CSC-587-W1-HOMEWORK 3

Q1:
(a)

Bin 1: 13, 15, 16 (mean = 14.66)
Bin 2: 16, 19, 20 (mean = 18.33)
Bin 3: 20, 21, 22 (mean = 21)
Bin 4: 22, 25, 25 (mean = 24)
Bin 5: 25, 25, 30 (mean = 26.66)
Bin 6: 33, 33, 35 (mean = 33.66)
Bin 7: 35, 35, 35 (mean = 35)
Bin 8: 36, 40, 45 (mean = 40.33)
Bin 9: 46, 52, 70 (mean = 56)

Smoothing by bin means:

Bin 1: 15, 15, 15
Bin 2: 18, 18, 18
Bin 3: 21, 21, 21
Bin 4: 24, 24, 24
Bin 5: 27, 27, 27
Bin 6: 34, 34, 34
Bin 7: 35, 35, 35
Bin 8: 40, 40, 40
Bin 9: 56, 56, 56

Comment : Smoothing by bin means reduces small fluctuations and makes the data appear smoother and easier to interpret. The data becomes simpler and less noisy which helps identify general trends but at the cost of losing detailed information and sensitivity to extreme values.

(b)

Q1 = 20
Q2 = 25
Q3 = 35

IQR = 35 - 20=15

Q1-1.5*IQR = 20-(1.5*15)= -2.5

Q3+1.5*IQR= 35+(1.5*15)= 57.5

70>56.75 => 70 is a outlier

(c)

min_max_norm= (35-13)/(70-13) = 0.386

(d)

μ= 809/27 = 29.96296296

σ= $\sqrt{167.4985755}$ = 12.94212407
ν'= (35-2 29.96296296)/ 12.70 = 0.389197

(e)

Max value: 70
Number of digits of max value : 2
x' = 35 / 10^2 = 0.35

Q3:
For Department:

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
Secretary	4	6	4/10	6/10
Total	52	113	52/165	113/165

Information_Gain= Info(D) - Info.department(D)

Info(D) = -(52/165)log2*(52/165)-(113/165)log2*(113/165) = **0.8990307712**
Info.department(D)=
(110/165)*Info.Sales+(31/165)*Info.Systems+(14/165)*Info.Marketing+(10/165)*Info.Secretary

Info.Sales= -(30/110)log2(30/110)-(80/110)log2(80/110)= 0.8453509366
Info.Systems= -(8/31)log2(8/31)-(23/31)log2(23/31)= 0.8238116333
Info.Marketing= -(10/14)log2(10/14)-(4/14)log2(4/14)= 0.8631205686
Info.Secretary= -(4/10)log2(4/10)-(6/10)log2(6/10)= 0.9709505945

Info.department(D)= **0.8504239852**
IG = 0.8990307712-0.8504= **0.04860678599**

For Salary:

Salary	**Senior**	**Junior**	**Psenior**	**Pjunior**
26-30K	0	46	0/46	46/46
31-35K	0	40	0/40	40/40
36-40K	4	0	4/4	0/4
41-45K	0	4	0/4	4/4
46-50K	40	23	40/63	23/63
66-70K	8	0	8/8	0/8

Info.salary(D)=(63/165)*Info.46_50K
Info.46_50K=-(40/63)log2(40/63)-(23/63)log2(23/63) = **0.9468188317**
Info.salary(D)=0.3615126448
IG = 0.8990307712-0.3615= **0.5375181264**

For Age:

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	0/10
41-45	3	0	3/3	0/3
46-50	4	0	4/4	0/4
Total	52	113	52/165	113/165

Info.Age(D)=(79/165)*Info.31-35
Info.31-35=-(35/79)log2(35/79)-(44/79)log2(44/79) = **0.9906174974**
Info.Age(D)=0.4742956503
IG = 0.8990307712-0.4742956503= **0.424735121**

Since the attribute with the highest information gain is salary, I would start the decision tree with salary as the root node.

Q4:

IF salary= 26-30K THEN status = junior
IF salary= 31-35K THEN status = junior
IF salary= 36-40K THEN status = senior
IF salary= 41-45K THEN status = junior
IF salary= 66-70K THEN status = senior
IF salary= 46-50K:
AND age= 21-25 THEN status: junior
AND age= 26-30 THEN status: junior
AND age= 31-35 THEN status: senior
AND age= 36-40 THEN status: senior