3, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70
(a) Use smoothing by bin means to smooth the above data, using a bin depth of 3. Illustrate your steps. Comment
on the effect of this technique for the given data.
(b) Use IQR measure to determine if there are any outliers in this data.
(c) Use min-max normalization to transform the value 35 for age onto the range [0.0, 1.0].
(d) Use z-score normalization to transform the value 35 for age? (you need to compute mean and standard deviation
first)
(e) Use normalization by decimal scaling to transform the value 35 for age.
2. Write a function in your preferred language which can take a data vector and do min-max normalization by transforming

a)
Bins
|3,15,16| Mean=(3+15+16)/3 = 11.3
16, 19, 20|, Mean=(16+19+20)/3=18.3
|20, 21, 22|, Mean=(20+21+22)/3=21
|22, 25, 25|, Mean=(22+25+25)/3=24
|25, 25, 30|, Mean=(25+25+30)/3=26.7
|33, 33, 35|, Mean=(33+33+35)/3=33.7
|35, 35, 35|, Mean=(35+35+35)/3=35
|36, 40, 45|, Mean=(36+40+45)/3=40.3
|46, 52, 70|, Mean=(46+52+70)/3=26.7
Result
|11.3,11.3,11.3|
|18.3,18.3,18.3|
|21,21,21|
|24,24,24|
|26.7,26.7,26.7|
|33.7,33.7,33.7|
|35,35,35|
|40.3,40.3,40.3|
|56,56,56|

This method does not have a great effect on this dataset because the data does not have many outliers or noise.

b)
According to R, the 1st quartile is 20.5 and the third quartile is 35, making the IQR=35-20.5=15.5

15(1.5)-20.5=2
15(1.5)+35=57.5

The only data point not in this range is 70, meaning it is an outlier

c)
Min=3
Max=70
Actual=35

(35-3)/(70-3)=0.4776

d)
Mean=29.59
Sd=13.5739

(35-29.59)/13.5739=0.3986

e)
35/100=0.35

Department	Senior	Junior	Psenior	Pjunior
Sales	30	80	30/110	80/110
Systems	8	23	8/31	23/31
Marketing	10	4	10/14	4/14

e)
35/100=0.35

Question 3

department	age	salary	status	count
sales	31_35	46K_50K	senior	30
sales	26_30	26K_30K	junior	40
sales	31_35	31K_35K	junior	40
systems	21_25	46K_50K	junior	20
systems	31_35	66K_70K	senior	5
systems	26_30	46K_50K	junior	3
systems	41_45	66K_70K	senior	3
marketing	36_40	46K_50K	senior	10
marketing	31_35	41K_45K	junior	4
secretary	46_50	36K_40K	senior	4
secretary	26_30	26K_30K	junior	6

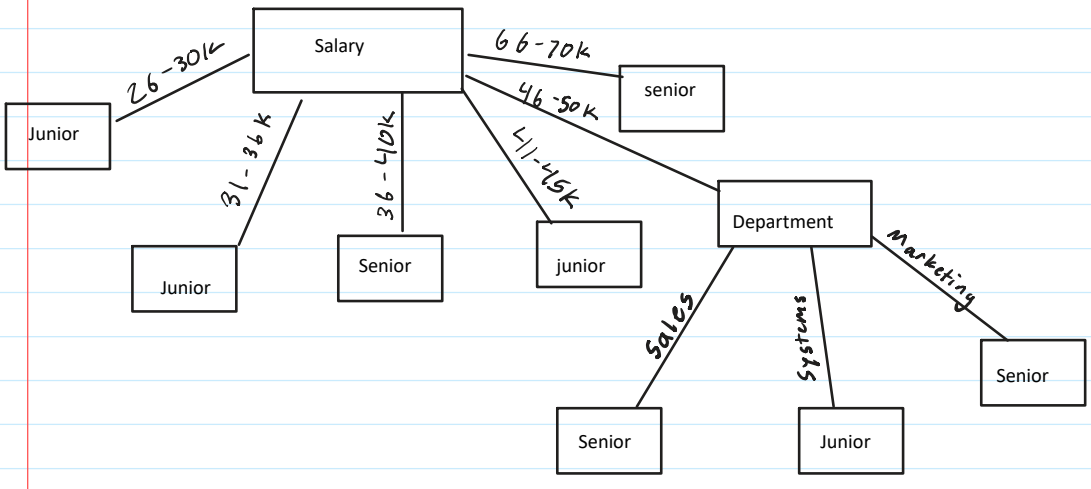
Info of Dataset = .8990308
Info(Department)= 0.1537002+0.03994238+0.05231034+0.0235382=0.26949112
Info(age) = 0+0+0.210131+0+0+0=0.210131
Info(salary)= 0+0+0+0+0.1146659+0=0.1146659

Level 1
InfoGain(Department) = .8990308-.26949112=0.62953968
InfoGain(age) = .8990308-0.210131=0.6888998
InfoGain(salary)= .8990308-0.1146659=0.7843649

Level 2
Entropy = 0.9468188

Info(age) = 0+0+0+0=0
Info(department) = 0+0+0+0=0

InfoGain for both = 0.9468188
Question 4



Sales	30	80	30/110	80/110
Systems	8	23	8/31	23/31
Marketing	10	4	10/14	4/14
Secretary	4	6	4/10	6/10
TOTAL	52	113	52/165	113/165

Age	Senior	Junior	Psenior	Pjunior
21_25	0	20	0/20	20/20
26_30	0	49	0/49	49/49
31_35	35	44	35/79	44/79
36_40	10	0	10/10	0
41_45	3	0	3/3	0/3
46_50	4	0	4/4	0/4
TOTAL	52	113	52/165	113/165

Salary	Senior	Junior	Psenior	Pjunior
26K_30K	0	46	0/46	46/46
31K_35K	0	40	0/40	40/40
36_40K	4	0	4/4	0/4
41K_45K	0	4	0/4	4/4
46K_50K	40	23	40/63	23/63
66K_70K	8	0	8/8	0/8
TOTAL	52	113	52/165	113/165

Department	Age	Status	Count
Sales	31-35	Senior	30
Systems	21-25	Junior	20
Systems	25-30	junior	3
Marketing	36-40	Senior	10

Department	Senior	Junior	Psenior	Pjunior
Sales	30	0	30/30	0/30
Systems	0	23	0/23	23/23
Marketing	10	0	10/10	0/10
TOTAL	40	23	40/63	23/63

Age	Senior	Junior	Psenior	Pjunior
21-25	0	20	0/20	20/20
25-30	0	3	0/3	3/3
31-35	30	0	30/30	0/30
36-40	10	0	10/10	0/10