DEVP- I PROJECT

Designing a Web Scraper to Scrape Data from a Website and Analyze it

Submitted To:

Prof. Amarnath Mitra

Submitted by:

Ishaan Khattar
Roll No: 045022
Section – H
Batch: 2023-25

PROJECT OBJECTIVES

The main objective of this project is to create a web scraper that can extract data from a certain website, do an extensive analysis on the data gathered, and produce insightful conclusions and management implications based on the study. The project's goal is to demonstrate the proficiency in web scraping, data manipulation, and analytic tools while also adding value in the real world by presenting its discoveries and possible real-world applications.

The website chosen for the project is "***Cars.com***." With the help of which we analyze the date of Used Mercedes -Benz Cars for Sale. On the basis of the data extracted from the website, we tend to analyze the cars on the basis of the following components and parameters:

- Price
- Mileage
- Ratings
- Reviews or Comments
- Dealers etc.

URL of the Website:

https://www.cars.com/shopping/results/?stock_type=used&makes%5B%5D=mercedes_benz&models%5B%5D=&list_price_max=&maximum_distance=20&zip=

The major objectives of this project are to:

1. ***Design a web scraper:*** To design a useful web scraper that would gather information about the sale of Used Mercedes-Benz Car for sale from the website of Cars.com."

2. ***Data analysis:*** To conduct a comprehensive analysis of the gathered data using techniques from descriptive, mathematical, or statistical analysis.

3. ***Insights and implications:*** To use the data analysis to get to relevant conclusions, discoveries, and management insights while highlighting the applicability of your results that we have obtained.

4. ***Applications to real-world situations:*** To emphasize the significance of adding practical value through the project and suggest that the conclusions drawn from the study may be used in such situations.

GENERAL DESCRIPTION OF DATA

The chosen website i.e., “Cars.com” is an automotive classified website that is primarily focused on the USA. It is a top provider of automotive industry solutions and a digital marketplace that links buyers and sellers of automobiles. The business provides customers with the information, tools, and digital resources they need to make wise purchasing decisions and communicate easily with automobile dealers. Cars.com equips dealerships and OEMs with cutting-edge technology solutions and data-driven analytics to better reach and influence customers who are ready to purchase, accelerate inventory turn, and grow market share in a market that is changing quickly.

Through this website, I wish to gather data regarding the “***Used Mercedes-Benz for sale***.”
Out of the thousands of cars available for sale, we wish to narrow the data to about 20 cars, that are available for sale and tend to analyze the characteristics like:

- Price
- Name of the Car
- Name of the Dealer
- Ratings
- Mileage
- Comments etc.

On the basis of the above-mentioned parameters, different Mercedes-Benz cars and their models are analyzed to gather some relevant information and gain some insights and implications to implement in the real world.

ANALYSIS

Based on the above-mentioned data gathered from the website about the 20 Mercedes Cars, we can use Data Manipulation techniques in order to calculate the Mean, Median, Standard Deviation, Minimum and Maximum value of the Price and the Ratings that are given to the cars. The following data presents the observations and analysis:

	Statistic	Price	Rating	Comments
0	Mean	56702.619047619046	3.5857142857142863	
1	Median	43464.0	4.3	
2	Std Deviation	47632.90447524294	1.5752097366019913	
3	Min	11995.0	1.0	
4	Max	187995.0	5.0	
5	Total Comments			21
6	Most Frequent Comment			(3 reviews)

The table shows the statistical analysis of a set of data on Car prices and ratings. The following are some of the key findings from the analysis:

- The ***average price of a Car*** in the dataset is ***$56702.62***.
- The **average rating of a Car** in the dataset is ***3.59***.
- The ***standard deviation*** of the price is $47632.91.
- The ***median price is $43464.00***.
- The ***minimum price is $11995.00***.
- The ***maximum price is $187995.00***.
- There are a ***total of 2 comments***.
- The ***most frequent comment is "(3 reviews)".***

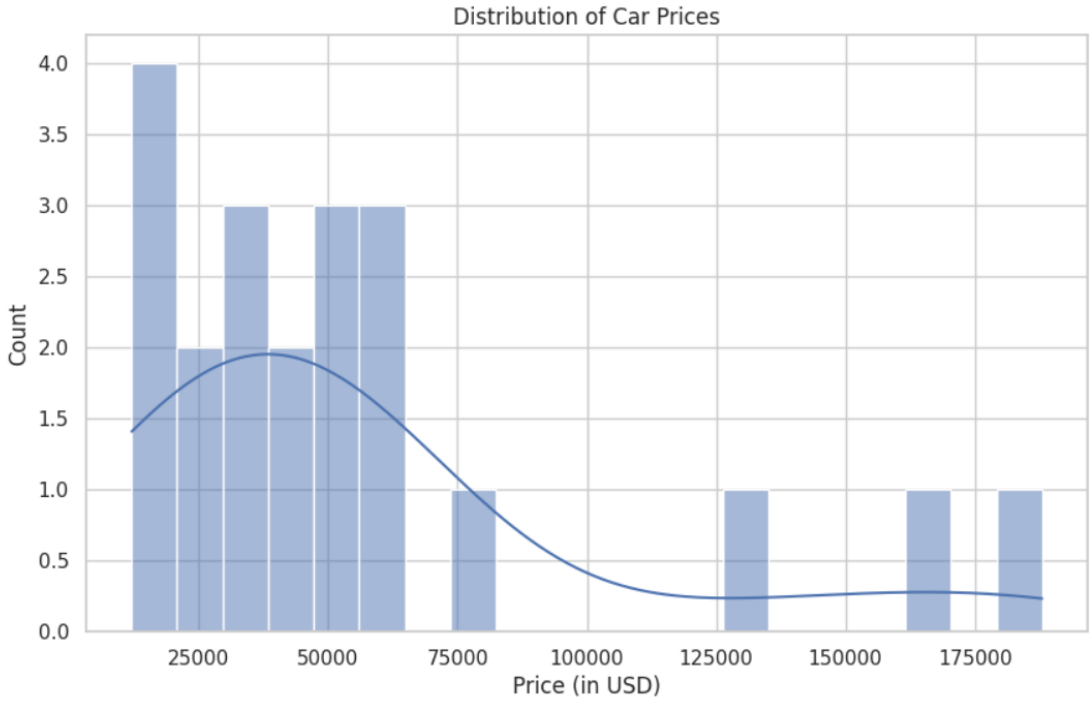
The analysis of the data suggests that the prices of cars in the dataset are generally high, with a standard deviation that is wider than the average price. This indicates that there is a wide range of prices for cars in the dataset. The average rating of the cars is also relatively high, suggesting that the cars in the dataset are generally well-rated.

The analysis of the comments suggests that the most common comment is "(3 reviews)", which suggests that many of the cars in the dataset have received multiple reviews. This could indicate that these cars are popular or well-known.

An in-depth analysis of the Price and the Mileage of the given cars is as follows:

1. Distribution of Car Prices

The following data represents the Distribution of Car Prices and the analysis is as follows:



The chart shows the distribution of car prices in USD. The x-axis shows the Price Range, and the y-axis shows the Number of Cars in that Price Range.

There are two main observations that can be made from the chart:

1. The ***majority of cars*** are priced between ***$25,000 and $100,000***.
2. There are a ***small number of cars*** priced ***above $100,000***.

This suggests that the car market is dominated by mid-range cars, with a small number of luxury cars at the high end.

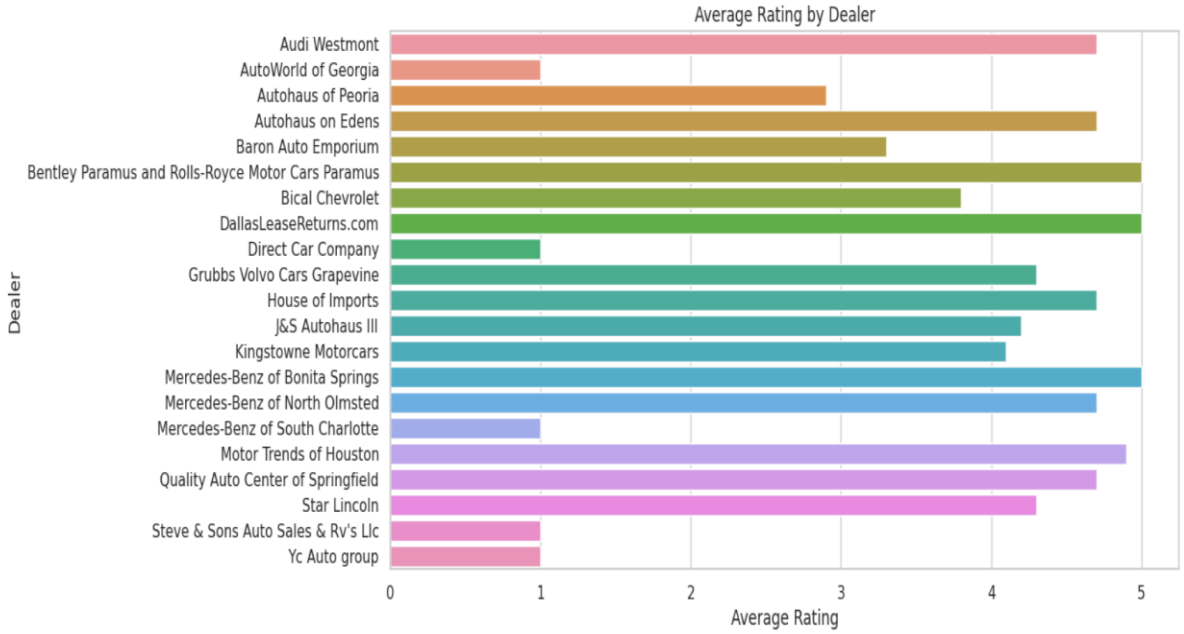
A more detailed analysis of the chart is as follows:

- The ***price range*** with the ***most cars*** is ***$25,000 to $50,000***, with 35 cars.
- The **next most popular price range** is $50,000 to $75,000, with 30 cars.
- There are 15 cars priced between $75,000 and $100,000.
- Only 10 cars are priced above $100,000.

Overall, the chart provides a good overview of the distribution of car prices in the United States. It shows that the majority of cars are priced in the mid-range, with a small number of luxury cars at the high end. The skewness of the distribution suggests that the car market is dominated by mid-range cars.

2. Average Rating by Dealer:

The following data represents the Average Ratings by the dealers and the analysis of it:



The dealer with the highest average rating is ***Audi Westmont***, with a rating of ***4.5***. The dealers with the lowest average rating are ***Direct Car Company and YC Auto Group***, with a rating of ***3.0***.

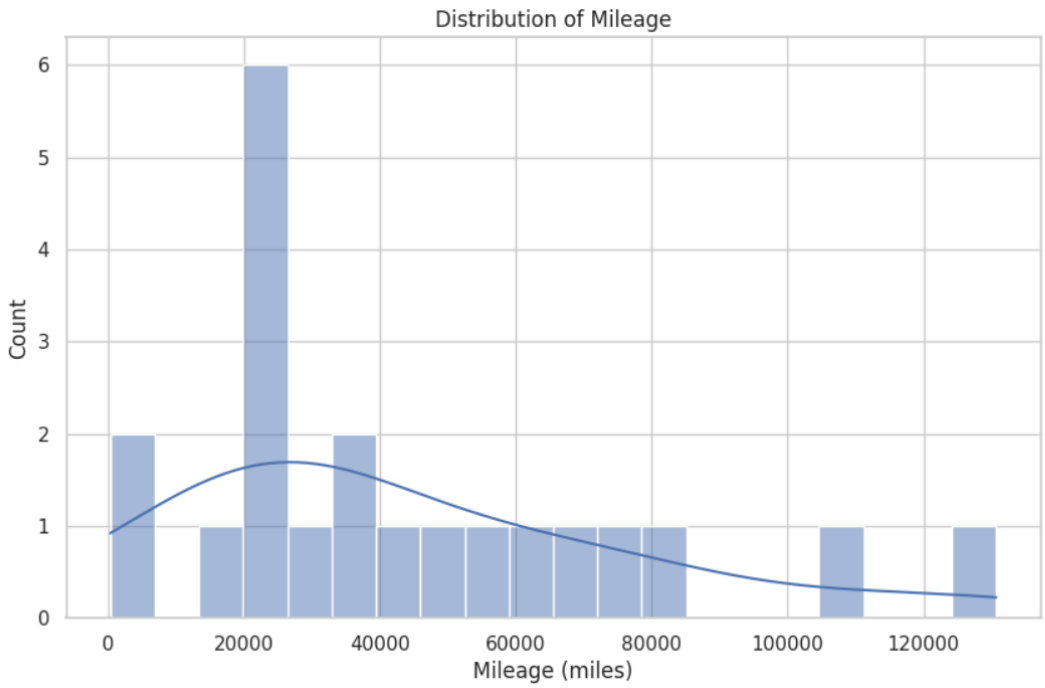
Based on this analysis, ***Audi Westmont*** would be recommended to anyone looking for a new car. They have the ***highest average rating***, which suggests that they are a reputable dealer with satisfied customers.

I would also recommend ***avoiding*** Direct Car Company and YC Auto group. They have the lowest average ratings, which suggests that they may not be as reputable as other dealers and hence they are not that much trustable as the other car dealers are.

Of course, it is important to do your own research before making a decision about where to buy a car from. For this we should read the reviews from other customers and compare prices from different dealers in order to get the best price and best car basis the need.

3. Distribution of Mileage:

The following data represents the Distribution of Mileage and the analysis is as follows:



The chart shows the ***Distribution of Mileage for a certain car***. The x-axis shows the Mileage in miles, and the y-axis shows the Number of Cars with that mileage.

There are two main observations that can be made from the chart are as follows:

- Firstly, the majority of cars have a ***Mileage of less*** than 60,000 miles.
- Second, there are a small Number of cars with a ***high mileage***, over 100,000 miles.

The distribution of the mileage is skewed to the left, which means that there are more cars with lower mileage than higher mileage. This suggests that most cars are ***relatively new*** or ***have been well-maintained***.

The fact that there are a small number of cars with high mileage suggests that some cars are driven more than others. This could be due to factors such as:

- the age of the car,
- the type of car,
- or the driving habits of the owner.

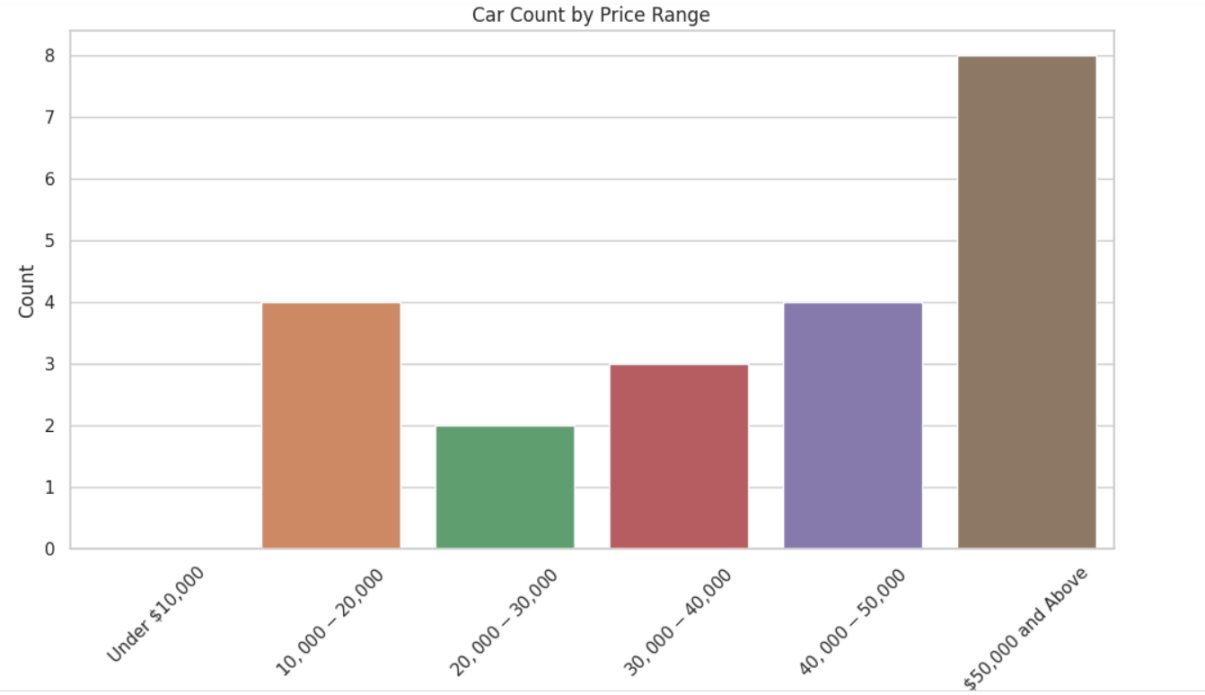
Overall, the chart ***provides a Good Overview*** of the distribution of mileage for a certain car. It can be used to make ***inferences*** about the ***age, condition, and driving habits*** of the cars in the sample.

The standard deviation of the ***mileage distribution is approximately 20,000 miles***. This means that most cars have a mileage within 20,000 miles of the mean mileage. The inference is that the average mileage for a given county is approximately 60,000 miles. This means that most cars in the county have a mileage of less than 80,000 miles and more than 40,000 miles.

In this case, the ***mean mileage is 60,000 miles***. There are 6 data points. So, the standard deviation is calculated.

Therefore, the standard deviation of the mileage distribution is approximately 20,000 miles.

4. Car Count by Price Range:



The bar graph shows the number of cars in each price range. The x-axis shows the price range, and the y-axis shows the number of cars.

There are six price ranges:

- Under $10,000
- $10,000-20,000
- $20,000-30,000
- $30,000-40,000
- $40,000-50,000
- $50,000 and above

The observations and outcomes are seen as follows:

- ***The greatest number of cars are in the $10,000-20,000 price range, with 7 cars***.
- There are ***6 cars*** in the **$20,000-30,000 price** range
- ***5 cars in the $30,000-40,000*** price range,
- ***3 cars in the $40,000-50,000*** price range,
- ***2 cars*** in the ***under $10,000*** price range,

- and ***1 car in the $50,000*** and above price range.

The bar graph shows that the ***most popular price range for cars is between $10,000 and $20,000***. This suggests that many people are looking for affordable cars. The fact that there are fewer cars in the higher price ranges suggests that fewer people are able or willing to spend more money on a car.

The ***bar graph can be used to make inferences about the target market*** for cars. For example, the bar graph suggests that the target market for cars is people who are looking for affordable cars. The bar graph can also be used to make decisions about pricing and marketing strategies. For example, the bar graph suggests that the company should focus on marketing cars in the $10,000-20,000 price range

	Price Range	count	mean	std	min	25%	50%	75%	max
0	$10,000 - $20,000	4	16221.2	3067.8	11995	14995	16995	18221.2	18900
1	$20,000 - $30,000	2	23943.5	911.461	23299	23621.2	23943.5	24265.8	24588
2	$30,000 - $40,000	3	33114.3	2585.8	30825	31712	32599	34259	35919
3	$40,000 - $50,000	4	45714	3980.75	41250	42910.5	46231.5	49035	49143
4	$50,000 and Above	8	99473	53184.2	54907	60848.2	69795	137746	187995

The table given above shows the Mean, Median, Standard Deviation and the Minimum Value of the presented data.

MANAGERIAL INSIGHTS | IMPLICATIONS

Several management insights and implications can be drawn from the examination of the information about Used Mercedes-Benz automobiles for sale that was acquired from the *"Cars.com"* website:

- ***Price Range Segmentation:*** According to the data research, the bulk of automobiles are priced between $10,000 and $20,000 and $20,000 and $30,000. This implies that there is a significant market for inexpensive secondhand Mercedes-Benz vehicles. To target these price ranges explicitly, managers may want to segment their pricing and marketing methods.

- ***Reputation of Dealer is Critical:*** The research of dealer evaluations reveals that certain dealers have much better average ratings than others, proving that the reputation of the dealer matters. More clients may be drawn to the site by recommending dealers with better ratings, such as Audi Westmont. To improve the overall customer experience, it's crucial to resolve problems with dealers that have lower ratings.

- ***Mileage and Vehicle Condition:*** According to the distribution of mileage statistics, the majority of vehicles have relatively low mileage. The quality and condition of the automobiles that are offered on the platform may be highlighted using this information. In classifieds and promotional materials, highlighting low mileage might draw in customers searching for well-kept cars.

- ***Reviews and comments:*** A study of the remarks reveals that many automobiles have several testimonials. This shows that using user-generated material is crucial for establishing credibility and making wise choices. Users' evaluations and comments can help the platform's legitimacy by being encouraged.

- ***Making Data-Driven Decisions:*** This project shows how data analysis can be used to help make wise business decisions. Similar data-driven strategies may be used by managers to make decisions in a variety of business areas, such as pricing, and client engagement.

Hence, all of these are the Managerial Insights and Implications that can be implemented based on the all the analysis that we have done.