



BY ANTON  
AND  
VALENTIN

# TOYOTA ANALYSIS

[HTTPS://WWW.TRUECAR.COM/](https://www.truecar.com/)

# CONTENT

**01**

HAVE YOU EVER WONDERED HOW TO BUY A USED CAR?  
OUR PROJECT IS BASED ON THE TOYOTA MARKET ONLY

**02**

THERE ARE MANY MODELS OF TOYOTA WITH DIFFERENT FEATURES.

**03**

ANALYSIS OF THE FEATURES OF TOYOTA CARS AND THEIR IMPACT ON  
COST.

**04**

ANALYTICS OF THE MOST POPULAR TOYOTA MODELS



# RESEARCH QUESTION

Is it possible to predict the fair price  
of a vehicle using a comprehensive  
database?

# PARSING



Here is an example of a webpage from which all the information for the project was extracted.



Parsing was performed using the BeautifulSoup library to gather interest data from various scattered locations on the page, taking into account that the location of the data may vary.

## 2021 Toyota Corolla

LE CVT | 26,782 miles

[Share](#) [Save](#)



List Price

**\$20,143**

Excellent Price

\$912 (4.5%) below avg. list price

Est. Monthly Payment

\$334/mo [View & Edit](#)

Mileage

26,782

Location Albany, GA

Upfront Price Available ⓘ

No accidents reported, 1 owner

Minimal Options

[Unlock Dealer Details](#)



**Prequalify for Financing Today**

Get prequalified with no impact to your credit score.

[Get Started](#)



[Vehicle Overview](#)

[Condition History](#)

[Popular Features](#)

[Price Summary](#)

[Pricing Context](#)

[Seller Notes](#)

### Vehicle Overview

**Style**  
Sedan

**Exterior Color**  
Super White

**Interior Color**  
Black

**MPG**  
30 city / 38 hwy

**Engine**  
1.8L Inline-4 Gas

**Drive Type**  
FWD

**Fuel Type**  
Gas

**Transmission**  
Automatic





# FEATURES

## 2021 Toyota Corolla

LE CVT | 26,782 miles









List Price

**\$20,143**

-  Albany, GA
-  Upfront Price Available ⓘ
-  No accidents reported, 1 owner
-  Minimal Options

- Each car has different characteristics. The characteristics can vary, and some of them are acquired by the car during its usage, such as mileage, the number of previous owners, and the number of accidents.

- There are also features that the car comes with originally from the manufacturer, such as engine, consumption, color, etc.

 <b>Style</b> Sedan	 <b>Exterior Color</b> Super White
 <b>Engine</b> 1.8L Inline-4 Gas	 <b>Drive Type</b> FWD
 <b>Interior Color</b> Black	 <b>MPG</b> 30 city / 38 hwy
 <b>Fuel Type</b> Gas	 <b>Transmission</b> Automatic





1

2

3

4

5

## CREATING DATAFRAME

Every row is a different Toyota car and each column represents different features.

## REMOVING CARS THAT HAVE MISSING DATA.

Due to the specificity of the data, we anticipate the missing features.

## REMOVING RARE INDIVIDUAL FEATURES FROM DATAFRAME

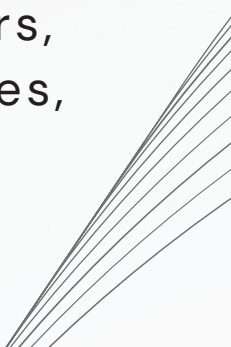
Since we are comparing a large number of cars, rare individual features will not provide us with useful information, so we cannot use them for comparison purposes.

## CLEANING OF DUPLICATES

Cleaning the data frame from duplicates that do not help us in any way in the research.

## PREPARATION DATA FRAME FOR MACHINE LEARNING

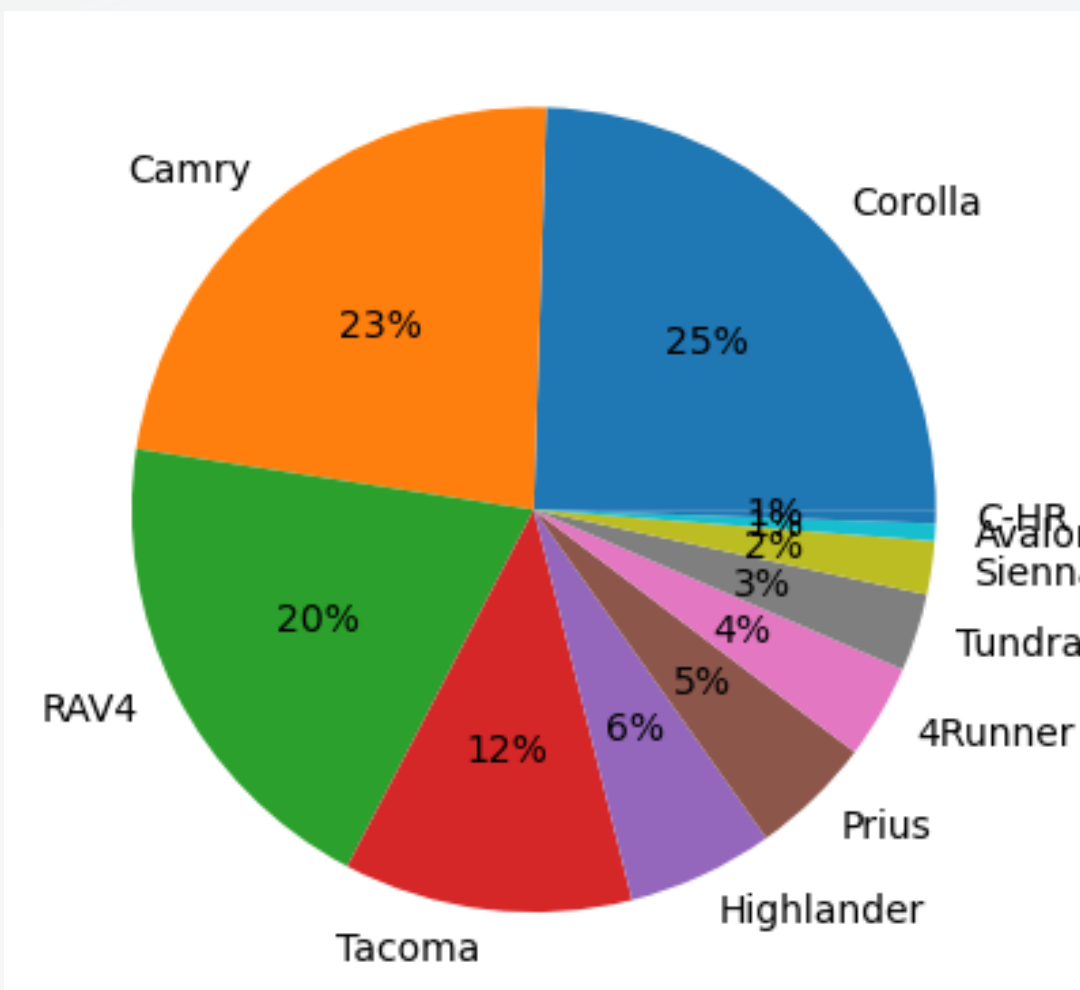
Cleaning up unnecessary characters in the data frame features, such as commas in numbers, dollar signs in prices, etc.



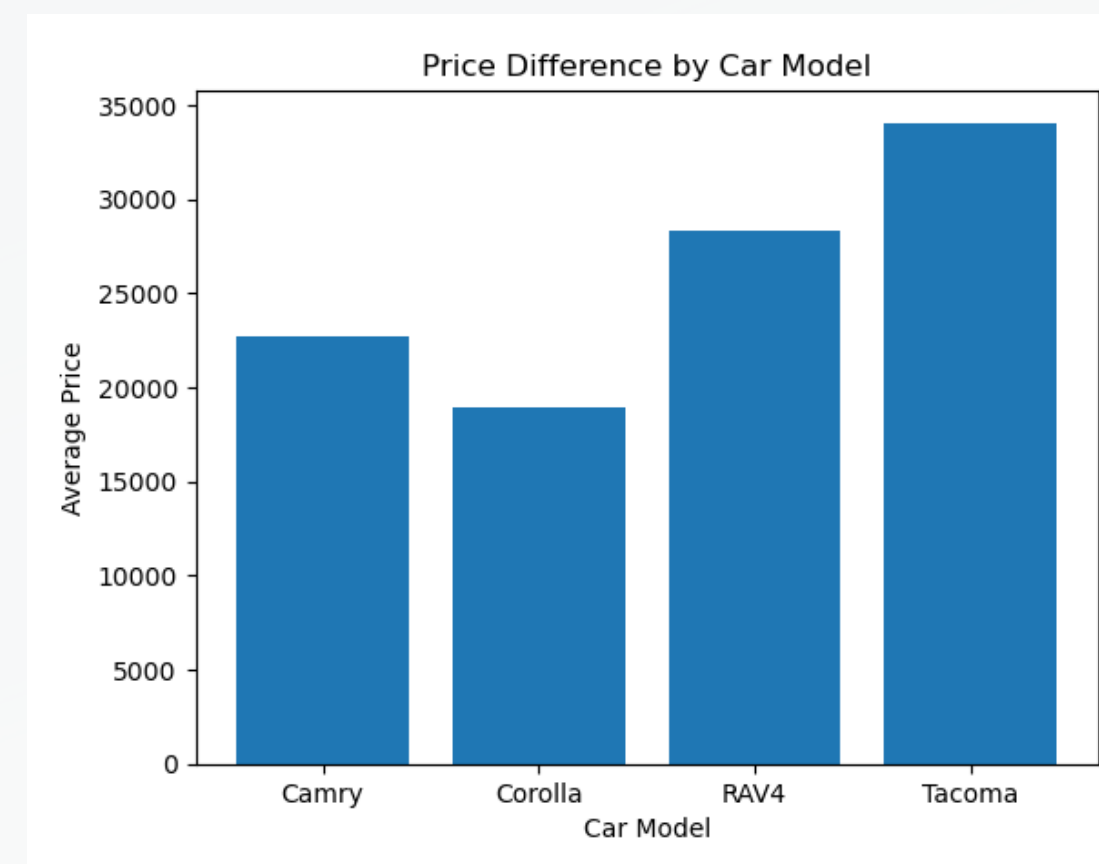
# EDA

With the help of the pie chart, we were able to clearly check the most popular models of the Toyota brand, in order to take exactly them for our research, as cars of which small quantities do not contribute to research.

Corolla	1399
Camry	1318
RAV4	1127
Tacoma	663
Highlander	336
Prius	280
4Runner	216
Tundra	178
Sienna	122
Avalon	40
C-HR	31

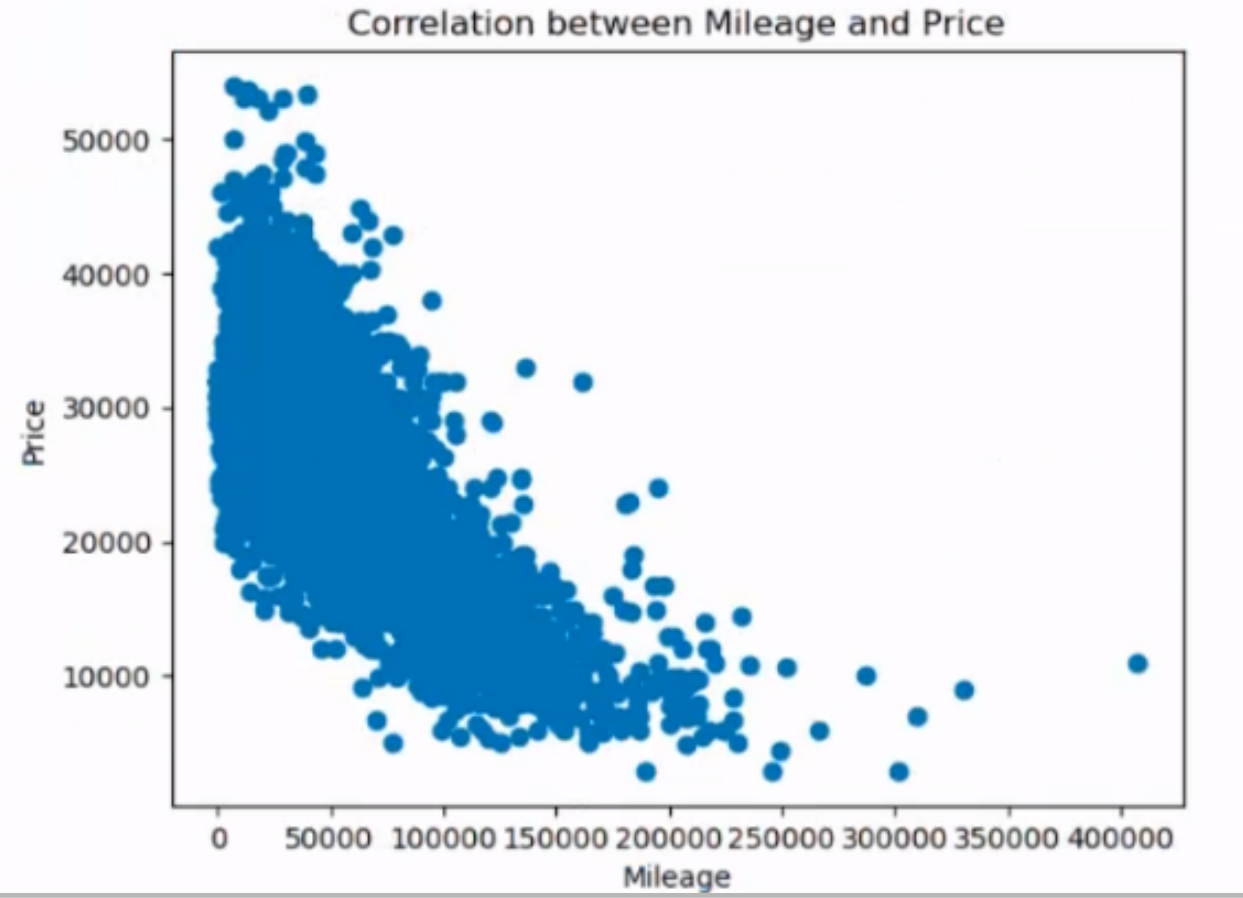


Next, we checked the correlations between the "price" feature and other features in the data frame in order to understand which features affect the price more and help us in our research.

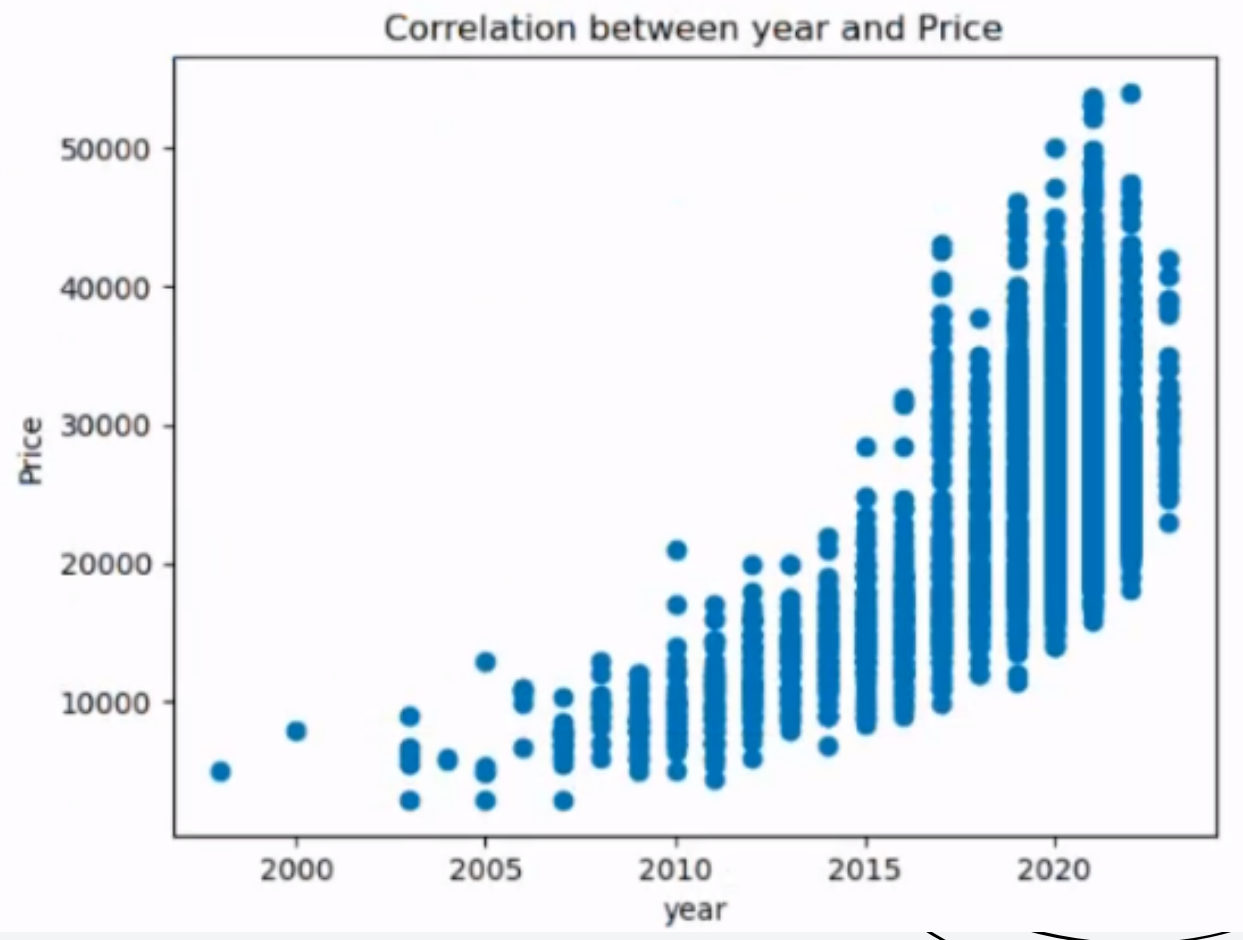


CORRELATIONS

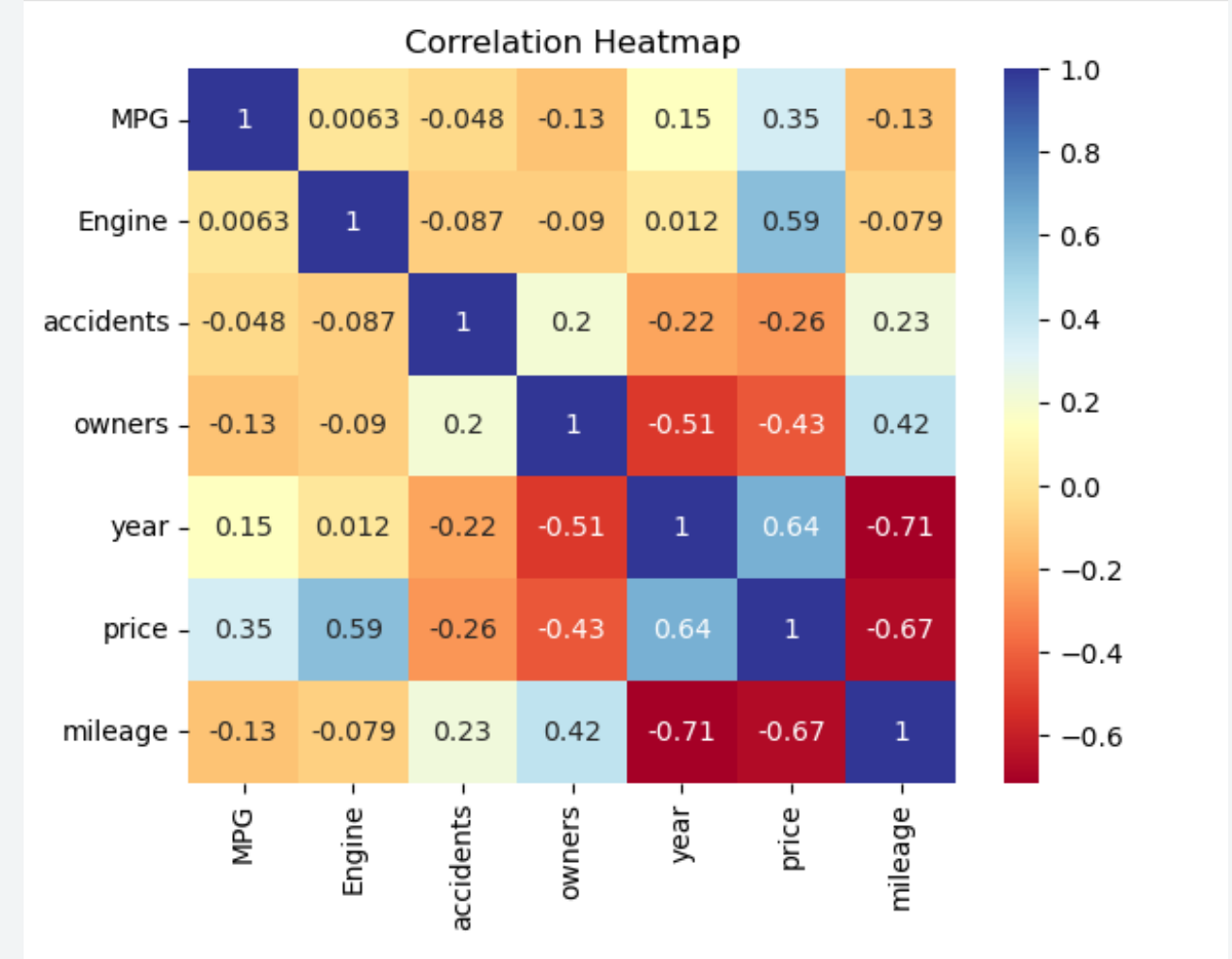
CORRELATIONS



Correlation Coefficient: -0.6727253896401031



Correlation Coefficient: 0.6462042567843962



Correlatin's coefficent: 0.04785943130816836

Name: Exterior Color, Length: 119, dtype: int64  
Correlation between 'price' and 'Exterior Color':  
color\_49035: -0.02119927646911944  
color\_Aloe Green Metallic: -0.03451042354037666  
color\_Attitude Black: -0.04494376548132127  
color\_Attitude Black Metallic: -0.08183815125379225  
color\_Balck: -0.01578195504815986  
color\_Barcelona Red Metallic: -0.06012503340277321  
color\_Beige: -0.052417302081316196  
color\_Black: -0.16149233783904468

color\_White: -0.0461003496014669  
color\_White W/Black Sand Pearl Roof: 0.005903577184425236  
color\_Wind Chill Pearl: 0.09514619663706052  
color\_Wind Chill Pearl/Midnight Black Metallic: 0.03582264793292066  
color\_Yellow: -0.02101695408874065  
Average correlation between price and Exterior Color: -0.005569395630008432



# ADVANCED DATA ANALYSIS

## New data frame

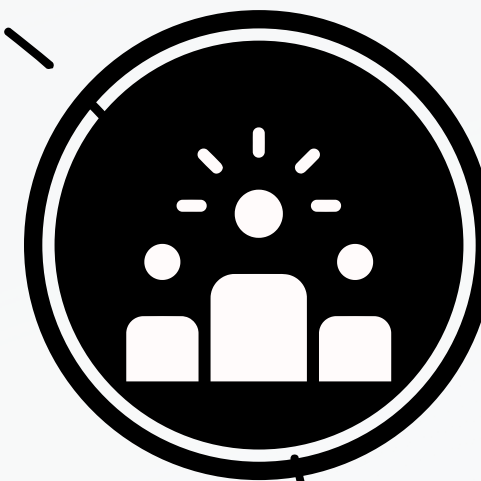
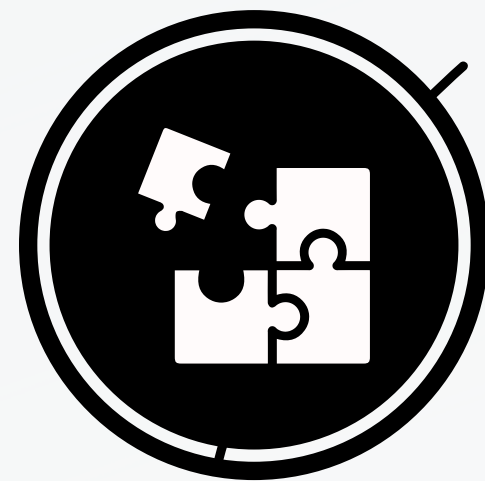
Created a data frame for the 4 most popular models, since models with a small number of cars can worsen the price forecast accuracy coefficient.

## Necessary features

In the new data frame, we took only the necessary features that will be used to predict prices

## Categorical features

Model features represent categorical variables that indicate the type of car model. Each variable can take the value 0 or 1, where 1 indicates that the car belongs to the corresponding model, and 0 indicates that it does not belong to this model.



	MPG	Engine	accidents	owners	year	mileage	model_Camry	model_Corolla	model_RAV4	model_Tacoma
1	0.717949	2.5	1	2	2019	103386	1	0	0	0
2	0.789474	1.8	0	1	2022	21320	0	1	0	0
4	0.789474	1.8	0	1	2020	57515	0	1	0	0
5	0.789474	1.8	1	2	2020	62134	0	1	0	0
7	0.789474	1.8	0	1	2021	57144	0	1	0	0
...	...	...	...	...	...	...	...	...	...	...
5954	0.777778	1.8	0	2	2017	67945	0	1	0	0
5955	0.714286	2.5	0	2	2016	37479	1	0	0	0
5956	0.714286	2.5	1	2	2016	129263	1	0	0	0
5957	0.717949	2.5	0	1	2021	30901	1	0	0	0
5958	0.771429	2.5	0	1	2021	11205	0	0	1	0

4376 rows × 10 columns

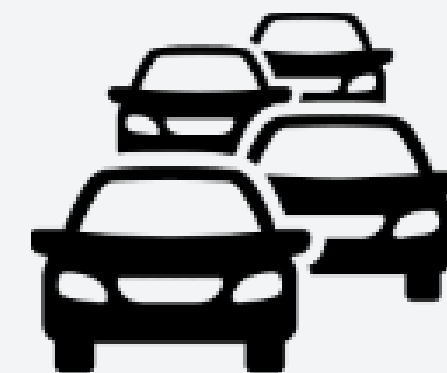
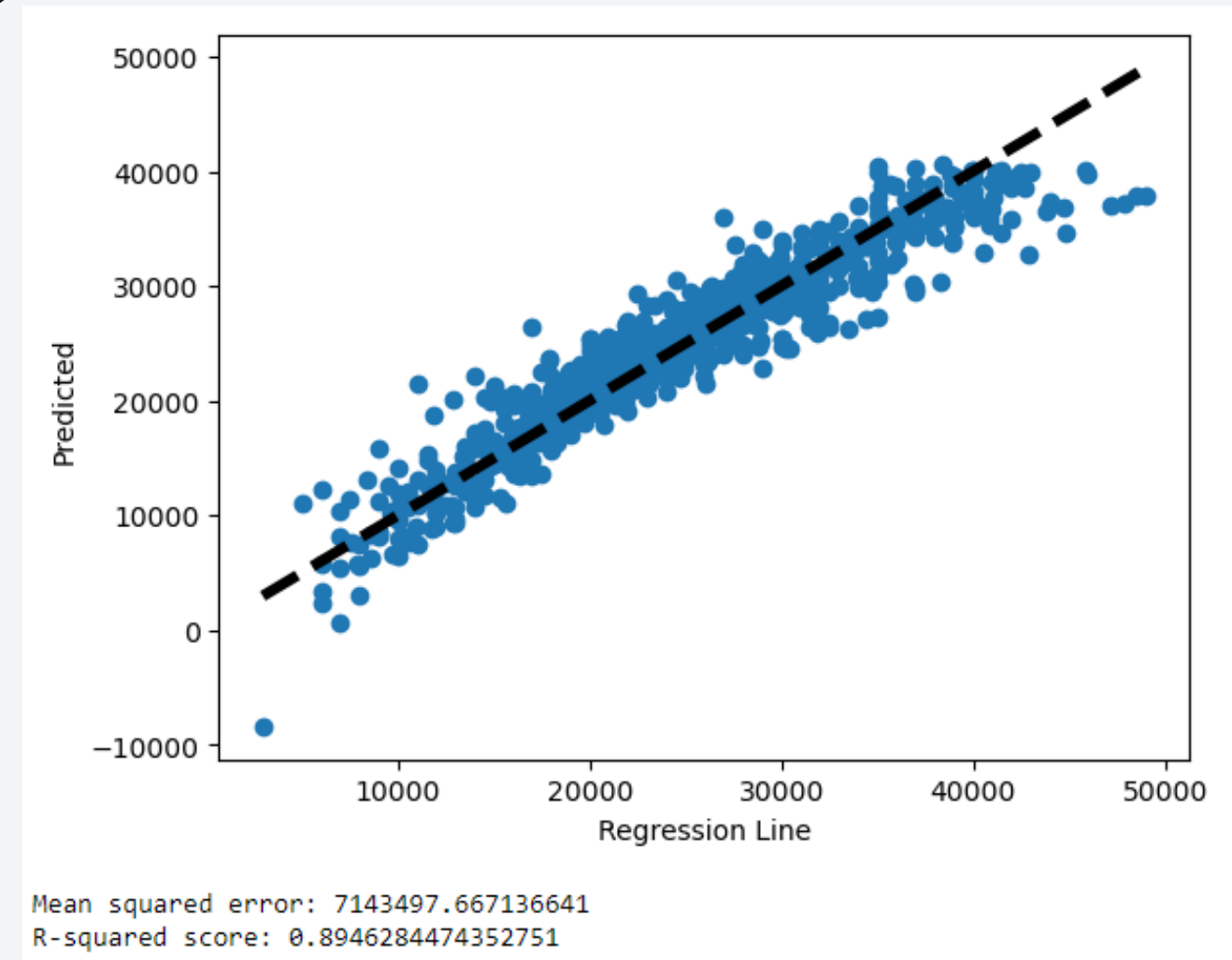
# METHOD

We had chosen an Linear Regression in order to predict price in the following reasons : Linear regression algorithm establish the relationship between dependent variable and independent variable, as for Logistic Regression it fits more for binary problems that answers the question Yes or No, for example an insurance companies. Also Logistic Regression often used for classification problems, as for us it does not fit completely therefore Linear Regression .

**LINEAR  
REGRESSION**

# ADDITIONAL CHECKING

Removing the most expensive cars with inflated prices allows us to train the machine more accurately for further price prediction. On the side, we can observe the R-score before and after the removal.



# 70%

Mean squared error: 2804277.433266712  
R-squared score: 0.9042248701479308



```
##### Cross-Validation: #####
```

```
Cross Validation Validity Values: [0.86098628 0.87893995 0.88912575 0.88865758 0.89087603]
```

```
Cross-validation is a valuable tool for model evaluation and can help prevent overfitting and aid in hyperparameter tuning.
```

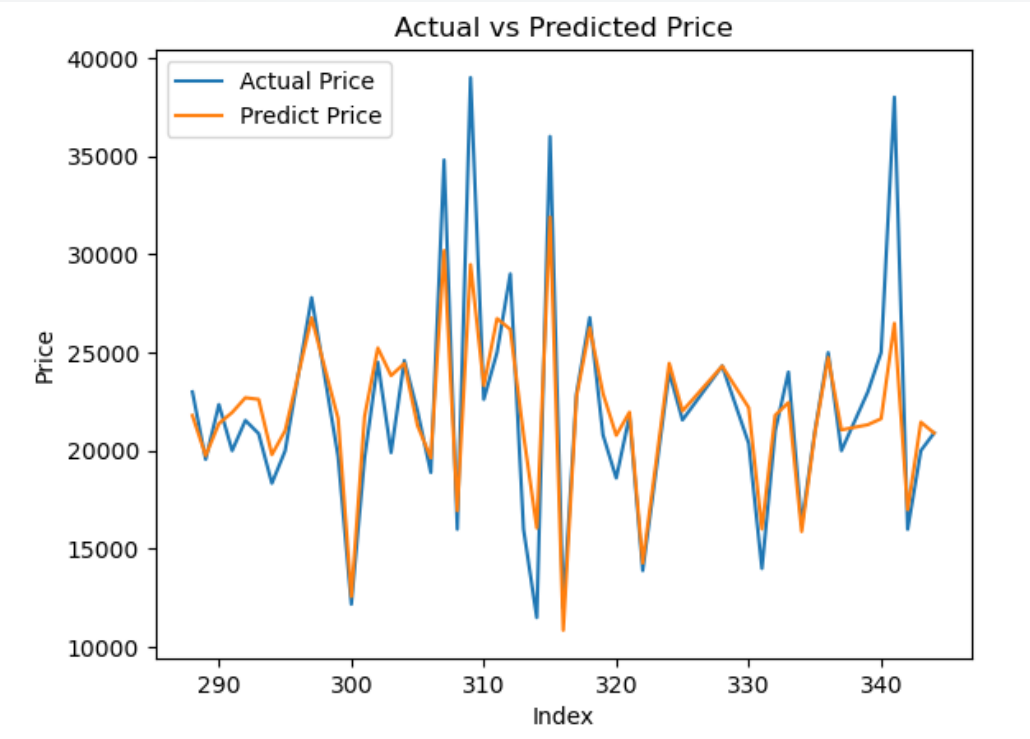
We utilized cross-validation to assess the performance of a model on unseen data. The data was divided into parts, with 5-fold cross-validation being employed in our case. The model was tested on different parts, providing an evaluation of its performance. Cross-validation aids in model selection, parameter tuning, and mitigating overfitting. Additionally, with the removal of 30 percent of the data, we can observe an improvement in the cross-validation results as well.

```
##### Cross-Validation for 70%: #####
```

```
Cross Validation Validity Values: [0.85723715 0.8957893 0.898146 0.90846899 0.9003063 ]
```

# CONCLUSION

*In conclusion: We are assuming that we managed to achieve our goal based on our research and machine learning algorithm we are able to predict the correct and fair price for our Toyota cars based on mechanical parts and price-dependable features. This gives us various functions for example to indicate a correct price for a potential buyer. Thank you for your attention and have a great day :)*



model	price	mileage	Prediction of price	Difference
Camry	18182	103386	18869	-687
Corolla	21950	21320	23648	-1698
Corolla	17749	57515	20045	-2296
Corolla	18979	62134	18963	16
Corolla	18999	57144	20907	-1908
...	...	...	...	...
Corolla	18999	67945	16644	2355