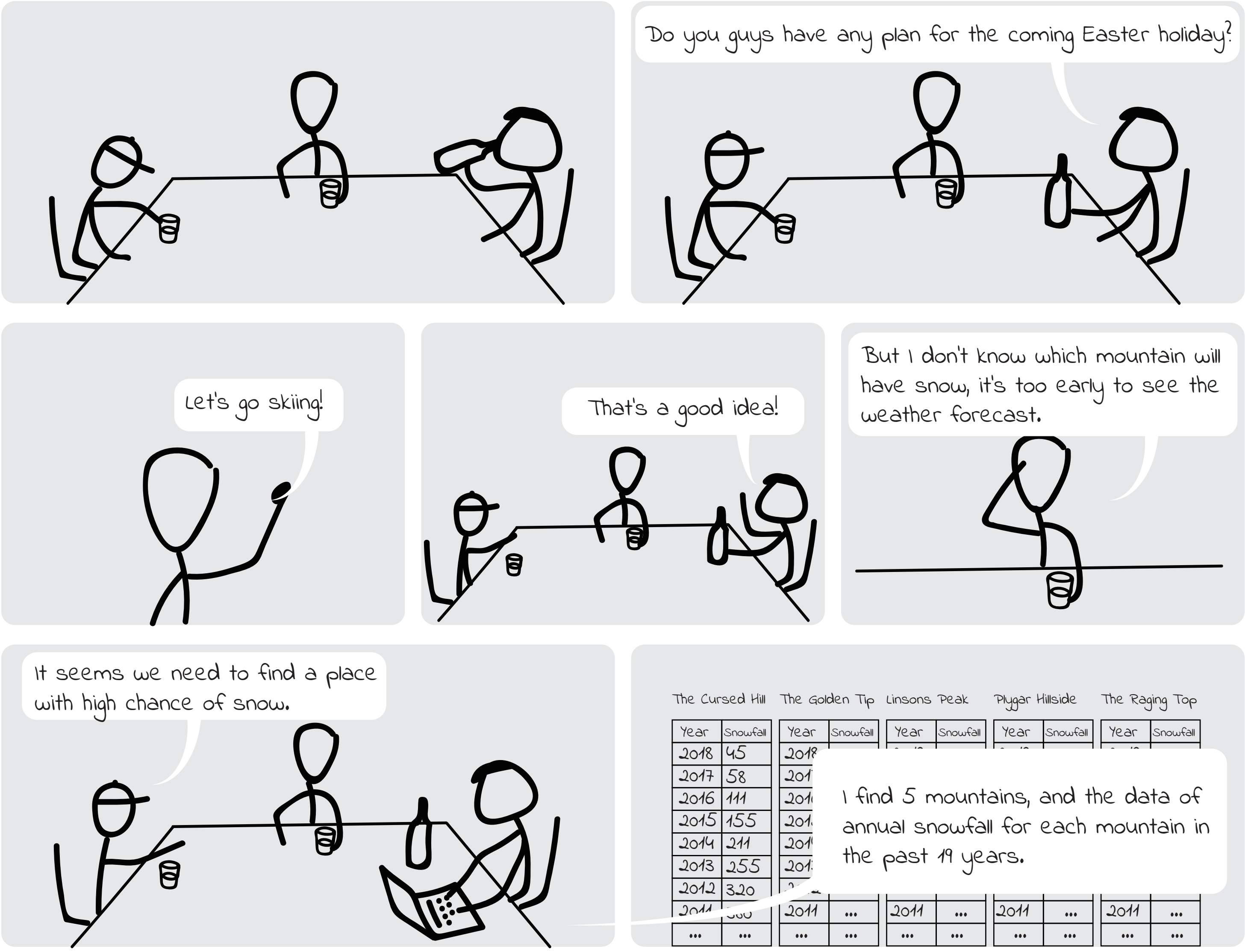
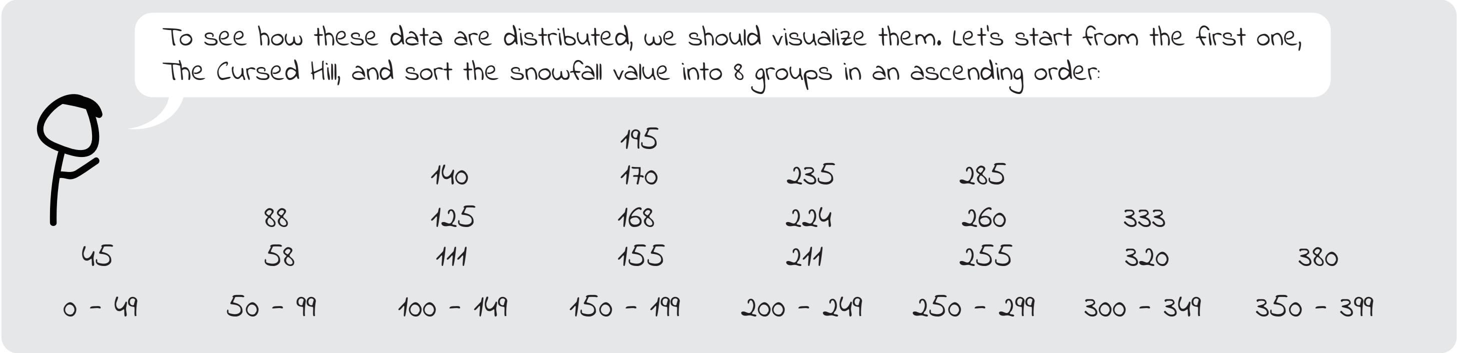
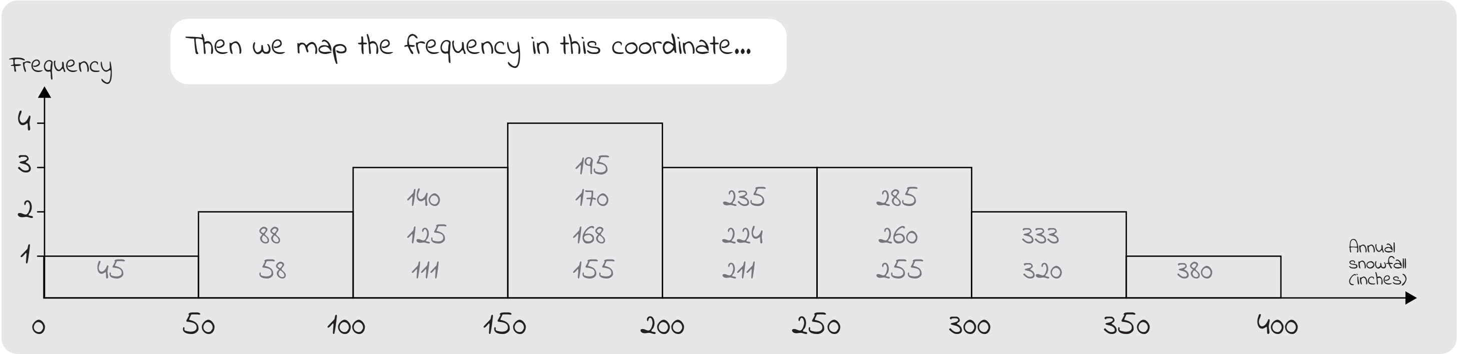
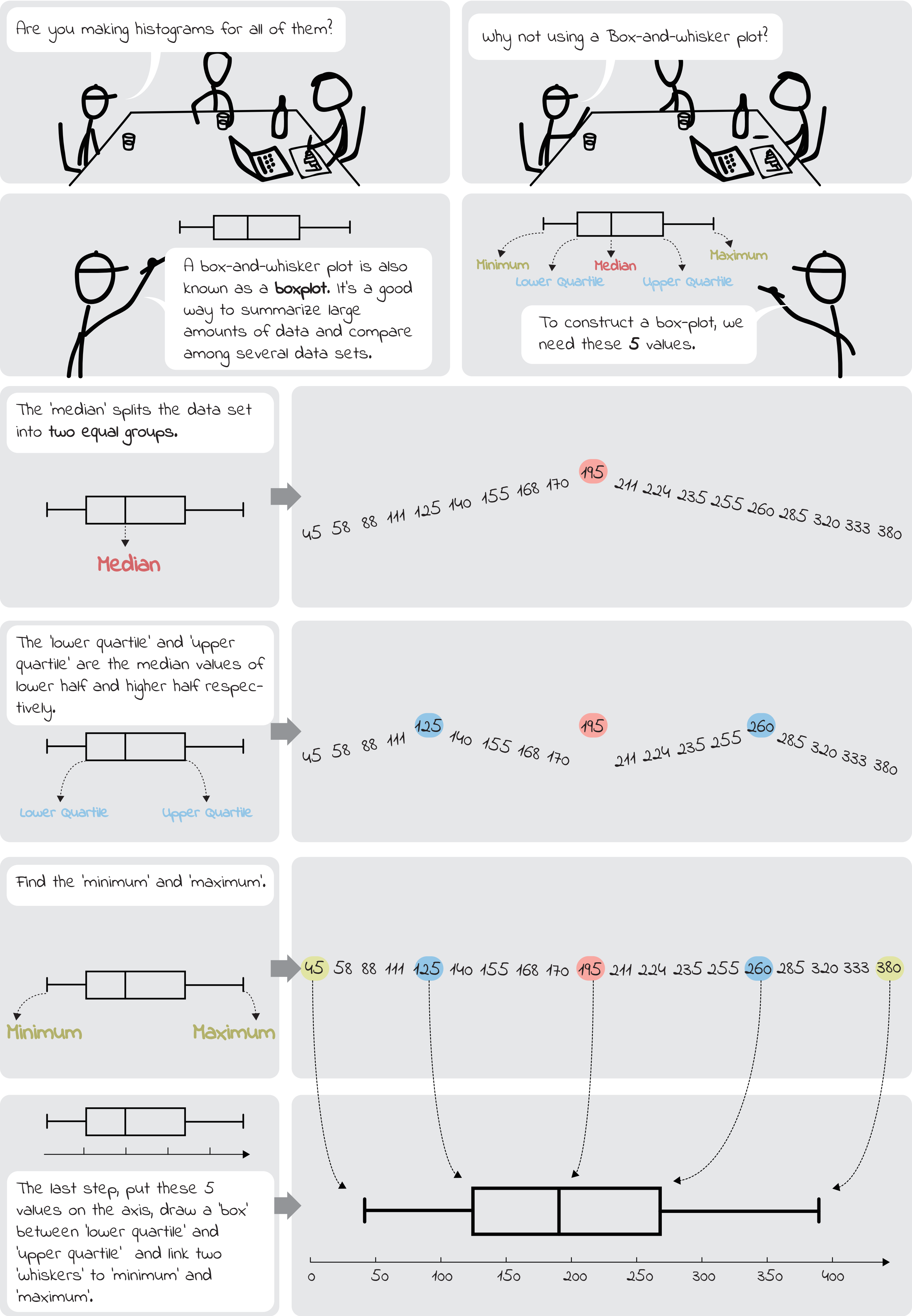
Boxplot

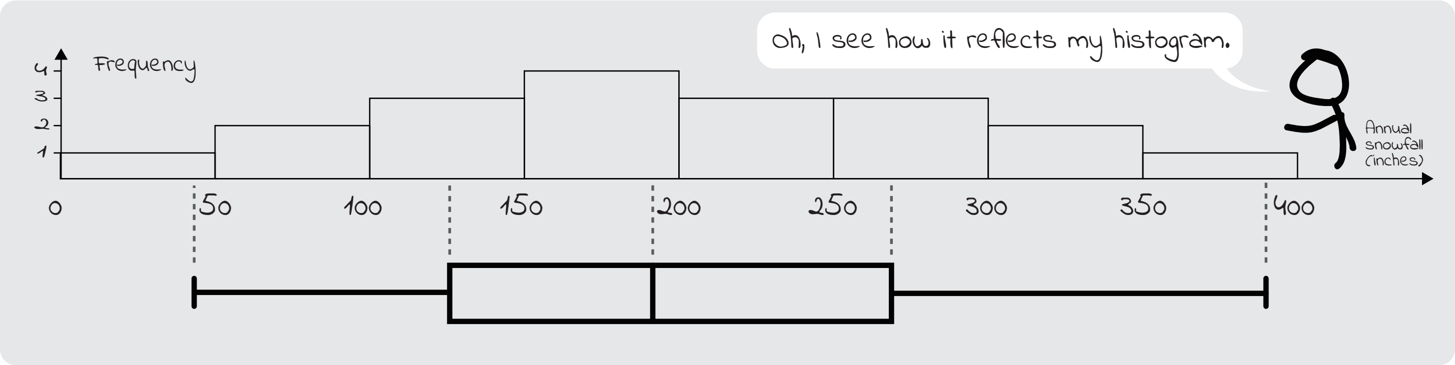
Introduction

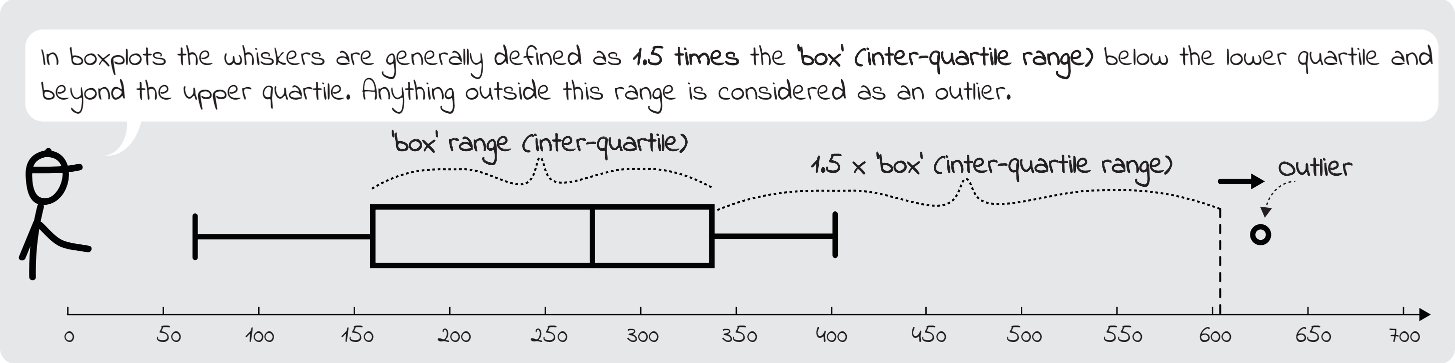

Do you guys have any plan for the coming Easter holiday?
Let's go skiing!
That's a good idea!
But I don't know which mountain will have snow, it's too early to see the weather forecast.
It seems we need to find a place with high chance of snow.
The Cursed Hill
The Golden Tip
Linsons Peak
Plygar Hillside
The Raging Top
Year
Snowfall
2018
45
2017
58
2016
111
2015
155
2014
211
2013
255
2012
320
2011
...
I find 5 mountains, and the data of annual snowfall for each mountain in the past 19 years.

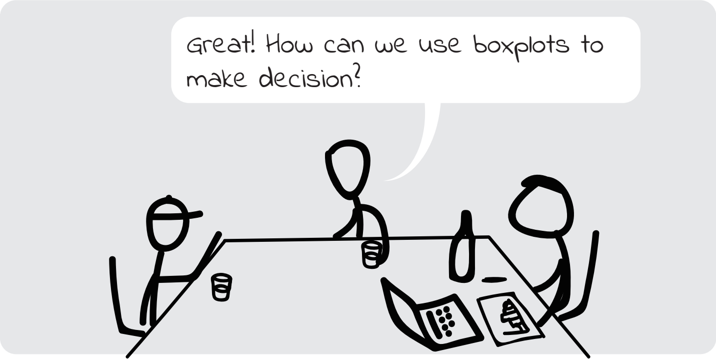

To see how these data are distributed, we should visualize them. Let's start from the first one, The Cursed Hill, and sort the snowfall value into 8 groups in an ascending order:
45
0 - 49
88
58
50 - 99
140
125
111
100 - 149
195
170
168
155
150 - 199
235
224
211
200 - 249
285
260
255
250 - 299
333
320
300 - 349
380
350 - 399

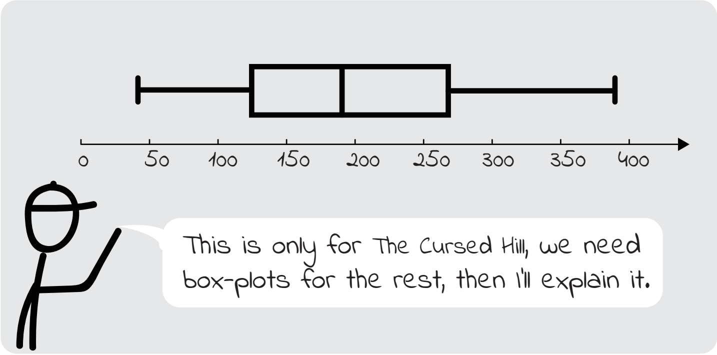


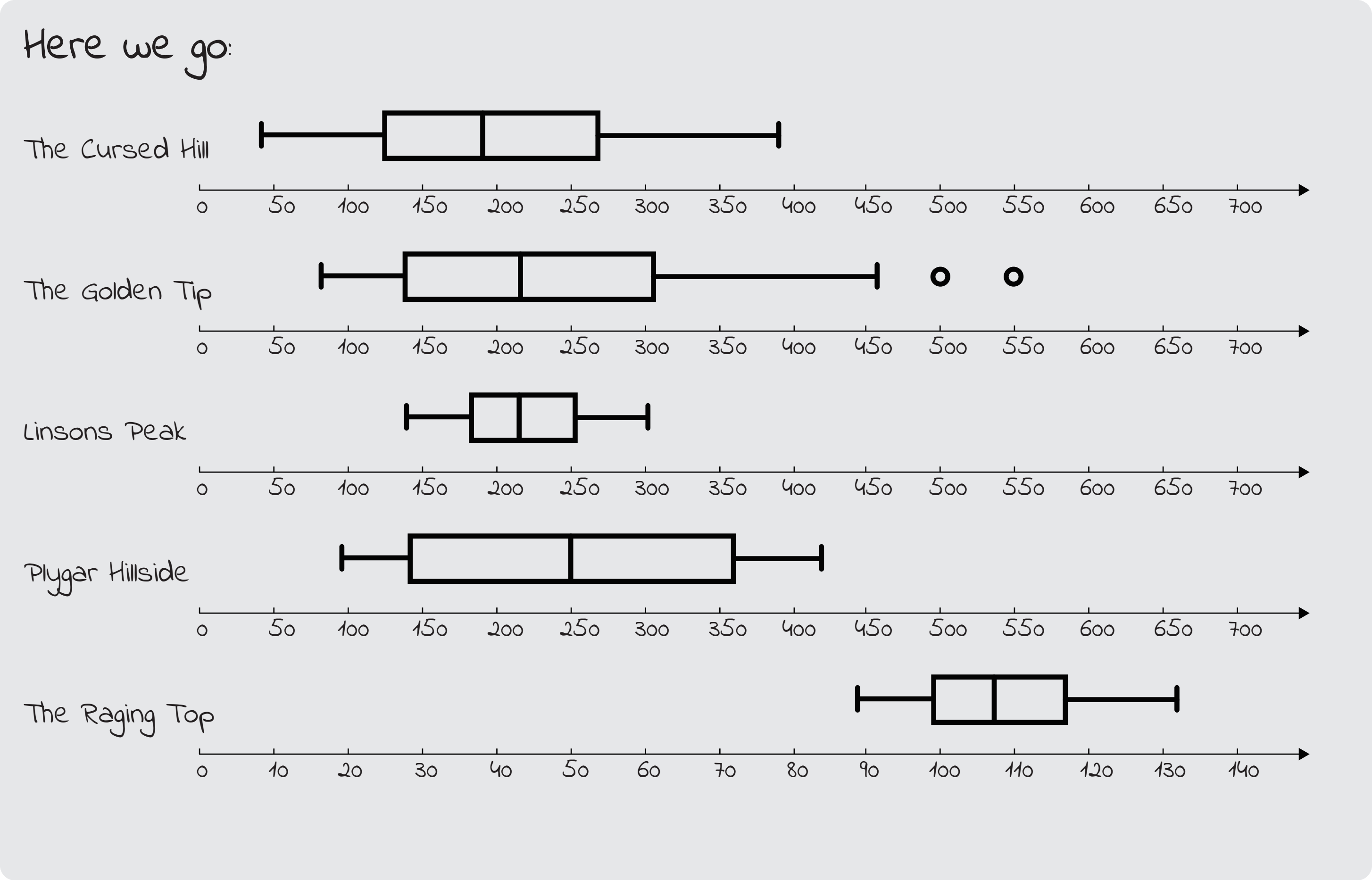

Are you making histograms for all of them?
why not using a Box-and-whisker plot?
A box-and-whisker plot is also known as a **boxplot.** It's a good way to summarize large amounts of data and compare among several data sets.
Minimum
Lower Quartile
Median
Upper Quartile
Maximum
To construct a box-plot, we need these **5** values.
The 'median' splits the data set into **two equal groups.**
Median
45 58 88 111 125 140 155 168 170 195 211 224 235 255 260 285 320 333 380
The 'lower quartile' and 'upper quartile' are the median values of lower half and higher half respectively.
Lower Quartile
Upper Quartile
45 58 88 111 125 140 155 168 170 195 211 224 235 255 260 285 320 333 380
Find the 'minimum' and 'maximum'.
Minimum
Maximum
45 58 88 111 125 140 155 168 170 195 211 224 235 255 260 285 320 333 380
The last step, put these 5 values on the axis, draw a 'box' between 'lower quartile' and 'upper quartile' and link two 'whiskers' to 'minimum' and 'maximum'.
0
50
100
150
200
250
300
350
400

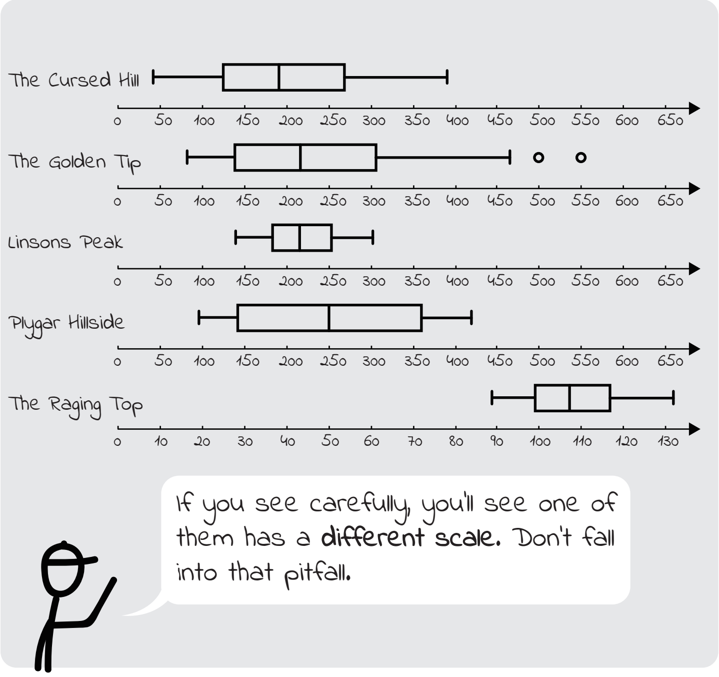

Oh, I see how it reflects my histogram.
Frequency
1
2
3
4
Annual snowfall (inches)
0
50
100
150
200
250
300
350
400

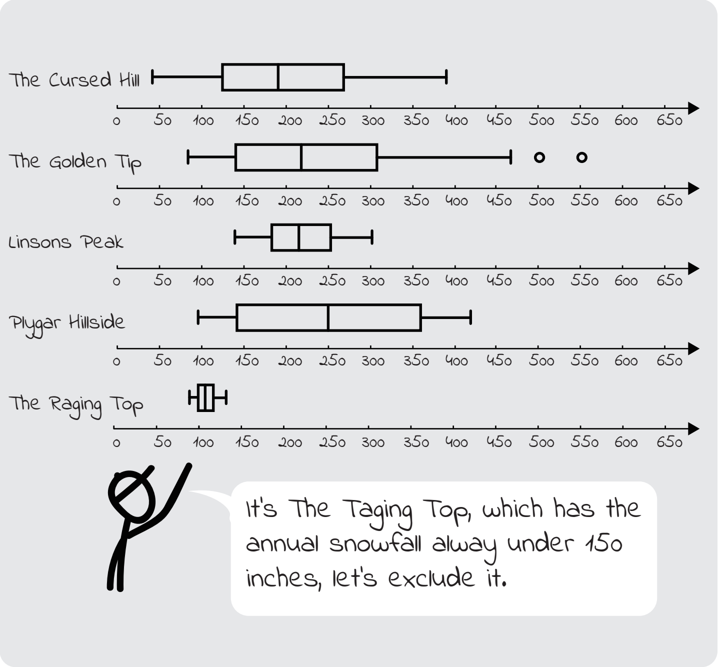

In boxplots the whiskers are generally defined as **1.5 times** the **'box' (inter-quartile range)** below the lower quartile and beyond the upper quartile. Anything outside this range is considered as an outlier.
'box' range (inter-quartile)
1.5 x 'box' (inter-quartile range)
outlier
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700

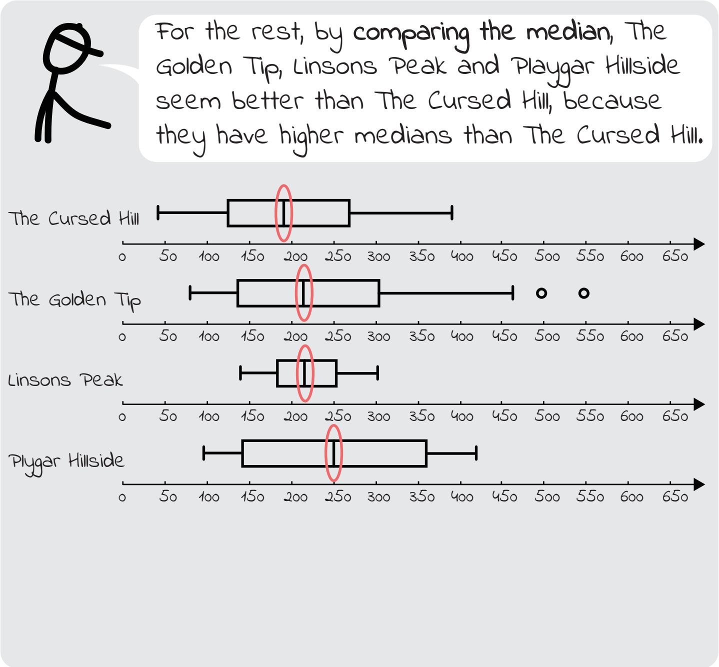

Great! How can we use boxplots to make decision?

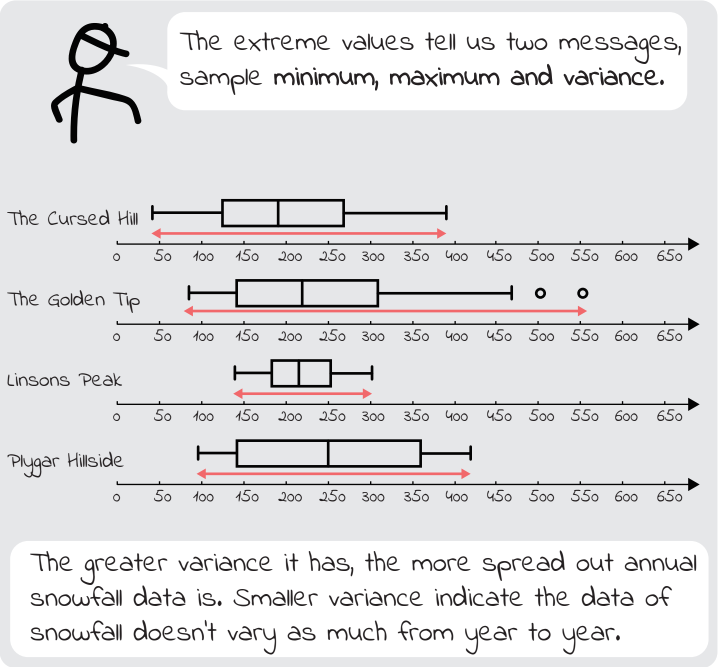

0
50
100
150
200
250
300
350
400
This is only for The Cursed Hill, we need box-plots for the rest, then I'll explain it.

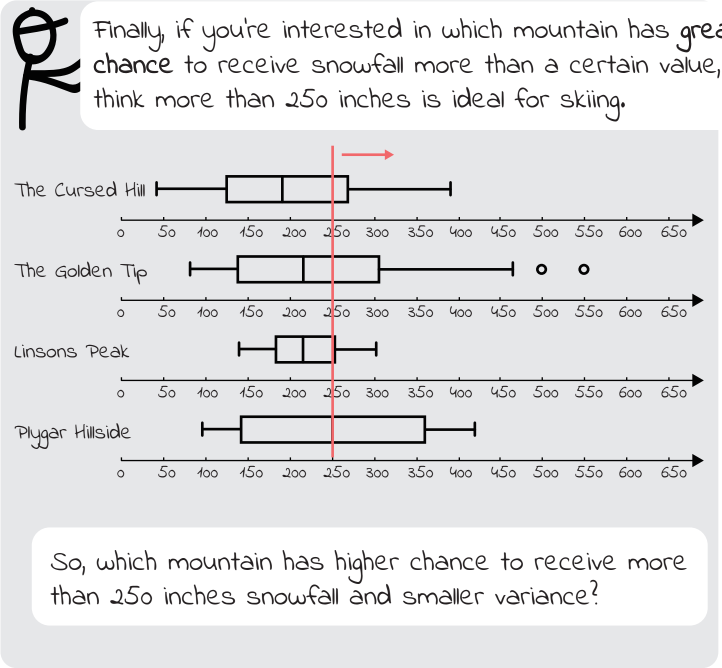

Here we go:
The Cursed Hill
The Golden Tip
Linsons Peak
Plygar Hillside
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
The Raging Top
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140

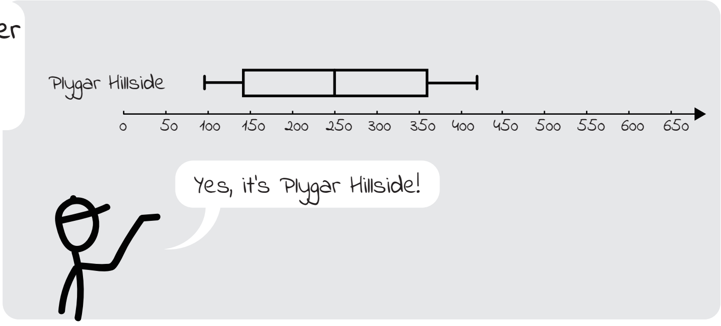

The Cursed Hill
The Golden Tip
Linsons Peak
Plygar Hillside
The Raging Top
0 50 100 150 200 250 300 350 400 450 500 550 600 650
0 10 20 30 40 50 60 70 80 90 100 110 120 130
If you see carefully, you'll see one of them has a **different scale**. Don't fall into that pitfall.


The Cursed Hill
The Golden Tip
Linsons Peak
Plygar Hillside
The Raging Top
0 50 100 150 200 250 300 350 400 450 500 550 600 650
It's The Taging Top, which has the annual snowfall alway under 150 inches, let's exclude it.

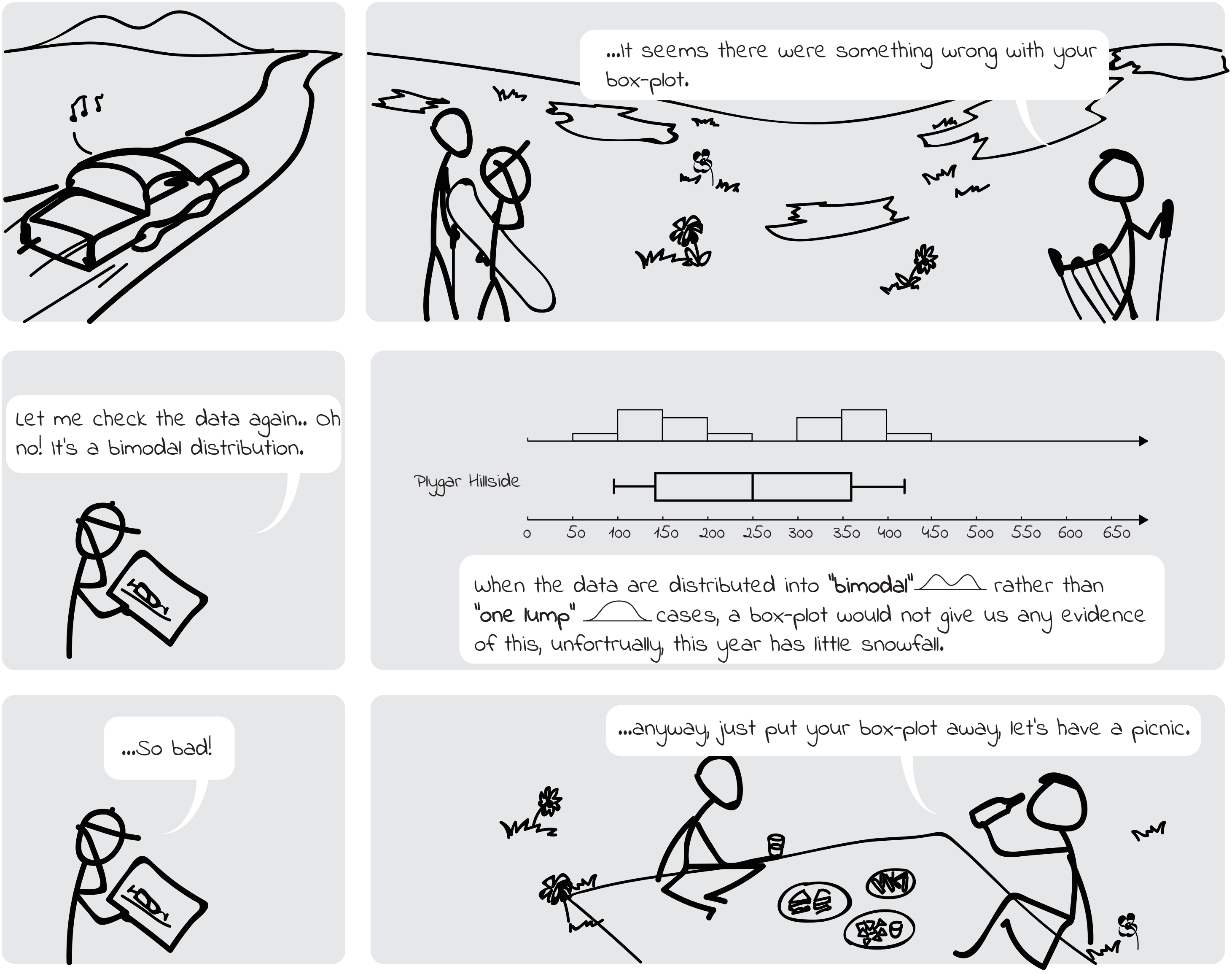

For the rest, by **comparing the median**, The Golden Tip, Linsons Peak and Playgar Hillside seem better than The Cursed Hill, because they have higher medians than The Cursed Hill.
The Cursed Hill
The Golden Tip
Linsons Peak
Plygar Hillside
0 50 100 150 200 250 300 350 400 450 500 550 600 650


The extreme values tell us two messages, sample **minimum, maximum and variance.**
The Cursed Hill
The Golden Tip
Linsons Peak
Plygar Hillside
0 50 100 150 200 250 300 350 400 450 500 550 600 650
The greater variance it has, the more spread out annual snowfall data is. Smaller variance indicate the data of snowfall doesn't vary as much from year to year.


Finally, if you're interested in which mountain has **greater chance** to receive snowfall more than a certain value, I think more than 250 inches is ideal for skiing.
The Cursed Hill
The Golden Tip
Linsons Peak
Plygar Hillside
0 50 100 150 200 250 300 350 400 450 500 550 600 650
So, which mountain has higher chance to receive more than 250 inches snowfall and smaller variance?


Plygar Hillside
0 50 100 150 200 250 300 350 400 450 500 550 600 650
Yes, it's Plygar Hillside!

Yah! I can't wait for Easter holiday!


...It seems there were something wrong with your box-plot.
Let me check the data again.. oh no! It's a bimodal distribution.
Plygar Hillside
0 50 100 150 200 250 300 350 400 450 500 550 600 650
when the data are distributed into "bimodal" rather than "one lump" cases, a box-plot would not give us any evidence of this, unfortrually, this year has little snowfall.
...So bad!
...anyway, just put your box-plot away, let's have a picnic.