Evan Kerivan

Mini Project One

Contents

Problem Statement

US Used Car Dealership Asset Appraisal

- US based used car dealership currently has a large stock of vehicles
- The dealership needs a new way to determine the value of their vehicles for stocktake and purchasing
- Current method involves driving to other car lots and trying to find similar cars with prices
- The dealership has +50 lots nationwide

Dataset:

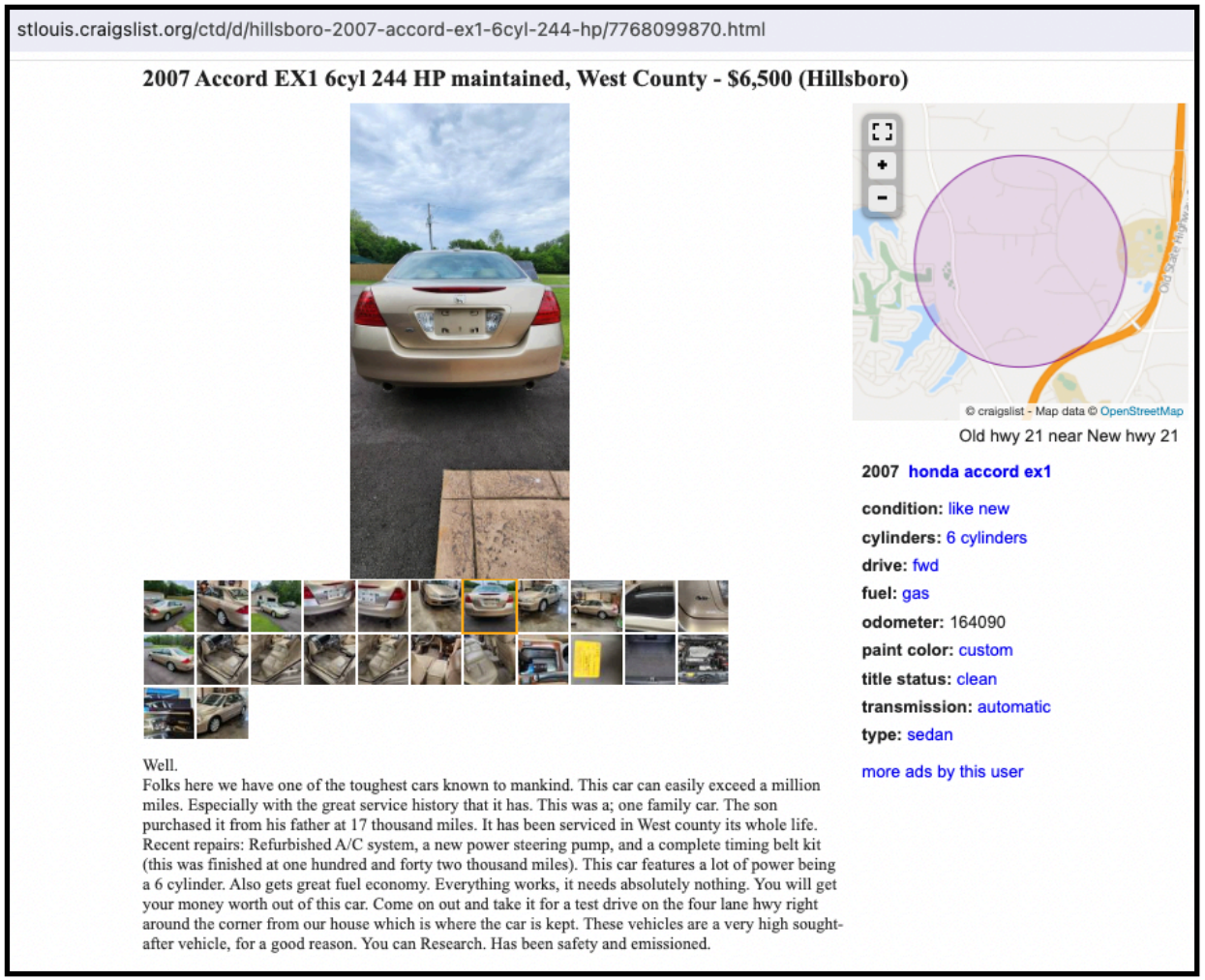
“Used Car Dataset”

Kaggel user built a web scrapper that scrapes cragslist.org's used car section nation wide. New data is published every few months.

25 features and 426,880 records

	id	426880	non-null	int64
1	url	426880	non-null	object
2	region	426880	non-null	object
3	region_url	426880	non-null	object
4	price	426880	non-null	int64
5	year	425675	non-null	float64
6	manufacturer	409234	non-null	object
7	model	421603	non-null	object
8	condition	252776	non-null	object
9	cylinders	249202	non-null	object
10	fuel	423867	non-null	object
11	odometer	422480	non-null	float64
12	title_status	418638	non-null	object
13	transmission	424324	non-null	object
14	VIN	265838	non-null	object
15	drive	296313	non-null	object
16	size	120519	non-null	object
17	type	334022	non-null	object
18	paint_color	296677	non-null	object
19	image_url	426812	non-null	object
20	description	426810	non-null	object
22	state	426880	non-null	object

Example Craigslist Listing

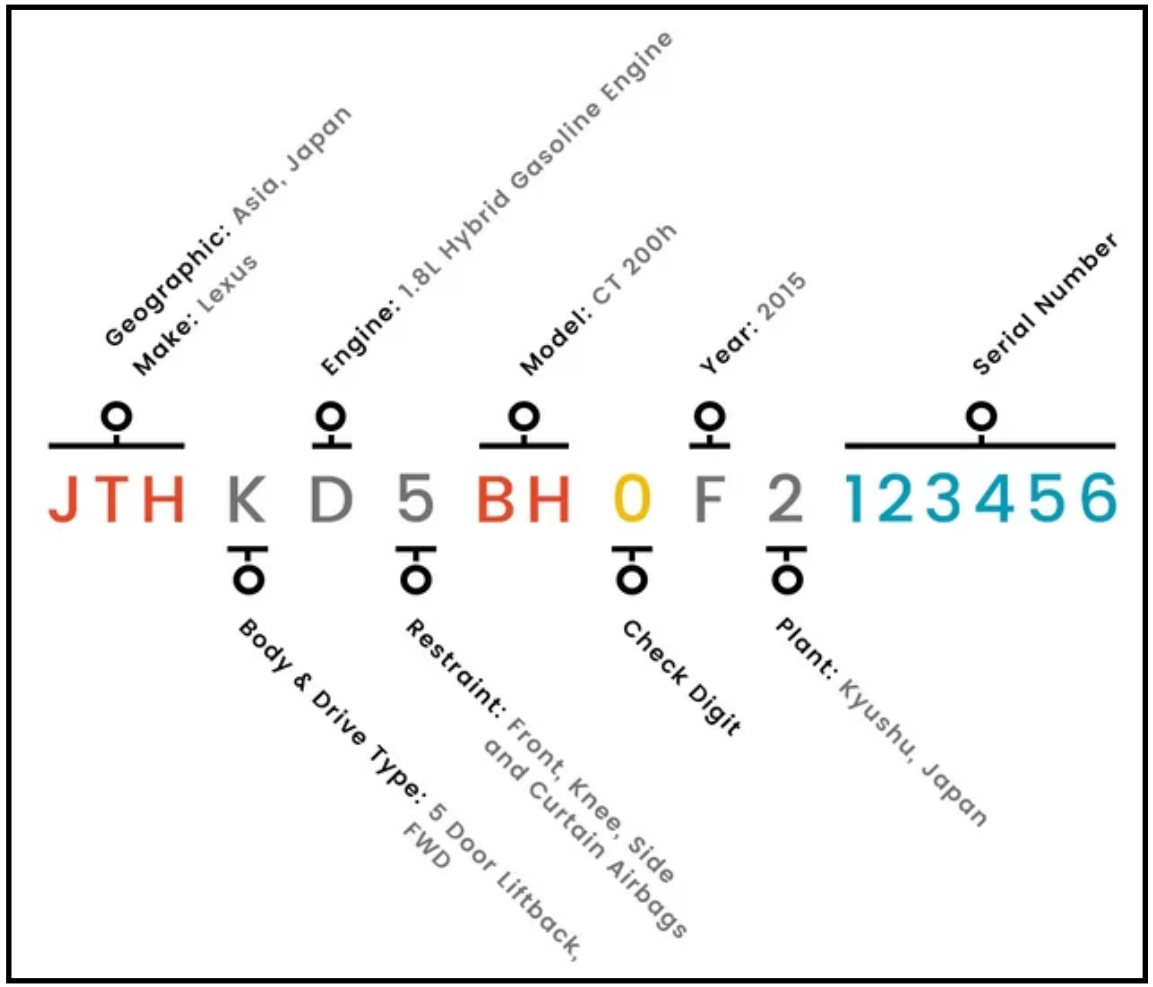


Data Cleansing:

What is a VIN (Vehicle Identification Number?

- 17 alphanumeric identifier required for all manufactured vehicles
- Encoded with information about the vehicle
- Standardised after 1981
- I,O,Q not allowed to prevent confusion with similar letters and numbers
- 9th digit check digit to prevent fraud
    - (sum 16 digits and divide by 11)

VIN Decoder Diagram

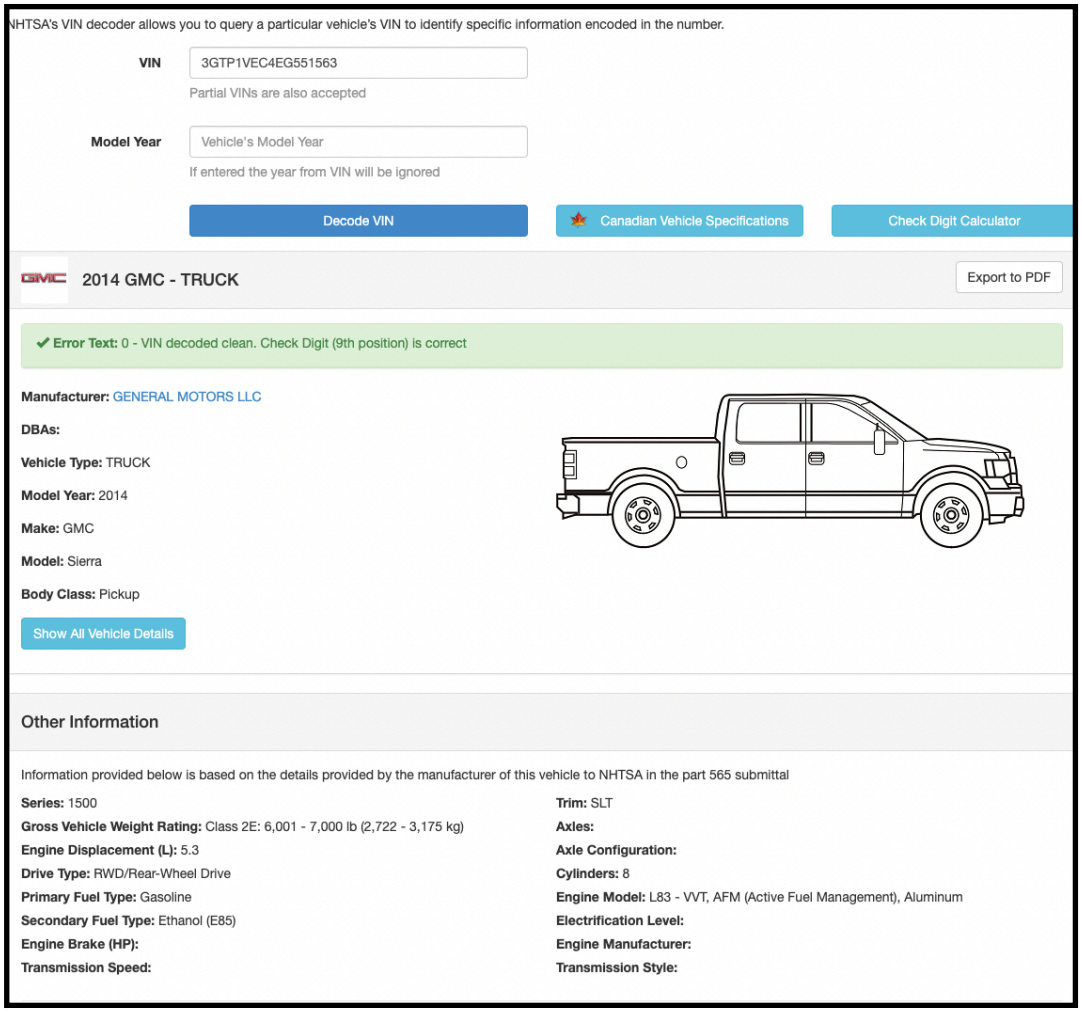


Data Cleansing:

Filling Missing Values with VIN Decoder

NHTSA Website VIN decoder

- NHTSA has a publicly available VIN decoder API
- Able to return missing values for all of the vehicle description features except color, title status and condition
- Built a multistep function to call API, deal with connection errors, store results, decode the results and fill the missing values
- Ran the function on segments of the data set to ensure function was working properly and could handle various errors
- Overwhelmed the API several times and needed to use backoff delays
- Used ThreadPool to decode multiple VIN simultaneously
- Used tqmd to display progress bar as the process took over 12 hours, needed to be sure it was running
- Extremely accurate but time consuming
    - It's possible to build a decoder locally and avoid the API



Data Cleansing:

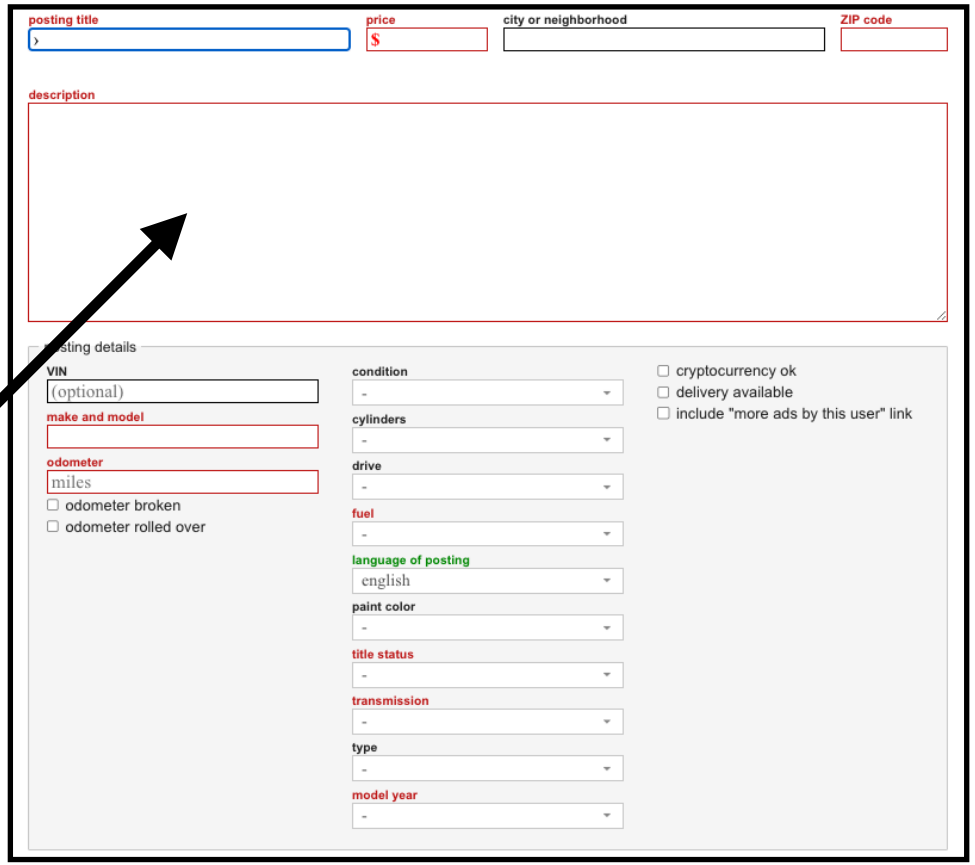
Filling Missing Description Reader

- Craigslist form has mandatory (red) and optional fields.
- All of the categorical fields (except make and model) are predefined drop down lists
- Often information for the categorical fields was written as part of the description.
- A list of the unique values(keywords) for a selection of features and the feature name were used to make a set of dictionaries i.e { paint colour: “red”, “blue”,”green”,….} A function was created to scan all of the “descriptions” in the data set and return the matching key value pairs i.e {paint colour:’red’}
- This was highly effective but more prone to errors

Actual description submitted by user Keyword

2011 **Toyota Prius Hybrid**, 153K Miles, Bluetooth, JBL- 6-CD, AC, Cruise2011 Toyota Prius Hybrid, 153K Miles, **Automatic** CVT Transmission, **Red** with Grey Cloth Interior, Climate Control with Ice Cold Air Conditioning, Bluetooth, Power Windows, Power Door Locks with Keyless Entry Remote, Power Mirrors, JBL Stereo System with 6 Disk CD Player, Aux Input, Steering Wheel Mounted Audio & Temp Controls, Cruise Control, On Board Computer, Tilt & Telescopic Steering Wheel. This 11 Prius **Sedan** is Capable of 50+ MPG and is Being Sold with a 20 Day Plate, NH Safety Inspection Sticker and Dealer Warranty!Call Rafferty Auto Sales Anytime at 603.263.0870Check out more inventory at http://www.raffertyauto.comRafferty Auto Sales LLC29 Laconia Road, Route 106Belmont, NH 03220Similar To Honda Insight, Nissan Leaf, Chevy Volt 386881

Craigslist posting form



Data Cleansing:

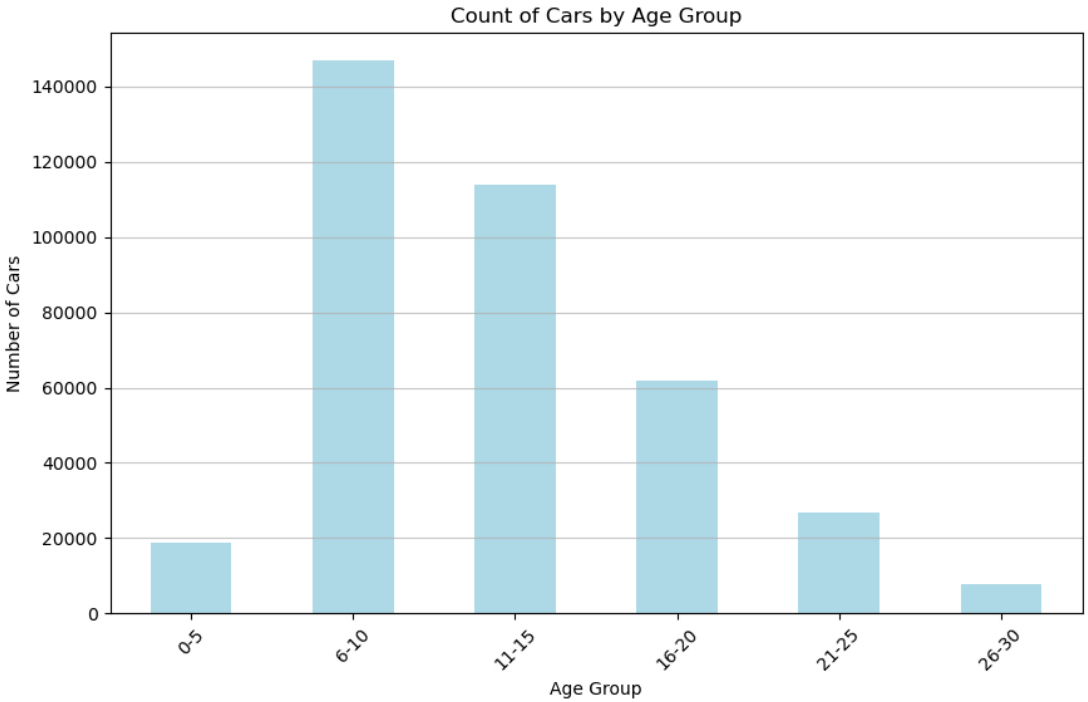
Wrap up

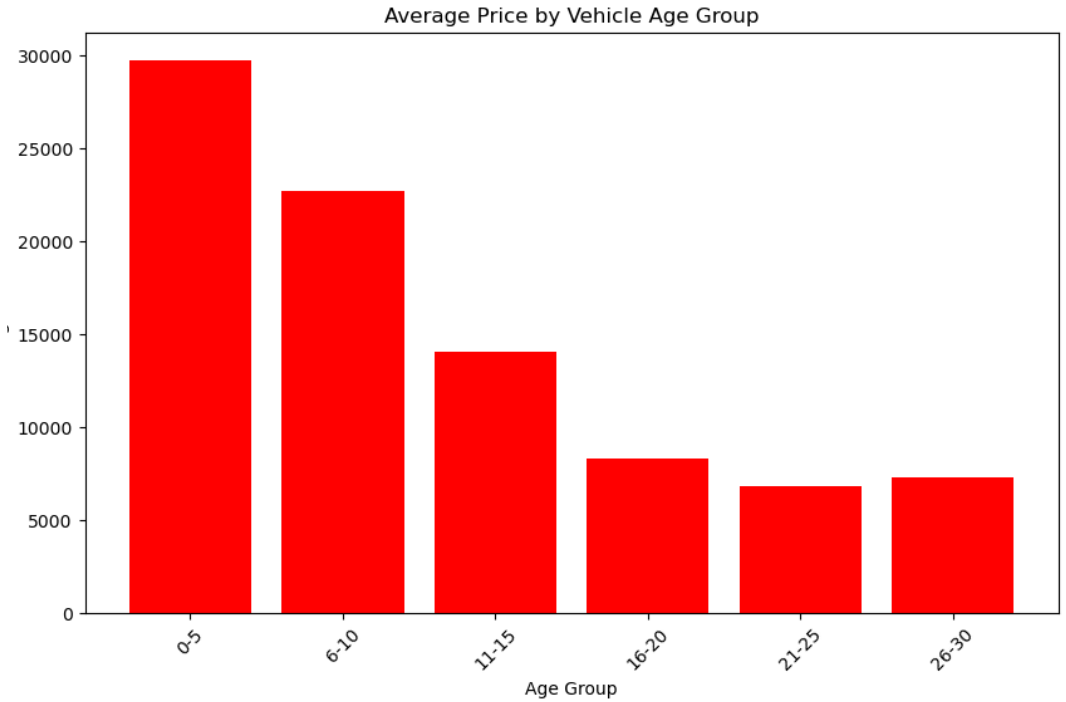
1. Dropped outliers
2. VIN Decoder API
3. Description dictionary
4. Grouped by “odometer” and “condition” to fill “condition”
5. Dropped unnecessary features
6. Dropped NAs
7. Used Label Encoder to transform categorical features into numerical features

Column name	Pre data cleaning % Missing	Post steps 1-2 % Missing	Improvement
country	100.000000	100.000000	0.000000
size	71.767476	66.711722	5.055754
VIN	41.622470	37.725356	3.897114
condition	40.785232	16.516117	24.269115
paint_color	37.725356	7.067326	30.65803
cylinders	30.586347	3.619518	26.966829
title_status	30.501078	1.930753	28.570325
drive	1.930753	1.032374	0.898379
odometer	1.534155	1.030735	0.50342
transmission	1.534155	0.233087	1.301068
type	1.236179	0.175928	1.060251
description	1.030735	0.016398	1.014337
manufacturer	0.282281	0.013118	0.269163
model	0.016398	0.010307	0.006091
fuel	0.015930	0.002343	0.013587
year	0.015930	0.002108	0.013822
price	0.000000	0.000000	0.000000
state	0.000000	0.000000	0.000000
region_url	0.000000	0.000000	0.000000
id	0.000000	0.000000	0.000000

EDA:

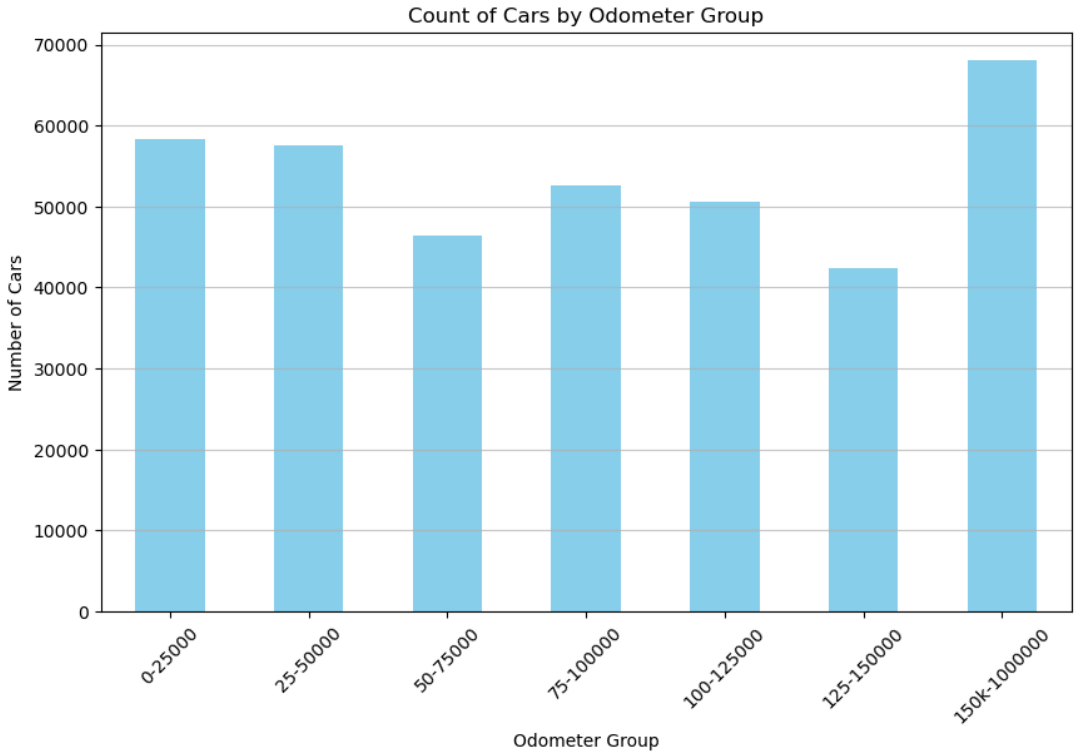
Vehicle Age

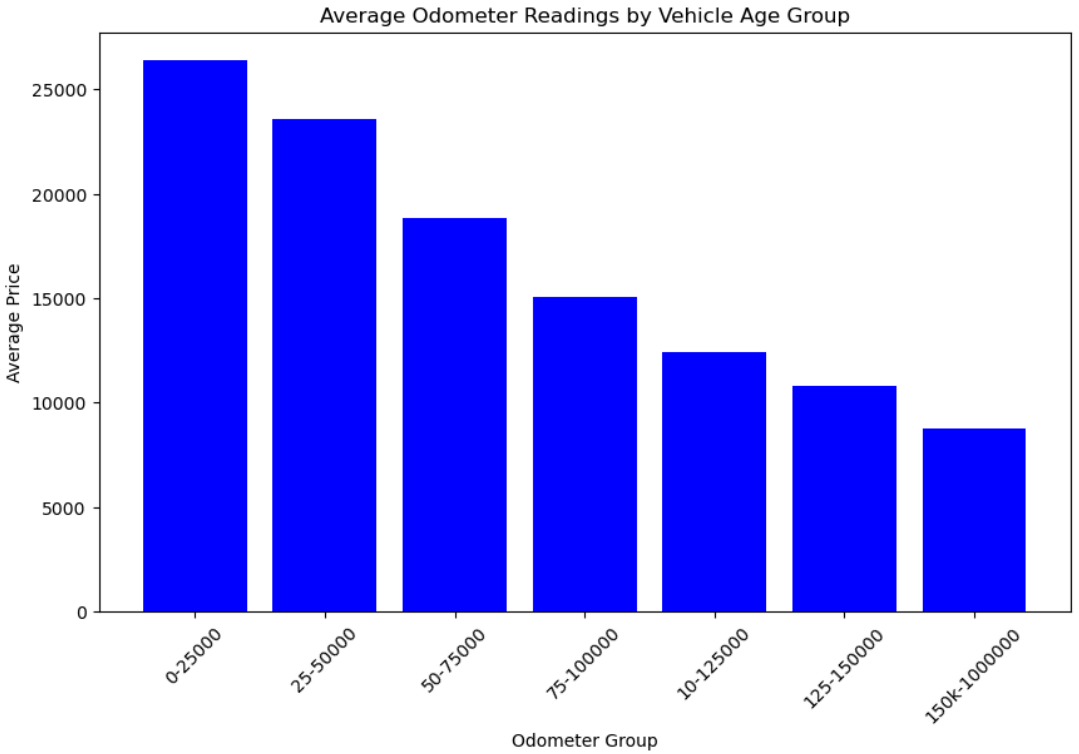




EDA:

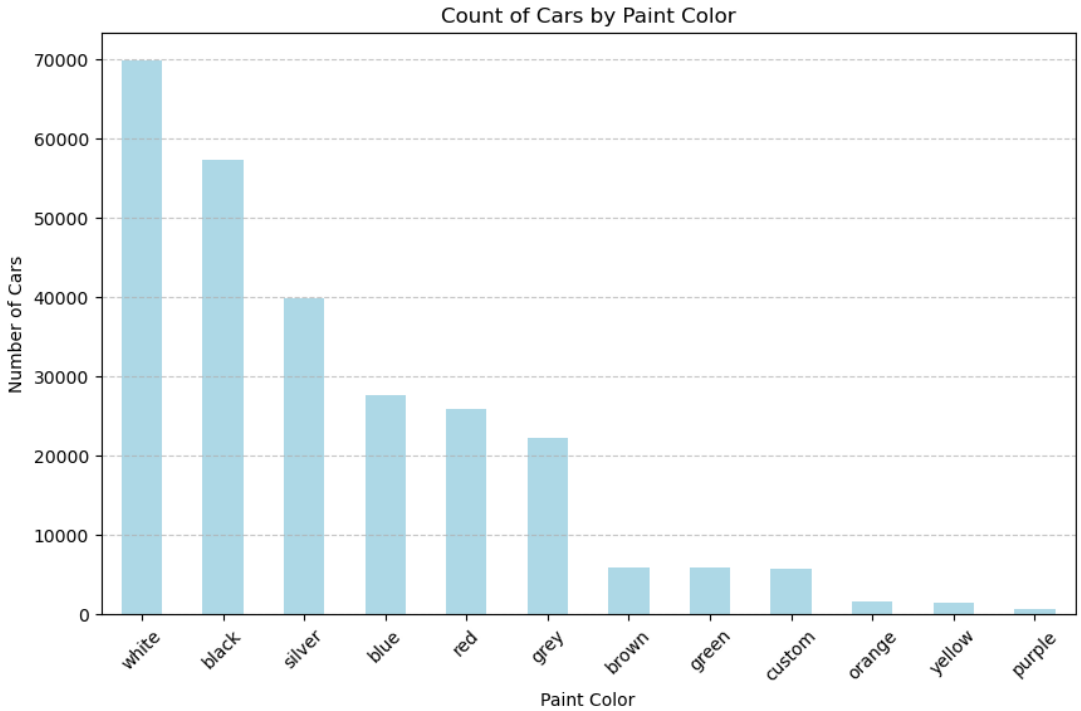
Vehicle Odometer Reading

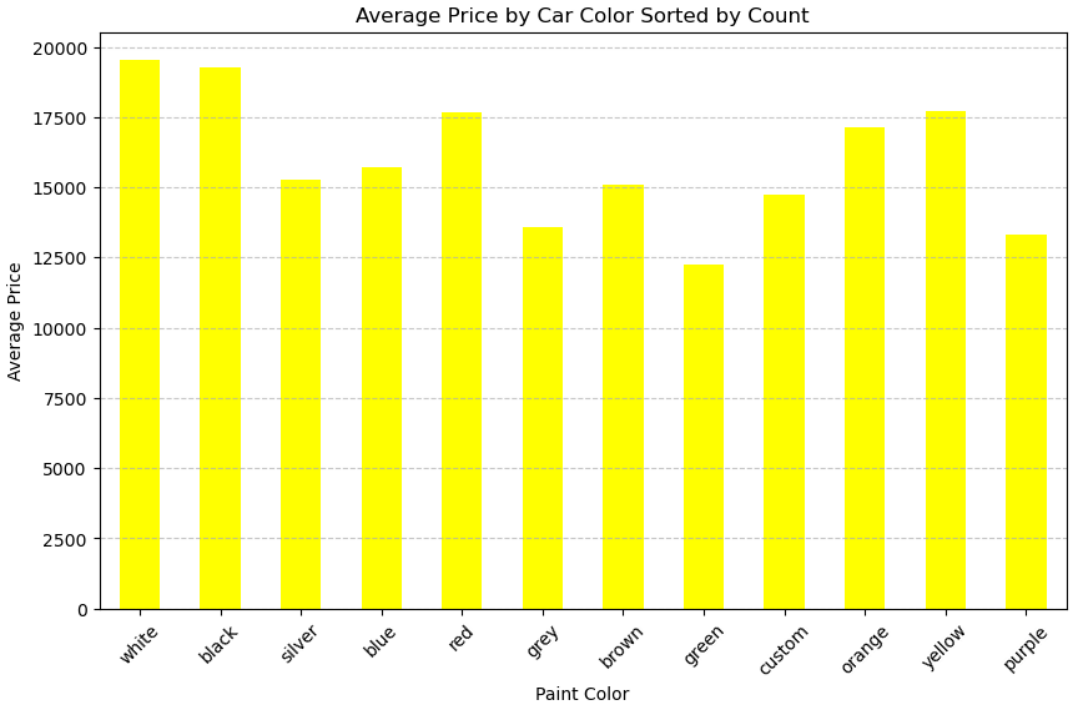




EDA:

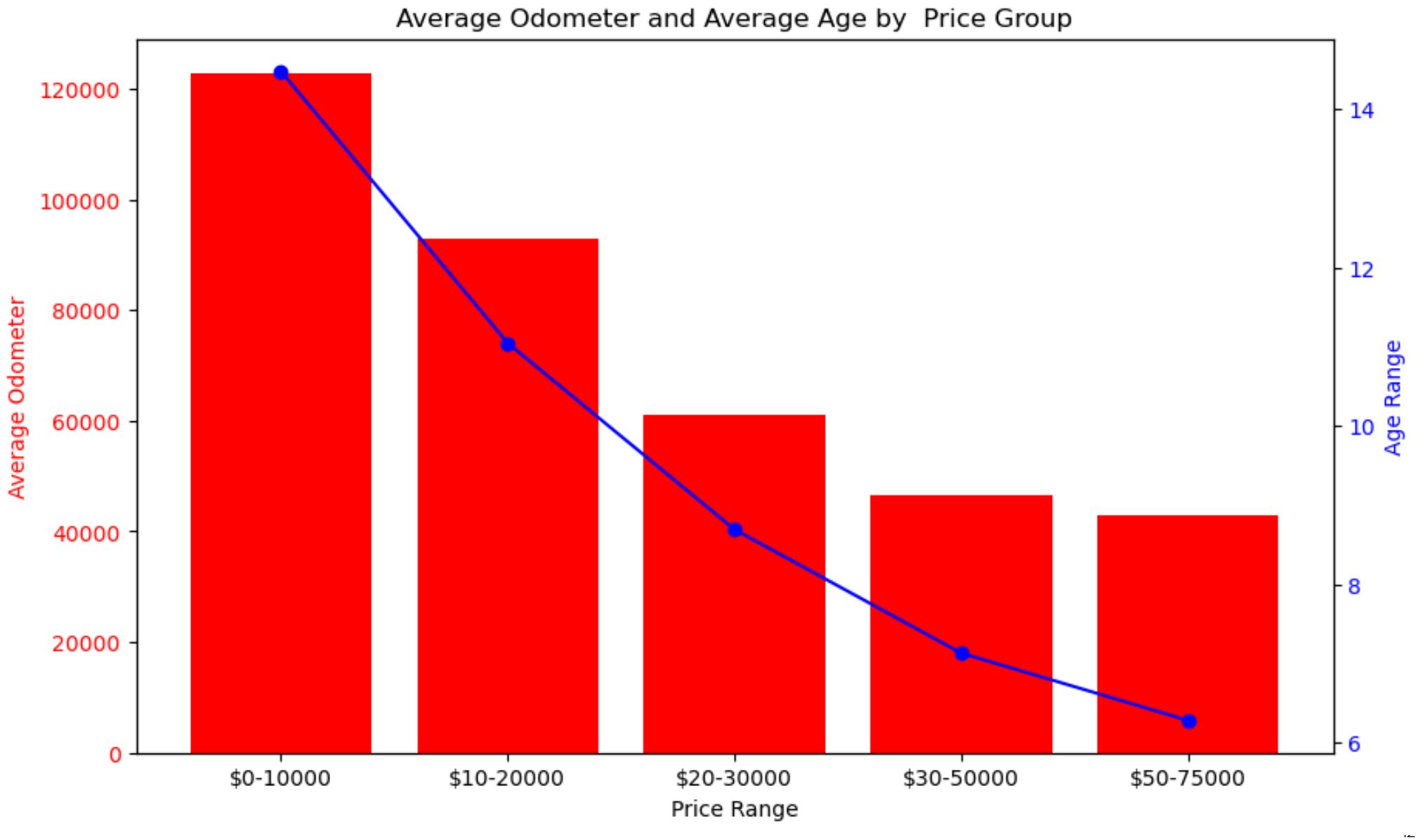
Vehicle Age





EDA:

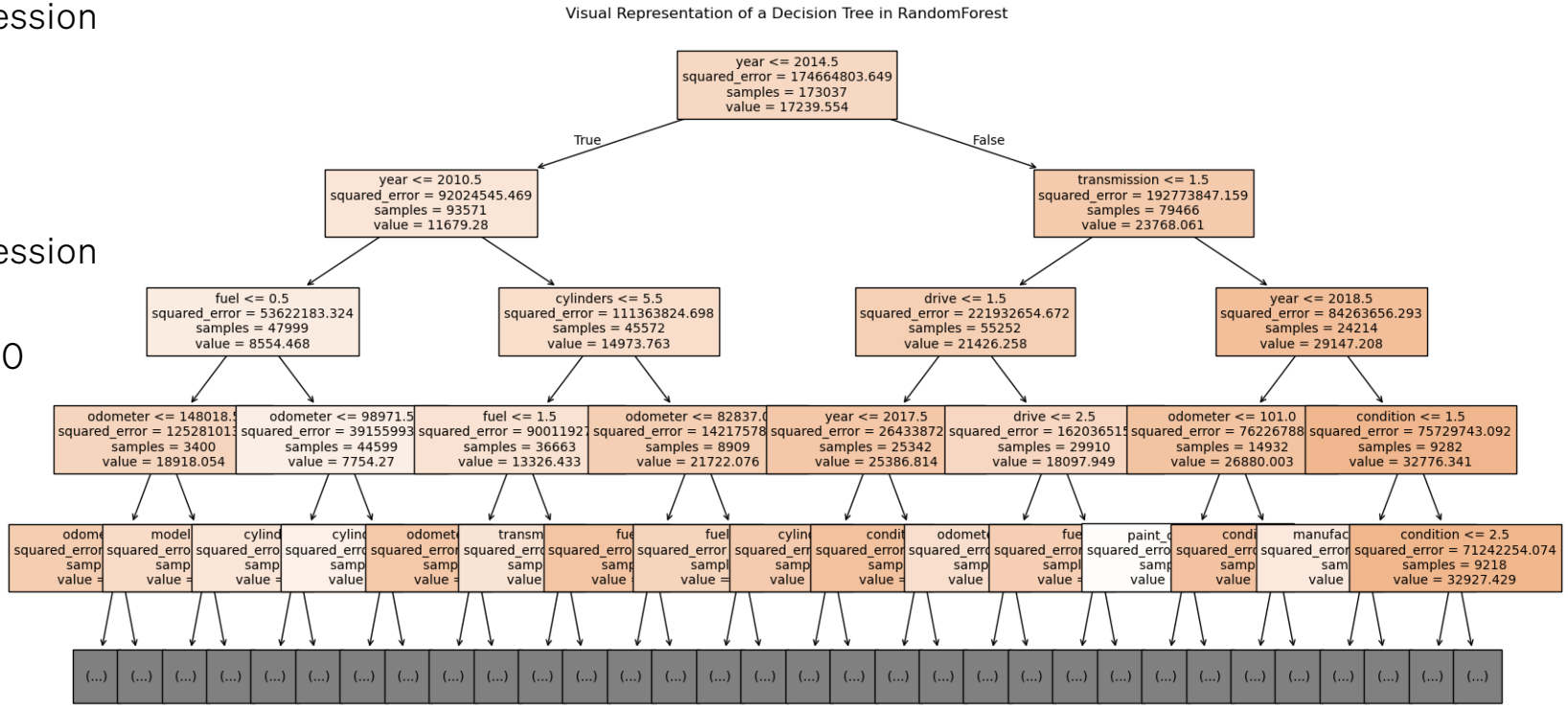
Odometer x Age x Price



Models:

Regression

- Base linear Linear Regression
  - r2= 32%
- Linear Regression with Feature selector and Standard Scalar
  - R2 = 33%
- Random Forest Regression
  - Estimators 20
  - Max Depth 5
  - r2= 81%
- Random Forest Regression depth
  - Estimators 200
  - Max depth 20
  - r2 train=92%
  - r2 test=80%



Conclusion

US Used Car Dealership Asset Appraisal

- The model performs well and can predict the price of a used car
- Far better outcome than employees driving to other car lots to check prices. Time savings and accuracy
- There is likely some overfitting but this could be addressed with further hyper tuning
- From the model results it is likely that the relationship is not linear and better modelled using a non-linear modelling technique i.e random forest
- The model will not perform well with outliers i.e supercars or custom cars
- The model will need to be updated with new data from craigslist on a regular basis
- Computationally heavy and time consuming to perform updates on a local computer. National head office should house on their server and allow locations to call the model.