### **QMBU 420/520**

**FİNAL Project Report**

**Team 2**

**01.06.2025**

### **1. Business Understanding**

#### **1.1 Project Objective**

The objective of this project is to analyze the quality of Emergency Assembly Zones (EAZs) in Istanbul using unsupervised learning techniques to geospatial data. The desire is to categorize the zones based on their physical and demographic attributes to determine which of them might be under increased risk within an earthquake scenario. Through the clustering of the zones, we aim to provide data-informed recommendations to policymakers regarding targeted enhancement and emergency planning.

#### **1.2 Problem Context**

#### Istanbul is one of the world's most seismically active cities, straddling the North Anatolian Fault. In the event of a devastating earthquake, Emergency Assembly Zones are designed to be first-gathering and safe zones for citizens. But numerous reports and academic studies\*\* have suggested a great many existing zones are either unsatisfactory, poorly located, or inaccessible. So there is a requirement for an evidence-based evaluation of today's EAZ network.

#### **1.3 Success Criteria**

The success of this project hinges on our ability to:

* Effectively retrieve and process geospatial EAZ data from official sources
* Engineer relevant features such as buffer-based building counts and zone area
* Cluster the zones meaningfully using unsupervised techniques like K-means
* Interpret the resulting clusters to identify low- and high-risk zones
* Provide actionable strategic insights based on the clustering results

The final measure of success is whether the analysis produces clearly interpretable groupings that align with real-world concerns and can guide improvements in disaster preparedness.

### **2. Data Understanding**

#### **2.1 Data Sources**

Our primary data source is the official "AFAD Emergency Assembly Area Inquiry" system. Due to the API of the system being able to support only afew zones to return per coordinate request (3 per lat-lon point),

In response to API limitations, we reverse-engineered data access mechanism and used a systematic sampling process to obtain full data. We enriched this data source with additional sources:

* **AFAD EAZ API**: Geographic coordinates, polygon boundaries, address information
* **OpenStreetMap (OSM)**: Building geometry and location
* **Atlas Big**: Neighborhood-level population data

#### These datasets enabled a multi-dimensional view of every zone, harmonizing spatial and demographic context to support stable clustering. In addition to further enhancing the dataset, we also performed feature engineering. Geospatial analysis was applied to create new features such as the number of buildings in 0–15m, 15–30m, and 30–50m of each zone. These enhanced features helped us create a multi-dimensional view of everyzone, balancing spatial proximity and demographic context to lead to a more stable process of clustering.

#### **2.2 Initial Exploration**

Upon retrieval, exploratory data analysis (EDA) was performed. We examined distributions of key variables, such as where we attach the graphical representations in the appendix:

* Area (in square meters)
* Population of the associated neighborhood
* Number of buildings within proximity buffers (0–15m, 15–30m, 30–50m)

These bar charts revealed notable difference in zone area and the resident population densityaround them, which justified clustering. Missing and invalid input entries (e.g., unmatchingneighborhood names) were also observed for cleaning purposes later on.

### **3. Data Preparation**

#### **3.1 Feature Engineering**

#### In order to assess the environment for each zone, we employed QGIS and OpenStreetMap (OSM) data to create buffer zones at 0–15m, 15–30m, and 30–50m ranges from each assembly area's polygon. We then counted the number of buildings in the buffers to create new numerical features recording surrounding density. We also calculated each zone's area (square meters) from its polygon geometry.

#### These artificially created details provide information on how packed or packed each region can be in a situation of crisis. An essential limitation, however, was identified to be: OSM lacks all buildings existing, specifically in under-mapped regions, which leads to potential underestimation of density indicators.

#### **3.2 Integration and Cleaning**

Neighborhood-level population data from Atlas Big was integrated into the dataset by matching zone addresses to official administrative divisions. This required name normalization and manual correction in some cases. In total, 23 zones were excluded due to unmatched or inconsistent neighborhood identifiers.

#### **3.3 Final Dataset Structure**

The final dataset combined spatial and demographic attributes for each EAZ:

* Area (m²)
* Buildings within 0–15m, 15–30m, 30–50m
* Neighborhood population
* Administrative identifiers (district, street, etc.)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Zone ID** | **Area (m²)** | **Bldgs 0–15m** | **Bldgs 15–30m** | **Bldgs 30–50m** | **Population** |
| 001 | 580 | 5 | 12 | 19 | 4,320 |
| 002 | 1320 | 0 | 3 | 8 | 7,890 |
| 003 | 2760 | 9 | 22 | 34 | 11,104 |

### **4. Modeling**

#### **4.1 Model Selection**

#### Since our data was unlabeled, it was resolved that clustering was the most appropriate method of modeling. It was to find naturally occurring clusters within the zones, based on their attributes. Out of the available clustering techniques, **K-means** was picked since it was easy to interpret and simple. It supports partitioning by Euclidean distance in feature space, which is ideally applicable in our scaled, numerical data.

#### **4.2 Model Building**

We applied standard scaling to scale the data before clustering. Although we first applied **Robust Scaler** to address skewness, we ultimately applied **Standard Scaler** because:

* Robust Scaler changed the nature of the data, especially affecting density-related features
* Our aim was to preserve the relative influence of area and building counts on cluster formation.

We used the **Elbow Method** to determine the optimal number of clusters. While k=5 was initially shown by the elbow chart, upon closer inspection, it was found that **k=4** provided more understandable and compact groups. See the appendix for the graphical plot of the Elbow Method.

#### **4.3 Model Performance**

To evaluate cluster quality, we employed three key metrics:

* **Silhouette Score**: 0.373 — indicates moderate separation between clusters
* **Calinski-Harabasz Index**: 2406.3 — high values denote compact, well-separated clusters
* **Davies-Bouldin Index**: 0.97 — values below 1 suggest tight and distinct clusters

### **5. Evaluation & Results**

#### **5.1 Cluster Interpretations**

The model produced four distinct clusters, each exhibiting unique spatial and demographic characteristics:

* **Cluster 0**: Small zones with low surrounding building density and low population — likely balanced zones with low stress in emergencies.
* **Cluster 1**: Moderately-sized zones with extremely high population — these areas pose serious risk of overcapacity and should be prioritized in urban planning.
* **Cluster 2**: Densely packed zones with high numbers of surrounding buildings and medium-to-high population — typical central urban areas with moderate risk.
* **Cluster 3**: Very large zones with numerous surrounding buildings and high population — may include large parks or misclassified zones and warrant further field verification. Also see the scatter plot in the appendix that demonstrate the distribution of the clusters, implying that our classification is distinct and valid.

#### **5.2 Strategic Recommendations**

* **Expand and Upgrade Cluster 1 Zones**High population and limited space make these zones prone to failure in emergencies. They should be urgently reviewed, expanded where possible, and prioritized for infrastructure improvements.
* **Monitor Cluster 2 and 3 Zones**These zones face moderate-to-high risk due to dense surroundings or large, complex layouts. Ongoing tracking of population and construction activity is essential to maintain readiness.
* **Preserve Cluster 0 Zones**Though low-risk, Cluster 0 zones serve as valuable reserve capacity. They should be protected from development and maintained for future needs.
* **Build a Dynamic Risk Assess** **ment Tool**The clustering model offers a solid foundation for a real-time EAZ monitoring system that can adapt to demographic and spatial changes.
* **Integrate Additional Risk Layers**Future work should include earthquake vulnerability, infrastructure access, and socioeconomic data to refine risk classification.

#### **5.3 Reflection & Lessons Learned**

This project showed the importance of combining open-source geospatial data with unsupervised learning techniques for public sector planning. Key insights:

* The importance of preprocessing and spatial feature engineering
* The sensitivity of model results to scaling decisions
* The real-world challenges of data completeness and administrative inconsistencies.

### **6. Team Contributions**

This project was the result of collaborative work across all stages of the CRISP-DM framework. Below is a summary of each member’s key contributions:

* **Mehmet Boran Göksel**: Led data scraping efforts using the reverse-engineered AFAD API, constructed the clustering pipeline in Python, and created visualizations.
* **Süha Nurhat**: Coordinated project planning and report writing, contributed to business understanding and evaluation sections, and led communication across team members along with the creation of the graphical representation.
* **Ali Kerem Uysal**: Handled buffer-based spatial analysis in QGIS, performed feature engineering, and assisted in data cleaning and integration.
* **Emir Bedirhan Aydın**: Oversaw population data merging, carried out exploratory data analysis (EDA), and refined model performance evaluation.

### **AppendixA chart of building counts Description automatically generated: Visuals:** Area and population distribution bar charts: A graph of a number of blue bars Description automatically generated with medium confidence

* A map of clustering data

  Description automatically generated**A graph with a line

  Description automatically generated**Elbow method plot:
* Cluster scatter plot:

**D. GitHub Repository** All code, figures, and modeling outputs can be accessed at:  
<https://github.com/devoidoflight/QMBU/blob/main/qmbu_clustering_v2.ipynb>