DATA-DRIVEN PRICING: PROMOTING FAIRNESS IN THE USED CAR MARKET

BY: TORRODJAE SOMERVILLE

INTRODUCTION

- **Topic:** Promoting Fairness in the Used Car Market through Data-Driven Pricing.
- **Research/Data Questions:** (You have these, just list them here)
  - What is the relationship between a vehicle's age, mileage, and its listing price?
  - How do vehicle type and condition influence pricing trends and volatility?
  - What are the key features that signal a "good deal" or a potential premium?
- **Objectives & Goals:** To uncover the true factors that determine a used car's value, identify pricing anomalies, and empower buyers and sellers with transparent, data-backed insights.
- **Why this topic?** The used car market is massive and often opaque. Consumers can easily overpay or undervalue their vehicles. Data can help demystify this process and promote fairer transactions.

DATA INGESTION - OVERVIEW

- **Source & Creator:** The dataset is called "Used Cars Dataset". It was put together by Austin Reese, a data scientist, and shared on Kaggle.
- **Where It Came From:** The data was scraped from Craigslist listings to give a clear look at the U.S. used car market.
- **How It Was Collected:** Reese used a web scraper to pull listing details, then cleaned and organized them into one dataset.
- **Frequency:** Craigslist updates all the time, but this dataset is just a snapshot from one scraping point in 2022.
- **Why It Was Collected:** To study prices, trends, and help build models that can predict car values. What's Inside: Info on prices, makes and models, car condition, mileage, fuel types, transmissions, and locations.

THE RAW DATA

```
================================================
DATA INGESTION PHASE
================================================
1. Original Dataset Shape (Rows, Columns): (426880, 26)

2. Original Columns:
 ['id', 'url', 'region', 'region_url', 'price', 'year', 'manufacturer', 'model', 'condition', 'cylinders', 'fuel', 'odometer', 'title_status',
'transmission', 'VIN', 'drive', 'size', 'type', 'paint_color', 'image_url', 'description', 'county', 'state', 'lat', 'long', 'posting_date']

3. Model Year Range in Raw Data: 1900.0 to 2022.0

4. The first row (instance) represents this:
 id                                                7222695916
url            https://prescott.craigslist.org/cto/d/prescott...
region                                              prescott
region_url                       https://prescott.craigslist.org
price                                                   6000
year                                                     NaN
manufacturer                                             NaN
model                                                    NaN
condition                                                NaN
cylinders                                                NaN
fuel                                                     NaN
odometer                                                 NaN
title_status                                             NaN
transmission                                             NaN
VIN                                                      NaN
drive                                                    NaN
size                                                     NaN
type                                                     NaN
paint_color                                              NaN
image_url                                                NaN
description                                              NaN
county                                                   NaN
state                                                     az
lat                                                      NaN
long                                                     NaN
posting_date                                             NaN
Name: 0, dtype: object
```

- **Size:** 426,880 rows, 26 columns
- **Scope:** Listings from all model years (1948-2023)
- **What's in a row?** Each row is one car listing.
- **Problem here:** Data is messy – contains errors, possible duplicates, and outliers.
- NaN = variable may need to be changed

FOCUSING THE DATASET FOR ANALYSIS

Problem: The full dataset is too broad and had errors from it's collection.

Solution: I created a focused, clean subset.

Filters Applied:

- **Model Years: 1948 to 2023** (excluded classic cars per Classic Car Club of America guidelines)
- **Price: $500 to $90,000** (focused on mainstream market, excluded larger outliers caused by classics)
- **Dropped rows missing critical data** (price, manufacturer, year, odometer)
- **Removed unrealistic odometer readings** (> 0 miles)
- **Standardized text fields** (manufacturer, model names to lowercase)
- **Consolidated vehicle type** names that tripped up data (mini-van and minivan)
- **Consolidated condition categories** from 8+ values to 5 clear levels (fair, salvage,.etc)
- **Converted variables to proper data types** (price, year, odometer to numeric)

Result: A higher-quality dataset for accurate analysis.

DATA WRANGLING – MY CLEANING

Handled Duplicates:

- Removed 0 identical listing entries initially. But found 739 when checking vins.
- Showed the vin values that needed to be fixed. Before used for comparisons.

Fixed Data Types:

- Converted price from text ($10,000) to a float (10000.0)
- Odometers converted to integers.

Filled Missing Values:

- Filled blank condition and fuel fields with "unknown".
- Dropped fields missing critical values.

Filtered Outliers:

- Removed junk listings ($0 cars, really high outlier prices).
- Kept age modern to avoid price alterations due to outside influence (classic cars)

KEY DATA FEATURES

- **Target Variable:**
  - **Price:** The listing price (USD). This is what we want to understand and predict.
- **Core Predictive Features:**
  - **Year:** Age is a primary driver of value. Newer models command higher prices.
  - **Manufacturer:** The brand (Toyota, BMW) is a key indicator of cost, reliability, and prestige.
  - **Model:** The specific model name (F-150, Camry) defines the vehicle's type and market segment.
  - **Condition:** The state of the vehicle (excellent, good, fair) directly impacts its worth.
  - **Odometer:** Mileage is a direct measure of use and wear, a critical factor for valuation.
  - **fuel:** Fuel type (gas, hybrid, electric) affects operating costs and desirability, influencing price.

THE FINAL PRODUCT

```
=== DATA CLEANING SUMMARY ===
Original dataset: 426,880 rows, 26 columns
Final dataset: 365,671 rows, 11 columns
Total rows removed: 61,209

=== BREAKDOWN OF REMOVED ROWS ===
• Rows removed due to missing data: 21,803
• Duplicate listings removed: 0
• Rows filtered by price (<$500 or >$90,000): 43,036
• Rows filtered by year (kept only 1948-2023): 1,024
```

- **Final Size:** Final dataset: 365,671 rows × 11 columns (14% reduction from original)
- **Data Quality:** High quality missing duplicates (found using the vin#) as well as now having uniform data a little more specific to what I'm doing. (Like modern model years)
- Much more manipulation friendly now.

SUMMARY STATISTICS: SHOWING MY DATA

```
=== FILTERING CRITERIA APPLIED ===
• Price range: $500 - $90,000
• Model years: 1948 - 2023
• Removed missing critical fields (price, year, manufacturer, odometer)
• Removed duplicate listings
• Selected only relevant columns for analysis

=== FEATURE ENGINEERING ===
Engineered Features Created:
- age (2023 - year)
- miles_per_year (odometer / age)
- condition_clean (5 standardized categories)
- price_category (Budget/Mid-Range/Luxury)

=== FINAL PRODUCT DATASET ===
Final Size: 365,671 rows × 15 columns
Data Quality: High quality - removed duplicates, missing data, and unrealistic values
Scope: Modern cars (1948-2023) at realistic prices ($500-$90,000)

=== SUMMARY STATISTICS ===
Average price: $19085.44
Average model year: 2011
Average mileage: 97424
Average vehicle age: 11.6 years
```

- **Caption:** Summary of key numerical fields after cleaning.
- **Insights:**
  - The average car price is ~$19,085.44
  - The average model year is 2011.
  - The average mileage is 97,460.
  - Prices found here will be slightly inflated compared to now due market change over the years.

THE ANALYTIC DATASET AND CLEANING

- **Source:** Austin Reeves' "Used Cars Dataset" from Kaggle.
- **Initial Size:** 426,880 rows × 26 columns
- **Final Size:** 365,671 rows × 15 columns
- **Cleaning Steps:**
  - Removed unrealistic prices (<$500, >$90,000)
  - Removed classic cars (model year ≤ 1948) per Classic Car Club of America guidelines
  - Filtered out extreme odometer readings
  - Dropped rows with missing critical data

FEATURE ENGINEERING

age: Created from year (2023 - year). A more intuitive measure for depreciation.

Why This Was Essential:

- **Business Insight:** Cars depreciate with age, not model year - a 2015 car is worth less in 2023 simply because it's older
- **Analytical Value:** Age creates a linear relationship with price that model year alone doesn't capture
- **Real-world Application:** Insurance companies and lenders use vehicle age, not model year, for valuation models

```
# Creating age feature for comparisons and insight
df_clean['age'] = 2023 - df_clean['year']
```

miles_per_year: Created from odometer / age. (Identifies excessively or sparingly used vehicles.)

Why I felt it was needed:

- **Critical Insight:** Total mileage alone is misleading - 100,000 miles over 2 years vs. 8 years indicates completely different wear patterns
- **Risk Assessment:** High miles_per_year (>15k) suggests highway/rental/fleet use with more mechanical stress
- **Value Preservation:** Low miles_per_year (<8k) often indicates well-maintained, garage-kept vehicles that hold value better

```
# Create miles_per_year (handle division by zero)
df_clean['miles_per_year'] = df_clean['odometer'] / df_clean['age'].replace(0, 1)
```

FEATURE ENGINEERING CONT.

- **condition_clean:** Consolidated 8+ original values into 5 clear categories (Salvage, Fair, Good, Excellent, Like New / New too many subsections of samething).
- **price_category:** Binned prices using Edmunds' used car classification.

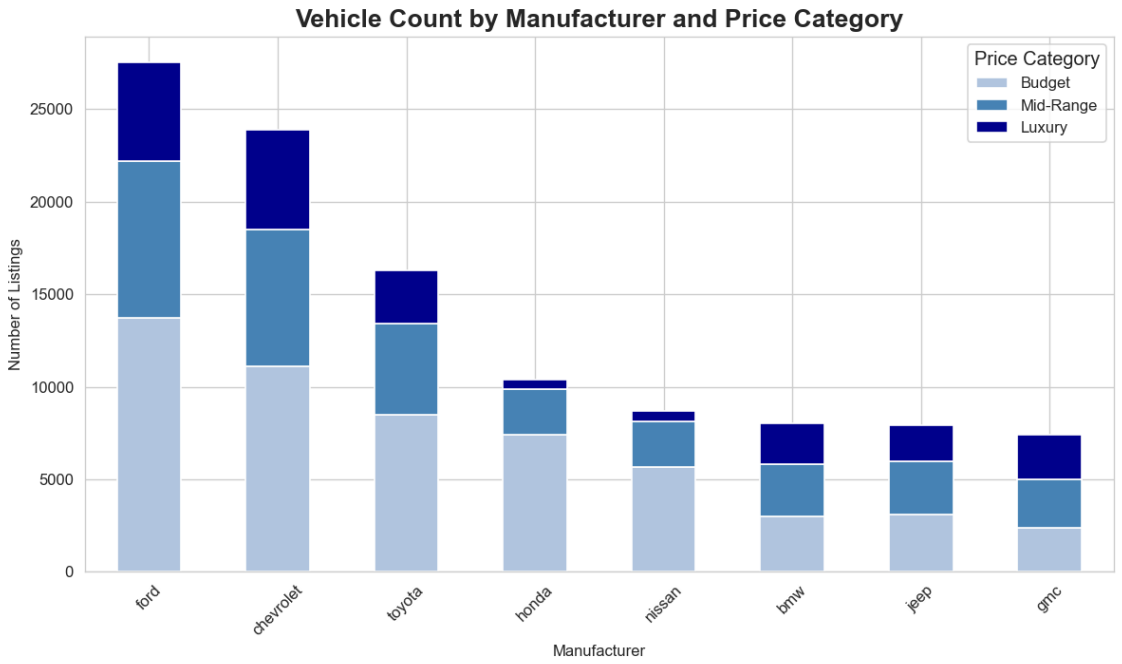
Why I felt categorizing the prices was needed

- **Binning Method:** Applied supervised binning using industry-standard price thresholds from Edmunds
- **Market Segmentation:** Created three strategic price bins (Budget, Mid-Range, Luxury) to analyze brand positioning
- **Analytical Power: This binning process** transformed continuous price data into categorical segments, enabling clear market tier analysis
- **Consumer Insight: The price binning** reveals how different buyer segments shop across price ranges

```python
# Create price categories
def categorize_price(price):
    if price <= 15000:
        return 'Budget'
    elif price <= 30000:
        return 'Mid-Range'
    else:
        return 'Luxury'

df_clean['price_category'] = df_clean['price'].apply(categorize_price)
```

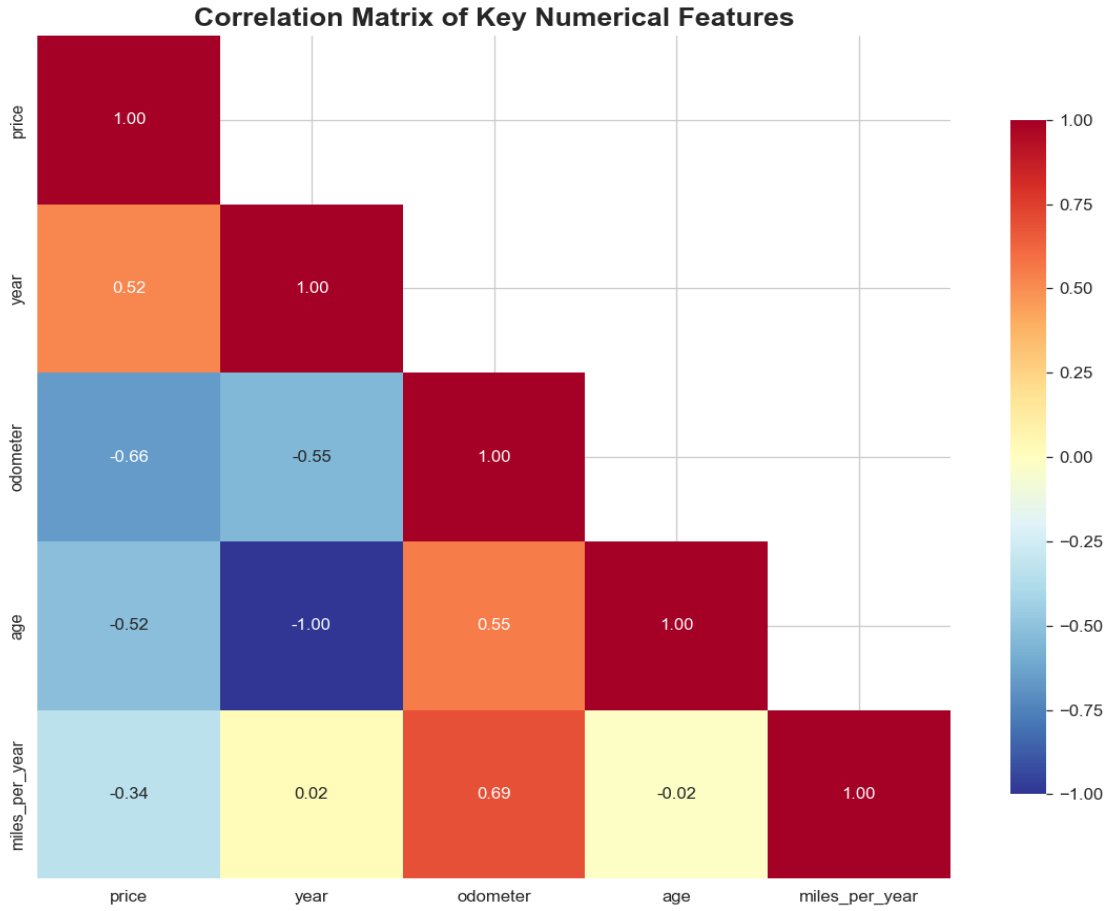
FEATURE ENGINEERING - PRICE CATEGORIES (EDMUNDS)



- created a new feature, **price_category**, by binning prices according to **Edmunds' used car classification**:
  - **Budget:** $0 - $15,000
  - **Mid-Range:** $15,000 - $30,000
  - **Luxury:** $30,000+
- *Source: Edmunds, "Used Car Valuation: How Much Is Your Car Worth?"*
- **Key Insight:** This chart reveals clear brand positioning. (top 8 brands)
  - **Ford** and **Chevrolet** dominate the **Budget** and **Mid-Range** segments.
  - **BMW** consists of a considerable luxury section given it's size the **Luxury** segment.
  - **Honda** and **Nissan** show a trend, presence across the categories. Showing they don't have many luxury portions but are made mostly of budget vehicles this could be a key area to give insight on what manufacturer to look into based on a customers price range. (combine mileage and price to see which vehicles hold more value over time)

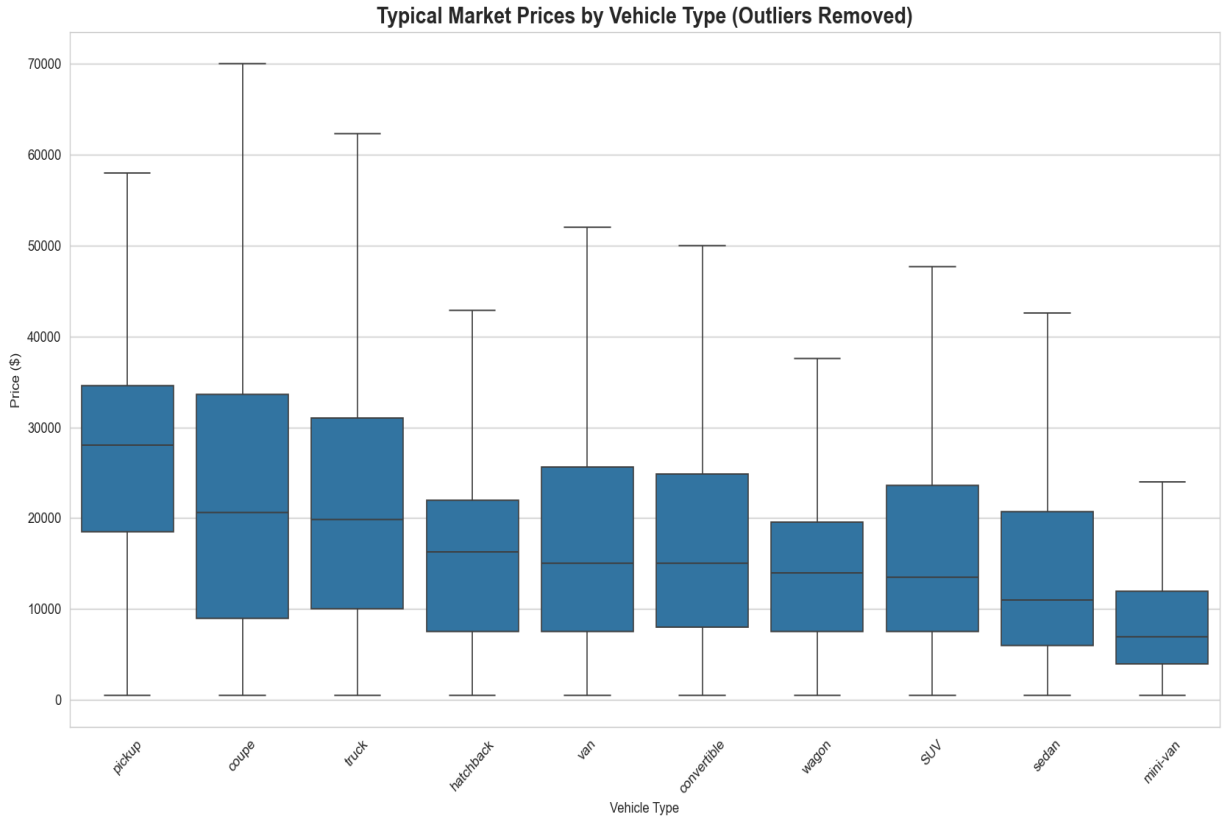
FINDING #2 - THE CORE RELATIONSHIPS

- **Price vs. Odometer (-0.66):** The strongest negative correlation. Higher mileage is the biggest driver of lower prices.
- **Price vs. Age (-0.52):** A strong negative correlation confirming vehicle depreciation over time.
- **Price vs. Miles/Year (-0.34):** Our engineered feature shows a clear negative relationship, showing overused vehicles. **A negative value means** that as the average number of miles driven per year **increases**, the vehicle's price **decreases**.
- This proves that our new feature successfully captures a car's **usage intensity**. It's no longer just about *how many* miles are on the odometer, but *how quickly* those miles were put on the car.



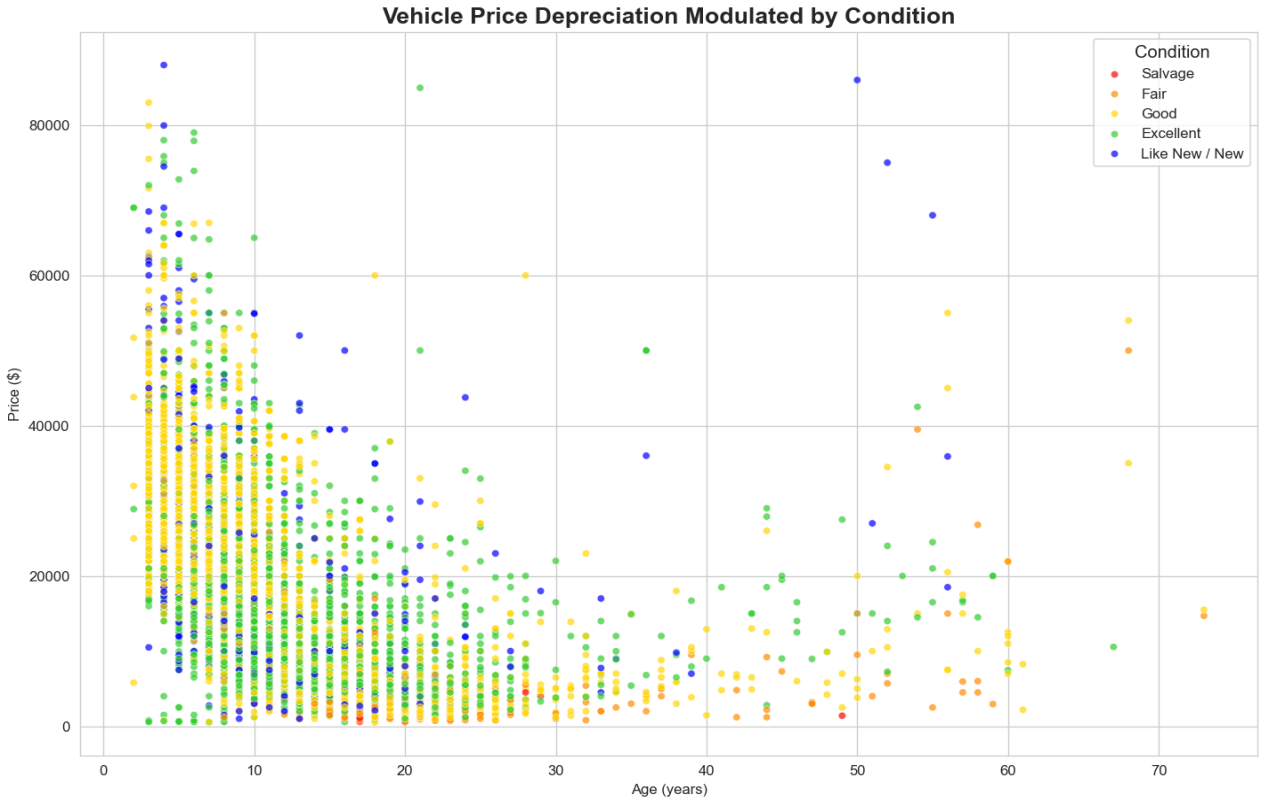
FINDING #3 - THE TYPICAL MARKET LANDSCAPE

- This view shows the **typical price range** (IQR) for each vehicle type with **outliers removed** for clarity. Using the IQR method any vehicle below Q1-1.5xIQR or above Q3 + 1.5xIQR was dropped from this visual,
- Done to allow focus on core price distribution where real purchases would occur, without the distortion from potential scams etc.
- It establishes a baseline: **trucks and SUVs** have a higher median value than **sedans and coupes** so you could expect to spend more there.
- I removed outliers here to understand the **core market dynamics** without the large values skewing the view.

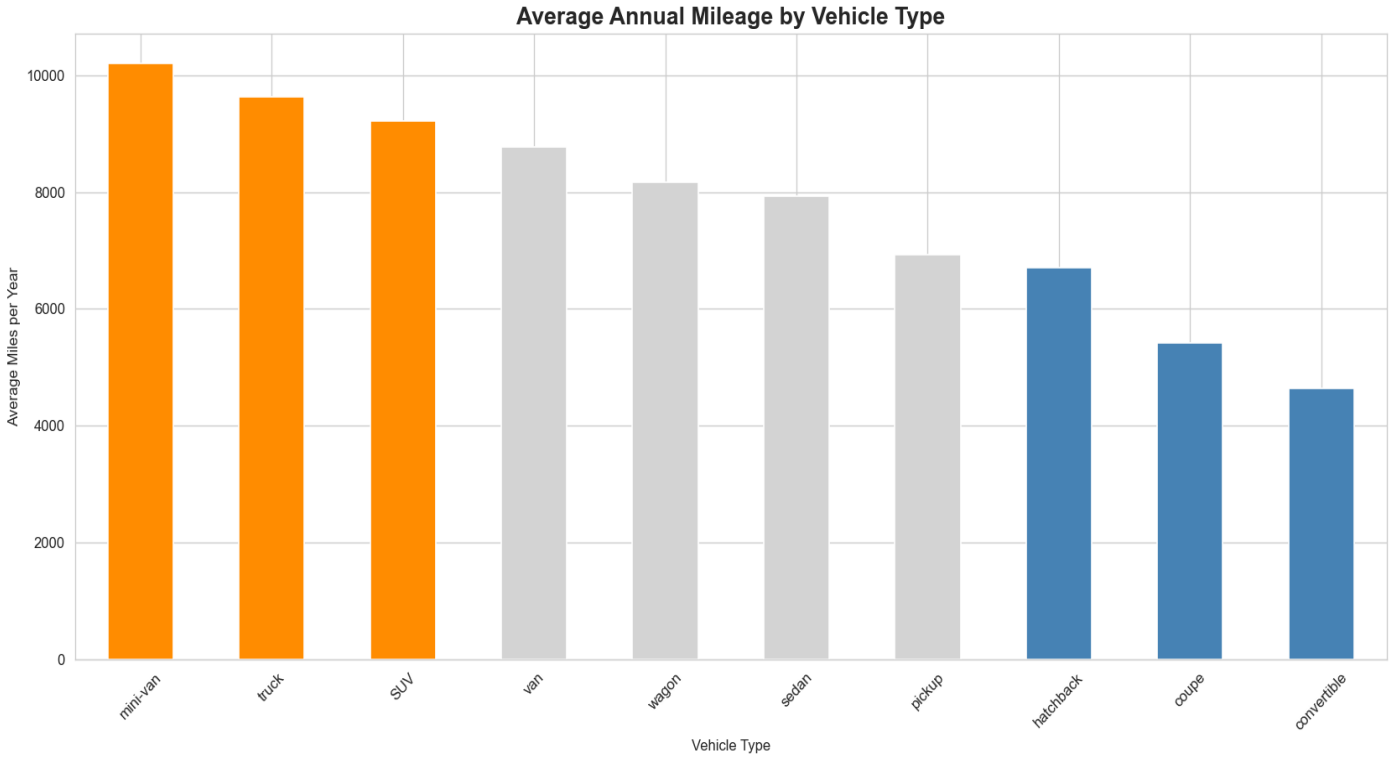


FINDING #4: CONDITION MODIFIES VALUE

- While age dictates the overall price trend, **condition a key modifier**.
- At any point on the depreciation curve, a car in **'Excellent' or 'Like New / New' condition (blue/green)** commands a premium.
- A **'Salvage' or 'Fair' title (red/orange)** significantly reduces value, regardless of age.



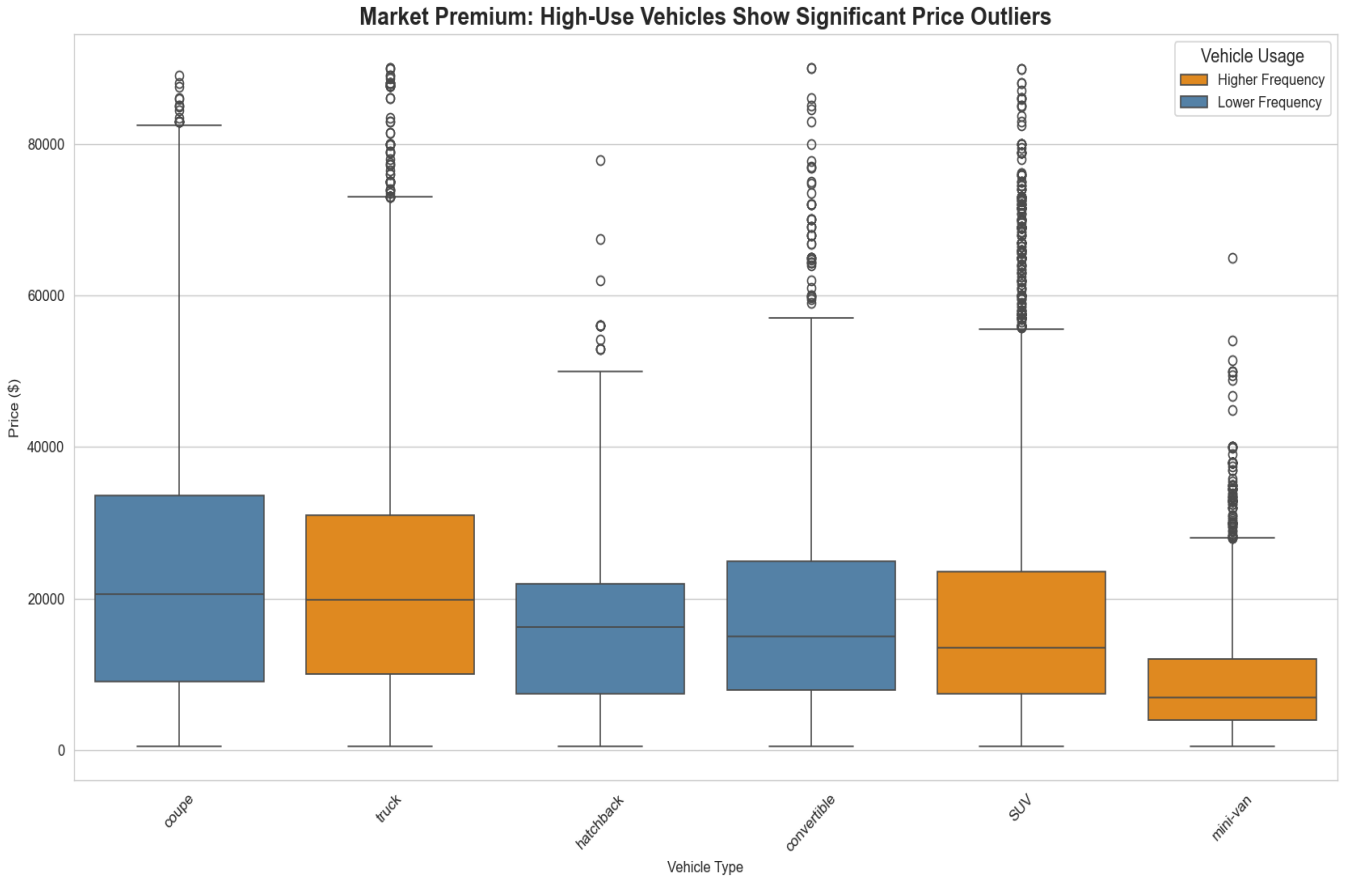
FINDING #5 REVEALING USAGE PATTERNS



- My engineered feature, **miles_per_year**, reveals how vehicles are used.
- **Mini-vans, Trucks, and SUVs (Orange)** are **high-frequency use** vehicles, often for families, work, and daily logistics.
- **Hatchbacks, Coupes, and Convertibles (Blue)** are **lower-frequency use** vehicles, often secondary cars or for leisure.
- This usage directly impacts their depreciation rates and value.

FINDING #6: THE HIGH-USE PREMIUM

- This plot tells the story of market demand. We isolated the 3 most-used and 3 least-used vehicle types.
- **High-use vehicles (Mini-vans, Trucks, SUVs - Orange)** show a **massive number of high-price outliers**.
- **Why?** High demand and necessity create a market where sellers can command a premium, leading to both justified high prices and potential scam listings.
- **Lower-use vehicles (Blue)** show a tighter and more consistent price range with it missing all the potential scams.



VALUE RETAINMENT

```
🏆 TOP 10 BRANDS FOR VALUE RETENTION (After 100,000+ Miles)
Higher = Better value maintained even at high mileage
------------------------------------------------------------------------
 1. ram           | $ 0.16 per mile | Avg Price: $22,761 | Avg Miles: 156,152
 2. porsche       | $ 0.14 per mile | Avg Price: $17,182 | Avg Miles: 131,742
 3. gmc           | $ 0.11 per mile | Avg Price: $15,787 | Avg Miles: 158,385
 4. ford          | $ 0.10 per mile | Avg Price: $14,564 | Avg Miles: 153,661
 5. rover         | $ 0.10 per mile | Avg Price: $13,020 | Avg Miles: 137,250
 6. audi          | $ 0.09 per mile | Avg Price: $10,360 | Avg Miles: 132,353
 7. chevrolet     | $ 0.09 per mile | Avg Price: $12,500 | Avg Miles: 152,912
 8. jeep          | $ 0.09 per mile | Avg Price: $11,112 | Avg Miles: 144,770
 9. mercedes-benz | $ 0.08 per mile | Avg Price: $10,299 | Avg Miles: 139,857
10. infiniti      | $ 0.08 per mile | Avg Price: $9,950 | Avg Miles: 146,210
```

DATA INSIGHTS

What I Learned from the Data:

- **Price is most strongly driven by mileage and usage intensity,** not just age. Our engineered feature miles_per_year proved to be a powerful predictor.
- **Condition is a critical value modifier.** A car in 'Excellent' condition can command a significant premium, while a 'Salvage' title drastically reduces value, regardless of age.
- **Market demand creates pricing anomalies.** High-use vehicles (Trucks, SUVs, Minivans) have a wider range of high-price outliers, reflecting both justified premiums and potential overpricing.
- **Brand and vehicle type establish a clear pricing baseline.** Understanding this landscape is the first step toward identifying a good deal.

Most Important Actionable Insights:

- **For Assessing Value:** Look at miles_per_year, not just total odometer reading. A car with 80k miles driven over 2 years is riskier than one driven over 8 years.
- **For Negotiating:** Condition is a powerful bargaining point. Use the documented value gap between 'Fair' and 'Excellent' to justify offers.
- **For Buying:** Be highly skeptical of high-price outliers for Trucks and SUVs; the market contains both real premiums and potential scams.

Realistic Recommendations:

- **For Buyers:** Use a checklist based on our key features (vehicle type, miles_per_year, condition) to quickly identify fair prices and flag potential "too-good-to-be-true" or overpriced listings.
- **For Sellers:** Price your vehicle according to its *usage intensity* and be honest about condition. Highlighting a low miles_per_year can justify a higher asking price.
- **For Platforms (like Craigslist):** Integrate a "Fair Market Value" estimate on listings using these factors to promote transparency and trust.

PROPOSED DATA PRODUCT: THE "FAIRPRICE" CAR VALUATION TOOL

- **What it would do:** A web-based application or dashboard where a user inputs their car's details (make, model, year, condition, mileage, fuel type, vehicle type). The tool then uses a machine learning model (trained on your cleaned dataset) to:
  - Predict a **fair market value**.
  - Flag if the car is a potential **"Good Deal"** (price significantly below predicted value) or a **"Premium Listing"** (price significantly above predicted value).
  - Show how the car compares to the typical market landscape for its type (like your box plot visualization).
- **Who would use it:**
  - **Used Car Buyers:** To avoid overpaying and identify the best value listings.
  - **Used Car Sellers:** To price their vehicle competitively and realistically for a faster sale.
  - **Car Dealerships:** To be a bit more in touch with the "real world" market.
- **How it could provide value:**
  - **Democratizes Information:** Puts powerful analytics into the hands of everyday consumers.
  - **Promotes Fairness:** Directly addresses your project's goal by creating a transparent benchmark for pricing.
  - **Reduces Risk:** Helps buyers avoid scams (e.g., the "high-price outliers" you identified for trucks/SUVs) and helps sellers understand the impact of condition and usage on their car's value)
  - Would give more insight on private party sales, than **KBB** (According to the website is tailored more to whole and dealership sales)

CHALLENGES AND POTENTIAL NEXT STEPS

- **Challenges & Pitfalls Encountered:**
- **Data Quality Issues:**
  - Original dataset contained inconsistent condition labels (8+ variations), missing critical fields, and unrealistic values
  - **Solution:** Standardized condition categories, removed incomplete records, and applied realistic price/mileage filters
- **Feature Engineering Complexity:**
  - Creating meaningful metrics like miles_per_year required handling edge cases (division by zero for new cars)
  - **Solution:** Implemented .replace(0, 1) to handle brand new vehicles and validate calculations
- **Outlier Management:**
  - Extreme values in price and mileage skewed initial analysis
  - **Solution:** Established data boundaries ($500-$90,000 prices, 100k-300k mileage ranges) based on market reality
- **Missing Context:**
  - Raw data lacked important depreciation context
  - **Solution:** Engineered age and price_per_mile features to better capture true value retentio

Potential Next Steps with More Time:

- **Machine Learning Integration:**
  - Build a predictive pricing model using Random Forest or XGBoost to estimate fair market value
  - Create a "Good Deal" algorithm that flags listings significantly below predicted value
- **Enhanced Feature Engineering:**
  - Incorporate geographic price variations by analyzing regional pricing differences
  - Add seasonal demand patterns to understand price fluctuations throughout the year
- **Data Product Development:**
  - Create a "FairPrice Calculator" web app for consumers to validate listing prices
  - Build a dealer dashboard for inventory valuation and pricing optimization
- **Additional Data Sources:**
  - Integrate maintenance history, accident reports, and ownership records
  - Add manufacturer reliability ratings and consumer satisfaction data
- **What i Would Do Differently:**
- Collect more structured condition data with standardized appraisal criteria
- Gather more detailed vehicle history and maintenance records
- Implement real-time data streaming for current market prices rather than snapshot data

THANK YOU

TORRODJAE SOMERVILLE