Data Mining Final Project

Vehicle Sale Data

Table of Contents

[Explanatory Data Analysis 3](#_Toc166762970)

[Data Set Information 3](#_Toc166762971)

[Data Set Cleaning 4](#_Toc166762972)

[Insights (Knowing Data) 6](#_Toc166762973)

[Feature: Year 6](#_Toc166762974)

[Feature: Make 7](#_Toc166762975)

[Feature: Model 9](#_Toc166762976)

[Feature: State 10](#_Toc166762977)

[Feature: Color 11](#_Toc166762978)

[Feature: Interior Color 11](#_Toc166762979)

[Feature: Selling Price 12](#_Toc166762980)

[Feature: Season (Sale Date) 13](#_Toc166762981)

[Feature: Odometer 14](#_Toc166762982)

[Feature: Body 15](#_Toc166762983)

[Feature: Make-Model 15](#_Toc166762984)

[Model Evaluation 19](#_Toc166762985)

[Step 1: Data Processing 19](#_Toc166762986)

[Step 2: Model Training and Evaluation 19](#_Toc166762987)

[Step 3: Ensemble Method 21](#_Toc166762988)

[Step 4: Cross Validation 22](#_Toc166762989)

[Step 5: Compare Results 22](#_Toc166762990)

[Step 6: Feature Engineering 23](#_Toc166762991)

[Summary 26](#_Toc166762992)

# Explanatory Data Analysis

## Data Set Information

Vehicle sale data information download from kaggle for the final project. Link as seen below.

<https://www.kaggle.com/datasets/syedanwarafridi/vehicle-sales-data/data?select=car_prices.csv>

Total 16 features and 558837 rows exist within the dataset. Feature names and the descriptions are also seen as seen below. Though the project we will be working on this features in order to accomplish the goal. Our aim is to predict sale price based on year, make, and model information. By the end of the project we will add condition parameter to see how this affect to the model evaluation metrics.

|  |  |  |
| --- | --- | --- |
| **Feature** | **Description** | **Data Type** |
| Year | The manufacturing year of the vehicle. | Numerical |
| Make | The brand or manufacturer of the vehicle. | Categorical |
| Model | The specific model of the vehicle. | Categorical |
| Trim | Additional designation for the vehicle model. | Categorical |
| Body | The body type of the vehicle (e.g., SUV, Sedan). | Categorical |
| Transmission | The type of transmission in the vehicle (e.g., automatic). | Categorical |
| Vin | Vehicle Identification Number, a unique code for each vehicle | Numerical |
| State | The state where the vehicle is registered | Categorical |
| Condition | Condition of the vehicle, possibly rated on a scale. | Numerical |
| Odometer | The mileage or distance traveled by the vehicle. | Numerical |
| Color | Exterior color of the vehicle. | Categorical |
| Interior | Interior color of the vehicle. | Categorical |
| Seller | The entity selling the vehicle | Categorical |
| Mmr | Manheim Market Report, possibly indicating the estimated market value of the vehicle. | Numerical |
| Selling price | The price at which the vehicle was sold. | Numerical |
| Sale data | The date and time when the vehicle was sold. | Numerical |

## Data Set Cleaning

This step will include several steps during the project but we believe to mention all removed items to mentioned in this section.

* Fresh dataset has 558837 rows
* 86512 null values exist within the dataset. After removing them 472325 rows left
* 6352 records was useless based on year information. There were lots of years and was not any car sales exist. After removing them dataset became 465953 rows
* 294 row are related with the very less car sales like Ferrari, Lamborghini. This was affecting the scale of the car sale disctribution and affecting the prediction. After removing them dataset has 465659 rows
* 9946 rows removed less than 100 car sales. This was affecting very special old cars like before year 2000 models. So it is not interested in our project. After that dataset became 455713
* Finally there were 13193 records had 1 USD sale price in the dataset. We could use the mean of the sale price but since the data set still plenty rows we removed thos rows. After that data set became 442520 rows

Special cars, 1 USD sale price, before year 2000 records were also having outliers in our dataset.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

Before and after data cleaning data set statistics:

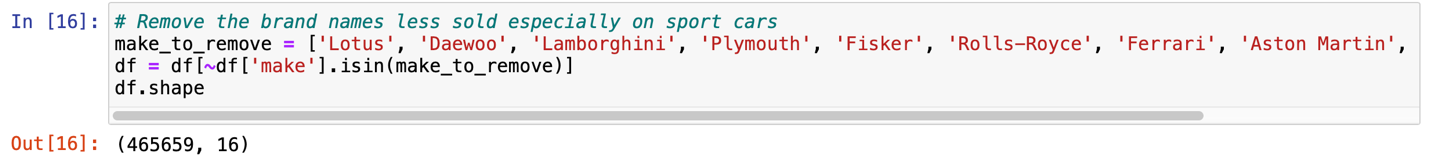
Better look on the statistical distribution perspective

A table with numbers and symbols

Description automatically generatedA table with numbers and a few ones

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generatedA close-up of a computer code

Description automatically generated

A close-up of a computer code

Description automatically generated

A screenshot of a phone

Description automatically generated

# Insights (Knowing Data)

## Feature: Year

As we can see from the below graphs there were not many car sale acitivies were exist and removed those values. Histogram gets better to show vehicle car sales increased during the years. Especially had a peak on 2012 and 2013. And started to reduce with 2014. Car sale activities seems affected by the economical situaitons

A graph of a number of years

Description automatically generated

A graph of a growing graph

Description automatically generated with medium confidence

## Feature: Make

As we can see from the below graphs most trusted car brands in USA are Ford, Checrolet, Nissan, Toyota, Dodge, and Honda. Sale data as seen below: Ford: 80247 – Chevrolet: 53650 – Nissan: 43726 – Toyota: 34473 – Dodge: 27087 – Honda: 23833

A graph of a bar graph

Description automatically generated

If we go further detail on the top 3 car brands which models are the most sold, graphs as seen below. Ford has Fusion, F-150, and Escape models. Chevroler has Impala, Malibu, and Silverado models. Nissan has Altima, Maxima, and Sentra.

We can have an insights from the graphs that sedan cars are mostly selected by the consumers except F-150 by Ford which is the only truck within the list.

A graph of a number of cars

Description automatically generatedA graph of a number of numbers

Description automatically generated

A graph of numbers and names

Description automatically generated with medium confidence

## Feature: Model

Comparing models non related with the brand look Nissan Altima is the top sold car in the US. It is followed with the Ford Fusion and Ford F-150.

A pie chart with numbers and a circle

Description automatically generated

A graph of sales

Description automatically generated

## Feature: State

Top 5 car sale activities exist in the US are Florida, California, Texas, Georgia, and Philadelphia. This insight sohws that car sale market especially related with the population’ economical situation. These are the richest states in the US.

A graph of sales

Description automatically generated

## Feature: Color

Customers are mostly selected the black, white, gray, and silver cars. Shiny colors are not selected for used. Our experiment is focusing mid level cars so super sport cars are not included in this subject.

A graph of sales

Description automatically generated

## Feature: Interior Color

Customers mostluy select the either black or gray interior colors. This insight shows that there is a relation between color car and the interior color. Such as white color car interior mostly black but black color can be either gray or black.

A graph of sales

Description automatically generated

## Feature: Selling Price

Selling price data was really up to 350K USD because of the some rarely sold cars. After cleaning outliers.

A graph with blue bars

Description automatically generated

A blue rectangular object with black dots

Description automatically generated

## Feature: Season (Sale Date)

Car sales mostly happened either December-February or summer time. This might be related with the new models comes to sale and sellers wants to sell early year models with discount. Summer time is also inetresting may be related with the customers want to change their car before holiday plan.

A graph of blue rectangular bars

Description automatically generated with medium confidence

## Feature: Odometer

Most customers select less than 100K miles on the car when they prefer to buy a car.Since o miles cars are also in the 0-50K group this might also says that to use customers prefer new or near new cars to buy.

A graph of cars sold based on odometer readings

Description automatically generated

## Feature: Body

From the body types perspective most selected body types of the cars are either sedan or SUV. We can say from the data set that cusomer mostly select the family-based big cars.

A graph of a car type

Description automatically generated

## Feature: Make-Model

We have liisted based on brand and model how many sales occurred.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **make** | **model** | **count** |  | **make** | **model** | **count** |  | **make** | **model** | **count** |
| 0 | Acura | ILX | 103 | 128 | Ford | Five Hundred | 340 | 256 | Mazda | MX-5 Miata | 183 |
| 1 | Acura | MDX | 1160 | 129 | Ford | Flex | 1097 | 257 | Mazda | Mazda2 | 341 |
| 2 | Acura | RDX | 297 | 130 | Ford | Focus | 9545 | 258 | Mazda | Mazda3 | 2378 |
| 3 | Acura | RSX | 130 | 131 | Ford | Freestar | 246 | 259 | Mazda | Mazda5 | 537 |
| 4 | Acura | TL | 1602 | 132 | Ford | Freestyle | 262 | 260 | Mazda | Mazda6 | 1323 |
| 5 | Acura | TSX | 789 | 133 | Ford | Fusion | 12113 | 261 | Mazda | Tribute | 204 |
| 6 | Audi | A3 | 214 | 134 | Ford | Fusion Hybrid | 613 | 262 | Mercedes-Benz | C-Class | 4358 |
| 7 | Audi | A4 | 1863 | 135 | Ford | Mustang | 4369 | 263 | Mercedes-Benz | CL-Class | 105 |
| 8 | Audi | A5 | 279 | 136 | Ford | Ranger | 1256 | 264 | Mercedes-Benz | CLA-Class | 110 |
| 9 | Audi | A6 | 536 | 137 | Ford | Taurus | 4018 | 265 | Mercedes-Benz | CLK-Class | 396 |
| 10 | Audi | A7 | 2 | 138 | Ford | Transit Connect | 595 | 266 | Mercedes-Benz | CLS-Class | 223 |
| 11 | Audi | A8 | 134 | 139 | Ford | Windstar | 120 | 267 | Mercedes-Benz | E-Class | 2670 |
| 12 | Audi | Q5 | 349 | 140 | GMC | Acadia | 1682 | 268 | Mercedes-Benz | GL-Class | 592 |
| 13 | Audi | Q7 | 243 | 141 | GMC | Canyon | 112 | 269 | Mercedes-Benz | GLK-Class | 659 |
| 14 | Audi | S4 | 126 | 142 | GMC | Envoy | 558 | 270 | Mercedes-Benz | M-Class | 1265 |
| 15 | Audi | S5 | 75 | 143 | GMC | Envoy XL | 230 | 271 | Mercedes-Benz | R-Class | 276 |
| 16 | Audi | TT | 109 | 144 | GMC | Savana Cargo | 554 | 272 | Mercedes-Benz | S-Class | 782 |
| 17 | BMW | 1 Series | 361 | 145 | GMC | Sierra 1500 | 1921 | 273 | Mercedes-Benz | SL-Class | 263 |
| 18 | BMW | 3 Series | 6597 | 146 | GMC | Sierra 2500HD | 292 | 274 | Mercedes-Benz | SLK-Class | 233 |
| 19 | BMW | 4 Series | 25 | 147 | GMC | Terrain | 992 | 275 | Mercury | Grand Marquis | 305 |
| 20 | BMW | 5 Series | 2895 | 148 | GMC | Yukon | 1052 | 276 | Mercury | Mariner | 252 |
| 21 | BMW | 5 Series Gran Turismo | 90 | 149 | GMC | Yukon XL | 1011 | 277 | Mercury | Milan | 261 |
| 22 | BMW | 6 Series | 264 | 150 | HUMMER | H2 | 286 | 278 | Mercury | Mountaineer | 284 |
| 23 | BMW | 7 Series | 836 | 151 | HUMMER | H3 | 406 | 279 | Mercury | Sable | 150 |
| 24 | BMW | M | 6 | 152 | Honda | Accord | 7947 | 280 | Mitsubishi | Eclipse | 349 |
| 25 | BMW | M3 | 133 | 153 | Honda | Accord Crosstour | 102 | 281 | Mitsubishi | Eclipse Spyder | 119 |
| 26 | BMW | X1 | 256 | 154 | Honda | CR-V | 2649 | 282 | Mitsubishi | Endeavor | 160 |
| 27 | BMW | X3 | 901 | 155 | Honda | CR-Z | 101 | 283 | Mitsubishi | Galant | 502 |
| 28 | BMW | X5 | 1485 | 156 | Honda | Civic | 6785 | 284 | Mitsubishi | Lancer | 974 |
| 29 | BMW | X6 | 58 | 157 | Honda | Crosstour | 290 | 285 | Mitsubishi | Lancer Evolution | 119 |
| 30 | BMW | Z4 | 190 | 158 | Honda | Element | 250 | 286 | Mitsubishi | Mirage | 201 |
| 31 | Buick | Century | 141 | 159 | Honda | Fit | 407 | 287 | Mitsubishi | Montero Sport | 122 |
| 32 | Buick | Enclave | 907 | 160 | Honda | Insight | 213 | 288 | Mitsubishi | Outlander | 490 |
| 33 | Buick | LaCrosse | 1053 | 161 | Honda | Odyssey | 2984 | 289 | Mitsubishi | Outlander Sport | 427 |
| 34 | Buick | LeSabre | 364 | 162 | Honda | Pilot | 1807 | 290 | Nissan | 370Z | 187 |
| 35 | Buick | Lucerne | 467 | 163 | Honda | Ridgeline | 185 | 291 | Nissan | Altima | 16248 |
| 36 | Buick | Regal | 471 | 164 | Hyundai | Accent | 1762 | 292 | Nissan | Armada | 720 |
| 37 | Buick | Rendezvous | 260 | 165 | Hyundai | Azera | 278 | 293 | Nissan | Cube | 230 |
| 38 | Buick | Verano | 300 | 166 | Hyundai | Elantra | 5343 | 294 | Nissan | Frontier | 1001 |
| 39 | Cadillac | ATS | 167 | 167 | Hyundai | Elantra GT | 222 | 295 | Nissan | Juke | 434 |
| 40 | Cadillac | CTS | 1627 | 168 | Hyundai | Elantra Touring | 106 | 296 | Nissan | Leaf | 1509 |
| 41 | Cadillac | CTS Coupe | 120 | 169 | Hyundai | Equus | 66 | 297 | Nissan | Maxima | 5364 |
| 42 | Cadillac | DTS | 484 | 170 | Hyundai | Genesis | 753 | 298 | Nissan | Murano | 2436 |
| 43 | Cadillac | DeVille | 460 | 171 | Hyundai | Genesis Coupe | 255 | 299 | Nissan | Pathfinder | 1586 |
| 44 | Cadillac | Escalade | 776 | 172 | Hyundai | Santa Fe | 1427 | 300 | Nissan | Quest | 610 |
| 45 | Cadillac | Escalade ESV | 381 | 173 | Hyundai | Santa Fe Sport | 262 | 301 | Nissan | Rogue | 4177 |
| 46 | Cadillac | Escalade EXT | 120 | 174 | Hyundai | Sonata | 5786 | 302 | Nissan | Sentra | 4180 |
| 47 | Cadillac | SRX | 947 | 175 | Hyundai | Sonata Hybrid | 509 | 303 | Nissan | Titan | 734 |
| 48 | Cadillac | STS | 256 | 176 | Hyundai | Tiburon | 168 | 304 | Nissan | Versa | 2420 |
| 49 | Cadillac | Seville | 101 | 177 | Hyundai | Tucson | 753 | 305 | Nissan | Versa Note | 517 |
| 50 | Cadillac | XTS | 108 | 178 | Hyundai | Veloster | 530 | 306 | Nissan | Xterra | 886 |
| 51 | Chevrolet | Avalanche | 1018 | 179 | Hyundai | Veracruz | 135 | 307 | Oldsmobile | Alero | 104 |
| 52 | Chevrolet | Aveo | 735 | 180 | Infiniti | EX | 124 | 308 | Pontiac | G5 | 122 |
| 53 | Chevrolet | Blazer | 198 | 181 | Infiniti | FX | 337 | 309 | Pontiac | G6 | 1572 |
| 54 | Chevrolet | Camaro | 1732 | 182 | Infiniti | FX35 | 261 | 310 | Pontiac | Grand Am | 337 |
| 55 | Chevrolet | Captiva Sport | 896 | 183 | Infiniti | G Convertible | 301 | 311 | Pontiac | Grand Prix | 868 |
| 56 | Chevrolet | Cavalier | 270 | 184 | Infiniti | G Coupe | 1503 | 312 | Pontiac | Torrent | 241 |
| 57 | Chevrolet | Cobalt | 1555 | 185 | Infiniti | G Sedan | 6938 | 313 | Pontiac | Vibe | 191 |
| 58 | Chevrolet | Colorado | 766 | 186 | Infiniti | G35 | 961 | 314 | Porsche | 911 | 22 |
| 59 | Chevrolet | Corvette | 338 | 187 | Infiniti | G37 | 192 | 315 | Porsche | Boxster | 140 |
| 60 | Chevrolet | Cruze | 5404 | 188 | Infiniti | JX | 273 | 316 | Porsche | Cayenne | 285 |
| 61 | Chevrolet | Equinox | 4335 | 189 | Infiniti | M | 801 | 317 | Porsche | Panamera | 1 |
| 62 | Chevrolet | Express | 264 | 190 | Infiniti | M35 | 222 | 318 | Ram | 1500 | 2703 |
| 63 | Chevrolet | Express Cargo | 982 | 191 | Infiniti | Q50 | 470 | 319 | Ram | 2500 | 438 |
| 64 | Chevrolet | HHR | 1040 | 192 | Infiniti | QX | 18 | 320 | Ram | 3500 | 104 |
| 65 | Chevrolet | Impala | 7491 | 193 | Infiniti | QX56 | 346 | 321 | Ram | Dakota | 32 |
| 66 | Chevrolet | Impala Limited | 1043 | 194 | Infiniti | QX60 | 93 | 322 | Saab | 9.Mar | 287 |
| 67 | Chevrolet | Malibu | 6172 | 195 | Jaguar | S-Type | 221 | 323 | Saturn | Aura | 426 |
| 68 | Chevrolet | Malibu Maxx | 159 | 196 | Jaguar | X-Type | 171 | 324 | Saturn | ION | 436 |
| 69 | Chevrolet | Monte Carlo | 344 | 197 | Jaguar | XF | 315 | 325 | Saturn | L-Series | 167 |
| 70 | Chevrolet | S-10 | 168 | 198 | Jaguar | XJ | 53 | 326 | Saturn | Outlook | 267 |
| 71 | Chevrolet | Silverado 1500 | 5521 | 199 | Jaguar | XJ-Series | 144 | 327 | Saturn | S-Series | 164 |
| 72 | Chevrolet | Silverado 1500 Classic | 171 | 200 | Jeep | Cherokee | 418 | 328 | Saturn | VUE | 781 |
| 73 | Chevrolet | Silverado 2500HD | 1012 | 201 | Jeep | Commander | 601 | 329 | Scion | FR-S | 159 |
| 74 | Chevrolet | Silverado 3500HD | 85 | 202 | Jeep | Compass | 1139 | 330 | Scion | iQ | 121 |
| 75 | Chevrolet | Sonic | 1602 | 203 | Jeep | Grand Cherokee | 3436 | 331 | Scion | tC | 641 |
| 76 | Chevrolet | Spark | 308 | 204 | Jeep | Liberty | 2171 | 332 | Scion | xB | 433 |
| 77 | Chevrolet | Suburban | 1659 | 205 | Jeep | Patriot | 1559 | 333 | Scion | xD | 170 |
| 78 | Chevrolet | Tahoe | 2257 | 206 | Jeep | Wrangler | 3031 | 334 | Subaru | Forester | 815 |
| 79 | Chevrolet | TrailBlazer | 1050 | 207 | Kia | Cadenza | 216 | 335 | Subaru | Impreza | 841 |
| 80 | Chevrolet | TrailBlazer EXT | 177 | 208 | Kia | Forte | 2470 | 336 | Subaru | Impreza WRX | 103 |
| 81 | Chevrolet | Traverse | 1758 | 209 | Kia | Optima | 3755 | 337 | Subaru | Legacy | 916 |
| 82 | Chevrolet | Uplander | 315 | 210 | Kia | Rio | 1305 | 338 | Subaru | Outback | 1207 |
| 83 | Chevrolet | Venture | 161 | 211 | Kia | Rondo | 147 | 339 | Subaru | XV Crosstrek | 177 |
| 84 | Chevrolet | Volt | 502 | 212 | Kia | Sedona | 833 | 340 | Suzuki | Forenza | 222 |
| 85 | Chrysler | 200 | 3714 | 213 | Kia | Sorento | 2914 | 341 | Suzuki | Grand Vitara | 179 |
| 86 | Chrysler | 300 | 3314 | 214 | Kia | Soul | 2437 | 342 | Suzuki | SX4 | 272 |
| 87 | Chrysler | Aspen | 207 | 215 | Kia | Spectra | 420 | 343 | Suzuki | XL7 | 134 |
| 88 | Chrysler | PT Cruiser | 1142 | 216 | Kia | Sportage | 1102 | 344 | Toyota | 4Runner | 1005 |
| 89 | Chrysler | Pacifica | 200 | 217 | Land Rover | LR2 | 100 | 345 | Toyota | Avalon | 852 |
| 90 | Chrysler | Sebring | 1214 | 218 | Land Rover | LR4 | 35 | 346 | Toyota | Camry | 10600 |
| 91 | Chrysler | Town and Country | 5053 | 219 | Land Rover | Range Rover | 154 | 347 | Toyota | Camry Hybrid | 569 |
| 92 | Dodge | Avenger | 3788 | 220 | Land Rover | Range Rover Evoque | 41 | 348 | Toyota | Camry Solara | 305 |
| 93 | Dodge | Caliber | 1249 | 221 | Land Rover | Range Rover Sport | 278 | 349 | Toyota | Corolla | 6583 |
| 94 | Dodge | Caravan | 189 | 222 | Lexus | CT 200h | 406 | 350 | Toyota | FJ Cruiser | 360 |
| 95 | Dodge | Challenger | 1054 | 223 | Lexus | ES 300 | 212 | 351 | Toyota | Highlander | 1656 |
| 96 | Dodge | Charger | 3983 | 224 | Lexus | ES 330 | 295 | 352 | Toyota | Highlander Hybrid | 190 |
| 97 | Dodge | Dakota | 499 | 225 | Lexus | ES 350 | 1683 | 353 | Toyota | Matrix | 290 |
| 98 | Dodge | Dart | 932 | 226 | Lexus | GS 300 | 159 | 354 | Toyota | Prius | 1754 |
| 99 | Dodge | Durango | 1611 | 227 | Lexus | GS 350 | 561 | 355 | Toyota | Prius c | 115 |
| 100 | Dodge | Grand Caravan | 6965 | 228 | Lexus | GX 460 | 68 | 356 | Toyota | RAV4 | 2980 |
| 101 | Dodge | Journey | 2693 | 229 | Lexus | GX 470 | 101 | 357 | Toyota | Sequoia | 381 |
| 102 | Dodge | Magnum | 348 | 230 | Lexus | IS 250 | 1775 | 358 | Toyota | Sienna | 2497 |
| 103 | Dodge | Neon | 209 | 231 | Lexus | IS 250 C | 109 | 359 | Toyota | Tacoma | 1064 |
| 104 | Dodge | Nitro | 585 | 232 | Lexus | IS 300 | 107 | 360 | Toyota | Tundra | 1014 |
| 105 | Dodge | Ram Pickup 1500 | 1831 | 233 | Lexus | IS 350 | 154 | 361 | Toyota | Venza | 468 |
| 106 | Dodge | Ram Pickup 2500 | 435 | 234 | Lexus | LS 430 | 345 | 362 | Toyota | Yaris | 1182 |
| 107 | Dodge | Ram Pickup 3500 | 159 | 235 | Lexus | LS 460 | 332 | 363 | Volkswagen | Beetle | 428 |
| 108 | Dodge | Sprinter Cargo | 103 | 236 | Lexus | RX 300 | 194 | 364 | Volkswagen | CC | 578 |
| 109 | Dodge | Stratus | 264 | 237 | Lexus | RX 330 | 389 | 365 | Volkswagen | Eos | 163 |
| 110 | FIAT | 500 | 484 | 238 | Lexus | RX 350 | 1843 | 366 | Volkswagen | GTI | 380 |
| 111 | FIAT | 500L | 238 | 239 | Lexus | RX 400h | 154 | 367 | Volkswagen | Golf | 268 |
| 112 | Ford | C-Max Energi | 230 | 240 | Lexus | RX 450h | 97 | 368 | Volkswagen | Jetta | 3951 |
| 113 | Ford | C-Max Hybrid | 303 | 241 | Lincoln | Aviator | 103 | 369 | Volkswagen | Jetta SportWagen | 171 |
| 114 | Ford | E-Series Van | 1107 | 242 | Lincoln | LS | 193 | 370 | Volkswagen | New Beetle | 665 |
| 115 | Ford | E-Series Wagon | 871 | 243 | Lincoln | MKS | 788 | 371 | Volkswagen | Passat | 2752 |
| 116 | Ford | Econoline Cargo | 468 | 244 | Lincoln | MKT | 139 | 372 | Volkswagen | Routan | 196 |
| 117 | Ford | Edge | 5098 | 245 | Lincoln | MKX | 941 | 373 | Volkswagen | Tiguan | 478 |
| 118 | Ford | Escape | 10652 | 246 | Lincoln | MKZ | 1505 | 374 | Volkswagen | Touareg | 261 |
| 119 | Ford | Escape Hybrid | 127 | 247 | Lincoln | Navigator | 552 | 375 | Volvo | C70 | 117 |
| 120 | Ford | Excursion | 107 | 248 | Lincoln | Town Car | 182 | 376 | Volvo | S40 | 343 |
| 121 | Ford | Expedition | 1898 | 249 | MINI | Cooper | 2016 | 377 | Volvo | S60 | 927 |
| 122 | Ford | Explorer | 5883 | 250 | MINI | Cooper Clubman | 337 | 378 | Volvo | S80 | 227 |
| 123 | Ford | Explorer Sport Trac | 283 | 251 | MINI | Cooper Countryman | 443 | 379 | Volvo | V70 | 115 |
| 124 | Ford | F-150 | 11120 | 252 | Mazda | CX-5 | 442 | 380 | Volvo | XC60 | 320 |
| 125 | Ford | F-250 Super Duty | 1844 | 253 | Mazda | CX-7 | 317 | 381 | Volvo | XC70 | 115 |
| 126 | Ford | F-350 Super Duty | 811 | 254 | Mazda | CX-9 | 638 | 382 | Volvo | XC90 | 600 |
| 127 | Ford | Fiesta | 2214 | 255 | Mazda | MPV | 111 | 383 | smart | fortwo | 343 |

# Model Evaluation

## Step 1: Data Processing

We have take below action on this part of the project:

* Encode all categorical variables. Only year, make, and model will be used as of now but encoded all of them in case we need later on. Label encoder used hence there are too many variables exist one-hot encoding may not be a good option.
* Set the features. We are targeting the predict sellingrpice which will be the Y value. Using year, make, and model categorical features those will be X
* Split the data train and test based on 80% - 20%

A screenshot of a computer program

Description automatically generated

## Step 2: Model Training and Evaluation

We have take following steps:

* Decision tree algorithm (DecisionTreeRegressor) evaulated. This is one of the good result algorithm and had the below model evaluation metrics:
  + MSE: 7337042.067300556
  + MAE: 1919.4265891344216
  + RMSE: 2708.6974853793763
  + R-Square: 0.8685518076078089
  + MAPE: 24.336220660333105
* K-Means algorithm (KMeans) evaluated and get the below results. By the way this is not a good option to select for this assignment that we understand
  + MSE: 55476844.996523894
  + MAE: 5984.487365244151
  + RMSE: 7448.2779886712
  + R-Square: 0.00609388258587007
  + MAPE: 127.2523078746145
* We also evaluated linear regression (LinearRegressor) as well
  + MSE: 32663756.36511696
  + MAE: 4506.224066639455
  + RMSE: 5715.221462473432
  + R-Square: 0.41480617235805917
  + MAPE: 57.16556456159733

As of now looks Decision Tree is the best algorithm for this regression prediction assignment hence it has the highest R-Square value which is 0.867

A screenshot of a computer program

Description automatically generated

A screenshot of a computer code

Description automatically generated

A screenshot of a computer program

Description automatically generated

## Step 3: Ensemble Method

* We used Random Forest Regressor here to use ensemble method top see how will impact the model evaluation results in our scores
  + MSE: 7340317.367543845
  + MAE: 1919.6639097005007
  + RMSE: 2709.3020074446936
  + R-Square: 0.8684931283345305
  + MAPE: 24.338146861735265
* Seems very similar R-Square value we have already with Decision tree algorithm here

A screenshot of a computer program

Description automatically generated

## Step 4: Cross Validation

We used CV for 2 algorithm (compare Decision Tree and Random Forest) including ensemble method, we see the below results. Results are almost same.

* Decision Tree Cross-Validation :
  + Mean RMSE: 2733.818554892202
  + Std RMSE: 70.19681518427721
* Random Forest Cross-Validation:
  + Mean RMSE: 2733.305758150658
  + Std RMSE: 69.79797573562276

A screenshot of a computer code

Description automatically generated

## Step 5: Compare Results

* We have the below table as a result.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **MSE** | **MAE** | **RMSE** | **R-Square** | **MAPE** |
| Decision Tree | 7.337042e+06 | 1919.426589 | 2708.697485 | 0.868552 | 24.336221 |
| K-Means | 5.547684e+07 | 5984.487365 | 7448.277989 | 0.006094 | 127.252308 |
| Linear Regression | 3.266376e+07 | 4506.224067 | 5715.221462 | 0.414806 | 57.165565 |
| Random Forest | 7.340317e+06 | 1919.663910 | 2709.302007 | 0.868493 | 24.338147 |

A screenshot of a computer code

Description automatically generated

## Step 6: Feature Engineering

* Lets add one more feature into our dataset and compare the model evaluation results if any impact occur. We added condition feature into prediction calculation.
* Condition has from 1 to 50 scale and divided into 5 categories:
  + Between 1.0 and 9.9 = Excellent
  + Between 10.0 and 19.9 = Above Average
  + Between 20.0 and 29.9 = Average
  + Between 30.0 and 39.9 = Below Average
  + Between 40.0 and 49.9 = Poor
* Total 6 steps applied during future engineering:
  + Step 1: Categorize the features like condition
  + Step 2: Encode categorical variables. I re-encoded again since I wanted to be the code can be work from here also
  + Step 3: Update features
  + Step 4: Train and evaluate models
  + Step 5: Cross Validation
  + Step 6: Compare results
* After feature extraction step, result as seen in the table:

A screenshot of a table

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

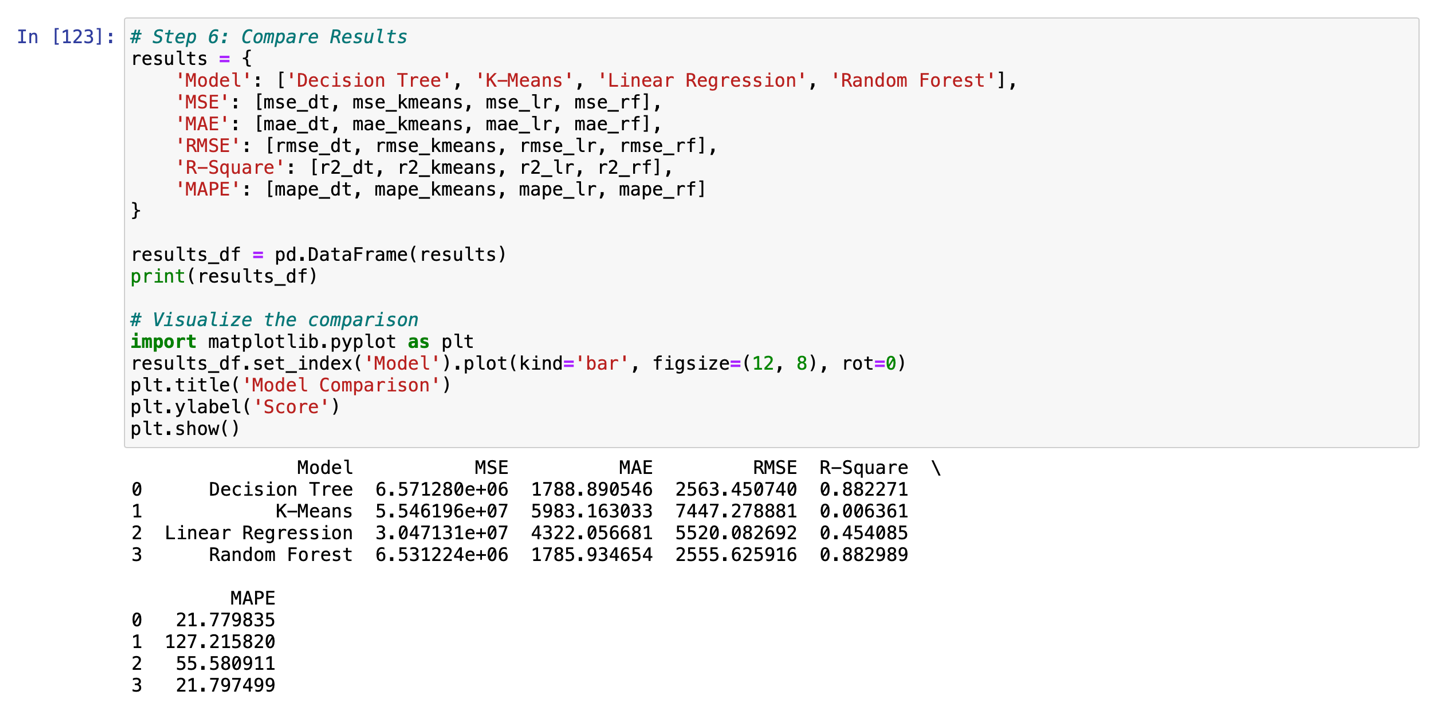
Description automatically generated

A screenshot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated



# Summary

* The Decision Tree model performs well with a high R-Square value (0.868) indicating that it explains 86.8% of the variance in the target variable. It has the lowest MAE and relatively low MAPE, suggesting it makes smaller errors on average.
* The K-Means model performs poorly in this regression task, which is expected since K-Means is primarily a clustering algorithm and not typically used for regression. It has the highest MSE, MAE, RMSE, and MAPE, and the R-Square value is close to zero, indicating it does not explain the variance in the target variable well.
* The Linear Regression model has moderate performance with an R-Square value of 0.415, explaining 41.5% of the variance in the target variable. The MSE, MAE, and RMSE values are higher than those of the Decision Tree and Random Forest, indicating larger prediction errors. The MAPE is also high at 79.72%, suggesting less reliable predictions.
* The Random Forest model performs similarly to the Decision Tree model with a high R-Square value (0.868), indicating that it explains 86.8% of the variance in the target variable. It has the lowest MAE and MAPE values, suggesting it makes small errors on average, comparable to the Decision Tree model.

Best Performers: Both the Decision Tree and Random Forest models perform the best with very similar metrics. They have high R-Square values, indicating good explanatory power, and low error metrics (MSE, MAE, RMSE, MAPE).

Moderate Performer: The Linear Regression model performs moderately with a lower R-Square value and higher error metrics compared to the Decision Tree and Random Forest models.

Poor Performer: The K-Means model performs poorly for this regression task with high error metrics and a very low R-Square value.

A graph of different values

Description automatically generated A graph of a graph with different colored bars

Description automatically generated with medium confidence

A graph of blue and orange bars

Description automatically generated A graph of blue and orange bars

Description automatically generated

Thank You