STA 546 : STATISTICAL DATA MINING 2

HOMEWORK 1

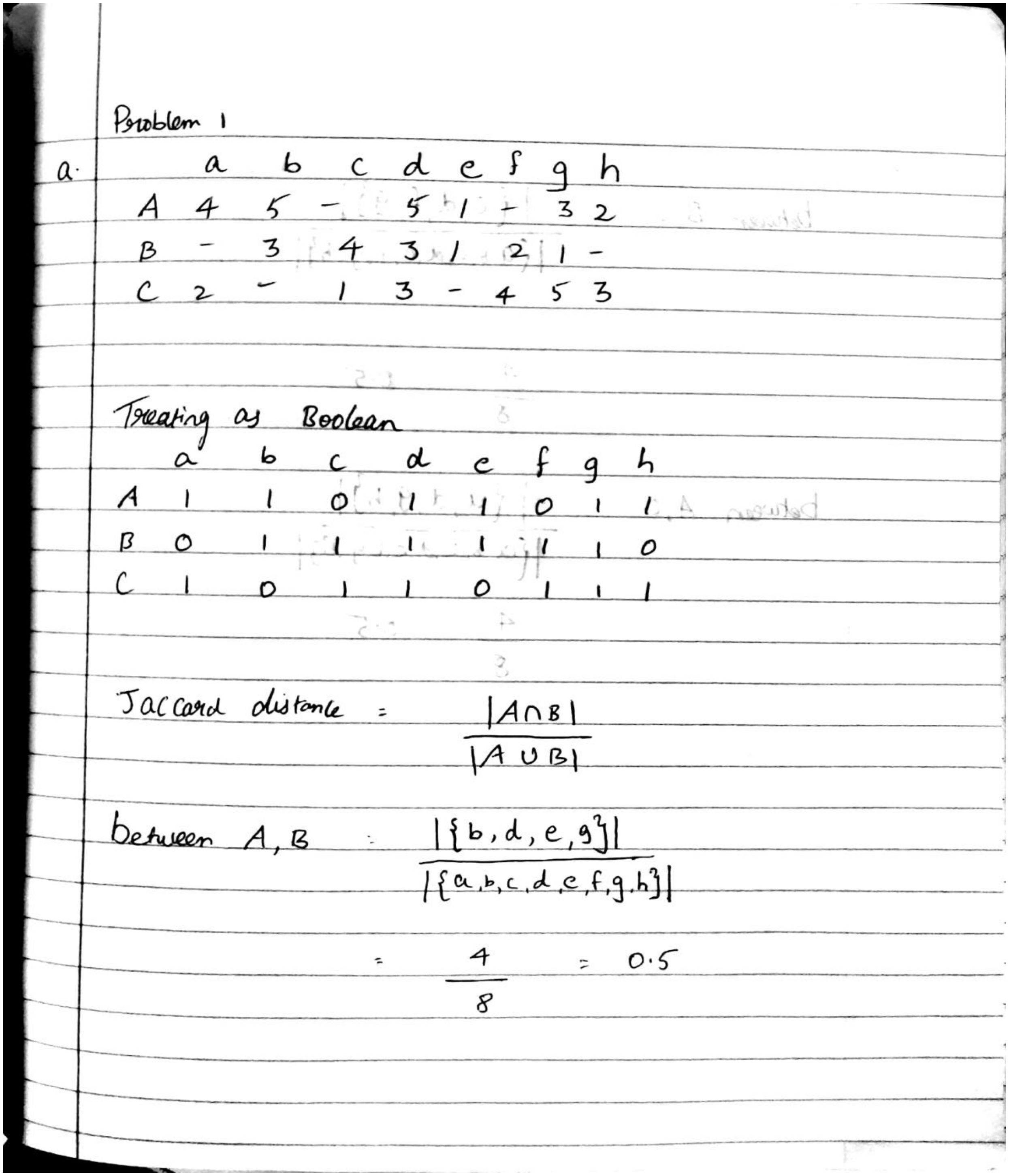
BY

ABINESH SENTHIL KUMAR

#50320934

Question 1

(a) Treat the utility matrix as Boolean and compute the Jaccard distance, and the cosine distance between users.



Problem 1

a.

$$\begin{array}{c|cccccccc} & a & b & c & d & e & f & g & h \\ A & 4 & 5 & - & 5 & 1 & - & 3 & 2 \\ B & - & 3 & 4 & 3 & 1 & 2 & 1 & - \\ C & 2 & - & 1 & 3 & - & 4 & 5 & 3 \end{array}$$

Treating as Boolean

$$\begin{array}{c|cccccccc} & a & b & c & d & e & f & g & h \\ A & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\ B & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ C & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \end{array}$$

Jaccard distance $= \dfrac{|A \cap B|}{|A \cup B|}$

between A, B $= \dfrac{|\{b, d, e, g\}|}{|\{a, b, c, d, e, f, g, h\}|}$

$= \dfrac{4}{8} = 0.5$

between B, C $= \dfrac{|\{c, d, f, g\}|}{|\{a, b, c, d, e, f, g, h\}|}$

$= \dfrac{4}{8} = 0.5$

between A, C $= \dfrac{|\{a, d, g, h\}|}{|\{a, b, c, d, e, f, g, h\}|}$

$= \dfrac{4}{8} = 0.5$

Cosine similarity

$$\text{Cosine similarity} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

for (A, B)

$A = \{1, 1, 0, 1, 1, 0, 1, 1\}$

$B = \{0, 1, 1, 1, 1, 1, 1, 0\}$ ×

$$\text{Cosine}(A,B) = \frac{(1\times 0)+(1\times 1)+(0\times 1)+(1\times 1)+(1\times 1)+(0\times 1)+(1\times 1)+(1\times 0)}{\sqrt{(1)^2+(1)^2+(1)^2+(1)^2+(1)^2+(1)^2} \times \sqrt{(1)^2+(1)^2+(1)^2+(1)^2+(1)^2+(1)^2}}$$

$$= \frac{1+1+1+1}{\sqrt{6} \times \sqrt{6}} = \frac{4}{6} = \frac{2}{3}$$

$$= 0.66$$

for (B,c)

B = {0, 1,1,1,1,1,1,0}

C = {1,0,1,1,0,1,1,1}

$$\text{Cosine}(B,C) = \frac{(0\times1)+(1\times0)+(1\times1)+(1\times1)+(1\times0)+(1\times1)+(1\times1)+(0\times1)}{\sqrt{(1)^2+(1)^2+(1)^2+(1)^2+(1)^2+(1)^2} \times \sqrt{(1)^2+(1)^2+(1)^2+(1)^2+(1)^2+(1)^2}}$$

$$\frac{4}{\sqrt{6}\times\sqrt{6}} = \frac{4}{6} = \frac{2}{3}$$

$= 0.66$

for (A,C)

$A = \{1,1,0,1,1,0,1,1\}$

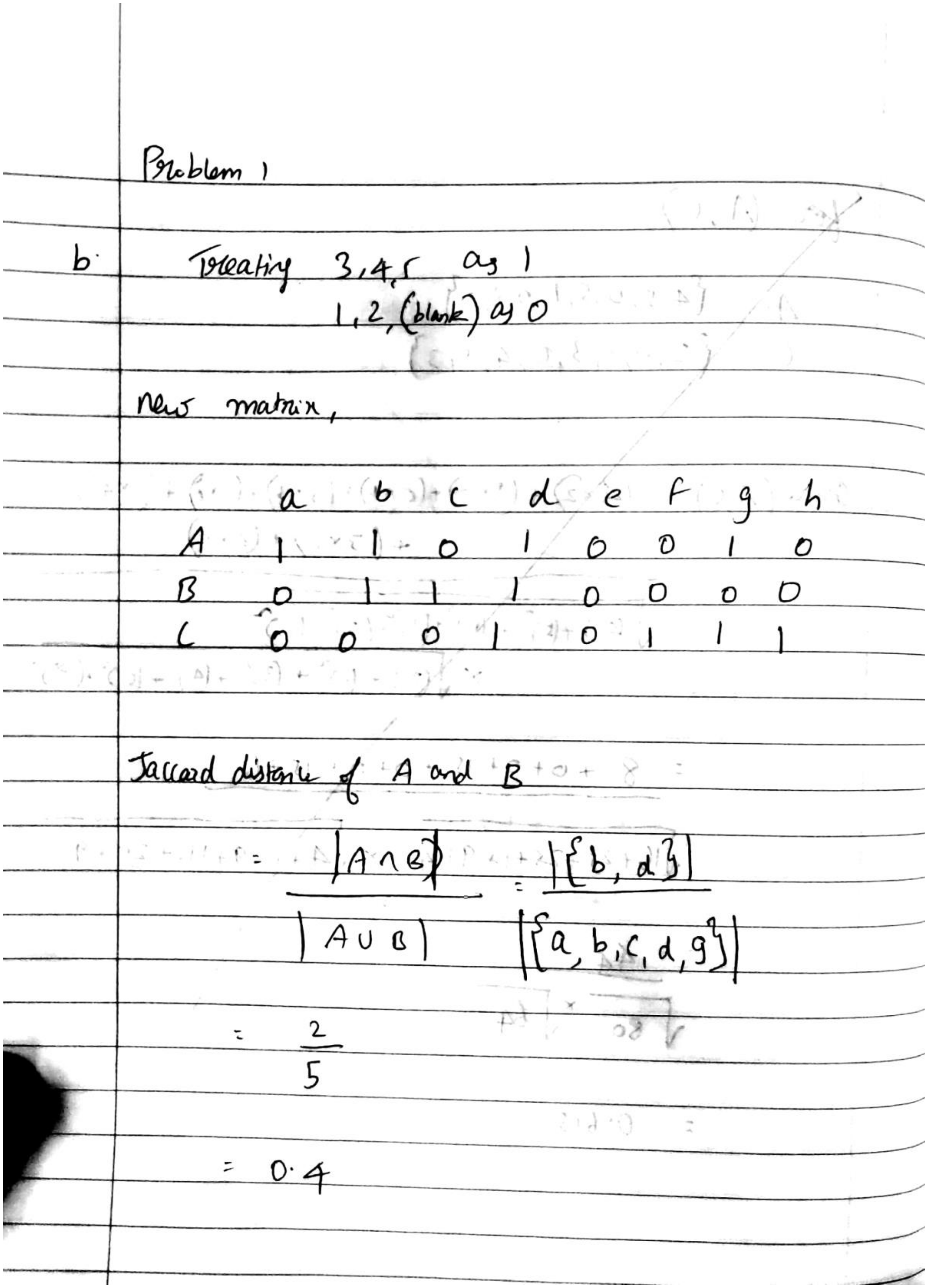
$C = \{1,0,1,1,0,1,1,1\}$

$$\text{Cosine}(A,C) = \frac{(1\times1)+(1\times0)+(0\times1)+(1\times1)+(1\times0)+(0\times1)+(1\times1)+(1\times1)}{\sqrt{6}\times\sqrt{6}}$$

$$= \frac{4}{6} = \frac{2}{3}$$

$= 0.66$

(b) Use a different discretization: treat ratings 3,4,5 as 1, and ratings 1, 2, and blank as 0. Compute the Jaccard distance and cosine distance and compare to that of part A.



Problem 1

b. Treating 3,4,5 as 1
1,2,(blank) as 0

new matrix,

$$\begin{array}{c|cccccccc} & a & b & c & d & e & f & g & h \\ \hline A & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ B & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ C & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{array}$$

Jaccard distance of A and B

$$= \frac{|A \cap B|}{|A \cup B|} = \frac{|\{b, d\}|}{|\{a, b, c, d, g\}|}$$

$$= \frac{2}{5}$$

$$= 0.4$$

Jaccard distance of B and C

~~= |{b, c, d~~

$$= \frac{|\{d\}|}{|\{b, c, d, f, g, h\}|} = \frac{1}{6}$$

= 0.1666

×

Jaccard distance of A and C

$$= \frac{|\{g, d\}|}{|\{a, b, d, f, g, h\}|} = \frac{2}{6} = \frac{1}{3}$$

= 0. 3333

Cosine distance of (A,B)

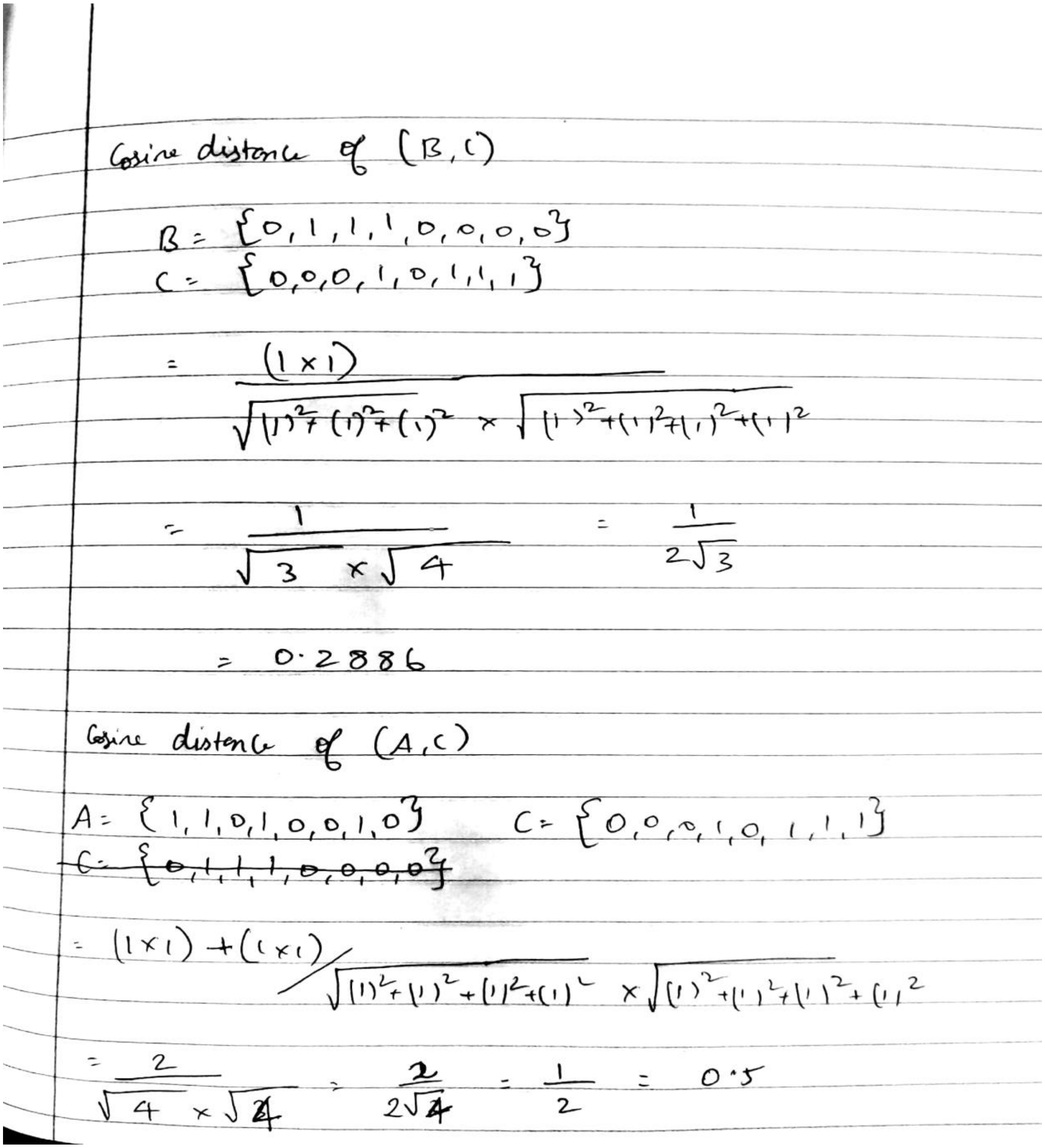
$A = \{1,1,0,1,0,0,1,0\}$

$B = \{0,1,1,1,0,0,0,0\}$

$$= \frac{(1\times0)+(1\times1)+(0\times1)+(1\times1)+(0\times0)+(0\times0)+(1\times0)+(0\times0)}{\sqrt{(1)^2+(1)^2+(1)^2+(1)^2}\times\sqrt{(1)^2+(1)^2+(1)^2}}$$

$$= \frac{1+1+0}{\sqrt{4}\times\sqrt{3}} = \frac{2}{2\sqrt{3}} = \frac{1}{\sqrt{3}}$$

$= 0.5773$



Cosine distance of (B, C)

$B = \{0,1,1,1,0,0,0,0\}$

$C = \{0,0,0,1,0,1,1,1\}$

$$= \frac{(1 \times 1)}{\sqrt{(1)^2+(1)^2+(1)^2} \times \sqrt{(1)^2+(1)^2+(1)^2+(1)^2}}$$

$$= \frac{1}{\sqrt{3} \times \sqrt{4}} = \frac{1}{2\sqrt{3}}$$

$$= 0.2886$$

Cosine distance of (A, C)

$A = \{1,1,0,1,0,0,1,0\}$ $C = \{0,0,0,1,0,1,1,1\}$

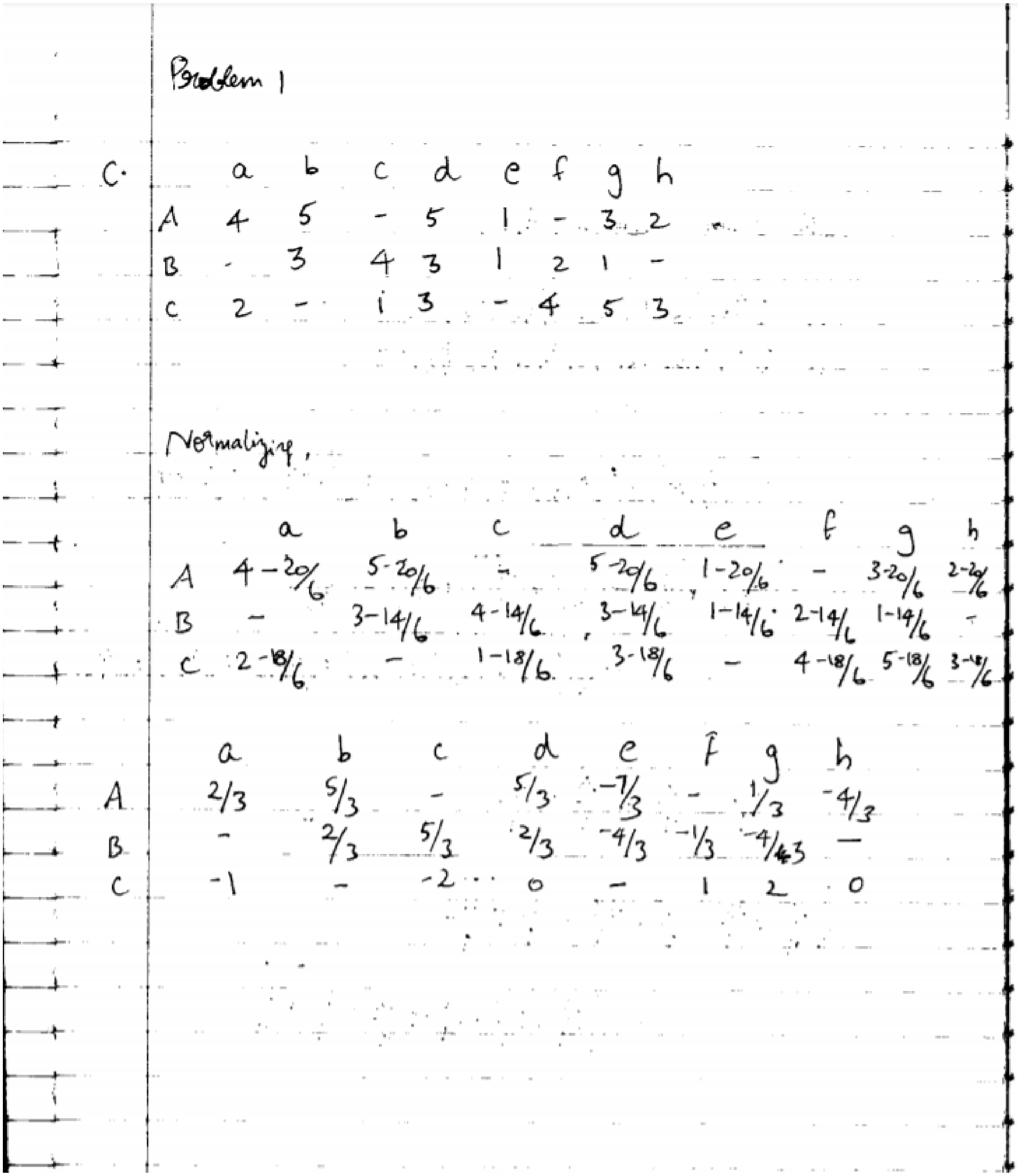
~~C = {0,1,1,1,0,0,0,0}~~

$$= \frac{(1 \times 1)+(1 \times 1)}{\sqrt{(1)^2+(1)^2+(1)^2+(1)^2} \times \sqrt{(1)^2+(1)^2+(1)^2+(1)^2}}$$

$$= \frac{2}{\sqrt{4} \times \sqrt{4}} = \frac{2}{2\sqrt{4}} = \frac{1}{2} = 0.5$$

Comparing the jaccard distance and cosine distance to that of part A, we can see that both cosine as well as the jaccard distance between the users is **less** when compared to that of part a.

(b) Normalize the matrix by subtracting from each nonblank entry the average value for its user. Using this matrix, compute the cosine distance between each pair of users.



Problem 1

C.

	a	b	c	d	e	f	g	h
A	4	5	-	5	1	-	3	2
B	-	3	4	3	1	2	1	-
C	2	-	1	3	-	4	5	3

Normalizing,

	a	b	c	d	e	f	g	h
A	4-20/6	5-20/6	-	5-20/6	1-20/6	-	3-20/6	2-20/6
B	-	3-14/6	4-14/6	3-14/6	1-14/6	2-14/6	1-14/6	-
C	2-18/6	-	1-18/6	3-18/6	-	4-18/6	5-18/6	3-18/6

	a	b	c	d	e	f	g	h
A	2/3	5/3	-	5/3	-7/3	-	1/3	-4/3
B	-	2/3	5/3	2/3	-4/3	-1/3	-4/3	—
C	-1	-	-2	0	-	1	2	0

Cosine distance between (A, B)

$$A = \left\{\frac{2}{3}, \frac{5}{3}, 0, \frac{5}{3}, \frac{-7}{3}, 0, \frac{1}{3}, \frac{-4}{3}\right\}$$

$$B = \left\{0, \frac{2}{3}, \frac{5}{3}, \frac{2}{3}, \frac{-4}{3}, \frac{-1}{3}, \frac{-4}{3}, 0\right\}$$

$$= \frac{\left(\frac{2}{3}\times 0\right) + \left(\frac{5}{3}\times\frac{2}{3}\right) + \left(\frac{5}{3}\times\frac{2}{3}\right) + \left(\frac{7}{3}\times\frac{4}{3}\right) - \left(\frac{1}{3}\times\frac{4}{3}\right)}{\sqrt{\left(\frac{2}{3}\right)^2 + \left(\frac{5}{3}\right)^2 + \left(\frac{5}{3}\right)^2 + \left(\frac{-7}{3}\right)^2 + \left(\frac{1}{3}\right)^2 + \left(\frac{-4}{3}\right)^2}}$$

$$\times \sqrt{\left(\frac{2}{3}\right)^2 + \left(\frac{5}{3}\right)^2 + \left(\frac{2}{3}\right)^2 + \left(\frac{-4}{3}\right)^2 + \left(\frac{-1}{3}\right)^2 + (\frac{-4}{3}}$$

$$= \frac{\frac{10}{9} + \frac{10}{9} + \frac{28}{9} - \frac{4}{9}}{\sqrt{\frac{4}{9} + \frac{25}{9} + \frac{25}{9} + \frac{49}{9} + \frac{1}{9} + \frac{16}{9}} \times \sqrt{\frac{4}{9} + \frac{25}{9} + \frac{4}{9} + \frac{16}{9} + \frac{1}{9} + \frac{16}{9}}}$$

$$= \frac{44/9}{\sqrt{120/9} \times \sqrt{66/9}}$$

$$= \frac{4.888}{9.888} = 0.4943$$

Cosine distance between (B, C)

$$B = \{0, 2/3, 5/3, 2/3, -4/3, -1/3, -4/3, 0\}$$
$$C = \{-1, 0, -2, 0, 0, 1, 2, 0\}$$

$$= \frac{(5/3 \times -2) + (-1/3 \times 1) + (-4/3 \times 2)}{\sqrt{(2/3)^2 + (5/3)^2 + (2/3)^2 + (-4/3)^2 + (-1/3)^2 + (-4/3)^2} \times \sqrt{(-1)^2 + (-2)^2 + (1)^2 + (2)^2}}$$

$$= \frac{-10/3 - 1/3 - 8/3}{\sqrt{4/9 + 25/9 + 4/4 + 16/9 + 1/9 + 16/9} \times \sqrt{1 + 4 + 1 + 4}}$$

$$= \frac{-19/3}{\sqrt{66/9} \times \sqrt{10}}$$

$$= \frac{-6.33}{\sqrt{7.33} \times \sqrt{10}} = \frac{-6.33}{2.707 \times 3.162}$$

$$= -6.33 / 8.559 = 0.739$$

Cosine distance between (A,C)

$A = \{2/3, 5/3, 0, 5/3, -7/3, 0, 1/3, -4/3\}$

$C = \{-1, 0, -2, 0, 0, 1, 2, 0\}$

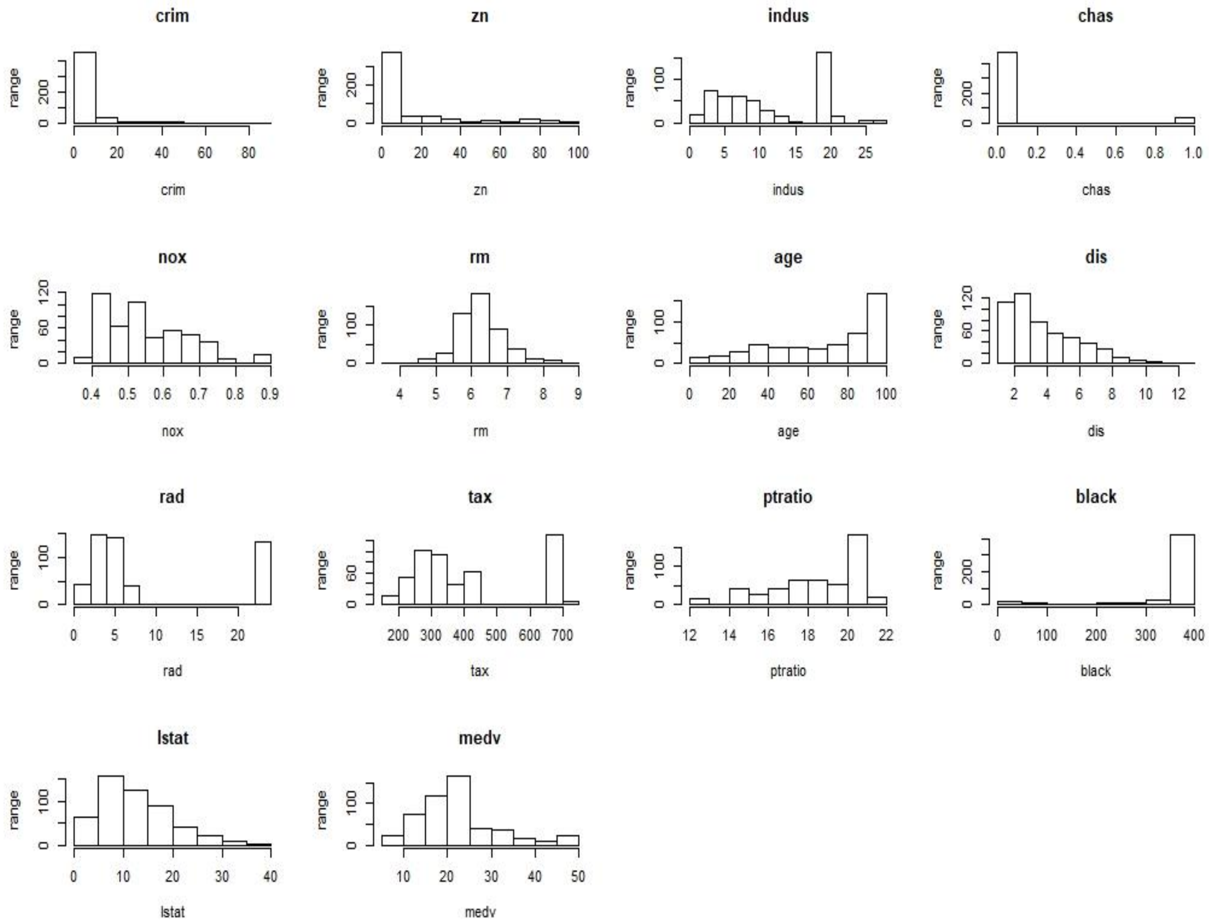
$$= \frac{(2/3 \times -1) + (1/3 \times 2)}{\sqrt{(2/3)^2 + (5/3)^2 + (5/3)^2 + (-7/3)^2 + (1/3)^2 + (-4/3)^2} \times \sqrt{(-1)^2 + (-2)^2 + (1)^2 + (2)^2}}$$

$$= \frac{-2/3 + 2/3}{\sqrt{120/9} \times \sqrt{10}} = 0$$

QUESTION 2

a) Visualize the data using histograms of the different variables in the data set. Transform the data into a binary incidence matrix, and justify the choices you make in grouping categories.

Data visualization of different variables using histograms is done using hist() function,



To transform into binary incidence matrix, first, we have to change the variables to ordered factors and split each factor level accordingly.

Splitting factors levels accordingly,

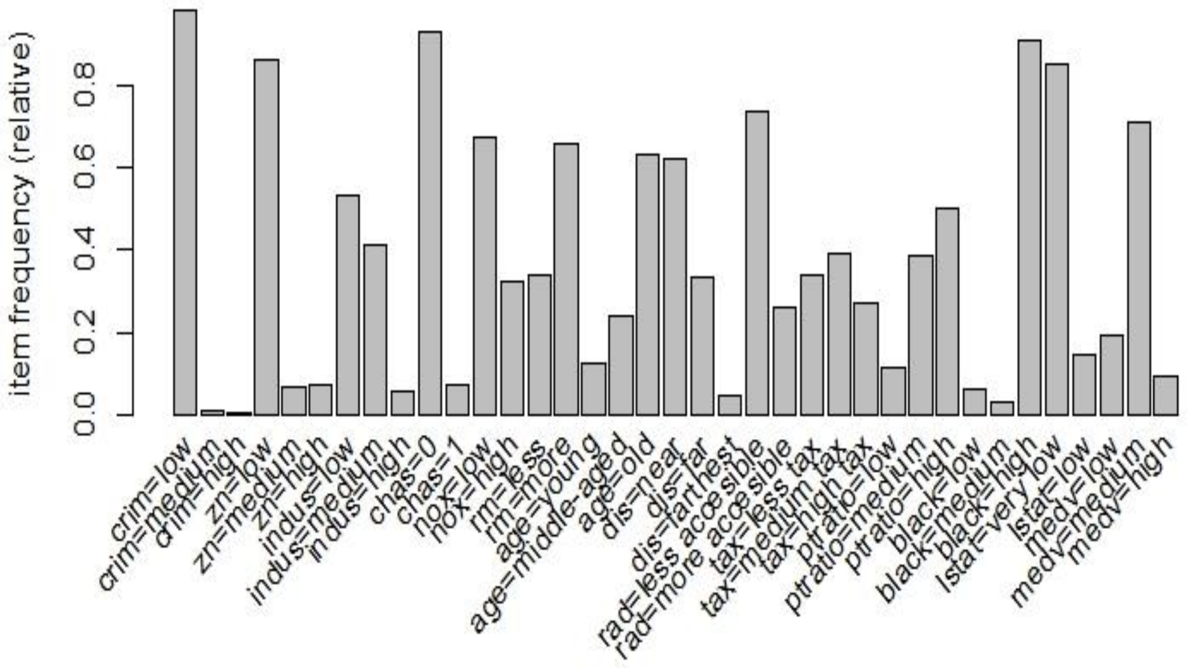
- The **crim** column in Boston data has been separated into three categories, low crime area (0-30), medium crime (30-60) and high crime areas (60-90). Here most of the data falls into the first category.

- The **indus** column has been separated into three categories, low proportion (0-10), medium proportion (10-20), high proportion areas (20-30).here most of the data falls into the second category.
- The **chas** column has been changed to ordered categorical variable using as.factor and as.ordered functions.here most of the data falls into the first category.
- The **zn** column has been transformed into low proportion (0-30), medium proportion (30-60), high proportion areas (60-100).
- The **nox** column has been changed into low (0.3-0.6) and high(0.6-0.9) nitrogen oxide concentration areas.
- The **rm** column has been changed into less (3-6) average number of rooms and more (6-9) average number of rooms. Here most of the data falls into the second category
- The **age** column has been changed into young (0-30), middle-aged (30-60) and old (60-100) aged prople.
- The **dis** column has been changed into near (1-4), far (4-8) and farthest (8-13) from the employment centers.here most of the data falls into the first category
- The **rad** column has been changed into less accessible (0-8) and more accessible (8-25) to radial highways.here most of the data falls into the first category.
- The **tax** column ahs been changed into less tax (180-300), medium tax (300-500) and high tax (500-720) categories.
- The **ptratio** column has been changed into low (11-15), medium (15-19) and high (19-23) pupil teacher ratio.
- The **black** column has been changed into low (0-100), medium (100-250) and high (250-400) population areas. Here most of the data falls into the third category.
- The **lstat** column has been changed into very low (1-20) and low (20-40) percentage categories.
- The **medv** column has been changed into low (0-15), medium (15-35) and high (35-50) value categories.

After changing the variables into ordered factors, we then convert the dataframe into binary incidence matrix using **as.** function with class as **“transaction”** class.

b) Visualize the data using the itemFrequencyPlot in the “arules” package. Apply the apriori algorithm (Do not forget to specify parameters in your write up).

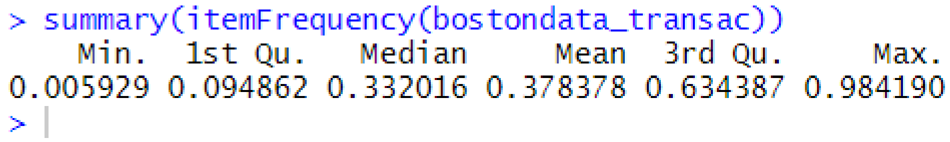
Visualizing the data using **itemFrequencyPlot()** from “**arules”** package,



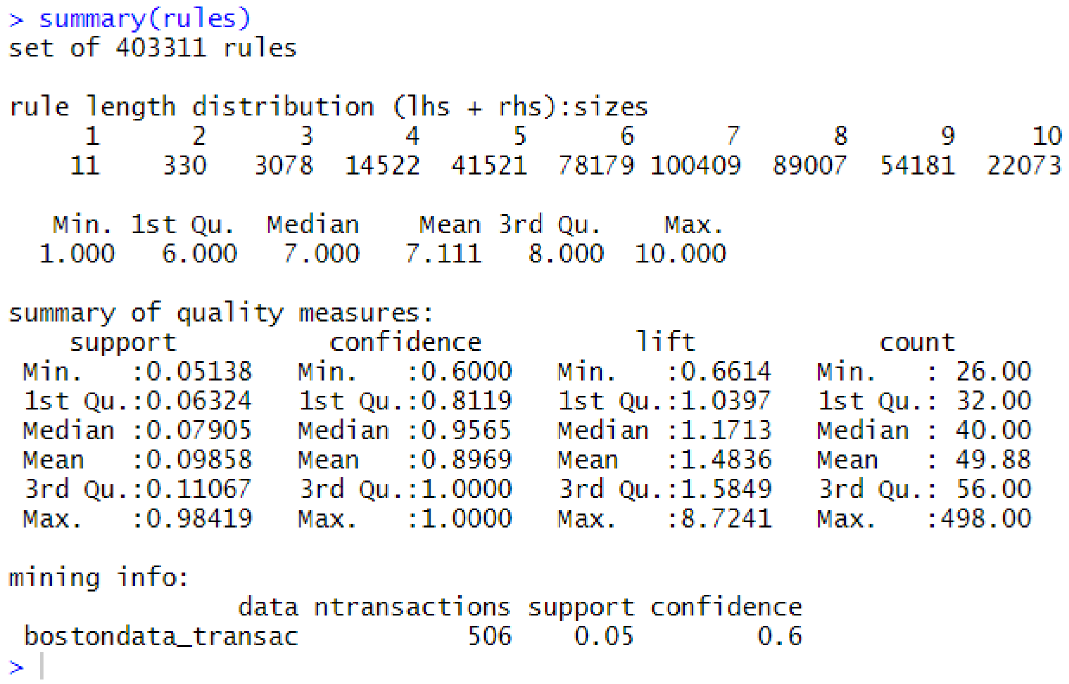
- The itemfrequency plot has been created with support of 0.001, since we don’t have much variables and can visualize all the categories of every variable.

Applying apriori algorithm,

- Apriori algorithm is then applied to the binary incidence matrix to get the rules using **apriori()** function from the **arules** package with support as 0.05 (The most frequent single item has a support of 0.9, so we choose a value below that and we have chosen 0.05 to include more categories) and we select confidence as 0.6.
- With this we get total 403311 rules.



```
> summary(itemFrequency(bostondata_transac))
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
0.005929 0.094862 0.332016 0.378378 0.634387 0.984190 
> 
```

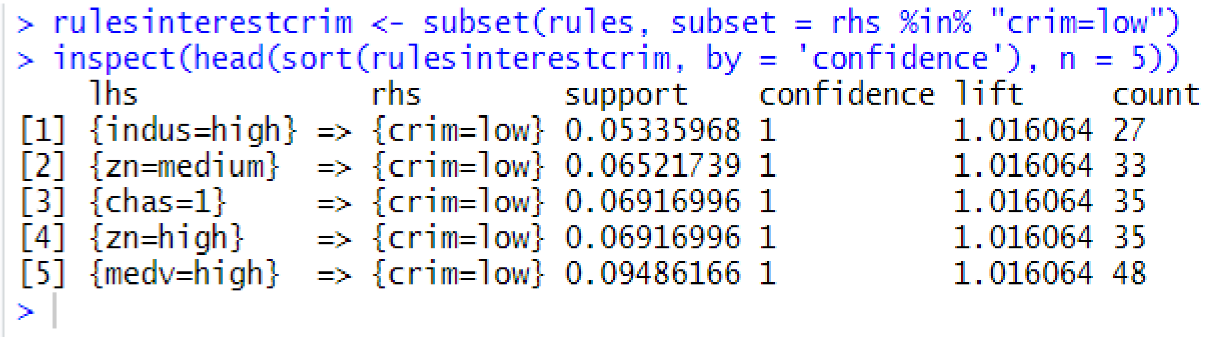


```
> summary(rules)
set of 403311 rules

rule length distribution (lhs + rhs):sizes
     1      2      3      4      5      6      7      8      9     10 
    11    330   3078  14522  41521  78179 100409  89007  54181  22073 

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.000   6.000   7.000   7.111   8.000  10.000 

summary of quality measures:
    support          confidence          lift             count       
 Min.   :0.05138   Min.   :0.6000   Min.   :0.6614   Min.   : 26.00  
 1st Qu.:0.06324   1st Qu.:0.8119   1st Qu.:1.0397   1st Qu.: 32.00  
 Median :0.07905   Median :0.9565   Median :1.1713   Median : 40.00  
 Mean   :0.09858   Mean   :0.8969   Mean   :1.4836   Mean   : 49.88  
 3rd Qu.:0.11067   3rd Qu.:1.0000   3rd Qu.:1.5849   3rd Qu.: 56.00  
 Max.   :0.98419   Max.   :1.0000   Max.   :8.7241   Max.   :498.00  

mining info:
                data ntransactions support confidence
 bostondata_transac           506    0.05        0.6
> 
```

c) A student is interested is a low crime area, but wants to be as close to the city as possible (as measured by "dis"). What can you advise on this matter through the mining of association rules?

For a student who is interested in low crime area but closer distance to city, we can subset the rules with this constrain and inspect them.

Sorting by confidence for **low crime area** and inspecting top 5 rules we can see,

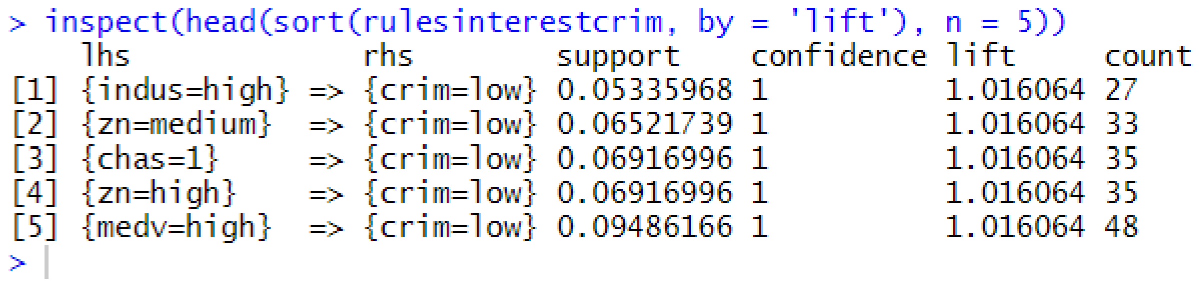


```
> rulesinterestcrim <- subset(rules, subset = rhs %in% "crim=low")
> inspect(head(sort(rulesinterestcrim, by = 'confidence'), n = 5))
    lhs              rhs        support    confidence lift     count
[1] {indus=high}  => {crim=low} 0.05335968 1          1.016064 27   
[2] {zn=medium}   => {crim=low} 0.06521739 1          1.016064 33   
[3] {chas=1}      => {crim=low} 0.06916996 1          1.016064 35   
[4] {zn=high}     => {crim=low} 0.06916996 1          1.016064 35   
[5] {medv=high}   => {crim=low} 0.09486166 1          1.016064 48   
> 
```

With this we can say that,

- To have low crime area, the proportion of non-retail business acres must be high.
- The proportion of residential land must be medium or high.
- The tract must bound a river.
- The median value of the houses should be high.

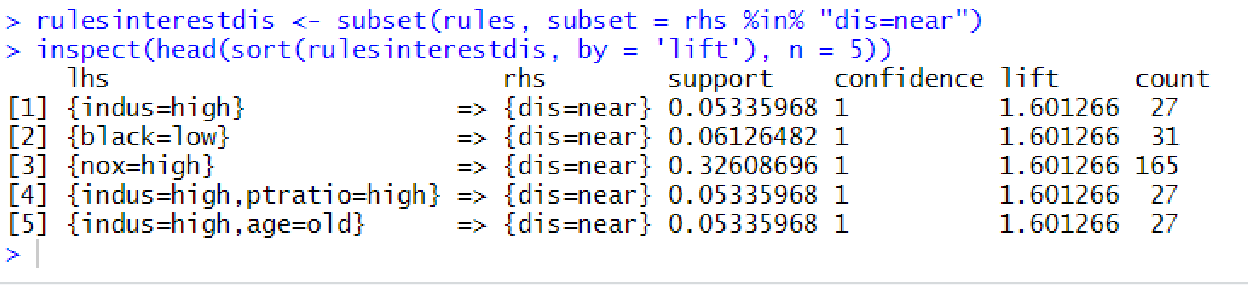
Sorting by lift for low crime area and inspecting top 5 rules we can see,



```
> inspect(head(sort(rulesinterestcrim, by = 'lift'), n = 5))
    lhs              rhs          support    confidence lift     count
[1] {indus=high}  => {crim=low} 0.05335968 1          1.016064 27
[2] {zn=medium}   => {crim=low} 0.06521739 1          1.016064 33
[3] {chas=1}      => {crim=low} 0.06916996 1          1.016064 35
[4] {zn=high}     => {crim=low} 0.06916996 1          1.016064 35
[5] {medv=high}   => {crim=low} 0.09486166 1          1.016064 48
> 
```

So, sorting both by **lift and confidence** gives **same result** thus we can conclude the same thing.

Sorting by lift for area with **closer distance** to Boston employment centers we can see,

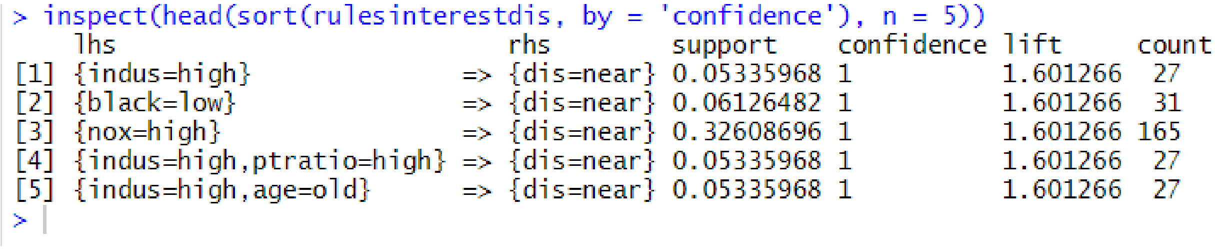


```
> rulesinterestdis <- subset(rules, subset = rhs %in% "dis=near")
> inspect(head(sort(rulesinterestdis, by = 'lift'), n = 5))
    lhs                          rhs          support    confidence lift     count
[1] {indus=high}              => {dis=near} 0.05335968 1          1.601266  27
[2] {black=low}               => {dis=near} 0.06126482 1          1.601266  31
[3] {nox=high}                => {dis=near} 0.32608696 1          1.601266 165
[4] {indus=high,ptratio=high} => {dis=near} 0.05335968 1          1.601266  27
[5] {indus=high,age=old}      => {dis=near} 0.05335968 1          1.601266  27
> 
```

From this we can say that,

- To have less distance from Boston employment centers, the proportion of non-retail business acres must be high.
- The black population must be low.
- The nitrogen oxide concentration must be high.
- The pupil-teacher ratio must be high, and the age should be old

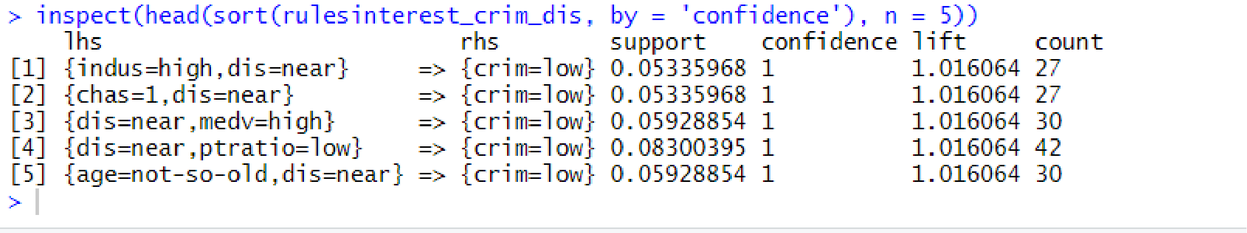
Sorting by confidence and inspecting for the top five rules,



```
> inspect(head(sort(rulesinterestdis, by = 'confidence'), n = 5))
    lhs                          rhs          support    confidence lift     count
[1] {indus=high}              => {dis=near} 0.05335968 1          1.601266  27
[2] {black=low}               => {dis=near} 0.06126482 1          1.601266  31
[3] {nox=high}                => {dis=near} 0.32608696 1          1.601266 165
[4] {indus=high,ptratio=high} => {dis=near} 0.05335968 1          1.601266  27
[5] {indus=high,age=old}      => {dis=near} 0.05335968 1          1.601266  27
> 
```

So, sorting both by **lift and confidence** gives **same result** thus we can conclude the same thing.

To have both constrain at once, we can subset the rules by having the **distance in the left-hand side** and **crime in the right-hand side**. By doing this we get,

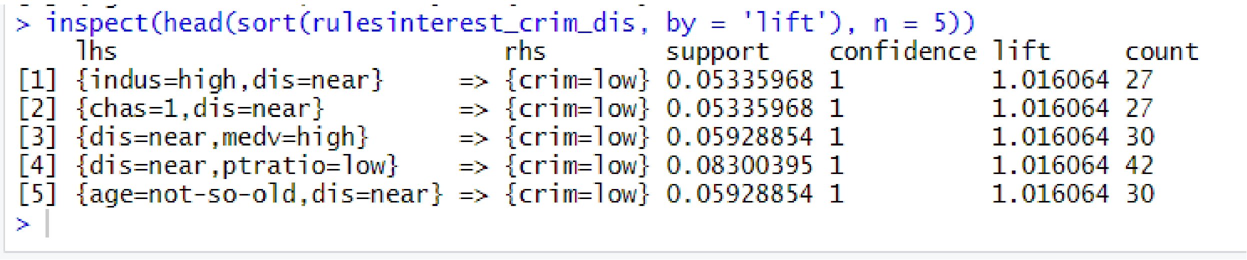


```
> inspect(head(sort(rulesinterest_crim_dis, by = 'confidence'), n = 5))
    lhs                            rhs          support    confidence lift     count
[1] {indus=high,dis=near}       => {crim=low} 0.05335968 1          1.016064 27
[2] {chas=1,dis=near}           => {crim=low} 0.05335968 1          1.016064 27
[3] {dis=near,medv=high}        => {crim=low} 0.05928854 1          1.016064 30
[4] {dis=near,ptratio=low}      => {crim=low} 0.08300395 1          1.016064 42
[5] {age=not-so-old,dis=near} => {crim=low} 0.05928854 1          1.016064 30
>
```

From this we can say that,

- To have low crime area and less distance from Boston employment centers, the proportion of non-retail business acres must be low.
- The tract must bound a river and the median value of the house must be high.
- The pupil-teacher ratio must be low, and the age should be not so old.

Sorting by **lift** and inspecting,



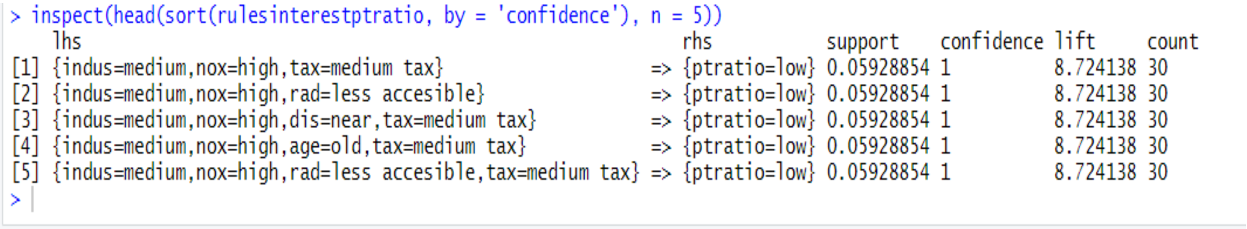
```
> inspect(head(sort(rulesinterest_crim_dis, by = 'lift'), n = 5))
    lhs                            rhs          support    confidence lift     count
[1] {indus=high,dis=near}       => {crim=low} 0.05335968 1          1.016064 27
[2] {chas=1,dis=near}           => {crim=low} 0.05335968 1          1.016064 27
[3] {dis=near,medv=high}        => {crim=low} 0.05928854 1          1.016064 30
[4] {dis=near,ptratio=low}      => {crim=low} 0.08300395 1          1.016064 42
[5] {age=not-so-old,dis=near} => {crim=low} 0.05928854 1          1.016064 30
>
```

Sorting by **lift and confidence** both gives **same result** and we can conclude the same thing.

d) A family is moving to the area, and has made schooling a priority. They want schools with low pupil-teacher ratios. What can you advise on this matter through the mining of association rules?

To get the rules for **low pupil teacher ratio**, we subset the rules with constrain as “ptratio=low” in the right-hand side.

Inspecting the rules and sorting by **confidence**,

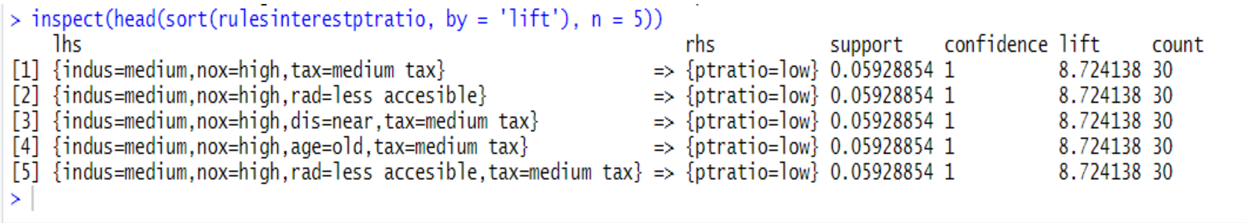


```
> inspect(head(sort(rulesinterestptratio, by = 'confidence'), n = 5))
    lhs                                                       rhs             support    confidence lift     count
[1] {indus=medium,nox=high,tax=medium tax}                 => {ptratio=low} 0.05928854 1          8.724138 30
[2] {indus=medium,nox=high,rad=less accesible}             => {ptratio=low} 0.05928854 1          8.724138 30
[3] {indus=medium,nox=high,dis=near,tax=medium tax}        => {ptratio=low} 0.05928854 1          8.724138 30
[4] {indus=medium,nox=high,age=old,tax=medium tax}         => {ptratio=low} 0.05928854 1          8.724138 30
[5] {indus=medium,nox=high,rad=less accesible,tax=medium tax} => {ptratio=low} 0.05928854 1          8.724138 30
>
```

From this we can say that,

- To have pupil-teacher ratio as low, the proportion of non-retail business acres must be low along with the low concentration of nitrogen oxide and the tax must be medium.
- The radial highway should be less accessible and the distance from the Boston employment centers must be near
- Finally, the age should be old.

Sorting the rules by **lift**,



```
> inspect(head(sort(rulesinterestptratio, by = 'lift'), n = 5))
    lhs                                                  rhs           support    confidence lift     count
[1] {indus=medium,nox=high,tax=medium tax}            => {ptratio=low} 0.05928854 1          8.724138 30
[2] {indus=medium,nox=high,rad=less accesible}        => {ptratio=low} 0.05928854 1          8.724138 30
[3] {indus=medium,nox=high,dis=near,tax=medium tax}   => {ptratio=low} 0.05928854 1          8.724138 30
[4] {indus=medium,nox=high,age=old,tax=medium tax}    => {ptratio=low} 0.05928854 1          8.724138 30
[5] {indus=medium,nox=high,rad=less accesible,tax=medium tax} => {ptratio=low} 0.05928854 1          8.724138 30
> 
```

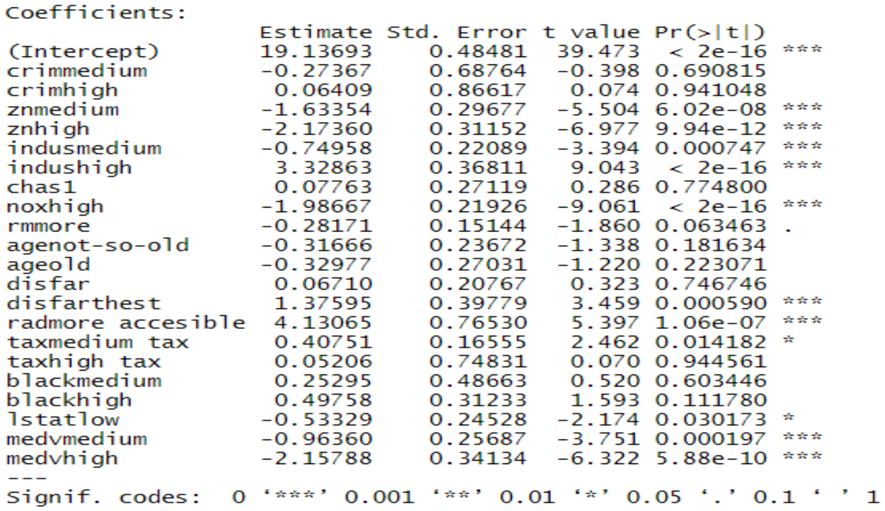
from this we can see that sorting both by **lift and confidence** gives the **same result** and we can conclude the same thing.

e) Use a regression model to solve part d. Are you results comparable? Which provides an easier interpretation? When would regression be preferred, and when would association models be preferred?

To treat the scenario as regression problem,

- we take the ptratio column from the original dataset and combine it with the factors from the modified dataset to create a new data frame.
- We then apply **linear regression** using lm() function with ptratio column as the response.

From the summary of the linear model we can see that,



```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       19.13693    0.48481  39.473  < 2e-16 ***
crimmedium        -0.27367    0.68764  -0.398 0.690815    
crimhigh           0.06409    0.86617   0.074 0.941048    
znmedium          -1.63354    0.29677  -5.504 6.02e-08 ***
znhigh            -2.17360    0.31152  -6.977 9.94e-12 ***
indusmedium       -0.74958    0.22089  -3.394 0.000747 ***
indushigh          3.32863    0.36811   9.043  < 2e-16 ***
chas1              0.07763    0.27119   0.286 0.774800    
noxhigh           -1.98667    0.21926  -9.061  < 2e-16 ***
rmmore            -0.28171    0.15144  -1.860 0.063463 .  
agenot-so-old     -0.31666    0.23672  -1.338 0.181634    
ageold            -0.32977    0.27031  -1.220 0.223071    
disfar             0.06710    0.20767   0.323 0.746746    
disfarthest        1.37595    0.39779   3.459 0.000590 ***
radmore accesible  4.13065    0.76530   5.397 1.06e-07 ***
taxmedium tax      0.40751    0.16555   2.462 0.014182 *  
taxhigh tax        0.05206    0.74831   0.070 0.944561    
blackmedium        0.25295    0.48663   0.520 0.603446    
blackhigh          0.49758    0.31233   1.593 0.111780    
lstatlow          -0.53329    0.24528  -2.174 0.030173 *  
medvmedium        -0.96360    0.25687  -3.751 0.000197 ***
medvhigh          -2.15788    0.34134  -6.322 5.88e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

So, from the regression model summary we can infer that,

- The proportion of **residential lands** zoned should be **medium or high**, the **proportion of non-retail business acres per town** should be **medium or high**, the **nitrogen oxide** concentration should be **high**, the **distance** from the employment centers should be **farthest**, the highway should be **more accesible**, the **tax** must be **medium**, the **lower status** of the population should be **low** and the **median house value** should be **medium** or **high**
- These are the important factors in deciding the **ptratio response**.

Comparable?

With the regression model we can only see the factors that are important in predicting the ptratio response, not the specific low ptratio factor level. So**, the results are not comparable**, and we cannot solve the original question of finding the rules that leads to low pupil teacher ratio with regression model.

From this we can clearly say that **association rules** lead to **easier interpretation**.

When would regression be preferred?

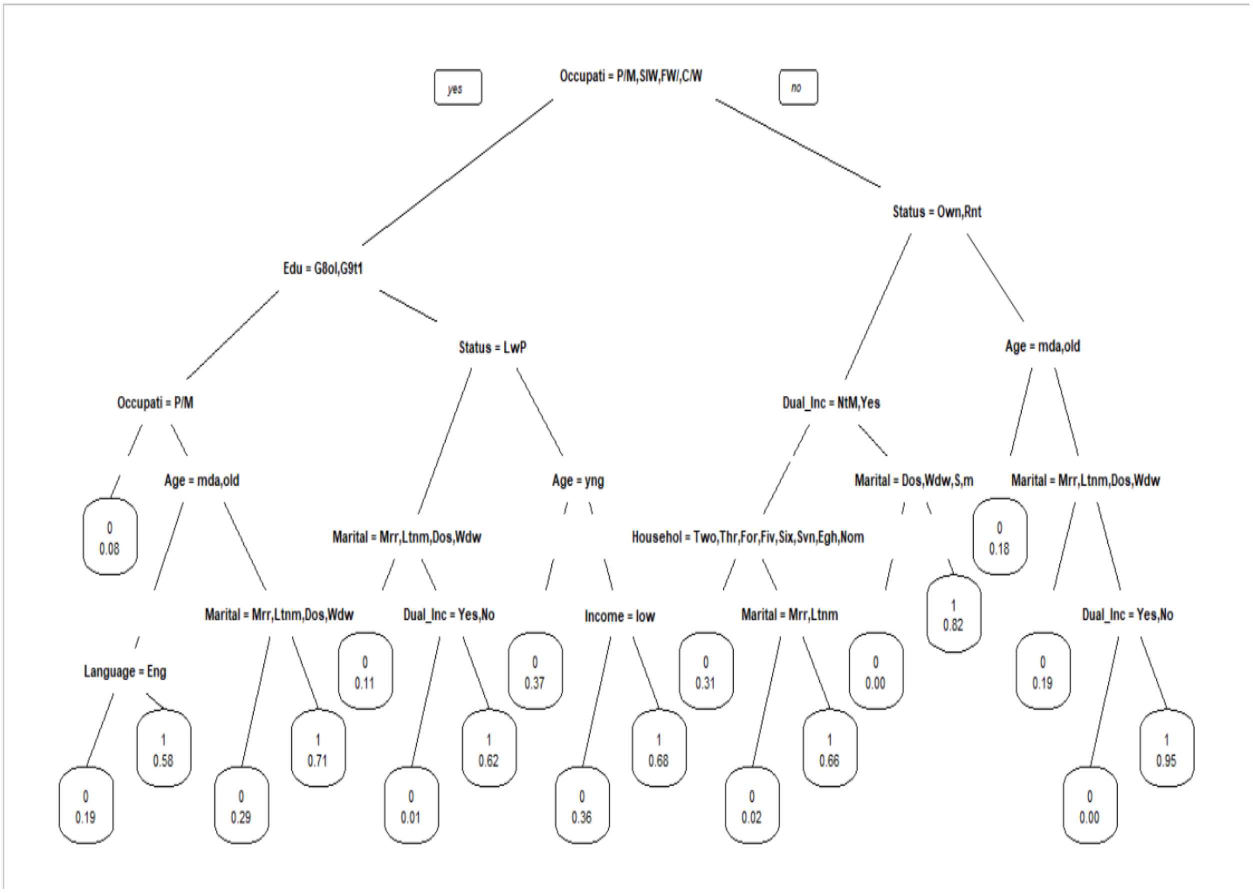
- Regression should be preferred when we have continuous response and to describe the relationship between the set of independent variables and the dependent variable. For example, predicting snowfall in inches given set of independent variables such as temperature, pressure, humidity etc.
- Association rules is preferred when we want to find or detect relationships between two products or specific values of categories in a factor. For example, in customer transaction data, we can use association rules to find what products are bought along with other products.

QUESTION 3

(3) (10 points) (Modified Exercise 14.4 in ESL) Cluster the demographic data (>data(marketing in ESL package)) of Table 14.1 using a classification tree. Specifically, generate a reference sample the same size as the training set, by randomly permuting the values within each feature. Build a classification tree to the training sample (class 1) and the reference sample (class 0) and describe the terminal nodes having highest estimated class 1 probability

Steps involved in building a classification tree,

- The data is first clustered into appropriate categories and a new column of class 1 is assigned.
- New reference sample is created by randomly sampling from the training sample and again a new column is generated but given class 0
- The training sample is named as marketing_data_1 and the reference sample is named as data1
- Both the data frames are combined using rbind() function and the new data frame is named as newdata_for_tree.
- A tree model is created using rpart() function from "rpart" library with max depth of 5.
- The created tree is pruned using optimal value of cp which is found by cross validation.
- The pruned tree is then plotted using prp() function from "rpart.plot" library.



The terminal node with class 1 has the highest probability of 0.95. it is given by,

- Dual income should be not married status and the marital status must be single and never married.
- Age should be young, and the status must be living with parents.
- Occupation must be one of the following,
    - Student/ high school/ college
    - Military
    - Retired
    - Unemployed
    - Homemaker

The terminal node with class 1 which has the second highest probability is 0.82 is given by,

- Marital status must be married or not married but living together
- Dual income must be no, and the status should be own or rent
- Occupation must be one of the following,
    - Student/ high school/ college
    - Military
    - Retired
    - Unemployed
    - Homemaker