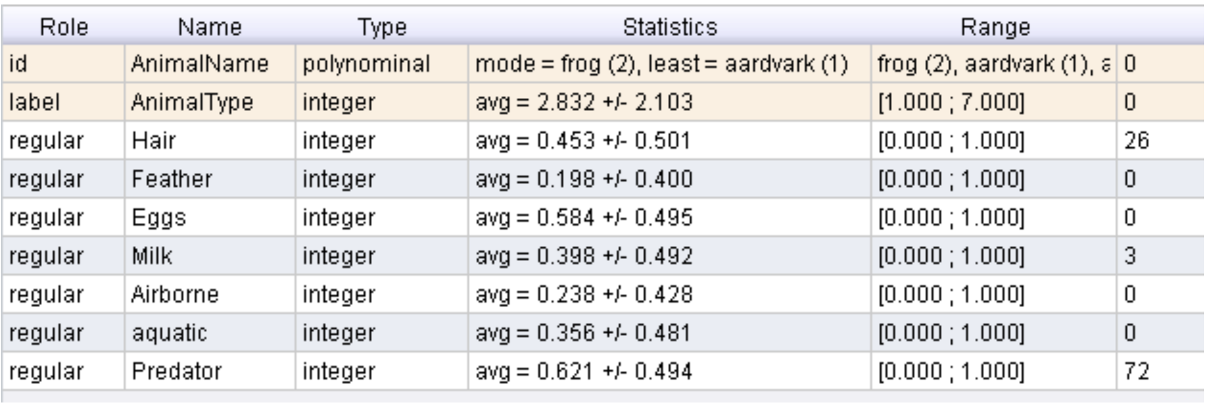
1. The following dataset has some quality issues, recommend how you would approach cleaning this dataset. The final column indicates number of missing values for that column. The dataset has 101 rows in total:



Role	Name	Type	Statistics	Range	
id	AnimalName	polynominal	mode = frog (2), least = aardvark (1)	frog (2), aardvark (1), a	0
label	AnimalType	integer	avg = 2.832 +/- 2.103	[1.000 ; 7.000]	0
regular	Hair	integer	avg = 0.453 +/- 0.501	[0.000 ; 1.000]	26
regular	Feather	integer	avg = 0.198 +/- 0.400	[0.000 ; 1.000]	0
regular	Eggs	integer	avg = 0.584 +/- 0.495	[0.000 ; 1.000]	0
regular	Milk	integer	avg = 0.398 +/- 0.492	[0.000 ; 1.000]	3
regular	Airborne	integer	avg = 0.238 +/- 0.428	[0.000 ; 1.000]	0
regular	aquatic	integer	avg = 0.356 +/- 0.481	[0.000 ; 1.000]	0
regular	Predator	integer	avg = 0.621 +/- 0.494	[0.000 ; 1.000]	72

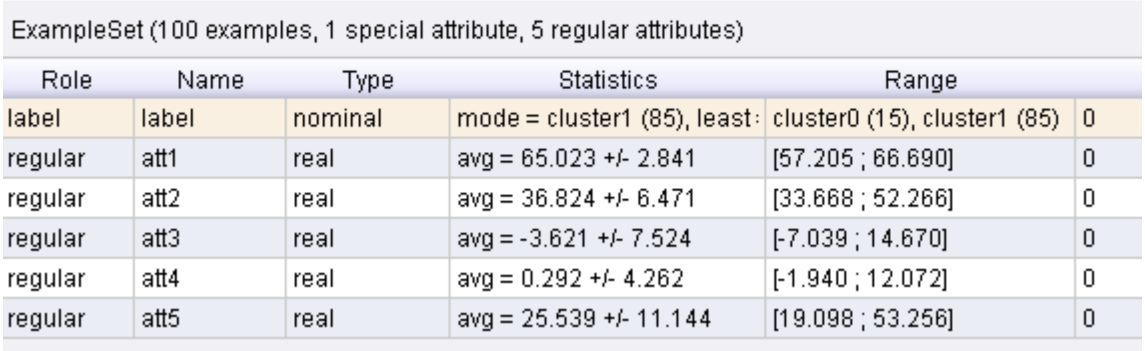
2. Is it better to fill missing values using an average value, or using imputation? Explain your answer.

3. Give an example of when it would be appropriate to use binning.

Convert the following list of numbers into three equi-width bins:

1	2	3	5	6	7	8	10	12	12

4. You have been asked to cluster the following dataset. What pre-processing needs to be done first:



ExampleSet (100 examples, 1 special attribute, 5 regular attributes)

Role	Name	Type	Statistics	Range	
label	label	nominal	mode = cluster1 (85), least:	cluster0 (15), cluster1 (85)	0
regular	att1	real	avg = 65.023 +/- 2.841	[57.205 ; 66.690]	0
regular	att2	real	avg = 36.824 +/- 6.471	[33.668 ; 52.266]	0
regular	att3	real	avg = -3.621 +/- 7.524	[-7.039 ; 14.670]	0
regular	att4	real	avg = 0.292 +/- 4.262	[-1.940 ; 12.072]	0
regular	att5	real	avg = 25.539 +/- 11.144	[19.098 ; 53.256]	0

5. When sampling data, outline a method for determining the optimal sample size.

6. What is PCA, and when would you use it?

Past Example questions:

2013: January paper. Question 2:

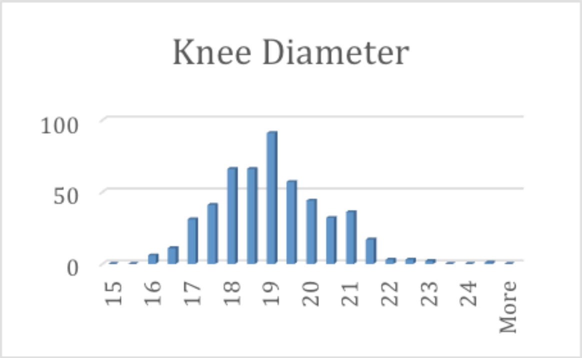
The table below shows the meta data for a dataset of skeletal measures, used to determine **gender**. The dataset has 8 attributes, and 2000 rows:

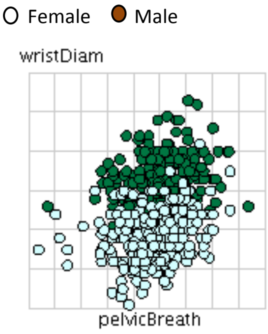
Role	Name	Type	Statistic	Range	Missing values
Label	Gender	Binominal	Mode=0(1012)	0 (1102) 1(988)	0
Regular	Age	Real	30±9.6	[18,67]	1300
Regular	Pelvic Breath	Real	27.83±2.2	[18.7,34.7]	2
Regular	Chest Depth	Real	19.2±2.5	[14.3,27.5	3
Regular	Chest Diameter	Real	27.9±22.7	[22.2,35.6]	6
Regular	Elbow Diameter	Real	13.38±1.3	[9.9,16.7]	200
Regular	Wrist Diameter	Real	10.54±0.9	[8.1,13.3]	0
Regular	Knee Diameter	Real	18.8±1.3	[15.7,24.3	0
Regular	Height	Real	171±9.3	[147.2,198.1]	0

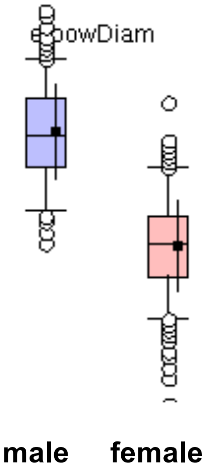
a) For each of the five attributes with missing data, recommend a suitable approach for handling their missing values. Justify each of your recommendations. **(9 marks)**

b) Recommend two other preprocessing techniques to use on the dataset above. Give a detailed explanation of each technique, and justify why they are an appropriate choice for this dataset. **(10 marks)**

c) Interpret each of the three plots below. **(11 marks)**
The histogram is for **Knee Diameter**. The scatter plot illustrates **Wrist Diameter** by **Pelvic Breath** and is colour coded by **Gender**. The box plots are for **Elbow Diameter**, split by **Gender**.







Total: 30 marks

2013: Repeat paper. Question 2:

The table below shows the meta data for a dataset of titanic passengers, and whether or not they survived. The dataset has 8 attributes, and 891 rows:

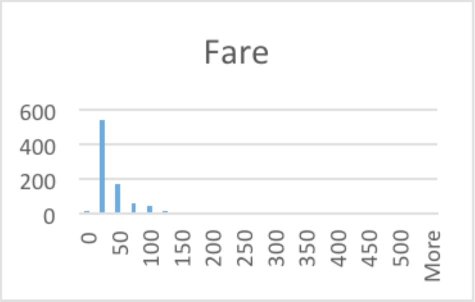
Role	Name	Data Type	Statistic	Range	Missing values
Label	Survived	Binominal	Mode=No (549)	No (549),Yes(342)	0
Regular	PClass	Integer	2.3±0.8	[1,3]	0
Regular	Name	Polynominal	Mode = Harris (1)		0
Regular	Gender	Binominal	Mode=male(577)	Male(577), Female(314)	0
Regular	Age	Real	29.7±14.5	[0,80]	177
Regular	Ticket	Polynominal	Mode=1601(4)		0
Regular	Fare	Real	32.2 ± 49.6	[0,512]	0
Regular	Cabin	Polynominal	Mode=G6(4)		687
Regular	Embarked	Polynominal	Mode=SouthHampton(644)	SouthHampton(644), Queenstown(77), Cherbourg(168)	2

a) Three of the attributes in the table above have missing values. Explain how you would handle the missing values in each case. Justify the choices you make. **(8 marks)**

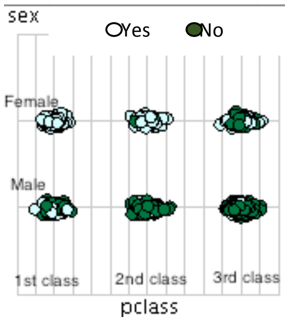
b) Discuss each of the data types in the table above with reference to how useful they are to a classification algorithm. Are there any attributes you would remove from the dataset at this point, based on the meta data? **(12 marks)**

c) Interpret each of the plots below. The first histogram is for **fare.** The scatter plot is **sex** by **passenger class (pclass)**, colour coded by the class label, **survived**, **yes** or **no**. The final histogram is for **age**.

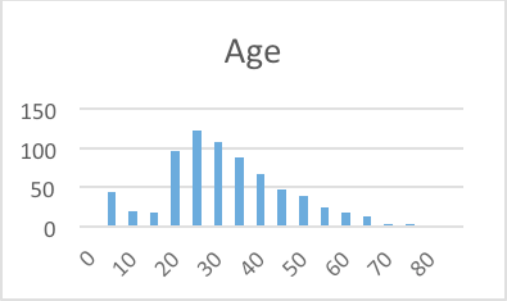
(10 marks)



(i)



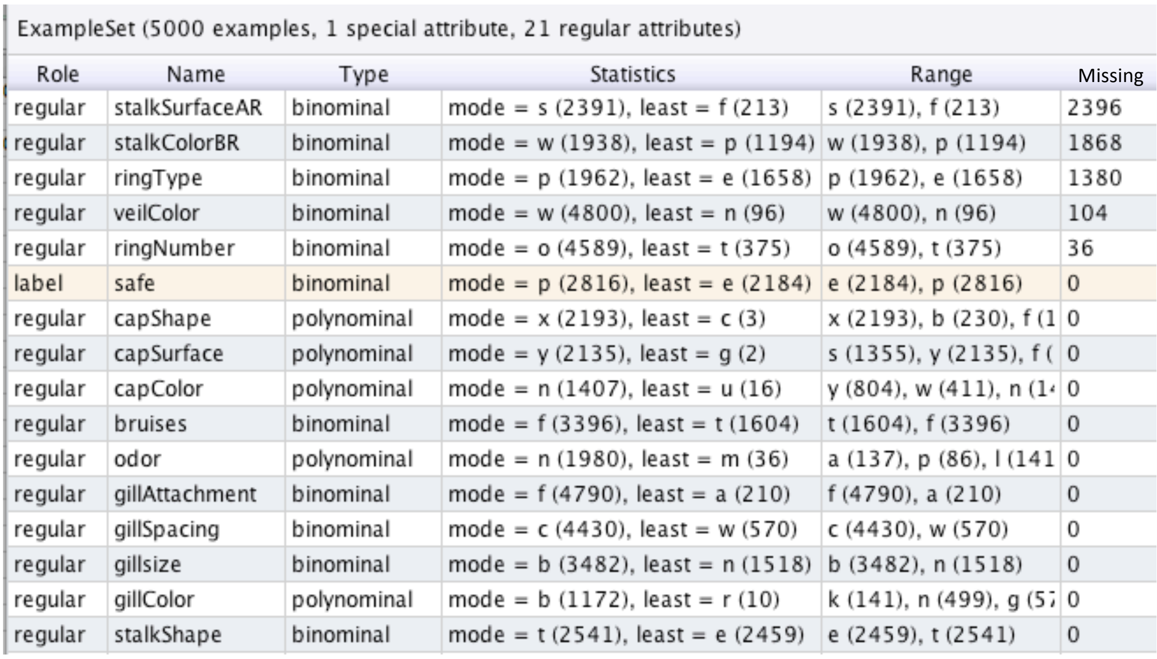
(ii)



(iii)

Total: 30 marks

2012: January paper. Question 2: Data Preparation



ExampleSet (5000 examples, 1 special attribute, 21 regular attributes)

Role	Name	Type	Statistics	Range	Missing
regular	stalkSurfaceAR	binominal	mode = s (2391), least = f (213)	s (2391), f (213)	2396
regular	stalkColorBR	binominal	mode = w (1938), least = p (1194)	w (1938), p (1194)	1868
regular	ringType	binominal	mode = p (1962), least = e (1658)	p (1962), e (1658)	1380
regular	veilColor	binominal	mode = w (4800), least = n (96)	w (4800), n (96)	104
regular	ringNumber	binominal	mode = o (4589), least = t (375)	o (4589), t (375)	36
label	safe	binominal	mode = p (2816), least = e (2184)	e (2184), p (2816)	0
regular	capShape	polynominal	mode = x (2193), least = c (3)	x (2193), b (230), f (1	0
regular	capSurface	polynominal	mode = y (2135), least = g (2)	s (1355), y (2135), f (	0
regular	capColor	polynominal	mode = n (1407), least = u (16)	y (804), w (411), n (1	0
regular	bruises	binominal	mode = f (3396), least = t (1604)	t (1604), f (3396)	0
regular	odor	polynominal	mode = n (1980), least = m (36)	a (137), p (86), l (141	0
regular	gillAttachment	binominal	mode = f (4790), least = a (210)	f (4790), a (210)	0
regular	gillSpacing	binominal	mode = c (4430), least = w (570)	c (4430), w (570)	0
regular	gillsize	binominal	mode = b (3482), least = n (1518)	b (3482), n (1518)	0
regular	gillColor	polynominal	mode = b (1172), least = r (10)	k (141), n (499), g (57	0
regular	stalkShape	binominal	mode = t (2541), least = e (2459)	e (2459), t (2541)	0

Figure 1. Meta data for the Mushroom dataset

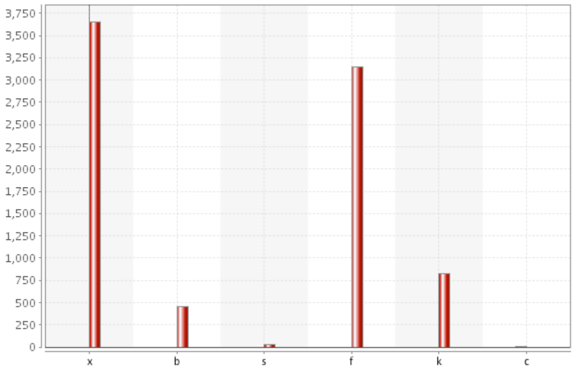
a) As is illustrated in Figure 1 above, five attributes listed in the meta data have missing values. For each attribute, explain what you would do to address the missing values. Justify all choices made. Where you recommend filling the missing values, explain two alternative techniques you could use.

12 marks

b) Explain how you would decide if sampling is appropriate for the mushroom dataset above. Also in your answer give details of two sampling techniques that could be used.

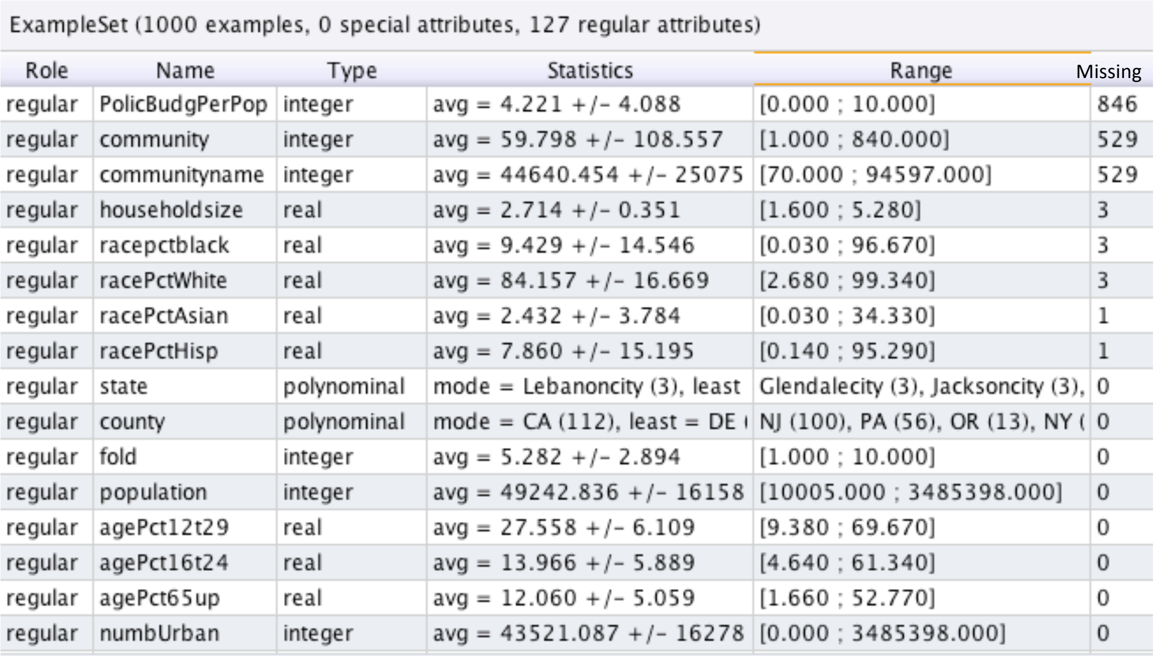
11 marks

c) Interpret the histogram below for the attribute ‘**capShape**’. Does it suggest any issues with this attribute?



7 marks

<u>2012: repeat paper</u>. Question 2: Data Preparation



ExampleSet (1000 examples, 0 special attributes, 127 regular attributes)

Role	Name	Type	Statistics	Range	Missing
regular	PolicBudgPerPop	integer	avg = 4.221 +/- 4.088	[0.000 ; 10.000]	846
regular	community	integer	avg = 59.798 +/- 108.557	[1.000 ; 840.000]	529
regular	communityname	integer	avg = 44640.454 +/- 25075	[70.000 ; 94597.000]	529
regular	householdsize	real	avg = 2.714 +/- 0.351	[1.600 ; 5.280]	3
regular	racepctblack	real	avg = 9.429 +/- 14.546	[0.030 ; 96.670]	3
regular	racePctWhite	real	avg = 84.157 +/- 16.669	[2.680 ; 99.340]	3
regular	racePctAsian	real	avg = 2.432 +/- 3.784	[0.030 ; 34.330]	1
regular	racePctHisp	real	avg = 7.860 +/- 15.195	[0.140 ; 95.290]	1
regular	state	polynominal	mode = Lebanoncity (3), least	Glendalecity (3), Jacksoncity (3),	0
regular	county	polynominal	mode = CA (112), least = DE (	NJ (100), PA (56), OR (13), NY (	0
regular	fold	integer	avg = 5.282 +/- 2.894	[1.000 ; 10.000]	0
regular	population	integer	avg = 49242.836 +/- 16158	[10005.000 ; 3485398.000]	0
regular	agePct12t29	real	avg = 27.558 +/- 6.109	[9.380 ; 69.670]	0
regular	agePct16t24	real	avg = 13.966 +/- 5.889	[4.640 ; 61.340]	0
regular	agePct65up	real	avg = 12.060 +/- 5.059	[1.660 ; 52.770]	0
regular	numbUrban	integer	avg = 43521.087 +/- 16278	[0.000 ; 3485398.000]	0

Figure 2. Meta data for the Crime & Community dataset

Figure 2 above is an extract from the meta data generated from the Crime&Community dataset, a US based dataset to investigate community related attributes and their relationship to Crime in that community. Answer the following questions based on this meta data:

Note: The dataset has 127 attributes in total.

a) Eight attributes listed in the meta data have missing values. Explain what you would do to address these missing values. Justify all choices made.

7 marks

b) The dataset above is to be used for cluster analysis. Apart from filling missing values, give details of TWO other preprocessing techniques you would recommend for the dataset. Explain the purpose of each technique, how it works, and justify why it is appropriate based on the metadata above.

14 marks

c) The histograms shown on the next page were generated as part of the Exploratory Data Analysis of the Crime&Community dataset. Disucss the two histograms with reference to:

i. Variable distribution

ii. Presence of outliers

6 marks

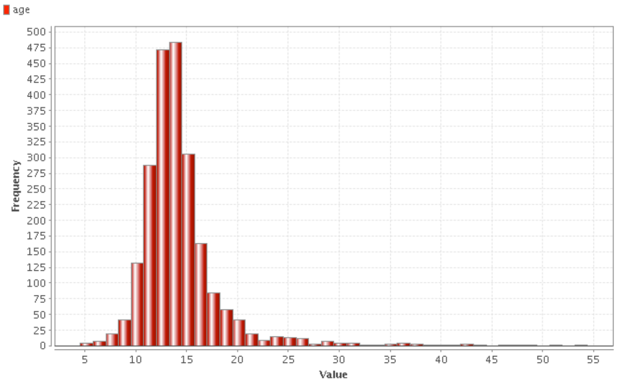
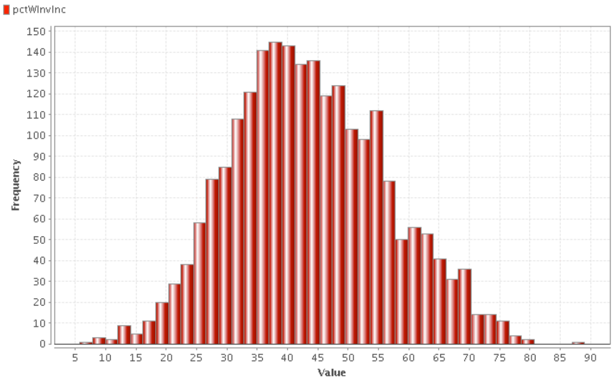


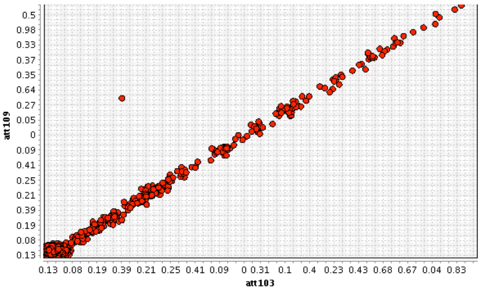
Figure 3. Histogram for Age



Histogram for Income

d) Below is a scatter matrix of two attributes from the Crime&Community dataset. What does this tell you about the relationship between the two attributes, and what is the significance of this in a data mining context?

3 marks


att109
att103