Gun Incidents in the USA

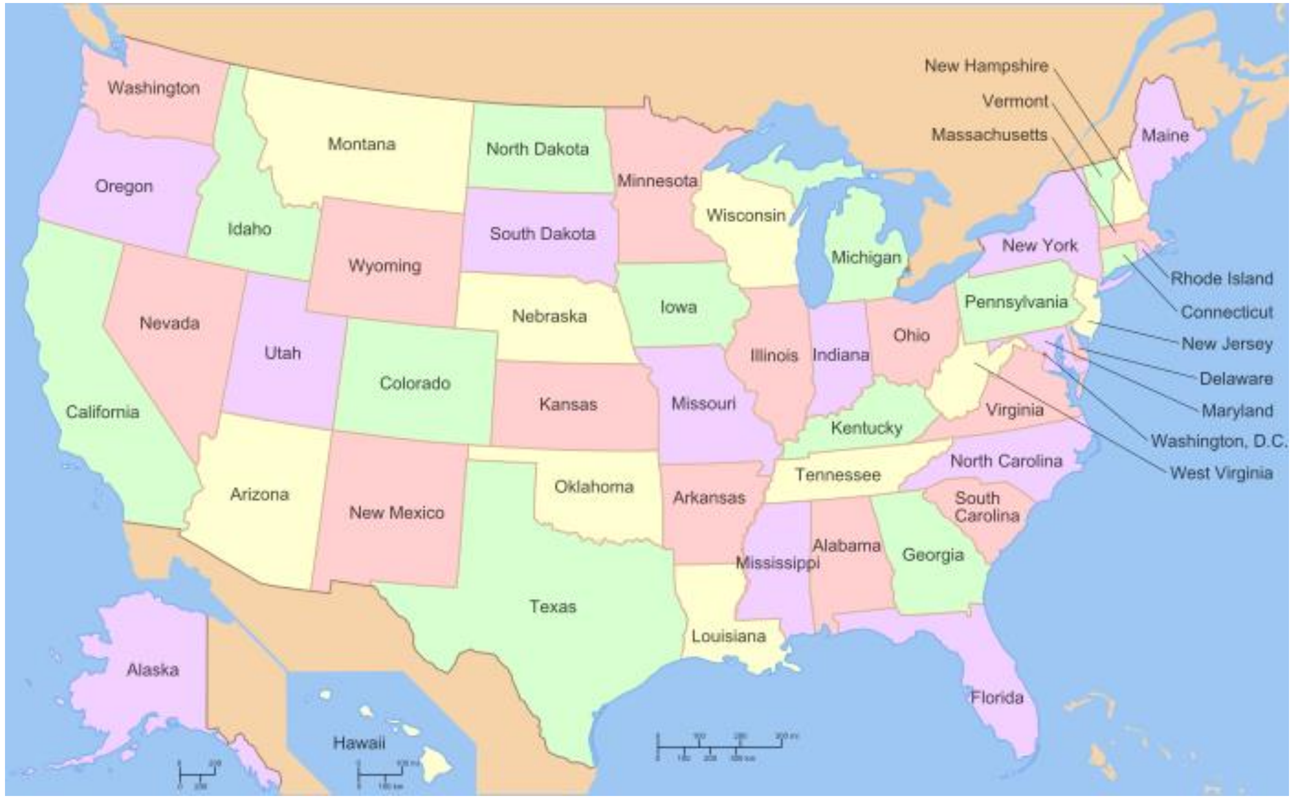
Data Mining, Università di Pisa

24/01/2024

Marc Llobera Villalonga & Patxi Juaristi Pagegi

Index

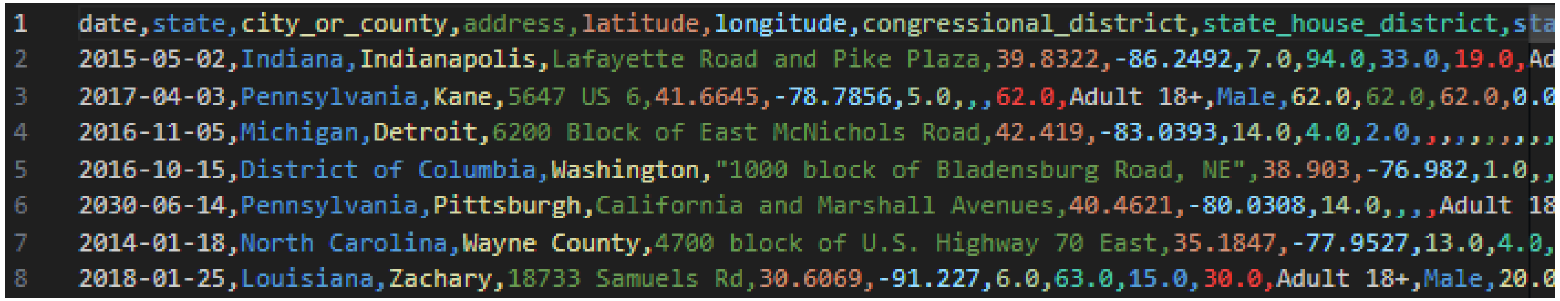
1. INTRODUCTION



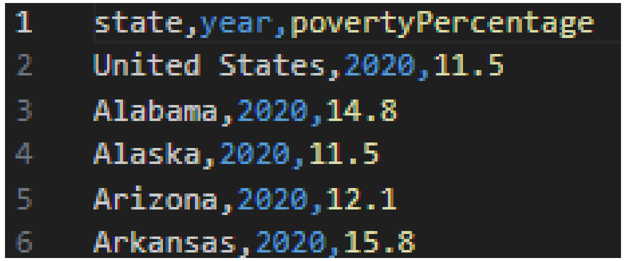
2. DATA UNDERSTANDING AND PREPARATION

Datasets:

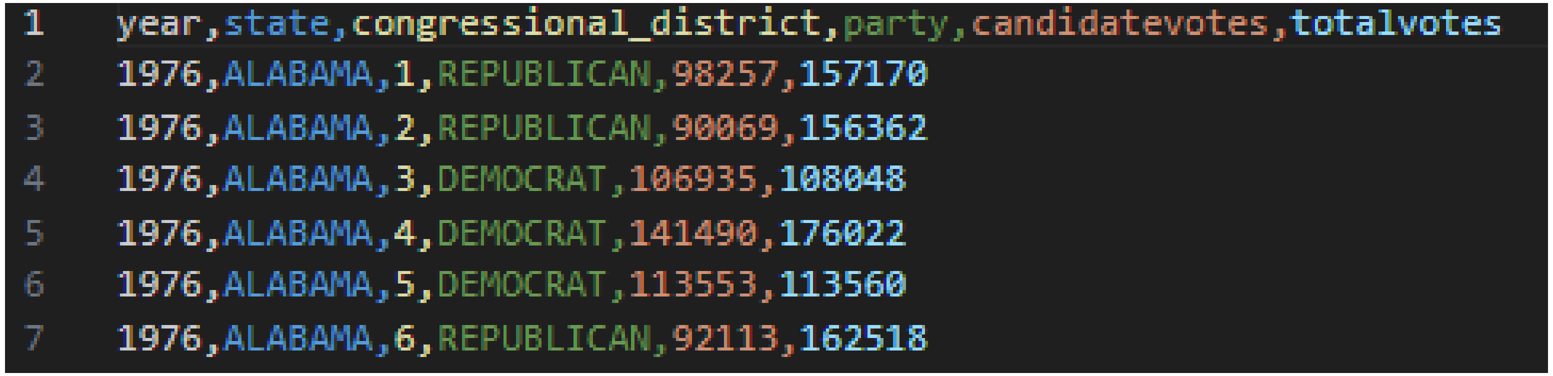
- Incidents Dataset
- Poverty rates dataset
- Congressional elections dataset



```
date,state,city_or_county,address,latitude,longitude,congressional_district,state_house_district,sta
2015-05-02,Indiana,Indianapolis,Lafayette Road and Pike Plaza,39.8322,-86.2492,7.0,94.0,33.0,19.0,Ad
2017-04-03,Pennsylvania,Kane,5647 US 6,41.6645,-78.7856,5.0,,,62.0,Adult 18+,Male,62.0,62.0,62.0,0.0
2016-11-05,Michigan,Detroit,6200 Block of East McNichols Road,42.419,-83.0393,14.0,4.0,2.0,,,,,,,,,,
2016-10-15,District of Columbia,Washington,"1000 block of Bladensburg Road, NE",38.903,-76.982,1.0,,
2030-06-14,Pennsylvania,Pittsburgh,California and Marshall Avenues,40.4621,-80.0308,14.0,,,,Adult 18
2014-01-18,North Carolina,Wayne County,4700 block of U.S. Highway 70 East,35.1847,-77.9527,13.0,4.0,
2018-01-25,Louisiana,Zachary,18733 Samuels Rd,30.6069,-91.227,6.0,63.0,15.0,30.0,Adult 18+,Male,20.0
```



```
state,year,povertyPercentage
United States,2020,11.5
Alabama,2020,14.8
Alaska,2020,11.5
Arizona,2020,12.1
Arkansas,2020,15.8
```

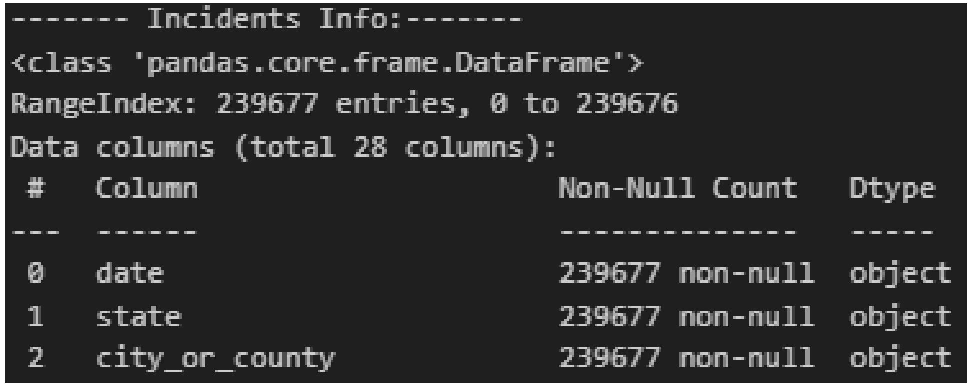


```
year,state,congressional_district,party,candidatevotes,totalvotes
1976,ALABAMA,1,REPUBLICAN,98257,157170
1976,ALABAMA,2,REPUBLICAN,90069,156362
1976,ALABAMA,3,DEMOCRAT,106935,108048
1976,ALABAMA,4,DEMOCRAT,141490,176022
1976,ALABAMA,5,DEMOCRAT,113553,113560
1976,ALABAMA,6,REPUBLICAN,92113,162518
```

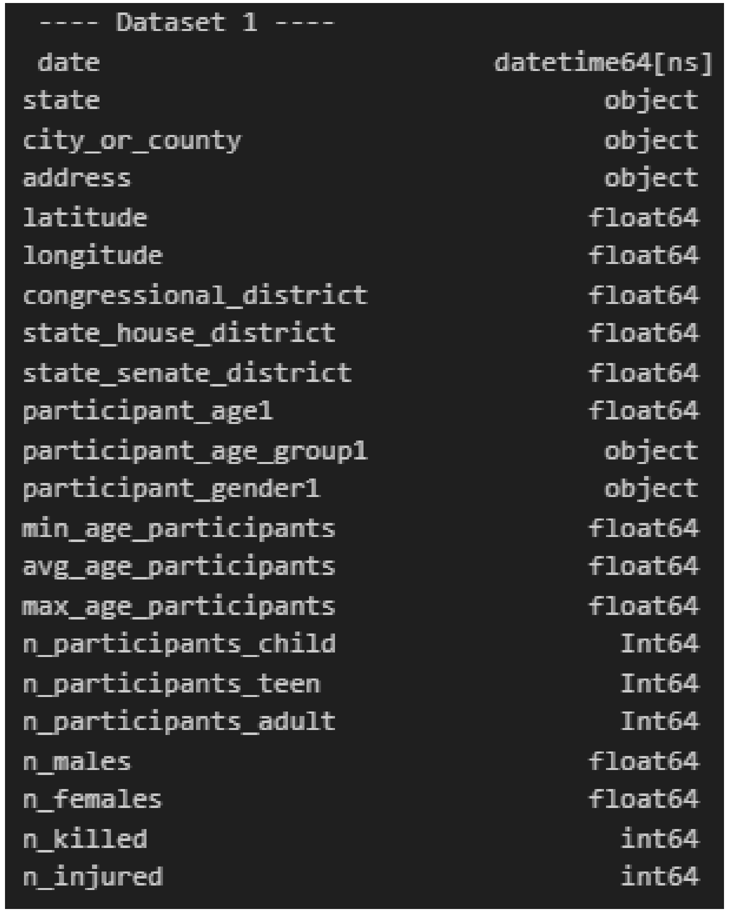
2.1 Data Understanding

info()
describe()
head()

Datetime format
Numeric format

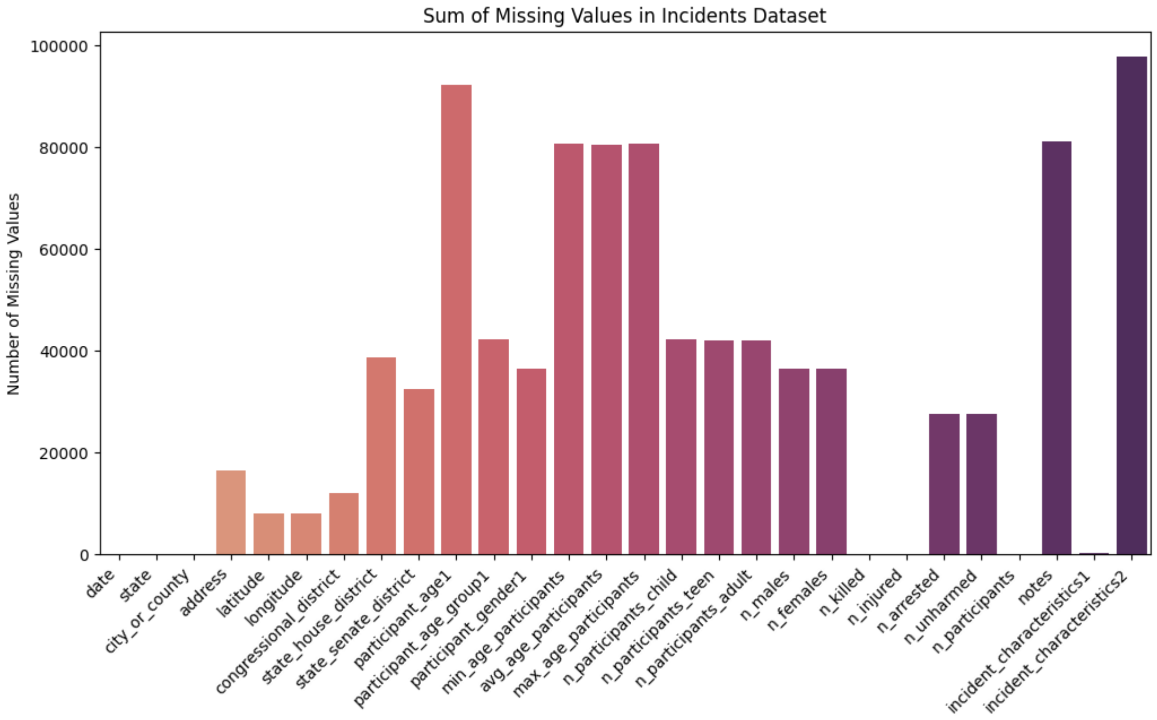


```
------- Incidents Info:-------
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 239677 entries, 0 to 239676
Data columns (total 28 columns):
 #   Column                     Non-Null Count   Dtype
---  ------                     --------------   -----
 0   date                       239677 non-null  object
 1   state                      239677 non-null  object
 2   city_or_county             239677 non-null  object
```

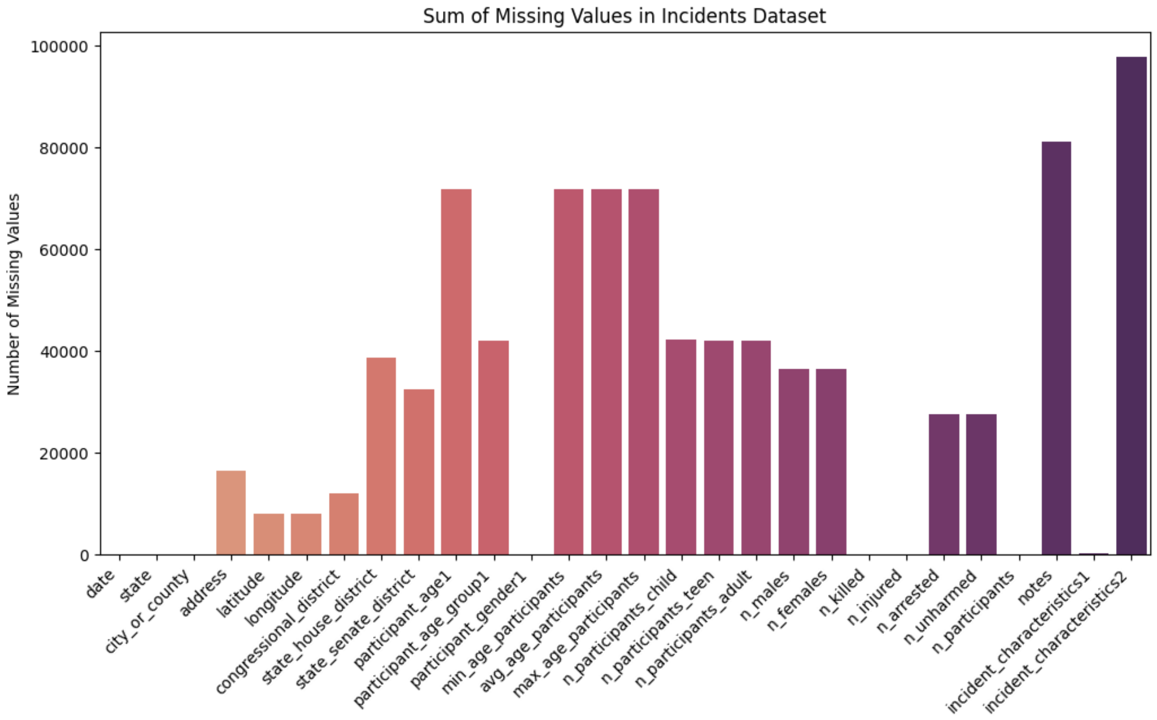


```
 ---- Dataset 1 ----
 date                          datetime64[ns]
state                                 object
city_or_county                        object
address                               object
latitude                             float64
longitude                            float64
congressional_district               float64
state_house_district                 float64
state_senate_district                float64
participant_age1                     float64
participant_age_group1                object
participant_gender1                   object
min_age_participants                 float64
avg_age_participants                 float64
max_age_participants                 float64
n_participants_child                   Int64
n_participants_teen                    Int64
n_participants_adult                   Int64
n_males                              float64
n_females                            float64
n_killed                               int64
n_injured                              int64
```

2.1.1 Data quality assessment



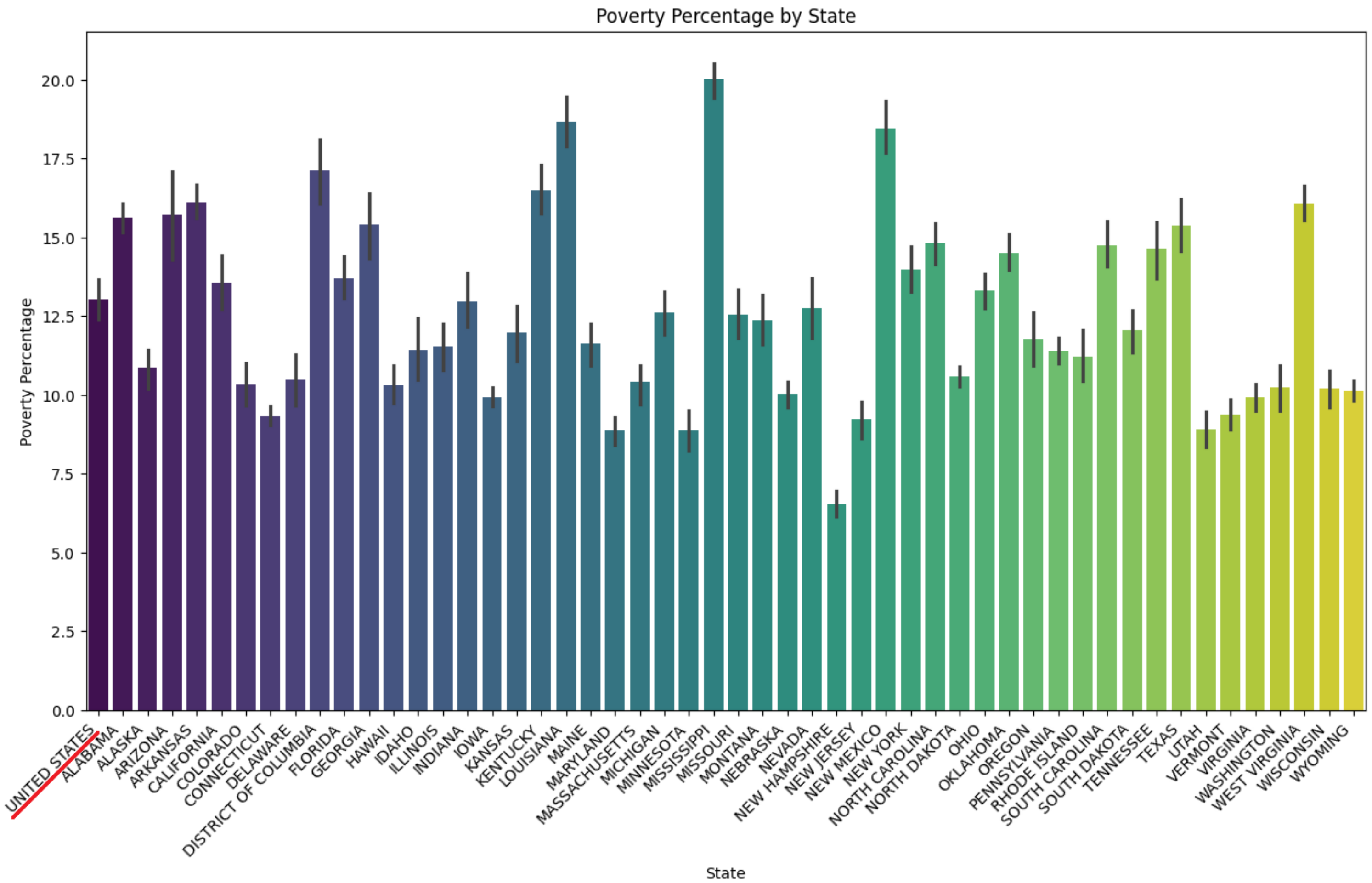
Sum of Missing Values in Incidents Dataset



Sum of Missing Values in Incidents Dataset

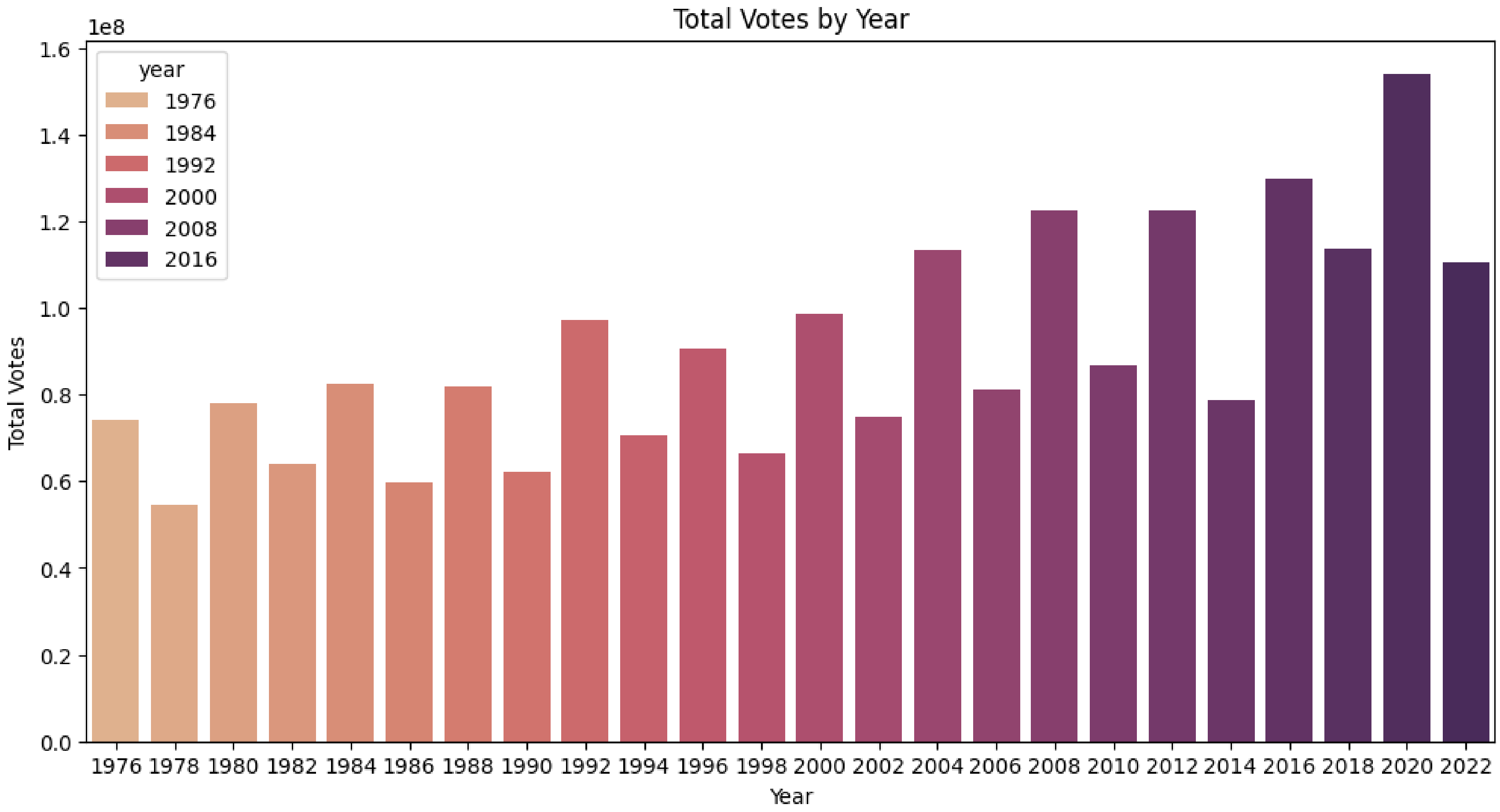
2.1.2 Distribution of variables

Evaluation of poverty percentages dataset



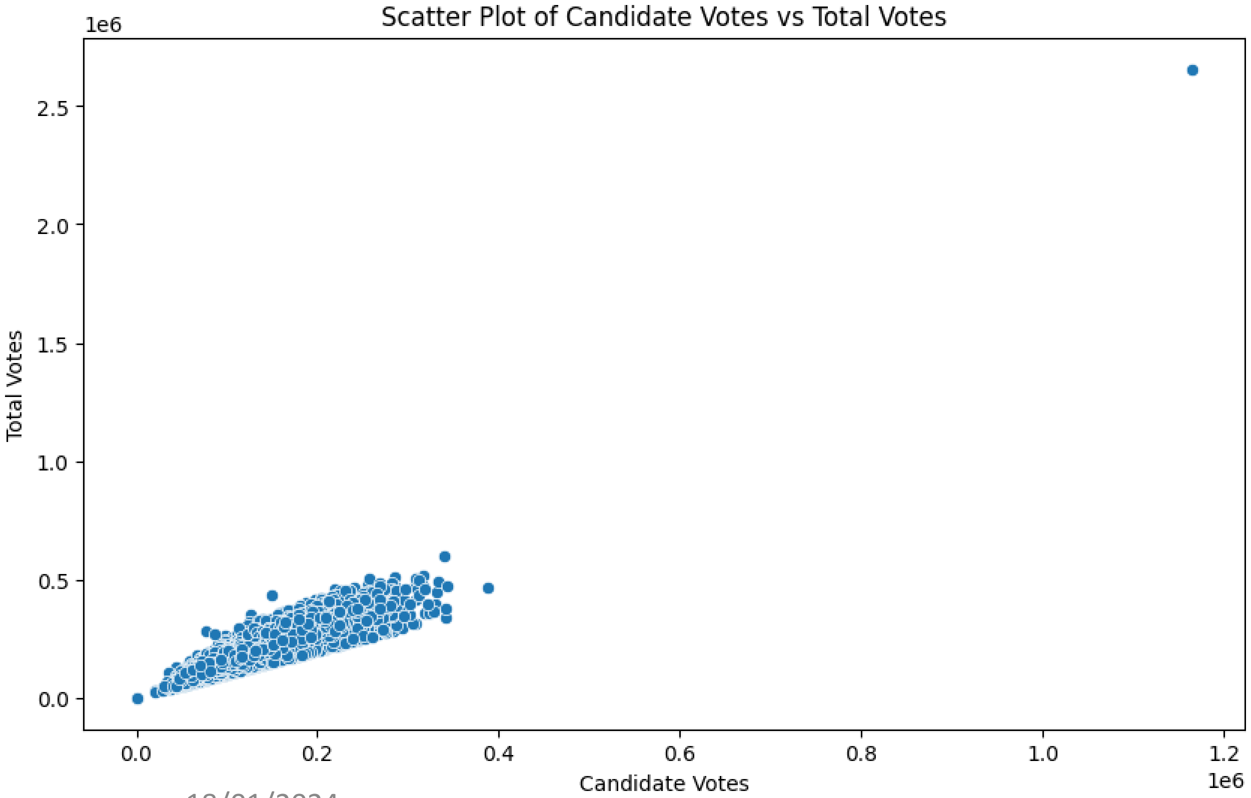
2.1.2 Distribution of variables

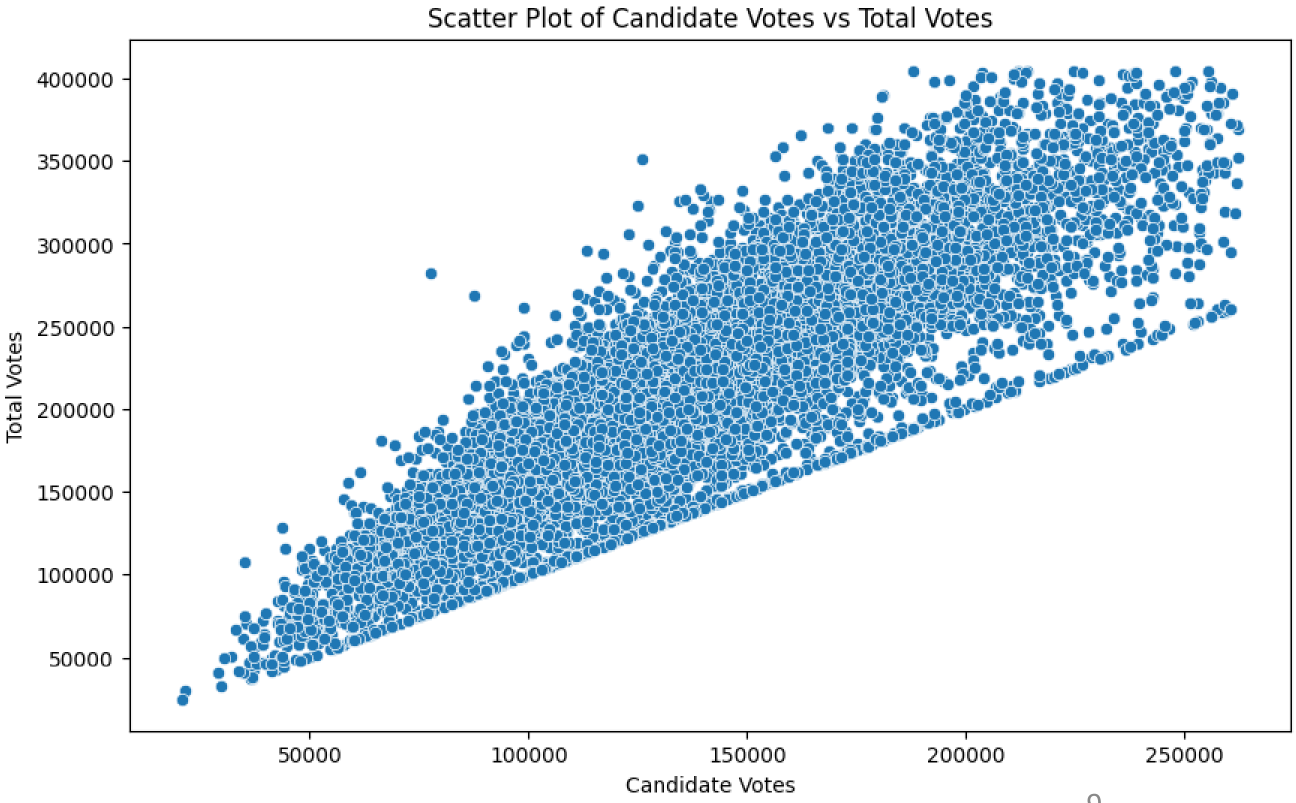
Evaluation of elections dataset



2.1.2 Distribution of variables

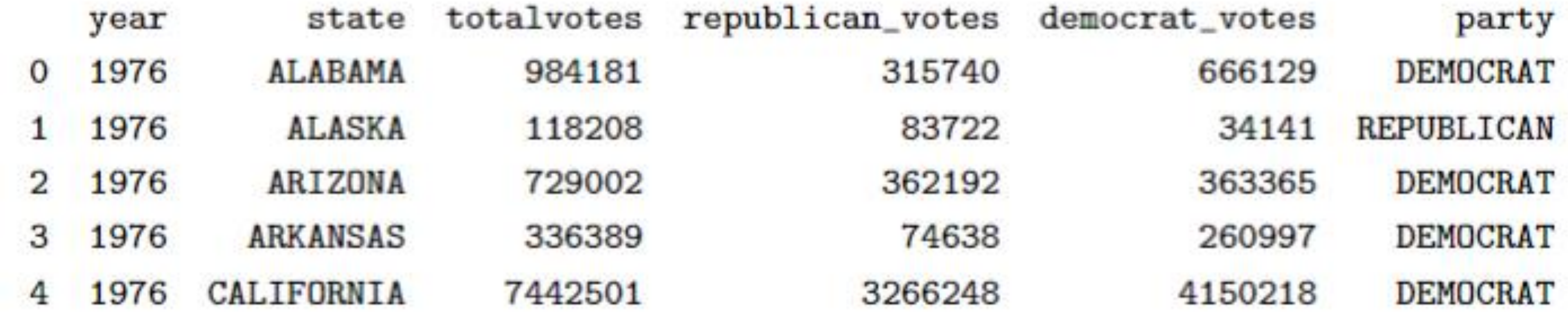
Evaluation of elections dataset





2.1.2 Distribution of variables

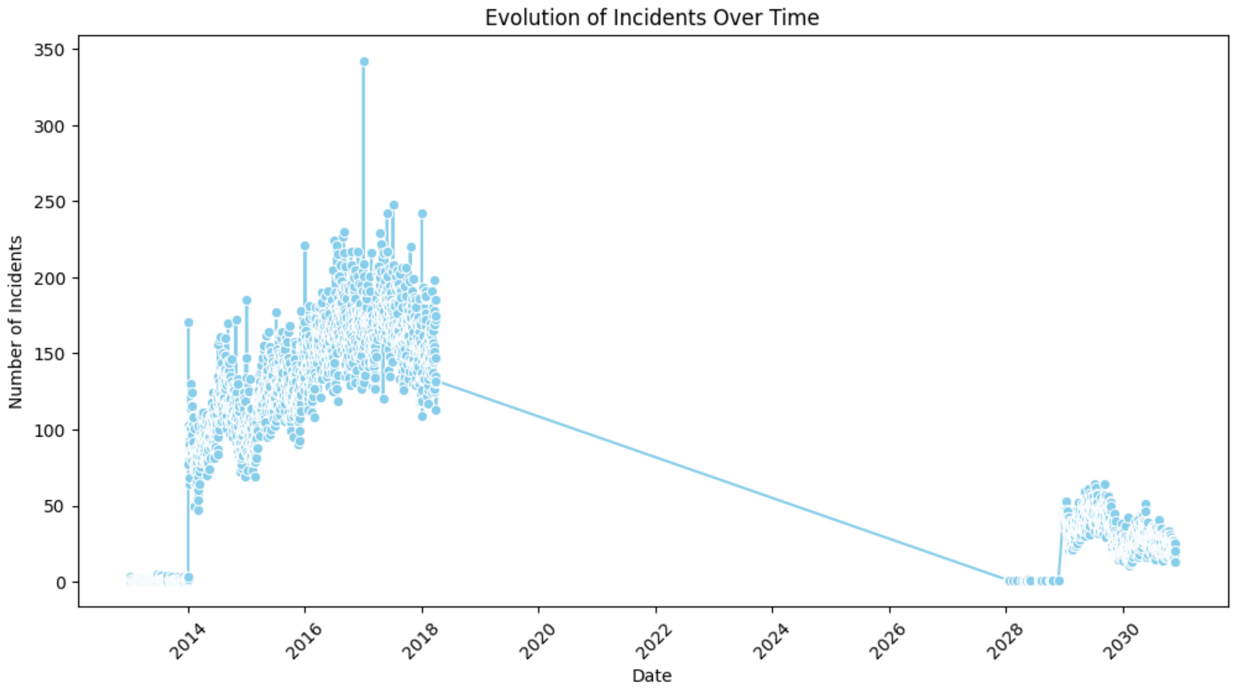
Evaluation of elections dataset

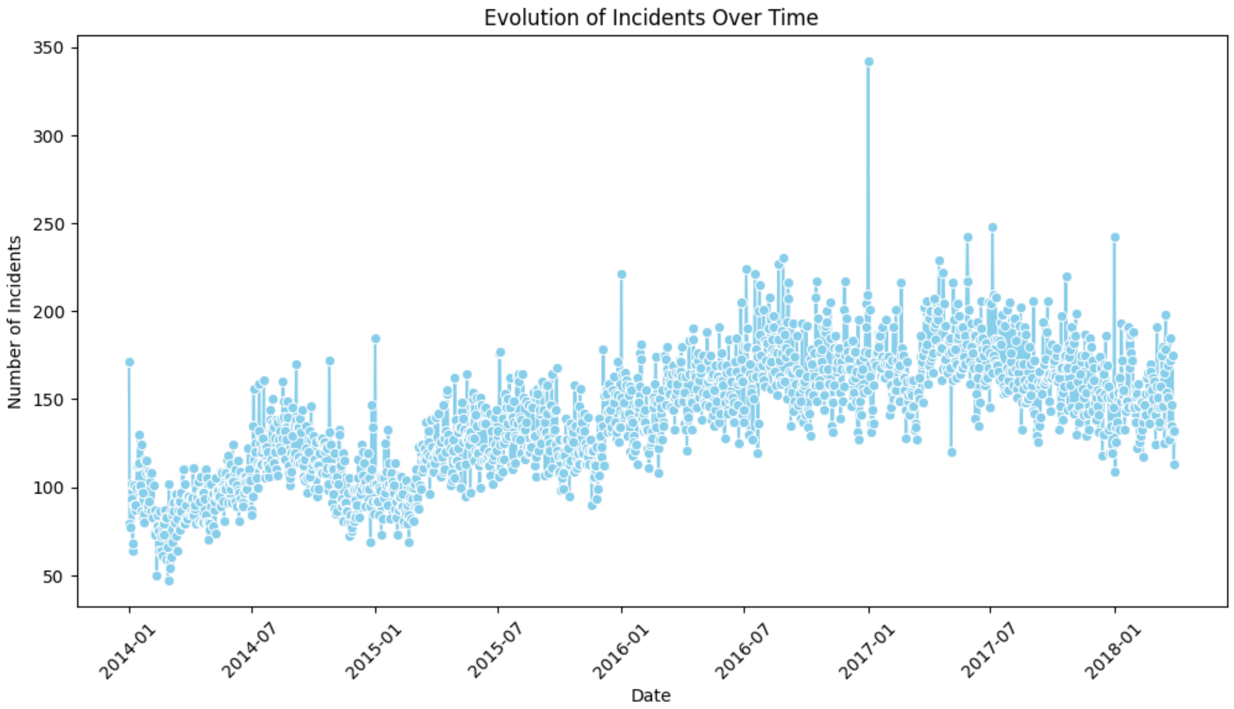


	year	state	totalvotes	republican_votes	democrat_votes	party
0	1976	ALABAMA	984181	315740	666129	DEMOCRAT
1	1976	ALASKA	118208	83722	34141	REPUBLICAN
2	1976	ARIZONA	729002	362192	363365	DEMOCRAT
3	1976	ARKANSAS	336389	74638	260997	DEMOCRAT
4	1976	CALIFORNIA	7442501	3266248	4150218	DEMOCRAT

2.1.2 Distribution of variables

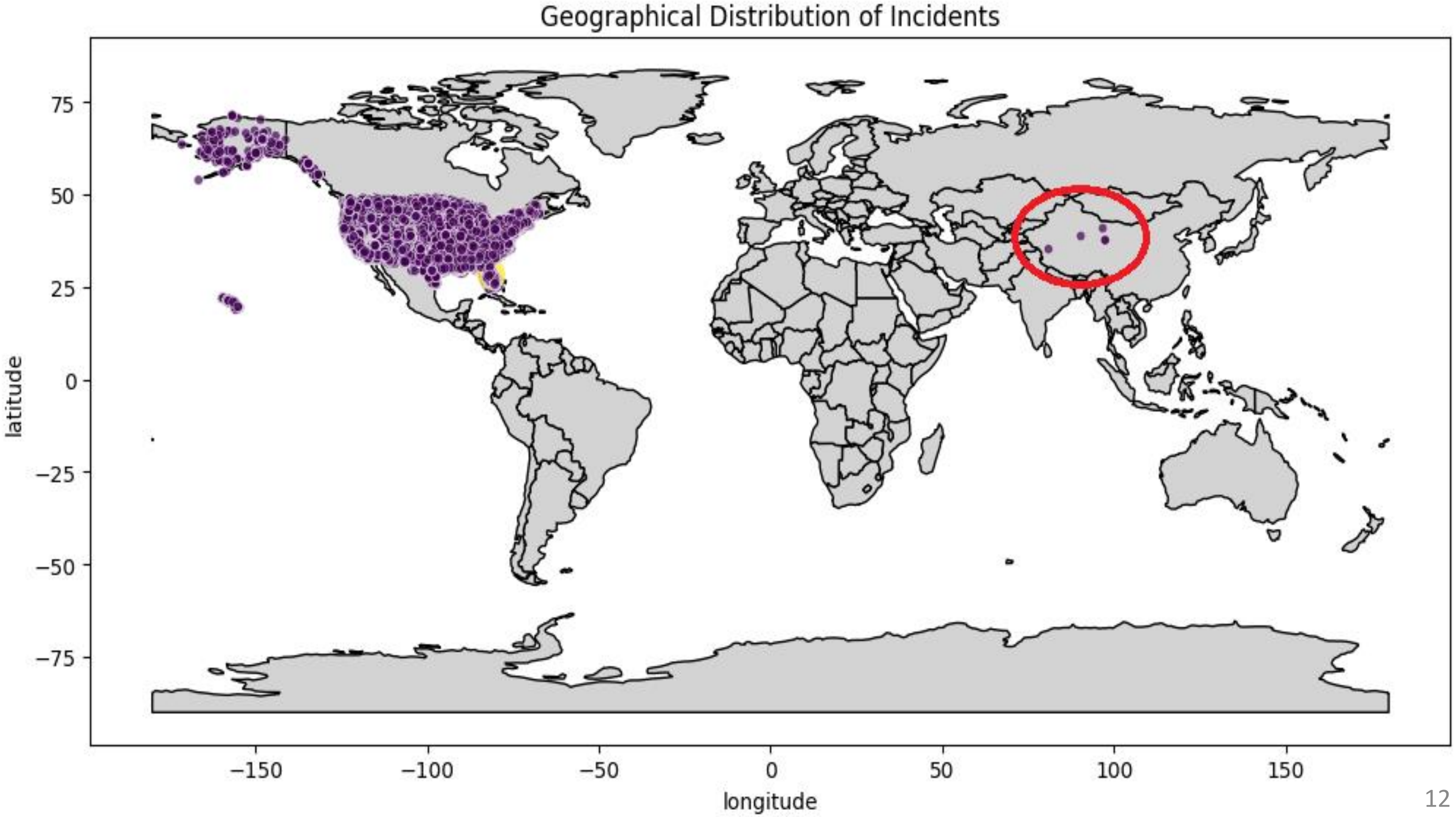
Evaluation of incidents over time





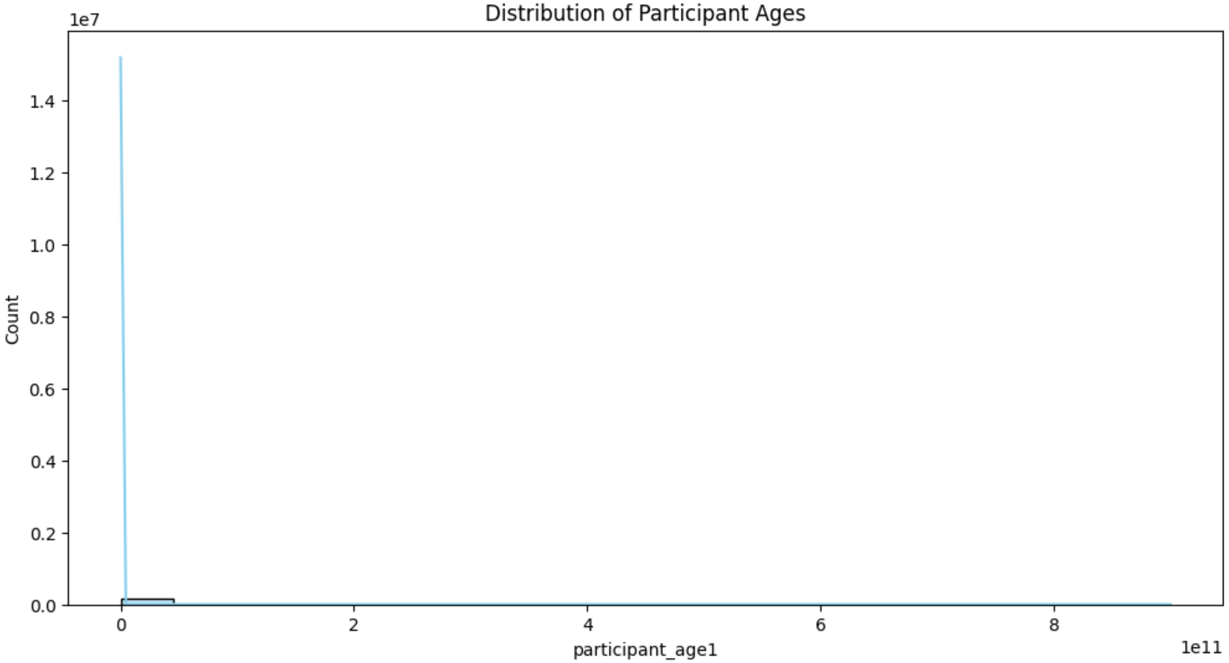
2.1.2 Distribution of variables

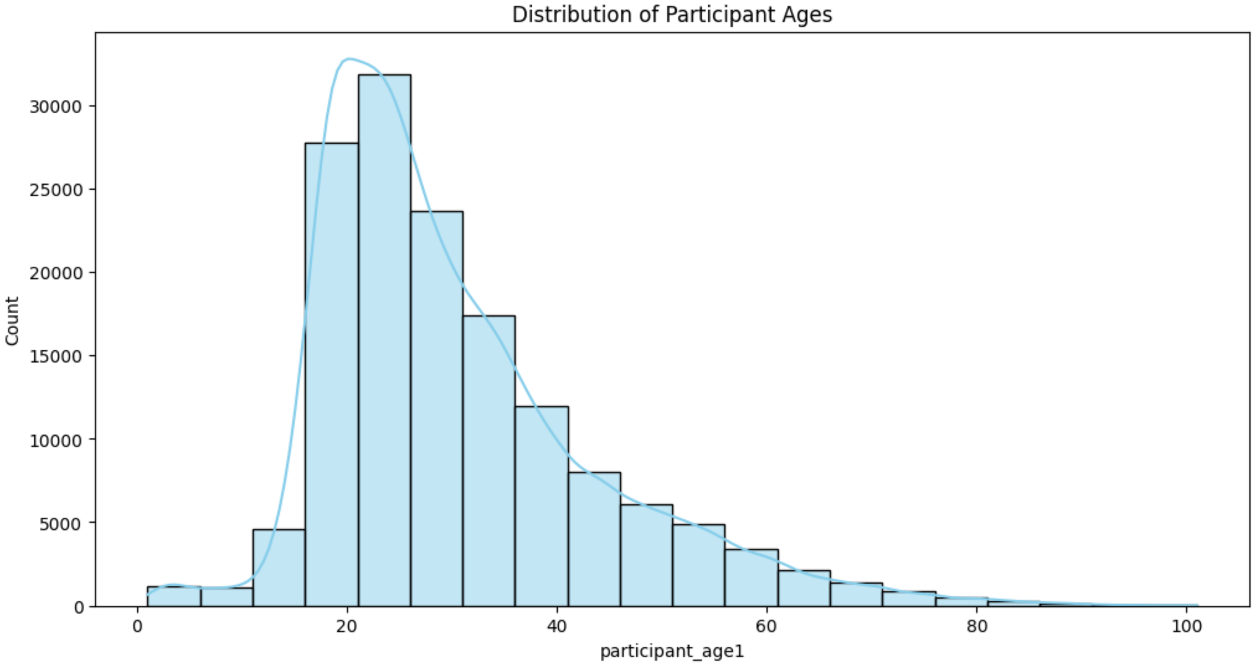
Geographical distribution of incidents



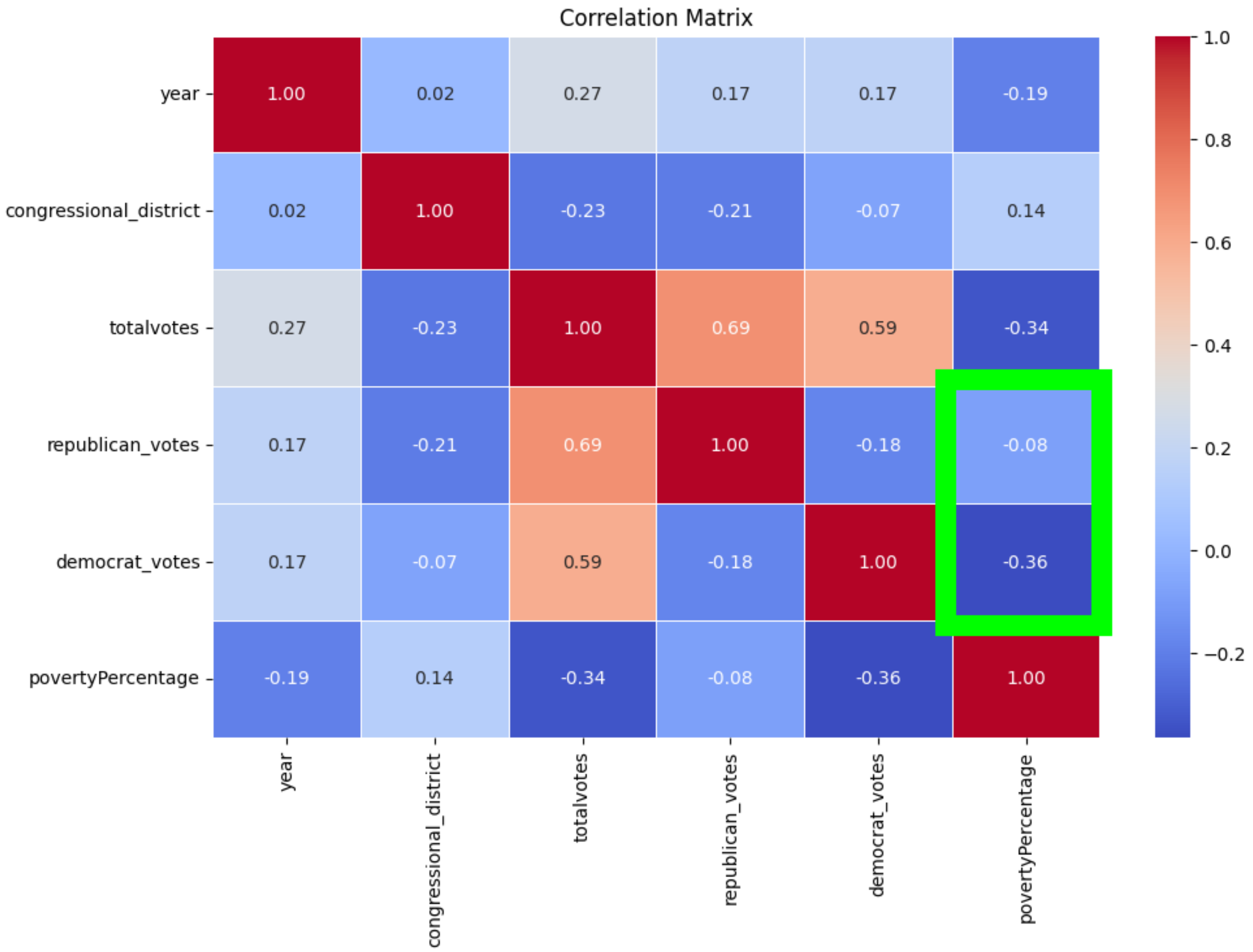
2.1.2 Distribution of variables

Distribution of participant age





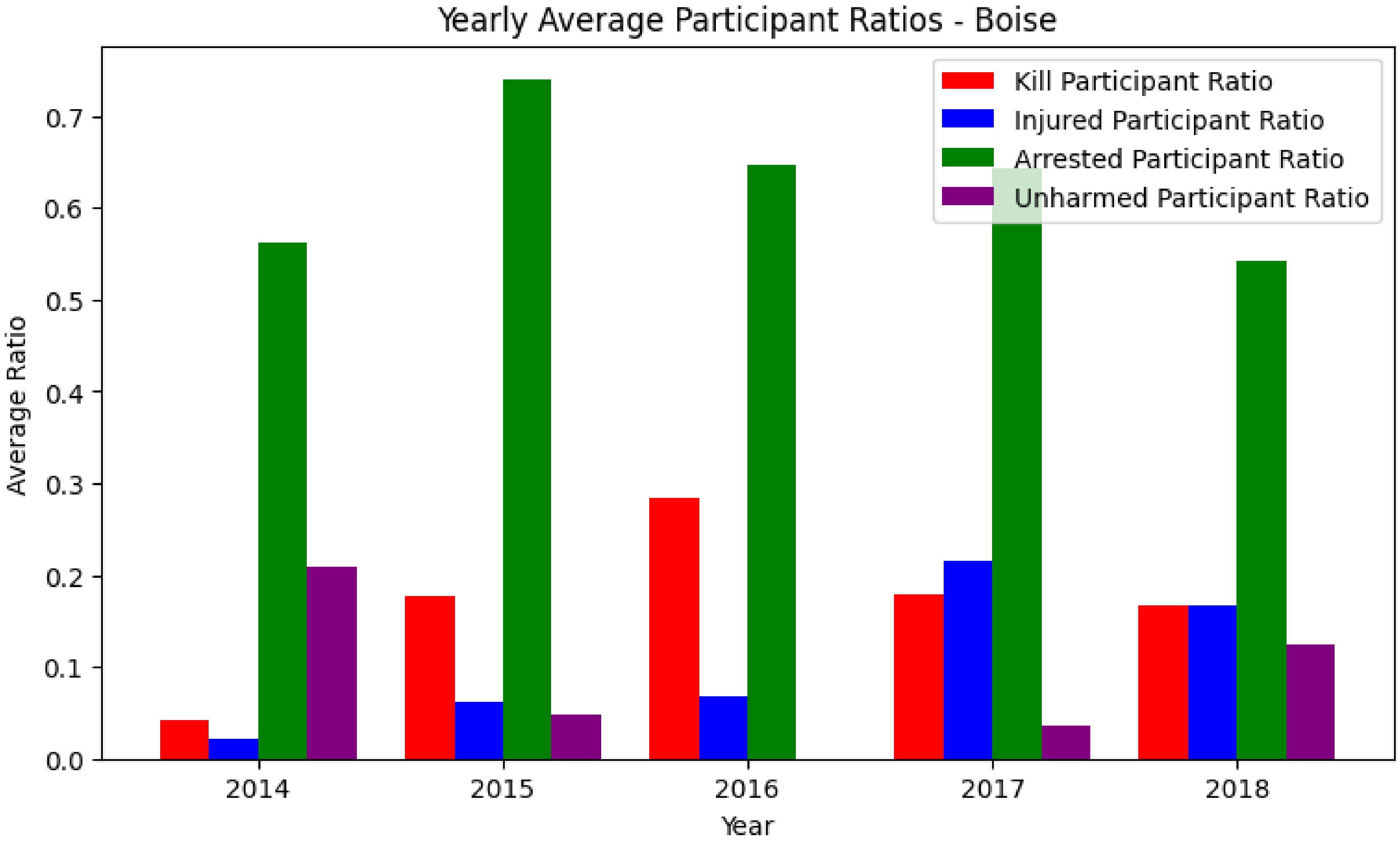
2.1.3 Pairwise correlation



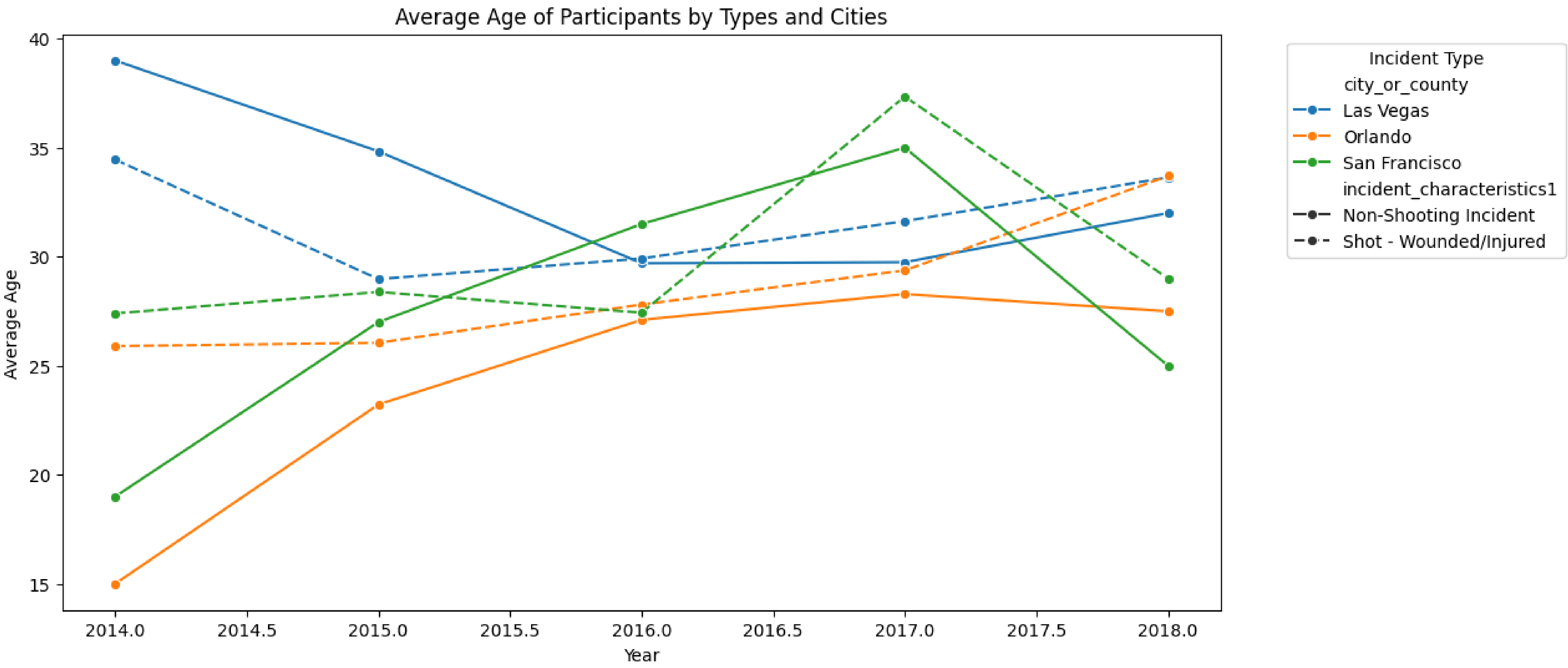
Correlation Matrix

	year	congressional_district	totalvotes	republican_votes	democrat_votes	povertyPercentage
year	1.00	0.02	0.27	0.17	0.17	-0.19
congressional_district	0.02	1.00	-0.23	-0.21	-0.07	0.14
totalvotes	0.27	-0.23	1.00	0.69	0.59	-0.34
republican_votes	0.17	-0.21	0.69	1.00	-0.18	-0.08
democrat_votes	0.17	-0.07	0.59	-0.18	1.00	-0.36
povertyPercentage	-0.19	0.14	-0.34	-0.08	-0.36	1.00

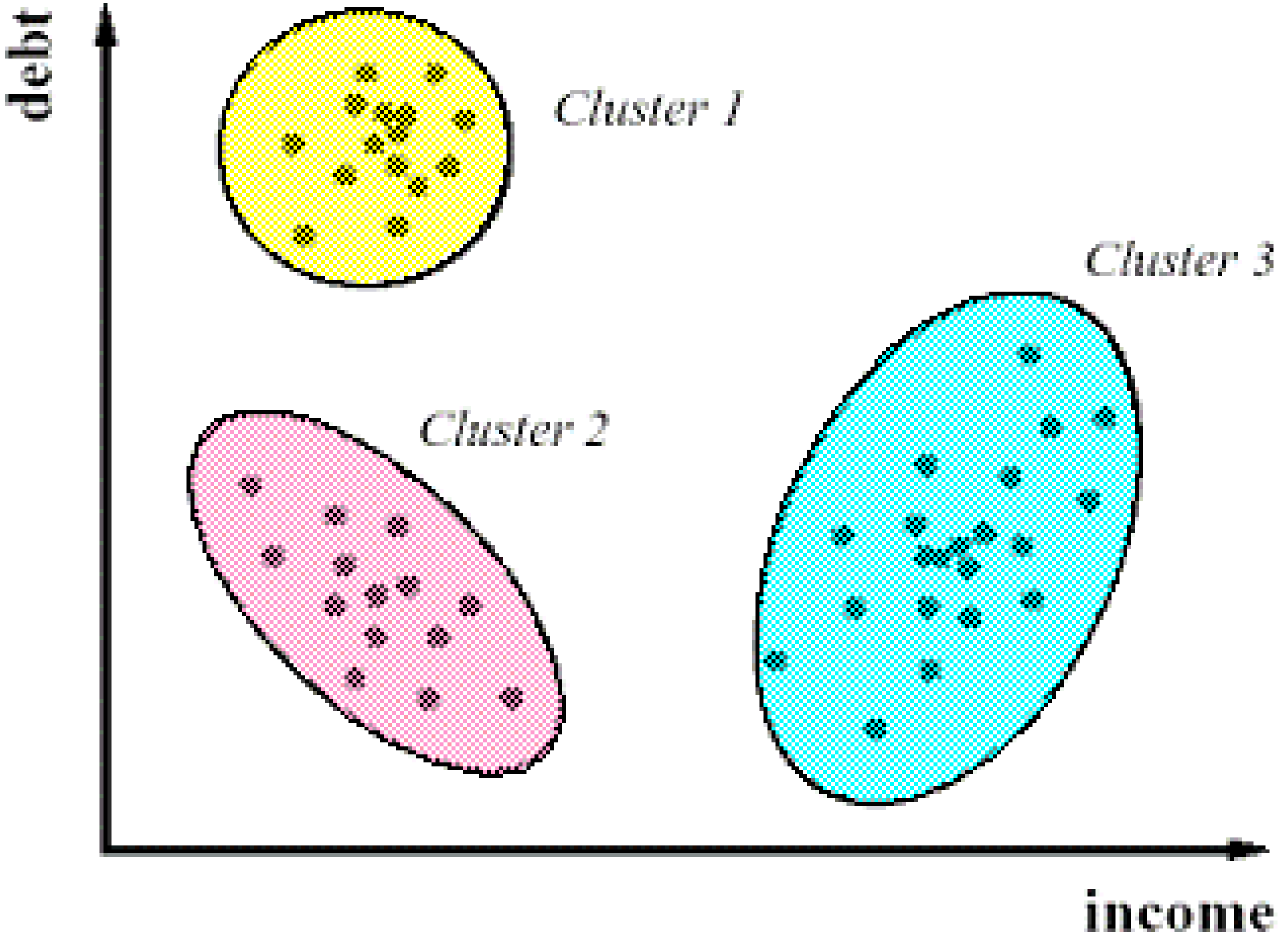
2.2 Data preparation



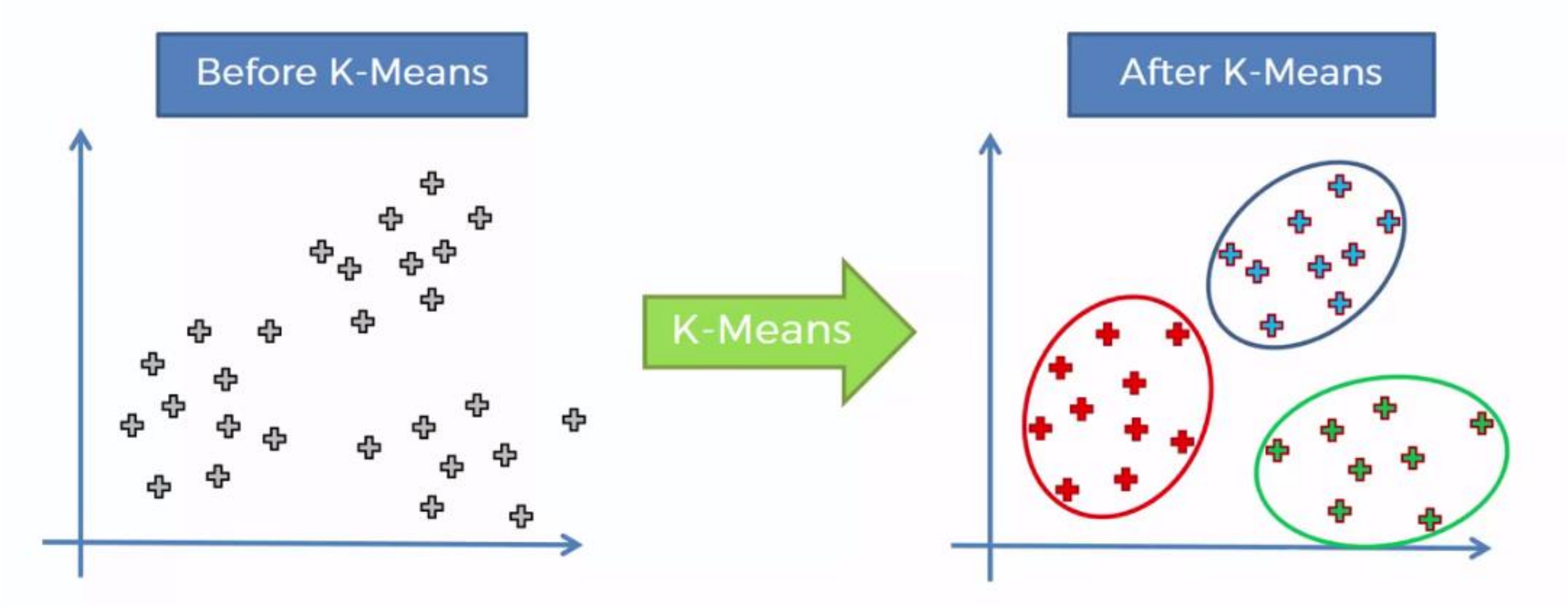
2.2 Data preparation



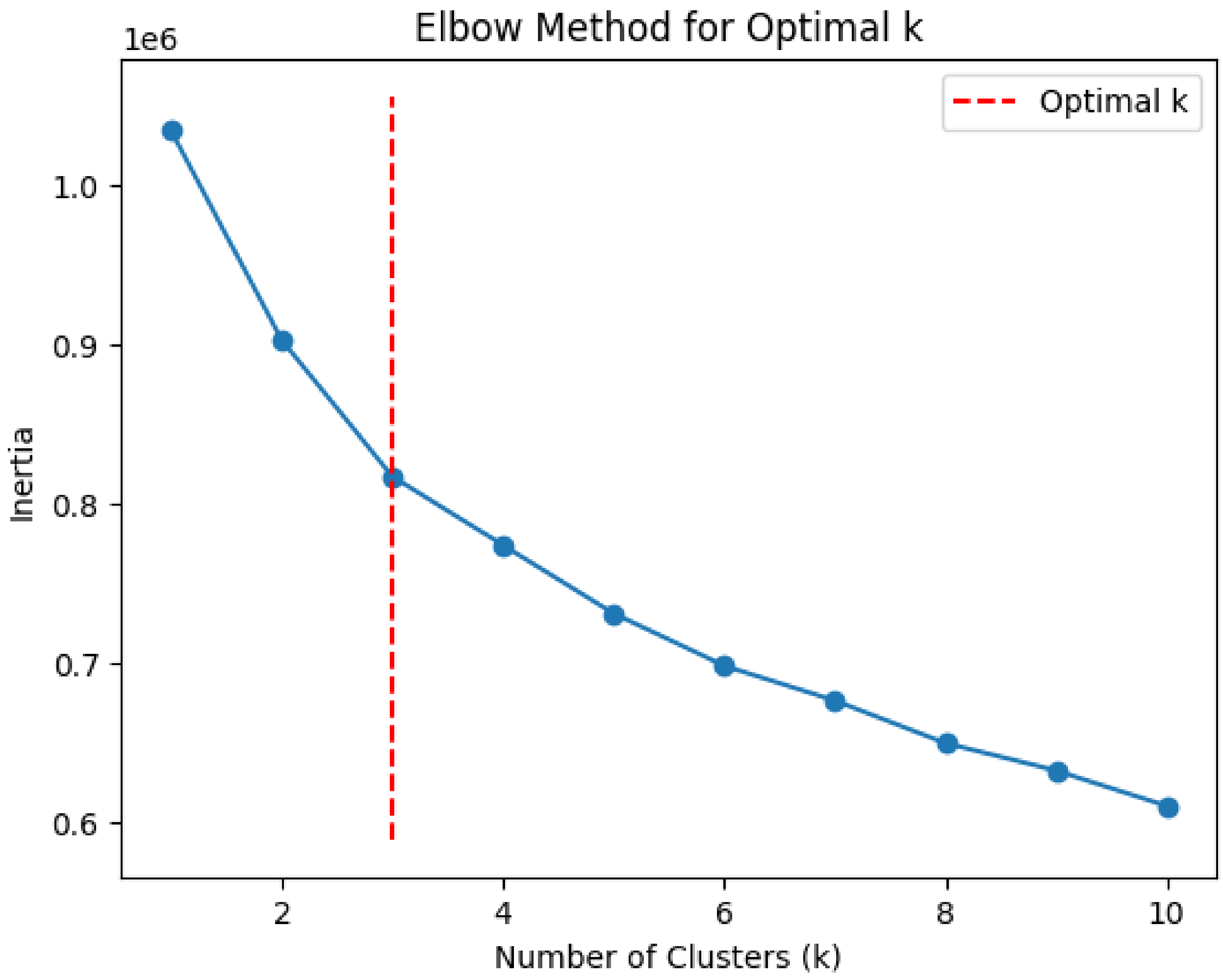
3. CLUSTERING ANALYSIS



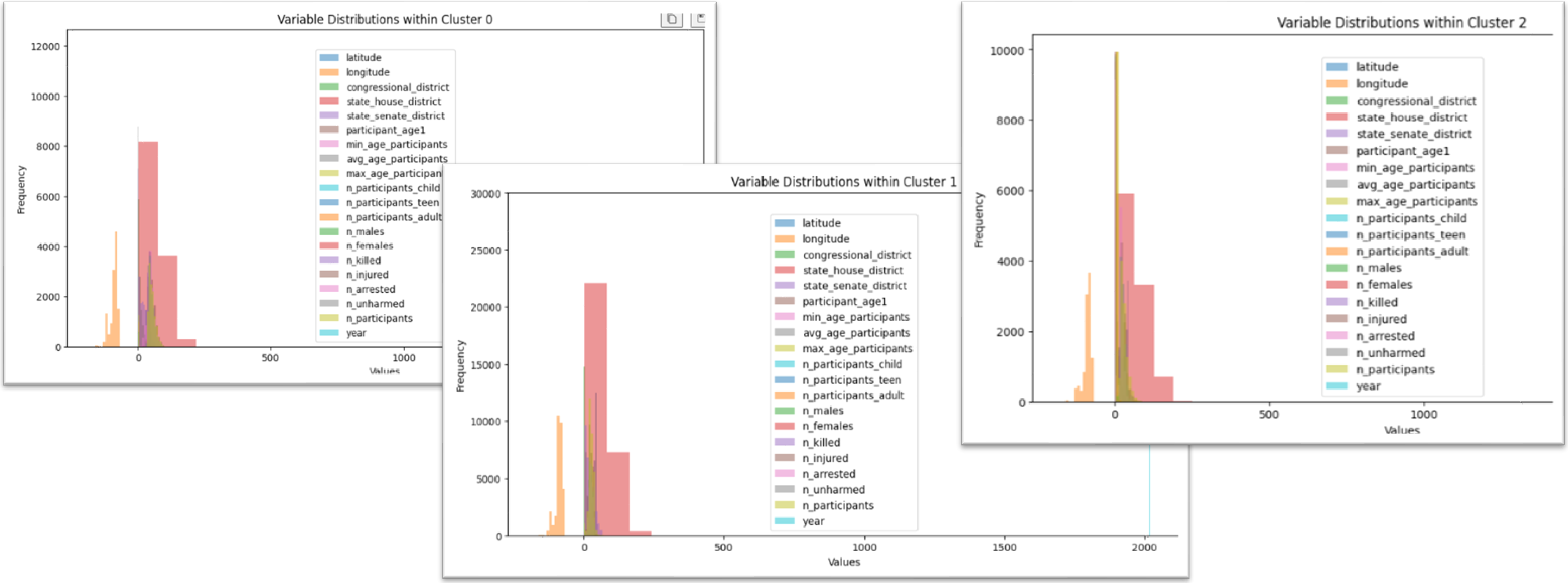
3.1 K-Means clustering



3.1.1 Identification of the best k value



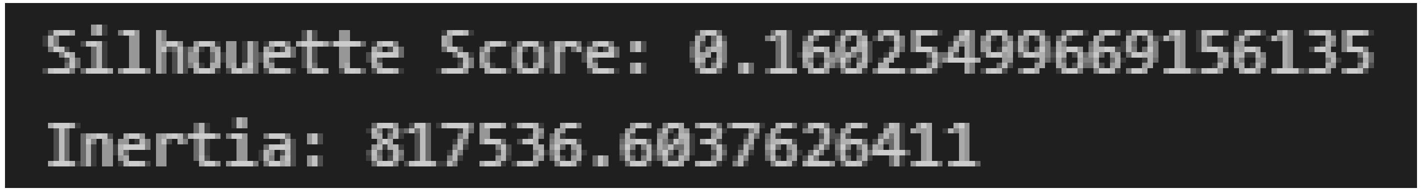
3.1.2 Cluster generation



3.1.3 Cluster evaluation

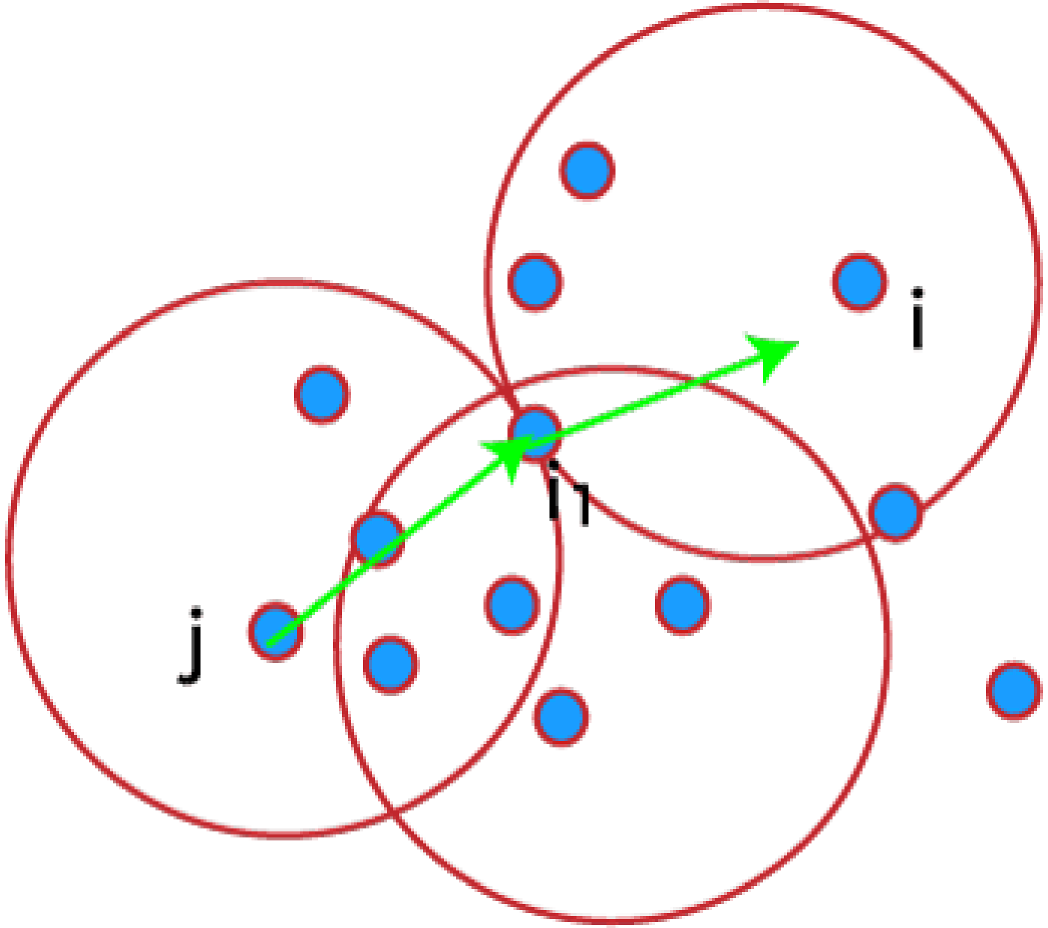
Evaluation techniques:

- **Silhouette Score:** Well separate clusters. Score = 0.160
- **Inertia:** Clusters compactness. Lower inertia values = more compact clusters. Inertia value = 817536 (relatively high)

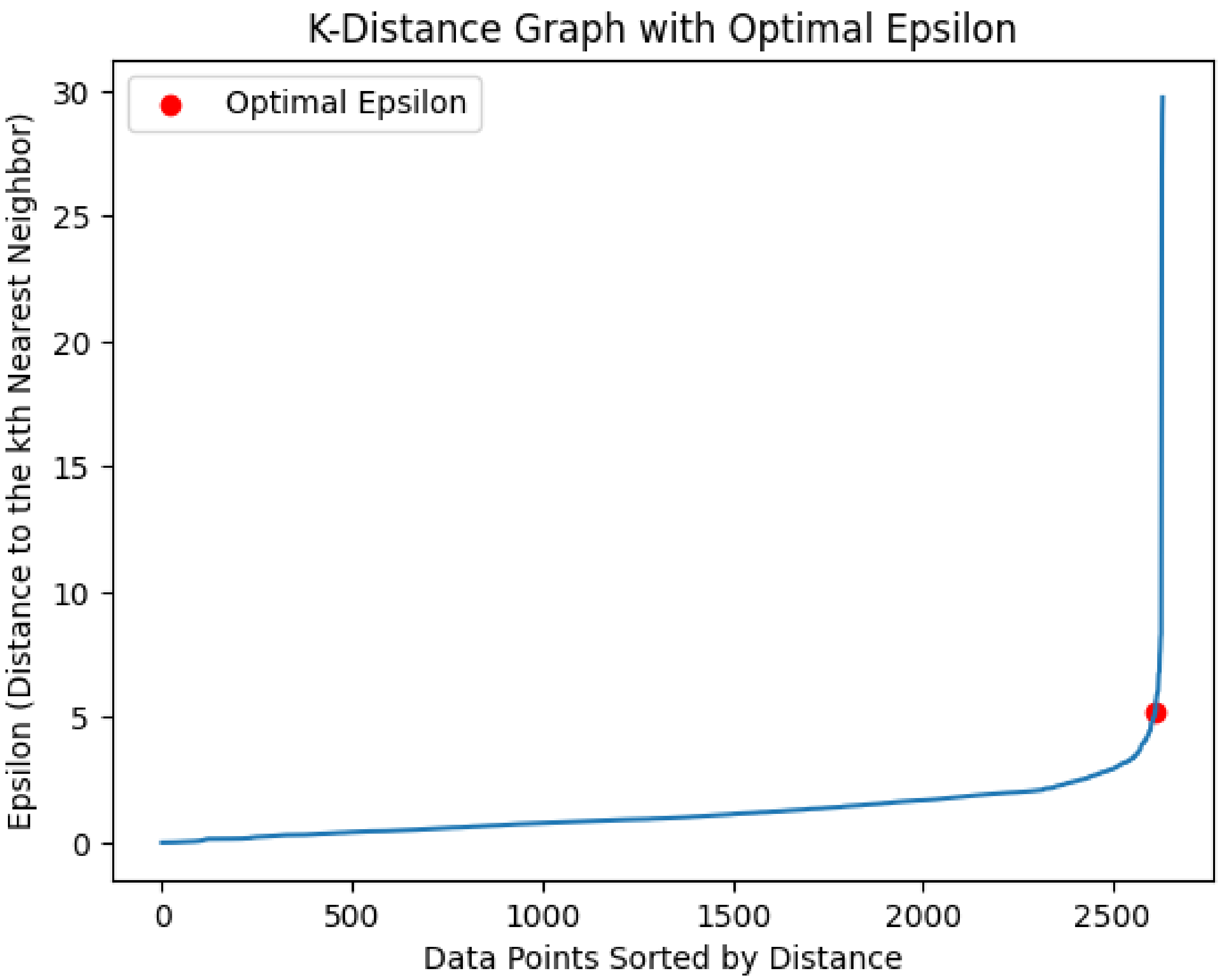


```
Silhouette Score: 0.16025499669156135
Inertia: 817536.6037626411
```

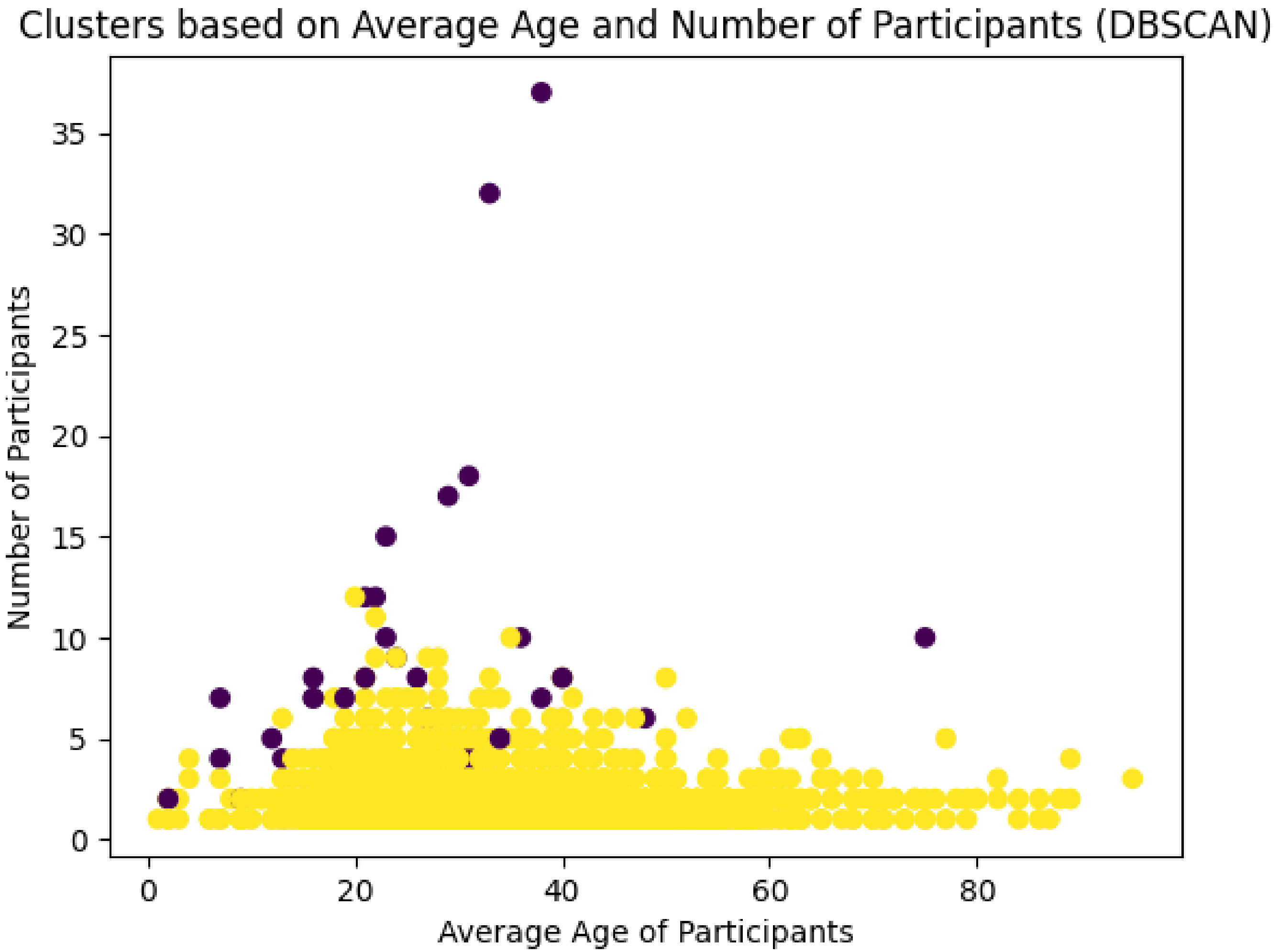
3.2 Density-based clustering



3.2.1 Study of clustering parameters



3.2.2 Characterization and interpretation of obtained clusters



3.3 Hierarchical clustering

Single Linkage: Shortest distance between points in two clusters.

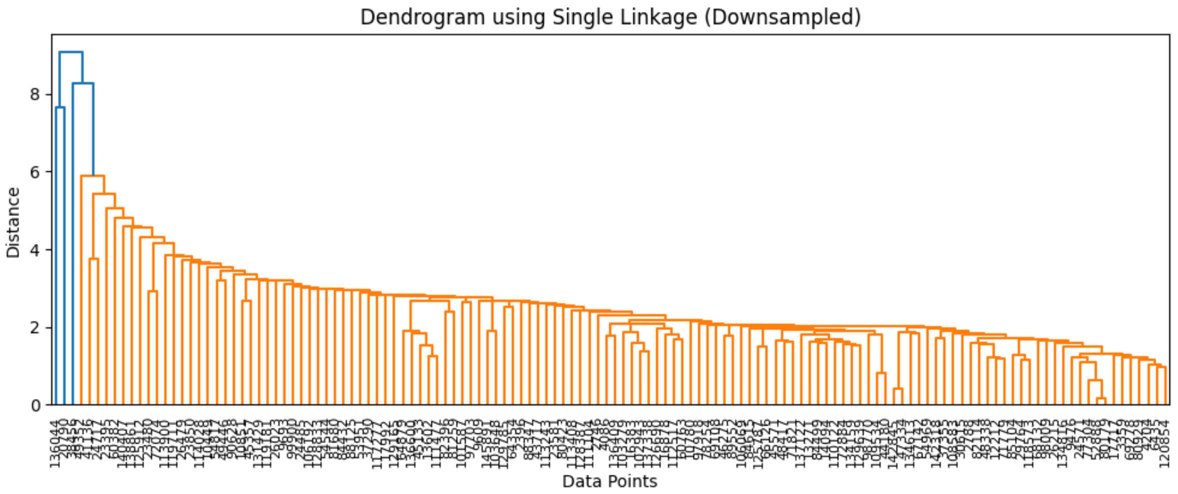
Complete Linkage: Longest distance between points in two clusters.

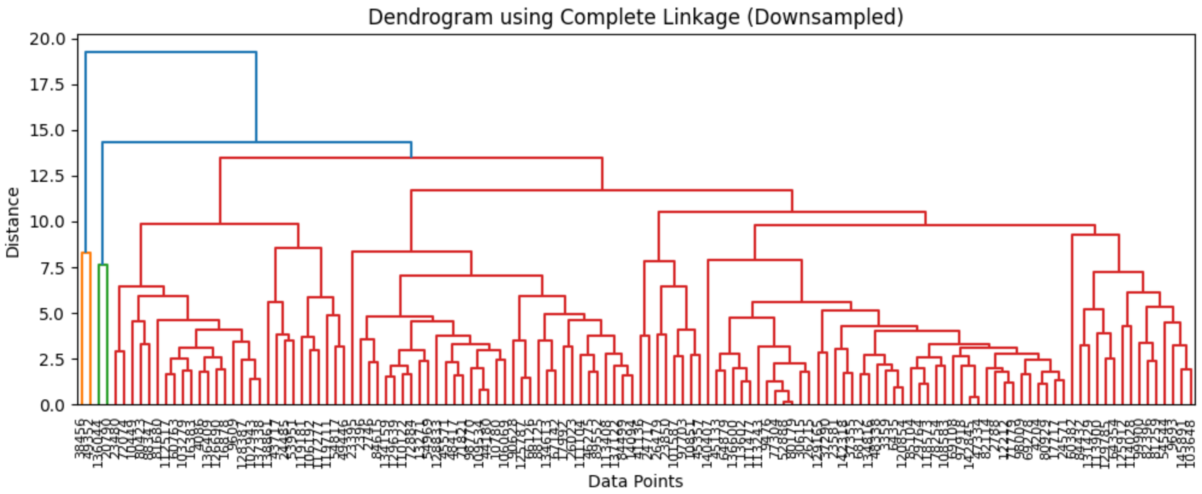
Average Linkage: Average distance between points in two clusters.

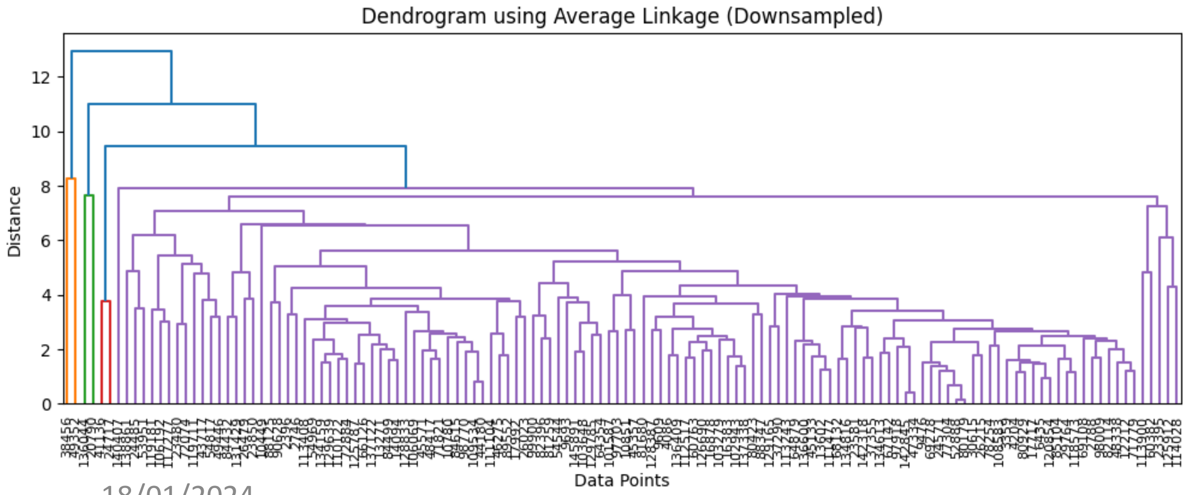
Ward Linkage: Minimizes the variance within clusters.

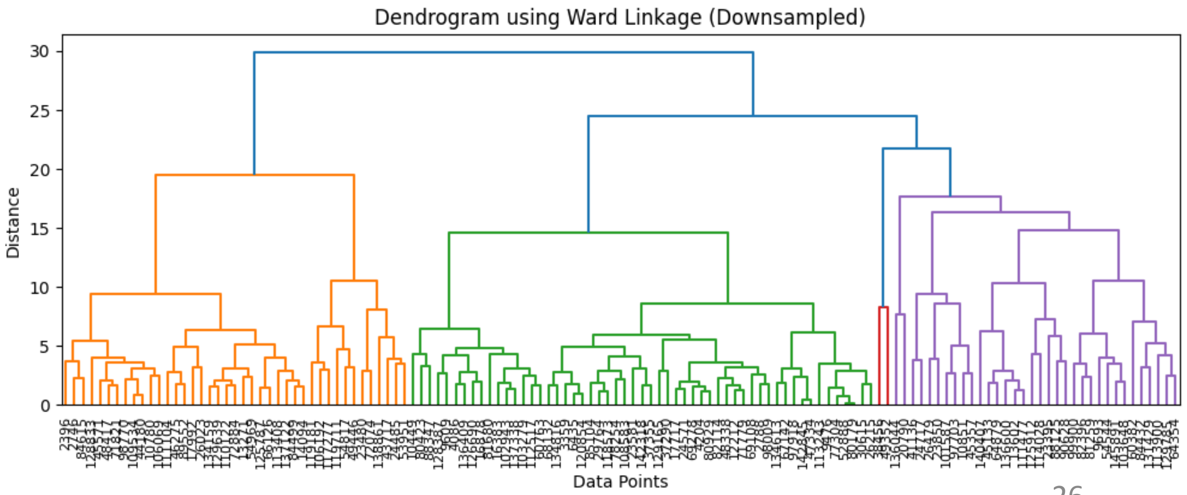
Single Linkage has the highest **Silhouette Score** with 0.314

3.3 Hierarchical clustering

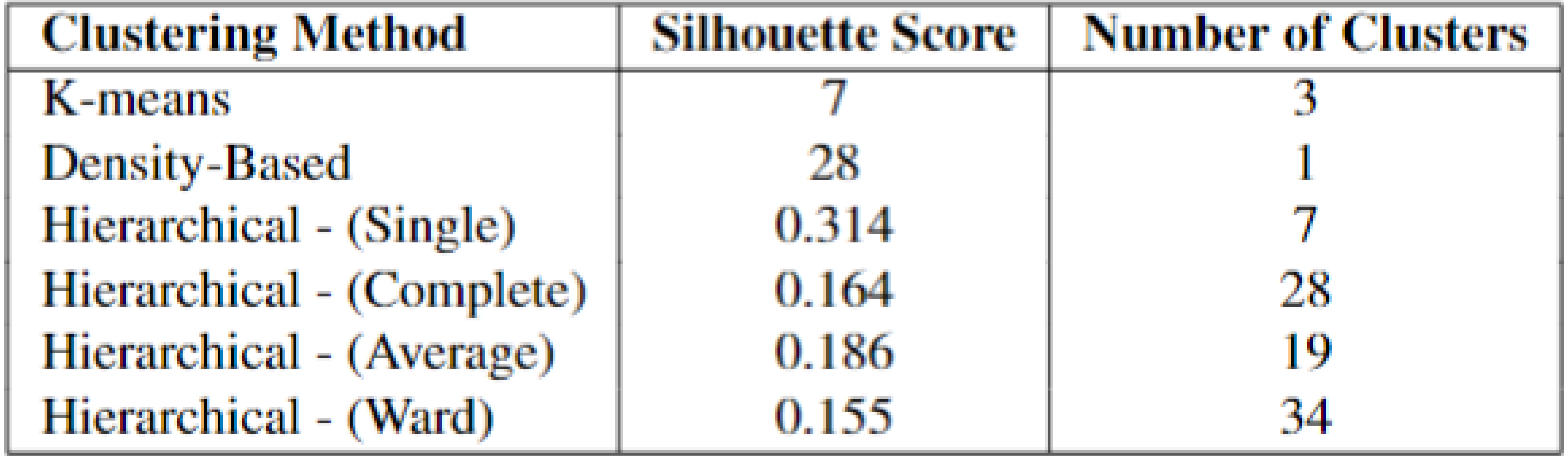








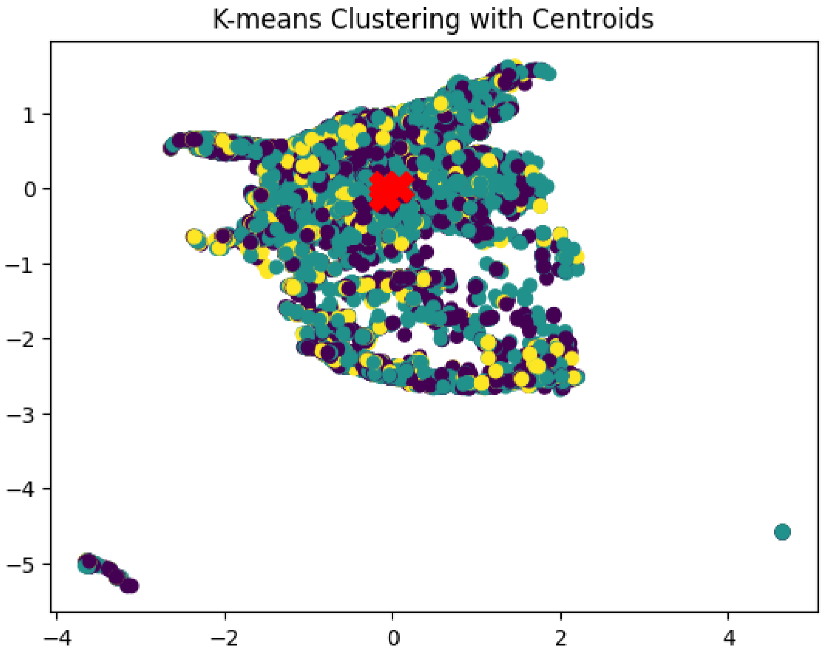
3.4 Evaluation of clustering approaches

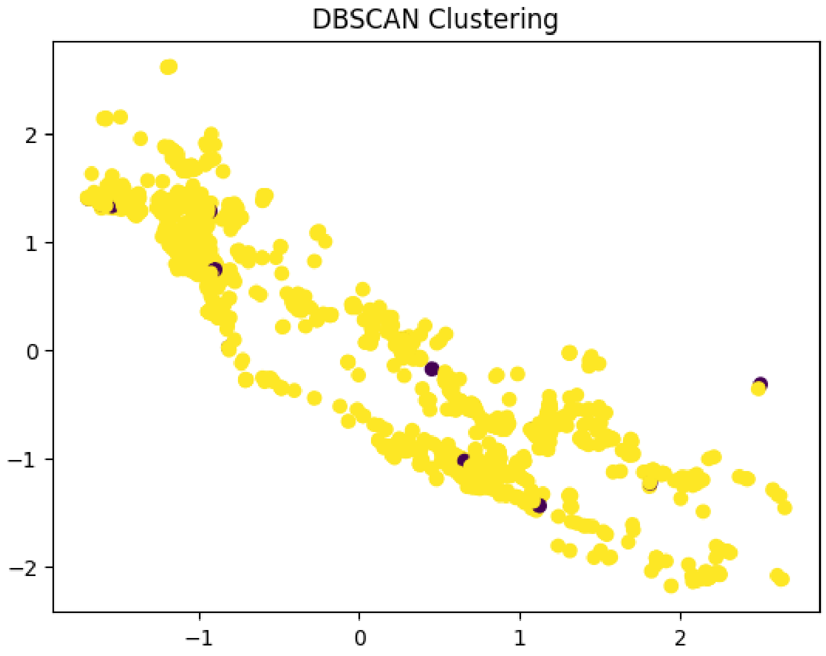


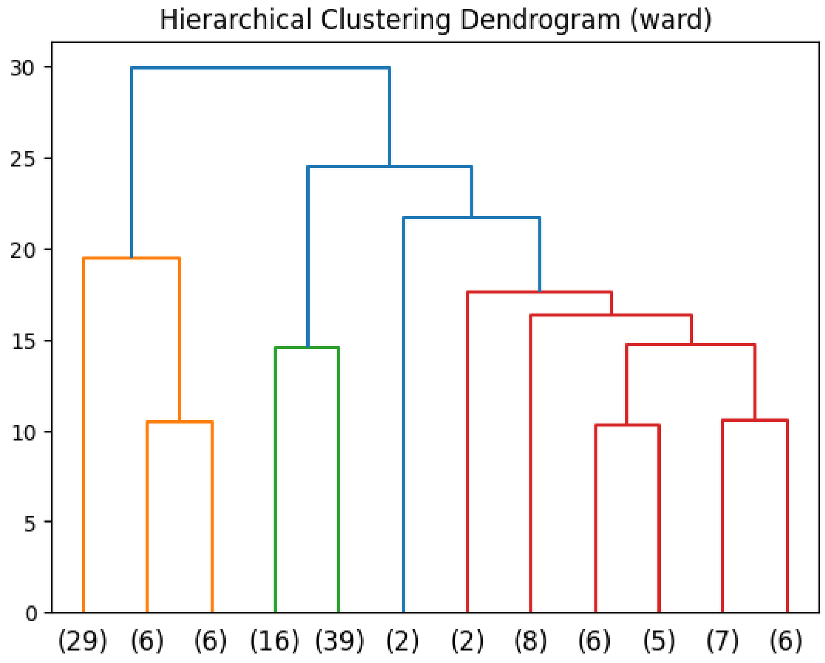
Clustering Method	Silhouette Score	Number of Clusters
K-means	7	3
Density-Based	28	1
Hierarchical - (Single)	0.314	7
Hierarchical - (Complete)	0.164	28
Hierarchical - (Average)	0.186	19
Hierarchical - (Ward)	0.155	34

Table 3.1: Silhouette scores of clustering methods

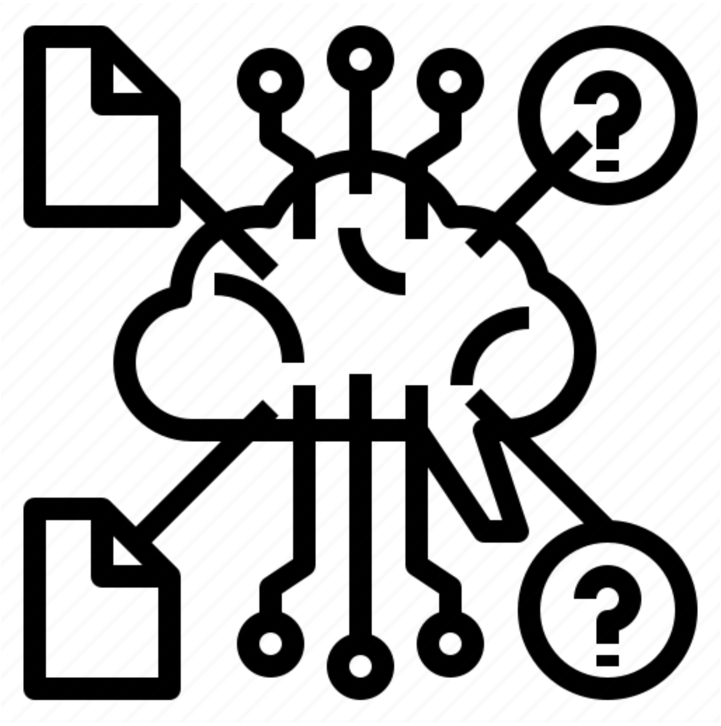
3.4 Evaluation of clustering approaches



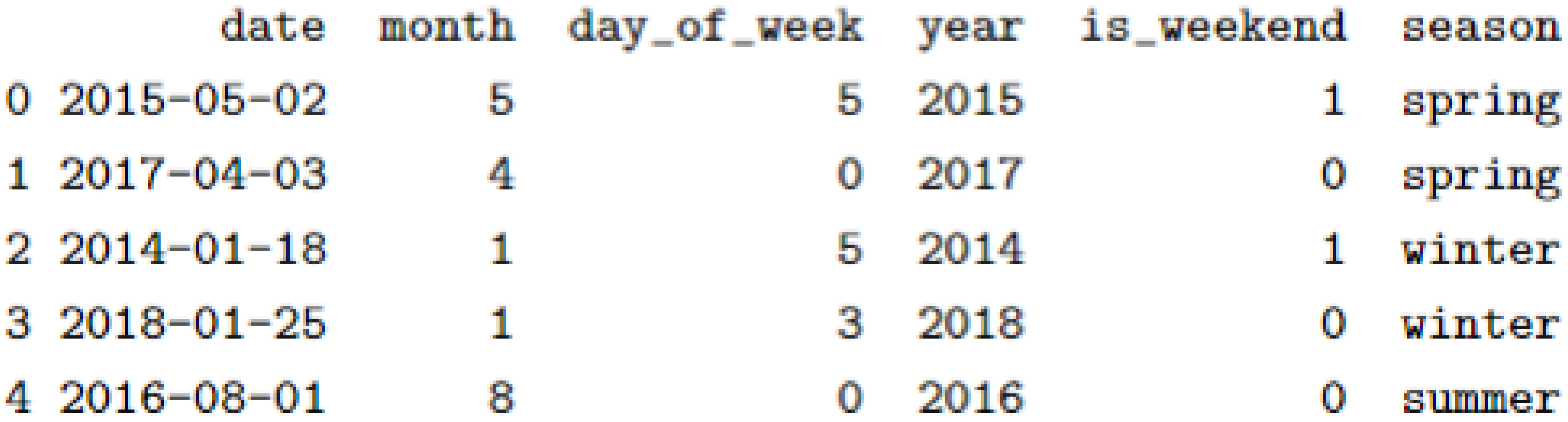




4. PREDICTIVE ANALYSIS

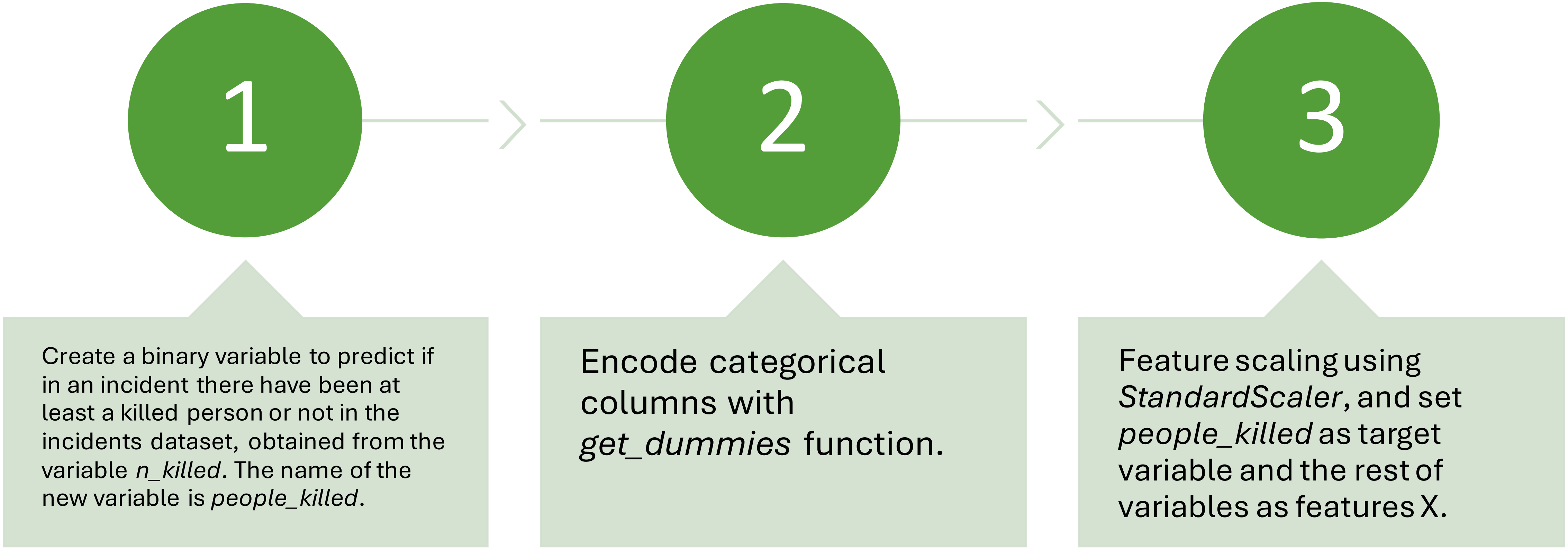


4.1 New feature definition



```
         date  month  day_of_week  year  is_weekend  season
0 2015-05-02      5            5  2015           1  spring
1 2017-04-03      4            0  2017           0  spring
2 2014-01-18      1            5  2014           1  winter
3 2018-01-25      1            3  2018           0  winter
4 2016-08-01      8            0  2016           0  summer
```

4.2 Preprocessing



1

Create a binary variable to predict if in an incident there have been at least a killed person or not in the incidents dataset, obtained from the variable *n_killed*. The name of the new variable is *people_killed*.

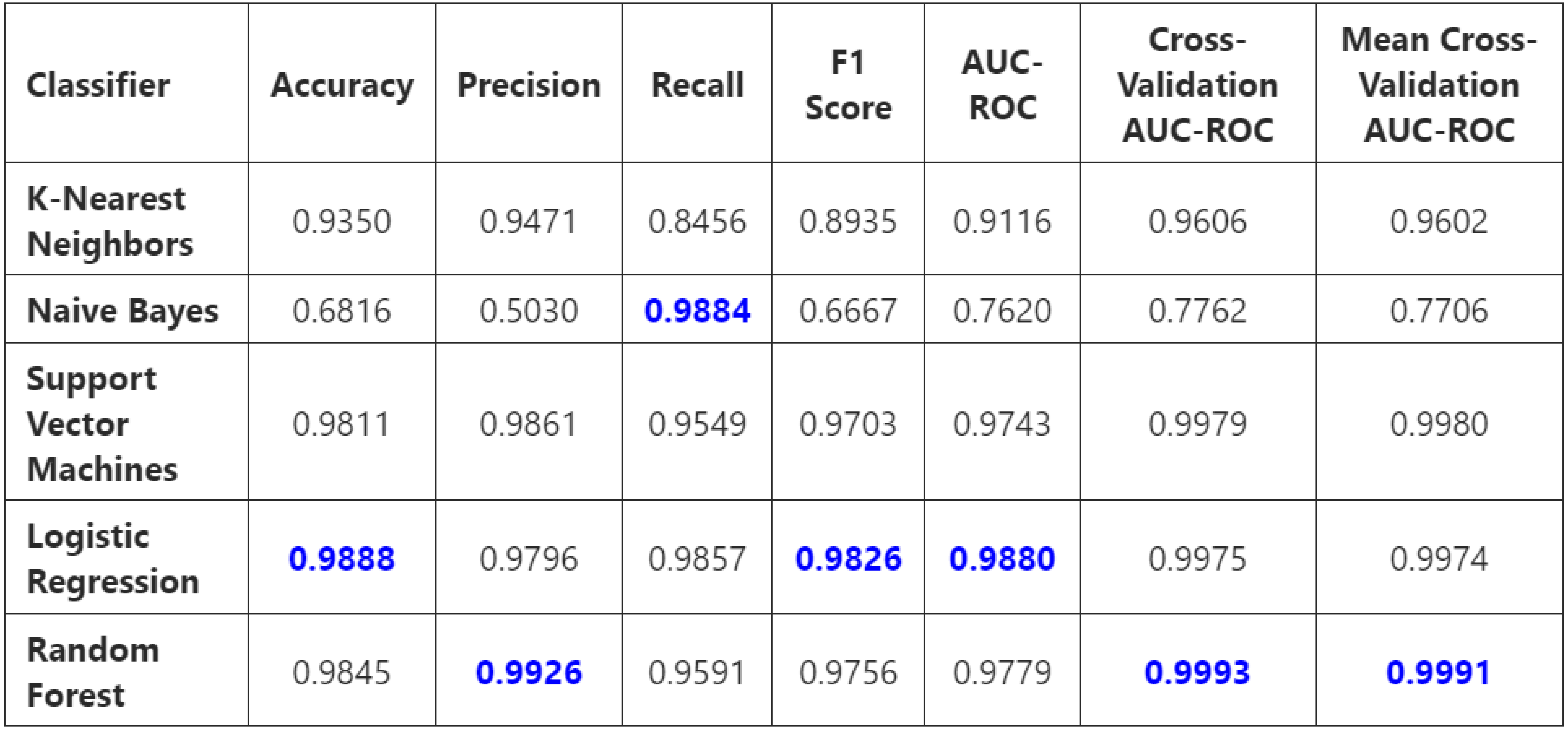
2

Encode categorical columns with *get_dummies* function.

3

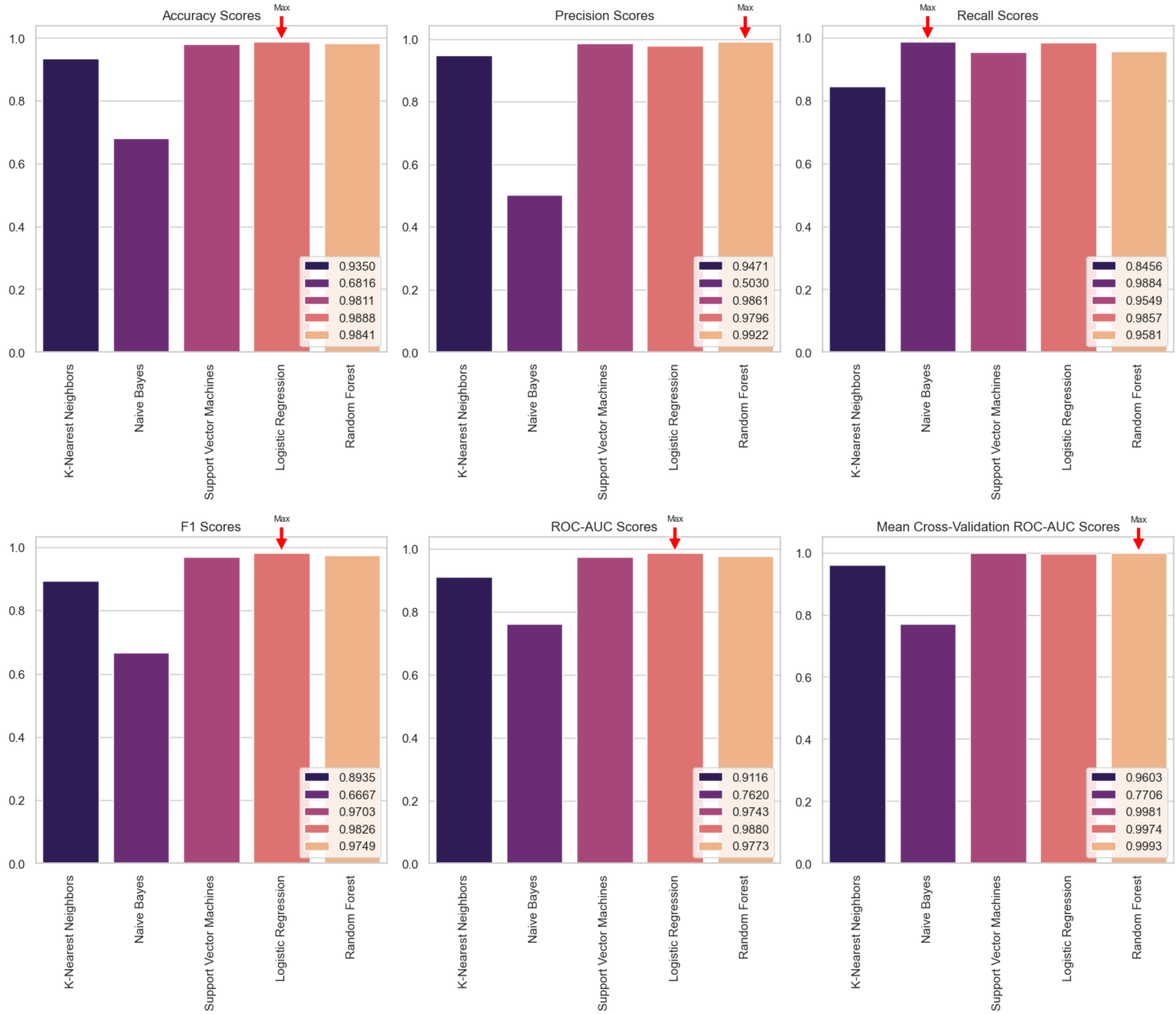
Feature scaling using *StandardScaler*, and set *people_killed* as target variable and the rest of variables as features X.

4.3 Model selection and evaluation

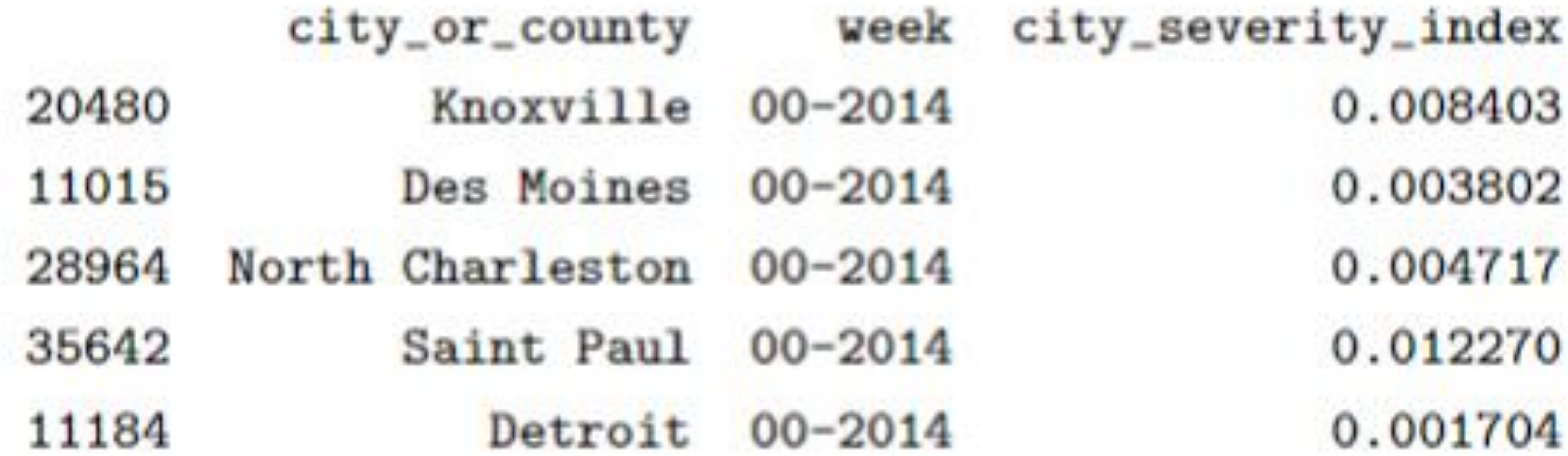


Classifier	Accuracy	Precision	Recall	F1 Score	AUC-ROC	Cross-Validation AUC-ROC	Mean Cross-Validation AUC-ROC
K-Nearest Neighbors	0.9350	0.9471	0.8456	0.8935	0.9116	0.9606	0.9602
Naive Bayes	0.6816	0.5030	**0.9884**	0.6667	0.7620	0.7762	0.7706
Support Vector Machines	0.9811	0.9861	0.9549	0.9703	0.9743	0.9979	0.9980
Logistic Regression	**0.9888**	0.9796	0.9857	**0.9826**	**0.9880**	0.9975	0.9974
Random Forest	0.9845	**0.9926**	0.9591	0.9756	0.9779	**0.9993**	**0.9991**

4.3 Model selection and evaluation

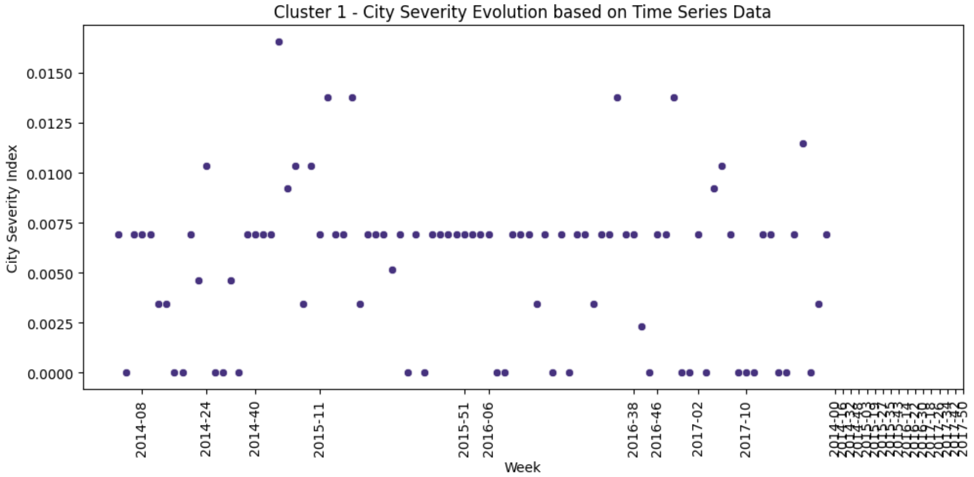


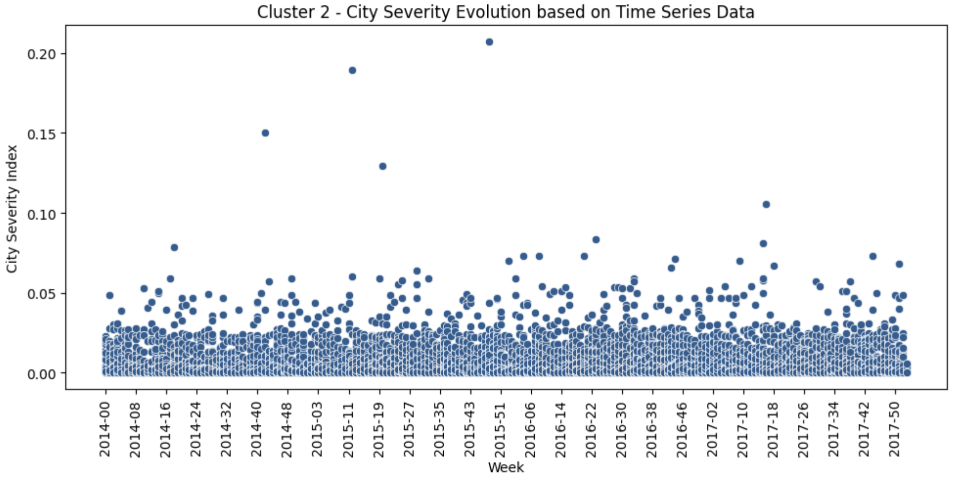
5. TIME SERIES ANALYSIS

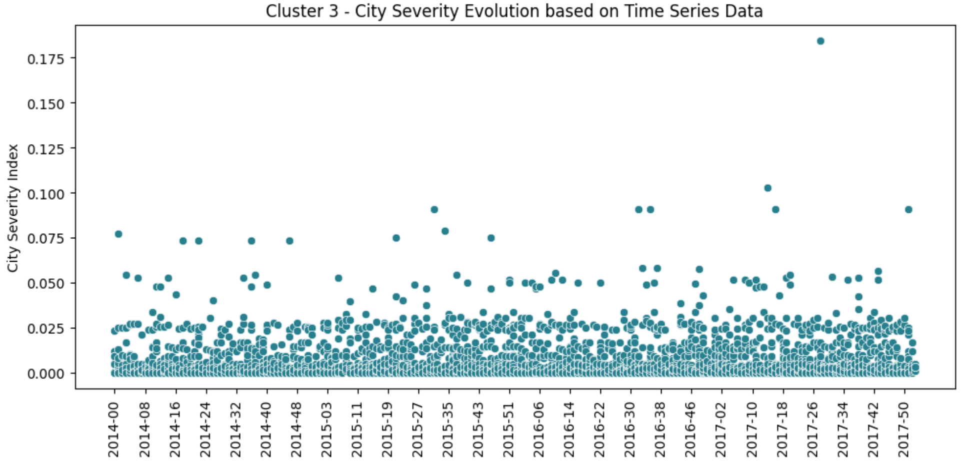


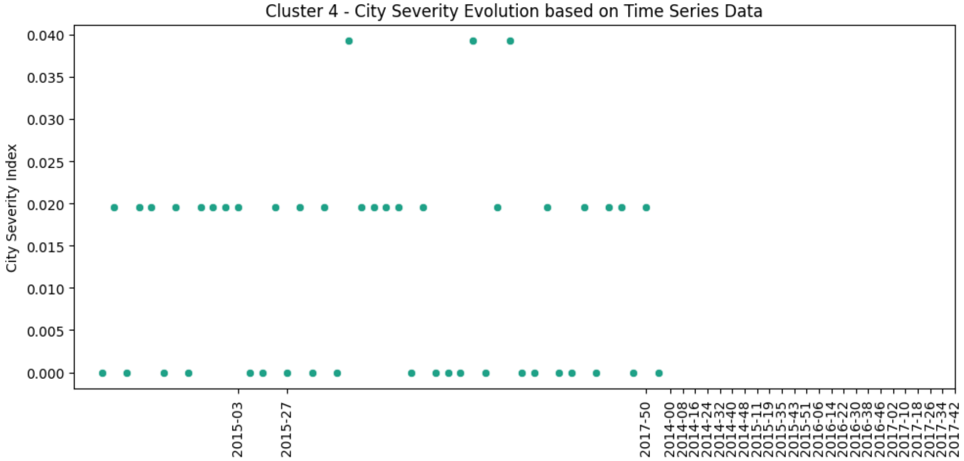
```
        city_or_county     week  city_severity_index
20480        Knoxville  00-2014             0.008403
11015       Des Moines  00-2014             0.003802
28964 North Charleston  00-2014             0.004717
35642       Saint Paul  00-2014             0.012270
11184          Detroit  00-2014             0.001704
```

5.1 Clustering

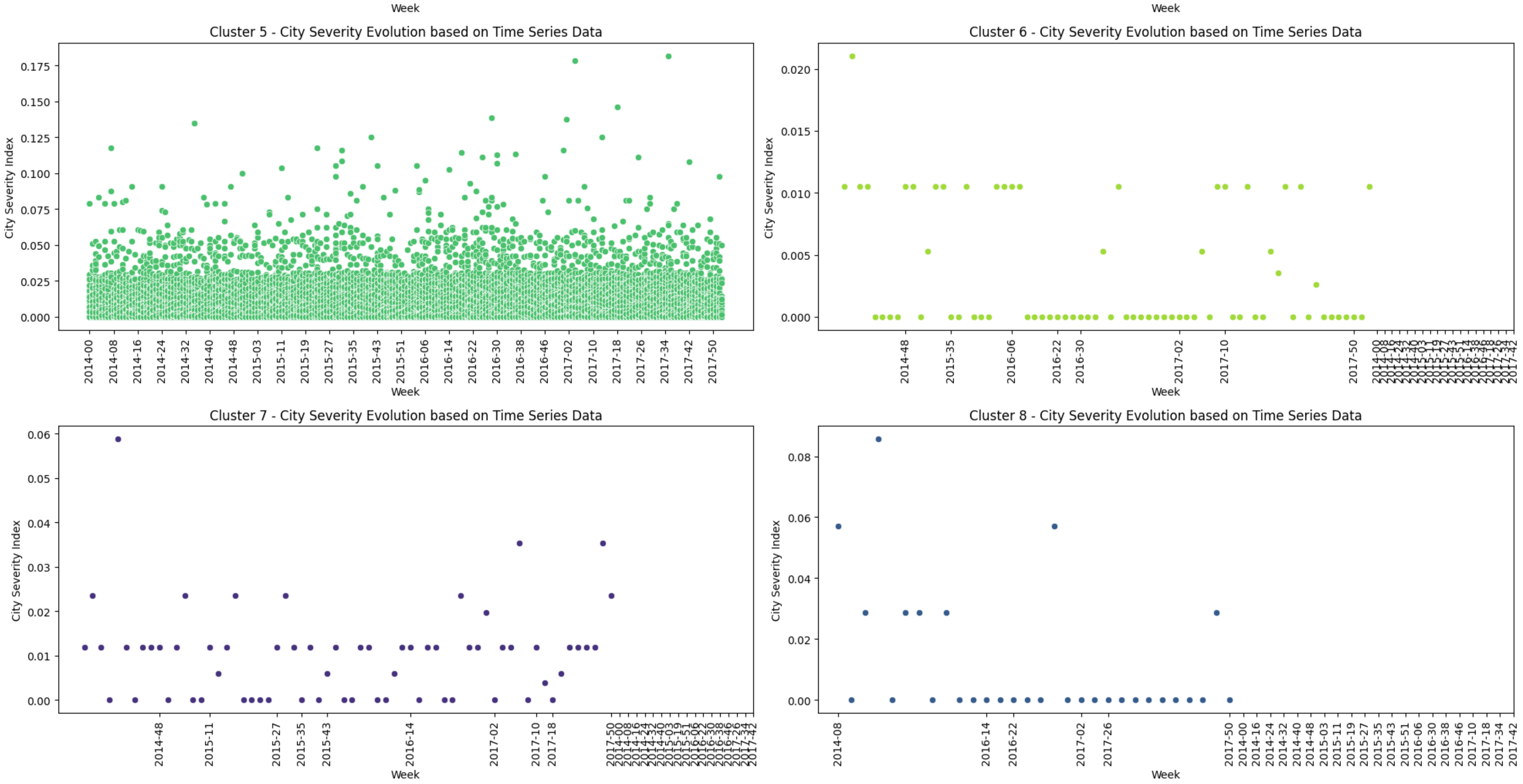

Cluster 1 - City Severity Evolution based on Time Series Data


Cluster 2 - City Severity Evolution based on Time Series Data


Cluster 3 - City Severity Evolution based on Time Series Data


Cluster 4 - City Severity Evolution based on Time Series Data

5.1 Clustering

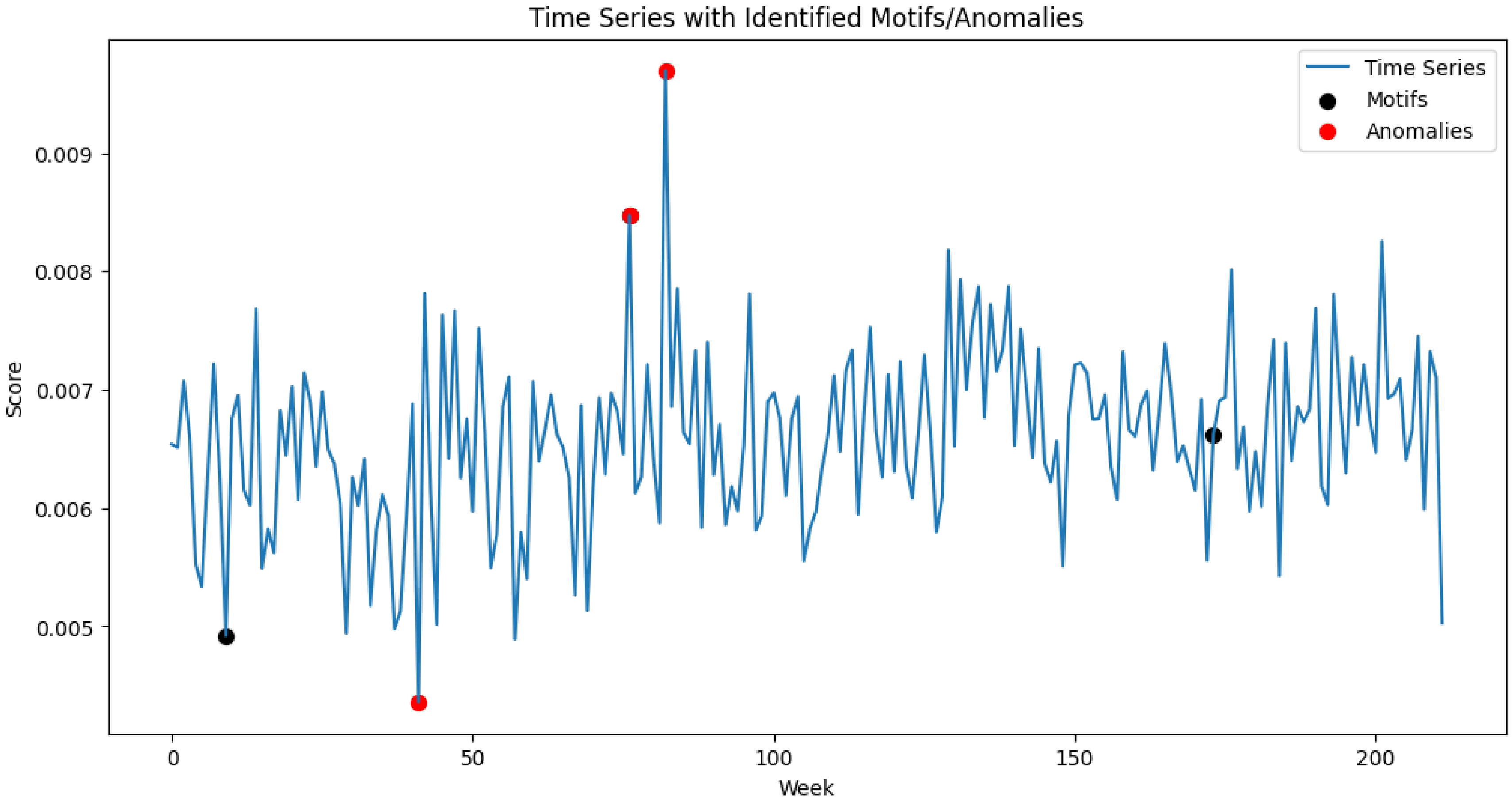

Cluster 5 - City Severity Evolution based on Time Series Data

Cluster 6 - City Severity Evolution based on Time Series Data

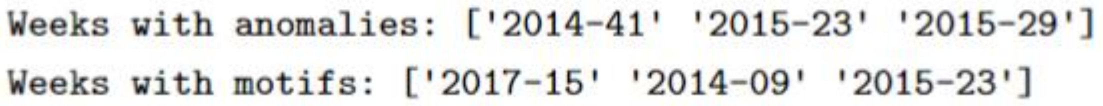
Cluster 7 - City Severity Evolution based on Time Series Data

Cluster 8 - City Severity Evolution based on Time Series Data

5.2 Motif and anomalies extraction

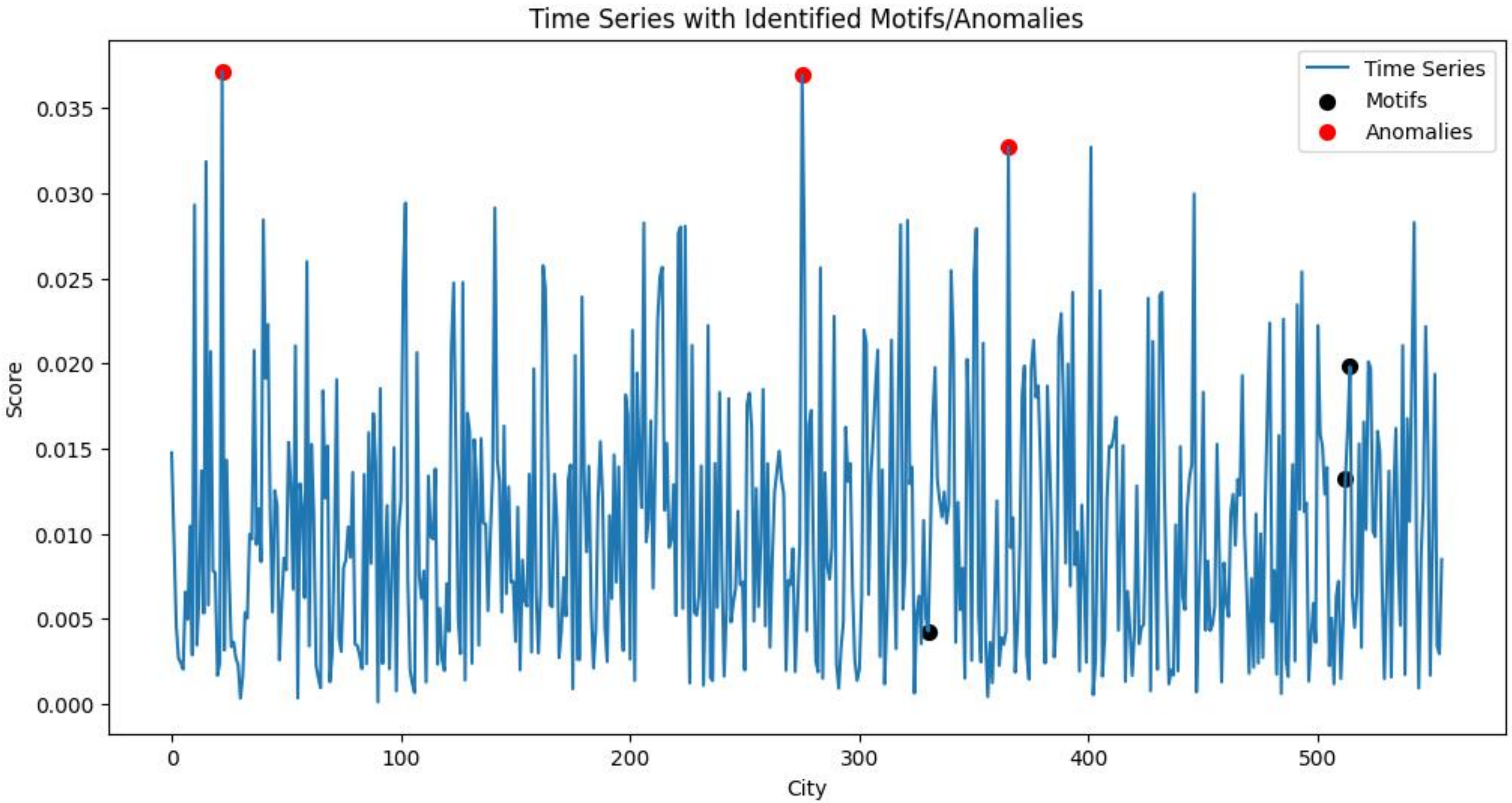


Time Series with Identified Motifs/Anomalies



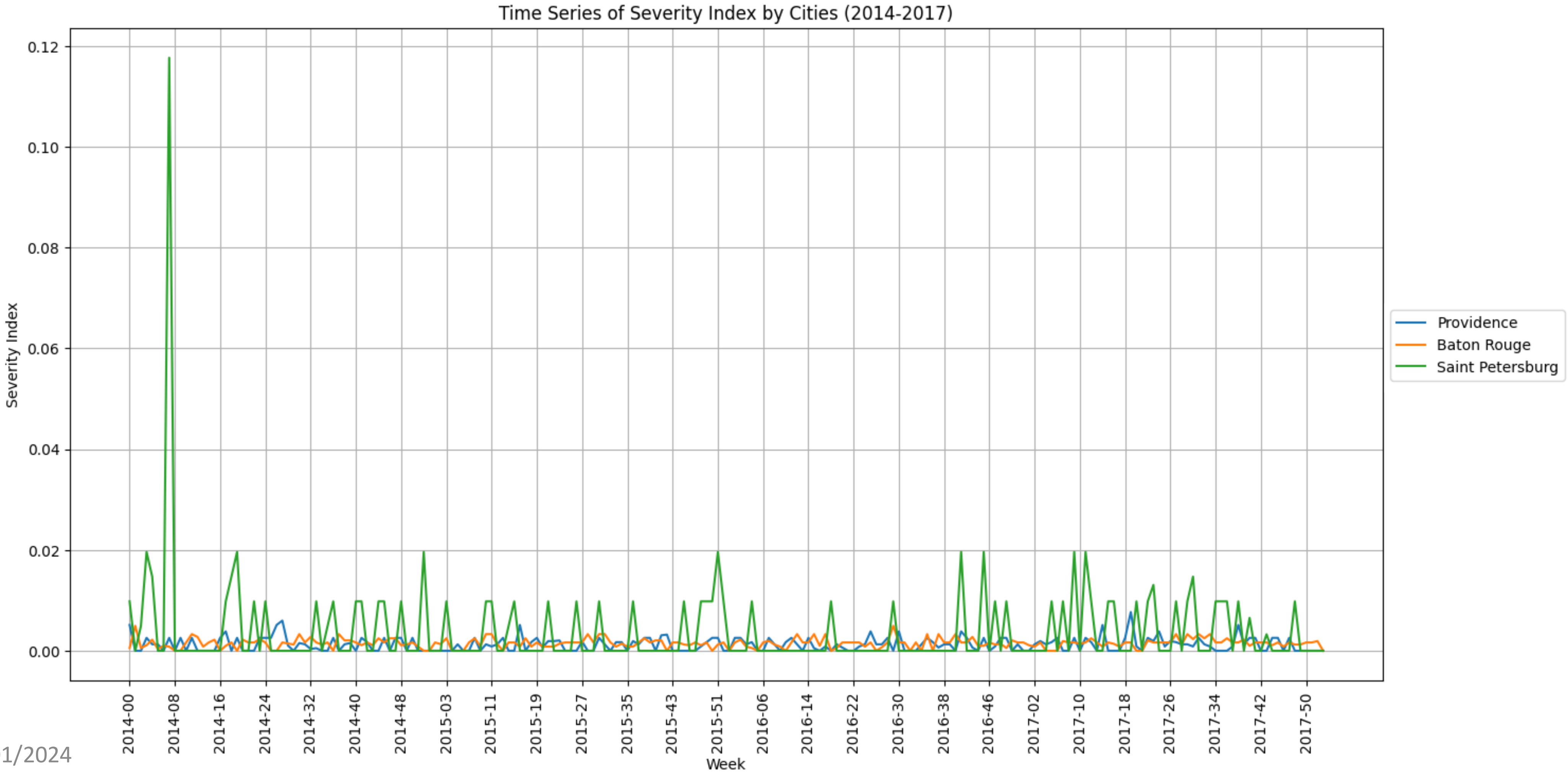
```
Weeks with anomalies: ['2014-41' '2015-23' '2015-29']
Weeks with motifs: ['2017-15' '2014-09' '2015-23']
```

5.2 Motif and anomalies extraction

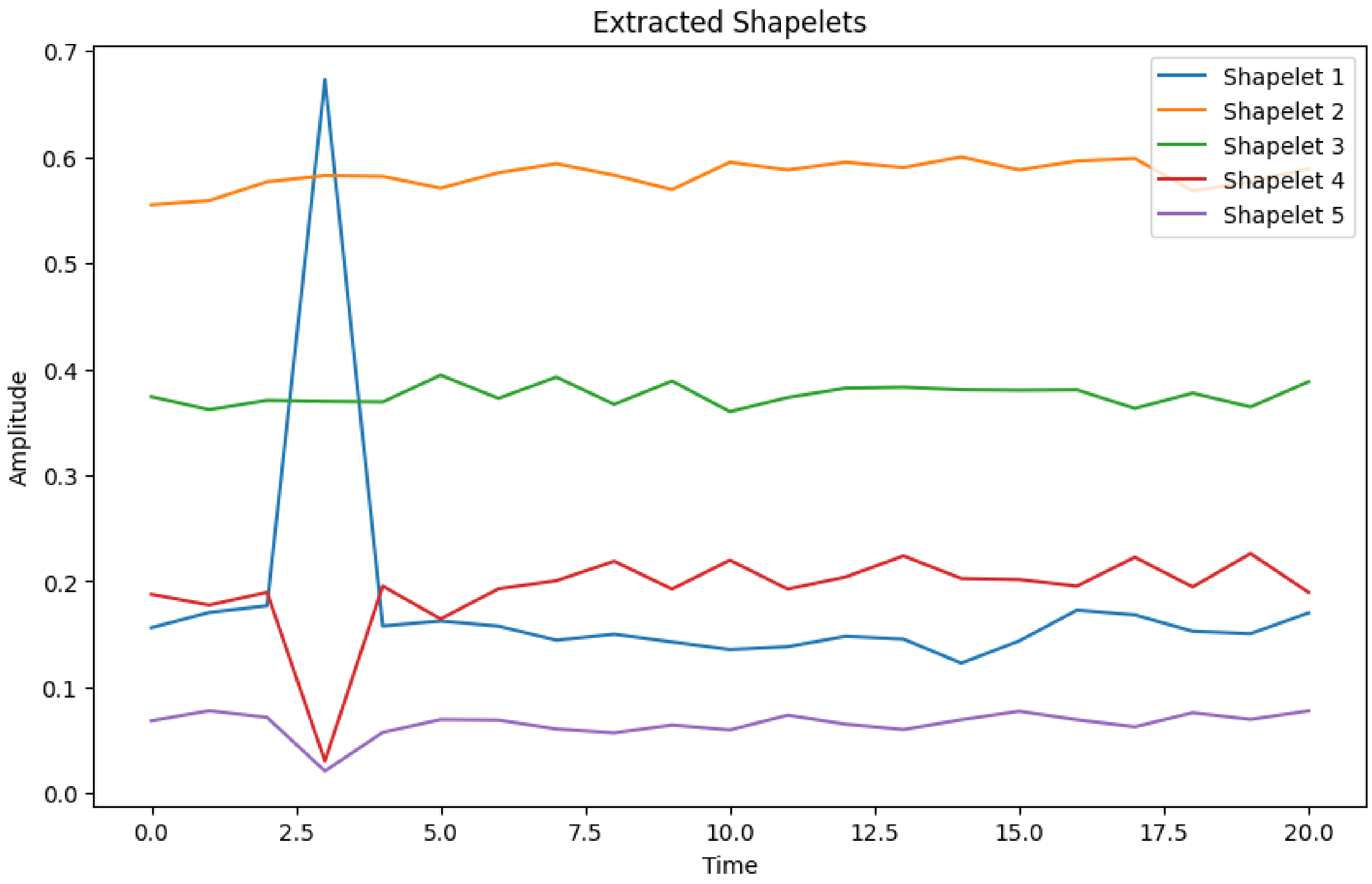


```
Cities with anomalies: ['Providence' 'Baton Rouge' 'Saint Petersburg']
Cities with motifs: ['Tucson' 'Long Beach' 'Springfield']
```

5.2 Motif and anomalies extraction



5.3 Shapelet extraction



Thank you for your attention