Milestone II Report
Predicting Used Car Prices using Machine Learning

CSCE 3602 - Fundamentals of Machine Learning

Omar Bahgat
Computer Engineering Department
The American University in Cairo
Cairo, Egypt
omar_bahgat@aucegypt.edu

Omar Saleh
Computer Engineering Department
The American University in Cairo
Cairo, Egypt
Omar_Anwar@aucegypt.edu

Dataset Description

The dataset chosen for our project, authored by Andrei Novikov, contains data on around 760,000 used cars which were scrapped from cars.com during April 2023.

With 20 distinct features, the dataset covers everything from basic car information like car manufacturer, model, and production year, to specifics like mileage, engine type, and transmission configuration. It also includes details like fuel type, MPG (miles per gallon), exterior and interior colors, accident history, and ownership status.

The main point of our analysis lies in the "Price" feature, serving as the label for our model. This feature signifies the listed price of each used car on cars.com which is the output our machine learning model will try to predict accurately.

This dataset can be found on Kaggle here [1].

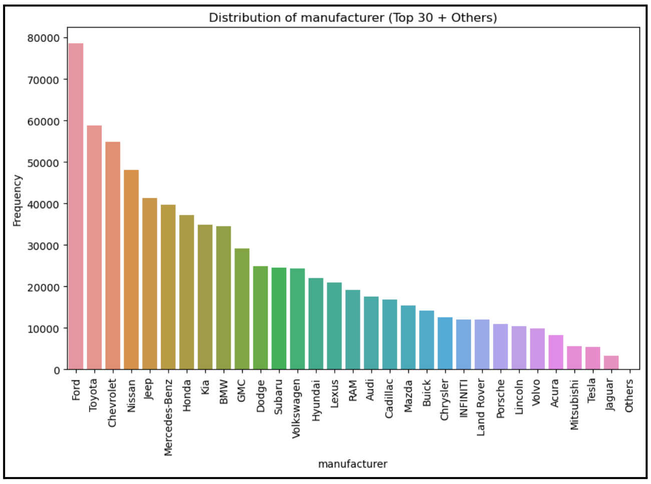
Feature Analysis

In our feature analysis, we will delve into the detailed description, cleaning, and preprocessing steps for each of the 20 features in our dataset. As a preliminary step, we pruned our dataset by dropping any rows with more than 25% of their values missing - around 14,000 rows - ensuring data integrity and quality.

[1] Manufacturer

The ‘manufacturer’ feature *[string]* represents the name of the car manufacturer.

Statistical Distribution:



Missing Values:

There are 0 missing values.

Correlation with Label:

-0.001679

Unique Values:

There are 3675 unique values.

Semantic Importance / Relevance:

The “Manufacturer” (brand) is semantically important because it encapsulates a wealth of information about a car. Reputation: Established brands have reputations for reliability, quality, and innovation. Brand Image: Each brand has a unique image (e.g., luxury, eco-friendly, affordable).

Cleaning and Preprocessing Steps:

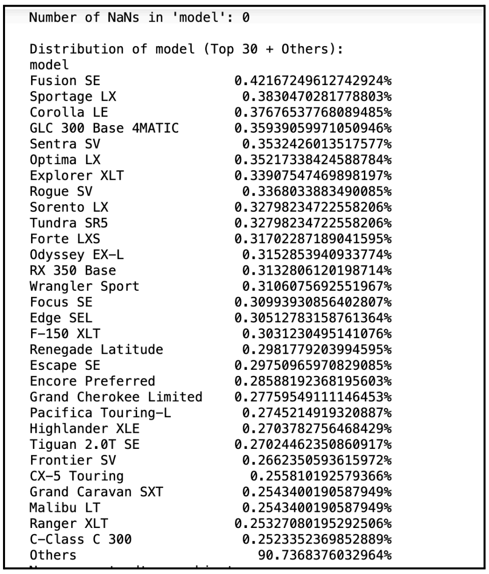
1. Encoding

Since the manufacturer is a categorical feature, we used one-hot-encoding to convert its values into a format suitable for machine learning algorithms.

[2] Model

The ‘model’ feature *[string]* represents the name of the car model.

Statistical Distribution:



```
Number of NaNs in 'model': 0

Distribution of model (Top 30 + Others):
model
Fusion SE                 0.42167249612742924%
Sportage LX                0.3830470281778803%
Corolla LE                0.37676537768089485%
GLC 300 Base 4MATIC       0.35939059971050946%
Sentra SV                  0.3532426013517577%
Optima LX                 0.35217338424588784%
Explorer XLT              0.33907547469898197%
Rogue SV                   0.3368033883490085%
Sorento LX                0.32798234722558206%
Tundra SR5                0.32798234722558206%
Forte LXS                 0.31702287189041595%
Odyssey EX-L               0.3152853940933774%
RX 350 Base                0.3132806120198714%
Wrangler Sport             0.3106075692551967%
Focus SE                  0.30993930856402807%
Edge SEL                  0.30512783158761364%
F-150 XLT                  0.3031230495141076%
Renegade Latitude          0.2981779203994595%
Escape SE                 0.29750965970829085%
Encore Preferred          0.28588192368195603%
Grand Cherokee Limited    0.27759549111146453%
Pacifica Touring-L         0.2745214919320887%
Highlander XLE             0.2703782756468429%
Tiguan 2.0T SE            0.27024462350860917%
Frontier SV                0.2662350593615972%
CX-5 Touring                0.255810192579366%
Grand Caravan SXT          0.2543400190587949%
Malibu LT                  0.2543400190587949%
Ranger XLT                0.25327080195292506%
C-Class C 300              0.2523352369852889%
Others                       90.7368376032964%
```

Missing Values:

There are 0 missing values.

Correlation with Label:

-0.001878

Unique Values:

There are 3675 unique values.

Semantic Importance / Relevance:

The model is important because each model has unique features, specifications, and characteristics which can significantly affect the price

Cleaning and Preprocessing Steps:

1. Categorizing the Data

Simplify Rare Models: If a car model name appears very infrequently (less than 25 times), the code simplifies it by keeping only the first word. This might be because rare models might not be informative for the model and could even confuse it.

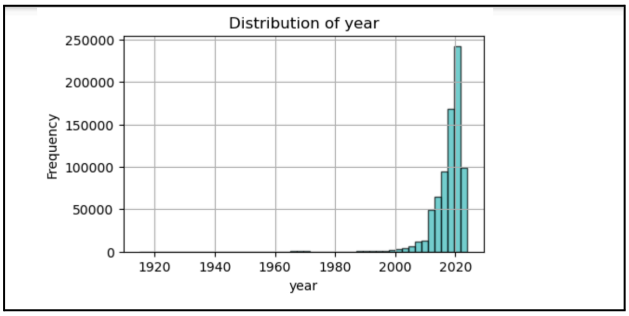
2. Encoding

Since the model name is a categorical feature, we used one-hot-encoding to convert its values into a format suitable for machine learning algorithms.

[3] Year

The ‘year’ feature *[int]* represents the year when the car was produced.

Statistical Distribution:



Missing Values:

There are 0 missing values.

Correlation with Label:

0.002154

Unique Values:

There are 98 unique values.

Semantic Importance / Relevance:

The year model plays a crucial role in determining a car's market value, as newer models generally have higher prices due to advancements in technology while older models may depreciate in value over time.

Cleaning and Preprocessing Steps:

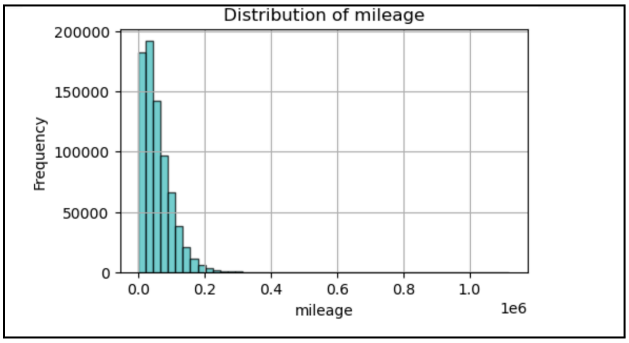
1. Encoding

Since the year model can be considered as a categorical feature, we used one-hot-encoding to convert its values into a format suitable for machine learning algorithms.

[4] Mileage

The ‘mileage’ feature *[float]* represents the number of miles the car has traveled since production.

Statistical Distribution:



Missing Values:

There are 485 missing values which constitute approximately 0.06 % of the feature values.

Correlation with Label:

-0.003170

Unique Values:

It’s a continuous feature so not applicable.

Semantic Importance / Relevance:

Mileage serves as a key indicator of a vehicle's history and is always considered in determining its value and condition.

Cleaning and Preprocessing Steps:

1. Handling Missing Values

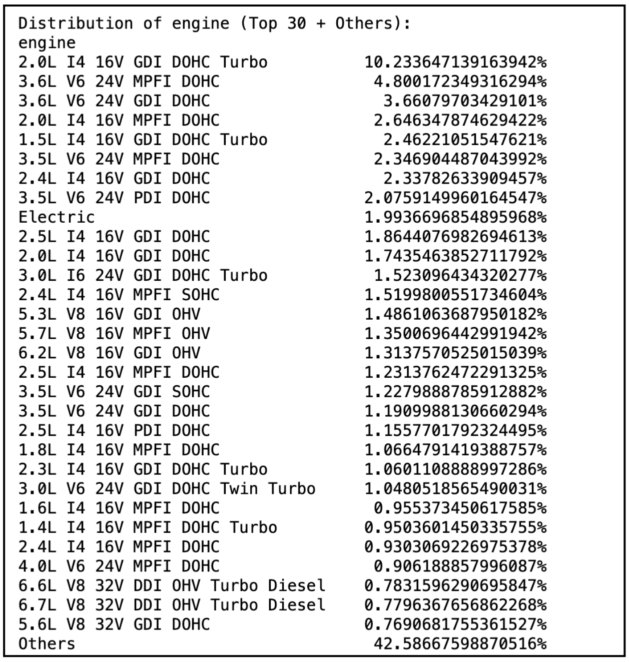
The missing values were handled by replacing them with the median value of the non-missing data. This approach was chosen due to the heavily skewed distribution *(skewness = 1.45)*, as the median is less sensitive to outliers compared to the mean.

2. Normalization

To ensure consistent scaling and comparability across features, the values were scaled using z-score normalization.

[5] Engine

Statistical Distribution:



```
Distribution of engine (Top 30 + Others):
engine
2.0L I4 16V GDI DOHC Turbo          10.233647139163942%
3.6L V6 24V MPFI DOHC                4.800172349316294%
3.6L V6 24V GDI DOHC                  3.66079703429101%
2.0L I4 16V MPFI DOHC                2.646347874629422%
1.5L I4 16V GDI DOHC Turbo            2.46221051547621%
3.5L V6 24V MPFI DOHC                2.346904487043992%
2.4L I4 16V GDI DOHC                  2.33782633909457%
3.5L V6 24V PDI DOHC                2.0759149960164547%
Electric                            1.9936696854895968%
2.5L I4 16V GDI DOHC                1.8644076982694613%
2.0L I4 16V GDI DOHC                1.7435463852711792%
3.0L I6 24V GDI DOHC Turbo           1.523096434320277%
2.4L I4 16V MPFI SOHC               1.5199800551734604%
5.3L V8 16V GDI OHV                 1.4861063687950182%
5.7L V8 16V MPFI OHV                1.3500696442991942%
6.2L V8 16V GDI OHV                 1.3137570525015039%
2.5L I4 16V MPFI DOHC               1.2313762472291325%
3.5L V6 24V GDI SOHC                1.2279888785912882%
3.5L V6 24V GDI DOHC                1.1909988130660294%
2.5L I4 16V PDI DOHC                1.1557701792324495%
1.8L I4 16V MPFI DOHC               1.0664791419388757%
2.3L I4 16V GDI DOHC Turbo          1.0601108888997286%
3.0L V6 24V GDI DOHC Twin Turbo     1.0480518565490031%
1.6L I4 16V MPFI DOHC                0.955373450617585%
1.4L I4 16V MPFI DOHC Turbo         0.9503601450335755%
2.4L I4 16V MPFI DOHC               0.9303069226975378%
4.0L V6 24V MPFI DOHC                0.906188857996087%
6.6L V8 32V DDI OHV Turbo Diesel    0.7831596290695847%
6.7L V8 32V DDI OHV Turbo Diesel    0.7796367656862268%
5.6L V8 32V GDI DOHC                0.7690681755361527%
Others                               42.58667598870516%
```

Missing Values:

There are 10175 missing values which constitute approximately 1.36 % of the feature values.

Correlation with Label:

0.003279

Unique Values:

There are 1536 unique values.

Semantic Importance / Relevance:

The engine specifications play a pivotal role in determining the price of a used car, with factors such as performance and reliability directly impacting its resale value.

Cleaning and Preprocessing Steps:

1. Categorizing the Data

Simplify Rare Engines: If a car Engine name appears very infrequently (less than 20 times), the code simplifies it by keeping only the first word.

[6] Transmission

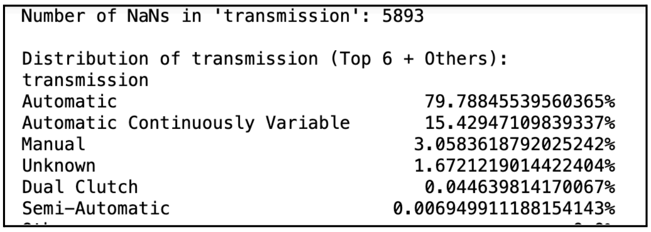
It will be split into two columns:

1. Transmission type

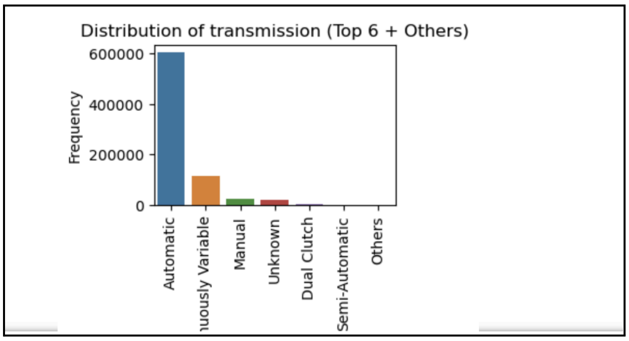
2. Number of speeds

Statistical Distribution:

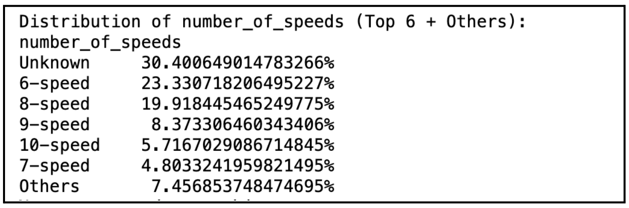
Transmission type



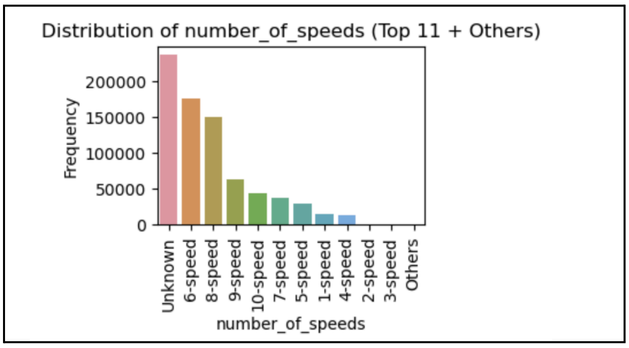
```
Number of NaNs in 'transmission': 5893

Distribution of transmission (Top 6 + Others):
transmission
Automatic                          79.78845539560365%
Automatic Continuously Variable    15.42947109839337%
Manual                             3.0583618792025242%
Unknown                            1.6721219014422404%
Dual Clutch                         0.044639814170067%
Semi-Automatic                    0.006949911188154143%
```



Number of Speeds



```
Distribution of number_of_speeds (Top 6 + Others):
number_of_speeds
Unknown      30.400649014783266%
6-speed      23.330718206495227%
8-speed      19.918445465249775%
9-speed       8.373306460343406%
10-speed     5.7167029086714845%
7-speed      4.8033241959821495%
Others        7.456853748474695%
```



Missing Values:

For Transmission type

There are 12,511 missing values. which constitute approximately 1.672% of the feature values.

For Number of Speeds

There are 227,461 missing values. which constitute approximately 30.4% of the feature values.

Correlation with Label:

-0.001664

Unique Values:

For Transmission type

There are 6 unique values post-cleaning

For Number of Speeds

There are 11 unique values post-cleaning

Semantic Importance / Relevance:

The type of transmission is an important factor influencing the price of a used car, as it directly affects driving experience, maintenance costs, and market demand, thereby impacting its resale value.

Cleaning and Preprocessing Steps:

1. Categorizing the Data

It extracts more specific details from the raw "transmission" data. Identifying the transmission type (e.g., "Automatic", "Manual", "Continuously Variable"). Extracting the number of speeds (e.g., "6-speed", "8-speed") if available. Handles cases where this information can't be confidently parsed, labeling them as "Unknown".

The extracted transmission type and number of speeds are stored as separate new columns in the dataset, creating distinct features for modeling.

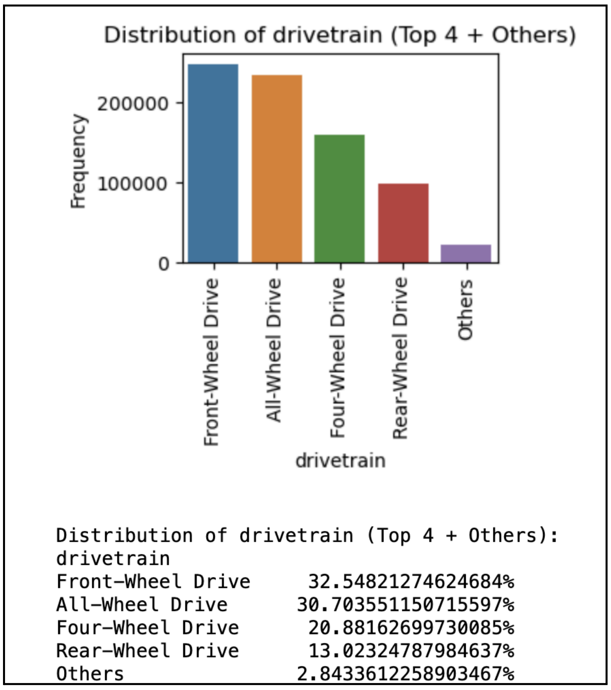
2. Encoding

Since the transmission type is a categorical feature, we used one-hot-encoding to convert its values into a format suitable for machine learning algorithms.

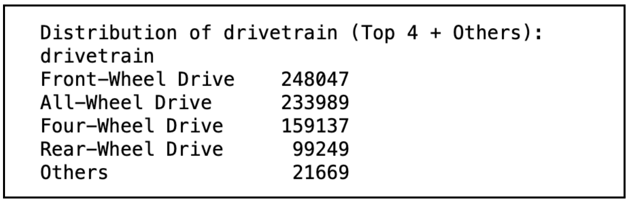
[7] Drivetrain

The ‘drivetrain’ feature *[string]* represents the car’s drivetrain.

Statistical Distribution:



```
Distribution of drivetrain (Top 4 + Others):
drivetrain
Front-Wheel Drive      32.54821274624684%
All-Wheel Drive       30.703551150715597%
Four-Wheel Drive       20.88162699730085%
Rear-Wheel Drive       13.02324787984637%
Others                2.8433612258903467%
```



```
Distribution of drivetrain (Top 4 + Others):
drivetrain
Front-Wheel Drive    248047
All-Wheel Drive      233989
Four-Wheel Drive     159137
Rear-Wheel Drive      99249
Others                21669
```

Missing Values:

There are 16,634 missing values which constitute approximately 2.22 % of the feature values.

Correlation with Label:

-0.002613

Unique Values:

There are 5 unique values post cleaning.

Semantic Importance / Relevance:

Drivetrain affects the selling price of a used car by influencing its performance, durability, and appeal to different buyers.

Cleaning and Preprocessing Steps:

1. Categorizing the Data

We identified the top 4 most common drivetrains and classified them accordingly. We then matched cars based on their drivetrains and assigned them to one of these categories.

2. Handling Missing and Irrelevant Values

The missing values were grouped into an 'other' category to ensure data integrity while accommodating their presence. Additionally, values with negligible occurrence rates - 105 values - were considered insignificant and also included in the 'other' category.

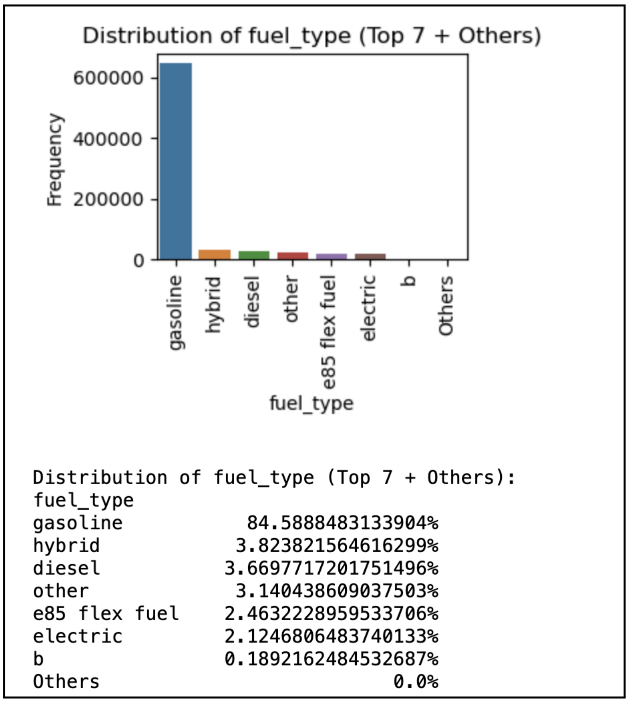
3. Encoding

Since drivetrain is a categorical feature, we used one-hot-encoding to convert its values into a format suitable for machine learning algorithms.

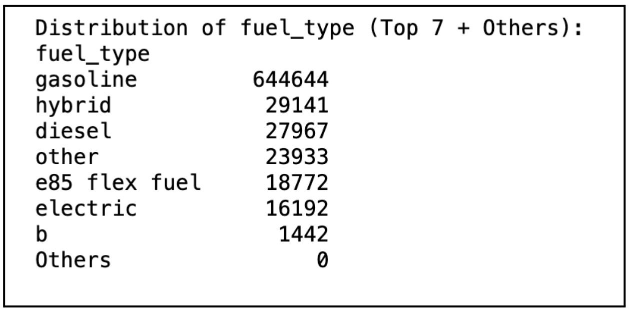
[8] Fuel Type

The ‘fuel_type’ feature *[string]* represents the type of fuel the car consumes.

Statistical Distribution:



```
Distribution of fuel_type (Top 7 + Others):
fuel_type
gasoline          84.5888483133904%
hybrid           3.823821564616299%
diesel          3.6697717201751496%
other            3.140438609037503%
e85 flex fuel   2.4632228959533706%
electric        2.1246806483740133%
b               0.1892162484532687%
Others                        0.0%
```



```
Distribution of fuel_type (Top 7 + Others):
fuel_type
gasoline         644644
hybrid            29141
diesel            27967
other             23933
e85 flex fuel     18772
electric          16192
b                  1442
Others                0
```

Missing Values:

There are 17,983 missing values which constitute approximately 2.4 % of the feature values.

Correlation with Label:

-0.000949

Unique Values:

There are 7 unique values post cleaning.

Semantic Importance / Relevance:

The type of fuel a car consumes directly affects its operational costs thereby influencing its desirability and, consequently, its selling price in the used car market.

Cleaning and Preprocessing Steps:

1. Categorizing the Data

We identified the top 6 most common fuel types and classified them accordingly. We then matched cars based on their fuel type and assigned them to one of these categories.

2. Handling Missing and Irrelevant Values

Missing values were grouped into an 'other' category to ensure data integrity while accommodating their presence. Additionally, values with negligible occurrence rates - 839 values - were considered insignificant and also included in the 'other' category.

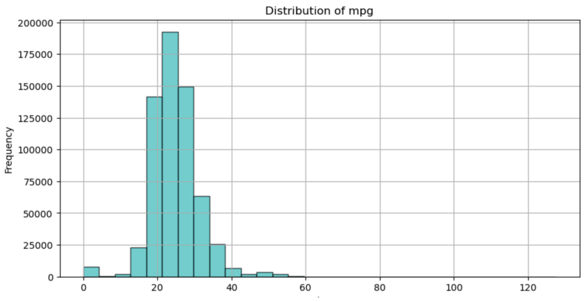
3. Encoding

Since fuel type is a categorical feature, we used one-hot-encoding to convert its values into a format suitable for machine learning algorithms.

[9] Miles Per Gallon (MPG)

The ‘mpg’ feature *[float]* represents the number of miles a car can travel using one gallon of fuel.

Statistical Distribution:



Missing Values:

There are 128,830 missing values which constitute approximately 17.16 % of the feature values.

Correlation with Label:

-0.004335

Unique Values:

We consider MPG to be a non-categorical feature as it is a float so ‘unique values’ is not applicable.

Semantic Importance / Relevance:

The miles per gallon (mpg) metric provides crucial insight into a car's fuel efficiency, which affects its running costs and shapes its selling price.

Cleaning and Preprocessing Steps:

1. Handling Missing Values

The missing values were handled by replacing them with the mean value of the non-missing data. This approach was chosen since the data distribution was only slightly skewed *(skewness = 0.46).*

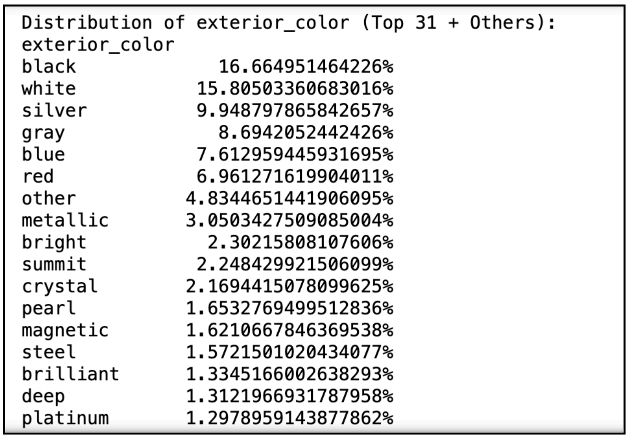
2. Normalization

To ensure consistent scaling and comparability across features, the values were scaled using z-score normalization.

[10] Exterior Color

The ‘exterior_color’ feature *[string]* represents the car’s exterior color.

Statistical Distribution:



```
Distribution of exterior_color (Top 31 + Others):
exterior_color
black            16.664951464226%
white           15.80503360683016%
silver           9.948797865842657%
gray              8.6942052442426%
blue             7.612959445931695%
red              6.961271619904011%
other           4.8344651441906095%
metallic        3.0503427509085004%
bright            2.30215808107606%
summit           2.248429921506099%
crystal         2.1694415078099625%
pearl           1.6532769499512836%
magnetic        1.6210667846369538%
steel           1.5721501020434077%
brilliant       1.3345166002638293%
deep            1.3121966931787958%
platinum        1.2978959143877862%
```

Missing Values:

There are 8,638 missing values which constitute approximately 1.15 % of the feature values.

Correlation with Label:

-0.002184

Unique Values:

There are 31 unique values post cleaning.

Semantic Importance / Relevance:

The choice of exterior color influences a car's marketability and resale price, reflecting consumer preferences and impacting its overall desirability.

Cleaning and Preprocessing Steps:

1. Categorizing the Data

We identified the top 30 most common exterior colors and classified them accordingly. We then matched cars based on their exterior colors and assigned them to one of these categories.

2. Handling Missing and Irrelevant Values

The missing values were grouped into an 'other' category to ensure data integrity while accommodating their presence. Additionally, exterior colors with negligible occurrence rates - 27534 values - were considered insignificant and also included in the 'other' category.

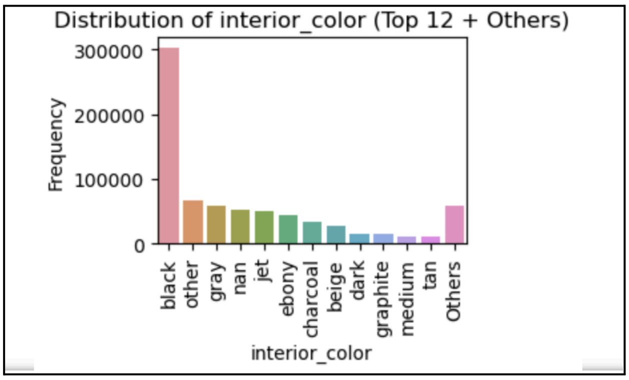
3. Encoding

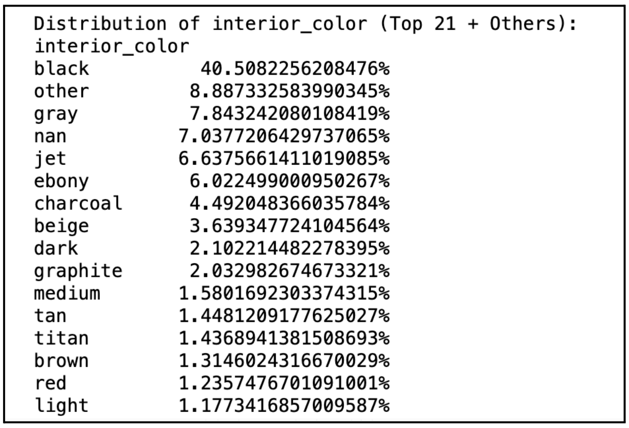
Since exterior color is a categorical feature, we used one-hot-encoding to convert its values into a format suitable for machine learning algorithms.

[11] Interior Color

The ‘interior_color’ feature *[string]* represents the car’s interior color.

Statistical Distribution:





```
Distribution of interior_color (Top 21 + Others):
interior_color
black            40.5082256208476%
other           8.887332583990345%
gray            7.843242080108419%
nan            7.0377206429737065%
jet            6.6375661411019085%
ebony           6.022499000950267%
charcoal        4.492048366035784%
beige           3.639347724104564%
dark            2.102214482278395%
graphite        2.032982674673321%
medium         1.5801692303374315%
tan            1.4481209177625027%
titan          1.4368941381508693%
brown          1.3146024316670029%
red            1.2357476701091001%
light          1.1773416857009587%
```

Missing Values:

There are 52,657 missing values which constitute approximately 7.03 % of the feature values.

Correlation with Label:

-0.000497

Unique Values:

There are 21 unique values post cleaning.

Semantic Importance / Relevance:

The specific hue of a car's interior affects its resale value, catering to individual tastes and preferences of potential buyers.

Cleaning and Preprocessing Steps:

1. Categorizing the Data

We identified the top 20 most common interior colors and classified them accordingly. We then matched cars based on their interior colors and assigned them to one of these categories.

2. Handling Missing and Irrelevant Values

The missing values were grouped into an 'other' category to ensure data integrity while accommodating their presence. Additionally, interior colors with negligible occurrence rates - 13839 values - were considered insignificant and also included in the 'other' category.

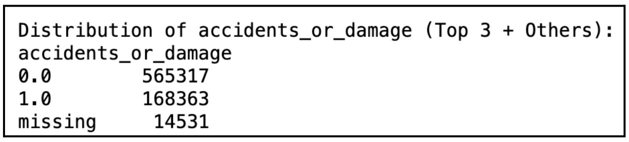
3. Encoding

Since interior color is a categorical feature, we used one-hot-encoding to convert its values into a format suitable for machine learning algorithms.

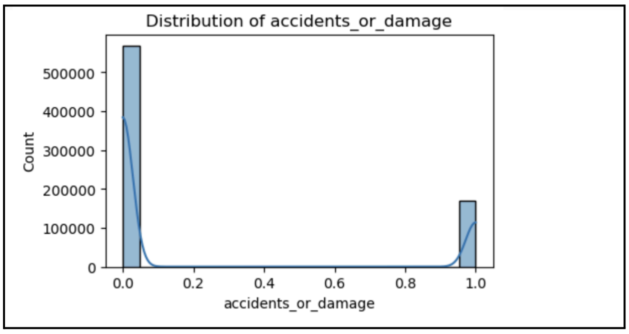
[12] Accidents or Damage

The 'accidents_or_damage' feature *[binary]* represents whether the car was involved in accidents or was damaged before or not.

Statistical Distribution:



```
Distribution of accidents_or_damage (Top 3 + Others):
accidents_or_damage
0.0        565317
1.0        168363
missing     14531
```



Missing Values:

There are 14,531 missing values which constitute approximately 1.94 % of the feature values.

Correlation with Label:

-0.003209

Unique Values:

There are 3 unique values post cleaning.

Semantic Importance / Relevance:

The presence or absence of an accident history can significantly impact the final selling price, as cars with a clean accident history generally have higher prices due to perceived lower risk and better condition.

Cleaning and Preprocessing Steps:

1. Handling Missing Values

The missing values were grouped into a third separate category to ensure data integrity while accommodating their presence.

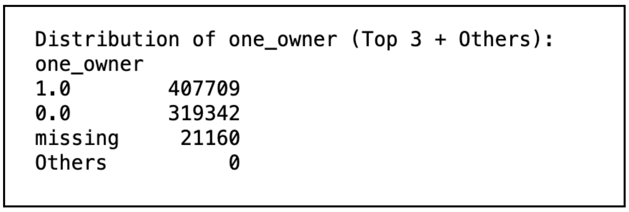
2. Encoding

The feature has been one-hot encoded into three columns: accidents (1), no_accidents (0), and missing_values.

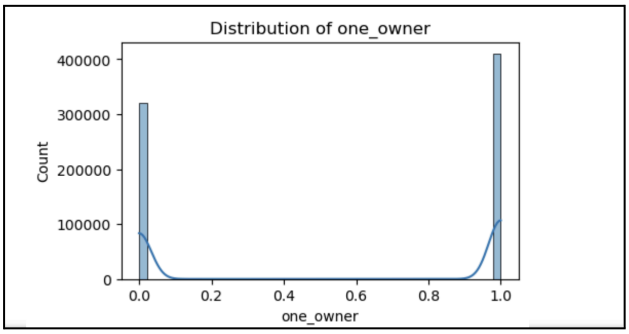
[13] One Owner

The 'one_owner' feature *[binary]* represents whether the car was owned by one person or not.

Statistical Distribution:



```
Distribution of one_owner (Top 3 + Others):
one_owner
1.0        407709
0.0        319342
missing     21160
Others          0
```



Missing Values:

There are 21,160 missing values which constitute approximately 2.83 % of the feature values.

Correlation with Label:

0.004003

Unique Values:

There are 3 unique values post cleaning.

Semantic Importance / Relevance:

Determining whether a car had only one previous owner or not may influence the final selling price, as cars with a single owner often have more consistent maintenance records, potentially leading to higher resale values.

Cleaning and Preprocessing Steps:

1. Handling Missing Values

The missing values were grouped into a third separate category to ensure data integrity while accommodating their presence.

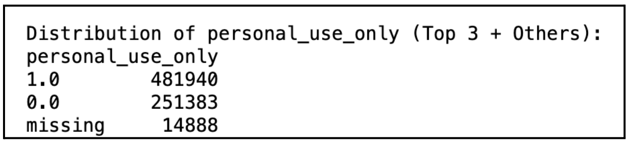
2. Encoding

The feature has been one-hot encoded into three columns: one_owner (1), several_owners (0), and missing_values.

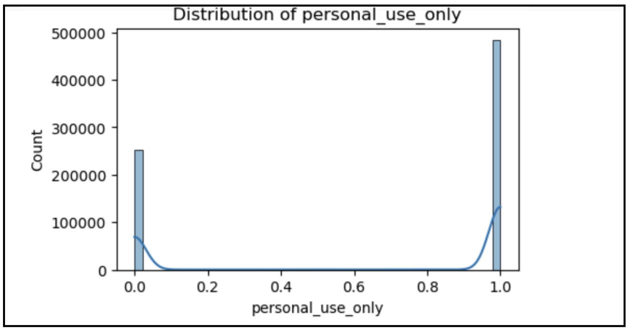
[14] Personal Use Only

The 'personal_use_only' feature *[binary]* represents whether the car was used for personal purposes only or not.

Statistical Distribution:



```
Distribution of personal_use_only (Top 3 + Others):
personal_use_only
1.0         481940
0.0         251383
missing      14888
```



Missing Values:

There are 14,888 missing values which constitute approximately 1.99 % of the feature values.

Correlation with Label:

0.001641

Unique Values:

There are 3 unique values post-cleaning.

Semantic Importance / Relevance:

A car used solely for personal purposes or not can affect the final selling price, as vehicles primarily used for personal use may have lower mileage, less wear, and a better overall condition, all of which can contribute to a higher resale value compared to cars used for commercial purposes

Cleaning and Preprocessing Steps:

1. Handling Missing Values

The missing values were grouped into a third separate category to ensure data integrity while accommodating their presence.

2. Encoding

The feature has been one-hot encoded into three columns: personal_use (1), not_personal_use (0), and missing_values.

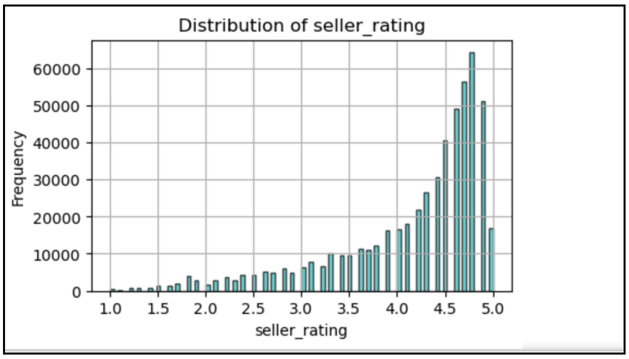
[15] Seller Name

The 'seller_name' feature *[string]* represents the car's seller name. However, we opted to remove the feature from the dataset due to several factors: a high proportion of missing values (28%), a large number of unique values (18k), and redundancy given the availability of seller ratings. With seller ratings already present, including the seller's name would be unnecessary and redundant for our analysis.

[16] Seller Rating

The 'seller_rating' feature *[float]* represents the car's seller rating.

Statistical Distribution:



Missing Values:

There are 211007 missing values which constitute approximately 27.89 % of the feature values.

Correlation with Label:

0.002311

Unique Values:

It’s a continuous feature so not applicable.

Semantic Importance / Relevance:

Cleaning and Preprocessing Steps:

1. Handling Missing Values

The missing values were handled by replacing them with zero Replacing missing values with 0 assumes that a missing rating signifies a neutral evaluation.

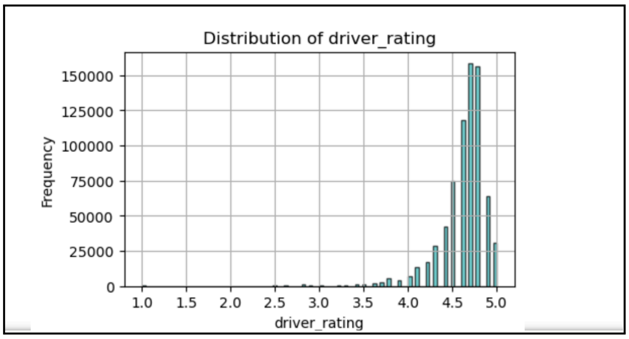
2. Normalization

To ensure consistent scaling and comparability across features, the values were scaled using z-score normalization.

[17] Driver Rating

The ‘driver_rating’ feature *[float]* represents the car rating given by drivers.

Statistical Distribution:



Missing Values:

There are 30,335 missing values which constitute approximately 4.05% of the feature values.

Correlation with Label:

0.000251

Unique Values:

It’s a continuous feature so not applicable.

Semantic Importance / Relevance:

The driver rating, reflecting the satisfaction level of car owners, is crucial for setting the price of a used car, directly impacting its perceived value and buyer confidence in the market.

Cleaning and Preprocessing Steps:

1. Handling Missing Values

The missing values were handled by replacing them with the median value of the non-missing data. This approach was chosen due to the heavily skewed distribution *(skewness = -2.82)*, as the median is less sensitive to outliers compared to the mean.

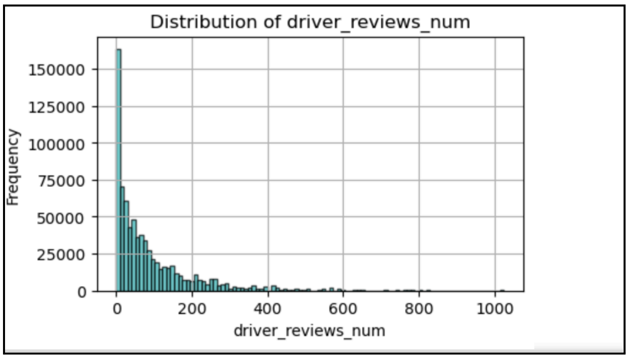
2. Normalization

To ensure consistent scaling and comparability across features, the values were scaled using z-score normalization.

[18] Driver Reviews Num

The ‘driver_review_num’ feature *[int]* represents the number of car reviews left by drivers

Statistical Distribution:



Missing Values:

There are 0 missing values.

Correlation with Label:

-0.002575

Unique Values:

This feature could be considered a categorical one; however, since the number of unique values is large, we will consider it as a continuous feature and will perform z-scaling. Therefore, 'unique values' is not applicable.

Semantic Importance / Relevance:

The number of driver reviews is important for setting the price of a used car, showing how satisfied buyers are and affecting their confidence in the purchase.

Cleaning and Preprocessing Steps:

1. Normalization

To ensure consistent scaling and comparability across features, the values were scaled using z-score normalization.

[19] Price Drop

The 'price_drop' feature *[float]* represents the car's price reduction from the initial price. However, we opted to remove the feature from the dataset due to its high 46% missing data rate. This feature's limited information and prevalence of missing values make it less valuable for analysis compared to other variables. Removing it would simplify our dataset, ensuring more reliable results.

[20] Price

The 'price' feature *[float]* is our label.

Note: All measurements (mean, median, std, unique, skew, etc..) are shown and printed in the code part.

Post Cleaning Metrics

Final list of chosen features:

1. Manufacturer
2. Model
3. Year
4. Mileage
5. Engine
6. Transmission
7. Drivetrain
8. Fuel Type
9. MPG
10. Exterior Color
11. Interior Color
12. Accidents or Damage
13. One owner
14. Personal use only
15. Seller rating
16. Driver rating
17. Driver reviews num

The dataset now has 748,211 rows and 5562 columns.

References

[1] Novikov, Andrei. 2023. Used Cars Dataset. Kaggle. Available at: https://www.kaggle.com/datasets/andreinovikov/used-cars-dataset