Learning Objective:

Primary task is to perform segmentation on the given dataset into optimal number of clusters and to understand the similarities in the records that are grouped together. Segmentation will help us into understanding what similarities are present between the records of a particular cluster. Secondary task is to perform the analysis on the dataset and understand the data.

Dataset:

The dataset on which the analysis was done and the findings were recorded was downloaded from the following link: https://archive.ics.uci.edu/ml/datasets/Census+Income

The Census Income dataset has five files. Following are the names of the files:

- Index
- adult.data
- adult.names
- adult.test
- old.adult.names

The main content is only present in the adult.data and adult.test files. The dataset contains total 48843 records and 15 columns. Following is the list of columns:

- age: This is a numerical feature.
- Workclass: This is a categorical feature.
- Fnlwgt: This is a numerical feature.
- Education: This is a categorical feature.
- education-num: This is a numerical feature.
- martial-status: This is a categorical feature.
- occupation: This is a categorical feature.
- relationship: This is a categorical feature.
- race: This is a categorical feature.
- sex: This is a categorical feature.
- capita-gain: This is a numerical feature.
- capital-loss: This is a numerical feature.
- hours-per-week: This is a numerical feature.
- native-country: This is a categorical feature.
- income: This is a categorical feature.

The old.adult.names and adult.names contains the general description of the dataset. The Index file just provides the indexing of the given dataset.

Tasks Performed:

The following tasks were performed on the given dataset:

- Exploratory Data Analysis: To better understand the data, it was necessary to do the required analysis on the dataset. Originally, this dataset was meant for classification where the target feature is **income.** The income feature classifies the records into two categories namely **'<=50K'** and '**>50k'**. For this task, we are not going to be doing classification but rather clustering, so we will drop this feature. The dataset was divided into two files namely adult.data and adult.test which we have combined for doing all the task. We are combining the dataset files as we are not performing the classification task rather clustering and the more data we have the better .As part of the task that were performed for doing exploratory data analysis are:

- ➢ Exploring the given files in the dataset
- ➢ Checking for missing values
- ➢ Univariate Analysis(Checking the composition of each feature)
- ➢ Checking for duplicate records
- ➢ Checking for outliers
- ➢ Multivariate Analysis (Checking the breakup of multiple columns with respect to other columns)
- ➢ Checking for correlation in the numerical features

- Data Preprocessing: To make the data in good condition, so that we can apply the clustering algorithm, we need to perform some necessary tasks. Following are the task that were performed as part of data preprocessing:
  - ➢ Removing the missing values(removing the records which contain NAN values)
  - ➢ Replacing the categories which don't make sense with the mode of that particular column
  - ➢ Removing the duplicate records
  - ➢ Encoding the categorical features
  - ➢ Performing Min-Max Normalization on the numerical features

- Model Building and Evaluation: After the data preprocessing part, comes the core part which is to build the model on the preprocessed dataset. In this problem, we need to build a clustering model. The model which we chose was K-Means clustering and to find the best parameter on which we get the best group of clustering, we selected the silhouette score as the performance metric. The silhouette score measures two things cohesion and separation. Cohesion represents the similarity between the data points in a particular cluster and separation refers to the separation of clusters. The silhouette score ranges from -1 to +1. The closer the score is to +1, the better the clustering.

- Cluster Analysis: After, we have gotten the clusters we will be doing analysis on the clusters to understand the similarities between the data points and the differences between the other clusters. This is the most important step as depending on this only, we will be able to understand the composition of different clusters meaning what is similarities between the data points in a particular cluster and what separates one cluster from other clusters.

Results and Learning Outcomes:

The results are compiled phase wise:

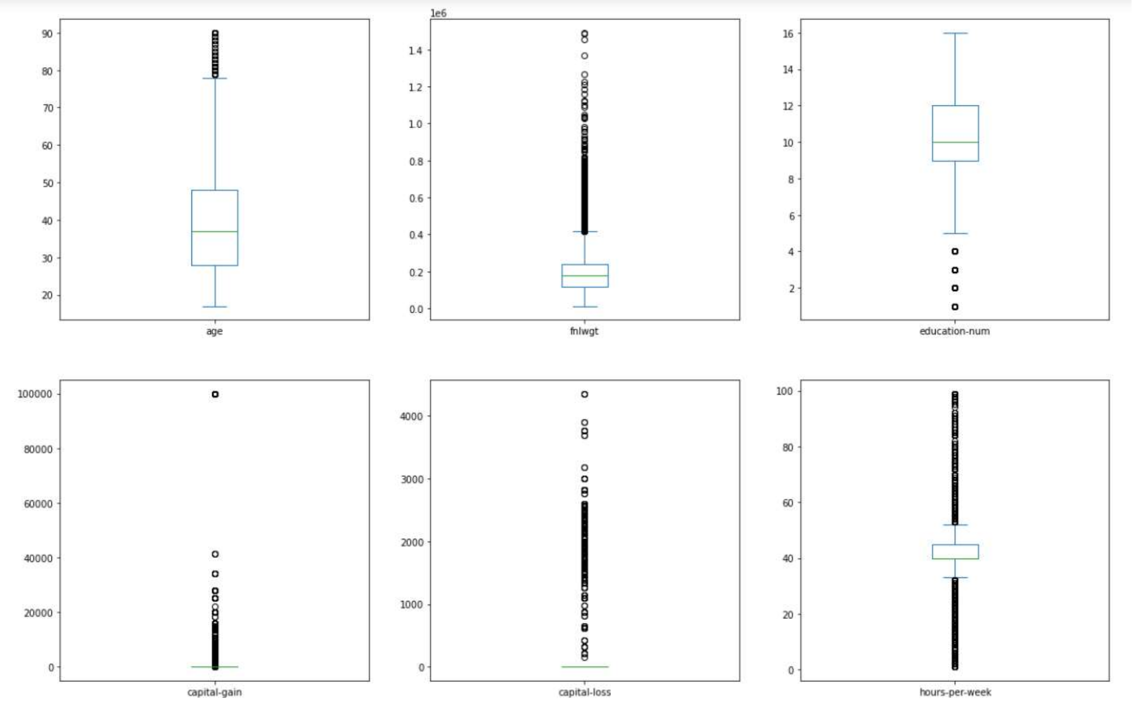
- Phase-I: This phase contains the exploratory data analysis and data preprocessing part. We first tried to look into the various measures of the whole dataset. Please refer to the **Jupyter notebook** as reference for detailed results. Following are the results highlighted:

First we combine the train and test set. We then check the number of rows and columns, some sample rows and look through the unique values for each column. While looking through the dataset, we found one column which contained NAN values, so we removed that record. While we were looking at the distinct values each column contains, we came across column namely workclass, occupation and native-country had a category which does essentially represents missing values. The category was marked as ‘ ?’, we replaced the values in these columns with its mode.

We also checked for duplicate rows in the dataset and found about 29 records. We dropped these duplicate records as they will not contribute anything in the database but may induce of adding more weights in the clusters.

We dropped the income column as we do not require it for now for performing the clustering.

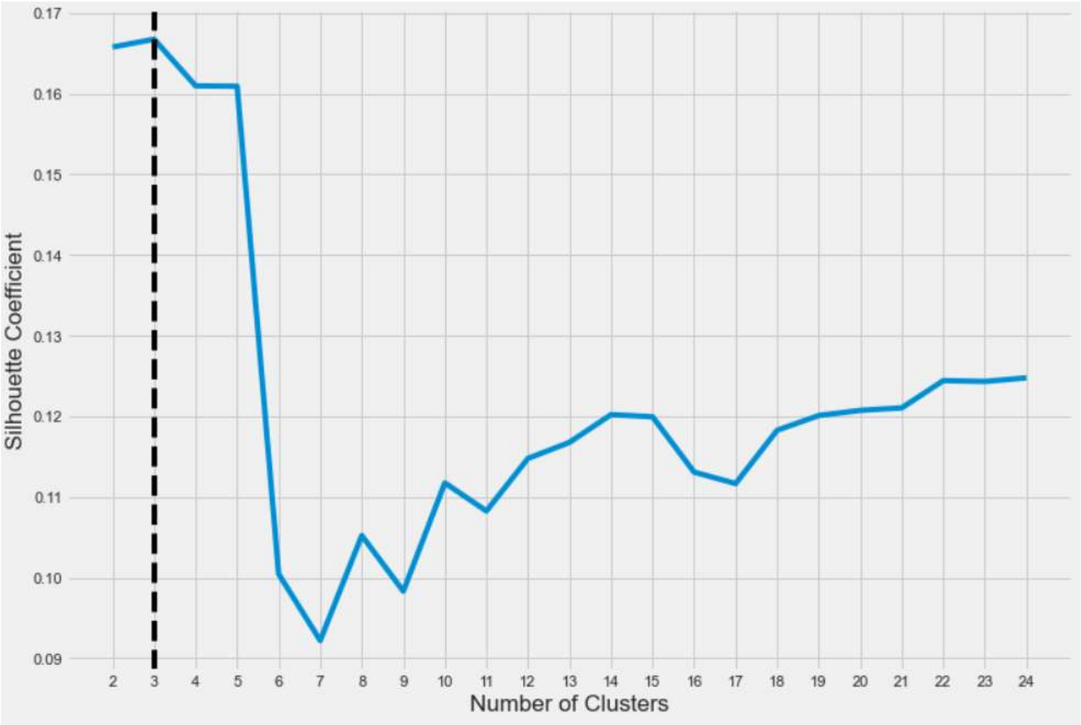
Outlier analysis was also done on the numerical features by plotting box plots. We found out that there were many outliers in all the numerical features. Following are the plots:



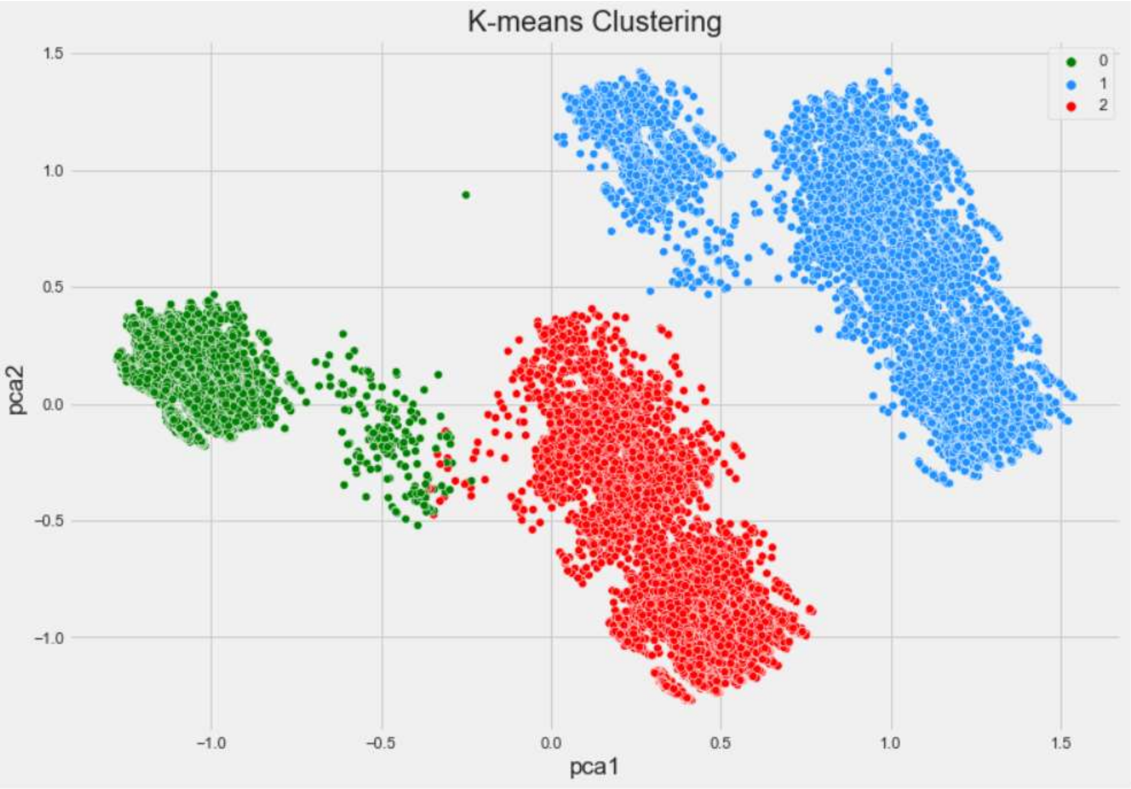
For dealing with outlier, we will be performing min-max normalization on the numerical features so that they will be in range of 0 – 1.

We also performed comparisons between different columns using bar graphs. Please refer to the Jupyter Notebook for detailed analysis.

- Phase-II: This phase comprises of the model building and the evaluating the performance of the model. We are using K-means clustering model and silhouette score for performance metrics. After measuring the performance, we came across the optimal number of clusters are 3. Refer to the following plot:

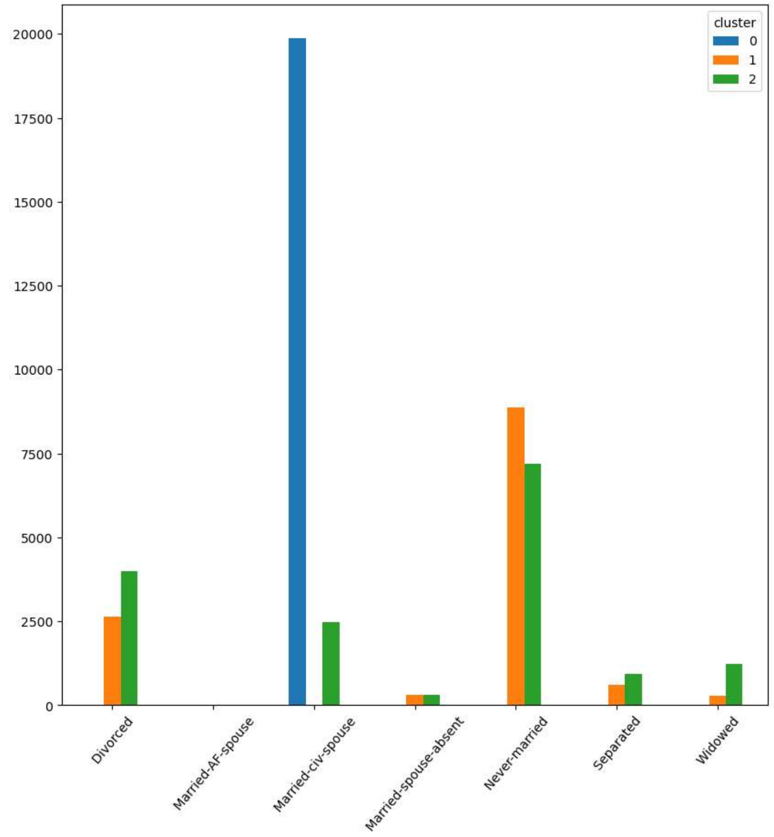


After training the model, we perform PCA on the dataset, so that we can visualize the data points in different clusters. Following is the plot representing the optimal number of cluster

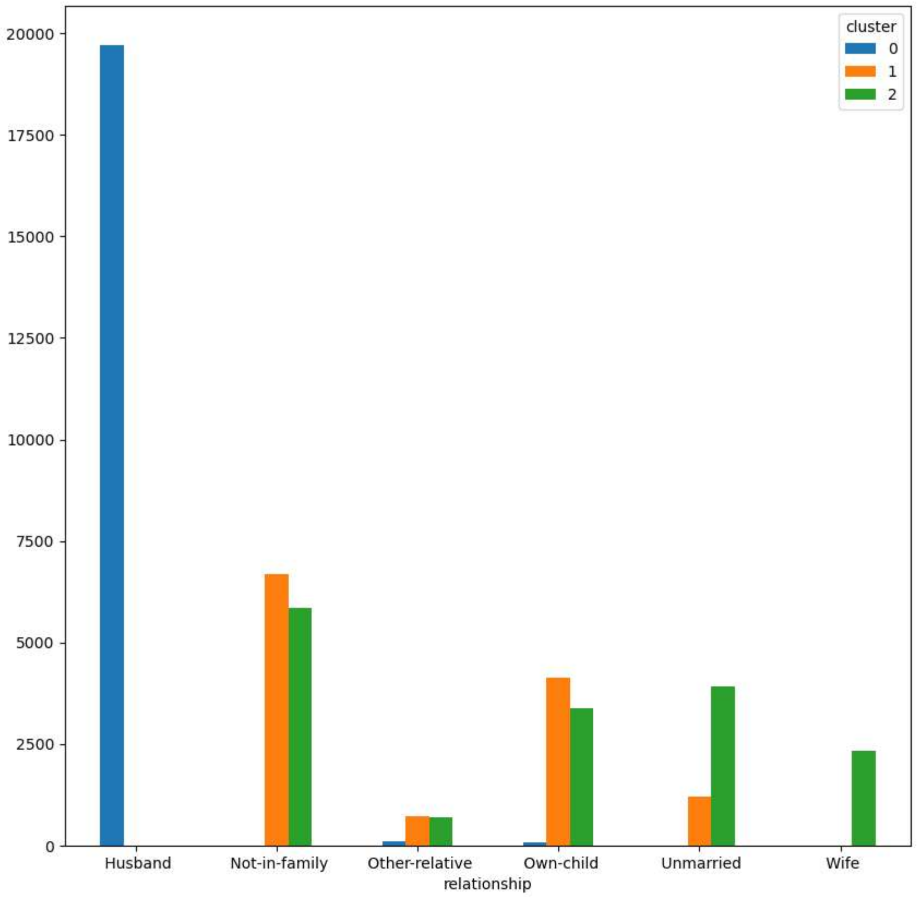


Looking at the graph, we can conclude that the clusters are well defined.

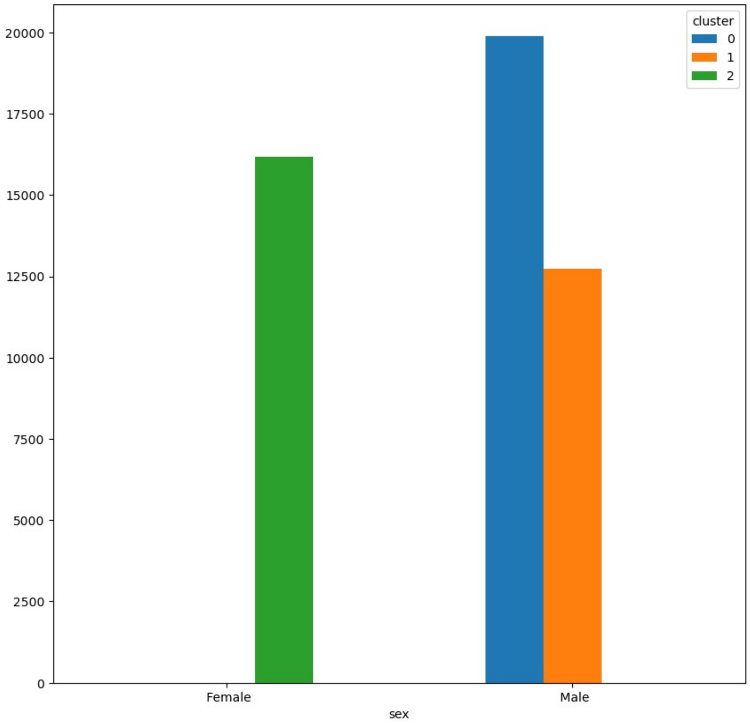
- Phase-III: This phase contains the cluster analysis part. The main observations are recorded here that we came across based on analysis on different clusters. If you want the full observations, please refer to the **Jupyter Notebook**: Following are the major observations:
  - Cluster 0 has the mean age of 43.81 years with median age being 43 years and mode being 38 years.
  - Cluster 0 has the mean capital-gain of 1765.46.
  - Cluster 0 has the mean capital-loss of 121.5.
  - Cluster 0 has the mean hours-per-week being 44 hours with a standard deviation of 11.77 hours
  - Cluster 1 has the mean age of 32 years with median age being 30 years and mode being 23 years.
  - Cluster 1 has the mean capital-gain of 644.23.
  - Cluster 1 has the mean capital-loss of 67.79.
  - Cluster 1 has the mean hours-per-week being 39.7 hours with a standard deviation of 12.16 hours.
  - Cluster 2 has the mean age of 36.9 years with median age being 35 years and mode being 23 years.
  - Cluster 2 has the mean capital-gain of 581.32.
  - Cluster 2 has the mean capital-loss of 61.54.
  - Cluster 2 has the mean hours-per-week being 36.4 hours with a standard deviation of 11.94 hours.



  - No category in martial-status is falling in Cluster 0 except Married-civ-spouse. Married-civ-spouse is divided in Cluster 0 and Cluster 2 with major being in Cluster 0.



- All having relationship as Husband are falling in cluster 0 and all having relationship as wife are falling in cluster 2.



- Male and Female have been divided into 2 parts where all the female are falling in cluster 2 and all the male are falling in cluster 0 and 1. Cluster 0 having more male compared to cluster 1.
- All the clusters are having the same three top countries namely United States, Mexico and Philippines.