Label all plots and diagrams with readable and visible fonts
  - Use same axis bounds when comparing plots
  - Use meaningful axis bounds to convey effect size
    (50-55 on a 100 point scale over-sells small differences)
  - Design so comparison/effect is clear, include description of axes

Next Time

THE UNIVERSITY OF SYDNEY

Lecture plan

- W1: Introductions and housekeeping
- W2: Data Acquisition & Exploration I
- W3: Data Exploration with Python
- W4: Cleaning and storing data
- W5: Querying and summarising data
- W6: Hypothesis testing
  Project stage 1 due
- W7: Data Mining
- W8: Machine learning
- W9: From data to decisions
- W10: Unstructured data
- W11: Analysing big data
- W12: Product Thinking and Ethics*
  Project stage 2 due
- W13: Review
- ***Exam***

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 16 and 17
- Explore and select project data

Project Stage 1

THE UNIVERSITY OF SYDNEY

Project stage 1: Explore, Clean, Pitch

Objective

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

Activities

- Choose a data set
- Explore, summarise and prepare data
- Define problem, specify requirements

Output

- 2-page report summarising problem, analysis and proposal (plus code)
- 1-page technical summary

Marking

- 10% of overall mark

Suggested timeline for Assignment 1 (Project Stage 1)

- W1: Identify possible data sets
- W2: Identify & Explore possible data sets
- W3: Select project data set
- W4: Draft summary (problem & exploratory analysis)
- W5: Clean and prepare data
- W6: Submit 2-page report (summary + stage 2 proposal)

Types of projects to consider

- Analyse and quantify difference between two populations
- Develop alternative visualisations and test effectiveness
- Test for correlation between populations or attributes
- Discover clusters in data or learn association rules
- Train a classifier and evaluate prediction accuracy

Project and discussion time

Time for you to talk
to tutors, instructors and each other
about data sets, data exploration
and possible research questions.