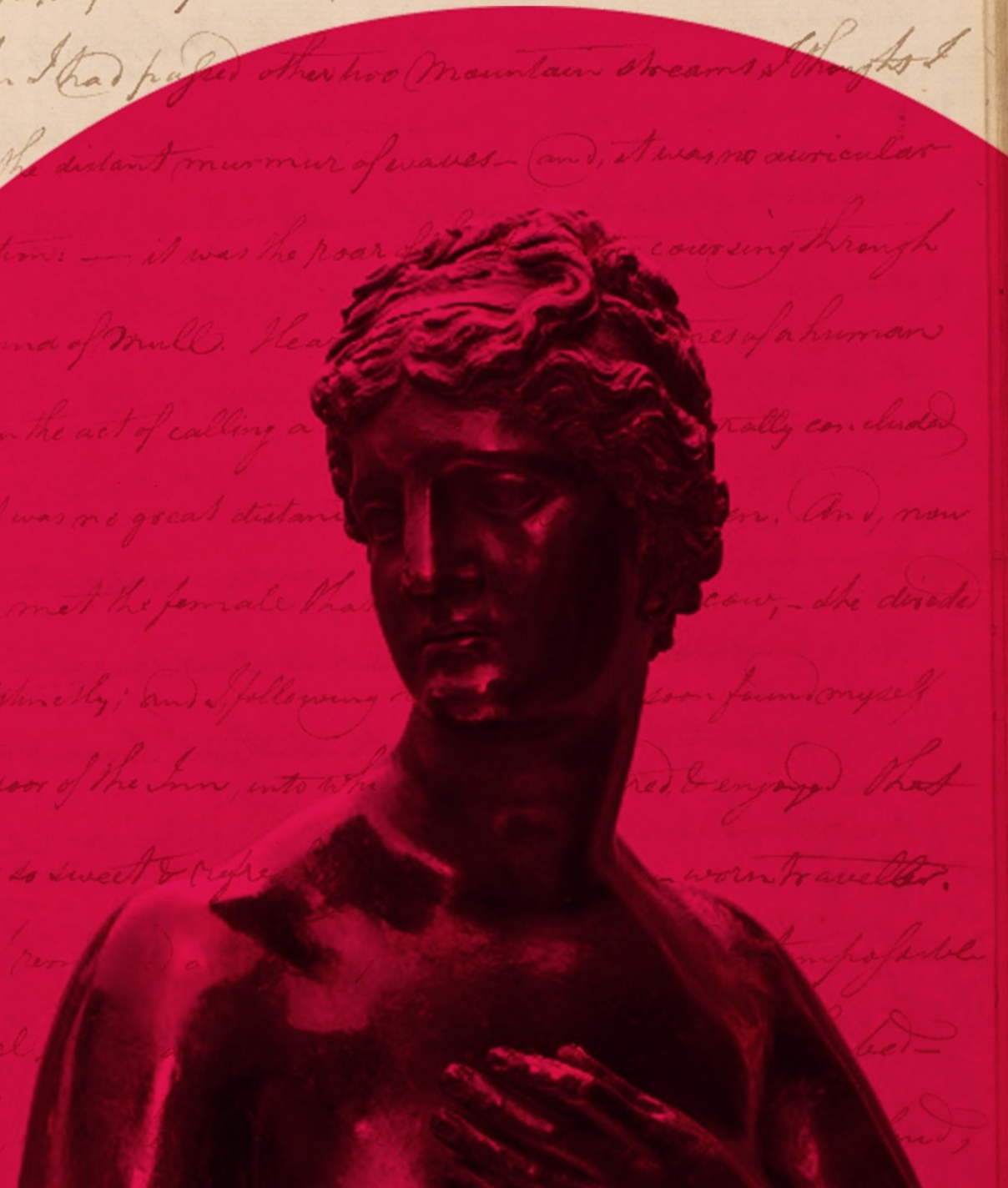
DATA CULTURE SOCIETY
CDCS.ED.AC.UK

SCOTTISH GRADUATE SCHOOL FOR ARTS & HUMANITIES
Sgoil Ceumnachaidh na h-Alba airson Ealain agus Daonnachdan
2018-2024

DAY 4



DATA CULTURE SOCIETY
CDCS.ED.AC.UK

SCOTTISH GRADUATE SCHOOL FOR ARTS & HUMANITIES
Sgoil Ceumnachaidh na h-Alba airson Ealain agus Daonnachdan
2014 - 2024

INTERLACING LOOPS, FUNCTIONS AND LISTS

WAKE UP TASK…

For each of the following in your pairs..

1. What is a list? Give an example of other types of similar data structures and where each have uses.

2. Where do we need to use functions?

3. How do we effectively use a loop?

Sometimes we want to use all this together.

```
def filter_above_average(lengths):
    avg = average(lengths)
    return [length for length in lengths if length > avg]
```

What does the above do?

What happens if all the data isn’t the same though…

```
lengths = [10, 15, 12, ‘gentoo’, 8]

def filter_above_average(lengths):
    avg = average(lengths)
    return [length for length in lengths if length > avg]

filtered_lengths = filter_above_average(lengths)
```

What happens here?

ERROR! But it’s all going to be ok, I promise!

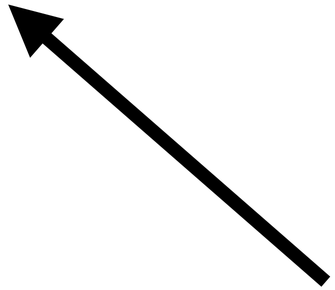
What might we instructively in English say to someone dealing with this kind of error?

```
lengths = [10, 15, 12, ‘gentoo’, 8]
```

“Give me the numbers in this list which are larger than the average.”

TRY…to do this thing…
EXCEPT…if you cannot then tell me!

```
try:
    result = 10 / 0
except ZeroDivisionError:
    result = None
print("Result:", result)
```



Notice here this is very specific. There are ‘built-in’ Error classes. You can build your own but we won’t go into that here. The built-in ones should cover most things you need!

So why do we need to combine loops, lists and functions?

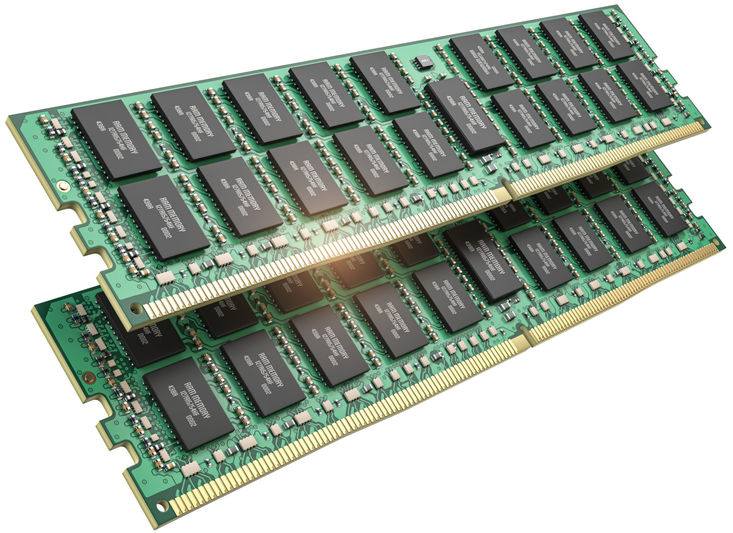
> Needed for complex data handling.
> Efficiently can analyse data.
> Your computer may be good, but it might not be that good…

TIME COMPLEXITY

Linear Time O(n): A loop over n items that performs a fixed amount of work for each item. This means we can just add the time is takes for one thing to happen in the loop n-times.

Quadratic Time O(n^2): Nested loops over the same dataset, like comparing each item with every other item. This type of time is bad as it scales ***exponentially*** and gets very slow very quick.

SPACE COMPLEXITY



Space Complexity refers to the amount of *memory space* required by an algorithm to run as a function of the length of the input.

LETS GET PROGRAMMING

Session 10: The Final Syntactic Frontier

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

COFFEE BREAK

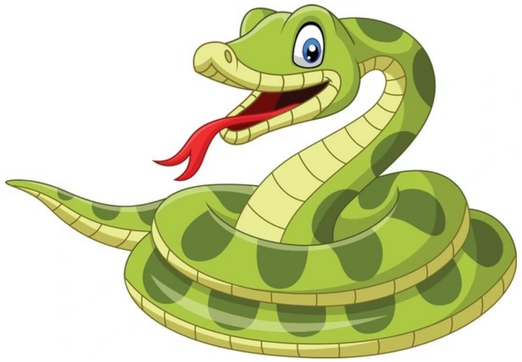
WE ARE GOING TO RESTART AT 11:00

DATA CULTURE SOCIETY
CDCS.ED.AC.UK

SCOTTISH GRADUATE SCHOOL FOR ARTS & HUMANITIES
Sgoil Ceumnachaidh na h-Alba airson Ealain agus Daonnachdan
2014 - 2024

DATA IMPORTING AND HANDLING

LETS MEET THE ANIMALS…



Python

Pandas

Penguins

WHAT IS PANDAS?

Origin: pandas was developed by Wes McKinney in 2008 while he was working at AQR Capital Management, a quantitative investment management firm.

Motivation: Wes McKinney aimed to create a powerful and flexible tool for financial data analysis, which led to the birth of pandas.

Open Source: Since its release, pandas has become a popular open-source library, maintained and improved by a large community of contributors.

WHAT IS PALMER PENGUINS?

Palmer Penguins is a dataset containing information about penguin species from the Palmer Archipelago in Antarctica.

Species	Island	Culmen Length (mm)	Culmen Depth (mm)	Flipper Length (mm)	Body Mass (g)	Sex
Adelie	Torgersen	39.1	18.7	181	3750	Male
Chinstrap	Dream	46.5	17.9	195	3650	Female
Gentoo	Biscoe	50.0	15.3	220	5550	Male

WHAT IS A DATAFRAME?

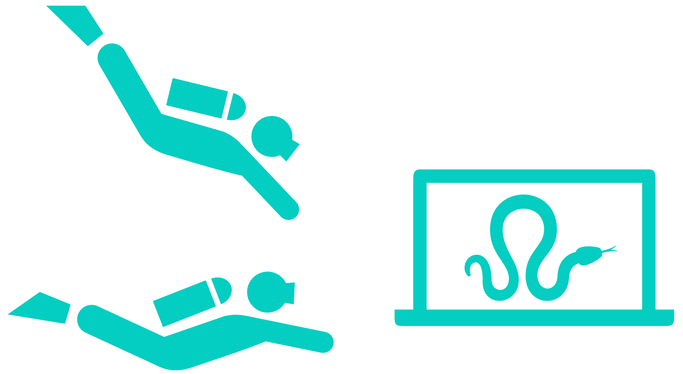
```
df = pd.DataFrame ( { 'num_legs' : [ 2, 4, 8, 0 ],
 'num_wings' : [ 2, 0, 0, 0 ],
 'num_specimen_seen' : [ 8, 2, 5, 6 ] },
 index = [ 'sparrow', 'fox', 'spider', 'snake' ] )
```

DataFrame

	num_legs	num_wings	num_specimen_seen
sparrow	2	2	8
fox	4	0	2
spider	8	0	5
snake	0	0	6

LETS GET PROGRAMMING

Session 11: Handle with Caution



THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

LUNCH BREAK

WE ARE GOING TO RESTART AT 13:30

DATA CULTURE SOCIETY
CDCS.ED.AC.UK

SCOTTISH GRADUATE SCHOOL FOR ARTS & HUMANITIES
Sgoil Ceumnachaidh na h-Alba airson Ealain agus Daonnachdan
2014 - 2024

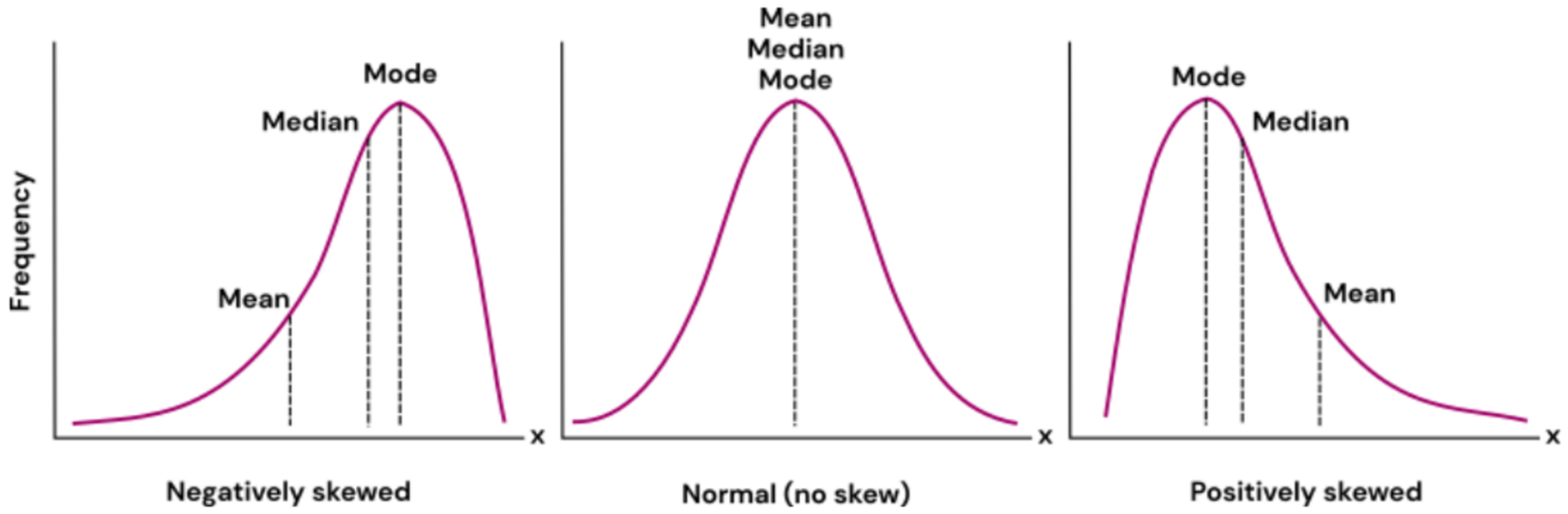
DATA SUMMARIES AND OVERVIEWS

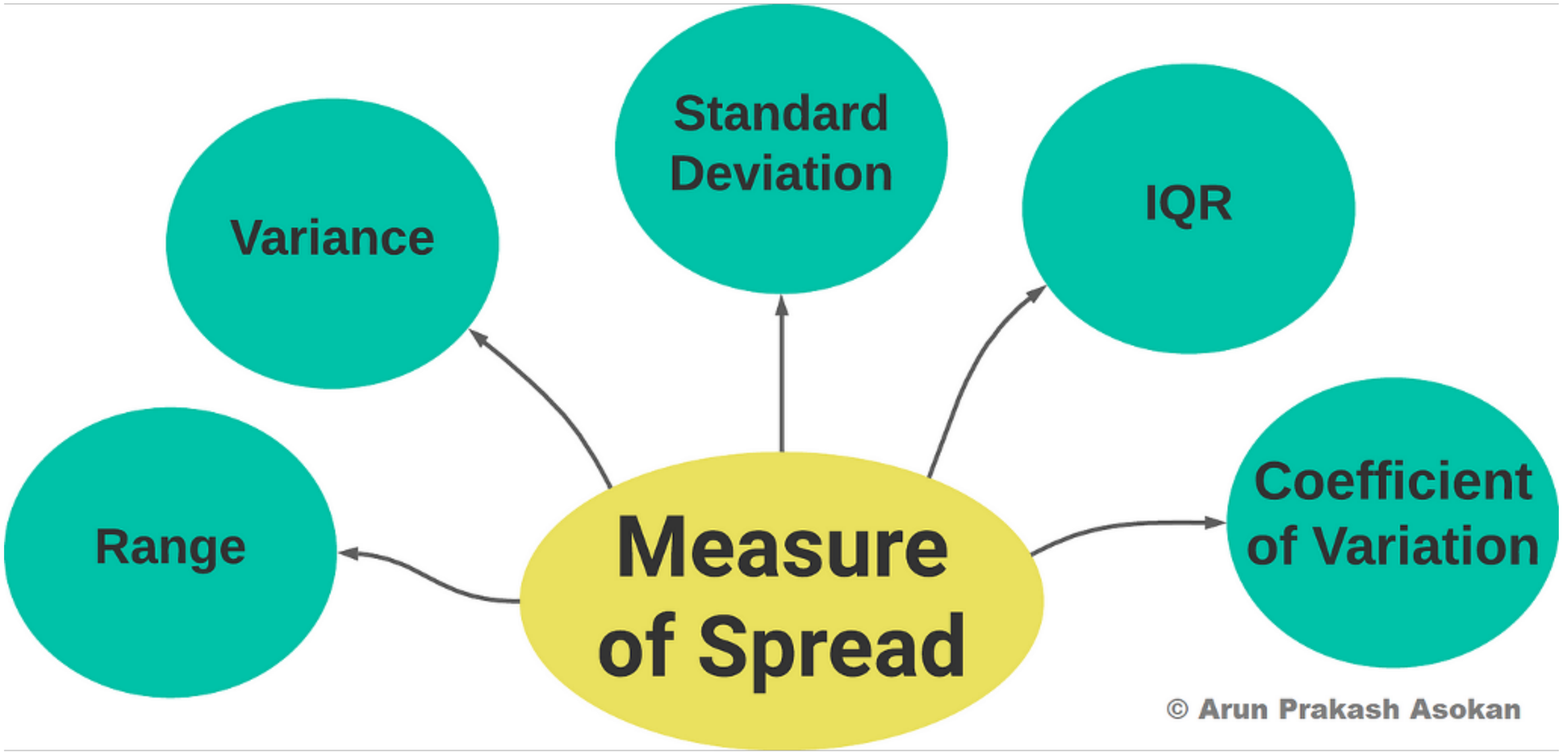
WHAT IS A DATA SUMMARY?

Ultimately we are talking about descriptive statistics!

Descriptive statistics involve summarising and presenting the main features of a dataset, such as measures of central tendency (mean, median, mode) and measures of variability (range, standard deviation).

MEASURES OF CENTRAL TENDENCY




Standard Deviation
Variance
IQR
Range
Measure of Spread
Coefficient of Variation
© Arun Prakash Asokan

In Python, we can get all this information very quickly!

```
summary_stats = penguins.describe()
```

Or individually instead.

```
mean_values = penguins.mean()
median_values = penguins.median()
```

We may also want to get descriptive statistics about specific subgroups of data

e.g. About one variable, grouped by another

```
group_by_species = penguins.groupby('species')
['flipper_length_mm'].mean()
```

Or maybe counting how many we have of a certain type.

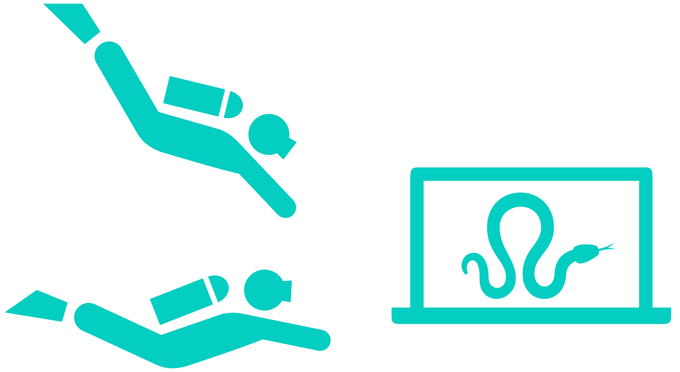
```
species_counts = penguins['species'].value_counts()
```

For this type of grouping, pivot tables are our friends!


```
pivot_table = penguins.pivot_table(values='body_mass_g',
index='species', columns='island', aggfunc='mean')
```

LETS GET PROGRAMMING

Session 12: Birdseye View



THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

COFFEE BREAK

WE ARE GOING TO RESTART AT 15:30

DATA CULTURE SOCIETY
CDCS.ED.AC.UK

SCOTTISH GRADUATE SCHOOL FOR ARTS & HUMANITIES
Sgoil Ceumnachaidh na h-Alba airson Ealain agus Daonnachdan
2014 - 2024

DATA CLEANING

LETS START AT MISSING DATA

- Missing data is a common issue and can occur for various reasons.

- It's important to identify and handle missing data appropriately to ensure accurate data analysis.

LETS START AT MISSING DATA

- Missing data is a common issue and can occur for various reasons.

Poor collection Data corruption Doesn’t apply to all records

- It's important to identify and handle missing data appropriately to ensure accurate data analysis.

The first step is always to identify if we have missing data.

```
missing_values_count = penguins.isnull().sum()
```

What we do next depends on ‘opinion’…
(Chris’ subtext: there is no opinion but statistical principals that dictate how we deal with missing data that are mostly followed…but we won’t get into that here)

We have a two immediate options to pick from…

1. Remove the missing data records.

```
penguins_dropped_rows = penguins.dropna()
```

2. Fill it in somehow.

```
penguins_filled_mean =
penguins.fillna(penguins.mean(numeric_only=True))
```

THE REST OF DATA CLEANING IS FAIRLY INTUITIVE

```
2 == 2
2 == 3

"penguin" == "penguin"
"penguin" == “whale"
"penguin" == "Penguin"

"2024" == 2024
```

Remember this on Day 1?

We can combine logic with pandas to filter our data.

```
adelie_penguins = penguins[penguins['species'] == 'adelie']
```

We have to select the whole data set first and use square brackets.

Then we select the specific column we want to base our condition on.

After this provide some logic condition.

Finally select what the condition needs to satisfy.

LETS GET PROGRAMMING

Session 13: my_data, my_way