Report for Second-Hand Car Dealers

1. Business Understanding

The goal is to help second-hand car dealers across the United States maximize sales and pricing by identifying the key factors that influence used car prices. This analysis provides recommendations based on a large dataset of 79,195 car sales.

2. Data Understanding

The dataset consists of sales records from second-hand car dealers nationwide.

Initial inspection revealed numerous missing values (NaN), which were removed, leaving a statistically meaningful sample size. The scatter plot of data reveals that a very small number of sales < 0.04% of samples are skewing the data. Outliers (extremely high prices, specialty cars) were excluded by applying a price cap of $125,000, removing only 33 rows and retaining 79,162 samples.

Specialty cars (classic, luxury, highly modified) are underrepresented, limiting recommendations for these categories.

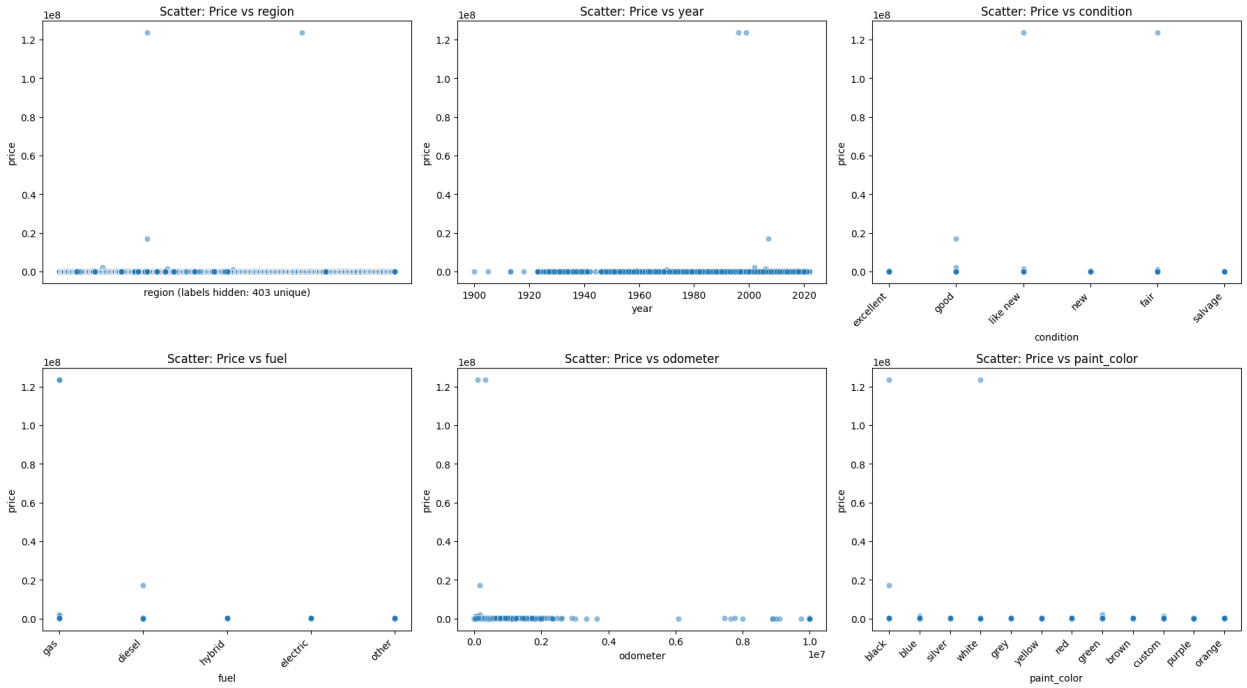

Figure 1: Scatter plot of the original data

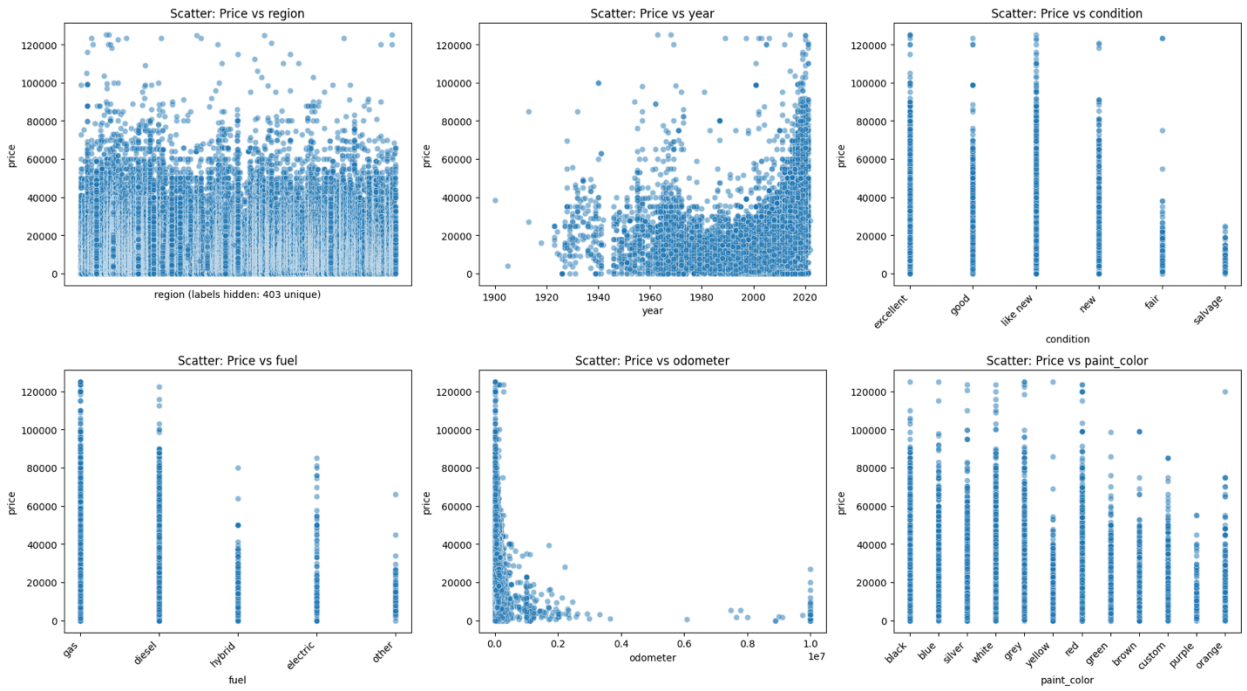

Figure 2: Scatter plot of data after removing data for specialty cars

3. Data Preparation

Data cleaning involved removing NaN entries and outliers. Encoding was performed using a blend of one-hot and probability-based methods. Many iterations (trial and error) were performed to find the best blend between one-hot encoding and probability-based encoding of string features. Probability encoding sped up processing. However, since ordinal information does not have a one-to-one correspondence with numbers, all information about the details of the feature was lost.

4. Modeling

Correlation Analysis: As an initial measure to understand the significance of various features, the correlation matrix of the data was calculated, and the cross-correlation with price was extracted and plotted. Even this simple technique reveals important information about the most popular cases: diesel cars fetch the best price, while blue, gas-fueled, front-wheel-drive cars can also bring high sale prices.

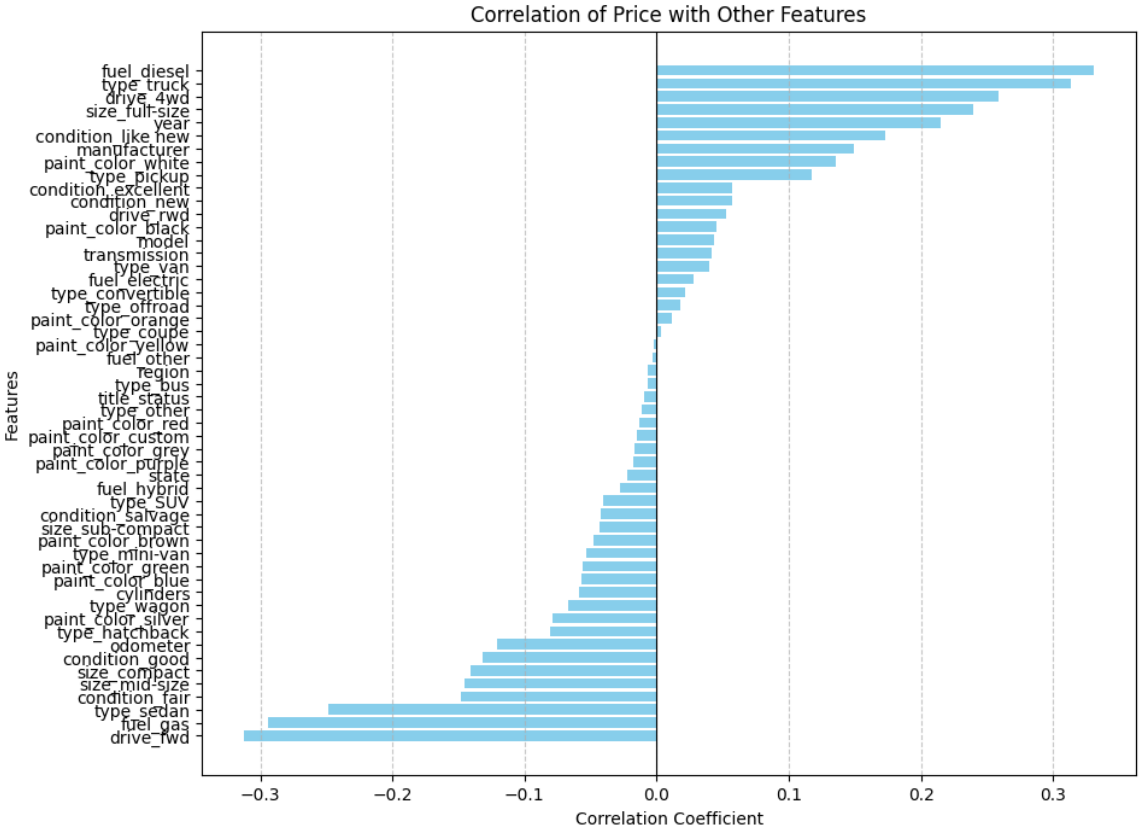

Figure 3: correlation of price with various features

Linear Regression: Given the high number of columns, running the cell even for a second-order polynomial was too slow and was abandoned. Only a first-order polynomial was used. As shown below, the mean squared error (MSE) steadily decreased as the number of features increased. The MSE converged to its lowest value after 17 features

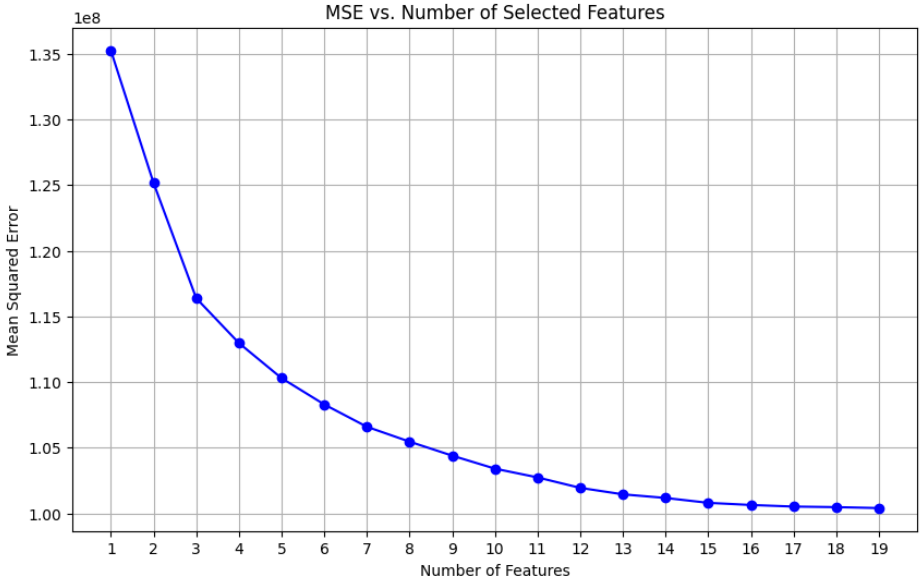

Figure 4: MSE for linear regression

Ridge Regression: For ridge regression, polynomial from order 1 to 2 were explored. The second order polynomial showed lower MSE as show in graph below.

A grid search to find the best hyperparameter (alpha) found the optimal value to be 20.4, which is consistent with what is shown in Figure 6.

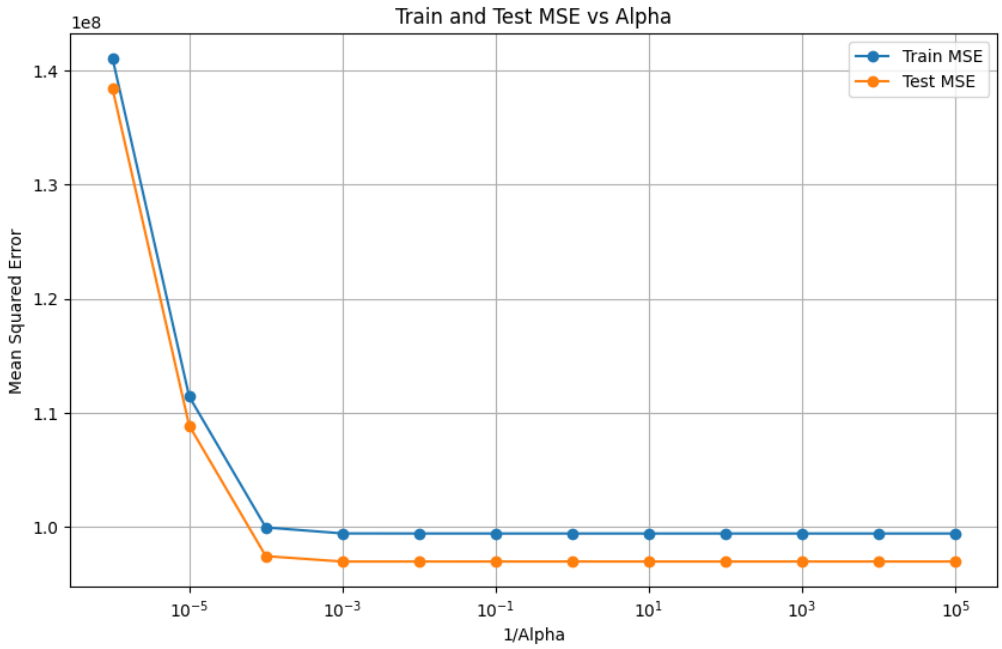

Figure 5: MSE for ridge regression of first order polynomial features

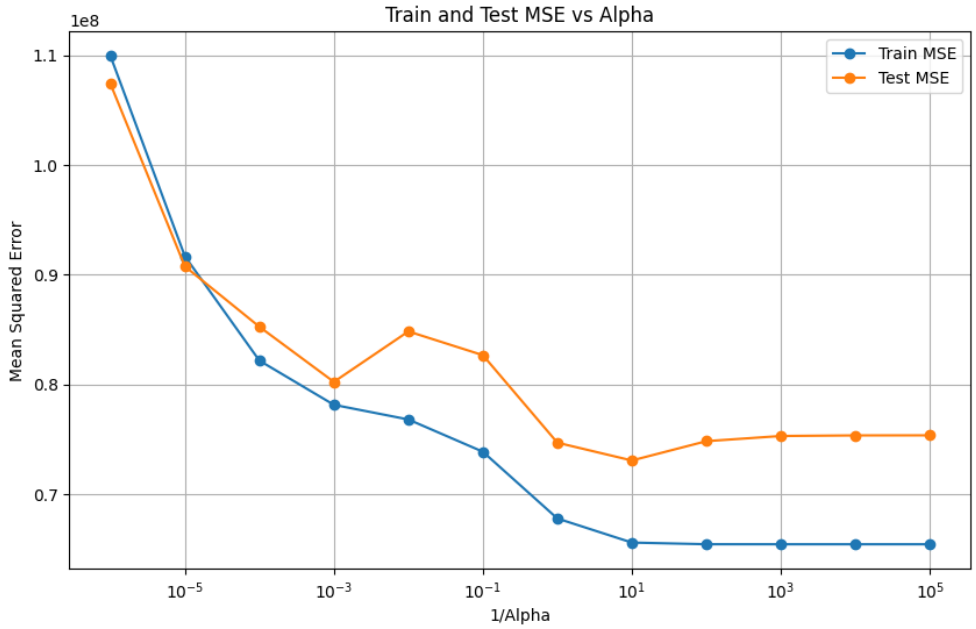

Figure 6: MSE for ridge regression of second order polynomial features

Coefficient of linear and ridge models are shown below:

Linear Regression 1st order		Ridge Regression 1st order		Ridge Regression 2nd order	
model	-144823.866084	**year**	2358.712762	**year^2**	395112.8942
fuel_electric	10101.267581	**type_truck**	1577.650993	**year**	-391471.0754
fuel_diesel	9760.841057	**condition_like new**	1550.329224	**year drive_4wd**	159145.2571
type_wagon	-7624.816603	**fuel_diesel**	1399.433486	**year fuel_gas**	-126459.6016
condition_like new	6738.101252	**drive_fwd**	-1343.446422	**year fuel_diesel**	120140.0413
type_sedan	-5731.785021	**type_sedan**	-1256.733562	**year drive_fwd**	-103666.733
type_hatchback	-5015.839991	**condition_fair**	-1212.692997	**year type_wagon**	-87247.45944
condition_salvage	-4886.893275	**fuel_gas**	-1161.969351	**year type_truck**	84607.4291
condition_fair	-4489.911427	**odometer**	-1136.468950	**year drive_rwd**	-74374.73829
drive_fwd	-4432.523297	**drive_4wd**	1071.071926	**year paint_color_grey**	52894.15179
type_SUV	-4400.915365	**condition_good**	-1013.347079	**year type_van**	51234.08166
type_mini-van	-4288.082825	**type_SUV**	-893.226747	**year cylinders**	50282.07074
condition_excellent	2693.324734	**type_convertible**	852.294376	**year condition_excellent**	-48196.27533
type_truck	2082.978919	**type_coupe**	825.024167	**year paint_color_black**	45496.69819
size_full-size	2080.687010	**size_full-size**	682.320034	**year condition_like new**	44815.63078
title_status	-1130.475373	**type_wagon**	-673.057437	**year odometer**	-44250.26584
paint_color_silver	-991.984836	**model**	-671.786348	**year type_hatchback**	-41756.05203
year	245.220895	**type_pickup**	592.908525	**year paint_color_blue**	-40509.54929
odometer	-0.005066	**condition_new**	582.567753	**year paint_color_red**	-39150.54252
		size_compact	-576.691804	**year type_sedan**	-38848.30961

5. Evaluation

Key recommendations from the regression and correlation analyses include:

- Newer cars command higher prices; model year is a strong predictor.
- 4WD vehicles are priced higher than FWD or RWD.
- Diesel vehicles generally sell for more.
- Trucks and vans tend to have higher sale prices, while wagons, hatchbacks, and sedans tend to have lower prices.
- Grey and black cars are the most popular and fetch higher prices, while blue and red cars bring in lower prices.
- Condition strongly influences price: ‘like new’ cars sell for more, but ‘excellent’ condition may reduce the price.
- The number of cylinders generally has a negative impact on price, except for newer cars, where it can be positive.
- Cars with high mileage (large odometer readings) should be avoided, as their sales prices are strongly and negatively impacted by mileage.

6. Deployment & Recommendations

- Dealers should focus on acquiring newer, low-mileage, diesel or hybrid 4WD trucks, vans, and cars in grey or black with 'like new' condition to achieve higher prices.
- Specialty cars require more data for concrete recommendations. Dealers should collect additional information on these categories

Summary:
This CRISP-DM report provides a structured overview of the key factors affecting used car prices, offering practical guidance for second-hand car dealers. By focusing on the highlighted vehicle types, features, and conditions, dealers can optimize their inventory and pricing strategies for maximum profitability.