Data Mining (IDS 472)
HOMEWORK 1

Oleh Antonyuk
Ravi Patel
Warren Svoboda

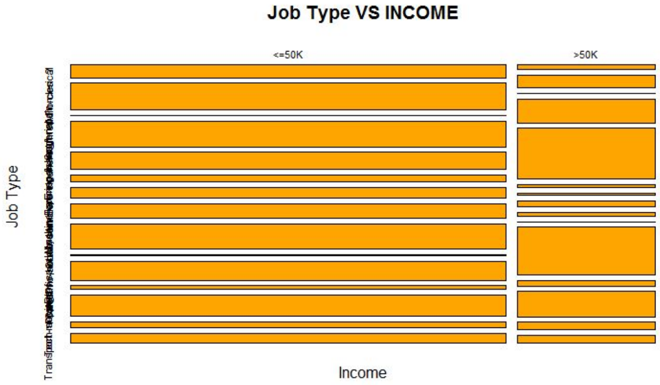
Problem 1:

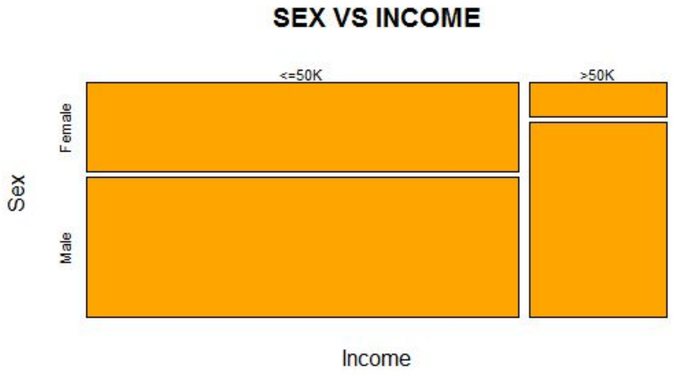
A) There are 32,561 records. The function in R that you can use to get this information is dim(salary_class)

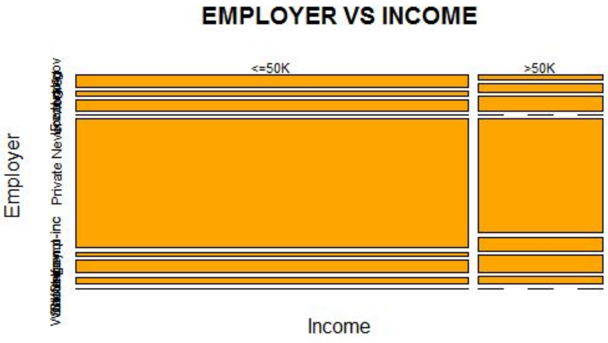
B) Measurements of all the variables:
AGE: Nominal; discrete
Employer: Categorical - Nominal
Degree: Ordinal-discrete
MStatus: Nominal
Jobtype: nominal
Sex: binary
C-Gain: Interval - Integer
C-Loss: Interval - Integer
Hours: Interval - Integer
Country: Nominal
Income: Binary

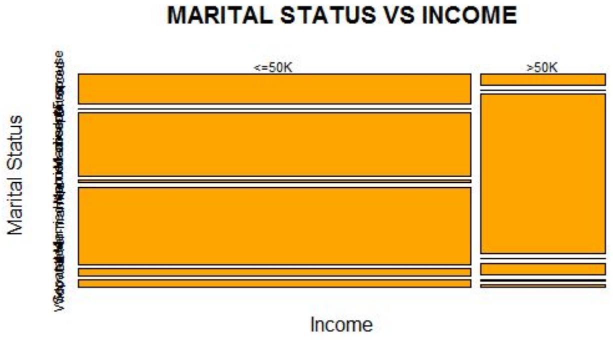
You can find these measurements by using the str(salary_class) function in R.

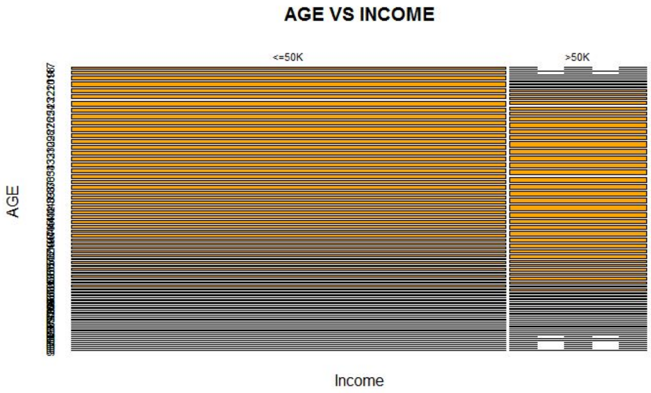
C) Age Group:
   a) Mean: 38.58
   b) Median: 37
   c) Variance: 186.06
   d) Standard Deviation: 13.64


Job Type VS INCOME
<=50K
>50K
Job Type
Income


SEX VS INCOME
<=50K
>50K
Sex
Female
Male
Income


EMPLOYER VS INCOME
<=50K
>50K
Employer
Private
Income

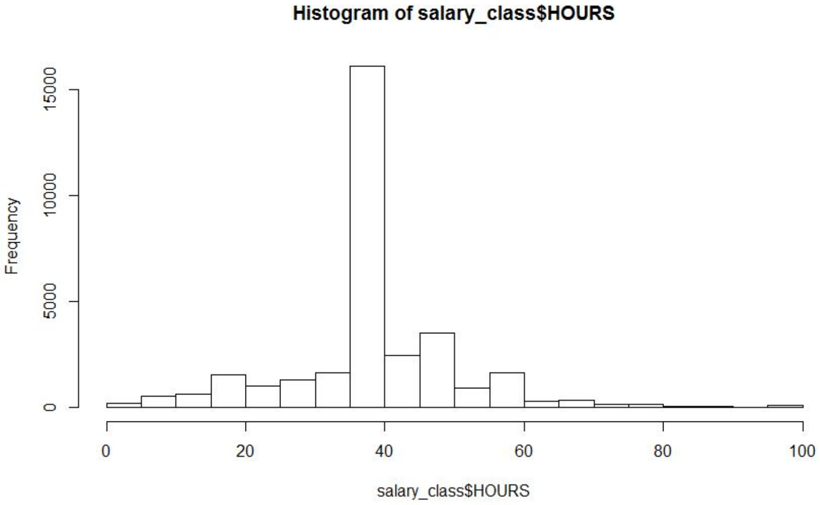

MARITAL STATUS VS INCOME
<=50K
>50K
Marital Status
Income


AGE VS INCOME
<=50K
>50K
AGE
Income

D)

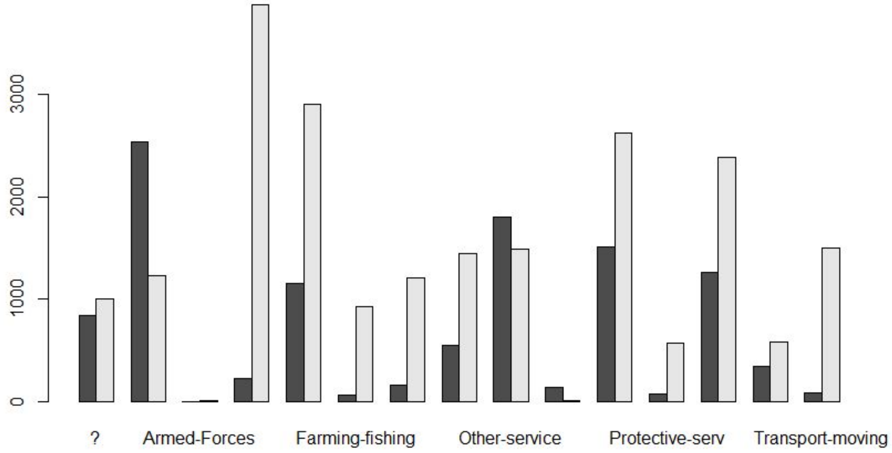
We can tell a few things about these graphs most of all that we can say is that the majority of people make less than 50k, males make more income than females, and the most of the high paying jobs are in the private sector along with being some sort of upper management. There are a majority of married people but if you are single you have a higher opportunity to have more income.The age that you will get the most income is around the 30’s to 40’s.

E) The hours have an asymmetrical, non-normal, distribution because there are majority who are working around the 40 hour margin, meanwhile there are others who are working quite less than that, or quite a lot more than the 40 hours

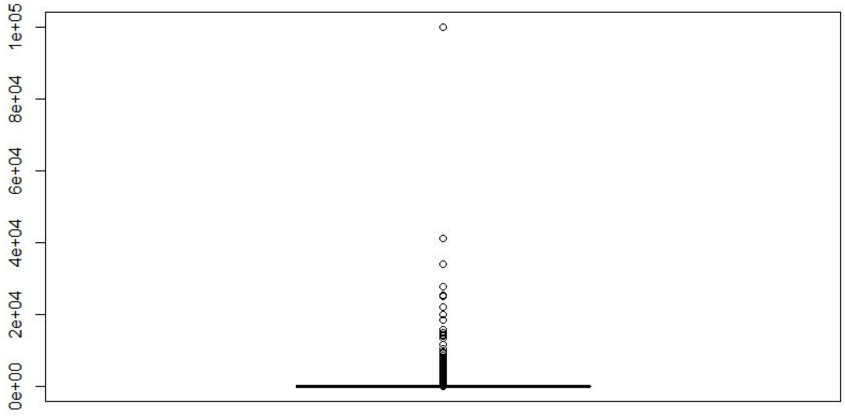


.

F)



G) Yes according to the boxplot below there are a few outliers within this dataset. You can see that majority of the data points are all the way at the bottom, versus a few that are very high up and alone.

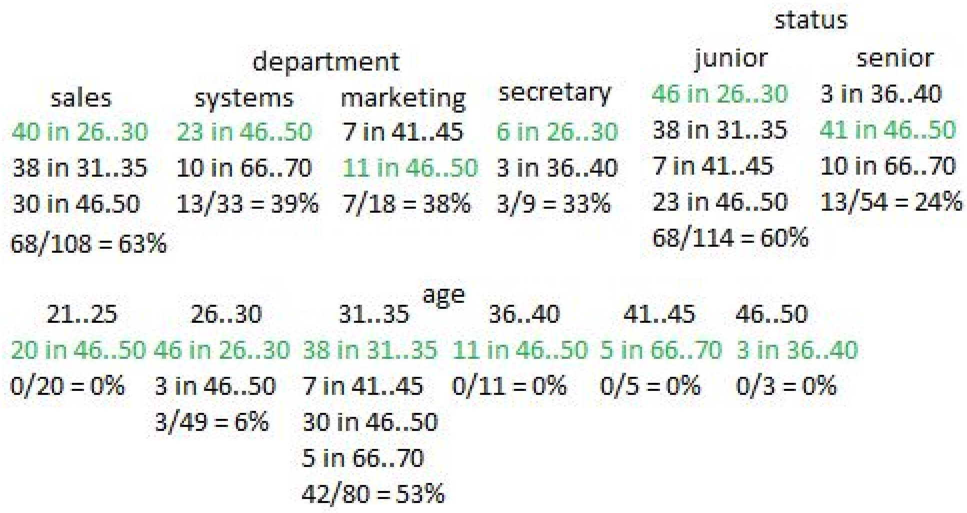


Problem 2

a. Support = 23/168 Confidence = 23/33

There are 168 total employees, 33 of which belong to the systems department, and 23 of which are in the 46K..50K salary range.

b.



department			
sales	systems	marketing	secretary
40 in 26..30	23 in 46..50	7 in 41..45	6 in 26..30
38 in 31..35	10 in 66..70	11 in 46..50	3 in 36..40
30 in 46.50	13/33 = 39%	7/18 = 38%	3/9 = 33%
68/108 = 63%			

status	
junior	senior
46 in 26..30	3 in 36..40
38 in 31..35	41 in 46..50
7 in 41..45	10 in 66..70
23 in 46..50	13/54 = 24%
68/114 = 60%	

age					
21..25	26..30	31..35	36..40	41..45	46..50
20 in 46..50	46 in 26..30	38 in 31..35	11 in 46..50	5 in 66..70	3 in 36..40
0/20 = 0%	3 in 46..50	7 in 41..45	0/11 = 0%	0/5 = 0%	0/3 = 0%
	3/49 = 6%	30 in 46..50			
		5 in 66..70			
		42/80 = 53%			

Note that in the above calculations, the green text indicates the most frequent (dominant) class. The calculation at the bottom of each column is the error for the given subset (complement of the dominant class/total observations in subset).

Dept: ((68/108)*(108/168))+((10/33)*(33/168))+((7/18)*(18/168))+((3/9)*(9/168)) =52.38%
Status: ((13/54)*(54/168))+((68/114)*(114/168)) = 48.21%
Age: (0)+((3/49)*(49/168))+((42/80)*(80/168))+(0)+(0)+(0) = 26.79%

The last three calculations are calculating total error of the three sets (misclassification rates). The zeroes for age indicate pure sets, as shown in the trees. These percent answers make it obvious the using the input attribute age will yield the lowest misclassification rate.