

2-7 Exploratory Data Analysis

NOTE: The exercises in this section may be done much more easily when ordered lists of the values are available. Exercises #1, #2, #6 and #7 were worked using such ordered lists and the method of the text to obtain the quartiles. The data for each of the remaining exercises exists as an Excel workbook, and this manual uses the Excel quartiles for those exercises.

1. Consider the 25 employee ages.

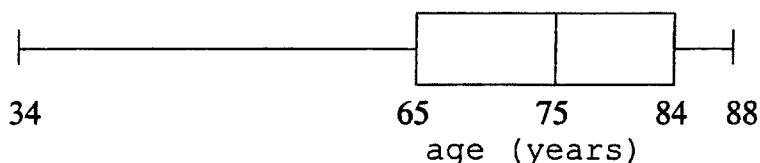
$$\min = x_1 = 34$$

$$Q_1 = x_7 = 65 \quad [L = (25/100) \cdot 25 = 6.25 \text{ rounded up to } 7]$$

$$Q_2 = x_{13} = 75 \quad [L = (50/100) \cdot 25 = 12.5 \text{ rounded up to } 13]$$

$$Q_3 = x_{19} = 84 \quad [L = (75/100) \cdot 25 = 18.75 \text{ rounded up to } 19]$$

$$\max = x_{25} = 88$$



The employees seem to be considerably older than most American workers.

3. Consider the 94 pulse rates.

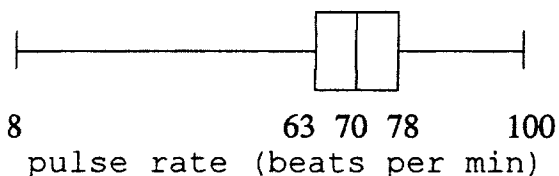
$$\min = \text{MIN}(G2:G95) = 8$$

$$Q_1 = \text{QUARTILE}(G2:G95,1) = 63$$

$$Q_2 = \text{QUARTILE}(G2:G95,2) = 70$$

$$Q_3 = \text{QUARTILE}(G2:G95,3) = 78$$

$$\max = \text{MAX}(G2:G95) = 100$$



The values 8 and 15 are far from the other values and appear to be errors.

NOTE: for exercise #3 and all future uses of this data set. The values 8 and 15 are obvious errors that will be eliminated from all subsequent analyses. Such obvious errors occur in many real life data sets. Often correct values can be deduced and the data adjusted accordingly. If, for example, all the other values were multiples of 4, one could infer that students monitored their pulse for 15 seconds and multiplied by 4 to obtain a per minute rate -- and that some students forgot to multiply by 4. It appears that the instructor gathered data by having each student monitor his own pulse rate -- and that there was not careful instruction or a re-take of suspicious values. It also appears that 6 students couldn't find their pulse. Although also questionable, the values in the 40's will be included in subsequent analyses of the pulse data, but the numbers are suspect and should not be taken as accurate renderings of student pulse rates.

5. Consider the 175 axial loads of the 0.0111 in. thick cans.

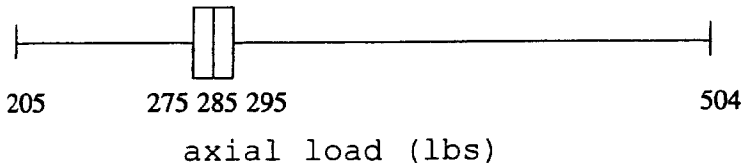
$$\text{min} = \text{MIN}(B2:B176) = 205$$

$$Q_1 = \text{QUARTILE}(B2:B176,1) = 275$$

$$Q_2 = \text{QUARTILE}(B2:B176,2) = 285$$

$$Q_3 = \text{QUARTILE}(B2:B176,3) = 294.5$$

$$\text{max} = \text{MAX}(B2:B176) = 504$$



The value 504 appears to be either an error or an anomaly.

7. Consider the 36 actor and 36 actress values.

$$\text{For } Q_1 = P_{25}, L = (50/100) \cdot 36 = 18 \text{ -- an integer, use } 18.5.$$

$$\text{For } \bar{x} = Q_2 = P_{50}, L = (50/100) \cdot 36 = 18 \text{ -- an integer, use } 18.5.$$

$$\text{For } Q_3 = P_{75}, L = (75/100) \cdot 36 = 27 \text{ -- an integer, use } 27.5.$$

For the actors

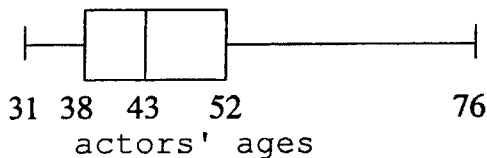
$$\text{min} = x_1 = 31$$

$$Q_1 = x_{9.5} = (37 + 38)/2 = 37.5$$

$$Q_2 = x_{18.5} = (43 + 43)/2 = 43$$

$$Q_3 = x_{27.5} = (51 + 53)/2 = 52$$

$$\text{max} = x_{36} = 76$$



For the actresses

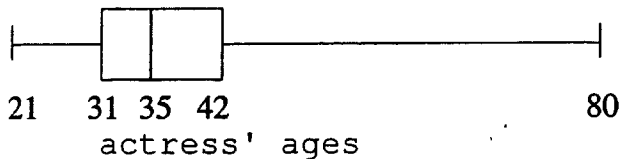
$$\text{min} = x_1 = 21$$

$$Q_1 = x_{9.5} = (30 + 31)/2 = 30.5$$

$$Q_2 = x_{18.5} = (35 + 35)/2 = 35$$

$$Q_3 = x_{27.5} = (41 + 42)/2 = 41.5$$

$$\text{max} = x_{36} = 80$$



The ages for the actresses cover a wider range and cluster around a lower value than do the ages of the actors.

9. Consider the 35 new textbook prices for the author's college.

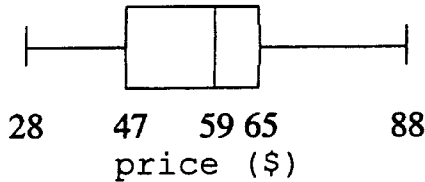
$$\text{min} = \text{MIN}(C2:C36) = 28.35$$

$$Q_1 = \text{QUARTILE}(C2:C36,1) = 47.4$$

$$Q_2 = \text{QUARTILE}(C2:C36,2) = 59.35$$

$$Q_3 = \text{QUARTILE}(C2:C36,3) = 65.10$$

$$\text{max} = \text{MAX}(C2:C36) = 88$$



Consider the 40 new textbook prices for UMASS.

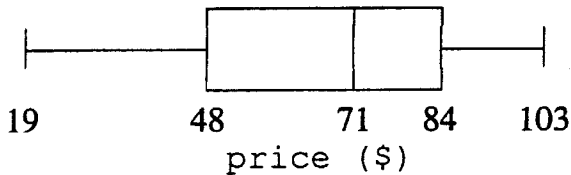
$$\text{min} = \text{MIN}(A2:A41) = 18.95$$

$$Q_1 = \text{QUARTILE}(A2:A41,1) = 47.925$$

$$Q_2 = \text{QUARTILE}(A2:A41,2) = 71.35$$

$$Q_3 = \text{QUARTILE}(A2:A41,3) = 83.5875$$

$$\text{max} = \text{MAX}(A2:A41) = 102.95$$



New textbooks at the University of Massachusetts appear to cost more than those at the author's school.

11. The following given values are also calculated in detail in exercise #5.

$$Q_1 = 275$$

$$Q_2 = 285$$

$$Q_3 = 295$$

a. $\text{IQR} = Q_3 - Q_1 = 295 - 275 = 20$

b. Since $1.5(\text{IQR}) = 1.5(20) = 30$, the modified boxplot line extends to the smallest value above $Q_1 - 30 = 275 - 30 = 245$, which is 246
the largest value below $Q_3 + 30 = 295 + 30 = 325$, which is 317

c. Since $3.0(\text{IQR}) = 3.0(30) = 60$, mild outliers are x values for which
 $Q_1 - 60 \leq x < Q_1 - 30$ or $Q_3 + 30 < x \leq Q_3 + 60$
 $215 \leq x < 245$ $325 < x \leq 355$

Those values are as follows.

lower end: 215,216,222,225,227,230,231,243,244

upper end: (none)

d. Extreme outliers in this exercise are x values for which $x < 215$ or $x > 355$.

Those values are as follows.

lower end: 205,210,210,211

upper end: 504