# What Drives the Price of a Car? - Analysis Findings

**Jupyter Notebook Link**: [prompt\_II.ipynb](https://github.com/michaeltuszynski/assignment-11.1/blob/main/practical_application_II_starter/prompt_II.ipynb)

## Executive Summary

This analysis explores key factors influencing used car prices using a dataset of 426,880 vehicles. Through data cleaning, feature engineering, and machine learning modeling, we identified crucial elements that impact vehicle valuation and developed actionable recommendations for used car dealerships to optimize their inventory decisions.

## Key Findings

### 1. Mileage Impact

* Low-mileage vehicles command premium prices, averaging $29,315
* Clear price depreciation pattern across mileage categories
* Very high mileage vehicles average $7,739 (73.6% reduction from low-mileage)
* Consistent distribution across mileage categories (~22,000 vehicles each)

### 2. Manufacturer Market Position

#### Volume Leaders

* Ford: Market leader with 19,478 units at $17,554 average
* Chevrolet: Strong second with 16,640 units at $18,270 average
* Toyota: Solid third position with 9,576 units at $14,369 average

#### Premium Segment

* RAM: Highest average price at $26,181
* GMC: Premium positioning at $22,534 with 4,572 units
* Jeep: Strong mid-premium at $18,875 with 5,466 units

#### Value Segment

* Honda: 6,452 units at $9,293 average
* Nissan: 5,897 units at $13,160 average

### 3. Market Segmentation

* Economy Segment ($4,695 average)
  + 36,574 vehicles
  + $1,892 standard deviation
  + Most consistent pricing
* Mid-Range Segment ($11,998 average)
  + 36,545 vehicles
  + $2,925 standard deviation
  + Balanced price variation
* Luxury Segment ($30,842 average)
  + 36,461 vehicles
  + $9,288 standard deviation
  + Highest price volatility

## Strategic Recommendations

### Inventory Strategy

1. Maintain balanced segment distribution (~36,500 units each)
2. Focus on proven volume leaders while keeping premium options
3. Use mileage-based pricing with 73.6% total depreciation curve

### Pricing Approach

1. Economy segment: Tight pricing (±$1,892 variation expected)
2. Mid-range: Moderate flexibility (±$2,925 variation expected)
3. Luxury segment: Wide pricing range (±$9,288 variation expected)

## Implementation Plan

### Short-term (0-3 months)

* Implement mileage-based pricing guidelines
* Establish manufacturer mix targets:
  + 40% volume leaders (Ford, Chevrolet, Toyota)
  + 30% premium segment (RAM, GMC, Jeep)
  + 30% value segment (Honda, Nissan)

### Medium-term (3-6 months)

* Develop segment-specific pricing strategies
* Create inventory monitoring dashboard tracking:
  + Mileage distribution
  + Manufacturer mix
  + Segment balance

### Long-term (6-12 months)

* Implement segment-specific acquisition strategies
* Optimize pricing based on:
  + Vehicle age
  + Manufacturer value retention
  + Local market segment demand

## Methodology

* Comprehensive data cleaning and preprocessing
* Feature engineering including mileage and age calculations
* Random Forest regression modeling
* Segment and manufacturer analysis
* Price variation analysis by category

## Data Overview

* Initial dataset: 426,880 vehicles
* Final cleaned dataset: ~109,580 vehicles after handling missing values
* Key features: price, mileage, manufacturer, year, condition
* Comprehensive coverage across all market segments

## Model Performance

* Best performing model: Random Forest (R² score: 0.729)
  + Significantly outperformed linear models (R² = 0.617)
  + Optimized with 500 trees, max depth of 30
  + Cross-validated across 5 folds
  + Explains ~73% of price variation

### Model Insights

1. Most Important Price Factors:
   * Vehicle age (25.2% importance)
   * Odometer reading (16.1% importance)
   * Year (14.5% importance)
   * Drive type (FWD) (14.2% importance)
   * Miles per year (5.2% importance)
2. Model Comparison:
   * Linear models (Linear Regression, Ridge, Lasso) performed similarly (R² = 0.617)
   * Random Forest showed superior performance (R² = 0.729)
   * Non-linear relationships better captured by Random Forest
3. Validation:
   * Consistent performance across cross-validation folds
   * Hyperparameter optimization via grid search
   * Robust feature importance analysis