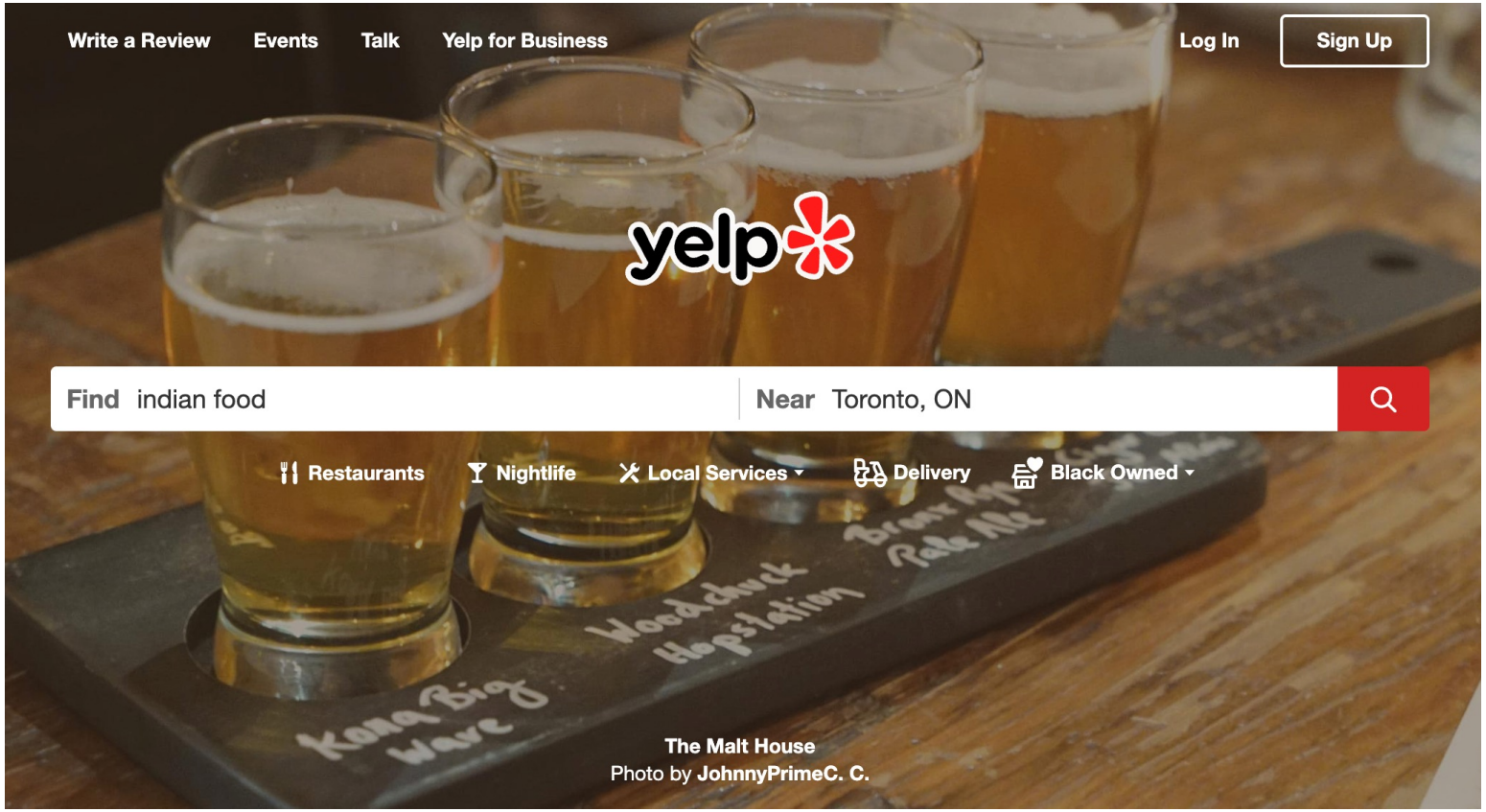
Capstone project #1 : Web scraping
Indian restaurants in Toronto



Founded in 2004 and headquartered in San Francisco, Yelp publishes crowd-sourced reviews about businesses

29th December, 2021

Manika Midha

Step by Step approach :

- Find Indian restaurants in Toronto : extracted their names, areas, star rating, reviews count, price range, cuisines offered, delivery option. Collected data over 24 pages in a csv.
  - ✓ Did analysis
  - ✓ With data analysis, found the most popular Indian restaurant among other things

- Did analysis of the best dishes offered by the most popular restaurant : extracted dish name, number of photos and reviews and found top dishes.

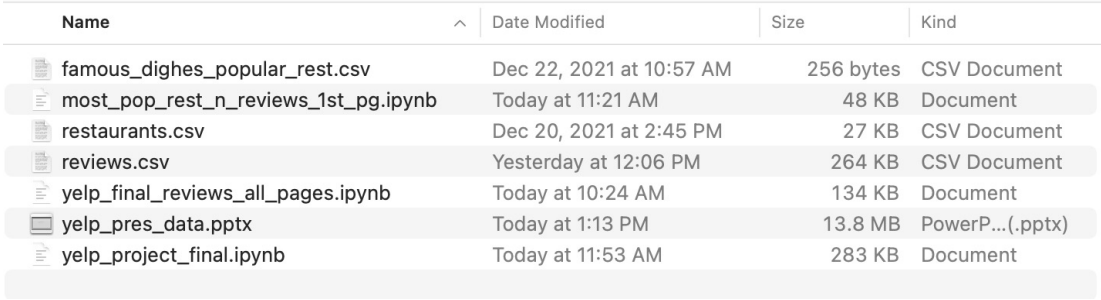
- Did analysis of the reviews : extracted review text, date of review, star rating, location of reviewer, count of useful, funny, cool reviews. Collected data over 42 pages in a csv.

Data analysis - Restaurants

- Average star rating
- Average count of reviews
- Number of restaurants with specific rating
- Area wise Indian restaurants
- Areas with highly rated Indian restaurants
- 5 star rating Indian restaurants with most reviews
- Top 10 Indian restaurants

Data analysis - Most popular restaurant

- Top dishes
- Average star rating of a review
- Number of reviews with specific rating
- Most useful review
- Most funny review
- Latest review
- Number of reviews per year



Name	Date Modified	Size	Kind
famous_dighes_popular_rest.csv	Dec 22, 2021 at 10:57 AM	256 bytes	CSV Document
most_pop_rest_n_reviews_1st_pg.ipynb	Today at 11:21 AM	48 KB	Document
restaurants.csv	Dec 20, 2021 at 2:45 PM	27 KB	CSV Document
reviews.csv	Yesterday at 12:06 PM	264 KB	CSV Document
yelp_final_reviews_all_pages.ipynb	Today at 10:24 AM	134 KB	Document
yelp_pres_data.pptx	Today at 1:13 PM	13.8 MB	PowerP...(.pptx)
yelp_project_final.ipynb	Today at 11:53 AM	283 KB	Document

Code for analysis of Indian restaurants in Toronto

```
from selenium import webdriver
from selenium.webdriver.common.keys import Keys

from bs4 import BeautifulSoup
import requests
import pandas as pd

driver = webdriver.Chrome()
url = "https://www.yelp.ca/toronto"
driver.get(url)
```

```
#find search bar and enter 'Indian food' and press enter
search_bar = driver.find_element_by_xpath('//*[@id="find_desc"]')
search_bar.send_keys('Indian food')
search_bar.send_keys(Keys.RETURN)
```

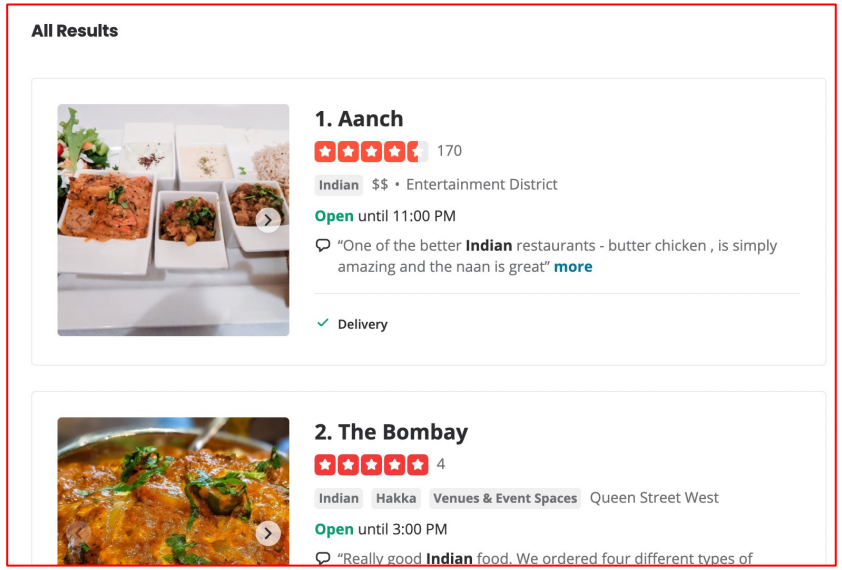
```
#Given a page number, returns soup for that page
import requests

def return_page_soup(page_num):
    url_page_num = (page_num - 1)*10
    page_url = f'https://www.yelp.ca/search?find_desc=Indian+food&find_loc=Toronto%2C+ON&ns=1&s
    try :
        response = requests.get(page_url)
        code = response.status_code
        if code == 200 :
            driver.get(page_url)
            page_source = driver.page_source
            soup = BeautifulSoup(page_source)
            return soup
    except :
        return 'Bad request'
```

```
#Given soup per page, returns dictionary per page
def get_page_info(soup1):

    dict_per_page = {
                    'rest_name' : [],
                    'star_rating' : [],
                    'restaurant_reviews_count' : [],
                    'restaurant_price' : [],
                    'restaurant_cuisines' : [],
                    'restaurant_area' : [],
                    'restaurant_delivery_options' : []

                    }

    #refers to all restaurants on one page
    containers = soup1.find_all('div', class_ = 'container__09f24__mpR8_')
```

```
#1 container refers to one restaurant
for container in containers :

    #restaurant name
    try:
        restaurant_name = container.find('a',class_ = 'css-1422juy').text
    except :
        restaurant_name = ''
    dict_per_page['rest_name'].append(restaurant_name)

    #star rating, get the number of rating and convert in float
    try:
        star_rating = container.find('span',class_ = 'display--inline__09f24__c6N_k border-
        star_rating_number = float(star_rating.split()[0])
    except :
        star_rating_number = 0.0
    dict_per_page['star_rating'].append(star_rating_number)

    #number of reviews - should be integer
    try:
        reviews_count = int(container.find('span', class_= 'reviewCount__09f24__tnBk4 css-1
    except :
        reviews_count = 0
    dict_per_page['restaurant_reviews_count'].append(reviews_count)

    #price range - $ sign
    try:
        rest_price = container.find('span', class_='priceRange__09f24__mmOuH css-18qxe2r').
    except :
        rest_price = ''
    dict_per_page['restaurant_price'].append(rest_price)
```

```
    # cuisine list
    try:
        cuisines = container.find_all('p',class_ = 'css-1p8aobs')
        cuisine_list = []
        for cuisine in cuisines :
            cuisine_list.append(cuisine.text)
    except :
        cuisine_list = []
    dict_per_page['restaurant_cuisines'].append(cuisine_list)

    #area of restaurant
    try:
        area = container.find_all('span', class_ = 'css-1e4fdj9')[-1].text
    except :
        area = ''
    dict_per_page['restaurant_area'].append(area)

    #delivery options
    try:
        delivery_options = container.find_all('span', class_ = 'raw__09f24__T4Ezm')
        delivery_options_list = []
        for delivery_option in delivery_options :
            delivery_options_list.append(delivery_option.text)
    except :
        delivery_options_list = []
    dict_per_page['restaurant_delivery_options'].append(delivery_options_list)

return dict_per_page
```

```
#get total num of pages :
total_pages = soup.find('div', class_ = 'pagination__09f24__VRjN4 border--top__09f24__exYYb bor
total_pages_list = total_pages.find('span', class_ = 'css-1e4fdj9').text.split() #['1','of','24
total_page_count_int = int(total_pages_list[2]) #24

master_df = pd.DataFrame()

for page in range(1,total_page_count_int+1):
    print(f'getting info from page {page}')

    df_page = pd.DataFrame()

    soup = return_page_soup(page) #func call 1, returns soup per page
    dict_for_page = get_page_info(soup) #func call 2, returns dict per page

    df_page = pd.DataFrame(dict_for_page) #create dataframe from dict per page
    master_df = master_df.append(df_page,ignore_index = True)
    master_df.to_csv('restaurants.csv')
```



Code for all reviews of the most popular restaurant

```
from selenium import webdriver
from selenium.webdriver.common.keys import Keys

from bs4 import BeautifulSoup
import requests
import pandas as pd
```

```
restaurants_df = pd.read_csv('restaurants.csv')
restaurants_df_drop_col = restaurants_df.drop(['Unnamed: 0'], axis=1)
restaurants_df_drop_col_remove_dups = restaurants_df_drop_col.drop_duplicates(keep = False)
top_ten_rests = restaurants_df_drop_col_remove_dups.sort_values(['restaurant_reviews_count','st
most_popular_rest = top_ten_rests.iloc[3,:]
```

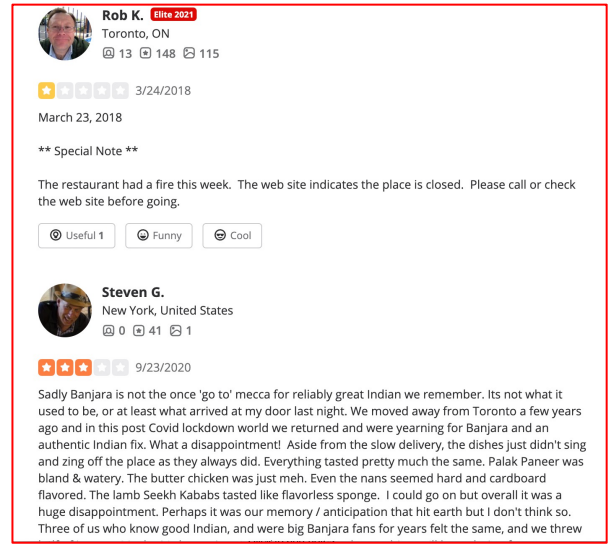
```
#returns url of most popular restauramt

def return_url(rest_name_lower, rest_name_area_lower):

    rest_name_area_lower_list = rest_name_area_lower.split()
    rest_name_list = rest_name_lower.split()

    x = '-'.join(rest_name_list)
    y = '+'.join(rest_name_area_lower_list)

    main_url_str = 'https://www.yelp.ca/biz/' + x + '-toronto?osq=' + y

    return f'{main_url_str}'
```

```
#get url of most popular restaurant :

import requests
driver = webdriver.Chrome()

rest_name_lower = most_popular_rest['rest_name'].lower()
rest_name_area_lower = most_popular_rest['rest_name'].lower() + ' ' + most_popular_rest['restau

page_url = return_url(rest_name_lower,rest_name_area_lower)
try :
    response = requests.get(page_url)
    code = response.status_code
    if code == 200 :
        driver.get(page_url)
        page_source = driver.page_source
        soup = BeautifulSoup(page_source)
except :
    print('Bad request')
```

```
#returns all information about all reviews on one page, of the most popular restaurant
#in a dict format

def get_page_review_info(soup):

    review_dict = {
        'review_text':[],
        'review_date':[],
        'review_star_rating': [],
        'reviewer_city': [],
        'review_useful_count':[],
        'review_funny_count':[],
        'review_cool_count':[]
    }
```

```
#each review container has class - review__09f24__oHr9V border-color--default__09f24__NPAKY
#information for a review is contained within this class
containers = soup.find_all('div', class_ = 'review__09f24__oHr9V border-color--default__09f

for container in containers :
    #get review text :
    try :
        comment = container.find('span', class_ = 'raw__09f24__T4Ezm').text
    except :
        comment = ''
    review_dict['review_text'].append(comment)

    #get review date :
    try:
        review_date_str = container.find('span', class_ = 'css-1e4fdj9').text
        review_date_formatted = pd.to_datetime(review_date_str,format = '%m/%d/%Y')
    except :
        review_date_formatted = ''
    review_dict['review_date'].append(review_date_formatted)

    #get star rating
    try:
        review_star_rating_str = container.find('span',class_='display--inline__09f24__c6N_
        review_star_rating_flt = float(review_star_rating_str.split()[0])
    except :
        review_star_rating_flt = 0.0
    review_dict['review_star_rating'].append(review_star_rating_flt)

    #get reviewer location
    try:
        reviewer_location = container.find('span', class_ = 'css-1sufhje').text
        reviewer_city = reviewer_location.split(',')[0]
    except:
        reviewer_city = ''
    review_dict['reviewer_city'].append(reviewer_city)
```

```
    #get emotions number : useful, funny, cool
    notes = container.find_all('span', class_ = 'css-1rw3tz3')
    try:
        useful_count =  int(notes[0].text.split()[1])
    except :
        useful_count = 0
    review_dict['review_useful_count'].append(useful_count)

    try:
        funny_count =  int(notes[1].text.split()[1])
    except :
        funny_count = 0
    review_dict['review_funny_count'].append(funny_count)

    try:
        cool_count =  int(notes[2].text.split()[1])
    except :
        cool_count = 0
    review_dict['review_cool_count'].append(cool_count)


return review_dict
```



Rob K. Elite 2021
Toronto, ON
13 148 115

3/24/2018

March 23, 2018

** Special Note **

The restaurant had a fire this week. The web site indicates the place is closed. Please call or check the web site before going.

Useful 1 Funny Cool

Steven G.
New York, United States
0 41 1

9/23/2020

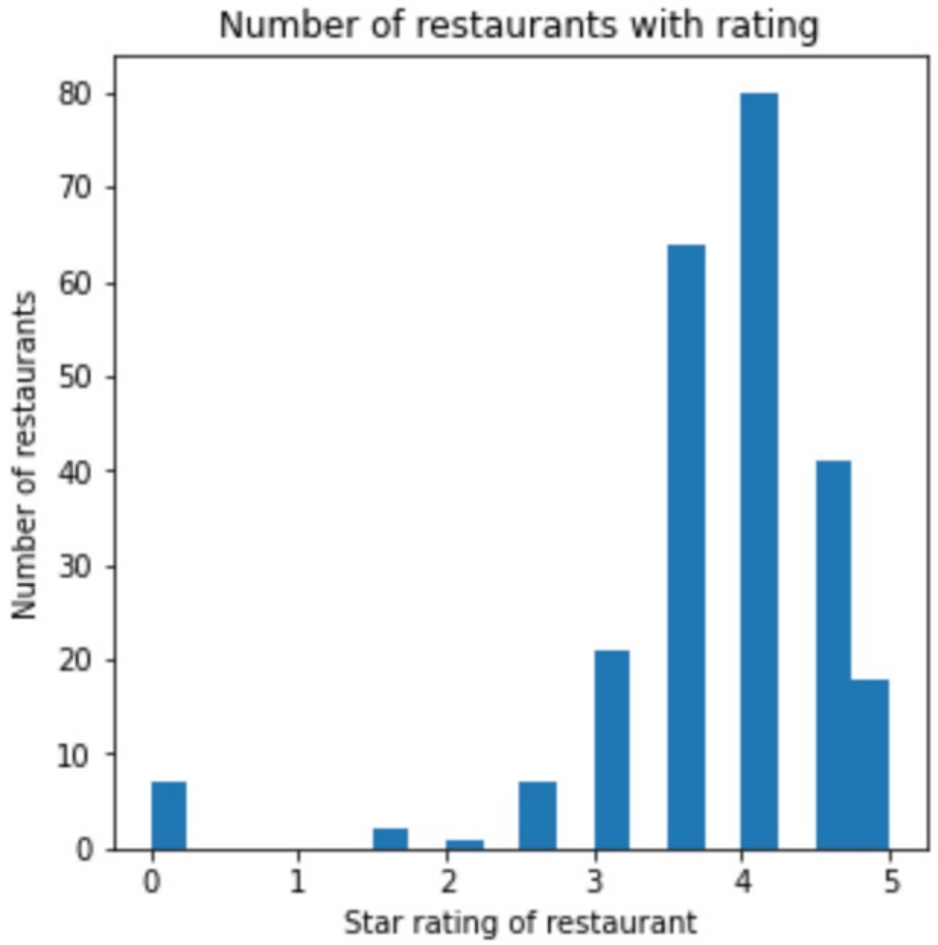
Sadly Banjara is not the once 'go to' mecca for reliably great Indian we remember. Its not what it used to be, or at least what arrived at my door last night. We moved away from Toronto a few years ago and in this post Covid lockdown world we returned and were yearning for Banjara and an authentic Indian fix. What a disappointment! Aside from the slow delivery, the dishes just didn't sing and zing off the place as they always did. Everything tasted pretty much the same. Palak Paneer was bland & watery. The butter chicken was just meh. Even the nans seemed hard and cardboard flavored. The lamb Seekh Kababs tasted like flavorless sponge. I could go on but overall it was a huge disappointment. Perhaps it was our memory / anticipation that hit earth but I don't think so. Three of us who know good Indian, and were big Banjara fans for years felt the same, and we threw

```
#Given a review page number, returns soup for that page
import requests
import time

def return_review_page_soup(page_num):
    url_page_num = (page_num - 1)*10
    page_url = f'https://www.yelp.ca/biz/banjara-indian-cuisine-toronto?osq=banjara+indian+cuis
    try :
        response = requests.get(page_url)
        code = response.status_code
        if code == 200 :
            driver.get(page_url)
            time.sleep(3)
            page_source = driver.page_source
            soup_rev_page = BeautifulSoup(page_source)
            return soup_rev_page
    except :
        return 'Bad request'
```

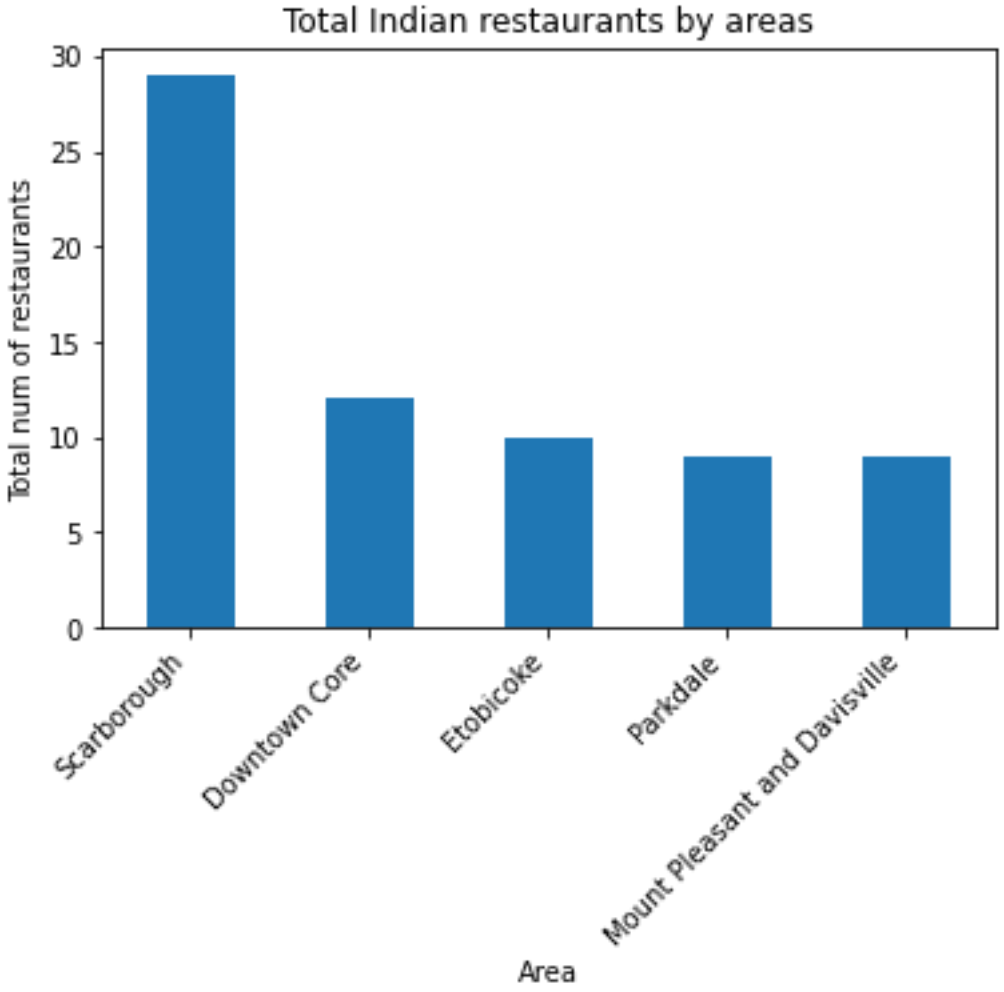
```
#gets reviews from each page, converts it into a dataframe and adds this info to a csv

#get total num of pages :
total_pages = soup.find('div', class_ = 'border-color--default__09f24__NPAKY text-align--center
total_pages_list = total_pages.find('span', class_ = 'css-1e4fdj9').text.split() #['1','of','24
total_page_count_int = int(total_pages_list[2]) #42


master_df_review = pd.DataFrame()
for page in range(1,total_page_count_int+1):

    print(f'getting info from page {page}')

    df_page_review = pd.DataFrame()

    soup_rev_page = return_review_page_soup(page) #func call 1, returns soup per page
    review_dict_for_page = get_page_review_info(soup_rev_page) #func call 2, returns dict per p


    df_page_review = pd.DataFrame(review_dict_for_page) #create dataframe from dict per page
    master_df_review = master_df_review.append(df_page_review,ignore_index = True)
    master_df_review.to_csv('reviews.csv')
```

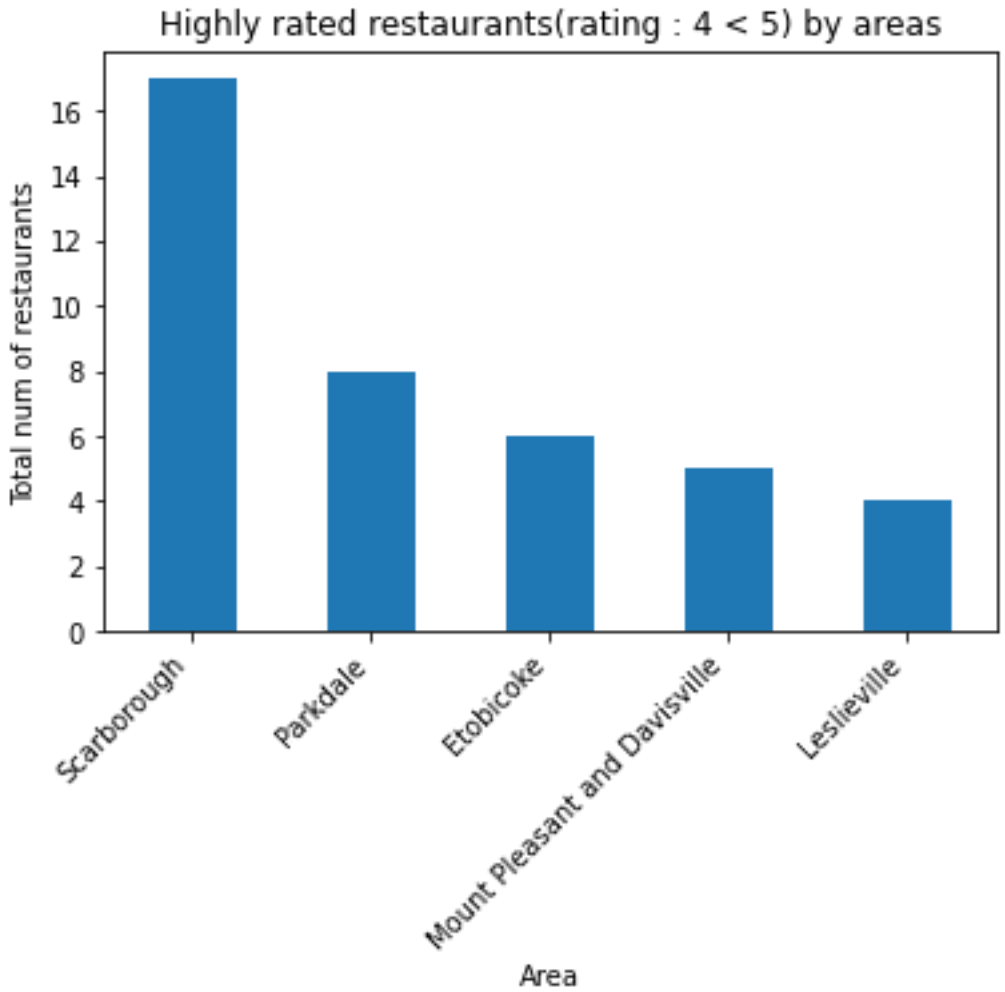
	star_rating	restaurant_reviews_count
count	241.000000	241.000000
mean	3.751037	55.726141
std	0.900375	91.633598
min	0.000000	0.000000
25%	3.500000	7.000000
50%	4.000000	27.000000
75%	4.000000	63.000000
max	5.000000	814.000000

For Indian restaurants, average star rating is 3.75 and average review count is 56



4.0	80
3.5	64
4.5	41
3.0	21
5.0	18
0.0	7
2.5	7
1.5	2
2.0	1





Parkdale	88.888889
Etobicoke	60.000000
Scarborough	58.620690
Mount Pleasant and Davisville	55.555556
Downtown Core	NaN
Leslieville	NaN

Percentage of highly rated restaurants in each area

restaurant_delivery_options	star_rating	
[]	4.0	29
	3.5	26
['Delivery', 'Takeout']	4.0	24
	3.5	16
['Outdoor dining', 'Delivery', 'Takeout']	4.0	13
[]	4.5	11
	3.0	11
['Delivery', 'Takeout']	4.5	10
['Delivery']	3.5	9
	4.5	9

Delivery does not affect rating

rest_name	restaurant_reviews_count
TVX	22
Pukka Pukka	11
Indian Desire	9
Maezo Indian Cuisine	9
Tadka - Sizzling Indian Spices	9

5 star restaurants with most reviews

rest_name	star_rating	restaurant_reviews_count	restaurant_price	restaurant_cuisines	restaurant_area	restaurant_deliver
Byblos	4.5	814	$$$	['Mediterranean', 'Middle Eastern']	Entertainment District	['Delivery',
Maha's	4.0	522	$$	['Egyptian', 'Vegan', 'Sandwiches']	Leslieville	['Outdoor dining',
Lahore Tikka House	3.5	508	$$	['Halal', 'Pakistani']	Leslieville	['Outdoor dining',
Banjara Indian Cuisine	3.5	417	$$	['Indian']	Christie Pits	
Fresh on Spadina	4.0	406	$$	['Vegan', 'Juice Bars & Smoothies', 'Burgers']	Entertainment District	['Proof of v required',
Little India Restaurant	3.5	390	$$	['Indian', 'Buffets']	Queen Street West	['Delivery',
The Halal Guys	3.0	257	$$	['Middle Eastern', 'Halal']	Church-Wellesley Village	['Delivery',
Tibet Kitchen	4.5	245	$$	['Himalayan/Nepalese']	Parkdale	[
Aroma Fine Indian Cuisine	3.5	210	$$	['Indian']	Entertainment District	[
Indian Roti House	4.0	208	$$	['Indian']	Harbourfront	['Delivery',

Top ten Indian restaurants, sorted by number of reviews, then by number of stars

restaurant_price	star_rating	
$$	4.0	32
	3.5	24
$	3.5	12
$$	3.0	9
	4.5	8
$	4.0	6
	4.5	5
$$	2.5	3
$	2.5	2
	3.0	2
$$$	4.0	2
$	1.5	1
$$$	3.5	1
	4.5	1
$$$$	3.5	1

A higher price does not convert into higher rating

$$	76
$	28
$$$	4
$$$$	1

Indian food is reasonable – low to mid price point.

$ sign refers to the price range

Most Popular Restaurant

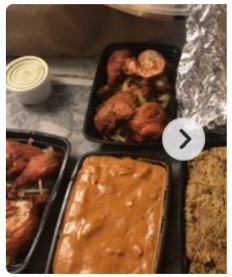
	rest_name	star_rating	restaurant_reviews_count	restaurant_price	restaurant_cuisines	restaurant_area	restaurant_delivery
72	Byblos	4.5	814	$$$	['Mediterranean', 'Middle Eastern']	Entertainment District	['Delivery',
185	Maha's	4.0	522	$$	['Egyptian', 'Vegan', 'Sandwiches']	Leslieville	['Outdoor dining',
28	Lahore Tikka House	3.5	508	$$	['Halal', 'Pakistani']	Leslieville	['Outdoor dining',
17	Banjara Indian Cuisine	3.5	417	$$	['Indian']	Christie Pits	
255	Fresh on Spadina	4.0	406	$$	['Vegan', 'Juice Bars & Smoothies', 'Burgers']	Entertainment District	['Proof of v required',
21	Little India Restaurant	3.5	390	$$	['Indian', 'Buffets']	Queen Street West	['Delivery',
292	The Halal Guys	3.0	257	$$	['Middle Eastern', 'Halal']	Church-Wellesley Village	['Delivery',
57	Tibet Kitchen	4.5	245	$$	['Himalayan/Nepalese']	Parkdale	[
36	Aroma Fine Indian Cuisine	3.5	210	$$	['Indian']	Entertainment District	[
15	Indian Roti House	4.0	208	$$	['Indian']	Harbourfront	['Delivery',

Most mentioned dishes View full menu >

Butter Chicken
22 Photos • 119 Reviews

Palak Paneer
2 Photos • 24 Reviews

Aloo Gobi
1 Photo • 22 Reviews

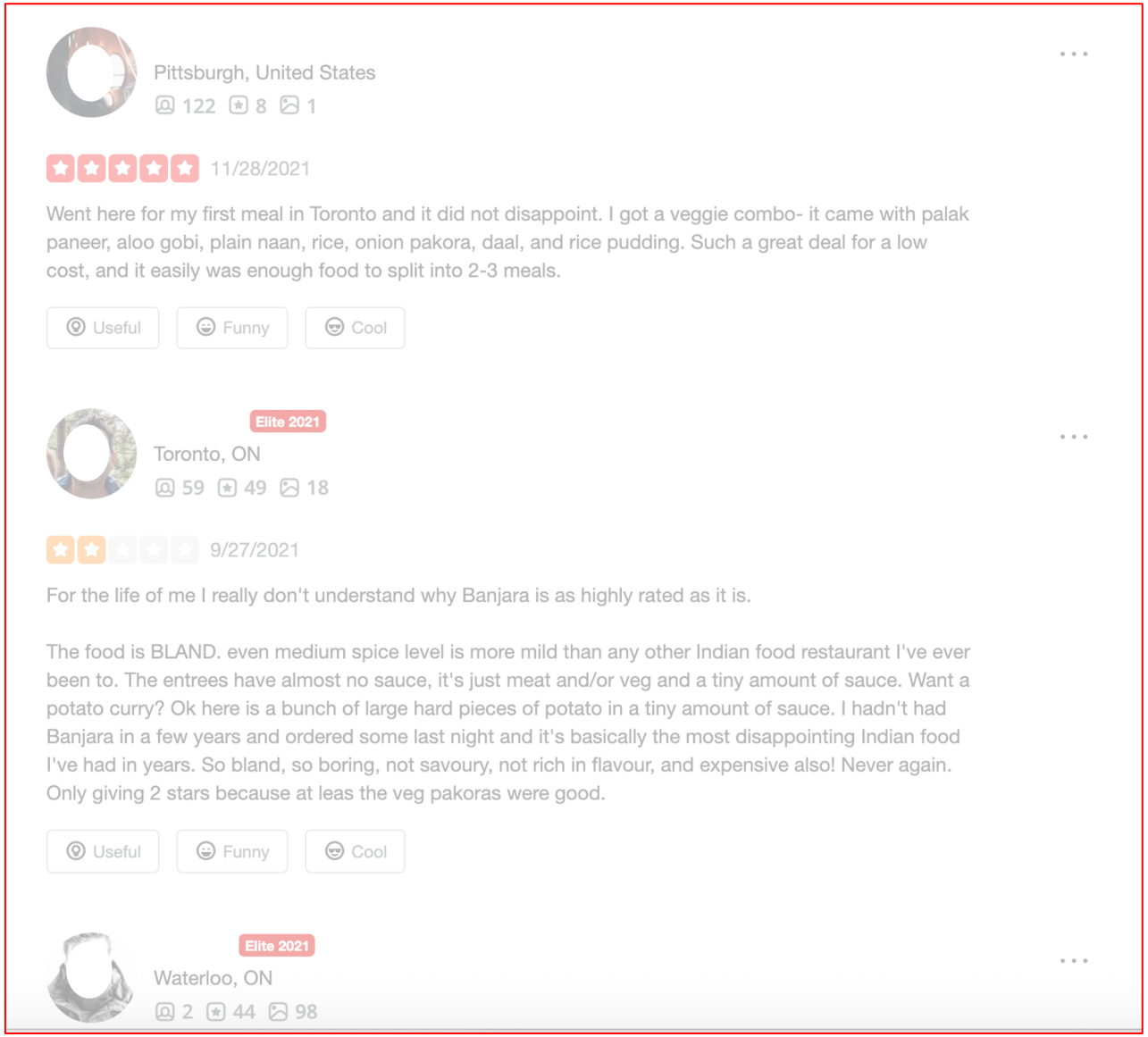


Tandoori Chicke
4 Photos • 13 Revie

dish_name	number_photos	number_reviews
Butter Chicken	22	119
Palak Paneer	2	24
Aloo Gobi	1	22
Tandoori Chicken	4	13
Chicken Korma	1	14

Top five dishes at the most popular Indian restaurant in Toronto

Analysis of reviews of the most popular Indian restaurant, Toronto



Pittsburgh, United States
122 8 1

11/28/2021

Went here for my first meal in Toronto and it did not disappoint. I got a veggie combo- it came with palak paneer, aloo gobi, plain naan, rice, onion pakora, daal, and rice pudding. Such a great deal for a low cost, and it easily was enough food to split into 2-3 meals.

Useful Funny Cool

Elite 2021

Toronto, ON
59 49 18

9/27/2021

For the life of me I really don't understand why Banjara is as highly rated as it is.

The food is BLAND. even medium spice level is more mild than any other Indian food restaurant I've ever been to. The entrees have almost no sauce, it's just meat and/or veg and a tiny amount of sauce. Want a potato curry? Ok here is a bunch of large hard pieces of potato in a tiny amount of sauce. I hadn't had Banjara in a few years and ordered some last night and it's basically the most disappointing Indian food I've had in years. So bland, so boring, not savoury, not rich in flavour, and expensive also! Never again. Only giving 2 stars because at leas the veg pakoras were good.

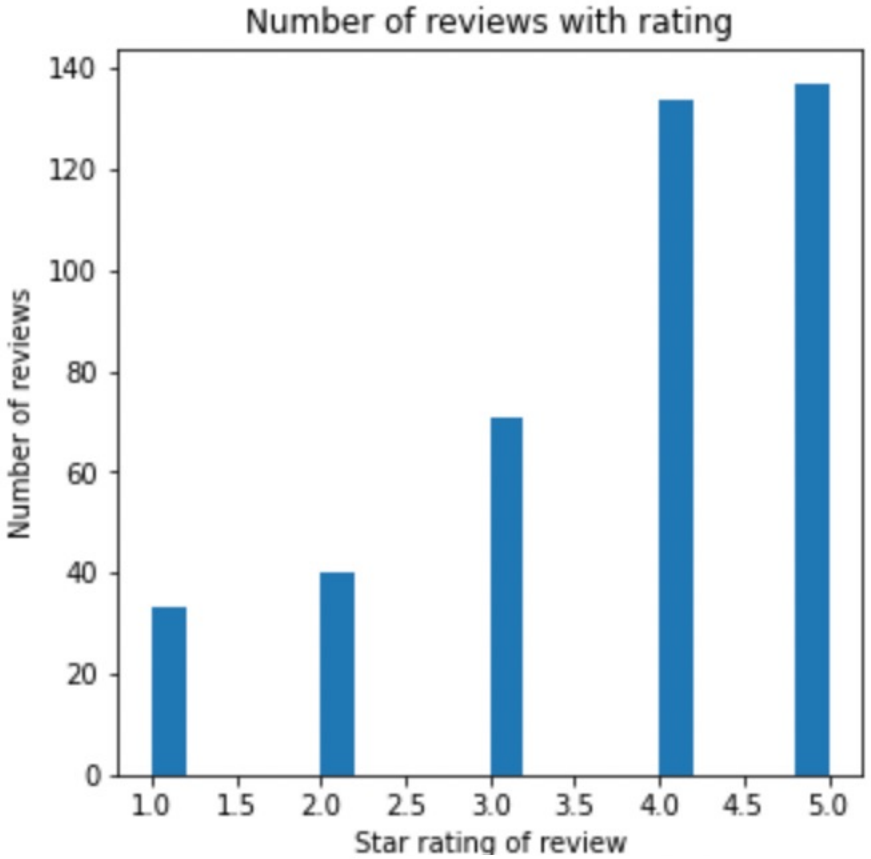
Useful Funny Cool

Elite 2021

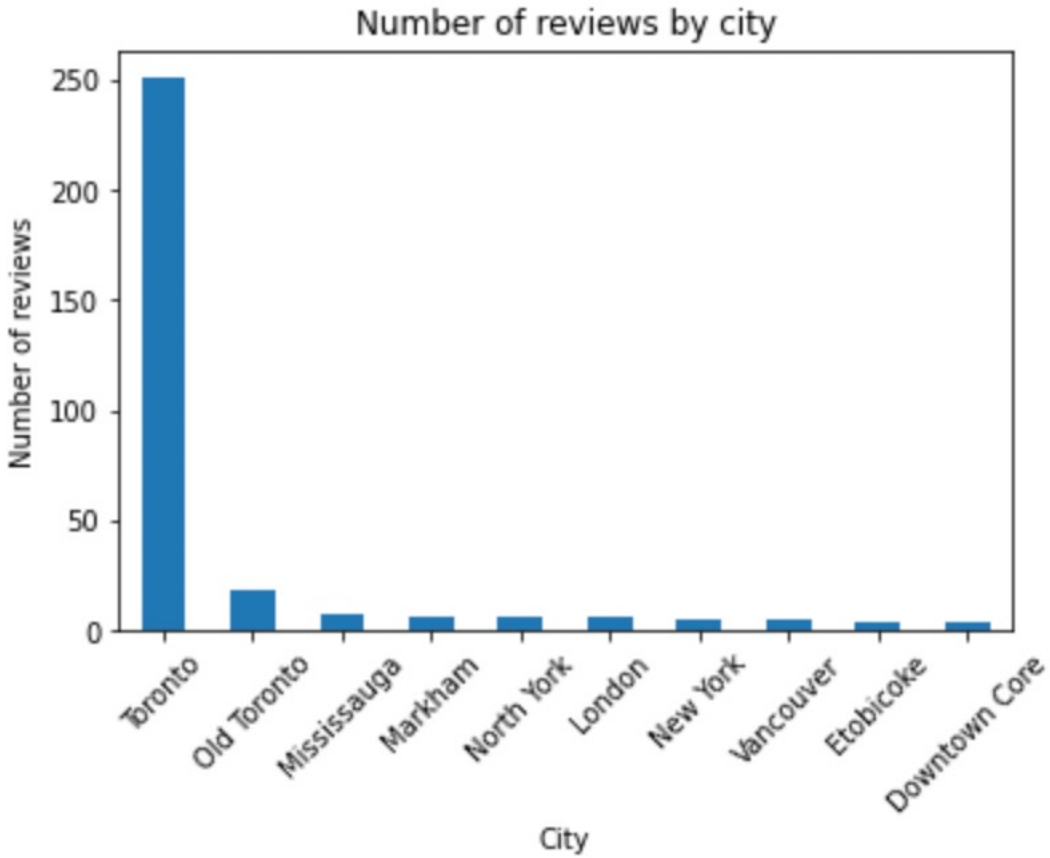
Waterloo, ON
2 44 98

	review_star_rating	review_useful_count	review_funny_count	review_cool_count
count	415.000000	415.000000	415.00000	415.000000
mean	3.727711	0.985542	0.59759	0.375904
std	1.237724	1.784199	1.56486	1.004049
min	1.000000	0.000000	0.00000	0.000000
25%	3.000000	0.000000	0.00000	0.000000
50%	4.000000	0.000000	0.00000	0.000000
75%	5.000000	1.000000	1.00000	0.000000
max	5.000000	13.000000	21.00000	8.000000

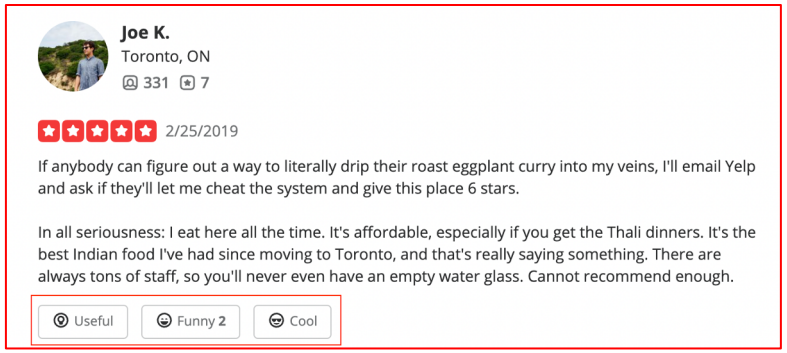
Average star rating of a review is 3.7, min rating is 1 and max is 5



5.0	137
4.0	134
3.0	71
2.0	40
1.0	33



Top ten cities by number of reviews posted



27 Banjara, the perfume of Christie Pits. The thing about Indian food for me, is that I crave it, and when I'm eating it the aromas are beautiful... but the smell of other peoples' vindaloos, curries and paneers often turn me off. So when I lived near Banjara and got a whiff when the wind changed, I wasn't into it... but now that it's not part of my day to day life, man do I want it. I want it all of the time. Their dine in experience, is a red velvet treat, but I usually go for delivery. Delicious, delicious tandoor straight to my door.

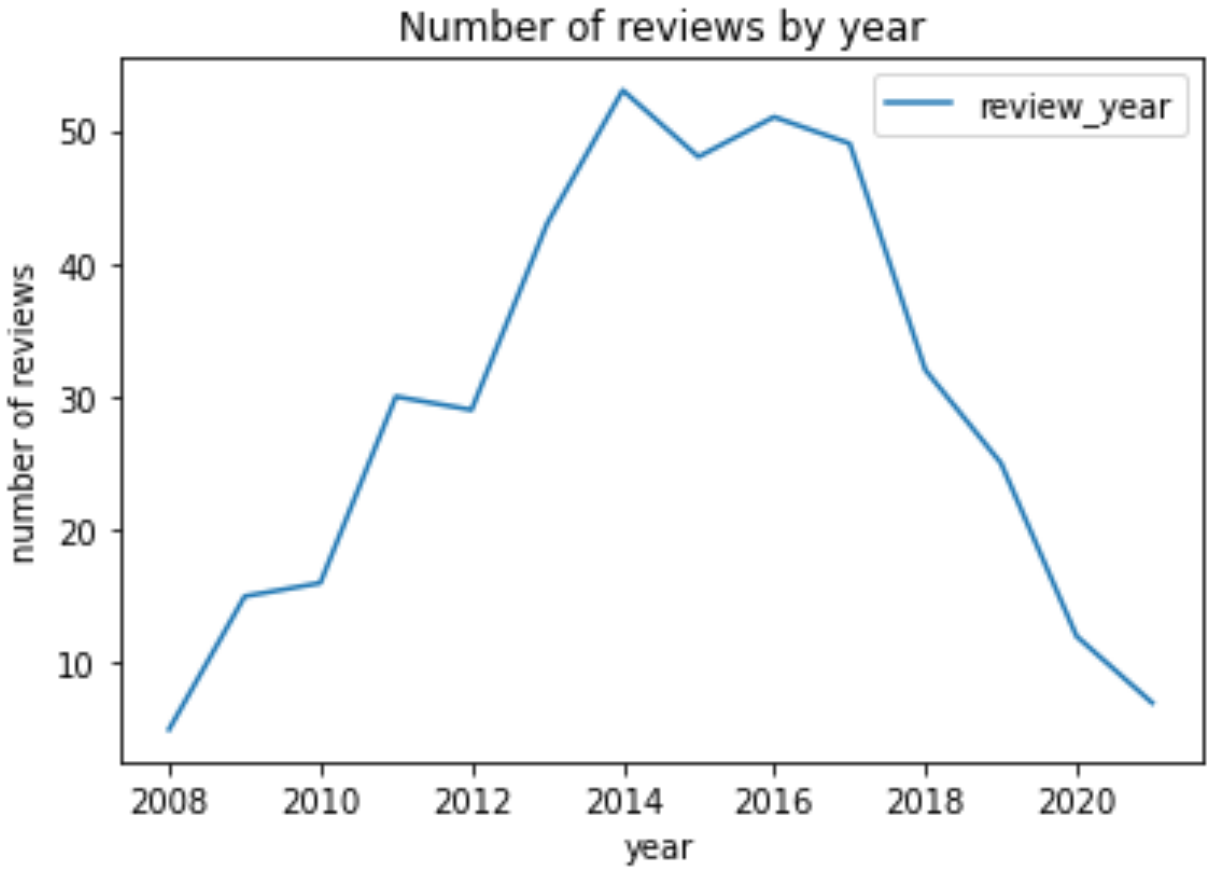
Most useful and cool review

FatCat watch many documentaries... One documentary showing how life is in India.. So oo many people everywhere... No white lines on roads.. no traffic lights anywhere.. people trying to sell you something on every corner... cows walking around streets like its no body business.. Total Chaos for FatCat... FatCat always wonder how country actually still functioning... Well.. Now FatCat get a small taste of it right here in Canada at Bajara restaurant, where entropy is in overdrive...Things FatCat like - Fooooood. Friend asking what they should try when they go.. FatCat say it doesn't matter, because everything is very good. - Have parking lot - but not try to park here if you have SUV or other big car.. Don't ask why.. Just trust FatCat.. Park on street.Things FatCat not like - Total chaos.. Order get messed up by servers.. Place totally rammed like yonge/bloor subway platform at 8:30am on Monday morning.. FatCat trapped behind table and cant go wee.. No one know what the heck going on with food... But FatCat realize that everything work out in the end somehow and everyone walk out or restaurant alive and smiling.. FatCat now understand how busy countries like India still function well, even with all chaos. FatCat learn new way to operate in life... FatCat call this "go with the flow". No more illusion of control over affairs FatCat not able to control.
Then FatCat and crew walk down the street to XO Karaoke (FatCat also do review of XO)

Most funny review

review_text	review_date	review_star_rating	reviewer_city	review_useful_count	review_funny_count	review_cool_count	revie
Went here for my first meal in Toronto and it did not disappoint. I got a veggie combo- it came with palak paneer, aloo gobi, plain naan, rice, onion pakora, daal, and rice pudding. Such a great deal for a low cost, and it easily was enough food to split into 2-3 meals.	2021-11-28	5.0	Pittsburgh	0	0	0	

Latest review was on 28th November,2021, which means people still frequent this restaurant



Number of reviews peaked in 2014 and have never regained. They have been declining since 2017, so Covid is not a factor.

Challenges :

- Thinking of a framework, structure : how to get data from website
- Data Collection and cleaning :
    1. Remove duplicates (sponsored listings)
    2. Understanding HTML tags
    3. How get data from HTML code
    4. Cleaning data : remove xa0, add default value when no data is returned from html page

Next steps

- Do Sentiment analysis – positive and negative reviews
- Do text mining : word cloud, ten most common words, relationship among different words
- Do Topic Modelling