

USED CAR PRICE PREDICTION

MSBA 503

Dylan Borchert, Cooper Shults

Problem Description

- Car Values Depreciate
- Prices Vary Widely
- Buyer Trust is Low

Importance

- Economic Importance
- Buyer Knowledge
- Data Backed Decisions

Stakeholders

- Dealerships
- Sellers
- Buyers
- Marketplaces
- Insurance Companies
- Manufacturers

Implications

- Secondary dataset creates more robust model
- Age has the most importance in price prediction
- Combined model will reduce dataset bias

Dataset

ORIGINAL AND NEW DATASET

Primary dataset: (4009, 12)

Secondary dataset: (5000, 10)

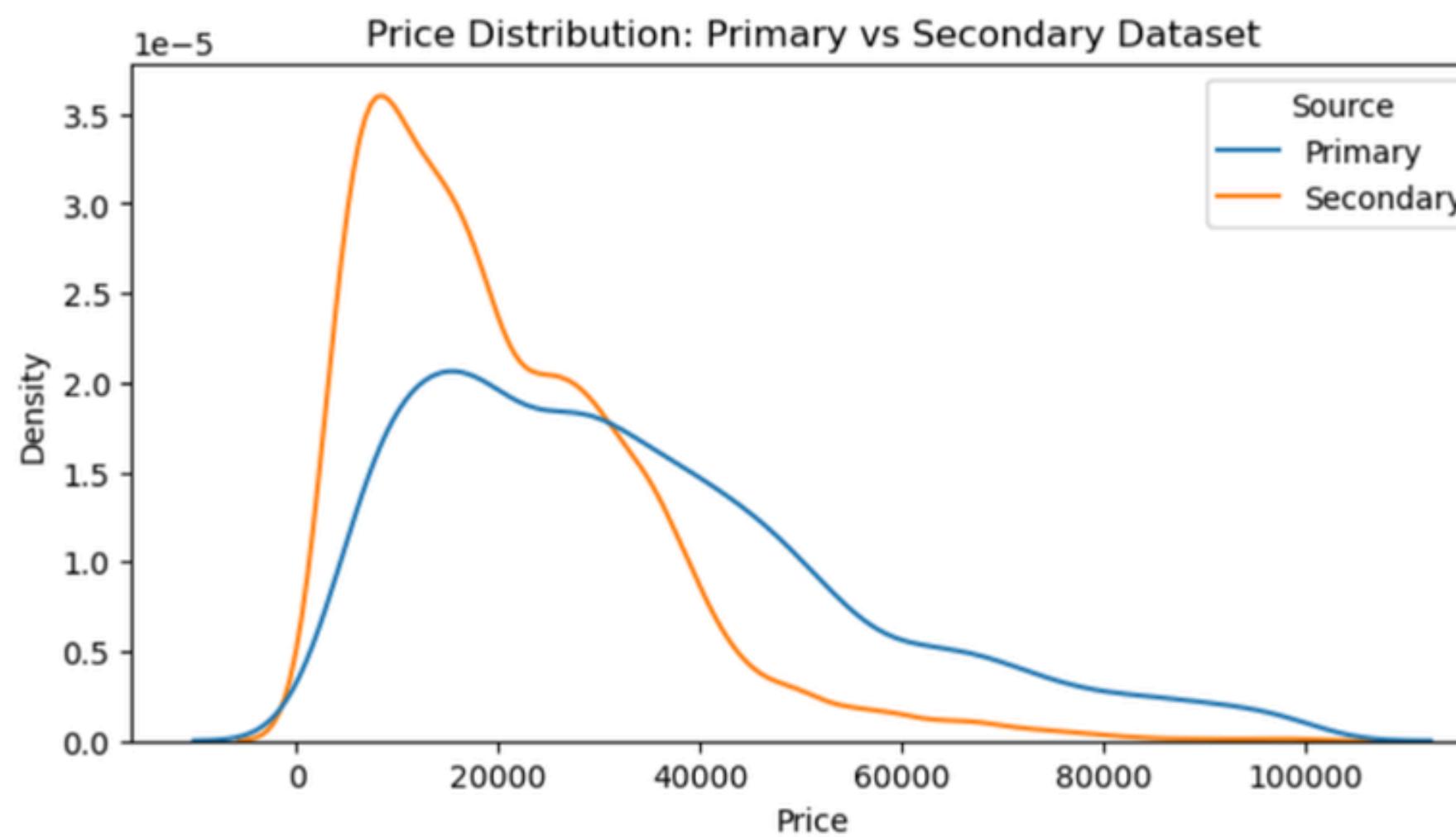
Combined dataset

(8725, 15)

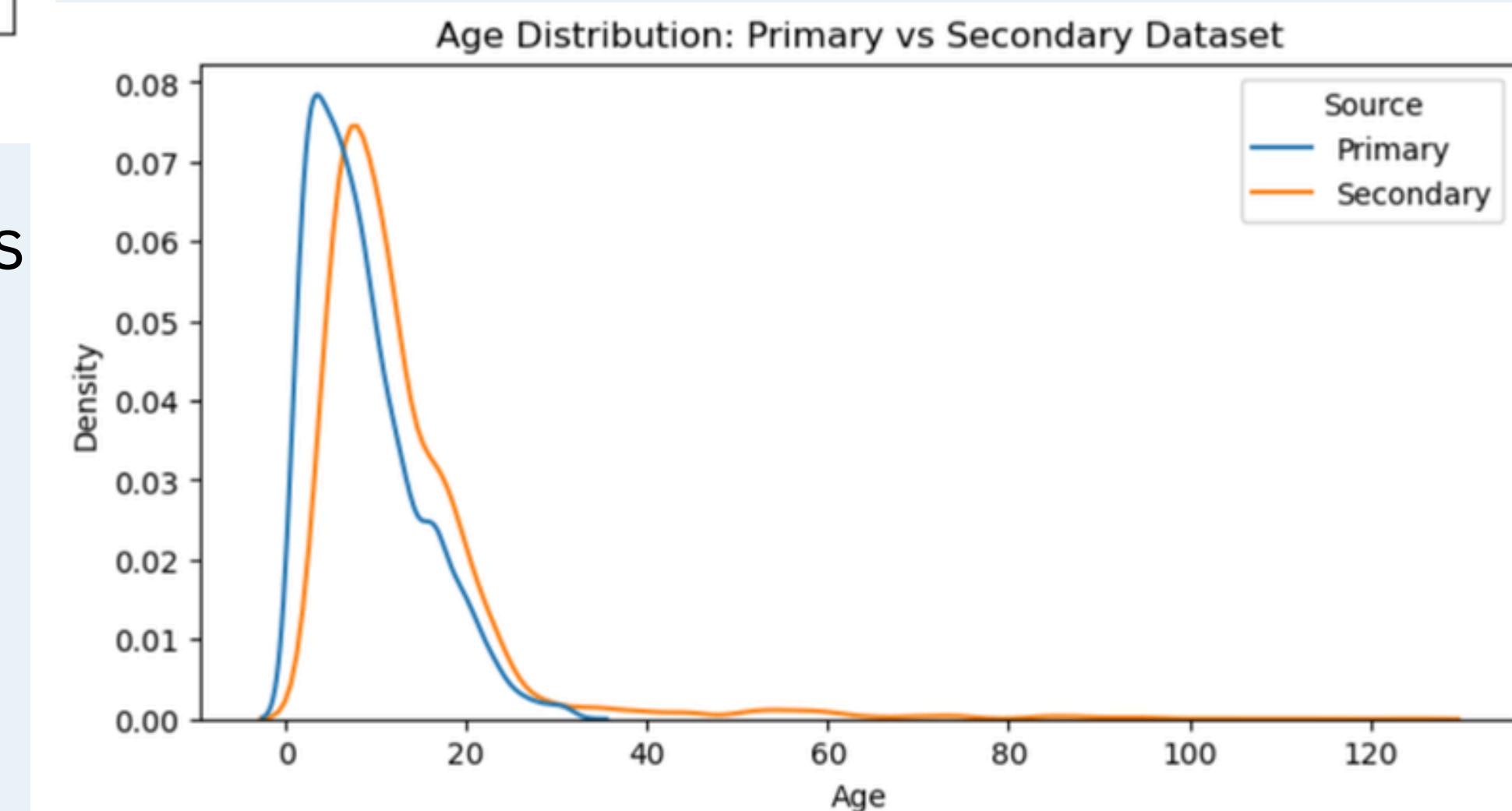
Data Cleaning

- \$2000 < Price > \$100,000
- Mileage < 300,000
- Age column: 2025 - model year

AGE & PRICE DISTRIBUTION COMPARISON

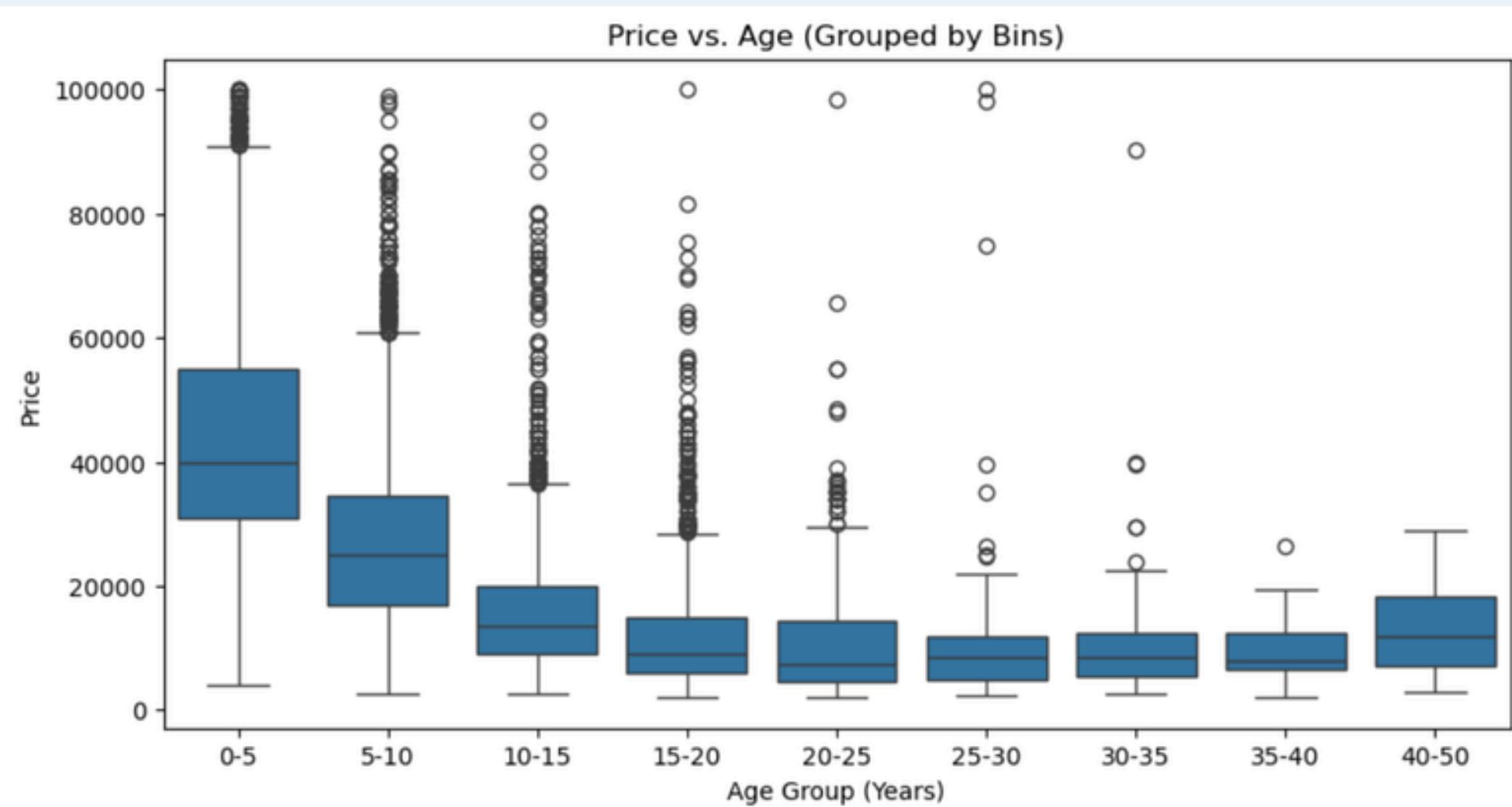


- New dataset has more typical used cars
- Age distribution is very similar

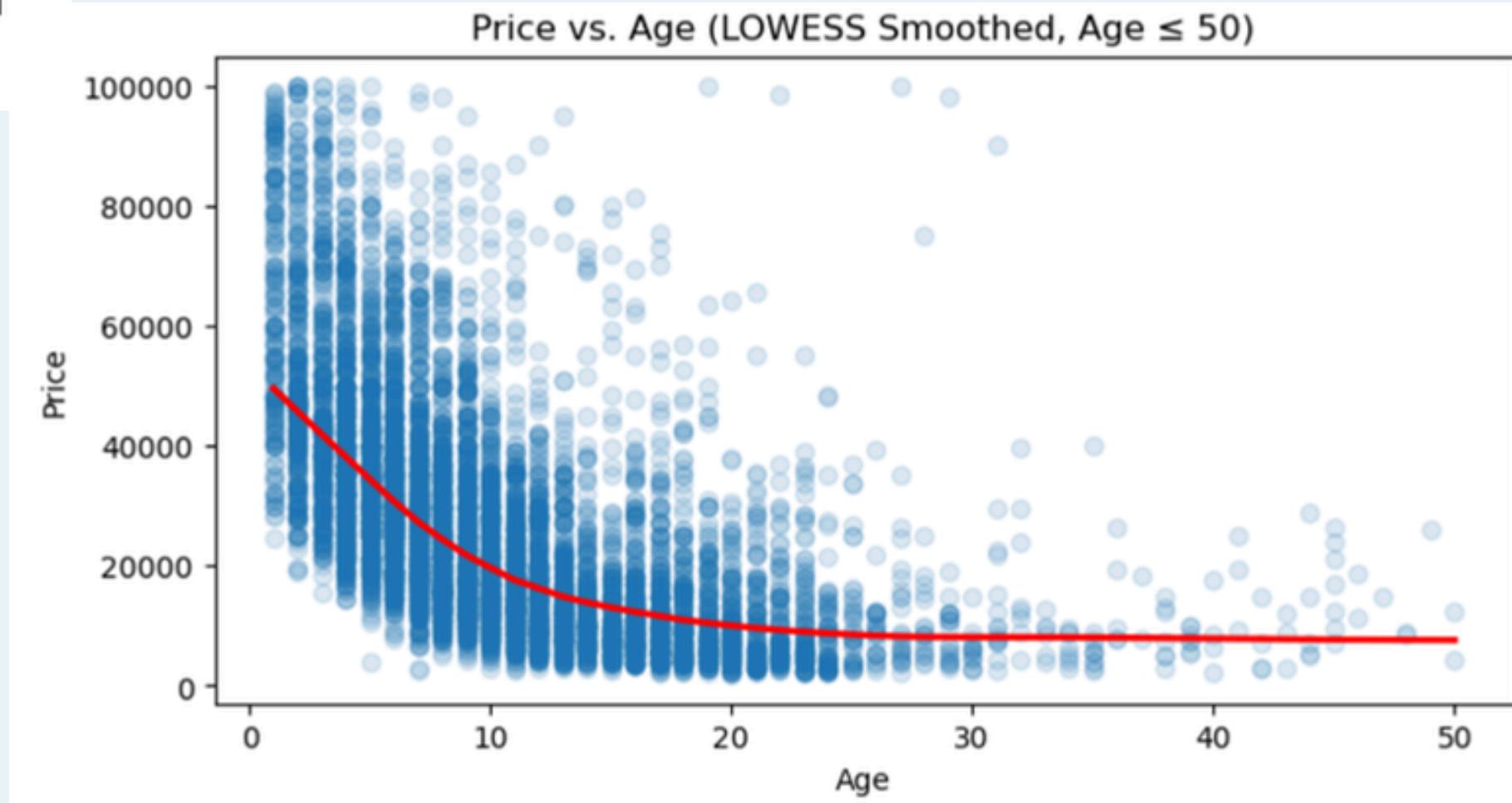


Findings

PRICE VS AGE CHARTS

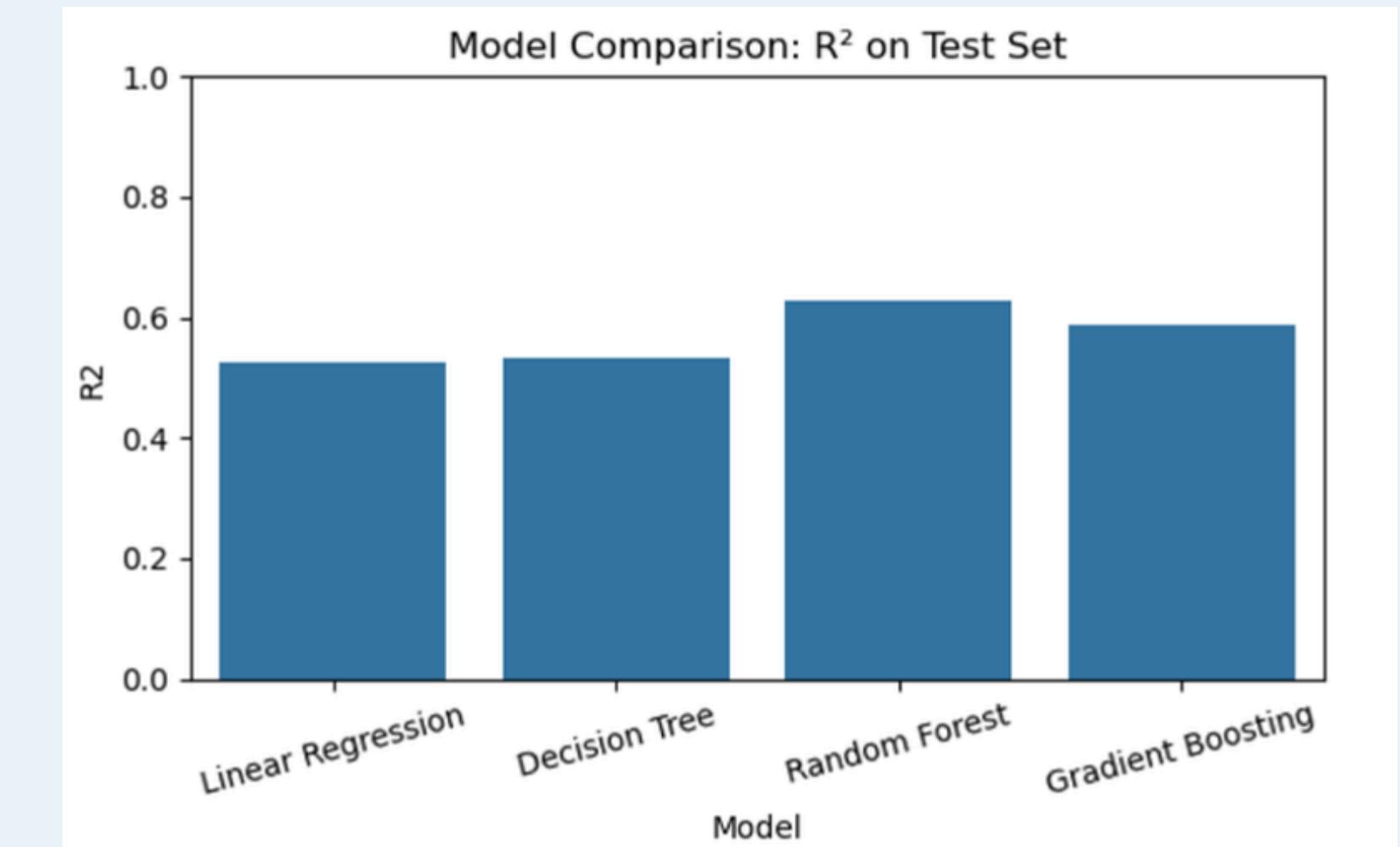


- After 10+ years price is fairly negligible
- Less outliers with each new bin
- Clear downward trendline in the first twenty years
- Some outliers 20+ due to old luxury cars



NEW MODELS

- 4 models including linear regression
- Linear regression performed the worst
- Random Forest was the best
 - Smallest average error
 - Makes fewer large errors
 - Explains more of the used car price variation than other models



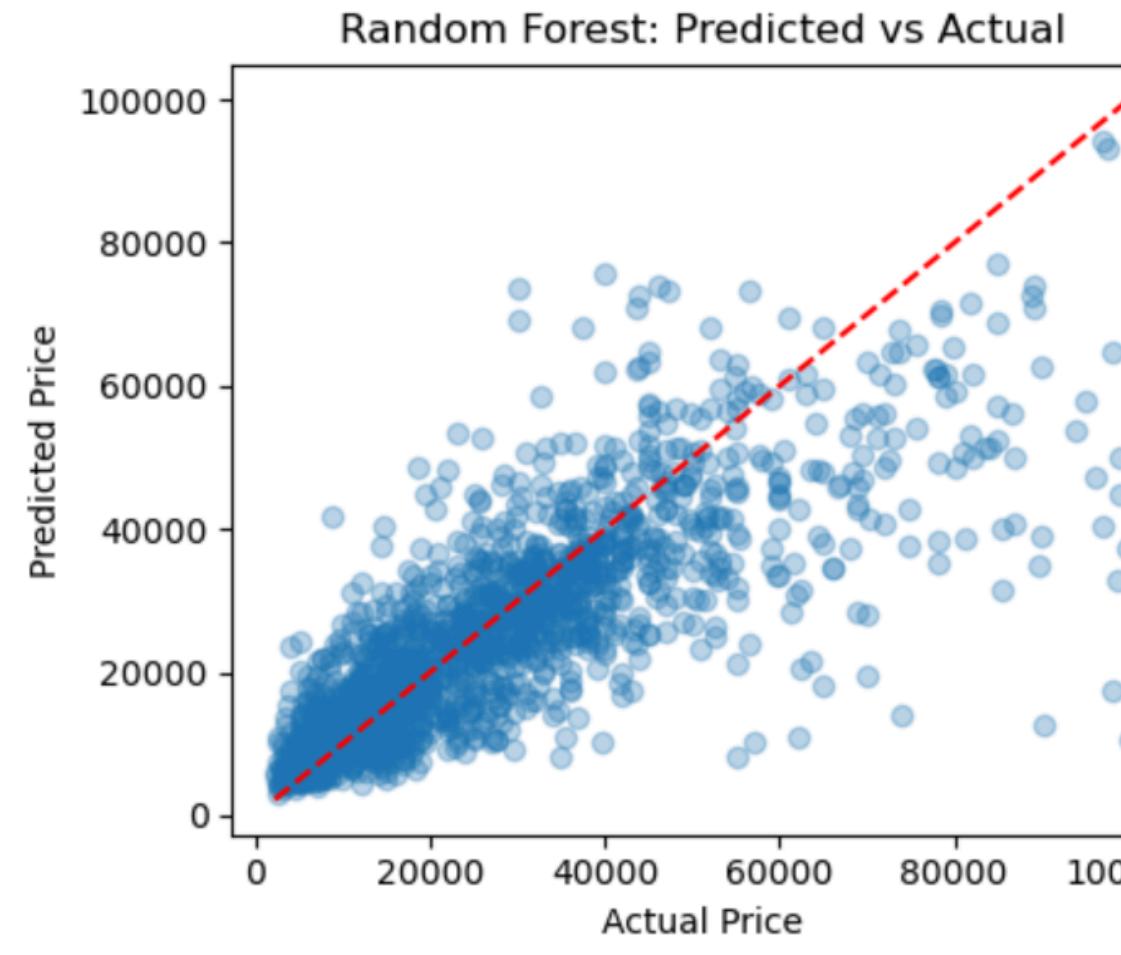
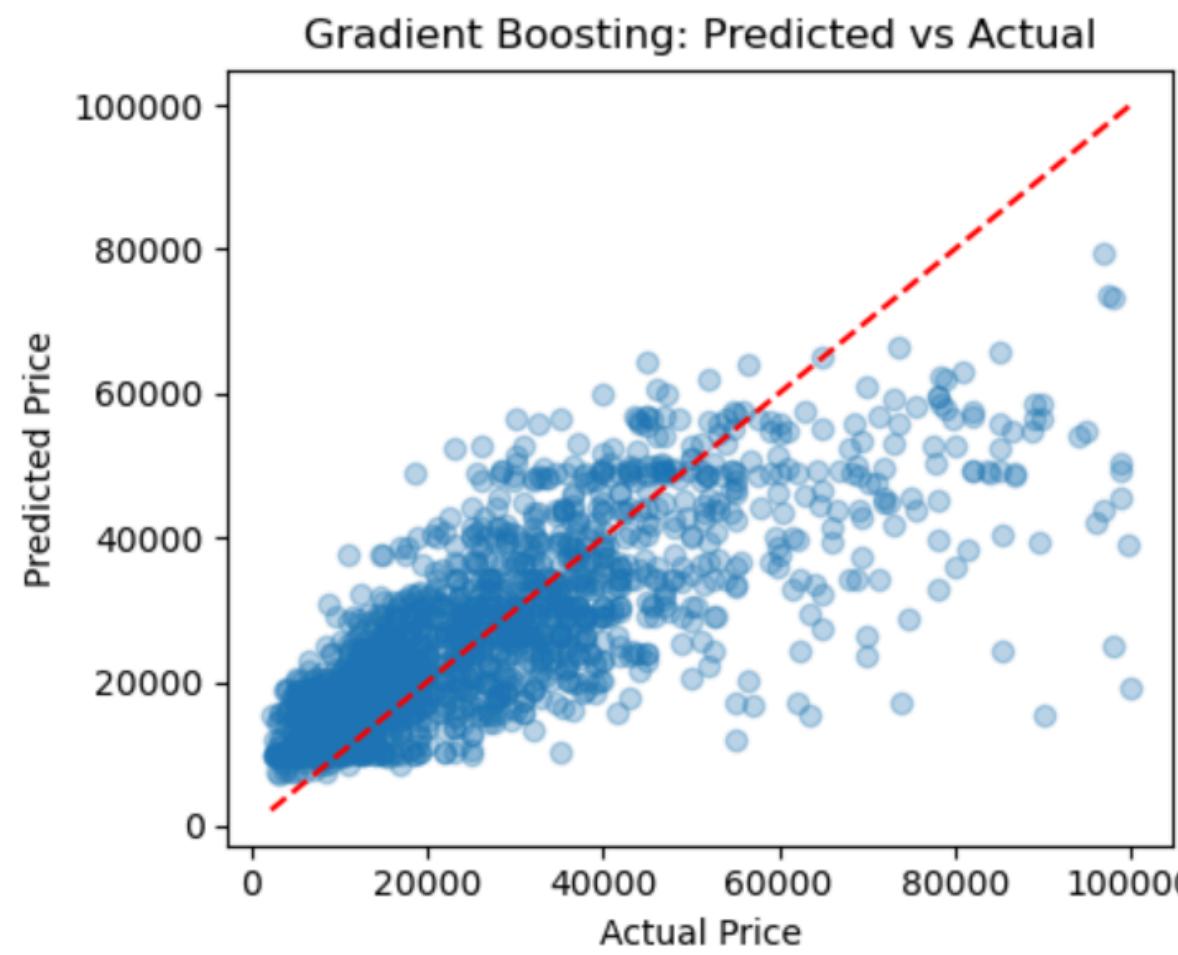
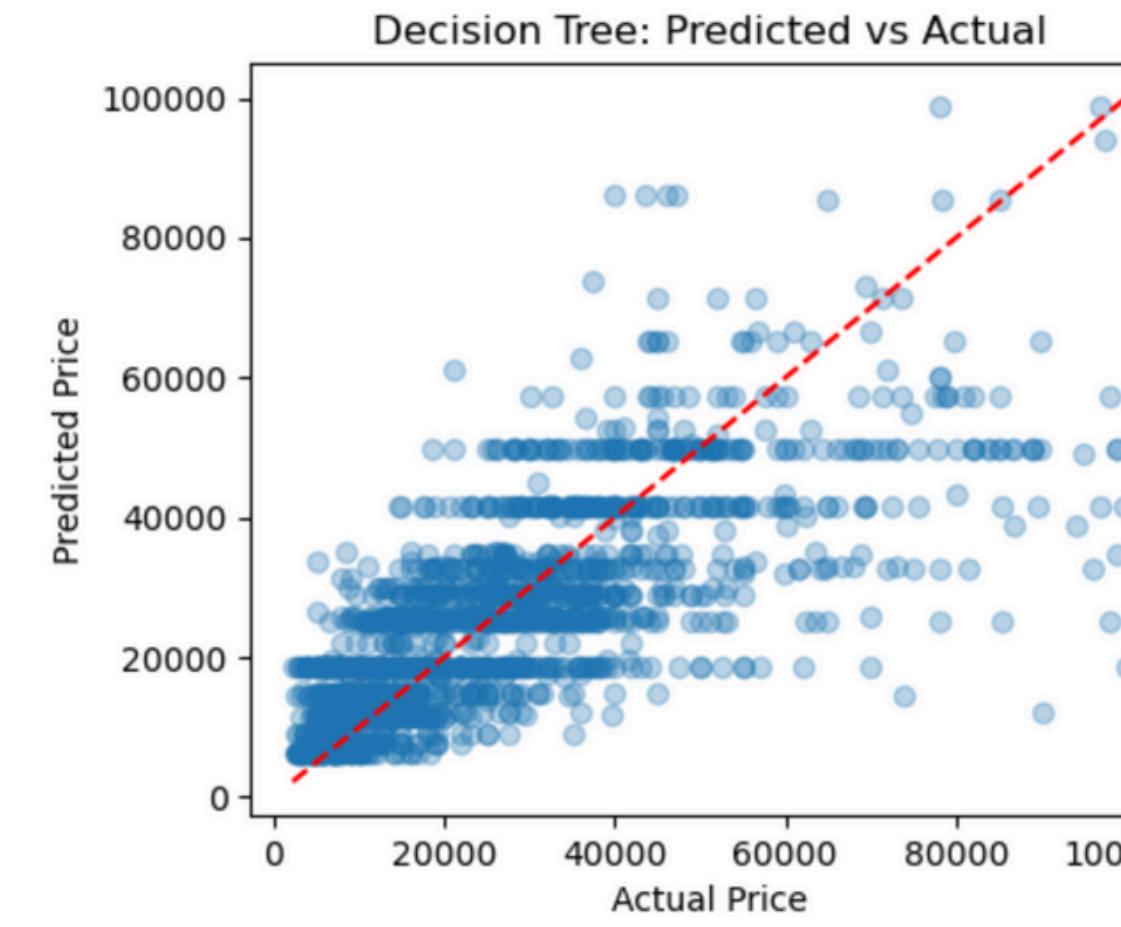
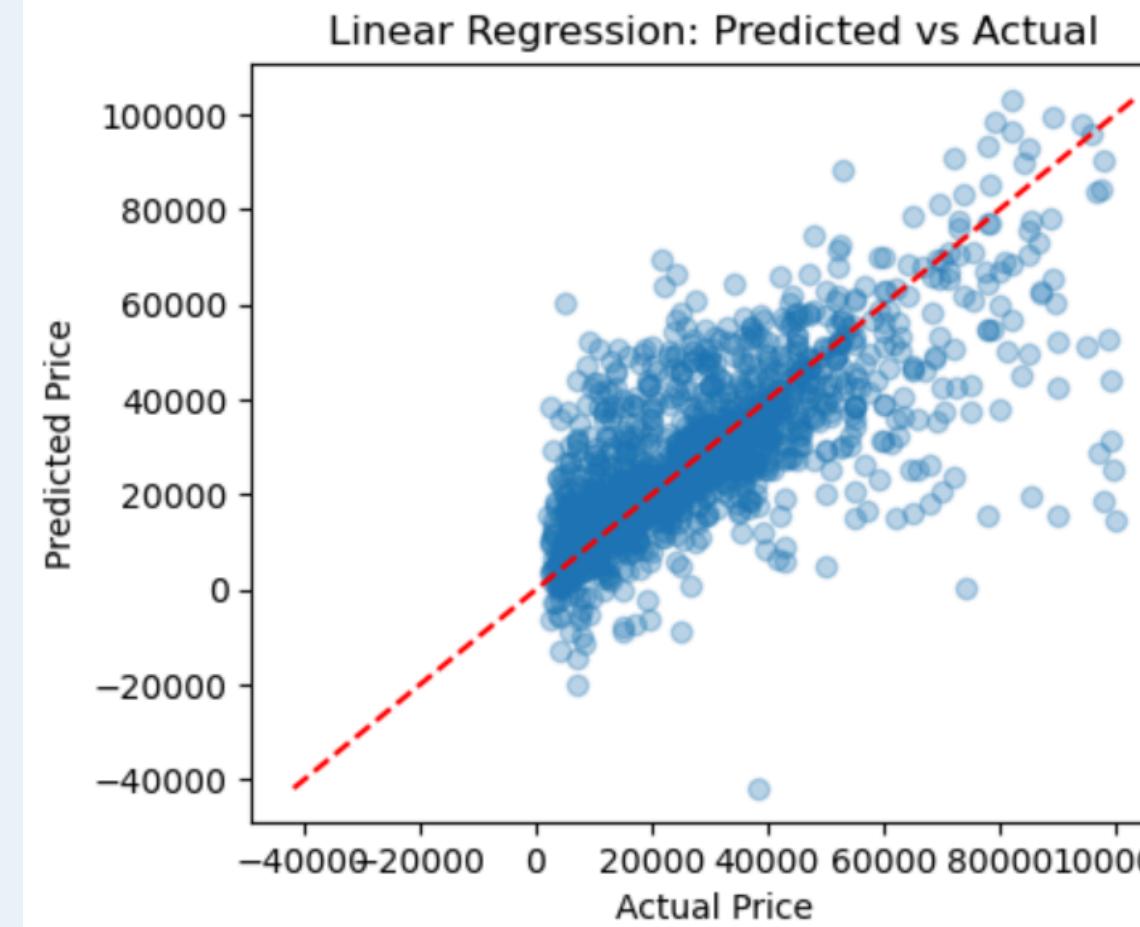
	Model	MAE	RMSE	R2
2	Random Forest	7346.102168	11577.860474	0.626929
3	Gradient Boosting	8680.415417	12179.899888	0.587122
1	Decision Tree	9118.015878	12972.050439	0.531670
0	Linear Regression	8230.758774	13046.863117	0.526253

DATA TYPES AND IMPORTANCE

	Feature	Importance
1	Age	0.681577
3559	Fuel_type_diesel	0.038823
3561	Fuel_type_gas	0.037226
0	Milage	0.026418
3620	Accident_None reported	0.020027
3621	Accident_Unknown	0.014094
418	Model_911 Carrera S	0.013546
3554	Fuel_type_Diesel	0.010748
1370	Model_R1S Adventure Package	0.007000
3585	Transmission_7-Speed A/T	0.006549
3613	Transmission_Transmission w/Dual Shift Mode	0.005663
705	Model_Corvette Stingray w/2LT	0.005394
3569	Transmission_10-Speed A/T	0.005388
2661	Model_i8 Base	0.004604
85	Brand_nissan	0.004596

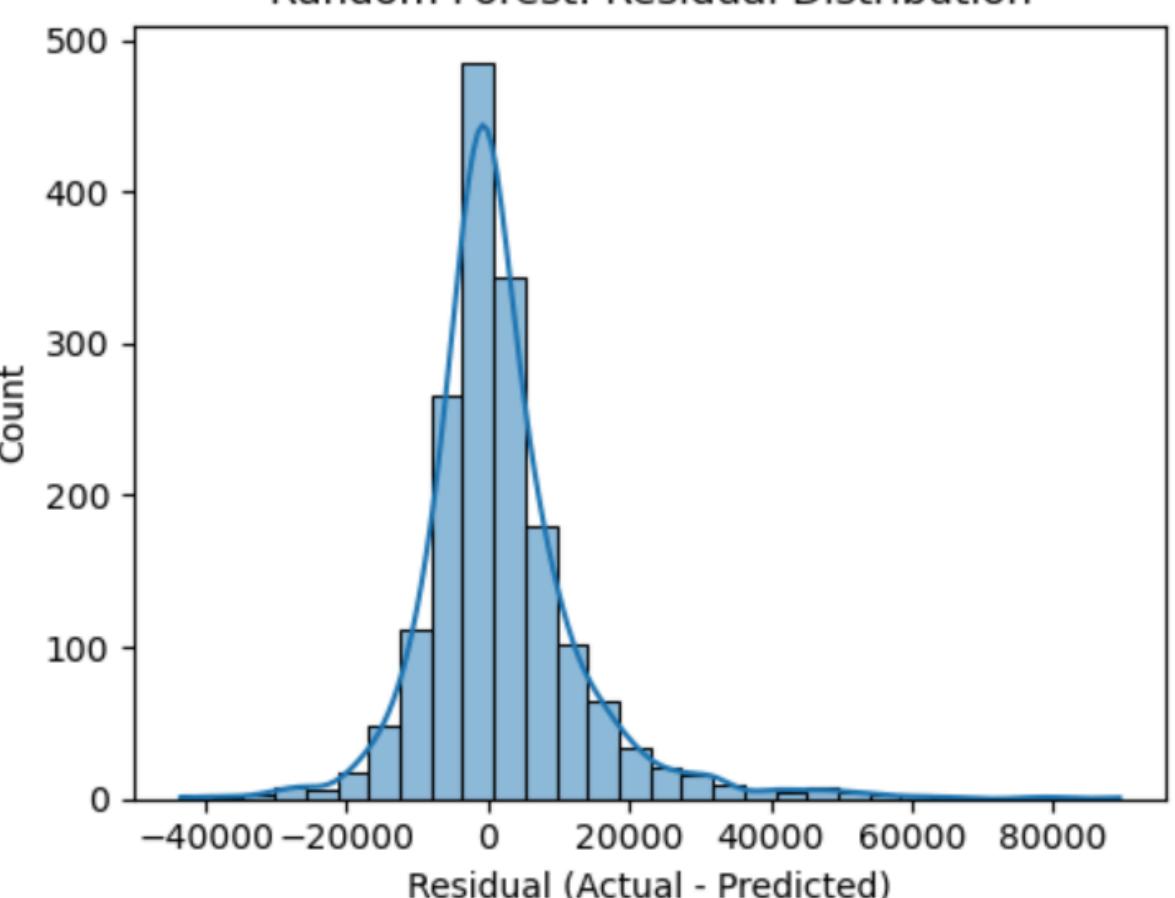
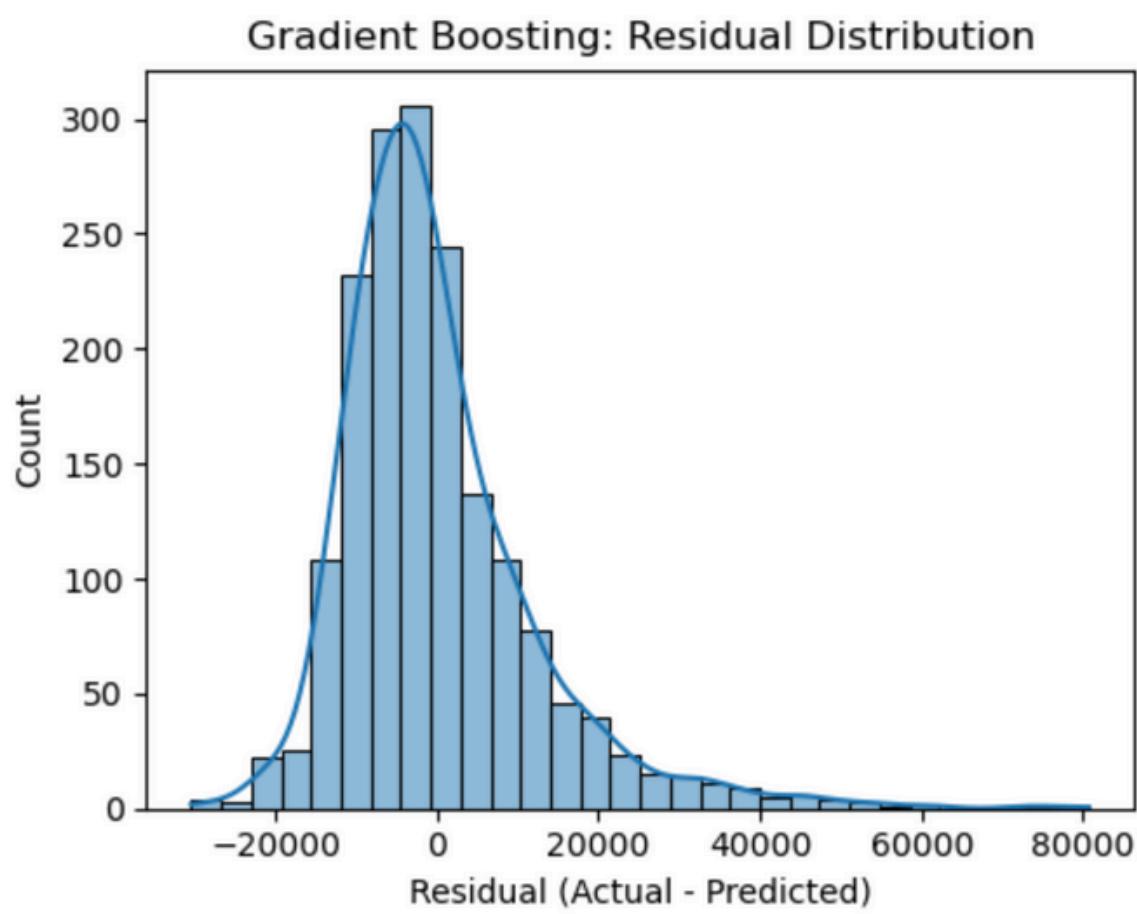
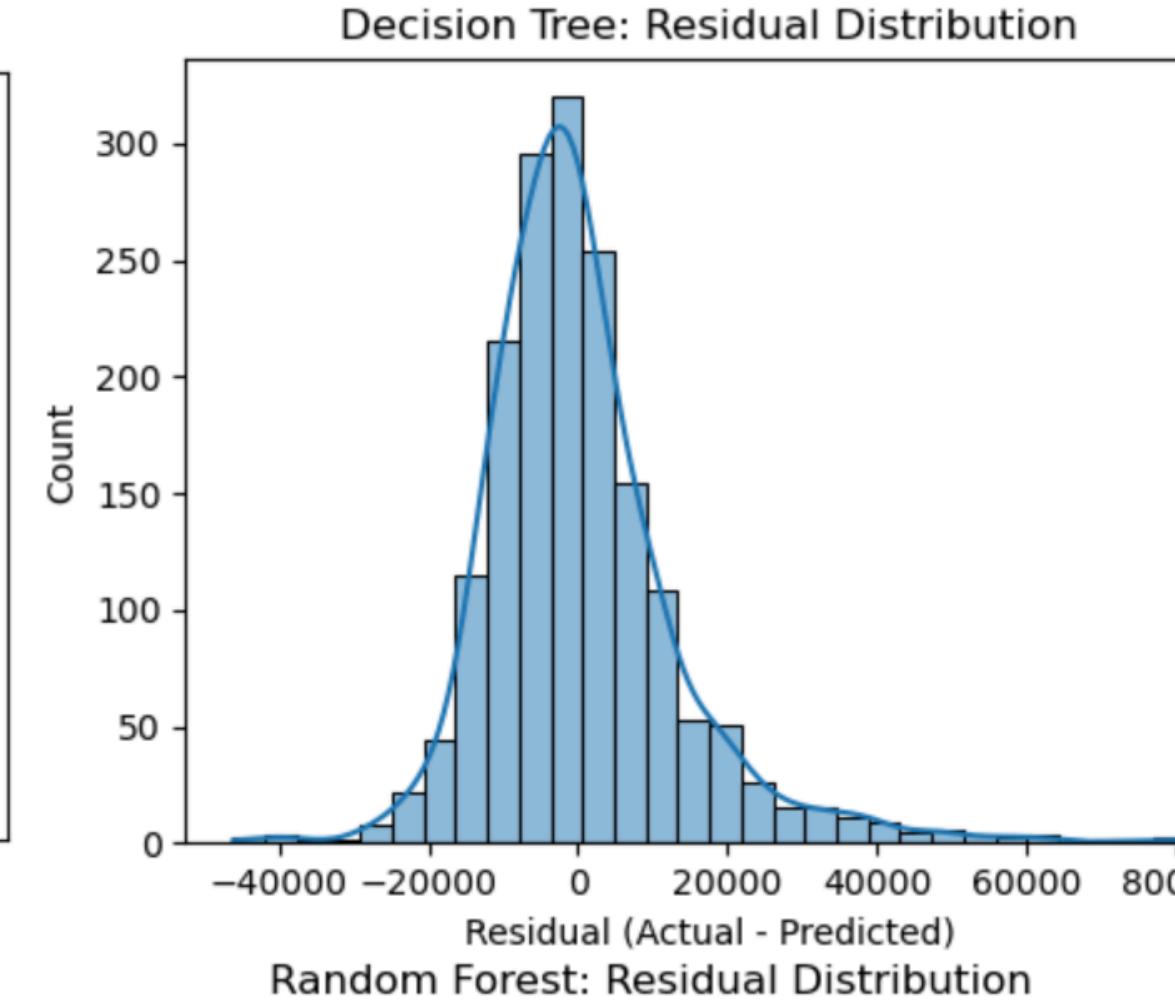
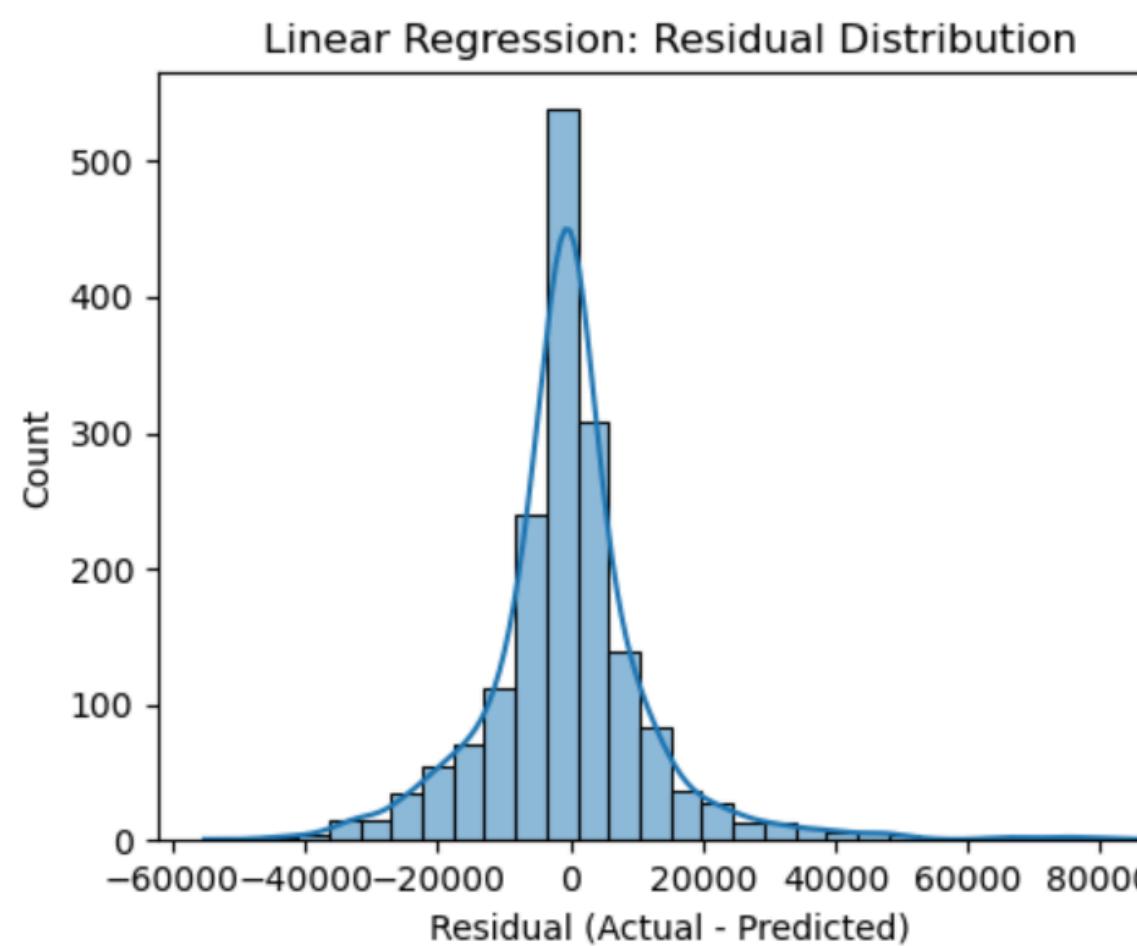
- **Age is by far the most important variable in pricing a car**
- **Diesel trucks and cars retain value longer**
- **Mileage is important but significantly less than age**
- **No accidents increase value**
- **Unknown accidents lowers buyer confidence and value**

PREDICTED VS ACTUAL



- Linear regression
 - Scattered
 - Too high and low
- Decision Tree
 - Horizontal Clusters
 - Overfitting
- Gradient Boosting
 - Good patterns
 - Slightly scattered
- Random Forest
 - Few large errors
 - Best across most prices

RESIDUAL DISTRIBUTION



- Linear regression
 - Not centered on 0
 - Large + - residuals
- Decision Tree
 - Overfit
 - Many 20k+ errors
- Gradient Boosting
 - Good model
 - Underpredicts lux cars
- Random Forest
 - Tightly centered on 0
 - Errors are small

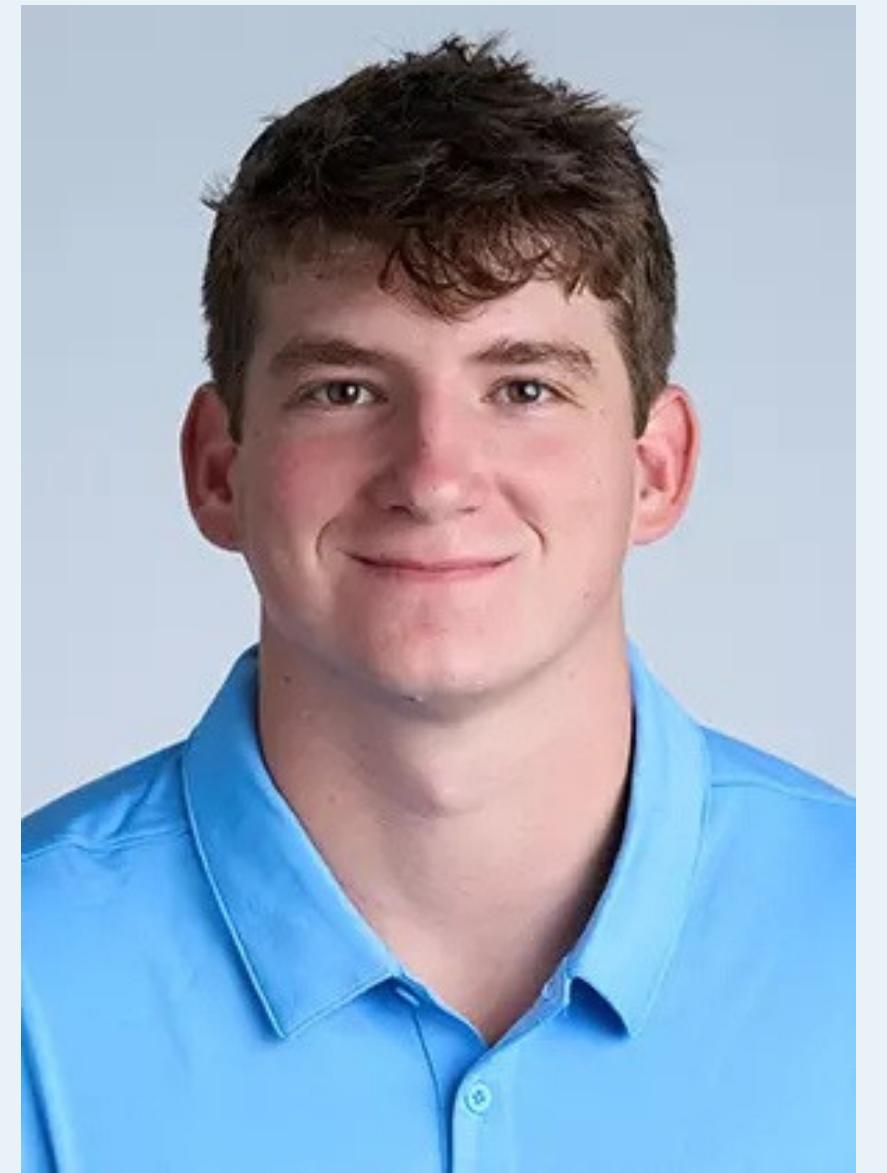
Summary

- Mileage, MPY ranks behind age in importance
- Tree-based models better reflect combined dataset training
- Combined dataset introduces broader variation in age, mileage, condition, and pricing behavior
- Gradient boosting is better calibrated than other models

Questions?



Slides 1-8
Coded cells 8-15



Slides 9-15
Coded cells 1-7