Advanced Python AI and ML Tools

Team Member Details:

Sneh Patel – C0869367
Malav Shah- C0870066
Dhruvraj Chavda – C0867457
Nandiniben Patel – C0869773

Introduction:

- Kijiji is a popular online marketplace for buying and selling cars in Canada.

- The Kijiji website contains vast data on vehicles, including make, model, year, price, and location.

- In this report, we will use advanced Python techniques to extract data from the Kijiji website and analyze it to gain insights into the car market in Toronto.

- Our Dataset Consist of various Features they are as follows.
  - 'brand'
  - 'model'
  - 'model_year' : Type: int
  - 'list_price' : Type: int
  - 'color'
  - 'configration'
  - 'condition'
  - 'body_type'
  - 'wheel_config'
  - 'transmission'
  - 'fuel_type'
  - 'mileage': Type: int
  - 'carfax_link'
  - 'vin_number'
  - 'image_link'
  - 'dealer_address'

About Data Scraping:

- Data extraction from websites is known as data scraping. In this instance, we are harvesting information from Toronto Kijiji's automobile listings.

- To do this, we extract the web page's HTML source code using the URL lib and BeautifulSoup Python tools and parse the pertinent information.

- It obtains the webpage's HTML code. The appropriate data is then extracted from the HTML code using Beautiful Soup's find () function, which searches for HTML elements matching specific criteria (like the item prop property) and removes the text from those components.

- To maintain a consistent data structure across all the objects we scrape, we set the value of any information we are trying to extract to "NA" (short for unavailable) if it cannot be discovered.

Data Wrangling:

Data Wrangling involves the various steps:

- Data collection: The data was scraped from the Kijiji website using Python packages like BeautifulSoup and Requests. For each automobile listing, the data is present as HTML text.

- Data cleaning: The HTML content for each car listing needs to be cleaned and parsed to extract relevant information. It removes unwanted HTML tags and other formatting elements irrelevant to the data analysis.

- Data structuring: It is necessary to organize the retrieved data for each car listing in a manner that can be quickly evaluated. It can entail building a database table or data frame with columns for every information item.

- Data validation: The extracted data needs to be validated to ensure accuracy and consistency. This could involve checking for missing or incorrect values and identifying any outliers or anomalies in the data.

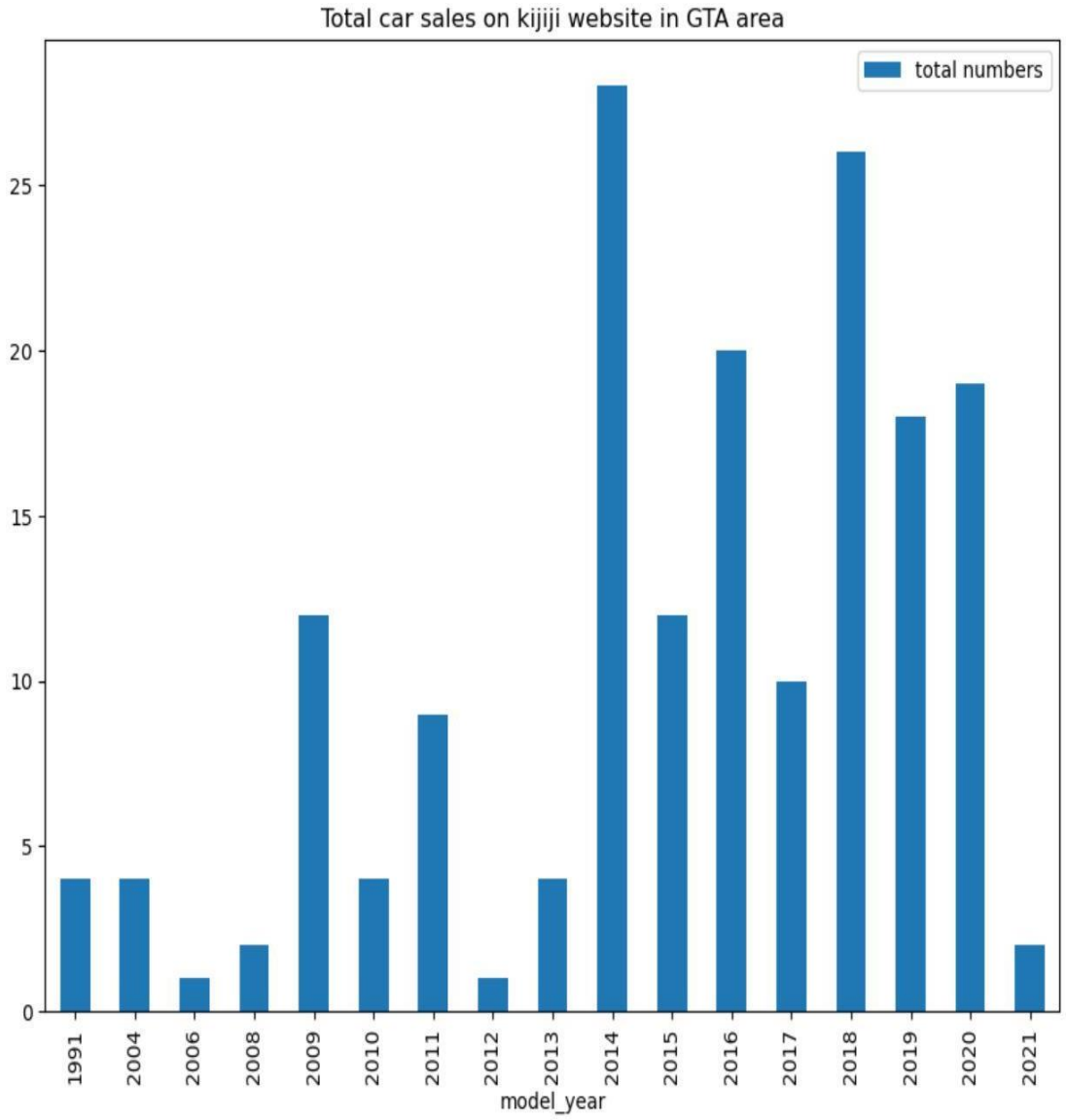
- Data formatting: To make the extracted data accepted for analysis, it can be necessary to format the data in a certain way. For instance, changing the price and mileage data to numeric values while maintaining the date field's original format could be required.

- Data enrichment: Additional information may need to be added to the dataset to make it more useful for analysis. It involves merging the car listing data with other data sources, such as information about the make and model of the vehicle.

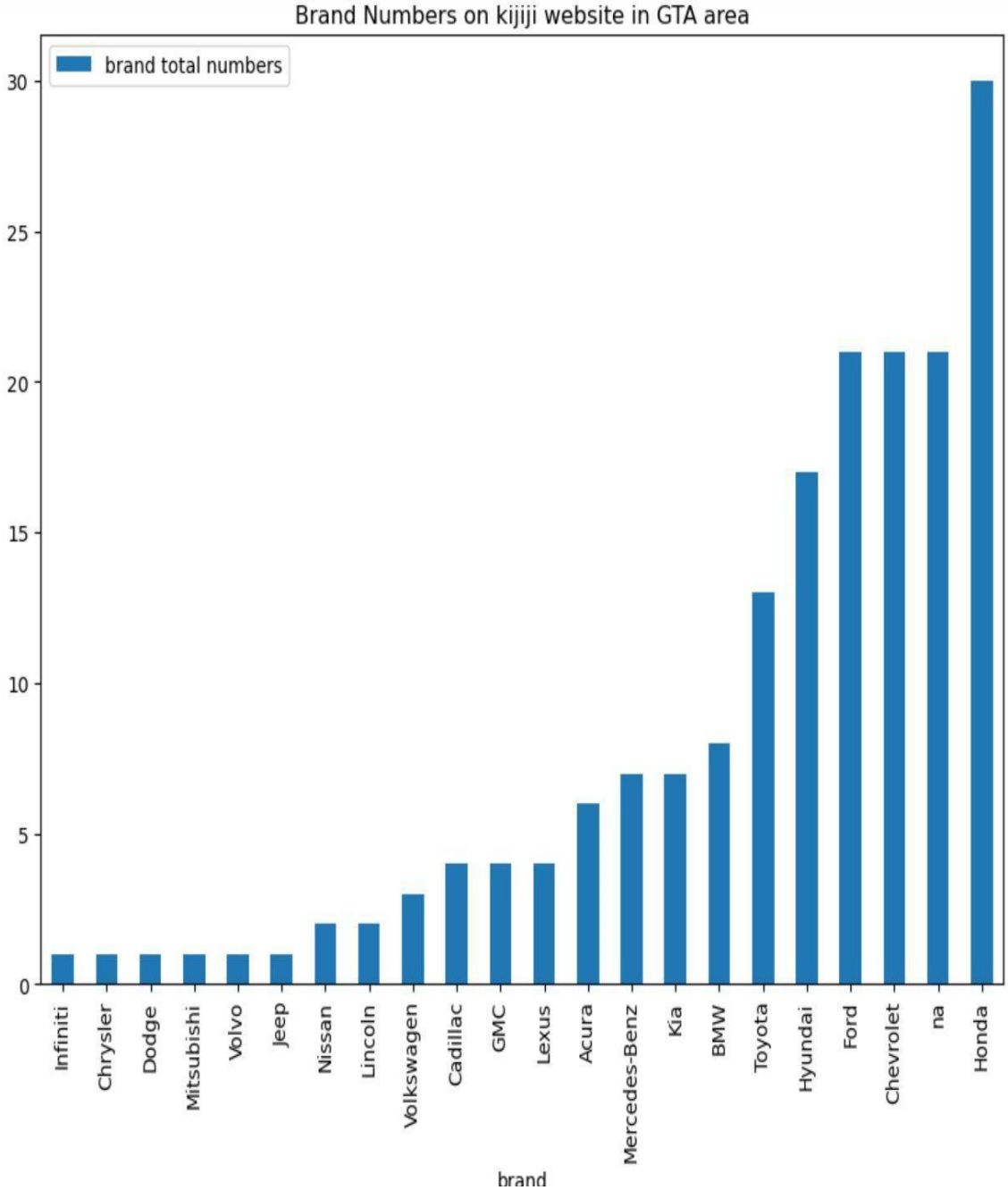
  Overall, the goal of data wrangling is to prepare the data in a way that makes it suitable for analysis and ensures that the results of the analysis are accurate and meaningful.

Data Visualization among various attributes:

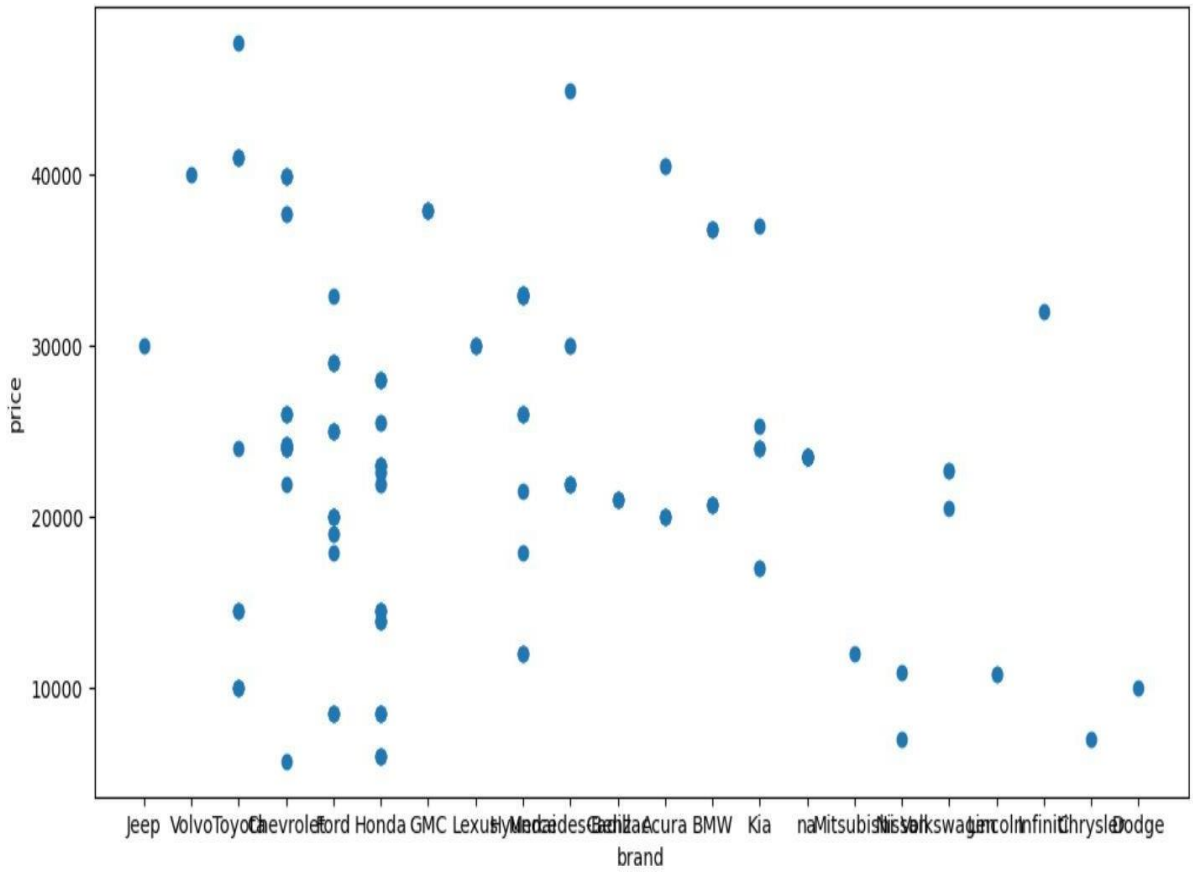
- Total Car Sales on Kijiji website in GTA Area VS Model Year.



- Brand Numbers on Kijiji Website in GTA Area.



- Scatter Plot of Brand Vs Price



Pandas Profiling:

- You need to change your data.csv and your report.html for Pandas Profiling to reflect the names of your CSV and report files, respectively.

- Once the report has been created, you may explore the data analysis findings by opening it in any web browser.

- The report thoroughly analyzes your data, including fundamental statistics, different variable kinds, missing values, correlations, and visualizations.

Encoding Methods:

- We discovered after examining the data that some of the columns in our dataset are categorical variables and are not appropriate for our machine-learning models.
- Hence, to translate these variables into numerical values, we need to encode them.
- Categorical variables in pandas are transformed into numerical variables using the pd. get dummies () method. In this instance, it transforms the car df dataframe's'model' column into a series of binary columns, each representing a different value in the model's column.

Box Plot And IQR

- To identify outliers in each column of the car_df dataset, we can use boxplots and the interquartile range (IQR) method.
- The IQR method is based on the range between the 75th and 25th percentiles of a column, and any data point outside of the range of 1.5 times the IQR below the 25th percentile or above the 75th percentile is considered an outlier.

Various techniques to address the outliers.

To address outliers in the car_df dataset, we can use various techniques, including quantile-based flooring and capping, trimming, and log transformation.

Quantile-based Flooring and Capping: With this approach, a column's lower and upper bounds are determined based on a certain percentile. Any values below or beyond those limits are replace with the corresponding limitations. We can put this into practice using the quantile () function in pandas.

Trimming: This method entails cutting off the distribution's top and lower tails to remove outliers from a column. We can put this into practice by defining a specific range of values to maintain Boolean indexing.

Log Transformation: This technique involves taking the natural logarithm of a column to reduce the impact of extreme values. This technique is beneficial when the data is heavily skew. We can implement this using the log () function in NumPy.

K Means Algorithm:

For this assignment, we can use K-means clustering, which is an unsupervised learning method used for clustering or grouping similar data points together. In our case, we can use it to cluster similar cars based on their attributes such as brand, model, year, price, etc.

We must preprocess our data and change all categorical columns to numerical ones before we can use K-means clustering. To change category columns into numerical ones, we may utilise pandas' get dummies () method.

After preprocessing the data, we may use the K-means clustering technique to sort the automobiles into comparable groups. The elbow approach may be applied to identify the number of clusters (K) we wish to construct.

After the number of clusters has been determined, we can fit the K-means model to our preprocessed data and forecast the clusters for each automobile in our dataset. Scatterplots may illustrate the clusters and show us the similarities and differences between the various groupings.

We may benefit from K-means clustering's results in several ways. It may assist us in establishing market categories based on consumer preferences, finding comparable autos based on their characteristics, and seeing any trends in the data that may have needed to be more prominent.

It is crucial to remember that K-means clustering is sensitive to the choice of the initial centroids and the number of clusters (K). To assure the stability of the groups, it is crucial to run numerous times with various initializations and compare the outcomes.

Reference Link

Jin, T. (2019, January 10). Kijiji Web Scraping Project. https://pyligent.github.io/2019-01-10-kijiji-Web-Scraping/