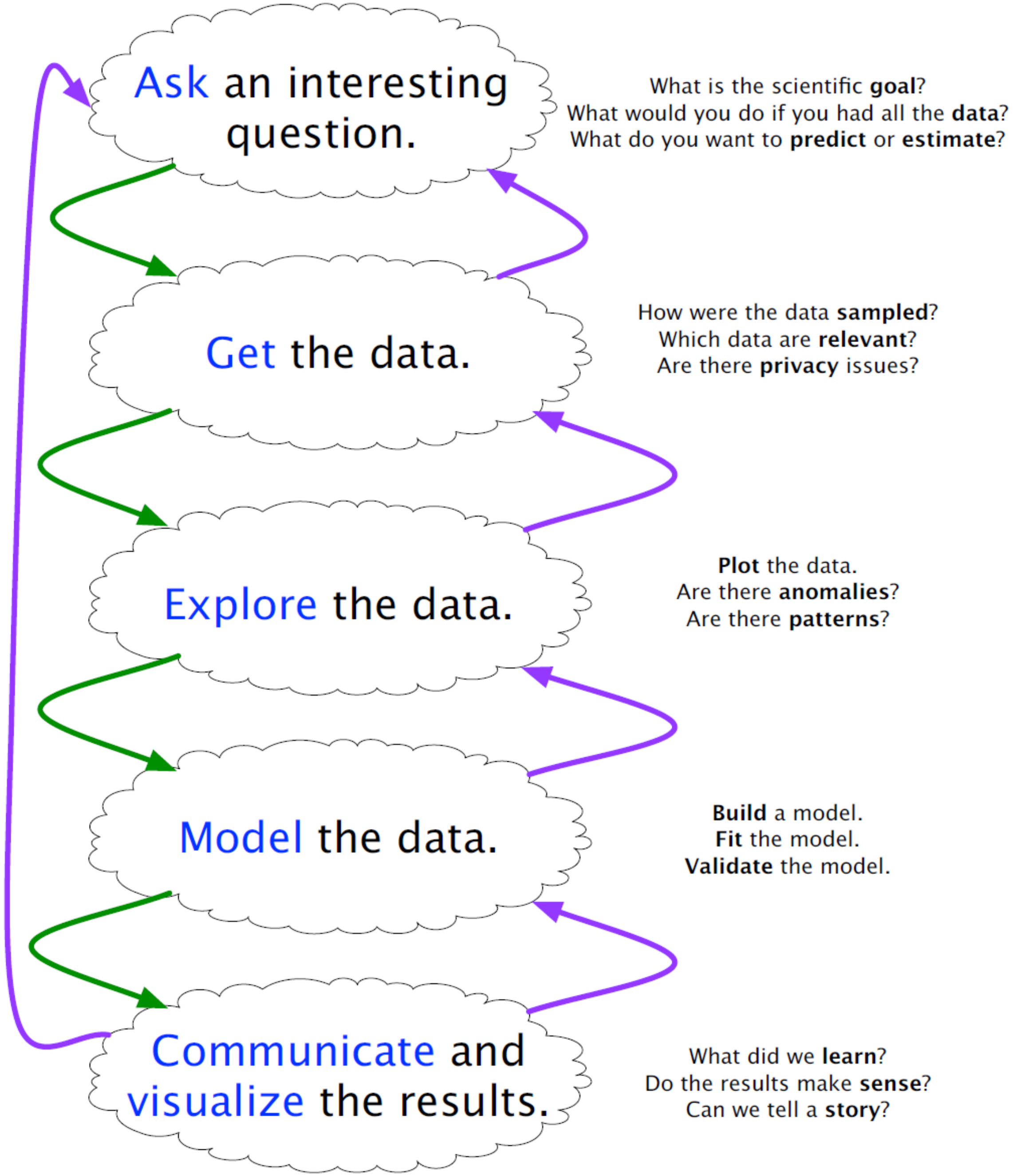
# Inf2 – Foundations of Data Science
Data

THE UNIVERSITY of EDINBURGH
informatics

FOUNDATIONS
OF
DATA
SCIENCE



Ask an interesting question.

What is the scientific **goal**?
What would you do if you had all the **data**?
What do you want to **predict** or **estimate**?

Get the data.

How were the data **sampled**?
Which data are **relevant**?
Are there **privacy** issues?

Explore the data.

**Plot** the data.
Are there **anomalies**?
Are there **patterns**?

Model the data.

**Build** a model.
**Fit** the model.
**Validate** the model.

Communicate and visualize the results.

What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?

Image taken from Harvard CS109

# How to ask interesting questions?

(From last year’s final project)

You can choose any dataset you like, provided that it has the following characteristics:

- Multivariate (at least 3 variables)
- Available for public download.

Choose a few representative samples of instances in your data and look at them closely to get some ideas for potential questions.

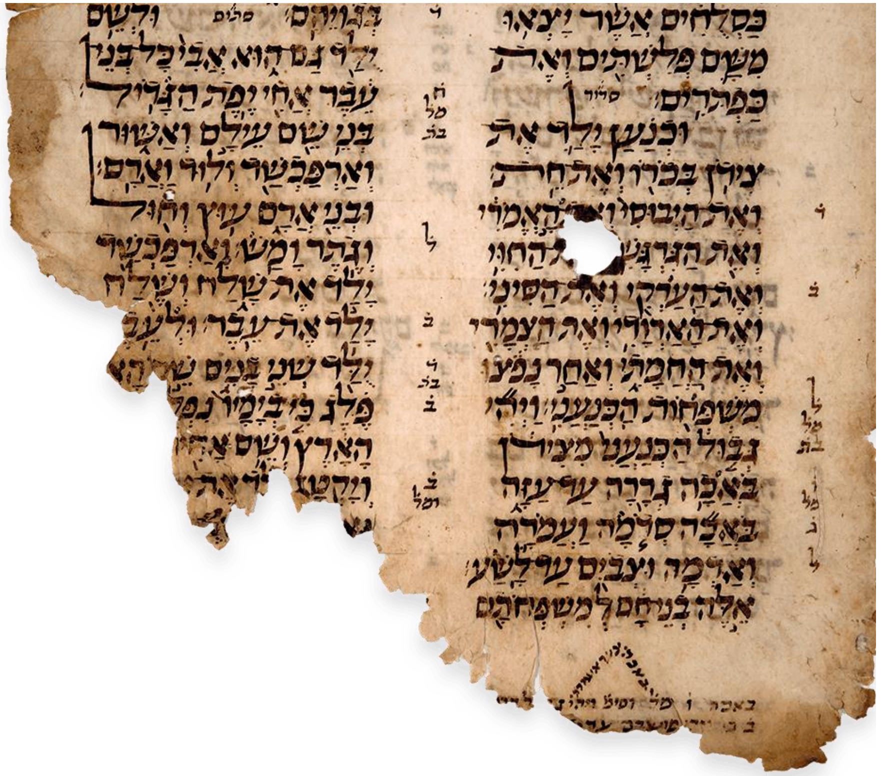
- QUESTIONS: What are the questions you wanted to explore? Why are they interesting to you?
- DATASET: Describe the dataset you use; Explain why it is appropriate for answering these questions.

# Example

- In this project I will be discussing the factors affecting the popularity of Munros.
- The main question I will be exploring is what factors make a Munro popular?
- What factors have the most influence in the popularity of Munros?
  How do the natural relief and cities' location affect the popularity?
- I have used data provided by
- https://www.walkhighlands.co.uk/munros/
- http://www.hills-database.co.uk/downloads

# What is data and where does it come from?

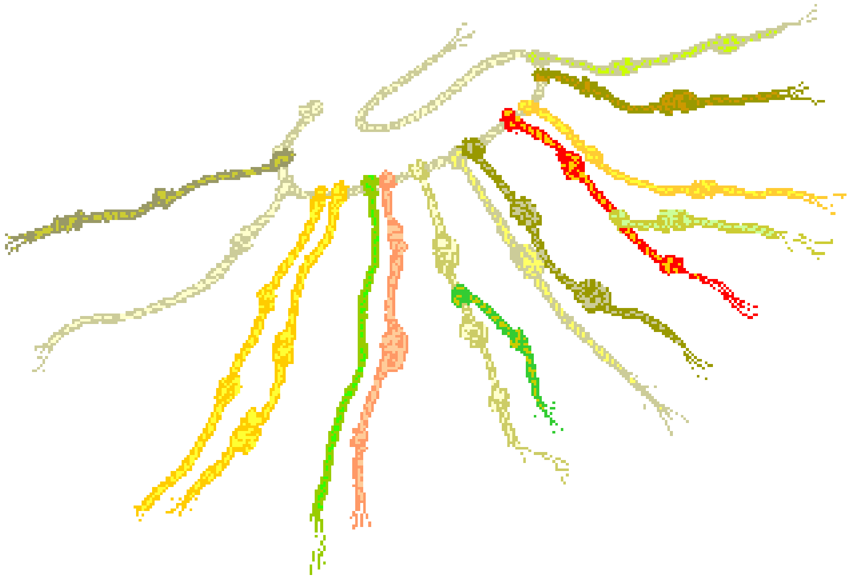
# Ancient methods



כסלחים אשר יצאו
משם פלשתים ואת
כפתרים
וכנען ילד את
צידן בכרו ואת חת
ואת היבוסי ואת האמרי
ואת הגרגשי ואת החוי
ואת הערקי ואת הסיני
ואת הארודי ואת הצמרי
ואת החמתי ואחר נפצו
משפחות הכנעני ויהי
גבול הכנעני מצידן
באכה גררה עד עזה
באכה סדמה ועמרה
ואדמה וצבים עד לשע
אלה בני חם למשפחתם

בגויהם ולשם
ילד גם הוא אבי כל בני
עבר אחי יפת הגדול
בני שם עילם ואשור
וארפכשד ולוד וארם
ובני ארם עוץ וחול
וגתר ומש וארפכשד
ילד את שלח ושלח
ילד את עבר ולעבר
ילד שני בנים שם הא
פלג כי בימיו נפ
הארץ ושם אחיו
ויקטן ילד את

https://www.scribesofthecairogeniza.org/

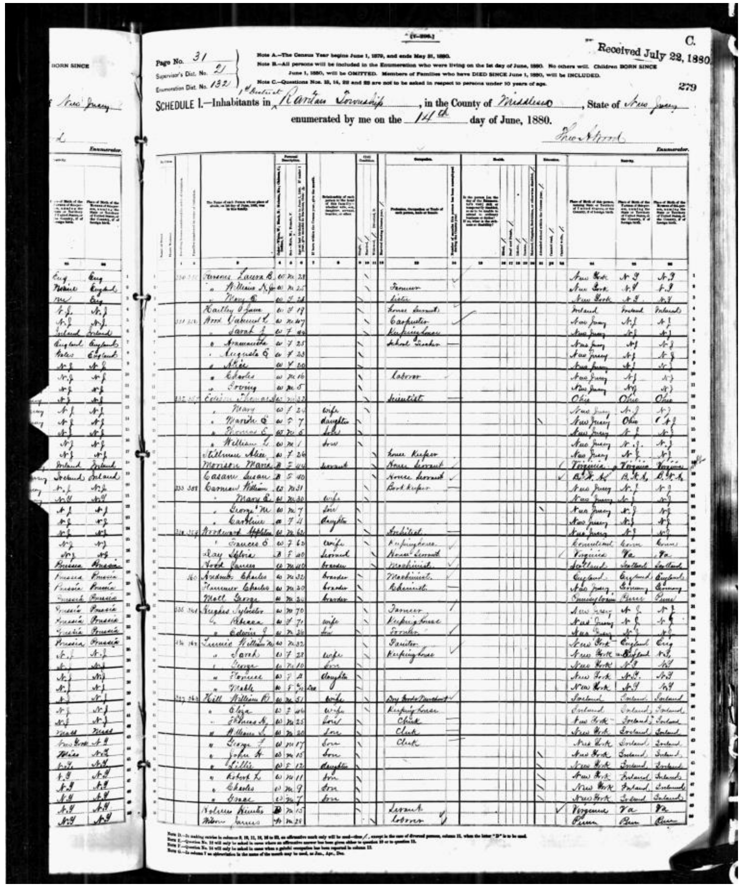


Quipu

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

# Late 19th century: US census

U.S. DEPARTMENT OF COMMERCE
BUREAU OF THE CENSUS



[V-296.]

Received July 22, 1880

C.

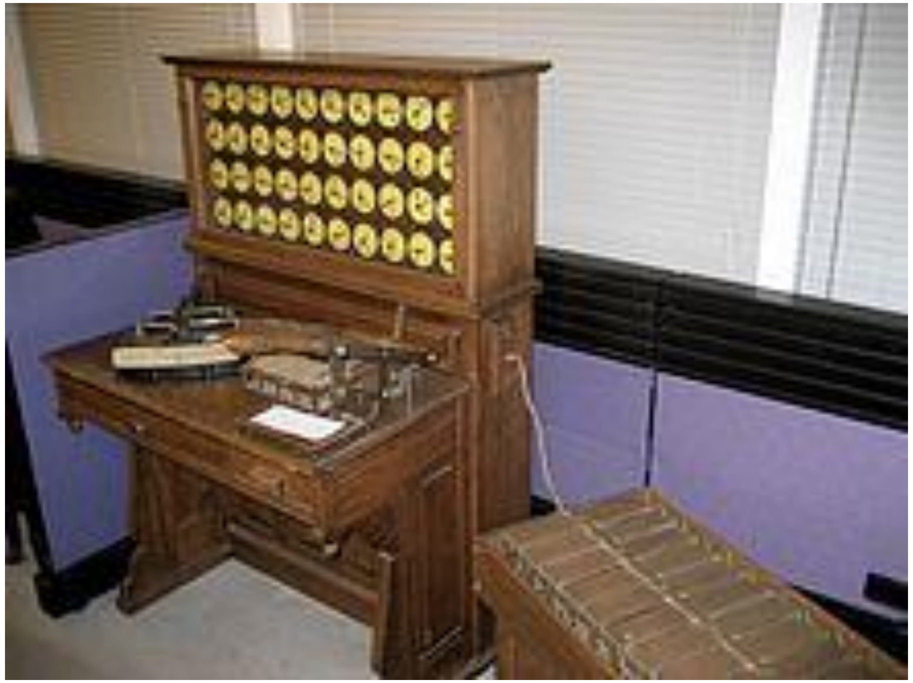
Page No. 31
Supervisor's Dist. No. 21
Enumeration Dist. No. 132

Note A.—The Census Year begins June 1, 1879, and ends May 31, 1880.
Note B.—All persons will be included in the Enumeration who were living on the 1st day of June, 1880. No others will. Children BORN SINCE June 1, 1880, will be OMITTED. Members of Families who have DIED SINCE June 1, 1880, will be INCLUDED.
Note C.—Questions Nos. 13, 14, 22 and 23 are not to be asked in respect to persons under 10 years of age.

279

SCHEDULE 1.—Inhabitants in 1st District Raritan Township, in the County of Middlesex, State of New Jersey, enumerated by me on the 14th day of June, 1880.

Theo. N. Wood
Enumerator.

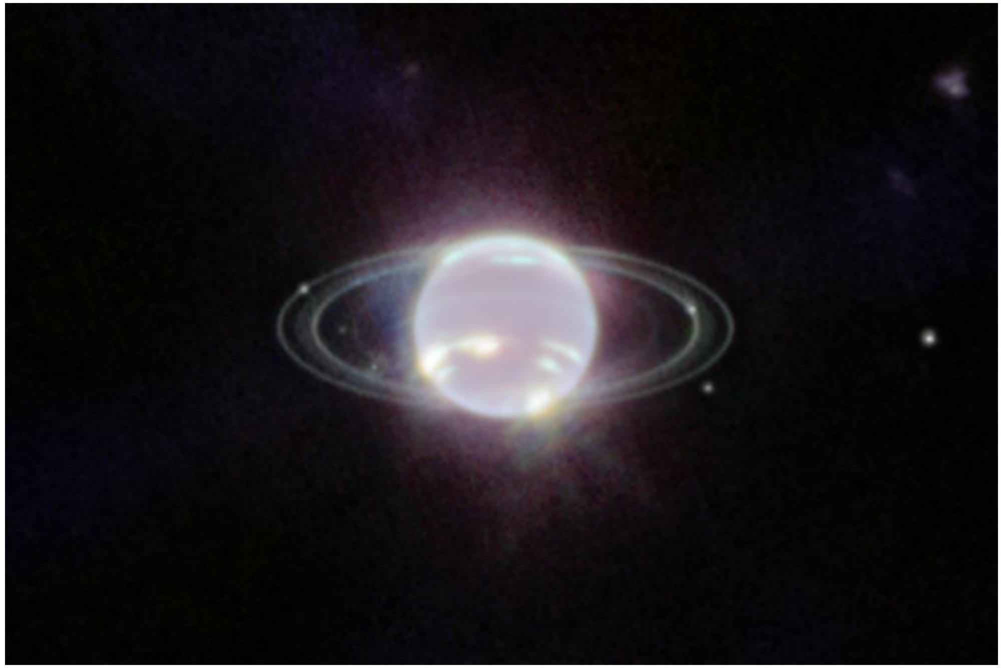


Hollerith 1890 tabulating machine

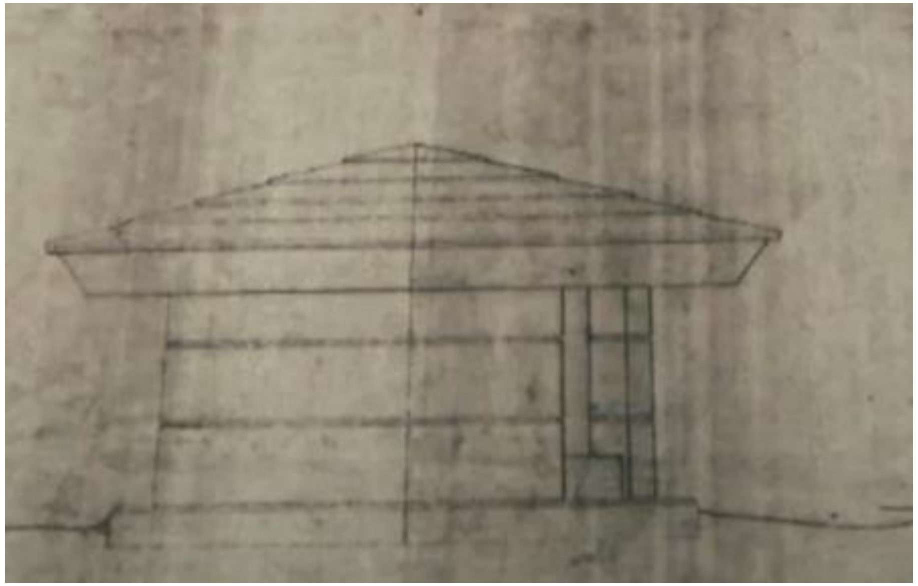
# 20th century: Storage devices

Fritz Pfleumer, with his magnetic tape machine

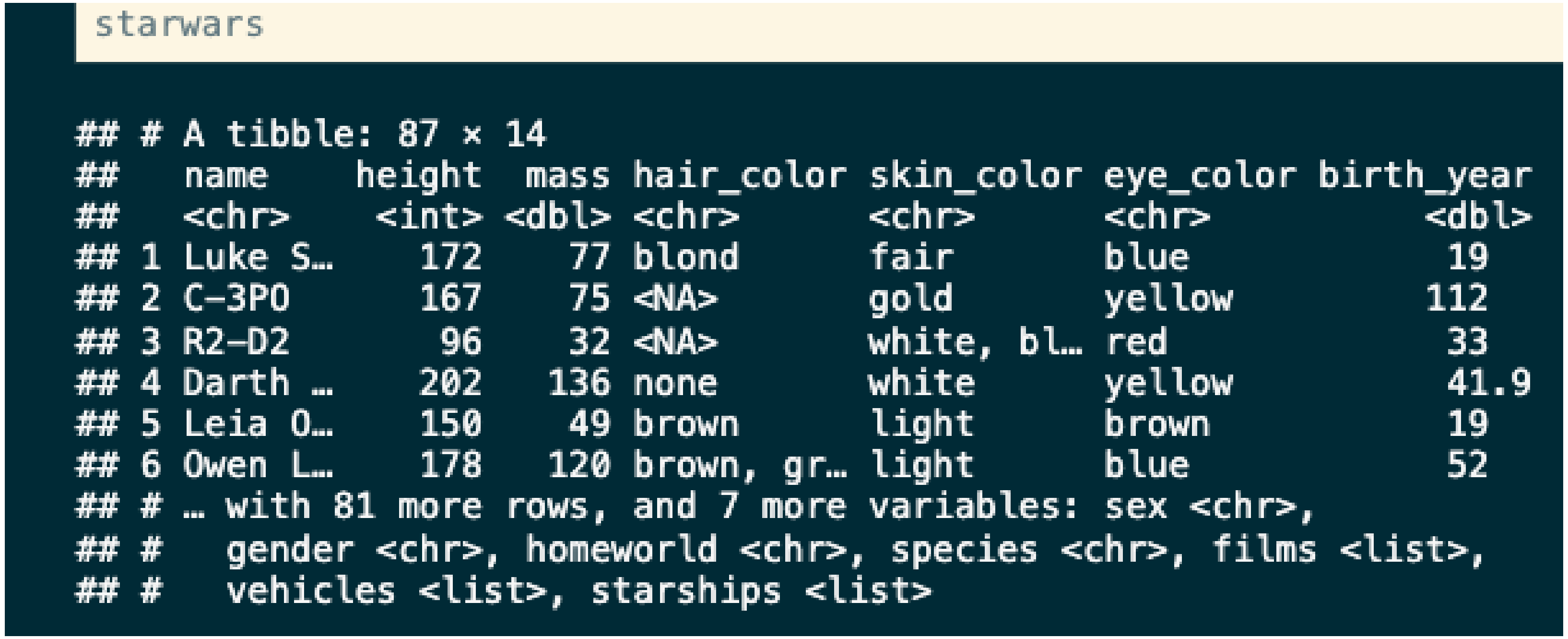
# Unstructured data



James Webb Telescope: Rings of Neptune



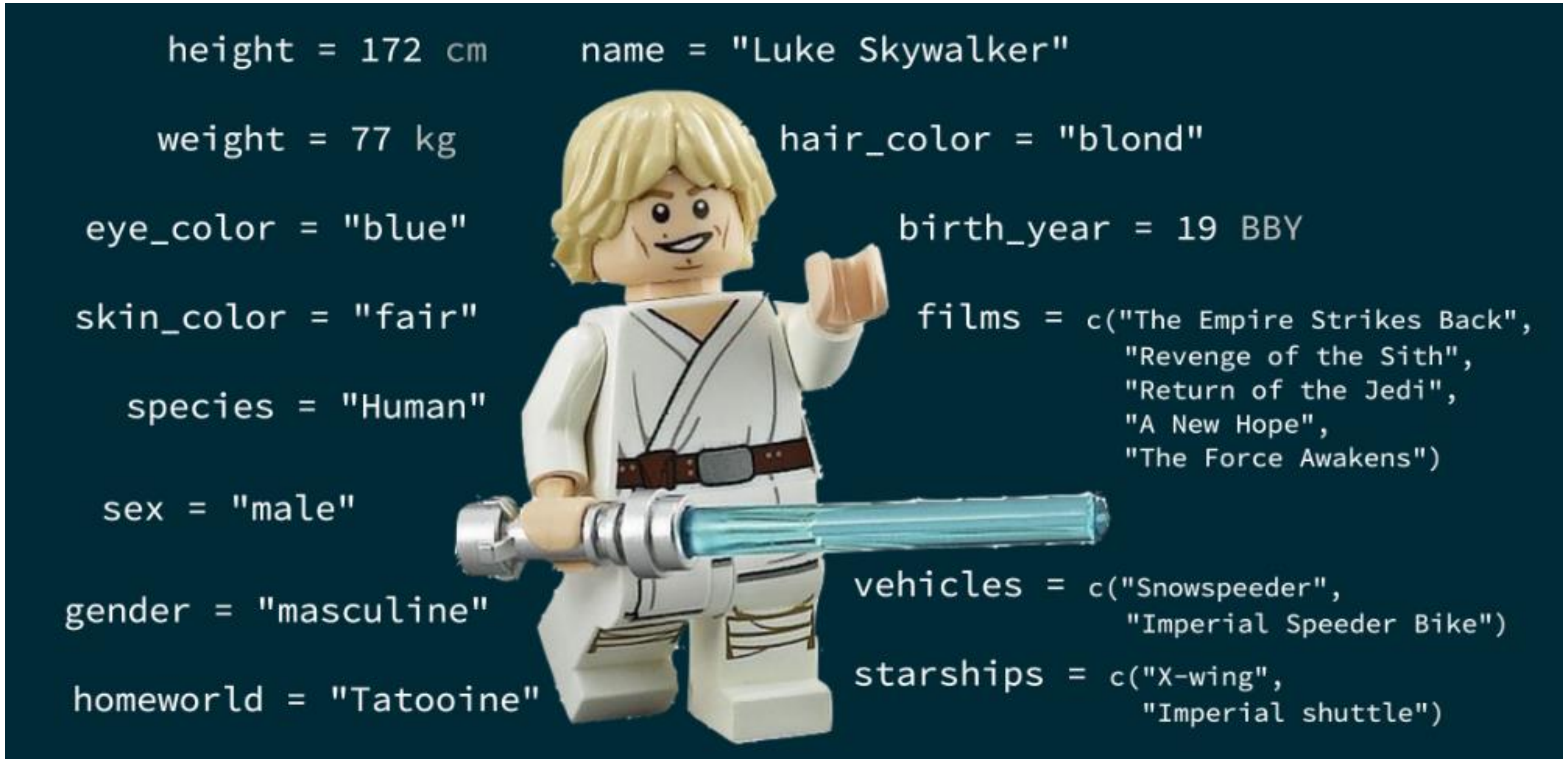
# Structured Data



```
starwars
```

```
## # A tibble: 87 × 14
##   name     height  mass hair_color skin_color eye_color birth_year
##   <chr>     <int> <dbl> <chr>      <chr>      <chr>          <dbl>
## 1 Luke S…     172    77 blond      fair       blue            19  
## 2 C-3PO       167    75 <NA>       gold       yellow         112  
## 3 R2-D2        96    32 <NA>       white, bl… red             33  
## 4 Darth …     202   136 none       white      yellow          41.9
## 5 Leia O…     150    49 brown      light      brown           19  
## 6 Owen L…     178   120 brown, gr… light      blue            52  
## # … with 81 more rows, and 7 more variables: sex <chr>,
## #   gender <chr>, homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

# Luke Skywalker



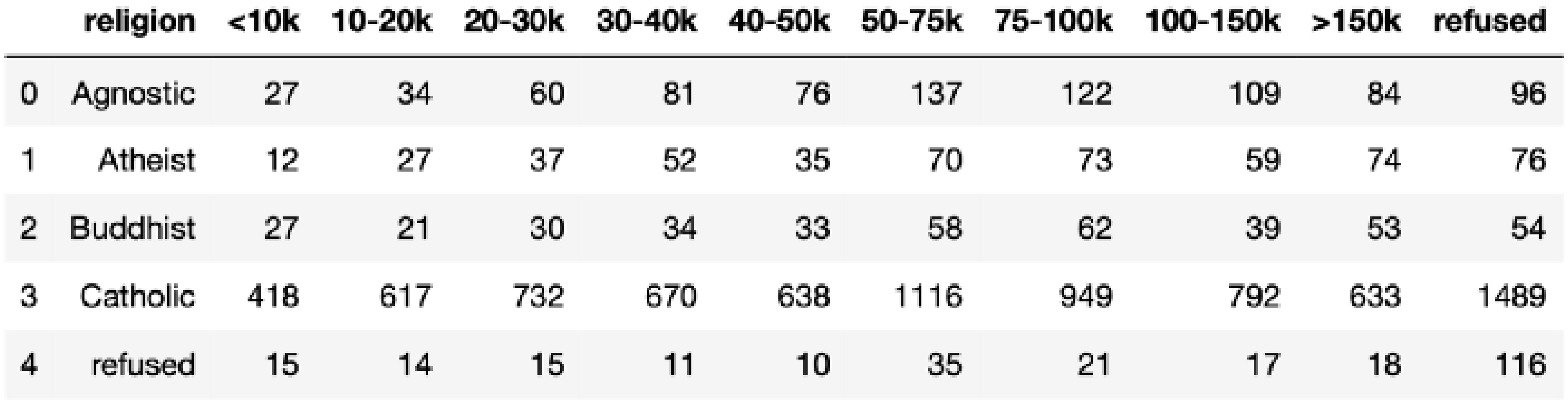
```
height = 172 cm
weight = 77 kg
eye_color = "blue"
skin_color = "fair"
species = "Human"
sex = "male"
gender = "masculine"
homeworld = "Tatooine"

name = "Luke Skywalker"
hair_color = "blond"
birth_year = 19 BBY
films = c("The Empire Strikes Back",
          "Revenge of the Sith",
          "Return of the Jedi",
          "A New Hope",
          "The Force Awakens")
vehicles = c("Snowspeeder",
             "Imperial Speeder Bike")
starships = c("X-wing",
              "Imperial shuttle")
```

From

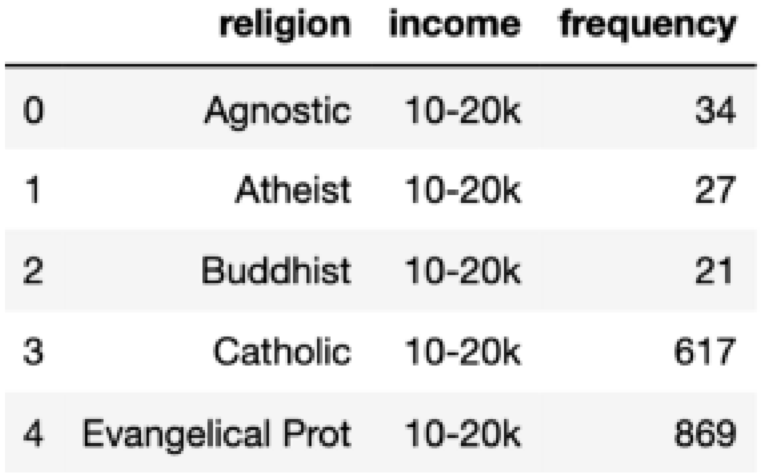
# Tidy Data

- Each variable forms a column.
- Each observation forms a row.
- Each unit forms a table.
- Tables may contain
  - Univariate data
  - Bivariate data
  - Multivariate data

# Messy vs. Tidy data



| | religion | <10k | 10-20k | 20-30k | 30-40k | 40-50k | 50-75k | 75-100k | 100-150k | >150k | refused |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Agnostic | 27 | 34 | 60 | 81 | 76 | 137 | 122 | 109 | 84 | 96 |
| 1 | Atheist | 12 | 27 | 37 | 52 | 35 | 70 | 73 | 59 | 74 | 76 |
| 2 | Buddhist | 27 | 21 | 30 | 34 | 33 | 58 | 62 | 39 | 53 | 54 |
| 3 | Catholic | 418 | 617 | 732 | 670 | 638 | 1116 | 949 | 792 | 633 | 1489 |
| 4 | refused | 15 | 14 | 15 | 11 | 10 | 35 | 21 | 17 | 18 | 116 |



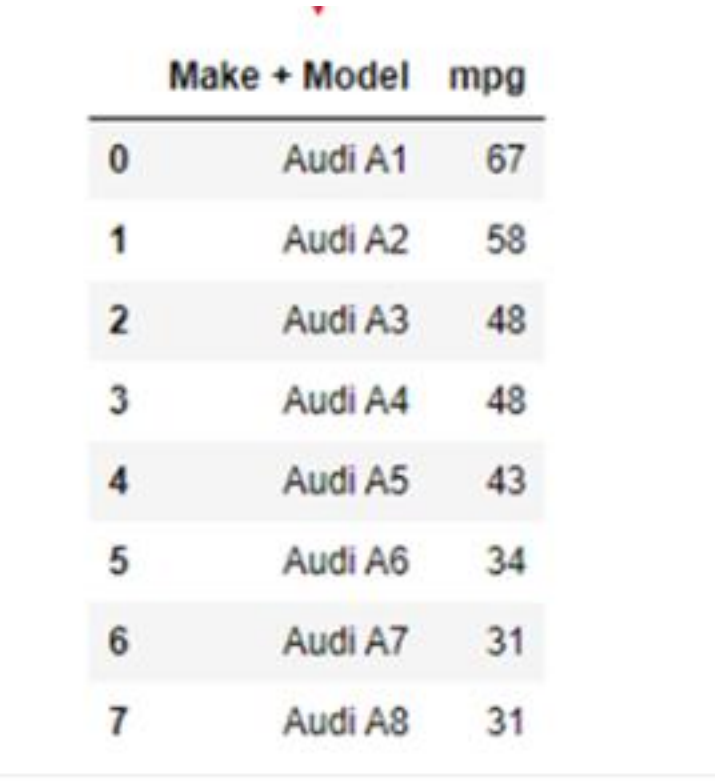
| | religion | income | frequency |
|---|---|---|---|
| 0 | Agnostic | 10-20k | 34 |
| 1 | Atheist | 10-20k | 27 |
| 2 | Buddhist | 10-20k | 21 |
| 3 | Catholic | 10-20k | 617 |
| 4 | Evangelical Prot | 10-20k | 869 |

https://towardsdatascience.com/whats-tidy-data-how-to-organize-messy-datasets-in-python-with-melt-and-pivotable-functions-5d52daa996c9

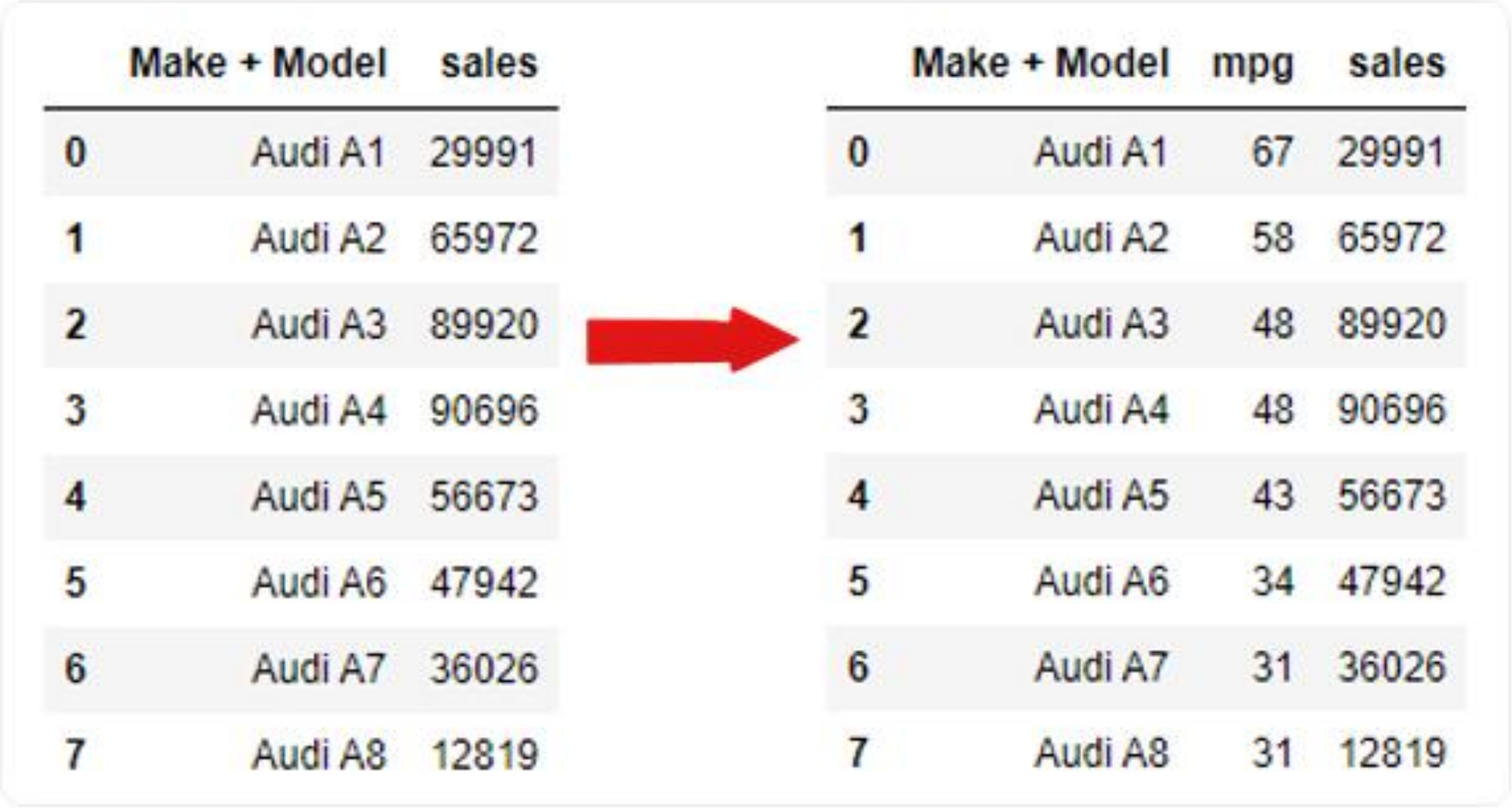
# Cleaning data

- Mixing text and numbers
- Missing data
- Mislabeled data
- Munging the data

# Merging two tables



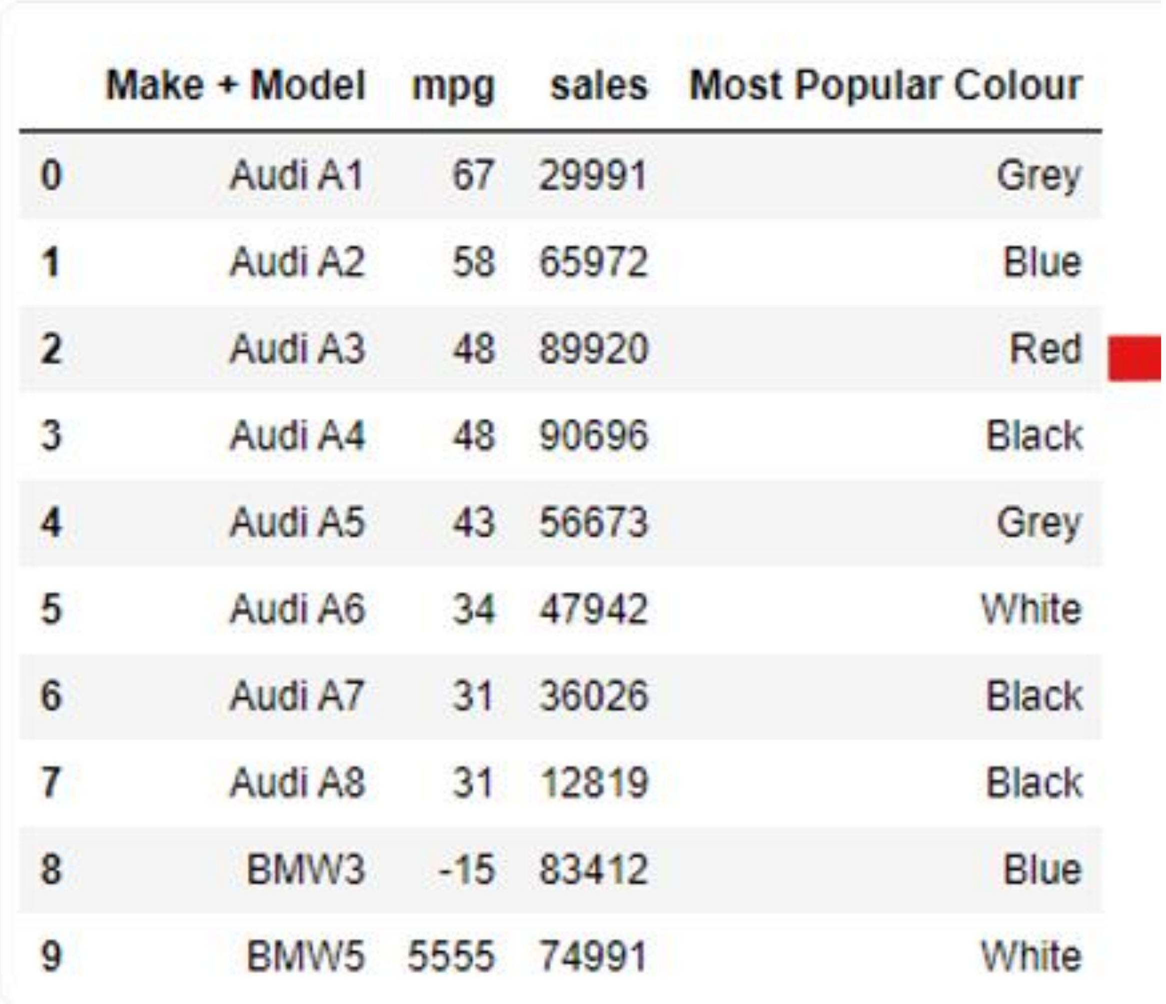
| | Make + Model | mpg |
|---|---|---|
| 0 | Audi A1 | 67 |
| 1 | Audi A2 | 58 |
| 2 | Audi A3 | 48 |
| 3 | Audi A4 | 48 |
| 4 | Audi A5 | 43 |
| 5 | Audi A6 | 34 |
| 6 | Audi A7 | 31 |
| 7 | Audi A8 | 31 |



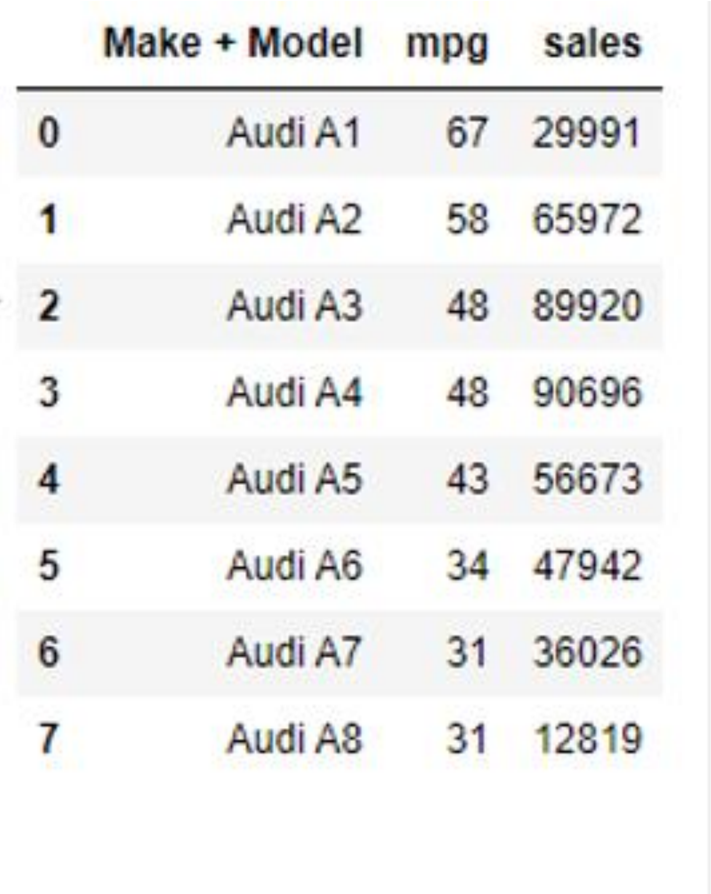
| | Make + Model | sales |
|---|---|---|
| 0 | Audi A1 | 29991 |
| 1 | Audi A2 | 65972 |
| 2 | Audi A3 | 89920 |
| 3 | Audi A4 | 90696 |
| 4 | Audi A5 | 56673 |
| 5 | Audi A6 | 47942 |
| 6 | Audi A7 | 36026 |
| 7 | Audi A8 | 12819 |

| | Make + Model | mpg | sales |
|---|---|---|---|
| 0 | Audi A1 | 67 | 29991 |
| 1 | Audi A2 | 58 | 65972 |
| 2 | Audi A3 | 48 | 89920 |
| 3 | Audi A4 | 48 | 90696 |
| 4 | Audi A5 | 43 | 56673 |
| 5 | Audi A6 | 34 | 47942 |
| 6 | Audi A7 | 31 | 36026 |
| 7 | Audi A8 | 31 | 12819 |

# Cleaning



| | Make + Model | mpg | sales | Most Popular Colour |
|---|---|---|---|---|
| 0 | Audi A1 | 67 | 29991 | Grey |
| 1 | Audi A2 | 58 | 65972 | Blue |
| 2 | Audi A3 | 48 | 89920 | Red |
| 3 | Audi A4 | 48 | 90696 | Black |
| 4 | Audi A5 | 43 | 56673 | Grey |
| 5 | Audi A6 | 34 | 47942 | White |
| 6 | Audi A7 | 31 | 36026 | Black |
| 7 | Audi A8 | 31 | 12819 | Black |
| 8 | BMW3 | -15 | 83412 | Blue |
| 9 | BMW5 | 5555 | 74991 | White |



| | Make + Model | mpg | sales |
|---|---|---|---|
| 0 | Audi A1 | 67 | 29991 |
| 1 | Audi A2 | 58 | 65972 |
| 2 | Audi A3 | 48 | 89920 |
| 3 | Audi A4 | 48 | 90696 |
| 4 | Audi A5 | 43 | 56673 |
| 5 | Audi A6 | 34 | 47942 |
| 6 | Audi A7 | 31 | 36026 |
| 7 | Audi A8 | 31 | 12819 |

# Joining Tables

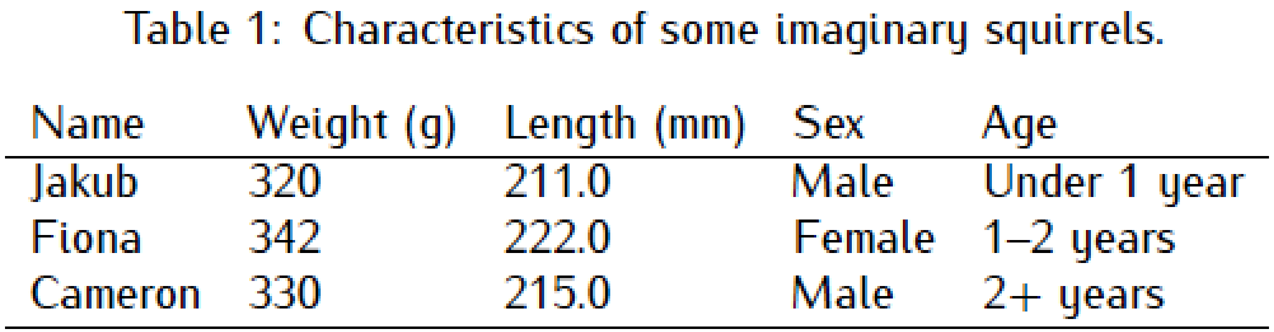


Table 1: Characteristics of some imaginary squirrels.

| Name | Weight (g) | Length (mm) | Sex | Age |
|---|---|---|---|---|
| Jakub | 320 | 211.0 | Male | Under 1 year |
| Fiona | 342 | 222.0 | Female | 1–2 years |
| Cameron | 330 | 215.0 | Male | 2+ years |

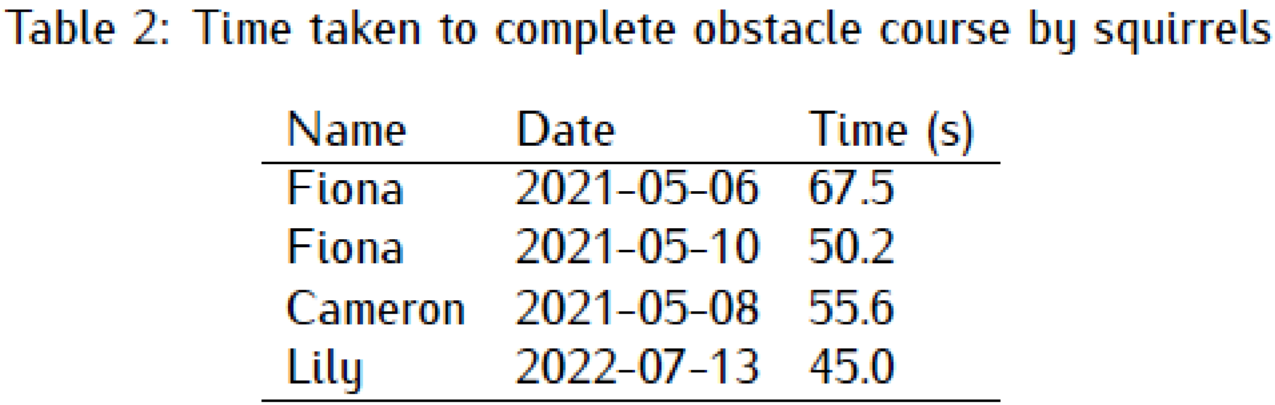


Table 2: Time taken to complete obstacle course by squirrels

| Name | Date | Time (s) |
|---|---|---|
| Fiona | 2021-05-06 | 67.5 |
| Fiona | 2021-05-10 | 50.2 |
| Cameron | 2021-05-08 | 55.6 |
| Lily | 2022-07-13 | 45.0 |

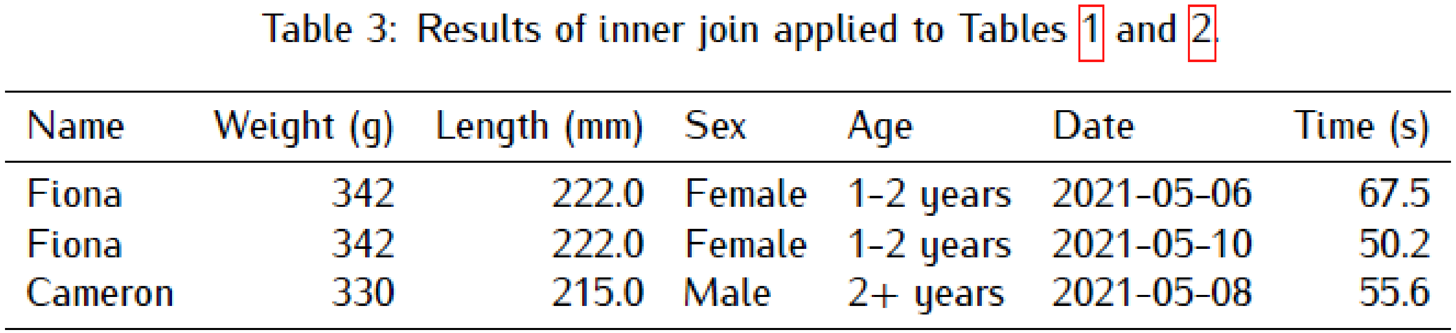


Table 3: Results of inner join applied to Tables 1 and 2.

| Name | Weight (g) | Length (mm) | Sex | Age | Date | Time (s) |
|---|---:|---:|---|---|---|---:|
| Fiona | 342 | 222.0 | Female | 1-2 years | 2021-05-06 | 67.5 |
| Fiona | 342 | 222.0 | Female | 1-2 years | 2021-05-10 | 50.2 |
| Cameron | 330 | 215.0 | Male | 2+ years | 2021-05-08 | 55.6 |

Only the squirrels in both datasets will be present in the joined dataset.

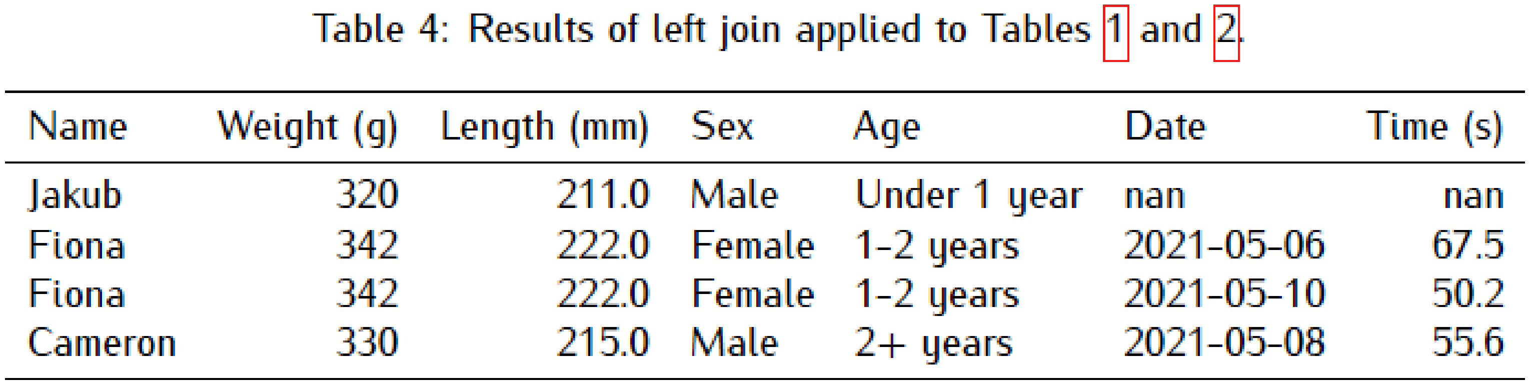


Table 4: Results of left join applied to Tables 1 and 2.

| Name | Weight (g) | Length (mm) | Sex | Age | Date | Time (s) |
|---|---:|---:|---|---|---|---:|
| Jakub | 320 | 211.0 | Male | Under 1 year | nan | nan |
| Fiona | 342 | 222.0 | Female | 1-2 years | 2021-05-06 | 67.5 |
| Fiona | 342 | 222.0 | Female | 1-2 years | 2021-05-10 | 50.2 |
| Cameron | 330 | 215.0 | Male | 2+ years | 2021-05-08 | 55.6 |

All squirrels present in the first table will be present

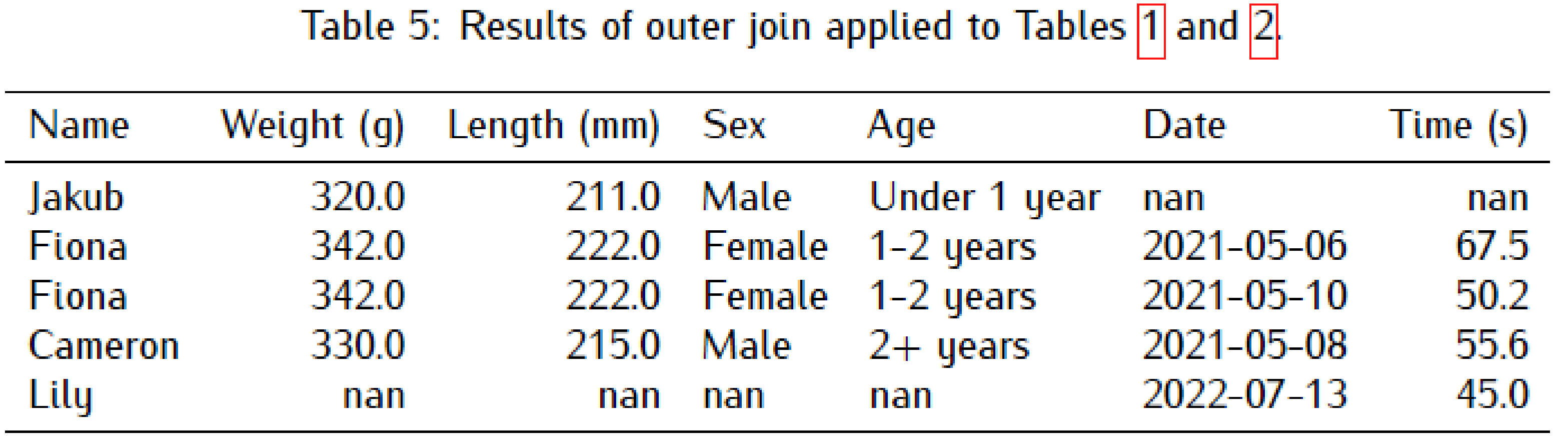


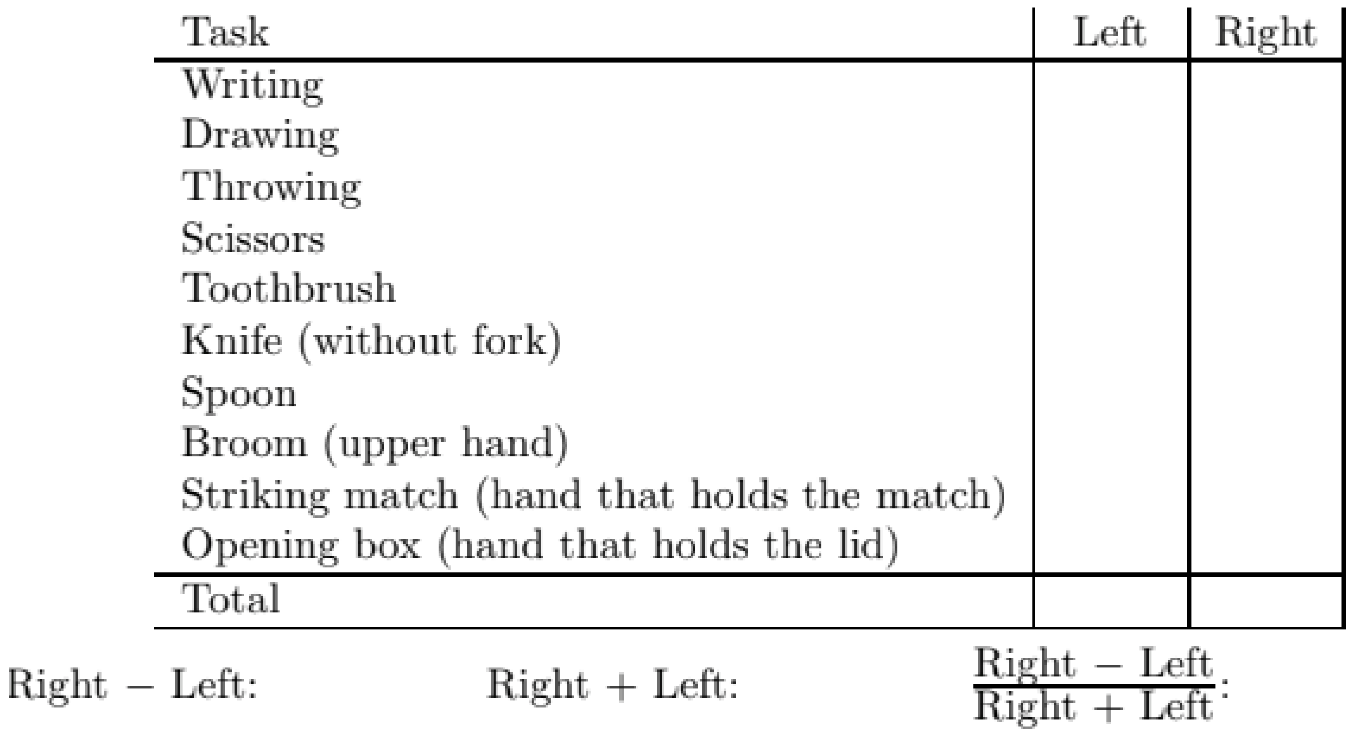
Table 5: Results of outer join applied to Tables 1 and 2.

| Name | Weight (g) | Length (mm) | Sex | Age | Date | Time (s) |
|---|---|---|---|---|---|---|
| Jakub | 320.0 | 211.0 | Male | Under 1 year | nan | nan |
| Fiona | 342.0 | 222.0 | Female | 1-2 years | 2021-05-06 | 67.5 |
| Fiona | 342.0 | 222.0 | Female | 1-2 years | 2021-05-10 | 50.2 |
| Cameron | 330.0 | 215.0 | Male | 2+ years | 2021-05-08 | 55.6 |
| Lily | nan | nan | nan | nan | 2022-07-13 | 45.0 |

All squirrels in both tables will be present

# Handedness Survey

# Handedness score



| Task | Left | Right |
|---|---|---|
| Writing | | |
| Drawing | | |
| Throwing | | |
| Scissors | | |
| Toothbrush | | |
| Knife (without fork) | | |
| Spoon | | |
| Broom (upper hand) | | |
| Striking match (hand that holds the match) | | |
| Opening box (hand that holds the lid) | | |
| Total | | |

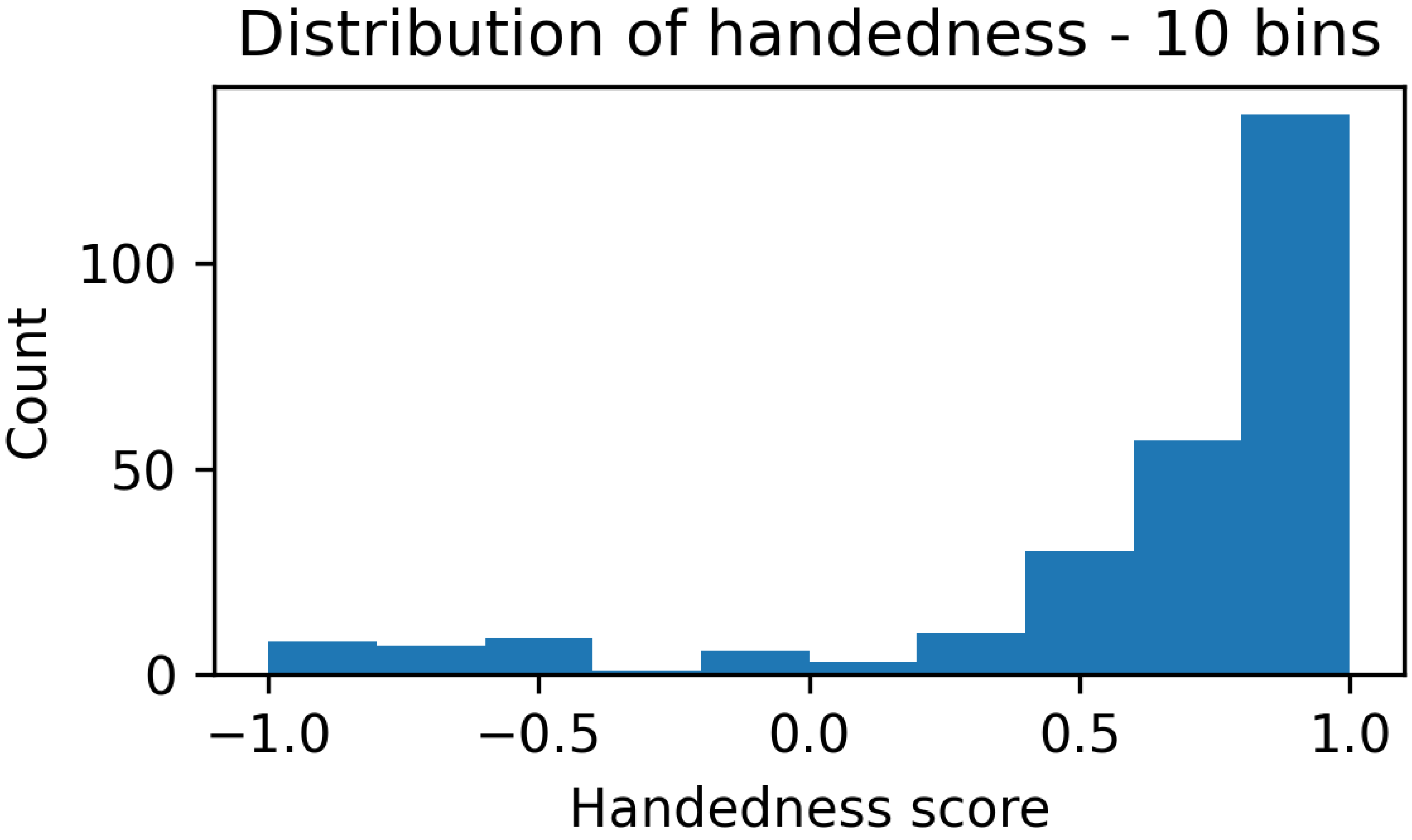
Right $-$ Left: Right $+$ Left: $\frac{\text{Right} - \text{Left}}{\text{Right} + \text{Left}}$:

# Handedness Histogram

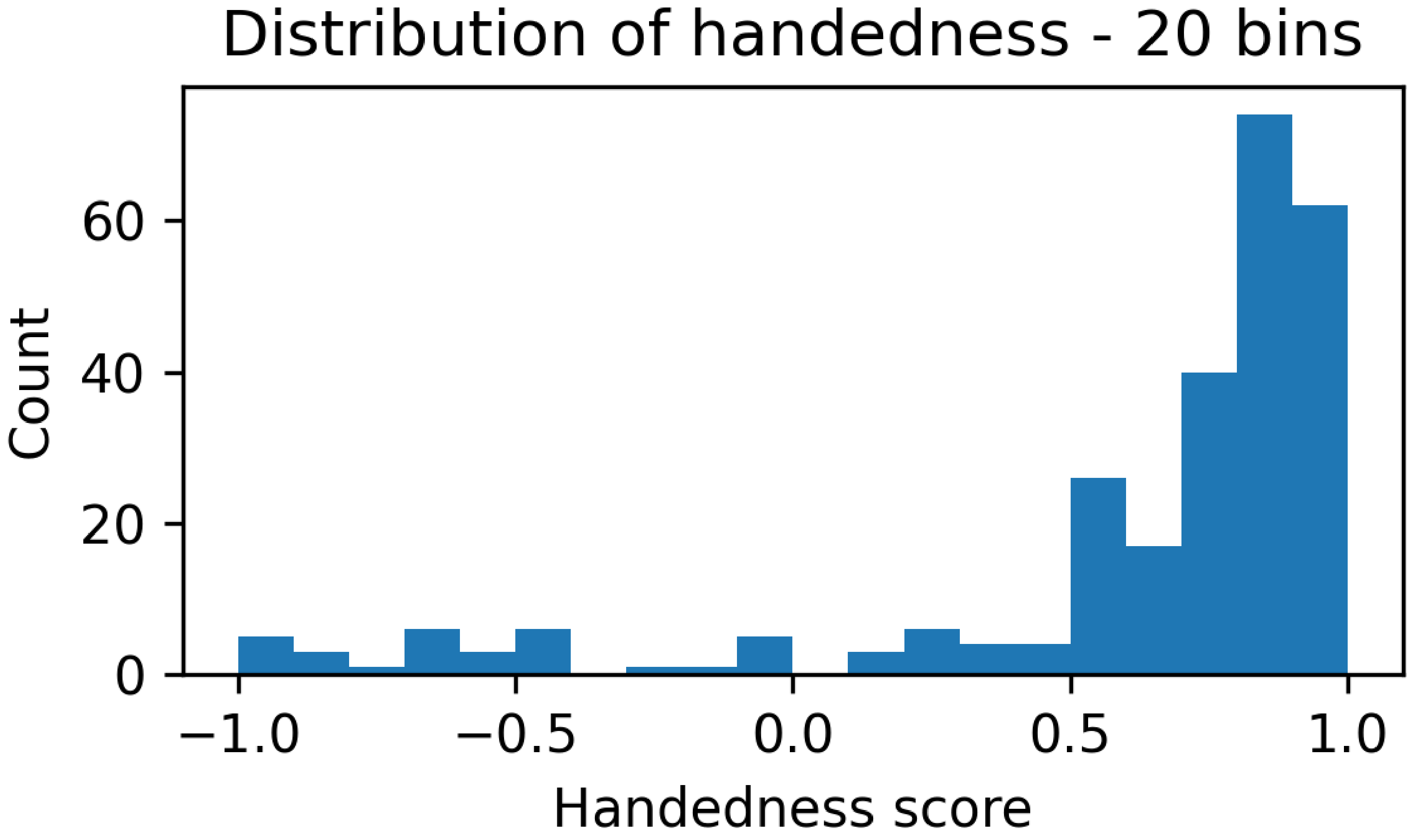
# students
-1
Handedness Score
+1

# Your data (10 bins)

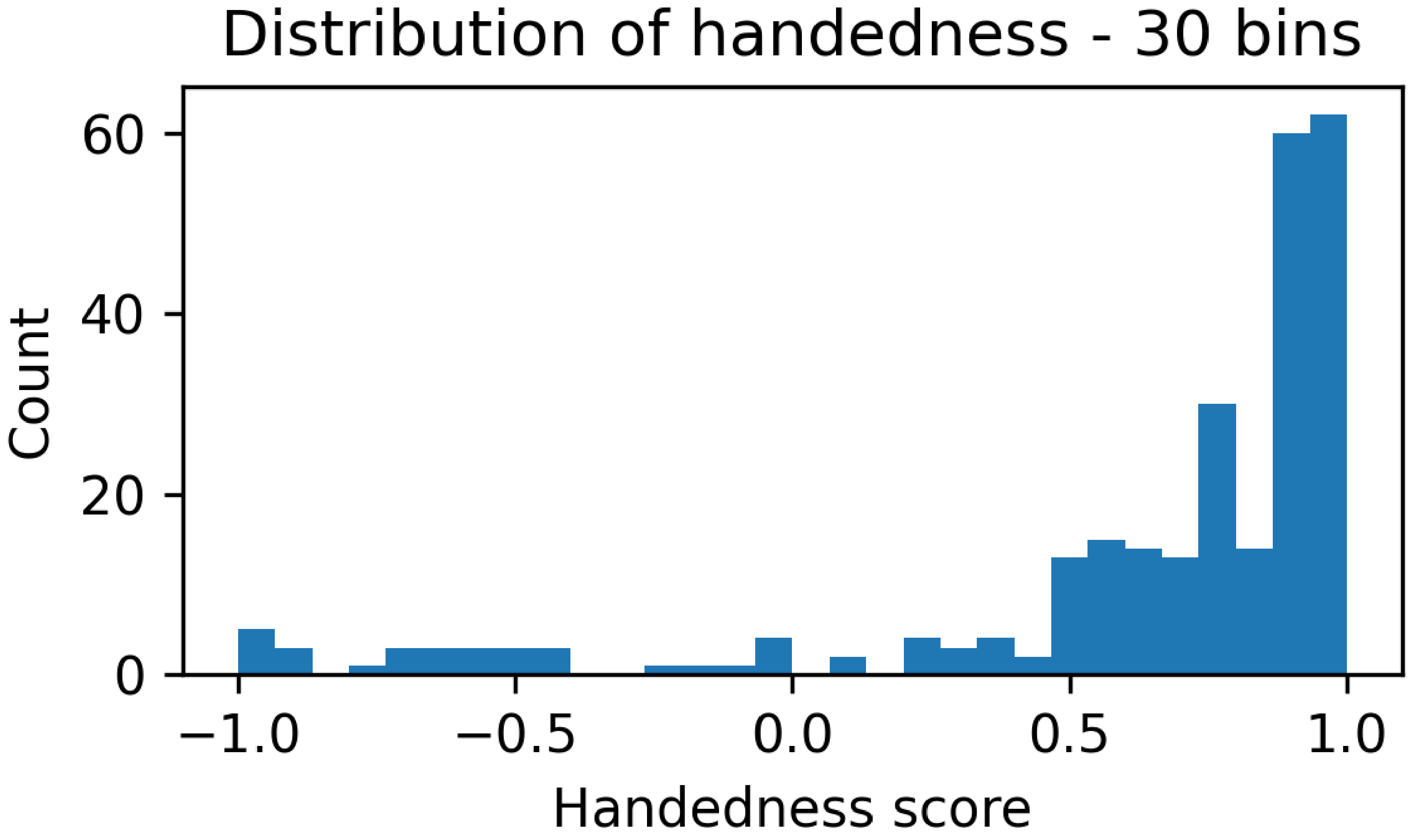
matplotlib.pyplot.hist()



# Your data (20 bins)



# Your data (30 bins)


Distribution of handedness - 30 bins
Count
Handedness score