Statistical Analysis of Housing Market Data in Stockholm

Author: Prashant Kumar Singh

September 22, 2025

Contents

1 Introduction

This project presents a statistical analysis of housing market data in Stockholm. The dataset (`dataset03.xlsx`) contains information on housing units, including starting prices, regions, housing types, presence of balconies, number of rooms, and area. The analysis was conducted in R using packages such as `dplyr`, `mosaic`, and `DescTools`. The study provides descriptive statistics, correlation analysis, and regression modeling to understand factors affecting housing prices.

2 Data Preparation

Categorical variables such as `REGION`, `TYPE`, and `BALCONY` were converted into factors. The dataset was checked for missing values and cleaned accordingly.

3 Analysis and Results

3.1 Descriptive Statistics of Starting Price

A detailed analysis of the variable `STARTING_PRICE` was performed, including mean, median, mode, quantiles, skewness, and outlier detection.

Key Findings

- Mean price: 4,273,799 SEK
- Median price: 3,495,000 SEK
- Skewness: 2.06 (right-skewed distribution)
- Outliers: 32 properties priced above 9,495,000 SEK

3.2 Housing Type by Region

Cross-tabulation of housing type and region shows clear dominance of apartments in all regions, especially Stockholm and Northwest.

3.3 Area by Region

Boxplots reveal that the Northeast region has the highest mean and median area, while Stockholm has smaller, more uniform housing sizes.

3.4 Correlation Between Area and Price

Scatterplot and correlation analysis show a moderate positive relationship between starting price and area. Pearson's correlation coefficient = 0.62.

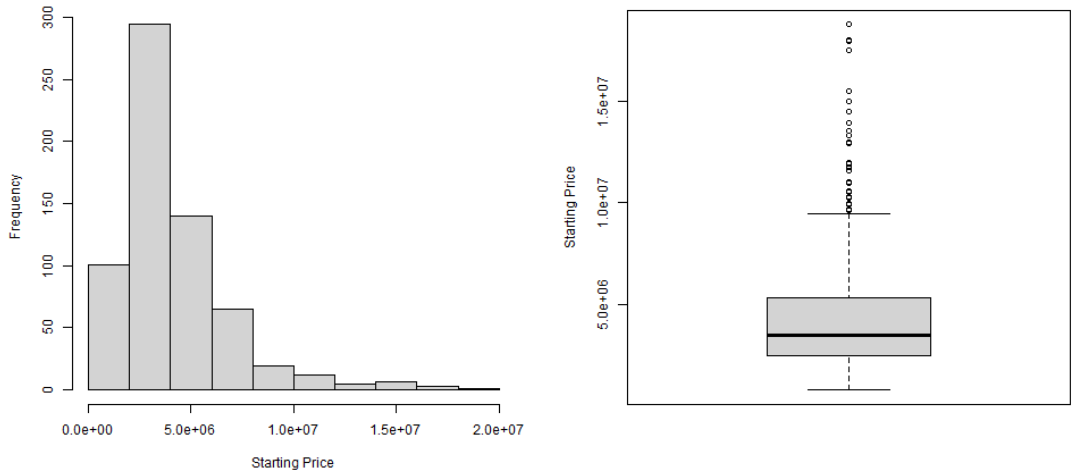

Figure 1: Histogram and Boxplot of Starting Price

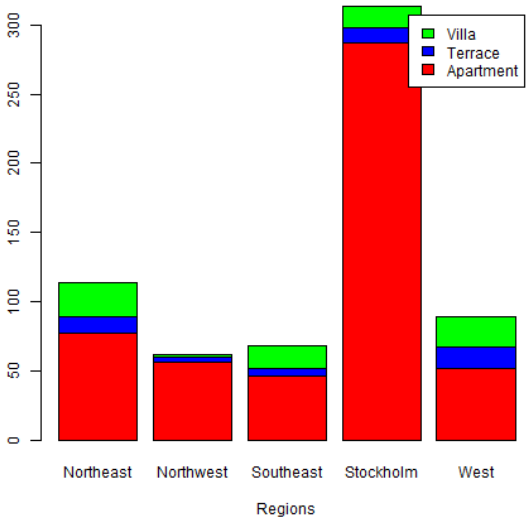

Figure 2: Housing Types by Region

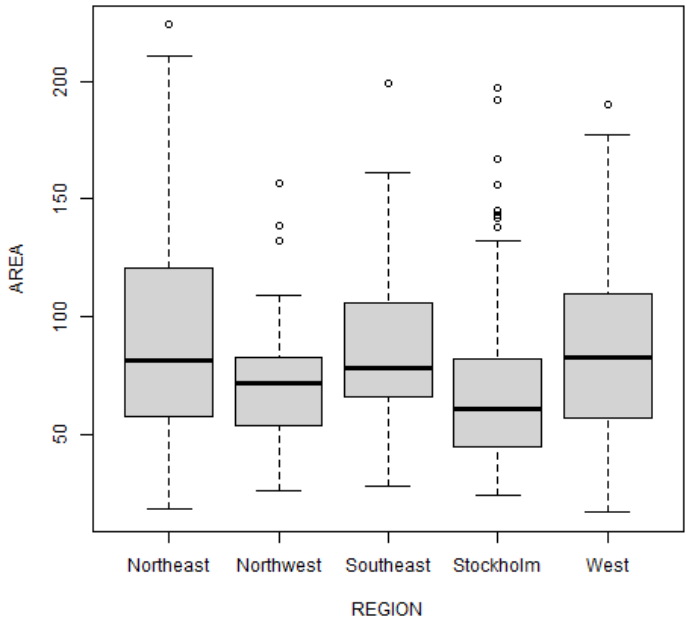

Figure 3: Distribution of Area by Region

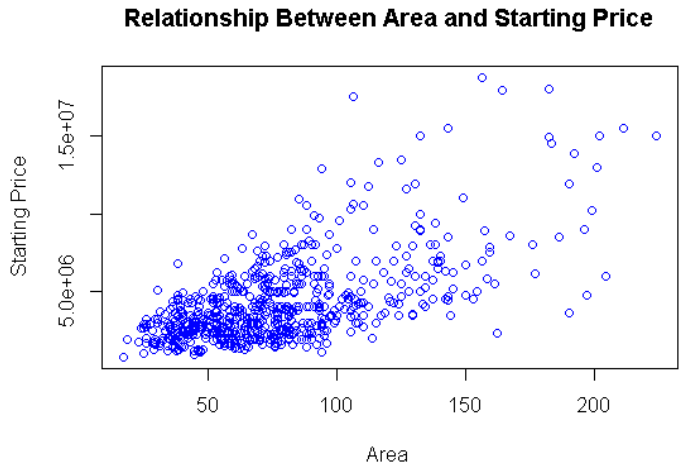

Figure 4: Scatterplot of Area vs Starting Price

3.5 Simple Linear Regression

The regression model:

$$\text{Price} = 646202.35 + 47132.73 \cdot \text{Area}$$

with $R^2 = 0.385$, meaning area explains 38.5% of the variation in price.

3.6 Multiple Regression

Including region, type, balcony, rooms, and area improves the model:

$$\text{Price} = -176230 + \ldots + 60419 \cdot \text{Area}$$

with $R^2 = 0.563$.

3.7 Prediction on Test Data

Using the multiple regression model, predictions were made on unseen test data.

4 Conclusion

The study finds that:

- Housing prices are highly skewed with significant outliers.
- Region and housing type strongly influence prices.
- Area is an important but not exclusive determinant of price.
- The multiple regression model explains about 56% of price variation.

ID	Region	Type	Rooms	Area	Balcony	Predicted Price (SEK)
629	Northwest	Apartment	3	74	Yes	3,950,728
718	Northeast	Apartment	4	99.5	Yes	4,934,974
1534	Stockholm	Apartment	6	115	Yes	7,061,187

Table 1: Predicted Starting Prices for Test Data