Data Analysis Report

Housing Price Predictor

by Zhaklin Yanakieva

Table of content

Versioning table

Time period(weeks)	Comments on changes	Version
5 — 7	Initial changes to the document — creating the structure	1.0.0
8 — 10	Filling up with the needed information, such as explanation to the graphs, etc., and summarizing	2.0.0
11 — 12	Adding the Data storage solution	3.0.0

Overview:

The data analysis report is to provide the findings of the research that was conducted during the collection of the data for the project.

Background:

In the DataCollectionLedger file is explained more about how and from where the data was collected. In the proposal of the project is explained the goal and for what models the data will be used.

Data storage solution:

An essential part of Machine Learning is the data storage solution for the selected data and machine learning model. In order to accomplish the most efficient manner of working with data during this project, the following tools were used:

- Git Version Control
- Data Version Control (DVC)

Git Version Control:

Git has been a popular tool among programmers and it is so for a reason. It allows tracking changes in any set of files, usually used for coordinating work among programmers collaboratively developing source code during software development.[1]

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

Data Version Control (DVC):

Data Version Control is a new type of data versioning, workflow, and experiment management software that builds upon Git (although it can work stand-alone). [2]Using Git and DVC, machine learning teams can version experiments, manage large datasets, and make projects reproducible. By utilizing DVC data will be tracked and stored in an effective and efficient way because the data is accessible from everywhere via internet connection for every contributor.

Summary:

- DVC will create reference files to data versions
- Git will store the DVC files

In this project, I decided to not use extraction from a csv file for the data, but to scrape it. Web scraping is the process of using bots to extract content and data from a website. Scraping extracts underlying HTML code and, with it, data stored in a database. The scraper can then replicate entire website content elsewhere.[3] After extracting the data from two different websites - 'Pararius' (https://www.pararius.com/apartments/eindhoven)and 'Friendly Housing' (https://www.friendlyhousing.nl/nl). Changes were made to like cleaning and processing it so as to make it more suitable to work with and acceptable to store.

Purpose of the extraction effort:

Understanding, preparing and cleaning the data require first to know what type of it will be used and for what reason. The purpose of the web scraping in the Housing Price Predictor is to extract data that will be updated all the time (as the website changes). In this case, the results will be up to date, which states that the predictions will be more accurate.

---

[2] https://dvc.org/doc/user-guide/what-is-dvc
[3] https://www.imperva.com/learn/application-security/web-scraping-attack/

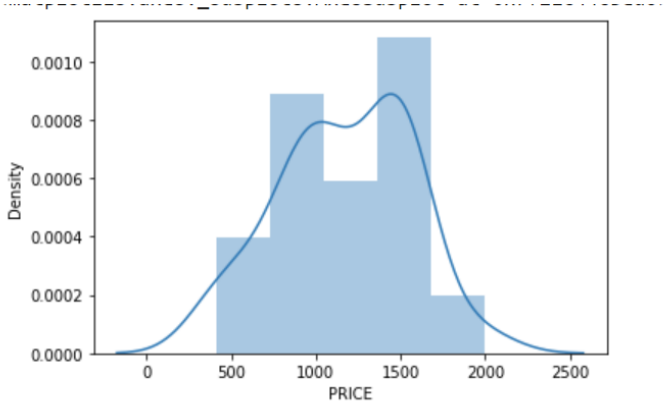
Data integration:

In the project, I decided to scrape data from two different websites in order to gather more information, and then saved the data from each website into a pandas dataframe. After that, data integration was applied in order to combine the information into one dataframe. The method that was used was 'Inner join', which gets me the intersection between the datasets.

Methods:

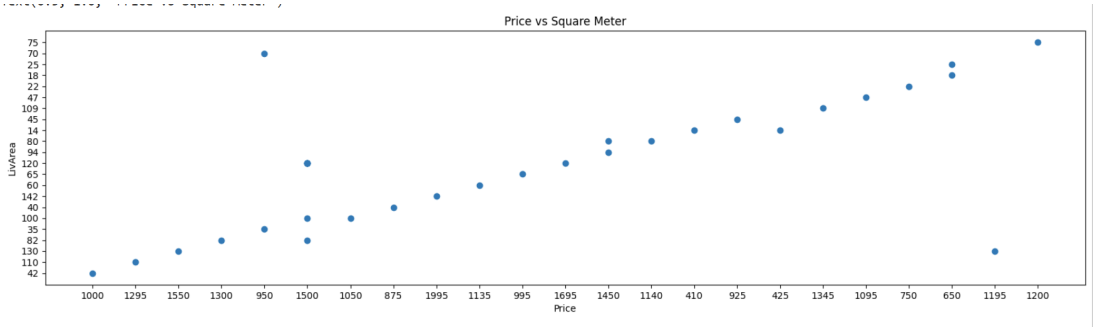
There are many ways to retrieve data from a source you selected. Common ways of extracting are from a file (csv or txt), from a database, JSON, API or web-scraping. In my project, I chose to extract the data from a website: (https://www.pararius.com/apartments/eindhoven; https://www.friendlyhousing.nl) by web scraping. I first extract the needed information about the housing properties and then, save the data into a pandas dataframe. In the end, I save the data into a csv-file so as to be able to apply data integration for at least three different pages of the website.

Results:

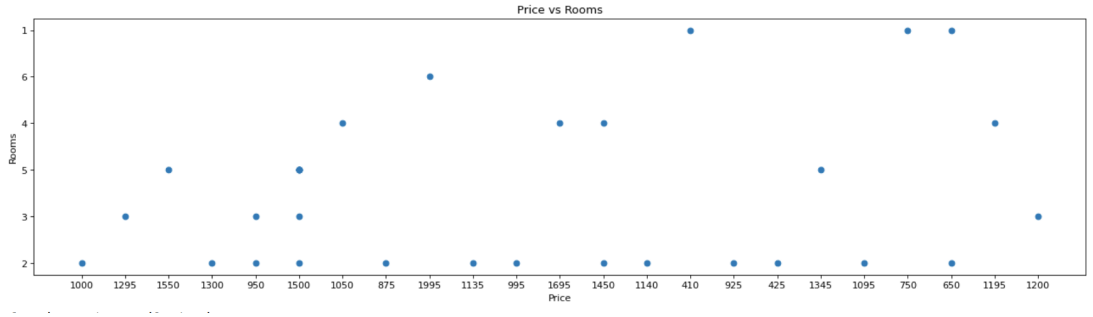


The histogram above represents the distribution of the targeted value - the price. It is noticeable that the 'PRICE' distribution is not skewed, but normally distributed, which means that the data is

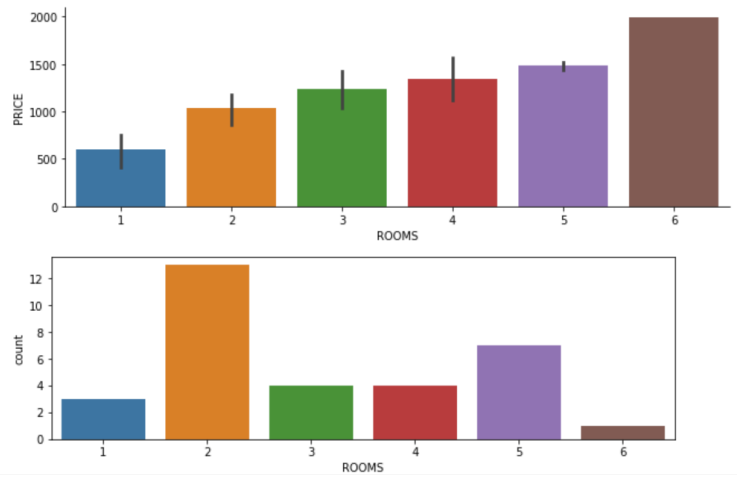
symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.

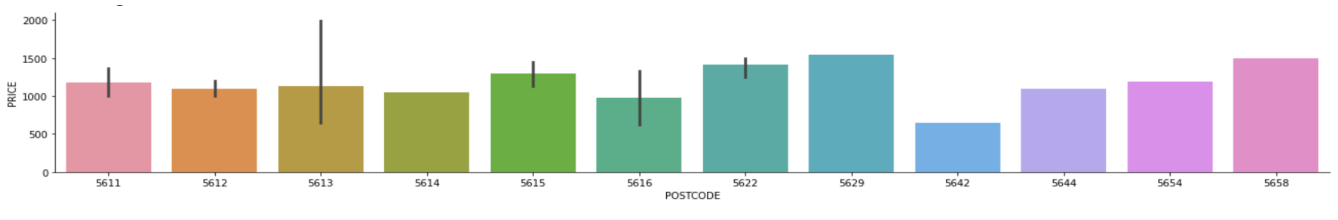


This diagram displays the price in accordance to the living area of a housing property. It can be noticed that there is a positive correlation between the price and the living area, which means that the variables move in tandem—that is, in the same direction. This means that whenever one variable increases, the other decreases. For instance, the price increases with the increase in the living area.

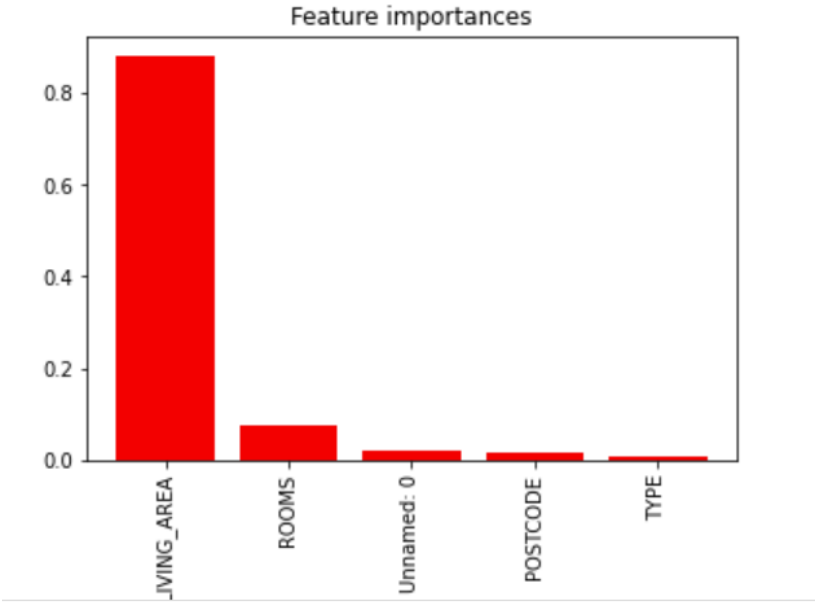


The same situation is with this diagram. There is a slightly positive correlation, which states that the more rooms a real estate has, the more the price will be.





Factor plot is informative when we have multiple groups to compare. There are many parameters under factor plot and I want to focus on basic parameters. In this case, it depicts how the rooms have an effect on the price. Also, it shows the variance in quantity of rooms taken. Concluding, real estate with 5 rooms has the highest Price while the sales of others with rooms of 2 is the most sold ones.



This diagram outlines the importance of each feature after the training of the 'Random forest model'. 'Random Forest' determined that overall the living area of a home is by far the most important predictor. Following are the size of above grade rooms and postcode.

Summary:

In conclusion, for the data collection part, I decided to use web scraping as e technique because it gives the opportunity to work with a data set that is up to date and therefore, makes more accurate summaries. For the data processing, different types of data transforms were used to expose the data structure better, so we may be able to improve model accuracy later. Standardizing was made to the

data set so as to reduce the effects of differing distributions. The skewness of the features was checked in order to see how distorted a data sample is from the normal distribution. Rescaling (normalizing) the dataset was also included to reduce the effects of differing scales

Then, for the modelling, I used two models to determine the accuracy - Linear Regression and Random Forest. Linear Regression turns out to be the more accurate model for predicting the house price. It scored an estimated accuracy of 68%, out performing the Random Forest - 66%. Random Forest determined that overall the living area of a home is by far the most important predictor. Following are the sizes of above rooms and postcode.

Bibliography

"DVC." https://dvc.org/doc/user-guide/what-is-dvc.

"Git." https://en.wikipedia.org/wiki/Git.