Executive Summary

This project investigates the key drivers of used-car prices using a large dataset of over **426,000 vehicles** sourced from a publicly available Kaggle dataset. The analysis follows the **CRISP-DM framework**, moving from business understanding through deployment, and aims to provide actionable insights for a used-car dealership seeking to optimize inventory decisions and pricing strategies.

Business Objective

Used-car dealerships operate in a dynamic market with significant variability in vehicle prices driven by age, mileage, condition, model type, and brand perception. The primary goal of this project is to determine **which attributes most strongly influence used-car prices** and to develop a **predictive model** that can provide consistent, data-driven valuation guidance for new inventory.

Data Preparation & Exploration

Initial exploration revealed substantial skew in price and mileage variables, missing values in key fields, and high-cardinality categorical features (such as specific model names). To prepare the dataset for analysis:

- Erroneous entries (e.g., unrealistic prices, mileage outliers) were removed.
- Missing values were imputed using median (numeric) and most frequent (categorical) strategies.
- New features such as **vehicle age** and **high-mileage indicators** were engineered.
- Target variable **price** was log-transformed to stabilize variance and improve model performance.
- Categorical data was encoded using **OneHotEncoding**, and numeric features were standardized for consistency.

This produced a clean, scalable dataset for modeling.

Modeling Approach

Due to the high dimensionality of the data—particularly after one-hot encoding—traditional linear regression models become unstable. To address this, the project employed a modeling pipeline consisting of:

1. **Preprocessing (imputation, scaling, encoding)**
2. **Dimensionality reduction using TruncatedSVD**
    - Handles sparse, high-cardinality categorical features efficiently
    - Provides dense latent components suitable for linear modeling
3. **Ridge Regression**
    - Reduces variance
    - Controls overfitting
    - Stabilizes predictions by shrinking coefficients

Hyperparameters for both SVD (number of components) and Ridge (regularization strength) were optimized using **GridSearchCV with 5-fold cross-validation**.

The final model achieved strong predictive performance, with errors typically within a reasonable range in dollar terms—sufficient for guiding pricing and acquisition decisions.

Key Insights

Analysis of the data and model results revealed several consistent price drivers:

- **Age and mileage** remain the most influential factors. Newer, lower-mileage vehicles command significantly higher prices.
- **Brand and model** matter: Toyota, Honda, and certain luxury brands retain value better than the market average.
- **Drive type and powertrain** contribute to resale value—4WD/AWD and hybrid options generally increase price.
- **Vehicle condition** and specific body types (SUVs and trucks) show meaningful price premiums.
- **High-mileage thresholds** (e.g., >150,000 miles) sharply reduce expected price, confirming industry expectations.

These insights help dealerships make informed decisions regarding **acquisition, pricing, and trade-in valuations**.

Deployment & Business Impact

The complete modeling pipeline can be deployed in several ways:

- **Batch scoring** new inventory to compare market-aligned prices against current listing prices.
- **Real-time valuation tools** for trade-ins or auction purchases.
- **Integration into dealership CRM or inventory systems** for automated alerts and pricing adjustments.

This empowers dealerships to:

- Reduce pricing uncertainty
- Improve profit margins
- Enhance purchasing decisions
- Understand which vehicles retain value and should be prioritized

---

Conclusion

By combining rigorous data preparation with dimensionality reduction and regularized regression, this project delivers a robust, interpretable model that supports accurate used-car pricing. The insights derived from the analysis align with market expectations and provide valuable guidance for optimizing dealership operations.

The final modeling pipeline is reproducible, scalable, and suitable for operational deployment, enabling ongoing data-driven decision-making in a competitive market.