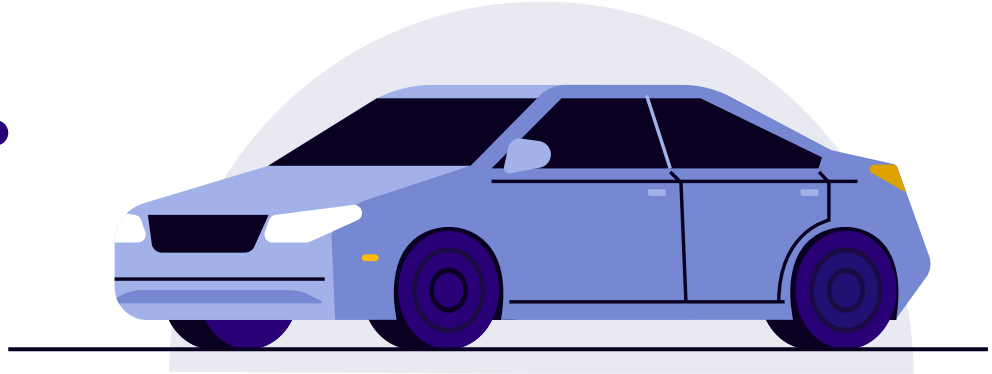
Decomposing Experiential Value from Car Prices

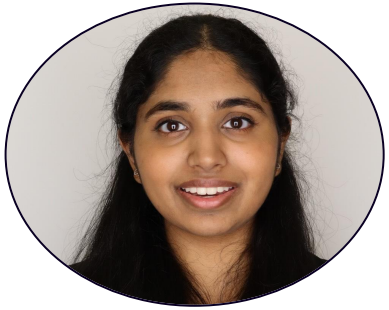
Ethan Wong, Timmy Ren, Neha Boinapalli,
Michael Crosson, Andy Ma



Meet the Team!

Ethan Wong

Timmy Ren



Neha Boinapalli

Michael Crosson

Andy Ma

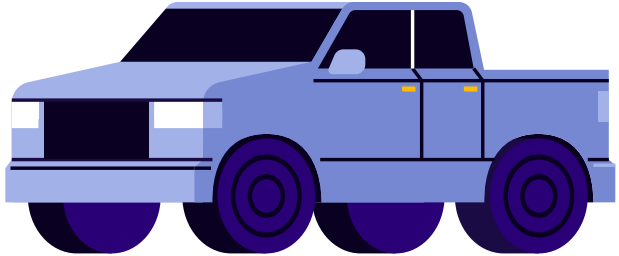
Problem Statement

Can we estimate experiential value component of price based off consumer sentiments?

Table of contents

01.

Data Collection



Where Did We Source Our Data?

Kelley Blue Book

"Since 1926, Kelley Blue Book has been one of the best-known names in the auto industry. Today, KBB.com extends the tradition, with trusted values and a reputation for innovation, including resources to help you research, price and shop for the car you've been looking for."

How Data Collection Was Achieved

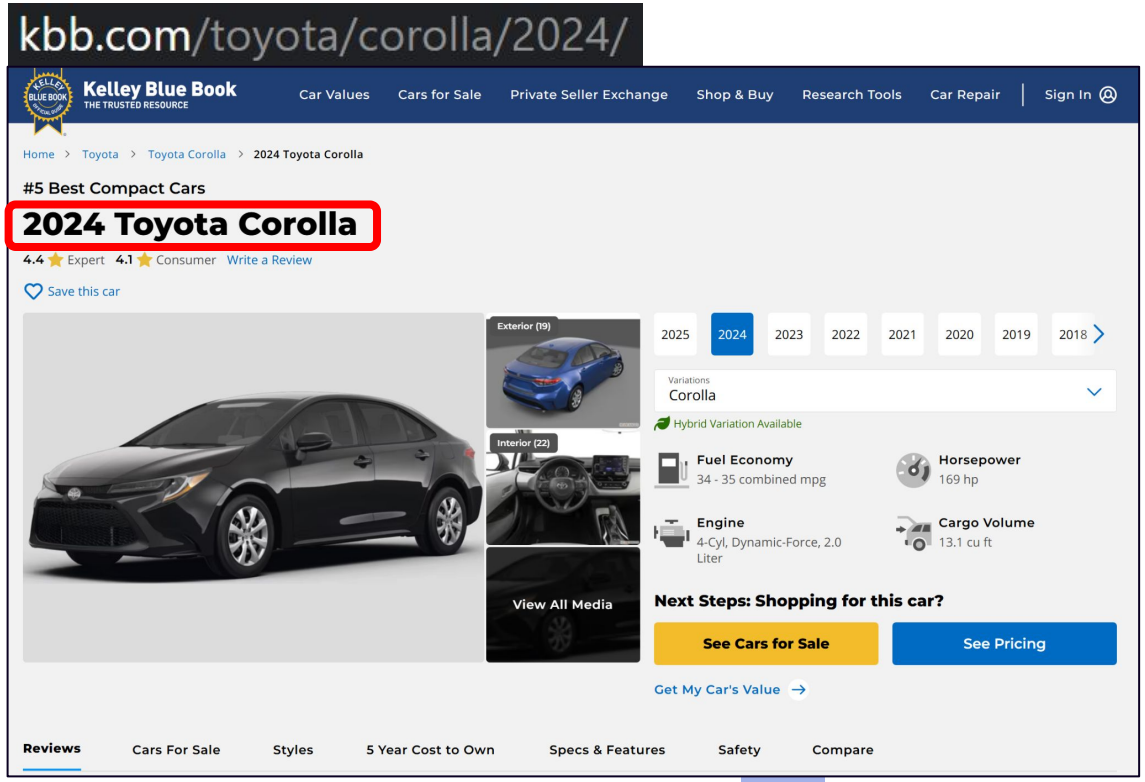
Python Web Scraping

We implemented two powerful libraries for extracting elements and text data from HTML-based websites:

- Selenium - Site task automation and direct text extraction using site content structure
- BeautifulSoup - Parses out relevant text within HTML body; not standalone so it requires a web scraper

***Understand** website content* → ***Analyze** website structure* → ***Input** target elements into web scraping loop* → ***Compile** CSV files of relevant automobile data*

Kelley Blue Book Website Layout



Expert Reviews and Price Data

- Data was extracted for over 175 common/popular cars on the road in the US

2024 Toyota Corolla Review

By **Eric Brandt**
Updated October 21, 2024

Pros

- Lots of standard safety tech
- Great fuel economy
- Good resale value

Cons

- Rivals are roomier

What's New?

- Nightshade trim

The Toyota Corolla upholds its reputation as a safe, reliable, and fuel-efficient compact car while offering modern tech features at an affordable price. Pricing starts at $21,900.

The Toyota Corolla has been in continuous production since 1966. In that time, it's become a benchmark for safe, reliable, and practical personal transportation. Now in its 12th generation, the Corolla builds on that reputation with a modern spin on a proven formula. The styling and technology are thoroughly modern, but the Corolla still impresses on its historic strong suits with excellent safety scores and outstanding fuel economy.

The 50-mpg Corolla Hybrid sedan and sporty, flexible Corolla Hatchback are reviewed separately.

We've spent hundreds of hours driving and evaluating the current collection of compact cars, including this Toyota Corolla.

What's New For 2024

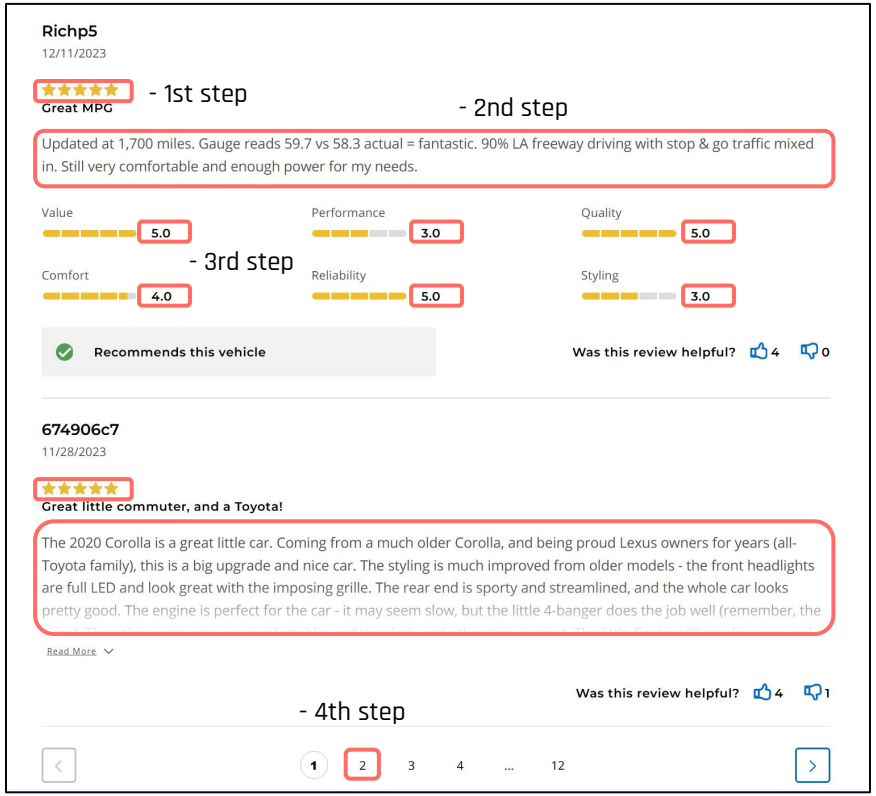
The Toyota Corolla adds the moody Nightshade trim to its lineup for 2024. It's essentially an appearance package for the SE trim with dark exterior accents and stylish bronze-colored wheels.

2024 Toyota Corolla Pricing

The 2024 Toyota Corolla has a starting sticker price of $23,145, with the range-topping Corolla XSE kicking off at $28,245. But Kelley Blue Book Fair Purchase Pricing currently suggests paying $1,167 to $1,293 less than MSRP, depending on trim and equipment. These prices are updated weekly.

	MSRP	KBB Fair Purchase Price (nat'l average)
LE	$23,145	$21,978
SE	$25,585	$24,292
Nightshade	$26,585	$25,404
XSE	$28,245	$26,990

Reviews and Car Ratings (Feature Ratings too!)



Finding Functional Attributes of Car Models

Fetching Car Specifications

- We took car information that seemed to be the most promising for inclusion in our analysis and models
- To ensure data collection within a reasonable time, only some of the specifications are kept
  - Horsepower
  - Curb Weight
  - Fuel Economy
  - And more . . .

Fuel Economy

City	32 mpg
Highway	41 mpg
Combined	35 mpg

Mechanical

Drivetrain	FWD
Transmission Type	Automatic
Recommended Fuel	Regular
Hill Start Assist	Available

Performance

Horsepower	169 @ 6600 RPM
Torque	151 @ 4400 rpm
Engine	4-Cyl, Dynamic-Force, 2.0 Liter

Dimensions, Weights & Capacities

Curb Weight	2955 lbs.

02.

Exploratory Data Analysis

Data Preprocessing

Merging

Web scraping Kelley Blue Book resulted in three CSV files that we were able to easily merge together by the full car model name and year (e.g. Honda Accord 2010)

Missing Values

Rows with missing 'Review' or 'Rating' were removed, and missing specifications were imputed with group medians or overall medians/modes as needed.

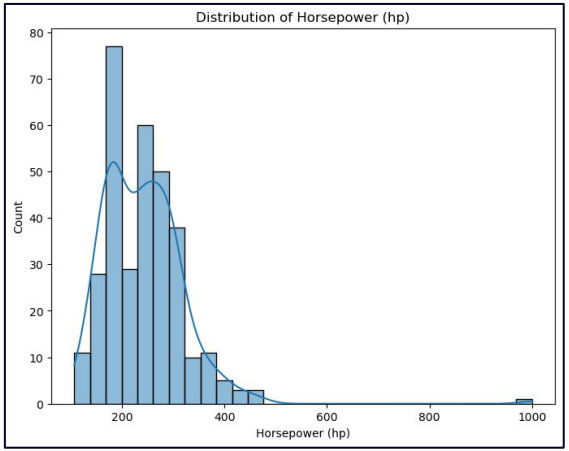
Dropped Columns

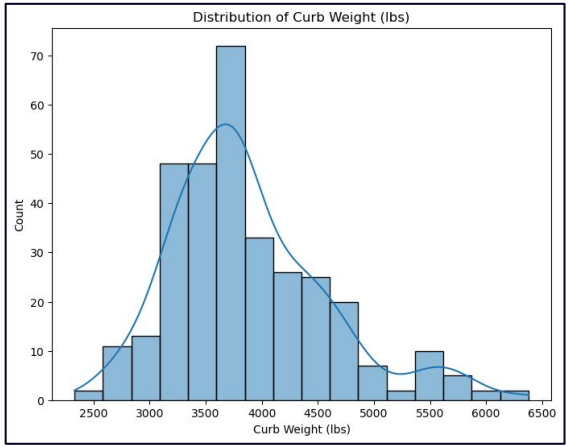
Columns we deemed too sparse or redundant, such as City/Highway MPG and Expert Review/Rating, were removed, simplifying the dataset to focus on key car specifications and ratings.

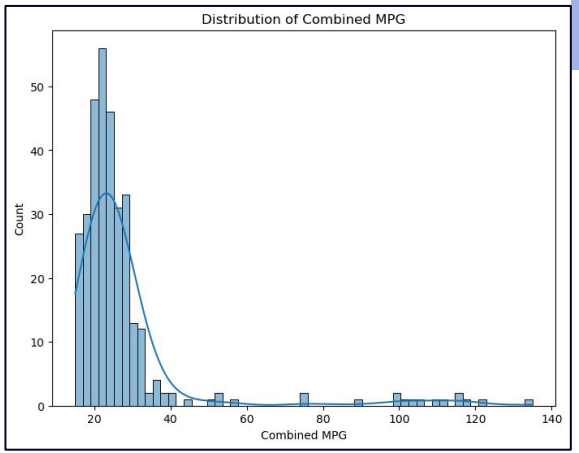
Cleaned Dataset

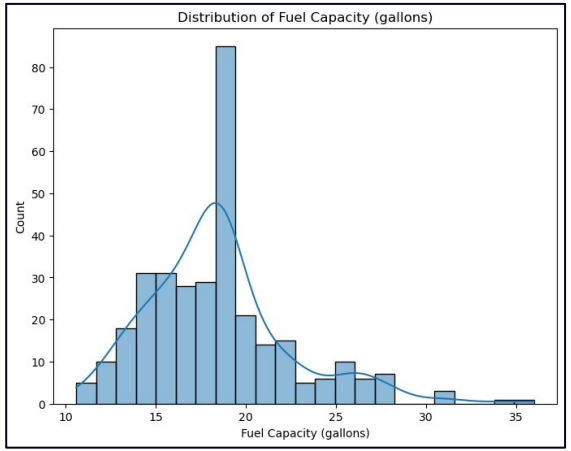
The final dataset consists of 15,260 rows and 22 columns, with each row corresponding to a review made for a given model full name and year.

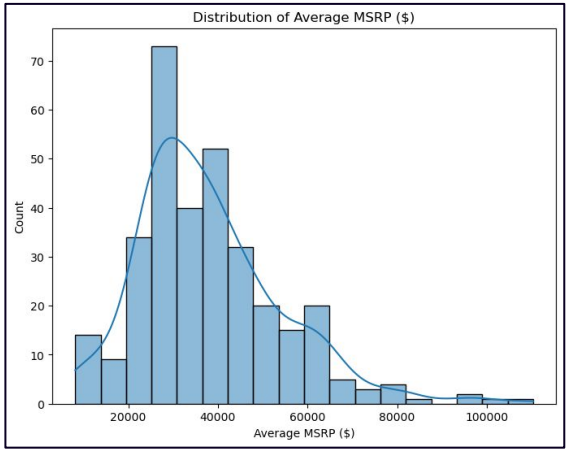
Data Distributions

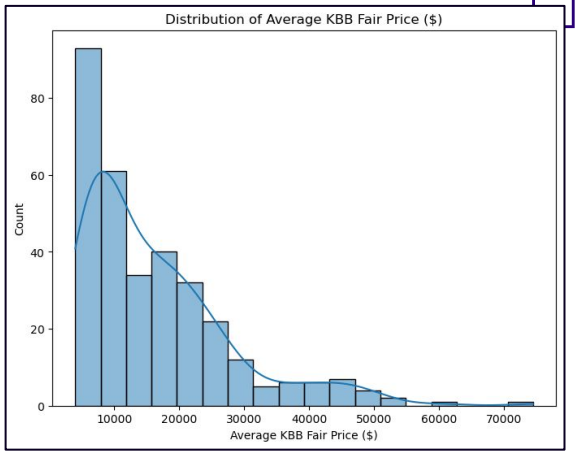




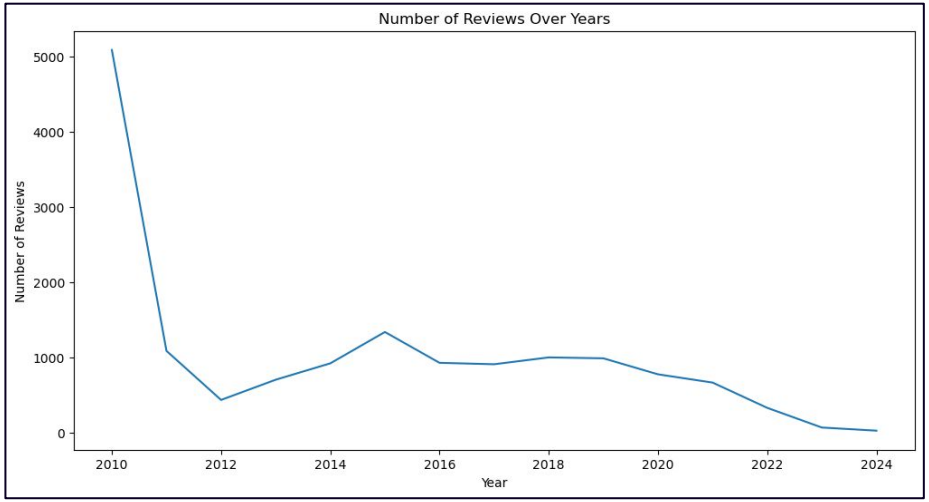






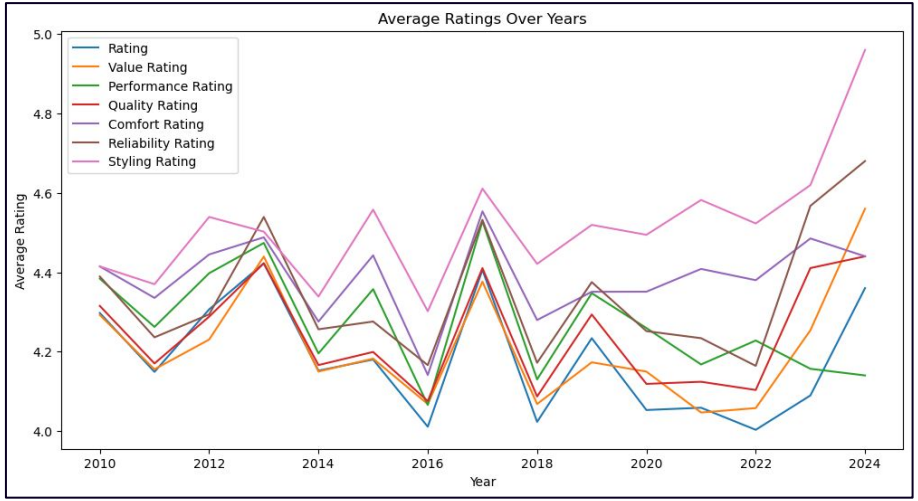


Reviews and Ratings Over the Years

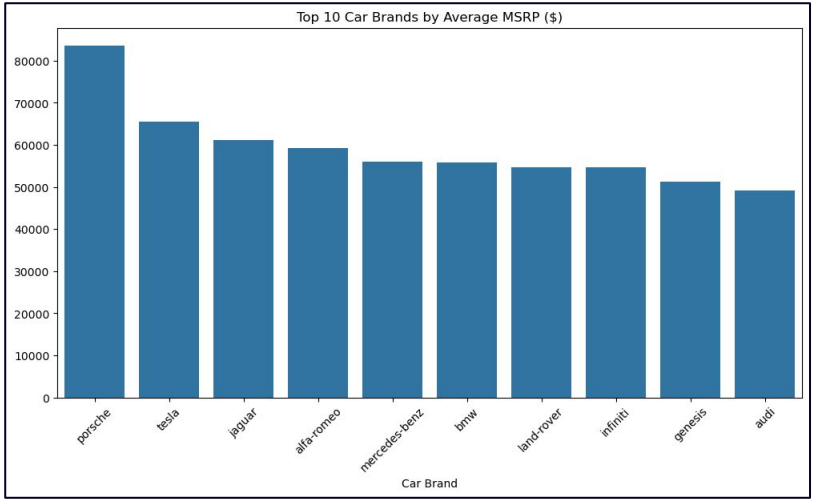


Most of the reviews we were able to scrape came from 2010, with sharp decline in the number of car reviews from 2010 to 2012, followed by a relatively stable but lower review count from 2014 onward.

Average ratings across different aspects (e.g., Value, Performance, Comfort) have fluctuated over the years, with notable peaks around 2012 and a general upward trend in recent years, especially for Styling Rating.



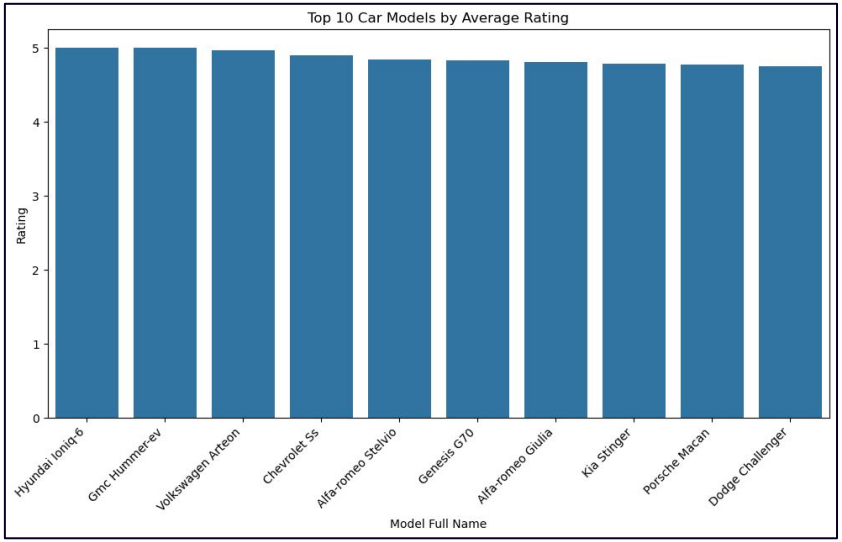
Top Car Brands by Average MSRP and Rating



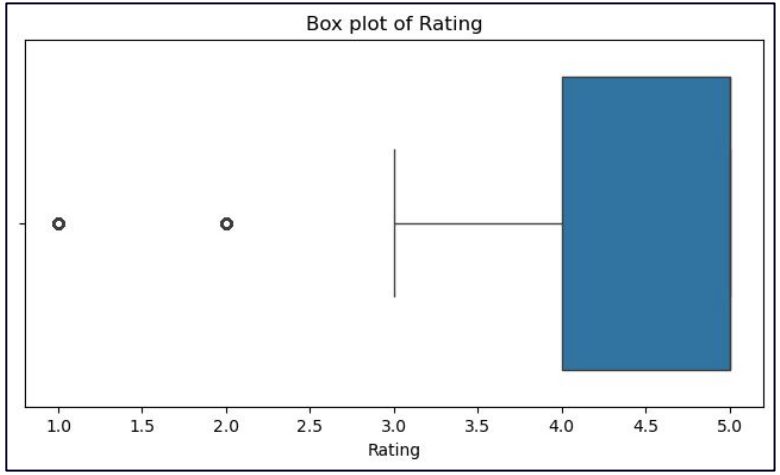
Porsche leads as the brand with the highest average MSRP, followed by Tesla, Jaguar, and other luxury brands, indicating a strong representation of high-end brands in the top 10.

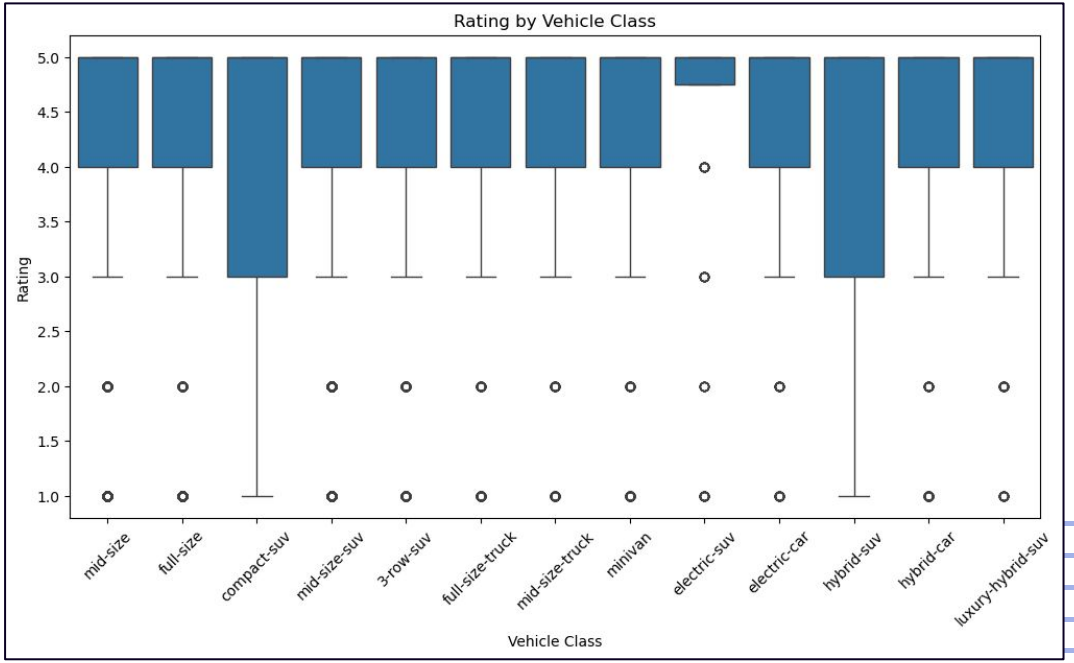
Models like the Hyundai Ioniq 6 and GMC Hummer EV achieve near-perfect ratings, with the top 10 car models all having consistently high average ratings close to 5.

However, this isn't limited to just these 10 car models...



Skewed Reviews



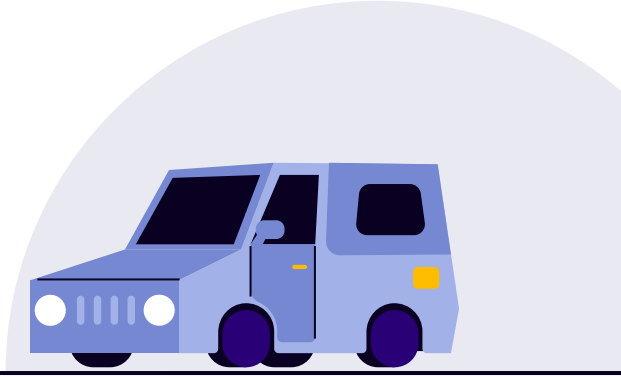


The majority of ratings are high, with a median around 4.5, while a few outliers exist below 3, showing that most vehicles receive favorable ratings. The same story applies when we break down the ratings by vehicle class.

The distribution of ratings across different aspects were essentially identical to the above plots.

03

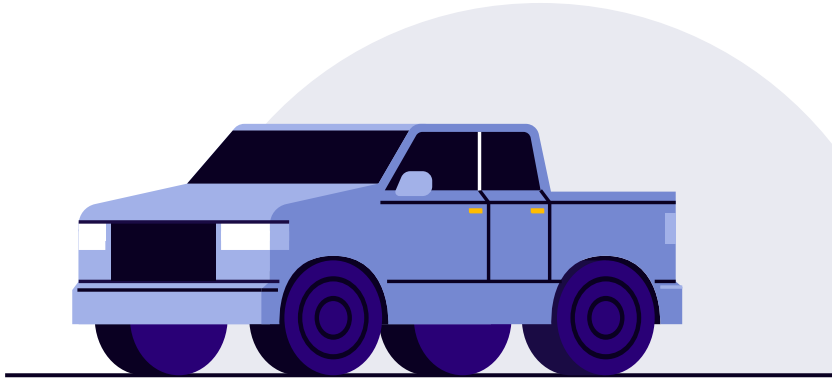
Topic Modeling



Topic Analysis - Steps

We have a lot of information trapped in reviews, how do we get it out?

1. Extract the Reviews
2. Remove Stopwords
3. Term Frequency-Inverse Document Frequency (TF-IDF)
4. Choose and Run the Model



Topic Analysis - Choose the Model

We have a lot of information trapped in reviews, how do we get it out?

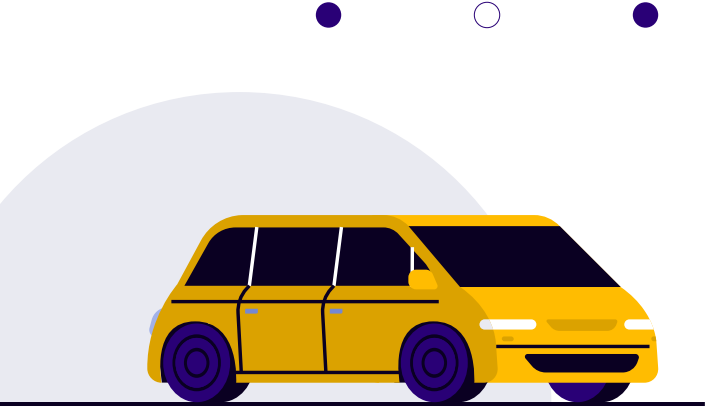
Latent Dirichlet Allocation (LDA)
Non-negative Matrix Factorization (NMF)

Bidirectional Encoder Representations from Transformers (BERT)

Latent Semantic Analysis (LSA)

Topic Analysis - Choose the Model

We have a lot of information trapped in reviews, how do we get it out?



Latent Dirichlet Allocation (LDA)

Non-negative Matrix Factorization (NMF)

Bidirectional Encoder Representations from Transformers (BERT)

Latent Semantic Analysis (LSA)

Topics After TF-IDF + NMF

0. **Interior/Features**
- Seats, Interior, System

3. **No Technical Issues**
- Oil, Years, Engine, Never, Issues

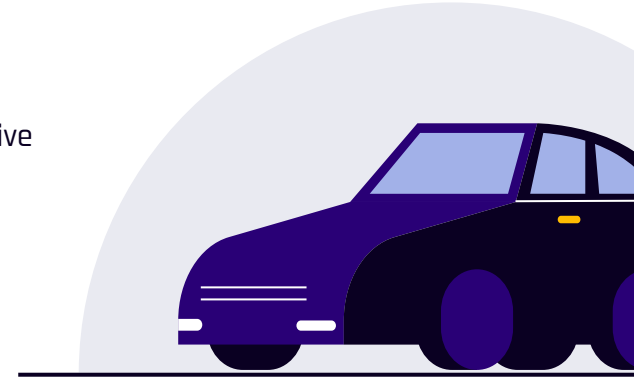
4. **Best Car Ever**
- Best Car, Ever Owned

5. **Trucks!**
- Great Truck, Truck, Bed, Cab

7. **Great Gas Mileage**
- Gas, Gas Mileage, Good Gas

8. **Recommend for Family**
- Recommend, Value, Family, Reliable

9. **Fun to Drive**
- Drive, Fun, Wheel Drive

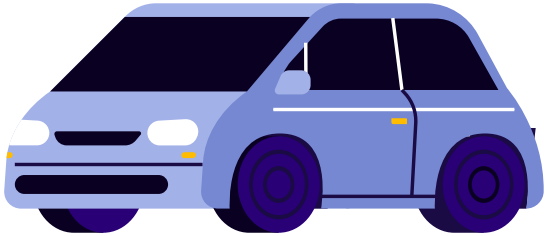


Example Output of Topic Analysis

	Car Brand	Car Name	Review	. . .	Topic 0	Topic 3	. . .	Topic 9
Row 1	Honda	Accord	"I love my car..."	. . .	0.00321	0.06835	. . .	0.04891
Row 2	Toyota	Corolla	"This car sucks..."	. . .	0.01023	0.00000	. . .	0.00265

04.

Regression Analysis

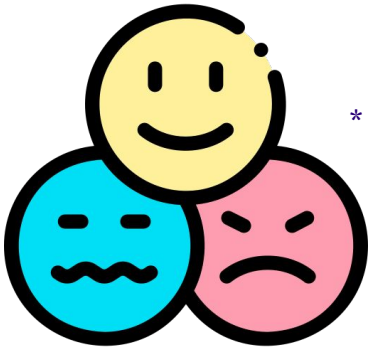


Data Split: Functional Features

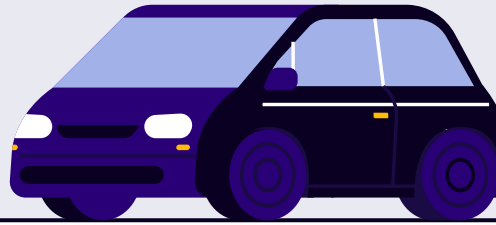
- **Horsepower** (hp)
- **Curb Weight** (lbs)
- **Combined MPG**
- **Fuel Capacity** (gallons)
- **Age**
- **Drivetrain** (RWD, FWD, AWD, 4WD, 2WD) - one-hot encoding
- **Vehicle Class** (electric SUV, hybrid car, mid-size SUV, etc.) - one-hot encoding

Data Split: Experiential Features

- Calculated **Sentiment Scores** on a per Review basis using *Vader*
- Determined **'topic_#_weighted'** (Topic 0 - 9)
  - Multiply Sentiment of Review by the Topic of the Review
- **Car Brand** - ex. Honda
  - Model Full Name ex. Honda Accord
    - Lead to overfitting on aggregated dataset (more on this later .....)

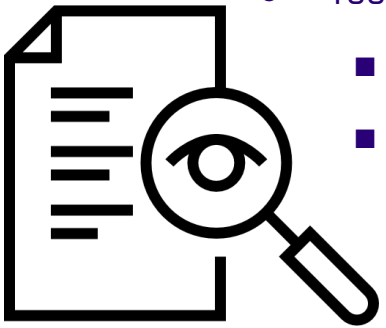


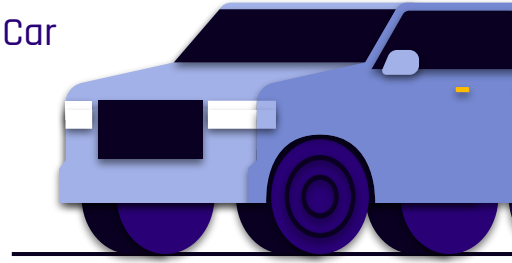
* *We did not include ratings as an experiential feature*



Unaggregated v. Aggregated Dataset

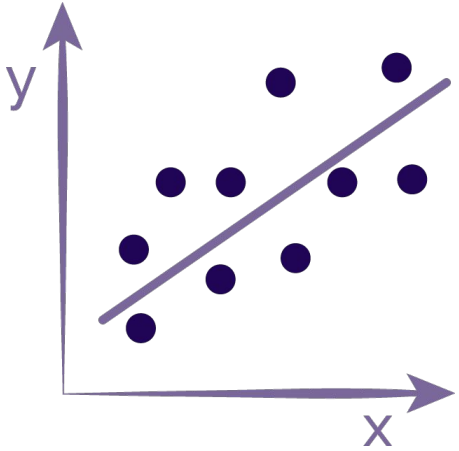
- **Unaggregated:**
  - This is on a per-review basis regardless of the Model Full Name
- ***Aggregated:**
  - Groupby *Model Full Name* and *Year*
  - Took the mean values/ classification for the rest of the parameters
    - Ended up with 326 rows of data (unique brand model year combos)
    - 62 potential predictors (including the dummy variables for Car Brand, Drivetrain, Vehicle Class)





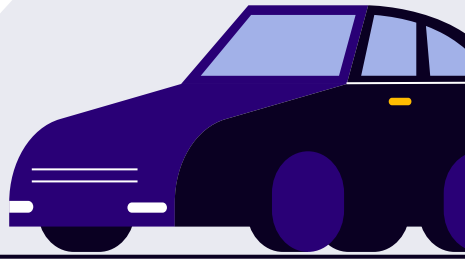
Linear Regression Calculations

- **3 Types of Linear Regression**
  - Linear Regression
  - Lasso (Regularization L1)
  - Ridge (Regularization L2)
- **Assumptions**
  - Did 80/20 Training/Test Set Split
  - Alpha Assumption = [0.1, 1.0, 10.0, 20, 50, 75, 100]
    - For loop through alphas to determine best alpha based on RMSE



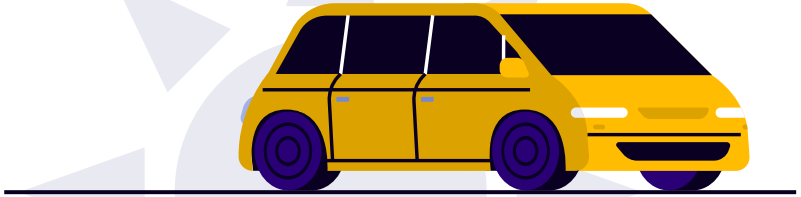
Results:

	Division Title	Model Type	Best Alpha	R^2	RMSE
0	Functional Unaggregated	Linear Regression	NaN	0.78	4123.80
1	Functional Unaggregated	Lasso	0.1	0.78	4123.79
2	Functional Unaggregated	Ridge	0.1	0.78	4123.80
3	Functional Aggregated	Linear Regression	NaN	0.64	5149.06
4	Functional Aggregated	Lasso	50.0	0.67	4940.99
5	Functional Aggregated	Ridge	10.0	0.66	4981.18
6	Functional & Experiential Unaggregated	Linear Regression	NaN	0.78	4119.96
7	Functional & Experiential Unaggregated	Lasso	0.1	0.78	4120.22
8	Functional & Experiential Unaggregated	Ridge	1.0	0.78	4119.90
9	Functional & Experiential Aggregated	Linear Regression	NaN	0.64	5125.43
10	Functional & Experiential Aggregated	Lasso	100.0	0.71	4650.99
11	Functional & Experiential Aggregated	Ridge	100.0	0.75	4284.02

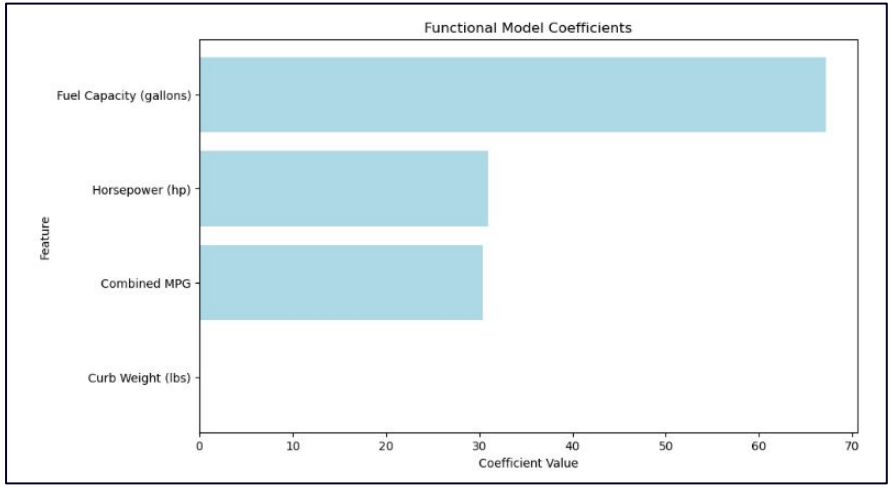


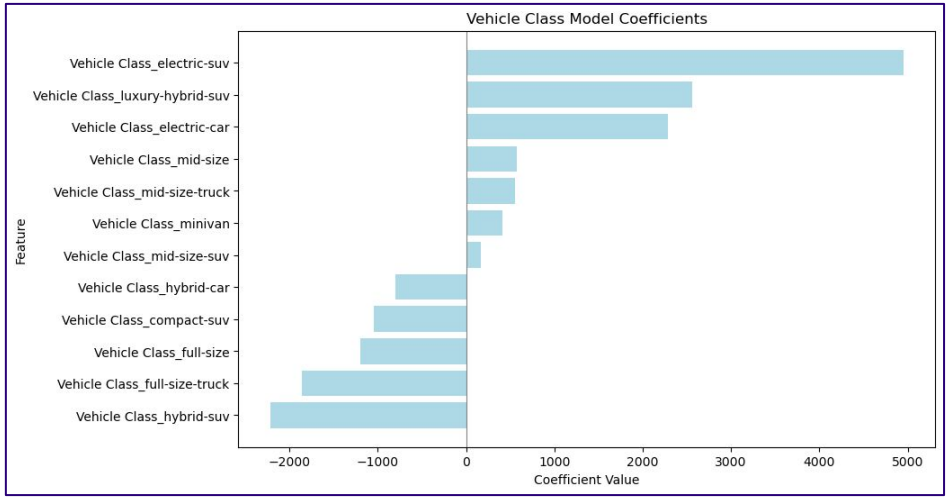
Aggregated Results:

	Division Title	Model Type	Best Alpha	R^2	RMSE
Row 3	Functional	Linear Regression	NaN	0.64	5149.06
Row 4	Functional	Lasso	50	0.67	4940.99
Row 5	Functional	Ridge	10	0.66	4981.18
Row 9	Functional + Experiential	Linear Regression	NaN	0.64	5125.43
Row 10	Functional + Experiential	Lasso	100	0.71	4650.99
Row 11	Functional + Experiential	Ridge	100	0.75	4284.02



Functional Coefficients

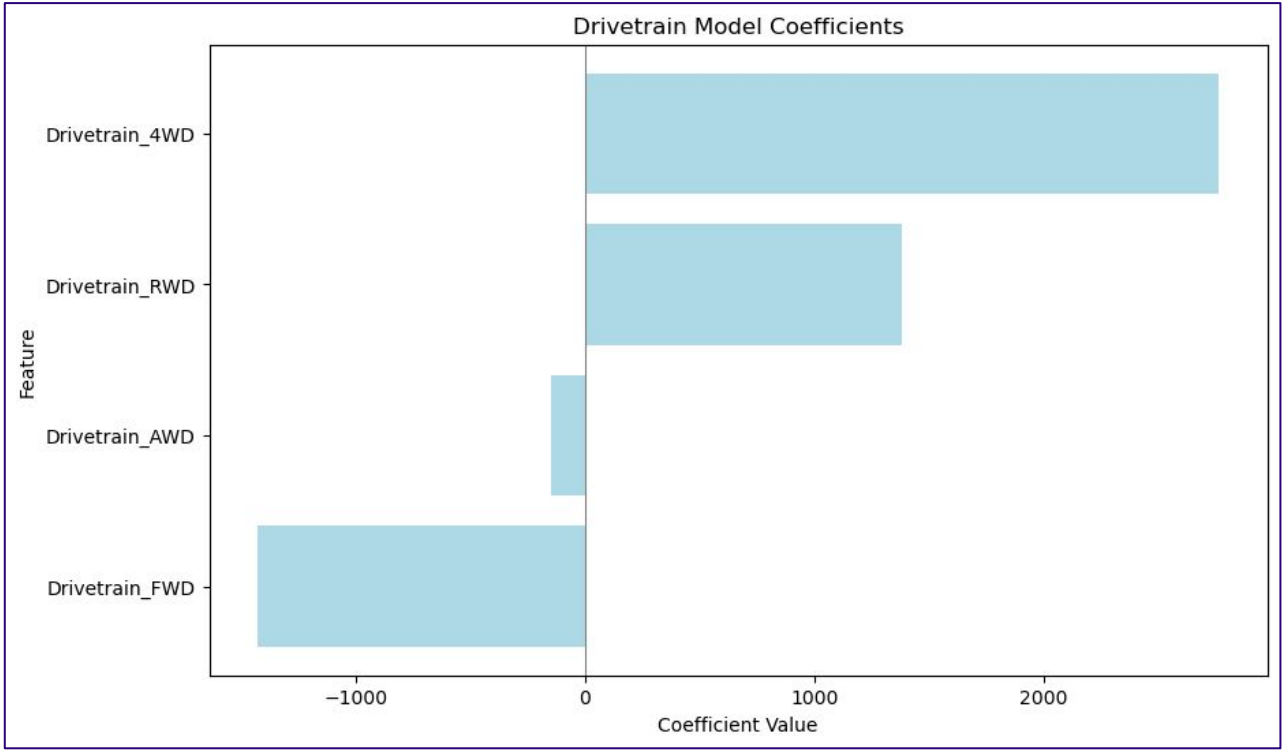




*Baseline value is 3-row-suv

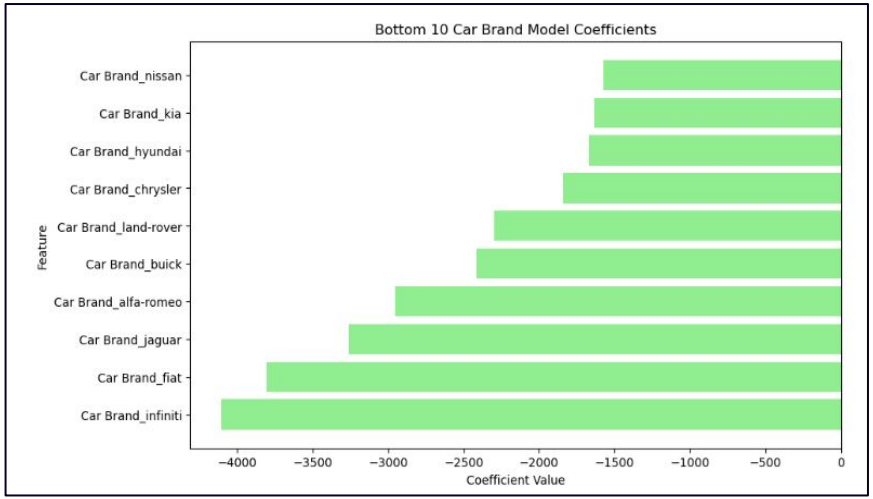
- Age Coefficient: -828.15

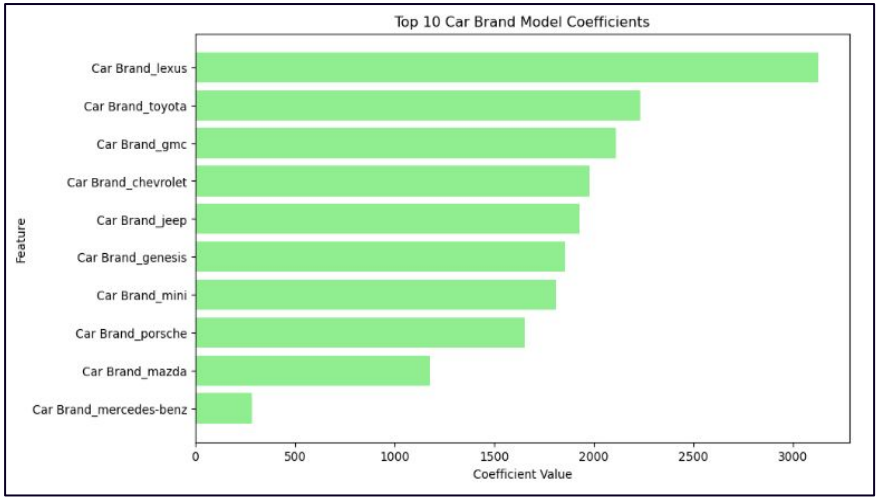
Functional Coefficients Pt. 2



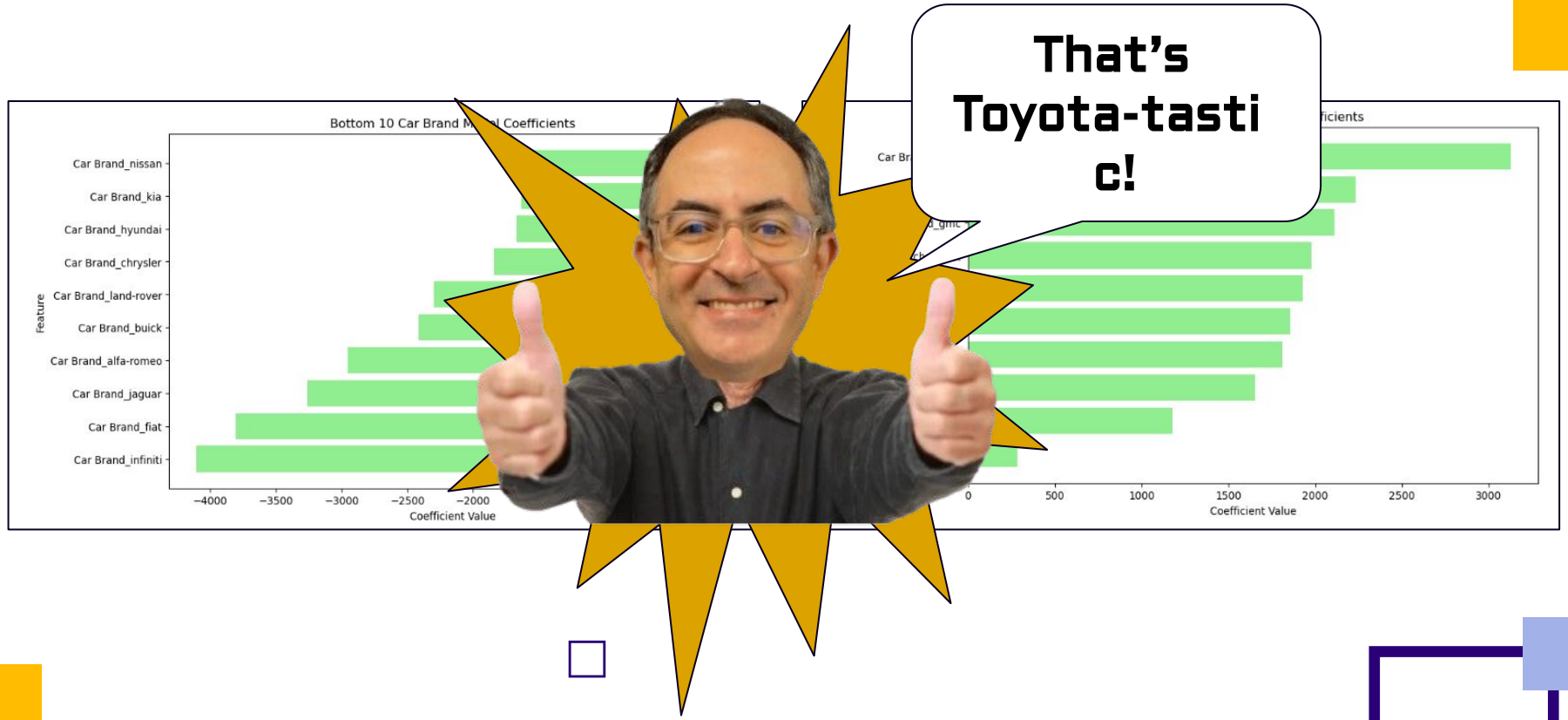
*Baseline value is 2WD

Experiential: Brand Coefficients

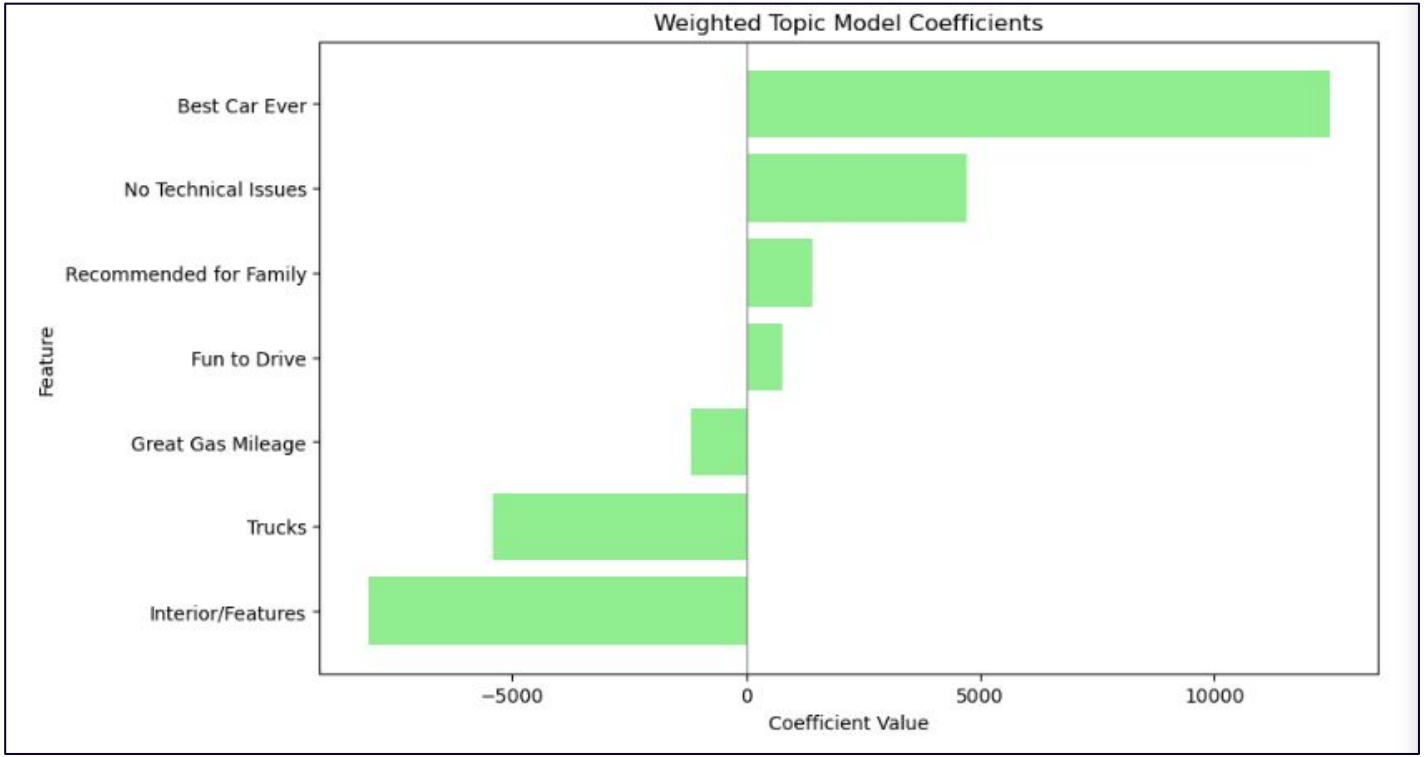




Experiential: Brand Coefficients



Experiential: Topic Sentiments

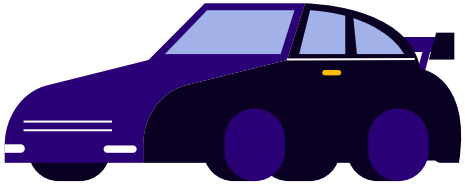


Limitations

- Voluntary Response Bias - Data is not representative of the true population, and the reviews are much more focused on people with more polarized sentiments
- General Sentiments Instead of Topic Based Sentiments
- Disproportionate Data
  - Some **cars** have very few **reviews** relative to each other
  - Some **brands** have very few **cars** relative to each other
- Reviews may be about specific trims, our data uses the base trim and averaged pricing

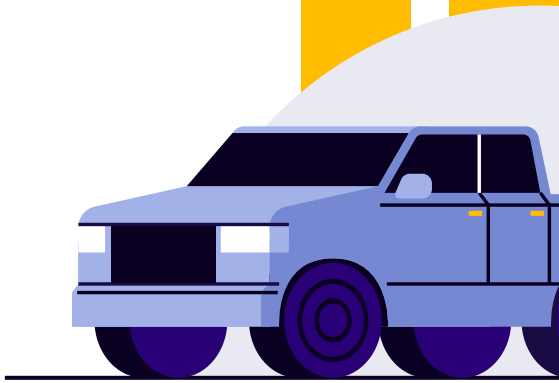
05.

Conclusion



Conclusion

- **Takeaways:**
  - Experiential Values Do Exist!
  - Beyond brand, the weighted topic sentiments do contribute a significant amount to the price
- **Future Application:**
  - Assign Prices to Concept Cars
  - Can be combined with a recommender system

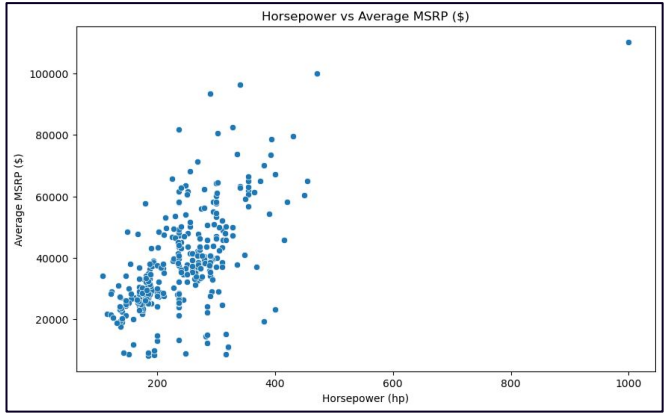


Thanks!

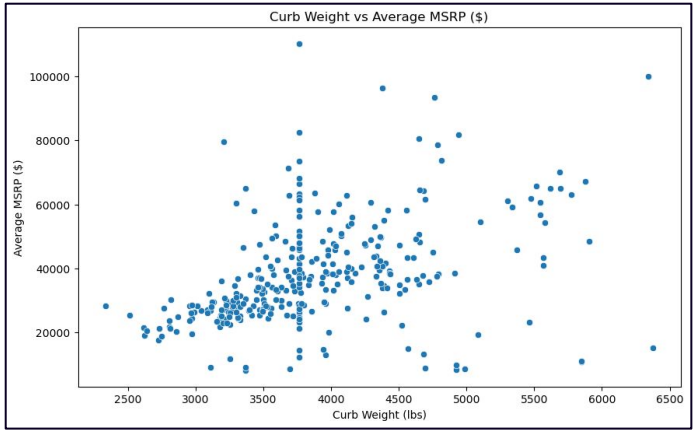
We are open to any questions!

Appendix/Additional Analyses

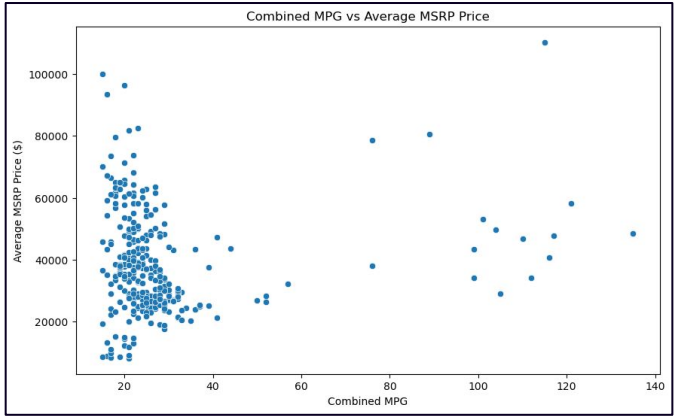
Average MSRP Scatterplots



There is an inverse relationship between combined MPG and MSRP, with most high-MPG vehicles having lower MSRP, while a few high-MPG outliers have a higher MSRP.

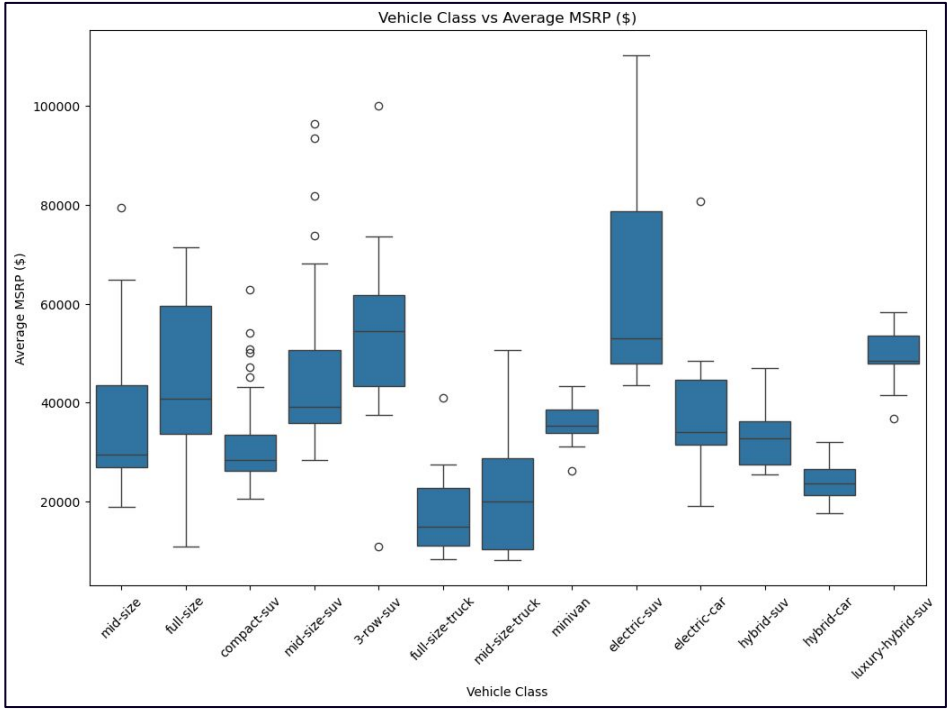


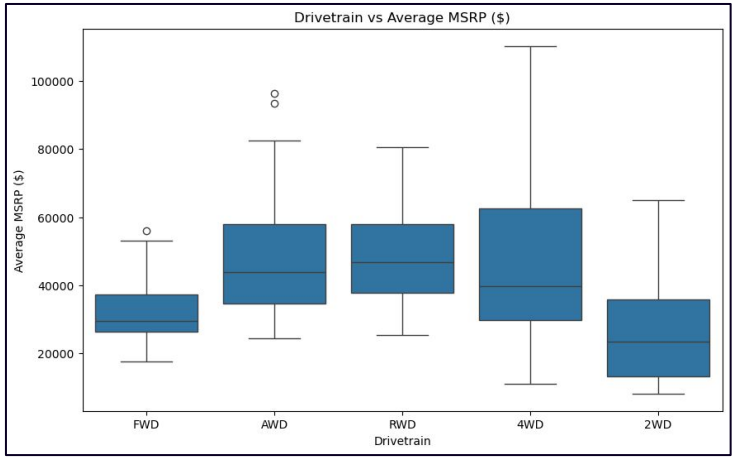
There is a positive correlation between horsepower and MSRP, with higher horsepower generally associated with higher MSRP, though there are a few high-MSRP outliers.

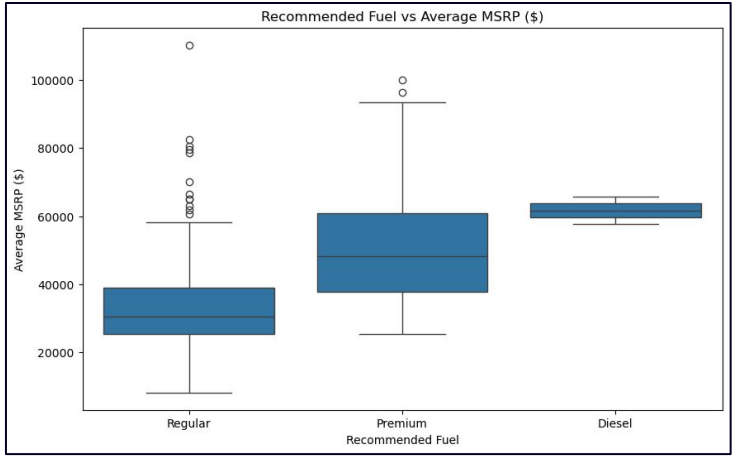


Curb weight shows a moderate positive relationship with MSRP, with more variation in price as curb weight increases, but there is no strong linear trend.

Average MSRP Boxplots







Correlation Matrices