

# Evan Kerivan

## Mini Project One

## Contents

1. Problem Statement
2. Data Cleaning
  - 2.1. VIN Decoder
  - 2.2. Description Reader
3. EDA
4. Models
  - 4.1. Linear Regression
  - 4.2. Random Forest Regression
5. Conclusion

## 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
- The dealership has +50 lots nationwide

## Dataset:

## “Used Car Dataset”

Kaggle user built a web scrapper that scrapes [craigslist.org](https://www.craigslist.org)'s used car section nationwide. New data is published every few months.




25 features and 426,880 records

## Example Craigslist Listing

	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

stlouis.craigslist.org/ctd/6/hillsboro-2007-accord-ax1-6cyl-244-hp/7768098870.html

**2007 Accord EX1 6cyl 244 HP maintained, West County - \$6,500 (Hillsboro)**

Old Hwy 21 near New Hwy 21

**2007 honda accord ex1**

condition: like new  
 cylinders: 6 cylinders  
 drive: fwd  
 fuel: gas  
 odometer: 164090  
 paint color: custom  
 title status: clean  
 transmission: automatic  
 type: sedan

more ads by this user

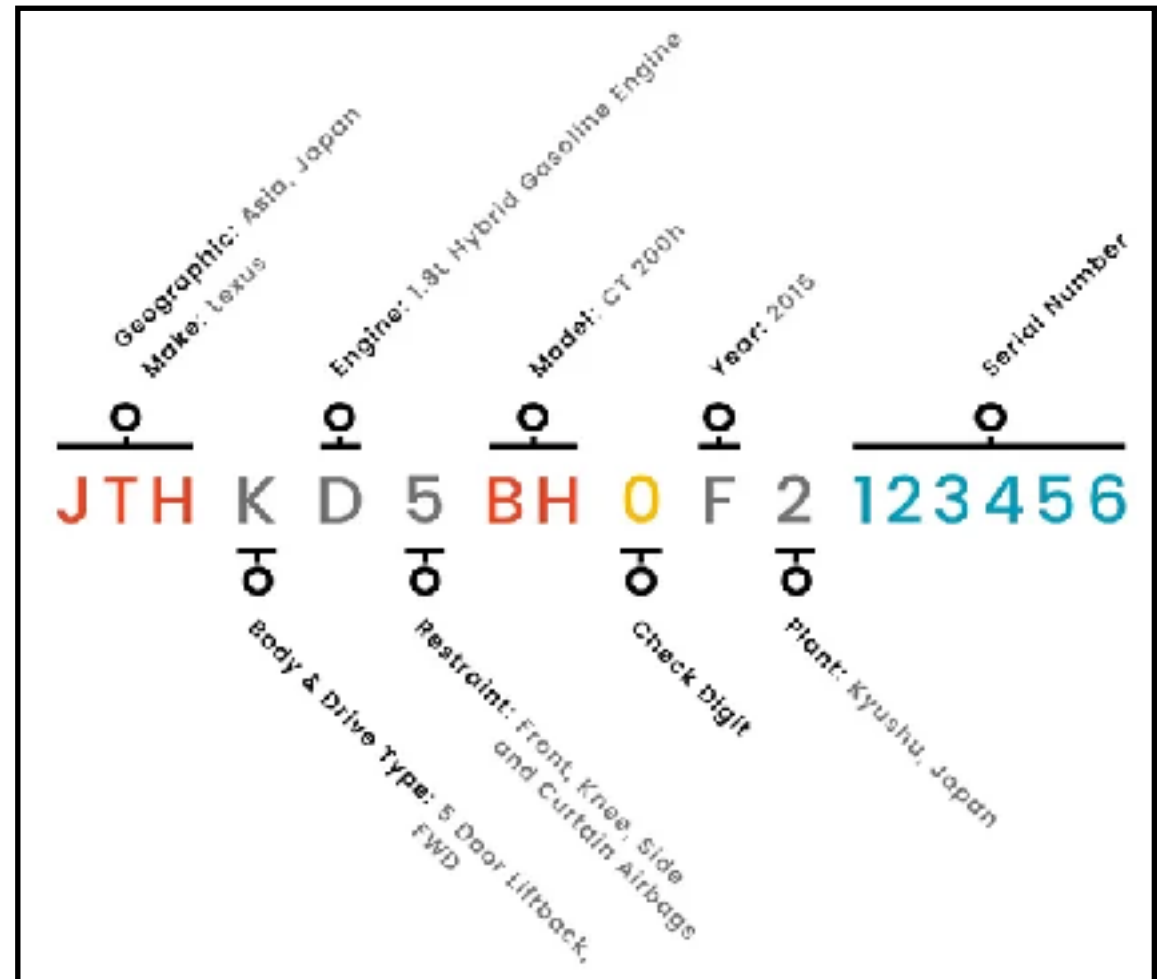
**Will**  
 Folks here we have one of the toughest cars known to mankind. This car can easily exceed a million miles. Especially with the great service history that it has. This was an one family car. The car purchased it from his father at 17 thousand miles. It has been serviced in West county its whole life. Recent repairs: Refurbished A/C system, a new power steering pump, and a complete timing belt kit. (this was finished at one hundred and forty two thousand miles). This car features a lot of power being a 6 cylinder. Also gets great fuel economy. Everything works, it needs absolutely nothing. You will get your money worth out of this car. Come on out, and take it for a test drive on the four lane hwy right around the corner from our house which is where the car is kept. These vehicles are a very high craigslist vehicle, for a good reason. You can Research, I has been safety and emissions.

## 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
  - (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 tqdm 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

NHTSA's VIN decoder allows you to query a particular vehicle's VIN to identify specific information encoded in the number.

VIN:

Partial VINs can also be queried

Model Year:

It entered the year that VIN will be ignored

**GMC 2014 GMC - TRUCK**

✓ Error Text: VIN decoded clean. Check Digit (9th position) is correct

Manufacturer: **GENERAL MOTORS LLC**

Color:

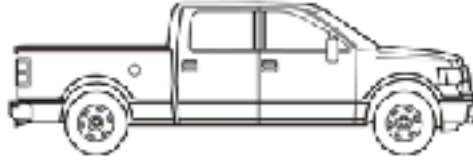
Vehicle Type: **TRUCK**

Model Year: **2014**

Make: **GMC**

Model: **Sierra**

Body Class: **Pickup**



**Other Information**

Information provided below is based on the details provided by the manufacturer of this vehicle to NHTSA in the past 555 submitted

Series: <b>1500</b>	Trim: <b>SLT</b>
Name: <b>Sierra</b>	Area:
Weight Rating: <b>Class 2F (4,001 - 7,000 lb (2,222 - 3,175 kg))</b>	Axle Configuration:
Engine Displacement (L): <b>5.0</b>	Cylinders: <b>3</b>
Drive Type: <b>Front-Wheel Drive</b>	Engine Model: <b>LS3 - VVT, AFM (Active Fuel Management), 4800ccm</b>
Primary Fuel Type: <b>Gasoline</b>	Evolution Level:
Secondary Fuel Type: <b>ETanol (E85)</b>	Engine Manufacturer:
Engine Cycle (RPM):	Transmission Style:
Transmission Speed:	

## Data Cleansing:

### Filling Missing Values with VIN Decoder

- 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.com>Rafferty Auto Sales LLC29 Laconia Road, Route 106Belmont, NH 03220Similar To Honda Insight, Nissan Leaf, Chevy Volt 386881

Craigslist posting form

The screenshot shows the Craigslist posting form. At the top, there are input fields for 'posting title', 'price' (with a '\$' symbol), 'city or neighborhood', and 'ZIP code'. Below these is a large 'description' text area, which is highlighted with a red border. A black arrow points from the word 'Keyword' in the text block to this description area. Under the description, there are several 'posting details' including 'VIN', 'make and model', 'odometer' (with a 'miles' dropdown), and checkboxes for 'odometer broken' and 'odometer rolled over'. To the right of these are dropdowns for 'condition', 'cylinders', 'drive', 'fuel', 'language of posting', 'paint color', 'title status', 'transmission', 'type', and 'model year'. On the far right, there are three checkboxes: 'cryptocurrency ok', 'delivery available', and 'include "more ads by this user" link'.

# 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

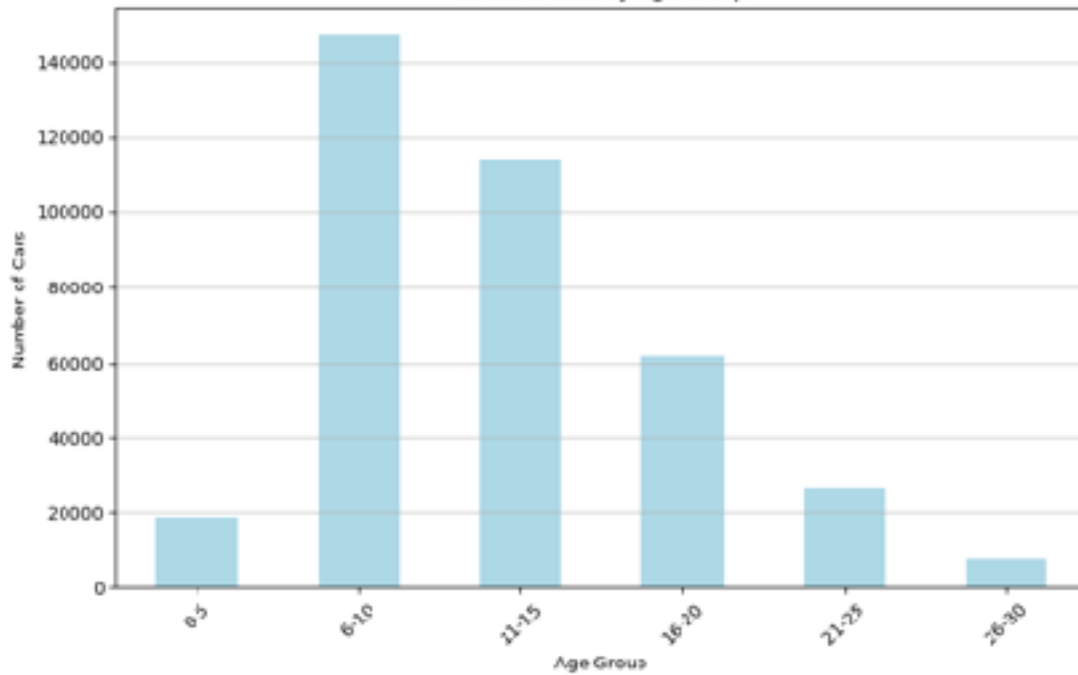
Column name	Pre data cleaning % Missing	Post steps 1-2 % Missing	Improvement
<b>country</b>	100.000000	100.000000	0.000000
<b>size</b>	71.767476	66.711722	5.055754
<b>VIN</b>	41.622470	37.725356	3.897114
<b>condition</b>	40.785232	16.516117	24.269115
<b>paint_color</b>	37.725356	7.067326	30.65803
<b>cylinders</b>	30.586347	3.619518	26.966829
<b>title_status</b>	30.501078	1.930753	28.570325
<b>drive</b>	1.930753	1.032374	0.898379
<b>odometer</b>	1.534155	1.030735	0.50342
<b>transmission</b>	1.534155	0.233087	1.301068
<b>type</b>	1.236179	0.175928	1.060251
<b>description</b>	1.030735	0.016398	1.014337
<b>manufacturer</b>	0.282281	0.013118	0.269163
<b>model</b>	0.016398	0.010307	0.006091
<b>fuel</b>	0.015930	0.002343	0.013587
<b>year</b>	0.015930	0.002108	0.013822
<b>price</b>	0.000000	0.000000	0.000000
<b>state</b>	0.000000	0.000000	0.000000
<b>region_url</b>	0.000000	0.000000	0.000000
<b>id</b>	0.000000	0.000000	0.000000 <sup>8</sup>



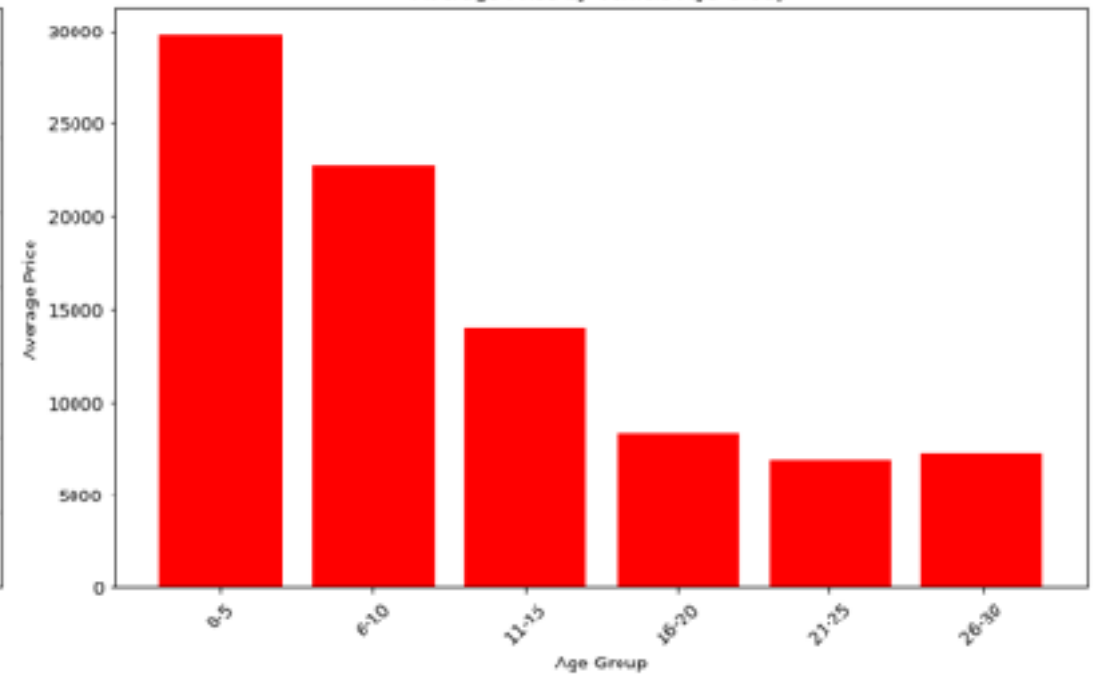
EDA:

## Vehicle Age

Count of Cars by Age Group

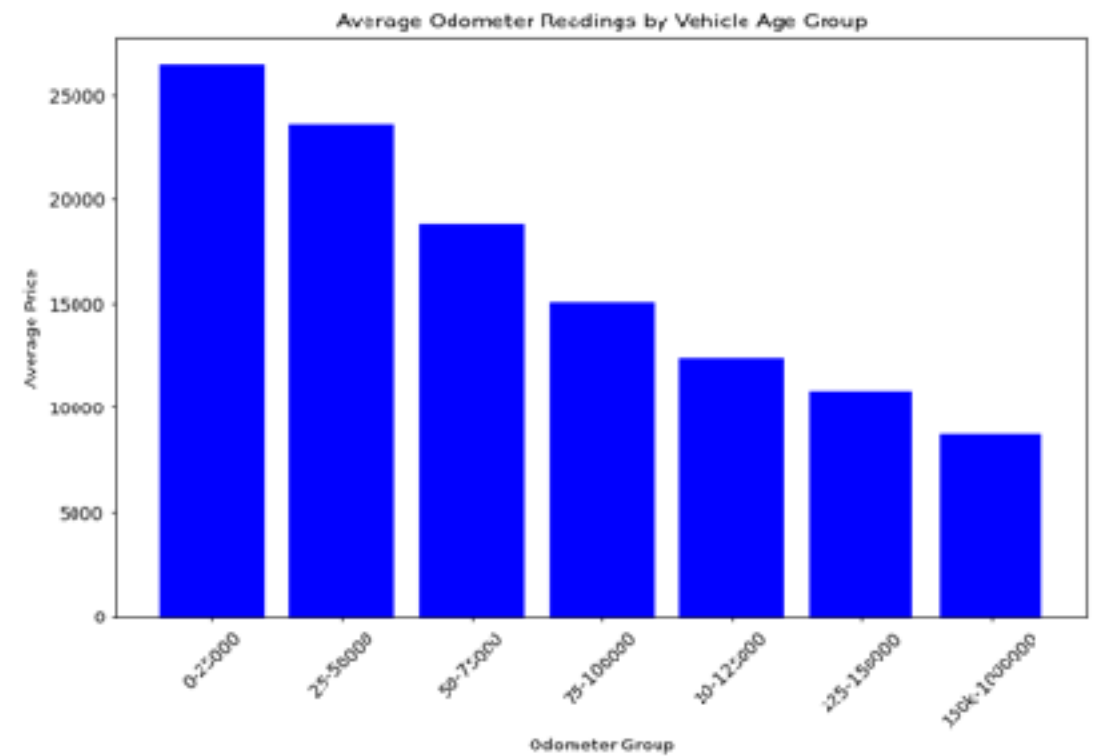
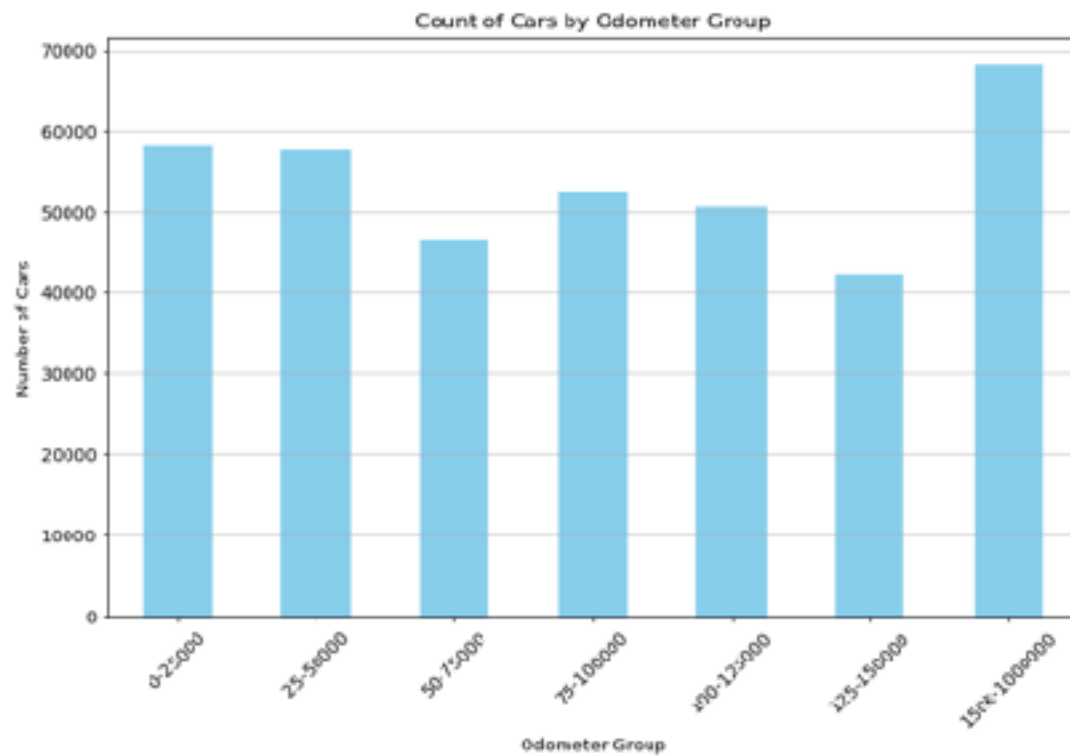


Average Price by Vehicle Age Group



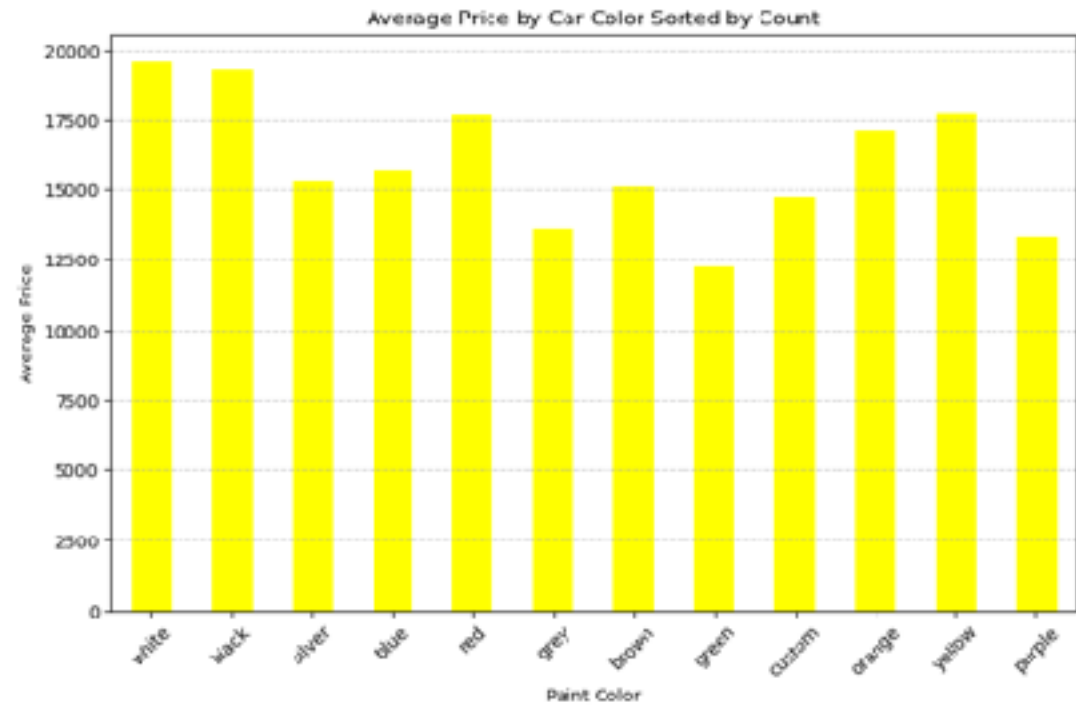
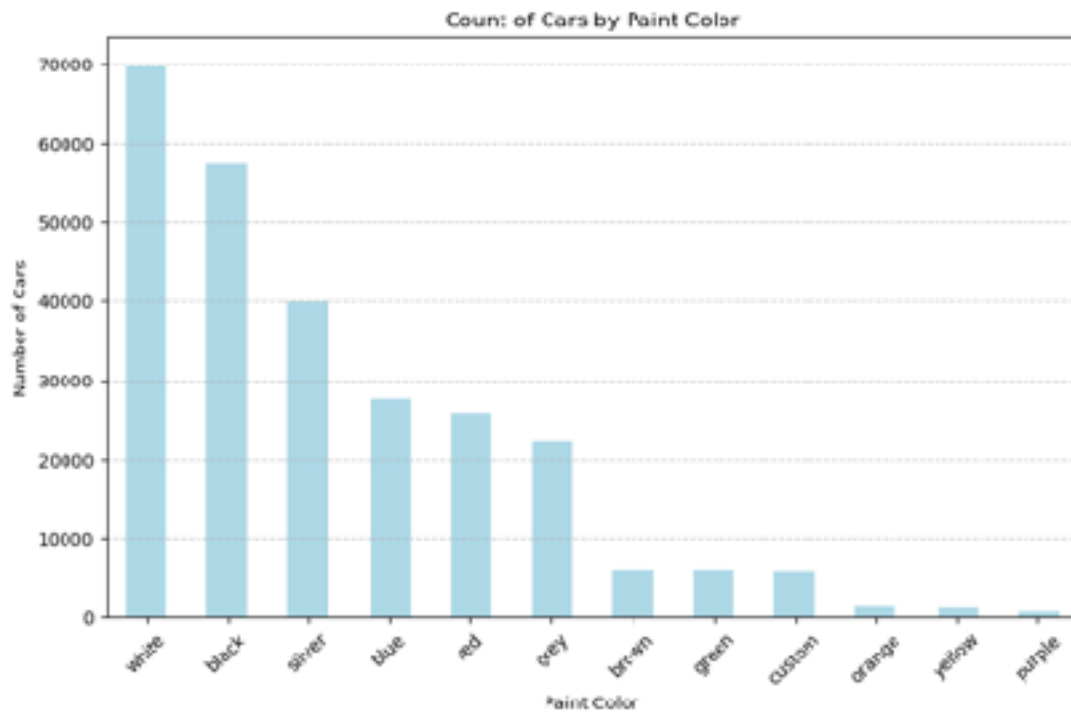
EDA:

## Vehicle Odometer Reading



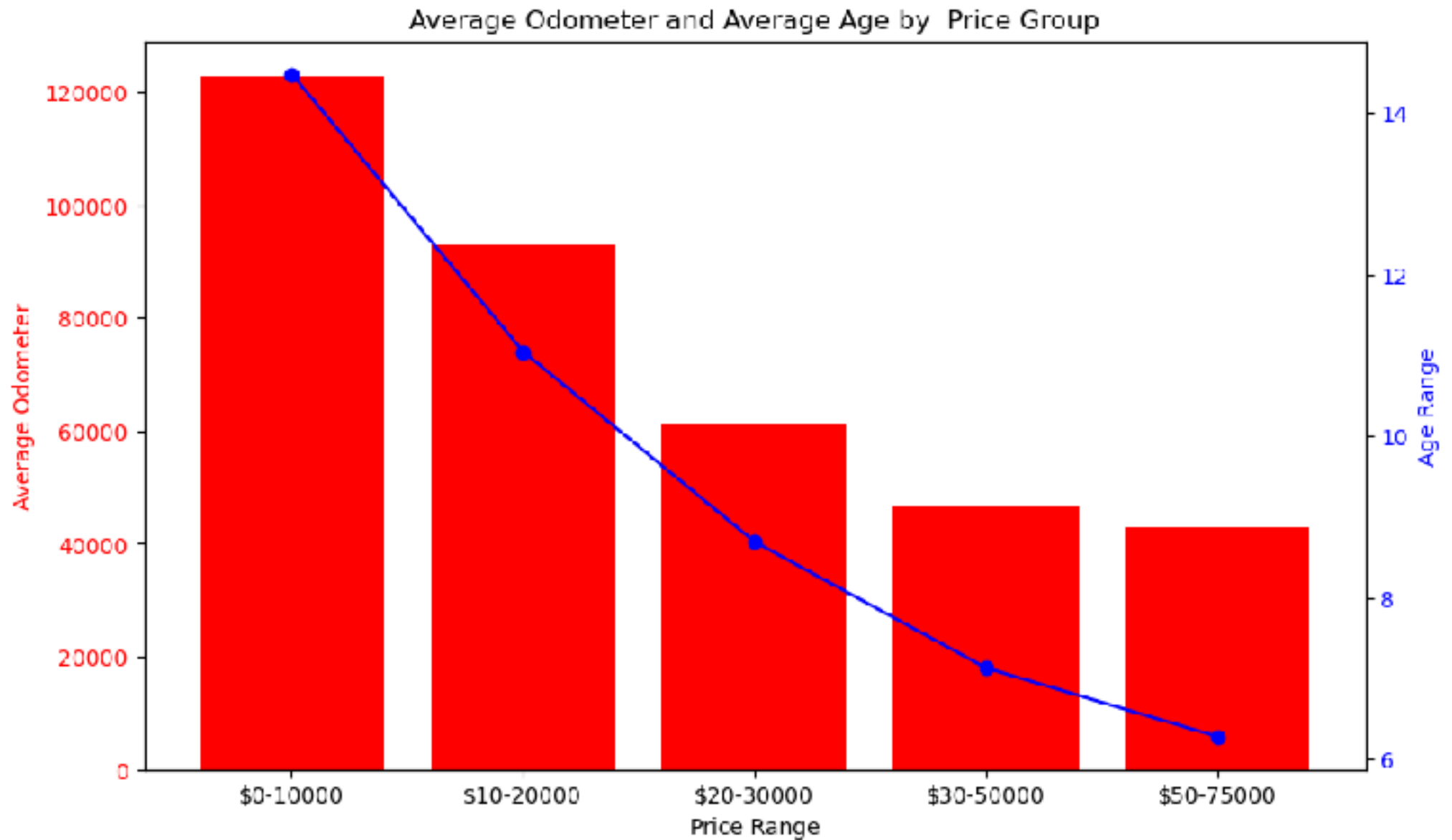
# EDA:

## Vehicle Age



EDA:

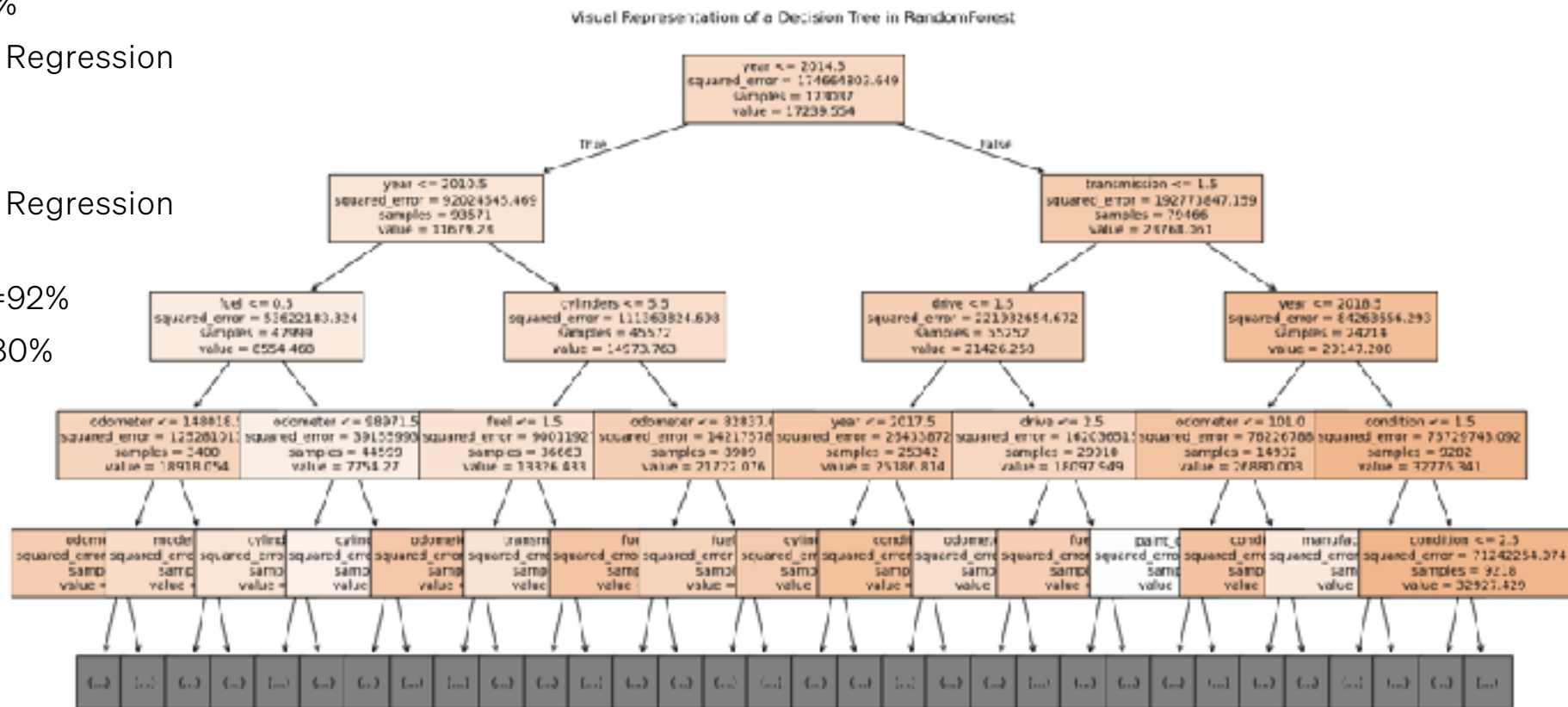
Odometer x Age x Price



# Models:

## Regression

- Base linear Linear Regression
  - $r^2 = 32\%$
- Linear Regression with Feature selector and Standard Scalar
  - $R^2 = 33\%$
- Random Forest Regression depth 3
  - $r^2 = 81\%$
- Random Forest Regression depth
  - $r^2_{\text{train}} = 92\%$
  - $r^2_{\text{test}} = 80\%$



## Conclusion

### US Used Car Dealership Asset Appraisal

- The model performs well and can predict the price of a used car based
- From the model results it is likely that the relationship is not linear and better modelled using a non-linear modelling technique
- The model will not perform well with outliers i.e supercars or custom cars