



MOTORBIKE
SELECTION

CAPSTONE

MOTORBIKE RANGE PRESENTATION

Nurul Liyana

Start



WHAT IS IN THIS PRESENTATION



OBJECTIVE



THE DATASET



THE APPROACH



DATA
PREPROCESSING



DATA UNDERSTANDING/
DATA EXPLORATION



TRAINING
MODELS



MODEL
EVALUATION



HYPERPARAMETER
TUNING



MODEL
PERFORMANCE



CONCLUSION

OBJECTIVE

- To assist with a client of mine who wants to open a small motorbike shop.
- Data from SGBike Mart
- Supporting data from:
 - Carousell
 - Kaggle
 - Data.gov
 - Bikemart.sg

Next





Motor name	Motor Price	Brand	Engine capacity	Classification	Vehicle Type	Model	Year	Category	Power (hp)	Engine cylinder	Engine stroke	Gearbox	Fuel capacity (lts)	Transmission type	Front brakes	Rear brakes	Front suspension	Rear suspension
Honda VFR800X CrossRunner	16800	honda	782	Class 2	Sport Tourers	vfr800x crossrunner	2015	Sport touring	107.0	V4	four-stroke	6-speed	20.8	Chain	Double disc. ABS. Floating discs. Hydraulic. F...	Single disc. ABS. Hydraulic. Two-piston calipe...	43mm HMAS cartridge-type telescopic fork with ...	Pro-Link with gas-charged HMAS damper, 7-step ...

Null values

```

1 Motor name          0
2 Motor Price        0
3 Brand              0
4 Engine capacity    0
5 Classification     0
6 Vehicle Type       0
7 Model              0
8 Year               0
9 Category            0
10 Power (hp)         1094
11 Engine cylinder    0
12 Engine stroke      0
13 Gearbox             209
14 Fuel capacity (lts) 149
15 Transmission type  148
16 Front brakes        14
17 Rear brakes          15
18 Front suspension    118
19 Rear suspension      109
20 dtype: int64

```

dtypes

```

1 Motor name          object
2 Motor Price         object
3 Brand              object
4 Engine capacity    int64
5 Classification     object
6 Vehicle Type       object
7 Model              object
8 Year               int64
9 Category            object
10 Power (hp)          object
11 Engine cylinder    object
12 Engine stroke      object
13 Gearbox             object
14 Fuel capacity (lts) float64
15 Transmission type  object
16 Front brakes        object
17 Rear brakes          object
18 Front suspension    object
19 Rear suspension      object
20 dtype: object

```

THE DATASET

www.reallygreatsite.com

22 columns
4127 rows



Next Slide



THE APPROACH

Regression Models:

1. Linear Regression
2. Logistic Regression
3. Ridge Regression
4. Lasso Regression
5. Random Forest Regression
6. Decision Tree Regression

Cross Validate:

1. Grid Search CV
2. Random Search CV

Next



DATA PREPROCESSING

1. Drop duplication rows
2. Fill in missing values
3. Reformat values(eg Yamaha XSR 700 to 'Yamaha' and 'xsr 700')
4. Merging supporting datasets to form one big merging data
5. Reformat dtypes and creating dummies for columns that is not numerical

Next



DATA UNDERSTANDING/ DATA EXPLORATION

Exploratory Data Analysis

Next





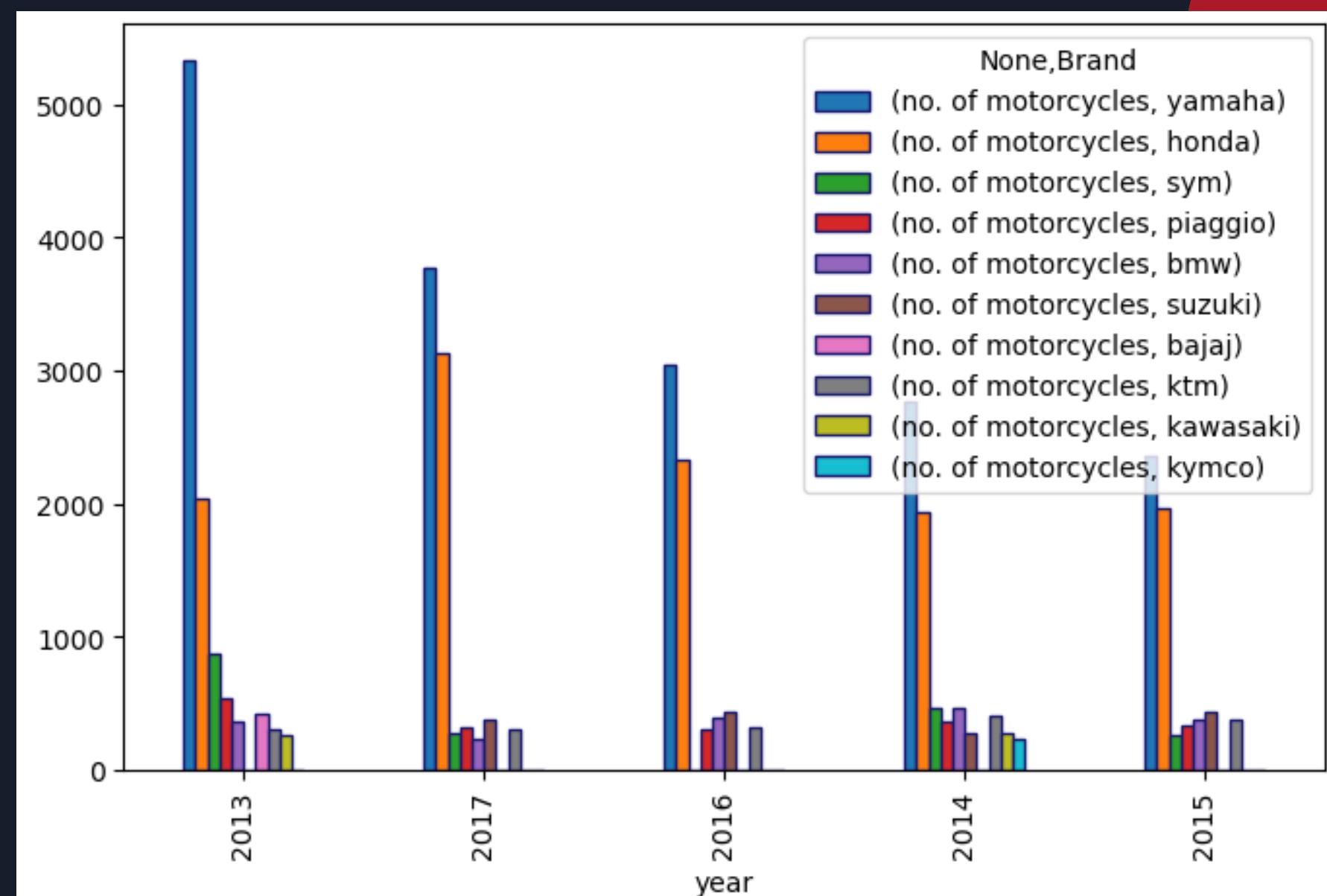
NUMBER OF MOTORCYCLES REGISTERED

DESCRIPTION

Top 5 Brands that is registered in Singapore from 2013 - 2017.

	2013	2017
Yamaha:	5,334	3,764
Honda:	2,046	3,125
Suzuki:	163	387
KTM:	310	312
Piaggio:	542	322

The registrations are usually done by the motorcycle shop where they register the rider's purchased motorcycle in order for them to use it on the road.





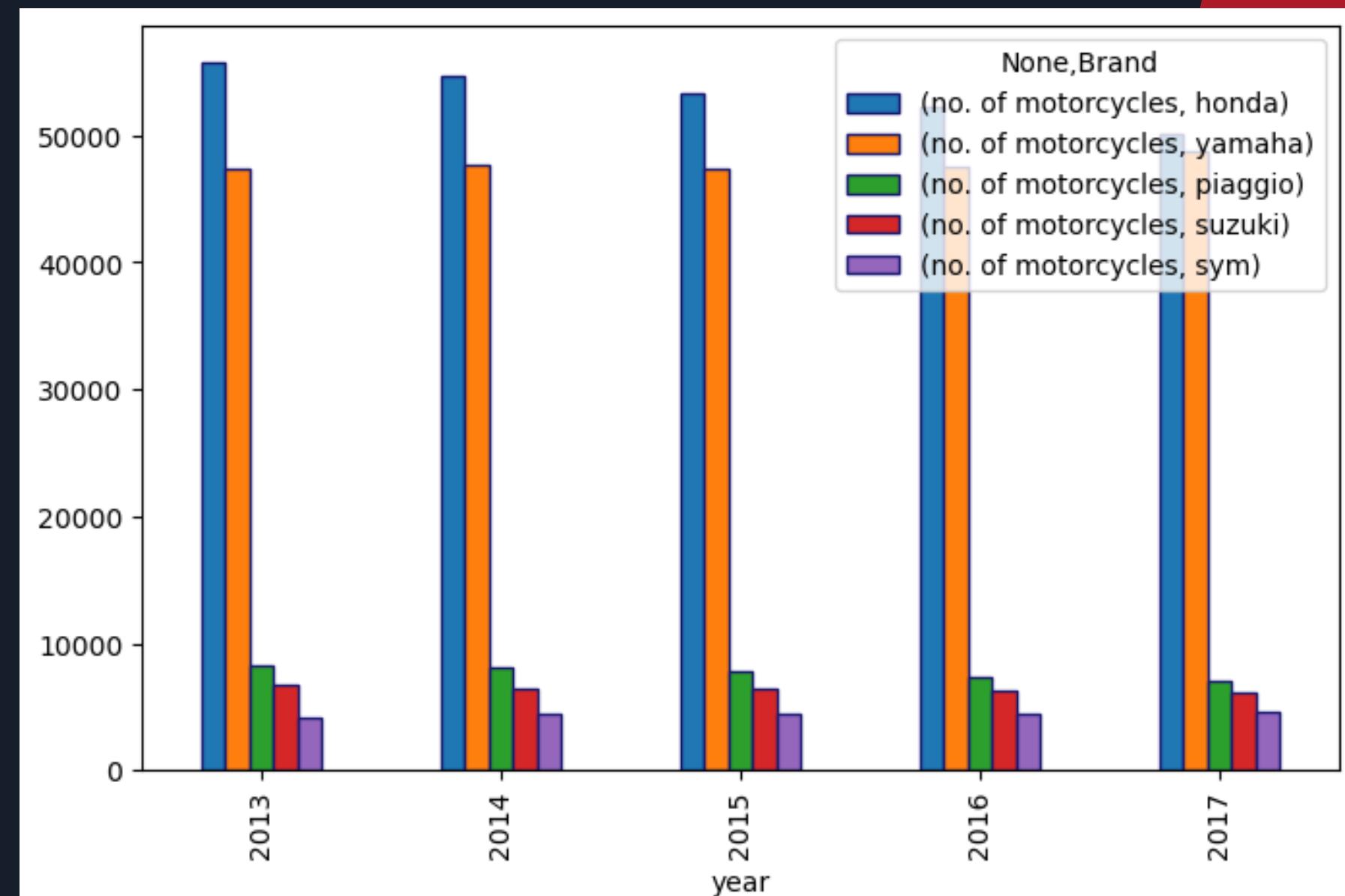
POPULATION OF MOTORCYCLES BY BRAND

DESCRIPTION

Total number of motorbike riders based on Brands they are using from 2013-2017.

	2013	2017
Honda:	55,797	50,118
Yamaha:	48,664	47,390
Piaggio:	8,298	6,962
Suzuki:	6,771	6,076
SYM:	4,142	4,598

Majority motorbike riders in Singapore are using Honda and Yamaha. And the remaining riders, using Piaggio, Suzuki and SYM





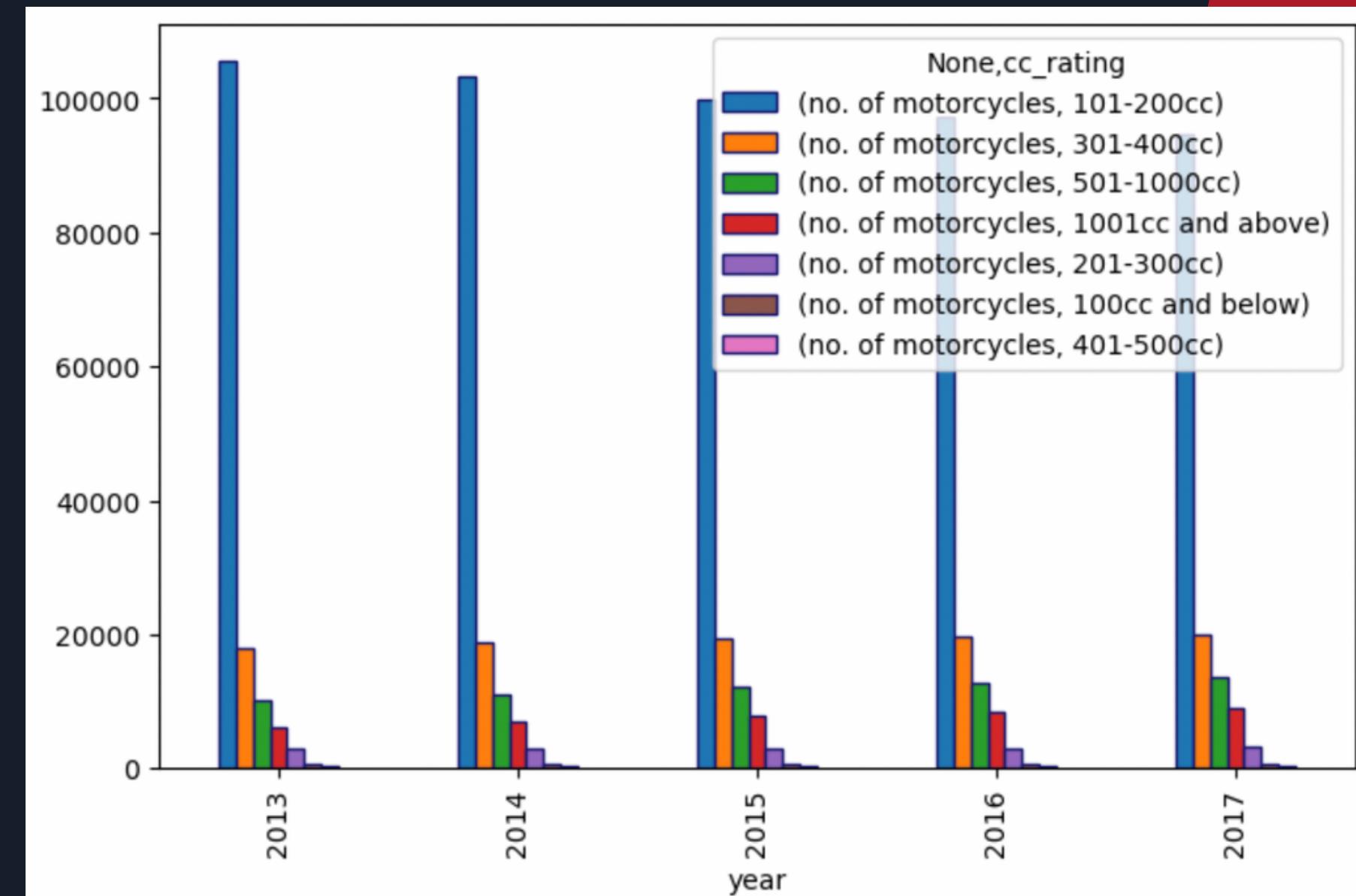
POPULATION OF MOTORCYCLES BY CC RATING

DESCRIPTION

	CC rate
Class 2B	200cc and below
Class 2A.	201cc to 400cc
Class 2	Above 400cc

	2013	2017
101cc to 200cc:	105,735.	94,609
301cc to 400cc:	17,993	19,937
501cc to 1000cc:	10,246	13,585
1001cc and above:	6,175	8,881

201-300cc and 401-500cc , they are the lower range of their individual bike range, so why choose the smaller cc of the range when u can just get the bigger cc of the same range



REASONS WHY RIDERS CHOSE THESE BRANDS POPULARITY

Buys similar motorbikes as friends/family members
and can be seen everywhere you go



YAMAHA AEROX 155

Class 2B



HONDA CB 400

Class 2A

REASONS WHY RIDERS CHOSE THESE BRANDS DAILY USE

Bikes that can easily move around with minimal specifications



YAMAHA XMAX 250-300

Class 2A



YAMAHA SNIPER T135

Class 2B

REASONS WHY RIDERS CHOSE THESE BRANDS OUTLOOK

They are at the higher price range compared to similar CC ratings.



HUSQVARNA SVARTPILEN 401

Class 2



DUCATI AGUSTA DRAGSTER RR

Stuntrider

REASONS WHY RIDERS CHOSE THESE BRANDS COLLECTION

Usually vintage or rare bikes



VINTAGE HONDA CB750K

Class 2



VESPA SIDECAR ANNO

Class 2B

REASONS WHY RIDERS CHOSE THESE BRANDS SERVICES

Companies which deliver foods or helps to deliver documents from point A to point B



FULLY ELECTRIC 3 WHEELS

Class 2B



YAMAHA YBR 125

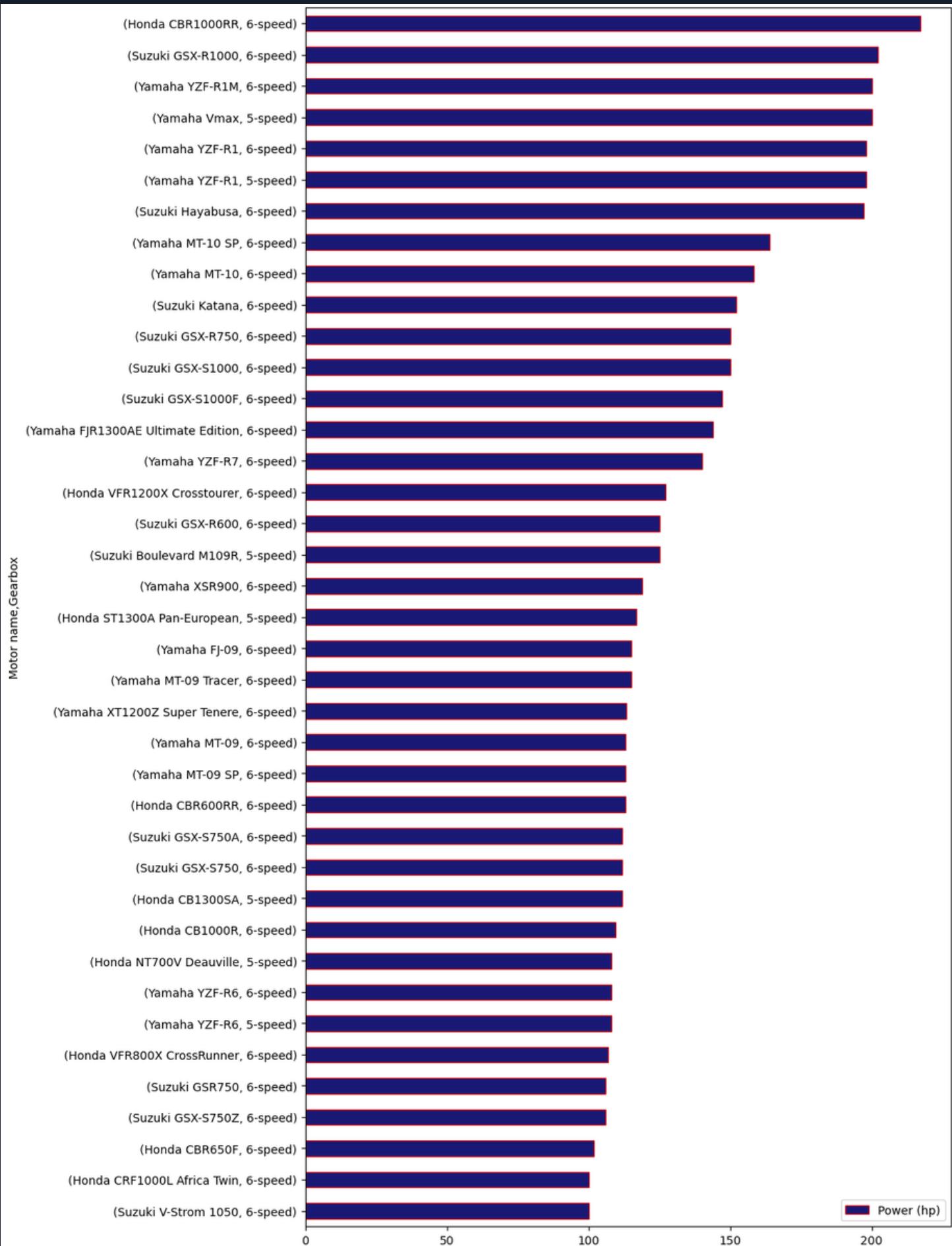
Class 2B



IMPROVE YOUR GENERAL RIDEABILITY

DESCRIPTION

Looking at the mean price for the mid price range, affordable for riders to consider buying from her shop.



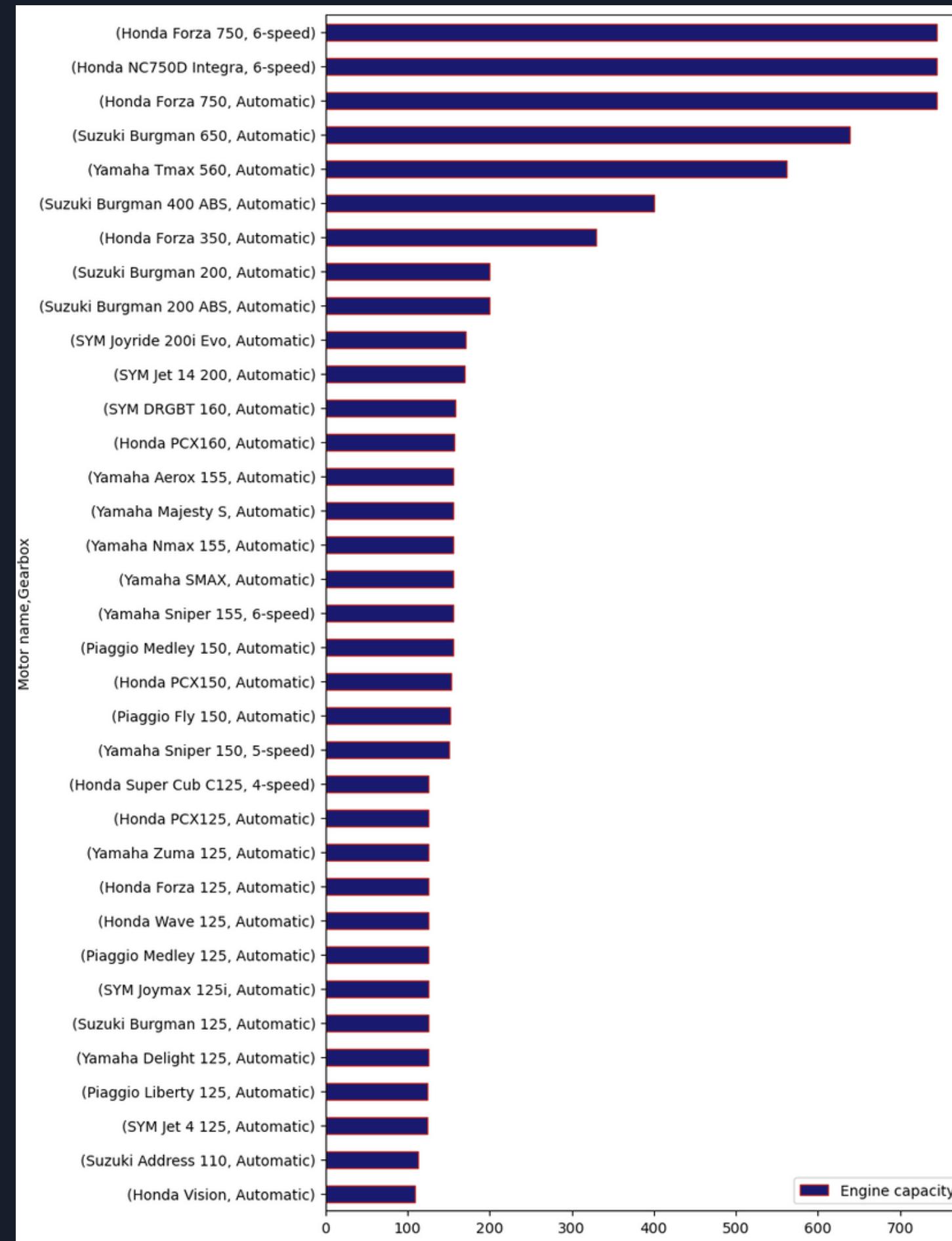
MOTORBIKE
SELECTION



IMPROVE YOUR GENERAL RIDEABILITY

DESCRIPTION

The current trend, popular motorcycle currently is the Scooter as majority of the scooter now are automatic gears, that cut most of the job with clutches and changing of gears.

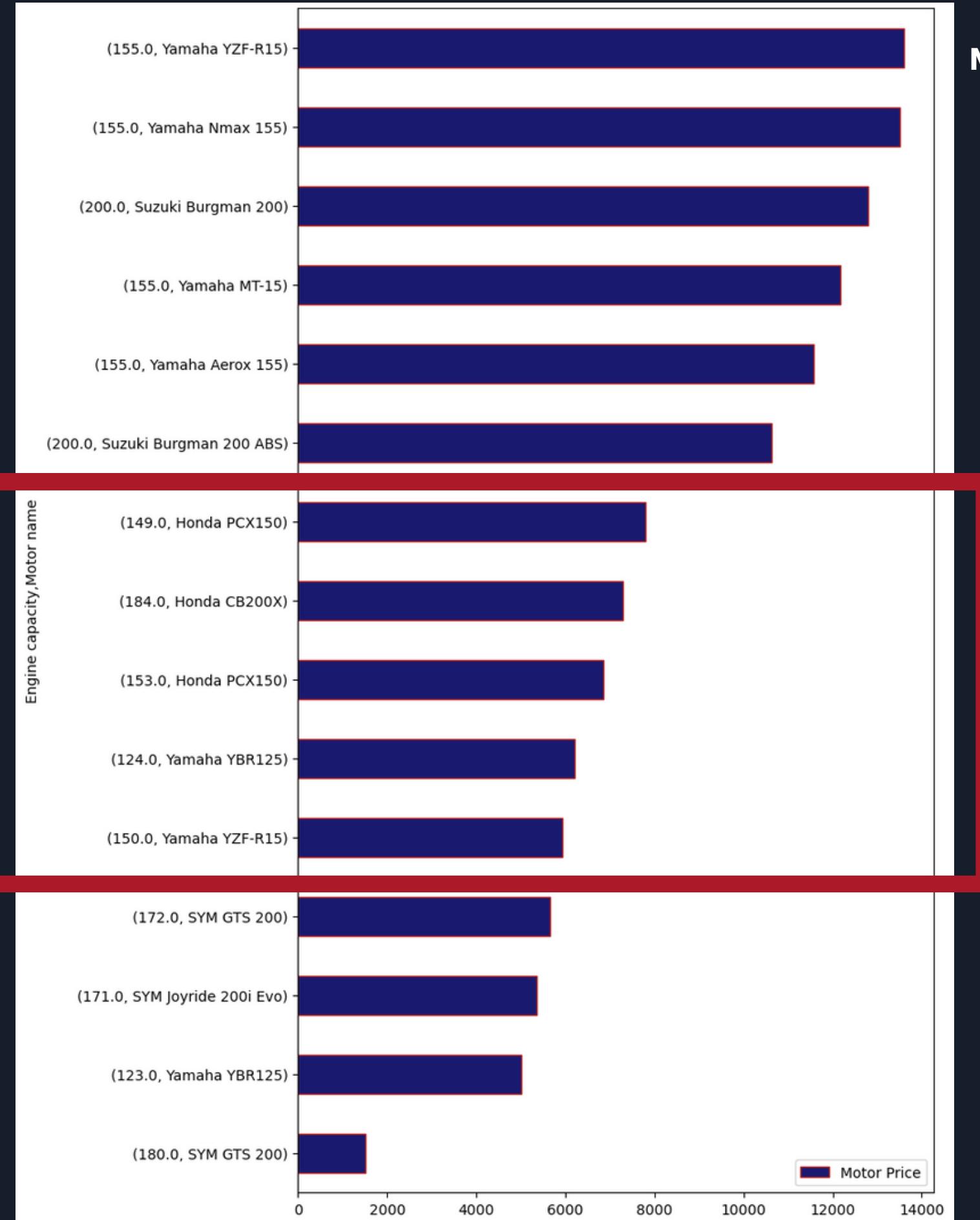




IMPROVE YOUR GENERAL RIDEABILITY

DESCRIPTION

Looking at the mean price for the mid price range, affordable for riders to consider buying from her shop.

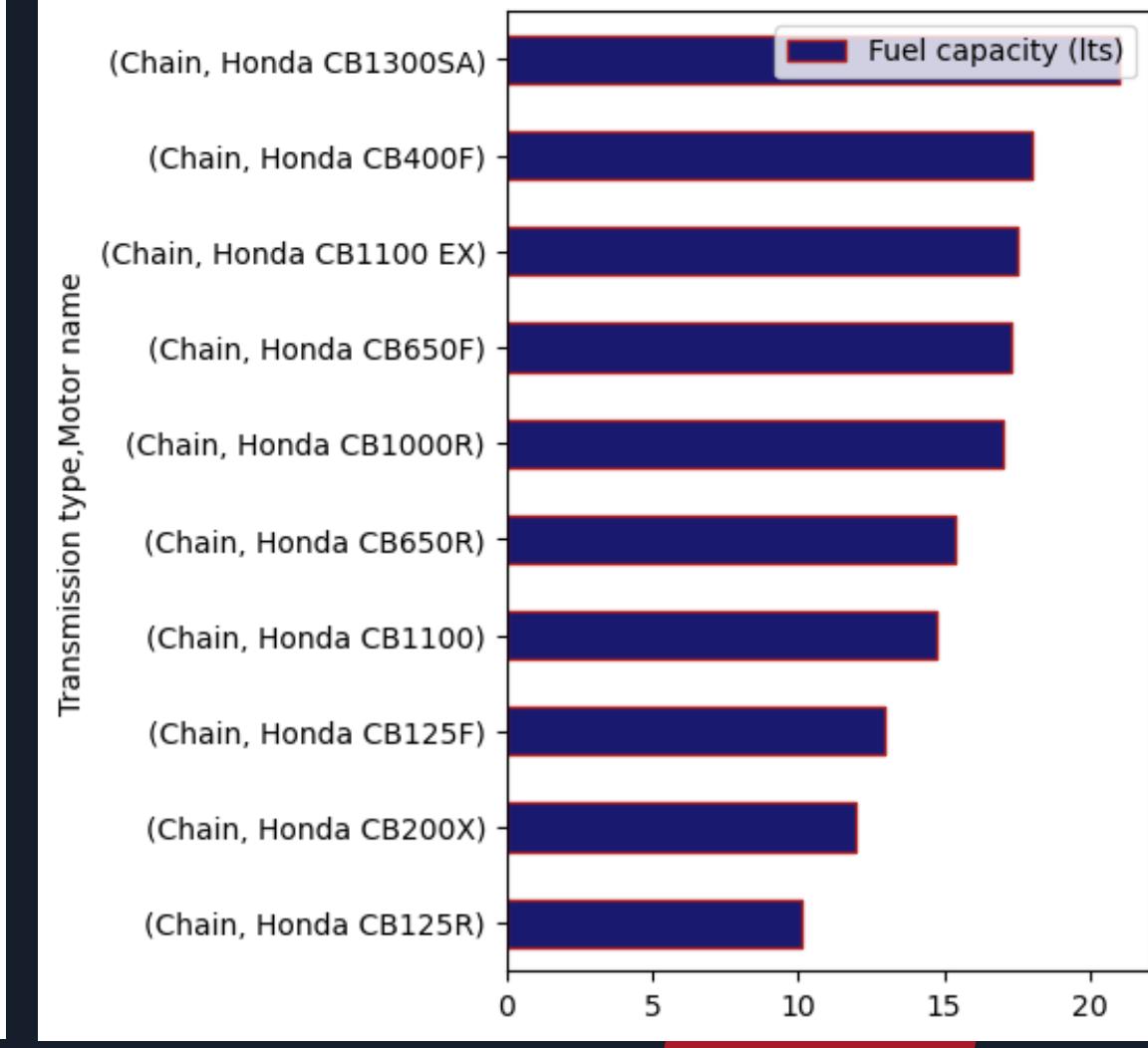
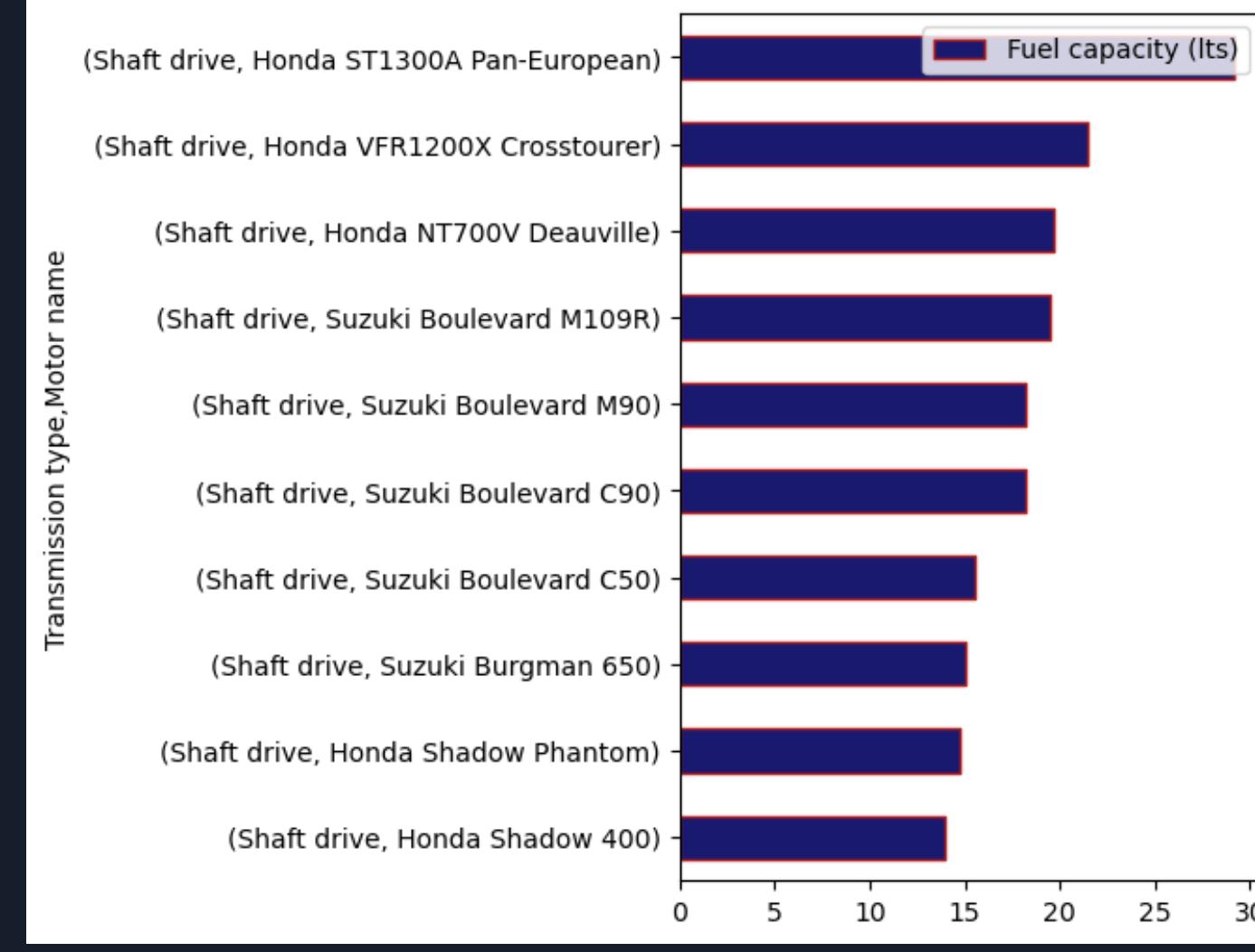
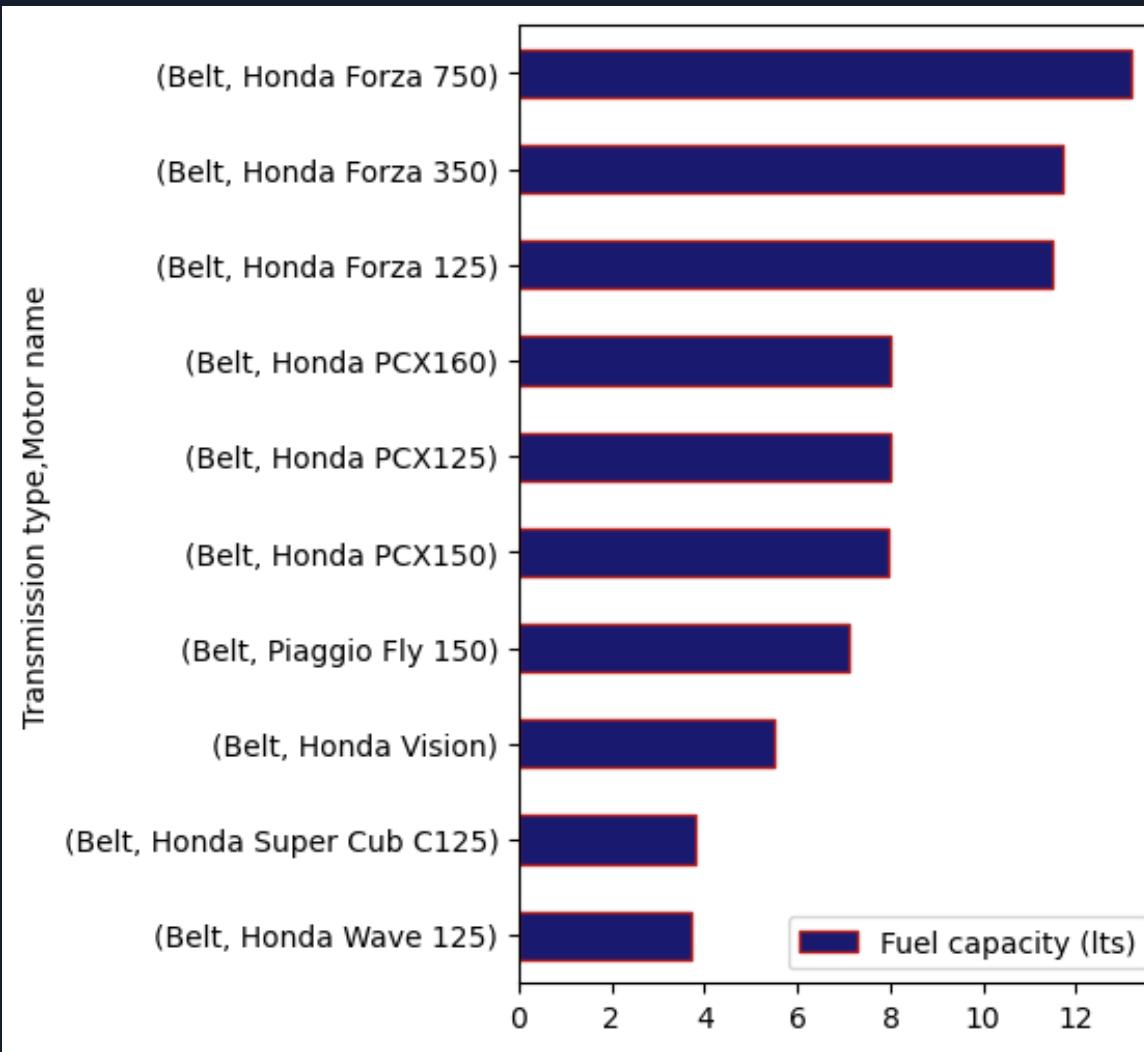




IMPROVE YOUR GENERAL RIDEABILITY

DESCRIPTION

Honda motorcycles has the biggest fuel capacity in the whole range of its transmission type.



TRAINING MODELS

```
X = data.drop(['Motor Price','Year'] , axis=1)  
Y = data["Motor Price"]
```

```
X_train , X_test , Y_train , Y_test = train_test_split(X ,  
Y , test_size = 0.2, random_state = 28)
```

```
model.fit(X_train, Y_train)
```

```
Y_prediction(X_test)
```

```
r2_score(Y_test, Y_prediction)
```

Next



REGRESSION MODEL

R2 SCORE

Random Forest has the highest R2 score with

	Model	R2 Score	Y_Pred
4	Random Forest Regression	6.838392e-01	[6040.320395160396, 8613.58486612516, 12084.40...
2	Ridge Regression	4.613442e-01	[5874.667235261509, 8352.282146566991, 12223.0...
3	Lasso Regression	4.481465e-01	[6087.28882729716, 8290.04835991612, 12163.786...
5	Elastic Net	1.195380e-01	[9157.49334934531, 9465.88891089178, 11247.177...
6	Decision Tree Regression	5.437219e-02	[10791.047619047618, 10791.047619047618, 14929...
1	Logistic Regression	-1.492792e+00	[8800.0, 32000.0, 12000.0, 8800.0, 32000.0, 32...
0	Linear Regression	-1.172806e+16	[6553.136474609375, 8364.540771484375, 12217.8...

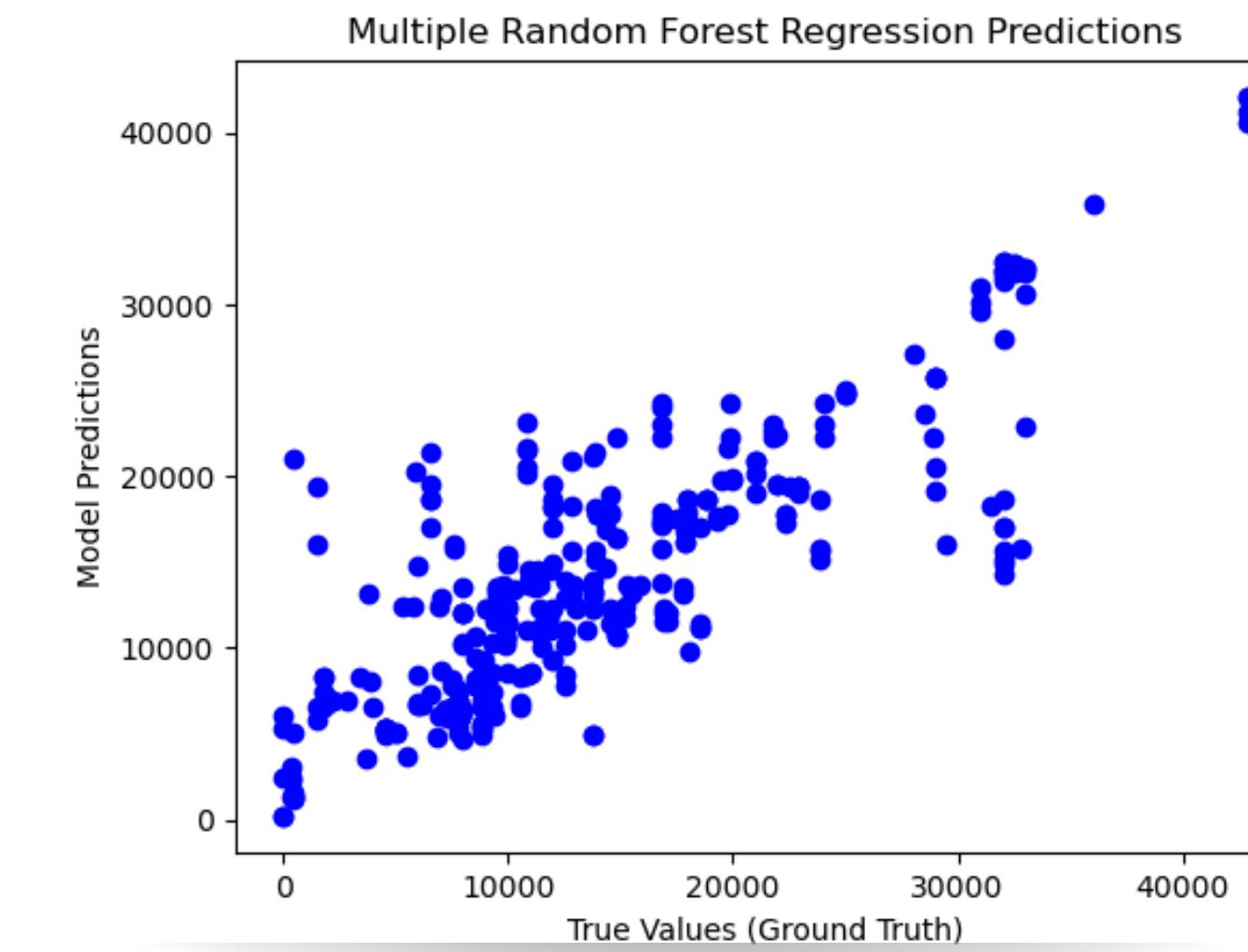
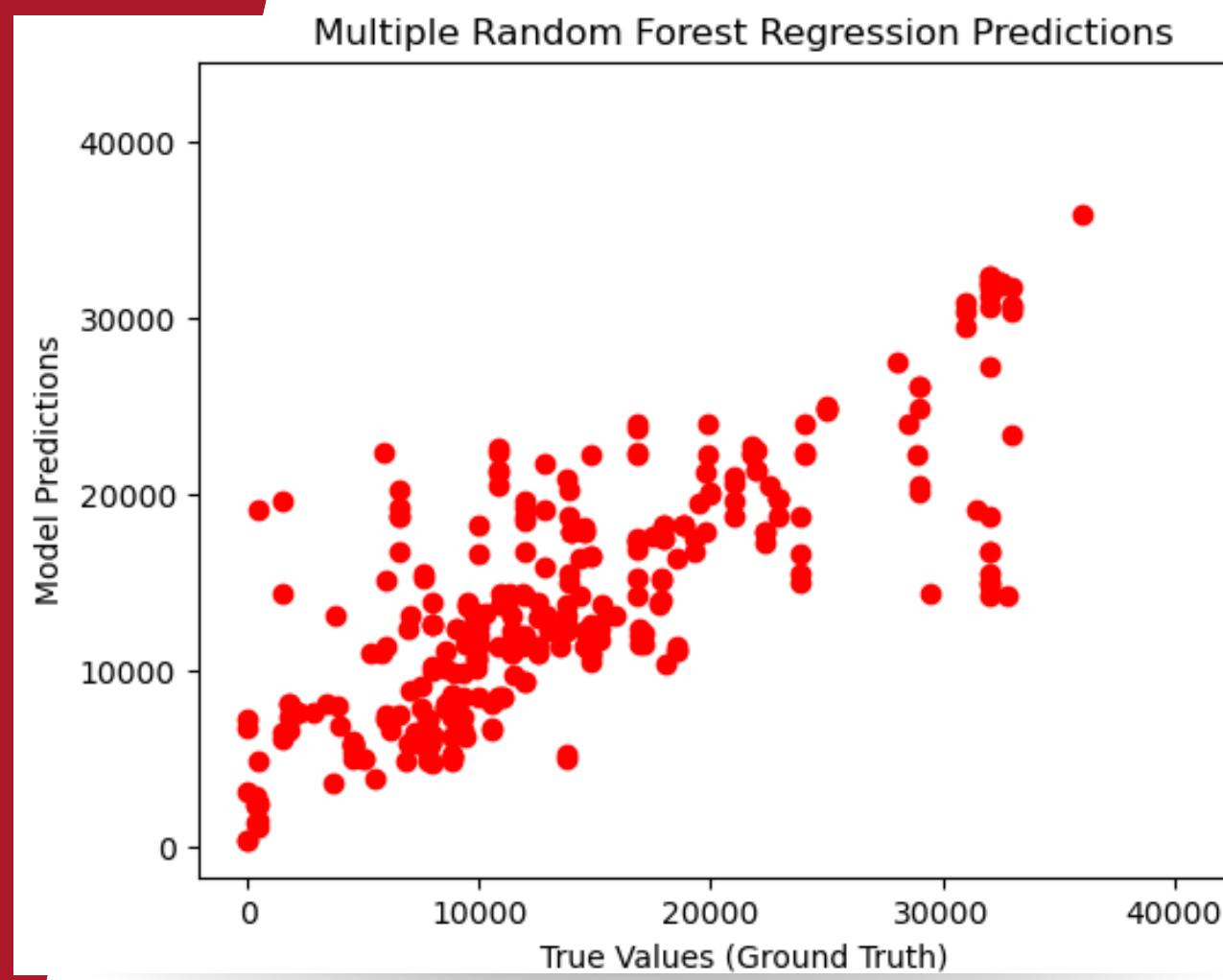
RANDOM FOREST REGRESSION HELPS TO:

1. Predict future prices/costs
2. Predict future revenue
3. Compare performance

ADVANTAGES OF RANDOM FOREST REGRESSION:

1. Extremely high accuracy
2. Scales well
3. Interpretable
4. Easy to use

MODEL PREDICTIONS SCATTER PLOT



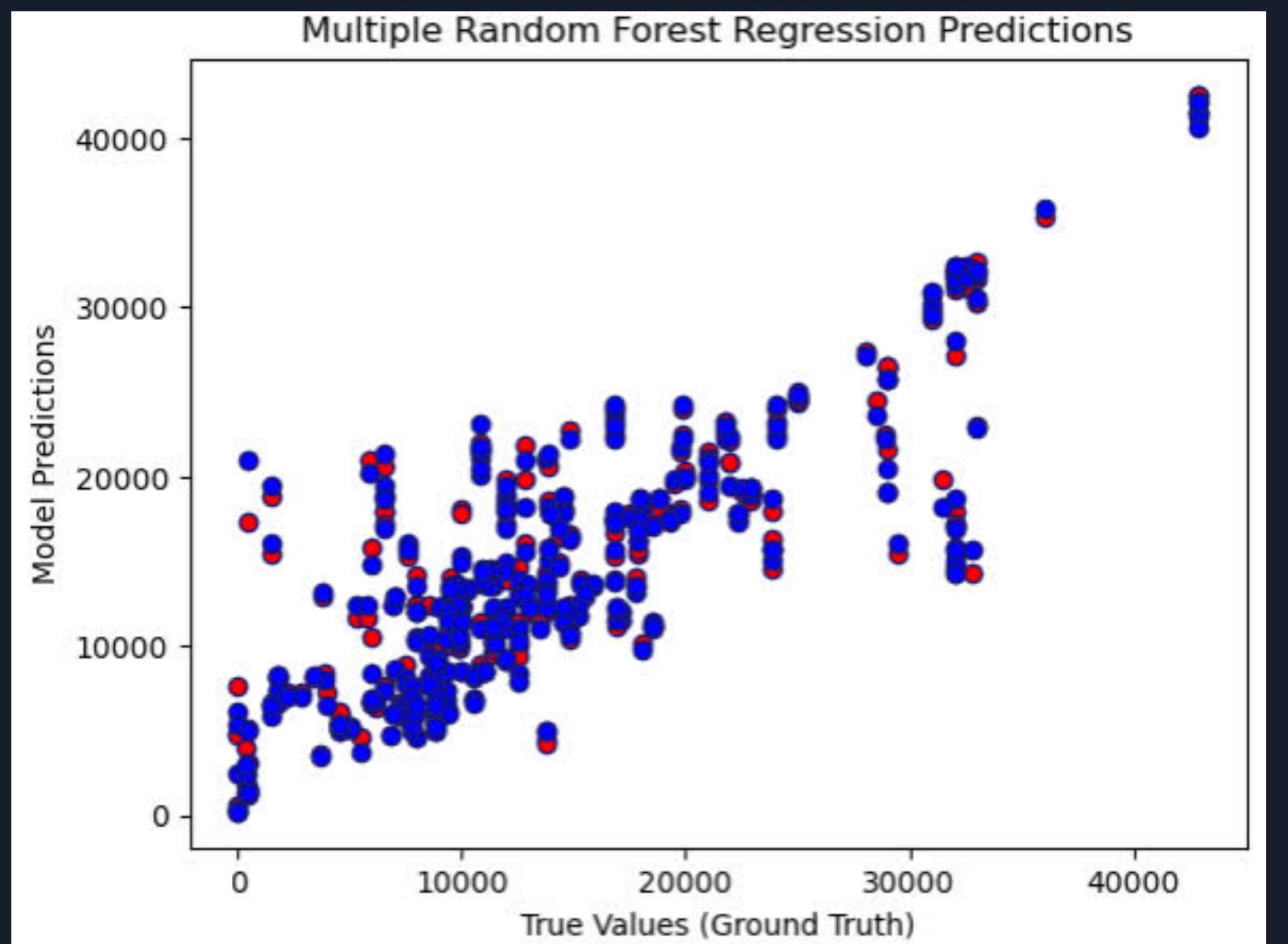
X-AXIS:

Y-test
'True value'

Y-AXIS:

Y_pred
Model predictions

MODEL EVALUATION



HYPERPARAMETER TUNING

- Provide parameters for GridSearchCV and RandomizedSearchCV
- Include Random Forest Regressor with the parameters in the cross validation model.
- Fit the model
- Get the best parameters
- Get R2 score

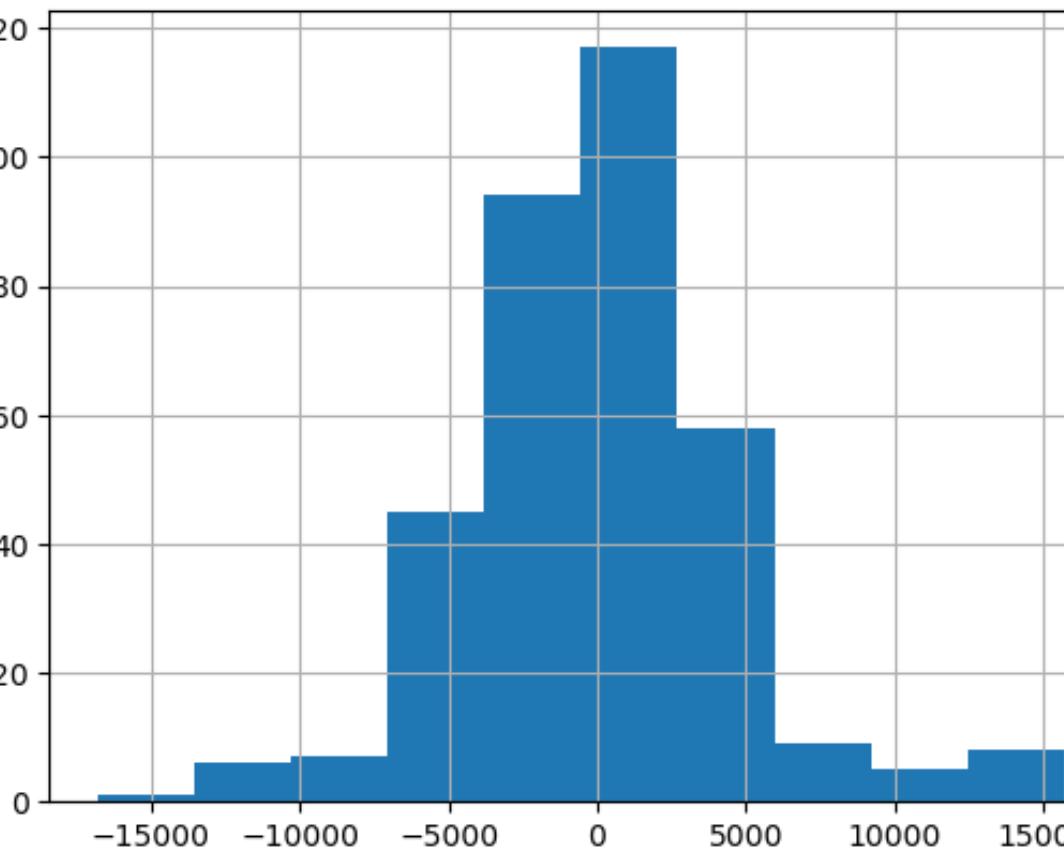
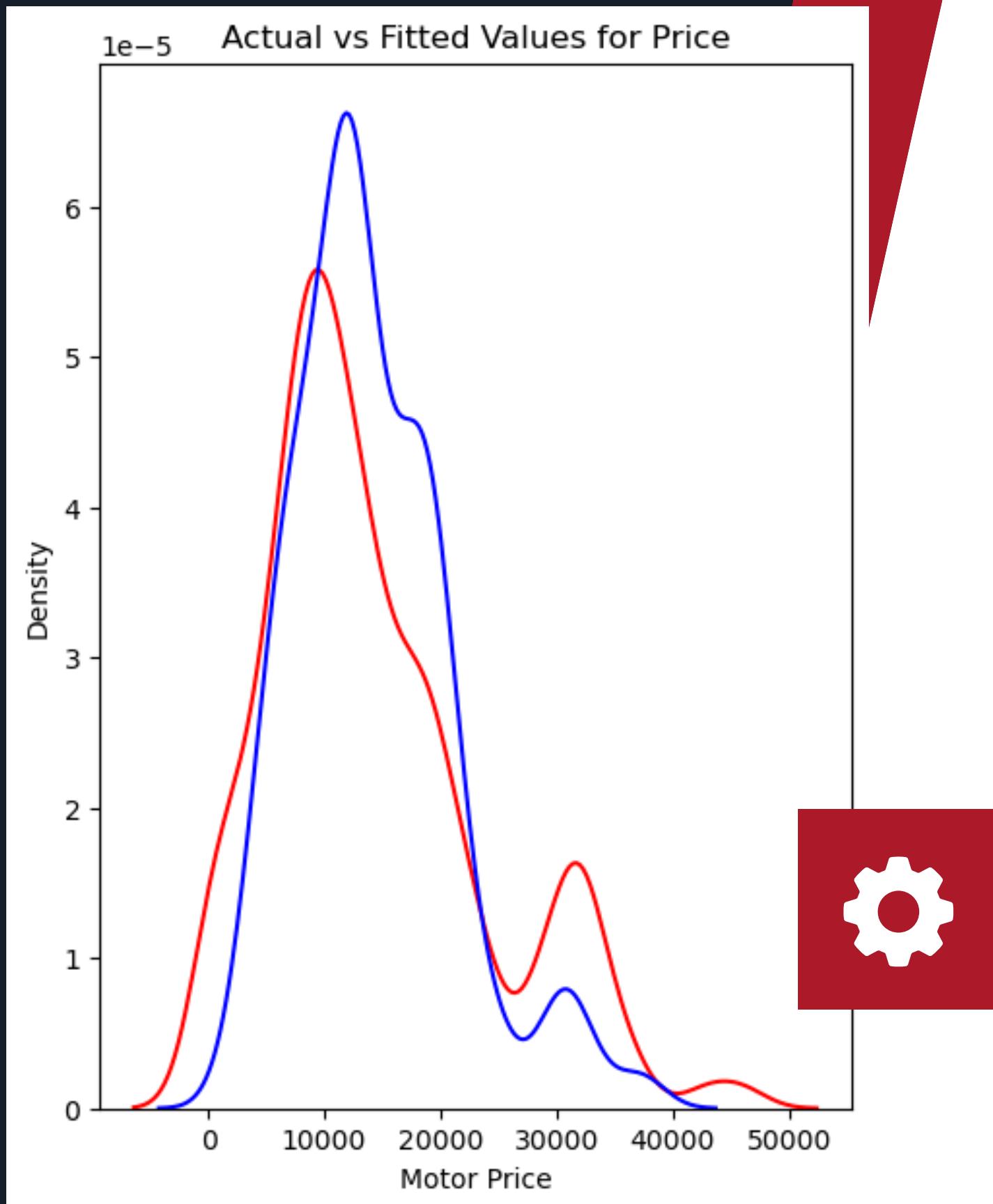


HYPERPARAMETER TUNE WITH GRIDSEARCHCV AND RANDOMIZEDSEARCHCV

Random Forest and RandomizedSearchCV gives the highest combination

	Random Forest and GridSearchCV	Random Forest and RandomizedSearchCV
Best parameters	{'max_depth': 9, 'n_estimators': 300}	{'bootstrap': True, 'max_depth': 10, 'max_features': 664, 'min_samples_split': 4, 'n_estimators': 249}
R2 Score	0.7369348118492023	0.7571842701063927
Percentage	74.0%	76.0%

MODEL PERFORMANCE WITH HYPERPARAMETER TUNING



DESCRIPTION

Mean Absolute Error:
3236.509718707003

Mean Squared Error:
20068072.668873265

Root Mean Squared Error:
4479.740245692072



IN CONCLUSION |

- Bring in The top 2 Brands which is Honda and Yamaha from various classification range.
- Random Forest Regression model fits this dataset on predicting the motorcycle brands and motor model
- Regression in Machine learning is always about predicting what's the next continuous outcome as a target to achieve.
- The only difference between both the approaches is in grid search we define the combinations and do training of the model whereas in RandomizedSearchCV the model selects the combinations randomly.
- Random forest regression and RandomizedSearchCV gives a good output, will want to try with other predictions if given the opportunity.



THANK YOU

NEW GENERATION OF MOTORCYCLES HERE

github link

End Slide

