STA 1013 : Statistics through Examples

Lecture 9: Visualization for quantitative variables

Hwiyoung Lee

September 20, 2019

Department of Statistics, Florida State University

Overview

1. Histogram

2. Boxplot

Data visualization

- Categorical variables : Bar charts, Pie charts
- Quantitative variable :
  - Stem and Leaf plot : Not appropriate for large data
  - The two most common types of graphics are Histogram and Box-plot

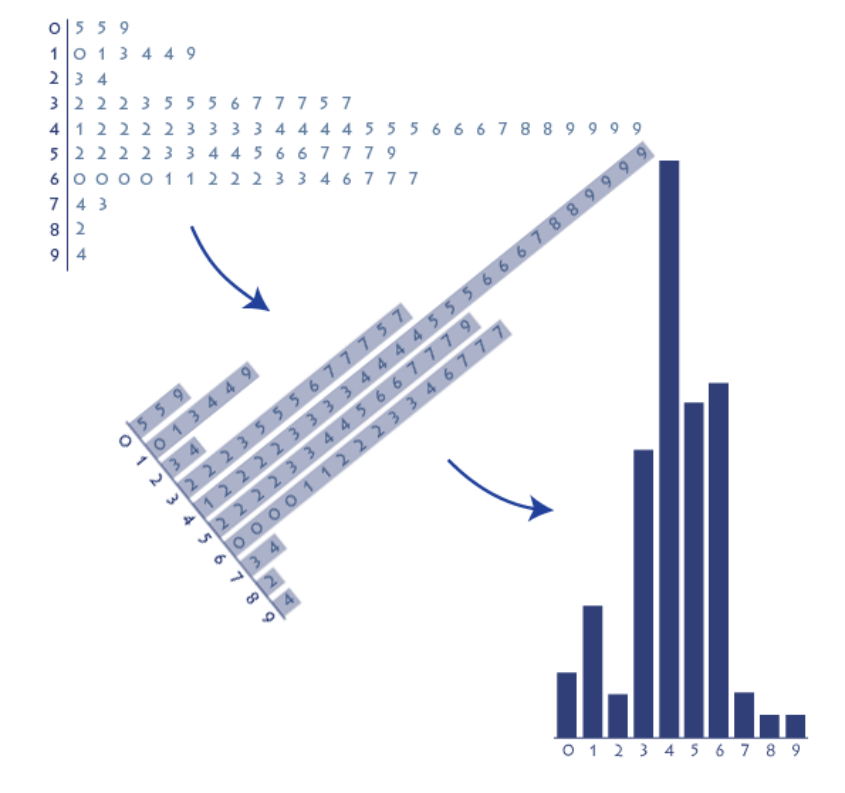
Histogram

Histogram

- A histogram is essentially a bar chart in which the data categories (bin) are quantitative
- The bars in a histogram must follow the natural order of the numerical categories
- The widths of the bars must be equal, and they must have a specific meaning
- The bars in the histogram touch each other because there are no gaps between the categories

Histogram

- Histogram is the generalized version of Stem and Leaf plot



How to draw a Histogram

- The first step is to bin the data
- Count the frequencies
- Do same steps in a bar chart

Example

Oscar-Winning Actress

The following data show the ages (at the time when they won the award) of all Academy Award–winning actresses through 2012, sorted into age order. Display the data using 10-year bins.

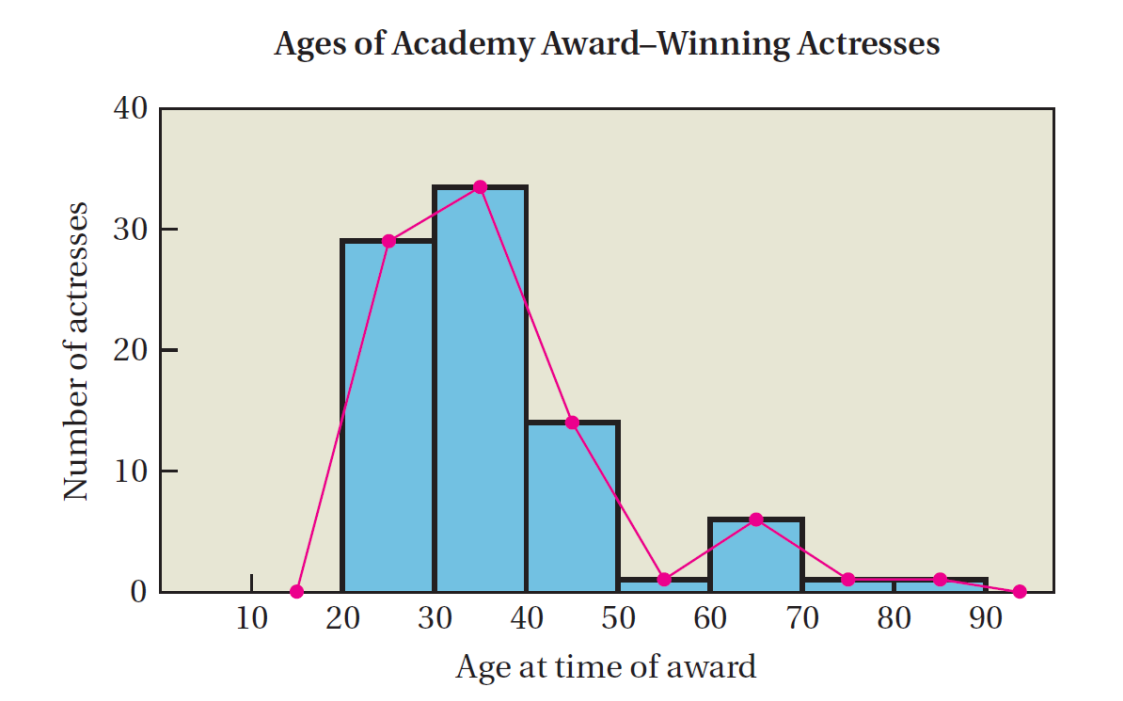
Ages of Actresses at Time of Academy Award (through 2012)

21	22	24	24	25	25	25	25	26	26	26
26	26	27	27	27	27	28	28	28	28	29
29	29	29	29	29	29	29	30	30	30	31
31	31	32	32	32	32	33	33	33	33	33
33	34	34	34	35	35	35	35	35	36	36
37	37	38	38	38	38	39	39	40	41	41
41	41	41	42	42	45	45	48	49	49	54
60	61	61	61	62	63	74	80			

Source: Academy of Motion Picture Arts and Sciences.

Example : Oscar-Winning Actresses

Age	Freq
20 - 29	29
30 - 39	34
40 - 49	13
50 - 59	1
60 - 69	6
70 - 79	1
80 - 89	1



Example : Weight data

Draw the histogram of the weight data given below

195.6	200.4	165.6	165.3	191.7	169.3	153.2
189.5	170.4	149.3	185.3	150.3	179.6	160.3
198.5	163.2	166.3	197.3	201.3	168.2	198.4

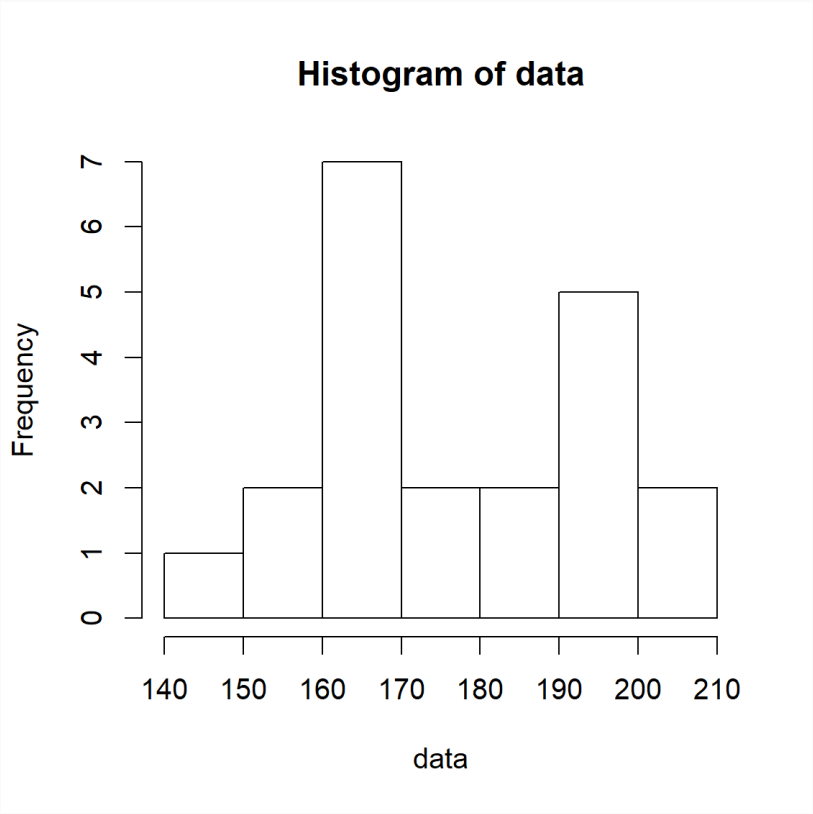
Use the following bins :

$$[140, 150), [150, 160), [160, 170), [170, 180), [180, 190), [190, 200), [200, 210)$$

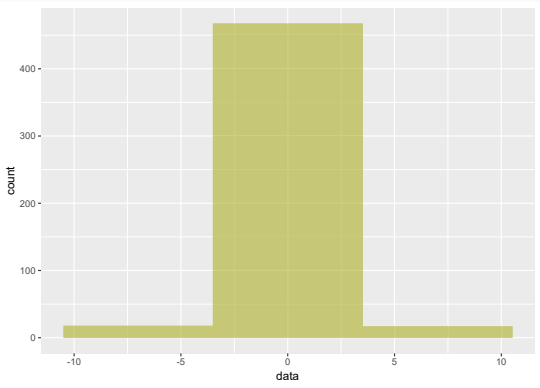
Histogram of the weight data

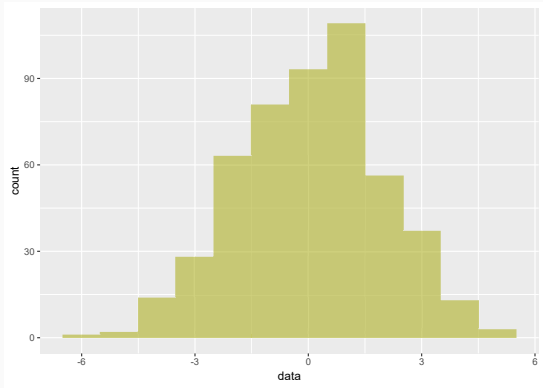
Binned weight data

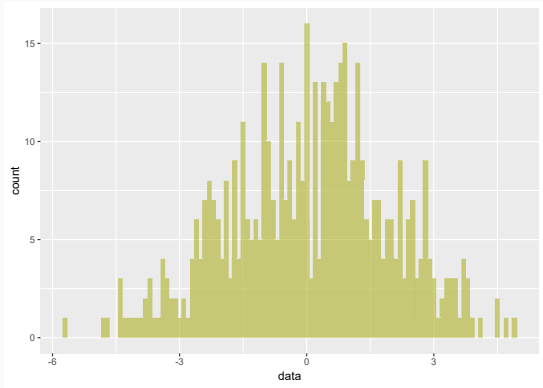
Weight	Count
140 ∼ 149.9	1
150 ∼ 159.9	2
160 ∼ 169.9	**7**
170 ∼ 179.9	2
180 ∼ 189.9	2
190 ∼ 199.9	5
200 ∼ 209.9	2



Bin number





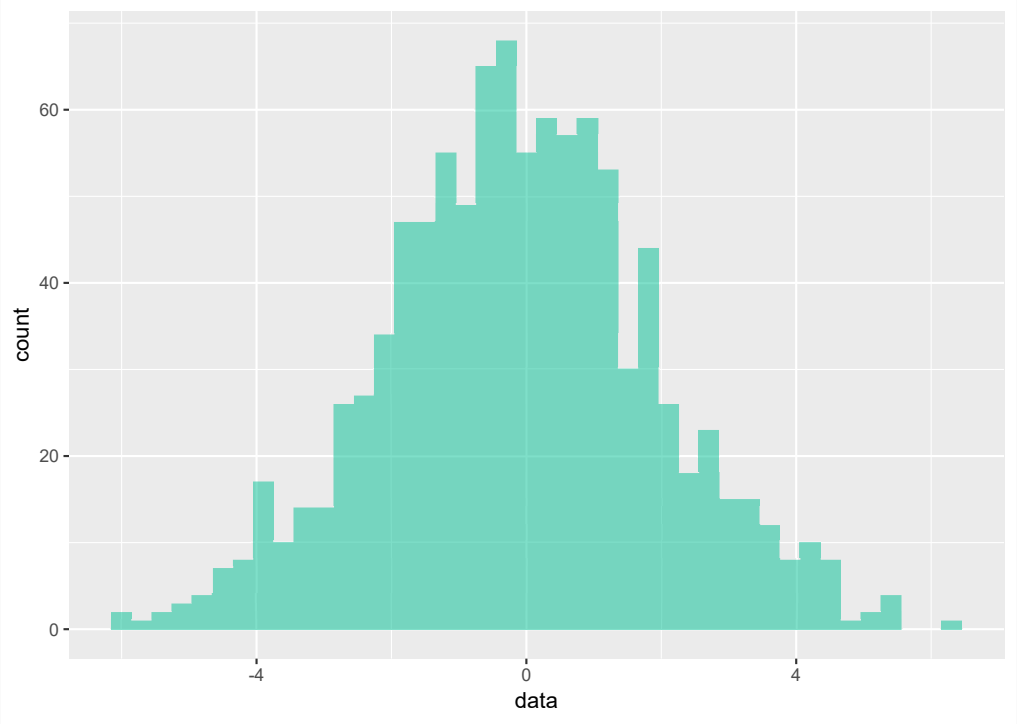


Bin width (Number of bins) is an important issue

- Large width of bin → small numbers of bins
  : Can't detect the shape of distribution
- Small width of bin → Large numbers of bins
  : Too wiggly histogram

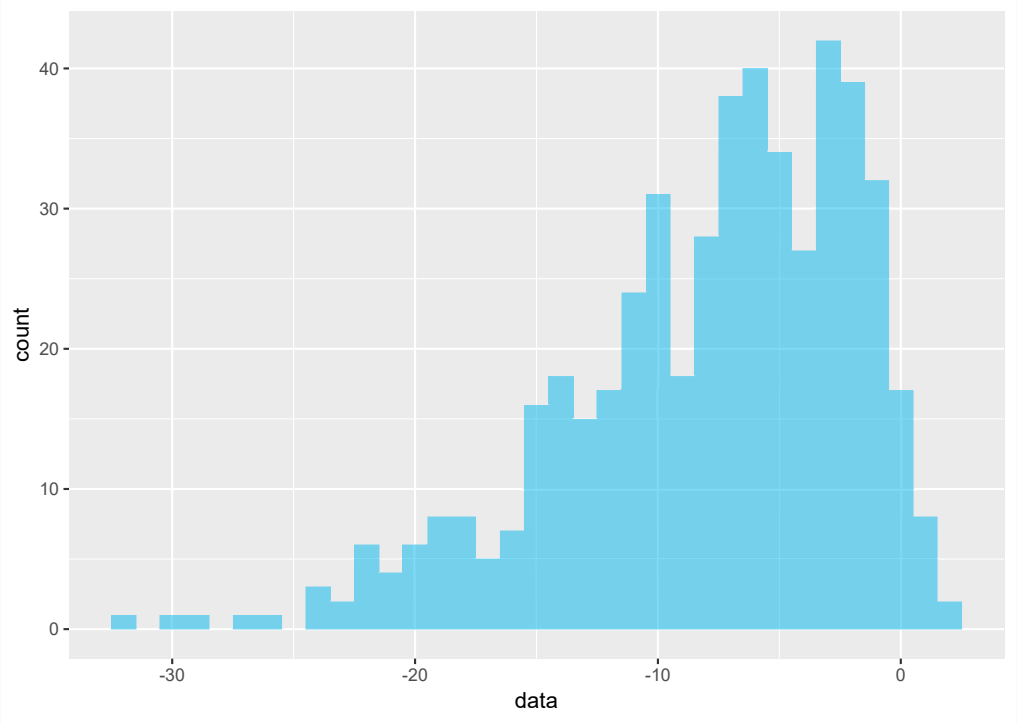
Choosing between 5 and 20 bins is recommended

Shape of data (Symmetric)



- Bell shape
- Left half is a mirror image of its right half
- Mode = Median = Mean

Shape of data (Left Skewed)



- Values are more spread out on the left side
- Values are concetrated on the right side (large value)
- Mode $>$ Median $>$ Mean

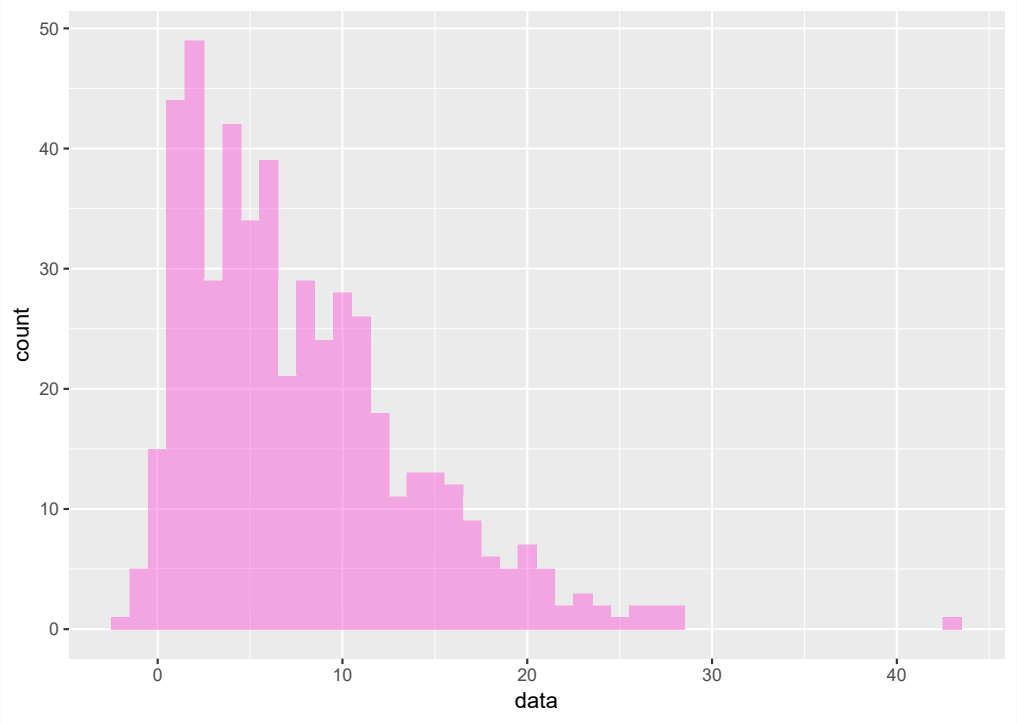
Left Skewed data

Skewness affects the relative positions of the mean, median, mode

Left skewed :

- By definition, the mode is at the peak in a single-peaked distribution (Located on the right side)
- A left-skewed distribution pulls both the mean and median to the left of the mode (meaning to values less than the mode)
- Outliers at the low end of the data set make the mean less than the median

Shape of data (Right Skewed)



- Values are more spread out on the right side
- Values are concetrated on the left side (small value)
- Mode $<$ Median $<$ Mean

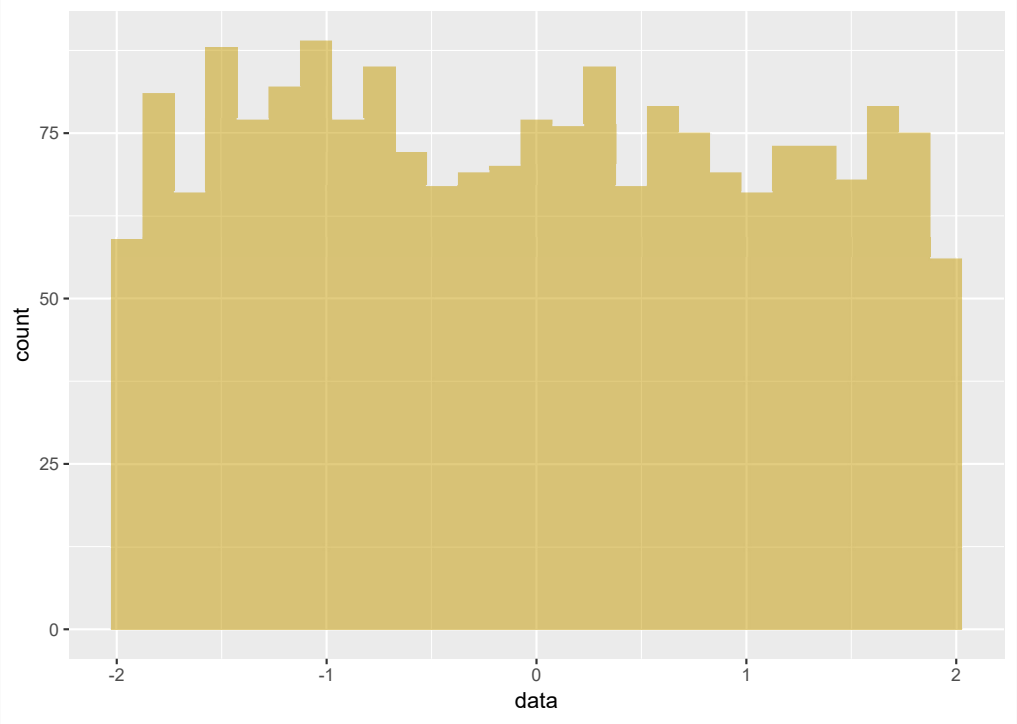
Right Skewed data

Skewness affects the relative positions of the mean, median, mode

Right skewed :

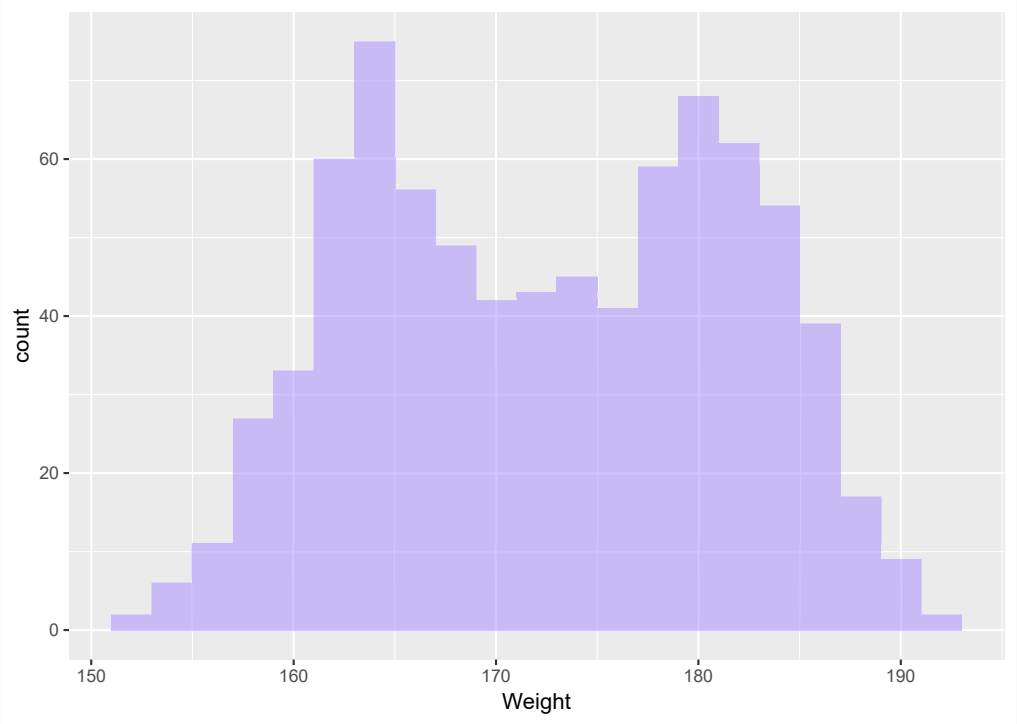
- By definition, the mode is at the peak in a single-peaked distribution (Located on the left side)
- A right-skewed distribution pulls the mean and median to the right of the mode (meaning to values greater than the mode)
- Outliers at the high end of the data set make the mean greater than the median

Shape of data (Uniform)



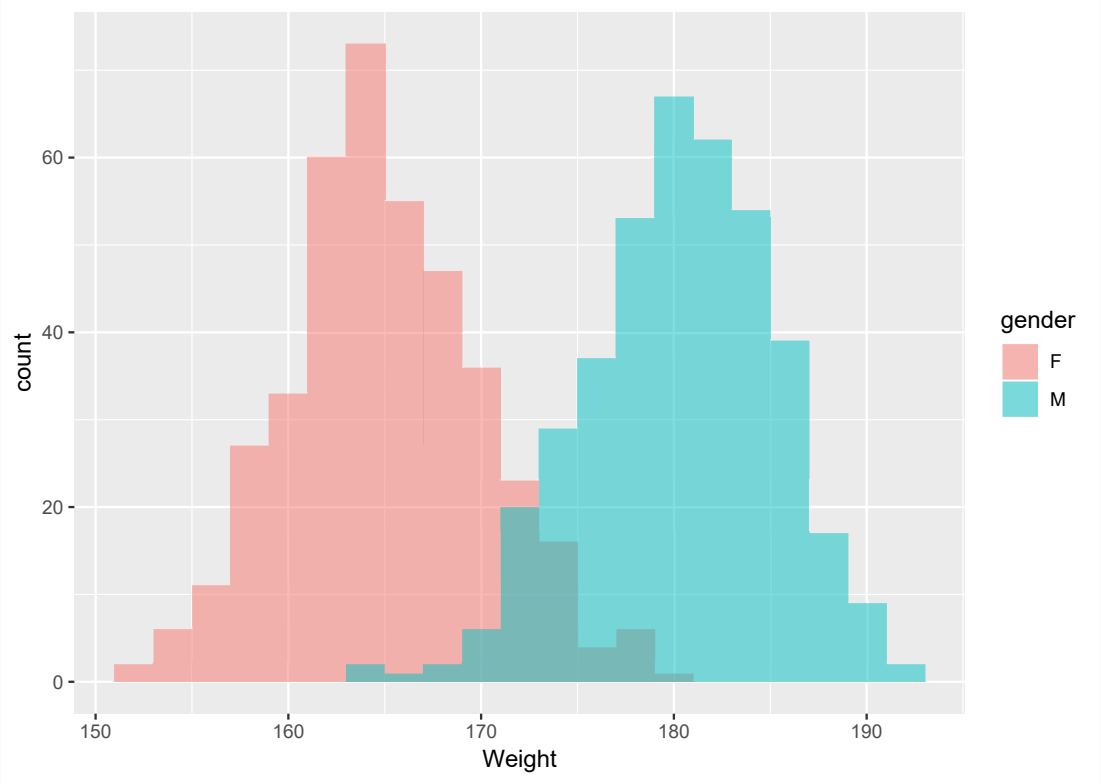
No dominant mode because all data values have the roughly same frequency

Shape of data



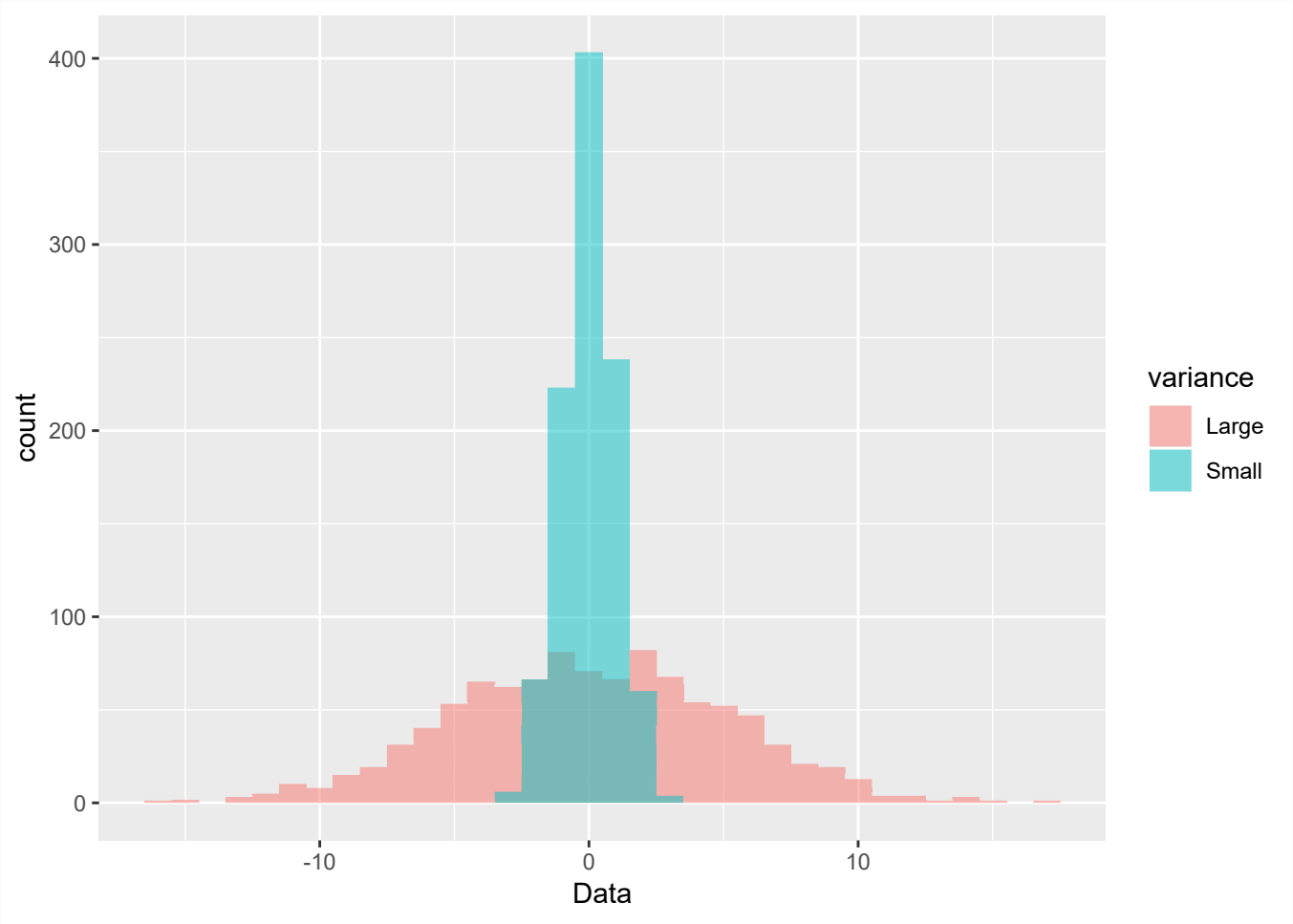
- Weight of FSU students
- Single mode ? No !!

Shape of data (Bimodal)



- Two dominant peak : Bimodal shape
- May show that the data has come from two different systems (or two different groups)

Shape of data (Variance)



Shape of data (Variance)

Small variance

- the data are concentrated arount the center
- Range of histogram is narrow
- sharp peak

Large variance

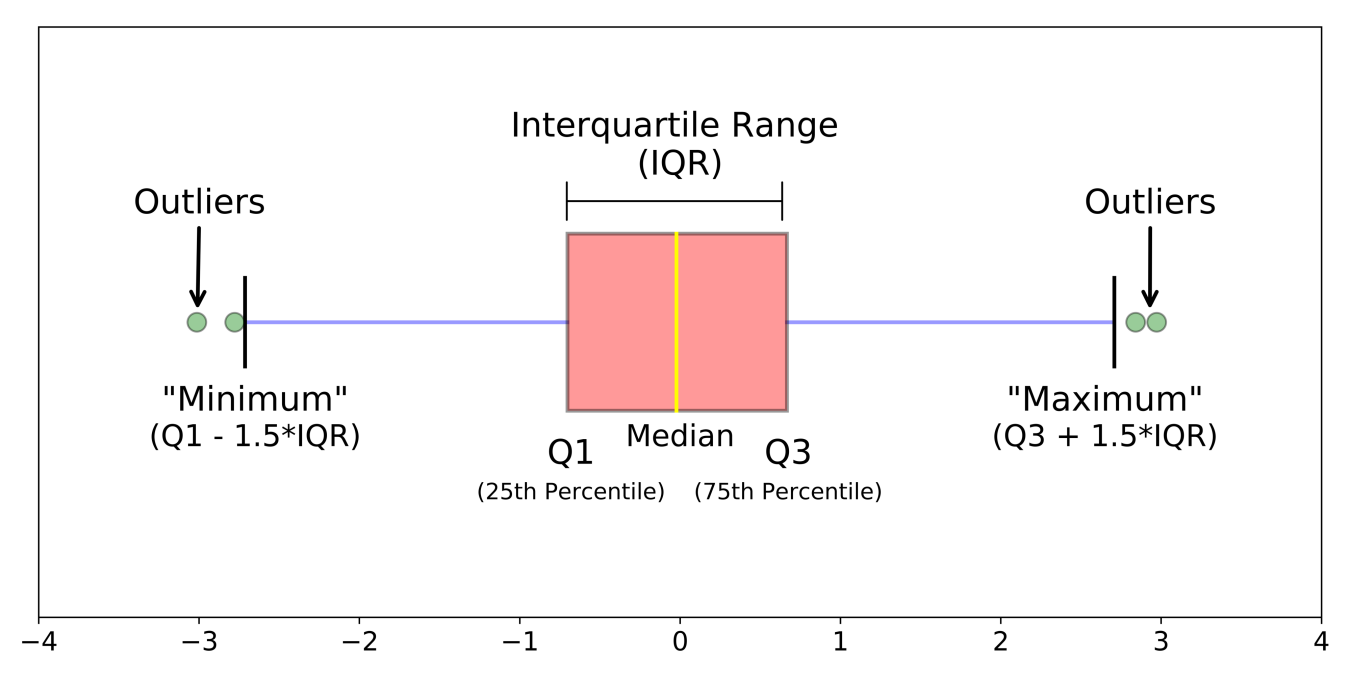
- the data are distributed widely around the center
- Range of hitogram is wide
- broad peak

Boxplot

Boxplot

- We can display the five-number summary with a graph called a boxplot
- Using a number line for reference, we enclose the values from the lower to the upper quartiles in a box.
- We then draw a line through the box at the median and add two "whiskers," extending from the box to the low and high values if there is no outlier.

Boxplot

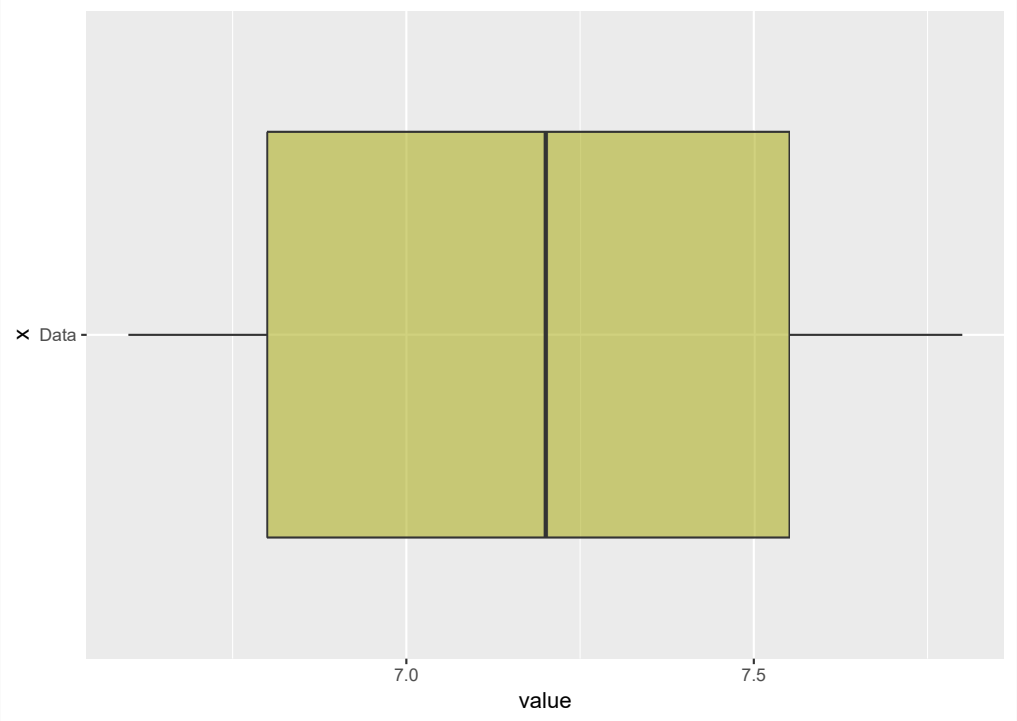


- First find the five-number summary
- Find the lower, upper fence (Detect outliers)
  - No outliers $\rightarrow$ whiskers : min and max of the data
  - Outliers exist $\rightarrow$ whiskers : min and max values inside the lower, and upper fence

Boxplot

Draw the box plot :

Obs	1	2	3	4	5	6	7	8	9	10	11
Data	6.6	6.7	6.7	6.9	7.1	7.2	7.3	7.4	7.7	7.8	7.8



Draw the box plot and label points with the numbers

blood Alcohol : Blood alcohol concentrations of drivers involved in fatal crashes: (n=11)

0.01	0.12	0.13	0.14	0.16	0.16
0.17	0.24	0.27	0.29	0.46	

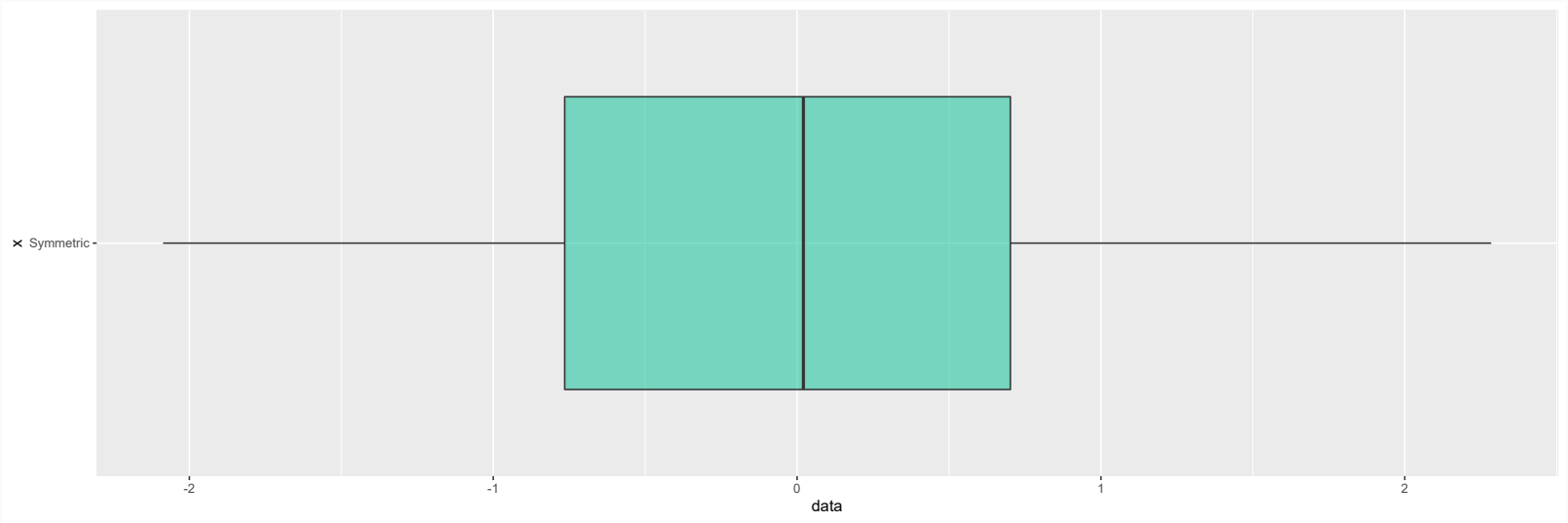
Draw the box plot and label points with the numbers

Space Shuttle Flights : Listed below are the durations (in hours) of a sample of all flights of NASA's Space Transport System (space shuttle): (n=15)

0	73	95	165	191	192	221	235
235	244	259	262	331	376	381	

Shape of Box plot (Symmetric)

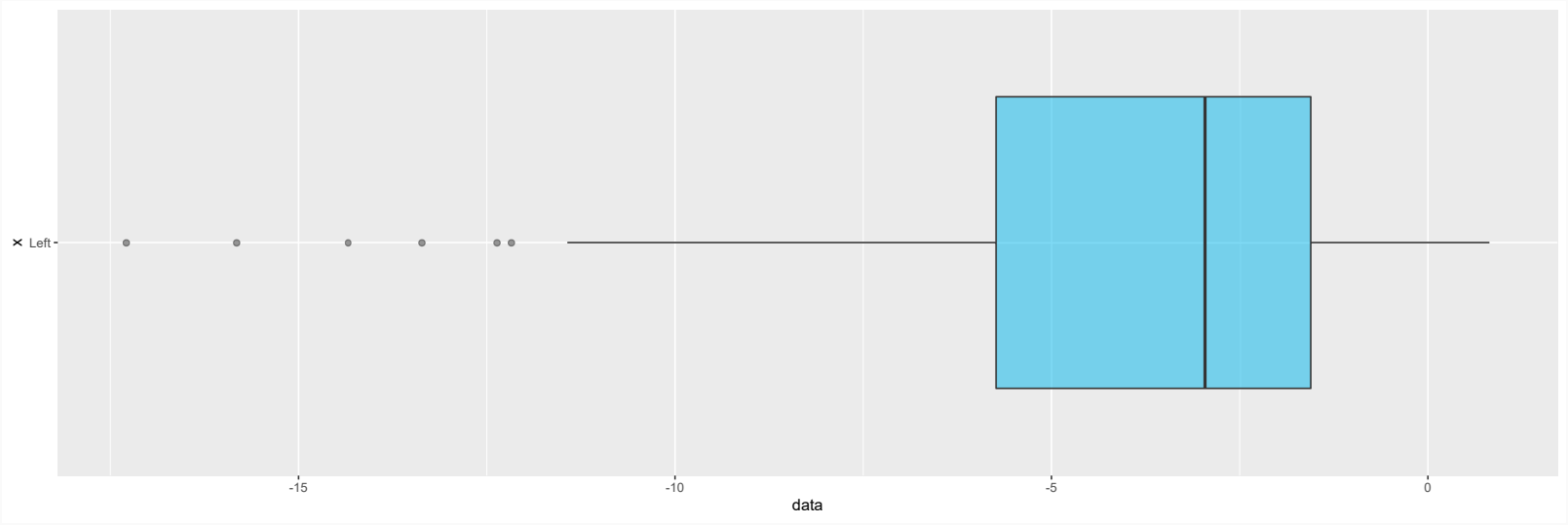
Box plot of Symmetric data



- Median located on the center of the box
- the left and right tails are equally balanced

Shape of Box plot (Left skewed)

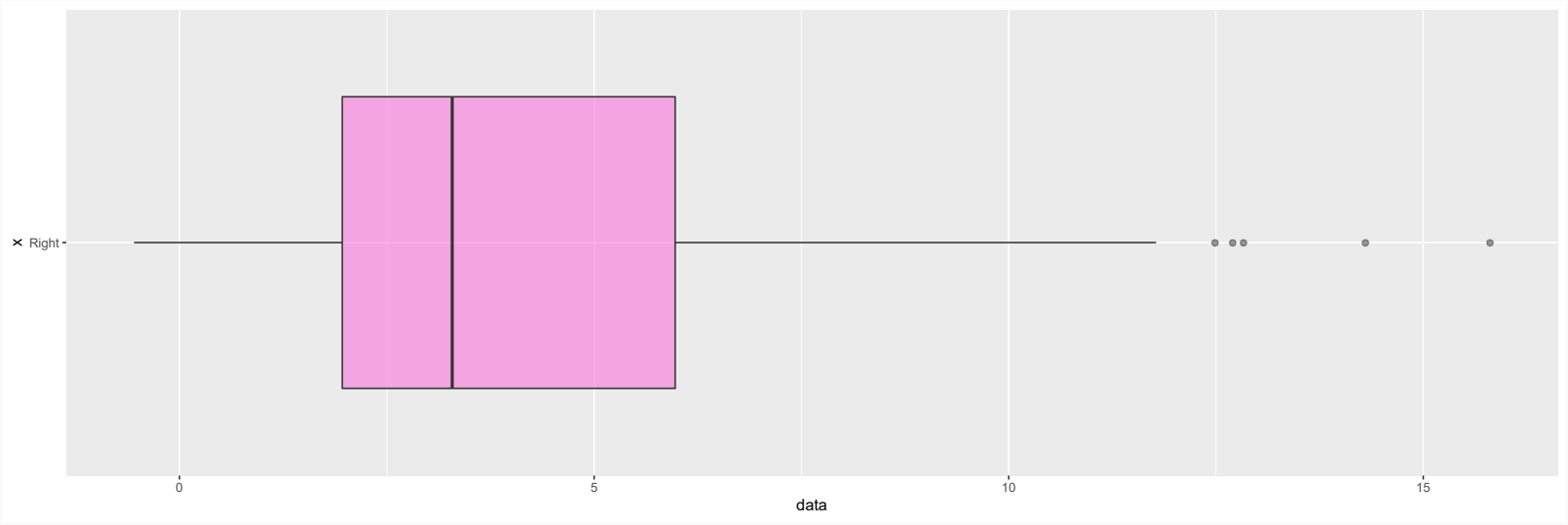
Box plot of the Left Skewed data



- Median closer to the upper quartile (Q_3)
- There are Low outliers (left side)
- Left whisker is longer than the right whisker

Shape of Box plot (Right skewed)

Box plot of the Right Skewed data

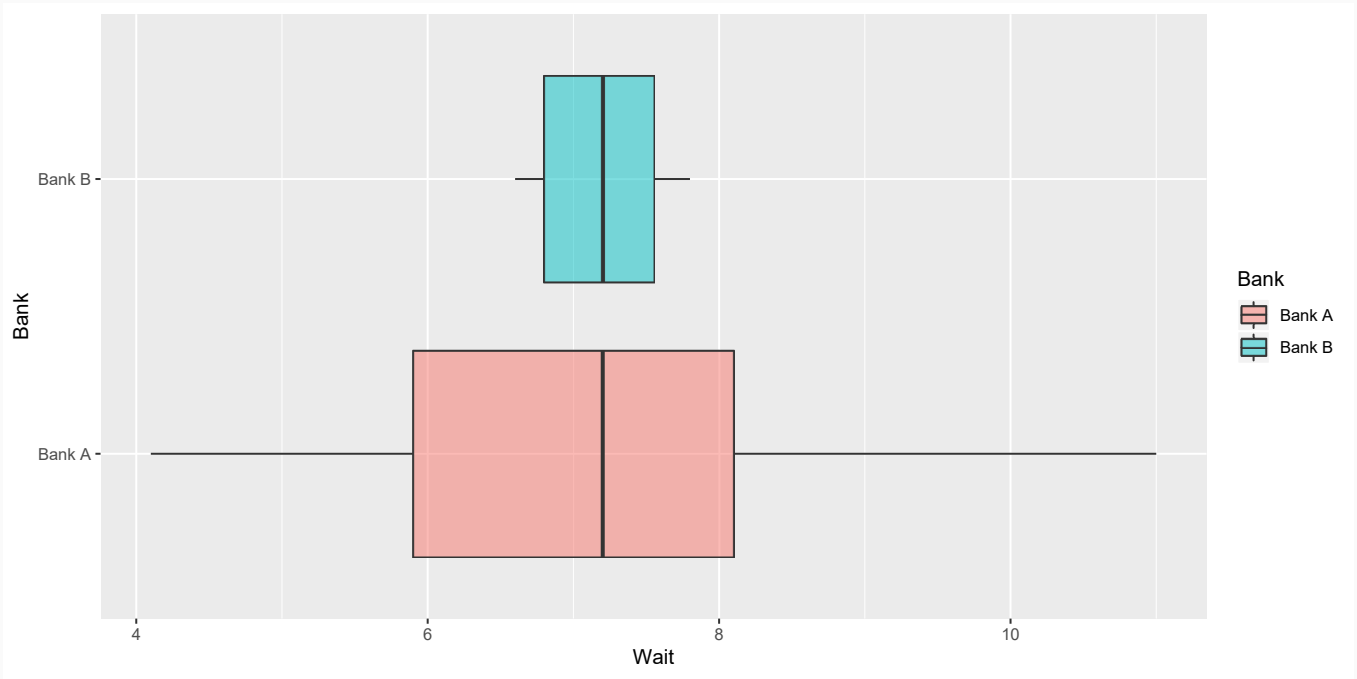


- Median closer to the lower quartile (Q_1)
- There are High outliers (right side)
- Right whisker is longer than the left whisker

Shape of Box plot (Variance)

Bank A	4.1	5.2	5.6	6.2	6.7	7.2	7.7	7.7	8.5	9.3	11.0
Bank B	6.6	6.7	6.7	6.9	7.1	7.2	7.3	7.4	7.7	7.8	7.8

Box-plot of the bank data



Shape of Box plot (Variance)

Small variance

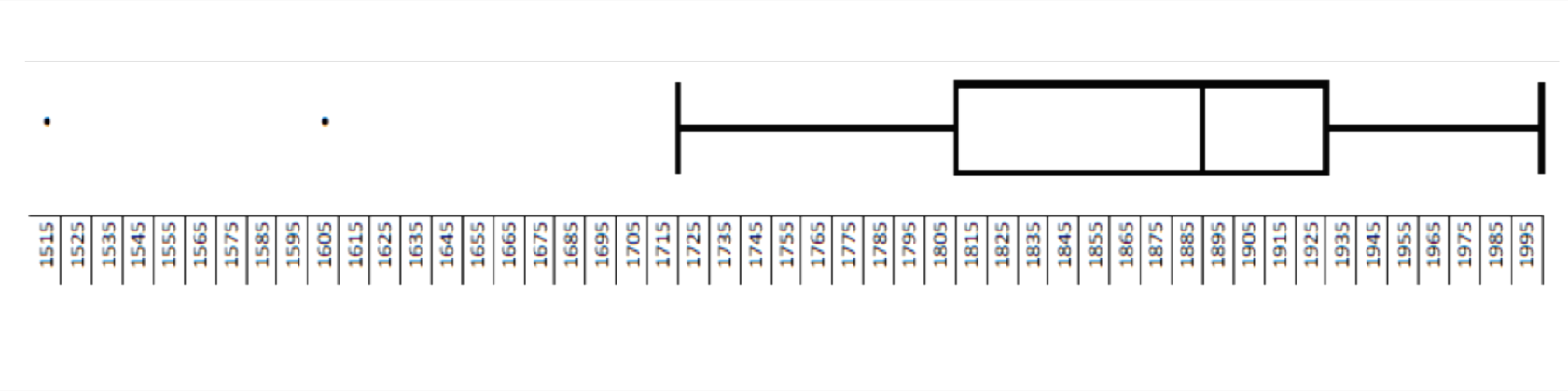
- Narrow box
- Short whiskers
- No ourliers

Large variance

- Wide box
- Long whiskers
- Sometimes Outliers exist

Exercise : Gas production

The boxplot shows the USA's monthly natural gas production for a certain period of time. The units are "thousand barrels per day" and the line inside the box is at 1,885.



- What is the shape of the distribution?
- This shape tells us that most of the production amounts were on the (higher / lower) side.

Exercise : Gas production

- What percent of the time was production more than 1,885 thou.barrels/day?
- 75% of the time production was less than _____ thousand barrels per day.
- We might expect the mean production to be (less than / roughly equal to / greater than) 1,885 thousand barrels per day.
- Determine the range and interquartile range of the production amounts
- What percent of the time was production between 1,885 and 1,925 thou.barrels/day?
- What percent of the time was production between 1,515 and 1,805 thou.barrels/day?
- What percent of the time was production between 1,925 and 1,995 thou.barrels/day?