UNIVERSITY OF MORATUWA

FACULTY OF ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

MBA/PG Diploma in Information Technology
2017 Intake Semester 4 Examination

CS5122 DESCRIPTIVE AND PREDICTIVE ANALYTICS

Time allowed: 2 Hours March 2018

ADDITIONAL MATERIAL: *None*

INSTRUCTIONS TO CANDIDATES:

1. This paper consists of **5** questions in **12** pages.
2. Answer **all** questions.
3. Answer the questions on the paper itself. **DO NOT** exceed the given space.
4. The maximum attainable mark for each question is given in brackets.
5. This examination accounts for 40% of the module assessment.
6. This is a closed book examination.

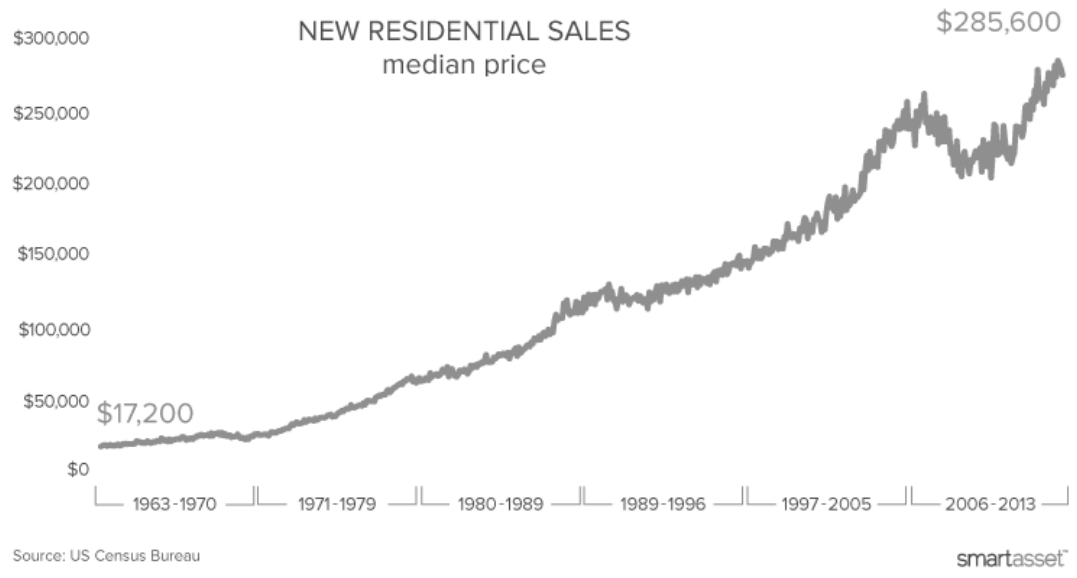
   NB: It is an offence to be in possession of unauthorised material during the examination.
7. Only calculators approved by the Faculty of Engineering are permitted.
8. Assume reasonable values for any data not given in or with the examination paper. Clearly state such assumptions made on the script.
9. In case of any doubt as to the interpretation of the wording of a question, make suitable assumptions and clearly state them on the script.
10. This paper should be answered only in English.

Question 1 (20 marks)

Answer the following questions based on the given table and figure related to the price of buying and renting houses of different cities in the USA.

Breakeven point is the point at which the total costs of renting become greater than the total costs of buying.

City	Breakeven Year	Average Monthly Mortgage Payment	Average Monthly Rent	Average Home Price
New York City	18.3	$ 2,399	$ 2,056	$ 591,895
San Jose	16.7	$ 3,162	$ 2,503	$ 779,975
Seattle	14.9	$ 1,751	$ 1,777	$ 431,939
San Francisco	14.6	$ 3,643	$ 2,915	$ 898,706
Orange County	10.8	$ 2,551	$ 2,300	$ 629,280
Los Angeles	8.8	$ 2,110	$ 2,061	$ 520,559
San Diego	8.6	$ 2,167	$ 2,134	$ 534,695
Honolulu	8.6	$ 2,970	$ 2,682	$ 732,637
Portland	6.9	$ 1,285	$ 1,428	$ 317,085
Washington, D.C.	6.5	$ 1,909	$ 2,127	$ 471,071
Boston	6.3	$ 1,563	$ 1,970	$ 385,521
Riverside	5.8	$ 1,304	$ 1,582	$ 321,665
Phoenix	5.7	$ 1,122	$ 1,410	$ 276,744
Denver	5.4	$ 1,114	$ 1,504	$ 274,835
Pittsburgh	4.3	$ 560	$ 1,069	$ 138,235
Minneapolis	4.2	$ 965	$ 1,419	$ 238,051
Chicago	4.2	$ 719	$ 1,331	$ 177,481



Source: smartasset.com

(i) Name 2 cities each where it makes sense to:

(a) Rent a house compared to buying [2]

(b) Buy a house compared to renting [2]

(ii) "An investor plans to buy and then rent a house. He/she could make more money in Washington, D.C. than in San Diego." Do you agree or disagree? [3]

(iii) "The boom in high-tech industry over the past few years has generally been concentrated in a relatively small number of cities such as San Francisco and Seattle." How is this change reflected in the real-estate price? [3]

(iv) "Portland is an outlier." Do you agree or disagree? Briefly discuss. [3]

(v) Discuss what can be learnt from the above time series while considering the Trend, Seasonal, Cyclical, and Irregular time series components. [7]

Question 2 (20 marks)

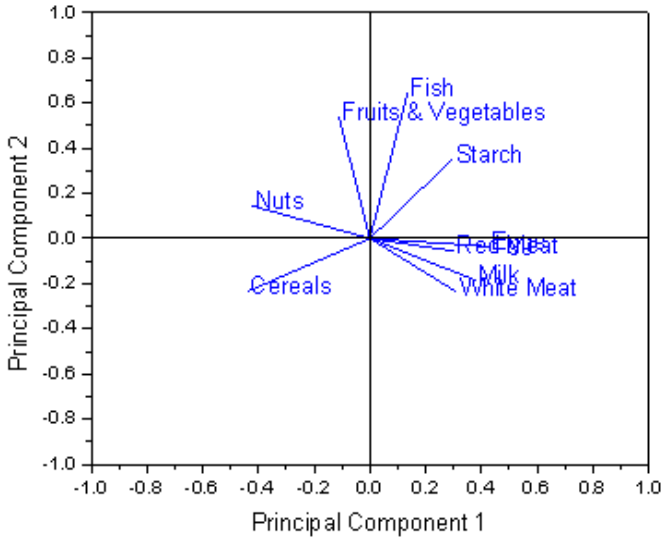
(i) Following results are based on Principle Component Analysis (PCA) of a dataset measuring protein consumption in 25 European countries for 9 food groups.

Correlation Matrix

	Red Meat	White Meat	Eggs	Milk	Fish	Cereals	Starch	Nuts	Fruits & Vegetables
Red Meat	1	0.153	0.58561	0.50293	0.06096	-0.49988	0.13543	-0.34945	-0.07422
White Meat	0.153	1	0.62041	0.28148	-0.23401	-0.4138	0.31377	-0.63496	-0.06132
Eggs	0.58561	0.62041	1	0.57553	0.06557	-0.71244	0.45223	-0.55978	-0.04552
Milk	0.50293	0.28148	0.57553	1	0.13788	-0.59274	0.22241	-0.62109	-0.40836
Fish	0.06096	-0.23401	0.06557	0.13788	1	-0.52423	0.40385	-0.14715	0.26614
Cereals	-0.49988	-0.4138	-0.71244	-0.59274	-0.52423	1	-0.53326	0.651	0.04655
Starch	0.13543	0.31377	0.45223	0.22241	0.40385	-0.53326	1	-0.47431	0.08441
Nuts	-0.34945	-0.63496	-0.55978	-0.62109	-0.14715	0.651	-0.47431	1	0.37497
Fruits & Vegetables	-0.07422	-0.06132	-0.04552	-0.40836	0.26614	0.04655	0.08441	0.37497	1

	Eigenvalue	Percentage of Variance	Cumulative
1	4.00644	44.52%	44.52%
2	1.635	18.17%	62.68%
3	1.12792	12.53%	75.22%
4	0.95466	10.61%	85.82%
5	0.46384	5.15%	90.98%
6	0.32513	3.61%	94.59%
7	0.27161	3.02%	97.61%
8	0.11629	1.29%	98.90%
9	0.09911	1.10%	100.00%

	Coefficients of PC1	Coefficients of PC2	Coefficients of PC3	Coefficients of PC4
Red Meat	0.30261	-0.05625	-0.29758	0.64648
White Meat	0.31056	-0.23685	0.6239	-0.03699
Eggs	0.42668	-0.03534	0.18153	0.31316
Milk	0.37773	-0.18459	-0.38566	-0.00332
Fish	0.13565	0.64682	-0.32127	-0.21596
Cereals	-0.43774	-0.23349	0.09592	-0.0062
Starch	0.29725	0.35283	0.24298	-0.33668
Nuts	-0.42033	0.14331	-0.05439	0.33029
Fruits & Vegetables	-0.11042	0.53619	0.40756	0.46206



Source: https://www.originlab.com/doc/Tutorials/Principal-Component-Analysis

(i) How many Principle Components are suitable to represent this dataset? [3]

(ii) What food groups contribute (i.e., loadings) to the 1st Principle Component? [3]

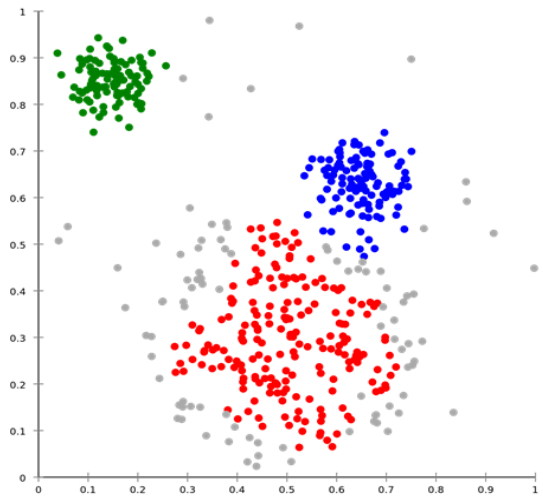
(iii) Identify 2 food groups with related protein consumption. Justify your selection. [6]

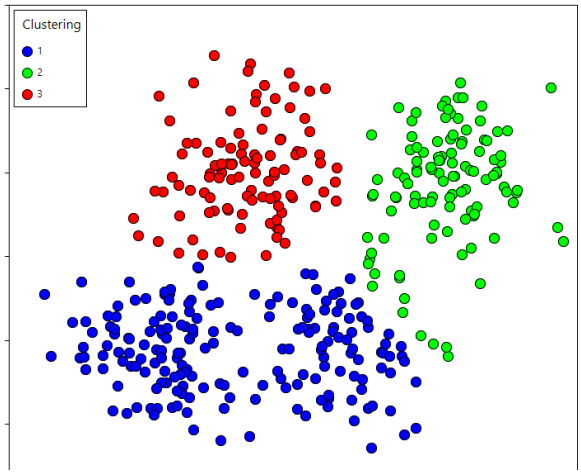
(iv) If you are to cluster the food groups, how many food groups can be formed? Briefly Discuss. [4]

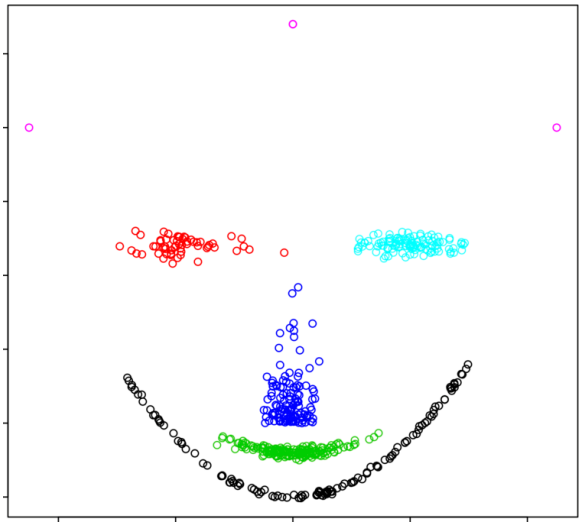
(iv) List 2 other observations from the above analysis. [4]

Question 3 (20 marks)

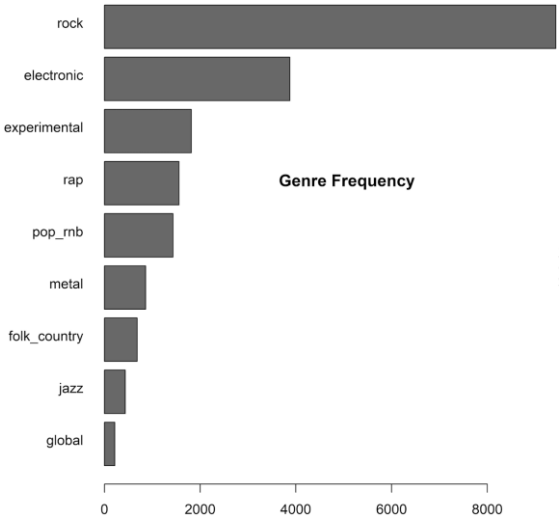
(i) Which clustering technique would you recommend to cluster each of the following datasets. Justify your answer. [3 × 3]

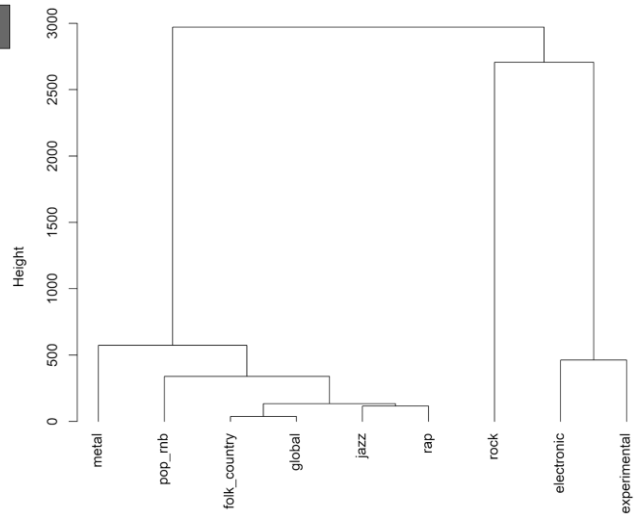






(ii) Following graphs are based on 18,393 music reviews from the Pitchfork website. Data cover reviews between Jan. 1999 and Jan. 2016.



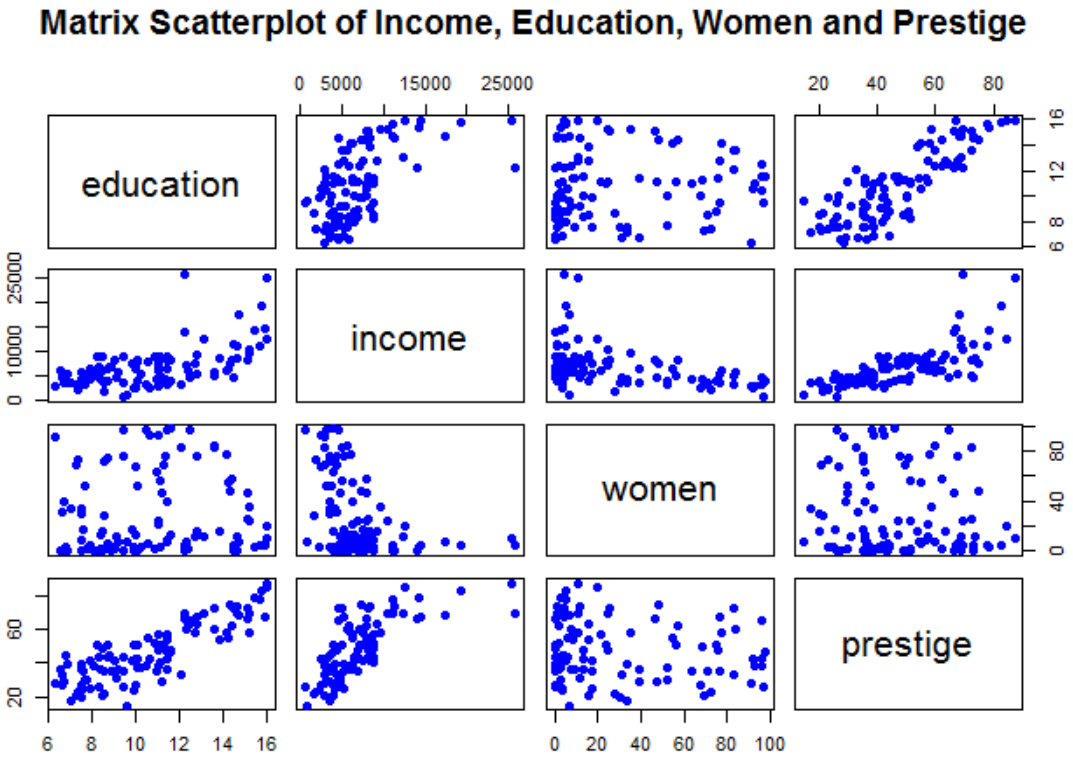


Source: www.r-bloggers.com

What can you conclude from the histogram and cluster analysis? Discuss. [11]

Question 4 (20 marks)

Following scatter plot matrix shows the relationship between average years of education, percentage of women in the occupation, prestige of the occupation education, and income for a set of occupations.




```
mod1 = lm(income ~ education.c + prestige.c + women.c, data=newdata)
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7715.3  -929.7  -231.2   689.7 14391.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6797.902    254.934  26.665  < 2e-16 ***
## education.c  177.199    187.632   0.944    0.347
## prestige.c   141.435     29.910   4.729 7.58e-06 ***
## women.c      -50.896      8.556  -5.948 4.19e-08 ***
```

(i) By analysing the scatter plots and regression model parameters what can you claim about the relationship among attributes. [8]

(ii) Write the corresponding Multiple Linear Regression equation. [4]

(iii) Predict the income of a person, given education = 13.11, women = 11.16, and prestige = 68.8. [3]

(iv) Discuss the accuracy of the fitted Multiple Linear Regression model and whether it can be improved further. [5]

Question 5 (20 marks)

In 2013 PwC (PricewaterhouseCoopers) estimated that "trade eBooks (excluding educational publications) will reach $8.2 billion in sales by 2017 to surpass printed book sales, which are expected to fall from $11.9 billion in 2012 to $7.9 billion in 2017."

(i) Based on the data presented in Figure 6, what are your conclusions about the status of printed book sales as of 2016. Discuss. [14]

(ii) Going forward, do you expect PwC prediction to be correct or not? Briefly Discuss. [6]

---- END OF PAPER ---