

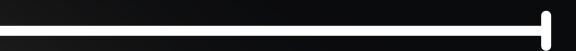


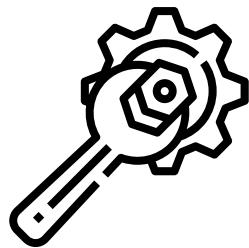
USED CAR VALUATION PLATFORM

Advanced Programming in R

Hubert Wojewoda 406420

Michał Grzyb 419327



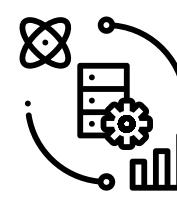


A powerful pricing tool - for the average consumer

Our project involves the creation of a used car valuation platform, where, using data scraped from the otomoto website, we created linear regression, random forest and xgboost powered pricing tool - capable of pricing a car with user-defined custom characteristics, or if provided with a link to an offer it can provide additional insight into the value of the listed car.

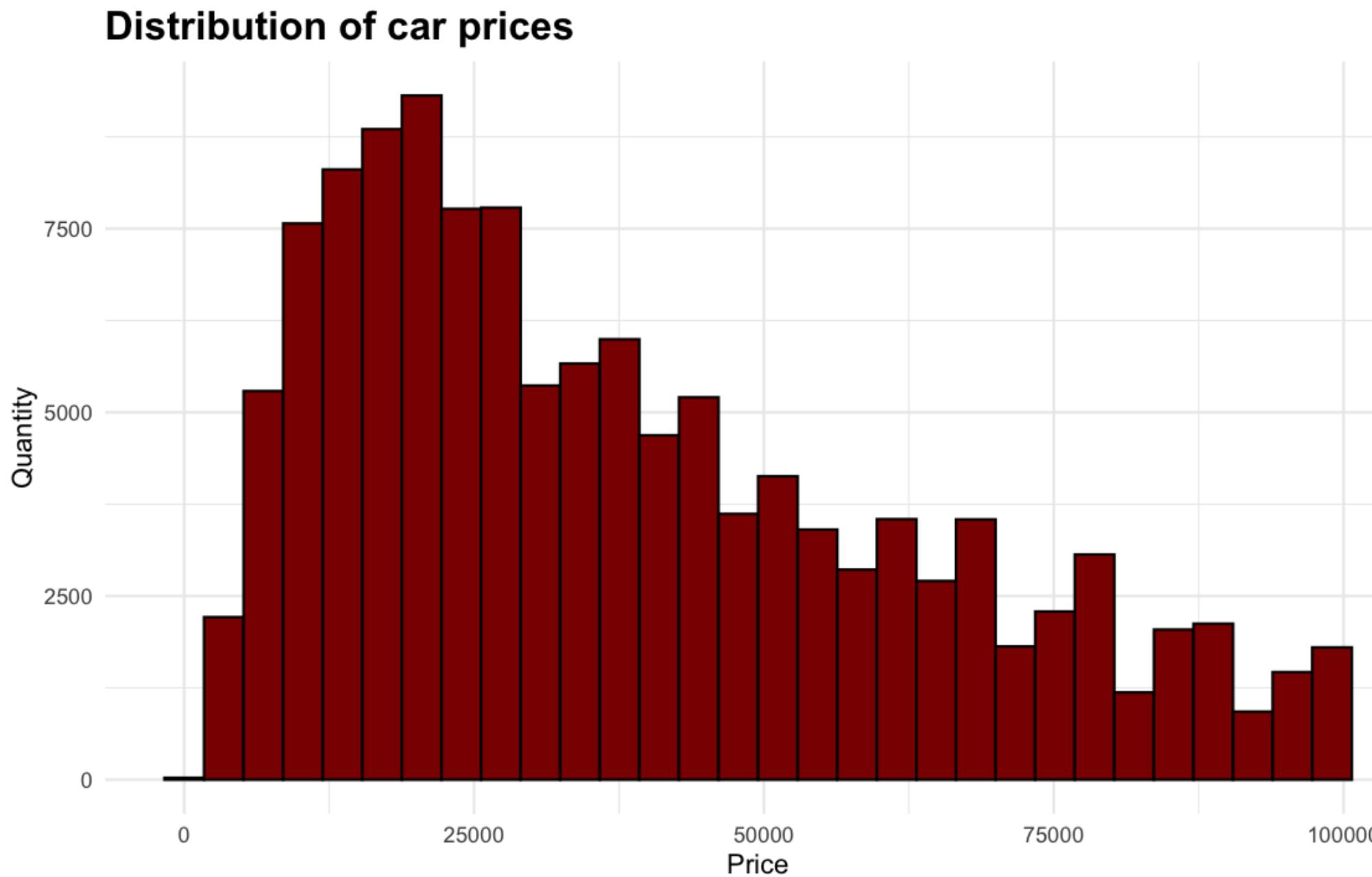
The tool aims to enhance and simplify the experience of buying or selling a car for the average consumer, that doesn't know everything about the car market.





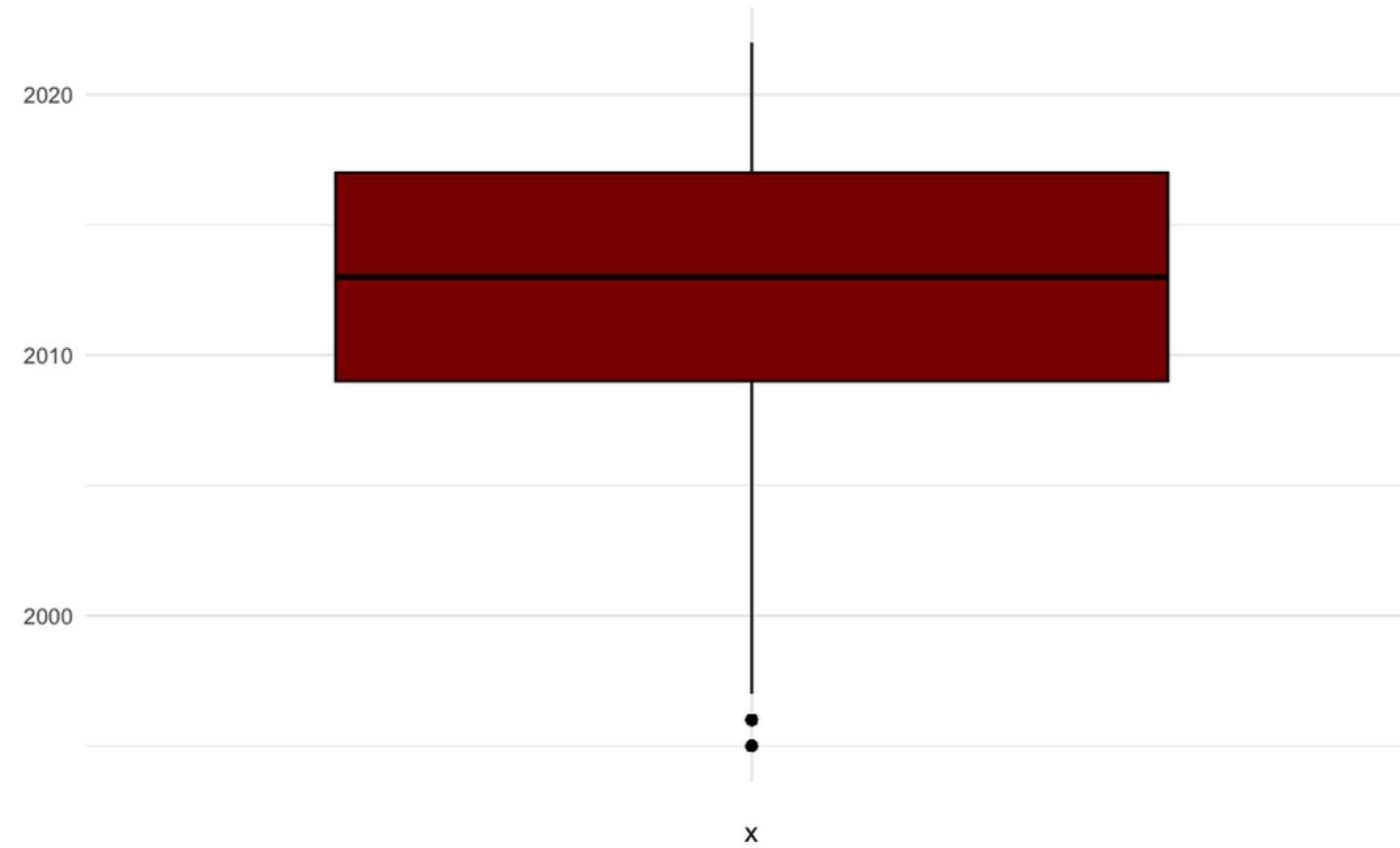
Scraping data and preprocessing

The data collected from the otomoto website includes more than 154,000 cars along with 25 attributes such as year of manufacture, power or drive type. We designed custom advanced preprocessing functions which take care of all the necessary data manipulations in our dataset.

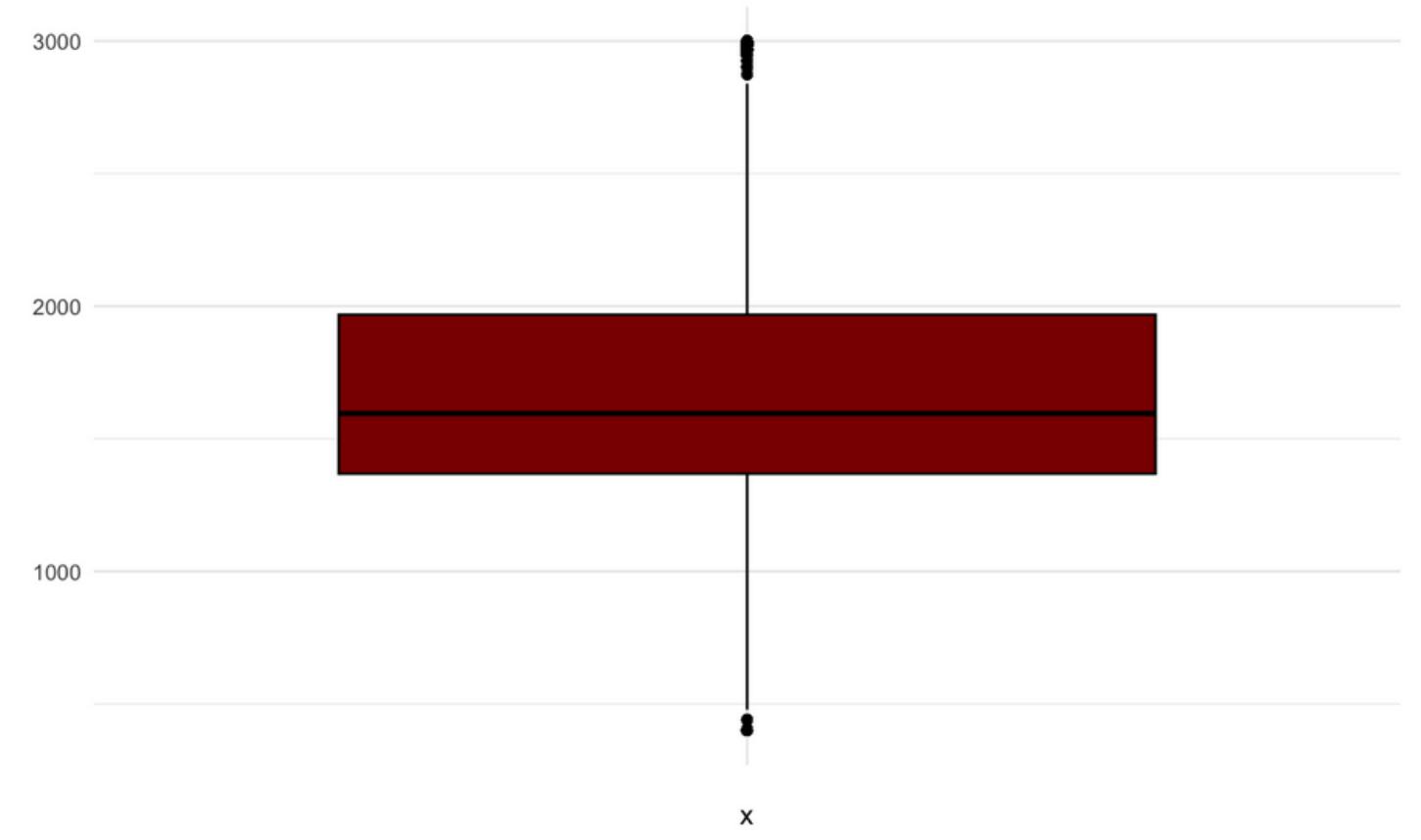


```
'data.frame': 154609 obs.  
 $ X :  
 $ Cena :  
 $ Waluta :  
 $ Wystawca :  
 ...  
 $ Marka :  
 $ RokProdukcji :  
 $ Przebieg_km :  
 $ Pojemnosc_cm3 :  
 $ Paliwo :  
 $ Moc_km :  
 $ Skrzynia :  
 $ Naped :  
 ...  
 $ Spalanie_miasto_l_100km:  
 $ Nadwozie :  
 $ Drzwi :  
 $ Miejsca :  
 $ Wyposazenie :  
 $ Kolor :  
 $ Rodzaj_koloru :  
 $ Pierwszy_wlasciciel :  
 $ ASO :  
 $ Bezwypadkowy :  
 $ Tuning :  
 $ Zabytek :  
 $ Kraj_pochodzenia :  
 ...
```

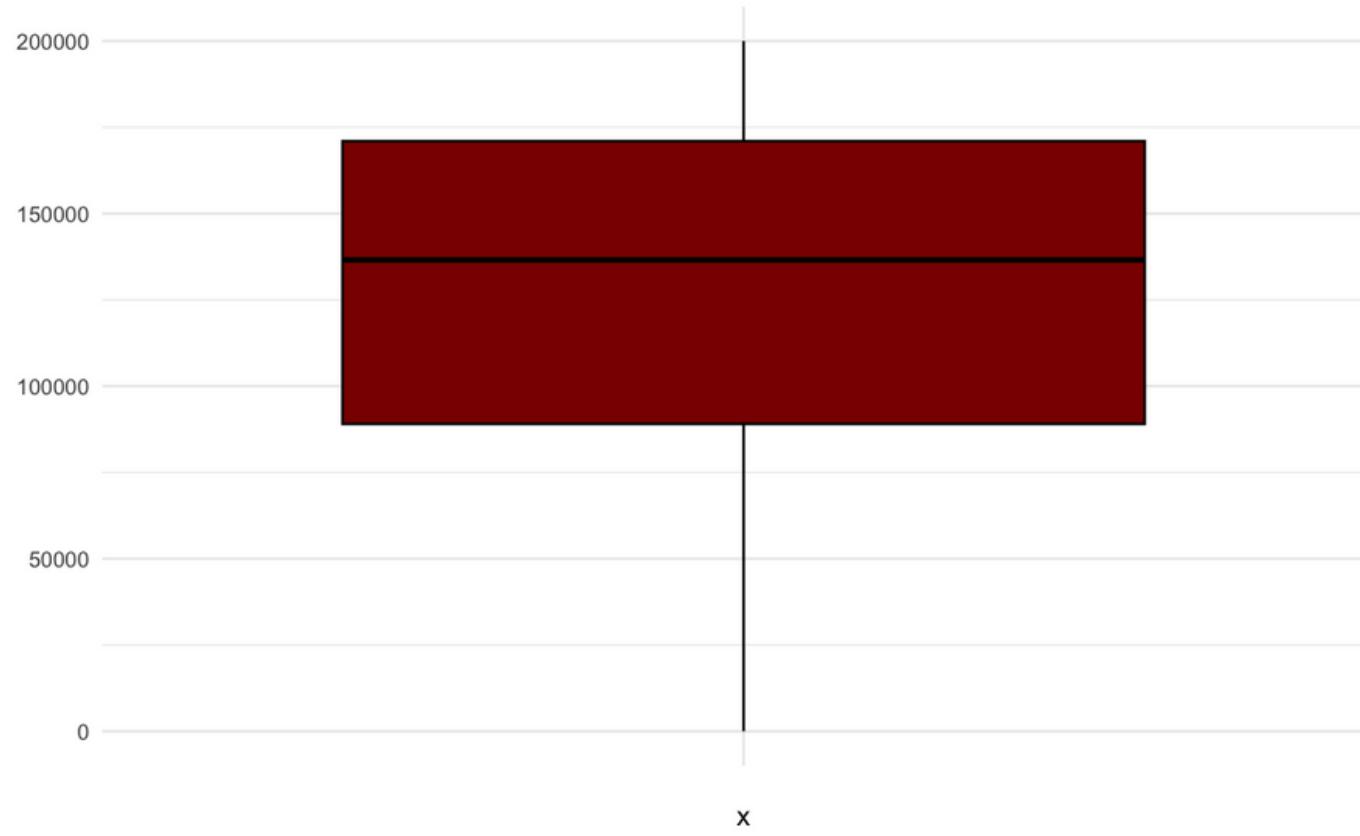
Year of production



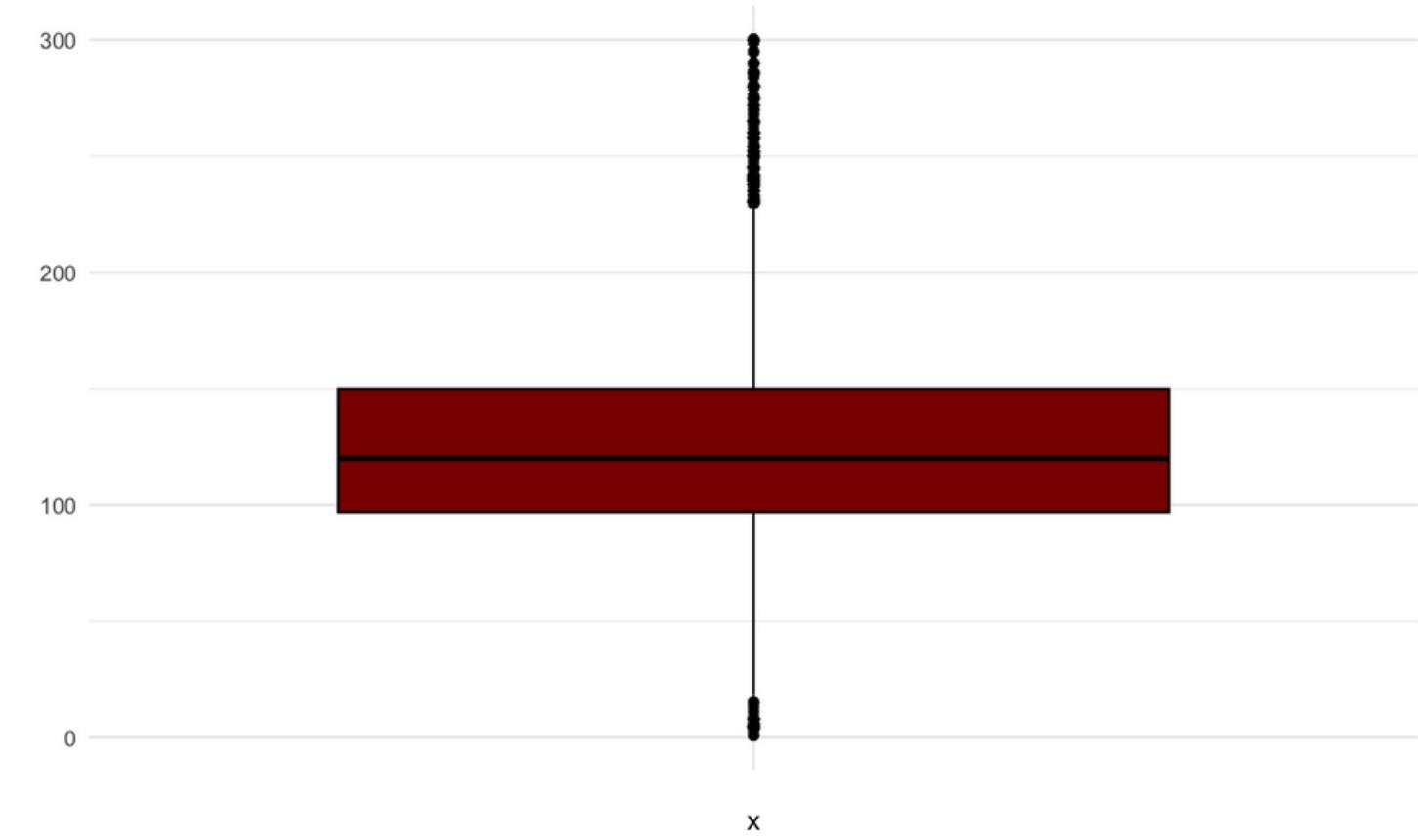
Box Plot of Engine capacity [cm³]

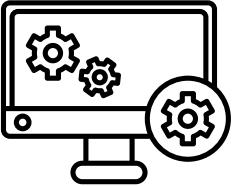


Box Plot of Mileage [km]



Box Plot of Power [km]





CarModel R6 Class - for modelling

Features:

- Data Preprocessing
- Splitting data into training and testing sets
- Model Training
- Prediction Generation
- Performance Evaluation
- Model Saving

Benefits:

- Complete workflow encapsulated in a single class
- Simplifies experimentation and comparison of different modeling approaches
- Enables efficient management of multiple instances for different datasets or configurations
- Promotes code reusability and streamlines the car price prediction process

```
# Instantiate the CarModel class
carModel ← CarModel$new()

# Preprocess data
data ← carModel$preprocess('otomoto_data.csv')
#####
# Train linear model
carModel$train(data$train_data, method = "lm")

# Make predictions
lm_predictions ← carModel$predict(data$test_data, method = "lm")

# Evaluate linear model
lm_metrics ← carModel$evaluate(lm_predictions, data$test_data)
lm_metrics

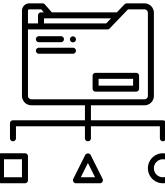
# Save linear model
carModel$save_model("lm_model", method = "lm")
#####

# Train xgboost model
carModel$train(data$train_data, method = "xgb")

# Make predictions
xgb_predictions ← carModel$predict(data$test_data, method = "xgb")

# Evaluate xgboost model
xgb_metrics ← carModel$evaluate(xgb_predictions, data$test_data)
xgb_metrics

# Save xgboost model
carModel$save_model("xgb_model", method = "xgb")
```



Custom Rcpp metric functions

We have designed and coded using C++ - two new custom scoring functions which are designed to be specific to the problems in our model's use case.

```
#include <Rcpp.h>
using namespace Rcpp;

// [[Rcpp::export]]
double price_accuracy_ratio(NumericVector predicted, NumericVector actual)
{
  int n = predicted.size();
  double sum = 0.0;

  for (int i = 0; i < n; i++)
  {
    sum += std::abs(predicted[i] - actual[i]) / actual[i];
  }

  return sum / n;
}

// [[Rcpp::export]]
List segment_specific_mae(NumericVector predicted, NumericVector actual)
{
  NumericVector segments = NumericVector::create(10000, 25000, 50000, 75000, 100000);
  NumericVector mae_per_segment(segments.size());

  for (int i = 0; i < segments.size(); ++i)
  {
    double sum = 0.0;
    int count = 0;
    for (int j = 0; j < actual.size(); ++j)
    {
      if (actual[j] <= segments[i] && (i == 0 || actual[j] > segments[i - 1]))
      {
        sum += std::abs(predicted[j] - actual[j]);
        ++count;
      }
    }
    mae_per_segment[i] = (count > 0) ? sum / count : NA_REAL;
  }

  return List::create(Named("segments") = segments,
                     Named("mae") = mae_per_segment);
}
```

Price accuracy ratio (PAR)

The purpose of the PAR is to measure how accurate the model's predictions are as a ratio of the actual price of the cars. This could be especially useful in a setting where the prices can vary significantly, like in the used car market.

Segment Specific Mean Absolute Error (SMAE)

SMAE extends the ordinary MAE concept to compute it separately for different price segments of the car market. It returns a list containing two vectors: one for the segments and one for the MAE associated with each segment. Each segment is defined by the upper limit of the car price for that segment.

Home

Predictions

Scrape and Predict

Car Price Predictions

Production Year

2013

Mileage (km)

100000

Engine Size (cm³)

2000

Power (km)

100

Fuel

Benzyna

Transmission

Automatyczna

Drive

Na przednie koła

Body

Auta małe

First owner

Nie

ASO

Nie

Accident free

Nie

Exhibitor

Firmy

Predict

Predicted Car Prices:

Lower confidence bound: PLN 41410.84**PLN 43590.36****Higher confidence bound: PLN 45769.88**

[Home](#)[Predictions](#)[Scrape and Predict](#)

Scrape Car Data and Predict Price

Enter the URL of the car listing

Scrape Data and Predict

Scraped Car Data:

	Rok_produkcji	Przebieg_km	Pojemnosc_cm3	Moc_kw	Paliwo	Skrzynia	Naped	Nadwozie
1	2020	45050	9993	100	Benzyna	Manualna	Na przednie koła	Kompakt
	Pierwszy_wlasciciel	AS0	Bezwypadkowy	Wystawca				
1	Tak	Tak	Tak	Firmy				

Predicted Car Price:

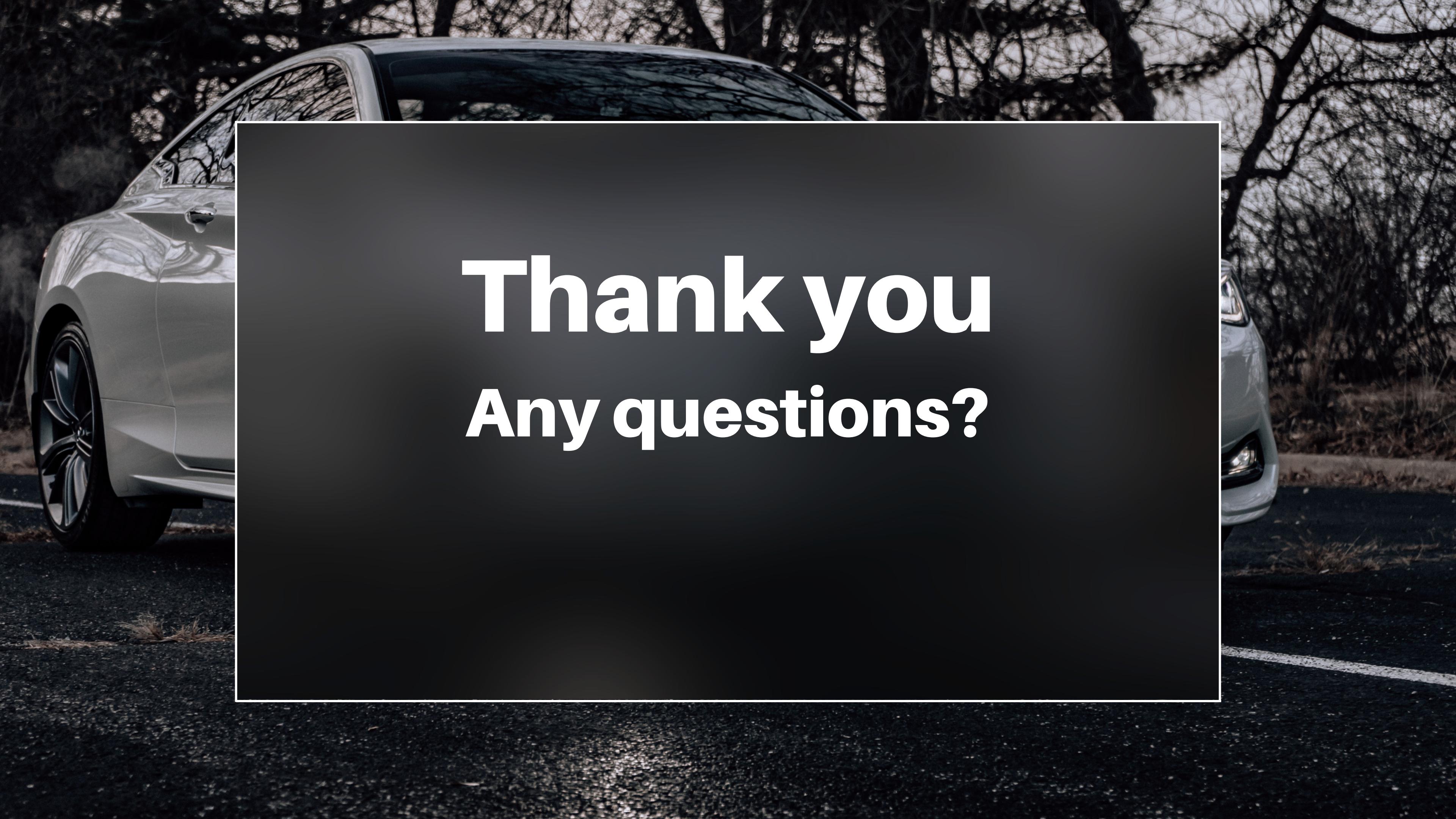
Listed Price: PLN 81900

Predicted Price: PLN 81084.54

Difference: PLN -815.46 (-1%)

A low-angle, close-up shot of the front of a dark-colored Mercedes-Benz AMG. The car's iconic grille with vertical slats and the AMG logo is visible on the left. To the right is the large, illuminated headlight. The background is dark, suggesting it might be nighttime or the car is parked in a dimly lit area.

Time for a live demo



Thank you
Any questions?