By : Roy Shlomo Chen & Yarin Akiva

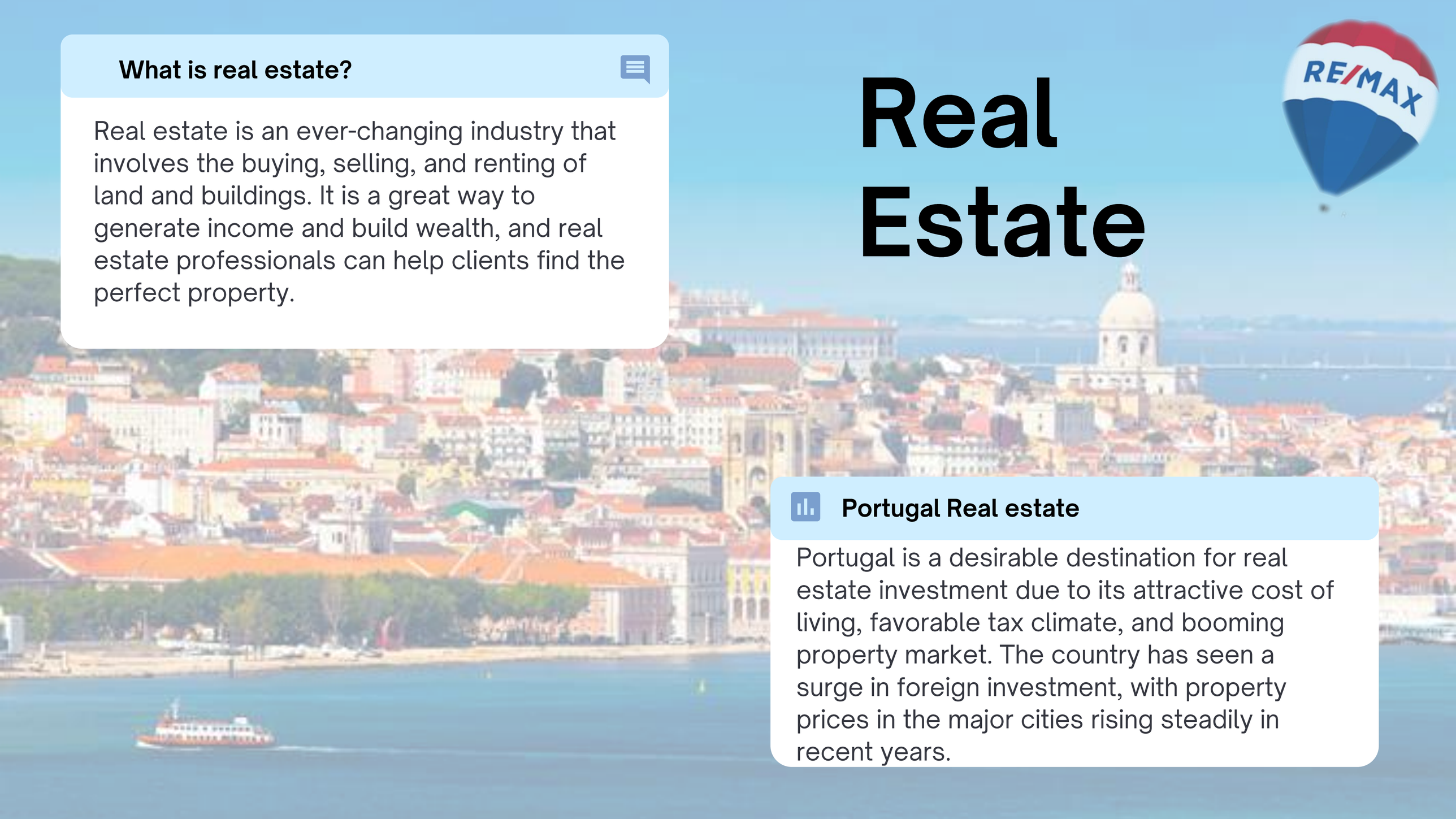
Data Science Project

Portugal Real Estate

Real Estate

What is real estate?

Real estate is an ever-changing industry that involves the buying, selling, and renting of land and buildings. It is a great way to generate income and build wealth, and real estate professionals can help clients find the perfect property.



Portugal Real estate

Portugal is a desirable destination for real estate investment due to its attractive cost of living, favorable tax climate, and booming property market. The country has seen a surge in foreign investment, with property prices in the major cities rising steadily in recent years.

Research Question

How to identify a good investment property in Portugal ?

Why Did We Choose This Subject?

Increase in Jewish applications for a Portuguese passport

In recent years, Jews who could prove that they were deported from Spain, could submit an application to the Portuguese authorities and the Jewish community living in Porto or Lisbon, and ask for citizenship in Portugal only by proving that they are descendants of those deported from Spain.

The hardening of the Portuguese government's conditions

At the end of 2022, the Portuguese government decided to tighten the conditions for obtaining citizenship and a Portuguese passport for the descendants of those deported from Spain, and from now on you also have to show affinity to the state of Portugal.

Why did we choose this topic?

The solution

One of these ways is to invest or purchase real estate in Portugal.

We decided to check whether it is possible to analyze the information on properties in Portugal using the Remax Portugal website, and to determine what is a good property for investment in Portugal.

The goal

The goal is to help those Jews who want to become citizens in Portugal, and they are now faced with the tightening of the conditions for obtaining citizenship/

Main Process Steps

1 Obtaining Data

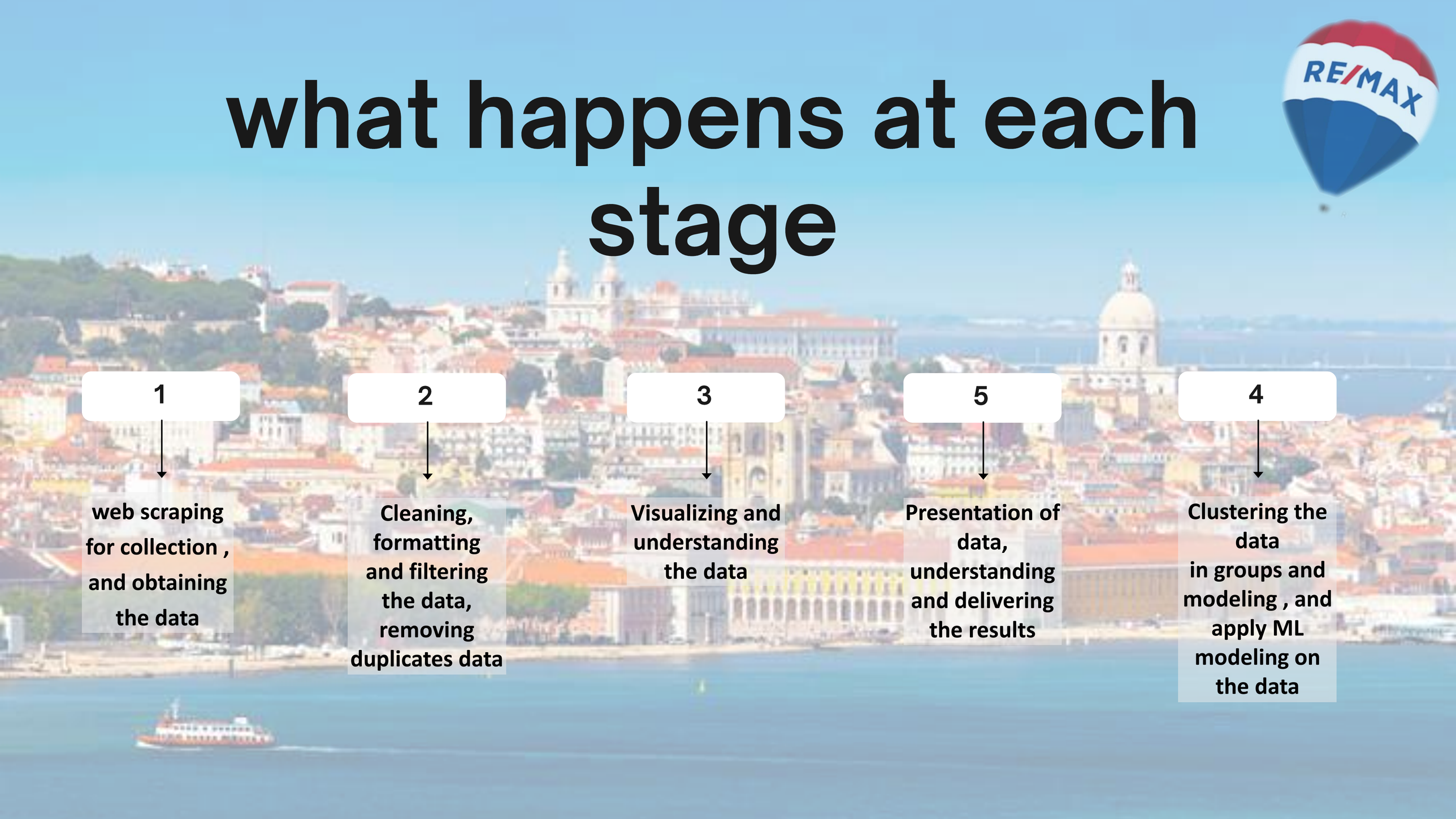
2 Data Handling

3 Exploring Data

4 Machine Learning

5 Interpreting Data

what happens at each stage



Step 1 Obtaining Data

Source

In this step, we will collect information of properties in Portugal in various big cities we will crawl along web pages and scrape information about properties features

Main tools :
BeautifulSoup, Selenium

Source Data :
https://www.remax.pt/en/

	Price	Address	Prop_type	Size	Bath	Bed
0	72 500 €	Lisboa - Vila Franca de Xira, Alhandra, São...	House	80 m2	1	3
1	680 000 €	Lisboa - Oeiras, Barcarena	House	119 m2	3	3
2	260 000 €	Lisboa - Torres Vedras, A dos Cunhados e Ma...	House	206 m2	2	5
3	425 000 €	Lisboa - Mafra, Azueira e Sobral da Abelhei...	House	233 m2	2	3
4	870 000 €	Lisboa - Sintra, Queluz e Belas	House	233 m2	4	4

The process

To collect each property, we used Selenium because this is a dynamic scrapping, After we passed the dynamic scrapping difficulty We used BeatifulSoup to store the data in dataframes

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11665 entries, 0 to 11664
Data columns (total 6 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Price       11665 non-null  object
 1   Address     11665 non-null  object
 2   Prop_type   11604 non-null  object
 3   Size        11665 non-null  object
 4   Bath        11665 non-null  object
 5   Bed         11665 non-null  object
dtypes: object(6)
memory usage: 546.9+ KB
```

	Price	Address	Prop_type	Size	Bath	Bed
0	4 500 € / Monthly	Porto - Porto, Lordelo do Ouro e Massarelos	Business Office	452 m2	6	--
1	50 € / Monthly	Porto - Valongo, Valongo	Garage	15 m2	0	0
2	700 € / Monthly	Porto - Marco de Canaveses, Várzea, Aliviad...	House	80 m2	0	2
3	1 600 € / Monthly	Porto - Porto, Cedofeita, Santo Ildefonso, ...	Condo/Apartment	145 m2	3	3
4	1 100 € / Monthly	Porto - Maia, Cidade da Maia	Condo/Apartment	110 m2	2	2
5	900 € / Monthly	Porto - Porto, Campanhã	Condo/Apartment	100 m2	2	3
6	950 € / Monthly	Porto - Porto, Cedofeita, Santo Ildefonso, ...	Condo/Apartment	49 m2	1	1
7	770 € / Monthly	Porto - Gondomar, Gondomar (São Cosme), Val...	Condo/Apartment	88 m2	1	2
8	850 € / Monthly	Porto - Gondomar, Rio Tinto	Condo/Apartment	110 m2	1	3
9	700 € / Monthly	Porto - Paredes, Paredes	Condo/Apartment	90 m2	2	2

Step 2 Data Handling

Real Estate shape: (11665, 6)

What we change

In this step, we cleaning, formatting and filtering the data.
-Splitting columns
-Remove duplicates rows and cols
-Remove unwanted marks (€ , '/')
-Edit Types of variables

```
buy['Price'] = buy['Price'].str.replace(' ', '')
buy['Price'] = buy['Price'].str.replace('€', '')
buy['Address'] = buy['Address'].str.replace('Porto ', '')
buy['Address'] = buy['Address'].str.replace('Lisboa ', '')
buy['Address'] = buy['Address'].str.replace('Braga ', '')
buy['Address'] = buy['Address'].str.replace('-', '')
buy['City'] = buy['Address'].str.split(',', expand=True)[0]
buy['Division'] = buy['Address'].str.split(',', expand=True)[1]
buy['City'] = buy['City'].str.strip()  # removing leading and tailing white spaces
buy['Size'] = buy['Size'].str.replace('m2', '')
```

```
buy['Price'] = pd.to_numeric(buy['Price'], errors='coerce')
buy['Size'] = pd.to_numeric(buy['Size'], errors='coerce')
buy['Bath'] = pd.to_numeric(buy['Bath'], errors='coerce')
buy['Bed'] = pd.to_numeric(buy['Bed'], errors='coerce')
```

Types Handling

For each column we change the type from category to numeric , help in the next step and make more easiest to work on the data

Prop_type
House
House
House
House
House
...
Condo/Apartment
Condo/Apartment
Condo/Apartment
Condo/Apartment
Condo/Apartment

Property Type

Using replace map - we change all the property type from category to numeric
'1' to Houses and '2' to apartments

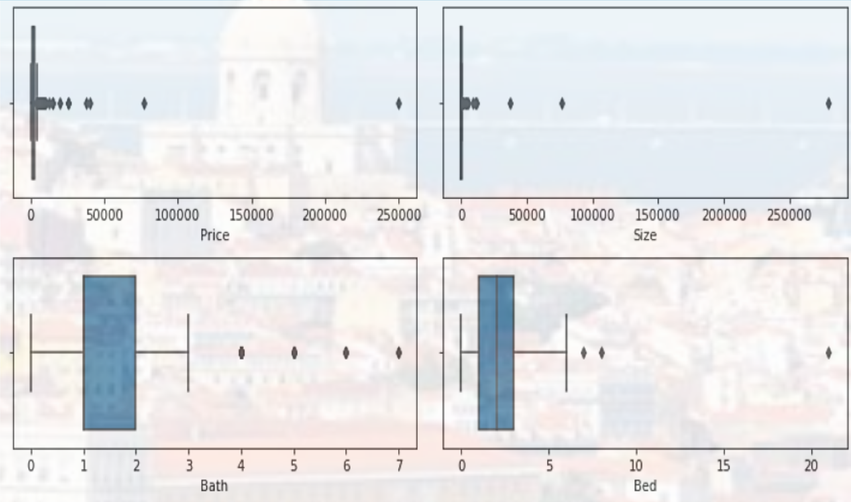
```
replace_map={'House':1,'Condo/Apartment':2}
buy2.replace(replace_map, inplace=True)
```

Prop_type
1
1
1
1
1
...
2
2
2
2
2

Step 3 Exploring Data

The Process

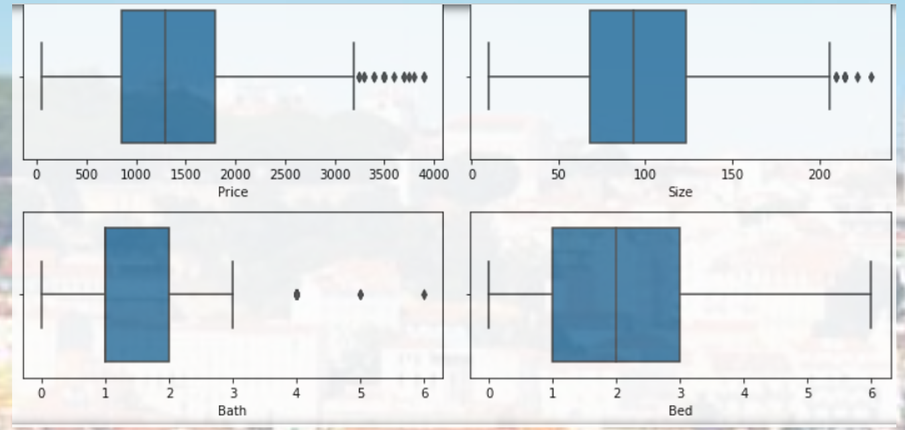
We did some visualization with our data after all the cleaning. With the EDA we did outliers treatment and remove unnecessary data.
We start from the data frame of the properties for rent

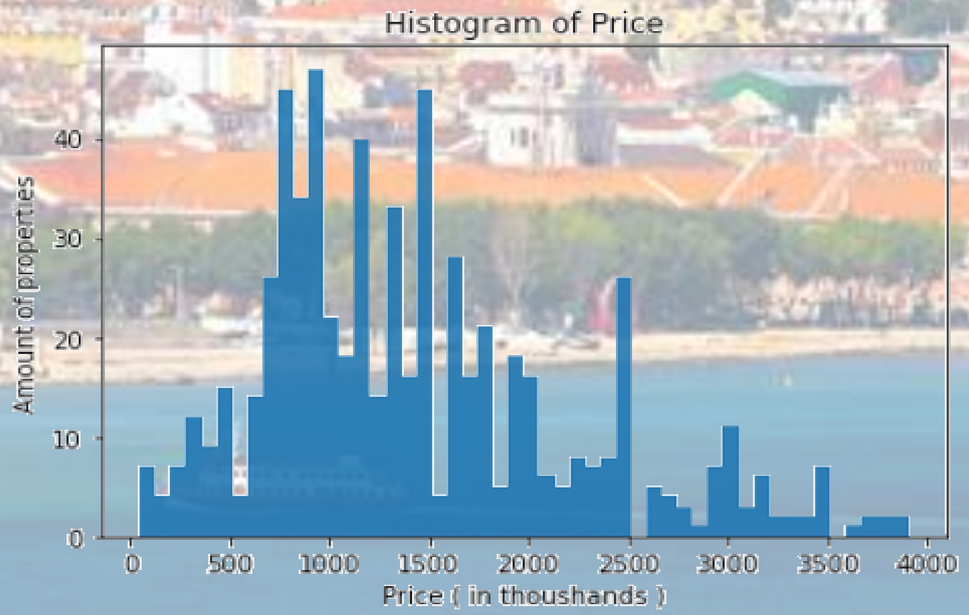


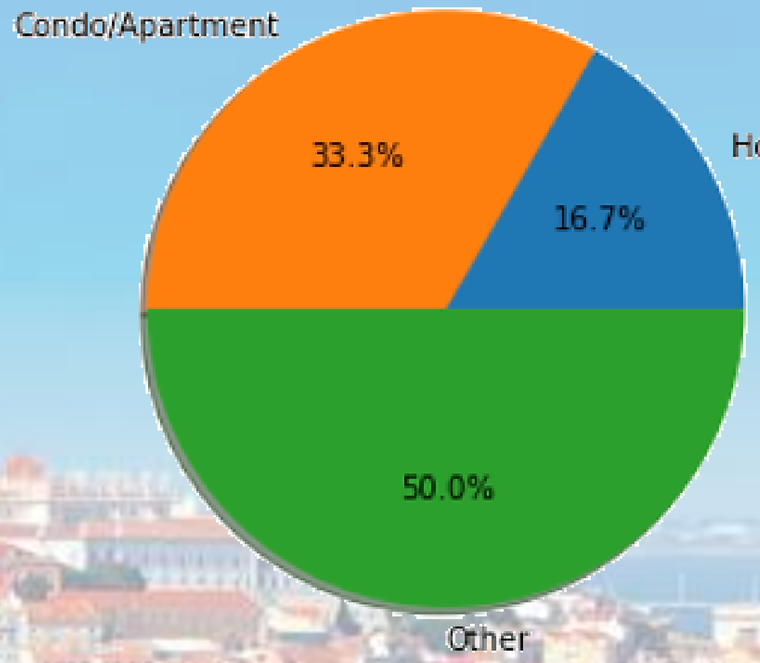
	Price	Prop_type	Size	Bath	Bed
count	740.000000	740.000000	740.000000	740.00000	740.000000
mean	2390.117568	2.150000	732.991892	1.55000	1.989189
std	9934.385044	0.541073	10768.645222	1.21066	1.575314
min	50.000000	1.000000	10.000000	0.00000	0.000000
25%	887.500000	2.000000	72.000000	1.00000	1.000000
50%	1397.500000	2.000000	100.000000	1.00000	2.000000
75%	2100.000000	2.000000	148.000000	2.00000	3.000000
max	250000.000000	3.000000	280000.000000	7.00000	21.000000

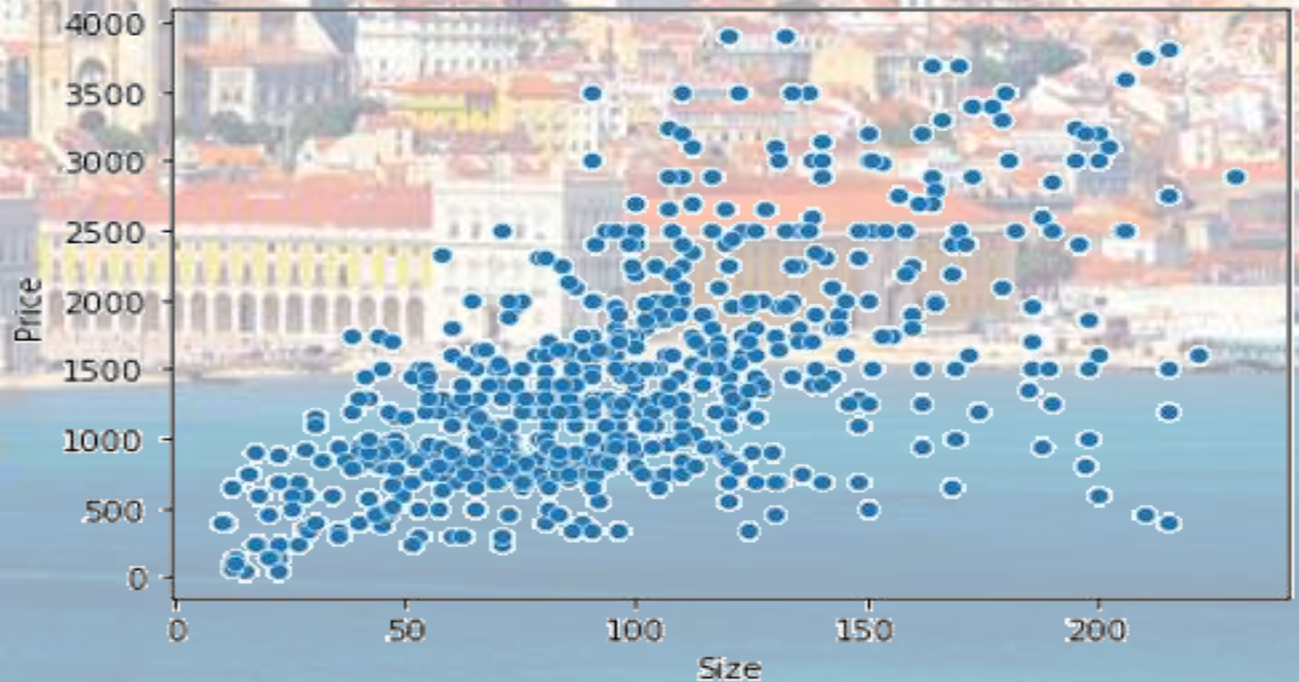
Visualization after the cleaning

We did some visualization with various of plots .

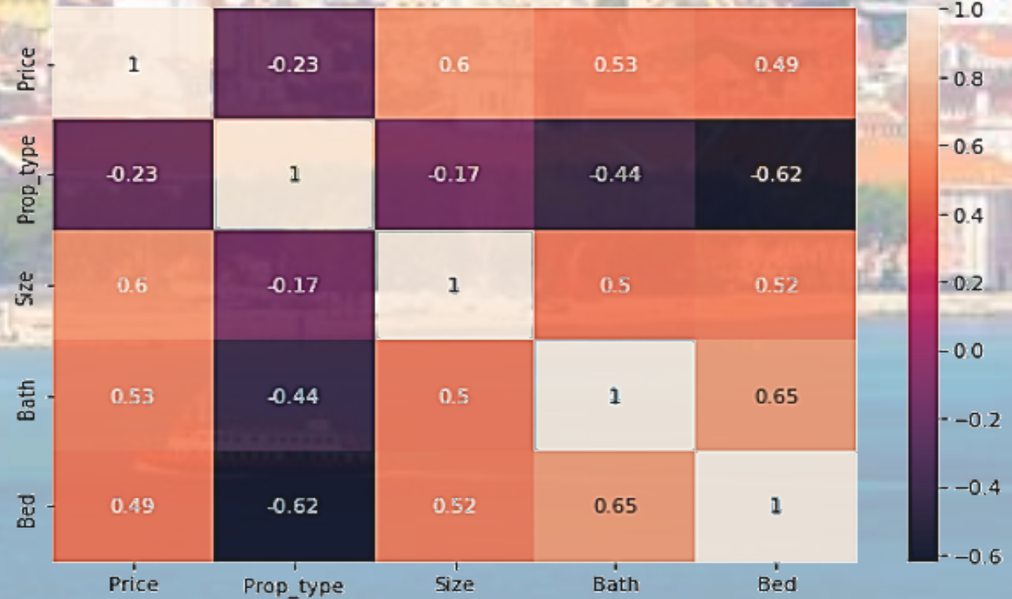


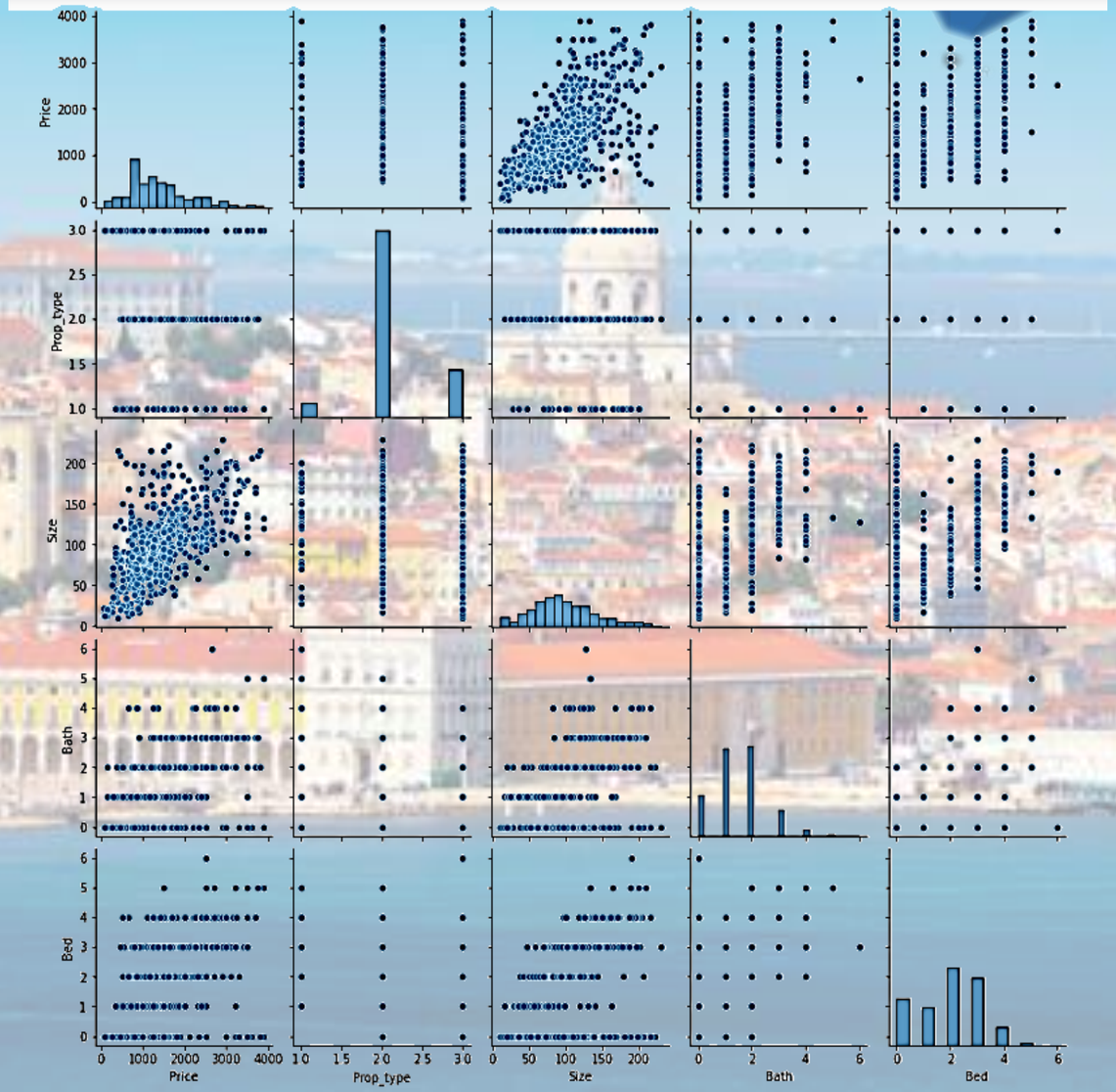


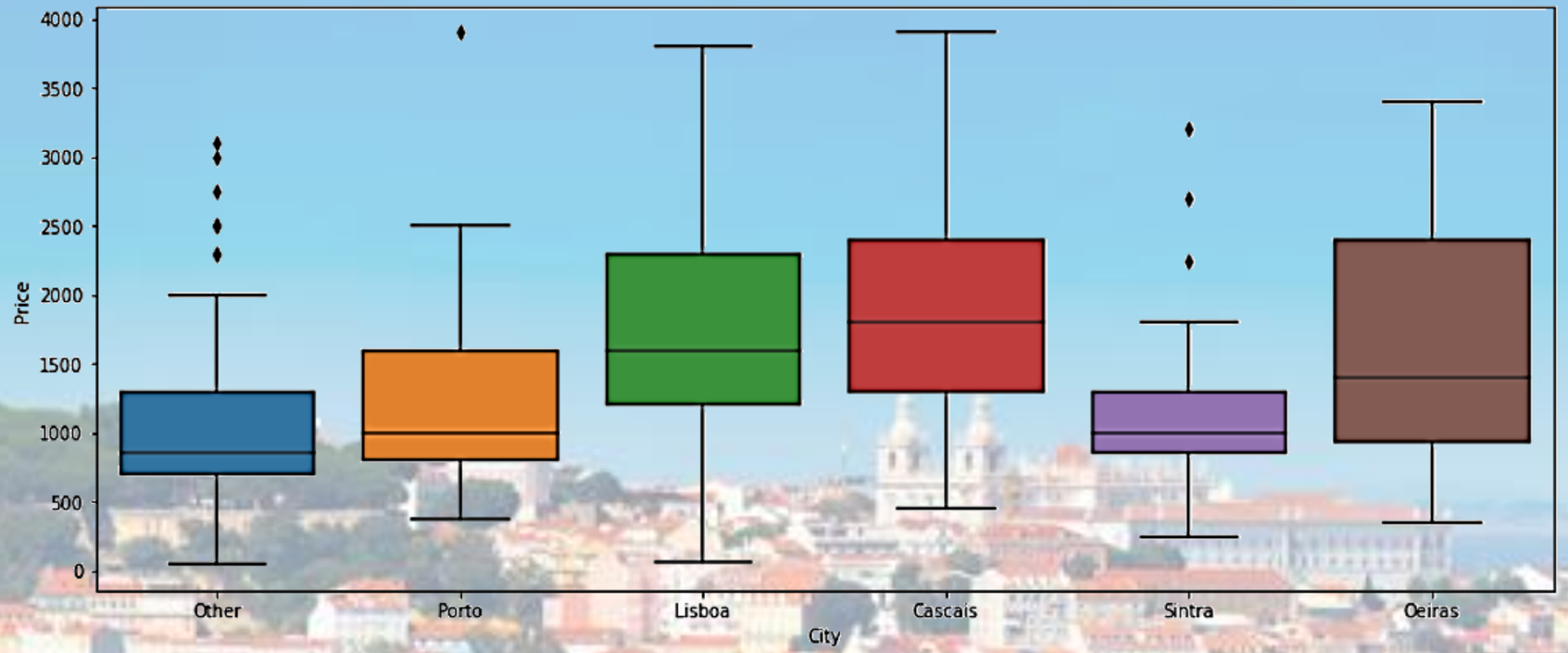


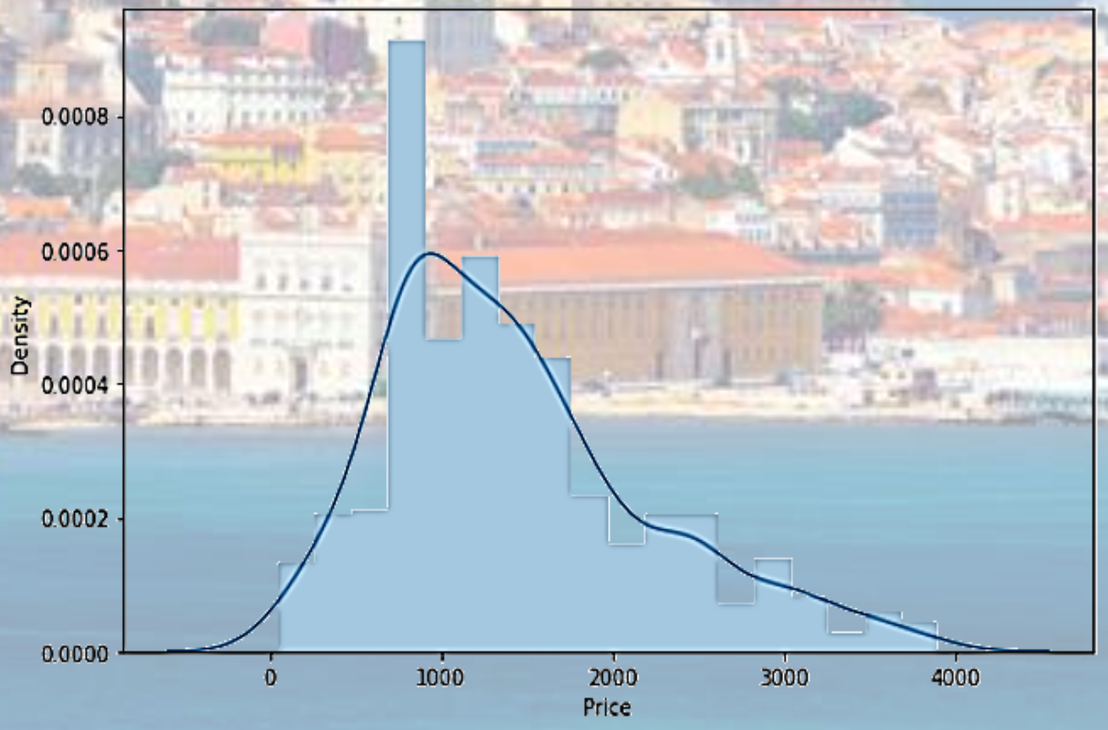



Price by Property Type
Price
4000
3500
3000
2500
2000
1500
1000
500
0
1
2
3
Prop_type


Price
Prop_type
Size
Bath
Bed
1
-0.23
0.6
0.53
0.49
-0.23
1
-0.17
-0.44
-0.62
0.6
-0.17
1
0.5
0.52
0.53
-0.44
0.5
1
0.65
0.49
-0.62
0.52
0.65
1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6


Price
Prop_type
Size
Bath
Bed


Price
4000
3500
3000
2500
2000
1500
1000
500
0
Other
Porto
Lisboa
Cascais
Sintra
Oeiras
City


Density
0.0008
0.0006
0.0004
0.0002
0.0000
0
1000
2000
3000
4000
Price

Step 4 Machine Learning

Supervised Learning

Supervised learning is a type of machine learning in which the computer is provided with labeled data and it is expected to generate a predictive model. This model is then used to make predictions about new data. Supervised learning can be used to classify data, predict outcomes, and identify patterns.

	coefficients
Size	484.158947
Bath	186.283048
Bed	24.404741
City_Lisboa	-16.698865
City_Oeiras	-201.125179
City_Other	-758.986941
City_Porto	-455.028673
City_Sintra	-678.834405

	coef	std err	t	P>|t|	[0.025	0.975]
const	1631.3350	88.140	18.509	0.000	1458.168	1804.502
Size	484.1589	32.615	14.845	0.000	420.081	548.237
Bath	186.2830	30.690	6.070	0.000	125.986	246.580
Bed	24.4047	49.194	0.496	0.620	-72.245	121.055
City_Lisboa	-16.6989	91.228	-0.183	0.855	-195.934	162.537
City_Oeiras	-201.1252	119.214	-1.687	0.092	-435.345	33.094
City_Other	-758.9869	93.834	-8.089	0.000	-943.342	-574.632
City_Porto	-455.0287	113.762	-4.000	0.000	-678.536	-231.522
City_Sintra	-678.8344	121.331	-5.595	0.000	-917.213	-440.456

The Process

we start to train and test when we scaled the data and after we did the same without scaling (without modeling on the Z-score)

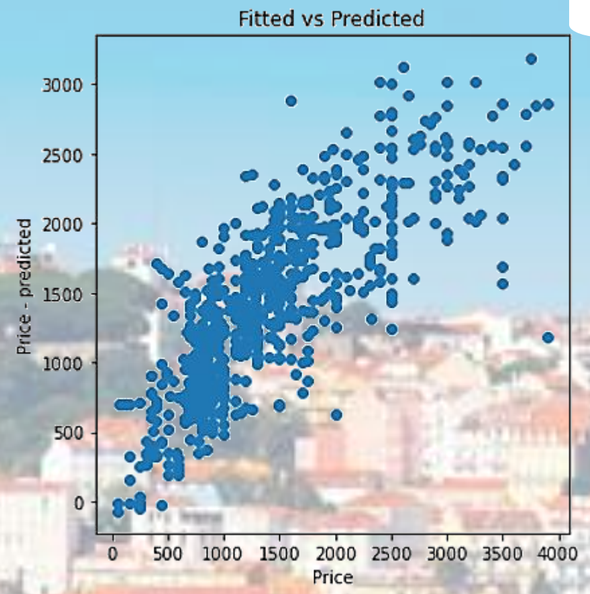
The Process

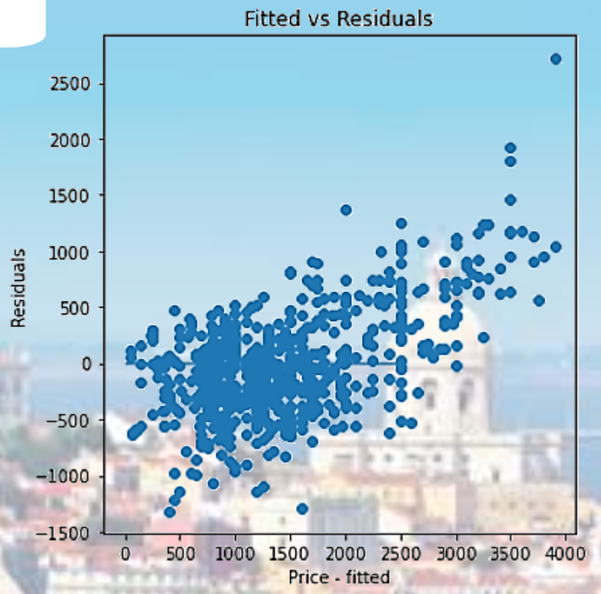
we use 20% of the data to test
and 80% to train
our X column was the Bed, Size and Bath
and our Y column was the Price

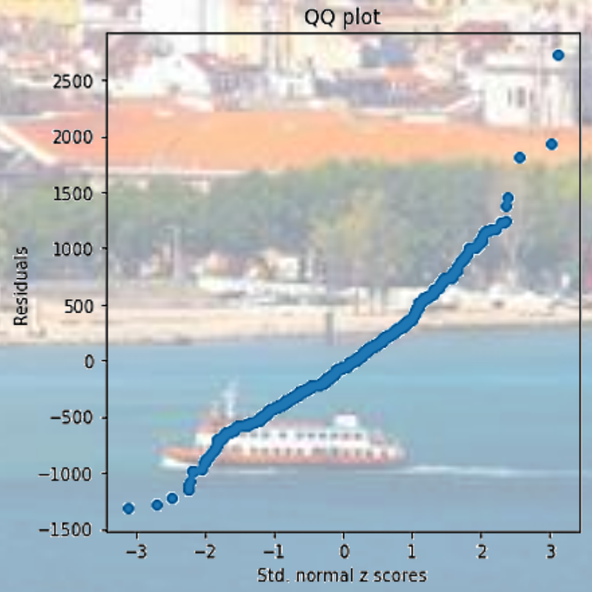
```
# strat train the model , splitting to 20/80

X = rent2.drop(columns = ['Price'])
y = rent2['Price']

# 20% of the data will be used for testing and 80% will be used for training.
X_train, X_test, y_train, y_test = train_test_split(X,y, test_size=0.2)

col = ['Size', 'Bath', 'Bed'] # used to scale data in the same range
scaler = RobustScaler().fit(X_train[col]) # RobustScaler() to scale the columns 'Size', 'Bath', and 'Bed' in the training set
X_train[col] = scaler.transform(X_train[col])
```

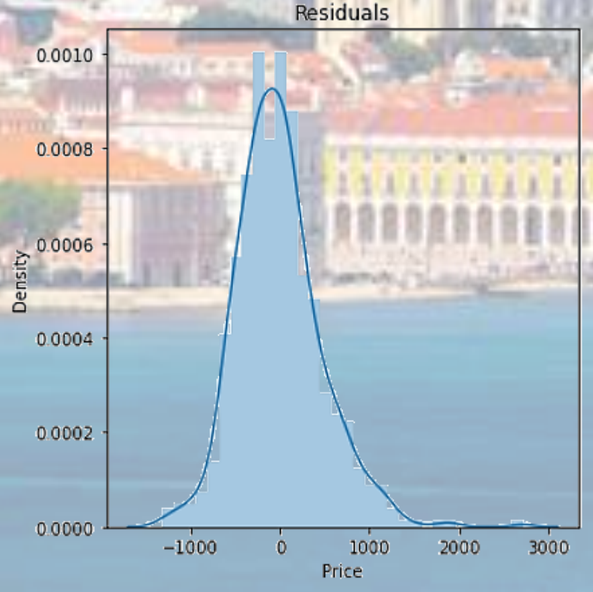
Linear Regression Equation

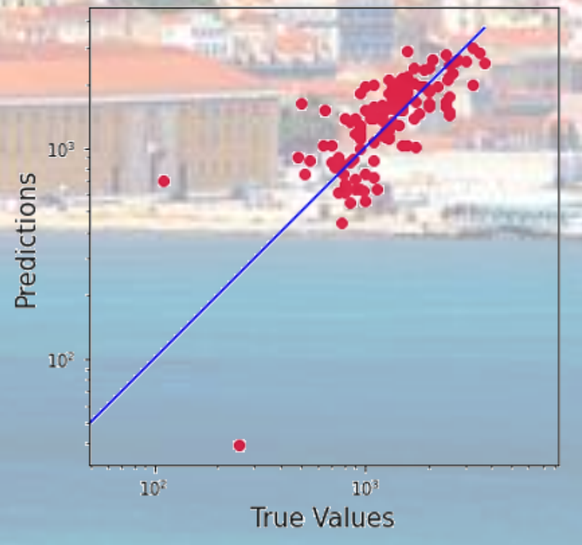
Some of the result











The Process

While we get our linear regression equation we load the buy data and fit the data to look like the same data as the rent .
After we train the Price data with the linear regression predict we got previously, and then we get the predicted rent column about the buy data

```
X = buydf.drop(columns = ['Price'])
buydf['predicted_rent'] = lm.predict(X)
```

```
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Price         10791 non-null  int64
 1   Size          10791 non-null  int64
 2   Bath          10791 non-null  int64
 3   Bed           10791 non-null  int64
 4   City_Cascais  10791 non-null  uint8
 5   City_Lisboa   10791 non-null  uint8
 6   City_Other    10791 non-null  uint8
 7   City_Porto    10791 non-null  uint8
 8   City_Sintra   10791 non-null  uint8
dtypes: int64(4), uint8(5)
```

	Price	Size	Bath	Bed	City_Cascais	City_Lisboa	City_Other	City_Porto	City_Sintra	predicted_rent
0	295000	140	2	3	0	0	0	1	0	1828.899446
1	22500	46	0	2	0	0	1	0	0	290.946897
2	485000	152	3	4	0	0	1	0	0	1838.769698
3	1250000	527	0	10	0	0	0	1	0	5222.931309
4	67000	75	1	2	0	0	1	0	0	723.019019

Start to working on the property to buy data

The last step
is to load the buy data after adding the predicted rent column.
we calculate with the Portugal formula to
calculate the monthly mortgage and other monthly expense , such as insurance or property tax and make one column that shows , how much profit the property owner stay with at the end of the month.

Price	Size	Bath	Bed	City_Cascais	City_Lisboa	City_Other	City_Porto	City_Sintra	predicted_rent	Monthly_Mortgage	Monthly_Other	Monthly_Profit
295000	140	2	3	0	0	0	1	0	1761.992029	1408	640.150	-286.157971
22500	46	0	2	0	0	1	0	0	387.169896	107	48.825	231.344896
485000	152	3	4	0	0	1	0	0	1831.693994	2315	1052.450	-1535.756006
1250000	527	0	10	0	0	0	1	0	5211.002405	5967	2712.500	-3468.497595
67000	75	1	2	0	0	1	0	0	766.781801	319	145.390	302.391801
...	...	...	...	...	...	...	...	...	...	...	...	...
135000	22	1	0	0	0	0	0	0	976.560974	644	292.950	39.610974
130000	18	1	0	0	0	0	0	0	943.395058	620	282.100	41.295058
139900	110	2	3	0	0	1	0	0	1270.217720	667	303.583	299.634720
799000	720	12	7	0	0	1	0	0	8015.910647	3814	1733.830	2468.080647
170000	113	2	2	0	0	0	0	0	2018.394865	811	368.900	838.494865

10791 rows × 13 columns

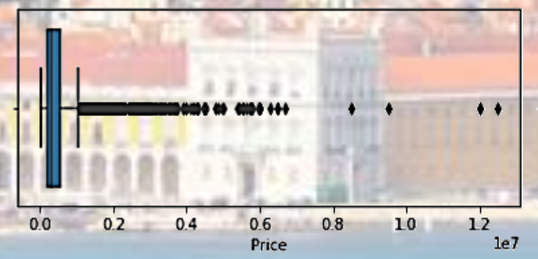
	Price	Size	Bath	Bed	City_Cascais	City_Lisboa	City_Other	City_Porto	City_Sintra	Monthly_Profit
0	295000	140	2	3	0	0	0	1	0	-286.157971
1	22500	46	0	2	0	0	1	0	0	231.344896
2	485000	152	3	4	0	0	1	0	0	-1535.756006
3	1250000	527	0	10	0	0	0	1	0	-3468.497595
4	67000	75	1	2	0	0	1	0	0	302.391801

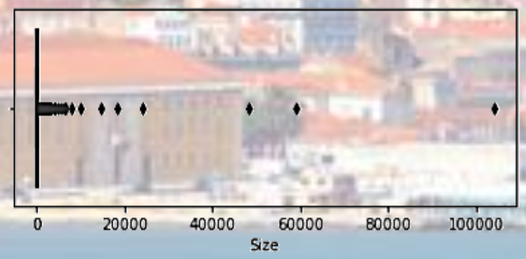
Eda And Data Handling

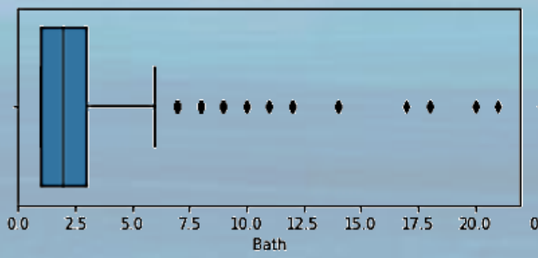
We did again
Eda and outliers treatment
on the new data
to make a cleanier data to
work on
with the machine learning
modules.

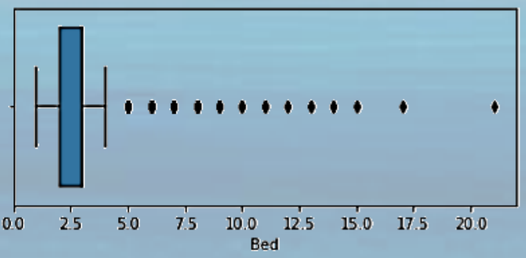
Price	Size	Bath	Bed	City_Cascais	City_Lisboa	City_Other	City_Porto	City_Sintra	predicted_rent	Monthly_Mortgage	Monthly_Other	Monthly_Profit
295000	140	2	3	0	0	0	1	0	1761.992029	1408	640.150	-286.157971
22500	46	0	2	0	0	1	0	0	387.169896	107	48.825	231.344896
485000	152	3	4	0	0	1	0	0	1831.693994	2315	1052.450	-1535.756006
1250000	527	0	10	0	0	0	1	0	5211.002405	5967	2712.500	-3468.497595
67000	75	1	2	0	0	1	0	0	766.781801	319	145.390	302.391801
...	...	...	...	...	...	...	...	...	...	...	...	...
135000	22	1	0	0	0	0	0	0	976.560974	644	292.950	39.610974
130000	18	1	0	0	0	0	0	0	943.395058	620	282.100	41.295058
139900	110	2	3	0	0	1	0	0	1270.217720	667	303.583	299.634720
799000	720	12	7	0	0	1	0	0	8015.910647	3814	1733.830	2468.080647
170000	113	2	2	0	0	0	0	0	2018.394865	811	368.900	838.494865

10791 rows × 13 columns



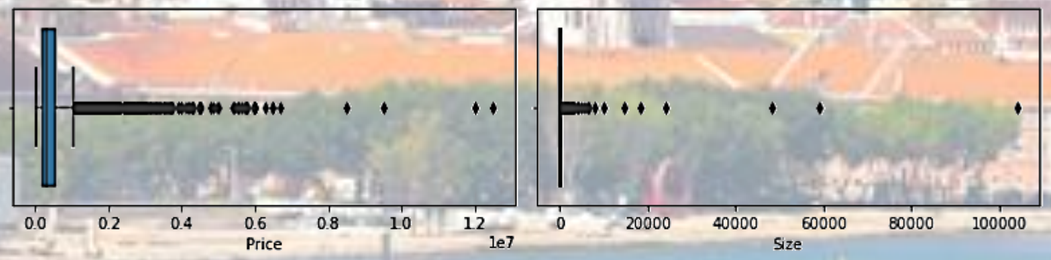


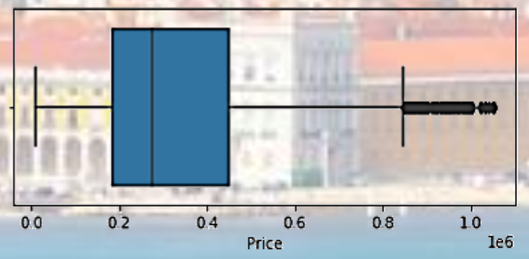


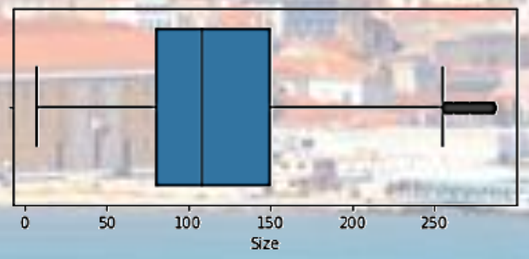


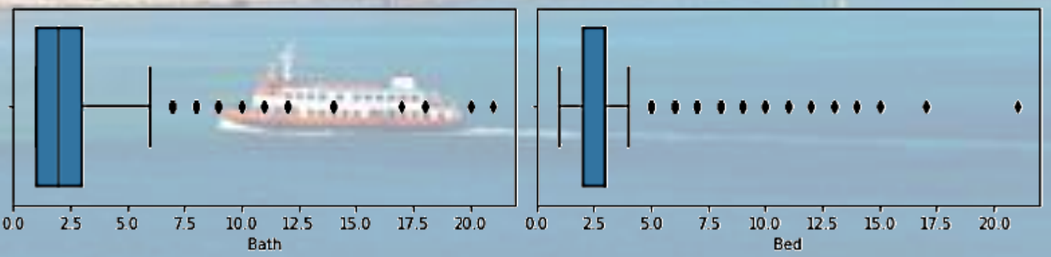
Eda And Data Handling

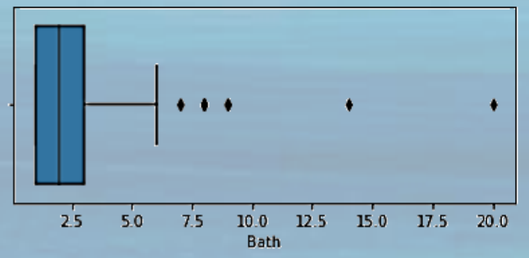

```
# outlier treatment for Size
plt.boxplot(df2.Size)
Q1 = df2.Size.quantile(0.25)
Q3 = df2.Size.quantile(0.75)
IQR = Q3 - Q1
df2 = df2[(df2.Size >= Q1 - 1.5*IQR) & (df2.Size <= Q3 + 1.5*IQR)]
```

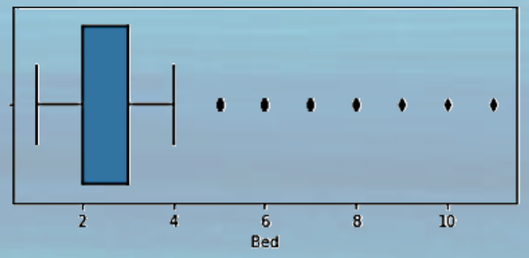






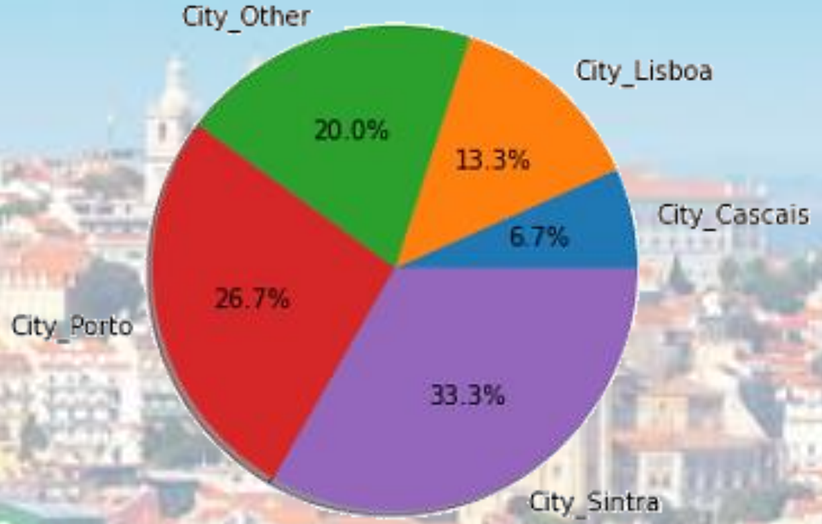


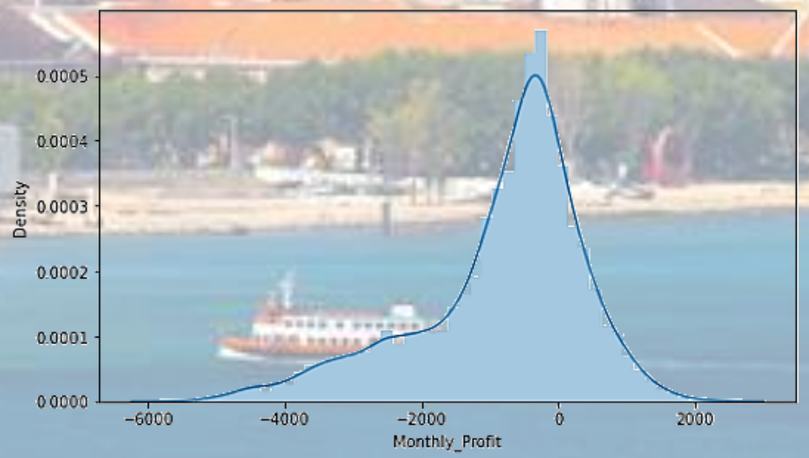


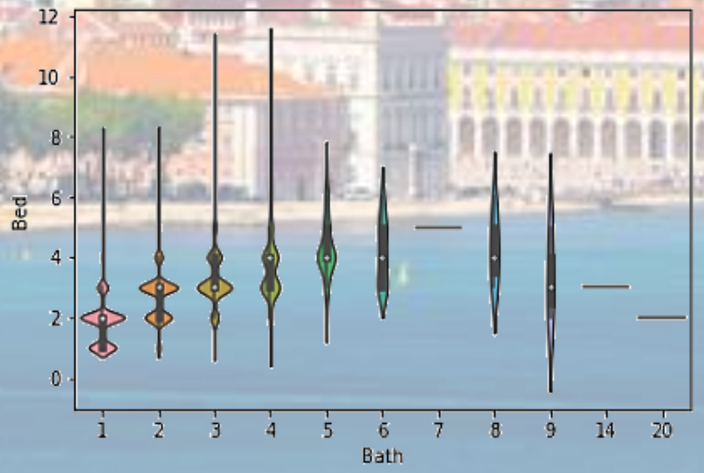


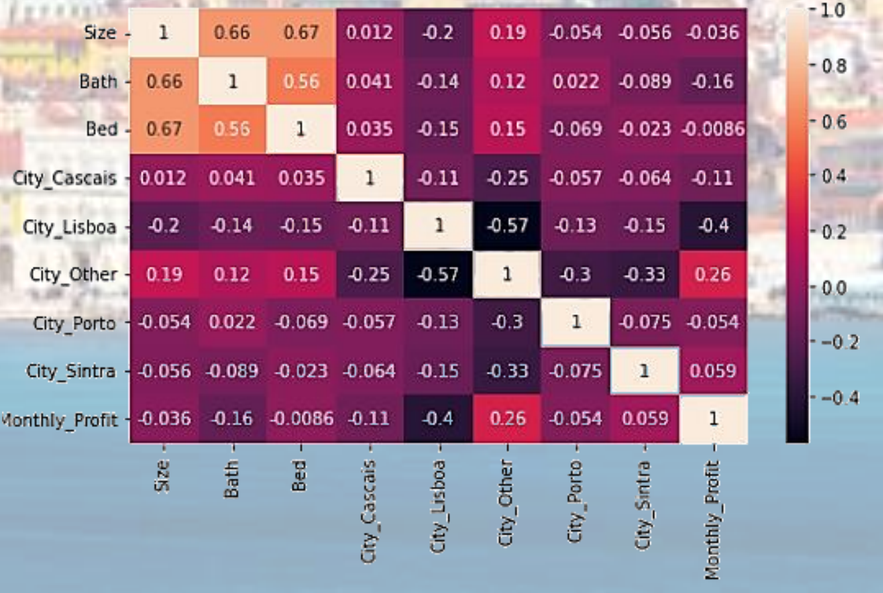
EDA

Some of the visualization we get after cleaning the data





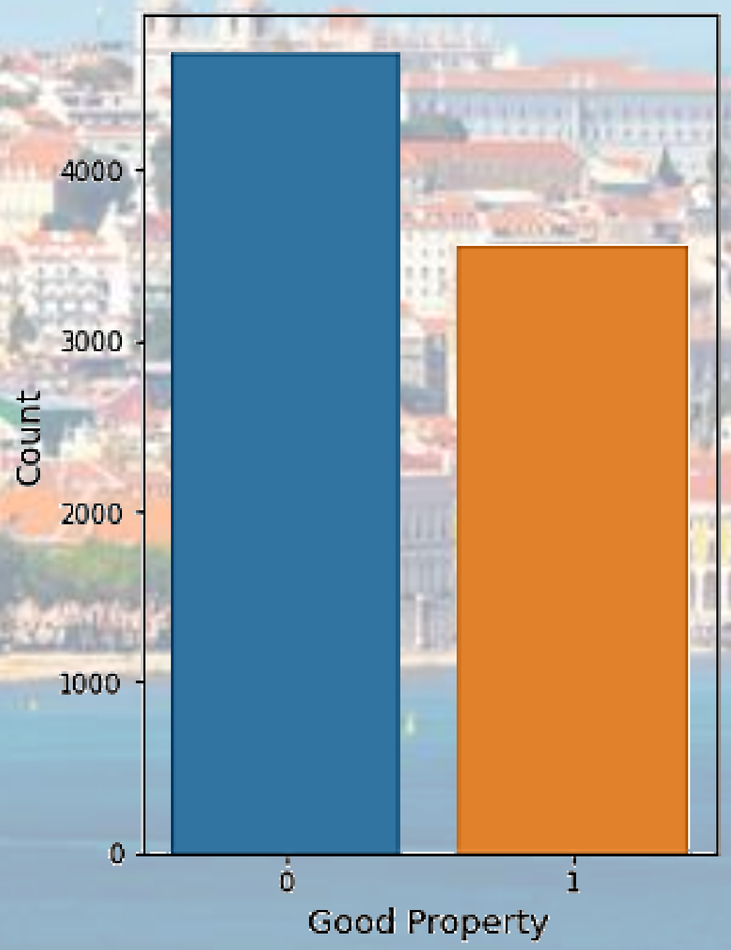




Good Property

We decide that good property
is a property that his owner doesn't lost more than 400 euro per month

```
not good: 56.99999999999999%
good: 43.0%
```



‘0’ = Not good property
‘1’ = Good Property

Step 5 Interpreting Data

Buy Machine Learning

in the last step
Everything we've learned so far from the machine learning we've done on rental properties
We operated on properties to buy, and finally presented the data, results and conclusions

Supervised Learning

Supervised Learning : Again , we did a machine learning, But now on the buy data frame , and the target column to predict and learning the machine on is the 'good prop' column.

```
accuracy_score(y_test, lr_pred)
```

```
0.6877278250303767
```

```
precision, recall, fscore, support = score(y_test, lr_pred, average='macro')
print(f'Precision : {precision}')
print(f'Recall    : {recall}')
print(f'F-score   : {fscore}')
```

```
Precision : 0.6895781637717122
Recall    : 0.6952276073544981
F-score   : 0.6859454924757047
```

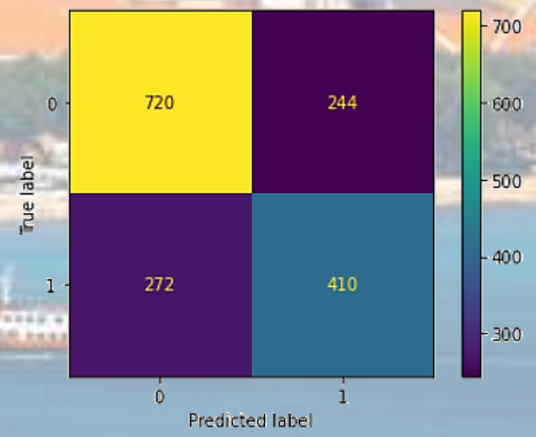
	Coefs	Names	Odds Coefs
0	0.115042	Size	1.121921
1	-0.441402	Bath	0.643134
2	0.157835	Bed	1.170973
3	-3.051520	City_Cascais	0.047287
4	-3.979764	City_Lisboa	0.018690
5	-1.853800	City_Other	0.156641
6	-3.305936	City_Porto	0.036665
7	-1.796576	City_Sintra	0.165866
8	2.462683	intercept	11.736263

```
X_train, X_test, y_train, y_test = train_test_split(X,y, test_size=0.2)

z = StandardScaler()
z.fit(X_train[['Size']])  # size is the only continuous var

X_train['Size'] = z.transform(X_train[['Size']]) # transform - to subtract the mean and divide by the standard deviation in
X_test['Size'] = z.transform(X_test[['Size']]) # order to get the standardized values

# להחסיר את הממוצע ולחלק בסטיית התקן על מנת לקבל את הערכים המתוקנים
```

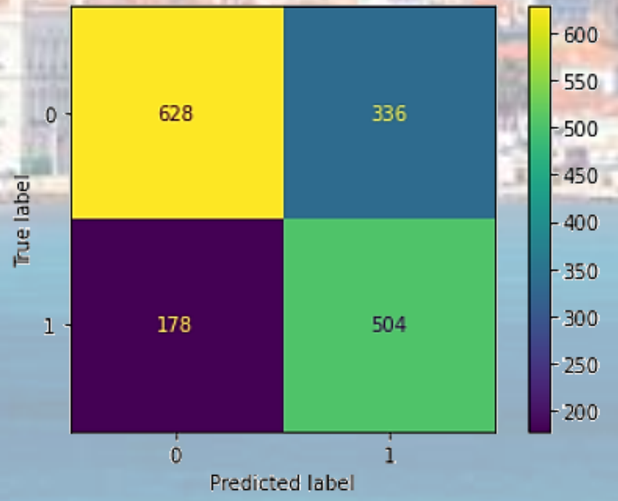
Result

The accuracy results
Visualizations for illustration

Decision Tree

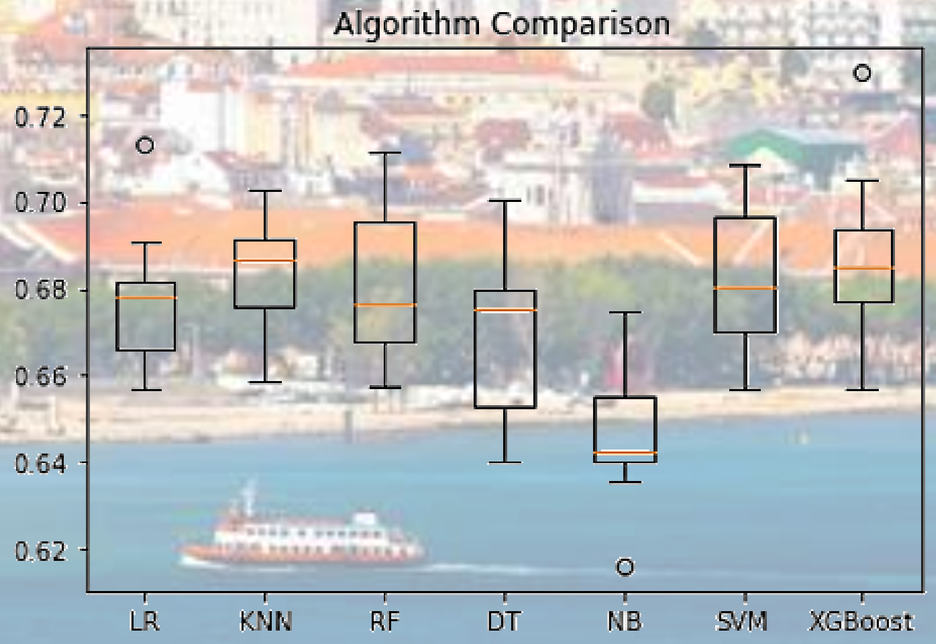


Linear Regression Accuracy



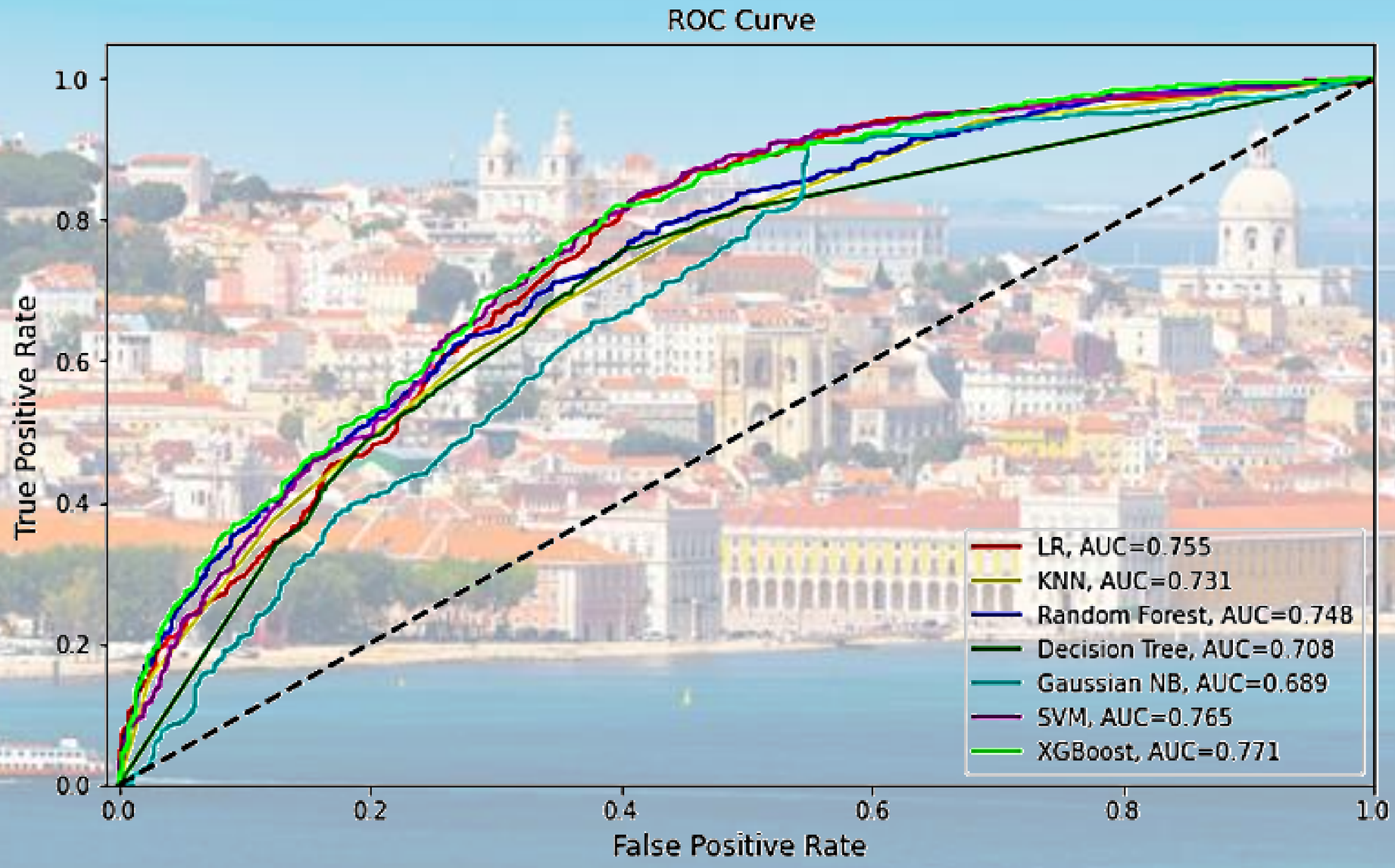
Model Comparison

	Model	Train Accuracy	Test Accuracy	Precision	Recall	F1
0	LR	0.678	0.683	0.682	0.684	0.682
1	KNN	0.688	0.671	0.666	0.665	0.665
2	Random Forest	0.675	0.683	0.678	0.678	0.678
3	Dec Tree	0.673	0.665	0.659	0.653	0.654
4	Gaussian NB	0.646	0.651	0.713	0.679	0.644
5	SVM	0.685	0.697	0.697	0.700	0.696
6	XGBoost	0.684	0.688	0.683	0.681	0.682




```
LR: 0.677552 (0.015389)
KNN: 0.683173 (0.012226)
RF: 0.680747 (0.016745)
DT: 0.670265 (0.019072)
NB: 0.645963 (0.015128)
SVM: 0.682868 (0.016269)
XGBoost: 0.685756 (0.020575)
```

Roc Curve Result



Conclusion

As you can see
the best score we get is
from the
SVM model.

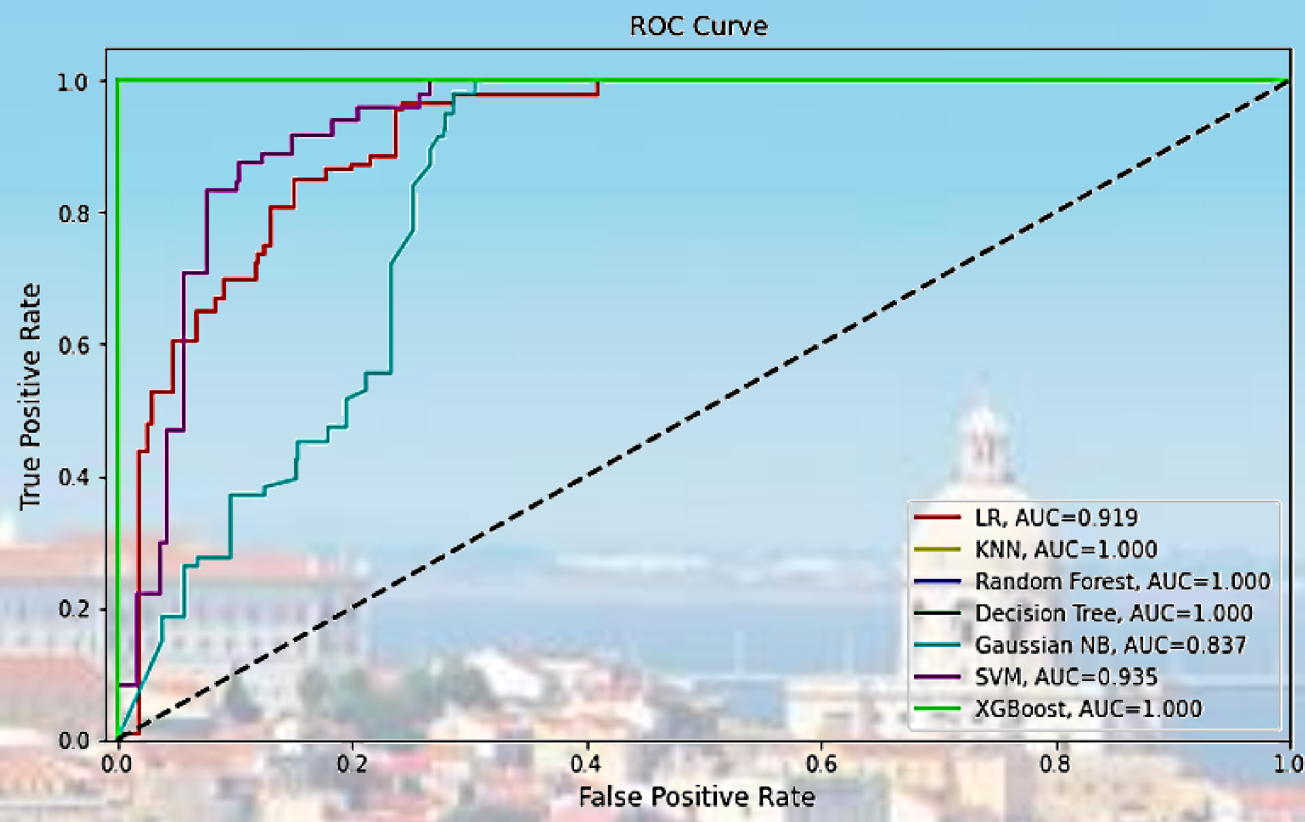
SVM	0.685	0.697	0.697	0.700	0.696

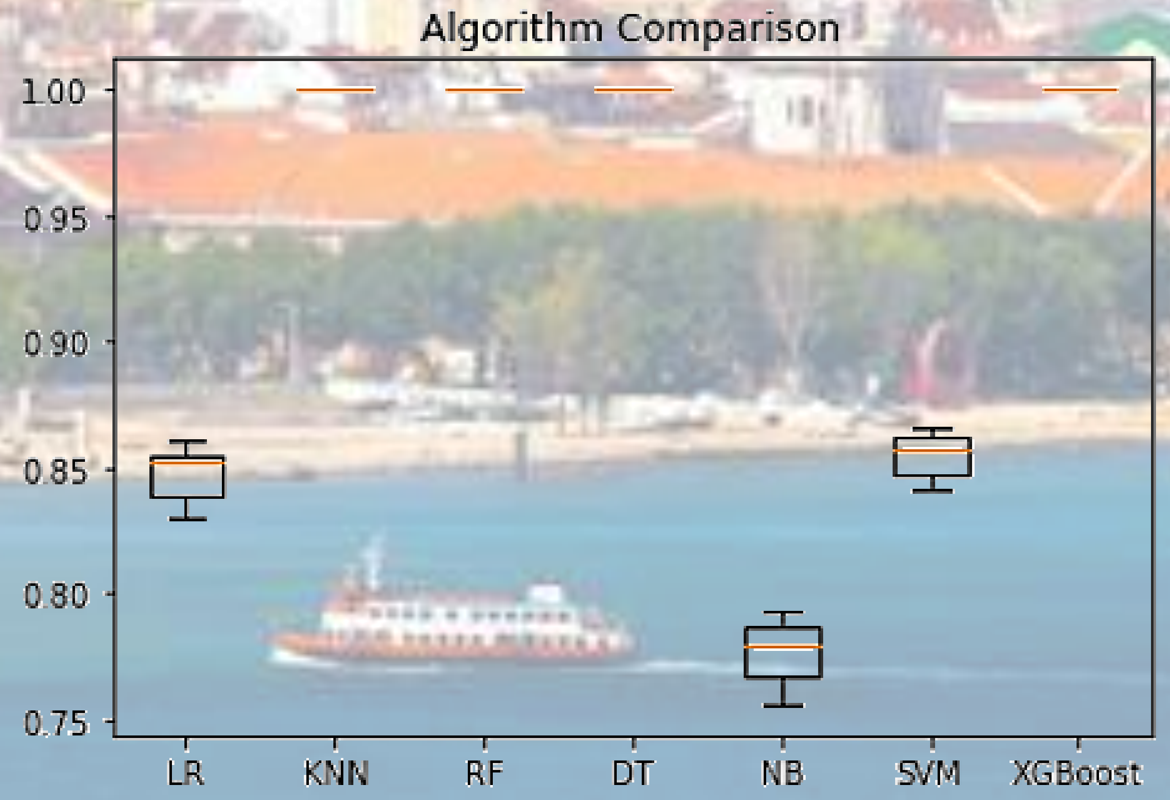
Our conclusion is that it is indeed possible to identify a good investment property in Portugal.
We have indeed seen that machine learning models achieve quite good results.
It was also said that to improve the accuracy more information is needed which cannot be obtained
Apartment sale and rental site such as the construction price, the quality of the neighborhood, average electricity and water prices for the property
The specific one being tested, and more

overfitting and bias

Our Failures

```
LR: 0.847653 (0.009985)
KNN: 1.000000 (0.000000)
RF: 1.000000 (0.000000)
DT: 1.000000 (0.000000)
NB: 0.776267 (0.012458)
SVM: 0.854606 (0.008567)
XGBoost: 1.000000 (0.000000)
```

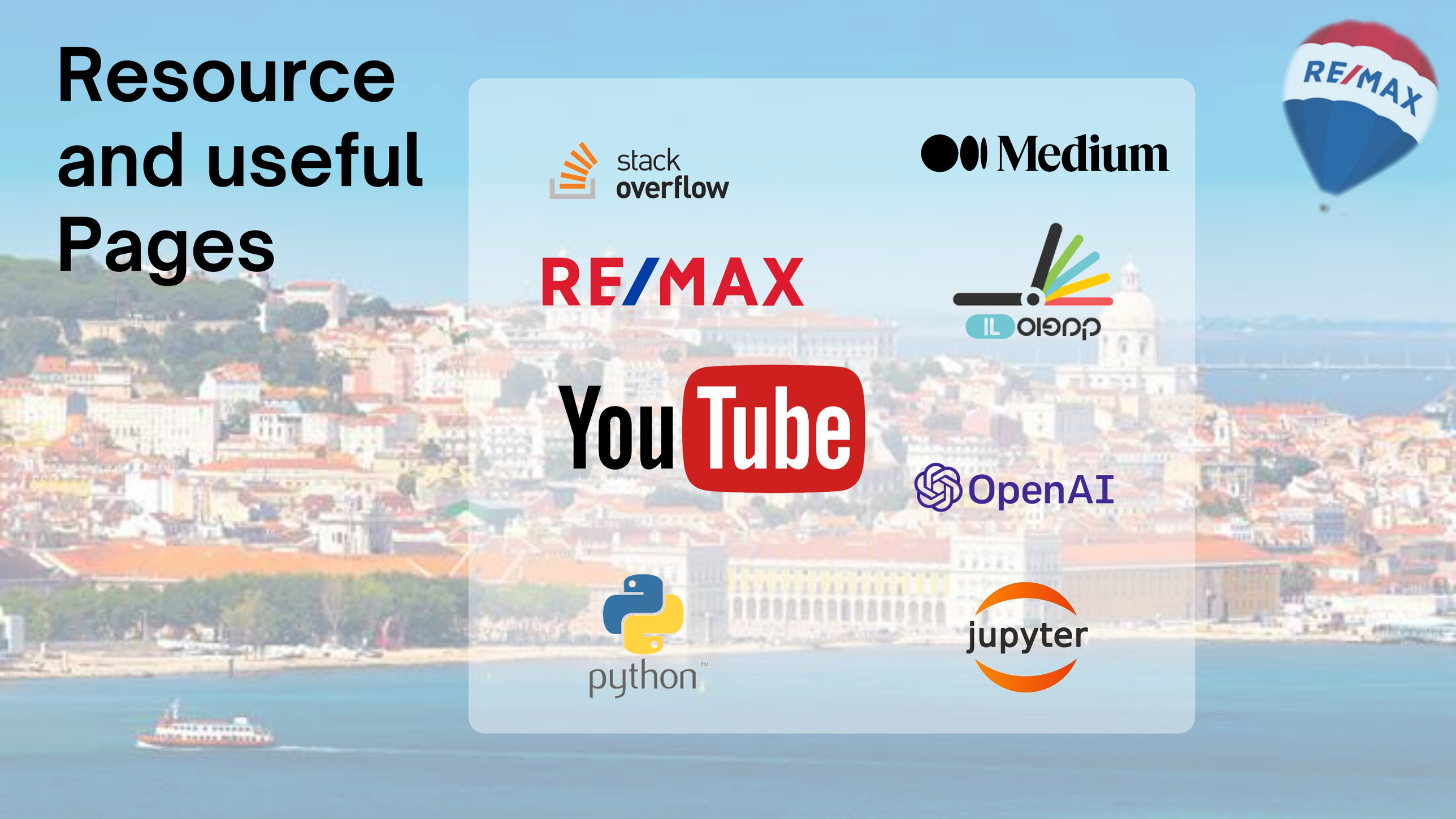




Scrapping Problems

3	2	100 m2	Condo/Apartment	Braga -‚ Braga, Braga (Maximinos, S© e Cividade)	198 500 ג‚¬
3	2	100 m2	Condo/Apartment	Braga -‚ Braga, Braga (Maximinos, S© e Cividade)	198 500 ג‚¬
3	2	100 m2	Condo/Apartment	Braga -‚ Braga, Braga (Maximinos, S© e Cividade)	198 500 ג‚¬
3	2	100 m2	Condo/Apartment	Braga -‚ Braga, Braga (Maximinos, S© e Cividade)	198 500 ג‚¬
3	2	100 m2	Condo/Apartment	Braga -‚ Braga, Braga (Maximinos, S© e Cividade)	198 500 ג‚¬
3	2	100 m2	Condo/Apartment	Braga -‚ Braga, Braga (Maximinos, S© e Cividade)	198 500 ג‚¬
3	2	100 m2	Condo/Apartment	Braga -‚ Braga, Braga (Maximinos, S© e Cividade)	198 500 ג‚¬
3	2	100 m2	Condo/Apartment	Braga -‚ Braga, Braga (Maximinos, S© e Cividade)	198 500 ג‚¬
3	2	100 m2	Condo/Apartment	Braga -‚ Braga, Braga (Maximinos, S© e Cividade)	198 500 ג‚¬
3	2	100 m2	Condo/Apartment	Braga -‚ Braga, Braga (Maximinos, S© e Cividade)	198 500 ג‚¬
3	2	100 m2	Condo/Apartment	Braga -‚ Braga, Braga (Maximinos, S© e Cividade)	198 500 ג‚¬
3	2	100 m2	Condo/Apartment	Braga -‚ Braga, Braga (Maximinos, S© e Cividade)	198 500 ג‚¬
3	2	100 m2	Condo/Apartment	Braga -‚ Braga, Braga (Maximinos, S© e Cividade)	198 500 ג‚¬
3	2	100 m2	Condo/Apartment	Braga -‚ Braga, Braga (Maximinos, S© e Cividade)	198 500 ג‚¬
3	2	100 m2	Condo/Apartment	Braga -‚ Braga, Braga (Maximinos, S© e Cividade)	198 500 ג‚¬
3	2	100 m2	Condo/Apartment	Braga -‚ Braga, Braga (Maximinos, S© e Cividade)	198 500 ג‚¬
3	2	100 m2	Condo/Apartment	Braga -‚ Braga, Braga (Maximinos, S© e Cividade)	198 500 ג‚¬
3	2	100 m2	Condo/Apartment	Braga -‚ Braga, Braga (Maximinos, S© e Cividade)	198 500 ג‚¬
3	2	100 m2	Condo/Apartment	Braga -‚ Braga, Braga (Maximinos, S© e Cividade)	198 500 ג‚¬
3	2	100 m2	Condo/Apartment	Braga -‚ Braga, Braga (Maximinos, S© e Cividade)	198 500 ג‚¬
3	2	100 m2	Condo/Apartment	Braga -‚ Braga, Braga (Maximinos, S© e Cividade)	198 500 ג‚¬
3	2	100 m2	Condo/Apartment	Braga -‚ Braga, Braga (Maximinos, S© e Cividade)	198 500 ג‚¬
3	2	100 m2	Condo/Apartment	Braga -‚ Braga, Braga (Maximinos, S© e Cividade)	198 500 ג‚¬
3	2	100 m2	Condo/Apartment	Braga -‚ Braga, Braga (Maximinos, S© e Cividade)	198 500 ג‚¬
3	2	100 m2	Condo/Apartment	Braga -‚ Braga, Braga (Maximinos, S© e Cividade)	198 500 ג‚¬
3	2	100 m2	Condo/Apartment	Braga -‚ Braga, Braga (Maximinos, S© e Cividade)	198 500 ג‚¬
3	2	100 m2	Condo/Apartment	Braga -‚ Braga, Braga (Maximinos, S© e Cividade)	198 500 ג‚¬
3	2	100 m2	Condo/Apartment	Braga -‚ Braga, Braga (Maximinos, S© e Cividade)	198 500 ג‚¬
3	2	100 m2	Condo/Apartment	Braga -‚ Braga, Braga (Maximinos, S© e Cividade)	198 500 ג‚¬
3	2	100 m2	Condo/Apartment	Braga -‚ Braga, Braga (Maximinos, S© e Cividade)	198 500 ג‚¬
3	2	100 m2	Condo/Apartment	Braga -‚ Braga, Braga (Maximinos, S© e Cividade)	198 500 ג‚¬

Resource and useful Pages



Thanks for listening!

Roy And Yarin