Understanding Municipal Resource Distribution in Israel: A Comprehensive Analysis of Local Authorities in 2018

Presented by:
Odeya Hazani - 207288457
Roni Epstein - 211645825

Abstract

Understanding the distribution of resources and characteristics across municipal authorities is crucial for effective governance and policy-making in Israel. This study analyzes comprehensive data from 255 local authorities in Israel (77 municipalities, 124 local councils, and 54 regional councils) to investigate patterns in budget allocation and identify key factors that differentiate various types of authorities. The dataset, sourced from data.gov.il and the Central Bureau of Statistics (CBS), encompasses multiple domains including demographics, education, welfare, infrastructure, and financial metrics. Through careful data preprocessing and analysis, this study aims to uncover meaningful patterns in how resources are distributed across different types of municipal authorities and what can be learned from this distribution. The findings have significant implications for policy makers and local governance.

Introduction

The purpose of this study is to analyze the relationships between different characteristics of local authorities and their budget allocation patterns. By examining this comprehensive dataset, we seek to identify factors that influence resource distribution and understand how different types of authorities manage their resources. This analysis can provide valuable insights for policy makers, municipal leaders, and researchers in the field of local governance.

Dataset and Features

The data was obtained from data.gov.il, Israel's official government data portal, which aggregates data from various government ministries and makes it freely available to the public. The Central Bureau of Statistics, established shortly after the state's founding, provided the underlying statistical information. The CBS continues to serve as Israel's primary statistical data collection and analysis organization, ensuring the reliability and comprehensiveness of the data.

Our dataset contains information on all 255 local authorities in Israel for the year 2018, covering 397 different variables organized into major categories including:

Basic authority information (status, district, religious character), Regular budget (income and expenses), Development budget, Municipal taxes (arnona), Demographics, Education, Health, Employment and welfare, Construction and housing, Infrastructure, Transportation, Crime and justice, Water management, Waste management, Land usage, Elections, Indices and rankings, Regional councils, Migration, Family and marriage, Institutional population ,Welfare services.

Data Preprocessing and Exploratory Data Analysis (EDA)

Initial Data Cleaning

The first stage of data preprocessing involved handling various types of missing values according to the official documentation. Special markers in the dataset were processed as follows:

- ".." (unknown or unpublishable data)
- "-" (no cases)
- "0" (value less than half the unit of measurement)
- "0.0" (value less than 0.1)
- Empty cells (statistically insignificant data)
- "*" (temporary name)
- "()" (estimated data or data with high sampling error)

Missing Value Treatment

We implemented a sophisticated approach to handling missing values, with different strategies based on the percentage of missing data in each column:

1. Columns with <10% missing values:
   - Filled using median values
   - This approach was chosen due to non-normal distributions and wide value ranges in the data
2. Columns with 10-30% missing values:
   - Implemented KNN (K-Nearest Neighbors) algorithm
   - Used carefully selected relevant features
   - Applied n_neighbors=5 for optimal imputation
   - Distribution of missing values was monitored before and after imputation
3. Columns with >30% missing values:
   - Removed from the dataset to maintain data quality
   - This resulted in 384 remaining variables from the original 397

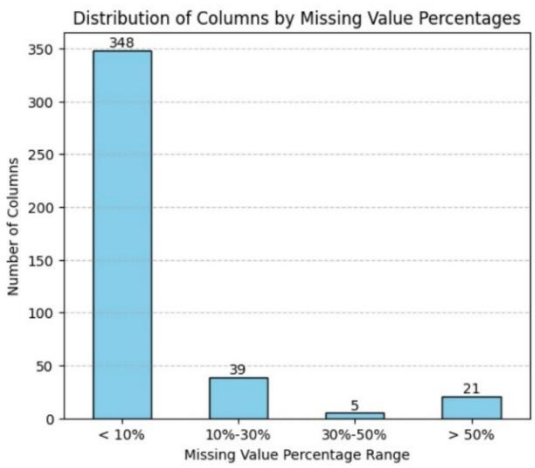
Outlier Detection and Analysis

Rather than automatically removing outliers, we employed a thoughtful approach to outlier analysis:

Initial Detection:

- Utilized Isolation Forest algorithm for anomaly detection
- Applied only to numerical columns
- Identified outliers based on probability thresholds

1. Outlier Evaluation:
   - Conducted in-depth analysis of identified outliers
   - Major cities like Jerusalem and Tel Aviv were identified as statistical outliers
   - Retained these cases as they represent meaningful variations in the data



Problematic Data Analysis

We implemented an IQR (Interquartile Range) based methodology to identify columns with high proportions of extreme values:

- Problematic columns were evaluated for their impact on model performance
- Outlier retention was determined by analytical merit
- Columns showing significant impact were retained for future analysis

Final Data Preparation

The dataset was standardized and prepared for analysis:

- Column names were translated from Hebrew to English
- Key numerical variables were validated and standardized
- The final cleaned dataset was exported to Excel format for subsequent analysis

This comprehensive data preparation process ensures the reliability and quality of our analysis while preserving important variations in the data that could provide valuable insights into municipal governance patterns.

Methodology: Unsupervised Learning Analysis

Data Preparation and Topic-Based Segmentation

Our analysis began with a systematic approach to handle the complex, multi-dimensional nature of municipal data. The initial phase involved two key steps:

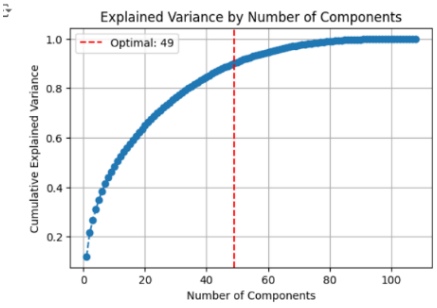
Data Type Processing

1. Categorical and numerical data separation
2. Strategic selection of critical categorical variables:
    - Authority status (regional/local council)
    - General district
    - Settlement Religion
3. Implementation of one-hot encoding for selected categorical variables

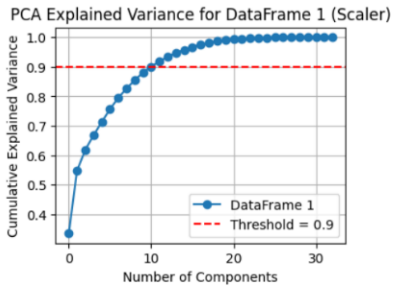
Topic-Based Data Division

Initial dimensionality reduction attempts on the complete dataset proved inadequate due to the diverse nature of the variables. This led to a strategic decision to segment the data into 22 distinct topic-based groups, including: Economic indicators, Transportation metrics, Demographic statistics, Educational measures, Welfare indicators.
This segmentation was implemented using column prefix patterns, allowing for more focused analysis within each domain.

Multi-Stage Dimensionality Reduction



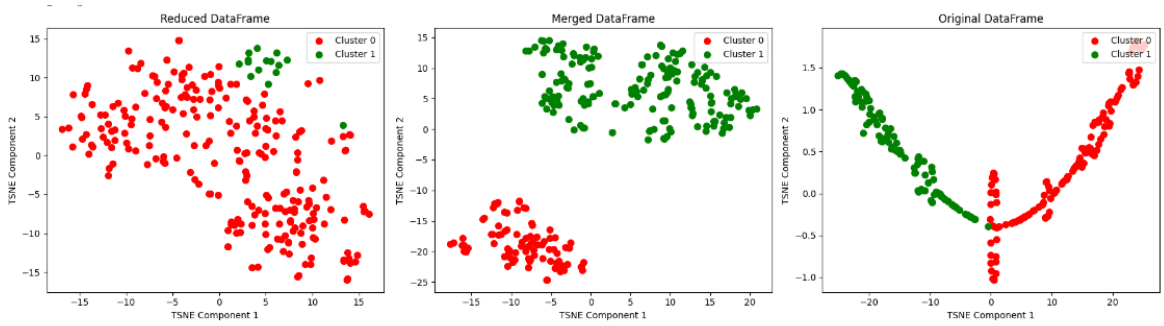
Optimal number of components: 49



DataFrame 1 (Scaler): Retained 11 components explaining 90.11% of variance.
Original columns: 33, Reduced columns: 11, Non-numeric columns: 0

Initial PCA Implementation

1. Standardization using MinMaxScaler (selected after comparative analysis with other scaling methods)
2. Individual PCA application to each topic group, maintaining 90% explained variance
3. Consolidation of PCA components from all groups into a unified dataset (108 columns)
4. Secondary dimensionality reduction on the consolidated data, resulting in 49 final columns

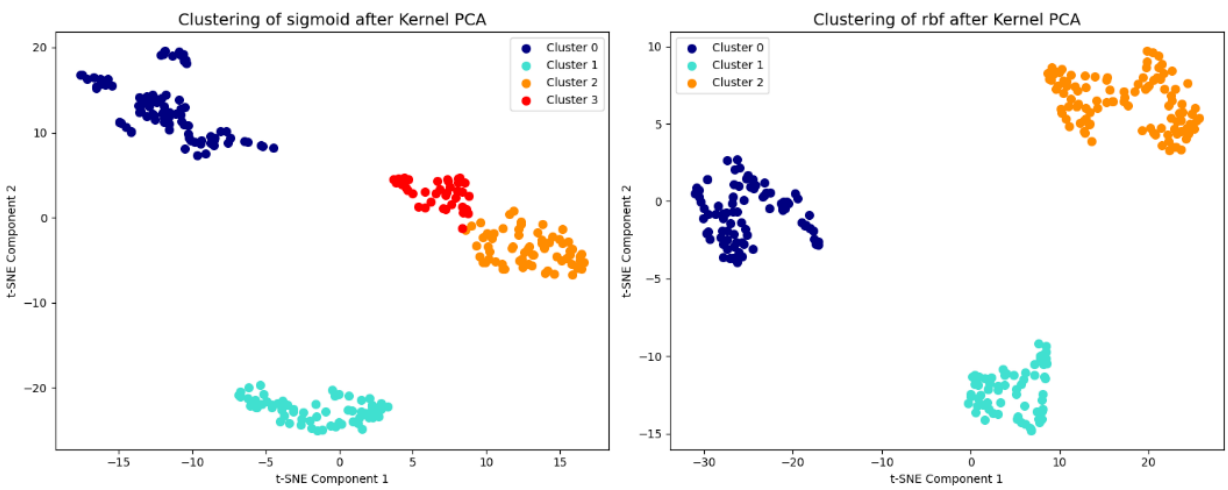


Advanced Dimensionality Reduction with Kernel PCA

The non-linear nature of our data led to the exploration of Kernel PCA methods:

1. Kernel Function Comparison:
   - RBF (Radial Basis Function)
   - Sigmoid function
2. Performance Evaluation Metrics:
   - Silhouette Score
   - Calinski-Harabasz Index
   - Davies-Bouldin Index
   - Inertia

RBF kernel demonstrated superior performance, maintaining 90% explained variance while capturing non-linear relationships more effectively than both Sigmoid kernel and standard PCA.

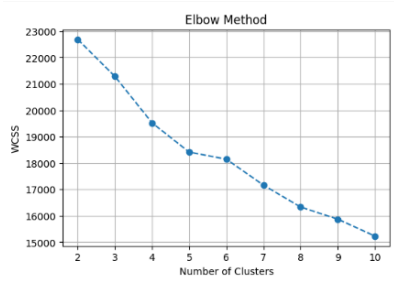


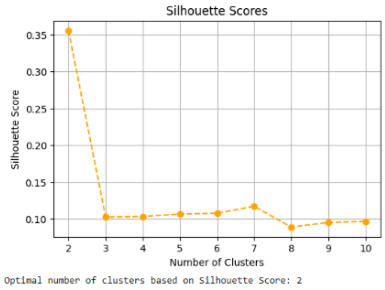
Clustering Analysis and Optimization

Cluster Number Optimization

Multiple methods were employed to determine the optimal number of clusters:

1. Elbow Method: Analysis of WCSS (Within-Cluster Sum of Squares) degradation
2. Silhouette Score: Evaluation of cluster separation quality
3. TSNE visualization for cluster quality assessment

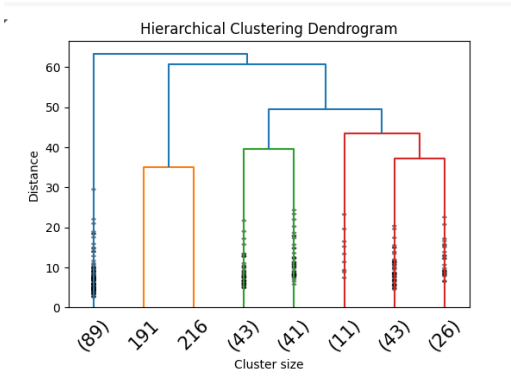


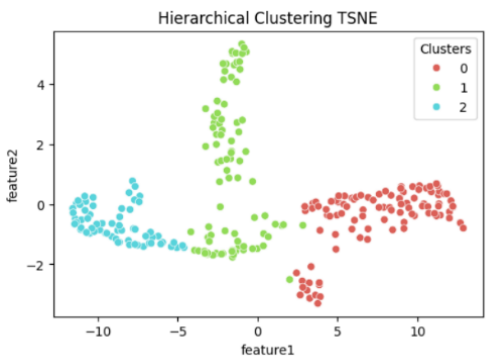


Hierarchical Clustering Validation

To validate our clustering approach:

1. Implementation of hierarchical clustering using Ward Linkage
2. Dendrogram analysis for cluster structure visualization
3. Comparative metric analysis:
    - Silhouette Score
    - Calinski-Harabasz Index
    - Davies-Bouldin Index





Final Model Selection

After comprehensive evaluation, the final methodology was selected based on:

1. Kernel PCA with RBF function
2. Optimization to three clusters
3. Validation through multiple quantitative metrics and visual TSNE analysis

This approach was chosen due to:

- Superior performance in quantitative metrics
- Clear cluster separation in visual analysis
- Robust performance across different validation methods
- Meaningful interpretation potential for municipal analysis

The final methodology provides a balance between dimensionality reduction effectiveness and clustering quality, while maintaining interpretability of the results.

Results and Discussion

Clustering Analysis Results

Model Selection and Validation

Our analysis revealed that Kernel PCA with RBF kernel provided superior clustering results compared to other methods. While the Elbow method and initial Silhouette analysis suggested four clusters for RBF, deeper analysis using multiple metrics supported a three-cluster solution:

- Silhouette Score: RBF (0.508) vs Sigmoid (0.544)
- Davies-Bouldin Index: RBF (0.868) vs Sigmoid (0.686)
- Calinski-Harabasz Index: RBF (303.32) vs Sigmoid (348.34)
- Inertia: RBF (11.51) vs Sigmoid (61.94)

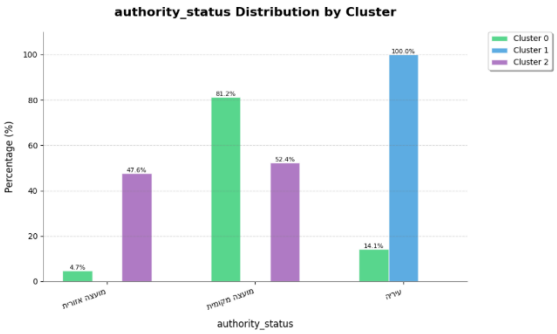
The three-cluster solution not only showed better metric performance but also provided more interpretable and meaningful groupings of municipalities.

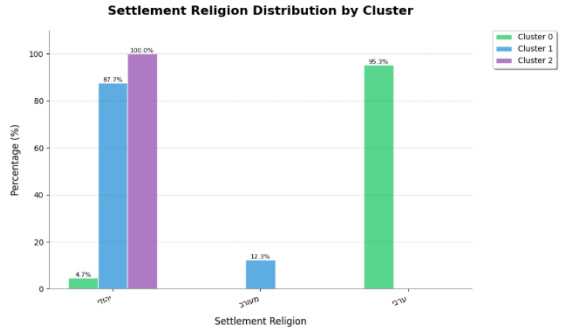
Cluster Distribution

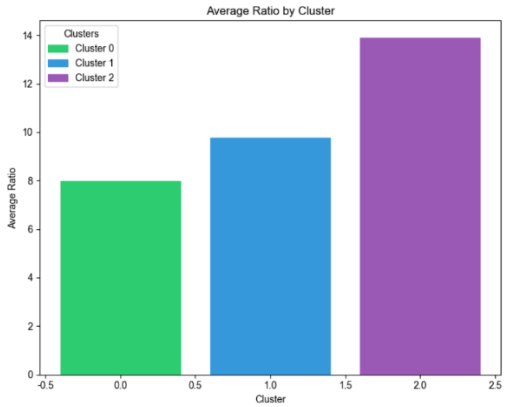
The final clustering resulted in the following distribution:

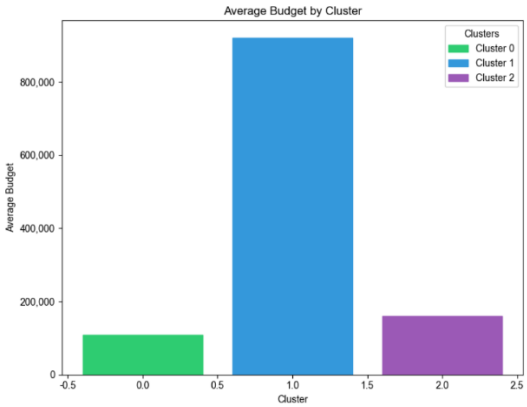
- Cluster 0: 85 municipalities (33.3%)
- Cluster 1: 65 municipalities (29.5%)
- Cluster 2: 105 municipalities (41.2%)

Cluster Characteristics and Insights









Cluster 0: Peripheral and Rural Arab Municipalities

Key Characteristics:

- Geographic: Furthest from Tel Aviv (88.8 km average)
- Demographics: Small population (average 15,035)
- Land Area: Smallest average area (9 km^2)
- Religious Composition: 95.3% Arab municipalities
- Budget: Lowest per capita budget ratio (7.98)

Challenges:

- Negative migration balance
- Limited infrastructure
- Restricted budget resources

Cluster 1: Urban Mixed Population Centers

Key Characteristics:

- Geographic: Moderate distance from Tel Aviv (52.2 km average)
- Demographics: Largest population (average 94,923)
- Land Area: Medium sized (26.7 km^2)
- Religious Composition: 87.7% Jewish, 12.3% mixed
- Budget: Highest total budget (921,573₪)

Notable Features:

- Strong positive migration (234.4)
- High accident rates
- Robust educational infrastructure
- Economic diversity within cities

Cluster 2: Established Central Jewish Municipalities

Key Characteristics:

- Geographic: Closest to Tel Aviv (29.7 km average)
- Demographics: Moderate population (average 13,877)
- Land Area: Largest area (174.3 km^2)
- Religious Composition: 100% Jewish
- Budget: Highest per capita budget ratio (12.59)

Strengths:

- Strong resource base
- Stable population
- Efficient infrastructure

Implications and Significance

Resource Distribution

The analysis reveals significant disparities in resource allocation:

- Cluster 0 municipalities face the most severe resource constraints
- Cluster 1 shows high total budgets but moderate per-capita resources
- Cluster 2 demonstrates the most efficient resource utilization per resident

Development Patterns

The clustering highlights three distinct development patterns:

1. Peripheral development challenges (Cluster 0)
2. Urban growth and diversity management (Cluster 1)
3. Suburban stability and resource efficiency (Cluster 2)

Policy Implications

The results suggest several policy considerations:

1. Need for targeted resource allocation to Cluster 0 municipalities
2. Infrastructure development focus for Cluster 1's growing populations
3. Maintenance of stability and efficiency in Cluster 2

The clustering analysis provides a data-driven framework for understanding municipal resource distribution and development patterns in Israel, offering valuable insights for policy makers and municipal leaders.

Contributions

Throughout the project, we divided our tasks based on our areas of expertise and interests, while maintaining continuous collaboration. One of us focused primarily on the development and implementation of classification processes in the structured analysis (Supervised Learning), while the other concentrated on unsupervised learning techniques (Unsupervised Learning). All decisions, including algorithm selection and workflow planning, were made collaboratively through joint discussions and teamwork. We were both actively involved in each other's progress, contributing to coding and problem-solving whenever needed.

The report and presentation were written together as a joint effort, with a fair division of tasks, frequent discussions, and a shared focus on combining our perspectives to create a high-quality final result. We spent a significant amount of time working on the report, dedicating many hours to ensure its thoroughness and quality. Our teamwork was excellent, characterized by open communication, mutual support, and a strong commitment to achieving the best possible outcome.

Appendices

https://towardsdatascience.com/an-approach-for-choosing-number-of-clusters-for-k-means-c28e614ecb2c

https://www.kaggle.com/code/marcinrutecki/outlier-detection-methods#7.-DBSCAN---Density-Based-Spatial-Clustering-of-Applications-with-Noise

https://www.kaggle.com/code/victorambonati/unsupervised-anomaly-detection#2.5-Isolation-Forest

https://www.kaggle.com/code/kashnitsky/topic-7-unsupervised-learning-pca-and-clustering