Dubizzle is a corporation in Dubai, operating under the B2B model in many fields such as real estate, and ecommerce, ...

Case: Dubizzle wants an overview of the used car business in recent times, and they want to focus more on car models. With the increasing need to buy cars, this is also a problem that every business is facing in the same field.

I Data information

II Exploratory Data Analyst (EDA)

III Tranning model and compare

IV Conclude

Data information

I

Data include:

- 20 columns
- 9170 rows

DATA DICTIONARY

title	Car’s name
price_in_aed	Car’s price by dirham
kilometers	Distance the vehicle has traveled
body_condition	Body status
mechanical_condition	Engine car status
seller_type	Seller type (dealer, oner, other)
body_type	Body type (SUV, Sedan, other)
no_of_cylinder	Number of cylinder
transmission_type	Transmission type (auto, manual)
regional_spec	The nation of the band car

Data information

I

DATA DICTIONARY

horsepower	Calculate power of engine
fuel_type	Fuel type
steering_side	Driver seat
year	Car debut
color	Body color
emirates	Arab Emirates
motor_trim	Engine type
company	Company manufacture
model	Car model
date_post	Date upload advertise

Exploratory Data Analyst

II

Prepare data

1. Check data type

Columm **no_of_cylinder** and **horsepower** have value "**Unknown**"
⇒ Change value "**Unknown**" to **null** can fill data after

2. **Check null and duplicated**

- About duplicated data have 282 rows

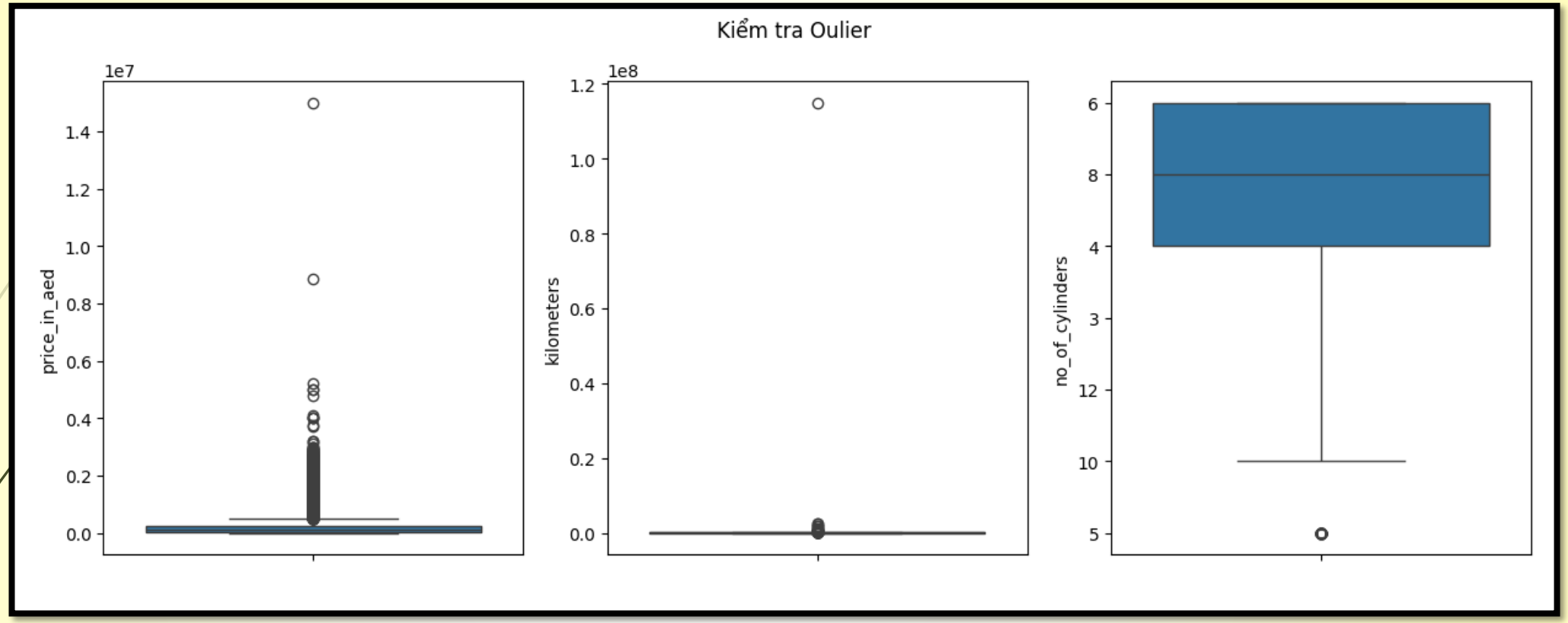
⇒ Remove duplicate data

- About null data:
  title: 5
  no_of_cylinder: 138 → fill median
  horsepower: 814 → fill median
  year: 970 → fill mode
  motor_trim: 28

⇒ After EDA, fill null to EDA data still origin data.

Exploratory Data Analyst

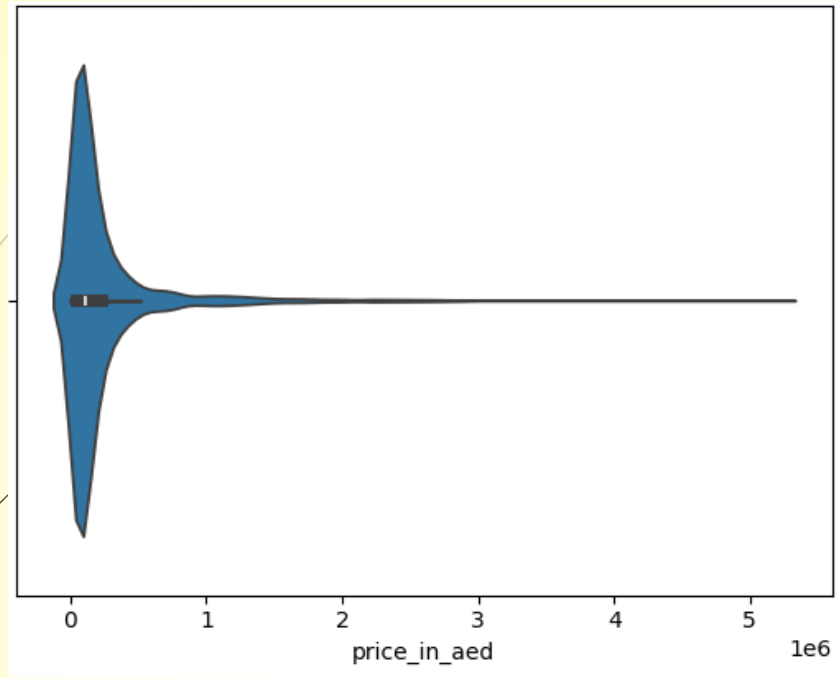
II

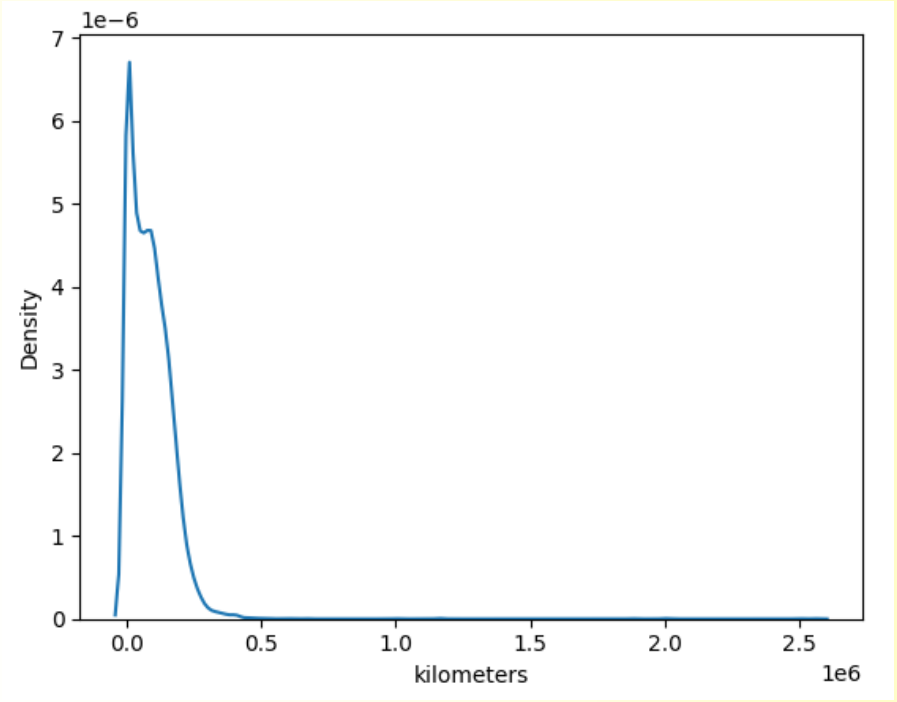


- Check outliers of columns with data (int, float) and remove outliers of price_in_aed and kilometers.

Exploratory Data Analyst

II

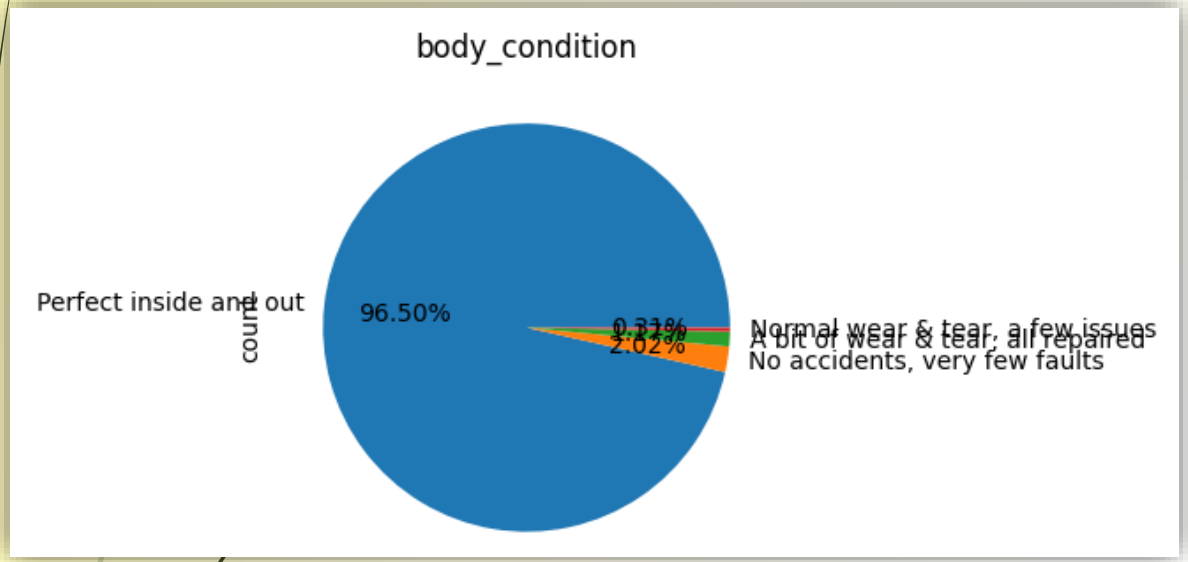


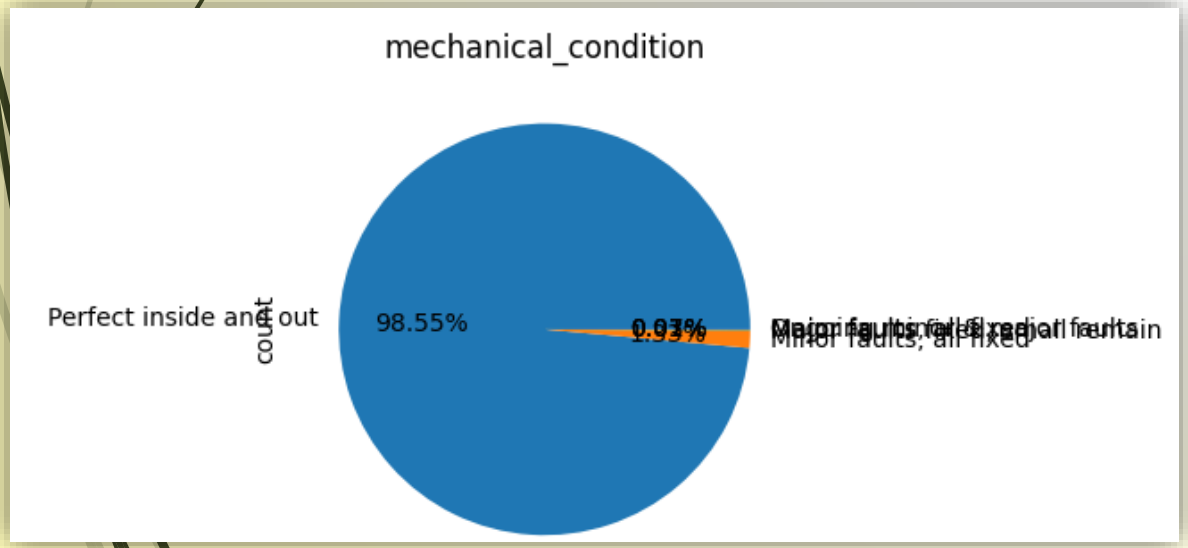


- Most car prices are under 1,000,000 VND
- The number of kilometers used by the car is mostly under 500,000km

Exploratory Data Analyst

II

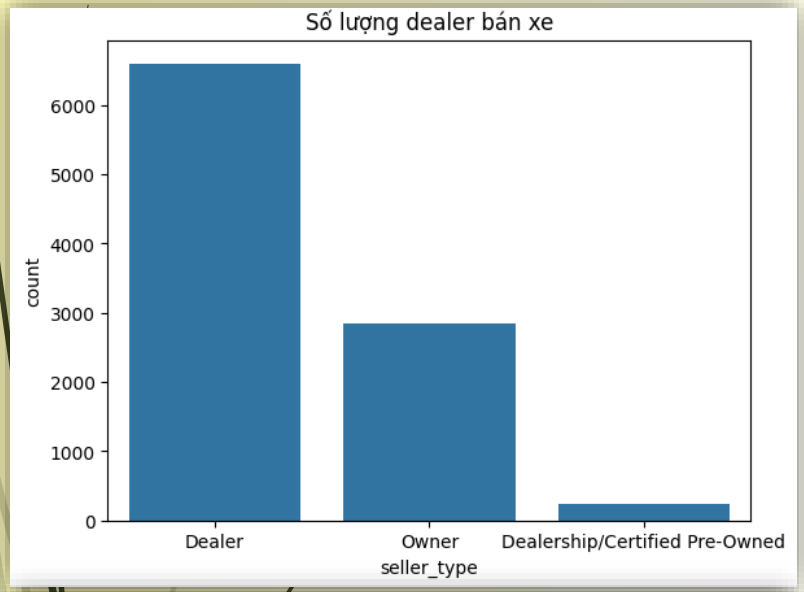


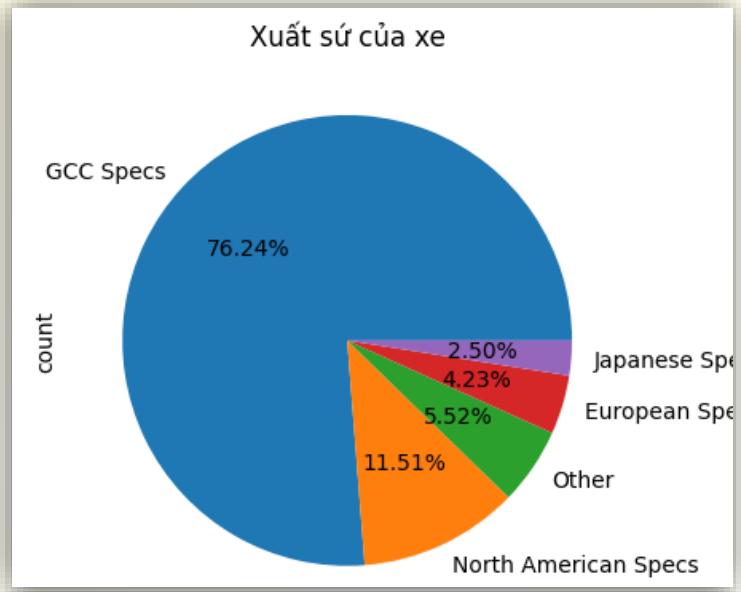


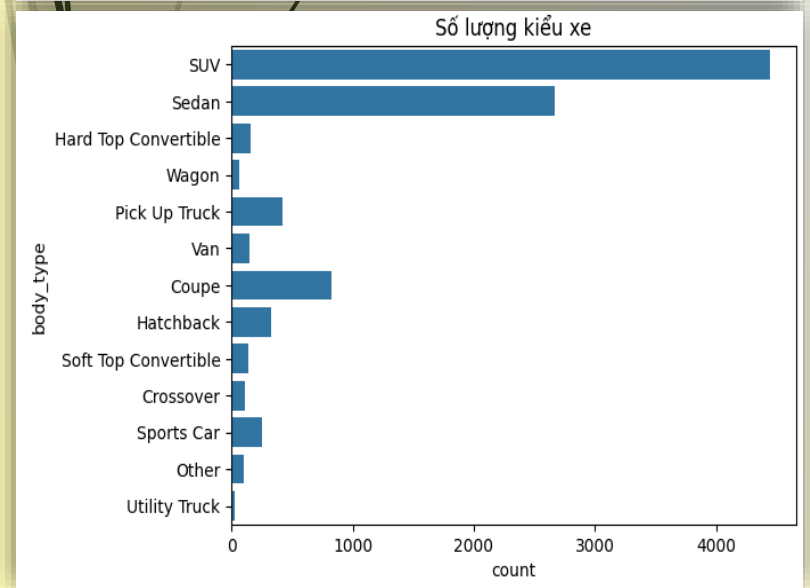
- The number of cars sold is almost intact without much damage.

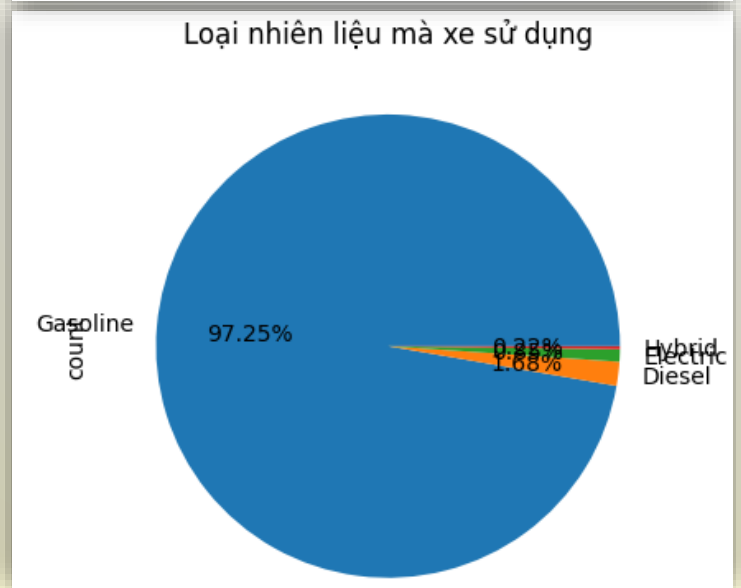
Exploratory Data Analyst

II





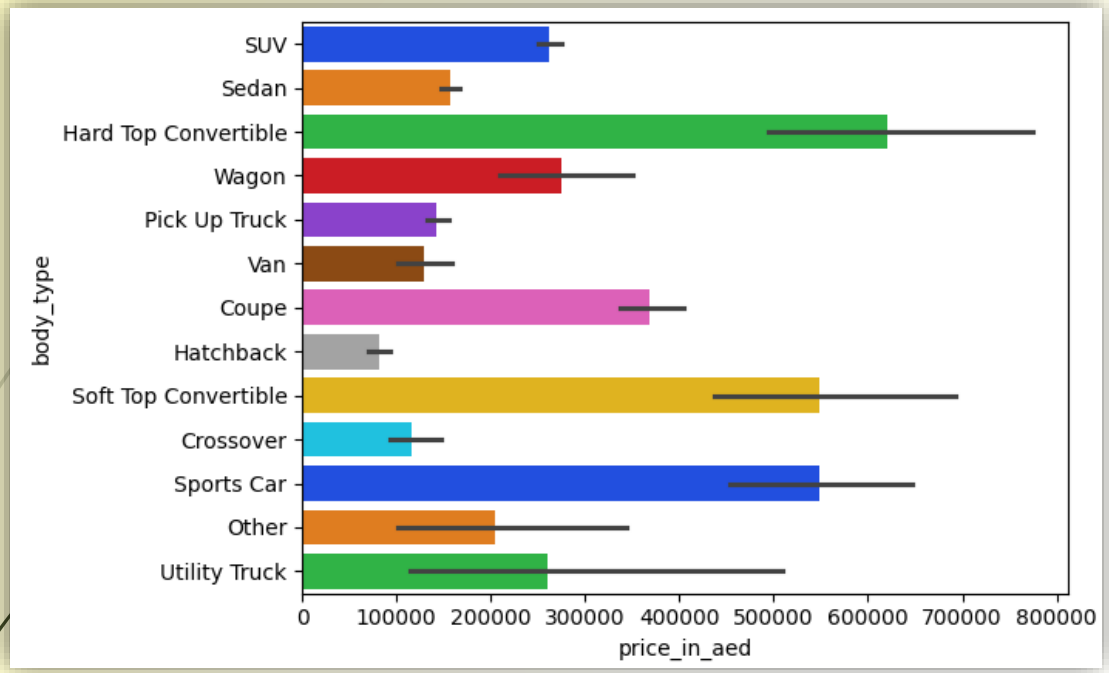


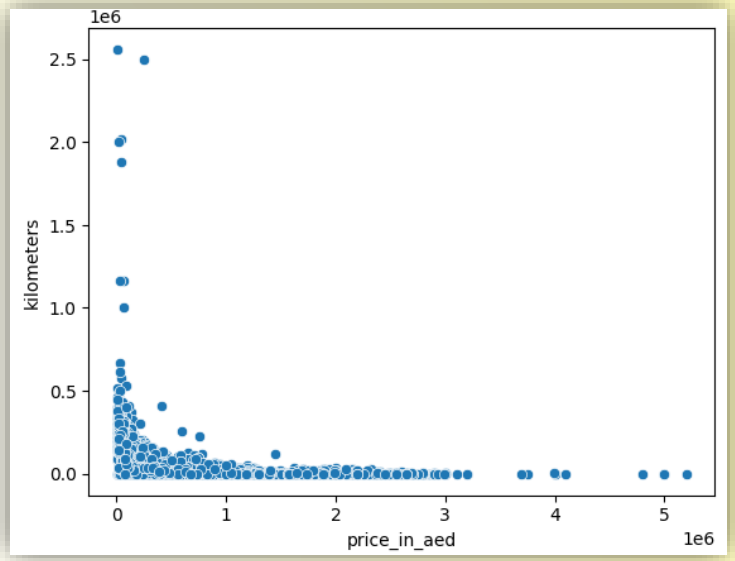


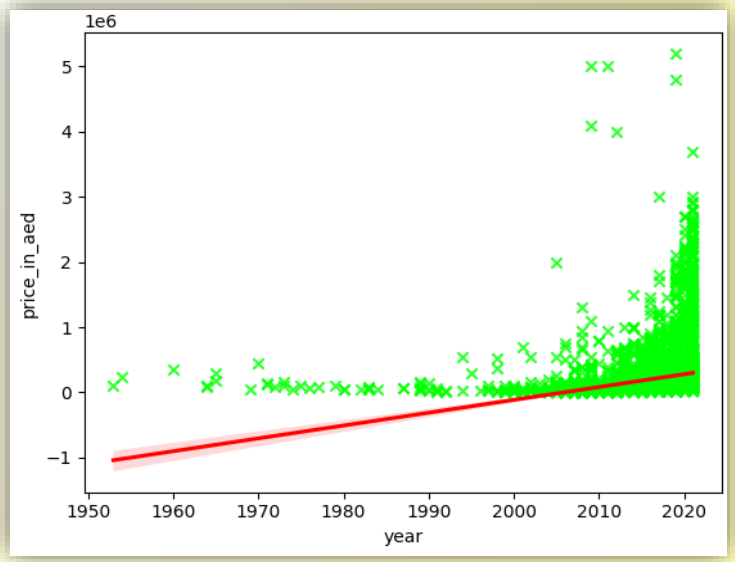
- Most car sellers are dealer.
- Sedan and SUV are the two most used vehicle types.
- The fuel that most cars use is gasoline
- Because of business in the Middle East, vehicles from the same region account for 76%, followed by North America and other countries.

Exploratory Data Analyst

II



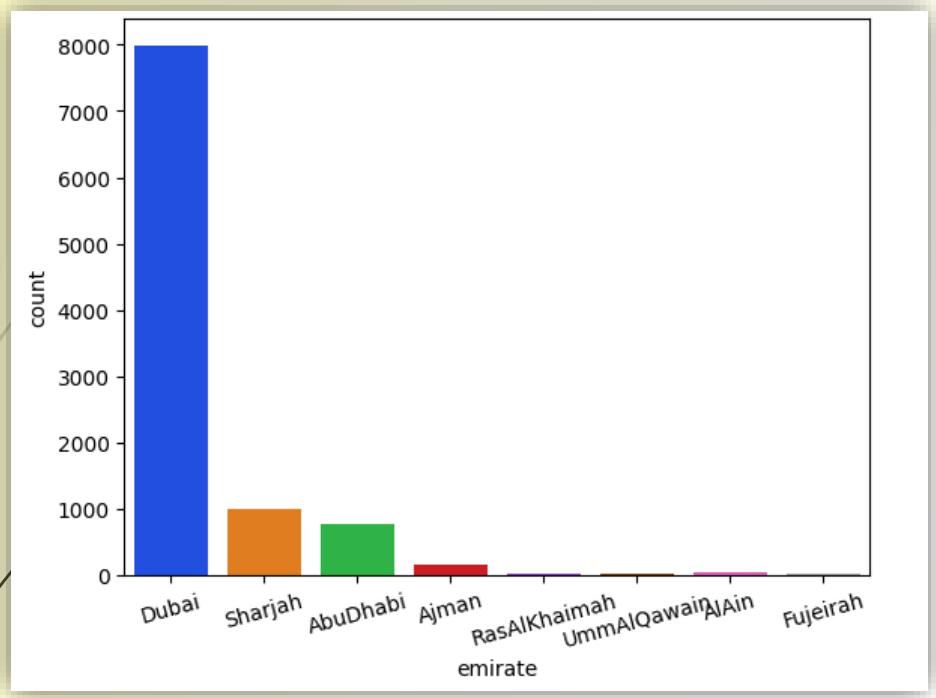


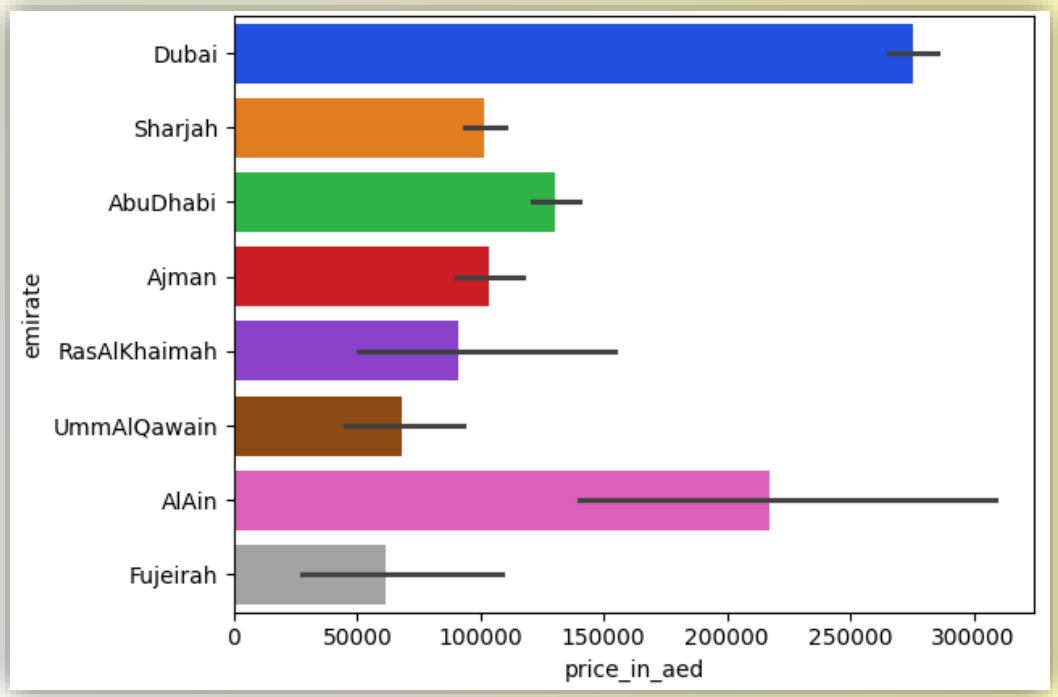


- The lower the number of kilometers, the higher the car price
- Cars debut since 2000 are selling more and more
- Although SUVs and Sedan sell a lot, the average price is about 100,000 to 200,000. Instead, sports cars and convertibles have very high prices.

Exploratory Data Analyst

II

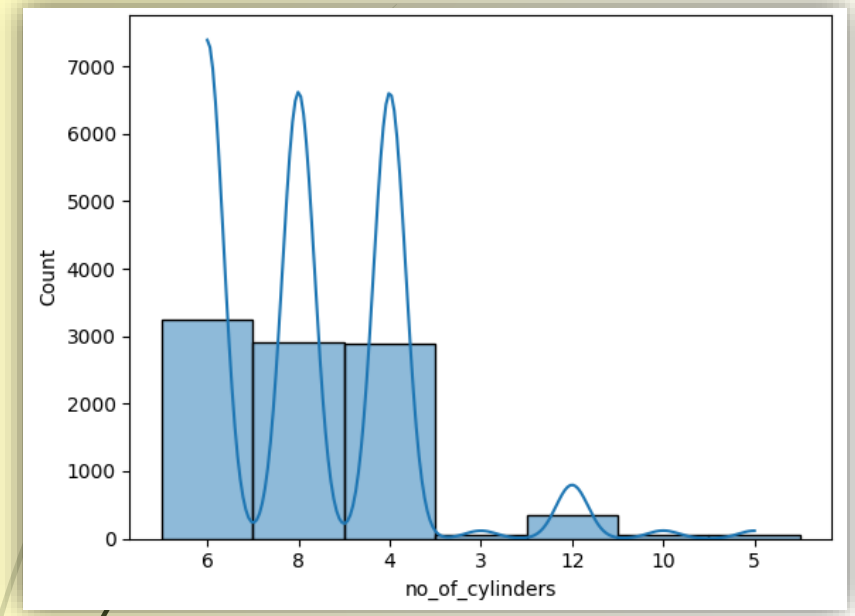


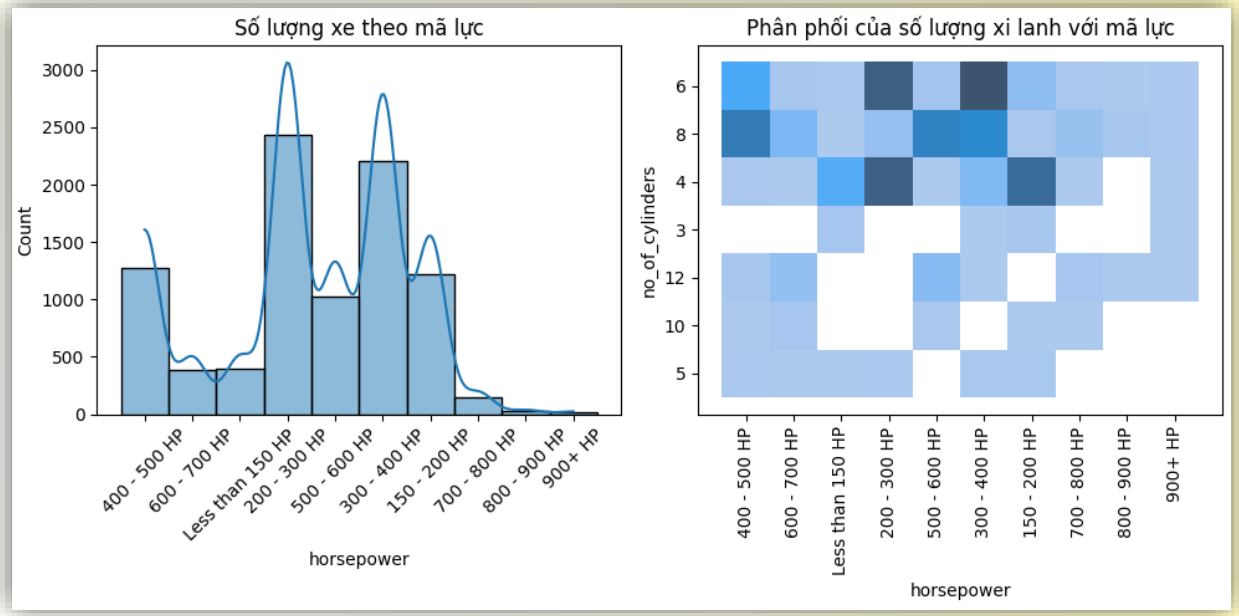


- Dubai is the most stable and popular car buying area

Exploratory Data Analyst

II

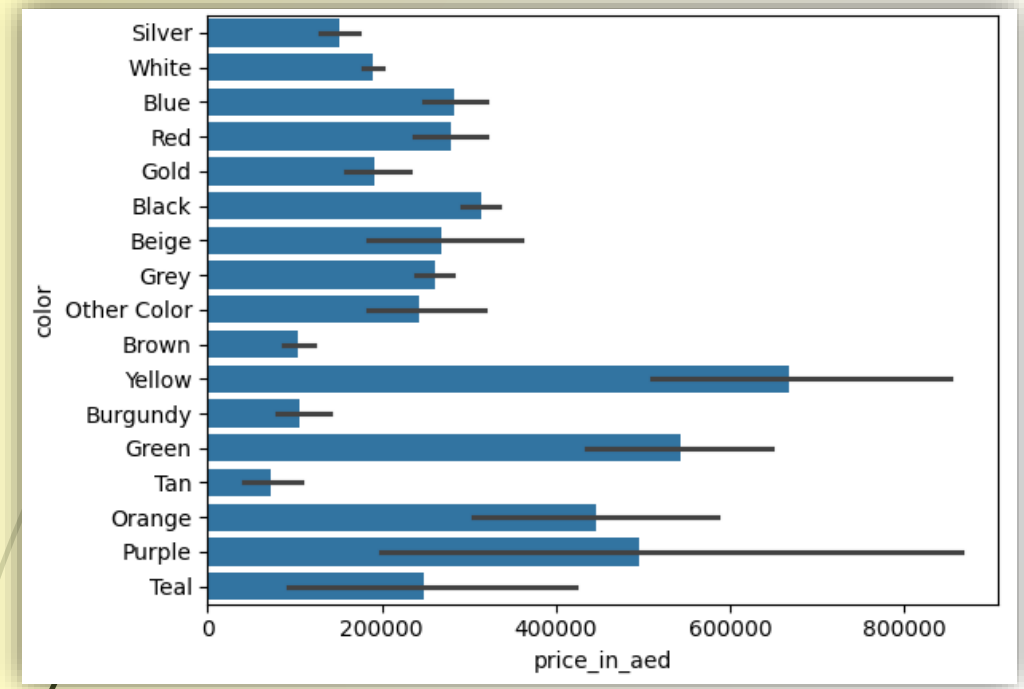


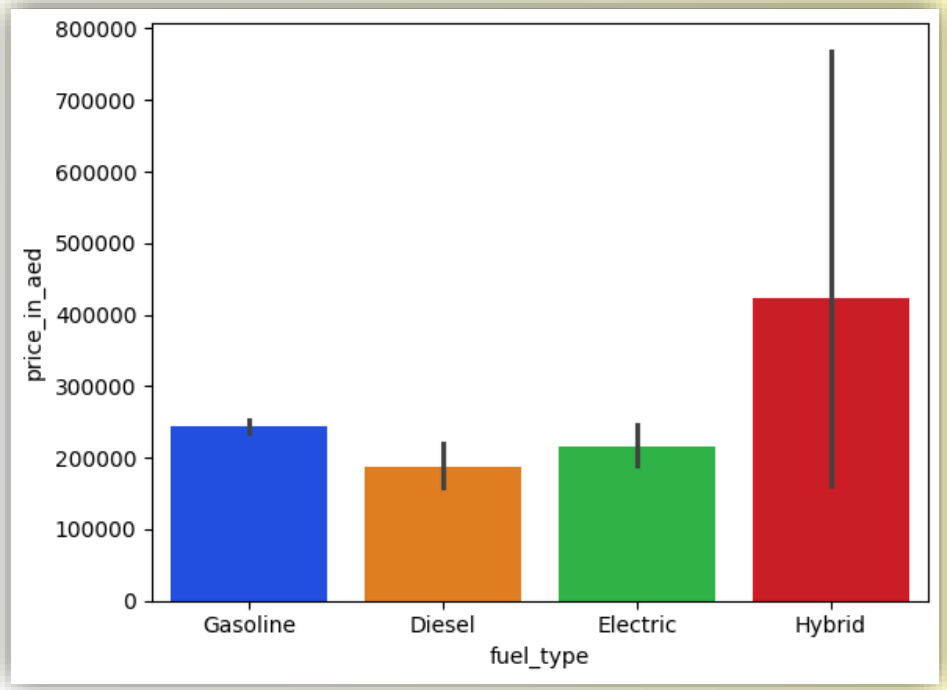


- The buyers cars with cylinders 4, 6, 8 is large, proportional to the number of horsepower of the same type, 200 - 300HP, 300 - 400HP and 400 - 500HP.

Exploratory Data Analyst

II

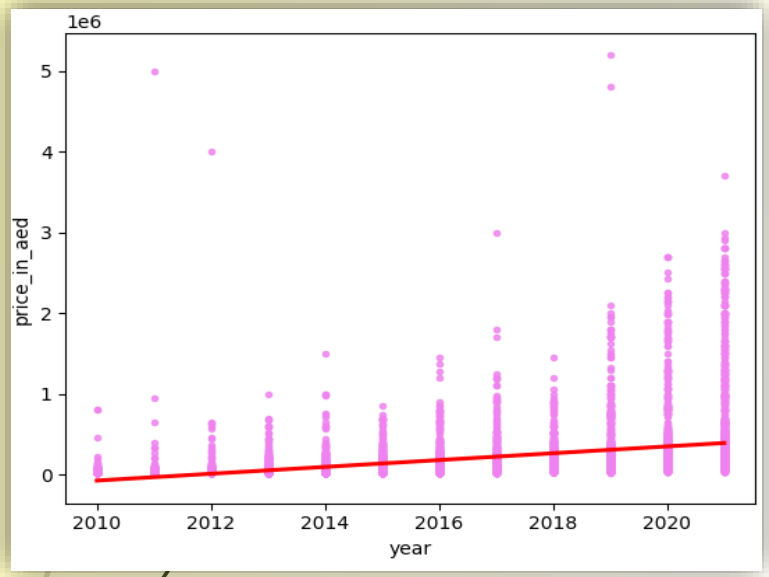


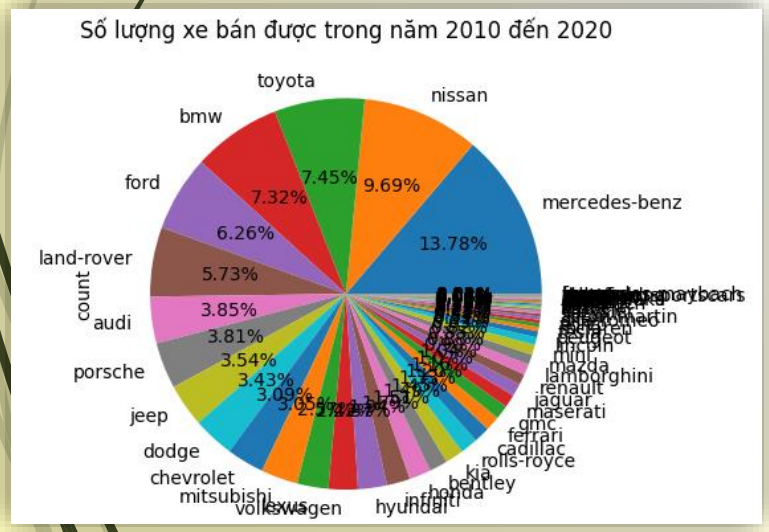


- Yellow is the color that sells for the highest price
- Fuel does not affect the price too much

Exploratory Data Analyst

II





- In general, cars debut from 2010 to 2020 sell more cars, most of them are still domestic models, besides there are Japanese and American car models. And expensive cars.

Traning data and compare

III

Data encoding

title	Car's name
price_in_aed	Car's price by dirham
kilometers	Distance the vehicle has traveled
body_condition	Body status
mechanical_condition	Engine car status
seller_type	Seller type (dealer, oner, other)
body_type	Body type (SUV, Sedan, other)
no_of_cylinder	Number of cylinder
transmission_type	Transmission type (auto, manual)
regional_spec	The nation of the band car

horsepower	Calculate power of engine
fuel_type	Fuel type
steering_side	Driver seat
year	Car debut
color	Body color
emirates	Arab Emirates
motor_trim	Engine type
company	Company manufacture
model	Car model
date_post	Date upload advertise

The red highlight lines are used to training model

Traning data and compare

III

Check correlation

	price_in_aed	kilometers	seller_type	body_type	no_of_cylinders	transmission_type	regional_specs	horsepower	fuel_type	year
price_in_aed	1.000000	-0.349631	-0.154449	0.068812	0.502244	-0.059099	0.200858	0.573126	-0.019033	0.221120
kilometers	-0.349631	1.000000	0.115975	-0.065792	-0.087253	0.031201	-0.104224	-0.192604	-0.090696	-0.405762
seller_type	-0.154449	0.115975	1.000000	0.018075	-0.131594	-0.019663	-0.112743	-0.092035	-0.053351	-0.106923
body_type	0.068812	-0.065792	0.018075	1.000000	-0.015603	0.151979	0.039958	0.066855	0.064395	-0.001938
no_of_cylinders	0.502244	-0.087253	-0.131594	-0.015603	1.000000	-0.119658	0.111202	0.724329	-0.052364	-0.086347
transmission_type	-0.059099	0.031201	-0.019663	0.151979	-0.119658	1.000000	0.045570	-0.093400	0.113587	-0.062163
regional_specs	0.200858	-0.104224	-0.112743	0.039958	0.111202	0.045570	1.000000	0.116555	0.145435	-0.033044
horsepower	0.573126	-0.192604	-0.092035	0.066855	0.724329	-0.093400	0.116555	1.000000	-0.046501	0.056031
fuel_type	-0.019033	-0.090696	-0.053351	0.064395	-0.052364	0.113587	0.145435	-0.046501	1.000000	0.059917
year	0.221120	-0.405762	-0.106923	-0.001938	-0.086347	-0.062163	-0.033044	0.056031	0.059917	1.000000

Traning data and compare

III

Train_Test: 8 : 2

StandardScaler

Metric: R2 score

Model name	R2 score
Linear Regression	-0.21
SVR	-1216
Dicision Tree Regressor	0.69
Random Forest Regressor	0.80
Ada Boost Regressor	0.66
Extra Tree Reegressor	0.81
Gradient Boosting Regressor	0.78
XGB Regressor	0.81

Conclude

IV

"SUV and sedans are selling very well."

Can expand into a new business starting from the previous two types of SUVs and Sedan.

Beside the Dubai area, the other areas have very low sales.

Promote marketing and expand services to increase vehicle usage in other areas.

The current trend is to use electric vehicles

Expanding markets to find new customers and marketing to reduce the use of fossil fuels.

Conclude

IV

Models are Tree and XGB provide highly accurate forecasting.

Suitable for forecasting revenue for the upcoming business period.

Random Forest Regressor	0.80
Extra Tree Reegressor	0.81
XGB Regressor	0.81