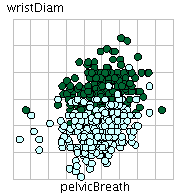
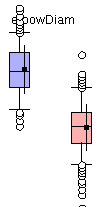
**Past exam questions. Question 2, exploration and preprocessing:**

**2013, Jan paper:**

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.  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. | (11 marks) |

Female Male

male female

*Total: 30 marks*

**2013, Repeat paper:**

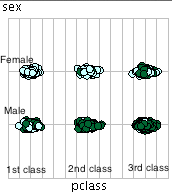
The table below shows the meta data for a dataset of titanic passengers, and whether or not they survived. The dataset has 8 attributes, and 891 rows:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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) |

1. 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)

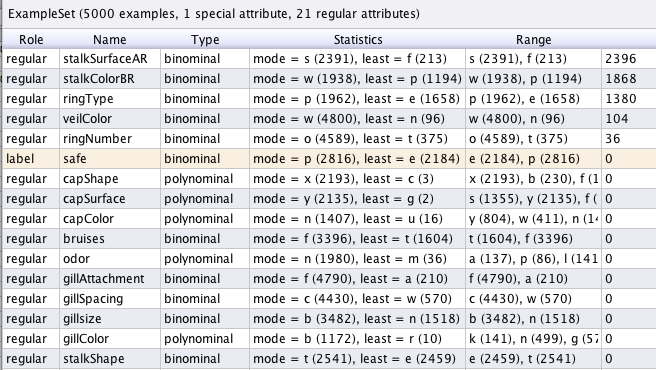
(i) (ii) 

Yes No

Female Male

(iii) *Total: 30 marks*

**2012, Jan paper:**



Missing

Figure 3. Meta data for the Mushroom dataset

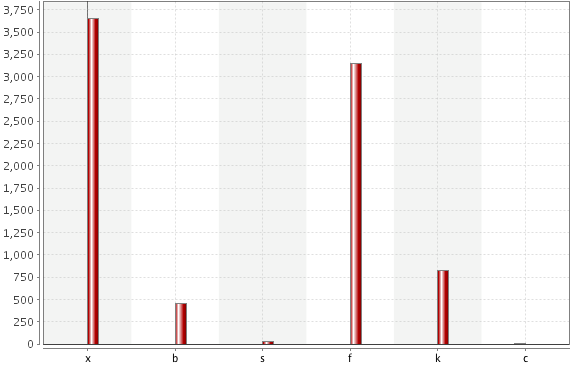
1. As is illustrated in Figure 3 above, five attributes listed in the meta data have missing values. For each attribute, explain what you would do to address the missing values. Justify all choices made. Where you recommend filling the missing values, explain two alternative techniques you could use.

*(12 marks)*

1. 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.

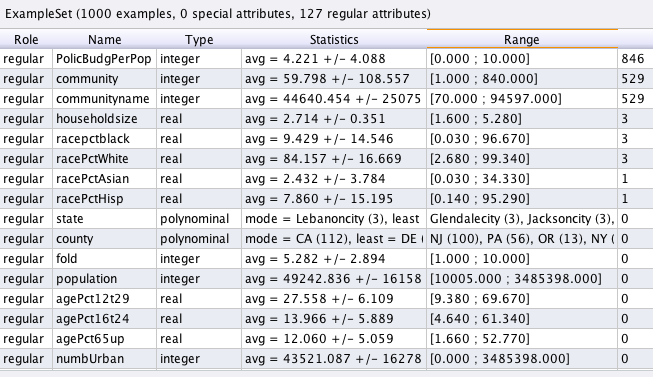
*(11 marks)*

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



*(7 marks)*

**2012, repeat paper:**



Missing

Figure 1 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:

Figure 1. Meta data for the Crime & Community dataset

Note: The dataset has 127 attributes in total.

1. 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)*

1. 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)*

1. The histograms shown on the next page were generated as part of the Exploratory Data Analysis of the Crime&Community dataset. Discuss the two histograms with reference to:
   1. Variable distribution
   2. Presence of outliers

*(6 marks)*

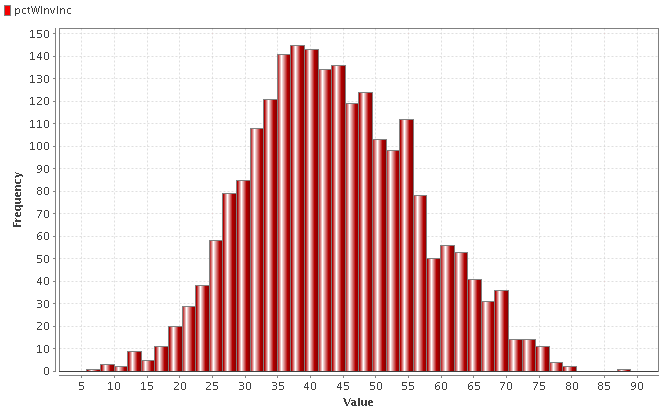
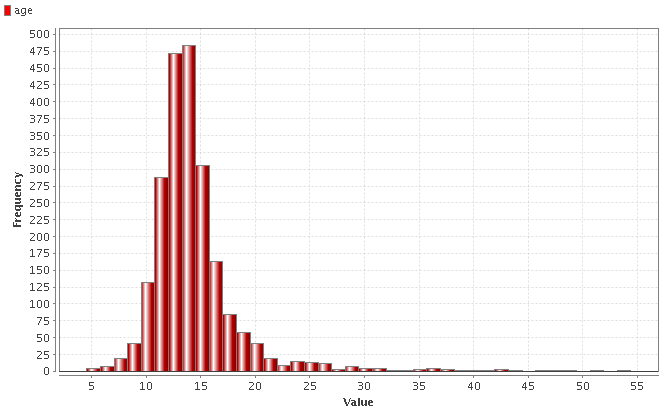
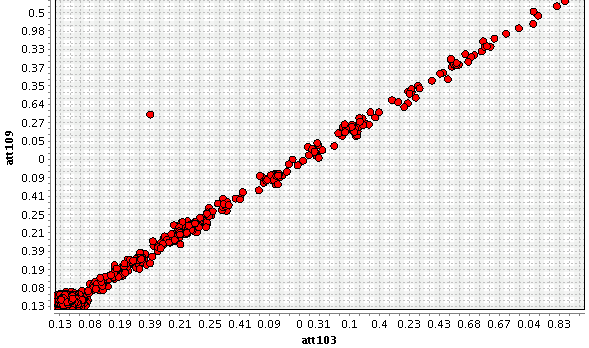


Figure 2. Histogram for Age Histogram for Income

1. 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)*

**2009, summer paper:**

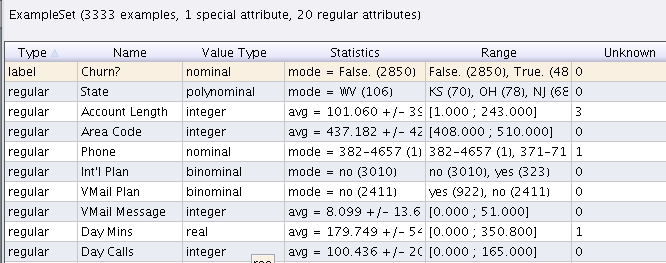


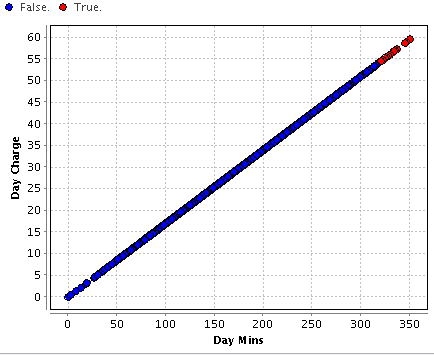
Figure 3. Dataset representing customer phone call patterns

1. Given the meta data displayed above, recommend THREE pre-processing techniques that would be appropriate to use on this dataset. In your answer, explain both why that technique would be advisable, and what would be the effect of applying the technique.

**12 marks**

1. Explain each of the data types listed in the meta data above. **5 marks**
2. Briefly explain the role of the data exploration phase of a data mining process

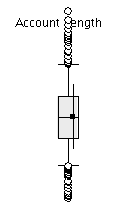
**3 marks**

1. Interpret the following three plots generated from the dataset above.

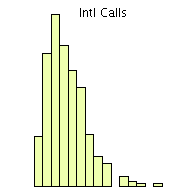
True

False

(i) Scatter plot of **day minutes** versus **day charge**, overlaid with the binary class label, **churn** (all true values of the class label are on the top right corner).



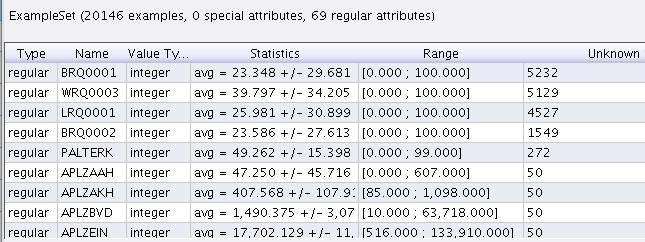
(ii) Box plot on **account length**



(iii) Histogram of **International calls made**.

**10 marks**

**2009, repeat paper:**



1. Given the meta data displayed above, recommend THREE pre-processing techniques that would be appropriate to use on this dataset. In your answer, explain both why that technique would be advisable, and what would be the effect of applying the technique.

**12 marks**

1. Explain the entries under **statistics** and **range** in the meta data above. Why is this information useful?

**12 marks**

1. Recommend two ways to visualize the data in the dataset above, explaining what information is portrayed by each.

**6 marks**