SCRAPING DATA FROM GOOGLE MAPS

[1] MR.Qamer Ahmed,[2] S. Akash Reddy[3] Y Manoj Kumar,[4] Jyothi Swaroopachari [5]V. Manoj Kumar, [6] R. Srikanth
[1]Assistant Professor, Department of DS, Sri Indu College Of Engineering & Technology.
[2,3,4,5,6] U.G.Scholor, Department of DS, Sri Indu College Of Engineering & Technology, Hyderabad

ABSTRACT

Data mining is an important step in the data science lifecycle, as it helps in extracting useful information from large datasets. In this research, Selenium, an open-source web automation tool, is used to collect data from Google Maps. The extracted data is then processed and converted into structured format using Python for further analysis. After preprocessing, the data is visualized using Tableau to present meaningful insights. For this study, the Mumbai Western Express Highway is selected as the case study. The results of the project help identify the most congested locations along the highway in both directions, analyze traffic patterns over time, and estimate the average travel time during different periods such as morning, evening, and night.

Keywords: Selenium, structured data, Python, Tableau.

I. INTRODUCTION

Over the years Google Maps have amassed billions of bytes of facts to map the complete world. With over 1 Billion plus month-to-month lively users, Google Maps is one of the popularly used client application. The most captivating feature is the real-time traffic information. This crucial information contains the time and distance to travel from one place to another, also on the map we can see various traffic zone coloured areas marked in green, yellow, and red denoting traffic intensity.

This available information is in an unstructured format that needs to be captured using Web scraping. Web scraping, also known as web harvesting is a method that can be used to capture and extract a large amount of data from a website and store it in a structured format such as an Excel sheet or a CSV file. Next, This Structure data can be easily manipulated using Python and Tableau for data mining. Data mining or knowledge discovery can be used to analyze hidden patterns of data into meaningful information.

II. METHODOLOGY

For this research, I have considered a 30km stretch in the western express highway with 30 points separated at a distance of 1km. Now there are 30 points from stretch A to B similarly, for the opposite direction from B to A there will be another 30 points. Hence, a total of 60 points will be analyzed in one loop.

	map	index	type	start_x	start_y	end_x	end_y	avg_x	avg_y	avg_main_x	avg_main_y	place
0	https://www.google.co.in/maps/dir/19.2849189,7...	1	fountain to bandra	19.284919	72.903945	19.286280	72.894924	19.285600	72.899434	19.285235	72.899931	varsova road ghodbunder
1	https://www.google.co.in/maps/dir/19.2862802,7...	2	fountain to bandra	19.286280	72.894924	19.280436	72.891395	19.283358	72.893160	19.284188	72.892105	varsova road
2	https://www.google.co.in/maps/dir/19.2804362,7...	3	fountain to bandra	19.280436	72.891395	19.273883	72.885305	19.277160	72.888350	19.276076	72.888680	varsova road sai palace
3	https://www.google.co.in/maps/dir/19.2738833,7...	4	fountain to bandra	19.273883	72.885305	19.268168	72.877863	19.271026	72.881584	19.272517	72.883146	kashimira flyover miraroad
4	https://www.google.co.in/maps/dir/19.2681678,7...	5	fountain to bandra	19.268168	72.877863	19.260761	72.872963	19.264465	72.875413	19.264213	72.874716	thakur mall

Figure 1: Top 5 elements of the dataset showing the link to google map and coordinates of the place

With the help of Selenium, we can open the links one by one and extract the necessary data which needs to be scrapped. As seen in figure 1 there is a column called map in which contains a link to Google Maps.

As shown in Figure 2, a link will be opened for each place and the code will scrap important data such as distance, time to travel, colour code which may be red, green, or yellow depending upon on traffic. I have used the DateTime module to get the current time and epoch value. Moreover, with the help of python, a column for the time zone is created showing morning, noon, night, etc.

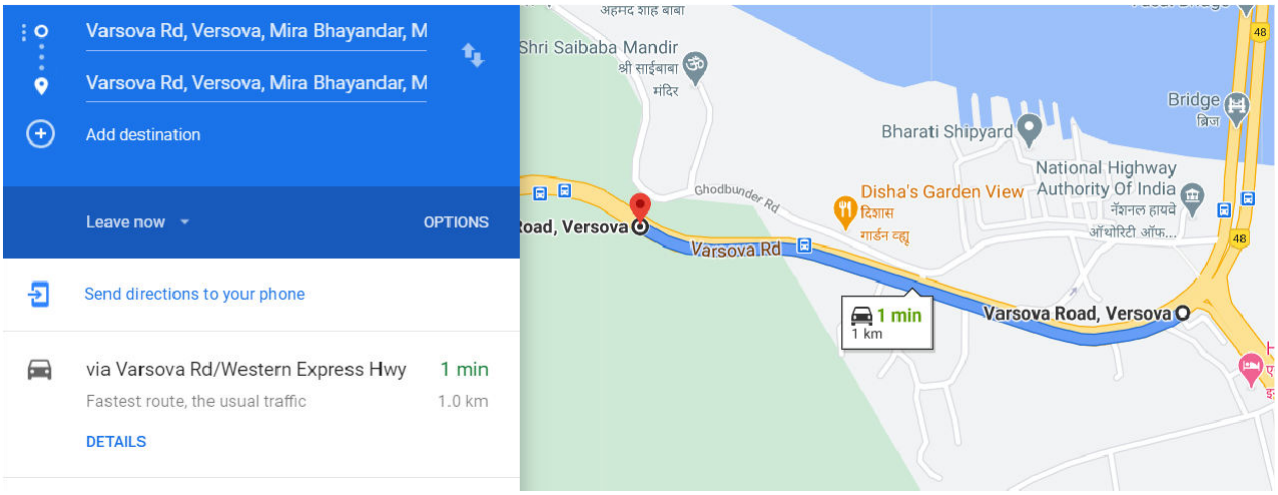

Figure 2: Map showing distance, time to travel, colour code ie green

Therefore, for the total of 60 points, the above data will be extracted and will be converted to a data frame using the Pandas library in python which can be saved to a CSV file. In this research, I have collected data for 24 hours straight by running the code in a recursive loop. As seen in Figure 3 we get the extracted data which I have analyzed using Tableau which is a Business Intelligence tool for visually analyzing the data. Users can create and distribute an interactive and shareable dashboard, which depicts the trends, variations, and density of the data in the form of graphs and charts. The data contains about 8160 rows and 18 columns.

map	index	type	start_x	start_y	end_x	end_y	avg_x	avg_y	travel_time	distance	traffic	date	epoch	avg_main_x	avg_main_y	place	time_zone
https://ww	1	fountain to	19.284919	72.903945	19.28628	72.894924	19.2856	72.899435	2	1	green	04:05:2021	1.618E+09	19.285235	72.899931	varsova roa	Noon
https://ww	2	fountain to	19.28628	72.894924	19.280436	72.891395	19.283358	72.89316	1	1	green	04:05:2021	1.618E+09	19.284188	72.892105	varsova roa	Noon
https://ww	3	fountain to	19.280436	72.891395	19.273883	72.885305	19.27716	72.88835	2	1	yellow	04:05:2021	1.618E+09	19.276076	72.88868	varsova roa	Noon
https://ww	4	fountain to	19.273883	72.885305	19.268168	72.877863	19.271026	72.881584	1	1	green	04:05:2021	1.618E+09	19.272517	72.883146	kashimira fl	Noon
https://ww	5	fountain to	19.268168	72.877863	19.260761	72.872963	19.264465	72.875413	2	1	red	04:05:2021	1.618E+09	19.264213	72.874716	thakur mall	Noon
https://ww	6	fountain to	19.260761	72.872963	19.252945	72.868548	19.256853	72.870756	5	1	red	04:05:2021	1.618E+09	19.257176	72.87127	toll naka da	Noon
https://ww	7	fountain to	19.252945	72.868548	19.244951	72.864676	19.248948	72.866612	1	1	green	04:05:2021	1.618E+09	19.250105	72.866513	dahisar east	Noon
https://ww	8	fountain to	19.244951	72.864676	19.236049	72.86336	19.2405	72.864018	1	1	green	04:05:2021	1.618E+09	19.241244	72.864065	borivali east	Noon
https://ww	9	fountain to	19.236049	72.86336	19.2271	72.863469	19.231574	72.863415	1	1	green	04:05:2021	1.618E+09	19.229887	72.863363	national par	Noon
https://ww	10	fountain to	19.2271	72.863469	19.218393	72.866495	19.222746	72.864982	1	1	green	04:05:2021	1.618E+09	19.223028	72.864824	magathane	Noon
https://ww	11	fountain to	19.218393	72.866495	19.209283	72.868239	19.213838	72.867367	1	1	green	04:05:2021	1.618E+09	19.214261	72.867872	thakur villag	Noon
https://ww	12	fountain to	19.209283	72.868239	19.202382	72.862241	19.205833	72.86524	1	1	green	04:05:2021	1.618E+09	19.205197	72.864915	mahindra ar	Noon

Figure 3: CSV file created after extracting data

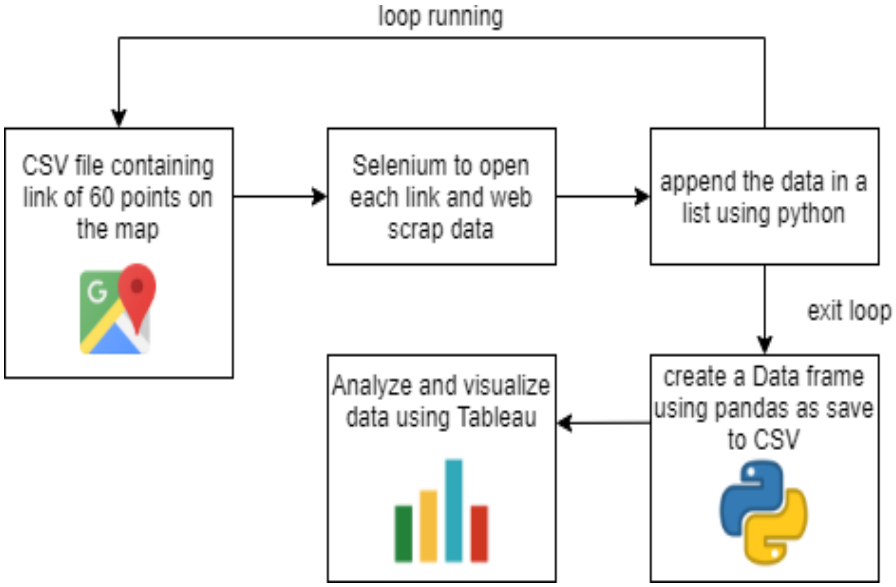

Figure 4: Flowchart of the project

III. RESULTS AND DISCUSSION

Figure 5: Dashboard created on Tableau

Figure 6: Dashboard displaying chart for a specific location when selected

As seen in the above figures 5 and 6 a dynamic dashboard was made using tableau. For creating this dashboard separate individual worksheets were first made and then combined.

For instance, in figure 5 we can see the visualization for the evening time zone which is a high-level detail. The dashboard displays the average time for every 30 points from road A to B and vice versa in the evening, various traffic zones displayed in the colour red, yellow and green. Also, the dashboard shows the avg, min, and max time required to travel, along with the area chart. With the help of longitude and latitude points, we can plot a map chart in tableau.

Next, by using actions I have created interactivity in the dashboard which allows navigation between high-level details to low-level details for more analysis. For instance, in figure 6 by selecting a particular location which in this case is the Rani Sati flyover, we can see the charts focused on that place only ie the low-level details.

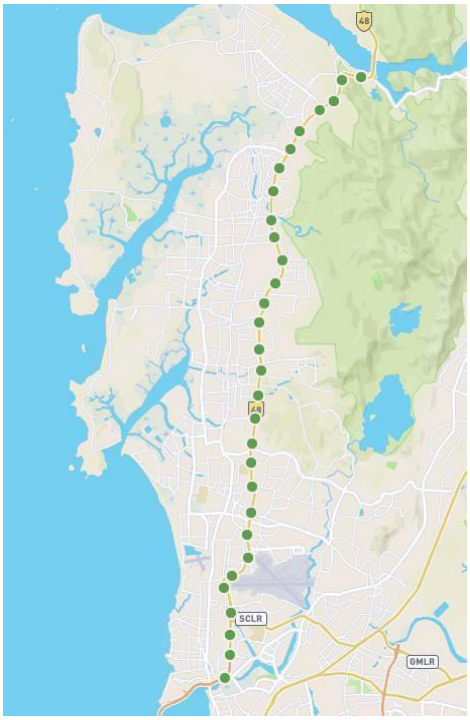

Figure 7: 30 points displayed on the map

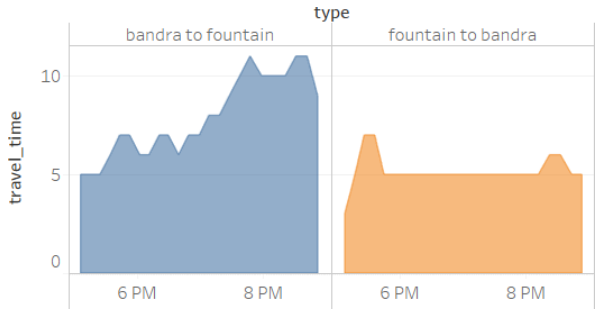

Figure 8: Area chart for Dahisar toll naka during the evening

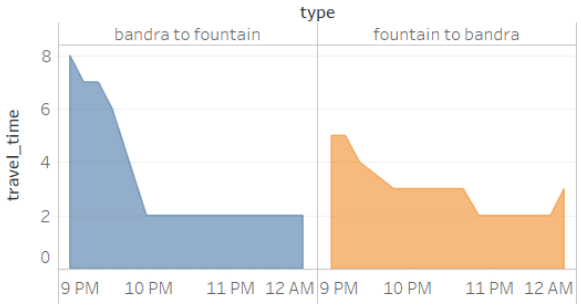

Figure 9: Area chart for Dahisar toll naka during the night

Table 1: Time in minutes for Bandra to Fountain

	whole day	early morning	morning	noon	Eve	night	late night
average	44.29	33.15	38.12	44.31	79.02	39.47	32.33
min	31.00	31.00	35.00	36.00	50.00	31.00	31.00
max	130.00	43.00	48.00	70.00	126.00	68.00	39.00

Table 2: Time in minutes for Fountain to Bandra

	whole day	early morning	morning	noon	Eve	night	late night
average	38.77	34.55	42.35	40.74	46.57	33.87	33.72
min	31.00	31.00	33.00	35.00	35.00	31.00	31.00
max	70.00	50.00	60.00	52.00	62.00	45.00	44.00

The above table 1 and 2 contains the average, minimum and maximum time across various time zone for both the directions. As seen in Table 1 we can see that during the evening the average time is about 79 minutes, the reasons are explained below.

- **Rush hour:** Many people travel from suburban areas of Mumbai for work to the main city and while return there is a traffic delay on the western express highway.

- **Metro Construction:** The majority of the metro line which runs on the western express highway is built on the Bandra to Fountain side. Due, to incomplete construction certain lanes, are blocked with barricades which can cause traffic delays. However, this project plans to reduce congestion on the road once completed.
- **Toll plaza:** It is clear from Figures 5 and 6 that the highest traffic is received on Dahisar toll naka as vehicles tend to slow down near this area. As seen in Figures 7 and 8 the travel time increases during evening rush hour and then the slope falls down at night for Bandra to Fountain route on the highway.
- **More increase in private vehicles:** Mumbai and its surrounding regions have added 9.9% of vehicles on the road in 2019 as per the Economic Survey of Maharashtra 2018-19.

IV. CONCLUSION

Data mining is the process useful for the discovery of informative and analysis of a raw collection of stats called data. In this Research project, I have successfully accumulated data from an unstructured format to a structured one with the help of a web scraping technique using selenium and python. Next using tableau, we can study the lower-level details in-depth with the help of visualizations. In this research, I found that on the entire route of 30kms the most congested traffic zone areas are Dahisar toll naka, Times of India Malad, Rani sati flyover, and Jogeshwari east. Various stats regarding the time it takes to travel in both the direction for various time zones were discussed. Moreover, we can interact with the data on tableau allowing for better understanding.

V. FUTURE SCOPE

Dataset can be further scraped from Google maps to collect weekly information about the traffic which can be used to create a supervised machine learning model. Supervised learning is when the model is getting trained on a labeled dataset. The labeled dataset is one that has both input and output parameters. Through this project, we can try to predict future traffic trends for various locations based on certain factors.

VI. REFERENCES

[1] D. M. Thomas and S. Mathur, "Data Analysis by Web Scraping using Python," 2019 3rd International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2019, pp. 450-454, DOI: 10.1109/ICECA.2019.8822022.

[2] R. Diouf, E. N. Sarr, O. Sall, B. Birregah, M. Bousso and S. N. Mbaye, "Web Scraping: State-of-the-Art and Areas of Application," 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 2019, pp. 6040-6042, DOI: 10.1109/BigData47090.2019.9005594.

[3] Akhtar, Nikhat & Tabassum, Nazia & Perwej, Dr.Asif & Perwej, Dr. Yusuf. (2020). Data analytics and visualization using Tableau utilitarian for COVID-19 (Coronavirus). Global Journal of Engineering and Technology Advances. Volume 3. Page 28-50. 10.30574/gjeta.2020.3.2.0029.

[4] Wu, Jiahao. (2019). Web Scraping Using Python: A Step By Step Guide.