

Week 2 (COMP5310, 2020s2)

=====

Main Goals:

- Differences between **types of data** and applicable measures of dispersion and central tendency
- Idea and 1st Experience with **Data Cleansing**
- **Data Exploration** (using spreadsheet here)

Note: This is less meant to be a week about "how to use spreadsheets" than about data cleansing and exploration.

Because we cannot use Python yet (as too early) we are going for Google Spreadsheets instead.

Exercise 1: Data Import into Google Spreadsheets and Data Cleansing

Uses real data from last week's survey

- 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/aezH7m>
- Load survey data
 - File > Import

Click on Upload

- Select and load: Survey COMP5310 2019s2 (Responses) - Form Responses 1.csv

Some observations:

- Ration Data for Years experiences:
 - not all as numbers, but some free text (Ten == 10?)
 - how to handle '<1 year' or less than 1 year?
 - 'kind of 1' ?
- Nominal Data with Other entry:
 - Some free text entries ('Magic!')
 - Some use different names:
 - "Public Transport" and "Transport"
 - "Property" vs. "Property and Construction"
 - special characters in free-text answers just as group names
- 'None' instead of empty cell

- Multi-valued data for Skills etc: semicolon-separated text string


Exercise 2: 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

Exercise 3: Summarising Nominal Data (Known and Future Industries)

Create histograms of known and future industries

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”

Discuss:

- What do we need to do to make these comparable?
- What is the mode?

Exercise 4: Summarising Ordinal Data (Importance of various skills for data analytics)

Important to use Histogram Chart here, NOT Column or Bar Chart


=> aggregations other than count not defined on ordinal data

also careful not to use first column as label data

=> check whether data management shows up as one series

Create a histogram diagram of area importance ratings

In Google Sheets:

- Select data range (e.g., G1:Ln*) *n is the total number of responses
- 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

Discuss:


- Which area gets the most high rankings?
- Are there interesting differences between areas?
- Do medians differ? Ranges?

Exercise 5: Summarising Ratio/Interval Data (years of prof. vs. prog., experience)

Create a scatter plot of professional vs. programming experience

- 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*) *n is the total number of responses
- Click “insert chart” icon 
- Select chart type “Scatter Chart”

Discuss/explore:

- Can you see any relation between the of professional and programming experience?

Exercise 6: Creating a Pivot Table

Create a table of average programming experience by industry

In Google Sheets:

- Select data range (e.g., C1:En*) *n is the total number of responses
- Go to Data > Pivot Table (should insert a new sheet)
- Select industry under row
- Select professional experience under value
- Summarise by average

Discuss:

- What other statistics can we calculate?
- What other variable combinations could we explore

Exercise 7: Complex Counting

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

- 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), "*"))`