Delineation of Residential Housing Submarkets Using Spatially Constrained Multivariate Clustering

Samuel Ngere Njoroge

151522

Master of Science in Data Science and Analytics

2024

Delineation of Residential Housing Submarkets Using Spatially Constrained Multivariate Clustering

Samuel Ngere Njoroge

151522

Submitted in Partial fulfilment of the Requirements for the Degree of Master of Science in Data Science and Analytics at Strathmore University

Strathmore Institute of Mathematical Sciences

Strathmore University

Nairobi, Kenya

July 2024

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Student's Name: Samuel Ngere Njoroge

22/01/2024

Sign: _____ Date: _____

Approval

The thesis of Samuel Ngere Njoroge was reviewed and approved for defense by the following:

Dr. Victor Odipo

Supervisor,

Institute of Mathematical Sciences,

Strathmore University

Sign: _____ Date: 23/01/2024

Abstract

Every housing market is made up of unique submarkets. Submarkets are areas or neighborhoods where houses have similar features, such as the age of the houses or price. Segmenting housing markets into submarkets is recommended for better understanding and more effective interventions in the housing market. While different submarket delineation approaches exist, many do not impose spatial constraints, overlooking the spatial relationships between houses. This oversight results in submarkets with poorly defined boundaries that do not match the urban layout, making accurate spatial inferences difficult and limiting stakeholders' ability to establish policy zones. To address these limitations, this study uses the SKATER clustering algorithm, which demarcates submarkets by taking into account the location of houses and ensuing spatial relationships alongside the structural attributes of the houses. The proposed method is implemented in a case study of King County, using house sale data from May 2014 to May 2015. It identifies four submarkets with boundaries closely aligned with the landscape, marking an improvement over previous research. The analysis reveals a notable housing market imbalance whereby northern cities like Bellevue feature high-priced, spacious, high-quality houses on large lots. At the same time, the southern region, including SeaTac and Federal Way, offers older, smaller houses at relatively lower prices. These findings help stakeholders and investors make accurate spatial inferences for addressing housing challenges, particularly market imbalances.

Keywords

housing market, spatial constraints, submarkets

Table of Contents

List of Figures

List of Tables

List of Abbreviations

SKATER	Spatial Kluster Analysis by Tree Edge Removal
GIS	Geographic Information Systems
REDCAP	Regionalization with Dynamically Constrained Agglomerative clustering
CRISP-DM	Cross Industry Standard Process for Data Mining
SSD	Sum of Squared Deviations

Acknowledgements

First and foremost, I would like to express my gratitude to God for His grace throughout this journey. I sincerely thank my supervisor, Dr. Victor Odipo, for His insightful feedback and patience throughout this project. I also thank Dr Olukuru for His valuable insights and constructive criticism, greatly enriching this work. I thank Strathmore University and my classmates for their shared experiences, which have enriched my learning experience and made this journey more enjoyable. I am indebted to the scientific community for making invaluable resources and research materials freely available. Lastly, I greatly appreciate my family's unwavering love, support, and understanding throughout this endeavor. Their sacrifices have been my source of strength.

Chapter 1: Introduction

1.1 Background to the Study

The current body of literature generally agrees that housing markets comprise discrete submarkets identified by geographic regions (Gale et al., 2022). These submarkets are geographic areas or neighborhoods with housing units with identical characteristics such as price range and size. For this reason, when examining housing markets, it is advised to divide them into submarkets. This is because of the intricate nature of the housing market, whose influences differ between locations. By breaking down the housing market into manageable segments, it becomes easier to comprehend the composition of the market and analyze its dynamics.

Many stakeholders benefit from having a deep understanding of the housing submarkets. Wu & Sharma (2012) argue that segmenting housing markets into submarkets enables stakeholders and real estate players to tailor their strategies to the particular requirements of each submarket, leading to more effective interventions. Additionally, accurate submarket segmentation improves the predictive power of housing price estimation models (Goodman & Thibodeau, 2007). With information on the extent and characteristics of each submarket, buyers can streamline their home searching process and make well-informed decisions when purchasing a house (Islam & Asami, 2009). Governments also use submarket knowledge to inform their initiatives and develop practical solutions, such as to alleviate the shortage of affordable housing or assess the results of public policy initiatives.

The present literature has a wide range of approaches to submarket delineation, including hierarchical clustering algorithms, as reviewed by Wu and Sharma in 2012. However, many of these methods overlook the location of the houses and thus fail to constrain the clusters to

neighboring features. This oversight leads to poorly defined submarket boundaries that are not spatially contiguous and do not match the urban layout. Chen et al., 2023 highlight that such submarkets may be challenging to interpret and have little real-world application.

Goodman (1981) suggests that submarkets in a housing market should be few, compact, and have clearly defined and contiguous borders. The transition between adjacent submarkets should be seamless and gradual. For example, there should be a smooth transition from large, pricey homes to smaller, more inexpensive ones when moving between two submarkets. This is because factors influencing submarket characteristics, such as zoning laws, are usually limited by geography and exhibit gradual variations.

The purpose of this study is to delineate well-defined submarkets. This study integrates both the structural attributes of the houses and their locations. The SKATER algorithm developed by Assunção et al. 2006, short for Spatial Kluster Analysis by Tree Edge Removal, is used to find geographically contiguous clusters representing submarkets in the housing market.

1.2 Problem Statement

Though there have been significant advancements in methods for delimiting submarkets, many of these methods fail to incorporate spatial constraints in the delineation process. This oversight often leads to disjointed segments that do not represent the cohesive nature of housing markets, especially within metropolitan areas. The absence of spatial constraints can compromise the localization and practicality of the derived submarket boundaries, hindering practical analysis, policy-making, and investment strategies in the real estate sector.

Given that housing markets function in a geographical context, with location influencing parameters such as price, it becomes necessary to employ models and algorithms that enforce spatial constraints during submarket delineation.

1.3 Research Objectives

1.3.1 General Objective

The primary goal of the study is to delineate well-defined residential submarkets.

1.3.2 Specific Objectives

1. To use spatially constrained multivariate clustering to delineate submarkets.

2. To investigate the features of the identified submarkets.

3. To create a map illustrating the geographic boundaries of the submarkets.

1.4 Justification of the Study

Understanding the dynamics of the housing market is essential for various interested parties, such as government agencies and real estate developers, as it is a fundamental component of the economy. This study attempts to broaden our understanding of the housing market and provide valuable insights. Furthermore, current submarket delineation methodologies need more comprehensive approaches, often due to inadequate consideration of all factors or ineffective incorporation of spatial constraints. This study aims to close these gaps and promote the creation of more effective techniques.

1.5 Scope and Limitations

The study will leverage Geographic Information System (GIS) technology and spatial statistical methodologies to provide valuable insights into the spatial segmentation of urban

housing markets. Although the primary dataset will focus on King County, the employed methodologies are expected to be adaptable and applicable to housing markets in other regions, given the availability of data.

This dataset contains information on approximately 21,000 houses, with 21variables including the location, price, size, grade and condition of the houses compiled from May 2014 to May 2015. It is important to note that local data lacks organization and is characterized by a limited number of attributes. This project is essential in bringing attention to the importance of having detailed and well-structured data for academic research purposes and for making well-informed decisions within the housing market.

The remaining sections are arranged as follows: Chapter 2 reviews previous studies on housing submarket segmentation, Chapter 3 outlines the methodology employed in this study, and Chapter 4 presents the analysis results.

Chapter 2: Literature Review

2.1 Introduction

This literature provides an overview of the methodologies and algorithms developed thus far for delineating submarkets, highlighting their capabilities and limitations. Several techniques for delineating housing submarkets have been implemented, discussed, and improved. However, as Keskin and Watkins (2016) point out, no universally accepted method exists.

The first studies on housing submarkets were conducted in the 1950s, but scientific research on this subject picked up in the early 2000s. According to Wu and Sharma (2012), these methodologies fall into two primary groups: data-driven and a priori. As Chen et al. (2023) highlight, a critical difference between these approaches is how they incorporate spatial constraints.

2.1 Delineating housing submarkets through A priori methods

A priori approaches rely on the experience and understanding of real estate agents and other industry experts to define the extent of submarkets. While a priori methods may sometimes employ statistical tests, these experts often use readily available boundaries when delineating submarkets. These borders may include government boundaries, census blocks, or natural features such as rivers (Bourassa et al., 2003; Goodman & Thibodeau, 2003). Though a priori methods offer defined submarkets, these submarkets can fail to capture the true complexities of the housing market. Chen et al., 2023 indicate that a priori methods often rely on a narrow set of housing characteristics, which can lead to an incomplete understanding of the housing market. Experts' insights are valuable, but their subjectivity and potential inconsistencies can limit the accuracy of these delineations.

2.2 Delineating housing submarkets through data-driven methodologies

Data-driven techniques rely on data analysis and can consider various factors, such as location, demographics, and housing structure. For instance, Bourassa et al. (1999) leveraged a two-step approach, using principal component analysis followed by K-means cluster analysis, to identify aspatial submarkets based on comparable structural elements and neighborhood traits. Other data-driven strategies commonly used for submarket demarcation include factor analysis, which can identify underlying latent factors influencing housing characteristics, and partitioning algorithms, which, like K-means, group similar data points together. Furthermore, hierarchical clustering, explored by Goodman and Thibodeau (2003), offers a different approach where data points are grouped in a nested structure based on their similarity. However, a fundamental limitation of many of these algorithms is that they do not explicitly consider the geographical context of the houses, potentially grouping houses in geographically separated areas together if they share similar characteristics (Liu et al., 2022).

According to Bourassa et al. (2003), "Not only do submarkets matter, but geography is what makes them matter" (p. 27). Supporting this notion, Wu and Sharma (2012) argue that housing units with similar characteristics are more likely to be concentrated in specific geographic areas. This is because houses in the same neighborhood often share access to similar amenities and local resources. Tobler's first law of geography states, "Everything is related to everything else, but near things are more related than distant things" (Waters, 2018, p. 1). Following this, houses close to each other should have comparable qualities, and the likelihood of distinction should increase as the distance between houses increases. Therefore, it is imperative to utilize algorithms that impose spatial constraints limiting clusters to adjacent or nearby properties.

2.3 Delineating housing submarkets through spatially constrained data-driven approaches

Identifying regional patterns within massive real-world datasets is a growing field thanks to advancements in computing power. This has spurred the development of various algorithms for regionalization. Regionalization involves segmenting data into distinct, neighboring regions that share internal similarities while contrasting with other. Despite notable progress, research comparing the effectiveness of various regionalization methods still needs to be expanded (Helbich et al., 2013). Researchers have also developed algorithms incorporating spatial constraints into the submarket delineation process. These constraints can be soft or hard (Chen et al., 2023).

Leveraging spatial clustering, Wu et al. (2018) utilize the Density-Based Spatial Clustering algorithm to determine submarkets in Shenzhen, China. However, their findings demonstrate that the algorithm produces fragmented submarkets even when spatial constraints are included. To identify submarkets within the housing market of Franklin County, Ohio, Chen et al. (2023) leverage two essential methods: ClustGeo and REDCAP (Regionalization with Dynamically Constrained Agglomerative Clustering and Partitioning). These techniques allow them to capture the spatial extents of submarkets and how they evolve and change over time.

Assunção et al. (2006) developed the SKATER algorithm which has since been employed in various studies. For example, in exploring urban housing submarkets within Shenyang, China, Liu et al. (2022) employed the SKATER algorithm for submarket delineation. Their findings revealed a considerable house price disparity in Shenyang, with higher housing prices in the city center and showing variations between the north and south.

The idea that submarkets should have spatially continuous boundaries implies that continuous representations better explain reality than rigid, sharply cut boundaries (Helbich et

al., 2013). According to Barrett and Alan (2022), the SKATER algorithm is particularly efficient when delineating geographically compact submarkets with contiguous borders, especially when dealing large datasets. SKATER naturally encourages smooth boundaries that have little overlap. This contrasts other methods, like agglomerative clustering, which might need help to achieve such clear boundaries. For this study, SKATER emerges as the best tool for submarket identification.

SKATER is fundamentally an unsupervised technique that utilizes graph partitioning. In order to guarantee that cluster membership is limited to contiguous or nearby features, the SKATER approach begins with a connectivity graph, more precisely, a minimal spanning tree. It transforms the data points representing housing units into nodes connected by weighted edges, representing the similarity of the features (such as size) between the connected units. The algorithm then splits the most appropriate edge to partition the tree into separate clusters representing submarkets. This decision is guided by two fundamental principles: minimizing dissimilarity within each submarket (increasing internal homogeneity) and, whenever possible, avoiding the creation of clusters containing only a single unit (singletons). This process of splitting continues until the desired number of submarkets is achieved.

While the ideal number of submarkets might be known in some cases, it is often not readily apparent. In these situations, SKATER relies on a strategy to select the number of submarket clusters that best balance internal similarities and external differences. To achieve this, a metric called the Calinski-Harabasz pseudo-F-statistic is employed. This statistic compares the variation within and between submarkets. A higher F-statistic signifies a submarket partition where clusters are more alike internally and distinct.

According to Islam and Asami (2009), the approaches used in submarket delineation are determined by the intended usage. The rationale behind employing this methodology is to establish well-defined spatially bounded submarkets. By integrating spatial constraints with housing attributes, this approach ensures that the derived submarkets capture the spatial organization of the housing market. This, in turn, facilitates practical spatial analysis and informed decision-making processes. In practice, location and house features influence buyer's housing preferences, and both must be considered when defining submarkets (Watkins, 2001). This approach is readily interpretable, can be replicated and is graphically presentable, which encourages confidence and acceptance.

2.4 Research Gap

Spatially constrained data-driven methodologies must be researched and implemented to delineate housing submarkets effectively. Conventional techniques fail to account for the housing units' location and the spatial relationships inherent in the housing market. This research gap must be addressed to improve submarket reliability and relevance in urban planning, decision making and policy-making scenarios.

Chapter 3: Methodology

3.1 Research Design

This study adopts the Cross Industry Standard Process for Data Mining (CRISP-DM) model to ensure a structured and rigorous approach. The following sections outline how each of the six sequential phases of the CRISP-DM model is applied in this study.

3.1.1 Business Understanding

In this initial phase, the research goals and the problems to be addressed are thoroughly understood and defined.

3.1.2 Data Understanding

This phase delves into the collected data to comprehensively assess its quality. Activities include identifying missing values and outliers, visualizing the data to uncover patterns and trends, and summarizing its essential characteristics through data profiling.

3.1.3 Data Preparation

During this phase, the data is prepared for analysis. Key tasks include feature selection, handling missing values, addressing outliers, normalizing the data, and merging relevant datasets as needed.

3.1.4 Modelling

This stage focuses on extracting valuable knowledge, patterns, and insights from the prepared data using appropriate data mining models.

3.1.5 Evaluation

In this phase, the chosen model's accuracy, generalizability, and interpretability is assessed. Evaluation is conducted using relevant metrics.

3.1.6 Deployment

The final stage involves integrating the chosen model into real-world applications. This integration enables the generation of actionable insights that can directly inform strategic decision-making processes.

Figure 3.1 illustrates how the six phases of the CRISP-DM is applied in this study.

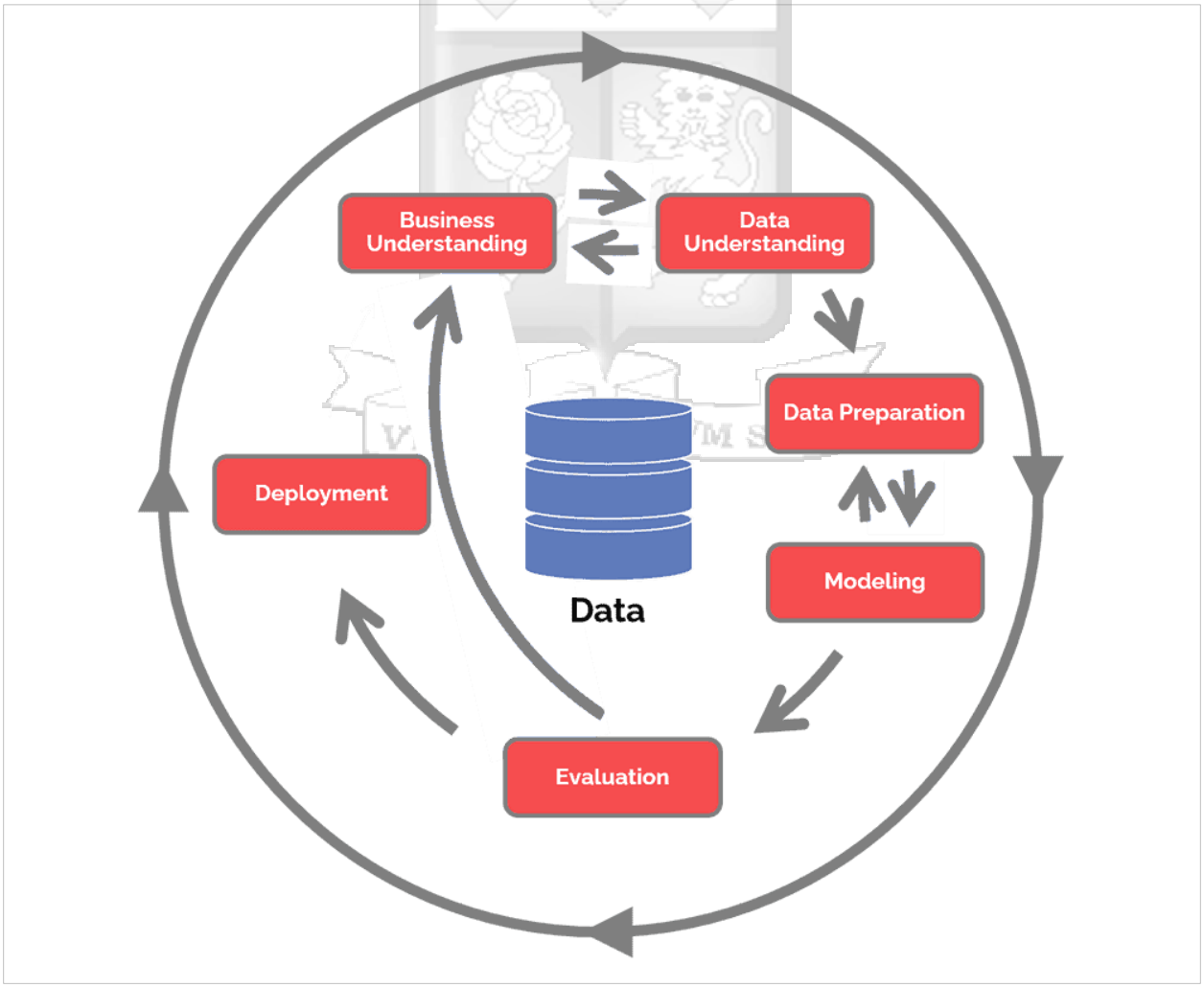


Figure 3.1 An outline of phases in CRISP-DM by Hotz (2018)

3.2 Data

According to the results of the 2020 census, King County was identified as the most populous county in Washington state, with Seattle, the state's largest city, situated within its boundaries. The suburban areas surrounding Seattle are where most of King County's inhabitants reside, while woodlands and farmlands characterize the county's eastern region. The Center for Spatial Data Science, based at the University of Chicago, compiled the dataset used in this analysis, which consists of 21 variables detailed in Table 3.1.

Table 3.1 Description of input variables

No.	**Variable Name**	**Description**
1	Id	A unique identifier
2	Date	Date of sale of house
3	Price	Price of the house
4	Bedrooms	Bedroom count
5	Bathrooms	Bathroom count
6	Sqft_Living	Size of the living spaces in square feet
7	Sqft_Lot	The size of the land on which the house is situated in square feet
8	Floors	Floor count
9	Waterfront	Does the property have a waterfront
10	View	Categorical variable for property view quality
11	Condition	House condition rating (1-5)
12	Grade	The assigned grade ranges from 1 to 13.
13	Sqft_Above	Total above-ground living area (square footage)
14	Sqft_Basement	Below-ground-level living area (square footage)
15	Yr_Built	Year built
16	Yr_Renovated	Last renovation year
17	Zip code	The zip code where the house is located
18	Lat	y coordinate of the house
19	Long	x coordinate
20	Sqft_Living15	average living area of 15 nearest neighbors (sq ft)

21	Sqft_Lot15	Average lot size of 15 nearest neighbors (sq ft)

3.3 Data Processing Techniques

Data preprocessing is a crucial initial step, laying the groundwork for analysis. This ensures the data is clean, consistent, and suitable for modeling. The data preprocessing techniques employed in this research address outliers, select the most pertinent features for analysis and scale the data. These processes are explained in detail below.

3.3.1 Managing Outliers

Outliers are data points that significantly deviate from the expected pattern. They exhibit values that fall well outside the typical range observed in the dataset, often much higher or lower than the average values. Research by Nowak-Brzezińska and Gaibei (2022) highlights the negative impact of outliers on clustering quality. Their findings emphasize the importance of proactively identifying and removing outliers, especially when dealing with large, real-world datasets. Outliers can hinder the formation of well-defined clusters and make it more challenging to explore the data effectively. Visual inspection is a well-established technique for outlier detection, as demonstrated by Dastjerdy et al. (2023). This method is used in this study and involves plotting the data using boxplots to reveal potential outliers visually across each variable.

3.3.2 Feature Selection

Feature selection involves choosing a subset of the most informative features from the initial dataset. These features are the ones that best differentiate between clusters, allowing the clustering algorithm to group similar data points effectively. When the number of features becomes excessive, it becomes more challenging to identify meaningful relationships between the data points, and the model can overfit because of multicollinearity. This phenomenon is

known as the curse of dimensionality. Correlation analysis is used in this study to combat the curse of dimensionality. This involves calculating a correlation matrix, which reveals the relationships between each pair of features in the dataset, and then removing one feature if the two are strongly correlated.

3.3.3 Feature Scaling

Feature scaling is a crucial preprocessing step that ensures all numerical features are transformed to a common scale. This is particularly important when dealing with features with naturally large value variations. Z-score normalization is a common technique used for feature scaling and is employed in this study. This technique standardizes the features by subtracting the mean value (average) from each data point and dividing the result by the standard deviation of all the values in that feature.

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

3.4 Clustering

This study will use the SKATER algorithm for clustering purposes. The first step involves constructing a dissimilarity matrix (i, j) representing the dissimilarity (often measured by Euclidean distance) between two data points, i and j, in the spatial dataset. SKATER only considers distances between neighboring observations, meaning those considered contiguous in space. This creates a sparse dissimilarity matrix, focusing on the relationships between spatially connected data points. The dissimilarity matrix is then transformed into a graph structure. The SKATER algorithm utilizes the graph to construct a minimum spanning tree that connects all the nodes (observations) in the graph with the minimal total edge weight (dissimilarity). This creates

an initial clustering solution where all observations are linked in a network, with the total edge weight representing the initial sum of squared deviations (SSD).

$$\sum_i (x_i - \overline{x})^2$$

The objective of SKATER lies in minimizing the overall SSD. SKATER achieves this by strategically cutting edges within the minimum spanning tree in an optimal way to partition the observations into distinct submarkets that exhibit high internal similarity and low inter-cluster dissimilarity.

$$SSD_T - (SSD_i + SSD_j)$$

SSDi and SSDj are the contributions of each subtree. To identify the most effective cut, SKATER employs a strategy that evaluates the impact of each potential cut on the overall SSD. It prioritizes cuts that maximize the difference above. This process continues iteratively until the desired number of submarkets is achieved.

3.5 Performance Evaluation Criteria

To assess the quality of the submarkets identified by SKATER, the Calinski-Harabasz pseudo-F-statistic will provide a quantitative measure of how well-separated the clusters are. In essence, it evaluates the balance between within-cluster variance and between-cluster variance. Within-cluster variance refers to how similar the data points are within each submarket, whereas between-cluster variance refers to how different the data points are between different submarkets.

$$F = \frac{\frac{R^2}{n_c} - 1}{(1 - \frac{R^2}{n} - n_c)}$$

where R^2 represents the R-squared value, a statistical measure of how well the cluster center explains the variations within a cluster, n denotes the total number of entities (data points) in the dataset and nc signifies the number of clusters (submarkets) identified by SKATER (Liu et al., 2022).

Figure 3.2 illustrates the workflow and the various stages of the methodology.

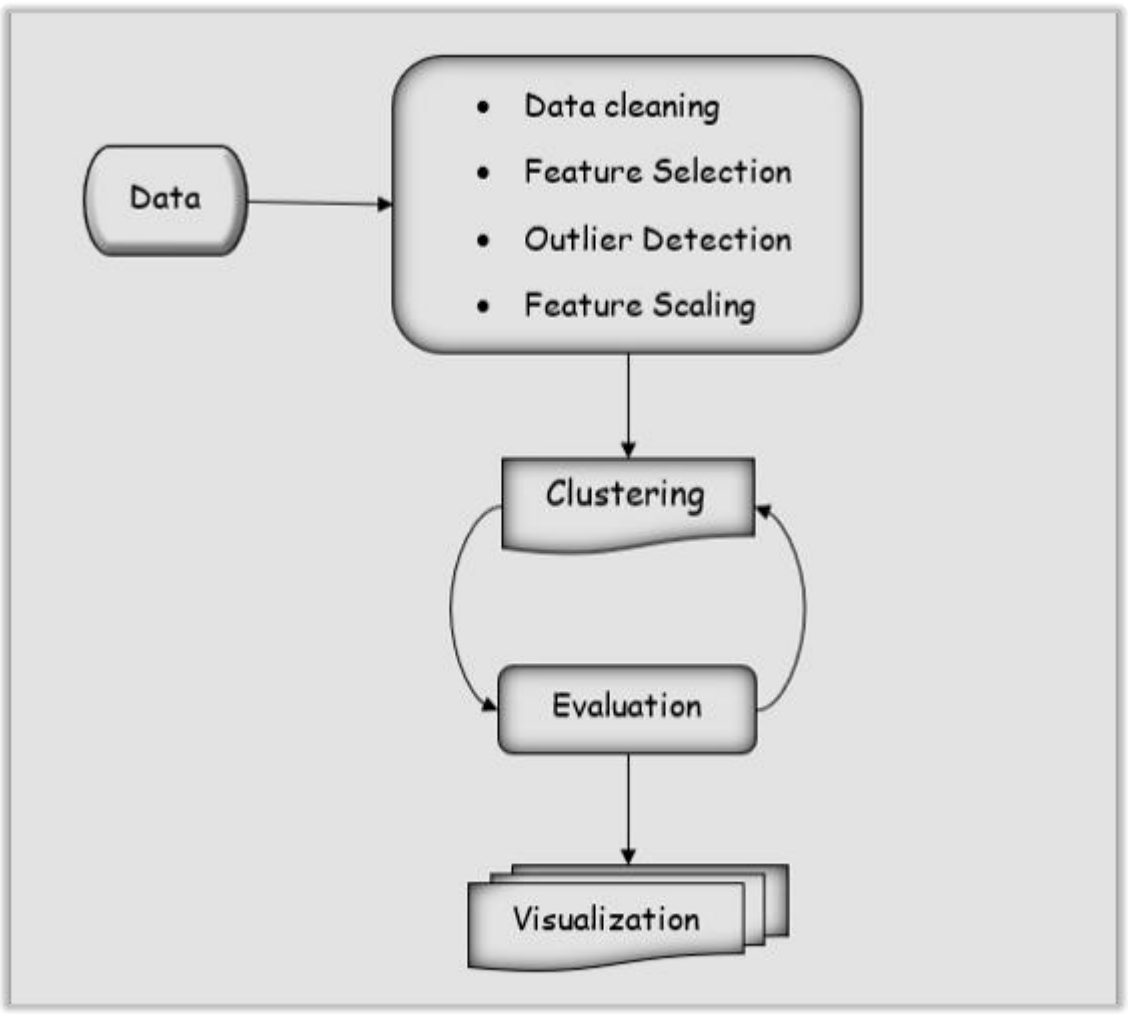


Figure 3.2 Workflow chart

Chapter 4: Results

4.1 Features and data preprocessing

The 'price', house 'living area' and house 'age' variables have been recognized in previous literature as pivotal factors for delineating submarkets, and these variables were employed in this study. Additional variables were incorporated to ensure a comprehensive exploration and to derive interpretable and meaningful submarkets, as demonstrated by (Chen et al., 2023).

There were no missing values across any of the features. The variable 'yr_renovated' was replaced with the temporal difference representing the difference in years between the renovation year and the year of sale. The 'id' column, used as an identifier, was removed due to its minimal impact on the analysis. Similarly, 'zipcode' was considered redundant given the presence of x and y coordinates. Additionally, 'sqft_above' and 'sqft_basement' were omitted as their details are encompassed by the 'sqft_living' feature, representing their summation.

All instances of duplicate houses resulting from multiple sales between 2014 and 2015 were removed. Houses with 0 bathrooms were removed, as residential properties are typically expected to include bathroom facilities, especially within the scope of this study. Additionally, a house with 33 bedrooms but only 1.75 bathrooms in a 1620-square-foot space was excluded due to concerns about data accuracy. The cleaned and tidied dataset consisted of 21,425 observations.

4.2 Feature selection

According to the correlation matrix (Figure 4.1), 'sqft_living' exhibits a high correlation with 'bathrooms' (0.75) and 'bedrooms' (0.65). The high correlation could be because larger living spaces often have more bathrooms and bedrooms. Accordingly, the features 'bathrooms' and 'bedrooms' were removed to prevent multicollinearity. Furthermore, 'sqft_lot15' shows a

significant correlation with 'sqft_lot' and was, therefore, also excluded. Similarly, 'sqft_living15' displayed a high correlation with the 'grade' variable and was also dropped.

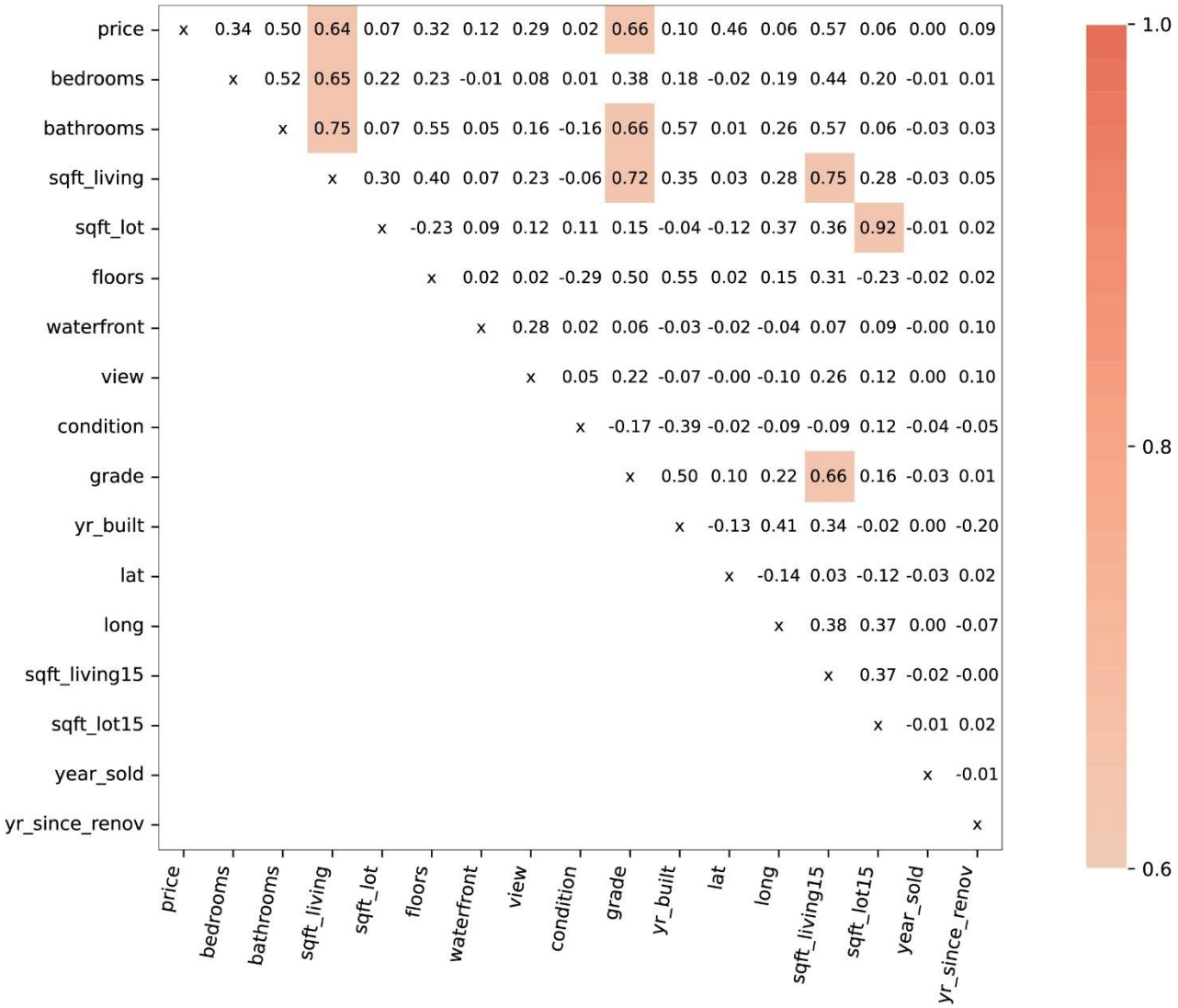


Figure 4.1 Correlation matrix of the variables

4.3 Outlier treatment

Outliers were present in the 'price', 'sqft_lot' and 'sqft_living' variables. Figure 4.2 illustrates how outliers were examined against the 'price' variable using both a violin plot and a box plot. As per Table 4.1, upon comparing the 75th percentile price value of $645,000 with the

maximum value of $7,700,000, it became evident that the 'price' feature contained outliers. A more in-depth analysis of the 'price' feature distribution revealed a right-skewed distribution with a heavy tail, confirming the presence of outliers (refer to Figure 4.3).

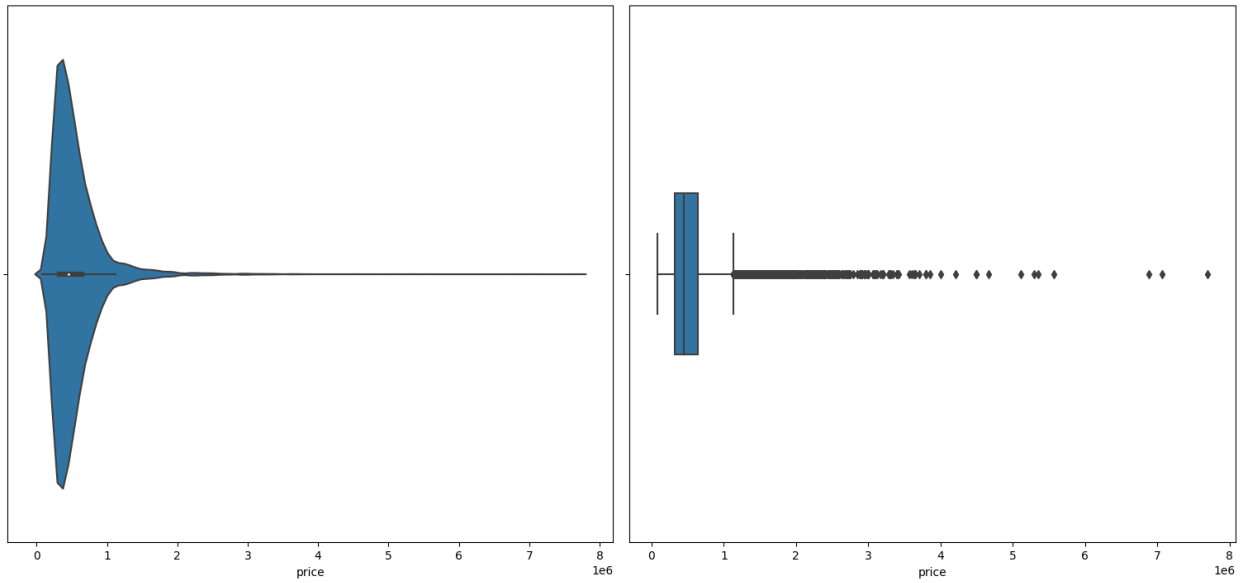


Figure 4.2 Existence of outliers in the price variable

Table 4.1 Summary statistics of the price variable

mean	540,088
std	367,127
min	75,000
Q1	321,950
Q2	450,000
Q3	645,000
max	7,700,000

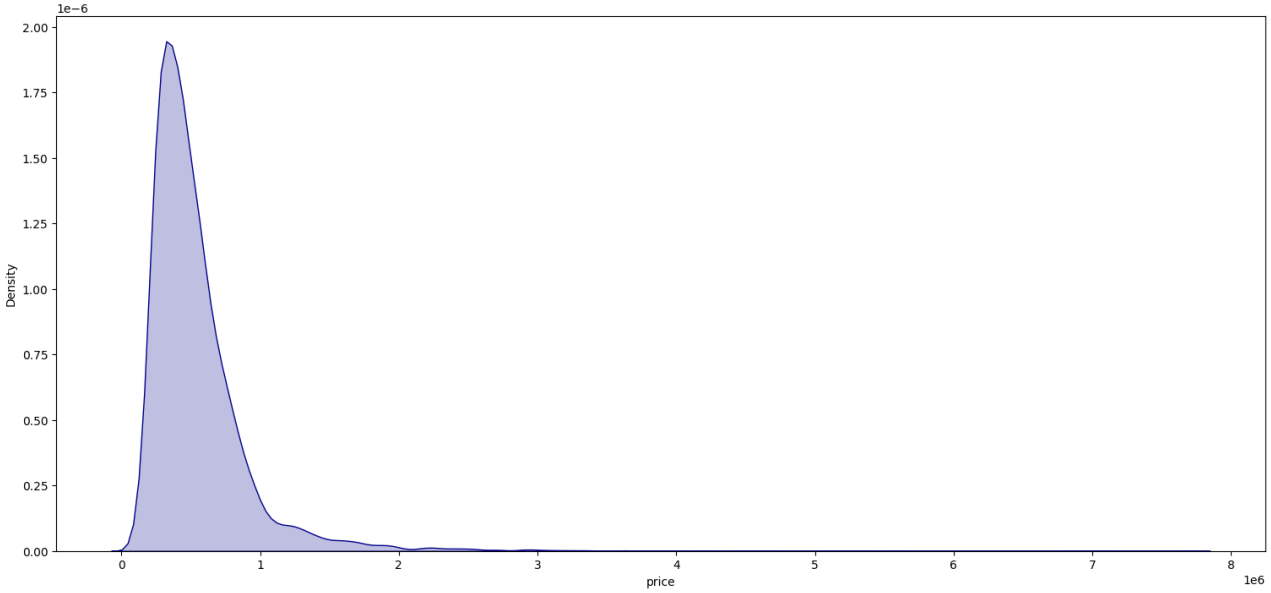


Figure 4.3 Distribution of the price feature

Rather than eliminating outliers from the dataset, a method involving flooring and capping was employed. This approach, as utilized by Wang and Zhao (2022), entails replacing outliers with computed floor and cap values. Capping involves setting a limit for a feature and assigning the value of this limit to any outliers surpassing it. Specifically, values more significant than the 95th percentile number were substituted for the' price' variable with the 95th percentile value. This identical approach was extended to the 'sqft_lot' and 'sqft_living' features. Figure 4.4 illustrates the distribution of the 'price' feature following the treatment of outliers.

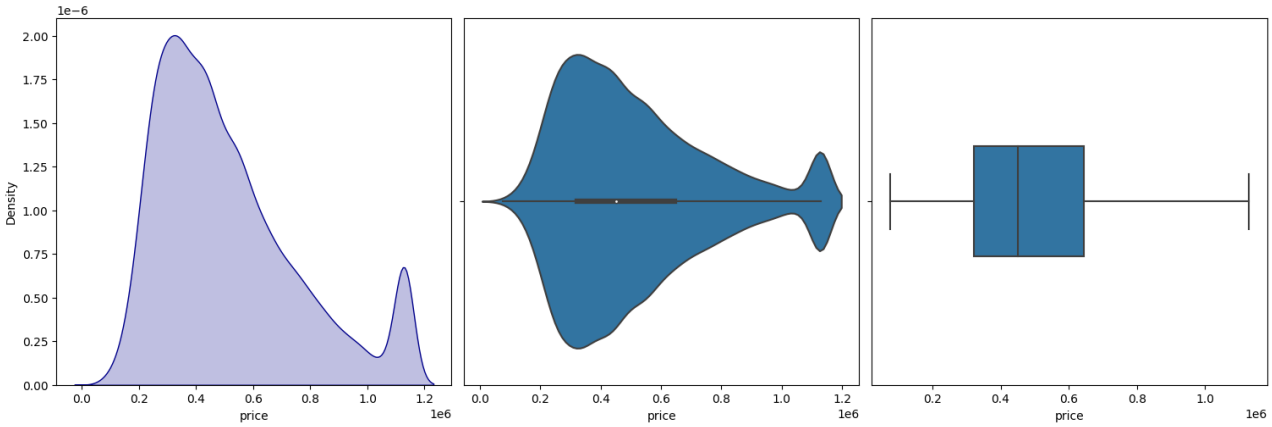


Figure 4.4 Price distribution after capping

4.4 Exploratory Data Analysis

This research employed exploratory data analysis to uncover the underlying insights within the raw data. The feature distribution highlights that single-floor houses were the predominant type (Figure 4.5), closely followed by two-floor houses. Most houses lacked a waterfront (Figure 4.6) and exhibited a less desirable property view (Figure 4.7).

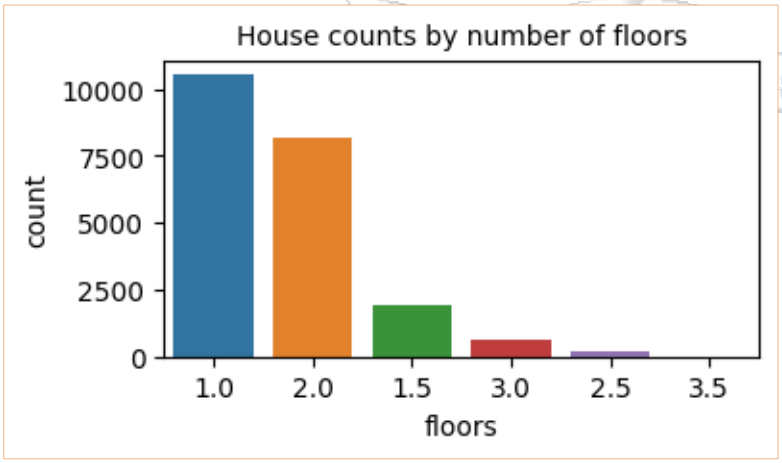


Figure 4.5 House counts based on the number of floors

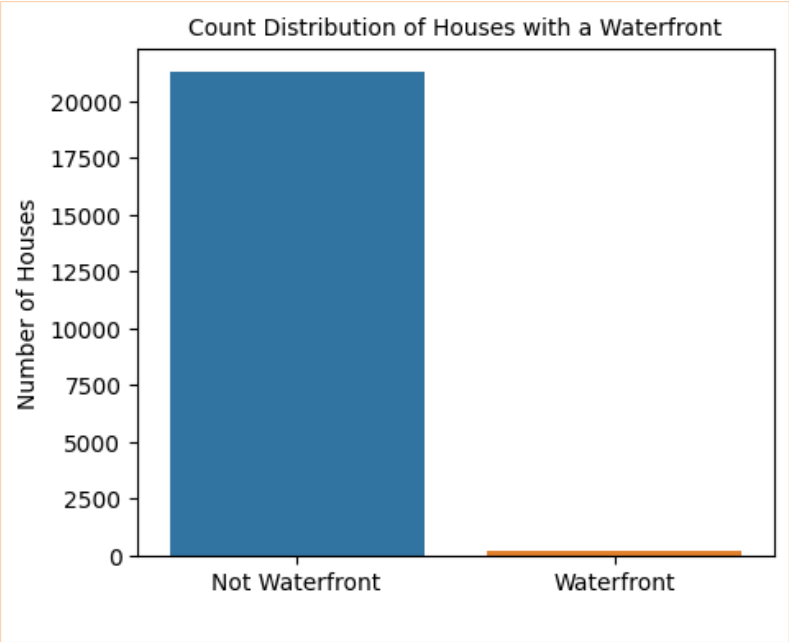


Figure 4.6 Number of houses with a waterfront

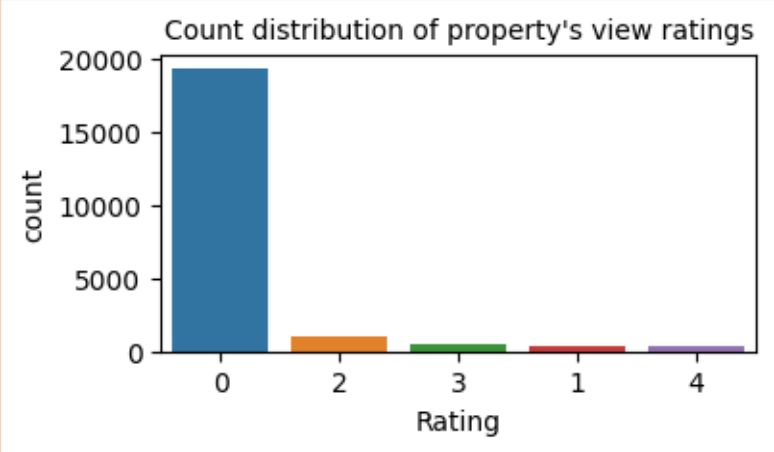


Figure 4.7 Count distribution of property's view ratings

Most houses were rated grade 7 and 8, as shown in Figure 4.8, indicating an average level of construction and design. Moreover, many of these houses were assigned a condition rating of 3 (Figure 4.9), suggesting that the houses in the dataset were of a mid-level classification.

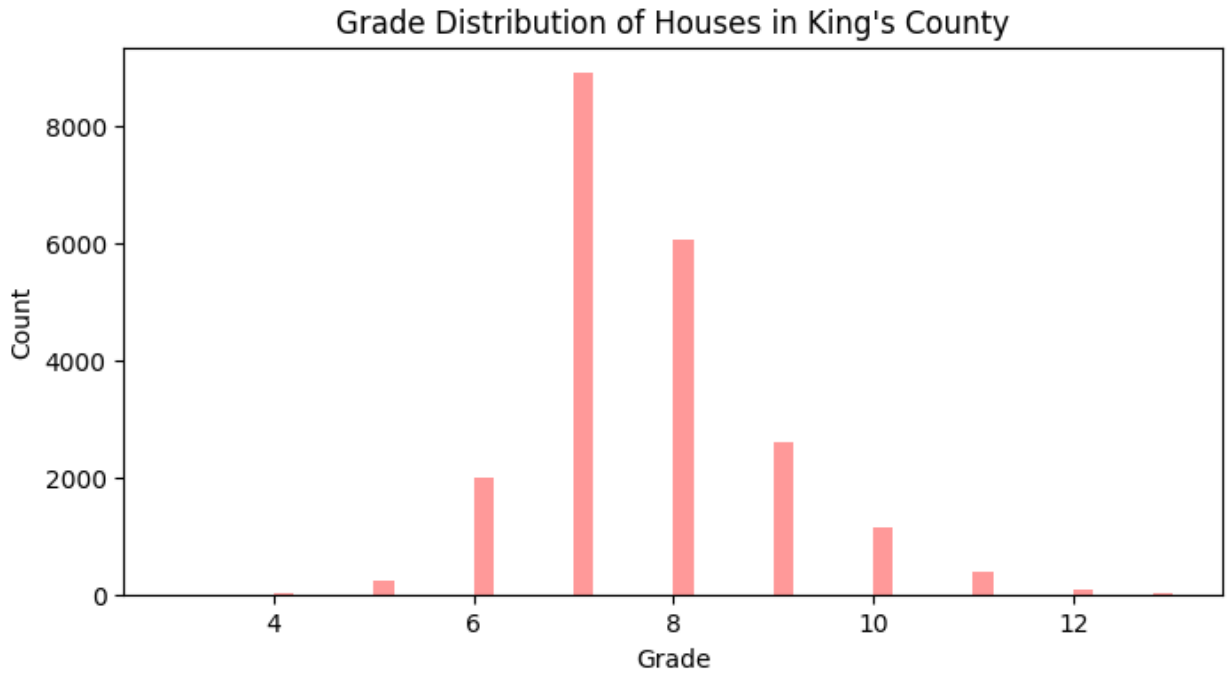


Figure 4.8 Count distribution of house grades.

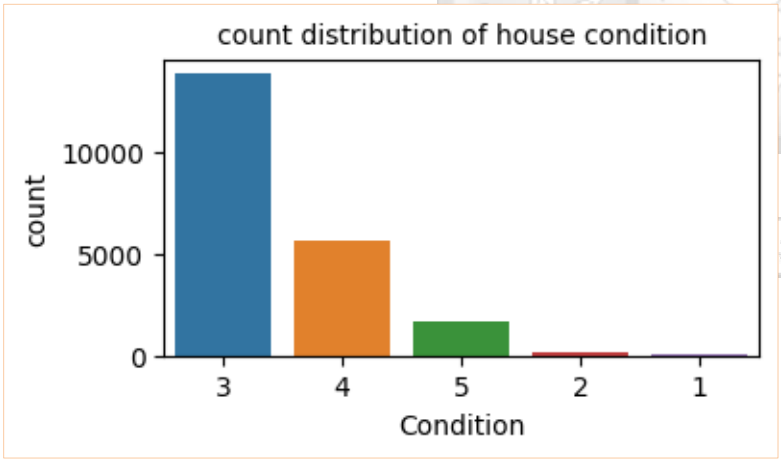


Figure 4.9 Count distribution of house conditions.

Examining the Figure 4.10, it was evident that renovations were relative. Most houses appear to fall into the 'Not Renovated' category, suggesting a limited occurrence of renovation activities.

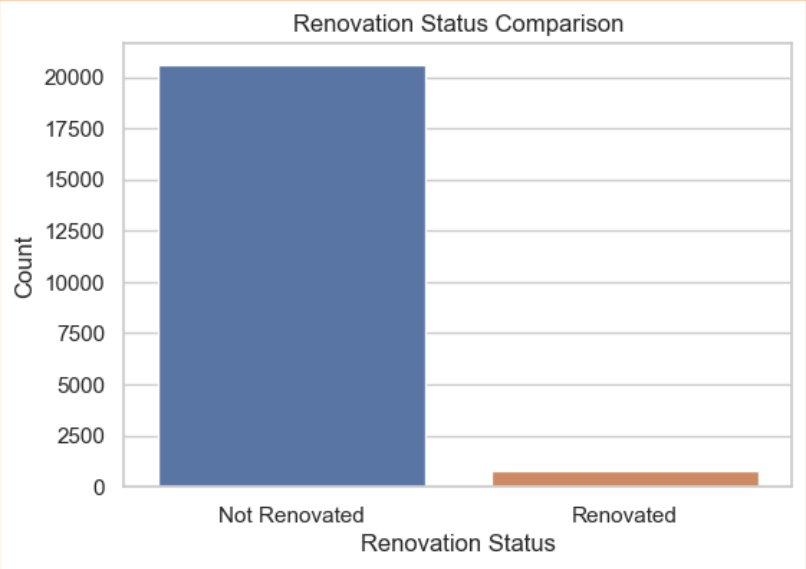


Figure 4.10 Count distribution of renovated houses.

The map in Figure 4.11 illustrates the distribution of houses in King County, with a predominant concentration of data observed in the western region. Conversely, data are scarce in the eastern cities such as Snoqualmie. This discrepancy arises due to the extensive farmland and forest coverage characterizing the majority of the eastern areas in King County.

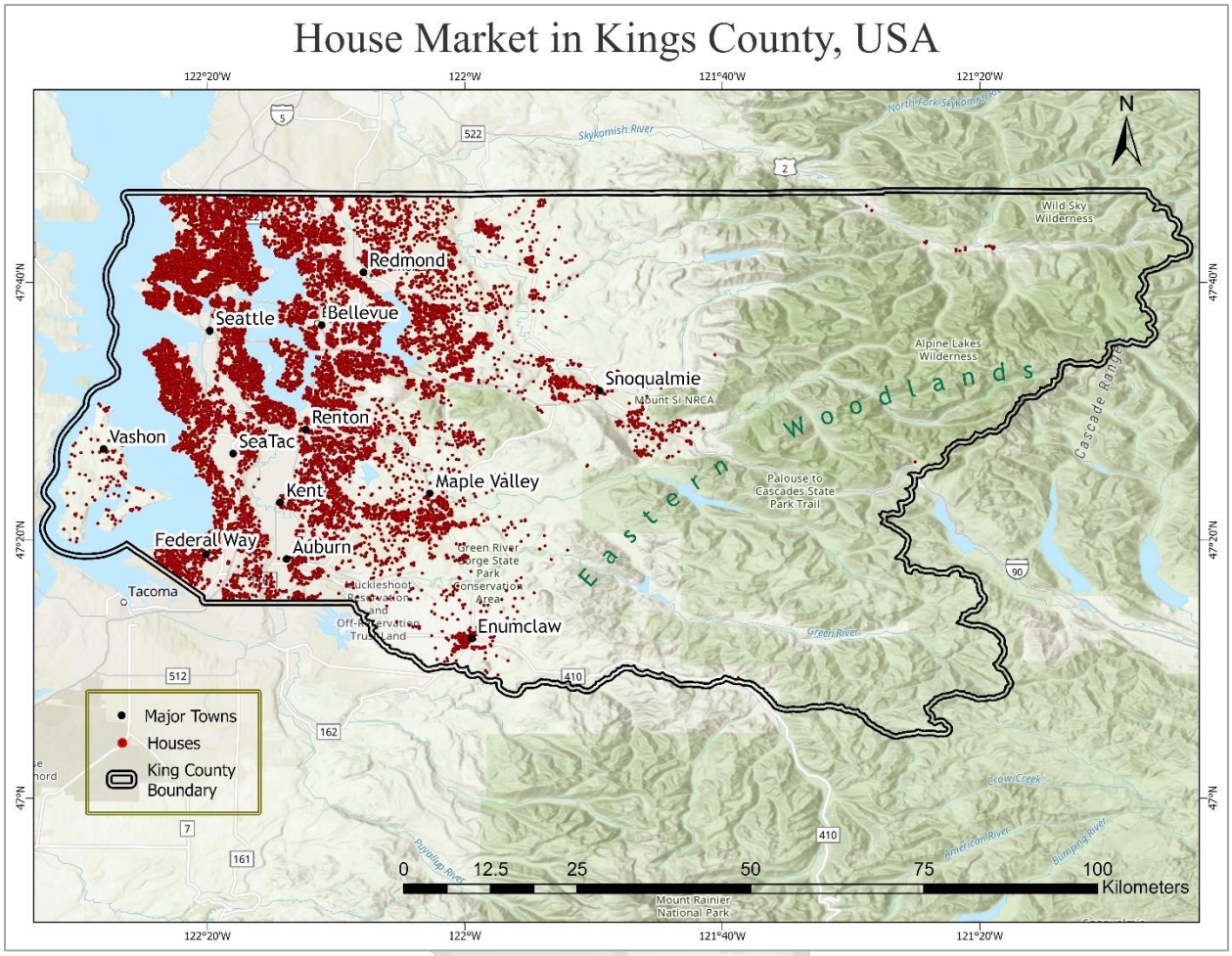


Figure 4.11 House market in Kings County

In the exploratory data analysis (EDA) phase, it became apparent that most houses share a condition rating of 3, have 1 or 2 floors, and only a tiny fraction underwent renovation. Recognizing the limited discriminative power of these specific columns for clustering purposes, condition, floors and yr_renovated were deliberately omitted from the dataset as their inclusion would not significantly contribute to the effectiveness of the clustering analysis. In this study, critical variables identified to influence housing market dynamics include yr_built, sqft_lot, sqft_living, price, and house grade, which were adopted in the delineation process.

4.5 Modelling

The Spatially Constrained Multivariate Clustering tool within ArcGIS Pro's Spatial Statistics toolbox was utilized to demarcate homogeneous and spatially contiguous submarkets in Kings County. ArcGIS Pro, developed by Esri, is a commercially available desktop geographic information system (GIS) software application. It allows users to create, analyze, visualize, and share geospatial data. The Spatial Statistics toolbox in ArcGIS Pro offers specialized statistical tools tailored for analyzing spatial distributions and relationships within geographic data. Unlike traditional statistical methods, spatial statistics directly integrate spatial relationships such as proximity and connectivity into their analysis, providing insights into spatially dependent phenomena. Figures 4.12 and 4.13 below show a screenshot of the tool and the requisite input parameters.

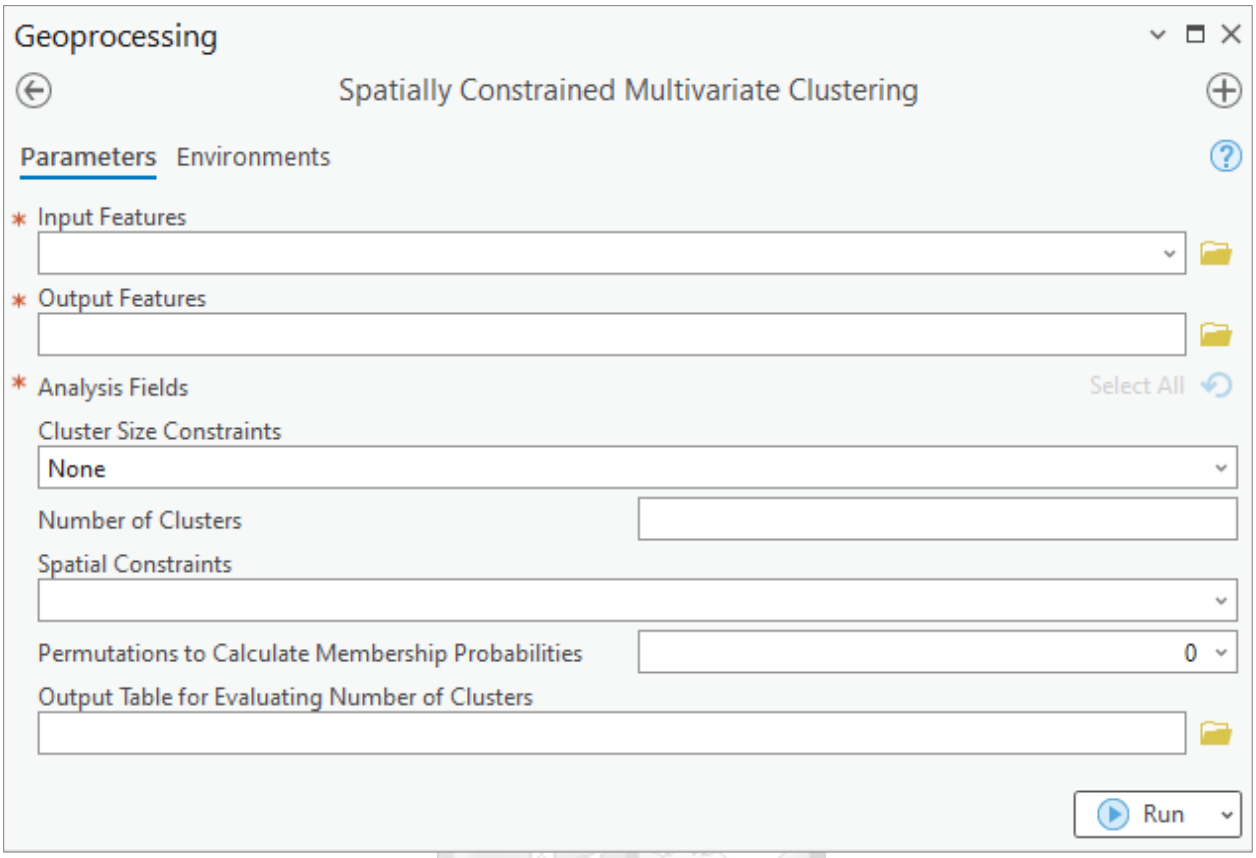


Figure 4.12 Screenshot of input parameters for the tool

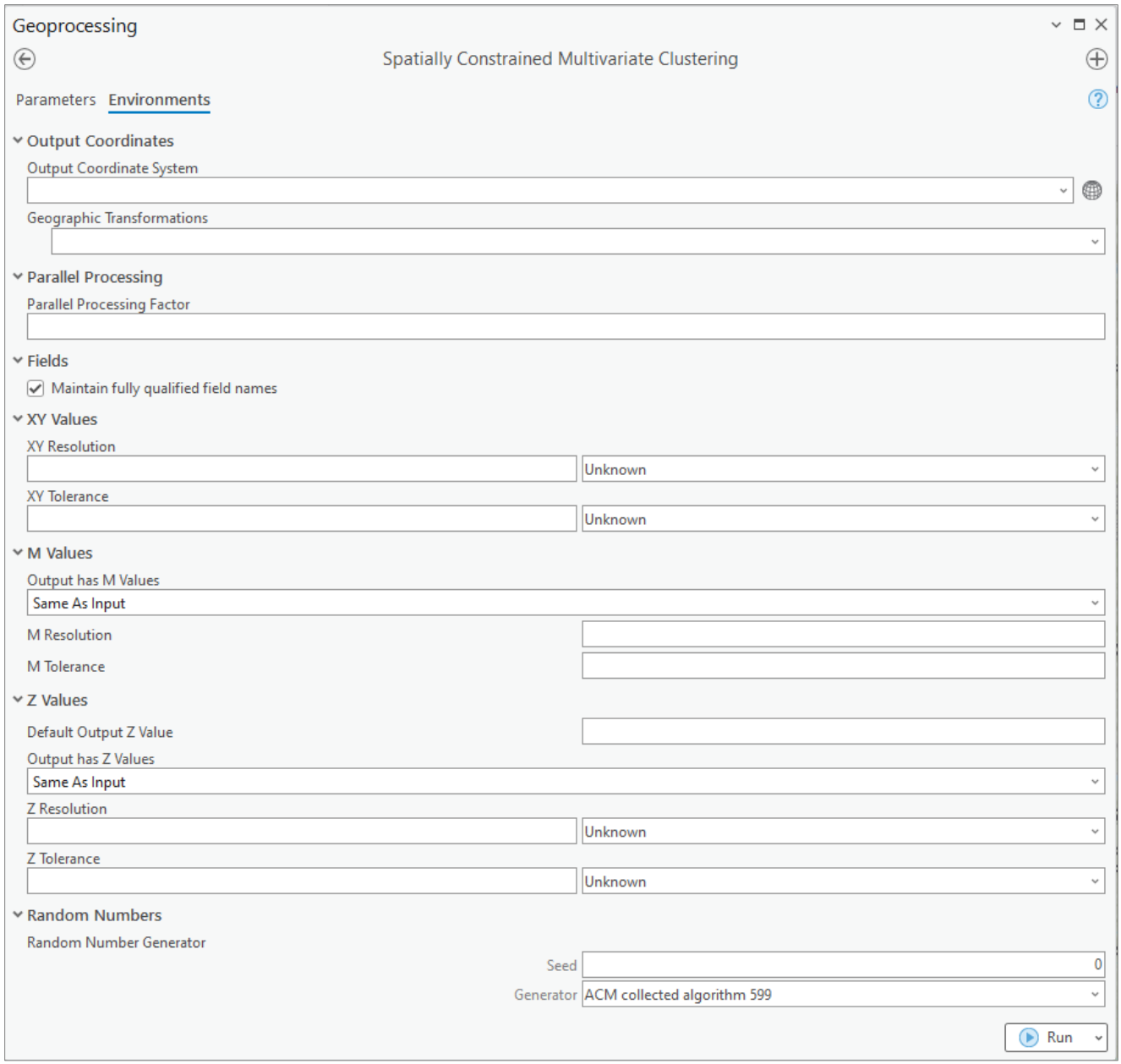


Figure 4.13 Screenshot of environment parameters for the tool

Four submarkets were derived, each portrayed in a different color, as shown in Figure 4.14.

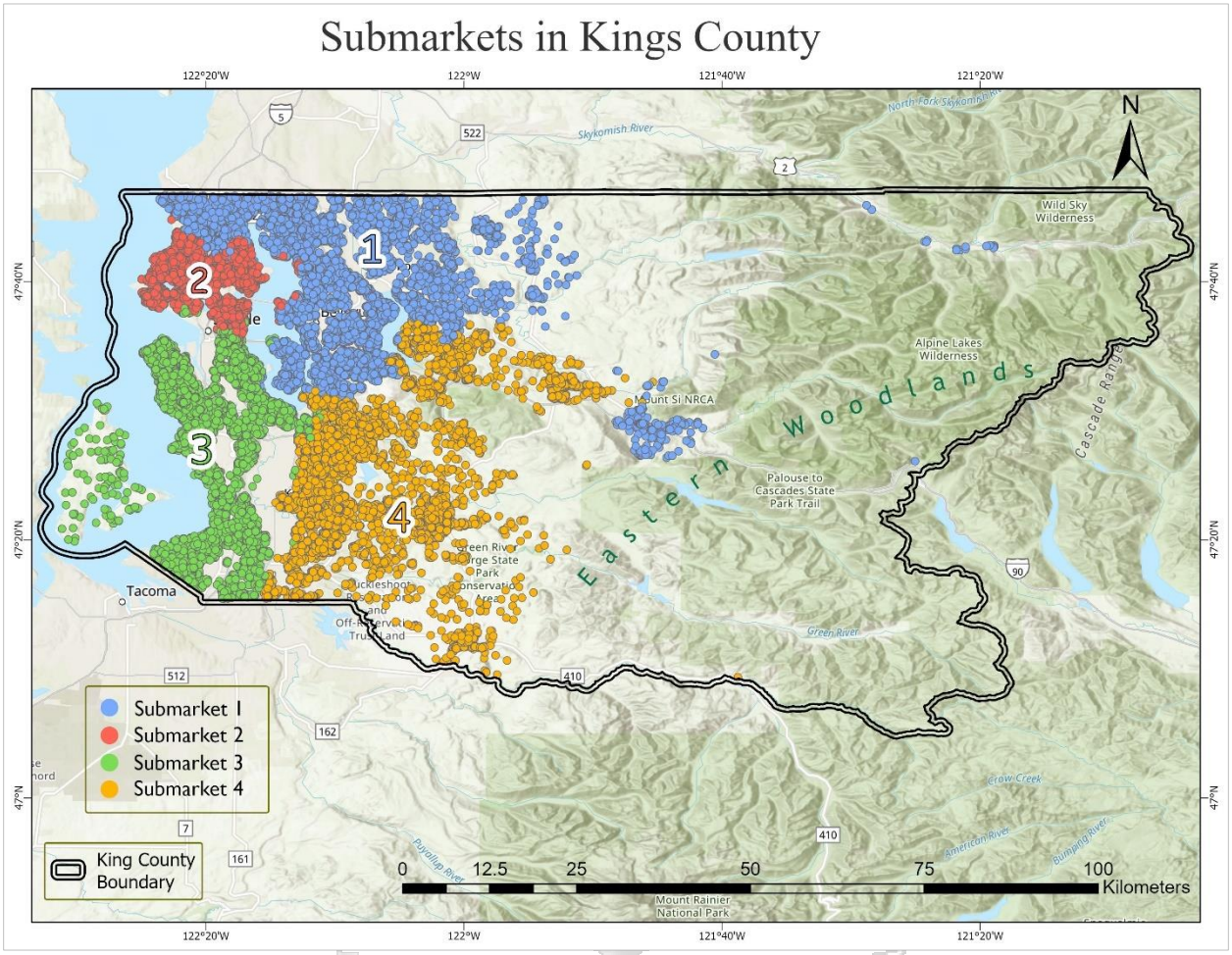


Figure 4.14 Submarkets in Kings County

Hard spatial constraints were applied, grouping only houses with high similarity in non-spatial attributes and sharing common spatial extents into a single submarket. Conversely, soft spatial constraints were applied, allowing observations with high similarity in non-spatial attributes to be grouped into one submarket even if they were not spatially adjacent. This explains why submarket 1 consisted of three non-contiguous spatial segments. The optimal number of submarkets was determined using the Pseudo-F-statistic. The Pseudo F-statistic is a ratio that assesses the similarity within and the difference between groups, providing insight into

how well-defined and distinct the submarkets are. Figure 4.15 below illustrates how the pseudo-F statistic varied when the number of clusters varied.

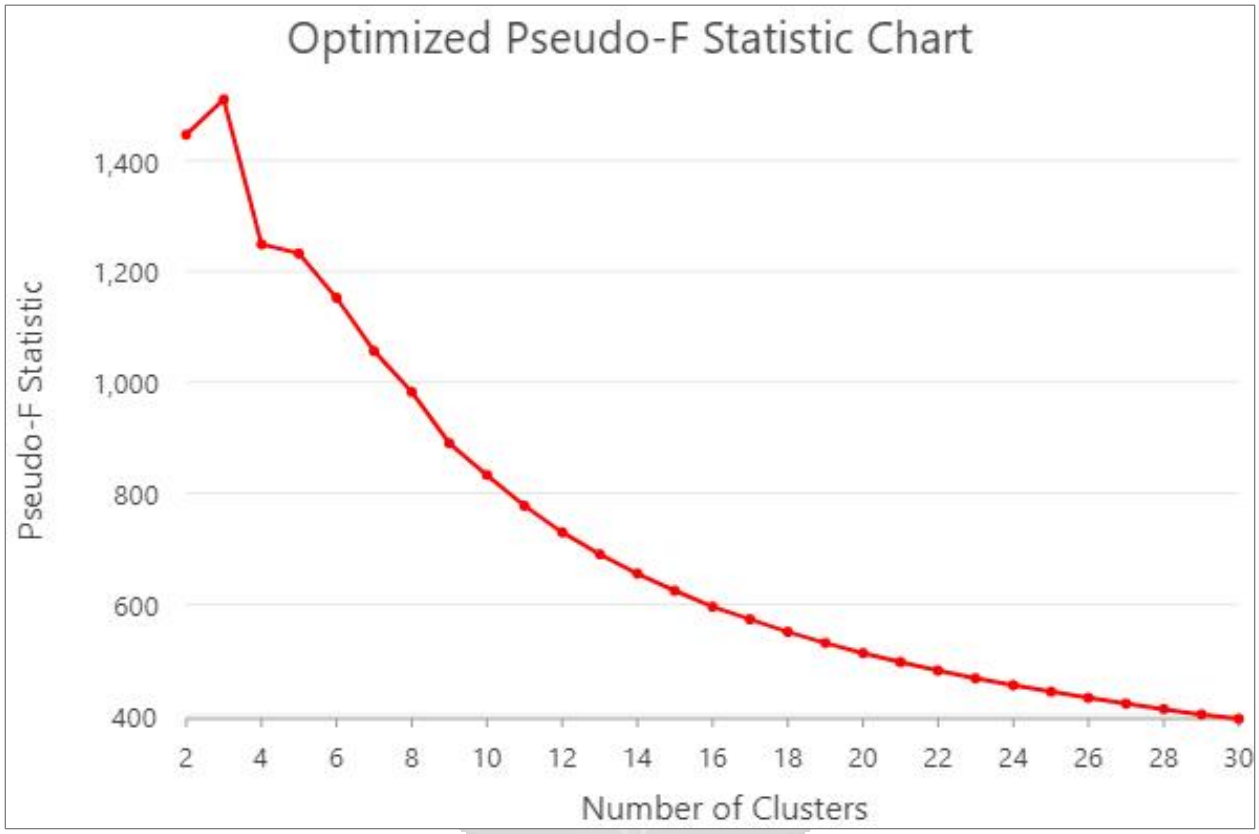


Figure 4.15 Variations in the Pseudo-F statistic across different cluster numbers

According to the chart, an increase in clusters led to a decrease in the Pseudo F-statistic, indicating that the clusters were becoming less distinct. The pseudo-F-statistic reached its peak at cluster 3. Nevertheless, a potential elbow point was identified at cluster count 4, where the compromise was less pronounced. The results aligned with the observation by Bourassa et al., 2003, who highlighted that excessive homogeneity may not be practical, but having a restricted number of submarkets at a macro level is preferable.

The silhouette analysis score assessed the clustering quality, revealing a mean value of 0.69 and a median of 0.72. The silhouette analysis score assesses the quality of clustering by measuring the similarity of a data point to its cluster compared to other clusters. A score close to 1 signifies effective clustering, while negative scores imply potential misassignment of a data point to a cluster. The observed mean value of 0.69 suggested a reasonable cluster separation, indicating that most data points had a good chance of belonging to their assigned clusters. Figure 4.16 illustrates the distribution of silhouette scores across the dataset.

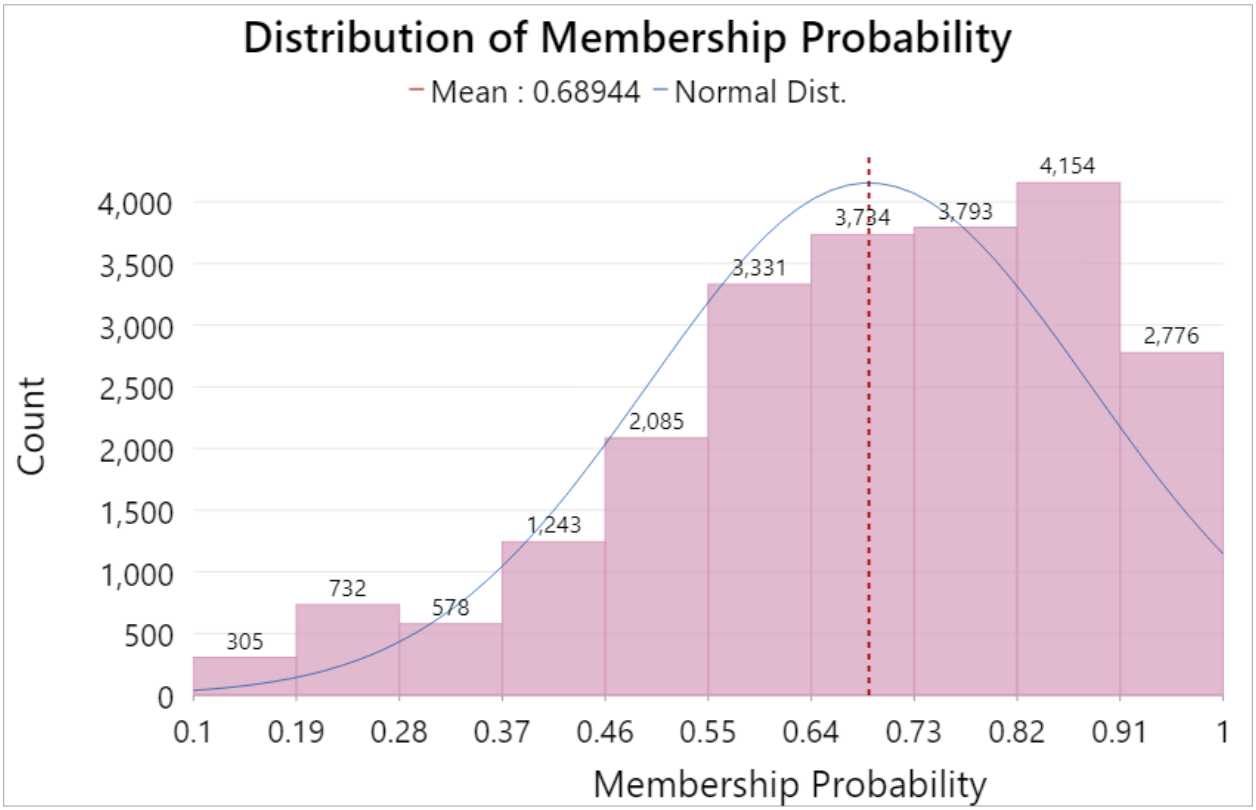


Figure 4.16 Distribution of membership probabilities

Figure 4.17 below presents a boxplot depicting how the five non-spatial attributes varied across the submarkets. This plot shows the unique characteristics of each identified submarket and allows for a direct comparison of the characteristics between different submarkets.

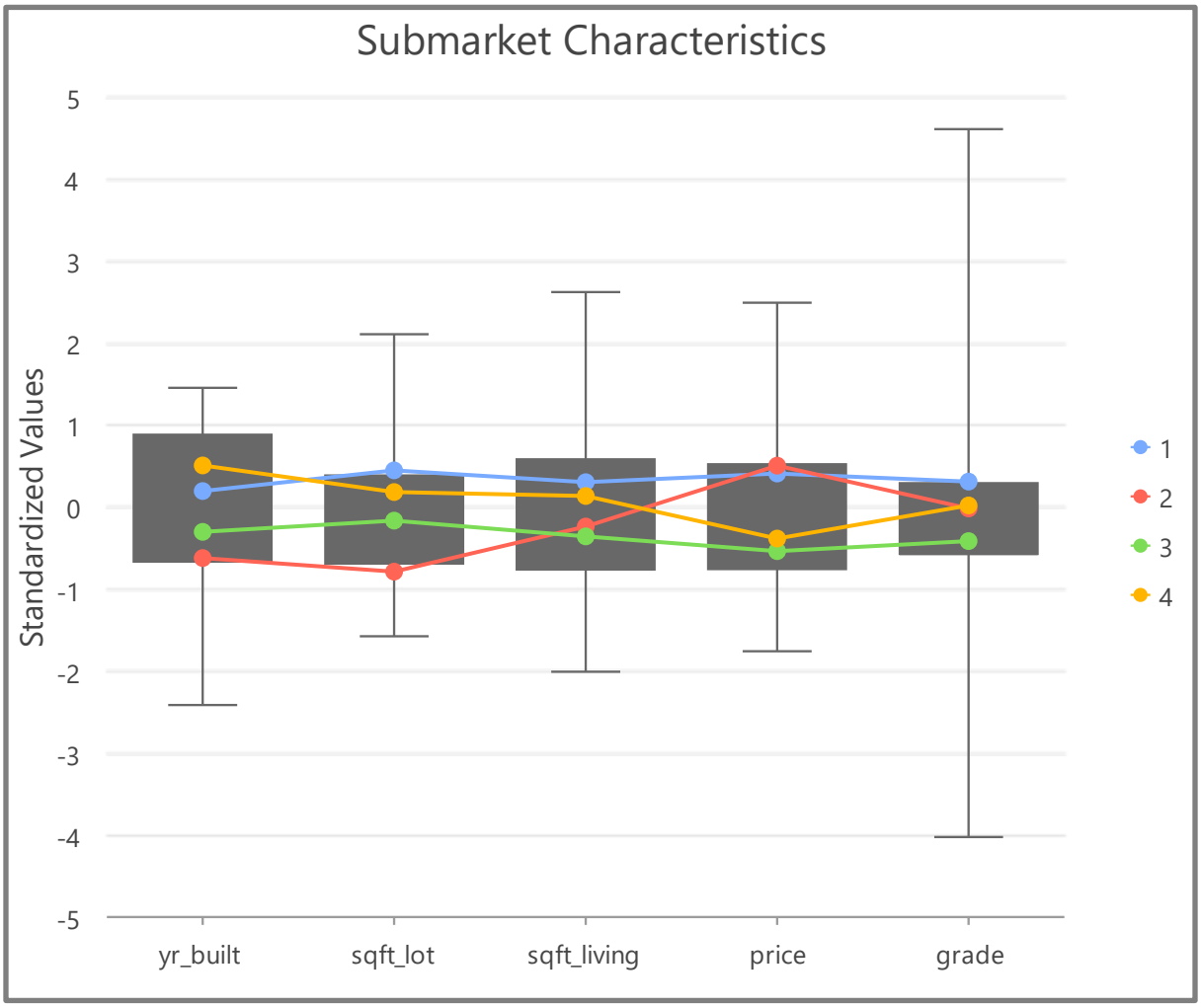


Figure 4.17 Boxplot showing variation in non-spatial attributes across submarkets

Linking the boxplots with their respective submarket maps revealed an imbalanced structure in the housing market within King County.

Submarket 1, encompassing Bellevue and Redmond cities, featured the most spacious, high-quality houses and ranked as the second priciest housing submarket in King County. These

houses had the most significant living spaces and occupied the most oversized land lots in King County. The high grade of these houses can be attributed to the strict enforcement of building codes and zoning regulations by local governments, ensuring high construction standards. Planning policies prioritize the conservation of expansive lots and open areas, which is evident in the substantial land parcels on which the houses are situated. These factors contributed to the high house prices in the area.

Submarket 2, the smallest among the submarkets, covered the areas surrounding Seattle and featured the highest housing prices. Notably, Submarket 2 stood out for having the oldest houses and the smallest land area on which the house is situated. This submarket had the second smallest house sizes with an average grade compared to other submarkets. The high housing prices in Submarket 2 can be attributed to Seattle's thriving tech industry, including giants like Amazon, which attracts high-paid employees to the region. However, the limited housing supply in the region fails to meet the soaring demand, leading to exorbitant house prices. As a result, individuals may choose smaller houses in Submarket 2, despite the higher prices, due to their accessibility to employment opportunities in the area.

Submarket 3 encompasses a notable portion of the county in the west region, including cities such as SeaTac, Federal Way, and the island on the left. The presence of waterfront properties characterizes this submarket, as it neighbours the water bodies in King County. Additionally, it features the second oldest houses, smallest houses, lowest prices, and lowest grades among the submarkets. In this submarket, cities like SeaTac and Federal Way have a sparser population than neighbouring cities, possibly resulting in lower housing demand and prices. The houses in these areas exhibit architectural styles from the 1970s, contributing to their low grades. The smaller houses in this submarket can be attributed to the challenge of purchasing

larger houses that are both costly and difficult to commute to and from, contributing to their affordability.

Submarket 4, including cities like Kent, Renton, Maple Valley, Auburn, Enumclaw, and Snoqualmie, features some of the oldest houses in King County. Despite their age, these houses offer affordable prices, average quality, and relatively large sizes in both living space and land occupied. The properties in this submarket have relatively big backyards, offering ample space for gardening and outdoor activities. Housing affordability in this area can be attributed to the fact that it has yet to undergo gentrification, resulting in more affordable options for buyers.

The figure illustrates the price distribution across submarkets, with Submarket 2 displaying the highest prices and Submarket 3 showing generally lower prices. Refer to Table 4.2 for the specific raw price values.

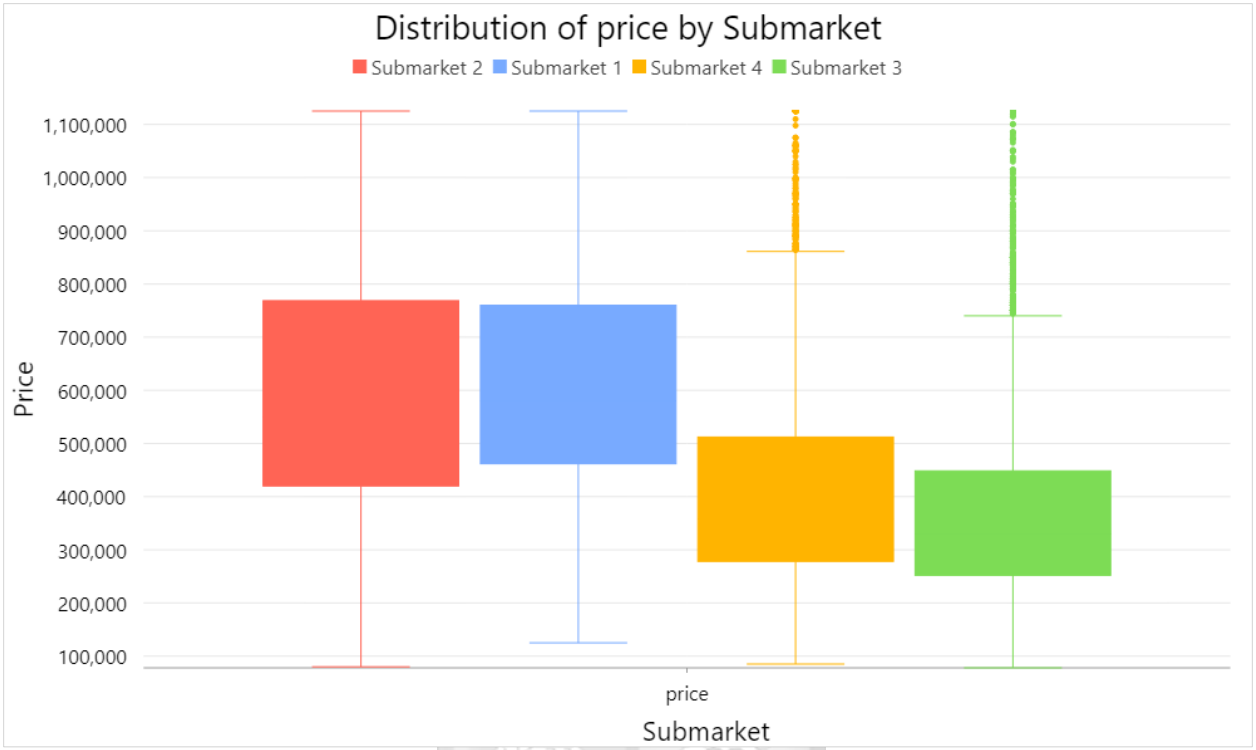


Figure 4.18 Distribution of the price by submarket

Table 4.2 House Prices Across Various Submarkets

Submarket	Min_Values	Median_Values	Max_Values
Submarket 3	78,000	330,000	1,125,150
Submarket 1	80,000	550,000	1,125,150
Submarket 4	85,000	358,500	1,125,150
Submarket 2	125,000	580,000	1,125,150

Similarly, the figure below illustrates the distribution of living space sizes across submarkets. Submarket 1 exhibits larger living spaces. Conversely, Submarket 3 comprises houses with the smallest sizes, potentially contributing to the lower prices observed earlier. Refer to Table 4.3 for the raw values of house sizes.

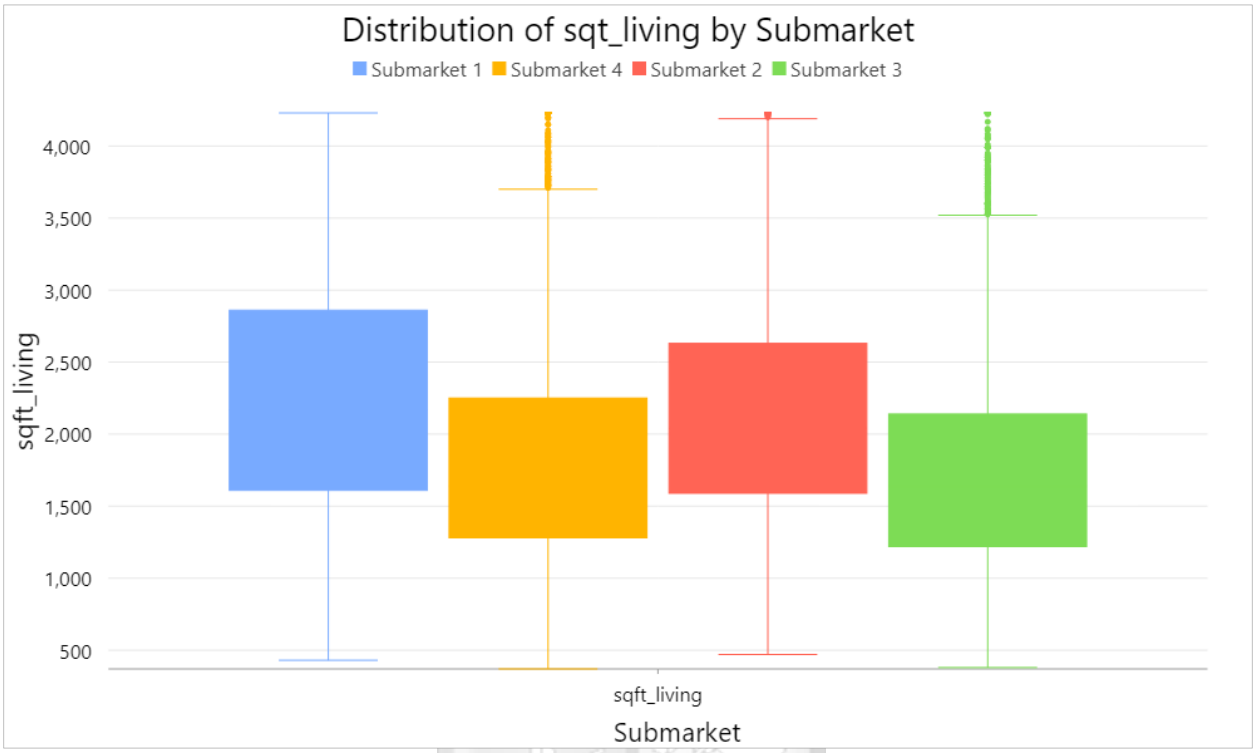


Figure 4.19 Distribution of living space sizes by submarkets

Table 4.3 House sizes in square feet across various submarkets

Submarket	Min_Values	Median_Values	Max_Values
Submarket 2	370	1,640	4,230
Submarket 3	380	1,640	4,230
Submarket 1	430	2,170	4,230
Submarket 4	470	2,040	4,230

According to Figure 4.20 below, despite high prices due to high demand, the oldest houses are situated in Submarket 2, having the earliest construction year. Conversely, newer houses are predominant in Submarket 3.

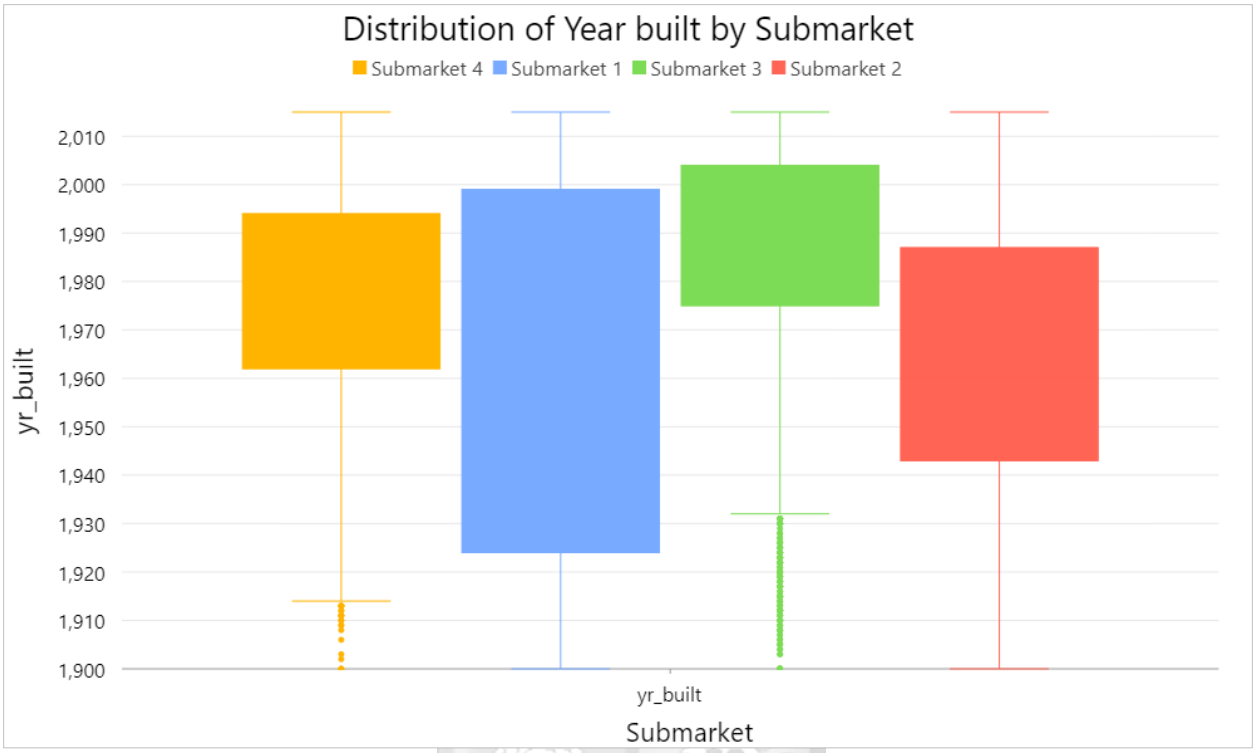


Figure 4.20 Distribution of year built by submarket

According to Figure 4.21 below, Submarket 1 stands out for its exceptionally high-grade houses, which correlate with their relatively high prices. Conversely, Submarket 3 exhibits the lowest grades among the submarkets, potentially contributing to the lower prices observed earlier.

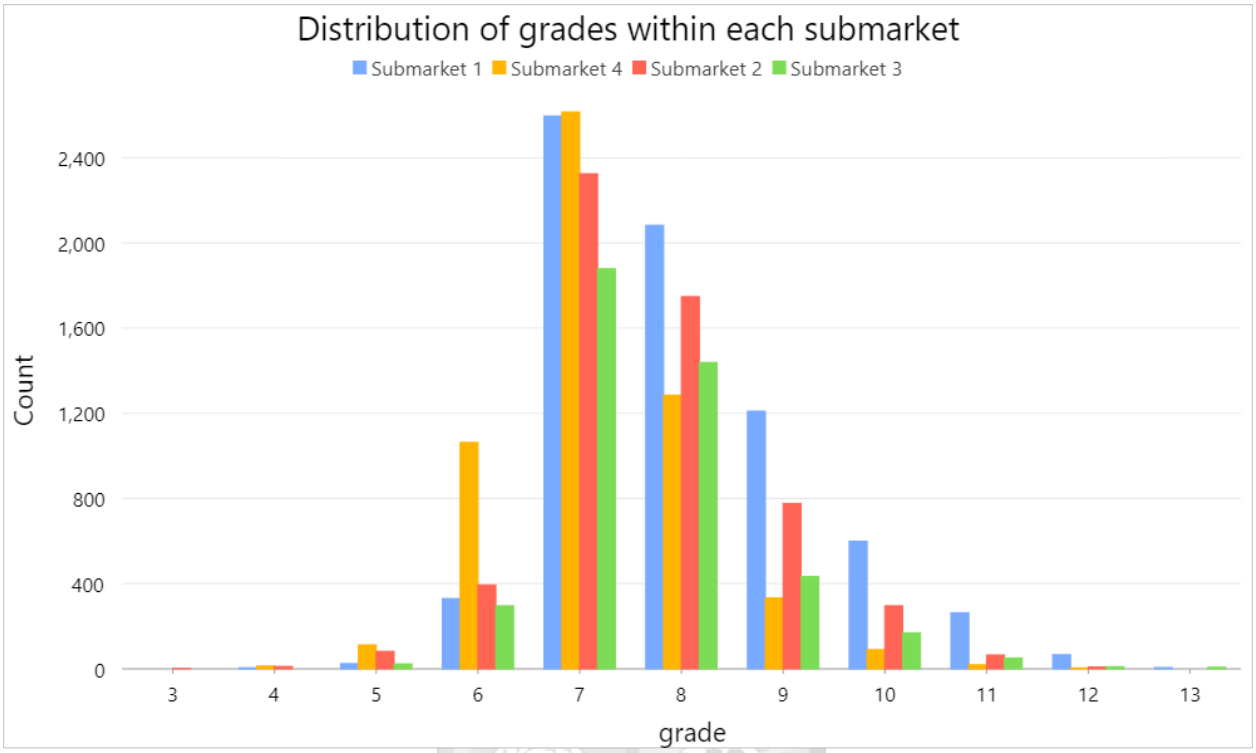


Figure 4.21 Distribution of grades by submarket

Chapter 5: Conclusion

This study utilizes data science and spatial analysis to delineate submarkets, focusing on a case study of King County, United States. Despite the recognized importance of location, existing algorithms for submarket delineation often overlook this crucial factor, potentially leading to inaccurate conclusions. This study addresses and attempts to rectify these limitations.

Firstly, the dataset is thoroughly examined. A method involving flooring and capping is utilized to address extreme values in numerical features like 'price'. Correlation coefficients are then used to identify and remove highly correlated features, and exploratory data analysis is conducted to omit features with limited discriminative power for clustering. The features ultimately chosen for delineating submarkets are price, size, grade, lot size', and year the house was built, all recognized as pivotal factors in the literature.

The SKATER algorithm is employed to investigate homogeneous and spatially contiguous housing submarkets, explicitly considering spatial effects, a factor often neglected in previous studies. By specifying the x and y coordinates of housing units in the SKATER algorithm, soft or hard spatial constraints are enforced, resulting in homogenous submarkets where houses are confined to adjacent or nearby features.

The appropriate number of submarkets is determined using the Pseudo F-statistic ratio, resulting in four submarkets deemed to accurately represent the housing market in King County. This finding aligns with previous research suggesting a limited number of submarkets at a macro level.

This study contributes to the literature by providing improved submarket boundaries compared to previous research. The resulting demarcations align well with the actual landscape,

revealing a significant imbalance in the housing market in King County. For instance, cities like Bellevue and Redmond in the north feature pricey, spacious, high-quality houses on large lots, while the southern region, including SeaTac and Federal Way, offers waterfront properties with older, smaller houses at comparatively lower prices.

These findings are invaluable for stakeholders addressing housing problems, enabling precise spatial analyses and effective policy formulation to tackle challenges such as housing imbalances in the market. Furthermore, this study explores the integration of location into machine learning algorithms, inspiring geospatial researchers to delve deeper into the role of space in machine learning.

Building upon these findings, future research could explore the utilization of advanced clustering and regionalization algorithms such as ClustGeo and REDCAP. These techniques have the potential to refine the delineation of housing submarkets.

References

AssunÇão, R. M., Neves, M. C., Câmara, G., & Da Costa Freitas, C. (2006a). Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science*, *20*(7), 797–811. https://doi.org/10.1080/13658810600665111

Barrett, L., & Alan, L. (2022). Creating Compact Regions of Social Determinants of Health. https://doi.org/10.48550/arXiv.2209.11836

Bourassa, S. C., Cantoni, E., & Hoesli, M. (2007). Spatial dependence, housing submarkets, and house price prediction. *The Journal of Real Estate Finance and Economics*, *35*(2), 143–160. https://doi.org/10.1007/s11146-007-9036-8

Bourassa, S. C., Hamelink, F., Hoesli, M., & MacGregor, B. D. (1999). Defining housing submarkets. *Journal of Housing Economics*, *8*(2), 160–183. https://doi.org/10.1006/jhec.1999.0246

Bourassa, S. C., Hoesli, M., & Peng, V. S. (2003). Do housing submarkets really matter? *Journal of Housing Economics*, *12*(1), 12–28. https://doi.org/10.1016/s1051-1377(03)00003-2

Chen, M., Chun, Y., & Griffith, D. A. (2023). Delineating housing submarkets using space–Time House sales data: Spatially constrained data-driven approaches. *Journal of Risk and Financial Management*, *16*(6), 291. https://doi.org/10.3390/jrfm16060291

Daoud, J. I. (2017). Multicollinearity and regression analysis. Journal of Physics: Conference Series, 949, 012009. https://doi.org/10.1088/1742-6596/949/1/012009

Gale, H., & Roy, S. S. (2022). Optimization of United States residential real estate investment through geospatial analysis and Market Timing. *Applied Spatial Analysis and Policy*, *16*(1), 315–328. https://doi.org/10.1007/s12061-022-09475-x

Goodman, A. C. (1981). Housing submarkets within urban areas: Definitions and evidence*. *Journal of Regional Science*, *21*(2), 175–185. https://doi.org/10.1111/j.1467-9787.1981.tb00693.x

Goodman, A. C., & Thibodeau, T. G. (2003). Housing market segmentation and hedonic prediction accuracy. *Journal of Housing Economics*, *12*(3), 181–201. https://doi.org/10.1016/s1051-1377(03)00031-7

Helbich, M., Brunauer, W., Hagenauer, J., & Leitner, M. (2013). Data-driven regionalization of housing markets. *Annals of the Association of American Geographers*, *103*(4), 871–889. https://doi.org/10.1080/00045608.2012.707587

Hotz, N. (2018, September 10). What is CRISP DM? Data Science Process Alliance. https://www.datascience-pm.com/crisp-dm-2/

Liu, X., Guo, R., Lei, G., & Liu, N. (2022). Classification of housing submarkets considering human preference: A case study in Shenyang, China. *Mathematical Problems in Engineering*, *2022*, 1–20. https://doi.org/10.1155/2022/2948352

Islam, K. S., & Asami, Y. (2009). Housing Market Segmentation: A Review. *Review of Urban & Regional Development Studies*, *21*(2–3), 93–109. https://doi.org/10.1111/j.1467-940x.2009.00161.x

Keskin, B., & Watkins, C. (2016). Defining spatial housing submarkets: Exploring the case for expert delineated boundaries. *Urban Studies*, *54*(6), 1446–1462. https://doi.org/10.1177/0042098015620351

Peeters, A., Zude, M., Käthner, J., Ünlü, M., Kanber, R., Hetzroni, A., Gebbers, R., & Ben-Gal, A. (2015). Getis–ord's hot- and cold-spot statistics as a basis for multivariate spatial clustering of Orchard Tree Data. *Computers and Electronics in Agriculture*, *111*, 140–150. https://doi.org/10.1016/j.compag.2014.12.011

Varghese, B., Unnikrishnan, A., & Jacob, K. (2014, 01). Spatial clustering algorithms- an overview. *Asian Journal of Computer Science And Information Technology*, *3*.

Wang, X., & Wang, J. (2009, 10). Using clustering methods in geospatial information systems. *Geomatica*, *64*. https://doi:10.1117/12.813150

Wang, Y., & Zhao, Q. (2022). House price prediction based on machine learning: A case of king county. Proceedings of the 2022 7th International Conference on Financial Innovation and Economic Development (ICFIED 2022). https://doi.org/10.2991/aebmr.k.220307.253

Waters, N. (2018). Tobler's first law of geography. *International Encyclopedia of Geography*, pp. 1–15. https://doi.org/10.1002/9781118786352.wbieg1011.pub2

Watkins, C. A. (2001). The definition and identification of housing submarkets. *Environment and Planning A: Economy and Space*, *33*(12), 2235–2253. https://doi.org/10.1068/a34162

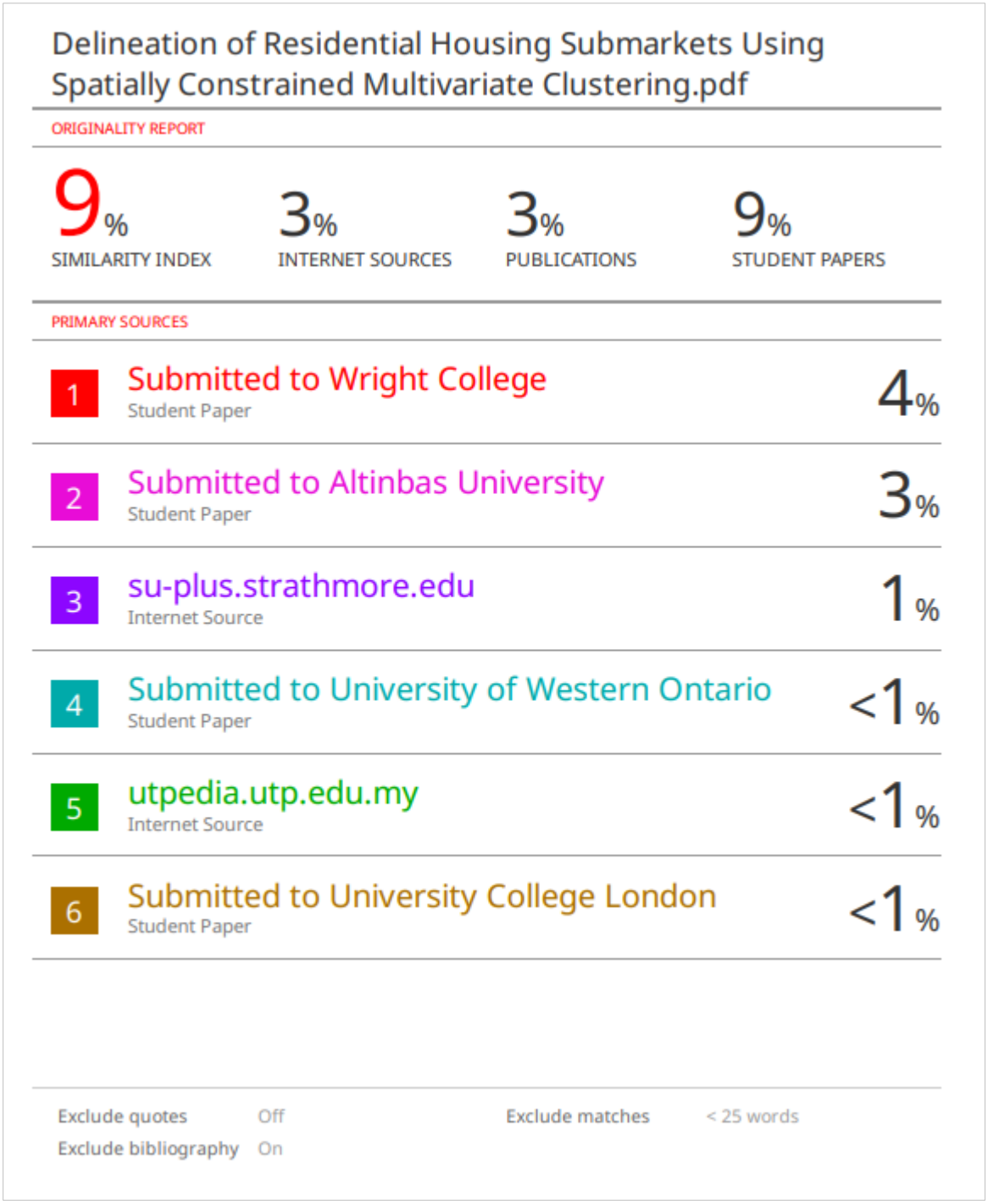
Wu, Changshan, & Sharma, R. (2012). Housing submarket classification: The role of spatial contiguity. *Applied Geography*, *32*(2), 746–756. https://doi.org/10.1016/j.apgeog.2011.08.011

Wu, Chao, Ye, X., Ren, F., & Du, Q. (2018). Modified data-driven framework for housing market segmentation. *Journal of Urban Planning and Development*, *144*(4). https://doi.org/10.1061/(asce)up.1943-5444.0000473

Wu, Chao, Ye, X., Ren, F., & Du, Q. (2018). Modified data-driven framework for housing market segmentation. *Journal of Urban Planning and Development*, *144*(4). https://doi.org/10.1061/(asce)up.1943-5444.0000473

Appendices

Appendix A: Similarity Report



Delineation of Residential Housing Submarkets Using Spatially Constrained Multivariate Clustering.pdf

ORIGINALITY REPORT

9%	3%	3%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Wright College Student Paper	4%
2	Submitted to Altinbas University Student Paper	3%
3	su-plus.strathmore.edu Internet Source	1%
4	Submitted to University of Western Ontario Student Paper	<1%
5	utpedia.utp.edu.my Internet Source	<1%
6	Submitted to University College London Student Paper	<1%

Exclude quotes	Off	Exclude matches	< 25 words
Exclude bibliography	On		

Appendix B: Ethical Clearance Confirmation

Strathmore UNIVERSITY

22nd March 2024

Mr Njoroge Samuel,
samuel.ngere@strathmore.edu

Dear Mr Njoroge,

RE: Delineation of Residential Housing Submarkets Using Spatially Constrained Multivariate Clustering

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** research proposal. Your application reference number is **SU-ISERC2049/24.** The approval period is from **22nd March 2024 to 21st March 2025.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

STRATHMORE UNIVERSITY INSTITUTIONAL SCIENTIFIC AND ETHICAL REVIEW COMMITTEE
(SU-ISERC)
25-Mar-2024
Email:ethicsreview@strathmore.edu
P.O BOX 59857-00200
NAIROBI-KENYA

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu