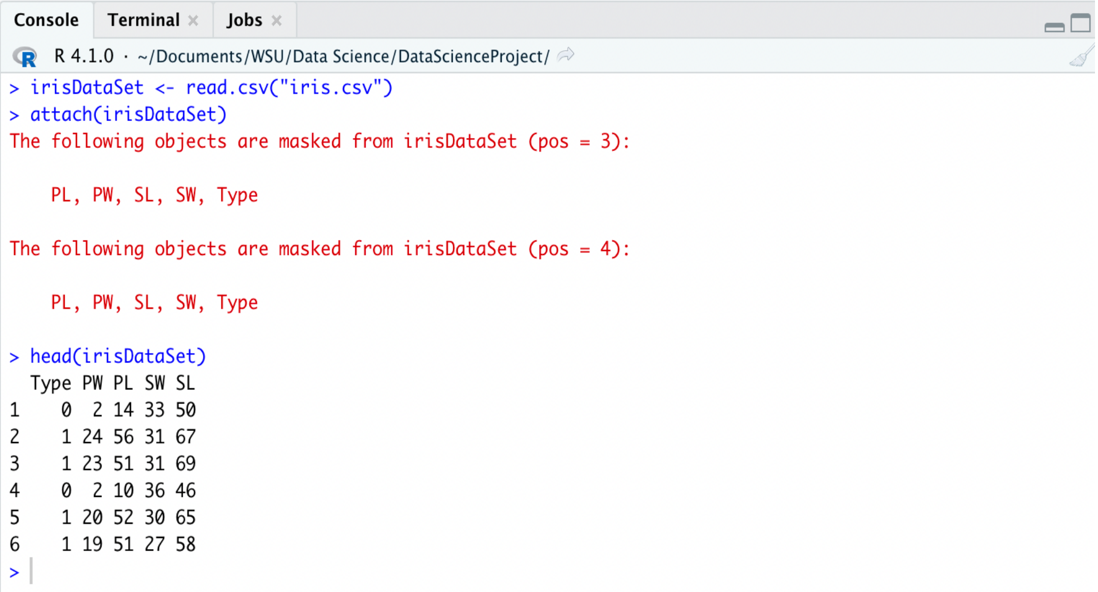
ASSIGNMENT – WEEK (1)

1. IRIS Dataset:

- Explore the variables:

Command:
irisDataSet <- read.csv("iris.csv")
attach(irisDataSet)
head(irisDataSet)

Output:



```
> irisDataSet <- read.csv("iris.csv")
> attach(irisDataSet)
The following objects are masked from irisDataSet (pos = 3):

    PL, PW, SL, SW, Type

The following objects are masked from irisDataSet (pos = 4):

    PL, PW, SL, SW, Type

> head(irisDataSet)
  Type PW PL SW SL
1    0  2 14 33 50
2    1 24 56 31 67
3    1 23 51 31 69
4    0  2 10 36 46
5    1 20 52 30 65
6    1 19 51 27 58
> 
```

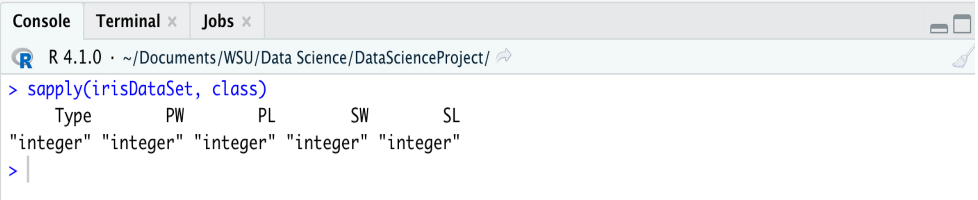
From this output we can say that Iris data set has 5 columns named Type, PW, PL, SW and SL.

Command:

sapply(irisDataSet, class)

This command gives us the data type for the variables.

Output:



```
> sapply(irisDataSet, class)
     Type        PW        PL        SW        SL 
"integer" "integer" "integer" "integer" "integer" 
> 
```

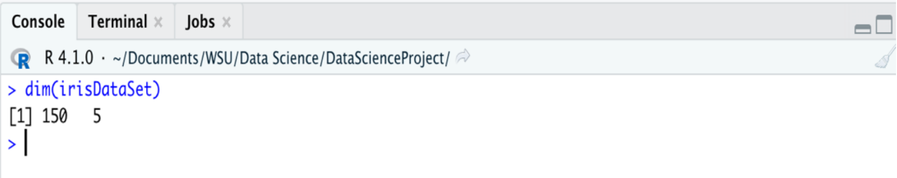
From the above output, it seems that every variable is integer variables.

Command:

dim(irisDataSet)

The above command gives us the dimensions of the data frame.

Output:



```
> dim(irisDataSet)
[1] 150   5
> 
```

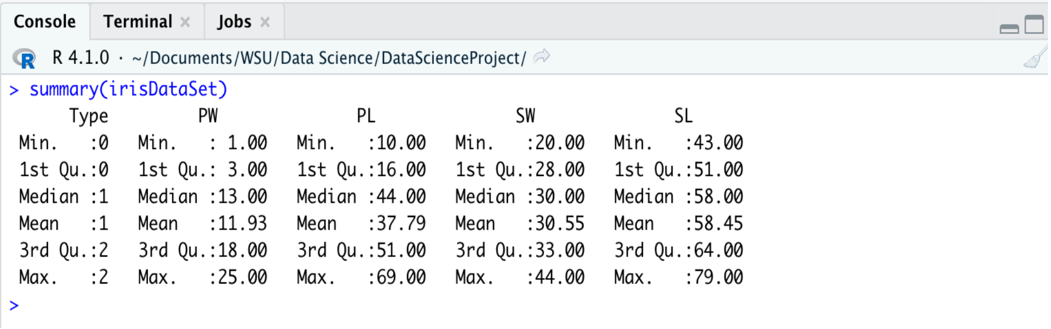
There are 150 observations for the 5 variables in the Iris dataset.

Command:

summary(irisDataSet)

data set. The summary includes details about min value, max value, 1st and 3rd Quadrant, median and mean value.

Output:



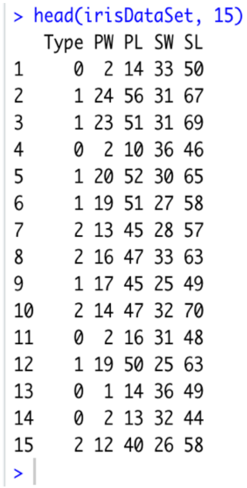
```
> summary(irisDataSet)
      Type          PW              PL              SW              SL       
 Min.   :0   Min.   : 1.00   Min.   :10.00   Min.   :20.00   Min.   :43.00  
 1st Qu.:0   1st Qu.: 3.00   1st Qu.:16.00   1st Qu.:28.00   1st Qu.:51.00  
 Median :1   Median :13.00   Median :44.00   Median :30.00   Median :58.00  
 Mean   :1   Mean   :11.93   Mean   :37.79   Mean   :30.55   Mean   :58.45  
 3rd Qu.:2   3rd Qu.:18.00   3rd Qu.:51.00   3rd Qu.:33.00   3rd Qu.:64.00  
 Max.   :2   Max.   :25.00   Max.   :69.00   Max.   :44.00   Max.   :79.00  
>
```

From the above summary command, it seems that Type variable has just three values 0,1 and 2 while other variables are well distributed.

- List the Quantitative variables and Qualitative variables:

  Since we have attached the data set at the beginning, we can directly call the irisDataSet's variable and use them.

  So, by looking at the data set in more deep let's just say first 15 rows by using command head(irisDataSet, 15) we can conclude that:



```
> head(irisDataSet, 15)
   Type PW PL SW SL
1     0  2 14 33 50
2     1 24 56 31 67
3     1 23 51 31 69
4     0  2 10 36 46
5     1 20 52 30 65
6     1 19 51 27 58
7     2 13 45 28 57
8     2 16 47 33 63
9     1 17 45 25 49
10    2 14 47 32 70
11    0  2 16 31 48
12    1 19 50 25 63
13    0  1 14 36 49
14    0  2 13 32 44
15    2 12 40 26 58
> 
```

Qualitative Variable – Type
Quantitative Variable – PW, PL, SW, SL

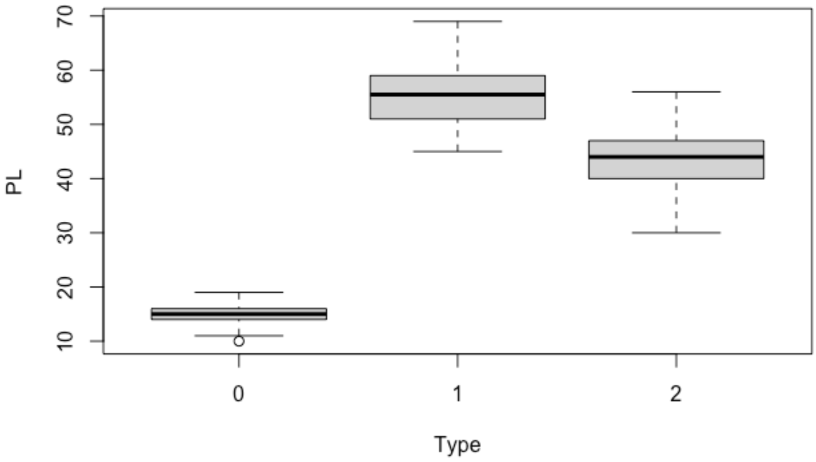
Type is Qualitative because it has only 3 set of values and can be looked as a factor/categorical data while other variables are numeric and thus identified as the Quantitative variables.

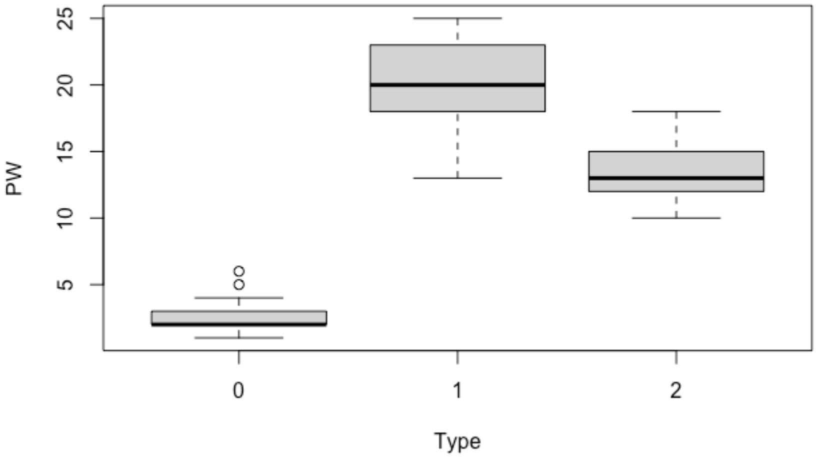
- State a research question and identify the target variable if applicable:

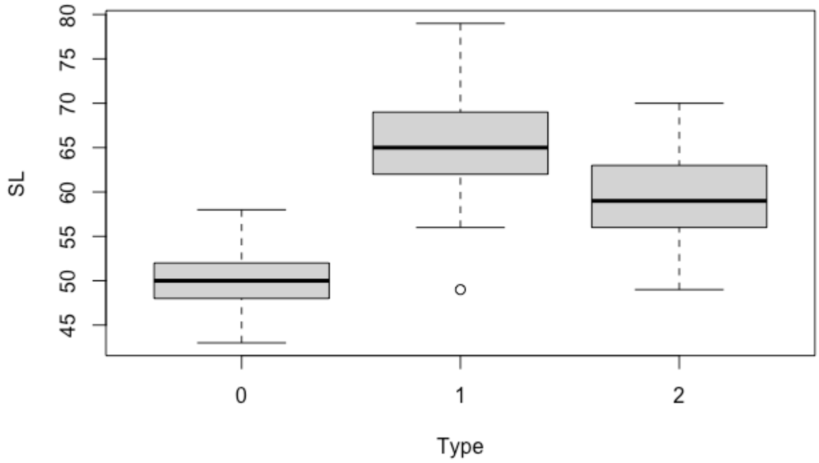
Research question: What type of the iris flowers are smallest and largest in terms of sepals and petals?

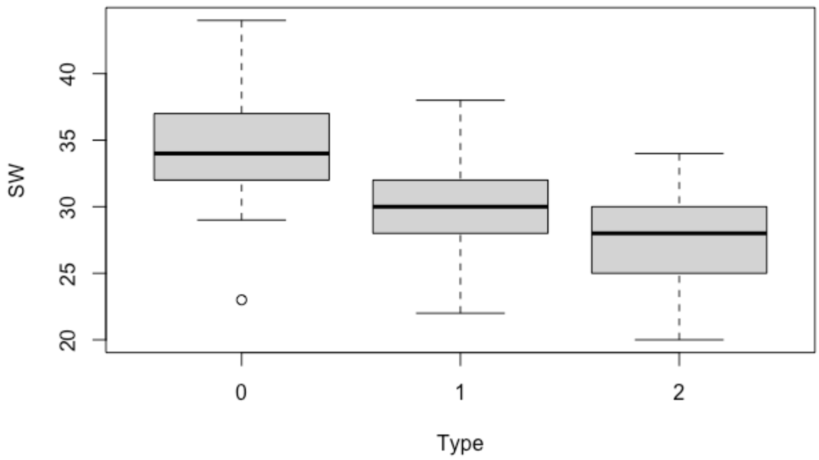
Target variable: The target variable suitable for the above question will be "**Type**".

Analysis:









From the above boxplots, following points are observed:

- Type 0 iris flowers are the smallest flowers, but they are much wider sepals as compared to others.
- Comparatively, Type 1 iris flowers are the largest flowers.
- While Type 2 flowers are in between both type.

- Comment if they are Supervised learning or unsupervised learning:

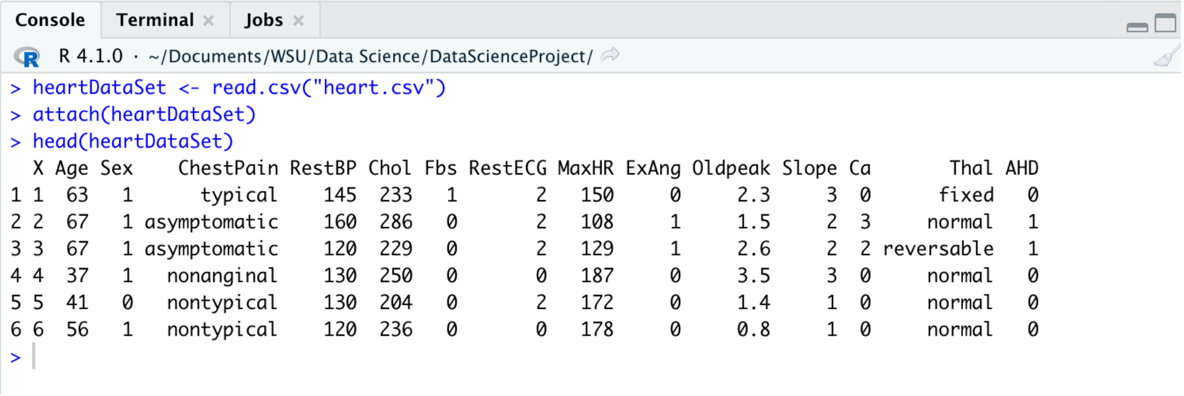
  By looking at the dataset, we can find that one special variable which can provide us the outcome and that is the variable “Type”. Therefore, this concludes that Iris dataset is supervised learning.

2. Heart DataSet:

- Explore the variables:

  Command:
  heartDataSet <- read.csv("heart.csv")
  attach(heartDataSet)
  head(heartDataSet)

  Output:



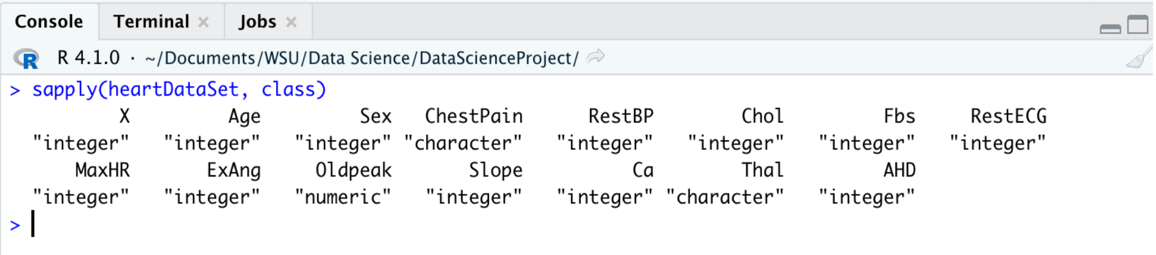
```
> heartDataSet <- read.csv("heart.csv")
> attach(heartDataSet)
> head(heartDataSet)
  X Age Sex    ChestPain RestBP Chol Fbs RestECG MaxHR ExAng Oldpeak Slope Ca       Thal AHD
1 1  63   1      typical    145  233   1       2   150     0     2.3     3  0      fixed   0
2 2  67   1 asymptomatic    160  286   0       2   108     1     1.5     2  3     normal   1
3 3  67   1 asymptomatic    120  229   0       2   129     1     2.6     2  2 reversable   1
4 4  37   1   nonanginal    130  250   0       0   187     0     3.5     3  0     normal   0
5 5  41   0   nontypical    130  204   0       2   172     0     1.4     1  0     normal   0
6 6  56   1   nontypical    120  236   0       0   178     0     0.8     1  0     normal   0
> 
```

  From this output we can say that Iris data set has 15 columns named X, Age, Sex, ChestPain, RestBP, Chol, Fbs, RestECG, MaxHR, ExAng, Oldpeak, Slope, Ca, Thal and AHD.

  Command:

  sapply(heartDataSet, class)

Output:



```
> sapply(heartDataSet, class)
          X         Age         Sex   ChestPain      RestBP        Chol         Fbs     RestECG 
  "integer"   "integer"   "integer" "character"   "integer"   "integer"   "integer"   "integer" 
      MaxHR       ExAng     Oldpeak       Slope          Ca        Thal         AHD 
  "integer"   "integer"   "numeric"   "integer"   "integer" "character"   "integer" 
> |
```

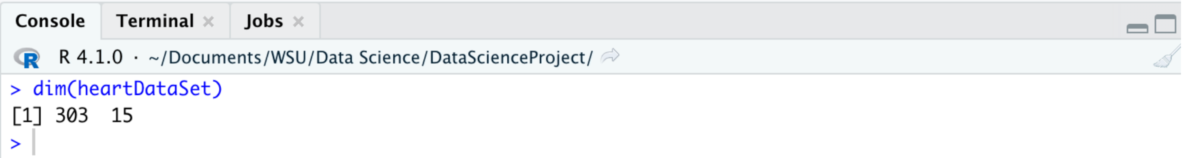
From the above output, it seems that we have a mix of data types in the data sets.

Command:

dim(heartDataSet)

The above command gives us the dimensions of the data frame.

Output:



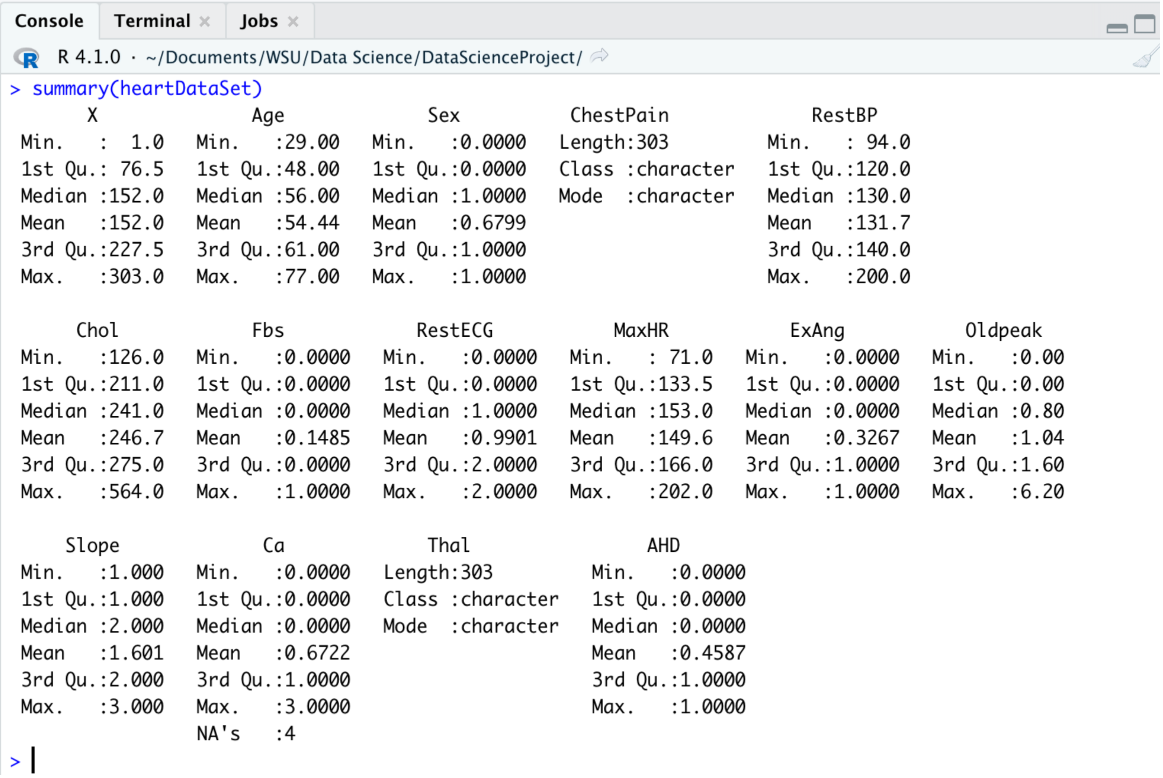
```
> dim(heartDataSet)
[1] 303  15
> |
```

There are 303 observations for 15 variables in the Heart dataset.

Command:

summary(heartDataSet)

Output:



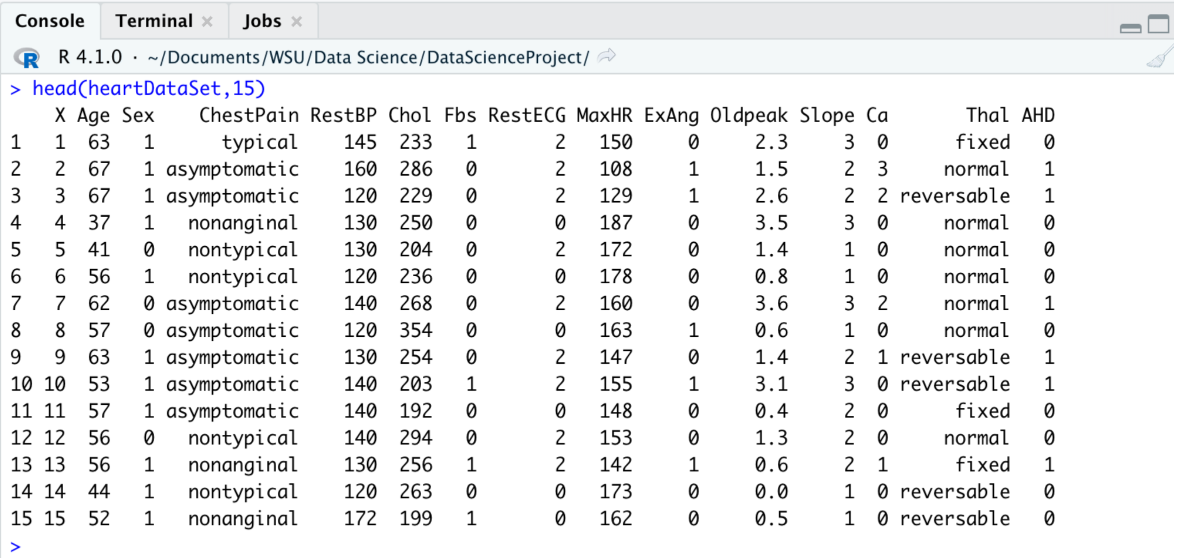
```
> summary(heartDataSet)
      X              Age             Sex          ChestPain             RestBP
Min.   :  1.0   Min.   :29.00   Min.   :0.0000   Length:303         Min.   : 94.0
1st Qu.: 76.5   1st Qu.:48.00   1st Qu.:0.0000   Class :character   1st Qu.:120.0
Median :152.0   Median :56.00   Median :1.0000   Mode  :character   Median :130.0
Mean   :152.0   Mean   :54.44   Mean   :0.6799                      Mean   :131.7
3rd Qu.:227.5   3rd Qu.:61.00   3rd Qu.:1.0000                      3rd Qu.:140.0
Max.   :303.0   Max.   :77.00   Max.   :1.0000                      Max.   :200.0

     Chol            Fbs            RestECG           MaxHR           ExAng           Oldpeak
Min.   :126.0   Min.   :0.0000   Min.   :0.0000   Min.   : 71.0   Min.   :0.0000   Min.   :0.00
1st Qu.:211.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:133.5   1st Qu.:0.0000   1st Qu.:0.00
Median :241.0   Median :0.0000   Median :1.0000   Median :153.0   Median :0.0000   Median :0.80
Mean   :246.7   Mean   :0.1485   Mean   :0.9901   Mean   :149.6   Mean   :0.3267   Mean   :1.04
3rd Qu.:275.0   3rd Qu.:0.0000   3rd Qu.:2.0000   3rd Qu.:166.0   3rd Qu.:1.0000   3rd Qu.:1.60
Max.   :564.0   Max.   :1.0000   Max.   :2.0000   Max.   :202.0   Max.   :1.0000   Max.   :6.20

    Slope             Ca             Thal                AHD
Min.   :1.000   Min.   :0.0000   Length:303         Min.   :0.0000
1st Qu.:1.000   1st Qu.:0.0000   Class :character   1st Qu.:0.0000
Median :2.000   Median :0.0000   Mode  :character   Median :0.0000
Mean   :1.601   Mean   :0.6722                      Mean   :0.4587
3rd Qu.:2.000   3rd Qu.:1.0000                      3rd Qu.:1.0000
Max.   :3.000   Max.   :3.0000                      Max.   :1.0000
                NA's   :4
>
```

From the above summary command, it seems that "AHD" variable has just two values as 0 and 1while other variables are well distributed.

- List the Quantitative variables and Qualitative variables:

Since we have attached the data set at the beginning, we can directly call the heartDataSet's variable and use them.

So, by looking at the data set in more deep let's just say first 15 rows by using command head(heartDataSet, 15) we can conclude that:



```
> head(heartDataSet,15)
    X Age Sex    ChestPain RestBP Chol Fbs RestECG MaxHR ExAng Oldpeak Slope Ca       Thal AHD
1   1  63   1      typical    145  233   1       2   150     0     2.3     3  0      fixed   0
2   2  67   1 asymptomatic    160  286   0       2   108     1     1.5     2  3     normal   1
3   3  67   1 asymptomatic    120  229   0       2   129     1     2.6     2  2 reversable   1
4   4  37   1   nonanginal    130  250   0       0   187     0     3.5     3  0     normal   0
5   5  41   0   nontypical    130  204   0       2   172     0     1.4     1  0     normal   0
6   6  56   1   nontypical    120  236   0       0   178     0     0.8     1  0     normal   0
7   7  62   0 asymptomatic    140  268   0       2   160     0     3.6     3  2     normal   1
8   8  57   0 asymptomatic    120  354   0       0   163     1     0.6     1  0     normal   0
9   9  63   1 asymptomatic    130  254   0       2   147     0     1.4     2  1 reversable   1
10 10  53   1 asymptomatic    140  203   1       2   155     1     3.1     3  0 reversable   1
11 11  57   1 asymptomatic    140  192   0       0   148     0     0.4     2  0      fixed   0
12 12  56   0   nontypical    140  294   0       2   153     0     1.3     2  0     normal   0
13 13  56   1   nonanginal    130  256   1       2   142     1     0.6     2  1      fixed   1
14 14  44   1   nontypical    120  263   0       0   173     0     0.0     1  0 reversable   0
15 15  52   1   nonanginal    172  199   1       0   162     0     0.5     1  0 reversable   0
>
```

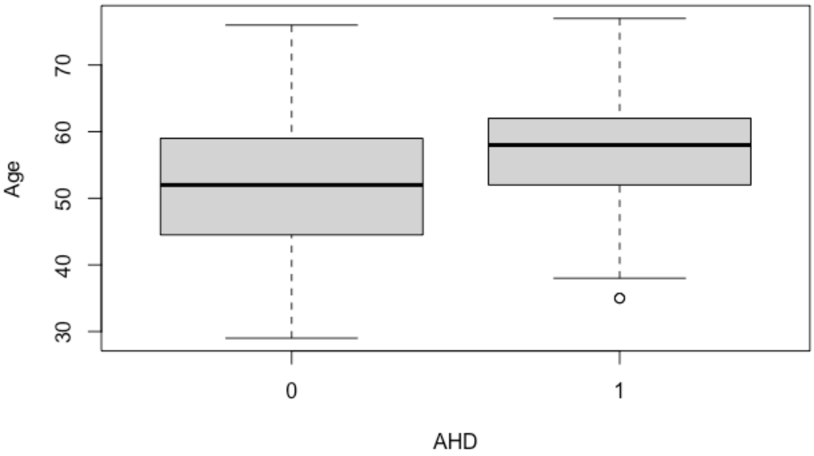
Qualitative Variable – Sex, ChestPain, Fbs, RestECG, ExAng, Ca, Thal, AHD
Quantitative Variable – X, Age, RestBP, Chol, MaxHR, Oldpeak

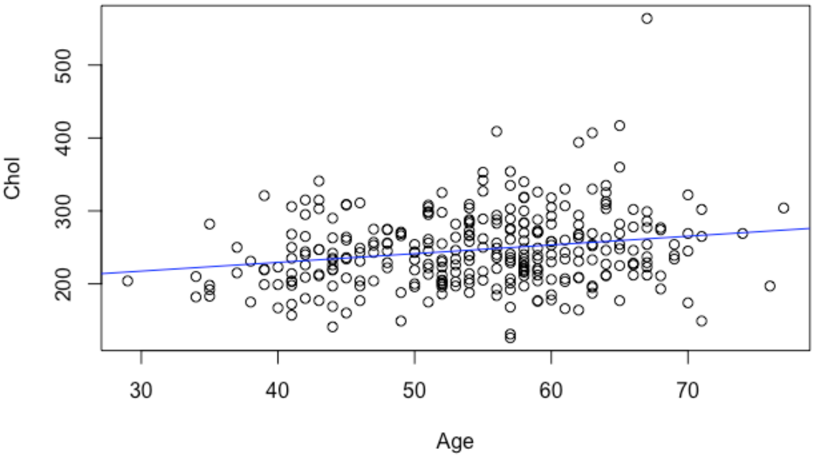
- State a research question and identify the target variable if applicable:

  Research question: With the increase of Age, are we more likely to have a heart disease?

  Target variable: The target variable suitable for the above question will be "**AHD**".

Analysis:





From the above graphs, we can conclude that as per the data if the Age is higher then there is more like to get a heart disease. It can also be confirmed by looking Cholesterol vs Age scatterplot. The linear line depicts that with the increase of Age, cholesterol level may also increase.

- Comment if they are Supervised learning or unsupervised learning:

By looking at the dataset, we can find that one special variable which can provide us the outcome and that is the variable “AHD”. Therefore, this concludes that Heart dataset is supervised learning.

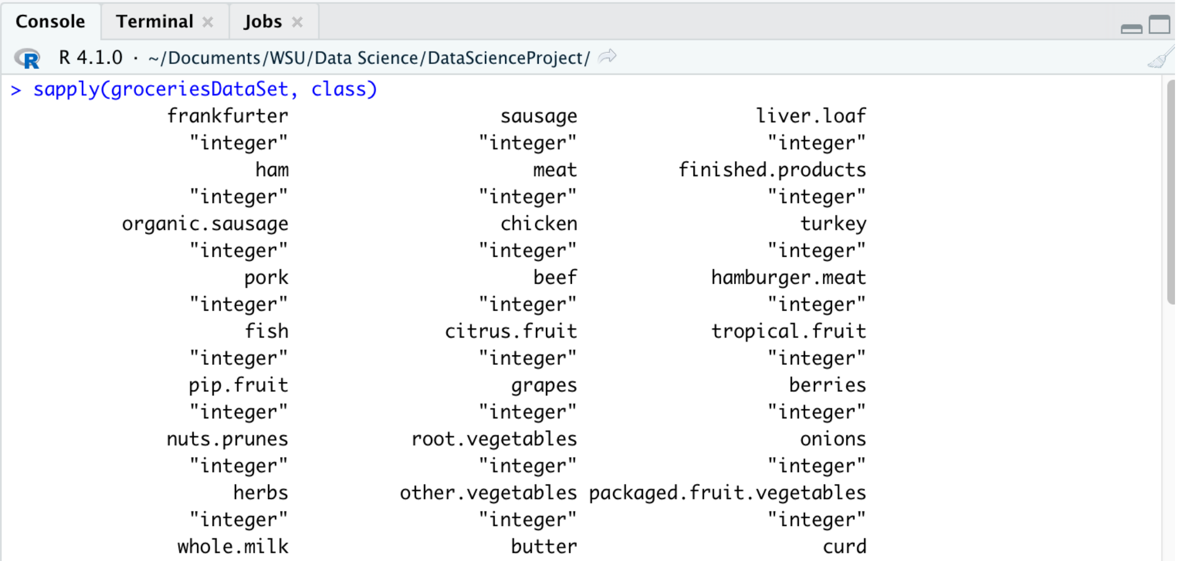
3. Groceries DataSet:

- Explore the variables:

Command:

```
groceriesDataSet <- read.csv("groceries.csv")
attach(groceriesDataSet)
head(groceriesDataSet)
sapply(groceriesDataSet, class)
```

Output:



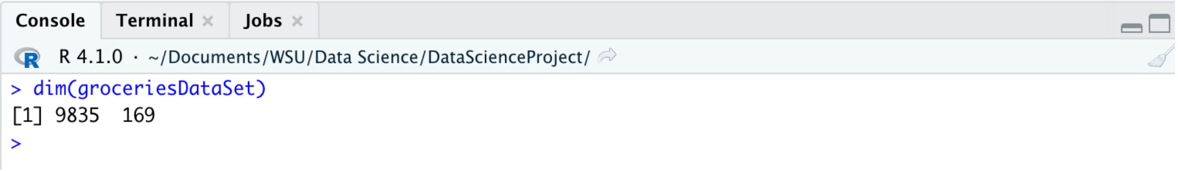
```
> sapply(groceriesDataSet, class)
              frankfurter                   sausage                liver.loaf
                "integer"                 "integer"                 "integer"
                      ham                      meat         finished.products
                "integer"                 "integer"                 "integer"
          organic.sausage                   chicken                    turkey
                "integer"                 "integer"                 "integer"
                     pork                      beef            hamburger.meat
                "integer"                 "integer"                 "integer"
                     fish              citrus.fruit            tropical.fruit
                "integer"                 "integer"                 "integer"
                pip.fruit                    grapes                   berries
                "integer"                 "integer"                 "integer"
              nuts.prunes           root.vegetables                    onions
                "integer"                 "integer"                 "integer"
                    herbs          other.vegetables packaged.fruit.vegetables
                "integer"                 "integer"                 "integer"
               whole.milk                    butter                      curd
```

From this sapply command, it is clear that each and every data in the dataset is in the form of integers.

Command:

```
dim(groceriesDataSet)
```

Output:



```
R 4.1.0 · ~/Documents/WSU/Data Science/DataScienceProject/
> dim(groceriesDataSet)
[1] 9835   169
>
```

From above output it is clear that there are 9835 observations for the 169 variables in the Groceries dataset.

- List the Quantitative variables and Qualitative variables:

  So, by looking at the data set in more deep let’s just say first 15 rows by using command head(groceriesDataSet, 15) we can conclude that each and every variables can be categorized because they all have either 0s or 1s as their values. This all the variables are **Qualitative** variables.

- State a research question and identify the target variable if applicable:

  Research question: What are the top 10 highly purchased products by consumer?

  Target variable: There are no such target variables since each and every variable is either 0s or 1s, thus we need to look for patterns.

Analysis:

Script:

```
        accumulatedTotal =
        colSums(groceriesDataSet[,-1])

        accumulatedTotalDescending = sort(x,
        decreasing = TRUE)

        accumulatedTotalDescending[0:10]

        barplot(accumulatedTotalDescending[0:10],
        las=2)
```

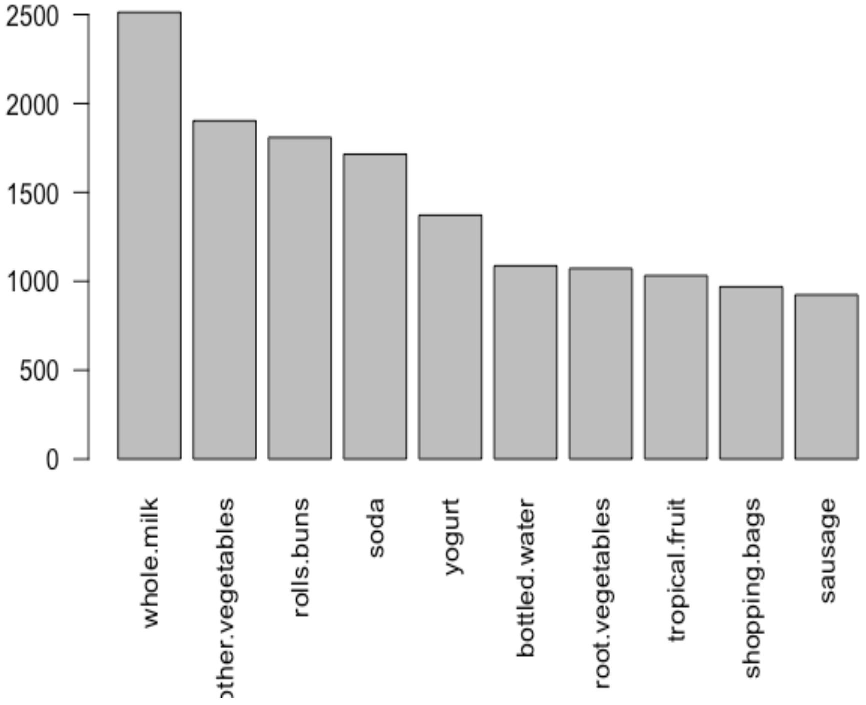
What these scripts are doing is that, taking a sum of all the values in each column. We are doing this because each variable consists of either 0 or 1 so purchased product means that value will be 1. Therefore, by adding all the values we can fetch total number of purchases for that variable and then will repeat this for all the columns.

After adding them, we will get a data frame for all the products and the number of times that product is purchased.

Next, we want to sort them in descending order so that highest purchased product comes at first.

Lastly, a bar graph to plot and see the top 10 purchased item.

Output:



From the above graph, we can see that milk was the most bought item followed by vegetables and so on.

- Comment if they are Supervised learning or unsupervised learning:

  By looking at the dataset, we can find that there is no such special variable that we can use to resolve our research question. Hence, we need to look for patterns and visualization to find the desired result.

  Therefore, we can say that Groceries dataset is an unsupervised learning.