Price Dispersion - Used Car Industry Case Study

Author: David Tian

Background

As a result of the pandemic, used car prices have skyrocketed to record levels. What has been historically viewed as a depreciating asset is now appreciating. The increase in used car prices can be explained through the laws of supply and demand: limited inventories of used cars, paired with strong demand from consumers, have created an upward price pressure.

When it comes to supply, the inventory of used cars has decreased for two main reasons. Firstly, rental car companies were forced to sell a third of their fleets last year to raise enough cash to survive the pandemic. With the current rebound in travel, rental car companies face a shortage of cars, and aren't selling their inventory even as demand for used cars has soared. Secondly, a major shortage of computer chips has resulted in a decrease of new car inventory. As a result, rental cars are holding on to existing fleets because they can't buy replacement cars.[1]

Meanwhile, demand for used cars has increased significantly during the pandemic, as people avoided mass transportation and became more price sensitive.[2] As the nation re-opens, new job hirings and lifts in work-from-home restrictions have further amplified the demand for used cars.[1]

While the above-mentioned factors have certainly influenced pricing in the used car marketplace, this analysis will focus on car type, model year, and mileage in order to provide pricing transparency and analyze price dispersion.

Data

Data was web scraped from CarMax, AutoNation, Berkshire and Sonic websites via Python's BeautifulSoup and Selenium libraries on August 17, 2021 and August 18, 2021. These retailers represent 4 of the top 5 major used car retailers.[3]

Data scraping was limited to:

- Geography: Nationwide (USA)
- Car type: SUVs and sedans
- Condition: Used/Pre-Owned only (does not include certified pre-owned)

Data Dictionary:

- price - price of car
- mileage - mileage of car
- model_year - model year of car
- retailer - retailer selling car
- car_type - type of car
- manufacturer - manufacturer of car (not included in regression)

Data Pre-Processing:

- Rows with missing data field for price or mileage were removed
- Model years with less than 30 cars were removed
- Prices lower than 0.5 percentile and higher than 99.5 percentile were removed
- Mileage lower than 0.5 percentile and higher than 99.5 percentile were removed

The final dataset used for analysis consisted of ~35K SUV records and ~22K sedan records.

Assumptions

1. Data scraped is assumed to reflect each retailer’s entire used car inventory.
    a. Violation of this assumption (i.e., there exists a lag-time from when a car is available and then posted online) would require additional research to understand causes (implications for any future time-series analysis) or modification to web scraping code.

2. Pricing data is assumed to closely reflect the final sale price to the consumer.
    a. While this assumption does not reflect the negotiable nature of car prices, it simplifies the analysis for comparative purposes.
        i. Retailers may have different rebate contracts with manufacturers, varying commissions structures for their salesmen, and hidden fees which all impact the negotiation process.[4]
    b. Violation of this assumption means pricing data could be inflated for some retailers and not provide a clear comparison between retailers.

3. Mileage is assumed to have a negative linear relationship with price: a unit increase in mileage is associated with some unit decrease in price.
    a. Violation of this assumption means that linear regression results are not valid and that non-linear models such be considered (i.e., tree-based models).

4. Retailer inventories are assumed to have comparable quality/manufacturer mix.
    a. Violation of this assumption means that regression results will need to be revisited and quality/manufacturer features need to be added.

5. Assumptions for linear regression are assumed to be satisfied. Refer to the Appendix for actual assumption testing.
    a. Violation of any of the linear regression assumptions means that regression results may be unreliable/misleading.

6. 95% confidence level is assumed to be statistically significant.
    a. Violation of this assumption means a higher confidence level is required and would broaden confidence intervals, which may impact statistical significance of regression results.

Findings

Exhibit 1 - SUV/Sedan Inventory Distribution by Model Year Between Retailers

This plot illustrates the distribution inventory by model year between retailers for SUV/sedans.

Summary of Findings

1. CarMax has the largest inventory of SUVs/sedans, with AutoNation a distant second. Berkshire/Sonic have similar inventory sizes.
2. Retailers have more SUVs than sedans.
    a. This can be explained by consumer preference shifting towards more expensive SUVs, rather than less expensive sedans.[1]
3. The majority of model years range from the time period: 2015 - 2020.
    a. This could be related to how these retailers purchase their inventory, or to the 7-year time frame around extended warranties.[5]
        i. Further research is required to support this.
4. There are a number of 2022 SUVs for sale, but none for sedan.
    a. It is interesting to note that there is inventory for 2022 used SUVs for sale as early as August. It raises the question of whether people are buying the latest model just to sell it several months later.
    b. In Exhibit 2, we'll investigate this further.

Exhibit 2 - 2022 SUV Counts by Manufacturer

This plot illustrates the 2022 SUV model counts and median mileage between manufacturers.

Summary of Findings

1. There is a disproportionate amount of BMWs.
    a. BMW has historically paid dealers to buy BMWs for their fleets of loaner vehicle in order to show off the latest models to service customers and boost sales.[6] This could explain how BMW, the most popular luxury car brand for 2021,[7] increases the affordability of their cars to capture additional sales without lowering their brand equity.[8] It seems that BMW incentivizes dealers to buy their new models as loaner cars to increase mileage and decrease price.
        i. Further research is required to support this.
    b. Given the affordability of used BMW cars, they could appeal to consumers who are looking to buy the latest luxury car model at a discount.

Exhibit 3 - Mileage and Price vs. Model Year - SUV

This plot illustrates median mileage/price between retailers by model year for SUVs.
Splits by manufacturer are available in Jupyter Notebook, but are not shown for conciseness.

Summary of Findings

1. CarMax SUVs have the least mileage from 2009 - 2020 and tend to be more expensive.
    a. CarMax's used cars could appeal to consumers with higher budgets looking to buy SUVs.
2. Albeit tiny inventory, only AutoNation and Berkshire offer a selection of old and cheap SUVs (2004, 2007).
    a. AutoNation and Berkshire's inventory of old and cheap SUVs could appeal to consumers who are extremely budget-conscious and don't mind buying an old SUV.

Exhibit 4 - Mileage and Price vs. Model Year - Sedan
This plot illustrates median mileage/price between retailers by model year for sedans.
Splits by manufacturer are available in Jupyter Notebook, but are not shown for conciseness.

Summary of Findings

1. CarMax offers the lowest mileage sedans from 2010 - 2020, while pricing is the highest from 2010 - 2016.
2. CarMax's 2018 sedan collection provides the best value (lowest mileage, lowest price) and must be explored by consumers looking for value.

Exhibit 5 - Linear Regression
These exhibits highlight results from linear regression. Linear regression was performed on all 4 retailers data to develop industry confidence intervals for mileage, car type and model year. Regression was then performed on retailer-specific data to compare confidence intervals to industry in order to draw conclusions.

Methodology:

- Mileage and Price were scaled down by a factor of 1,000.
- Retailer - AutoNation was dropped after dummying to avoid the dummy variable trap.
- Car Type - Sedan was dropped after dummying to avoid the dummy variable trap.
- Model Year - Model years with less than 1,000 data points were removed.
    - 2013 was dropped after dummying to avoid the dummy variable trap.
- Manufacturer was not included due to time constraints.

The final dataset used for regression consisted of 55K SUVs/sedans.

Summary of Findings

1. Variation in Pricing
    a. Mileage, car type, and model year explain the most variation in AutoNation's pricing (37.8%), and explain the least variation in CarMax's pricing (27.4%).
        i. Pricing is dependent on numerous factors not present in this analysis. Please refer to next-steps for future improvements.
2. Mileage
    a. For increases in mileage, CarMax decreases pricing less than the industry average, while AutoNation increases pricing more.
        i. Industry average: every 1,000 increase in mileage is associated with a $100 decrease in price.
            1. For CarMax, this associated price decrease is $90.
            2. For Autonation, this associated price decrease is $120.
3. Car Type
    a. Berkshire's SUV and sedan pricing is further apart, while Sonic's SUV and sedan pricing is closer together.
4. Model Year
    a. Industry pricing decreases sharply for newer models and tails-off for older models

Full regression results can be viewed in the Jupyter Notebook.

Next Steps

Outside of the further research noted, potential next-steps could be:

1. Ensuring linear regression assumptions are met should be prioritized.
   a. Generalized linear models (GLM) can be explored due to relaxed normality assumptions.[9]

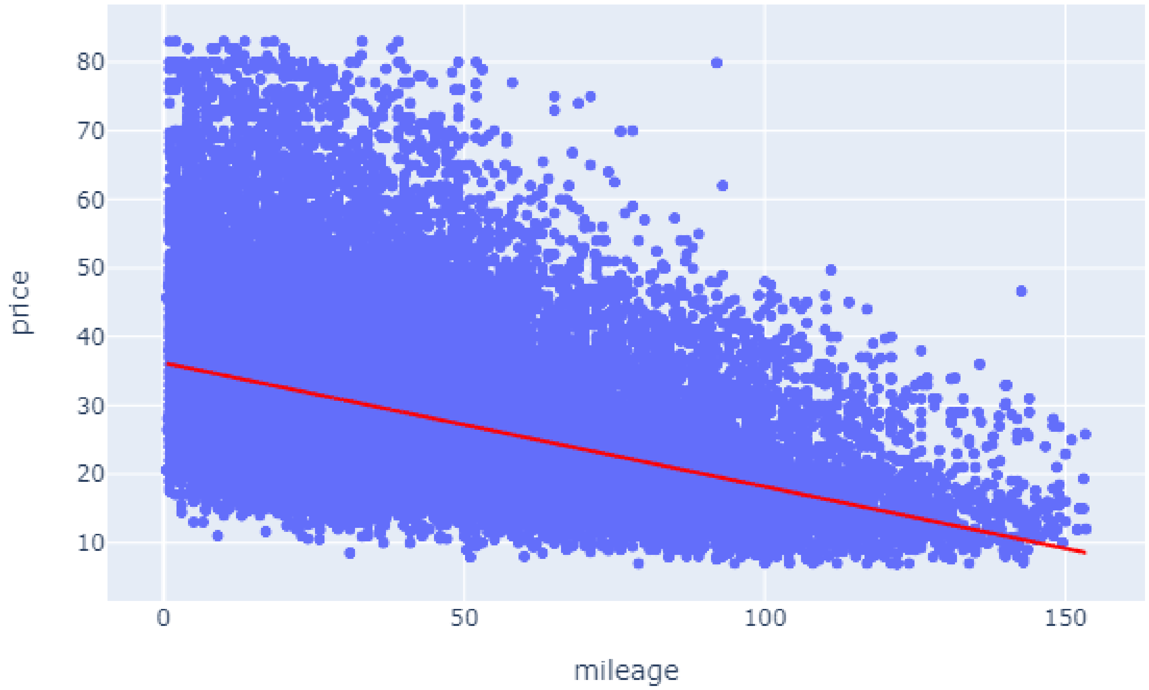
2. Regression can incorporate manufacturer data, which would explain an additional 17.1% in the variation in industry pricing (adj. r-squared of 47.6%).

3. Expanding the underlying data to capture additional highly predictive variables.
   a. Additional car types and capturing geographical data would be a priority.
      i. For instance, convertibles and sports cars command higher prices along the coasts and in warmer climates, while four-wheel-drive trucks and SUVs do best in the Northeast, Midwest, and other areas that get a lot of snow.[10]
      ii. Geographical differences between retailers and inventory of car types can be analyzed to draw insights into which retailer is positioned best for the upcoming Fall/Winter seasons.
   b. Interaction terms could be explored in linear regression (i.e., should mileage have a different effect on pricing depending on manufacturer).[11]

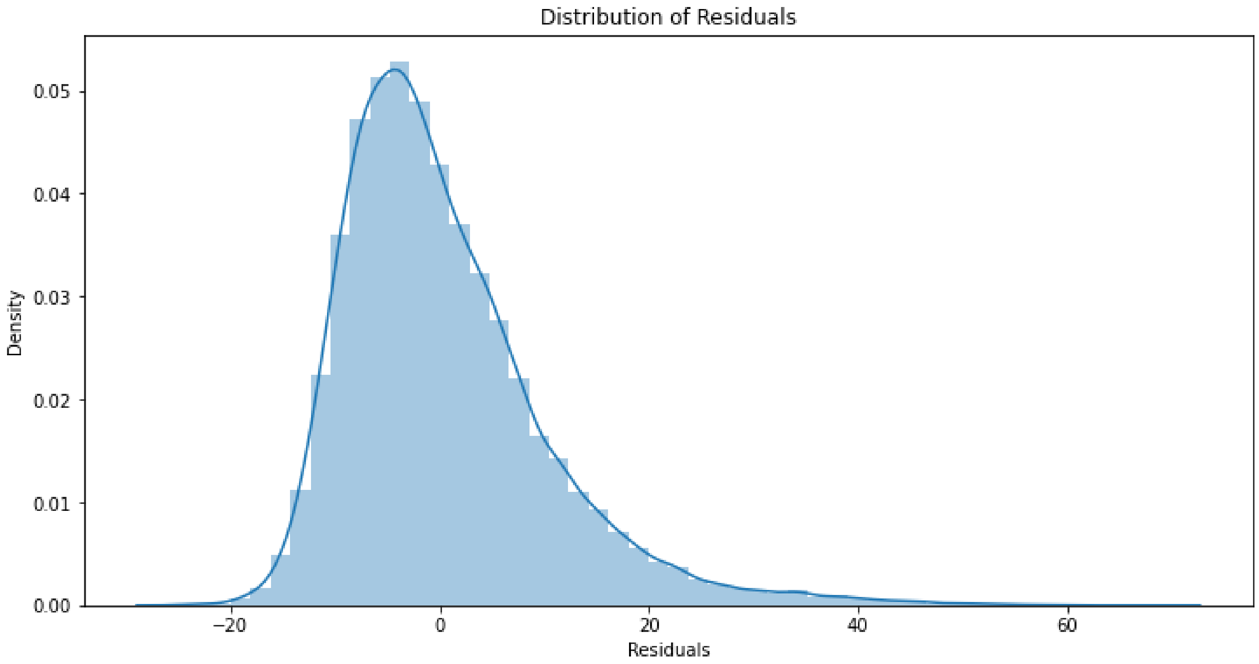
4. Additional variables to consider scraping:
   a. Gas/Electric/Hybrid or Economy/Luxury classifications
      i. Further research to understand customer and market segments to help distinguish retailers.
   b. Physical condition of car.
      i. Physical condition is more subjective than mileage, but it is as important as mileage in assessing value.[11]
   c. Options.
      i. Diesel engines / All-wheel drive / Panoramic moon roofs / Premium factory sound system / Leather seats
         1. To avoid the curse of dimensionality, counts of options can be shown.
   d. Expanding the number of retailers from 4 to 10.

Appendix - Linear Regression Assumption Testing

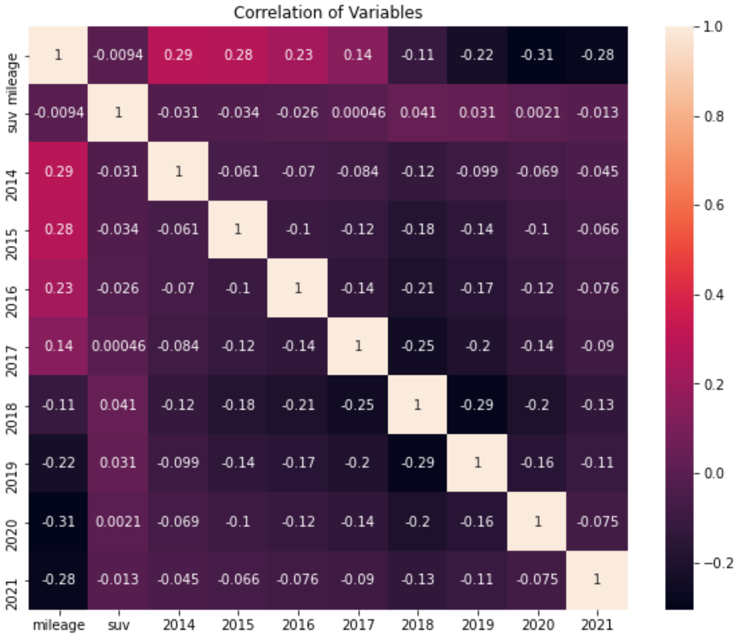
1. Linearity
   a. Data points generally follow a linear pattern.
      i. Could add polynomial terms, apply nonlinear transformations, or add additional variables to improve linearity.



2. Normality of the residuals
   a. Errors are not normally distributed. Confidence intervals will likely be affected.
      i. Could perform nonlinear transformations on variables or remove outliers.

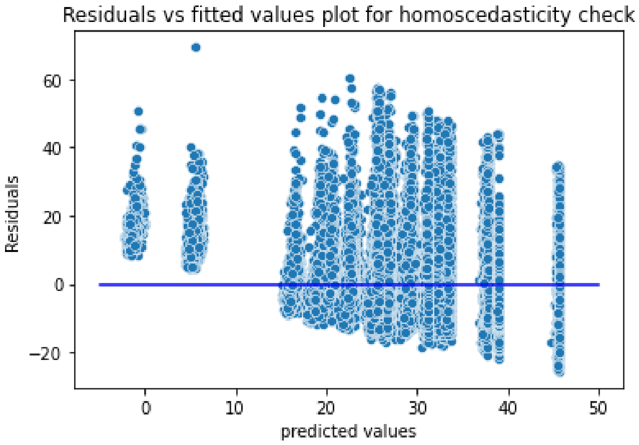


3. Independence (no multicollinearity among predictors)
    a. Satisfied.



4. No residual autocorrelation
    a. Signs of autocorrelation was shown through the Durbin-Watson Test.
        i. Values of 1.5 < d < 2.5 generally show that there is no autocorrelation in the data.
            1. 0 to 2< is positive autocorrelation.
            2. >2 to 4 is negative autocorrelation.
        ii. Could add interaction terms, additional variables or additional transformations.

5. Homoscedasticity of residuals
    a. Satisfied.
        i. No definite pattern can be seen (linear/quadratic/funnel-shape).



        ii. Goldfeld Quandt Test was performed and p value = 0.89 > 0.05.
            1. Cannot reject null hypothesis that residuals are homoscedastic.