Required Assignment 11.1
What Drives the Price of a Car ?

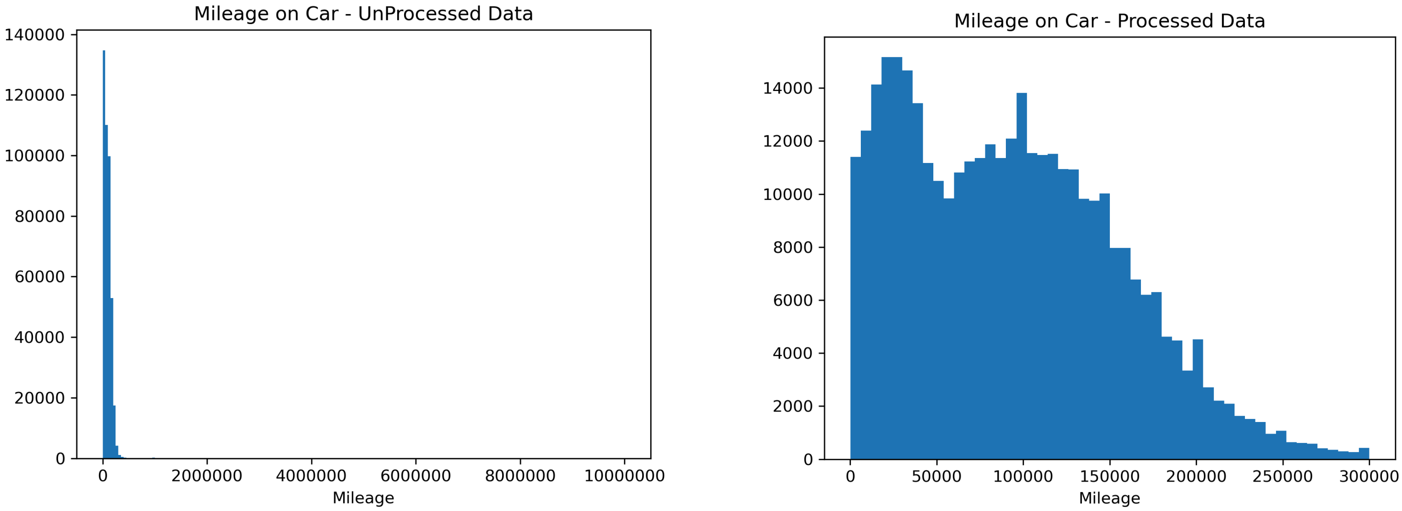
Github repository
https://github.com/devasidgmail/car_prise.git

Overview

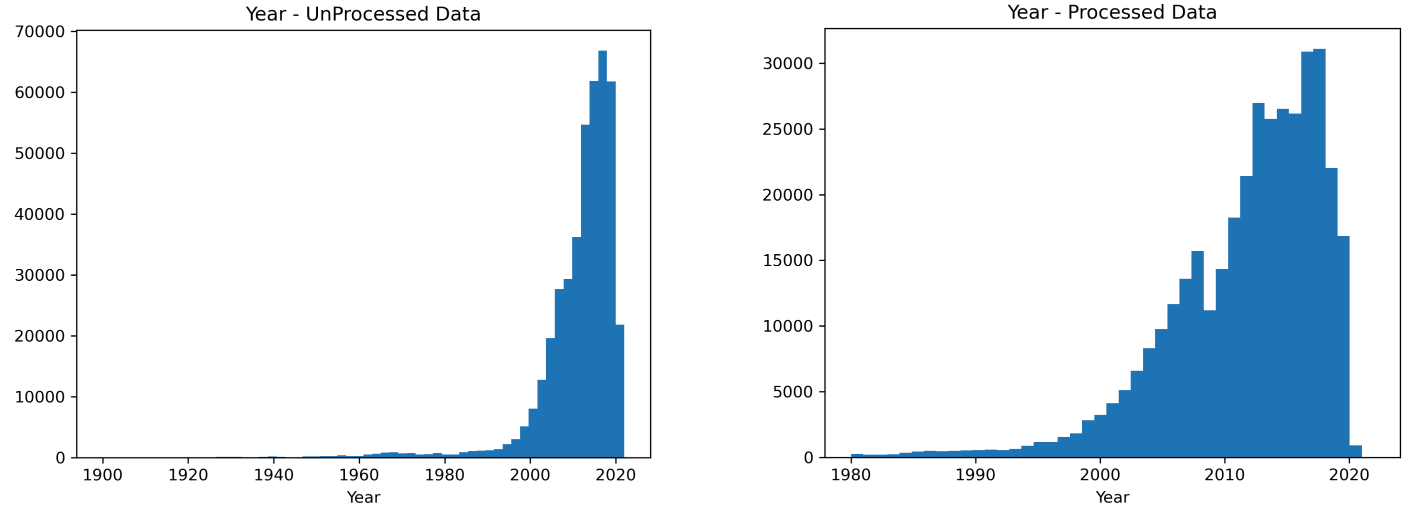
The aim of this project is to determine what factors make the price of a used car more or less expensive and which features are most important in determining the price of the car.

Understanding and cleaning up the data.

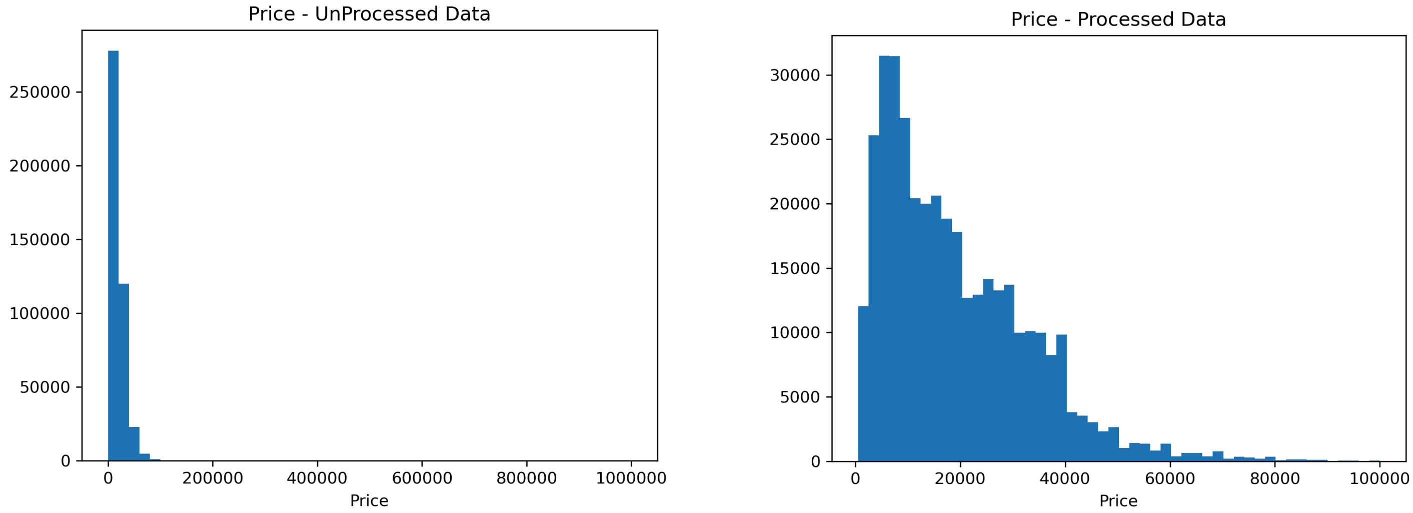
In order to understand the data I plotted histograms of the data and noticed that there was a **large variation in price, age and mileage on the cars. These outliers were removed.**



Mileage on Car : UnProcessed and Processed data.



Year of Car : UnProcessed and Processed Data



Price of the car : UnProcessed and Processed Data

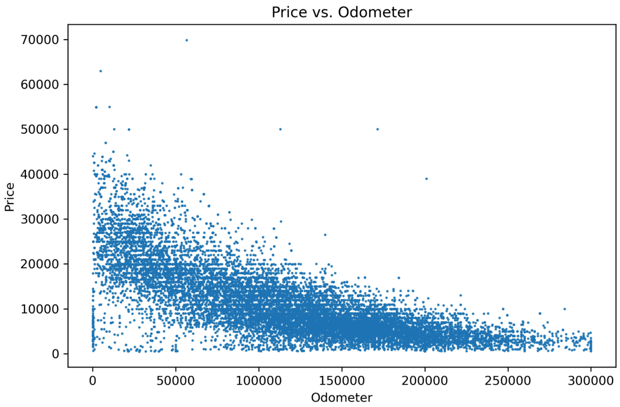
Add new Feature : Car_Age

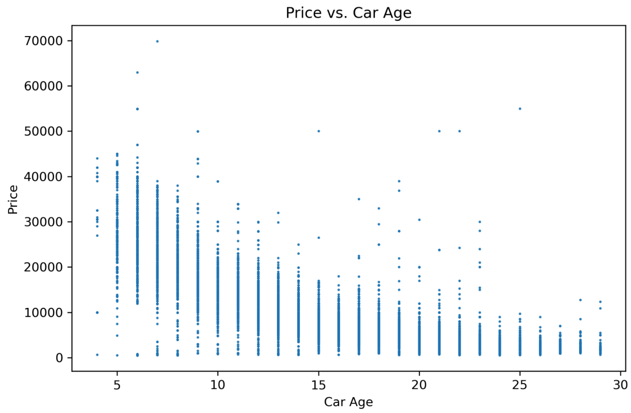
Instead of working with the “Year” of the car, I added a **new feature “Car_Age”.**

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.





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

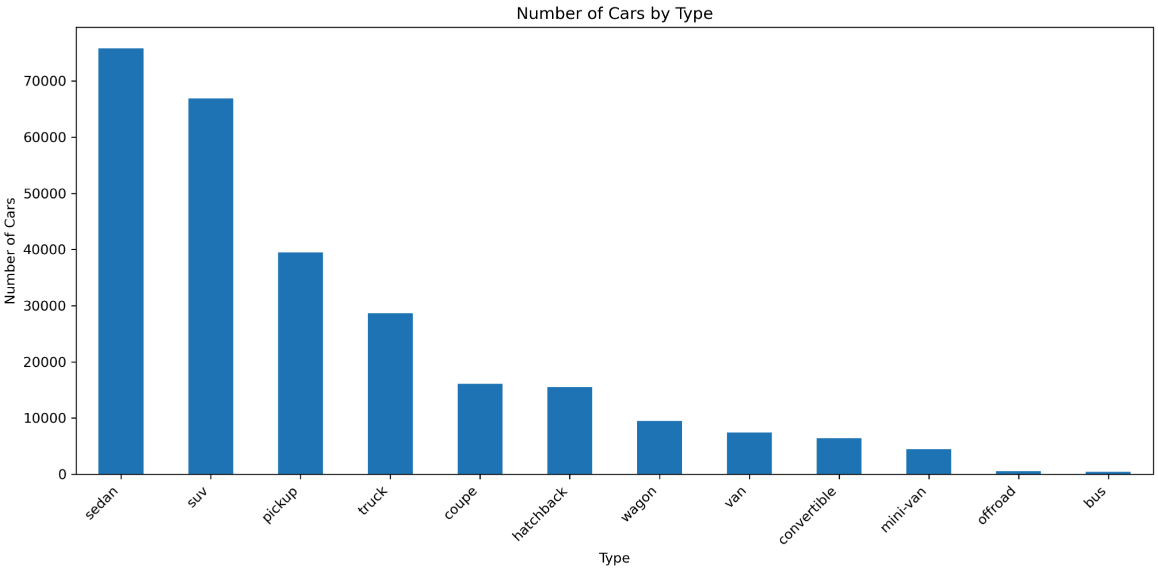
Computational limitations

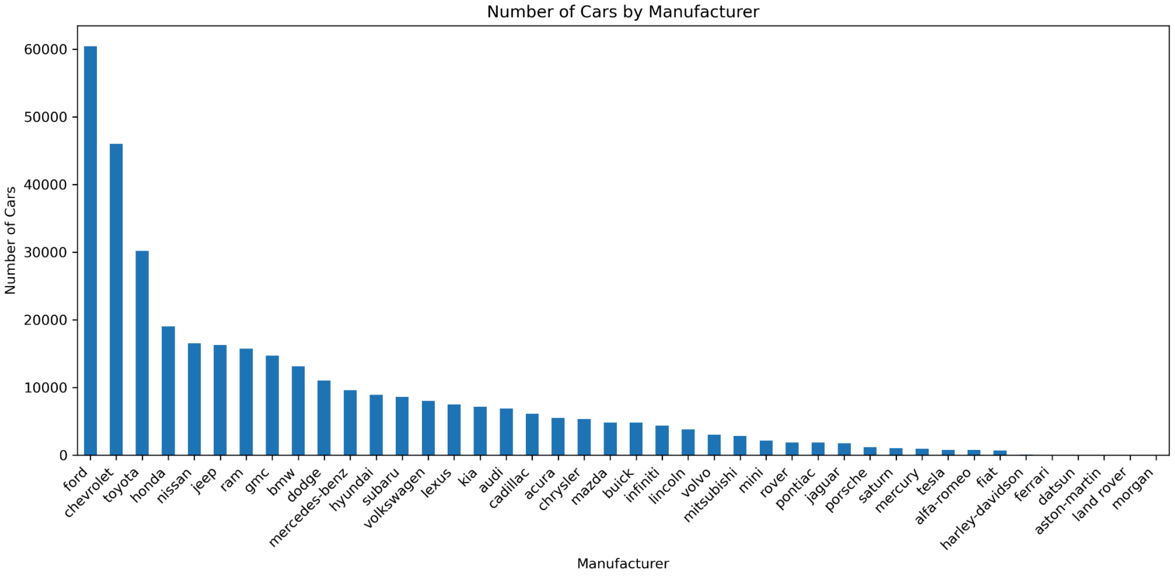
Due to the computational power limitations on my home PC, a lot of features , like **“vin”**, **”id”** **were dropped** . Also,some lesser relevant features like **cylinders, fuel, transmission were also dropped.**

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

More Analysis

Number of cars sold by type.





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

Regressions and methods used

1. **Linear Regression with Forward selection.**

   The first method employed was LinearRegression with forward selection.
   The selection was cross validated with R^2 score and 5 folds.

   The top 5 features from the LinearRegression Model and their coefficients are

   1. car_age : -757.9168886459646
   2. odometer : -0.08500296772393606
   3. type_truck: 11240.699406410857
   4. type_pickup: 8774.613900365743
   5. type_sedan: -4891.545054972943

      Intercept: 36173.51416554385
      R2 = 0.5008167896526374

Feature	Coefficient	Interpretation
type_truck	+11,240.70	Being a truck increases price by about $11,240 compared to the baseline category (likely SUV or other).
type_pickup	+8,774.61	Being a pickup increases price by about $8,775 vs. baseline.
type_sedan	−4,891.55	A sedan decreases price by about $4,892 compared to baseline.
car_age	−757.92	Every additional year of age reduces price by about $758.
odometer	−0.085	Each extra mile reduces price by about $0.085

2. **Ridge Regression with hyperparameter and grid search**

   1. car_age   -757.916900
   2. odometer   -0.085003
   3. type_truck 11240.694788
   4. type_pickup   8774.610992
   5. type_sedan -4891.545091

      cross-validated R2: 0.49725250620694894

3. **Lasso**
   1. car_age   -757.916891
   2. odometer   -0.085003
   3. type_truck 11240.684955
   4. type_pickup   8774.603020
   5. type_sedan -4891.541854

   cross-validated R2: 0.49725250620694894

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.