**TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI**

**TRƯỜNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG**

**------------\*------------**

****

**CAPSTONE PROJECT REPORT**

**SUBJECT:**

**Housing price prediction**

|  |  |
| --- | --- |
| **Giáo viên hướng dẫn:** | Trần Việt Trung |
| **Sinh viên thực hiện:** | Hoàng Thanh Lâm  Phạm Thành Biên  Phạm Hải Đăng  Lê Đức Huy  Hoàng Quang Mạnh |

**TABLE OF CONTENTS**

[**I.**](#_heading=h.30j0zll) **Problem Statement** 2

[**II.**](#_heading=h.1fob9te) **Data Science methodology** 3

[**1.**](#_heading=h.3znysh7) **Data Crawling** 3

[**2.**](#_heading=h.2et92p0) **Data Preprocessing** 6

[**3.**](#_heading=h.tyjcwt) **Data Visualization** 11

[**4.**](#_heading=h.3dy6vkm) **Model Training** 15

[**III.**](#_heading=h.1t3h5sf) **Result** 16

[**1.**](#_heading=h.4d34og8) **Final Result** 16

[**2.**](#_heading=h.2s8eyo1) **Difficulties:** 16

[**3.**](#_heading=h.17dp8vu) **Future Improvement** 17

[**4. Conclusions** 17](#_heading=h.lnxbz9)

**Foreword**

In the dynamic realm of data science, where algorithms and analytics converge to uncover patterns and insights, the topic of housing price prediction stands as a testament to the transformative power of data-driven decision-making. As we navigate an era increasingly defined by technology, the ability to forecast housing prices has become a pivotal aspect of real estate, economics, and urban planning.

The Housing Price Prediction in Data Science subject represents an exciting exploration into the intersection of advanced statistical modeling, machine learning, and domain expertise. In this era of big data, where information flows in torrents, the ability to distill meaningful predictions from the data deluge has become paramount. Predicting housing prices extends beyond the confines of academic curiosity—it has tangible implications for homeowners, investors, policymakers, and the broader community.

This subject delves into a multifaceted landscape, addressing challenges that range from feature engineering and model selection to ethical considerations in algorithmic decision-making. Through hands-on exploration and theoretical understanding, students embarking on this journey will develop a keen sense of how to navigate the complexities inherent in housing market dynamics.

|  |
| --- |
| Our Team |
| Hoàng Thanh Lâm  Lê Đức Huy  Phạm Hải Đăng  Phạm Thành Biên  Hoàng Quang Mạnh |

# **Problem Statement**

• People looking to buy a new home tend to be more conservative with their budgets and market strategies.  
• This project aims to analyse various parameters like average income, average area etc. and predict the house price accordingly.  
• This application will help customers to invest in an estate without approaching an agent  
• To provide a better and fast way of performing operations.  
• To provide proper house price to the customers.  
• To eliminate need of real estate agent to gain information regarding house prices.  
• To provide best price to user without getting cheated.  
• To enable user to search home as per the budget.  
• The aim is to predict the efficient house pricing for real estate customers with respect to their budgets and priorities. By analyzing previous market trends and price ranges, and also upcoming developments future prices will be predicted.  
• House prices increase every year, so there is a need for a system to predict house prices in the future.  
• House price prediction can help the developer determine the selling price of a  
house and can help the customer to arrange the right time to purchase a house.  
• We use linear regression algorithm in machine learning for predicting the house price trends.

# **Data Science methodology**

## **Data Crawling**

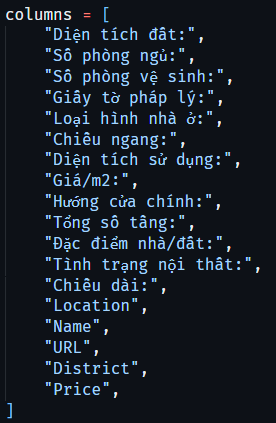
Our team using webdriver from selenium to doing crawling task

First, we find “https//www.nhatot.com” is good web about housing price information and it contain enough information for us to crawl. In this website, we calculate that maximum of a web in each city by visiting last page of that and agree only information in Ha Noi is enough and data also good for learning purpose. So come to final, we use Ha Noi as default city and nhatot is default url and maximum page of hanoi have is 500

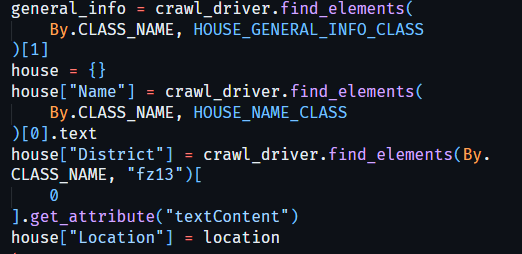


Afraid of losing data when we crawl because of device so when we have information about 5 page ( each page have 20 housing information ) we write into file and save.

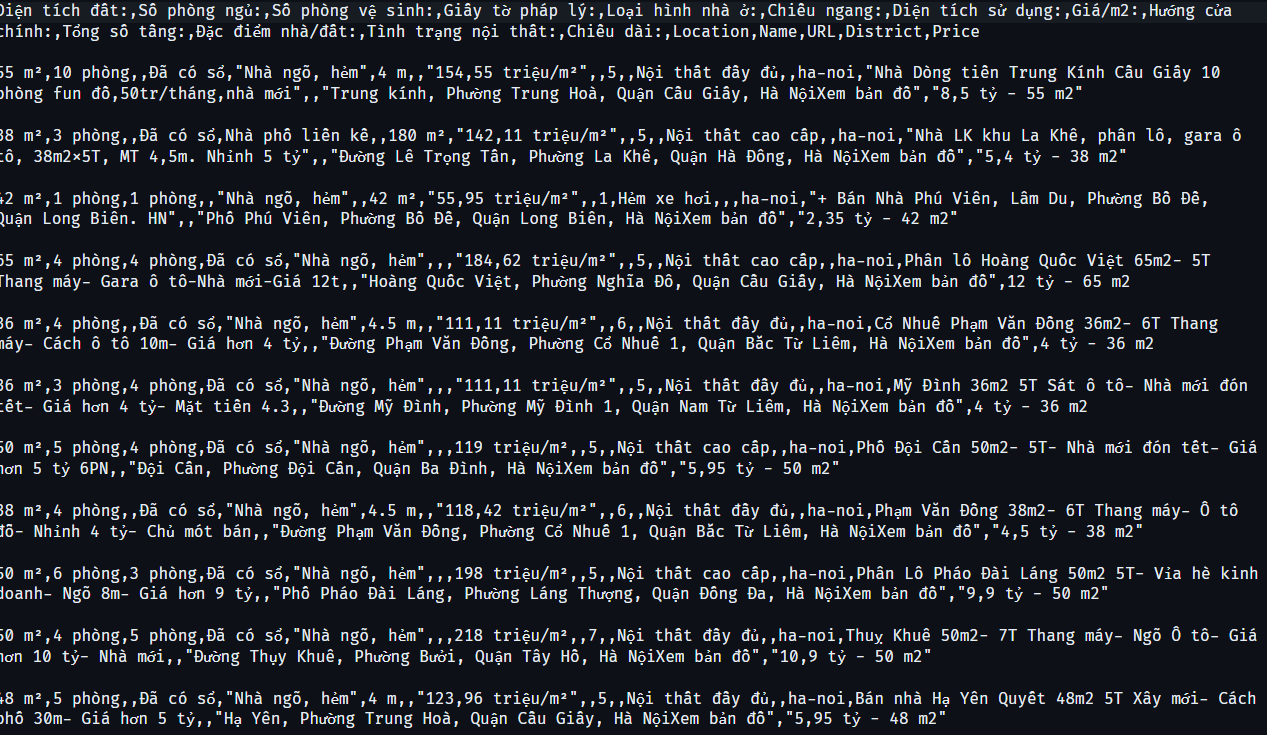
After consider all appropiate field we decided to take only some data in this page



Webdriver is automatic tools to crawl, so we run loop to open each page of website, get data by function from webdriver that get contetnt by some class name or id element in that web.



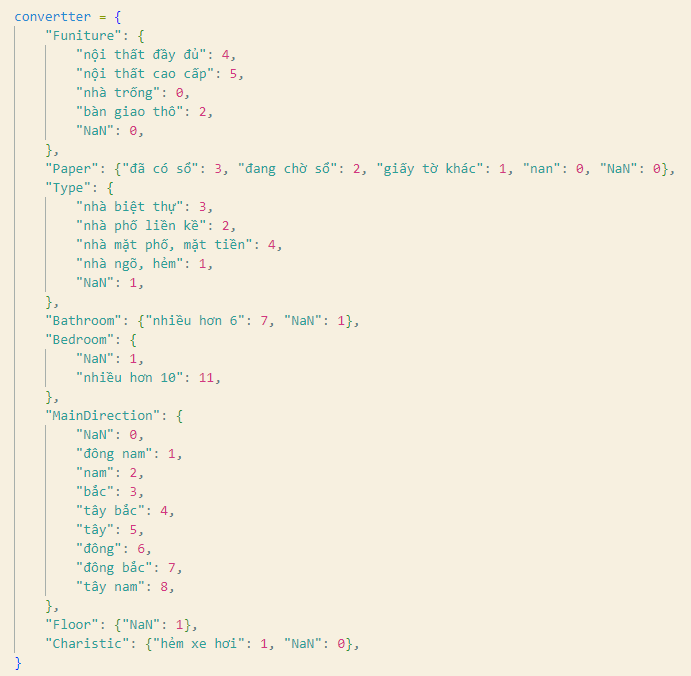
After long time we finally have data that we crawl



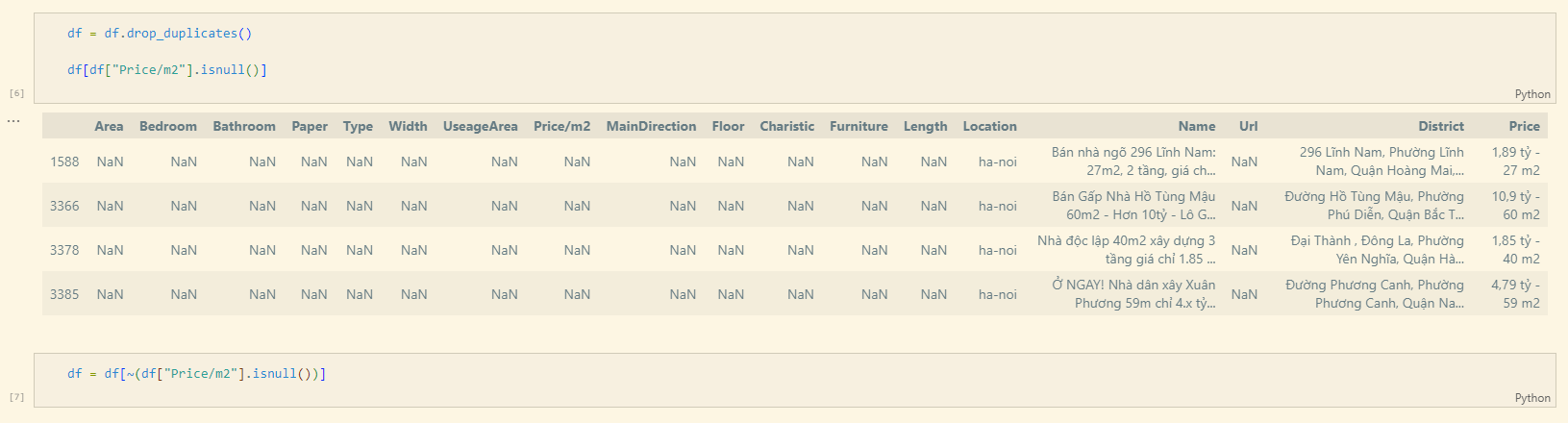
## **Data Preprocessing**

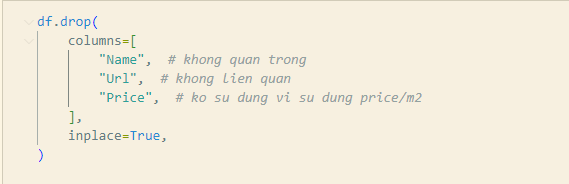
After the completion of the data crawling stage, the team proceeds to preprocess the data using the following steps [2]:

\*Data normalization:

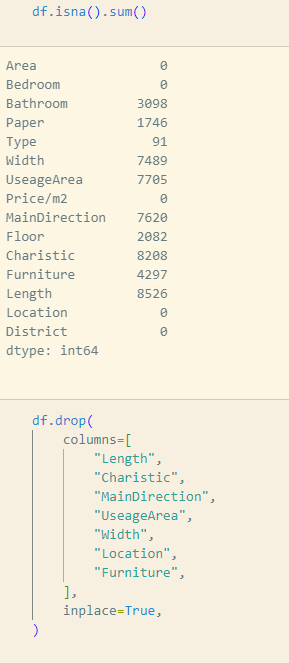


\*Remove all duplicate , and null variable at “Price/m2”:

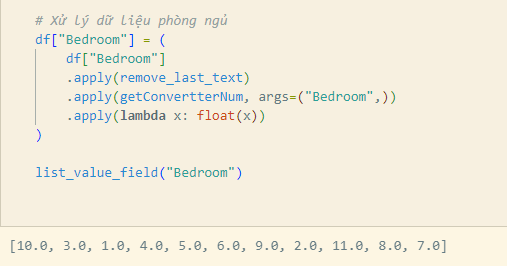


\*Remove some field are not important, irrelevant:  


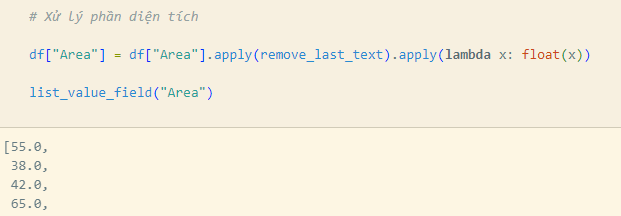
\*Remove columns have too much null, lost information:

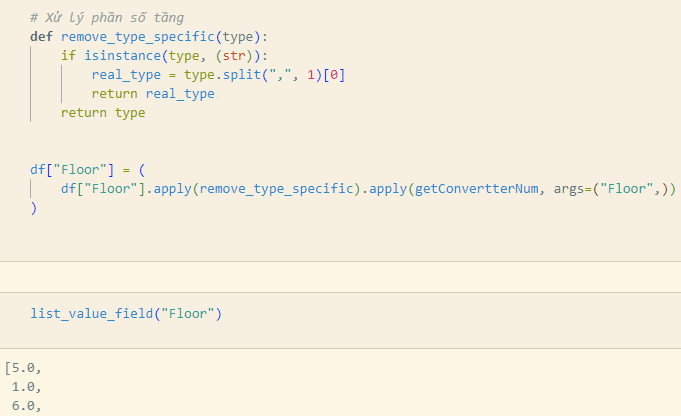


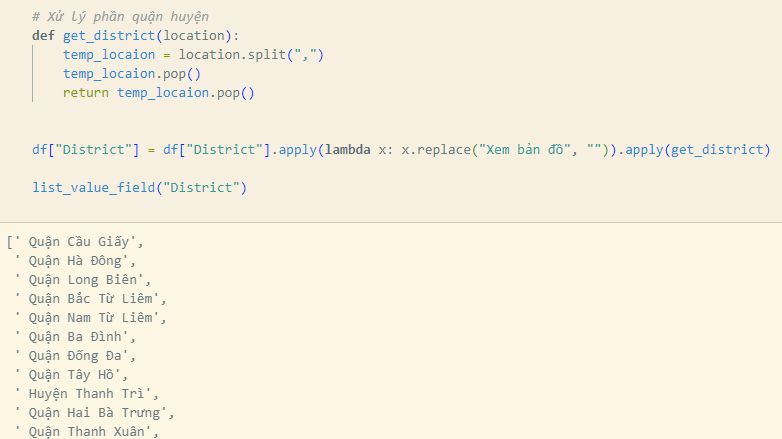
\*Process data : convert to number:







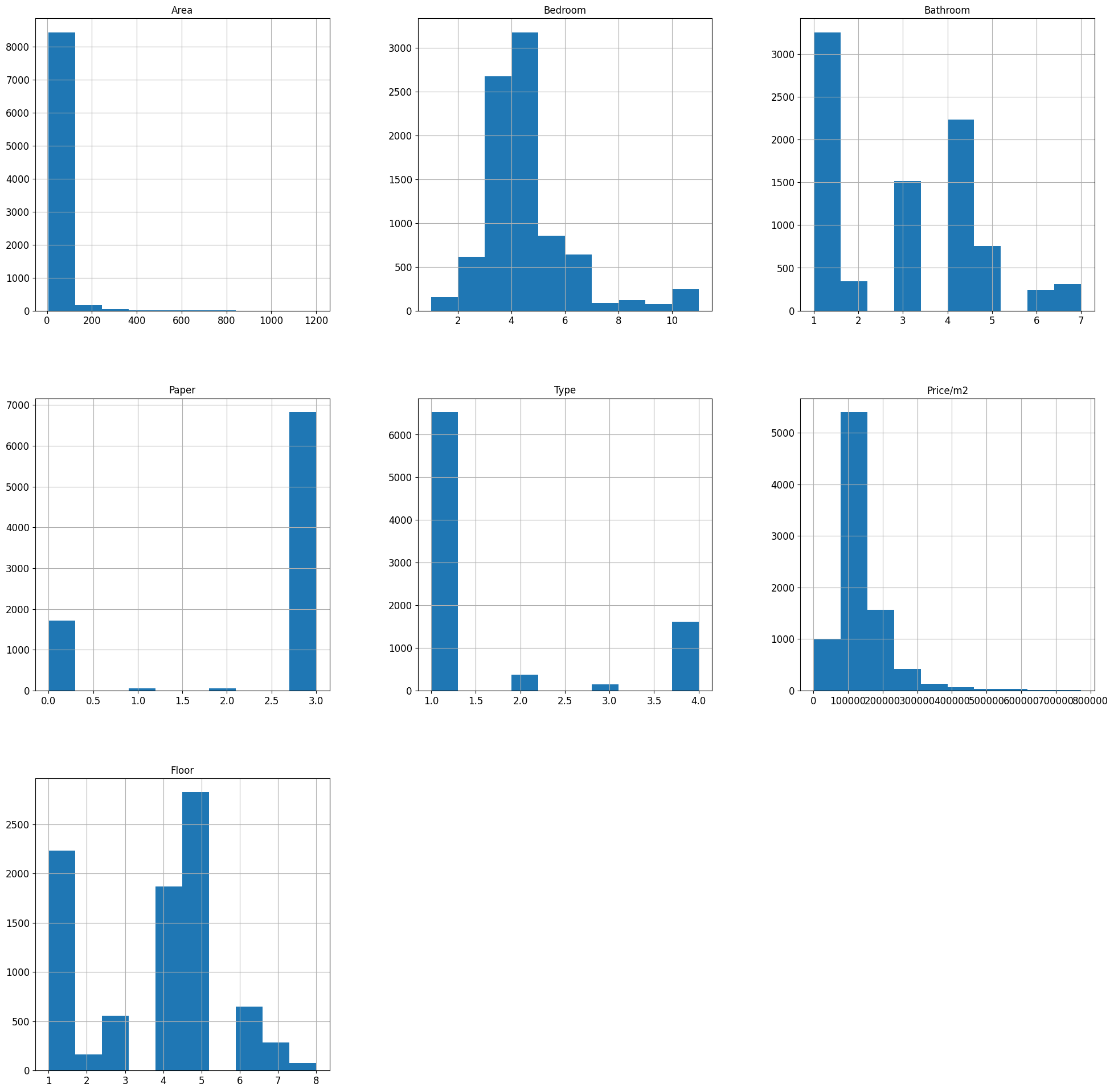




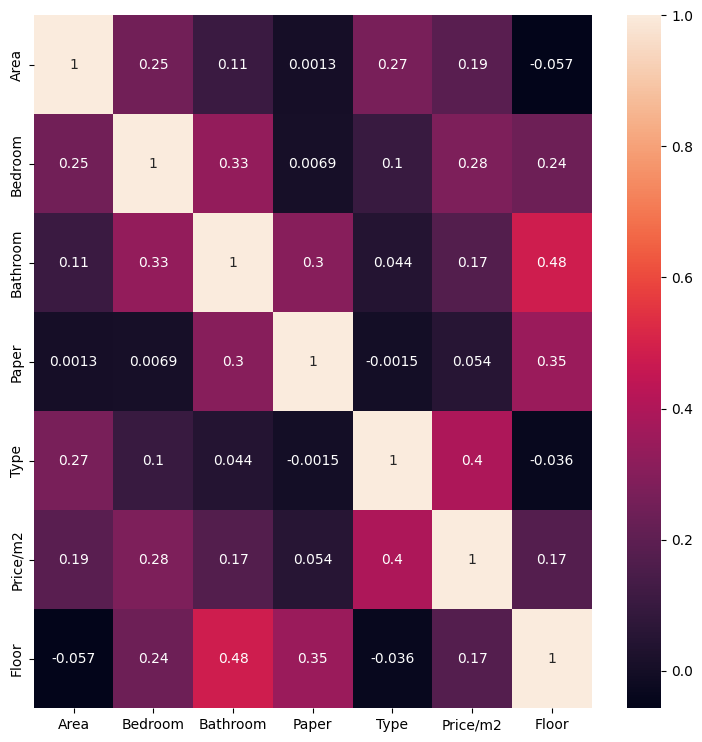
## **Data Visualization**

Our team visualizes the data after it has been cleaned and normalized in order to provide a first assessment of the obtained data set.

1. **The histogram is an accurate representation of the distribution of numerical data**

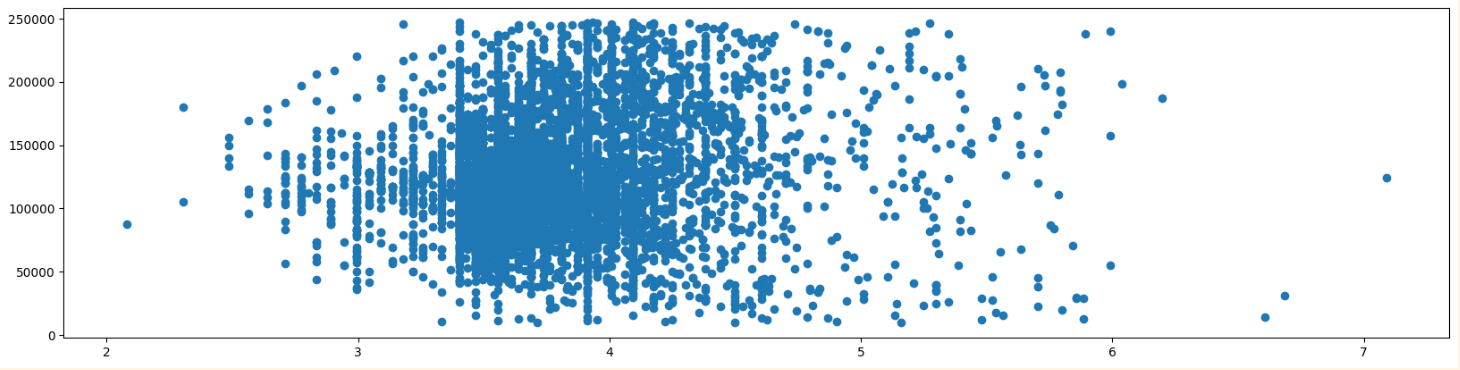


* Data is most popular with Area < 500m2 , from 2-> 5 bedroom, 1,2 bathroom, already have certificates of land use rights or pending, on the alley or on the street
* The data is skewed.

1. **The heat map provide an intuitive way to explore and understand the patterns and relationships within a dataset**

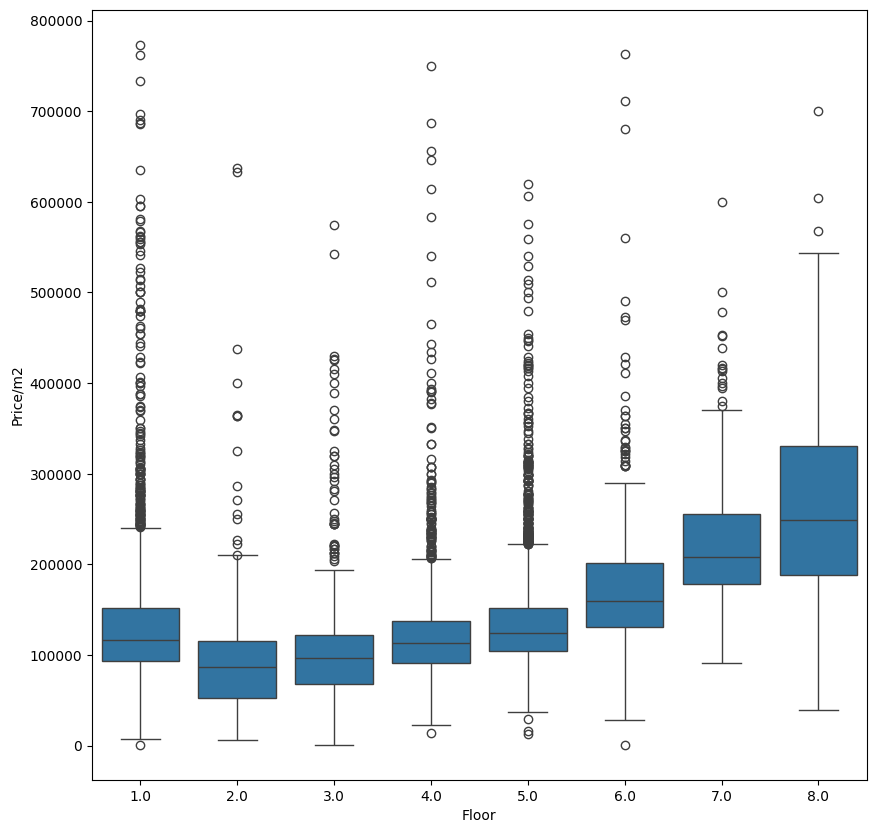
* Positive correlations:
  + Price/m2: Has strong positive correlations with Bedroom, Bathroom, Type.
  + Area: Has strong positive correlations with Bedroom, Bathroom, Type.
  + Floor: Has a strong negative correlation with Bathroom.
* No significant correlations:
  + Paper has no significant correlations with any other variables (except Floor, Bathroom)

1. **The scatter plot chart displays individual data points on a two-dimensional graph :**

****

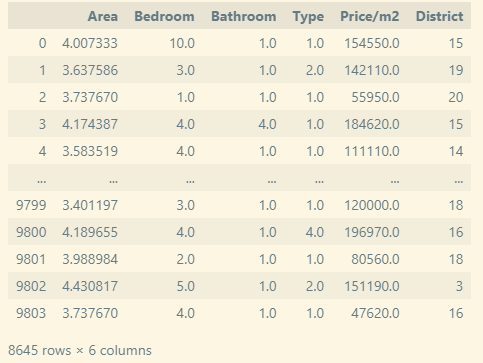
* Outliers, or data points that deviate significantly from the general pattern, can be easily identified in a scatter plot. These points often stand out from the main distribution and may indicate errors, anomalies.
* One way to remove outliers.
* The points on the plot form a clear trend.

1. **Box-plot is present the relation between Price and Floor:**



* Analysis:
  + The distribution of property prices is uneven, with many apartments having prices higher than the median.
  + There are a few luxury apartments with prices significantly higher than the rest of the market.
  + Hard to predict price from floor information.

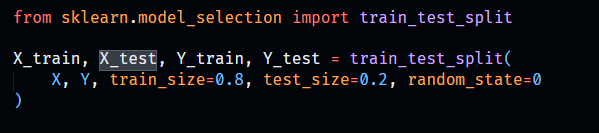
1. **Data after cleaning:**

****

## **Model Training**

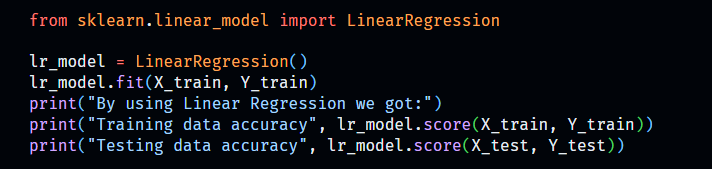
* 1. **Split dataset:**

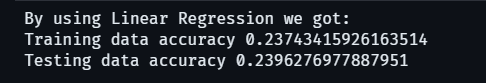
We split the dataset into train set and test set (with ratio 8:2) and also keep the ratio between categories in training set and test set similar to the original ratio between them



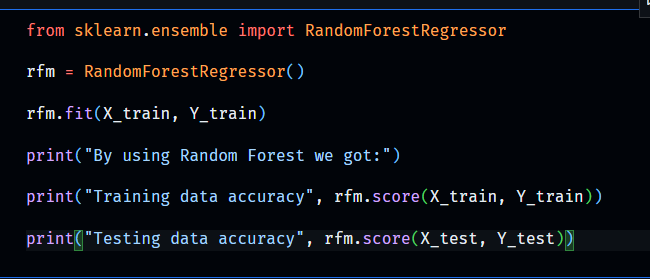
Train set include 6916 records while this number in test set is 1729.

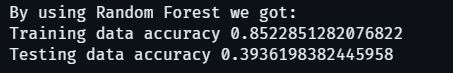
**4.2. Apply LinearRegression:**





**4.3 Apply RandomForestRegression**





1. **Conclusion**

The reason why our model train’s result is quite bad because we think that our data crawling in web’s that is some marketing’s price and not real and we also collect many type of house like apartment, basement, …

# **Result**

## **Final Result**

**a)Training:**

\*Accuracy: 0.85

**b)Test:**

\*Accuracy:0.39

## **Difficulties:**

- The collected data is not large but it is time-consuming to collect. Difficult to continue if there is an error while crawling.

- The model accuracy is not bad but also not good. Especially , art-and-entertainment category always have a low precision in test.

## **Future Improvement**

## **-** In the future, for improvement, we are going to collect more data to increase the accuracy and detect if there are any noises in our dataset

## - In addition, we will try our best to handle the multi-class problem . Because some website we have had crawled may belong to more than 1 category and this can be the reason that make our model not work so well

-We will also try another model to make more improvement .

## 

## **Conclusions**

The report summarized the process our team did, with data science methodology. From the steps of problem statement, data collection, cleaning, model training as well as model evaluation with test set.

In this project, we tried our best to collect data, although we had to deal with a lot of faulty websites and websites that preventing us from crawling. We also learned RandomForestRegression -an easy-to-implement algorithm.

Finally, completing this project in accordance with the data science methodology gave the team a true understanding of the process. It is not a process of downloading an existing dataset and then training to refine the model, but a practical process.