Spatial Analysis of the 2023 Earthquakes in Turkiye

Autor : Gozde Yazganoglu Delgado Doblas

Tutor: Diego José Bodas Sagi

September, 2023

Thesis presented for the Degree of
Master of Data Science

Abstract

Earthquakes in Turkey has caused 56,840 fatalities, 130,000 injured and affected somehow 1,900,000 people. Up to now it is biggest disaster of this year. In addition, it is estimated to have 34.2 billion economic lost due to damaged and destroyed buildings.

In this study we aim to see spatial patterns, regions, and their effect on Machine Learning models, to understand which factors should be considered to reconstruct the region. This is important because scientists expect a bigger Istanbul earthquake potentially will affect a much more populated region and important business centers and due to land structure it will be impossible to deliver help in case of a disaster. For these reasons, we have examined a building dataset by EU Copernicus Disaster database, merged with faults information and regional socioeconomic variables form the region.

In this study, supervised machine learning techniques were employed to understand the factors affecting damage levels. The data was modeled as a Multiclass Classification problem. An important aspect of the research was the handling of geospatial data, which required specialized skills and tools. The analysis revealed that across all regions, residential buildings, and critical infrastructure such as roads, railroads, and airfields sustained the most damage. The extent of the damage was found to be closely related to spatial variables, particularly distances to significant locations.

Table of Contents

Abstract 2
List of Graphics: 4
1. Introduction 6
1.1. Objectives of the study 6
2. State of Art 7
2.1. Geospatial Data Basics 7
2.2. Studies on Spatial Data 7
2.3. Studies of Machine Learning on Earthquake 8
3. Methodology 9
3.1. Data Collection 10
3.1.1. EU Copernicus Disaster Data 10
3.1.2. AFAD Earthquake Data information 11
3.1.3. TURKSTAT 11
3.1.4. Fault Data 11
3.2. Spatial Analysis Methods 11
3.2.1. Point Pattern Analysis 12
3.2.2 Ball Tree Algorithm and Its Use in Data Preparation Process 12
3.2.3. Spatial Lags, Weights and Autocorrelation 13
3.3. Machine Learning Algorithms and Tools 15
3.3.1. Supervised Modeling 16
3.3.1.1.Random Forest Classification 16
3.3.2. Unsupervised Modeling 16
3.3.2.1. K-Means Clustering 17
3.3.3. Interpretation 17
4. Spatial Analysis and Machine Learning Models of Buildings and Facilities 18
4.1. EDA and ESDA 18
4.1.1 EDA (Explanatory Data Analysis) 19
4.1.2. ESDA (Explanatory Spatial Data Analysis) 24
4.1.2.1. Point Patern Analysis 25
4.1.2.2. Spatial Autocorrelation 30
4.1.3. Insights 31
4.2. Machine Learning Models 33
4.2.1. Supervised Machine Learning 33
4.2.2. Unsupervised Machine Learning 38
Conclusion 42
References 44

List of Graphics:

Figure 1: Diagram of CRISP-DM Methodology 10
Figure 2: Schema of Ball Tree approach. 13
Figure 3 Random Forrest 16
Figure 4 K Means Clustering 17
Figure 5 Overview of the Dataset 19
Figure 6 Destroyed Buildings in Cities 20
Figure 7 Damaged Buildings in Cities 20
Figure 8 Damaged buildings according to object type 21
Figure 9 Destroyed Buildings according to obj_type 21
Figure 10 Boxplot of nearest_earthquake_distance 22
Figure 11 Boxplot Nearest Fault distance 22
Figure 12 Descriptive Statistics and new variables created using Ball Tree method. 23
Figure 13 Correlation Matrix of the dataset. 24
Figure 14 Map Representation 24
Figure 15 Scatter Point of All Observations. 25
Figure 16 Scatter points for destroyed and damaged buildings 26
Figure 17 KDE of all data 26
Figure 18 Quadrant Count for Real Data 27
Figure 19 Quadrant Count of randomized simulation of same data 27
Figure 20 Ripley's G(d) distribution of Randomness. 28
Figure 21 Ripley's F(d) Distribution of Randomness 28
Figure 22 Mean, Median and Center points and comparison with important events on 6th of February 29
Figure 23 Spatial Mapping for the percentage variable 31
Figure 24 Spatial Mapping for Destroyed and Damaged Regions 32
Figure 26 Feature Importance Plot for the Random Forest Model 35
Figure 27 Confussion Matrix for Random Forest Model 35
Figure 28 Classification Matrix of the Random Forest Model 36
Figure 29 Class Prediction Error for Random Forest Classifier 36
Figure 30 Validation Curve for Random Forrest Model 37
Figure 31 Mean SHAP Value average impact on model output. 37
Figure 32 Shap Dependency for Info- Damaged Percentage 38
Figure 33 TSNE Plot refletting clusters in 3D visualization 39
Figure 34 Silhouette Plot of KMeans Clustering 40
Figure 35 Elbow Graph to determine which number of clusters are the optimum. 40
Figure 36 Clusters Reflected on the Map 42

List of Tables

Table 1 Moran’s values for spatial variables and interpretation 30
Table 2 Supervised Machine Learning Setup for the Classification Model 33
Table 3 Classification Experiment Results 34
Table 4 Optimization Results obtained with PyCaret tune_model 34
Table 5 Clustering Experience Details. 39

List of Equations

Equation 1 Matrix representation 14
Equation 2 Individual Notation 14
Equation 3Moran's I 14

List of Acronyms

A

AFAD Disaster and Emergency Management Division of Ministery of Interior.

ANN (Artificial Neural Network

ArcGIS – The Family of Geographical Information Systems by ESRI

ATAG (Active Tectonic Research Group

C

CRISP-DM (CRoss Industry Standard Process for Data Mining)

D

DBSCAN (Density Based Spatial Clustering)

E

EDA (Explanatory Data Analysis)

EERI (Earthquake Engineering Research Institute)

ERDAS- A raster data Type by Leica

ESDA (Explanatory Spatial Data Analysis)

ESRI- (Environmental Systems Research Institute)

G

GeoTIFF- A georeferencing TIFF file

GIS (Geographical Information Systems)

GPU Graphical Processing Unit

I

IDRISI- An Integrated GIS and remote sensing raster data type

ITU(Istanbul Technical University)

K

KDE (Kernel Density Estimation)

L

LightGBM- Light Gradient Boosting Machine Learning Model

Q

QGIS- Open Source Alternative of ArcGIS

S

SHAP- Shapley Additive exPlanations

T

TIFF (Tag Image File Format)

Turkstat (Turkish Statistical Institution)

1. Introduction

Earthquakes, like the massive one that hit Kahramanmaraş, Turkey in 2023, remind us of nature's unpredictable power. These events, despite being purely natural, can reshape landscapes, cities, and lives in moments. With advanced tools like GIS (Geographical Information Systems) and the power of spatial data science, we are improving our ability to predict and prepare for these earth-shaking moments.

Kahramanmaraş was not the only city affected; this earthquake's ripples also reached Syria. According to trusted sources like Wikipedia and ReliefWeb, the event had a heart-wrenching human toll: almost 57,000 lives lost and about 130,000 people left with serious injuries. By March 19th, the aftermath was evident across 17 cities, touching the lives of approximately 19 million individuals.

Beyond the emotional and physical pain, the financial strain was palpable. Damages soared to an astounding $34.2 billion, causing a significant 4% drop in Turkey's overall economy, as highlighted by the World Bank. This kind of economic impact underscores that the repercussions of earthquakes are not limited to just toppled buildings and cracked roads; they have the potential to destabilize national economies.

Rebuilding in such scenarios is a layered process. It is not just about bricks, cement, and steel. It's about reviving neighborhoods, reinstating the rhythms of daily routines, and nurturing the shared bonds that form the essence of communities. Schools, hospitals, parks, and even local coffee shops become focal points of restoration, emphasizing the broader dimensions of recovery. Rebuilding is a testament to human resilience and the interwoven fabric of geography, culture, and commerce. It's a journey from ruins back to routines, and from loss to a renewed sense of belonging.

Given the scale and implications of such natural disasters, it becomes essential to really get a grip on their effects. We need to pinpoint where the most damage occurred and predict potential future hotspots. Is a specific location more susceptible to damage than others? If we could identify certain areas as more vulnerable based on their location—perhaps they are closer to a fault line or in a valley that amplifies seismic waves. It would be like understanding that a house on the coast might get hit harder by a storm than one inland. By knowing this, communities can better prepare, construct sturdier buildings, or even reconsider urban planning strategies. It's not just about reacting to the aftermath, but also about anticipating and preparing for what might come next. This pro-active approach, driven by a mix of geography, science, and local insights, could be a game-changer in safeguarding lives and infrastructures.

1.1. Objectives of the study

This study focuses on using the data collected from Kaggle (Dincer, 2023) that is sourced from EU Copernicus disaster information (European Commission - Copernicus Emergency Management Service, 2023) in a machine learning model to predict patterns of the damaged buildings just after the disaster. Of course, we do not hold more information about the land properties or building materials. However, sometimes, most of the time buildings collapse due to bad business decisions. In this study we aim to find answers to following questions.

- If there is a relationship with the geophysical fault information?
- How are clustered buildings and different level of damages?
- By using a machine learning algorithm can we find what makes an area more vulnerable?
- Which regions are affected worse compared to other regions?

For the objectives have been set, there are serious sub-objectives that could be established to achieve main goals. As all data science projects needs to start with, a proper data collection should be made.

Here the data set is consisting of several dataset and in orders to make a proper analysis we need external information to have.

2. State of Art

Examining earthquake data has created interest of numerous disciplines, spanning from geology, civil engineering, urban planning, to data science and economics. Geologists often are into the root causes, studying tectonic plate movements and subsurface structures to gauge where the next big quake might strike. Civil engineers, on the other hand, scrutinize this data to design resilient infrastructures that can withstand seismic shocks. Urban planners use the information to make informed decisions about land use, ensuring that vital facilities, like hospitals and fire stations, are positioned in areas less susceptible to significant damage. Meanwhile, data scientists harness advanced algorithms to predict patterns and probabilities, contributing to our evolving understanding of these natural phenomena. Economists analyze the economic impact, both immediate and long-term, to help governments plan for future contingencies. Together, these diverse fields collaborate, using earthquake data as a touchstone, to make our world safer and more prepared for seismic challenges.

Spatial methods in data science, on the other hand, are like advanced map tools that we use to solve real-world problems. For instance, in urban planning, these techniques help in determining optimal locations for bus stops, ensuring that residents have easy access to transportation. Environmental scientists use them to monitor deforestation rates, ensuring sustainable management of our forests. In healthcare, spatial analysis can predict potential disease outbreaks, allowing early interventions. Agriculturists can employ these methods to assess soil health and optimize irrigation, ensuring maximum yield. In business, determining the ideal location for a new store or café can be driven by spatial data analysis, weighing factors like population density and nearby competitors. As technology advances, the potential applications of these 'smart mapping' techniques will continue to grow, shaping various sectors in nuanced ways.

2.1. Geospatial Data Basics

John Graunt was the first demographic in the history who tried to use statistics to use mortal rates of plague. In his famous book ‘Natural and Political Observations Made upon the Bills of Mortality’, he has dedicated a chapter on differences between parishes of London.(1662). Other early studies of the geospatial data were used to find most commonly used routes.

Geo-spatial data is the kind of data informs us about the localization of the observation. This information can be in two formats which are raster data and vector data. While raster data is a numeric representation of the surface in equal areas such as pixels, vector data can be a representation of point, line-string or polygons which could be either observed natural shapes like rivers or mountains or administrative units such as cities, villages countries or even just the borders. Either of these types can be useful depending on the research problem of subject.

Raster data could be found in formats such as GeoTIFF, ERDAS and IDRISI. GeoTIFF needs to be composed of .tif, .tiff and .ovr files while Erdas composed on .img file and IDRISI Raster is composed of .rst and rdc files. (Anello, E, 2019).

Vector data is differently found and commonly used in Shape file (.shp) which needs to be supported by .dbf and .shx files. Vector data could be found also in geojson or json document which have all related necessary information (Anello, E, 2019). In this study we are dealing with vector data that shows point, line and polygons and can be kept in tabular format whenever is necessary.

2.2. Studies on Spatial Data

Spatial analysis can be done with several tools including applications for non-programmers and libraries in programming languages.

GIS stands for Geographical Information Systems, which handles storing sharing manipulating and analyzing geographic information. Before the revolution of cloud computing and data storage tools

introduced by Google Earth, it has been also important to store it by using GIS tools (Bivand, Pebesma, and Gómez-Rubio, 2013). GIS can be deployed with specail software programmes such as ArcGIS which is developed by ESRI or QGIS an open source alternative that offer easy interface to the users for several basic tasks. For perhaps many uses programming might be harder than using a GIS software interface.

On the other hand, programming offers more control to the analyst over algorithms and parameters. R is practical specifically for the analysts coming from a statistical background for its ease in use (Bivand, Pebesma, and Gómez-Rubio, 2013)

Python offers a object oriented programming approach to spatial analysis without abandoning functional programming. Libraries of geospatial analysis in Python are usually offer both to the analysts. In addition, although objects are abstract concepts in python, here in geo spatial analysis it is easier to relation with real time objects such as rivers, trees, or buildings.

2.3. Studies of Machine Learning on Earthquake

Earthquake hazard detection was reviewed in a paper by the Earthquake Engineering Research Institute (EERI) in 2020. They suggested hazard detection using sensors to detect seismic activity and attempted to identify damage resulting from such activity (Xie et al., 2020). The authors consulted various machine learning models and utilized artificial neural network (ANN) modeling to detect damage. Furthermore, they proposed that Machine Learning could provide insights into earthquake damage comparable to those offered by traditional physics-based approaches.

A temporal-spatial analysis made by Bagus Priambodo, Wayan Firdaus Mahmudy, and Muh Arif Rahman that analyzes seismic activity data of Indonesia (2020). They have worked on a dataset that concsist the variables of latitude, longitude, timestamp, and magnitude. They have divided Indonesia to 16 different grids and applied an ANN model that consist of 16 neurons as out layers and only one hidden layer. Although they have achieved satisfactory results, they admit that more geo specific information about the grids is necessary to predict better an upcoming earthquake.

A type of data we would have liked to have been analyzed by Garg, Masih, and Sharma (2021). They were able to work with a dataset of damage level of bridges which have material, age, structure and distance from earthquake. Although they failed to reflect which variables have affected the more in their paper, they have good scores on classification models. In addition, this study is important as bridges are important connection points in case of emergency. We are rather focused on a continental dataset that doesn't have a lot of bridges, in our data set we have included roads that would have a similar type of importance. Another idea can be taken from here is that distance to the earthquake center is an important variable to be considered and it is worth to include in a machine learning model.

Mangalathu et al. (2020) utilized machine learning to classify earthquake damage to buildings in South Napa 2014 earthquake. According to their random forest model the most important variables are found as age, spatial acceleration variable fault distance. Their model reached an accuracy of 65%.

Sheibani and Ou (2021) introduced Gaussian process regression for efficient regional post-earthquake building damage inference on a dataset that have both variables about building and earthquake. Among several variables year of built, floor area, first vibration and location for the buildings and inter quartile range. Kurtosis, spatial acceleration arias and fajfar has been found as the most important variables that was found in the best model.

Ahmad and colleagues (2020) utilized earthquake insurance data, applying actuarial techniques to gain insights into their portfolio. They employed machine learning and Bayesian estimates to determine the risk premium associated with the portfolio. While their research is not directly related to this study, findings of it may be relevant for the insurance sector, aiming to mitigate risks for both clients and insurance providers and offer more pertinent services.

3. Methodology

In this project, we combined standard Data Science techniques with geospatial tools to get the job done. While a lot of our data was analyzed using common methods, we recognized that some information needed a more specialized touch due to its location-based nature. So, we ensured we used the right tool for each specific task, striking a balance between formal methodology and practical application.

For the required visualizations, analyses, and modeling in this study, Python and its essential libraries were utilized. Python, a versatile programming language, offers a vast ecosystem of specialized libraries for geospatial analysis, making it a premier choice for researchers and analysts. The chosen libraries provided functionalities ranging from data preprocessing to advanced spatial statistical computations. Additionally, Python's flexibility and integration capabilities facilitated the seamless combination of data from diverse sources and formats, ensuring a comprehensive and rigorous examination of the study's objectives.

Project in overall follows general methodology of CRISP-DM (CRoss Industry Standard Process for Data Mining) that was proposed before by Wirth and Hipp (2000). This methodology follows below stages iteratively and wherever necessary returning to the previous stages upon needs.

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

In this project, the stages of business comprehension and data preparation were of paramount importance. Given the gravity of the situation, our available data was limited. It was essential to have a clear understanding of our objectives and ensure that the data we had was meticulously organized and used to its fullest potential. Proper data handling was crucial to achieve accurate insights. Perhaps future studies will have better understanding when more information is available.

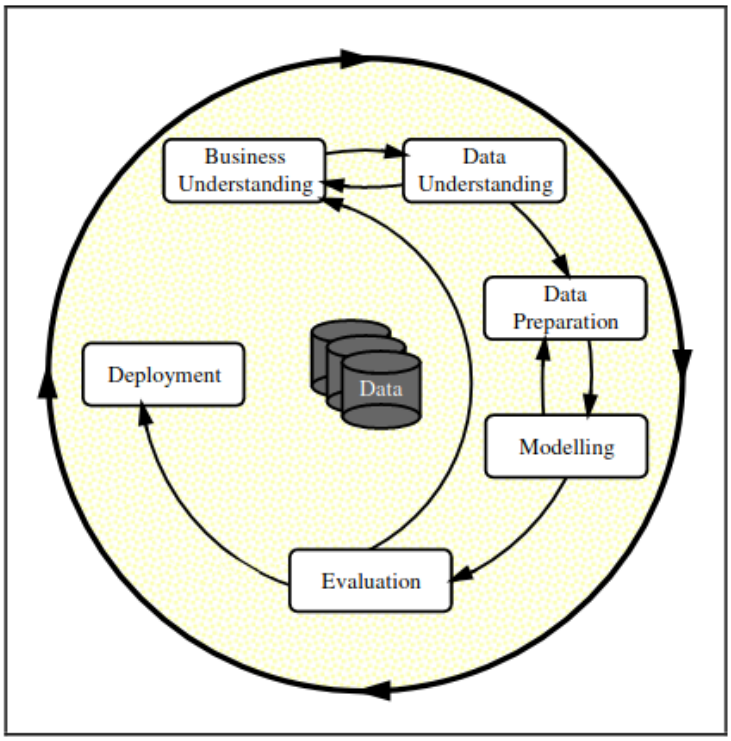

Figure 1: Diagram of CRISP-DM Methodology

Source: Wirth and Hipp (2000)

3.1. Data Collection

Data collection was crucial part for this project. The main objective was to predict damage level for the given values in a data set. Unfortunately, the data we have do not offer us building properties and yet do not offer some of the details we would like to have specifically although the damage situation is directly related to these variables as well. We are hoping that in future more data collection could be available for city planing and emergency planing purposes to avoid more problems.

Our data is a merge of several data sets that were carefully joined by spatial join (sjoin) method offered by Geopandas library. As demonstrated in the "geopandas.sjoin" function documentation (GeoPandas, 2023), this helps us join data frames in a spatial approach since sometimes data formats are different and not directly addressing a common variable for an ordinary join. We are able to do a geographic approximation to find points inside, in the border or 'within' the area. The datasets that we have used to aggregate main data frame can be listed as follows.

3.1.1. EU Copernicus Disaster Data

This is the main dataset we have damage information about the buildings, facilities and roads. This data set was retrieved from Kaggle (Dincer, 2023). Data consist of several shp and json files organized according to cities in different folders. Later in the same folders apart from mentioned information, in the city files there are also information about elevetion levels, water souces, campsite areas.

The original dataset has 20 different locals. However, since the local 'Cumhuriyet' didn't have other information, this area was not included in the study. This data for almost every local has below datasets that includes vector:

- Area of Interest (Polygon)
- Buildings (Point or Polygon)
- Facilities (Point or Polygon)
- Transportation (Line or Multiline string)
- Hydrography (Water Sources, Lines or Multiline strings)
- Physiography (Elevation contours Line or Multiline strings)
- Campsites (Point or Polygon)

In this study since it can directly affect the geographic properties, we have used all the information somehow. Buildings, Facilities and Transportation files have our target data “damage_gra” they make our main dataset. For each unit it was curious to know properties of the location. There for we have included hydrography and Campsites (In case of a disaster/evacuation are there options?)

Physiography data is also included to calculate an approximation of height from the sea level. In order to include all these information indirectly add to the dataset, we have used Euclidean distance to find closest point to the reference point. Details of the gathering will be explained more in the Geospatial methods.

3.1.2. AFAD Earthquake Data information

AFAD is governmental institution in Turkey that manages Disaster information mostly Earthquake related information. In the databank, it is possible to retrieve time series data of all earthquakes including their location, and magnitude (Event Catalog, 2023)

3.1.3. TURKSTAT

Turkstat,(TUIK in Turkish) is the statistical institution that collect and analyze several socioeconomic data. In this study in order to find clusters in the end it was interesting to include also socioeconomic data like income per capita, hh size unemployment etc. In addition, since buildings are closely related to the construction sector and to have some relevant information, we have added house selling data as well.

Required data was retrieved from regional data bank according to provinces and municipal regions manually after merging the main data.

3.1.4. Fault Data

Finally, just as the we and Campsites, we are interested also in the distance to fault as they are the main point that tectonic energy release. Therefore, buildings that are close to these areas likely to get damage more.

Data used for fault information was derived from data base of ATAG (Active Tectonic Research Group) ITU (Istanbul Technical University) database. Data consist of line string information belong to fault lines.

3.2. Spatial Analysis Methods

In the section dedicated to data, it's highlighted that we utilized multiple sources to gather information regarding the buildings' locations. For this research, we employed two primary analytical methods: Point Pattern Analysis and Spatial Autocorrelation Analysis. These analytical tools allow us to understand the spatial distribution and relationships of the buildings more thoroughly. Furthermore, the insights obtained from these analyses were instrumental in guiding and refining our subsequent modeling efforts, ensuring a comprehensive approach in our study.

3.2.1. Point Pattern Analysis

The main purpose of point pattern analysis is to understand if certain locations are clustered or randomly distributed. Apart from earthquake analysis, this kind of analysis can be conducted to problems as understanding clientele, regions with more crimes, disease spread etc… This analysis can be conducted to the data frames that has latitude and longitude information (Alam, 2020).

The analysis presented is closely tied to this study, as it seeks to understand the randomness of damage levels. Understanding this randomness is essential, as it can guide decision-making, particularly when assessing risk factors within the business domain. If the damage appears random, businesses can strategize based on these risk factors. For instance, they might contemplate insuring properties, utilizing more resilient building materials for new projects, or even strengthening current structures using advanced reinforcement methods. Conversely, if specific regions consistently show higher vulnerability due to inherent natural factors, it might be prudent for future developments to sidestep these zones to minimize potential damage in subsequent events. This not only ensures the safety of the structures but also offers a strategic approach to sustainable urban development.

In our analysis, visualization emerges as an indispensable tool, offering an intuitive grasp of data patterns. Even basic tools like scatter plots can yield significant information, revealing patterns like clustering or potential randomness in the data. Expanding on this, techniques like hexbin plots and Kernel Density Maps serve to highlight variations in data density, providing comparative magnitudes. To delve deeper into the numerical aspects, we've explored metrics such as Centrography, Tendency, Dispersion, and Distances. Within this study, particular emphasis was placed on Ripley’s G and Ripley’s F statistics, which served as our primary test statistics.

In their detailed book about using Python for geographic data, Rey, Arribas-Bel, and Wolf (2020) touch upon many key topics. One interesting part is about Ripley's G function, which basically tells you how far you'd typically have to go from any random spot to get to the nearest marked point. Then there's Ripley's F function, which is like asking, if you're standing on a marked point, how far is it to the next closest one? Both these concepts help us understand patterns and distances in space.

For the purpose of rendering these visualizations, we employed the capabilities of libraries like contextily, matplotlib, and Geopandas. Meanwhile, for computational heavy lifting and metric calculations, we leaned on the resources offered by libpysal and pointpats libraries.

3.2.2 Ball Tree Algorithm and Its Use in Data Preparation Process

In order we are able to calculate distances and, in the elevation, variable case the closest point’s height from the seaside, it was necessary to find the closest point in between different geometries. For this reason in the date preparation process, we have created distance variables and in order to calculate this we have used Ball Tree method hoping that since the earth is a sphere, since we use latitude longitude variable to use these and the locations are on earth the best approach would be applying ball trees while calculating Euclidean distances. Moreover, considering the faults and water sources can be deep, a 3D approach to calculate distances make a lot of sense.

b

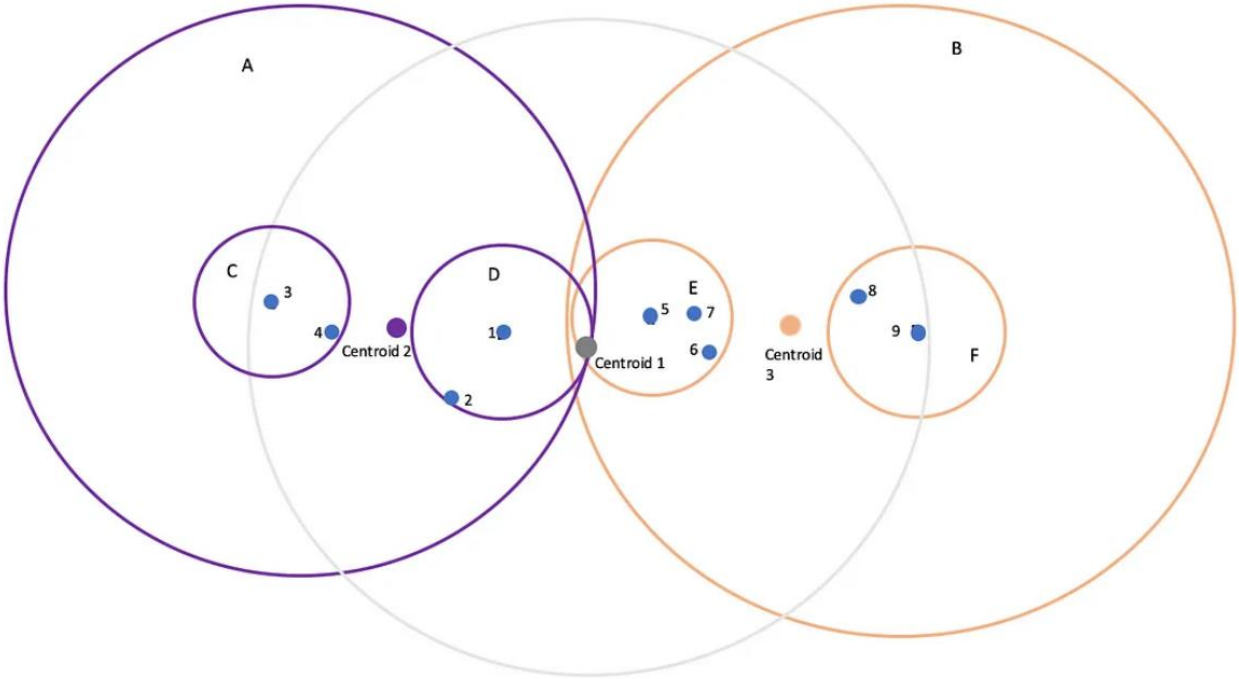

Figure 2: Schema of Ball Tree approach.

Source: Hucker, M (2020)

The Ball Tree Algorithm starts y identifying the central point of all data points. From there, the point farthest from this central point becomes the core of the first cluster. Next, the point most distant from this first cluster's core establishes the center of the second cluster. All remaining points are sorted based on their proximity to these two central points, joining either the first or the second cluster. Even though the boundaries of these clusters might overlap, every point distinctly belongs to just one cluster. In the rare event a point is equidistant from both centers, it still gets allocated to a single cluster, ensuring clusters can be uneven. Essentially, the algorithm divides data into increasingly smaller clusters by repeating this process within each cluster until reaching a predetermined depth, leading to a hierarchy of nested clusters (Hucker, 2020)

For practical purposes, Sci-kit learn library help us to create trees easily. In this study we use this method in our calculations.

3.2.3. Spatial Lags, Weights and Autocorrelation

Indicators of global spatial autocorrelation have two main goals. Firstly, they use the might of statistics to give a concise summary of the spatial patterns we see on a map. Secondly, they provide a formal measure of how much these patterns stray from pure randomness. Think of these indicators as tools that help us get the essence of a map's clustering features, be it through a visual representation or just a number. But to truly grasp these statistics, there's a foundational concept to know: the spatial lag. Once we've got a handle on that, it sets the stage for a deeper dive into global spatial autocorrelation. We'll start by looking at simple scenarios with just two possible values on the map and then delve into more detailed scenarios, exploring tools like the Moran Plot and Moran's I.

$$Y_{sl} = \mathbf{W}Y$$

Equation 1 Matrix representation

Source: Rey, S. J., Arribas-Bel, D., & Wolf, L. J. (2020)

Where:

- Ys is the spatial lag of the variable.
- W is the spatial weights matrix.
- Y is the vector of the given variable.

For individual notation, for each observation i:

$$y_{sl-i} = \sum_j w_{ij} y_j$$

Equation 2 Individual Notation
Source: Rey, S. J., Arribas-Bel, D., & Wolf, L. J. (2020)

Here:

- ysi represents the spatial lag for the observation i.
- wij is the weight from the spatial weight matrix corresponding to the relationship between observation i and observation j.
- yj is the value of the variable for observation j.

After calculation of weights/weight matrix we are now able to calculate Moran's I statistic that can be represented as follows:

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \times \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

Equation 3 Moran's I

Source: Rey, S. J., Arribas-Bel, D., & Wolf, L. J(2020)

Where:

- n is the number of spatial units (observations).
- wij is the spatial weight between observation i and observation j.
- xi is the value of the variable of interest for observation i.
- $\bar{x}$ is the mean of the variable of interest across all observations.

A positive Moran's I indicates that neighboring observations tend to have similar values (positive spatial autocorrelation), a negative Moran's I indicates that neighboring observations tend to have dissimilar values (negative spatial autocorrelation), and a Moran's I near zero suggests a spatially random pattern (Rey, Arribas-Bel, and Wolf , 2020)

It's worth noting that the interpretation of Moran's I also depends on the spatial weights matrix W, and the significance of the value should be tested against a null hypothesis of spatial randomness using appropriate methods (like Monte Carlo simulations).

In our research, we primarily focused on specific local regions, as data for these areas were accessible to us. Each of these local areas received a rating based on the extent of damage they sustained and other relevant factors. For the computational aspect of our investigation, we employed the 'libpysal' and 'esda' software libraries to derive crucial statistical measures. Beyond the primary statistical evaluations conducted, we also determined the spatial lags. This additional data will be incorporated into future machine learning algorithms that don't inherently account for spatial information.

3.3. Machine Learning Algorithms and Tools

In this study, our primary objective is to determine the damage levels of buildings. Given that we possess a dataset with a target variable, we can apply a supervised machine learning model. However, since we're working with geographic data and anticipate clustering, we are also keen on using an unsupervised clustering model. In the spatial analysis section, we managed to cluster regionally. However, since the dataset also contains other variables, we aim to identify clusters in their entirety. Thus, we employed non-spatial clustering using the spatial features we had previously developed.

Undoubtedly, the challenge in this research was the aspiration to utilize a more automated machine learning tool, especially in this era where AI tools are becoming increasingly prevalent. For this reason, we opted for PyCaret as our software tool. According to the official PyCaret documentation (PyCaret, 2022), PyCaret is a low-code Auto ML tool. With it, we can handle data cleaning, standardization, feature engineering, and feature selection. Furthermore, we can test a plethora of models, even those typically overlooked. Apart from data processing and modeling, PyCaret offers functionalities for optimization and interpretation, which is invaluable given the complexity of these tasks. Fortunately, PyCaret automates these processes.

However, conducting geospatial analysis with this library comes with its own challenges. Firstly, geometric data (Point, Line, Polygon) isn't supported, necessitating the removal of these variables before modeling. Secondly, there are potential compatibility issues between spatial analysis and visualization tool dependencies in contrast with PyCaret's dependencies. Hence, for this study, we strongly recommend using Conda as an environment and dependency manager, and maintaining two separate environments. Further details can be found in the GitHub repository of this study (Yazganoglu, G, 2023). It's worth noting that not all of PyCaret's functions are available upon initial installation. One needs to either install the entire library or make specific installations to access certain functions. Additionally, some algorithms, like LightGBM and feature selection methods using LightGBM, may necessitate the standardization of string values and might leverage the local machine's GPU.

3.3.1. Supervised Modeling

As mentioned above, we have used PyCaret for modeling and model selection process. The target variable “damage_gra” has 5 classes one of which is no-info therefore 4 classes which are destroyed, damaged, possibly damaged, no visual damage. Therefore, the model needs to be a multi-class classification model that guess the corresponding value of damage_gra on other given variables.

Among all the variables that can be setup in a ClassificationExperiment, the best value later we will see is Random Forest Classification Model.

3.3.1.1.Random Forest Classification

The Random Forest algorithm is an ensemble machine learning technique that integrates multiple decision trees to produce a more robust and generalizable model. Drawing inspiration from the bootstrap aggregating (or "bagging") methodology, the algorithm constructs each decision tree from a random subset of the data and a random subset of features. This ensures diversity among the trees, thereby reducing the variance associated with individual trees and mitigating overfitting. (Geron, A 2022) When making predictions, in the case of classification, each tree in the ensemble casts a vote, and the majority vote is taken as the final prediction; for regression tasks, the average prediction across all trees is considered. This consensus-driven approach typically enhances prediction accuracy and model stability. Despite its advantages, one of the challenges associated with the Random Forest is interpretability, given the complexity arising from the combination of multiple decision pathways. Nonetheless, its ability to handle large datasets, importance ranking of features, and resistance to overfitting make it a widely employed algorithm in various academic and practical applications.

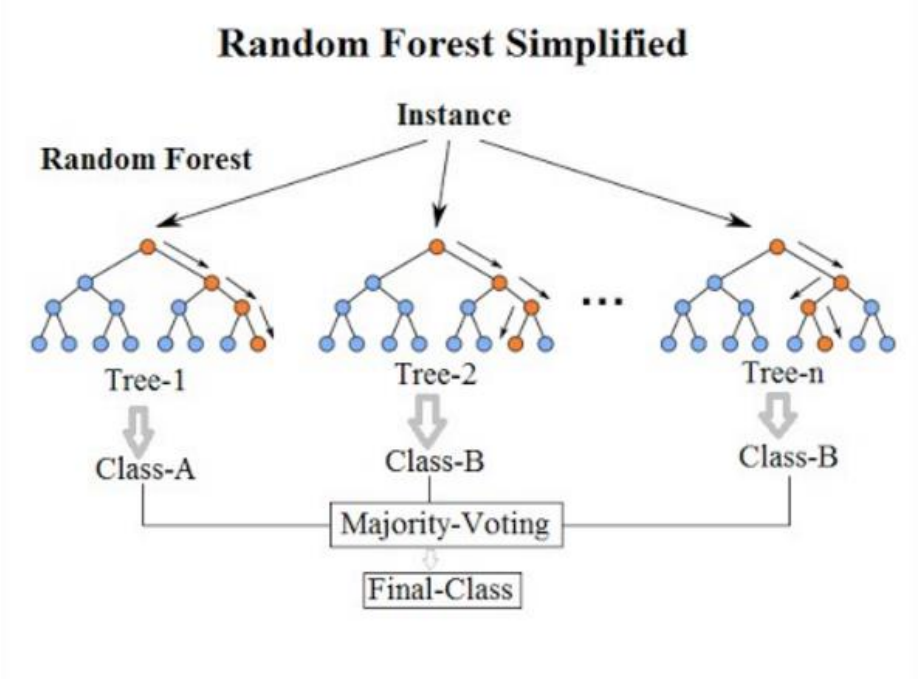

Figure 3 Random Forrest

Source: (Wikipedia, 2023)

3.3.2. Unsupervised Modeling

Incorporating unsupervised learning models is crucial when making informed business decisions regarding buildings that have suffered damage or complete destruction. Clustering, an unsupervised technique, offers an insightful way to categorize and understand building types based on their characteristics. Imagine a city planner who, after a natural disaster, wants to prioritize reconstruction

efforts. By using clustering, they could group buildings not just by the extent of their damage, but also by other features like their age, construction materials, or importance to infrastructure. For instance, two buildings might be equally damaged, but one might be a critical hospital while the other is a vacant warehouse. Clustering can help identify such nuances, enabling decision-makers to allocate resources more strategically and rebuild more efficiently. This methodological approach, which hinges on discerning patterns and similarities within the data, can serve as a foundational tool for urban development and disaster response strategies.

In order to cluster PyCaret can be used as well. However, this module is rather under development that we are not able to create a "Clustering Object" and we are not able to run several methods and choose one of them. Therefore, in modeling K-Means and DBSCAN models are deployed. K Means were the winner after all.

3.3.2.1. K-Means Clustering

K-Means is one of the most used unsupervised machine learning algorithms for partitioning a given data set into a set of k groups, where k represents the number of clusters pre-specified by the analyst. It classifies objects in multiple clusters based on the features of these objects, such that objects in the same cluster are more similar to each other than those in different clusters. The main objective is to segregate groups with similar traits and assign them into clusters.

In the K-means method within data mining, the initial step involves choosing a set of centroids at random, representing the starting point for each cluster. The algorithm then carries out iterative computations to refine and improve the positions of these centroids (LEDU, 2018). Later this process

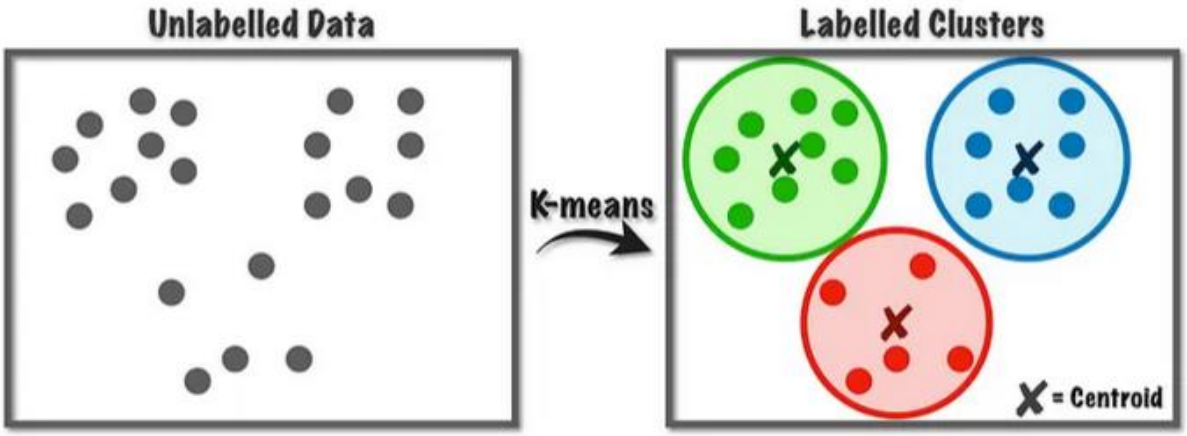

Figure 4 K Means Clustering

Source: Jeffares, A (2019)

needs to be repeated to find optimum number of clusters to make the model stable. In application using PyCaret, finding the optimum number of clusters process is also autotomized.

3.3.3. Interpretation

For Random Forrest model interpretation is a very important problem as the model itself doesn't explain how the variables affect the target value. For this reason in this study, we have consulted to SHAP values in order to get better insights.

SHAP values provide insights into model decisions, helping in both global and local interpretability. For example, in a banking scenario, they can reveal general trends like which factors lead to loan approvals or rejections. This information helps the bank provide guidelines to future applicants to increase their chances of loan approval. On a local level, if a specific loan application is rejected, SHAP values can pinpoint the exact reasons for this decision. This is vital, especially when a customer disputes a rejection. By analyzing individual cases, one can show customers the specific parts of their application that led to the model's decision (Bex, T 2022)

Implementation was through PyCaret library and this was also easy to demonstrate with a few line of codes. The tricky part here is that this modul is not directly included in the PyCaret package. It should be installed again with analysis module or directly by installing shap library to our environment.

4. Spatial Analysis and Machine Learning Models of Buildings and Facilities

In this section of this study results of the work is reflected according to the methodology that was explained in the previous section.

Data collection process was clearly explained previously, here we will follow with EDA, ESDA and modeling part and comments about the results.

4.1. EDA and ESDA

EDA stands for Explanatory Data Analysis which is necessary to decide how to clean the data to be used in machine learning algorithms. As the data set has several socioeconomic variables which worth to be checked before modeling as much as other data science projects. Therefore, in this project it is a must as well. In this study we first have a glance to the data set to understand which variables are important, if there are null values or any other values should be treated. We will see that some of the variables of the dataset does not give us any information at all and needs to be removed.

As mentioned below our data set is a spatial dataset and requires another treatment. ESDA stands for Explanatory Spatial Data Analysis which is the process of understanding patterns and relationships in spatial (geographical) data through various techniques and visual representations, like maps, plots, and statistics. ESDA is vital for several reasons. One of the primary goals of ESDA is to identify and visualize spatial patterns. For example, we might want to understand if a particular disease is localized to a certain region, if there's a hotspot of criminal activity, or if a species of animal is found more densely in specific habitats or is there a place the earthquake is likely to create damage or destruction.

In the first part of our ESDA we try to understand if there exist patterns for the locations and damage level. We are interested in damaged and destroyed values as these buildings for sure needs to be replaced and probably has caused losses of several kinds. We should understand if damage is random or happens in the specific places. Centrography, Tendency and Randomness are key things aspects in Point Pattern analysis.

In the second part of the analysis, we try to understand the homogeneity in a different way. As a tool we use spatial autocorrelation and Moran’s I statistic. In addition, we will create spatial lags to later to understand the spatial implications in the multiclass classification model.

4.1.1 EDA (Explanatory Data Analysis)

We have applied profiling that was offered by pandas_profiling library to understand very basic knowledge about the dataset. The dataset had initially 38 variables.

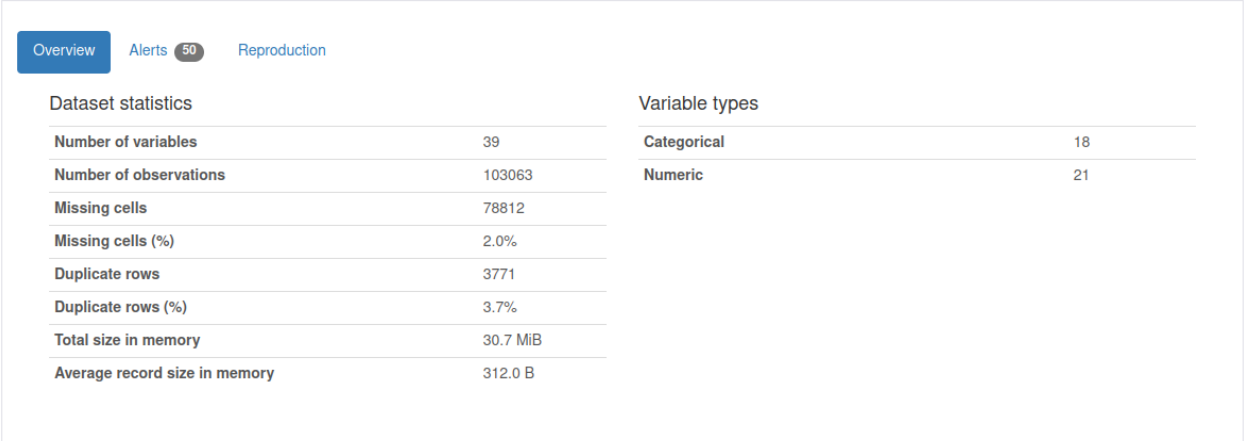

Figure 5 Overview of the Dataset

Source: Yazganoglu, G (2023)

Although there seems like a problem of missing values in a closer view we observe that there are several variables that doesn't really give us a lot of information. These variables that has values such as 'None' or 'Not Applicable' are as useless as null values. In addition, index values or defining variables such as 'name' cannot be used in the modeling either. Therefore, we can omit these variables. Deleted variables are name, det_method, notation, cd_value, real, index_right, esmr_id, glide_no, map_type. After removing these variables and duplicates we do not anymore have null values. Duplicate values probably occur due to data merging or some places maybe counted both as facility and building.

When further analyzed we are able to observe which group of buildings are affected the most bar plots show that Kirikhan and Kahramanmaras are the locations with the most destroyed and damaged buildings and most of the building that destroyed or damaged either residential buildings or roads. All kind of damages are important and may cause the loss of life or money meantime damages in road cause also barrier to access after disaster which means problem in delivering help.

(damage_gra 4: destroyed, damage_gra = 3: damaged, damage_gra =2, possibly damaged, damage_gra = 1: no visual damage, damage_gra = 0: no damage information)

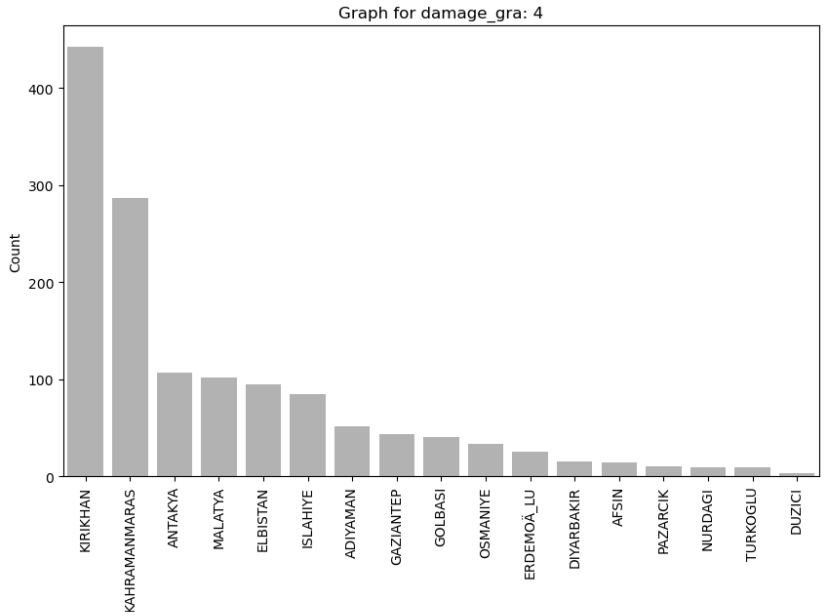

Figure 6 Destroyed Buildings in Cities

Source: Yazganoglu, G (2023)

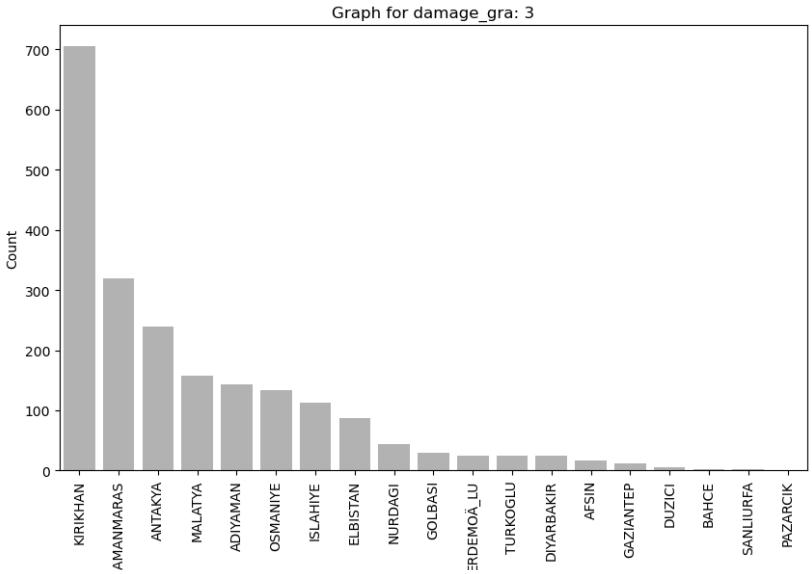

Figure 7 Damaged Buildings in Cities

Source: Yazganoglu, G (2023)

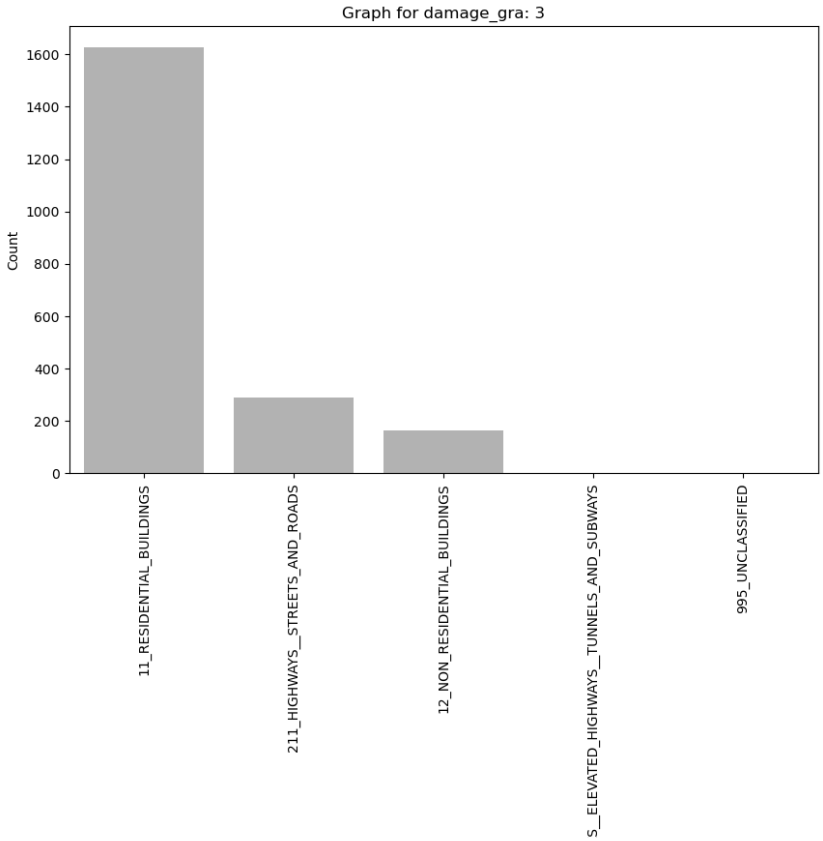

Figure 8 Damaged buildings according to object type

Source: Yazganoglu G (2023)

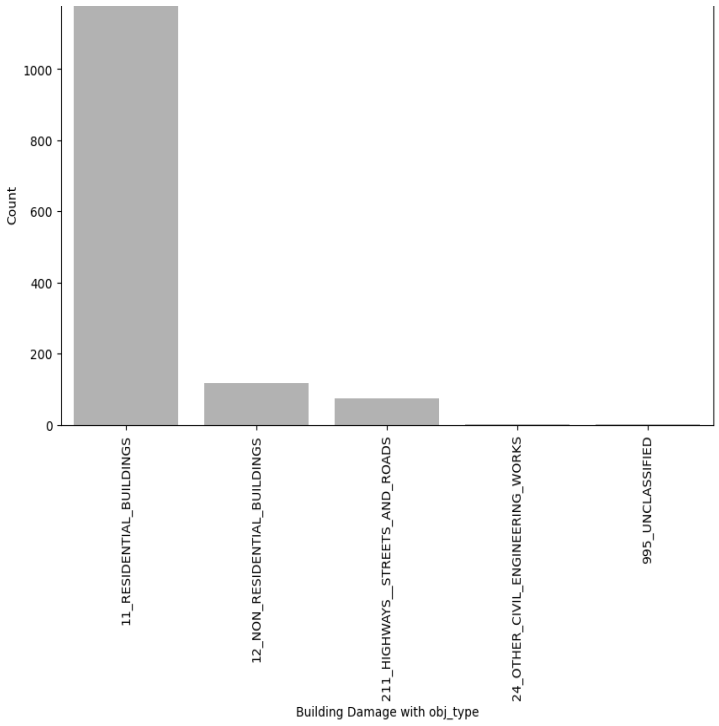

Figure 9 Destroyed Buildings according to obj_type

Upon closely examining the boxplots of numerical features segmented by various classes, we discerned distinct distribution characteristics for each class. These boxplots visually represent the

distribution's central tendency, variability, and skewness, highlighting any potential outliers. The discrepancies between the distribution metrics for different classes could provide insightful information, emphasizing the unique statistical properties and potential patterns within each class.

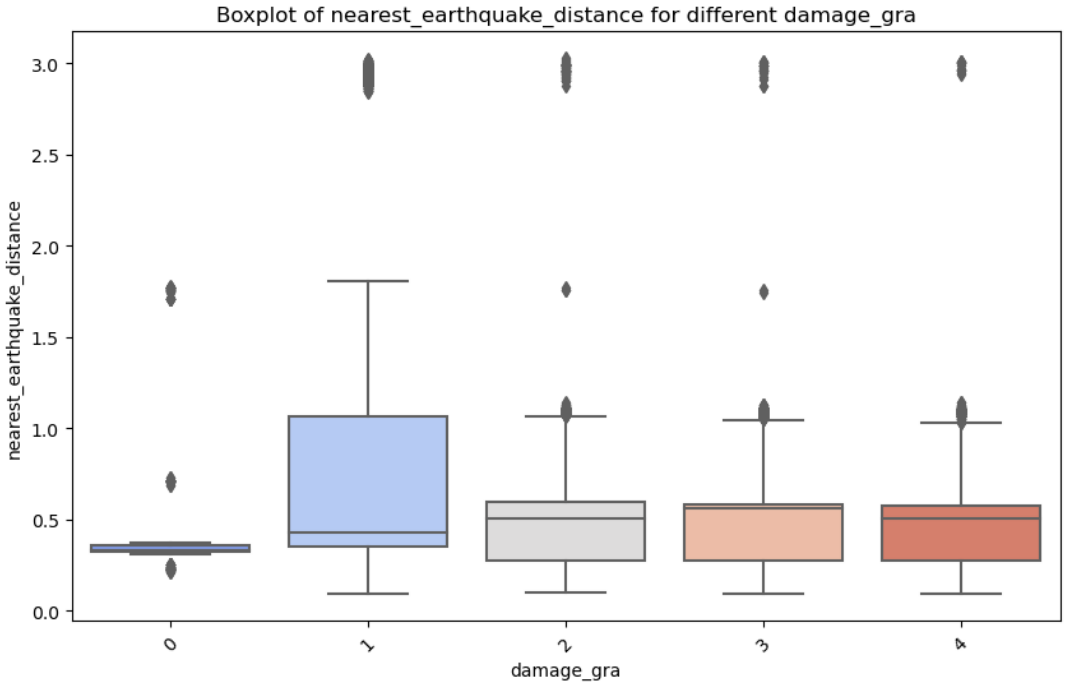

Figure 10 Boxplot of nearest_earthquake_distance

Source: Yazganoglu, G 2023

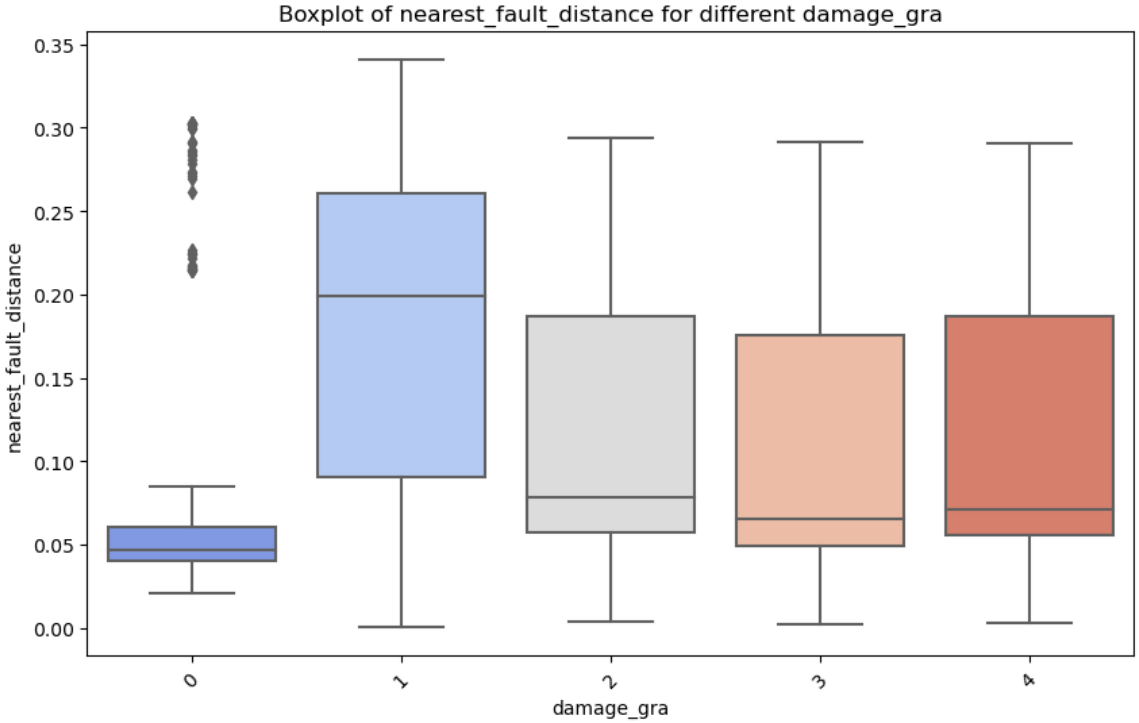

Figure 11 Boxplot Nearest Fault distance

Source: Yazganoglu, G, (2023)

	Column	Dtype	Missing	Missing Percentage	Distinct Count	Min	Max	Average	Std Dev
0	obj_type	object	0	0.000000	10	NaN	NaN	NaN	NaN
1	name	object	10	0.009703	5082	NaN	NaN	NaN	NaN
2	info	object	0	0.000000	52	NaN	NaN	NaN	NaN
3	damage_gra	object	0	0.000000	5	NaN	NaN	NaN	NaN
4	det_method	object	0	0.000000	2	NaN	NaN	NaN	NaN
5	notation	object	0	0.000000	3	NaN	NaN	NaN	NaN
6	or_src_id	int64	0	0.000000	6	1.000000e+00	9.970000e+02	6.042655e+02	4.848318e+02
7	dmg_src_id	int64	0	0.000000	4	2.000000e+00	9.970000e+02	1.317369e+01	1.019061e+02
8	cd_value	object	0	0.000000	1	NaN	NaN	NaN	NaN
9	real	object	78802	76.460029	2	NaN	NaN	NaN	NaN
10	index_right	int64	0	0.000000	1	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
11	emsr_id	object	0	0.000000	1	NaN	NaN	NaN	NaN
12	glide_no	object	0	0.000000	2	NaN	NaN	NaN	NaN
13	area_id	object	0	0.000000	19	NaN	NaN	NaN	NaN
14	locality	object	0	0.000000	19	NaN	NaN	NaN	NaN
15	map_type	object	0	0.000000	2	NaN	NaN	NaN	NaN
16	population	int64	0	0.000000	17	2.290400e+04	2.170110e+06	1.030682e+06	8.724357e+05
17	income	int64	0	0.000000	7	3.012000e+03	7.819000e+03	6.185708e+03	1.627163e+03
18	total_sales	int64	0	0.000000	19	1.588000e+03	2.481210e+05	1.083212e+05	9.221601e+04
19	second_sales	int64	0	0.000000	19	5.360000e+02	1.414340e+05	6.071037e+04	5.280406e+04
20	water_access	float64	0	0.000000	4	9.500000e-01	1.000000e+00	9.746440e-01	2.280118e-02
21	elec_cons	int64	0	0.000000	7	1.631000e+03	7.413000e+03	3.972904e+03	1.827892e+03
22	building_perm	int64	0	0.000000	8	5.830000e+02	2.959000e+03	1.660737e+03	8.814879e+02
23	land_permited	int64	0	0.000000	8	6.957180e+05	3.019546e+06	1.866989e+06	8.794762e+05
24	labour_fource	float64	0	0.000000	4	4.060000e+01	5.000000e+01	4.788620e+01	2.885233e+00
25	unemployment	float64	0	0.000000	4	1.010000e+01	1.710000e+01	1.368046e+01	3.302016e+00
26	agricultural	int64	0	0.000000	8	1.233061e+06	3.535085e+08	4.008357e+07	1.066765e+08
27	life_time	float64	0	0.000000	8	7.690000e+01	7.970000e+01	7.811588e+01	9.676588e-01
28	hb_per100000	int64	0	0.000000	8	1.930000e+02	3.690000e+02	2.669623e+02	4.670803e+01
29	fertility	float64	0	0.000000	8	1.630000e+00	3.810000e+00	2.376413e+00	5.862124e-01
30	hh_size	float64	0	0.000000	9	3.400000e+00	5.120000e+00	3.902629e+00	4.935363e-01
31	longitude	float64	0	0.000000	98600	3.611427e+01	4.025491e+01	3.735722e+01	8.896378e-01
32	latitude	float64	0	0.000000	98567	3.612831e+01	3.839672e+01	3.729042e+01	5.234444e-01
33	nearest_water_source_distance	float64	0	0.000000	98670	2.067003e-06	3.714001e-01	2.354838e-02	3.965855e-02
34	nearest_camping_distance	float64	0	0.000000	98670	5.749022e-07	8.413472e-01	1.000509e-01	2.467072e-01
35	nearest_earthquake_distance	float64	0	0.000000	98670	8.891392e-02	3.025931e+00	6.889907e-01	5.302256e-01
36	nearest_fault_distance	float64	0	0.000000	98670	1.675678e-04	3.405915e-01	1.703268e-01	9.564513e-02
37	elev	float64	0	0.000000	123	7.000000e+01	1.270000e+03	6.264720e+02	2.805116e+02
38	geometry	geometry	0	0.000000	98671	NaN	NaN	NaN	NaN

Figure 12 Descriptive Statistics and new variables created using Ball Tree method.

Source: Yazganoglu, G (2023)

As per correlation heatmap, existing variables show low level of correlation with other variables. As expected spatial variables show negative correlation which means high level of damages occur when distance to fault and earthquake is smaller. We can also observe it from histograms different damage values have different distributions. This has been evident but similar correlations is also observed for the elevation variable meaning higher damages are observed with lower altitude.

Some variables also show correlations with each other such as total house sales, secondhand house sales or household size and fertility. In the multiclass classification model we will let model to choose which one of the variables serve more.

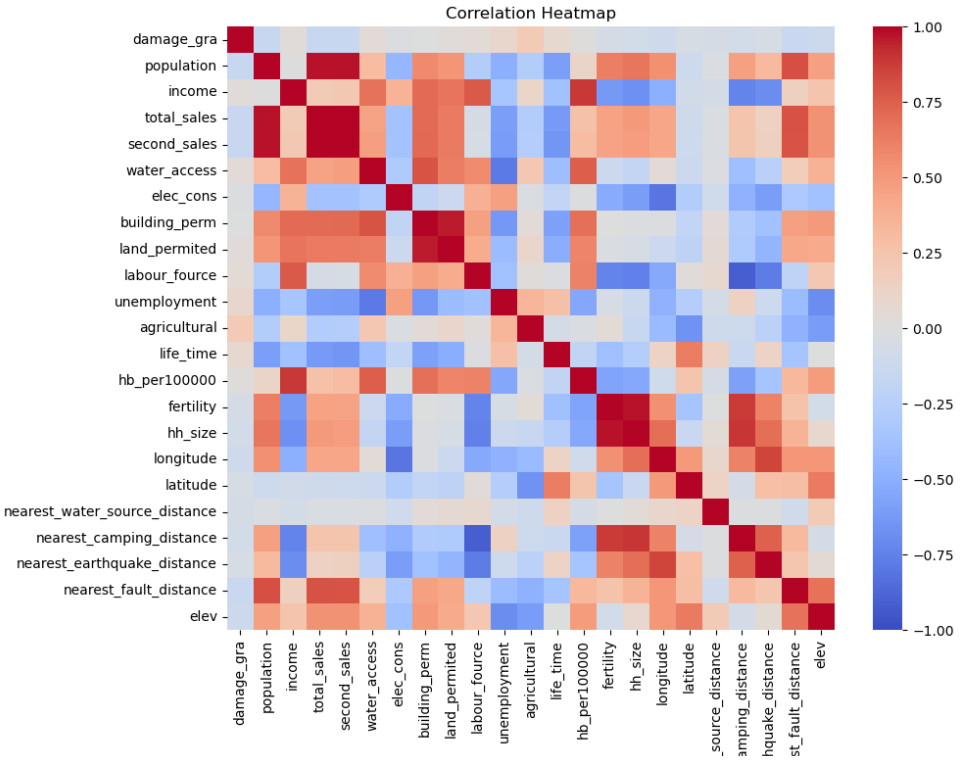

Figure 13 Correlation Matrix of the dataset.

Source: Yazganoglu, G (2023)

4.1.2. ESDA (Explanatory Spatial Data Analysis)

As mentioned in methodology part, a point pattern analysis has been conducted to this study and more spatial variables were created to see. As can be seen in final correlation matrix, other variables are able to explain a very little part of the changes in damage_gra class. Perhaps this is about to change with the implementation of spatial variables.

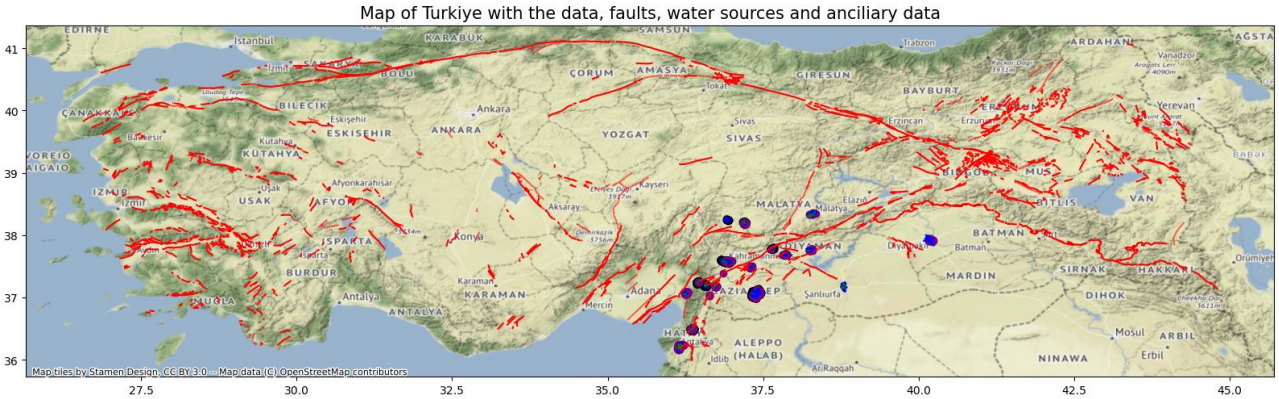

Figure 14 Map Representation

Source: Yazganoglu, G (2023)

4.1.2.1. Point Patern Analysis

After the creation of the main dataset in the beginning, faults, buildings, we can locate the event zone in the Turkiye's map as in the Map1. As can be recognized from the map, the disaster zone is not the only place with dangerous faults.

As the

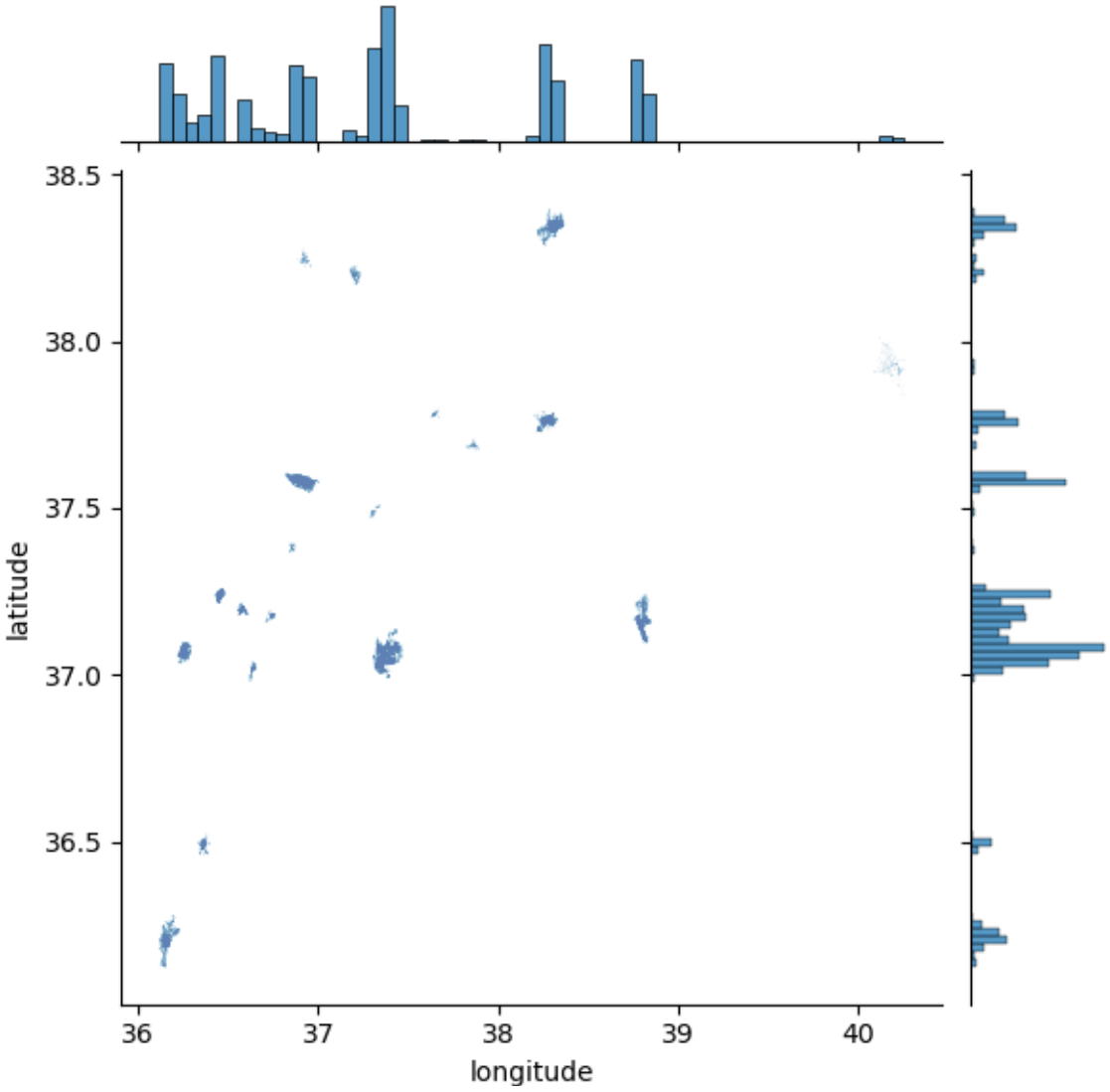

Figure 15 Scatter Point of All Observations.

Source: Yazganoglu, G (2023)

dataset consist of data from several cities, it is clearly observed that buildings and facilities tend to group in several regions. However, when we compare with the same analysis with the observations of only damaged and destroyed buildings, we observed they are centered differently than whole observations.

Kernel Density Estimation KDE is another way we can see the distribution of the all building dataset and damaged destroyed observations. As can be seen in the figure 17 and 18 damaged and destroyed buildings are differently centered.

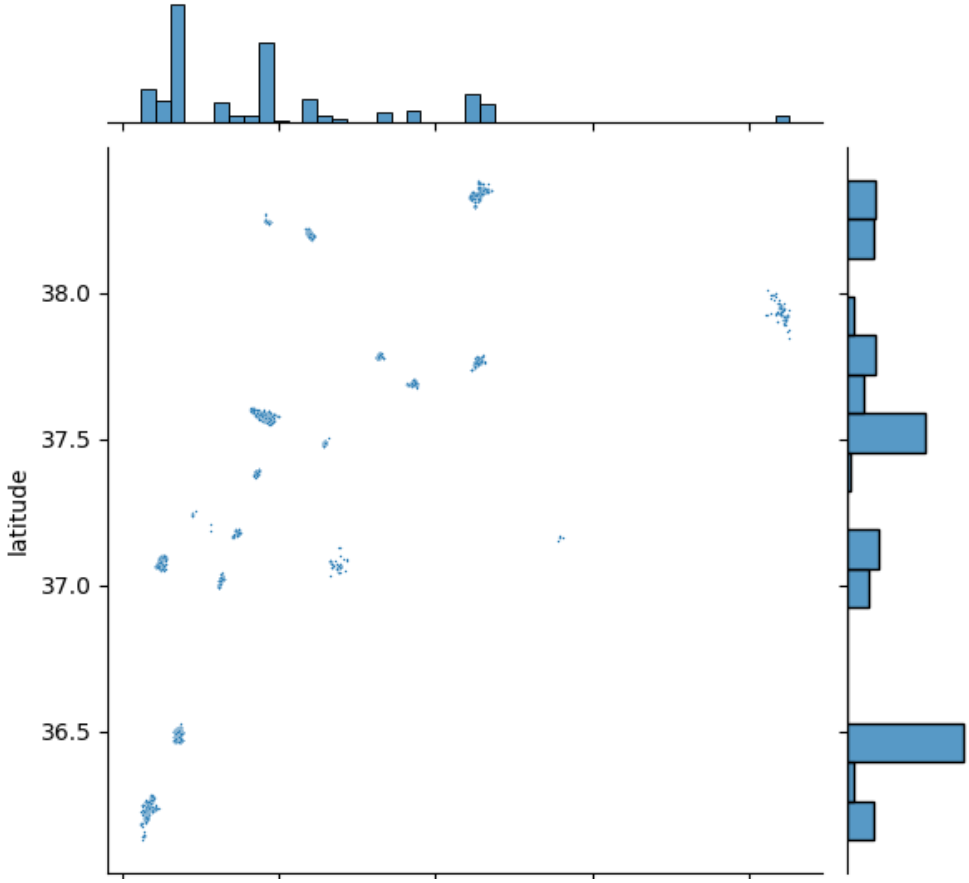

Figure 16 Scatter points for destroyed and damaged buildings

Source: Yazganoglu, G (2023)

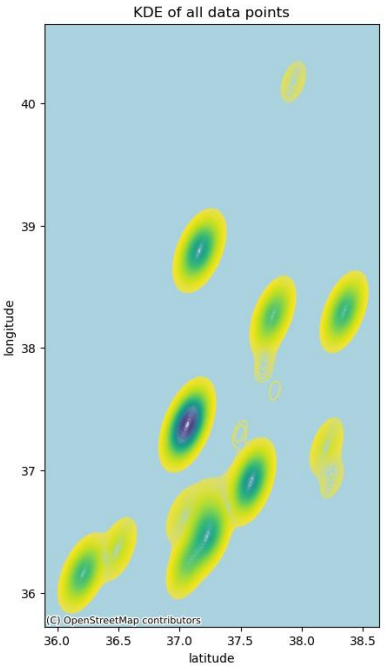

Figure 17 KDE of all data

Source: Yazganoglu, G (2023)

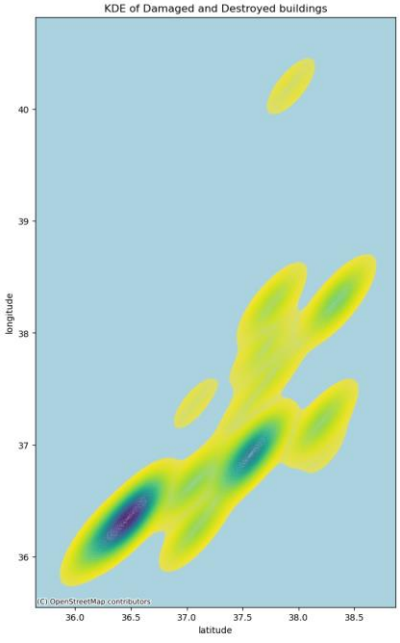

Figure 18 KDE of Damaged and Destroyed Buildings.

Source: Yazganoglu, G (2023)

Another aspect we would like to understand is randomness in point and pattern analysis. One of the ways is measuring quadrant count statistic, mapping observations and making another simulation of random observations and comparing with each other.

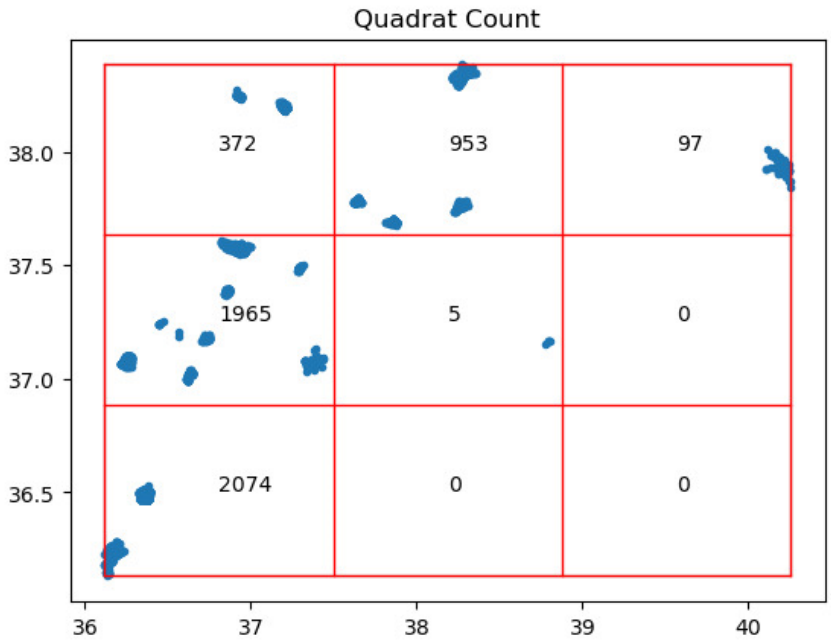

Figure 18 Quadrant Count for Real Data

Source: Yazganoglu, G (2023)

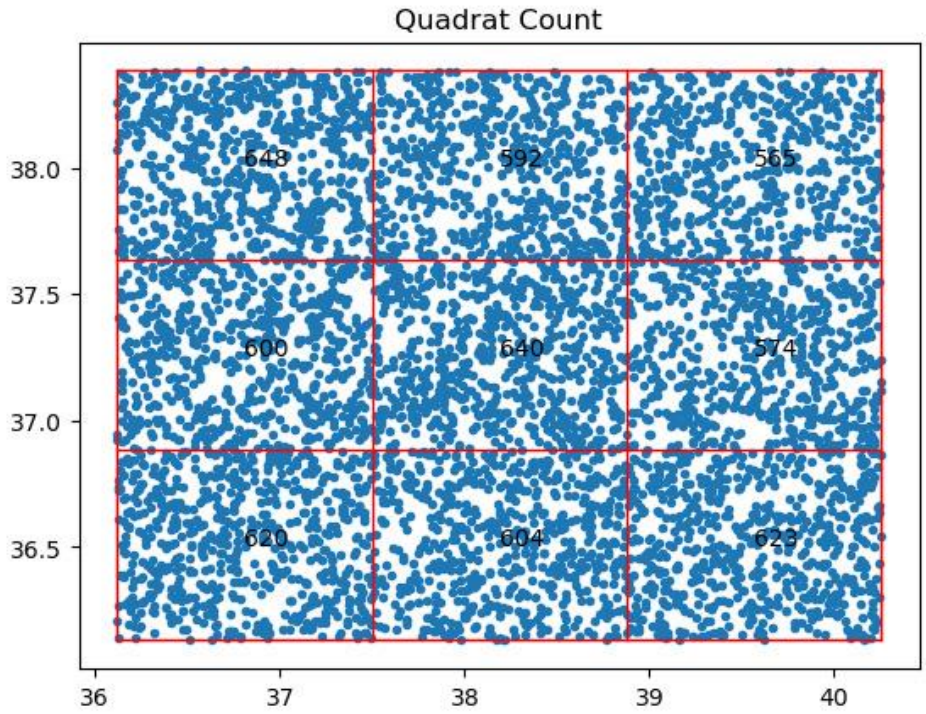

Figure 19 Quadrant Count of randomized simulation of same data

Source: Yazganoglu, G (2023)

In this study it is observed that p value for q statistic found very close to 0 resulting that we have to reject the hypothesis of damaged and destroyed buildings are randomly distributed. We can conclude

that they are clustered. This is because of 2 reasons. Firstly overall buildings are clustered in city centers. Secondly damaged and destroyed buildings are tend to stay close to Pazarcik and Nurdagi earthquake points. Can be clearly inferred from random and real quadrant count maps (Yazganoglu, 2023).

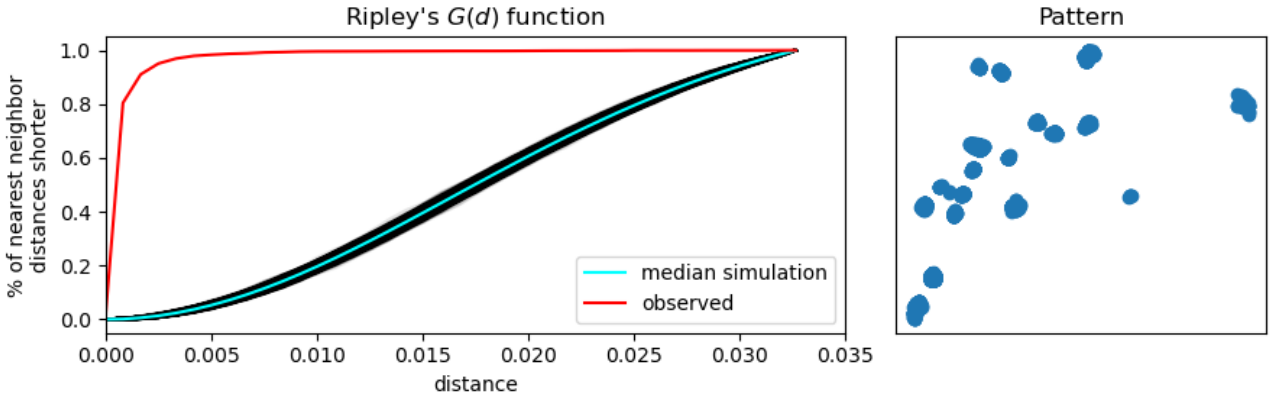

Figure 20 Ripley's G(d) distribution of Randomness.

Source: Yazganoglu, G (2023)

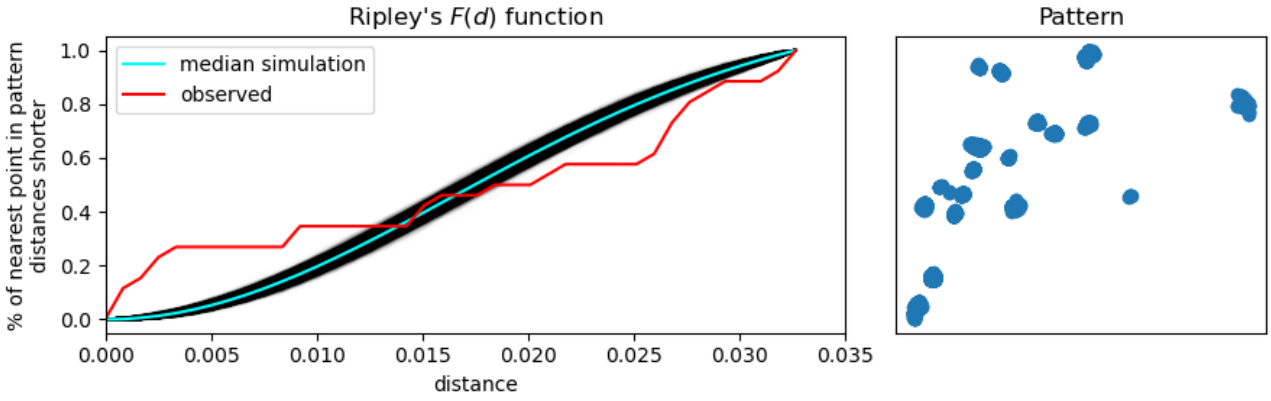

Figure 21 Ripley's F(d) Distribution of Randomness

Source: Yazganoglu, G (2023)

Another way to check randomness is checking with Ripley's G and F functions. As reflected on G(d) and F(d) functions' graphs point pattern seems to exhibit a mixed behavior. At smaller scales, there's clustering (events tend to be closer together), but as we consider larger scales, this clustering disappears, and there's a tendency towards regularity (events tend to be more evenly spaced). Specifically, F(d) function suggests that in the closest at smaller scales (distances up to 0.015), there's a tendency for events to be closer to any random location in space than what would be expected under CSR (Complete Spatial Randomness). This again indicates clustering at these scales. However, at larger scales (distances greater than 0.015), there's a tendency for events to be farther from any random location than expected under CSR, suggesting regularity or inhibition at these scales.

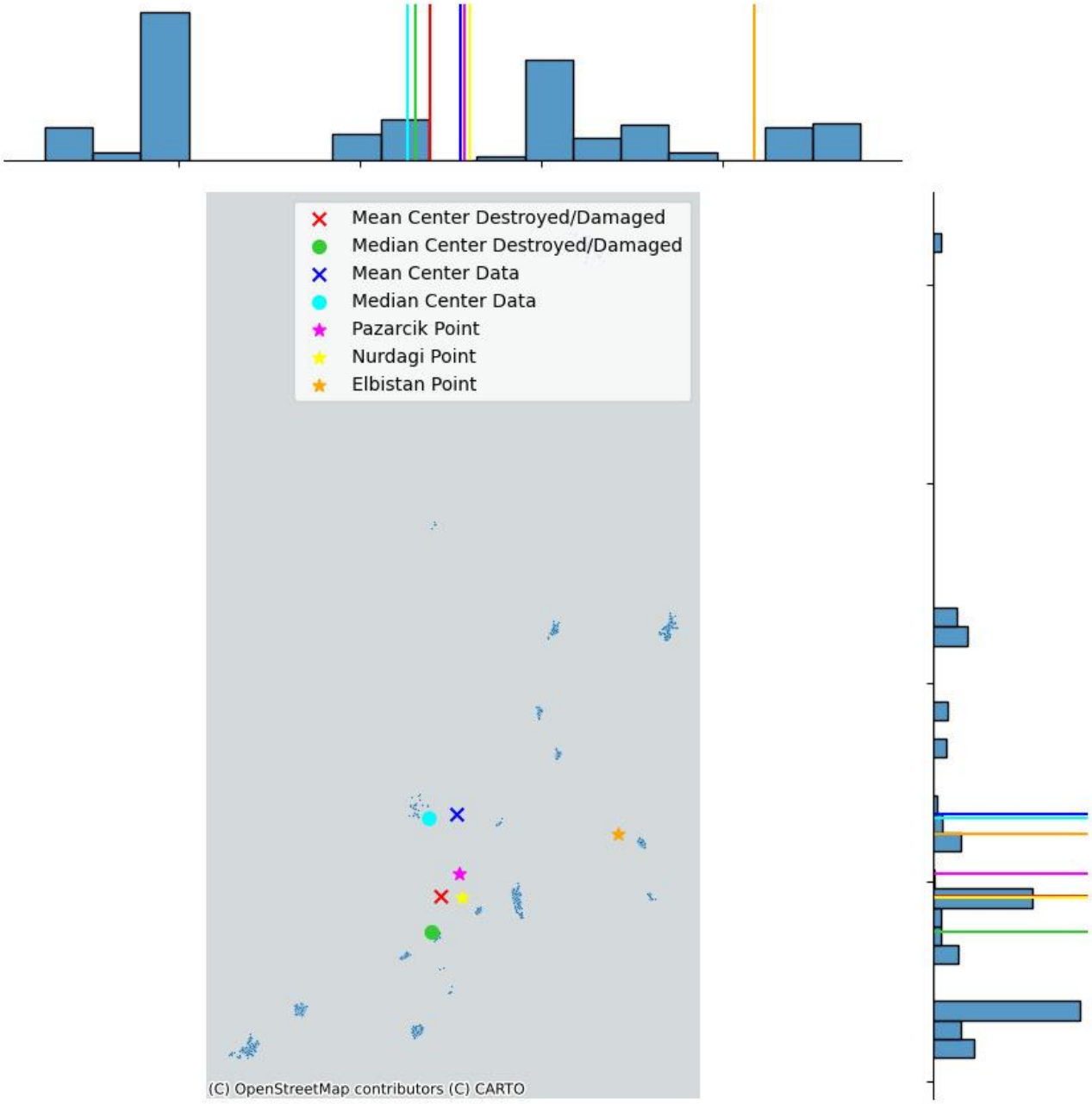

Figure 22 Mean, Median and Center points and comparison with important events on 6th of February

Source: Yazganoglu, G (2023)

When we reflect earthquake points on the map it is found that Nurdagi and Pazarcik earthquakes are closer to the damaged destroyed buildings in average, meantime Elbistan point is further than mean, median of the center. Meantime, center points for all buildings is in a different area. We can conclude that Nurdagi and Pazarcik earthquakes has significant effect on damage level.

The average dispersion is calculated as 1.046 degrees considering latitude and longitude (Yazganoglu, G 2023). This value shows that in this distance buildings are likely to have same kind of properties in terms of damage level.

4.1.2.2. Spatial Autocorrelation

By calculating spatial autocorrelation and spatial weights the objective is to understand if similar values are surrounded by similar values as in we tried to understand in point pattern analysis.

As reflected in Table, Moran's I value are suggesting that there is strong spatial relationship with variables damaged_percentage, destroyed_percentage, nearest_water_distance, nearest_camping distance, nearest_earthquake distance and nearest_fault_distance. This is understandable as these variables already have been created considering a locational relationship. However, damage_gra also suggest a moderate Moran's I value as the target value. This suggest that although there are factors that geografically affecting the damage level, there are other factors that are not related to the earthquake. Perhaps the way it is constructed that we do not have it in our data at all could be also affecting damage level.

Column	**Moran's I Value**	**Interpretation**
percentage	0.658767	Moderate positive spatial autocorrelation. Similar values tend to cluster together.
damaged_percentage	1	Strong positive spatial autocorrelation. Almost perfect clustering of similar values.
destroyed_percentage	1	Strong positive spatial autocorrelation. Almost perfect clustering of similar values.
nearest_water_source_distance	0.999846	Very strong positive spatial autocorrelation. Highly clustered similar values.
nearest_camping_distance	1.000024	Strong positive spatial autocorrelation. Almost perfect clustering of similar values.
nearest_earthquake_distance	1.000005	Strong positive spatial autocorrelation. Almost perfect clustering of similar values.
nearest_fault_distance	0.999976	Very strong positive spatial autocorrelation. Highly clustered similar values.
damage_gra	0.532654	Positive spatial autocorrelation but not as strong as others

Table 1Moran's values for spatial variables and interpretation

Source: Yazganoglu, G (2023)

4.1.3. Insights

The study presents an insightful analysis of the distribution and patterns of the disaster zone in Turkey, focusing particularly on building clusters and fault lines. The visualization from Map1 unequivocally demonstrates that Turkey, especially areas outside the current disaster zone, harbors dangerous fault lines. This information substantiates the persistent warnings from Turkish scientists regarding a potential large-scale earthquake in Istanbul. Given Istanbul's strategic location, demographic, and economic importance, such a disaster would be catastrophic, both in human and economic terms.

Our dataset, encapsulating various cities, vividly showcases the clustering of buildings and facilities in particular regions. Interestingly, there is a noticeable divergence in the distribution pattern when comparing all buildings to just the damaged or destroyed ones. This suggests that the impact of the disaster is somewhat selective and not uniformly spread across the regions.

The point pattern analysis, as elucidated by the G(d) and F(d) function graphs, indicates a dual nature in the spatial distribution of events. At smaller scales, a clustering pattern emerges, with events occurring close to one another. However, as the scale broadens, this clustered pattern fades, leading to a more regular distribution, with events being evenly spaced.

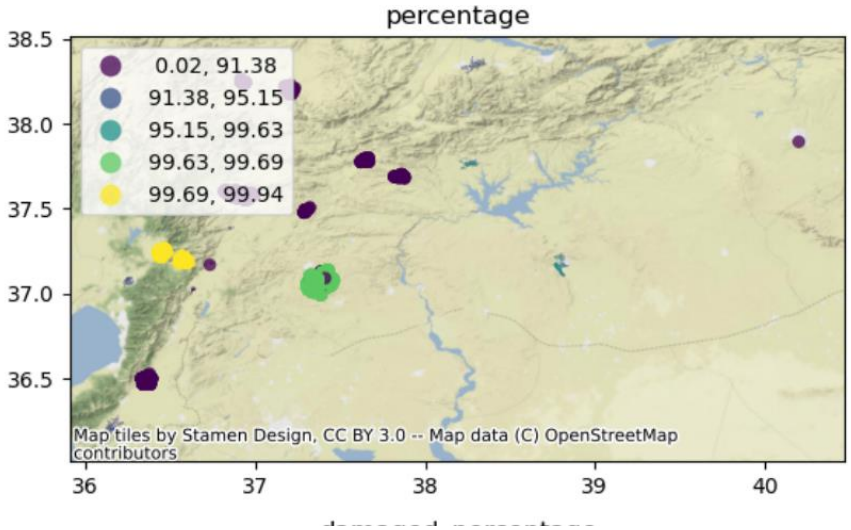

Figure 23 Spatial Mapping for the percentage variable

Source: Yazganoglu, G (2023)

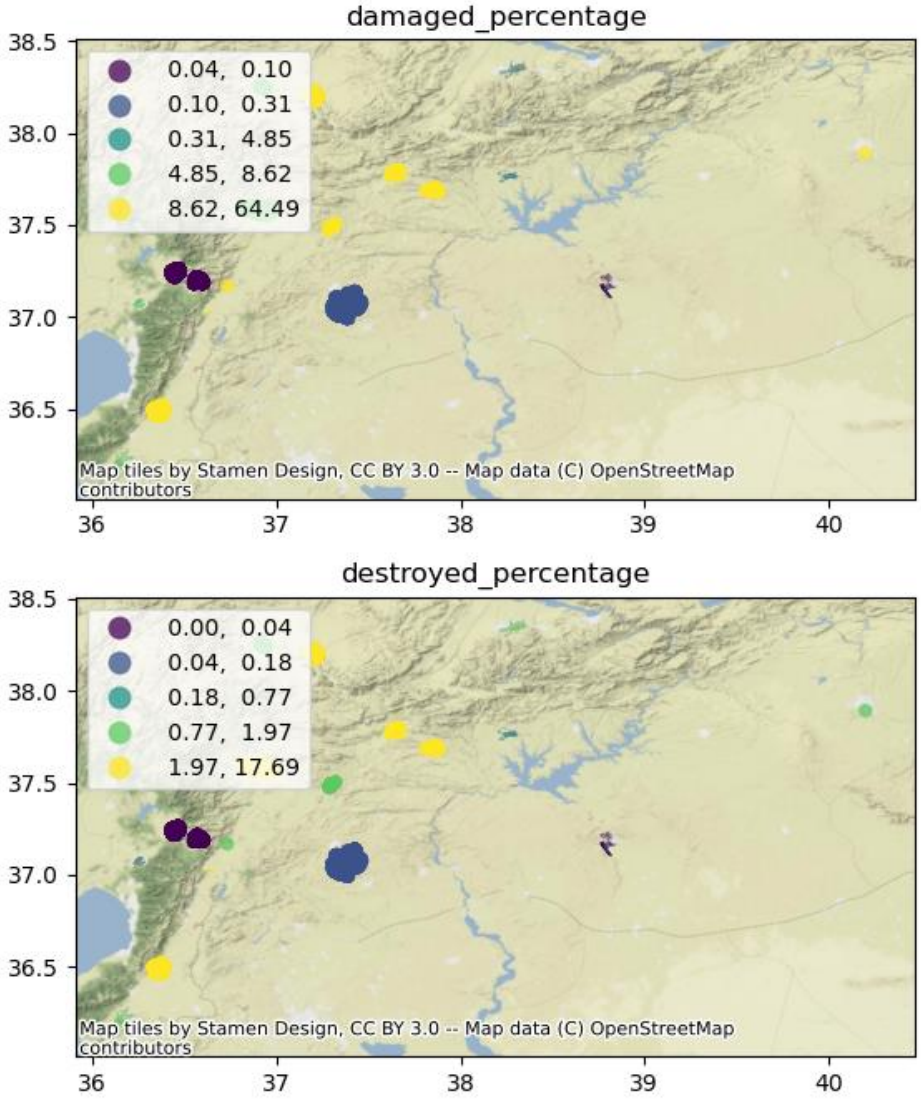

Figure 24 Spatial Mapping for Destroyed and Damaged Regions

Source: Yazganoglu, G (2023)

Moreover, the spatial autocorrelation analysis provides valuable insights. With Moran's I values serving as the benchmark, a strong spatial relationship emerges for variables like damaged_percentage, destroyed_percentage, and proximity metrics like nearest_water_distance and nearest_fault_distance. These high Moran's I values denote a profound clustering of similar values, underscoring the importance of locational relationships in the context of disasters. Surprisingly, the damage_gra column also exhibited a moderate spatial autocorrelation, suggesting other factors beyond the geographical ones might influence the extent of damage. Potential variables like construction quality, which are not present in our dataset, could also play a pivotal role in dictating damage levels.

In summary, while the geographical distribution of buildings and fault lines plays a significant role in determining the impact of disasters, other latent factors, not immediately obvious, might also have a consequential influence. As Turkey gears up to address its vulnerabilities, such comprehensive analyses are vital in developing robust disaster preparedness and response strategies.

4.2. Machine Learning Models

As mentioned earlier, after the data analysis supervised and unsupervised machine learning models have been conducted in following sections insights will follow

4.2.1. Supervised Machine Learning

A automatic Classification experiment was conducted to the dataset that has been parameters of the experiment has been as follows.

	Description	Value
0	Session id	123
1	Target	damage_gra
2	Target type	Multiclass
3	Target mapping	1: 0, 2: 1, 3: 2, 4: 3
4	Original data shape	(98272, 49)
5	Transformed data shape	(98272, 10)
6	Transformed train set shape	(68790, 10)
7	Transformed test set shape	(29482, 10)
8	Numeric features	45
9	Categorical features	3
10	Preprocess	True
11	Imputation type	simple
12	Numeric imputation	mean
13	Categorical imputation	mode
14	Maximum one-hot encoding	25
15	Encoding method	None
16	Remove multicollinearity	True
17	Multicollinearity threshold	0.900000
18	Feature selection	True
19	Feature selection method	classic
20	Feature selection estimator	lightgbm
21	Number of features selected	0.200000
22	Fold Generator	StratifiedKFold
23	Fold Number	10
24	CPU Jobs	-1
25	Use GPU	True

Table 2 Supervised Machine Learning Setup for the Classification Model

Source: Yazganoglu, G (2023)

Using these parameters the auto ML model check all available classification models also applying feature selection and validation to the models.

After the experiment, we end up with the classification results according to which random forest has the best scores and selected as the model to be optimized.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
rf	Random Forest Classifier	0.9636	0.9748	0.9636	0.9587	0.9607	0.6216	0.6263	0.5020
lightgbm	Light Gradient Boosting Machine	0.9621	0.9830	0.9621	0.9553	0.9579	0.5928	0.6009	0.4480
et	Extra Trees Classifier	0.9613	0.9688	0.9613	0.9577	0.9593	0.6116	0.6137	0.5820
gbc	Gradient Boosting Classifier	0.9612	0.9786	0.9612	0.9533	0.9562	0.5736	0.5843	0.4730
ada	Ada Boost Classifier	0.9586	0.9672	0.9586	0.9499	0.9533	0.5463	0.5562	0.3740
knn	K Neighbors Classifier	0.9548	0.9157	0.9548	0.9449	0.9486	0.4929	0.5051	0.5950
dt	Decision Tree Classifier	0.9536	0.8494	0.9536	0.9538	0.9537	0.5671	0.5672	0.3090
lr	Logistic Regression	0.9535	0.9611	0.9535	0.9422	0.9463	0.4645	0.4806	0.4330
ridge	Ridge Classifier	0.9503	0.0000	0.9503	0.9323	0.9345	0.2924	0.3553	0.2690
dummy	Dummy Classifier	0.9444	0.5000	0.9444	0.8919	0.9174	0.0000	0.0000	0.3340
lda	Linear Discriminant Analysis	0.9390	0.9627	0.9390	0.9453	0.9414	0.4791	0.4816	0.3800
qda	Quadratic Discriminant Analysis	0.9210	0.9461	0.9210	0.9457	0.9317	0.4219	0.4343	0.4870
nb	Naive Bayes	0.9182	0.9380	0.9182	0.9436	0.9292	0.4018	0.4139	0.3090
svm	SVM - Linear Kernel	0.8846	0.0000	0.8846	0.9356	0.8903	0.3330	0.3522	0.2420

Table 3Classification Experiment Results

Source: Yazganoglu, G (2023)

Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.9594	0.9778	0.9594	0.9505	0.9532	0.5510	0.5623
1	0.9589	0.9727	0.9589	0.9494	0.9527	0.5367	0.5504
2	0.9596	0.9739	0.9596	0.9532	0.9532	0.5499	0.5622
3	0.9602	0.9777	0.9602	0.9537	0.9549	0.5731	0.5806
4	0.9597	0.9722	0.9597	0.9532	0.9535	0.5496	0.5629
5	0.9580	0.9787	0.9580	0.9513	0.9517	0.5363	0.5474
6	0.9597	0.9744	0.9597	0.9511	0.9528	0.5536	0.5657
7	0.9564	0.9714	0.9564	0.9501	0.9499	0.5151	0.5267
8	0.9570	0.9749	0.9570	0.9503	0.9503	0.5172	0.5302
9	0.9581	0.9739	0.9581	0.9521	0.9506	0.5274	0.5419
Mean	0.9587	0.9747	0.9587	0.9515	0.9523	0.5410	0.5530
Std	0.0012	0.0024	0.0012	0.0014	0.0015	0.0170	0.0160

Table 4 Optimization Results obtained with PyCaret tune_model

Source: Yazganoglu, G (2023)

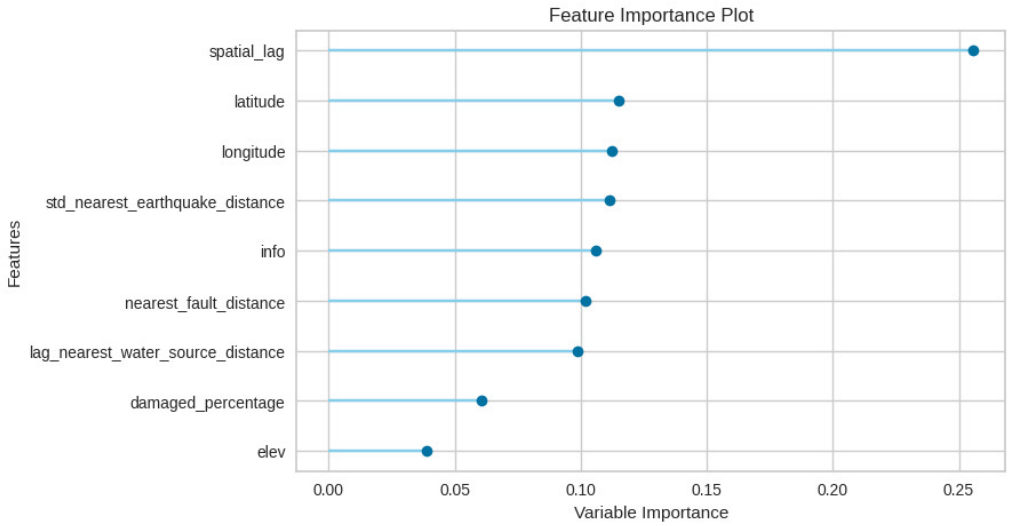

Figure 25 Feature Importance Plot for the Random Forest Model

Source: Yazganoglu, G (2023)

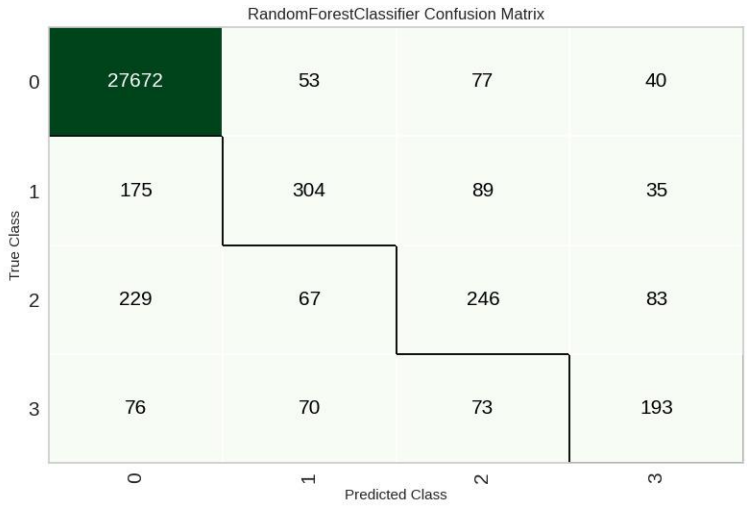

Figure 26 Confussion Matrix for Random Forest Model

Source: Yazganoglu, G (2023)

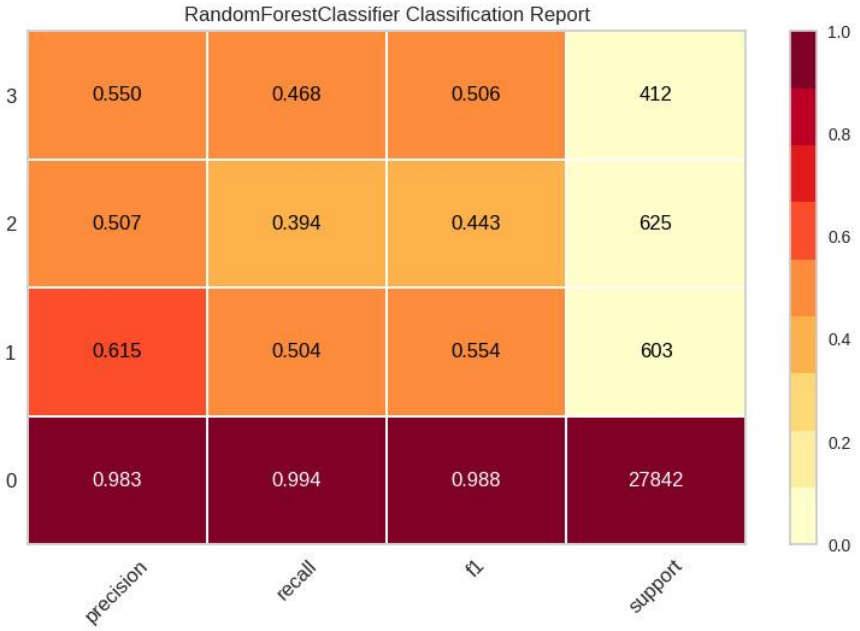

Figure 27 Classification Matrix of the Random Forest Model

Source: Yazganoglu, G (2023)

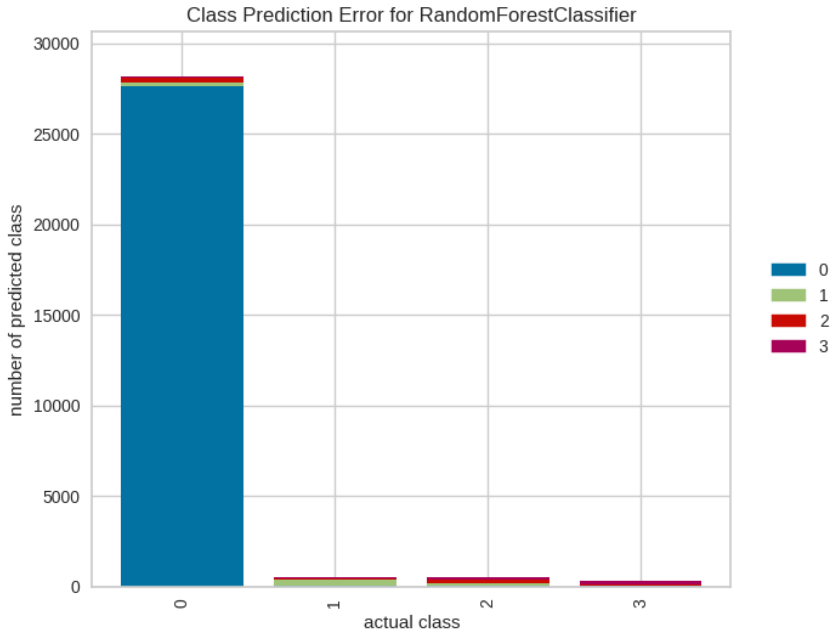

Figure 28 Class Prediction Error for Random Forest Classifier

Source : Yazganoglu, G (2023)

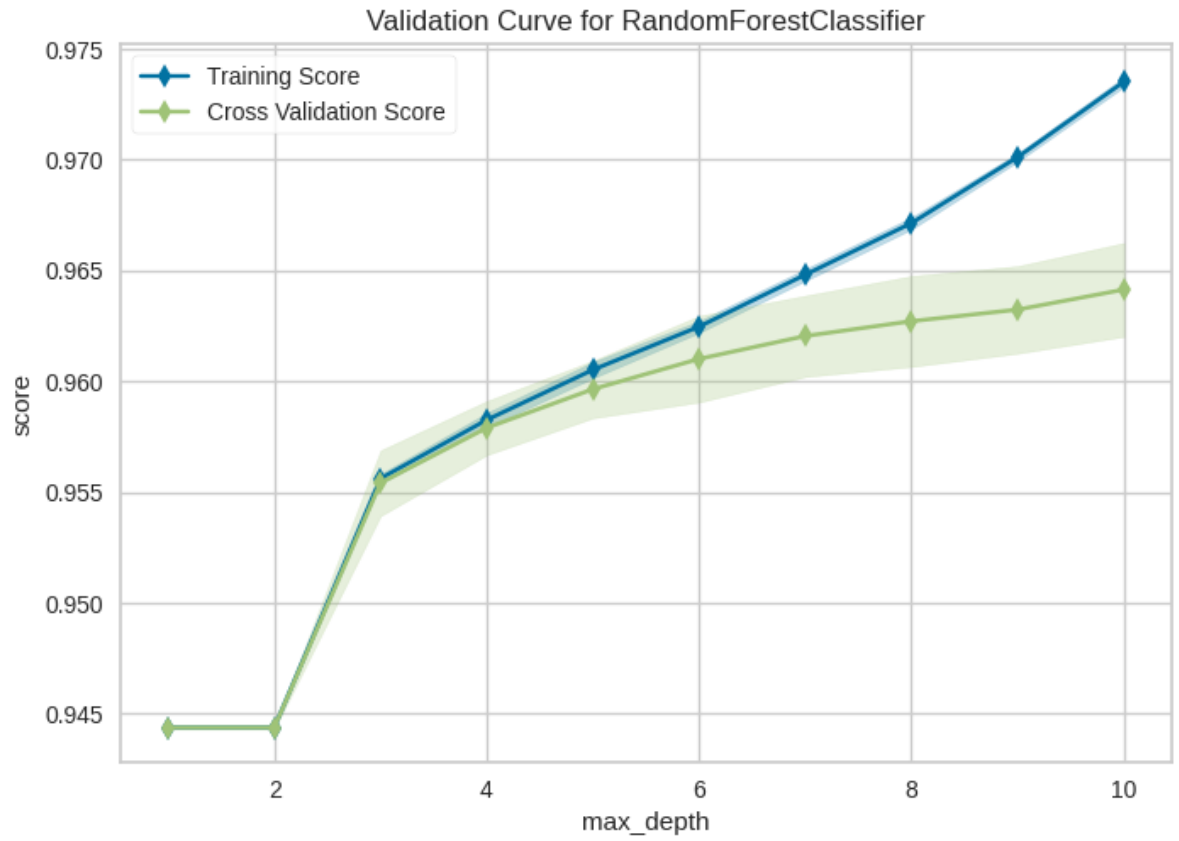

Figure 29 Validation Curve for Random Forrest Model

Source: Yazganoglu, G (2023)

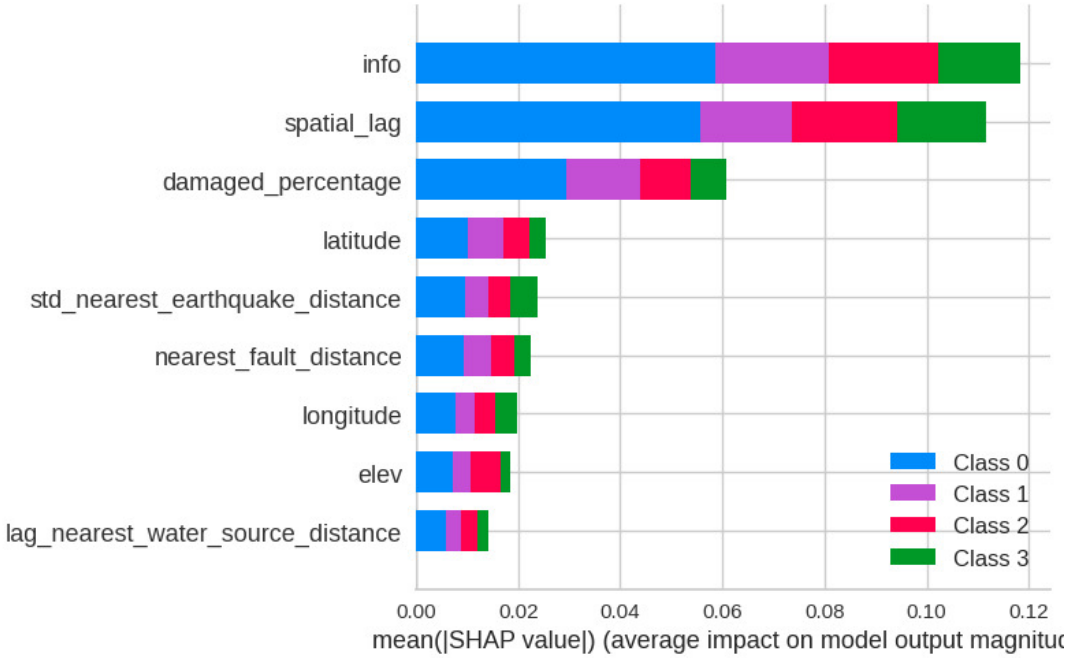

Figure 30 Mean SHAP Value average impact on model output.

Source: Yazganoglu, G (2023)

Interpreting SHAP values in a multiclass classification model adds an extra layer of complexity compared to binary classification or regression models, as we need to consider the contribution of features towards each class prediction.

As figure reflects, building 'info' and 'spatial_lag' s are the important values to determine building damage level in overall. We also observe fault distance and earthquake distance are also following important values. These 2 variables also is the most used ones for all classes.

A partial dependence plot can show whether the relationship between the target and feature is linear, monotonic, or more complex (Cohen, 2021). In our example, info is a categorical variable and perhaps do not have many different unique values for some values even the value doesn't change the impact to the result changes a lot. We suspect specific values of info outsizes the impact of model's predictions.

In summary we conclude that among spatial variables spatial_lag of damage gra, damaged_percentage in the locality, distance to the earthquake and fault are important and have a high impact on all classes.

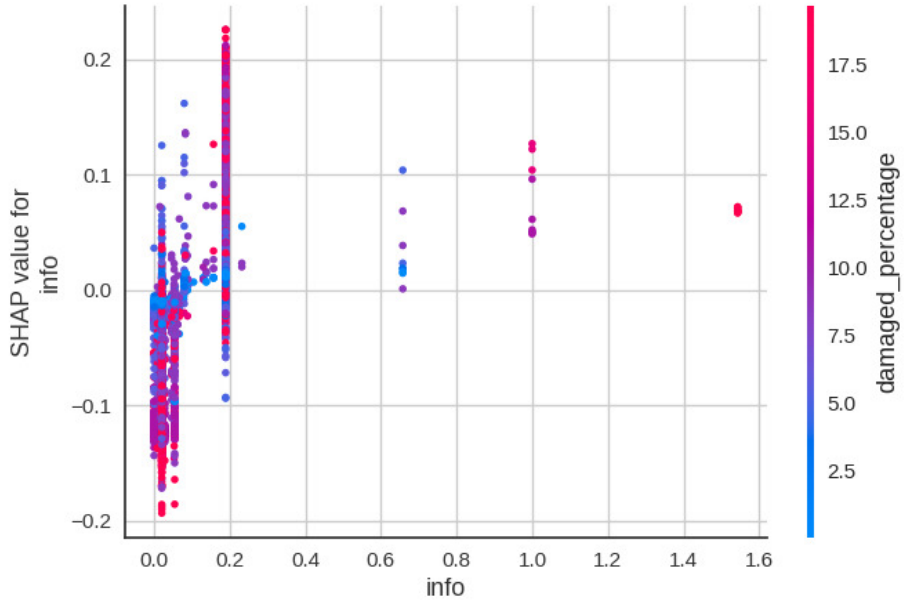

Figure 31 Shap Dependency for Info- Damaged Percentage

Source: Yazganoglu, G (2023)

4.2.2. Unsupervised Machine Learning

After conducting a clustering experiment, we have found optimum according to the metrics was K-Means clustering into 4 different clusters. K-Means has superior Silhouette, Calinski-Harabasz, and Davies-Bouldin scores, suggesting better, more distinct, and well-separated clusters compared to DBSCAN. Both algorithms perform poorly in terms of homogeneity, Rand Index, and completeness. This might indicate that if there are ground-truth class labels, neither clustering algorithm aligns well with them. This is due to buildings do not spread homogenously.

Given the data, K-Means seems to be the better clustering model in terms of defining distinct clusters. We can try to profile what is the profile for these clusters.

	Description	Value
0	Session id	6993
1	Original data shape	(98272, 49)
2	Transformed data shape	(98272, 120)
3	Numeric features	46
4	Categorical features	3
5	Preprocess	True
6	Imputation type	simple
7	Numeric imputation	mean
8	Categorical imputation	mode
9	Maximum one-hot encoding	-1
10	Encoding method	None
11	CPU Jobs	-1
12	Use GPU	False
13	Log Experiment	False
14	Experiment Name	cluster-default-name
15	USI	8f7c

	Silhouette	Calinski-Harabasz	Davies-Bouldin	Homogeneity	Rand Index	Completeness
0	0.7315	645515079.2386	0.2921	0	0	0

Table 5 Clustering Experience Details.

Source: Yazganoglu, G (2023)

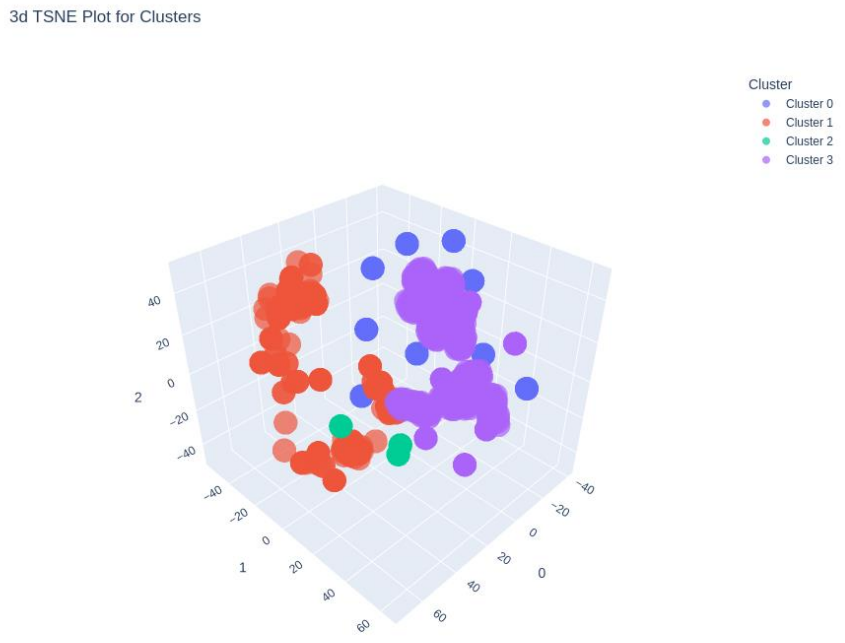

Figure 32 TSNE Plot refletting clusters in 3D visualization

Source: Yazganoglu, G (2023)

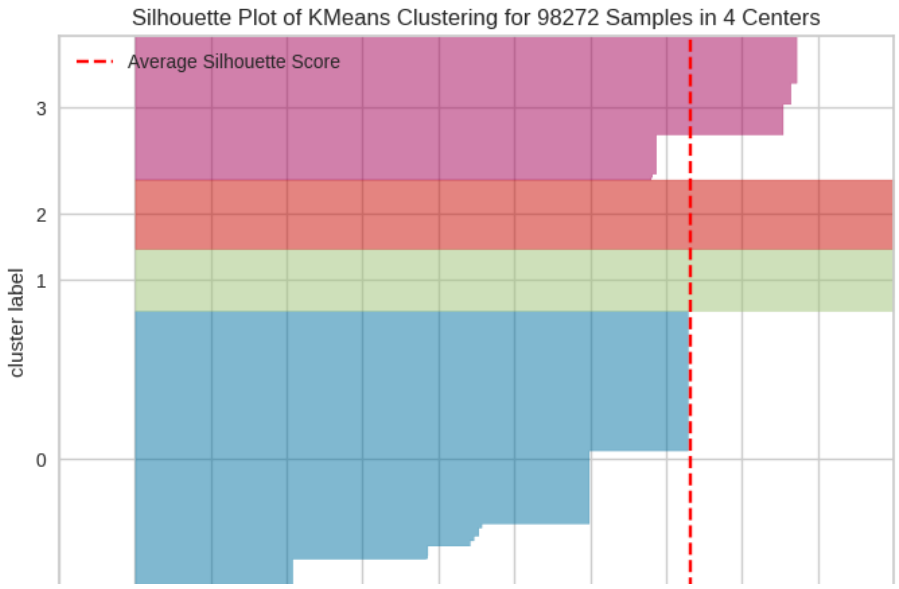

Figure 33 Silhouette Plot of KMeans Clustering

Source: Yazganoglu, G (2023)

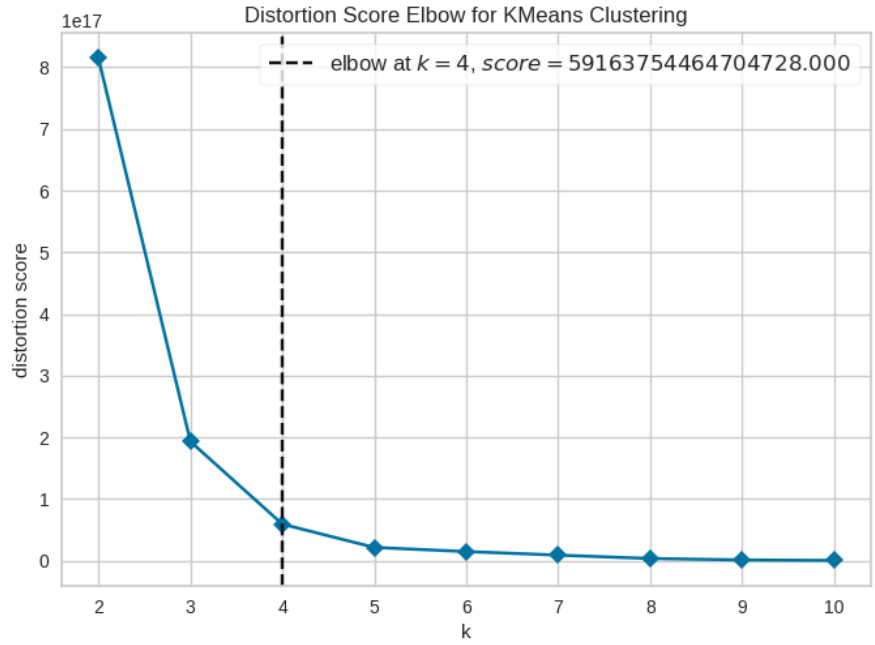

Figure 34 Elbow Graph to determine which number of clusters are the optimum.

Source: Yazganoglu, G (2023)

Clustering Metrics and Insights

After dividing into groups another explanatory analysis is needed to understand what the properties of these clusters are. We will name clusters after the numbers assigned by the algorithm to each.

Cluster 0

This cluster is predominantly present in 'ANTAKYA' (8196 counts) followed by 'KIRIKHAN' (2498 counts). No representation in the majority of other localities.

This cluster contains the second highest number of observations. It primarily experienced moderate damages. The most affected building types in this cluster were residential buildings, followed closely by highways, streets, and roads, and non-residential buildings. Despite this, the cluster had the lowest population and the third highest income. Furthermore, it reported the lowest figures in both total sales and secondhand sales.

Meantime, this cluster also has damages in building types such as railways, electricity lines, bridges types of buildings which require extra attention in case of an emergency. This types of buildings should be replaced again with reinforcement as in this cluster it is more likely that these kind of buildings get damages.

Cluster 1

This cluster majorly concentrated in 'GAZIANTEP' with 24025 counts.Other significant localities include 'KAHRAMANMARAS' (12584 counts) and 'MALATYA' (8269 counts).

Holding the fourth largest number of observations, Cluster 1 encountered the most severe damages. Residential buildings were the most impacted, trailed by highways, streets, and roads, and non-residential buildings. The population density here was relatively low, ranking second lowest among all clusters. Economically, it stood second in income levels and third in both total and secondhand sales metrics. Apart from these damages we do not observe damages in a weighted damage model.

Cluster 2

Highly localized in 'SANLIURFA' with 11941 counts. Almost no representation in other localities.

This cluster, with the third highest observations, interestingly reported no complete destructions and had the least amount of damage. Residential buildings, highways, streets, and roads, along with other civil engineering works, took the most damges. It was characterized by a densely populated locality, the lowest income levels, but paradoxically, had the highest numbers in total sales and secondhand sales.

Cluster 3

Cluster spread across multiple localities with high counts in 'ADIYAMAN' (6772), 'DUZICI' (7995), and 'BAHCE' (3632).The most populated in terms of observations, Cluster 3 experienced moderate damages. The main structures affected here were highways, streets, and roads. Residential buildings and unclassified structures also reported significant damages. The locality was quite crowded, and it boasted the highest income levels. In terms of sales, it held the second position in both total and secondhand sales categories. Therefore we can conclude that before the earthquake there has been a market in this cluster and yet will there be more after reconsturction as it has been hichly populated-

Unfortunately using this dataset, we are not able to conclude much about schools, theaters which are important about culture of the city. However considering the habits and damages have been received so far, one can conclude that preserving the culture as it is important. Buildings should be reinforced in a way without gentrification. Crowded households with low income level should be able to afford the

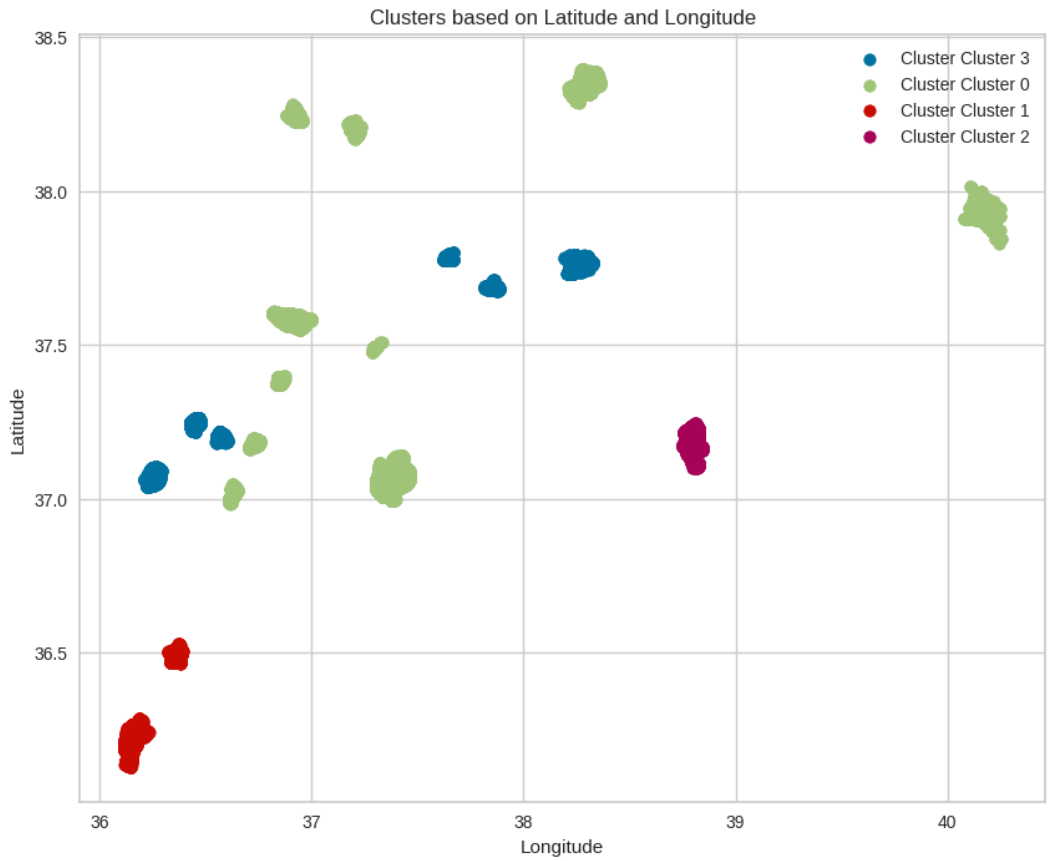

Figure 35 Clusters Reflected on the Map

Source: Yazganoglu, G (2023)

Conclusion

In this study, the journey had started with the biggest disaster of the year, ambition to understand and avoid damages, losses, destruction and a group of satellite datasets from EU Copernicus database reflecting damage grades after the 6th of February 2023 earthquakes for buildings, roads and facilities within 19 locations. Later layers such as distances to water sources, faults and earthquake locations and emergency campsites were added and merged with socioeconomic information from TUIK.

After the creation of main dataset point-pattern analysis and spatial autocorrelation analysis was made to understand how geography affecting the damage level. Results have showcased that location of Pazarcik and Nurdagi earthquakes are closer to the center point of all damaged and destroyed buildings than the other important earthquake of the day in Elbistan. All observations have another central point which reflects that these earthquakes are closer to the damaged buildings. In addition to this, Both all observations and subset of damaged and destroyed buildings are clustered in certain points that we have to reject random distribution of this dataset. Findings were parallel with the analysis of spatial autocorrelation which demonstrated that damaged level spatially autocorrelated around 65% which shows that damaged, destroyed and not damaged buildings tend to appear close to each other. Zone can give some idea about the buildings.

Results of multiclass supervised classification experiment was also curious. The best fitting model is found as Random Forest classifier. This model is has good results in several criteria as well as do its best to predict all classes. According to this model the best explainers are building type and spatial variables such as distance to earthquake and distance to fault. We are able to conclude that in general spatial variables explain this phenomena better than the socioeconomic variables.

In order to map problematic areas, K-Means and DBSCAN models were also experimented, According to silhouette scores K-Means model was a better fit and this model is optimum at 4

clusters. It is observed that socioeconomic variables were more active in this model as descriptive statistics for clusters suggests.

These steps were relevant to understand the problem and answer the objective questions in the beginning we have. Results showcase that distance to fault and distance to earthquake is important. Distance to fault is a factor we can know for all existing buildings in the world but as of today's technology we are not able to determine time, place, and magnitude of an earthquake. We know that they happen around fault areas because of the geological movements.

Another question in mind was how the clusters are. Point pattern analysis show that destroyed and damaged points are likely to be close to each other and places with more density are close the disastrous region. As the nature of the dataset, localities are far from each other resulting in we have to reject randomness for all dataset and filtered dataset for damaged and destroyed buildings.

Machine learning techniques are able to detect which regions are more vulnerable according to others. In this dataset we haven't observed hard clustering due to lack of specific local information but the aggregate information we have, we are able to see which regions have more buildings, which one has more damaged buildings which are more close to the disastrous regions.

This is a very important business problem as the life of people and the regional economy as at stake. Future efforts should be trying to avoid these kinds of disasters such as earthquake is inevitable. In addition, it is not only about avoiding but also it is about the recovery of the affected regions.

Research into earthquake dynamics necessitates a comprehensive, interdisciplinary approach, calling for the expertise of not just seismologists but also economists, architects, civil engineers, and data scientists. A thorough examination of existing literature and the models employed in this study indicates that the extent of damage from earthquakes is affected by the geographical location of the area. Additionally, there are other pivotal factors that, unfortunately, went unobserved in this research. Given the heightened seismic risk observed across many cities in Turkey, there's a pressing need for more in-depth investigations backed by richer datasets. Possessing a more detailed dataset would undeniably enable us to extract deeper, more nuanced insights into the dynamics and impact of earthquakes.

Our efforts aimed at mapping earthquake impacts were largely based on broader locality data due to the absence of more granular information. A methodology that delves into more localized areas, rather than general localities, would likely yield richer insights. For a more nuanced understanding, the incorporation of geocoding could be invaluable, especially when one is interested in meticulous mapping within city subdivisions. Moreover, employing specific spatial weights would further refine the granularity and precision of the study's findings.

Considering the earthquake as a whole concept the city planning should be considered very carefully, materials should be elected carefully, ground should be examined in details and even after all these precautions, possible economic damages should be covered. This demonstrate that there are a lot of business opportunities considering unaffected zones in Turkey. There should be a lot of buildings in different areas which might represent the same kind of risks and very highly to end up demolished. These kinds of buildings should be reinforced or rebuilt and even after that the possible losses could be covered.

The housing market is very high which comes time to time with a loan market. As guarantee for bankruptcy in a mortgage is usually only the building itself, banks offer, to the clients products such as life insurance or housing insurance...etc. In case of destruction not only the loan holders lose their place to live but also bank loses future collectibles and assets. A comprehensive insurance for the earthquake should be considered as a side product and perhaps both insurance customers and banks consider auditing the building with the civil engineers.

References

- AFAD. (2023). Event Catalog. https://deprem.afad.gov.tr/event-catalog
- Ahmad, Z., Mahmoudi, E., & Kharazmi, O. (2020). On Modeling the Earthquake Insurance Data via a New Member of the T-X Family. *Computational Intelligence and Neuroscience*, 2020, Article ID 7631495. DOI: 10.1155/2020/7631495. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7520674/
- Alam, M. (2020). Point pattern analysis of location data. *Towards Data Science*. https://medium.com/p/1346f318865d
- Anello, E. (2019). A Practical Introduction to Geospatial Data Analysis using QGIS. Retrieved from https://towardsdatascience.com/a-practical-introduction-to-geospatial-data-analysis-using-qgis-a6f82105b30e
- Bex T. (2022). *A Complete SHAP Tutorial: How to Explain Any Black-box ML Model in Python. Towards Data Science.* Retrieved from https://towardsdatascience.com/a-complete-shap-tutorial-how-to-explain-any-black-box-ml-model-in-python-7538d11fae94
- Bivand, R. S., Pebesma, E., & Gómez-Rubio, V. (2013). *Applied Spatial Data Analysis with R* (2nd ed.). Springer Science+Business Media. https://doi.org/10.1007/978-1-4614-7618-4_1
- Bivand, R. (2022). R Packages for Analyzing Spatial Data: A Comparative Case Study with Areal Data. Geographical Analysis, 54(4), 488-518.
- Cohen, I. (2021, July 12). Explainable AI (XAI) with SHAP -Multi-Class Classification Problem: A practical guide for XAI analysis with SHAP for a Multi-class classification problem. Towards Data Science. https://towardsdatascience.com/explainable-ai-xai-with-shap-multi-class-classification-problem-64dd30f97cea
- Dincer, B. (2023). Feb 06-23 earthquake turkey cities vectors. Kaggle. https://www.kaggle.com/datasets/brsdincer/feb-06-23earthquake-turkey-citiesvectors
- Education Ecosystem (LEDU). (2018). Understanding K-means Clustering in Machine Learning. *Towards Data Science*. Retrieved from https://medium.com/towards-data-science/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1.
- European Commission - Copernicus Emergency Management Service. (2023, February 13). Report EMSR648: Earthquake Türkiye. https://emergency.copernicus.eu/mapping/download/201003/Report_EMSR648_Earthquake_T%C3%BCrkiye_20230213.pdf?redirect=list-of-components/EMSR648/GRADING/ALL
- Garg, Y., Masih, A., & Sharma, U. (2021). *Predicting Bridge Damage During Earthquake Using Machine Learning Algorithms*. B. Tech CSE. ASET, Amity University, Noida, India retrieved from.https://ieeexplore.ieee.org/abstract/document/9377100
- Géron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 3rd Edition*. O'Reilly Media, Inc. ISBN: 9781098125974.
- GeoPandas Development Team. (2023). `geopandas.sjoin`.(01 August 2023) from https://geopandas.org/en/stable/docs/reference/api/geopandas.sjoin.html
- Graunt, J. (1662). Natural and Political Observations Made upon the Bills of Mortality. London: T. Roycroft. Retrieved from edstephan.org
- Hucker, M . (2020)."Tree algorithms explained: Ball Tree Algorithm vs. KD Tree vs. Brute Force - Understand what's behind the algorithms for structuring Data for Nearest Neighbour Search." *Towards Data Science*. Jun 15, 2020.

https://towardsdatascience.com/tree-algorithms-explained-ball-tree-algorithm-vs-kd-tree-vs-brute-force-9746debcd940

- Istanbul Technical University (ITU). (2023). [ATAG]. https://atag.itu.edu.tr/v4/?p=135
- Jeffares, A. (2019, November 19). K-means: A Complete Introduction. Towards Data Science. Retrieved August 20, 2023 Retrieved from https://towardsdatascience.com/k-means-a-complete-introduction-1702af9cd8c
- Lawhead, J. (2015). *Learning Geospatial Analysis with Python* (2nd ed.). Packt Publishing.
- Liu, C., Fang, D., & Zhao, L. (2021). *Reflection on earthquake damage of buildings in 2015 Nepal earthquake and seismic measures for post-earthquake reconstruction. Structures, 30*, 647–658. Elsevier Ltd. https://doi.org/10.1016/j.istruc.2020.12.089
- Mangalathu, S., Sun, H., Nweke, C. C., Yi, Z., & Burton, H. V. (2020). Classifying earthquake damage to buildings using machine learning. *Earthquake Spectra, 36*(1), 183-208. https://doi.org/10.1177/8755293019878137
- Priambodo, B., Mahmudy, W. F., & Rahman, M. A. (2020). Earthquake Magnitude and Grid-Based Location Prediction using Backpropagation Neural Network. *Knowledge Engineering and Data Science (KEDS)*, 3(1), 28–39. DOI: 10.17977/um018v3i12020p28-39. This article is available under the CC BY-SA license on http://journal2.um.ac.id/index.php/keds/article/download/14682/6239
- PyCaret. (2022). *PyCaret 2.3.5 documentation*. Retrieved from https://pycaret.readthedocs.io/en/stable/
- ReliefWeb. (2023). Earthquakes in North-West Syria - Situation Report No. 2 (15 March 2023). ReliefWeb. Retrieved March 19, 2023, from https://reliefweb.int/report/syrian-arab-republic/earthquakes-north-west-syria-situation-report-no-2-15-march-2023
- Rey, S. J., Arribas-Bel, D., & Wolf, L. J. (2020). *Geographic Data Science with Python*. Retrieved from https://geographicdata.science/book/intro.html#
- Orcun, A (2023). Turkey 6 February disaster and related datas. Kaggle.

TÜİK (Turkish Statistical Institute). (2023). Geographic Information Portal. https://cip.tuik.gov.tr/

- Sheibani, M., & Ou, G. (2021). The development of Gaussian process regression for effective regional post-earthquake building damage inference. *Earthquake Engineering & Structural Dynamics*. Advance online publication. doi: 10.1111/mice.12630

https://www.kaggle.com/datasets/ardaorcun/turkey-6-february-disaster-and-related-datas

- Xie, Y., Sichani, M. E., Padgett, J. E., & DesRoches, R. (2020). The promise of implementing machine learning in earthquake engineering: A state-of-the-art review. *Earthquake Spectra*. DOI: 10.1177/8755293020919419. Available at https://www.researchgate.net/profile/Yazhou-Xie/publication/341892878_The_promise_of_implementing_machine_learning_in_earthquake_engineering_A_state-of-the-art_review/links/5edb9eb945851529453cb373/The-promise-of-implementing-machine-learning-in-earthquake-engineering-A-state-of-the-art-review.pdf
- Wikipedia contributors. (2023). 2023 Turkey-Syria earthquake. In Wikipedia. Retrieved March 19, 2023, from https://en.wikipedia.org/wiki/2023_Turkey%E2%80%93Syria_earthquake
- Wikipedia contributors. (2023). Random forest. In Wikipedia. Retrieved August 15, 2023 from https://en.wikipedia.org/wiki/Random_forest
- Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. Retrieved from https://www.semanticscholar.org/paper/Crisp-dm%3A-towards-a-standard-process-modell-for-Wirth-Hipp/48b9293cfd4297f855867ca278f7069abc6a9c24

- World Bank. (2023, February 27). Earthquake Damage in Türkiye Estimated to Exceed $34 Billion: World Bank Disaster Assessment Report. Retrieved from https://www.worldbank.org/en/news/press-release/2023/02/27/earthquake-damage-in-turkiye-estimated-to-exceed-34-billion-world-bank-disaster-ass
- Yazganoglu, G. (2023). *TFM*. GitHub repository. Retrieved from https://github.com/gozdeydd/tfm/tree/main