FORECASTING PRICES: STRATEGIES AND TECHNIQUES

INTRODUCTION

The Research Question: Can we predict Real Estate prices based on a large DataSet by analyzing crucial points in data and using various techniques of analysis and machine learning

INFORMATION GATHERING

The first step in every project is to gather the data for the analysis.
Some of the sources include open API for gathering data.
I will use technique called "Crawling" to harvest my dataset and won't use any open source API.
To achieve my goal I will use Selenium library.
The Domain : https://www.grekodom.com

Data Introduction

After gathering my data set, as you can see I choose my crucial fields as :
Price, Sqm, Region and etc. My raw data includes : 54,194 values in total.
The next step is to clean the dataset to remove irrational values, duplicates and do a small research of what we got.

	Price	Sqm	Region	Year Of Construction	Dist From Sea in m	Dist From Air in km	Type
0	350000	159 m²	Thessaloniki	1962	100 m	15 km	Flat
1	235000	118 m²	Athens	1975	2400 m	26 km	Flat
2	137000	108 m²	Athens	2002	1400 m	46 km	Duplex
3	170000	160 m²	Olympic coast	1999	10000 m	110 km	Detached house
4	400000	800 m²	Western Peloponnese	No Const Land	Missing	Missing	Land
...	...	...	...	...	...	...	...
7737	280000	120 m²	Thessaloniki/suburbs	2007	10000 m	6 km	Flat
7738	260000	120 m²	Thessaloniki/suburbs	2007	10000 m	6 km	Flat
7739	260000	120 m²	Thessaloniki/suburbs	2007	10000 m	6 km	Flat
7740	240000	120 m²	Thessaloniki/suburbs	2007	10000 m	6 km	Flat
7741	240000	120 m²	Thessaloniki/suburbs	2007	10000 m	6 km	Flat

7742 rows × 7 columns

Cleaning the Data

The steps I took to clean my data:

1. Remove the duplicates + missing values - gives false statistics
2. Drop irrational values - dispose logical mistakes
3. Drop irrelevant "Type" - drop the "land" type because valuation of land is different.
4. Get dataset ready for machine learning - drop the strings and special chars.

	Price	Sqm	Region	Year Of Construction	Dist From Sea in m	Dist From Air in km	Type
0	350000	159	Thessaloniki	1962	100	15	Flat
1	235000	118	Athens	1975	2400	26	Flat
2	137000	108	Athens	2002	1400	46	Duplex
3	170000	160	Olympic coast	1999	10000	110	Detached house
6	142000	78	Athens	2006	1700	45	Flat
...	...	...	...	...	...	...	...
7734	290000	107	Crete	2012	80	16	Maisonette
7735	1500000	720	Thessaloniki/suburbs	2007	10000	6	Flat
7736	280000	120	Thessaloniki/suburbs	2007	10000	6	Flat
7738	260000	120	Thessaloniki/suburbs	2007	10000	6	Flat
7740	240000	120	Thessaloniki/suburbs	2007	10000	6	Flat

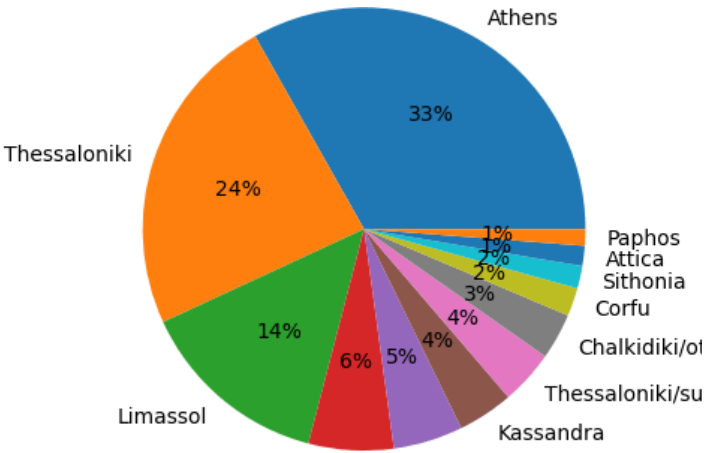
3299 rows × 7 columns

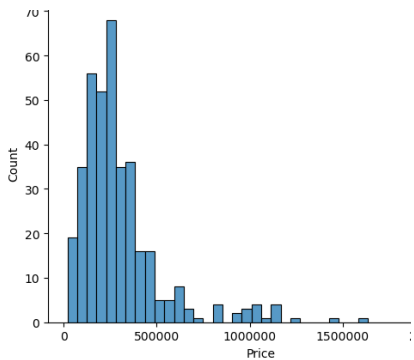
Analyzing the Data

By these EDA tools we can see the distribution of Real Estate across the cities of Greece. As we can see the Athens has the highest distribution value. In this case we will focus on Athens because of the following reason :

The evaluation of one city to another might be different.

The influence parameters can change depending on the city.





Athens	385
Thessaloniki	275
Limassol	163
Olympic coast	72
Crete	59
Kassandra	47
Thessaloniki/suburbs	46
Chalkidiki/other	39
Corfu	24
Sithonia	19
Attica	17
Paphos	14
Loutraki	11
Peloponnese	7
Euboea	6
Larnaka	6
Kavala	4
Eastern Peloponnese	4
Athos	3
Protaras	3
Cyclades	2
Central Greece	2
North Greece	2
Asprovalta	2
Nicosia	2
Islands	1
Western Peloponnese	1
Thrace	1
Serres	1
Xanthi	1

Name: Region, dtype: int64

MACHINE LEARNING

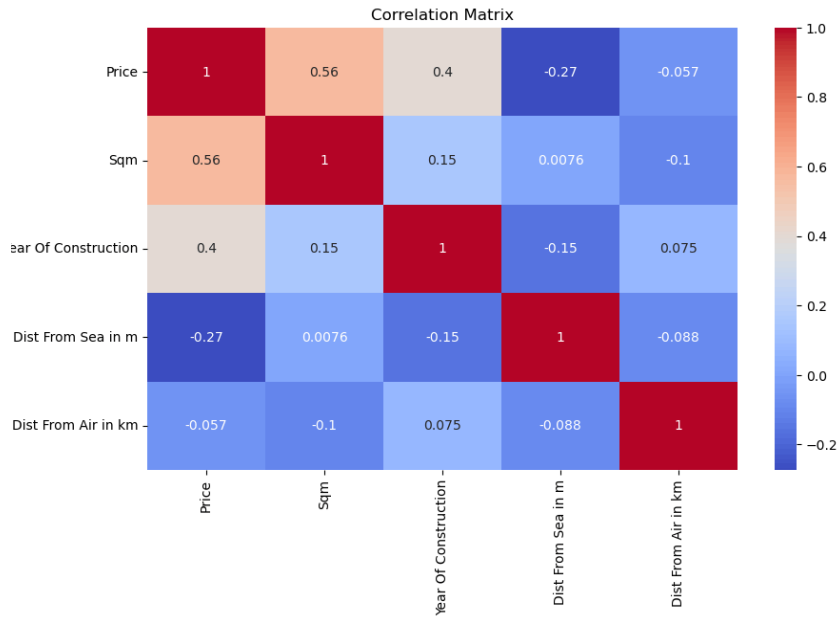
Regression analysis is a statistical technique that can be used to forecast prices. This technique involves analyzing the relationship between two or more variables to predict future outcomes. For my project I will use linear regression model.

Machine Learning Setup

In order to implement machine learning correctly we should find the most significant influence points in our data, I will be using correlation coefficient function in order to find the strongest relationship between the predicting values "Price" and the independable values.

We will take a look at 1st column named price :
We can see clearly that our most significant values are:

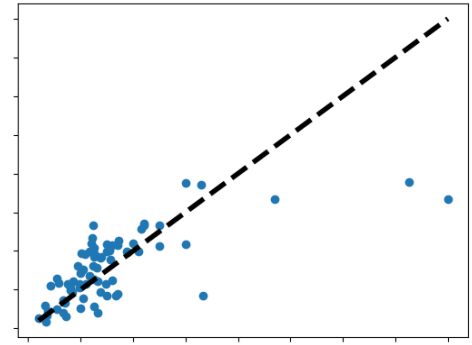
1. Sqm with the score of 0.56
2. Year of construction with score 0.4

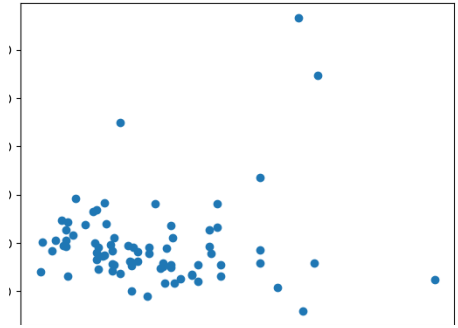


Machine Learning Model Test

After implementing the training for the machine learning using linear regression we need to test it's performance. The 1st score I would test is R^2 score. In my case my R^2 score in this model is : 0.48
Not good, not terrible.
Another test is the regression line test:
Shortly this diagonal line represent where the perfect predictions occur and the blue dots are predictions of the machine. If the blue dot is upper of the diagonal it means that machine overpriced the actual price other case is underpriced.

The second test is residual plot test:
Shortly it represents the residuals of the actual value and predicted ones. One of the methods to see that machine performs effectively is that residual plot should not have any clear patterns and spreaded as much as possible.





Actual Predictions

	Price	Sqm	Region	Year Of Construction	Dist From Sea in m	Dist From Air in km	Type	Machine Predictions	Difference
11	175000	70	Athens	1968	6000	40	Flat	156085	18915
13	64000	70	Athens	1966	7000	34	Flat	140093	-76093
14	165000	60	Athens	1962	9000	30	Flat	69535	95465
15	46000	50	Athens	1970	8000	38	Flat	74018	-28018
16	295000	84	Athens	1982	1400	24	Flat	323040	-28040
...	...	...	...	...	...	...	...	...	...
1425	355000	135	Athens	2000	1000	19	Flat	571236	-216236
1427	310000	121	Athens	2009	3000	2	Flat	545947	-235947
1436	670000	213	Athens	1988	30	32	Flat	788988	-118988
1437	370000	160	Athens	2009	4500	28	Flat	641462	-271462
1439	220000	78	Athens	2014	100	10	Flat	451508	-231508

The following data frame is generated on based on our trained model on the new data set that it has never seen before. As we can see we got an actual price column, the predicted price and the difference between them.
By the difference column we can see and decide if the flat is worth checking or is it highly overpriced. For some people that does not have a proper information about a real estate it can be handy tool to use.

CONCLUSION

Forecasting prices is a complex process that requires a combination of techniques and strategies. By using gathering data tools, regression analysis, machine learning, market research and all other crucial researches. In my opinion I had achieved my goal for predicting approximately correct price based on my research .

It takes time and effort to forecast correctly and train machine models to the maximum but it's worth every minute. Strong analytical tools will help students to achieve various victories in any field possible.

THANK YOU

Do you have any questions?

dima933@walla.com
2nd Year C.S. Student HIT