1.

(a)

> x=c(53,39,39,33,69,30,25,67,130,94,40)

> x

[1] 53 39 39 33 69 30 25 67 130 94 40

(b)

> mean(x)

[1] 56.27273

> sd(x)

[1] 31.96589

Round to the first decimal place. i.e., mean(x)=56.3 and sd(x)=32.0.

(c)

> median(x)

[1] 40

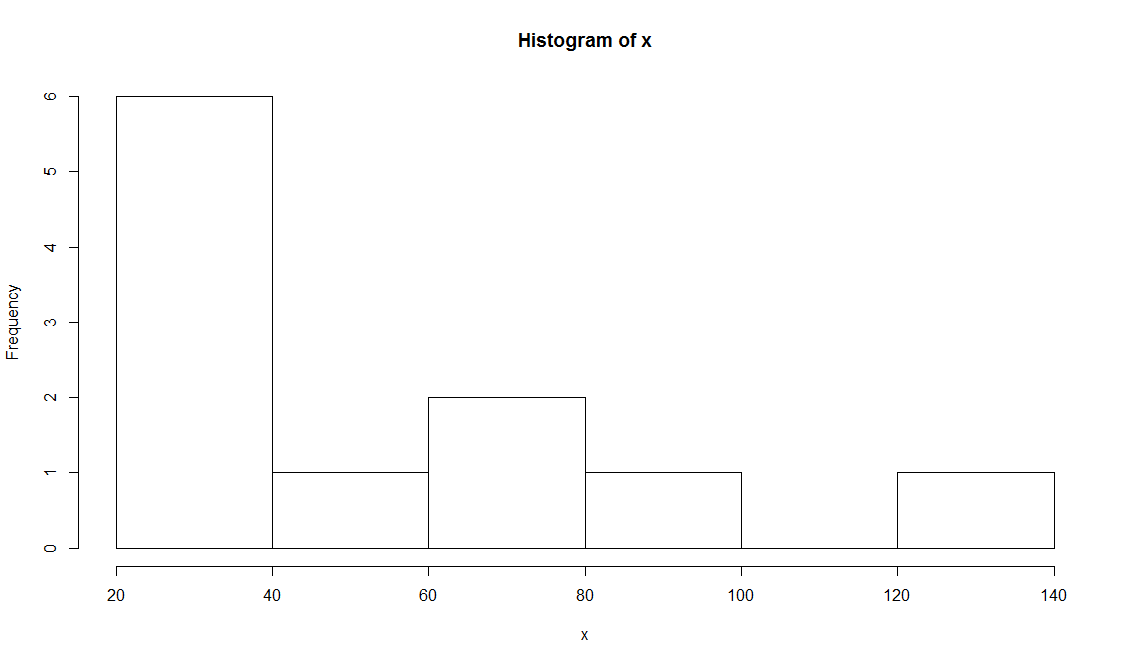
> quantile(x)

0% 25% 50% 75% 100%

25 36 40 68 130

(d)

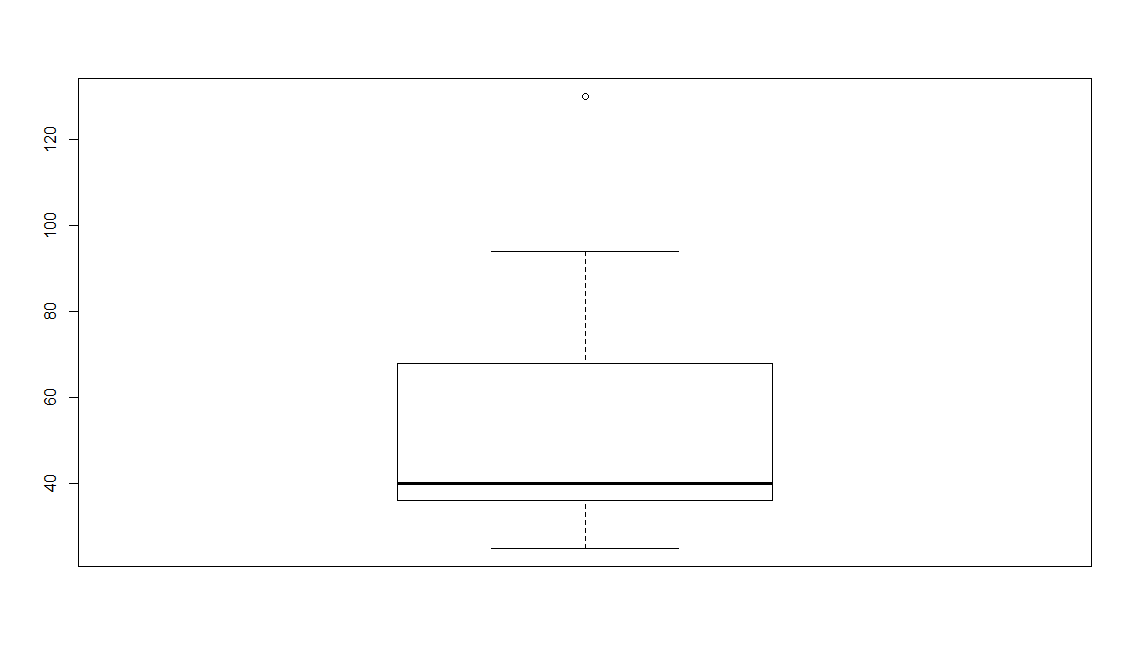
> hist(x)



The shape is skewed to right, and there is outlier in the histogram.

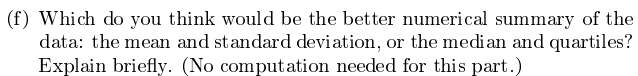
(e)

> boxplot(x)



Yes, the boxplox has shown that the data is skewed to the right. And median=40, first quartile=36 and third quartile=68 which are the same as in the histogram. We also can see a outlier in the figure.

(f)



Depends on whether there are outliers or skewness. If the data has outliers or is skewed, median and quartiles is a better choice; If the data has no outliers and is symmetric, then mean and standard deviation is much better. Hence according to the data in the question, I think median and quartiles is the better choice.

2.

(a)

**CODE:**

> data mydata;

> input mugtype $ tempdiff;

> cards;

> sigg 12

> sigg 16

> sigg 9

> sigg 23

> sigg 11

> sigg 20.5

> sigg 12.5

> sigg 24.5

> starbucks 13

> starbucks 7

> starbucks 7

> starbucks 17.5

> starbucks 10

> starbucks 15.5

> starbucks 6

> starbucks 6

> cupps 6

> cupps 6

> cupps 6

> cupps 18.5

> cupps 10

> cupps 17.5

> cupps 11

> cupps 6.5

> nissan 1.5

> nissan 2

> nissan 3

> nissan 0

> nissan 7

> nissan 0.5

> nissan 6

> nissan 2

> ;

>

> proc print;

>

> run;

**RESULTS:**

| **Obs** | **mugtype** | **tempdiff** |
| --- | --- | --- |
| **1** | sigg | 12.0 |
| **2** | sigg | 16.0 |
| **3** | sigg | 9.0 |
| **4** | sigg | 23.0 |
| **5** | sigg | 11.0 |
| **6** | sigg | 20.5 |
| **7** | sigg | 12.5 |
| **8** | sigg | 24.5 |
| **9** | starbuck | 13.0 |
| **10** | starbuck | 7.0 |
| **11** | starbuck | 7.0 |
| **12** | starbuck | 17.5 |
| **13** | starbuck | 10.0 |
| **14** | starbuck | 15.5 |
| **15** | starbuck | 6.0 |
| **16** | starbuck | 6.0 |
| **17** | cupps | 6.0 |
| **18** | cupps | 6.0 |
| **19** | cupps | 6.0 |
| **20** | cupps | 18.5 |
| **21** | cupps | 10.0 |
| **22** | cupps | 17.5 |
| **23** | cupps | 11.0 |
| **24** | cupps | 6.5 |
| **25** | nissan | 1.5 |
| **26** | nissan | 2.0 |
| **27** | nissan | 3.0 |
| **28** | nissan | 0.0 |
| **29** | nissan | 7.0 |
| **30** | nissan | 0.5 |
| **31** | nissan | 6.0 |
| **32** | nissan | 2.0 |

(b)

**CODE:**

> proc means;

> class mugtype;

> var tempdiff;

**RESULTS:**

**The MEANS Procedure**

| **Analysis Variable : tempdiff** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **mugtype** | **N Obs** | **N** | **Mean** | **Std Dev** | **Minimum** | **Maximum** |
| cupps | 8 | 8 | 10.1875000 | 5.2025921 | 6.0000000 | 18.5000000 |
| nissan | 8 | 8 | 2.7500000 | 2.5071327 | 0 | 7.0000000 |
| sigg | 8 | 8 | 16.0625000 | 5.9005902 | 9.0000000 | 24.5000000 |
| starbuck | 8 | 8 | 10.2500000 | 4.5512949 | 6.0000000 | 17.5000000 |

Since Nissan mug has the smallest mean and standard deviation, it would be the best one. On the contrary, Sigg mug has the biggest mean and standard deviation, so it appears to be the worse one.

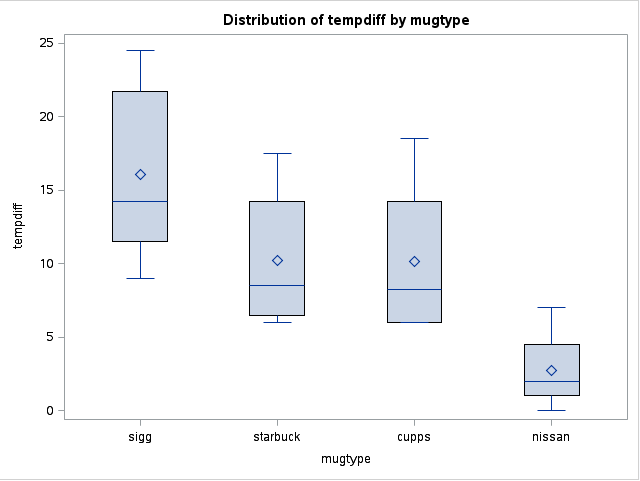
(c)

**CODE:**

> proc boxplot;

> plot tempdiff\*mugtype / boxstyle=schematic;

**RESULTS:**



Yes, from the boxplot we can see that Sigg mug has the biggest mean and Nissan has the smallest mean; All the means are above the medians, and the upper whiskers are longer than the lower whiskers which imply the right-skewness.

3.

(a)

**CODE:**

> data pop;

> infile '/home/shu.tu/pop.txt';

> input pop region $ ;

>

> proc print;

**RESULTS:**

| **Obs** | **pop** | **region** |
| --- | --- | --- |
| **1** | 1 | NE/MW |
| **2** | 3 | NE/MW |
| **3** | 4 | NE/MW |
| **4** | 4 | NE/MW |
| **5** | 5 | NE/MW |
| **6** | 5 | NE/MW |
| **7** | 6 | NE/MW |
| **8** | 6 | NE/MW |
| **9** | 7 | NE/MW |
| **10** | 8 | NE/MW |
| **11** | 8 | NE/MW |
| **12** | 9 | NE/MW |
| **13** | 9 | NE/MW |
| **14** | 9 | NE/MW |
| **15** | 9 | NE/MW |
| **16** | 10 | NE/MW |
| **17** | 10 | NE/MW |
| **18** | 11 | NE/MW |
| **19** | 12 | NE/MW |
| **20** | 1 | S/W |
| **21** | 6 | S/W |
| **22** | 9 | S/W |
| **23** | 9 | S/W |
| **24** | 9 | S/W |
| **25** | 10 | S/W |
| **26** | 10 | S/W |
| **27** | 11 | S/W |
| **28** | 11 | S/W |
| **29** | 13 | S/W |
| **30** | 14 | S/W |
| **31** | 14 | S/W |
| **32** | 14 | S/W |
| **33** | 14 | S/W |
| **34** | 15 | S/W |
| **35** | 17 | S/W |
| **36** | 18 | S/W |
| **37** | 20 | S/W |
| **38** | 20 | S/W |
| **39** | 21 | S/W |
| **40** | 21 | S/W |
| **41** | 23 | S/W |
| **42** | 24 | S/W |
| **43** | 26 | S/W |
| **44** | 30 | S/W |
| **45** | 30 | S/W |
| **46** | 31 | S/W |
| **47** | 40 | S/W |
| **48** | 66 | S/W |

(b)

**CODE:**

> proc boxplot;

> plot pop\*region / boxstyle=schematic;

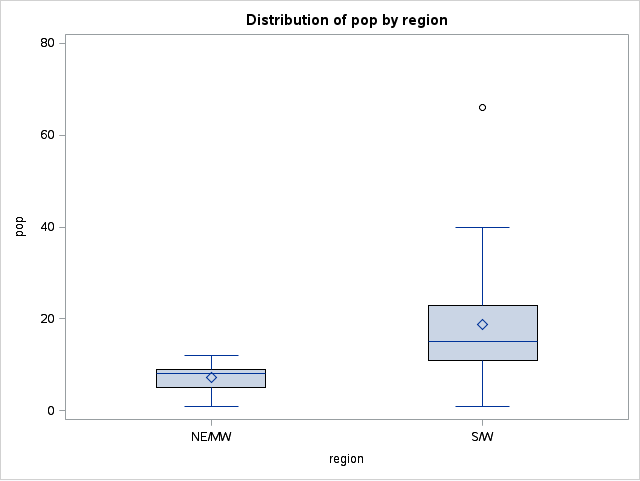
>

> proc means;

> class region;

> var pop;

**RESULTS:**



**The MEANS Procedure**

| **Analysis Variable : pop** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **region** | **N Obs** | **N** | **Mean** | **Std Dev** | **Minimum** | **Maximum** |
| NE/MW | 19 | 19 | 7.1578947 | 2.9488822 | 1.0000000 | 12.0000000 |
| S/W | 29 | 29 | 18.8620690 | 12.5006404 | 1.0000000 | 66.0000000 |

Both the mean and the standard deviation of the "S/W" are way larger than those of "NE/MW"; There is an outlier in "S/W", and the mean of "S/W" is bigger than its median, therefore, "S/W" would be skewed to the right.

4.

(a)

> mydata=read.table("wines.txt",header=T)

> mydata

CasePrice Location

1 123 Seneca

2 80 Cayuga

3 52 Seneca

4 112 Seneca

5 118 Keuka

6 95 Seneca

7 151 Seneca

8 100 Cayuga

9 66 Seneca

10 143 Keuka

11 110 Seneca

12 115 Seneca

13 70 Seneca

14 78 Seneca

15 103 Seneca

16 92 Seneca

17 118 Seneca

18 131 Keuka

19 115 Cayuga

20 128 Keuka

21 135 Cayuga

22 100 Cayuga

23 75 Seneca

24 105 Seneca

25 90 Seneca

26 72 Cayuga

27 138 Keuka

28 93 Seneca

29 106 Seneca

30 59 Seneca

31 97 Seneca

32 130 Cayuga

33 76 Cayuga

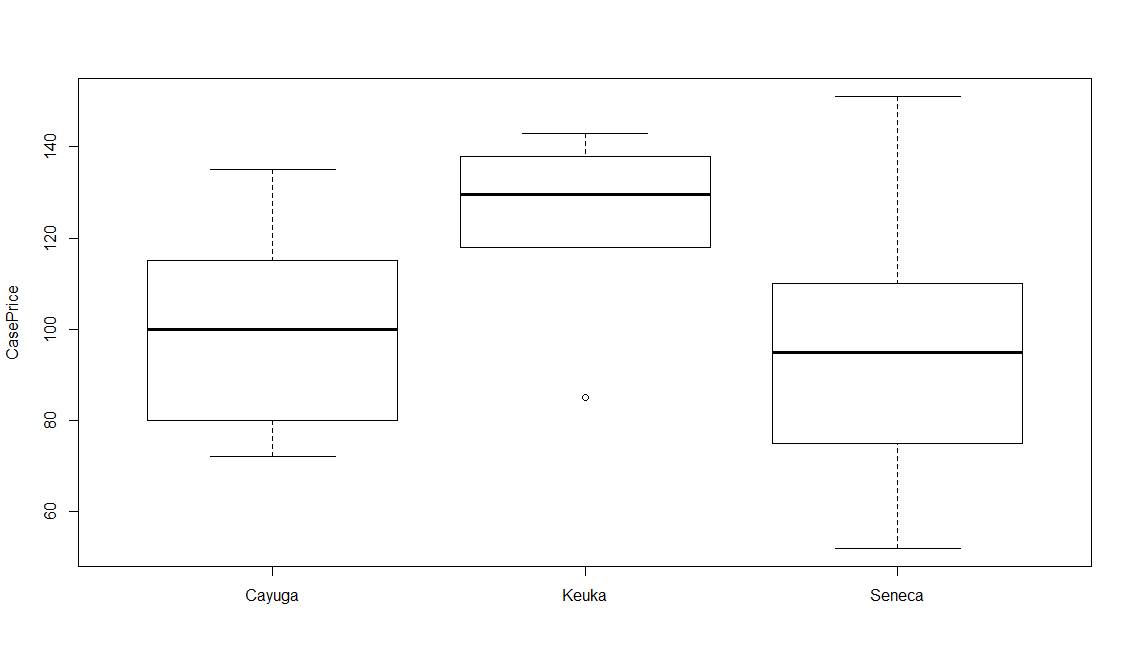
34 88 Cayuga

35 60 Seneca

36 85 Keuka

(b)

> boxplot(CasePrice~Location,data=mydata,ylab="CasePrice")



(c)

> aggregate(CasePrice~Location,mydata, mean)

Location CasePrice

1 Cayuga 99.55556

2 Keuka 123.83333

3 Seneca 93.80952

(d)

The boxplots have shown that Keuka has the highest caseprice on average, and the caseprices for Cayuga and Seneca on average are approximately the same. The means tell the same story. Therefore, the mean provides a fair comparison between the wines. However, the boxplots provides more information: there is an outlier in Keuka, and the whiskers of Seneca are much longer than those of the other two wines.

5.

(a)

**CODE:**

> data crime;

> infile '/home/shu.tu/crime.txt' expandtabs;

> input city $ homecide aggassault robbery BnE motortheft theft country $;

>

> proc print;

**RESULTS:**

| **Obs** | **city** | **homecide** | **aggassault** | **robbery** | **BnE** | **motortheft** | **theft** | **country** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | Toronto | 1.7 | 157 | 107 | 553 | 365 | 1692 | Canada |
| **2** | Montreal | 2.1 | 152 | 173 | 1195 | 800 | 2068 | Canada |
| **3** | Vancouve | 2.0 | 164 | 187 | 1430 | 1058 | 4415 | Canada |
| **4** | Calgary | 1.7 | 132 | 105 | 814 | 580 | 2616 | Canada |
| **5** | Edmonton | 2.0 | 180 | 134 | 986 | 539 | 2559 | Canada |
| **6** | Ottawa | 1.0 | 123 | 96 | 690 | 558 | 1835 | Canada |
| **7** | Quebec | 1.7 | 76 | 70 | 925 | 230 | 1771 | Canada |
| **8** | Winnipeg | 2.5 | 276 | 251 | 1228 | 1425 | 2779 | Canada |
| **9** | Hamilton | 1.3 | 144 | 86 | 815 | 698 | 1832 | Canada |
| **10** | LA | 10.6 | 607 | 298 | 636 | 674 | 1726 | USA |
| **11** | NYC | 7.8 | 474 | 372 | 453 | 428 | 1785 | USA |
| **12** | Philadel | 8.1 | 350 | 270 | 507 | 492 | 2199 | USA |
| **13** | Washingt | 7.4 | 265 | 171 | 452 | 484 | 2223 | USA |
| **14** | Detroit | 10.6 | 472 | 229 | 735 | 919 | 2280 | USA |
| **15** | Houston | 7.7 | 433 | 242 | 960 | 645 | 2724 | USA |
| **16** | Boston | 2.1 | . | 106 | 408 | 426 | 1640 | USA |
| **17** | Dallas | 8.5 | 377 | 252 | 1081 | 747 | 3146 | USA |
| **18** | Phoenix | 7.2 | 354 | 170 | 1111 | 1010 | 3524 | USA |

The expandtabs converts tabs to spaces so that list input can be used to read the tab-separated values.

(b)

Some of the city names are truncated, for example, Vancouver lost its "r", Washington lost its "on" etc.,

The "\*" in the "aggravated assault" column has turned into "." in SAS.

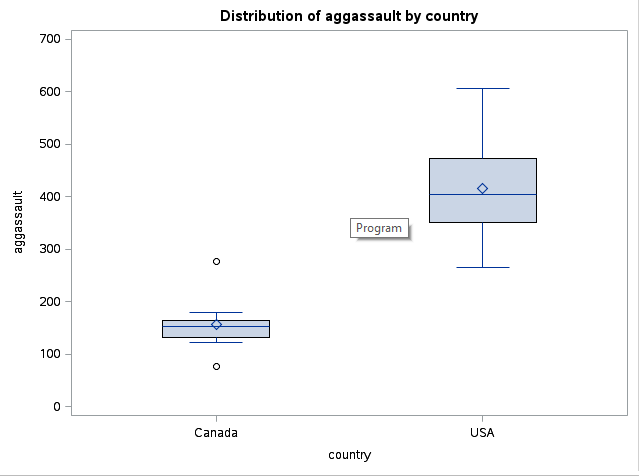
(c)

**CODE:**

> proc boxplot;

> plot aggassault\*country / boxstyle=schematic;

**RESULTS:**



From the boxplots we can see that the aggravated assault rates in Canada are much lower than in the U.S. And there is only one exception in Canada whose aggravated assault rate is relatively high which is almost up to 300 (i.e. the upper outlier).

(d)

**CODE:**

> proc means;

> var homecide;

> class country;

**RESULTS:**

**The MEANS Procedure**

| **Analysis Variable : homecide** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **country** | **N Obs** | **N** | **Mean** | **Std Dev** | **Minimum** | **Maximum** |
| Canada | 9 | 9 | 1.7777778 | 0.4437842 | 1.0000000 | 2.5000000 |
| USA | 9 | 9 | 7.7777778 | 2.4818228 | 2.1000000 | 10.6000000 |

The homicide rates on average are much higher in the U.S. than in Canada. The standard deviation of homicide rates for American cities is also higher than for Canadian cities which implies that there is one or more cities in the U.S. Which has a low homicide rate.



6.

(a)

> cr=read.csv("CrimeRates.csv",header=T)

> cr

City Homocide AggravatedAssault Robbery BreakAndEnter MotorVehicleTheft Theft Country

1 Toronto 1.7 157 107 553 365 1692 Canada

2 Montreal 2.1 152 173 1195 800 2068 Canada

3 Vancouver 2.0 164 187 1430 1058 4415 Canada

4 Calgary 1.7 132 105 814 580 2616 Canada

5 Edmonton 2.0 180 134 986 539 2559 Canada

6 Ottawa 1.0 123 96 690 558 1835 Canada

7 Quebec 1.7 76 70 925 230 1771 Canada

8 Winnipeg 2.5 276 251 1228 1425 2779 Canada

9 Hamilton 1.3 144 86 815 698 1832 Canada

10 LA 10.6 607 298 636 674 1726 USA

11 NYC 7.8 474 372 453 428 1785 USA

12 Philadelphia 8.1 350 270 507 492 2199 USA

13 Washington 7.4 265 171 452 484 2223 USA

14 Detroit 10.6 472 229 735 919 2280 USA

15 Houston 7.7 433 242 960 645 2724 USA

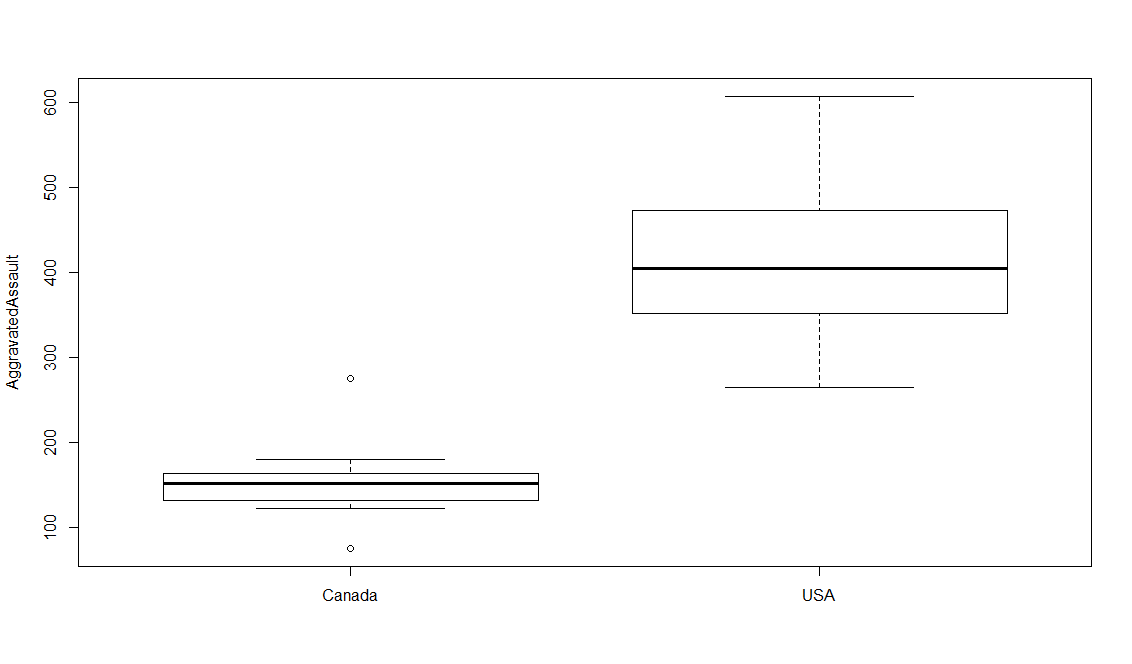
16 Boston 2.1 NA 106 408 426 1640 USA

17 Dallas 8.5 377 252 1081 747 3146 USA

18 Phoenix 7.2 354 170 1111 1010 3524 USA

(b)

>boxplot(AggravatedAssault~Country,data=cr,ylab="AggravatedAssault")



(c)

> aggregate(Homocide~Country,data=cr,mean)

Country Homocide

1 Canada 1.777778

2 USA 7.777778

(d)

> aggregate(Homocide~Country,data=cr,quantile)

Country Homocide.0% Homocide.25% Homocide.50%

1 Canada 1.0 1.7 1.7

2 USA 2.1 7.4 7.8

Homocide.75% Homocide.100%

1 2.0 2.5

2 8.5 10.6

(e)

Yes, they both give 1.777778 and 7.777778 for means, 1.0 and 2.1 for the minimums, 2.5 and 10.6 for the maximums.

7.

(a)

> chol=read.csv("Cholesterol.csv",header=T)

> chol

Smokers Ex.Smokers

1 225 250

2 211 134

3 209 300

4 284 249

5 258 213

6 216 310

7 196 175

8 288 174

9 250 328

10 200 160

11 209 188

12 280 321

13 225 213

14 256 257

15 243 292

16 200 200

17 213 271

18 246 227

19 225 238

20 237 163

21 232 263

22 267 192

23 232 242

24 216 249

25 216 242

26 243 267

27 200 243

28 155 217

29 216 267

30 271 218

31 230 217

32 309 183

33 183 228

34 280 NA

35 217 NA

36 305 NA

37 287 NA

38 217 NA

39 246 NA

40 351 NA

41 200 NA

42 280 NA

43 209 NA

There are 2 columns of cholesterol levels for Smokers and Ex.Smokers respectively.

(b)

> chol2=stack(chol)

> chol2

values ind

1 225 Smokers

2 211 Smokers

3 209 Smokers

4 284 Smokers

5 258 Smokers

6 216 Smokers

7 196 Smokers

8 288 Smokers

9 250 Smokers

10 200 Smokers

11 209 Smokers

12 280 Smokers

13 225 Smokers

14 256 Smokers

15 243 Smokers

16 200 Smokers

17 213 Smokers

18 246 Smokers

19 225 Smokers

20 237 Smokers

21 232 Smokers

22 267 Smokers

23 232 Smokers

24 216 Smokers

25 216 Smokers

26 243 Smokers

27 200 Smokers

28 155 Smokers

29 216 Smokers

30 271 Smokers

31 230 Smokers

32 309 Smokers

33 183 Smokers

34 280 Smokers

35 217 Smokers

36 305 Smokers

37 287 Smokers

38 217 Smokers

39 246 Smokers

40 351 Smokers

41 200 Smokers

42 280 Smokers

43 209 Smokers

44 250 Ex.Smokers

45 134 Ex.Smokers

46 300 Ex.Smokers

47 249 Ex.Smokers

48 213 Ex.Smokers

49 310 Ex.Smokers

50 175 Ex.Smokers

51 174 Ex.Smokers

52 328 Ex.Smokers

53 160 Ex.Smokers

54 188 Ex.Smokers

55 321 Ex.Smokers

56 213 Ex.Smokers

57 257 Ex.Smokers

58 292 Ex.Smokers

59 200 Ex.Smokers

60 271 Ex.Smokers

61 227 Ex.Smokers

62 238 Ex.Smokers

63 163 Ex.Smokers

64 263 Ex.Smokers

65 192 Ex.Smokers

66 242 Ex.Smokers

67 249 Ex.Smokers

68 242 Ex.Smokers

69 267 Ex.Smokers

70 243 Ex.Smokers

71 217 Ex.Smokers

72 267 Ex.Smokers

73 218 Ex.Smokers

74 217 Ex.Smokers

75 183 Ex.Smokers

76 228 Ex.Smokers

77 NA Ex.Smokers

78 NA Ex.Smokers

79 NA Ex.Smokers

80 NA Ex.Smokers

81 NA Ex.Smokers

82 NA Ex.Smokers

83 NA Ex.Smokers

84 NA Ex.Smokers

85 NA Ex.Smokers

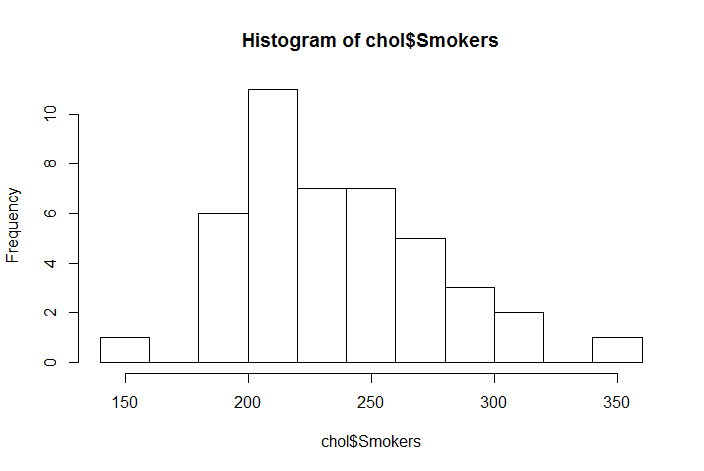
86 NA Ex.Smokers

This data frame has all the cholesterol values in one column and types of smokers in another column.

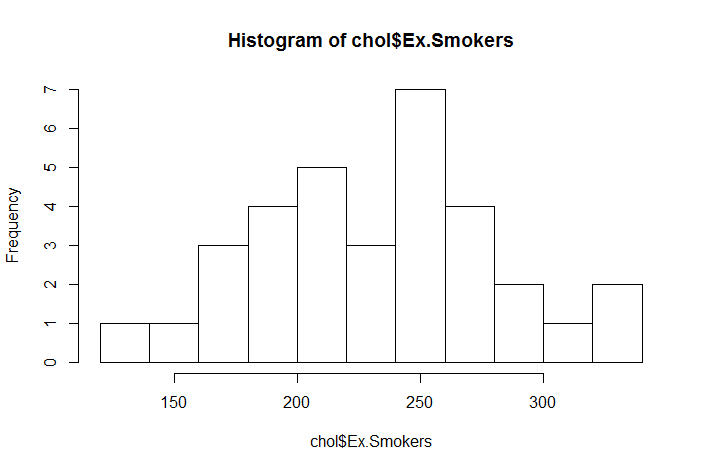


(c)

(i) > hist(chol$Smokers)

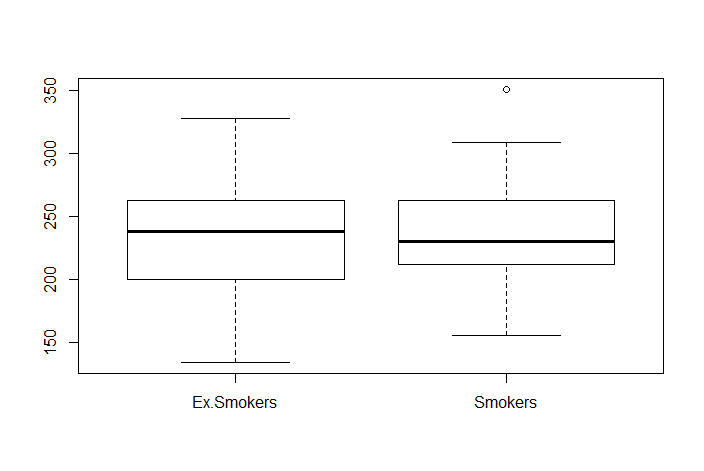


(ii) > hist(chol$Ex.Smokers)



(d)

> boxplot(values~ind,data=chol2)



(e)

I do not see any notable difference between these two groups. The only differences are that Ex.Smokers has a slightly wider spread and Smokers has an outlier.

(f)

For Smokers:

The shape is slightly skewed to the right. The values range from approximately 150 to 350, and the center is between 200 to 250. From the histogram we also can see that there is an outlier in Smokers.

For Ex.Smokers:

The shape is almost symmetric. The center is also between 200 to 250, and the value range from rougly 140 to 340.

Therefore , it is consistent with the ones from the boxplots.

(g) The boxplots might be easier to use for comparison. Because from the boxplots we can clearly figure out what median, mean, quartiles, shape and spread are. But from histogram, it is not that easy to get these information.