

Boston Houses Pricing

Predictive Model using Linear Regression
Machine Learning I

Professor: Concepción Díaz

Camilla Perotti | Hector Marmol | Lucia Sarobe | Tomás Silva | Vedant Agrawal

01

Problem Statement

Slide 3

02

Exploratory Data Analysis

Slide 4 – 5

03

Data Model

Slide 6 – 7

04

Model Deployment (*Streamlit* App)

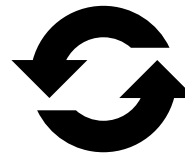
Slide 8

Boston residents' face multiple challenges when navigating through housing options

Some of the features in the dataset:



- **N° rooms** – tailored to meet your specific requirements
- **Crime rate per capita** – providing insights into neighborhood safety
- **Pollution levels** – helping you evaluate the environmental quality
- **Pupil-teacher ratio** – supporting decisions for families with children
- **Tax rate and industrial proportion** – offering economic perspectives



This **user-friendly platform simplifies the journey** of house hunting by allowing users to **input their preferences**, visualize data trends, and instantly **receive a price estimate** tailored to their dream home.

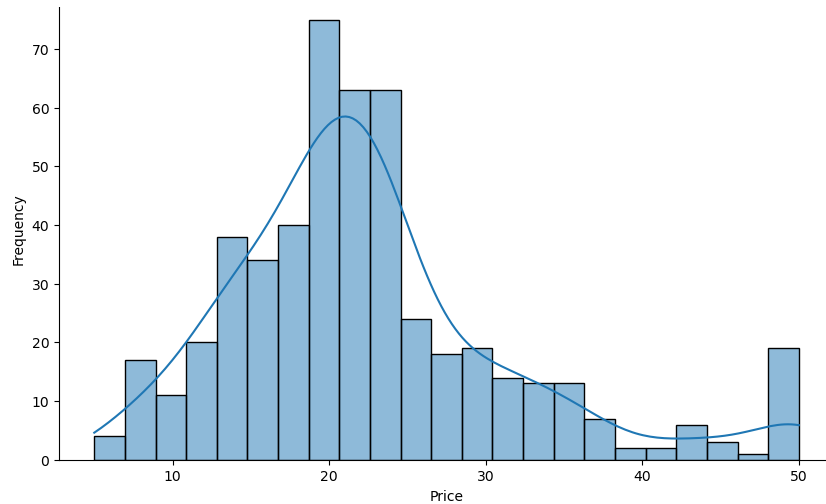
Boston Housing Predictor → *Value Proposition*



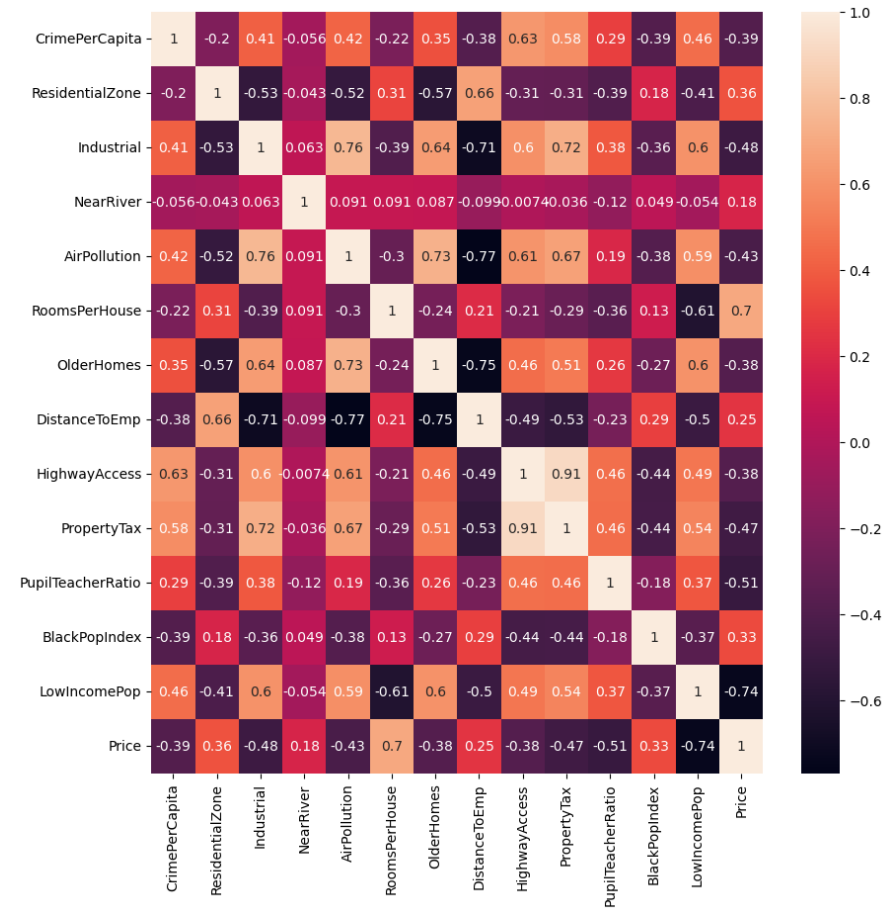
Transparency, simplicity and confidence to the housing market, empowering residents to find homes that meet their needs and align with their budgets

All the features are seemingly correlated with the dependent variable

Distribution of House Pricing



Correlation Heatmap



Key Takeaways

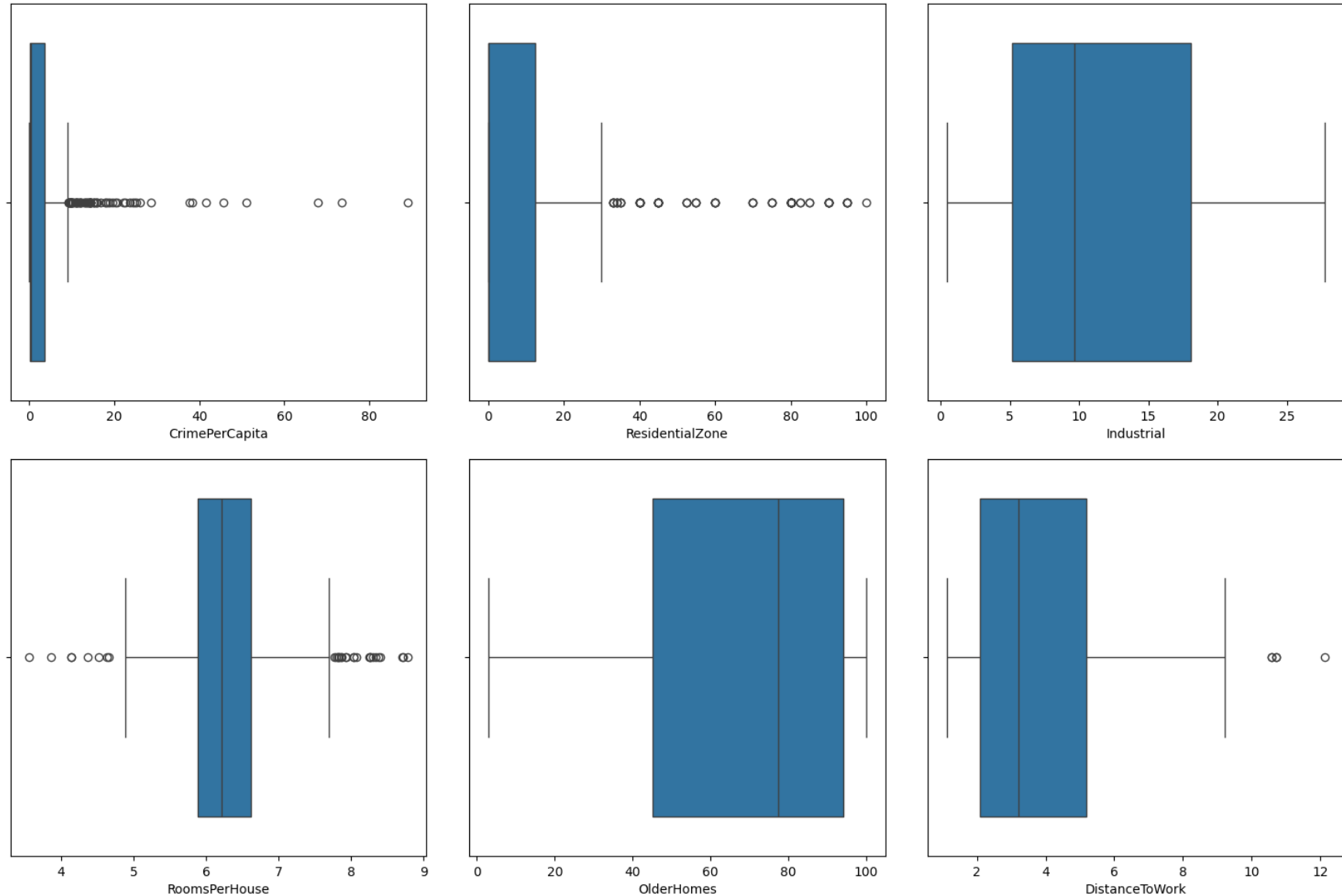
Distribution of House Pricing:

- Most houses are priced between **14K and 26K**, although there is a considerable amount in the 50 thousands' (**right-heavy tail**)

Correlation Heatmap:

- We will use the features with a **correlation above 0.4** (absolute value)
- PropertyTax** and **HighwayAccess** are highly correlated (**0.91**) between each other, which would possibly lead to multicollinearity, if they were variable of the same type

No outliers were detected



Key Takeaways

Boxplots for Outlier Detection:

- Example: **6 Charts**
- **Machine Learning's** outliers are considered differently than **Statistics**
- Charts that have plenty of circles outside the IQR means that the dataset has a **wide distribution**, hence are not considered outliers
- **DistanceToWork** required further investigation but lead to be considered as no outlier
- **No outliers detected**

3 models were trained: **df** – every feature contemplated; **df1** – correlation threshold of 0.4; **df2** – correlation threshold of 0.2

Macro Process for a Linear Regression Model

Understand the problem and have a clear value proposition



Dependent variable:
Price
Test proportion:
10%



Scale the features:
Fit and transform for training
Transform for test



Elastic Net model for regularization
Cross Validation through Grid Search

df | 14 features

df1 | 8 features

df2 | 12 features

Model I (df) every feature

Parameter Grid Definition – 210 fits

Best hyperparameters: alpha = 0.1; l1_ratio = 0.7



Model II (df1) >0.4 correlation

Parameter Grid Definition – 210 fits

Best hyperparameters : alpha = 0.1;
l1_ratio = 0.5



Worse results

Model III (df2) >0.2 correlation

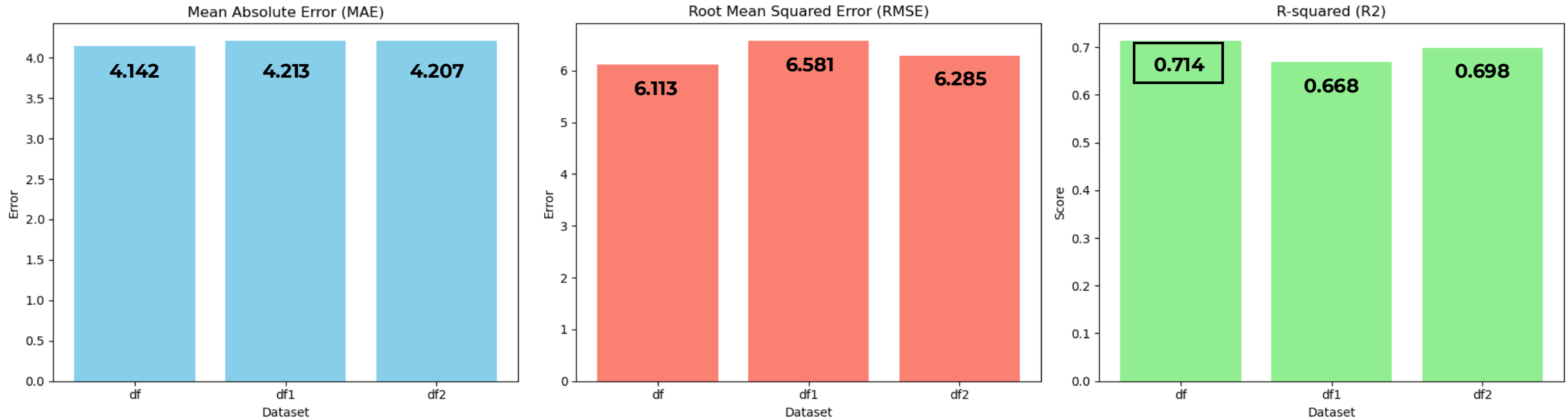
Parameter Grid Definition – 210 fits

Best hyperparameters : alpha = 0.1;
l1_ratio = 0.9



Worse results than df,
but better than df1

The 'best' model is the one with all the features with a R-square of 0.714. We should use a more complex model (Neural Networks e.g.) to improve results



- **The results indicate that DF performs best across all metrics** → including all features yields better predictive accuracy and explains more variance in the target variable.
- While **DF1** simplifies the model by using fewer features, it **sacrifices performance**, as shown by its higher RMSE and lower R-squared.
- **DF2 balances feature reduction and performance** but still underperforms compared to DF.
- Conclusion: **Linear regression not suitable as ML model**, opt to use a more complex model, e.g. neural network

The following app in *Streamlit* predicts the house price (in Boston) based on user preferences

ie

THE LOUVER CITY

Boston

1020

CITY OF DREAMS

Welcome to Boston!

The Mayor's Office of Housing is here to help you on your journey to find your home 🏡

House Price Predictor: How does it work?

We know finding a house at the right price in a new city can be overwhelming, therefore we have created this page for you to input your criterias and we will provide you with an price estimate.

If you would like to receive additional tips & tricks on how to find your dream home, we can send you a brochure directly to your inbox.

Enter your e-mail here:

Submit your info

Percentage of residential zones in your desired area

1

0

100

The higher the percentage is the lower is the house density and the more space you may expect in that area. You selected a percentage of 1

Do you wish your house to be closes to the Charles River?

1

1 is yes and 0 is no. Your selection is: 1

Percentage of older homes in your desired area.

1

0

100

Old houses are houses built before 1940. You selected a ratio of 1

Higher importance means choosing an area close to highways. Lower importance means living further away from highways. You selected an importance level of 4

Your criteria for your dream home

	CrimePerCapita	ResidentialZone	Industrial	NearRiver	AirPollution	RoomsPerHouse	OlderHomes
0	8.01	40	19.71	1	0.01	1	79

Calculate House Price

🏡 Estimated House Price

\$32,467

Application interface

Check your dream house price here!

