NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

R Lesson 3 - Solutions
MSPA 401 – Introduction to Statistical Analysis

1) For each weight class determine the mean and standard deviation of MPG. What can you conclude from these calculations?

	CLASS	MEAN_MPG	STD_DEV
1	C1	39.6777778	1.3608617
2	C2	35.5500000	0.5291503
3	C3	32.0166667	0.6293335
4	C4	29.6583333	2.1124989
5	C5	23.8500000	0.7728342
6	C6	19.1857143	2.7008817

Mean MPG changes with CLASS. The standard deviation is small in relation to MPG.

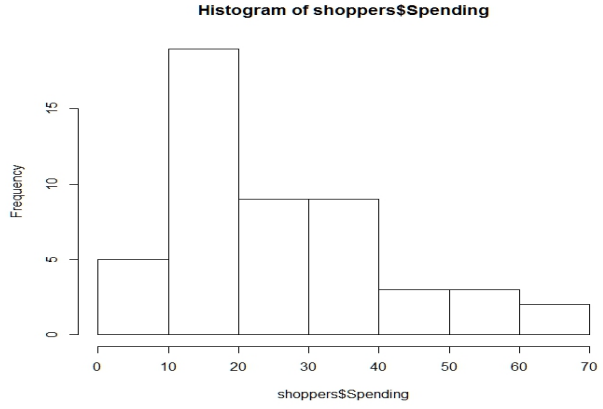
2) For each weight class determine the mean and standard deviation of HP. What can you conclude from these calculations?

	CLASS	MEAN_HP	STD_DEV
1	C1	89.222222	7.049429
2	C2	92.000000	9.086882
3	C3	103.500000	12.767145
4	C4	123.833333	25.672176
5	C5	171.583333	45.350069
6	C6	224.714286	74.017372

The standard deviation increases with increasing horsepower.

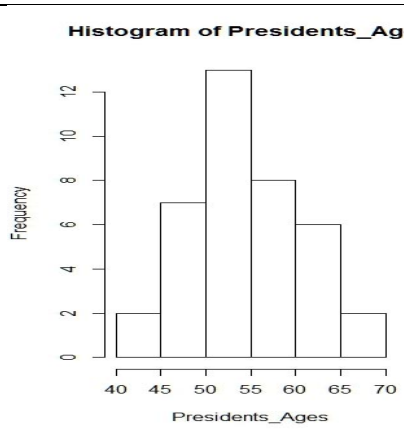
shoppers.csv contains the dollar amounts spent in a store by individual shoppers during one day. Find the mean, median, range, standard deviation, variance, Q1, Q3 and P10. Plot the histogram and describe the distribution.

mean: 25.4364
median: 20.735
range: 61.53
sd: 15.20959
variance: 231.3317
Q1: 14.3925
Q3: 33.665
P10: 10.179



The distribution is skewed right.

3) Find the mean, median, range, standard deviation, Q1, Q3 and P10 of the Presidents' ages. Plot hist() and comment on the distribution.

mean: 54.76316 median: 54.5 range: 27 sd: 6.561288 variance: 43.0505 Q1: 51 Q3: 57.75 P10: 46.7	

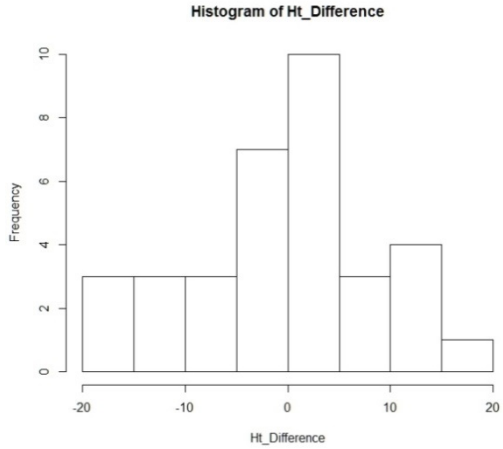
The distribution is relatively symmetric. The mean and median are close to each other.

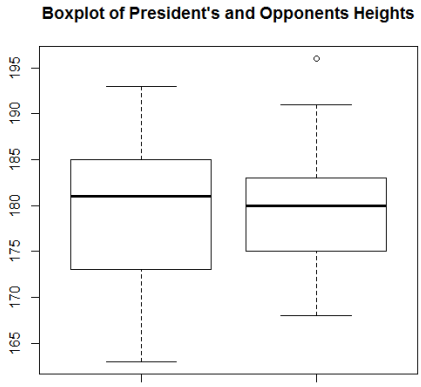
4) Find the mean, median, range, standard deviation, Q1, Q3 and P10 of the heights of the Presidents and also their opponents. Comment on what you find.

Presidents' Heights	Opponents' Heights
mean: 179.6842	mean: 179.9706
median: 181	median: 180
range: 30	range: 28
sd: 7.308289	sd: 6.201101
variance: 53.4111	variance: 38.45365
Q1: 173	Q1: 175.5
Q3: 184.5	Q3: 182.75
P10: 170	P10: 173

Heights are very similar. The difference is minor in comparison to the standard deviation.

5) Calculate the difference between each President's height and that of his opponent. Plot a histogram and construct a boxplot of these differences. Why is the difference of average heights calculated in (2) different from the average of the pairwise differences calculated in (3)?



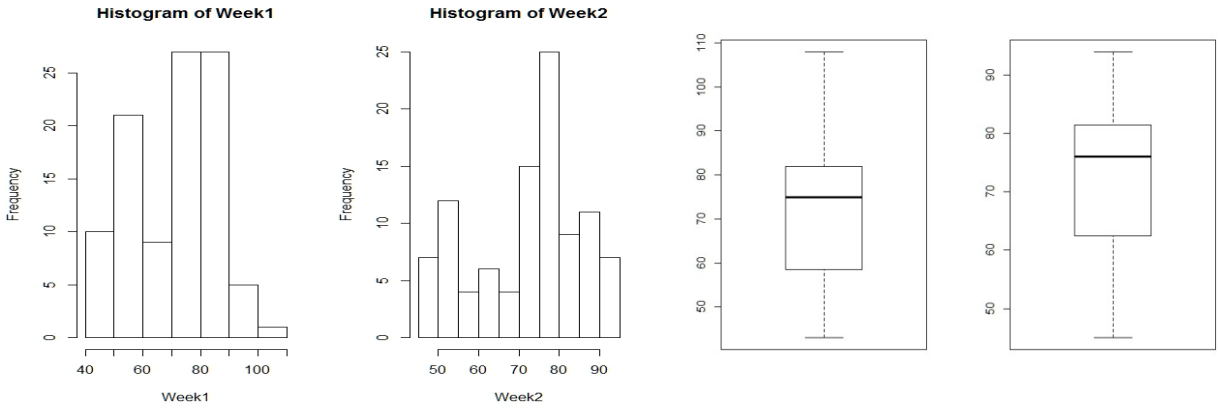


Immaterial difference in heights observed; difference in calculations dues to missing data (i.e. NA) for some opponents.

geyser.csv contains the intervals (in minutes) between eruptions of Old Faithful Geyser in Yellowstone National Park. The data were taken on two consecutive weeks: WEEK1 and WEEK2. Compare the two sets of data using summary() and hist(). What do you conclude?

```
     WEEK1                WEEK2
Min.   : 43.00   Min.   :45.00
1st Qu.: 58.75   1st Qu.:63.25
Median : 75.00   Median :76.00
Mean   : 71.62   Mean   :72.76
3rd Qu.: 82.00   3rd Qu.:81.25
Max.   :108.00   Max.   :94.00
```

The mean and median are comparable between WEEK1 and WEEK2. The two histograms suggest a multi-modal distribution. Note the difference in scales for the boxplots. This is one of the difficulties when base R boxplot() is used.



No apparent average difference between weeks, but multi-modal distribution.