

FINAL PROJECT REPORT

Nida KARATAŞ

Ali TAŞBAŞ

İsmet LOĞOĞLU

Azim Yunus ALPARSLAN

Onur KEKLİK

ABSTRACT

[INTRODUCTION 3](#_Toc125135126)

[DATA TIDYING AND CLEANING STEPS 3](#_Toc125135127)

[District 3](#_Toc125135128)

[The District’s Socioeconomic Status 3](#_Toc125135129)

[Numerical Data 4](#_Toc125135130)

[Price 4](#_Toc125135131)

[Age 4](#_Toc125135132)

[Floor 4](#_Toc125135133)

[Heating System 4](#_Toc125135134)

[Number of Room 4](#_Toc125135135)

[Area 5](#_Toc125135136)

[Unit 5](#_Toc125135137)

[RESEARCH QUESTIONS 5](#_Toc125135138)

[RESEARCH QUESTION 1 5](#_Toc125135139)

[RESEARCH QUESTION 2 7](#_Toc125135140)

[RESEARCH QUESTION 3 7](#_Toc125135142)

[RESEARCH QUESTION 4 8](#_Toc125135143)

[RESEARCH QUESTION 5 8](#_Toc125135144)

[RESEARCH QUESTION 6 9](#_Toc125135145)

[CONCLUSION 11](#_Toc125135146)

[REFERENCES 11](#_Toc125135147)

# **INTRODUCTION**

The housing market is a complex and dynamic field that is affected by various factors, including the physical properties of the houses, location, and socioeconomic status of the surrounding area. In this project, we will be analyzing a dataset that includes housing prices in three districts of Ankara, along with information about the houses' physical properties and the districts' socioeconomic status. We aim to use this data to gain insights into how these factors affect housing prices and answer a series of research questions. We will begin by cleaning the dataset, ensuring that the data is accurate and complete, and then use visualizations and statistical analysis to interpret the data and draw conclusions about the housing market in Ankara. By the end of the project, we would like to understand better the factors that influence housing prices and how they vary across different districts.

# DATA TIDYING AND CLEANING STEPS

To effectively analyze our data set, it is necessary to clean and rectify any defects present within the raw data. The subsequent section will detail the identified inaccuracies and outline the steps taken to obtain a straight and accurate data set.

Firstly we need to check duplicates; if there are any, we should remove them to avoid misunderstandings.

## District

Upon examination of the district data, several variables were identified as needing consistent or correct spellings. To standardize the data, it was determined that the three main districts of Çankaya, Keçiören, and Mamak would be the focus of the analysis. To this end, all variables were standardized to align with one of these three districts. This involved converting all data to lowercase, standardizing the formatting of variable names, and removing any extraneous white space.

## The District’s Socioeconomic Status

We applied the same operations to the district variable in this variable due to similar reasons. The variable is divided into three main levels: High, Middle, and Low. Empty data within the district\_socioeconomic\_level variable was replaced with the following values: Çankaya representing High Level, Keçiören representing Middle Level, and Mamak representing Low Level.

## Numerical Data

During the numeric cleaning process, we calculated the mean and median values. For the Age column, there was no significant difference between the districts and the overall mean. Therefore, we filled in any missing data with the overall mean. However, for the Price, Floor, and Area columns, there was a significant difference between the district's mean and the overall mean, so we employed the groupby() method to fill any missing data with the appropriate district's median or district's mean value.

### Price

We converted any values of 5.000000e+15 to NaN, as we did not want these values to impact the mean price calculation. We determined the mean for each district separately and subsequently filled any missing data with the appropriate district's mean.

### Age

The Age column contains 22 null values. As the variable has outliers, it was determined that the median would be a more robust measure than the mean. Additionally, using the mean would result in decimal values that would not be meaningful. These null values were filled with the median value. Additionally, 1000 outlier values were also replaced with the median value.

### Floor

The total number of null values in the Floor column is 29. Values of 0 were converted to NaN to allow a single code block to be used for replacement. As the variable contains outliers, the median was chosen as a more robust measure than the mean. Missing values were subsequently filled with the median value.

### Heating System

This column does not contain any values and will be removed from the data set.

### Number of Room

We converted any values of 0 to NaN, as it is not possible for a house to have 0 rooms. We created subsets for each district, Mamak, Keçiören, and Çankaya, and found the median values to be 2, 3, and 4, respectively. Since the median values of these districts have a significant difference, we filled any missing values with the median value specific to that district. This process was done using a for loop and updating the appropriate subset values.

### Area

We converted the areas whose unit is cm2 to m2 by dividing the value by 10000. We also split the mean area concerning districts. Subsets of areas were created. Since there is a significant difference between their mean, assigning the overall mean to all missing values is inappropriate.

### Unit

All area data was converted from cm2 to m2. As a result, there is no longer any data in cm2 units. Since all area data is now in m2 units, any missing data were filled with the appropriate m2 value.

Finally, we completed the data cleaning process by adequately formatting the names of all columns.

# RESEARCH QUESTIONS

## RESEARCH QUESTION 1

**How do the price, area, and rooms of flats in Çankaya compare with those in the other districts?**

A picture containing radar chart

Description automatically generatedAs can be seen from the violine chart, Mamak district has the lowest price distribution and the most expensive price distribution for Çankaya. We can say that the number of houses is the highest in the price range of 490k in the Mamak region, 580k in the Çankaya region, and 530k in the Keçiören region. We can say that the difference between house prices in Keçiören is the region with the slightest difference. Besides that, Keçiören has the lowest range with house prices. This represents that house prices in Keçiören are close to each other and exhibit consistency.

Chart, box and whisker chart

Description automatically generatedThe area distribution of the houses for each district is shown with a Box plot. While the houses with the highest area in the distribution are located in Çankaya, the district with the lowest area distribution is the Mamak. The highest range was observed in Keçiören, while the lowest range was observed in Mamak. In other words, we can say that all the houses in Mamak, except the outliers, are between 92 m2 and 96 m2. While the houses in Keçiören show a normal distribution, Mamak; is negatively skewed, and Çankaya is a positively skewed distribution.

Chart, bar chart

Description automatically generatedOur dataset’s highest number of houses is Keçiören, with 67 houses. Çankaya, 59 houses, is in the 2nd place, and Mamak, 33 houses, is in the 3rd place. In our dataset, the maximum and minimum observations, that is, range values, are not very high in percentage. This shows that the values in the dataset are reliable.

## RESEARCH QUESTION 2

Chart, treemap chart

Description automatically generated**Which variable correlates highest with the price between the flats in all the districts? Interpret the result, and give your opinion.**

0,06 0,67 0,61 0,63

The variable with the strongest positive correlation with house prices is the Floor variable with 0.67. Immediately after, it is the Area variable with 0.63. The variable Price has a weak relationship with the correlation value of the variable Age of 0.06. Ultimately, the floor number is the most effective variable on prices. The floor might have the highest correlation in theory, but that is not a reliable measure. Looking at the range of floors generally, it takes much work to accept it. Out of these variables, the area is the most logical one, and its correlation coefficient isn't that different from the floor's one. The area is overall the most logical and reliable indicator of the flat's price. (We will later see that the district it is located in can significantly impact the price.)

## Chart, scatter chart Description automatically generated

## RESEARCH QUESTION 3

**How related are the prices and respective areas of flats in Kecioren?**

The relationship between the areas of the houses in Keçiören and the prices presented on the scattered plot. When the data points are plotted, we observe that all the values cannot come together at a point close to each other and cannot form a solid straight line. Values that can come together on very few data have a weak relationship in the negative direction. The correlation coefficient value of area and price values in the Keçiören district is -0.16. This value shows us a fragile negative relationship between these two values.

## RESEARCH QUESTION 4

Chart

Description automatically generated**Does the older a building get in Çankaya increase its value? Compare it with the other districts.**

The scatter plot above provides us with the most appropriate graph for whether an increase in the house’s value can be observed as the house’s age increases in the Çankaya district. It is not possible to say that the prices and ages of houses in the Çankaya district have increased similarly. It is quite possible to say there is a weak relationship between them, even close to 0. The correlation coefficient between the ages and wages of the houses located only in the Çankaya district is -0.02. This value explains the weak relationship between these two quantitative variables.

## RESEARCH QUESTION 5

Chart

Description automatically generated**Is the floor number of a flat a good indicator of the price for flats in Mamak? Compare it with the other districts.**

The drawing above shows the prices and floor numbers of the houses in the whole district. Although there was only a weak positive relationship for Mamak in the previous graph, the effect of outliers on the cumulative data decreased with the increase in the data. However, a positive relationship continues from the bottom right to the top left of the Chart. This relationship, which could be more robust, may have a moderately positive relationship with a correlation coefficient value of 0.67. The correlation strength is moderate because since both graphs are right-skewed, the distribution is more concentrated in the lower left, so there is no regular increase.

Chart, box and whisker chart

Description automatically generated

The number of floors and prices in the Mamak district; is visualized with a scatter plot containing two histogram graphs on the side. While the horizontal histogram graph shows the independent floor variable distribution frequency, We can observe the independent price variable distribution in the "y" label. The distribution of these two variables is plotted with a scatter plot. There is a positive relationship between values. It can be observed that the values ​​concentrated in the left, middle and upper right corners are in a not very strong positive relationship.

## RESEARCH QUESTION 6

Chart, line chart

Description automatically generated

Examining the dumbbell plot, it is evident that most flats with similar designs are concentrated in Mamak and Keçiören districts. Notably, the flats in Keçiören tend to have slightly higher prices compared to those in Mamak. This pattern is also observed for other types of flats; those located in districts with higher socioeconomic status are generally more expensive than those in other districts. Furthermore, the dumbbell plot illustrates the significant price disparities between houses with identical characteristics in the Çankaya and Keçiören districts, with a price difference of approximately 80k for homes in the same category. Additionally, the price difference between houses of the same type between Keçiören and Mamak is also clearly visible in the plot.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Age | Floor | Number of Room | Price Differences | Districts |
| 4 | **3** | **3** | 45.518 | Keçiören-Mamak |
| 5 | **3** | **3** | 32.233 | Keçiören-Mamak |
| 6 | **3** | **2** | 74.525 | Çankaya-Keçiören |
| 7 | **3** | **2** | 68.659 | Çankaya-Mamak |
| 8 | **1** | **3** | 2.185 | Keçiören-Mamak |
| 8 | **2** | **2** | 21.895 | Keçiören-Mamak |
| 8 | **3** | **1** | 48.881 | Keçiören-Mamak |
| 8 | **3** | **2** | 34.583 | Keçiören-Mamak |
| 8 | **3** | **4** | 63.819 | Çankaya-Keçiören |
| 9 | **3** | **5** | 60.051 | Çankaya-Keçiören |

The table illustrates the apparent disparities in prices between the districts. As can be seen, only one flat has a comparable price between the districts, with the majority being significantly more expensive in certain areas. However, it should be noted that this small sample size does not allow for a clear pattern to be identified. There are various factors that contribute to the prices, and the district is a significant factor. It is essential to keep in mind that numbers alone may not provide a complete understanding of the situation. Additionally, It is clear that the price difference between houses of the same age, the same number of rooms, and the same number of floors is high between Çankaya and Keçiören districts. Also, the price difference between the houses with the same characteristics between Keçiören and Mamak is much less than the difference between Çankaya and Mamak. It can also be said that it is possible to find houses with the slightest price difference between Keçiören and Mamak.

# 

# CONCLUSION

To summarize our operations, we first cleaned our data set using NumPy and pandas. Then, in order to understand our data set, we formulated six research questions and plotted relevant graphs using seaborn and mathplotlib. As a result, it is observed that the Çankaya district is the most expensive among these three districts, while the Mamak district is the least costly. In addition to these, we investigated which variable has a greater impact on price. From this, we found that although the floor affected the price, we thought the area variable would be more logical. To investigate this, we compared area and price in the Keçiören district and found a very weak negative linear relationship. Since it was a very weak relationship, it did not lead us to a definite conclusion. Later, when we observed the relationship between age and price variables, we observed that although it does not affect much, the price may decrease as the age of the building increases.

# 

# REFERENCES

<https://dribbble.com/shots/10696566-Ankara-Icon-Set>

<https://tr.depositphotos.com/vector-images/ankara-turkey.html>

**GitHub Links**

<https://github.com/nidarkstone/House_Price_Data_Tidying_and_Cleaning.git>

<https://github.com/nidarkstone/House_Price_Research_Questions_Graphs.git>

**THANKS FOR YOUR ATTENTION AND PATIENCE**