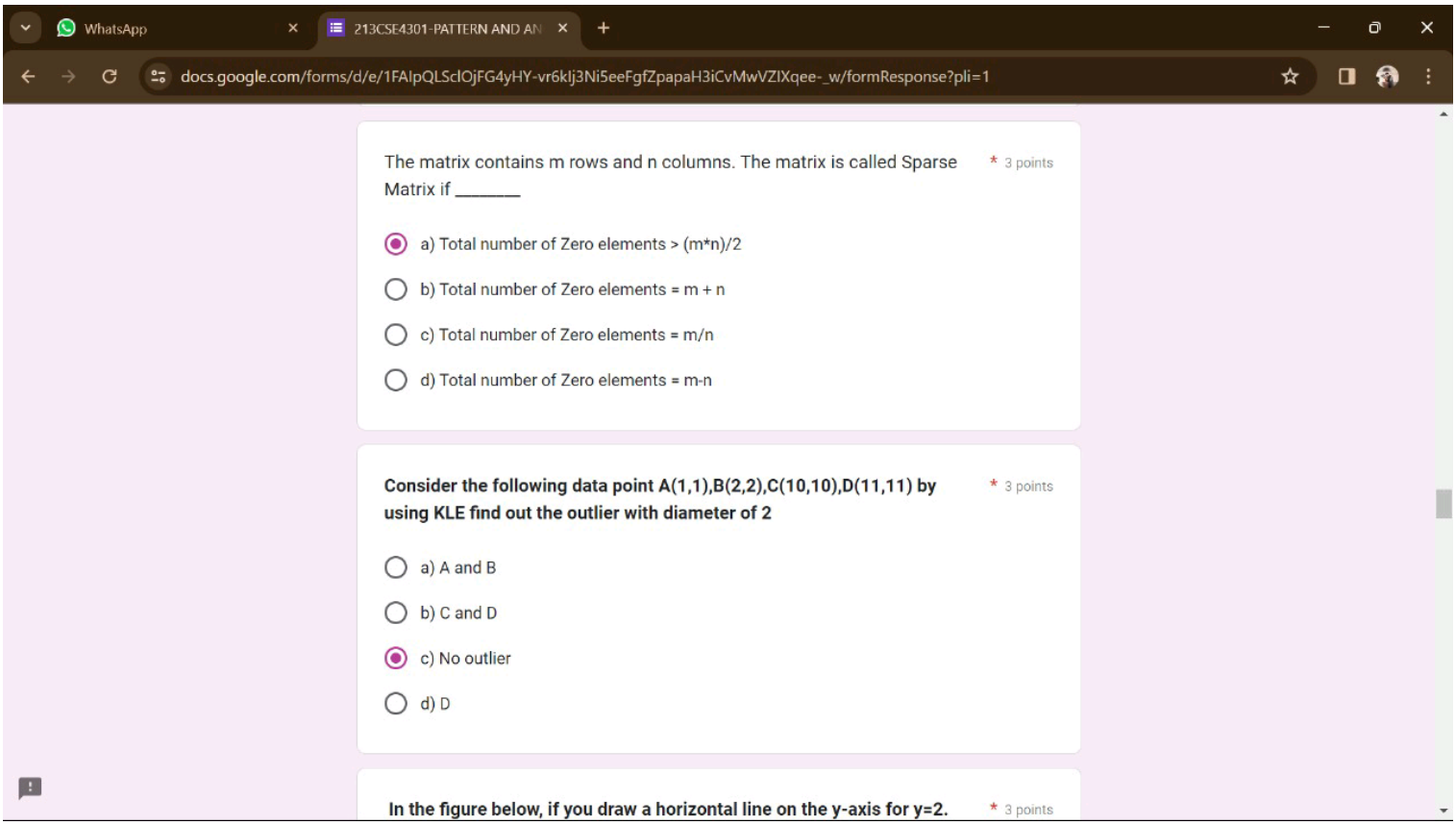
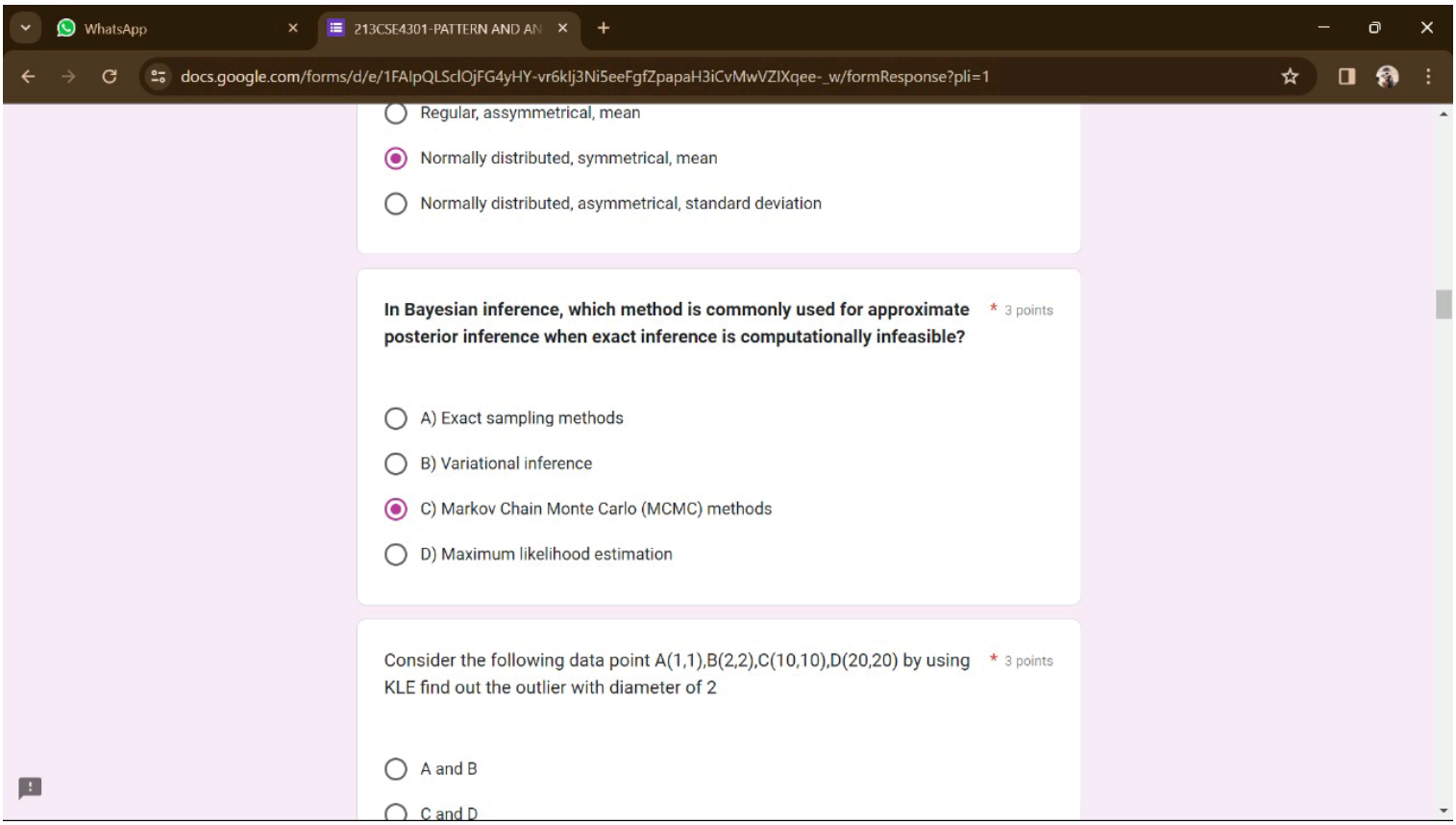
Answer all the following questions:

Which of the following can act as possible termination conditions in K-Means? * 3 points

1.For a fixed number of iterations.
2.**The assignment of observations to clusters does not change between iterations, except for cases with a bad local minimum.**
3.**Centroids do not change between successive iterations.**
4. **Terminate when RSS falls below a threshold.**

- A. 1, 3 and 4
- B. 1, 2 and 3
- C. 1, 2 and 4
- D. All of the above


WhatsApp
213CSE4301-PATTERN AND AN
docs.google.com/forms/d/e/1FAIpQLSclOjFG4yHY-vr6klj3Ni5eeFgfZpapaH3iCvMwVZIXqee-_w/formResponse?pli=1
The matrix contains m rows and n columns. The matrix is called Sparse Matrix if _____
* 3 points
a) Total number of Zero elements > (m*n)/2
b) Total number of Zero elements = m + n
c) Total number of Zero elements = m/n
d) Total number of Zero elements = m-n
Consider the following data point A(1,1),B(2,2),C(10,10),D(11,11) by using KLE find out the outlier with diameter of 2
* 3 points
a) A and B
b) C and D
c) No outlier
d) D
In the figure below, if you draw a horizontal line on the y-axis for y=2.
* 3 points


WhatsApp
213CSE4301-PATTERN AND AN
docs.google.com/forms/d/e/1FAIpQLSclOjFG4yHY-vr6klj3Ni5eeFgfZpapaH3iCvMwVZIXqee-_w/formResponse?pli=1
Regular, assymmetrical, mean
Normally distributed, symmetrical, mean
Normally distributed, asymmetrical, standard deviation
In Bayesian inference, which method is commonly used for approximate posterior inference when exact inference is computationally infeasible? * 3 points
A) Exact sampling methods
B) Variational inference
C) Markov Chain Monte Carlo (MCMC) methods
D) Maximum likelihood estimation
Consider the following data point A(1,1),B(2,2),C(10,10),D(20,20) by using KLE find out the outlier with diameter of 2 * 3 points
A and B
C and D

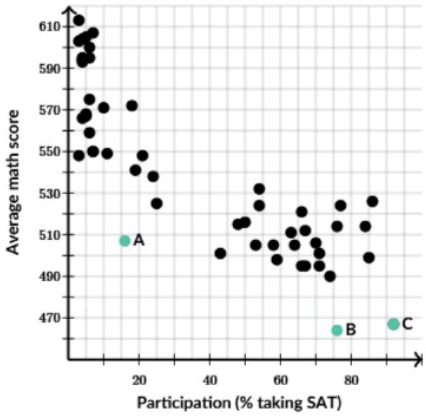
WhatsApp | 213CSE4301-PATTERN AND AN

docs.google.com/forms/d/e/1FAIpQLSclOjFG4yHY-vr6klj3Ni5eeFgfZpapaH3iCvMwVZIXqee-_w/formResponse?pli=1

Some high school students in the U.S. take a test called the SAT before applying to colleges. The scatter plot to the right shows what percent of each state's college-bound graduates took the SAT in along with that state's average score on the math section. The three labeled points could be considered outliers.

* 4 points

Why might these points be considered outliers?



- (●) 1. These states scored lower than other states with similar participation rates.
- ( ) 2. These states scored higher than other states with similar participation rates.

d) Support vector machine

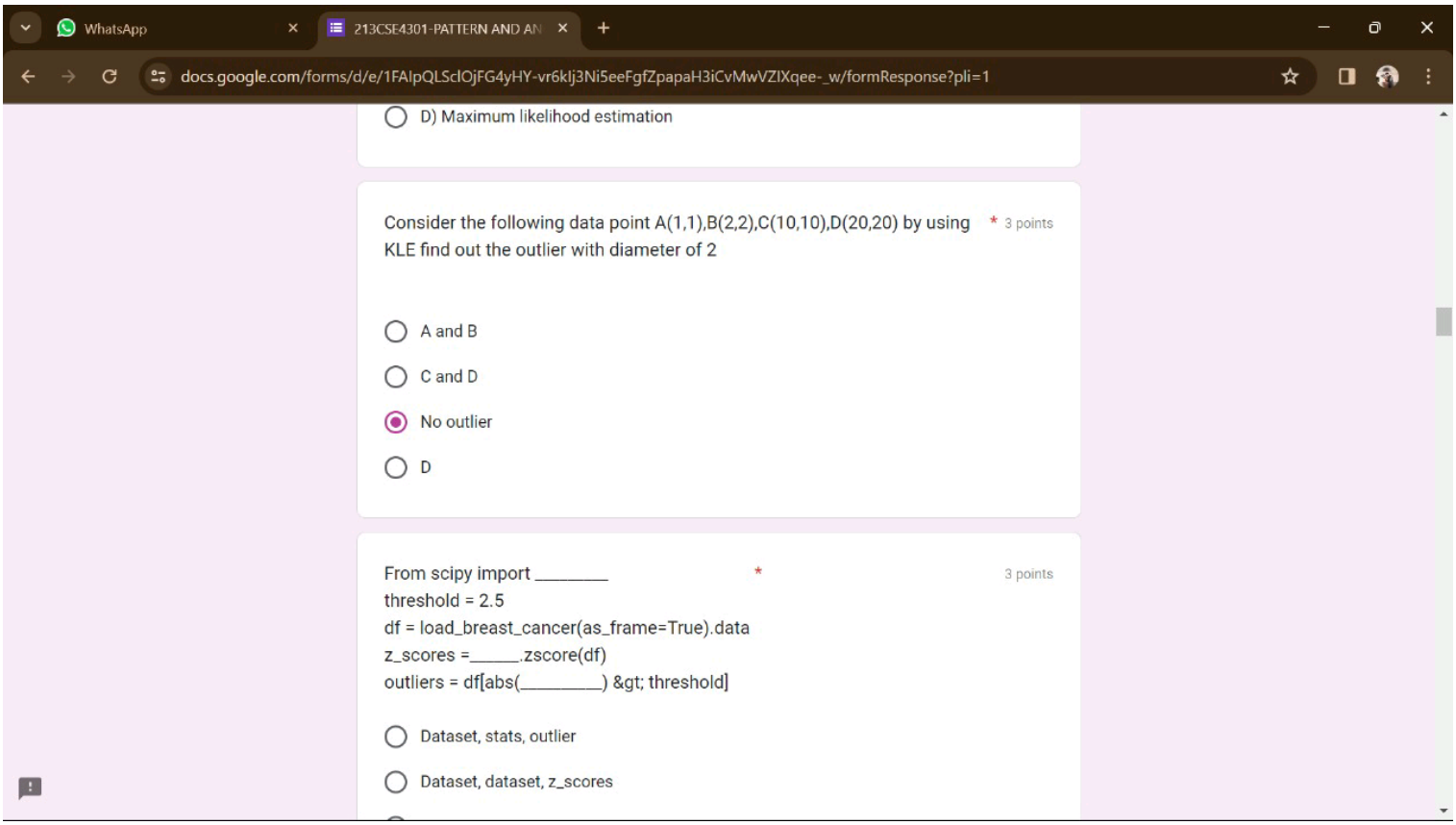
Which of the following values states outliers as per sklearn.clusters.DBSCAN * * 3 points

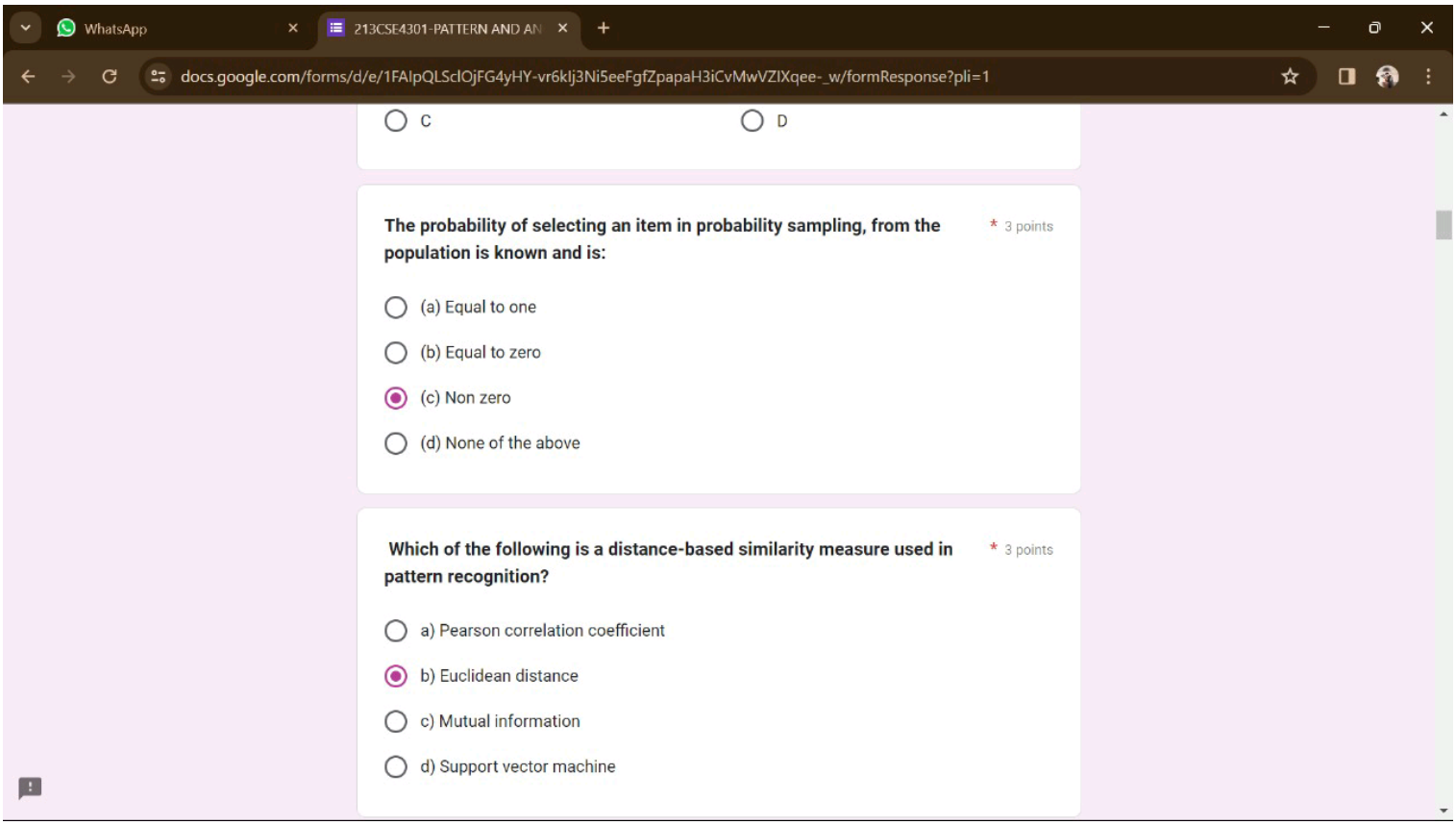
- a) -3
- b) -2
- c) -1
- d) Values >3

Following are the results observed for clustering 6000 data points into 3 clusters: A, B, and C: * 5 points

What is the F1-Score with respect to cluster B?

		Actual			
		A	B	C	SUM
Predicted	A	600	400	200	1200
	B	1000	1200	200	2400

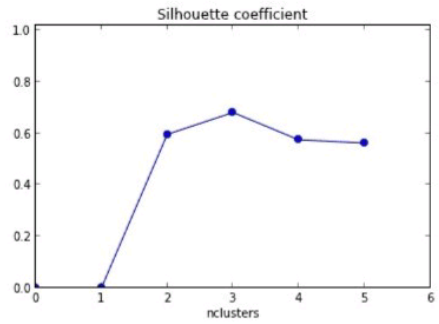

WhatsApp
213CSE4301-PATTERN AND AN
docs.google.com/forms/d/e/1FAIpQLSclOjFG4yHY-vr6klj3Ni5eeFgfZpapaH3iCvMwVZIXqee-_w/formResponse?pli=1
D) Maximum likelihood estimation
Consider the following data point A(1,1),B(2,2),C(10,10),D(20,20) by using KLE find out the outlier with diameter of 2
* 3 points
A and B
C and D
No outlier
D
From scipy import _____
threshold = 2.5
df = load_breast_cancer(as_frame=True).data
z_scores =_____.zscore(df)
outliers = df[abs(_____) > threshold]
* 3 points
Dataset, stats, outlier
Dataset, dataset, z_scores


WhatsApp
213CSE4301-PATTERN AND AN
docs.google.com/forms/d/e/1FAIpQLSclOjFG4yHY-vr6klj3Ni5eeFgfZpapaH3iCvMwVZIXqee-_w/formResponse?pli=1
C
D
The probability of selecting an item in probability sampling, from the population is known and is:
* 3 points
(a) Equal to one
(b) Equal to zero
(c) Non zero
(d) None of the above
Which of the following is a distance-based similarity measure used in pattern recognition?
* 3 points
a) Pearson correlation coefficient
b) Euclidean distance
c) Mutual information
d) Support vector machine

(1) WhatsApp | 213CSE4301-PATTERN AND AN

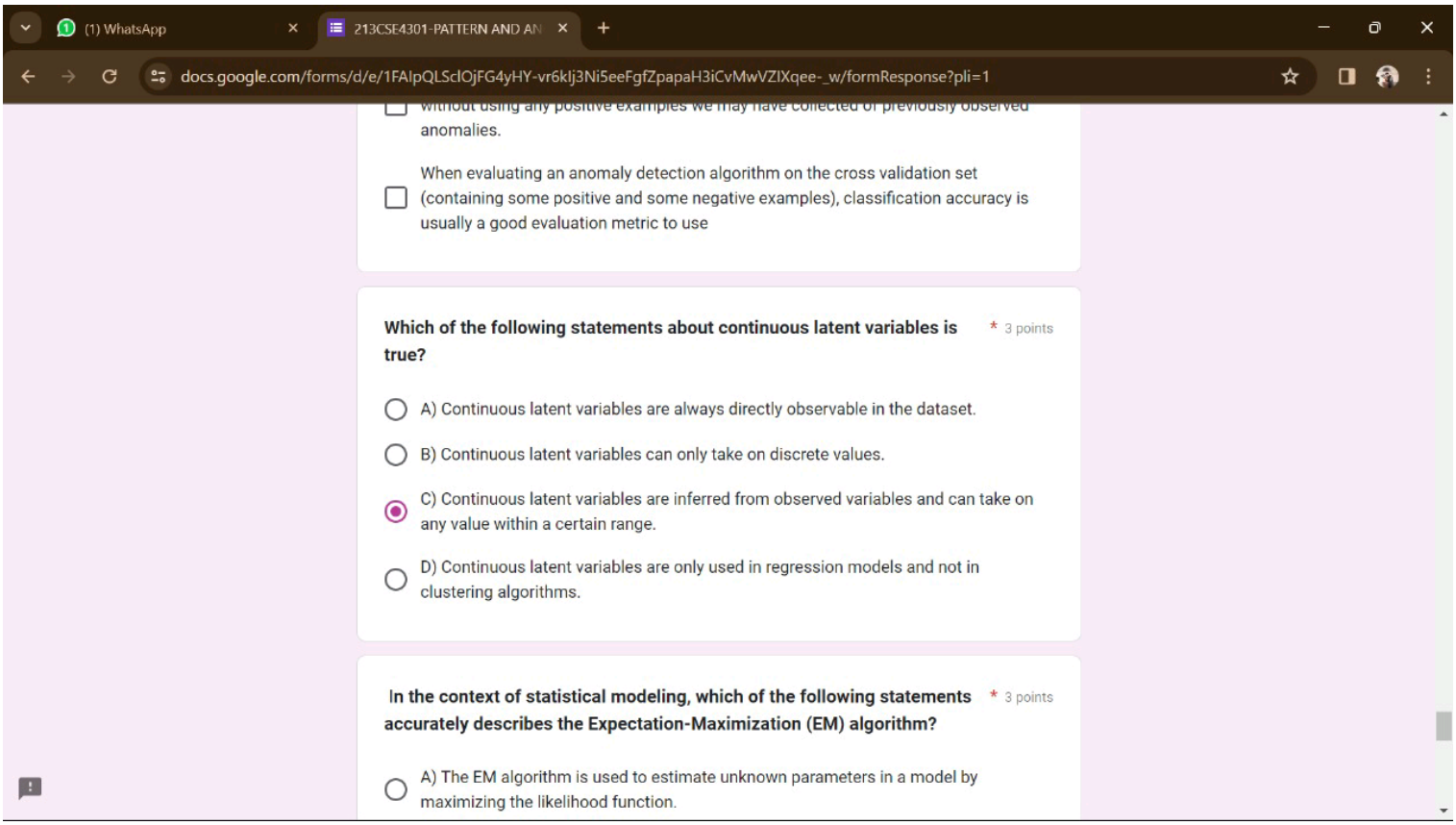
docs.google.com/forms/d/e/1FAIpQLSclOjFG4yHY-vr6klj3Ni5eeFgfZpapaH3iCvMwVZIXqee-_w/formResponse?pli=1

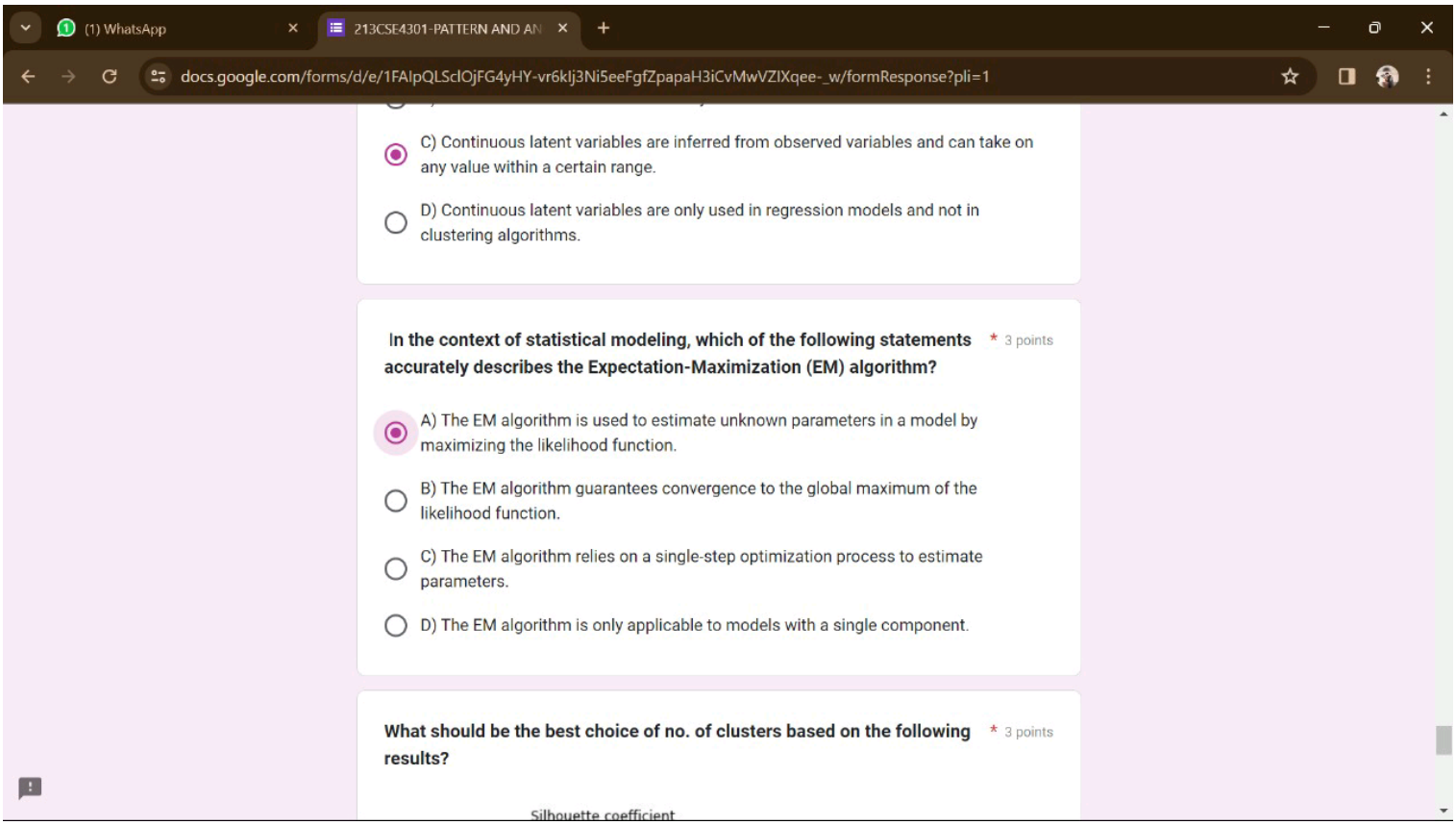
What should be the best choice of no. of clusters based on the following results? * 3 points

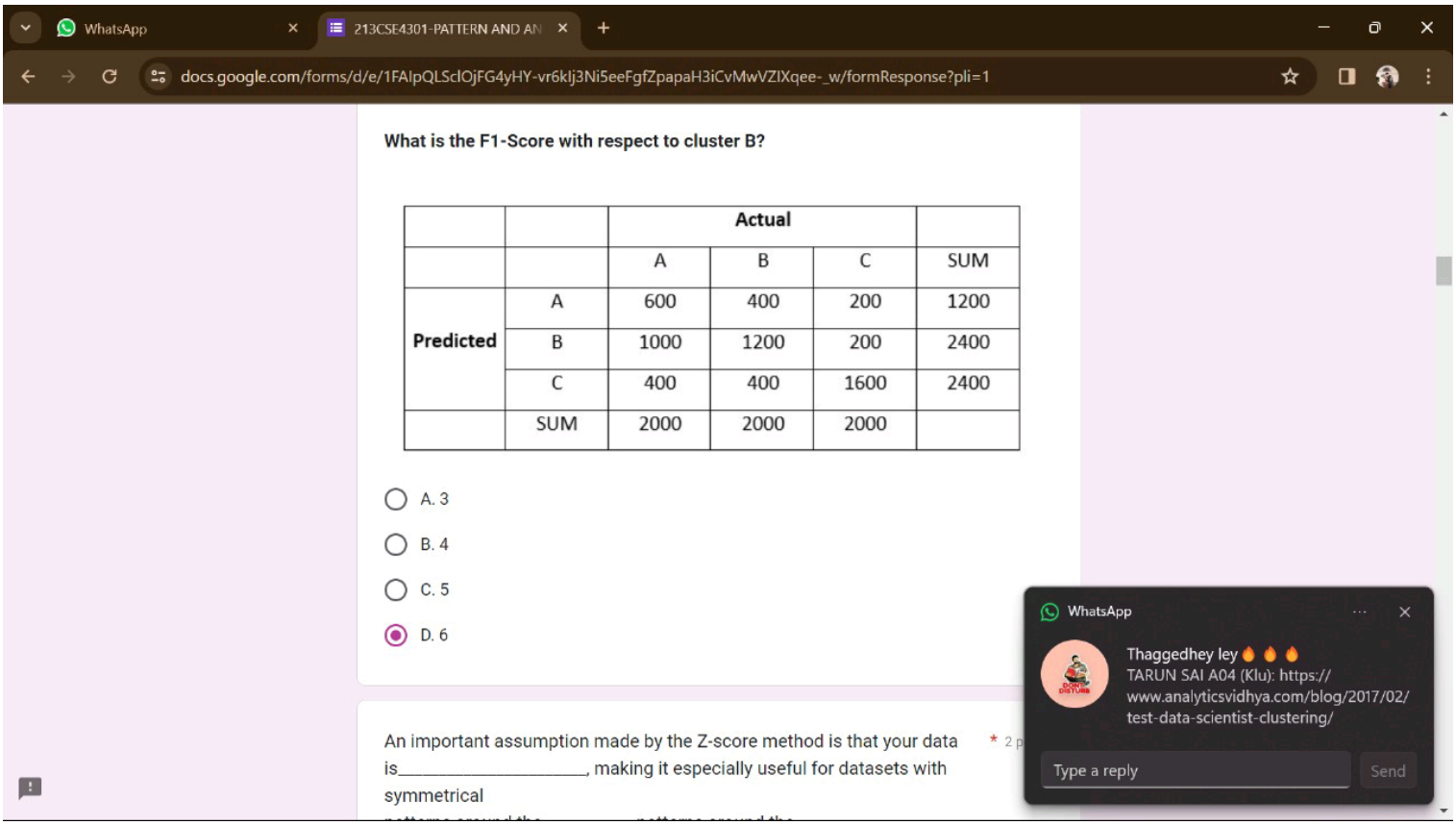


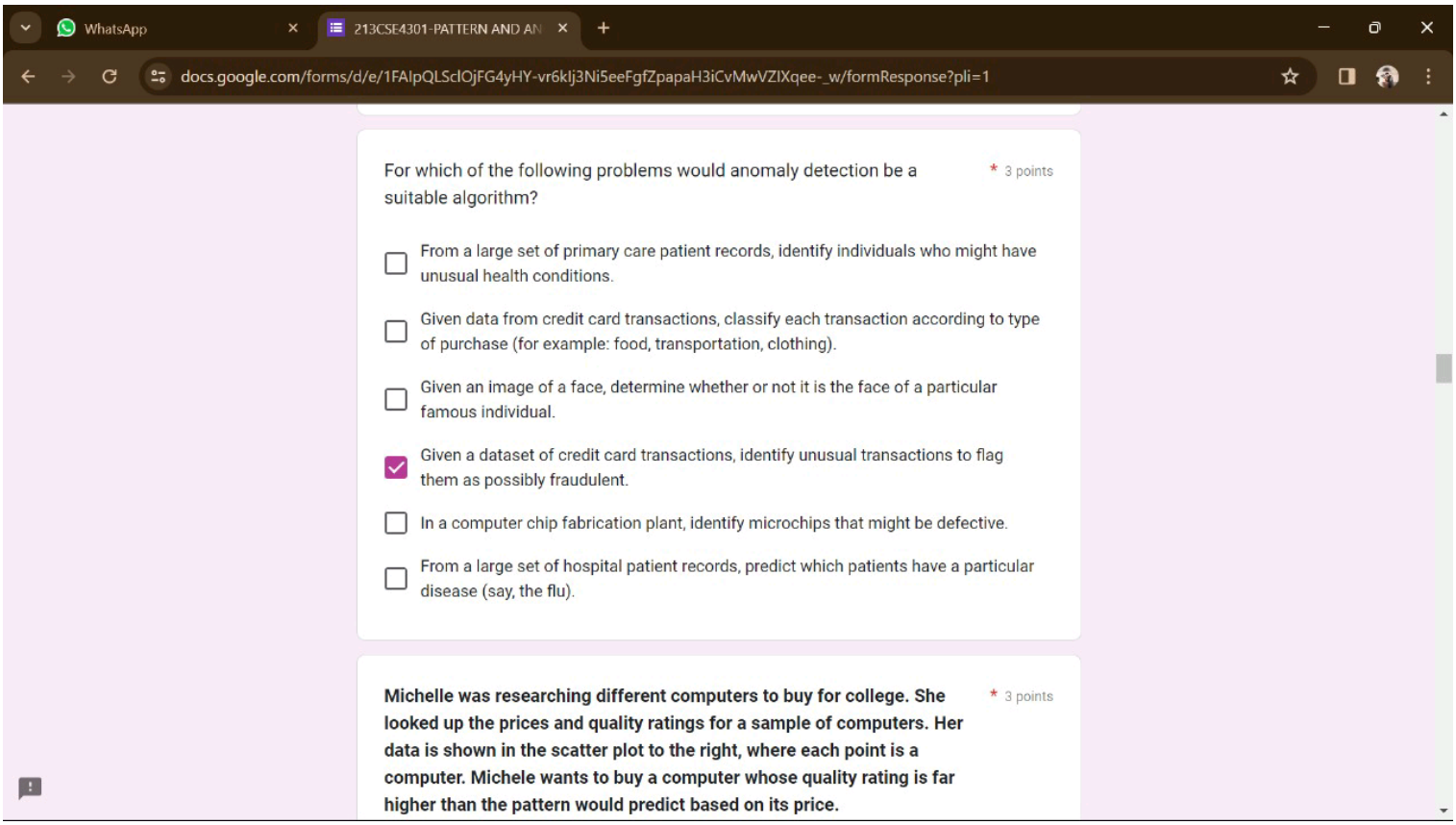
- A. 1
- B. 2
- C. 3
- D. 4

What are the outcomes of DBSCAN techniques? * 3 points


(1) WhatsApp
213CSE4301-PATTERN AND AN
docs.google.com/forms/d/e/1FAIpQLSclOjFG4yHY-vr6klj3Ni5eeFgfZpapaH3iCvMwVZIXqee-_w/formResponse?pli=1
without using any positive examples we may have collected of previously observed anomalies.
When evaluating an anomaly detection algorithm on the cross validation set (containing some positive and some negative examples), classification accuracy is usually a good evaluation metric to use
Which of the following statements about continuous latent variables is true? * 3 points
A) Continuous latent variables are always directly observable in the dataset.
B) Continuous latent variables can only take on discrete values.
C) Continuous latent variables are inferred from observed variables and can take on any value within a certain range.
D) Continuous latent variables are only used in regression models and not in clustering algorithms.
In the context of statistical modeling, which of the following statements accurately describes the Expectation-Maximization (EM) algorithm? * 3 points
A) The EM algorithm is used to estimate unknown parameters in a model by maximizing the likelihood function.

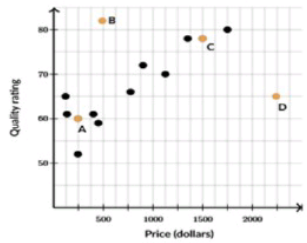

(1) WhatsApp
213CSE4301-PATTERN AND AN
docs.google.com/forms/d/e/1FAIpQLSclOjFG4yHY-vr6klj3Ni5eeFgfZpapaH3iCvMwVZIXqee-_w/formResponse?pli=1
C) Continuous latent variables are inferred from observed variables and can take on any value within a certain range.
D) Continuous latent variables are only used in regression models and not in clustering algorithms.
In the context of statistical modeling, which of the following statements accurately describes the Expectation-Maximization (EM) algorithm? * 3 points
A) The EM algorithm is used to estimate unknown parameters in a model by maximizing the likelihood function.
B) The EM algorithm guarantees convergence to the global maximum of the likelihood function.
C) The EM algorithm relies on a single-step optimization process to estimate parameters.
D) The EM algorithm is only applicable to models with a single component.
What should be the best choice of no. of clusters based on the following results? * 3 points
Silhouette coefficient


WhatsApp
213CSE4301-PATTERN AND AN
docs.google.com/forms/d/e/1FAIpQLSclOjFG4yHY-vr6klj3Ni5eeFgfZpapaH3iCvMwVZIXqee-_w/formResponse?pli=1
What is the F1-Score with respect to cluster B?
Actual
A
B
C
SUM
Predicted
A
600
400
200
1200
B
1000
1200
200
2400
C
400
400
1600
2400
SUM
2000
2000
2000
A. 3
B. 4
C. 5
D. 6
WhatsApp
Thaggedhey ley
TARUN SAI A04 (Klu): https://www.analyticsvidhya.com/blog/2017/02/test-data-scientist-clustering/
Type a reply
Send
An important assumption made by the Z-score method is that your data is_____, making it especially useful for datasets with symmetrical

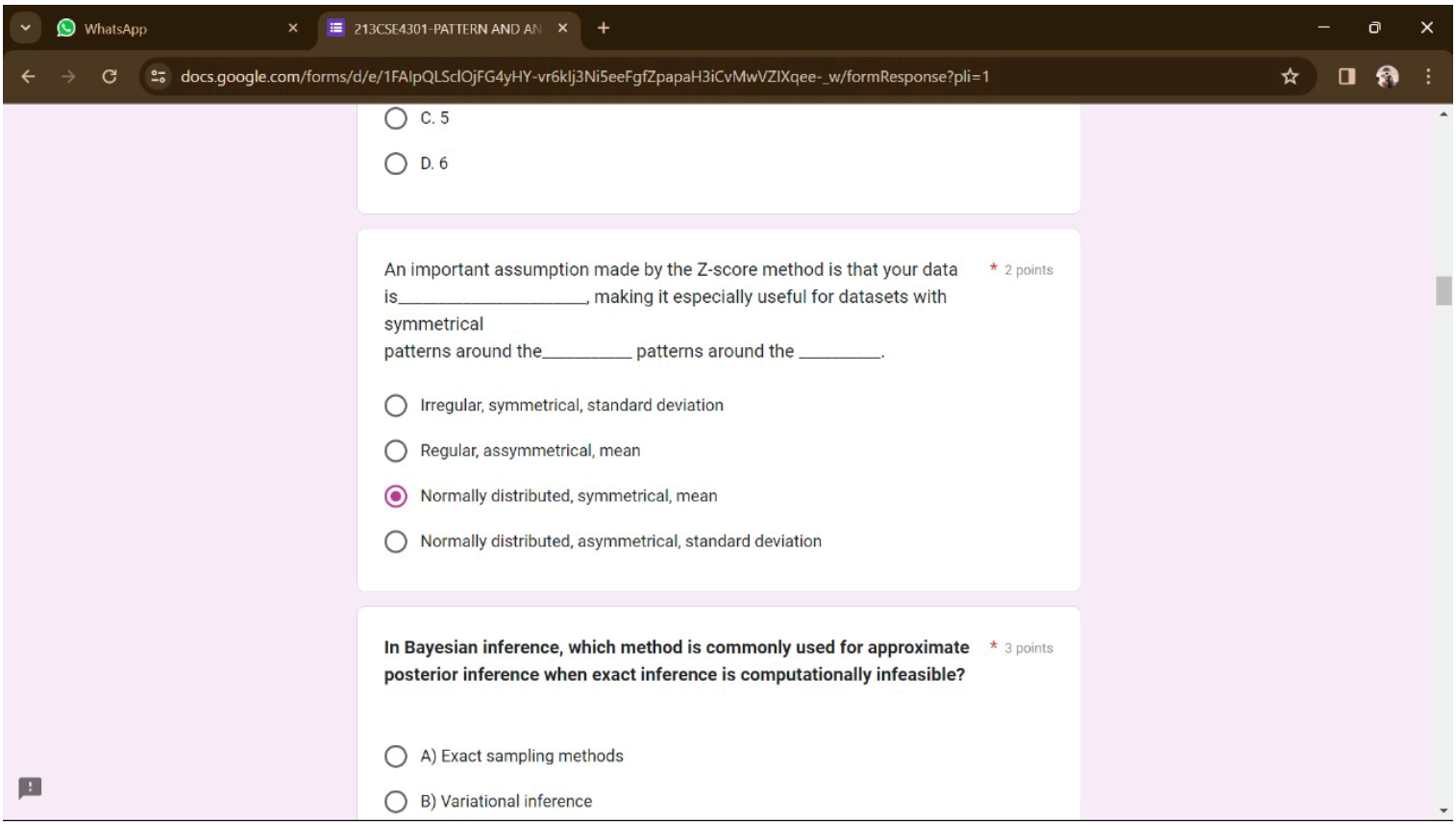

WhatsApp
213CSE4301-PATTERN AND AN
docs.google.com/forms/d/e/1FAIpQLSclOjFG4yHY-vr6klj3Ni5eeFgfZpapaH3iCvMwVZIXqee-_w/formResponse?pli=1
For which of the following problems would anomaly detection be a suitable algorithm?
* 3 points
From a large set of primary care patient records, identify individuals who might have unusual health conditions.
Given data from credit card transactions, classify each transaction according to type of purchase (for example: food, transportation, clothing).
Given an image of a face, determine whether or not it is the face of a particular famous individual.
Given a dataset of credit card transactions, identify unusual transactions to flag them as possibly fraudulent.
In a computer chip fabrication plant, identify microchips that might be defective.
From a large set of hospital patient records, predict which patients have a particular disease (say, the flu).
Michelle was researching different computers to buy for college. She looked up the prices and quality ratings for a sample of computers. Her data is shown in the scatter plot to the right, where each point is a computer. Michele wants to buy a computer whose quality rating is far higher than the pattern would predict based on its price.
* 3 points

Michelle was researching different computers to buy for college. She looked up the prices and quality ratings for a sample of computers. Her data is shown in the scatter plot to the right, where each point is a computer. Michele wants to buy a computer whose quality rating is far higher than the pattern would predict based on its price.
Which of the labeled points represents a computer that Michele wants to buy?

* 3 points

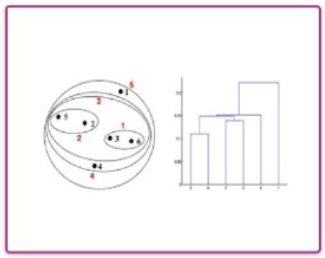


- a. D
- b. C
- c. B
- d. A

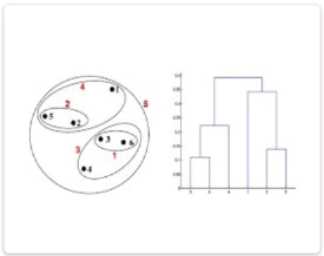

WhatsApp
213CSE4301-PATTERN AND AN
docs.google.com/forms/d/e/1FAIpQLSclOjFG4yHY-vr6klj3Ni5eeFgfZpapaH3iCvMwVZIXqee-_w/formResponse?pli=1
C. 5
D. 6
An important assumption made by the Z-score method is that your data is_____, making it especially useful for datasets with symmetrical
patterns around the_____ patterns around the _____.
* 2 points
Irregular, symmetrical, standard deviation
Regular, assymmetrical, mean
Normally distributed, symmetrical, mean
Normally distributed, asymmetrical, standard deviation
In Bayesian inference, which method is commonly used for approximate posterior inference when exact inference is computationally infeasible?
* 3 points
A) Exact sampling methods
B) Variational inference

p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

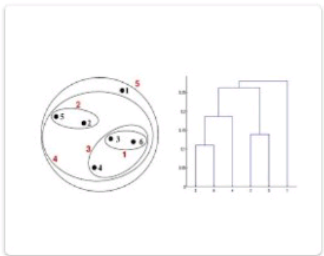
Table : Distance Matrix for Six Points

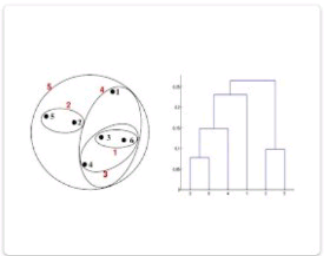


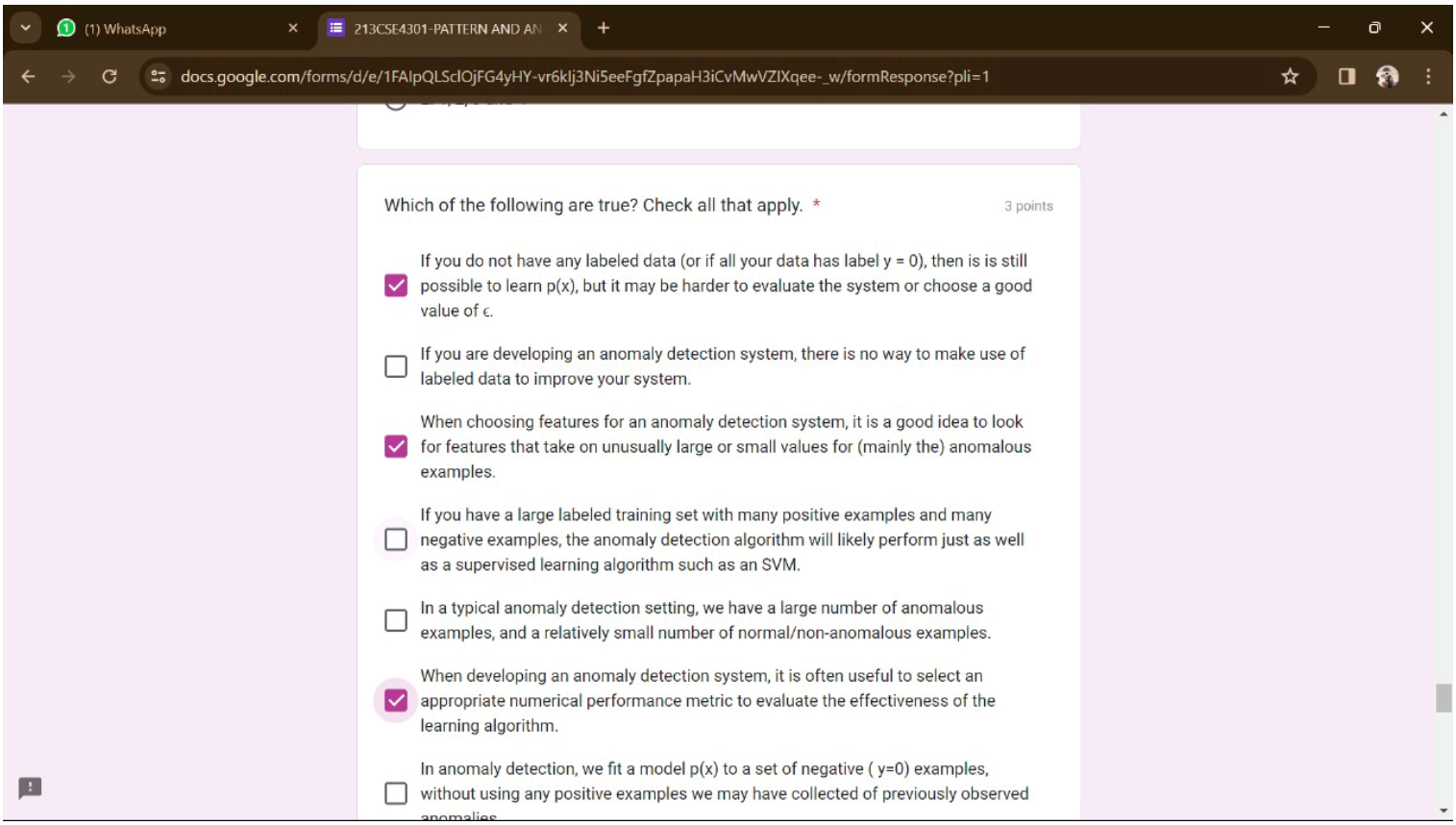
A

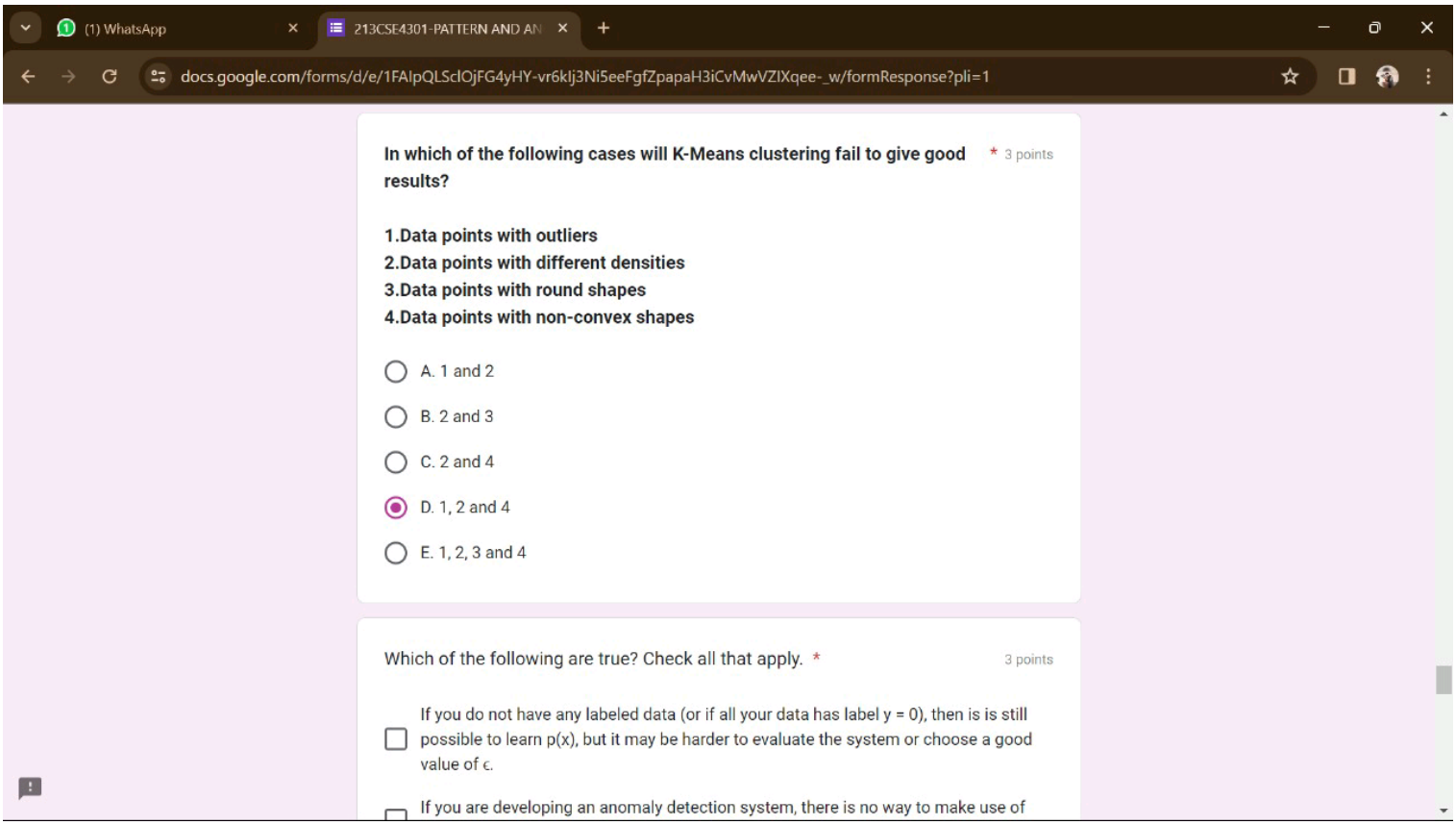


B





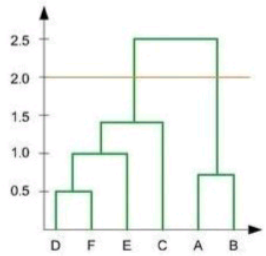

(1) WhatsApp
213CSE4301-PATTERN AND AN
docs.google.com/forms/d/e/1FAIpQLSclOjFG4yHY-vr6klj3Ni5eeFgfZpapaH3iCvMwVZIXqee-_w/formResponse?pli=1
Which of the following are true? Check all that apply. *
3 points
If you do not have any labeled data (or if all your data has label y = 0), then is is still possible to learn p(x), but it may be harder to evaluate the system or choose a good value of ϵ.
If you are developing an anomaly detection system, there is no way to make use of labeled data to improve your system.
When choosing features for an anomaly detection system, it is a good idea to look for features that take on unusually large or small values for (mainly the) anomalous examples.
If you have a large labeled training set with many positive examples and many negative examples, the anomaly detection algorithm will likely perform just as well as a supervised learning algorithm such as an SVM.
In a typical anomaly detection setting, we have a large number of anomalous examples, and a relatively small number of normal/non-anomalous examples.
When developing an anomaly detection system, it is often useful to select an appropriate numerical performance metric to evaluate the effectiveness of the learning algorithm.
In anomaly detection, we fit a model p(x) to a set of negative (y=0) examples, without using any positive examples we may have collected of previously observed


(1) WhatsApp
213CSE4301-PATTERN AND AN
docs.google.com/forms/d/e/1FAIpQLSclOjFG4yHY-vr6klj3Ni5eeFgfZpapaH3iCvMwVZIXqee-_w/formResponse?pli=1
In which of the following cases will K-Means clustering fail to give good results? * 3 points
1.Data points with outliers
2.Data points with different densities
3.Data points with round shapes
4.Data points with non-convex shapes
A. 1 and 2
B. 2 and 3
C. 2 and 4
D. 1, 2 and 4
E. 1, 2, 3 and 4
Which of the following are true? Check all that apply. * 3 points
If you do not have any labeled data (or if all your data has label y = 0), then is is still possible to learn p(x), but it may be harder to evaluate the system or choose a good value of ϵ.
If you are developing an anomaly detection system, there is no way to make use of

WhatsApp | 213CSE4301-PATTERN AND AN

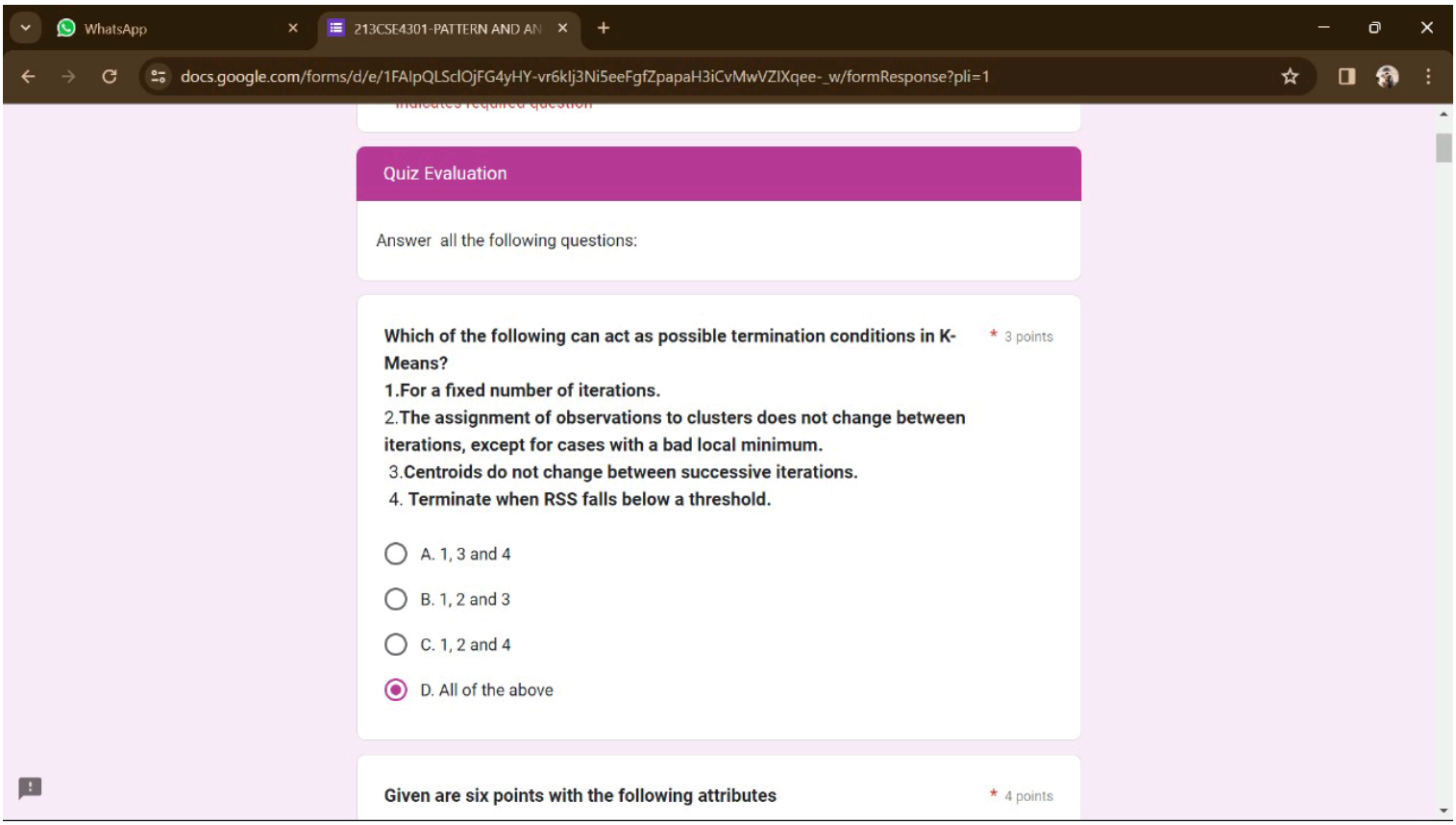
docs.google.com/forms/d/e/1FAIpQLSclOjFG4yHY-vr6klj3Ni5eeFgfZpapaH3iCvMwVZIXqee-_w/formResponse?pli=1

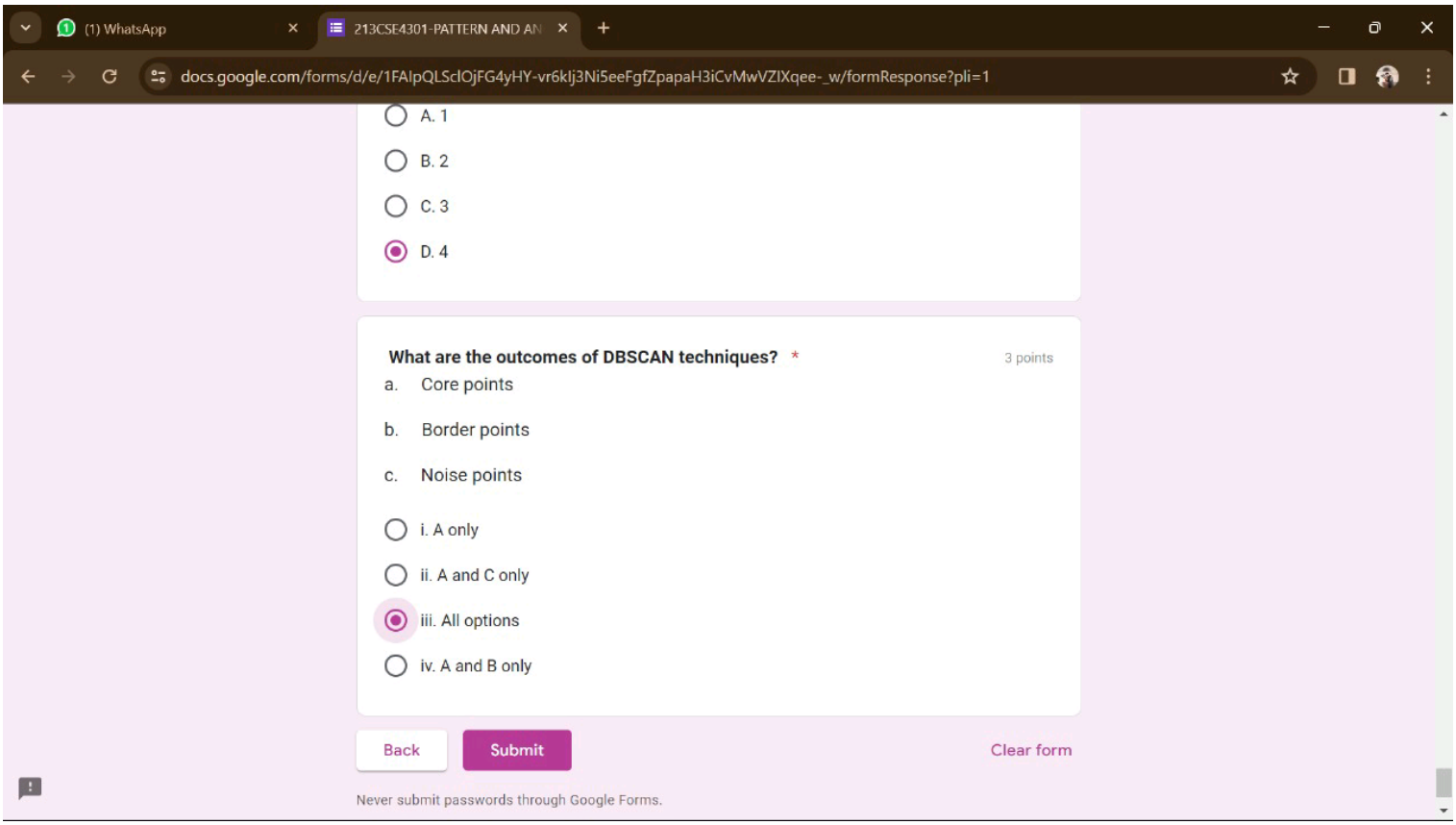
In the figure below, if you draw a horizontal line on the y-axis for y=2. What will be the number of clusters formed? * 3 points

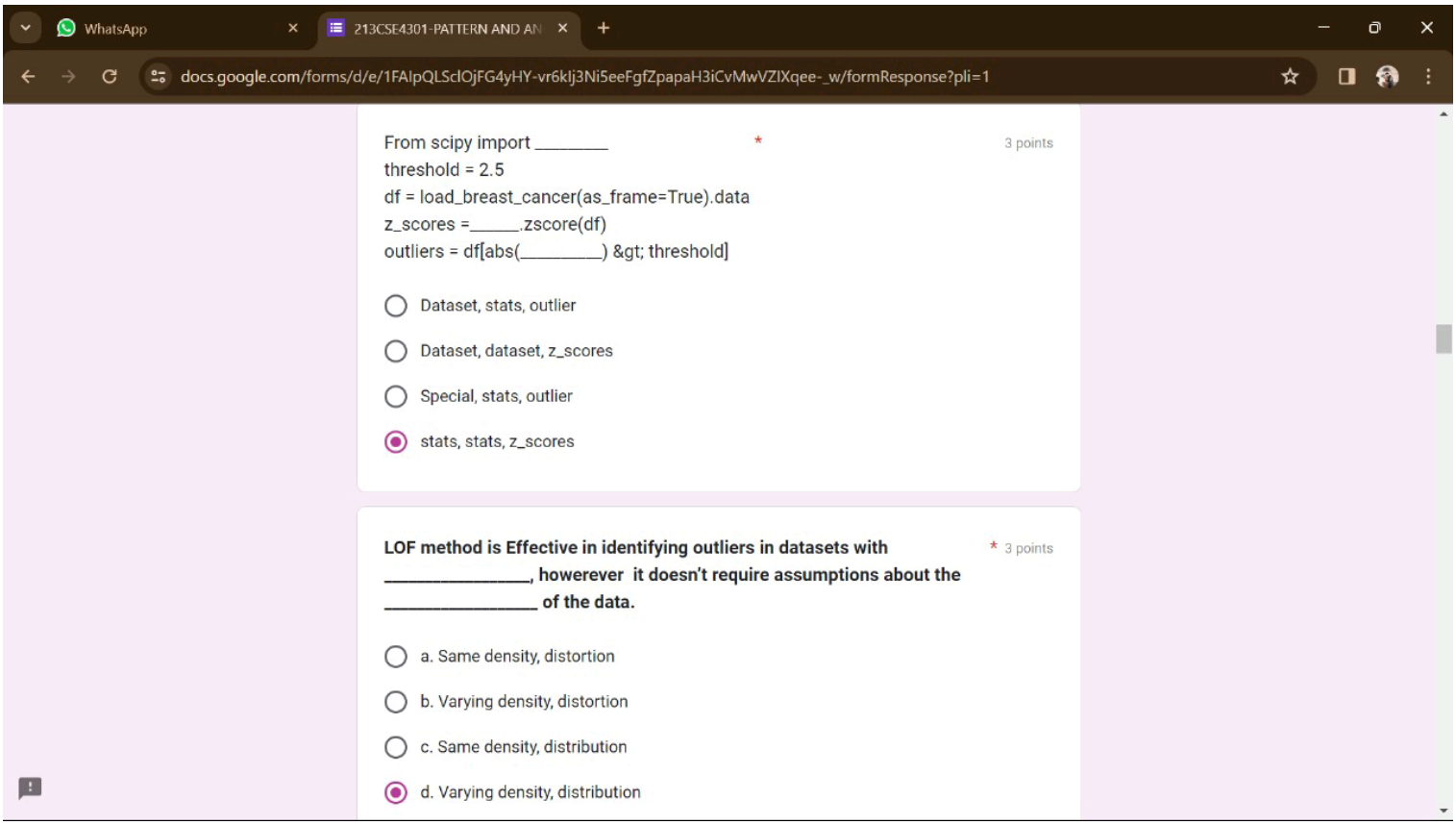


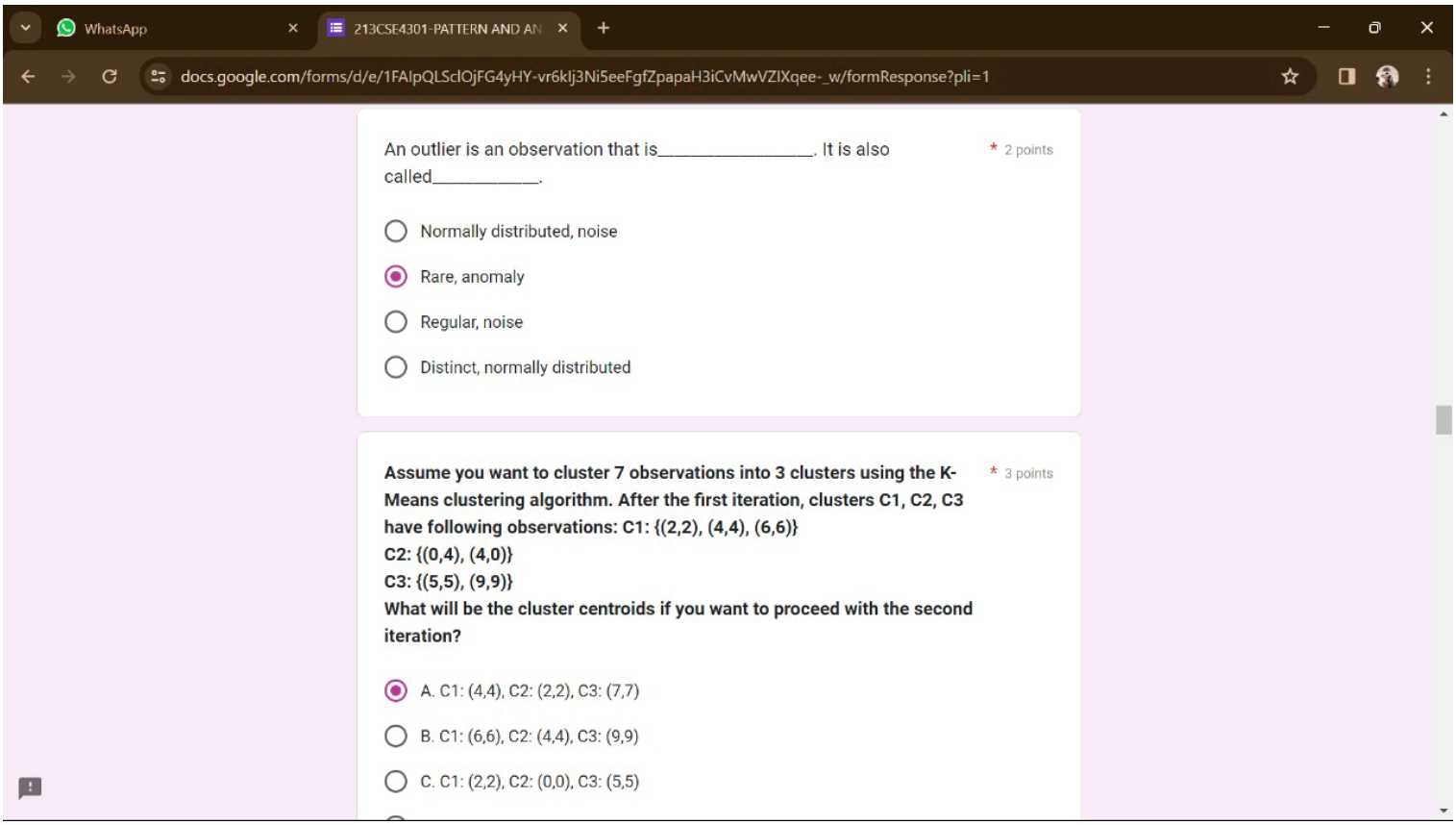
- A. 1
- B. 2
- C. 3
- D. 4

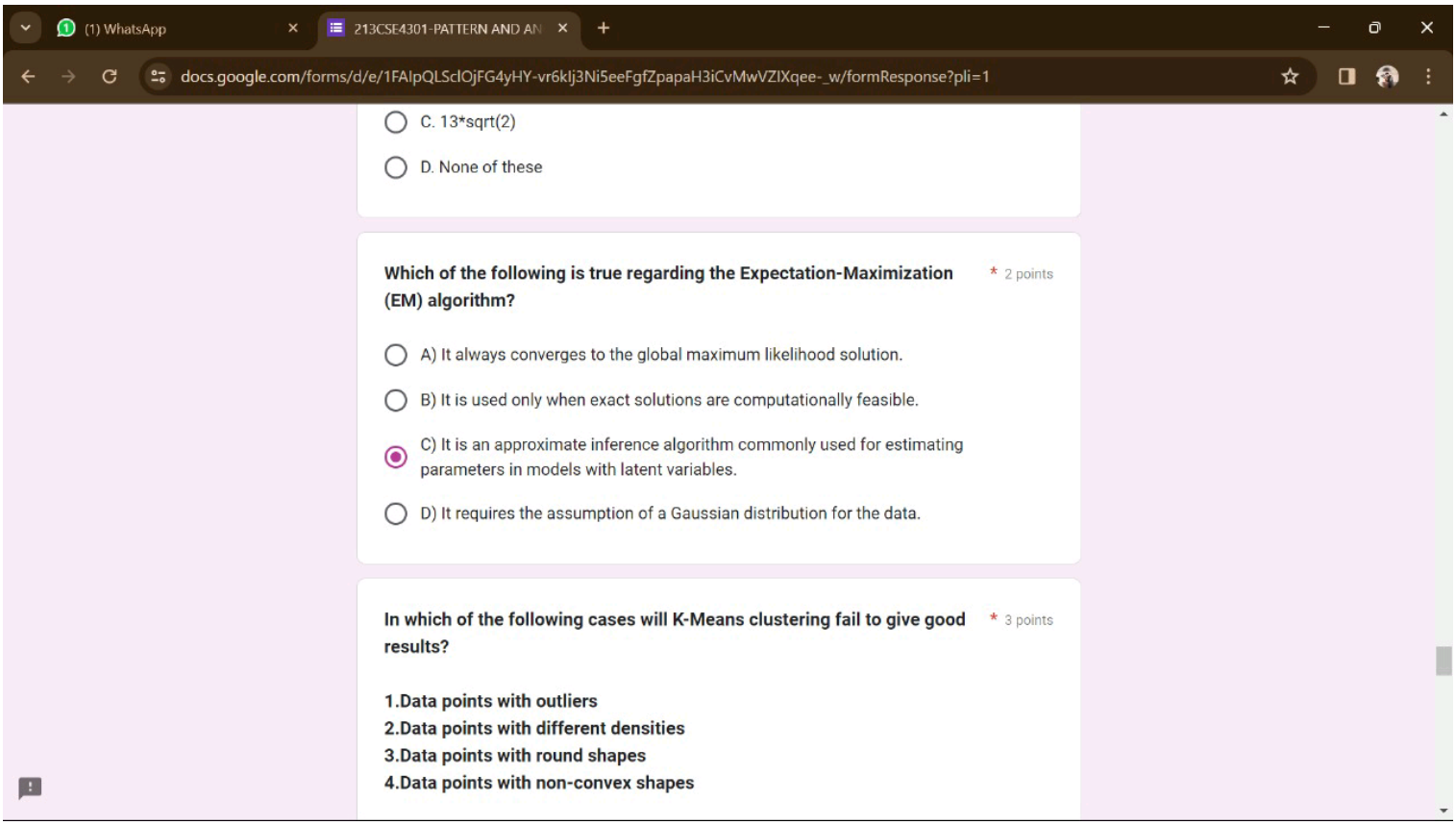
Suppose you have trained an anomaly detection system for fraud detection, and your system that flags anomalies when p(x) is less than ε * 3 points

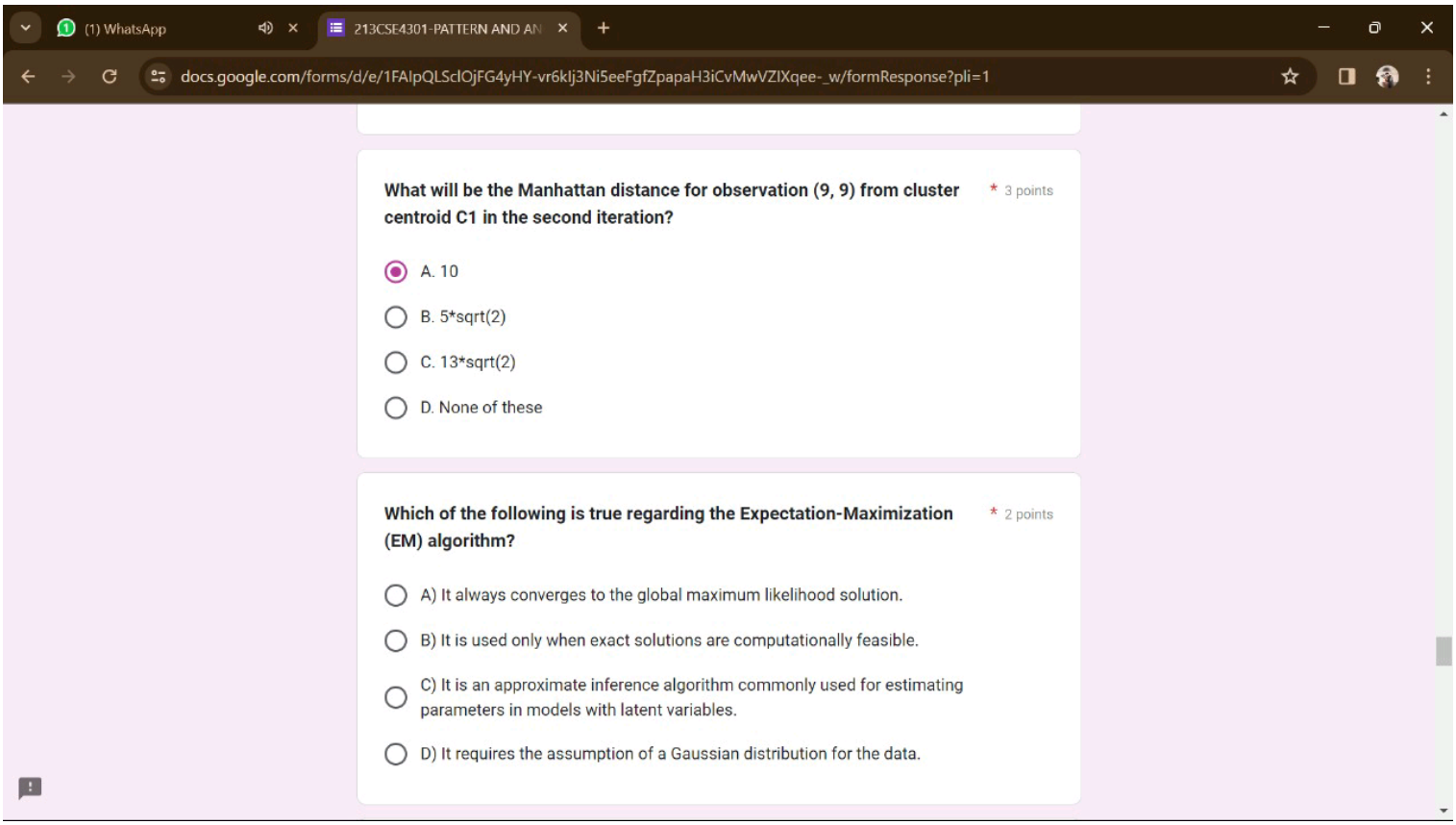

WhatsApp
213CSE4301-PATTERN AND AN
docs.google.com/forms/d/e/1FAIpQLSclOjFG4yHY-vr6klj3Ni5eeFgfZpapaH3iCvMwVZIXqee-_w/formResponse?pli=1
Quiz Evaluation
Answer all the following questions:
Which of the following can act as possible termination conditions in K-Means?
* 3 points
1.For a fixed number of iterations.
2.The assignment of observations to clusters does not change between iterations, except for cases with a bad local minimum.
3.Centroids do not change between successive iterations.
4. Terminate when RSS falls below a threshold.
A. 1, 3 and 4
B. 1, 2 and 3
C. 1, 2 and 4
D. All of the above
Given are six points with the following attributes
* 4 points

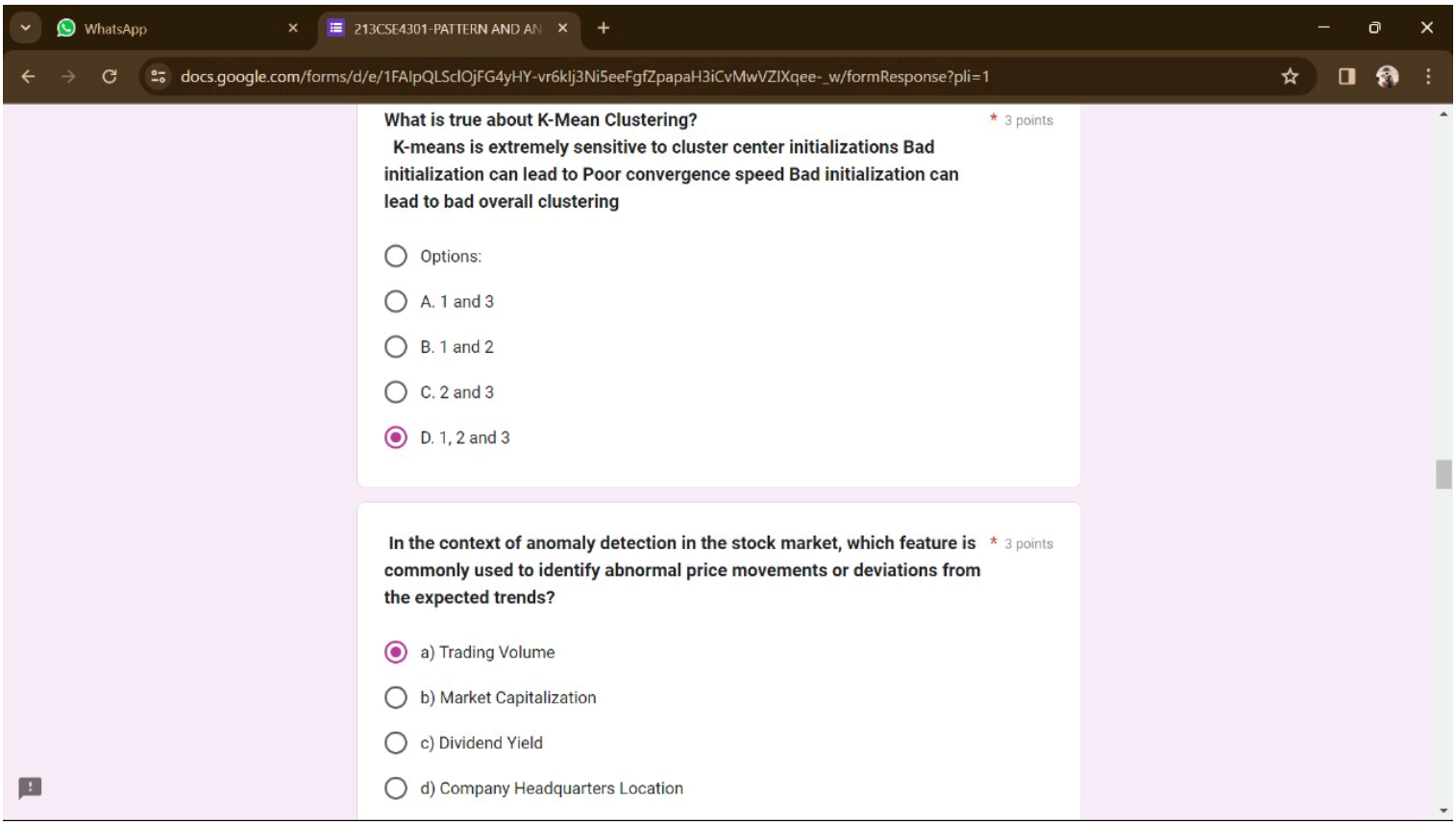

(1) WhatsApp
213CSE4301-PATTERN AND AN
docs.google.com/forms/d/e/1FAIpQLSclOjFG4yHY-vr6klj3Ni5eeFgfZpapaH3iCvMwVZIXqee-_w/formResponse?pli=1
A. 1
B. 2
C. 3
D. 4
What are the outcomes of DBSCAN techniques? *
3 points
a. Core points
b. Border points
c. Noise points
i. A only
ii. A and C only
iii. All options
iv. A and B only
Back
Submit
Clear form
Never submit passwords through Google Forms.

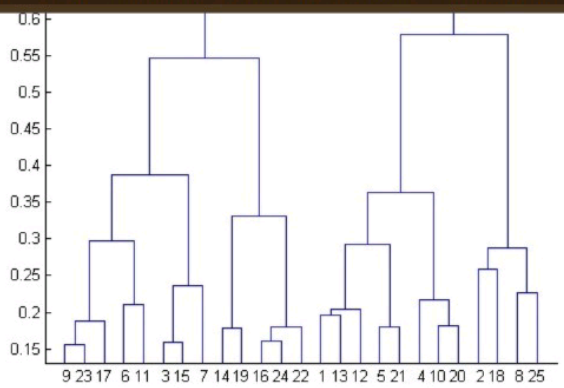

WhatsApp
213CSE4301-PATTERN AND AN
docs.google.com/forms/d/e/1FAIpQLSclOjFG4yHY-vr6klj3Ni5eeFgfZpapaH3iCvMwVZIXqee-_w/formResponse?pli=1
From scipy import _____ *
3 points
threshold = 2.5
df = load_breast_cancer(as_frame=True).data
z_scores =_____.zscore(df)
outliers = df[abs(_____) > threshold]
Dataset, stats, outlier
Dataset, dataset, z_scores
Special, stats, outlier
stats, stats, z_scores
LOF method is Effective in identifying outliers in datasets with _____, howerever it doesn't require assumptions about the _____ of the data. * 3 points
a. Same density, distortion
b. Varying density, distortion
c. Same density, distribution
d. Varying density, distribution


WhatsApp
213CSE4301-PATTERN AND AN
docs.google.com/forms/d/e/1FAIpQLSclOjFG4yHY-vr6klj3Ni5eeFgfZpapaH3iCvMwVZIXqee-_w/formResponse?pli=1
An outlier is an observation that is_____. It is also called_____.
* 2 points
Normally distributed, noise
Rare, anomaly
Regular, noise
Distinct, normally distributed
Assume you want to cluster 7 observations into 3 clusters using the K-Means clustering algorithm. After the first iteration, clusters C1, C2, C3 have following observations: C1: {(2,2), (4,4), (6,6)}
C2: {(0,4), (4,0)}
C3: {(5,5), (9,9)}
What will be the cluster centroids if you want to proceed with the second iteration?
* 3 points
A. C1: (4,4), C2: (2,2), C3: (7,7)
B. C1: (6,6), C2: (4,4), C3: (9,9)
C. C1: (2,2), C2: (0,0), C3: (5,5)


(1) WhatsApp
213CSE4301-PATTERN AND AN
docs.google.com/forms/d/e/1FAIpQLSclOjFG4yHY-vr6klj3Ni5eeFgfZpapaH3iCvMwVZIXqee-_w/formResponse?pli=1
C. 13*sqrt(2)
D. None of these
Which of the following is true regarding the Expectation-Maximization (EM) algorithm?
* 2 points
A) It always converges to the global maximum likelihood solution.
B) It is used only when exact solutions are computationally feasible.
C) It is an approximate inference algorithm commonly used for estimating parameters in models with latent variables.
D) It requires the assumption of a Gaussian distribution for the data.
In which of the following cases will K-Means clustering fail to give good results?
* 3 points
1.Data points with outliers
2.Data points with different densities
3.Data points with round shapes
4.Data points with non-convex shapes

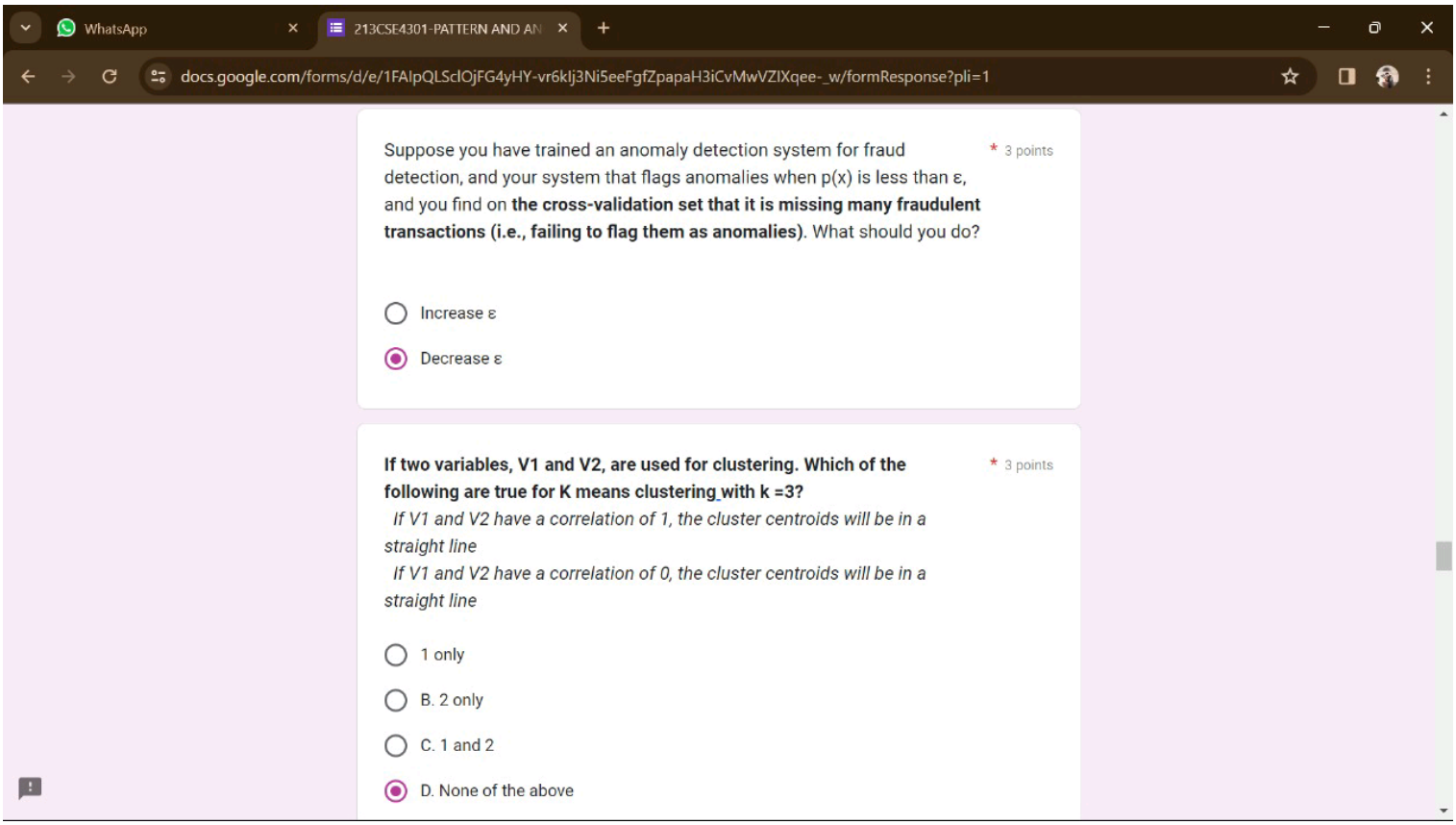

(1) WhatsApp
213CSE4301-PATTERN AND AN
docs.google.com/forms/d/e/1FAIpQLSclOjFG4yHY-vr6klj3Ni5eeFgfZpapaH3iCvMwVZIXqee-_w/formResponse?pli=1
What will be the Manhattan distance for observation (9, 9) from cluster centroid C1 in the second iteration?
* 3 points
A. 10
B. 5*sqrt(2)
C. 13*sqrt(2)
D. None of these
Which of the following is true regarding the Expectation-Maximization (EM) algorithm?
* 2 points
A) It always converges to the global maximum likelihood solution.
B) It is used only when exact solutions are computationally feasible.
C) It is an approximate inference algorithm commonly used for estimating parameters in models with latent variables.
D) It requires the assumption of a Gaussian distribution for the data.

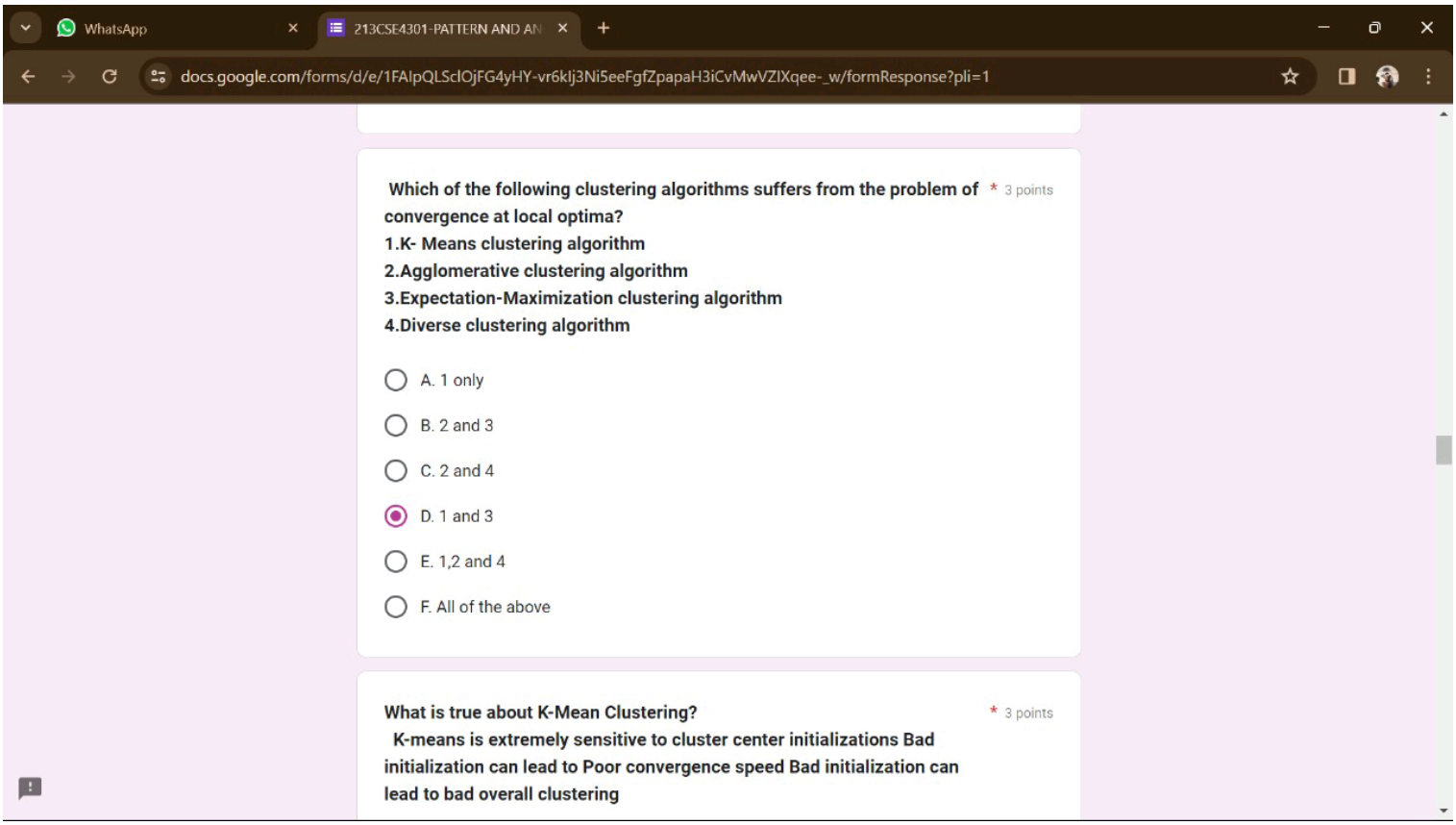

WhatsApp
213CSE4301-PATTERN AND AN
docs.google.com/forms/d/e/1FAIpQLSclOjFG4yHY-vr6klj3Ni5eeFgfZpapaH3iCvMwVZIXqee-_w/formResponse?pli=1
What is true about K-Mean Clustering?
* 3 points
K-means is extremely sensitive to cluster center initializations Bad initialization can lead to Poor convergence speed Bad initialization can lead to bad overall clustering
Options:
A. 1 and 3
B. 1 and 2
C. 2 and 3
D. 1, 2 and 3
In the context of anomaly detection in the stock market, which feature is commonly used to identify abnormal price movements or deviations from the expected trends?
* 3 points
a) Trading Volume
b) Market Capitalization
c) Dividend Yield
d) Company Headquarters Location



- ◯ There were 28 data points in the clustering analysis
- ◯ B. The best no. of clusters for the analyzed data points is 4
- ◯ C. The proximity function used is Average-link clustering
- ◉ D. The above dendrogram interpretation is not possible for K-Means clustering analysis

What will be the Manhattan distance for observation (9, 9) from cluster centroid C1 in the second iteration? * 3 points


WhatsApp
213CSE4301-PATTERN AND AN
docs.google.com/forms/d/e/1FAIpQLSclOjFG4yHY-vr6klj3Ni5eeFgfZpapaH3iCvMwVZIXqee-_w/formResponse?pli=1
Suppose you have trained an anomaly detection system for fraud detection, and your system that flags anomalies when p(x) is less than ε, and you find on **the cross-validation set that it is missing many fraudulent transactions (i.e., failing to flag them as anomalies)**. What should you do?
* 3 points
Increase ε
Decrease ε
If two variables, V1 and V2, are used for clustering. Which of the following are true for K means clustering with k =3?
If V1 and V2 have a correlation of 1, the cluster centroids will be in a straight line
If V1 and V2 have a correlation of 0, the cluster centroids will be in a straight line
* 3 points
1 only
B. 2 only
C. 1 and 2
D. None of the above


WhatsApp
213CSE4301-PATTERN AND AN
docs.google.com/forms/d/e/1FAIpQLSclOjFG4yHY-vr6klj3Ni5eeFgfZpapaH3iCvMwVZIXqee-_w/formResponse?pli=1
Which of the following clustering algorithms suffers from the problem of convergence at local optima? * 3 points
1.K- Means clustering algorithm
2.Agglomerative clustering algorithm
3.Expectation-Maximization clustering algorithm
4.Diverse clustering algorithm
A. 1 only
B. 2 and 3
C. 2 and 4
D. 1 and 3
E. 1,2 and 4
F. All of the above
What is true about K-Mean Clustering? * 3 points
K-means is extremely sensitive to cluster center initializations Bad initialization can lead to Poor convergence speed Bad initialization can lead to bad overall clustering

(1) WhatsApp | 213CSE4301-PATTERN AND AN | docs.google.com/forms/d/e/1FAIpQLSclOjFG4yHY-vr6klj3Ni5eeFgfZpapaH3iCvMwVZIXqee-_w/formResponse?pli=1

C) Continuous latent variables are inferred from observed variables and can take on any value within a certain range.

D) Continuous latent variables are only used in regression models and not in clustering algorithms.

In the context of statistical modeling, which of the following statements accurately describes the Expectation-Maximization (EM) algorithm? * 3 points

A) The EM algorithm is used to estimate unknown parameters in a model by maximizing the likelihood function.

B) The EM algorithm guarantees convergence to the global maximum of the likelihood function.

C) The EM algorithm relies on a single-step optimization process to estimate parameters.

D) The EM algorithm is only applicable to models with a single component.

What should be the best choice of no. of clusters based on the following results? * 3 points

Silhouette coefficient