Chris Troop
2024-09-23
CSC6510 Project M1

Used Car Data Report

The used car data set is a compilation of information about Toyota Corollas . This data set includes 1436 individual cars that have 37 attributes. The attributes names and their inferred data types by pandas are listed below:

- Id - int64
- Model - Category
- Price - int64
- Age_08_04 - int64
- KM - int64
- Fuel_Type - Object
- HP - int64
- Color - Object
- Automatic - Object
- CC - int64
- Doors - Object
- Cylinders - Object
- Gears - Object
- Quarterly_Tax - int64
- Weight - int64
- Mfr_Guarantee - Object
- BOVAG_Guarantee - Object
- Guarantee_Period - Object
- ABS - Object
- Airbag_1 - Object
- Airbag_2 - Object
- Airco - Object
- Automatic_airco- Object
- Boardcomputer- Object
- CD_Player- Object
- Central_Lock- Object
- Powered_Windows- Object
- Power_Steering- Object
- Radio- Object
- Mistlamps- Object
- Sport_Model- Object
- Backseat_Divider - Object
- Metallic_Rim - Object
- Radio_cassette - Object
- Parking_Assistant- Object
- Tow_Bar - Object
- Mfg_Date - Object

To begin cleaning the data, the category variable types were changed to better match what each represented.

- Id - int64
- Model - Category
- Price - float64
- Age_08_04 - int64
- KM - int64
- Fuel_Type - Category
- HP - int64
- Color - Category
- Automatic - Category
- CC - int64
- Doors - Category
- Cylinders - Category
- Gears - Category
- Quarterly_Tax - float64
- Weight - int64
- Mfr_Guarantee - Category
- BOVAG_Guarantee - Category
- Guarantee_Period - Category

- ABS - Category
- Airbag_1 - Category
- Airbag_2 - Category
- Airco - Category
- Automatic_airco- Category
- Boardcomputer- Category
- CD_Player- Category
- Central_Lock- Category
- Powered_Windows- Category
- Power_Steering- Category
- Radio- Category
- Mistlamps- Category
- Sport_Model- Category
- Backseat_Divider - Category
- Metallic_Rim - Category
- Radio_cassette - Category
- Parking_Assistant- Category
- Tow_Bar - Category
- Mfg_Date - datetime64

The variable type changes that were changed from object to category were driven by that variable having a few repeating values that most if not all data points shared. The int64 to float64 changes for “Price” and “Quarterly_Tax” were driven by cash values rarely being round numbers and the common practice of including two decimal places for cents when writing out dollar amounts. “Mfg_Date” was changed from Object to datetime64 to better reflect the type of variable.

Eight vehicles were dropped from the dataset for missing data or a value outside of realistic possibility. Row 80 was removed for a 16000 CC engine, which is far too large to be a real vehicle. Rows 171, 244, and 301 were removed for having “?” for “Automatic”. Rows 203, 271, 277, 355 were removed for having 0.0 for “Quarterly_Tax”. All vehicles with -1 in the cylinder column were replaced with pd.NA. All vehicles with “3” for “Gaurantee_Period” were changed to “36”. The belief is that these values are un years and the remaining values are in months, thus the conversion of 3 years to 36 months.

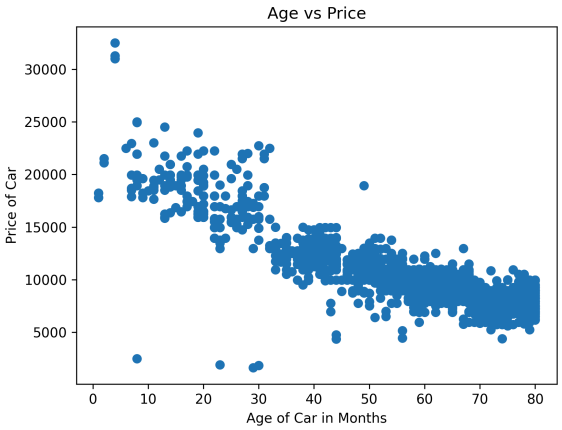
Below is a brief summary of the numerical variables from the data set using the describe() method in pandas.

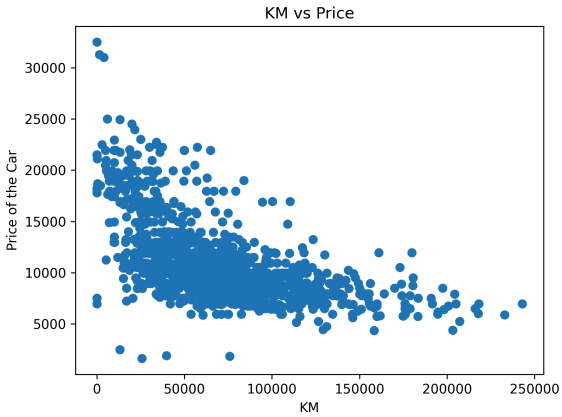
3]:

	Id	Price	Age_08_04	KM	HP	CC	Doors	Gears	Quarterly_Tax	Weight	Guarantee_Period
count	1428.000000	1428.000000	1428.000000	1428.000000	1428.000000	1428.000000	1428.000000	1428.000000	1428.000000	1428.000000	1428.000000
mean	724.268908	10661.277311	56.082633	68674.696779	101.483193	1566.432073	4.033613	5.026611	87.087535	1072.378852	33.084734
std	416.016249	3606.998601	18.548073	37523.730773	14.986530	187.484257	0.952401	0.189028	41.152340	52.661086	8.405322
min	1.000000	1625.000000	1.000000	1.000000	69.000000	1300.000000	2.000000	3.000000	19.000000	1000.000000	6.000000
25%	367.750000	8450.000000	44.000000	43000.000000	89.000000	1400.000000	3.000000	5.000000	69.000000	1040.000000	36.000000
50%	725.500000	9897.500000	61.000000	63567.000000	110.000000	1600.000000	4.000000	5.000000	85.000000	1070.000000	36.000000
75%	1083.250000	11900.000000	70.000000	87130.750000	110.000000	1600.000000	5.000000	5.000000	85.000000	1085.000000	36.000000
max	1442.000000	32500.000000	80.000000	243000.000000	192.000000	2000.000000	5.000000	6.000000	283.000000	1615.000000	36.000000

The price of cars ranges from 1625 to 32500 which is a fairly large range considering all the cars in the data set are the same make and model. This requires further variables to explain such as age and KM. The Age of cars ranges from 80 months down to 1 month. The KMs is a similar story with a range of 243000 down to 1. The motor variables, HP and CC, are other variables that see a significant range. HP has a range of 69 to 192 and CC ranges from 1300 to 2000.

The “so what” from the data set is that there seems to be a relationship between age and price as well as km and price. Plots below.





Both plots show a negative relationship between price and the variable graphed.