CARS DATASET

Cars Dataset is taken from Kaggle and the dataset has lots of Null values and outliers because the data was Web Scraped for more than a year in Czech Republic and Germany from different web pages.

Also, the dataset should be cleaned before performing any Analysis or Algorithms. The Dataset has many columns, describing the features of the car. The dataset can mainly be used in Predictive Analytics.

The Dataset has around 10 lakh rows and 16 columns namely, maker, model, mileage - in KM, manufacture year, engine displacement - CCM, engine power - KW, body type, color slug, door count, stk control, transmission, seat count, fuel type, data created, date last seen and price - in EUR.

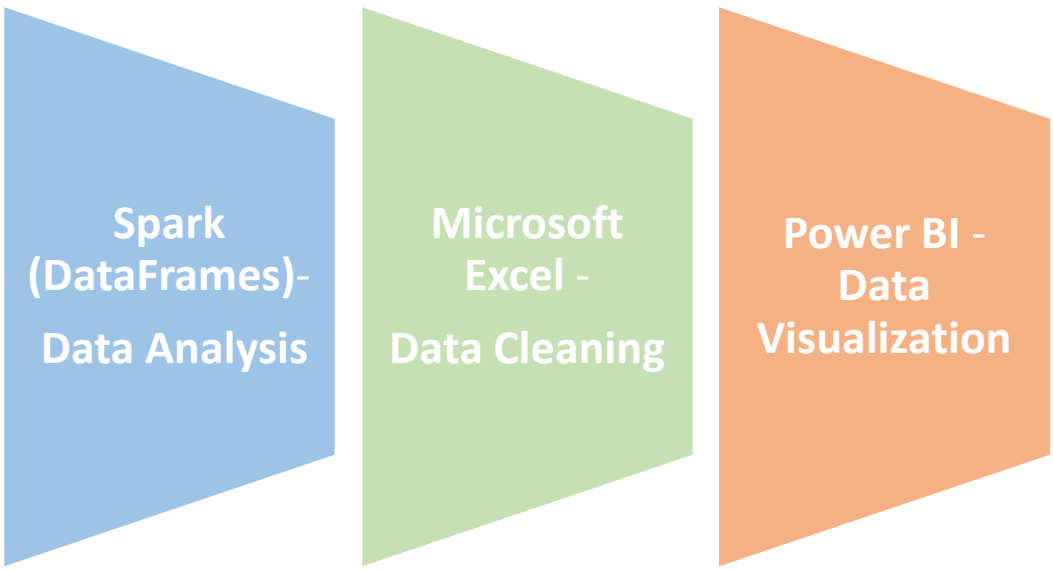
OBJECTIVE:

- The Main Objective of the Analysis is to help (Used Car Retail Company) to predict the Top 5 car models for Personal use and Top 5 car models for Commercial use (Taxi/Cab) for the customers using Spark.

- The other objectives are to see the relationship between maker, model and price to find the cheapest car makers and car models.

- Final objective of the analysis is to compare the price difference between gasoline and diesel engine cars.

TOOLS AND TECHNOLOGIES USED:



DATA CLEANING

The Cars Dataset is cleaned using Microsoft Excel. The Dataset had more than 10 lakh rows and now its reduced to around 2.5 lakhs. Also, the Dataset had 16 columns and now its reduced to 11 columns. The Dataset had a plethora of Outliers and NULL Values and so the outliers are replaced, and NULL values are removed in order to get clean and clear Data.

Moreover, columns like, body_type, color_slug, stk_year, door_count, date_created, datelastseen, transmission and seat_count are removed and new columns like, pollution_test and car_type are added. In maker and model column, the blank values are removed. Later, in mileage column, values less than 15000 KM are removed because it's impossible to recommend cars with mileage less than that to the customers. Later, in manufacture_year column, year less than 2001 are removed because it's impossible to recommend cars older than 15 years to the customers.

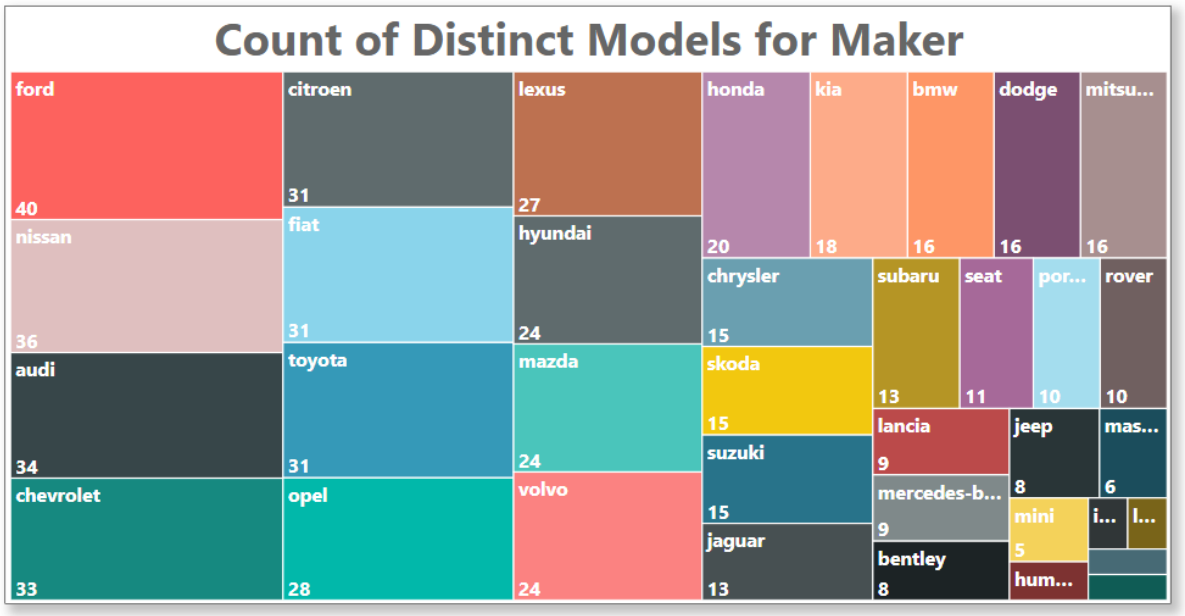
After that, in engine_displacement column, removed values less than 1500 CCM and greater than 8000 CCM. Later, in the engine_power column, removed all values less than 40 KW because the car needs at least the basic engine power. After that, derived a new column called pollution_test with the reference of stk_control column. Later, derived another new column called car_type with the reference of seat_count column. Finally, the price column had lots of outliers and missing values and so replaced it by taking average price of same model cars.

First 20 Rows and All Columns of Cleaned Dataset:

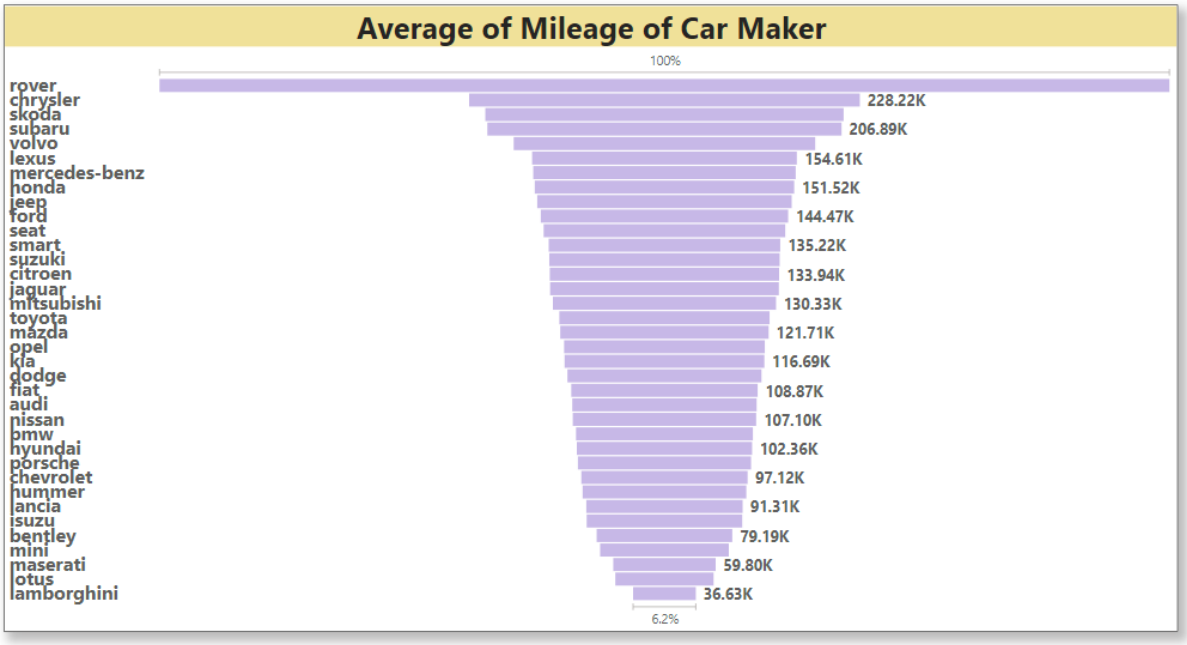
	maker	model	mileage	manufacture_year	engine_displacement	engine_power	pollution_test	auto_facility	car_type	fuel_type	price
2	ford	galaxy	151000	2011	2000	103	FALSE	FALSE	SUV	diesel	10584.75
3	skoda	octavia	143476	2012	2000	81	FALSE	FALSE	COMPACT	diesel	13161.05
4	skoda	fabia	111970	2004	1600	47	FALSE	FALSE	COMPACT	gasoline	14339.33
5	skoda	fabia	128886	2004	1800	47	FALSE	FALSE	COMPACT	gasoline	14339.33
6	skoda	fabia	140932	2003	1700	40	FALSE	FALSE	COMPACT	gasoline	14339.33
7	skoda	fabia	167220	2001	1600	74	FALSE	FALSE	COMPACT	gasoline	12833.54
8	skoda	octavia	105389	2003	1900	81	FALSE	FALSE	COMPACT	diesel	13161.05
9	skoda	favorit	41250	2006	1800	44	FALSE	TRUE	COMPACT	gasoline	14339.33
10	suzuki	swift	122100	2003	1800	99	FALSE	FALSE	COMPACT	gasoline	14339.33
11	nissan	x-trail	149465	2005	2500	121	FALSE	FALSE	COMPACT	gasoline	18717.21
12	opel	astra	316054	2005	1700	74	FALSE	FALSE	COMPACT	diesel	12833.54
13	skoda	superb	269398	2005	1900	96	FALSE	FALSE	COMPACT	diesel	14267.76
14	skoda	fabia	87257	2008	1900	44	FALSE	FALSE	COMPACT	gasoline	14339.33
15	skoda	fabia	130340	2001	1900	50	FALSE	FALSE	COMPACT	gasoline	14339.33
16	ford	focus	227415	2002	1800	85	FALSE	FALSE	COMPACT	diesel	13161.05
17	ford	fiesta	84476	2005	1700	44	FALSE	TRUE	COMPACT	gasoline	14339.33
18	citroen	c4-picasso	112313	2007	1700	92	FALSE	FALSE	COMPACT	gasoline	14267.76
19	seat	ibiza	86484	2007	1600	51	FALSE	FALSE	COMPACT	gasoline	14339.33
20	audi	a6	207427	2007	2700	132	FALSE	FALSE	COMPACT	diesel	18717.21

DATA ANALYSIS

Analysis using Power BI:



- The above analysis shows the count of distinct car models for different car makers.
- Car makers like, Ford, Nissan, Audi, Chevrolet and Citroen are the top 5 cars which released many models.



- The above analysis shows the Average mileage of distinct car maker.
- Car makers like Rover, Chrysler, Skoda, Volvo and Lexus are the top 5 makers with best average mileage.

Analysis using Spark (Data Frames):

1.How car maker, car model and car price are related to each other? And find the top 5 cheapest car makers and car models.

QUERY - 1:

```
val cars_1 = cars_cols.groupBy(col("maker")).agg(avg(col("price"))
.alias("average_price"), countDistinct(col("model"))
.alias("model_count")).orderBy(col("average_price")
.asc).limit(5).show()
```

```
scala> val cars_1 =cars_cols.groupBy(col("maker")).agg(avg(col("price")).alias("average_price"), countDistinct(col("
model")).alias("model_count")).orderBy(col("average_price").asc).limit(5).show()
+------+-----------------+-----------+
| maker|    average_price|model_count|
+------+-----------------+-----------+
|suzuki| 13393.89837398374|         15|
| smart|13761.886287625419|          2|
|  fiat|14171.987372598538|         31|
| skoda|14305.561757208594|         15|
|  opel|14487.711921656935|         28|
+------+-----------------+-----------+
```

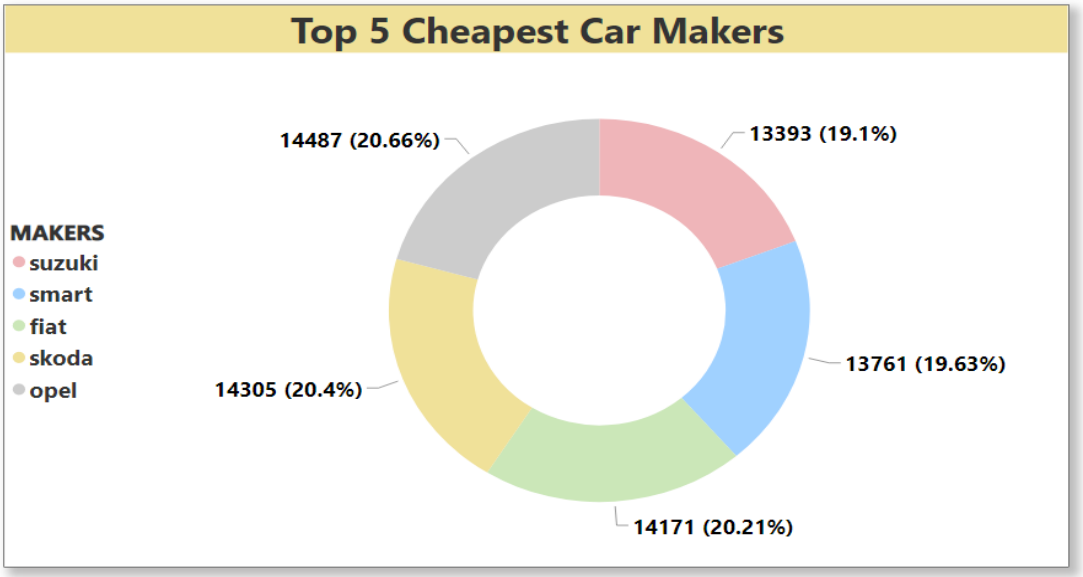
QUERY - 2:

```
val cars_2 = cars_cols.groupBy(col("model"),col("maker"))
.agg(avg(col("price")).alias("average_price"))
.orderBy(col("average_price").asc).limit(5).show()
```

```
scala> val cars_2 =cars_cols.groupBy(col("model"),col("maker")).agg(avg(col("price")).alias("average_price")).orderB
y(col("average_price").asc).limit(5).show()
+-------+------+-----------------+
|  model| maker|    average_price|
+-------+------+-----------------+
| 280-zx|nissan|          11201.0|
|   adam|  opel|11665.329470198676|
|insight| honda|11885.851851851852|
|sequoia|toyota|          12001.0|
|   pony|hyundai|          12057.0|
+-------+------+-----------------+
```

Explanation for the Question - 1:

- Both queries of the above question deal with the relationship between maker, model and price.
- The first query gives the top 5 cheapest makers with its average price and number of models released by the maker.
- The second query gives the top 5 cheapest models with its average price and maker of the model.



2.Find the top 5 car makers and give reasons to prove it.

QUERY:

```
val cars_3 = cars_cols.where("mileage >= 200000")
.where("manufacture_year >= 2005")
.where("engine_displacement >= 1500")
.where("engine_power >= 100")
.where("pollution_test = true")
.where("fuel_type == \"gasoline\" ")
.orderBy(col("price").asc).select(col("maker"))
.distinct().show(5)
```

```
scala> :paste
// Entering paste mode (ctrl-D to finish)

val cars_3 = cars_cols.where("mileage >= 200000").where("manufacture_year >= 2005").where("engine_displacement >= 15
00")
.where("engine_power >= 100").where("pollution_test = true").where("fuel_type == \"gasoline\" ")
.orderBy(col("price").asc).select(col("maker")).distinct().show(5)

// Exiting paste mode, now interpreting.

+----------+
|     maker|
+----------+
|mitsubishi|
|     lexus|
|    toyota|
|      seat|
|  chrysler|
+----------+
only showing top 5 rows
```

Explanation for the Question - 2:

- The above query gives the top 5 manufactures of car (Mitsubishi, Lexus, Toyota, Seat, Chrysler).
- It is based on the mileage greater than or equal to 2 lakh kms, manufacture year greater than or equal to 2005 because no customer will wish in buying a 10 years old car.
- Engine displacement greater than or equal to 1500 CCM as it should have basic engine displacement.
- Engine power greater than or equal to 100 kw as it is common to have good engine power, pollution test should be passed, fuel type should be petrol as it is Environment friendly.
- Price should be the minimal as it is welcomed by all customers who wishes to buy used cars.

3.Compare the gasoline and diesel engine car's price and give comments.

QUERY - 1:

```
val cars_4 = cars_cols
.filter("fuel_type == \"gasoline\"")
.select(avg(col("price"))).show()
```

```
scala> val cars_4 = cars_cols.filter("fuel_type == \"gasoline\"").select(avg(col("price"))).show()
+-----------------+
|       avg(price)|
+-----------------+
|16773.572693007158|
+-----------------+
```

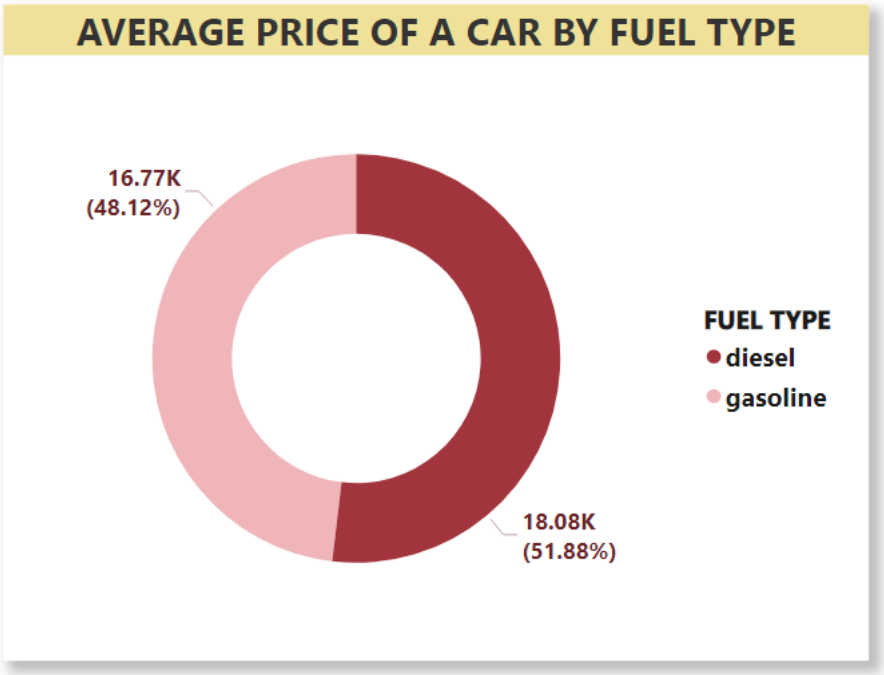
QUERY - 2:

```
val cars_5 = cars_cols
.filter("fuel_type == \"diesel\"")
.select(avg(col("price"))).show()
```

```
scala> val cars_5 = cars_cols.filter("fuel_type == \"diesel\"").select(avg(col("price"))).show()
+----------------+
|      avg(price)|
+----------------+
|18083.76350118398|
+----------------+
```

Explanation for the Question - 3:

- Both the queries of the above question deal with the average price of all makers.
- The first query gives the average price of makers with petrol engine and second query gives the average price of makers with diesel engine.
- In comparison between gasoline and diesel engine, the average car price of diesel engine (18083 EUR) is greater than the average car price of petrol engine (16773 EUR).



4.What are the top five car models for personal use? and Why?

QUERY:

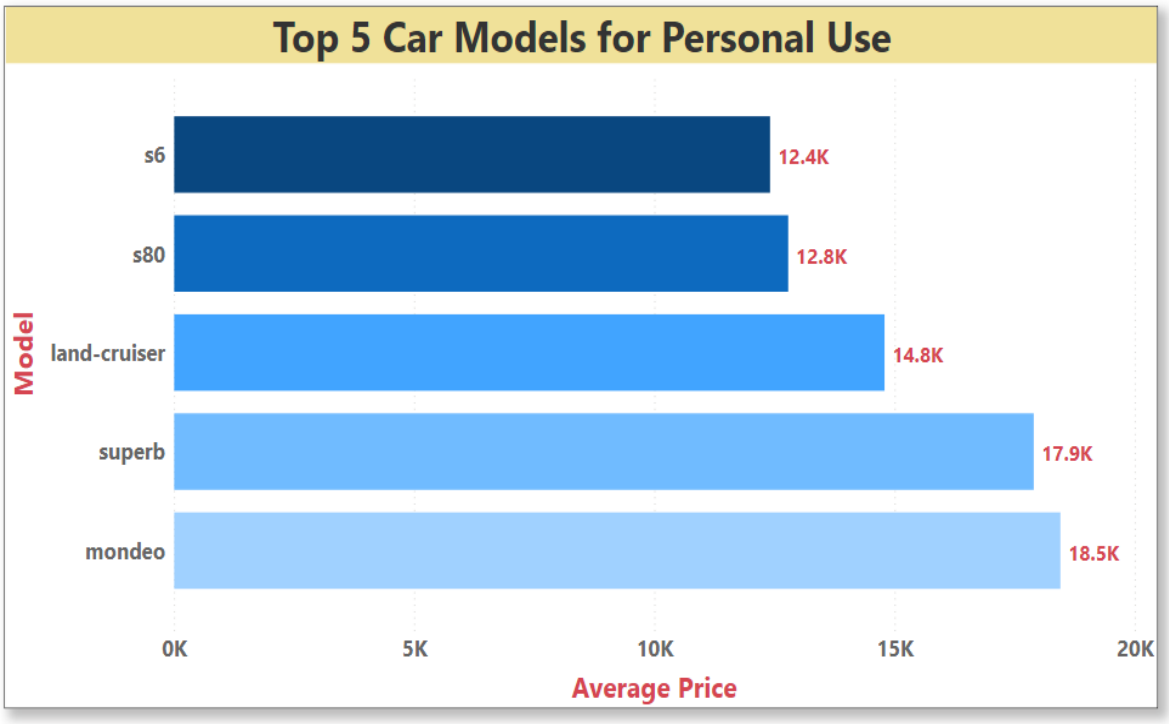
```
val cars_6 = cars_cols.where("mileage >= 200000")
.where("manufacture_year >= 2005")
.where("engine_displacement >= 1500")
.where("engine_power >= 100")
.where("pollution_test == true").where("auto_facility == true").where("fuel_type == \"gasoline\" ")
.where("car_type == \"COMPACT\" ")
.orderBy(col("price").asc)
.select(col("model")).distinct().show(5)
```

```
scala> :paste
// Entering paste mode (ctrl-D to finish)

val cars_6 = cars_cols.where("mileage >= 200000").where("manufacture_year >= 2005").where("engine_displacement >= 1500")
.where("engine_power >= 100").where("pollution_test = true").where("auto_facility = true").where("fuel_type == \"gasoline\" ").where("car_type == \"COMPACT\" ")
.orderBy(col("price").asc).select(col("model")).distinct().show(5)

// Exiting paste mode, now interpreting.

+------------+
|       model|
+------------+
|          s6|
|         s80|
|land-cruiser|
|      superb|
|      mondeo|
+------------+
only showing top 5 rows
```

Explanation for the Question - 4:

- The above query gives the top 5 car models for personal use (s6, s80, land-cruiser, superb, Mondeo).
- It is based on mileage greater than or equal to 2lakh kms, manufacture year greater than or equal to 2005 because not to recommend cars older than 10 years for the customers.
- Engine displacement greater than or equal to 1500 CCM as it should have basic requirements.

- Engine power greater than or equal to 100 kw as it normal for customers to ask for high power, should have Automatic gear facility, Pollution test should be passed, should be petrol engine as it is used in controlling pollution, car type should be compact as its for personal or family use.
- Finally, price should be low as its should not be the concern for the customers.



5.What are the top five car models for Commercial use (taxi/cab)? and Why?

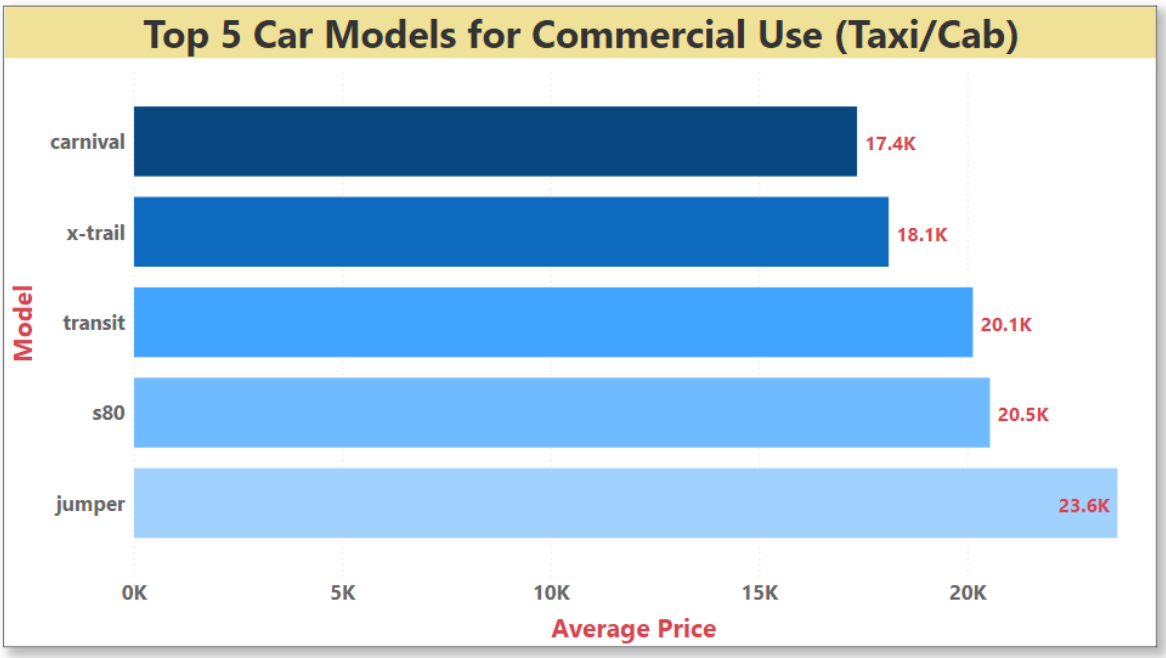
QUERY:

```
val cars_7 = cars_cols.where("mileage >= 200000")
.where("manufacture_year >= 2005")
.where("engine_displacement >= 1500")
.where("engine_power >= 100").where("fuel_type == \"diesel\" ")
.where("car_type == \"SUV\" ")
.orderBy(col("price").asc)
.select(col("model")).distinct().show(5)
```

```
scala> :paste
// Entering paste mode (ctrl-D to finish)

val cars_7 = cars_cols.where("mileage >= 200000").where("manufacture_year >= 2005").where("engine_displacement >= 15
00")
.where("engine_power >= 100").where("fuel_type == \"diesel\" ").where("car_type == \"SUV\" ")
.orderBy(col("price").asc).select(col("model")).distinct().show(5)

// Exiting paste mode, now interpreting.

+--------+
|   model|
+--------+
|carnival|
| x-trail|
| transit|
|     s80|
|  jumper|
+--------+
only showing top 5 rows
```

Explanation for the Question - 5:

- The above query gives the top 5 car models for personal use (Carnival, X-trail, transit, s80, jumper) based on mileage greater than or equal to 2lakh kms, manufacture year greater than or equal to 2005 because not to recommend cars older than 10 years for the customers.
- Engine displacement greater than or equal to 1500 CCM as it should have basic requirements.
- Engine power greater than or equal to 100 kw as it should have basic requirements, should be diesel engine as mileage is significant for the customers running taxi/cab.
- Car type should be SUV as seat count is important for the customers running taxi/cab.
- Finally, price should be low as it will be welcomed by all the customers.



REFERENCE

https://www.kaggle.com/mirosval/personal-cars-classifieds