1. Association rules mining

Based on my ANU ID u5326448,

(1) Database:

Trans id	Item-set	
T1	a, b, e ,f	
T2	a, b, c, e ,f	
T3	b, c ,e	
T4	a, b, c, e ,f	
T5	b, c ,e	
T6	c, d, e	
T7	b, c ,f	
T8	a, b, d, f	[digit for ANU ID：4]
T9	c, d, e	[digit for ANU ID：4]
T10	a, b, c	[digit for ANU ID：8]

(2) Large 2 item-sets:

Item-set	Count
a, b	5
a, c	3
a, e	3
a, f	4
b, c	6
b, e	5
b, f	5
c, e	6
c, f	3
e, f	3

(3) Large 3 item-sets:

Item-set	Count
a, b, c	3
a, b, e	3
a, b, f	4
a, e, f	3
b, c, e	3
b, c, f	3
b, e, f	3

(4) Candidate 4 item-sets:

Item-set
a, b, e, f
a, b, c, e <- Prunned in Apriori prune step

Some length 3 item-sets included in {a, b, c, e} are not in Large 3 item-sets, for

example, {a, c, e} is not in L3. Thus, {a, b, c, e} need to be pruned.

Large 4 item-sets:

Item-set	Count
a, b, e, f	3

(5) Frequent rules of length 3 from first two large 3 item-sets:

Rule	Support	Confidence	Lift
(a, b) -> c	30	60	0.75
(a, c) -> b	30	100	1.25
(b, c) -> a	30	50	1.00
(a, b) -> e	40	60	0.85
(a, e) -> b	40	100	1.25
(b,e) -> a	40	60	1.20

4.5/5
Missing entries in C4.

2. Characteristics of clustering algorithms

(a) AGNES
(1). Arbitrary shape.
(2). Input final stop numbers of clusters *k*
(3). Time complexity is at least $O(n^2 \log n)$, n is the number of data points. No object function may be minimized directly.

(b) CLARA
(1). Connectivity models (data point with linking).
(2). *K* samples and applies PAM on each sample.
(3). Efficiency based on the size of samples. A good sample based clustering might not necessarily represent a good clustering of the whole data set if the sample is biased.

(c) DBSCAN
(1). Arbitrary shape.
(2). ε (eps) and the minimum number of points.
(3). It is not entirely deterministic. It is difficult to choose a distance threshold ε if the scale and data are not well understood.

(d) *k*-means
(1). Spherical shape.
(2). Input numbers of clusters *k*
(3). It may have problems when the data contains outliers. It also when clusters are of differing size and densities.

2/4
Missing distance measures function in (2) and example application.

3. Classifier accuracy measures

(a) Uni ID: 5326448

TP = 53264
FP = 26448
TN = 326448
FN = 532

(1) Confusion matrix:

	Pred Pos	Pred Neg	
True Pos	53264	532	
True Neg	26448	326448	Total: 406692

(2) Normalised confusion matrix:

	Pred Pos	Pred Neg
True Pos	0.13097	0.00131
True Neg	0.06503	0.80269

(3) Accuracy = (true_pos + true_neg) / (all_class_pos + all_class_neg)
= (53264 + 326448) / (53264 + 26448 + 326448 + 532)
= 93.37%
Error rate = 1 – Accuracy = 1 – 93.37% = 6.63%

(4) and (5)

Specificity = true_neg / all_true_neg = 326448 / (26448 + 326448) = 92.51%
Precision = true_pos / (true_pos + false_pos)
= 53264 / (53264 + 26448) = 66.82%
Recall = true_pos / all_true_pos = 53264 / (53264 + 532) = 99.01%
F-measure = 2 * Precision * Recall / (Precision + Recall)
= 2 * 66.82% * 99.01% / (66.82% + 99.01%)
= 79.79% (Wikipedia, 2016)

3/3

(b) Balanced Classification Rate = 1/2(Specificity + Recall)
= 1/2(92.51% + 99.01%) = 95.76%
Since True_pos << True_neg, thus we consider that harmonic mean of precisi and recall would be more accurate. I choose F-measure as measurement. F score is 79.79%, so I think this classification problem is balanced.

0/1
Incorrect balance ratio + explanation.

4. Decision tree classification

(a) Gender = Male, Age = Young, it is true positive and class= Pre_Neg, thus the record is a false negative.
(b) Gender = Male, Age = Old, Has_car = No, it is true positive and class= Pre_Neg, thus the record is a false negative.

(c) Gender = Female, Student = No, it is true negative and class= Pre_Neg, thus the record is a true negative.

(d) Gender = Female, Student = Yes, Employed = Yes, it is true negative and clas Pre_Pos, thus the record is a false positive .

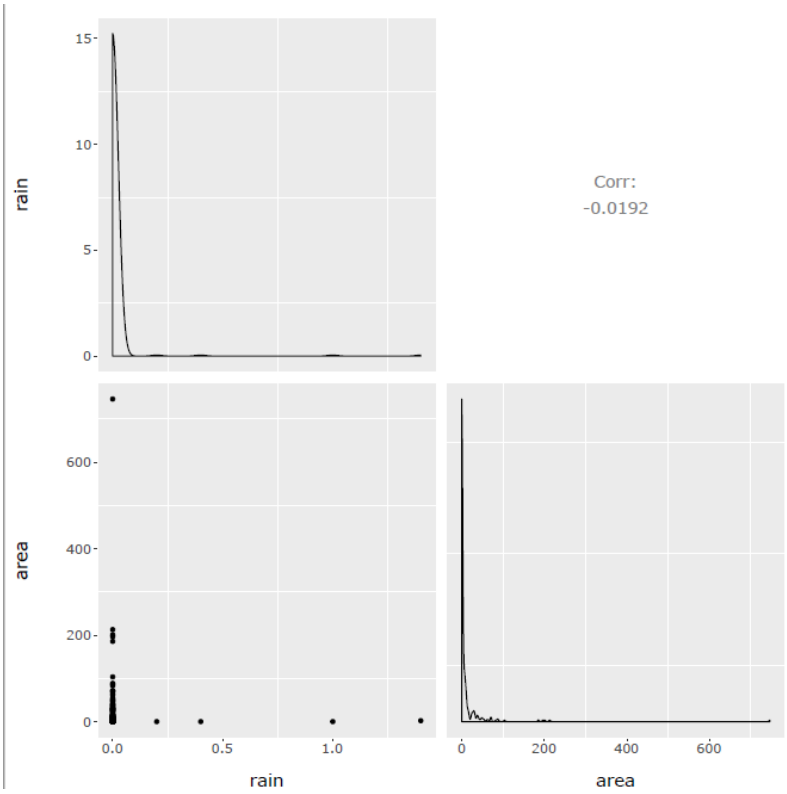
0.5/2 Incorrect (a),(b) and (d). Need to revise on relevant lectures.

5. Data mining project

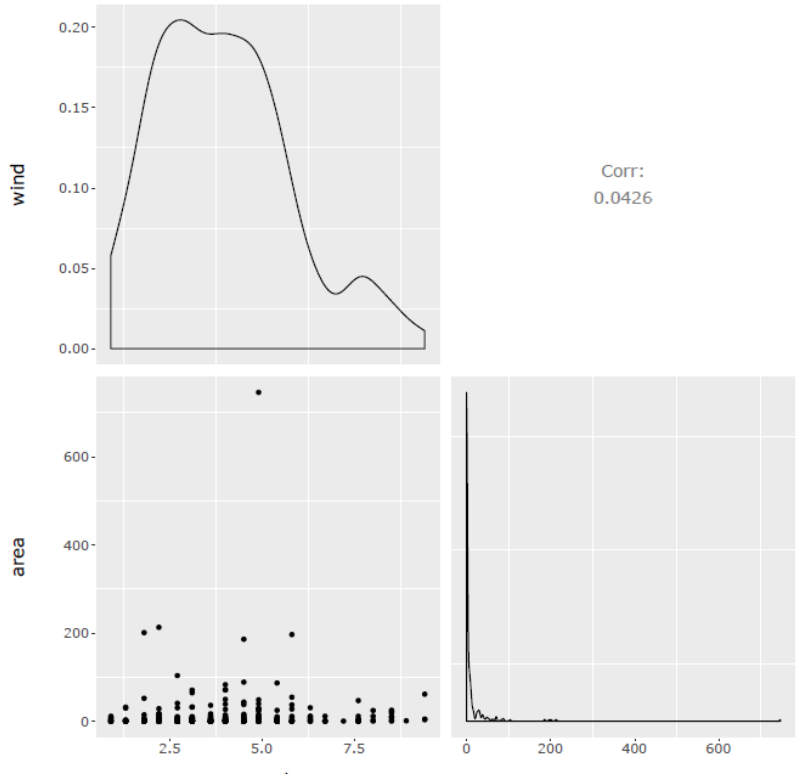
Data set: http://archive.ics.uci.edu/ml/machine-learning-databases/forest-fires/

Variable	Data Type	Comment	Maximum	Minimum	Mean	Media
X	Numeric	9	9.000	1.000	4.598	4.000
Y	Numeric	7	9.000	2.000	4.244	4.000
month	Categoric	12	August	June	/	/
day	Categoric	7	Sunday	Wednesday	/	/
Temp	Numeric	192	33.30	2.20	18.81	19.30
RH	Numeric	75	100.0	15.0	44.5	42.0
Wind	Numeric	21	9.400	0.900	3.963	4.000
Rain	Numeric	7	1.40000	0.00000	0.00831	0.00000
area	Numeric	251	746.28	0.00	10.94	0.00

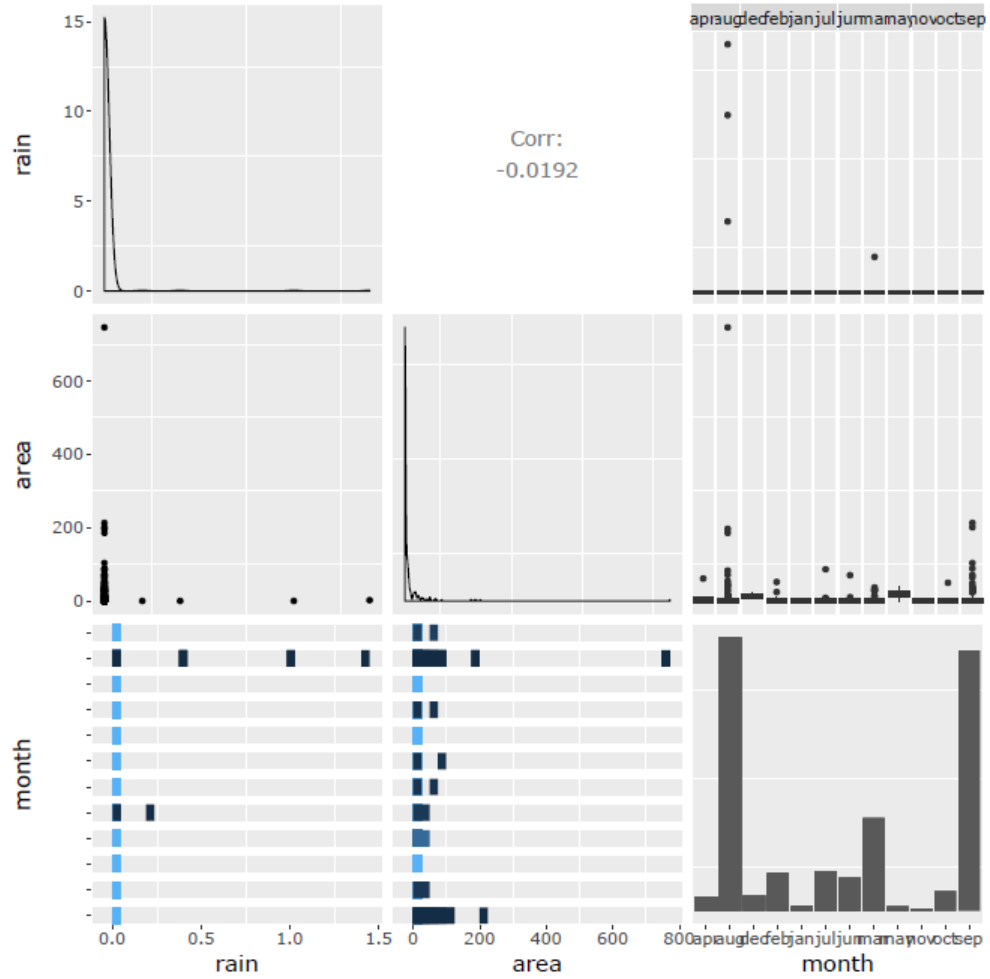
When we choose rain and area:



When we choose wind and area:



When we choose rain, area and month together:



0.5/5
Missing answers + need more explanations.

10.5/20

Missing answers + need to revise on Q4 but otherwise OK.