Group 7 presentation

EXPLORING CAR PRICE THROUGH ANALYTICAL INSIGHT

BITI2513 INTRODUCTION TO DATA SCIENCE

Project Overview

- Explores the relationship between car features and pricing in the automotive market.
- Focuses on analyzing a comprehensive dataset of various car models, brands, and their corresponding features.
- Aim to uncover insights into how different features impact the Manufacturer's Suggested Retail Price (MSRP) of vehicles.

Objective of The Project

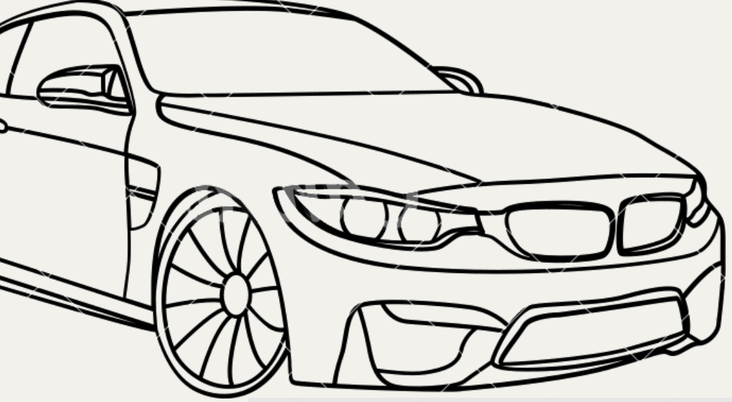
Understanding of how car features influence the pricing of vehicles.

Identify the key factors that significantly contribute to the pricing of vehicles

To guide car buyers in finding the right balance between desired features and affordability.

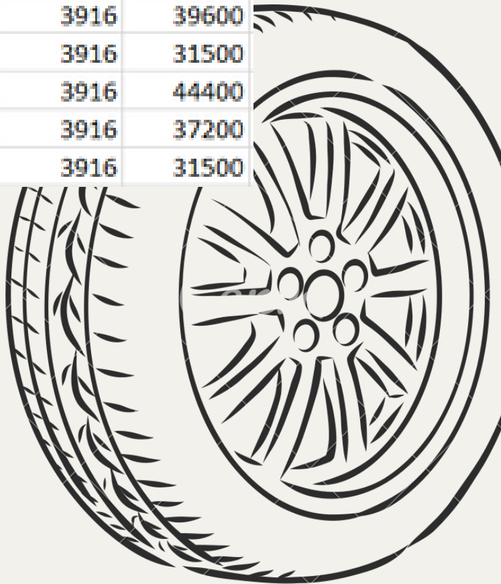
Gain valuable insights into market demand and customer preferences

DATA WRANGLING



Data Collection: Collect data from Kaggle

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Make	Model	Year	Engine I	Engine I	Engine (	Transm	Driven_	Numbe	Market	Vehicle	Vehicle	highwa	city mp	Popular	MSRP
2	BMW	1 Series M	2011	premium u	335	6	MANUAL	rear wheel	2	Factory Tu	Compact	Coupe	26	19	3916	46135
3	BMW	1 Series	2011	premium u	300	6	MANUAL	rear wheel	2	Luxury,Per	Compact	Convertibl	28	19	3916	40650
4	BMW	1 Series	2011	premium u	300	6	MANUAL	rear wheel	2	Luxury,Hig	Compact	Coupe	28	20	3916	36350
5	BMW	1 Series	2011	premium u	230	6	MANUAL	rear wheel	2	Luxury,Per	Compact	Coupe	28	18	3916	29450
6	BMW	1 Series	2011	premium u	230	6	MANUAL	rear wheel	2	Luxury	Compact	Convertibl	28	18	3916	34500
7	BMW	1 Series	2012	premium u	230	6	MANUAL	rear wheel	2	Luxury,Per	Compact	Coupe	28	18	3916	31200
8	BMW	1 Series	2012	premium u	300	6	MANUAL	rear wheel	2	Luxury,Per	Compact	Convertibl	26	17	3916	44100
9	BMW	1 Series	2012	premium u	300	6	MANUAL	rear wheel	2	Luxury,Hig	Compact	Coupe	28	20	3916	39300
10	BMW	1 Series	2012	premium u	230	6	MANUAL	rear wheel	2	Luxury	Compact	Convertibl	28	18	3916	36900
11	BMW	1 Series	2013	premium u	230	6	MANUAL	rear wheel	2	Luxury	Compact	Convertibl	27	18	3916	37200
12	BMW	1 Series	2013	premium u	300	6	MANUAL	rear wheel	2	Luxury,Hig	Compact	Coupe	28	20	3916	39600
13	BMW	1 Series	2013	premium u	230	6	MANUAL	rear wheel	2	Luxury,Per	Compact	Coupe	28	19	3916	31500
14	BMW	1 Series	2013	premium u	300	6	MANUAL	rear wheel	2	Luxury,Per	Compact	Convertibl	28	19	3916	44400
15	BMW	1 Series	2013	premium u	230	6	MANUAL	rear wheel	2	Luxury	Compact	Convertibl	28	19	3916	37200
16	BMW	1 Series	2013	premium u	230	6	MANUAL	rear wheel	2	Luxury,Per	Compact	Coupe	28	19	3916	31500



Data Inspecting: Checking Missing Data

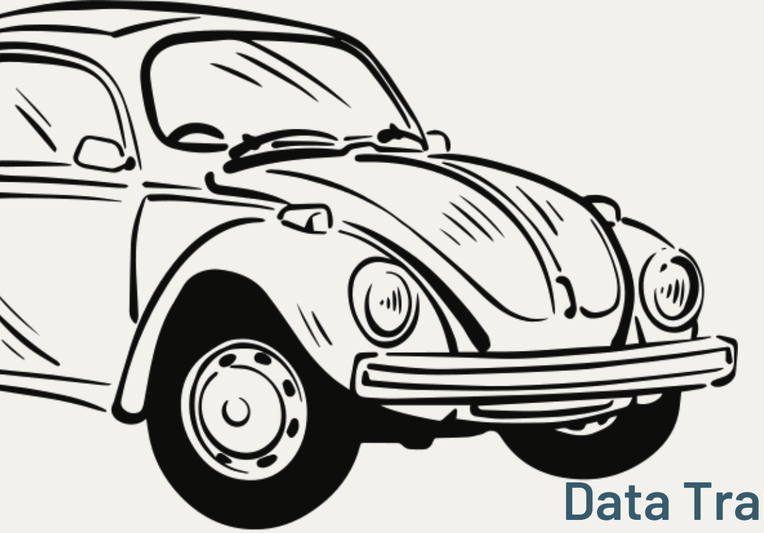
```
#missing data
df.isnull().sum().sort_values(ascending=False)
```

```
Market Category      3742
Engine HP              69
Engine Cylinders       30
Number of Doors         6
Engine Fuel Type        3
Make                    0
Model                   0
Year                    0
Transmission Type       0
Driven_Wheels           0
Vehicle Size            0
Vehicle Style           0
highway MPG             0
city mpg                0
Popularity              0
MSRP                    0
dtype: int64
```

Data Cleaning: Handling Missing Values

```
# Handling missing values
df['Market Category'] = df['Market Category'].fillna(df['Market Category'].mode()[0])
df['Engine Fuel Type'] = df['Engine Fuel Type'].fillna(df['Engine Fuel Type'].mode()[0])
```

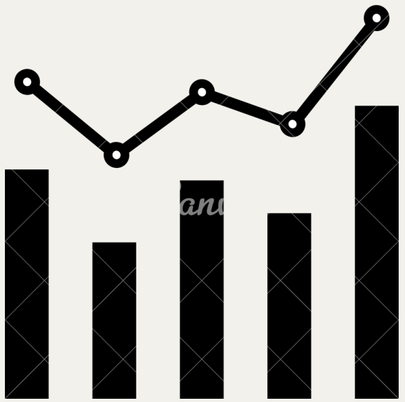
```
# Calculate the mean
mean_val1 = df['Engine HP'].mean()
mean_val2 = df['Engine Cylinders'].mean()
mean_val3 = df['Number of Doors'].mean()

# Replace NaN values with the mean
df['Engine HP'].fillna(mean_val1, inplace=True)
df['Engine Cylinders'].fillna(mean_val2, inplace=True)
df['Number of Doors'].fillna(mean_val2, inplace=True)
```



Data Transformation: Separate the feature and target column

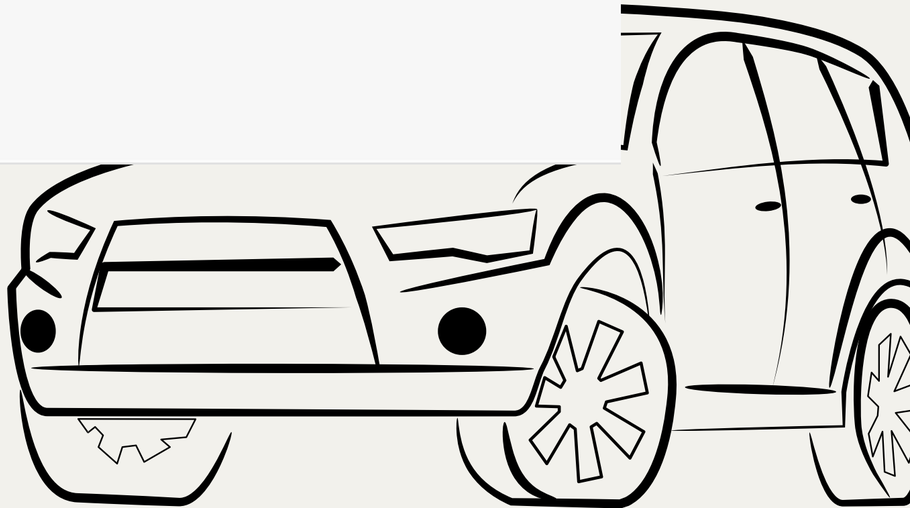
```
# Separate the feature and target columns
# To compare the data
X = df.drop('MSRP', axis = 1)
y = df['MSRP']
```

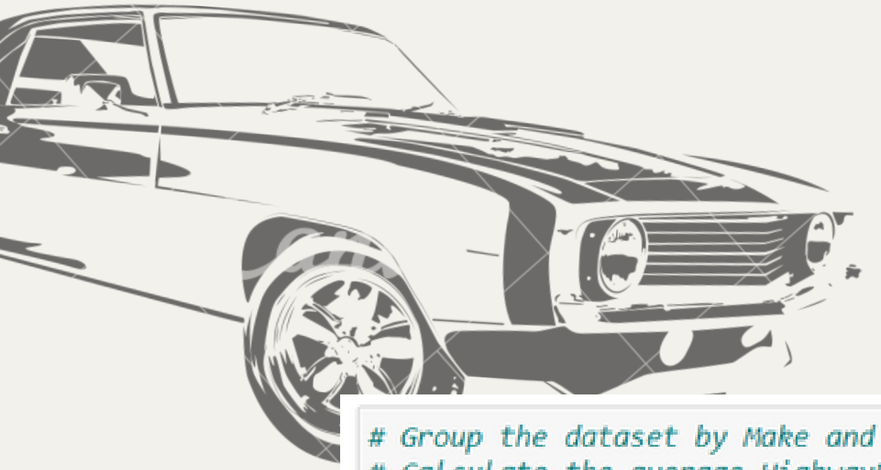
```
# Split columns into data types numerical and object (categorical)
num_cols = X.select_dtypes(include=['int64', 'float64']).columns.tolist()
cat_cols = X.select_dtypes(include=['object']).columns.tolist()
```



Data Formatting: Label Encoding

```python
# Perform data pre-processing by importing standardscaler
# To standardize and normalize numerical input variables for classification
from sklearn.preprocessing import StandardScaler

def scale_and_encode(df):
    # Split columns into numerical and categorical
    num_cols = df.select_dtypes(include=['int64', 'float64']).columns.tolist()
    cat_cols = df.select_dtypes(include=['object']).columns.tolist()

    # Standardize numerical columns
    scaler = StandardScaler()
    df[num_cols] = scaler.fit_transform(df[num_cols])

    # One-hot encode categorical columns
    df = pd.get_dummies(df, columns=cat_cols,drop_first = True)

    return df
```



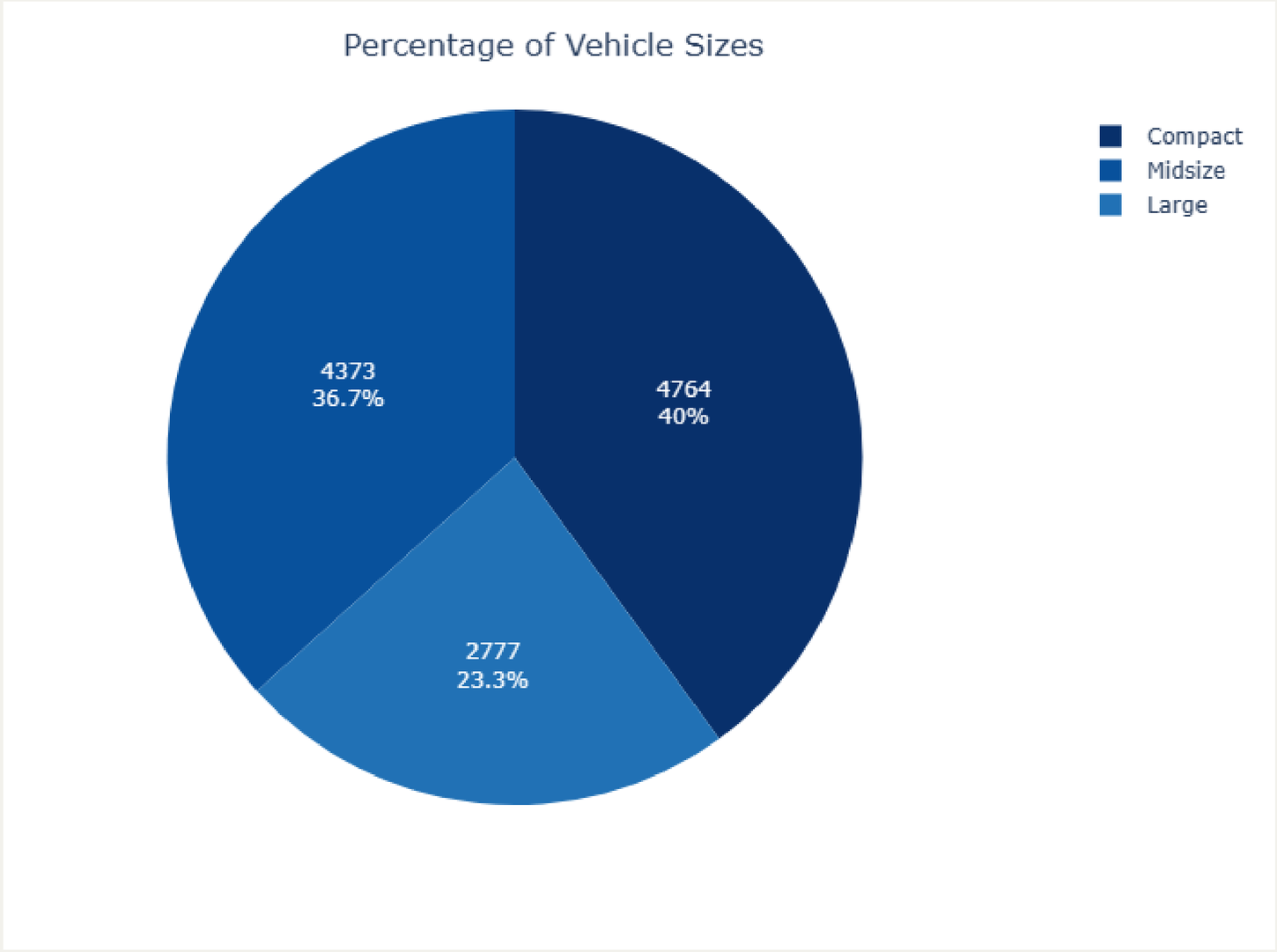


Data Reshaping: Grouping

```python
# Group the dataset by Make and calculate the average HighwayMPG
# Calculate the average HighwayMPG and CityMPG by Make
average_mpg_by_make = df.groupby('Make').agg({'highway MPG': 'mean', 'city mpg': 'mean'}).reset_index()
average_mpg_by_make['CombinedMPG'] = average_mpg_by_make['highway MPG'] + average_mpg_by_make['city mpg']

# Print the results
print("\nAverage MPG by Make:")
print(average_mpg_by_make)
```

```
Average MPG by Make:
             Make  highway MPG   city mpg  CombinedMPG
0           Acura    28.111111  19.940476    48.051587
1      Alfa Romeo    34.000000  24.000000    58.000000
2    Aston Martin    18.892473  12.526882    31.419355
3            Audi    28.823171  19.585366    48.408537
4             BMW    29.245509  20.739521    49.985030
5         Bentley    18.905405  11.554054    30.459459
```

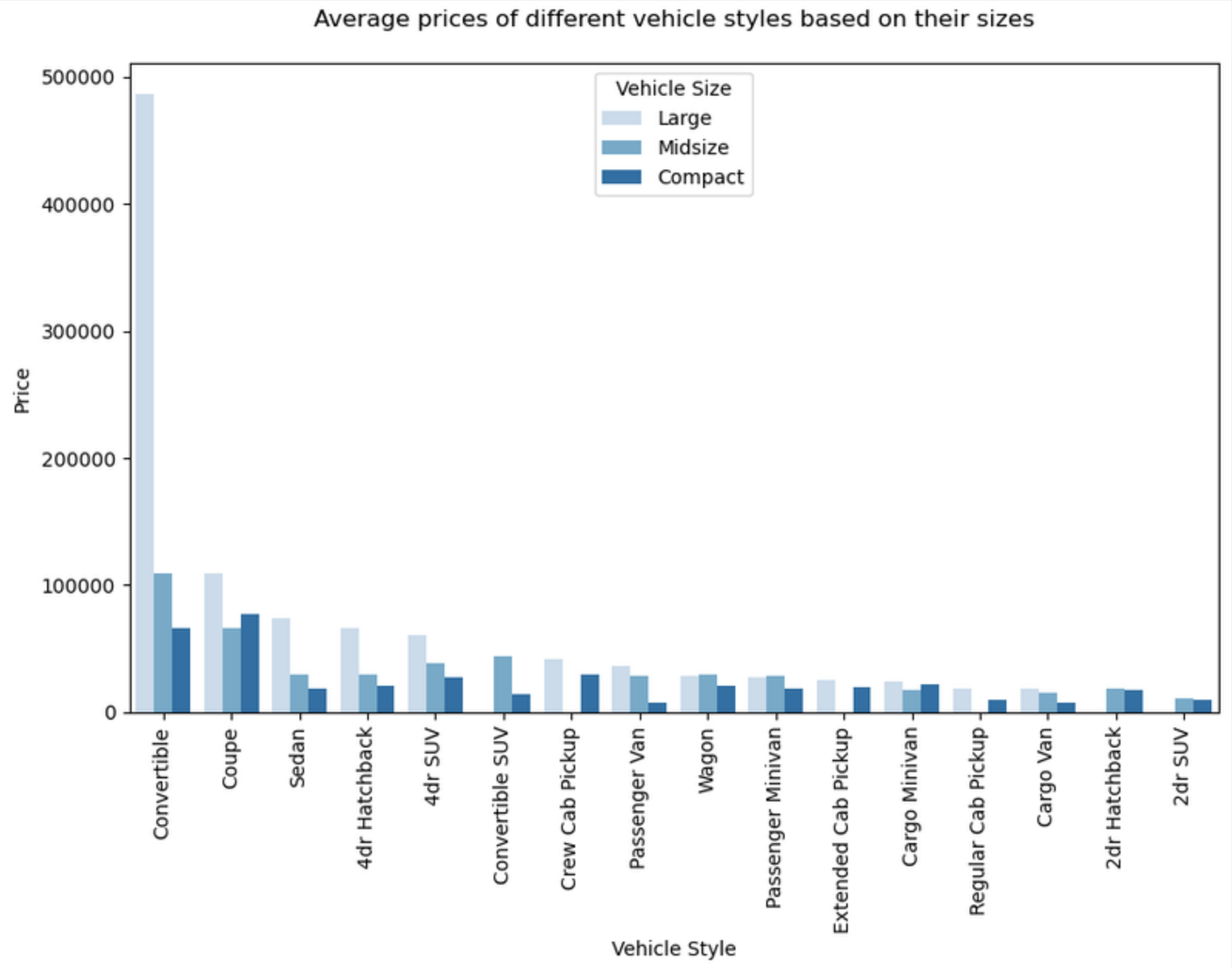
EXPLORITARY DATA ANALYSIS



The insight:

There 3 sizes which are compact, medium and large.

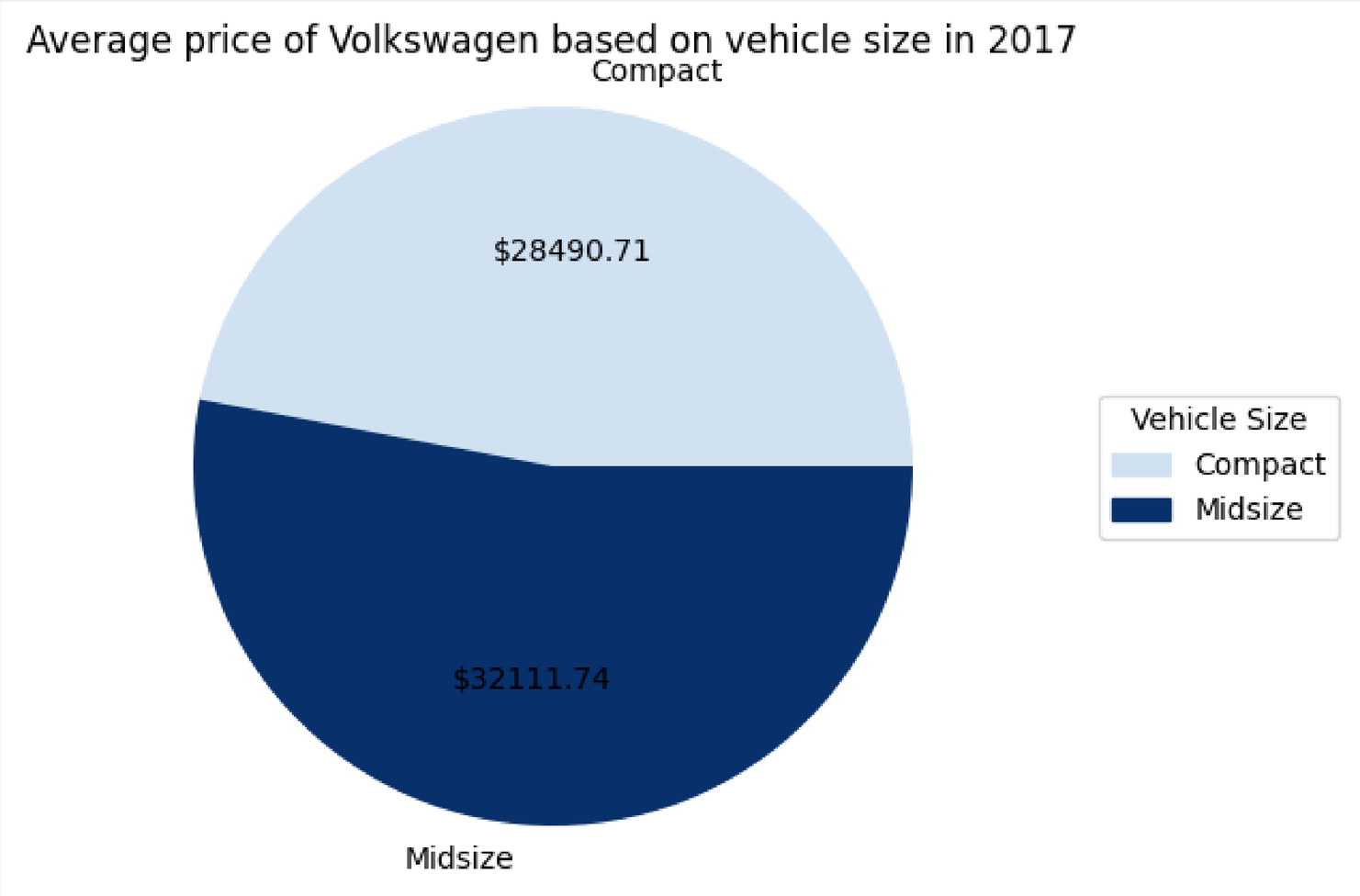
Compact size has highest number of cars (4764) followed by midsize (4373) and large (2777)



The insight:

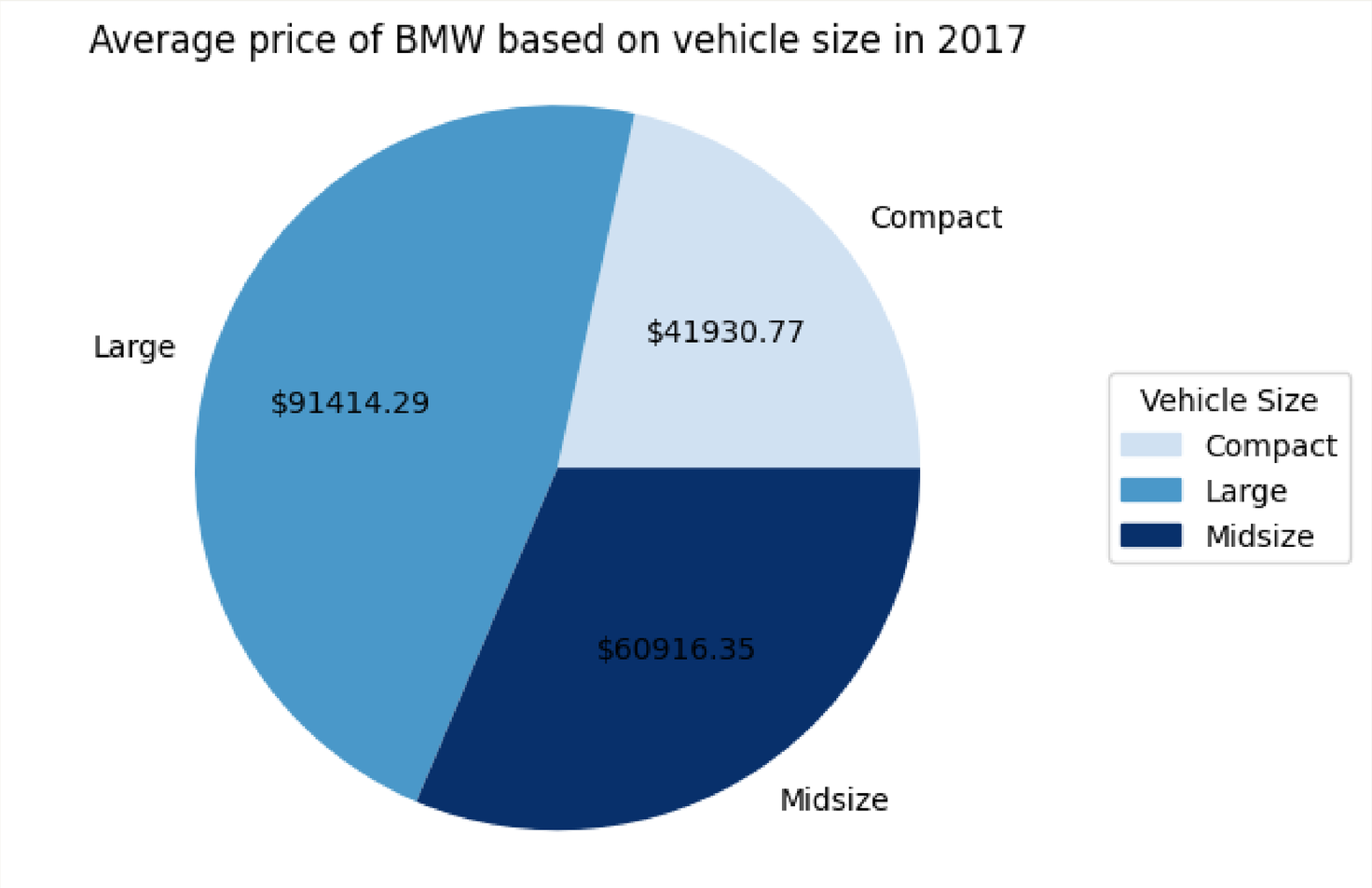
Convertible (large size) has the highest price, which is closest to 50k USD.

Most large vehicles have the highest price compared to medium and compact vehicles.



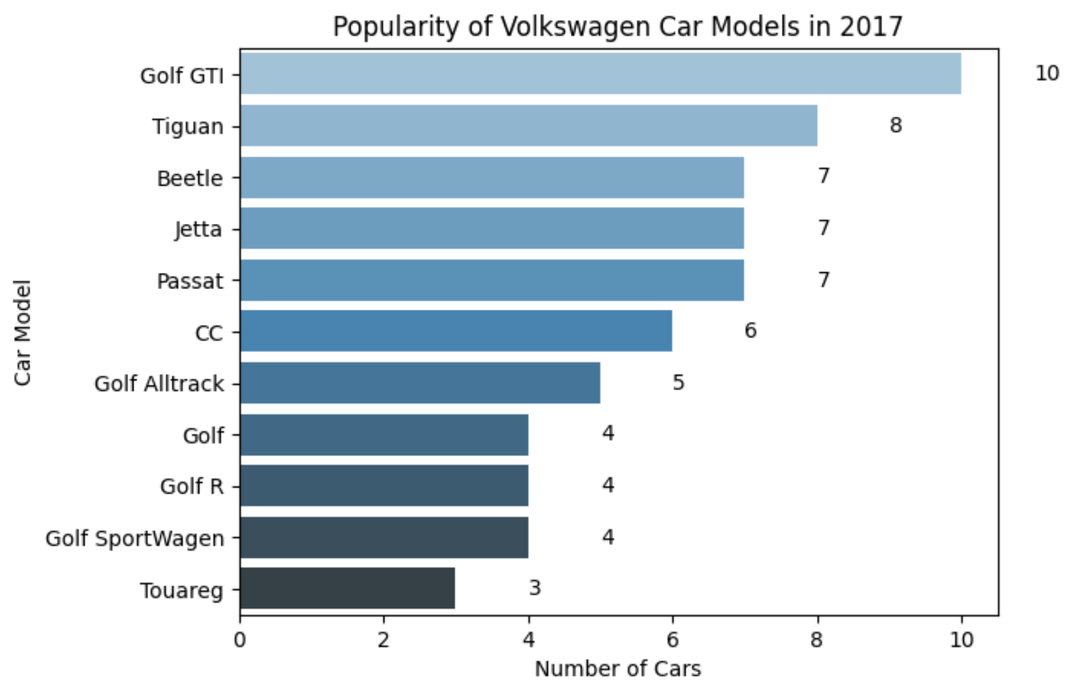
The insight:

Medium-sized cars are more expensive than compact cars.



The insight:

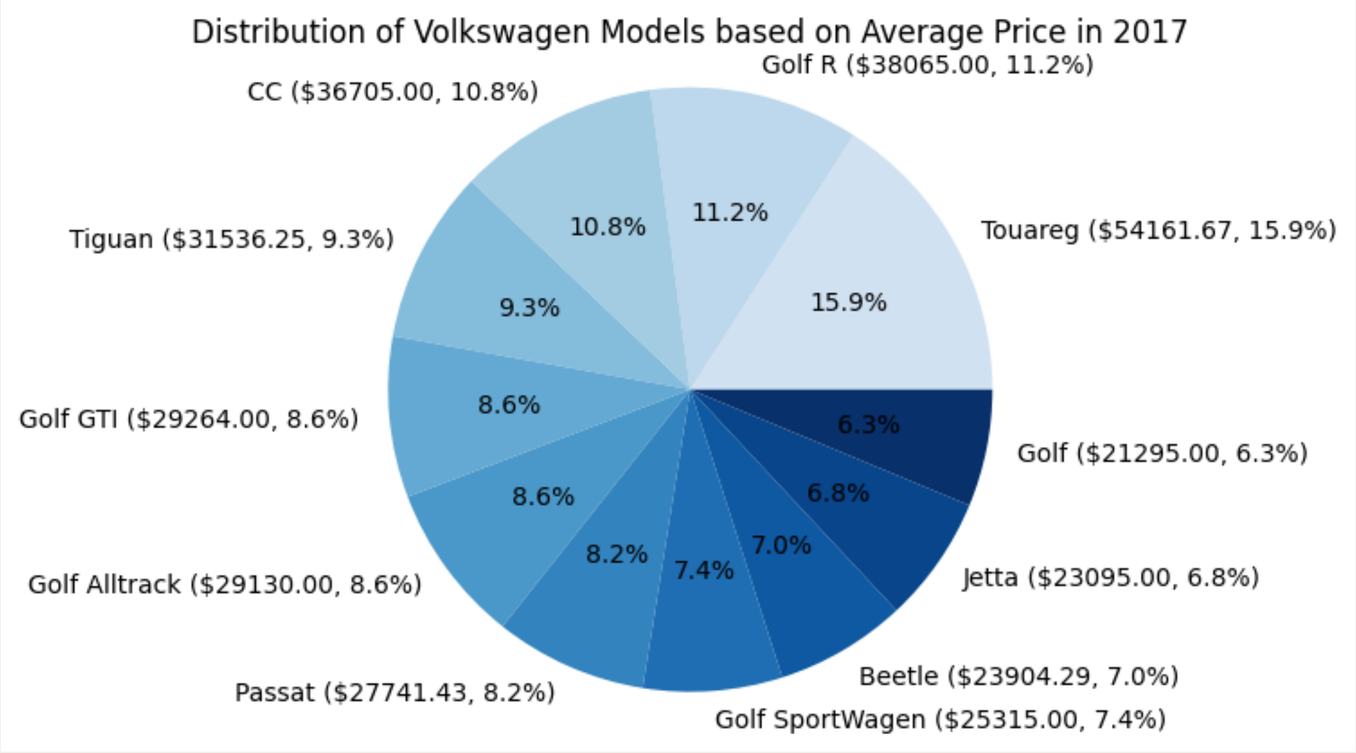
Large size cars have greatest average price followed by medium and compact

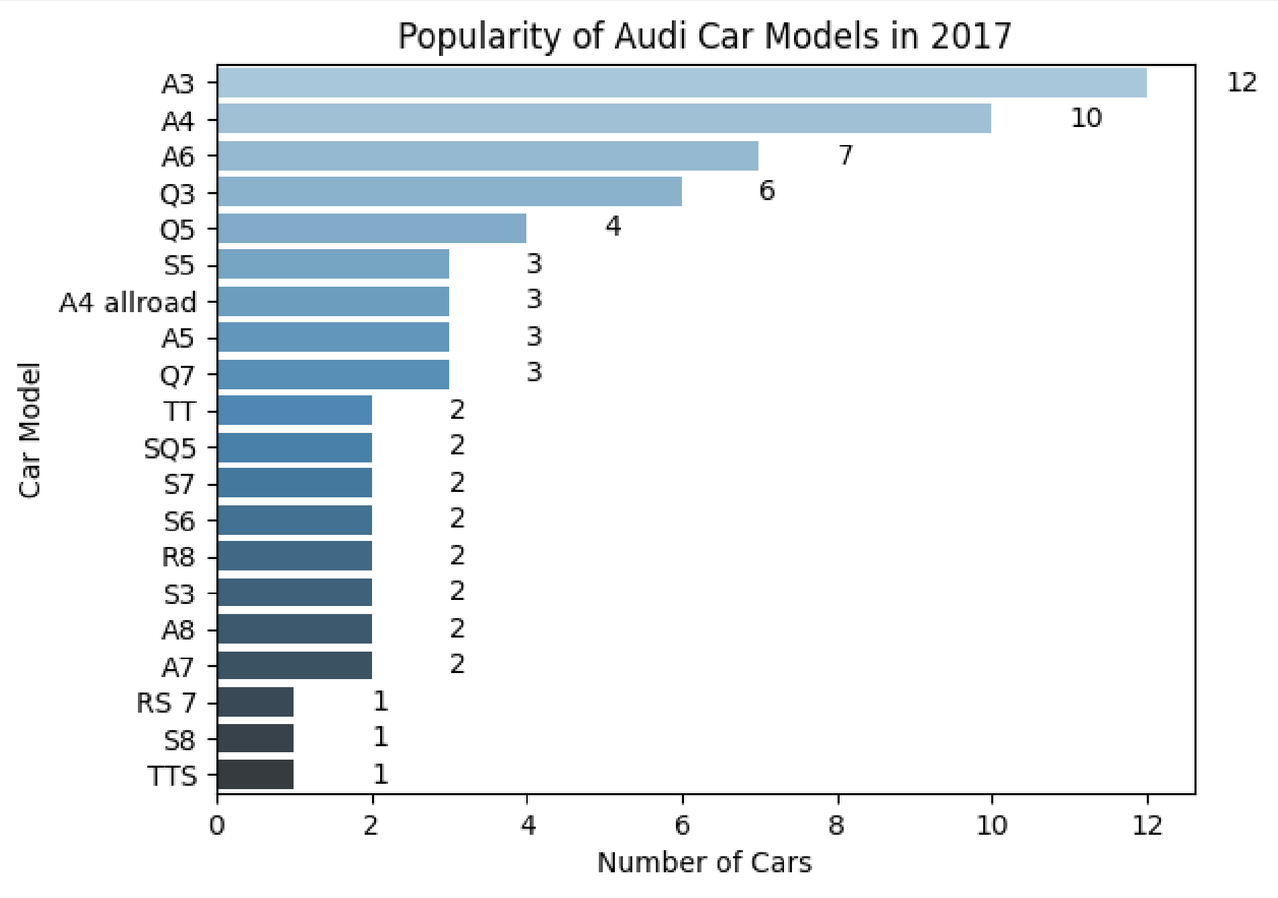


The insight:

The Golf GTI model was the most popular car model.

It was among the most expensive car models.

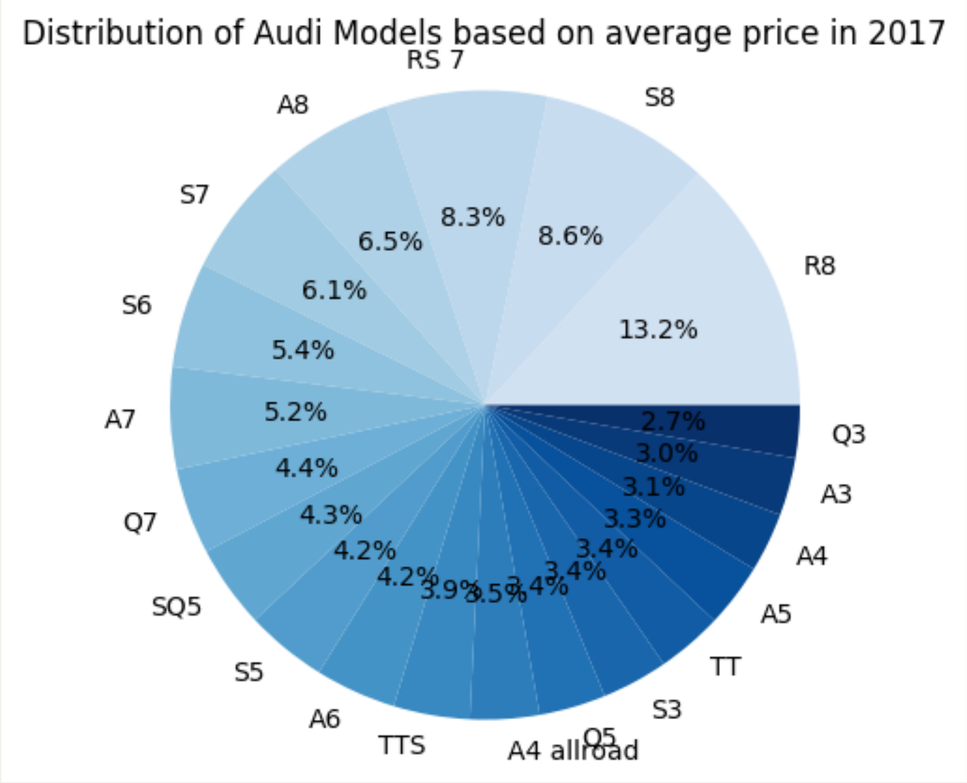


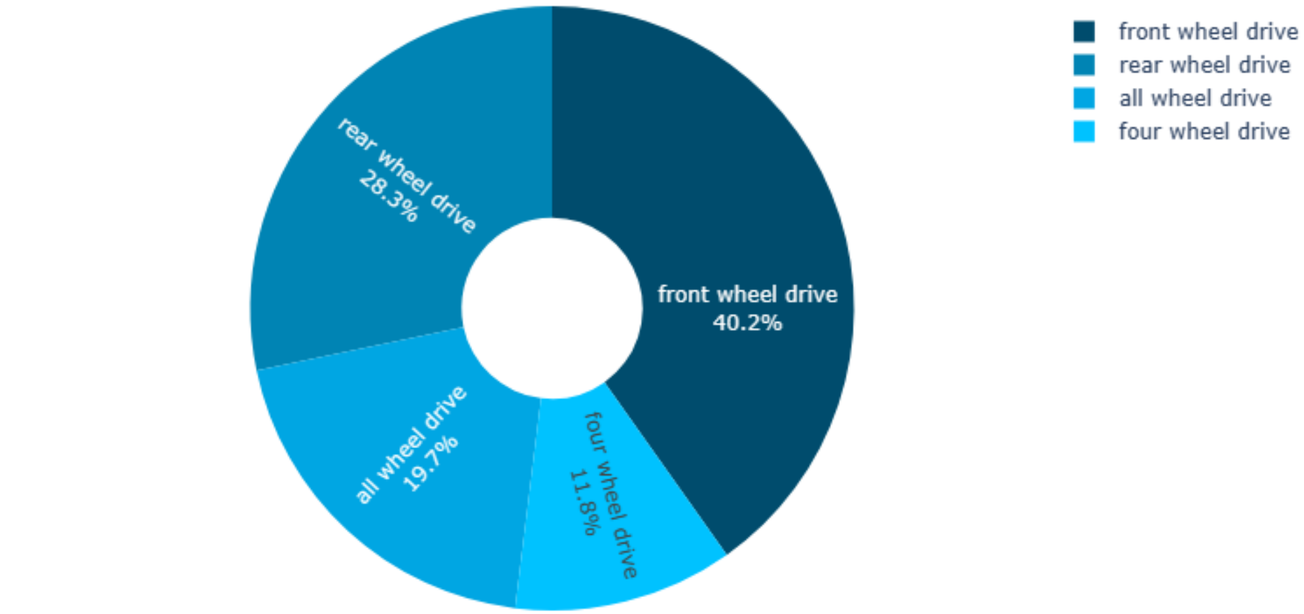


The insight:

A3 model has highest number of cars sold.

It was among the cheapest car models which is only 3% of distributions.



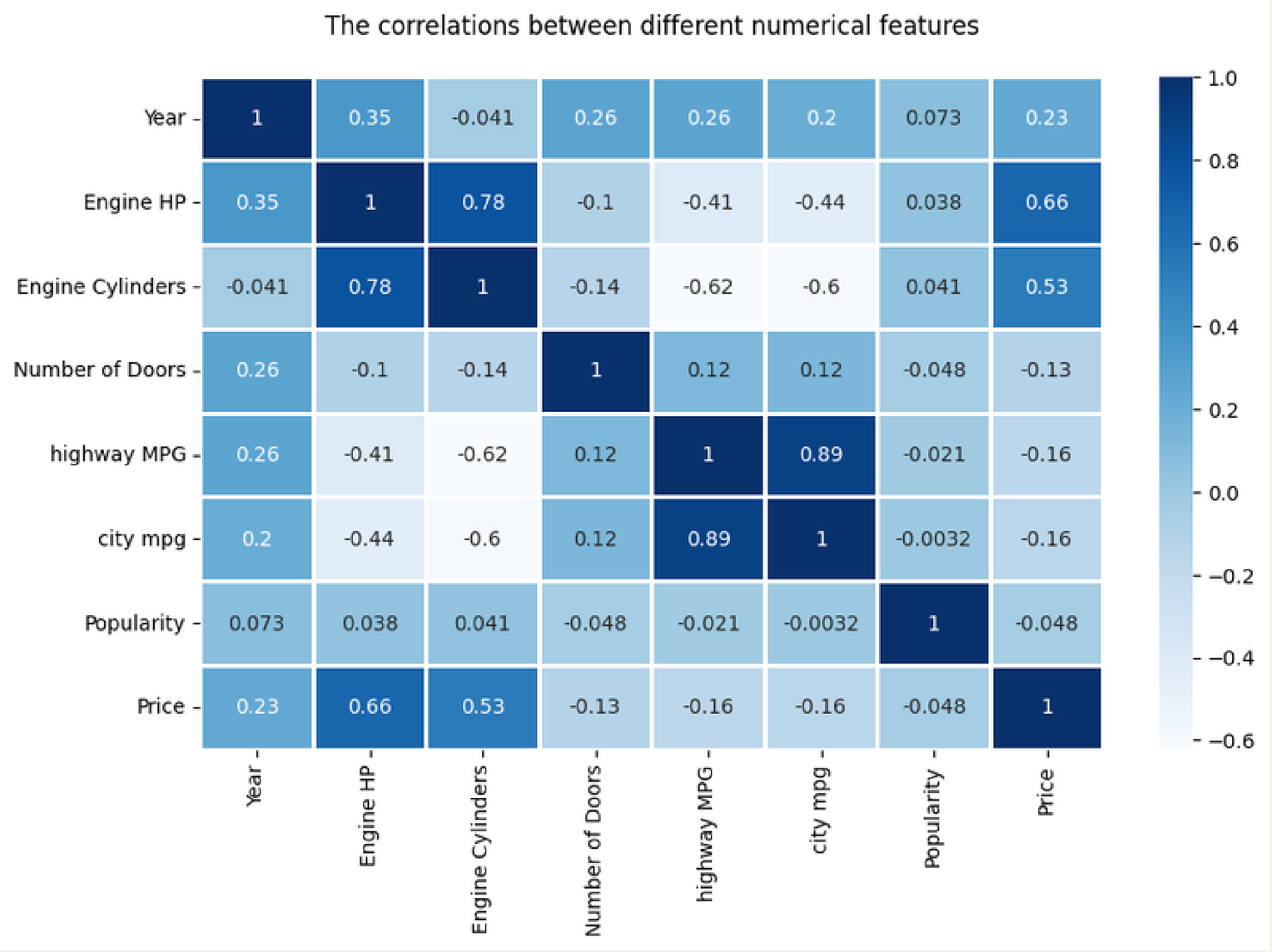


The insight:

Front-wheel drive is the most common driven wheel type.

Real-wheel drive (RWD) has a noticeable but comparatively smaller representation.

All-wheel drive (AWD) and four-wheel drive (4WD) have a relatively smaller percentage in the dataset.



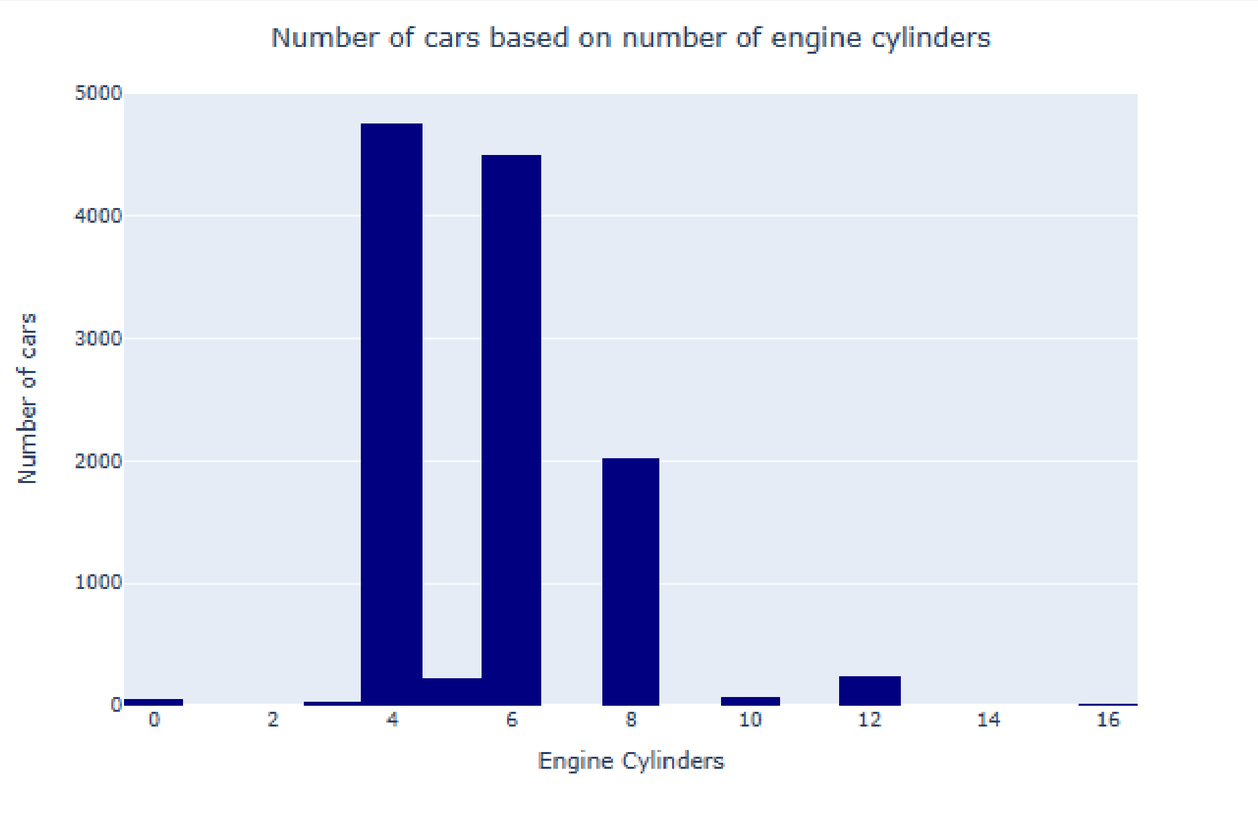
The insight:

city MPG and highway MPG has highest ratio of correlation.

City driving involves frequent stops and congested traffic. Highway consistent speeds and fewer stops.

Having lower fuel efficiency in city driving conditions.

Engine HP and engine cylinders are the most significant features that affect the price.

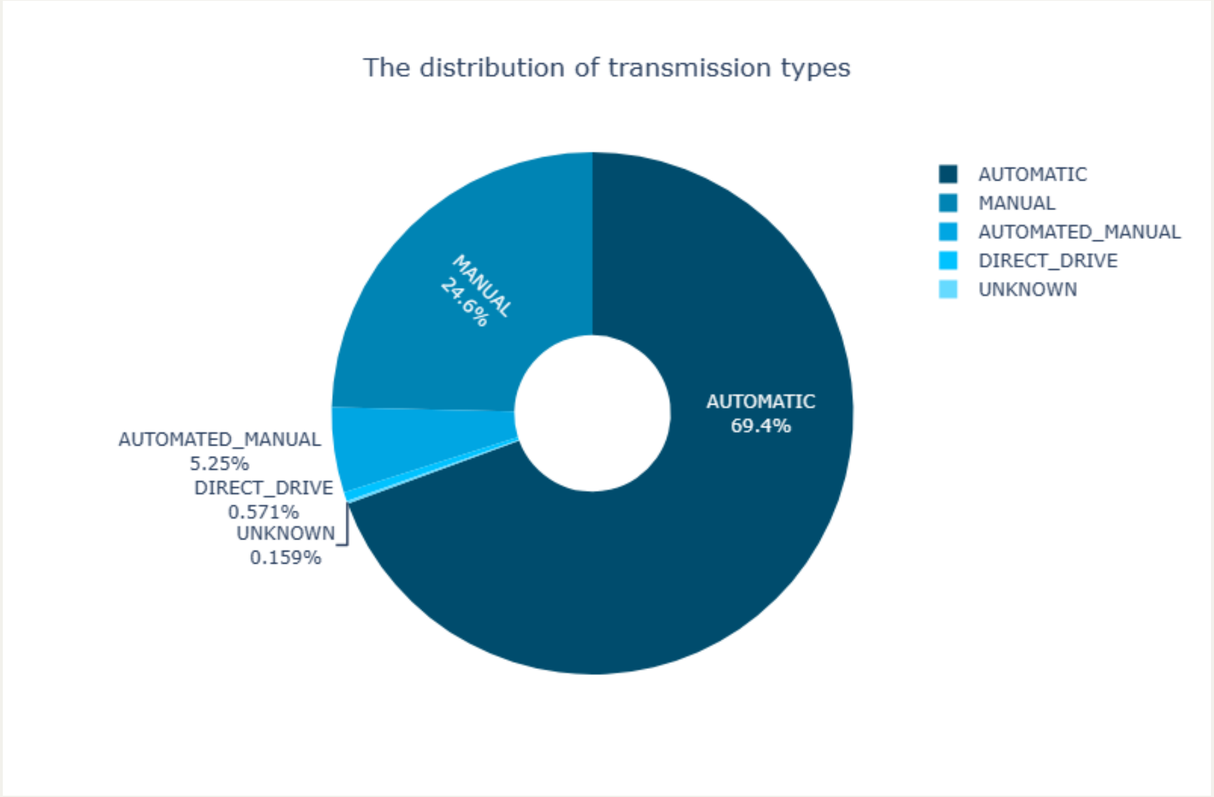


The insight:

Bugatti model has highest number of engine cylinders (16).

The specifications as follows:

```
Company                                 Bugatti
Model                                Veyron 16.4
Year                                        2008
Engine Fuel Type     premium unleaded (required)
Engine HP                                1001.0
Engine Cylinders                           16.0
Transmission Type                AUTOMATED_MANUAL
Driven_Wheels                 all wheel drive
Number of Doors                            2.0
Market Category          Exotic,High-Performance
Vehicle Size                            Compact
Vehicle Style                             Coupe
highway MPG                                  14
city mpg                                      8
Popularity                                  820
Price                                   2065902
Name: 11362, dtype: object
```

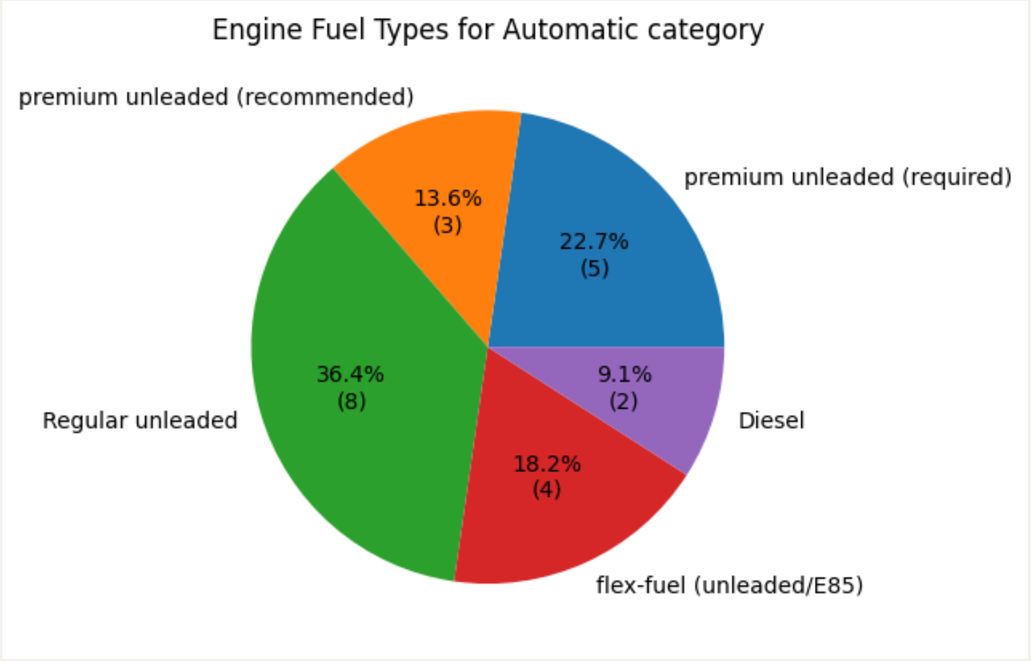


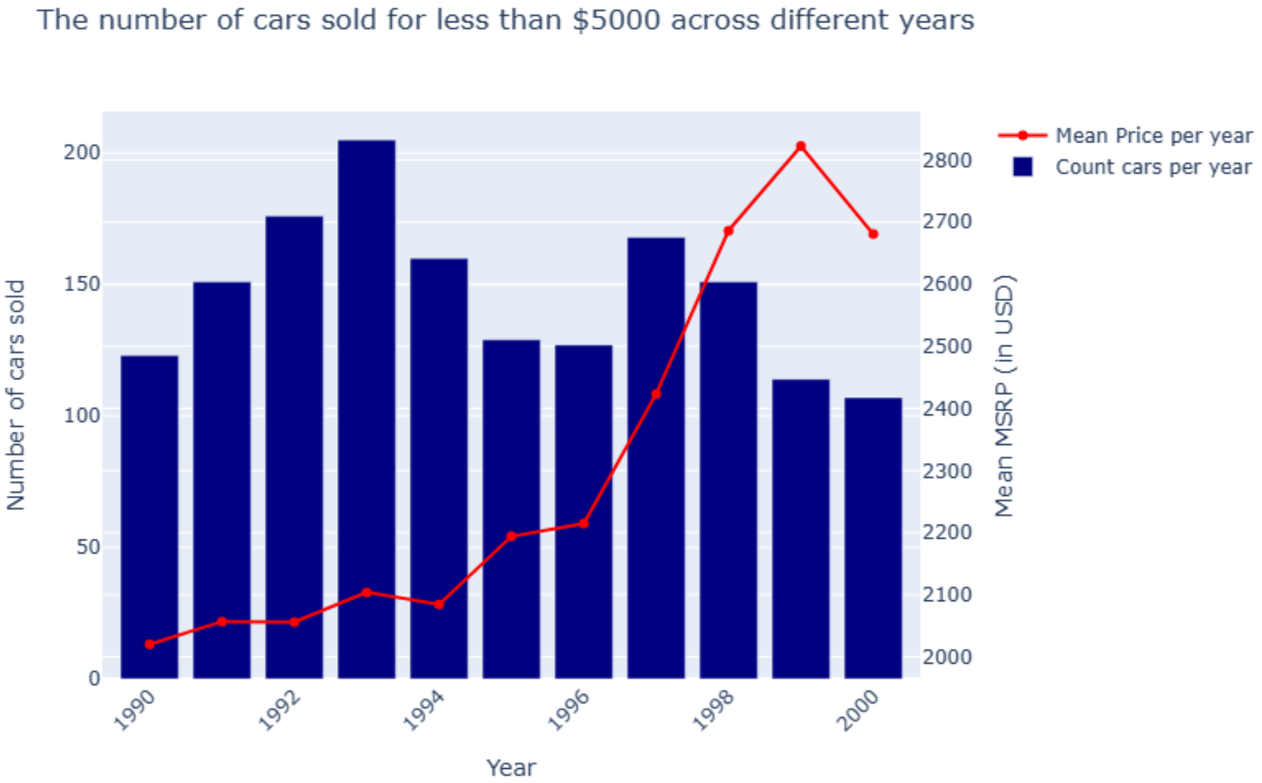
The insight:

Almost 70% of cars are automatic.

For decades automatic transmission have been considered much more convenient in the USA.

The American car industry was more competitive, with more powerful engines and cheap fuel.





The insight:

An increase in the average price of MSRP.

No models marketed after the 2000s appear.

MODELLING AND EVALUATION

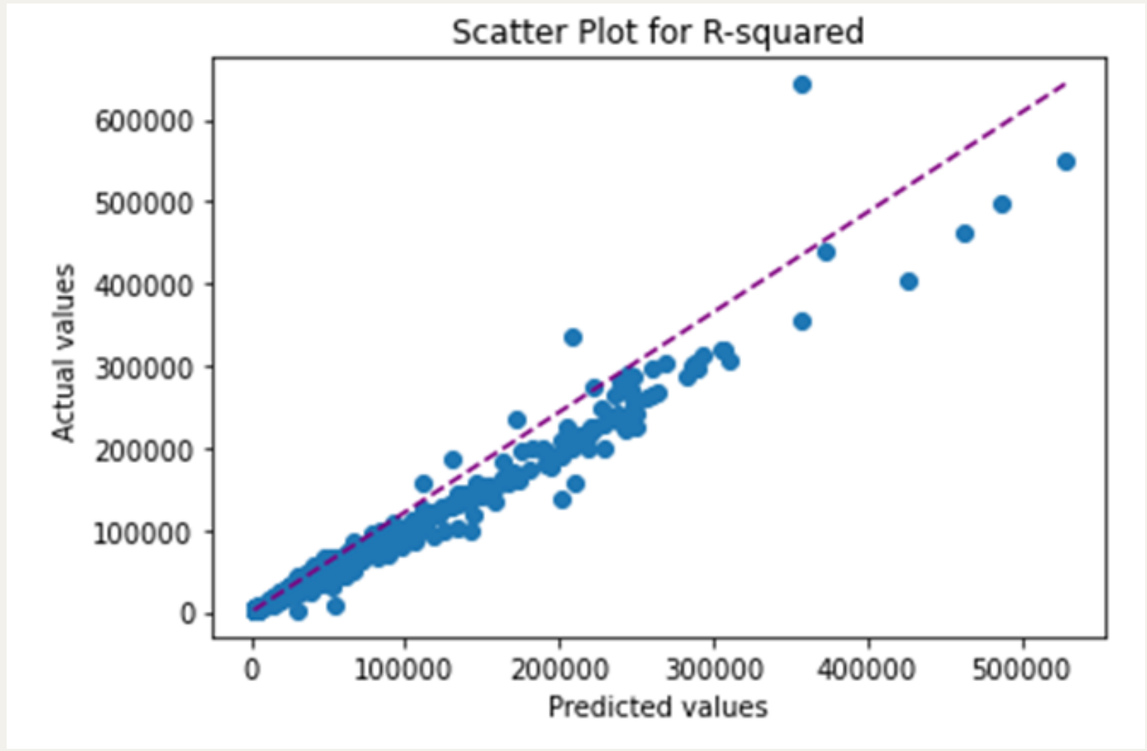
RANDOM FOREST

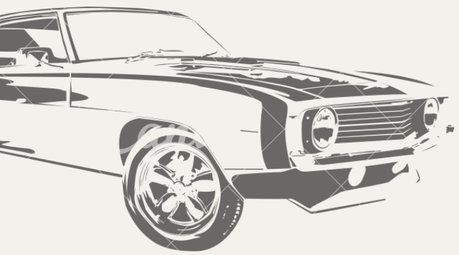
```
# Step 5: Evaluate the model
r2 = r2_score(y_test, y_pred)
print('R-squared (R2):', r2)
```

```
R-squared (R2): 0.968904336402638
```

R2 score of car price = 0.968904336402638

Scatter plot of R2 score



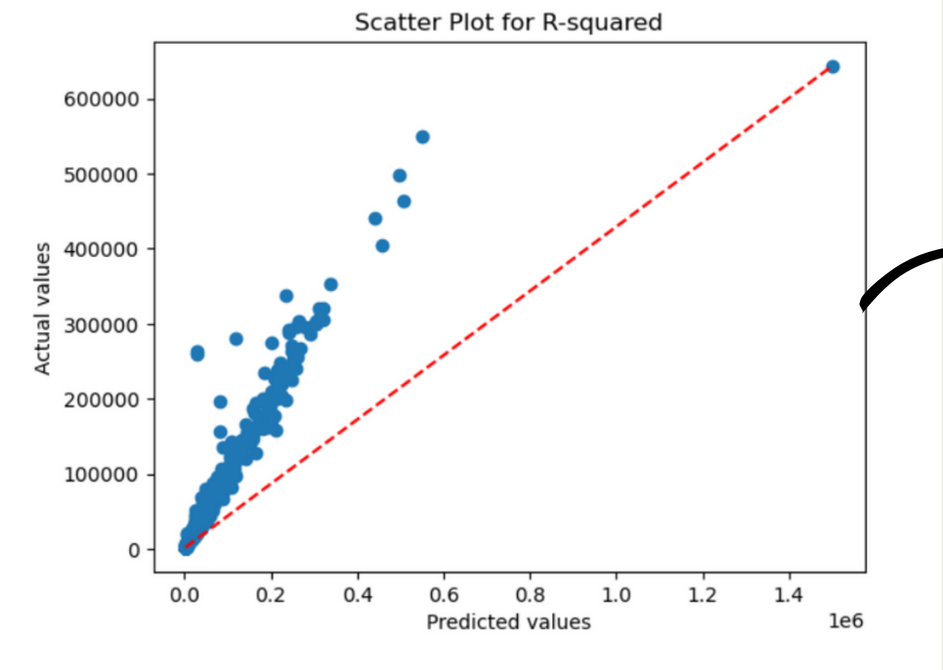


DECISION TREE

R2 score of car price = 0.8259107418283362

```
# Evaluate the model
r2 = r2_score(y_test, y_pred)

# Print the evaluation metrics
print("R-squared:", r2)
```

```
R-squared: 0.8259107418283362
```



Scatter plot of R2 score

Commercial Benefits

Allows manufacturers and dealers to determine the optimal pricing strategy for their vehicles

Manufacturers to identify specific customer segments based on their preferences, needs, and price sensitivity.

Identify customer preferences and demands

Car buyers extensively research and compare different models before making a purchase decision.

THANK YOU

CAPSTONE PROJECT (GROUP 7)

PRESENTED BY:
AINA SYAZZWEEN SURAYA
THVEYA A/P MAHENDRAN
NURSYAZA NISSA
NUR SYUHADA