IBM Coursera Advance Data Science Capstone

by

Ashiqur Rahman Khan

Github: https://github.com/ashiqurrahmankhan21st/CapstoneProject

Table of Content

DataSet

- DataSet Shape:

  column - 26

  row - 203

- Data types:

  float64 - 8, int64 - 8, object - 10

data.head()

	car_ID	symboling	CarName	fueltype	aspiration	doornumber	carbody	drivewheel	enginelocation	wheelbase	...	e
0	1	3	alfa-romero giulia	gas	std	two	convertible	rwd	front	88.6	...	
1	2	3	alfa-romero stelvio	gas	std	two	convertible	rwd	front	88.6	...	
2	3	1	alfa-romero Quadrifoglio	gas	std	two	hatchback	rwd	front	94.5	...	
3	4	2	audi 100 ls	gas	std	four	sedan	fwd	front	99.8	...	
4	5	2	audi 100ls	gas	std	four	sedan	4wd	front	99.4	...	

5 rows × 26 columns

Features

data.info()

as data will be labeled based on the car volume. thats why calculating the car volume.

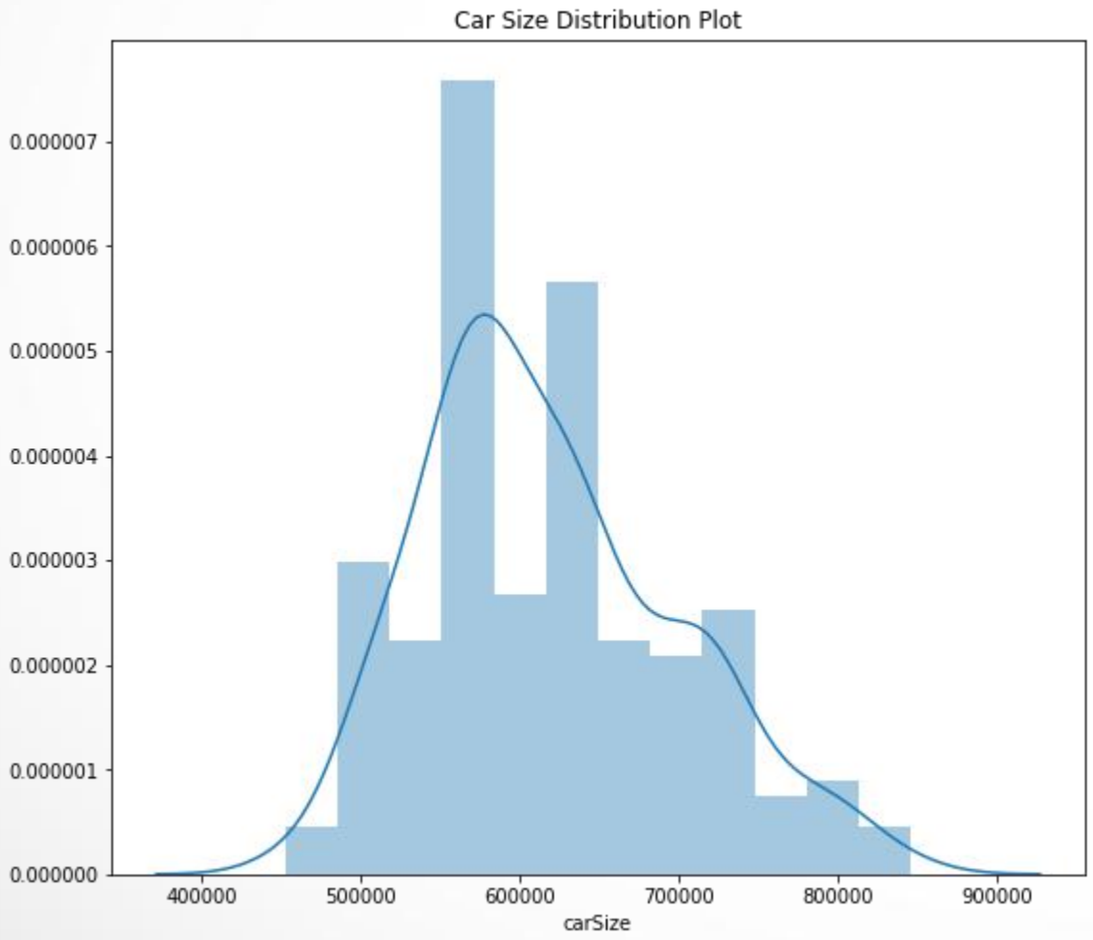
#cars volume assigned in carSize column:

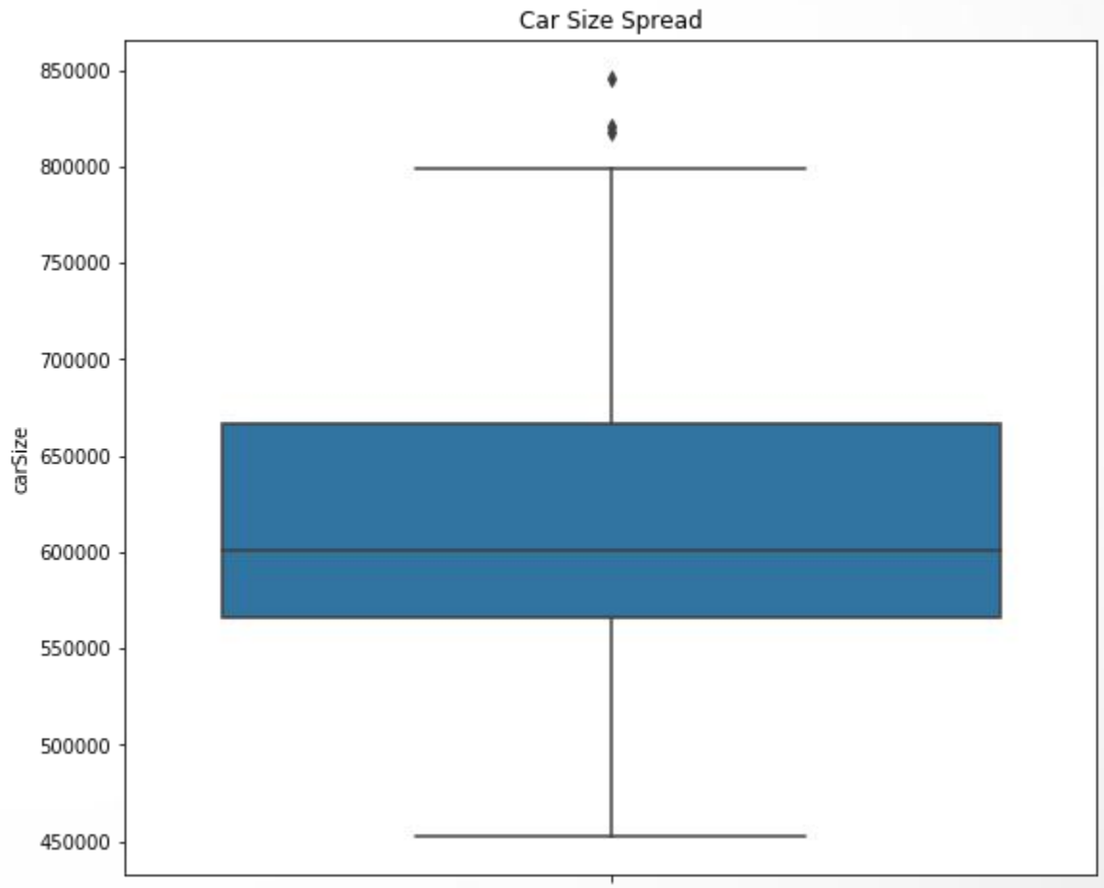
df['carSize'] = df['carlength'] * df['carwidth'] * df['carheight']

```
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   car_ID            205 non-null    int64
 1   symboling         205 non-null    int64
 2   CarName           205 non-null    object
 3   fueltype          205 non-null    object
 4   aspiration        205 non-null    object
 5   doornumber        205 non-null    object
 6   carbody           205 non-null    object
 7   drivewheel        205 non-null    object
 8   enginelocation    205 non-null    object
 9   wheelbase         205 non-null    float64
 10  curbweight        205 non-null    int64
 11  enginetype        205 non-null    object
 12  cylindernumber    205 non-null    object
 13  enginesize        205 non-null    int64
 14  fuelsystem        205 non-null    object
 15  boreratio         205 non-null    float64
 16  stroke            205 non-null    float64
 17  compressionratio  205 non-null    float64
 18  horsepower        205 non-null    int64
 19  peakrpm           205 non-null    int64
 20  citympg           205 non-null    int64
 21  highwaympg        205 non-null    int64
 22  price             205 non-null    float64
 23  carSize           205 non-null    float64
dtypes: float64(6), int64(8), object(10)
```

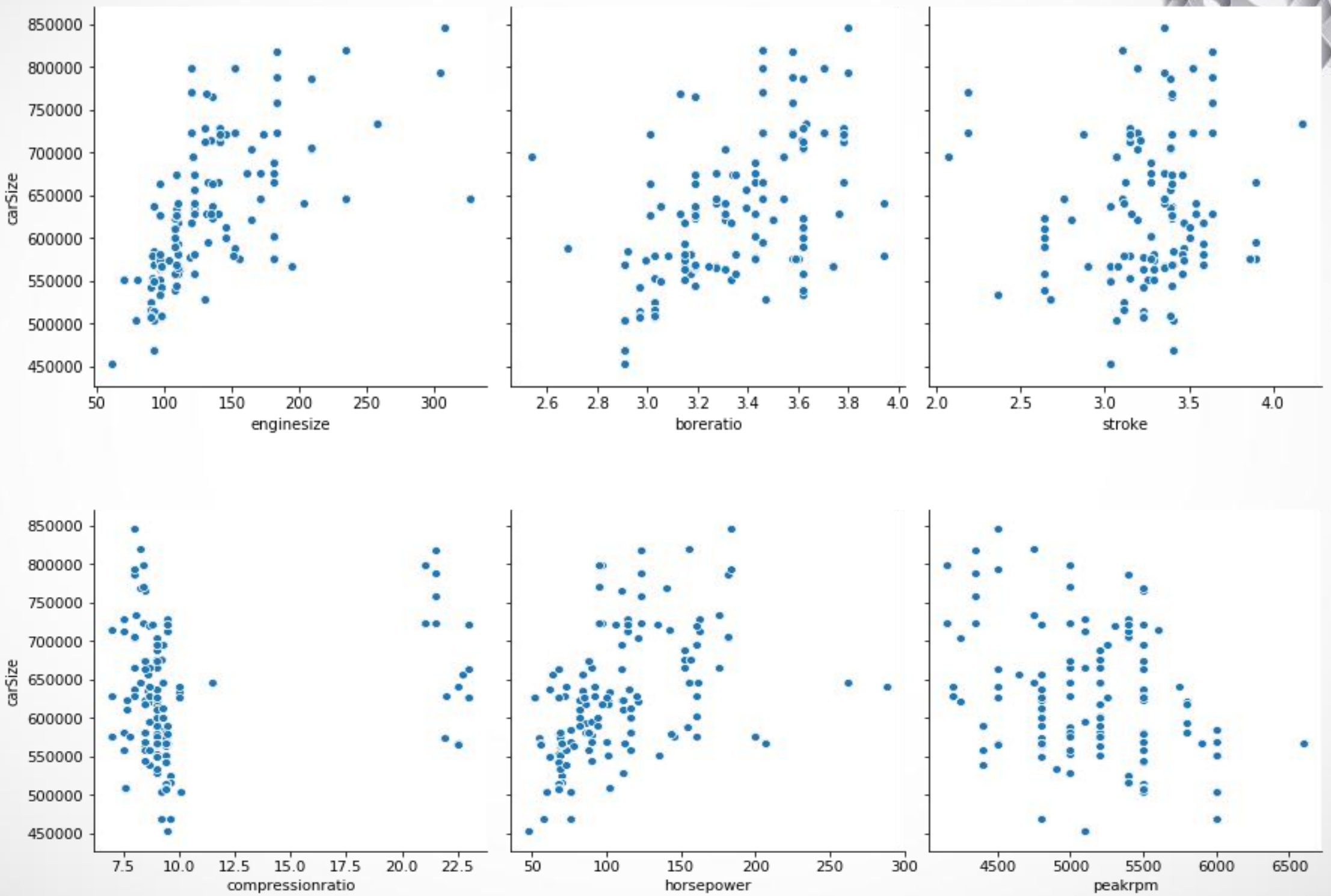
Visualization

Cars Volume/Size Column

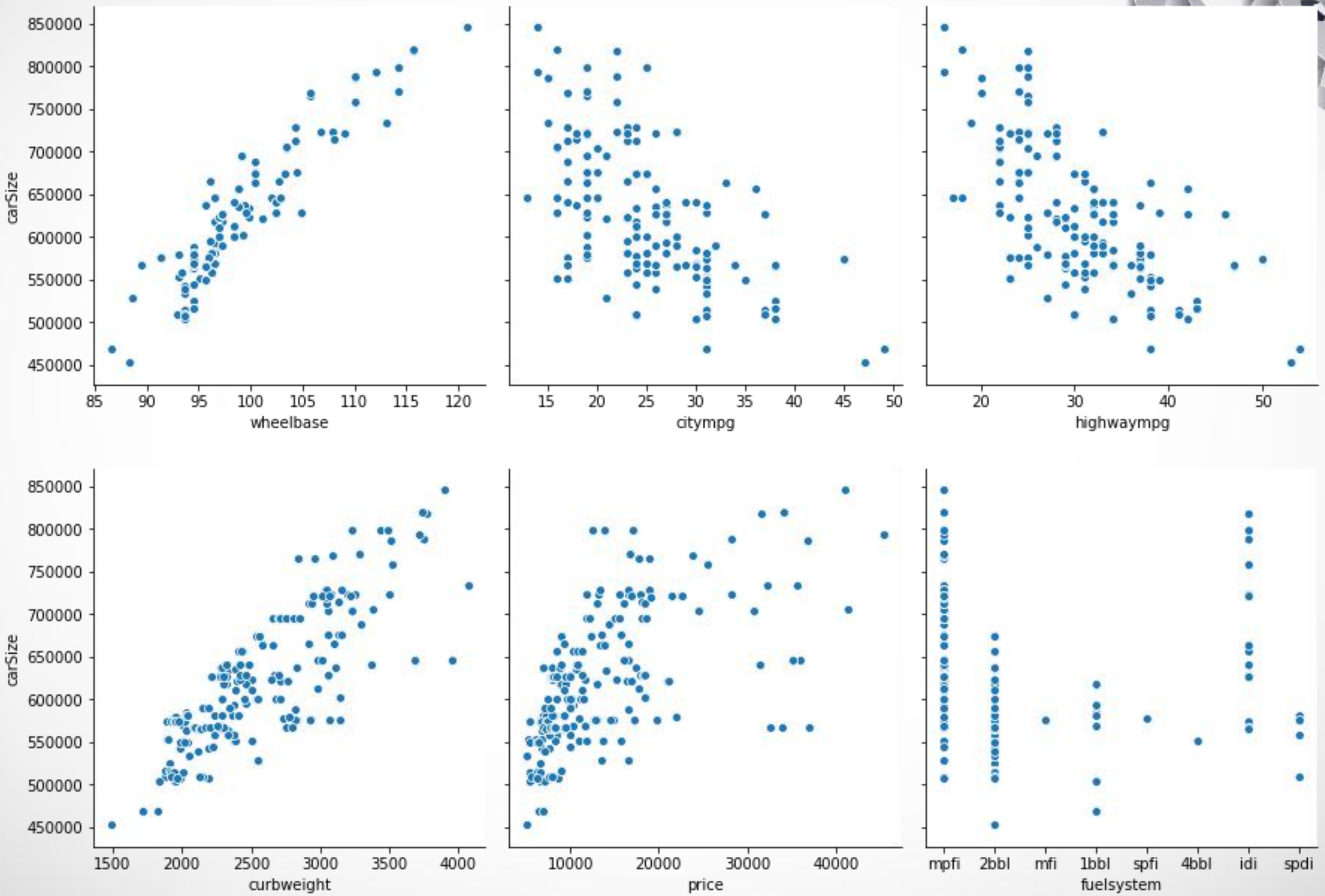




carSize vs other features



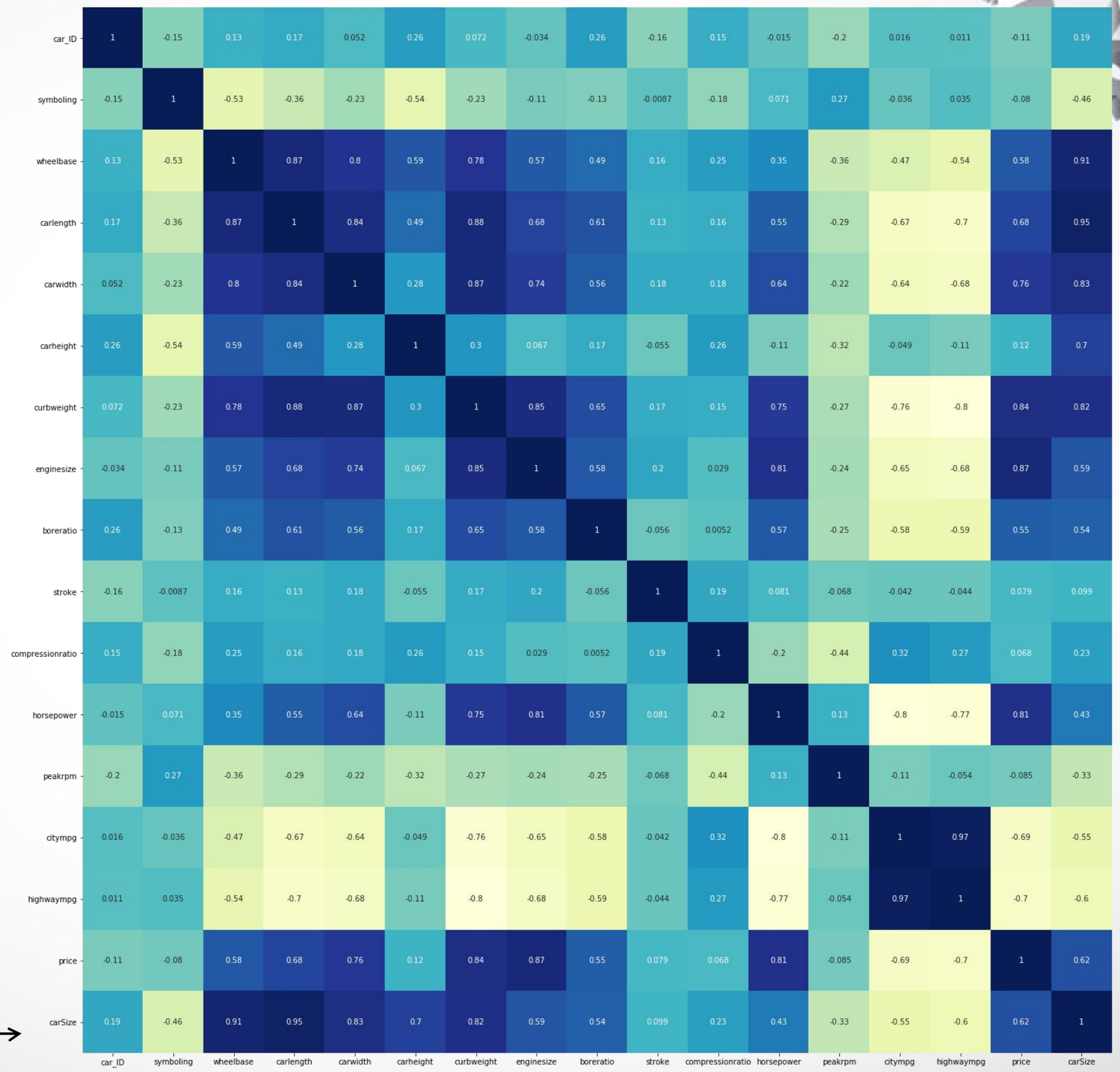
carSize vs other features



Correlation Matrics

highly corelated features
1 > x >= 0.5

- wheelbase
- curbweight
- enginesize
- boreratio
- price



Use Case

- in thes project a car dataset is used with many car's features
- in a car inventory or car showroom cars can be arranged in any manner.
- if the cars are arraned in a random way or without any analysis then it wil inefficient inventory arrangement and verry messy.

Solution to the Use Case

- different cars have different feature measures
- here each car volume/size was measured
- as car inventory management is all about space management so significant features should be the main focus which affects the car size
- then pair wise correlation was measured to identify the significant features for further calculations

Solution to the Use Case

- then highly correlated features were identified which do affect the Car Size. these are:

1. wheelbase
2. curbweight
3. enginesize
4. boreratio
5. price

- then labels were assigned to the data

- then by applying some ML and NN models best result can be achived

Architectural Choice

- Environment :

  Jupyter Notebook

- Frameworks :

  Apache Spark (pyspark)

Models

MultiClass Classification - A, B & C :

1. LogisticRegression
2. RandomForestClassifier
3. DecisionTreeClassifier

Binary Classification - A & B :

1. LinearSVC
2. GBTClassifier
3. RandomForestClassifier

Data Quality assessment

- as we saw there is no null value in the dataset and the carSize feature has 2 outliars only and some there were some features had positive correlation with the carSize so the Data Quality is good enough.

Data Pre-processing

- As it is a classification problem we must assign some class/label to the dataset for each row.
- we are solving space related probelm so we must must assign class based on car volume/size.
- so ABC analysis was done based on carSize (car volume) column and each row got a class A,B or C,
- for binary classification model only A and B class was assigned.

Feature Engineering

- Based on the correlation matrics 5 features were selected which have correlation with carSize grater than 0.5 and less than 1.0

- then Class column was assigned as label column and that 5 columns as feature columns.

Model Performance Indicator

- Model Performance Indicators
  - Accuracy
  - Precision
  - Recall

	RandomForestMultiClassifier	DecisionTreeClassifier	LogisticRegression	LinearSVC	GBTClassifier	RandomForestBinaryClassifier
TrainAccuracy	1	0.951515	0.5375	0.892374	1	1
TestAccuracy	1	0.975	0.577778	0.909722	1	1
Accuracy	0.711111	0.7	0.577778	0.529412	1	0.960784
Precision	1	1	0.742857	0.529412	1	0.913043
Recall	1	0.863636	1	1	1	1

Model Algorithm

Based on the Model Performance Indicators DecisionTreeClassifier Algorithm should be selected as the Model Algorithm.