Exploratory Data Analysis.

Pedro Vinícius

pedrovgasparotti@gmail.com

2023 07 25

1 Feature analysis

The goal is to develop price predictions and answer market questions based on the provided dataset.

In order to identify the main dataframe features and statistics, it is necessary to focus in the numerical features, hence the target variable is the price of the car. Thus, the situation requires a regression to achieve the numerical predictions.

First, it is necessary to load and describe the features of the dataset:

1.1 Numerical Features Description

	id	num_fotos	ano_de_fabricacao	ano_modelo	\
count	2.958400e+04	29407.000000	29584.000000	29584.000000	
mean	1.705650e+38	10.323834	2016.758552	2017.808985	
std	9.814219e+37	3.487334	4.062422	2.673930	
min	1.332600e+34	8.000000	1985.000000	1997.000000	
25%	8.617510e+37	8.000000	2015.000000	2016.000000	
50%	1.706530e+38	8.000000	2018.000000	2018.000000	
75%	2.554710e+38	14.000000	2019.000000	2020.000000	
max	3.402560e+38	21.000000	2022.000000	2023.000000	

	hodometro	num_portas	veiculo_alienado	preco
count	29584.000000	29584.000000	0.0	2.958400e+04
mean	58430.592077	3.940677	NaN	1.330239e+05
std	32561.769309	0.338360	NaN	8.166287e+04
min	100.000000	2.000000	NaN	9.869951e+03
25%	31214.000000	4.000000	NaN	7.657177e+04
50%	57434.000000	4.000000	NaN	1.143558e+05
75%	81953.500000	4.000000	NaN	1.636796e+05
max	390065.000000	4.000000	NaN	1.359813e+06

Figure 1: Numeric Features Description

This description provides only a general count of the numerical features.

1.2 Correlation Matrix

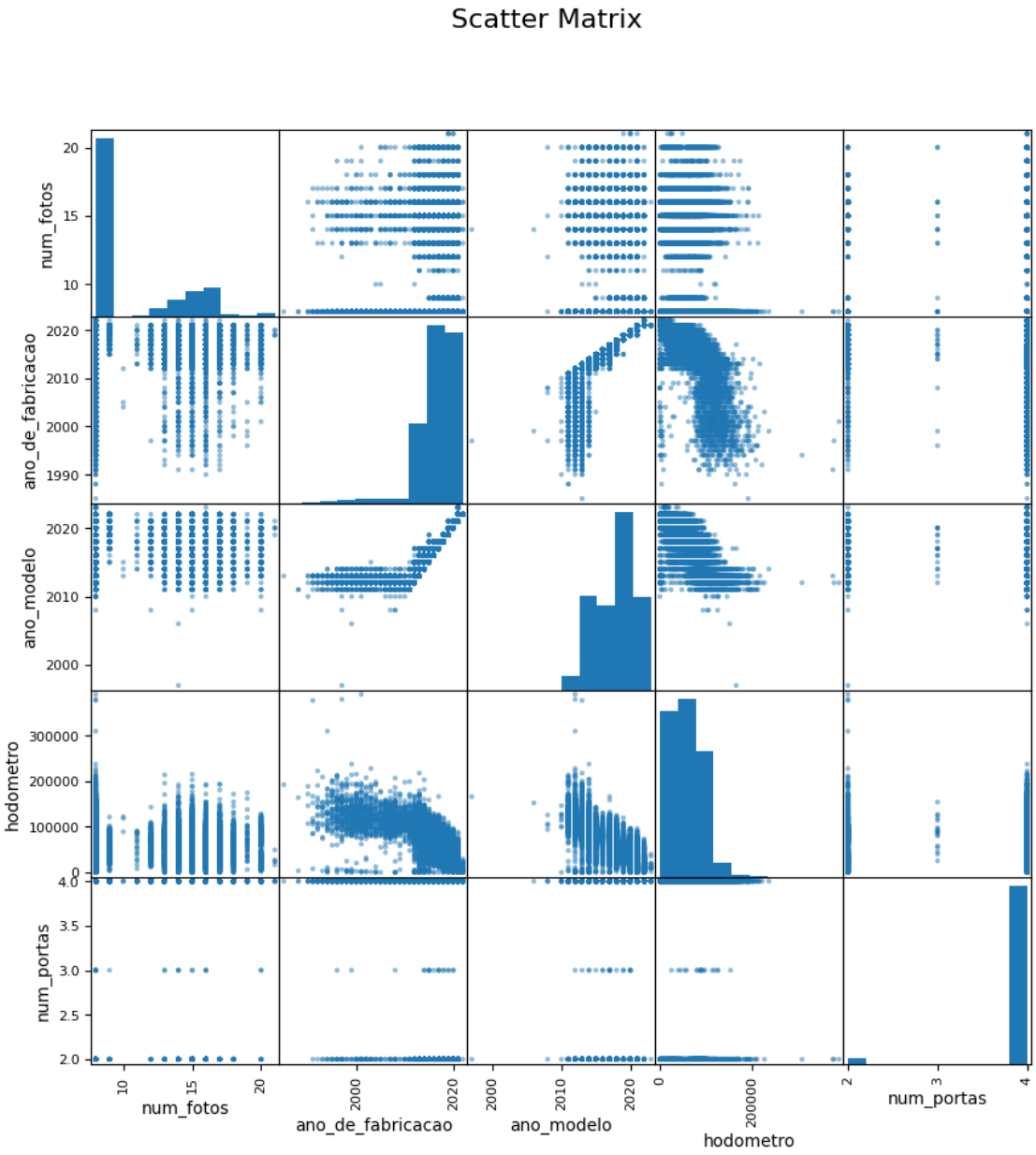

Figure 2: Correlation Matrix

The correlation matrix exhibit linear tendencies between 'hodometro', 'ano_modelo' and 'ano_de_fabricacao', so those variables are suited for a regression prediction.

1.3 Categorical Features

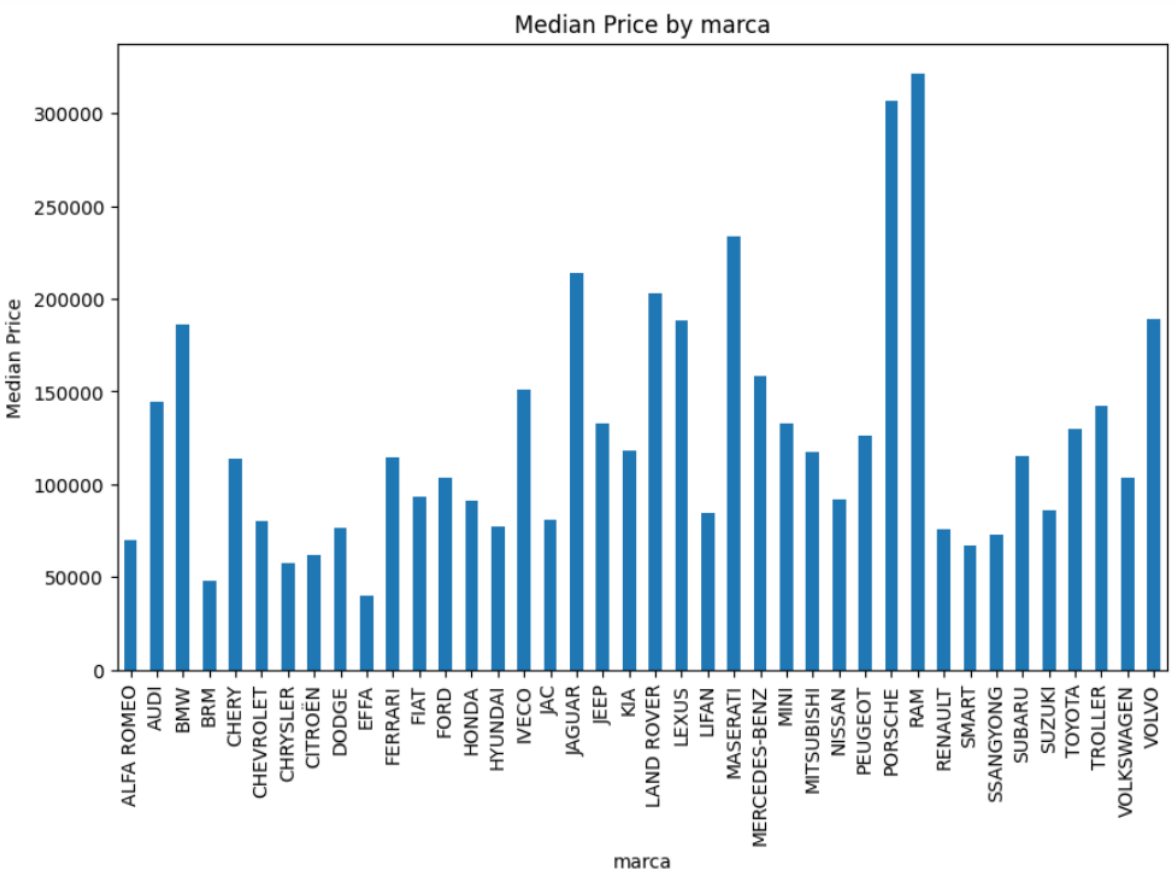

Figure 3: Cathegorical Feature Investigation

The prices vary for each vehicle brand, so, the model should also consider the relevant categorical features to make predictions. It has been observed that the price varies significantly according to the vehicle model and it's technical specification, altough it cannot be expressed visually.

That said, the most important features are 'marca', 'modelo' for categorical and 'hodometro', 'ano_modelo' and 'ano_de_fabricacao' for numerical.

2 Business Hypothesis

2.1 What is the best state to sell a popular brand car and why?

First, it is necessary to define what are the popular brands. To achieve that, the dataframe was filtered by cars bellow the average price, and only brands with enough number of samples where considered, as it follows:

```
# Filter by average, and eliminate samples with a small count

filtered_classes = summary_statistics_df[summary_statistics_df['mean'] < average_price]
filtered_classes = filtered_classes[summary_statistics_df['count'] > average_sample]

# Display the filtered DataFrame
print(filtered_classes)
```

```
         marca           mean         median  count
5    CHEVROLET   93187.683964   79934.798235   3020
11        FIAT   99711.164582   93202.070185   1918
13       HONDA  100620.715073   91506.027175   1586
14     HYUNDAI   84419.639626   77118.762160   2043
28     PEUGEOT  122797.835087  126420.198200   1675
38  VOLKSWAGEN  117940.087380  103350.092100   4594
```

Figure 4: Obtaining Popular Brands

The next step is identify the state in wich those brands are more valuable. First, a new dataframe is isolated, containing only the popular cars instances. Then, it is calculated the average price of the popular brands cars for each one of the states:

```
average_prices = df_marcas_populares.groupby('estado_vendedor')['preco'].mean()

# Print the average prices
print(average_prices)

print("State with highest average_price is: ", average_prices.idxmax())
```

```
estado_vendedor
Acre (AC)                    76202.462150
Alagoas (AL)                115331.244507
Amazonas (AM)                85600.767352
Bahia (BA)                  103447.392531
Ceará (CE)                  101979.092115
Espírito Santo (ES)          96095.233233
Goiás (GO)                  123187.094372
Maranhão (MA)               121041.689124
Mato Grosso (MT)            131479.497304
Mato Grosso do Sul (MS)      91542.079854
Minas Gerais (MG)           106611.617111
Paraná (PR)                 112967.354766
Paraíba (PB)                 91066.777475
Pará (PA)                   117677.240212
Pernambuco (PE)              97301.329251
Piauí (PI)                  166998.772860
Rio Grande do Norte (RN)    109105.107498
Rio Grande do Sul (RS)      111356.887024
Rio de Janeiro (RJ)         104746.253524
Rondônia (RO)               118363.820725
Roraima (RR)                 63613.691190
Santa Catarina (SC)         100559.480464
Sergipe (SE)                108353.825845
São Paulo (SP)              100082.412685
Tocantins (TO)              103968.621731
Name: preco, dtype: float64
State with highest average_price is:  Piauí (PI)
```

The state with the highest average price for popular brands is Piauí (PI), which qualifies this state as the best place to sell a popular brand car.

2.2 What is the best state to buy a pickup truck with automatic transmission?

The procedure is similar to the one presented above, but now filtering by 'cambio' and 'tipo', then looking for the lower price:

```
average_prices = picape_df.groupby('estado_vendedor')['preco'].mean()

# Print the average prices
print(average_prices)

print("State with lowest average_price is: ", average_prices.idxmin())
```

```
estado_vendedor
Acre (AC)                  145256.693662
Alagoas (AL)               207186.510773
Bahia (BA)                 191059.367089
Goiás (GO)                 176032.071975
Mato Grosso (MT)           214102.315650
Mato Grosso do Sul (MS)    144700.247632
Minas Gerais (MG)          175063.807123
Paraná (PR)                178192.445987
Paraíba (PB)                93157.035253
Pernambuco (PE)            182464.583045
Piauí (PI)                 208181.077750
Rio Grande do Norte (RN)   179961.694800
Rio Grande do Sul (RS)     171557.555193
Rio de Janeiro (RJ)        162029.423074
Santa Catarina (SC)        165469.196214
Sergipe (SE)               254108.597633
São Paulo (SP)             163663.940588
Tocantins (TO)             187717.127533
Name: preco, dtype: float64
State with lowest average_price is:  Paraíba (PB)
```

State with lowest average price is: Paraíba (PB), which qualifies this state as the best place to buy a pickup truck with automatic transmission.

2.3 What is the best state to buy cars that still have a manufacturer's warranty?

The procedure is similar to the previous ones , but now filtering by 'garantia_de_fábrica', then looking for the lower price:

```
average_price_garantia = garantia_df.groupby('estado_vendedor')['preco'].mean()

# Print the average prices
print(average_price_garantia)

print("State with lowest average_price is: ", average_price_garantia.idxmin())
```

```
estado_vendedor
Acre (AC)                   150416.911340
Alagoas (AL)                154268.676542
Amazonas (AM)                99617.303340
Bahia (BA)                  165221.236195
Ceará (CE)                  123939.878800
Espírito Santo (ES)         104030.208124
Goiás (GO)                  161709.106729
Mato Grosso (MT)            197657.066550
Mato Grosso do Sul (MS)     121709.589842
Minas Gerais (MG)           157531.959478
Paraná (PR)                 170214.863647
Paraíba (PB)                 95762.746630
Pará (PA)                    98156.615279
Pernambuco (PE)             149898.416932
Rio Grande do Norte (RN)    133120.393897
Rio Grande do Sul (RS)      169001.736437
Rio de Janeiro (RJ)         174742.392511
Santa Catarina (SC)         163290.559023
Sergipe (SE)                318314.436800
São Paulo (SP)              161694.077183
Tocantins (TO)              243002.217000
Name: preco, dtype: float64
State with lowest average_price is:  Paraíba (PB)
```

State with lowest average price is: Paraíba (PB), which qualifies this state as the best place to buy cars that still have a manufacturer's warranty.

3 Predict Prices

The prices will be predicted by a Multi-Input Neural Network. It is called "multi-input" because it can accept multiple different types of inputs, such as numerical features and categorical features, then process them together to predict numerical values. Multi-Input Neural Network are flexible models that can handle different types of features, capture complex relationships, and adapt well to regression tasks, making it a good choice for price prediction and similar regression problems.