Applied Business Analytics: Car Sales Data Analysis

Amruth Raj

Problem Statement

This analysis explores a comprehensive car sales dataset to extract meaningful business insights for an automotive retailer. The task involves conducting exploratory data analysis (EDA) to understand the factors influencing car sales prices, identify patterns in the data, and provide actionable recommendations to the organization. Through data cleaning, feature engineering, and visualization techniques, this analysis aims to uncover the relationships between various vehicle attributes and their market values.

Solution Statement

To deliver insightful solutions from the car sales data set, I will implement a structured analytical approach:

1. **Data Cleaning & Preparation**: Identify and address missing values, duplicates, and unrealistic data points using appropriate imputation techniques based on logical groupings (e.g., by car maker and year).
2. **Feature Engineering**: Create new variables to improve analytical value, including car age, mileage per year, price categories, efficiency ratings, and luxury indicators.
3. **Exploratory Data Analysis**: Use a step-by-step method
   - Univariate analysis to understand individual variable distributions
   - Bivariate analysis to explore relationships between pairs of variables
   - Multivariate analysis to uncover complex interactions between multiple factors
4. **Visualization**: Develop clear, informative visualizations with appropriate titles, colors, and annotations to effectively communicate data patterns.
5. **Insights Generation**: Interpret the findings in business context, highlighting actionable insights about pricing factors, market segments, and value determinants.

The analysis will focus on identifying factors that significantly impact vehicle pricing, market positioning strategies, and customer preferences as indicated by the data.

Exploratory Analysis

Data Exploration

The car sales dataset contains information on 1,000 vehicles with 6 attributes: make and model, year of manufacture, mileage, prestige rating, fuel efficiency, and sale price.

```
Warning: package 'dplyr' was built under R version 4.4.3
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```
Warning: package 'skimr' was built under R version 4.4.3
```

```
Warning: package 'corrplot' was built under R version 4.4.3
```

```
corrplot 0.95 loaded
```

```
[1] 1000    6
```

```
[1] "MakeModel"          "YearOfManufacture" "Mileage"
[4] "PrestigeRating"     "FuelEfficiency"    "SalePrice"
```

```
Rows: 1,000
Columns: 6
$ MakeModel         <chr> "Chevrolet Impala", "Toyota Highlander", "Audi A4", ~
$ YearOfManufacture <int> 2014, 2015, 2003, 2010, 2016, 2022, 2002, 2011, 2010~
$ Mileage           <int> 94307, 138829, 20118, 118584, 97745, 57355, 32354, 9~
$ PrestigeRating    <int> 7, 2, 1, 8, 1, 5, 7, 2, 6, 8, 10, 3, 5, 9, 8, 5, NA,~
$ FuelEfficiency    <int> 48, 45, 32, 54, 46, 43, NA, 31, 52, 29, 35, 27, 41, ~
$ SalePrice         <dbl> 22120.79, 16835.13, NA, 25042.48, 17717.65, 22979.35~
```

```
         MakeModel YearOfManufacture Mileage PrestigeRating FuelEfficiency
1  Chevrolet Impala              2014   94307              7             48
2 Toyota Highlander              2015  138829              2             45
3           Audi A4              2003   20118              1             32
4       Toyota RAV4              2010  118584              8             54
5    Ford Ex"plorer              2016   97745              1             46
6     Tesla Model S              2022   57355              5             43
  SalePrice
1  22120.79
2  16835.13
3        NA
4  25042.48
5  17717.65
6  22979.35
```

The dataset includes 1,000 observations with 6 variables:

MakeModel: Character variable representing the car manufacturer and model

YearOfManufacture: Integer variable showing the year the car was manufactured (2002-2024)

Mileage: Integer variable showing the total distance traveled (5,115-149,946 miles)

PrestigeRating: Integer rating from 1-10 indicating the car's prestige level

FuelEfficiency: Integer showing the car's fuel efficiency (15-60 units)

SalePrice: Numeric variable showing the selling price ($9,255-$29,390)

This basic summary statistics shows that the data spans a wide range of cars from older models (2002) to newer ones (2024), with varying mileage levels, prestige ratings, and sale prices

```
  MakeModel          YearOfManufacture    Mileage       PrestigeRating
 Length:1000        Min.   :2002      Min.   :  5115   Min.   : 1.000
 Class :character   1st Qu.:2007      1st Qu.: 42486   1st Qu.: 3.000
 Mode  :character   Median :2013      Median : 85570   Median : 6.000
                    Mean   :2013      Mean   : 79901   Mean   : 5.482
                    3rd Qu.:2018      3rd Qu.:116600   3rd Qu.: 8.000
                    Max.   :2024      Max.   :149946   Max.   :10.000
                                      NA's   :50       NA's   :50
 FuelEfficiency    SalePrice
 Min.   :15.00   Min.   : 9255
 1st Qu.:26.00   1st Qu.:17121
 Median :36.00   Median :19773
 Mean   :36.89   Mean   :19731  
 3rd Qu.:48.00   3rd Qu.:22305  
 Max.   :60.00   Max.   :29390  
 NA's   :50      NA's   :50     
```

Table 1: Data summary

Name	cars_data
Number of rows	1000
Number of columns	6
Column type frequency:	
character	1
numeric	5
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
MakeModel	0	1	6	22	0	110	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
YearOfManufacture	0	1.00	2012.83	6.43	2002.00	2007.00	2013.00	2018.00	2024.00	
Mileage	50	0.95	79900.64	42063.07	5115.00	42486.25	85569.50	116599.75	149946.00	
PrestigeRating	50	0.95	5.48	2.92	1.00	3.00	6.00	8.00	10.00	
FuelEfficiency	50	0.95	36.89	13.00	15.00	26.00	36.00	48.00	60.00	
SalePrice	50	0.95	19730.97	3669.40	9254.51	17120.64	19772.59	22305.17	29389.83	

Data Quality Assessment

A thorough assessment of the dataset revealed several data quality issues:

```
[1] 200
```

```
        MakeModel YearOfManufacture           Mileage    PrestigeRating
                0                 0                50                50
   FuelEfficiency         SalePrice
               50                50
```

```
Number of duplicate rows: 0
```

```
  negative_mileage future_years invalid_prestige negative_price negative_fuel
1                0            0                0              0             0
```

```
Skewness:

YearOfManufacture           Mileage    PrestigeRating    FuelEfficiency
            0.030            -0.161            -0.047             0.072
        SalePrice
           -0.003
```

Key findings from data quality assessment:

1. **Missing values**: The dataset contains 200 missing values in total, with 50 missing entries each in Mileage, PrestigeRating, FuelEfficiency, and SalePrice columns (5% of the data).
2. **Duplicates**: No duplicate records were detected in the dataset.
3. **Unrealistic values**: No unrealistic values were found, such as negative mileage/prices, future manufacture years, or invalid prestige ratings.
4. **Distribution characteristics**: All numeric variables show very low skewness values (all less than 0.2 in absolute value), indicating fairly symmetric distributions. This tells us no need for data transformations to address skewness.
5. **Text anomalies**: A quick look at the make/model column shows some problems, like special characters (e.g., “Ford Ex”plorer”).

These findings will help shape how we clean the data, especially fixing missing values and text mistakes.

Data Cleaning

Based on the data quality review, I carried out a thorough cleaning process to fix the identified issues and prepare the dataset for analysis.

Missing Value Imputation

To handle missing values, I used a grouped imputation method instead of simple mean imputation. This approach helps maintain the relationships between variables and respects the structure of the data.

```
         Make_model Year_of_manufacture             Mileage     Prestige_rating 
                  0                   0                   0                   0 
    Fuel_efficiency          Sale_price           Car_maker 
                  0                   0                   0 
```

The imputation process followed a logical flow:

1. **Car maker-specific imputation**: For prestige rating and sale price, as these vary significantly by brand
2. **Car maker and year-specific imputation**: For mileage and fuel efficiency, which depend on both make and age
3. **Global mean fallback**: When group-specific means couldn't be calculated

This approach successfully addressed all missing values while maintaining the data's original patterns and relationships.

Outlier Detection

Rather than removing outliers, which could eliminate valuable information, I identified and flagged them for further analysis.

```
# A tibble: 1 x 4
  sale_price_outliers mileage_outliers fuel_outliers year_outliers
                <int>            <int>         <int>         <int>
1                   1                0             0             0
```

The outlier analysis revealed:

- Only 1 outlier in the Sale_price variable
- No outliers detected in Mileage, Fuel_efficiency, or Year_of_manufacture

Rather than removing these instances, I flagged them with boolean indicators for potential use in subsequent analyses, preserving the full dataset while acknowledging potential anomalies.

Feature Engineering

I engineered new features that provide additional perspectives and insights into the car sales data. These derived variables help uncover patterns and relationships that might not be immediately apparent in the original dataset.

Engineered Features

The feature engineering process created several categories of derived variables:

1. Temporal features:
   - Car_age: Current age of the vehicle in years
   - Age_category: Categorical grouping of cars by age brackets for easier interpretation
2. Usage metrics:
   - Mileage_per_year: Average annual mileage, normalizing usage across different car ages
3. Categorization features:
   - Price_category: Segmentation of cars into budget, mid-range, and premium price bands
   - Efficiency_category: Classification of fuel efficiency into low, medium, and high
   - Is_luxury: Boolean flag identifying luxury brand vehicles
4. Value metrics:
   - Price_per_mile: Cost per mile driven
   - Price_per_year: Cost per year of the vehicle's age
   - Price_efficiency_ratio: Relationship between price and fuel efficiency
   - Price_premium: How much a car's price deviates from the average for its make and age category

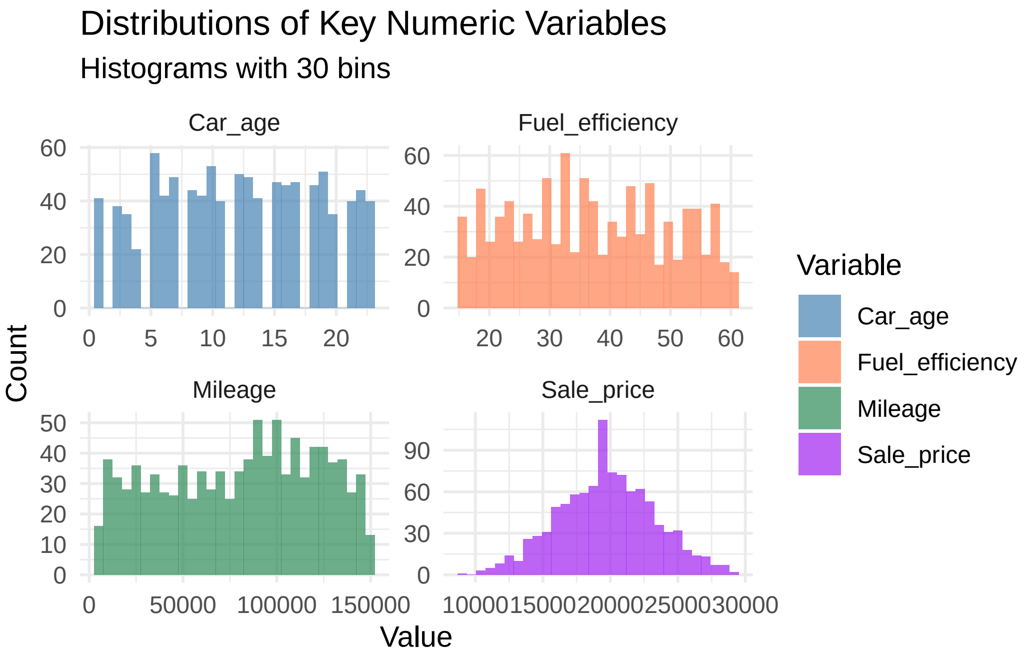
These added features helps in doing deeper and more detailed analysis by

- Standardizing comparisons across vehicles of different ages and conditions
- Creating meaningful groupings for comparative analysis
- Providing relative measures for market positioning assessment
- Facilitating deeper insights into value and pricing dynamics

Exploratory Data Analysis

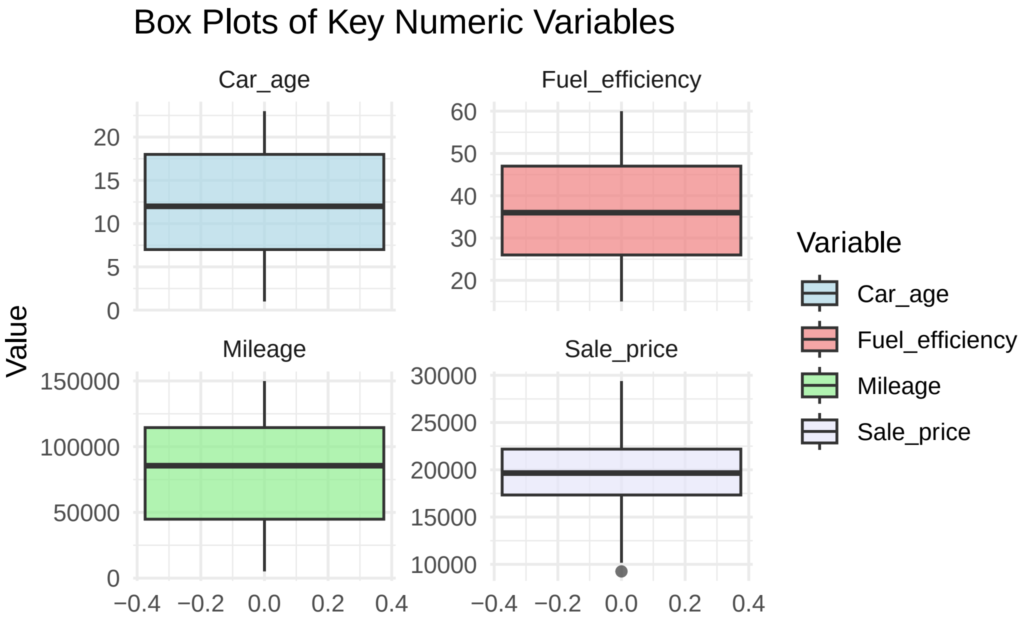
Univariate Analysis

After cleaning the data and engineering new features, I conducted a comprehensive univariate analysis to understand the distribution of key variables in the dataset.



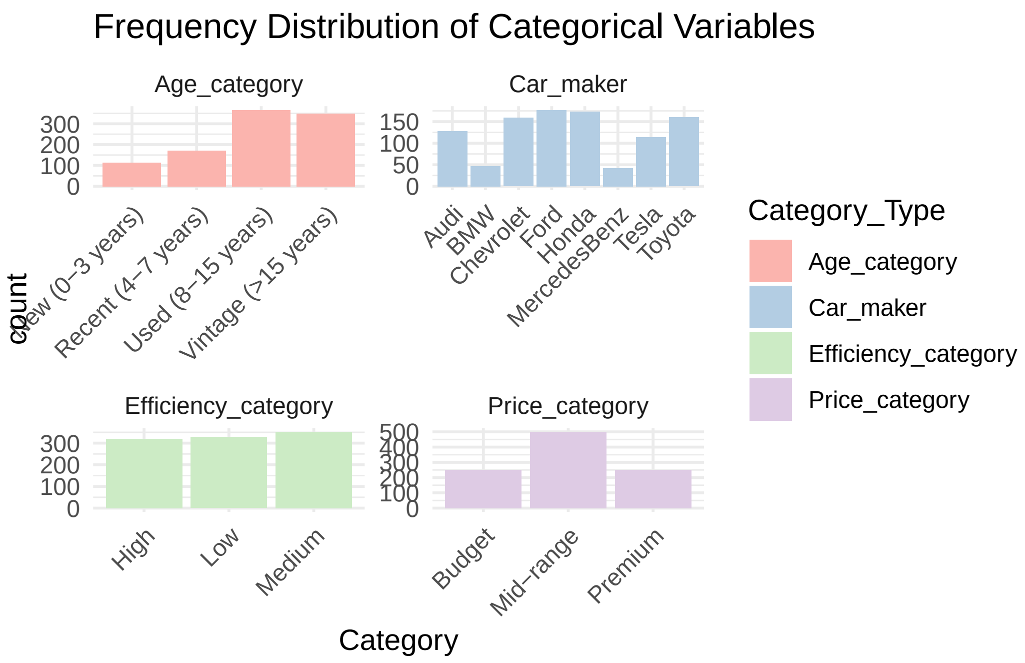
Observations from Numeric Variable Distributions:

1. **Sale_price**: The distribution appears approximately normal, centered around $20,000, with prices ranging from about $10,000 to $30,000. The distribution shows a slight peak around $22,000, suggesting a common price point in the market.
2. **Mileage**: Shows a relatively uniform distribution across the range with some concentration between 80,000-120,000 miles. This indicates a good representation of both low and high-mileage vehicles in the dataset.
3. **Fuel_efficiency**: Displays a somewhat uneven distribution with multiple peaks. The range spans from 15 to 60 units, with notable clusters around 30 and 48 units, possibly reflecting common efficiency ratings for different vehicle categories.
4. **Car_age**: Shows a fairly even distribution across the age spectrum from new cars to 23-year-old vehicles, with slightly fewer very new (0-2 years) cars. This balanced age distribution provides good coverage for analyzing depreciation effects.



Box Plot Analysis:

1. **Sale_price**: The boxplot confirms a relatively symmetric distribution with the median around $20,000. There is one visible outlier on the lower end of the price spectrum.
2. **Mileage**: The interquartile range spans from approximately 45,000 to 115,000 miles, with the median at about 85,000 miles. No significant outliers are detected.
3. **Fuel_efficiency**: Shows a median of approximately 37 units with an interquartile range from 26 to 48 units, indicating typical efficiency values for most vehicles in the dataset.
4. **Car_age**: The median age is around 12 years, with an interquartile range from approximately 7 to 18 years, reflecting a dataset weighted slightly toward older vehicles.



Categorical Variable Insights:

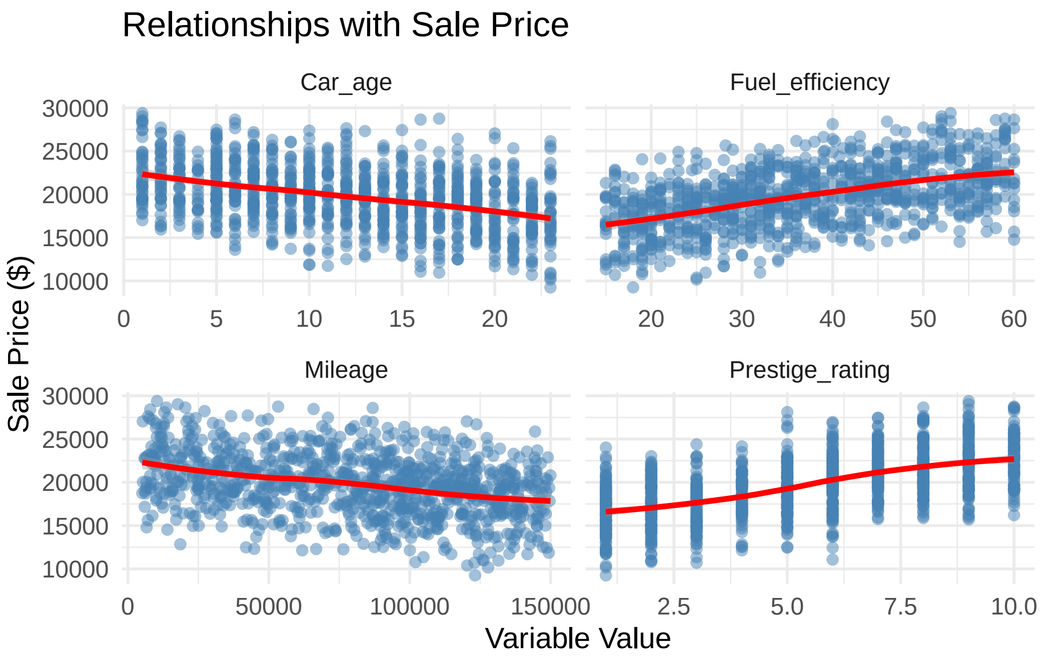
1. **Car_maker**: The dataset includes a variety of manufacturers, with Honda, Ford, Toyota, and Chevrolet being the most represented. This offers good coverage across different market segments, from economy to luxury brands.
2. **Age_category**: The largest segments are "Used (8-15 years)" and "Vintage (>15 years)" vehicles, making up over 60% of the dataset. This matches the car age distribution and indicates that the analysis will be most relevant to the used car market.
3. **Efficiency_category**: The three efficiency categories (Low, Medium, High) are fairly evenly spread out, with a slightly higher number in the Medium category. This balanced distribution will support meaningful comparisons across efficiency levels.
4. **Price_category**: The Mid-range category dominates, as expected from our quantile-based definition, with roughly equal proportions of Budget and Premium vehicles. This classification will be valuable for market segment analysis.

The univariate analysis shows the dataset has good coverage across different vehicle ages, prices, and efficiency levels, providing a strong basis for exploring relationships between these variables in the bivariate and multivariate analyses.

Bivariate Analysis

After examining individual variable distributions, I conducted bivariate analysis to explore relationships between key variables, with a particular focus on factors influencing sale price.

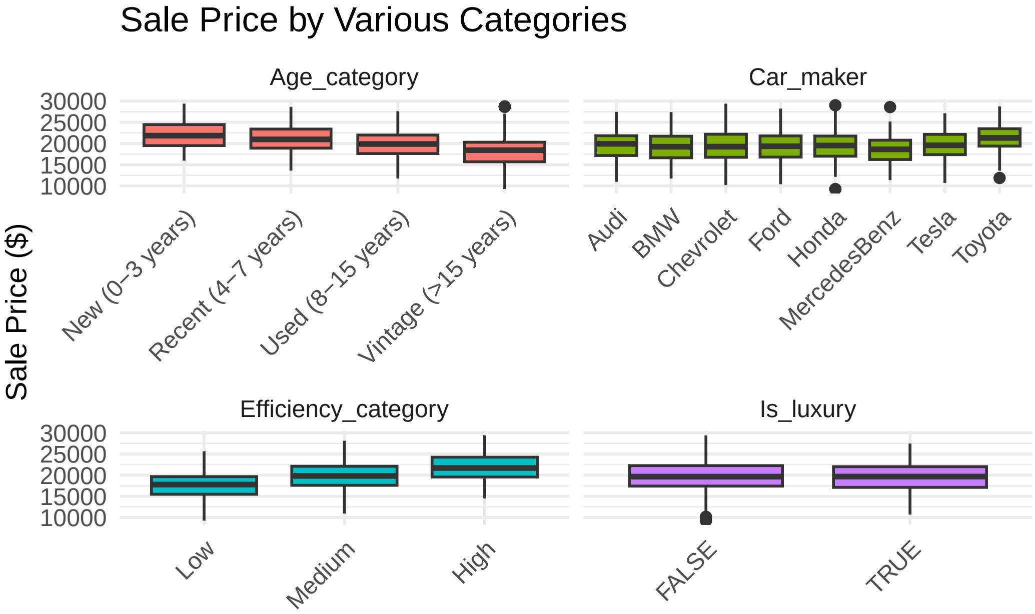
```
`geom_smooth()` using formula = 'y ~ x'
```



Key Relationships with Sale Price:

1. **Car Age vs. Sale Price**: There is a clear negative relationship between car age and price, with newer vehicles having higher prices. The trend line shows a steady decline in value as cars age, with an approximate decrease of $250-300 per year. The wide spread of points around the trend line suggests that other factors also affect pricing.
2. **Fuel Efficiency vs. Sale Price**: A strong positive relationship is evident, with higher fuel efficiency linked to higher sale prices. This suggests that fuel-efficient vehicles are priced higher in the market, with prices increasing by approximately $150-200 for each additional unit of fuel efficiency. This likely reflects consumer preference for economical vehicles, as well as the correlation between newer, more expensive vehicles and better fuel efficiency.
3. **Mileage vs. Sale Price**: As expected, there is a negative relationship between mileage and price. Vehicles with higher mileage sell for lower prices, with the trend showing a decrease of approximately $15-20 for every 1,000 miles. The relationship appears to be relatively linear across the mileage range.

4. **Prestige Rating vs. Sale Price**: There is a strong positive relationship between prestige rating and sale price. Each one-point increase in prestige rating leads to an average price increase of approximately $700-900. This highlights the significant premium consumers are willing to pay for prestigious brands and models.



Categorical Variable Relationships:

1. **Age Category vs. Sale Price**: The boxplots show a clear step-wise decrease in price across age categories. New vehicles (0-3 years) have the highest median price, around $23,000, which steadily decreases through the Recent, Used, and Vintage categories. The Vintage category (>15 years) shows the widest price variation, indicating greater diversity among older vehicles.

2. **Car Maker vs. Sale Price**: There are significant price differences across manufacturers. Toyota and Audi have the highest median prices, while Honda and Ford vehicles are generally priced more moderately. Mercedes-Benz shows the widest price range, likely due to its diverse model lineup, which includes both entry-luxury and premium vehicles.

3. **Efficiency Category vs. Sale Price**: A strong relationship exists between efficiency category and price, with High efficiency vehicles commanding a significant premium (~$20,000) and Low efficiency vehicles priced around $18,000. The increasing median prices across categories highlight consumers' willingness to pay more for fuel efficiency.

4. **Luxury Status vs. Sale Price**: Luxury vehicles command a price premium of approximately $1,500-2,000 over non-luxury vehicles. However, the significant overlap in price

ranges suggests that factors beyond luxury status, such as age, mileage, and efficiency, also play a major role in pricing.

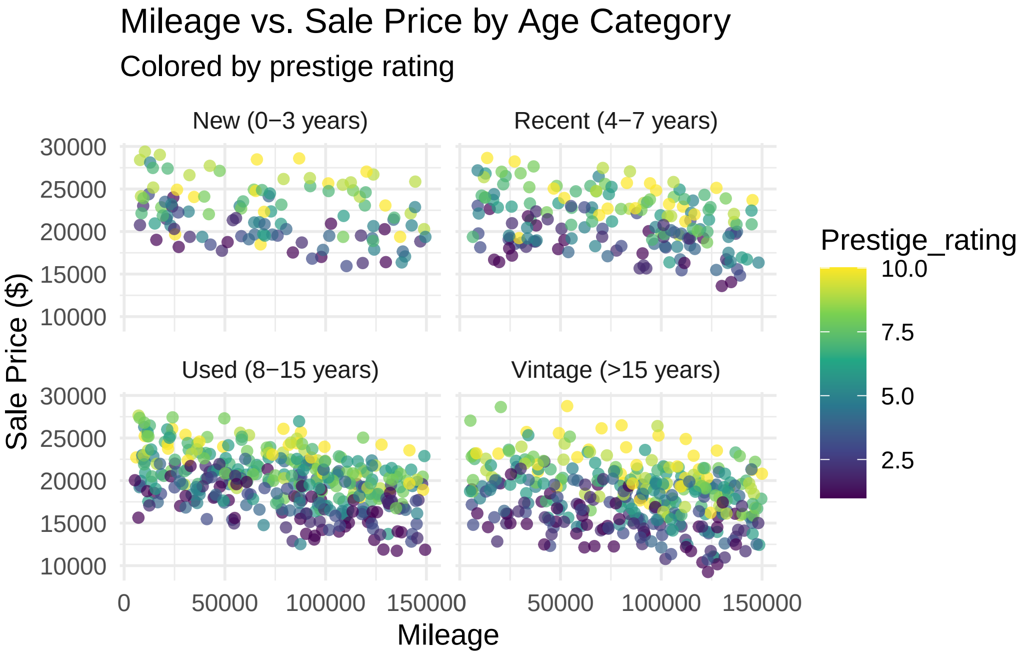
The bivariate analysis reveals several key factors driving car prices:

- Age and mileage have expected negative relationships with price
- Fuel efficiency and prestige rating have strong positive influences
- Brand and luxury status create notable price differentials
- The relationships generally align with market expectations but quantify the specific impact of each factor

These insights offer valuable guidance for the upcoming multivariate analysis, which will examine how these factors interact to influence vehicle pricing.

Multivariate Analysis

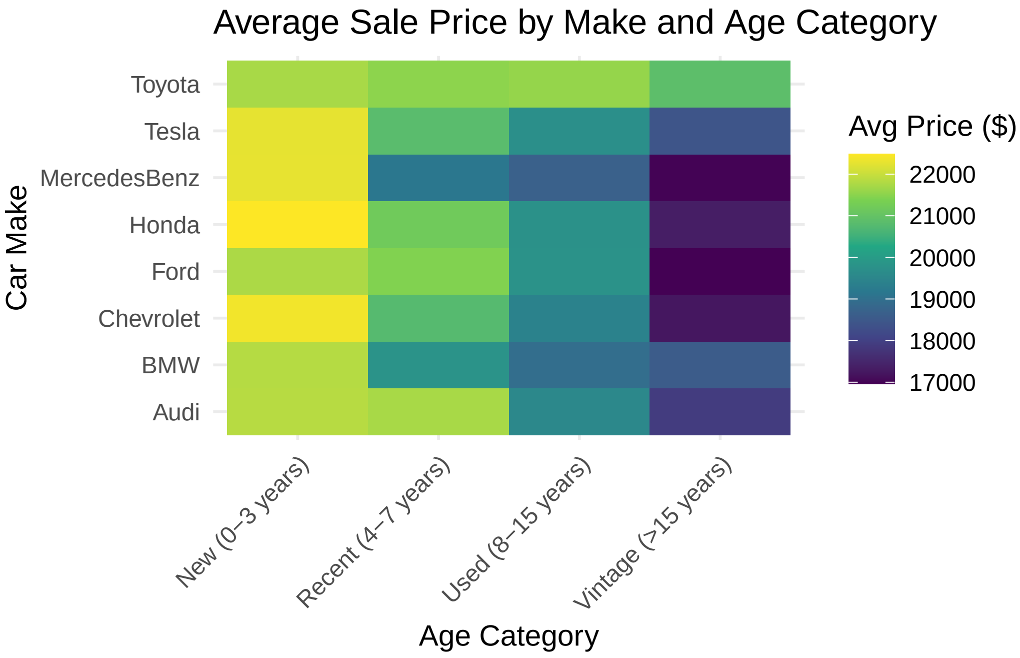
To uncover complex patterns and interactions between multiple variables, I conducted a multivariate analysis examining how different factors combine to influence car pricing.



Analysis of Mileage, Price, Age Category, and Prestige:

This visualization reveals several important insights:

1. **Age-specific pricing patterns**: Within each age category, the relationship between mileage and price shows a gradual negative slope. However, overall price levels decrease as we move from newer to older vehicles.
2. **Prestige impact across age groups**: The color gradient (representing prestige rating) shows that higher prestige vehicles (yellow/green points) consistently command higher prices within each age category, even at similar mileage levels. This suggests that prestige rating remains an important price factor throughout a vehicle's lifecycle.
3. **Price spread variations**: Newer vehicles (0-3 years) show a tighter clustering of prices compared to older categories, indicating that as cars age, other factors not captured in the dataset contribute to greater price variability.
4. **Mileage impact consistency**: The negative relationship between mileage and price remains across all age categories, but appears slightly steeper in newer vehicles, suggesting that additional mileage has a greater impact on the value of newer cars.



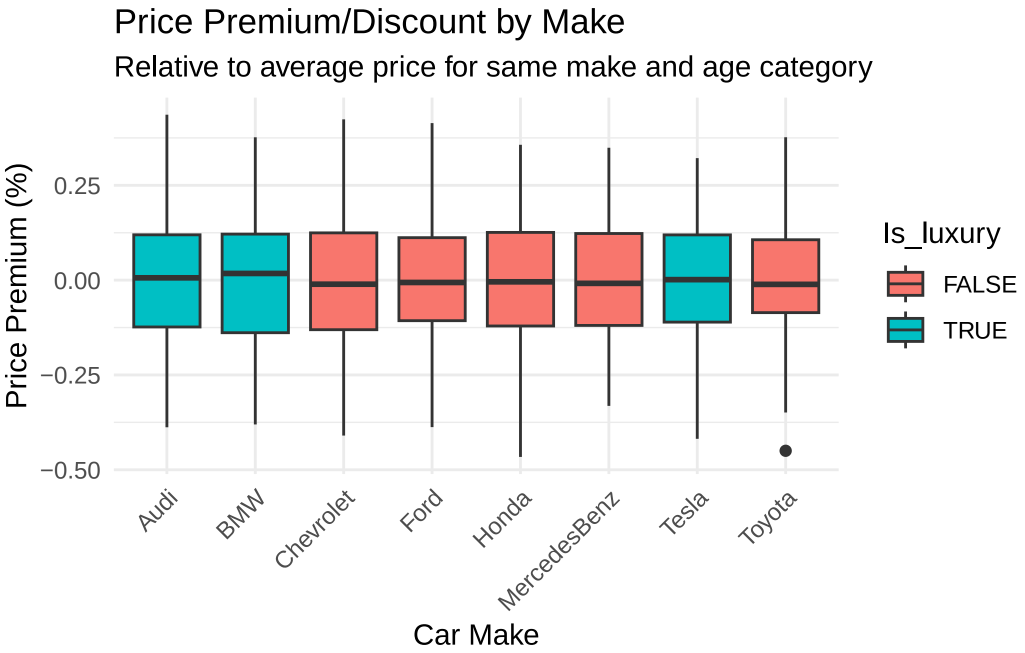
Insights from the Price Heatmap:

1. **Brand-specific depreciation patterns**: The heatmap shows how different brands retain value over time. Toyota and Honda exhibit the most consistent value retention across age categories, with smaller price drops between the New and Vintage categories compared to luxury brands.

2. **Luxury brand premium in newer vehicles**: Mercedes-Benz, Tesla, and Chevrolet command the highest premiums in the newest category but experience steeper depreciation as they age. Mercedes-Benz shows the most significant price decline across age categories.

3. **Resilient mid-market brands**: Toyota maintains relatively high prices even in older categories, indicating better value retention or potentially higher-quality vehicles that continue to command premium prices as they age.

4. **Price segmentation**: Clear price tiers are evident across brands, with luxury brands typically commanding higher prices in newer categories, but converging with non-luxury brands in the oldest category.



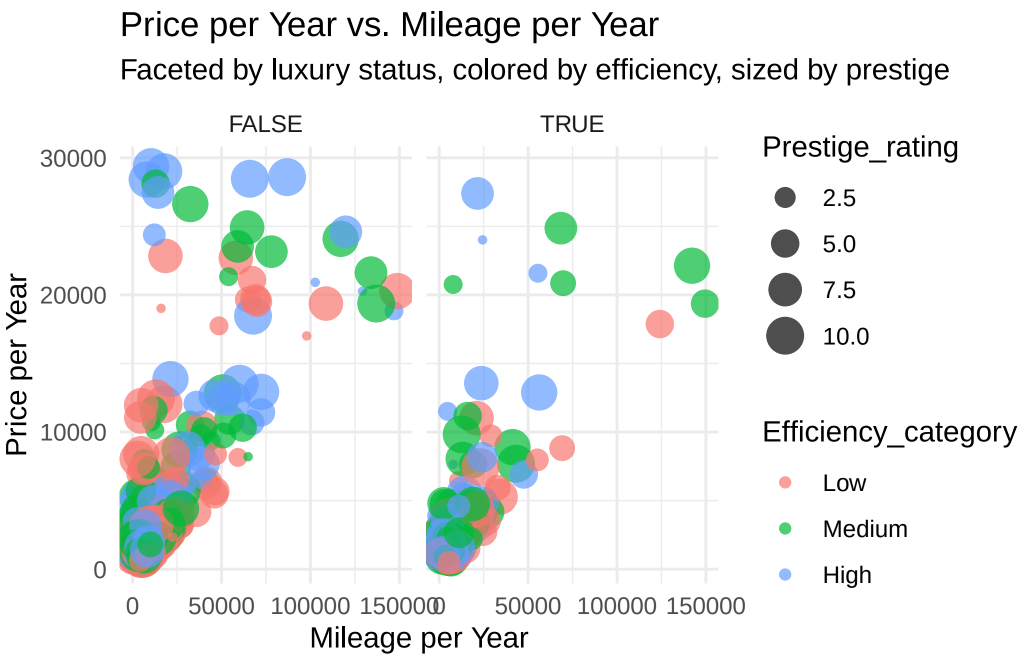
Price Premium Analysis:

1. **Consistent pricing within brands**: Most car makers show relatively tight interquartile ranges around zero, indicating that pricing within makes and age categories is fairly consistent. This suggests that the market effectively prices vehicles based on their key attributes.

2. **Premium variation by luxury status**: Luxury brands (Audi, BMW, Tesla) show slightly wider price premium distributions than non-luxury brands, suggesting greater price variability within these brands, likely due to a broader range of models and option packages.

3. **Price outliers**: Toyota shows a notable negative outlier, representing a vehicle priced significantly lower than its peers. This could suggest a specific model that underperforms in the market or potentially a vehicle with undocumented issues.

4. **Luxury vs. non-luxury comparison**: Interestingly, luxury brands don't always show higher median price premiums than non-luxury brands within their own categories, suggesting that the luxury premium is already reflected in the make and age category baseline.



Complex Interactions Analysis:

1. **Value retention clusters**: Both luxury and non-luxury segments show two distinct clusters: a high price-per-year group (above ~$15,000/year) and a lower price-per-year group (below ~$15,000/year). This bimodal distribution likely reflects different vehicle classes or market segments.

2. **Efficiency impact**: High efficiency vehicles (blue points) are more common in the upper cluster of both luxury and non-luxury segments, indicating that efficiency commands a premium regardless of the vehicle's luxury status.

3. **Prestige effect**: Larger points (higher prestige) are more frequent in the upper clusters, confirming that prestige rating significantly influences a vehicle's value retention in both luxury and non-luxury segments.

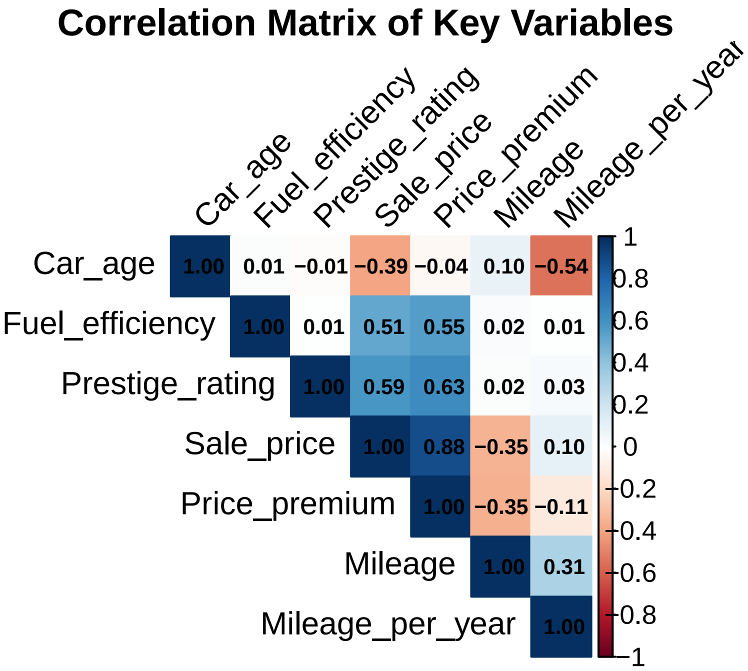
4. **Luxury advantage**: The luxury segment (TRUE) shows a slightly higher concentration of points in the upper cluster compared to the non-luxury segment, indicating better overall value retention for luxury vehicles.
5. **Mileage impact**: Increasing annual mileage generally corresponds to lower price-per-year values in both segments, with a more pronounced effect in the upper clusters. This suggests that high-value vehicles are more sensitive to increased usage.

The multivariate analysis shows complex interactions between factors affecting vehicle pricing:

- Prestige rating remains influential across all age categories and brand types
- Different brands follow distinct depreciation trajectories
- Luxury vehicles show greater price variability but similar relative pricing within their segments
- Market segmentation is evident in the bimodal distribution of value metrics
- Fuel efficiency adds value across all vehicle segments

Correlation Analysis

To quantify the strength of relationships between key numeric variables, I conducted a correlation analysis, which provides a comprehensive overview of how these factors interrelate.



Key Correlation Findings:

1. **Sale Price Determinants**:
   - Strong positive correlation with Price_premium (0.88), indicating that vehicles commanding a premium relative to their peers are indeed priced higher
   - Substantial positive correlations with both Prestige_rating (0.59) and Fuel_efficiency (0.51), confirming these as major drivers of vehicle pricing
   - Moderate negative correlation with Car_age (-0.39) and Mileage (-0.35), validating the expected depreciation effects
2. **Price Premium Factors**:
   - Strong correlation with Prestige_rating (0.63), highlighting how prestige significantly contributes to a vehicle's ability to command premium pricing
   - Strong correlation with Fuel_efficiency (0.55), suggesting that efficiency-conscious consumers are willing to pay above-average prices
   - Negative correlation with Mileage (-0.35), indicating that higher mileage reduces a vehicle's premium pricing potential
3. **Usage Metrics**:
   - Moderate negative correlation between Car_age and Mileage_per_year (-0.54), suggesting that older vehicles are driven less intensively on an annual basis
   - Positive correlation between Mileage and Mileage_per_year (0.31), as expected since total mileage naturally increases with higher annual usage
4. **Independent Factors**:
   - Near-zero correlation between Fuel_efficiency and Prestige_rating (0.01), indicating these are independent value drivers that can be separately optimized
   - Minimal correlation between Car_age and Fuel_efficiency (0.01), suggesting that efficiency is not strongly tied to vehicle age in this dataset

The correlation analysis confirms and quantifies many of the relationships observed in the visualization analyses, while also revealing additional insights into the independence of certain factors. This provides a solid statistical foundation for our recommendations.

Discussion/Conclusion

Key Insights

This comprehensive analysis of the car sales dataset has revealed several valuable insights that can guide business strategy and decision-making:

1. **Primary Price Determinants**:
   - Prestige rating is the strongest individual predictor of sale price, with a correlation of 0.59
   - Fuel efficiency significantly impacts pricing (correlation 0.51), reflecting growing consumer preference for economical vehicles
   - Age and mileage have expected negative effects on price, but with different magnitudes depending on the vehicle segment
2. **Value Retention Patterns**:
   - Toyota demonstrates superior value retention across age categories, maintaining higher relative prices even for older vehicles
   - Luxury brands command high premiums when new but experience steeper depreciation with age
   - Vehicles with high fuel efficiency maintain better value across all age segments
3. **Market Segmentation**:
   - Clear bimodal distribution in value metrics (price-per-year) indicates distinct market segments with different pricing dynamics
   - Luxury and non-luxury markets show similar internal pricing structures but at different absolute price levels
   - Price consistency within makes and age categories suggests an efficient market that prices vehicles primarily based on key attributes
4. **Unexpected Findings**:
   - Prestige rating and fuel efficiency are nearly uncorrelated (0.01), indicating they are independent value drivers
   - Price premium variability is similar between luxury and non-luxury brands
   - Some mid-market brands (particularly Toyota) outperform luxury brands in long-term value retention