**MACHINE LEARNING**

**PROJECT: CUSTOMER SEGMENTATION**

Aaiza Irfan (SP20-BCS-001)

Aliza Tanweer (SP20-BCS-013)

Kulsoom Khurshid (SP20-BCS-044)

**Customer Segmentation**

**Segmentation** refers to grouping entities that have similar properties or features. Entities could be anything. In our project, we are grouping customers on some similar features. For example, you are buying shoes online from any brand. Have you ever thought about who else got the same shoes? Definitely people similar to your age group, interest, gender, etc. We need to have data for grouping the customers based on these similar characteristics to market them effectively.

Data scientists in the companies segment the customers for them before launching a new product in the market. Therefore, this saves time and money as the company invests its energy and markets the product to that specific group of people.

**Dataset**

We have used K-Means and PCA algorithms for this project to group the customers. But before coding, we need to understand the problem and study the dataset that will be used in it.

Variable Description

Following are the features of our dataset.

1. ID is the unique identification number of each customer.
2. Sex refers to the biological gender of the customer. In our dataset,

* 0 is for male
* 1 is for female

1. Marital status of the customer.

* 0 represent single
* 1 represent non-single

1. Age of the customers in years. 18 is the lowest age group in the dataset where 76 is the highest.
2. Education level of the customer. Following numbers represent the levels of education.

* 0 is for other
* 1 for high school
* 2 for under graduate
* 3 for post graduate

1. Annual income in US-dollars of the customer. 35832 is the lowest income in our data whereas 309364 is the highest.
2. Occupation category of customers.

* 0 unemployed
* 1 represent the employed customer
* 2 is for high qualified employee.

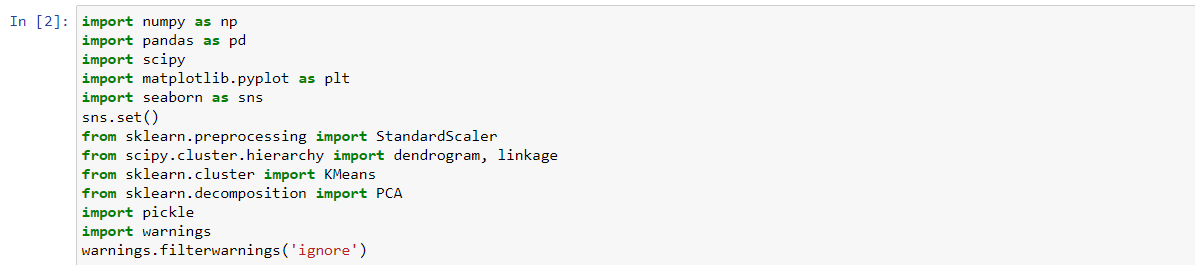
1. The size of the customer settlement where he lives.

* 0 represent small towns
* 1 small cities
* 2 for big cities

**Libraries**

Following are the libraries that we have used in our project.

* Pandas as pd, scipy and numpy as np are used for data analysis and manipulation.
* Matplolib.pyplot as plt and seaborn as sns are used for data visualization. Matplotlib is standard library used by many people whereas seaborn works on top of it.
* Sns.set() is used for styling purposes.
* Sklearn.preprocessing is used for standardizing features. We have used the standardScalar module of it.
* Scipy.cluster.heirarchy is used for hierarchal with the help of Sci py library. Dendrogram and linkage modules are used.
* Sklearn is the most used library for machine learning so we have used k-mean PCA module of it for our project.



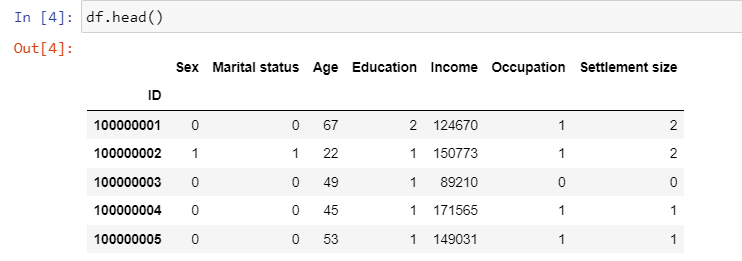
**Import the dataset**

First of all, we need to load the data stored in the CustomerSegmentation.csv file in our project.

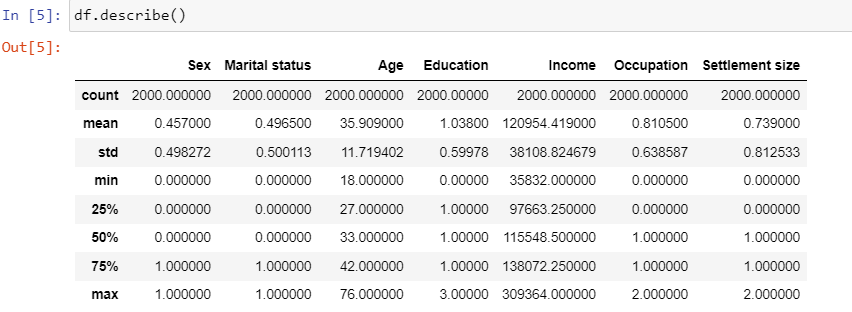


**Explore the dataset**

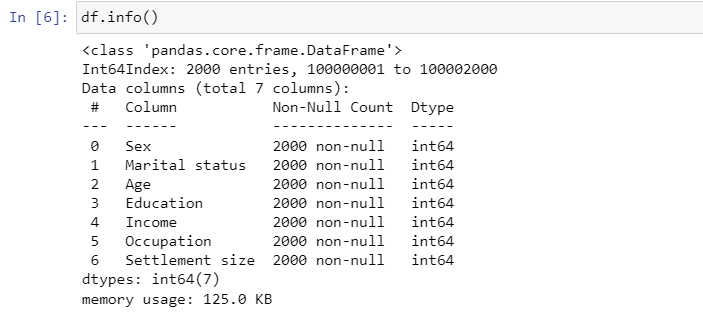
To gain insight into the dataset we need a descriptive analysis of it. No transformation or changes were done to the data.



To know the descriptive statistics of the columns we applied describe method on data. This method is very useful for numerical columns.



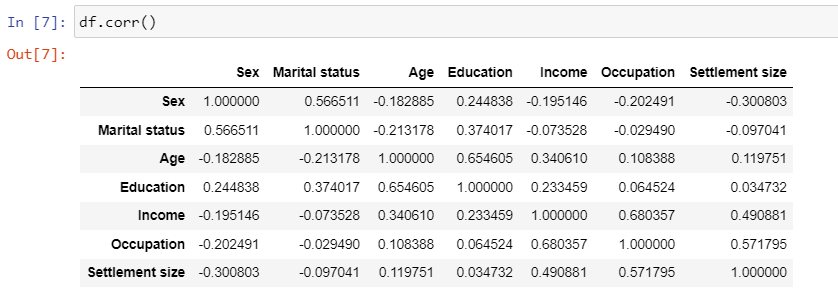
To gain information about the DataFrame that includes the index, data type, columns, memory usage and non-null values we have used the info method over the data.



This shows that no null values are there in our data and all the variables are number.

**Correlation Estimate**

To understand the relationship between the variables we need to find how they correlate for our better initial understanding. Corr method available in pandas library has the Pearson correlation set as default is best for this purpose.

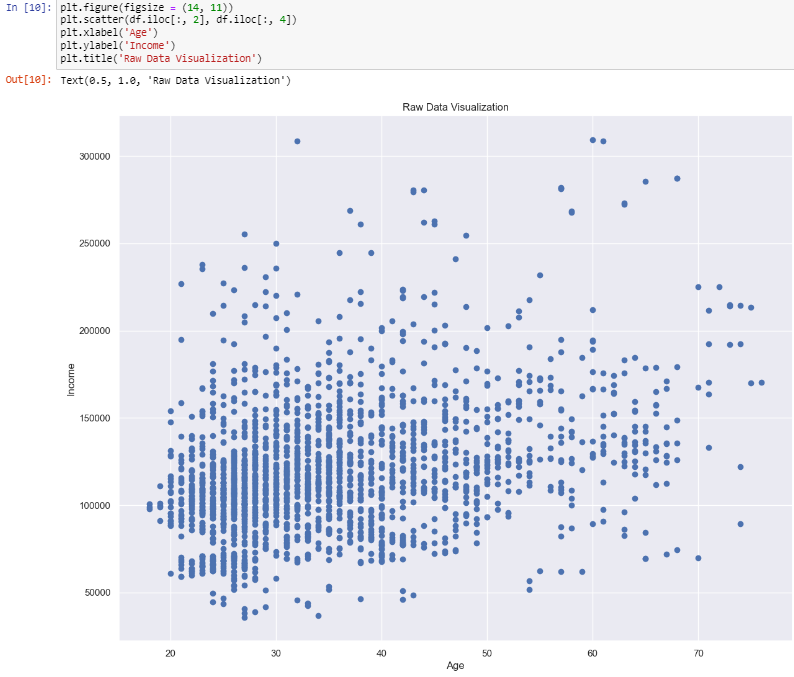




As we can see from the map, there is a very strong correlation of 0.65 between education and age, which means that elderly people are more educated. Similarly, income and occupation have a correlation of 0.68 means more salary is more likely to have a higher level job. This matrix is very important for analyzing the features relationship.

**Raw data visualization**

To plot the data we have a 14 by 11 inches figure that contains the 2000 data points that are scattered across the age and income (since it has the highest correlation value) that is located at position 2 and 4 in our dataset.



**Data Processing**

We have enough data insight so far, so the next step is to process this data. Before applying any distance-based machine learning algorithm such as K-mean, KNN we need to standardize our data. The purpose of doing this is to treat all the features equally by transforming them in a way that their value is in the same numerical range. Otherwise, income would be more important than education but we are not sure about it if it is the case or not. This is known as standardization.

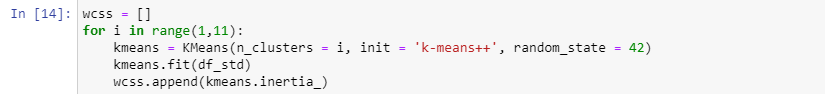
**Standardization**



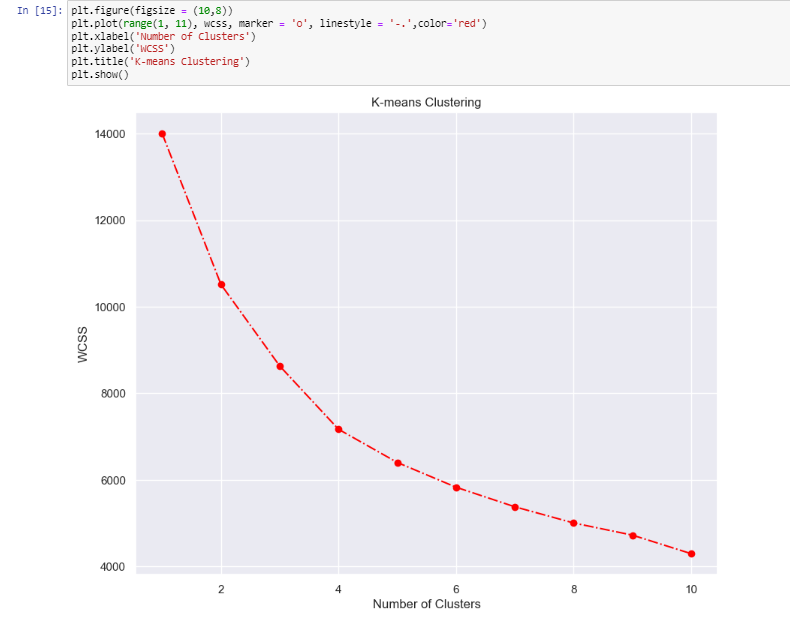


**Segmentation Model**

For selecting the number of clusters, we have two components one is WCSS (Within Clusters Sum of Squares) and Elbow method. Before applying the K-mean algorithm for clustering we have decided that we would consider 1-10 clusters which means for loop will iterate 10 times. The algorithm is executed at different starting points and we have the random state for reproducibility.



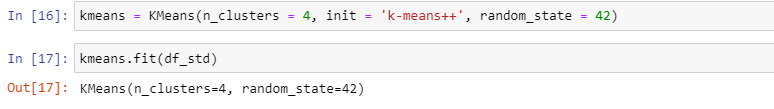
Visualization of WCSS list



As the elbow in the graph is for the four clusters mark, the only place where the graph is steeply declining while smoothing after this point.

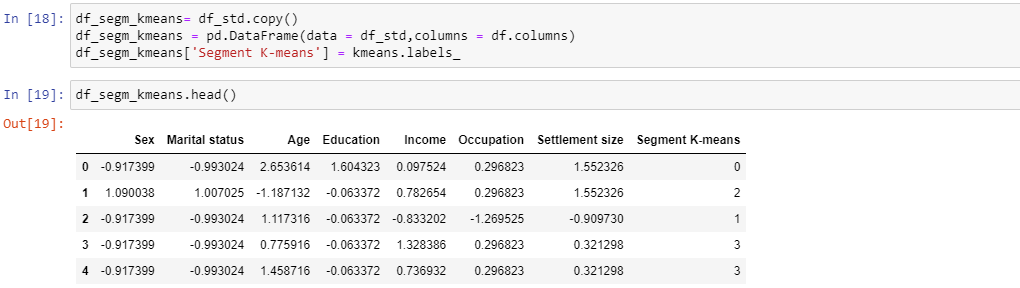
**Model Fitting on Dataset**

Hence we will run K-mean algorithm for four clusters.

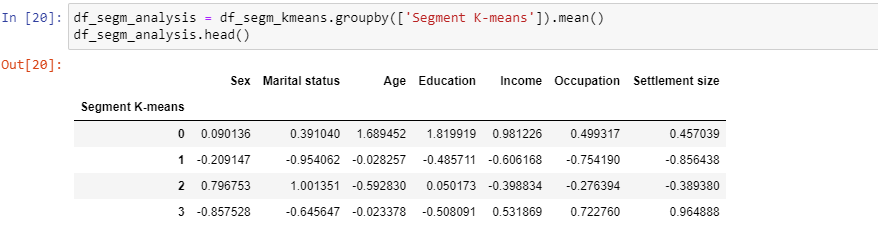


**Results**

A new dataframe is created with original features and adds a new column with the assigned clusters at each point.



Customers are grouped together in clusters and their average values are calculated the mean value for each variable.



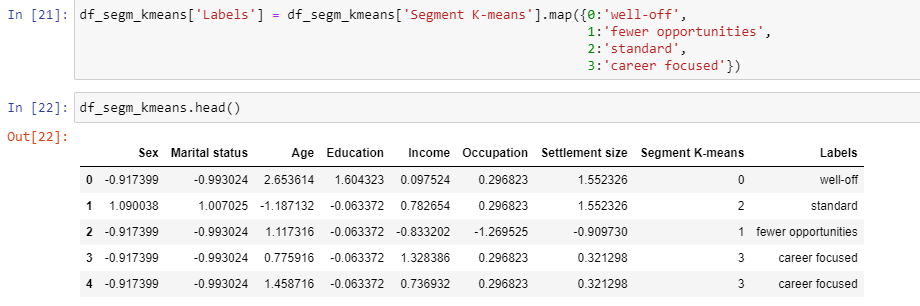
**Interpreting the new dataset**

The first segment is considered the oldest segment as it has almost the same number of men and women with an average age of 56.

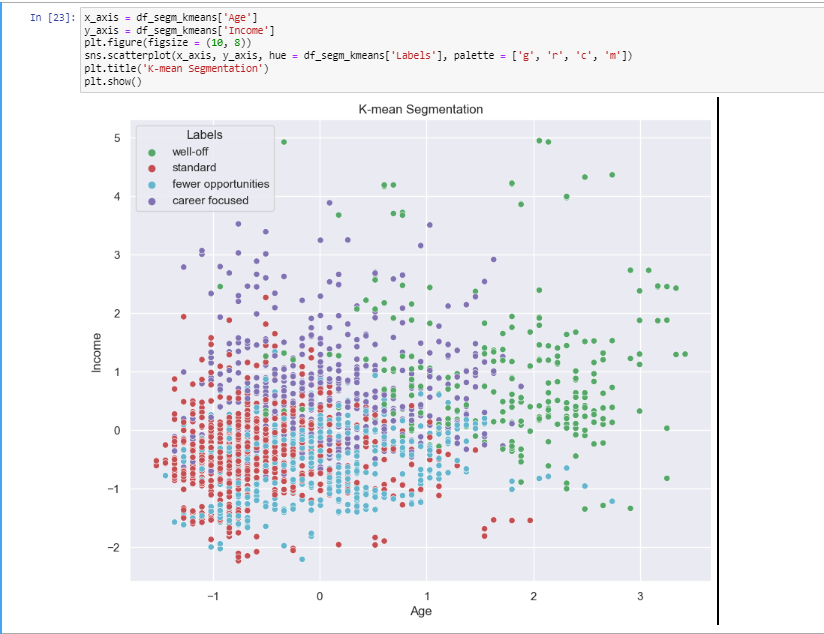
The second segment has the people living in small towns, having low incomes with fewer opportunities.

The third segment can be labeled as standard or average as it has the younger people of an average age of 29 having a medium level of education with an average amount of income.

The fourth and last segment mostly comprises of men out of which 20% of them are in relationships having low value for education and higher value for income and occupation. They lived in small cities or big cities.



K-mean plot for visualization. Each point has its own color that it has been assigned to its cluster



Green color representing the well-off cluster is clearly separated from the rest 3 as it is the highest in terms of age and income. Through K-mean it is hard to separate segments from each other.

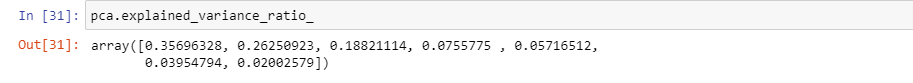
Hence we have used PCA along with it for better understanding.

**PCA with K-mean**

PCA is used to get the subset of components that explained the variance in data. So for that our standardized data was fitted in the PCA. Then we calculated the variance ratio for each of the seven components.

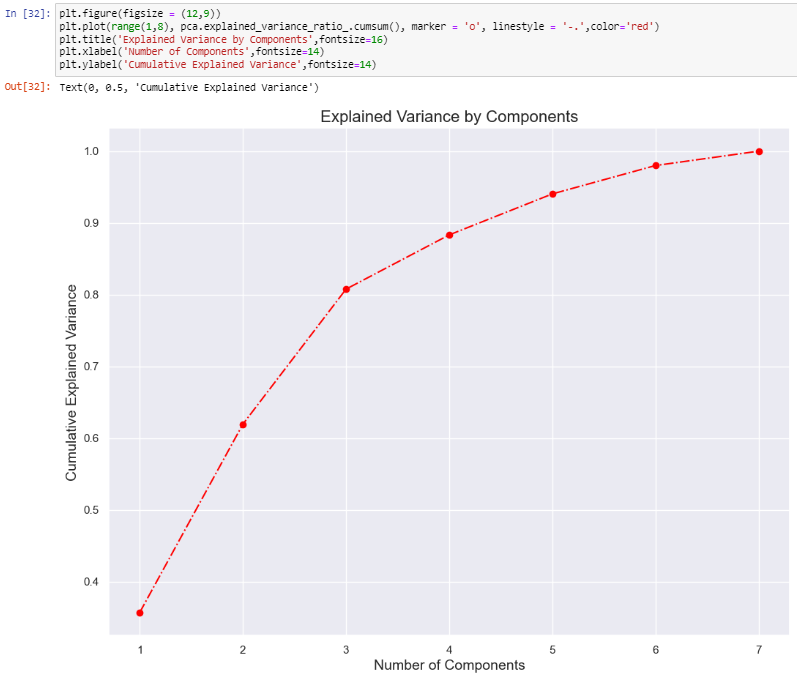






The first component has about 36% of variability of data, second component has 26%, 18% for third component and so on.

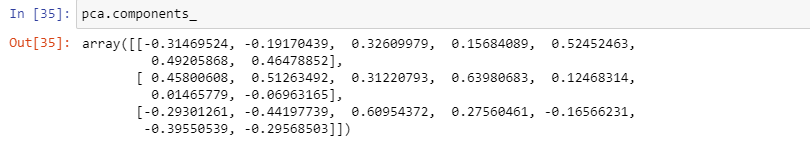
For plotting the cumulative we have considered around 80% of the explained variance and we have only those subset of components that we want to keep. There is no specific method to choose the components. But according to the rule of thumb there should be atleast 70%-80% of the explained variance



As we know that 80% of variance is explained in the first three components, so further analysis will be done on these three components. So we have fitted the data with the selected number of components in the PCA model.



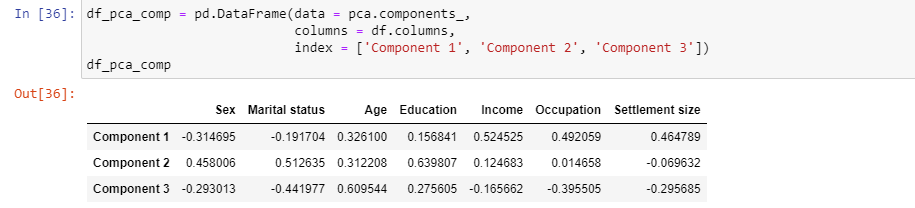
**PCA Result**



These attributes of the component shows the loading of each components on each of the seven original features. Here loading is the correlation between the original features and components.

As in the result we can see it is a 3 by 7 matrix which means that features are reduced from seven original values known as loadings to three. The first value corresponds to the loading of the first feature on the first component.

To view the data nicely we have fitted this information in the pandas data frame where columns refers to the seven original features and rows corresponds to the three components that are given by PCA.



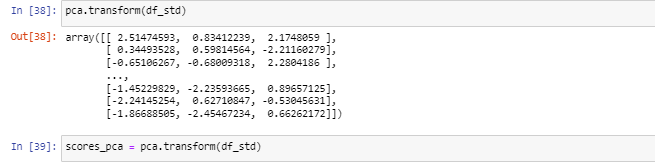
For the heatmap of PCA we have used the RdBU color scheme and set the borders at 1 and -1.

From this map it is visible that there exist a positive correlation between Component1 and income, age, Settlement and occupation. As it is concerned with the career of a person, so this component refers to the career focus of a person.

In second component, education, martial status and sex are the most prominent features.

In the third component, martial status and occupation loads negatively but still they are prominent features of this component along with age.

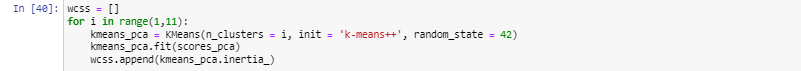




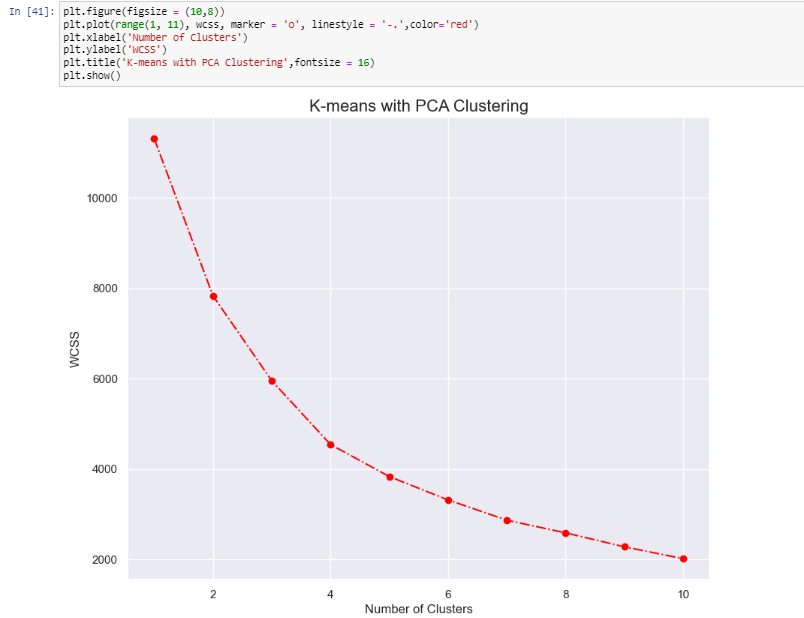
This shows the correlation between the new components and the variables.

**K-Mean Clustering with PCA**

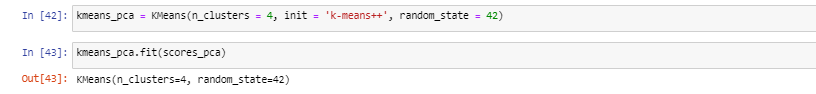
After attaining the new dataset we can apply K-Mean to this dataset having three components. Same steps are followed that we used for standard K-Means.



Plot for WCSS for K-means PCA model is shown below.

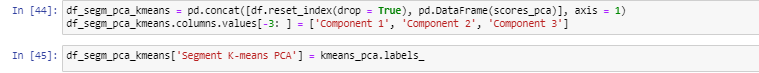


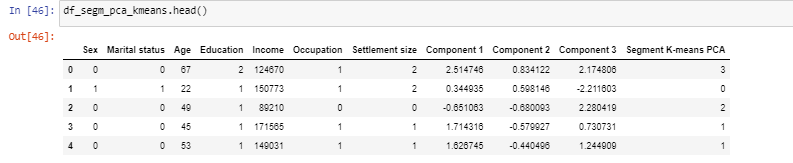
It is evident that four cluster are optimal by WCSS. K-Mean with four number of clusters is executed with the same initializer and random state as it was used.



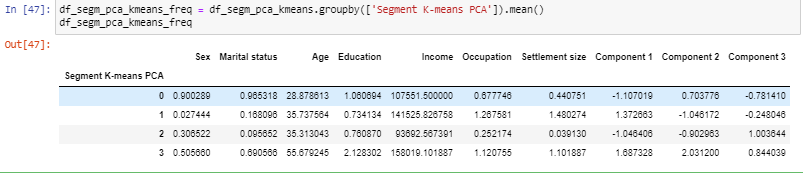
**K-Mean clustering with PCA results**

A new data frame was created with the original features and adds PCA scores to the assigned clusters. The last columns contains the PCA K-mean clustering labels.

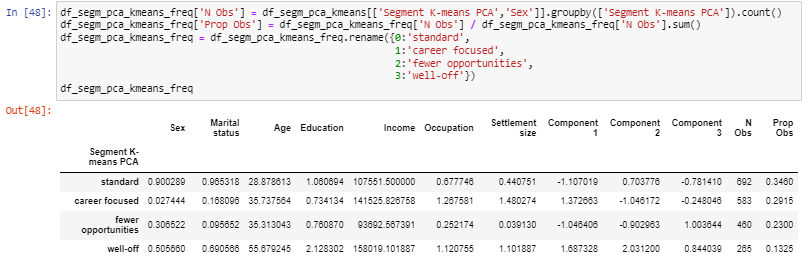




This shows the old dataset along with new components and labels.



In above line of code the data is grouped by K-mean segment but we can also convert the segment numbers as the labels and can view the total observation by seeing each observation and proportion of each segment.



This calculates the size of each cluster and its proportion to the entire dataset.

The previous four clusters were career focused, fewer opportunities, standard and well-off. This remains same.





This shows the data plotted by PCA component. Y axis represent the first component whereas the x-axis represent the second component.

In the above plot where we used only K-mean we were only able to differentiate well-off i.e., the green segment, but the divisions based on the components is much more noticeable. This is what PCA does, reduce the variables into bigger ones.