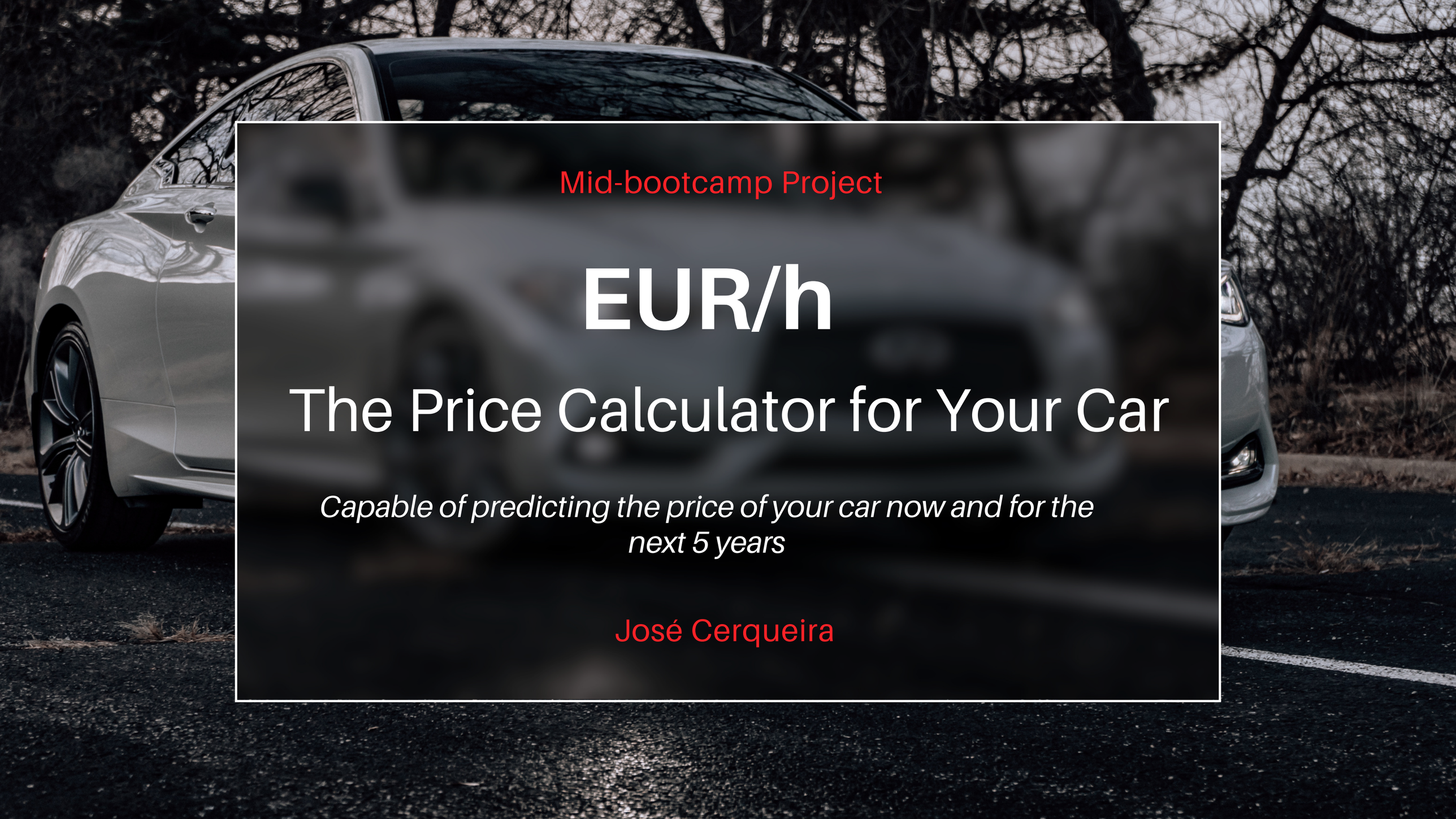


Mid-bootcamp Project

EUR/h

The Price Calculator for Your Car

Capable of predicting the price of your car now and for the next 5 years

José Cerqueira

MAIN STEPS

- Getting Reliable Data
- Preparing the Data for EDA
- EDA - Correlation Between Data
- Prediction Models
- Hypothesis Testing
- Calculator Interface

Getting Reliable Data

STAND VIRTUAL Web Site

- A website where individuals and companies sell their used cars.
- Only able to post a sale after verification by the website. Its **Reliable**!
- After web scraping we get:

1. Brand, Model and Engine Size
2. Sale Price
3. Km (How Many)
4. Registration Year (Age)

Preparing the Data for EDA

Goes through standvirtural.df

- Get the car model and brand out of a **single index**.
- Defining the columns **type**.
- Removing **duplicates**.
- After looking at the data frame :
    1. Deciding which **brands** are going to EDA

EDA - Correlation Between Data

Citroen C3 and Audi A1

- On the top of the **most common** models.
- Most **similar models** in terms of their characteristics.

Why Citroen C3 and Audi A1?

I want to know with one has a strong correlation with **price variation**:

- C3 km or A1 km?
- C3 Age or A1 Age?

Correlation with price variation

C3

```
registration_year     0.904930
km_to_date           -0.807842
sale_price            1.000000
car_age              -0.904930
km_age_ratio          0.325978
depreciation          0.741353
Name: sale_price, dtype: float64
```

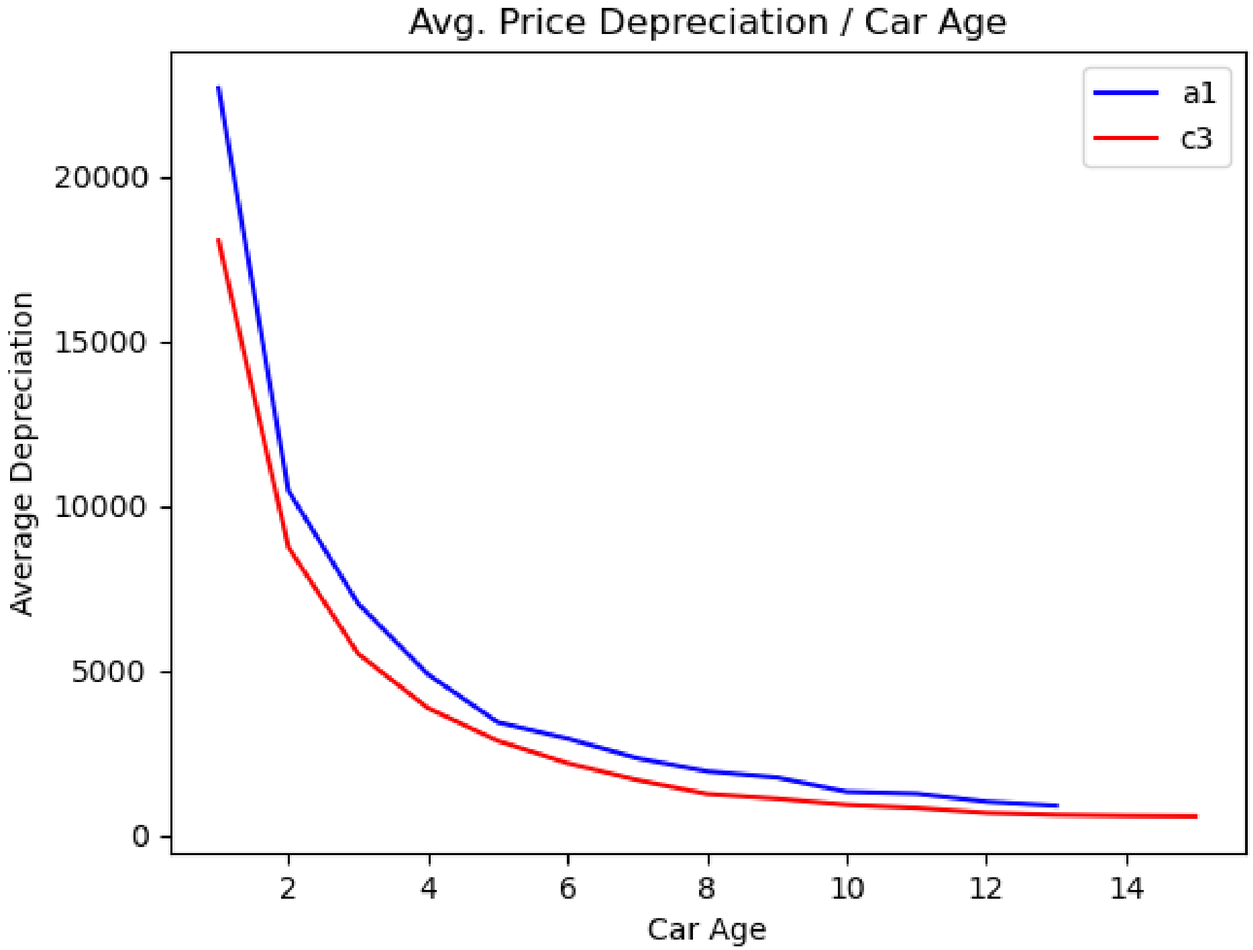
A1

```
registration_year     0.864625
km_to_date           -0.765919
sale_price            1.000000
car_age              -0.864625
km_age_ratio          0.208475
depreciation          0.748585
Name: sale_price, dtype: float64
```

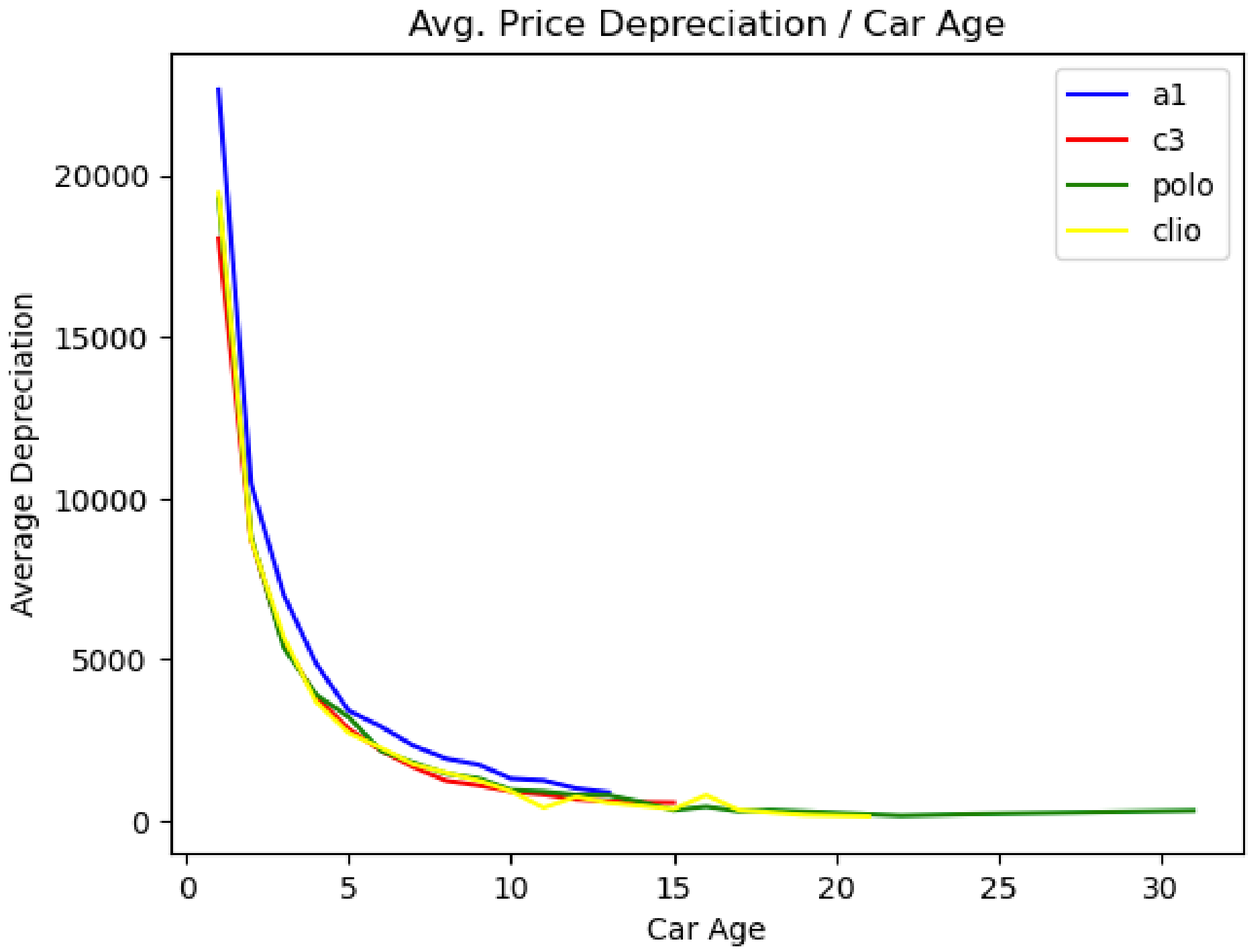
Why Citroen C3 and Audi A1?

I want to know if **brand** has an impact on **depreciation**, even for similar models.

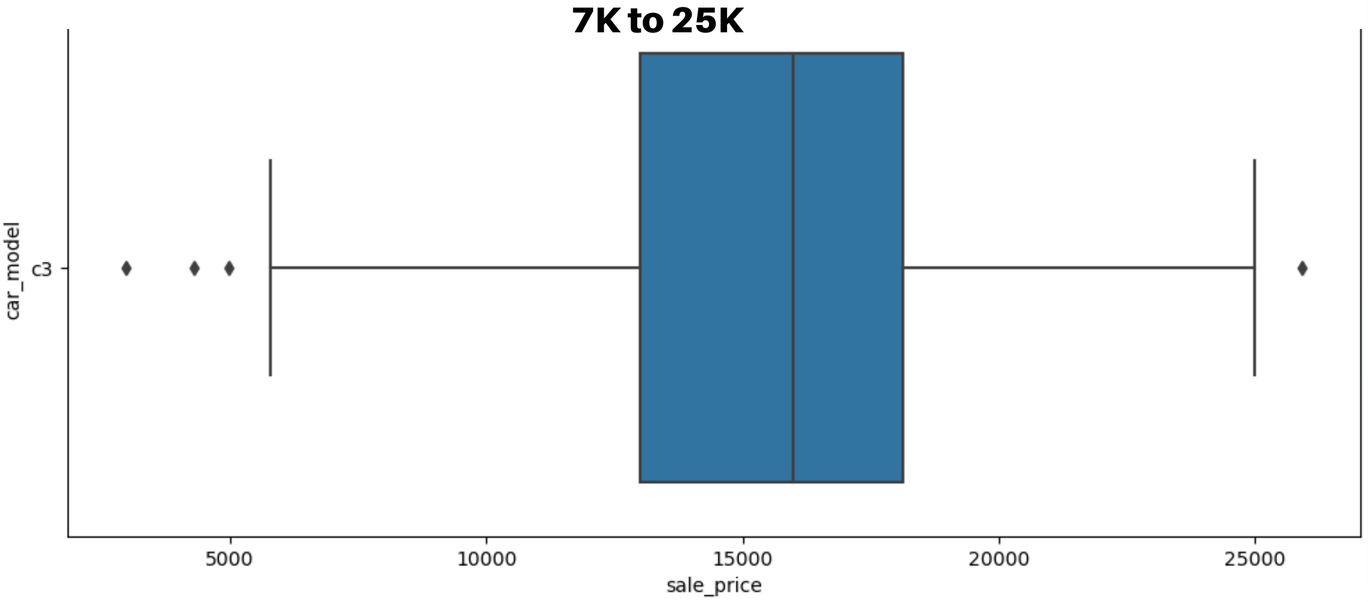
Price Depreciation:

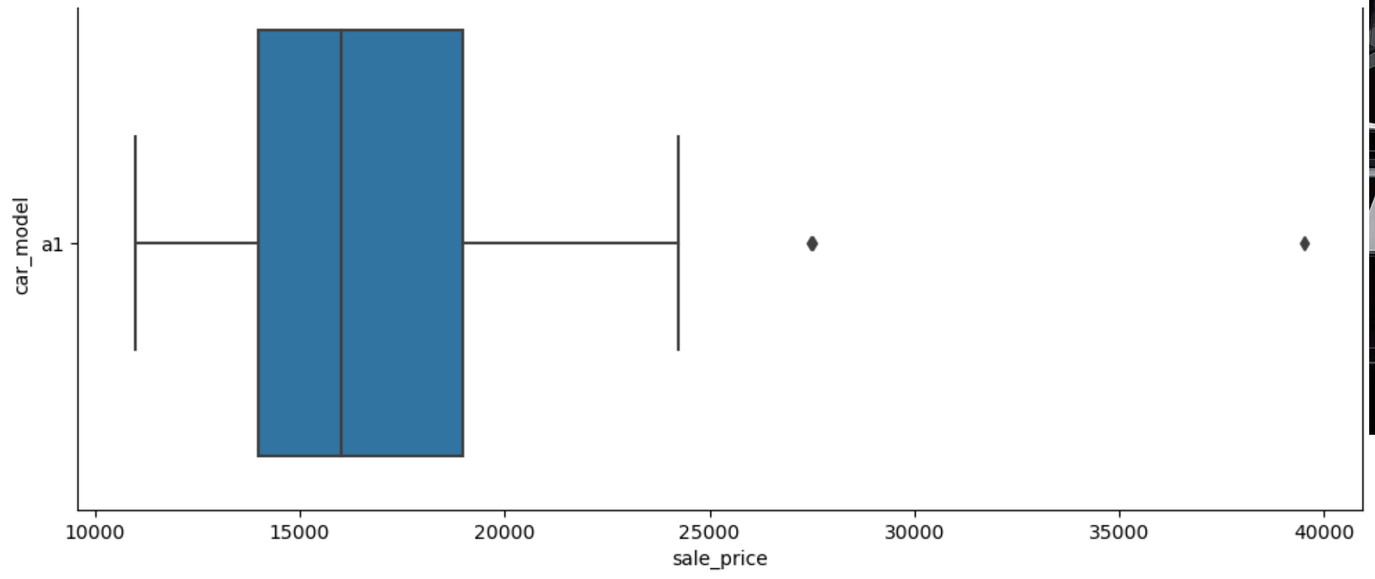


Price Depreciation:



Setting the Price Range





Prediction Models

Only for Citroen C3

- Multiple Regression
- KN Regressor
- Random Forest
- ADR Model

Prediction Models

Car details used for prediction:

Registration year - 2015

km to date - 90.000km

Multi Linear Regression - 11905.83€

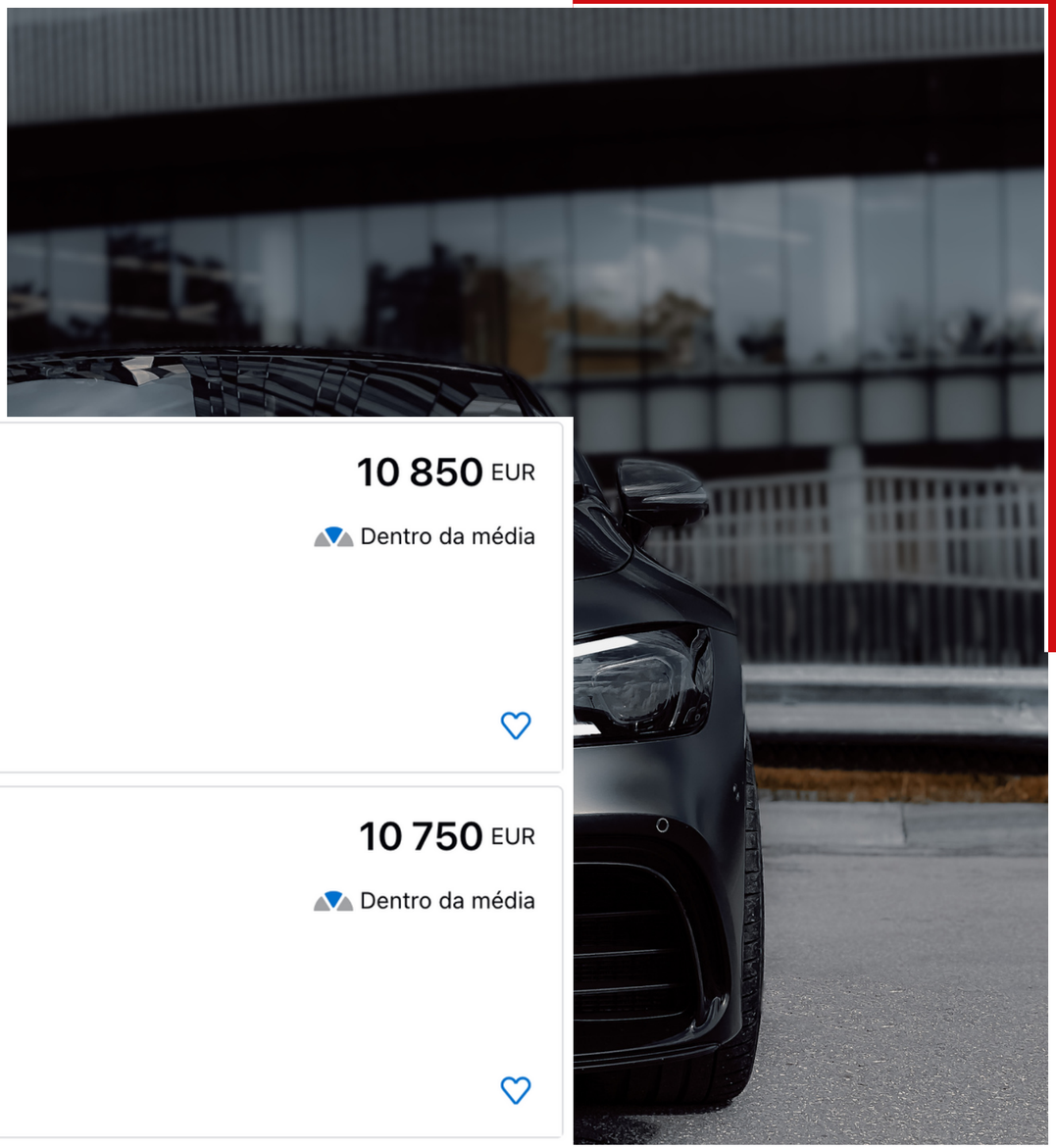
KN Regressor - 13900.00€

R. Forest - 10188.78€

ADR Model - 11489.35€

Prediction Models

We can say that R. Forest performs the best and KN Regressor the worst.



Citroën C3 1.2 VTi Exclusive
1 199 cm3 • 82 cv

89 000 km | Gasolina | Manual | 2014

Torres Novas (Santa Maria, Salvador e Santiago) (Santarém)
Publicado há 6 dias

Profissional — Ver anúncios

10 850 EUR
Dentro da média

Citroën C3 1.4 HDi Attraction
1 398 cm3 • 70 cv

94 917 km | Diesel | Manual | 2015

Santa Maria da Feira, Travanca, Sanfins e Espargo (Aveiro)
Para o topo há 2 dias

Feira Car — Ver anúncios

10 750 EUR
Dentro da média

Prediction Models

Performance

Linear Regression Model Metrics:

- MAE - The average absolute error is €1491.50, which says how far the predictions are from the actual prices.
- R-squared - An R2 value of **0.74** says that the model explains about 74% of the variance in the DF, which is good.

KNN Regressor Model Metrics:

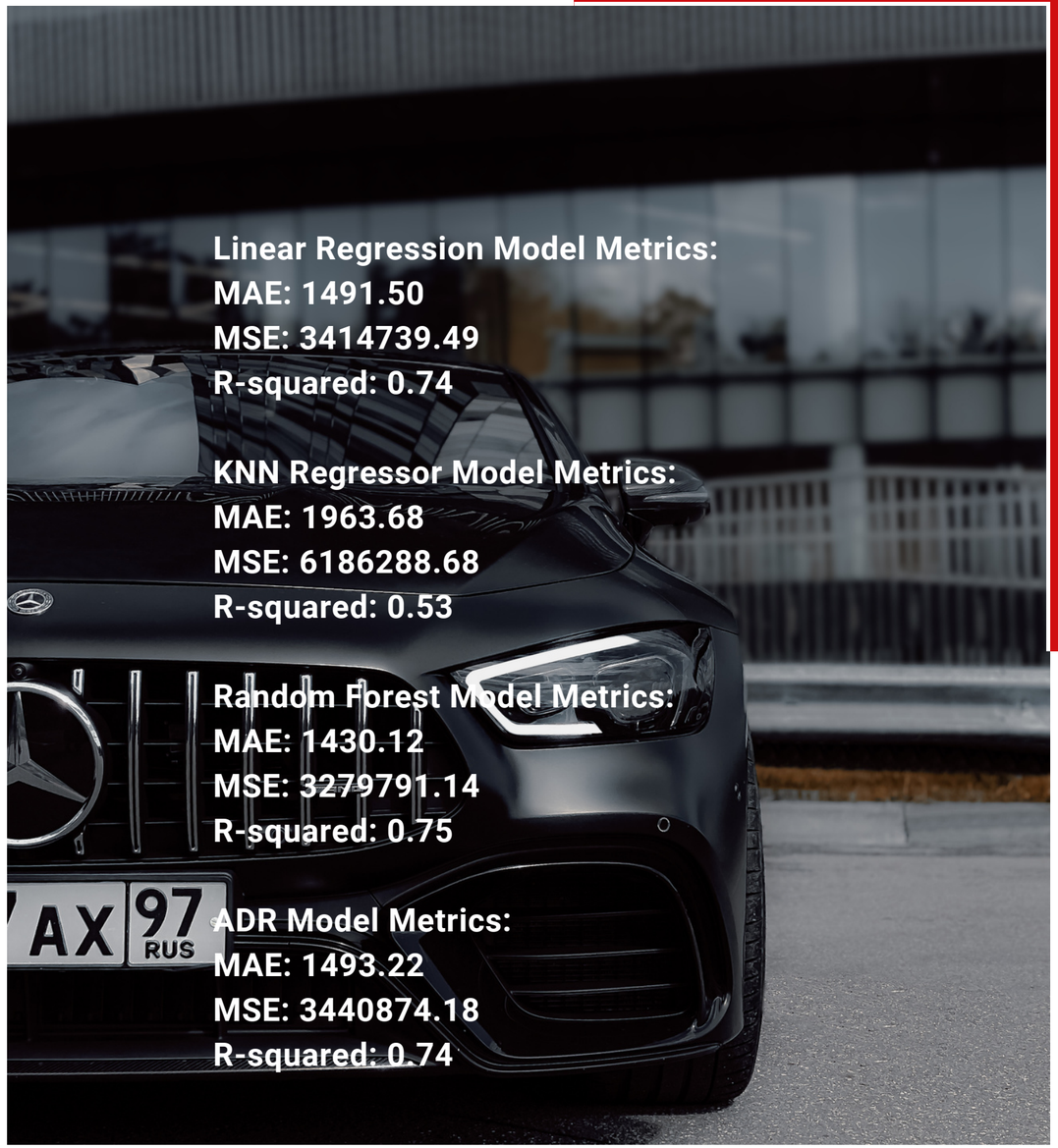
- A R-squared value of **0.53** suggests that it explains less of the variance in the DF compared to the Linear Regression model. Also MAE and MSE are higher.

Random Forest Model Metrics:

- A R-squared value of **0.75** indicating that it explains about 75% of the variance in the DF, which is a good fit.
- Also a lower MAE and MSE then the Linear Regression says it performs better.

ADR Model Metrics:

- Performance is similar to the Linear Model.



Citroen C3 and Audi A1

From 7K to 25K

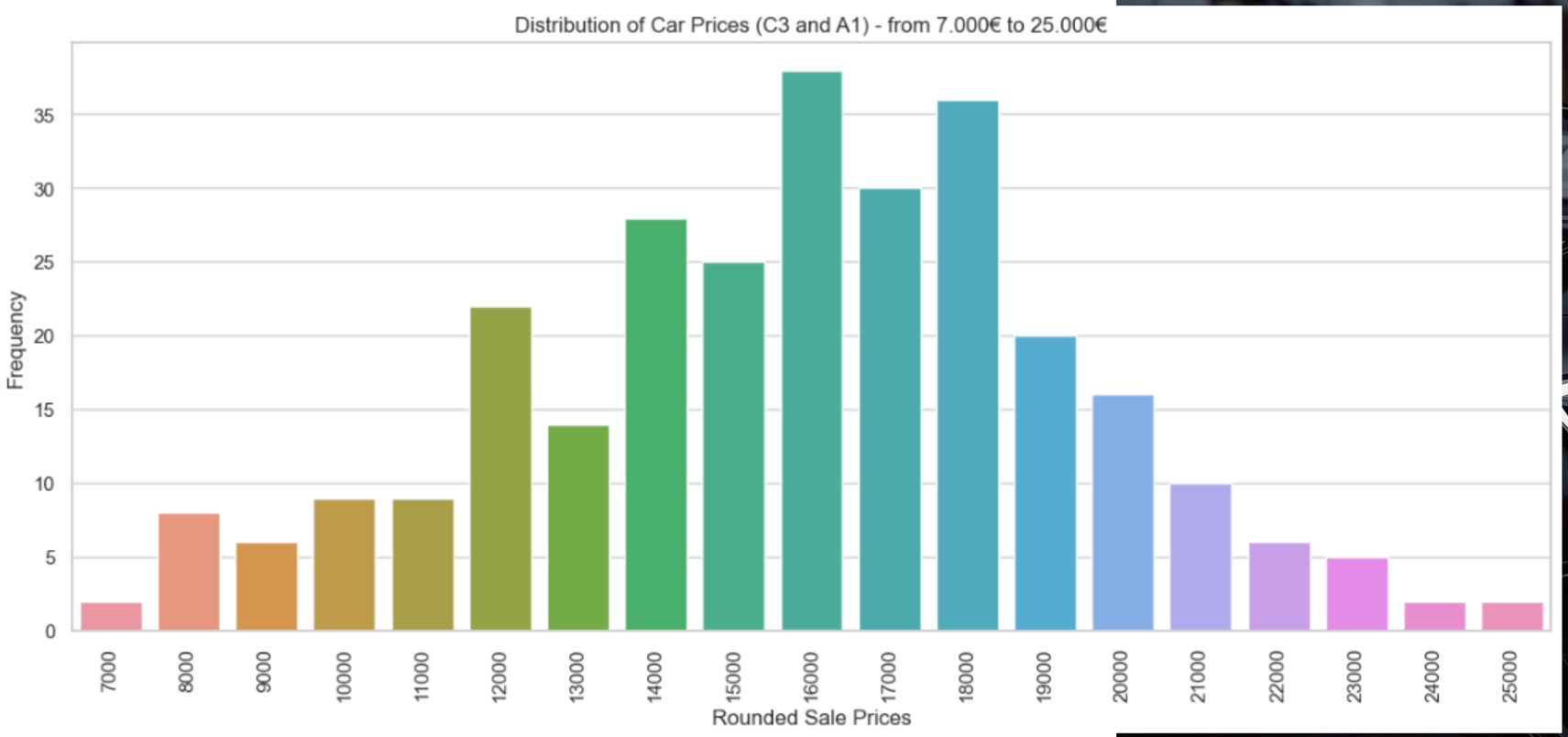
Check if the **km range** can influence the **price** of a car.

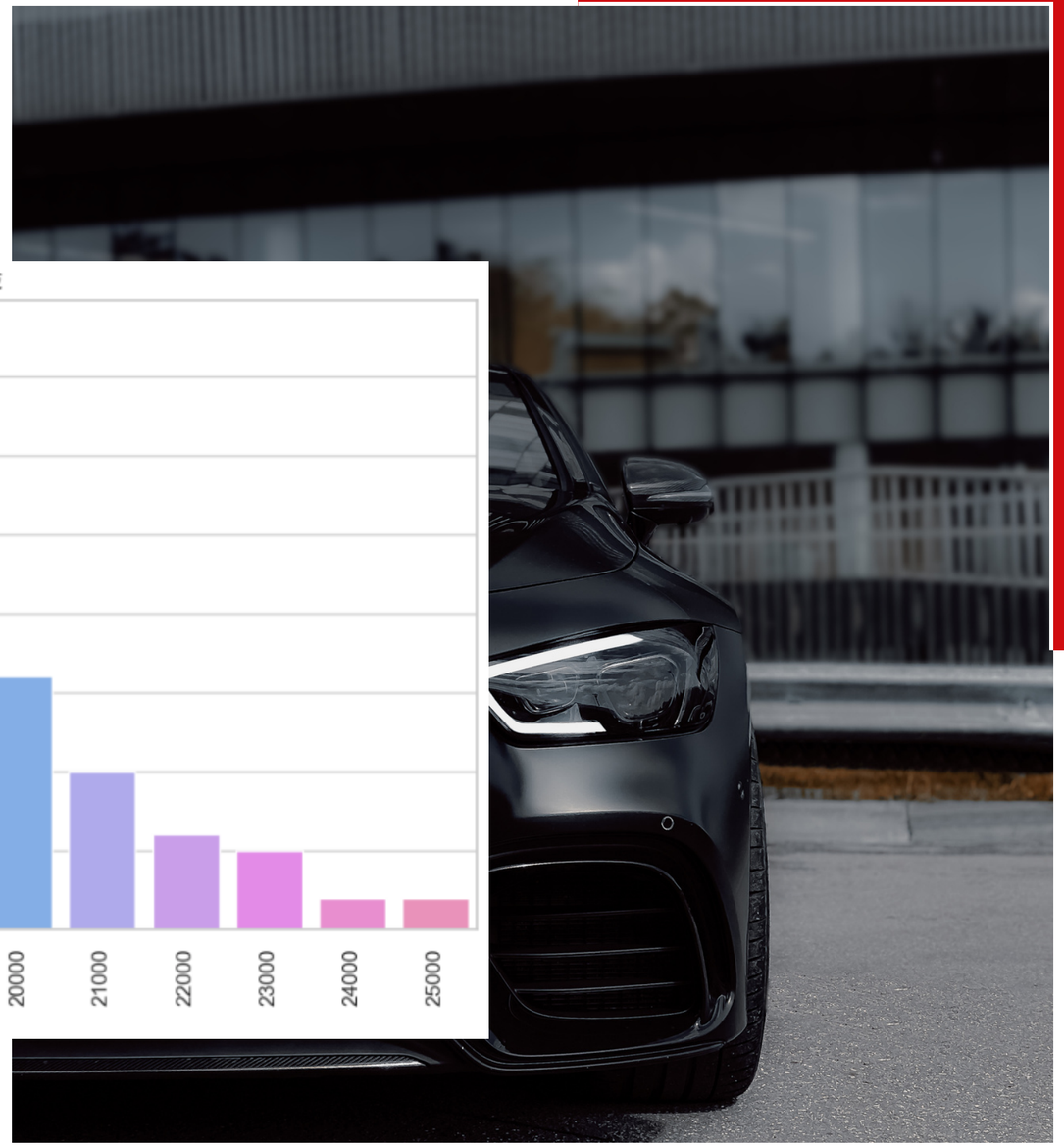
- Null-hypothesis - Km range has no influence on car price

Hypothesis Testing

Hypothesis Testing

Setting up a t_test





Rounded prices for every 1K

Hypothesis Testing

Two samples of 30:

- **Sample 1** - 0km to 90.000km
- **Sample 2** - 90.000km to 120.000km

The p-value is approx. 1.05e-05.

This indicates a **strong evidence against** the null-hypothesis.

This value can suggest that **there is a significant difference** in the average sale price between the two samples generated randomly.

If p-value is low, null must go!

Only for Citroen C3

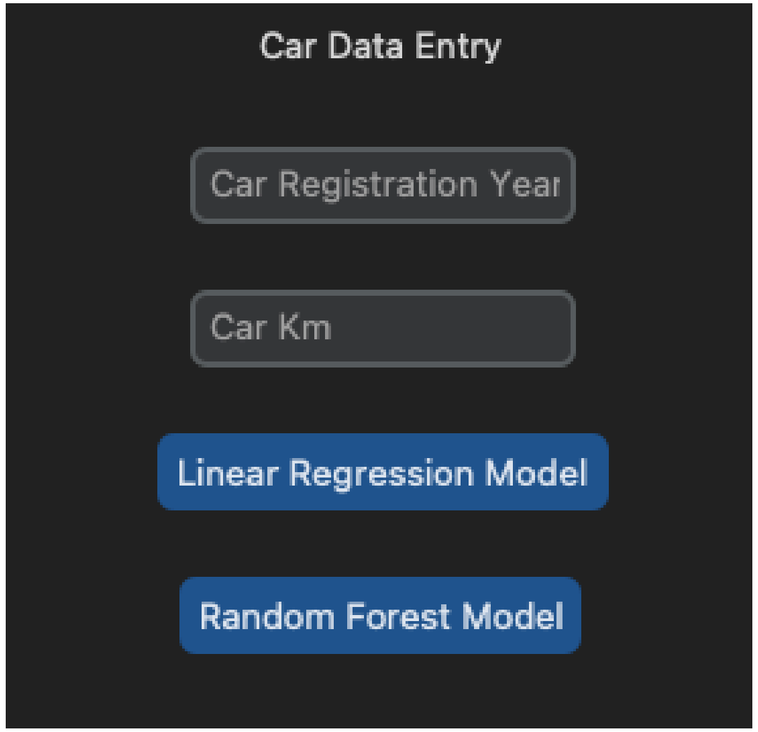
Asks the **user to input** the car registration year and the Km driven so far.

Used Models:

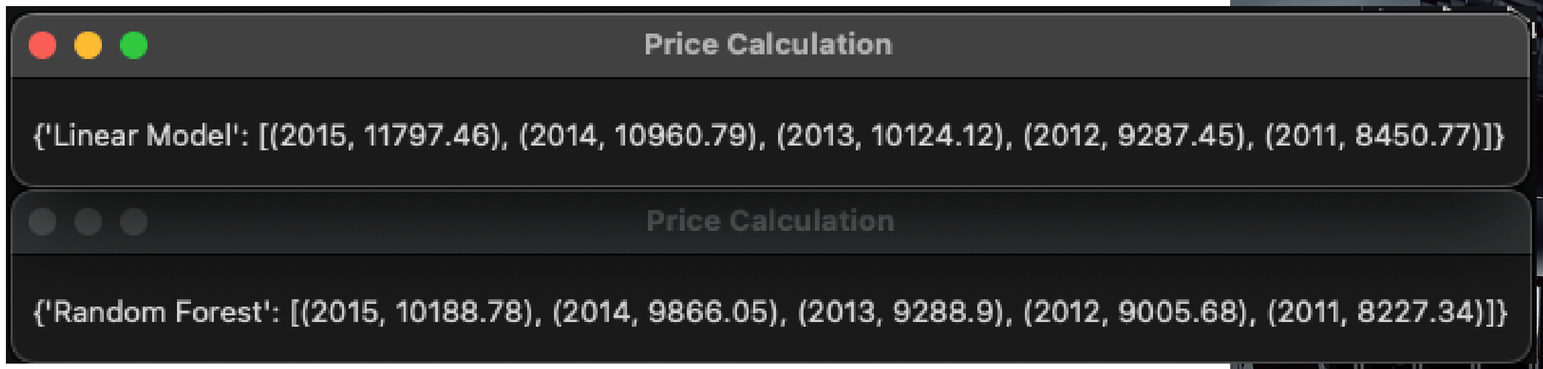
- Multi Linear Regression
- Random Forest

Calculator Interface

Calculator Interface



Calculator Interface




THANK YOU