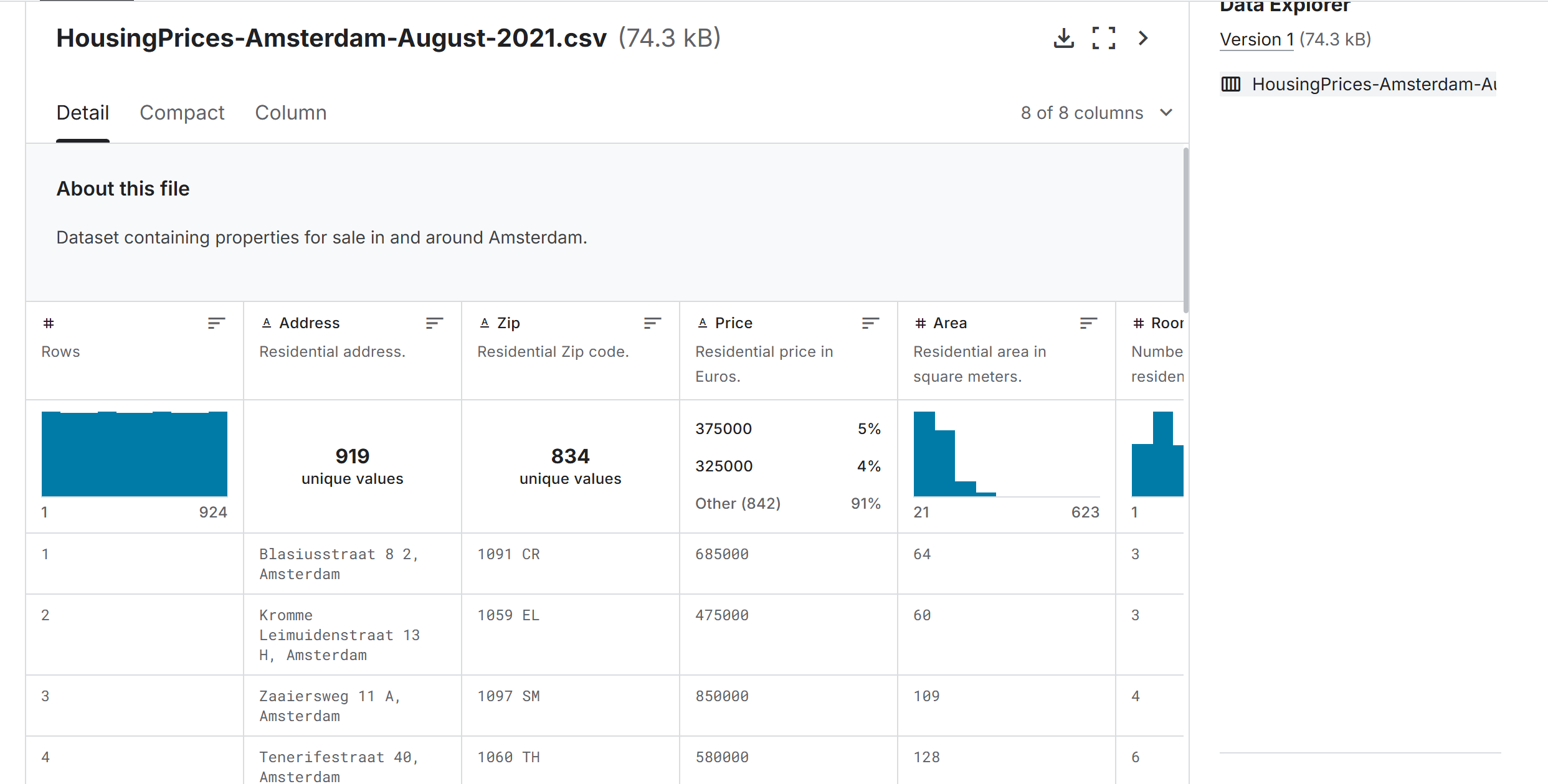
Task2 : **Predicting House Prices**

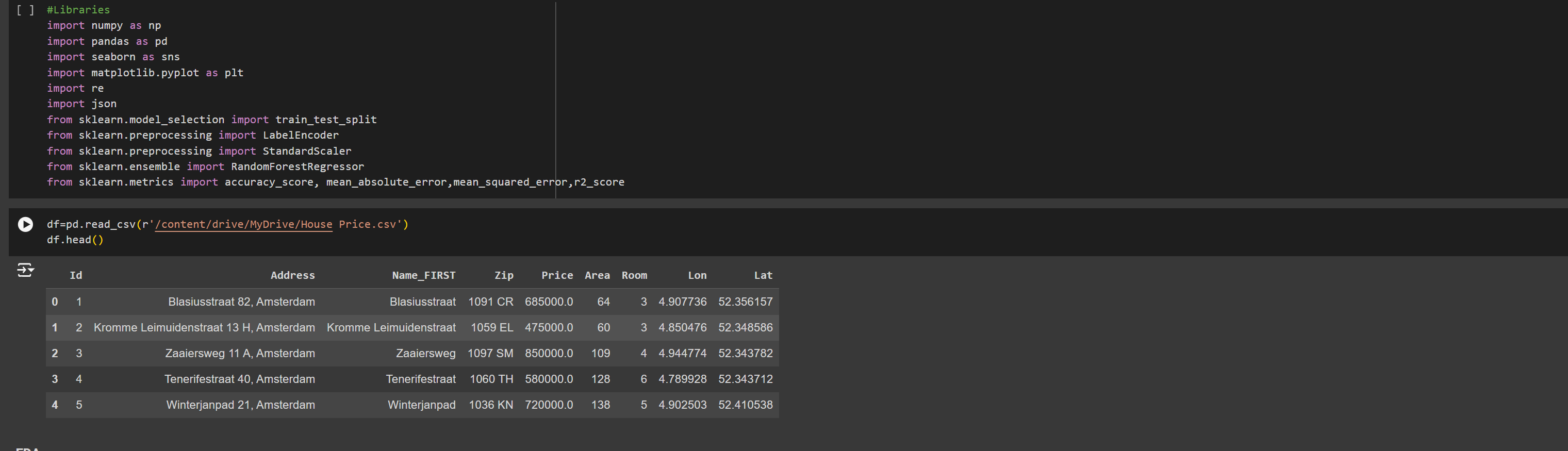
**Project Description**: Build a simple machine learning model to predict house prices based on various features such as square footage, number of bedrooms, neighborhood, and more. This project will introduce you to regression analysis, which is a fundamental concept in data science.

* **Data Collection**
* Data was collected from Kaggle and coding was done on Google Colab



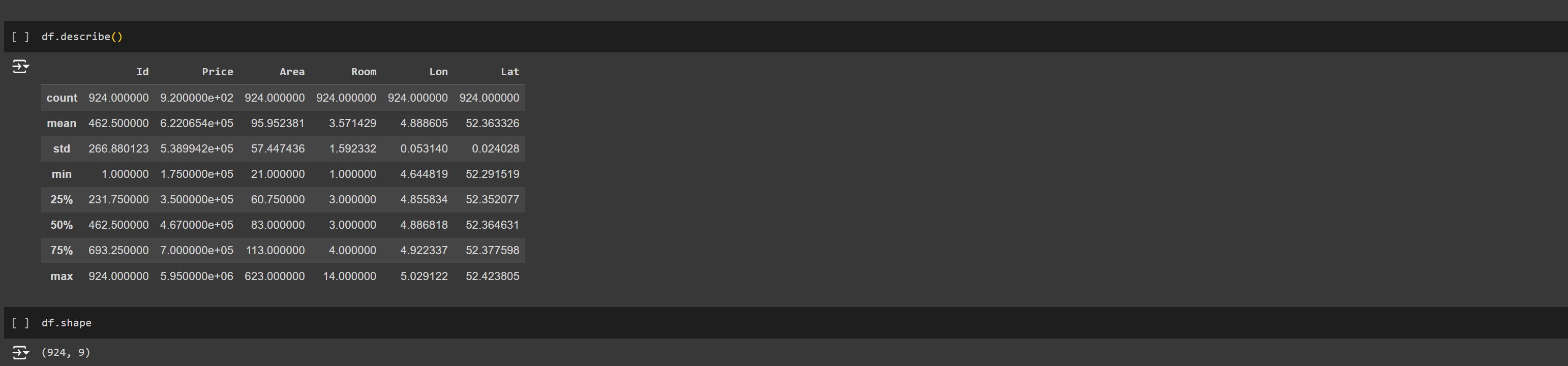
* **Import necessary Libraries and Read Data**

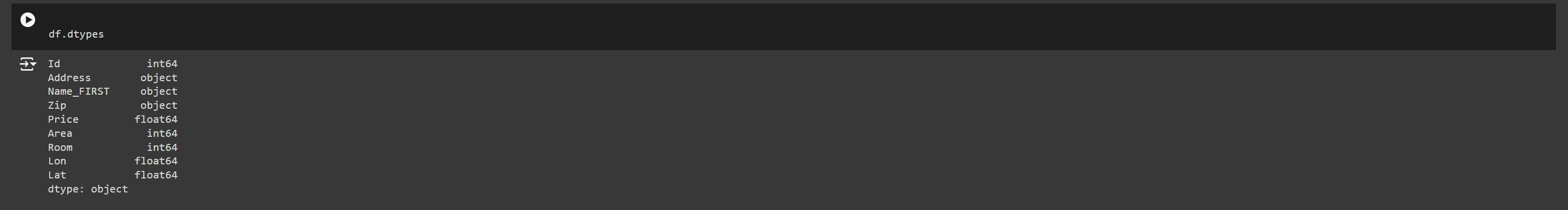
Necessary libraries were imported before starting the project and data was read .



* **Exploratory Data Analysis(EDA)**

EDA was done to understand the data in detail regarding the underlying patterns and type of data in it.This step was done to get the descriptive statistics of the data.





* **Handling missing Values**

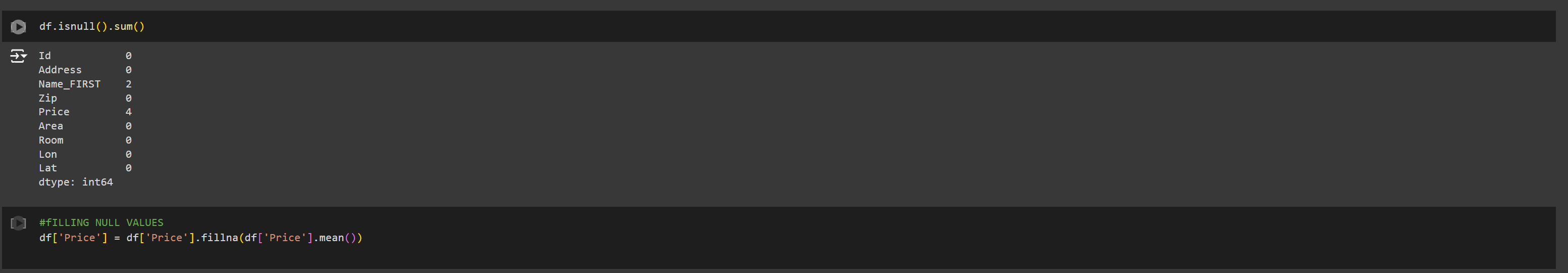
In this project I have done

**Identifying Missing Values:** Determining which columns had missing data.

**Removing Columns with Too Many Missing Values**: Dropping columns with more than 50% missing values.

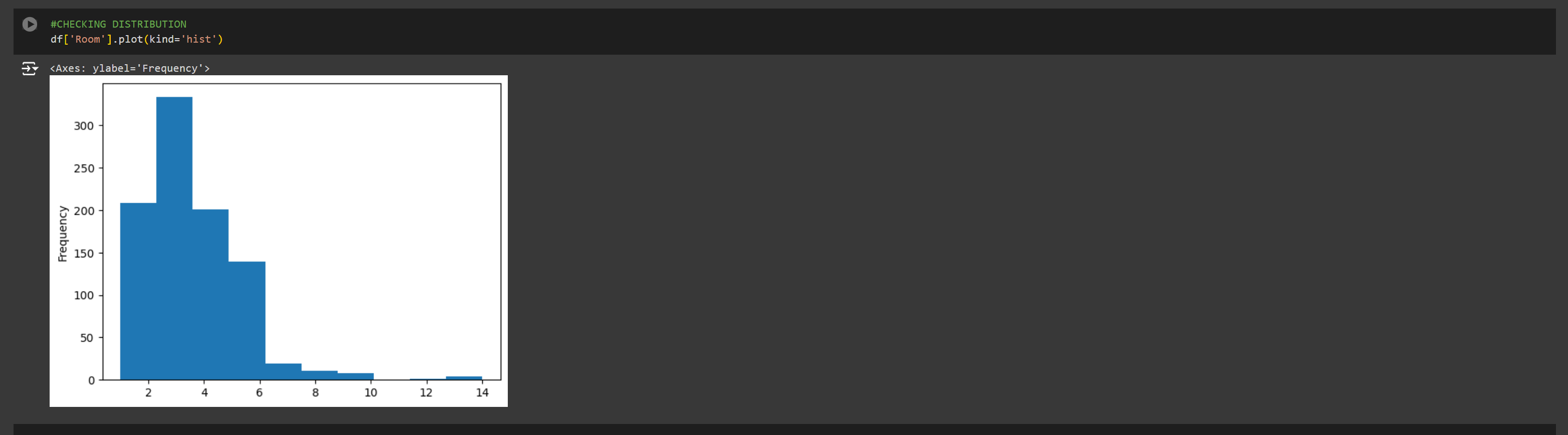
**Imputing Missing Values**: This was done to complete the data as the model wont train with missing data.

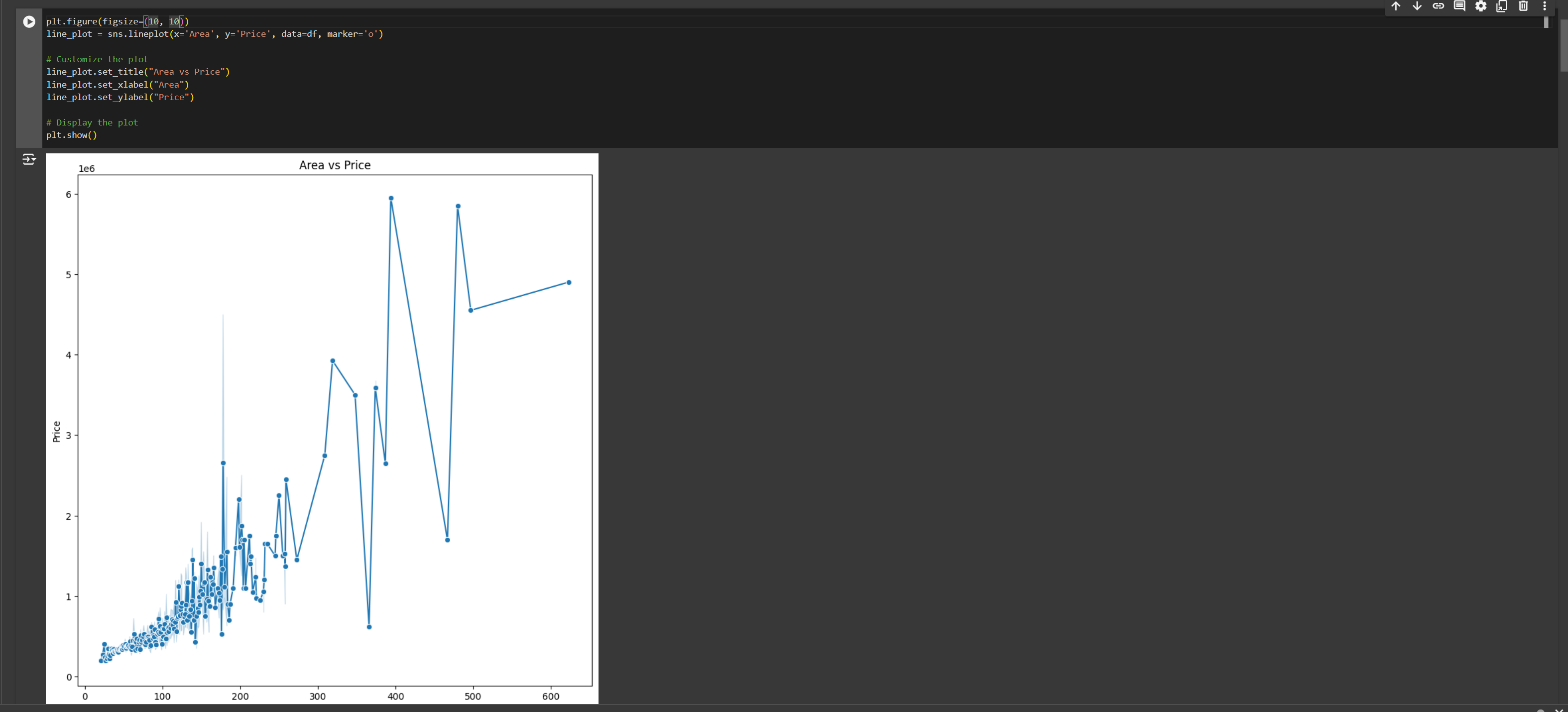
Ive used mean imputation to fill the data



* **Data visualisation**

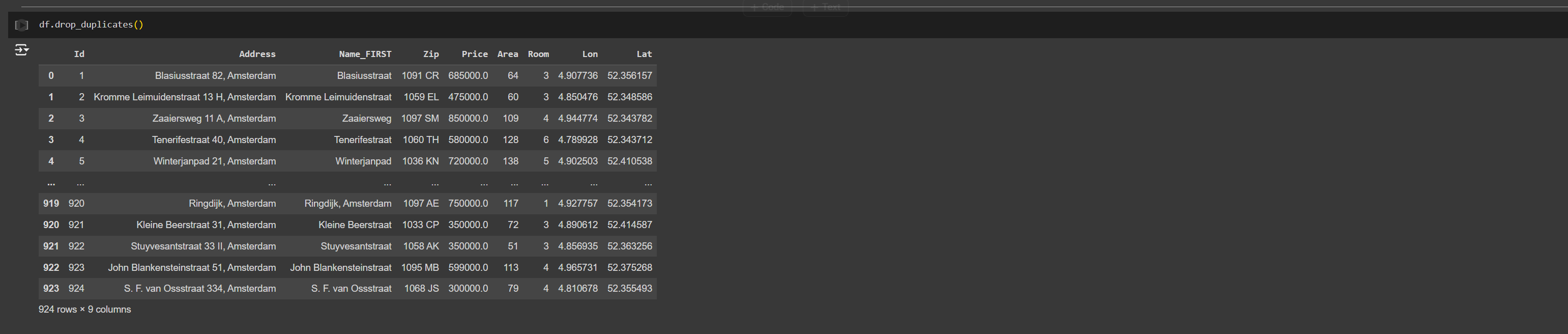
In this stage various visualisations were done to understand more on the data and the distributions .





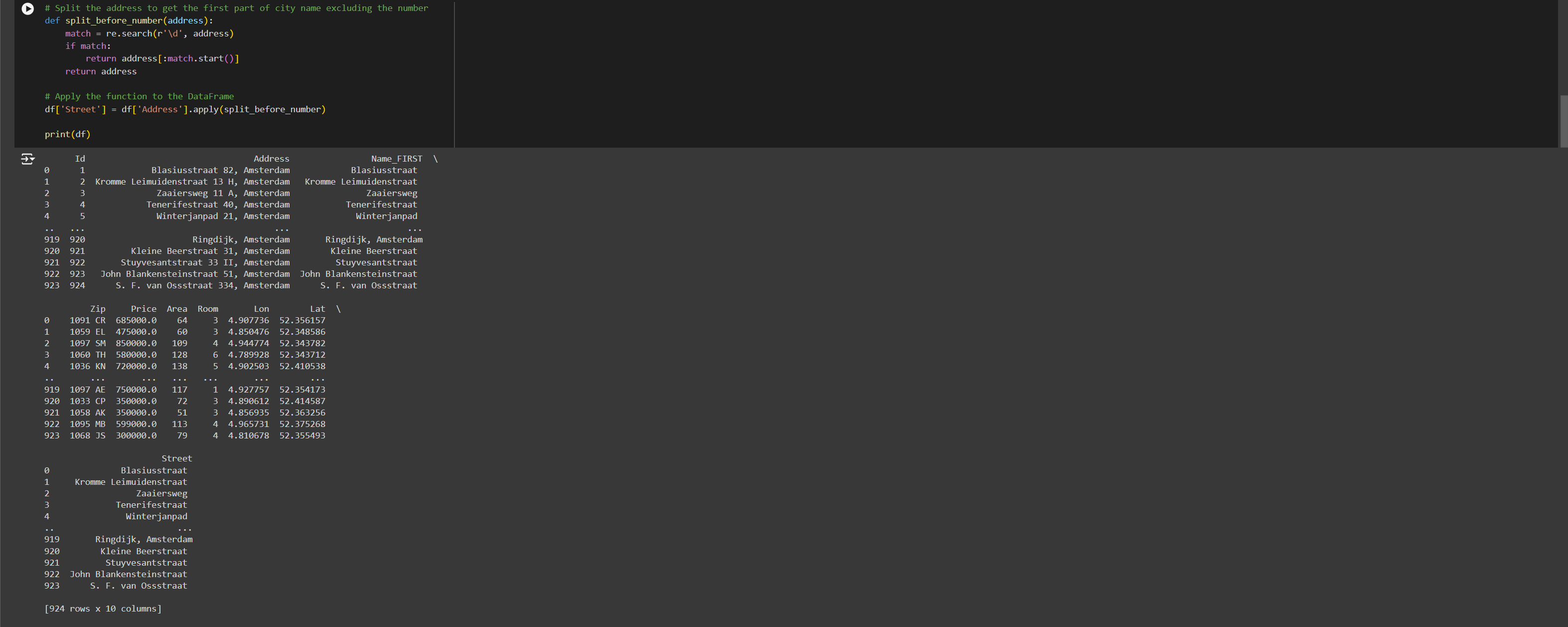
* **Drop Duplicates**

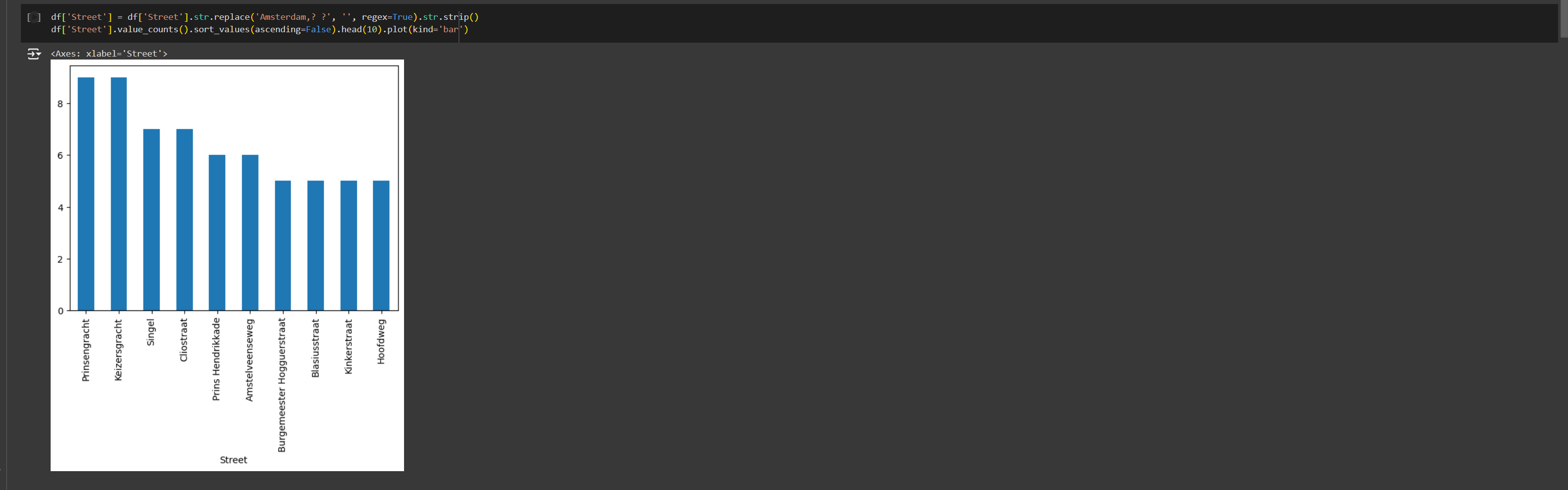
Duplicate data was dropped to reduce overfitting while training the data.



* **Feature Extraction**

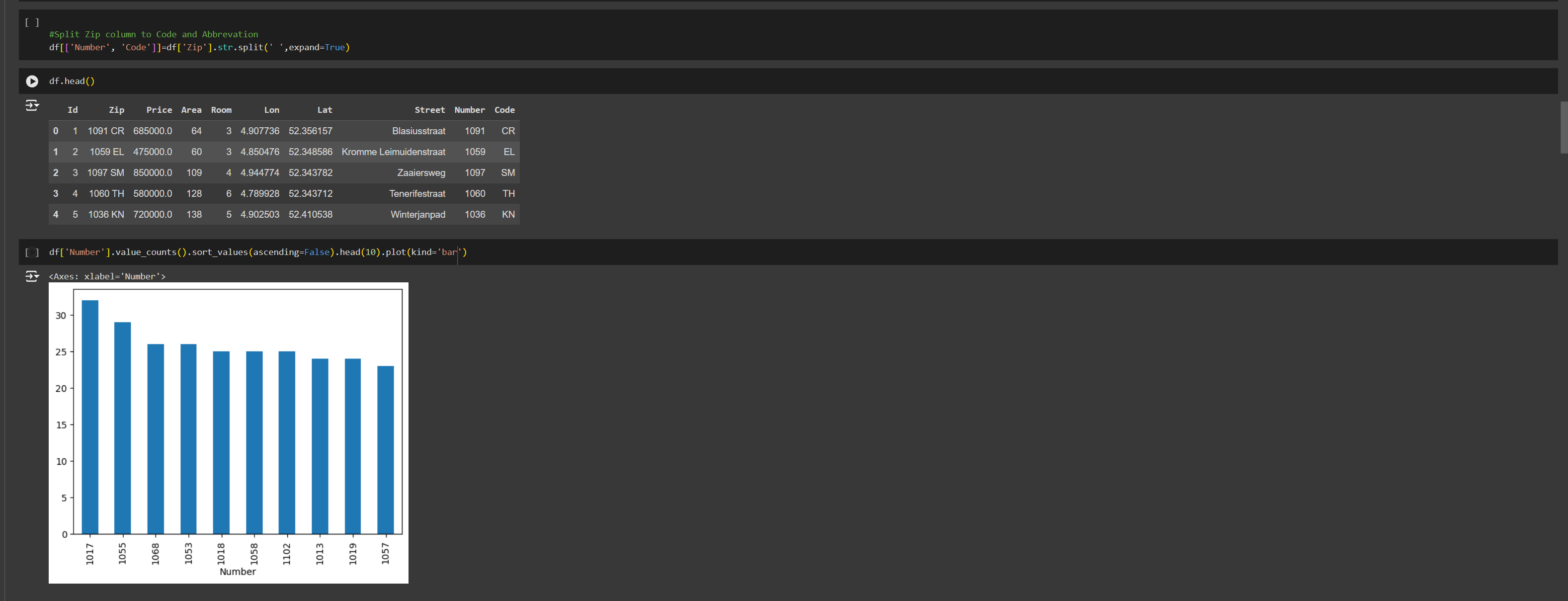
In this stage the address was split into 2 components – street and city which provided necessary insight into the data.





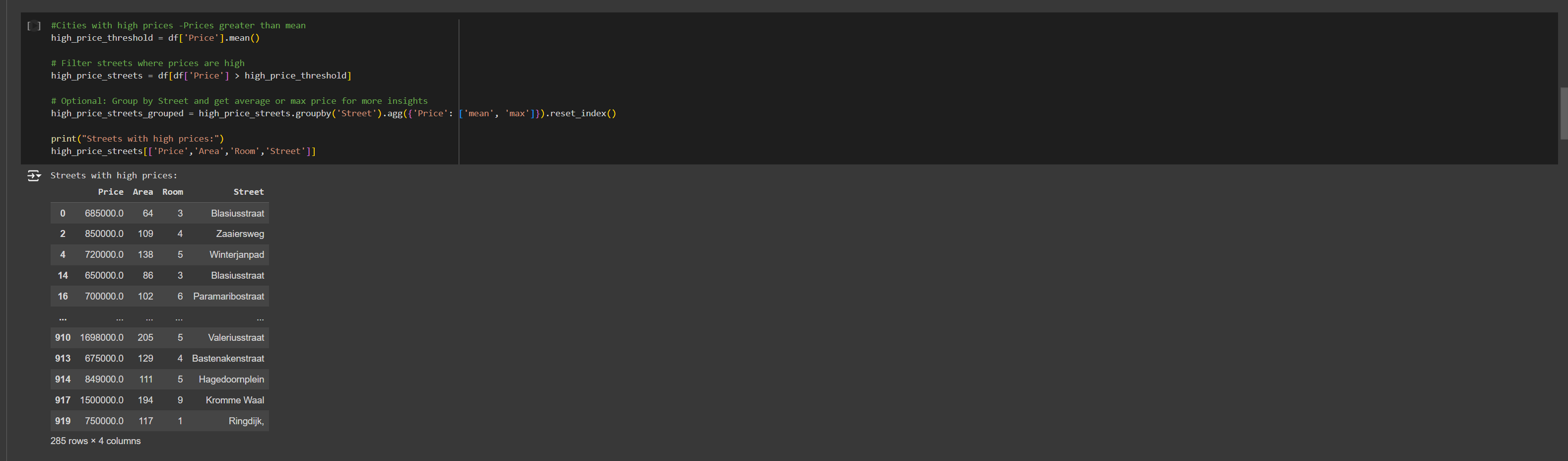
* **Data cleaning**

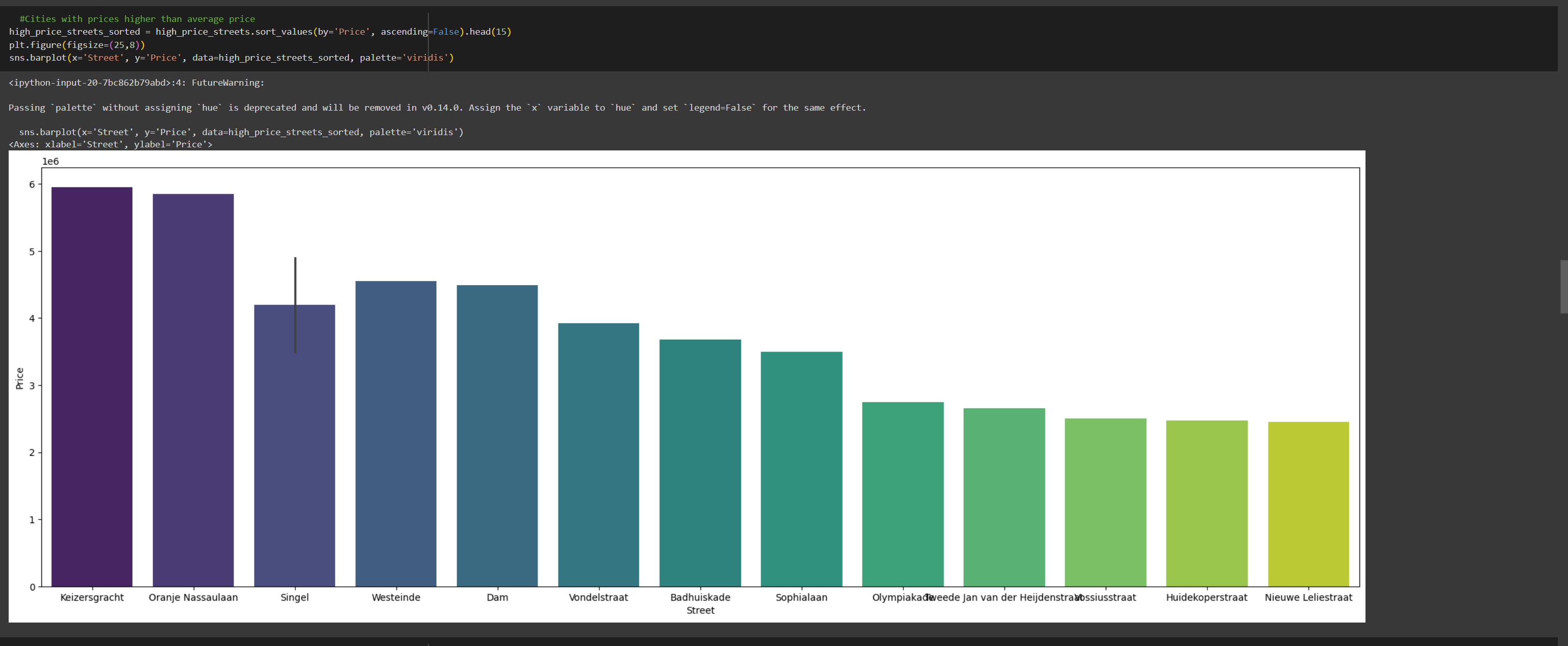
In this stage data was cleaned to make into proper format and remove unwanted data .This involved removing text from numerical format, removing data with high classes etc.



* **Finding houses which has prices above the mean price**

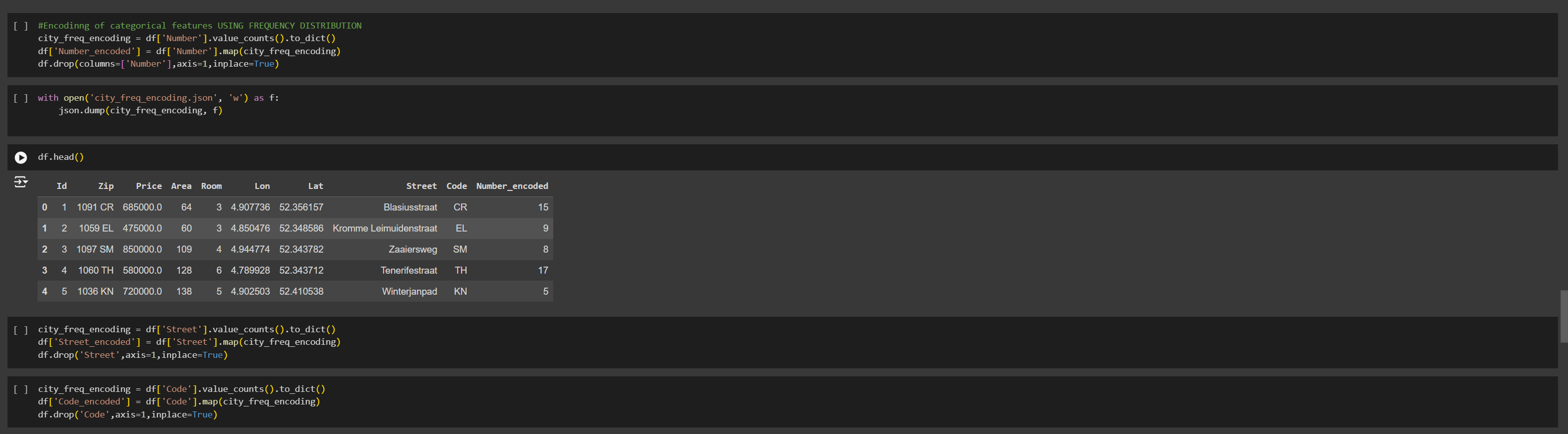
This helped to understand the pricing structure of the houses.about **265** houses had prices above the mean price.





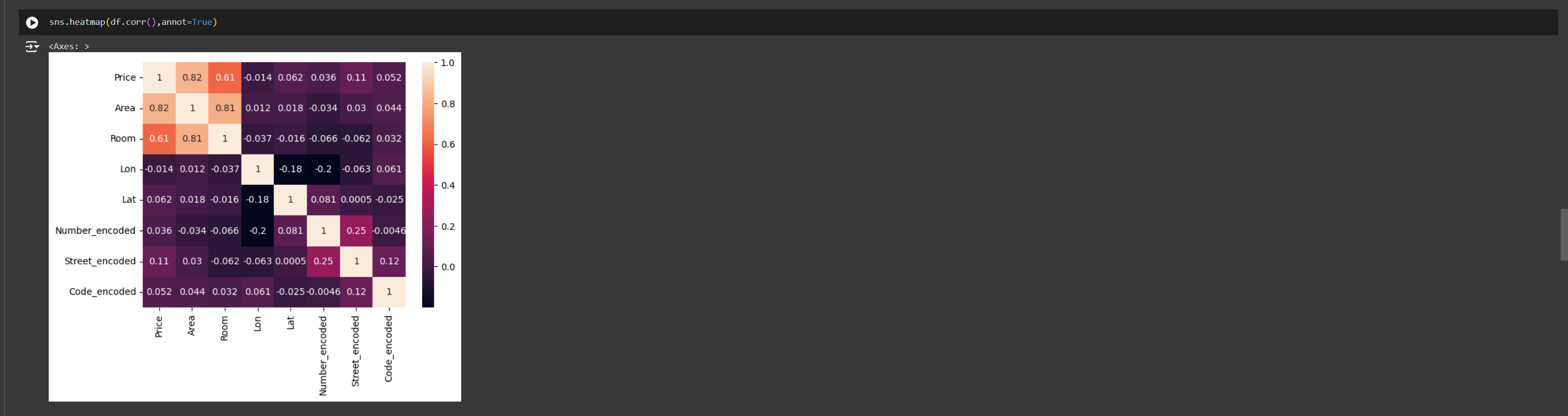
* **Encoding of categorical Features**

Encoding was done to convert categorical data into numerical data which could be used for machine learning.Ive done frequency enoding since the categories were large and replaced it with frequency of occurance.Larger the number of columns and fewer rows will result in overfitting of the model.



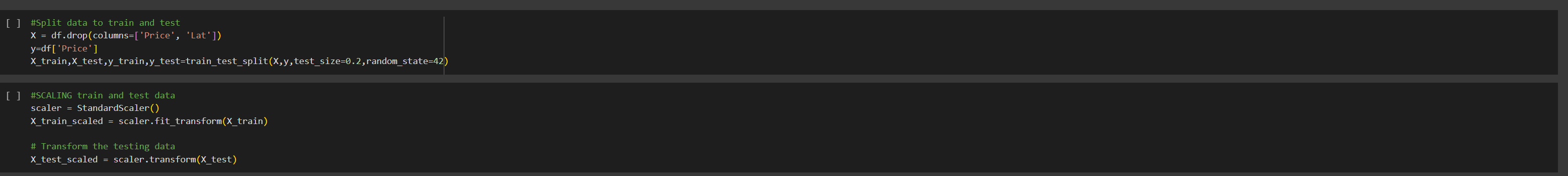
* **Feature Importance**

To find the important features,I’ve used heatmap to find the corelation of each features in the data.

****

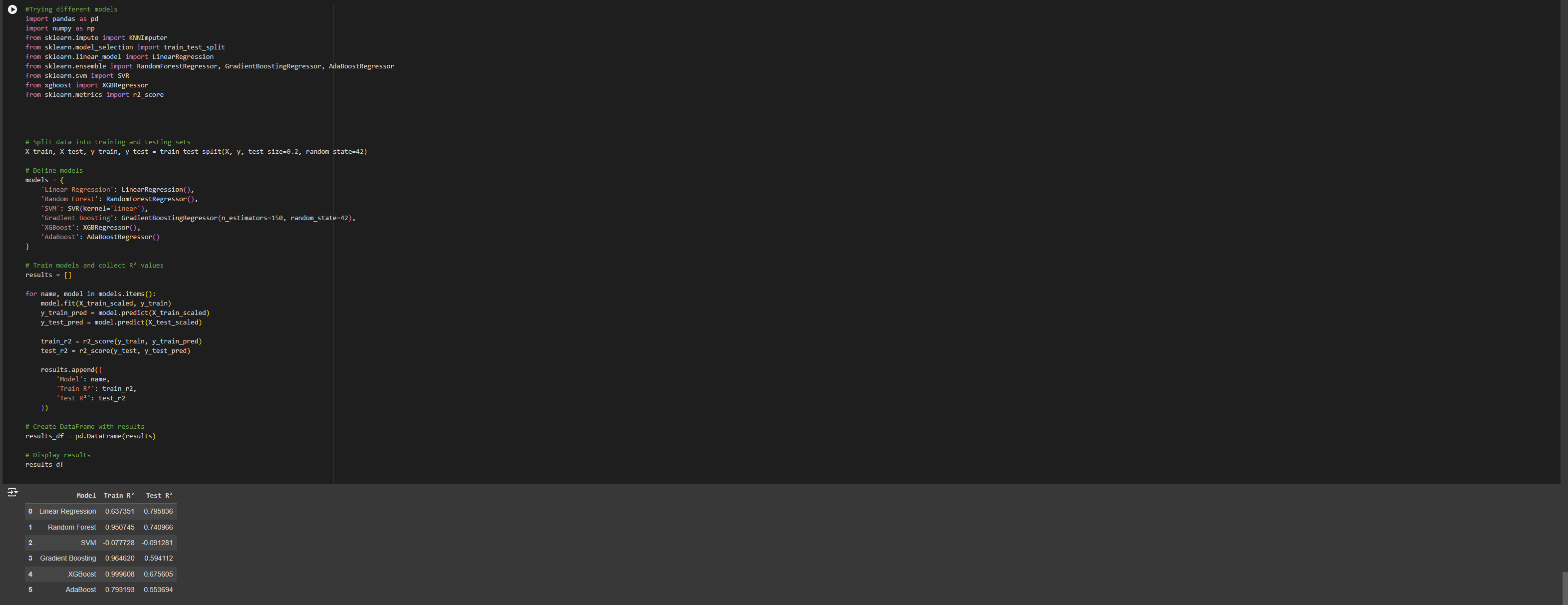
* **Split the data and Scaling**

Data was split into train and test data for training process.The data was further scaled using Standard Scalar.



* **Training and Testing the data**

For the regression process, Ive used many machine learning models available to analyze the performance of each model.From the process RandomForest provided best R2 value and performed well compared to other models.



* **HyperParameter Tuning**

The model was tuned to get best results using GridSearchCV .

