Analysis of House & Rent Prices in Germany

A. Introduction

A.1. Description & Disscusion of the Background

Germany is one of the most important countries not only in Europe but also in the whole world. According to the overall ranking of Best Countries measure global performance, Germany is the fourth Best Country of the World in 2020. (US-News)

"Germany, the most populous nation in the European Union, possesses one of the largest economies in the world and has seen its role in the international community grow steadily since reunification. The Central European country borders nine nations, and its landscape varies, from the northern plains that reach to the North and Baltic seas to the Bavarian Alps in the south. (Usnews, s. 2020)"

Nowadays, one of the most important investment items for both individual investors and companies is real estate investments. Considering the importance of Germany and Housing Trade together, I decided to make an enlightening analysis for those who are considering to trade housing in Germany.

The aim of this study is to show **the investors** in which cities they can invest in which price ranges. In addition, they will have the opportunity to compare their investment with the average rental fee.

A.2. Defining Data

In this analysis, to be used the Datasets listed below.

- Germany Housing - Rent and Price data set - Apr 20 from Kaggle
    - Kaggle is one of the popular platforms that offers sample data sets and sample data studies on data science and data analysis.Kaggle allows users to find and publish data sets , explore and build models in a web-based data-science env ironment, work with other data scientists and machine lear ning engineers, and enter competitions to solve data scien ce challenges.
    - In the kaggle data set used in the analysis; It is possibl e to reach many parameters such as prices, number of rooms , usage area, construction date, heating system according to the states, cities and districts where the houses in Ge rmany are located.
- Shape File and Locations of Cities

    The shapefile format is a digital vector storage format for storing geometric location and associated attribute in formation. In order to visualize price density map of Germ any this data set was needed. Furthermore, geometrical loc ations of districts was used for clustering analysis.

- Forsquare API was used to get the most common venues of given Borough of Cities.

B. Methodology

B.1 Data Cleaning

Data sets used primarily in this project were cleaned and edited. At this stage, columns not required for analysis were dropped. Necessary data frames were created for specific analyzes. Data types are made compatible with each other for combining operations. For example, merge operations were done over the “post code” column. However, the data types of this column were not the same in all used datasets. This problem has been fixed.

Incompatible data values between data sets have been made compatible with each other. For example, while 4-digit postal code was used in Price and Rent datasets for the state of Sachsen, 5-digit postal code was used by taking the first digit **'0'** in other datasets. In order to fix it ‘0’ was added to Post codes of Price and Rent datasets.

Out[11]:

	State	Year	Post_Code	Rooms	LocalArea	Price
0	Sachsen	2021.0	4571.0	4.0	Espenhain	445900.0
1	Berlin	1934.0	12357.0	2.0	Rudow_Neukölln	545000.0

In order to avoid data loss in table merges, **'0'** should be added to the beginning of the Sachsen State PostalCode values.

Lets fix it.

```
named_rent['Post_Code'] = named_rent['Post_Code'].apply(lambda x: '{0:0>5}'.format(x))
named_price['Post_Code'] = named_price['Post_Code'].apply(lambda x: '{0:0>5}'.format(x))
```

```
named_price.head()
```

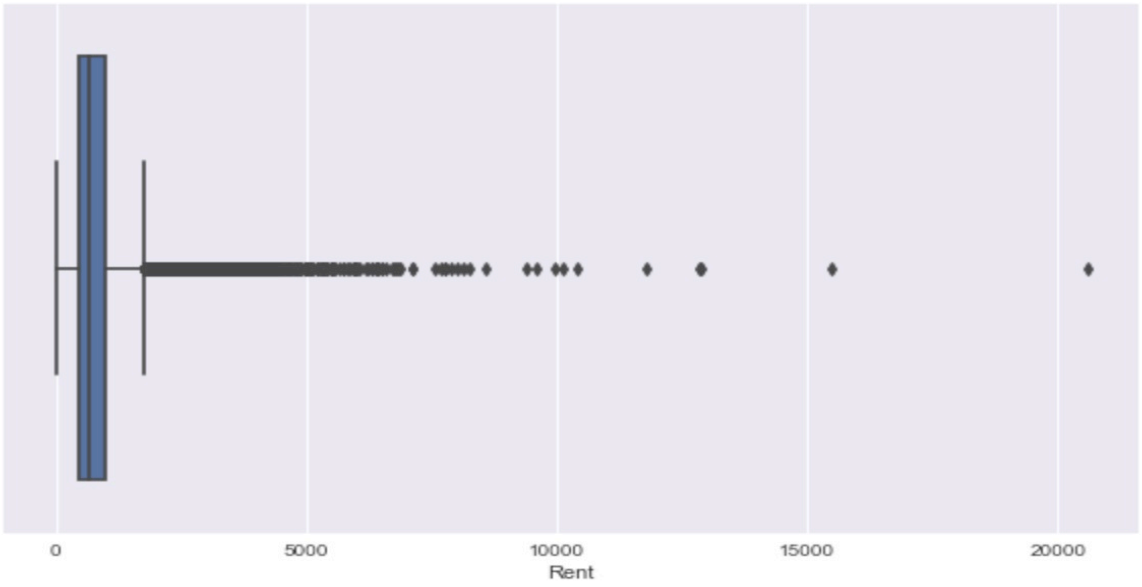
	State	Year	Post_Code	Rooms	LocalArea	Price
0	Sachsen	2021.0	04571	4.0	Espenhain	445900.0
1	Berlin	1934.0	12357	2.0	Rudow_Neukölln	545000.0
2	Baden_Württemberg	1920.0	69434	6.5	Heddesbach	195000.0
3	Schleswig_Holstein	1972.0	24558	4.0	Henstedt_Ulzburg	309000.0
4	Niedersachsen	NaN	49774	5.0	Lähden	267777.0

Outliers of Prices and Rents

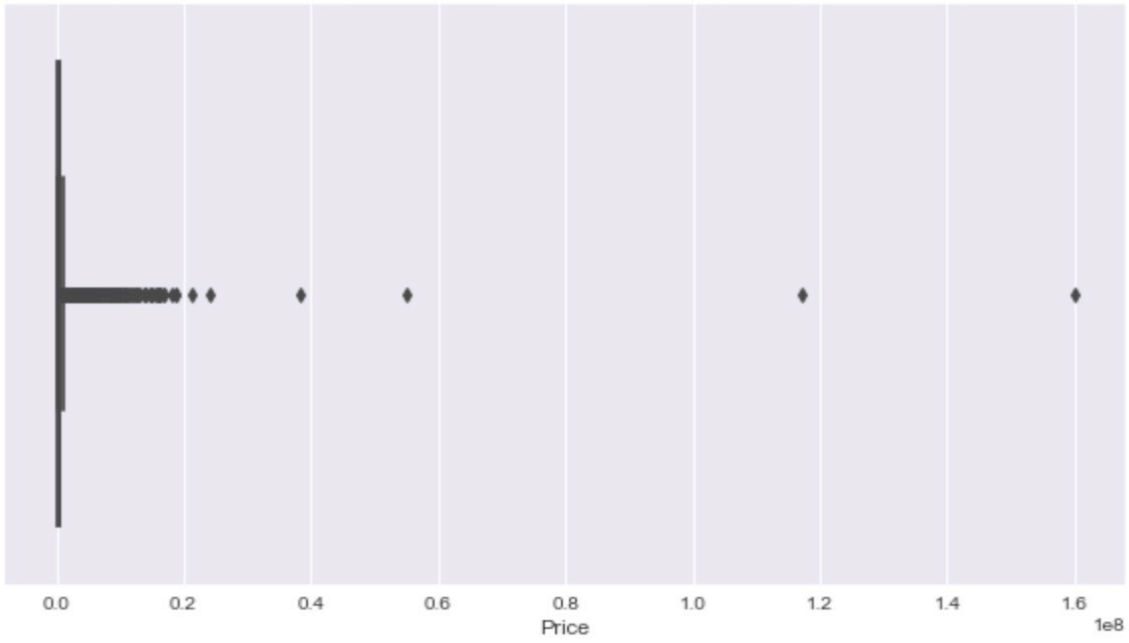
The other value that needed to be repaired was outlier values of Prices and Rents. According to Wikipedia[1], Outlier is a data point in the dataset that differs **significantly** from the other data or observations. Working on average values, these extreme values can be cause data analysis to give biased results. To prevent this, extremely large price values were replaced with the maximum value.To examine the values, seoborns boxplot was used to visualize them.

[1] https://en.wikipedia.org/wiki/Outlier

```
ax = sns.boxplot(x=named_rent["Rent"])
```



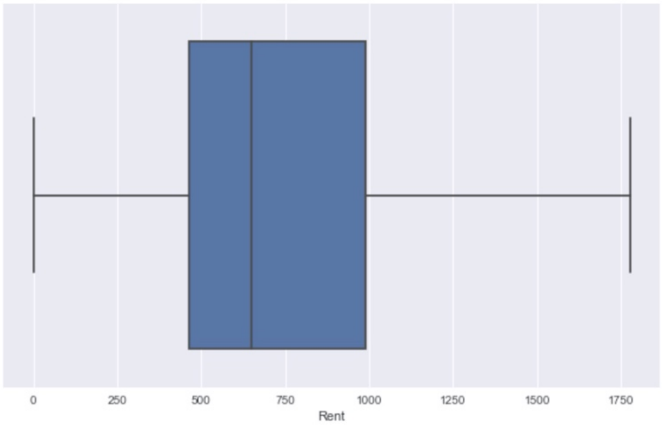

```
ax = sns.boxplot(x=named_price["Price"])
```



According to the Boxplots, that shows Price and Rent distribution, there is irrational Outlier values. Lets make them reasonable.

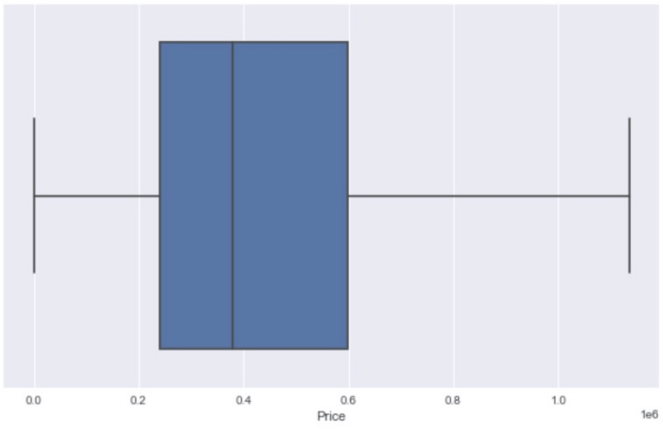
```
def max_outlier_setting(clmn):
    Q1 = clmn.quantile(0.25)
    Q3 = clmn.quantile(0.75)
    IQR = Q3 - Q1
    max_value = Q3 + 1.5 * IQR
    return max_value
```

```
ax = sns.boxplot(x=rent_map["Rent"])
```




```
price_map.Price.loc[price_map.Price > max_outlier_setting(price_map['Price'])]  = max_outlier_setting(price_map['Price'
rent_map.Rent.loc[rent_map.Rent > max_outlier_setting(rent_map['Rent'])]  = max_outlier_setting(rent_map['Rent'])
```

```
ax = sns.boxplot(x=price_map["Price"])
```



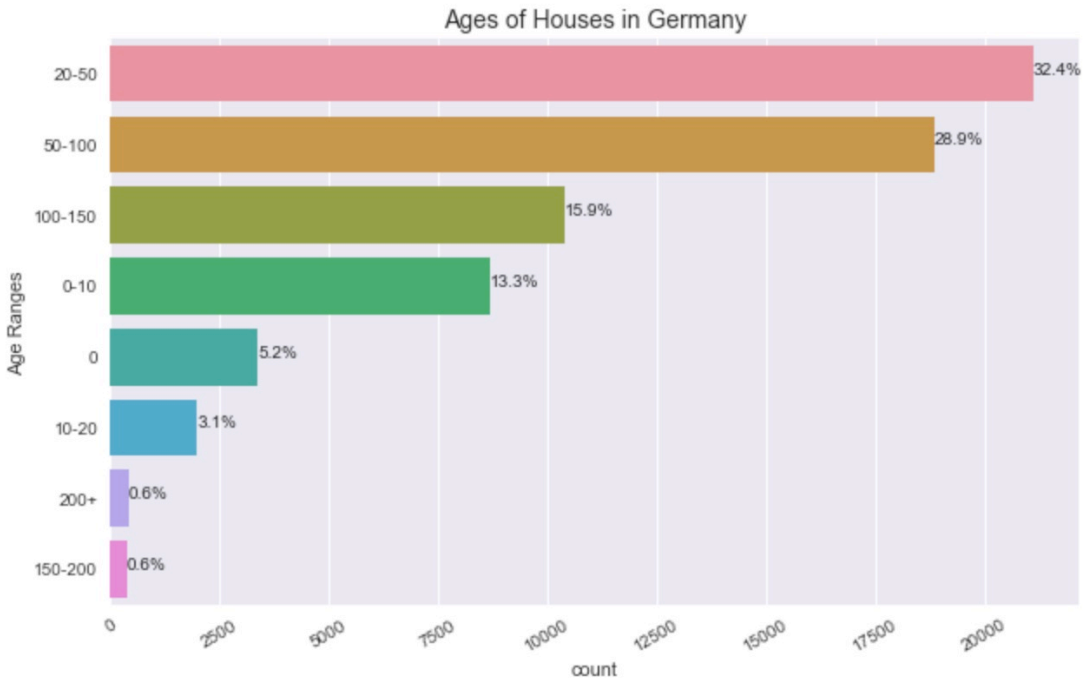
It can be seen from the boxplots that the distribution of the data is more reasonable after suppressing the extremely high values to the maximum value.

B.2 Explotary Data Analysis

What parameters can a property investor consider when buying a home?

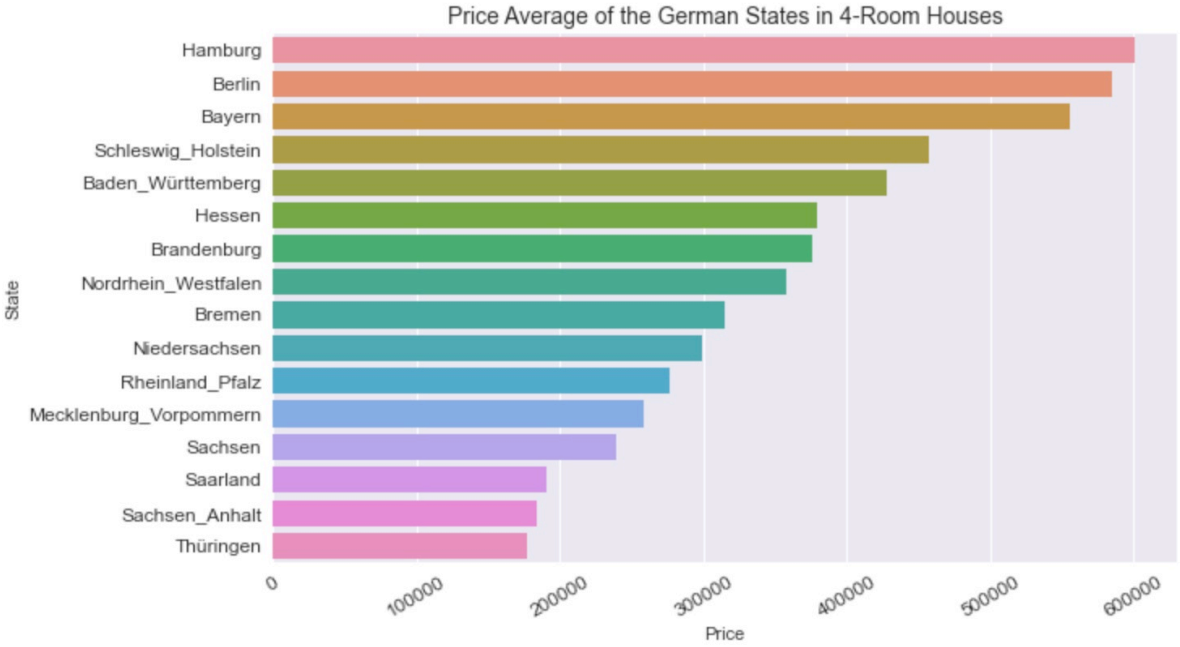
We tried to shed light on investors with analyzes and visualizations such as age ranges of houses, price and rent comparisons of states and cities with each other, and how many years the price of a house corresponds to the rent amount.

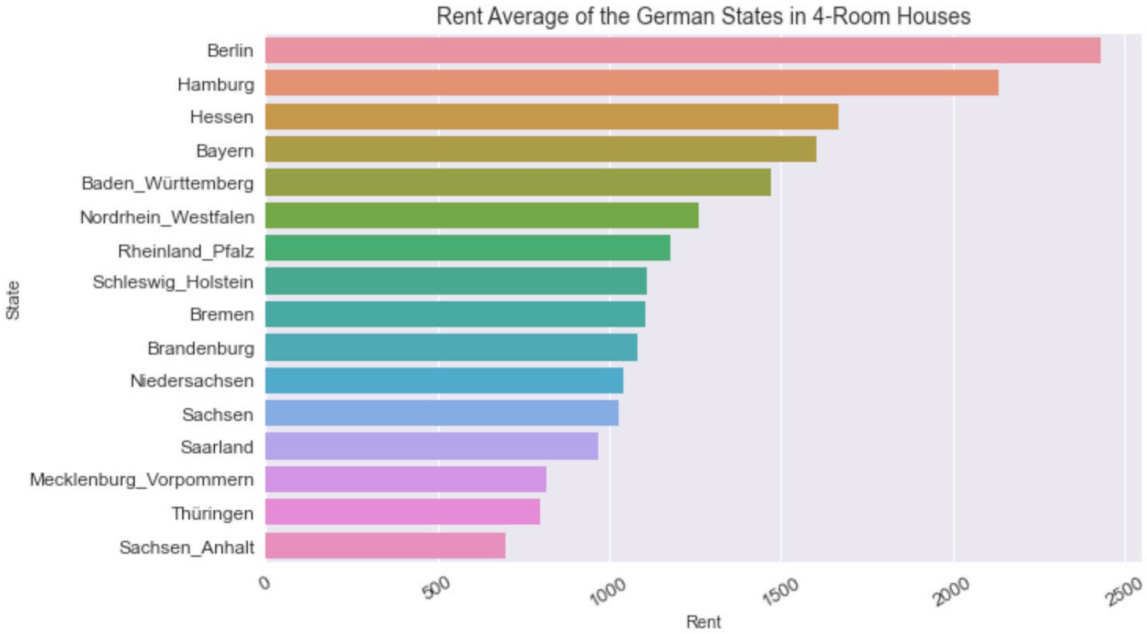
Lets have a look to the age ranges of the houses in Germany.



As can be seen in the graph, 32.4% of the houses in Germany are between 20 and 50 years old, while 18.5% of all houses are in the under 10 age group.

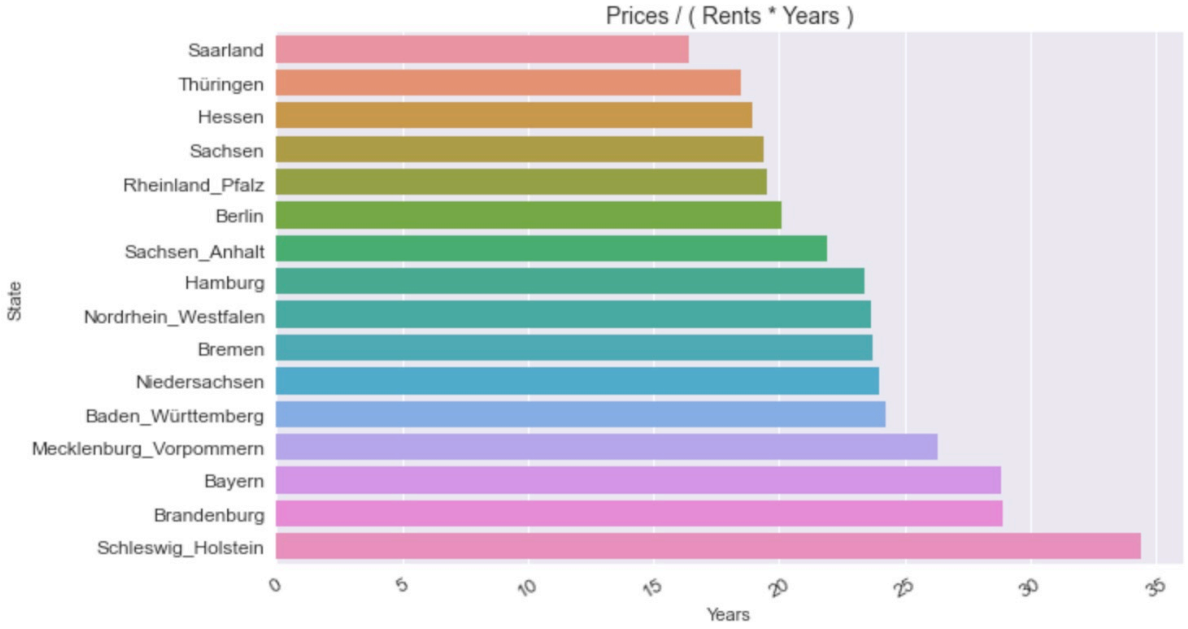
In the next stage, a new dataframe was created so that only the 4-room houses for rent and sale will be taken. By taking equal number of rooms, an attempt was made to prevent uneven distribution in the data from biased the analysis result. The state averages were visualized sequentially with the help of the seaborn library.





The most expensive states are Berlin and Hamburg.Thüringen and Sachsen-Anhalt states are relatively cheaper.

Another important parameter for an investor is the rental value of a house purchased. For this, again, on average, it is visualized in which state the price of a house is covered by how many annual rents.



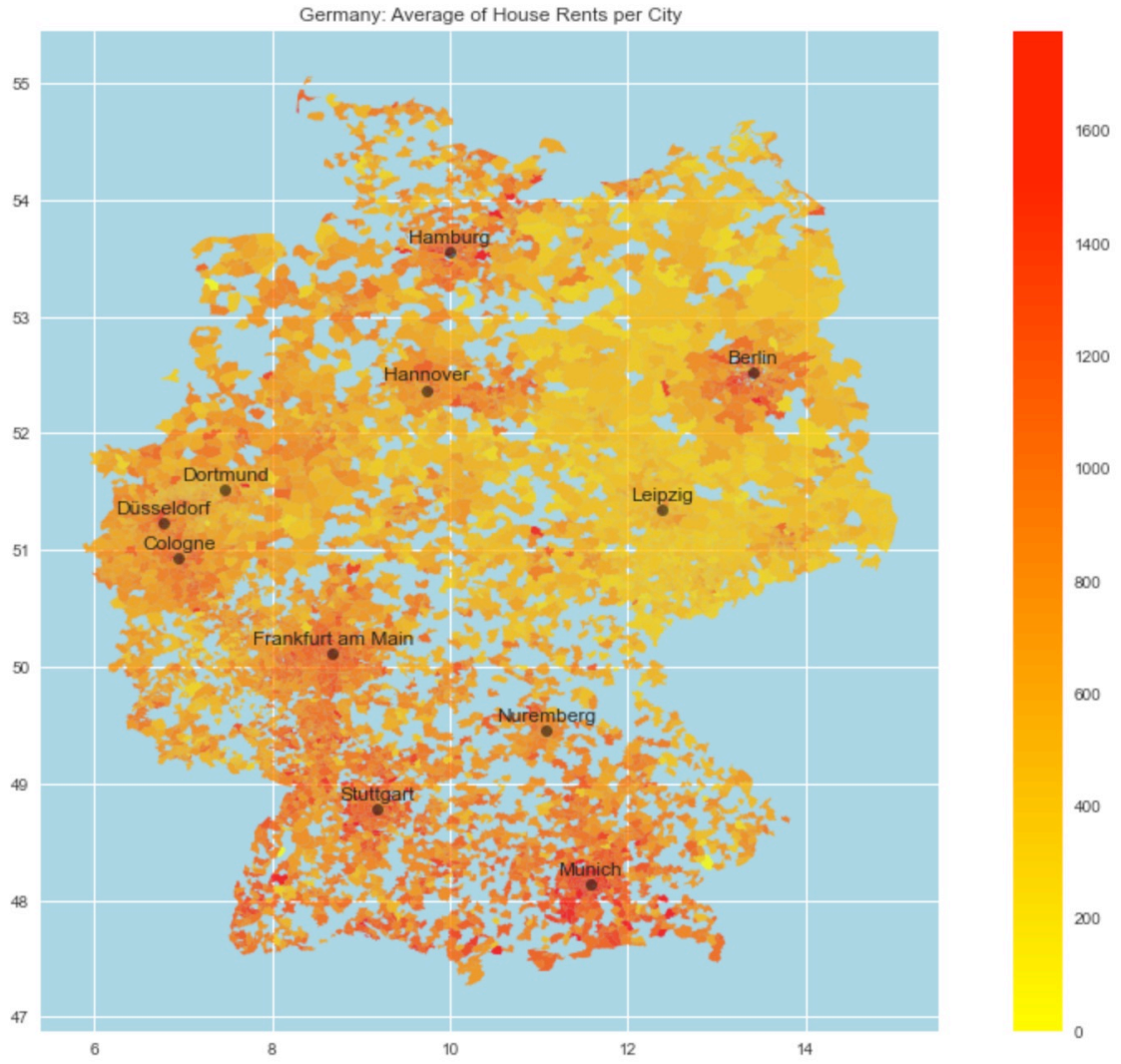
According to the data we have, the houses that meet the purchase price with rental income fastest are in the state of Saarland.

B.3 Visualization of price and rent distributions on the map

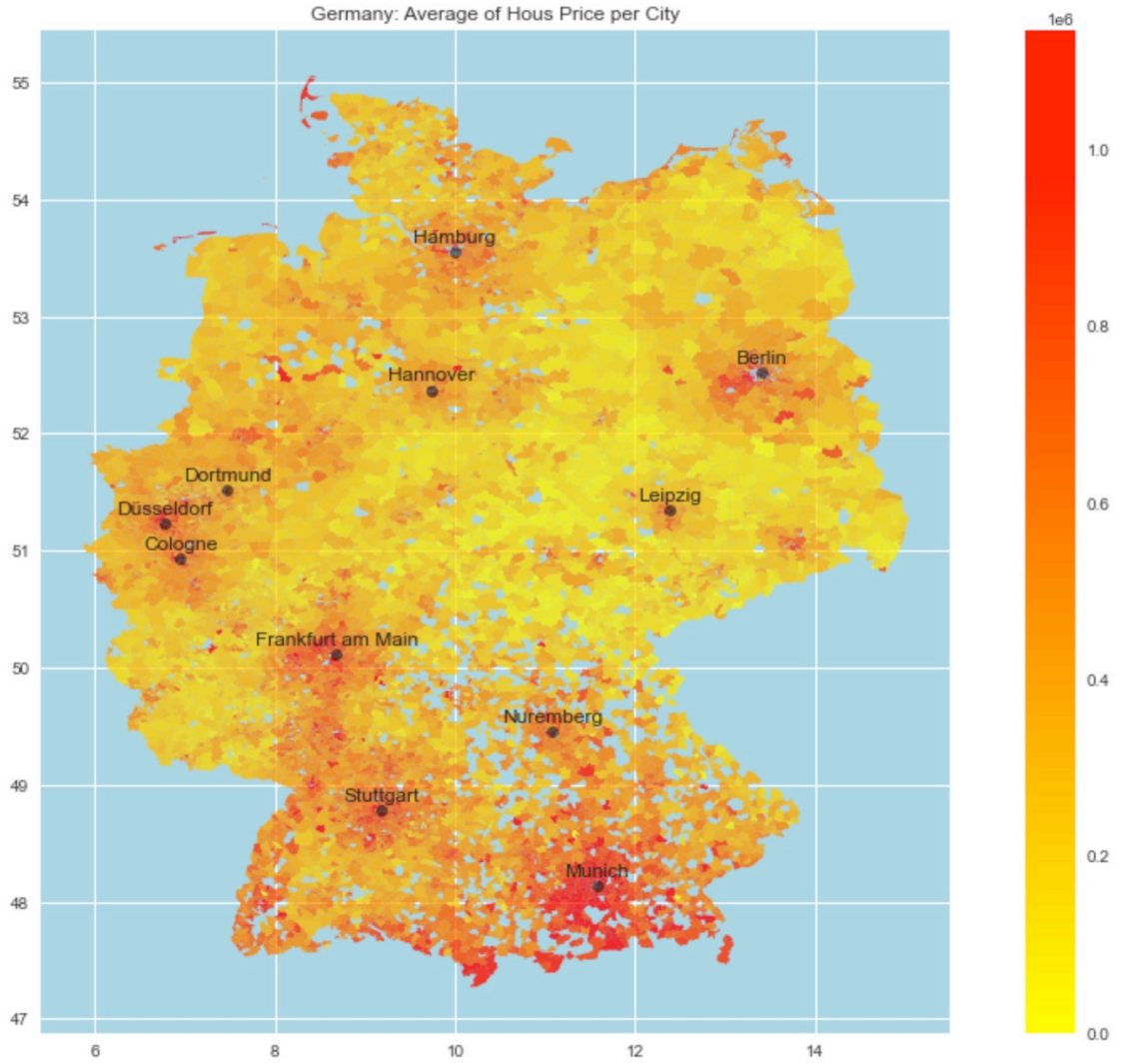
The price distribution was visualized on the map with the geopandas library.

GeoPandas is an open source project to make working with geospatial data in python easier. GeoPandas extends the datatypes used by pandas to allow spatial operations on geometric types. Geometric operations are performed by shapely[2].

In this way, it was aimed to give investors who want to benefit from the analysis, an idea about the geographical distribution of rents and prices over the entire country map. Major cities are indicated by writing their names on the map. Thus, it was aimed to make the geographical location easier to notice.



[2] https://geopandas.org/



Findings

Looking at the map, it can be concluded that house prices towards the South and West regions of Germany, as well as in Hamburg and Berlin, are relatively high.

It can also be said that rental prices higher around top cities.

B.3 Neighborhoods and House Price Distribution in Hamburg

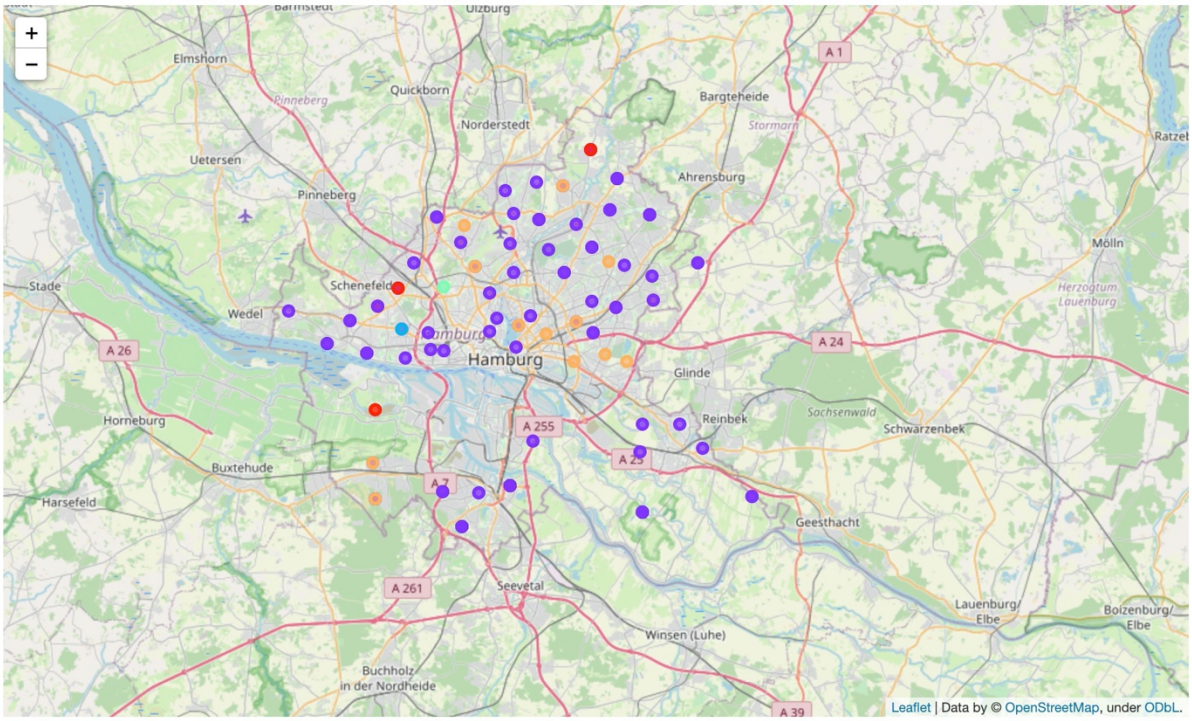
Considering the price and rent distributions, the investment distribution in the neighborhood of the city of Hamburg has been classified in order to serve as an example for the investors who want to make any commercial venture in the vicinity.

The Forsquare API was used to access neighborhood information. K-Means, one of the machine learning algorithms, was used for this clustering.

The K-Means algorithm is one of the most popular unsupervised machine learning algorithms. Normally, the unsupervised algorithms make inferences using unlabelled dataset. The goal of this unsupervised machine learning technique is to find similarities in the data point and group similar data points together. The K-Means clustering algorithm, it aims to

partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

Python's ***Folium*** library was used to visualize the clusters obtained by machine learning. With different color points, it was visualized which regions had similar characteristics with each other.



C. Results

The results obtained were briefly expressed under the analysis. In summary;

- The southern regions of Germany and the metropolitan areas were found to be more expensive in terms of both purchase prices and rents. In addition, the age distribution of the houses in Germany was seen.
- An idea has been obtained about the age distribution of the houses in Germany.
- To serve as an example, clusters of investment regions of the city of Hamburg were seen.

D. Discussion

As mentioned before, Germany is one of the important countries of Europe and the World. With this analysis, it was tried to shed light on the real estate market in Germany.

Technically, how successful were the visualization methods and machine learning algorithm used?

Financially, do the reviews in the analysis offer sufficient contribution?

These two issues can be developed with the contribution of consensus and better results can be obtained. Therefore, after this study is published, it will continue to be developed in line with the valid feedback obtained.

E. Conclusion

This analysis can be improved over time as the quality and quantity of data increases. Other variables that characterize a neighborhood such as population demographics, population density, crime statistics, etc., could be incorporated in a future analysis of the subject. Data from Google Maps on venues in each neighborhood is a promising option to complement the Foursquare data that we have now. Also, dynamic updating of our data on average rent prices from an online portal would be a more robust option than the static dataset that we are using now.

Taner Öztaş