Ontario Vehicle Pricing Predictions

1. Introduction

a) Problem Statement

Searching for vehicle pricing information can be a daunting task. At times listing prices can be deceiving for unexperienced individuals looking to buy, sell or trade their vehicle. The market is vast and the information can be inconsistent.

b) Background

Over the last five years the Ontario vehicle market has shown a somewhat consistent trend. With the exception of a few months impacted by the COVID-19 pandemic (Mar 2020-Jun 2020) the market seems to follow similar seasonal peaks and troughs. Peak sales occur in the summer months while the yearly minimums seem to occur around winter months.

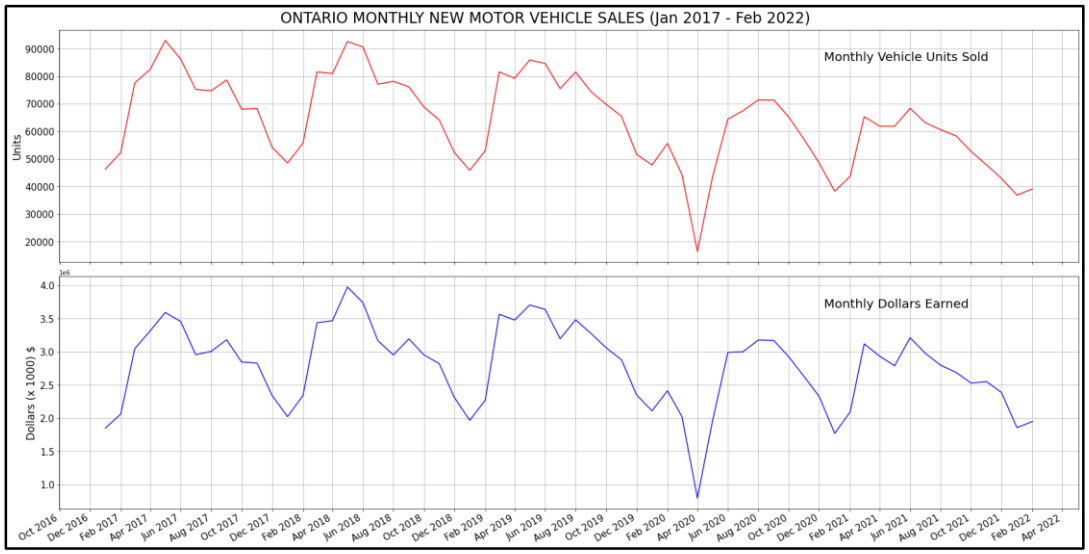

Figure 1 -Line graph of units sold and dollars earned for vehicle sales in Ontario from Jan 2017 to Feb 2022

c) Goal

This project aims to provide a more thorough prediction for vehicle pricing information backed up by data-driven analysis. The analysis combines vehicle pricing information from various retailers across Ontario. Based on relevant features, such as vehicle make, model, year and mileage; we provide a predicted price point for the specified vehicle.

2. Data Wrangling and Cleaning

The data was gathered primarily from one vehicle retail site Ontariocars.ca . This site is an online marketplace that brings together buyers and dealers in Ontario. They deal specifically with Ontario licensed motor vehicle dealers. Vehicle data was extracted from this site every 6 hours for a period of 2.5 months (Jan – Mar 2022).

	Title	View_URL	odolbl	Price	vehdetailslbl	colorlbl	Name	Location	Current_time
0	2019 Kia Forte	https://www.ontariocars.ca/kia/forte/ex/2019/3...	48500.0	23995.0	6 Speed Automatic, 4-Cylinder, FWD, Gasoline Fuel	CARBONIZED GREY Sedan	Eckert Auto Sales	Barrie, ON	2022-02-18 16:14:19.593
1	2016 Kia Forte Koup	https://www.ontariocars.ca/kia/forte-koup/sx/2...	130000.0	12995.0	6 Speed Automatic, 4-Cylinder, FWD, Gasoline Fuel	Alpine White [White] Coupe	Eckert Auto Sales	Barrie, ON	2022-02-18 16:14:19.679

Figure 2 - Sample screenshot of raw data retrieved from the retailer site.

a) Removing Duplicate Data

As the data was scraped over a period of months and the website continued to post new listings, the length of the data was quite vast and included quite a number of duplicate listings. Approximately 2% of the raw data was kept after collection.

```
Length of data before duplicates = 8602126
Length of data after duplicates = 195029
```

b) Feature Extraction/Creation

A few features were found and generated from the raw data. The URL offered information that was specific to the vehicle make, model, year and other relevant vehicle details. This information was used to create 4 new columns.

```
https://www.ontariocars.ca/ford/escape/xlt/2012/469022?VehicleTypeCategory=0&
BodyStyle=SUV&dealerId=1276&page=1
```

The odometer value also gave information about a vehicle's condition. A new feature column for vehicle condition was created based on vehicles with greater than 200 km of mileage and less than 200 km.

c) Outlier detection

After some preliminary plotting of values there were some outliers noticed in the data.

```
Max odometer value = 123456790.0    Min odometer value = 1.0
Max price value = 2399500.0         Min price value = 1.0
```

The interquartile range was used to identify outliers. For the 'Price' column, the low range was set at 0.05 and the high range was set at 0.95. For the 'mileage' column, only the high range was filtered for to account for new vehicles that may be labeled as 0 mileage.

d) ID Creation

Once the data was free of duplicates a hash value was created and assigned for each unique listing. The final dataset length was 139101 unique listings.

	Vehicle_Name	Vehicle_Model	Condition	odolbl	Year	Price
130144	Toyota	Tacoma	Used	94,754.00	2016	36,998.00
96003	Chevrolet	Silverado-1500	Used	22,848.00	2021	74,990.00
13122	Honda	Accord-Sedan	Used	50,025.00	2018	29,990.00

Figure 3- Sample screenshot of major features used for modeling

3. Exploratory Data Analysis and Initial Findings

a) Most common vehicle features

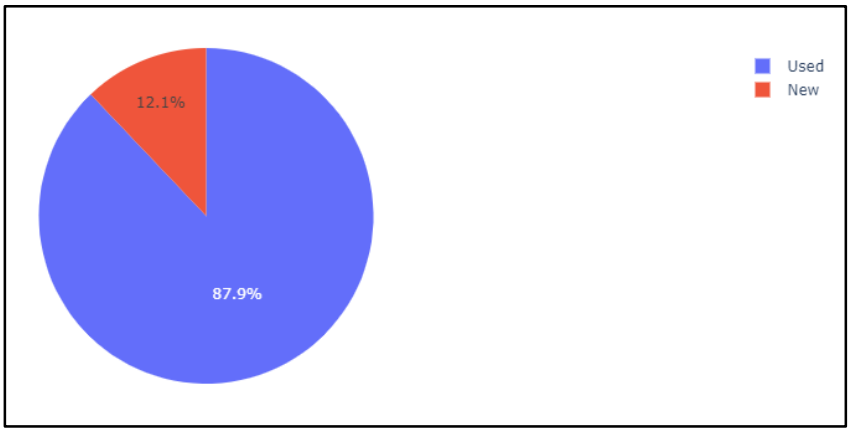

Figure 4- The dataset consisted primarily of used vehicles

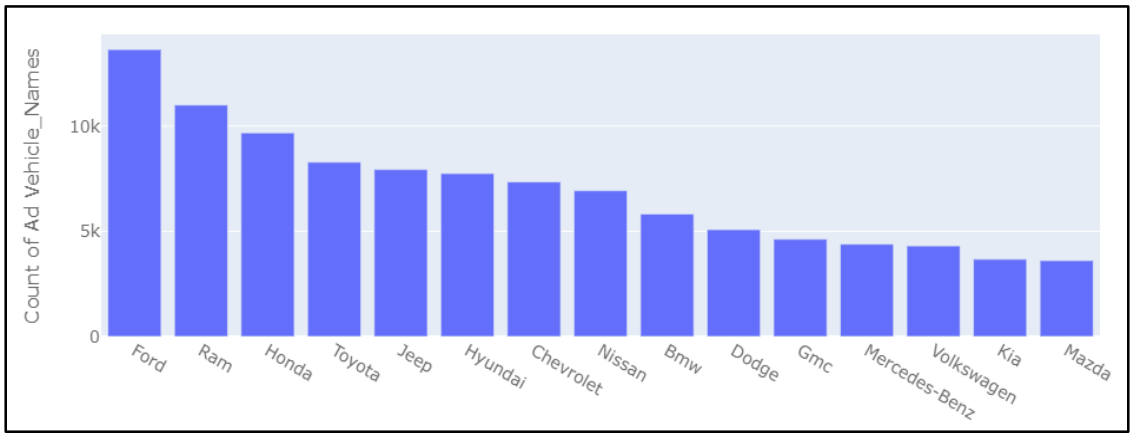

Figure 5 - Approximately 13% of all vehicle makes were from Ford. Ram (10.6%), Honda (9.3%) and Toyota (7.9%) were the next most represented makes. One of the least represented vehicle models were Volkswagen (4.1%), Kia (3.5%) and Mazda (3.4%).

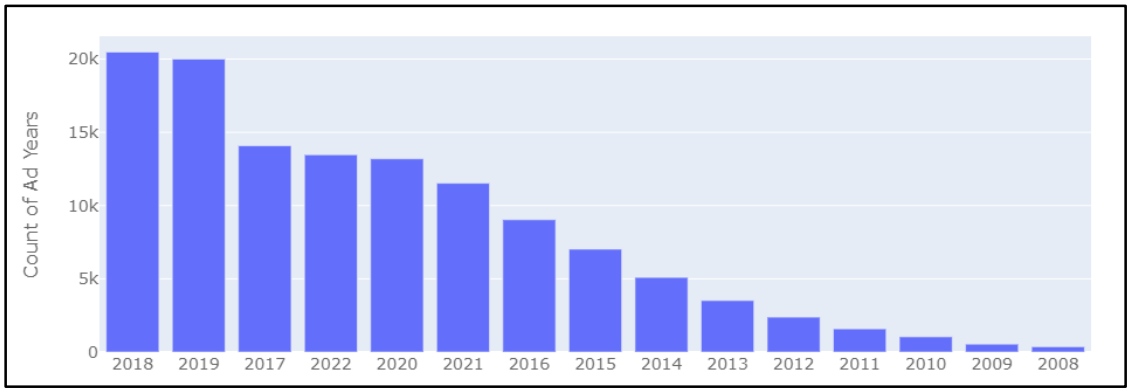

Figure 6 - Approximately 32% of all listings were either 2018 or 2019 vehicles. This suggests that at 3-4 years is a popular time to sell a vehicle.

b) Numeric Feature Association

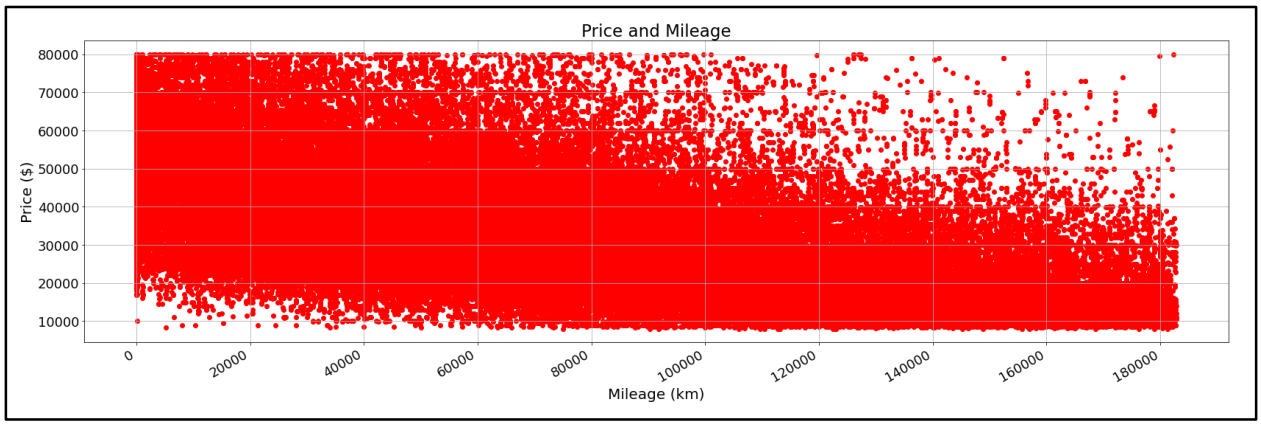

Figure 7 – The relationship between price and mileage seems to be quite distinct. Vehicle mileage seems to be a strong predictor of price.

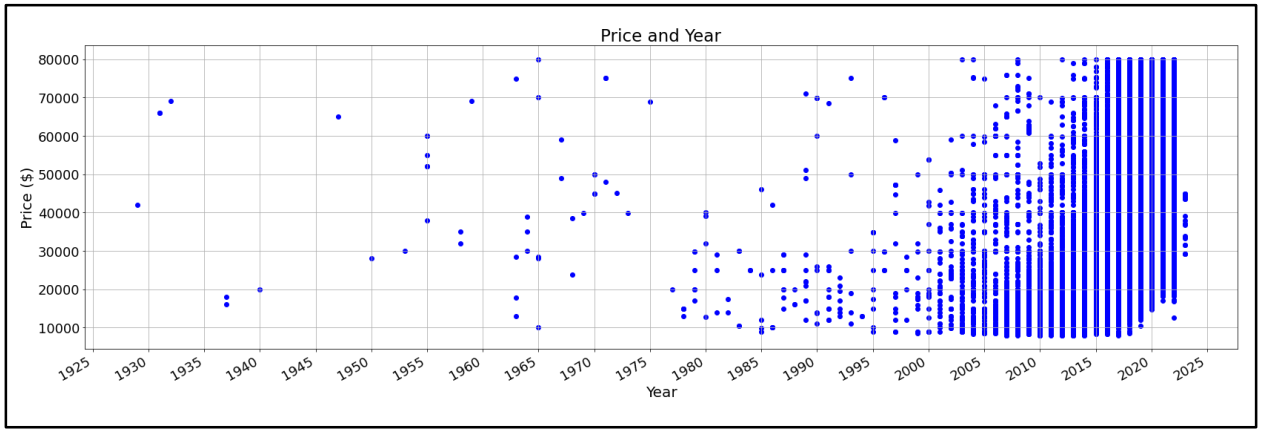

Figure 8 - The relationship between price and mileage seems to be somewhat definite. Selling price decreases overall for older/ageing vehicles.

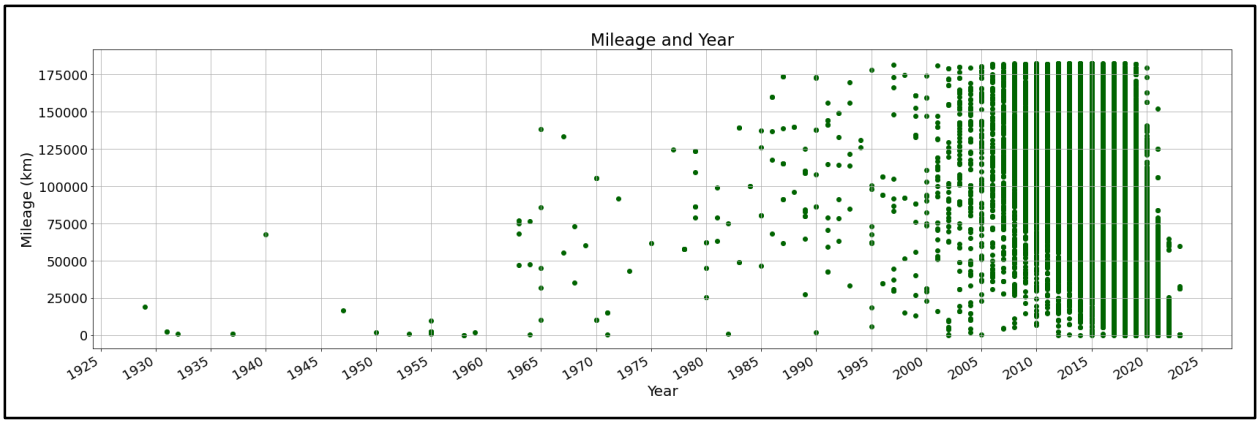

Figure 9 – An expected relationship distribution as older vehicles generally have higher mileage. Interesting to note that older, low-mileage vehicles are most likely vintage or classic cars.

4. Future Work

This project revealed that there is quite a bit more to be explored in analysis of this data. The current analysis and modeling served as a starting point for future work. There are quite a few areas for further research that are worthwhile to delve into.

a) More vehicle features

Given that only a few vehicle features were used and extracted to create the current model, there are a number of vehicle elements that could be explored and added to enhance the model. A few of them are listed below.

- Color
- Body style
- Engine
- Transmission
- Fuel type
- Cylinders
- Trim
- Doors

b) Sale location

Although not explored in this analysis and model. The effects of geographic location for vehicle sale may have an effect on vehicle pricing. This is especially true in secondary markets or with private retailers. Vehicles sold in or nearby larger cities may be sold for a different price when compared to smaller ones. Generally, prices tend to be higher in larger cities due to consumer demand.

c) Time of year

As noted from the Statistics Canada new motor vehicle sales report, there are more units sold and dollars earned in the summer months when compared to the winter months. This suggests that the time of year could have an effect on vehicle sale price as well.

d) Explore more automotive dealers

There are a variety of online vehicle dealer sites that offer similar serves within Ontario. A more robust model would use vehicle data captured from these sites as well. Some of these sites would include *ucda.ca*, *kijijiautos*, *cars.ca*, *autotrader.ca* etc.

5. References

Data Table without a Digital Object Identifier (DOI)

Statistics Canada. (2019). *Table 20-10-0001-01 (formerly CANSIM 079-0003) New motor vehicle sales* [Data table].
https://doi.org/10.25318/2010000101-eng