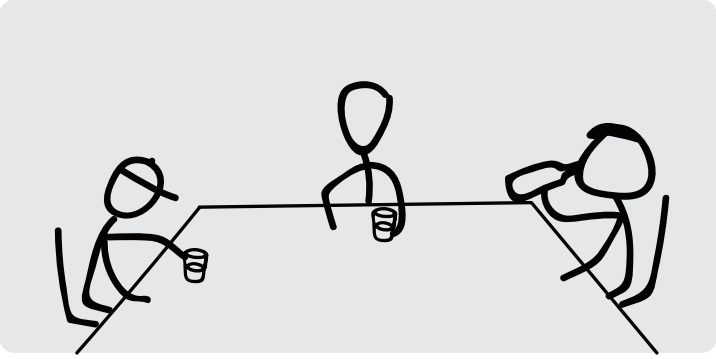
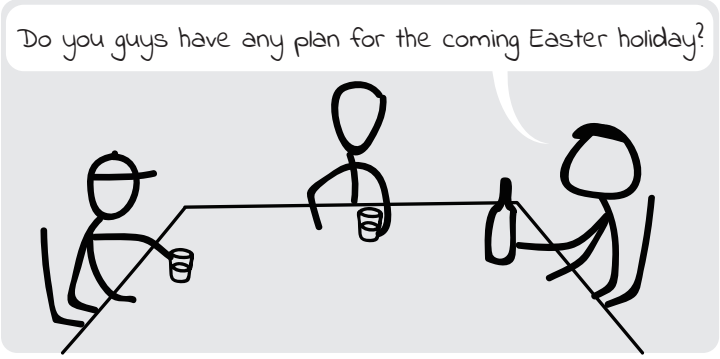
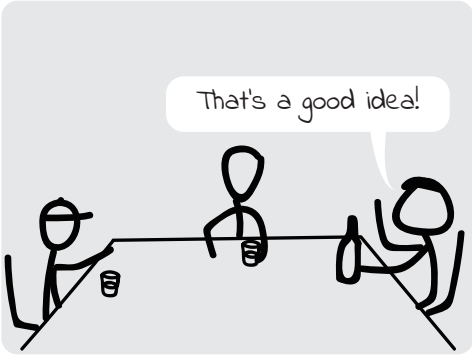
Boxplot

Introduction







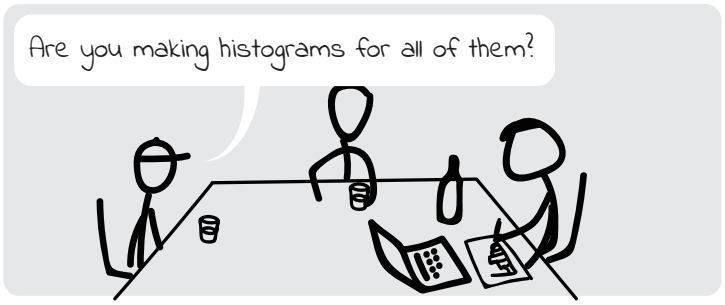
I find 5 mountains, and the data of annual snowfall for each mountain in the past 19 years.

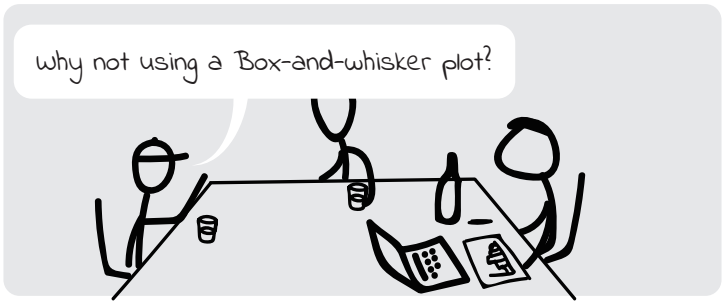
The Cursed Hill		The Golden Tip		Linsons Peak		Plygar Hillside		The Raging Top	
Year	Snowfall	Year	Snowfall	Year	Snowfall	Year	Snowfall	Year	Snowfall
2018	45	2018							
2017	58	201							
2016	111	201							
2015	155	201							
2014	211	201							
2013	255	201							
2012	320								
2011		2011	...	2011	...	2011	...	2011	...
...	...	...	...	...	...	...	...	...	...

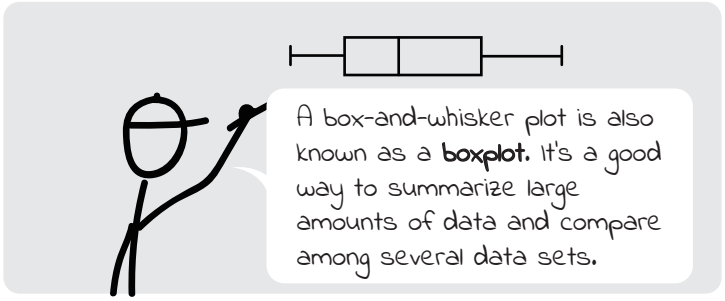
To see how these data are distributed, we should visualize them. Let's start from the first one, The Cursed Hill, and sort the snowfall value into 8 groups in an ascending order:

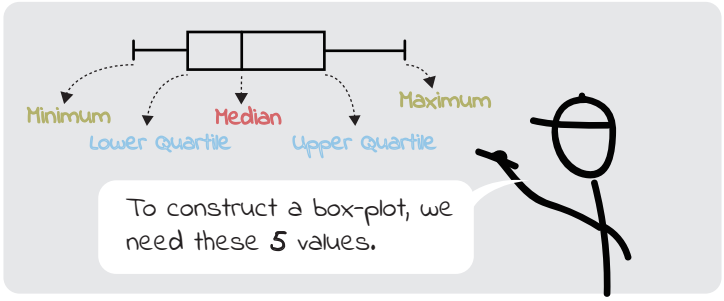
0 - 49	50 - 99	100 - 149	150 - 199	200 - 249	250 - 299	300 - 349	350 - 399
45	58	111	155	211	255	320	380
	88	125	168	224	260	333	
		140	170	235	285		
			195				

Then we map the frequency in this coordinate...

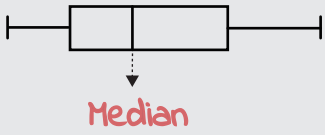






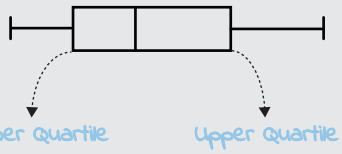


The 'median' splits the data set into **two equal groups.**



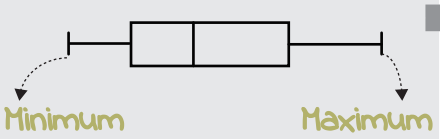
45 58 88 111 125 140 155 168 170 **195** 211 224 235 255 260 285 320 333 380

The 'lower quartile' and 'upper quartile' are the median values of lower half and higher half respectively.

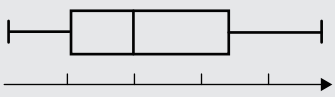


45 58 88 111 **125** 140 155 168 170 **195** 211 224 235 255 **260** 285 320 333 380

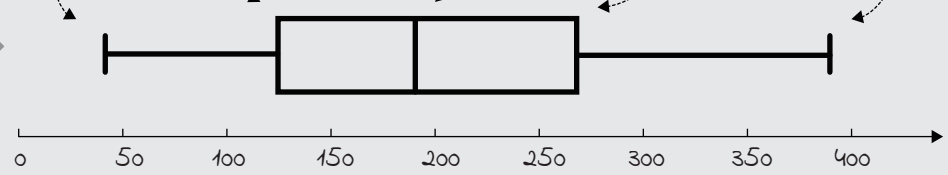
Find the 'minimum' and 'maximum'.

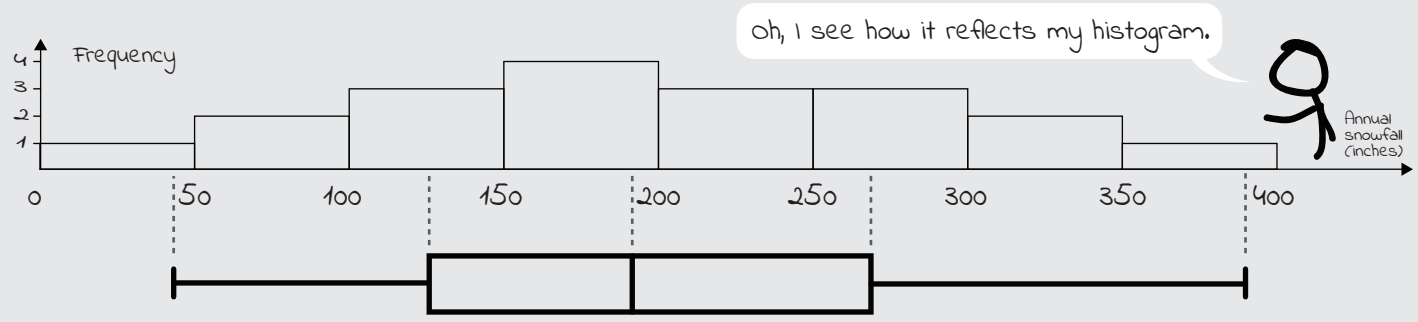


45 58 88 111 **125** 140 155 168 170 **195** 211 224 235 255 **260** 285 320 333 **380**

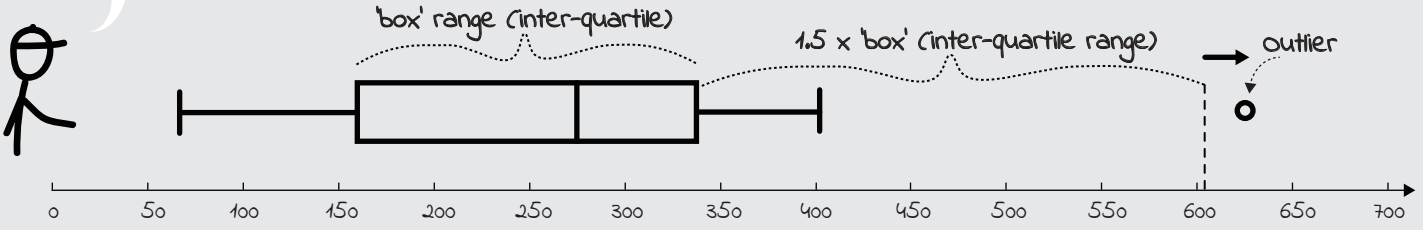


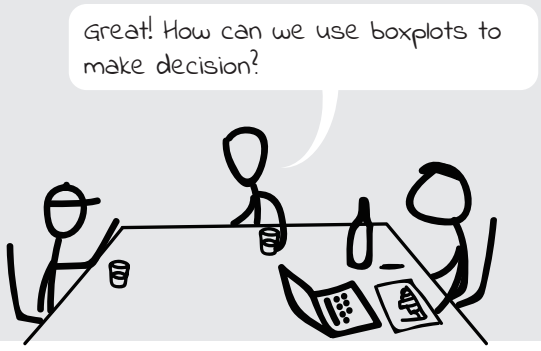
The last step, put these 5 values on the axis, draw a 'box' between 'lower quartile' and 'upper quartile' and link two 'whiskers' to 'minimum' and 'maximum'.

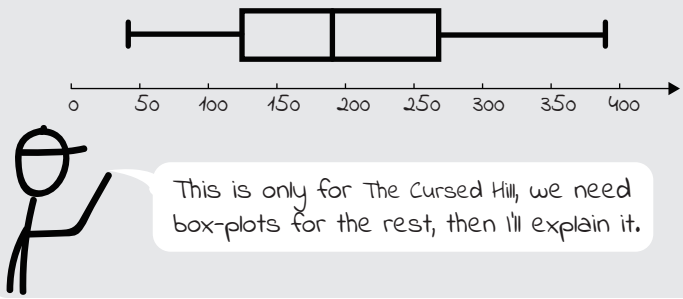




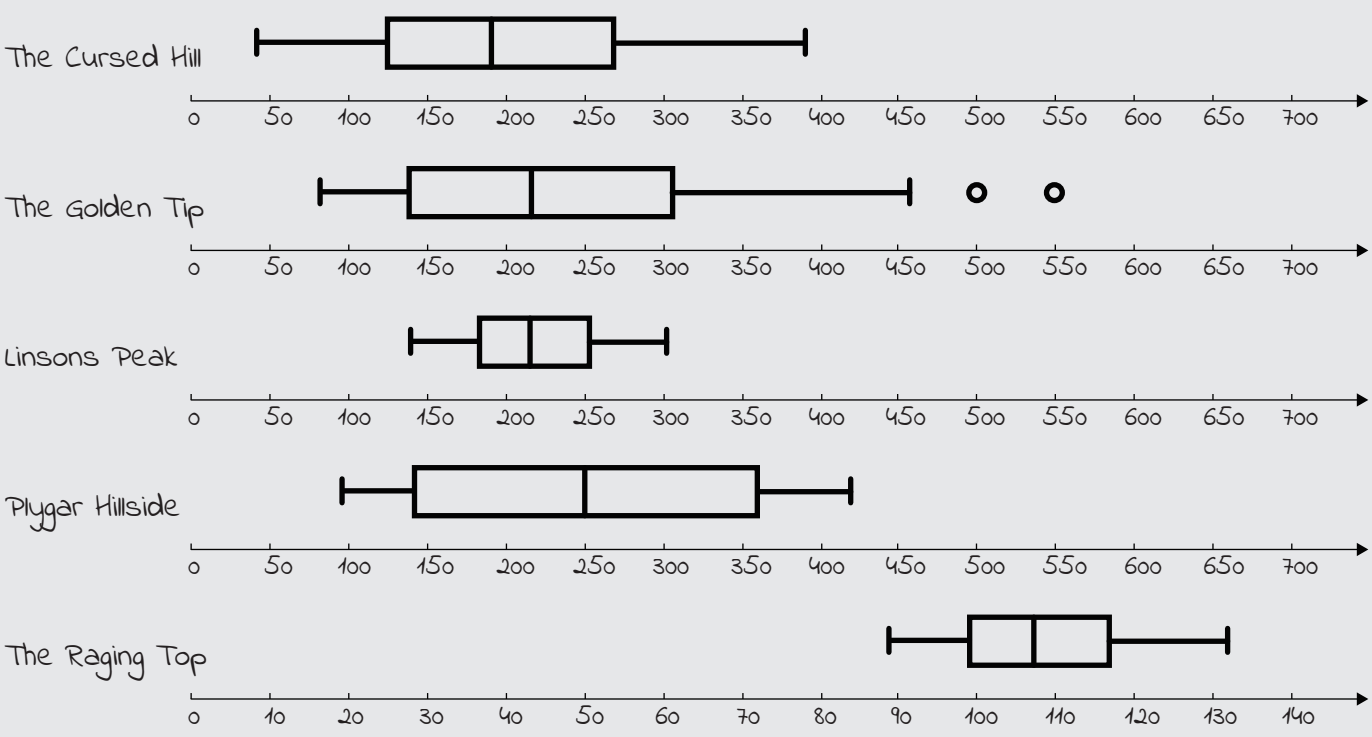
In boxplots the whiskers are generally defined as **1.5 times** the **'box' (inter-quartile range)** below the lower quartile and beyond the upper quartile. Anything outside this range is considered as an outlier.

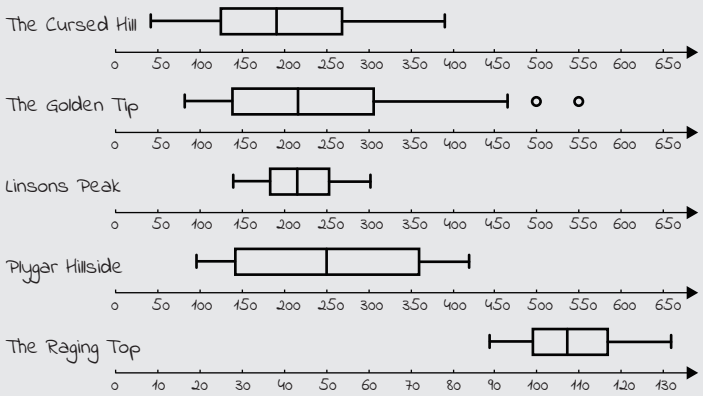




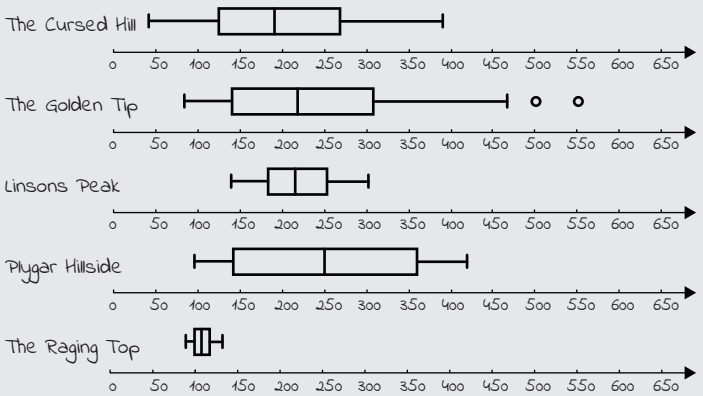


Here we go:



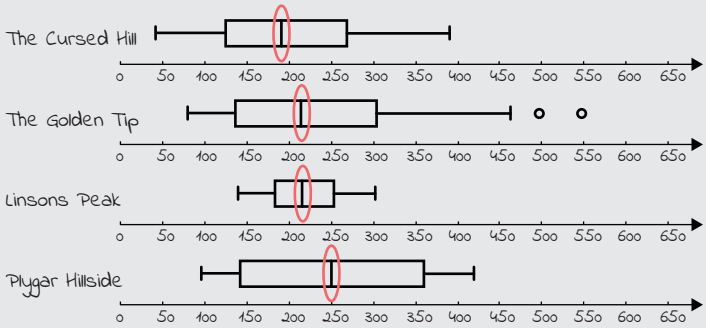


If you see carefully, you'll see one of them has a **different scale.** Don't fall into that pitfall.

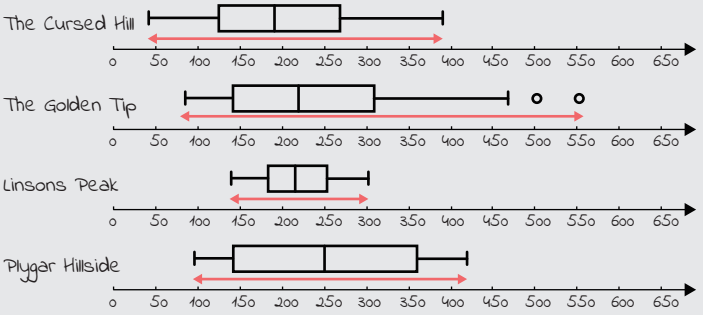


It's The Taging Top, which has the annual snowfall alway under 150 inches, let's exclude it.

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.

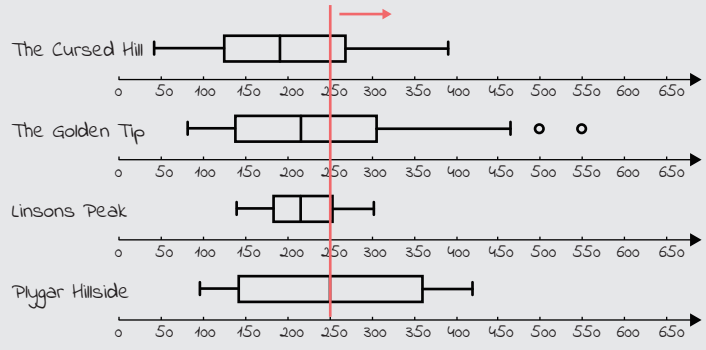


The extreme values tell us two messages, sample **minimum, maximum and variance.**

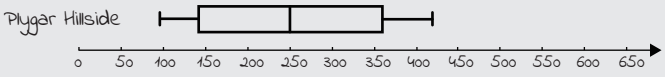


The greater variance it has, the more spread out annual snowfall data is. Smaller variance indicate the data of snowfall doesn't vary as much from year to year.

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.

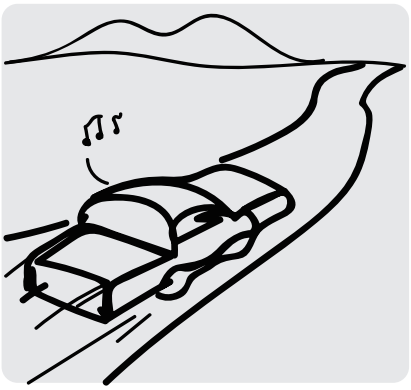


So, which mountain has higher chance to receive more than 250 inches snowfall and smaller variance?

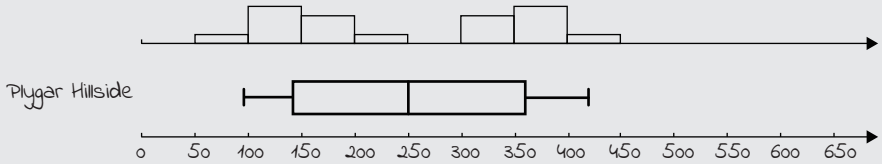


Yes, it's Plygar Hillside!

Yah! I can't wait for Easter holiday!



Let me check the data again.. oh no! It's a bimodal distribution.



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.