|  |  |
| --- | --- |
| **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** |

* **Project Description:**

This project aims at providing bus routes details of 12 Government RTC bus services available on the redbus website using

* **Data collection** : Selenium as web scraping tool using python selenium library to scrape bus information
* **Data storage** : Collect the scraped data in a csv file and load to SQL database
* **Data cleansing** : Clean the data to filter out relevant information using pandaSQL Library
* **Data Visualization Tool**: Streamlit for displaying scraped on web app with filters to provide
* Easy and better user experience with respective RTC option menu to filter out bus routes.
* Bus routes as per convenient bus schedules along with travel duration.
* Provide better seat availability search.
* Bus services based on star rating.
* Bus schedules to determine customer needs based on starting and reaching time.
* Bus fare range to choose price as per customer requirements.
* **Data collection using Python selenium library**

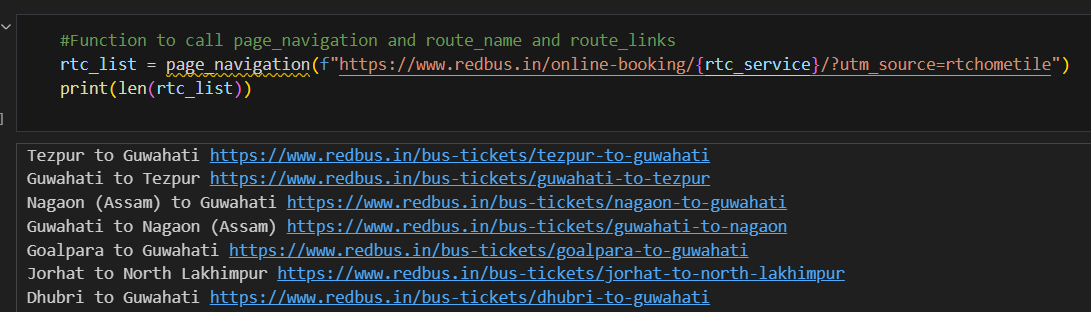
**Scraping process steps** :

1. Function **scroll\_to\_bottom():** To achieve page scrolling to bottom of the page and move to next page to scrape list of route names from the entire page.
2. Function **page\_navigation():** Scrapes route names matching the xpath expression for each government bus rtc service on the redbus website and route link URL from hyperlink (href ) attribute which is used in next step to fetch the bus detail.
3. Function**bus\_route\_details()** : This function fetches all the bus information for a given route link. If there are multiple Govt Bus services present , it will unhide the bus information of all Govt Bus Service, then it will scrape all the bus details mentioned below for both Govt and Private Bus Services and creates a list for each column.
   1. Bus name
   2. Bus type(Sleeper/Seater/AC/Non-AC
   3. Departing time
   4. Duration
   5. Reaching Time
   6. Bus Ticket Price
   7. Bus Seat Availability ,
   8. Bus Star Rating

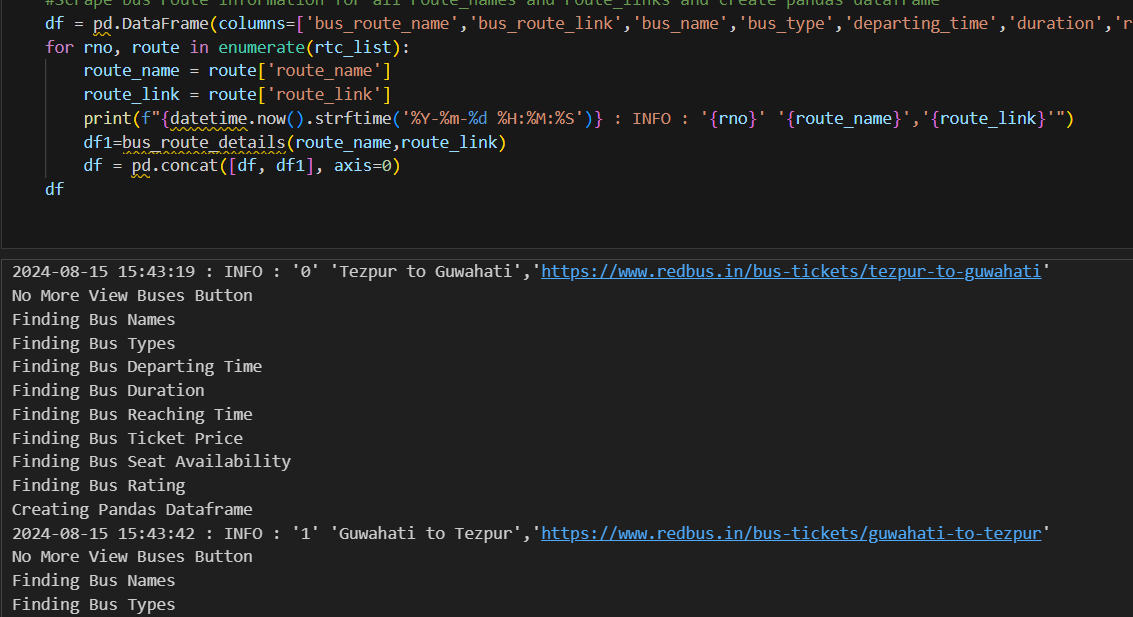
Added logic for handling exception if star rating is not present for any of the bus service. Convert above lists to a pandas dataframe using dictionary.

1. For each Govt bus service, a separate jupyter notebook has been created. It performs below steps,

* Find the route name and route links using page\_navigation function and store it in a list.



1. Loop through the list and find the bus details using bus\_route\_details function. A dataframe is created for each route which is finally concatenated to a single master dataframe.



* This master dataframe is written to a csv file with respective govt bus service name.
* This process is repeated in separate jupyter notebook for all the govt bus services.
* **Database creation and storage :**

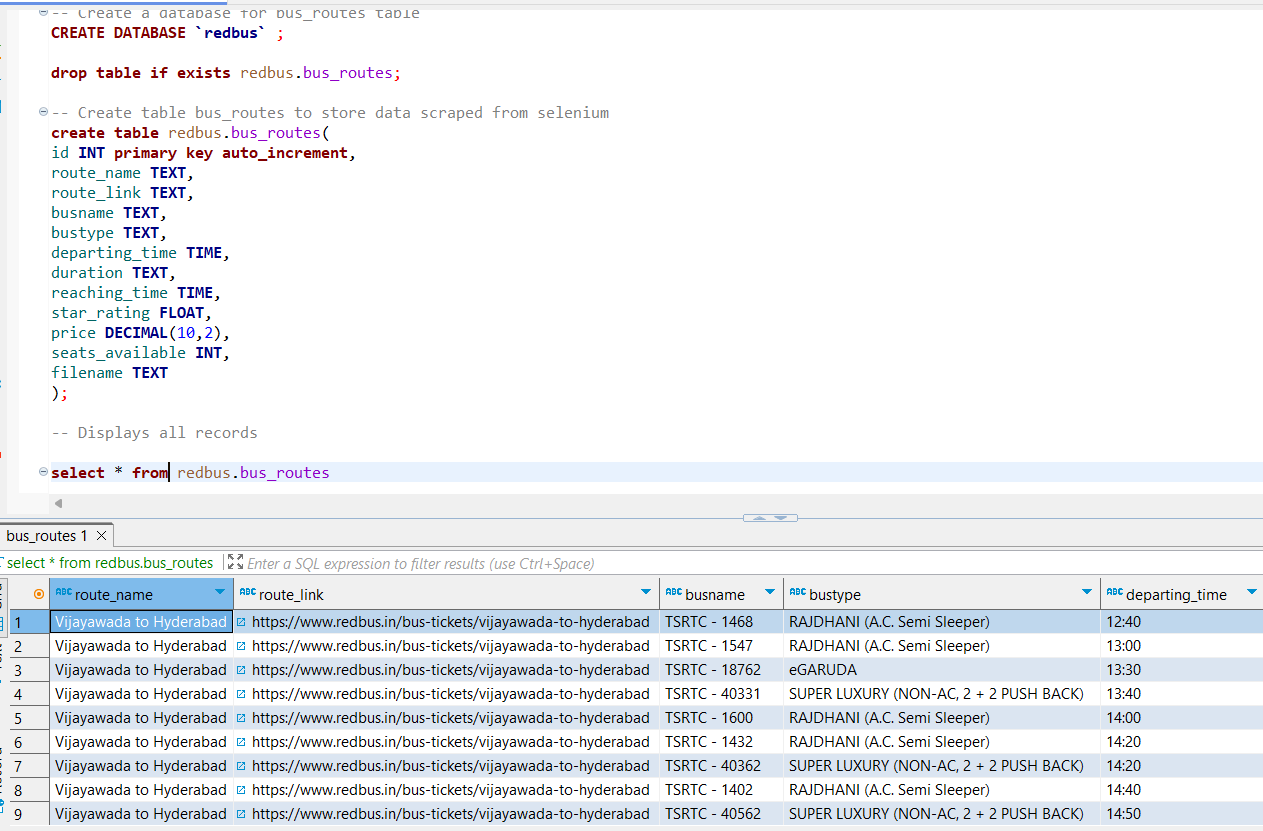
1. In MySQL, create a database schema (**Redbus**)and create a table (**bus\_routes**) to store the bus details.
2. Read the csv file of each govt bus service and create a consolidated pandas dataframe.

* **Data preprocessing and cleaning :**

As part of data cleansing below operations are performed on this dataframe

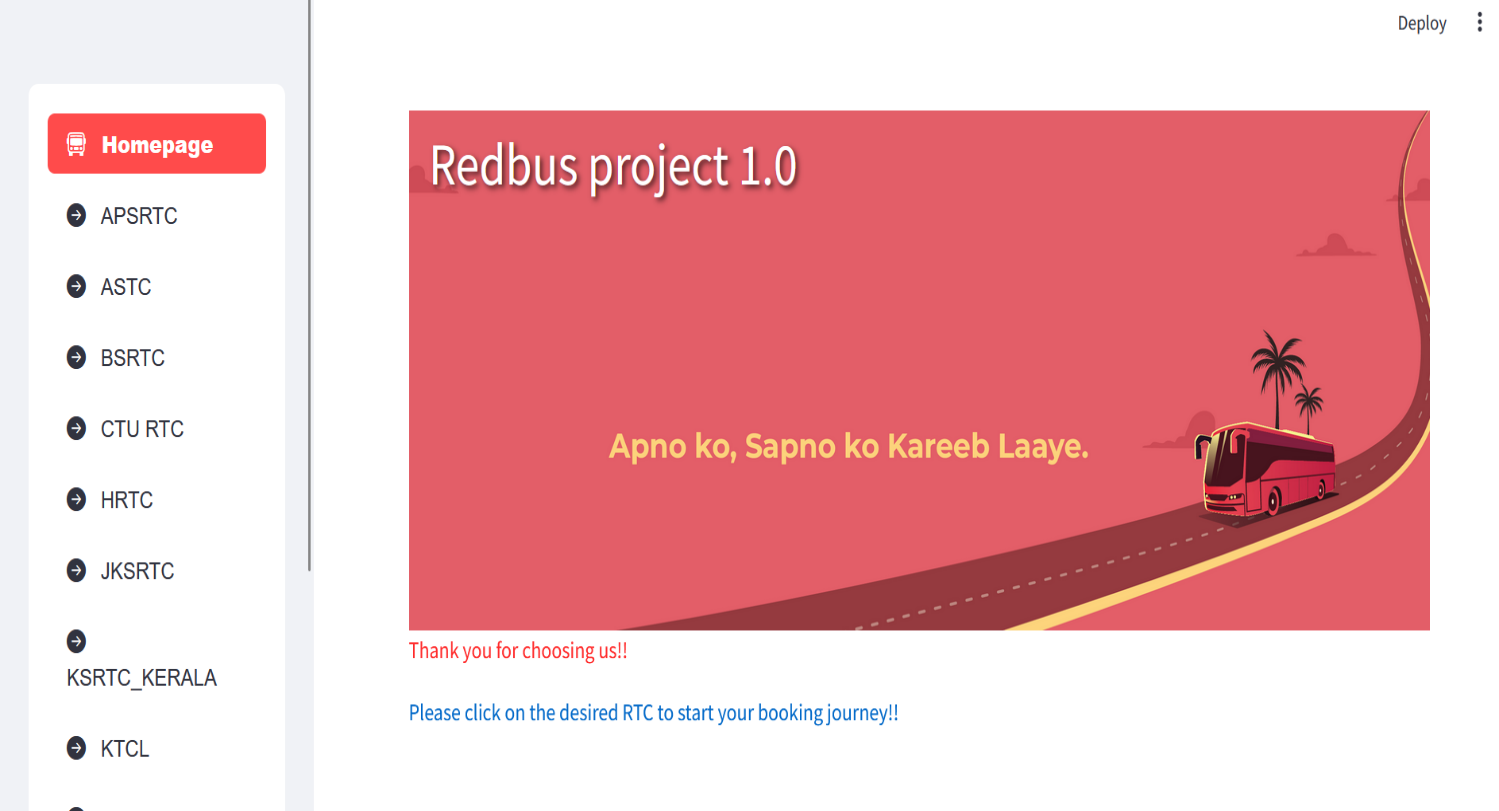
1. Column Name Correction
2. Fetch Numeric value for Seat Availability Column
3. Null record handling
4. Formatted the RTC Name using file name column

Use the cleansed dataframe to load the data into **redbus.bus\_route** table in MySQL.



* **Data Filtering and display in Streamlit Web app:**

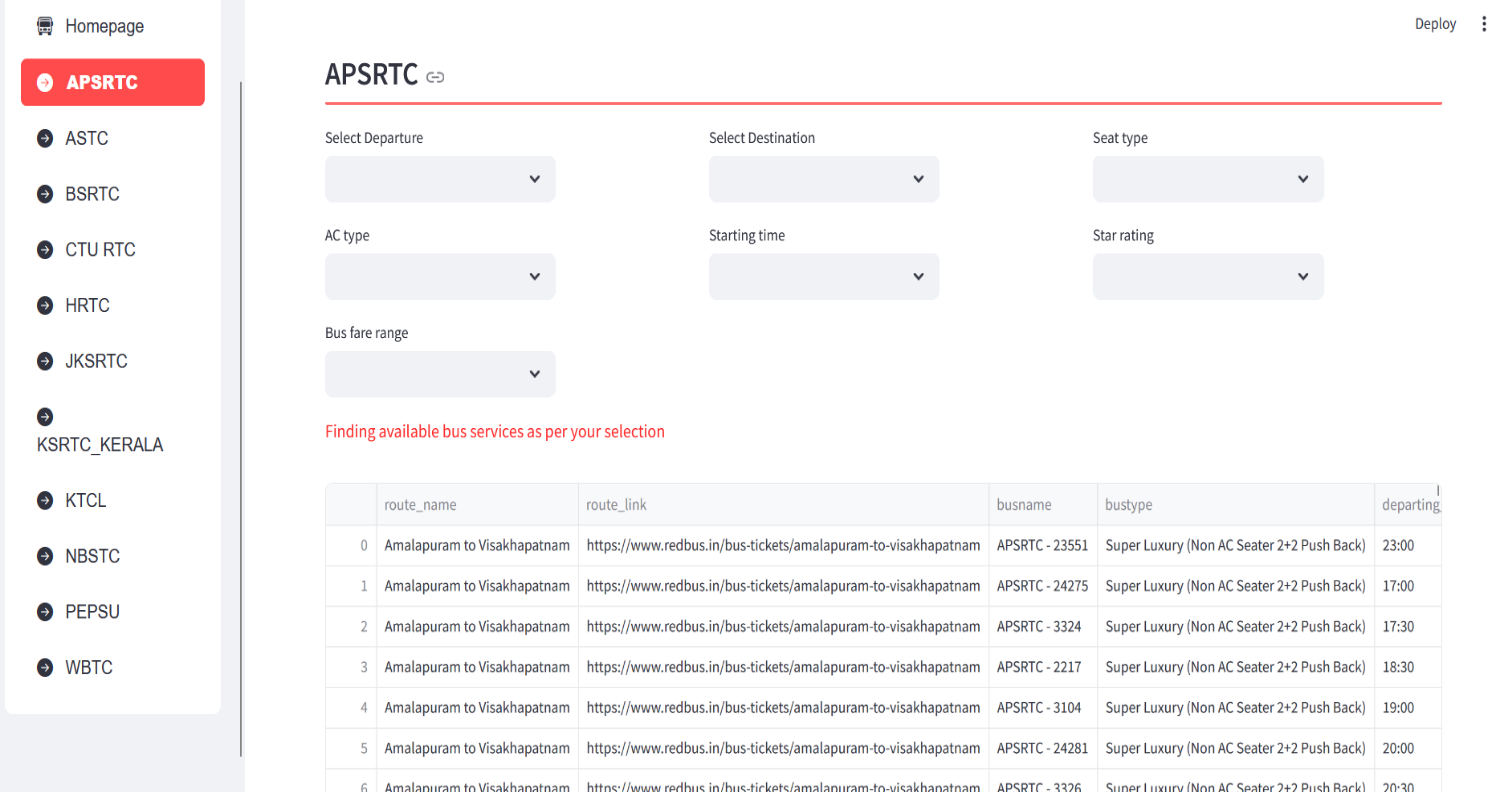
1. Function **get\_mysqldata(query)**  is to establish MYSQL connection and pull data using SQL alchemy engine to pandas dataframe
2. Streamlit styling options like wide mode used for better user experience



1. Sidebar selection menu added for displaying multiple RTC using Streamlit option menu.
2. Function **route\_name\_filter(df,rtc\_name):** Applies filter on dataframe based on RTC Option selected and displays page with different filters for Bus names, schedule, different price range, seat availability, and star rating options to get wide range of options for user.

* Respective RTC wise bus details for customized selection

Example1: APSRTC



Example 2: PEPSU

