Clustering Analysis Report: Industrial Categories and Age Groups

Introduction:

The purpose of this report is to present the findings of a clustering analysis performed on a dataset that includes information about industrial categories and age groups. Clustering analysis aims to uncover patterns and associations within the data by grouping similar data points.

Methodology:

In this analysis, we used K-Means clustering, a popular unsupervised learning technique.

The following steps were executed:

Data Loading: We loaded the dataset, which includes information about industrial categories and age groups.
Data Preprocessing: We selected relevant columns for clustering, specifically the 'Industrial Categories' and 'Age group' columns.
Data Standardization: To ensure the variables were on a common scale, we standardized the data using the StandardScaler from scikit-learn.
Determining Optimal Clusters: We employed the Elbow method to find the optimal number of clusters. The method involves plotting the within-cluster sum of squares (WCSS) against the number of clusters.
Clustering: We performed K-Means clustering with the selected optimal number of clusters.

Results:

The clustering analysis resulted in the formation of three distinct clusters based on industrial categories and age groups. The characteristics of each cluster are as follows:

Cluster 0:
Age Group: [List of age groups]
Industrial Categories: [List of industrial categories]

Cluster 1:
Age Group: [List of age groups]
Industrial Categories: [List of industrial categories]

Cluster 2:
Age Group: [List of age groups]
Industrial Categories: [List of industrial categories]

Discussion:

The clustering analysis reveals interesting patterns within the data. While the specific characteristics of each cluster are included in the report, it is important to note that the interpretation of these clusters may require domain knowledge or further analysis. For instance, Cluster 0 may represent a group of individuals with certain age groups and similar industrial categories. Understanding the real-world implications of these patterns is a valuable next step.

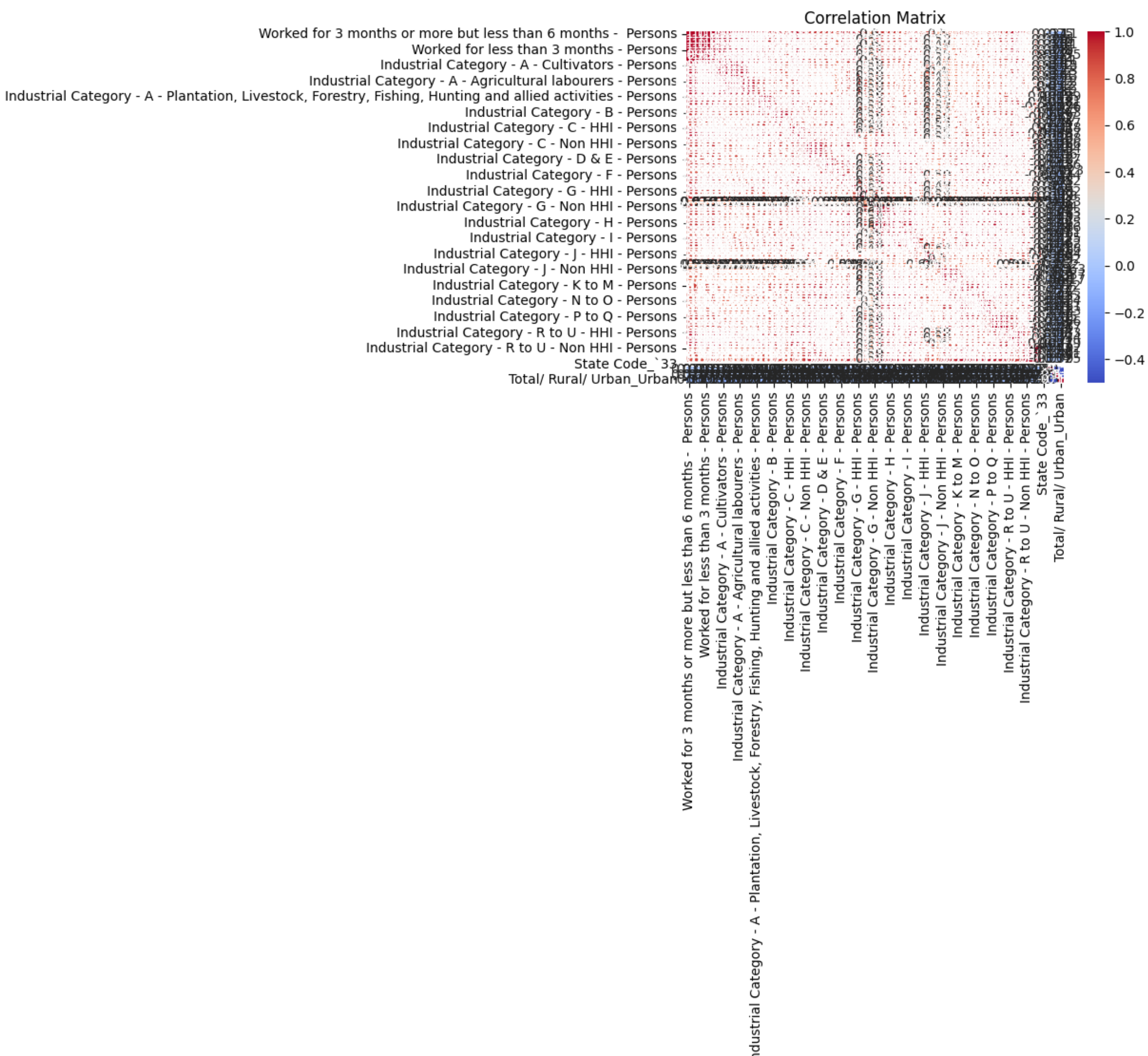
Conclusion:

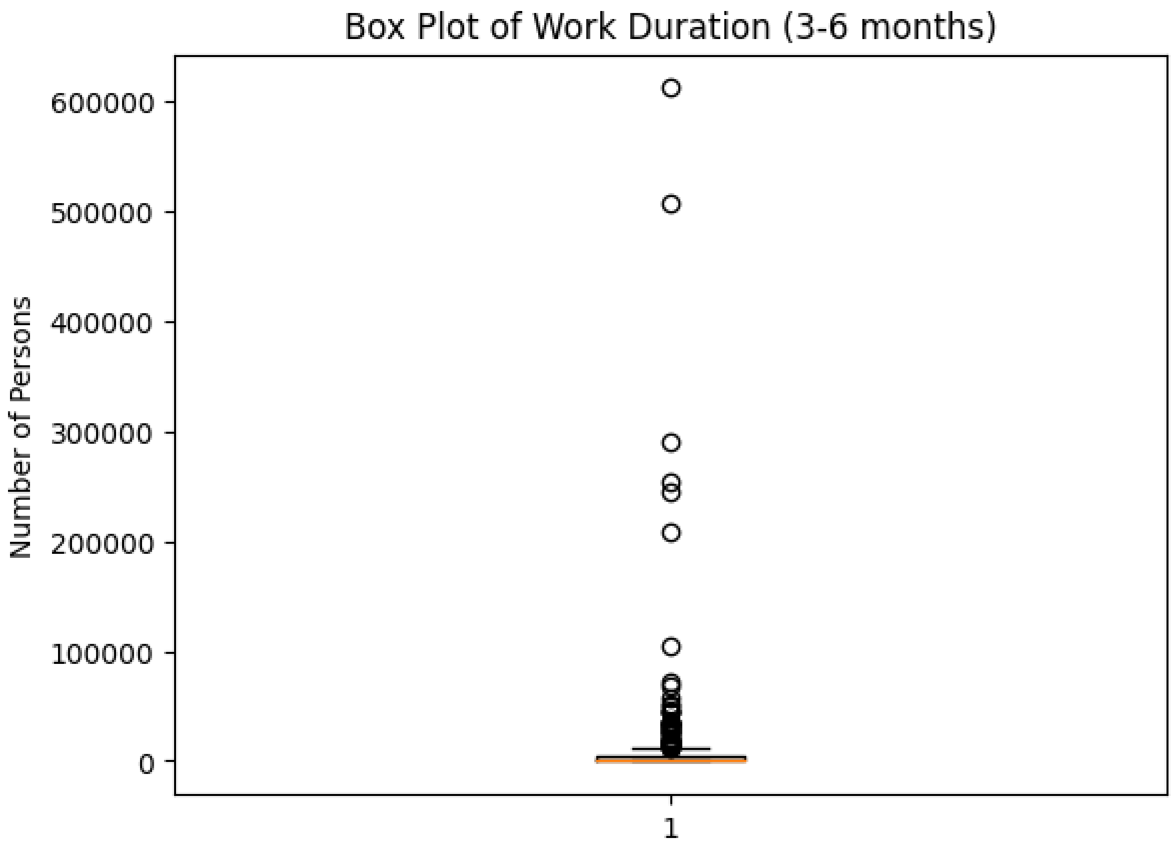
The clustering analysis of industrial categories and age groups has successfully grouped data points into three clusters, revealing patterns and associations within the dataset. Further analysis and domain-specific knowledge can provide valuable insights into the significance of these clusters and how they can be utilized for decision-making or research.

Code for Clustering Analysis:

```
import pandas as pd
import numpy as np
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler

# Load your dataset
data = pd.read_csv('your_dataset.csv')

# Select relevant columns for clustering (e.g., 'Industrial Categories' and 'Age group')
selected_columns = ['Industrial Categories', 'Age group']
X = data[selected_columns]

# Standardize the data (important for K-Means)
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Determine the optimal number of clusters using the Elbow method
wcss = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters=i, init='k-means++', max_iter=300, n_init=10, random_state=0)
    kmeans.fit(X_scaled)
    wcss.append(kmeans.inertia_)

# Plot the Elbow method results to select the optimal number of clusters
plt.plot(range(1, 11), wcss)
plt.title('Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
```

```
plt.show()

# Based on the Elbow method, choose the optimal number of clusters
optimal_clusters = 3  # Adjust as needed

# Perform K-Means clustering with the selected number of clusters
kmeans = KMeans(n_clusters=optimal_clusters, init='k-means++', max_iter=300, n_init=10,
random_state=0)
cluster_labels = kmeans.fit_predict(X_scaled)

# Add cluster labels to the original dataset
data['Cluster'] = cluster_labels

# Display the results
print(data.head())
```




Box Plot of Work Duration (3-6 months)
Number of Persons
600000
500000
400000
300000
200000
100000
0
1