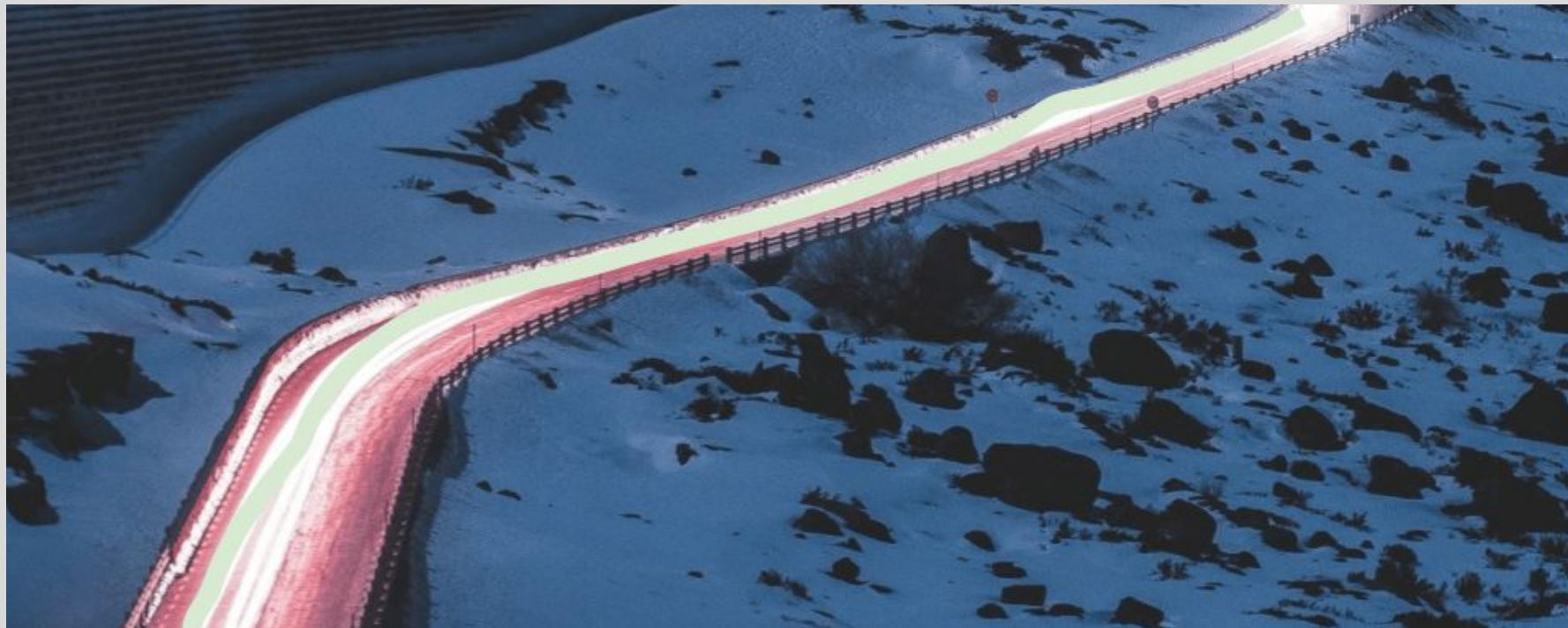


Exploring worlds through data

Prediction focus



Teaching focus rather than assessment ...

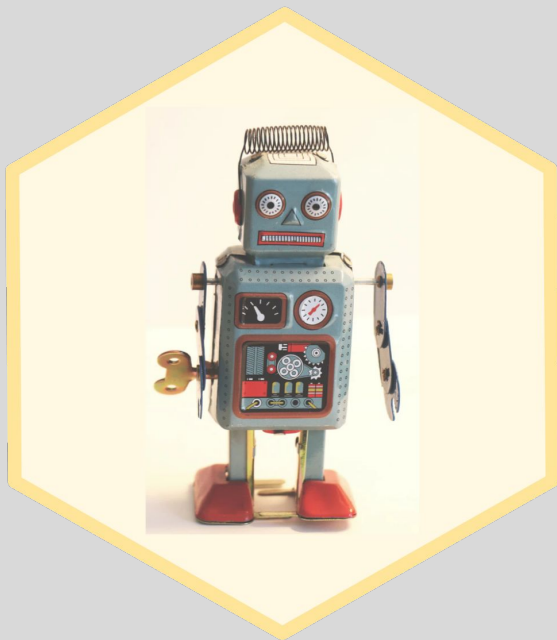
There's lots of things that are uncertain about what we may be teaching in the near future, due to the NCEA change programme (RAS) and the curriculum refresh project.

You might be looking for something definitive, like a set of steps that explain exactly how to assess one of the new proposed NCEA Level One Mathematics and Statistics achievement standards. This is totally understandable, but you won't find that in our materials :-)

What we have focused on for our workshop are what we believe are core ideas related to prediction that could both benefit our teaching right now but also inform how we could be teaching in the future.

We hope that through our examples and notes that you do obtain a clear understanding of what would be important to teach, and so assess, and gain some new ideas for data contexts and ways to engage ākonga with their learning from data.

... with an unashamedly data science influence!



Foundations for predictive modelling at Y12/13 (e.g. image recognition)

Dynamic sources of data

Access, explore, and use data about ourselves

Build awareness of digital technologies & related data technologies

Create awesome things from data!

Clare Nelson
Tangaroa College



Jim Davis
Westlake Boys'
High School



SCIENCE
DEPARTMENT OF STATISTICS

Amy Hooper
Cashmere High
School



Hanna Reid
St Peter's
Cambridge



Jacqui Hammond
Ormiston Senior
College



Ash Rambhai
Botany Downs Secondary
College



Marion Steel
Epsom Girls
Grammar



Lisa Mulvey
Selwyn College



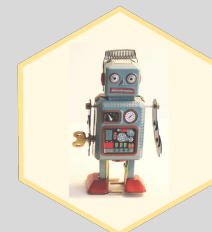
Chris Wild
University
of Auckland



Anna Fergusson
University
of Auckland



The prediction team!



1. A key statistical modelling approach is to use the data you have to predict an outcome you don't (yet) have.

2. Prediction is a broad core idea and exploration provides an accessible, flexible, creative way of supporting predictive modelling ideas

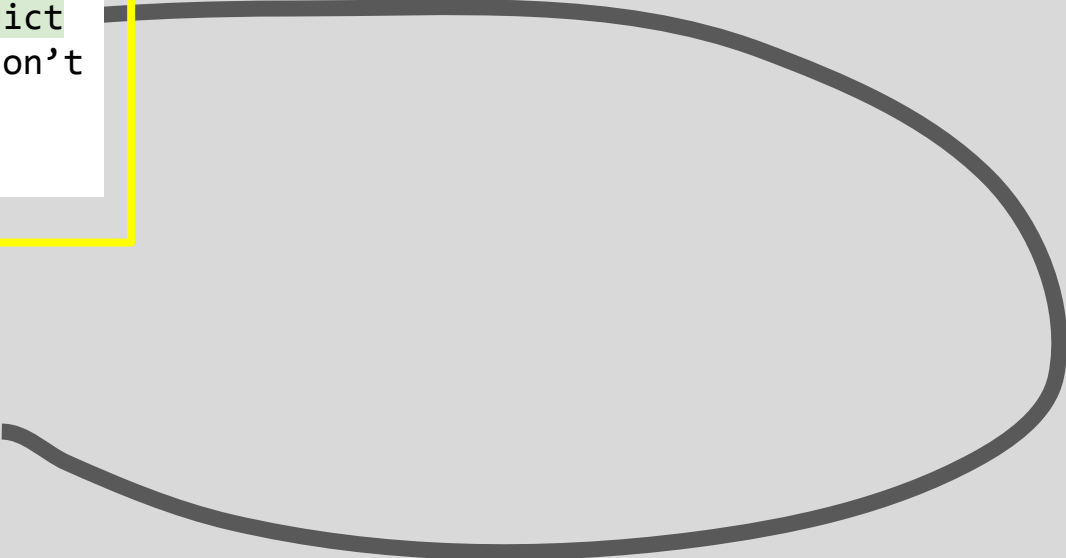
3. Time series data great intro for weaving storytelling and informal predictive modelling (& for working more closely/personally with digital data + CT)

6. Putting it all together with ePPDAC

5. Learning from features of scatterplots, using informal methods for prediction, evaluating models in terms of accuracy and precision

4. Different purposes/goals for prediction and different ways of designing data (e.g. experiments)

1. A key statistical modelling approach is to use the data you have to predict an outcome you don't (yet) have.

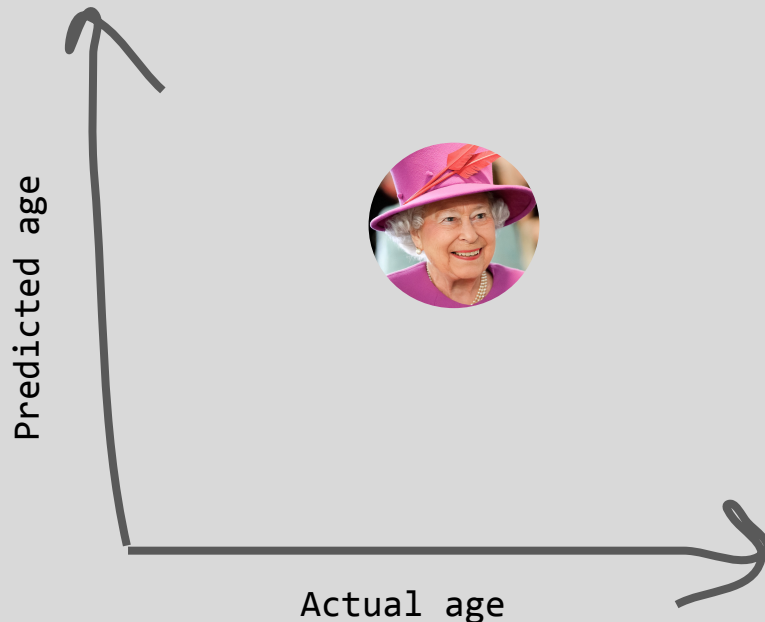


Predicting ages ...

“Oldie but a goodie”

<https://teaching.statistics-is-awesome.org/exploring-statistical-measures-by-estimating-the-ages-of-famous-people/>

- Show photos of famous people
- Students predict/guess age of each person



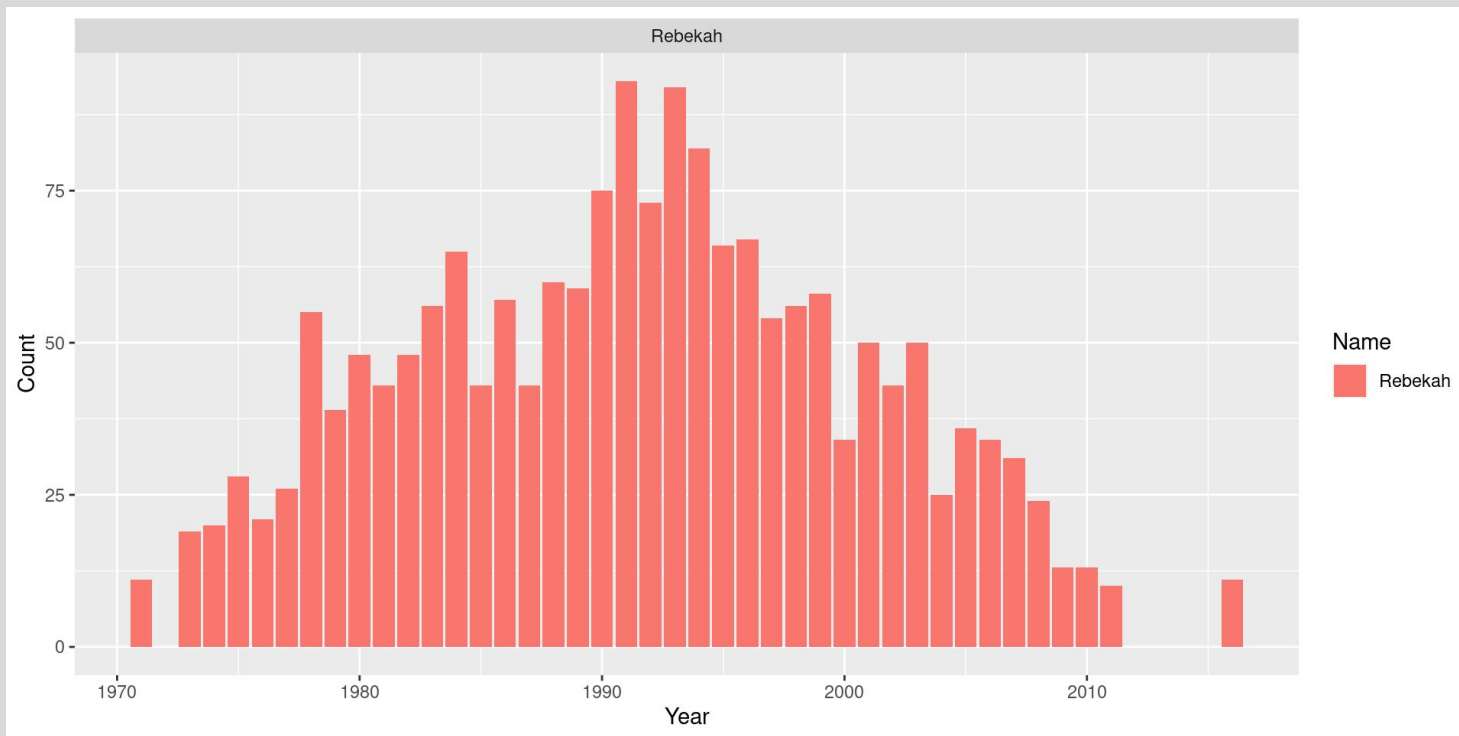
Jacinda

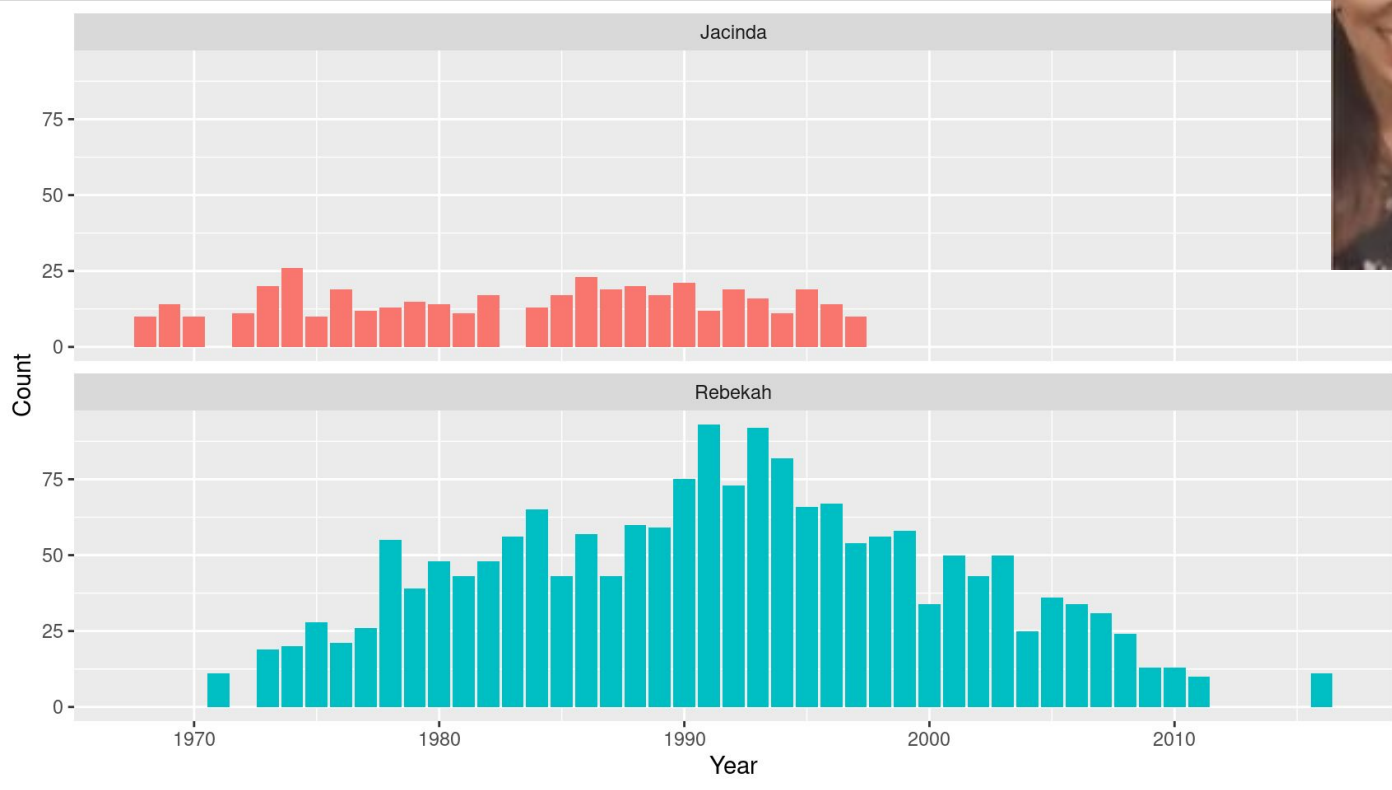


Rebekah

Complete the Google form
with your prediction for
the age of Rebekah

Also provide some reasons
why you think Rebekah is
this age



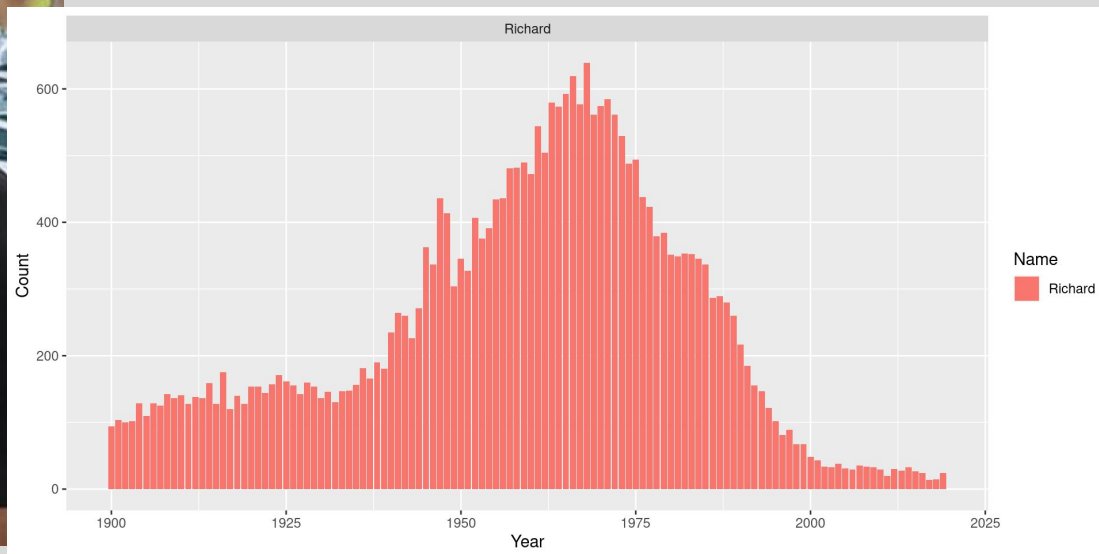


Richard

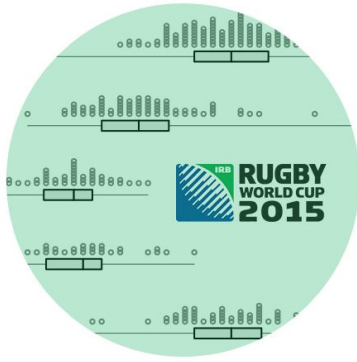


Complete the Google form with
you prediction for the age of
Richard

Also provide some reasons why
you think Richard is this age

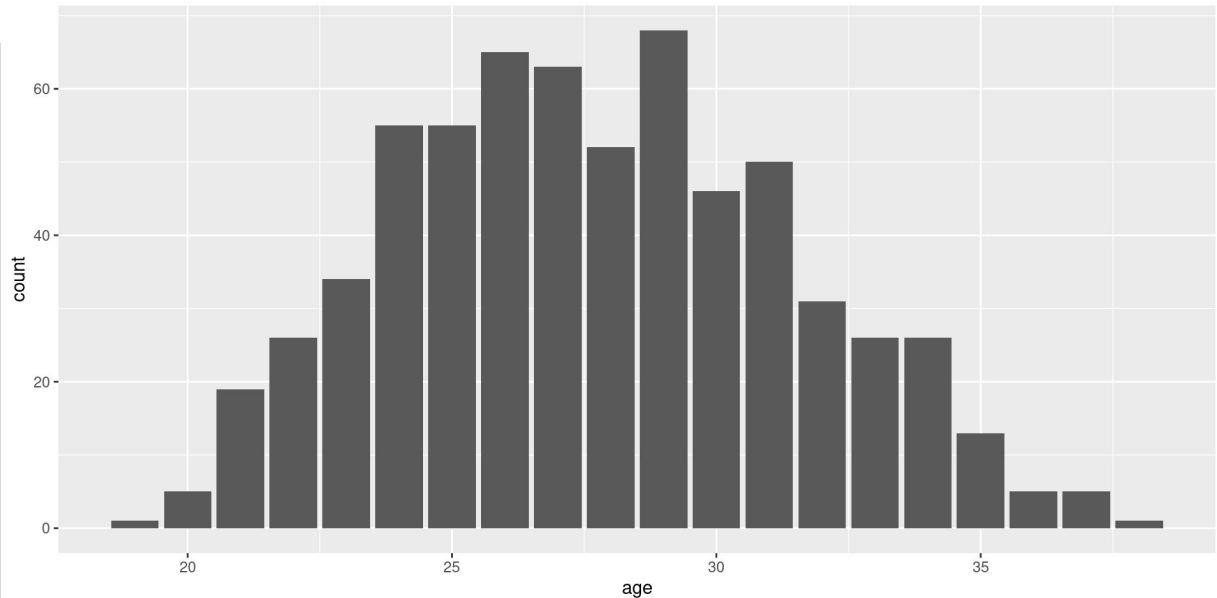


Rugby World Cup 2015 players (population data)



Complete the Google form with
you prediction for the age of
Richard

Also provide some reasons why
you think Richard is this age



Who performed better across the two situations?

Person A who got:

- one age correct
- one age wrong by five years



Predicted 31



Predicted 32

Person B who got:

- one age wrong by one year
- one age wrong by two years

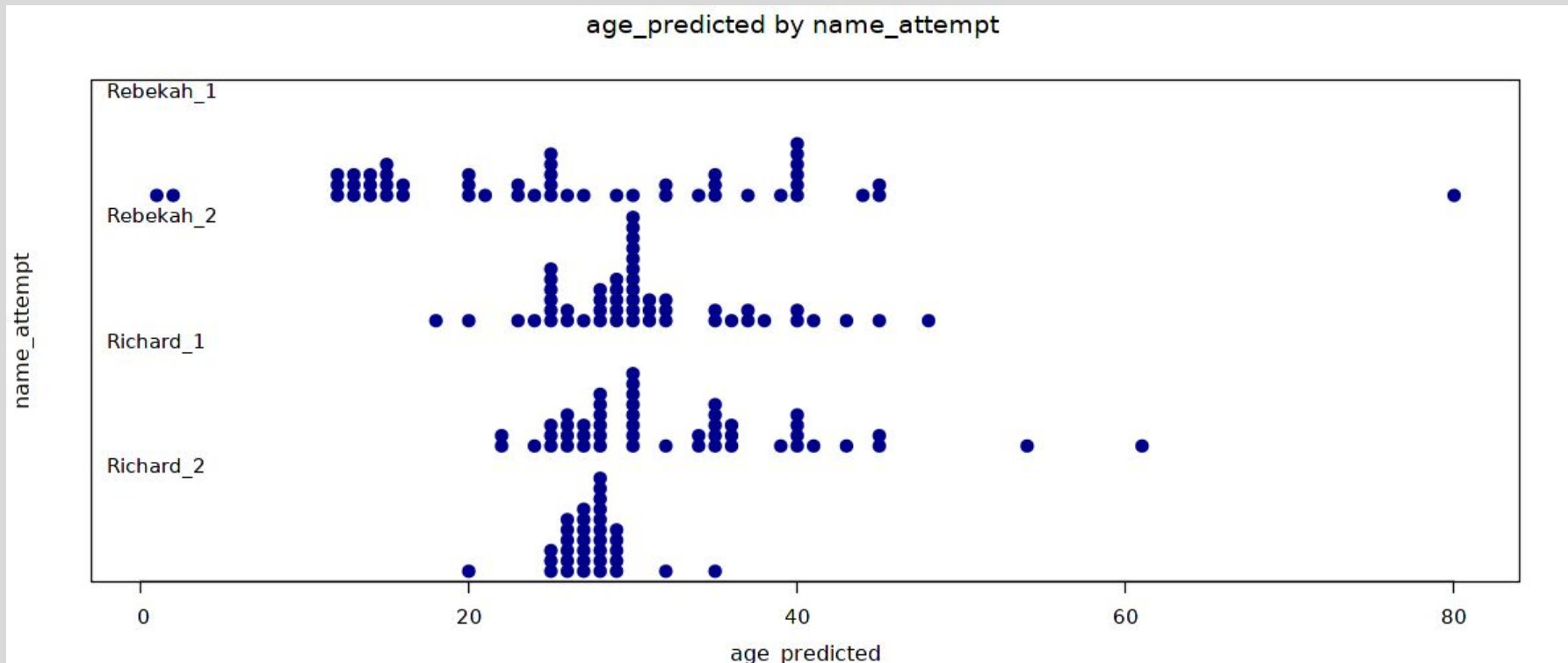


Predicted 30



Predicted 29

Results from teachers at the workshop



Core features of the activity

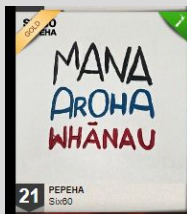
1. Ask students to predict the number value for something, relying only on existing knowledge
2. Ask for values *and reasons*, compare both, make the point that modelling is about turning reasons into structure that can be reused
3. Show data that *might* help students make a “better” prediction
4. Ask for values and reasons again, with reasons *based on reasoning with features of data*
5. Predictions made using second approach typically end up being more similar to each other
6. Compare predictions to actual value, discuss ideas of how to measure “getting it right” e.g. the exact number, or some measure of closeness?
7. Plant “seeds” of ideas for using intervals

Further down the track e.g. CL7/8, models can be “scored” in terms of (1) percentage correct (e.g. classification, PCC) or (2) how close (e.g. prediction, RMSE)

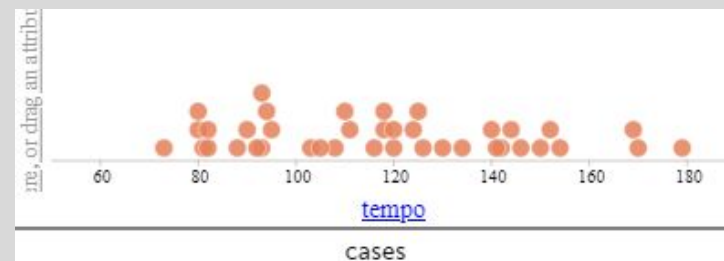
Same activity, different data context

Predict the beats per minute (tempo) for this song

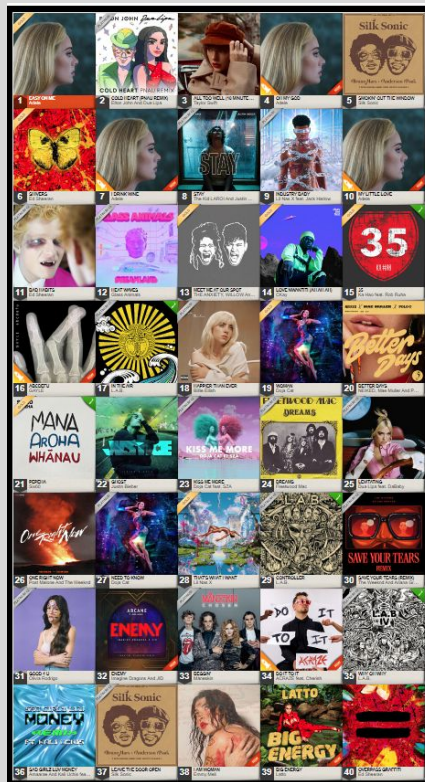
Actual tempo:
62 beats per
minute



Predict the beats per minute (tempo) for this song after seeing plot of tempos of other songs on the Top 40



Predict the beats per minute (tempo) for this song listening to the first 10 seconds



Limitations algorithm/humans



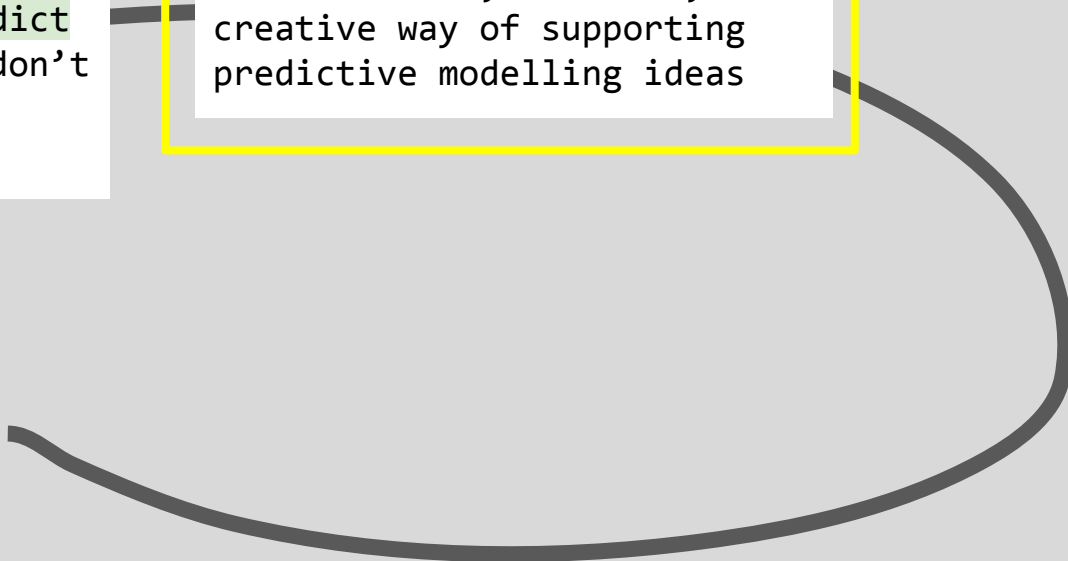
SIX60
THE GREATEST

Use a Spotify playlist
<https://open.spotify.com/playlist/3lBqBc4gEUcWpsfM5Ke8aF> +
 an online app
<http://sortyourmusic.playlistmachinery.com/> to get audio
 features of songs

But you do want to know more about the purpose of the Spotify algorithm, how the model was trained, etc. before using this source of data :-) Spoiler alert: It's not suitable for exploring relationships between variables!

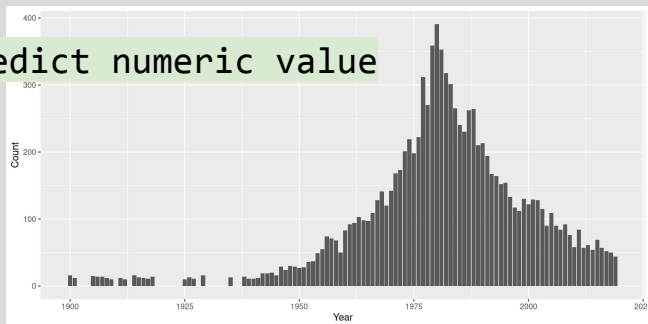
1. A key statistical modelling approach is to use the data you have to predict an outcome you don't (yet) have.

2. Prediction is a broad core idea and exploration provides an accessible, flexible, creative way of supporting predictive modelling ideas

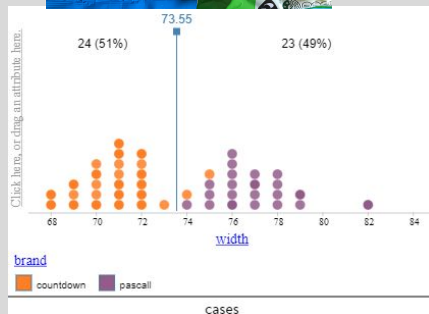


Prediction is a very large bucket of ideas!

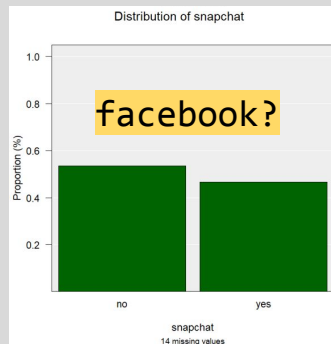
Predict numeric value



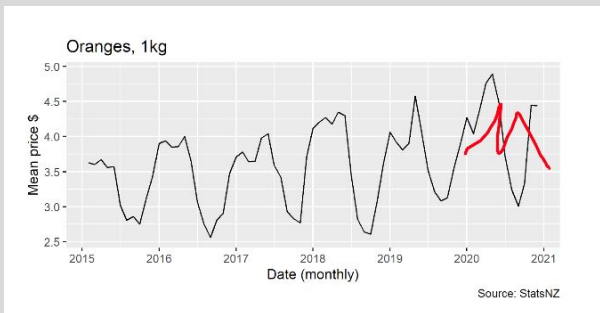
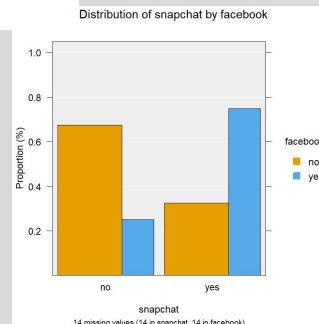
Predict categorical value



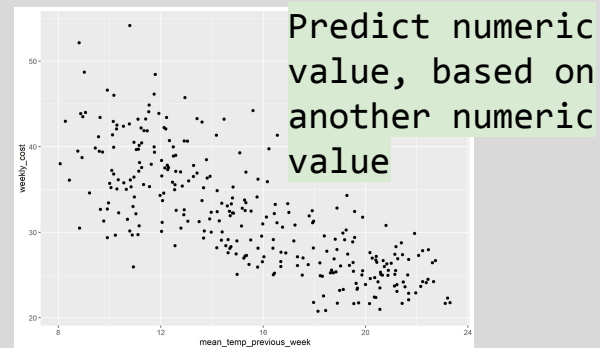
Predict categorical value, based on another numeric value



Predict categorical value, based on another categorical value



Predict numeric value, taking into account date/time value



Predict numeric value, based on another numeric value

Data challenge with prediction focus!

NOVEMBER 30, 2015 BY ANNA FERGUSSON (MARTIN)

Using data challenges to encourage statistical thinking

This post provides the notes for a workshop I ran at the Christchurch Mathematical Association (CMA) and Auckland Mathematical Association (AMA) 2015 Statistics Days about using data challenges to encourage statistical thinking.



What are data challenges?

DataFest
The American Statistical Association (ASA) DataFest is a celebration of data in which teams of undergraduate students work around the clock to find and share meaning in a large, real, and complex data set.

New Zealand's Next Top Engineering Scientist
Has your team got what it takes to win \$5000?

What is a data challenge? These are just the words I am using to describe a competition that involves (big) data. Some good examples are the [ASA datafest](#) and the [Hudson Data Jam](#). Closer to home (NZ) we have the [ISLP statistical poster competition](#) and [New Zealand's Next Top Engineering Scientist](#). The key ingredients are an interesting dataset that can be explored to find stories and the use of presentations or visualisations of this data to tell stories and communicate insight.

What is a statistical poster?
It is a challenge to communicate research in a clear and concise way.

Competitions with data
that involve thinking statistically

2016 Hudson Data Jam Competition
Making other "big" things through creative exploration

How in the first year, the Hudson Data Jam Competition challenges students to creatively tell stories for a general audience using data from the Hudson River watershed.

- Take a large multivariate data set
- Remove a chunk of the rows (cases) from the data
- Set challenges to try to predict a specific value for one of these “missing” cases that require students to learn from the data they do have access to
- Use a software tool so that these predictions are based on interpreting visual features of plots

Famous people and data about their social media accounts (e.g. followers, number of posts, etc.)

Songs and data about different features (e.g. tempo, length, genre of music, age of song, etc.)

Dude, where's my car?

Use the NZ police stolen vehicle database.

Filtered to only have “cars”:

- Sedan
- Hatchback
- Utility
- Stationwagon
- Light van

Stolen vehicles

Check to see if a vehicle is listed as stolen.

Enter a registration number

Rego number

Or VIN, engine or chassis number

Alternative number

☐

I'm not a robot



reCAPTCHA
Privacy • Terms

Search

Limitations of the data

Stolen vehicle information listed here is a snapshot of data taken from the Police vehicle of interest database. Police are unable to guarantee the accuracy of this information. For example, there can be a delay in stolen vehicles appearing and in recovered vehicles being cleared from the list. Some vehicles listed as stolen may have been located but Police haven't been advised.

Stolen vehicle information from these lists should not be relied upon exclusively when assessing whether or not a vehicle might have been stolen. Suggestions for steps prospective purchasers could take to check whether a vehicle is stolen are available from organisations such as Consumer Affairs and Neighbourhood Support.

This data does not include lost or stolen registration plates;

Download List

Note, the download is a zip file of data in CSV (comma separated values) format suitable for a spreadsheet application. If your device is a phone or tablet that does not handle .zip files, the result of attempting this download is unpredictable.

Download a file of stolen vehicles from the past 6 months

- ☐ All of New Zealand
- ☐ Auckland City
- ☐ Bay of Plenty
- ☐ Canterbury
- ☐ Central
- ☐ Counties/Manukau
- ☐ Eastern
- ☐ Northland
- ☐ Southern
- ☐ Tasman
- ☐ Waikato
- ☐ Waitematā
- ☐ Wellington

Download

Persons reporting stolen vehicles

Try searching by...

Lifestyle

Body Style

Affordability



Hatchback



Sedan



Coupe



Convertible



Station wagon



SUV



Start search



- You will be placed into breakout rooms
- Each breakout group has a number
- There is a slide for that group number in the collaborative set of Google slides
- Also on that slide is a link to CODAP that will open up a set of “training data” for you to use for the challenge
- Work through the questions on the Google slide in the order presented
- **Make sure you say what your final prediction is!**
- It helps to have one person in the group share their screen so you have a common thing to talk about
- Have fun!!!

Example of slide for a breakout group

Breakout group 1 *Feel free to edit this slide however you want to show your exploration*

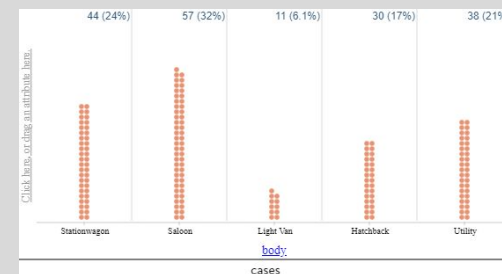
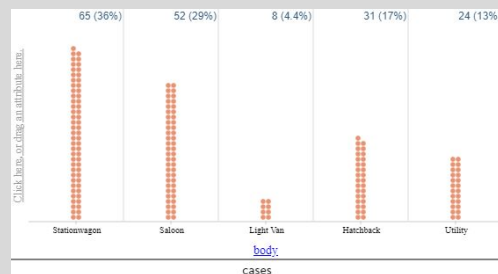
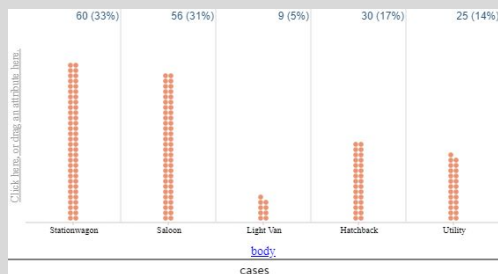
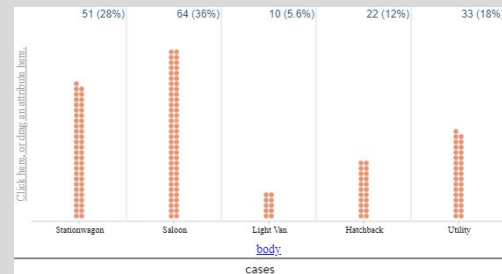
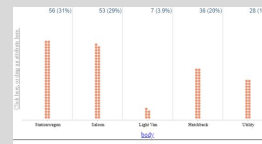
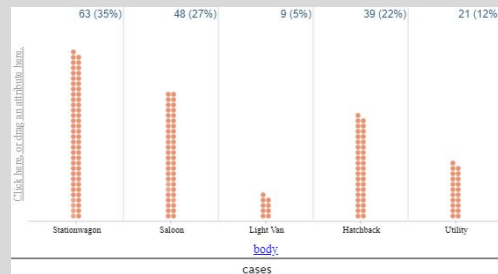
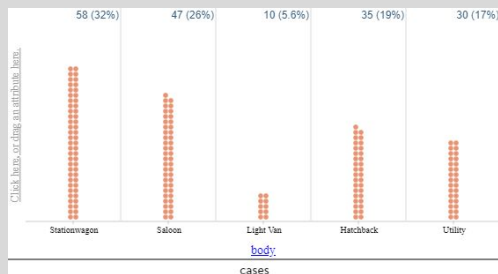
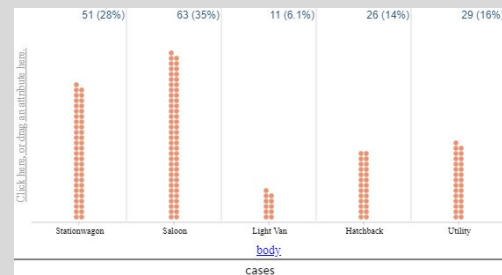
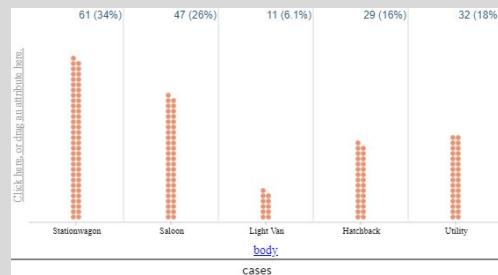
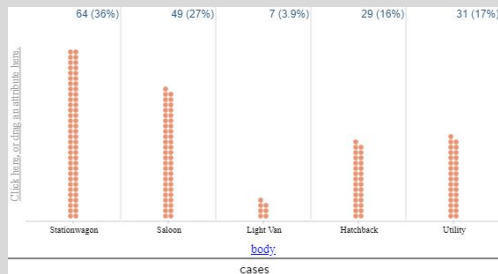
Link to CODAP:

<https://codap.concord.org/releases/latest/static/dg/en/cert/index.html#shared=https%3A%2F%2Fcfm-shared.concord.org%2F3Mx4q1lCUhfSS9EGZrH3%2Ffile.json>

1. What body type do you predict for the stolen car and why? stationwagon
2. What if you knew the age of the car stolen, would this help you predict the body type of the stolen car? Yes
3. What if? Try out some other variables in the data set and see if they could help you predict the body type of the stolen car!
4. Challenge! Predict the body type of the stolen car, given that you know it is/was:
 - Stolen in Wellington
 - 21 years old
 - 35 days since it was reported stolen
 - Mitsubishi
 - Green
 - License plate starts with Z

We predict the body type of the stolen car is: Stationwagon

Same but different ... what do you notice?



So, what was the body type of the car stolen?

Stationwagon



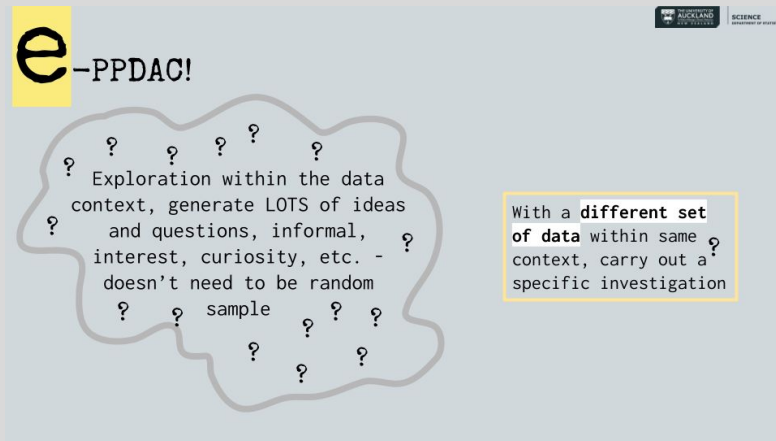
- Colour: Green
- License plate: ZH7355
- Make: Mitsubishi
- Year: 2000 (Age 21)
- Days since reported: 35
(23/10/2021)
- Region: Wellington

How is this different from sample-to-population inference?

We are not trying to estimate the true value of a population characteristic e.g. median, mean, proportion etc.

We are not posing one specific investigative question about a population

But exploring data can lead to focused investigations



Investigate car parking using Google maps and streetview

Use random coordinates or streets within suburbs to “sample”

Count how many cars parked “in the open”

Compare two suburbs

Need to come up with way to “standardise” counting regions

1. A key statistical modelling approach is to use the data you have to predict an outcome you don't (yet) have.

2. Prediction is a broad core idea and exploration provides an accessible, flexible, creative way of supporting predictive modelling ideas

3. Time series data great intro for weaving storytelling and informal predictive modelling (& for working more closely/personally with digital data + CT)

From cars to buses

In the chat box, enter the number that you think is the closest to the count for how many buses are on the roads in Tāmaki Makaurau right now!



Twitter bot setup by Professor Thomas Lumley (UoA), uses the Auckland Transport API to get the information!

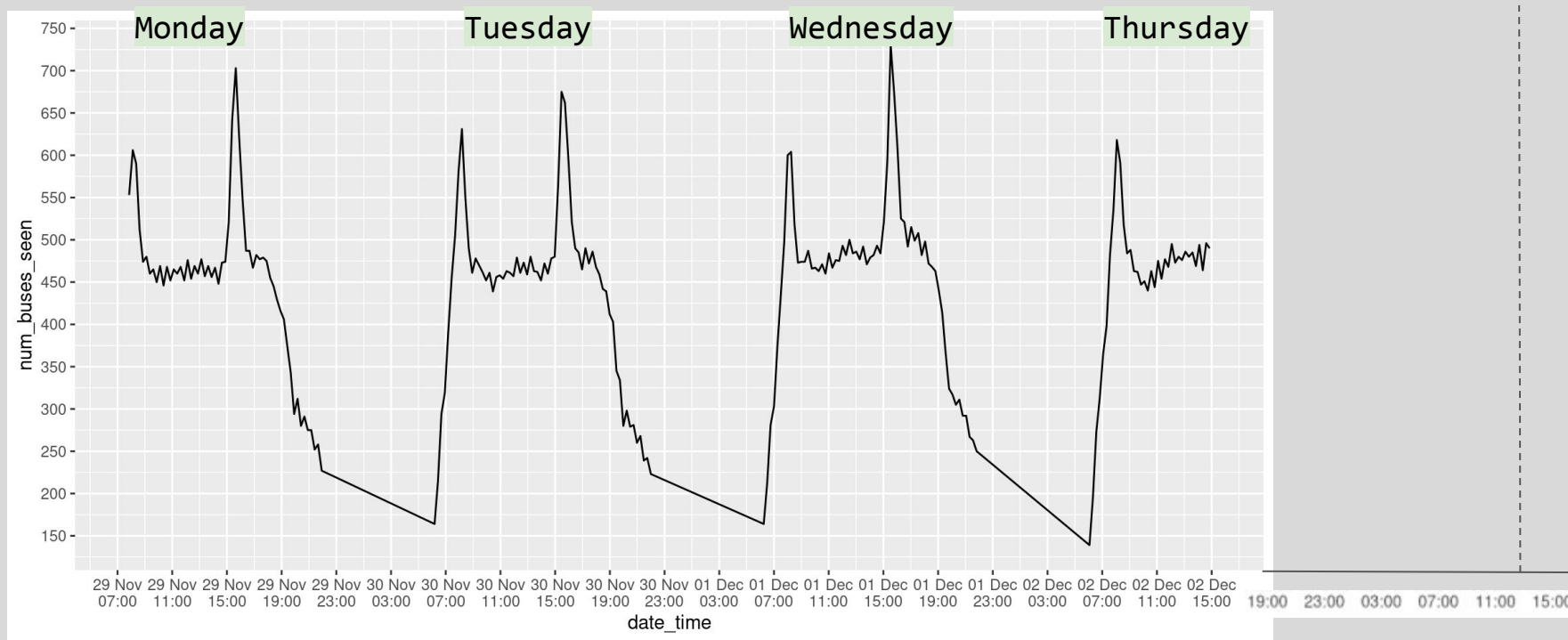


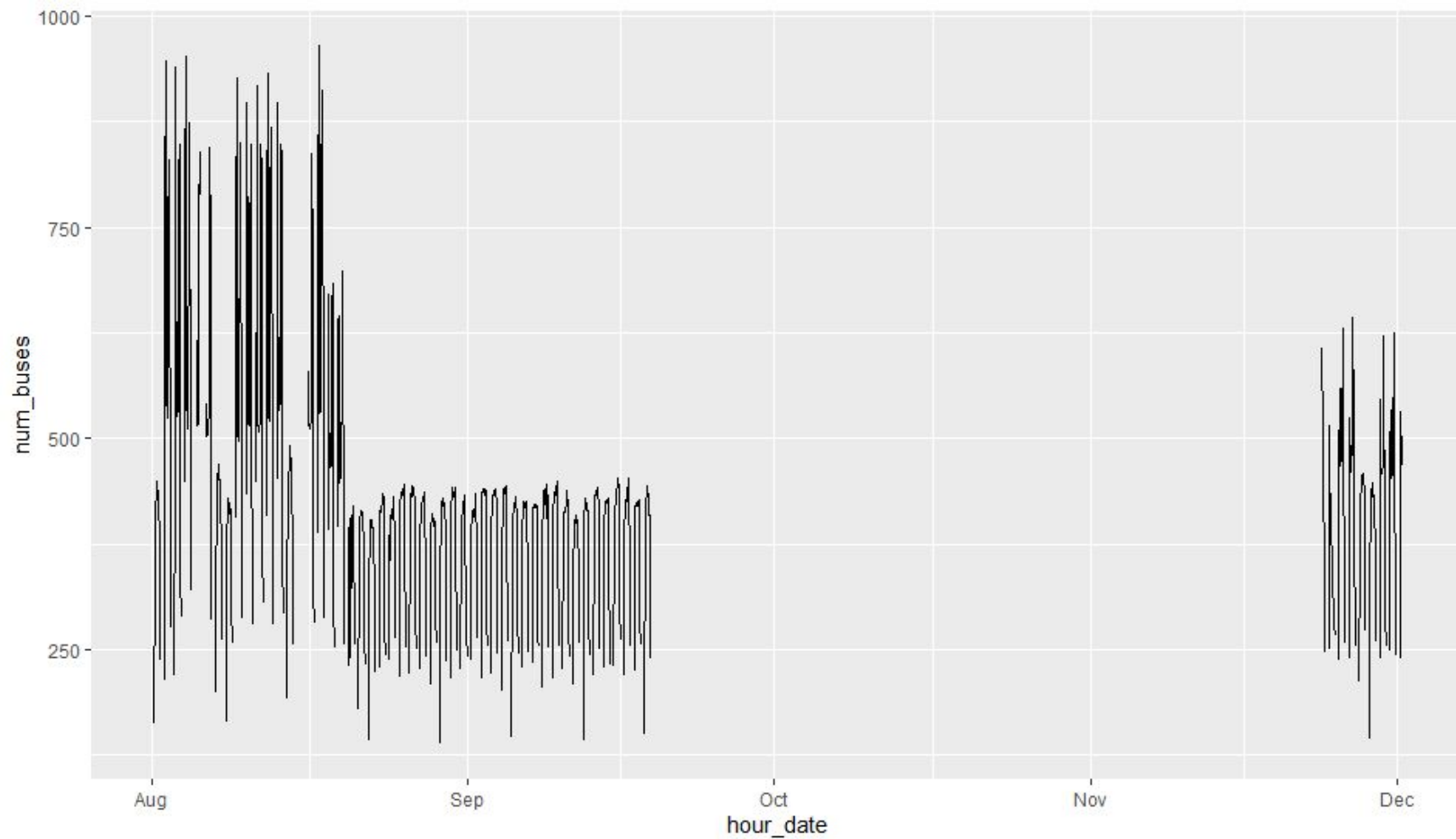
How many buses will be on the road at 1:00pm today?

Make a copy of the Google slide shared.

Use the scribble tool to annotate the time series plot.

Submit your prediction with your name into the Google form - there will be a prize!





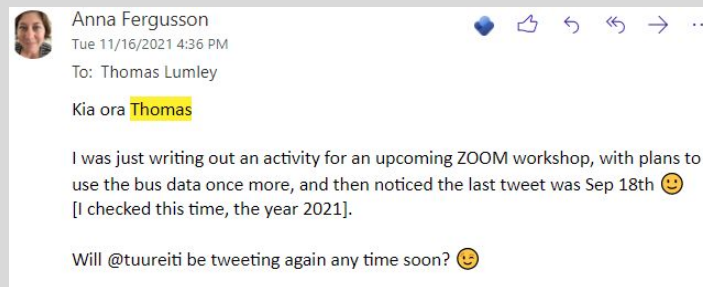
Collecting data from Twitter (AKA set and forget!)

To collect the bus tweets, I set up a “recipe” using the service IFTTT.



Every time a tweet is made by tūreiti @tuureiti, a new row is added to a Google sheet.

I set up up so long ago, I keep forgetting about the data - until workshops like these!



Not a focus within this workshop but ...

- Time series great place to start with prediction, natural storytelling, sketch past patterns into the future
- Also useful for monitoring dynamic systems and exploring a variety of visualisations
- Students can collecting/using own data, introduce some digital tech data skills
- Includes dynamic sources of data not just static spreadsheets

AUGUST 23, 2021

Explorations in variation



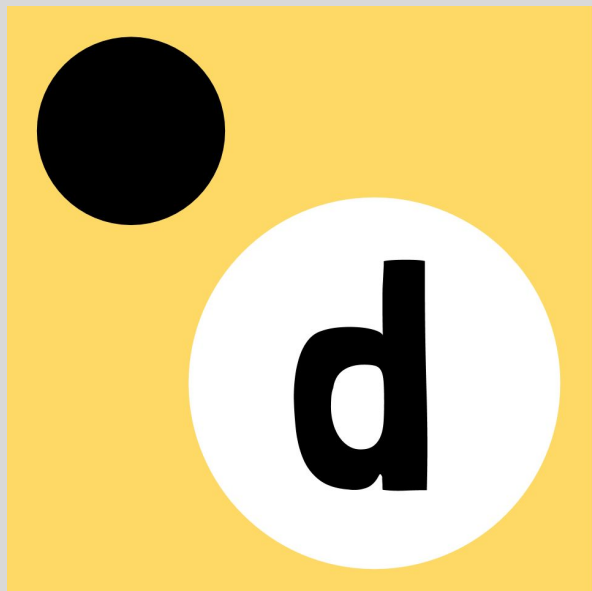
A practical introduction to data science

ExploRing time series data



In this two-hour workshop, you'll get to try out some of the learning tasks from the first week of STATS 100 Concepts in Statistics AKA the Year 13 Statistics course of my dreams! With a focus on modern data contexts and integrating statistical and computational thinking, we'll explore time series data using a range of digital technologies and statistical methods. Come for a fun introduction to data science, including accessing and using a wide range of modern engaging data sources applicable beyond time series; leave with some new ideas and resources for teaching time series. You will need to bring a WIFI-enabled laptop for this workshop.

Something to try out later ...



docactive is a little tool I've been working on

You can try it out here:

<https://docactive.online/G12152FK296/>

1. A key statistical modelling approach is to use the data you have to predict an outcome you don't (yet) have.

2. Prediction is a broad core idea and exploration provides an accessible, flexible, creative way of supporting predictive modelling ideas

3. Time series data great intro for weaving storytelling and informal predictive modelling (& for working more closely/personally with digital data + CT)

4. Different purposes/goals for prediction and different ways of designing data (e.g. experiments)

Akonga seeing themselves in the data



Sarah Carter @mathequalslove · Sep 2

...

We made our first **dot** plots today in statistics to answer the question "How many states have you visited?" Each student designed their own custom "**dot**" that we laminated and added magnets to. Will use the magnets throughout the year.

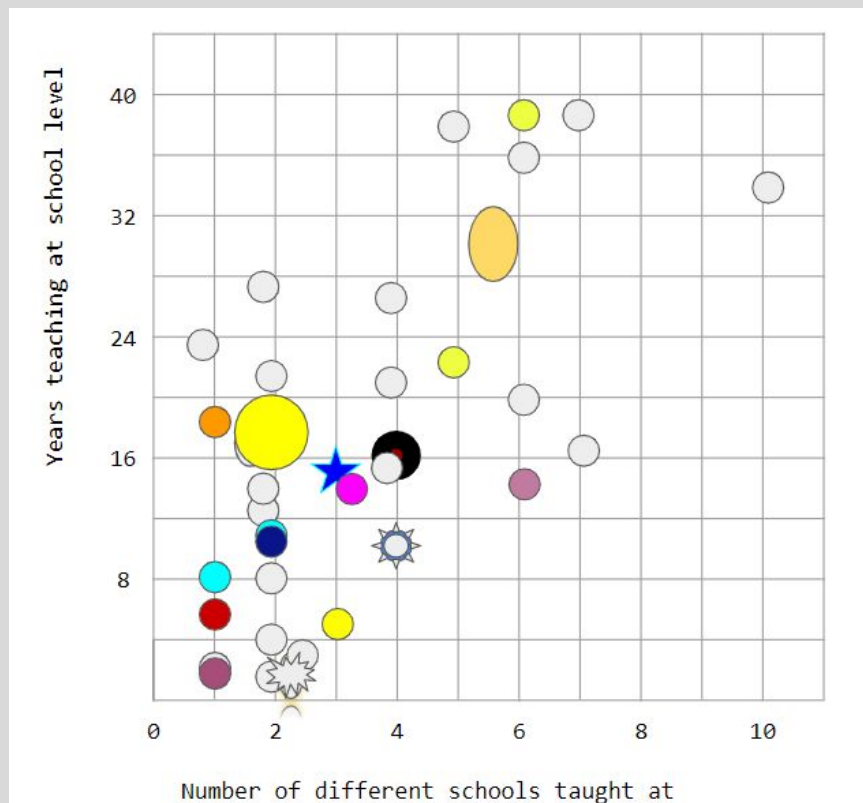
#mtbos #iteachmath #statschat #statchat



Grab a dot and locate yourself in the scatter plot!

Number of years you have taught at school level **vs** Number of schools you have taught at

Our dot plot



Experiment-based data collection

Situations where we can control the value of the x-variable (explanatory) and then measure the value of the y-variable (response).

Often these are “science” based, and the data collected can help to illuminate a known relationship between the two variables e.g., [Hooke's law](#)



We can also get creative with the **mystery box** challenge!

Find the slide that matches your breakout room number and try to design an experiment involving the objects so you can explore the relationship between two numeric variables.

For more “random” objects:
<https://perchance.org/object>

Example of break out room slide

Breakout group 1

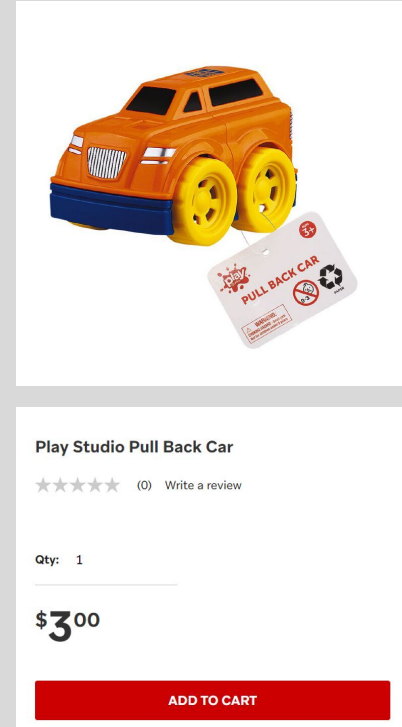


What do you want to explore?

- Time it takes to melt chocolate over candles used / distance
- Time to melt candle wax / volume of wax
- Time to boil water over candles / number / distance
- Number of popsicle sticks to be absorbed water volume / Time

Other ideas

- Pulling a wind-up car back a certain distance and then recording how far it travels
 - Trying this out on different surfaces
 - Trying this out with different cars
 - Taking into account the car doesn't always travel in a straight line
- 'Pinging' a rubber band - stretch length vs. distance it travels
 - Different weight rubber bands
 - Does fatigue (when using the same rubber band) affect results?
- Throwing a tennis ball with a ball thrower held at different places up the handle



1. A key statistical modelling approach is to use the data you have to predict an outcome you don't (yet) have.

2. Prediction is a broad core idea and exploration provides an accessible, flexible, creative way of supporting predictive modelling ideas

3. Time series data great intro for weaving storytelling and informal predictive modelling (& for working more closely/personally with digital data + CT)

5. Learning from features of scatterplots, using informal methods for prediction, evaluating models in terms of accuracy and precision

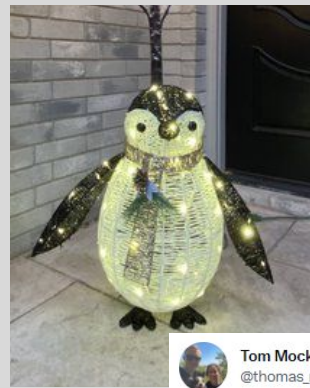
4. Different purposes/goals for prediction and different ways of designing data (e.g. experiments)

Survey-based data collection

Observational in nature, we measure/derive data from cases/individuals/animals “as is”, with no manipulations

Value of prediction model can sometimes be hard to convey, as if you have to ask for/measure one variable, you might as well ask for/measure the other

But still interesting to learn about relationships between variables!



Meet Marvin Palmer Mock, he's here to spread Holiday Cheer!



🍁🍁🍁 **Andrew Heiss** 🍁🍁🍁 @andrewheiss · Nov 28

Replying to @thomas_mock
quick measure its bill depth so i can predict its weight

1 1 5



Tom Mock 🍁🍁🍁 @thomas_mock · Nov 28

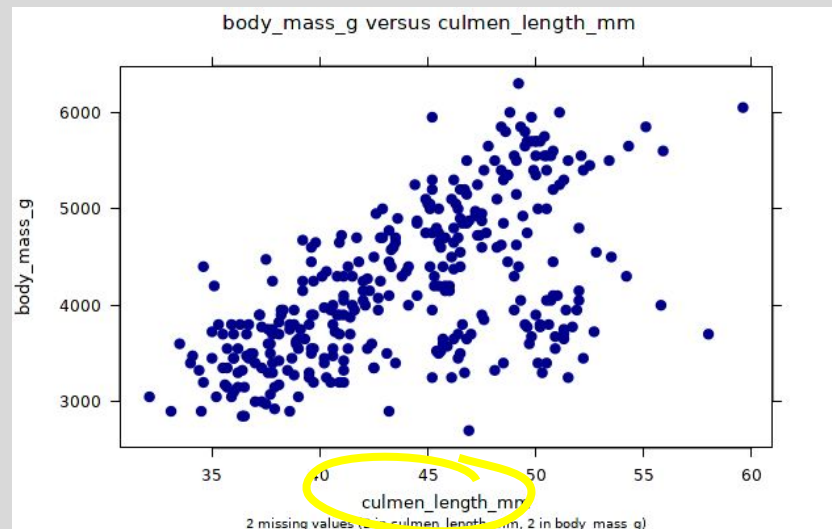
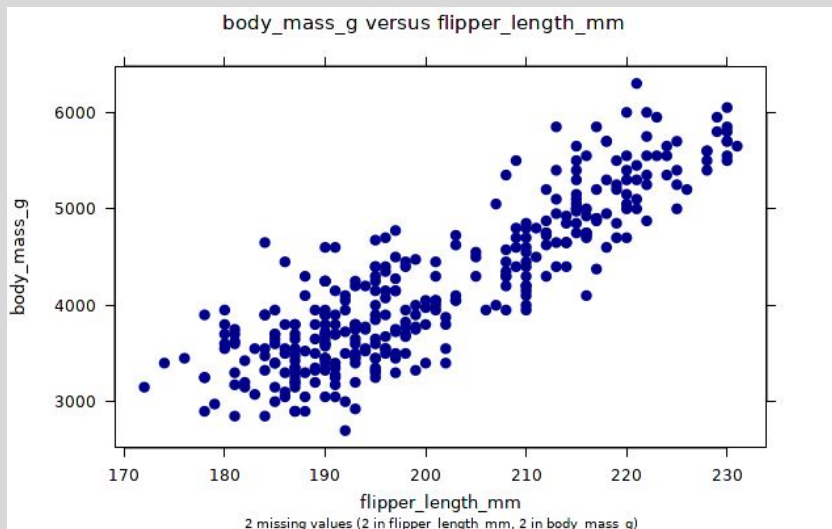
2.21 in beak and 12.36 in flipper - go!

1 1

56mm

314mm

Some data



beak_length

- How many species of penguin do you think are in this data set?
- Would you rather predict weight (body_mass) using flipper length or beak length?

Data source: [Dr. Kristen Gorman and the Palmer Station, Antarctica LTER](#)

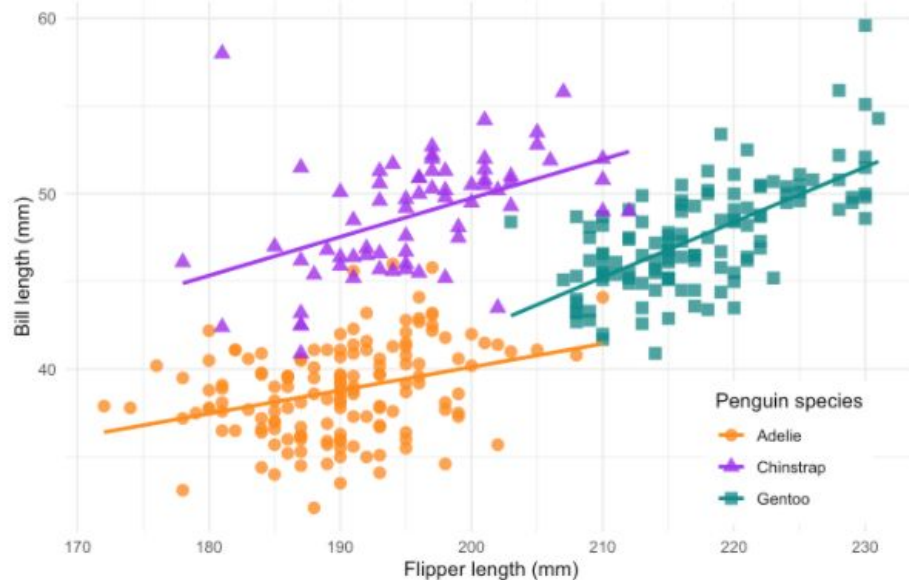
palmerpenguins



The goal of palmerpenguins is to provide a great dataset for data exploration & visualization, as an alternative to `iris`.

Flipper and bill length

Dimensions for Adelie, Chinstrap and Gentoo Penguins at Palmer Station LTER



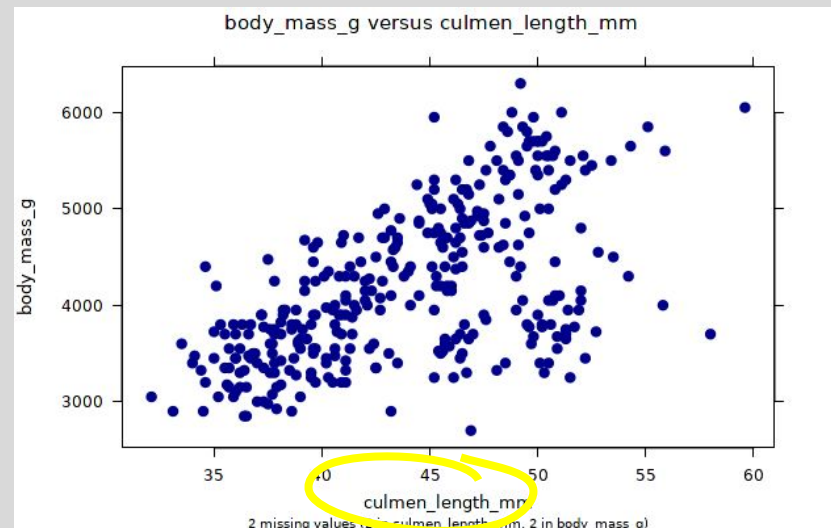
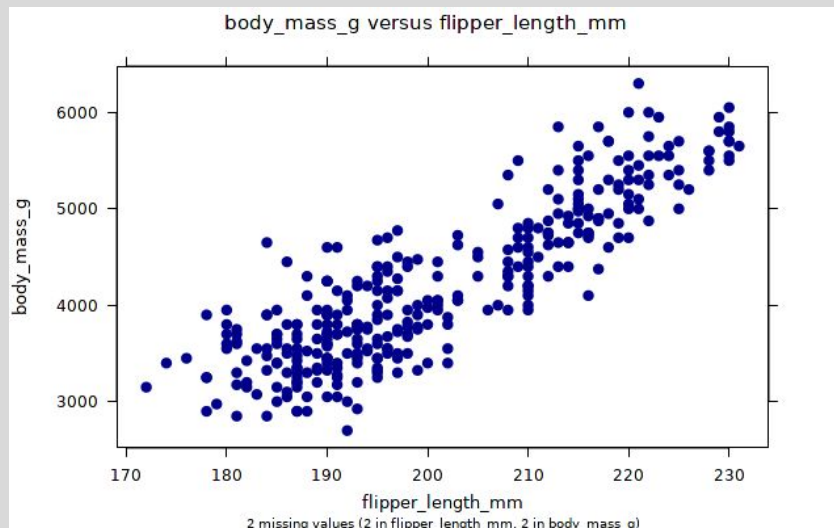
[https://allisonhors
t.github.io/palmerp
enguins/index.html](https://allisonhors.t.github.io/palmerpenguins/index.html)

Some data



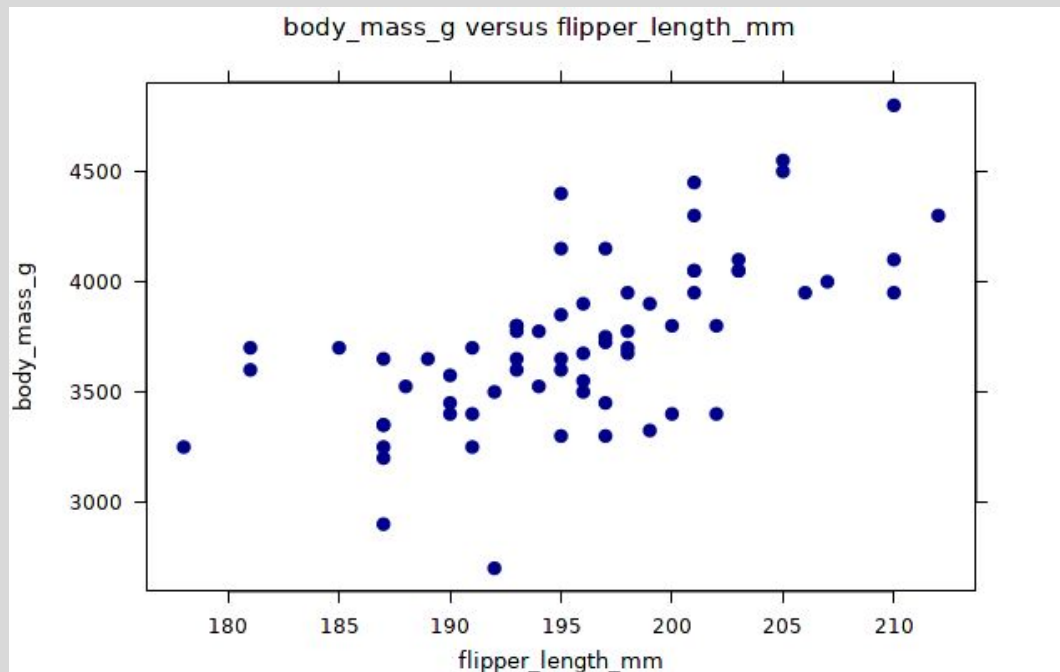
beak_length: 56mm

flipper_length: 314mm



beak_length

A more realistic Chinstrap penguin

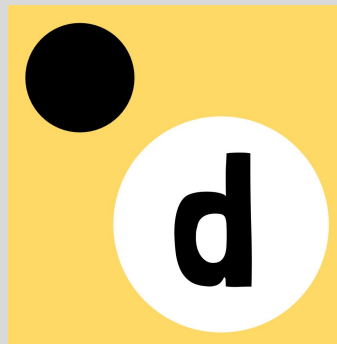


Tom Mock
@thomas_mock


Meet Marvin Palmer Mock, he's here to spread Holiday Cheer!

As a chinstrap penguin, he also wants you to go hang with his remote coworkers, the Palmer Penguins

Number of
strides to
walk 30
metres and
height



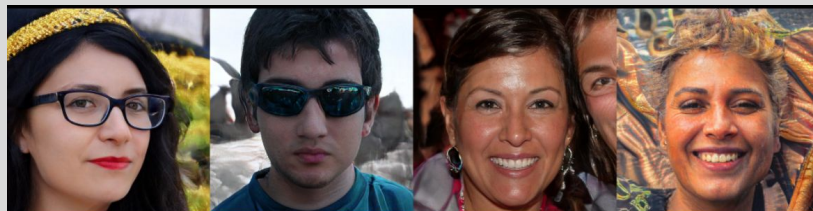
<https://docactive.online/G12Y121QN1/>



Count how many people
are in the crowd!

Crowd counting

I've created
"synthetic crowds"
using AI generated
faces



- In the app, you'll be shown one of these crowds for five seconds
- Scroll down, enter your prediction
- Then press the "show crowd" button and write down the actual "crowd size"
- Scroll down, enter the actual count
- Press submit, then click "Submit another response"
- Repeat three more times!

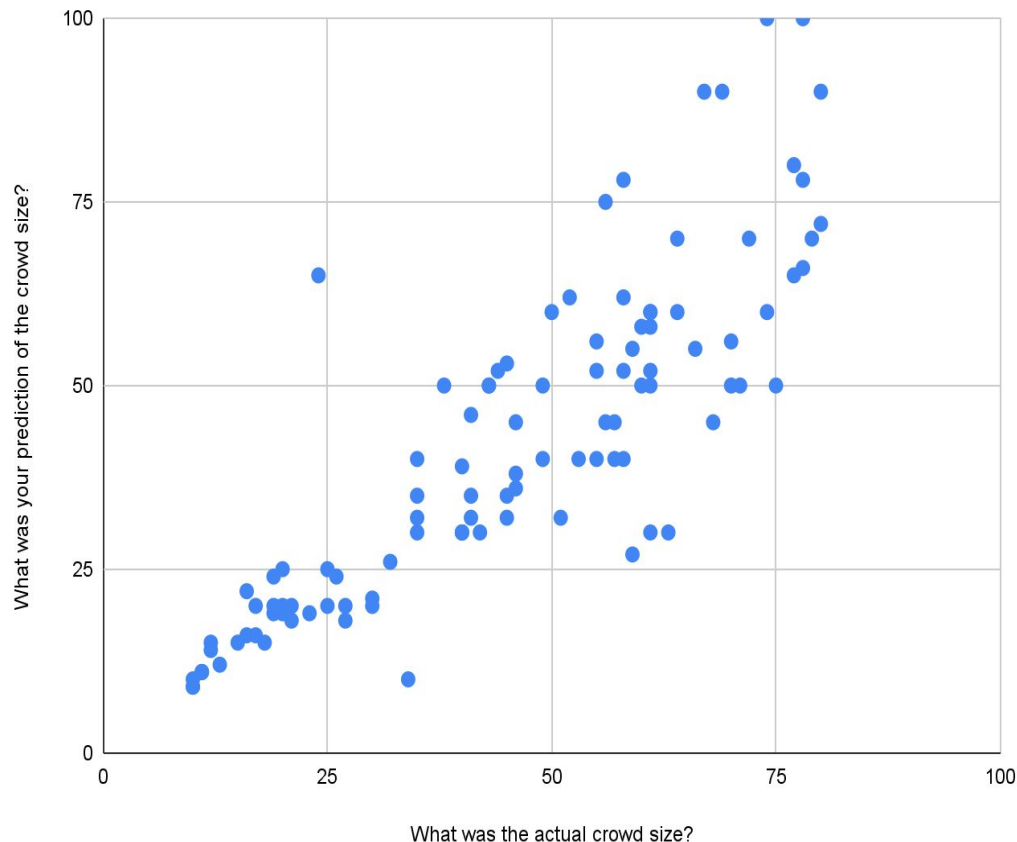
https://learning.statistics-is-awesome.org/crowd_counting/

How did we do?

In this case (not often the case), it makes sense to think about the line $y = x$, as this represents when the predicted counts were the same as the actual counts.

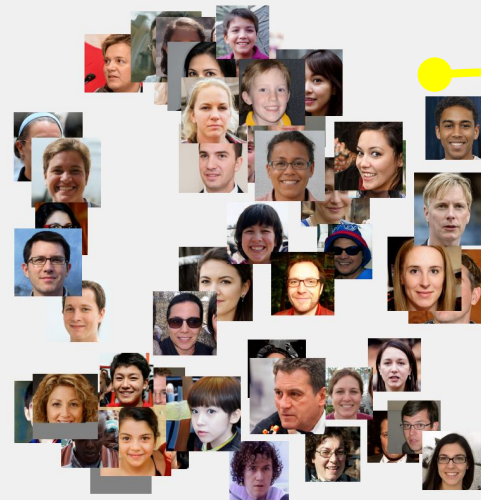
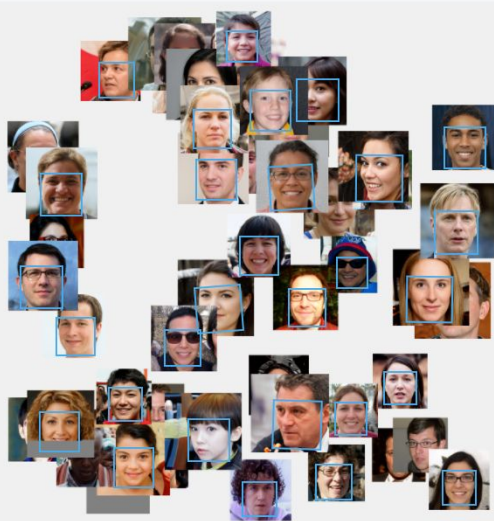
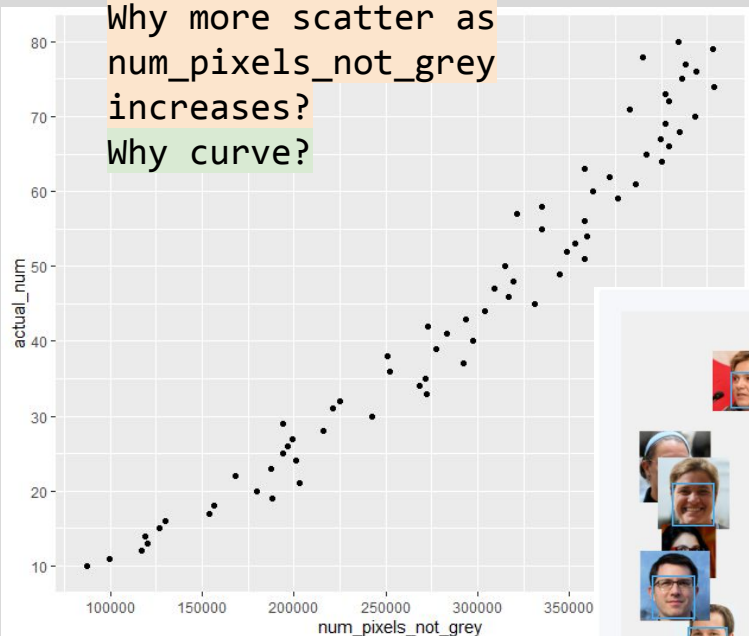
Should be nice example of “more uncertainty” as the values of x increase.

What was your prediction of the crowd size? vs What was the actual crowd size?



Not for this workshop but ...

Why more scatter as
num_pixels_not_grey
increases?
Why curve?



SCIENCE
DEPARTMENT OF STATISTICS

How many
pixels NOT the
background
grey colour?

Results

Response JSON



Supposed to be 52 faces!

29 faces have been detected, each of them got unique face_token, click the face image to check detect results. You can pass these informations to other APIs for further processing. We strongly recommend using Face Attributes and Face Landmark APIs.

Algorithmic bias



<https://www.youtube.com/watch?v=TWwSw1w-BVo>

1. A key statistical modelling approach is to use the data you have to predict an outcome you don't (yet) have.

2. Prediction is a broad core idea and exploration provides an accessible, flexible, creative way of supporting predictive modelling ideas

3. Time series data great intro for weaving storytelling and informal predictive modelling (& for working more closely/personally with digital data + CT)

6. Putting it all together with ePPDAC

5. Learning from features of scatterplots, using informal methods for prediction, evaluating models in terms of accuracy and precision

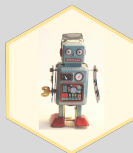
4. Different purposes/goals for prediction and different ways of designing data (e.g. experiments)

Opportunistic data collection

Data sources that can be “tapped into” - maybe it’s data that was collected for a different purpose, maybe it’s data “copied” from a website, data collected via phones or other sensors, maybe it’s data that hasn’t been “extracted” yet e.g. analysing sounds, images, text, etc.



Collecting
data from
YouTube videos



With great data comes great responsibility!

As students are encouraged to interact more closely with data from a wider range of sources, we need to teach about data ethics.

We can draw on the mahi and expertise of Māori scientists and researchers, for example, the [principles of Māori data sovereignty developed by Te Mana Raraunga](#).

How can data collected from the web help us predict ...



... how many words
are in a book?

e-PPDAC!

? ? ? ?
? Exploration within the data
context, generate LOTS of ideas
? and questions, informal,
interest, curiosity, etc. - ?
? doesn't need to be random
? sample ? ?
? ?

With a different set
of data within same ?
context, carry out a
specific investigation

& the bookable task

Exploration

Explore existing data set that has word counts and number of pages for a small set of fiction books written in English

Use a different set of fiction books to explore if the relationship “holds”

Then explore what other variables could be used to predict word count, based on information from Amazon about books

Finish with writing a specific investigative question based on the relationship between word count and another numeric variable

Investigation

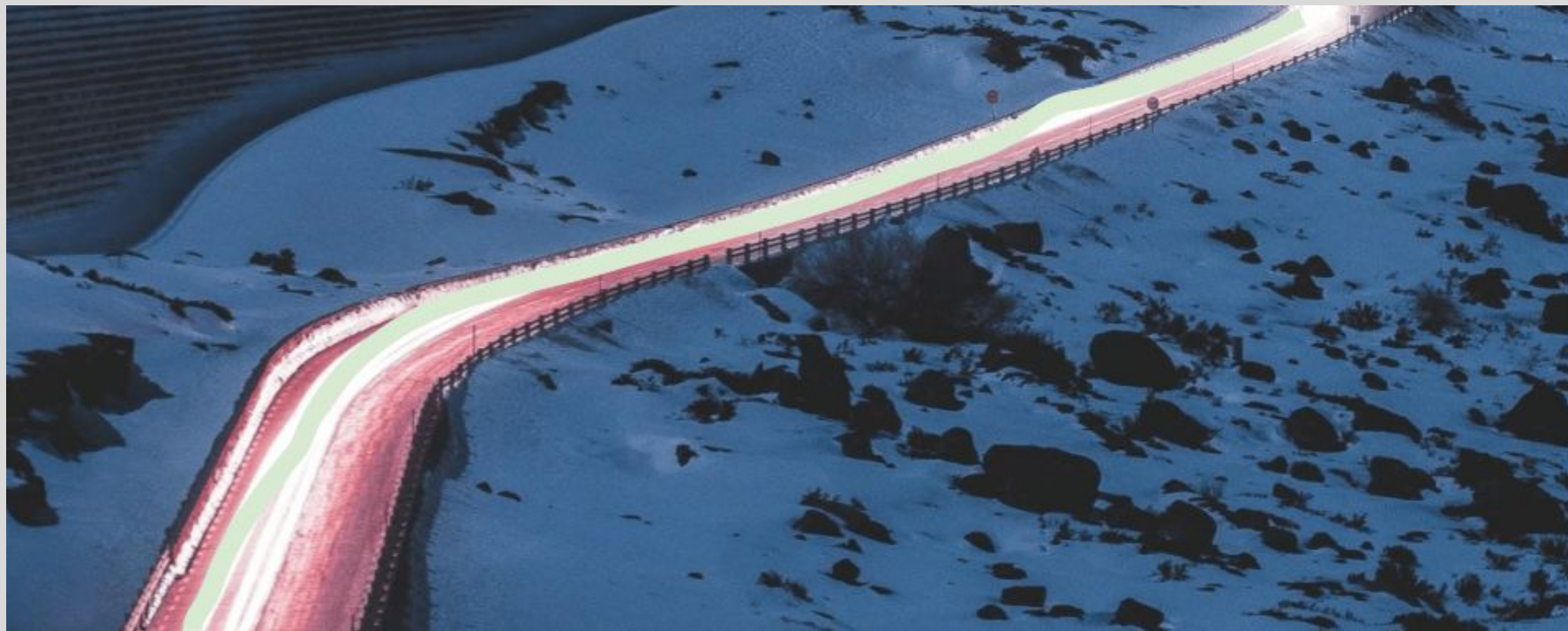
Work in small groups to “crowdsource” additional numeric variables from Amazon

Create a scatter plot with the two variables relevant to your investigative question and describe the features of the data in context

Develop an **informal prediction model** by hand sketching onto the plot and discuss how and how well the model will work to make predictions

Try out the **informal prediction model** with books NOT YET SEEN, predict word count for these books, and then compare predicted to the actual counts.

Exploring worlds with data



Clare Nelson
Tangaroa College



Jim Davis
Westlake Boys'
High School



SCIENCE
DEPARTMENT OF STATISTICS

Amy Hooper
Cashmere High
School



Hanna Reid
St Peter's
Cambridge



Jacqui Hammond
Ormiston Senior
College



Ash Rambhai
Botany Downs Secondary
College



Marion Steel
Epsom Girls
Grammar



Lisa Mulvey
Selwyn College



Chris Wild
University
of Auckland



Anna Fergusson
University
of Auckland



The prediction team!

