What Drives the Price of a Car ?

Github repository https://github.com/devasidgmail/car_prise.git

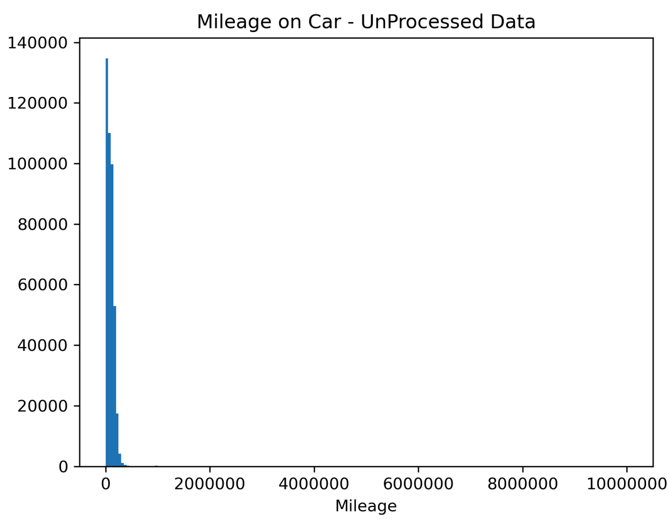
Sidd Devalapalli

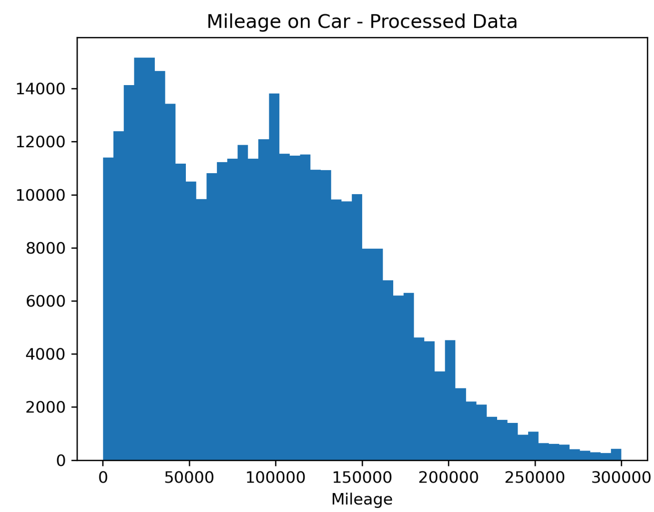
Overview

The aim of this project is to determine what factors make the price of a used car more or less expensive and which features are most important in determining the price of the car.

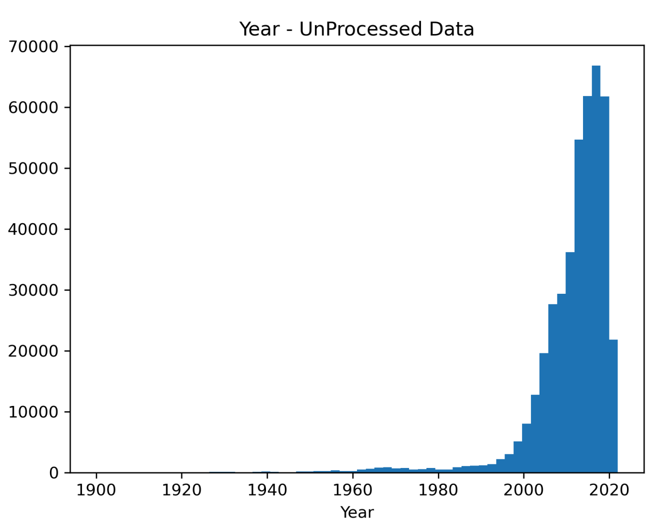
Understanding and cleaning up the data.

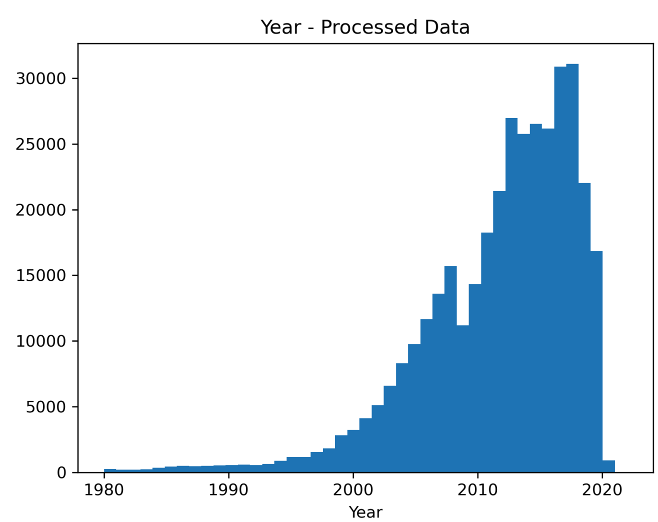
In order to understand the data I plotted histograms of the data and noticed that there was a **large variation in price, age and mileage on the cars. These outliers were removed.**



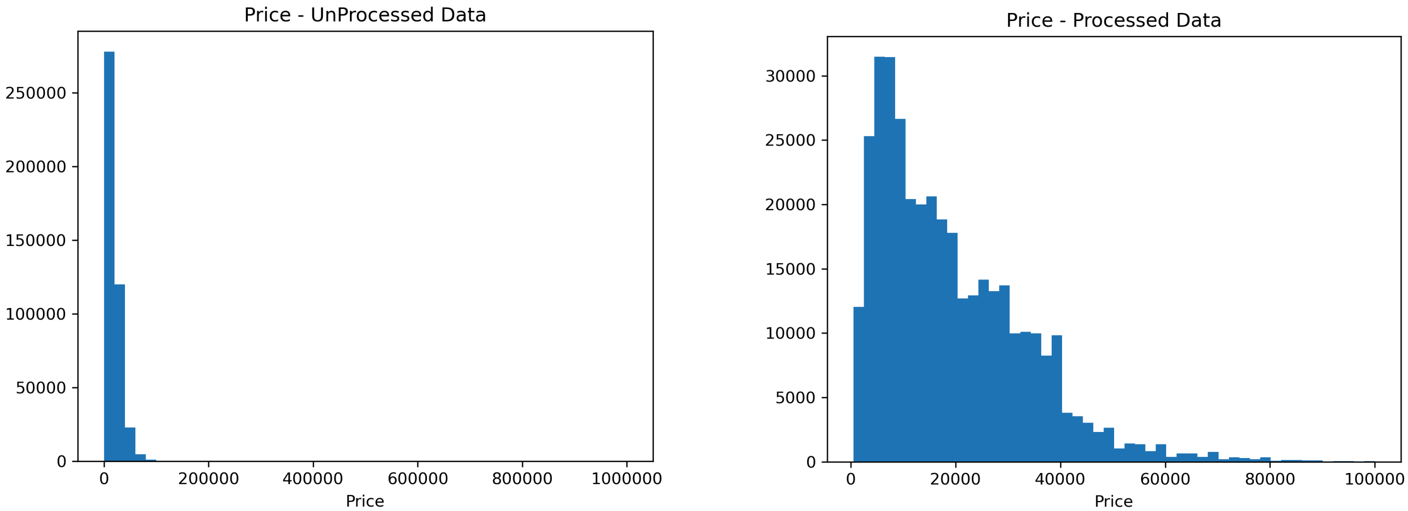


Picture - Mileage on Car : UnProcessed and Processed data.





Picture - Year of Car : UnProcessed and Processed Data



Picture - Price of the car : UnProcessed and Processed Data

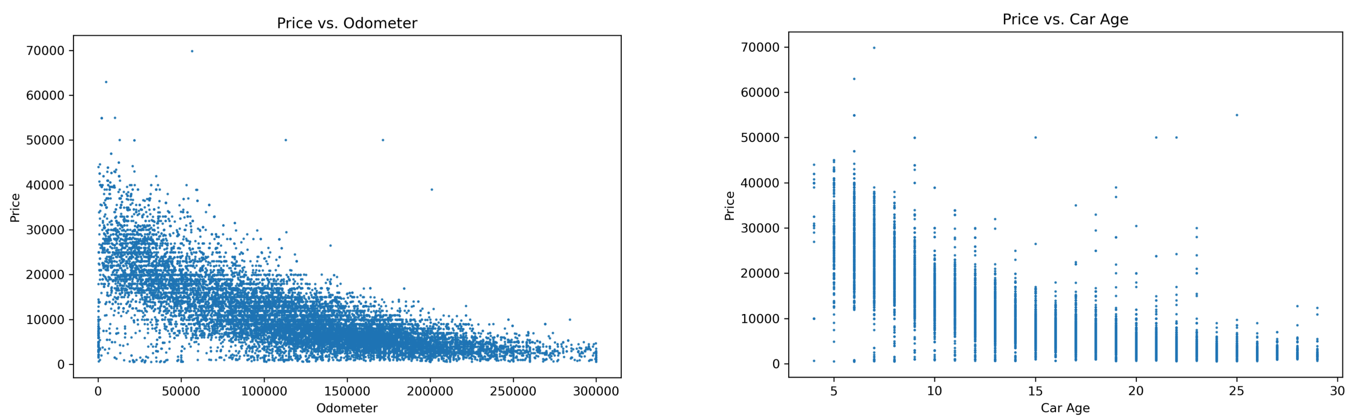
Add new Feature : Car_Age

Instead of working with the "Year" of the car, I added a **new feature "Car_Age".**

Car_Age = 2025 - "Year"

Understand linearity

I also plotted scatter plots to understand the linearity between the price of the car and age and mileage on the car.



Picture - Scatter plots of Odometer and Age of car vs Price of the car

I observed some linearity between the price of the car and both mileage and age of the car.

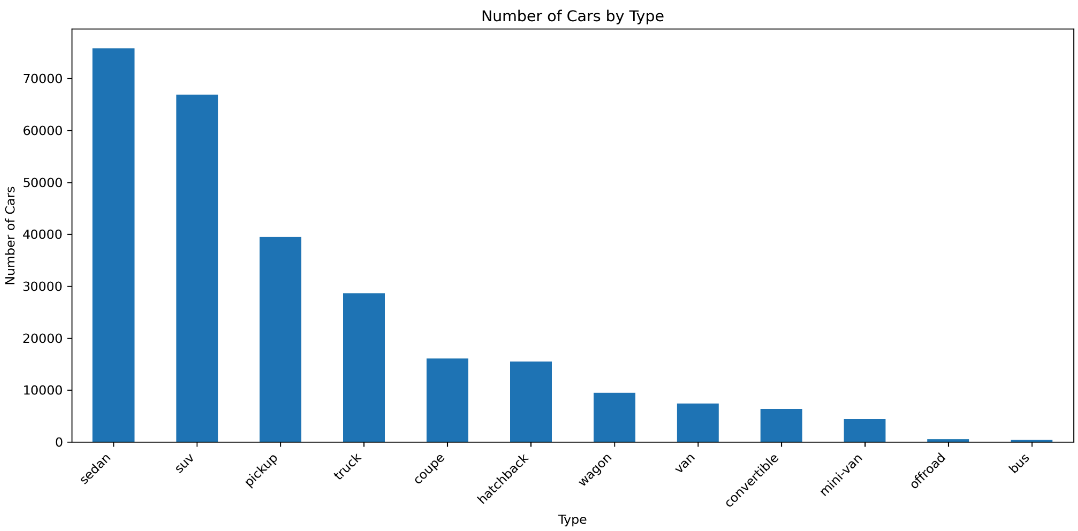
Computational limitations

Due to the computational power limitations on my home PC, a lot of features , like **"vin", "id" were dropped** . Also,some lesser relevant features like **cylinders, fuel, transmission were also dropped.**

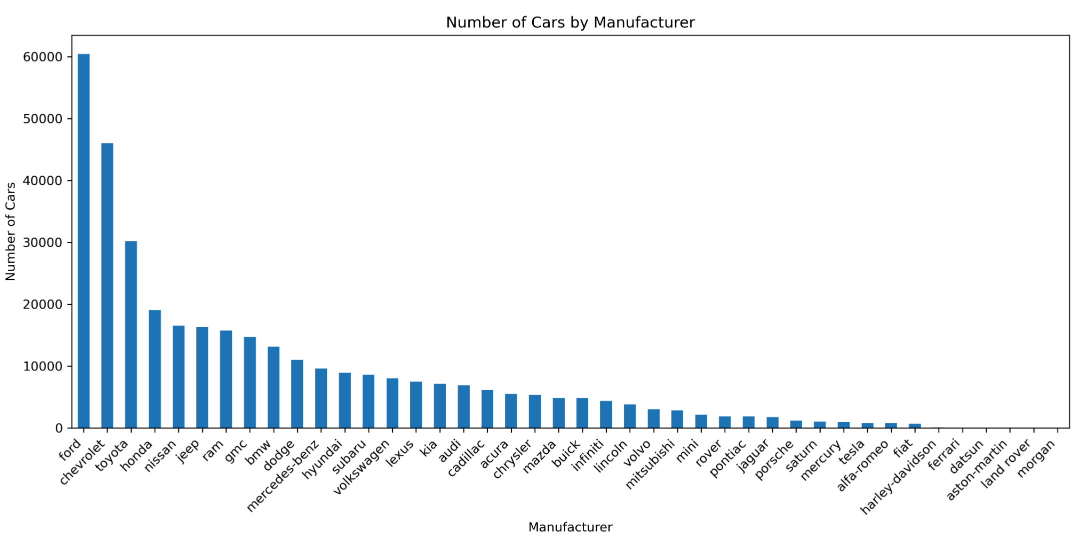
After cleanup a new csv file was created and saved which is named "df_clean.csv" in the data folder.

More Analysis

- Sedans , SUVs , pickup and trucks are the most sold vehicle types.
- Ford, Chevrolet , Toyota and Honda are the most sold vehicle manufacturers.



Picture : Number of cars sold by type.



Picture : Number of cars sold by manufacturers

Regressions and methods used.

Dummies were created on the categorical data and the data was split into training and test in the ratio 8 : 2

1. **Linear Regression**

   The first method employed was LinearRegression. The top 5 features in driving the cost of a car according to this method and their coefficient value are .

Feature	Coefficient
manufacturer_ferrari	52474.182452
manufacturer_datsun	19496.387434
manufacturer_aston-martin	16484.272674
manufacturer_tesla	14103.098848
manufacturer_porsche	10835.412427

Picture - Top 5 features , and their coefficients.

Intercept = 39971.861

Test R2 score = 0.5738324121815643

Training R^2 Score: 0.575328296514303

From Linear regression being a **Ferrari** drives the cost of the car.
Next are **Datsun, Aston-Martin, Tesla, Porsche**

Feature	**Coefficient**	**Price implication**
Ferrari	52474	A Ferrari adds approximately $52k to the price compared to the baseline brand.
Datsun	19496	Datsun adds ~$19k to the price
Aston-Martin	16484	Aston Martin cars increase the price by ~$16k.
Tesla	14103	Tesla cars increase the price by ~$14k.
Porsche	10835	Porsche adds ~$11k to price.

None of the other features given in the data figure in the top 5 features. Being a “Truck” is the 6 feature on my list , with a coefficient value of 8739.

2. Linear Regression with Forward Selection and cross validation.

I performed linear regression with forward selection and cross validation with fold = 5

The top 5 features selected with this method and their coefficient values are

Feature	Coefficient
type_truck	11303.309939
type_pickup	8804.053341
type_sedan	-4883.452399
car_age	-757.842619
odometer	-0.084855

Train R2 : 0.5011719829121948

Test R2 : 0.49937563759256665

3. Regularization with Lasso

I performed regularization with Lasso and the following parameters.

alpha = 10 , max_iter = 10000 , tol = 0.0001)

These are the results

Lasso Test R2 : 0.5688252785955601

Lasso Train R2 : 0.5708030174146866

4. Regularization with Ridge

I performed regularization with Ridge and the following parameters.

alpha = 10 , max_iter = 10000 , tol = 0.0001)

These are the results

Ridge Regression Test R2: 0.5730533171610614

Ridge Regression Train R2: 0.5748613129548226

5. Regularization with HyperParameter selection (Ridge) and 5 fold cross validation

R2 score : 0.5738264708199561

Best alpha : 0.3

Summary of all R2 test scores

Model / Method	R2 Score on Test data
1. Linear Regression	0.5738
2. Linear Regression with Forward Selection and cross validation, and selecting top 5 features.	0.4993
3. Regularization with Lasso	0.5688
4. Regularization with Ridge	0.5748
5. Ridge Regularization with HyperParameter selection and 5 fold cross validation Best alpha = 0.3	0.5738

Summary

Almost all the models said that the price of the used car depends on the model of the car.
The model of the car is what drives the price of the car and Ferrari being the top most expensive car.

The Linear regression model did a good job initially , but all the models cannot be used for production since because of the low R2 score.

Based on this can we make recommendations to used car dealers ?
No.
With this analysis car dealers would have to be selling very expensive cars like Ferrari , Aston Martin and so on. The number of customers buying these cars are very few. The database provided and the models used will give you the features that impact the cost of the car. They are not actually taking the fact as to how many cars of these models are being sold.

We need a better model to make recommendations to used-car dealers on what car to stock in their inventory.

I looked up the internet and came across the “RandomForestRegressor” model

RandomForestRegressor Model

Training R2 score : 0.9764

Testing R2 score : 0.8418

The top 5 important features according to this model are

1. **Car_age**
2. **Odometer**
3. **Truck**
4. **Pickup**
5. **Sedan**

Linear and regularized regression identify features with the highest direct price impact (luxury brands), while Random Forest identifies features that most frequently reduce prediction error across the full dataset (car age and mileage), which is why the most important features differ between the two approaches.

Findings

- The age of the car, the odometer reading , what type of vehicle (truck or pickup or sedan) are the most influential features in determining the price of the car.
- Car age strongly reduces prices.
  - Stocking newer used cars will give dealers more margins
- Odometer mileage constantly reduces prices.
  - More the mileage on a car, lesser the cost of the car.
- The dealership should stock more trucks and pickups because:
  - They may have much higher margins.
  - Buyers value utility vehicles (towing, payload, durability).
- Sedans are the most sold , so car dealers must stock them for inventory turn-around.

Future Work.

- The current model tells what are the features that influence the price of a vehicle. Instead a model must be build which will maximize the profits of the car dealer.
- Consider the omitted features for a more accurate model .
  - This will require a more powerful machine at my end.
- Considering “State” and “City” will help in a better targeted model catering specific locations.