Final Project: New York City Case Study using CRSIP-DM

Report 3: Cluster Analysis

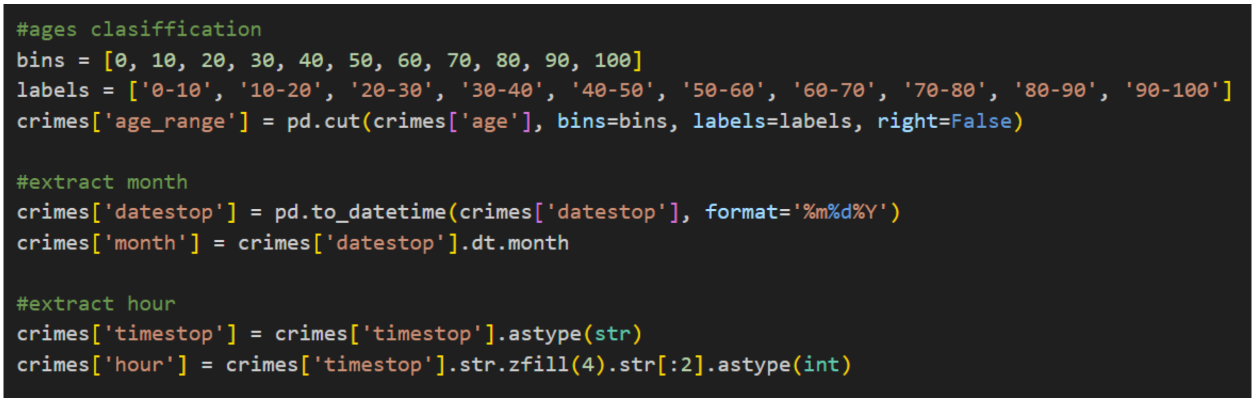
Juan J. Holguin

The Southern Alberta Institute of Technology

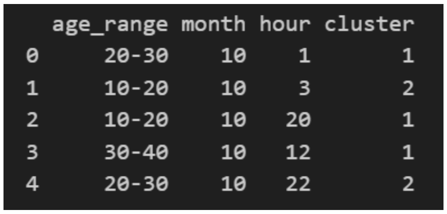
DATA 475: Advanced Concepts in Data Analytics

Data Preparation

The following facts changed to categories



```
#ages clasiffication
bins = [0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100]
labels = ['0-10', '10-20', '20-30', '30-40', '40-50', '50-60', '60-70', '70-80', '80-90', '90-100']
crimes['age_range'] = pd.cut(crimes['age'], bins=bins, labels=labels, right=False)

#extract month
crimes['datestop'] = pd.to_datetime(crimes['datestop'], format='%m%d%Y')
crimes['month'] = crimes['datestop'].dt.month

#extract hour
crimes['timestop'] = crimes['timestop'].astype(str)
crimes['hour'] = crimes['timestop'].str.zfill(4).str[:2].astype(int)
```



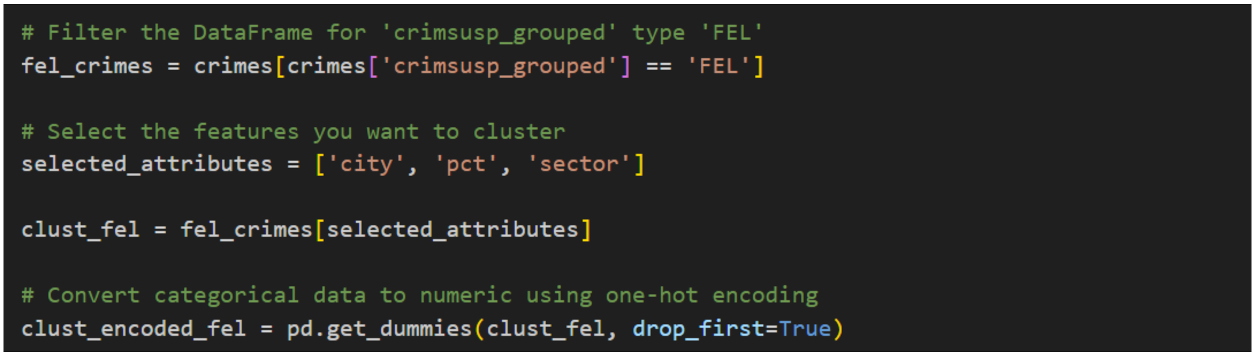
```
  age_range month hour cluster
0     20-30    10    1       1
1     10-20    10    3       2
2     10-20    10   20       1
3     30-40    10   12       1
4     20-30    10   22       2
```

Data modelling

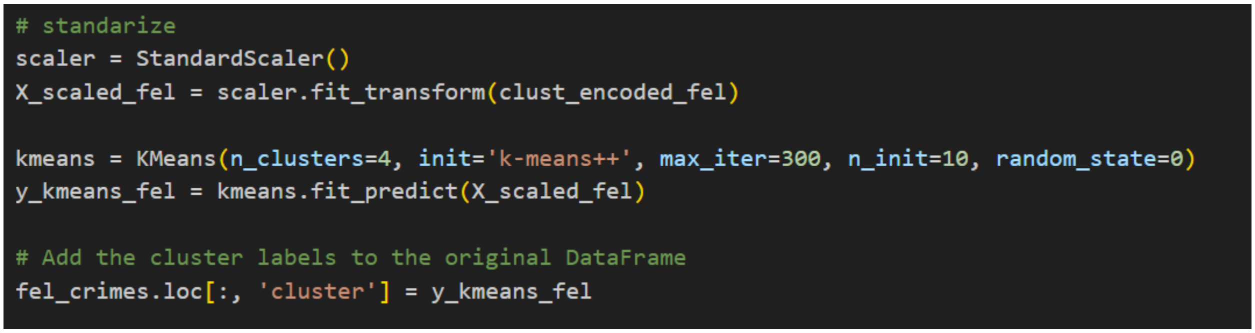
Clustering models to answer business question:

Q1. Cluster location and type of crime:

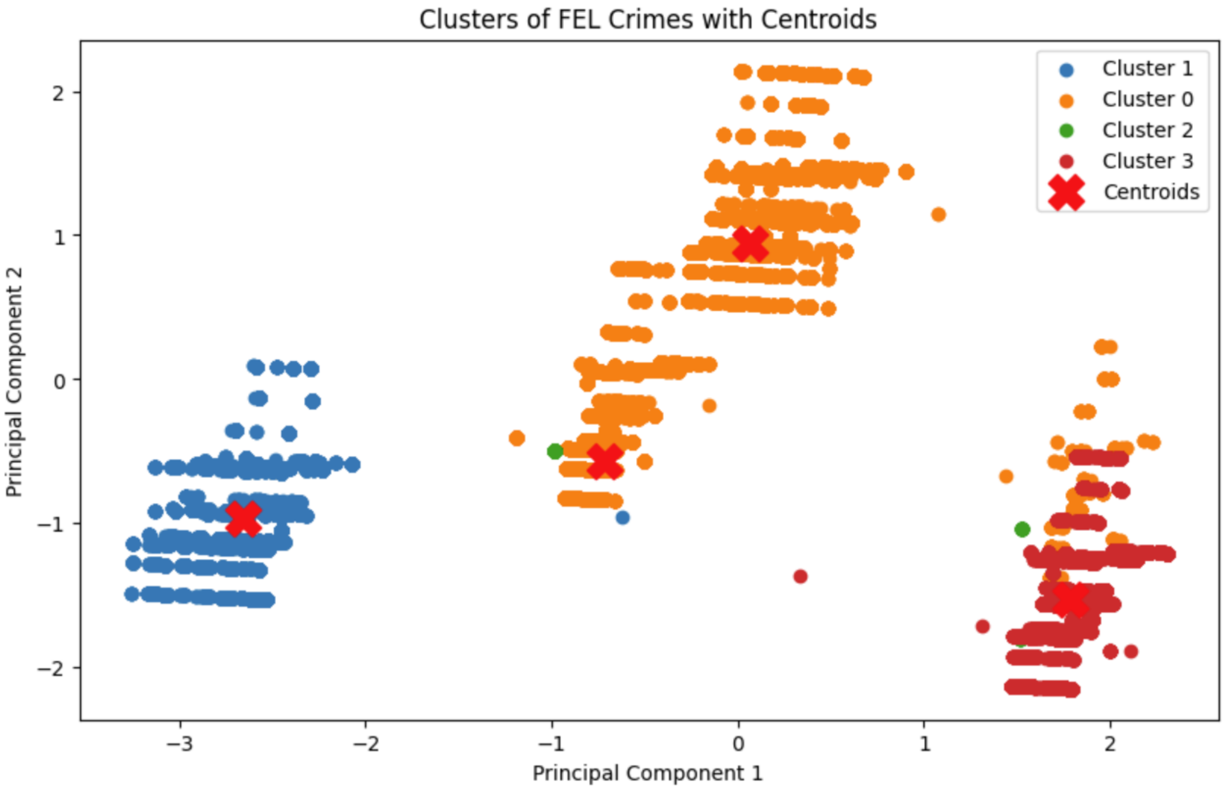
Crime = ‘FEL’ since it is the larger type of crime.



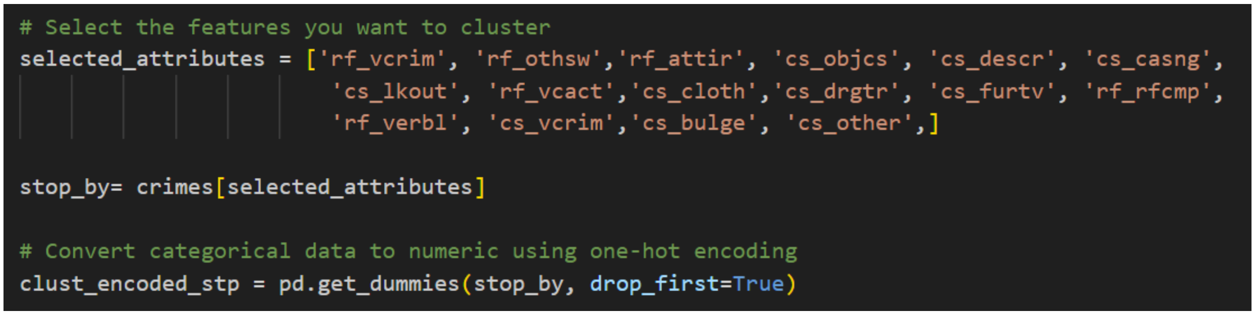
```
# Filter the DataFrame for 'crimsusp_grouped' type 'FEL'
fel_crimes = crimes[crimes['crimsusp_grouped'] == 'FEL']

# Select the features you want to cluster
selected_attributes = ['city', 'pct', 'sector']

clust_fel = fel_crimes[selected_attributes]

# Convert categorical data to numeric using one-hot encoding
clust_encoded_fel = pd.get_dummies(clust_fel, drop_first=True)
```



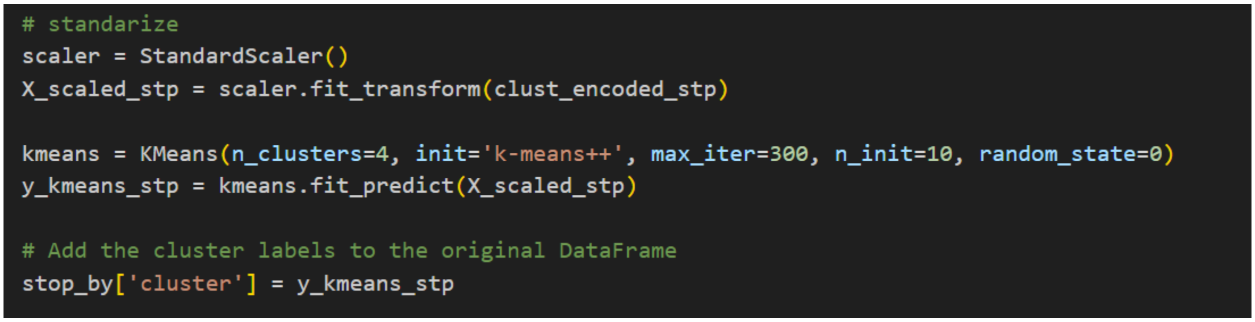
```
# standarize
scaler = StandardScaler()
X_scaled_fel = scaler.fit_transform(clust_encoded_fel)

kmeans = KMeans(n_clusters=4, init='k-means++', max_iter=300, n_init=10, random_state=0)
y_kmeans_fel = kmeans.fit_predict(X_scaled_fel)

# Add the cluster labels to the original DataFrame
fel_crimes.loc[:, 'cluster'] = y_kmeans_fel
```



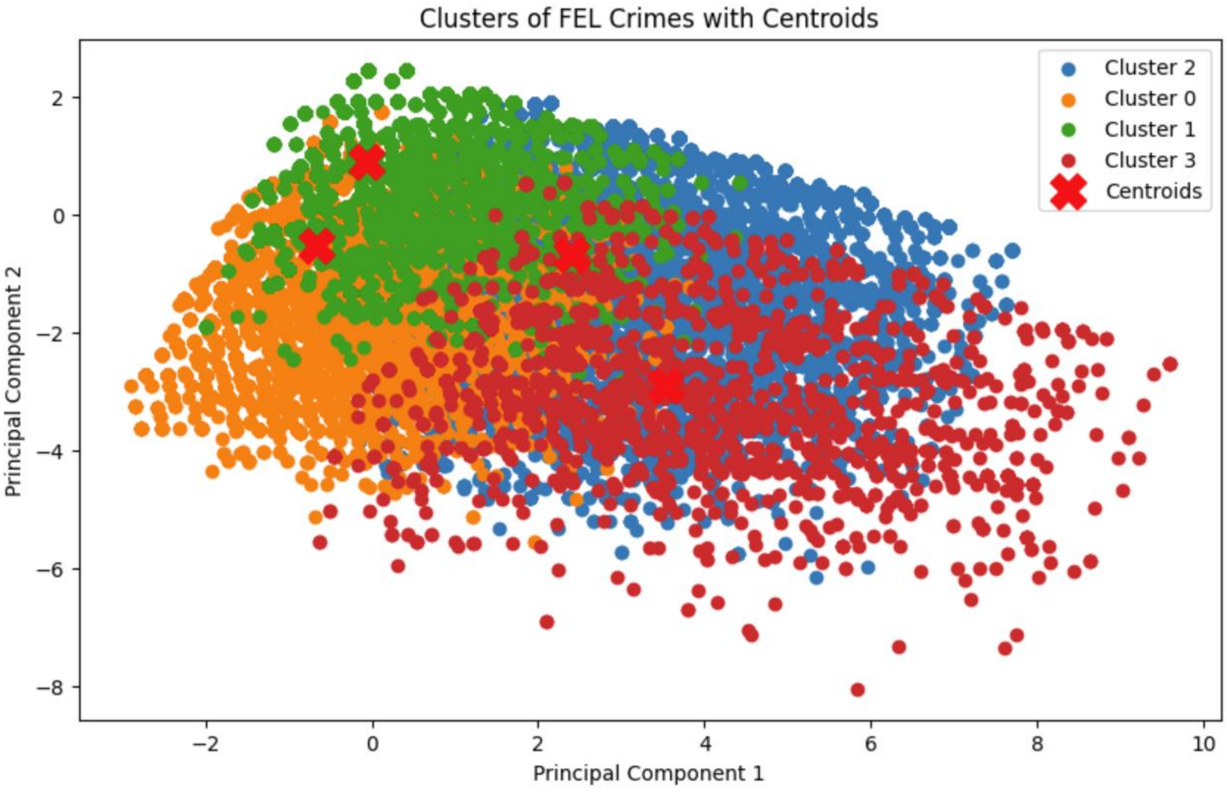
Q2. Stopped people by reason



```python
# Select the features you want to cluster
selected_attributes = ['rf_vcrim', 'rf_othsw','rf_attir', 'cs_objcs', 'cs_descr', 'cs_casng',
                        'cs_lkout', 'rf_vcact','cs_cloth','cs_drgtr', 'cs_furtv', 'rf_rfcmp',
                        'rf_verbl', 'cs_vcrim','cs_bulge', 'cs_other',]

stop_by= crimes[selected_attributes]

# Convert categorical data to numeric using one-hot encoding
clust_encoded_stp = pd.get_dummies(stop_by, drop_first=True)
```

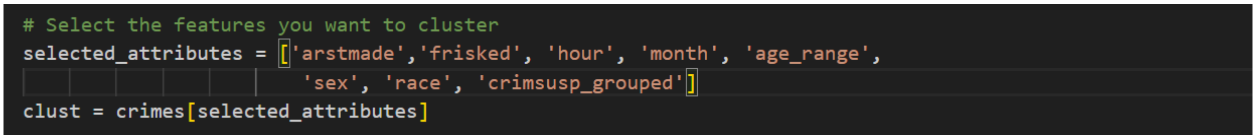


```python
# standarize
scaler = StandardScaler()
X_scaled_stp = scaler.fit_transform(clust_encoded_stp)

kmeans = KMeans(n_clusters=4, init='k-means++', max_iter=300, n_init=10, random_state=0)
y_kmeans_stp = kmeans.fit_predict(X_scaled_stp)

# Add the cluster labels to the original DataFrame
stop_by['cluster'] = y_kmeans_stp
```

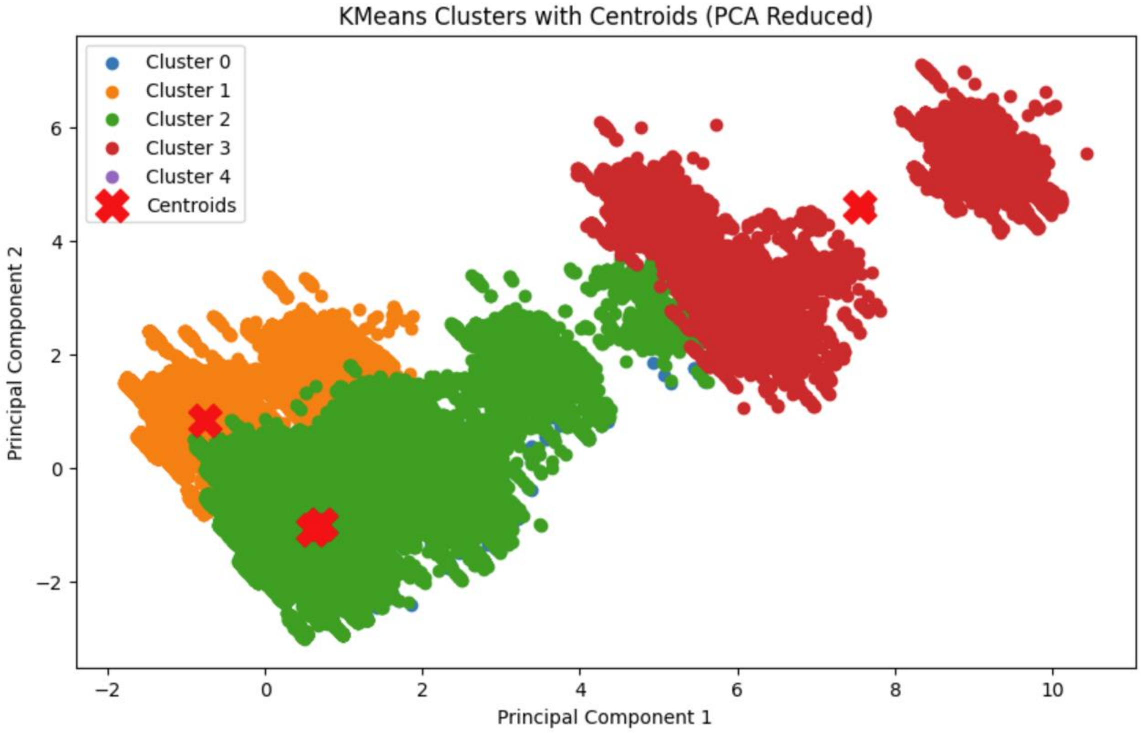


Q3. What else can be clustered

Attributes related to arrests by type of crimes, person and frisked

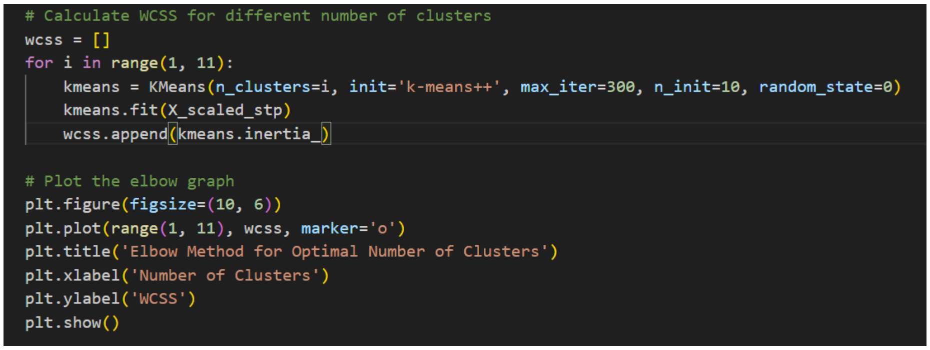


```
# Select the features you want to cluster
selected_attributes = ['arstmade','frisked', 'hour', 'month', 'age_range',
                       'sex', 'race', 'crimsusp_grouped']
clust = crimes[selected_attributes]
```

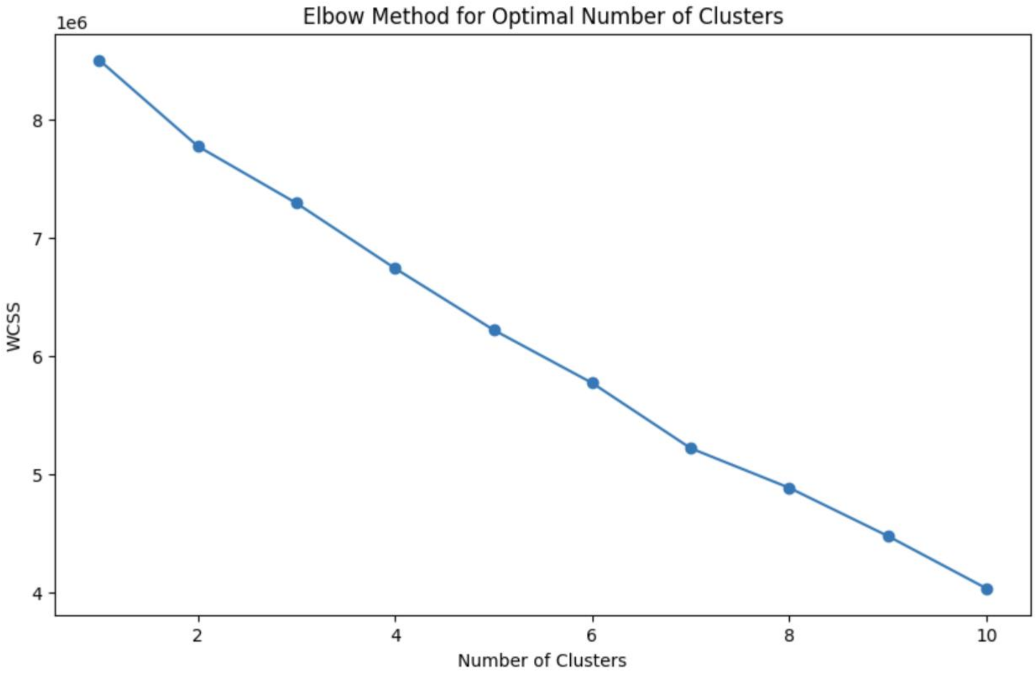


How did you determine a suitable number of clusters for each method?

Elbow method was used to define the number of clusters



```
# Calculate WCSS for different number of clusters
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_stp)
    wcss.append(kmeans.inertia_)

# Plot the elbow graph
plt.figure(figsize=(10, 6))
plt.plot(range(1, 11), wcss, marker='o')
plt.title('Elbow Method for Optimal Number of Clusters')
plt.xlabel('Number of Clusters')
plt.ylabel('WCSS')
plt.show()
```



Not a clear answer. # of clusters were tried by trial.

Evaluation

By clustering data, trends and patterns can be identified. For instance, stops-and-frisks can be prevalent in a certain area, or at certain time. Clusters can be used for predictive analysis to anticipate future incidents. Finally, clusters can highlight areas where additional training might be needed for law enforcement officers. For example, if a cluster shows a high rate of stops based on certain characteristics, it might indicate a need for training on bias and profiling.