

Car Sales EDA



INTRODUCTION

1. Finding patterns in Data
2. Determining relationships in Data
3. Checking of assumptions
4. Preliminary selection of appropriate models
5. Detection of mistakes

EDA ESSENTIALS

1. Problem Statement
2. Importing Packages
3. Shortlisting Data and Columns
4. Purification of Data
5. Analysis (Finding patterns in Data)
 - a. Univariate
 - b. Bivariate
 - c. Multivariate
6. Uncover insights (Answer questions in the data)
7. Conclude (Story Telling)
8. Actionable Insights

PROBLEM STATEMENT

1. Which brand of the cars sold most?
2. What all the factors are affecting to the car buyers?
3. What all features most likely for high profile buyers?

Got this data from INSAID to prepare Exploratory Data Analysis (EDA) with basic use of Pandas which includes cleaning, munging, combining, reshaping, slicing, dicing and transforming data for analysis purpose.

IMPORTING PACKAGES

1. **Numpy** - Implementing multi-dimensional array and matrices.
2. **Pandas** - For data manipulation and analysis.
3. **Pandas Profiling** - Providing analysis like type, unique values, missing values, quantile statistics, mean, mode, median, standard deviation, sum, skewness, frequent values, histograms, correlation between variables, count, heatmap visualization, etc.
4. **Matplotlib** - Plotting library for Python programming language and it's numerical mathematics extension NumPy.
5. **Seaborn** - Provides a high level interface for drawing attractive and informative statistical graphics.

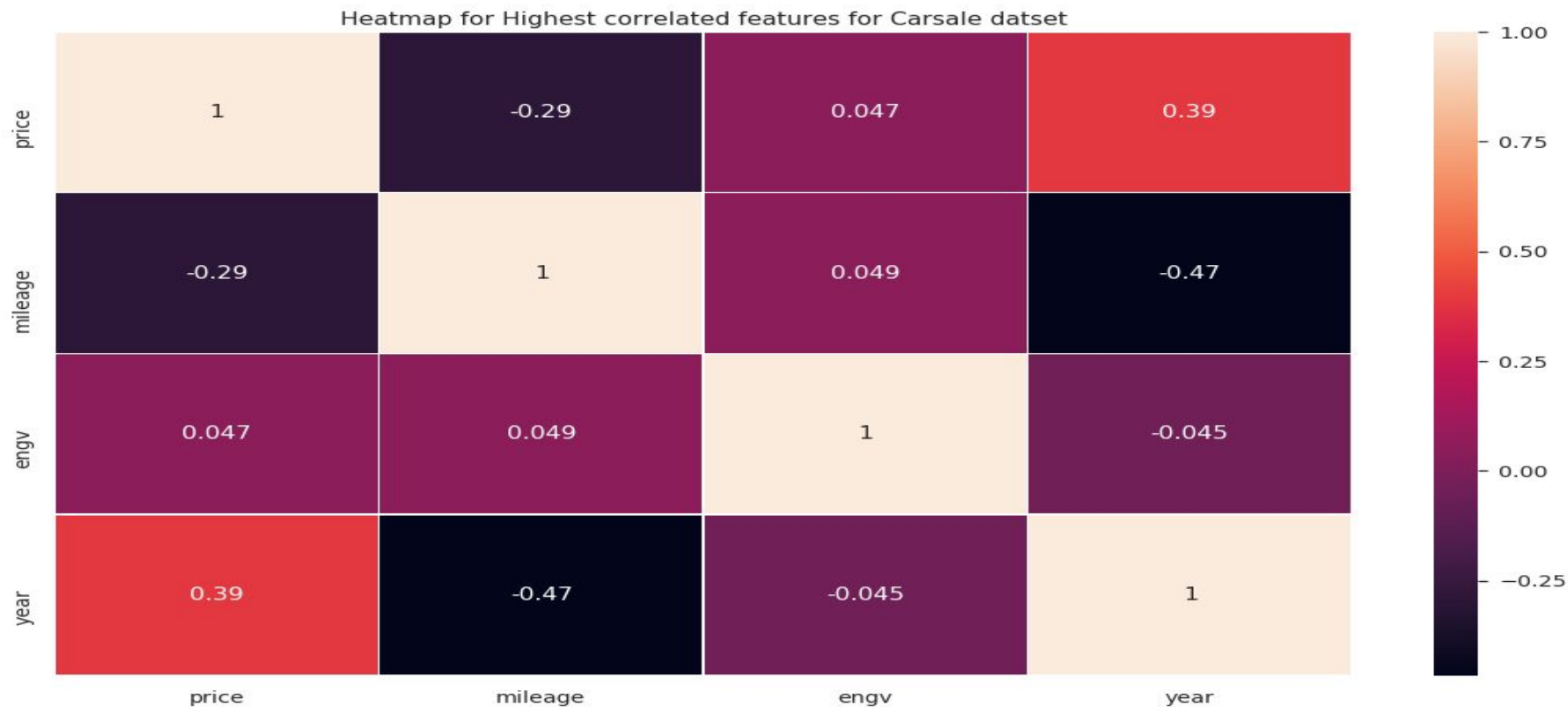
INITIAL OBSERVATIONS

1. This dataset has **9576** rows and **10** columns.
2. Summary of data types in this dataset:
 - a. **Numeric:** price (Float), Mileage (Integer), engV (Float), Year (Integer)
 - b. **Boolean:** registration
 - c. **Object:** Everything else
3. The following variables have null and zero values that may need to be addressed.
 - a. **Car** has a high cardinality **87 distinct values**.
 - b. **Drive** has 511 (5.3%) missing values.
 - c. **Eng Volume** has 434 (4.5%) missing values.
 - d. **Mileage** has 348 (3.6%) zeros.
 - e. **Model** has a high cardinality with 888 distinct values.
 - f. **Price** has 267 (2.4%) zeros.

Final Observations

1. **Price & Mileage:** Remove outliers as the columns have outlier as 0 and it is too low to be real value.
2. **Drive & engV:** Replaced the null values with the column mean value.
3. **Car** column looks good as we can expect **87** distinct values for Brand.
4. **Model** seems not useful for analysis and need to remove or ignore this column.
5. This dataset has **201** duplicate rows, which was removed here and now **9375** unique rows are present in dataset.

HEATMAP FOR HIGHEST CORRELATED FEATURES FOR car sale DATA SET (CONTINUED...)

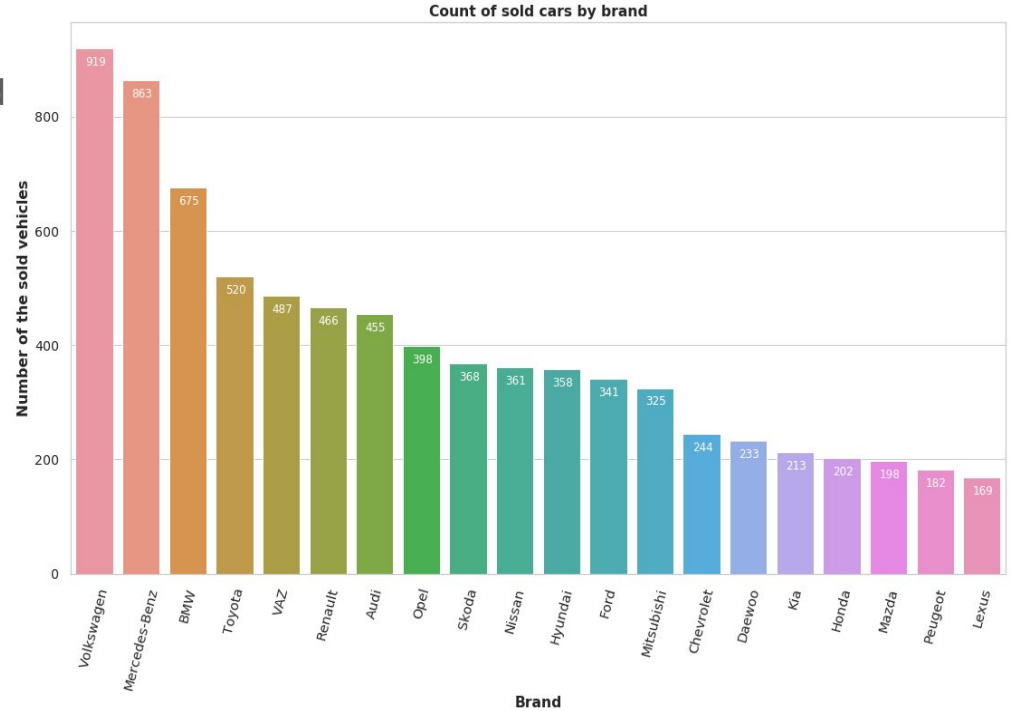


1. **price** and **production** year variables have **strong positive correlation**, i.e. latest production year cars have high price and older production year cars have lower price.
2. **price** and **engine volume (engV)** have **mild positive correlation**.
3. **engine volume (engV)** and **mileage** also have **some mild positive correlation**, generally engine volume and mileage do not have positive correlation, but in our data most of the cars have engine volume in between 1.5 CC to 2.5 CC, due to that mileage and engine volume have mild positive correlation.
4. **production year** and **mileage** variables have **strong negative correlation**, i.e. latest production year cars have high mileage and older production year cars have low mileage.
5. **price** and **mileage** have **negative correlation**, i.e. lower price cars have higher mileage and higher price cars have lower mileage.
6. **production year** and **engine volume (engV)** also have **some negative correlation**.

Univariate Analysis

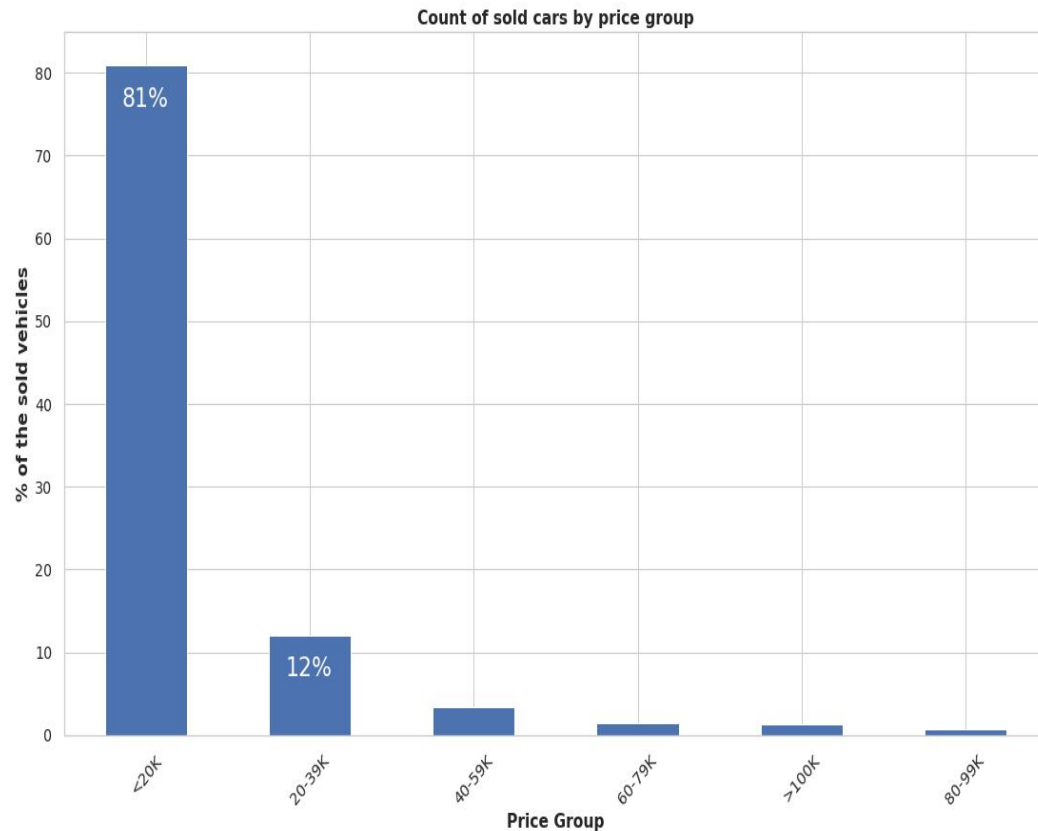
WHICH BRAND OF cars SOLD THE MOST OF THE car?

The most cars sold by the company
Volkswagen followed by **Mercedes-Benz** and
BMW.



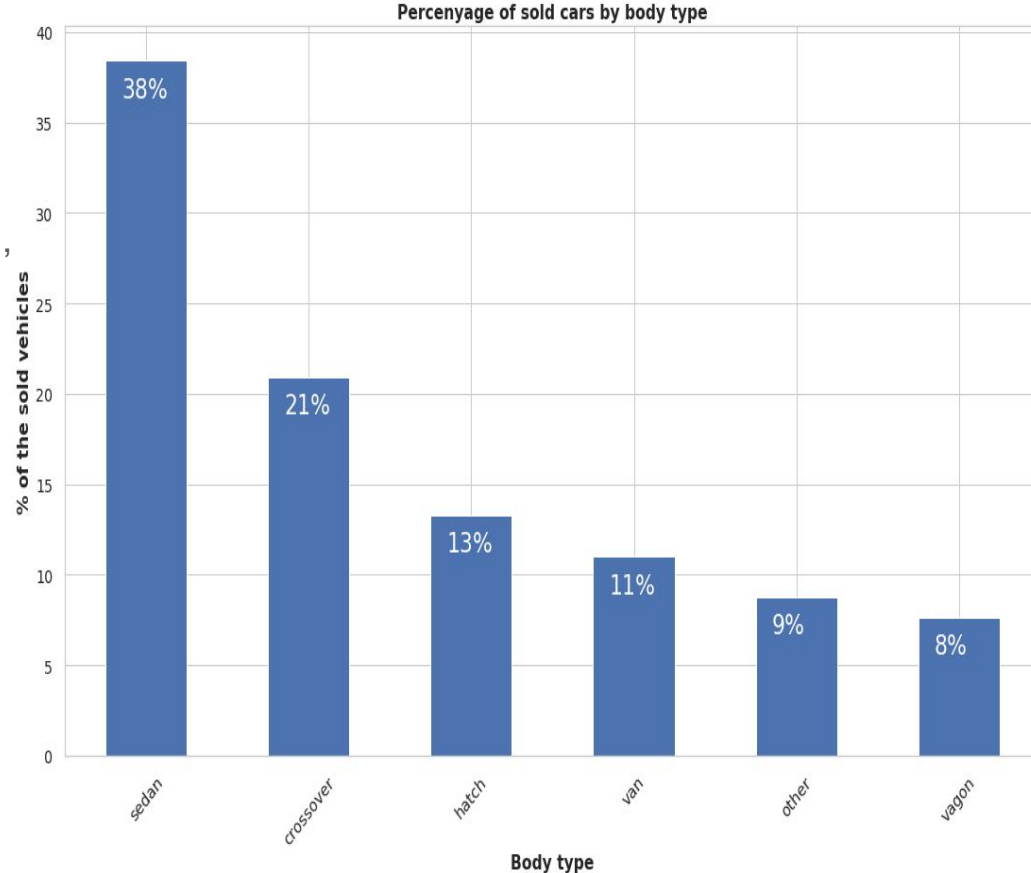
WHICH PRICE range OF cars SOLD THE MOST OF THE car?

Approx **81%** of cars sold in the **price range of less than or equal to \$20000**.



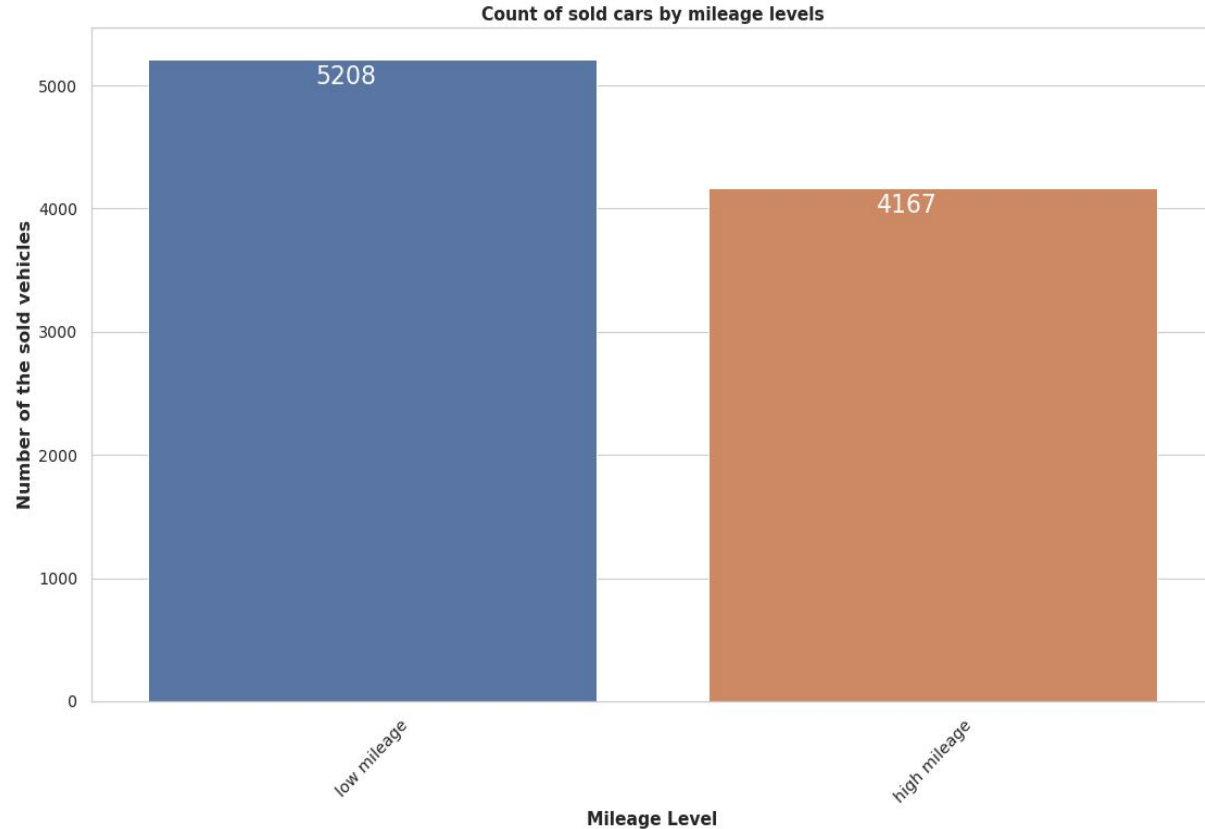
WHICH BODY TYPE OF CARS SOLD THE MOST OF THE car?

1. **Sedan** body type cars are the most sold cars.
2. We can say this in other terms that the most preferred choice for people is **Sedan**, because it is the perfect and comfortable car for small size family within lower budget.
3. Approx **38%** of total cars sold in **Sedan** body type, **21%** in **Crossover** body type and **13%** in **Hatch** body type.
4. **Sedan** are most preferred private cars among public vehicles vans are preferred.



WHAT IS VARIATION OF THE CAR SOLD IN TERMS OF MILEAGE OF THE CAR?

Low mileage vehicles are preferred.

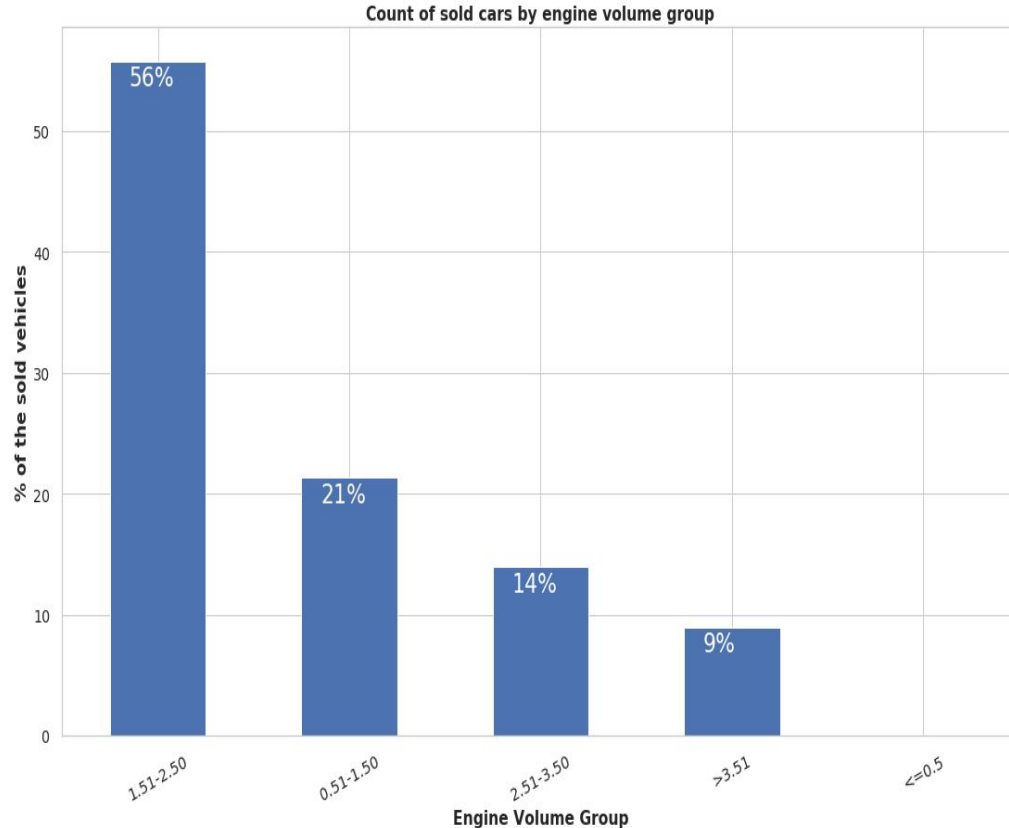


WHICH engine volume cars SOLD THE MOST OF THE car?

Below are the **top three categories of Engine Volume (engV)** cars sold.

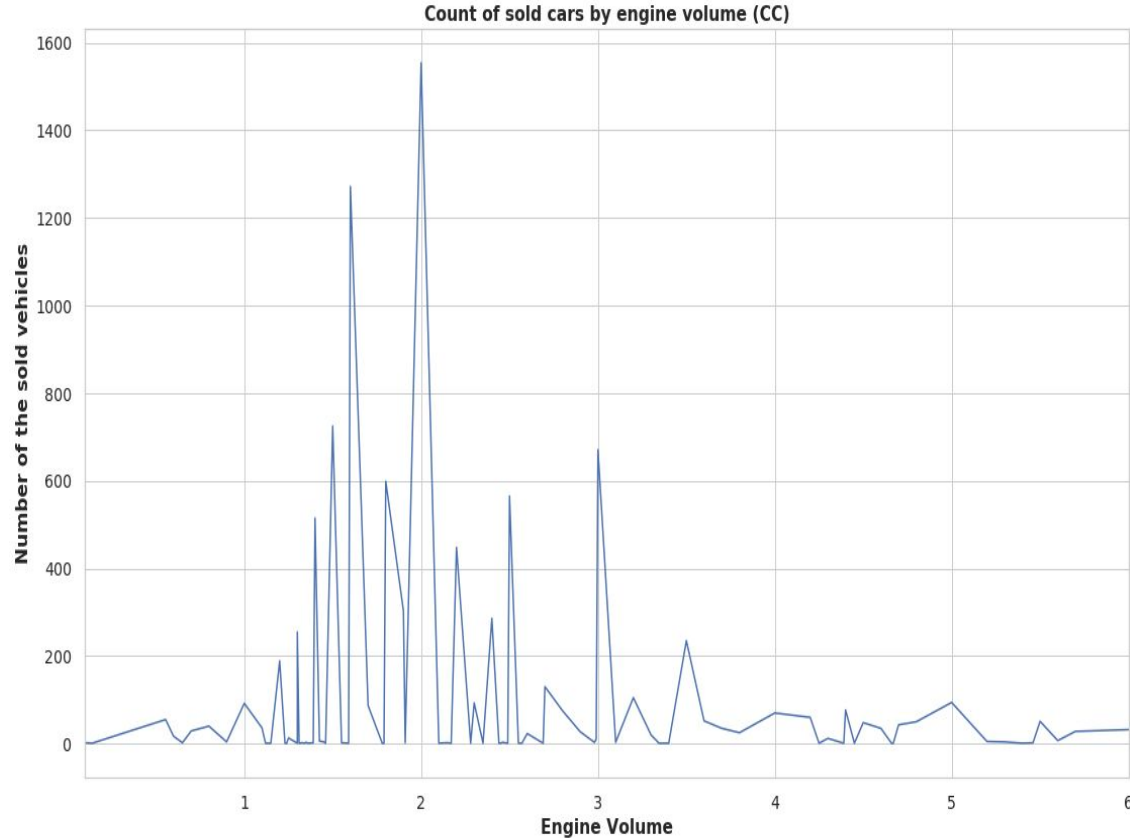
- **Ist Category** : > 1.5 & ≤ 2.5 (In this category approx **56%** of cars are available)
- **IInd Category** : > 0.5 & ≤ 1.5 (In this category approx **21%** of cars are available)
- **IIIRD Category** : > 2.5 & ≤ 3.5 (In this category approx **14%** of cars are available)

Out of total cars sold, approx **91%** cars fall in range of engine volume more than **1.5 CC to 3.5 CC**. Continued.....



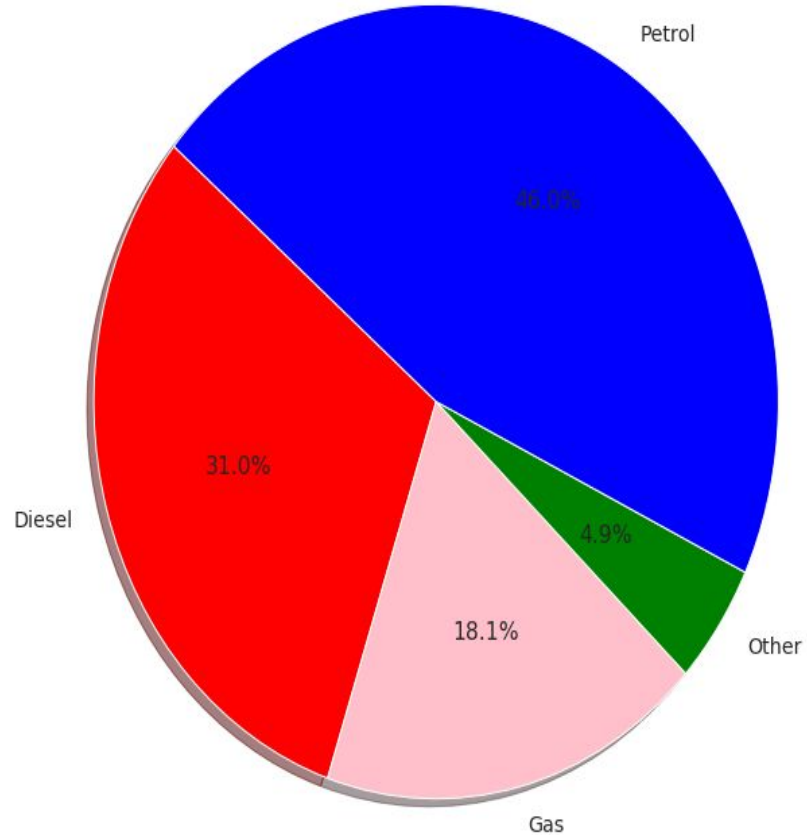
WHICH engine volume cars SOLD THE MOST OF THE car?

Most of the cars sold have **2.0 CC** engine volume followed by **1.5 CC**, there are very **few cars sold** with engine volume **less than equal to 1.0 CC** and **greater than equal to 4.0 CC**.



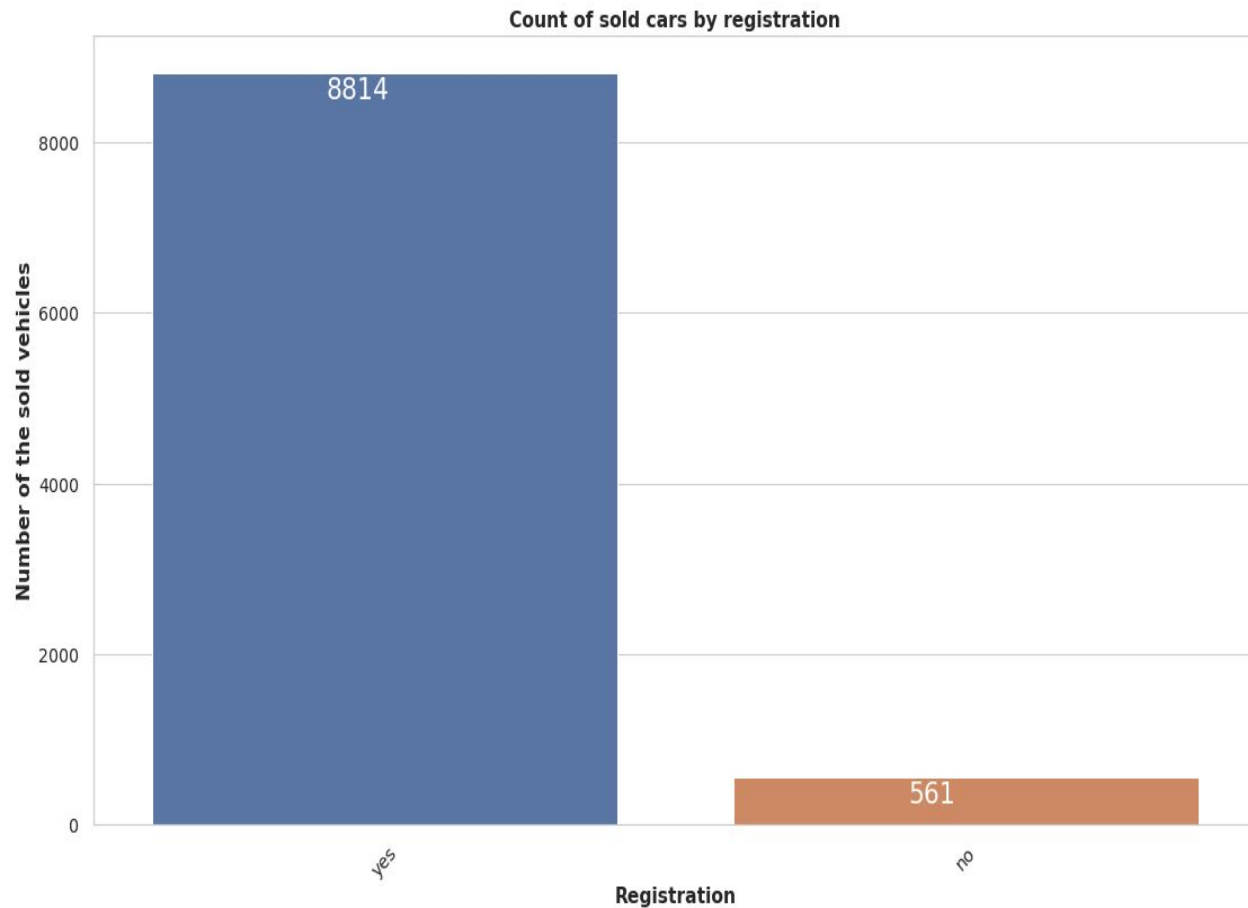
WHICH engine TYPE cars SOLD THE MOST OF THE car?

Petrol & Diesel cars are the most widely sold Engine Type. Gas and Other fuels are not so commonly used.



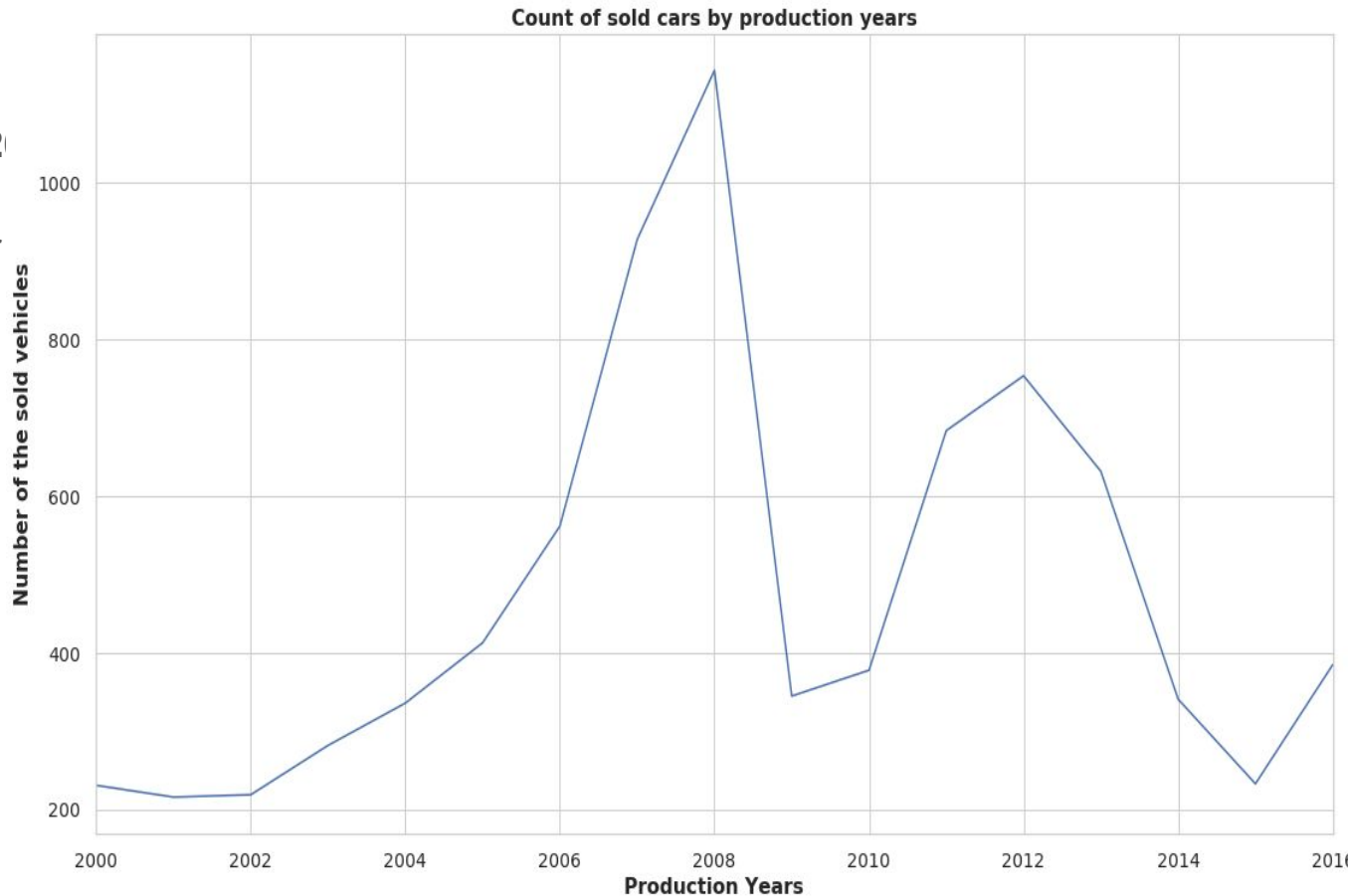
WHAT IS EFFECT OF THE CAR SOLD IN TERMS OF REGISTRATION OF THE CAR?

Registered vehicles are preferred.



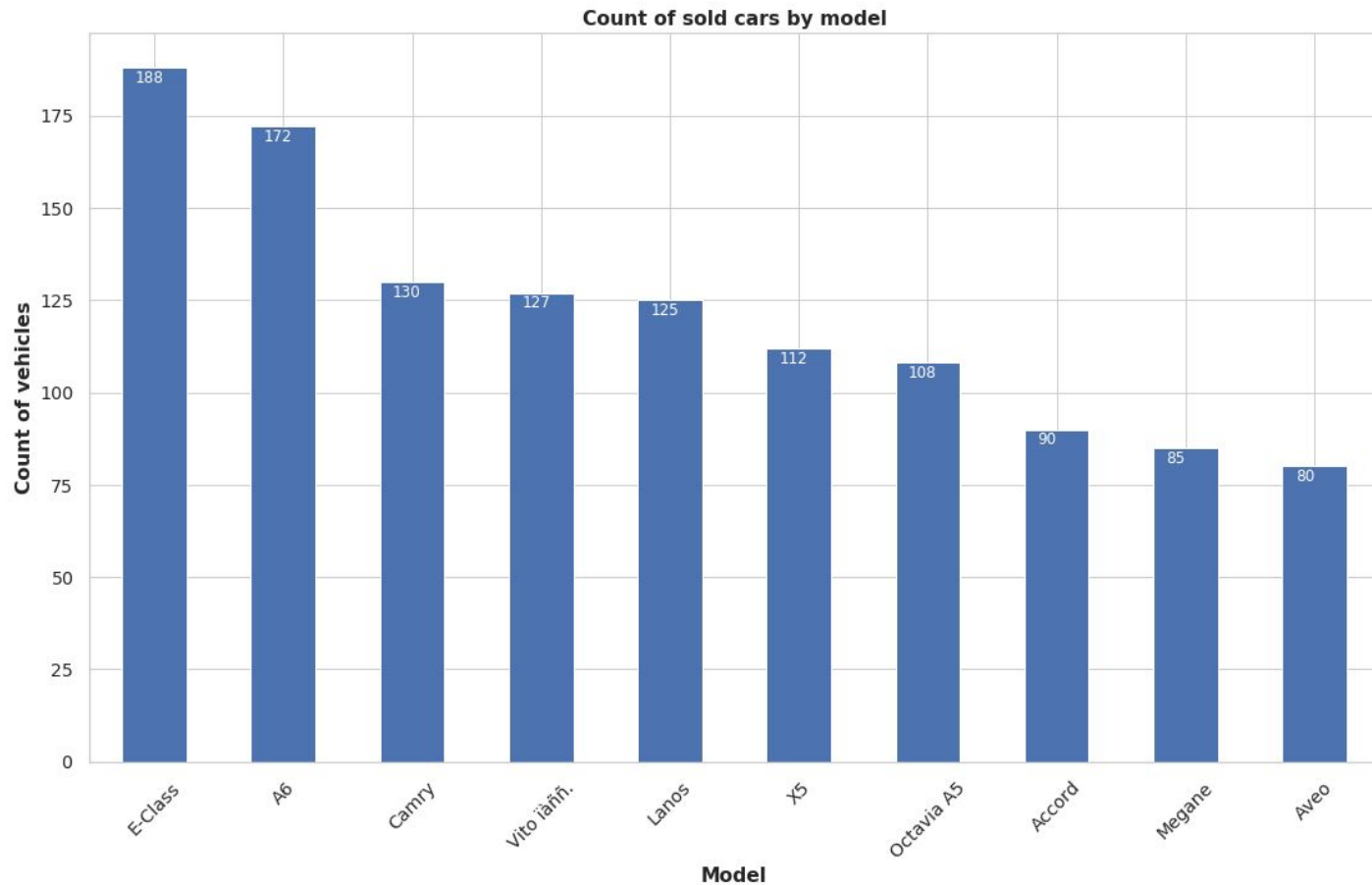
IN WHICH PRODUCTION year OF cars SOLD THE MOST OF THE car?

Production year 2008 and 2012 cars are sold the most. Year 2008 has the highest number car sales record.



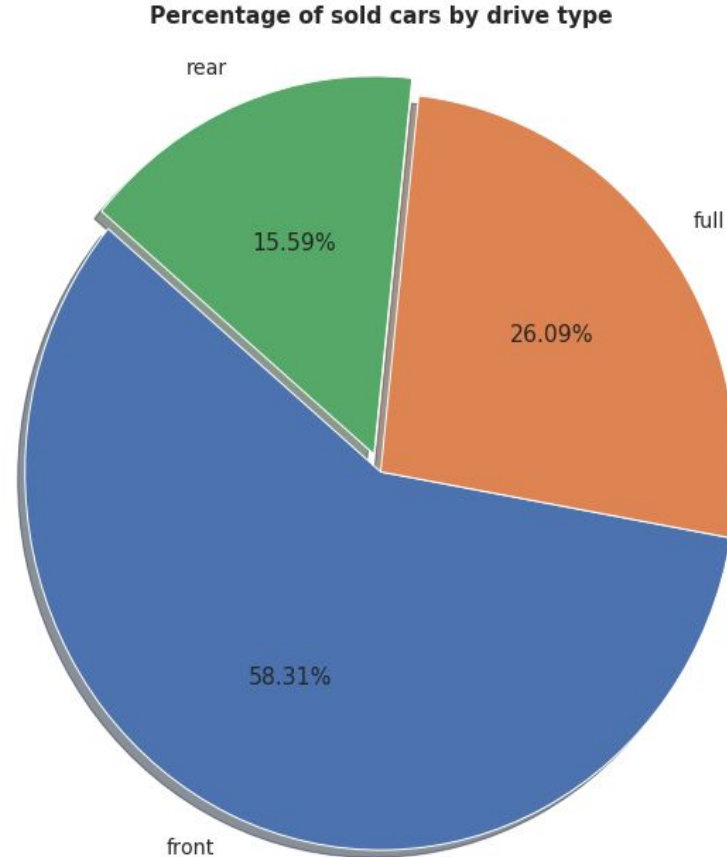
WHICH MODEL TYPE OF cars SOLD THE MOST OF THE car OF THE car?

E-Class car sold most.



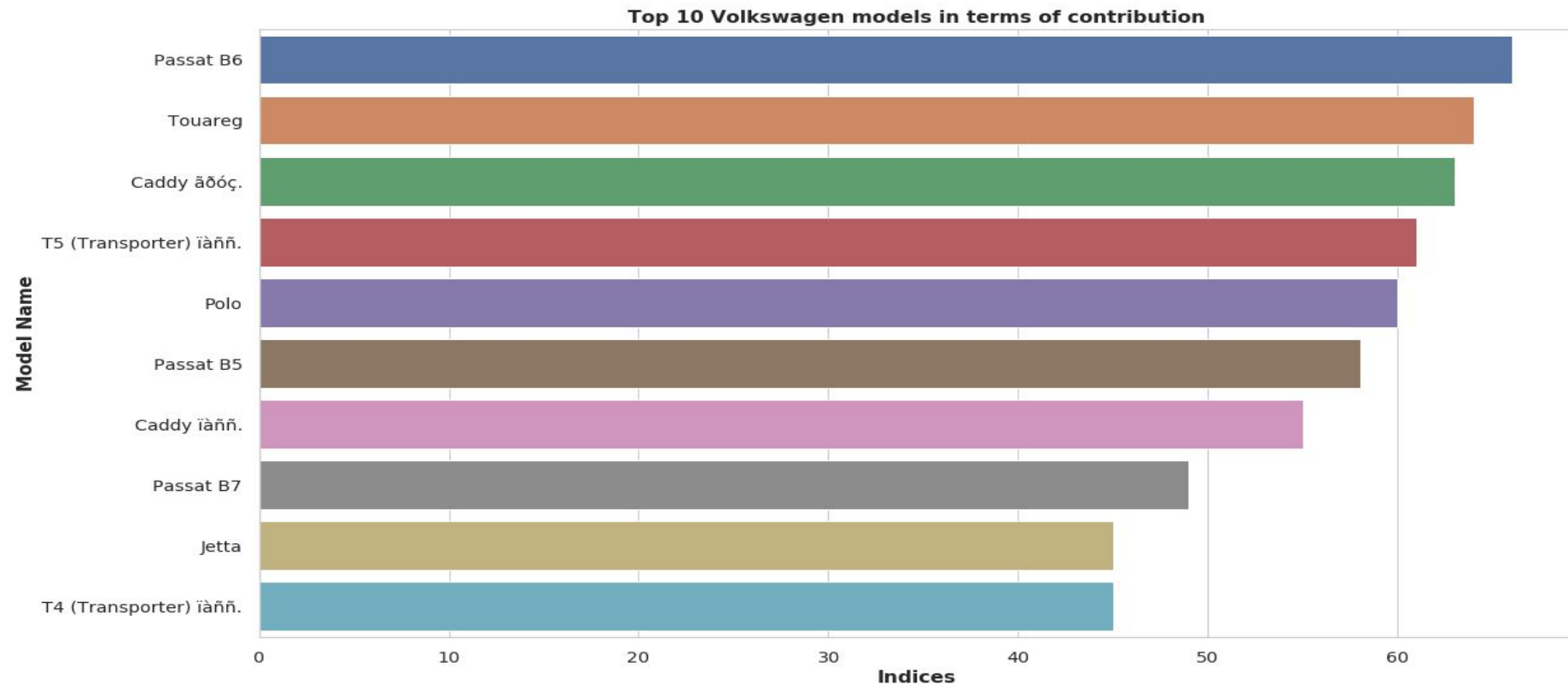
WHICH DRIVE TYPE OF cars SOLD THE MOST OF THE car?

Most sold cars in **front** drive type, in terms of percentage we can say that, approx 59% of total cars sold in **front** drive type, 26.09% in **full** drive type and 15.60% in **rear** drive type.

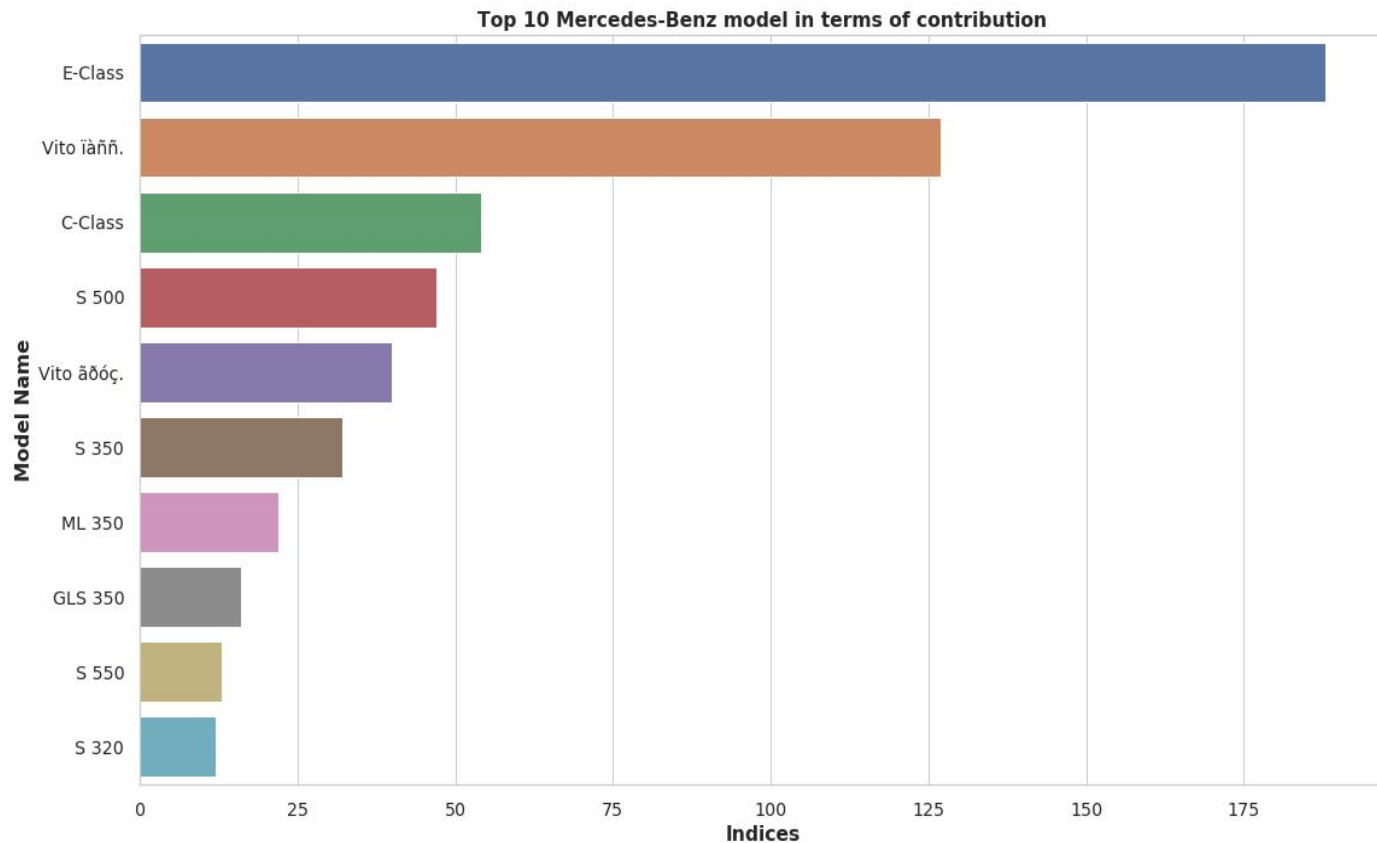


Bivariate Analysis

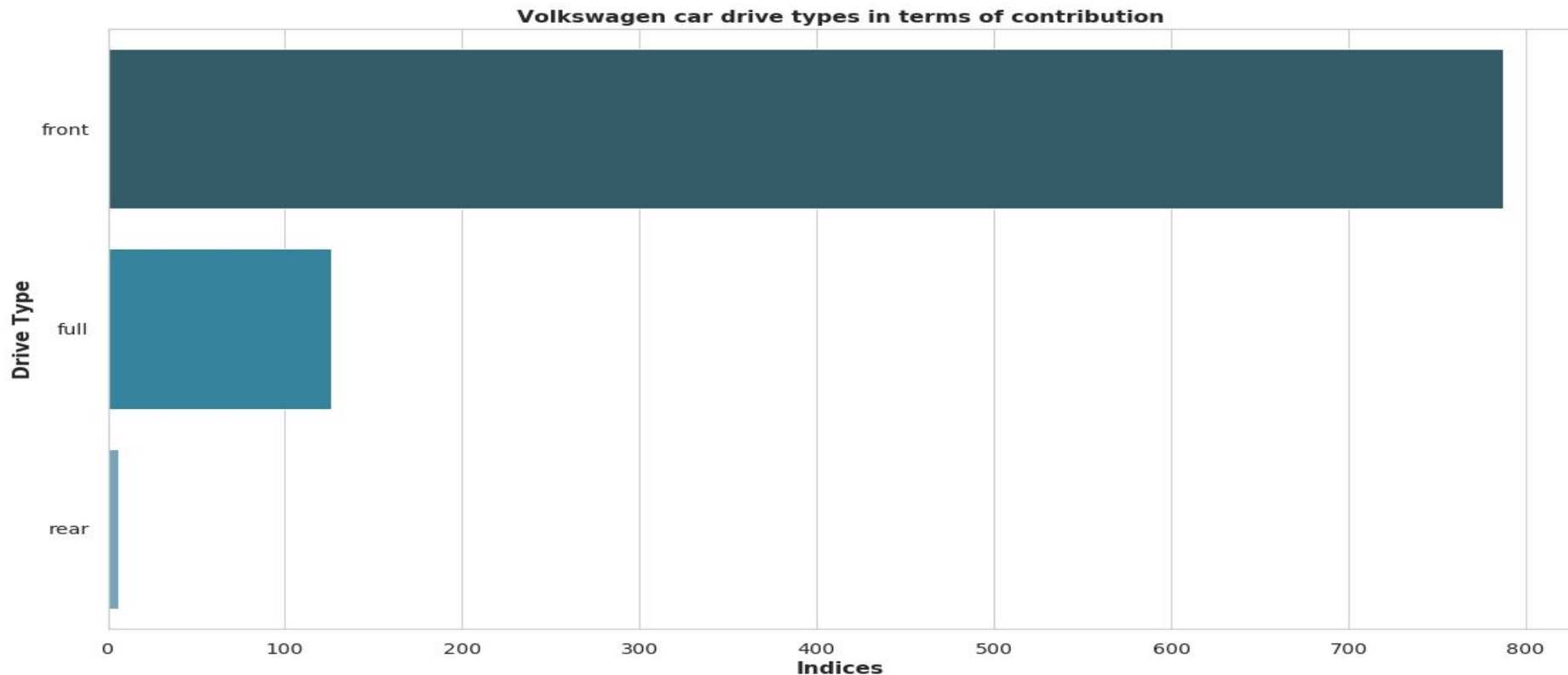
WHICH model IS THE MOST Preferred IN TOP Brands OF THE car? CONTINUED...



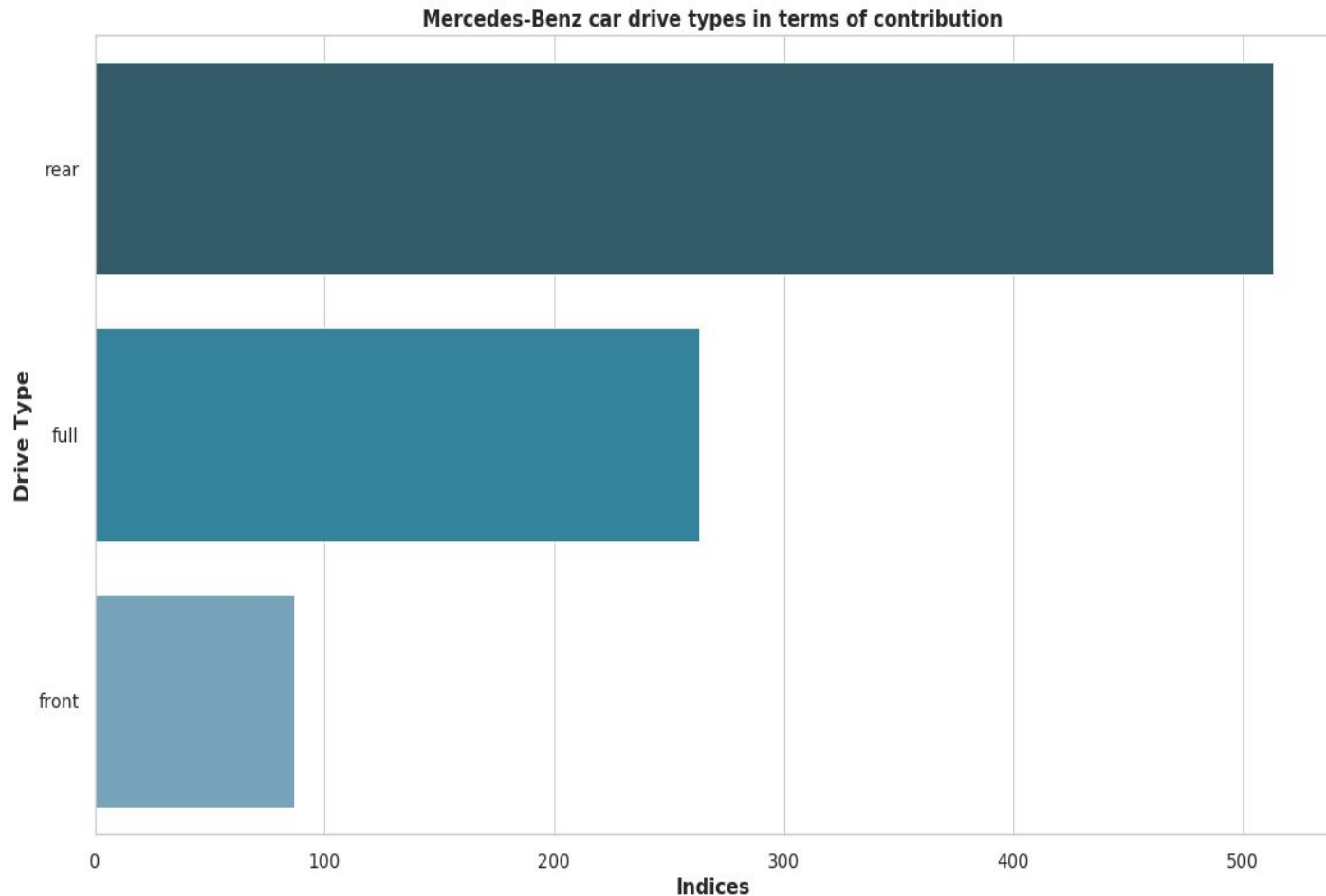
Passat & E-Class is the most preferred model in **Volkswagen** and **Mercedes-Benz** companies respectively.



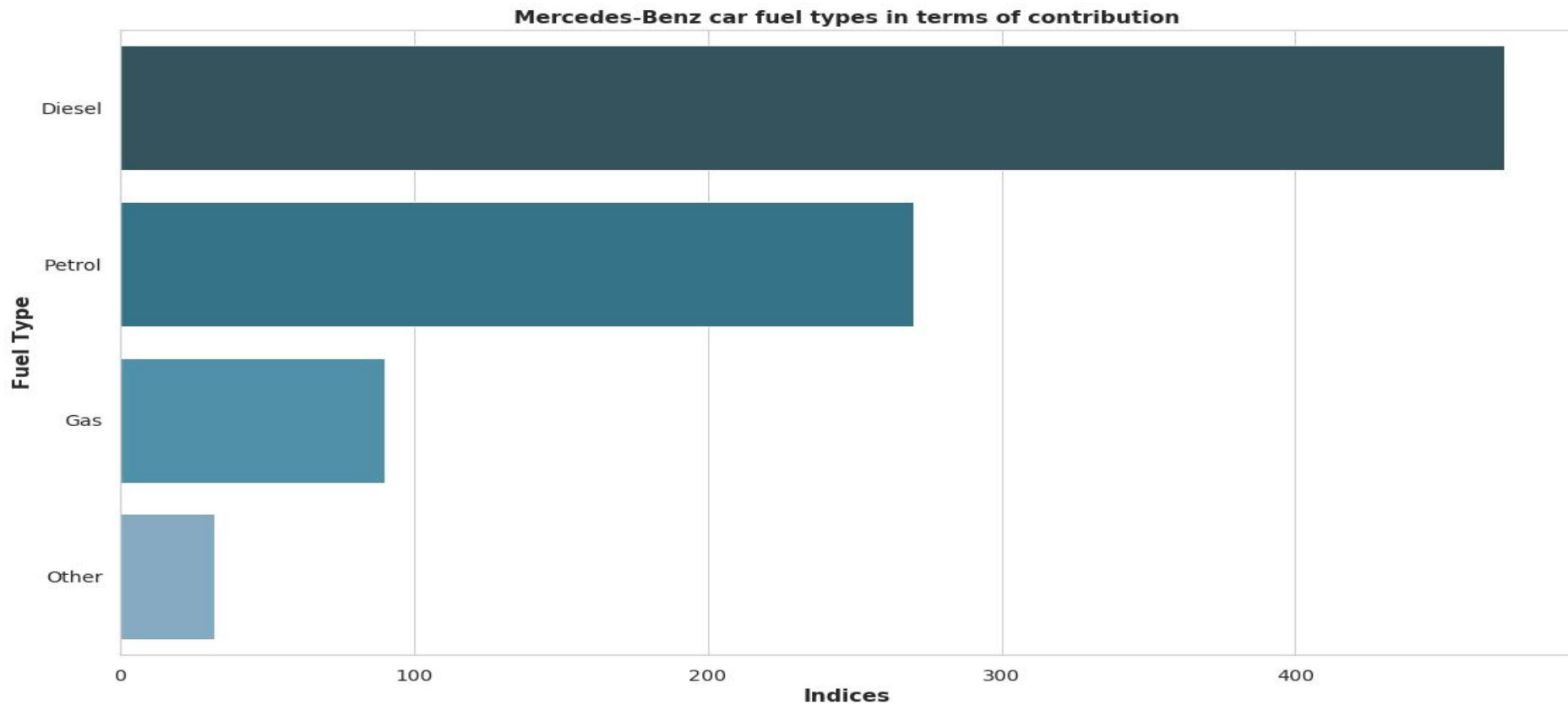
WHICH DRIVE TYPE IS THE MOST Preferred IN TOP BRANDS OF THE car? CONTINUED...



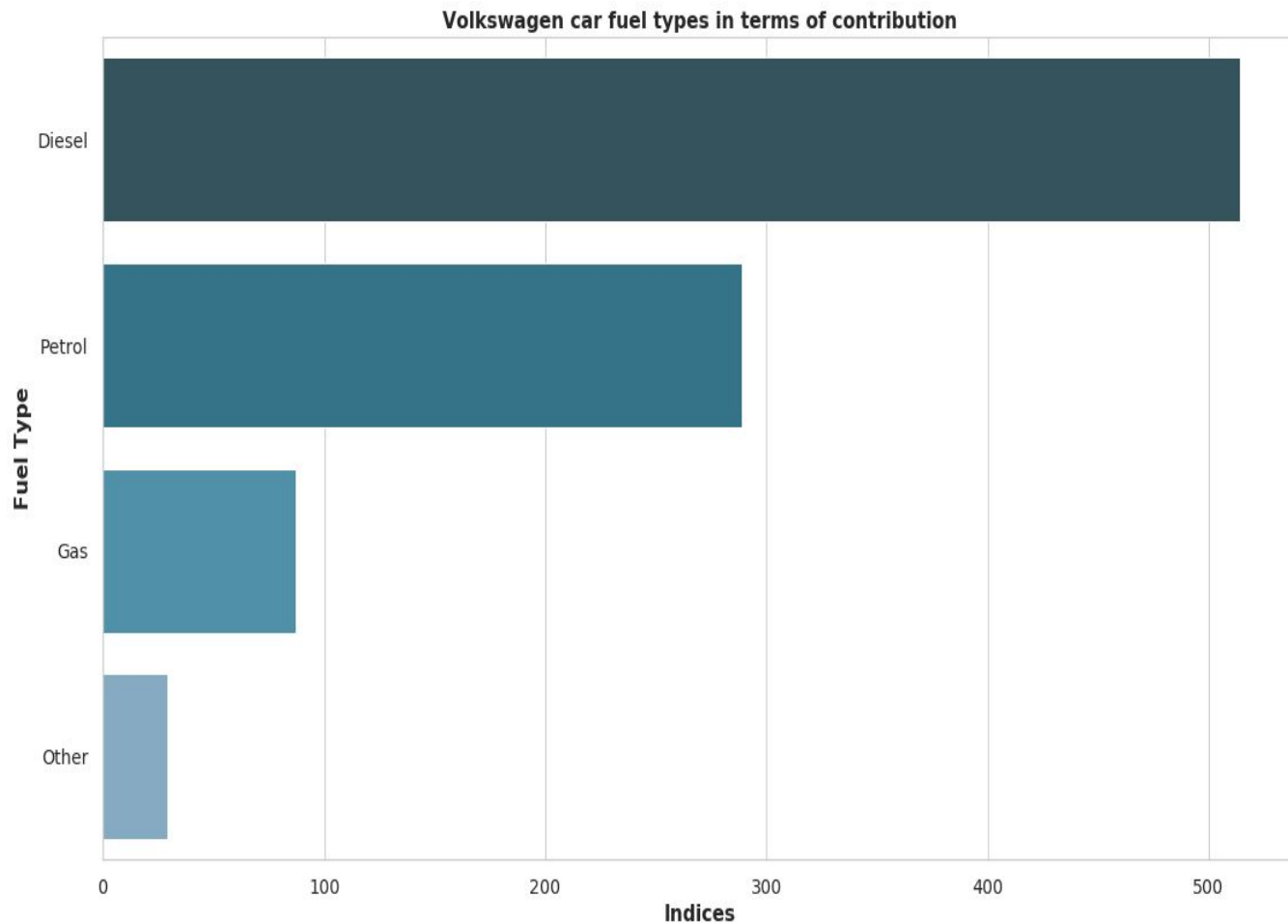
Front and rear drive is most preferred in **Volkswagen and Mercedes-Benz** companies respectively.



WHICH FUEL TYPE IS THE MOST PREFERRED IN TOP BRANDS OF THE car? CONTINUED...

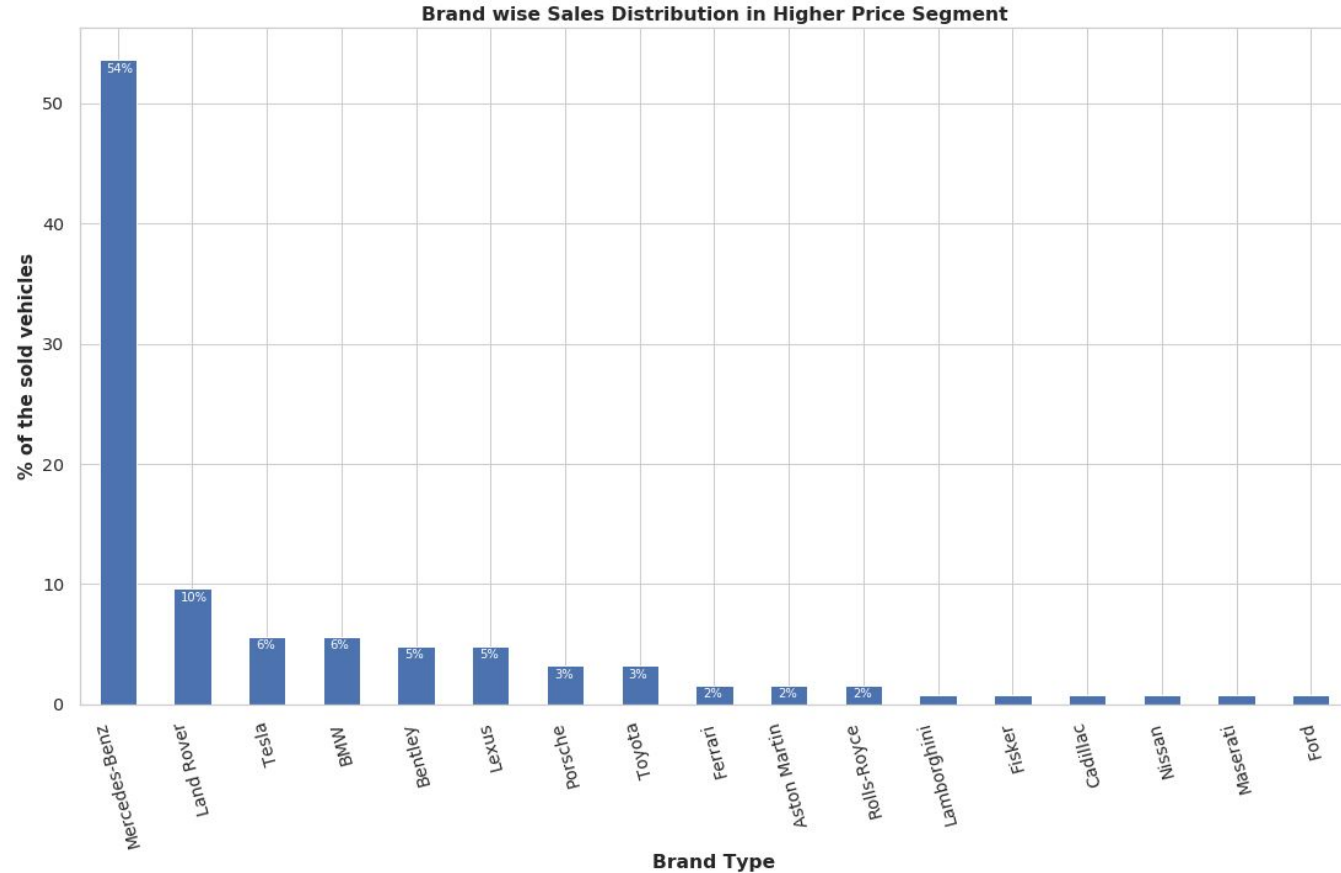


Diesel engines are most preferred in **Volkswagen and Mercedes-Benz** car companies.



AVERAGE PRICE OF VEHICLE BY BRANDS?

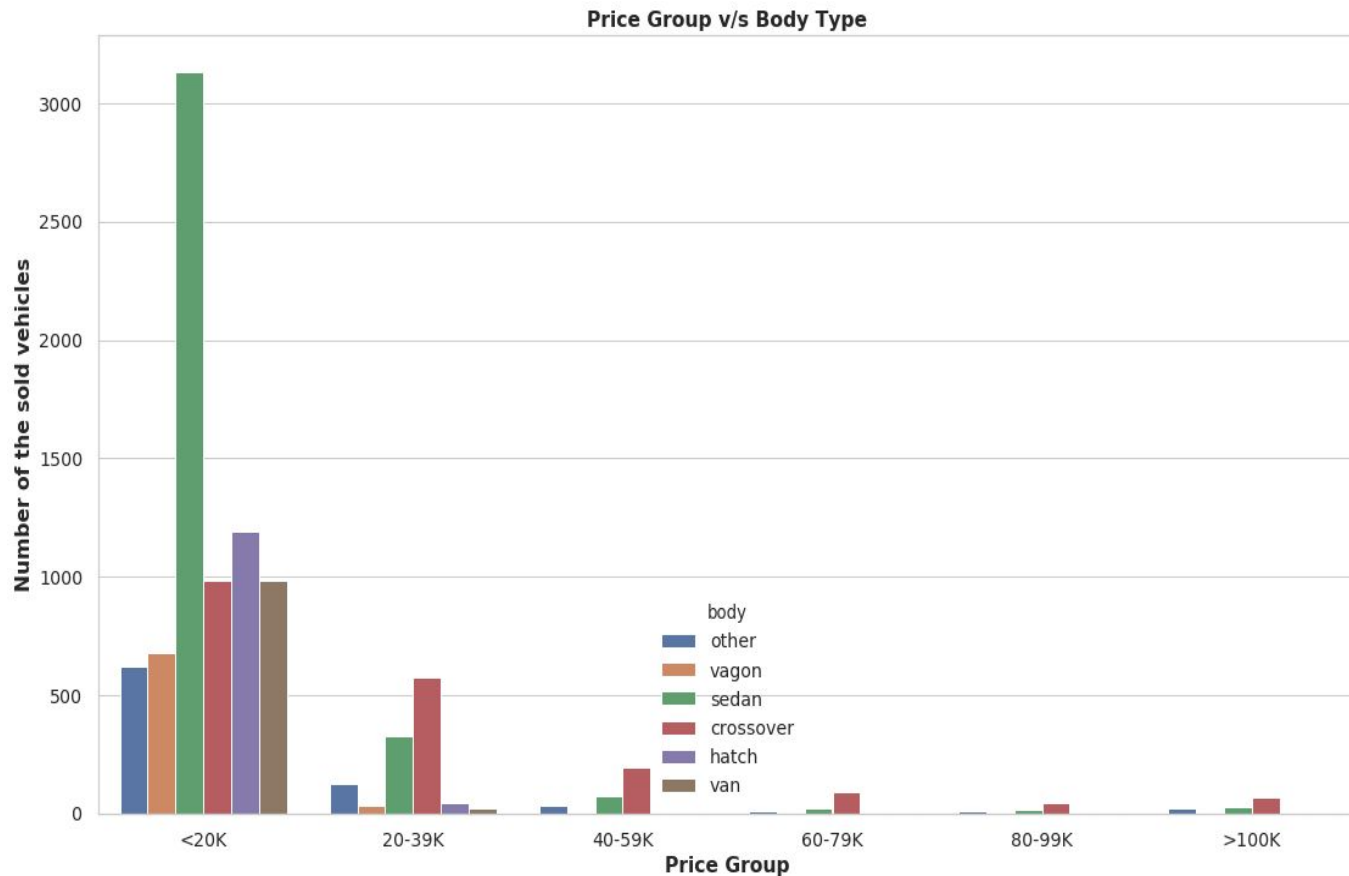
Mercedes-Benz is the leading brand in higher price segment cars.



AVERAGE PRICE OF VEHICLE BY BODY TYPE?

The most cars sold in price range of **less than 20000** and majority of these are in **Sedan** body type.

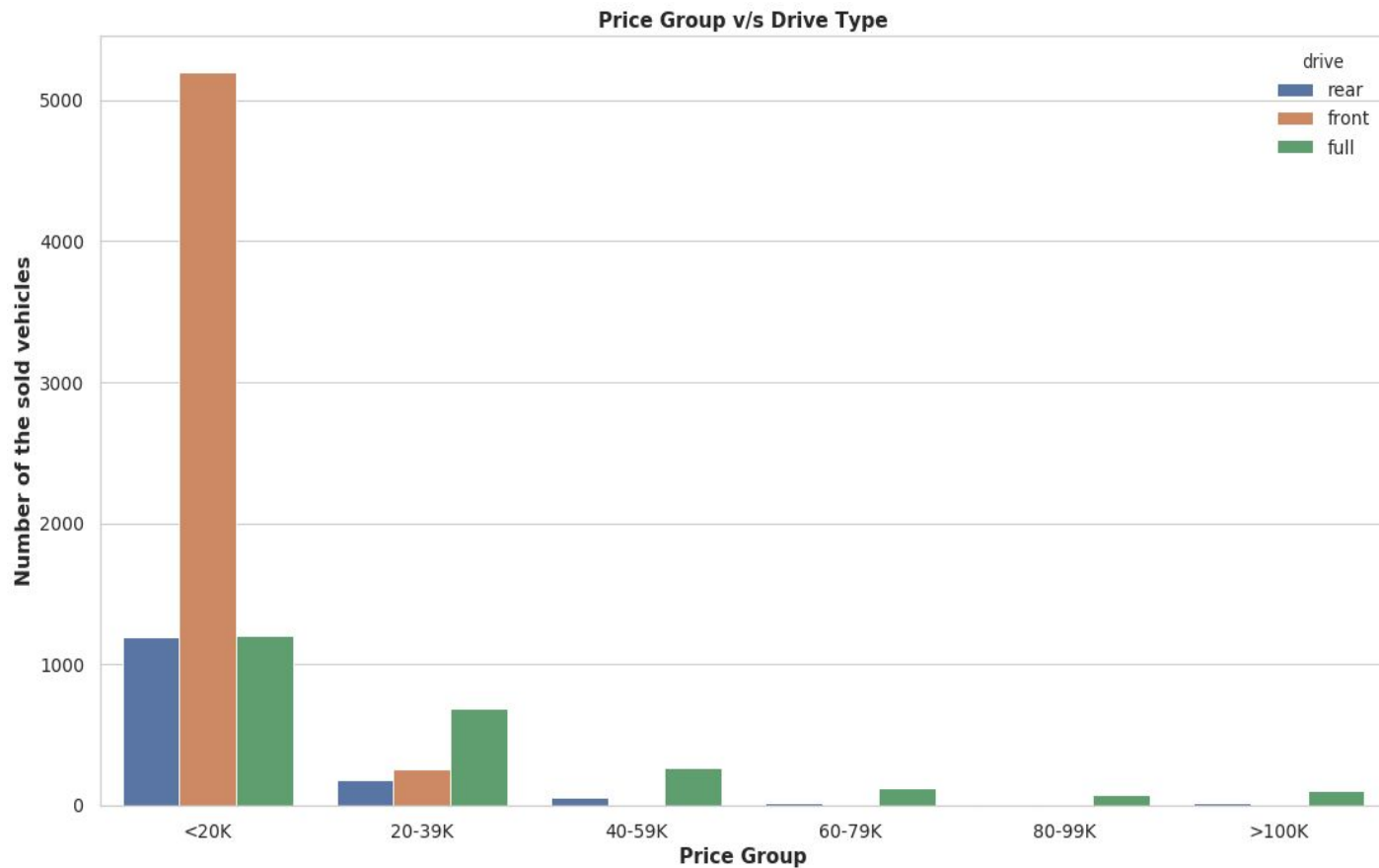
The majority is **Crossover** body type in Cars sold at a price **20000 or above**.



AVERAGE PRICE OF VEHICLE BY DRIVE TYPE?

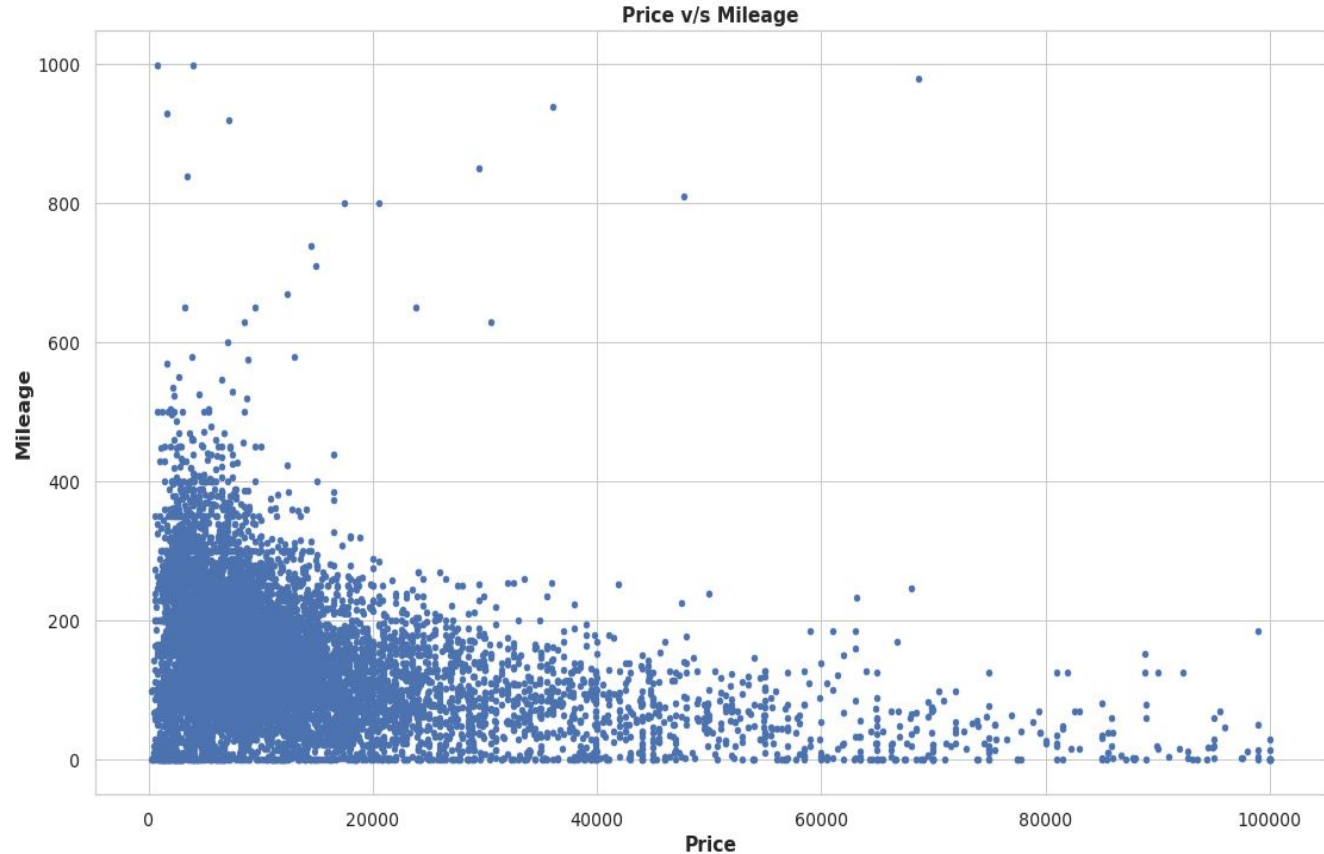
The most cars sold in price range of **less than 20000** and majority of these are in **Front** drive type.

The majority is **Full** drive type in Cars sold at a price **20000 or above**.



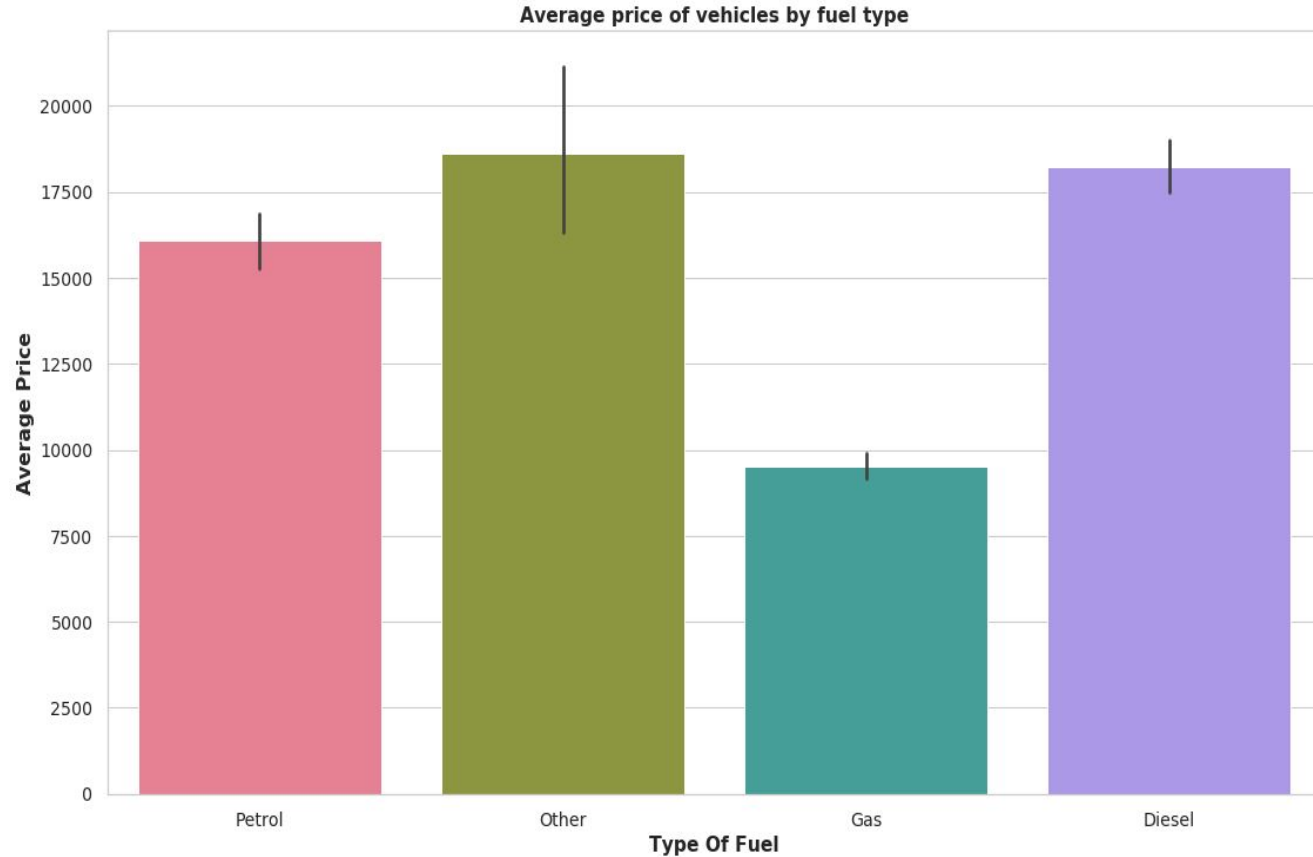
AVERAGE PRICE OF VEHICLE BY mileAge?

Price and **mileage** have **strongly negative correlation** in sold car dataset , i.e. lower price segment has higher mileage and higher price segment has lower mileage.



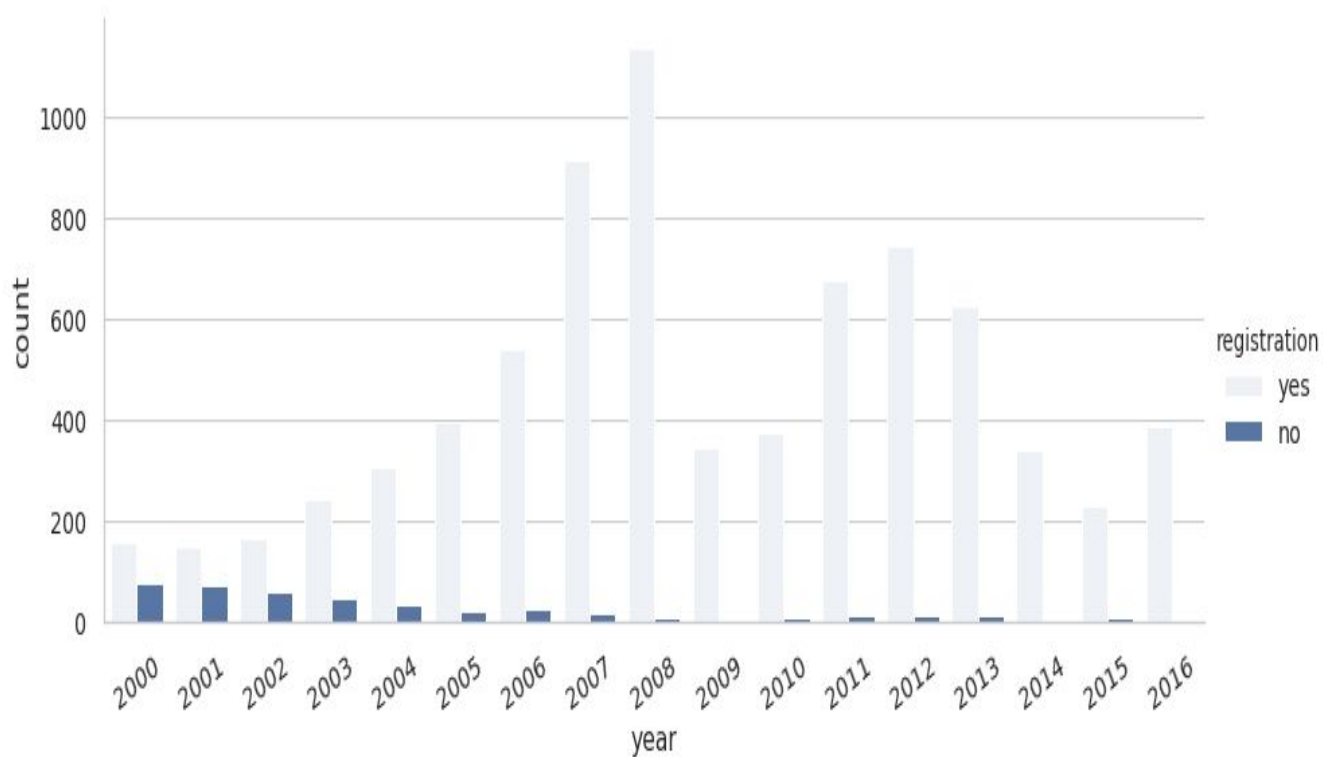
AVERAGE PRICE OF VEHICLE BY FUEL TYPE?

Diesel vehicles are highly priced than **Petrol** and **Gas**.



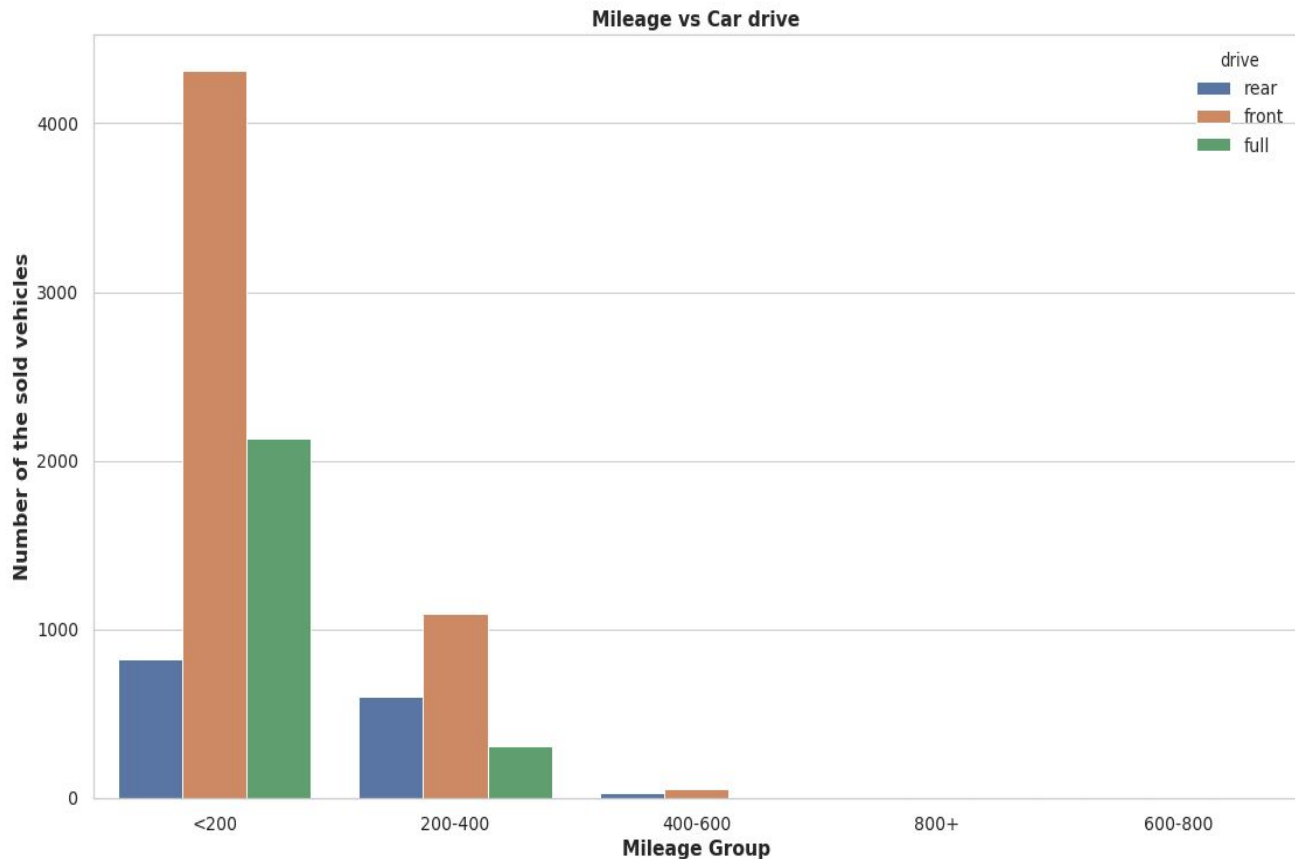
MOST Preferred registration mode over the years of the car?

Majority of the cars sold with **production year 2008 or later** were registered.



MOST Preferred DRIVE TYPE WITH mileAge OF THE car?

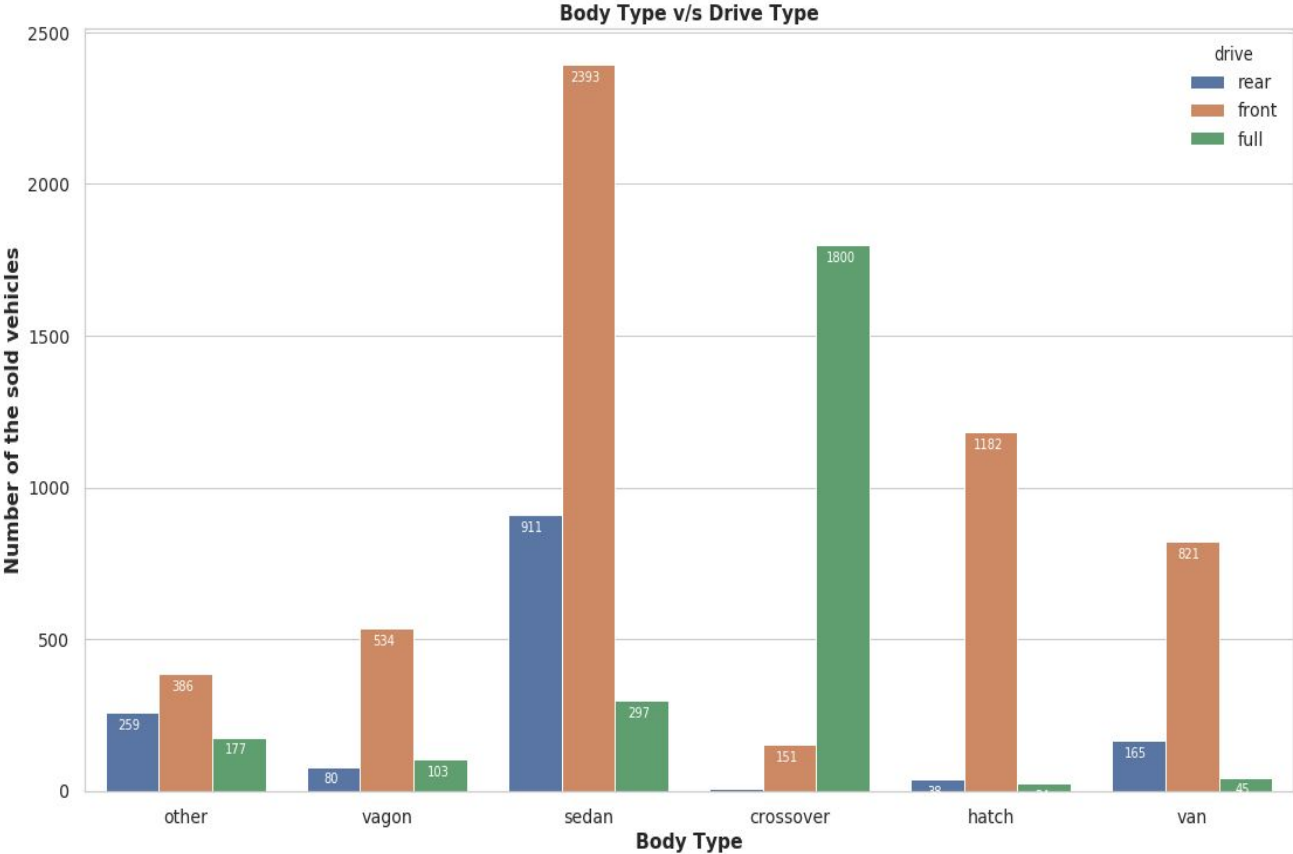
Front drive cars are most preferred because of the high mileage they provide followed by **Full drive** car with very low mileage(<200).



MOST Preferred DRIVE TYPE WITH BODY TYPE OF THE car?

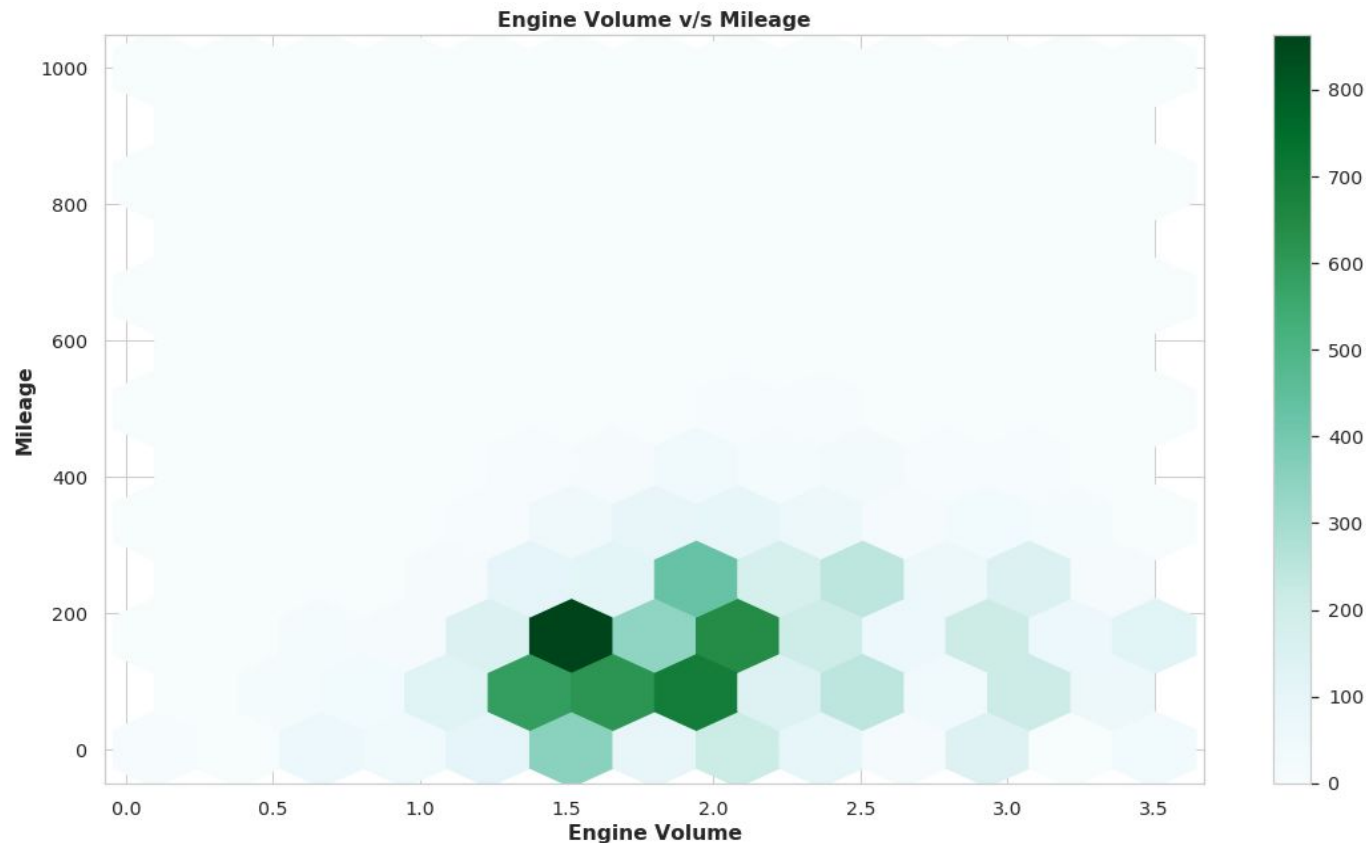
Front drive cars are the top most cars in all body type except in **Crossover** body type.

In **Crossover** body type **Full** drive cars are leading.



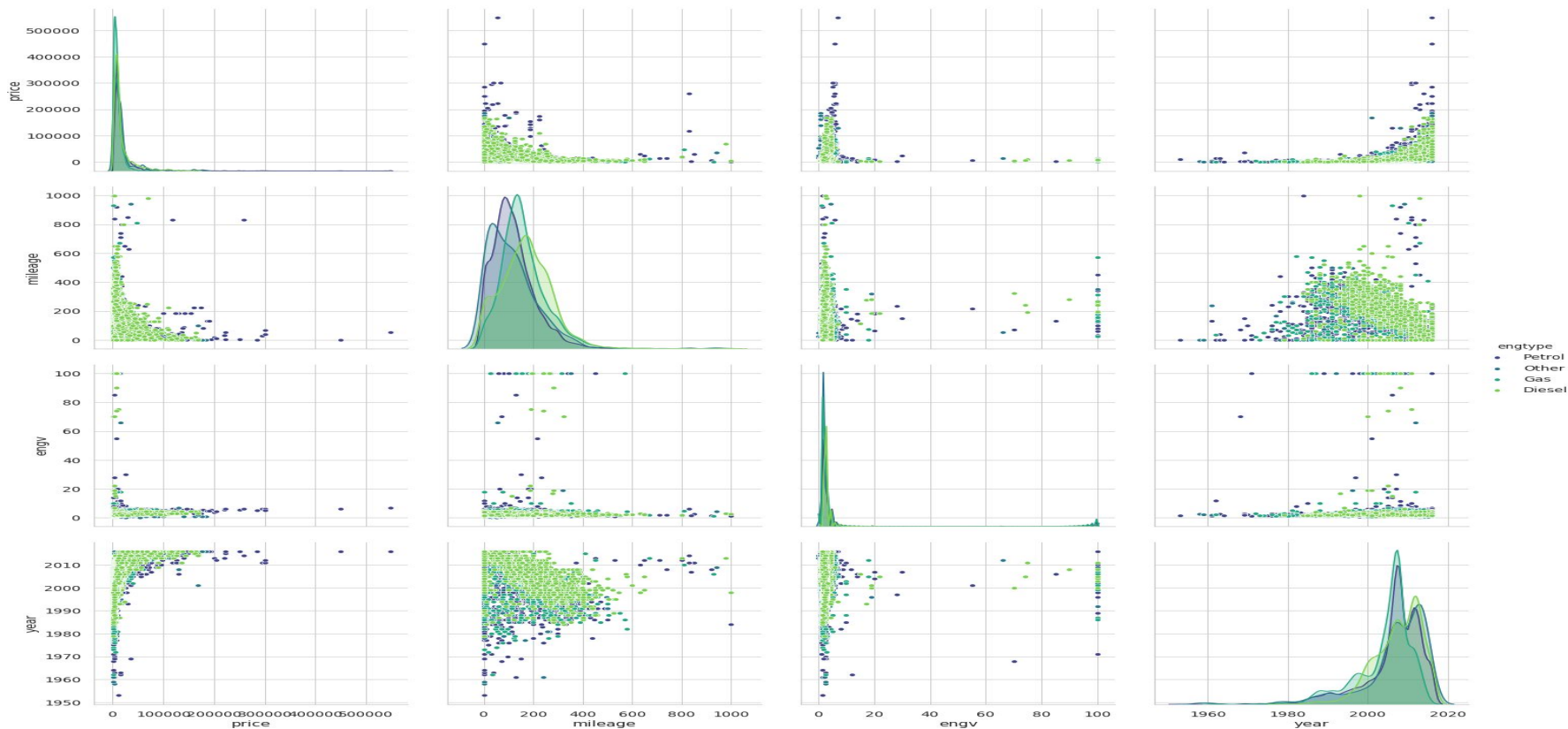
MOST Preferred engine volume WITH mileage OF THE car?

Most of the cars sold with **engine volume (engV)** as 1.5 CC and **mileage** as approx 200.



Multivariate Analysis

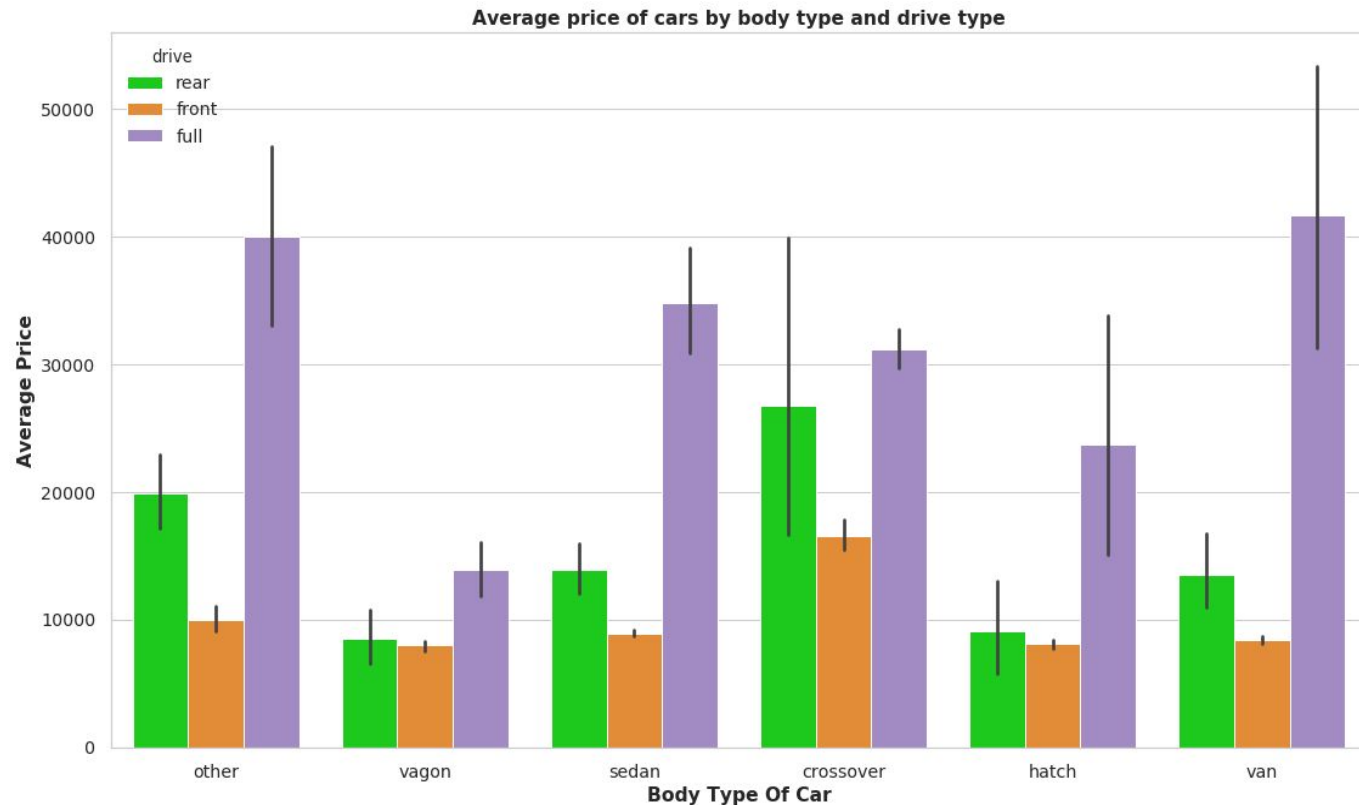
PAIR PLOT - WHAT ALL VARIABLES PAIRED WITH ALL THE OTHER VARIABLES OF THE CAR? CONTINUED...



1. Price varying based on **Year** and **Mileage**
2. As Year increases there is increase in **Petrol** engine type vehicles prices which also depending on **mileage** too.
3. Majority of the cars with **production year 2008 or later** sold in price **50000 or lower**.
4. All the cars with **production year 2010 or later** sold in **price more than 50000**.
5. All the cars with production year before 2010 sold in price 50000 or lower.
6. The most cars sold in price range of **less than 20000** and majority of these are in **Petrol** engine type.
7. The majority is **Diesel** engine type in Cars sold at a price **20000 or above**.
8. Mileage and price are indirectly related. high priced cars give low mileage.

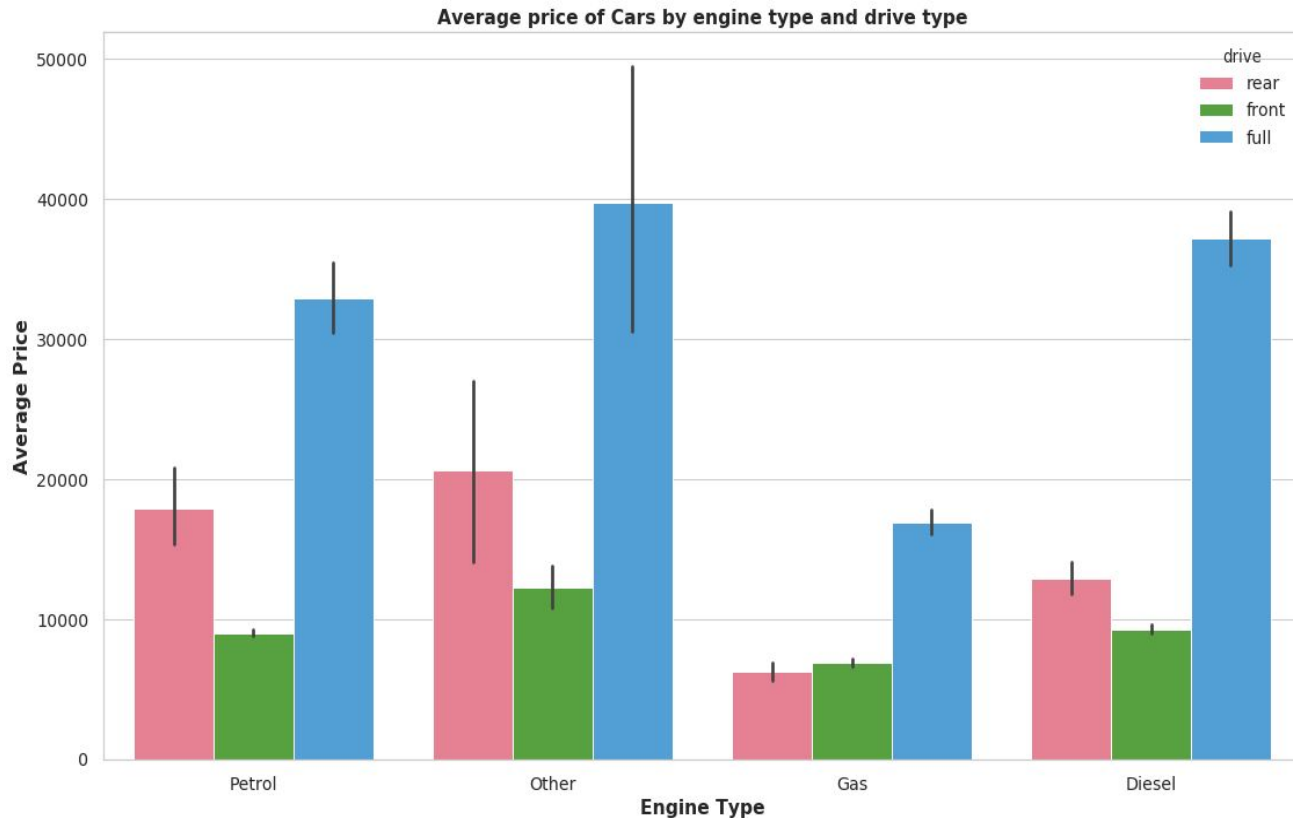
Average Price of Vehicle by Body Type and Drive Type of the Car?

All body types of cars have more average price for full drive type cars compare to other drive types.



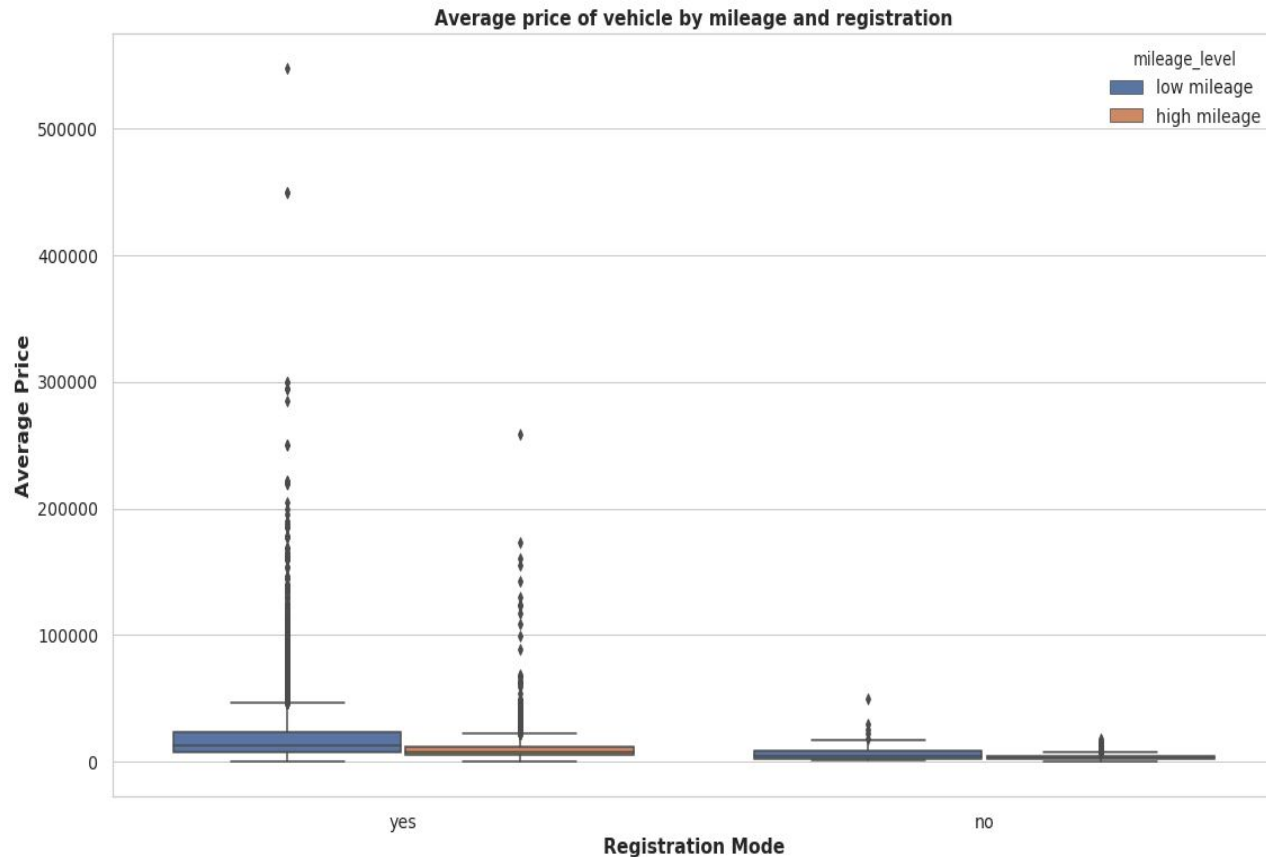
AVERAGE PRICE OF VEHICLE BY engine TYPE and DRIVE TYPE OF THE car?

The cars sold in **all engine type** have **more average price** for **full drive type** cars compare to **other drive type** cars.



Average Price of Vehicle BY mileage and registration OF THE car?

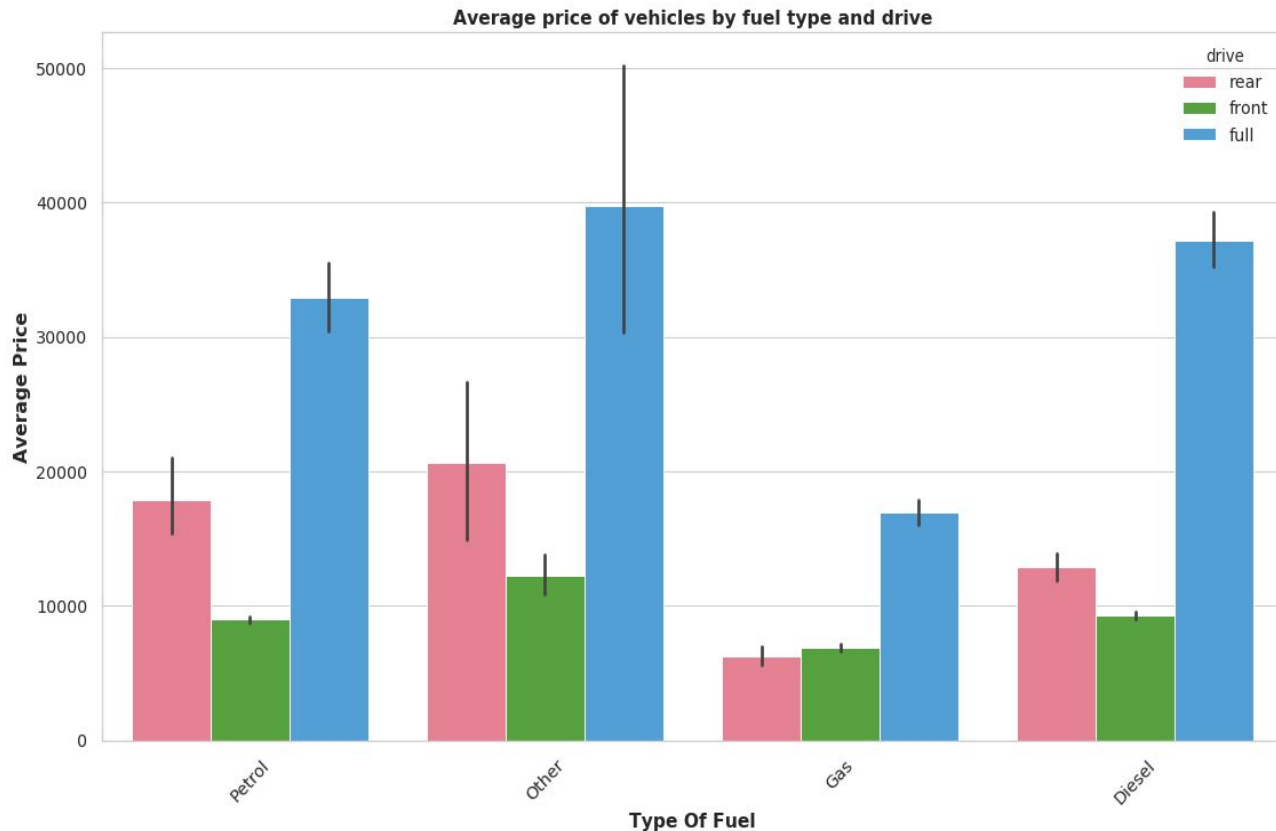
High mileage cars are low priced and vice-versa.



Average Price of Vehicle by Vehicles by Fuel Type and Drive?

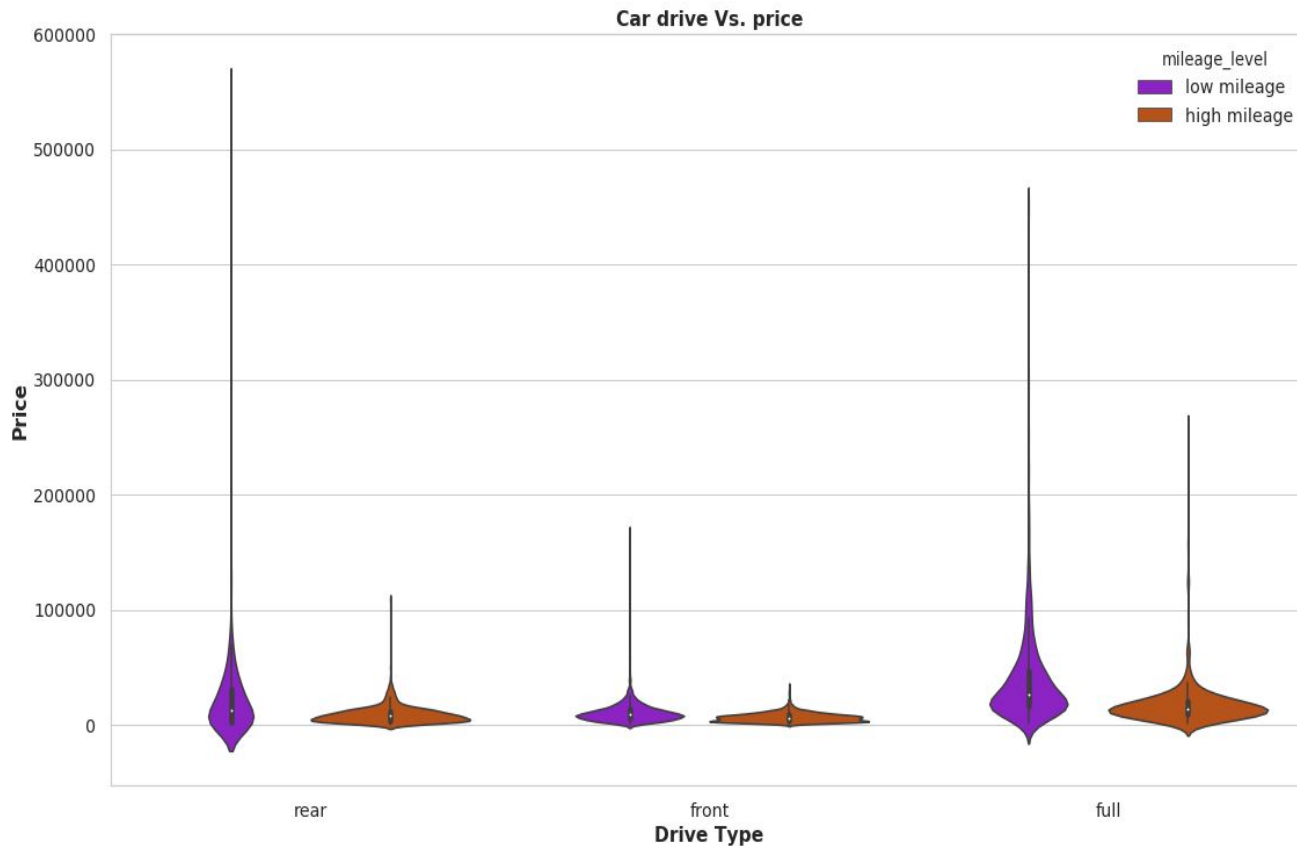
Full drive cars of all engine type cars are high priced compared to other fuel types followed by the **Rear drive** cars in all fuel types.

In Petrol and Diesel Front drive cars are almost equally priced.



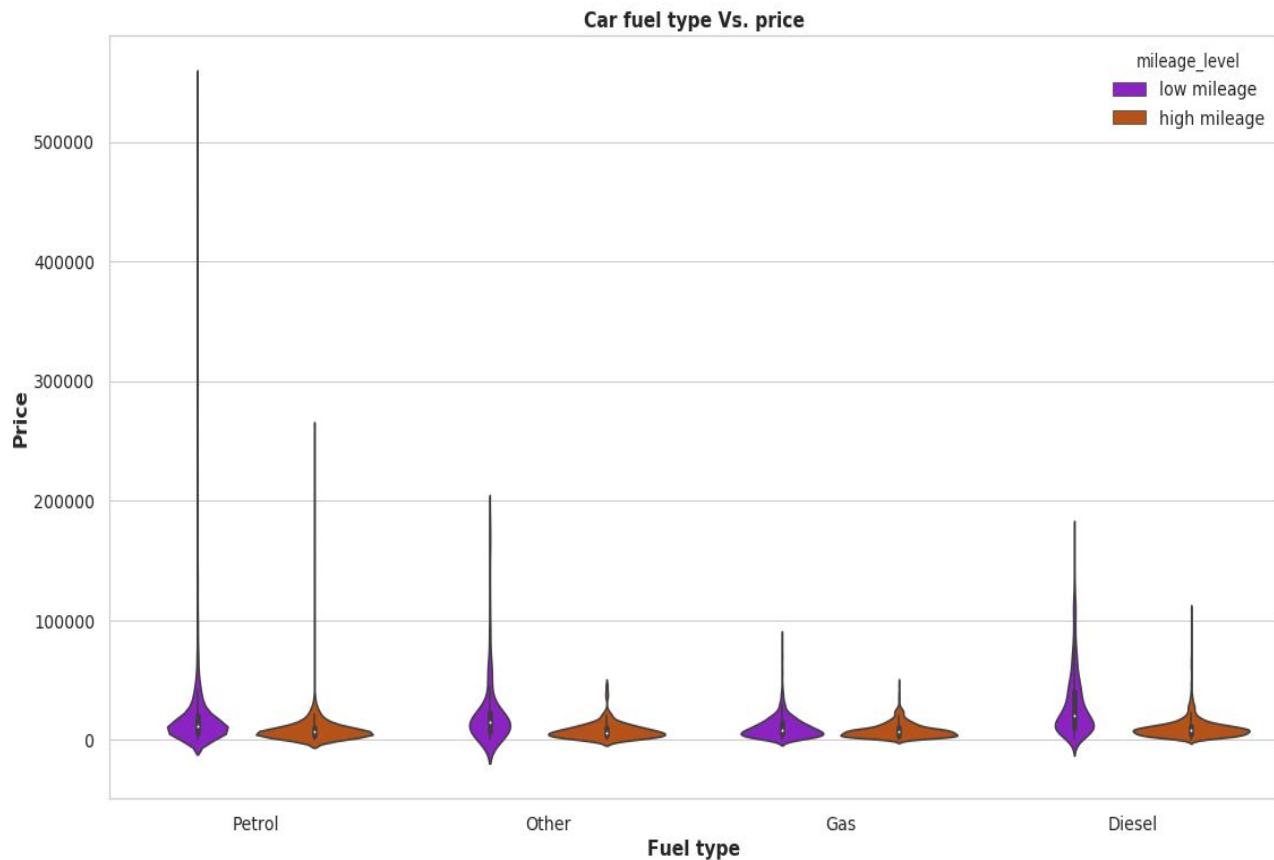
AVERAGE PRICE OF VEHICLE BY mileAge and DRIVE OF THE car?

In all the car drive types, low mileage car are highly priced. where as the high mileage cars are priced low.



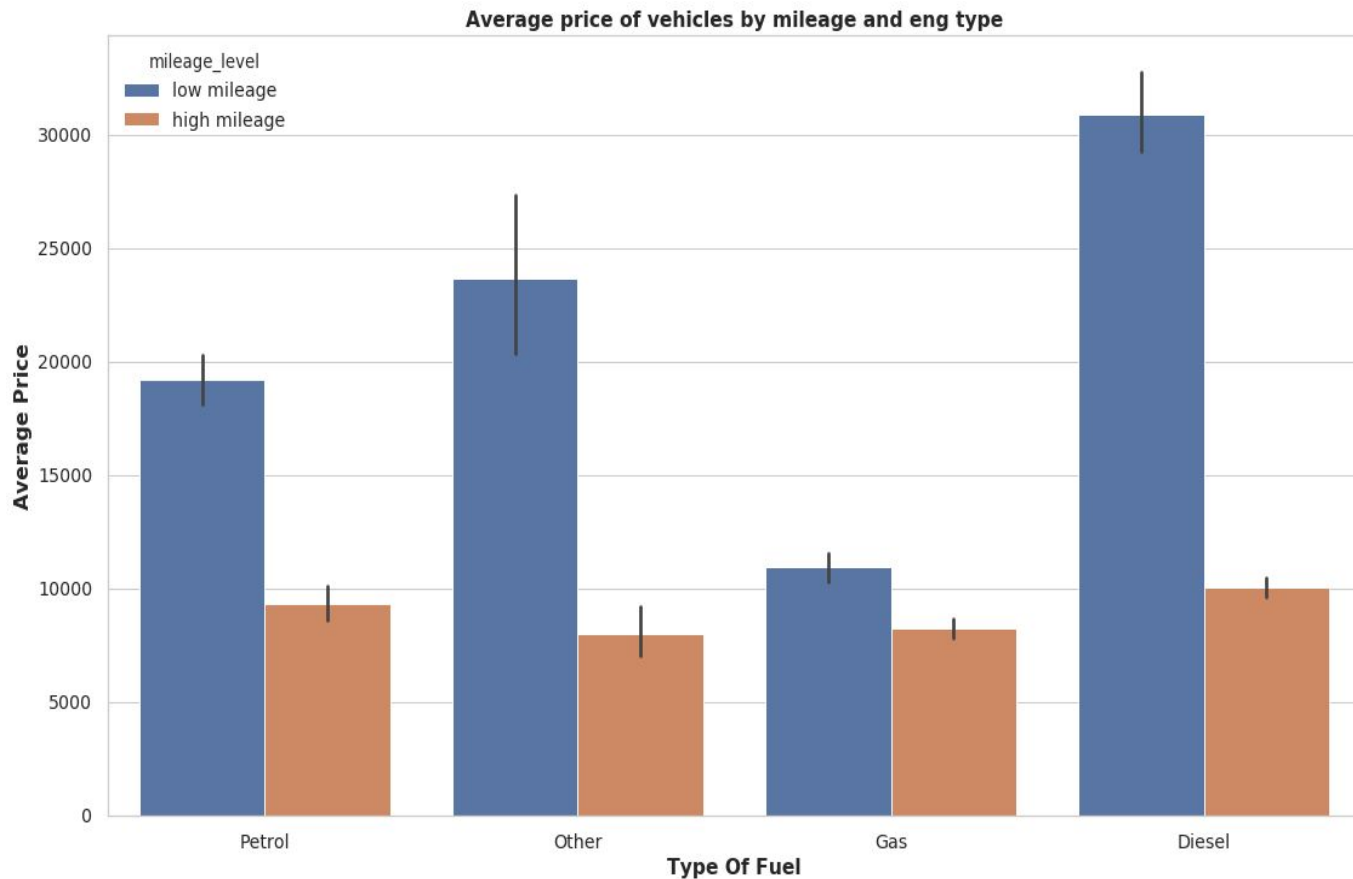
AVERAGE PRICE OF VEHICLE BY mileage and FUEL TYPE OF THE car?

In all the fuel types, low mileage car are highly priced. where as the high mileage cars are priced low.



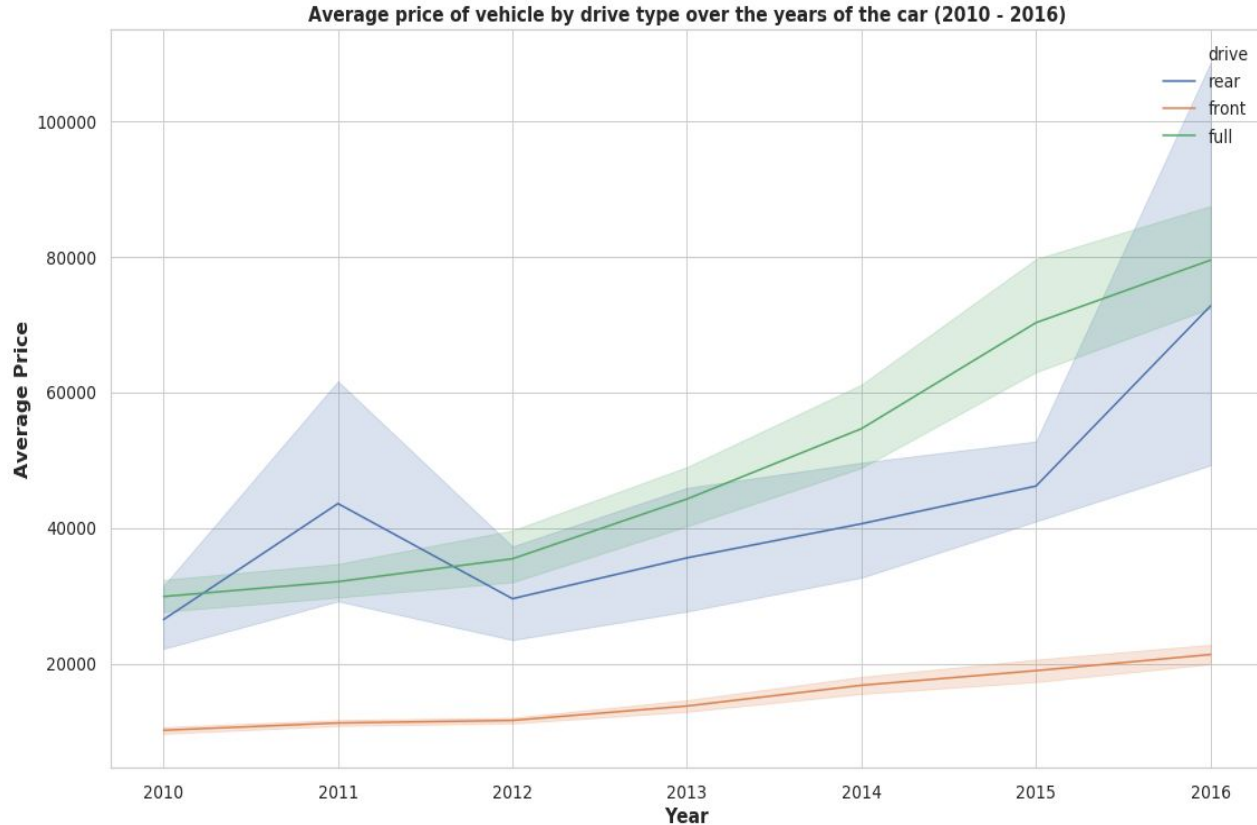
AVERAGE PRICE OF VEHICLES BY mileage and eng TYPE?

Diesel and Gas cars provide highest mileage followed by petrol cars, **Diesel** cars are highly priced with low mileage.



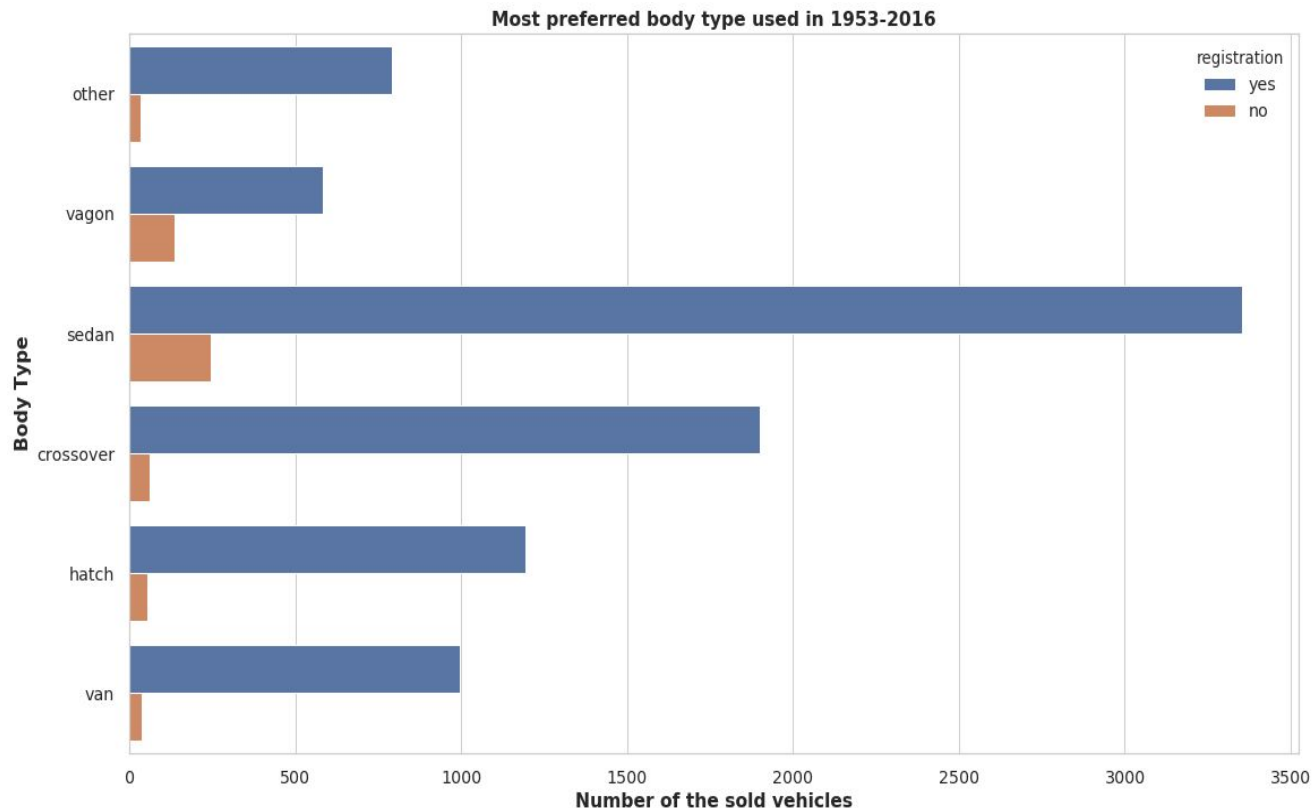
Average Price of Vehicle BY Drive Type Over The Years Of The Car (2010 - 2016)?

As the years increase, we cannot comment on the price increase, but in general, **there has been an increase in price in recent years.**



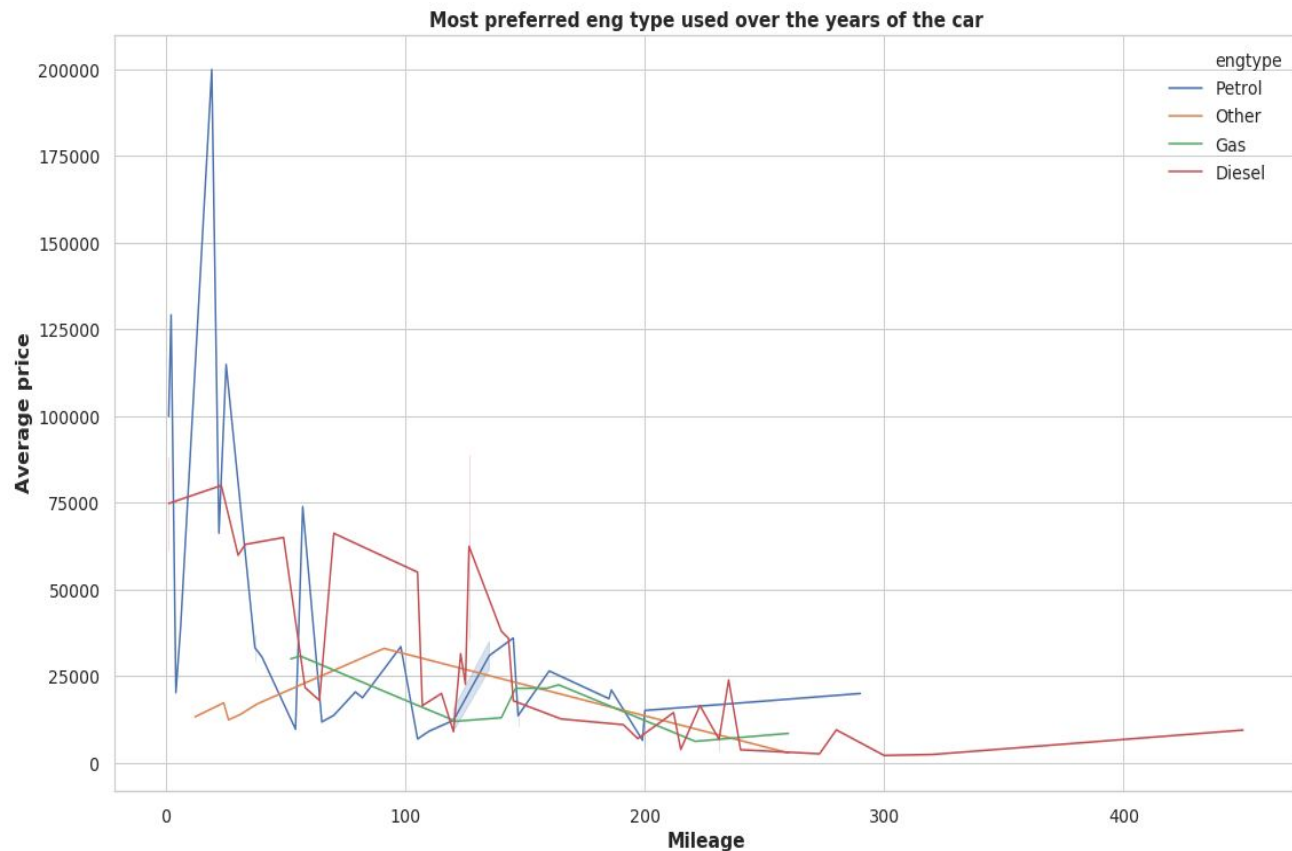
MOST Preferred BODY TYPE USED OVER THE years OF THE car?

The car brands having **Sedan type of body having maximum registration/sale** over the years. This shows People prefers **Sedan** type of body mostly and hence this information can be use for achieving max sale and to figure out production of units.



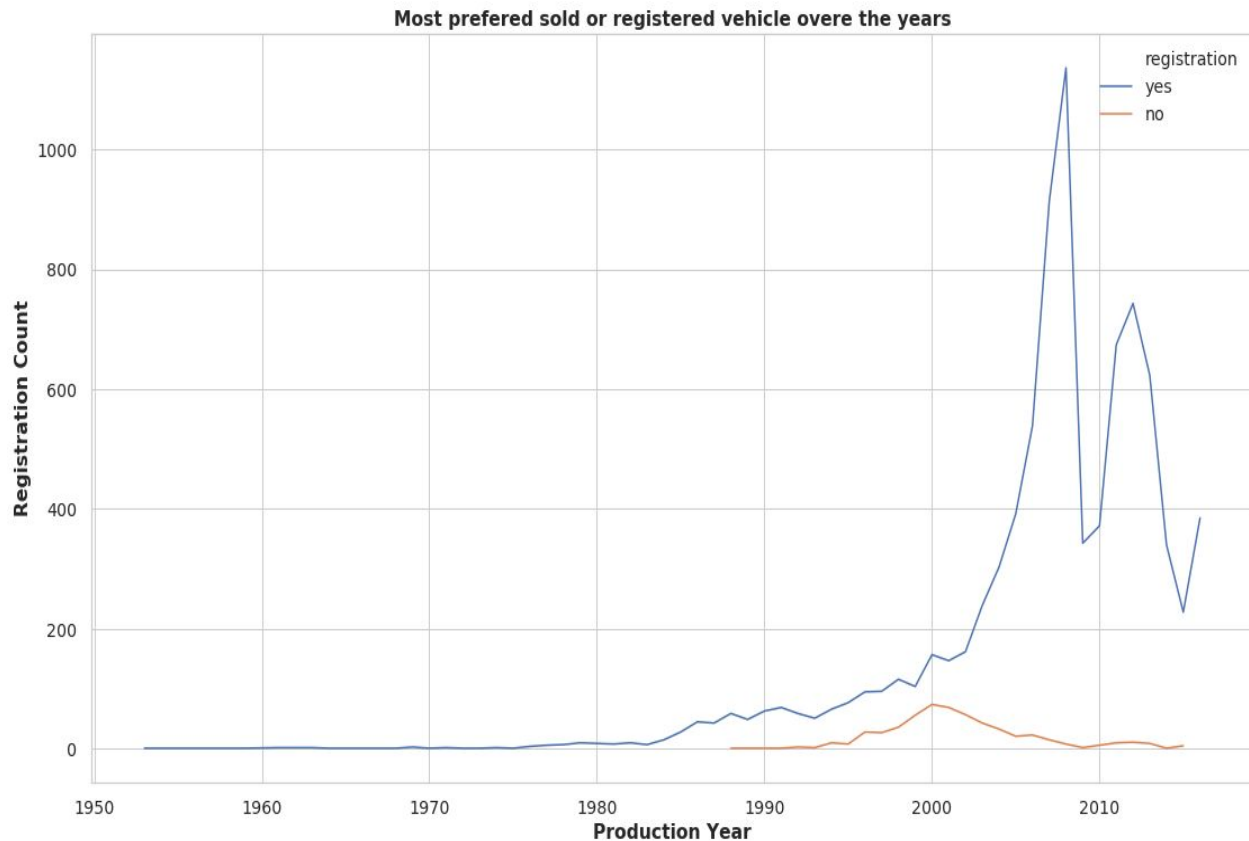
MOST Preferred eng TYPE USED OVER THE years OF THE car?

Price changing accordingly based on mileage value. So **price is varying based on mileage** too and this should be considered a factor for the calculation.



MOST Preferred SOLD or registered VEHICLE over THE years?

Shows max sale was
done in the Year **2008**.



conclusion

- The most cars sold by the company **Volkswagen** followed by **Mercedes-Benz** and **BMW**.
- **Passat & E-Class** is the most preferred model in **Volkswagen** and **Mercedes-Benz** companies respectively.
- **Front and rear** drive is most preferred in **Volkswagen** and **Mercedes-Benz** companies respectively.
- **Diesel engines** are most preferred in **Volkswagen** and **Mercedes-Benz** car companies.
- **Mercedes-Benz** is the leading brand in higher price segment cars.
- Approx **81%** of cars sold in the **price range of less than or equal to \$20000** and majority of these are in **Sedan** body type with **Front drive type**.
- The majority is **Crossover** body type in Cars sold at price between **20000 or above**.
- The majority is **Full drive type** in Cars sold at a price **20000 or above**.
- **Diesel** vehicles are highly priced than **Petrol** and **Gas**.
- Price varying based on **Year** and **Mileage**

Continued.....

- As **Year** increases there is increase in **Petrol** engine type vehicles prices which also depending on **mileage** too.
- Majority of the cars with **production year 2008 or later** sold in **price 50000 or lower**.
- All the cars with **production year 2010 or later** sold in **price more than 50000**.
- All the cars with **production year before 2010** sold in **price 50000 or lower**.
- The most cars sold in price range of **less than 20000** and majority of these are in **Petrol** engine type.
- The majority is **Diesel** engine type in Cars sold at a price **20000 or above**.
- Mileage and price are indirectly related. high priced cars give low mileage.
 - a. **High mileage** cars are **low priced** and vise-versa.
 - b. **Full drive** cars of all engine type cars are high priced compared to other fuel types followed by the **Rear drive** cars in all fuel types.
- In Petrol and Diesel Front drive cars are almost equally priced.
- **Diesel and Gas** cars provide highest mileage followed by petrol cars, **Diesel** cars are highly priced with low mileage.

Continued.....

- Approx 38% of total cars sold in **Sedan** body type, 21% in **Crossover** body type and 13% in **Hatch** body type.
- **Front** drive cars are the top most cars in all body type except in **Crossover** body type.
- In **Crossover** body type **Full** drive cars are leading.
- The **car brands having "sedan" type of body having maximum registration/sale** over the years. This shows People prefers **sedan** type of body mostly and hence this information can be use for achieving max sale and to figure out production of units.
- **Low** mileage vehicles are preferred.
- Majority of the cars sold with **production year 2008 or later were registered**.
- **Front drive** cars are most preferred because of the high mileage they provide followed by **Full drive** car with very low mileage(<200).
- Most of the cars sold with **engine volume (engV)** as 1.5 CC and **mileage** as approx 200.

Continued.....

- Out of total cars sold, approx **91%** cars fall in range of engine volume more than **1.5 CC to 3.5 CC**.
- Most of the cars sold have **2.0 CC** engine volume followed by **1.5 CC**, there are very **few cars sold** with engine volume **less than equal to 1.0 CC and greater than equal to 4.0 CC**.
- **Petrol & Diesel cars** are the most widely sold Engine Type. Gas and Other fuels are not so commonly used.
- **Registered** vehicles are preferred.
- **Production year 2008 and 2012** cars are sold the most. Year **2008** has the highest number of car sales record.

ACTIONABLE INSIGHTS

- a. Low mileage cars with high priced can be recommended for highly profiled customer.
- b. **Crossover** body type is preferred in **Full drive type** thus the same experiment can be applied with **Sedan**.