Andy Qin and Brandon Salgado

CSCI 182

Professor Ghosh

5 December 2021

Final Project Report

Part1: Web and Data Mining Ethics

Data are considered private equity that can be given for public usages only if they are appropriate according to guidelines. There are no clear policies stating that web scraping/crawling is illegal, but that does not mean we can do anything with the data we collected. There are cases where data is used inappropriately that cause damage to organizations --- using data to make a profit or extracting data that are associated with the customer's account, or data that are sensitive and crucial. Violation of the Digital Millennium Copyright Act (DMCA), Violation of the Computer Fraud and Abuse Act (CFAA) is put in place to warn people from selling confidential data.

Many websites are protected against web scraping for some level, but the protection does not prevent the websites from scraping. So technically, if you know really well how to code a scraper, you can scrape any published websites. All web scrapers are Artificial Intelligence that makes HTTP requests to a target website and extracts the data from a page. This does not include private websites, which are protected with a username and password. All public sites can be scraped. Public websites are any location on the web that is accessible to anyone with an internet browser and access to the internet. All advanced websites are published with links, IP addresses, embedded data, responding front-end and back-end. As long as that information is extractable, people can do web scraping by requesting access to the content of the HTML. The most popular scraping sites include Twitter, Yelp, eBay, Walmart, YellowPages, Walmart, Google, and Amazon. Twitter, Yelp, and YellowPages are scraping-friendly --- they allow appropriate scraping with minimal protection against them. On the other hand, Google, Amazon, and professional websites like Linkedin implement complex algorithms to prevent web scraping success -- You can still do the scraping, but it is a lot harder to extract data from those websites.

None of the scraping action is illegal or unethical --- it is about what you extract and what you do with it. Performing research and data analysis on those data are supported and are considered an efficient use of resources. On the other hand, using those data for self profit and with no citations is considered unethical and illegal. It may be true that everything on the internet is public, some of the data are still partially private because they contain sensitive information and personal information that directly relate to agents and profits. Thus, it is also inappropriate to extract data like personal account information, password, and contacts. We should also read the "robots.txt" file of the website before we start scraping --- it contains the lists of data we should not extract.

Resource:

1)https://www.edureka.co/blog/web-scraping-with-python/

https://realpython.com/beautiful-soup-web-scraper-python/

2)https://www.parsehub.com/blog/beginners-guide-to-web-scraping/

3)https://www.toptal.com/python/twitter-data-mining-using-python

4)https://medium.com/edureka/scrapy-tutorial-5584517658fb

5)https://www.dataquest.io/blog/web-scraping-python-using-beautiful-soup/

6)https://www.octoparse.com/blog/10-myths-about-web-scraping
https://towardsdatascience.com/scraping-multiple-urls-with-python-tutorial-2b74432d085f

Part 2: Project Proposal

Our Research:

For our final project, we are choosing to research the Popularity of Sedans Among Americans. Specifically, we want to look at the most popular consumer sedans and collect data through the ratings on the models. We want to look more closely at what the consumers like, dislike, purchase, and do not purchase. In addition, we will look at price and location correlation among purchasing consumers. Through this, we hope to derive a general understanding of what the most popular aspects of a sedan are among the average consumer.

The potential webpages we wish to scrape include anything from car retail sites, to review pages.

Sources:

1. https://www.kbb.com/(Kelly Blue Book)
2. https://www.carmax.com/ (CarMax)
3. https://www.autotrader.com/ (AutoTrader)
4. https://automobiles.honda.com/ (Honda Official)
5. https://www.cars.com/ (Car Reviews)
6. https://smartpath.toyota.com/inventory(Toyota Official)
7. https://www.ford.com/ (Ford Official)
8. https://www.nissanusa.com/ (Nissan Official)
9. https://www.hyundaiusa.com/us/en/vehicles (Hyundai Official)
10. https://www.lexus.com/ (Lexus Official)

Sample list of sedan cars we are researching for our project(increasing and ongoing...):

Brand	Model	Year	Price(manufacturer's suggestion)	Review Link
Toyota	Camry	2020	$25,420	
Honda	Civic	2020	$20,955	
Toyota	Corolla	2020	$20,430	
Honda	Accord	2020	$24,800	
Nissan	Altima	2020	$24,800	
Ford	Fusion	2020	$23,170	
Hyundai	Elantra	2020	$18,950.	
Chevrolet	Malibu	2020	$23,000	
Nissan	Sentra	2020	$19,090	
Kia	Forte	2020	$17,790.	
Volkswagen	Jetta	2021	$18,895	
Hyundai	Sonata	2020	$23,400	
Dodge	Charger	2020	$29,995	
Kia	Optima	2020	$23,390	
Nissan	Versa	2020	$16,400	
Subaru	Impreza	2020	$18,695	
Lexus	ES	2020	$40,925	
….				
….				
Volkswagen	Passat	2020		

The specific data mining tasks that we will perform on these sites depend on what kind of site we look at. For the review sites, we want to collect keywords and phrases on positive and negative feedback of the specific product. We will do this by organizing the reviews into positive (5 stars) and negative (1 star) and collecting the keywords associated with each. On the official brand site, we want to look at the price of products and the product details to help us find the correlation with popularity among consumers and what features are most important when purchasing to the average consumer. From the car retail sites, we can extract data for both the price and the ratings.

After we collect all the data, we want to integrate the data and normalize, smooth them so they are consistent. We will then perform analysis on all the data to find any association rules, important keywords queries, and correlations.

Part 3: Project Implementation and Report

Tools used:

- **Python, Python Libraries, Google Colab, Excel,** and **the internet.**
- Specific web scraping tools**: BeautifulSoup, Requests, Selenium(second option),Time**

Procedures:

For our project, we implemented our proposed idea of consumer research on popular sedans. We started by looking at 25 sedans which are generally well-liked among consumers. We used Kelly Blue Book's *25 most Popular Sedans in 2020* article as that starting point(https://www.kbb.com/best-cars/most-popular-sedans/16/). We decided to crawl and scrape different sites such as review sites and ranking sites to gather our data. The three most relevant sites we used were https://www.carmax.com, https://www.cars.com, https://www.kbb.com. Our mission for this project was to collect and interpret this data from the web to help understand why these cars are generally well-received and what consumers value or dislike in a sedan.

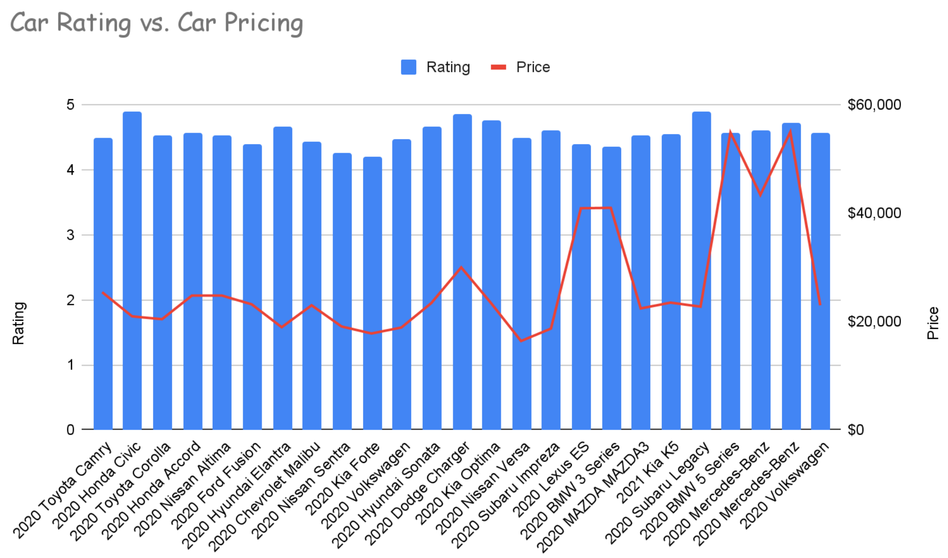
Once we had our list of cars, We organized them in a CSV to help aid, especially, in looping through the web pages while crawling. This way, we can use them to create the URL for a web page to search into. For example, with one of the review sites, Carmax, the URL for Toyota Camry was: https://www.carmax.com/reviews/toyota/camry/2020. Using our CSV we are able to run through and replace the car make, model, and year for each car we use to crawl through each site. Here is the car list final_project_cars (this can be edited at any time)

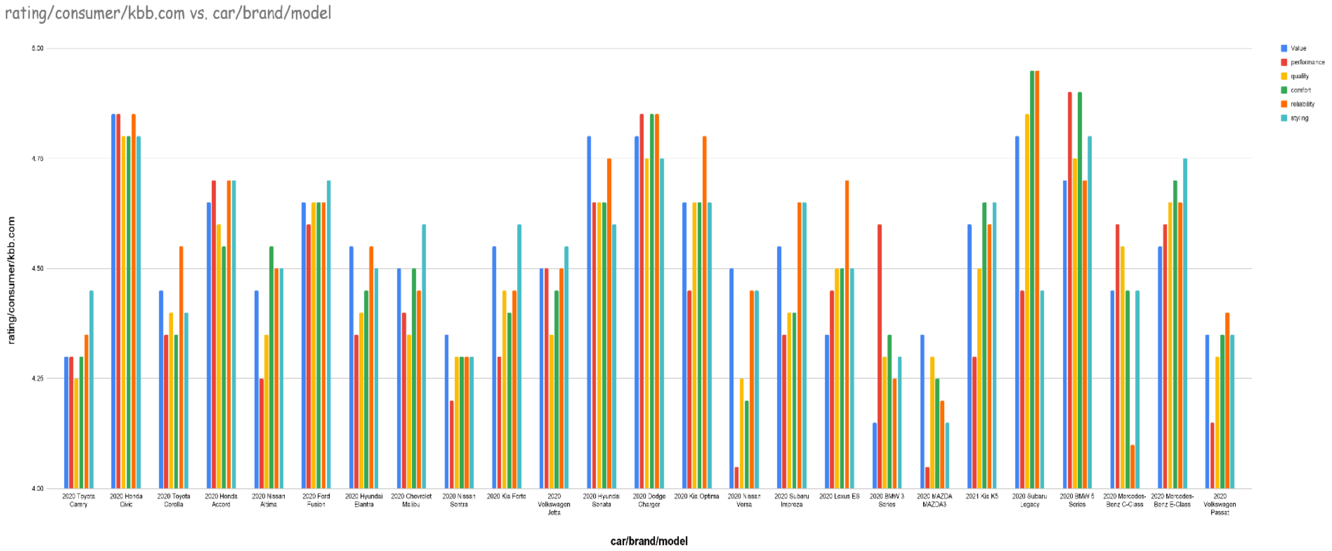
We then scraped each site for our key information. This included the general consumer ratings from each page, the specific ratings on each attribute of the car, and the reviews. For the reviews, we were able to find keywords and phrases that were associated with either a bad review or a good review and considered the pros and cons. We then organized all of this data into a CSV where we have the consumer ratings and comments for each car. We averaged the ratings from each site to help us with an overall opinion. This makes it a simple process to visualize the data and draw conclusions from it.

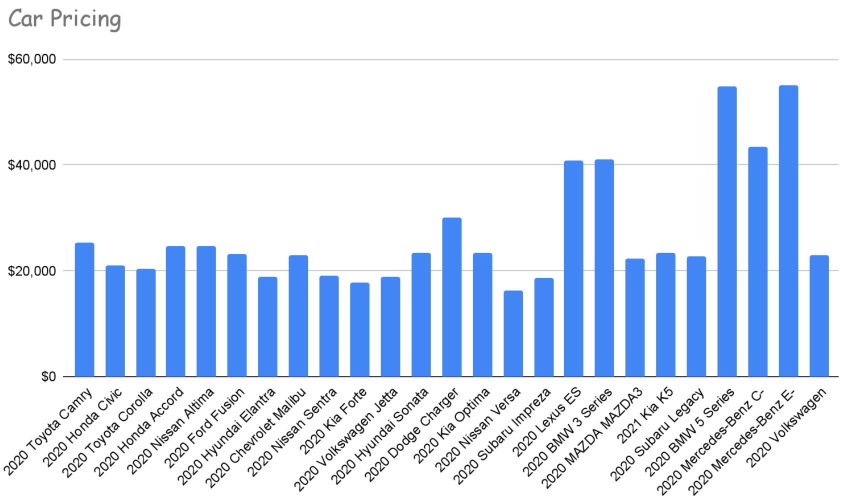
A brief overview of our python code(submitted with this report):

- We read the CSV file containing the list of cars with want to scrape.
- We attach the make, model, and year to the car retailer sites to form a specific car model search and review URL
- We then requested/selenium drove the HTML contents of that page, parsing it with BeautifulSoup, and finding specific values, strings that were needed for car reviews analysis.
- We stored those values in a pandas data frame and then wrote them to another CSV file.

Our Data:




rating/consumer/kbb.com vs. car/brand/model
5.00
4.75
4.50
4.25
4.00
rating/consumer/kbb.com
2020 Toyota Camry
2020 Honda Civic
2020 Toyota Corolla
2020 Honda Accord
2020 Nissan Altima
2020 Ford Fusion
2020 Hyundai Elantra
2020 Chevrolet Malibu
2020 Nissan Sentra
2020 Kia Forte
2020 Volkswagen Jetta
2020 Hyundai Sonata
2020 Dodge Charger
2020 Kia Optima
2020 Nissan Versa
2020 Subaru Impreza
2020 Lexus ES
2020 BMW 3 Series
2020 MAZDA MAZDA3
2021 Kia K5
2020 Subaru Legacy
2020 BMW 5 Series
2020 Mercedes-Benz C-Class
2020 Mercedes-Benz E-Class
2020 Volkswagen Passat
car/brand/model
Value
performance
quality
comfort
reliability
styling


Car Pricing
$60,000
$40,000
$20,000
$0
2020 Toyota Camry
2020 Honda Civic
2020 Toyota Corolla
2020 Honda Accord
2020 Nissan Altima
2020 Ford Fusion
2020 Hyundai Elantra
2020 Chevrolet Malibu
2020 Nissan Sentra
2020 Kia Forte
2020 Volkswagen Jetta
2020 Hyundai Sonata
2020 Dodge Charger
2020 Kia Optima
2020 Nissan Versa
2020 Subaru Impreza
2020 Lexus ES
2020 BMW 3 Series
2020 MAZDA MAZDA3
2021 Kia K5
2020 Subaru Legacy
2020 BMW 5 Series
2020 Mercedes-Benz C-
2020 Mercedes-Benz E-
2020 Volkswagen

cars

Conclusion:

Based on the data we collected, we were able to create charts of important points. Firstly, we created a chart of the pricing which was an easy way to distinguish the more expensive and less expensive, when looking at a budget. Then, we created a new graph where we could see the rating alongside the pricing to see if there was a clear correlation between the two. When looking at the overall rating, we drew the conclusion that the price and the rating had no correlation. We hypothesized the reason for this to be, as consumers write reviews, they consciously take the price into account. Our last graph shows the “individual attribute ratings”. These are ratings based on specific attributes of the car which include: value, performance, quality, comfort, reliability, and styling. We averaged these values from different sites to give ourselves a good estimate. The highest overall rating is tied at 4.9 out of 5 with the 2020 Honda Civic and the 2020 Subaru Legacy. The best value is the 2020 Honda Civic as well, with a 4.85 out of 5. So we would consider the budget-friendly option to be the Honda Civic along with the Nissan Versa, with a good rating to price ratio. The car with the best performance is the BMW 5 Series with a 4.9 out of 5. The car with the best quality is the Subaru Legacy with a 4.85 out of 5. The car with the most comfort is the Subaru Legacy, with a 4.95 out of 5. The Subaru Legacy also wins the reliability rating at 4.95 out of 5 stars. The BMW 5 Series wins the styling category along with the Honda Civic at 4.8 out of 5 stars. So our pick for the sporty car of the bunch is the BMW 5 Series.

Some findings:

1. The top 25 sedans by kbb.com have prices that are user-friendly(lower than $60,000)
2. The overall rating is not affected by the price, it is more related to specific keyword ratings

3. The specific keyword ratings are correlated with price because some of them include "value", "quality", 'performance", which correspond to cost-efficiency.
4. "Styling" and "comfort" seem to be the most important reason for people to buy sedans.

Reference

11. https://towardsdatascience.com/data-scraping-how-to-leverage-it-for-your-ecommerce-business-7320e8f82358
12. https://www.crawlnow.com/blog/7-ways-web-scraping-helps-your-ecommerce-business
13. https://medium.com/analytics-vidhya/web-scraping-e-commerce-sites-using-selenium-python-55fd980fe2fc
14. https://www.freecodecamp.org/news/scraping-ecommerce-website-with-python/