RED BUS WEB SCRAPPING PROJECT

Project Overview:

The "RedBus Data Scraping and Filtering with Streamlit Application" is designed to collect, analyze, and visualize bus travel data from RedBus. The project utilizes **Selenium** for web scraping, **MySQL** for storing data and **Streamlit** for building a user-friendly interface that enables users to filter, explore, and visualize bus travel information. This application automates the extraction of crucial data, such as bus routes, schedules, prices, and seat availability, and provides an intuitive dashboard for data analysis and decision-making.

Key Features:

1. **Data Scraping**: The application uses Selenium to scrape detailed bus information from the RedBus website, including:
   - Bus Routes
   - Schedules
   - Prices
   - Seat Availability
   - Ratings and Reviews

2. **Data Filtering**:
   - Users can apply dynamic filters to explore data based on bus type (AC/Non-AC, Sleeper/Seater), route name, price range, star rating, and seat availability.
   - Filters allow fine-grained control over the data for decision-making.

3. **Data Analysis**:
   - The app allows the user to analyze the scraped data, such as finding the most popular routes, average prices, and seat availability trends.

4. **Visualization**:
   - The Streamlit application displays the data in a clean, interactive dashboard, which can include tables, charts, and graphs for better data interpretation.
5. **Database Integration**:
   - A MySQL database is used to store and manage the scraped data. The database is queried dynamically to retrieve relevant information and support the filtering functionality.

Architecture:

The architecture of the "RedBus Data Scraping and Filtering with Streamlit Application" can be done by following steps:

1. **Web Scraping**:
   - **Selenium**: A web scraping tool used to automate the extraction of bus data from the RedBus website.
   - **Scraping Process**: The script navigates through the website, extracts relevant details (routes, prices, schedules, seat availability), and stores them in the MySQL database.
2. **Backend (Database)**:
   - **MySQL Database**: Used to store scraped bus data.
   - **Tables**: Each table stores specific types of data (e.g., bus routes, seat availability, pricing).
   - **SQL Queries**: The backend dynamically queries the database based on the user-selected filters (route, bus type, price, etc.).
3. **Frontend (Streamlit)**:
   - **Streamlit Application**: Provides an interactive UI for users to filter, analyze, and visualize the data.
   - **User Interface**: A sidebar for navigation, filter controls (dropdowns, sliders), and data display in tables.

Working Process:

1. Web Scraping Process

The web scraping module uses Selenium to automate browsing and extraction of bus data from the RedBus website. The data fetched includes:

- **Bus Routes**: The name of the route, origin, destination, etc.
- **Bus Schedules**: Departure and arrival times, frequency of buses.
- **Pricing**: Ticket prices for different bus types (AC, Non-AC, Sleeper, Seater).
- **Seat Availability**: Available seats for each bus.
- **Ratings**: User ratings for buses, including stars and reviews.

Scraping Workflow:

1. Initialize Selenium WebDriver and navigate to the RedBus website.
2. Parse relevant web pages (such as bus routes and schedules) using Selenium
3. Extract data for bus details, routes, prices, schedules, etc.
4. Store the extracted data in a SQL Database.
5. Ensure proper error handling for edge cases (e.g., failed scraping, no data).

---

2. Streamlit Application: Filtering and Displaying Data

The application allows users to interact with the scraped data via a simple interface with various filters:

Sidebar Navigation

- **Home**: Where users can filter bus data.
- **About Us**: Information about the application.

Filters:

- **Bus Type**: Select from options like AC, Non-AC, Sleeper, Seater.
- **Route Name**: Select bus routes based on origin and destination.
- **Price Range**: Adjust using a slider to filter buses within a specified price range.
- **Star Rating**: Filter buses based on user ratings.
- **Seat Availability**: Filter buses that have a certain number of available seats.

Data Display:

Once the filters are applied, the filtered bus data is fetched from the database and displayed as a table in the Streamlit app. It includes:

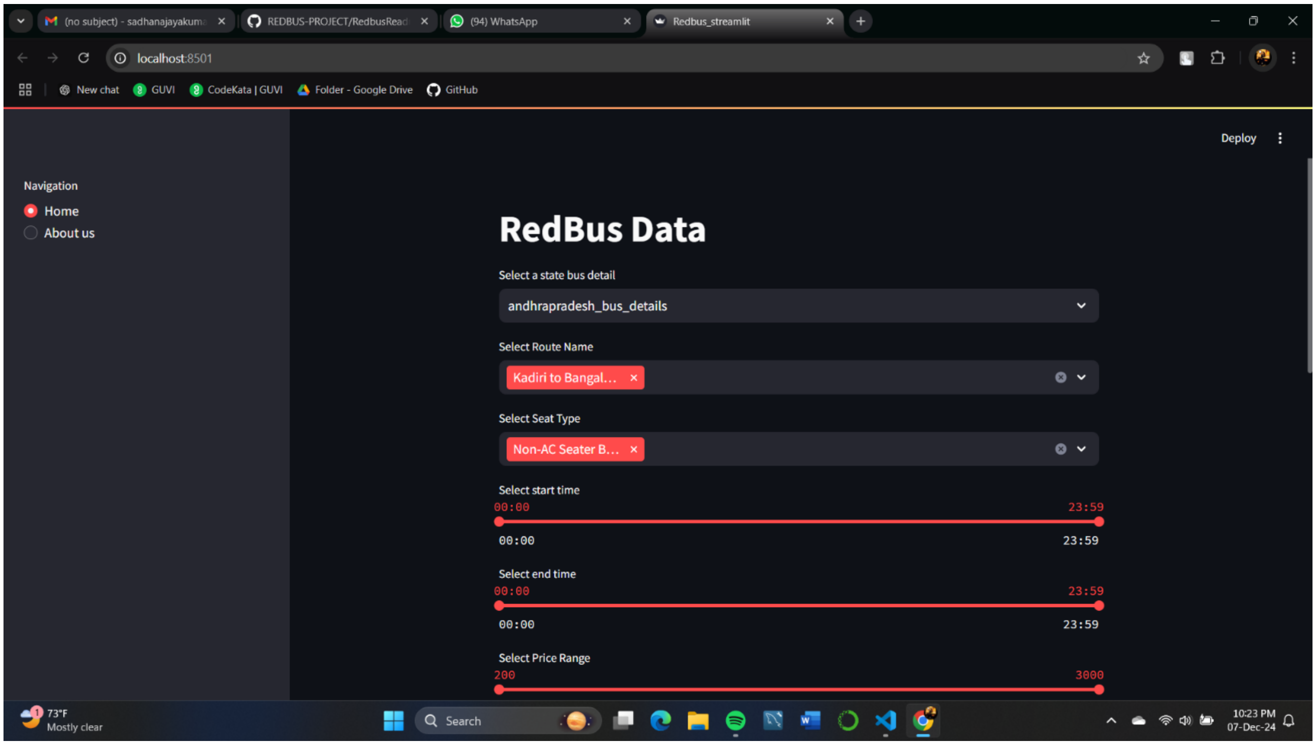
- Bus details (route, type, price, seats)
- Derived categories such as Seating Type (Seater/Sleeper) and Comfort Type (AC/Non-AC).

Database Schema:

Tables

1. **Bus_Routes**:
   - **id**: Primary key.
   - **route_name**: Name of the bus route.
   - **origin**: Starting location of the route.
   - **destination**: End location of the route.
2. **Bus_Details**:
   - **id**: Primary key.
   - **bus_type**: Type of the bus (AC/Non-AC, Sleeper/Seater).
   - **price**: Price of the ticket.
   - **departure_time**: Departure time of the bus.
   - **arrival_time**: Arrival time of the bus.
   - **rating**: Star rating (e.g., 4.5 stars).
3. **Seat_Availability**:
   - **id**: Primary key.
   - **bus_id**: Foreign key referring to Bus_Details.
   - **available_seats**: Number of seats available for the bus

OUTPUT:



Navigation

● Home

○ About us

RedBus Data

Select a state bus detail

andhrapradesh_bus_details

Select Route Name

Kadiri to Bangal... ×

Select Seat Type

Non-AC Seater B... ×

Select start time

00:00 — 23:59

00:00 23:59

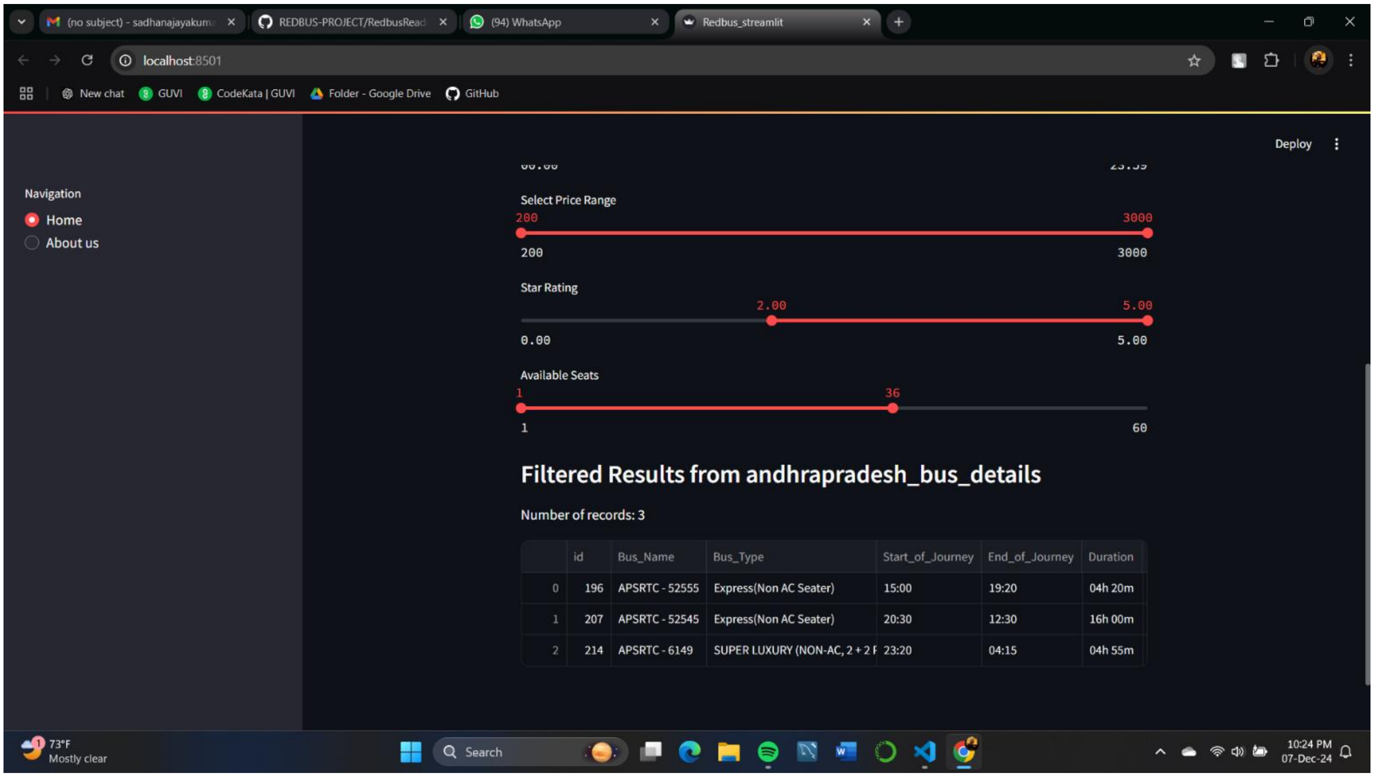
Select end time

00:00 — 23:59

00:00 23:59

Select Price Range

200 — 3000



Navigation

● Home

○ About us

Select Price Range

200 — 3000

200 3000

Star Rating

2.00 — 5.00

0.00 5.00

Available Seats

1 — 36

1 60

Filtered Results from andhrapradesh_bus_details

Number of records: 3

	id	Bus_Name	Bus_Type	Start_of_Journey	End_of_Journey	Duration
0	196	APSRTC - 52555	Express(Non AC Seater)	15:00	19:20	04h 20m
1	207	APSRTC - 52545	Express(Non AC Seater)	20:30	12:30	16h 00m
2	214	APSRTC - 6149	SUPER LUXURY (NON-AC, 2 + 2 F	23:20	04:15	04h 55m

Conclusion:

This application provides a powerful and automated way to collect and filter bus travel data from RedBus. With the combination of web scraping, database management, and an interactive Streamlit interface, it enables users to make data-driven decisions and gain insights into the bus travel market.