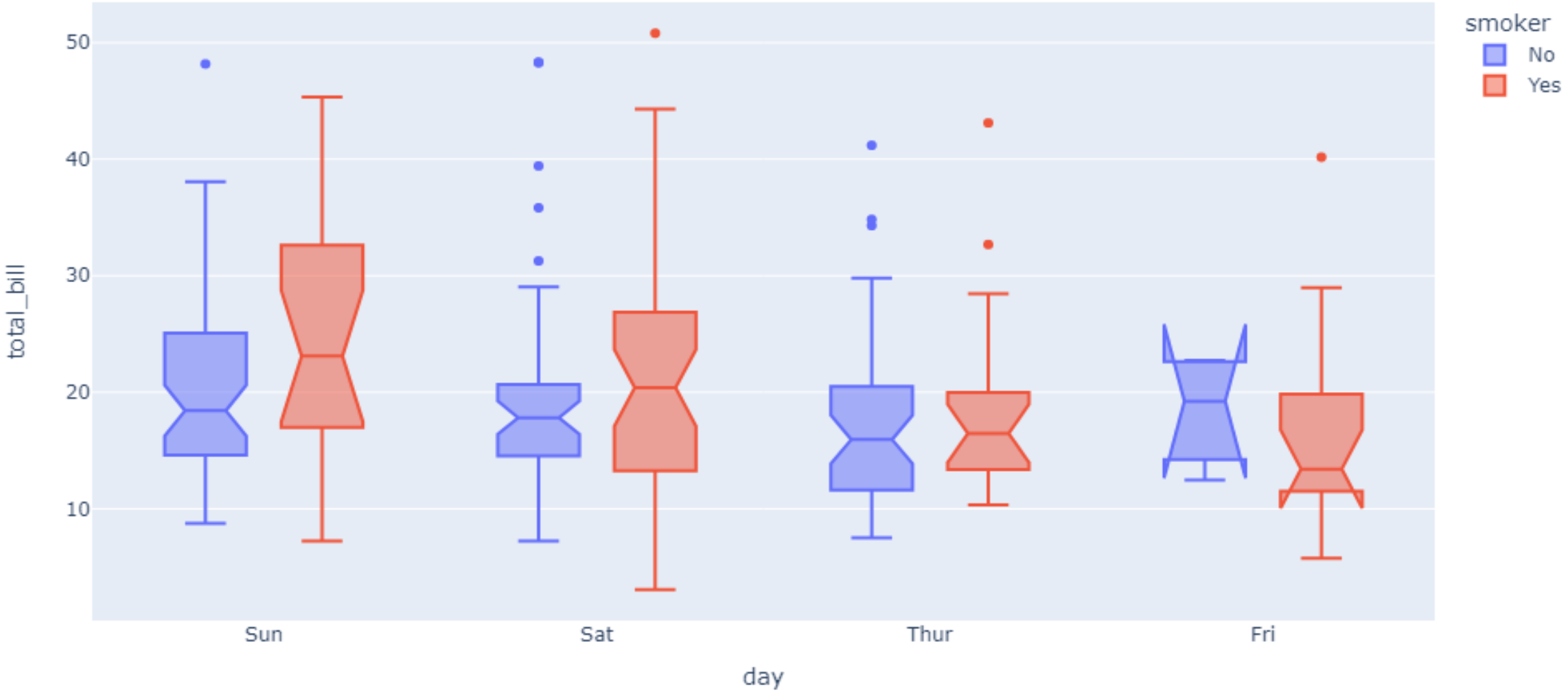
Outliers in data set is always a pain, lets figure out how to detect the outliers ... a bit Manually.!

BOX PLOTS are an amazing way to visualize the outliers.

An Example of box plots from *tips* dataset of seaborn library



Let’s now consider a data set of age of all the participants of a survey as below

Age={19, 19 ,20 ,20 ,20 ,21 ,25 ,26 ,35 , 150 ,42 ,41 ,39 ,54 ,67 ,70 ,15 ,15 ,16 ,17 ,18 ,23 ,24 ,52 ,99}

Can you identify the outliers ?

Let's sort the data for starter

The sorted data set looks like below.

{15, 15, 16, 17, 18, 19, 19, 20, 20, 20, 21, 23, 24, 25, 26, 35, 39, 41, 42, 52, 54, 67, 70, 99, 150 }

Well, now it gives some idea, but let's move on and find out

Let's see how we can find out by plotting a box plot manually with 5 number summary technique.

We will find out the below values from our data

- Minimum
- First Quartile Q1
- Median
- Third Quartile Q3
- Maximum

Data =

{15, 15, 16, 17, 18, 19, 19, 20, 20, 20, 21, 23, 24, 25, 26, 35, 39, 41, 42, 52, 54, 67, 70, 99, 150 }

We can find out that

24 is the median of our dataset

(the value at $\frac{n}{2}$ th position for n = even and average of $\frac{n}{2}$ th and $\frac{n}{2}$+1 th element if n = odd)

n = number of elements

Data =

{15, 15, 16, 17, 18, 19, 19, 20, 20, 20, 21, 23, 24, 25, 26, 35, 39, 41, 42, 52, 54, 67, 70, 99, 150 }

Let's find out now the 1st and 3rd Quartile.

Percentile is calculated by $\frac{x^{th}}{100} \times (n+1)$

(where x is percentile and n = total number of values)

Hence $Q1 = \frac{25}{100} \times (25+1)$ = 6.5th index = 19

And $Q3 = \frac{75}{100} \times (25+1)$ = 19.5th index = 47

Data =

{15, 15, 16, 17, 18, 19, 19, 20, 20, 20, 21, 23, 24, 25, 26, 35, 39, 41, 42, 52, 54, 67, 70, 99, 150 }

Now let's define a lower fence and higher fence as:

Lower fence(L.F) = Q1 – 1.5 x IQR

Higher fence(H.F) = Q3 + 1.5 x IQR

Where IQR = Q3 – Q1 = 47 – 19 = 28

Hence replacing the values

L.F = 19 – 1.5 x 28 = -23

H.F = 47 + 1.5 x 28 = 89

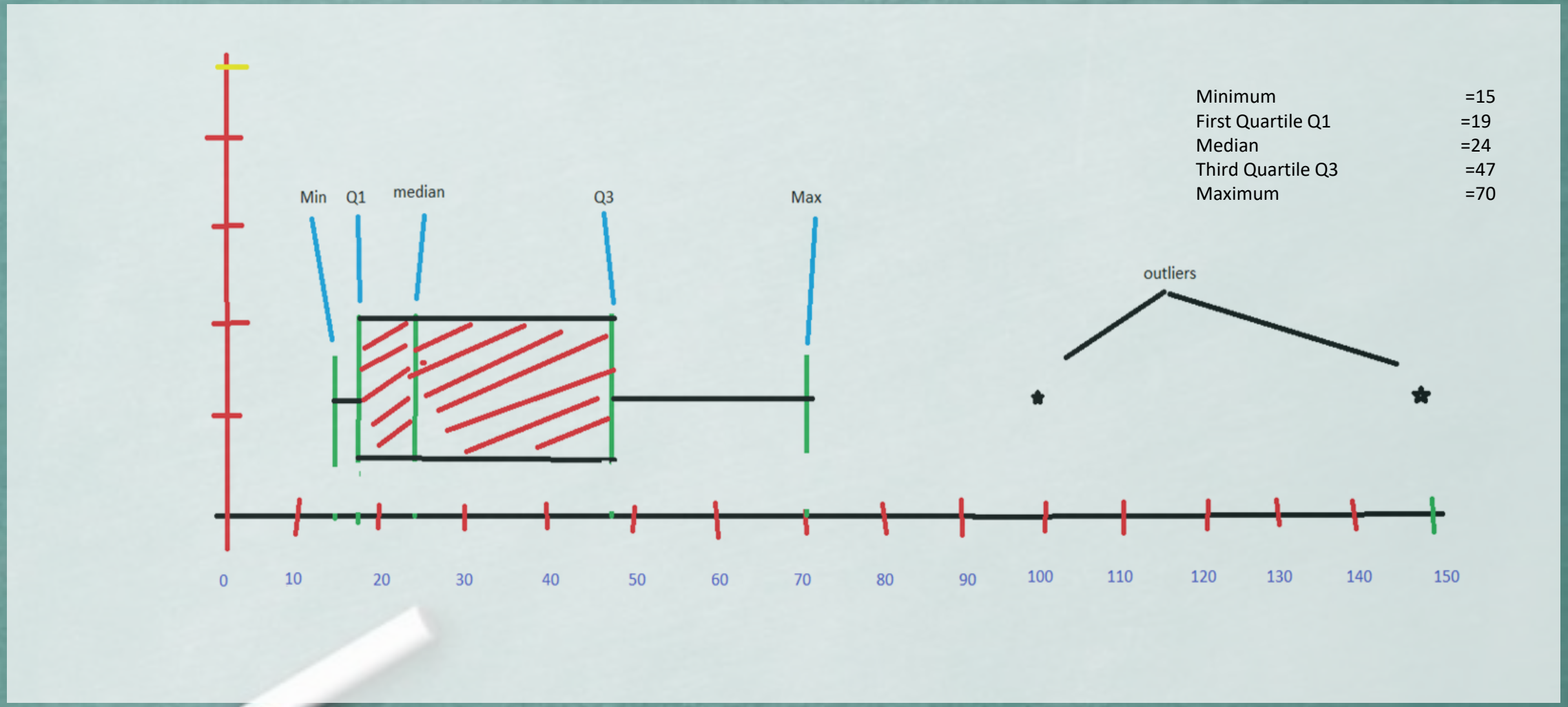
So, we get a range from -23 to 89, any values beyond this range of our dataset are outlier.

- Data =
- {15, 15, 16, 17, 18, 19, 19, 20, 20, 20, 21, 23, 24, 25, 26, 35, 39, 41, 42, 52, 54, 67, 70, 99, 150

- We get all our values as

- Minimum — 15 is the minimum value from our range -23 to 89
- First Quartile Q1 — 19
- Median — 24
- Third Quartile Q3 — 47
- Maximum — 70 is the maximum value from our range -23 to 89

Now let's plot this values along X axis and make a box between Q1 and Q3,



Yeah!! Our box plot is done.

A snap of box plot of same through python is shown here