

# Recommendations for buying Used Vehicle Based on regression analysis

---

Web Scraping and Linear Regression Analysis

Wei Zhao

# Introduction

---

Vehicles are important tools that allow great flexibility in our daily life. While our life is facilitated by vehicles, you may want to ask when you are buying a car:

- ❖ How do I choose it?
- ❖ What year, make, and model?
- ❖ With limited budget and buying a used car, what features of should I focus on?

# Data Collection

|   | make          | year | mileage | fuel_type | drive_type | transmission | engine_size | engine_type | cty_mpg | hwy_mpg | price |
|---|---------------|------|---------|-----------|------------|--------------|-------------|-------------|---------|---------|-------|
| 0 | INFINITI      | 2017 | 46423   | gas       | awd        | automatic    | 3.50        | regular     | 19      | 26      | 27690 |
| 1 | Chevrolet     | 2017 | 40653   | gas       | awd        | automatic    | 3.60        | regular     | 17      | 24      | 28589 |
| 2 | Jeep          | 2020 | 11994   | gas       | awd        | automatic    | 2.40        | regular     | 22      | 30      | 24900 |
| 5 | Mercedes-Benz | 2013 | 76867   | gas       | awd        | automatic    | 3.50        | regular     | 20      | 28      | 14995 |
| 6 | Nissan        | 2019 | 29537   | gas       | fwd        | automatic    | 2.50        | regular     | 27      | 37      | 17995 |
| 7 | Ford          | 2014 | 98388   | gas       | awd        | automatic    | 1.60        | turbo       | 22      | 30      | 11995 |
| 8 | Ford          | 2015 | 69024   | gas       | awd        | automatic    | 3.50        | regular     | 17      | 23      | 19995 |
| 9 | Subaru        | 2012 | 98464   | gas       | awd        | automatic    | 2.00        | regular     | 27      | 36      | 8995  |

Clean Data  
6557 cars  
in total

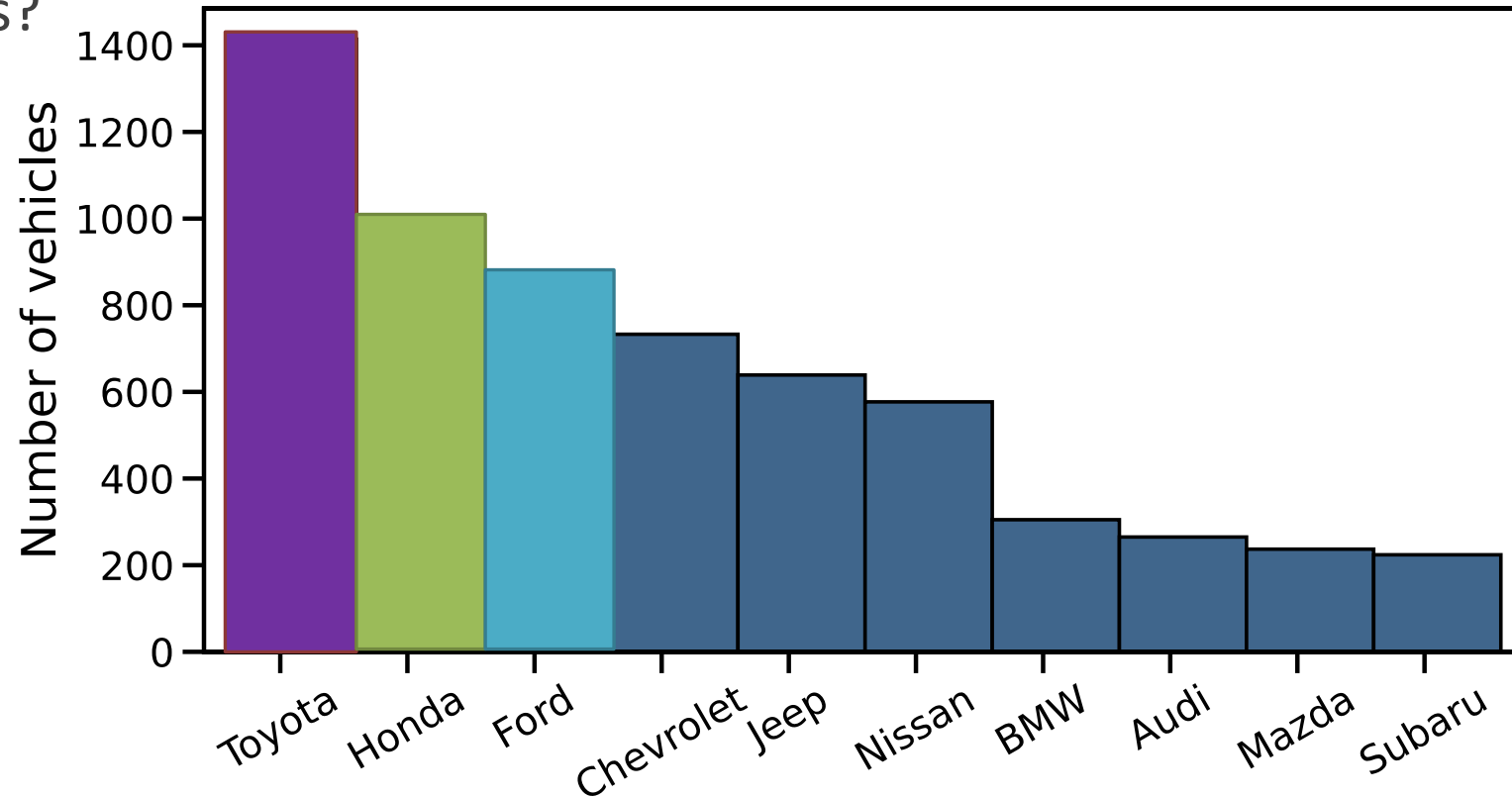
Within 25 mi  
from Boston, MA



# Exploratory analysis

---

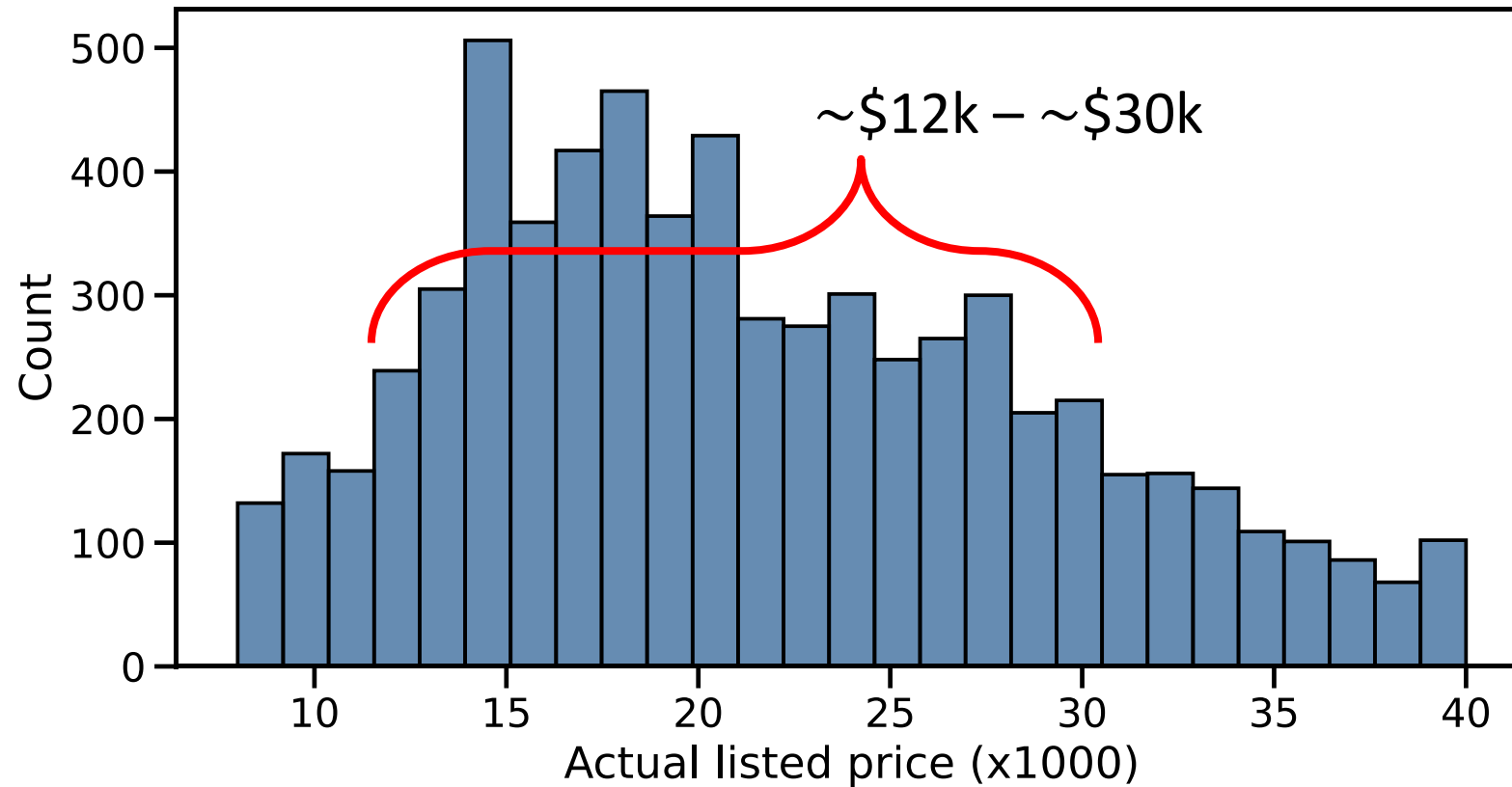
What makes?



# Exploratory analysis

---

Plan your  
budget



# Features and Labels

---

## Features:

### Continuous variables:

- ❖ Year
- ❖ Mileage
- ❖ Engine size (# liters)
- ❖ City MPG

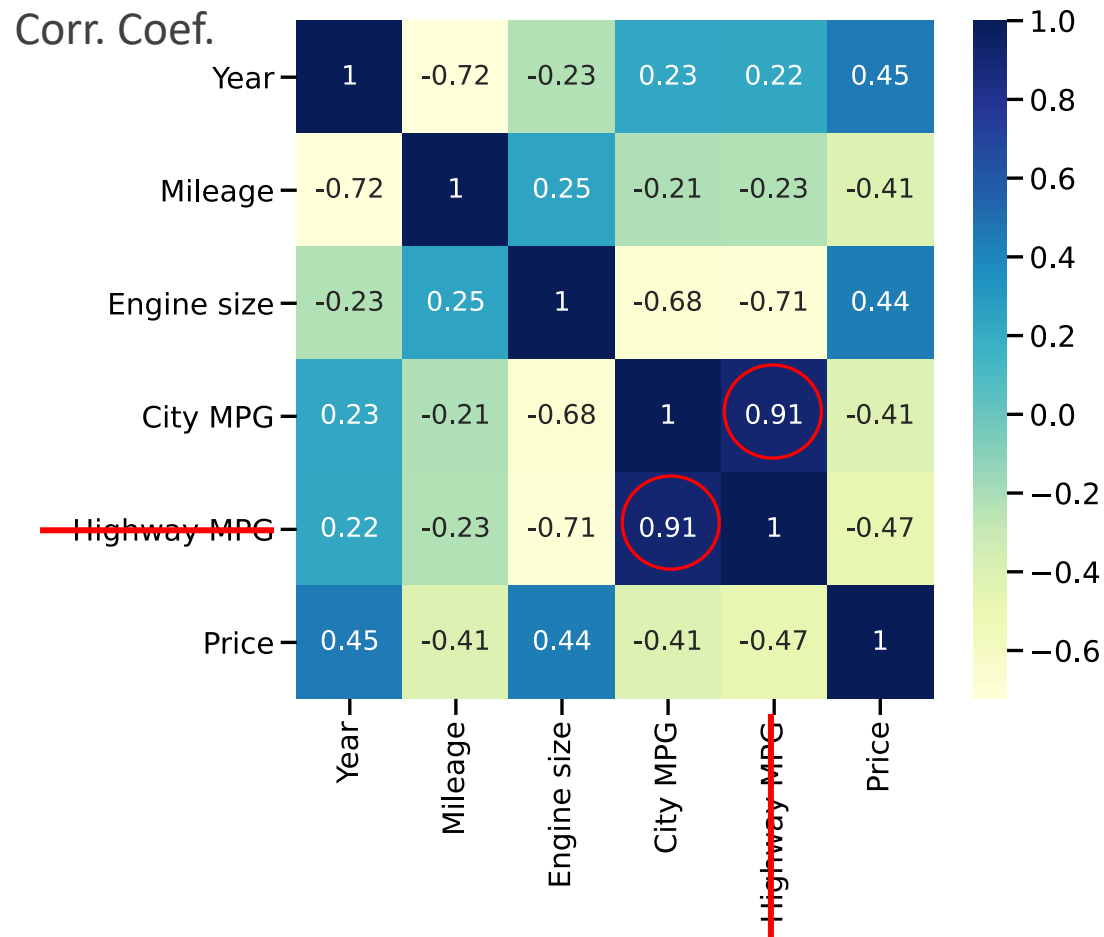
### Categorical variables:

- ❖ Fuel type (gas, diesel, hybrid)
- ❖ Drive type (AWD, FWD, RWD)
- ❖ Transmission (Manual and Auto.)
- ❖ Engine type (Regular and Turbo.)

## Label:

Listed price

# Feature Engineering



Redundant feature:

- ❖ Highway MPG

Interaction features:

- ❖ Drive type × City MPG

- RWD → same gas, more power (typically on luxury vehicles)

- ❖ Transmission × City MPG

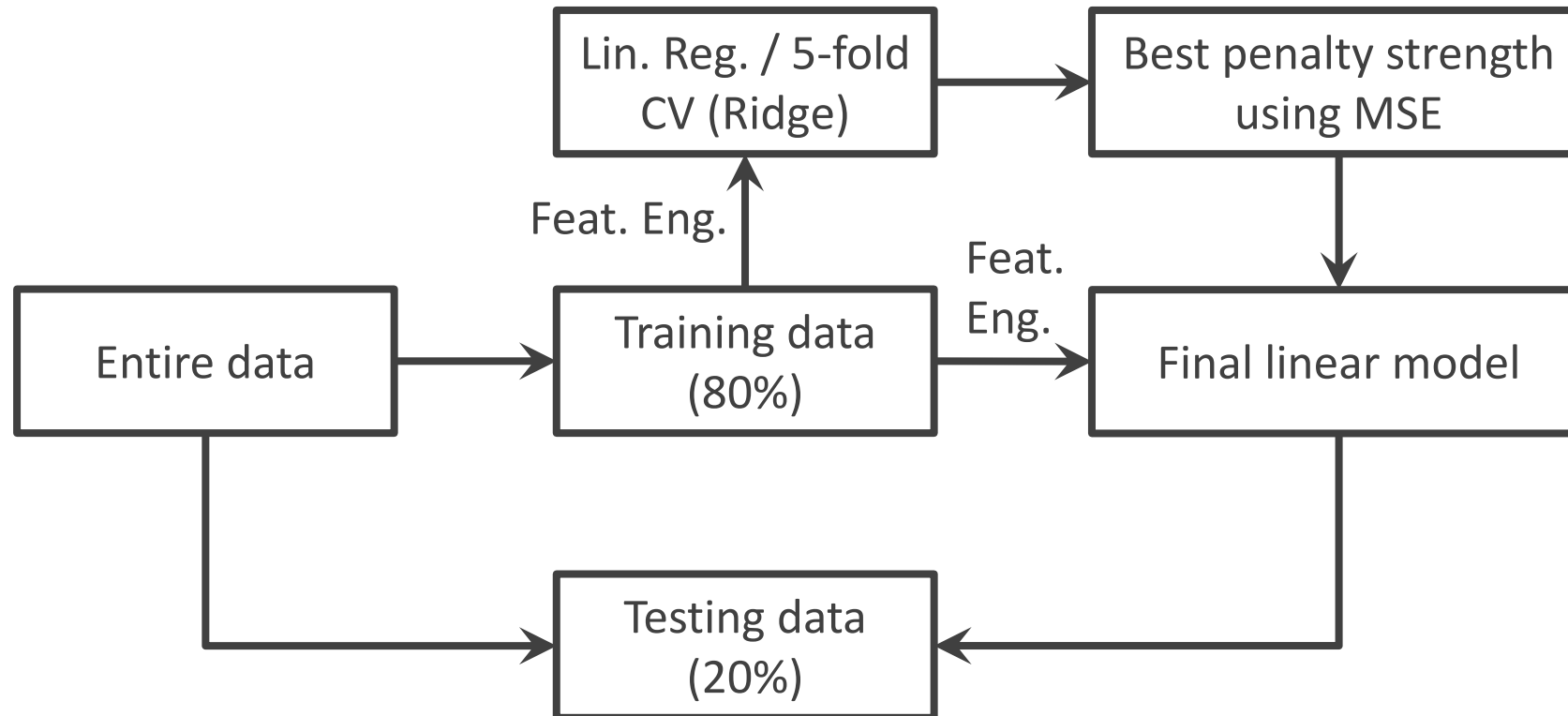
- M.T. on low trim and high MPG → low price

- ❖ Fuel type × City MPG

- Hybrid and high MPG → high price

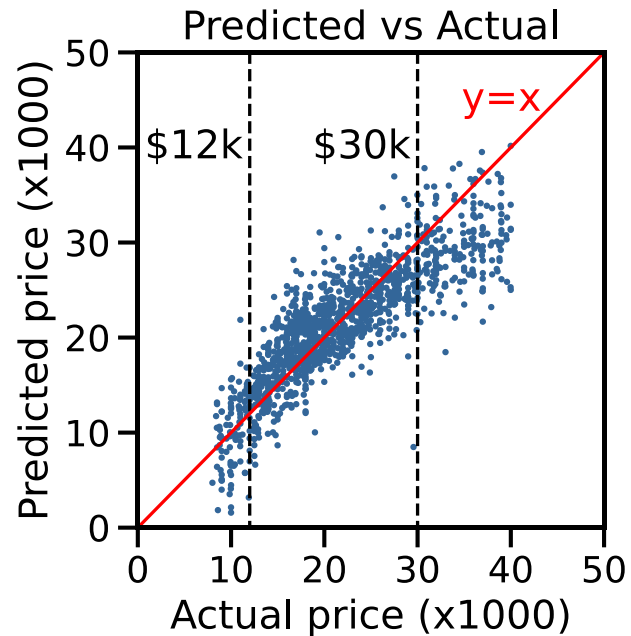
# Linear Regression modeling workflow

---





# Results

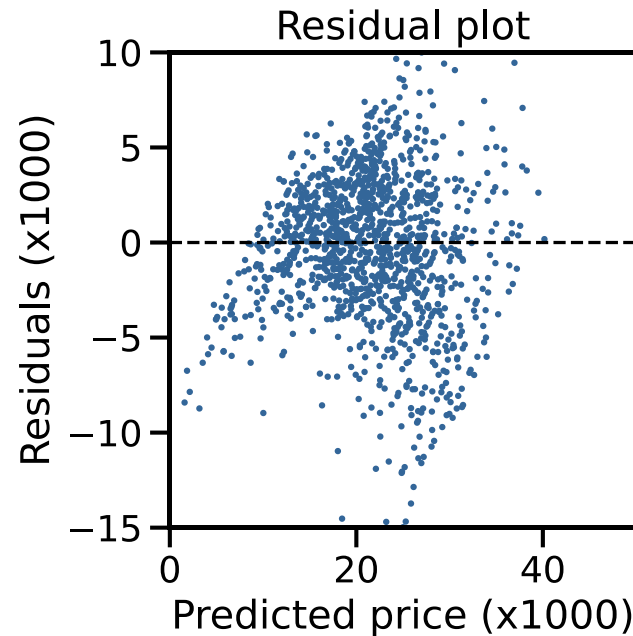


Training  $R^2 = 0.74$

Testing  $R^2 = 0.72$

MAE = \$3008

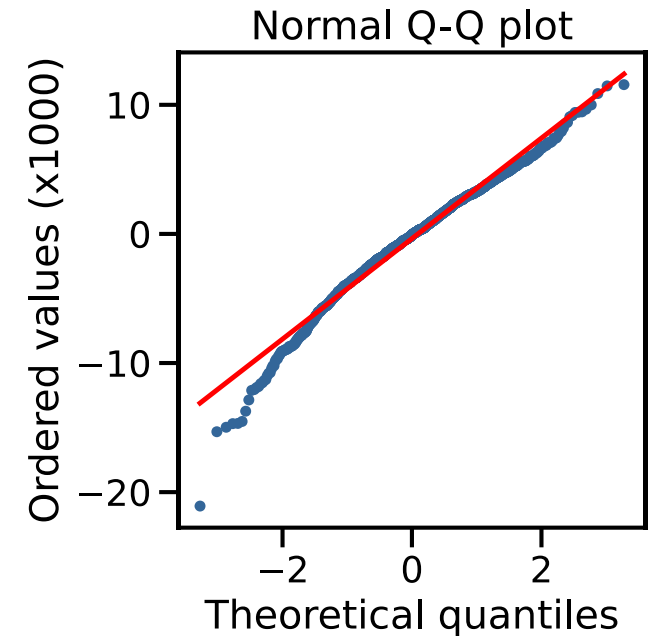
F-test:  $p$ -value = 0



Lin. Reg. between residuals  
and predicted price:

$R^2 = 0$ ; Slope = -0.01

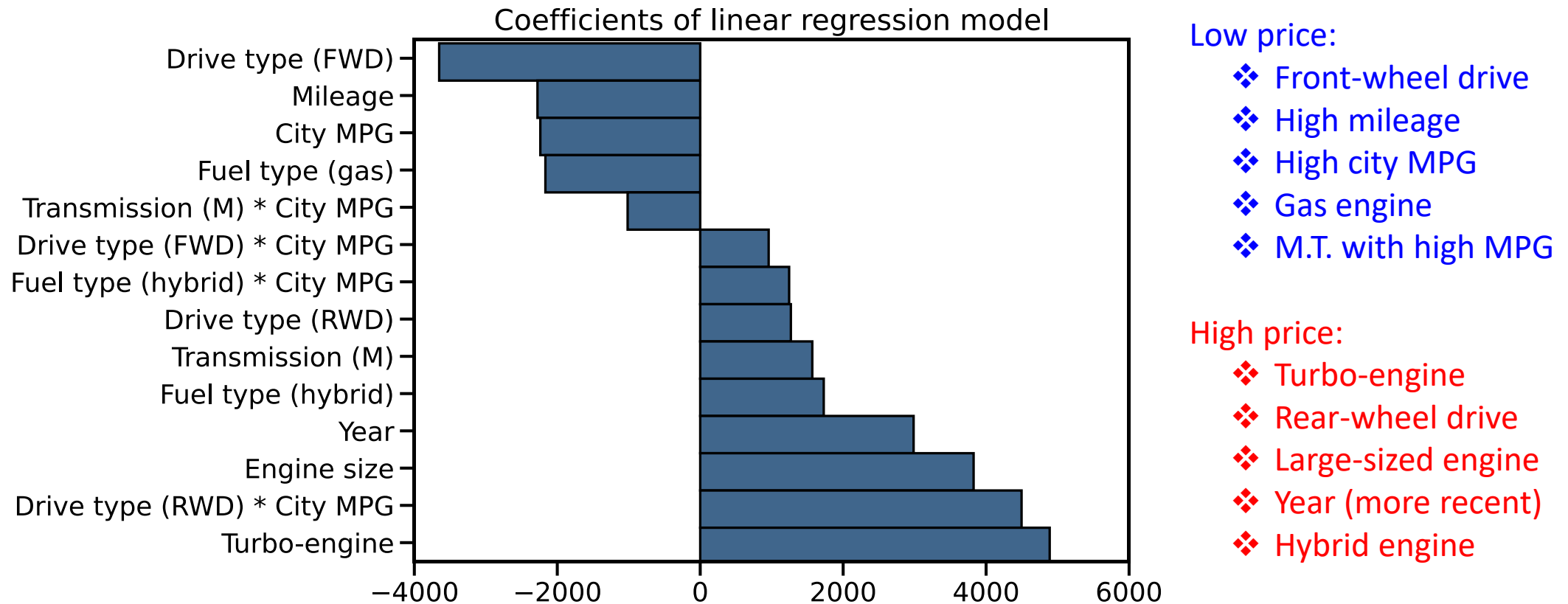
No specific pattern and trend



Residuals are normally  
distributed.

Left skewed  $\rightarrow$  price likely  
underestimated

# Features affecting price



# Summary

---

1. Top 3 popular makes:



2. Price range of most vehicles:

~\$12k – ~\$30k

3. Features to look at when buying a cheap used car:

- ❖ Front-wheel drive
- ❖ Certain amount of mileage
- ❖ Small-sized engine
- ❖ High city MPG
- ❖ Gas engine, not hybrid
- ❖ Not too recent

# Future Work

---

1. More specific price range
2. A specific make, e.g., Toyota, Honda, or Ford
3. More sophisticated and advanced deep neural network

Questions?