Quality of Life

Clustering of Italian Provinces

UNIVERSITÀ DI PISA

Data Mining and Machine Learning

By Iacopo Bicchierini

Highlights

Business Objectives

- Make local and state administrators aware about the situation
- Give hints to entrepreneurs for their activities

Data Mining Goals

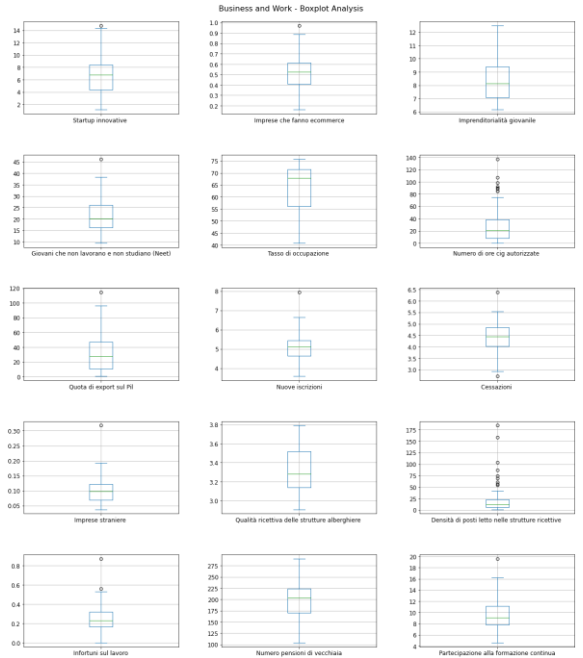
- Cluster Analysis
- Temporal Cluster Analysis

Dataset

Source: *IlSole24Ore* year 2020, 2021 and 2022

Exploratory Data Analyis

Box Plot Analysis



Correlation Analysis

	Cessazioni	Densità di posti letto nelle strutture ricettive	Giovani che non lavorano e non studiano (Neet)	Imprenditorialità giovanile	Imprese che fanno ecommerce	Imprese straniere	Infortuni sul lavoro	Numero di ore cig autorizzate	Numero pensioni di vecchiaia
Cessazioni									
Densità di posti letto nelle strutture ricettive									
Giovani che non lavorano e non studiano (Neet)									-0.73334
Imprenditorialità giovanile									
Imprese che fanno ecommerce									
Imprese straniere									
Infortuni sul lavoro									
Numero di ore cig autorizzate									
Numero pensioni di vecchiaia			-0.73334						
Nuove iscrizioni									
Partecipazione alla formazione continua									
Qualità ricettiva delle strutture alberghiere									
Quota di export sul Pil									
Startup innovative									
Tasso di occupazione	0.701139		-0.893498	-0.77364					0.832873

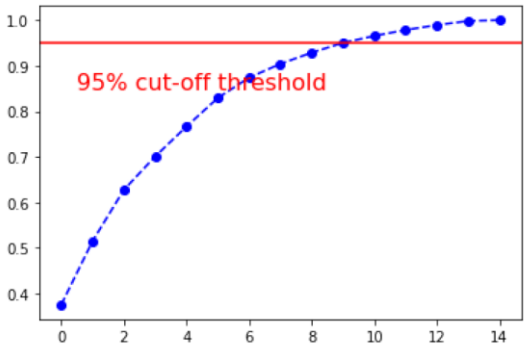
Preprocessing

Normalization

$$z = \frac{x - \mu}{\sigma}$$

Z-Score

Feature Reduction



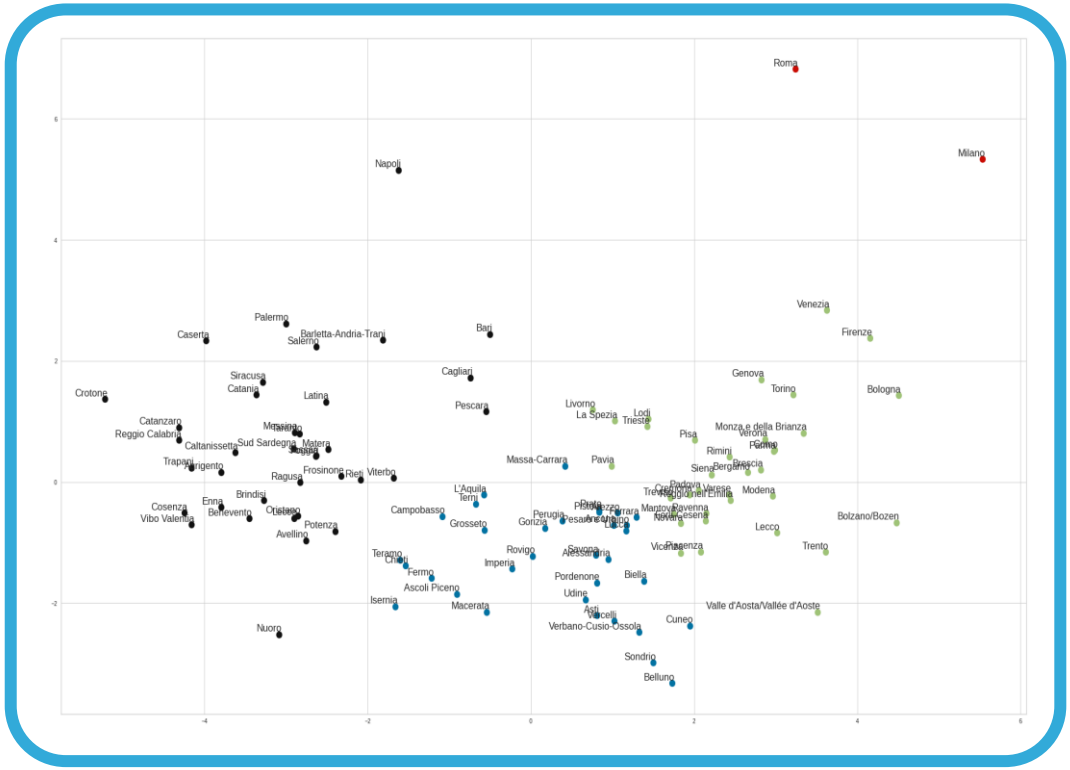
PCA with 95% variance explained

Cluster Tendency

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$$

Hopkins Statistic
of around 0.7 for all groups

Partitioning Clustering



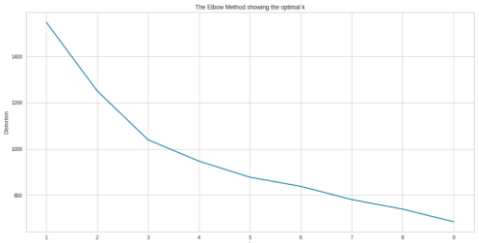
Wealth and Consumption 2022

Methods

- K-Means
- CLARANS

Choosing Number of Clusters

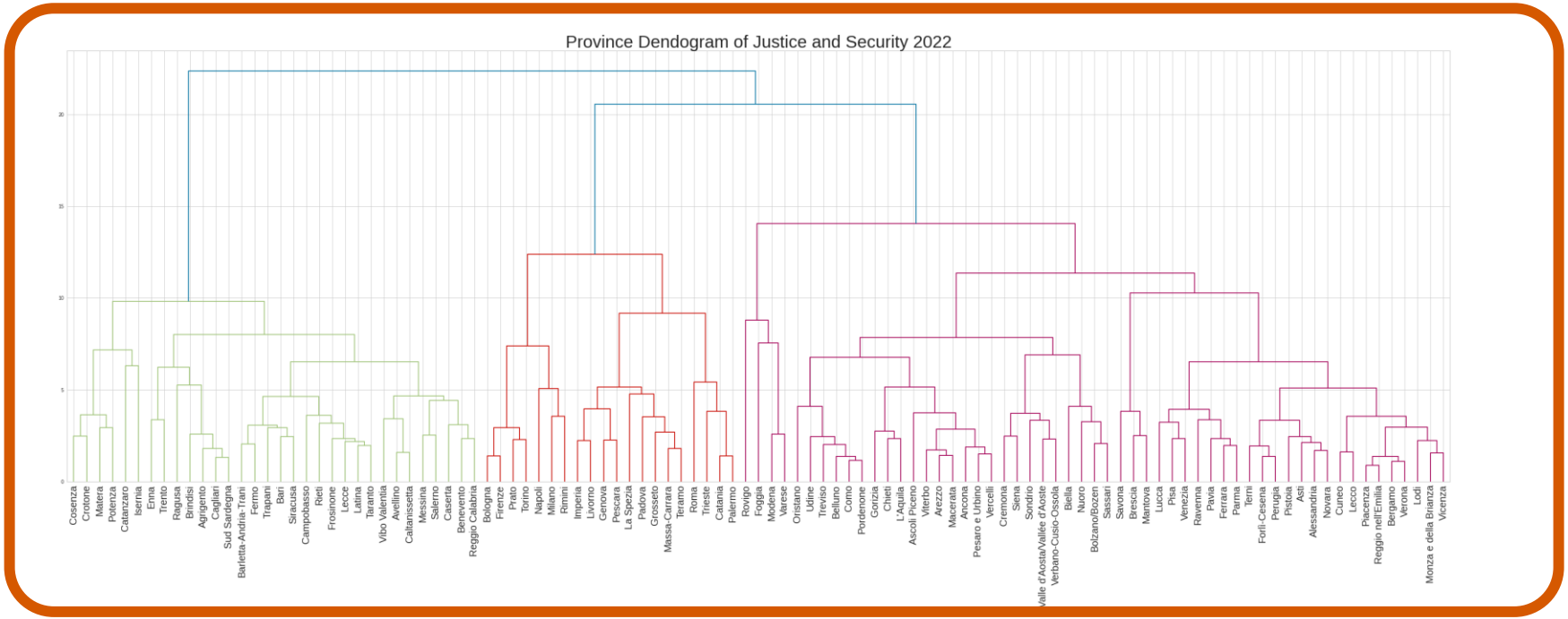
- Elbow Method
- Silhouette Analysis



Hierarchical Clustering

Methods

- Agglomerative
- BIRCH



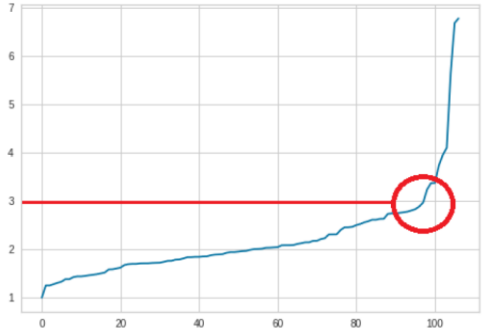
Density-based Clustering

Method

DBSCAN

Determine Parameters Eps and MinPts

Using k-dist graph



Total Fail

No global density values
for this dataset

Temporal Cluster Analysis

Matching Matrix

How many object from cluster i of window t-1 goes to cluster j of window t by the total number of objects

Clustering Method

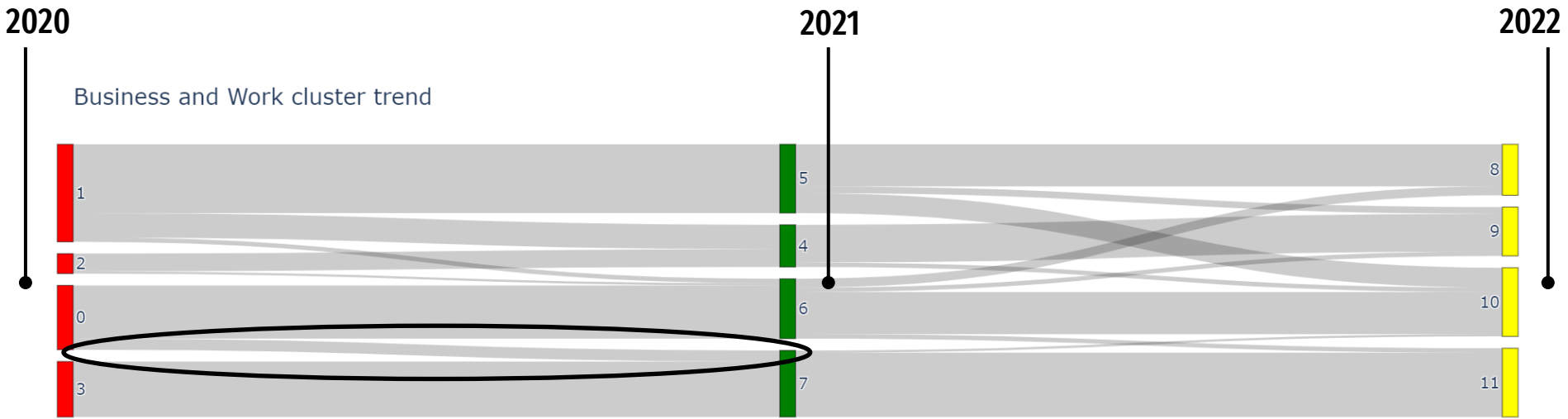
Same Clustering method used for the 3 years

Visualization

Using **Sankey Diagram**

Result

Intersection between clusters reveals who goes where



Evaluation

Silhouette Coefficent

Poor values

Best Clustering

K-Means

Best Indicator

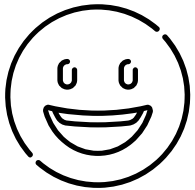
Demography and Society

Worst Clustering

CLARANS

	K-Means	CLARANS	Agglomerative	BIRCH
Business and Work	0.1567	0,096	0.1582	0.1614
Culture and Leisure	0.1913	0.1753	0.1816	0.1519
Demography and Society	0.199	0.1748	0.184	0.1796
Environment and Services	0.1557	0.085	0.1182	0.1097
Justice and Security	0.17	0.1	0.133	0.136
Wealth and Consumption	0.1965	0.1815	0.1785	0.123
Average	0.1782	0.1354	0.159	0.1436

THANK FOR YOUR ATTENTION



GitHub: https://github.com/Bicchie/Quality-of-Life-Italian-Provinces-Clustering