Housing Price Analysis and Prediction

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.

Data Collecting

- ▶ A dataset of **50k+** real estate's observations, Collect from Kaggle. Which is published as Belgium real estate industry

- ▶ 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.

Data Cleaning

Identifying the needs:

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

- Prices, postal code and per sqft price

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:

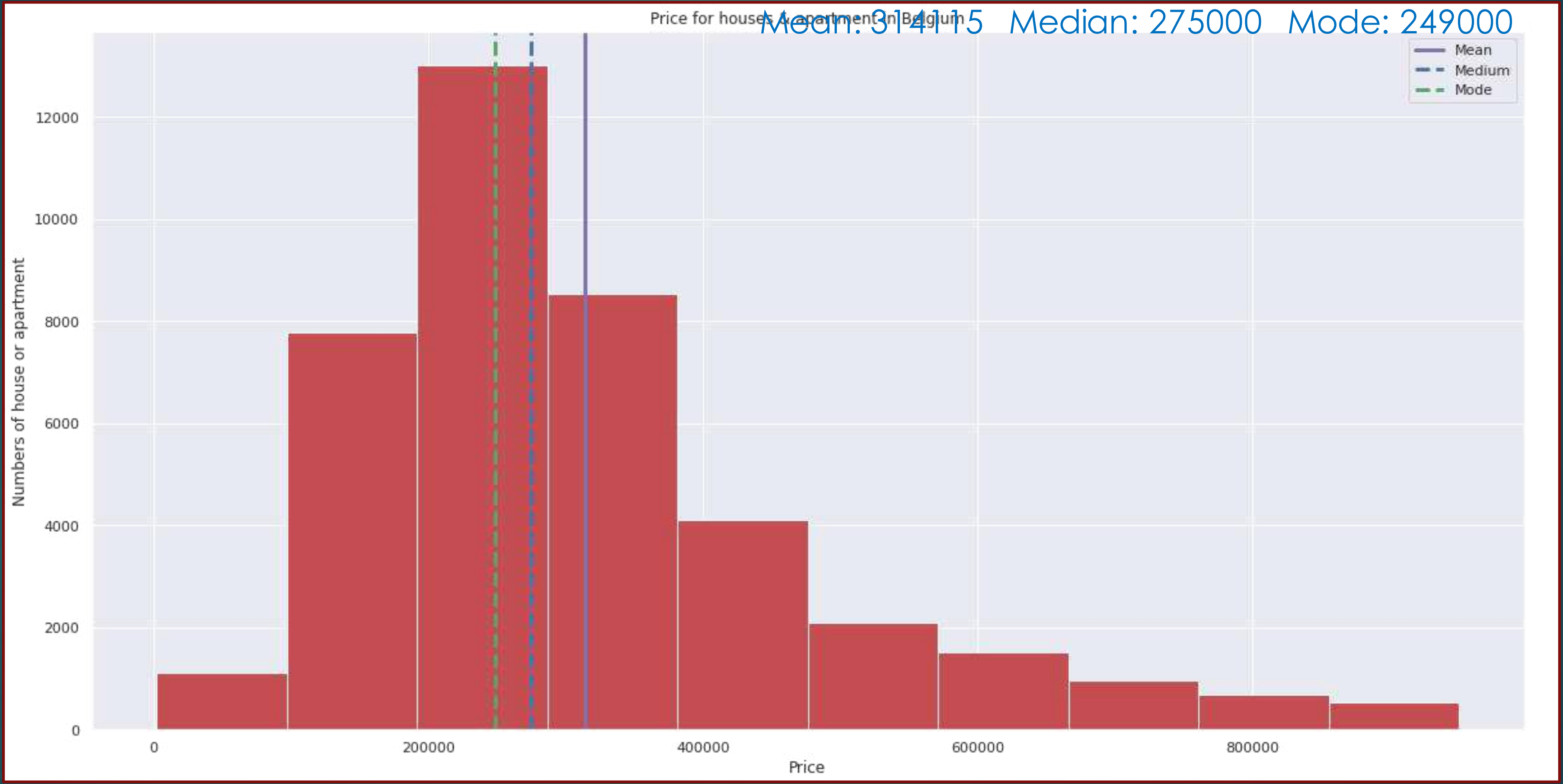
- Dropping "terrace_area" column
  - It has more than 30% of None.
- Dropping "garden_area" column
  - It has more than 50% of None.
- Dropping "subtype" column
  - Lots of property subtype. Some with less than 100 entries, in a dataset of 50.000.
  - This column was not relevant.
- Removing the "Apartment 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

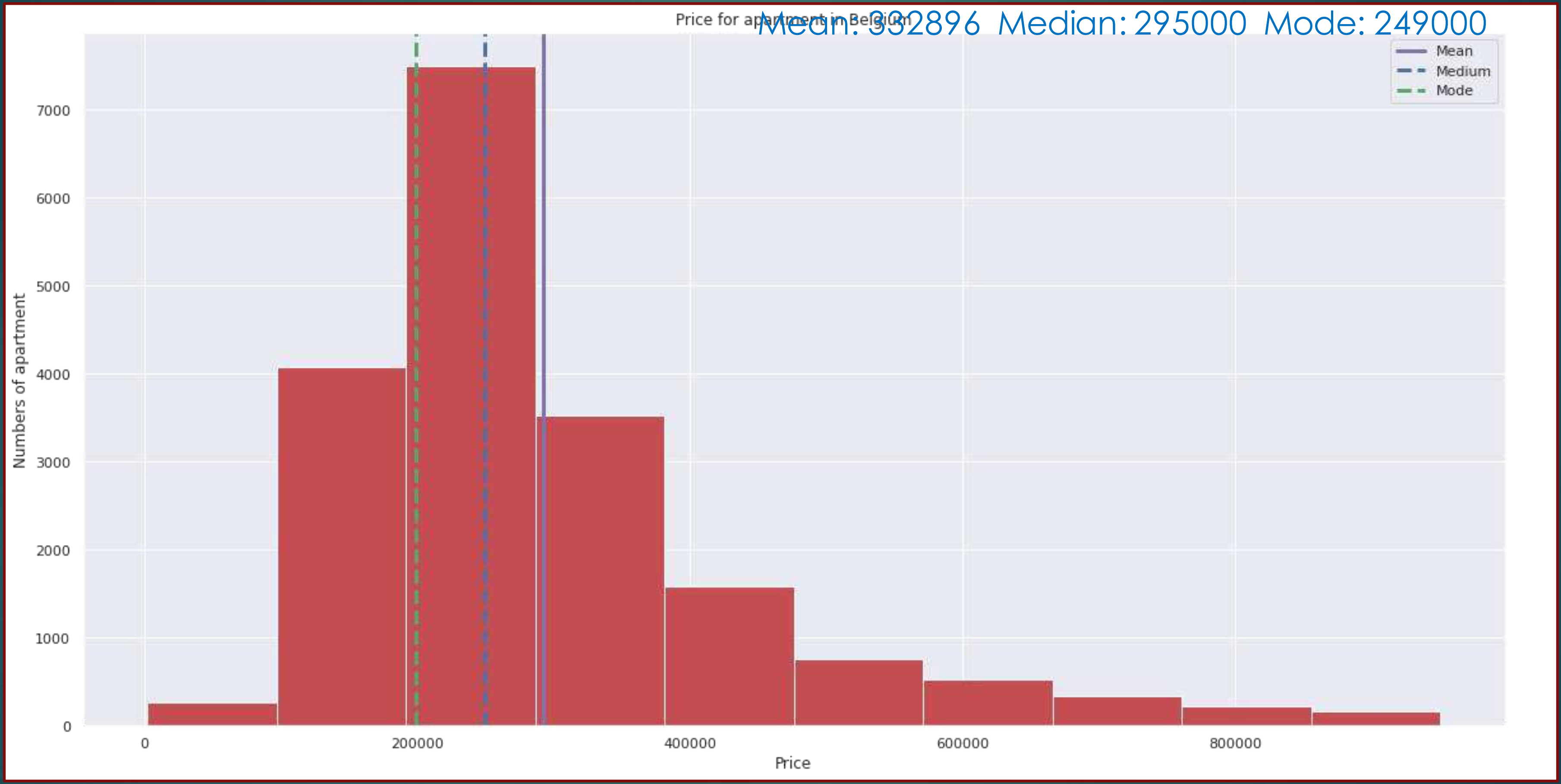


Mean: 314115   Median: 275000   Mode: 249000

Data Visualisation

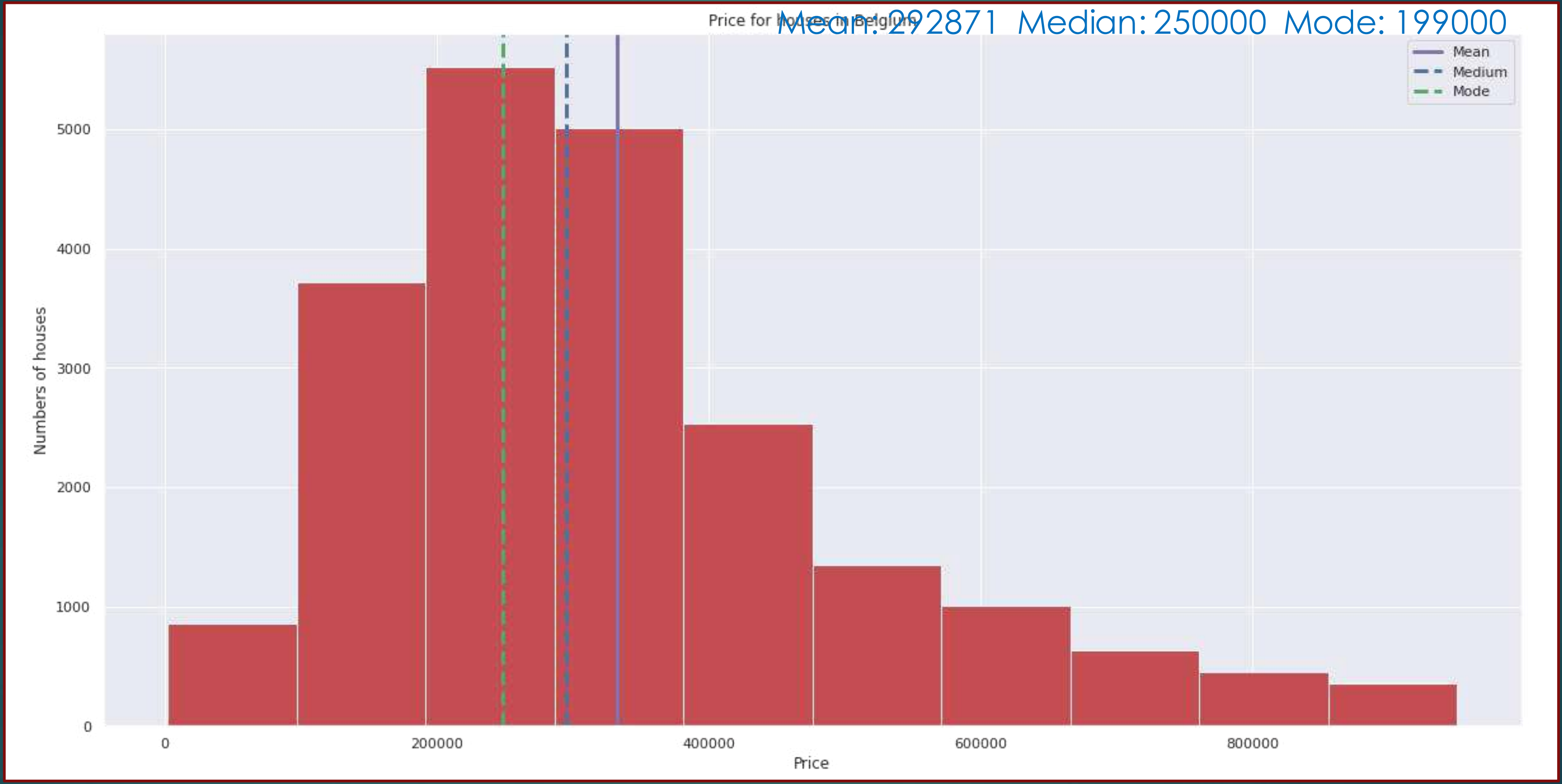
Our target: The Price

Mean: 332896 Median: 295000 Mode: 249000



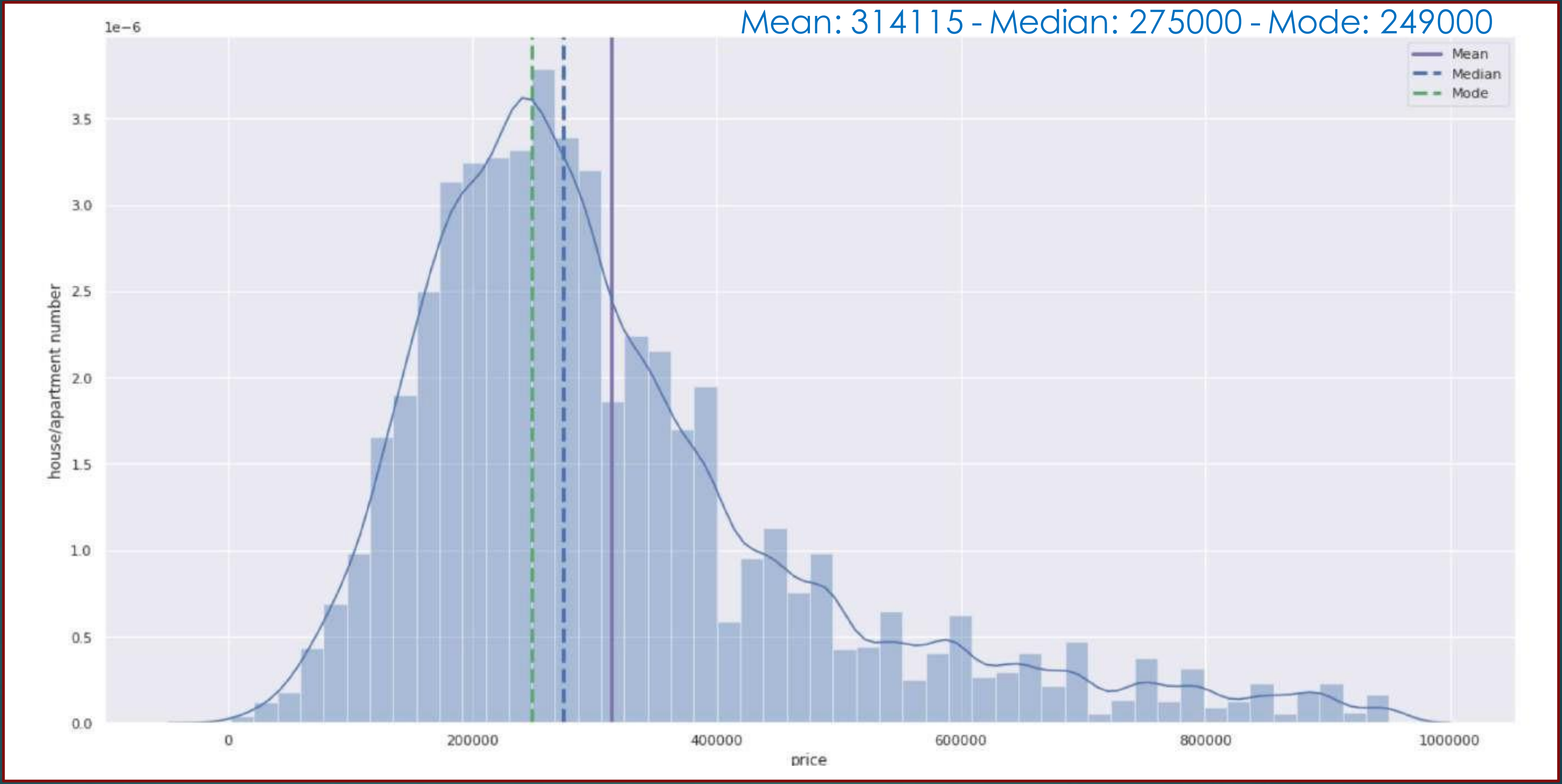
Data Visualisation

Our target: The Price

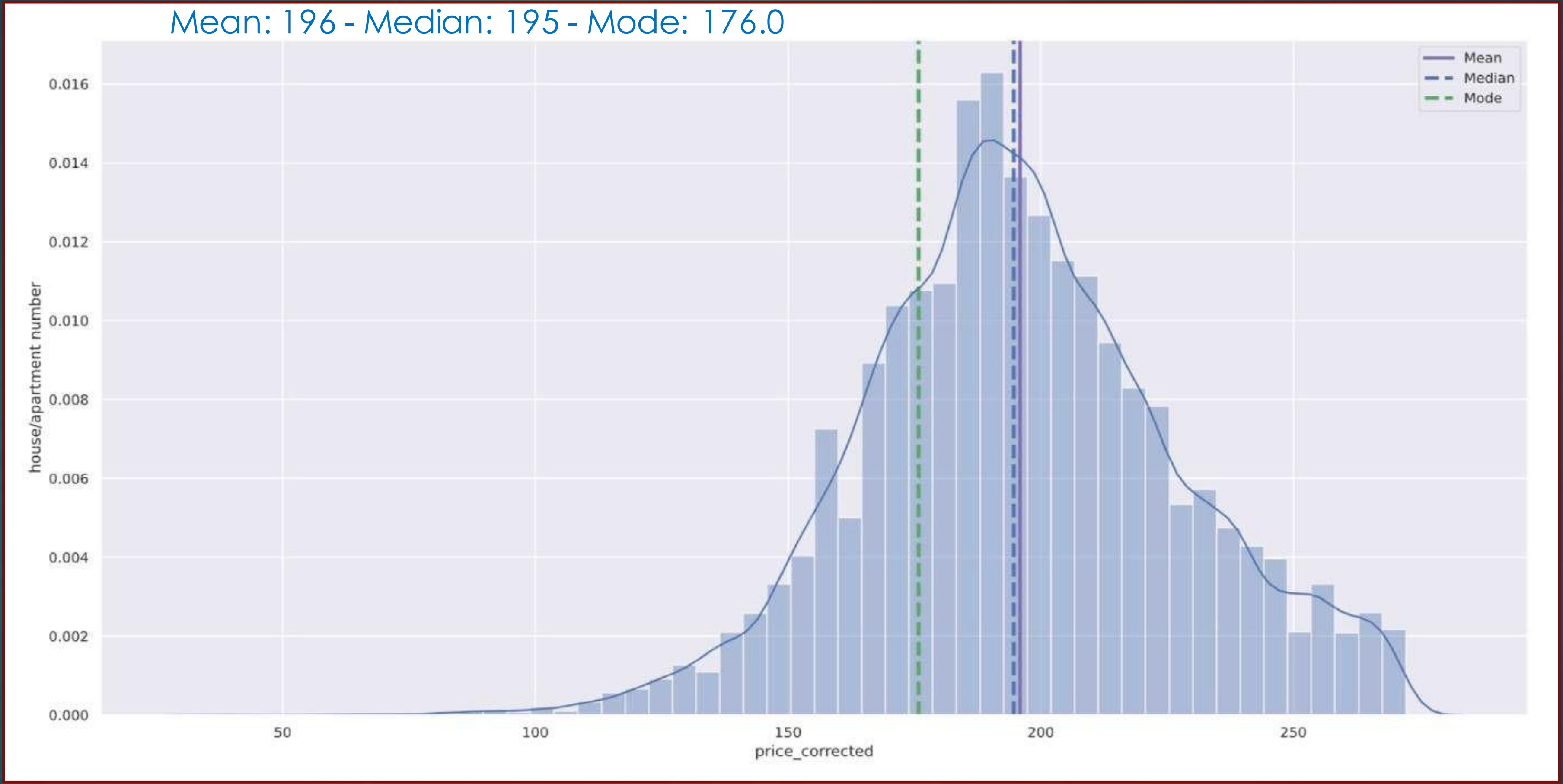


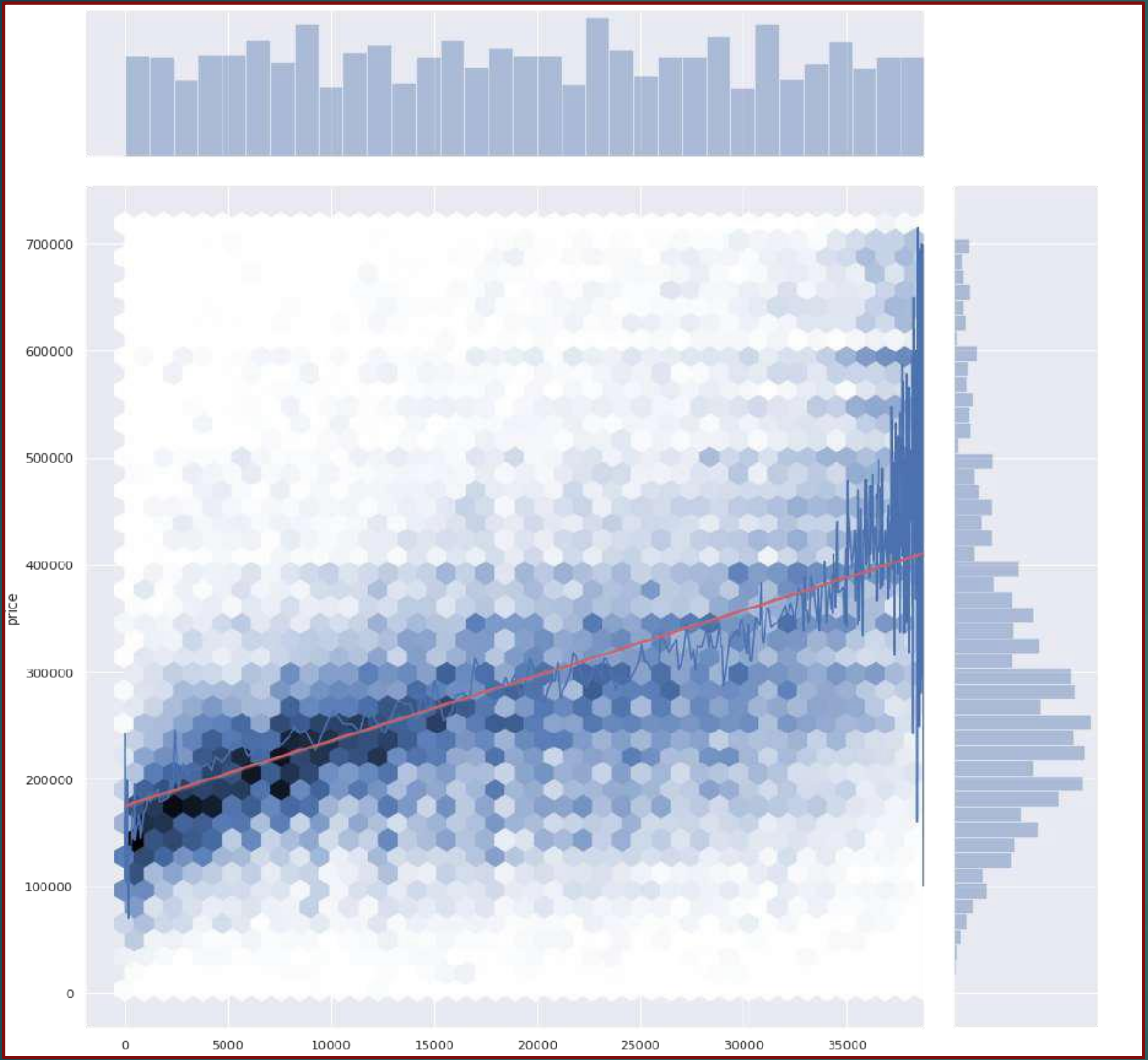
Mean: 292871 Median: 250000 Mode: 199000

Data Visualisation



Data Visualisation

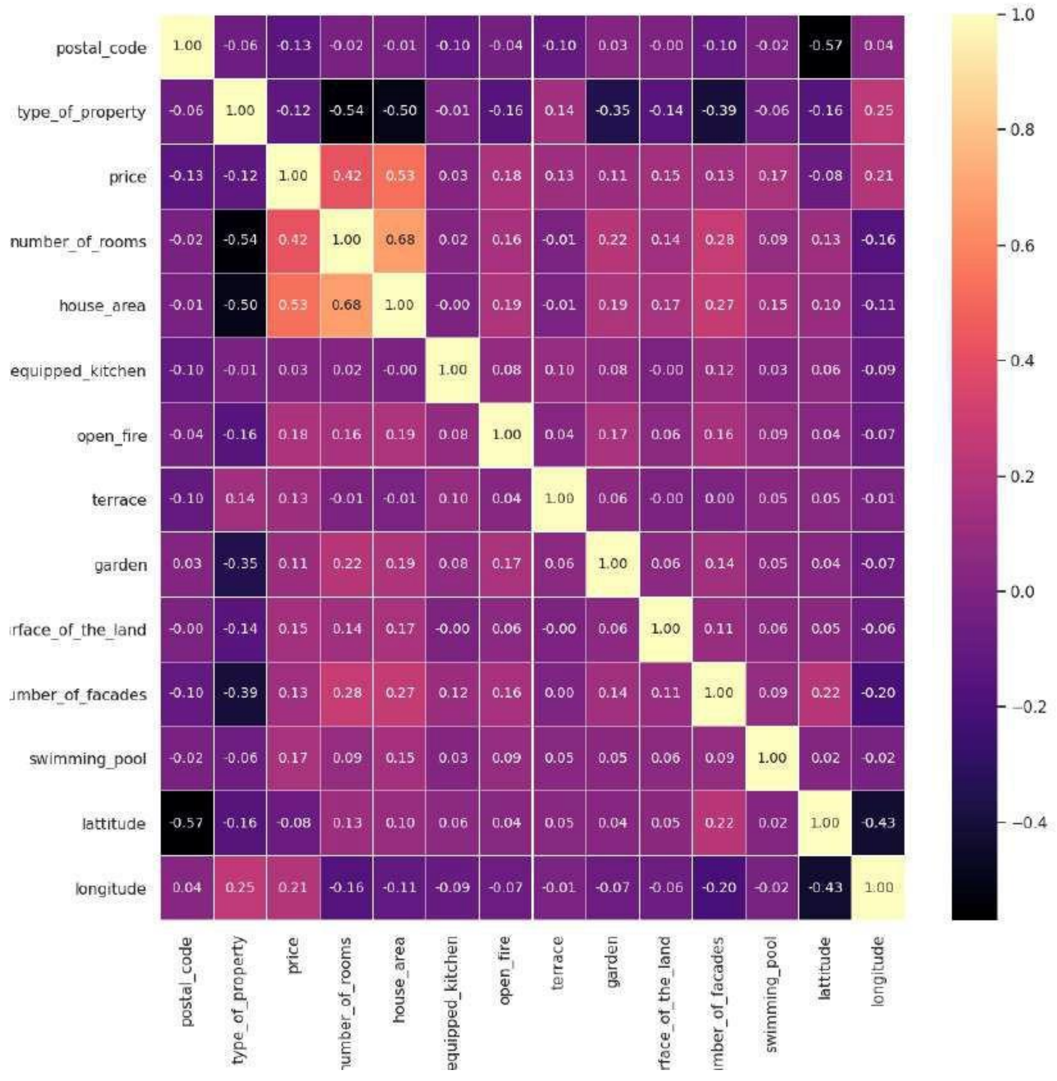




Data Visualisation

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.



	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.

Correlation does not imply causation

Links

Github Repository

- https://https://github.com/mdimran1/Real-Estate

Linkedin :

- https://www.linkedin.com/in/imran-pro/