Team member:

Kai Yung (Adam)
Joffrey
Mathieu

Challenge: Data Analysis

USE PANDAS, DATA VISUALISATION LIBRARIES

TO ESTABLISH CONCLUSIONS ABOUT A DATASET.

Mission & Objectives

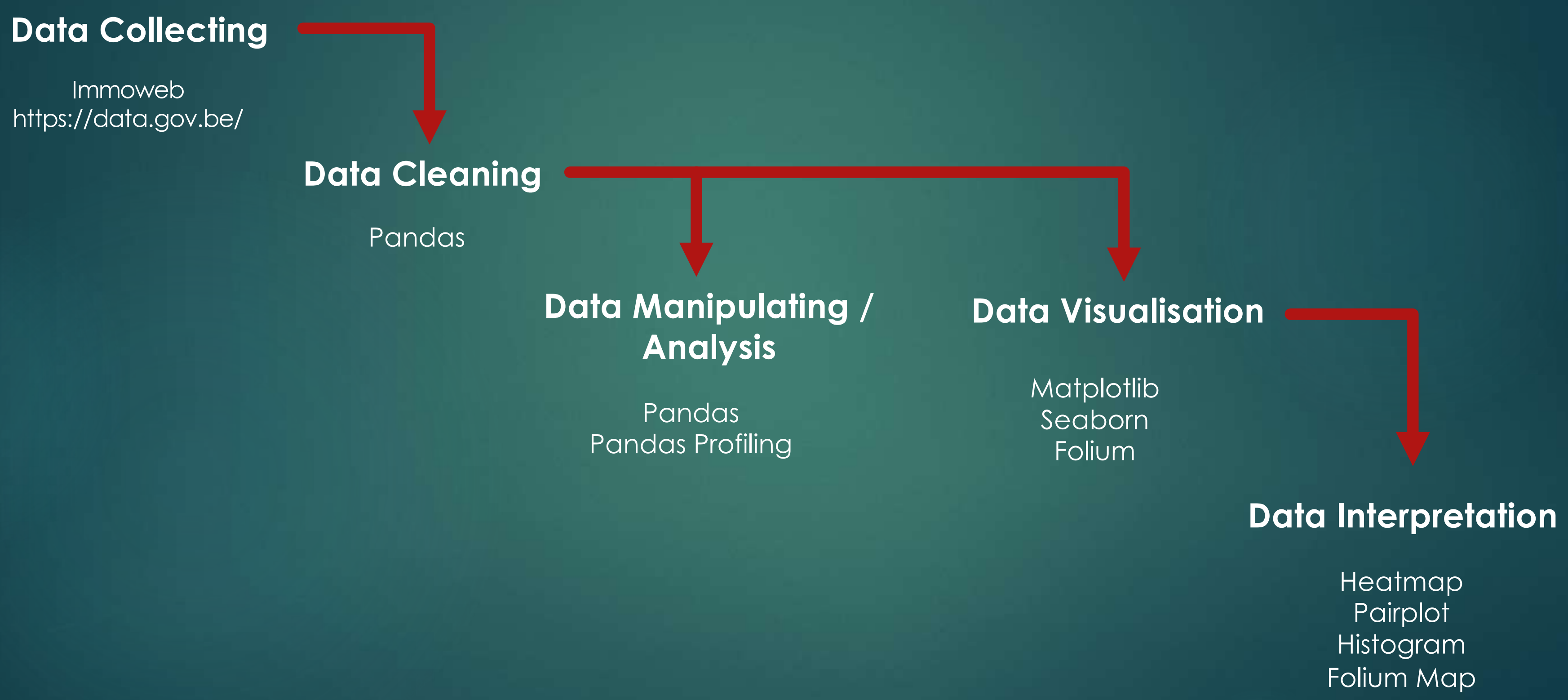
Mission:

- Cleaning and doing a complete analysis and interpretation of the dataset created during the previous challenge.
- In order to create a machine learning model to predict prices on Belgium's real estate's sales.

Objectives:

- Using Pandas for data manipulation.
- Using Matplotlib and/or Seaborn for plotting.
- Finding and understanding correlations between dataset's variables.

Flow Process



Data Collecting

Immoweb
https://data.gov.be/

Data Cleaning

Pandas

Data Manipulating / Analysis

Pandas
Pandas Profiling

Data Visualisation

Matplotlib
Seaborn
Folium

Data Interpretation

Heatmap
Pairplot
Histogram
Folium Map

Data Collecting

- A dataset of 52077 real estate's observations, previously scrapped by Becode colleagues during a previous challenge were used for this data analysis challenge.

Why this dataset ?

- It has a lot of entries : more than 50k ! By having the maximum amount of data to discover interesting correlations, and have a meaningful analyse.
- As there are raw data scrapped from web, there were lots of duplicated, null values in the dataset. This enable the team to clean the uncomplete/corrupted data.
- And at the same time, It was scrapped from ImmoWeb: probably the biggest real estate website of Belgium.

Data Cleaning

Identifying the needs:

To proceed to the analysis, we needed a clean dataset containing at least:

- Prices, postal code and city names.
- A price/m2 column

Removing the outliers (error, incorrect or absurd).

- It's good to have a lot of columns, as it can create more correlations between them. However, it's bad to have columns with errors, incorrect, missing or absurd values.

Data Cleaning

Two phases of data cleaning:

1. **Cleaning the raw:**

- A very first clean to the raw data. We were focused on "**dropping the big lies**":
- **Dropping** the duplicated rows
- **Dropping** columns with unique value
- **Checking** each columns' properties

2. Refining the values

- Some tweaks were made on the dataset to **remove outliers and useless columns**, due to their high rate of *None* value. This step required deeper investigation in top the data.

Data Cleaning

Details:

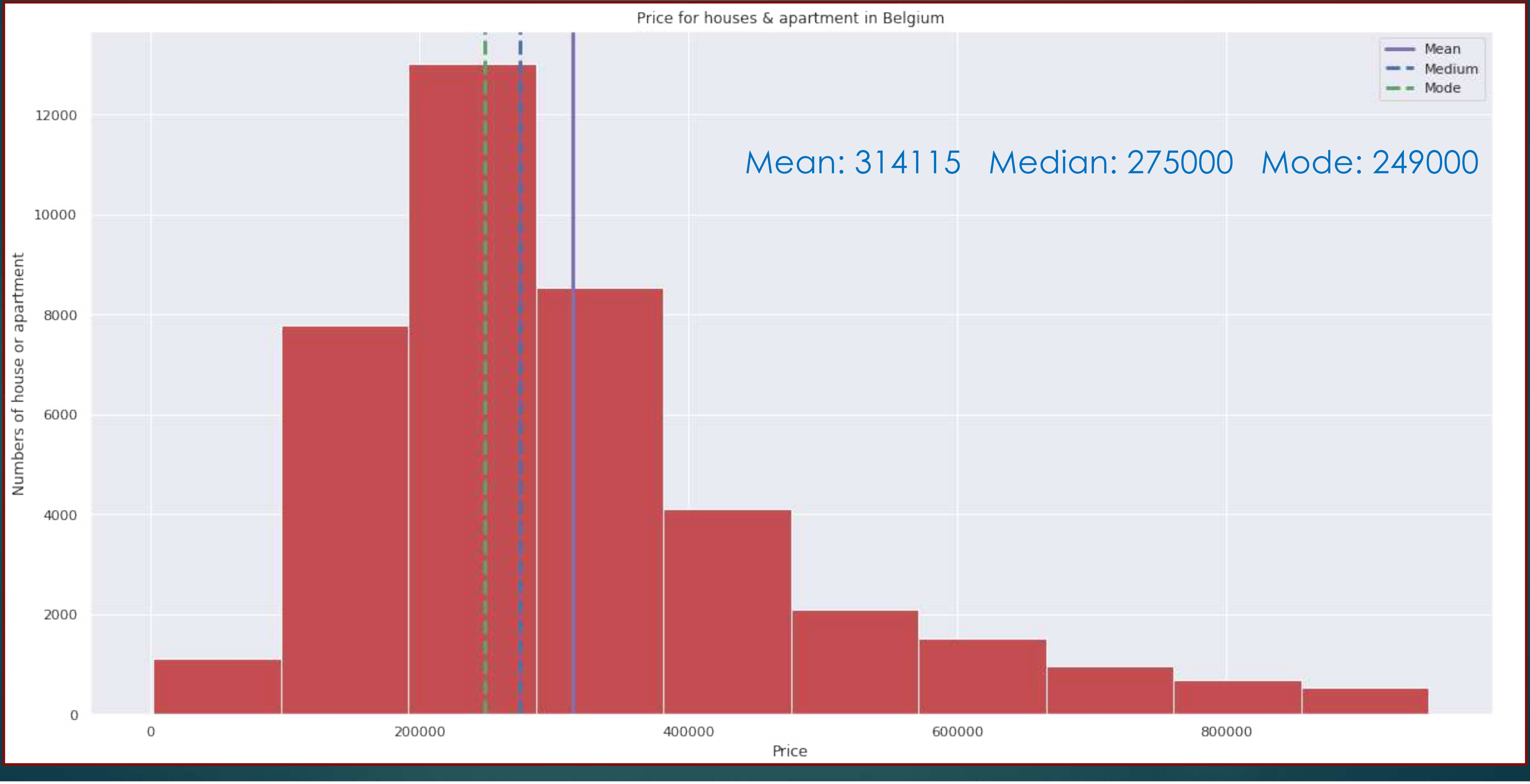
- Dropping "terrace_area" column
  - It has more than 30% of None.
- Dropping "garden_area" column
  - It has more than 50% of None.
- Dropping "subtype" column
  - Lots of property subtype. Some with less than 100 entries, in a dataset of 50.000.
  - This column was not relevant.
- Removing the "Apartment blocks" entries
  - Apartment blocks are a whole building. It's not the kind of real estate sales we want here.
- Changing None to "unknow"

- We also refactored all *float* to *int*. At the end of the cleaning, **we merged our dataframe with the two other ones created during the request study**.

40395 rows , 18 columns

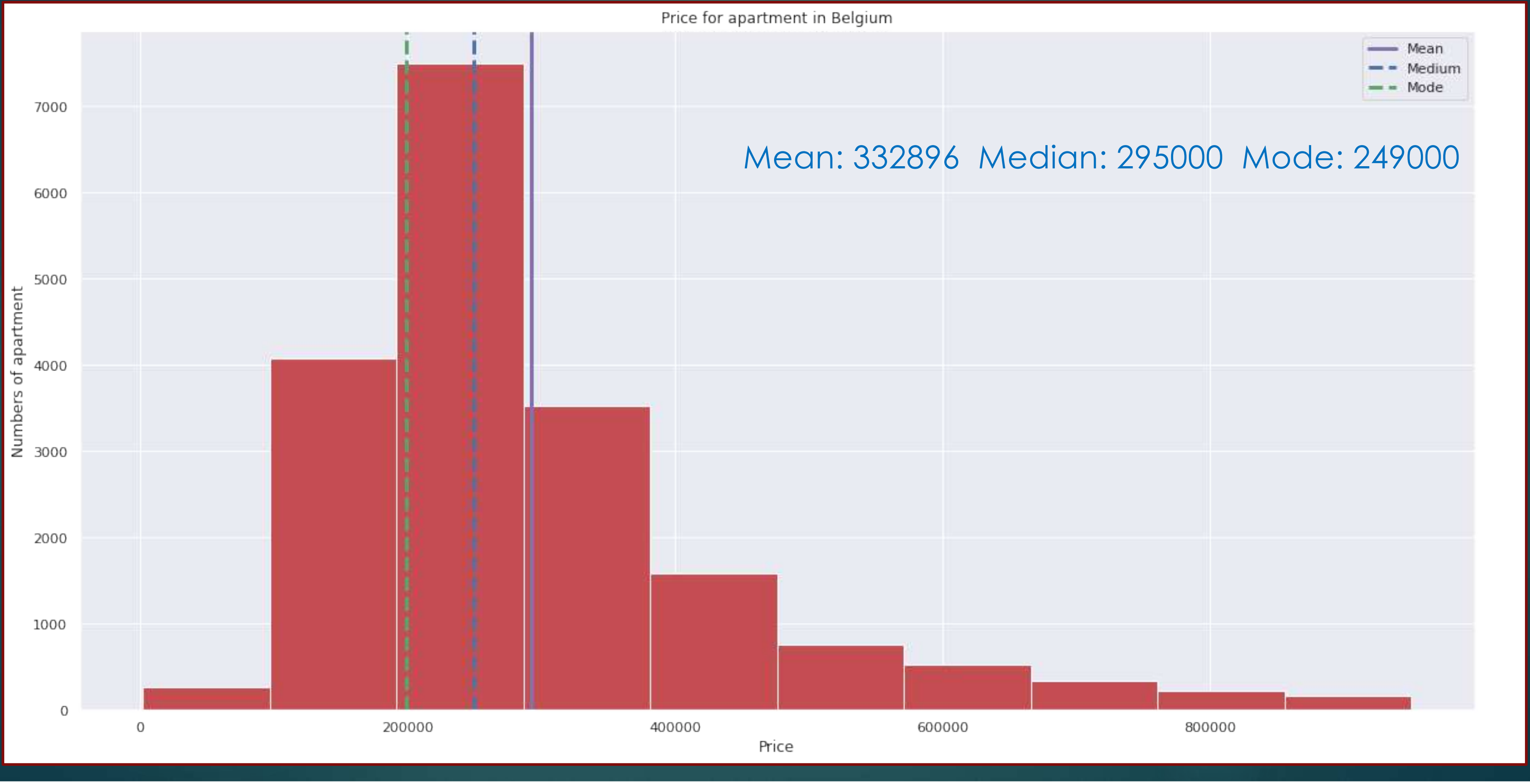
Data Visualisation

Our target: The Price



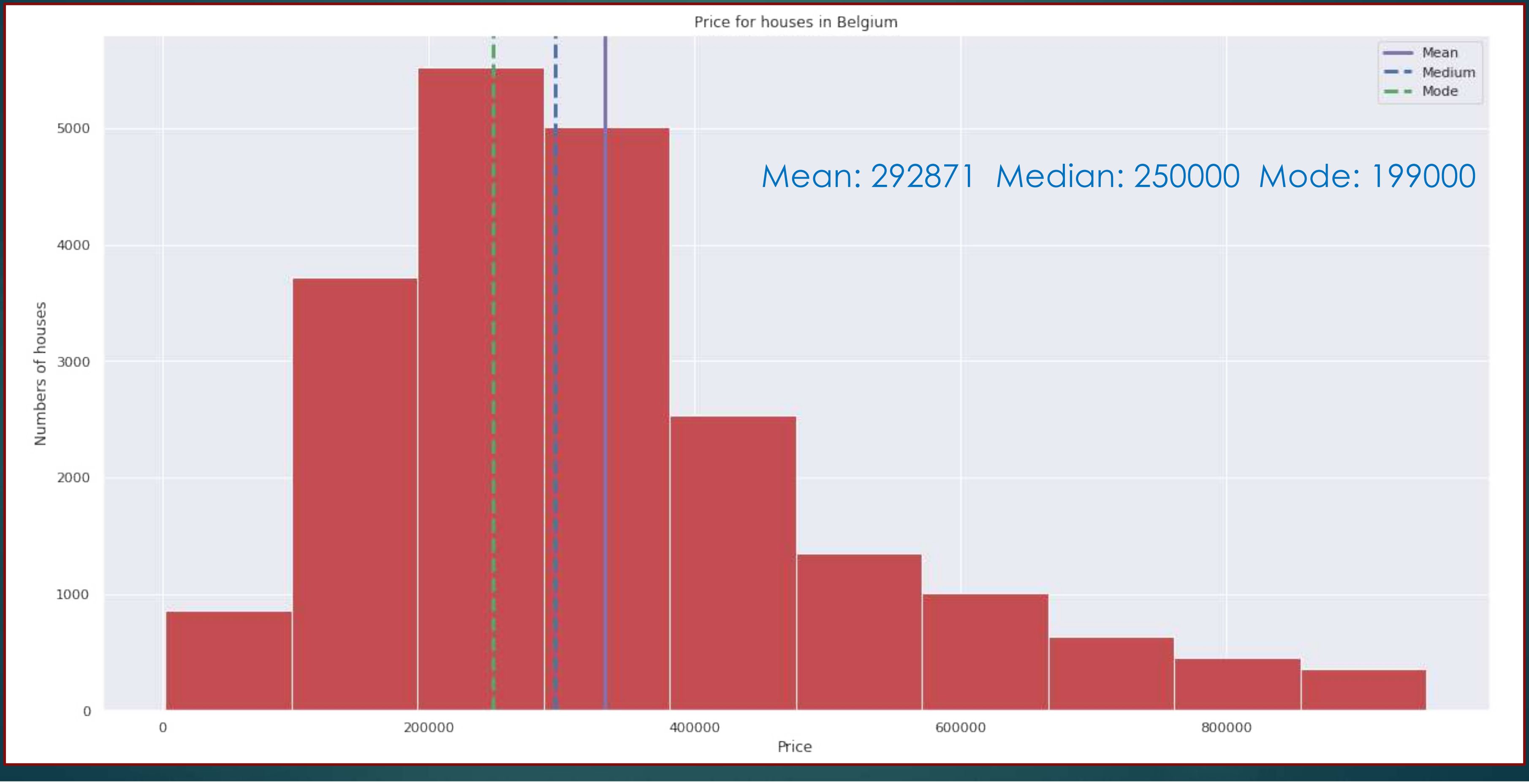
Data Visualisation

Our target: The Price



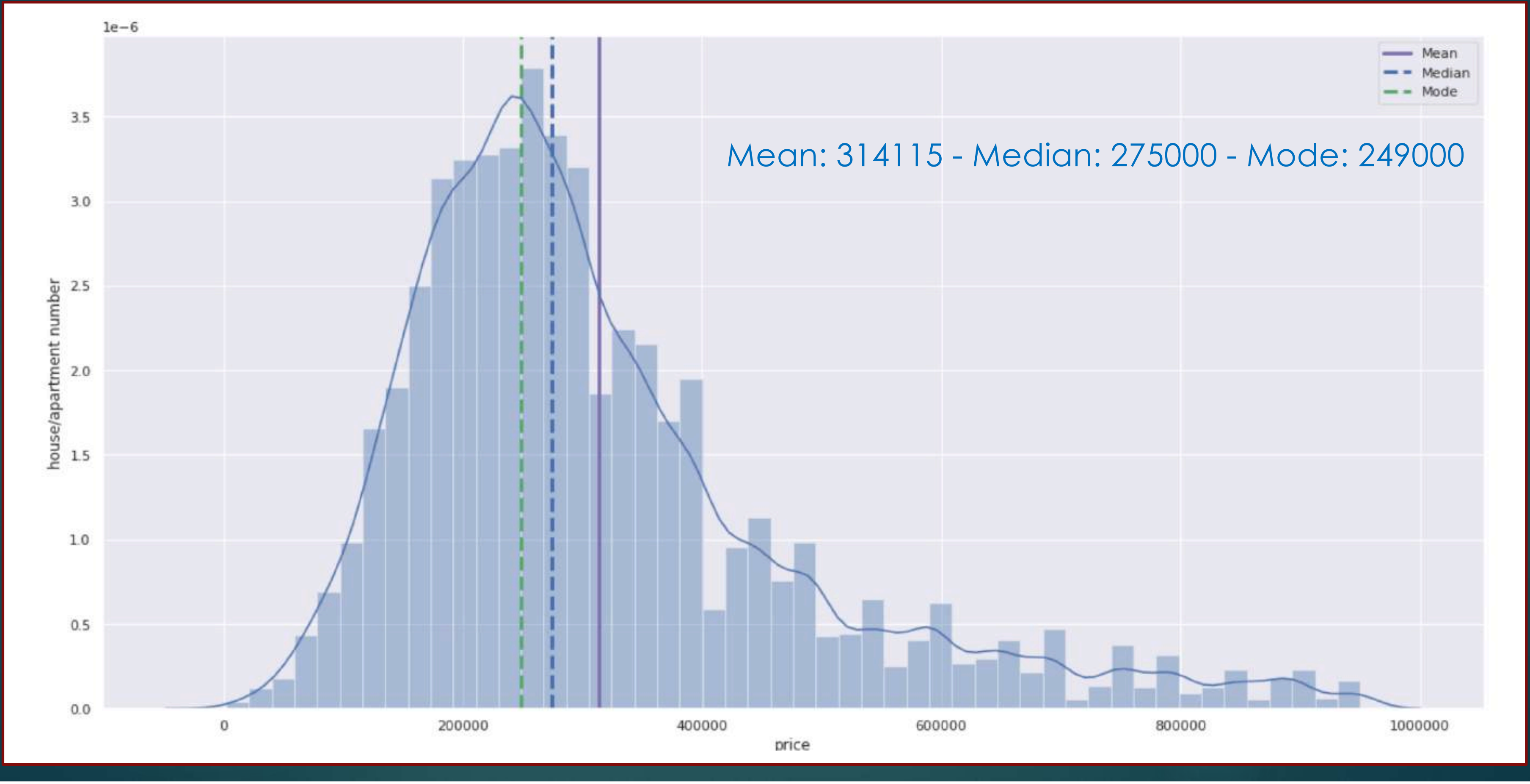
Data Visualisation

Our target: The Price

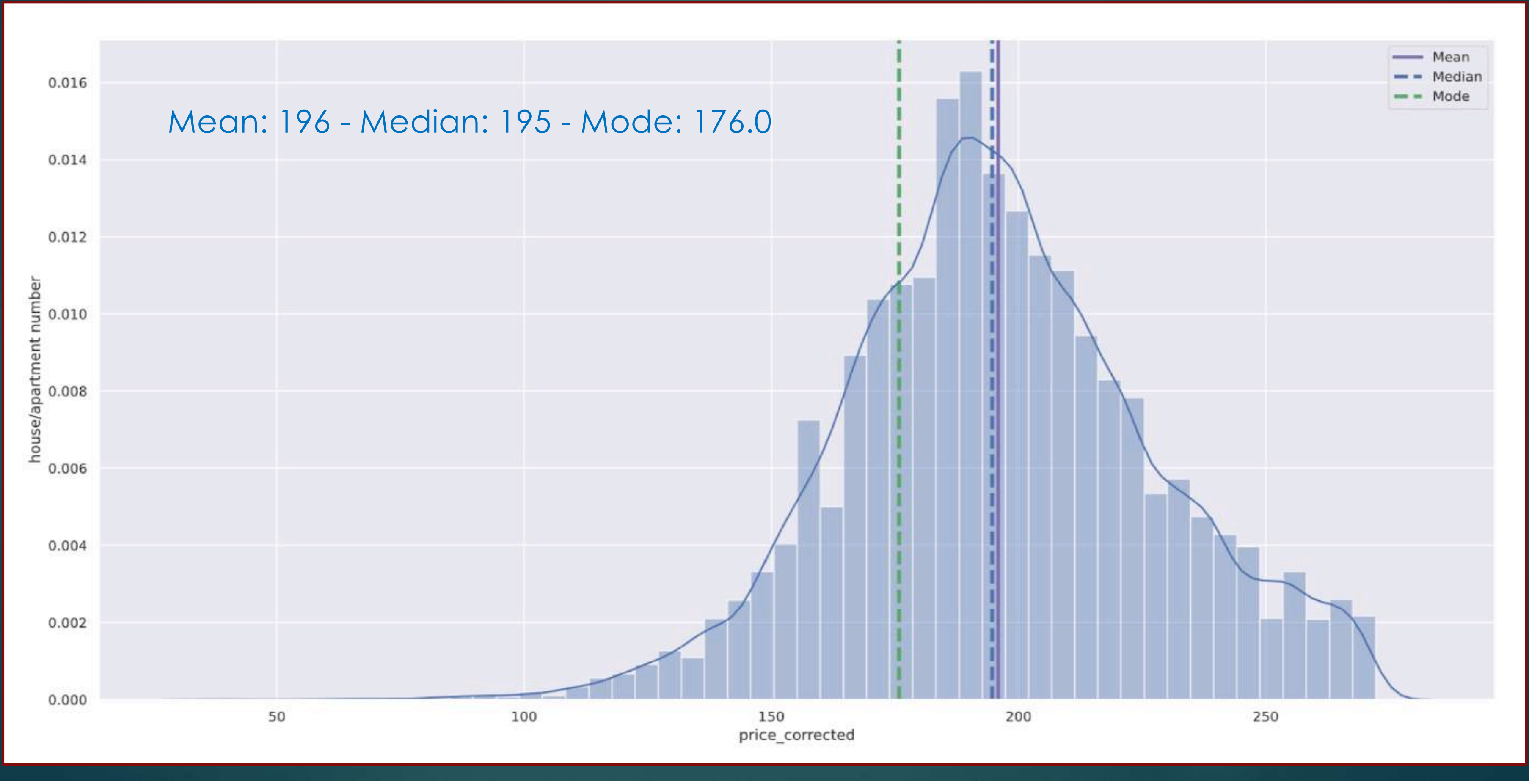


Mean: 292871 Median: 250000 Mode: 199000

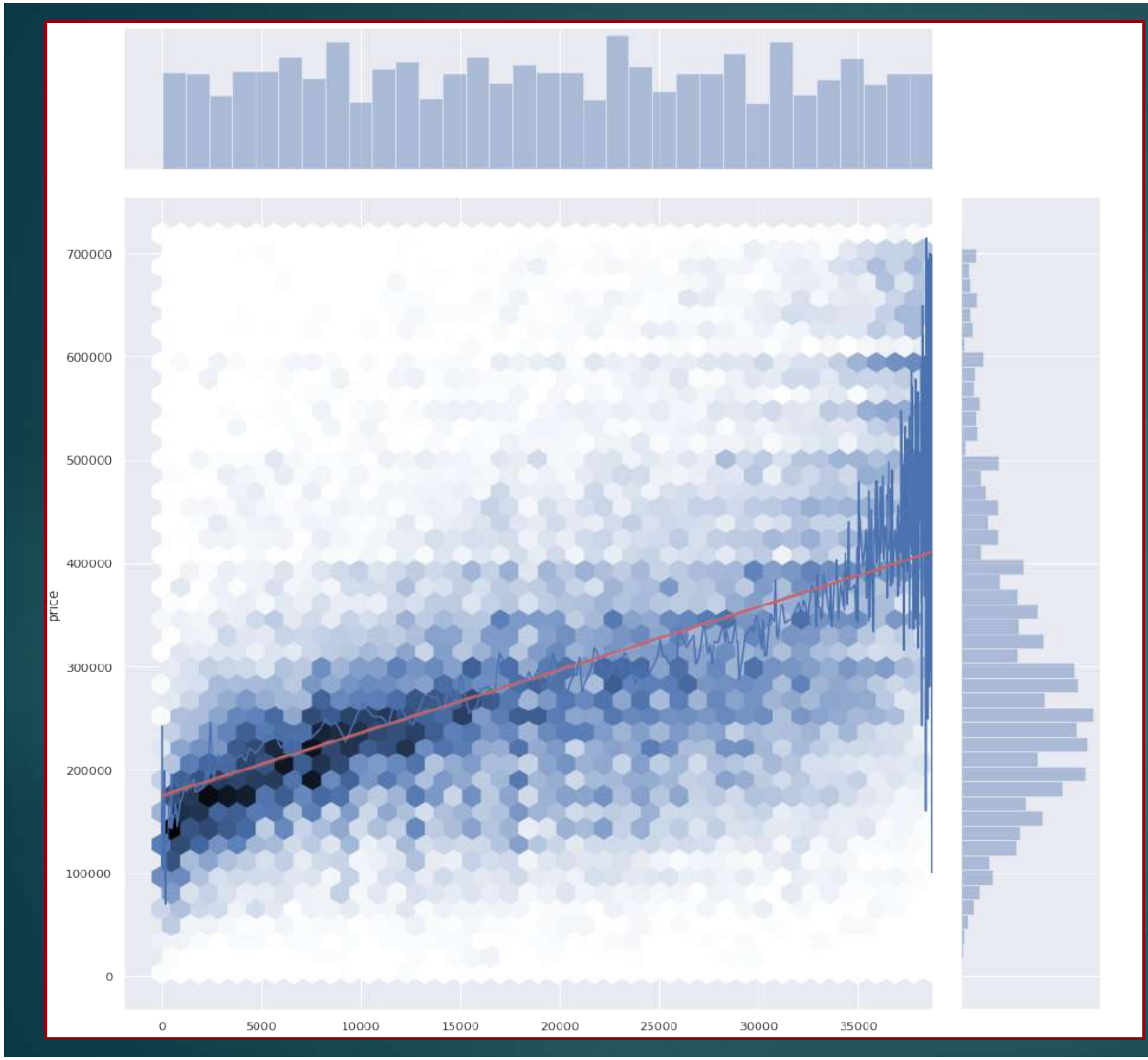
Data Visualisation



Data Visualisation



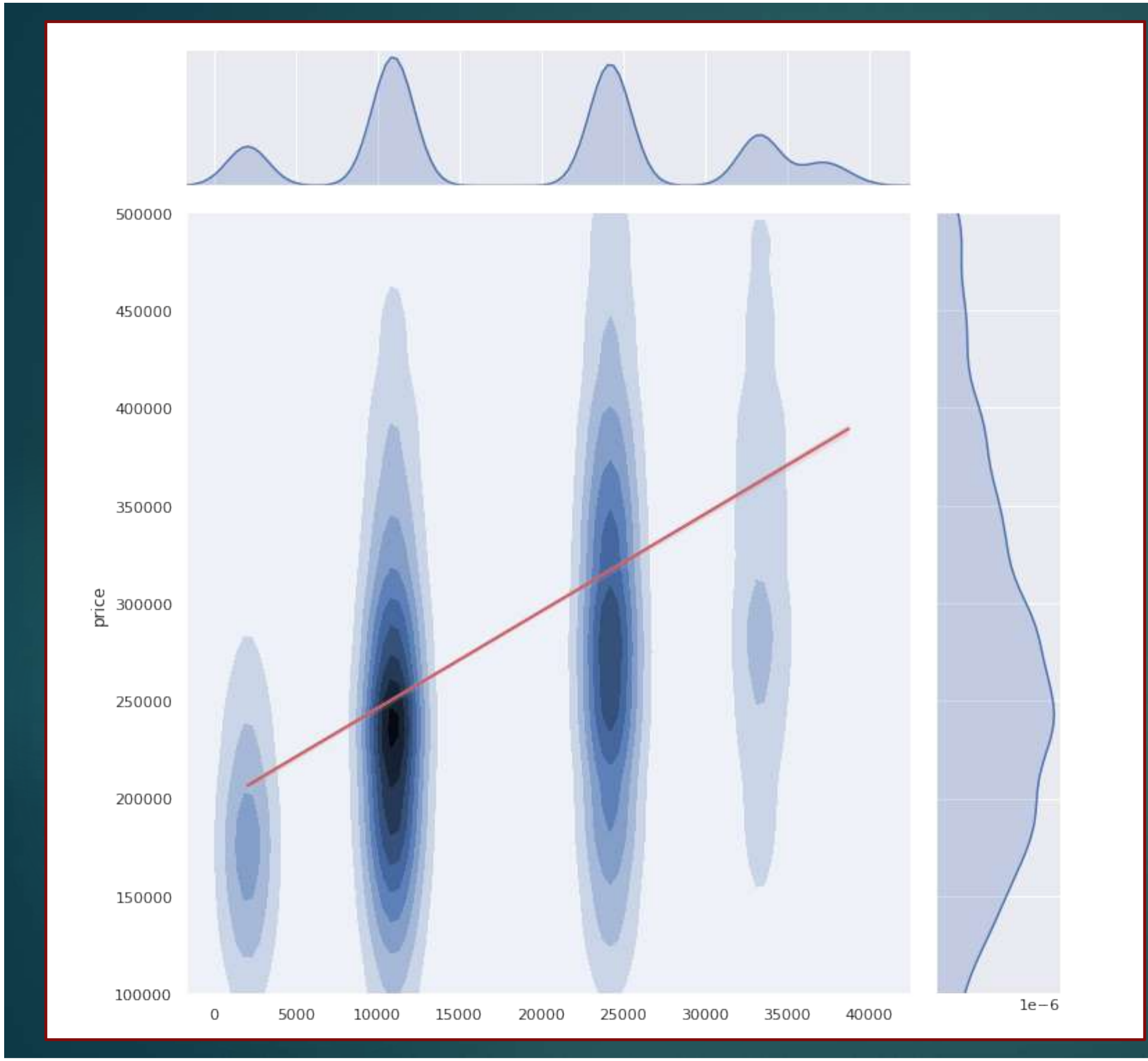
Mean: 196 - Median: 195 - Mode: 176.0



Data Visualisation

1.The Price of a house is correlated with its area: **The higher is the area, higher is the price.**

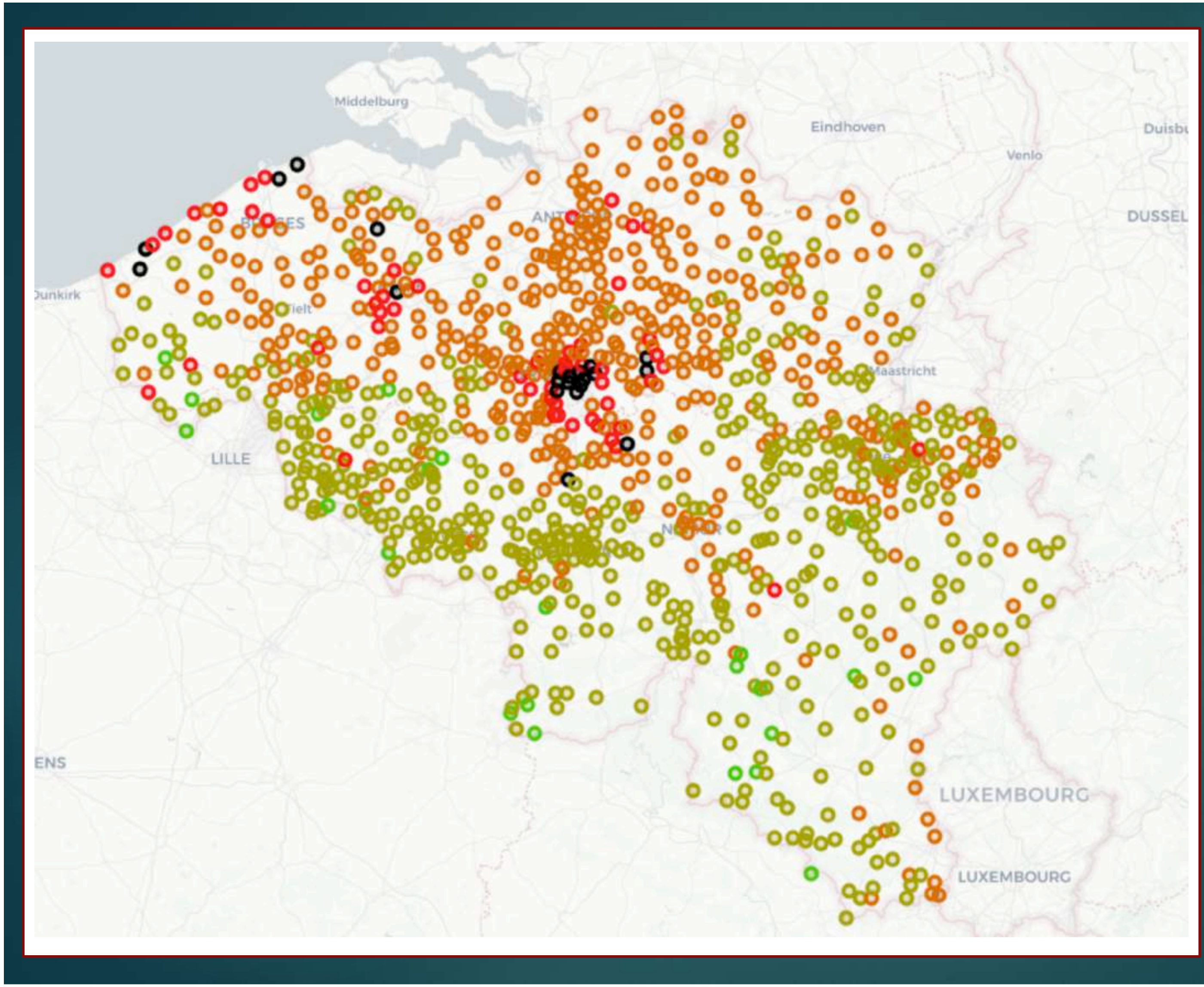
2. However, this correlation is not very strong, especially for big houses (houses with a area bigger than 35000 m2): The Price may vary a lot ! It may have other factor that influence the price of "big" houses.



Data Visualisation

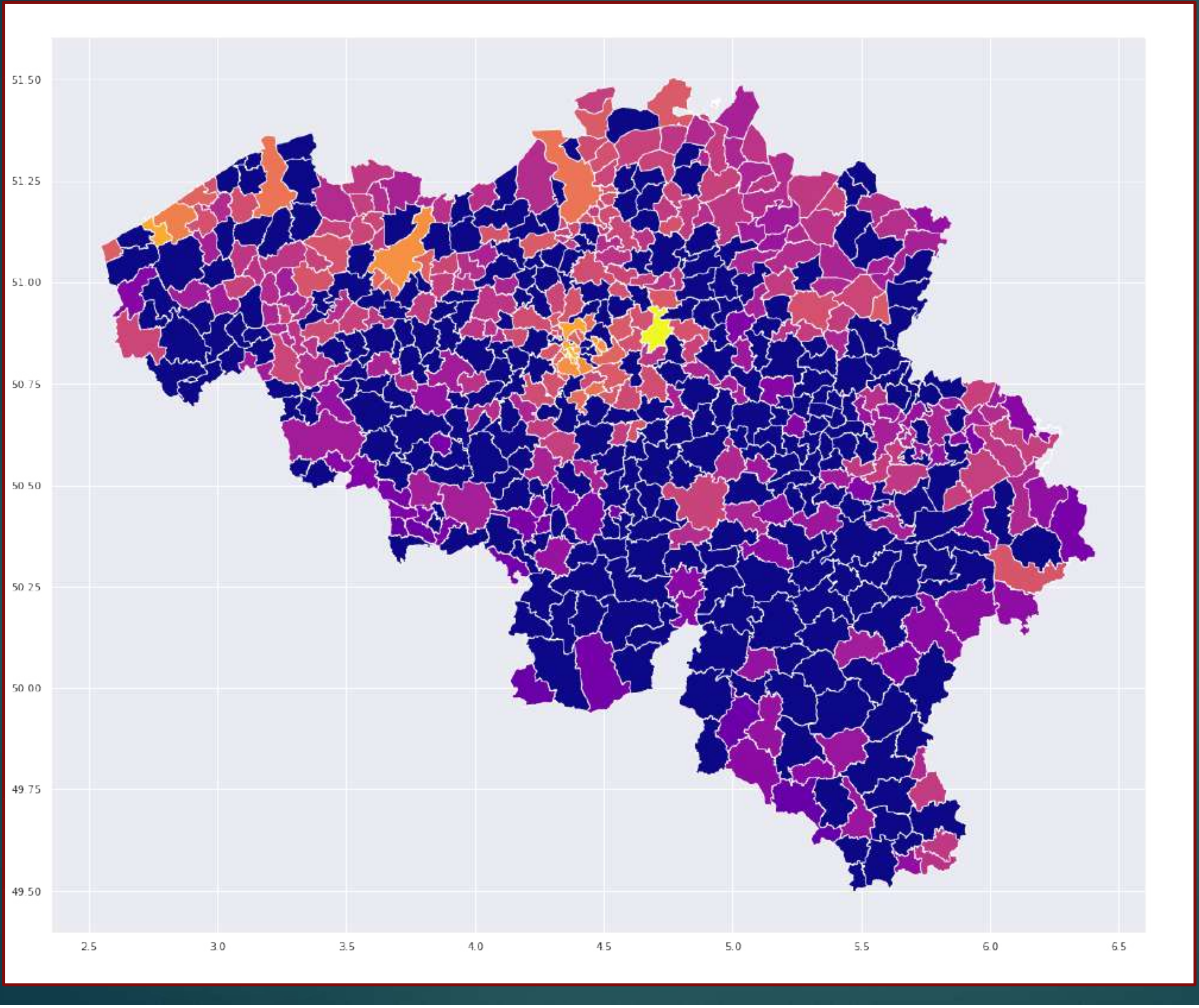
1. The Price of a house is correlated with the number of rooms: **More rooms tend to increase the price.**

2. However, this correlation is weak. Maybe the number of rooms is correlated with the house area ?



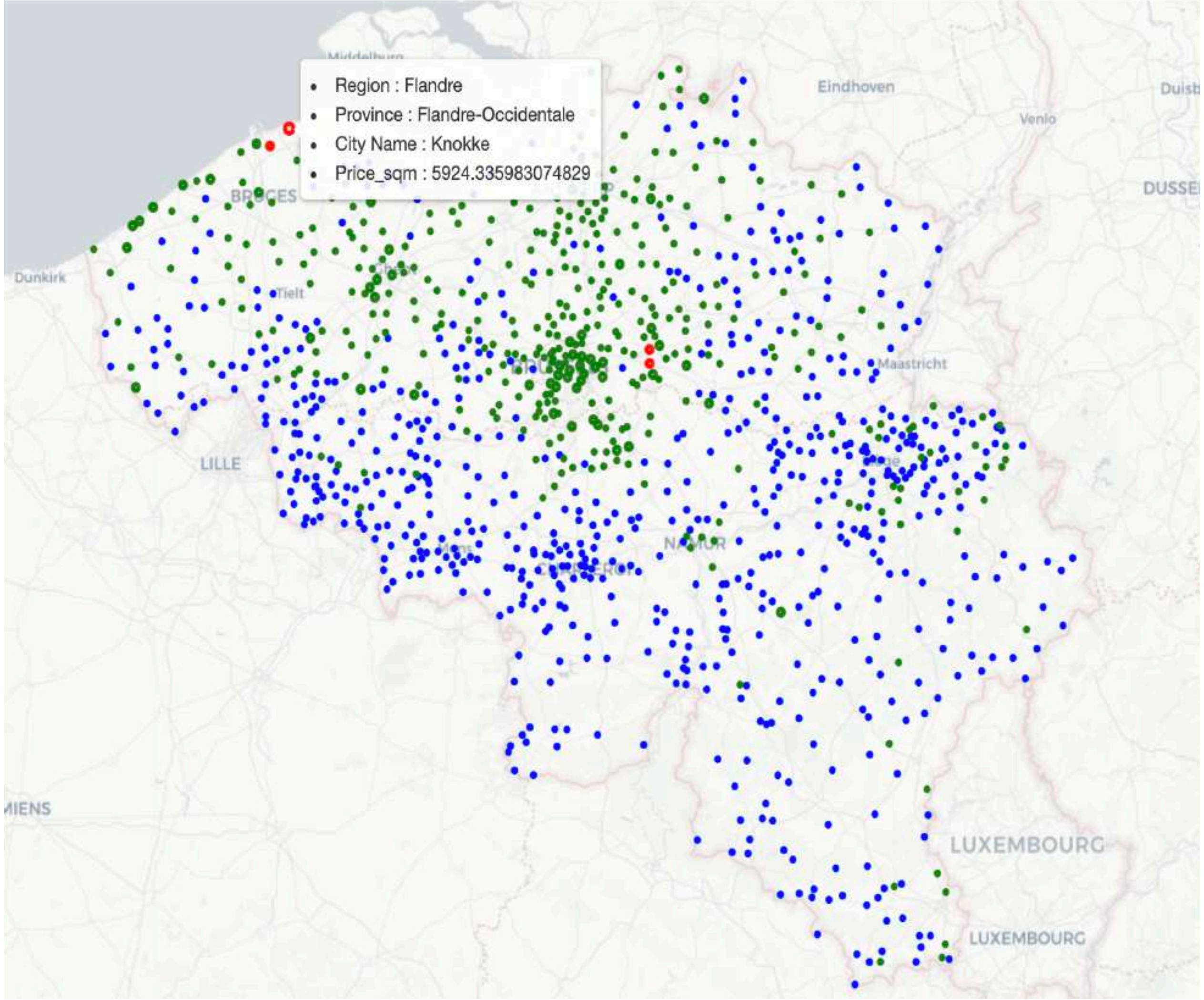
Data Visualisation

A map of all Belgian's municipalities with their median price/m2.



Data Visualisation

A map of all Belgian's municipalities with their median price/m2.



- Region : Flandre
- Province : Flandre-Occidentale
- City Name : Knokke
- Price_sqm : 5924.335983074829

Data Visualisation

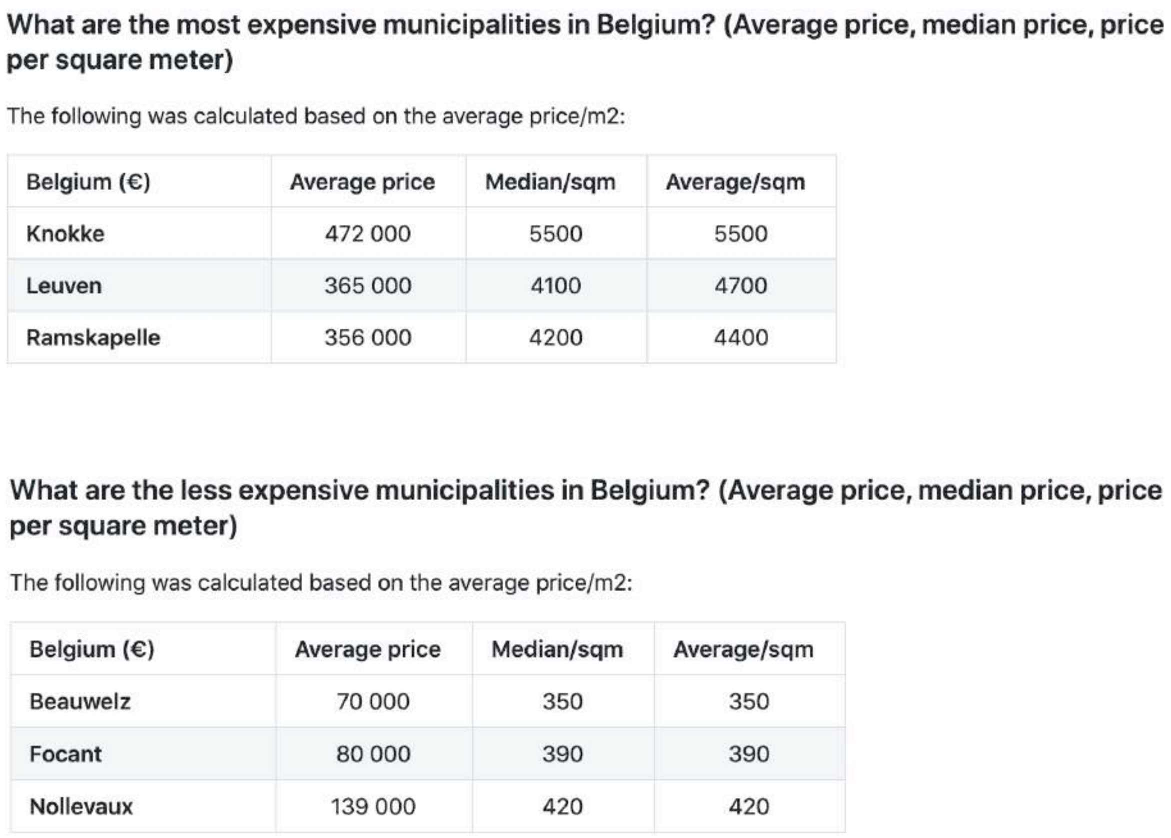
Average price/sqm for houses & apartments in Belgium:
https://kaiyungtan.github.io/challenge-data-analysis/Visualisation/average_price_per_sqm_belgium_house&apartment.html

Average price/sqm for apartments in Belgium:
https://kaiyungtan.github.io/challenge-data-analysis/Visualisation/average_price_per_sqm_belgium_apartment.html

Average price/sqm for houses in Belgium:
https://kaiyungtan.github.io/challenge-data-analysis/Visualisation/average_price_per_sqm_belgium_house.html

Data Analysis

Challenge's answers



What are the most expensive municipalities in Belgium? (Average price, median price, price per square meter)

The following was calculated based on the average price/m2:

Belgium (€)	Average price	Median/sqm	Average/sqm
Knokke	472 000	5500	5500
Leuven	365 000	4100	4700
Ramskapelle	356 000	4200	4400

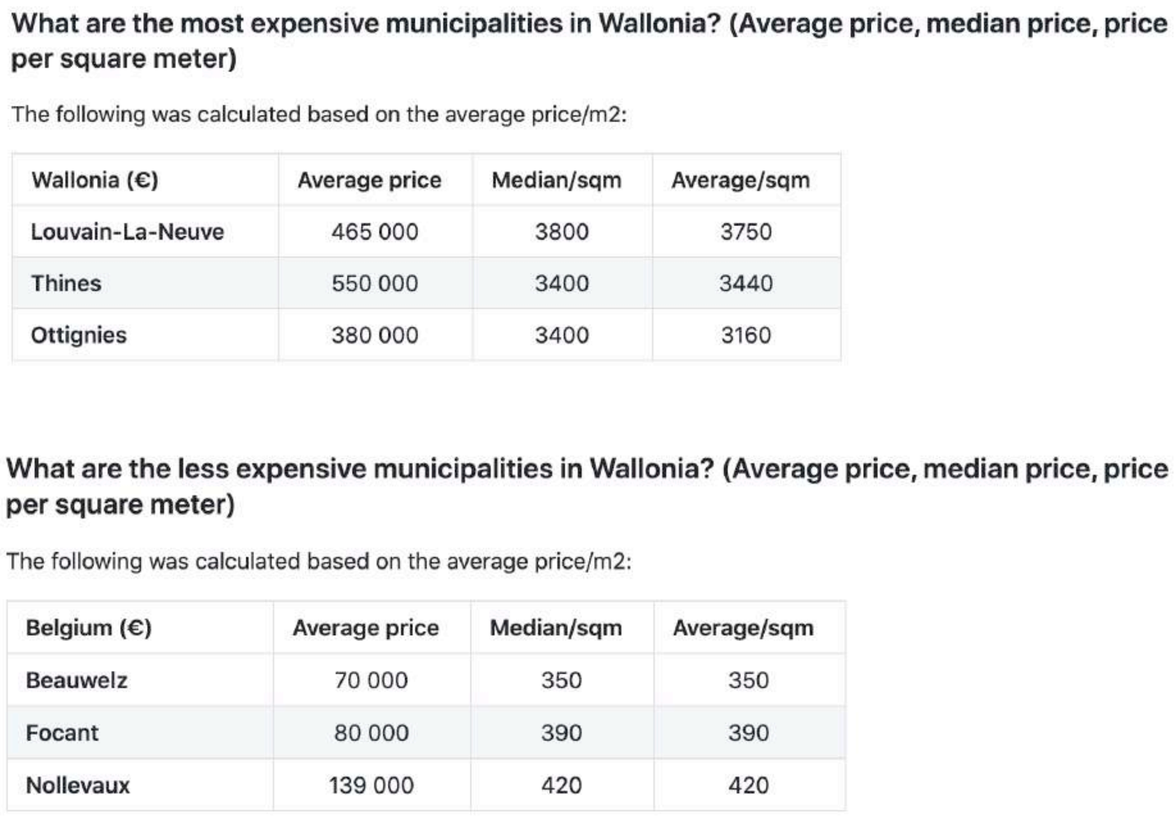
What are the less expensive municipalities in Belgium? (Average price, median price, price per square meter)

The following was calculated based on the average price/m2:

Belgium (€)	Average price	Median/sqm	Average/sqm
Beauwelz	70 000	350	350
Focant	80 000	390	390
Nollevaux	139 000	420	420

Data Analysis

Challenge's answers



What are the most expensive municipalities in Wallonia? (Average price, median price, price per square meter)

The following was calculated based on the average price/m2:

Wallonia (€)	Average price	Median/sqm	Average/sqm
Louvain-La-Neuve	465 000	3800	3750
Thines	550 000	3400	3440
Ottignies	380 000	3400	3160

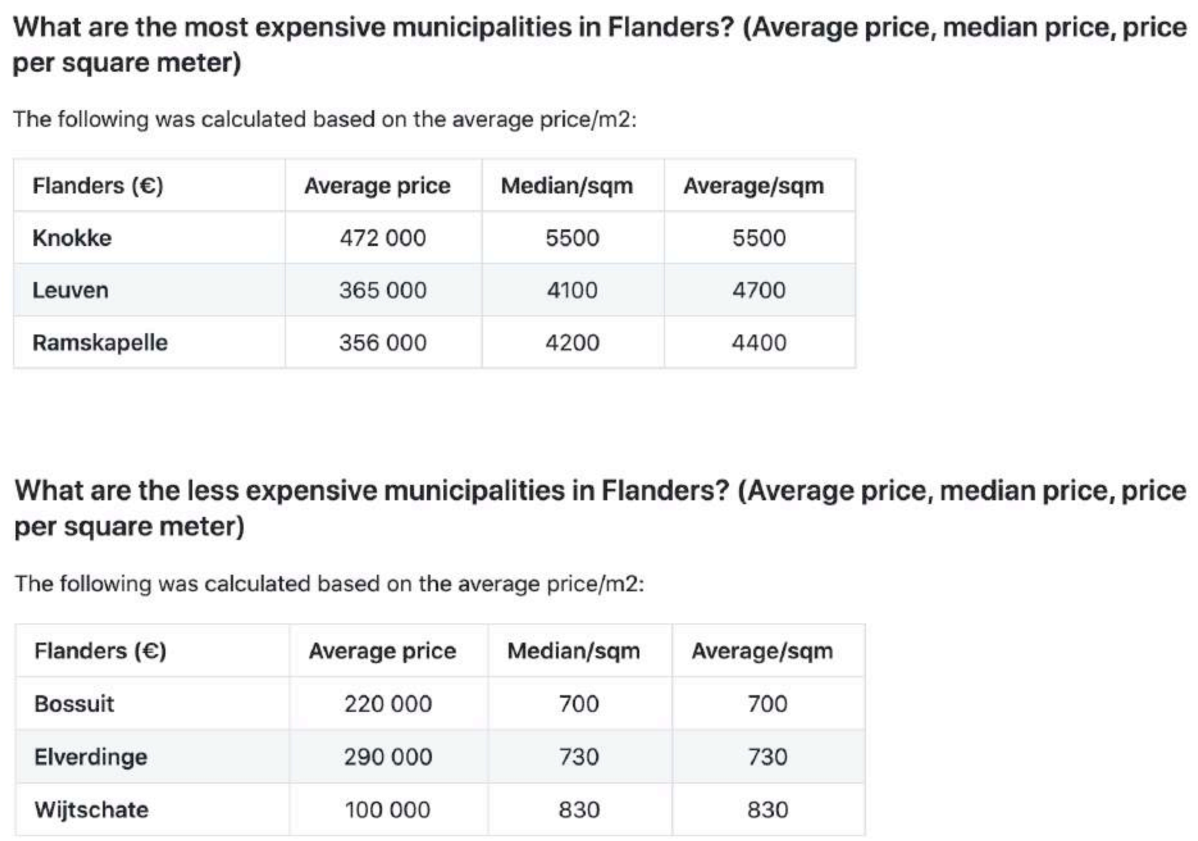
What are the less expensive municipalities in Wallonia? (Average price, median price, price per square meter)

The following was calculated based on the average price/m2:

Belgium (€)	Average price	Median/sqm	Average/sqm
Beauwelz	70 000	350	350
Focant	80 000	390	390
Nollevaux	139 000	420	420

Data Analysis

Challenge's answers



What are the most expensive municipalities in Flanders? (Average price, median price, price per square meter)

The following was calculated based on the average price/m2:

Flanders (€)	Average price	Median/sqm	Average/sqm
Knokke	472 000	5500	5500
Leuven	365 000	4100	4700
Ramskapelle	356 000	4200	4400

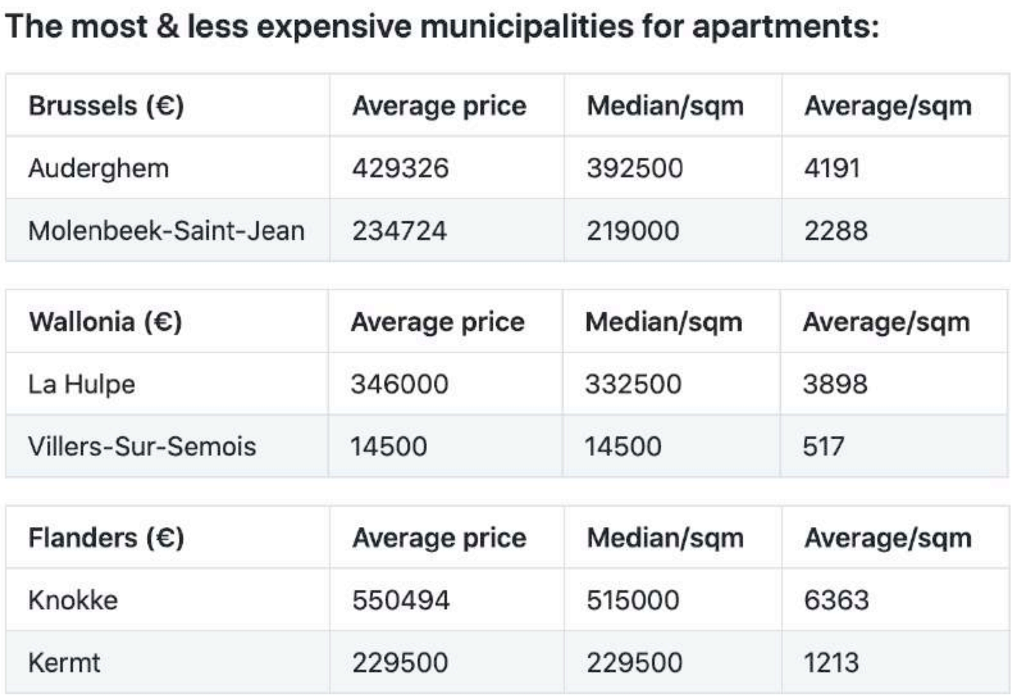
What are the less expensive municipalities in Flanders? (Average price, median price, price per square meter)

The following was calculated based on the average price/m2:

Flanders (€)	Average price	Median/sqm	Average/sqm
Bossuit	220 000	700	700
Elverdinge	290 000	730	730
Wijtschate	100 000	830	830

Data Analysis

Challenge's answers



The most & less expensive municipalities for apartments:

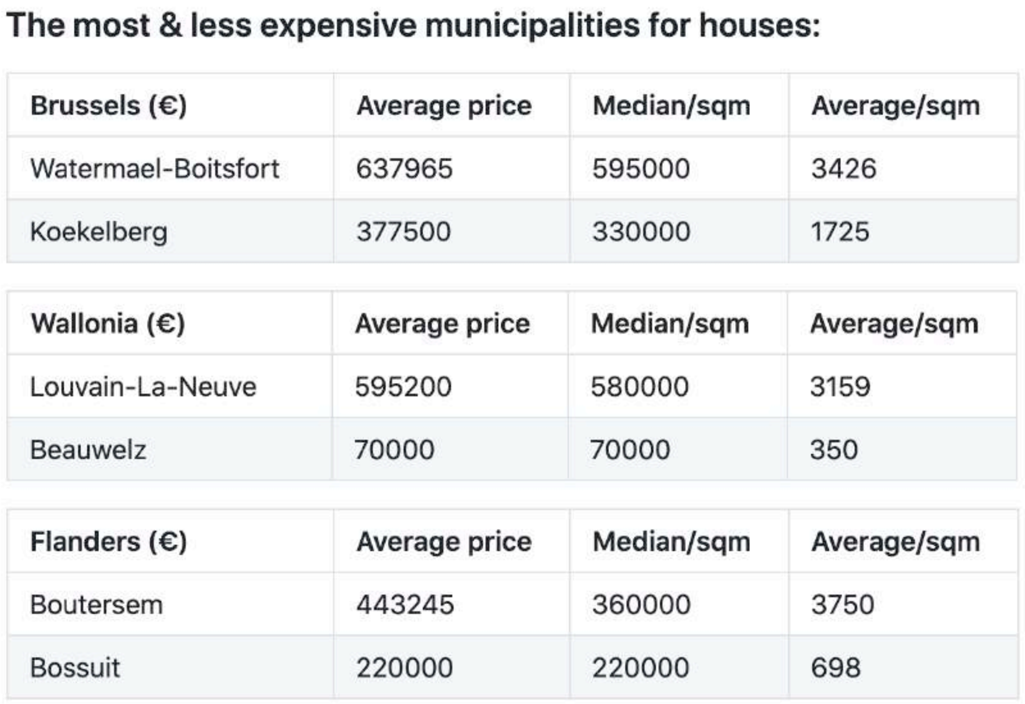
Brussels (€)	Average price	Median/sqm	Average/sqm
Auderghem	429326	392500	4191
Molenbeek-Saint-Jean	234724	219000	2288

Wallonia (€)	Average price	Median/sqm	Average/sqm
La Hulpe	346000	332500	3898
Villers-Sur-Semois	14500	14500	517

Flanders (€)	Average price	Median/sqm	Average/sqm
Knokke	550494	515000	6363
Kermt	229500	229500	1213

Data Analysis

Challenge's answers

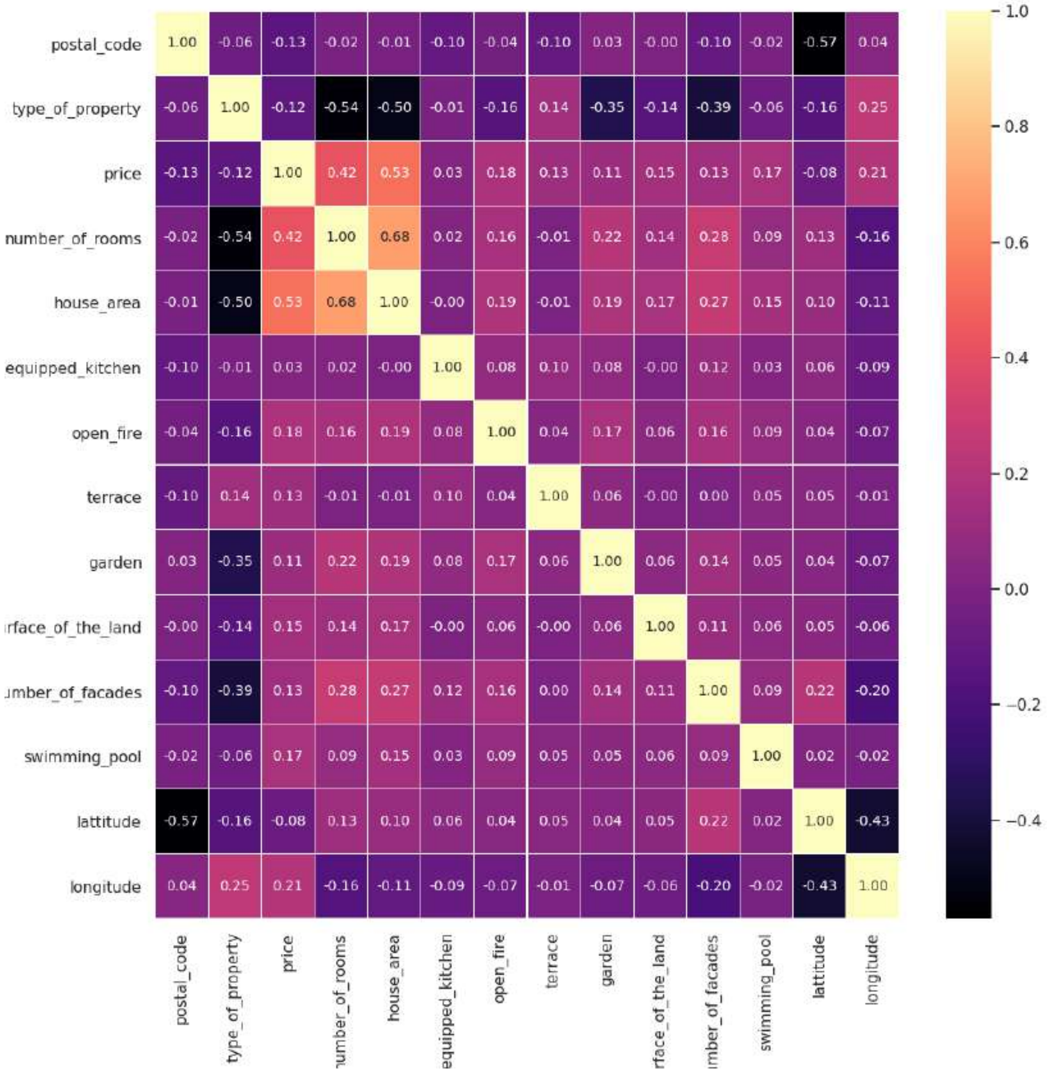


The most & less expensive municipalities for houses:

Brussels (€)	Average price	Median/sqm	Average/sqm
Watermael-Boitsfort	637965	595000	3426
Koekelberg	377500	330000	1725

Wallonia (€)	Average price	Median/sqm	Average/sqm
Louvain-La-Neuve	595200	580000	3159
Beauwelz	70000	70000	350

Flanders (€)	Average price	Median/sqm	Average/sqm
Boutersem	443245	360000	3750
Bossuit	220000	220000	698



	postal_code	type_of_property	price	umber_of_rooms	house_area	equipped_kitchen	open_fire	terrace	garden	rface_of_the_land	mber_of_facades	swimming_pool	lattitude	longitude
postal_code	1.00	-0.06	-0.13	-0.02	-0.01	-0.10	-0.04	-0.10	0.03	-0.00	-0.10	-0.02	-0.57	0.04
type_of_property	-0.06	1.00	-0.12	-0.54	-0.50	-0.01	-0.16	0.14	-0.35	-0.14	-0.39	-0.06	-0.16	0.25
price	-0.13	-0.12	1.00	0.42	0.53	0.03	0.18	0.13	0.11	0.15	0.13	0.17	-0.08	0.21
number_of_rooms	-0.02	-0.54	0.42	1.00	0.68	0.02	0.16	-0.01	0.22	0.14	0.28	0.09	0.13	-0.16
house_area	-0.01	-0.50	0.53	0.68	1.00	-0.00	0.19	-0.01	0.19	0.17	0.27	0.15	0.10	-0.11
equipped_kitchen	-0.10	-0.01	0.03	0.02	-0.00	1.00	0.08	0.10	0.08	-0.00	0.12	0.03	0.06	-0.09
open_fire	-0.04	-0.16	0.18	0.16	0.19	0.08	1.00	0.04	0.17	0.06	0.16	0.09	0.04	-0.07
terrace	-0.10	0.14	0.13	-0.01	-0.01	0.10	0.04	1.00	0.06	-0.00	0.00	0.05	0.05	-0.01
garden	0.03	-0.35	0.11	0.22	0.19	0.08	0.17	0.06	1.00	0.06	0.14	0.05	0.04	-0.07
rface_of_the_land	-0.00	-0.14	0.15	0.14	0.17	-0.00	0.06	-0.00	0.06	1.00	0.11	0.06	0.05	-0.06
umber_of_facades	-0.10	-0.39	0.13	0.28	0.27	0.12	0.16	0.00	0.14	0.11	1.00	0.09	0.22	-0.20
swimming_pool	-0.02	-0.06	0.17	0.09	0.15	0.03	0.09	0.05	0.05	0.06	0.09	1.00	0.02	-0.02
lattitude	-0.57	-0.16	-0.08	0.13	0.10	0.06	0.04	0.05	0.04	0.05	0.22	0.02	1.00	-0.43
longitude	0.04	0.25	0.21	-0.16	-0.11	-0.09	-0.07	-0.01	-0.07	-0.06	-0.20	-0.02	-0.43	1.00

Data Interpretation

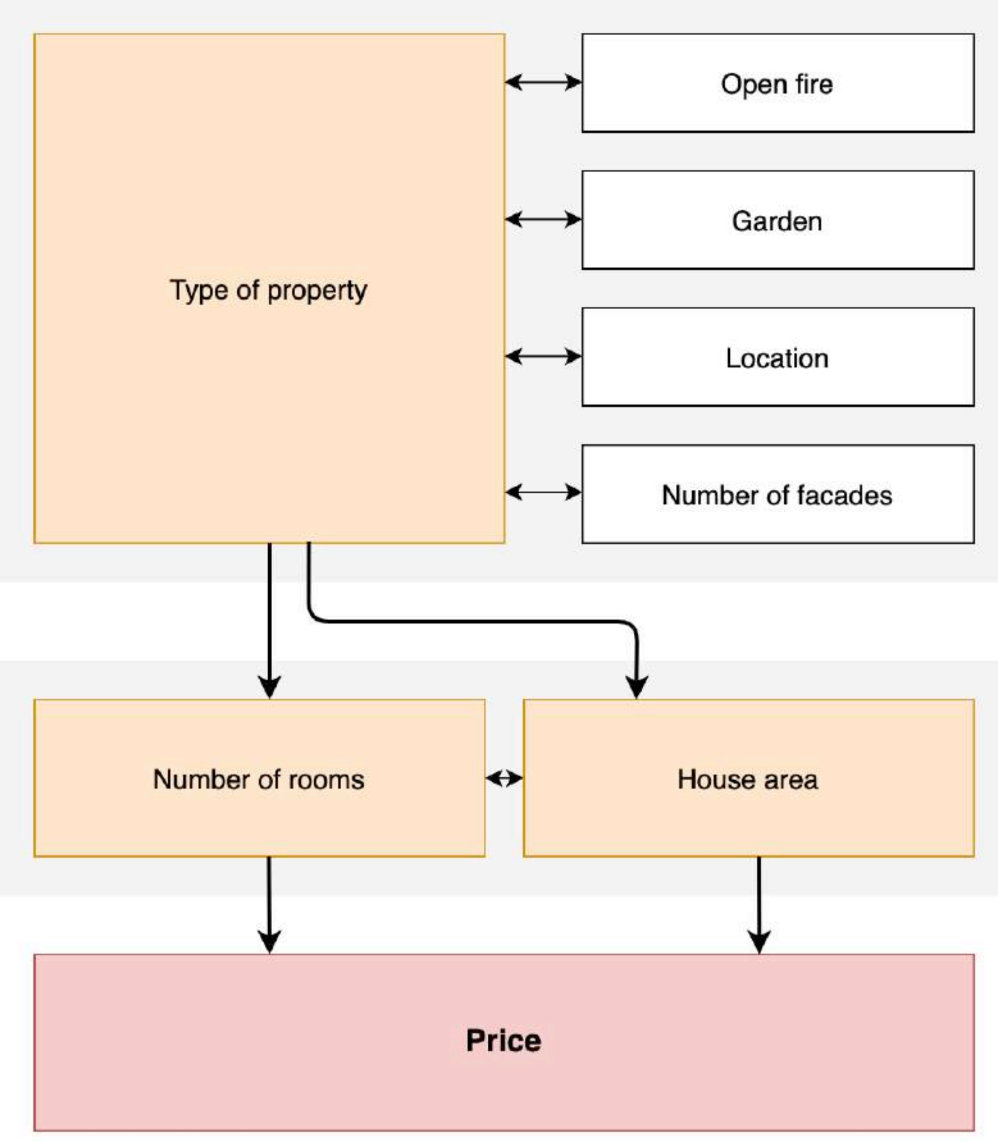
Correlation Heatmap

Observations:

1.The **Price** is mainly correlated with the *Number of rooms* and the *House area*.

2.The **Number of rooms** and *House area* seems mainly correlated with each other.

1.The **Type of property** is the variables which has the most correlation with other variables.



Conclusions:

- The **Open fire**, the **Garden**, the **Location** of the house (municipality) and its **Number of facades** determine the **Type of property**. Which influence greatly on the **Number of rooms** and the **House area**: An apartement will have less space and less rooms than a house.

- **Number of rooms** and **house area** are two variables based on the size of the property. And they are the main influence on the **Price**.

- A larger house/apartment is more expensive than a smaller house/apartment.

Debriefing

1. The first difficulty was to **find a method of collaborative work** adapted to our desires.

   ➢ We chose to work via google colab pages and to gather valid cells on this github repo.

2. The second "big" difficulty was to **quickly learn how to use tools** such as matplotlib or seaborn.

   ➢ This was solved by working on our own, and sharing our work with each other.

3. Another challenge has been the **fair distribution of work**.

   ➢ We do not all have the same ease of understanding in statistics and programming... So everyone was doing their best, and then we merged our result.

4. We encountered a small problem about the file name due to the fact that we don't work on the same environment (Win/Ubuntu...).

Links

Github Repository

- https://github.com/kaiyungtan/challenge-data-analysis

Github Page

- https://kaiyungtan.github.io/challenge-data-analysis/