

# @SGCarPrice

By: Asyraf Taha



# 01

Problem  
Statement

# 02

Background of  
Data

# 03

Exploratory Data  
Analysis

# 04

Model  
Selection,  
Tuning &  
Validation

# 05

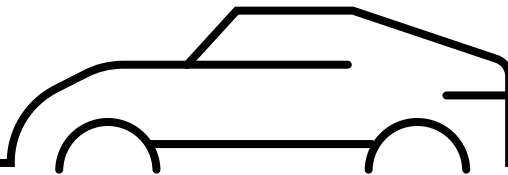
Try it!

# 06

Conclusion

# Problem Statement

\$????



Uncertainty over prices of their vehicle leaves car owners guessing an acceptable selling price. Another way to know the price is to get quotations from car dealers, which leaves car owners with no control nor bargain over their vehicle.

**Car Owners**  
Car wanting to sell their cars have two options,  
Sell Directly or Sell to Dealers



### Consignment / Sell to Dealers

Dealers buy off cars at a lower price and markets them to sell.

### Successful Sale

New owner purchases the car.

### Direct Selling

Involves marketing and communicating sales to 3<sup>rd</sup> parties.

Knowing the market value allows car owners to weigh the trade off between selling to dealers and selling directly. It gives car owners leverage and understanding of the value their vehicle possesses.

That is the aim of this project.

# Data Background

## Used Cars for Sale (29,333 listings)

### Kia Cerato Forte Koup 1.6A SX Sunroof (New 10- COE)



**\$58,600**

(70% loan at \$599/mth)

26-Jan-2011 (10yrs COE left)



140,000 km

DEPRE \$5,860 / yr

Golden Charter Pte Ltd

Bank Promo 2.88%! Showroom Condition! Warranty Provided, Buy With Confidence

### Volkswagen Golf 1.4A TSI (COE till 02/2025)



**\$32,000**

SGCarmart.com is Singapore's top car listing page in Singapore with over 30,000 listed cars on the website.

It has a mix of direct sellers and car dealers listing on the website. Gives an accurate picture of the Used Cars Market.

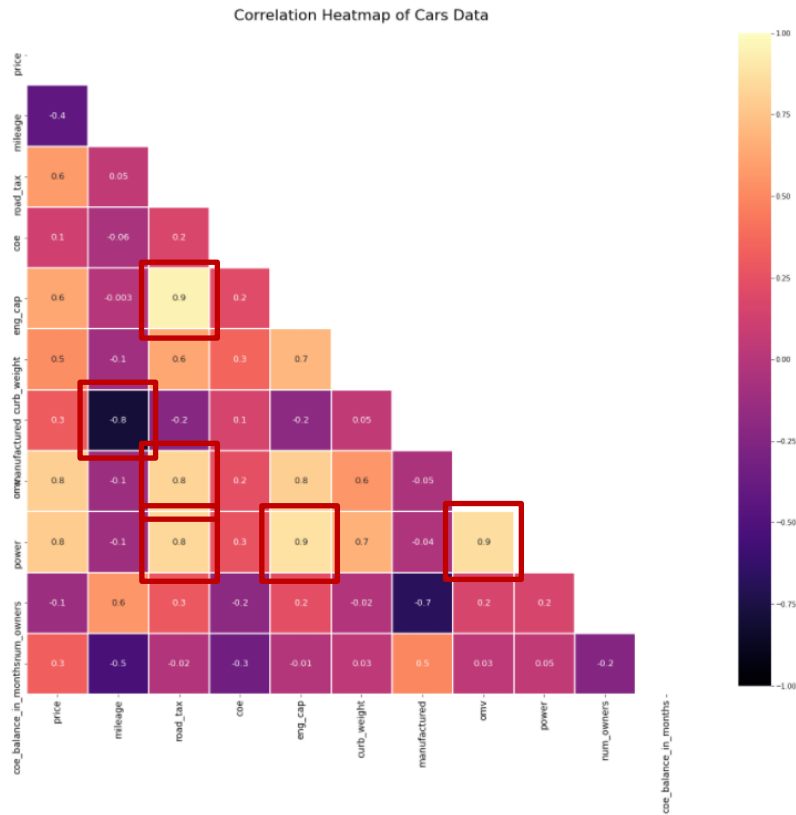
Data was scraped from here.

# Data Dictionary

Feature	Type	Description
model	object	Make & Model of the car
price	object	Price of Cars Listed
mileage	integer	The total distance the car has travelled
road_tax	integer	The amount of tax payable for a vehicle to use the roads in a year
coe	integer	The price of COE (latest) that was registered to the vehicle : note that COE expires every 10 years and subject to renewal or termination
eng_cap	integer	The engine capacity of a car
curb_weight	integer	The weight of the vehicle
manufactured	integer	The year the car was manufactured
transmission	object	Gear transmission of the car: Automatic or Manual
omv	integer	The machine price of the car (coe not included)
power	integer	Engine power
num_owners	integer	Toal number of owners (past & present): [1: 1 owner, 2: 2 owners, 3: 3 owners, 4: 4 owners, 5: 5 owners, 6: 6 owners, 7: 7 or more owners]
type	object	The vehicle types: [Hatchback, MPV, SUV, Luxury Sedan, Mid-Size Sedan, Sports Car, Stationwagon]
category	object	The categories of cars listing: [Premium Ad Car, COE Car, PARF Car, Almost New Car, Direct Owner Sale, STA Evaluated Car, sgCarMart Warranty Cars, Consignment Car, Rare & Exotic, Low Mileage Car]



# Pairwise Multi-Collinearity



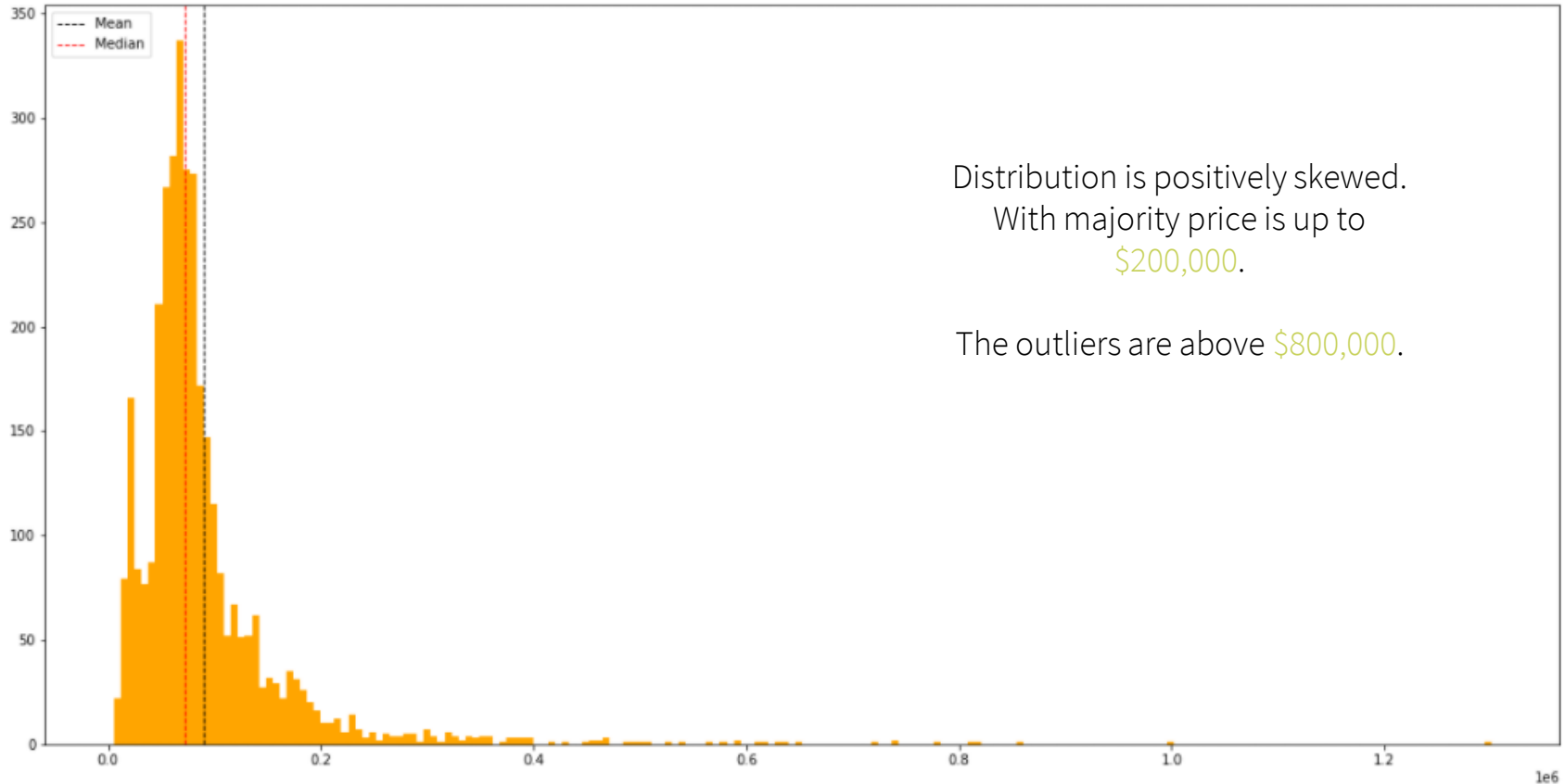
Quick plot of the correlation heatmap shows the existence of Pairwise Multi-Collinearity between some numeric features.

Features such as **manufactured year**, **engine capacity**, **open market value**, **power** and **road tax** show signs of the multi-collinearity with correlation of **+/- 0.8** and above.

Dropping ones with lower correlation to price.



# Price Distribution



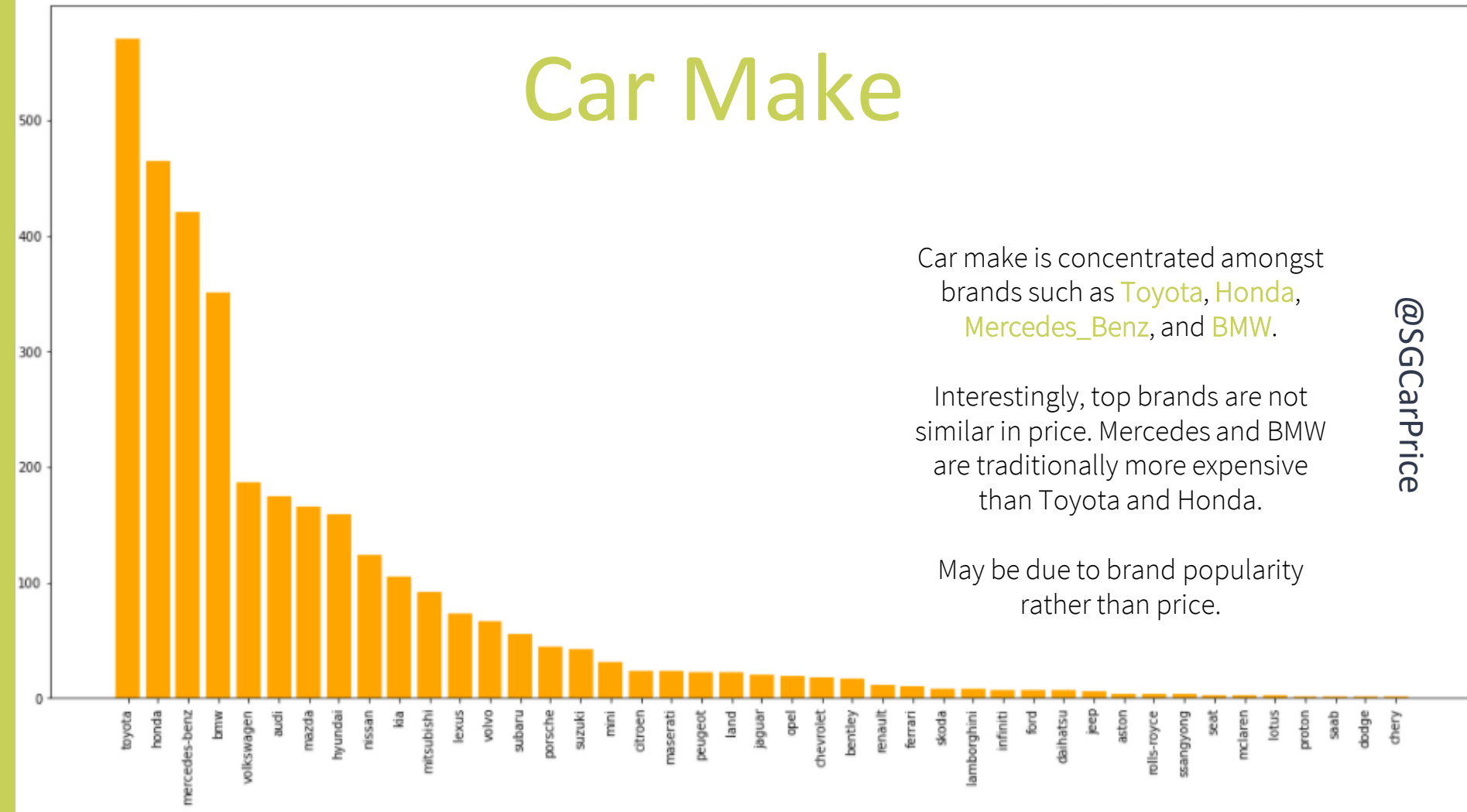
# Car Make

@SGCarPrice

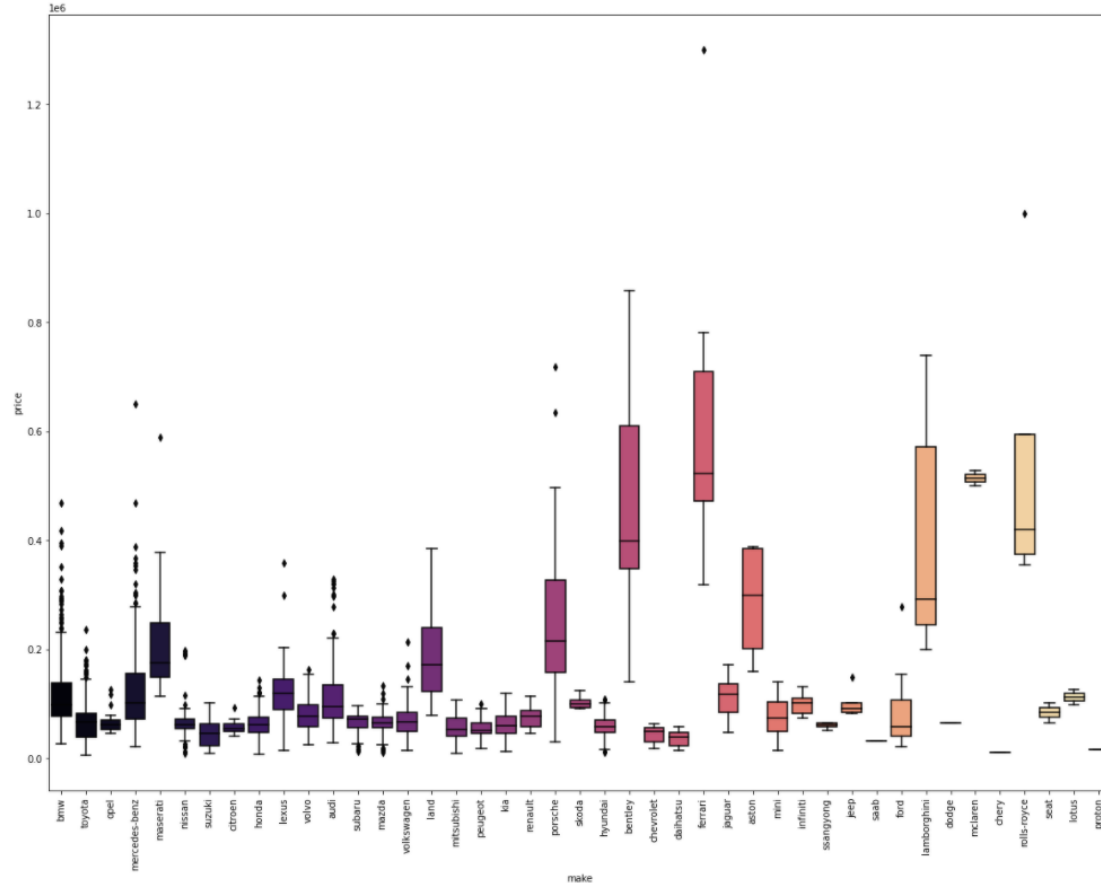
Car make is concentrated amongst brands such as **Toyota**, **Honda**, **Mercedes-Benz**, and **BMW**.

Interestingly, top brands are not similar in price. Mercedes and BMW are traditionally more expensive than Toyota and Honda.

May be due to brand popularity rather than price.

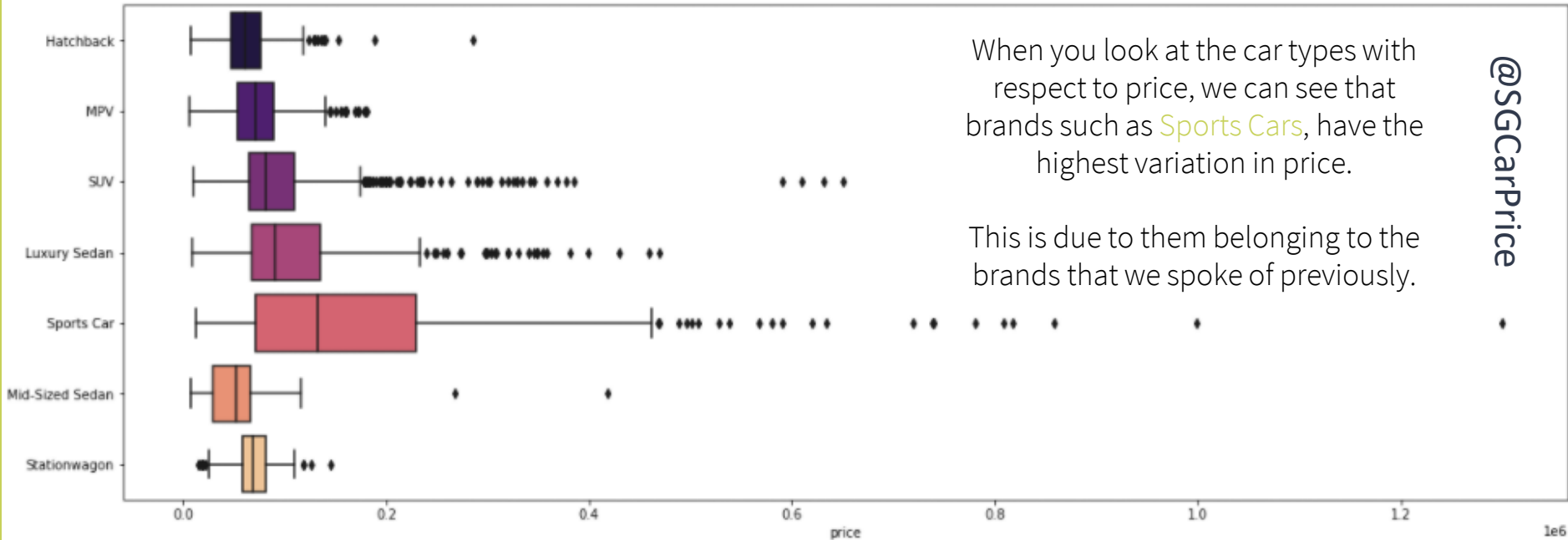


# Car Make



When you look at the distribution of car make with respect to price, we can see that brands such as **Ferrari**, **Bentley**, **Lamborghini**, **Rolls-Royce**, **McLaren**, **Porsche** and **Aston Martin**, are small in number but high in price.

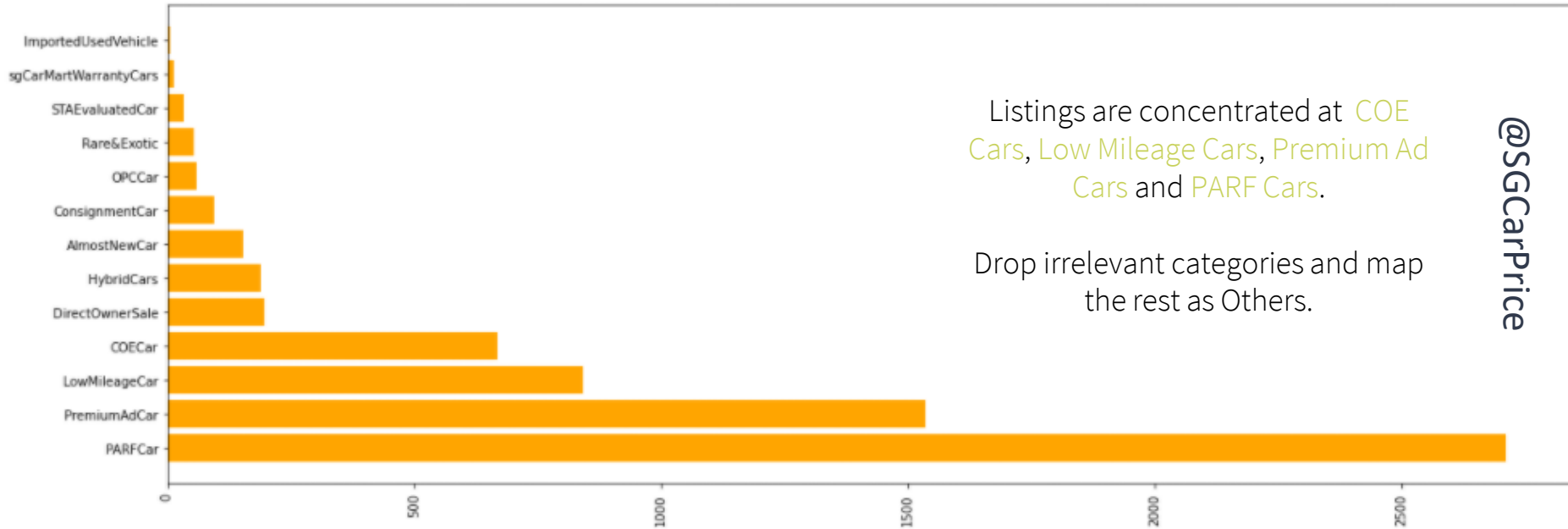
# Car Types



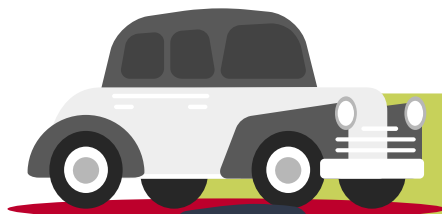
When you look at the car types with respect to price, we can see that brands such as Sports Cars, have the highest variation in price.

This is due to them belonging to the brands that we spoke of previously.

# Car Categories



# Initial Model



Linear

Train R Score: **0.8066**  
CV Score: **0.8055**

Test R Score: **0.8168**  
RMSE: **38,464**



Lasso

Train R Score: **0.8066**  
CV Score: **0.8057**

Test R Score: **0.8168**  
RMSE: **38,470**



Gradient  
Boost

Train R Score: **1.0**  
CV Score: **0.9145**

Test R Score: **0.8748**  
RMSE: **31,801**

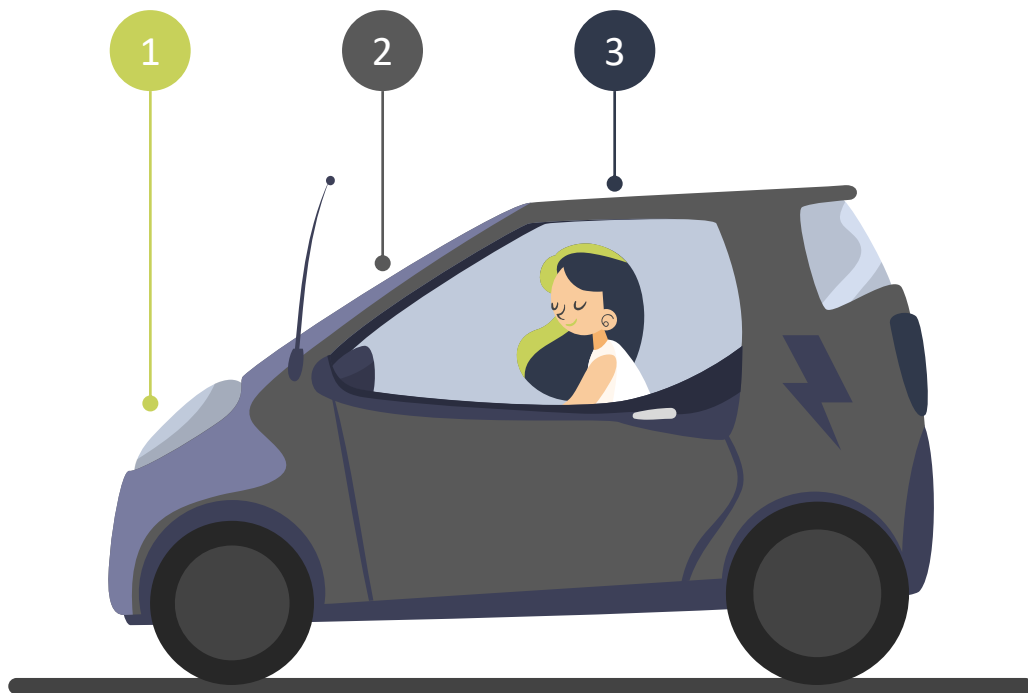


Random  
Forest

Train R Score: **0.9854**  
CV Score: **0.9134**

Test R Score: **0.9054**  
RMSE: **27,636**

# Tuning it up



1

## Limiting Price

To 2X the median  
(\$143,600)

2

## Limiting Mileage

To 2X the median  
(138,000km)

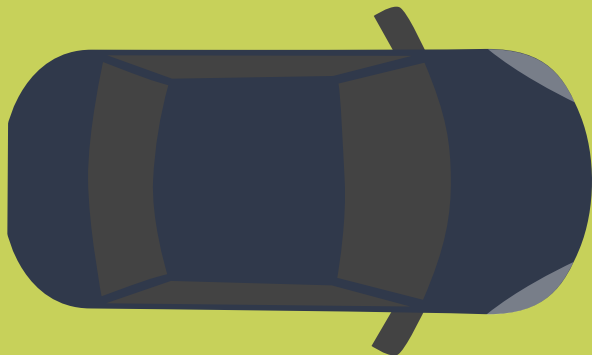
3

## Tune Model Parameters

Increase depth and  
estimators



# Model Head to Head



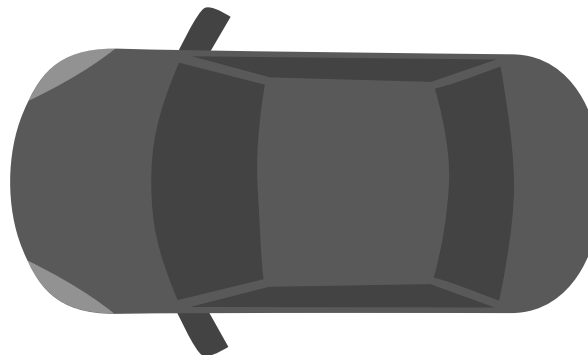
Gradient Boost

Train R Score: **0.9999**

Test R Score: **0.9432**

CV Score: **0.9287**

RMSE: **6,587**



Random Forest

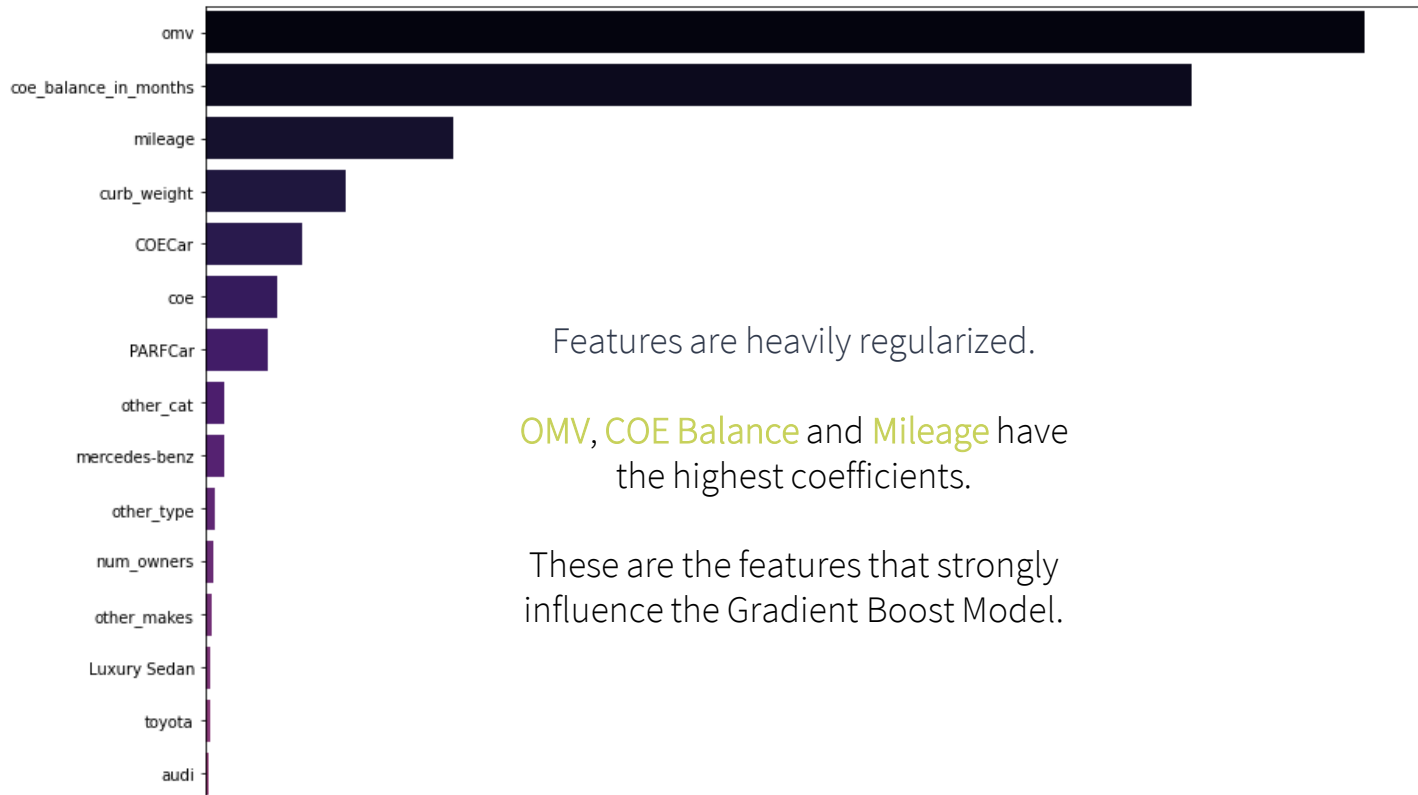
Train R Score: **0.9895**

Test R Score: **0.9311**

CV Score: **0.9135**

RMSE: **7,253**

# Feature Importance

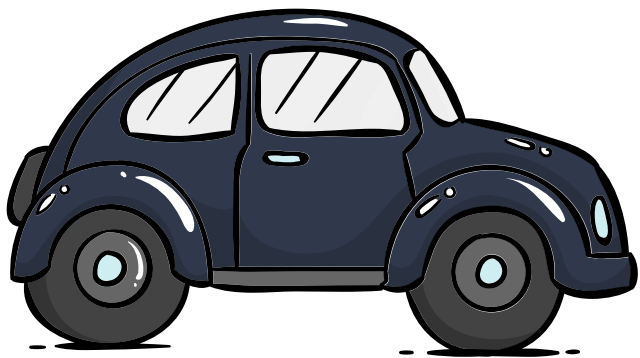


Features are heavily regularized.

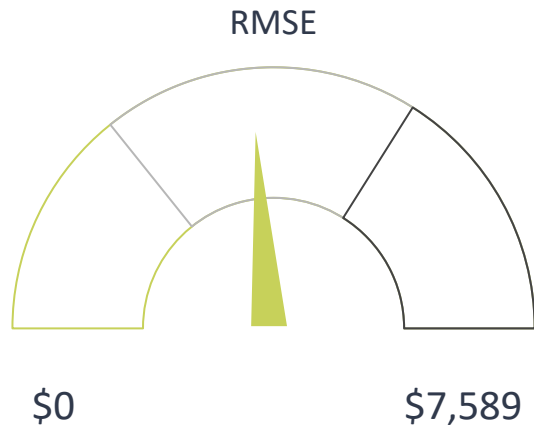
OMV, COE Balance and Mileage have the highest coefficients.

These are the features that strongly influence the Gradient Boost Model.

# Model Validation



Model was used on validated set of data scraped in January 2021.



**92.11%**

Model is able to account for 92.11% of validation data.

# Try it!

Download Telegram

1

2

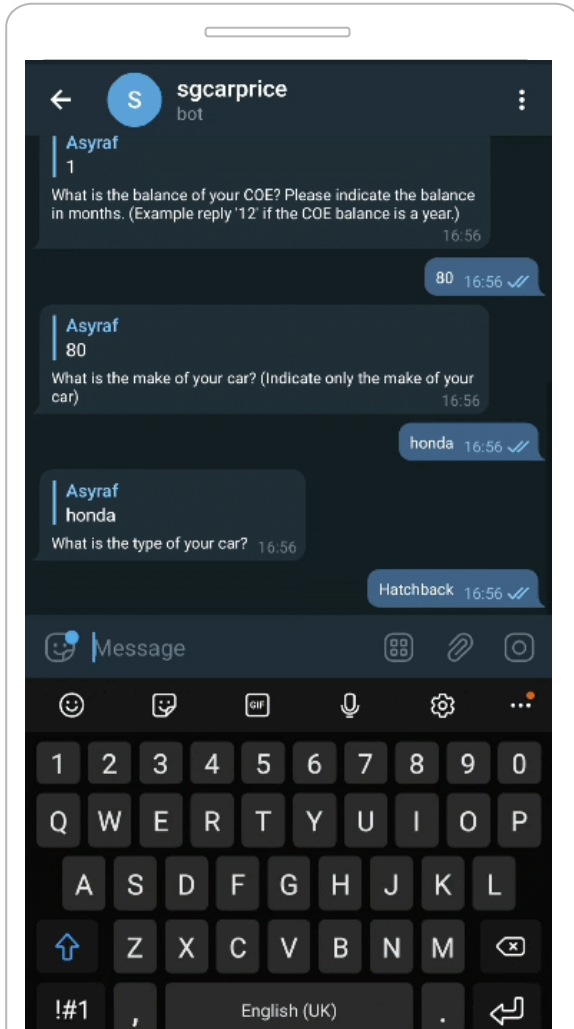
Find @SGCarPrice

Send /start to begin!

3

Note that the bot only is workable during this presentation as it is deployed locally.

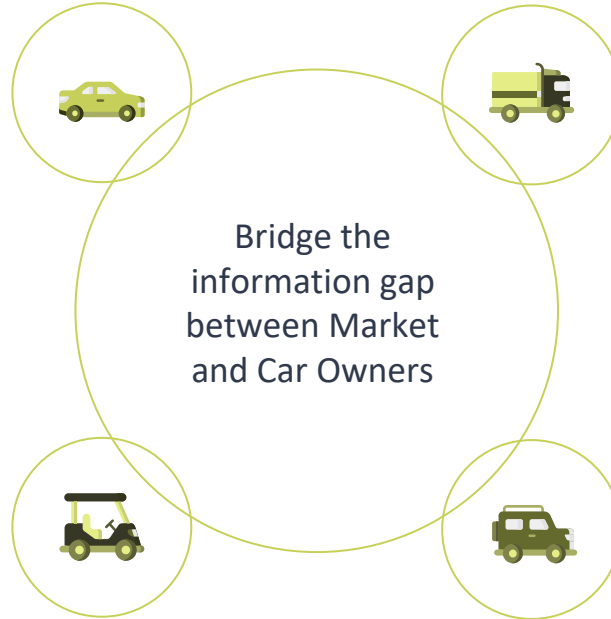
@SGCarPrice



# To Summarise...

**Model Performance**  
Model is able to predict with accuracy of **92.11%** and error of **\$7,589**

**Not effective predicting Luxury Cars**  
Luxury Brands such as Porsche, Ferrari and other Sports Car Brands, are considered outliers and there is not enough data to be able to predict accurately.



**COE Balance and OMV the main Influence**

Despite having so many features, these two features are the main factors influencing your car price.

**Not effective with High Mileage Cars**

Model is only effective for mileage 138,00km. More data is needed.

# Further Research



## More Luxury Car Data

Model could benefit from car listings data with valuation above \$143,600

## More Mileage Data

Model could benefit from car listings with above 138,000km in mileage

## Account for Car Condition

Condition of car or involvement in accidents may affect the price of the car.

## Public Collaboration to Continue to Scrape Data

Car prices are seasonal and requires constant update. Efforts by collaborators may help to improve the model.



Thank  
You



CREDITS: This presentation template was created by [Slidesgo](#), including icons by [Flaticon](#), and infographics & images by [Freepik](#).

**Please keep this slide for attribution.**

