EXPLORATORY DATA ANALYSIS

Data Visualization
and Wrangling

HOW DO WE MAKE SENSE OF DATA?

Humans understand visual information 60,000 times faster than any other form of information. Think about how we use hand gestures to increase understanding when we communicate. That's no coincidence. When it comes to processing information, visualizing data is king.

You are hired as a consultant to answer a question and/or predict the future.

You first gather information and collect data.

Now look at the dataset.
What does it reveal?
Does staring at the numbers help?

You want to learn about any patterns or randomness in the set.

Your job is to explore the dataset and gain insight into any phenomena present.

216	108	200	137	272	173	282	216	117	261
110	235	252	105	282	130	105	288	96	255
108	105	207	184	272	216	118	245	231	266
258	268	202	242	230	121	112	290	110	287
261	113	274	105	272	199	230	126	278	120
288	283	110	290	104	293	223	100	274	259
134	270	105	288	109	264	250	282	124	282
242	118	270	240	119	304	121	274	233	216
248	260	246	158	244	296	237	271	130	240
132	260	112	289	110	258	280	225	112	294
149	262	126	270	243	112	282	107	291	221
284	138	294	265	102	278	139	276	109	265
157	244	255	118	276	226	115	270	136	279
112	250	168	260	110	263	113	296	122	224
254	134	272	289	260	119	278	121	306	108
302	240	144	276	214	240	270	245	108	238
132	249	120	230	210	275	142	300	116	277
115	125	275	200	250	260	270	145	240	250
113	275	255	226	122	266	245	110	265	131
288	110	288	246	238	254	210	262	135	280
126	261	248	112	276	107	262	231	116	270
143	282	112	230	205	254	144	288	120	249
112	256	105	269	240	247	245	256	235	273
245	145	251	133	267	113	111	257	237	140
249	141	296	174	275	230	125	262	128	261
132	267	214	270	249	229	235	267	120	257
286	272	111	255	119	135	285	247	129	265
109	268								

In practice, data sets contain so many elements that it is absolutely necessary to condense the data for easy visual comprehension of general characteristics.

Today we discuss the histogram which is great for representing univariate datasets.

Univariate: consisting of repeated measurements of **one** particular quantity.

We will also consider scatterplots and box-and-whisker plots.

216	108	200	137	272	173	282	216	117	261
110	235	252	105	282	130	105	288	96	255
108	105	207	184	272	216	118	245	231	266
258	268	202	242	230	121	112	290	110	287
261	113	274	105	272	199	230	126	278	120
288	283	110	290	104	293	223	100	274	259
134	270	105	288	109	264	250	282	124	282
242	118	270	240	119	304	121	274	233	216
248	260	246	158	244	296	237	271	130	240
132	260	112	289	110	258	280	225	112	294
149	262	126	270	243	112	282	107	291	221
284	138	294	265	102	278	139	276	109	265
157	244	255	118	276	226	115	270	136	279
112	250	168	260	110	263	113	296	122	224
254	134	272	289	260	119	278	121	306	108
302	240	144	276	214	240	270	245	108	238
132	249	120	230	210	275	142	300	116	277
115	125	275	200	250	260	270	145	240	250
113	275	255	226	122	266	245	110	265	131
288	110	288	246	238	254	210	262	135	280
126	261	248	112	276	107	262	231	116	270
143	282	112	230	205	254	144	288	120	249
112	256	105	269	240	247	245	256	235	273
245	145	251	133	267	113	111	257	237	140
249	141	296	174	275	230	125	262	128	261
132	267	214	270	249	229	235	267	120	257
286	272	111	255	119	135	285	247	129	265
109	268								

You have been hired by National Parks to determine the health of Old Faithful Geyser in Yellow Stone.

You watch 272 consecutive eruptions and record their duration in seconds.

Old Faithful

What does the data reveal?
Looking at the dataset is a poor summary of what is happening.
How can we derive information from the data?

Perhaps rearrange it. Put it in order.

216	108	200	137	272	173	282	216	117	261
110	235	252	105	282	130	105	288	96	255
108	105	207	184	272	216	118	245	231	266
258	268	202	242	230	121	112	290	110	287
261	113	274	105	272	199	230	126	278	120
288	283	110	290	104	293	223	100	274	259
134	270	105	288	109	264	250	282	124	282
242	118	270	240	119	304	121	274	233	216
248	260	246	158	244	296	237	271	130	240
132	260	112	289	110	258	280	225	112	294
149	262	126	270	243	112	282	107	291	221
284	138	294	265	102	278	139	276	109	265
157	244	255	118	276	226	115	270	136	279
112	250	168	260	110	263	113	296	122	224
254	134	272	289	260	119	278	121	306	108
302	240	144	276	214	240	270	245	108	238
132	249	120	230	210	275	142	300	116	277
115	125	275	200	250	260	270	145	240	250
113	275	255	226	122	266	245	110	265	131
288	110	288	246	238	254	210	262	135	280
126	261	248	112	276	107	262	231	116	270
143	282	112	230	205	254	144	288	120	249
112	256	105	269	240	247	245	256	235	273
245	145	251	133	267	113	111	257	237	140
249	141	296	174	275	230	125	262	128	261
132	267	214	270	249	229	235	267	120	257
286	272	111	255	119	135	285	247	129	265
109	268								

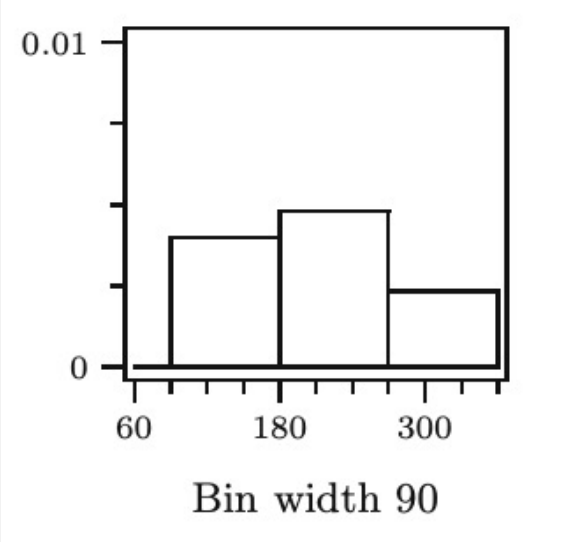
- Order immediately reveals the range 96 to 306.

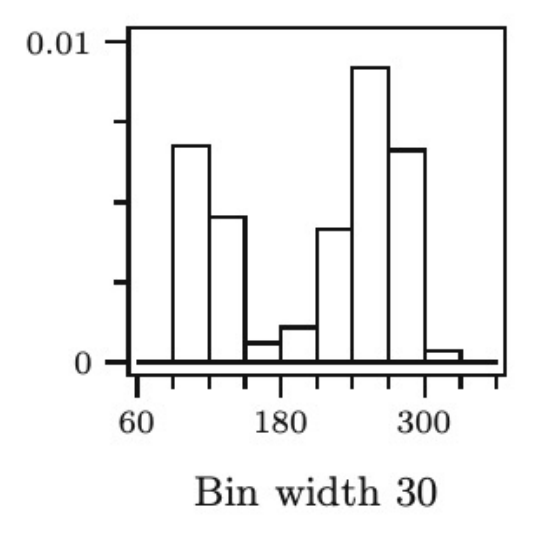
- And looking at the middle elements (136th & 137th) reveals they are closer to the maximum than they are to the minimum.

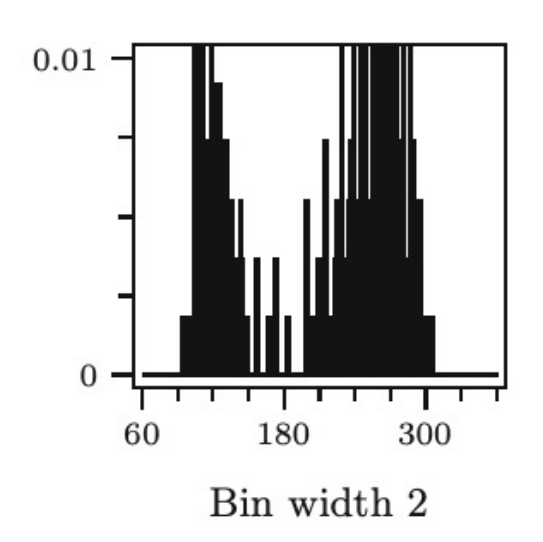
- Q: So what?

  A: Asymmetric, but how?

96	100	102	104	105	105	105	105	105	105
107	107	108	108	108	108	109	109	109	110
110	110	110	110	110	110	111	111	112	112
112	112	112	112	112	112	113	113	113	113
115	115	116	116	117	118	118	118	119	119
119	120	120	120	120	121	121	121	122	122
124	125	125	126	126	126	128	129	130	130
131	132	132	132	133	134	134	135	135	136
137	138	139	140	141	142	143	144	144	145
145	149	157	158	168	173	174	184	199	200
200	202	205	207	210	210	214	214	216	216
216	216	221	223	224	225	226	226	229	230
230	230	230	230	231	231	233	235	235	235
237	237	238	238	240	240	240	240	240	240
242	242	243	244	244	245	245	245	245	245
246	246	247	247	248	248	249	249	249	249
250	250	250	250	251	252	254	254	254	255
255	255	255	256	256	257	257	258	258	259
260	260	260	260	260	261	261	261	261	262
262	262	262	263	264	265	265	265	265	266
266	267	267	267	268	268	269	270	270	270
270	270	270	270	270	271	272	272	272	272
272	273	274	274	274	275	275	275	275	276
276	276	276	277	278	278	278	279	280	280
282	282	282	282	282	282	283	284	285	286
287	288	288	288	288	288	288	289	289	290
290	291	293	294	294	296	296	296	300	302
304	306								







- In determining the health of the geyser, a Geologist on your team suggests that there are actually two types of eruptions that are possible, and you could be witnessing both in the data.

- Does the data bear this out? At this point, it is not clear; maybe…

- Let's look at some histograms of the data (different amounts of bins).
- A histogram groups the data into bins and then you see how many of each 'type' there are.

Of course, we will use Python to create our histograms but we also have to understand what they are and how they work.

So, let's make a histogram by hand.
data: {2, 9, 10, 9, 4, 4, 5, 3, 8, 5, 8, 5, 7, 8, 6}
ordered data: {2, 3, 4, 4, 5, 5, 5, 6, 7, 8, 8, 8, 9, 9, 10, }

How many bins should be used?

(pg. 211, $m = 1 + 3.3\log_{10}(n)$ or $b = 3.49sn^{-1/3}$)

Guess and test: 5

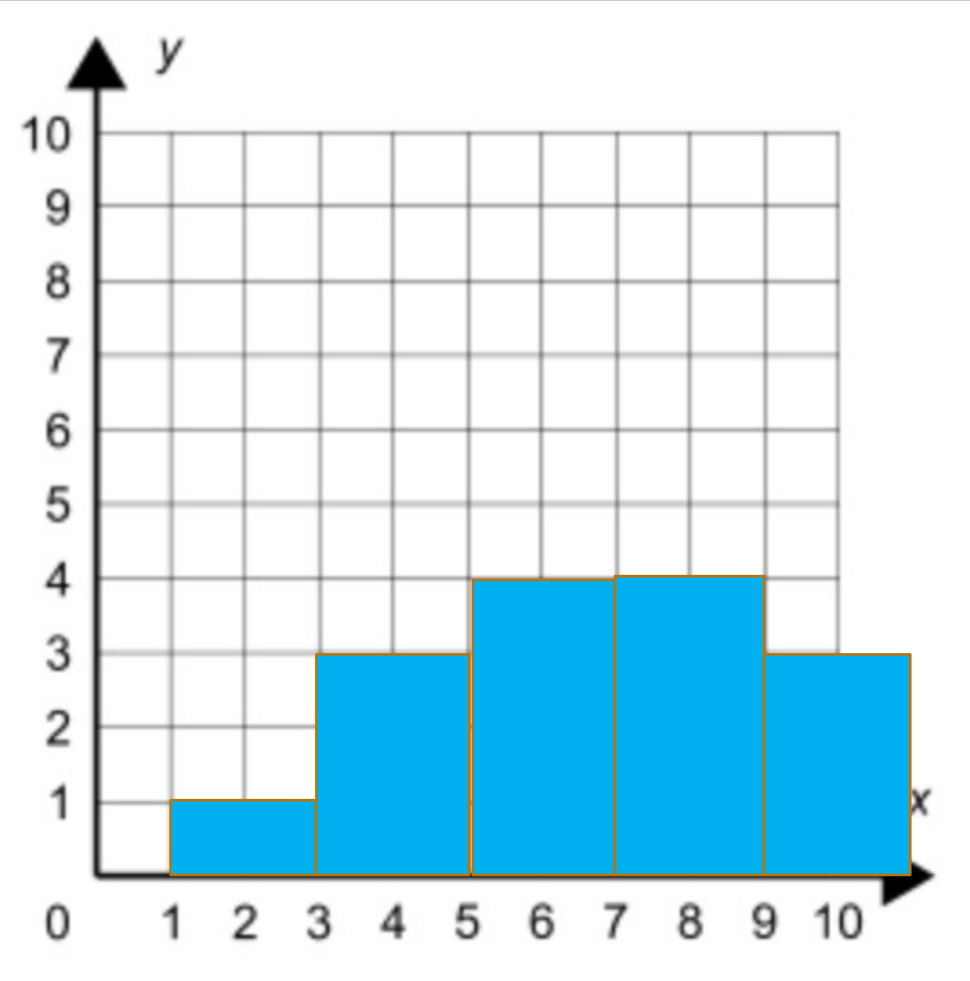
bin1: $[1, 3)$ contains 1 piece of data
bin 2: $[3,5)$ contains 3 pieces of data
bin 3: $[5, 7)$ contains 4 pieces of data
bin 4: $[7, 9)$ contains 4 pieces of data
bin 5: $[9, 11)$ contains 3 pieces of data

Now, we are ready to make a frequency histogram or a density histogram.

Frequency Histogram

ordered data: {2, 3, 4, 4, 5, 5, 5, 6, 7, 8, 8, 8, 9, 9, 10, }

bin1: [1, 3) 1
bin 2: [3,5) 3
bin 3: [5, 7) 4
bin 4: [7, 9) 4
bin 5: [9, 11) 3

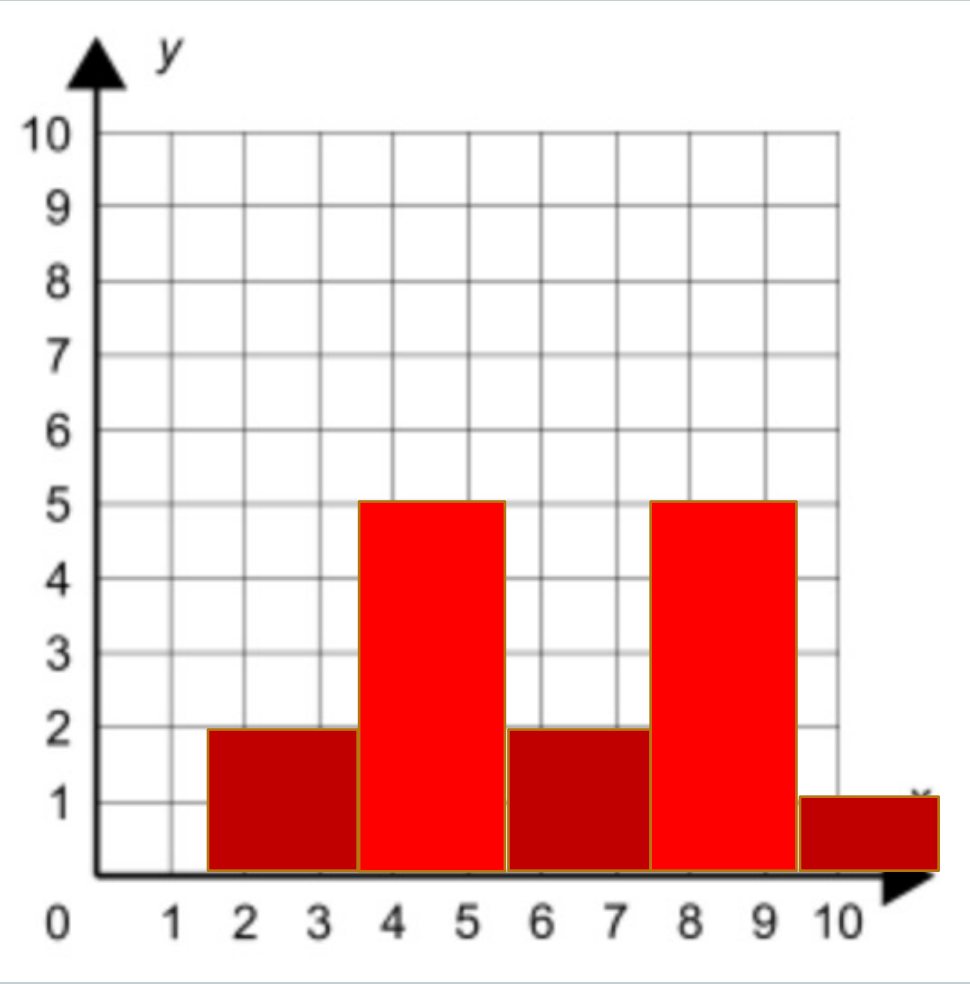


Should bins include actual data points?
Should bins go a bit below and a bit above the actual data?

Frequency Histogram

ordered data: {2, 3, 4, 4, 5, 5, 5, 6, 7, 8, 8, 8, 9, 9, 10, }

bin1: [1.5, 3.5) 2
bin 2: [3.5, 5.5) 5
bin 3: [5.5, 7.5) 2
bin 4: [7.5, 9.5) 5
bin 5: [9.5, 11.5) 1

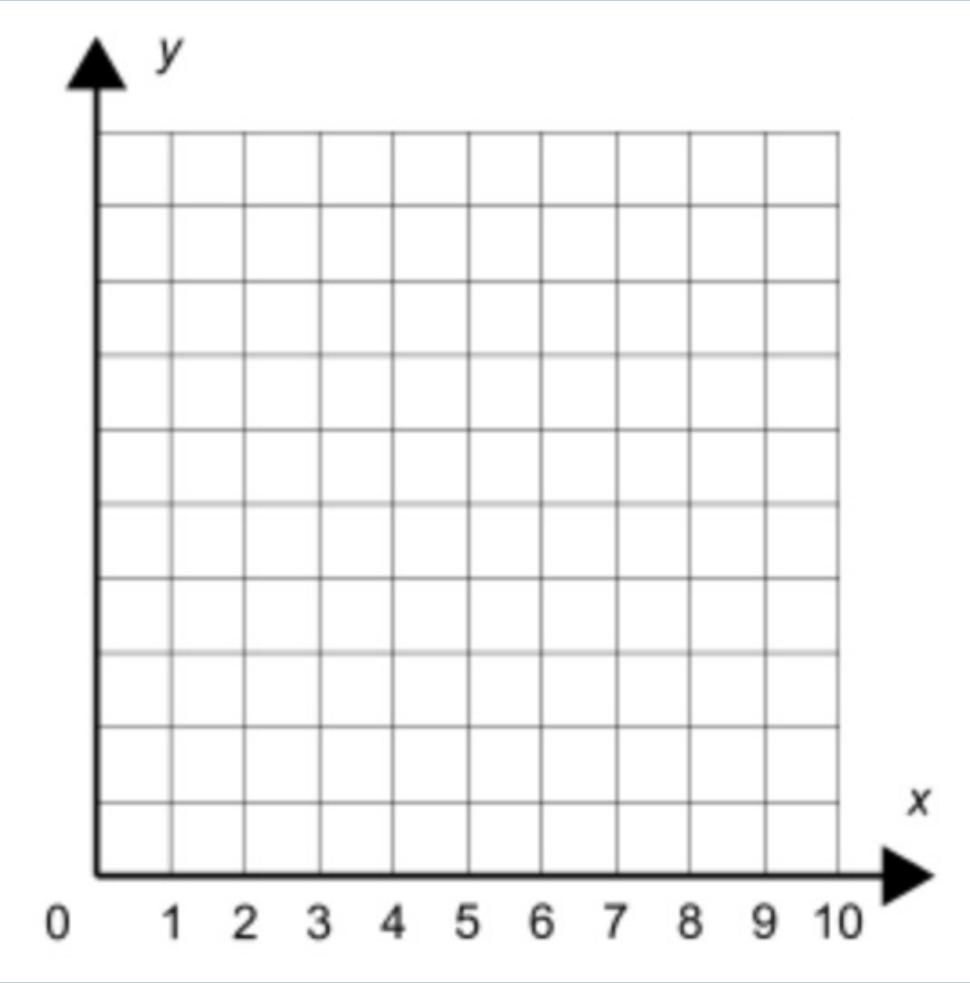


Denisty Histogram

ordered data: {2, 3, 4, 4, 5, 5, 5, 6, 7, 8, 8, 8, 9, 9, 10, }

bin1: [1, 3) 1/15
bin 2: [3,5) 3/15
bin 3: [5, 7) 4/15
bin 4: [7, 9) 4/15
bin 5: [9, 11) 3/15

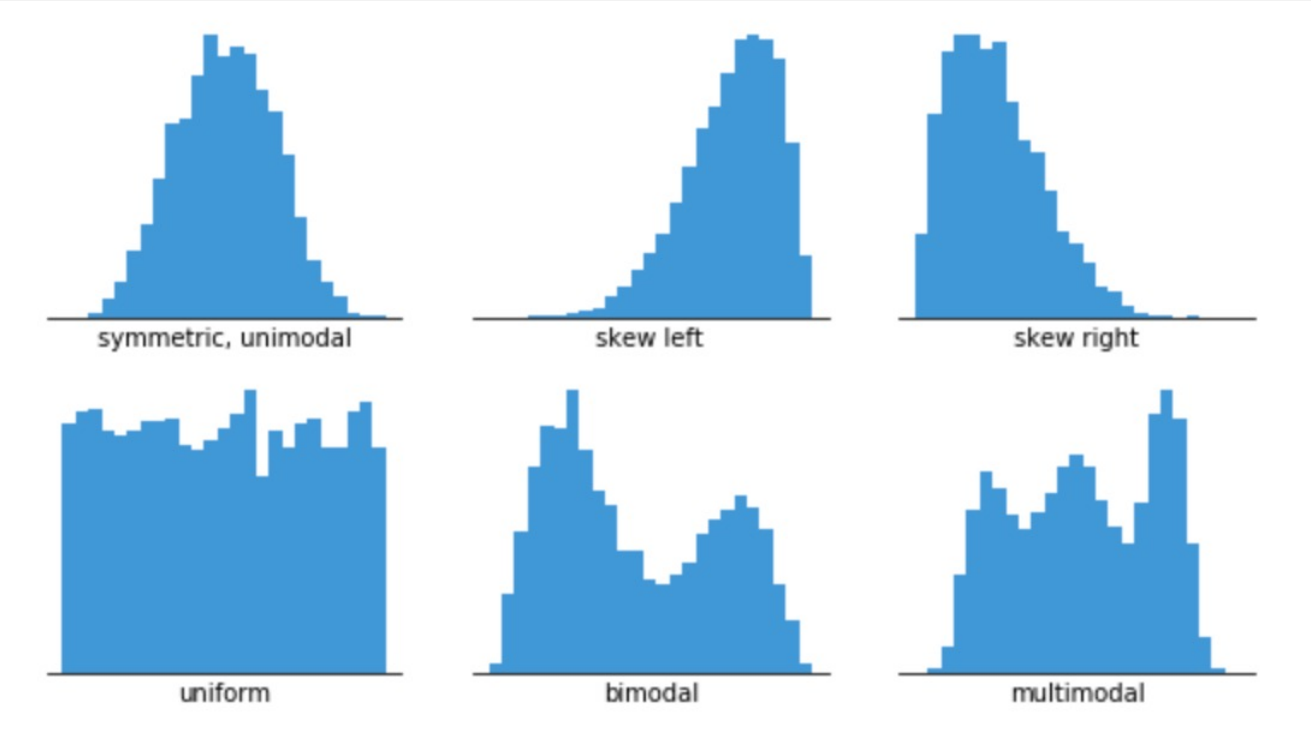
or bin1: [1.5, 3.5) 2/15
bin 2: [3.5, 5.5) 5/15
bin 3: [5.5, 7.5) 2/15
bin 4: [7.5, 9.5) 5/15
bin 5: [9.5, 11.5) 1/15



bin widths create different pictures, but only the y-axis will change from the previous pictures

ALTHOUGH ALL HISTOGRAMS WILL SHOW DIFFERENT SHAPES, THERE ARE SOME STANDARD SHAPES THAT SHOULD BE RECOGNIZED:

- UNIMODAL,
- BIMODAL,
- MULTIMODAL,
- UNIFORM,
- SKEW



VEHICLE	FUEL ECONOMY	VEHICLE	FUEL ECONOMY	VEHICLE	FUEL ECONOMY	VEHICLE	FUEL ECONOMY
Aston Martin V8 Vantage S	16	Hyundai Elantra	32	Kia Forte	31	Honda Accord	31
BMW 528i	27	Chevrolet Cruze	30	Buick Regal	24	Mazda 6	32
Subaru Legacy	30	Toyota Corolla	31	Lincoln MKZ Hybrid	40	Dodge Challenger	23
Dodge Dart	31	Lexus GS 450h	31	Subaru Impreza	31	Volvo V60 FWD	29

There are, of course, numerous ways to picture data.

In the previous set of slides, we talked about providing a 5-number summary of a set of data. This 5-number summary leads to a graphic called the box-and-whisker plot.

Suppose you are now attempting to get a picture/understanding of fuel efficiency in a variety of vehicles.

Get the 5-number summary and then choose a number line.

Data from chart: {16, 27, 30, 31, 32, 30, 31, 31, 31, 24, 40, 31, 31, 32, 23, 29}

Order the data: {16, 23, 24, 27, 29, 30, 30, 31, 31, 31, 31, 31, 31, 32, 32, 40}

There are 16 pieces of data.

Compute the quartiles and the IQR, list the min and max:

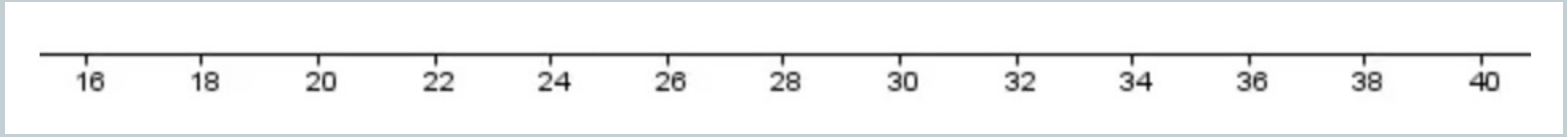
min: 16

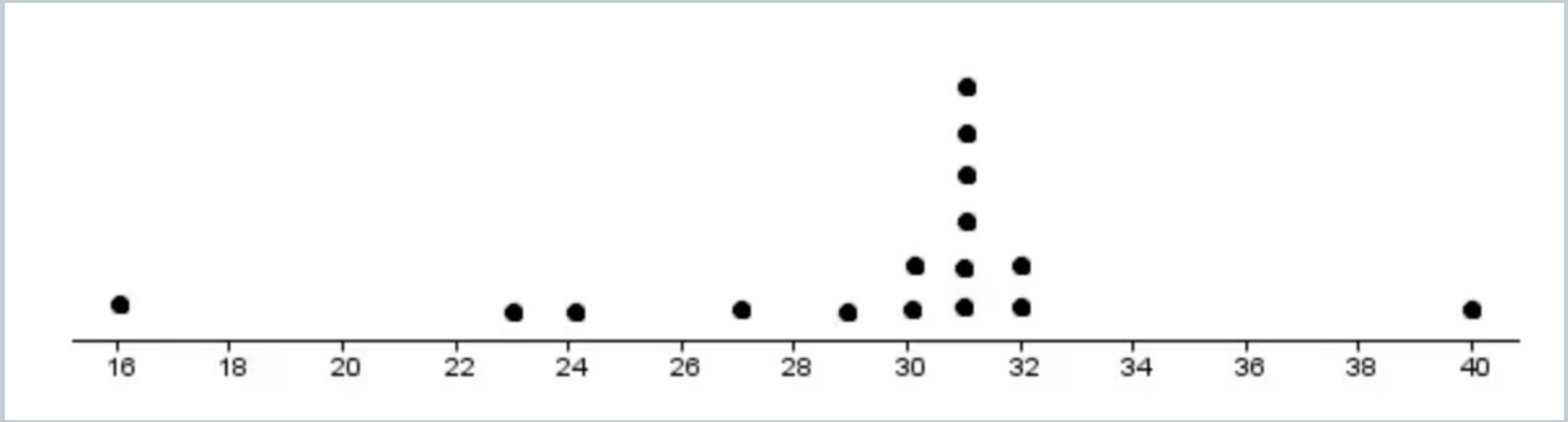
Q_1: $\frac{27+29}{2} = 28$

Q_2: $\frac{31+31}{2} = 31$

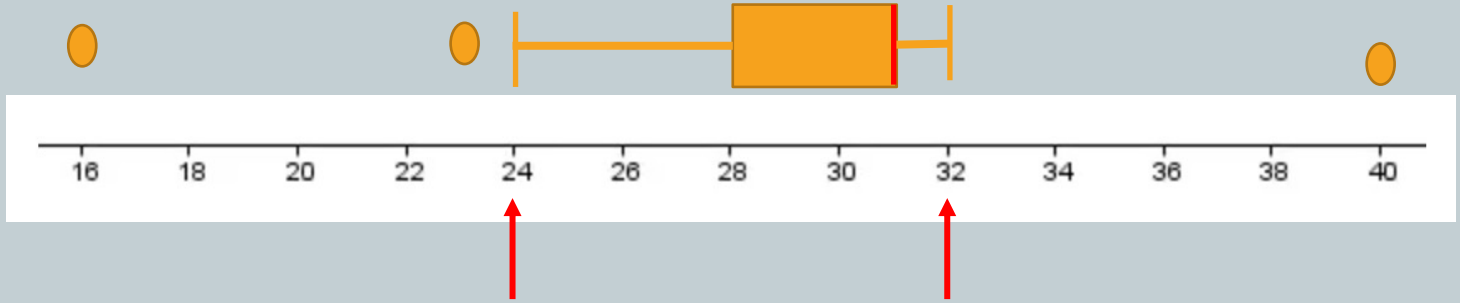
Q_3: $\frac{31+31}{2} = 31$

Max: 40





Box-and-Whisker Plot



{16, 23, 24, 27, 29, 30, 30, 31, 31, 31, 31, 31, 31, 32, 32, 40}

Recall the median (aka Q_2) is 31. Is it on the plot?

Q_1 is 28

Q_3 is 31

The IQR is $31 - 28 = 3$

outliers lay $1.5 \cdot IQR = 4.5$ beyond Q_1 and Q_3

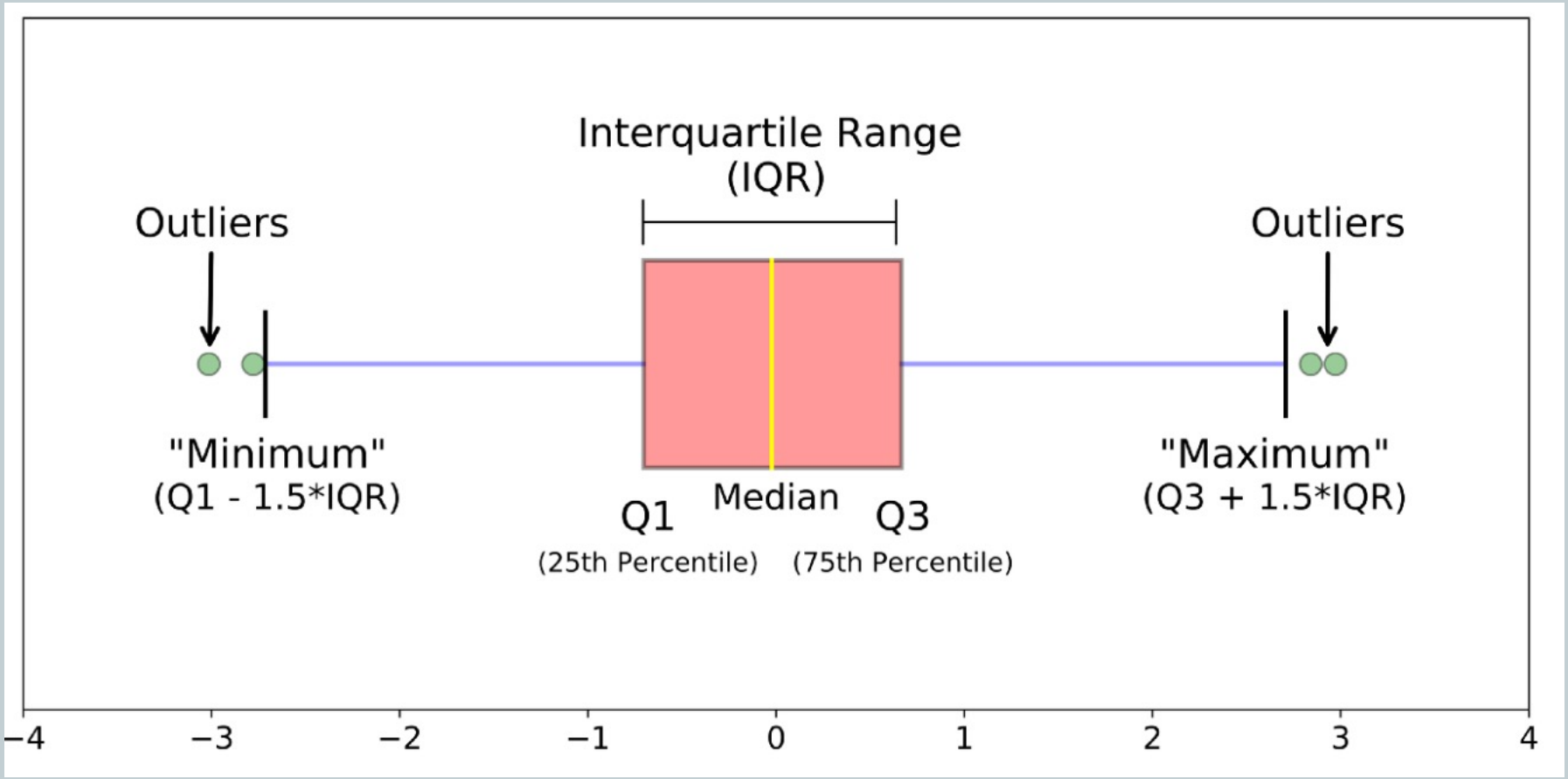
At Q_1 we have $28 - 4.5 = 23.5$

At Q_3 we have $31 + 4.5 = 35.5$

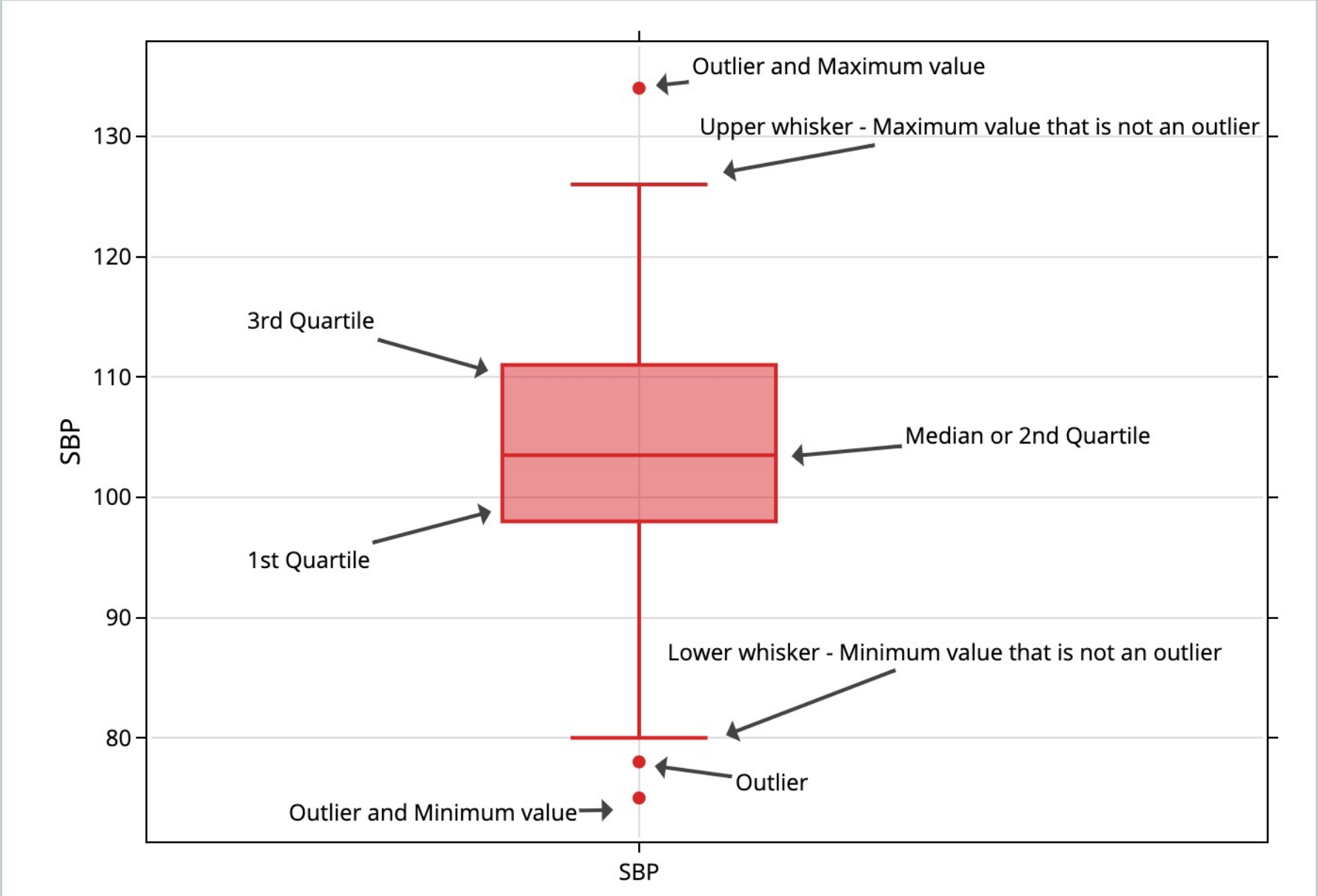
But the whiskers do not go to 23.5 and 35.5.
They go instead to 24 and 32. Why?

And what are the circles for?

Anatomy of a (horizontal) box plot



Anatomy of a (vertical) box plot



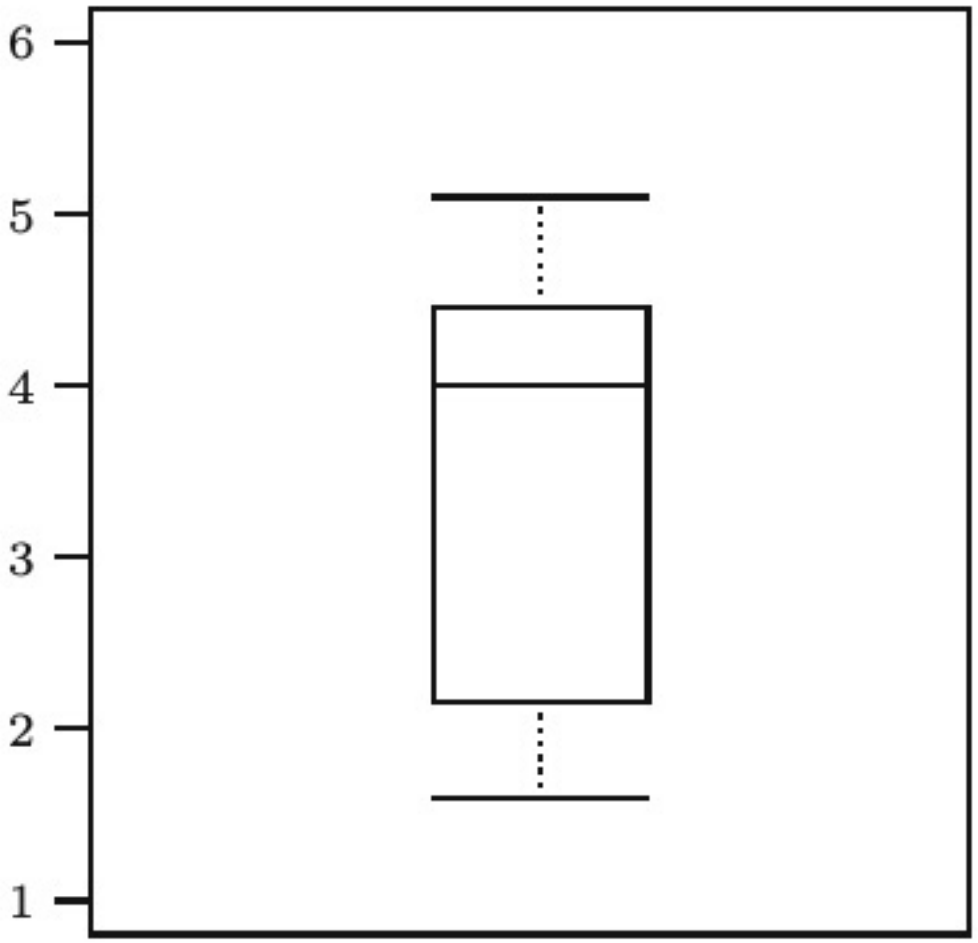
BOX AND WHISKER PLOT FOR OLD FAITHFUL DATA

A SHORTCOMING IS:

THE TWO SEPARATE PEAKS ARE HIDDEN

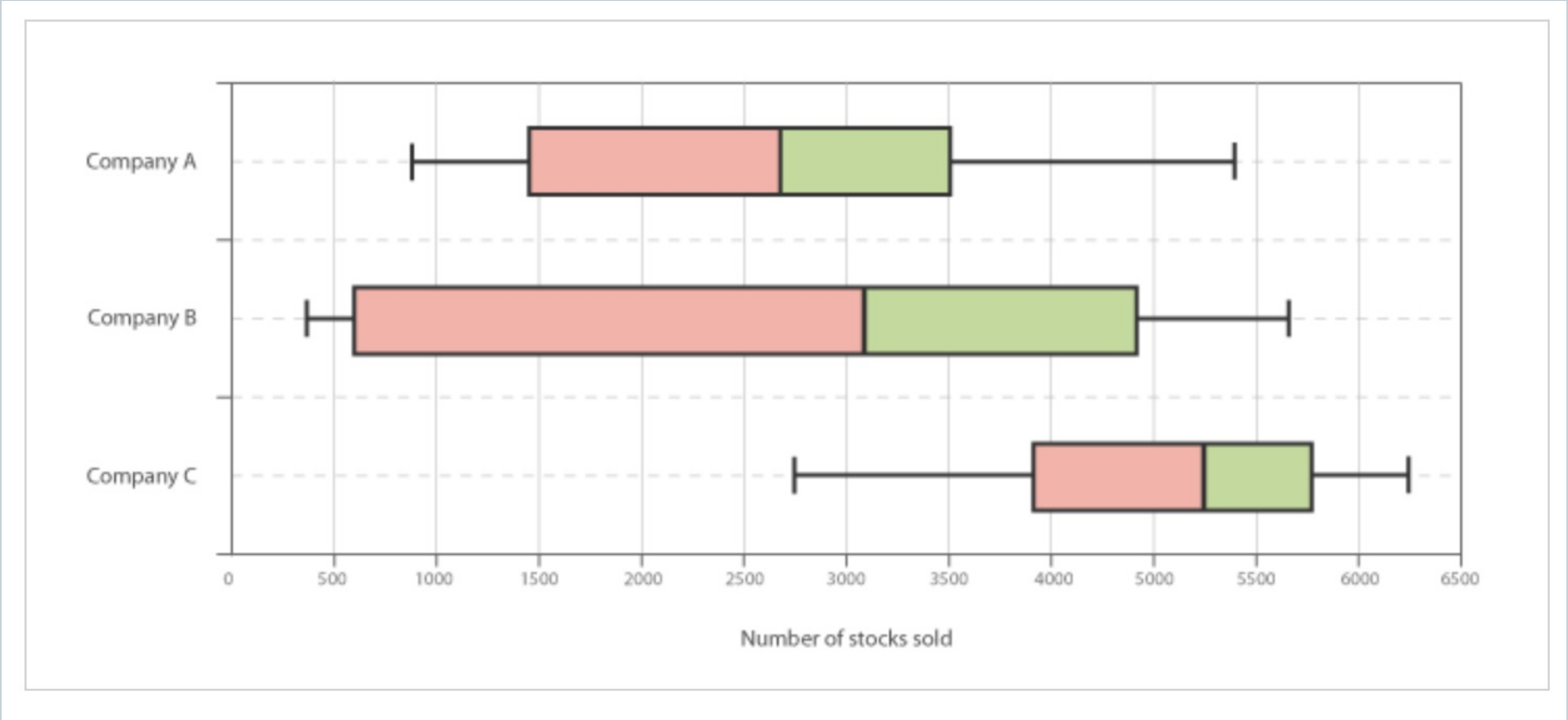
HOWEVER, THE POSITION OF THE SAMPLE MEDIAN INSIDE THE BOX SUGGESTS THAT THE DATASET IS SKEWED

ALL SAID, HISTOGRAMS ARE MORE INFORMATIVE AS GRAPHICAL SUMMARIES

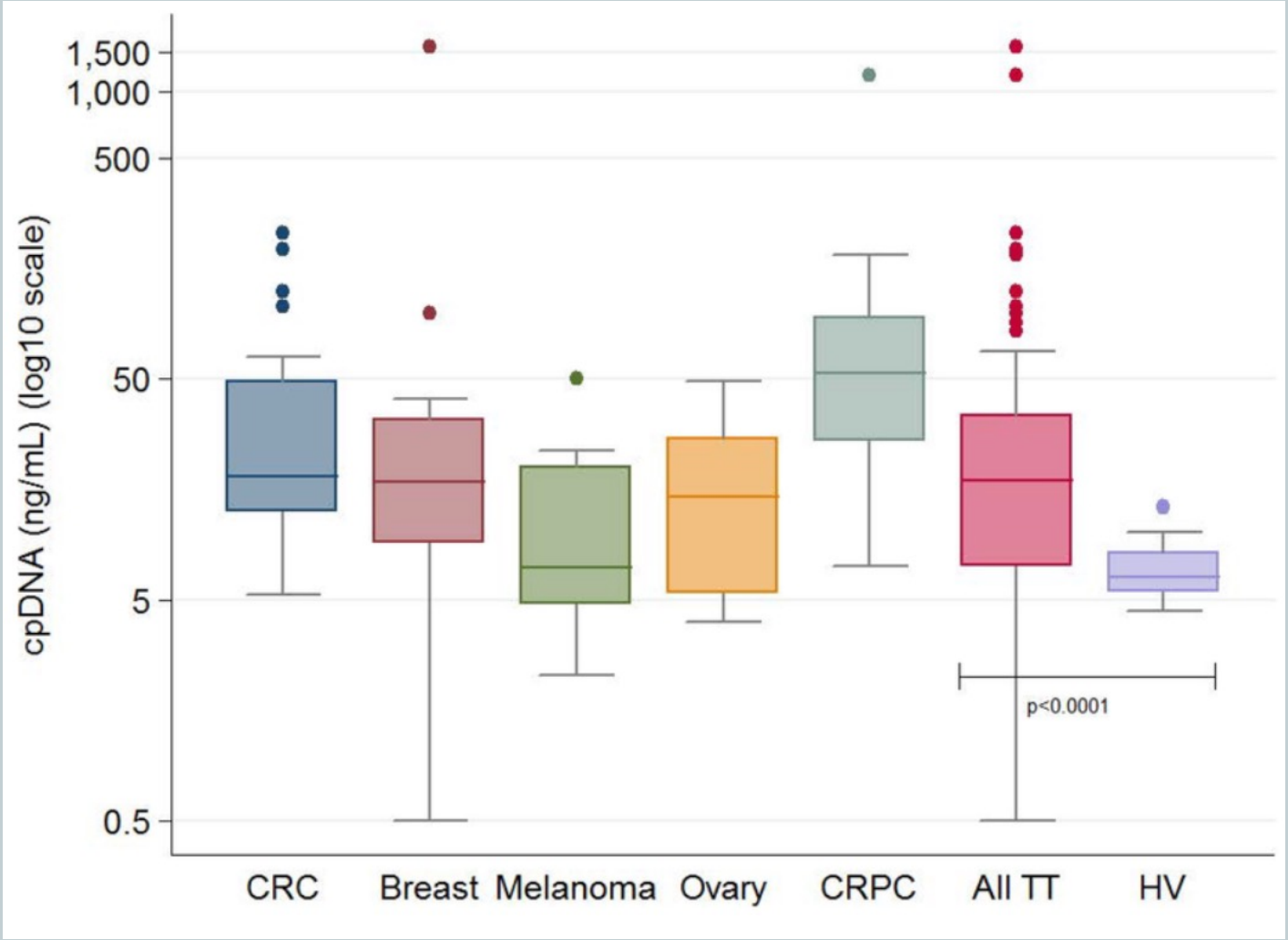


Old Faithful data

Box-and-Whisker plots are useful for side-by-side comparisons of different sets of data



Side-by-side comparison using vertical box plots



k