Question 1:

a)

Feature selection is the process of removing redundant features from our data set. We "select" only the features that are of importance and we discard the features that we deem irrelevant to our model. The advantages of using this process is that we now have a more computational dataset to work with making our model more efficient in terms of training. A disadvantage that we might encounter with Feature selection would be a more noticed bias. This is because we individually select features which might be a biased choice.

Feature extraction is the process of transforming our data into a lower dimension for better interpretability. This is done by using well known techniques such as LDA (Linear Discriminant Analysis) and PCA (Principal Component Analysis). An advantage would be a reduction of noise and a disadvantage would be an information loss. Since we are breaking our data down and making it more simple, there is going to be an information loss that may affect the accuracy of our model.

b)

Forward selection is the process of adding variables one by one for the purpose of decreasing error. We keep adding variables until there is no greater change in the error and we finally stop. An advantage of using Forward selection is that it is a more simpler approach and as we usually observe in machine learning, the simpler approach is usually the better one. A disadvantage would be the case where we might not find the best subset since this process does not consider the removal of predictors once they are added

Backward selection is the opposite where we start off with all out variables and we keep remove them one by one and consider the reduction in error. We keep removing until we observe that there is marginally less betterment in the error. An advatange of this process would be the fact that we have a higher potential of reaching our desired number of variables since we start off considering all of them. A disadvantage would be that this process is more computationally exhaustive because we consider more variables compared to Forward selection.

c)

Number of possible subsets of features would be 10C20 which is 184756 possible subsets.

Question 2:

a)

PCA (Principal Component Analysis) is the process of mapping from the inputs in the original d-dimensional space to a new dimensional space for the purpose of simplifying the dimensionality of our data set. This is an unsupervised method. PCA is performed by executing the following steps:

- Standardize the data
- Compute the Covariance Matrix

- Calculate the eigenvectors and eignevalues
- Sort the eigenvectors
- Select the top k eigenvectors where k becomes the number of dimensions

b)
The significance of PCA is a vital one as it lets us train out model with a lower dimension dataset that is considered to be better as it may deal with the complication of extremely noisy data and it also allows our model to learn easier because the complexity of the model can be correleted to the dimensionality of our data and as we have learned, dealing with complex data is not always the most advisable compared to simpler data. The optimal number to consider can be obtained by looking at a screen plot which depicts a graph with decreasing eigenvalues. The point where the variance does not seem to change would be our k value.
The advantage of using PCA is the efficiency and the simplicity that it offers and the downside would be the loss of data.

c)
The first line would be the most optimal as it has the highest spread which means it captures the data and allows for better understanding compared to the other lines.

Question 3

Question 4

a)
A full rank covariance matrix in PCA is extremely important as it allows us to reduce the dimensionality of our dataset while capturing as much variance as possible, ensuring that all our principal components are maximized in terms of usage.

b)
It is common practice to neglect the later eigenvectors with smaller eigenvalues in PCA because these eigenvalues correspond to a dimensionality that captures the lower levels of variance which defeats the whole purpose of performing PCA.
It is reasonable to consider less significant eigenvectors when the larger eigenvectors are show to capture below 90% variance which implies that they are not as helpful compared to the lower values.
It is reasonable to not consider less significant eigenvectors when we can observe that the larger eigenvectors are capable of accounting for 90% of variance in a dataset.

c)
Variance = (15 + 12)/50 = 54%

This shows that the first two components only explain about 54% of the variance in the dataset. Since we aim to capture a variance between 90% to 95% we must include the first 6 to 7 eigenvectors as well.

Question 5

Question 6

b)

2 Challenges associated with the dependency are:

- We might occur with inconsistent values returned from repeating the process of clustering because we may encounter many local maximus which makes the task much more difficult.

2 different initialization methods used are:

- We can start by randomly choosing points to be our starting reference vectors.
- We can take advantage of the data's principle components and divide the range into intervals of equal lengths thus giving us points by taking the mean of each of the given lengths.

In the context of Principal Component Analysis (PCA), a full-rank covariance matrix is important for several reasons:

Neglecting the later eigenvectors with smaller eigenvalues in Principal Component Analysis (PCA) is a common practice because these eigenvectors correspond to directions in the data with the least variance. The core idea of PCA is dimensionality reduction, where the aim is to capture as much of the variability in the data as possible using fewer dimensions. The eigenvectors of the covariance matrix (also known as principal components) are sorted by their corresponding eigenvalues in descending order, where the magnitude of an eigenvalue indicates the amount of variance captured by its associated eigenvector.

Reasons for Neglecting Smaller Eigenvalues:

- Noise Reduction: The dimensions associated with smaller eigenvalues often capture noise rather than meaningful information. By neglecting these, PCA helps in denoising the data, leading to a more robust analysis.
- Dimensionality Reduction: The main purpose of PCA is to reduce the number of variables while preserving as much information as possible. Focusing on eigenvectors with larger eigenvalues allows for a compact representation that retains the most significant variance.
- Interpretability: Models built on reduced dimensions are often easier to understand and interpret. By concentrating on the most significant directions of variance, it becomes simpler to analyze the relationships between variables.

Situations to Consider or Not Consider Less Significant Eigenvectors:

Consider:

- High-Dimensional Data with Subtle Signals: In datasets where important signals are spread out and not dominant, the variance explained by smaller eigenvalues might still hold relevant information. Especially in fields like genomics or signal processing, subtle signals could be critical.
- Comprehensive Variance Coverage: If the objective is to capture a very high percentage of the total variance (e.g., over 95%), including eigenvectors associated with smaller eigenvalues might be necessary.
- Anomaly Detection: Sometimes, the variance captured by smaller eigenvalues corresponds to anomalies or outliers. In such cases, including these components can be crucial for tasks like fraud detection.

Not Consider:

- Clear Dominance of Principal Components: If the first few principal components capture a vast majority of the variance, the additional complexity of including more components might not justify the marginal gain in explained variance.
- Computational Efficiency: Including more components requires more computational resources. If computational efficiency is a concern, focusing on the most significant eigenvectors is preferable.
- Overfitting Risk: Including too many dimensions can lead to overfitting, especially in predictive modeling, where the model might learn noise in the training data, leading to poor generalization.

In conclusion, the decision to consider or neglect less significant eigenvectors in PCA depends on the specific goals of the analysis, the nature of the dataset, and the trade-offs between interpretability, computational efficiency, and the need to capture subtle patterns or anomalies.

The first two principal components explain 54% of the total variance in the dataset. This value suggests that by choosing just the first two components, more than half of the total variance is captured, which is a significant amount considering the reduction from a higher-dimensional space to just two dimensions.

The decision on how many components to choose depends on the specific requirements for variance explanation and the application at hand. If the goal is dimensionality reduction for visualization or to achieve a simple yet reasonably informative representation of the data, then selecting the first two components might be adequate, especially if computational efficiency and model simplicity are priorities.

However, if the objective is to capture a more substantial portion of the variance (e.g., more than 80-90%), then including more components would be necessary. In this case, considering the third and possibly the fourth component could be valuable, depending on their individual contributions to the total variance and the diminishing returns of adding more components. This decision should be balanced with the considerations of overfitting, computational complexity, and the interpretability of the resulting model or analysis.

Question 3

$$\mu_n = w^T \mu$$

$$\Rightarrow \begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 1 \cdot 2 + 2 \cdot 3 \end{bmatrix} = \begin{bmatrix} 8 \end{bmatrix}$$

$$\Sigma_n = w^T \Sigma w = \begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 5 \\ 5 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$= \begin{bmatrix} 11 \end{bmatrix}$$

$$\therefore f(X) \sim N(\mu, \Sigma) = f(x) \sim N(8, 11)$$

Question 5

$$\Psi = cov(X) - V \cdot V^T \quad \text{(Must consider the first 2 rows)}$$

$$V \cdot V^T = \begin{bmatrix} 0.41 & -0.14 \\ 0.08 & 0.02 \\ -0.03 & -0.07 \\ 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0.41 & 0.08 & -0.03 & 0 \\ -0.14 & 0.02 & -0.07 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0.175 & -0.44 \\ -0.44 & 0.65 \end{bmatrix}$$

$$\Rightarrow \Psi = \begin{bmatrix} 0.15 & 1 & 0.2 & 0.08 \\ 0.23 & 0.1 & 0.43 & 0.32 \\ 0.19 & 0.6 & 0.5 & 0.07 \\ 0.3 & 0.4 & 0.5 & 0.07 \end{bmatrix} - \begin{bmatrix} 0.41 & -0.14 \\ 0.08 & 0.2 \\ 0.3 & -0.7 \\ 0 & 0 \end{bmatrix} \times \begin{bmatrix} 0.175 & -0.44 \\ -0.44 & 0.65 \end{bmatrix} \begin{bmatrix} 0.41 & 0.08 & 0.07 & 0 \\ -0.145 & 0.2 & -0.7 & 0 \end{bmatrix}$$

$$\therefore \Psi = \begin{bmatrix} -0.038 & 0.995 & 0.203 & 0.08 \\ 0.225 & 0.054 & 0.446 & 0.32 \\ 0.173 & 0.616 & 0.444 & 0.07 \\ 0.3 & 0.4 & 0.5 & 0.07 \end{bmatrix}$$

Question 6

$c_1 = 30$ $c_2 = 45$ $c_3 = 41$

$a = 21$ $b = 35$ $c = 10$ $d = 28$ $e = 41$

Round 1 :

distances from all points to each cluster :

a : distance(c_1) = abs(30 - 21) = 9 ★ c_1

distance(c_2) = abs(45 - 21) = 24

distance(c_3) = 17

b : distance(c_1) = 5 ★ c_1

distance(c_2) = 10

distance(c_3) = 31

c : distance(c_1) = 20

distance(c_2) = 35

distance(c_3) = 6 ★ c_3

d : distance(c_1) = 2 ★ c_1

distance(c_2) = 17

distance(c_3) = 24

e : distance(c_1) = 11

distance(c_2) = 4 ★ c_2

distance(c_3) = 37

Round 1 : $c_1 = \frac{21+35+28}{3} = 28$, $c_2 = 41$ $c_3 = 10$

Round 2 :

a : distance (c_1) = abs $(28 - 21) = 7$ ★ c_1
distance (c_2) = abs $(41 - 21) = 20$
distance (c_3) = abs $(10 - 21) = 11$

b : distance (c_1) = 7
distance (c_2) = 6 ★ c_2
distance (c_3) = 25

c : distance (c_1) = 18
: distance (c_2) = 31
distance (c_3) = 0 ★ c_3

d : distance (c_1) = 0 ★ c_1
: distance (c_2) = 13
distance (c_3) = 18

e : distance (c_1) = 13
distance (c_2) = 0 ★ c_2
distance (c_3) = 31

Round 2 :

$c_1 = \frac{21+28}{2} = \underline{24.5}$, $c_2 = \frac{35+41}{2} = \underline{38}$

$c_3 = \underline{10}$