

**INFO3406**

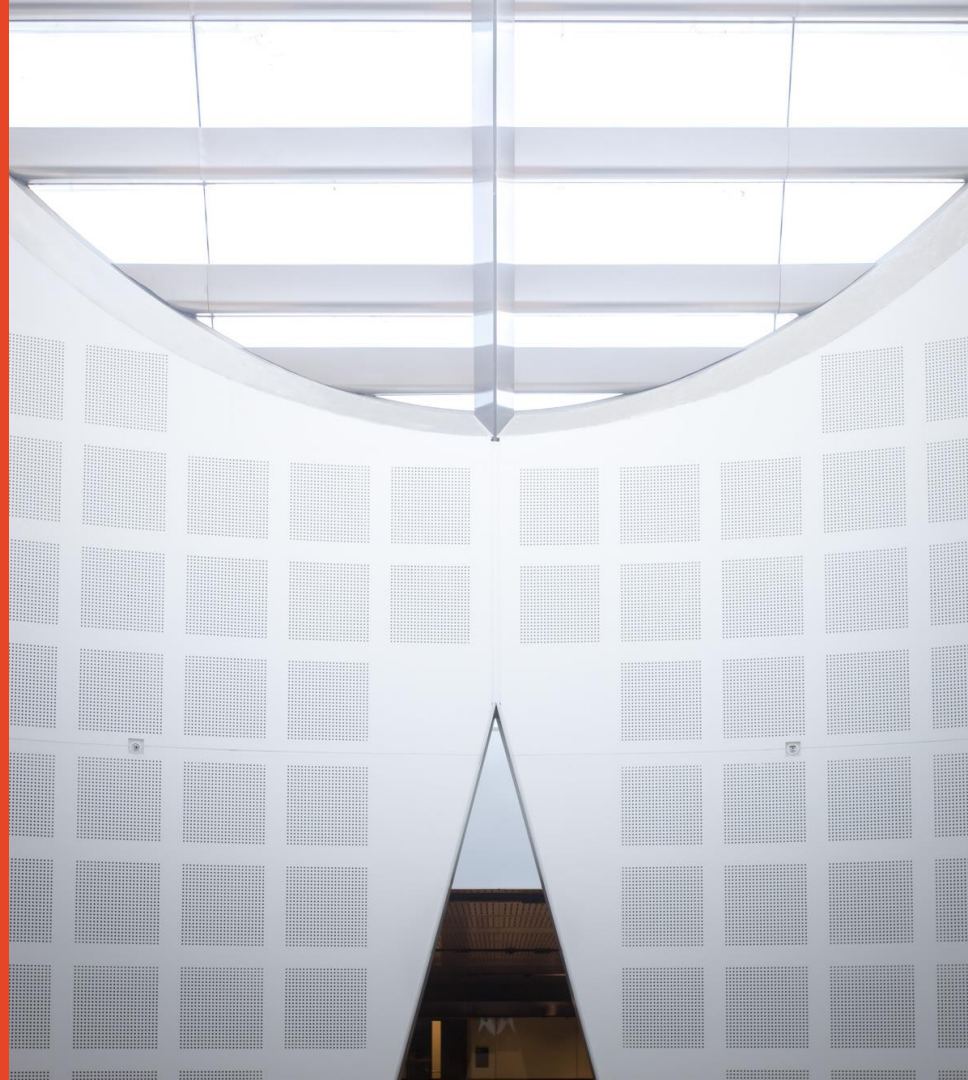
# **Introduction to Data Analytics**

## **W2: Data Acquisition and Exploration**

**Presented by**

Dr Ali Anaissi

School of IT



# 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 1-4
- 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

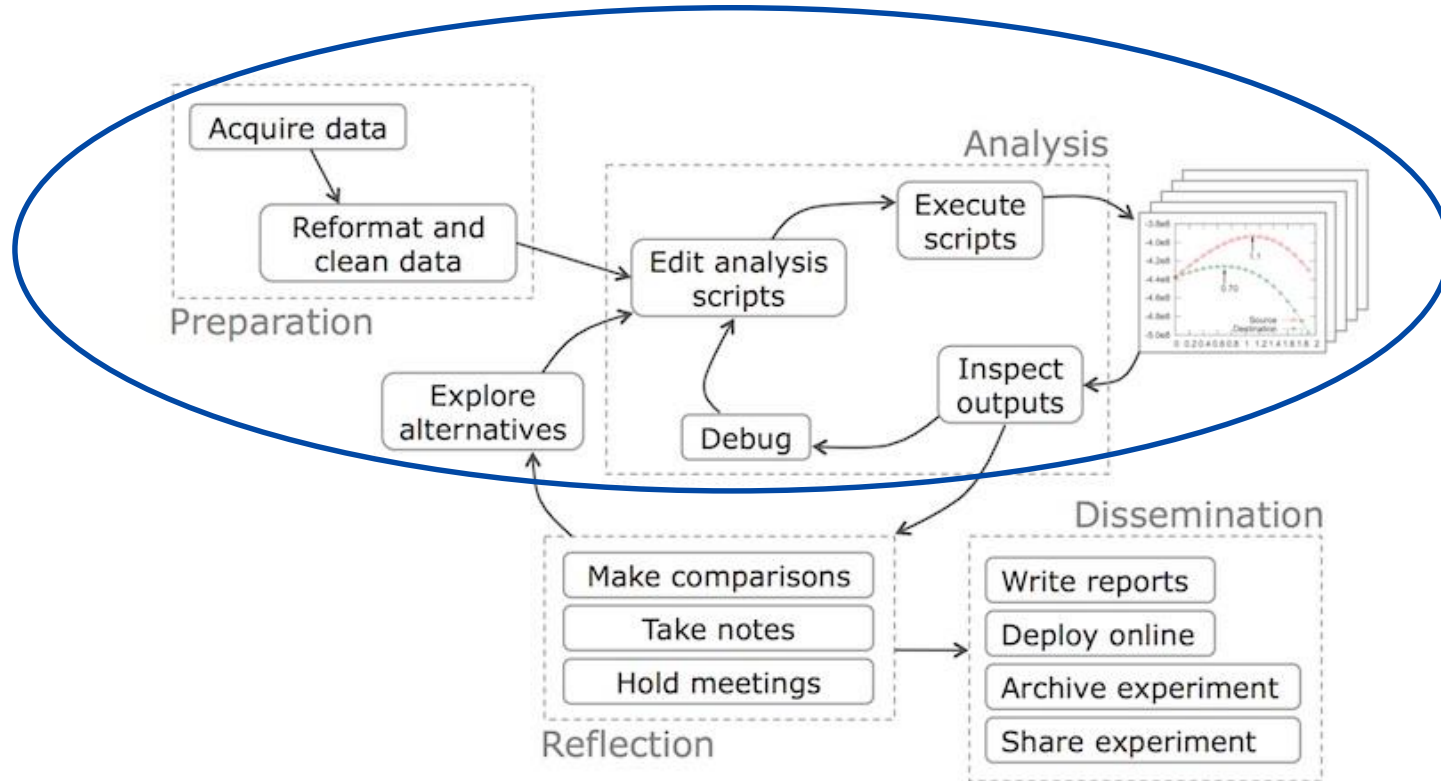
## Exercises

- Google Sheets: Visualisation
- Google Sheets: Descriptive stats

## TODO in W2

- Grok Python modules 1-4
- Grok SQL modules 1-4
- Explore project data

# Exploratory Analysis Workflow



# Preliminaries: Types of Data

# Nominal Data

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

# Ordinal Data

## How important are the following?

Data management \*

	1	2	3	4	5	
Not important in the slightest	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Extremely important

Statistics \*

	1	2	3	4	5	
Not important in the slightest	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Extremely important

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

<sup>1</sup>The mode is the number that is repeated more often than any other



# Ordinal Data

- 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**,3,3,3,3,4,4,4,4,4,4,4,4,4,4,4,4,4,4,5,5,5,5,5,5]

- 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]  
[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]

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

- The IQR is the difference between the first and third quartile. i.e:  
 $Q3 - Q1 = 4 - 3 = 1.$

# Interval Data



- 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

“[Thermometer](#)” by [Christer Edvartsen](#) is licenced under [CC BY 2.0](#)

# Ratio Data

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

# 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? \*

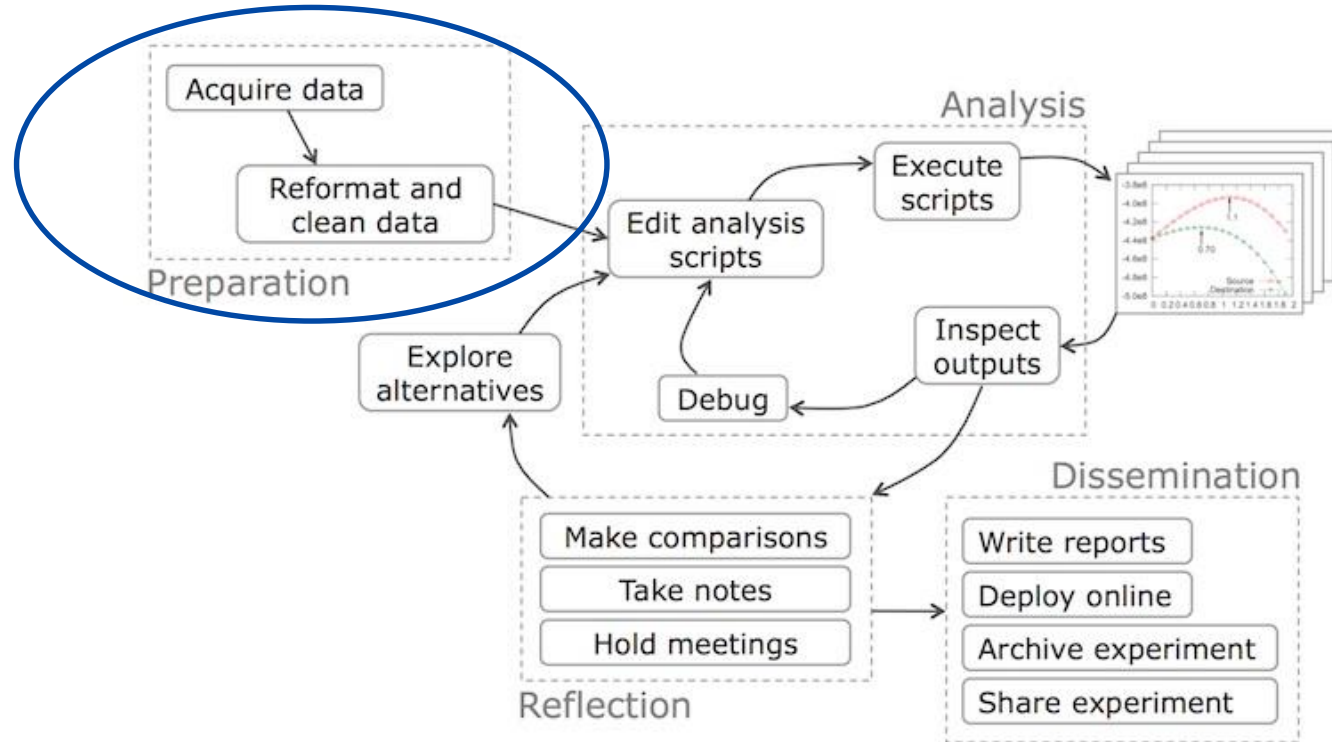
Your answer

---

- 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

- Scrapping the web (HTML)
  - or using APIs of Web Services (XML/JSON)
    - > Cf. textbook, Ch 9

- Database Access -> Week 4 onwards

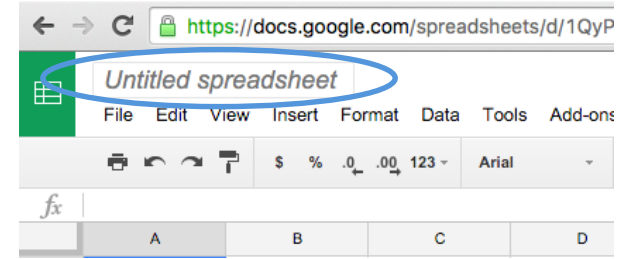
- Collect data yourself, eg. via a survey

**This week:** Using data from our online survey from Week 1



# Acquire data

- Create new Google spreadsheet
  - Go to <https://docs.google.com/spreadsheets>
  - File > New > Spreadsheet
  - Rename “INFO3406 Survey analysis (NAME, UNIKEY)”
- Download response data: <https://goo.gl/n8U7Mb>  
(link on Canvas)
- Load survey data
  - File > Import
  - Click on Upload
  - Select and load: Survey INFO3406 2018s2 (Responses) - Form Responses 1.csv

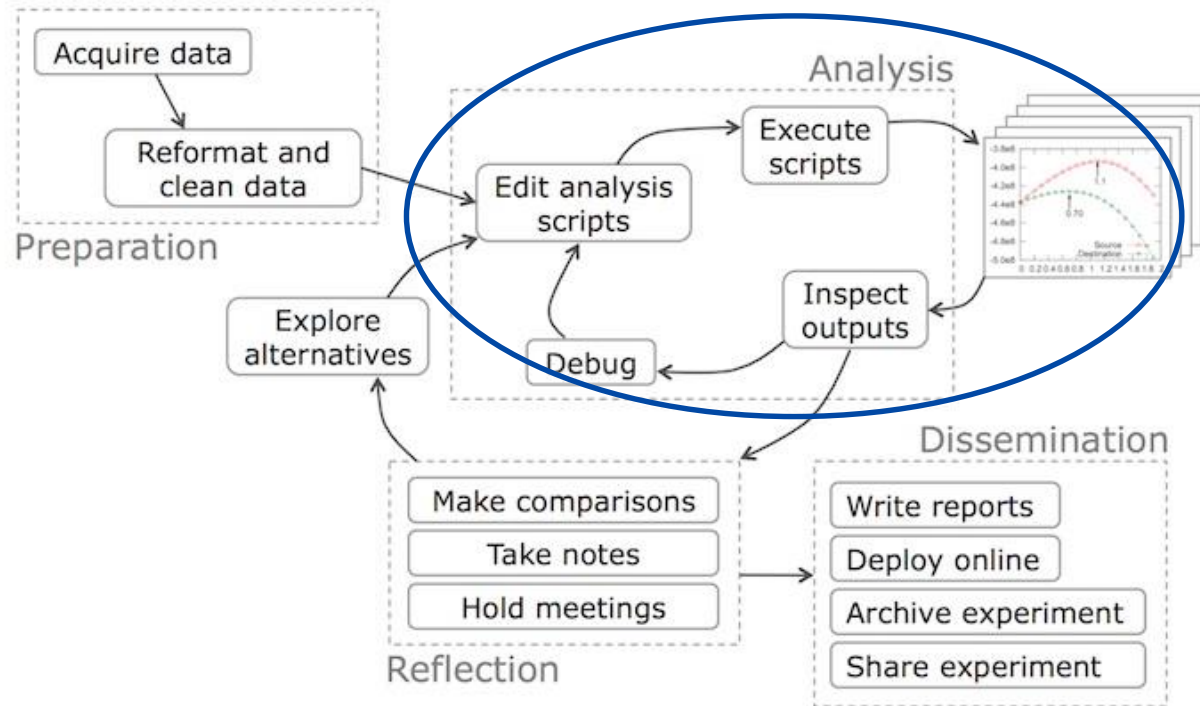


# Cleaning and Transforming Data

- 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

# What Questions Can We Answer?

# Exploratory Analysis Workflow



## Some descriptive questions

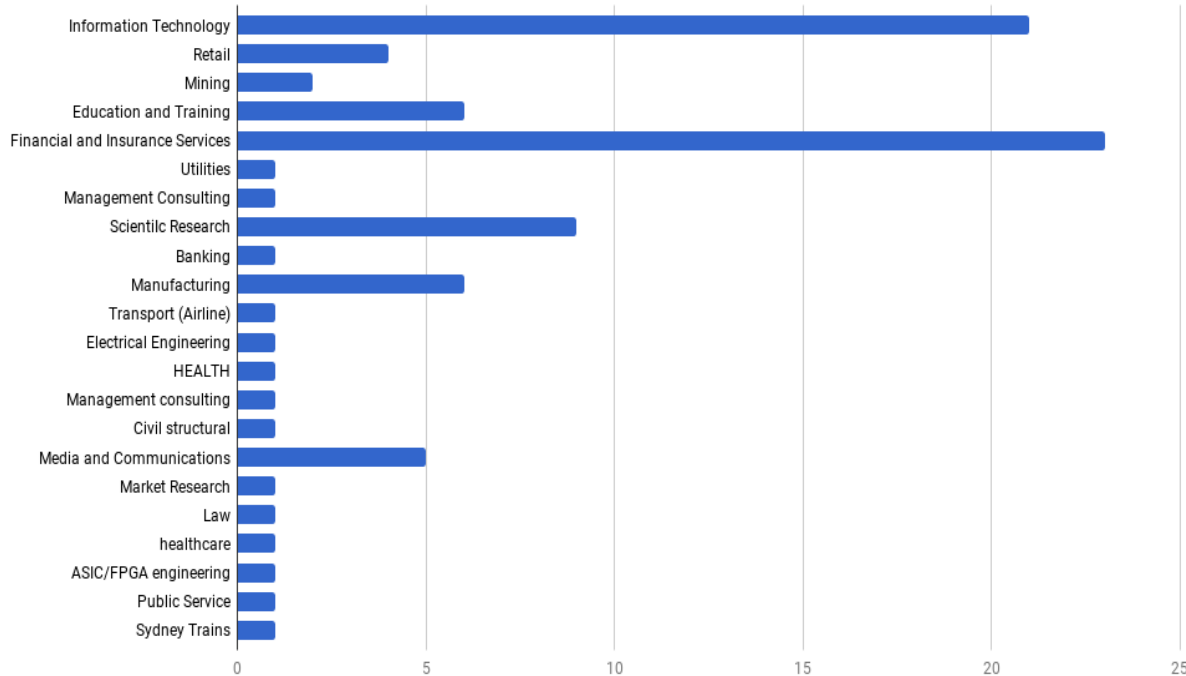
- What areas of data analytics are considered important?
- How do professional/programming experience compare?
- How does programming experience differ across industries?
- What skills do we know? What would we like to learn?
- Which industries are most desirable? Do past/future differ?
- What skills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

Known Industries



**Measures of central tendency:**

— mode

**Measures of dispersion:**


— counts/distribution%

# Calculating the Mode

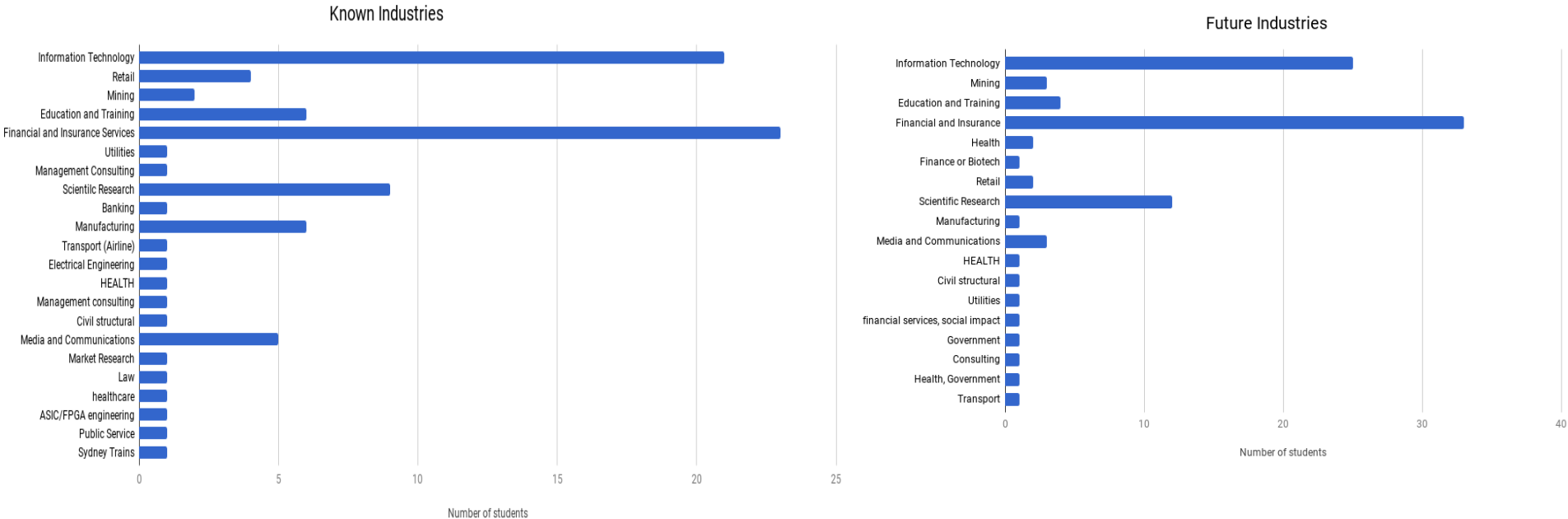
- The most frequent value
- Defined for nominal data, but spreadsheets might not compute
- Can read from a histogram



# Creating Histograms (bar charts)

- Count frequency of each category
- Display on bar chart
- In Google Sheets
  - Select data range (e.g., C1:Cn\*)      \*n is the total number of responses
  - Click “insert chart” icon 
  - On “DATA” tab under "Series", select “Aggregate column C” and "Use row 1 as headers"
  - Change the chart type, select “bar chart”

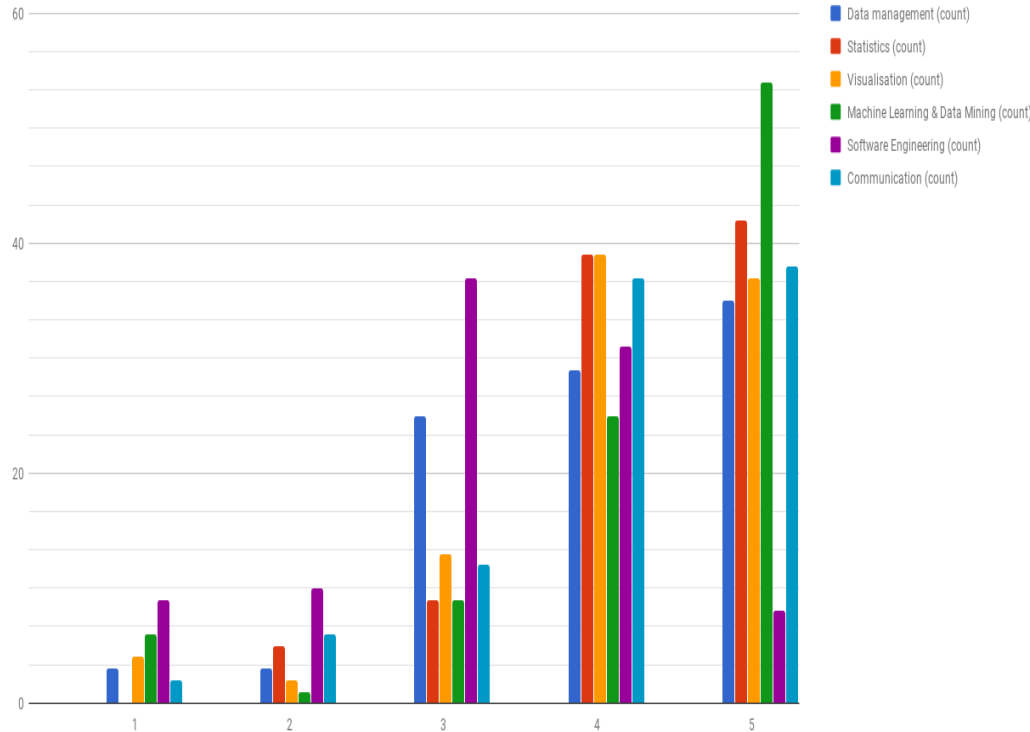
# Histograms comparing known and future industries



## Summarising Ordinal Data:

*What areas of data analytics are considered important?*

# Summarise ordinal data: histograms, median, percentiles



## Measures of central tendency:

- median, mode


## Measures of dispersion:

- counts/distribution
- min/max/range
- percentiles

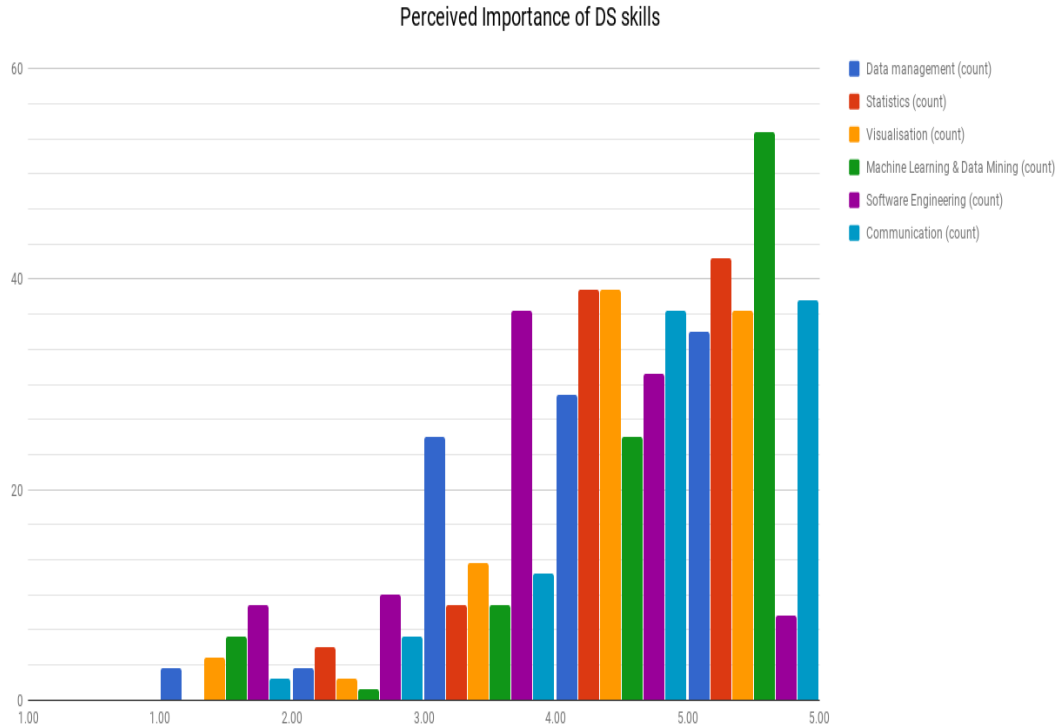
# Calculating descriptive statistics

- 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
  - **10<sup>th</sup> percentile** is item at index  $0.1 * N$
  - **90<sup>th</sup> 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:Ln)
  - 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

# 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)

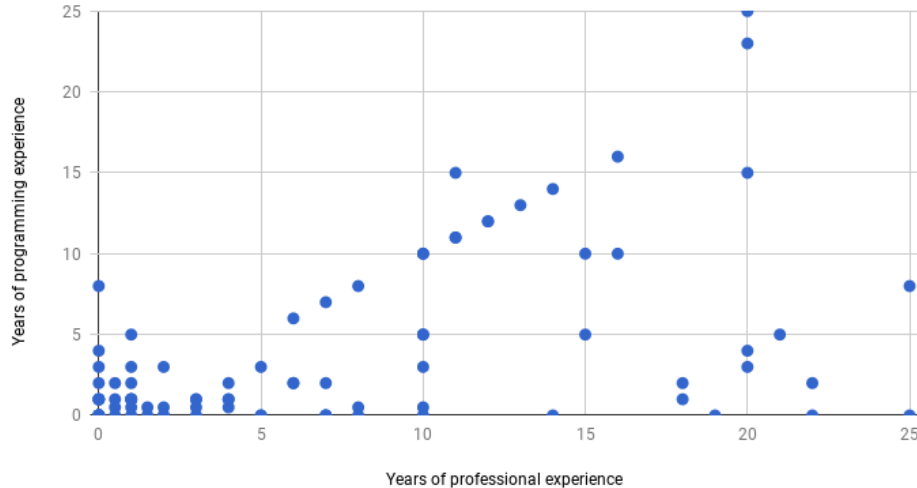
## Summarising Ratio Data:

*How do professional/programming experience compare?*



# Ratio (and interval) data

Professional vs programming experience



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

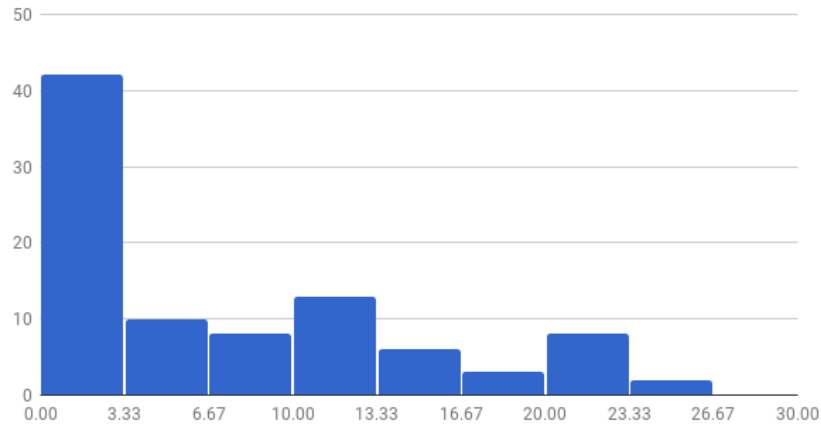
- **Standard deviation:** 
$$\sqrt{\text{variance}}$$

# Creating a Scatterplot

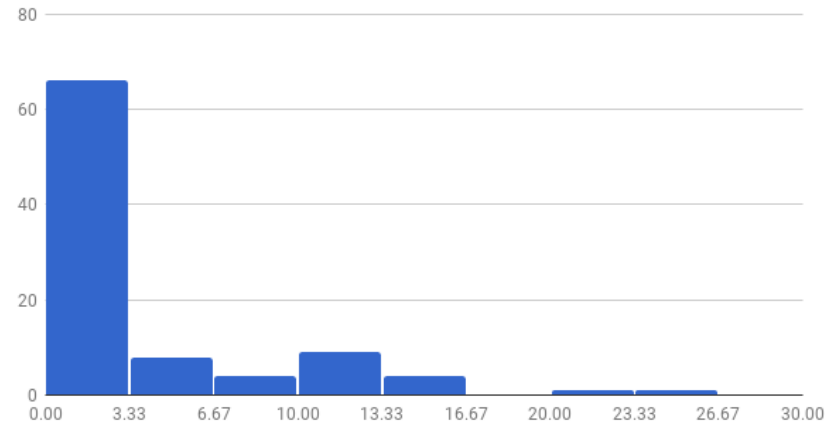
- Plots relationship between two different variables
- Display, e.g., professional experience on x-axis vs. programming experience on y-axis for each respondent
- In Google Sheets
  - Select data range (e.g., D1:En)
  - Click “insert chart” icon 
  - Select chart type “Scatter Chart”

# Binned histograms for experience

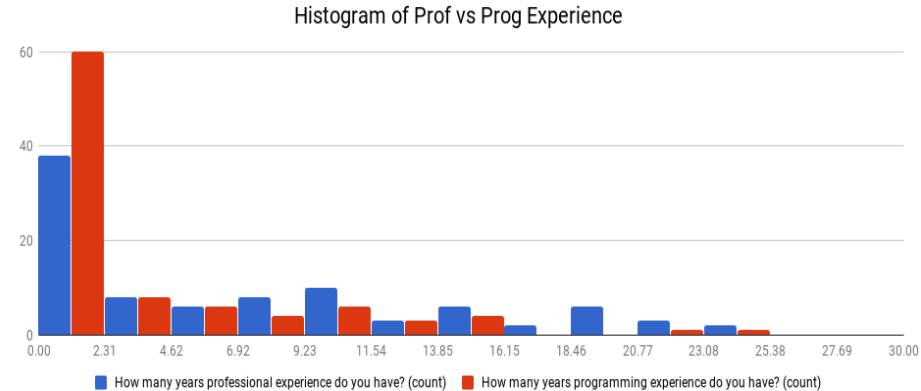
Histogram of professional experience



Histogram of programming experience



# Comparison with scatterplot and histogram overlays



## Pivot Tables:

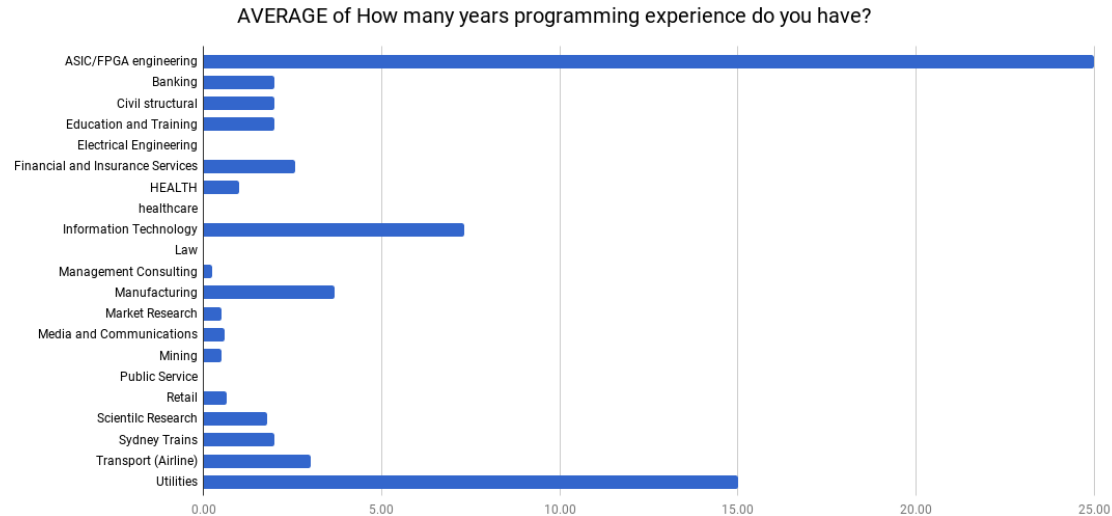
*How does programming experience differ across industries?*

# 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:En)
  - 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
Scientific Research	1.78
Sydney Trains	2.00
Transport (Airline)	3.00
Utilities	15.00
<b>Grand Total</b>	<b>3.57</b>





# Complex Counting:

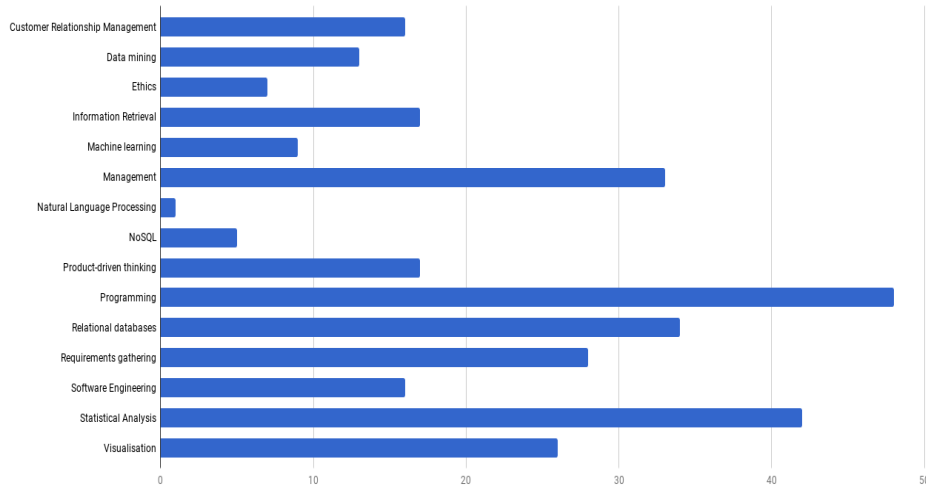
*What skills do we know? What would we like to learn?*

# How create a histogram of skills?

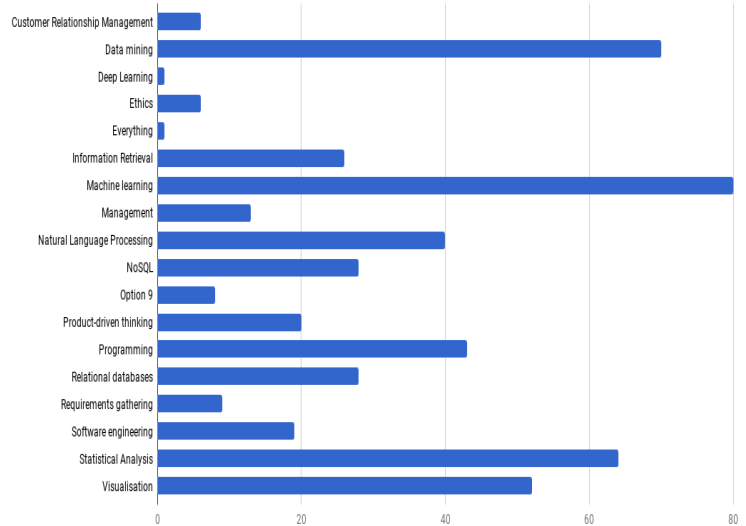
- Multiple values in cells within the skills column, e.g.:  
“Software engineering, Requirements gathering, Product-driven thinking”
- Need to split possible values:  
`=sort(unique(transpose(split(join(", ", N2:Nn), ", ", False))))`
- Then count:  
`=countif(N$2:N$n, concat(concat("*", T1), "*"))`
- Could use similar to get word counts
- Better to use programming language (clarity, reusability, etc)

# Histograms of current and desired skills (as of 2016...)

Current Skills



Desired Skills



# Review

# Participation

## Objective

Ensure everybody is keeping up.

## Requirements

- Submit code at end of each exercise
- Complete Grok (Python+SQL) by Week 6

## Marked:

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

## Marking

- 10% of overall mark

***TODO until Friday next week:***

***Export full workbook as PDF and upload to Canvas***

# 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 1-4 + SQL modules 1-2
- 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

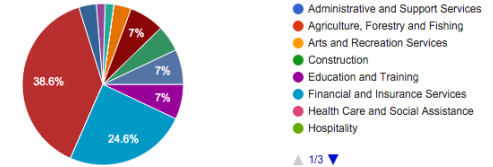
57 responses

SUMMARY

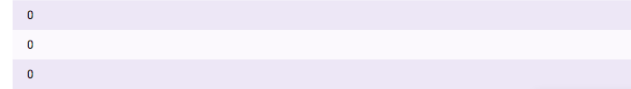
INDIVIDUAL

Accepting responses

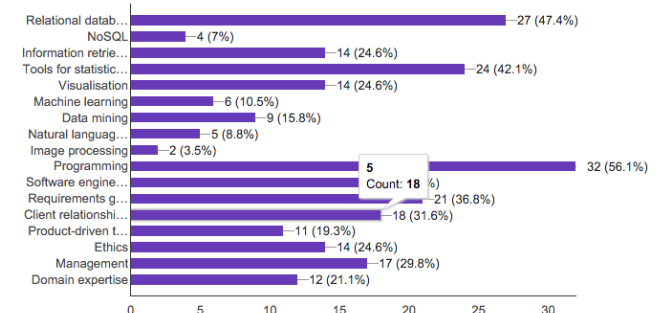
What main industry have you worked in? (57 responses)



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



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 tendency (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

# 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 1-4
- Grok SQL modules 1-4
- Explore and select project data

# Project Stage 1

# 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)

## Marking

- 13% 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: Problem & Exploratory analysis
- W5: Clean and prepare data
- 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