HW1.R

110077443

2/14/2022

Download the data file

```
customers <- read.table(file = "customers.txt", header = T)
```

1. What is the 5th element in the original list of ages?

```
ages <- customers$age
```

2. What is the fifth lowest age?

```
sort_ages <- sort(customers$age) #Creating variable with sorted ages - to solve later queries
sort_ages #Checking values of variable
```

```
##   [1] 18 19 19 19 19 19 19 19 19 20 20 20 20 20 21 21 21 21 21 21 21 22 22 23 23
##  [26] 23 23 23 23 24 24 24 25 25 25 25 25 25 25 26 26 26 26 26 26 26 26 26 27 27
##  [51] 27 27 27 28 28 28 28 28 29 29 29 29 29 29 30 30 30 30 30 30 30 30 31 31 31
##  [76] 31 31 31 31 31 32 32 32 32 32 32 32 32 33 33 33 33 33 34 34 34 34 34 34 34
## [101] 34 34 35 35 35 35 35 35 36 36 36 36 36 37 37 37 37 37 37 37 37 38 38 38 38
## [126] 38 38 39 39 39 39 39 40 40 40 40 40 40 40 41 41 41 41 41 41 42 42 42 42 42
## [151] 42 42 42 43 43 43 43 43 43 44 44 44 44 45 45 45 45 45 45 45 45 45 45 45 45
## [176] 45 45 45 45 45 45 45 45 45 45 46 46 46 46 46 46 46 46 46 47 47 47 47 47 47
## [201] 47 47 47 47 47 47 47 47 47 47 47 47 47 48 48 48 48 48 48 48 48 48 48 48 48
## [226] 48 48 48 48 48 49 49 49 49 49 49 49 49 49 49 49 49 49 49 49 49 49 49 49 49
## [251] 49 49 49 49 49 49 49 49 49 49 49 49 49 50 50 50 50 50 50 50 50 50 50 50 50
## [276] 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 51 51 51 51 51 51 52 52 52 53
## [301] 53 53 53 54 55 56 56 57 57 57 57 58 58 59 60 60 62 62 62 62 62 63 63 63 64
## [326] 64 65 66 67 67 67 68 68 69 70 70 70 70 70 70 71 71 71 71 71 71 72 72 72 72
## [351] 72 72 72 72 73 73 73 73 73 73 73 73 74 74 74 74 74 74 75 75 75 75 75 75 76
## [376] 76 76 76 76 76 77 77 77 77 78 78 78 78 79 79 79 79 80 80 81 82 82 83 85
```

```
sort(unique(sort_ages), decreasing = F)[5] #22 is the fifth lowest age - unique function is used
```

```
## [1] 22
```

```
#to remove repetitions in observations
```

3. Extract the five lowest ages together

```
sort(customers$age, decreasing = F)[1:5] #five lowest ages in data frame (with repeated ages)
```

```
## [1] 18 19 19 19 19
```

```
sort(unique(sort_ages), decreasing = F)[1:5] #Using the unique function to order and index (without rep
```

```
## [1] 18 19 20 21 22
```

4. Get the five highest ages by first sorting them in decreasing order first.

```
sort(customers$age, decreasing = T)[1:5] ##five highest ages in data frame (with repeated ages)
```

```
## [1] 85 83 82 82 81
```

```
sort(unique(sort_ages), decreasing = T)[1:5] #Using the unique function to order and index (without rep
```

```
## [1] 85 83 82 81 80
```

5. What is the average (mean) age?

```
mean(sort_ages) #46.8 is the average age - using the mean function to calculate
```

```
## [1] 46.80702
```

6. What is the standard deviation of ages?

```
require(pastecs) #Using this package to use a function that will provide detailed summary statistics
```

```
## Loading required package: pastecs
```

```
stat.desc(sort_ages) #Using function from package to get detailed summary statistics
```

```
##      nbr.val     nbr.null       nbr.na          min          max        range 
## 3.990000e+02 0.000000e+00 0.000000e+00 1.800000e+01 8.500000e+01 6.700000e+01 
##          sum       median         mean      SE.mean CI.mean.0.95          var 
## 1.867600e+04 4.700000e+01 4.680702e+01 8.195148e-01 1.611119e+00 2.679702e+02 
##      std.dev     coef.var 
## 1.636980e+01 3.497295e-01
```

```
sd(sort_ages) ##16.3698 is Standard deviation for ages
```

```
## [1] 16.3698
```

7. Make a new variable called age_diff, with the difference between each age and the mean age

```
customers$mean_ages <- mean(sort_ages) #Create column for mean age in data frame
customers$age_diff <- (customers$age - customers$mean_ages) #Create column for age_diff in data frame
head(customers) #Check data frame
```

```
##   age mean_ages  age_diff
## 1  49  46.80702  2.192982
## 2  69  46.80702 22.192982
## 3  41  46.80702 -5.807018
## 4  73  46.80702 26.192982
## 5  45  46.80702 -1.807018
## 6  71  46.80702 24.192982
```

```
age_diff <- (customers$age - customers$mean_ages) #Create a separate variable in global environment
age_diff #Check data
```

```
##   [1]   2.1929825  22.1929825  -5.8070175  26.1929825  -1.8070175  24.1929825
##   [7]   3.1929825  -3.8070175  23.1929825 -14.8070175   0.1929825  30.1929825
##  [13]  17.1929825   3.1929825   3.1929825  -1.8070175   2.1929825   0.1929825
##  [19]  15.1929825   3.1929825   0.1929825  25.1929825   0.1929825  16.1929825
##  [25] -25.8070175   2.1929825   3.1929825   1.1929825 -11.8070175  30.1929825
##  [31]   1.1929825   1.1929825   3.1929825   0.1929825 -17.8070175  -4.8070175
##  [37]  -4.8070175  38.1929825  -1.8070175   2.1929825  -1.8070175  -3.8070175
##  [43]   2.1929825  21.1929825  -4.8070175   1.1929825  25.1929825  32.1929825
##  [49]   1.1929825   3.1929825   0.1929825  -1.8070175 -16.8070175  29.1929825
##  [55] -15.8070175   2.1929825  27.1929825  25.1929825   1.1929825   2.1929825
##  [61]  26.1929825   3.1929825   0.1929825   0.1929825  36.1929825  25.1929825
##  [67]  28.1929825   3.1929825   3.1929825   2.1929825   1.1929825  -1.8070175
##  [73]   2.1929825   2.1929825   2.1929825  25.1929825   3.1929825  28.1929825
##  [79]  27.1929825  25.1929825  27.1929825  29.1929825   2.1929825   3.1929825
##  [85]  29.1929825 -10.8070175  -1.8070175 -11.8070175 -22.8070175  -1.8070175
##  [91]   3.1929825  -4.8070175 -24.8070175  13.1929825  12.1929825  -1.8070175
##  [97]   4.1929825  -0.8070175   0.1929825 -12.8070175  16.1929825  24.1929825
## [103]  -9.8070175 -25.8070175  -3.8070175 -14.8070175   0.1929825 -11.8070175
## [109]  23.1929825 -20.8070175  16.1929825   7.1929825  -1.8070175   0.1929825
## [115] -20.8070175 -11.8070175 -24.8070175 -15.8070175  23.1929825   4.1929825
## [121]  -9.8070175  -5.8070175   6.1929825 -12.8070175  -1.8070175 -12.8070175
## [127]  -3.8070175   3.1929825 -17.8070175   2.1929825  -0.8070175  -2.8070175
## [133] -20.8070175   2.1929825   1.1929825 -20.8070175 -12.8070175 -21.8070175
## [139]  -8.8070175 -21.8070175  31.1929825  -1.8070175 -15.8070175   0.1929825
## [145]  10.1929825 -18.8070175  28.1929825   2.1929825 -20.8070175   2.1929825
## [151] -12.8070175 -21.8070175   2.1929825 -12.8070175 -27.8070175 -14.8070175
## [157]   5.1929825  26.1929825  -7.8070175 -15.8070175   1.1929825  35.1929825
## [163] -13.8070175 -16.8070175  -9.8070175 -13.8070175   0.1929825 -17.8070175
## [169]   0.1929825  -9.8070175 -17.8070175  -6.8070175  15.1929825   1.1929825
## [175] -10.8070175  -5.8070175  10.1929825  10.1929825 -12.8070175 -21.8070175
## [181]  31.1929825 -23.8070175 -14.8070175  -5.8070175 -26.8070175  26.1929825
## [187]   2.1929825   3.1929825  -0.8070175   3.1929825 -19.8070175  -1.8070175
## [193] -17.8070175   9.1929825  28.1929825   6.1929825   0.1929825  -7.8070175
## [199]  31.1929825  -3.8070175  -1.8070175   5.1929825   1.1929825 -10.8070175
## [205]  31.1929825   0.1929825 -23.8070175 -12.8070175   2.1929825 -21.8070175
## [211]  -0.8070175  -6.8070175   3.1929825  -9.8070175   4.1929825 -11.8070175
## [217]  -1.8070175   2.1929825 -25.8070175  -9.8070175  -4.8070175  10.1929825
## [223]   2.1929825  -6.8070175   0.1929825   2.1929825   5.1929825  -4.8070175
## [229]   1.1929825 -18.8070175 -13.8070175   2.1929825   6.1929825 -25.8070175
## [235]  -8.8070175 -26.8070175 -14.8070175  30.1929825  -1.8070175   2.1929825
## [241] -25.8070175   1.1929825   3.1929825  15.1929825  -7.8070175  -1.8070175
## [247]  -2.8070175  -0.8070175  28.1929825  -4.8070175  -0.8070175   3.1929825
## [253]  23.1929825  -9.8070175  -8.8070175  -0.8070175 -14.8070175  -2.8070175
## [259]   3.1929825 -16.8070175  -7.8070175  -8.8070175 -19.8070175 -27.8070175
## [265]   2.1929825 -27.8070175  -6.8070175  18.1929825 -19.8070175   3.1929825
## [271]   1.1929825  -8.8070175  -2.8070175  23.1929825  17.1929825  25.1929825
## [277]   2.1929825 -14.8070175   2.1929825   2.1929825  26.1929825 -16.8070175
## [283] -16.8070175   8.1929825  -4.8070175  11.1929825  32.1929825 -18.8070175
## [289] -13.8070175 -20.8070175 -10.8070175 -15.8070175  26.1929825 -15.8070175
## [295]  24.1929825  21.1929825   3.1929825 -21.8070175  34.1929825 -22.8070175
## [301]   3.1929825 -25.8070175 -23.8070175   4.1929825  15.1929825  33.1929825
## [307]  19.1929825 -17.8070175 -16.8070175  -6.8070175 -27.8070175  24.1929825
## [313]  32.1929825  30.1929825 -14.8070175  -6.8070175   2.1929825 -20.8070175
## [319]   2.1929825  20.1929825   9.1929825 -22.8070175   0.1929825 -18.8070175
## [325]  11.1929825  -1.8070175 -27.8070175  25.1929825 -12.8070175  -0.8070175
## [331] -27.8070175 -13.8070175  33.1929825 -16.8070175  26.1929825 -26.8070175
## [337] -27.8070175  -6.8070175  29.1929825   1.1929825  -8.8070175  29.1929825
## [343]  29.1929825  -9.8070175 -11.8070175 -20.8070175 -21.8070175  20.1929825
## [349] -15.8070175 -19.8070175 -10.8070175 -25.8070175 -18.8070175  -7.8070175
## [355]   2.1929825  -1.8070175  13.1929825   1.1929825  -1.8070175   0.1929825
## [361] -19.8070175  32.1929825  -1.8070175   4.1929825 -23.8070175  27.1929825
## [367] -15.8070175 -26.8070175   3.1929825 -16.8070175  35.1929825  23.1929825
## [373]  -3.8070175 -26.8070175   3.1929825   1.1929825 -28.8070175  -1.8070175
## [379]  15.1929825  -5.8070175  24.1929825 -27.8070175  26.1929825 -20.8070175
## [385]  28.1929825  -5.8070175  -0.8070175   2.1929825   2.1929825 -23.8070175
## [391]  27.1929825   6.1929825 -23.8070175   4.1929825  24.1929825   3.1929825
## [397]   3.1929825  20.1929825  27.1929825
```

8. What is the average "difference between each age and the mean age"?

```
mean(age_diff) # Using the mean function to calculate - Instead of using the 'sd' function.
```

```
## [1] -1.623275e-15
```

```
# There is no relevance to calculating only this value although it is similar to calculating variance.
# It also doesn't make sense to work with negative numbers for age (absolute values would work)
# We could manually calculate the standard deviation by taking each number in the data set,
# subtracting the mean and squaring the results. We would then work out the sum
# of those squared differences to get the variance of the data.
# We would then take the square root of that to calculate the standard deviation.
```

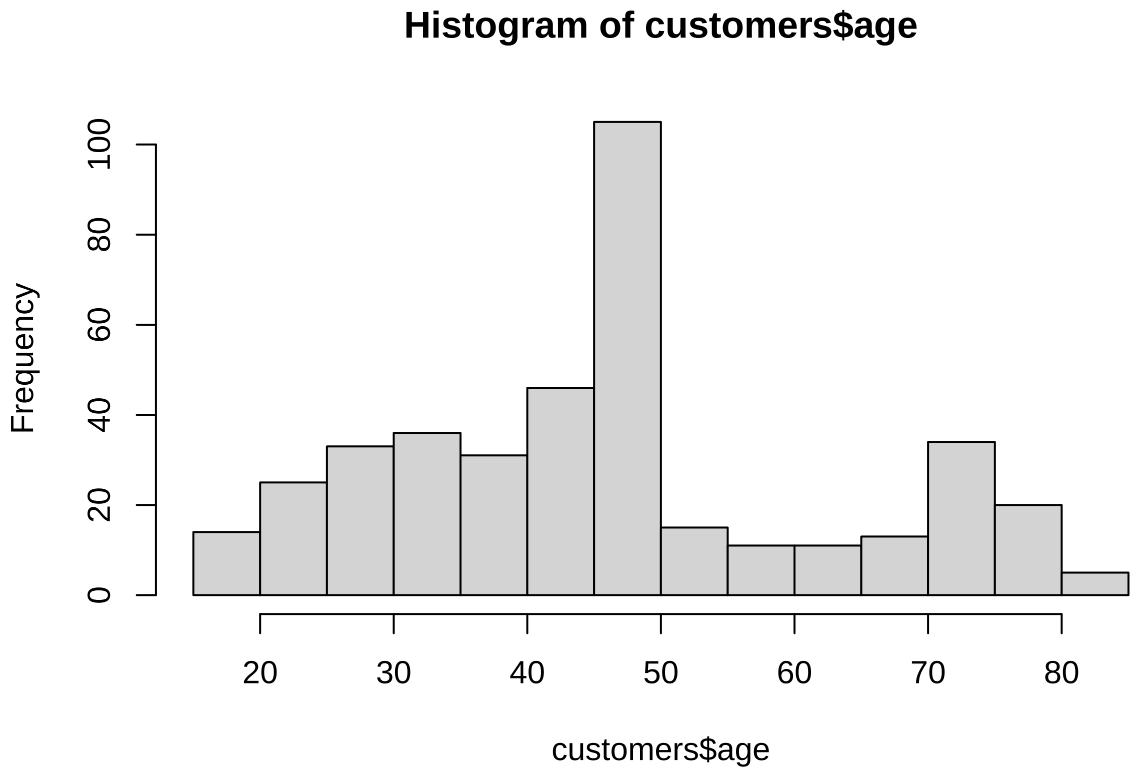
9. Visualize the raw data as we did in class: (a) histogram, (b) density plot, (c) boxplot+stripchart

a. histogram

```
customers$age <- as.numeric(customers$age) #Converting data type to numeric for visualization
class(customers$age) #Check class
```

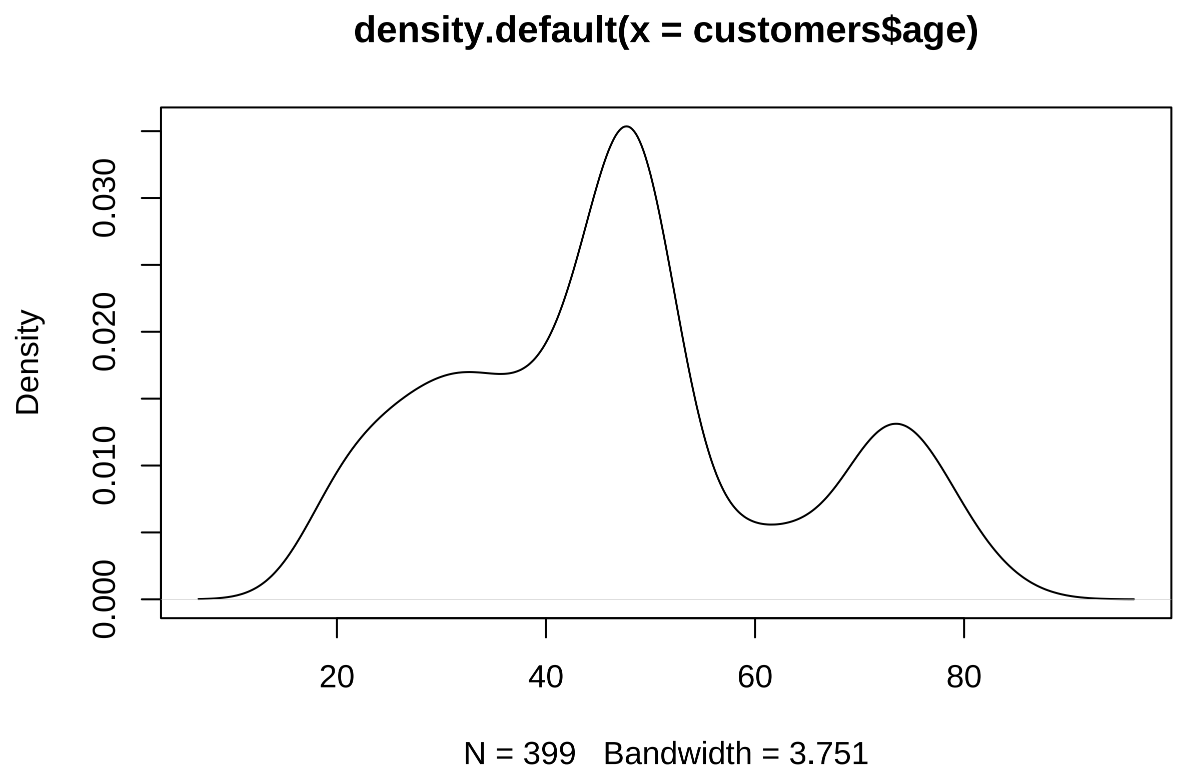
```
## [1] "numeric"
```

```
hist(customers$age) #histogram of raw data for age variable
```



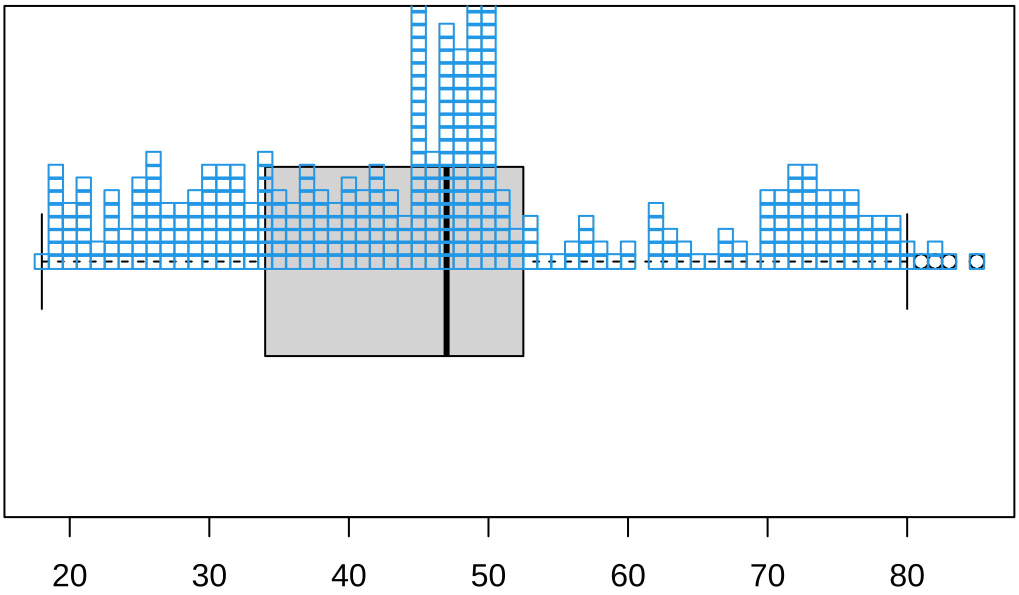
b. Density plot

```
plot(density(customers$age)) #density plot of raw data for age variable
```



c. Boxplot + stripchart

```
{boxplot(customers$age, horizontal = TRUE)
stripchart(customers$age, method = "stack", add = TRUE, col = 4)}
```



```
#Adding additional aesthetics to make the strip chart more visible.
```