1. Intro (What topic? Why topic?) 200
2. Describe our data 200
3. How do we get them 500--- 3 parts
4. How do we store them 500
5. What problem we met, how did we overcome or change our process 800-1000 --- 3 parts
6. How did we manage the project as team 500
7. What can be done with our data 500
8. Conclusion and limitation 500

Problem with web scraping:

The first issue we met was trying to find APIs for commercial restaurant ranking websites. We checked the API policy for Tripadvisor, Opentable, and Zomato and all of them only provide APIs for commercial purposes or in exchange for traffic. Thus, we will have to get information by web scraping.

However, commercial websites like Tripadvisor seem to have a strict policy against web scraping. At first, we can hardly get the website content due to connection issues. After changing a few, we finally found a header with a relatively higher success rate. The chance of successfully getting website content is about 1/3.

The second issue we had was the URL would not change if we go to the second/third page. At first, we thought it was because they are using Ajax, so we checked the network section for these websites. However, we didn’t find anything different between the first and second pages. Then we tried to use selenium to simulate clicks on the next page button. As a result, the success rate for a request is even lower. Finally, we found that each page has its unique URL, but it was hidden on the top.

There are some pages that have less content than others. In order to get content as much as possible, we added a try and expect function to get these parts of the content.

In the end, we still need to handle the problem of the low success rate of getting requests. The approach we used was to try three times before it fails to connect and if we still can’t get anything, we will then leave a blank row in our dataframe. We will then rerun the function to get information to fill in the blank. After a few rounds of reruns, we managed to get about 250 content for 300 records.

Since TripAdvisor only provide APIs for non-commercial purpose, data are retrieved through web scraping. Python packages such as BeautifulSoup and requests are used in web scraping.

However, TripAdvisor has a strict policy against web scraping, thus there is only about 1/3 chance to establish a successful connection with it even after trying different headers.

At most 3 try and except are used to gain a higher chance of establishing a successful connection with TripAdvisor. Additionally, multiple web scraping results are combined. As a result, about 250/300 rows of data are successfully retrieved.

The limitation of using web scraping is that content on the websites may not be arranged in the same way. In this case, some pages have less information than the others. To get the most of the information, a try and except method is implemented.