

# Web-Scraping Project: Used Cars on U.S. Market

[www.carfax.com](http://www.carfax.com)

Zhengqing (James) Chen

02/14/2018

# Outline

- ① Data (Source, Post-Scraping Processing)
- ② Observations:
  - ① Top fives
  - ② Comparison of a certain class of cars
  - ③ Correlations between quantities

- Select from 1000 most populous U.S. zip codes; 25 cars from each
- Key info. to be scraped **#done with scrapy**  
Year, Make, Model, Price, Mileage, Body, Color, Engine

The screenshot shows a web browser displaying the Carfax website. The URL is https://www.carfax.com/Used-Cars-in-New-York-NY\_c88636. The page features a header with the Carfax logo and navigation links for Used Cars for Sale, CARFAX Reports, My Car Maintenance, Used Car Values, Research, Help, Sign In, and Sign Up.

**2015 Toyota RAV4 LE**  
\$15,490 (\$254/mo est.)  
GREAT VALUE \$3,630 below \$19,120 CARFAX Value

Mileage: 17,569 miles Body Type: SUV  
Color: Gray Engine: 4 Cyl  
Description: Used 2015 Toyota RAV4 LE with AWD, Keyless Entry, Cruise Control, CD (Single Disc), Rear-View Camera, Air Conditioning, MP3, Cloth Seats, Power Steering, Power Windows, AM/FM  
Certified Luxury Motors: Great Neck, NY (15 miles from New York, NY)

**2015 Audi A6 2.0T**  
\$22,980 (\$376/mo est.)  
GREAT VALUE \$5,400 below \$28,380 CARFAX Value

Mileage: 31,501 miles Body Type: Sedan  
Color: Black Engine: 4 Cyl  
Description: Used 2015 Audi A6 2.0T with AWD, Leather Seats, Fog Lights, Alloy Wheels, Navigation System, Keyless Entry, Cruise Control, Heated Seats, CD (Single Disc), Air Conditioning, MP3  
Certified Luxury Motors: Great Neck, NY (15 miles from New York, NY)

# Post-Scraping Data Processing

- All the usual data cleaning; Add field: Manufacturer Country (e.g. Audi: Germany)
- For practical consideration, only retain
$$\{Car_j : 5,000 \text{ mi} < \text{Car}_j[\text{'Mileage'}] \leq 90,000 \text{ mi}\}$$
- Table ready:

```
In [86]: df = pd.merge(df, make_country_df, on='Make');

df = df[(df.Mileage>5000) & (df.Mileage<=90000)];

print(df.shape)

df.sample(10)
```

(17356, 9)

Out[86]:

|       | Year | Make          | Model    | Mileage | Price | Body      | Color  | Engine | Country |
|-------|------|---------------|----------|---------|-------|-----------|--------|--------|---------|
| 7581  | 2015 | Chevrolet     | Equinox  | 14296   | 18884 | SUV       | White  | 4 Cyl  | USA     |
| 17506 | 2004 | Mini          | Cooper   | 79801   | 4999  | Hatchback | Silver | 4 Cyl  | UK      |
| 13815 | 2015 | Mercedes-Benz | M-Class  | 37984   | 33991 | SUV       | Gray   | 6 Cyl  | Germany |
| 14315 | 2014 | Mercedes-Benz | E-Class  | 32697   | 23999 | Coupe     | Blue   | 6 Cyl  | Germany |
| 15753 | 2011 | Cadillac      | CTS      | 71462   | 15640 | Coupe     | Silver | 6 Cyl  | USA     |
| 17960 | 2015 | Lexus         | RX       | 18781   | 30996 | SUV       | Black  | 6 Cyl  | Japan   |
| 5968  | 2015 | Nissan        | Altima   | 57027   | 8995  | Sedan     | Gray   | 4 Cyl  | Japan   |
| 6497  | 2015 | Nissan        | Rogue    | 37849   | 16995 | SUV       | White  | 4 Cyl  | Japan   |
| 13764 | 2015 | Mercedes-Benz | GL-Class | 22167   | 44484 | SUV       | Gray   | 6 Cyl  | Germany |
| 12311 | 2017 | Toyota        | Camry    | 9606    | 17995 | Sedan     | Red    | 4 Cyl  | Japan   |

## Some Top Fives

In terms of *volumes* on market: `#desc( n() )`

- Top 5 Makes

|             |  |
|-------------|--|
| Makes       | Toyota, Ford, Nissan, Chevrolet, Mercedes-Benz |
| U.S. Makes  | Ford, Chevrolet, Jeep, Dodge, GMC              |
| Int'l Makes | Toyota, Nissan, Mercedes-Benz, BMW, Hyundai    |

- Top 5 Models

|              |  |
|--------------|--|
| Models       | Toyota Corolla, Toyota Camry, Nissan Altima, Toyota RAV4, Nissan Rogue         |
| U.S. Models  | Chevrolet Silverado, Jeep Wrangler, Ford Explorer, Chevrolet Cruze, Ford F-150 |
| Int'l Models | Toyota Corolla, Toyota Camry, Nissan Altima, Toyota RAV4, Nissan Rogue         |

## Some Other Top Fives

- In terms of *volumes* on market: `#desc( n() )`

|              |                                 |
|--------------|---------------------------------|
| Top 5 Colors | Black, White, Gray, Silver, Red |
| Top 5 Audi   | Q5, A4, A6, A3, A5              |

- In terms of *price* on market: `#desc( median(Price) )`

|                    |  |
|--------------------|--|
| Top 5 U.S. Models  | Dodge SRT, Cadillac Escalade, Lincoln Continental, Ford F-250, GMC Yukon |
| Top 5 Int'l Models | Audi R8, Porsche 911, BMW i8, Toyota Land Cruiser, Mercedes-Benz GLS     |

# Comparison of Three Models of the Same Class {Audi A4, BMW 3, Mercedes-Benz C-Class}

(1/3)

Note: Audi A4 =  $\mathcal{O}$  (BMW 3) =  $\mathcal{O}$  (Benz C-Class)



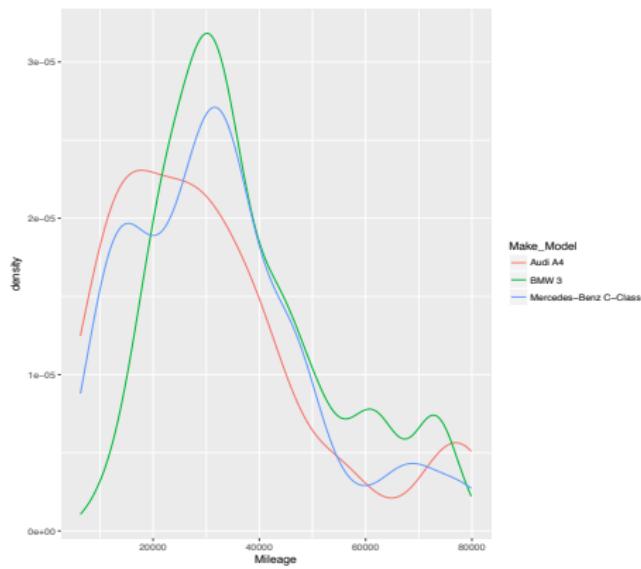
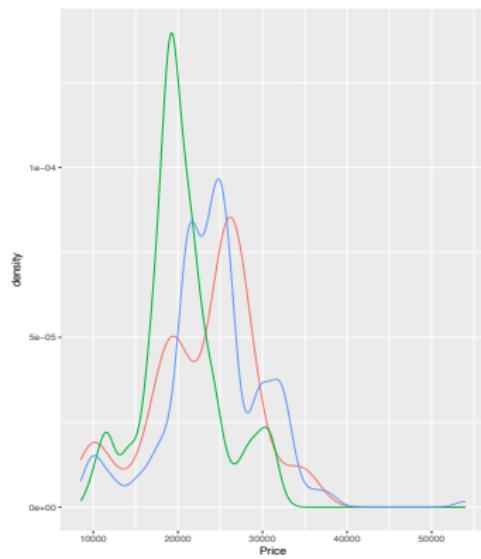
# Comparison of Three Models of the Same Class {Audi A4, BMW 3, Mercedes-Benz C-Class}

(2/3)

Note: Audi A4 =  $\mathcal{O}$  (BMW 3) =  $\mathcal{O}$  (Benz C-Class)

Left: Price distributions; Right: Mileage distributions

(not very informative)



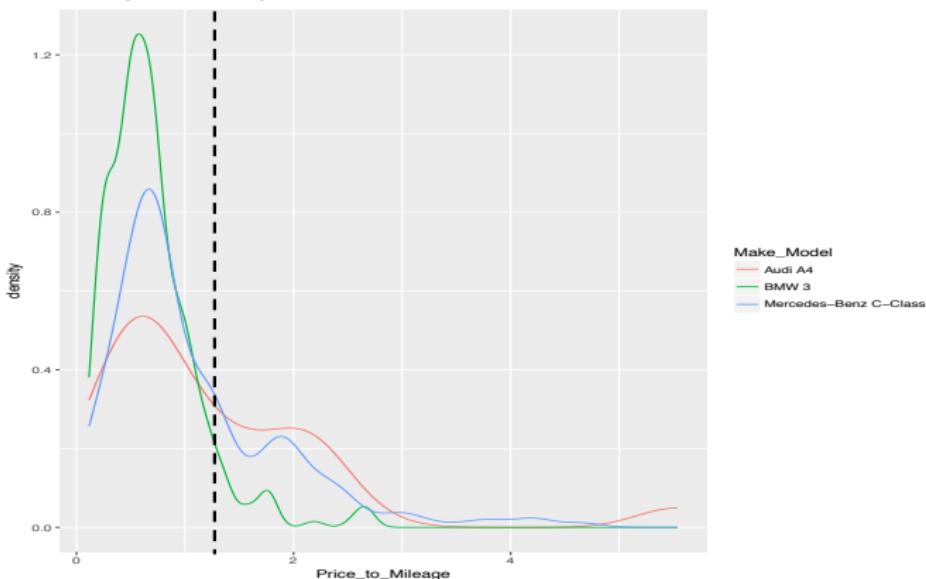
# Comparison of Three Models of the Same Class {Audi A4, BMW 3, Mercedes-Benz C-Class}

(3/3)

Think about:

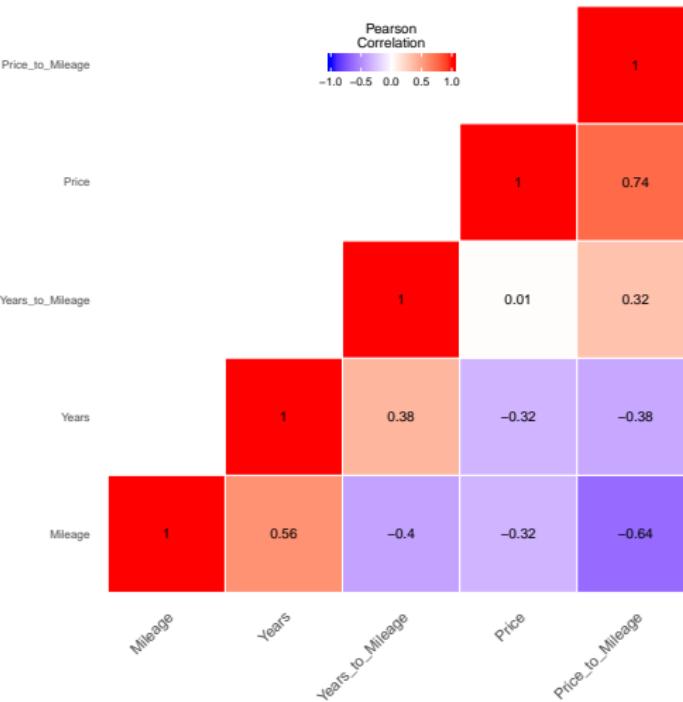
|                                    |     |                                    |
|------------------------------------|-----|------------------------------------|
| <u>Same \$\$\$</u><br>More Mileage | vs. | <u>Same \$\$\$</u><br>Less Mileage |
| <u>Less \$\$\$</u><br>Same Mileage | vs. | <u>More \$\$\$</u><br>Same Mileage |

Postulate: Two categories of potential consumers for this class of cars



# Heat Map of Pearson Correlations

Note: Quantity  $\frac{\text{Years}}{\text{Mileage}}$   $\sim$  Gentleness of Usage, e.g. consider:  $\frac{5 \text{ years}}{5000 \text{ miles}}$  vs.  $\frac{5 \text{ years}}{50000 \text{ miles}}$



# Thank You

- Possible Future Work:
  - ① More refined study wrt. locations, e.g., Mid-West market, NYC area market, etc
  - ② Incorporate other data, e.g., area demographic statistics, etc
  - ③ Discover more interesting patterns and try to explain them
  - ④ ...
- Questions?
- Thank you!