Statistical Report for the sold housing in an agency

Group A - 04

Home assignment for statistics and data analysis

1. **Ghazaleh Ranji** 1990-03-23 ghazaleh.ranji90@gmail.com
2. **Negin Ebrahimitofighi** 1989-03-29 Ebrahiminegin67@gmail.com
3. **Peeyush Kelkar** 1998-07-11 peeyushkelkar@gmail.com

Exercise 1

Analysis of variable: STARTING_PRICE

Measures of Location	Measures of Variability	Measures of Shape
mean =4,358,600 SEK	Range = 21,205,000 SEK	skewness = 2.252064
median = 3,495,000 SEK	IQR = 3,023,750 SEK	Outlier
mode = 2,995,000 SEK	Variance = 8.78 trillion SEK	
First quantile = 2,437,500	Standard deviation = 2,962,645	
Third quantile = 5,461,250		

Numerical description

For the first part of the analysis, we calculated measures of location to summarize where the observations of the variable “STARTING_PRICE” are located and will describe it both numerically and graphically.

Measures of Location

Starting with the *mean*, the center of gravity of the observation. It indicates the average starting price across all observations. In other words, the starting prices of the houses sold by the real estate agency hover around 4,358,600 SEK.

The Median gives us a more realistic picture of the middle starting price in the market which equals 3,495,000 SEK. It also indicates that half of the sold units have a starting price below this value and the other half has a starting price above this value.

Mode is another parameter of location that indicates the most frequently occurring value, in our case, 21 units with a starting price of 2,995,000 SEK has the most frequent value.

To get a better picture of the distribution of the starting prices in our data set, we calculated another location parameter called *quartiles* which divides the observation into four quarters and helps us to conclude that:

- Q1 = 2,437,500. meaning 25% of the starting prices are below this price.

- Q2 = 3,495,000. Equals to the median and indicates that half of the starting prices are below, and the other half is above this price.
- Q3 = 5,461,250. About 25% of the starting prices are above this price.

Measures of variability

For the second part of the analysis, we calculated measures of variability that allows us for measuring the variability in the variable "STARTING_PRICE".

The parameter *range* shows that the Maximum Starting Price of the units equals to 21,995,000 SEK, the Minimum Starting Price: 790,000 SEK and difference between the lowest and highest starting prices, 21,205,000 SEK**,** indicates a large variability in housing prices.

Because range is not a robust parameter to extreme observations, we measured the *interquartile range*. In this case, which is the difference between the third and first quartiles. It means that the middle 50% of starting prices are between 2,437,500 SEK and 5,461,250 SEK, spanning a range of 3,023,750 SEK.

Variance is another measure of variability that tells us how spread the starting prices in our data set is from the average price of 4,358,600 SEK. We can conclude from this number, 8.777267e+12 (8,777,267,000,000), that we have a high variability in the starting prices and many starting prices are far from the average starting price (4,358,600 SEK) by a few million SEK.

Due to the difficulty in interpreting the variance, we should use the *standard deviation.* This indicates the average distance from the starting prices to the mean is 2,962,645 SEK.

Measures of Shape

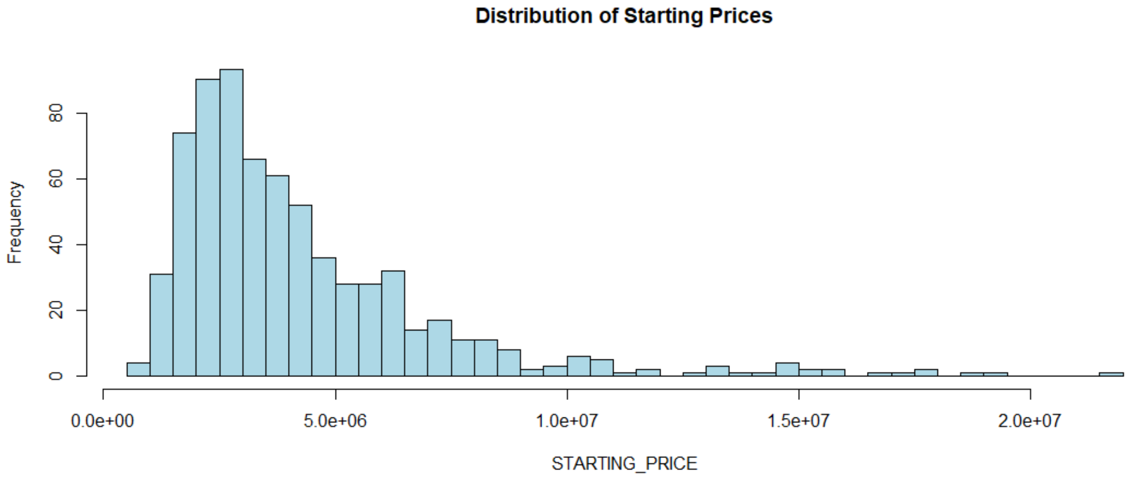
The last parameter which is measured to determine the shape of data is *Skewness* . The skewness of 2.252064 determines that most prices are concentrated at lower levels. As we can observe in the histogram, a few very high price houses stretch the tail to the right.

Outlier Detection

Another measure of location, is outliers. By using the IQR method for calculating the outliers, there are 35 outliers in the data set.

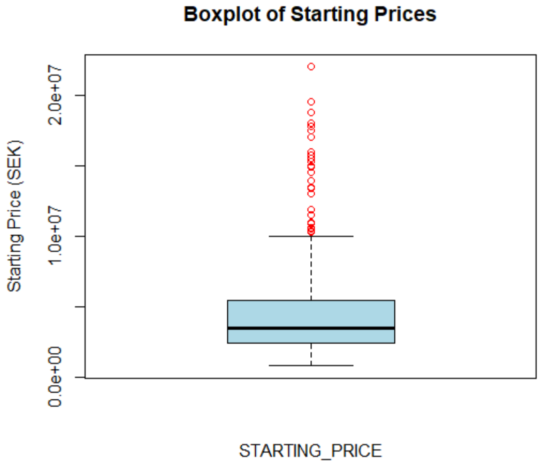
Outliers are data points significantly different from the rest of the dataset .In other words , outliers are units with unusual and extremely higher or lower price compared to the typical range of starting prices (IQR) for most of the units.

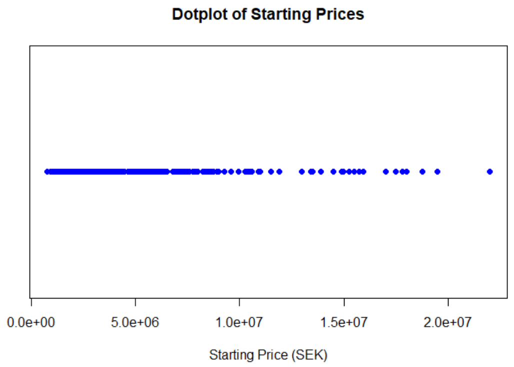
Graphical Summary



2. **Boxplot**

In the boxplot of starting prices, outliers are visible as red dots. It means that many properties' prices are above the typical range.





3. **Dot plot**

In the dot plot of starting prices, it is obvious data spread and density and outliers as well.

The chart shows that most starting prices are concentrated at the lower range (closer to 0), indicating that many houses have relatively low starting prices.
It is shown that most properties fall within a low price range.

A few dots are located on the right. These represent outliers.

Exercise 2

Analysis Variables of Region and Type

Numerical description

contingency table

	Apartment	Terrace	Villa	Sum
Northeast	85	13	26	124
Northwest	69	3	2	74
Southeast	44	11	20	75
Stockholm	308	12	14	334
West	58	12	19	89
Sum	564	51	81	696

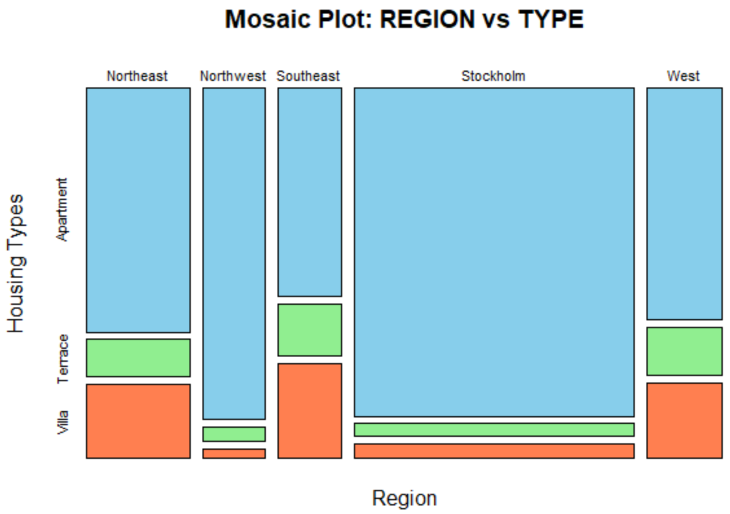
Conditional distribution table

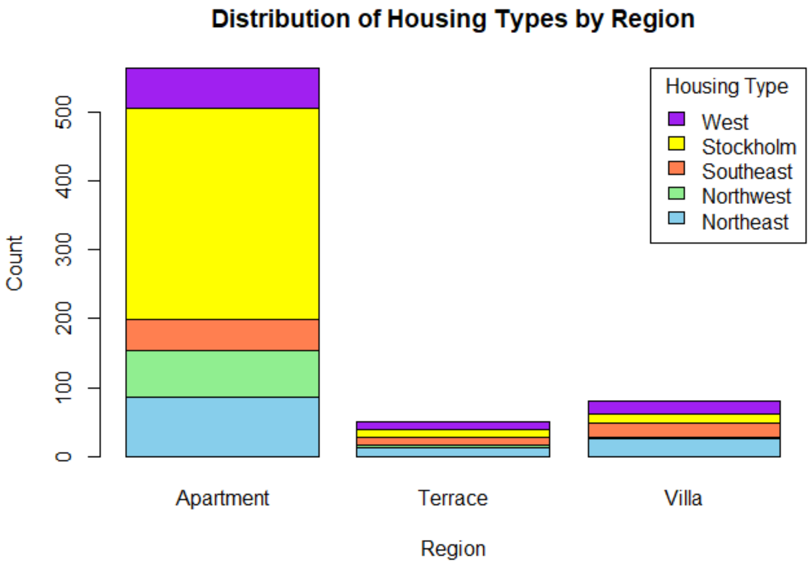
	Apartment	Terrace	Villa
Northeast	0.15070922	0.25490196	0.32098765
Northwest	0.12234043	0.05882353	0.02469136
Southeast	0.07801418	0.21568627	0.24691358
Stockholm	0.54609929	0.23529412	0.17283951
West	0.10283688	0.23529412	0.23456790
Sum	1	1	1

The first table (*contingency table*) shows how many apartments, terraced houses, and villas are located in each region and the summation of the units in each region is represented in the Sum column.

The another table (conditional *distribution table Column wise*) shows the proportional contribution of each region to each housing type. The values represent the proportion of properties from each region that are of that type. For example: 0.5460 of all apartments are in Stockholm, making Stockholm the most common region for Apartments.

Graphical Summary:





As shown in the mosaic plot, the width of each region block shows the total number of housing units in the region, while the height of each color segment represents the proportion of each housing type.

We can observe that Regions with a larger proportion of apartments like Stockholm have larger blue sections , while regions with more villa types like Northeast and Southeast will have more orange sections.

Northwest is also dominated by apartments (large blue area), with very little contribution from terraced houses and villas.

Northeast, Southeast, and West regions have the most balanced distributions, with large proportions of villas and terraced houses.

The stacked bar plot displays the counts of each housing type across regions. Apartments are the most common housing type among other types of housing and with a large contribution by

Stockholm (Yellow) and terraces are the least common housing type .

Most of the Villas are respectively concentrated in the Northeast (light blue), Southeast (orange), and West (purple) regions.

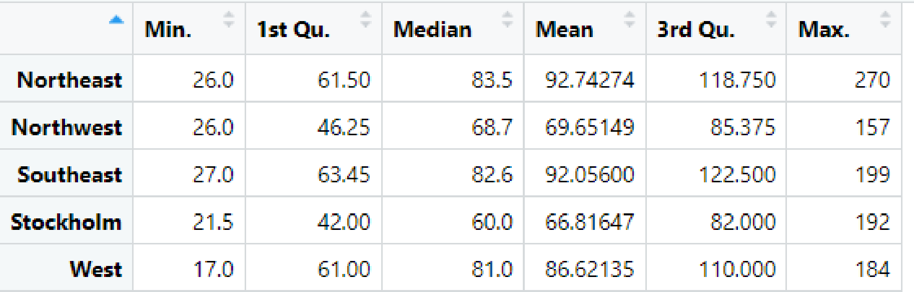
Northwest has the least concentration of Terrace compared to other regions.

Exercise 3

Analysis Variables of Region and Area

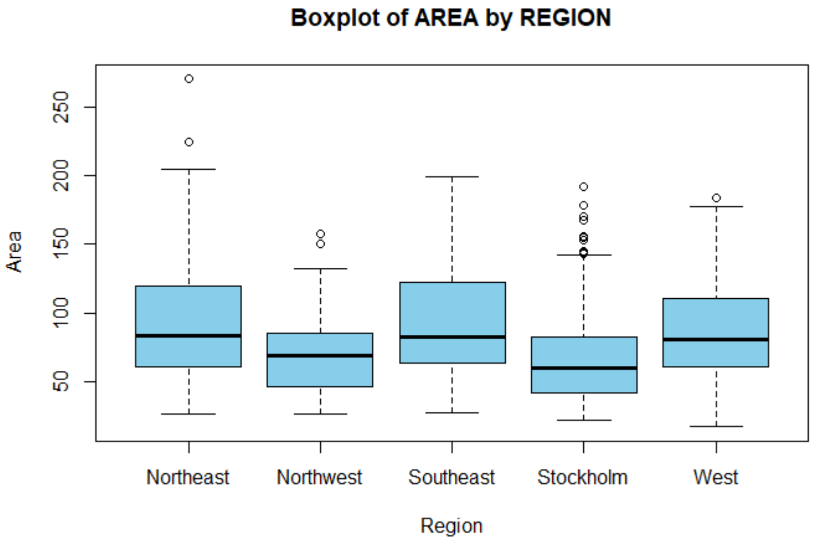
Numerical description

"Descriptive Statistics of Housing Unit Sizes (Area) Across Regions"



	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Northeast	26.0	61.50	83.5	92.74274	118.750	270
Northwest	26.0	46.25	68.7	69.65149	85.375	157
Southeast	27.0	63.45	82.6	92.05600	122.500	199
Stockholm	21.5	42.00	60.0	66.81647	82.000	192
West	17.0	61.00	81.0	86.62135	110.000	184

Graphical Summary:



This table and box plot summarizes the distribution of housing unit sizes for each region. We can extract these facts from the table and the box plot :

1. The smallest units are found in Stockholm, and the largest units are found in Northeast.
2. Northeast has the widest range of housing sizes, while Northwest has a narrower range.

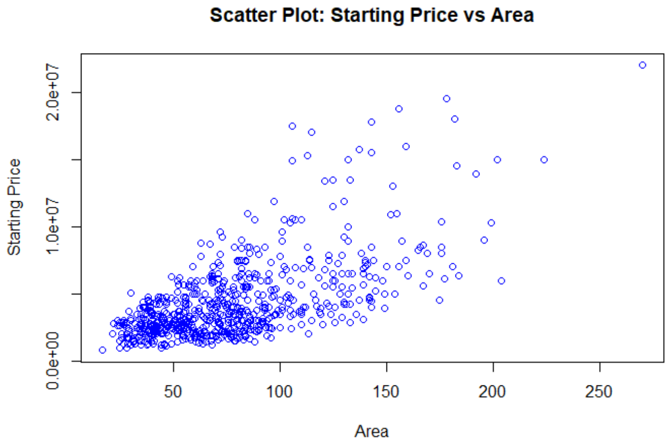
3. Stockholm has smaller housing units, while Northeast, West and Southeast have larger and more varied housing options.
4. The southeast region suggests a balance between and slightly higher variability compared to regions like Northwest and Stockholm.
5. In Northeast region we observe larger housing units dominate, with some extremely large units as outliers.
6. As mentioned before ,Northwest region generally has smaller housing units with less variability. It doesn't include significant outliers (extreme prices), indicating a consistent housing market.
7. Stockholm has the lowest variability and several small outliers.

Exercise 4

Analysis Variables of Starting Price and Area

Numerical and graphical description

Describing two numerical variables that starting price is dependent on Area (independent variable)



As shown in the scatter plot, there is a correlation between the starting price and area. To be able to clarify this correlation we calculated a measure called Pearson correlation coefficient which equals to 0.6348584.

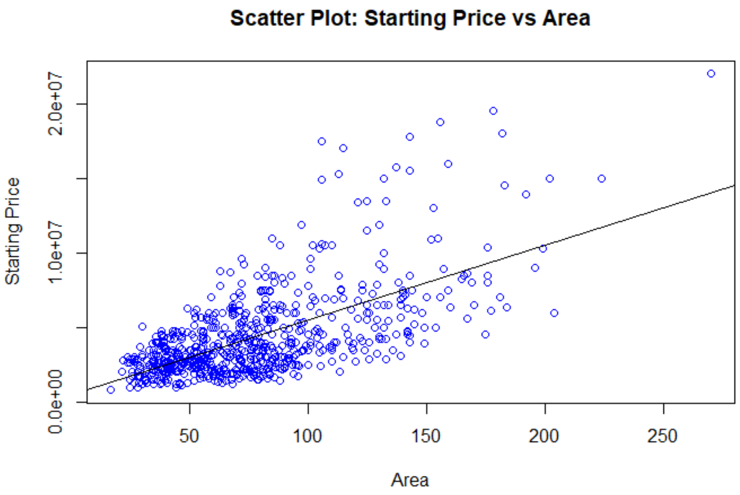
This positive correlation coefficient indicates that there is a positive and strong correlation between the area and the starting price. This means that increment in the size of houses are associated with increments in starting prices.

We also observe it in the scatter plot above that units with larger area has almost a higher starting price.

Exercise 5

Relationship between Starting price and Area

We have examined the relationship between a housing unit's starting price and its size (area) using a simple linear regression model. This statistical approach allowed us to estimate how much the starting price increases, on average, with each additional square meter of housing.



To describe the relationship between Starting price (y) and area (x): we need to calculate two key components of the linear relationship:

1- Intercept equals to 502,645 SEK. This parameter means that when the area is 0 square meters, we can expect a starting price of 502,645 SEK. This value does not make sense, but it is necessary for the regression equation.

2- Slope (50,084 SEK per sqm): This means that for every additional square meter of area, the starting price increases by 50,084 SEK on average.

So the regression line can be described by this equation:

$$StartingPrice = 502,645 + 50,084 * Area$$

How good the area is at explaining the price?

- **R-squared (0.403)** Shows the 40.3% of variability of the starting price that it explained by the area .
- **Adjusted R-squared (0.402)** adjusts for the number of predictors in the model and is slightly lower. It still indicates that area explains a moderate portion of the price variation.
- **Residual Standard Error (2,291,000)**: On average, the predicted starting prices differ from the actual prices by approximately 2,291,000 SEK.

Conclusion

The analysis confirms a **positive relationship** between size and starting price: larger houses tend to cost more. However, the moderate R-squared value suggests that size alone does not fully explain price variations, and additional factors need consideration.

Exercise 6

The fitted multiple linear regression model explaining starting price by region, type, area, balcony and rooms is the following:

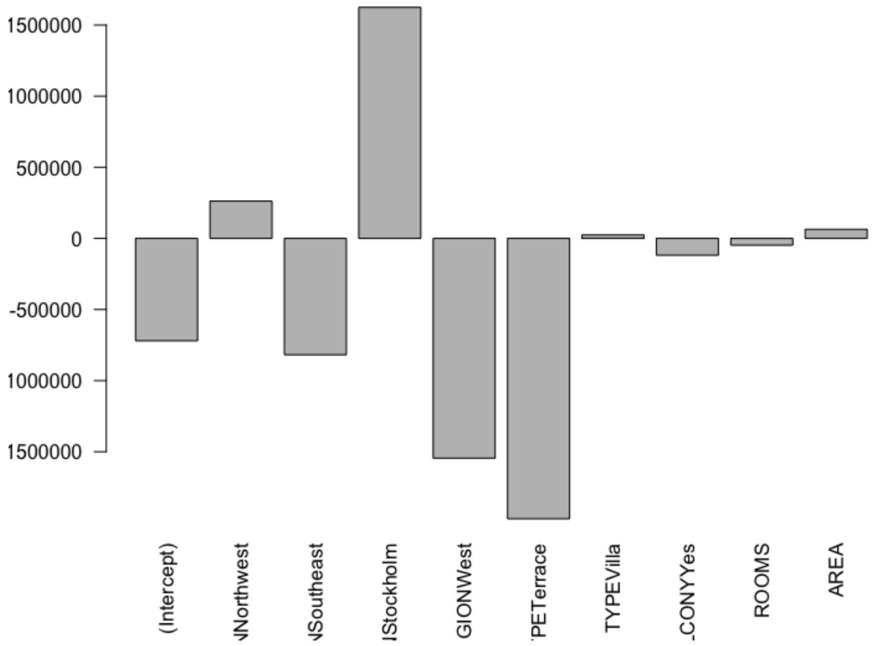
$$STARTING_{P}RICE = -720267.22 - 47969.61 \cdot ROOMS + 63806.94 \cdot AREA + 261971.25 \cdot REGION_{N}orthwest - 818160.53 \cdot REGION_{S}outheast + 1624294.95 \cdot REGION_{S}tockholm - 1545151.32 \cdot REGION_{W}est - 1971245.22 \cdot TYPE_{T}errace + 24423.70 \cdot TYPE_{V}illa - 118401.34 \cdot BALCONY_{Y}es$$

As an example of how we can extract the relation between different slope, If we consider all other variables constant except Area, we can expect that for each unit increment in size of the unit , the starting price will averagely increase by 63806.94 SEK.

Being in the Northwest increases the price by 261,971.25, while being in the Southeast or West decreases the price by 818,160.53 and 1,545,151.32, respectively.

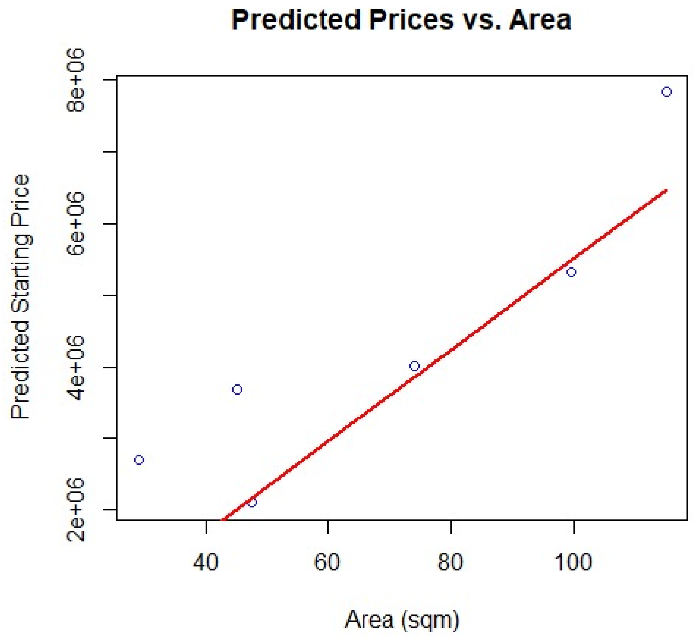
On the other hand, a unit with a balcony is expected to cost 118401.34 less.

The bar plot bellow explains the impact of each variable on starting price.



Exercise 7

Prediction the starting price of these housing units in the test data set



Based on the regression model applied earlier and the plot above , we can estimate the price of the housing units with the mentioned sizes 29, 45, 47.5 , 74 , 99.5 and 115 sqm are expected to be :

REGION	TYPE	ROOMS	AREA (sqm)	BALCONY	PREDICTED_PRICE (SEK)
Northwest	Apartment	3	74.0	Yes	4,001,107
Northeast	Apartment	4	99.5	Yes	5,318,243
Stockholm	Apartment	6	115	Yes	7,835,607
Stockholm	Apartment	2	45	No	3,679,401
Stockholm	Apartment	1	29	No	2,706,459
Northeast	Apartment	2	47.5	Yes	2,096,222