Baseline Questions (TCX3212)

Q1

Which of the following is/are of nominal data type?

i. Hotel class from 1-star, 2-stars… to 5 stars where the higher number of stars indicate a more luxurious hotel

ii. Ticket prices (in $) to different tourist attractions

iii. Types of tourist attractions (art, adventure, nature, culture, etc.)

iv. Location of attractions (e.g. Jurong, Mandai, Sentosa, Orchard, etc.)

v. Distance from MRT for the tourist attractions (in meters)

a) i only.

b) i, iii and iv only.

c) ii and v only.

d) iii only.

e) iii and iv only.

Q2

Suppose an analyst collected data from a supermarket on the daily sales (in dollars) of vegetables in the last year. His analysis of the data produces the following results:
- N = 365 (i.e. one observation for each day of the year)
- mean = $2350
- standard deviation = $50
- standard error = $2.617
- Prediction Interval = ($2251.54, $2448.46)

Assume that the sales of vegetable is relatively stable throughout the year.
Which of the following is a correct interpretation of the results?

a) We can predict with 95% confidence level that the sales of vegetables on a given day would be between $2251.54 to $2248.46.
b) We can predict with 95% confidence level that the mean daily sales for vegetables will be between $2251.54 to $2248.46.
c) We can predict that there is a 95% probability that the average daily sales for vegetables will lie between $2251.54 to $2248.46.
d) We can be 95% sure that the average daily sales of vegetables would be between $2251.54 to $2248.46.
e) We can predict there is a 95% chance that a randomly chosen sample mean will lie between $2251.54 and $2448.46.

Q3
A study was conducted on the 2020 cohort of graduates to gather their salary data. A survey was administered to the graduates one year after graduation and again two years after graduation. Graduates who responded to both surveys were kept in the dataset for analyses. In order to compare if there is a significant increase in salary from after one year after graduation to after the second year, which of the following would be the most appropriate test to conduct?

a) Paired 2-sample t-test
b) Independent 2-sample t-test
c) Anova test
d) Paired z-test
e) One sample t-test

(Modified – replace question & image

Q4

The table below shows the descriptive statistics for the variable `Weekly Expenditure` computed using the psych package.

Descriptive Statistics for Weekly Expenditure

n	mean	sd	median	min	max	skew	kurtosis	se
40	328.82	146.70	294.90	114.96	775.02	2.025	0.994	23.19

Which of the following best describes the distribution of this variable?

a) Highly left skewed with shorter and thinner tails than normal distribution.

b) Highly positively skewed with longer and thicker tails than normal distribution.

c) Moderately positively skewed with shorter and thinner tails than normal distribution.

d) Relatively symmetrical with longer and thicker tails than normal distribution.

e) Perfectly symmetrical and follows a normal distribution.

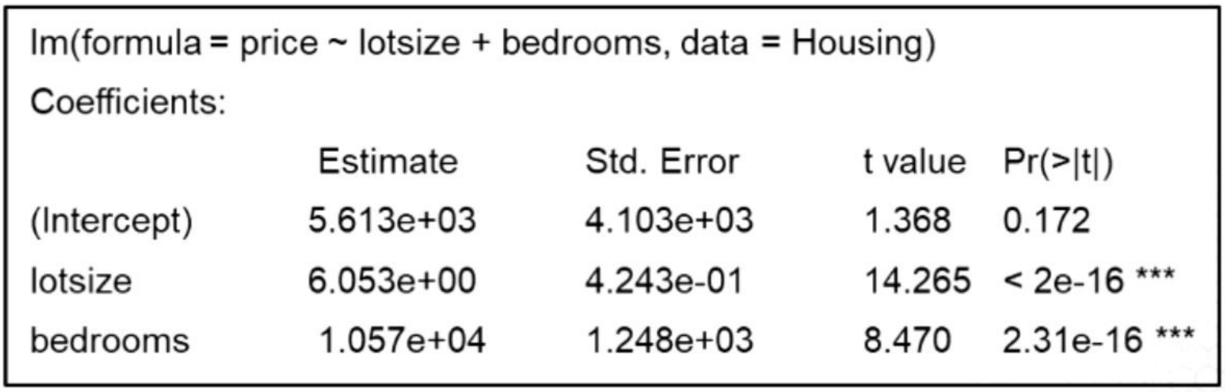
(Modified – replace question & image)

Q5

The housing data has been modelled using a multivariate regression function with the independent variables, lot size and number of bedrooms. The regression function is represented as follows:

Price ~ lotsize + bedrooms

Review the outputs of the results and choose the correct statement(s).



```
lm(formula = price ~ lotsize + bedrooms, data = Housing)
Coefficients:
               Estimate    Std. Error    t value   Pr(>|t|)
(Intercept)    5.613e+03   4.103e+03     1.368     0.172
lotsize        6.053e+00   4.243e-01     14.265    < 2e-16 ***
bedrooms        1.057e+04  1.248e+03     8.470     2.31e-16 ***
```

i. Adding more variables decreases the overall R-Square value of the multiple regression.
ii. In the regression output, a p-value of <2e-16 *** means that there is not much evidence for the regression coefficient of lot size to be different from zero.
iii. In the regression output, a p-value of 2.31e-16 *** means that there is evidence for the regression coefficient of bedrooms to be different from zero.
iv. For one more bedroom, we can expect the average price of the home to increase by $5613, when all other independent variables are held constant.
v. The fitted multivariate regression function is given by Price = 5613 + 6.053 lot size + 10570 bedrooms.

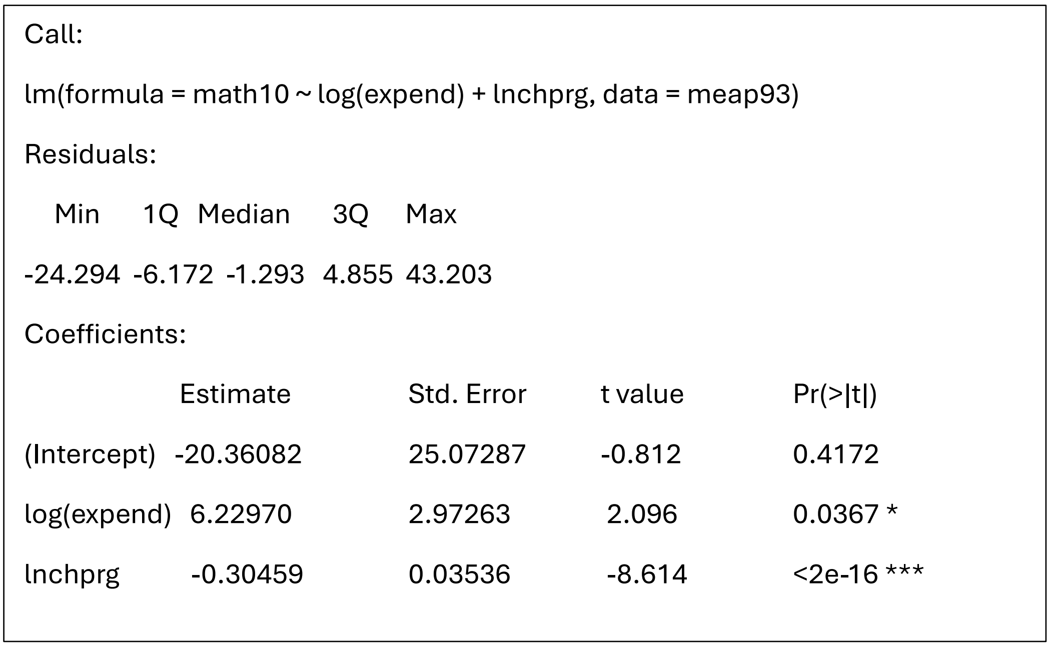
a) i and ii only
b) ii, iii and v only
c) i, iii and v only
d) iii and v only
e) i, ii, and v only

(Modified – replace image)

Q6

The multivariate linear regression output shows a multivariate linear regression for the percentage of students passing MEAP math. The variables are:

- math10: percentage of students passing MEAP math
- expend: expenditure per student in dollars
- lnchprg: percentage of students in school lunch programme which is linked with poverty rates
- droprate: school dropout rate, in percentage
- gradrate: school graduation rate, in percentage



```
Call:

lm(formula = math10 ~ log(expend) + lnchprg, data = meap93)

Residuals:

   Min     1Q  Median     3Q    Max

-24.294 -6.172  -1.293  4.855  43.203

Coefficients:

               Estimate          Std. Error      t value        Pr(>|t|)

(Intercept)  -20.36082          25.07287        -0.812          0.4172

log(expend)   6.22970           2.97263          2.096          0.0367 *

lnchprg       -0.30459          0.03536         -8.614          <2e-16 ***
```

What is the correct interpretation of the coefficient estimate for ‘lnchprg’ in this regression model?

a) For every 1-percentage-point increase in percentage of students in school lunch programme, the predicted percentage of students passing MEAP math decreases by 0.30%, holding other variables constant.

b) Schools with higher lnchprg always have lower math10 scores, regardless of all other factors.

c) A 1% increase in lnchprg causes a 30% decline in math10.

d) The relationship between lnchprg and math10 is not statistically significant at any reasonable confidence level.

e) lnchprg explains all variation in math10 once log(expend) is controlled for.

Samantha Sow sowjs@nus.edu.sg

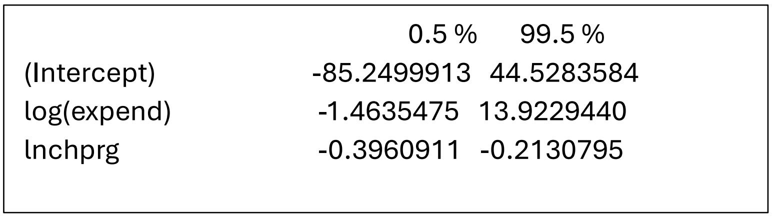
Modified –image)

Q7.

The multivariate linear regression output shows a multivariate linear regression for the percentage of students passing MEAP math. The variables are:

- math10: percentage of students passing MEAP math
- expend: expenditure per student in dollars
- lnchprg: percentage of students in school lunch programme which is linked with poverty rates
- droprate: school dropout rate, in percentage
- gradrate: school graduation rate, in percentage

The results of the confidence interval for the estimated OLS model is provided with confint (lm, level=0.99).



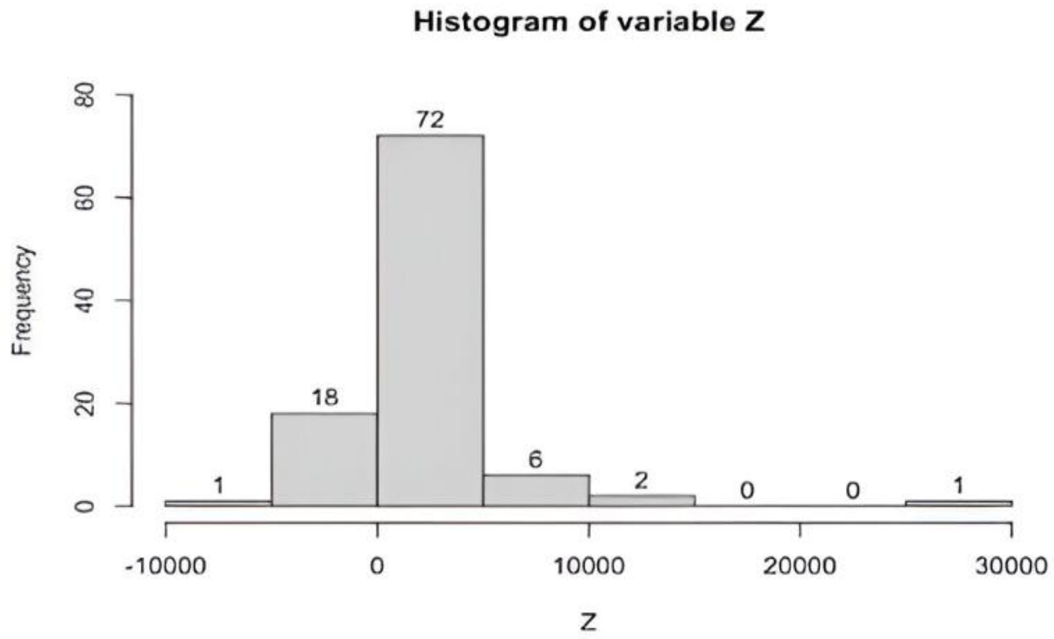
```
                              0.5 %      99.5 %
(Intercept)             -85.2499913  44.5283584
log(expend)              -1.4635475  13.9229440
lnchprg                  -0.3960911  -0.2130795
```

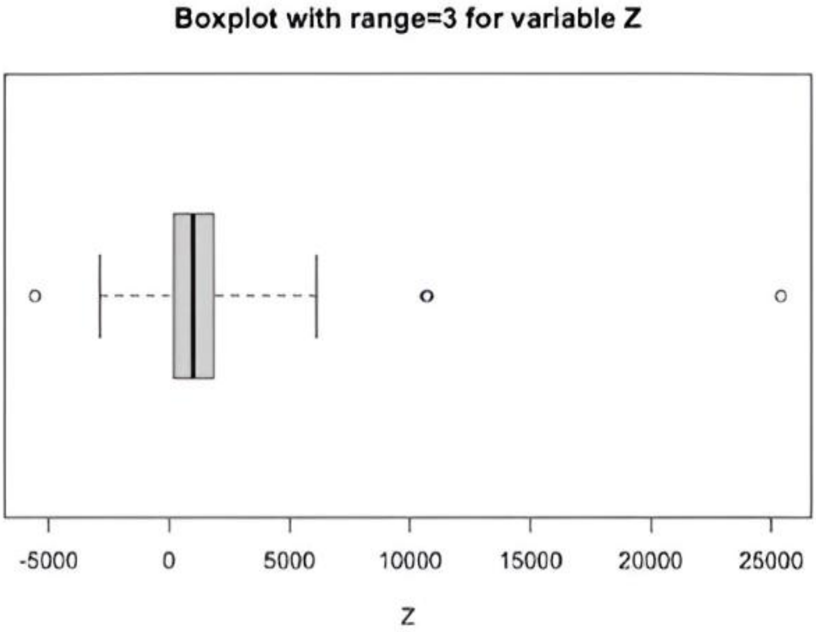
Based on the 99% confidence interval, which statement is correct?

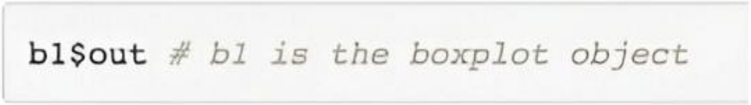
a) Because the 99% CI for lnchprg does not include zero, there is strong evidence that lnchprg is associated with lower math performance.
b) Although the coefficient is negative, the wide confidence interval suggests the relationship is too uncertain to be interpreted in policy contexts.
c) Since the 99% CI is narrow and fully above zero, lnchprg is a positive predictor of math10 after adjusting for expenditure.
d) The CI includes zero at the 99% level, so lnchprg is not statistically significant once log(expend) is controlled for.
e) The coefficient is significant only at the 95% level but not at the 99% level, indicating moderate evidence against the null hypothesis.

Q8.

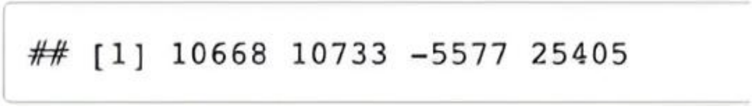
The histogram and box plot (with range=3) for a variable Z are shown below.







```
b1$out # b1 is the boxplot object
```



```
## [1] 10668 10733 -5577 25405
```

Using the above information, which of the following can you conclude about Z?
i) Z has a left skewed distribution.
ii) Z has one outlier based on the histogram.
iii) Z has three outliers based on the boxplot output.
iv) Z has a right skewed distribution.
v) Z is normally distributed.

a) ii and iv only.
b) ii and iii only.
c) v only.
d) i, ii and iii only.
e) ii, iii and iv only.

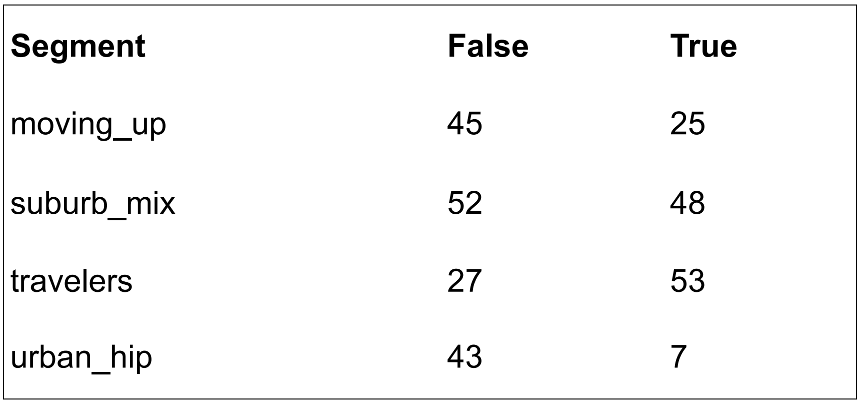
Q9
Given the following DataFrame store_sales, which of the following code snippets will correctly calculate the **mean of each row** across columns 3 to 8?

a) store_sales.iloc[3:9].mean(axis=0)
b) store_sales.iloc[:, 3:9].mean(axis=0)
c) store_sales.iloc[:, 3:9].mean(axis=1)
d) store_sales.iloc[3:9].mean(axis=1)
e) store_sales.iloc[:, 3:9].mean()

Modified – replaced image)

Q10

You have the following dataset, segment_data, with the columns Segment, own_home, and subscribe. You want to compute the frequency of different combinations of the Segment and own_home columns and get a breakdown of the number of customers who own their homes (True) and those who do not own their homes (False) in each segment. The desired output is as follows:



Segment	False	True
moving_up	45	25
suburb_mix	52	48
travelers	27	53
urban_hip	43	7

Which of the following code snippets will produce this result?

a) segment_data.groupby(['Segment', 'own_home'])['subscribe'].count().unstack()
b) segment_data.groupby(['Segment', 'own_home'])['subscribe'].sum().unstack()
c) segment_data['own_home'].value_counts().unstack()
d) segment_data.pivot(index='Segment', columns='own_home', values='subscribe').count()
e) segment_data.groupby(['Segment', 'own_home'])['subscribe'].count().pivot_table(columns='own_home')

Modified – replace image

Q11

You are working with a dataset, Auto, that contains information about cars, including their year, weight, and origin. The dataset is loaded from a CSV file as follows:

python

```
Auto = pd.read_csv( 'Auto.data',  na_values=['?'],  delim_whitespace = True)
```

You want to create a new DataFrame that consists of the weight and origin columns for only the cars built **after 1980** (i.e., those with a year greater than 80).
Which of the following code snippets will achieve this?

a) idx_80 = Auto['year'] > 1980
Auto_re = Auto.loc[idx_80, ['weight', 'origin']]
b) idx_80 = Auto['year'] > 80
Auto_re = Auto.loc[idx_80, ['weight', 'origin']]
c) idx_80 = Auto['year'] <= 80
Auto_re = Auto.loc[idx_80, ['weight', 'origin']]
d) idx_80 = Auto['year'] > 80
Auto_re.loc[idx_80, ['weight', 'origin']]
e)Auto_re = Auto[Auto['year'] > 80][['weight', 'origin']]

Modified – replace mage

Q12.
Consider the following hypothetical data concerning student characteristics and whether or not each student should be hired.

Name	GPA	Effort	Hirable
Sarah	poor	lots	Yes
Dana	average	some	No
Alex	average	some	No
Annie	average	lots	Yes
Emily	excellent	lots	Yes
Pete	excellent	lots	No
John	excellent	lots	No
Kathy	poor	some	No

Using a Naive Bayes classifier, compute the score for Hirable = Yes, given that GPA = poor, and Effort= lots which is P(Hirable = Yes | GPA = poor, Effort = lots).

a) $\text{Score(Yes)} = P(\text{Yes}) \times P(\text{GPA=poor} \mid \text{Yes}) \times P(\text{Effort=lots} \mid \text{Yes}) = \frac{3}{8} \times \frac{1}{3} \times 1 = \frac{1}{8}$

b) $\text{Score(Yes)} = P(\text{No}) \times P(\text{GPA=poor} \mid \text{No}) \times P(\text{Effort=lots} \mid \text{No}) = \frac{5}{8} \times \frac{1}{5} \times \frac{2}{5} = \frac{1}{20}$

c) $\text{Score(Yes)} = \frac{\frac{1}{8}}{\frac{1}{8}+\frac{1}{20}} = \frac{5}{7}$

d) Score(Yes)=P(Yes)×P(GPA=poor|Yes)×P(Effort=lots|Yes)=1/2×1/3×1=1/6

e) Score(Yes)=P(Yes)×P(GPA=poor|Yes)×P(Effort=lots|Yes)= 5/8×1/3×1=5/24

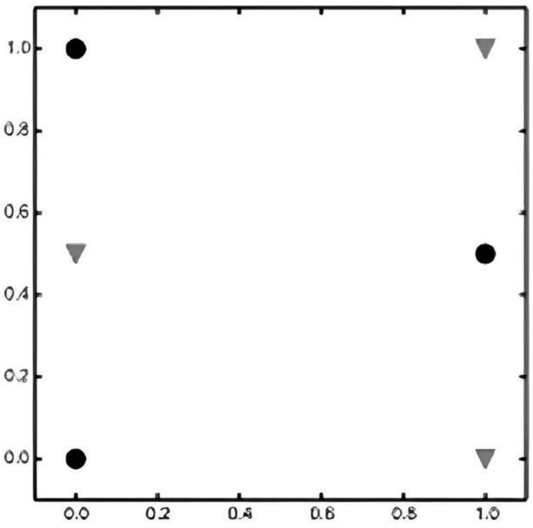
Q13 An analyst is working on a genomic classification problem with 5,000 predictors and 200 samples. She first selects the 100 predictors most correlated with the class labels using all samples and then performs 5-fold cross-validation to estimate prediction error using these 100 predictors. The cross-validation error comes out as 3%, while the true test error is 50%.

Which of the following statement(s) best explain this **discrepancy** between the cross-validation and true test errors?
i. The predictors were chosen using all samples, causing information from the test folds to leak into the model.
ii. Cross-validation was applied only after predictor selection, leading to an overly optimistic error estimate.
iii. The sample size was too small to prevent overfitting during cross-validation.
iv. The classifier inherently performs poorly with standardized data.
v. The predictors were not filtered by variance before model fitting.

a) ii and iv only.
b) iv only.
c) v only.
d) i and ii only.
e) i only.

Q14 Check all the binary classifiers that are able to correctly separate the training data (circles vs. triangles) given in the Figure.

a) Logistic regression
b) Support Vector Machines (SVM) with linear kernel
c) Support Vector Machines (SVM) with RBF kernel
d) Decision tree
e) 3-nearest-neighbor classifier (with Euclidean distance)



Q15: In Bayesian learning, what does the posterior probability represent?

a) The probability of the observed data given the model parameters.
b) The initial probability of the model before observing data.
c) The updated probability of a model, after having seen the data.
d) The probability of the data marginalized over all parameter values.
e) The probability of the model being true without considering data

Q16: What are support vectors in the context of Support Vector Machines (SVM)?

a) The training examples farthest from the decision boundary.
b) The only training examples necessary to compute the decision function $f(x)$ in an SVM.
c) The class centroids (mean points of each class).
d) The directions that maximize the separation between classes
e) All training examples that are correctly classified by the model.

Q17 According to ensemble methods, which of the following statements about bagging, random forests, and boosting is **FALSE**?

a) Bagging reduces variance by averaging predictions from multiple trees fit to bootstrapped samples of the training data.
b) Random forests improve upon bagging by decorrelating trees through feature randomization at each split.
c) Boosting builds trees sequentially, with each new tree trained on a modified version of the original data based on previous trees' residuals.
d) Out-of-bag (OOB) error estimation is available for boosting but not for bagging or random forests.
e) e) In bagging, increasing the number of trees B generally does not lead to overfitting.

Q18 Which statement about decision trees is **FALSE**?

a) Regression trees predict using the mean response in each region.
b) Gini index and entropy measure node purity for classification trees; RSS is for regression trees.
c) Cost-complexity pruning grows a large tree first, then prunes based on α.
d) Recursive binary splitting is a bottom-up approach.
e) Decision trees handle both quantitative and qualitative predictors.

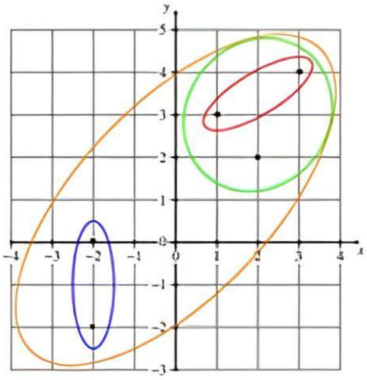
Q19 You are a set of feature vectors {(-2, -2), (-2, 0), (1, 3), (2, 2), (3, 4)}

We apply **greedy agglomerative hierarchical clustering** using **centroid linkage**.

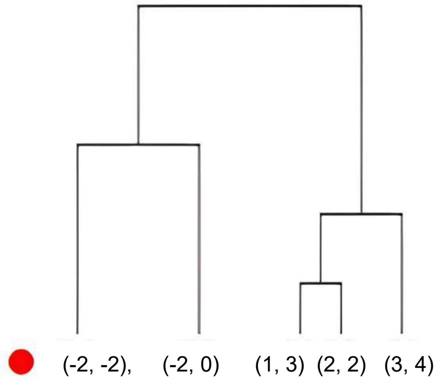
Two possible visualisations are provided:

- Each row shows a **cluster diagram (left)** and a **dendrogram (right)**.
- Only **one row** correctly shows both the clustering behaviour **and** the matching dendrogram.

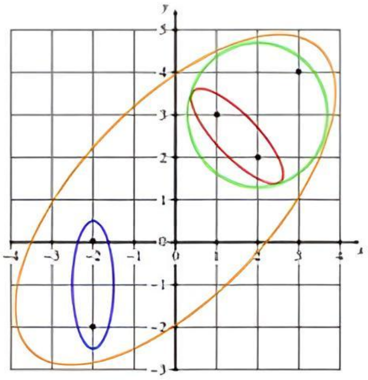
Which option shows the correct pair (cluster diagram **and** dendrogram) for centroid linkage?



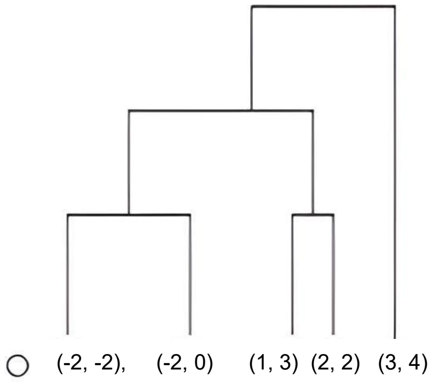
A



B



C

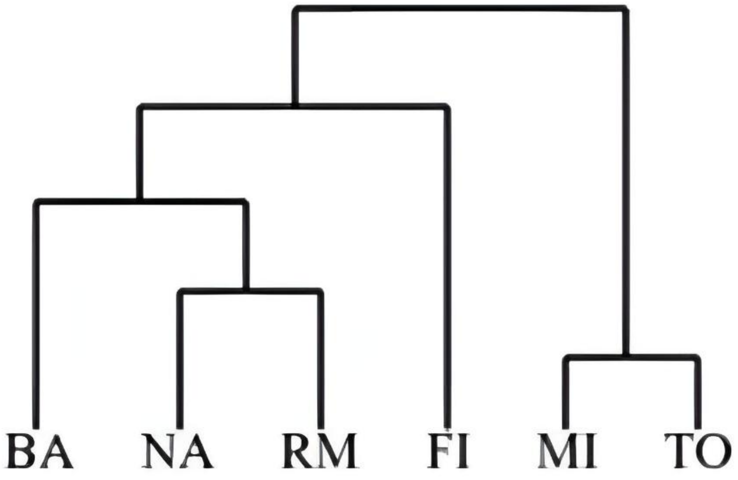


D

a) A and C
b) A and D
c) B and C
d) B and D
e) None of the pairs

Samantha Sow sowjs@nus.edu.sg

Q20 Consider the dendrogram. Using this dendrogram to create 3 clusters, what would the clusters be?



a) {BA, NA}, {RM, FI}, {MI, TO}
b) {NA, RM}, {BA, FI}, {MI, TO}
c) {BA, NA, RM, FI}, {MI}, {TO}
d) {BA, NA, RM}, {FI}, {MI, TO}
e) None of these

Q21. In simple exponential smoothing forecasting to give higher weightage to recent demand information, the smoothing constant must be close to

a) -1
b) Zero
c) 0.5
d) 1
e) None of the above

Q22. For a hotel, the actual demand for disposable cup was 600 units in January, and 700 units in February. The forecast for month of January was 500 units. What will be the forecast for month of March. Use simple exponential smoothening method (Smoothening coefficient = 0.8)

a) 676
b) 576
c) 680
d) 580

e) 612

--------------------------The end ---------------------------

Samantha Sow sowjs@nus.edu.sg

Explanation to Q12

Solution

Total instances: $N = 8$

Count of Hirable = Yes: $N_{Yes} = 3$

Count of Hirable = No: $N_{No} = 5$

$$P(\text{Yes}) = \frac{3}{8}$$

$$P(\text{No}) = \frac{5}{8}$$

Likelihoods for Hirable = Yes

$$P(\text{GPA=poor} \mid Yes) = \frac{1}{3}$$

$$P(\text{GPA=average} \mid Yes) = \frac{1}{3}$$

$$P(\text{GPA=excellent} \mid Yes) = \frac{1}{3}$$

$$P(\text{Effort=lots} \mid Yes) = \frac{3}{3} = 1$$

$$P(\text{Effort=some} \mid Yes) = \frac{0}{3} = 1$$

Likelihoods for Hirable = No

P(GPA=poor|No) = 1/5

P(GPA=average|No) = 2/5

P(GPA=excellent|No) = 2/5

P(Effort=lots|No) = 3/5

P(GPA=some|No) = 2/5

For Hirable = Yes:

$$\text{Score(Yes)} = P(\text{Yes}) \times P(\text{GPA=poor} \mid Yes) \times P(\text{Effort=lots} \mid Yes)$$

$$= \frac{3}{8} \times \frac{1}{3} \times 1$$

$$= \frac{1}{8}$$

For Hirable = No:

$$\text{Score(No)} = P(\text{No}) \times P(\text{GPA=poor} \mid No) \times P(\text{Effort=lots} \mid No)$$

$$= \frac{5}{8} \times \frac{1}{5} \times \frac{2}{5}$$

$$= \frac{1}{8} \times \frac{2}{5} = \frac{2}{40} = \frac{1}{20}$$

$\text{Since } 0.125 > 0.05$, Naive Bayes predicts **Hirable = Yes**.

Samantha Sow sowjs@nus.edu.sg