



Cape Peninsula
University of Technology

2024

PROPERTY RECOMMENDER

A Technical Report

Presented to

The Department of Mathematics and Physics

by

Sakhile Cedric Gumede

STUDENT NO.: 221356622

As an Assessment for the Module

MATHEMATICAL SCIENCES PROJECT 4 (MSP470S)

Within The Qualification

ADVANCED DIPLOMA IN MATHEMATICAL SCIENCES

ACADEMIC SUPERVISOR: Dr. Milaine SS Tchamga

Cape Peninsula University of Technology

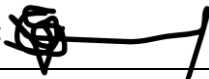
September 2024

DECLARATION

I, **Sakhile Cedric Gumede**, declare that the contents of this research report present my own work. I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is my own. Each contribution to, and quotation in this report from the work(s) of other people has been attributed and has been cited and referenced.

I have not allowed and will not allow anyone to copy my work with the intention of passing it off as his or her own work.

Date of Declaration: 21 October 2024

Signature: 

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor, **Dr. Milaine SS Tchamga** from the **Department of Civil Engineering and Geomatics** at Cape Peninsula University of Technology for the guidance of my research. Without his guidance, I would not have been able to complete this project on time.

I am also grateful to **Cape Peninsula University of Technology** for providing access to the necessary resources and tools required to conduct my analysis. Special thanks to the **Department of Mathematics and Physics** for their technical support and assistance in managing the data used in this study.

I thank my family for always believing in me and supporting me throughout my studies.

Finally, I appreciate the contributions of all those whose work inspired and informed this research, as well as the respondents/participants who provided the data necessary for analysis.

Thank you all for your support and contributions.

TABLE OF CONTENTS

DECLARATION	i
ACKNOWLEDGEMENTS	ii
LIST OF ABBREVIATIONS	vi
EXECUTIVE SUMMARY	1
CHAPTER 1: INTRODUCTION	2
1.1 Background	2
1.2 Research Problem	2
1.3 Research Objectives	3
1.4 Research Hypothesis	3
1.5 Significance of Study	4
CHAPTER 2: METHODOLOGY	5
2.1 Introduction	5
2.2 Data Source and Description	5
2.3 Data Preparation	6
2.4 Data Analysis	6
2.4.1 K-Means Clustering	7
2.4.2 Collaborative Filtering (Matrix Factorization)	7
CHAPTER 3: RESULTS and DISCUSSION	8
3.1 Introduction	8
3.2 Exploratory Data Analysis	8
3.3 Correlation Analysis	40
3.4 K-Means Clustering	42
3.5 Collaborative Filtering	45
CHAPTER 4: CONCLUSION and RECOMMENDATIONS	46
APPENDIX	47
REFERENCES	48

LIST OF FIGURES

Figure 3.1: Number of Properties between Gauteng and Western Cape	8
Figure 3.2: Number of properties with different number of bedrooms.....	9
Figure 3.3: Number of properties with different number of bathrooms.....	10
Figure 3.4: Number of properties in different cities.	11
Figure 3.5: Number of properties in different suburbs.	12
Figure 3.6: Number of properties for different property types.	13
Figure 3.7: Frequency of houses with different sizes, measured in square units.	14
Figure 3.8: Boxplot of house size	15
Figure 3.9: Frequency of erf sizes, measured in square units.	16
Figure 3.10: Boxplot of erf size measured in square units.....	17
Figure 3.11: Frequency distribution of properties listed at different prices in Rands.....	18
Figure 3.12: Distribution of property listing prices in Rands	19
Figure 3.13: Distribution of house sizes for different numbers of bedrooms.	20
Figure 3.14: Distribution of erf sizes for different numbers of bedrooms.	21
Figure 3.15: Distribution of listing prices for different numbers of bedrooms.....	22
Figure 3.16: Distribution of house sizes for different numbers of bathrooms.	23
Figure 3.17: Distribution of erf sizes for different numbers of bathrooms.	24
Figure 3.18: Distribution of listing prices for different numbers of bathrooms.....	25
Figure 3.19: Distribution of house sizes across different suburbs.	26
Figure 3.20: Distribution of erf sizes across different suburbs.	27
Figure 3.21: Distribution of listing prices across different suburbs.	28
Figure 3.22: Distribution of house sizes across two provinces: Gauteng and Western Cape.	29
Figure 3.23: Distribution of erf sizes across two provinces: Gauteng and Western Cape.	30
Figure 3.24: Distribution of listing prices across two provinces: Gauteng and Western Cape.....	31
Figure 3.25: Distribution of house sizes for different property types.	32
Figure 3.26: Distribution of erf sizes for different property types.	33
Figure 3.27: Distribution of listing prices for different property types.....	35
Figure 3.28: Distribution of house sizes across different cities.	36
Figure 3.29: Distribution of erf sizes across different cities.	37
Figure 3.30: Distribution of List Price across different cities.	38
Figure 3.31: Pair Plot.	39
Figure 3.32: Correlation matrix	40
Figure 3.33: Elbow Method plot.	42
Figure 3.34: The scatter plot displays List Price versus House Size Clusters	43
Figure 3.35: The boxplots of 6 numerical variables subplots.	44
Figure 3.36: Distribution across the 5 clusters.....	44

LIST OF TABLES

Table 1: List of variables and description	5
Appendix 1: Project Calendar	47

LIST OF ABBREVIATIONS

- EDA – Exploratory Data Analysis
- Erf – Yard
- H0 – Null Hypothesis
- H1 – Alternative Hypothesis
- WCSS – Within-Cluster Sum of Squares
- RMSE – Root Mean Square Error
- IQR – Interquartile Range
- SVD – Singular Value Decomposition

EXECUTIVE SUMMARY

This project aims to develop a Property Recommender System using property listings dataset from Property24 with details such as the number of bedrooms, bathrooms, property size, and suburb name. The objective is to create a recommendation engine that suggests similar properties to users based on their past interactions on the website. By employing data science techniques such as clustering, exploratory data analysis, and machine learning, the system can provide personalized recommendations to enhance the user experience and improve engagement.

Through extensive data analysis and modelling, this report outlines the methods used, including data preprocessing, clustering techniques, and recommendation algorithms. A thorough evaluation of the models was performed, and the final model's performance was optimized using hyperparameter tuning. The system efficiently recommends properties that match user preferences, with potential to add value to the business by increasing user satisfaction and time spent on the website.

CHAPTER 1: INTRODUCTION

1.1 Background

Property24 is a leading real estate platform in South Africa, offering a wide range of property listings for sale and rent, including houses, apartments, and vacant land or plot, with advanced search filters for price, size, location, and specific attributes like pet-friendliness and security. The platform also provides tools like property valuations, loan calculators, and real estate advice. A paper by Gharahighehi et al. (2021) explored the impact of e-commerce on the real estate industry, highlighting the role of recommendation systems in personalizing property suggestions. The property market is highly dynamic, and users often browse several listings before deciding. However, finding similar properties manually can be time-consuming. By building a recommendation system, we aim to automate this process and improve user satisfaction by suggesting relevant listings based on previous interests.

1.2 Research Problem

Property 24, a leading real estate platform in the country, features a vast array of property listing with diverse attributes. Despite offering an extensive selection, platform users often face difficulties navigating through the property listings to find their property of preference. This can yield in a less satisfying user experience, reduced platform engagement, and missed opportunities for sales.

The core issue in Property 24 is the absence of a highly personalized and efficient recommendation system to match users with relevant properties. Users feel overwhelmed when navigating the platform by the sheer number of listings, and majority does not match with their preferences.

Developing an advanced property recommendation system for Property24 can greatly enhance the platform's usability by making it easier for users to find properties that suit their needs. A more streamlined search process will increase user satisfaction, boost time spent on the platform, and potentially raise conversion rates. This research also contributes to the field of recommendation systems in real estate, showcasing how machine learning can optimize user experience on platforms like Property24.

1.3 Research Objectives

To solve this problem, a K-Means Clustering-based approach will be used. Properties with similar attributes (e.g., bedrooms, bathrooms, size, and suburb) will be grouped together, and build the recommendation system known as collaborative filtering. Data analysis, visualization, and machine learning techniques will be applied to build the model.

- **Identify Distinct Property Segments:** Utilize K-Means Clustering to categorize properties into segments based on list prices and house sizes.
- **Analyze the Relationship Between House Size and List Price:** Examine the correlation between house sizes and their list prices across clusters to understand their mutual influence.
- **Evaluate Cluster Characteristics and Amenities:** Assess differences in key property characteristics among clusters and their relationship to pricing.
- **Investigate Variability Within Clusters:** Analyze variability in property features within each cluster to identify outliers or unique properties that indicate market trends.
- **Provide Insights for Market Strategies:** Derive actionable insights from clustering results to inform targeted marketing strategies, investment opportunities, and pricing models for real estate stakeholders.
- **Contribute to Real Estate Analytics:** Enhance understanding of property market dynamics through clustering techniques, contributing to the broader field of real estate analytics and data-driven decision-making.

1.4 Research Hypothesis

- **Null hypothesis (H_0):** There is no significant relationship between house size and list price across the different clusters.
- **Alternative hypothesis (H_1):** There is a significant relationship between house size and list price across the different clusters, with larger houses generally having higher list prices.

1.5 Significance of Study

The significance of this study lies in its ability to enhance understanding of property segmentation by categorizing properties based on list prices and sizes, which enables targeted marketing strategies. It improves decision-making for stakeholders, helping investors and real estate professionals make informed pricing and investment choices. Additionally, the study identifies emerging market trends by analysing variability within property clusters and offers actionable insights for tailored marketing approaches. By contributing to the field of real estate analytics and providing a foundation for future research, the study addresses practical implications for policymakers in tackling housing market issues related to affordability and urban development, ultimately promoting strategic decision-making in the sector.

CHAPTER 2: METHODOLOGY

2.1 Introduction

This chapter outlines the steps taken to analyse and preprocess the property listings dataset for developing a recommendation model. The analysis focused on key features that are essential for users when searching for properties, including province name, suburb name, property type bedrooms, bathrooms, house size erf size (in square meters), and listing price. These variables were selected as they summarize the primary factors influencing property selection.

2.2 Data Source and Description

This study utilizes the 2024 Real Estate Listings Dataset, from Property 24. The data was compiled by data capturing team and provides information on residential properties listed in Gauteng and Western Cape from March 2023 to July 2024, 2023. Key features of the dataset include the number of bedrooms, bathrooms, property size in square meters, suburb, and listing price. The primary dataset for this project, Property 24 Listing (2024), provides property listings with different features. The shape of the data is (986379, 18) before data cleaning with key attributes such as:

Table 1: List of variables and description

Variables	Description
Suburb Name	Categorical variable, indicating location.
City Name	Categorical variable, indicating location.
Province Name	Categorical variable indicating location.
Bedrooms	Integer, representing the number of bedrooms.
Bathrooms	Integer, representing the number of bathrooms.
Property Type	Categorical variable, indicating different types of properties.
Erf Size	Continuous variable, in square meters.
House Size	Continuous variable, in square meters.
Number of Parking Spaces	Integer, representing number of parking spaces.
List Price	Continuous variable, indicating property prize.

2.3 Data Preparation

Data preparation is a crucial step in ensuring the quality and integrity of the dataset used for modelling. The following actions were taken to prepare the property listings dataset for analysis and clustering.

The dataset initially contained some missing values in key features such as price listing, house size and erf size. Since these are critical variables for modelling, properties with missing data in any of these fields were removed. This reduced the dataset to about 85% of its original size but ensured that the remaining data was complete and reliable for analysis.

To ensure that features like price listing, house size and erf size were on comparable scales, normalization was performed on these numerical variables. This step was important for K-Means clustering, as the algorithm is sensitive to the scale of input features. Without normalization, variables with larger scales (such as price) could disproportionately influence the clustering results.

2.4 Data Analysis

In the initial analysis phase, Exploratory Data Analysis (EDA) was conducted to understand the dataset. Histograms and boxplots revealed the distribution and characteristics of numerical features (e.g., size and price), while bar plots showed property distribution across suburbs. Pearson correlation analysis identified relationships between key numerical variables, such as house size, erf size, and list price. A pair plot was also used to visually assess clustering patterns among property features.

To identifying groups of similar properties, K-Means Clustering was selected as the primary model. K-Means, an unsupervised learning algorithm, is well-suited for this task as it effectively groups data points based on similarity, making it ideal for clustering property listings according to their features.

Collaborative Filtering is a popular recommendation technique that suggests items (in this case, properties) based on the preferences of similar users or items. Collaborative Filtering model is implemented to leverage user behaviour data like properties viewed or liked by users to offer personalized property recommendations. This approach complements K-Means clustering by focusing not only on property similarities but also on patterns in user preferences, making recommendations even more tailored.

2.4.1 K-Means Clustering

K-Means Clustering is an unsupervised machine learning algorithm commonly used for clustering data points into groups, where each group, or "cluster," is defined by a centroid that minimizes the distance between data points within the cluster. Below is the K-Means Clustering expression:

$$\sum_{i=1}^n \sum_{k=1}^K \|x_i - \mu_k\|^2$$

Where:

- n : The total number of clusters.
- K : The number of data points in each cluster.
- x_i : Represents the data point i in each cluster.
- μ_k : represents the centroid of each cluster.
- $\|x_i - \mu_k\|^2$: The squared Euclidean distance between the data point x_i and the cluster centroid μ_k .

2.4.2 Collaborative Filtering (Matrix Factorization)

Collaborative Filtering with Matrix Factorization is a popular recommendation method that predicts user preferences by breaking down a large, sparse user-item interaction matrix into smaller matrices, allowing for efficient prediction of unknown ratings. This approach is well-suited for data on user interactions like movie viewing or product purchases. Below is the collaborative equation filtering equation:

$$R_{ui} = \mu + b_u + b_i + p_u^T q_i$$

Where:

- R_{ui} is the predicted rating,
- μ the global mean,
- b_u and b_i are user and item biases,
- p_u and q_i are latent feature vectors for user and items

CHAPTER 3: RESULTS and DISCUSSION

3.1 Introduction

The primary objective of this project was to develop an effective property recommendation system using a combination of clustering and collaborative filtering. By analysing property attributes and user behaviour, the project aimed to provide end users with personalized property suggestions that align with their preferences.

This chapter summarizes the study's findings from applying Machine Learning models and data visualization techniques. The property recommender system project successfully utilized various visualizations, that improve dataset comprehension. The following sections outline key findings, interpretations, and their alignment with the project objectives.

3.2 Exploratory Data Analysis

Figure 3.1: Number of Properties between Gauteng and Western Cape

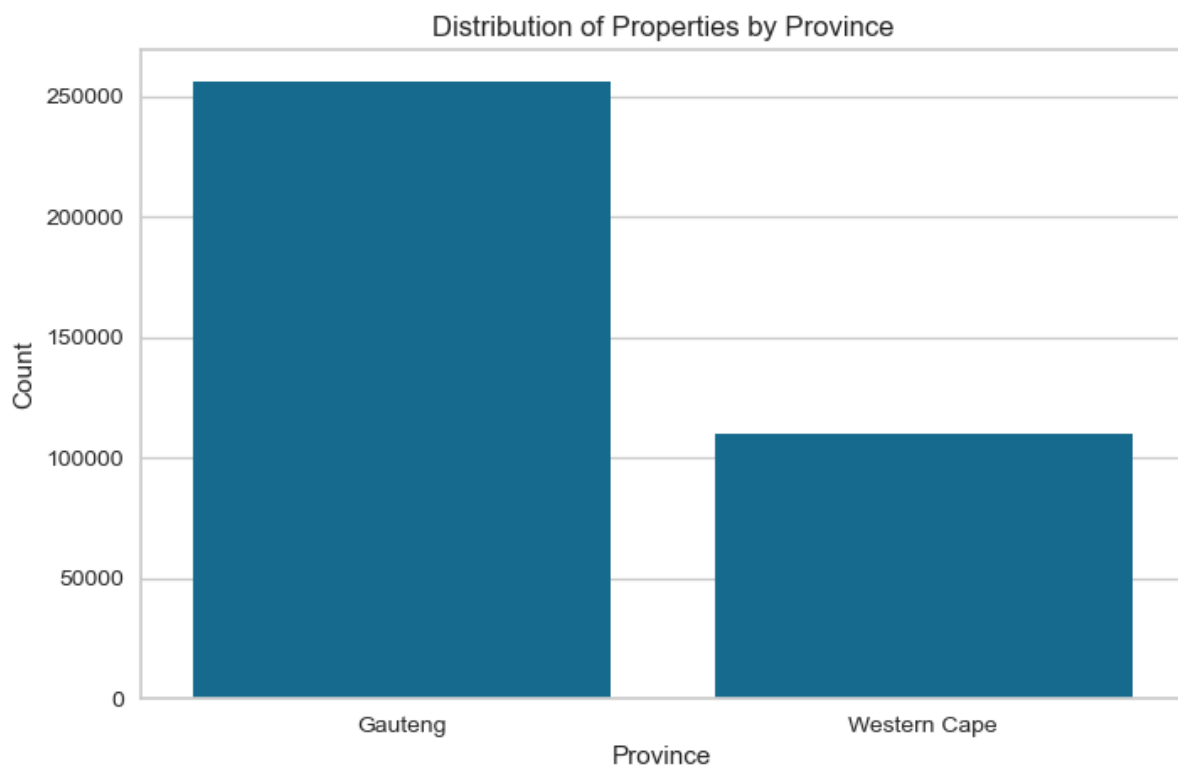


Figure 3.1 shows the distribution of properties by province. Gauteng has significantly more properties (around 250,000+) compared to the Western Cape, which has roughly half the amount, just over 100,000.

Figure3.2: Number of properties with different number of bedrooms.

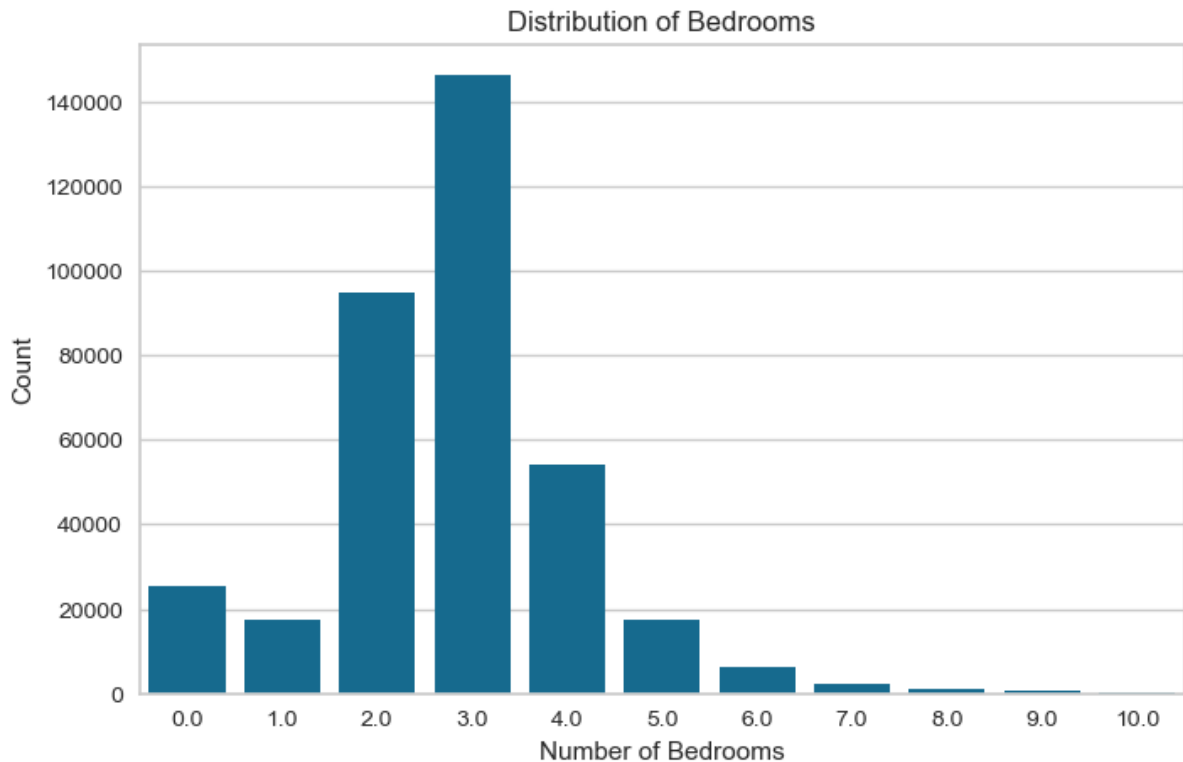


Figure 3.2 shows that the distribution of number of bedrooms is right-skewed, with most properties having a smaller number of bedrooms and fewer properties having a larger number of bedrooms.

- The plot is indicating that most properties have 2 and 3 bedrooms but 3 is the highest.
- The number of bedrooms decreases rapidly after 3. As the number of bedrooms increases, the bars become progressively shorter, suggesting a decline in the number of properties with more bedrooms and properties with 10 bedrooms are very few.

Figure 3.3: Number of properties with different number of bathrooms.

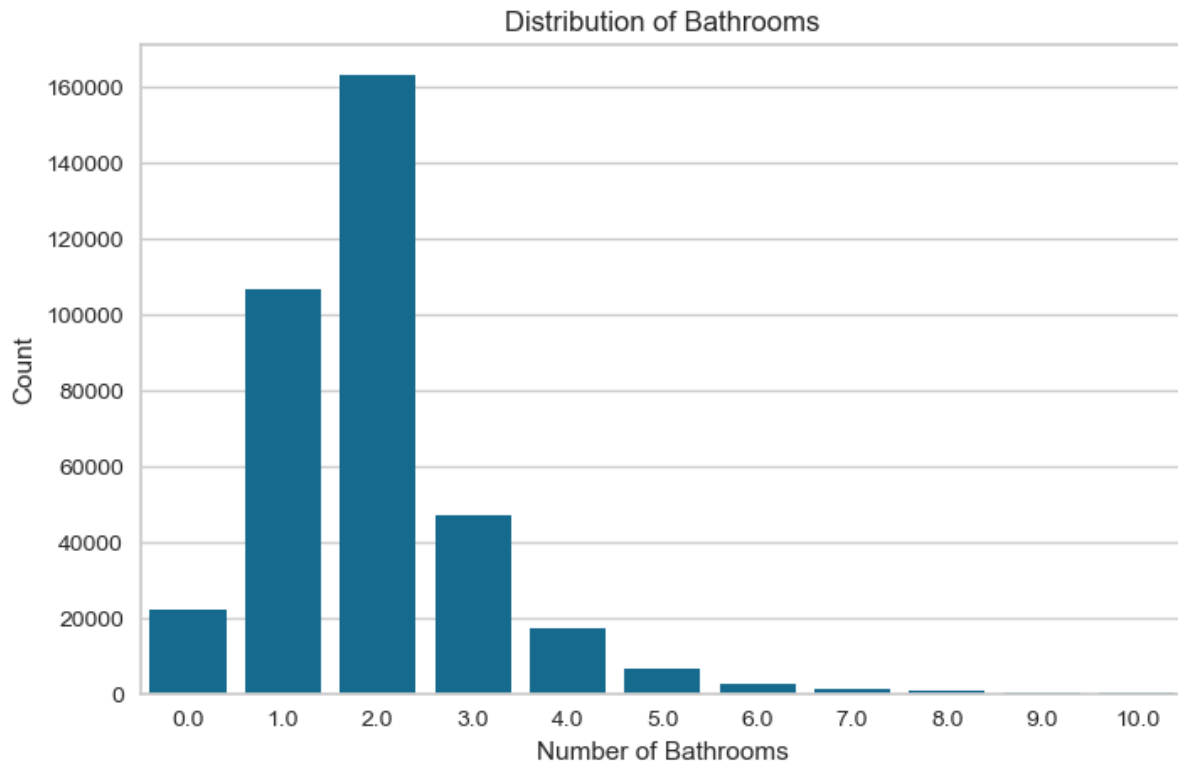


Figure 3.3 shows that the distribution of bathrooms is right-skewed, with most properties having a smaller number of bathrooms and fewer properties having a larger number of bathrooms. Like the bedroom distribution, there's a concentration in the lower range of bathroom counts.

- Most properties have 1 and 2 bathrooms but 2 is the highest.
- The number of bathrooms decreases rapidly after 2.
- As the number of bathrooms increases, the bars become progressively shorter, suggesting a decline in the number of properties with more bathrooms and properties with 10 bathrooms are rare.

Figure 3.4: Number of properties in different cities.

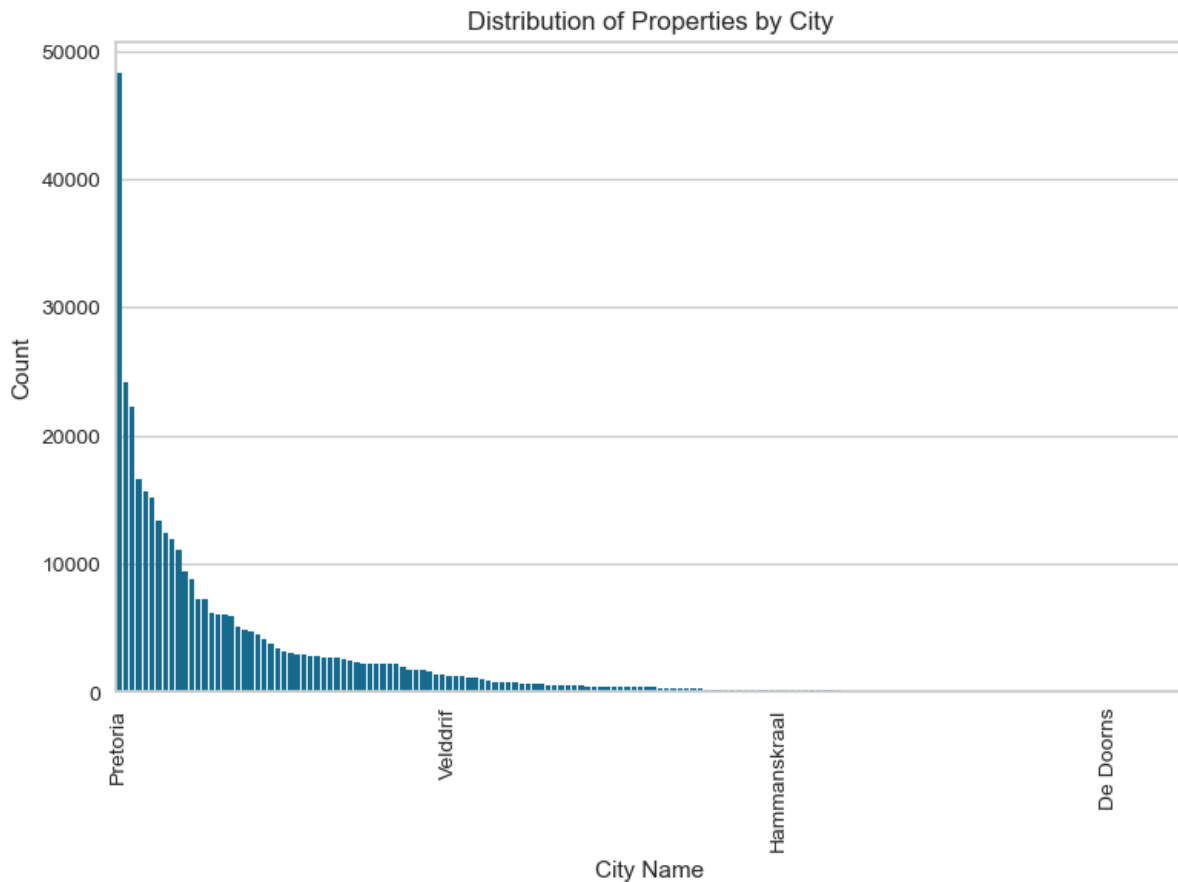


Figure 3.4 shows a highly skewed distribution, with Pretoria having a dominant share of properties and most other cities having significantly fewer properties.

- Pretoria has the most properties: The tallest bar in the chart is for Pretoria, indicating that it has the highest number of properties among the cities shown.
- The number of properties decreases rapidly for other cities: After Pretoria, the bars become progressively shorter, suggesting a significant drop in the number of properties in other cities.
- Many cities have a small number of properties: Most of the bars are quite short, indicating that most cities have a relatively low number of properties compared to Pretoria.

Figure 3.5: Number of properties in different suburbs.

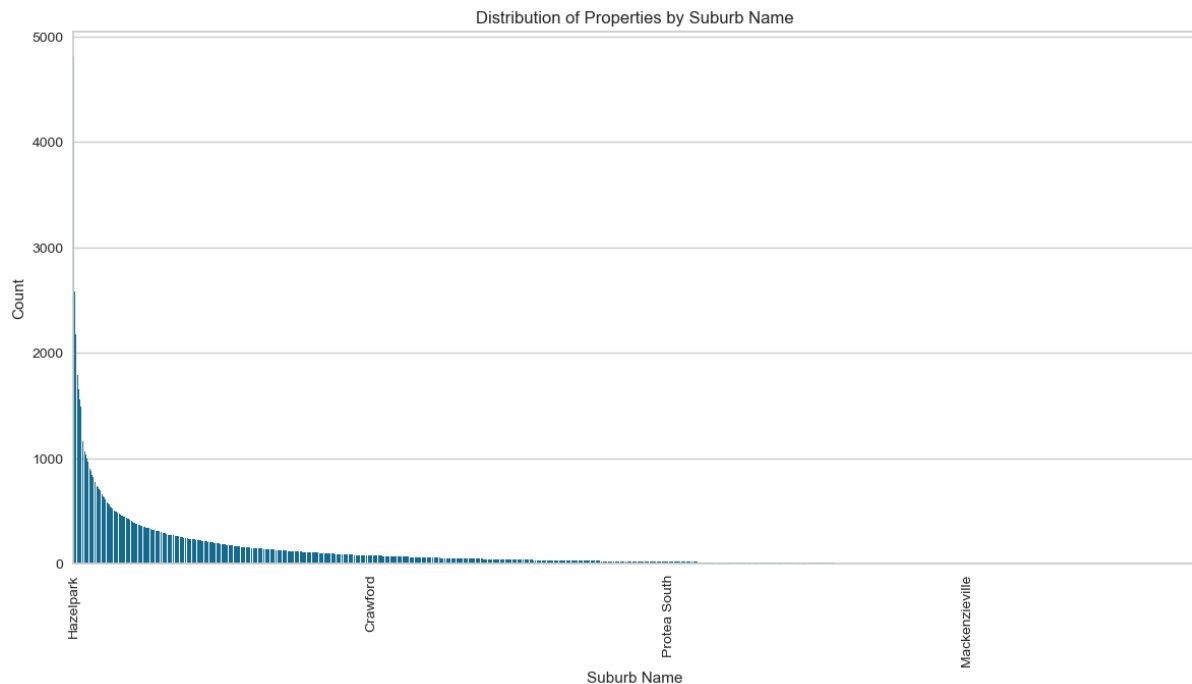


Figure 3.5 shows a highly skewed distribution, with Hazelpark having a dominant share of properties and most other suburbs having significantly fewer properties.

Here are some key observations from the plot:

- Hazelpark has the most properties, the tallest bar in the chart is for Hazelpark, indicating that it has the highest number of properties among the suburbs shown.
- The number of properties decreases rapidly for other suburbs. After Hazelpark, the bars become progressively shorter, suggesting a significant drop in the number of properties in other suburbs.
- Many suburbs have a small number of properties. Most of the bars are quite short, indicating that most suburbs have a relatively low number of properties compared to Hazelpark.

Figure 3.6: Number of properties for different property types.

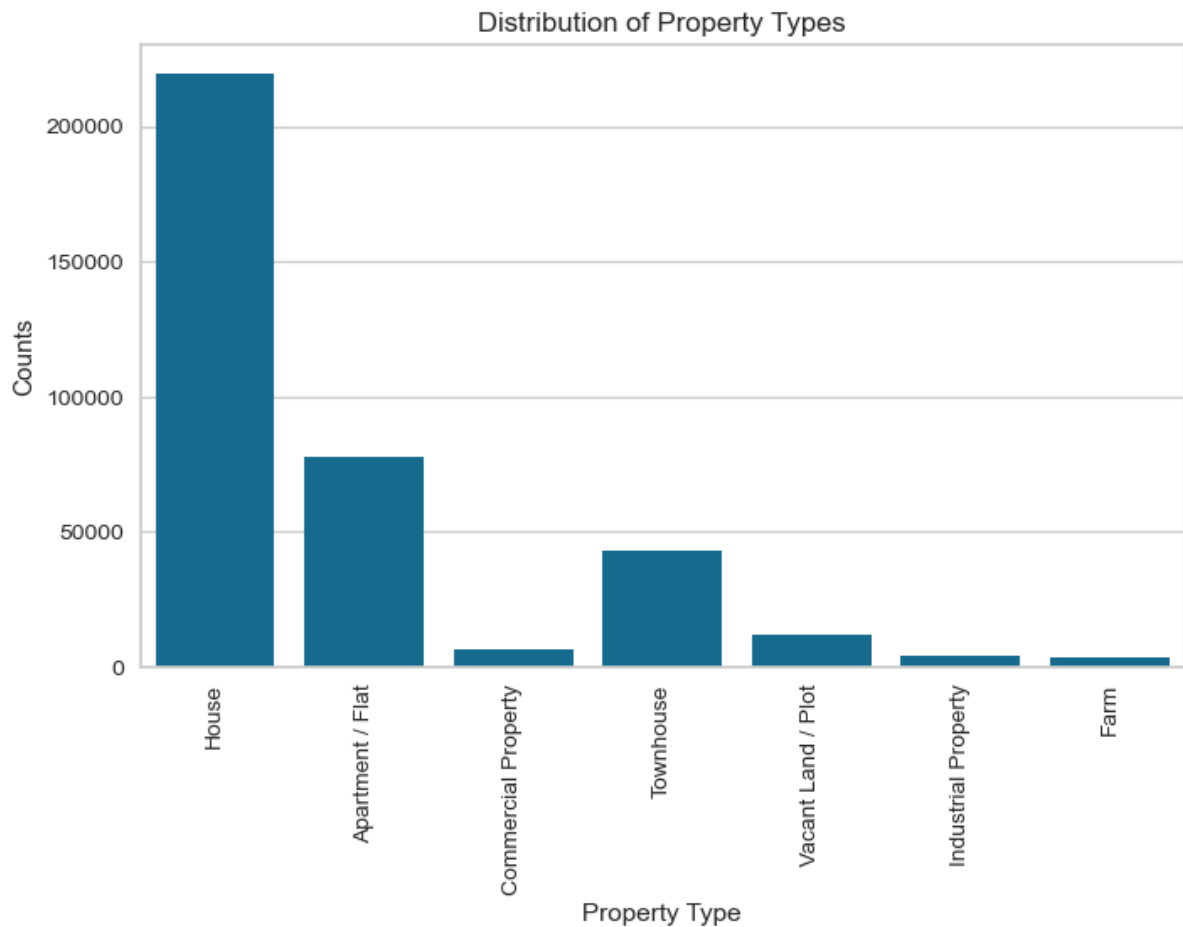


Figure 3.6 shows that houses and apartments/flats dominate the property types in the dataset, with other types being less prevalent.

- The property type is indicating that houses are the most prevalent property type in the dataset.
- Apartments/Flats are the second most common, the plot is suggesting that apartments or flats are also a significant property type.
- Townhouse is significantly less than the first House and Apartment / Flat.
- Commercial Property, Vacant Land / Plot, Industrial Property, and Farm are all relatively short on plot, suggesting that these property types are less common compared to houses, apartments and townhouse.

Figure 3.7: Frequency of houses with different sizes, measured in square units.

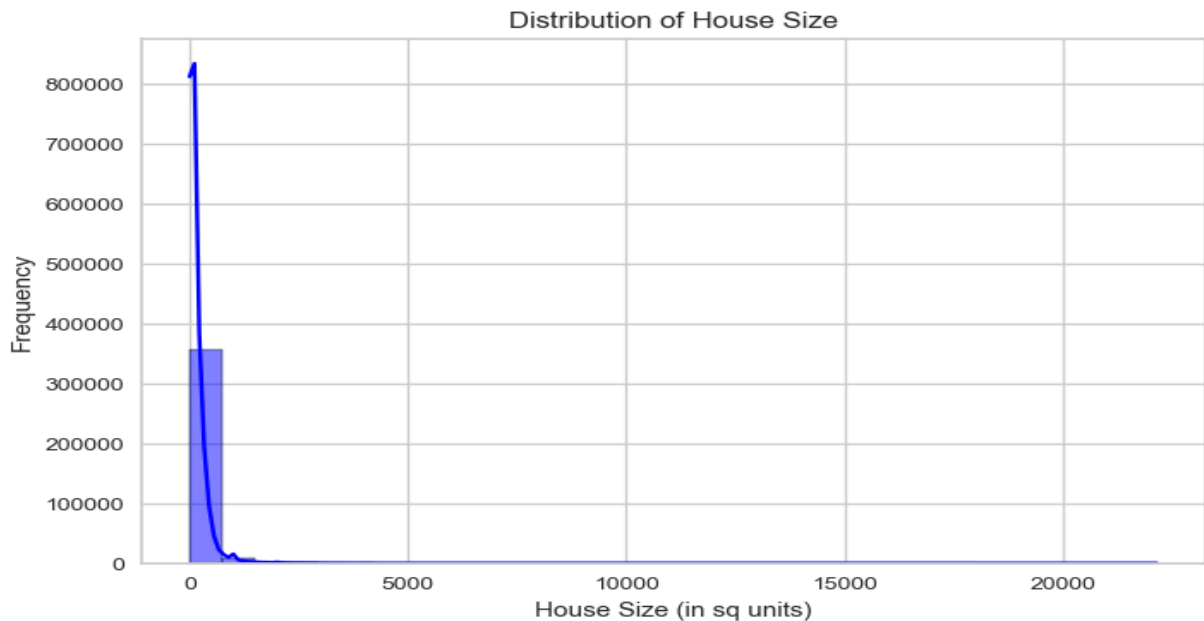


Figure 3.7 shows that the distribution of house sizes is right-skewed, with most houses having a smaller size and a few houses having a larger size.

Key observations:

- Most house sizes are below 1000 square units.
- There is a long tail extending to the right, indicating a small number of much larger houses (up to 20,000 square units or more).
- The high frequency of small house sizes suggests a clustered or dense population of smaller homes, with fewer large properties.

Figure3. 8: Boxplot of house size

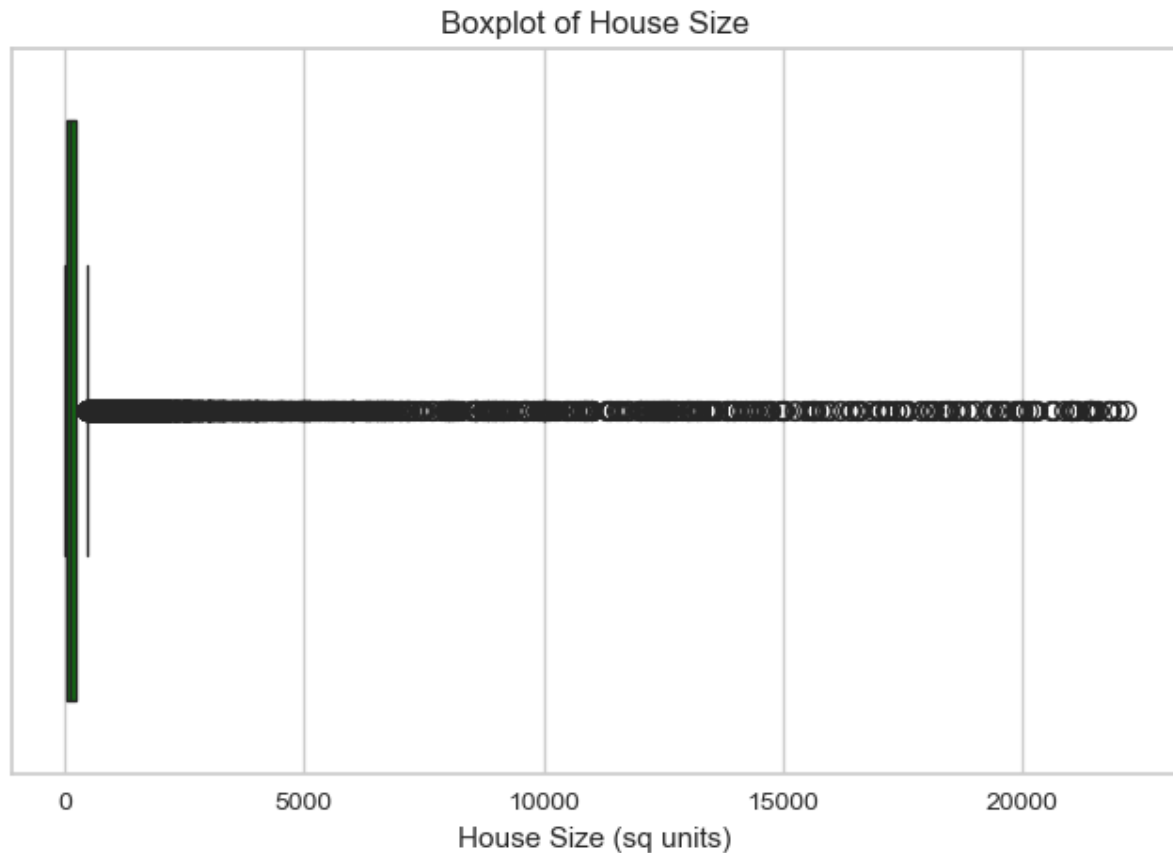


Figure 3.8 visually reinforces the same distribution shape seen in the previous histogram, where many house sizes are clustered at the lower end, but there are numerous large houses that create significant outliers. Here is a breakdown:

- **Median and Quartiles:** The narrow interquartile range (IQR) close to the left of the plot shows that most houses are small, with the median near the lower end, indicating a low central tendency.
- **Outliers:** Numerous data points extend far beyond the right whisker, revealing many large outliers and a long right tail, which suggests a highly skewed distribution.
- **Whiskers:** The short whiskers relative to the outlier spread indicate limited variation among typical house sizes, with a few extreme values contributing to the rightward skew.

Figure 3.9: Frequency of erf sizes, measured in square units.

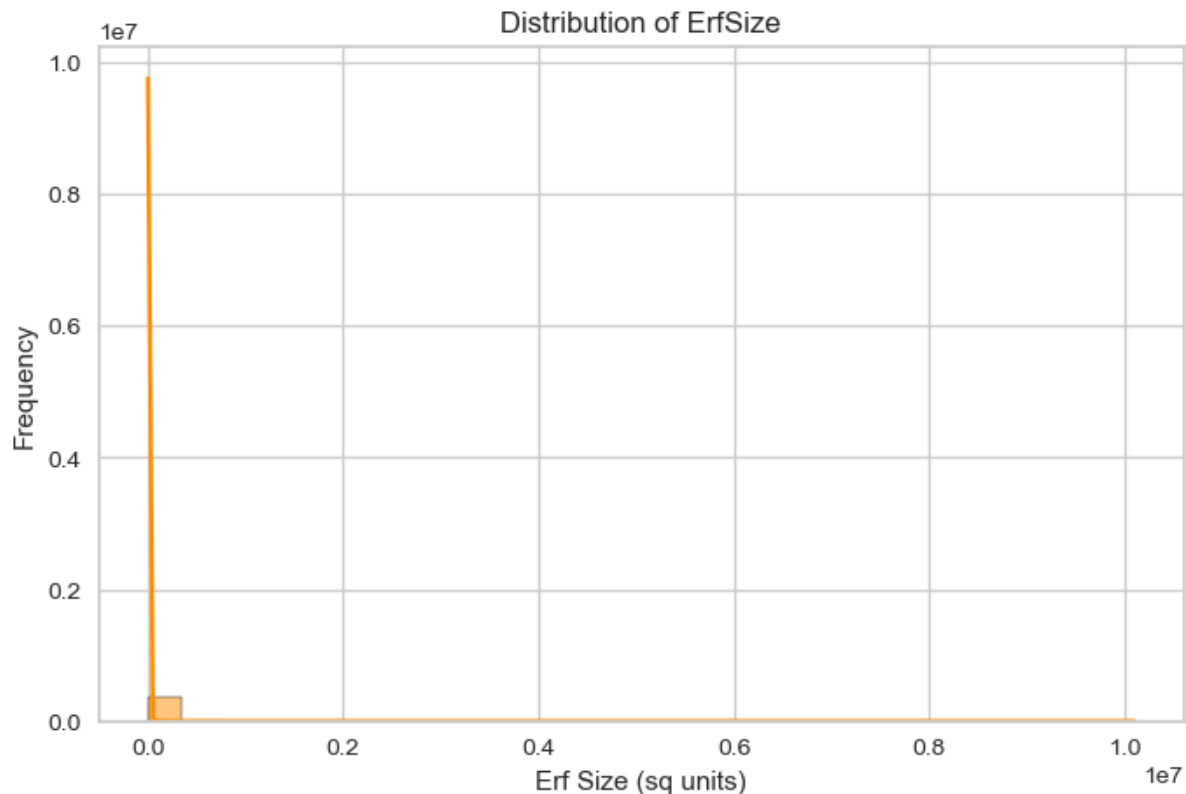


Figure 3.9 shows that the frequency of erf sizes appears highly concentrated around smaller values, with a few exceptionally large plots stretching the distribution to the right. This extreme skewness aligns with typical real estate data, where most properties are small with only large outliers.

- **Heavy Right Skew:** Most erf sizes are clustered near zero, with a peak around smaller sizes, while the x-axis extends up to 10 million square units, indicating that most erfs are very small.
- **Extreme Outliers:** A long right tail shows the presence of a few significantly larger erfs, likely large plots of land that are uncommon in the dataset.
- **Scale and Frequency:** The y-axis has a high frequency scale, nearing 10 million, showing a dense concentration of small erf sizes, with large properties being rare.

Figure 3.10: Boxplot of erf size measured in square units.

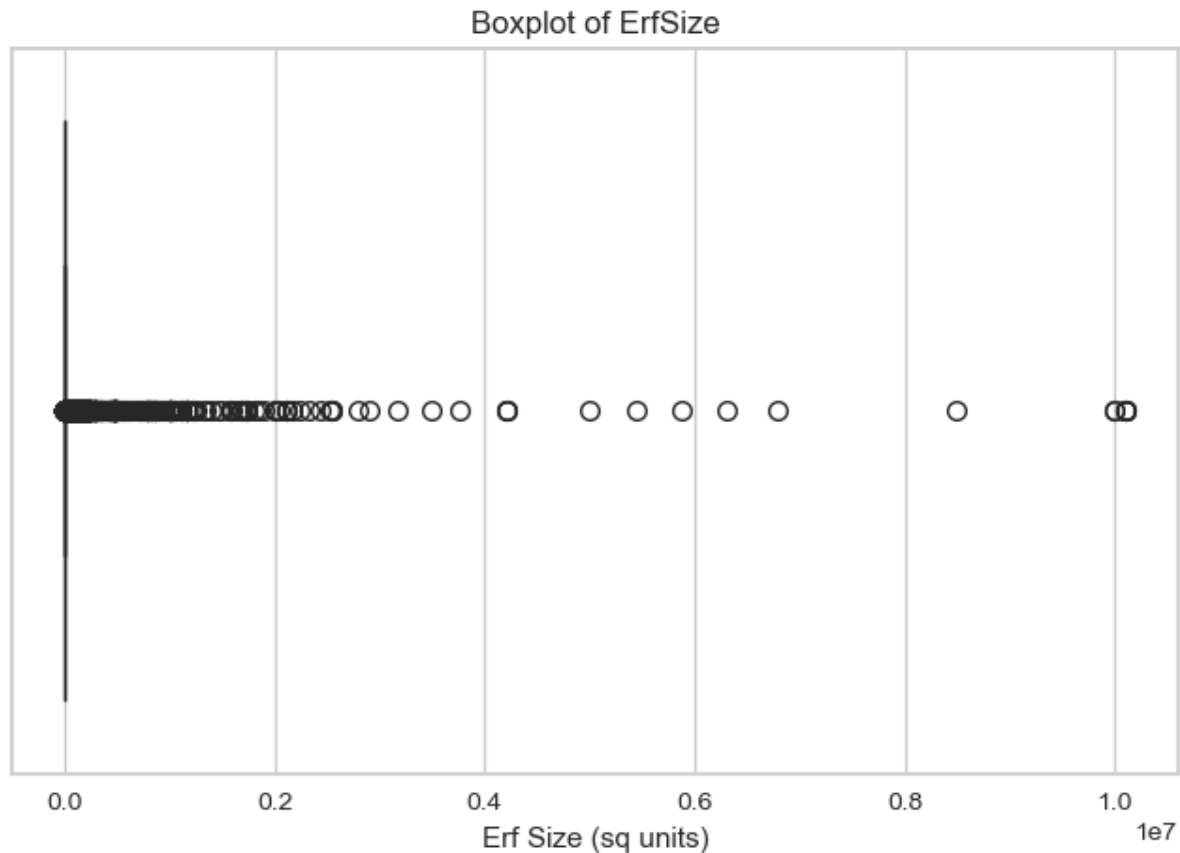


Figure 3.10 provides a visual representation of the distribution of erf sizes, highlighting the skewness and the presence of outliers. The boxplot is right-skewed, with most erf sizes clustered on the lower end and a few large values extending to the right. The plot suggests that most erfs are small, with a few significantly larger properties, indicating a marked difference in erf sizes across the dataset

Key observations from the plot:

- **Median:** Close to 0, indicating that half of the erf sizes are small.
- **Quartiles:** The interquartile range (IQR) spans from 0 to 0.2×10^7 square units, suggesting that 50% of erf sizes fall within this small size range.
- **Outliers:** Many points extend beyond the whiskers, representing large erfs compared to the majority.

Figure 3.11: Frequency distribution of properties listed at different prices in Rands

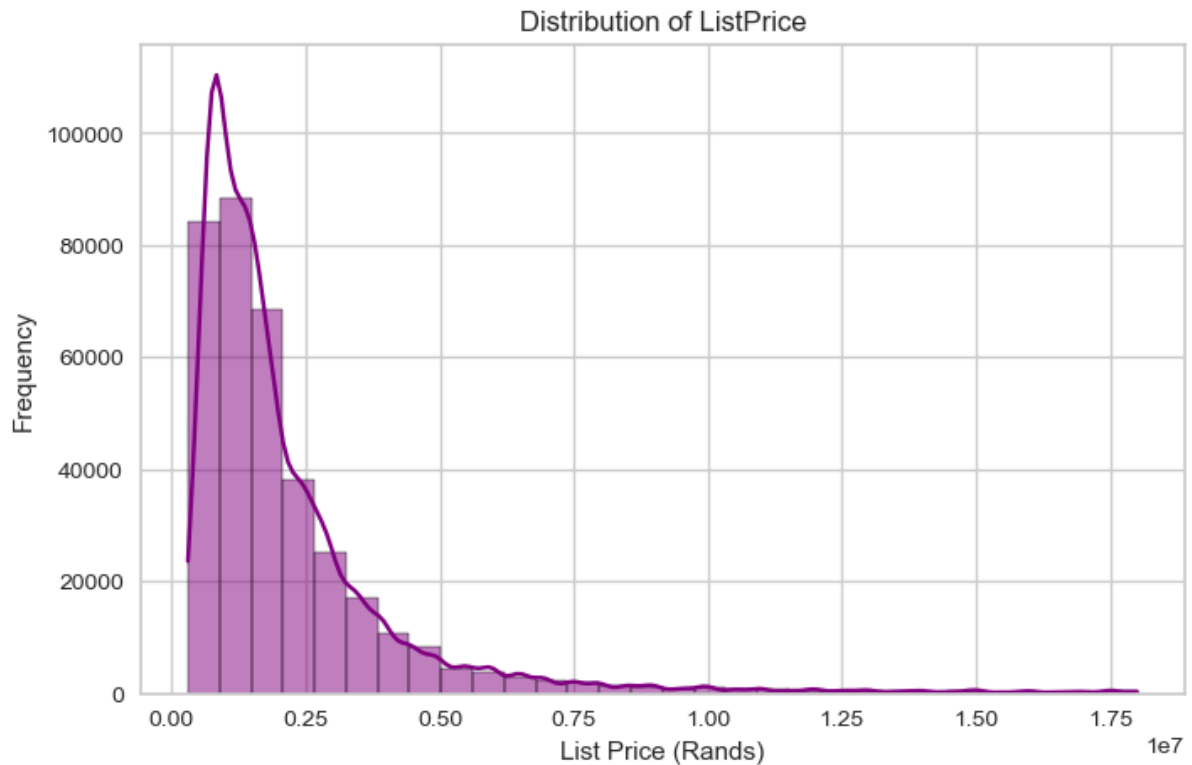


Figure 3.11 shows the frequency of properties listed at different prices in Rands. The histogram is heavily right-skewed, with most properties clustered at the lower end of the price range and a few with very high listing prices.

Key Features:

- **Peak:** The highest frequency of listings is around 0.2×10^7 million Rands, showing many properties in this price range.
- **Long Tail:** Extending to the right, indicating some properties have significantly high prices.

Figure 3.12: Distribution of property listing prices in Rands

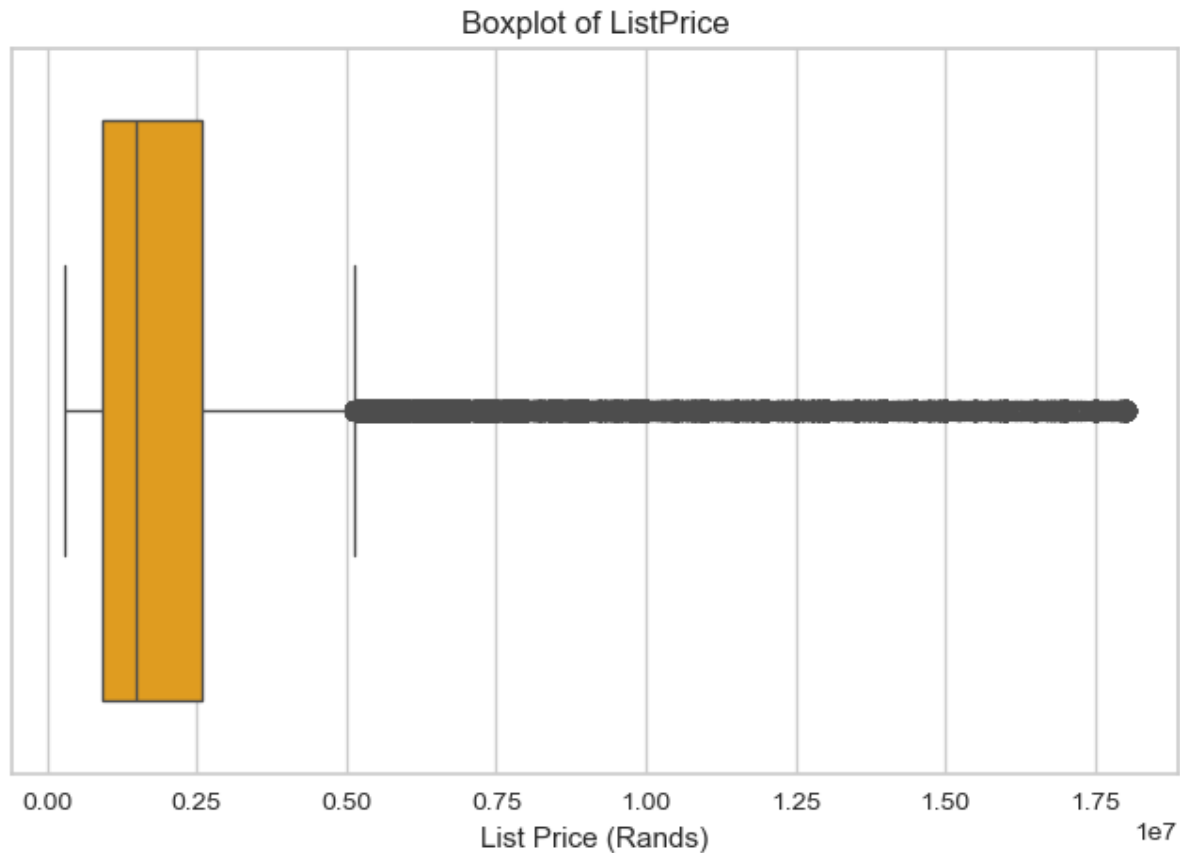


Figure 3.12 shows the distribution of listing prices in Rands. The boxplot is heavily skewed to the right, with most of the data concentrated on the lower end of the price range and a few properties with very high listing prices. The boxplot highlights a concentration of properties at lower prices, with some high-end properties creating a noticeable price difference.

- **Median:** The median listing price is around 0.2 million Rands, meaning half of the properties are listed at or below this price.
- **Quartiles:** The interquartile range (IQR) spans 0.1 to 0.3 million Rands, indicating that 50% of properties are priced within this range.
- **Whiskers:** The lower whisker reaches around 0.05 million Rands, while the upper whisker extends to about 0.6 million Rands.
- **Outliers:** Numerous outliers beyond the upper whisker show properties with significantly higher prices.

Figure 3.13: Distribution of house sizes for different numbers of bedrooms.

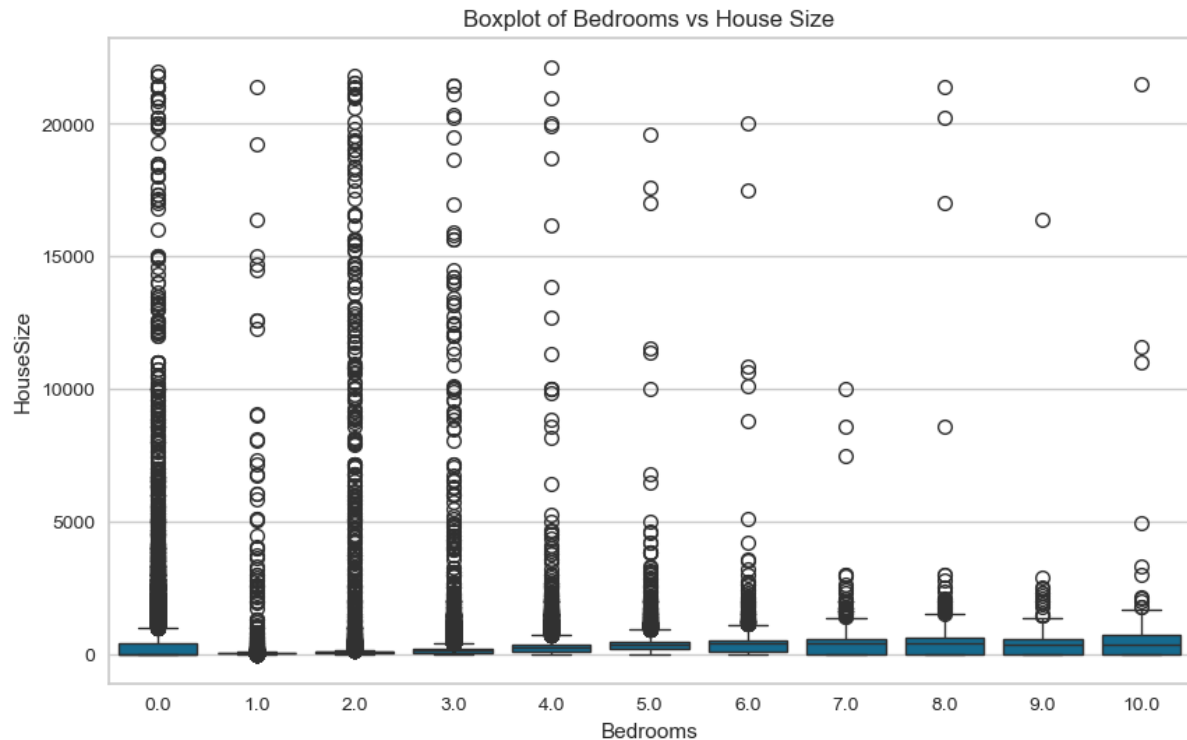


Figure 3.13 shows a strong positive relationship between the number of bedrooms and house size. The variability in house sizes within each category suggests that some houses deviate substantially in size from others with the same bedroom count.

Key Features:

- **Median:** House size median increases as the number of bedrooms rises, showing a trend of larger house sizes with more bedrooms.
- **Quartiles:** The interquartile range (IQR) broadens with more bedrooms, indicating greater size variation among houses with additional bedrooms.
- **Outliers:** Outliers are present, showing some houses are significantly larger or smaller than others within the same bedroom category.

Figure 3.14: Distribution of erf sizes for different numbers of bedrooms.

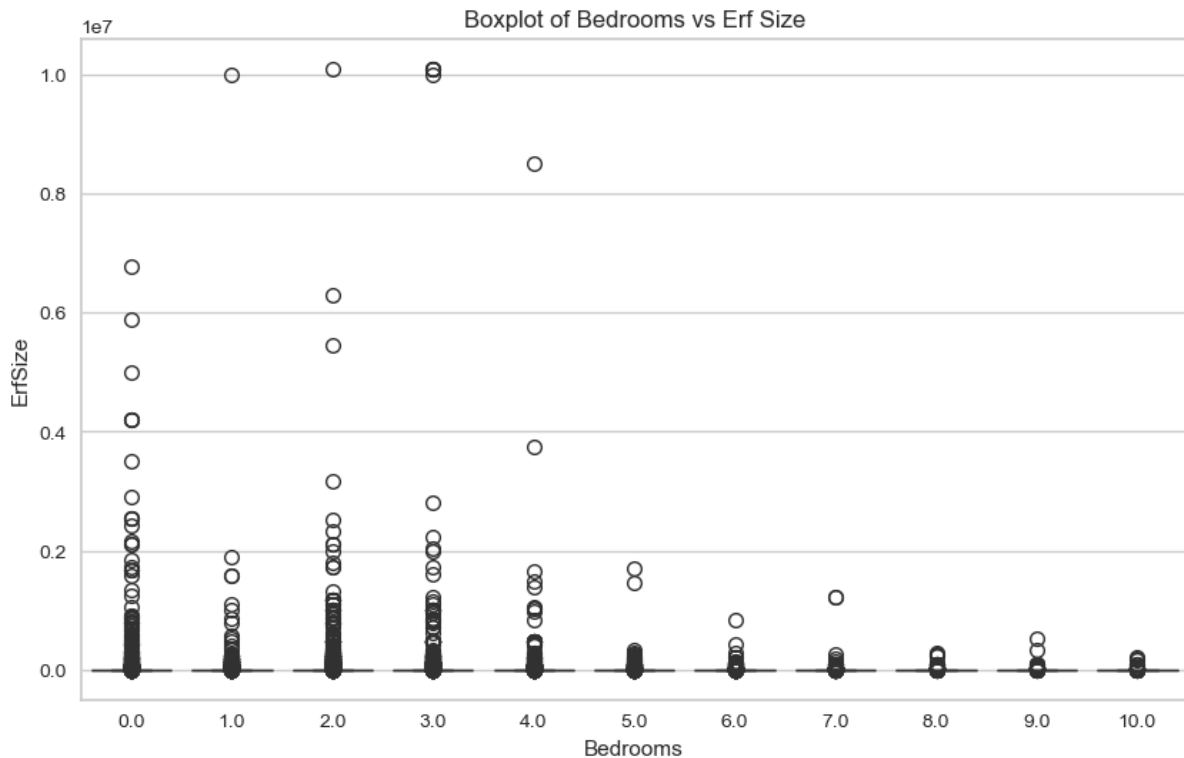


Figure 3.14 shows that there is a strong positive relationship between bedroom count and erf size, with typical erf size increasing as the number of bedrooms grows.

Key Features:

- **Median:** The median erf size rises with more bedrooms, as indicated by the upward shift of the lines within each box.
- **Quartiles:** The interquartile range (IQR) broadens with more bedrooms, showing greater variability in erf sizes.
- **Outliers:** Numerous outliers indicate erfs that are significantly larger or smaller than typical sizes for each bedroom category.

Figure 3.15: Distribution of listing prices for different numbers of bedrooms.

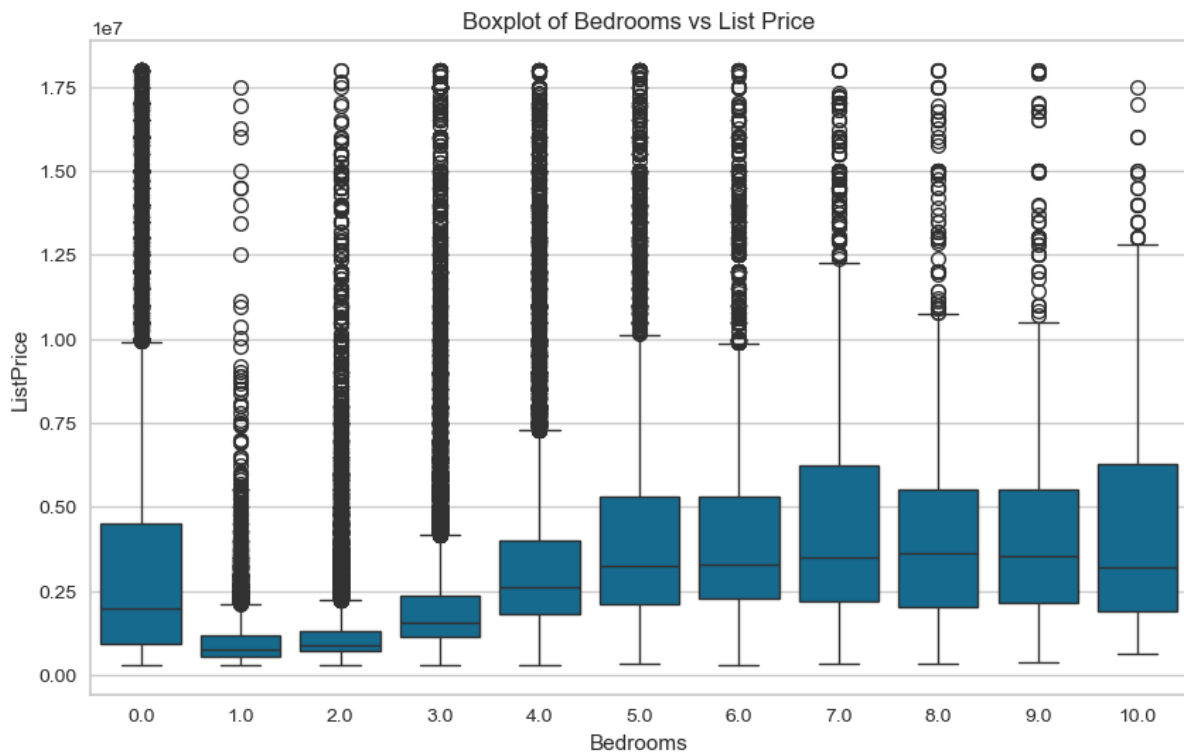


Figure 3.15 shows that there is a strong positive relationship between bedroom count and listing price, with higher number of bedrooms linked to higher prices. Outliers indicate price variability within each bedroom category.

Key Features:

- **Median:** The median listing price increases with more bedrooms, shown by the upward shift of the lines in each box.
- **Quartiles:** The interquartile range (IQR) widens with additional bedrooms, indicating greater price variability.
- **Outliers:** Numerous outliers represent properties with significantly higher or lower prices than typical for each bedroom category.

Figure 3.16: Distribution of house sizes for different numbers of bathrooms.

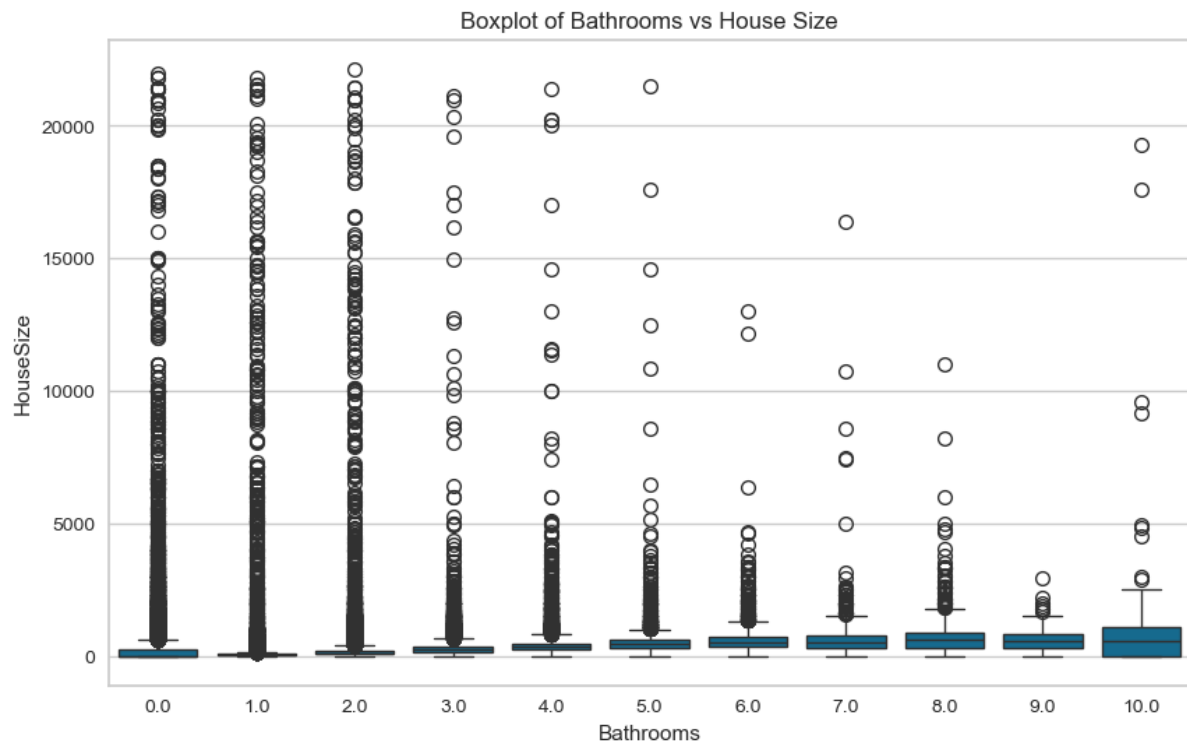


Figure 3.16 demonstrates a strong positive correlation between the number of bathrooms and house size, suggesting that more bathrooms generally mean larger houses. Additionally, the presence of outliers indicates variability in house sizes within each bathroom category, with some houses differing significantly in size.

Key Features:

- **Median:** The median house size rises with an increasing number of bathrooms, indicated by the upward shift of the lines within the boxes.
- **Quartiles:** The interquartile range (IQR) also grows as the number of bathrooms increases, shown by the widening of the boxes.
- **Outliers:** Numerous outlier points appear beyond the whiskers, indicating houses that are much larger or smaller compared to others with the same number of bathrooms.

Figure 3.17: Distribution of erf sizes for different numbers of bathrooms.

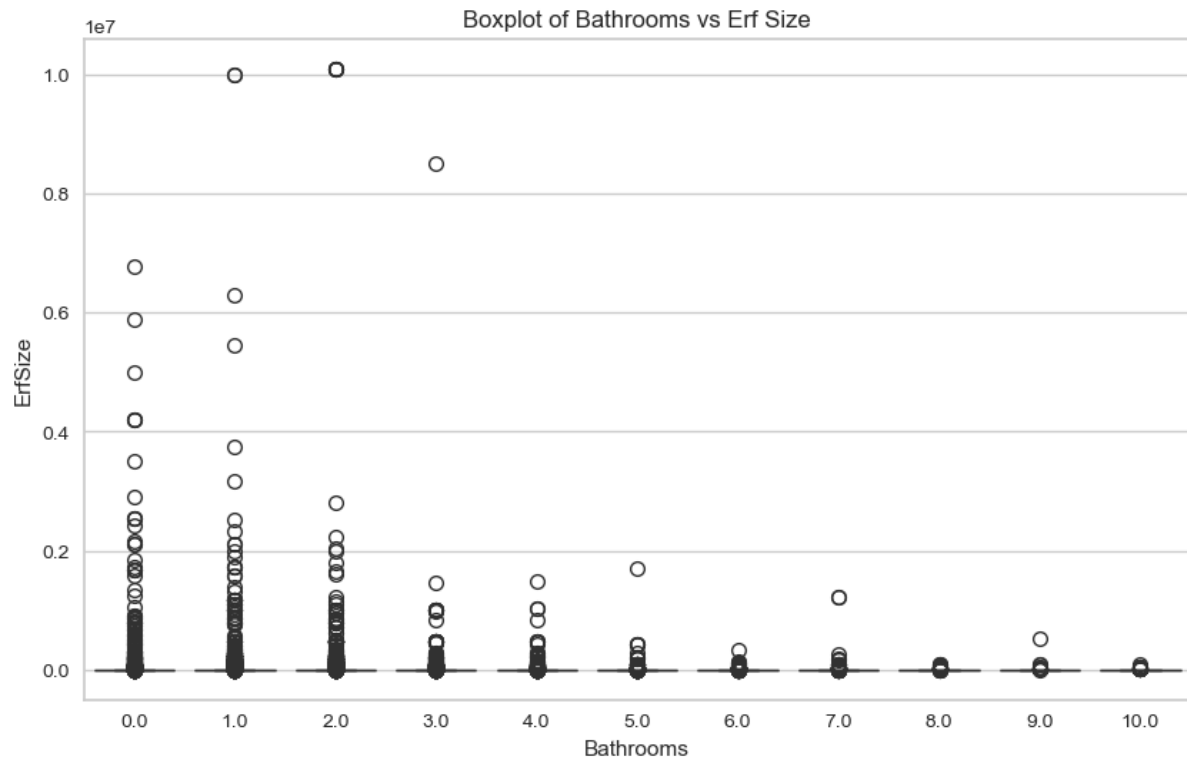


Figure 3.17 demonstrates a strong positive correlation between the number of bathrooms and erf size, suggesting that more bathrooms generally correspond to larger erfs. The presence of outliers indicates variability in erf sizes within each bathroom category, with some erfs differing significantly in size.

Key Features:

- **Median:** The median erf size increases with the number of bathrooms, indicated by the upward shift of the lines within the boxes.
- **Quartiles:** The interquartile range (IQR) also increases with the number of bathrooms, as shown by the widening of the boxes.
- **Outliers:** Numerous outlier points beyond the whiskers indicate erfs that are significantly larger or smaller than others with the same number of bathrooms.

Figure 3.18: Distribution of listing prices for different numbers of bathrooms.

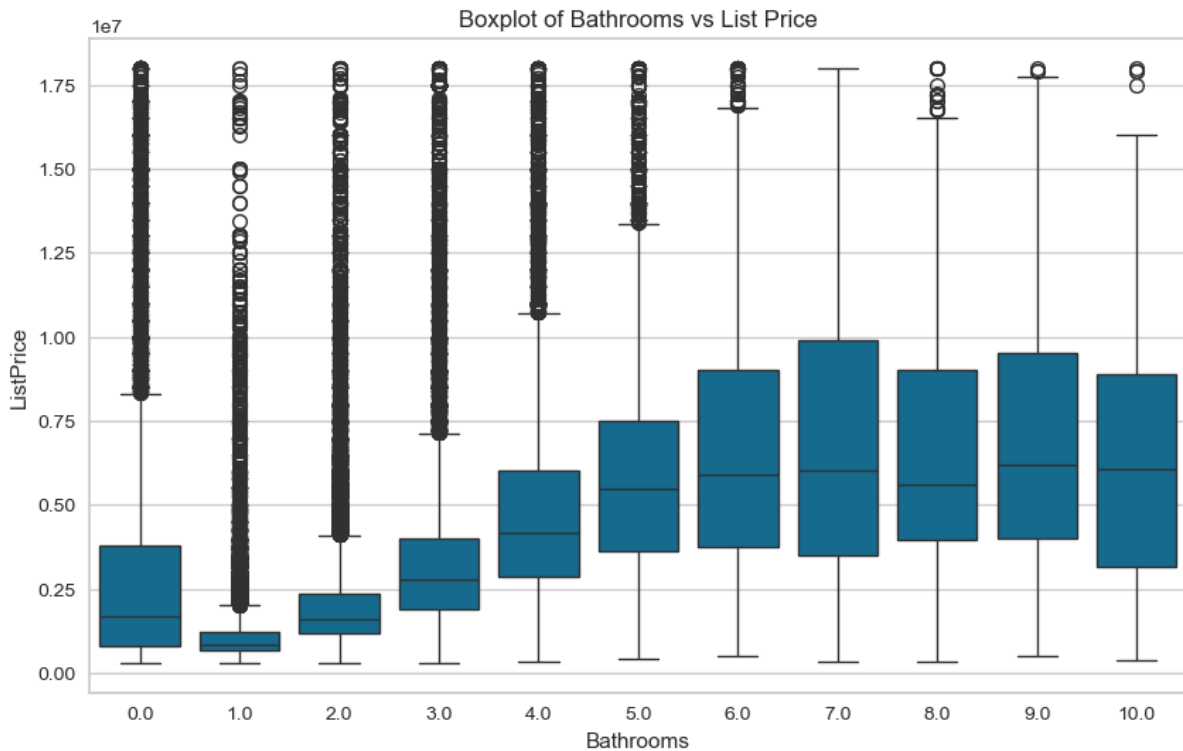


Figure 3.18 indicates a strong positive correlation between the number of bathrooms and listing price, suggesting that properties with more bathrooms generally have higher listing prices. The presence of outliers highlights variability in listing prices within each bathroom category, with some properties being significantly more or less expensive than others in the same category.

Key Features:

- **Median:** The median listing price rises with the number of bathrooms, as indicated by the upward shift of the lines within the boxes.
- **Quartiles:** The interquartile range (IQR) increases with the number of bathrooms, shown by the widening of the boxes.
- **Outliers:** Numerous outlier points beyond the whiskers represent properties with significantly higher or lower listing prices compared to others with the same number of bathrooms.

Figure 3.19: Distribution of house sizes across different suburbs.

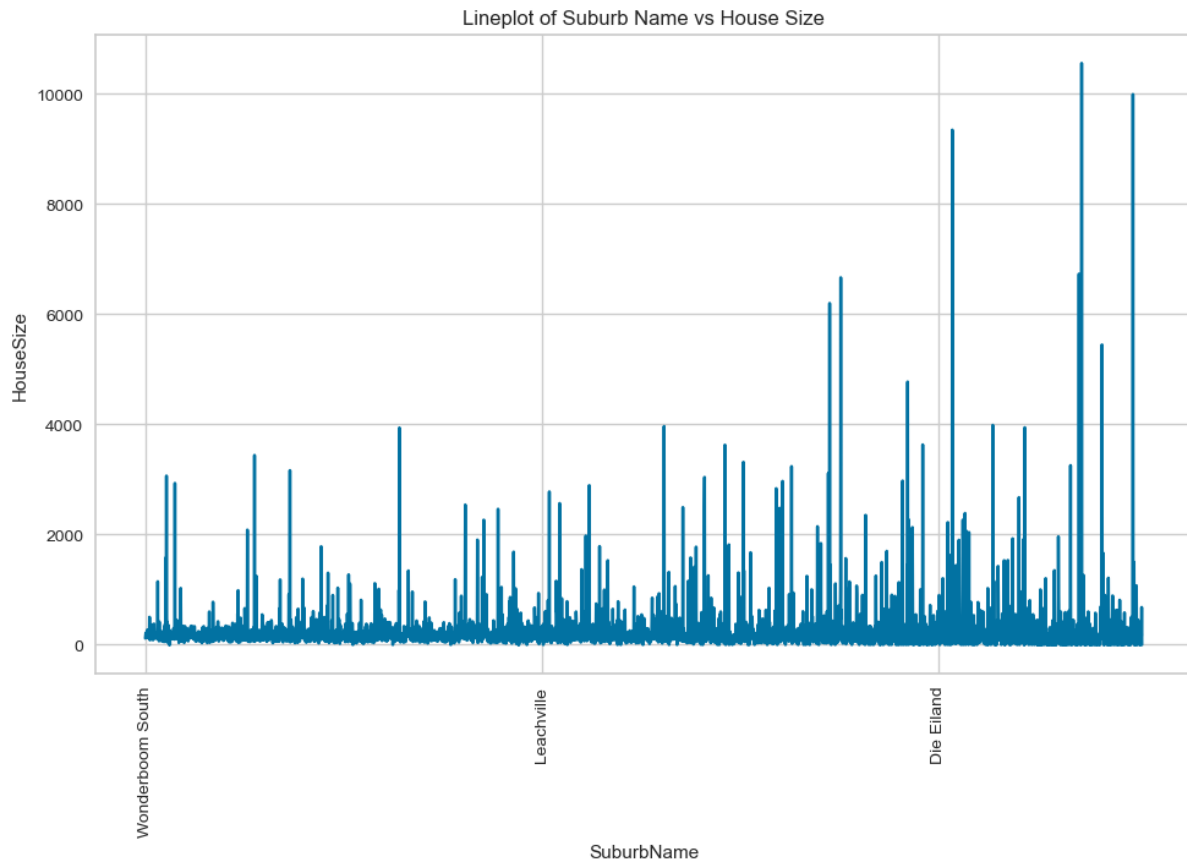


Figure 3.19 shows that house sizes vary significantly between suburbs, with some suburbs exhibiting a higher concentration of larger homes and others favoring smaller ones. The presence of outliers indicates that certain suburbs may have a few exceptionally large or small houses, likely influenced by factors such as zoning regulations, historical development, or local market dynamics.

Key Features:

- **Variability:** There is considerable variation in house sizes among different suburbs, with some suburbs featuring many large houses and others having many smaller ones.
- **Outliers:** Outliers are evident as lines that are significantly taller or shorter than others, indicating suburbs with a few very large or very small houses.

Figure 3.20: Distribution of erf sizes across different suburbs.

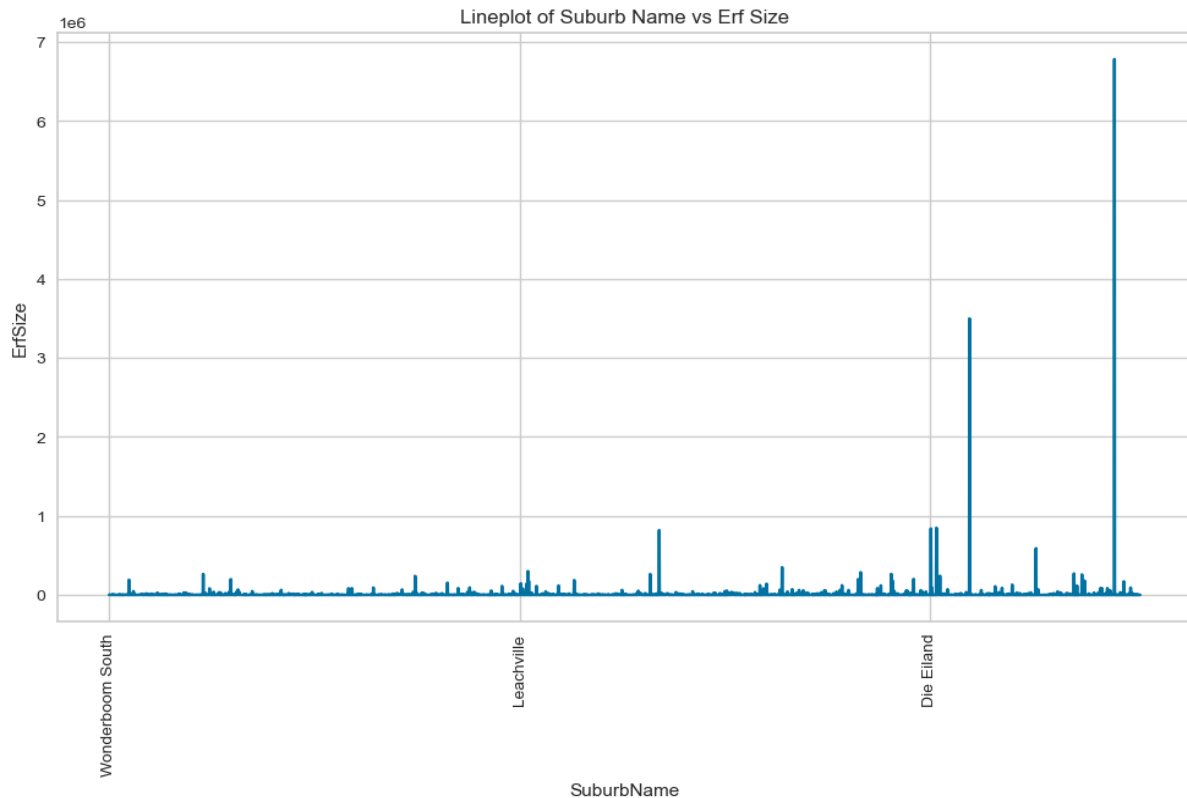


Figure 3.20 demonstrates that erf sizes can vary significantly between suburbs, with some areas having a higher concentration of larger erfs and others favoring smaller ones. The presence of outliers suggests that certain suburbs may have exceptionally large or small erfs, potentially influenced by factors such as zoning regulations, historical development, or local market dynamics.

Key Features:

- **Variability:** There is considerable variation in erf sizes among different suburbs, with some suburbs showcasing many large erfs and others having many smaller ones.
- **Outliers:** Outliers appear as lines significantly taller or shorter than others, indicating suburbs with a few very large or very small erfs.

Figure 3.21: Distribution of listing prices across different suburbs.

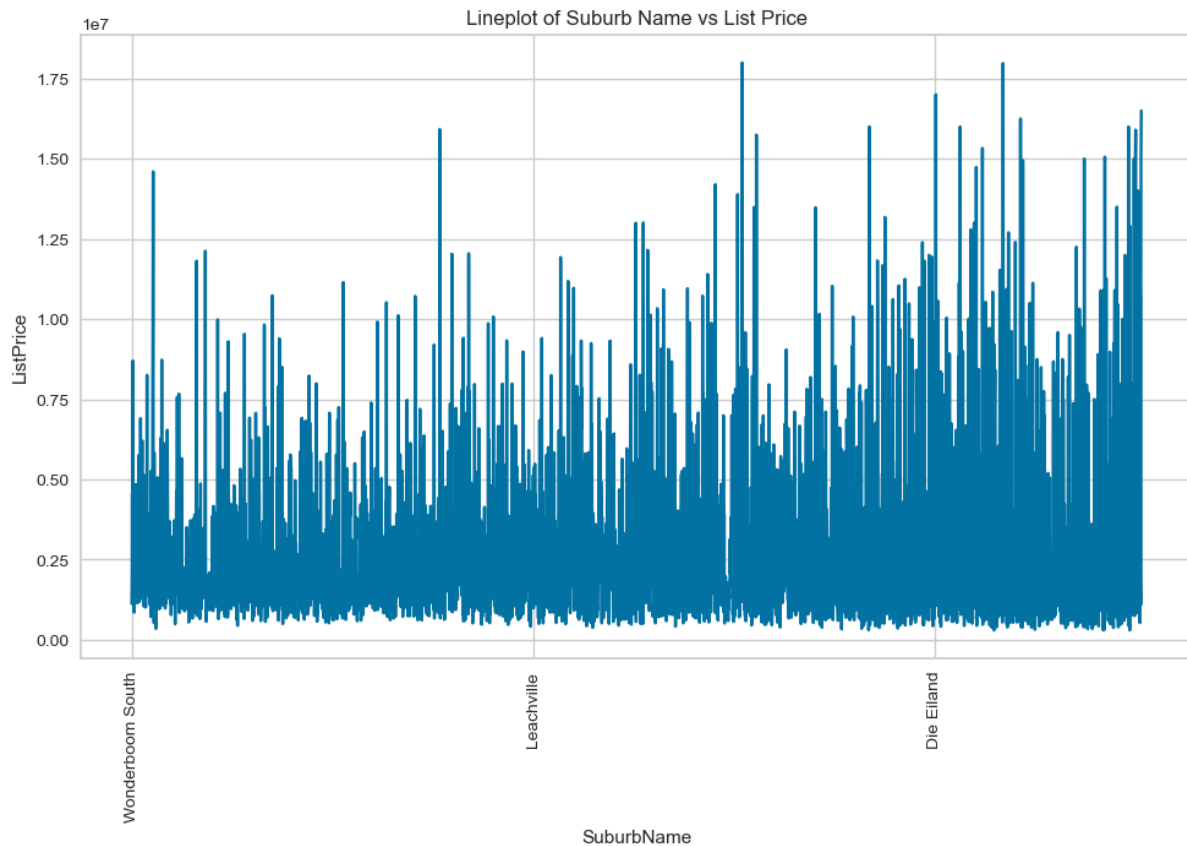


Figure 3.21 demonstrates that erf sizes can vary significantly between suburbs, with some areas having a higher concentration of larger erfs and others favoring smaller ones. The presence of outliers suggests that certain suburbs may have exceptionally large or small erfs, potentially influenced by factors such as zoning regulations, historical development, or local market dynamics.

Key Features:

- **Variability:** There is considerable variation in erf sizes among different suburbs, with some suburbs showcasing many large erfs and others having many smaller ones.
- **Outliers:** Outliers appear as lines significantly taller or shorter than others, indicating suburbs with a few very large or very small erfs.

Figure 3.22: Distribution of house sizes across two provinces: Gauteng and Western Cape.

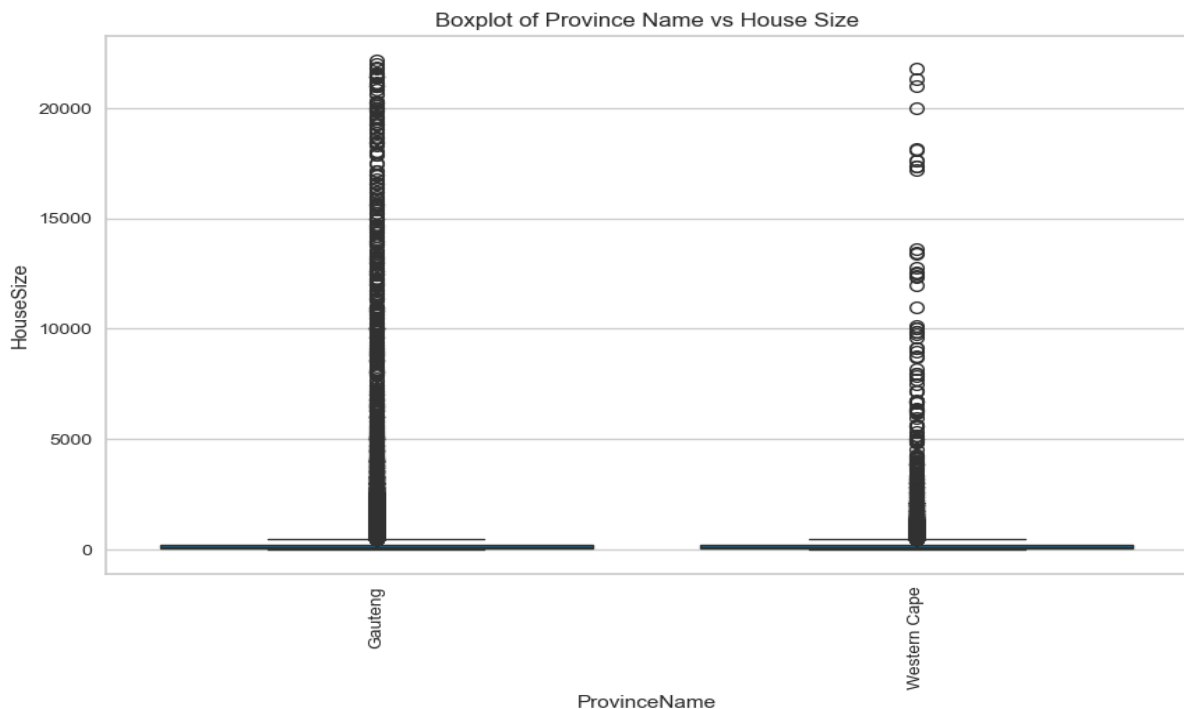


Figure 3.22 highlights significant differences in house size distributions between Gauteng and the Western Cape. Gauteng has a higher median house size and greater variability, indicating a prevalence of larger houses. In contrast, the Western Cape features a lower median house size and less variability, suggesting that smaller houses are more common in this province.

Key Features:

Gauteng:

- The median house size is significantly higher than in the Western Cape.
- The interquartile range (IQR) is larger, indicating greater variability in house sizes.
- Numerous outliers represent houses that are significantly larger than the majority.

Western Cape:

- The median house size is lower compared to Gauteng.
- The IQR is smaller, indicating less variability in house sizes.
- Fewer outliers suggest house sizes are more concentrated around the median.

Figure 3.23: Distribution of erf sizes across two provinces: Gauteng and Western Cape.

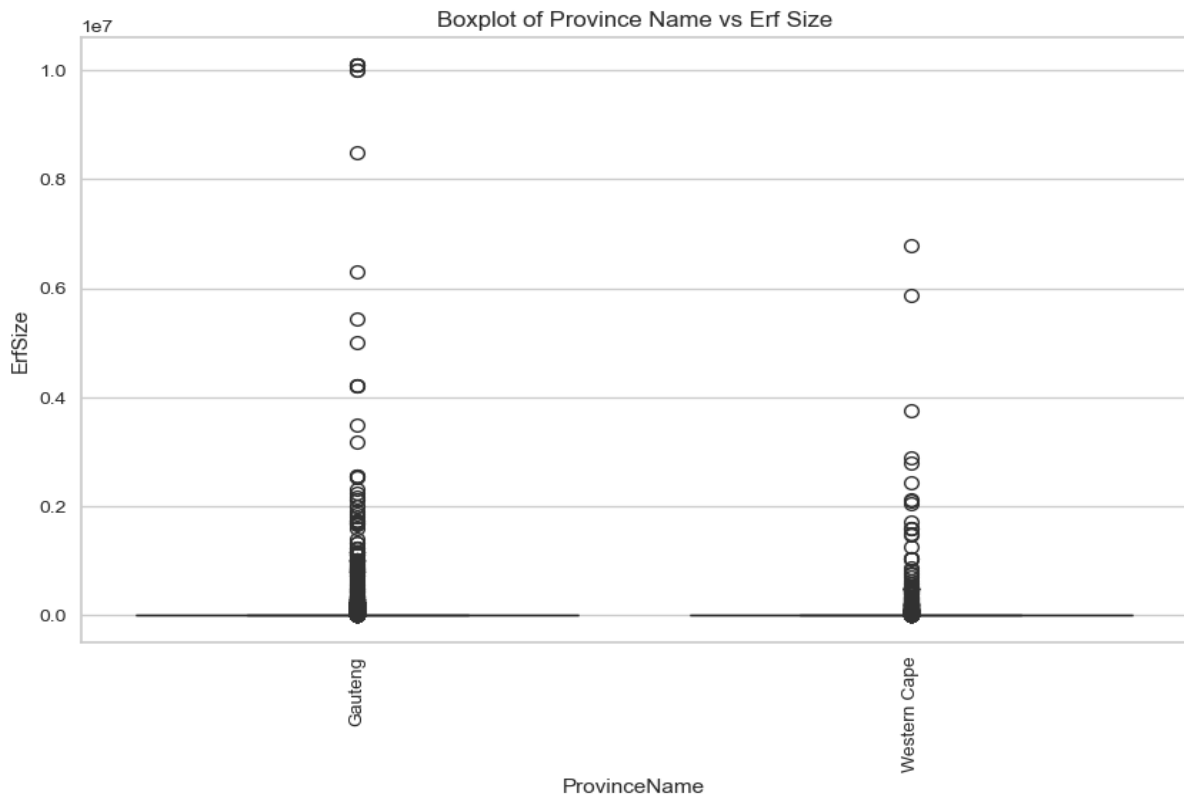


Figure 3.23 illustrates significant differences in erf size distributions between Gauteng and the Western Cape. Gauteng features a higher median erf size and greater variability, indicating a prevalence of larger erfs than Western Cape.

Key Features:

Gauteng:

- The median erf size of Gauteng is significantly higher than in the Western Cape.
- The interquartile range (IQR) is larger, indicating greater variability in erf sizes.
- Numerous outliers represent erfs that are significantly larger than the majority.

Western Cape:

- The median erf size is lower compared to Gauteng.
- The IQR is smaller, indicating less variability in erf sizes.
- Fewer outliers suggest that erf sizes are more concentrated around the median.

Figure 3.24: Distribution of listing prices across two provinces: Gauteng and Western Cape.

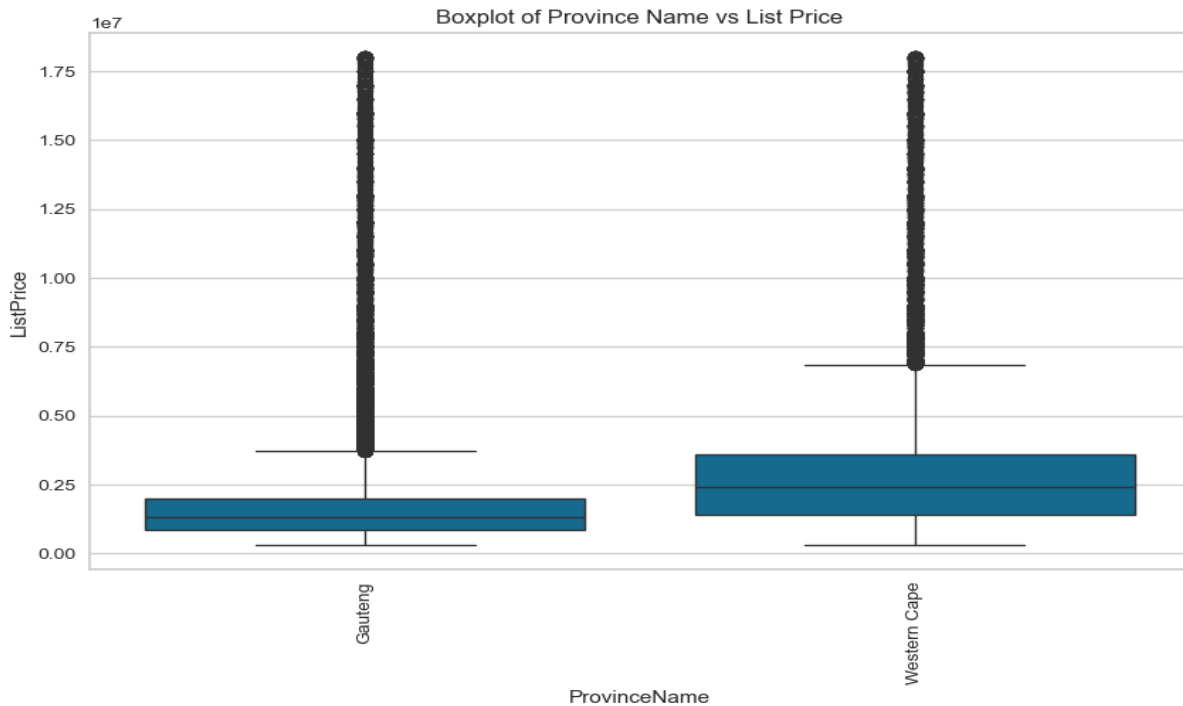


Figure 3.24 demonstrates significant differences in listing price distributions between Gauteng and the Western Cape. Gauteng has a higher median listing price and greater variability, indicating a prevalence of more expensive properties. In contrast, the Western Cape features a lower median listing price and less variability, suggesting that more affordable properties are more common in this province.

Key Features:

Gauteng:

- The median listing price is significantly higher than in the Western Cape.
- The interquartile range (IQR) is larger, reflecting greater variability in listing prices.
- Numerous outliers represent properties with significantly higher listing prices compared to the majority.

Western Cape:

- The median listing price is lower compared to Gauteng.
- The IQR is smaller, indicating less variability in listing prices.
- Fewer outliers suggest that listing prices are more concentrated around the median.

Figure 3.25: Distribution of house sizes for different property types.

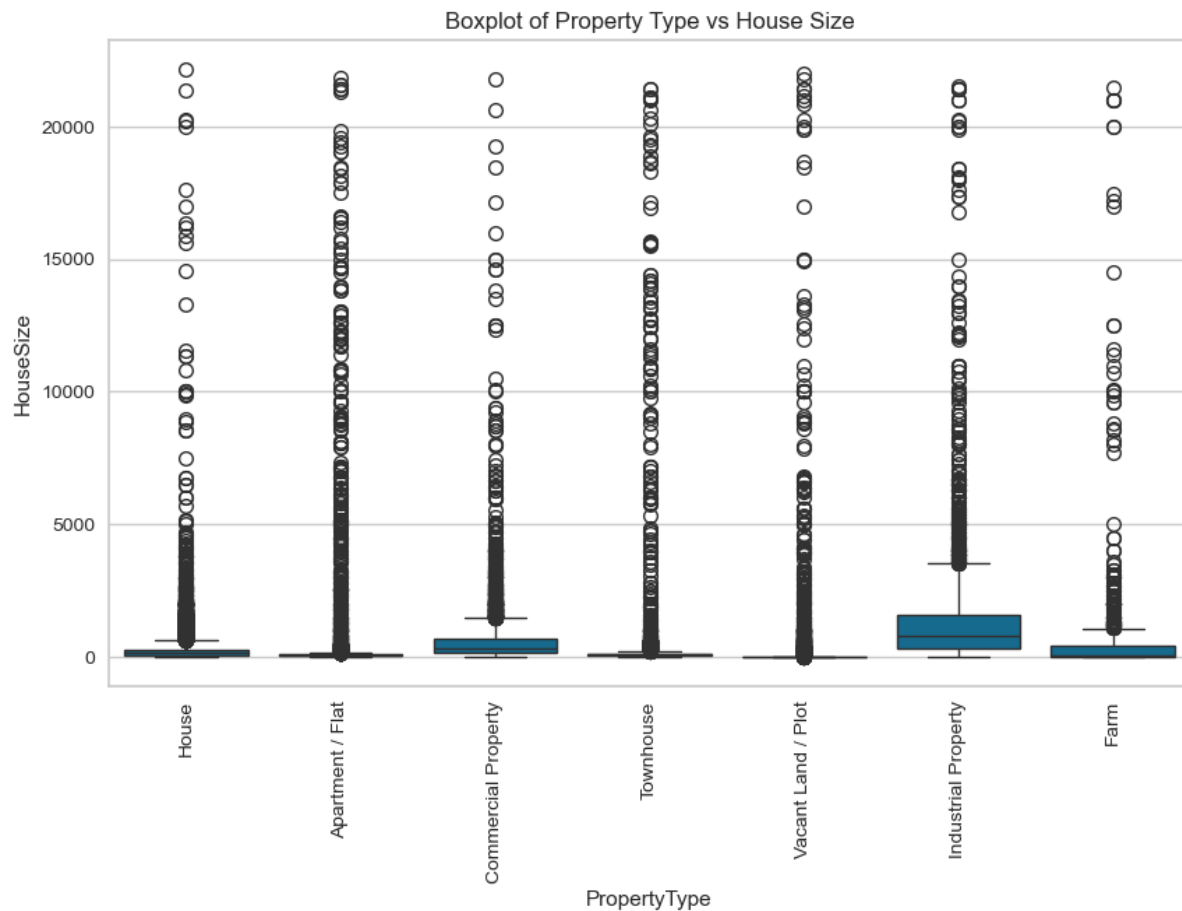


Figure 3.25 illustrates significant variations in house sizes across different property types. Houses and farms typically have larger sizes, while apartments/flats, commercial properties, and vacant land/plots feature smaller sizes. The presence of outliers indicates variability within each property type, with some properties being significantly larger or smaller than others in the same category.

Key Features:

- **Houses:** Largest median house size with a relatively large interquartile range (IQR), indicating significant variability in sizes.
- **Apartments/Flats:** Smaller median house size compared to houses with a smaller IQR, suggesting less variability.
- **Commercial Property:** Very small median house size and small IQR, indicating similar sizes among properties.

- **Townhouses:** Median house size smaller than houses but larger than apartments/flats, with moderate IQR reflecting some variability.
- **Vacant Land/Plot:** Very small median house size, with a small IQR indicating similar sizes.
- **Industrial Property:** Slightly larger median house size compared to vacant land/plots, with a slightly larger IQR showing some variability.
- **Farms:** Largest median house size among all property types, with a large IQR indicating significant variability.

Figure 3.26: Distribution of erf sizes for different property types.

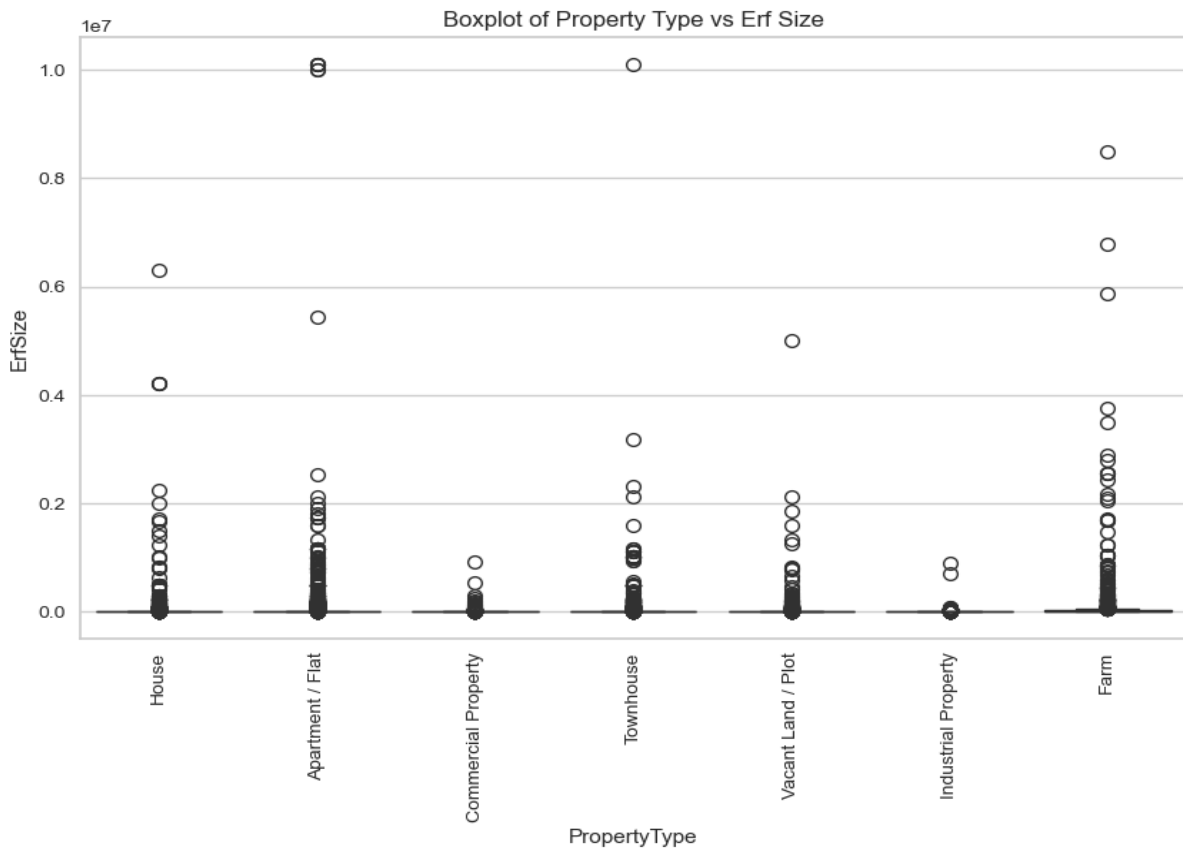


Figure 3.26 shows significant variations in erf sizes across different property types. Houses and farms generally have larger erf sizes, while apartments/flats, commercial properties, and vacant land/plots tend to have smaller sizes. The presence of outliers indicates variability within each property type, with some properties having significantly larger or smaller erfs than others in the same category.

Key Features:

- **Houses:** Larger median erf size compared to other property types, with a relatively large interquartile range (IQR) indicating significant variability.
- **Apartments/Flats:** Smaller median erf size than houses, with a smaller IQR, suggesting less variability.
- **Commercial Property:** Very small median erf size and small IQR, indicating similar sizes among properties.
- **Townhouses:** Median erf size is smaller than houses but larger than apartments/flats, with a moderate IQR indicating some variability.
- **Vacant Land/Plot:** Very small median erf size, with a small IQR suggesting similar sizes.
- **Industrial Property:** Slightly larger median erf size compared to vacant land/plots, with a slightly larger IQR indicating some variability.
- **Farms:** Largest median erf size among all property types, with a large IQR indicating significant variability.

Figure 3.27: Distribution of listing prices for different property types.

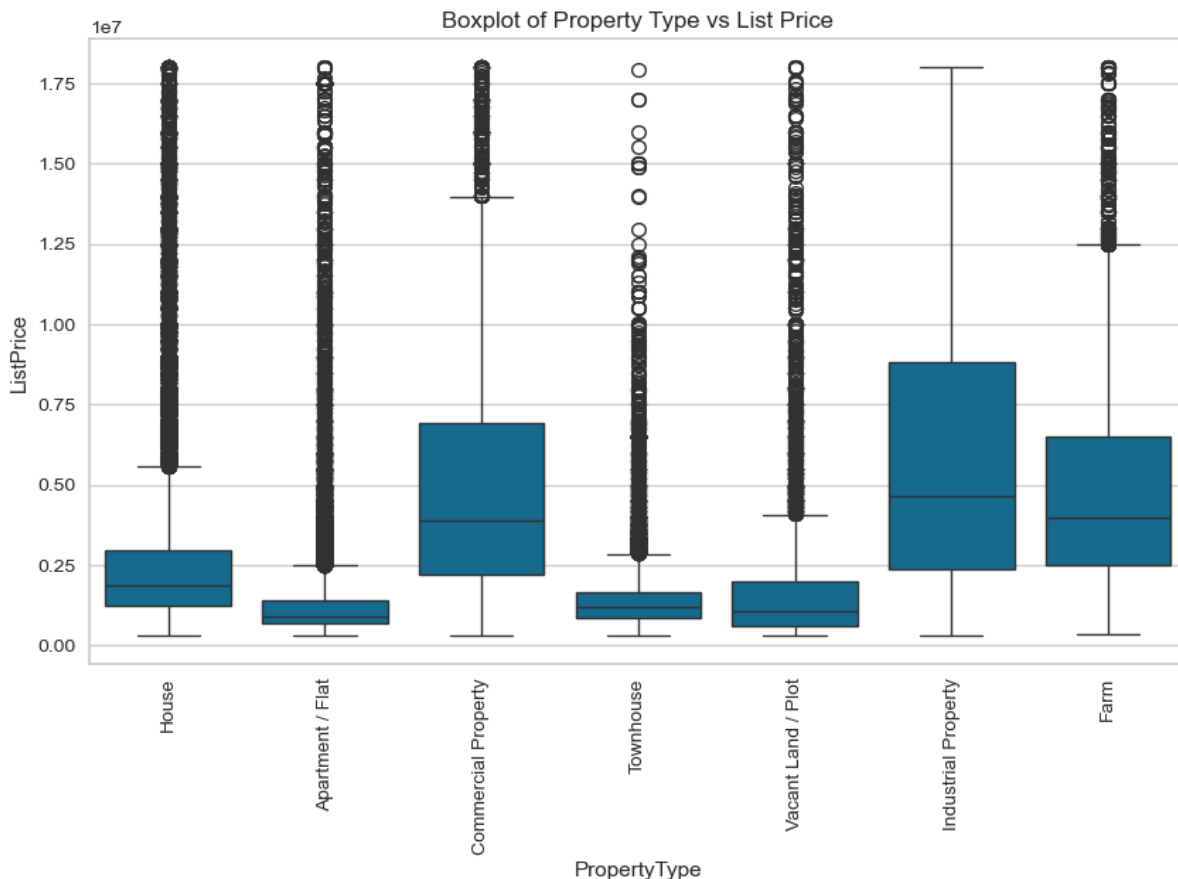


Figure 3.27 illustrates significant variations in listing prices across different property types. Commercial property, Industrial Property and farms generally have higher listing prices, while apartments/flats, house, townhouse and vacant land/plots tend to have lower prices. The presence of outliers indicates variability within each property type, with some properties being significantly more or less expensive than others in the same category.

Key Features:

- **Houses:** Largest median listing price among property types, with a large interquartile range (IQR) indicating significant variability.
- **Apartments/Flats:** Smaller median listing price than houses and a smaller IQR, suggesting less price variability.
- **Commercial Property:** Smaller median listing price than houses, with a small IQR indicating low variability.
- **Townhouses:** Median listing price smaller than houses but larger than apartments/flats, with a moderate IQR indicating some variability.

- **Vacant Land/Plot:** Very small median listing price and a small IQR, suggesting similar low prices.
- **Industrial Property:** Larger median listing price compared to vacant land/plots, with a larger IQR indicating some variability.
- **Farms:** Largest median listing price of all property types, with a large IQR indicating significant variability.

Figure 3.28: Distribution of house sizes across different cities.

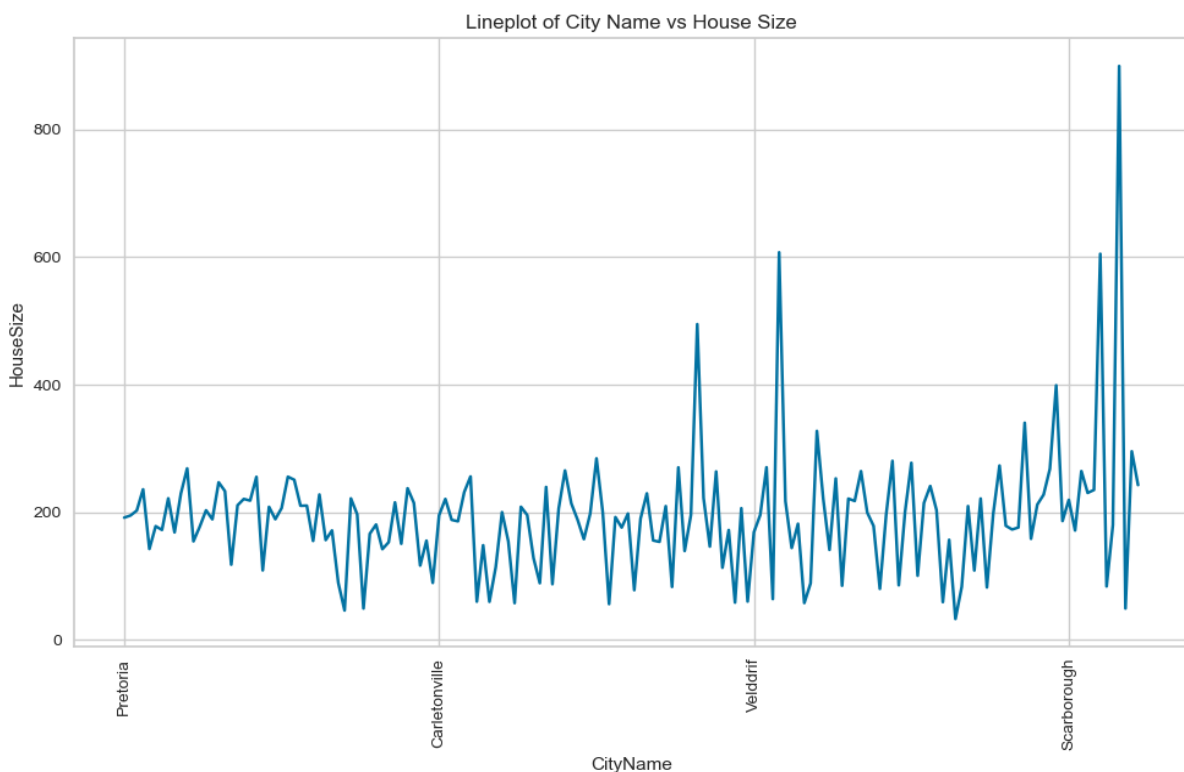


Figure 3.28 demonstrates substantial differences in house sizes among cities. Certain cities may have a higher concentration of larger houses, while others focus more on smaller houses. The outliers suggest that specific cities may contain exceptionally large or small houses, potentially influenced by factors like location, amenities, or local market conditions.

Key Features:

- **Variability:** There is significant variation in house sizes across cities, with some cities having many large houses while others feature smaller ones.
- **Outliers:** Some cities exhibit outlier lines that are much taller or shorter than the rest, indicating the presence of a few very large or very small houses.

Figure 3.29: Distribution of erf sizes across different cities.

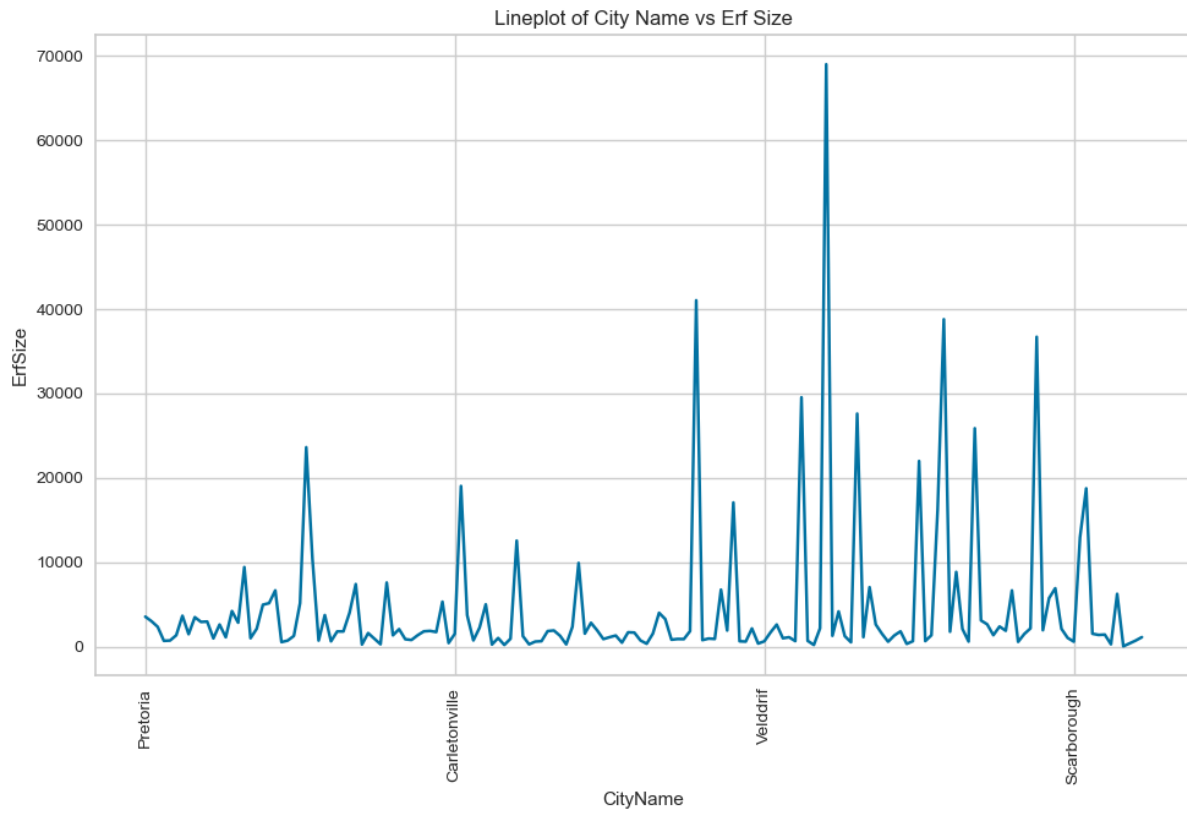


Figure 3.29 illustrates considerable differences in erf sizes among cities. Certain cities may feature a higher concentration of larger erfs, while others focus on smaller erfs. The outliers suggest that specific cities may contain exceptionally large or small erfs, likely influenced by factors such as location, amenities, or local market dynamics.

Key Features:

- **Variability:** There is substantial variation in erf sizes across cities, with some cities showing a high concentration of large erfs while others have more small erfs.
- **Outliers:** Some cities have outlier lines that are significantly taller or shorter than others, indicating the presence of a few very large or very small erfs.

Figure 3.30: Distribution of List Price across different cities.

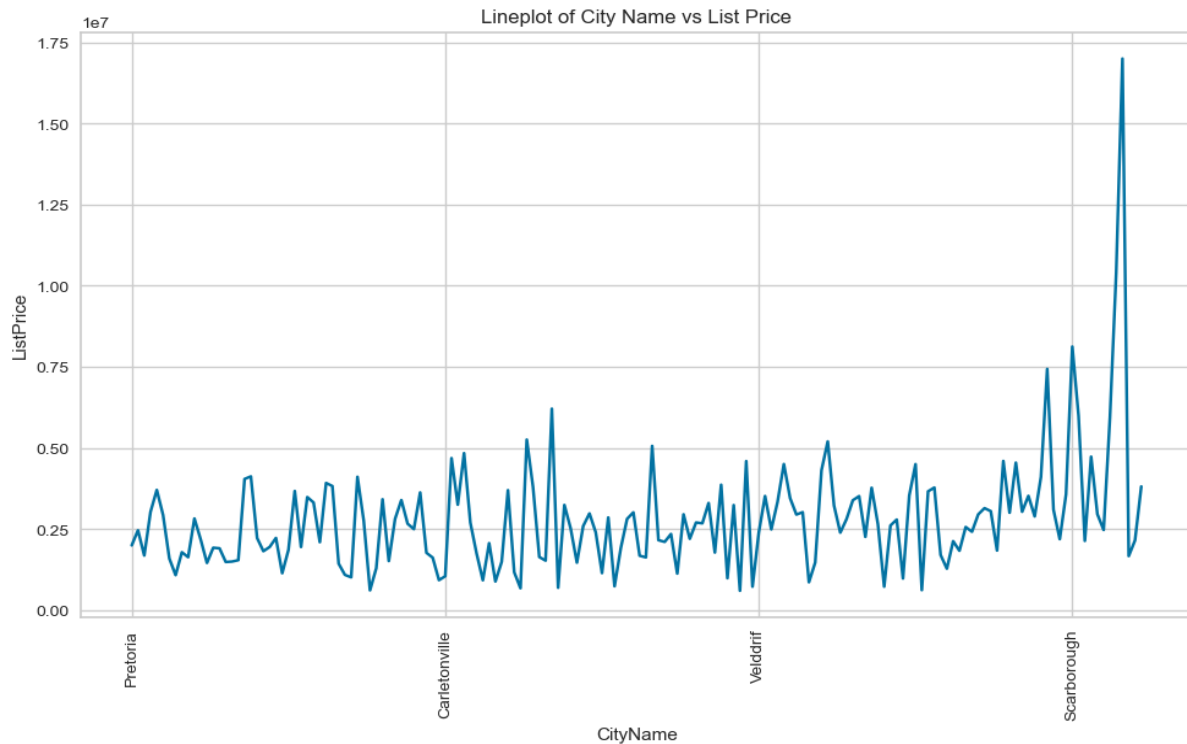


Figure 3.30 highlights considerable differences in listing prices between cities. Some cities have a higher concentration of expensive properties, while others have more affordable options. The presence of outliers suggests that some cities may include exceptionally high or low-priced properties, potentially influenced by factors such as location, amenities, or local market conditions.

Key Features:

- **Variability:** There is significant variation in listing prices among cities, with some cities having a high concentration of expensive properties and others featuring more affordable ones.
- **Outliers:** Certain cities exhibit outlier lines that are much taller or shorter than others, indicating the presence of a few very expensive or very cheap properties.

Figure 3.31: Pair Plot.

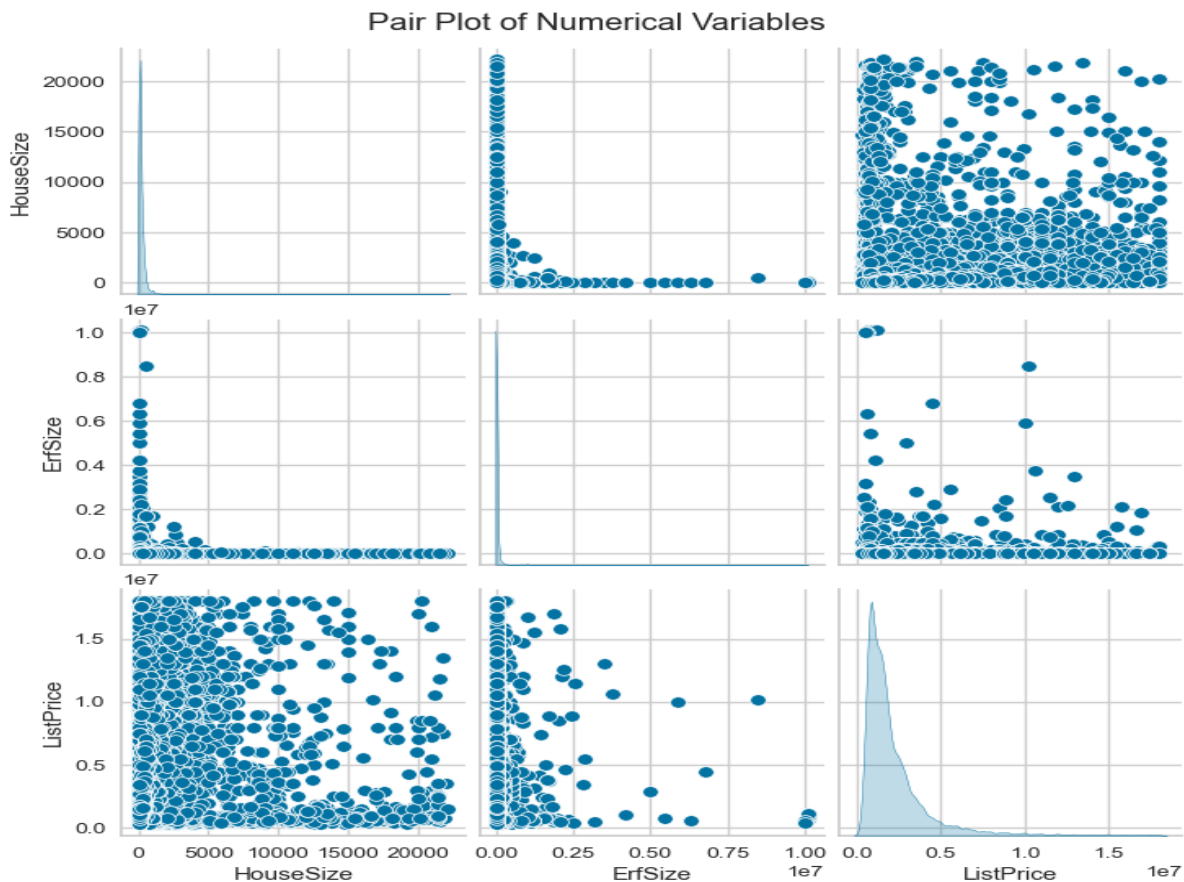


Figure 3.31 indicates some relationships among the numerical variables, though they are not particularly strong. The strongest correlation appears between house size and list price. The presence of outliers suggests that certain properties possess unique characteristics that deviate from the general trends observed in the data.

Key Features:

- **House Size vs. Erf Size:** There is a weak positive correlation, suggesting that larger erf sizes are somewhat associated with larger houses. However, some outliers exist where very large erf sizes correspond to relatively small houses.
- **House Size vs. List Price:** A moderate positive correlation indicates that as house size increases, list price generally tends to increase as well. Notably, some outliers show large house sizes with lower list prices.
- **Erf Size vs. List Price:** A weak positive relationship is present, indicating a slight increase in list price with larger erf sizes. Again, there are outliers where large erf sizes are paired with lower list prices.

3.3 Correlation Analysis

Figure 3.32: Correlation matrix

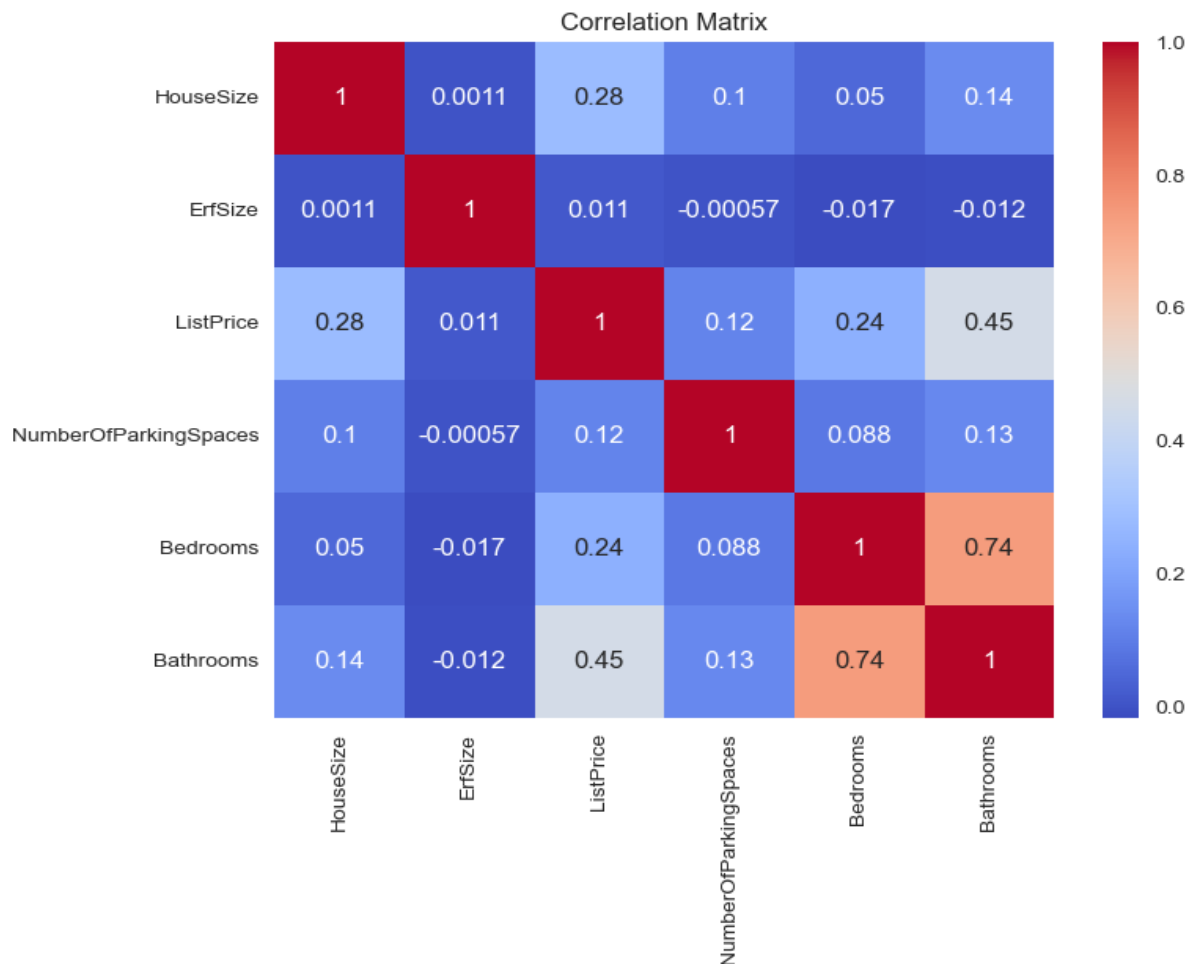


Figure 3.32 highlights moderate positive correlations between certain variables, particularly house size and list price, as well as bedrooms and bathrooms. However, most correlations are weak, indicating that the relationships between the variables are not strong.

Key Features:

House Size, Erf Size, and List Price:

- A moderate positive correlation (0.28) between house size and list price suggests that larger houses generally have higher listing prices.
- A weak positive correlation (0.12) between erf size and list price indicates that larger erfs tend to be associated with slightly higher prices.

- A very weak correlation (0.0011) between house size and erf size shows nearly no linear relationship between these two variables.

Bedrooms and Bathrooms:

- A strong positive correlation (0.74) suggests that properties with more bedrooms typically have more bathrooms.

Other Variables:

- Correlation coefficients for additional variables, such as the number of parking spaces, are generally weak and not statistically significant.

3.4 K-Means Clustering

After running K-Means clustering, each property listing was assigned a cluster label, grouping similar properties based on features like bedrooms, bathrooms, and location. The Elbow Method was employed to select an optimal number of clusters, with K=5 chosen as the best balance between within-cluster variance and computational efficiency.

Figure 3.33: Elbow Method plot.

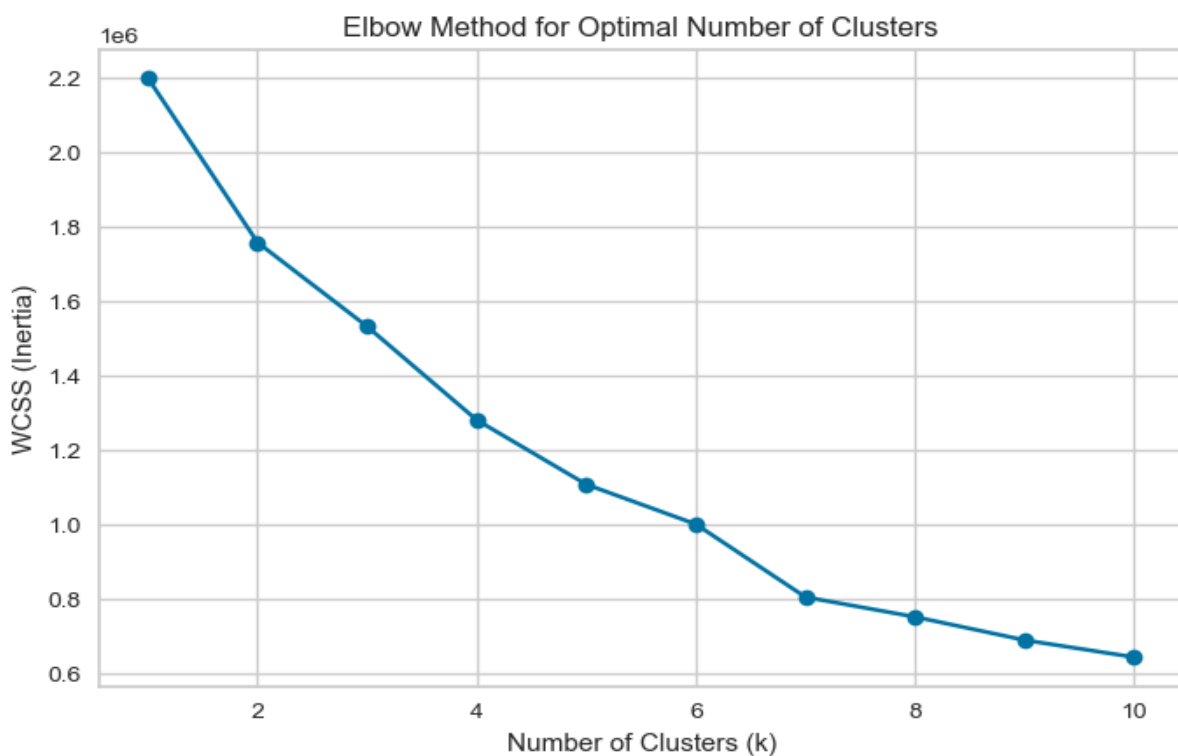


Figure 3.33 shows an elbow plot using 5 clusters balances reduced within-cluster variation with a manageable number of clusters. Adding more clusters beyond 5 may yield minimal clustering improvement and risks overfitting.

Summary:

- **WCSS:** The y-axis shows Within-Cluster Sum of Squares, where lower values indicate better clustering.
- **Number of Clusters (k):** The x-axis displays the number of clusters, with WCSS generally decreasing as k increases.
- **Elbow Point:** Around $k = 5$, the WCSS decrease rate levels off, suggesting 5 clusters as a good balance between cluster quality and quantity.

Figure 3.34: The scatter plot displays List Price versus House Size Clusters

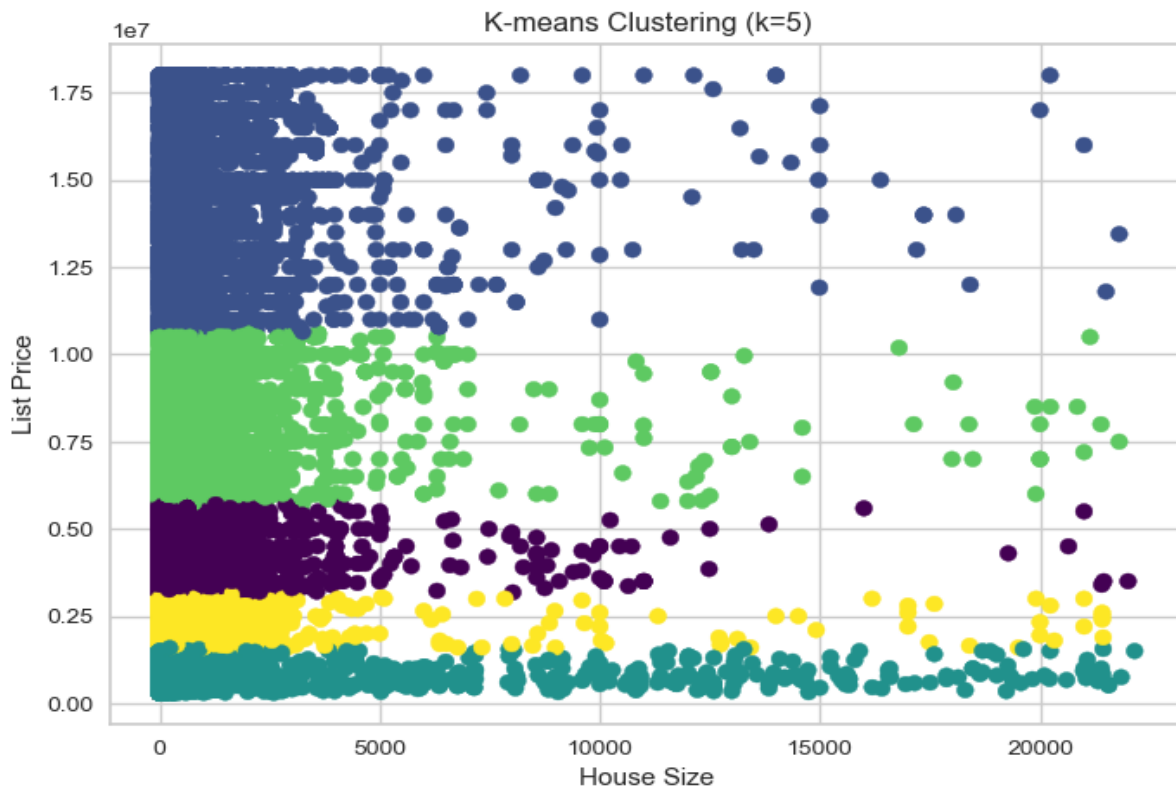


Figure 3.34 shows a K-means clustering successfully that identifies 5 property groups, offering valuable insights for understanding market trends and guiding targeted marketing or investment strategies.

Key Observations:

Cluster Characteristics:

- **Cluster 1 (Teal):** Properties with low list prices and small house sizes.
- **Cluster 2 (Yellow):** Properties with moderate list prices and medium house sizes.
- **Cluster 3 (Green):** Properties with higher list prices and larger house sizes.
- **Cluster 4 (Purple):** Properties with very high list prices and very large house sizes.
- **Cluster 5 (Blue):** Properties with a broad range of house sizes and higher list prices.

Clustering Insights:

- The clusters reveal distinct property segments based on list price and house size.
- Larger house sizes generally correspond to higher prices, but some larger homes have relatively low prices, indicating potential outliers or unique features.

Figure 3.35: The boxplots of 6 numerical variables subplots.

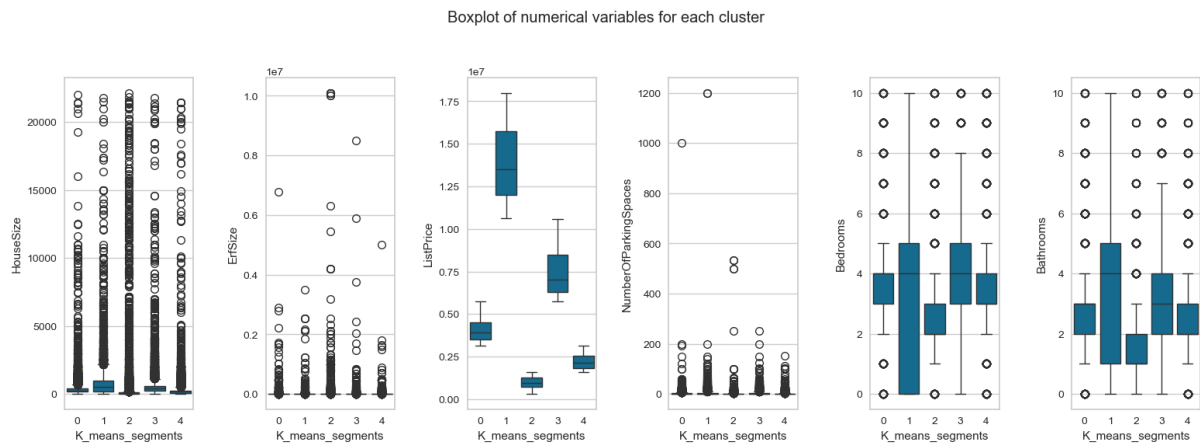


Figure 3.35 reveals that Clusters 4 and 5 generally represent larger, more expensive properties with more amenities. Clusters 1 and 2, in contrast, consist of smaller, more affordable properties with fewer amenities. This segmentation highlights distinct property types across clusters, useful for targeted marketing or investment decisions.

Figure 3.36: Distribution across the 5 clusters

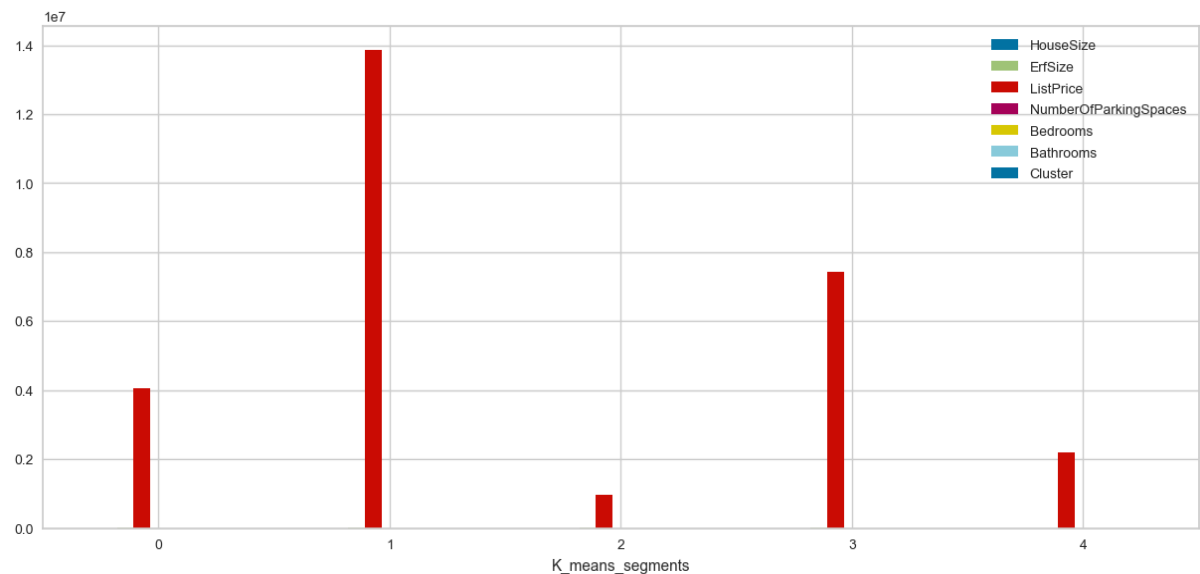


Figure3.36 provides a clear visual comparison of the numerical variable distributions across the 5 clusters. It helps to identify the unique characteristics of each cluster and understand how the properties within each cluster differ from one another.

3.5 Collaborative Filtering

“A Collaborative Filtering model was implemented to improve recommendations based on user behaviour data”, (Majumdar, 2022, p. 10). Using Singular Value Decomposition (SVD), the model provided personalized recommendations by learning from users' interactions with properties.

Top 5 Recommended Properties:

- **House** – This property type appears frequently in the recommendations, suggesting a strong preference alignment.
- **Commercial Property** – Based on similar user profiles, commercial properties align with user interests, likely due to a trend in mixed-use or investment opportunities.
- **Townhouse** – Townhouses are recommended due to their popularity among users with similar preferences, possibly due to their balance between space and affordability.
- **Apartment** – Apartments rank highly, appealing to users looking for urban or low-maintenance living spaces.
- **House** (second recommendation) – Indicates a high preference for standalone homes, possibly due to recurring interest in larger or private properties.

CHAPTER 4: CONCLUSION and RECOMMENDATIONS

This study examined the relationship between property features, such as bedroom count, and listing prices to develop a property recommendation model and K-Means clustering. The findings revealed a moderate positive correlation (0.28) between house size and list price, suggests that larger houses generally have higher listing prices. Outliers were observed, indicating considerable variability in list prices within each house size, possibly due to location, amenities, or other factors not fully explored in this study. The data analysis also highlighted the predominance of smaller properties in the dataset, with only a few high-end properties creating a skewed distribution.

Despite achieving reliable predictions for typical properties, the study's limitations include potential bias from unaccounted features and limited geographical scope. Overall, the research contributes valuable insights into property trends and sets a foundation for property recommendation models tailored to buyer preferences.

Recommendations

- **Enhance Data Collection:** Add features like property location, age, and neighborhood amenities to improve model accuracy and better address price variability.
- **Expand Geographical Scope:** Include data from more locations to create a more generalized model for diverse property markets.
- **Implement in Real-Time Systems:** Deploy the model on real estate platforms to enable dynamic updates with new listings and market trends.
- **Further Research on Outliers:** Study properties outside typical price ranges to understand luxury market trends and improve high-end property predictions.
- **Consider Buyer Behavior Patterns:** Use collaborative filtering based on user viewing and purchasing behaviors to personalize recommendations.

APPENDIX

Appendix 1: Project Calendar

Week	Activities	Output
Week 1 (Sep 1 - Sep 7)	Project Kick-off & Requirement Gathering	Defined project goals, key data sources, and objectives. Discussed client needs.
Week 2 (Sep 8 - Sep 14)	Data Collection and Preliminary Analysis	Collected user interaction data and property feature data. Initial data quality checks and cleaning performed.
Week 3: Sep 15 - Sep 21	Literature Review and Model Selection	Conducted research on collaborative filtering, content-based filtering, and hybrid systems. Selected initial models for testing.
Week 4: Sep 22 - Sep 28	Data Preprocessing and EDA	Processed user and property data, handling missing values, normalization, and feature extraction. Prepared data for training.
Week 5: Sep 29 - Oct 5	Model Development (Collaborative Filtering)	Developed and tested collaborative filtering models using matrix factorization (SVD, ALS).
Week 6: Oct 6 - Oct 12	Model Development (Content-Based Filtering)	Develop content-based filtering models using property features and cosine similarity. Analyse performance.
Week 7: Oct 13 - Oct 19	Hybrid Model Integration	Combine collaborative and content-based models into a hybrid recommendation system. Fine-tune weighting of models.
Week 8: Oct 20 - Oct 26	Hyper-Parameter Tuning and Optimization	Optimize model parameters using grid search or random search to improve performance of the hybrid system.
Week 9: Oct 27 - Nov 2	Model Evaluation and Testing	Evaluate model accuracy, precision, recall, and other metrics using test and validation data. Perform A/B testing on a sample of users.
Week 10: Nov 3 - Nov 9	System Deployment Preparation	Finalize system architecture for deployment. Prepare cloud infrastructure (e.g., AWS) and set up real-time data pipelines.
Week 11: Nov 10 - Nov 16	System Deployment and Monitoring	Deploy the hybrid recommendation system to Property24 platform. Set up monitoring for system performance and user feedback.
Week 12: Nov 17 - Nov 23	Post-Deployment Analysis and Final Adjustments	Analyse system performance post-deployment, gather feedback from client, and make final adjustments.
Week 13: Nov 24 - Nov 30	Final Report and Client Handover	Prepare final project report, including insights and recommendations for long-term model improvement. Conduct client handover.

REFERENCES

1. Badriyah, T. Azvy, S. Yuwono W. and Syarif, I. *Recommendation system for property search using content-based filtering method*. 2018. International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, Indonesia, 2018, pp. 25-29. <https://ieeexplore.ieee.org/document/8350801> [20 September 2024].
2. Bobadilla, J, Ortega, F, Hernando, A, and Gutiérrez, A. 2013. *Recommender systems survey*. Knowledge-based systems 46 (2013), 109–132.
3. Devore, J.L. and Berk, K.N., 2017. *Modern Mathematical Statistics with Applications*. Softcover reprint of the original 2nd ed. 2012. New York: Springer-Verlag.
4. Gharahighehi, A., Pliakos, K., Vens, C. *Recommender Systems in the Real Estate Market—A Survey*. Appl. Sci. 2021, 11, 7502. <https://doi.org/10.3390/app11167502> [20 September 2024].
5. Harwani, M. 2012. *Introduction to Python® Programming and Developing GUI Applications with PyQt*. Boston: Stacy L. Hiquet.
6. Kesalika, B. Doshi, M. and Doshi, H, 2021. *Real Estate Recommendation System*. Ijariit. <https://www.ijariit.com> [20 September 2024].
7. Majumdar, A. (2022). Collaborative Filtering: Recommender Systems. CRC Press, Taylor & Francis Group.
8. Parambath, S.A.P. and Chawla, S. *Simple and effective neural-free soft-cluster embeddings for item cold-start recommendations*. Data Min. Knowl. Discov. 2020, 34, 1560–1588.
9. Gumede.S. Code. https://drive.google.com/file/d/1hfMiTe_5YYifl-zt1IcL-ScbzmlKbWfA/view?usp=drive_link