Final Report: Price Sensitivity to Mileage Across Car Brands & Predictive Modeling of Second-Hand Car Prices

1. Project Overview

This comprehensive study explores the relationship between mileage and car pricing across different brands, especially distinguishing between luxury and economy segments. Using a dataset of 531 used car listings, we conducted both exploratory data analysis (EDA) and machine learning modeling to:

- Test whether mileage impacts car price differently across brand classes.
- Predict second-hand car prices based on various vehicle attributes.

---

2. Dataset Summary

- **Total Records**: 531 used car listings
- **Key Features**:
    - Car brand, model, manufacture year
    - Mileage (in km), fuel type
    - Engineered features: car_age, log_price, log_mileage, price_per_km, brand_classification (luxury/economy), mileage_per_year, mileage_penalty, brand/model encodings

---

3. Hypotheses & Exploratory Analysis

Hypotheses

- **H1 (Alternative Hypothesis)**: Luxury brands lose value more rapidly with mileage than economy brands.
- **H0 (Null Hypothesis)**: Price sensitivity to mileage is uniform across all brands.

Findings

- **Pearson Correlation**: Mileage vs. Price:
    - Overall: -0.38 (p = 2.90e-16)
    - Economy Brands: -0.34 (p = 2.62e-04)

- **Car Age**: Strongest individual predictor of price
- **Brand Sensitivity**:
    - BMW (luxury) shows steep value depreciation with mileage.
    - Kia (economy) holds value well even with increasing mileage.

Conclusion:

We reject the null hypothesis. Luxury vehicles show significantly more price sensitivity to mileage than economy vehicles, confirmed both statistically and visually through correlation plots and regression analyses.

---

4. Machine Learning Model Performance

Model: Regression-based predictive model using 15 features.

Metric	Training Set	Test Set
R^2	0.8855	0.3976
MAE (€)	€1,505.93	€3,222.20
MAE (% of avg price)	9.4%	20.4%

- **Interpretation**: High performance on training data shows good model fit, but low test performance suggests **overfitting**. More data or regularization is needed.

---

5. Key Predictive Features

Rank	Feature	Importance
1	Brand Classification Encoded	0.435

2	Car Age	0.117
3	Mileage per Year	0.081
4–7	Encoded Brand/Model Features	~0.06 each

- Brand classification (luxury vs. economy) is by far the most influential variable.

---

6. Case Study: Mileage Depreciation

BMW X5 (Luxury):

- Depreciation from 10k to 90k km: **~€20,314**

SEAT Leon (Economy):

- Depreciation from 10k to 90k km: **~€5,791**

These case studies affirm that **luxury vehicles depreciate faster** per kilometer than economy ones.

---

7. Key Takeaways

- **Mileage sensitivity varies significantly across car brands**, with luxury brands being more affected.
- **Car age and brand prestige** play crucial roles in determining used car prices.
- **Model performance issues on the test set** indicate room for improvement in generalization—cross-validation and more robust feature selection are recommended.

---

8. Recommendations for Future Work

- **Enhance Model Generalizability**:

    - Use regularization (e.g., Lasso/Ridge)
    - Employ cross-validation techniques
- **Data Expansion**:
    - Include features like accident history, number of owners, regional price differences
- **Granular Brand Categorization**:
    - Subdivide by brand reliability rankings or fuel efficiency to better segment market impact

---

Conclusion

This project successfully validated the hypothesis that brand classification impacts mileage sensitivity in second-hand car pricing. Additionally, the regression model demonstrated strong potential for price prediction, though future work is needed to refine its predictive accuracy and robustness.