Histograms and Distributions

So far…

Scatterplots - pairs of attributes

How does a single attribute behave?

Example

Life expectancy in GapMinder data set

Example

Life expectancy in GapMinder data set

Varies between countries

1967 Average Life Expectancy in Morocco, Bangladesh = 50.34, 43.45

Example

Life expectancy in GapMinder data set

Varies between countries

1967 Average Life Expectancy in Morocco, Bangladesh = 50.34, 43.45

Varies over time

Average life expectancy in Namibia in 1952, 2007 = 41.73, 52.90

Aside - life expectancy for country itself an estimate

Varies according to

Year of birth

Socioeconomic group (wealthy/poor)

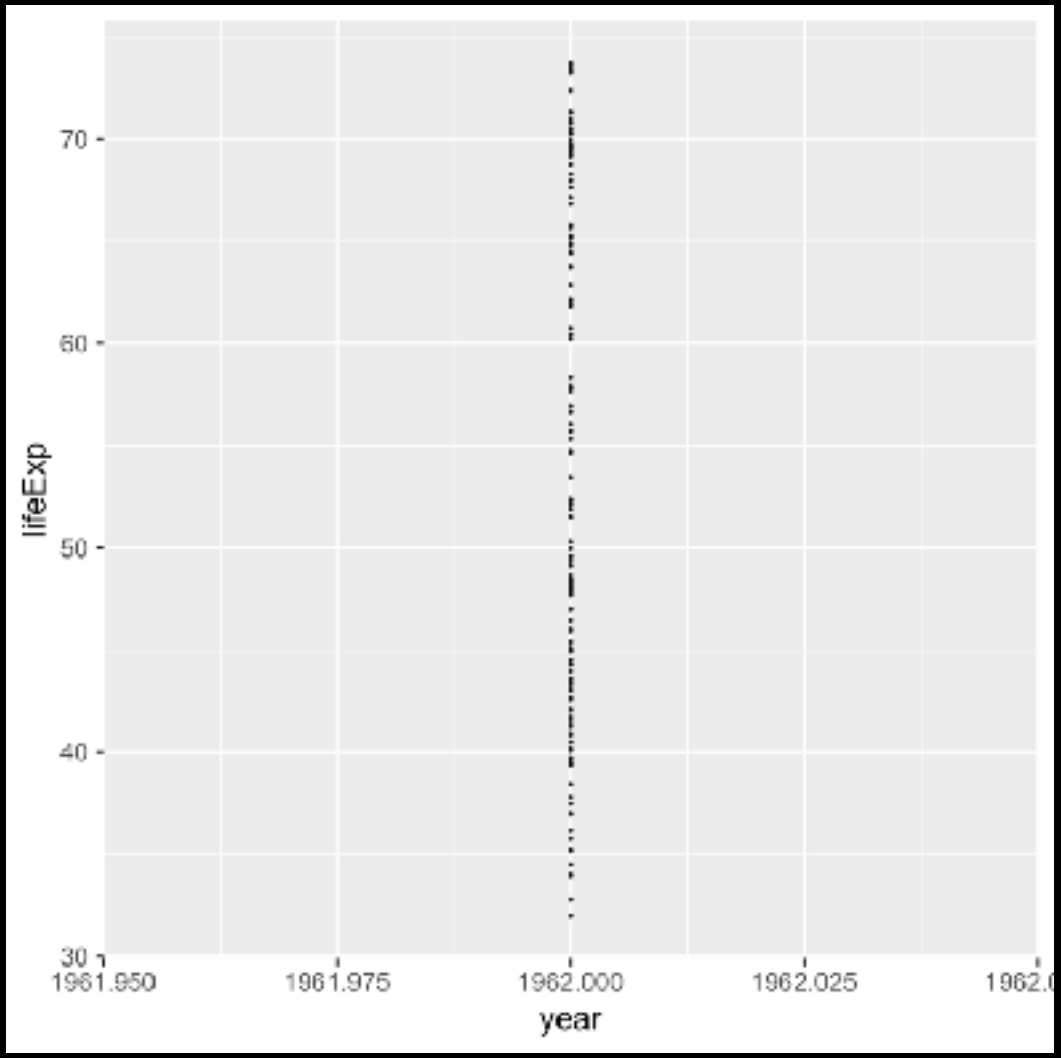
Geographical location

First attempt - just plot points

Example - Life Expectancy in 1962 for all countries

First attempt - just plot points

Example - Life Expectancy in 1962 for all countries

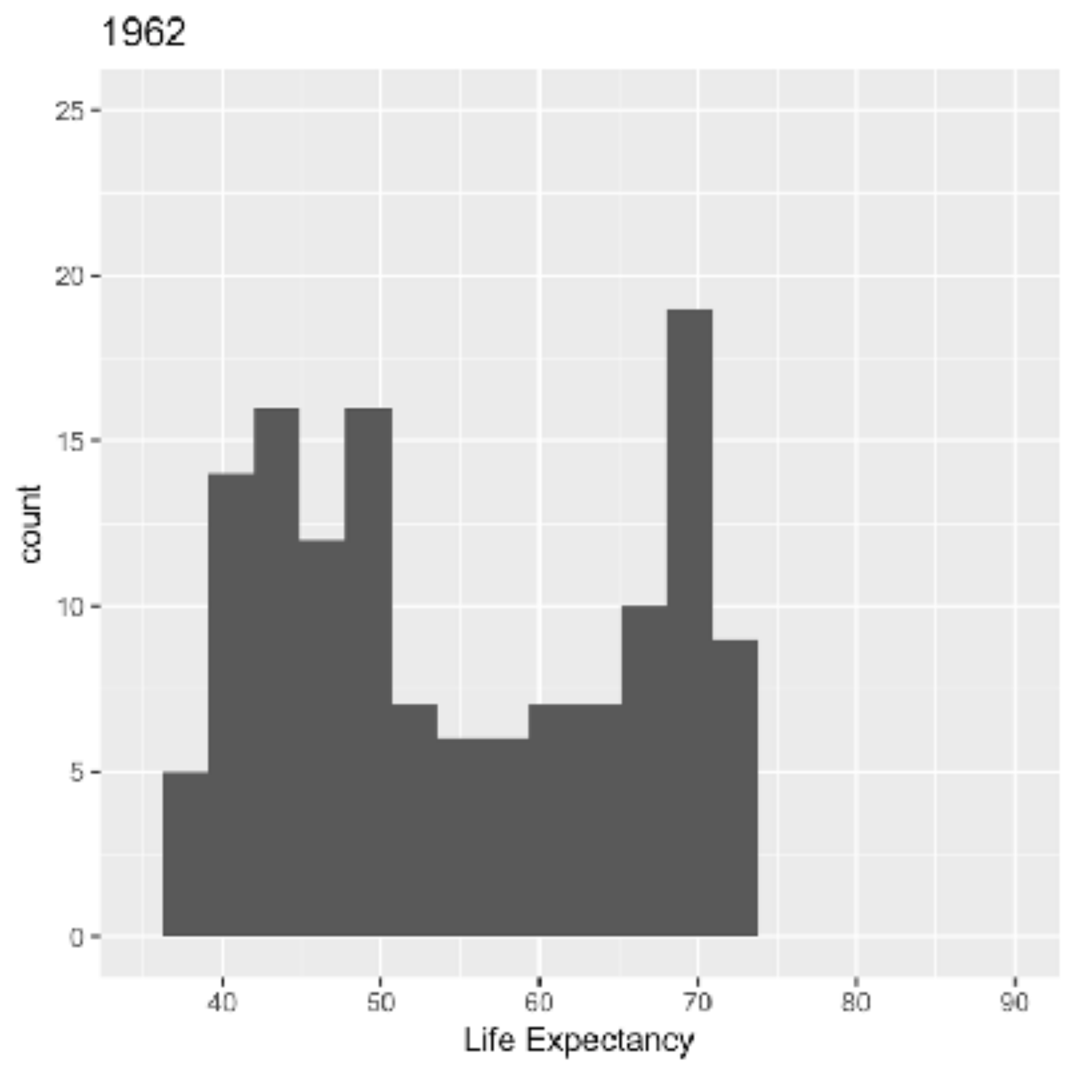


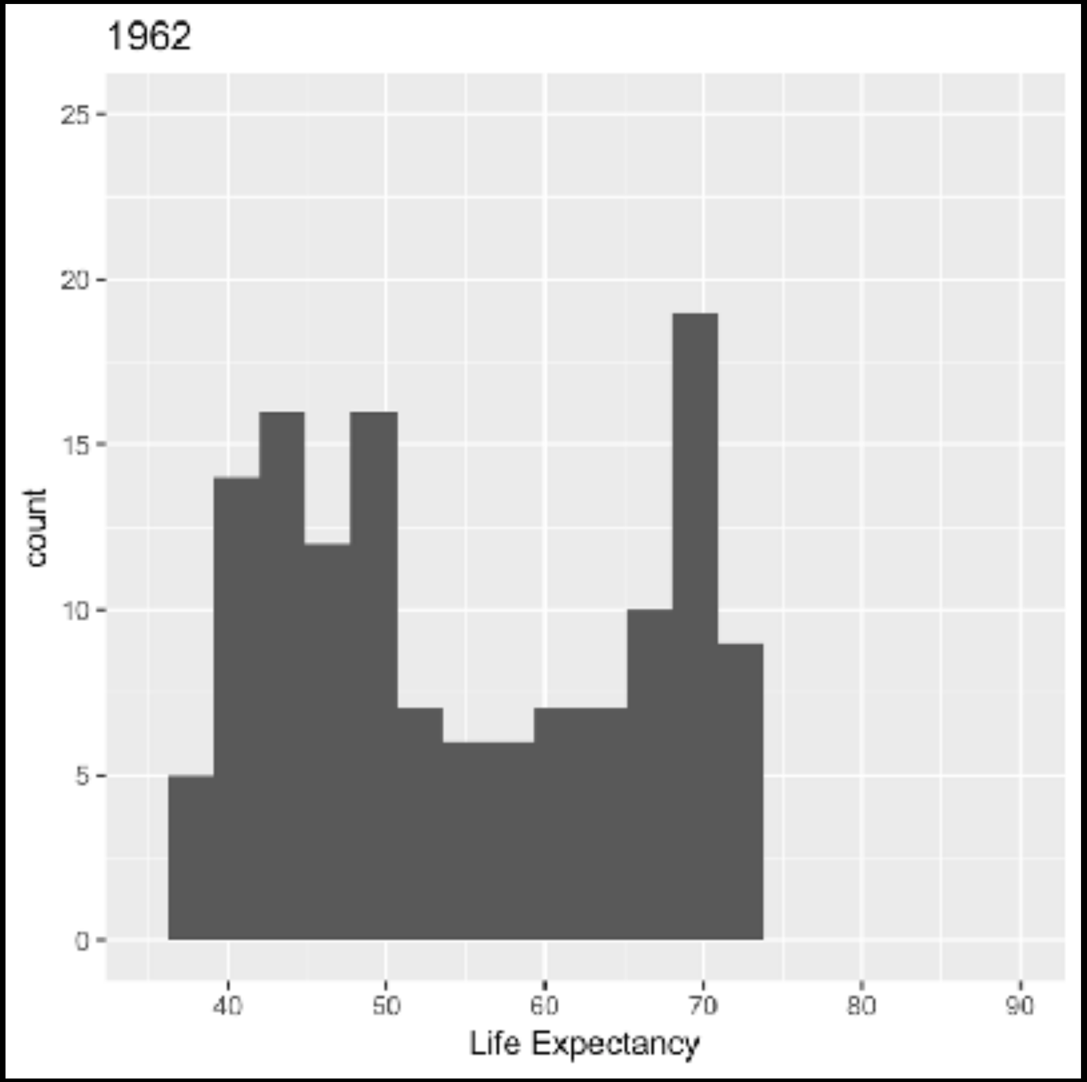
Histogram

Means of converting Ordinal to Categorical data

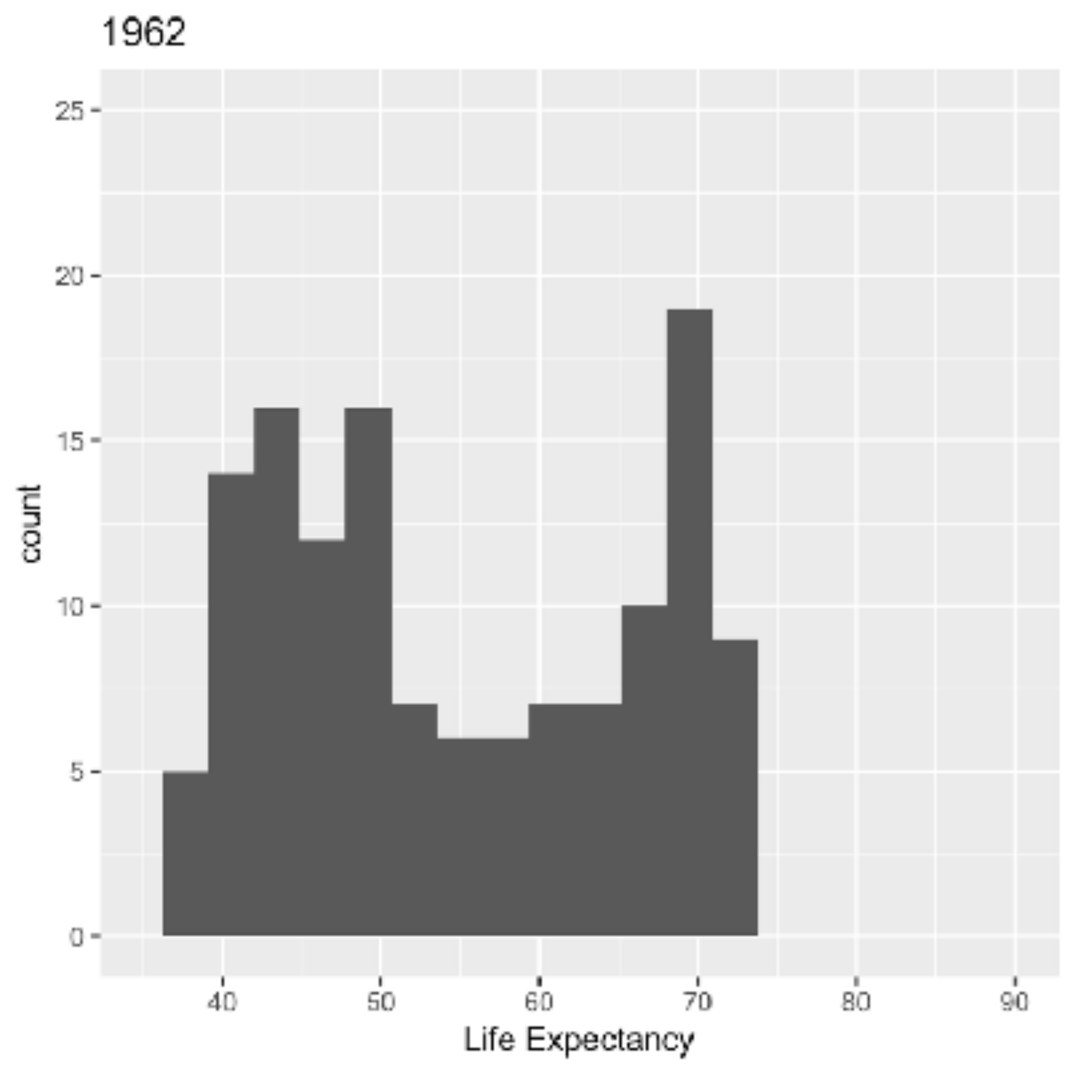
Create ‘bins’ -
how many countries have an Life Expectancy of 40-45 etc.

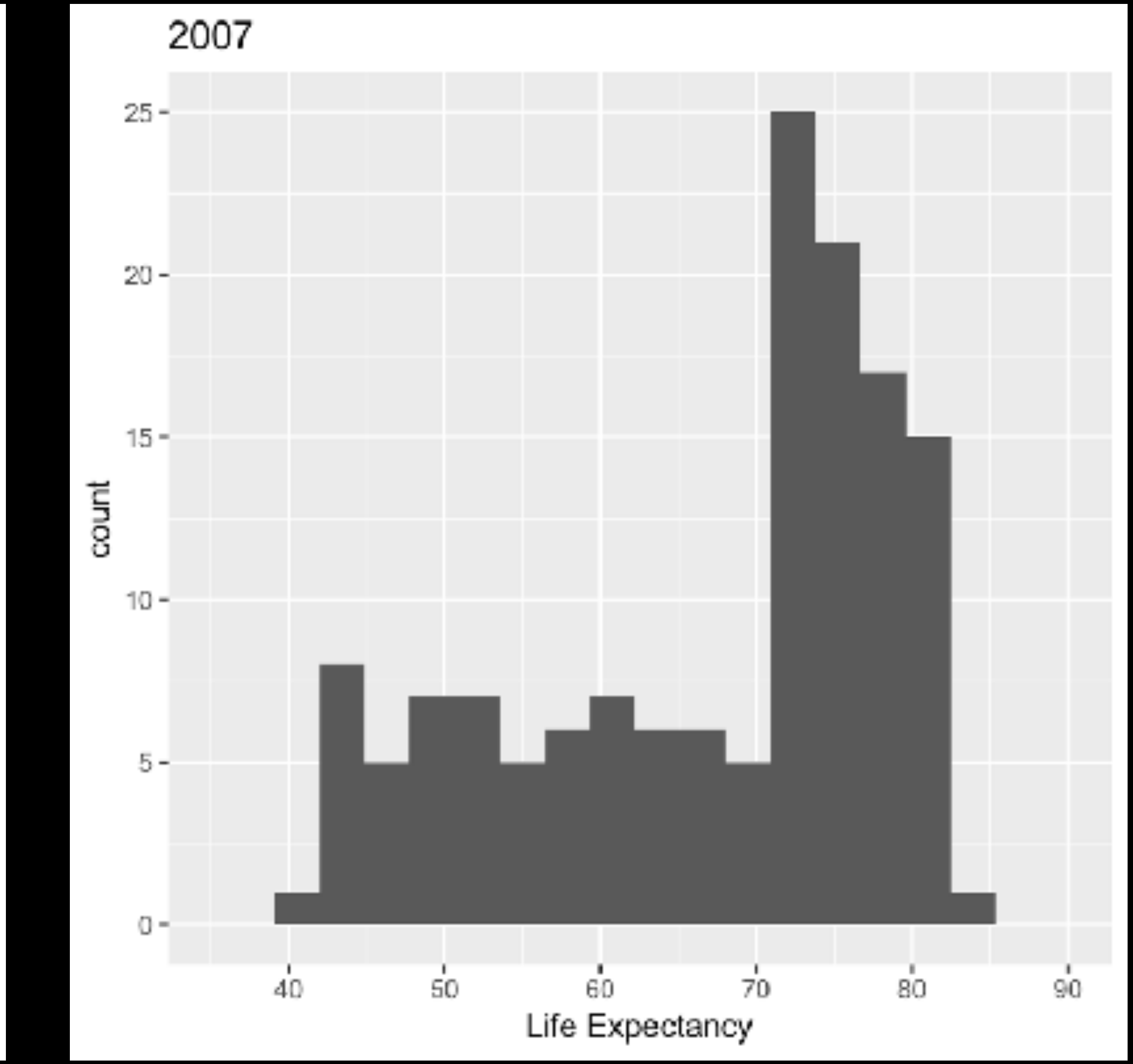
Create bar plot of counts


1962
count
25
20
15
10
5
0
40
50
60
70
80
90
Life Expectancy

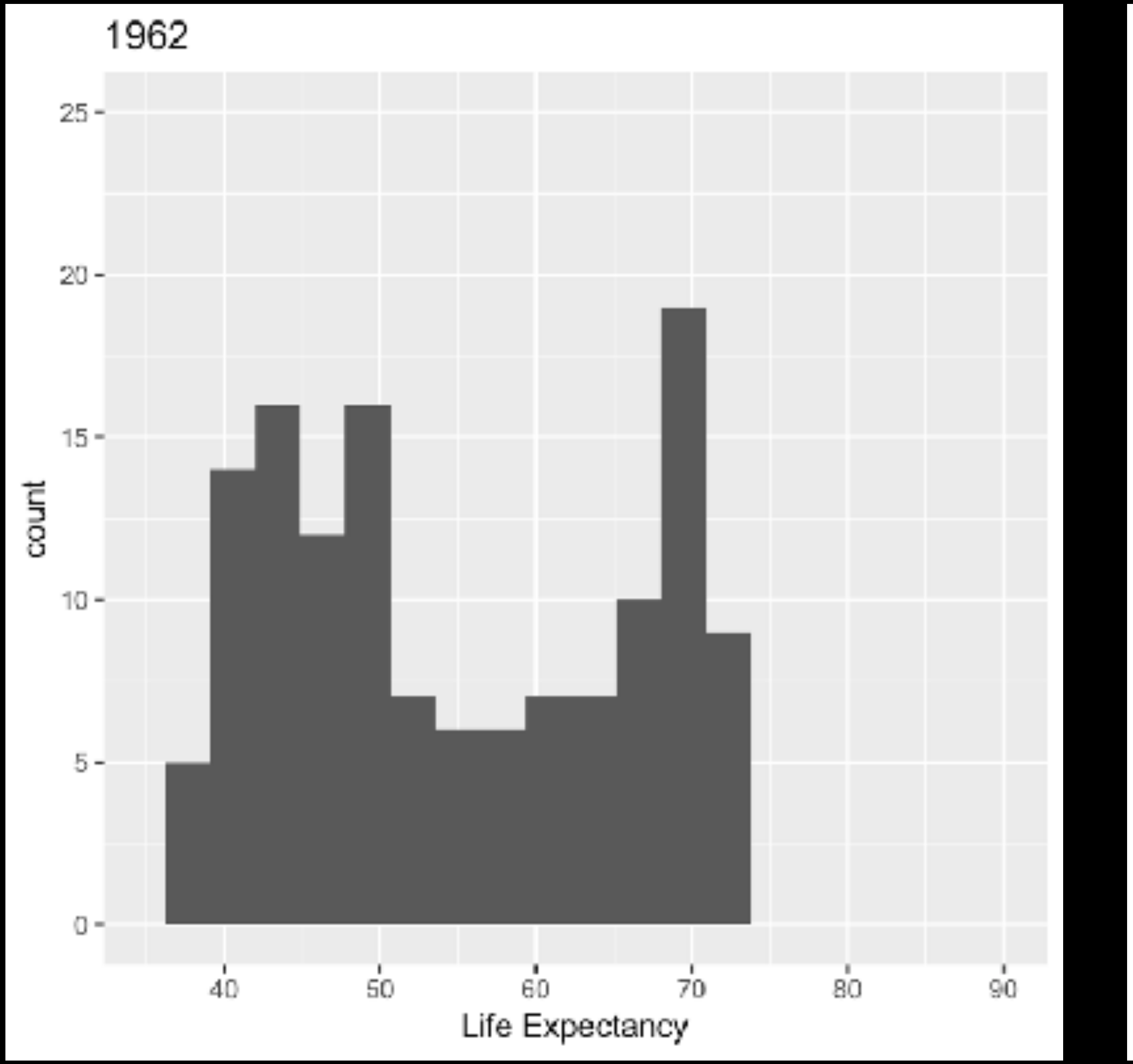


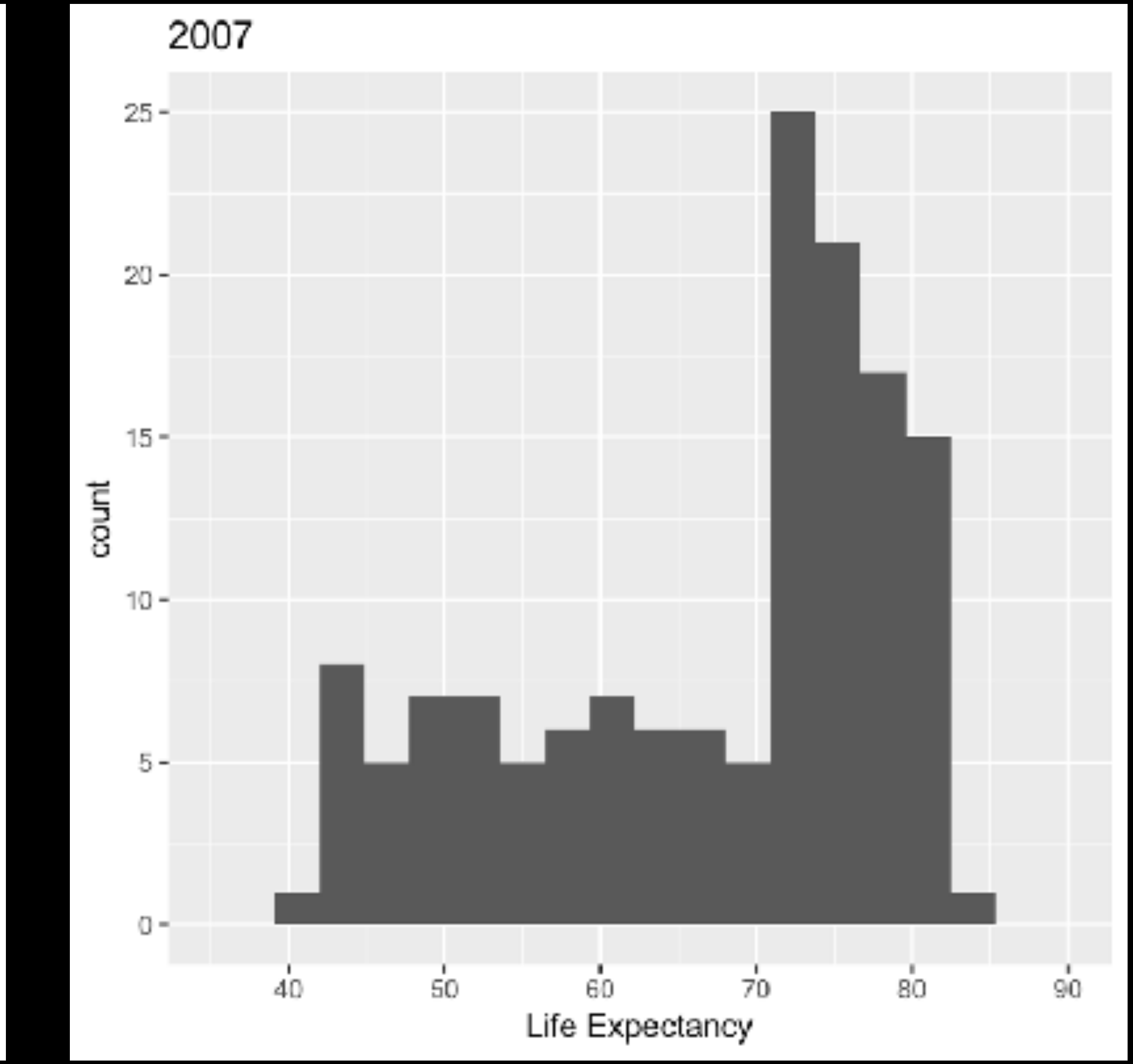
What are seeing here?





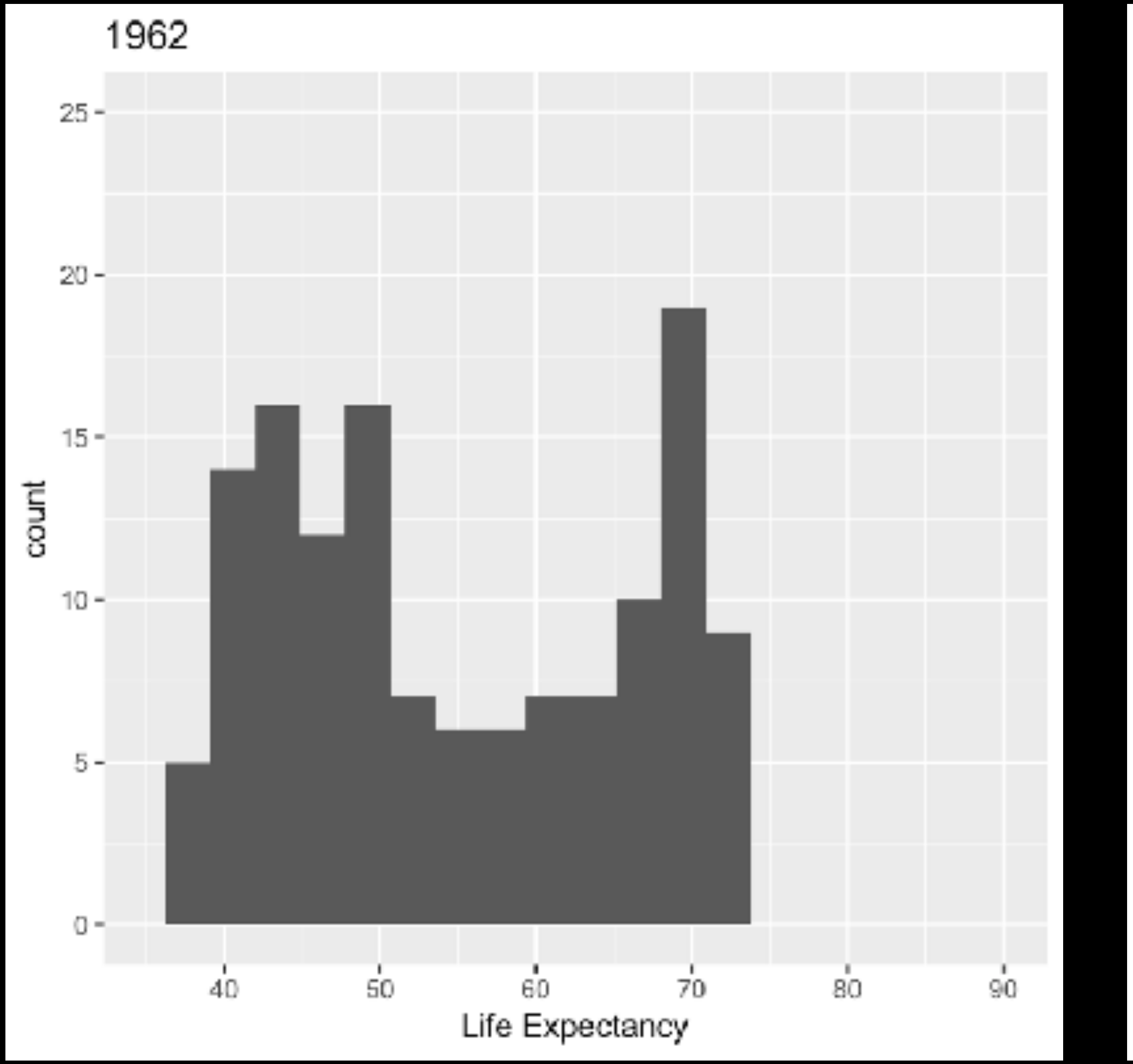
What are seeing here?

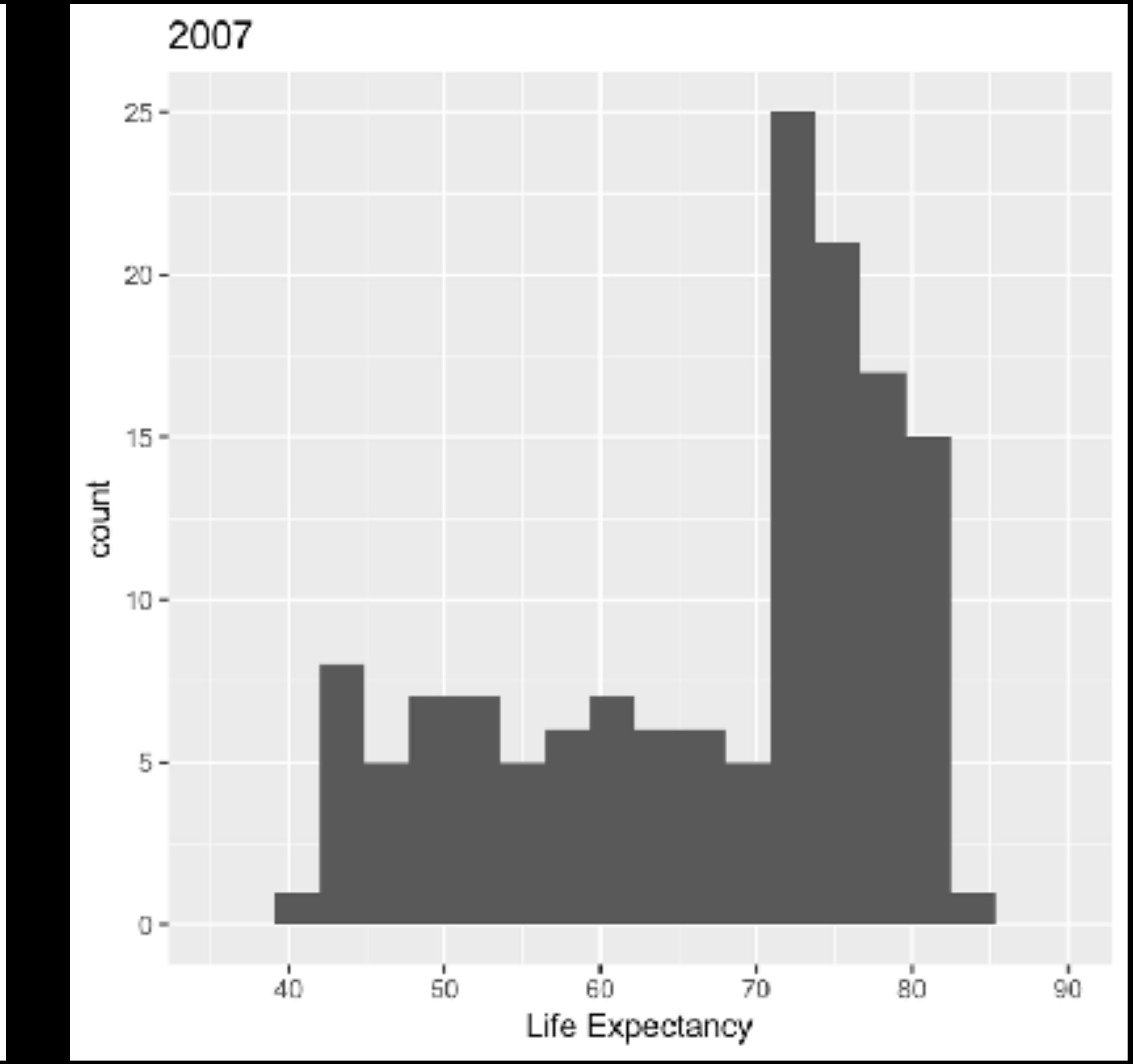




What are seeing here?

What’s different?





What are seeing here?

What’s different?

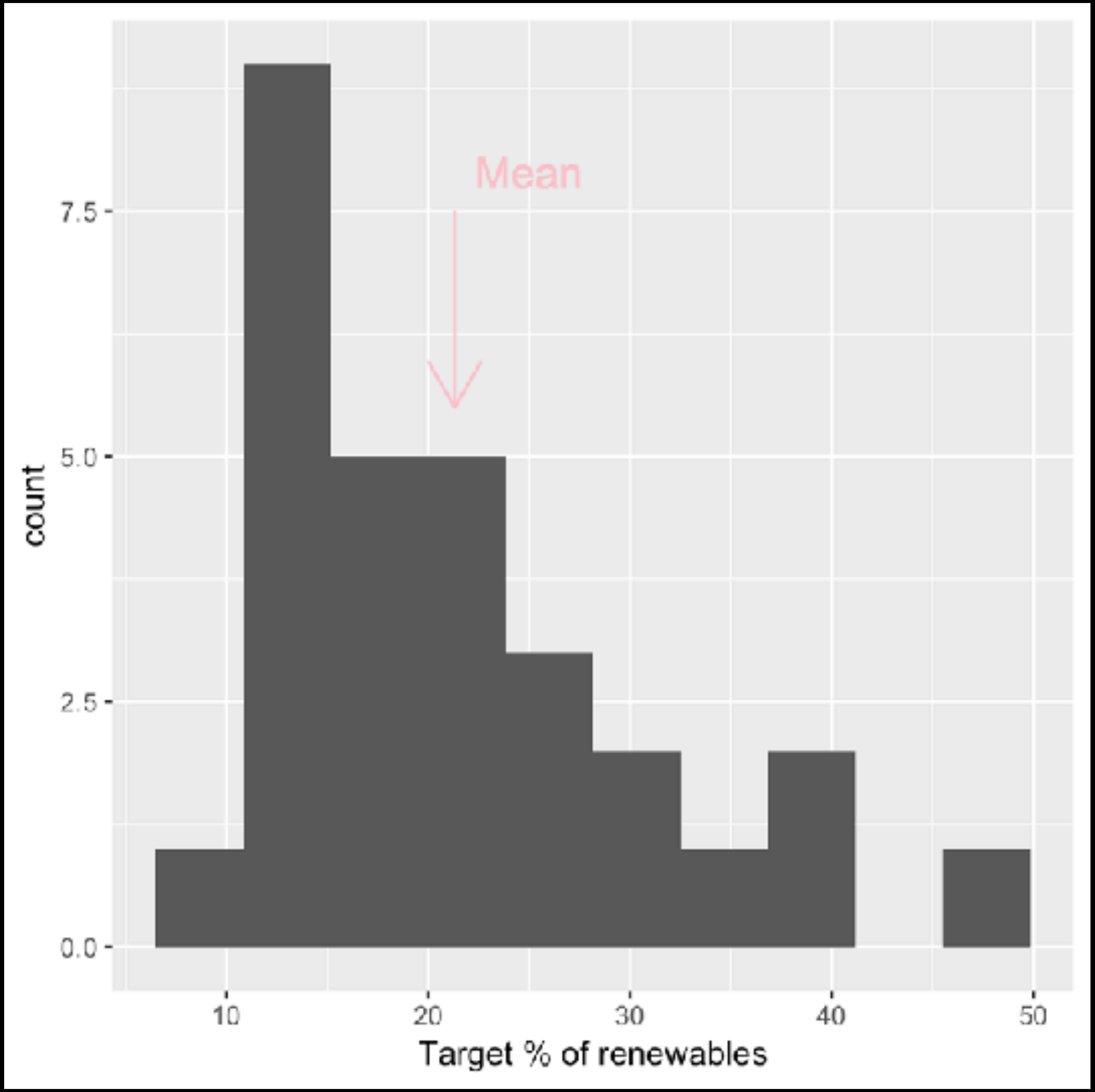
What did I do to make this comparison?

Number of bins

Too many - just see noise

Too few - don’t see any features

What about this ?

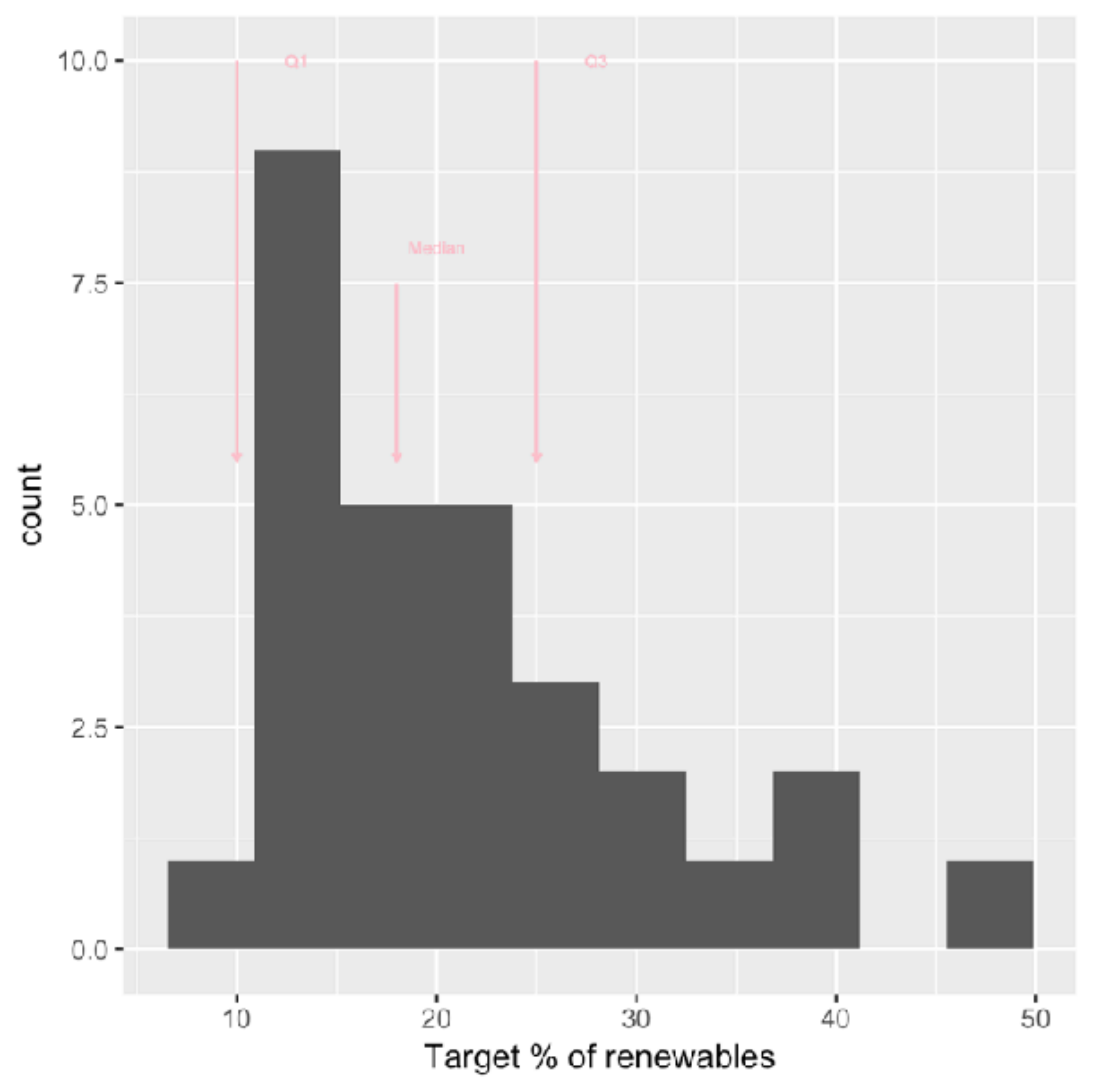


Better to think of quartiles

1st quartile (Q1) - value of y which is greater than 25% of the y_i

Median - value of y which is greater than 50% of the y_i

3rd quartile (Q3) - value of y which is greater than 75% of the y_i


Q1
Q3
Median
10.0
7.5
5.0
2.5
0.0
count
10
20
30
40
50
Target % of renewables

What if we want to compare distributions?

Life expectancy of countries between years?

What if we want to compare distributions?

Life expectancy of countries between years?

Could do a facet plot of histograms

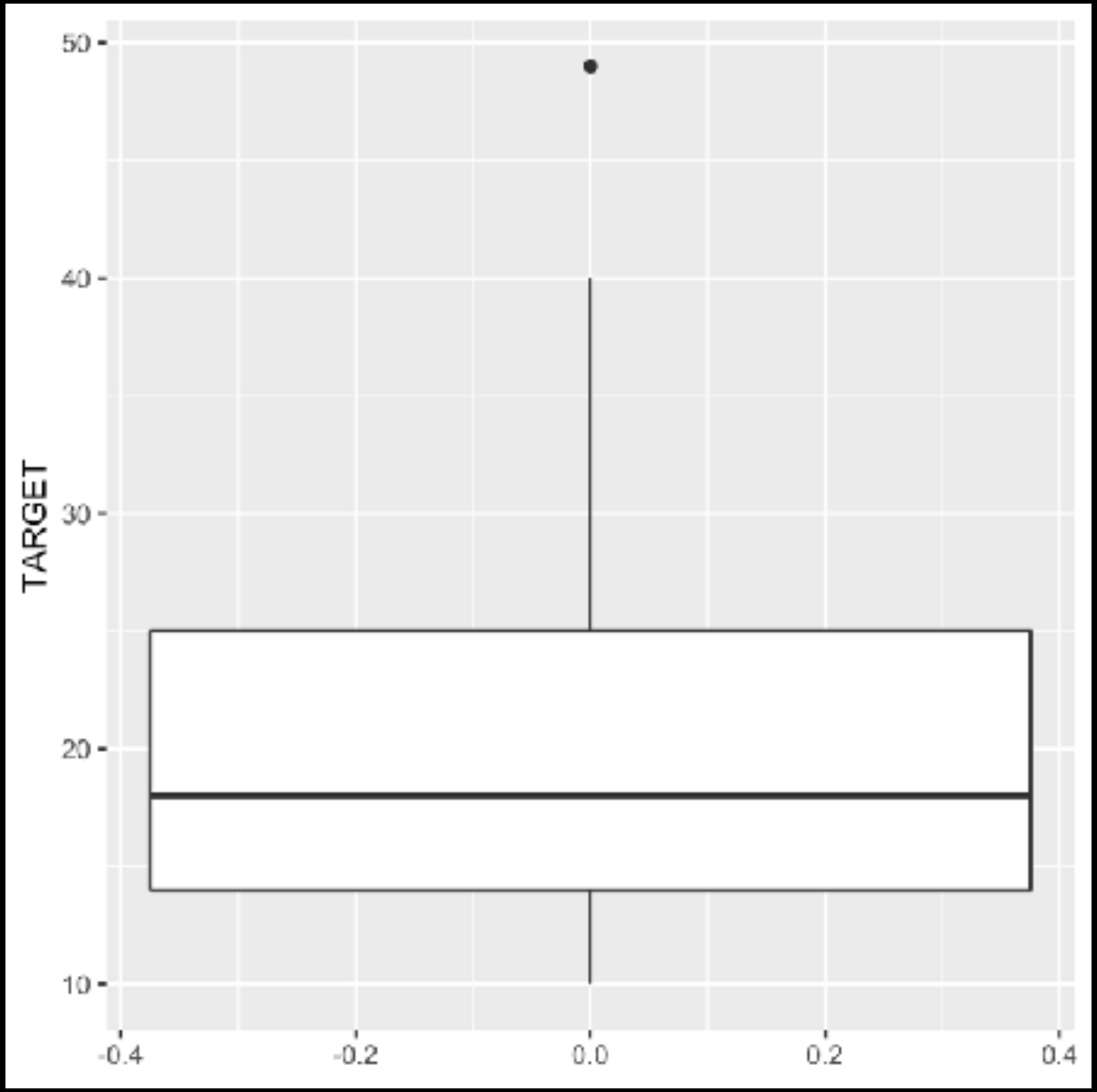
What if we want to compare distributions?

Life expectancy of countries between years?

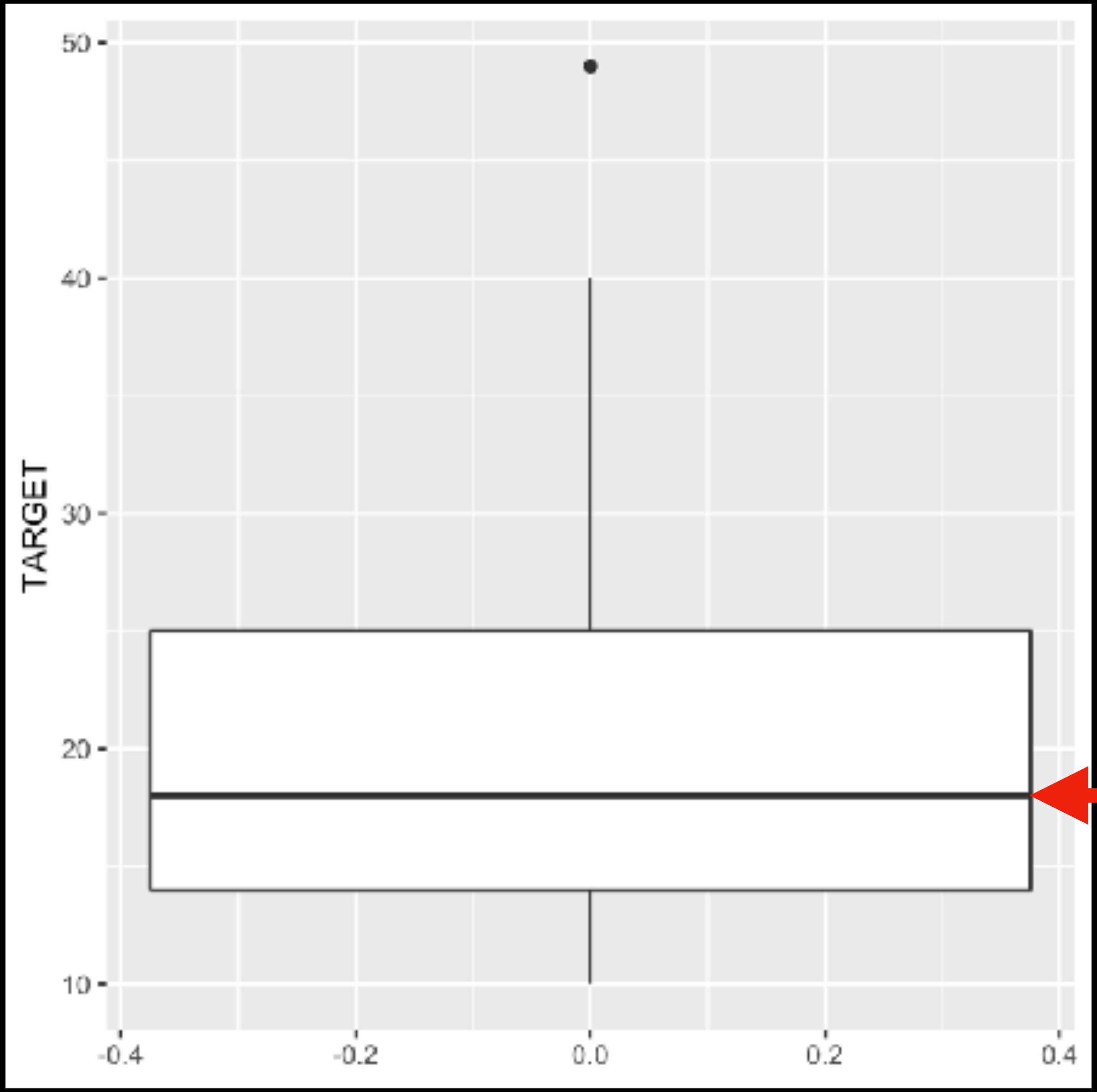
Could do a facet plot of histograms

But can also other comparisons

Box plot of renewables target

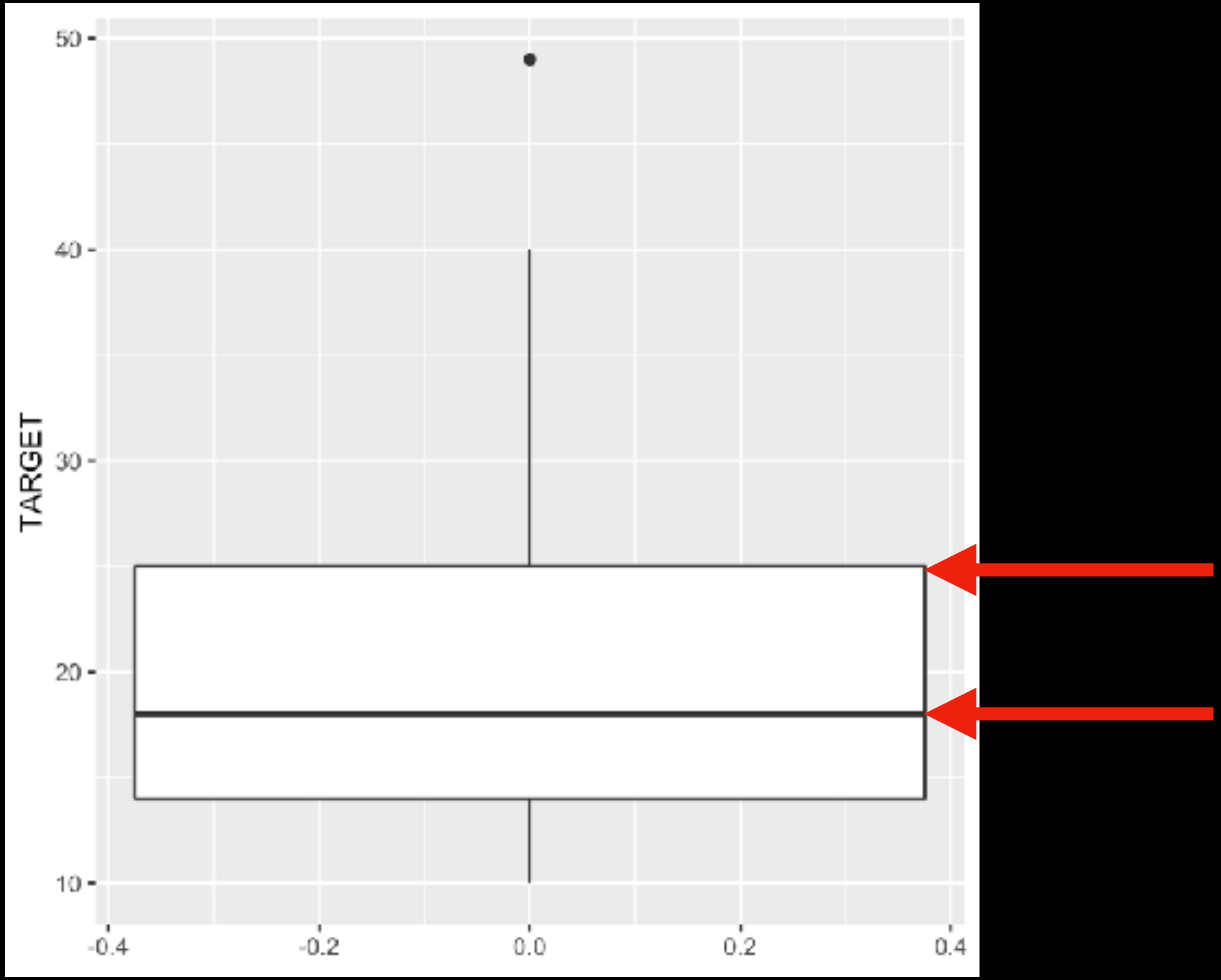


Box plot of renewables target



Median

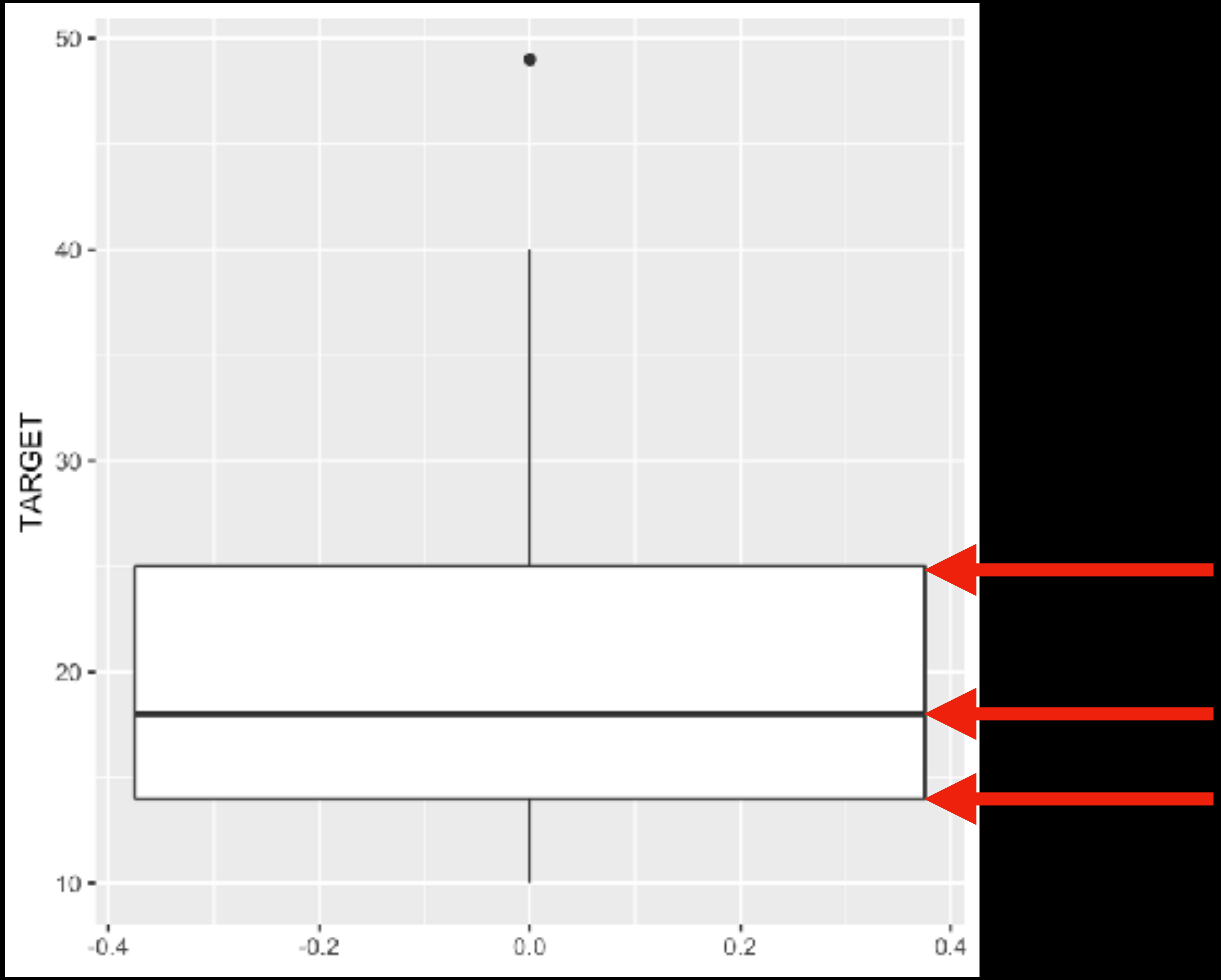
Box plot of renewables target



Q3

Median

Box plot of renewables target

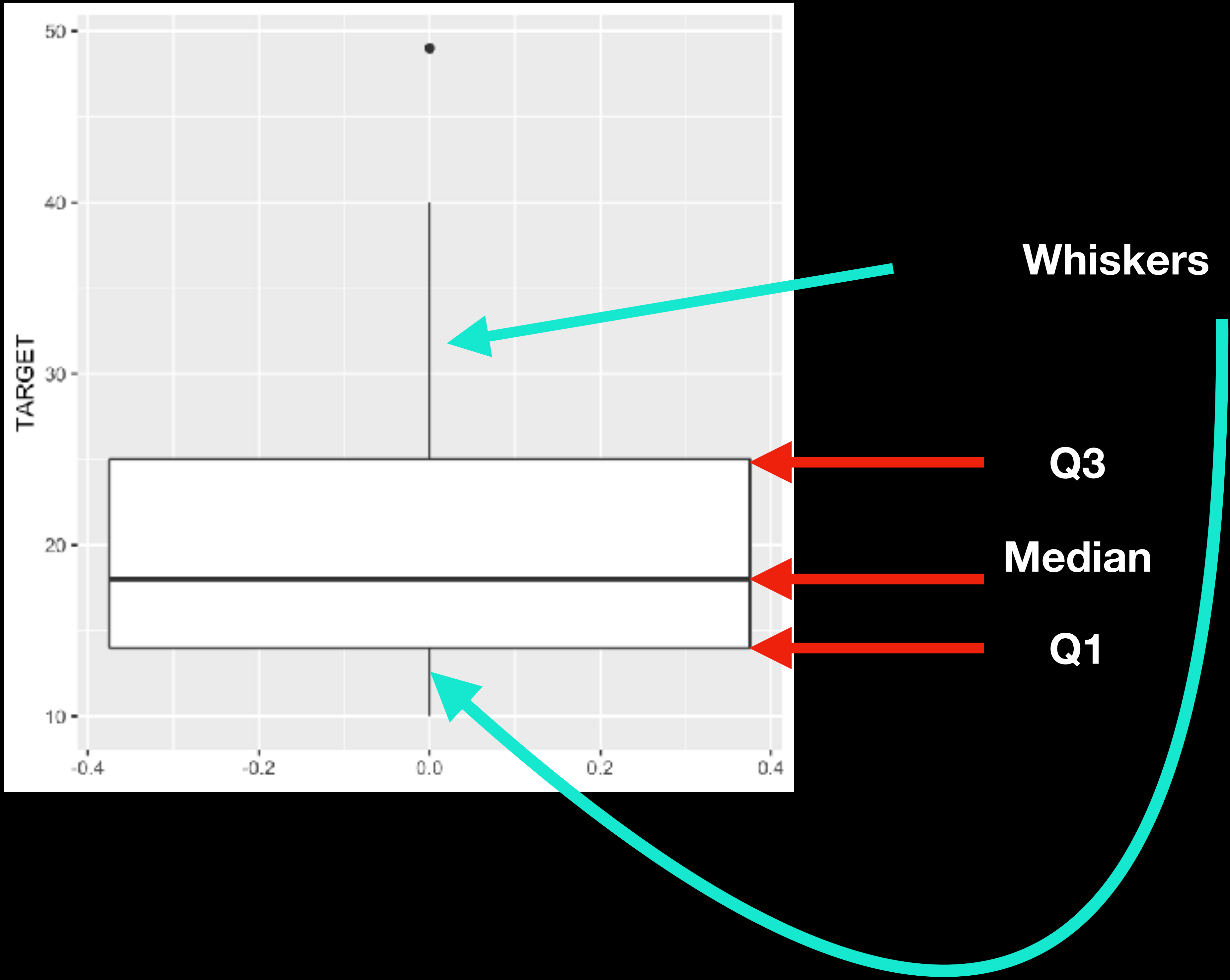


Q3

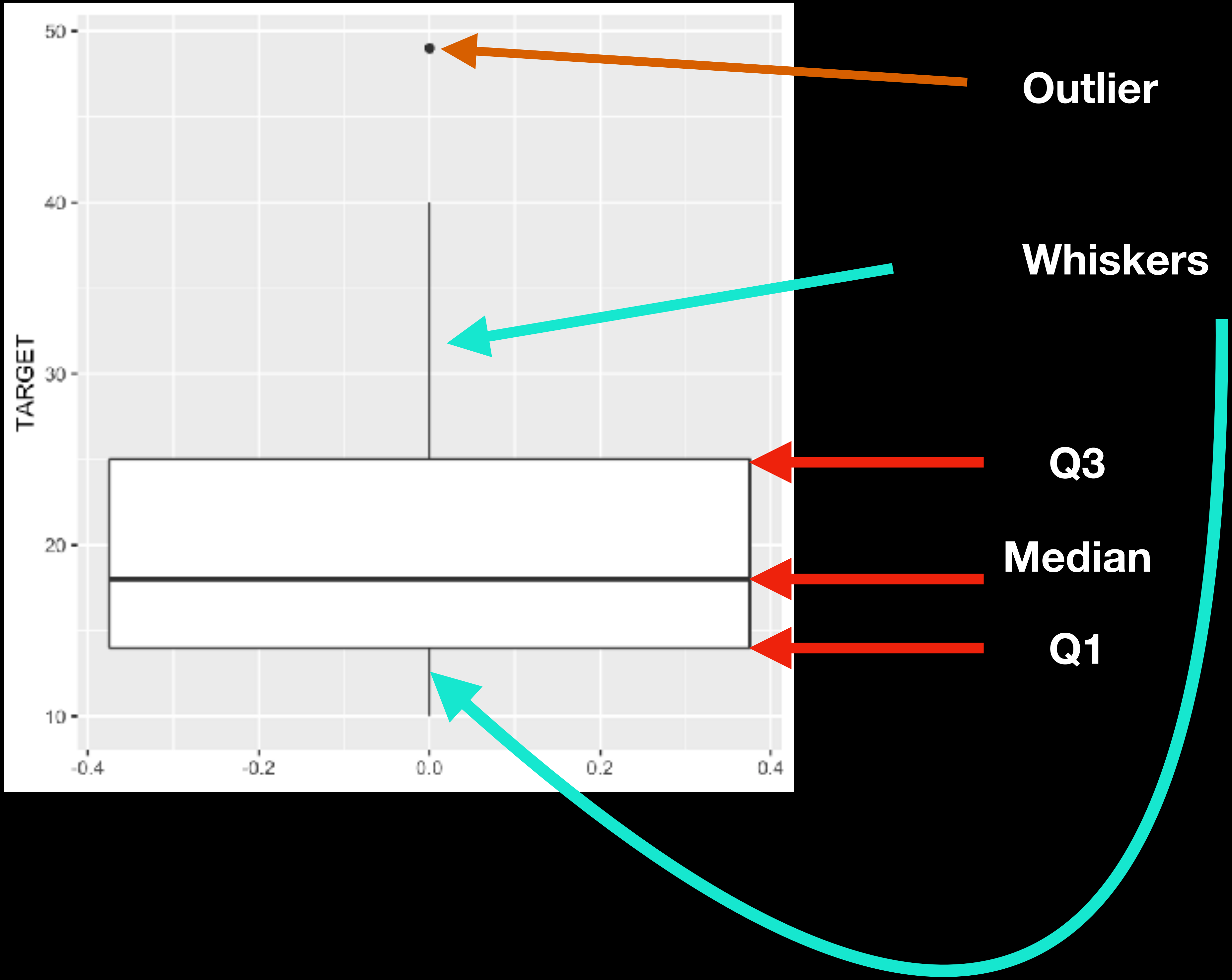
Median

Q1

Box plot of renewables target



Box plot of renewables target



Outlier

Whiskers

Q3

Median

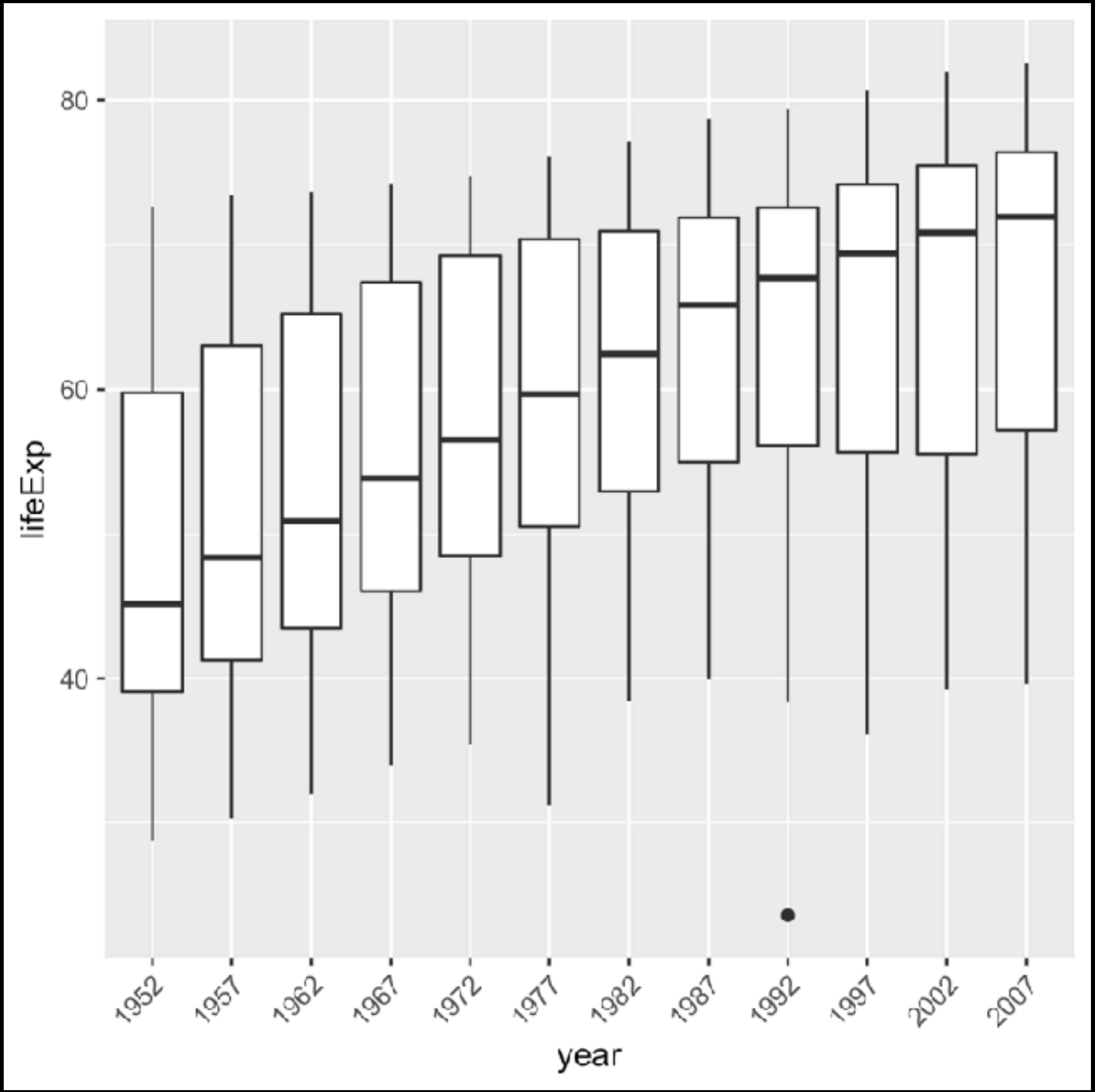
Q1

Whisker length

No bigger than 1.5 x IQR

IQR = Q3 - Q1

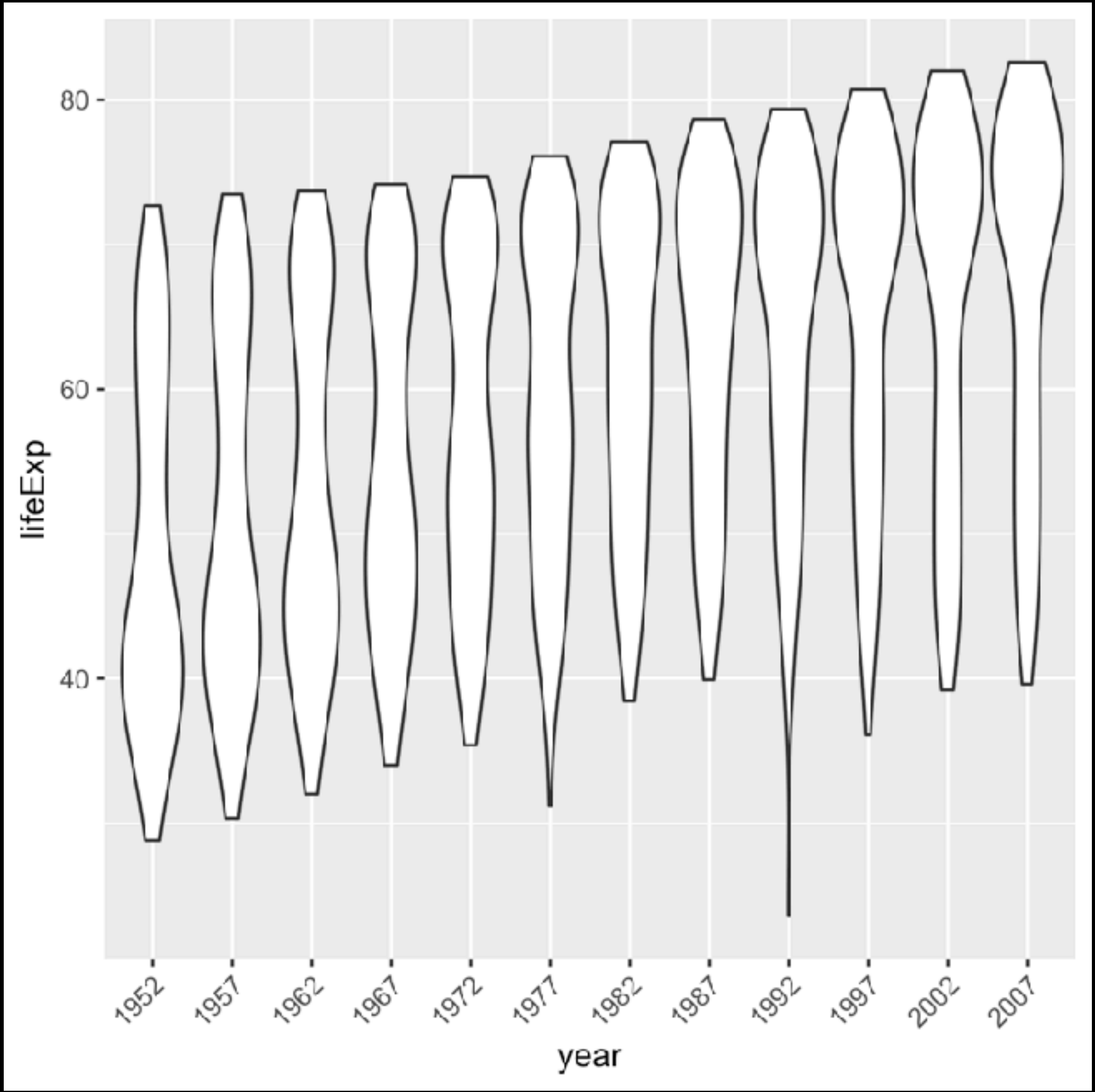
If any data is greater than maximum whisker length
then plotted as point


80
60
40
lifeExp
1952
1957
1962
1967
1972
1977
1982
1987
1992
1997
2002
2007
year

But Life Expectancy isn’t unimodal !

Try Violin plots

Estimate distribution - width of shape indicates height of distribution


80
60
40
lifeExp
1952
1957
1962
1967
1972
1977
1982
1987
1992
1997
2002
2007
year