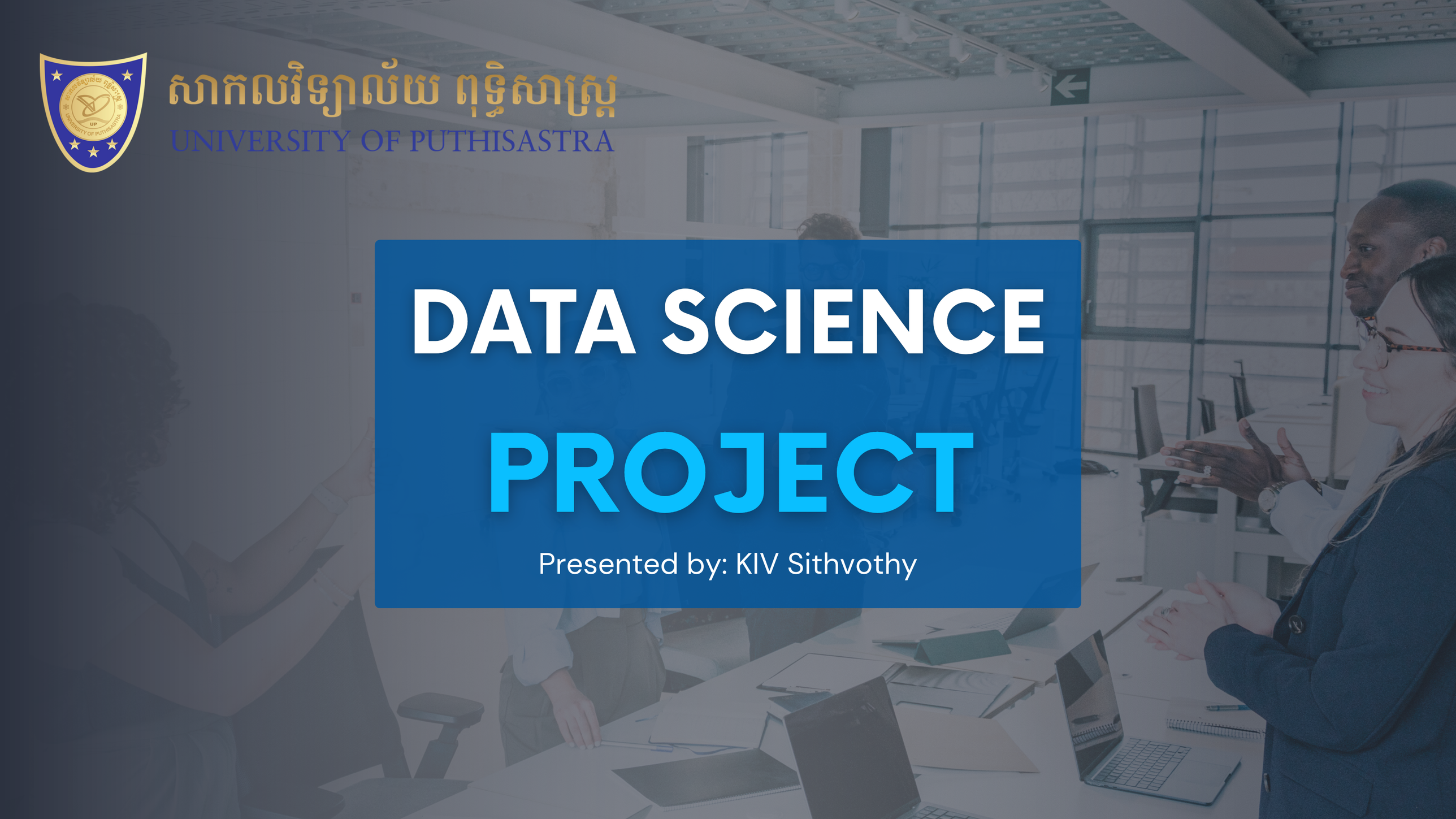


សាកលវិទ្យាល័យ ពុទ្ធិសាស្ត្រ
UNIVERSITY OF PUTHISASTRA

DATA SCIENCE PROJECT

Presented by: KIV Sithvothy

OVERVIEW

- Introduction
- Descriptive Analysis
- Data Exploration
- EDA
- Report
- Reference

INTRODUCTION

This assignment's goal is to do exploratory data analysis (EDA) on a given **automobile.csv** dataset. I will be able to obtain practical experience in data preprocessing, visualization, modeling, and interpretation—all important skills for a data scientist—through this assignment.

DESCRIPTIVE ANALYSIS

UNDERSTANDING THE BASIC STATISTICS OF DATASET

```
df.describe()
```

✓ 0.0s Pyth

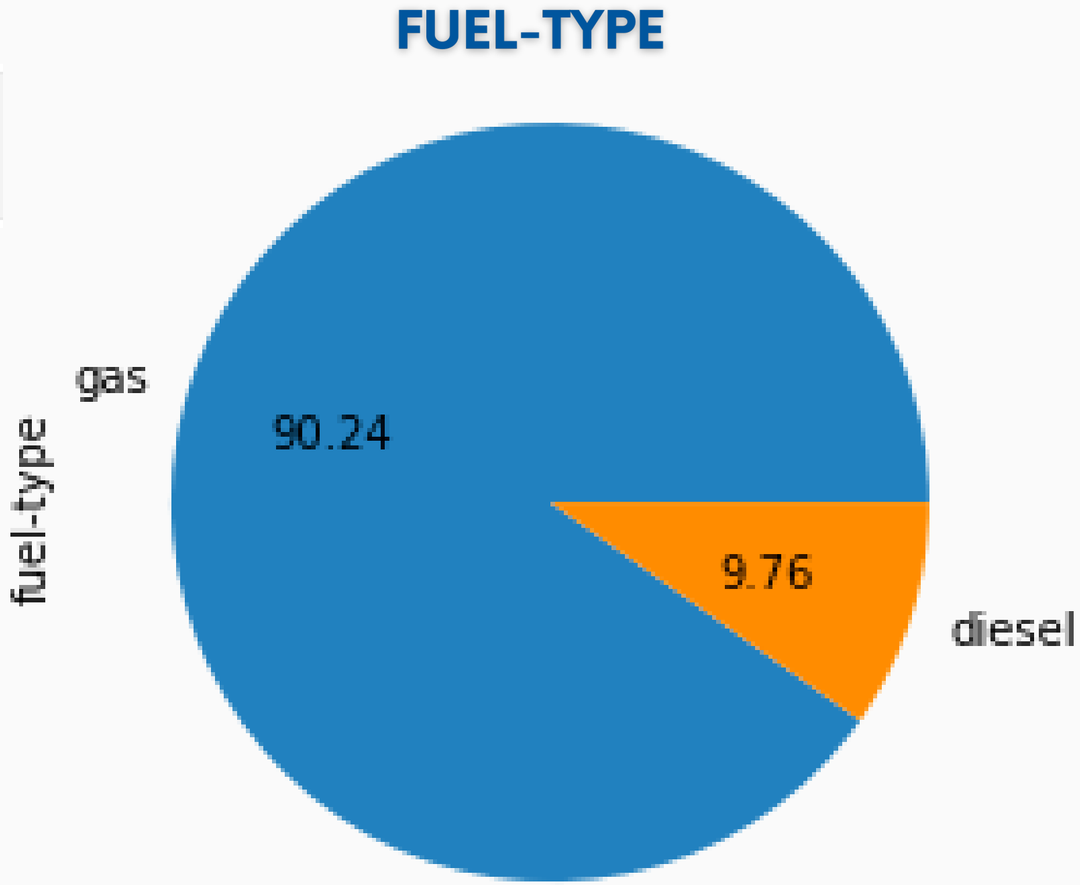
	symboling	wheel-base	length	width	height	curb-weight	engine-size	compression-ratio	city-mpg	highway-mpg
count	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000
mean	0.834146	98.756585	174.049268	65.907805	53.724878	2555.565854	126.907317	10.142537	25.219512	30.751220
std	1.245307	6.021776	12.337289	2.145204	2.443522	520.680204	41.642693	3.972040	6.542142	6.886443
min	-2.000000	86.600000	141.100000	60.300000	47.800000	1488.000000	61.000000	7.000000	13.000000	16.000000
25%	0.000000	94.500000	166.300000	64.100000	52.000000	2145.000000	97.000000	8.600000	19.000000	25.000000
50%	1.000000	97.000000	173.200000	65.500000	54.100000	2414.000000	120.000000	9.000000	24.000000	30.000000
75%	2.000000	102.400000	183.100000	66.900000	55.500000	2935.000000	141.000000	9.400000	30.000000	34.000000
max	3.000000	120.900000	208.100000	72.300000	59.800000	4066.000000	326.000000	23.000000	49.000000	54.000000

DESCRIPTIVE ANALYSIS

DISTRIBUTION OF CATEGORICAL VARIABLES

```
df["fuel-type"].value_counts().plot.pie(autopct="%.2f")
```

✓ 0.1s



DATA EXPLORATION AND PREPROCESSING

MISSING VALUES

```
df.head()
```

✓ 0.0s Python

normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	...	engine-size	fuel-system	bore	stroke	compression-ratio	horsepower
?	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68	9.0	111
?	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68	9.0	111
?	alfa-romero	gas	std	two	hatchback	rwd	front	94.5	...	152	mpfi	2.68	3.47	9.0	154
164	audi	gas	std	four	sedan	fwd	front	99.8	...	109	mpfi	3.19	3.4	10.0	102
164	audi	gas	std	four	sedan	4wd	front	99.4	...	136	mpfi	3.19	3.4	8.0	115

nns

```
df["normalized-losses"]=df["normalized-losses"].str.replace("?",str(mean)).astype(dtype="float")
```

✓ 0.2s Python

DATA EXPLORATION AND PREPROCESSING

NORMALIZE NUMERICAL FEATURES

['price', 'engine-size']

```
# Select the columns of interest
columns_of_interest = ['price', 'engine-size']
data = df[columns_of_interest]
data.head()
```

✓ 0.0s

	price	engine-size
0	13495.0	130
1	16500.0	130
2	16500.0	152
3	13950.0	109
4	17450.0	136

DATA EXPLORATION AND PREPROCESSING

NORMALIZE NUMERICAL FEATURES

Using pandas and scikit-learn

['price', 'engine-size']

```python
from sklearn.preprocessing import StandardScaler

# Standardize the data
scaler = StandardScaler()
standardized_data = scaler.fit_transform(data)
standardized_df = pd.DataFrame(standardized_data, columns=columns_of_interest)

# Display the standardized data
print(standardized_df)
```

✓ 0.0s

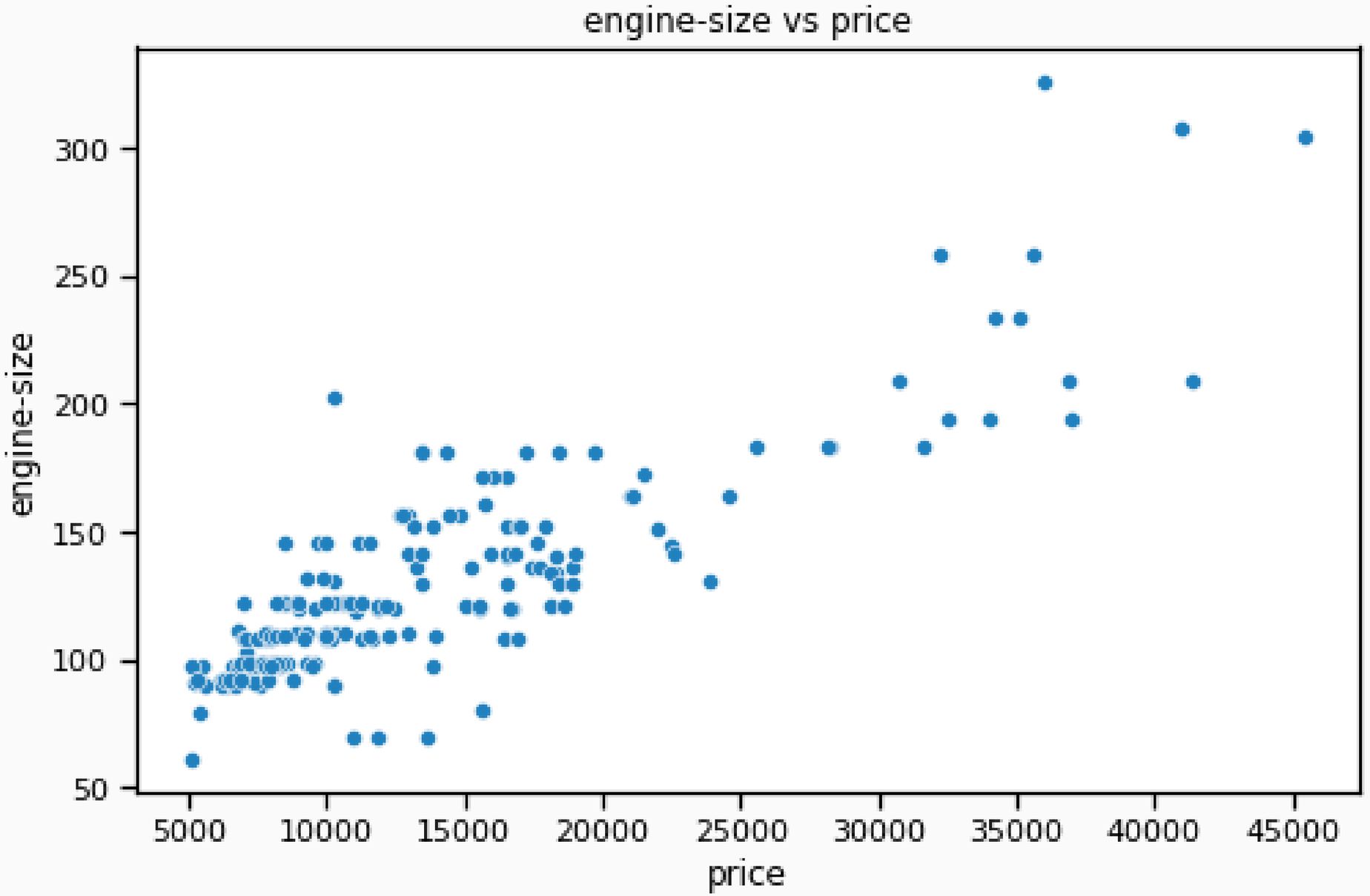
```
        price  engine-size
0    0.043855     0.074449
1    0.426176     0.074449
2    0.426176     0.604046
3    0.101744    -0.431076
4    0.547043     0.218885
..        ...          ...
200  0.470070     0.339248
201  0.749972     0.339248
202  1.060410     1.109571
203  1.185730     0.435538
204  1.205450     0.339248

[205 rows x 2 columns]
```

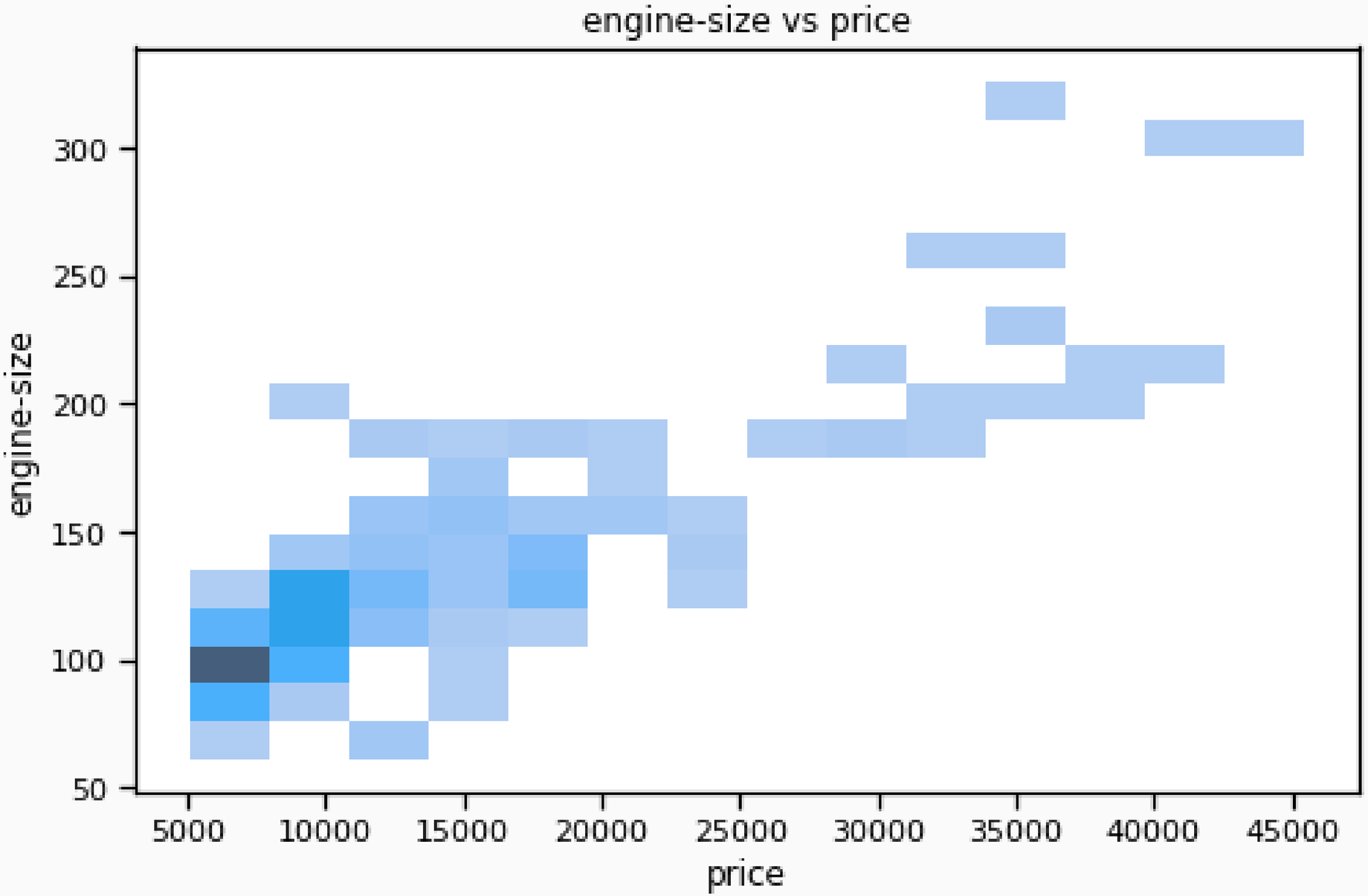
EXPLORATORY DATA ANALYSIS

(EDA)

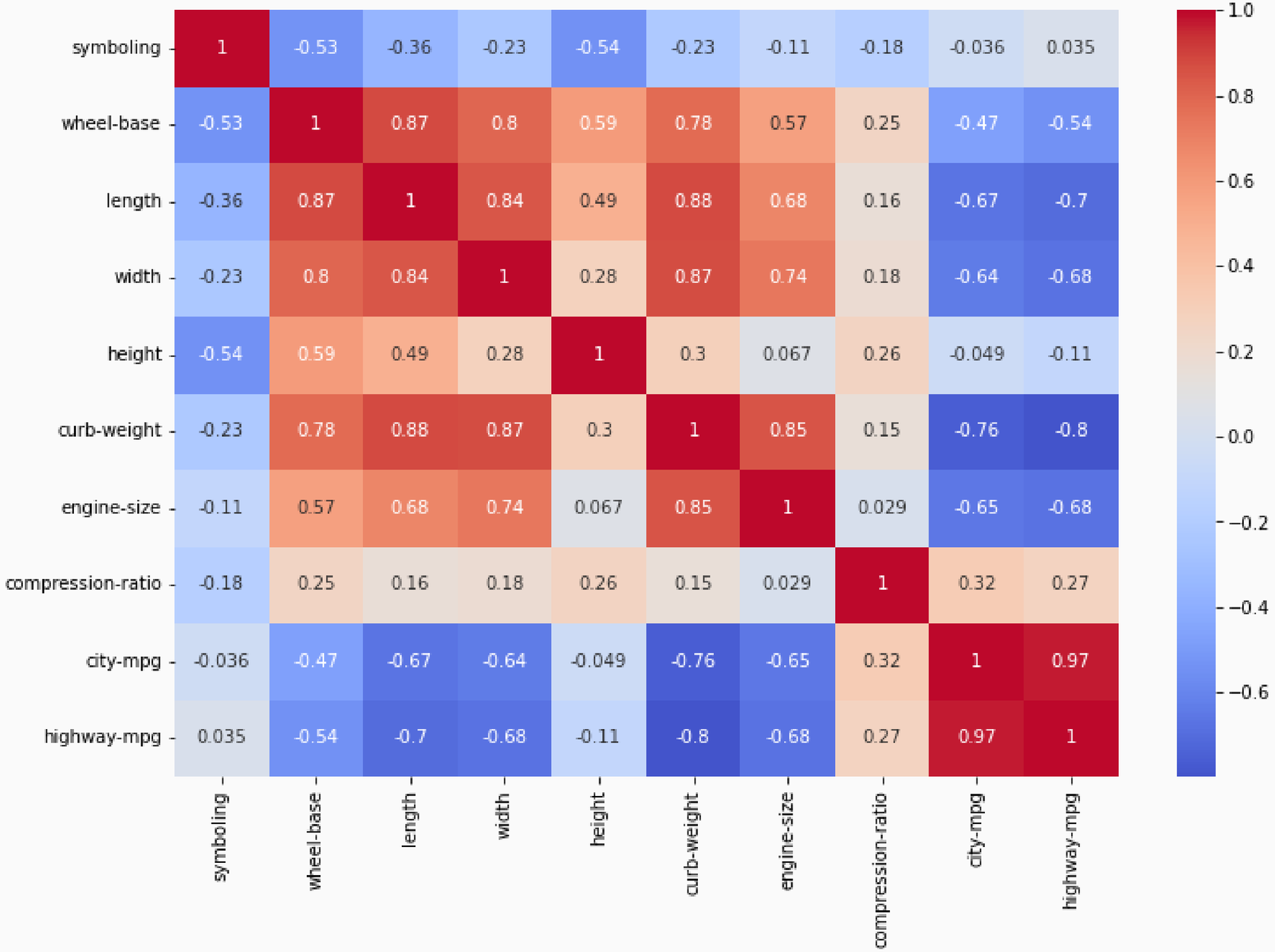
VISUALIZE THE RELATIONSHIP AND EXPLORE DISTRIBUTION TARGET VARIABLE



VISUALIZE THE RELATIONSHIP AND EXPLORE DISTRIBUTION TARGET VARIABLE



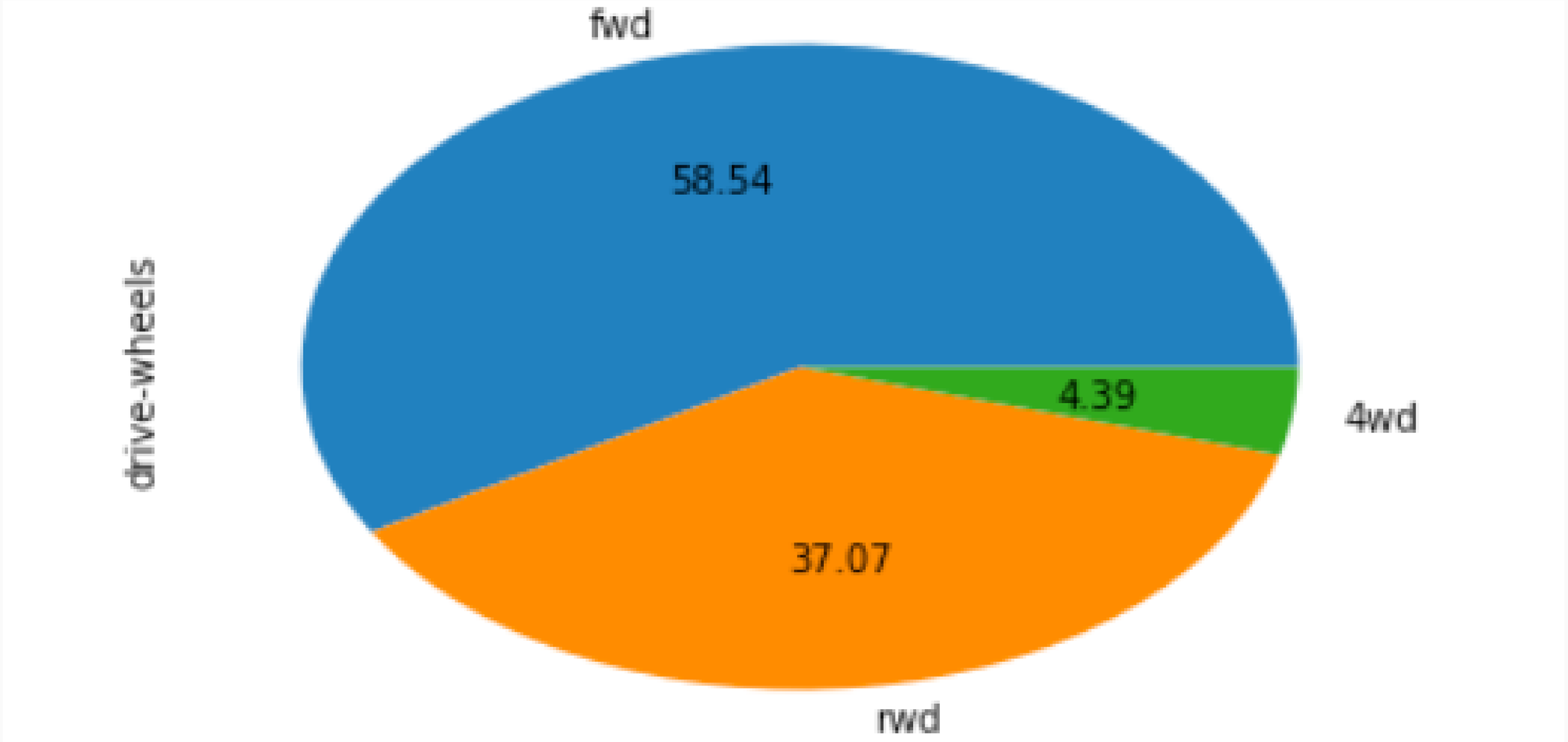
IDENTIFY ANY PATTERNS OR ANOMALIES



REPORT

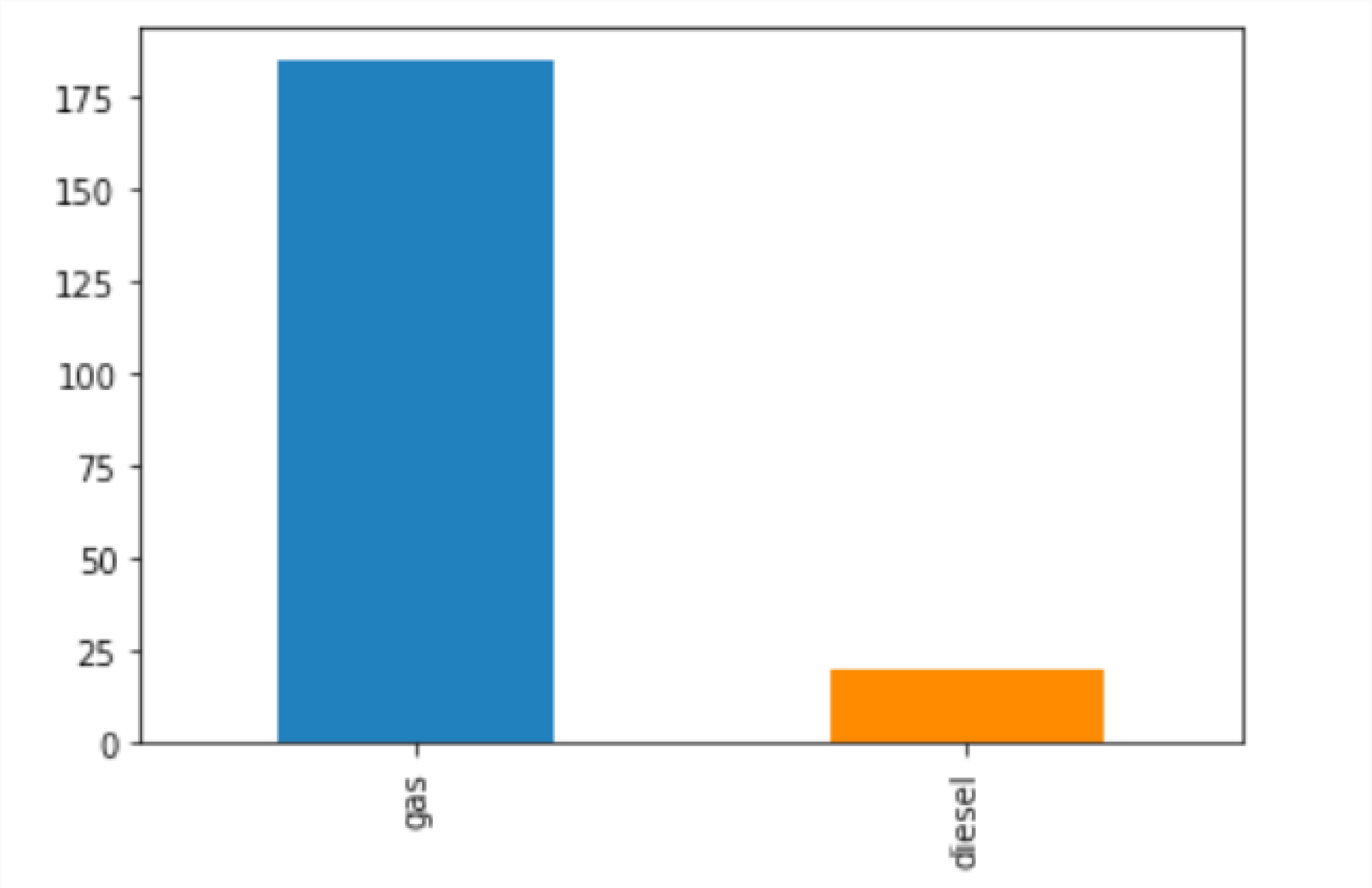
VISUALISATIONS AND THE CORRESPONDING INSIGHTS

Not a lot of people have 4 wheel drive cars



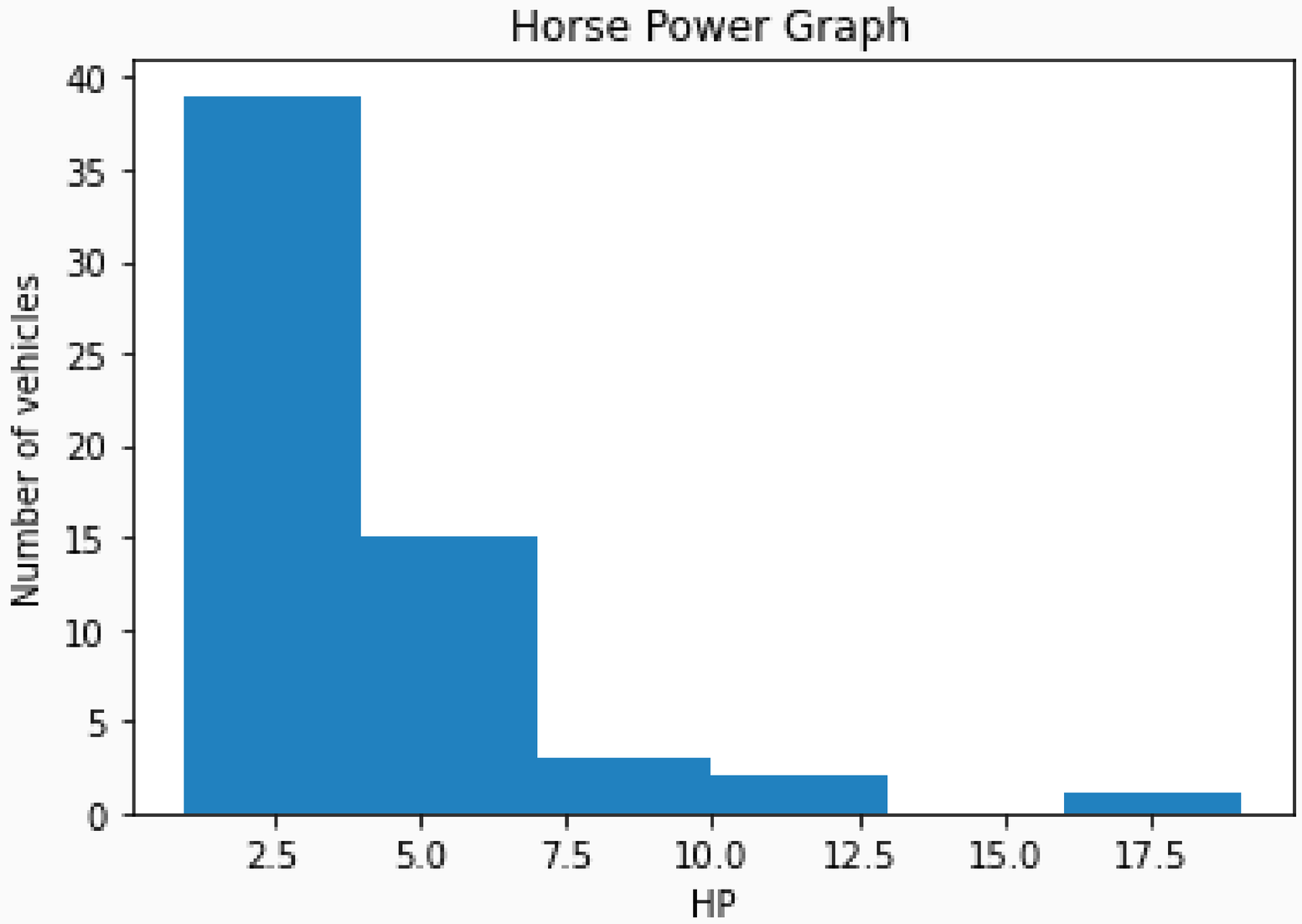
VISUALISATIONS AND THE CORRESPONDING INSIGHTS

90% of the people use cars that run of gas rather than diesel



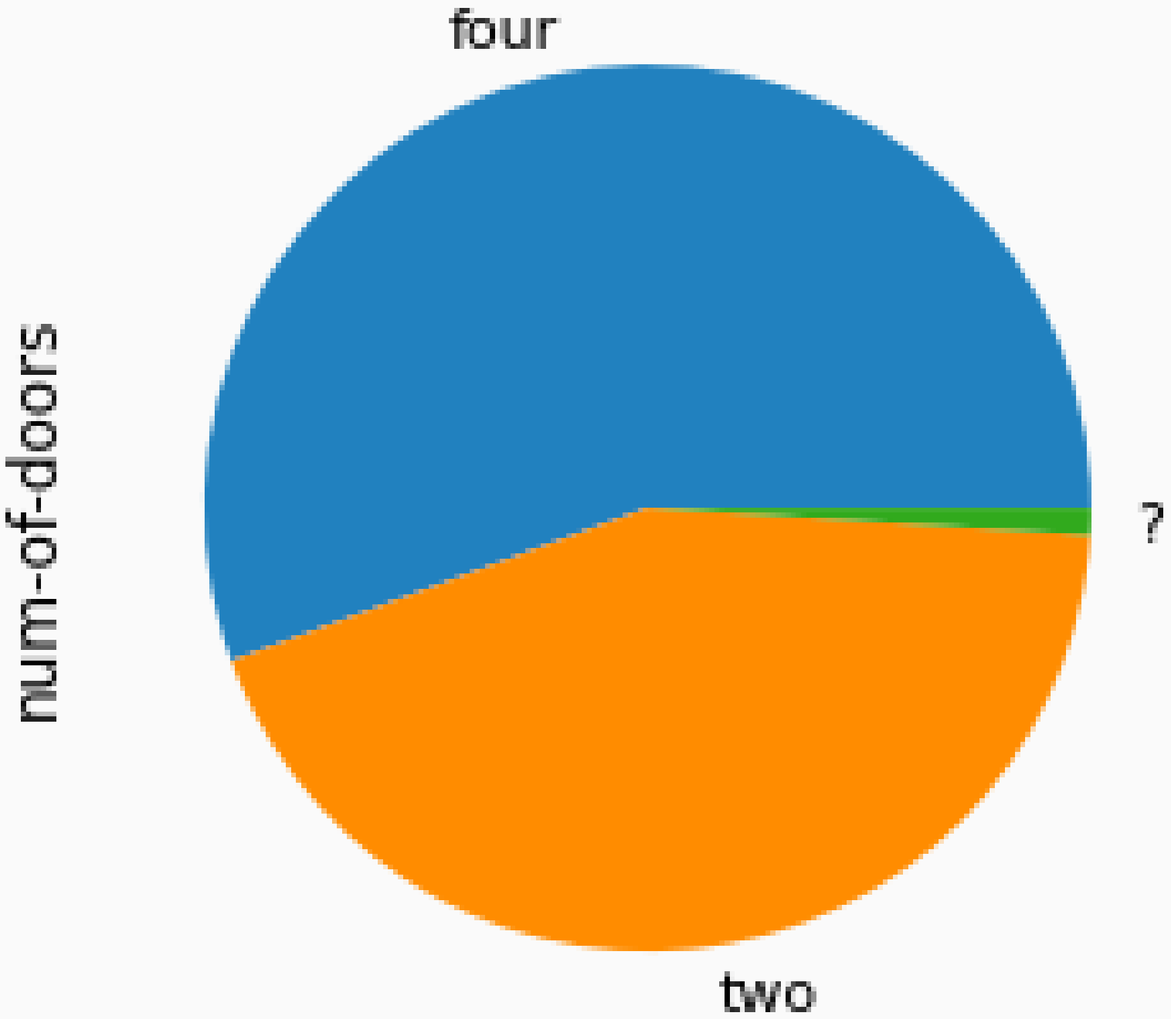
VISUALISATIONS AND THE CORRESPONDING INSIGHTS

Visuallize vehicles horse power



VISUALISATIONS AND THE CORRESPONDING INSIGHTS

number of the vehicle's doors



REFERENCE

REFERENCES

- https://github.com/UP-MSIT/EDA-on-Automobile-Dataset
- https://pandas.pydata.org/
- https://numpy.org/
- https://matplotlib.org/
- https://seaborn.pydata.org/