Classifying Buenos Aires Neighborhoods by Quality of Life

A Data Science project using clustering, PCA, and geospatial data

AUTHOR: IVAN OSIPOV

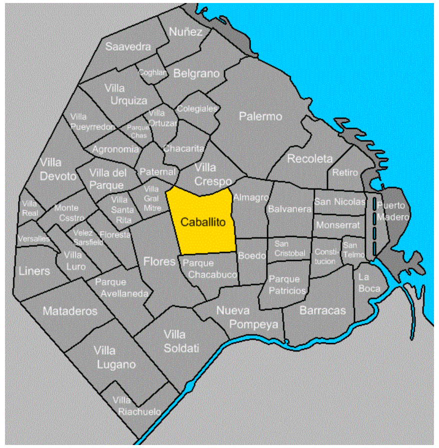
LOCATION: BUENOS AIRES, ARGENTINA

PERSONAL DATA SCIENCE PORTFOLIO

21.04.2025

Introduction: Why Classify Neighborhoods?

- Buenos Aires is a diverse city with contrasting neighborhoods.
- Urban planning, resource allocation, and social programs benefit from zone-based analysis.
- **Goal:** Group neighborhoods into meaningful zones using open data and machine learning.

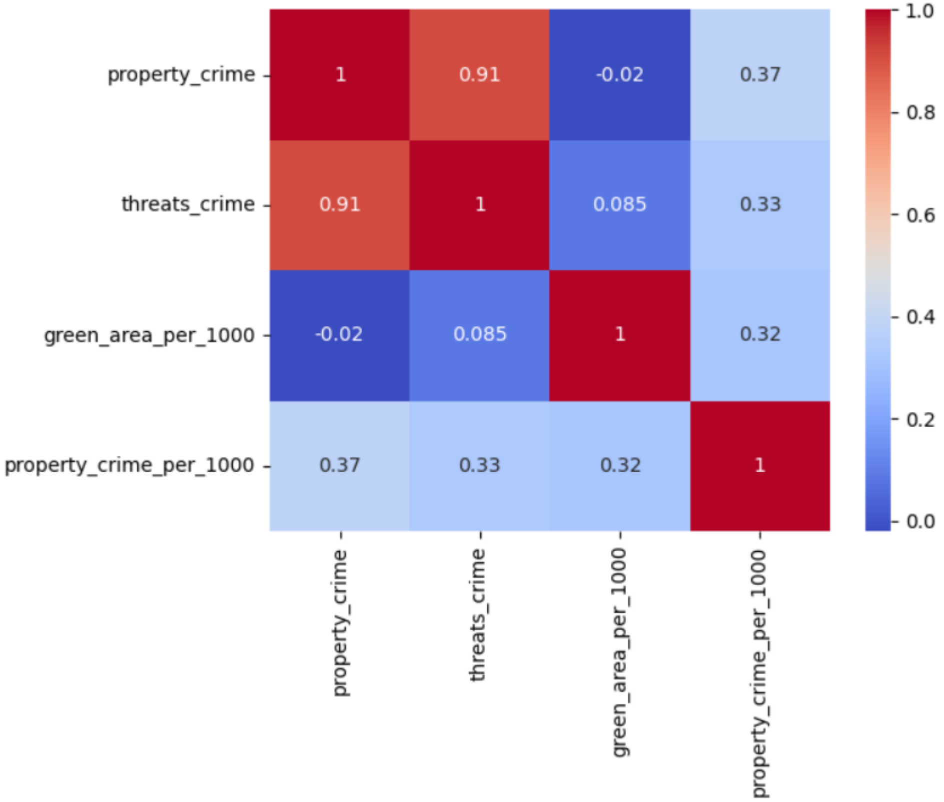


Data & Features: What Data Did We Use?

- Data sources:
  - BA Open Data Portal
  - INDEC Population Census
  - Slum and crime datasets
- Features:
  - Slum density, crime per 1000, green area %, hospitals per 1000, noise levels, schools per 1000, population density, etc.
- 24 features engineered → 10 principal components (PCA)
- Pipeline scheme:
  - Raw Data → Feature Engineering → PCA → Clustering

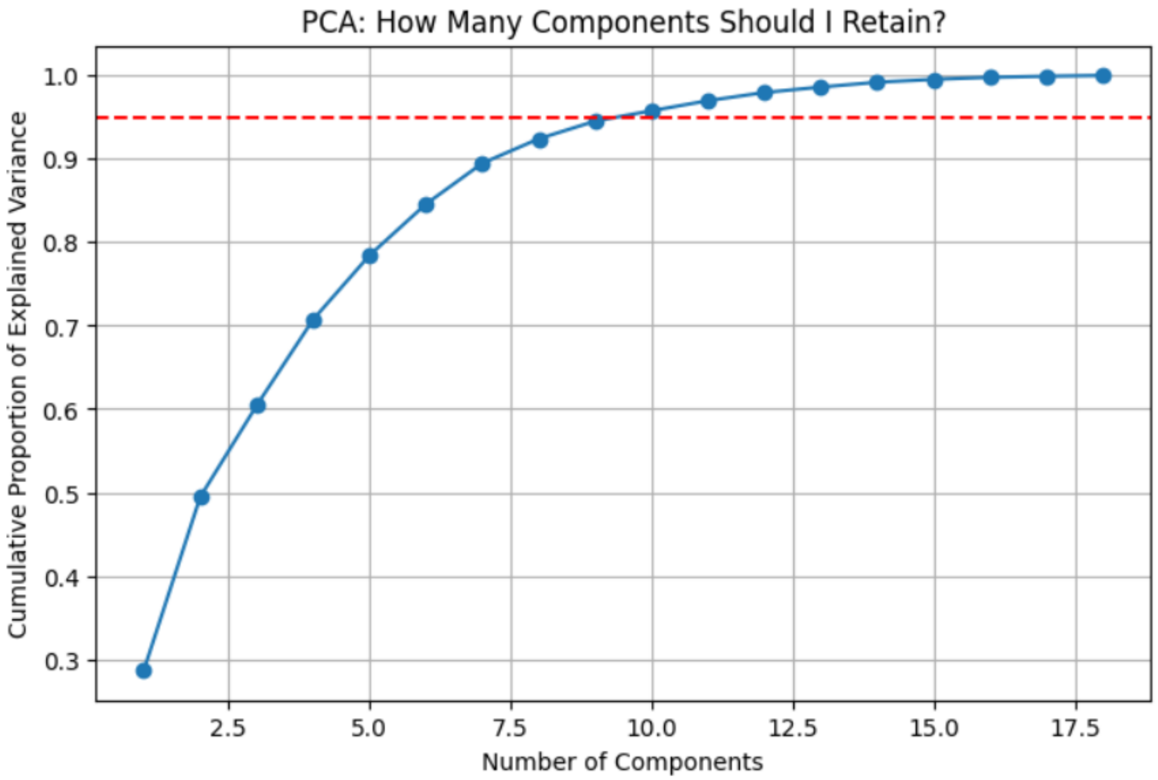
Methodology: From Raw Data to Feature selection.

- Standardization of features using StandardScaler to ensure fair comparison between features with different units and scales.
- Feature selection begins by computing the matrix of pairwise Pearson correlation coefficients between all standardized features. This allows us identify multicollinearity — features that are strongly correlated (> 0.9). Then drop one feature from each highly correlated pair.



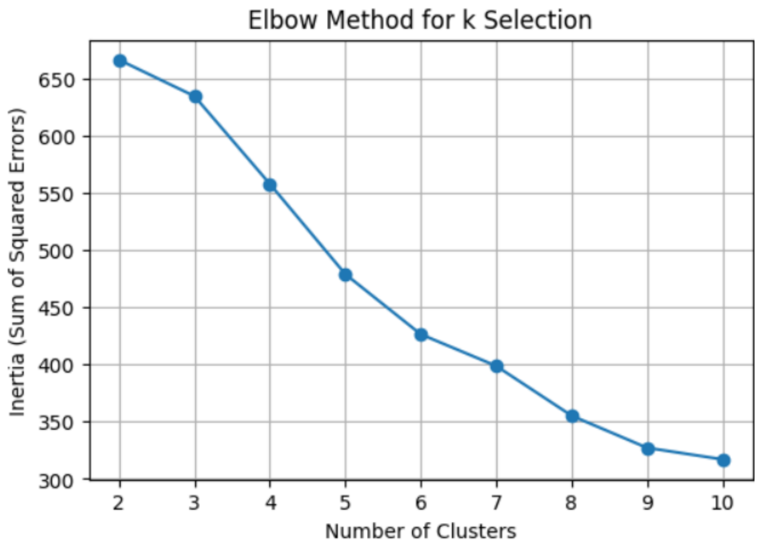
Methodology: Principal Component Analysis

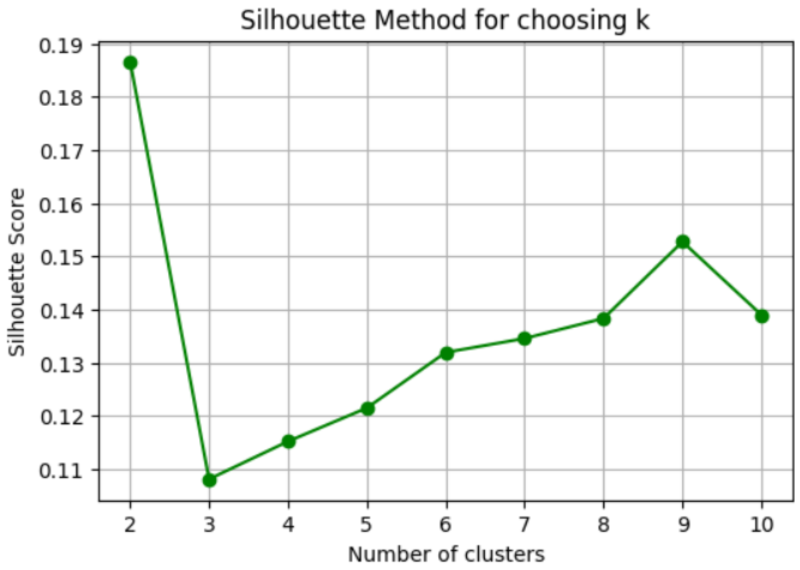
- Dimensionality reduction with PCA
  - Cumulative proportion of explained variance by each principal component
  - The number of components that explain at least 95% of the total variance in the dataset was selected.



Methodology: Optimal Number of Clusters

- Elbow Method
  - We analyze the inertia to find the "elbow" point — the value of **k** after which inertia decreases more slowly
- Silhouette Method
  - We calculate the silhouette score for each **k** — a metric that measures how well each object lies within its cluster

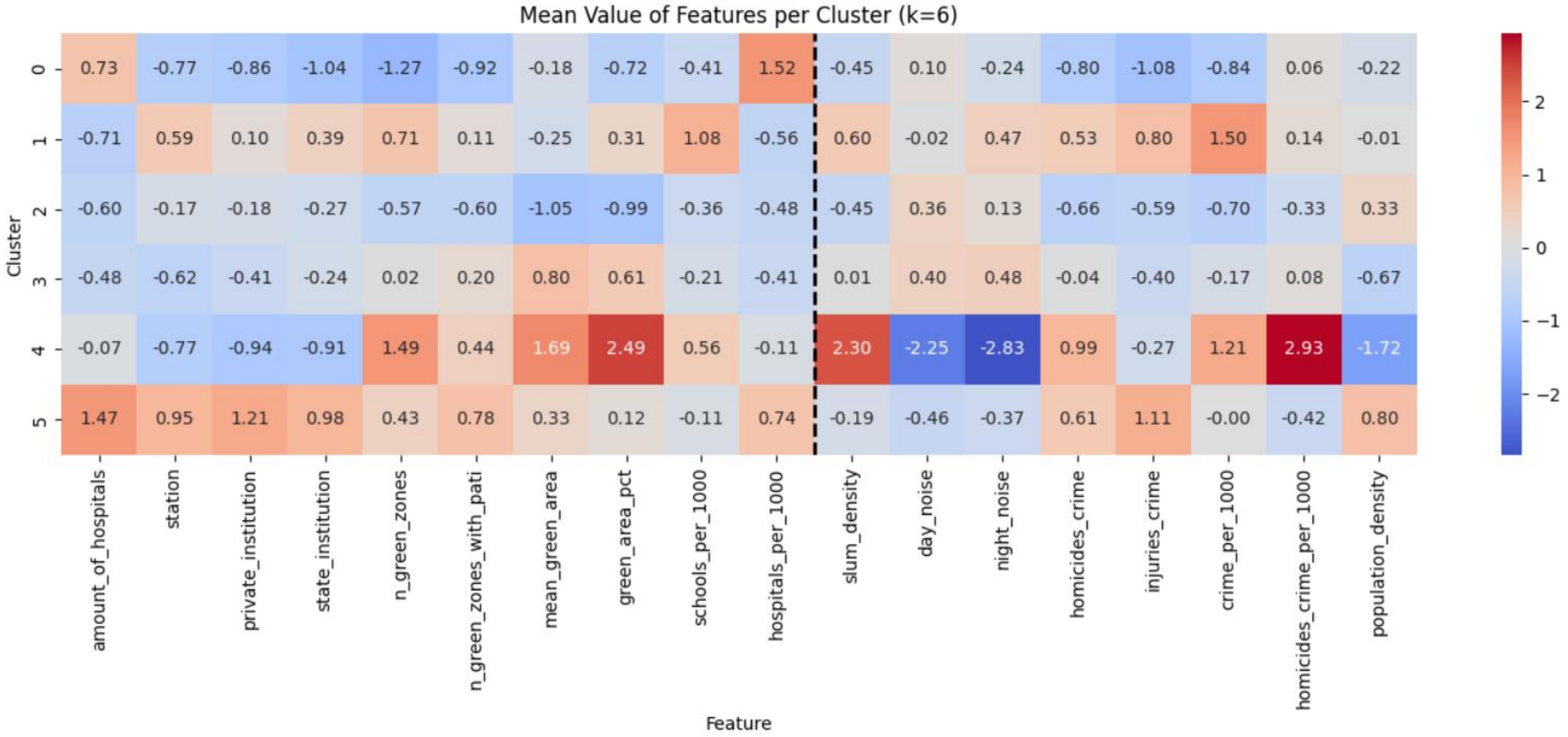




Methodology: Final Clustering with KMeans (k = 6)

- Based on the analysis of the Elbow and Silhouette methods, we decided to use k = 6 as the optimal number of clusters for KMeans.

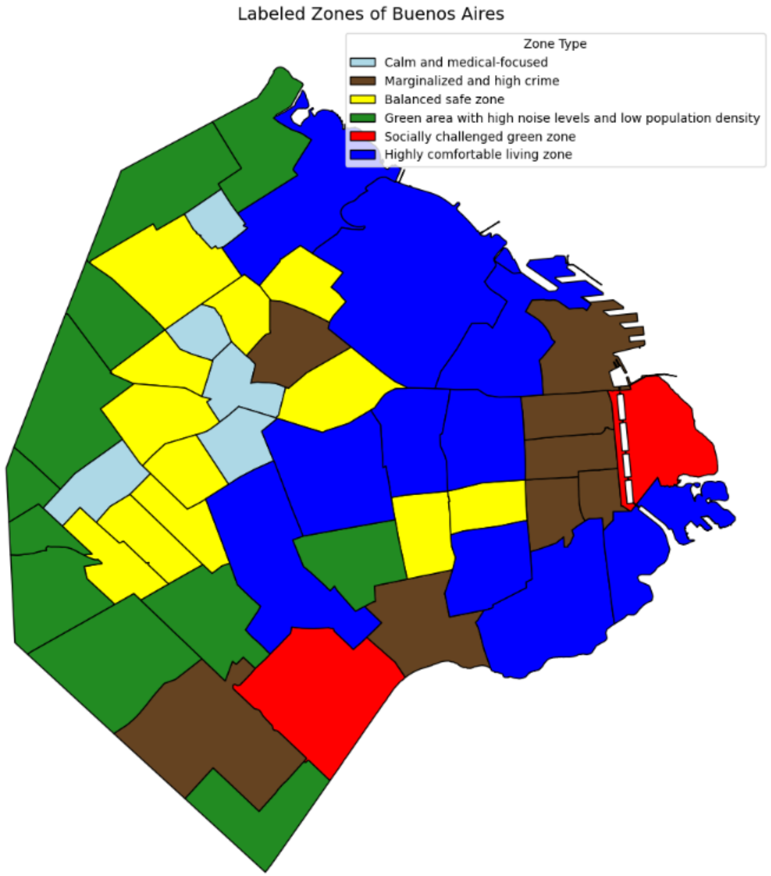
- In order to generalize and highlight key traits of each cluster, we compute and visualize the mean values of all features per cluster.



Visualizing the City: Labeled Clusters of Buenos Aires on the Map

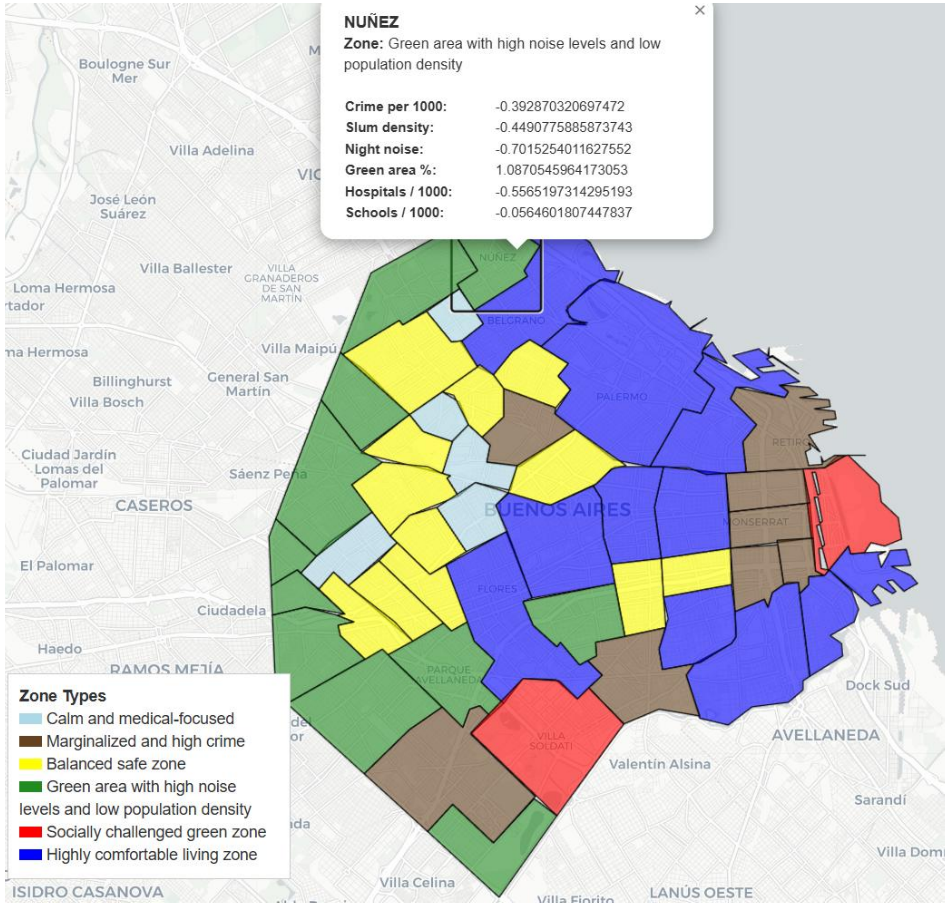
- Here we are able to see a visual result of our clustering

- Map was created using matplotlib.pyplot of manually labeled clusters



Visualizing the City: Interactive Folium Map

- Final map includes clickable neighborhoods
- Popups show cluster label + selected indicators
- Fully interactive (Folium)
- [Link to my GitHub with interactive map]



Conclusion & Future Work: What Did We Learn?

- Data-driven clustering reveals useful city zoning
- Socioeconomic and environmental variables align with perceived quality of life
- This approach can support urban planning, social research, and public communication
- The methodology is scalable and can be applied to other cities

Contact information

- LinkedIn
- GitHub
- Ivan.Osipov.job.mail@gmail.com