Census Income Study

Clustering Analysis

Andrew Graham – Fall 2022

UNIVERSITY of DENVER
DANIEL FELIX RITCHIE SCHOOL OF ENGINEERING & COMPUTER SCIENCE

PCA

Build PCA

```
from sklearn.decomposition import PCA
pca = PCA(n_components = 2)
df_PCA = pca.fit_transform(df_x.drop(columns=['ZZ_SPLIT']))
```

✓ 0.6s

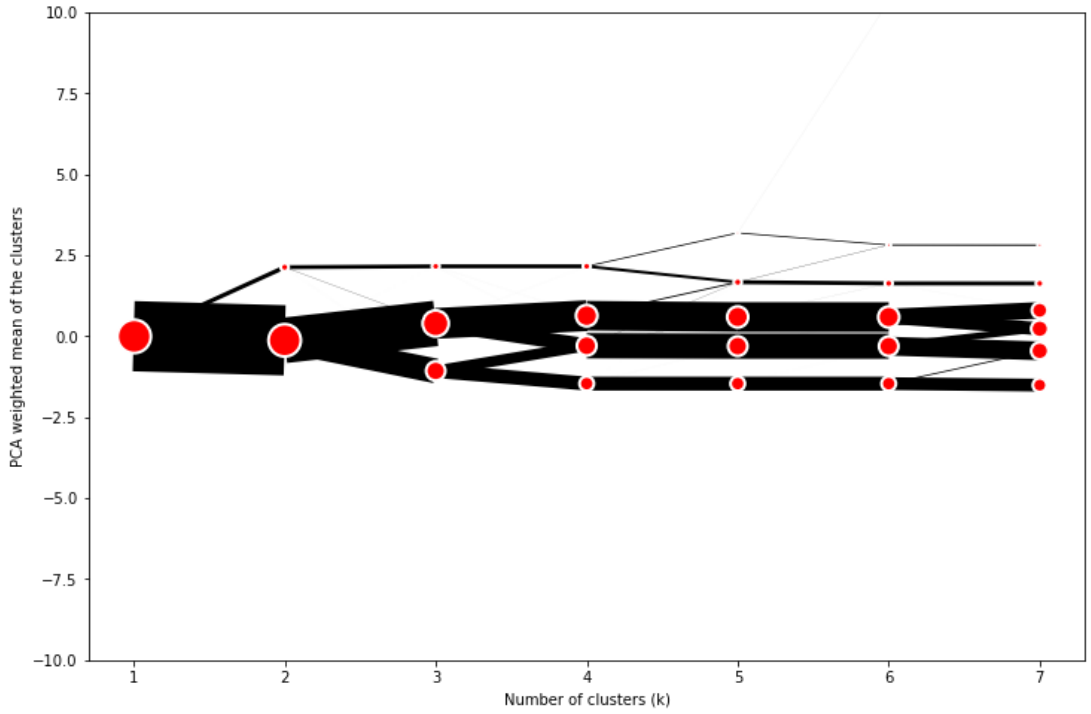
```
df_PCA = pd.DataFrame(df_PCA, columns=['PC1','PC2'])
_PCA
```

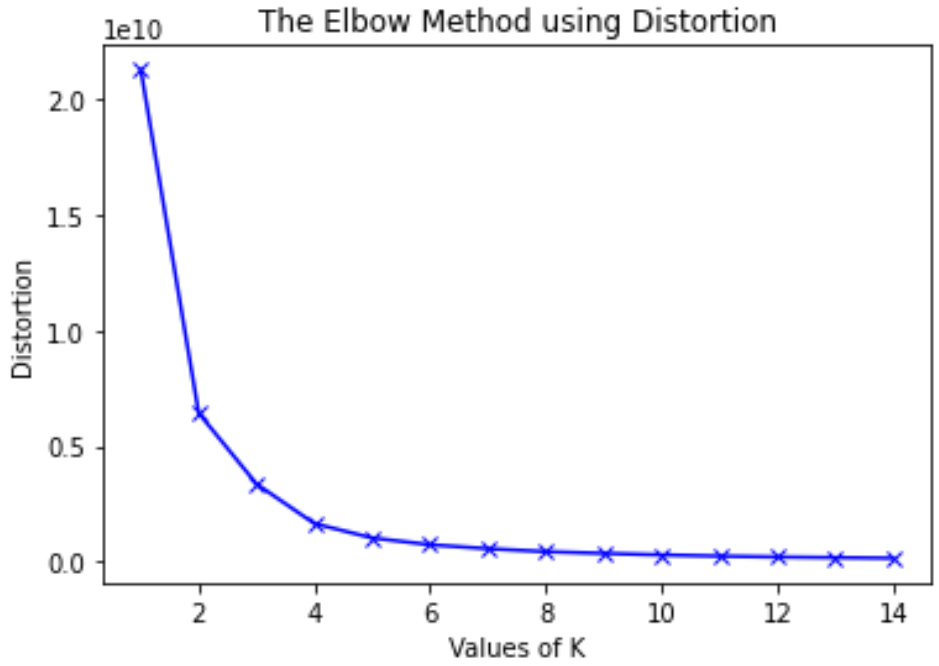
✓ 0.3s

	PC1	PC2
0	-73.241173	38.348610
1	-72.925397	11.461817
2	-73.200980	9.285858

- Clean data from the census dataset used
- Target Income removed from dataset
- PCA used to determine 2 principal components
- The PCA data set will be used for the rest of the analysis

Optimal Clusters





- Dendrogram shows optimal clusters at 4
- Elbow Diagram shows optimal clusters at 4
- 4 clusters will be used.

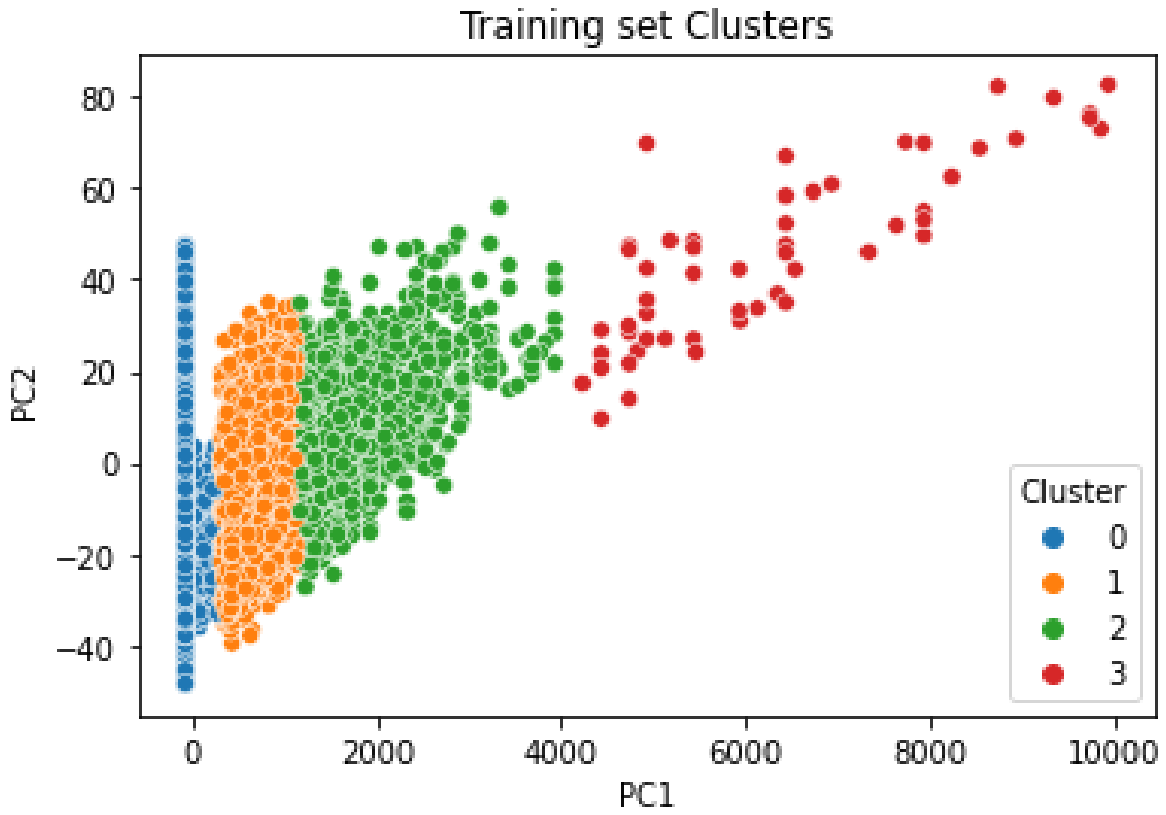
K-Means Clustering

K-Means - 4 clusters

```
# Building the model with 4 clusters
kmean = KMeans(n_clusters=4, init='k-means++', random_state=42)
clusters = kmean.fit_predict(df_PCA)
df_PCA.insert(0, "Cluster", clusters, True)
```

- K-means used with number of clusters equal to 4
- Cluster results added to PCA data as Target

PCA Clustering Plot



What do these Cluster represent?

- As seen from prior slide, the 4 clusters do differ and have recognizable groups.
- 2 clusters in the middle have some merging along the edge so the dividing line could be changed
- The clusters represent similar demographics from the study.
  - A smaller cluster may be Educated white people from the US.
  - These 4 clusters represent similar groups of these smaller clusters

2 Cluster Analysis

```
# Predicting training set
kmean = KMeans(n_clusters=2, init='k-means++', random_state=42)
y_train_pred = 1-kmean.fit_predict(train)


# Predicting test set
kmean = KMeans(n_clusters=2, init='k-means++', random_state=42)
y_test_pred = 1-kmean.fit_predict(test)
```

```
Train acc:  0.8684002056898253
Test acc:  0.8684002056898253
```

- 2 cluster analysis performed to see if there are 2 clusters similar to the income groups
- Accuracy of these clusters is 86%
- Indicates a significant difference between demographics of people making >50k and those making <50k
- Considering other models achieve >90% this model should not be used.