University of
South Australia

Feature Engineering & Data Exploration

Dr Srećko Joksimović

Feature Engineering

Applying **domain knowledge** to create new features that allow machine learning algorithms to **improve performance (or work at all)**

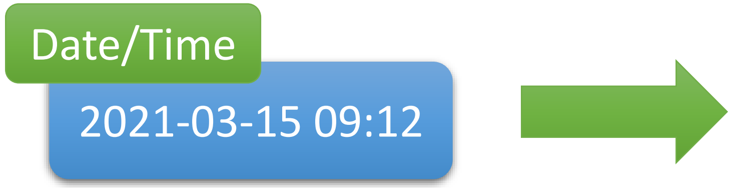
Why so important?

- More important than algorithm selection

- Algorithms have no ***deeper*** understanding of data

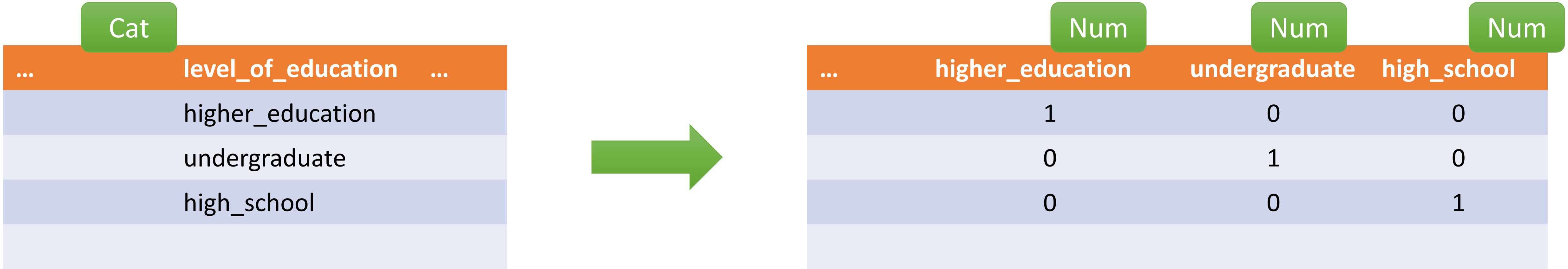
- Intuition is good, but risky – remember to evaluate
  - For example, predicting health status or wellbeing:
    - Your weight in kilograms to the first decimal place (e.g., 86.4)
    - Your height in centimetres (e.g., 172).
    - Why not BMI, instead?

Example: Date/Time Fields



	Num	Cat	Num	Num	Num
…	**Year**	**Month**	**Day**	**Hour**	**Minute**
	2021	Mar	15	9	12

Example: Dummy variables



	Cat	
…	level_of_education	…
	higher_education	
	undergraduate	
	high_school	

	Num	Num	Num
…	higher_education	undergraduate	high_school
	1	0	0
	0	1	0
	0	0	1

Example: Text analysis

Text analysis, also known as **text** mining, is a machine learning technique used to automatically extract value from **text** data. With the help of natural language processing (NLP), **text** analysis tools are able to understand, analyze, and extract insights from your unstructured data.

TEXT appears 4 times

Example: Text analysis

Text analysis, also known as **text** mining, is a machine learning technique used to automatically extract value from **text** data. With the help of natural language processing (NLP), **text** analysis tools are able to understand, analyze, and extract insights from your unstructured data.

…	text	learn	process
	4	1	1

Example: Binning

…	value	bin
	2	Low
	45	Mid
	3	Low
	85	High
	28	Low

Example: Computed from Existing Features

debt	income
10,134	100,000
85,234	134,000
8,112	21,500
0	45,900
17,534	52,000

→

debt_income_ratio
0.10
0.64
0.38
0
0.34

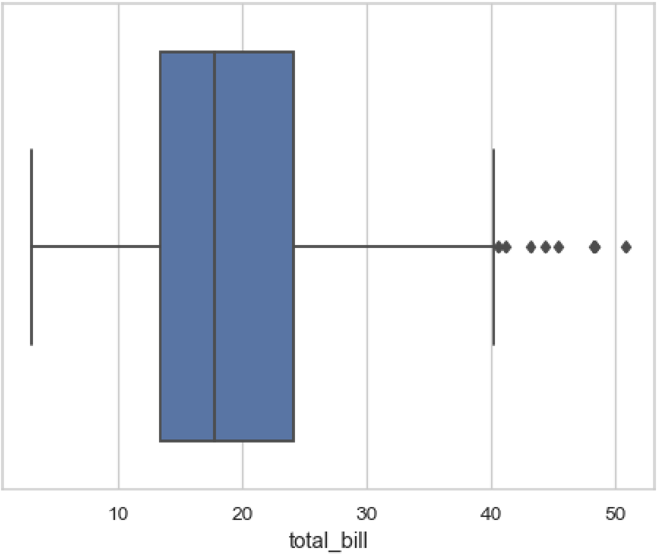
Outliers

What are outliers?

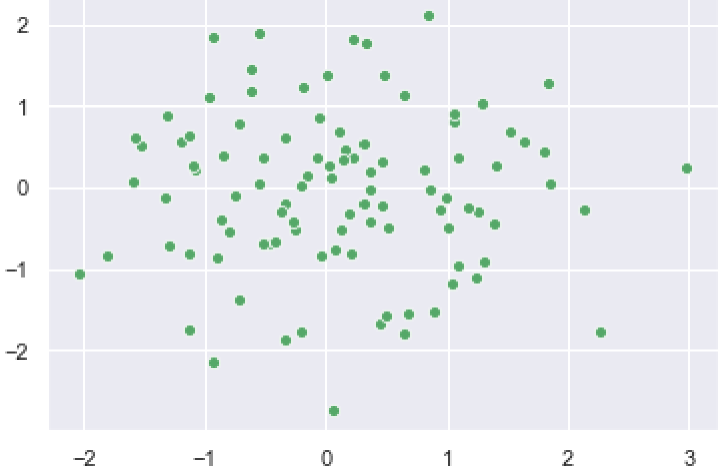
An **outlier** is a data point that’s significantly different from the remaining data

Source: https://heartbeat.fritz.ai/hands-on-with-feature-engineering-techniques-dealing-with-outliers-fcc9f57cb63b

Detecting outliers



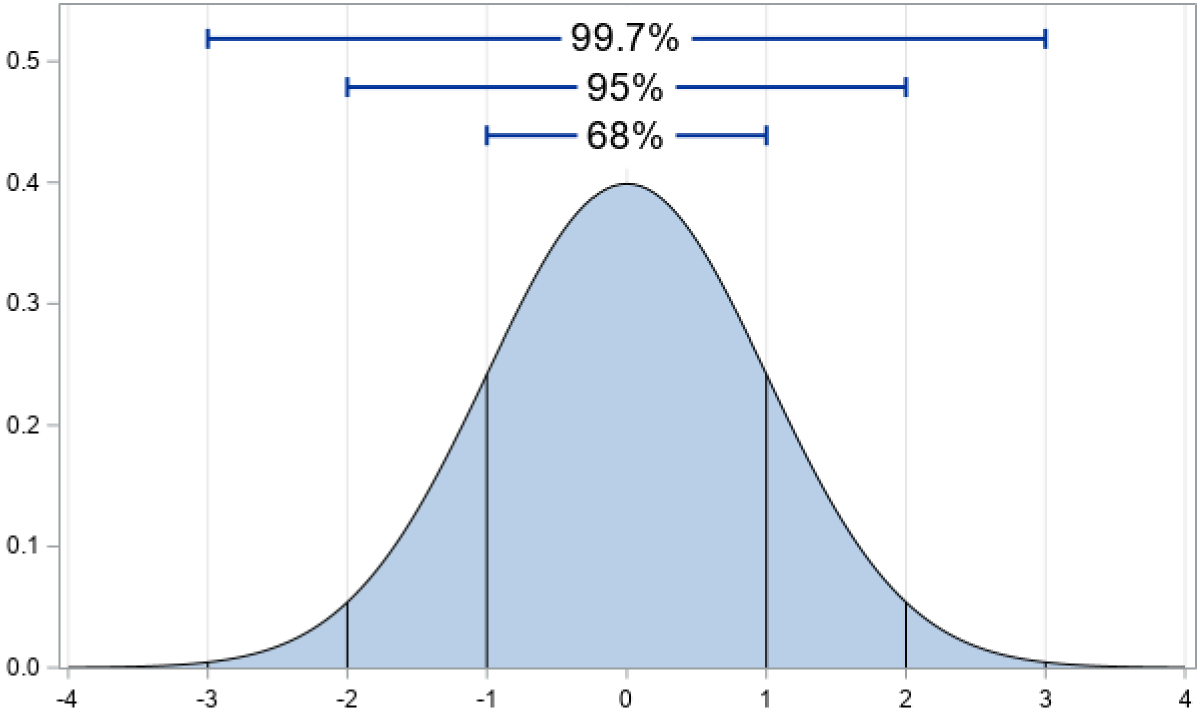
Using visualization plots like **boxplot** and **scatterplot**



Source: https://heartbeat.fritz.ai/hands-on-with-feature-engineering-techniques-dealing-with-outliers-fcc9f57cb63b

Detecting outliers

Using a normal distribution (mean and std)



Source: https://heartbeat.fritz.ai/hands-on-with-feature-engineering-techniques-dealing-with-outliers-fcc9f57cb63b

Handling outliers

- **Trimming:** Simply removing the outliers from our dataset.
- **Imputing:** We treat outliers as missing data, and we apply missing data imputation techniques.
- **Discretization:** We place outliers in edge bins with higher or lower values of the distribution.
- **Censoring:** Capping the variable distribution at the maximum and minimum values.

Source: https://heartbeat.fritz.ai/hands-on-with-feature-engineering-techniques-dealing-with-outliers-fcc9f57cb63b

Data Exploration

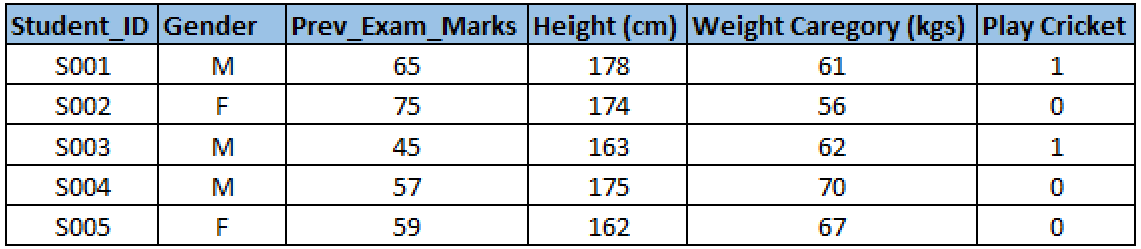
Data Exploration

- Variable identification
- Univariate analysis
- Bi-variate analysis

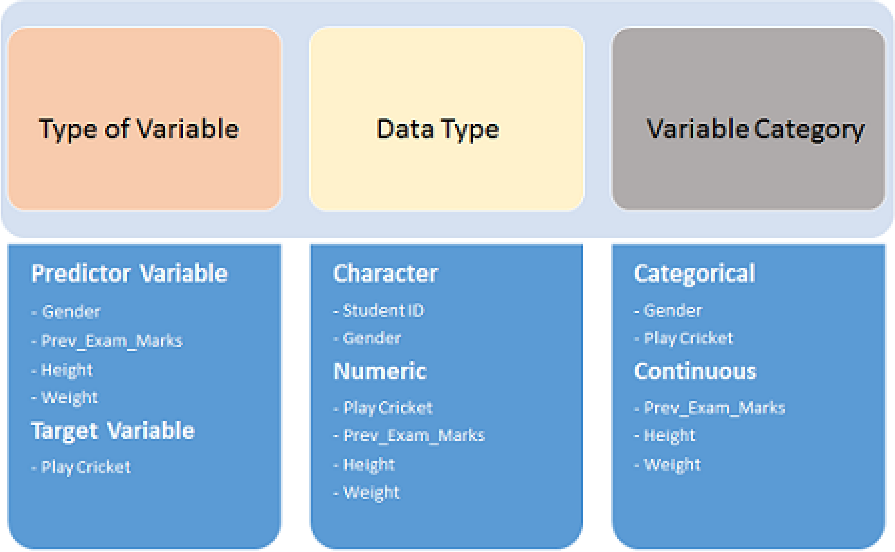
Variable identification

- Identify **Predictor** (Input) and **Target** (output) variables.

- Identify the data type and category of the variables.



Student_ID	Gender	Prev_Exam_Marks	Height (cm)	Weight Caregory (kgs)	Play Cricket
S001	M	65	178	61	1
S002	F	75	174	56	0
S003	M	45	163	62	1
S004	M	57	175	70	0
S005	F	59	162	67	0



Type of Variable

Predictor Variable
- Gender
- Prev_Exam_Marks
- Height
- Weight

Target Variable
- Play Cricket

Data Type

Character
- Student ID
- Gender

Numeric
- Play Cricket
- Prev_Exam_Marks
- Height
- Weight

Variable Category

Categorical
- Gender
- Play Cricket

Continuous
- Prev_Exam_Marks
- Height
- Weight

Source: https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/

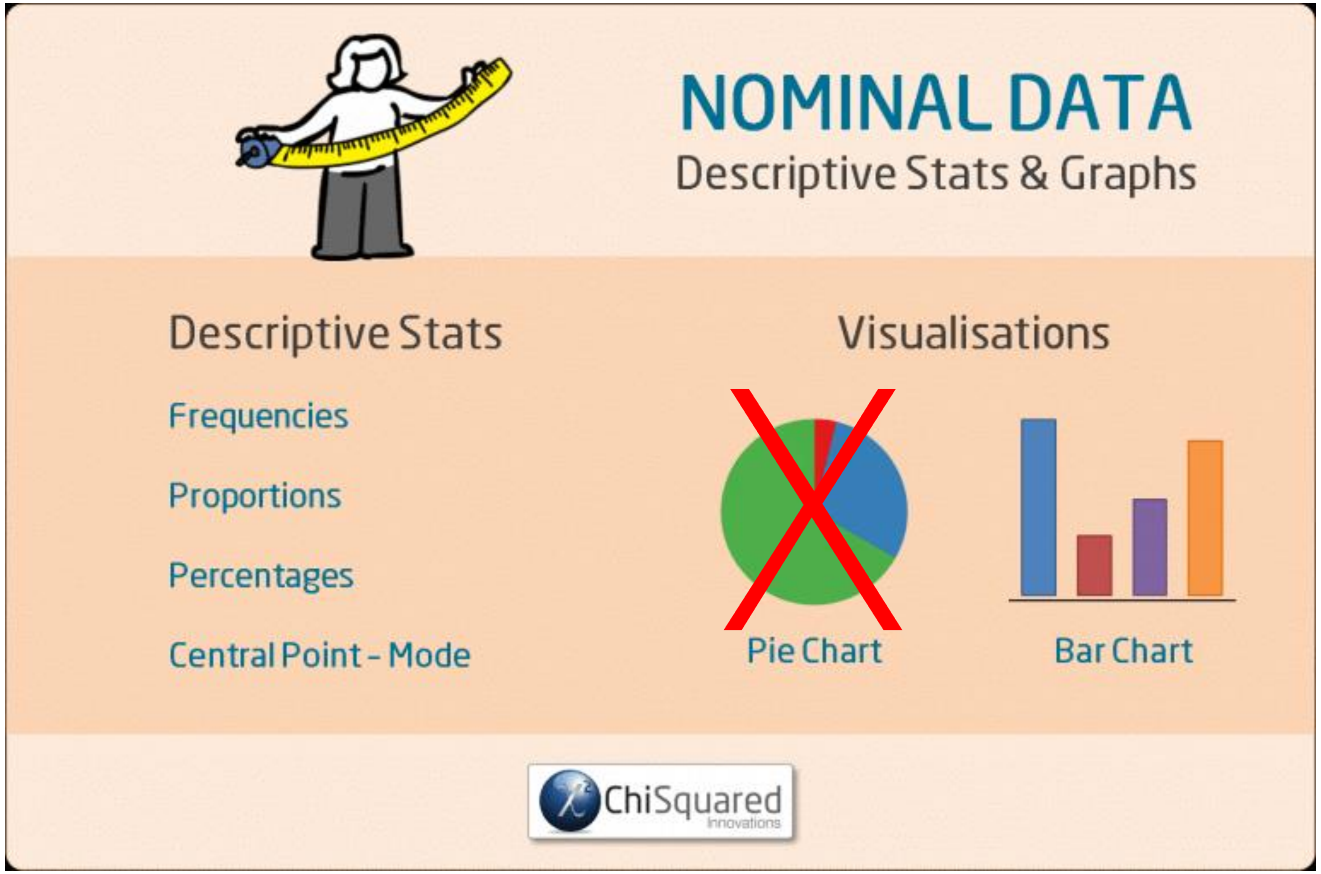
Univariate analysis

- Exploring variables one by one
- Available methods – depending on the variable category

Source: https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/

Nominal data

Descriptive stats
and visualisations



NOMINAL DATA

Descriptive Stats & Graphs

Descriptive Stats

- Frequencies
- Proportions
- Percentages
- Central Point - Mode

Visualisations

Pie Chart

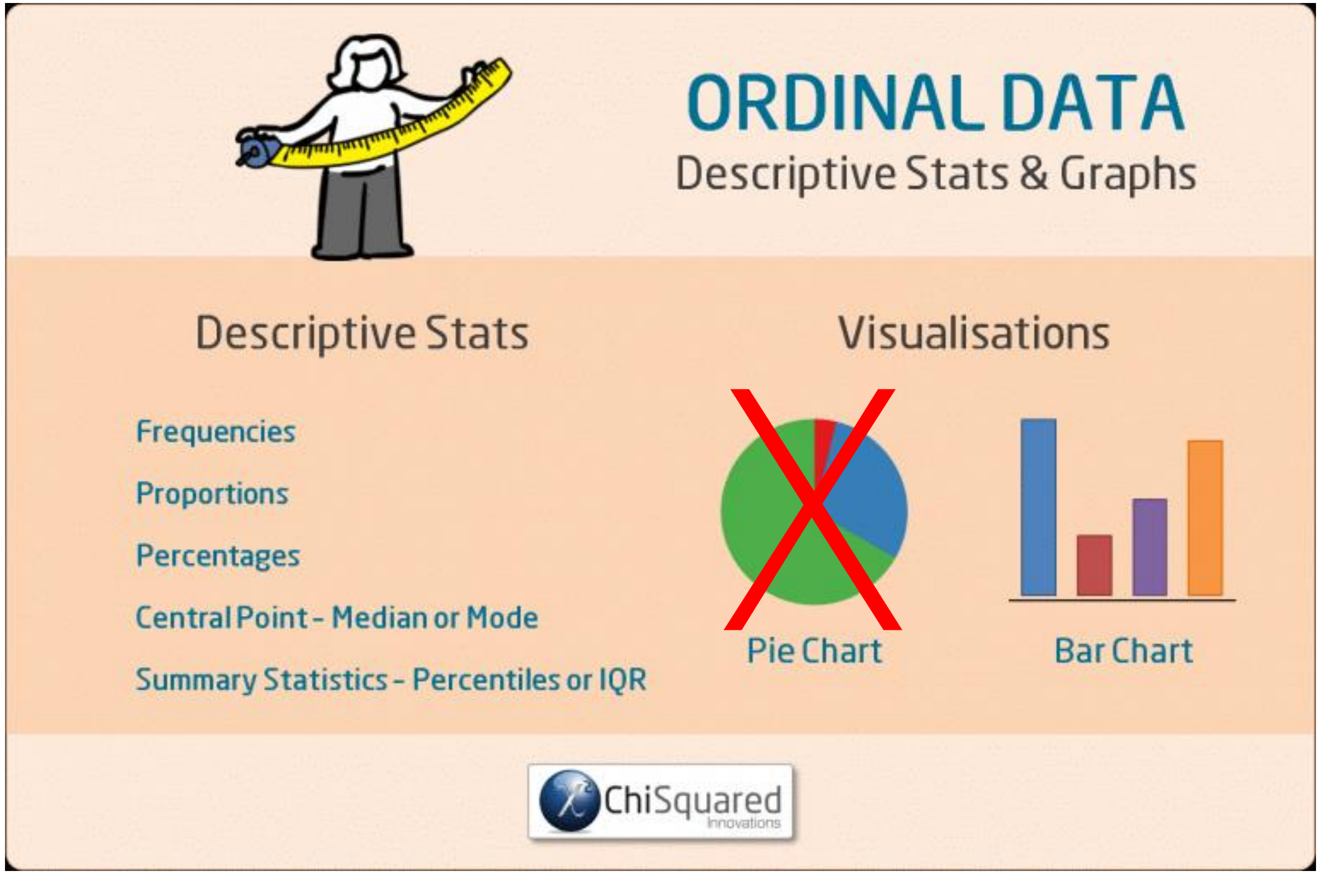
Bar Chart

ChiSquared Innovations

Source: https://www.chi2innovations.com/blog/discover-data-blog-series/nominal-data/

Ordinal data

Descriptive stats
and visualisations



ORDINAL DATA

Descriptive Stats & Graphs

Descriptive Stats

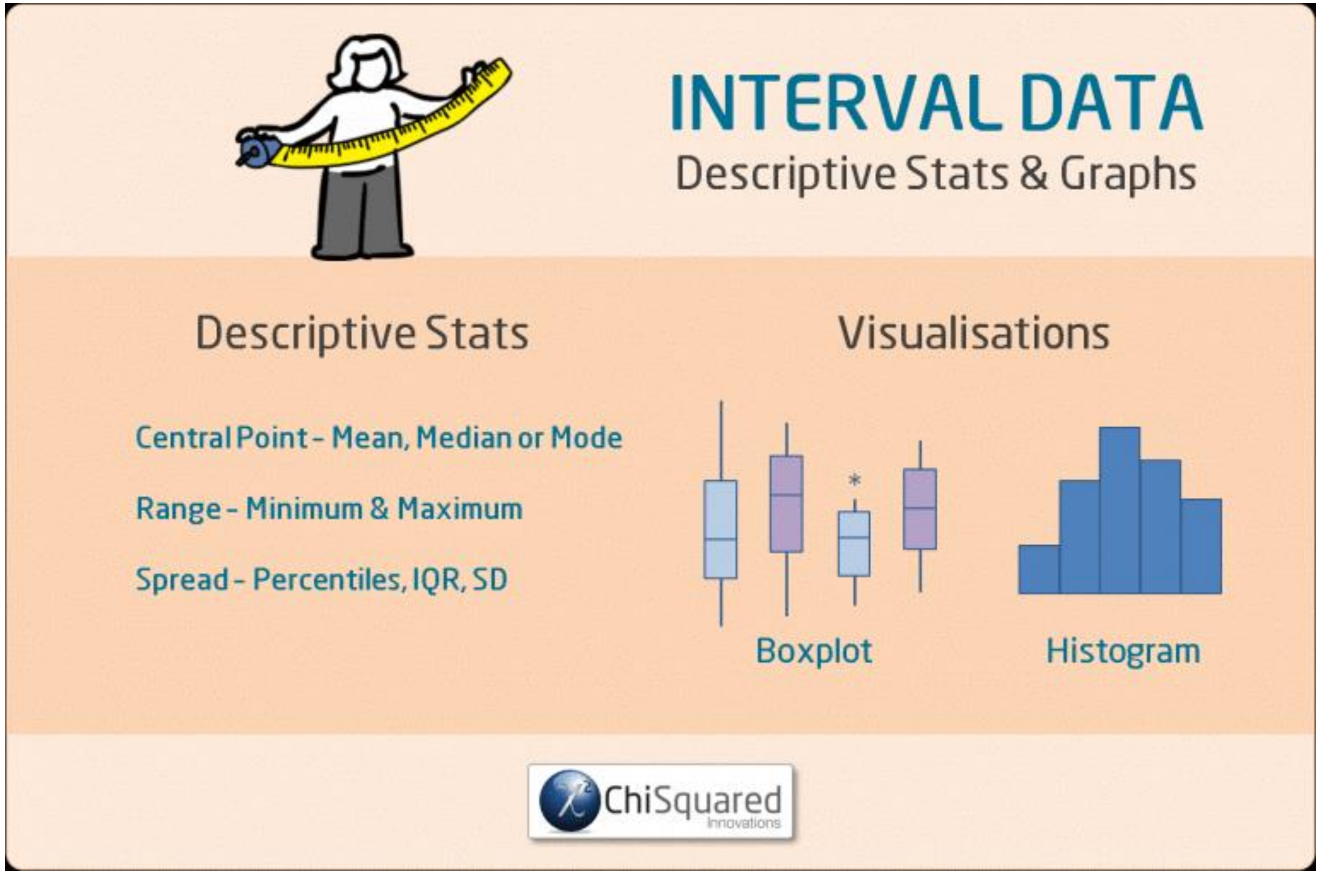
- Frequencies
- Proportions
- Percentages
- Central Point - Median or Mode
- Summary Statistics - Percentiles or IQR

Visualisations

Source: https://www.chi2innovations.com/blog/discover-data-blog-series/nominal-data/

Interval data

Descriptive stats
and visualisations



INTERVAL DATA

Descriptive Stats & Graphs

Descriptive Stats

Central Point - Mean, Median or Mode

Range - Minimum & Maximum

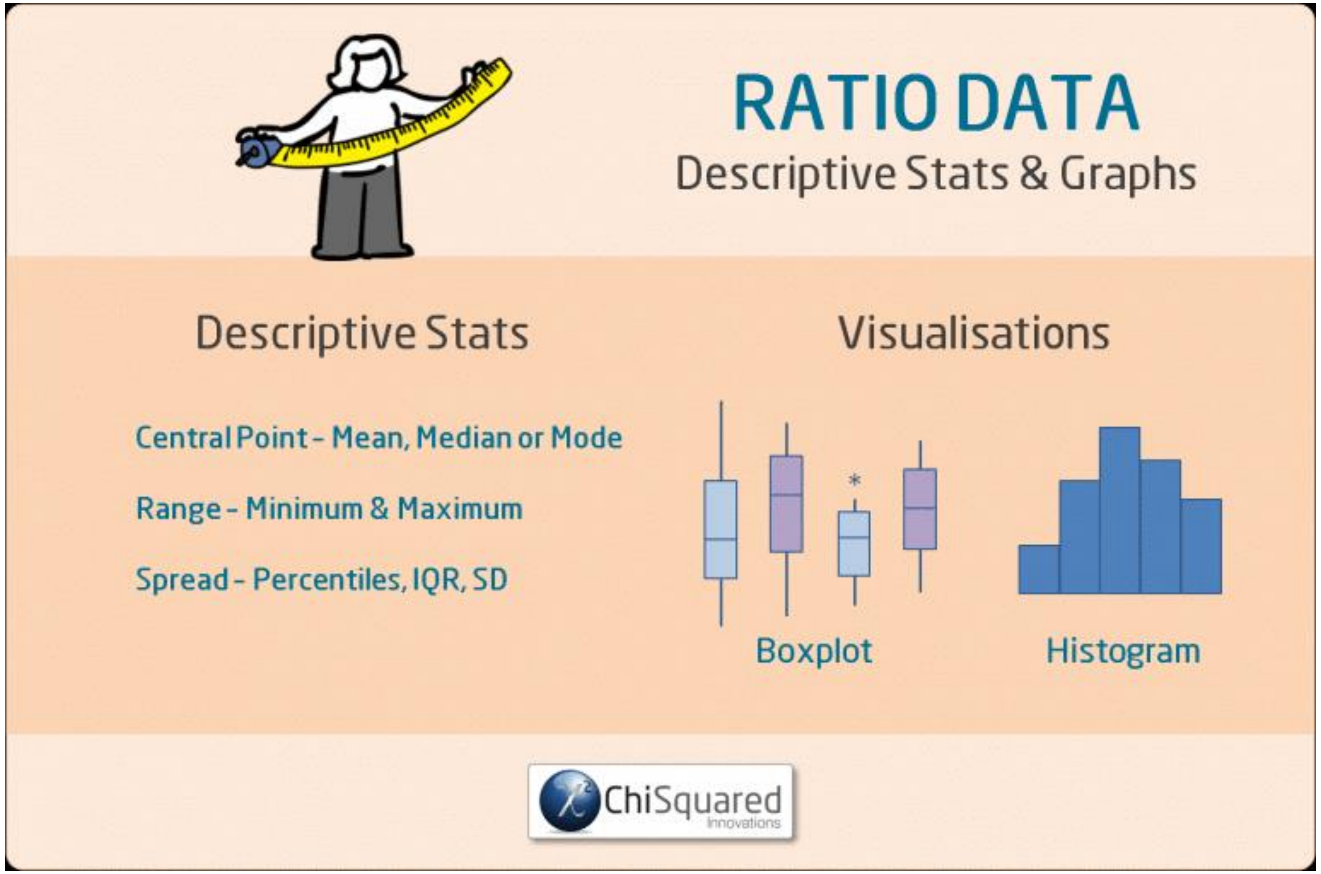
Spread - Percentiles, IQR, SD

Visualisations

Source: https://www.chi2innovations.com/blog/discover-data-blog-series/nominal-data/

Ratio data

Descriptive stats
and visualisations



RATIO DATA
Descriptive Stats & Graphs

Descriptive Stats

Central Point - Mean, Median or Mode

Range - Minimum & Maximum

Spread - Percentiles, IQR, SD

Visualisations

Source: https://www.chi2innovations.com/blog/discover-data-blog-series/nominal-data/

Bi-variate analysis

- Exploring the association between two (or more) variables

- Common combinations:
  - Categorical & categorical
  - Continuous & continuous
  - Categorical & continuous

Source: https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/

Two categorical variables

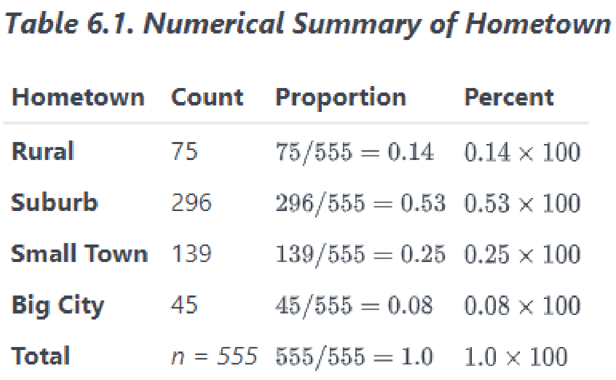
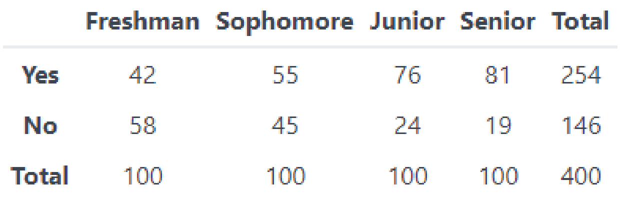


Table 6.1. Numerical Summary of Hometown Description

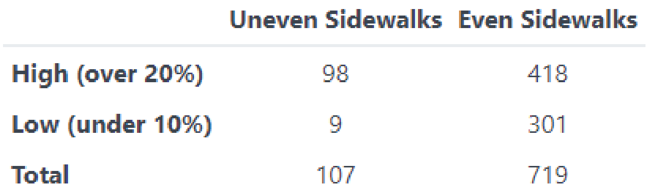
Hometown	Count	Proportion	Percent
Rural	75	$75/555 = 0.14$	0.14×100
Suburb	296	$296/555 = 0.53$	0.53×100
Small Town	139	$139/555 = 0.25$	0.25×100
Big City	45	$45/555 = 0.08$	0.08×100
Total	n = 555	$555/555 = 1.0$	1.0×100

Numerical summary



	Freshman	Sophomore	Junior	Senior	Total
Yes	42	55	76	81	254
No	58	45	24	19	146
Total	100	100	100	100	400

Contingency table



	Uneven Sidewalks	Even Sidewalks
High (over 20%)	98	418
Low (under 10%)	9	301
Total	107	719

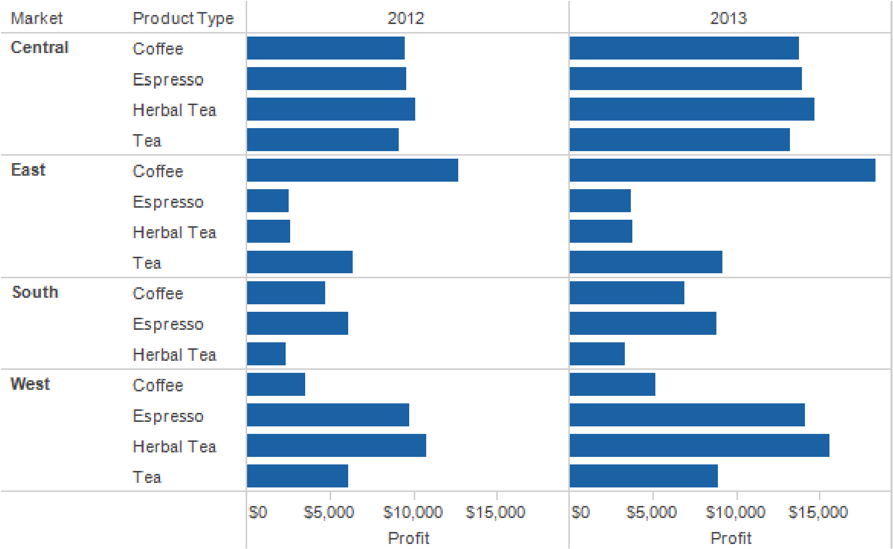
2x2 table

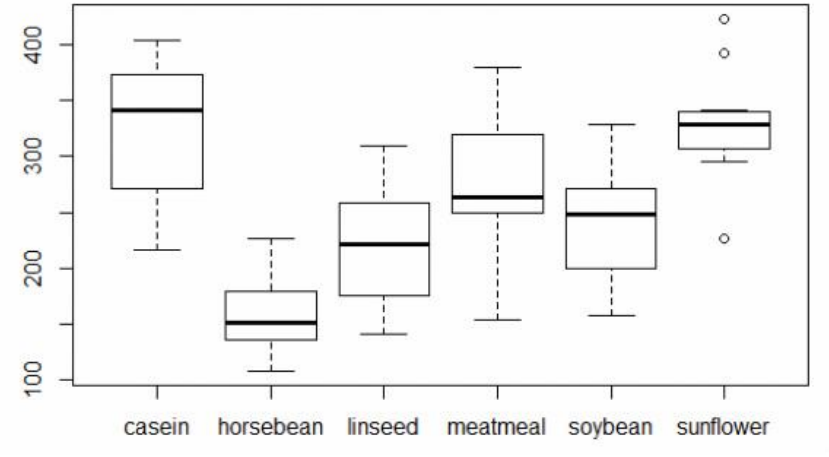
Statistical analysis:

- » **Chi-square test**
- » **Crammer's V**

https://online.stat.psu.edu/stat100/lesson/6

Categorical and Continuous variable



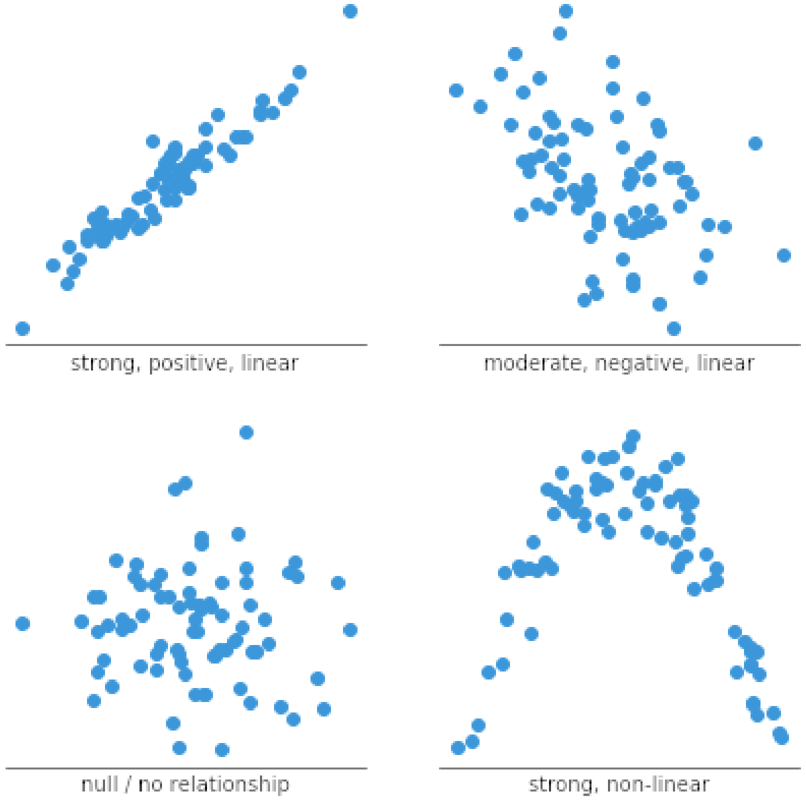


Statistical analysis:

- » **T-test/Z-test**
- » **ANOVA**

Source: https://eagereyes.org/basics/data-continuous-vs-categorical

Two continuous variables



Statistical analysis:

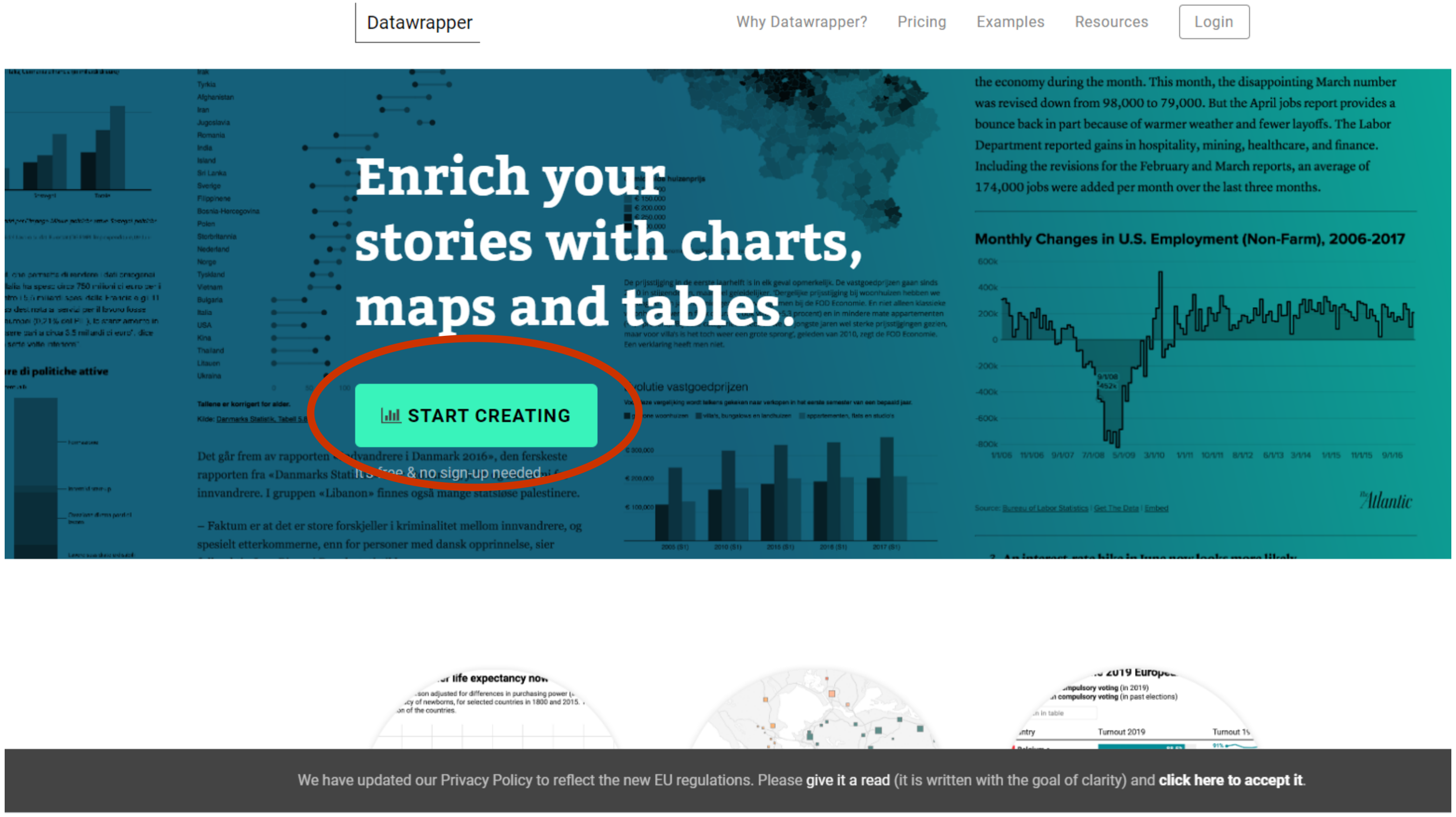
- **Correlation**
- **Regression**

Source: https://chartio.com/learn/charts/what-is-a-scatter-plot/

Hands-on

Explore visualisations (individually or in groups)

- Go to https://www.datawrapper.de/



Enrich your stories with charts, maps and tables.

It's free & no sign-up needed.

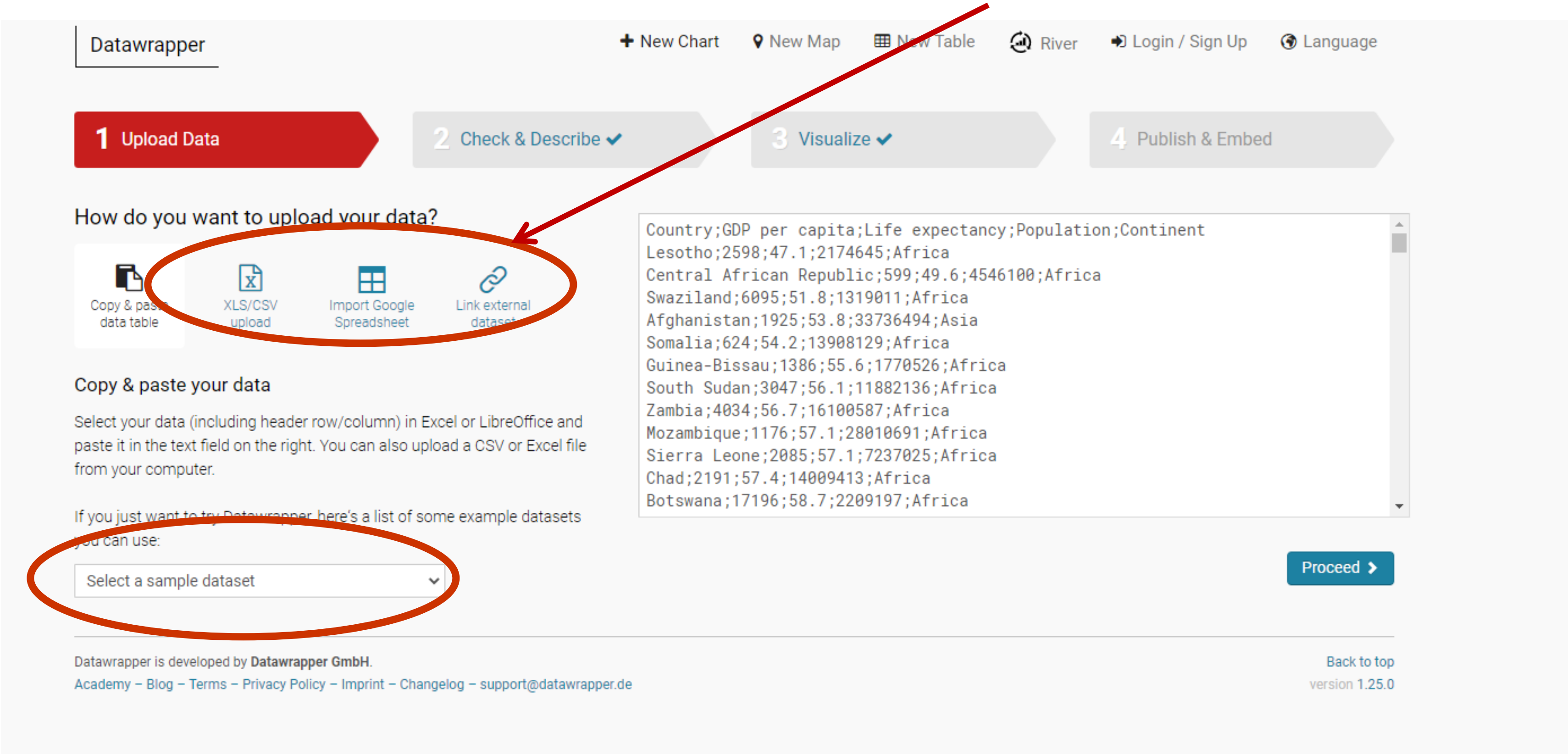
the economy during the month. This month, the disappointing March number was revised down from 98,000 to 79,000. But the April jobs report provides a bounce back in part because of warmer weather and fewer layoffs. The Labor Department reported gains in hospitality, mining, healthcare, and finance. Including the revisions for the February and March reports, an average of 174,000 jobs were added per month over the last three months.

Monthly Changes in U.S. Employment (Non-Farm), 2006-2017

We have updated our Privacy Policy to reflect the new EU regulations. Please **give it a read** (it is written with the goal of clarity) and **click here to accept it**.

Upload data or select a sample dataset

Choose one of the datasets provided in **Slides and Other Resources** folder



How do you want to upload your data?

Copy & paste your data

Select your data (including header row/column) in Excel or LibreOffice and paste it in the text field on the right. You can also upload a CSV or Excel file from your computer.

If you just want to try Datawrapper here's a list of some example datasets you can use:

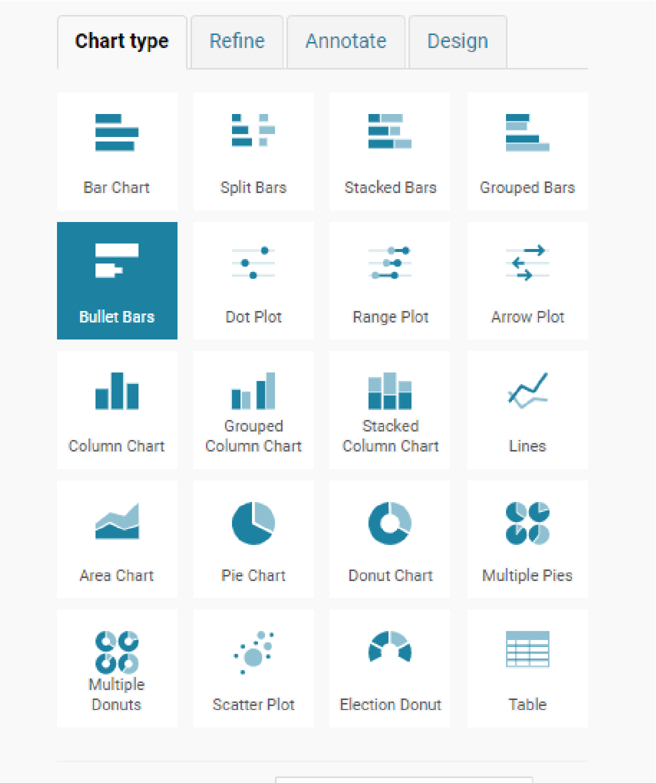
```
Country;GDP per capita;Life expectancy;Population;Continent
Lesotho;2598;47.1;2174645;Africa
Central African Republic;599;49.6;4546100;Africa
Swaziland;6095;51.8;1319011;Africa
Afghanistan;1925;53.8;33736494;Asia
Somalia;624;54.2;13908129;Africa
Guinea-Bissau;1386;55.6;1770526;Africa
South Sudan;3047;56.1;11882136;Africa
Zambia;4034;56.7;16100587;Africa
Mozambique;1176;57.1;28010691;Africa
Sierra Leone;2085;57.1;7237025;Africa
Chad;2191;57.4;14009413;Africa
Botswana;17196;58.7;2209197;Africa
```

Datawrapper is developed by **Datawrapper GmbH**.

Academy – Blog – Terms – Privacy Policy – Imprint – Changelog – support@datawrapper.de

version 1.25.0

Explore



- Choose at least **2 Chart types** to explore selected dataset.
- Explain why you selected these charts – what kind of information they provide?
- Discuss (compare and contrast) insights you obtained with each of the selected chart types. What those charts tell you about your dataset?

University of
South Australia

INFS 5100 Predictive Analytics

Q&A