Used Cars For Final In Machine Learning
Exploratory Analysis

Drew Buhr, dbuhr@bellarmine.edu
Joshua Dow, jdow@bellarmine.edu

I. INTRODUCTION

On Kaggle, we were able to find a used car dataset that described over 20,000 used cars for sale. The dataset included information from a car's manufacturer, price, year, and transmission. We decided to choose this dataset because it offered a fair mix of categorical and numerical data, and we thought it would be an interesting dataset to work with.

II. DATA SET DESCRIPTION

The data set originally contained over 10,000 samples with 25 columns, but after cleaning unnecessary columns and samples with incomplete entries we narrowed the data set down to slightly over 4000 samples with 18 columns. A complete listing is shown in **Table 1**.

Table 1: Data Types and Missing Data

Variable Name	*Data Type*	*Missing Data (%)*
1) Region	Nominal / object	0%
2) Price	Ratio / int64	0%
3) Year	Interval / int64	0%
4) Manufacturer	Nominal / object	0%
5) Model	Nominal / object	0%
6) Condition	Ordinal / object	0%
7) Cylinders	Nominal / object	0%
8) Fuel	Nominal / object	0%
9) Odometer	Ratio / int64	0%
10) Title	Nominal / object	0%
11) Transmission	Nominal / object	0%
12) Drive	Nominal / object	0%
13) Size	Nominal / object	0%
14) Type	Nominal / object	0%
15) Color	Nominal / object	0%
16) Description	Nominal / object	0%
17) State	Nominal / object	0%

III. Data Set Summary Statistics

Narrative introduction to the section.

Table 2: Summary Statistics for Used_Cars

Variable Name	*Count*	*Mean*	*Standard Deviation*	*Min*	*25th*	*50th*	*75th*	*Max*
Price	4125	10,945	9,227	265	4,950	7,990	14,989	135,000
Year	4125	2008	7	1917	2005	2009	2013	2020
Odometer	4125	115,418	59,858	22	74,097	116,783	152,000	417,000

There should be a table for **EACH** categorical variable.

Table 3: Proportions for XXX (n=yyy)

Category	*Frequency*	*Proportion (%)*
Region	4125	*100%*
Manufacturer	4125	100%
Model	4125	100%

Condition	4125	100%
Cylinders	4125	100%
Fuel	4125	100%
Title	4125	100%
Transmission	4125	100%
Drive	4125	100%
Size	4125	100%
Type	4125	100%
Color	4125	100%
Description	4125	100%
State	4125	100%

After you summarize the categorical variables, generate a correlation matrix for all continuous variables (not categorical – this doesn't make sense)

Table 4: Correlation Table/Tables

	Price	Year	Odometer
Price	1.00000	.409356	-.502527
Year	.409356	1.00000	-.340117
Odometer	-.502527	-.340117	1.00000

We found a heatmap to be nonhelpful with this particular dataset. A heatmap was very difficult to read and did not show very distinct information about the dataset.

IV. DATA SET GRAPHICAL EXPLORATION

Narrative introduction to the section. In each section below, indicate any interesting distributions, anomalies, imbalance, etc. that you notice.

Figure 1: Distribution of the Cars' Manufactured Year

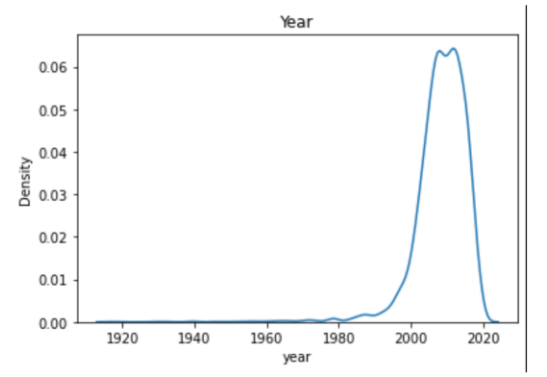


Figure 2: Scatterplot of Price Compared to miles on the Odometer

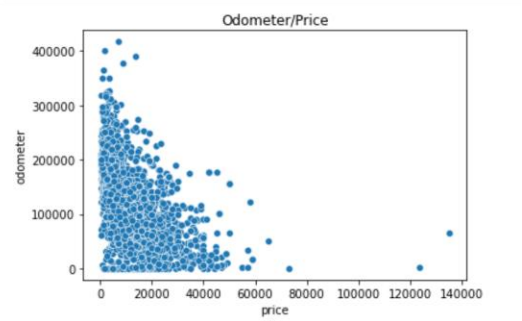


Figure 3: Scatterplot of Price Compared to the Cars' Manufacturer

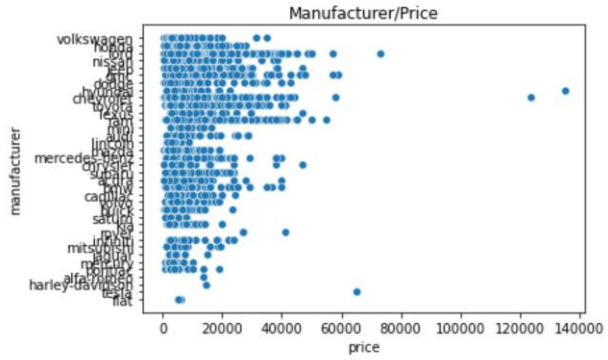


Figure 4: Bar Graph Showing Max Price of Each Condition Type

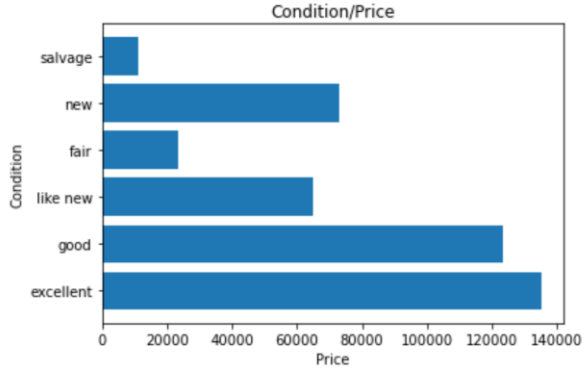


Figure 5: Bar Graph Showing Max Price For Each Transmission Type

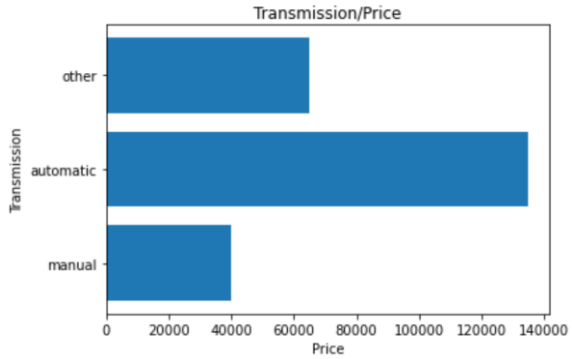


Figure 6: Distribution of Prices

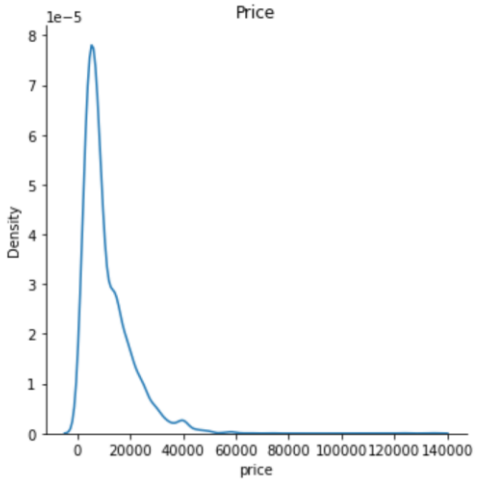
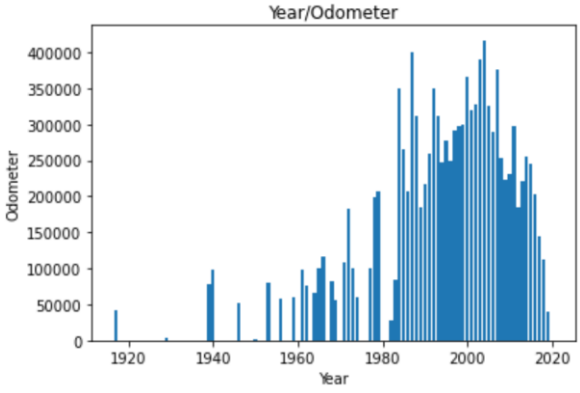


Figure 7: Bar Graph Showing Comparison Between a Car's Manufactured Year and Odometer Reading



V. SUMMARY OF FINDINGS

In conclusion, we found that the price of a used car is affected by multiple aspects of the car. The year a car was manufactured, the number of miles driven on the car, and the transmission of the car had the effects of the cars' prices that we expected. For the most part, the more recently the car was produced the higher the resale price would be. However, in some rare instances if a vintage or old collection car were being resold, the price could still be higher than the mean and the median. We also found the smaller number of miles on the cars' odometer readings the seller was able to sell the car for a higher price and vice versa. Automatic transmitted cars were also sold at a higher value than manually transmitted cars. The description of the cars' conditions impacted the price as we expected, the better the condition resulted in a higher price. We surprisingly found that the manufacturer of the car did not have a noticeable impact on the resale price of the car. Overall, we found that the resale price of a car is impacted mostly by when it was produced, the number of miles on the car, and the type of transmission.