

Project Title	Redbus Data Scraping with Selenium & Dynamic Filtering using Streamlit
Skills take away From This Project	Web Scraping using Selenium, Python, Streamlit , SQL
Domain	Transportation

Problem Statement:

The "Redbus Data Scraping and Filtering with Streamlit Application" aims to revolutionize the transportation industry by providing a comprehensive solution for collecting, analyzing, and visualizing bus travel data. By utilizing Selenium for web scraping, this project automates the extraction of detailed information from Redbus, including bus routes, schedules, prices, and seat availability. By streamlining data collection and providing powerful tools for data-driven decision-making, this project can significantly improve operational efficiency and strategic planning in the transportation industry.

Business Use Cases:

The solution can be applied to various business scenarios including:

- **Travel Aggregators:** Providing real-time bus schedules and seat availability for customers.
- **Market Analysis:** Analyzing travel patterns and preferences for market research.
- **Customer Service:** Enhancing user experience by offering customized travel options based on data insights.
- **Competitor Analysis:** Comparing pricing and service levels with competitors.

Approach:

1. **Data Scraping:**
 - Use Selenium to automate the extraction of Redbus data including routes, schedules, prices, and seat availability.
2. **Data Storage:**
 - Store the scraped data in a SQL database.
3. **Streamlit Application:**

- Develop a Streamlit application to display and filter the scraped data.
- Implement various filters such as bustype, route, price range, star rating, availability.

4. **Data Analysis/Filtering using Streamlit:**

- Use SQL queries to retrieve and filter data based on user inputs.
- Use Streamlit to allow users to interact with and filter the data through the application.

Results:

You should aim to:

- Successfully scrape a minimum of 10 Government State Bus Transport data from Redbus website using Selenium. Also include the private bus information for the selected routes.
- Store the data in a structured SQL database.
- Develop an interactive Streamlit application for data filtering.
- Ensure the application is user-friendly and efficient.

Project Evaluation Metrics:

- **Data Scraping Accuracy:** Completeness and correctness of the scraped data.
- **Database Design:** Effective and efficient database schema.
- **Application Usability:** User experience and ease of use of the Streamlit application.
- **Filter Functionality:** Effectiveness and responsiveness of data filters.
- **Code Quality:** Adherence to coding standards and best practices.

Technical Tags:

- Web Scraping
- Selenium
- Streamlit
- SQL
- Data Analysis
- Python
- Interactive Application

Data Set:

- **Source:** Data will be scraped from the Redbus website.

Link- <https://www.redbus.in/>

- **Format:** The scraped data will be stored in a SQL database.
- **Required Fields:** Bus routes Link, Bus route Name, Bus name, Bus Type(Sleeper/Seater), Departing Time, Duration, Reaching_Time, Star-rating, Price, Seat_availability.

Data Set Requirements & Explanation:

The scraped dataset for this project should contain detailed information about bus services available on Redbus, covering various aspects critical to travelers and service providers. Here is a breakdown of the fields required:

- **Bus Routes Name:** This field captures the start and end locations of each bus journey, providing crucial information about the routes serviced.
- **Bus Routes Link:** Link for all the route details.
- **Bus Name:** The name of the bus or the service provider, which helps in identifying the specific operator.
- **Bus Type (Sleeper/Seater/AC/Non-AC):** This field specifies whether the bus is a sleeper or seater type, indicating the seating arrangements and comfort level offered.
- **Departing Time:** The scheduled departure time of the bus, essential for planning travel schedules.
- **Duration:** The total duration of the journey from the departure point to the destination, helping passengers estimate travel time.
- **Reaching Time:** The expected arrival time at the destination, allowing for better planning of onward travel or activities.
- **Star Rating:** A rating provided by passengers, indicating the quality of service based on factors such as comfort, punctuality, and staff behavior.
- **Price:** The cost of the ticket for the journey, which can vary based on factors like bus type and demand.
- **Seat Availability:** The number of seats available at the time of data scraping, giving real-time insight into the occupancy levels.

Database Schema: Table: bus_routes

Primary Key: To ensure each record is unique, an auto-incrementing primary key (**id**) is used.

Column Name	Data Type	Description
id	INT	Primary Key (Auto-increment)
route_name	TEXT	Bus Route information for each state transport
route_link	TEXT	Link to the route details
busname	TEXT	Name of the bus
bustype	TEXT	Type of the bus
departing_time	TIME	Departure time
duration	TEXT	Duration of the journey
reaching_time	TIME	Arrival time
star_rating	FLOAT	Rating of the bus

price	DECIMAL	Price of the ticket
seats_available	INT	Number of seats available

Project Deliverables:

- **Source Code:** Python scripts for data scraping, SQL database interaction, and Streamlit application.
- **Documentation:** Detailed documentation explaining the code, data collection, and application usage.
- **Database Schema:** SQL scripts to create and populate the database.
- **Application using Streamlit :** Screenshots or links to the Streamlit application showing data filtering/ analysis.

Project Guidelines:

- **Coding Standards:** Follow PEP 8 guidelines for Python code.
- **Version Control:** Use Git for version control and maintain regular commits.
- **Best Practices:** Write modular, reusable code, and include comments and docstrings.

Timeline:

Check your mail for the submission deadline of the project.

REFERENCES:

1)SELENIUM WEB SCRAPING RECORDING:

[Webscraping using Selenium- Special Session](#)

[Webscraping using Selenium- Special Session\(TAMIL\)](#)

2)PROJECT APPROACH GUIDE/RECORDING:

[Redbus_Project Orientation](#)

[Redbus_Project Orientation\(Tamil\)](#)

3)STREAMLIT RECORDING:

 **Special session for STREAMLIT(11/08/2024) (ENGLISH)**

https://us06web.zoom.us/rec/share/JTr7DywhE1-SarjyIHBSCn4qnI7_uvJH6IGk06qAlkE0Ny1o_rqcq5FRFKuo93dm.iyM2o6l0h9aTUkNE (TAMIL)

4)STREAMLIT DOCUMENTATION:<https://docs.streamlit.io/get-started/installation>

5)SELENIUM DOCUMENTATION:

<https://www.selenium.dev/documentation/webdriver/elements/locators/>

6) HOW TO ESTABLISH SQL CONNECTION:

<https://drive.google.com/file/d/1ZdgP0OgpxamIKJzAPspZkrExD0JvZb-A/view?usp=sharing>

7) PROJECT FAQs- RedBus_FAQs

PROJECT DOUBT CLARIFICATION SESSION (PROJECT AND CLASS DOUBTS)

About Session: The Project Doubt Clarification Session is a helpful resource for resolving questions and concerns about projects and class topics. It provides support in understanding project requirements, addressing code issues, and clarifying class concepts. The session aims to enhance comprehension and provide guidance to overcome challenges effectively.

Note: Book the slot at least before 12:00 Pm on the same day

Timing: Tuesday, Thursday, Saturday (5:00PM to 7:00PM)

Booking link : <https://forms.gle/XC553oSbMJ2Gcfug9>

LIVE EVALUATION SESSION (CAPSTONE AND FINAL PROJECT)

About Session: After completing the project, it is mandatory to get it evaluated by the internal Tech Mentor. The Live Evaluation Session for Capstone and Final Projects allows participants to showcase their projects and receive real-time feedback for improvement. It assesses project quality and provides an opportunity for discussion and evaluation.

Note: This form will Open on Saturday and Sunday Only on Every Week

Timing: Monday-Saturday (11:30PM to 12:30PM)

Booking link : <https://forms.gle/1m2Gsro41fLtZurRA>

Evaluation Metrics :  Project Live Evaluation