Readme File

In this file, we explain our data source by describing how and from where we collect, clean, and organize the data with step-by-step screenshots. At the end of this report, we test our database by presenting Transaction 4 as stated in our proposal.

- **Collecting Data**

  We collected data using web-scraping with Python. The four main packages we used are Google API, Yelp API, Selenium, and Request. Due to the time limit, we decided to only collect three reviews from each restaurant. Followings are the steps we take:

  A. Using Google API to collect restaurants' information, for restaurants in College Park:

  Code

```
API_KEY = 'AIzaSyALGu4SsQ0B9AQCG8Ch9CTlqTgIAvpWDsg'

gmaps = googlemaps.Client(key=API_KEY)

loc = []

ids = []

stores_info = []


# open file to read road name
with open('CollegeParkRoadName.csv', encoding='utf-8-sig') as csvfile:

    rows = csv.reader(csvfile)
    roadlist = []
    for row in rows:
        roadlist.append(row[0].replace('\xa0',''))

for road in roadlist:
    geocode_result = gmaps.geocode(road)
    loc = geocode_result[0]['geometry']['location']

    for place in gmaps.places_nearby(keyword='restaurant', location= loc, radius= 300)['results']:
        ids.append(place['place_id'])

ids = list(set(ids))

for id in ids:
    stores_info.append(gmaps.place(place_id=id)['result'])

oput = pd.DataFrame.from_dict(stores_info)

oput
```

Results

	address_components	adr_address	business_status	formatted_address	formatted_phone_number	geometry	icon	i
0	[{'long_name': '9905', 'short_name': '9905', '...	9905 Rhode Island...	OPERATIONAL	9905 Rhode Island Ave, College Park, MD 20740,...	(301) 254-9537	{'location': {'lat': 39.0163923, 'lng': -76.92...	https://maps.gstatic.com/mapfiles/place_api/ic...	
1	[{'long_name': '6094', 'short_name': '6094', '...	6094 Greenbelt Rd...	OPERATIONAL	6094 Greenbelt Rd, Greenbelt, MD 20770, USA	(301) 441-3233	{'location': {'lat': 38.999374, 'lng': -76.908...	https://maps.gstatic.com/mapfiles/place_api/ic...	
2	[{'long_name': '8147-D', 'short_name': '8147-D...	8147-D, ...	OPERATIONAL	8147-D, Baltimore Ave, College Park, MD 20740,...	(240) 553-7221	{'location': {'lat': 38.99146700000001, 'lng':...	https://maps.gstatic.com/mapfiles/place_api/ic...	
3	[{'long_name': 'suite b', 'short_name': 'suite...	8321 Baltimore Av...	OPERATIONAL	8321 Baltimore Ave suite b, College Park, MD 2...	(301) 474-4745	{'location': {'lat': 38.9931793, 'lng': -76.93...	https://maps.gstatic.com/mapfiles/place_api/ic...	

B. Using URL in *stores_info_CoolegePark* file to collect more information on Google Maps through web-scraping:

Code

Reading *stores_info_CoolegePark* file:

```
#read file stores_info_CollegePark file
with open('stores_info_CollegePark.json' , 'r',encoding="utf-8") as reader:
    jf = json.loads(reader.read())

unique ={each['url'] : each for each in jf} .values()

df = pd.DataFrame(unique)
df = df[['name','url','user_ratings_total']]

df = df.dropna()
df
```

Extracting URL:

```
#store restaurants' name and URL in lists
name = df["name"].tolist()
link = df["url"].tolist()

rests_name = []
url_link = []


for i in range(len(df)):
    rests_name.append(name[i])
    url_link.append(link[i])

oput_name = pd.DataFrame({'name':rests_name})
oput_link = pd.DataFrame({'url':url_link})
oput_rests = pd.concat([oput_name, oput_link], axis=1)
oput_rests
```

Web-scraping:

```python
#Web-scrapping
def visit():
    driver.get(url)
    sleep(5)
    html = driver.page_source
    soup = BeautifulSoup(html, 'html.parser')

    try:
        title_element = soup.find('div', class_='x3AX1-LfntMc-header-title-ij8cu')
        poiName_ele = title_element.find('h1', class_='x3AX1-LfntMc-header-title-title gm2-headline-5')
        name = poiName_ele.find('span').getText()
        genre = soup.select('button.Yr7JMd-pane-hSRGPd')[1].getText()
        totalReviews_section = title_element.find('span', class_='h0ySl-wcwwM-E70qVe-list')

        totalNumReview_text = totalReviews_section.find_all('button')[0].getText()
        print(totalNumReview_text)
        if "(" in totalNumReview_text and ")" in totalNumReview_text:
            totalNumReview = int(re.sub('[(),]', '', totalNumReview_text))
        else:
            totalNumReview = int(re.sub('[,]', '', totalNumReview_text.split(" ")[0]))


        listPost.append({'URL':url,
                         'Restaurant':name,
                         'Genre': genre,
                         'Total_reviews': totalNumReview,
                         'SeriviceOption':[],
                         'Review':[]
                        })


        for i, element in enumerate(soup.select('div.uxOu9-sTGRBb-p83tee')):
                listPost[len(listPost)-1]['SeriviceOption'].append({'Index':i+1,
                                                                    'SeriviceOption':element['aria-label']})


    except:
        print(f'unable to collect poi information on {url}')
        traceback.print_exc()

    pag.moveTo(200, 400)
    pag.scroll(-1000)
    sleep(10)
```

```python
    try:
        div = driver.find_elements(By.CSS_SELECTOR,'div.ODSEW-ShBeI.NIyLF-haAclf.gm2-body-2')
        soup = BeautifulSoup(driver.page_source, 'html.parser')

        stars=[]
        names=[]

        for a in soup.select('span.ODSEW-ShBeI-H1e3jb'):
            stars.append(a['aria-label'])

        for b in soup.select('div.ODSEW-ShBeI-title'):
            names.append(b)

        for i, element in enumerate(soup.select('div.ODSEW-ShBeI.NIyLF-haAclf.gm2-body-2')):

            Content = element.find('span',{'class':'ODSEW-ShBeI-text'})
            Like = element.find('button',{'class':'ODSEW-ShBeI-Sc2xXc-LgbsSe'})
            rvNumber = element.find('div',{'class':'ODSEW-ShBeI-VdSJob'}).find_all('span')


            listPost[len(listPost)-1]['Review'].append({'Index':i+1,
                                'Author_name':names[i].text,
                                'Review_number':rvNumber[1].getText(),
                                'Stars':stars[i],
                                'Time':element.find('span',{'class':'ODSEW-ShBeI-RgZmSc-date'}).text,
                                'Content':Content.text,
                                'Review_like':Like['aria-label'] })

    except Exception:
        traceback.print_exc()
        pass
```

```python
def end():
    driver.quit()


options = webdriver.ChromeOptions()
driver = webdriver.Chrome(executable_path='/Users/fanyunjung/Desktop/BA/Python/Google/chromedriver',options = options)


if __name__ == '__main__':

    listPost = []

    for url in url_link:
        visit()

    end()

    with open("google_review_CollegePark_rvNumber.json", "w",encoding="utf-8") as f:
        json.dump(listPost, f, ensure_ascii=False, indent=4)
```

- **Cleaning & Organizing Data**

  After collecting the needed information, in the next step, we began to clean and combine data from different sources by using Python, and further format the data as required in our database. Followings are the python code:

Code

In [1]:
```
import pandas as pd
import json
```

In [2]:
```
ty=pd.read_json('google_review_CollegePark_rvNumber.json')
```

In [3]:
```
ind1 = ty.columns
ind1 = ind1.delete([-1])
#ind1 = ind1.delete([-1])
ind1
```

Out[3]:
```
Index(['URL', 'Restaurant', 'Genre', 'Total_reviews', 'SeriviceOption'], dtype='object')
```

In [4]:
```
# ind2 = ty.iloc[0]['SeriviceOption'][0].keys()
# ind2 = pd.Index(ind2)
# ind2
```

In [5]:
```
ind3 = ty.iloc[0]['Review'][0].keys()
ind3 = pd.Index(ind3)
ind3
```

Out[5]:
```
Index(['Index', 'Author_name', 'Review_number', 'Stars', 'Time', 'Content',
       'Review_like'],
      dtype='object')
```

In [6]:
```
# save index to pass in expanded dataframe below
#ind = ind1.append(ind2)
ind = ind1.append(ind3)
ind
```

Out[6]:
```
Index(['URL', 'Restaurant', 'Genre', 'Total_reviews', 'SeriviceOption',
       'Index', 'Author_name', 'Review_number', 'Stars', 'Time', 'Content',
       'Review_like'],
      dtype='object')
```

In [7]:
```
# supposed row num in expanded daraframe
ty['Total_reviews'].sum()
```

Out[7]:
```
64592
```

In [8]:
```
# intertuples to interate rows in a dataframe

expand = pd.DataFrame([(tup.URL, tup.Restaurant, tup.Genre, tup.Total_reviews, tup.SeriviceOption, rev['Index'],
                        rev['Author_name'], rev['Review_number'], rev['Stars'], rev['Time'], rev['Content'],
                        rev['Review_like']) \
                          for tup in ty.itertuples() for rev in tup.Review], columns = ind)

expand.count() # rows in expanded df should be less than total_rev sum (above cell), not more
```

Out[8]:
```
URL               401
Restaurant        401
Genre             401
Total_reviews     401
SeriviceOption    401
Index             401
Author_name       401
Review_number     401
Stars             401
Time              401
Content           401
Review_like       401
dtype: int64
```

In [9]:
```
# expanded dataframe
expand.tail()
```

Out[9]:

	URL	Restaurant	Genre	Total_reviews	SeriviceOption	Index	Author_name	Review_number	Stars	Time	Content	Review
396	https://maps.google.com/?cid=141422348181345645	Krazi Kebob	Pakistani restaurant	410	[{'Index': 1, 'SeriviceOption': 'Offers takeou...	2	Richard Tan	2 reviews	5 stars	2 weeks ago	Got the chicken naan wrap and the chicken naan...	Ric Tan's re as he
397	https://maps.google.com/?cid=141422348181345645	Krazi Kebob	Pakistani restaurant	410	[{'Index': 1, 'SeriviceOption': 'Offers takeou...	3	fahd majiduddin	· 95 reviews	5 stars	a week ago	This place is awesome despite its location. T...	Mark majidud revie he
398	https://maps.google.com/?cid=7380797848103903593	MISSION BBQ	Barbecue restaurant	1038	[{'Index': 1, 'SeriviceOption': 'Serves dine-i...	1	Brittany Jeffries	· 131 reviews	5 stars	3 months ago	Very clean restaurant, fast service and tasty ...	1 pe f Bri Jeffi review

In [10]:
```
expand_woSeriviceOption = expand.drop(['SeriviceOption'], axis=1)
expand_woSeriviceOption
```

Out[10]:

	URL	Restaurant	Genre	Total_reviews	Index	Author_name	Review_number	Stars	Time	Content	Review_like
0	https://maps.google.com/?cid=2486673317011209047	Nene's Restaurant	Restaurant	22	1	E A	· 50 reviews	5 stars	a month ago	Absolutely delicious food, the jollof rice, eg...	Mark E A's review as helpful.
1	https://maps.google.com/?cid=2486673317011209047	Nene's Restaurant	Restaurant	22	2	Ray Weil	· 40 reviews	5 stars	4 months ago	I drove by this place several times during the...	3 people found Ray Weil's review helpful. Mar...
2	https://maps.google.com/?cid=2486673317011209047	Nene's Restaurant	Restaurant	22	3	Melissa Stefun	· 64 reviews	5 stars	4 months ago	Great food! I don't have much experience with ...	Mark Melissa Stefun's review as helpful.
3	https://maps.google.com/?cid=6705066049573664365	Subway	Sandwich shop	77	1	Jenny Deer	6 reviews	1 star	5 months ago	This is the worst subway I ordered a salad and...	Mark Jenny Deer's review as helpful.
4	https://maps.google.com/?cid=6705066049573664365	Subway	Sandwich shop	77	2	Javar Doogles	· 85 reviews	5 stars	8 months ago	This subway is the best\nMy Chipotle steak and...	Mark Javar Doogles's review as helpful.
...	...	...	...	...	...	...	...	...	...	...	...
396	https://maps.google.com/?cid=141422348181345645	Krazi Kebob	Pakistani restaurant	410	2	Richard Tan	2 reviews	5 stars	2 weeks ago	Got the chicken naan wrap and the chicken naan...	Mark Richard Tan's review as helpful.

In [11]:
```
expand_woSeriviceOption['Review_like'] = expand_woSeriviceOption['Review_like'].str[1:2]
expand_woSeriviceOption['Review_like'] = expand_woSeriviceOption['Review_like'].replace(['M'],'0')
expand_woSeriviceOption
```

Out[11]:

	URL	Restaurant	Genre	Total_reviews	Index	Author_name	Review_number	Stars	Time	Content	Review_like
0	https://maps.google.com/?cid=2486673317011209047	Nene's Restaurant	Restaurant	22	1	E A	· 50 reviews	5 stars	a month ago	Absolutely delicious food, the jollof rice, eg...	0
1	https://maps.google.com/?cid=2486673317011209047	Nene's Restaurant	Restaurant	22	2	Ray Weil	· 40 reviews	5 stars	4 months ago	I drove by this place several times during the...	3
2	https://maps.google.com/?cid=2486673317011209047	Nene's Restaurant	Restaurant	22	3	Melissa Stefun	· 64 reviews	5 stars	4 months ago	Great food! I don't have much experience with ...	0
3	https://maps.google.com/?cid=6705066049573664365	Subway	Sandwich shop	77	1	Jenny Deer	6 reviews	1 star	5 months ago	This is the worst subway I ordered a salad and...	0
4	https://maps.google.com/?cid=6705066049573664365	Subway	Sandwich shop	77	2	Javar Doogles	· 85 reviews	5 stars	8 months ago	This subway is the best\nMy Chipotle steak and...	0
...	...	...	...	...	...	...	...	...	...	...	...
396	https://maps.google.com/?cid=141422348181345645	Krazi Kebob	Pakistani restaurant	410	2	Richard Tan	2 reviews	5 stars	2 weeks ago	Got the chicken naan wrap and the chicken naan...	0

In [12]:
```
SeriviceOption = (expand.set_index(['URL','Restaurant'])['SeriviceOption']
                  .apply(pd.Series).stack().apply(pd.Series).reset_index().drop('level_2',1))
SeriviceOption = SeriviceOption.drop_duplicates()
SeriviceOption
```

Out[12]:

	URL	Restaurant	Index	SeriviceOption
0	https://maps.google.com/?cid=2486673317011209047	Nene's Restaurant	1	Serves dine-in
1	https://maps.google.com/?cid=2486673317011209047	Nene's Restaurant	2	Offers takeout
6	https://maps.google.com/?cid=6705066049573664365	Subway	1	Serves dine-in
7	https://maps.google.com/?cid=6705066049573664365	Subway	2	Offers takeout
8	https://maps.google.com/?cid=6705066049573664365	Subway	3	Offers delivery
...	...	...	...	...
1137	https://maps.google.com/?cid=14142234818134564 5	Krazi Kebob	1	Offers takeout
1138	https://maps.google.com/?cid=14142234818134564 5	Krazi Kebob	2	Offers delivery
1143	https://maps.google.com/?cid=7380797848103903593	MISSION BBQ	1	Serves dine-in
1144	https://maps.google.com/?cid=7380797848103903593	MISSION BBQ	2	Offers curbside pickup
1145	https://maps.google.com/?cid=7380797848103903593	MISSION BBQ	3	Offers delivery

385 rows × 4 columns

In [13]:
```
# save SeriviceOption
with open ('SeriviceOption.json', 'w', encoding='utf-8') as file:
    SeriviceOption.to_json(file, indent=4, force_ascii=False, orient='records')
```

- **Testing**

  Lastly, to test our database, we use Tableau to present Transaction 4. Followings are the results:

  A. Using the statement in Transaction 4, we can see all the reviews at first:

What are the reviews that come with the highest review like?

Res ID	Res Name	Rev Content	Rev Star Rate	Rev Like
R08	KOITÉ GRILL	Let me first say, that we did not visit this restaurant location. But, we stopped by Koite Grill Stand at RFK Stadium Open Air Farmers' Market in DC and we ordered Lamb, Chicken with JOLLOF Rice, Purple Tea and a upside down Pineapple ...	5.000	8.000
R17	Ledo Pizza	First and foremost, this is NO longer the original Ledo(s). This is now a chain Ledo(s) restaurant, plain and simple. Now if that's your thing, good on you. But those of us who patronized the original and have eaten at the chains, know the ...	2.000	5.000
R11	Azteca Restaurant & Cantina	Delicious, inexpensive food! The food is great! The drinks are a little on the strong side (for me), but still good! The service here was a little slow even though there weren't many people in there, but they may have had a lot of carry out ...	5.000	4.000
R01	Nene's Restaurant	I drove by this place several times during the pandemic in 2020 but didn't stop in until 2021. Now I'm hooked...a regular, so to speak. Delicious Nigeria food cooked to perfection. I enjoyed the spicy goat meat with Jollof rice and ...	5.000	3.000
R06	The Jerk Pit	There are a rack (a lot) of Jamaican culinary spots in the DC area. I've definitely reviewed a couple of the more popular eateries in the area...except this one @jerkpit !! Located at 9078 Baltimore Ave., College Park M.D., this place can be ...	5.000	3.000
R25	College Park Diner	Clean place, friendly, good food and service. You won't be disappointed.	5.000	3.000
R52	LaTao Hotpot College park	Lovely place and delicious food. Very entertaining to cook the food yourself, mix the sauces and experiment. The staff is very friendly and quick. They also have karaoke rooms, in case you love to sing and want to spend couple of hours ...	5.000	3.000

Restaurant Name
(All)

  B. If the users want to check a specific restaurant's review with the highest review like, then we can use the dropdown menu to filter the information:

What are the reviews that come with the highest review like?

Res ID	Res Name	Rev Content	Rev Star Rate	Rev Like
R73	sweetgreen	The audacity for this location to charge $2 to "add hot roasted sweet potatoes" and put three (3) pieces, below lukewarm ...	2.000	2.000

Restaurant Name

sweetgreen