Question 1: Assignment Summary

Problem Statement:

HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.

After the recent funding programmers, they have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.

Solution Methodology:

I first had a look at the dataset and the data dictionary in Microsoft Excel, there I found that some columns – imports, exports and health are in percentage form of Total GDP. As there might be cases where actual values of these columns are less but total percentage of GDP is more, in those cases my analysis will not be accurate if I go ahead with these %age values.

To convert these percentage values to raw values, I have used GDPP column as after doing a little research on Google I found that GDPP is a method of measuring GDP of a country.

Below are the steps which I have followed in Python Notebook:

1. First, I imported Country-data.csv dataset into python.
2. Then I looked at basic information like – shape, info, describe, Data Type Check of dataset.
3. **Missing Values Check** – Then I proceeded towards missing values check in our dataset. There I found that there is no data missing in our dataset.
4. **Duplicate Value Check** – Then I checked for any duplicate records and found no duplicate records present in the dataset.
5. **Data Preparation:** After doing this, I went ahead and converted columns – exports, imports and health from percentage to their actual values as I had found this during my analysis through data-dictionary.
6. **Outlier Detection:** After this, I plotted distribution of each column using boxplot and distplot. I also checked skewness and kurtosis present in each column.
   There I realized that my data is highly skewed and has a lot of outliers for almost all of the columns. But I didn't remove these outliers at this moment because these outliers are there may be because of poor condition of most needed country.
7. **Data Visualization:** Then I plotted different columns against each other. Basically I wanted to visualize if there is any pattern being followed in between different columns.
   There I found that countries whose income is low, there child_mort rate is highest and Countries with low income have low GDPP and vice-versa.
8. **MutiCollinearity Detection** – After above steps, I checked with correlation between different feature variables of our dataset and plotted a heatmap.

There I found that, most of the columns can be described using child_mort, GDPP and income. So, I went ahead for dimensionality reduction.

9. **Feature Scaling:** As it is important to scale feature variables before applying PCA, so I used Standard Scaler to scale my features.
10. **Dimensionality Reduction using PCA:** Then I applied PCA algorithm on my featured variables. Now here my question was how many Principle Components I should choose which can explain most of the variance in my dataset.
    - For this, I plotted **scree plot and check for cumulative explained variance ratio**. There I found that almost 87% of the variance can be explained by 3 PCs and 93% of the variance can be explained via 4 PCS.
    - Then **I checked for Eigen values**. There I realized that PC1, PC2 and PC3 has eigen value >=1 but PC4 has eigen value <1. As I know that PCs with Eigen Values >=1 are most optimal PCs. So, I decided to take 3 PCs.
11. **Feature Visualization:** Then I plotted Actual features using PC1 and PC2 and I found that all my features are nicely explained by these principle components.
12. **Multicollinearity Detection**: As before PCA, I had found that my feature variables are in a strong relationship with each other, I check for correlation ship between PCs. I found that all my PCs are independent of each other.
13. **Outlier Detection and Treatment:** Later I again checked for outliers in my dataset and treated them as below:
    - I calculated 5% and 95% quartile values of each PC and removed them from my dataset.
14. **Visualizing the data points:** Now, I plotted a scatter plot to see if I can see any cluster formation between my data points. And Yes, My data points were clearly taking the shapes of nice clusters.
15. **Clustering:**
    - Before proceeding on clustering, I wanted to cross-check if my data points are suitable for making clusters. So, I used for Hopkins Statistics and found that H value for my dataset is 0.82 which means I am good to go ahead and create clusters.
      I decided to apply both KMeans and Hierarchical clustering for making clusters.
    - **KMeans Clustering:** Now, I know that for Kmeans clustering, number of clusters should be known beforehand.
    - **Choosing Optimal Number of Clusters:** Now, to know optimal value of k, I plotted **Elbow curve and silhouette score**. There I found that k=3 and k=5 are both optimal number of clusters. So, I decided using both values 1 by 1 and see if I get desired result.
        - **KMeans Clustering using k=3:** After this, I applied KMeans clustering technique with k=3 and intial centroid selection criteria using Kmeans++.
            - **Data Visualization**: After getting cluster labels for my data points, I plotted a scatter plot and visualized different clusters which were formed after KMeans.
            - My data points are nicely distributed among different clusters.
            - Then I did **Mean Analysis of my clusters**. There I found that there was a huge gap between mean values of all feature variables between 2 clusters and 3rd cluster which means my data points are distributed properly in these clusters.
            - SO, I decided to apply KMeans with k=5: my second optimal k value.

- **Kmeans using K-5:** Same way, I applied KMeans algorithm with k=5 and initial centroid selection technique as KMeans++.
    - **Data Visualization:** After I got cluster labels, I again visualized my data points of all clusters. This time my there was a clear inter cluster distance and nice clusters were created.
    - **Mean Analysis of Clusters**: Then I plotted bar graphs for mean value of different features across all clusters. There I got cluster=4 which had 22 countries as my desired result.
- **Hierarchical Clustering:** Then I went ahead and plotted Dendrogram with single as well as complete linkage. There also I got k=3 as optimal number of clusters.
    - Same way, I sis data visualization and mean analysis for my data points among these clusters. Like KMean k=3, here also I found huge gap in mean values between my clusters. Which means There should be more clusters to group my data points.
    - So, I decided to go with the result of KMeans algorithm with k=5.

16. **Cluster Analysis:** Now, we need to **add cluster labels to our dropped records from outlier treatment.**
    - Now, I checked for minimum values of child_mort across the clusters and maximum values of income and gdpp. With the concept of binning, I checked for incoming record's child_mort value and cross-checked minimum child_mort values in each cluster and assigned cluster label to incoming record.
    - Below is my criteria:
      if(child_mort >= 55.2): then assign cluster as 4

      else if(child_mort value >=17.4): assign cluster as 2

      else if(child_mort value >=5.5): assign cluster as 0

      else if(child_mort value >=3.2): assign cluster as 3

      else if(child_mort value >=2.6): assign cluster as 1

    - After assigning clusters to dropped records, I verified income values of incoming records against min and max value of income in my clusters. All my results looked good.
17. **Analyzing countries which are in direst need of help:** By now, I had 39 countries which are in most need of help. But I wanted to look for countries which are in crucial need of help.
    So, I decided to go with below criteria to look for those countries where:
    - **Child_mort is > 50% of child_mort in cluster4(Cluster which has countries who needs help) and**
    - **Income < 50% of income in cluster 4 or**
    - **GDPP < 50% of GDPP in cluster4**

      Note: Here I have applied 'or' condition between income and GDPP because after taking a close look at data set, I found that there are countries where Income is little more than mean but GDPP is very less.

As countries health can be determined by GDPP, I have taken all such countries into account.

As we have a limited budget, so want to know Top 10 countries first which needs help urgently.

To find the same, I have sorted the list of countries on the basis of Child_mort in Descending order and GDPP in Ascending order, I could find TOP 10 countries which are in urgent need of help.

1. **Haiti**
2. **Sierra Leone**
3. **Chad**
4. **Central African Republic**
5. **Mali**
6. **Nigeria**
7. **Niger**
8. **Congo, Dem. Rep.**
9. **Burkina Faso**
10. **Guinea-Bissau**

If HELP NGO wants to invest $10 Million, that should be invested in above listed countries as they are in direst need of help.

Question 2: Clustering

a) **Compare and contrast K-means Clustering and Hierarchical Clustering.**

KMeans vs Hierarchical:

S.No.	KMeans	Hierarchical
1.	K-means can handle big data as time complexity of K-Means is linear: O(n)	Hierarchical clustering technique can not handle big data as its time complexity is quadratic: $O(n^2)$
2.	To use KMeans, we should know value of K i.e. number of clusters to be created beforehand.	To use this technique, there is no need to provide the value of k beforehand. Infact, we can plot a dendrogram and choose the number of clusters we need.
3.	K-Means algorithm gives different result on different runs for the same dataset as the result is fully dependent on initial random choice of centroids.	This algorithm provides same result even after multiple runs of algorithm if we don't alter the data in any way.

		But this method can also provide different solution on reordering the data or dropping a few records.
4.	Data points can be assigned to different cluster in re-computation step.	Once the data point is assigned to a cluster, it cannot be undone.
5.	There is no transparency in how the algorithm assigned data points to different clusters.	We can plot a dendrogram to see how the algorithm worked. It is good for presentation as well.

b) Briefly explain the steps of the K-means clustering algorithm.

K-Means Clustering Algorithm:

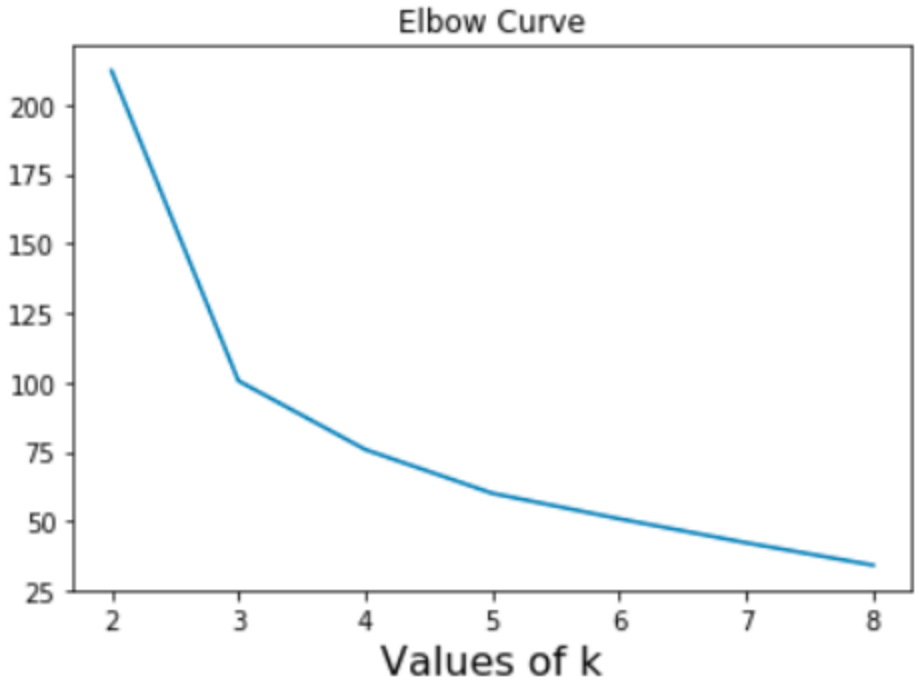
KMeans clustering algorithm is the process of dividing N data points into K groups or clusters. Here are the steps of K-Means algorithm:

1. Start by choosing K random points as the initial cluster centroids.
2. Assign each data point to their nearest cluster center. The most common way of measuring the distance between the points is Euclidean distance.
3. For each cluster, compute the new cluster center which will be the mean of all cluster members.
4. No re-assign the data points to different clusters by taking into account the new cluster centers.
5. Keep iterating through step 3 & 4 until there are no further changes possible.

c) How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

There are a number of pointers which helps us in choosing optimal number of k for KMeans clustering.

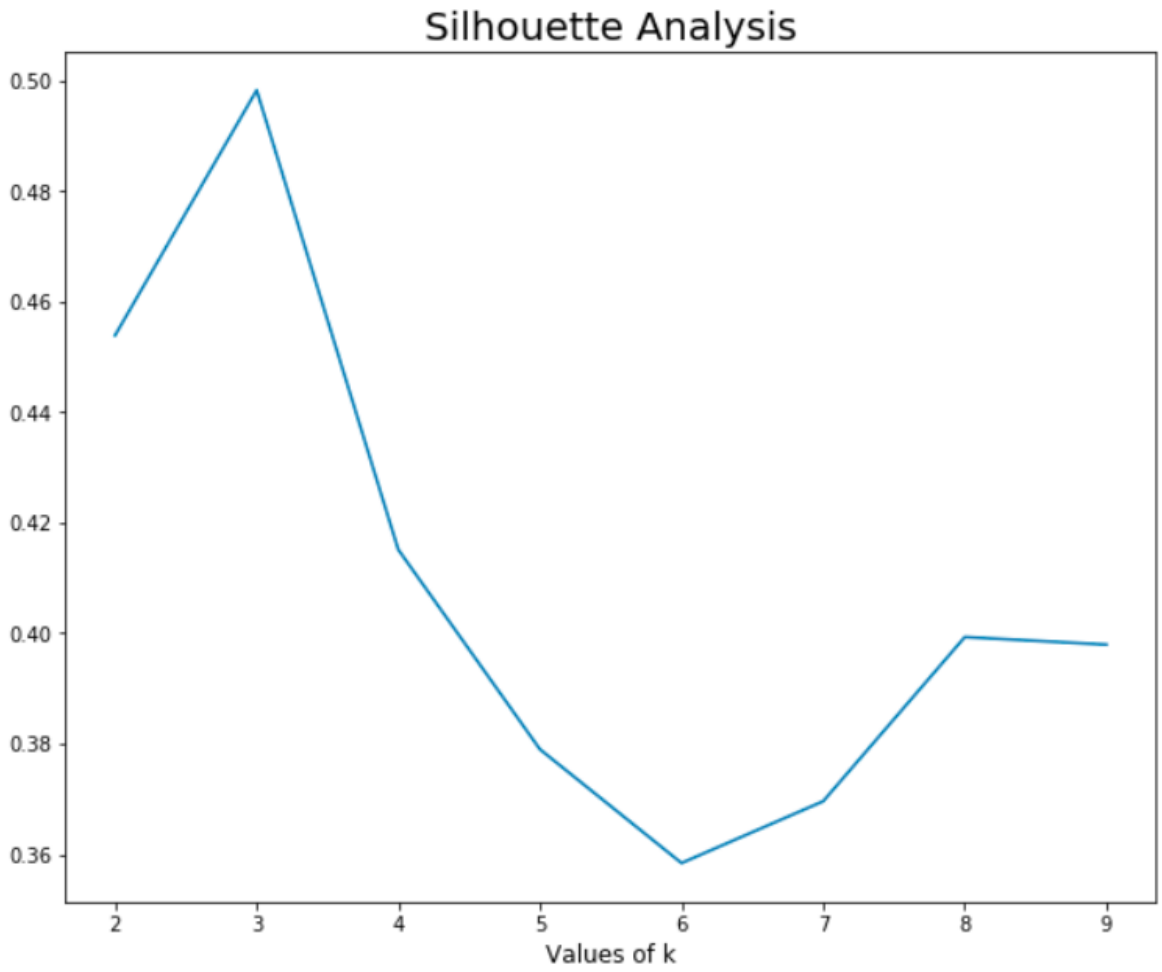
1. Elbow Method:
   - Compute Clustering algorithm(KMeans) for different values of k.
   - For each value of k, calculate total within-cluster sum of square(wss).
   - Plot a curve of these wss against its k value. It looks as below:



- Here, location of bend in the plot is generally considered as an indicator of appropriate number of clusters.

2. Average Silhouette Method:

- Compute clustering algorithm(KMeans in our case) for different values of k.
- For each k, calculate average silhouette score of observations.
- Plot the curve of avg silhouette score according to number of clusters. It may look as below:



- The location of maximum is considered as appropriate number of clusters.

Business Aspect: In real world, we always want to create cluster in such a way that distance between all data points of a cluster is minimum and distance between data points against data points of other clusters is maximum.

This is done so that data points of a cluster behave similar but at the same time they behave different in comparison with data points of other clusters. Only then, we will be able to come up with correct segment of target customers/data points.

Statistical Aspect: Here, Elbow curve actually gives us distance in between data points of a cluster. That is why we choose k against minimum significant distance difference.

And Silhouette average gives us average distance of datapoints of a cluster against that of another cluster. That is why we chose k with maximum significant distance difference.

d) Explain the necessity for scaling/standardisation before performing Clustering.

In real world, most of the times we are dealing with data having features with different measuring units. We apply different scaling methods to rescale these variables in our dataset so that the variables can share a same scale.

The reason behind its high importance in clustering is because we create different groups in this analysis based on distance between points in mathematical space.

When working with data where each variable means something different, the fields are not directly comparable and may not have the same importance. In a situation where one field has a much greater range of values than another, it may end up being the primary driver of defined clusters and we will end up making non-useful predictions.

Standardization helps to make the relative weight of each variable equal by converting each variable to a unitless measure or relative distance.

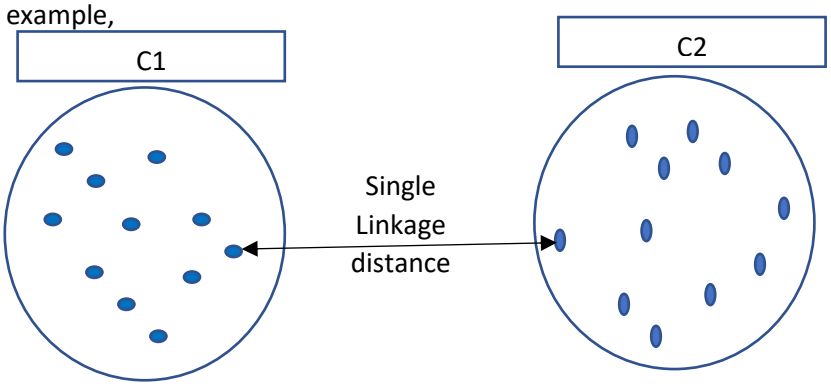
Example: Let's say amazon wants to see the most popular mobile among the customers. Now in their database, they will have quantity as well as price. Both has different scale of measure. If they don't scale and proceed towards clustering, price will overpower the importance of quantity and we will end up clustering based on highly priced mobiles rather than highly sold mobiles.

e) Explain the different linkages used in Hierarchical Clustering.

Before any clustering is performed, it is required to determine the proximity matrix containing the distance between each point using a distance function. Then, the matric is updated to display the distance between each cluster. The following 3 methods differ in how the distance between each cluster is measured:

1. Single Linkage: In single linkage hierarchical clustering, the distance between 2 clusters is defined as the shortest distance between 2 data points in each cluster.

   For example,

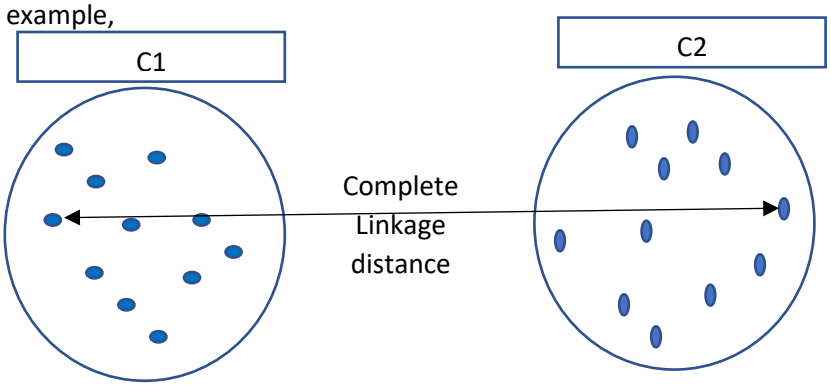


Lets say, there are 2 clusters named as C1 and C2, then single linkage distance refers to the shortest distance between 2 data points of that cluster.

L(C1, C2) = MIN(D(xc_{1i}, xc_{2j})

2. Complete Linkage: In complete linkage hierarchical clustering, the distance between 2 clusters is defined as the longest distance between 2 data points in each cluster. For example –

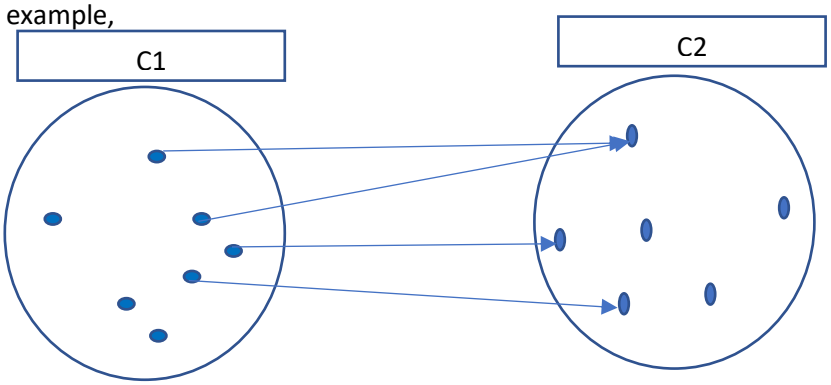
For example,



L(C1, C2) = MAX(D(xc_{1i}, xc_{2j})

3. Average Linkage: In average linkage hierarchical method, the distance between two clusters is defined as the average distance between each point in one cluster to every point in other cluster. For example, the siatnce beyween clusters C1 and C2 is equal to average length of each ararow between connecting the points of one cluster to the other.

For example,



L(C1,C2) = $1/Nc1Nc2\ {}^{N}{}_{c1}\sum_{1} \sum D(xc1i, xc2j)$

Question 3: Principal Component Analysis

a) Give at least three applications of using PCA.

Applications of PCA:

1. **MARKET Segmentation:** Market segmentation is a process used in marketing to divide customers into different clusters based on their characteristics (demographics, shopping behaviors, preferences etc.) Customers in the same market segment tend to respond similarly to market strategy.
   Therefore, segmentation process helps companies to understand their customer groups, target the right groups and tailor effective marketing strategies for different target groups.

   For example – A company named amazon, conducts a survey to understand its customers. Survey is now consisting of 4 types of questions:
   a. Attitudinal
   b. Demographic
   c. Purchase Process & Usage behavior
   d. Brand Awareness

   Lets see now that why PCA is used in such a case: If we directly go with any algorithm without using PCA, we end up dropping variables with comparatively low variance which may also cause important information loss. But in PCA, we create PCs such that all information is captured. We can even use first few PCs to describe vast majority of dataset without needing to compare and contrast every single feature. By using PC1 and PC2, we can visualize in 2D and inspect for clusters.

   Same way, we use PCA technique for all below applications:

2. **Vision:** Image Segmentation – Each image is a linear combination of pixels. Using PCA, we create clusters with pixels having same properties.
3. **Biology:** Discovering gene clusters with similar expression patterns, grouping homologous DNA sequences etc.
4. **Search Engine Grouping:** clusty.com
5. **Social Network Analysis:** Discovering user communities with similar interests
6. **Crime Analysis:** identification of “hot spots”
7. **Filtering emails**

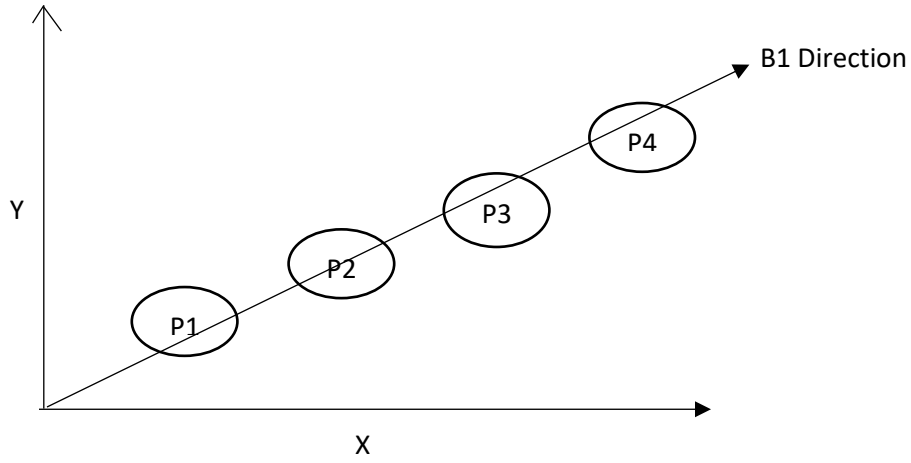
3b. Briefly discuss the 2 important building blocks of PCA - Basis transformation and variance as information.

Basis Transformation: It is the process of converting any information from one set of basis to another. Or, representing any data in new columns different from the original. Often for convenience and efficiency. As we know that standard axis is not always the axis which can provide us the base result.

Let's understand this with the help of an example:

Example:

Let's say you are gone to an unknown city and you want to go to different places to visit. Now, your friend has given you a map where all places are indicated in that map. Below is the map:



Now, If you ahead in X and Y direction, it will take a lot of your time to cover all places. But if you start going ahead in B1 direction, you will cover up all the places in much lesser time and no place will be left to visit.

Variance as information: When we deal with real life data, we see that there are a lot of columns available to us. Now, not all the feature variables are important. SO, we want to drop only those feature variables which are not important to us i.e. who doesn't contain much information.

Example: If you have a dataset of a grocery store available and you want to predict if which gender of customers returns product the most.

Dataset available:

Date	City	Gender	Product
29/09/2019	Bangalore	F	P1
29/09/2019	Bangalore	M	P2
29/09/2019	Bangalore	F	P3
29/09/2019	Bangalore	M	P4
29/09/2019	Bangalore	F	P5
29/09/2019	Bangalore	M	P6
29/09/2019	Bangalore	F	P7

Here, I can drop Date, and City column as their values are constant and these columns doesn't contain much variation.

3c. State at least three shortcomings of using Principal Component Analysis.

Shortcoming of PCA

1. PCA is limited to linearity. Means it is a linear transformation method and works well in tandem with linear models such as linear regression and logistic regression.
2. Orthogonality: PCA needs the component to be perpendicular, though in some cases, that may not be the best solution. The alternative technique is to use Independent Components Analysis.
3. Large variance implies more structure: PCA assumes that columns with low variance are not useful, which may not be true in prediction setups (especially classification problems with a high class imbalance).
4. Linearity: PCA assumes that the principle components are a linear combination of the original features. If this is not true, PCA will not give you sensible results.