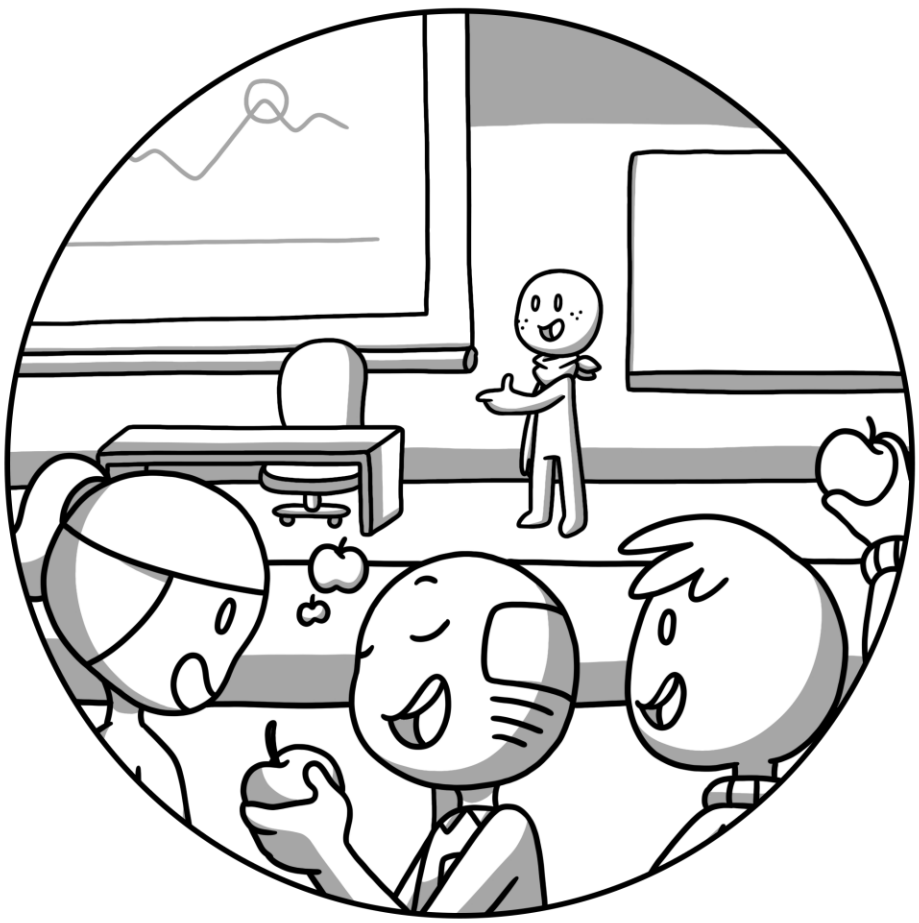
Central Tendency



COMIXPLAIN

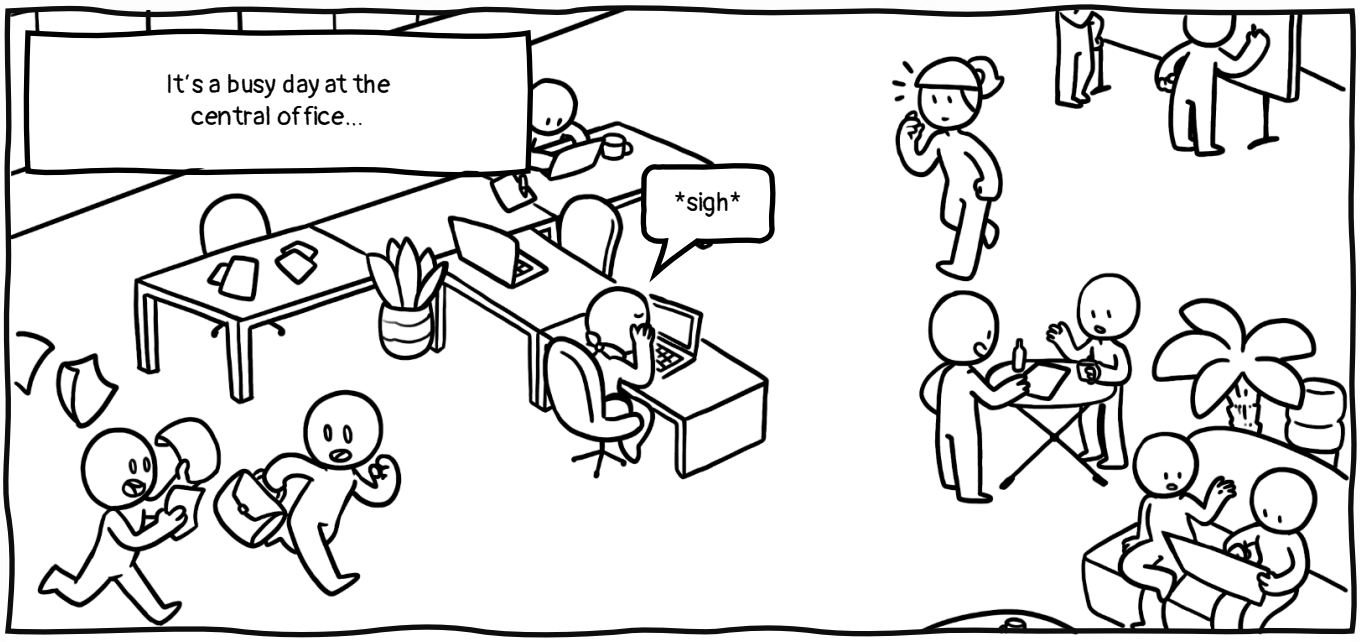
This comic was created in the course of the research project Comixplain, funded by St. Pölten UAS in the course of the Innovation Call 2022.

Project Team:
Victor-Adriel De-Jesus-Oliveira
Hsiang-Yun Wu
Christina Stoiber
Magdalena Boucher
Alena Ertl

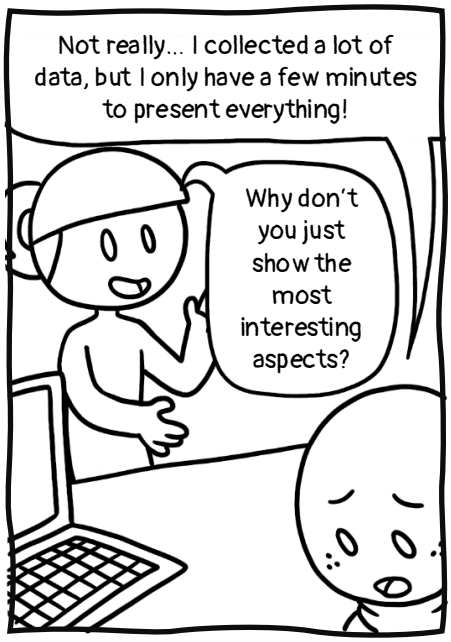
Contact:
victor.oliveira@fhstp.ac.at

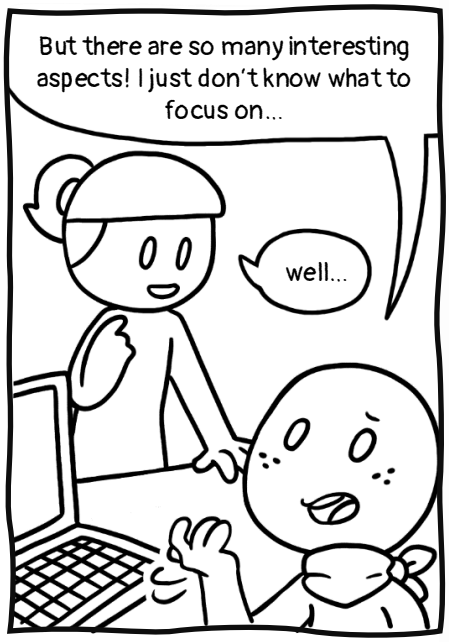
Illustrations:
Magdalena Boucher & Alena Ertl

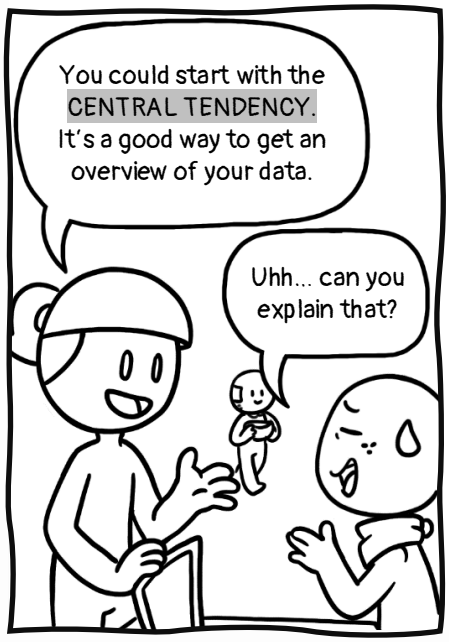
https://fhstp.github.io/comixplain


It's a busy day at the central office...
sigh

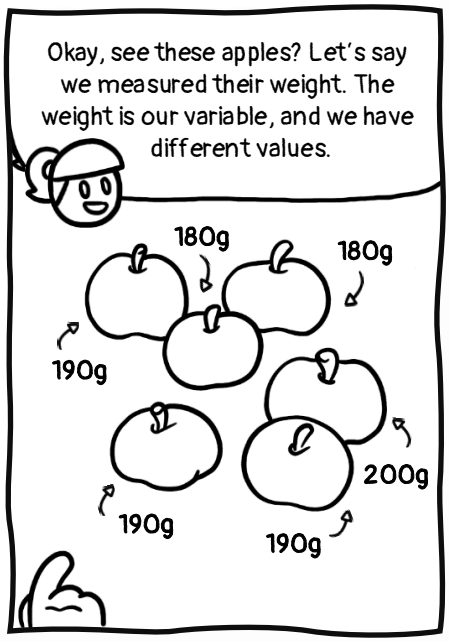

Hey newbie, are you ready for your presentation today?


Not really... I collected a lot of data, but I only have a few minutes to present everything!
Why don't you just show the most interesting aspects?

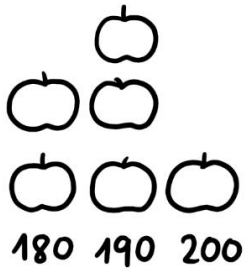

But there are so many interesting aspects! I just don't know what to focus on...
well...


You could start with the CENTRAL TENDENCY. It's a good way to get an overview of your data.
Uhh... can you explain that?

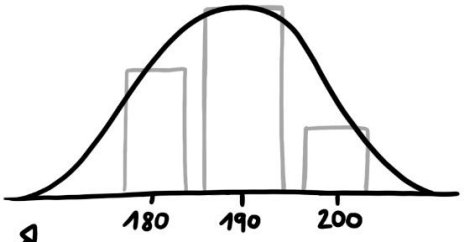

Hey, anybody want some apples?
Oh yes, these will do!


Okay, see these apples? Let's say we measured their weight. The weight is our variable, and we have different values.
180g
180g
190g
200g
190g
190g

The best way to describe a variable is to report the values, and how often each value appears. That's called the DISTRIBUTION of the variable.



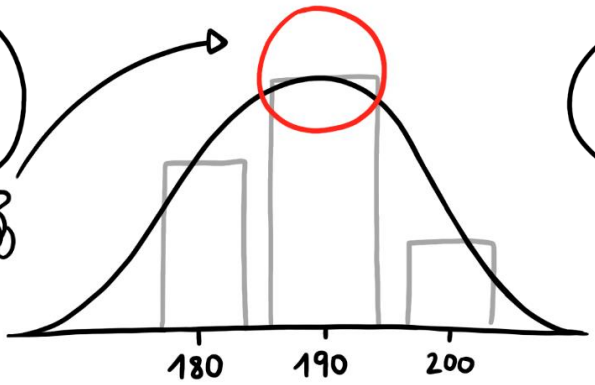
If we visualize the weight of our apples, the distribution would look like this, since all apples in this basket have around the same weight.



Weight of apples

It's kind of shaped like a bell...

Yes. If there were other apples in the basket, our best guess for their weight would be a value around that same spot.



Ah, okay. But... how would I calculate it?

This central point of our weight distribution represents well our data - that's why we call it the CENTRAL TENDENCY.

We can summarize the central tendency in multiple ways. The MEAN value is the most common way, and it's easy to calculate, too.

These are our six apples here:

To calculate the mean, we add all the weight values together...

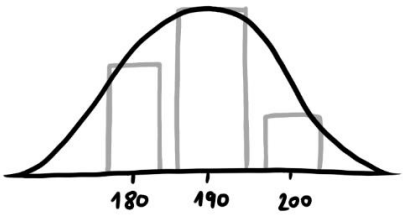
$$200 + 180 + 190 + 190 + 190 + 180$$

...and then divide by the number of apples we have...

$$\frac{(200 + 180 + 190 + 190 + 190 + 180)}{6}$$

That gives us a mean value of

188.3g



Oh that's really almost the top of our bell-shaped graph!

Exactly. But the mean is not always the best way to describe the data.

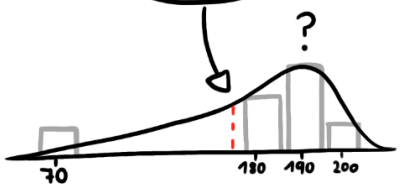
Guys, you won't believe how tiny this apple is!

It weighs, like, only 70 grams!

Once again, perfect example!

If we add the tiny apple to our dataset and calculate the mean...

$$\frac{(200+180+190+190+190+180+70)}{7 \text{ apples}} = 171.4\,g$$

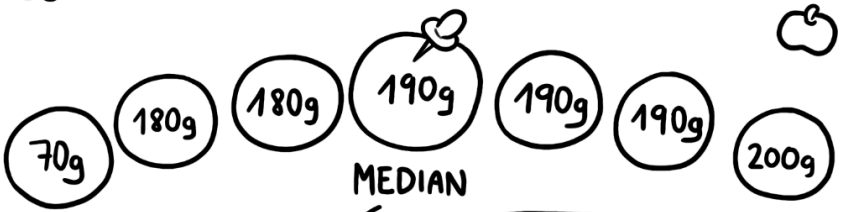


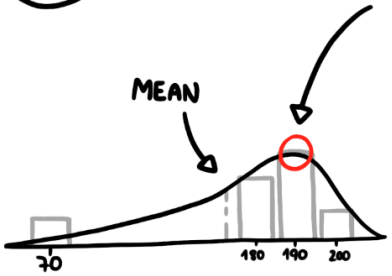
...the graph now looks more like a rat , and the mean is far away from the top, and closer to the tail.

Huh, you're right. Now the mean is not very representative...

True. But you can still describe the central tendency through the MEDIAN.

If we order all our apples by ascending weight, the median is the middle value.





The median here is 190g, which better summarizes this skewed distribution.

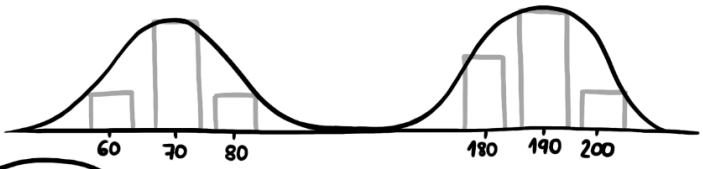
I see you found my tiny apples. Please help yourselves, I'm glad people like them!

Actually, can we borrow that basket for a moment?

There is one more thing to explain: If we add this whole basket of tiny apples to our set, the first tiny apple is not an outlier* anymore.

* Outliers are extreme values that can be errors in measurement, or accurate reports of rare events.

In this case, our graph suddenly has two hills:



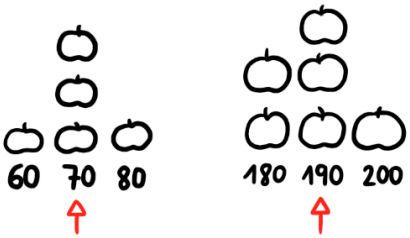
Like two rats in love!

Uhh, sure...

weight of tiny & big apples

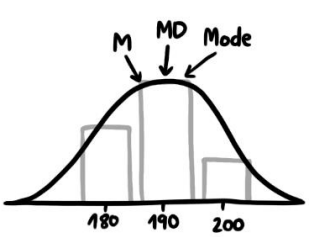
Anyway, we can use something called MODE to describe the central tendency if our distribution has multiple hills.

The mode defines the most frequently occuring value(s) in a dataset.



In this case, we have multiple modes, but there can also be just one, or even no mode at all.

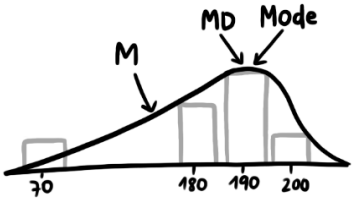
You can apply mean, median, and mode to different samples of apples.
But often, some will represent the data better than others.



180, 180, 190, 190, 190, 200

M = 188.3
MD = 190
Mode = 190

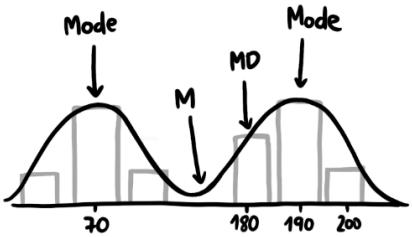
good parameters



70, 180 ,181, 190, 191, 191, 200

M = 171.8
MD = 190
Mode = 191

good parameters



60, 70, 70, 70, 80, 180, 180, 190, 190, 190, 200

M = 134.5
MD = 180
Mode = 70 & 190

good parameter

Okay, thank you... I've learned a lot. Now I just have to apply this to the data I have to present. It's from an app that tracks heart rate measurements.

User ID	Heart Rate (bpm)	Time of Use	User Rating
1	45	13:00	1
2	50	9:00	5
3	55	10:00	3
4	57	9:00	4
5	63	14:00	5
6	70	15:00	5
7	65	16:00	4
8	75	15:00	2

That should be doable - take a look at your data and follow the same steps we just did with the apples! You can use the next page for your calculations.

Before turning the page, try to calculate mean, median and mode for each variable, and check which parameter is most adequate for describing the central tendency. You can take notes on this page!

HEART RATE

Calculating the MEAN:

$$\frac{45+50+55+57+63+70+65+75}{8 \text{ users}} = \frac{480}{8} = 60 \text{ bpm}$$

Calculating the MEDIAN:

45 50 55 | 57 63 | 65 70 75

If there are two central values, the mean of the two values is the median:
(57+63)/2 = 60bpm)

Calculating the MODE:

45, 50, 55, 57, 63, 70, 65, 75

Each value only exists once - there is no mode!

If the distribution of the values is symmetrical, without any distortions, the mean is equal to the median.

MOST FREQUENT TIME OF USE

Calculating the MEAN:

$$\frac{9+9+10+13+14+15+15+16}{8 \text{ users}} = \frac{101}{8} = 12.6?$$

Time of use is not a quantitative value - so calculating mean and median does not make any sense!

Calculating the MEDIAN:

9 9 10 | 13 14 | 15 15 16 = 13.5?

Calculating the MODE:

9:00	10:00	13:00	14:00	15:00	16:00
2x	1x	1x	1x	2x	1x

= 2 modes: 9:00 & 15:00

Mode is not only suited for multimodal distributions, but also when working with ordinal and categorical data.

STAR RATING

Calculating the MEAN:

$$\frac{1+2+3+4+4+5+5+5}{8 \text{ users}} = \frac{29}{8} = 3.6 \text{ stars}$$

Calculating the MEDIAN:

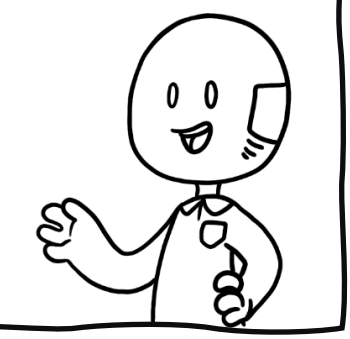
1 2 3 | 4 4 | 5 5 5 = 4 stars

Calculating the MODE:

1	2	3	4	5
1x	1x	1x	2x	3x

= 5 stars

For datasets with a skewed distribution, the median is a better way to describe central tendency.



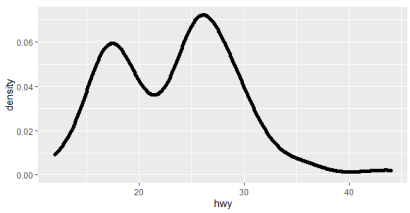
Programming languages like **R**, can help you calculate the central tendency of attributes in large datasets. With R libraries, like **tidyverse**, you can quickly visualize the data distribution.

	model	year	hwy
1	jetta	1999	44
2	corolla	2008	37
3	civic	2008	36
4	civic	2008	36
5	corolla	1999	35
6	altima	2008	32
7	sonata	2008	31
+ other 227 entries			

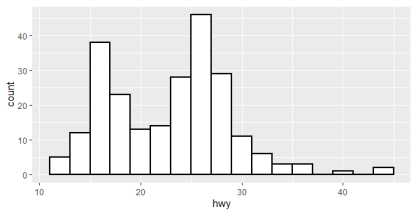
In tidyverse, you have access to datasets such as **mpg** with fuel economy data. It includes 11 attributes, such as car model, year of manufacture, and highway miles per gallon (hwy).

You can use **ggplot**, which is included in tidyverse, to visualize the data distribution of highway miles per gallon (hwy) using a histogram, a density curve, or both.

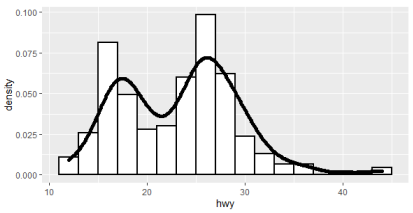
```
install.packages("tidyverse")  # Install it only the first time using the library

library(tidyverse)
plot <- ggplot(mpg, aes(x=hwy))
plot +
 geom_density()
```




```
plot +
 geom_histogram(
  colour="black",
  fill="white")
```



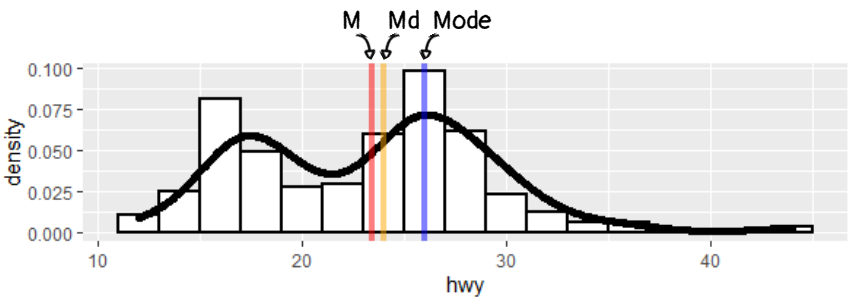

```
plot +
 geom_histogram(aes(y=..density..),
  colour="black",
  fill="white") +
 geom_density()
```



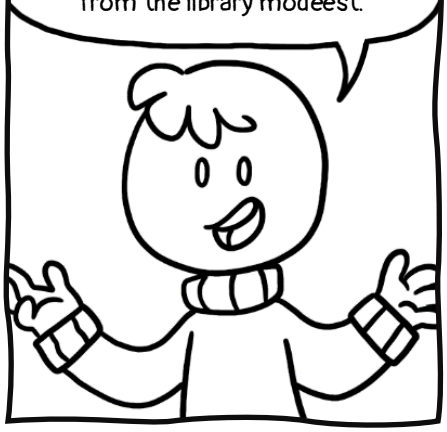
mean(mpg$hwy) = 23.4

median(mpg$hwy) = 24

library(**modeest**)
mlv(mpg$hwy) = 26



R includes native functions to calculate mean and median. For mode, you can build your own function or use the Most Likely Values (mlv) from the library modeest.



Sources:

Downey, A. (2014). Think stats: exploratory data analysis. O'Reilly Media, Inc.

Field, A. (2022). An adventure in statistics: The reality enigma. Sage.