A Statistical Study of Used Cars in the US: The correlation of Price vs. Mileage by Brand

By: Tasnim A. Raisa

Introduction

The U.S. used car market has undergone significant changes from 2015 to 2020s, marked by steady growth, shifting consumer behaviors, and external disruptions. In the mid-2010s, the used vehicle sector was already massive in scale, and over the past decade it has further expanded in volume and value. By the mid-2020s, it remained a pivotal part of the auto industry, influenced by new technologies and economic events.

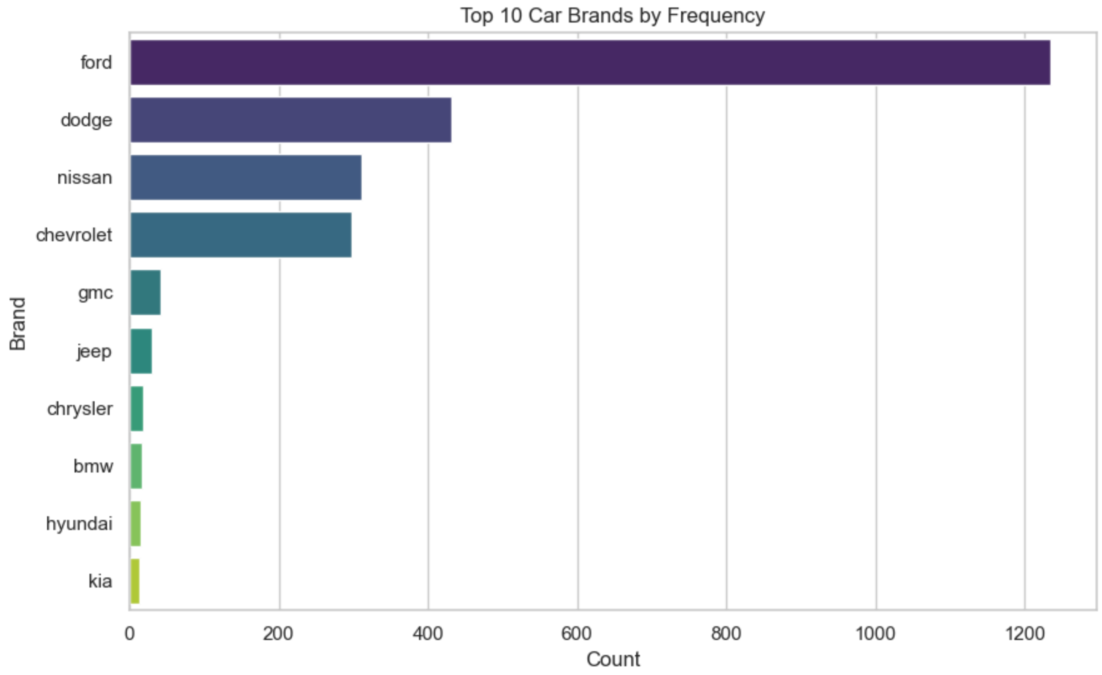
Used car prices have risen dramatically between 2015 and 2020s. These high prices have strained affordability for consumers but also increased trade-in values for sellers. The period from 2015 to 2025 saw a revolution in *how* used cars are bought and sold, with online platforms growing rapidly. Traditional dealerships increasingly embraced e-commerce, and dedicated online used-car retailers exploded in popularity. The paper aims to look at a comprehensive dataset of 2,499 used cars in the US from a dealership inventory to study trends, reach notable conclusions of brand reputation and take data analysis measures to further understand and study the data to conclude findings. The dataset was collected through the kaggle website as an open source datasat.

Dataset Overview

The dataset comprises 2,499 used cars with the following key attributes:

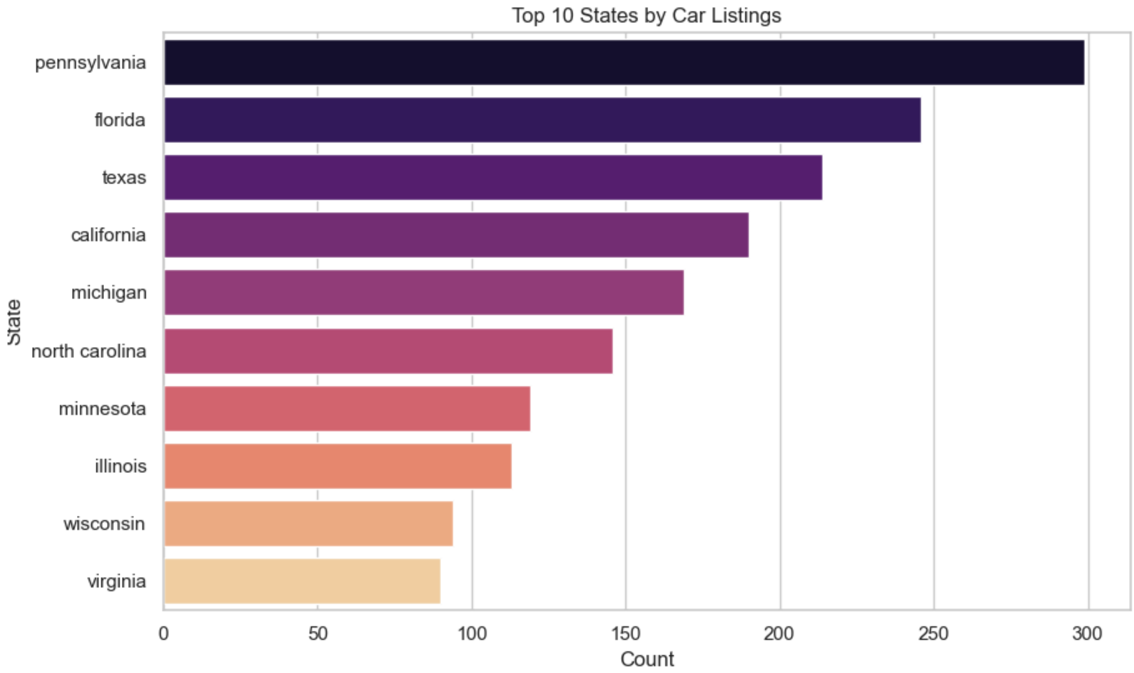
- Brands: Ford (49.42%,1,235 cars),Chevrolet (11.88%, 297 cars),Dodge(17.29%, 432 cars), and others (38.70%).
- Colors: Black (20.65%), red (7.68% per 100 listings), silver (300 cars), and 49 total color options.
- Condition: Urgent (41.22%), non-clean title (163 cars, 6.52% defective).
- Price: Uniformly distributed between $2,000and $42,030 (mean $18,767.67, standard deviation $12,116.09).
- Mileage: Exponentially distributed with a mean of 48,922.93 miles

It was observed Ford vehicles dominate the inventory, accounting for nearly 50% of the cars in this particular inventory. Other brands like Dodge and Chevrolet are also frequent but not nearly as dominant as Ford.

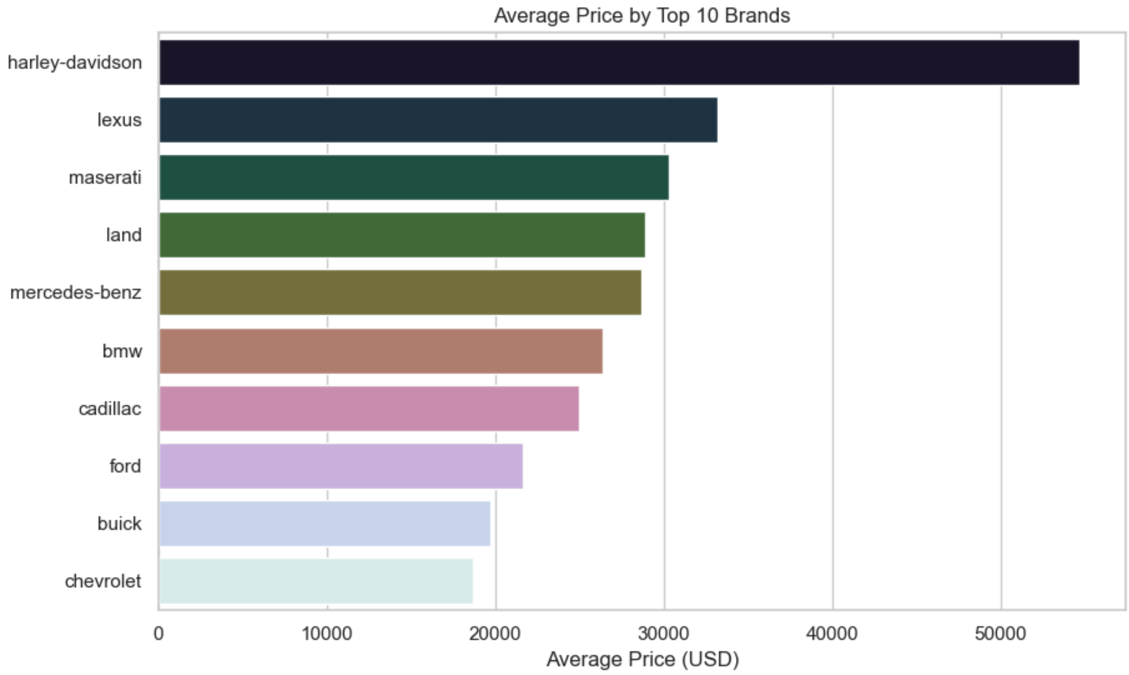


Top 10 car brands in the dataset by frequency.

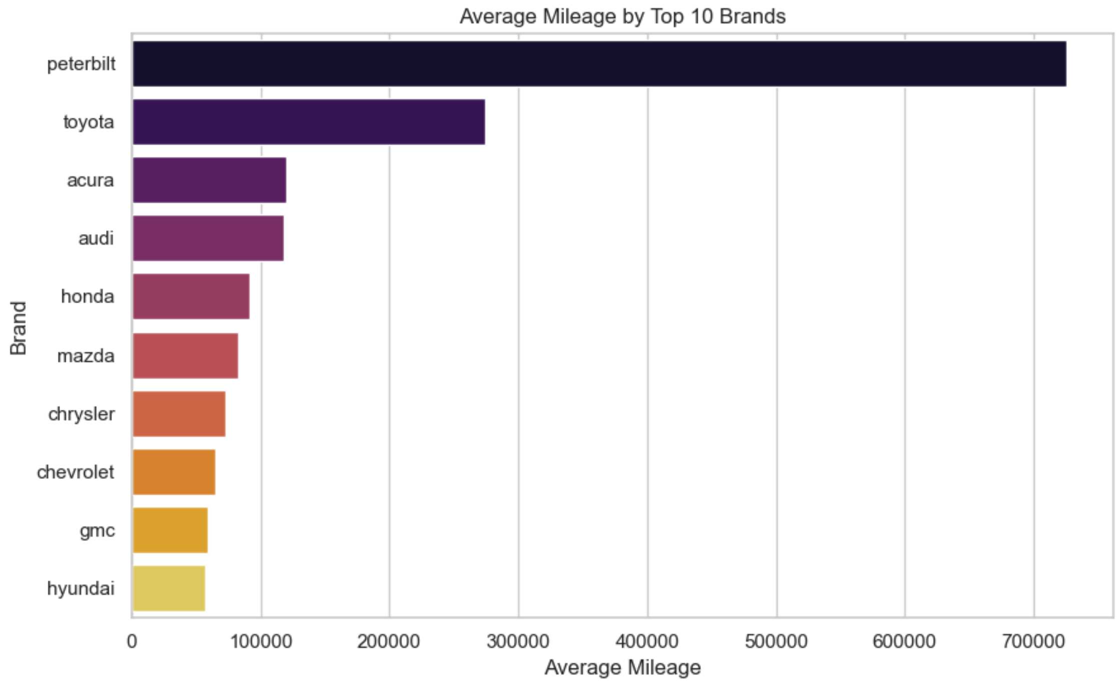
Observation: Ford is overrepresented, implying it disproportionately influences overall statistics and sampling probabilities. The implications can be observed throughout the statistics based analysis, however also begs the question of ford preferences in the USA market for both new and used car markets.



The graph above illustrates the number of car listings across ten U.S. states, with the count on the x-axis and states on the y-axis. Pennsylvania tops the list with around 300 listings, followed closely by Florida at approximately 275, Texas at about 250, and California with roughly 225 listings. Michigan has around 200 listings, while North Carolina has about 175, Minnesota around 150, and Illinois approximately 125. Wisconsin and Virginia have the fewest listings, with about 100 and 75 respectively.

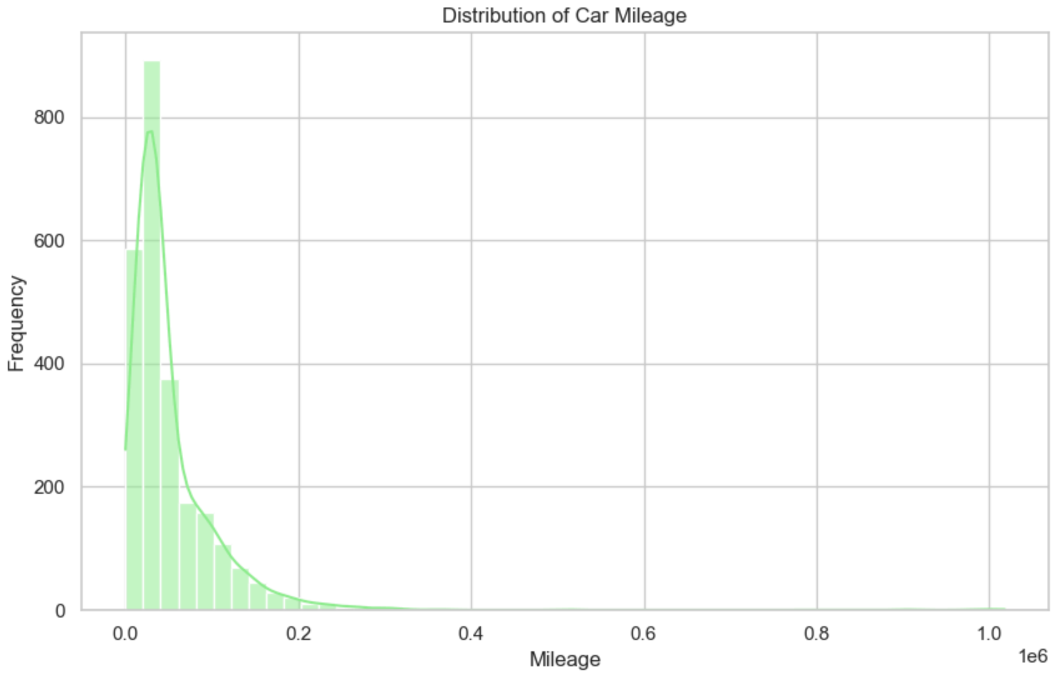


The bar chart titled "Average Price by Top 10 Brands" above displays the average price (in USD) of vehicles across ten car brands, with the price on the x-axis and brands on the y-axis. Harley-Davidson leads with the highest average price, nearing $50,000, possibly due to the limited edition nature of their cars. This is followed by Lexus at around $45,000, Maserati at approximately $40,000, and Land [Rover] at about $35,000. Mercedes-Benz and BMW both have average prices around $30,000, while Cadillac and Ford are close to $25,000. Buick averages around $20,000, and Chevrolet has the lowest average price, just above $15,000.

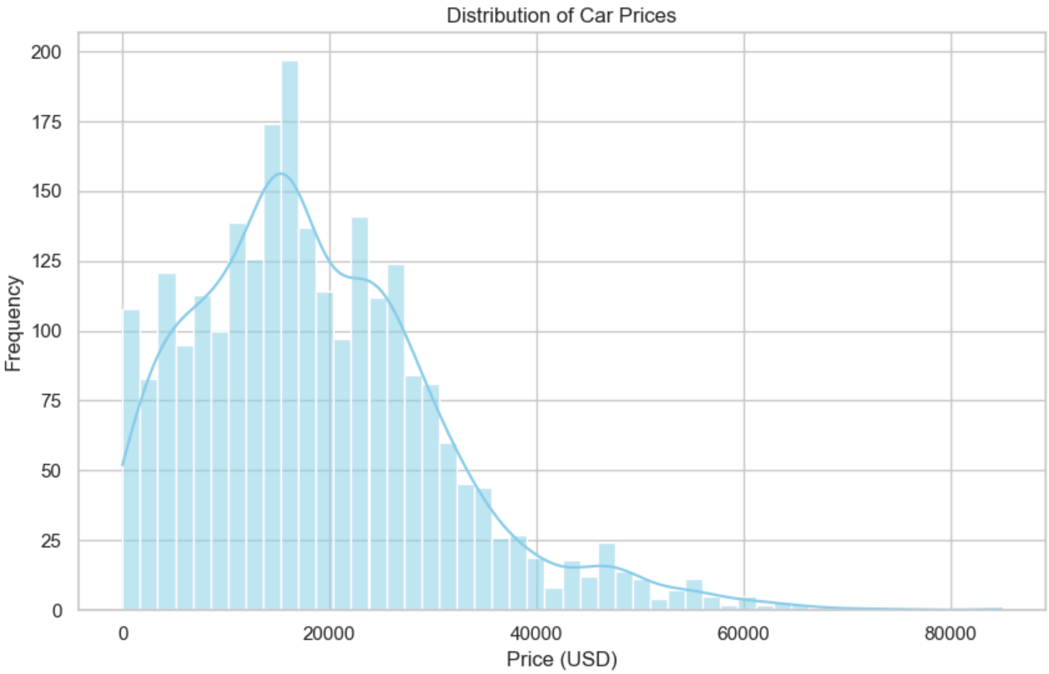


Average mileage per brand with significant outliers like Peterbilt and Toyota.

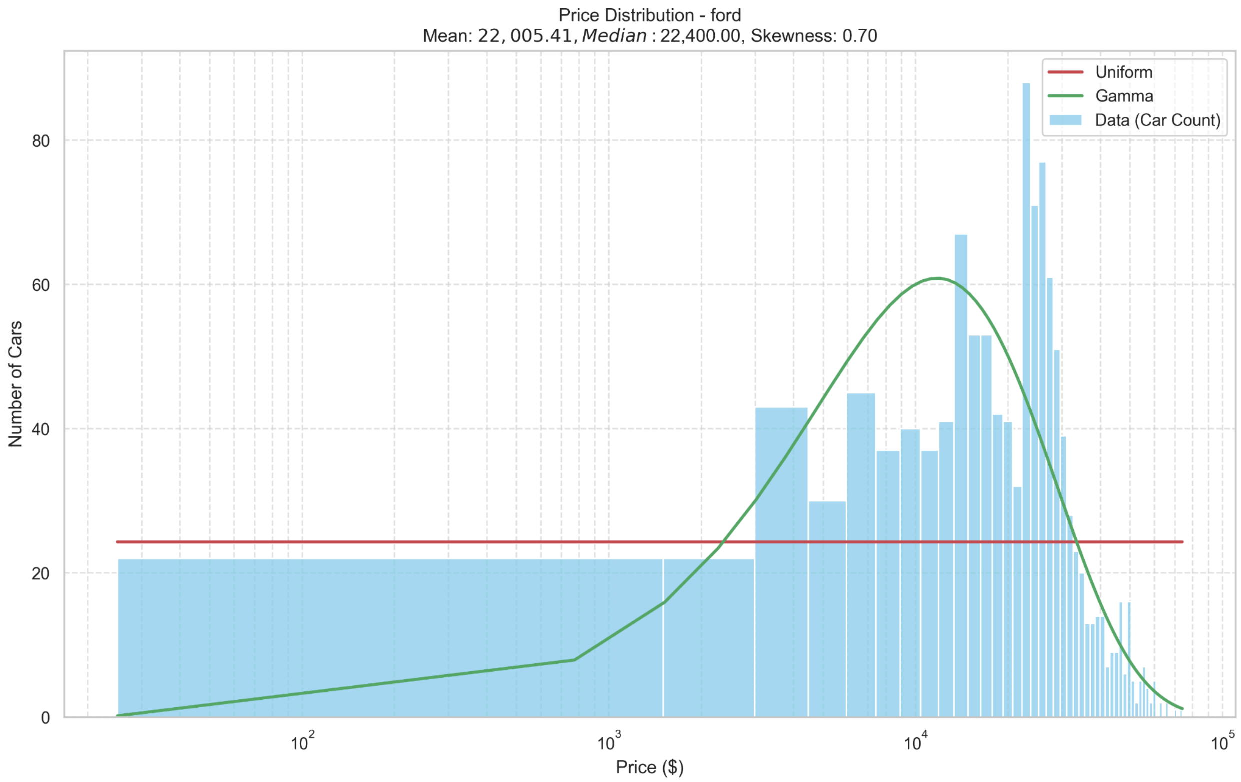
Some brands have much higher average mileage, such as Peterbilt and Toyota, which suggests these vehicles are retained longer or are used for more intensive driving. Brands like GMC and Hyundai exhibit lower average mileages, which may imply newer inventory or limited usage.



"Distribution of Car Mileage," which shows the frequency of cars across different mileage ranges. The x-axis represents mileage in millions (ranging from 0 to 1 million, denoted as "1e6"), while the y-axis represents frequency, ranging from 0 to 800. The data is depicted with green bars, forming a right-skewed distribution. The highest frequency, peaking at around 800, occurs at a mileage close to 0, indicating that most cars have low mileage. The frequency drops sharply as mileage increases, with a gradual decline beyond 0.2 million, tapering off to nearly 0 by 1 million, suggesting that fewer cars have high mileage.



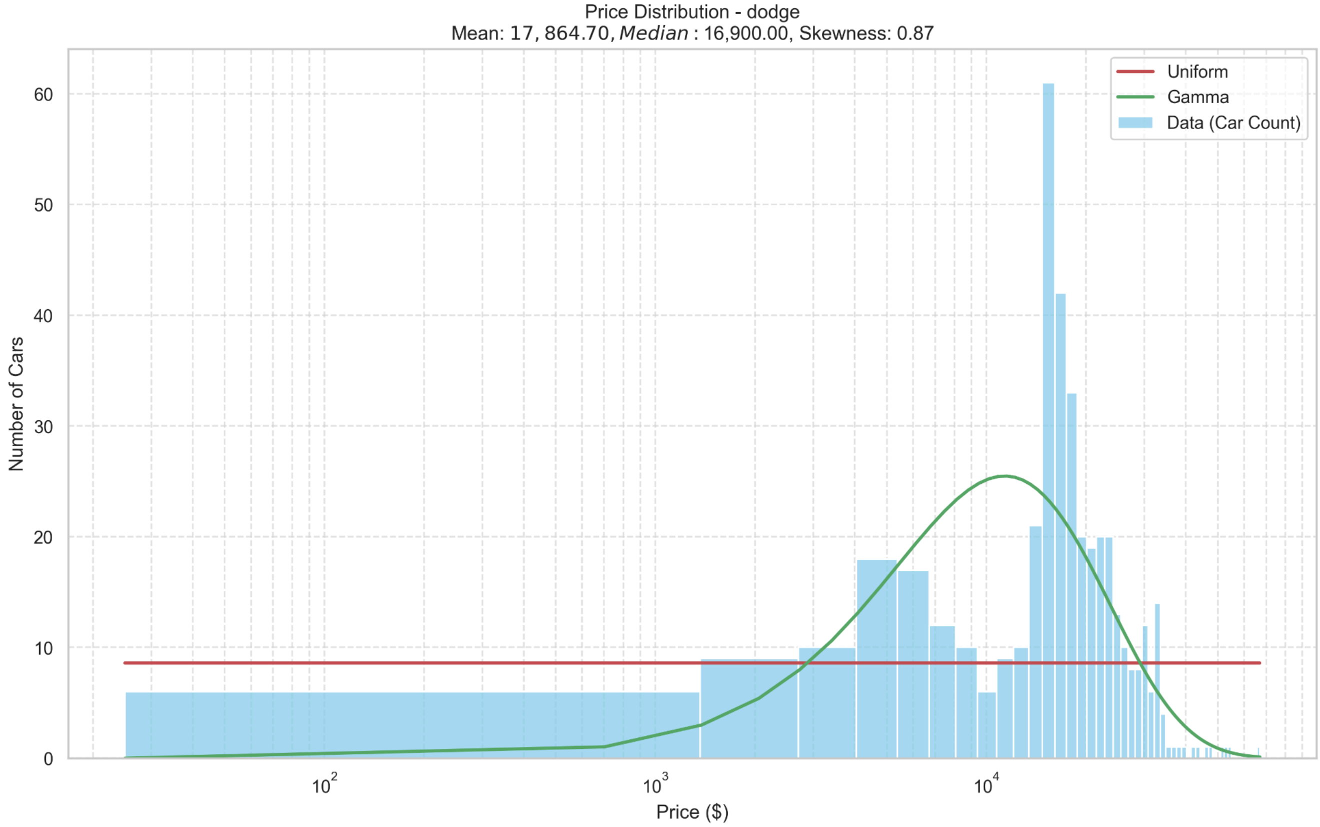
The graph is a histogram titled "Distribution of Car Prices," illustrating the frequency of car prices in USD. The x-axis represents the price range from 0 to 80,000 USD, while the y-axis shows the frequency, ranging from 0 to 200. The data is depicted with light blue bars and overlaid with a smooth curve, forming a bell-shaped distribution. The peak frequency, reaching approximately 175, occurs around 2,000 to 3,000 USD, indicating that most cars are priced in this range. The frequency decreases symmetrically on both sides, tapering off to nearly 0 by 8,000 USD, with a slight tail extending toward higher prices, suggesting a smaller number of cars exceed this range.



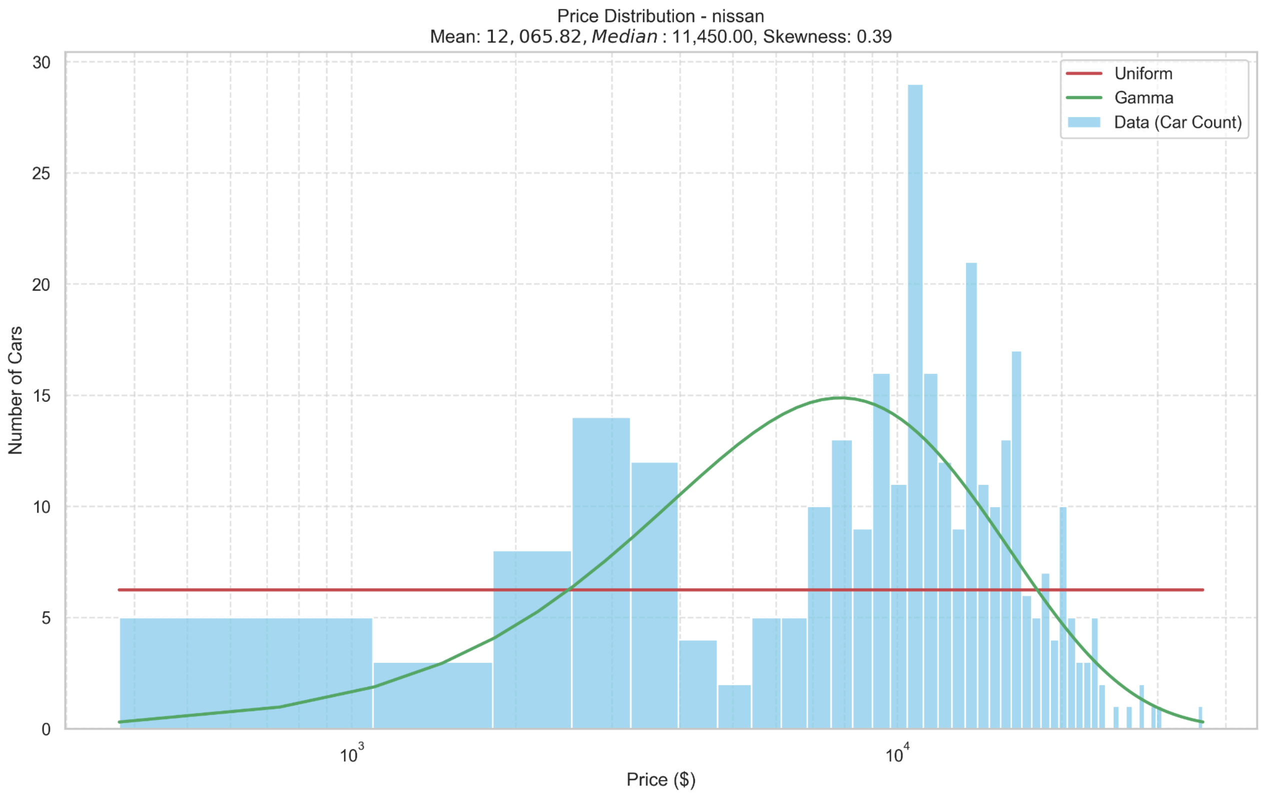
The graph, "Price Distribution - ford," is a histogram illustrating the distribution of car prices for Ford vehicles, with additional fitted curves for comparison. The x-axis represents price in dollars on a logarithmic scale, ranging from 10^2 (100) to 10^5 (100,000), while the y-axis shows the number of cars, ranging from 0 to 80. The histogram, depicted in light blue, peaks around 10^3 (1,000) with a frequency of about 60, indicating the most common price range, and shows a right-skewed distribution with a long tail extending toward 10^4 (10,000).

A green curve representing a Gamma distribution fits the data, peaking similarly around 10^3, while a red horizontal line at approximately 20 cars indicates a Uniform distribution for reference. Statistical measures include a mean of $22,005.41, a median of $22,400.00, and a skewness of 0.70, suggesting a moderately right-skewed distribution.

The following price distribution graphs follow the same trend as well illustrating a brand specific price distribution.



In the case of Dodge,the price distribution graph measures include a mean of $17,864.70, a median of $16,900.00, and a skewness of 0.87, demonstrating a right-skewed distribution. For Nissan on the other hand, a mean of $12,065.85, a median of $11,450.00, and a skewness of 0.39, suggesting a moderately left-skewed distribution.



Research Motivation and Hypothesis

This analysis of the dataset comprising 2,499 used cars from a dealership inventory seeks to examine how vehicle mileage influences resale prices and whether this effect varies across different car brands. The hypothesis posits that mileage negatively impacts price, with the effect being more pronounced for certain brands than others.

Furthermore, the study explores the distributions of key attributes and assesses the reliability of features such as car color in predicting variables like brand.

Additional hypotheses formed over research were:

- Ford's dominance skews inventory selection probabilities.
- Car colors are weak predictors of brand.
- Car prices follow a uniform distribution; mileage follows an exponential distribution.
- Normalized price and mileage are statistically independent.

Methodology Summary

To conduct this investigation,questions were formed from the statistics topics covered over the semester and to further analyse the dataset, it was initially narrowed to focus on the most frequent car brands for enhanced clarity. Linear regression models were created and fitted for each selected brand, with price as the dependent variable and mileage as the independent variable. The regression lines were visualized, and R^2 values were extracted to evaluate the strength of the relationship between price and mileage across brands.This was achieved through running python code on Jupyter Notebook to produce graphs and summary of data for seeming correlations. Additionally, the overall distributions of prices and mileage were analyzed, and assumptions were tested using probability models and independence checks.

- Filtered to top 5–10 brands based on frequency.
- Fitted `Price ~ Mileage` linear models and visualized R^2 scores.

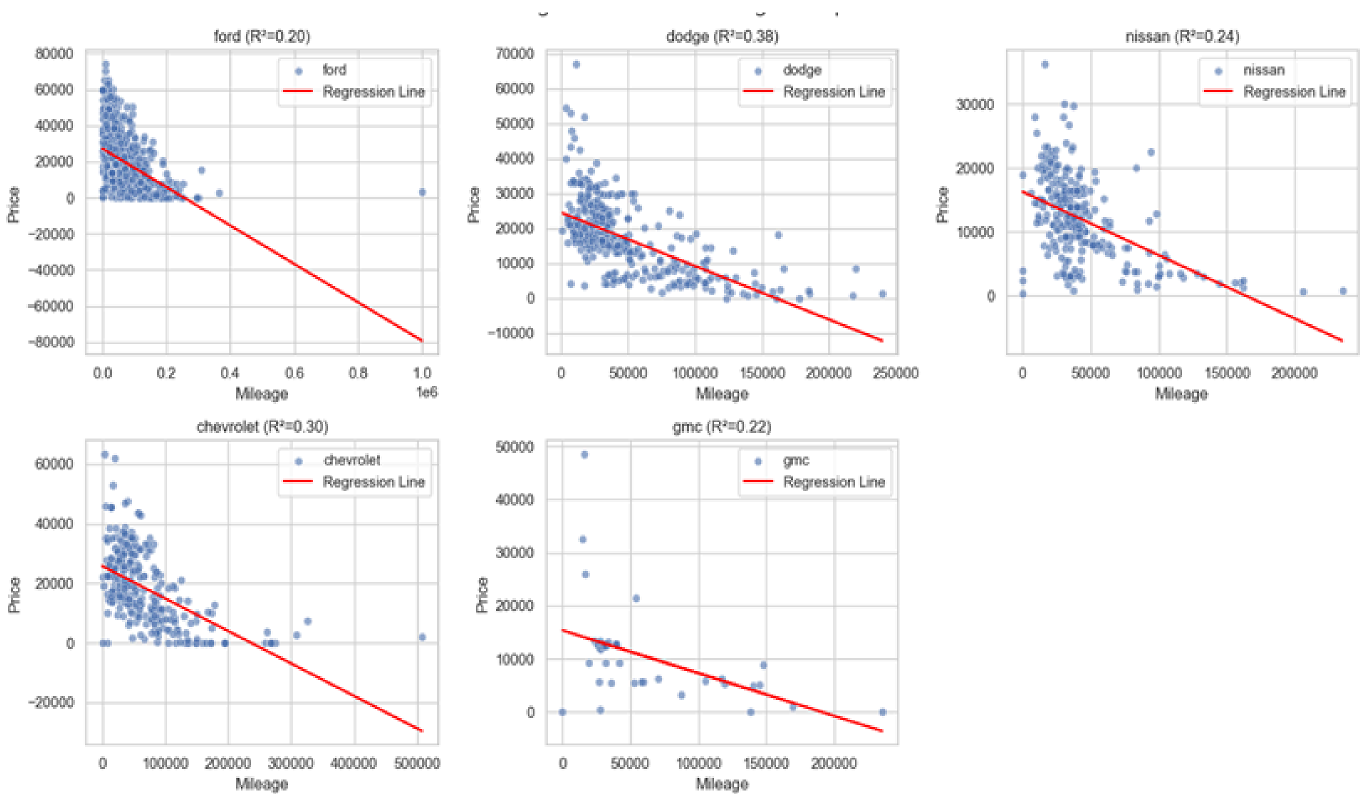
- Applied exponential model to mileage and uniform distribution to price.
- Used probability distributions to evaluate quality control and brand-color dependencies.

Visual Data Analysis

Linear Regression (Price vs Mileage by Brand)

Further analysis was performed through linear regression with the assistance of Jupyter Notebook and Statsmodel. The analysis revealed that all brands exhibit a negative slope, indicating that an increase in mileage generally leads to a decrease in price. However, the rate of depreciation varies across brands. Dodge displayed the steepest slope, suggesting that its price declines most rapidly with increased mileage. Conversely, Ford, despite being the most prevalent brand in the dataset, exhibited the flattest slope and the weakest correlation, indicating that other factors may influence its pricing variability.

See figure below showing linear regression trends by each brand.



The analysis revealed that all brands exhibit a negative slope, indicating that an increase in mileage generally leads to a decrease in price. However, the rate of depreciation varies across brands. Dodge displayed the steepest slope, suggesting that its price declines most rapidly with increased mileage. Conversely, Ford, despite being the most prevalent brand in the dataset, exhibited the flattest slope and the weakest correlation, indicating that other factors may influence its pricing variability.

Statistical Modeling Highlights

- Prices are uniformly distributed between $2,000 and $42,030, with a mean of $18,767.67.
- Mileage follows an exponential distribution with a mean of 48,922.93 miles.
- Normalized price and mileage are statistically independent (MSE: 7.3×10^{-5}).

To explore broader trends, car prices were modeled using a uniform distribution, while mileage was modeled with an exponential distribution. The findings were consistent: prices ranged evenly between $2,000 and $42,030, averaging approximately $18,767.67, and mileage was heavily right-skewed with a mean of 48,922.93 miles. Only about 13% of vehicles had mileage exceeding 100,000 miles. These distributions align well with the theoretical models applied.

The independence of normalized price and mileage was also tested, revealing a very small mean squared error between the joint and marginal distributions, indicating that they are approximately independent.

- Black is not a strong predictor for Dodge (only 13.37% of black cars are Dodge).
- A Ford car has a 33.52% chance of being marked as 'urgent'.
- There is a 71.35% probability of finding 5 clean-title cars in a sample.
- It takes approximately 39 scans to find 3 red cars.
- Price below $15,000: 32.48%; Mileage over 100,000 miles: 12.95%.
- Quality control plans that reject lots with any defects accept only 16.81% of lots if the defect rate is 30%.

Integrated Conclusions

This investigation confirms that mileage generally reduces car prices, though the strength of this relationship varies significantly across brands. Ford’s pricing behavior is notably inconsistent, potentially due to its high market volume or varying conditions.This statistical analysis of inventory reveals critical probabilistic patterns in brand, color, condition, price, and mileage. Ford’s dominance (49.42%) underscores its central role in inventory dynamics. Car color proves to be an unreliable indicator of brand, while probabilistic tools such as hypergeometric and Poisson distributions provide valuable insights into quality control and rarity-based scanning tasks.

Uniform price and exponential mileage distributions provide reliable models for financial and maintenance planning. The strict quality control plan highlights the need for balanced defect detection strategies. These findings support data-driven decision-making in inventory management, pricing, and quality assurance, with implications for operational efficiency and customer satisfaction.

- Mileage reduces price, but brand-specific trends matter.
- Brand influences price more than mileage due to Ford’s prevalence.
- Color is not a reliable predictor for a brand.
- Uniform and exponential distributions fit price and mileage well.
- Quality control is strict and should be reviewed.

Recommendations

Based on the analysis, it is recommended to prioritize Ford vehicles for volume deals while applying depreciation models with careful consideration of brand-specific trends. Inspection strategies should strike a balance between defect sensitivity and practicality. Additionally, statistical distributions should be integrated into financial and planning tools to enhance inventory decision-making processes.

- Prioritize Ford for volume deals and monitor Dodge depreciation trends closely.
- Utilize brand-specific regression models for pricing.
- Adjust quality control thresholds to avoid unnecessary rejections.
- Incorporate statistical distributions into pricing tools.
- Avoid relying on color for brand identification.

References

- Kaggle. (n.d.). *Used Car Dataset*. Retrieved from https://www.kaggle.com/datasets
- Statista. (n.d.). *Automotive Industry Statistics*. Retrieved from https://www.statista.com