

**Project:**

**Data Analysis of Trip Advisor Website**



Collect, Store, Retrieve Data

Data Analytics DA5020

**Master of Science in Data Analytics**

**Hemanth Lakshman Raju**

Under the guidance of Prof. Kathleen Durant

****

**Introduction:**

Travelling is an integral part of life, especially for people who have travelling as a passion. Not only they provide the experience, but they also give a quick view on the culture and the food they consume. Travelling also provide a break from the rat race that everyone runs. The most important player in this field is tripadvisor.com - the website which helps you make plan from the flight to the restaurant that you want to stay in.

Tripadvisor.com is one of the most reliable website that provides services like finding restaurants, booking hotels and flight booking. The website gives us suggestions while we search for the hotels, restaurants and flights which have the best deals in each of the service. It is the largest travel site in the world, with more than 315 million members and over 500 million reviews and opinions of hotels, restaurants, attractions and other travel-related businesses. The website services are free to users, who provide most of the content, and the website is supported by a hotel booking facility and an advertising business model.

**The idea of the project is to create a bird’s eye view of a locality showing Hotels, Restaurants and Tourist Spots in the locality. Trip advisor does not have a consolidated view of all the services it offers. We plan to achieve that with this project. For this Project, we have selected one of the most popular tourist spots, Paris.**

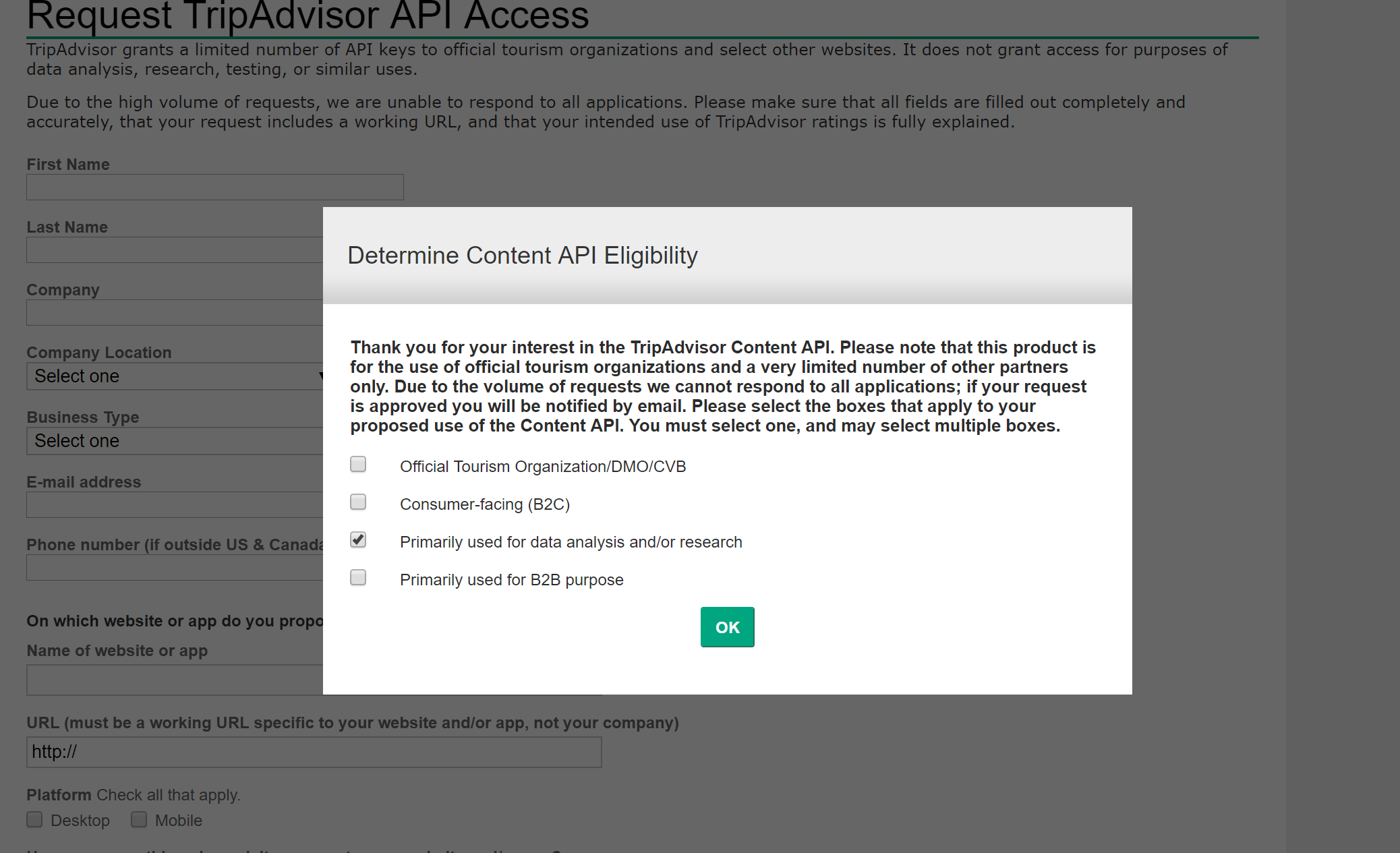
The objective of the proposed project is to scrape data from tripadvisor.com about the restaurants, hotels and popular tourist spots and visualize the data using R packages such as ggmap() and ggplot(). For this project we will consider one particular city and perform Analysis on the data which we get from TripAdvisor. The database will be created in SQLite.

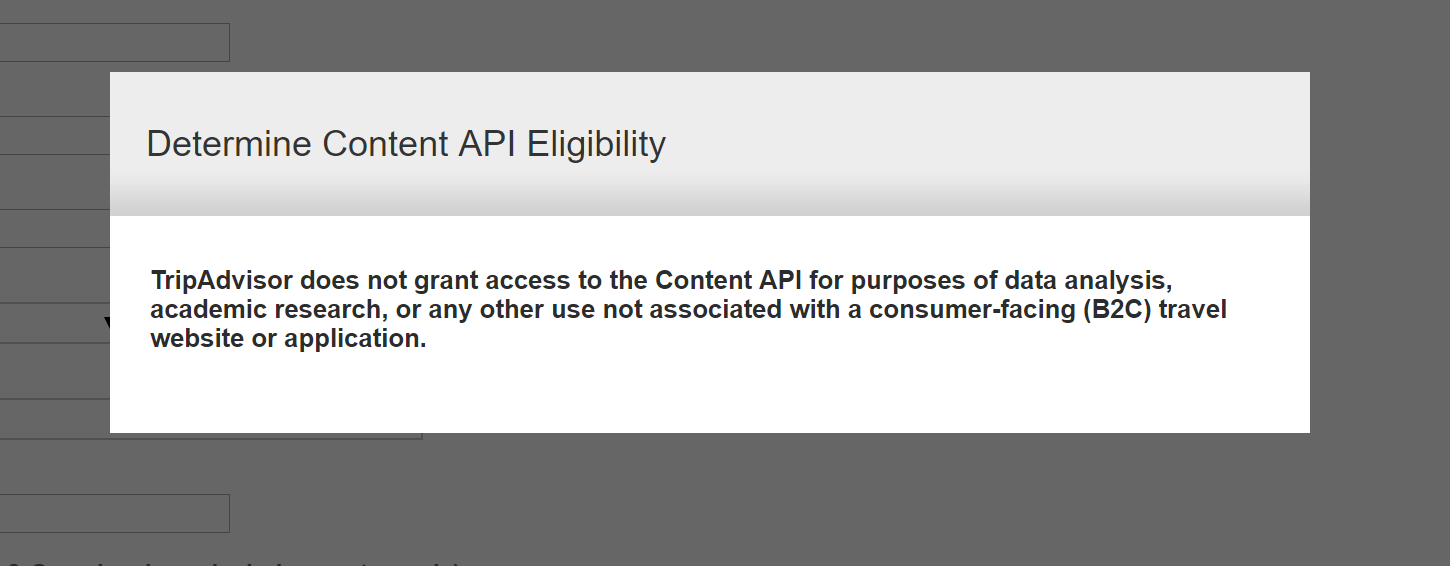
Web Scraping will be done on the TripAdvisor Website to gather the information. After the data scraping we perform multiple analysis of data in various forms such as sentimental analysis and analysis based on location of the best services provided. These analyses are done to ease the selection of restaurants, hotels and tourist spots for the tourists.

**Step 1: Data Collection**

We planned on using the TripAdvisor API for our project but when we tried to sign-up as a developer,

we were faced with this issue.



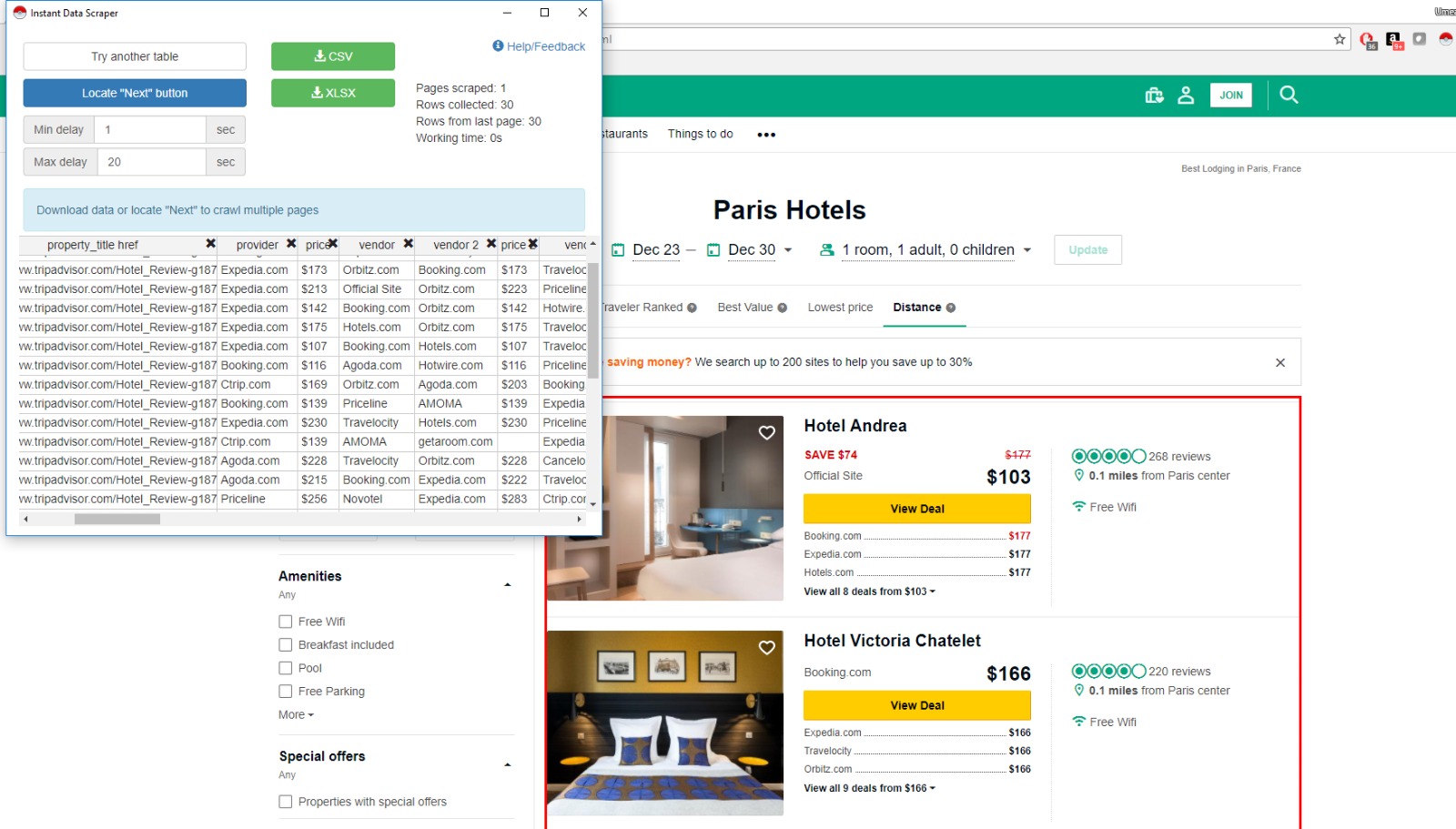


So, after discussing the issue with the professor, we decided to scrape the data using Google Chrome Extensions namely **‘Instant Data Scraper’** and **‘Web Scraper’**

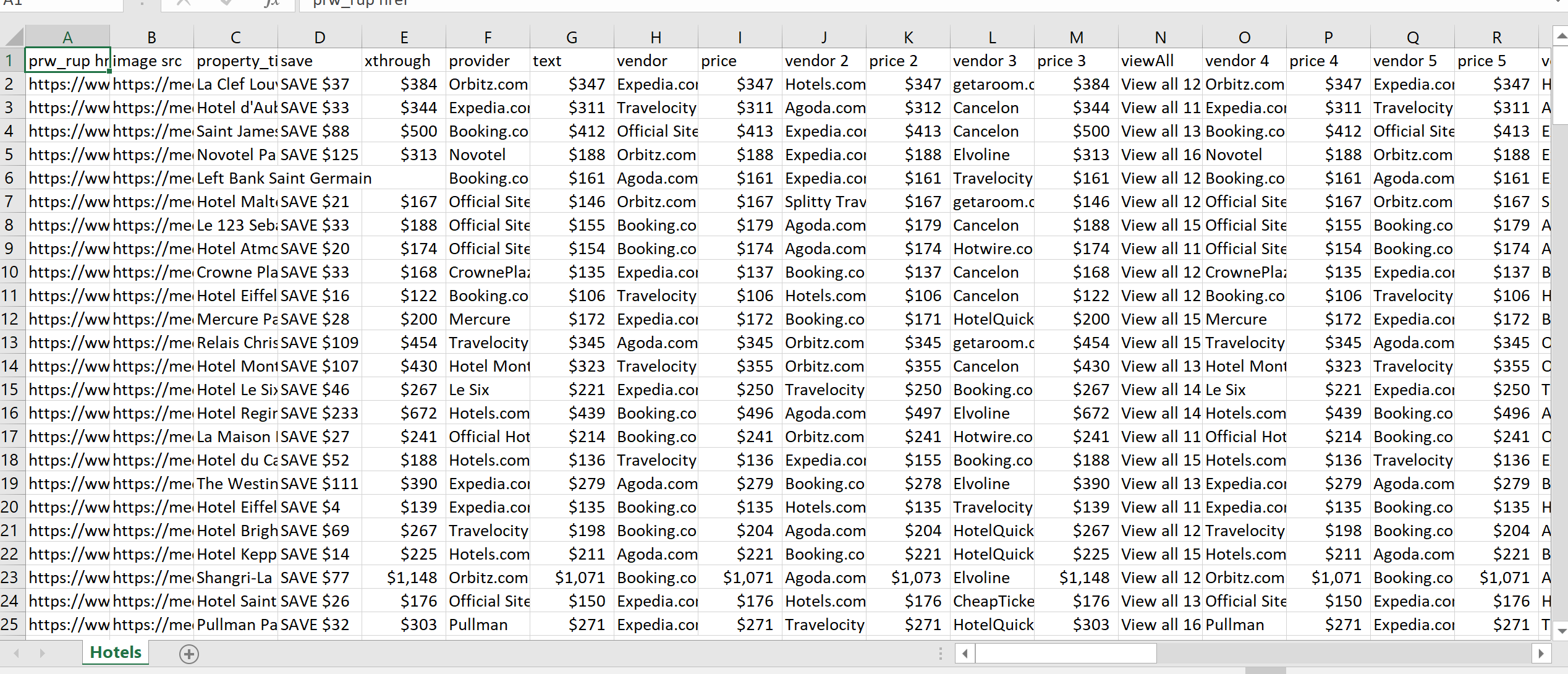
Collection of data from the tripadvisor.com website on the **Hotels** in Paris. The data collected from the website are,

* Hotel Name
* Hotel web link
* Cheapest provider
* Price
* Reviews
* Review count
* Distance
* Amenities

Screenshot showing the data scraping



Scraped Data



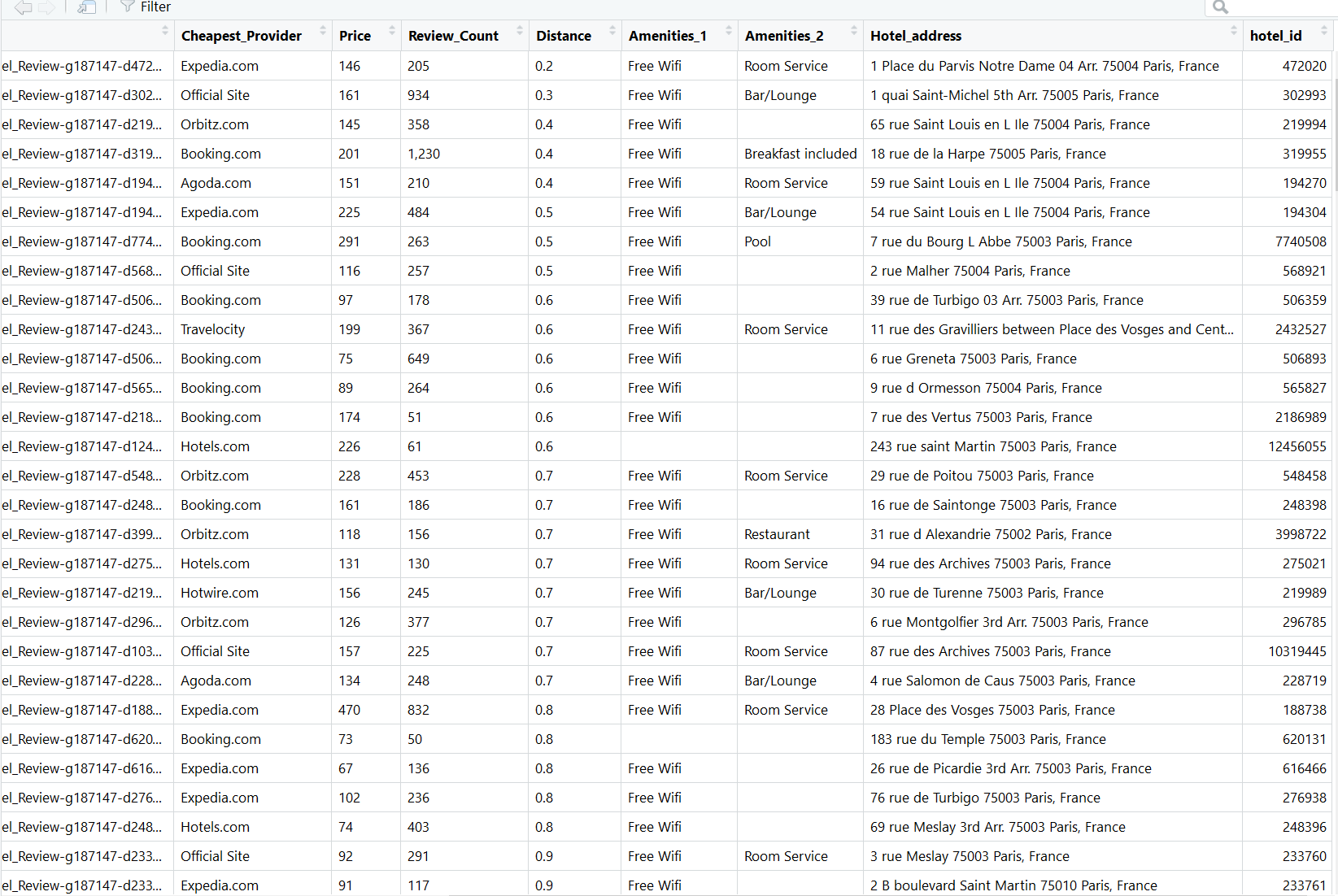
However, these data scrapers have their limitations. It was unable to scrape the address of the hotels because it was not readily available on the front page. (seen in screenshot)

We had to get the link of each Hotel and fetch the address data from that hyperlink. We wrote the code in R to do that.

We got the web link from the scraped data. Using that web link for each hotel, we ran a ‘for’ loop to fetch the address of the Hotels using html nodes.

Here is the code.

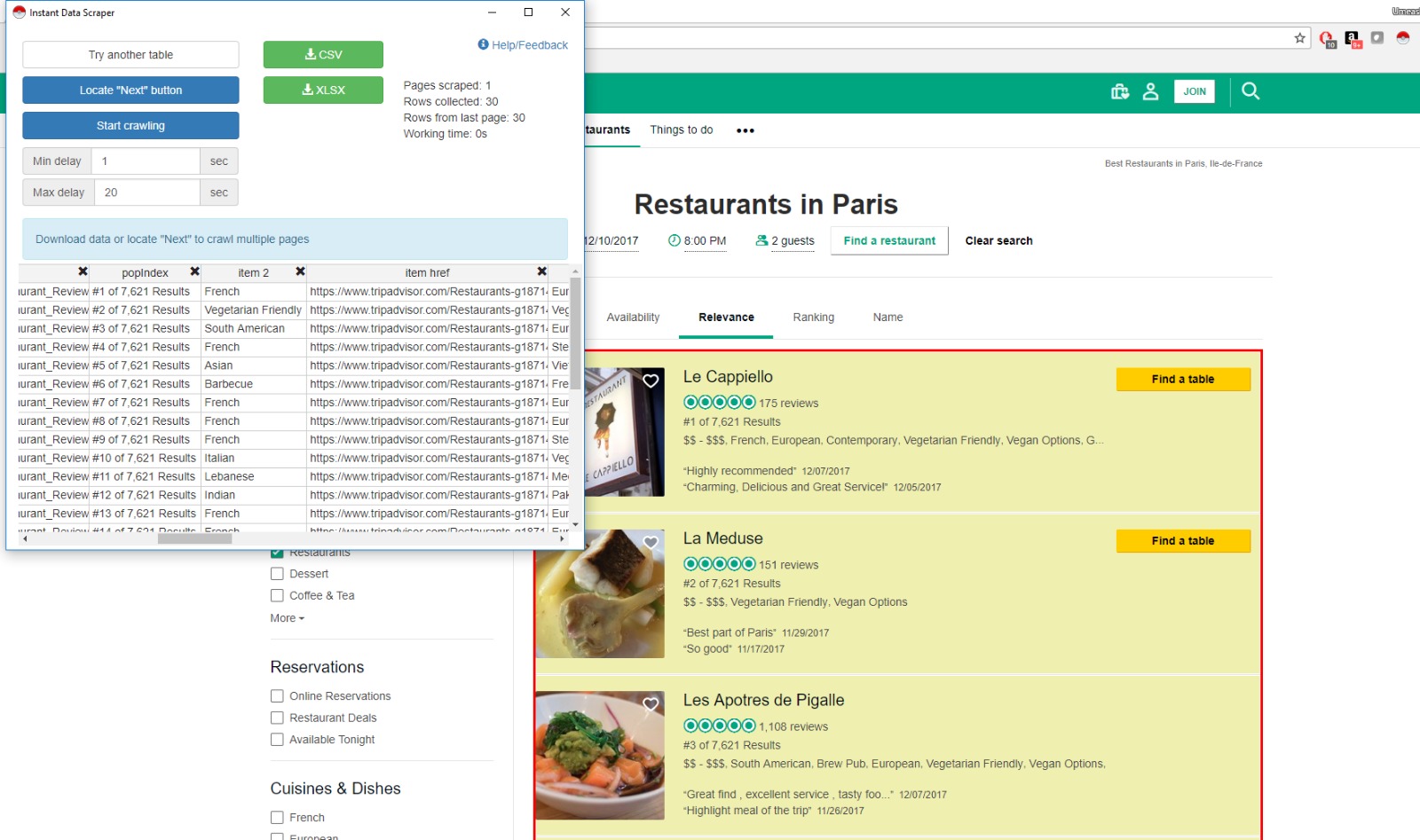




Collection of data from the tripadvisor.com website on **Restaurants** in Paris. The data collected from the website are,

* Restaurant name
* Restaurant web link
* Price range
* Cuisines offered
* Reviews
* Review date

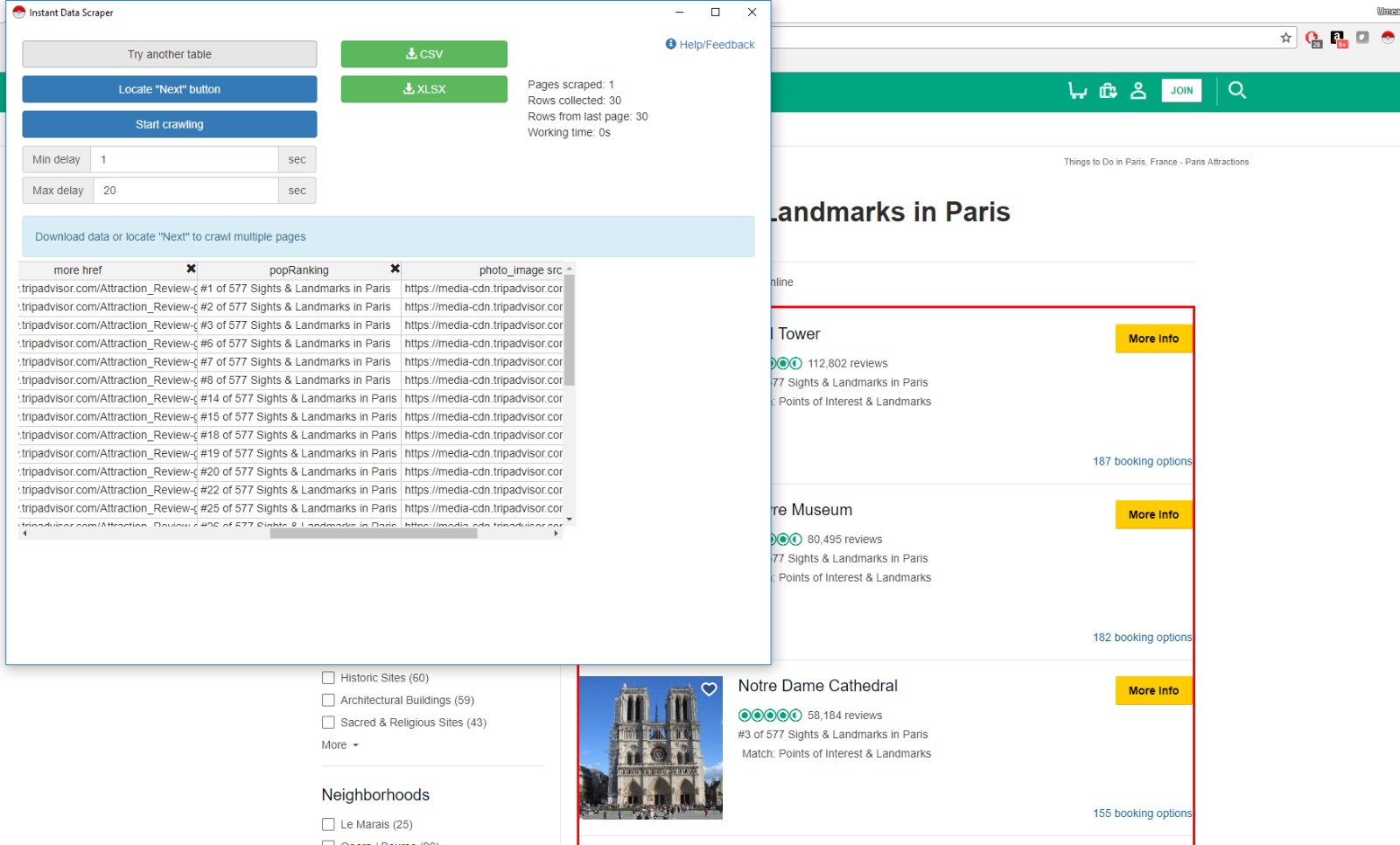
Screenshot of the data collected,



Collection of data from the tripadvisor.com website on **Tourist Spots** in Paris. The data collected from the website are,

* Name
* Web link
* Number of reviews
* Ranking
* Number of booking options

Screenshot of the data collected,

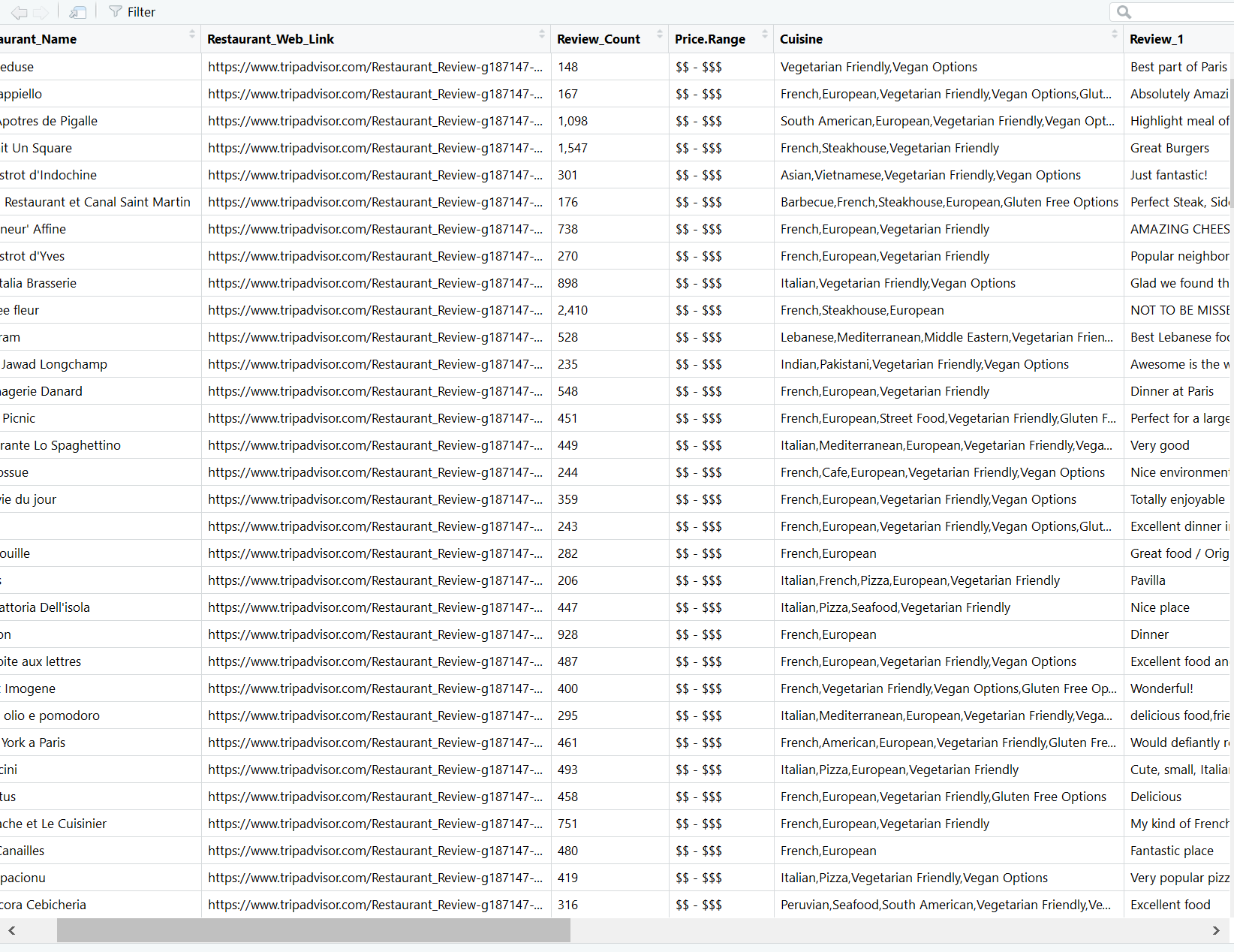


**Step 2: Cleaning the Data**

We got a lot of irrelevant and unnecessary data which was not needed for our Analysis. Hence we cleaned up the data as per our requirements.

* We removed the unwanted columns.
* We changed the Column Names
* Combined the cuisines offered into one column.
* Retrieved ID and Address using the Web Link.
* Removed unnecessary commas, words and Null Values from the dataset

**Cleaned Data – Restaurants**



We performed the same steps for Hotels and Tourist spots.

**Step 3: Storing the Data**

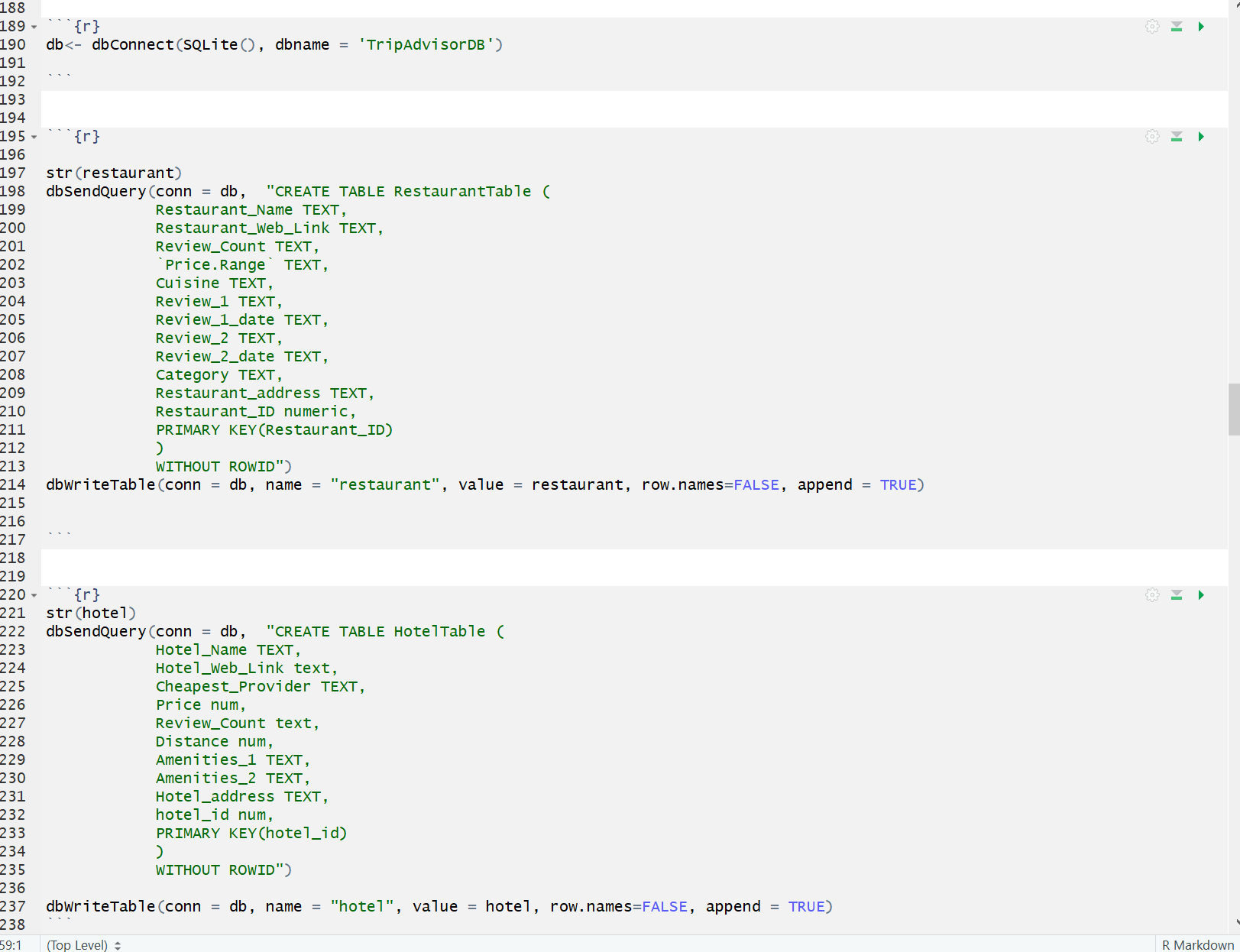
For storing the data, we used the RSQLite package to use SQLite as our database.

We created a database called **TripAdvisorDB**.

After the collection of data from the website, we created three Tables namely

* **RestaurantTable**
* **HotelTable**
* **POI\_Table**

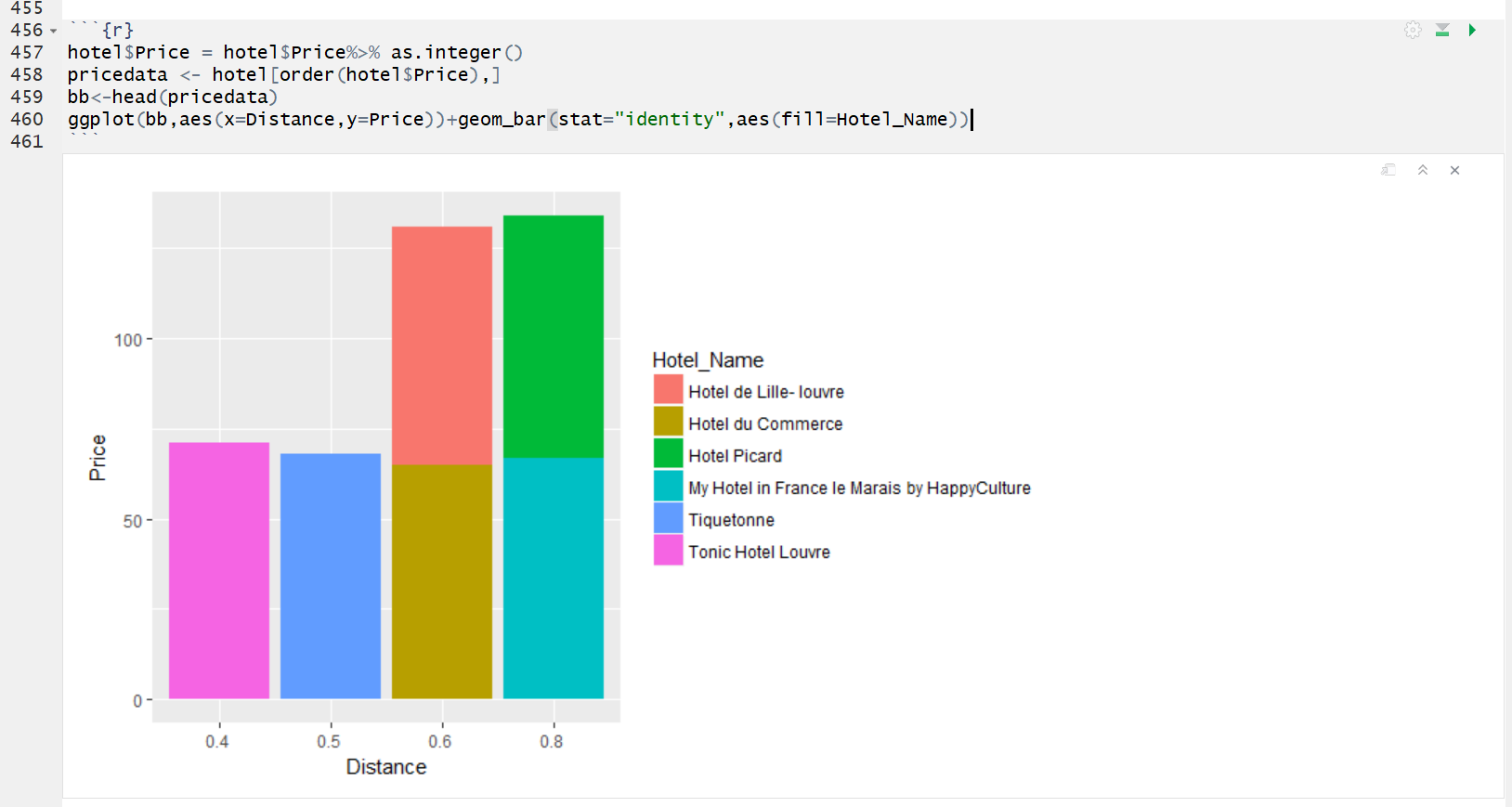
While creating the tables we needed a primary key for each Table, so, we used the weblink to get a unique ID for each entity and assigned it as the primary key for each table.



**Step 4: Data Analysis**

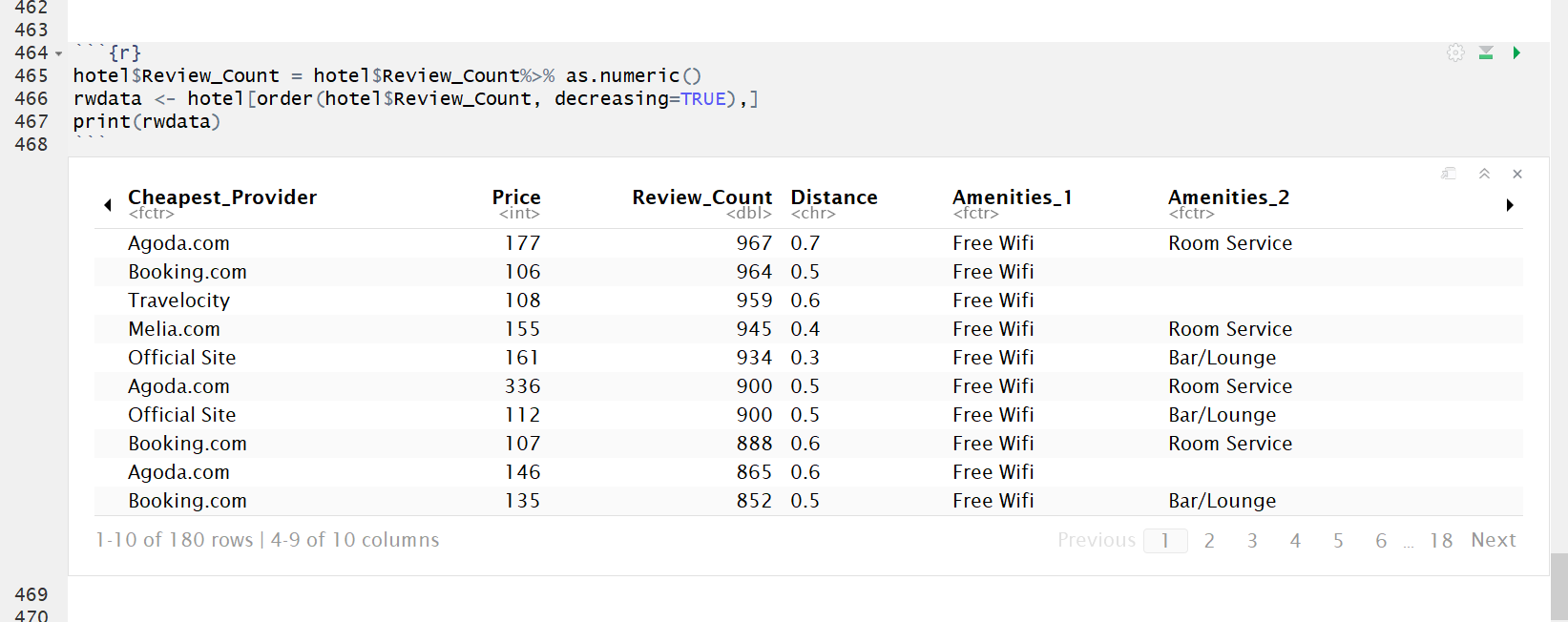
We did various types of analysis with the data that we collected.

1. Sentimental Analysis of the reviews.
2. Analysis of Hotels – How Price varies with distance from Paris center.
3. Highest Reviews, Highest bookings, Cheapest Hotels etc
4. Plotting location of hotel, restaurant and tourist spots on a map.
5. A bird’s eye view of the location showing the hotels, restaurants and tourist spots in the area for ease of planning for the tourist.



Price in dollars vs Distance in miles

Hotels Ranked according to highest reviews.



We performed sentimental analysis on the reviews by the travelers for each of the three entities.

For example, here is the sentimental analysis of the reviews of each restaurant.

We decided to create a wordcloud of the words picked up during sentimental analysis:

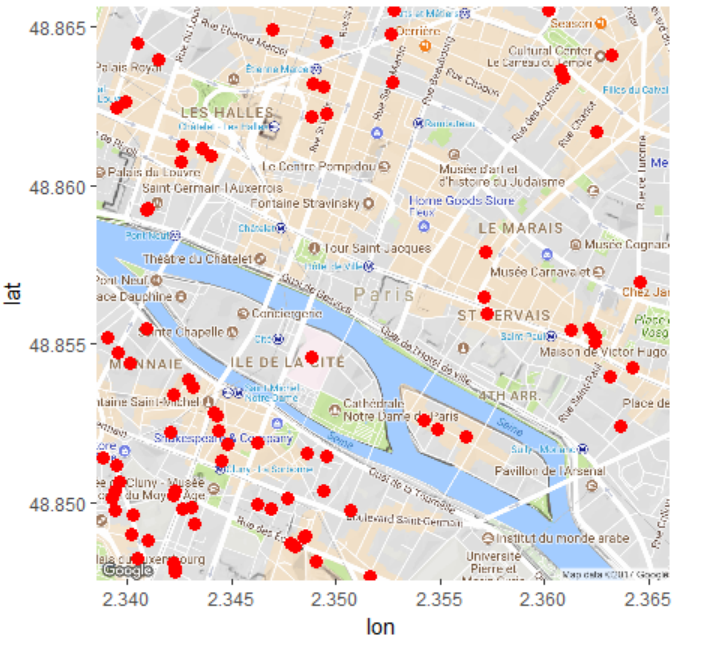


**WordCloud**



* Plotting the Location of the Hotels on the map.





* Bird’s Eye View of all the Restaurants, Hotels and Tourist Spots in the area.



**Red- Hotels Green – Restaurants Purple – Tourist Spots**

**Insights**

**Learning Outcomes:**

We chose to do Analysis on TripAdvisor.com because the data offered for a search is very elaborate and often very confusing for the user. With this project, we aimed to give a very concise and straightforward response to a user’s query.

During the course of this project, we learnt a lot of new techniques for analysis and visualization.

* We learnt to effectively use ggmap function.
* How to use HTML nodes to our advantage while scraping the data.
* How to assign primary keys to a database.
* Text analysis – Sentiment Analysis, Wordcloud.

We had difficulty to get the address of the hotel as it was inside a follow-up link. To overcome this problem, we tried using a variety of web scrapers and Chrome Extensions. We even tried to build our own web scraper using BeautifulSoup in Python.

Finally, we figured out that we could use the hyperlink extracted during our initial web scraping to get to the web page of each hotel and individually scrape the address.

**Future Possibilities:**

* We could further enhance the project by building a ‘Shiny’ app.
* We could use the location specified by the user to get the Hotels, Restaurants and Tourist spots in the locality.
* We can extend our project to include flight details and form a tour package for the specified budget.

**References:**

1. TripAdvisor.com
2. Wikipedia.com
3. R for Data Science – Hadley Wickham