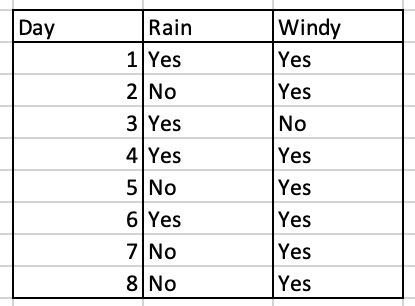
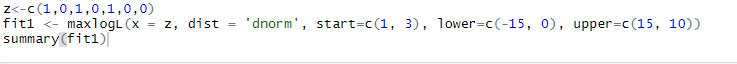
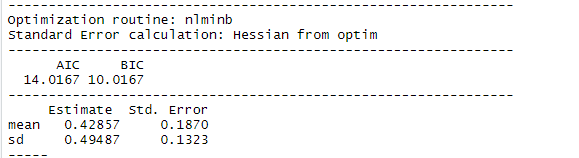
1.1 Given the following data, what is the Maximum Likelihood Estimate of the conditional probability P(Rain | Windy)?



Answer: 0.42857 Calculated with R programming and consider about rain when we have windy 1 if rain = “Yes” and 0 if rain = “NO”





1.2 From the previous data, what is the chance of rain and not windy?

Answer: 0.125 or 1/8

1.3 From the previous data, are the rain and wind conditions independent? Justify your answer.

Answer: Independent

From the sample above we can see that rain can be yes regard windy yes or no (case day1 and day 3 by example) and we can have wind yes regard rain yes or no (case 1 and case 2 by example) so the conditions is independent

1.4 From the data, it always rains when there's no wind. Is this a good conclusion? Why?

Answer: Bad conclusion we can observe only 1 day that have no wind from statistic theory we must have far much than 1 observation to have a good conclusion maybe if we observe day 9 we can see rain and wind are both no and the conclusion is invalid by just only add 1 observation it is no way this can be good conclusion

1.5 Which statement is the most related to the Central Limit Theorem?

A.Somchai thinks the sampled mean should be close to the actual mean if he collects a lot of data

B.Somjai models the number of people in his company that comes to work at any given day with a Gaussian distribution.

C.Sompong thinks the noise from the electronic truck scale (truck weighting device) as a Gaussian distribution.

Answer: A

The key concept of Central Limit Theorem is if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population **with replacementtext annotation indicator**, then the distribution of the sample means will be approximately normally distributed

1.6 Which statement is the most correct?

A.Prapard conducted an experiment with 150 samples and gets a significant value that is very close to significance. He then collect 5 more samples to get a significant result

B.Pravid tested his hypothesis whether COVID-20 is transmitted via the internet. The null hypothesis is no transmission. His hypothesis testing comes out not significant. So he concluded that COVID-20 does not transmit via the internet.

C.Prayud tested his hypothesis whether COVID-20 is transmitted via the internet. The null hypothesis is no transmission. His hypothesis testing comes out significant. So he concluded that COVID-20 transmits via the internet.

D.Prada thinks an experiment with p = 0.001 with n = 100 has a stronger conclusion than an experiment with p = 0.001 with n = 10.

F.Pranee concludes that p-value should be lower with a larger n given everything else the same.

Answer: D)

1. Is wrong collect more data does not always increase your significant value it depend on your variable
2. Is wrong Null hypotheses are never accepted. We either reject them or fail to reject them. The example show we accept null hypothesis so it is wrong
3. Is acceptable but I think it depends on your methodology. If your methodology is wrong your control is not proper. The conclusion may be invalid
4. True with same p-value more sample size has a stronger conclusion than smaller sample size
5. P value can go anyway (upper or lower) if H0 is true p value can go upper with sample size increase

1.7 If two random variables are uncorrelated, they are:

A.Independent

B.Dependent

C.Inconclusive

Answer: C

In statistic 2 variable are uncorrelated if their [covariance](https://en.wikipedia.org/wiki/Covariance) is zero and no linear relationship between them

If two variables are [independent](https://en.wikipedia.org/wiki/Statistical_independence), with finite [second moments](https://en.wikipedia.org/wiki/Second_moment), then they are uncorrelated. However, not all uncorrelated variables are independent. So the answer is C.) Inconclusive

2.1 Matrix A is of size [3,4]. Matrix B is of size [4,3]. Which is(are) correct? (Select all that apply)

A.AB is of size [3,3]

B.BA is of size [3,3]

C.(ABA)' is of size [4,3] (' is transpose)

D.A'B' is of size [3,3]

F.A'B is of size [4,3]

Answer: A,C

1. Is correct [3,4] matrix multiply by [4,3] matric result is matrix with [3,3]size
2. Is wrong [4,3] matrix multiply by [3,4] matric result should be [4,4] matrix
3. (ABA)’ is [3,4] \*[4,3] \*[3,4] matrix the result is [3,4] matrix and we transpose it to become [4,3] size matrix so it is correct
4. A transpose is [4,3] multiply by B transpose is [3,4] the result is [4,4] size matrix and it is wrong

F) A transpose is [4,3] cannot multiply with B which is [4,3] matric cause number of column in a transpose is not equal to row of matrix B

2.2 If a real-valued matrix is square and full rank, (Select all that apply)

A.it is invertible

B.it has a non-zero determinant \_

C.the number of rows and columns are the same \_

D.it is a diagonal matrix

E.its eigenvalues are non-zero

F.the columns are linearly independent \_

G.the rows are linearly independent \_

H.its eigenvectors are real-valued

I.its eigenvalues are real-valued

Answer: B,C,F,G,H,I

2.3 Somnuke collects 10 independent data points each with 20 features. Which statements is(are) true or proper? (Select all that apply)

A.The covariance matrix of this data is invertible.

B.He can reduce the dimension of the data to 15 dimensional using PCA

C.If he wants to predict the 20th feature with the first 19 features, the model is underdefined.

Answer: B,C

PCA is technique for select the feature of model while minimizing information loss so it can be reduce dimension of data

I think if we do not collect 20th feature but we use 19 features to predict 20 features it is not a way to fully define a model so it is underdefined. Which is okay in some case since we cannot collect 20th feature of the sample

3.1 F(x) = |x| for any real-valued X. F(x) is (Select all that apply)

A.Continuous

B.Differentiable

C.Bounded

D.Convex

E.Concave

F.Lipschitz continuous

Answer: A,F

1. This function is continuous cause it define for any real-valued
2. is wrong this function is not differentiable at x=0
3. This Function is Bounded Below since it cannot go below 0 F(x) >=0 always so it is Bounded
4. It cannot be convex since f(x) cannot be decrease when x is increase
5. I don’t sure we can accept straight line as concave since F(x) = |x| is straight line but i think straight line is not concave so E) is wrong
6. This function is Lipschitz continuous cause for every pair of points on the graph of this function, the absolute value of the slope of the line connecting them is not greater than this real number

3.2 Backpropagation is an application of

A.The sum-product rule

B.The chain rule

C.The fundamental theorem of Calculus

D.The integration by parts

Answer: B)

The Backpropagation is used to effectively train a neural network through a method called chain rule. In simple terms, after each forward pass through a network, backpropagation performs a backward pass while adjusting the model’s parameters

4.1 Which technique(s) is(are) used for reducing overfitting in neural networks? (Select all that apply)

A.Weight decay

B.L2 regularization

C.Dropout

D.Batchnorm

E.Residual connection

F.Data augmentation

G.Learning rate decay

H.ADAM

I.Label smoothing

J.Early stopping

K.Stack more layers

Answer: A,B,C,F

4.2 If a neural network has 5 inputs, 2 hidden layers (feed forward layers of 10 neurons each), and 1 output neuron. How many parameters are there in this model? Don't forget the biases.

Answer: 181

We can calculate parameters by connections between layers + biases in every layer so the answer is (5\*10 +10\*10 + 10\*1) + (10+10+1) = 181 parameters

4.3 If the input is an image with size 200x200x3, the input is passed through a 2D convolution with 20 filters of size 3x3, stride 2, padding='valid', what is the size of the resulting output? (Answer format: ??x??x??)

Answer: 98.5\*98.5\*20

The formula [(W−K+2P)/S]+1 in this case is [200-3+ 0 /2]+1 so the output shape is 98.5\*98.5\*20

4.4 Which statement is proper?

A.Maythinee backpropagate to the input layer to find which feature is the most responsible for the output of the model

B.Junepimath picks the model with the lowest training loss for production

C.Julinee uses the weight of the model to tell why the input is classified as class 10

D.Augarith thinks that a lower cross entropy loss on the validation set will give a better accuracy on the validation set.

Answer: C)

1. Is wrong backpropagate is not the way to find most responsile feature
2. Is wrong training loss does not always mean good model , overfit problem should be consider
3. You can interpret the result of prediction (or classification) with the weight of the model so C) is proper
4. Is wrong lower cross entropy does not mean better accuracy with the same dataset

5.1 Which is(are) NOT true about how to prepare train and test? (Select all that apply)

A.Missing values should be imputed before splitting train and test.

B.A random split is not suitable for a time series forecasting task.

C.A stratification should be considered for a classification task.

D.A stratification is not possible for a regression task.

Answer: C,D

1. Is true but not always maybe we can delete observation that contain missing values if we have large dataset and only small data contain missing value imputed it before splitting train test reduce the process we impute it after splitting
2. Time series forecast should not split random it is true since it continuity
3. Is False some classification task is no need to use stratification
4. Stratification can do in a regression task

5.2 Which is NOT true about how to split train and test in sklearn?

A.The command can split data into train, test, and validate altogether.

B.Inputs and target must be prepared into separated variables.

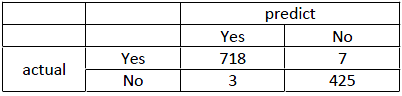
C.Once the option “shuffle” is False, a stratification cannot be performed.

D.The option “random\_state” should always be defined.

Answer: A)

1. Train\_test\_split is split data into 2 dataset which is train test dataset not train,test and validate dataset
2. True in order to have x\_train,x\_test,y\_train,y\_test you must prepared separated variables
3. True in help(train\_test\_split) it says shuffle boolean, optional (default=True)Whether or not to shuffle the data before splitting. If shuffle=False then stratify must be None.
4. Random\_state should be defined since in real work we need model that produce same result of train test dataset even with we run the code again

5.3 The table below is a confusion matrix of a prediction model to classify Covid‐19’s patients (Yes = Covid detected and No = undetected). Please compute F1 for both classes.



A.F1(Yes) = 0.9884, F1(No) = 0.9523

B.F1(Yes) = 0.9884, F1(No) = 0.9931

C.F1(Yes) = 0.9931, F1(No) = 0.9523

D.F1(Yes) = 0.9931, F1(No) = 0.9884

Answer: D)

For this answer we need to calculate precision and recall for each class since F1 score Formula is required precision and recall

Precision of Yes = 718/721 = 0.9958

Recall of Yes = 718/725 = 0.9903

Precision of No = 425/432 = 0.9838

Recall of No = 425/428 = 0.9929

So the f1(yes) is( 2\*(0.9958\*0.9903))/(0.9958+0.9903) = 0.9931

And f1(no) is ( 2\*(0.9838\*0.9929))/(0.9838+0.9929) = 0.9884

So the answer is D)

5.4 From the confusion matrix above, what is macro averaging F1?

A.0.99133

B.0.99073

C.0.99132

D.0.99072

Answer: B)

We have f1 score for each class so it’s simple to calculate macro averageing F1

As F1(yes) is 0.993084

And F1(no) is 0.988372

Macro F1 = (0.993084+0.988372)/2 = 0.990728 round as 0.990723

5.5 Which of the following measures should be considered first?

A.F1(Yes)

B.F1(No)

C.Accuracy

D.Macro average F1

Answer: A)

Actually i want to answer Recall since it’s covid data and covid is pandamic we need model to capture all covid patient as much as possible we don’t need precision that much since we can double check the patient again after model predict yes.

We can accept type I error (the patient model predict yes but actual is no) but we cant accept type II error (the patient model predict no but actual is yes) since as a pandemic it can cause much damage than type I error so i will consider recall of the model first

But on above choice i will consider F1(Yes) since It’s a way to combine precision and recall into a single number and patient with actual covid (Yes) is most consider in this case

6.1 Which is(are) NOT true about the assumption of Linear Regression below? (Select all that apply)



A.The relationship between Spending and Income should be linear.

B.The relationship between Spending and Age should be linear.

C.The relationship between Income and Age should be linear.

D.All variables should be normally distributed.

Answer: C and D is not true

1. A) From equation above a is true since spending up 1 when income up 5 so it will be linear
2. B) is true with the same reason as A) spending minus 10 when age up 1 so it will be linear
3. It not always true income and age relationship should be linear actually even it have no relationship between income and age it still be fine
4. Variable can be any distributed in linear regression model

6.2 Which is NOT true about regression techniques?

A.Each case must not contain missing values.

B.All inputs must be numeric.

C.Missing values in inputs and targets should be imputed.

D.There can be only one target.

Answer: D

A,B is cleary True

C) yes if you have missing targets you should delete it and since Each case must not contain missing values. So the missing inputs should be imputed (mean,mode imputation or delete )

D) There can be more than one target in regression it is called Multi target regression.

6.3 Which is NOT a method to alleviate a skewness issue?

A.Remove outliers

B.Take log

C.Binning

D.Min/max scaling

Answer: C)

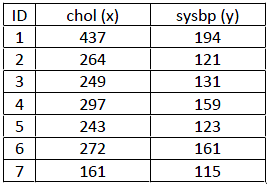
Binning is not a method to alleviate a skewness it will be the same

1. Remove outliers help data become more normal distributed since extream data is remove
2. Take log is a technique to normalization data with 1,000,000 and 10 will become closer so it help skewness issue

D) Min max scaling is another technique to normalization data so it will help skewness issue

So the answer is C)

6.4 From the input data below, compute w0 and w1 of Linear Regression to predict SystolicBlood Pressure (sysbp) from Choresterol (chol)



A.w0 = 57.8355, w1 = 0.5273

B.w0 = 57.8355, w1 = 0.3116

C.w0 = 0.5273, w1 = 57.8355

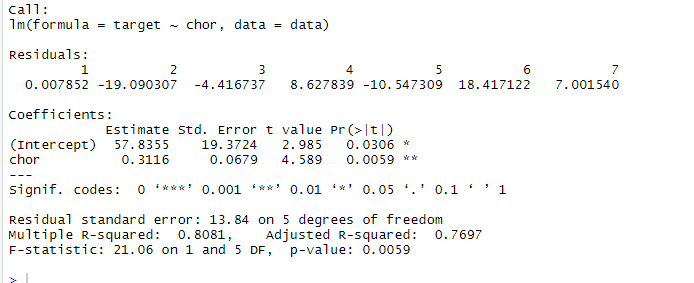
D.w0 = 0.3116, w 1 =57.8355

Answer: B)

For this answer we can compute it by create linear regression model and the result are as picture below so the equation of this model is sysbp = 57.8355 + 0.3116(chol)

So the w0 = 57.8355 and w1 = 0.3116

(see picture below)



6.5 From the above question, compute R2 and (RMSE) on the same data

A.R2 = 0.8052, RMSE = 11.7853

B.R2 = 0.8052, RMSE = 6.2891

C.R2 = 0.7213, RMSE = 11.7853

D.R2 = 0.7213, RMSE = 6.2891

Answer: A)

I think the choice is wrong but I choose A as it is close . We see above that R square is 0.8081

And if w0 = 57.8355, w1 = 0.3116 the predict value of the same data

Will be  and rmse will be 11.69765 not 11.7853 which is not in the choice but I choose A since it Close.

6.6 Which is NOT true about Linear Regression?

A.R2 is always higher in an equation with more variables (superset).

B.Adjusted R2 is always higher in an equation with more variables (superset).

C.Adjusted R2 is always smaller than R2 in the same equation, whose inputs are more than or equal to 2.

D.R2 can be calculated in both train and test data sets.

Answer: B

B is False it is not necessary of Adjusted R2 when we add more variables The adjusted R-squared compensates for the addition of variables and only increases if the new predictor enhances the model above what would be obtained by probability. Conversely, it will decrease when a predictor improves the model less than what is predicted by chance

6.7 Which is NOT true about a sequential selection in Linear Regression?

A.Forward, backward, and stepwise can never have the same result.

B.Stepwise is superior to forward and backward selections

C.Forward, backward, and stepwise have a regularization mechanism embedded.

D.All above

Answer: D)

1. Is not true in general, forward and backward stepwise regression can give you the same result
2. Is not true even though Stepwise can be superior than forward selection. Since it is modification of the forward selection so that after each step in which a variable was added, all candidate variables in the model are checked to see if their significance has been reduced below the specified tolerance level. If a nonsignificant variable is found, it is removed from the model. But they may be inferior to backward selection in some case
3. Is not True all of the above does not do anything with regualrization

7.1 Which task is NOT suitable for Logistic Regression?

A.Credit Scoring

B.House Price Prediction

C.Direct Target Marketing

D.Disease Classification

Answer: B) House Price Prediction

House Price Prediction is suit for linear regression or other advanced technique since logistic regression aim to predict probability to something it will not suit for House Price Prediction

7.2 The equation below is used for a loan approval prediction (bad (0) and good (1)). Use it to predict a customer with (x1, x2) = (0.3, 0.7).



A.0.243

B.0.56

C.Good

D.Bad

Answer: D) Bad

-0.81+0.276+0.777 = 0.243 . i think it will depend on threshold we choose but since we not know threshold i will use 0.5 as normal so 0.243 is classifiy as 0 and probability is bad

7.3 From the question above, what is a suitable value for ranking a customer?

A.0.243

B.0.56

C.Yes

D.Cannot answer

Answer: A) 0.243

Since we want to ranking customer from good to bad the probability value should be use

7.4 Which is NOT true about Odds Ratio (OR)?

A.OR(Age) = 1.75; For each increment of age by 1, the odds is increased by 75%.

B.For a variable with a negative weight, OR is always less than 0.

C.If OR is 1, it means that the variable is not important.

D.OR can be computed for both numeric and categorical variables.

Answer: B)

if the OR is less than 1, then A and B are negatively correlated, and the presence of one event reduces the odds of the other event. So it not necessary less than 0 with negative weight

8.1 Which is NOT true about k‐means clustering?

A.Each case cannot contain missing values.

B.All inputs must be numeric.

C.All inputs should have the same range.

D.The number of inputs should be large.

Answer: C)

1. Is true we must deal with missing values before use k-means clustering technique(remove null observation or just computed it with mean , mode etc)
2. Is true it a character of k-means consider dummy variable when we have categorical data
3. Not always true
4. Yeah it should be I thinks it is true but large is a word with no clear definition depend on the object of the task . But if we have small dataset (very small ie n=10) there will be a better way to group observation than k-means clustering so i think it is true

8.2 Which problem is suitable for DBSCAN?

A.Fraud detection

B.Employee profile segmentation

C.Student performance segmentation

D.Calling behavior segmentation

Answer: A.

I think it Calling behavior since DBSCAN technique suit for the data that is not distributed as a group and can be handle data with many outliers and noise make us can see which transactions is outliers so i think Fraud detection which we aim to detect anomaly detection and define as fraud can be use this technique

8.3 Which is NOT true about clustering?

A.The number of k in k‐means can be considered using elbow method.

B.The number of k in k‐means can be considered by applying hierarchical clustering.

C.Both k‐means and DBSCAN are needed user‐defined parameters.

D.The clustering techniques in sklearn do not have the prediction method since it is an unsupervised learning.

Answer: B.)

We don't use hierarchical clustering to define number of k

8.4 Please map the techniques below into these tasks: Association Rule Mining, Recommendation, Time Series Forecasting

A.A, B, C

B.C, B, A

C.B, A, C

D.C, A, B

Answer: B) C,B,A

Association rule is sure to be apriori

Collaborative filtering aim to defined user preference by user-based or content-based so it will be suit to use on recommendation purpose

Exponential smoothing is technique use for time series forecasting so it will be c

9.1 What is INCORRECT about ALTER TABLE?

A.Used to create a new row

B.Used to create a new column

C.Used to edit a column

D.Used to set the default value for a new column

E.None of the above

Answer: A)

Alter table can use for b c d but not a mostly when we want to add a new row we use INSERT INTO

9.2 Which command deletes every row in the table PRODUCTS?

A.DROP TABLE products;

B.DELETE TABLE products;

C.TRUNCATE TABLE products;

D.DELETE ALL ROWS TABLE products;

E.None of the above

Answer: C)

Truncate table use to delete all row , drop is delete table not delete row , delete is delete mostly use as delete from not delete table

9.3 Which of the below is INCORRECT regarding adding a new row to a table?

A.INSERT INTO adds a new row to a table

B.A new row cannot contain NULL values

C.We can add a row with all its values defined with a single command

D.While adding a new row, we can selectively define values for columns

E.We can import table A as rows into table B

Answer: B

B is clearly False we can add null values when create new rows such as   
INSERT INTO agents

VALUES ("A001","Jodi","London",.12,NULL);

9.4 SUBSTR(‘UTC AI ACADEMY’, 2, 8) would produce:

A.UTC AI A

B.TC AI A

C.TC AI AC

D.C AI AC

E.C AI ACA

Answer: C

We substring start with character 2 which is “T” and 8 length with is TC AI AC (space count as character ) so the answer is c

9.5 Which of the below is not correct regarding the command BEGIN TRAN TX1; ?

A.It previews the effects of a command without executing it

B.It guarantees the data consistency of the transaction

C.It bundles many commands as one

D.It needs to be used with COMMIT

E.It can only affect a single table

Answer: D)

It is not always used with COMMIT you use BEGIN TRAN with COMMIT when you're sure the effect of a command is like you want to do and give it a go. You can just use BEGIN TRAN only to preview the effect or gp ROLL BACK when it have something wrong

9.6 Which of the following is CORRECT regarding indexing?

A.Index can be made from many columns

B.Index provides data consistency at the cost of speed

C.As a default, an index is created automatically

D.Index should be used with small databases

E.One table can only contain one index

Answer: A)

Index can be made from many columns it is call Multiple-Column Indexes

9.7 To sum-aggregate the precipitation levels across all the weather stations, what command should we use?

A.select sum(nvl(precipitation, 0)) from weather\_station;

B.select sum(precipitation, 0) from weather\_station;

C.select nvl(sum(precipitation), 0) from weather\_station;

D.select nvl(precipitation, 0) from weather\_station;

E.None of the above

Answer: A)

A is the syntax fit most with this

B) will be an error. C) will not produce results as we want . and D) is not even sum

9.8 We want to query member’s first & last name, birthdate, and membership level for all those members born between April 1, 1990 and December 31, 1995. Which command should we use based on the following schema?

A.select first\_name, last name, date\_of\_birth, member\_level from members where date\_of\_birth between ’31-DEC-1995’ and ’01-APR-1990’;

B.select first\_name, last name, date\_of\_birth, member\_level from members where date\_of\_birth in( ’31-DEC-1995’, ’01-APR-1990’);

C.select first\_name, last name, date\_of\_birth, member\_level from members where date\_of\_birth like ’31-DEC-1995’ and ’01-APR-1990’;

D.select first\_name, last name, date\_of\_birth, member\_level from members where date\_of\_birth range( ’31-DEC-1995’, ’01-APR-1990’);

E.select first\_name, last name, date\_of\_birth, member\_level from members where date\_of\_birth between ’01-APR-1990’ and ’31-DEC-1995’;

Answer: E)

Answer E is correct since b and c we got only 2 days of query not between . D is error no function call range() in sql . A) is another error since later date is call before former date

9.9 We want to query all the member’s information, sorted by their points from highest to lowest. Which command should we use based on the following schema?



A.select first\_name, last name, current\_pointsl from members order by current\_points;

B.select first\_name, last name, current\_pointsl order by current\_points desc from members;

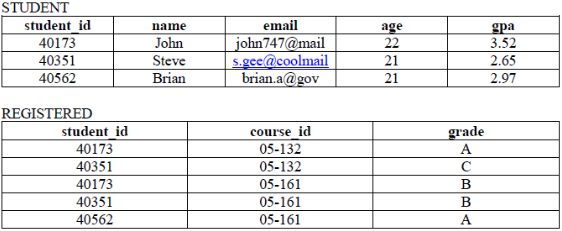
C.select first\_name, last name, current\_pointsl from members order by current\_points desc;

D.select first\_name, last name, current\_pointsl from members descending order by current\_points;

E.select first\_name, last name, current\_pointsl from members order by current\_points descending;

Answer: C)

Only ASC or DESC can type after order by no need to type descending that make D) and E) is wrong A) is not define DESC so it will order by ASC by default and B) is syntax error

Use the following schema to answer the questions 9.10 to 9.14

9.10 What is the expected result from the following command?



A.The result has one line

B.The result has two lines

C.The result has 3 lines

D.The result has 5 lines

E.There is no result

F.There is an error

Answer: B)

The result has two lines of ‘05-161’

9.11 What is the expected result from the following command?

A.The result has one line

B.The result has two lines

C.The result has 3 lines

D.The result has 5 lines

E.There is no result

F.There is an error

Answer: D)

Although the way he code is very strange he’s not call left join and query from 2 table at same times but it’s not error and have result as left join so it will become 5 lines

9.12 Which of the following command uniquely lists the IDs of all students who registered for more than a one subject?

A.SELECT DISTINCT r1.student\_id FROM registered AS r1, registered AS r2 WHERE r1.student\_id != r2.student\_id AND r1.course\_id != r2.course\_id

B.SELECT DISTINCT r1.student\_id FROM registered AS r1, registered AS r2 WHERE r1.student\_id = r2.student\_id AND r1.course\_id != r2.course\_id

C.SELECT DISTINCT r1.student\_id FROM registered AS r1, registered AS r2 WHERE r1.student\_id != r2.student\_id AND r1.course\_id = r2.course\_id

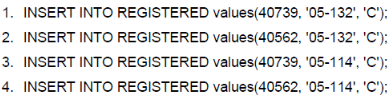
D.SELECT DISTINCT r1.student\_id FROM registered AS r1, registered AS r2 WHERE r1.student\_id = r2.student\_id AND r1.course\_id = r2.course\_id

E.None of the above

Answer: B)

The query B) command is appropriate way to detect student id with more than one course id .

9.13 Given that student\_id is the foreign key for STUDENT in the table REGISTERED, which of the following commands will successfully add a new row into REGISTERED?



A.1 and 2

B.1 and 3

C.1 and 4

D.2 and 3

E.2 and 4

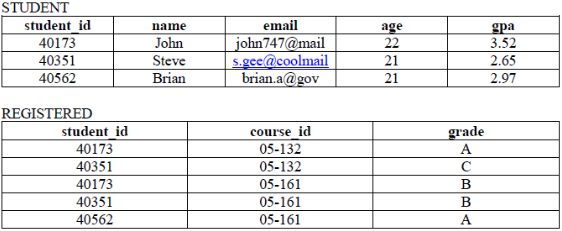
F.3 and 4

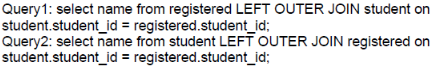
G.All of the above

Answer: G)

Every command is valid

9.14 Which of the following statement is TRUE regarding the following commands?





A.Query 1 and 2 produce the same results

B.Query 1 produces more rows than does Query 2

C.Query 2 produces more rows than does Query 1

D.Query 1 and 2 produce the same number of rows with different values

E.Query 1 runs successfully while Query 2 results in an error

F.Query 2 runs successfully while Query 1 results in an error

Answer: A)

This is tricky question but since join table is the same so the results from query name column will be john,steve,john,steve,brian bot Query 1 and Query 2

9.15 Which of the following is(are) INCORRECT regarding unstructured data? (Select all that apply)

A.It does not have a structure

B.It is easy to use

C.It comes in great volume and grows in a great rate

D.It is more useful than structured data

Answer: A),B),C),D)

1. Is wrong unstructured data has internal structure but is not structured via pre-defined data models or schema

B)It is not easy to use since it can be image ,video,sound,text so we must do some process before we can use compare to structure data

C) It not necessary of unstructured data come in great volume 1 image is unstructured data even it not grows in a great rate

D) It is not always more useful than structured data. It depend on your business purpose

9.16 Which of the following is(are) NOT unstructured data? (Select all that apply)

A.Comments in Facebook

B.Pictures in Instagram

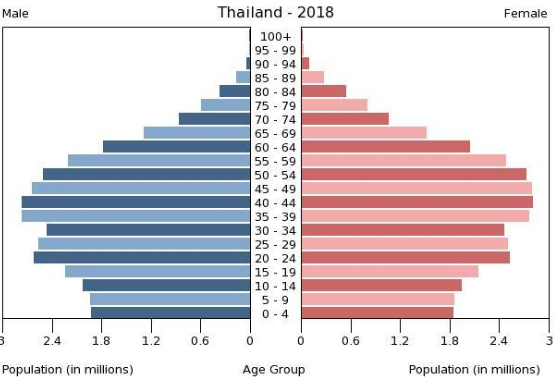
C.Precipitation levels in AccuWeather

D.Grades report in PDF

Answer: C,D

1. Is sure unstructured it is a text or maybe some sticker or feeling
2. Picture is unstructured
3. Precipitation levels can be store as structure in sql like this Bangkok , 2.00PM , 50 mm Precipitation
4. Display Data as PDF does not make unstructured data . It just make it cannot be edit

The following chart shows the population of Thailand statistics. Use this chart to answer the question 9.17



9.17 What are the number of measures and dimensions of this piece of data?

A.1 measure 1 dimension

B.1 measure 2 dimensions

C.1 measure 3 dimensions

D.1 measure 8 dimensions

E.2 measures 1 dimension

F.2 measures 2 dimensions

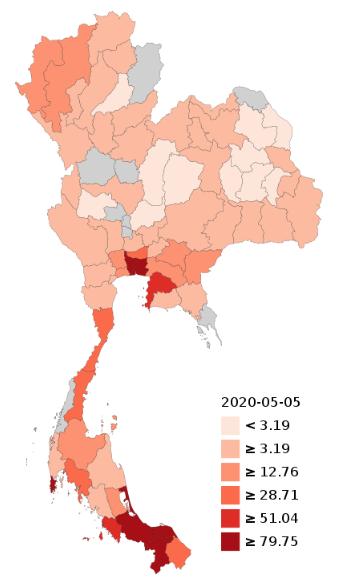
G.2 measures 3 dimensions

H.2 measures 8 dimensions

Answer: C)

I thinks the answer can be A),B),C) depend on what we think 1 measure for sure is age the and dimensions is gender for sure but another dimension can be country and years (if you think header is data to make this visualization cause this visualization define thailand as country and year is 2018) so I count it as dimension too cause I think header is a piece of data too so i answer C)

This figure shows the confirmed cases of COVID-19 per million residents in Thailand. Use this figure to answer the question 9.18.



9.18 From the above figure, what are the number of measures and dimensions of this piece of data?

A.1 measure 1 dimension

B.1 measure 2 dimensions

C.1 measure 3 dimensions

D.1 measure 8 dimensions

E.2 measures 1 dimension

F.2 measures 2 dimensions

G.2 measures 3 dimensions

H.2 measures 8 dimensions

Answer: B) 1 measure 2 dimension

The measure is COVID-19 per residents (divide by 1,000,000 to make per millons) and Dimension is Province and Dates which state by (2020-05-05)

9.19 What kind of chart is suitable for continuous-value dimension?

A.Simple Text

B.Line Chart

C.Bar Chart

D.Pie Chart

E.Heatmap

F.Scatter Plot

Answer: B) Line Chart

Example stock market index is continuous-value and mostly represent as line chart



9.20 What kind of chart is suitable for analyzing the relationship between two variables?

A.Simple Text

B.Line Chart

C.Bar Chart

D.Pie Chart

E.Heatmap

F.Scatter Plot

Answer: F) Scatter Plot

10.1 Graph (as a data structure) is most suitable to be represented with:

A.Circular Queue

B.1D Array

C.2D Array

D.None of the above

Answer: C) 2D Array

10.2 Which of the following data structures is(are) used to speed up search in large databases? (Select all that apply)

A.Graph

B.Skip List

C.Hash

D.Multiple Dimension Array

Answer: B,C

Hash is for sure speed up search . Skip list allow the process of item look up in efficient manner so it should be speed up

10.3 Can we implement the step “merge” in merge sort in an in-place manner?

A.No, Merge sort algorithm is not designed for in-place manner

B.No, there is no viable data structure to allow “merge” in-place

C.Yes, a skip-list can be used to implement the “merge” operation

D.Yes, an array can be used to implement the “merge” operation

Answer: D.

Yes it is call In-Place Merge Sort and input is array

11.1 Which of the following are TRUE? (Select all that apply)

A.PCA is invariant to affine transform

B.The Canny edge detection is a linear filter

C.Images in a pinhole camera are a parallel projection

D.PCA can be used for lossy image compression

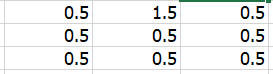
E.We can blur the image using linear filter

Answer: D),E)

D) Is true PCA allows us to represent high-dimensional data in a lower dimensional form

1. Is true technique such as box blur or gaussian blur is linear filter

11.2 Give a 3x3 convolution filter that shifts an image 1 pixel upwards and also reduces the image brightness by 50% (Answer format: [? ? ?; ? ? ?; ? ? ?])

Answer: 

(I answer with picture cause i don't know in the format which is row and which is column)

11.3 We can smooth an image by convolving the 1D vector [1/4, 1/2, 1/4] with the rows and then the columns of the image. Give a 2D matrix that combines these row and column operations into a single 3x3 convolution filter. (Answer format: [?, ?, ?; ?, ?, ?; ?, ?, ?])

Answer:

11.4 Which of the following are properties that cause an edge in an image? (Select all that apply)

A.A depth discontinuity

B.A surface discontinuity

C.A light source discontinuity

D.Illumination boundaries

Answer: A,B,D

Light source discontinuity is not propeties of image edge the others is yes

11.5 What is the role of pooling layers in a CNN?

Answer: pooling layers in cnn is building block to progressively reduce the spatial size of the representation to reduce the amount of parameters and computation in the network.

11.6 What is data augmentation? Why do we need it?

Answer: data augmentation is a technique that can be used to artificially expand the size of a training dataset by creating modified versions of images in the dataset. techniques such as cropping, padding, and horizontal flipping are commonly use

We need it when we want to increase the diversity of data available for training models, without actually collecting new data. May be the cost for collecting new data is so high maybe your have only small data