**Assignment 1 – Clustering For Web Scraped Data**

R. B. C. M. W. Thulana Vimukthi Abeywardana (C0861333)

Praveen Mahaulpatha (C0860583)

Tomson George (C0857730)

Jaskaran Sign

Artificial Intelligence and Machine Learning

2022F AML 2203 2 - Natural Language Processing

Vahid Hadavi

November 16, 2022

**Table of Contents**

[1. Introduction 2](#_Toc119254938)

[2. Methodology 3](#_Toc119254939)

[3. Web Scraping 3](#_Toc119254940)

[4. Exploratory Analysis & Data Wrangling 3](#_Toc119254941)

[5. Outlier Detection and Removal 9](#_Toc119254942)

[6. Encoding and Normalization 9](#_Toc119254943)

[7. Clustering 9](#_Toc119254944)

[6. Time series Prediction 9](#_Toc119254945)

[6. Conclusion 12](#_Toc119254946)

[7. References 12](#_Toc119254947)

# 1. Introduction

In this report, data related to vehicle sales were scraped from kijiji.com and Data wrangling processes were followed to get the data in the desired format and finally clustered with respect to Price, Age and Mileage of a corresponding vehicle. The goal was to find similar clusters hence enabling us to target market-specific groups of interest. In this report, we will be discussing the main steps, difficulties encountered and solutions we have taken to overcome them, the conclusion of the project and the use cases of the results. Further, natural language processing steps were carried out to find out the most commonly used words in the description feature of vehicles to further understand the market place.

# 2. Methodology

First, it was necessary to understand the dataset. We have followed the basic steps of data exploratory analysis in order to achieve this. By that, it was evident the dataset contained many categorical features compared to numerical features. Also some of the features contained inconsistent data which had to be structured for the desired format. Further, some of the categorical features contained a high number of unique values making it inefficient to encode hence they were dropped under the note of consideration at a later phase of the project. Once these features were dropped, data was plotted in graphs for a deeper understanding of distribution and relationships between instances. Following the visualization, we identified outliers and they were treated in 3 different methods for selected features. Then the selected categorical variables were encoded to fit the mode and clustering was performed for normalized data. Kmeans clustering and Meanshift clustering methods were used in this step and the results will be further discussed in the following sections of the report.

# 3. Web Scraping

The first requirement for the project was to produce a dataset by scraping the web. For that, we have selected the below link which contains advertisements relevant to vehicle sales. From the web page, we identified there were close to 44 advertisements on each page and these advertisements contained below features and they are being mentioned in the Figure 3.1.

URL: <https://www.kijiji.ca/b-cars-trucks/gta-greater-toronto-area/new__used/c174l1700272a49>

* Vehicle Name
* Location
* Price
* Description
* Mileage
* Transmission

Name

Graphical user interface, website

Description automatically generated

Location

Advertisements

Price

Description

Transmission & Mileage

Figure 3.1

Even though we could extract all these features of a vehicle, more features were available once you clicked on the vehicle name, such some information available are shown in the figure 3.2. But in order to fully view the information, it was necessary that we clicked on the ‘see more’ button. This information was extracted as well.

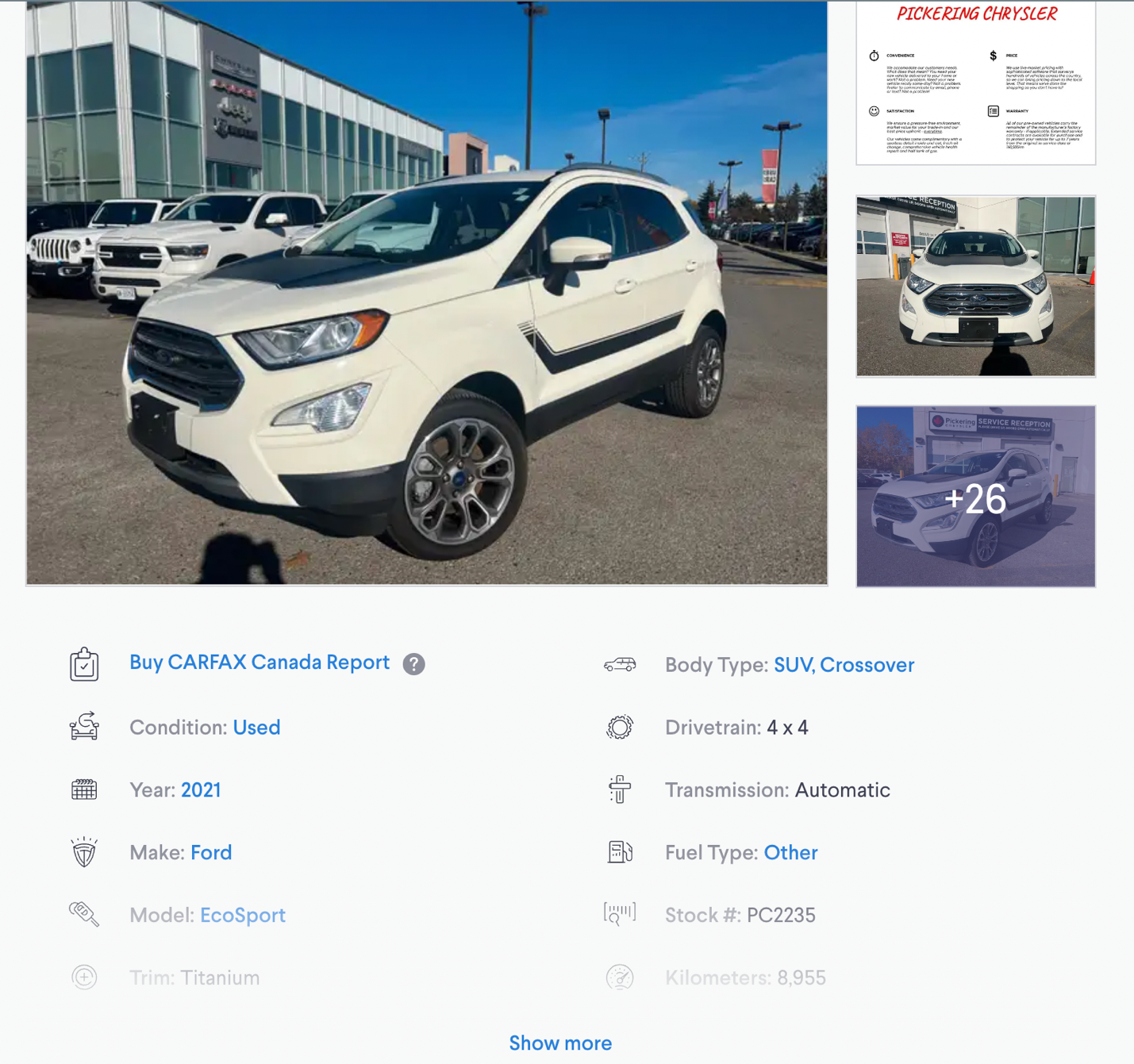


Figure 3.2

More Information

‘Show more’ Button

The overall scraping program run through a loop defined by the number of pages we want to extract and then get 44 advertisement elements to a list using selenium and iterate through each advertisement to get the features in the first layer. Secondly, it clicks the vehicle name and navigates to the more information page and clicks on the ‘Show more’ button. Then the rest of the information is extracted and saved in a dictionary. We encountered difficulties in waiting time for the google chrome driver to load a page and also some exceptions were thrown when some of the attributes did not exist. These were dealt with by using the sleep function by time module in python and exceptions were caught and a Nan value was saved for non-existing attributes. The figure 3.3 shows the first few rows of the data frame that was created using the scraped data. The data frame was then converted to CSV and saved to a local folder for the usage of clustering in the latter phase.

Graphical user interface, text, application, email

Description automatically generated

Figure 3.3

# 4. Exploratory Analysis & Data Wrangling

The main objective of this step was to gain a better insight into the dataset that we have in hand. The dataset contained 421,993 instances and 19 features. And the corresponding data types for features are shown in Figure 4.1.

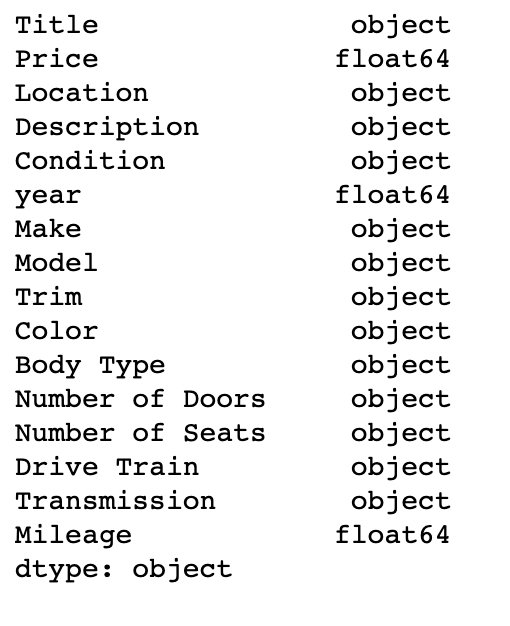


Figure 4.1

As we required some features to be in numerical format, we converted the features to get the desired format and the data types were obtained as below in Figure 4.2. (Price, Number of Doors, Number of Seats, Mileage were changed to float type)

A screenshot of a computer

Description automatically generated with low confidence

Figure 4.2

Later, the number of unique values was checked for each categorical feature. The results were as shown in the Figure 4.3.

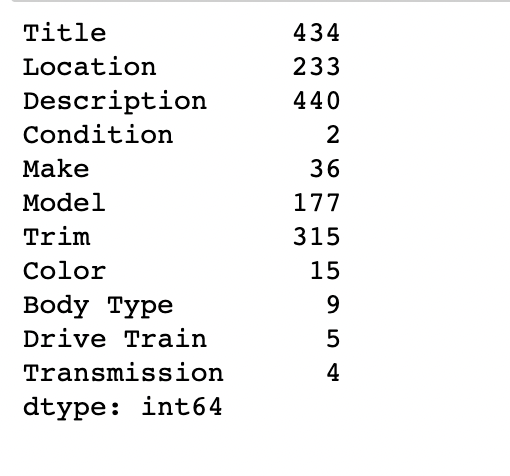


Figure 4.3

We decided to remove ‘Title’, ‘ Location’ , ‘Description’, ‘Make’, ‘Model’, ‘Trim’ as they had too many unique values under the assumption that the gain received by encoding them would be less compared to the efficiency gained by dropping them for the clustering.

**4.1 Identifying and Imputing Null values**

In this section, each feature was checked for the null value percentage it had against the total instances. The figure 4.4 shows the null value percentages for each feature in a bar chart.

Chart, bar chart

Description automatically generated

Figure 4.4

Null values were treated in two different ways,

* + 1. Imputation using KNN Imputer for Numerical Features

For numerical features, we have decided to KNN imputer with setting the n\_neighbors as 3. KNN imputer utilizes the k-Nearest Neighbors method to replace the missing values in the datasets with the mean value from the parameter ‘n\_neighbors’ nearest neighbours found in the training set. By default, it uses a Euclidean distance metric to impute the missing values.

We have performed this for below features

* Price
* Year
* Mileage
* Number of Doors
* Number of Seats
  + 1. Mode imputation for categorical Features

The most common class was imputed for a categorical feature assuming the missing values have a higher chance of falling in for the most common class.

We have performed this for below features

* Condition
* Color
* Body Type
* Drive Train
* Transmission

**4.2 Feature Engineering**

In this step, in order to get an intuitive meaning out of the column 'Year' feature, we can find out how old the vehicle is by subtracting the model year from current year. This new feature will be added as 'Age'.

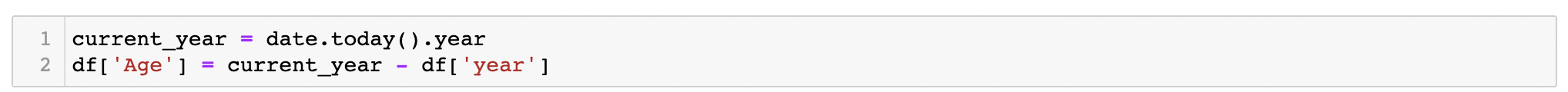


Figure 4.5

# 5. Visualization

**5.1 Numerical Feature Visualization**

**Histogram**

Out of many visualization techniques, we selected the histograms to understand the distribution of the data.

Chart, histogram

Description automatically generated

Figure 5.1

Figure 5.2 gives below discoveries:

* The ‘Price’ and ‘Age’ features are uniformly distributed
* The ‘Mileage’ feature is right skewed

The log transformation for the Mileage feature to reduce the skewness has been performed as well as a method for outlier treatment which will be discussed in the following sections of the report.

**Correlogram**

Secondly, we wanted to check the relationship between each numerical variable. For that, a correlogram was plotted. A correlogram plots scatter plots for each feature against a different feature and also plots a histogram along the diagonal.

Chart, scatter chart

Description automatically generated

Figure 5.2

From figure 5.2, we have discovered below,

* There seems to have a positive correlation between Mileage and Age,
* There seems to have a slight negative correlation between Price and Mileage

Both of these will be checked by plotting the correlation coefficient between them in the next section.

**Heatmap**

To further understand the correlation observed in the correlogram, the below heatmap was plotted with the correlation coefficient among the numerical features. As observed it concludes that there is a positive correlation between Age and Mileage and also between the Number of doors and the Number of seats. Similarly, a negative correlation of -0.5 was observed between Price and Mileage.

Chart

Description automatically generated with medium confidence

Figure 5.3

**5.2 Categorical Feature Visualization**

For visualizing categorical features, we have selected a count plot to get an idea of the occurrence of each class.

Graphical user interface, chart, bar chart, waterfall chart

Description automatically generated

Figure 5.4

Figure 5.4 shows below observations,

‘Condition’ and ‘Transmission’ features seem to have an imbalanced distribution. We could under-sample these two features to get a balanced dataset but since the number of instances is less, we will not proceed with this step at the moment.

# 6. Pandas Profiling

What pandas profiling does is save all the work of visualizing and understanding the distribution of each variable. It generates a report with the information available regarding the data frame.

Graphical user interface

Description automatically generated

Figure 6.1

# 7. Outlier Detection and Treatment

**7.1 Outlier Detection**

Apart from the observation, we got by looking at skewed distributions, another straightforward method to detect outliers is by plotting boxplots. The idea is that once the boxplots are plotted, the data points that fall outside of the lower and the upper whiskers are considered outliers.

Chart, box and whisker chart

Description automatically generated

Figure 7.1

The lower whisker is defined by : Quantile 1 – (IQR\*1.5)

The upper whisker is defined by : Quantile 3 + (IQR\*1.5)

From figure 7.1 we observed that Price, Mileage and Age have data points falling outside of upper whisker.

**7.2 Outlier Treatment**

Since outliers are considered as extreme values which could hinder the performance of the machine learning model, it is necessary that we treat these. There are 3 main methods for outlier treatment which will be discussed in this report. They are,

* + 1. Log Transformation
    2. Trimming
    3. Quantile Based Flooring and Capping

***7.2.1 Log Transformation***

Log transformation is a common technique used to remove the skewness of a feature thus making it closer to a normal distribution. By this, data gets closer to the mean and possibly brings outliers close to the center. This has the probability that that specific outlier no longer exists in the extreme value range. Figure 7.2 shows the distribution for the feature ‘Price’ before the log transformation with the corresponding skewness value.

**Chart, histogram

Description automatically generated**

Figure 7.2

In below figure 7.3, the distribution for the log-transformed Price column is been shown with clearly more normally distributed data and a lower value of skewness.

Chart, histogram

Description automatically generated

Figure 7.3

Each of the 3 main features log transformed was plotted against their original feature distributions in Figure 7.4 where we can observe a significant reduction in skewness for features Price and Age but for Mileage the skewness seems to be increasing. For this matter, we will have to continue with other methods as well.

Chart, histogram

Description automatically generated

Figure 7.4

***7.2.2 Trimming***

Trimming is a method to remove data that falls out of a specified range or a condition. In this project, we have demonstrated the trimming based on whiskers. Figure 7.5 shows the shapes of features before and after trimming.

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Figure 7.5

***7.2.2 Quantile-Based Flooring and Capping***

This method is used to replace the outliers which fall outside of the lower and the upper whiskers to be replaced by the whisker values. Outliers that are lower than the lower whisker will be replaced by the value of the lower whisker and outliers that are larger than the upper whisker will be replaced by the value of the upper whisker.

Chart, histogram

Description automatically generated

Chart

Description automatically generated

Chart

Description automatically generated

Figure 7.6

Figure 7.6 shows the boxplots relevant to the 3 features and we can observe that there are no outliers beyond the whiskers.

# 8. Encoding Categorical Features

Categorical features must be converted before being fitted to a machine-learning model. Even though we do not use them in this project for clustering. But they can certainly be used for further experimentation of this project. Figure 8.1 shows the categorical feature and its respective number of unique values. It is necessary that we first understand what are nominal and ordinal features out of the list and also their number of unique values. We have identified that all of the categorical features we have at hand are nominal.

Text

Description automatically generated

Figure 8.1

Out of the above, the ‘Condition’ feature has only two distinct values hence we have performed binary encoding to assign 0 for ‘used’ and 1 for ‘new’ respectively. The results are shown in figure 8.2.

Text

Description automatically generated with low confidence

Figure 8.2

Also for features which has more than 2 unique values and nominal, we have proceeded with one hot encoding. This was carried out for ‘Color’ and ‘ Drive Train’ features. The output new columns appended are shown in figures 8.3.

A picture containing text

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

Figure 8.3

# 9. Normalization

The selected features for the clustering are ‘Price’ , ‘Age’ , and ‘Mileage’. Since these features have distinct value ranges, the clustering model will not perform effectively without some degree of rescaling to a common range. This is achieved by the minmax scaler offered by sklearn library. Figure 9.1 represents the data distribution for the normalized features and the x axis ranges between 0 and 1 for all three.

Chart, histogram

Description automatically generated

Figure 9.1

# 10. Kmeans – Clustering

**10.1 Finding ideal number of clusters**

The parameter that we need to define in K means is the number of clusters the algorithm should divide the dataset into. For this the commonly used method is the elbow method where the error is plotted against a given number of clusters. Figure 10.1 shows the error plotted for number of clusters varies from 1 to 10.

Chart, line chart

Description automatically generated

Figure 10.1

From the above plot, we can conclude that after 4 clusters the reduction in error is low hence we take the ideal number of clusters as 4.

**10.1 Splitting data into Training and Testing set**

Once the number of clusters is identified, we split the data into training and testing. And we have taken 75% of the data for training and 25% for testing.

Figure 10.2 represents the 4 clusters found on the testing dataset and the 4 cluster points are represented by red star.

Chart, scatter chart

Description automatically generated

Figure 10.2

Also, an interactive 3D plot was created using plotly library for the 2 features which is represented in figure 10.3

Chart, bubble chart

Description automatically generated

Figure 10.3

# 11. Mean Shift Clustering

By using the mean shift clustering only two clusters were identifiable and the centroids are marked with a red cross in figure 11.1.

Chart, scatter chart

Description automatically generated

Figure 11.1

# 12. Finding Trending words in the Description

To further understand the commonly used keywords in the description feature of a vehicle, in the project we have applied below NLP concepts.

* Regular expression to extract alphabetical characters
* Converting to lowercase for dimensionality reduction
* Tokenizing
* Stop word removal
* Lemmatization
* Frequent Words count

The results were plotted in a line graph for better visualization as shown in figure 12.1

Chart, histogram

Description automatically generated

Figure 12.1

# 12. Conclusion