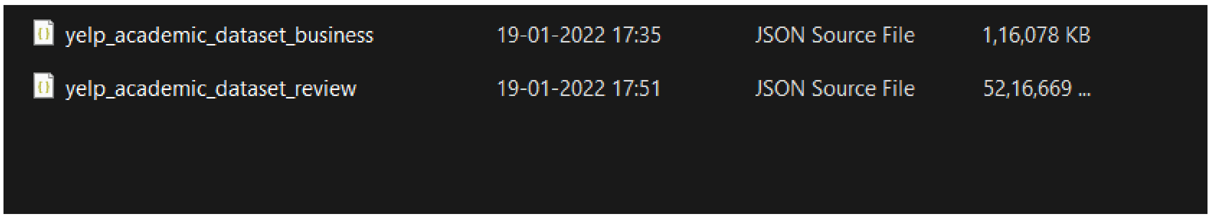
Lab 3_4

CIS612 Big Data and Parallel data processing systems

Semi-Structured Data Processing with MongoDB

Create a database with two Collections from the given Yelp Business.json and Review.json data files in the semi-structured database server MongoDB
For this lab, you can use 5 JSON files in YelpDataSets.zip from the Yelp site (data set in 2017) in the CIS612 Lab section. Or you can directly download from the Yelp site at https://www.yelp.com/dataset



yelp_academic_dataset_business	19-01-2022 17:35	JSON Source File	1,16,078 KB
yelp_academic_dataset_review	19-01-2022 17:51	JSON Source File	52,16,669 ...

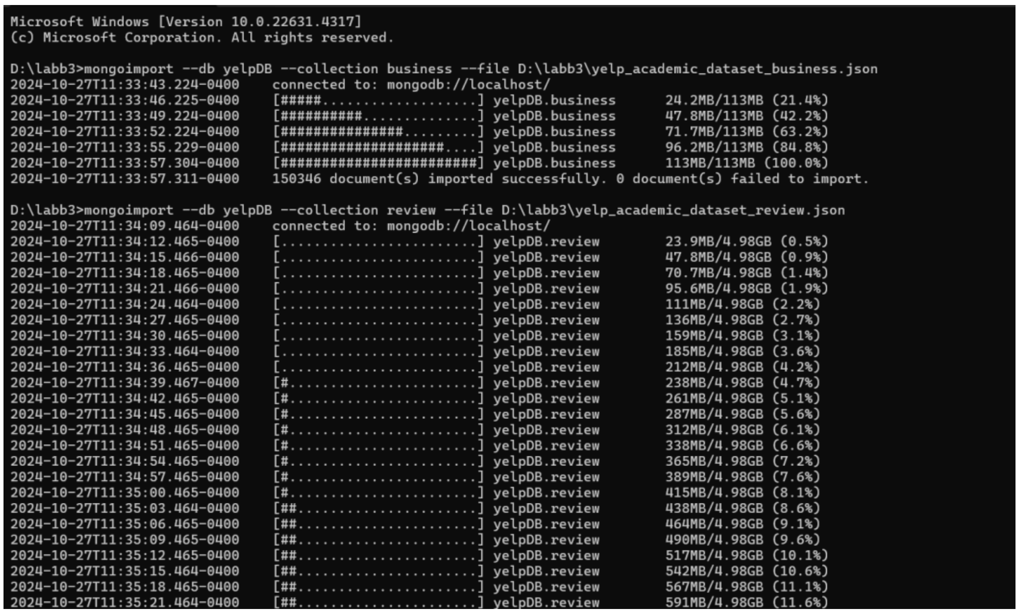
Here is yelpdata sets required

Creating Two Collections in MongoDB and Write a Data Pipelining to Retrieval Analytic Information Using Aggregate Pipelining and Join
1. Import Business.json and Review.json Data files from the Yelp site into MongoDB to create a Collection from each json file named business and review respectively.
2. Find the following information Using MongoDB Aggregation Pipelining:
Q1: For those business whose category is either "Fast Food" or "Restaurants", count the number of the business by each "city" and "stars".
"by each city and stars" means Group By city, stars in SQL.
Q2_1: For those business whose category is either "Fast Food" or "Restaurants" and review_count > 10 and stars >= 4, find all the reviews with stars >= 4 by performing Join with two Collections business and review.

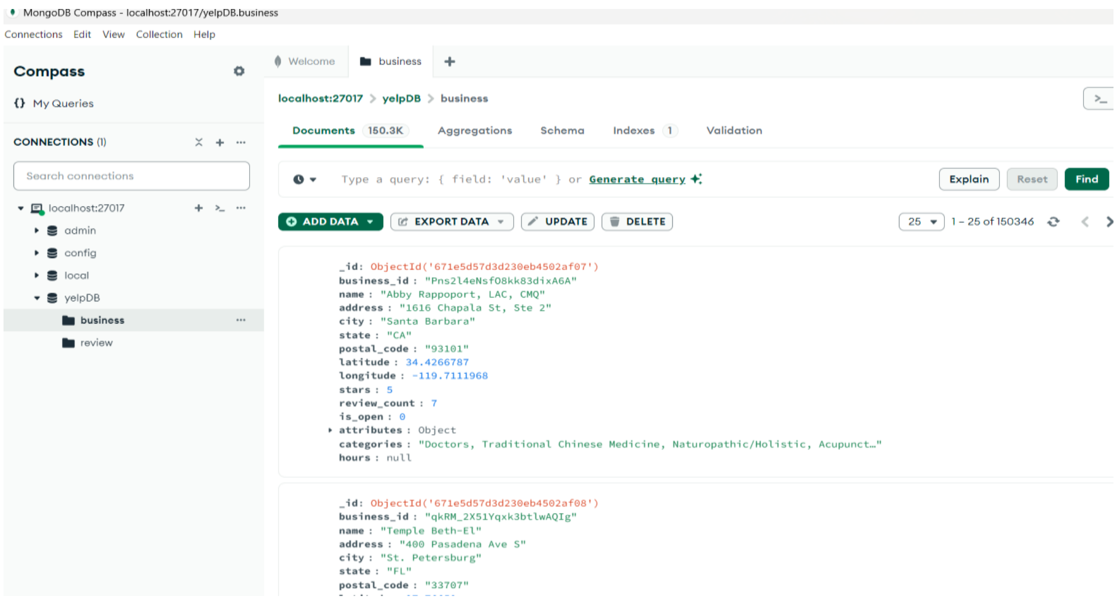
MongoDB query to retrieve the data from JSON files

bash

```
mongoimport --db yelpDB --collection business --file D:\labb3\yelp_academic_dataset_business.json

mongoimport --db yelpDB --collection review --file D:\labb3\yelp_academic_dataset_review.json
```



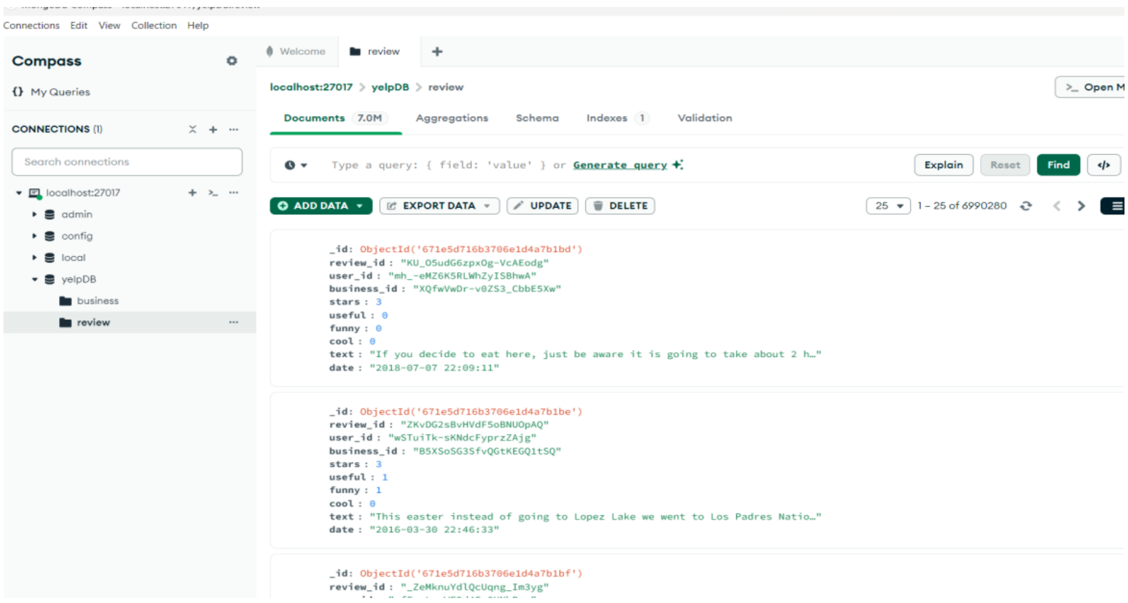
```
Microsoft Windows [Version 10.0.22631.4317]
(c) Microsoft Corporation. All rights reserved.

D:\labb3>mongoimport --db yelpDB --collection business --file D:\labb3\yelp_academic_dataset_business.json
2024-10-27T11:33:43.224-0400    connected to: mongodb://localhost/
2024-10-27T11:33:46.225-0400    [#####...................] yelpDB.business      24.2MB/113MB (21.4%)
2024-10-27T11:33:49.224-0400    [##########..............] yelpDB.business      47.8MB/113MB (42.2%)
2024-10-27T11:33:52.224-0400    [###############.........] yelpDB.business      71.7MB/113MB (63.2%)
2024-10-27T11:33:55.229-0400    [####################....] yelpDB.business      96.2MB/113MB (84.8%)
2024-10-27T11:33:57.304-0400    [########################] yelpDB.business      113MB/113MB (100.0%)
2024-10-27T11:33:57.311-0400    150346 document(s) imported successfully. 0 document(s) failed to import.

D:\labb3>mongoimport --db yelpDB --collection review --file D:\labb3\yelp_academic_dataset_review.json
2024-10-27T11:34:09.464-0400    connected to: mongodb://localhost/
2024-10-27T11:34:12.465-0400    [........................] yelpDB.review        23.9MB/4.98GB (0.5%)
2024-10-27T11:34:15.466-0400    [........................] yelpDB.review        47.8MB/4.98GB (0.9%)
2024-10-27T11:34:18.465-0400    [........................] yelpDB.review        70.7MB/4.98GB (1.4%)
2024-10-27T11:34:21.466-0400    [........................] yelpDB.review        95.6MB/4.98GB (1.9%)
2024-10-27T11:34:24.464-0400    [........................] yelpDB.review        111MB/4.98GB (2.2%)
2024-10-27T11:34:27.465-0400    [........................] yelpDB.review        136MB/4.98GB (2.7%)
2024-10-27T11:34:30.465-0400    [........................] yelpDB.review        159MB/4.98GB (3.1%)
2024-10-27T11:34:33.464-0400    [........................] yelpDB.review        185MB/4.98GB (3.6%)
2024-10-27T11:34:36.465-0400    [........................] yelpDB.review        212MB/4.98GB (4.2%)
2024-10-27T11:34:39.467-0400    [#.......................] yelpDB.review        238MB/4.98GB (4.7%)
2024-10-27T11:34:42.465-0400    [#.......................] yelpDB.review        261MB/4.98GB (5.1%)
2024-10-27T11:34:45.465-0400    [#.......................] yelpDB.review        287MB/4.98GB (5.6%)
2024-10-27T11:34:48.465-0400    [#.......................] yelpDB.review        312MB/4.98GB (6.1%)
2024-10-27T11:34:51.465-0400    [#.......................] yelpDB.review        338MB/4.98GB (6.6%)
2024-10-27T11:34:54.465-0400    [#.......................] yelpDB.review        365MB/4.98GB (7.2%)
2024-10-27T11:34:57.465-0400    [#.......................] yelpDB.review        389MB/4.98GB (7.6%)
2024-10-27T11:35:00.465-0400    [#.......................] yelpDB.review        415MB/4.98GB (8.1%)
2024-10-27T11:35:03.464-0400    [##......................] yelpDB.review        438MB/4.98GB (8.6%)
2024-10-27T11:35:06.465-0400    [##......................] yelpDB.review        464MB/4.98GB (9.1%)
2024-10-27T11:35:09.465-0400    [##......................] yelpDB.review        490MB/4.98GB (9.6%)
2024-10-27T11:35:12.465-0400    [##......................] yelpDB.review        517MB/4.98GB (10.1%)
2024-10-27T11:35:15.464-0400    [##......................] yelpDB.review        542MB/4.98GB (10.6%)
2024-10-27T11:35:18.465-0400    [##......................] yelpDB.review        567MB/4.98GB (11.1%)
2024-10-27T11:35:21.464-0400    [##......................] yelpDB.review        591MB/4.98GB (11.6%)
```



MongoDB Compass - localhost:27017/yelpDB.business

localhost:27017 > yelpDB > business — Documents 150.3K — 1 – 25 of 150346

```
_id: ObjectId('671e5d57d3d230eb4502af07')
business_id : "Pns2l4eNsfO8kk83dixA6A"
name : "Abby Rappoport, LAC, CMQ"
address : "1616 Chapala St, Ste 2"
city : "Santa Barbara"
state : "CA"
postal_code : "93101"
latitude : 34.4266787
longitude : -119.7111968
stars : 5
review_count : 7
is_open : 0
attributes : Object
categories : "Doctors, Traditional Chinese Medicine, Naturopathic/Holistic, Acupunct…"
hours : null

_id: ObjectId('671e5d57d3d230eb4502af08')
business_id : "qkRM_2X51Yqxk3btlwAQIg"
name : "Temple Beth-El"
address : "400 Pasadena Ave S"
city : "St. Petersburg"
state : "FL"
postal_code : "33707"
```



localhost:27017 > yelpDB > review — Documents 7.0M — 1 – 25 of 6990280

```
_id: ObjectId('671e5d716b3706e1d4a7b1bd')
review_id : "KU_O5udG6zpxOg-VcAEodg"
user_id : "mh_-eMZ6K5RLWhZyISBhwA"
business_id : "XQfwVwDr-v0ZS3_CbbE5Xw"
stars : 3
useful : 0
funny : 0
cool : 0
text : "If you decide to eat here, just be aware it is going to take about 2 h…"
date : "2018-07-07 22:09:11"

_id: ObjectId('671e5d716b3706e1d4a7b1be')
review_id : "ZKvDG2sBvHVdF5oBNUOpAQ"
user_id : "wSTuiTk-sKNdcFyprzZAjg"
business_id : "B5XSoSG3SfvQGtKEGQ1tSQ"
stars : 3
useful : 1
funny : 1
cool : 0
text : "This easter instead of going to Lopez Lake we went to Los Padres Natio…"
date : "2016-03-30 22:46:33"

_id: ObjectId('671e5d716b3706e1d4a7b1bf')
review_id : "_ZeMknuYdlQcUqng_Im3yg"
```

Importing the Business.josn and Review.json into MongoDB server

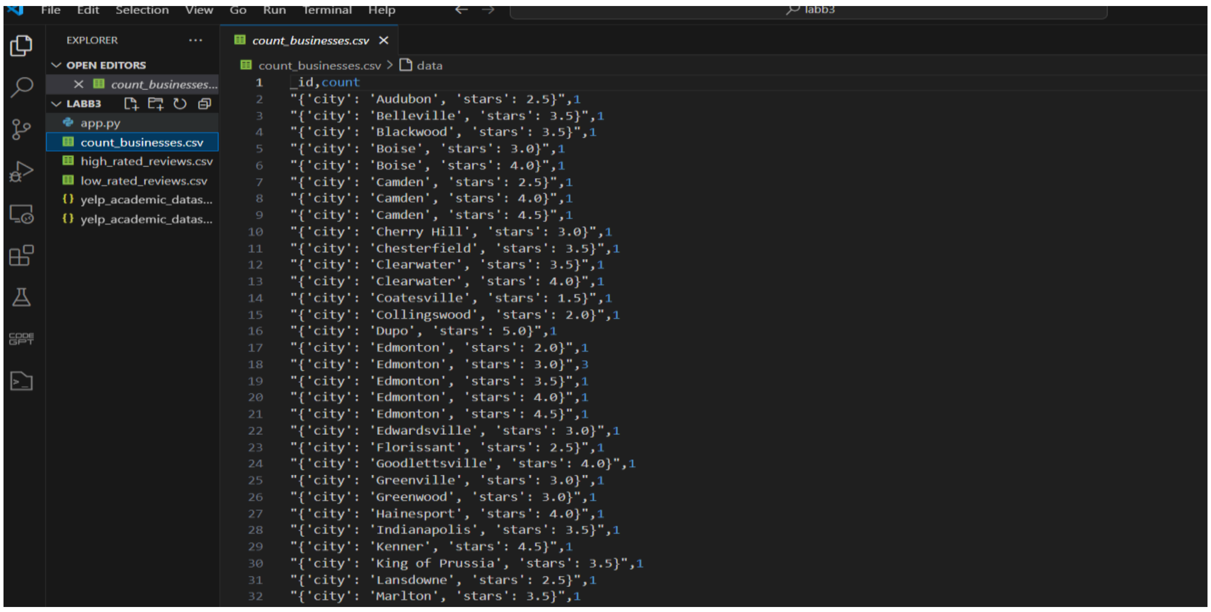
Python code:

```
from flask import Flask, jsonify, request, make_response
from pymongo import MongoClient
import pandas as pd

app = Flask(__name__)
client = MongoClient("mongodb://localhost:27017/")
db = client["yelpDB"]
business_collection = db["business"]
review_collection = db["review"]

@app.route('/count_businesses', methods=['GET'])
def count_businesses():
    pipeline = [
        {"$match": {"categories": {"$in": ["Fast Food", "Restaurants"]}}},
        {"$group": {"_id": {"city": "$city", "stars": "$stars"}, "count": {"$sum": 1}}},
        {"$sort": {"_id.city": 1, "_id.stars": 1}}
    ]
    result = list(business_collection.aggregate(pipeline))
    if request.args.get('format') == 'csv':
        return convert_to_csv(result, 'count_businesses')
    return jsonify(result)

@app.route('/high_rated_reviews', methods=['GET'])
def high_rated_reviews():
    pipeline = [
        {"$match": {"categories": {"$in": ["Fast Food", "Restaurants"]}, "review_count": {"$gt": 10}, "stars":
{"$gte": 4}}},
        {"$lookup": {
            "from": "review",
            "localField": "business_id",
            "foreignField": "business_id",
            "as": "high_reviews"
        }},
        {"$unwind": "$high_reviews"},
        {"$match": {"high_reviews.stars": {"$gte": 4}}},
         {"$project": {"review_id": "$high_reviews.review_id", "business_id": "$business_id", "stars":
"$high_reviews.stars", "review_text": "$high_reviews.text"}}
    ]
    result = list(business_collection.aggregate(pipeline))
    for document in result:
        if "_id" in document:
            document["_id"] = str(document["_id"])
        if "review_id" in document:
            document["review_id"] = str(document["review_id"])
```

```python
        if "business_id" in document:
            document["business_id"] = str(document["business_id"])
    if request.args.get('format') == 'csv':
        return convert_to_csv(result, 'high_rated_reviews')
    return jsonify(result)

@app.route('/low_rated_reviews', methods=['GET'])
def low_rated_reviews():
    pipeline = [
        {"$match": {"categories": {"$in": ["Fast Food", "Restaurants"]}, "review_count": {"$gt": 10}, "stars":
{"$lt": 2}}},
        {"$lookup": {
            "from": "review",
            "localField": "business_id",
            "foreignField": "business_id",
            "as": "low_reviews"
        }},
        {"$unwind": "$low_reviews"},
        {"$match": {"low_reviews.stars": {"$lt": 2}}},
        {"$project": {"review_id": "$low_reviews.review_id", "business_id": "$business_id", "stars":
"$low_reviews.stars", "review_text": "$low_reviews.text"}}
    ]
    result = list(business_collection.aggregate(pipeline))
    for document in result:
        if "_id" in document:
            document["_id"] = str(document["_id"])
        if "review_id" in document:
            document["review_id"] = str(document["review_id"])
        if "business_id" in document:
            document["business_id"] = str(document["business_id"])
    if request.args.get('format') == 'csv':
        return convert_to_csv(result, 'low_rated_reviews')
    return jsonify(result)

def convert_to_csv(data, filename):
    df = pd.DataFrame(data)
    print(df.head())  # Debug: print the first few rows of the DataFrame
    response = make_response(df.to_csv(index=False))
    response.headers["Content-Disposition"] = f"attachment; filename={filename}.csv"
    response.headers["Content-Type"] = "text/csv"
    return response

if __name__ == '__main__':
    app.run()
```

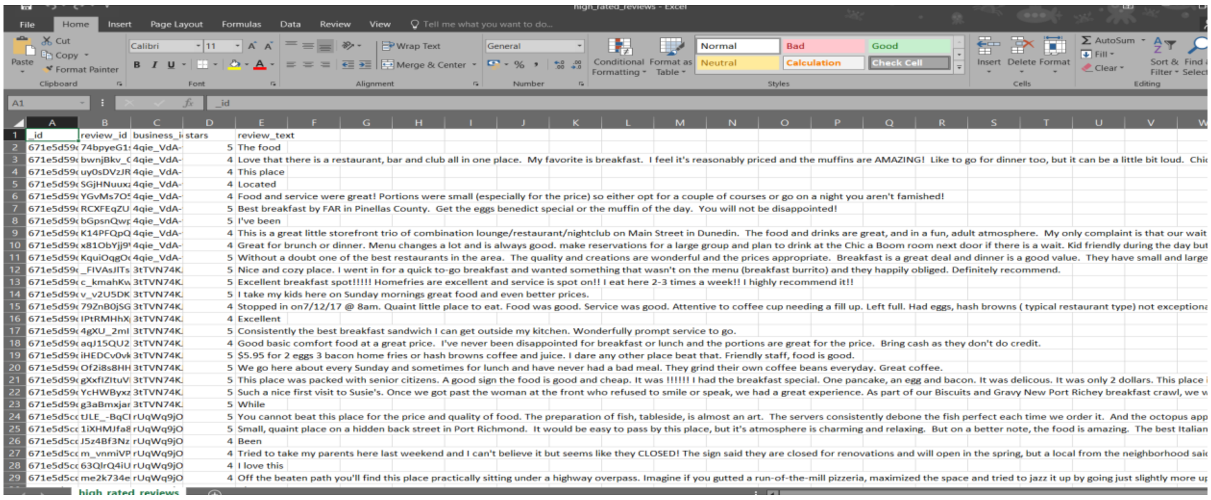
Getting the files in CSV format:

Count_business:



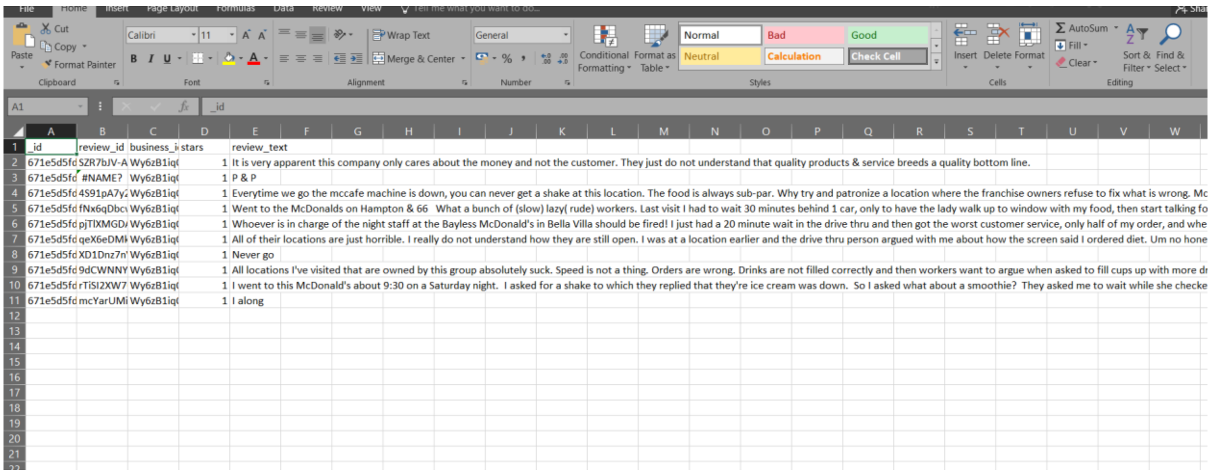
```
_id,count
"{'city': 'Audubon', 'stars': 2.5}",1
"{'city': 'Belleville', 'stars': 3.5}",1
"{'city': 'Blackwood', 'stars': 3.5}",1
"{'city': 'Boise', 'stars': 3.0}",1
"{'city': 'Boise', 'stars': 4.0}",1
"{'city': 'Camden', 'stars': 2.5}",1
"{'city': 'Camden', 'stars': 4.0}",1
"{'city': 'Camden', 'stars': 4.5}",1
"{'city': 'Cherry Hill', 'stars': 3.0}",1
"{'city': 'Chesterfield', 'stars': 3.5}",1
"{'city': 'Clearwater', 'stars': 3.5}",1
"{'city': 'Clearwater', 'stars': 4.0}",1
"{'city': 'Coatesville', 'stars': 1.5}",1
"{'city': 'Collingswood', 'stars': 2.0}",1
"{'city': 'Dupo', 'stars': 5.0}",1
"{'city': 'Edmonton', 'stars': 2.0}",1
"{'city': 'Edmonton', 'stars': 3.0}",3
"{'city': 'Edmonton', 'stars': 3.5}",1
"{'city': 'Edmonton', 'stars': 4.0}",1
"{'city': 'Edmonton', 'stars': 4.5}",1
"{'city': 'Edwardsville', 'stars': 3.0}",1
"{'city': 'Florissant', 'stars': 2.5}",1
"{'city': 'Goodlettsville', 'stars': 4.0}",1
"{'city': 'Greenville', 'stars': 3.0}",1
"{'city': 'Greenwood', 'stars': 3.0}",1
"{'city': 'Hainesport', 'stars': 4.0}",1
"{'city': 'Indianapolis', 'stars': 3.5}",1
"{'city': 'Kenner', 'stars': 4.5}",1
"{'city': 'King of Prussia', 'stars': 3.5}",1
"{'city': 'Lansdowne', 'stars': 2.5}",1
"{'city': 'Marlton', 'stars': 3.5}",1
```

High rated reviews:



_id	review_id	business_i	stars	review_text
671e5d59	74bpyeG1	4qie_VdA-	5	The food
671e5d59	bwnjBkv_C	4qie_VdA-	4	Love that there is a restaurant, bar and club all in one place. My favorite is breakfast. I feel it's reasonably priced and the muffins are AMAZING! Like to go for dinner too, but it can be a little bit loud. Chic
671e5d59	uy0sDVzJR	4qie_VdA-	4	This place
671e5d59	SGjHNuux	4qie_VdA-	4	Located
671e5d59	YGvMs7O5	4qie_VdA-	4	Food and service were great! Portions were small (especially for the price) so either opt for a couple of courses or go on a night you aren't famished!
671e5d59	RCXFEqZU	4qie_VdA-	5	Best breakfast by FAR in Pinellas County. Get the eggs benedict special or the muffin of the day. You will not be disappointed!
671e5d59	bGpsnQwp	4qie_VdA-	5	I've been
671e5d59	K14PFQpQ	4qie_VdA-	4	This is a great little storefront trio of combination lounge/restaurant/nightclub on Main Street in Dunedin. The food and drinks are great, and in a fun, adult atmosphere. My only complaint is that our wait
671e5d59	x81ObYjj9	4qie_VdA-	4	Great for brunch or dinner. Menu changes a lot and is always good. make reservations for a large group and plan to drink at the Chic a Boom room next door if there is a wait. Kid friendly during the day but
671e5d59	KquiOqgO	4qie_VdA-	5	Without a doubt one of the best restaurants in the area. The quality and creations are wonderful and the prices appropriate. Breakfast is a great deal and dinner is a good value. They have small and large
671e5d59	_FIVAsJITs	3tTVN74K	5	Nice and cozy place. I went in for a quick to-go breakfast and wanted something that wasn't on the menu (breakfast burrito) and they happily obliged. Definitely recommend.
671e5d59	c_kmahKw	3tTVN74K	5	Excellent breakfast spot!!!!! Homefries are excellent and service is spot on!! I eat here 2-3 times a week!! I highly recommend it!!
671e5d59	v_v2U5DK	3tTVN74K	5	I take my kids here on Sunday mornings great food and even better prices.
671e5d59	79ZnB0jSG	3tTVN74K	4	Stopped in on7/12/17 @ 8am. Quaint little place to eat. Food was good. Service was good. Attentive to coffee cup needing a fill up. Left full. Had eggs, hash browns (typical restaurant type) not exceptiona
671e5d59	IPtRMHhX	3tTVN74K	4	Excellent
671e5d59	4gXU_2mI	3tTVN74K	5	Consistently the best breakfast sandwich I can get outside my kitchen. Wonderfully prompt service to go.
671e5d59	aqJ15QU2	3tTVN74K	4	Good basic comfort food at a great price. I've never been disappointed for breakfast or lunch and the portions are great for the price. Bring cash as they don't do credit.
671e5d59	iHEDCv0vk	3tTVN74K	5	$5.95 for 2 eggs 3 bacon home fries or hash browns coffee and juice. I dare any other place beat that. Friendly staff, food is good.
671e5d59	Of2i8s8HH	3tTVN74K	5	We go here about every Sunday and sometimes for lunch and have never had a bad meal. They grind their own coffee beans everyday. Great coffee.
671e5d59	gXxfIZItuV	3tTVN74K	5	This place was packed with senior citizens. A good sign the food is good and cheap. It was !!!!!! I had the breakfast special. One pancake, an egg and bacon. It was delicous. It was only 2 dollars. This place i
671e5d59	YcHWByxz	3tTVN74K	5	Such a nice first visit to Susie's. Once we got past the woman at the front who refused to smile or speak, we had a great experience. As part of our Biscuits and Gravy New Port Richey breakfast crawl, we w
671e5d59	g3aBmxjar	3tTVN74K	5	While
671e5d5cc	tJLE_-BqCI	rUqWq9jO	5	You cannot beat this place for the price and quality of food. The preparation of fish, tableside, is almost an art. The servers consistently debone the fish perfect each time we order it. And the octopus app
671e5d5cc	1iXHMJfa8	rUqWq9jO	5	Small, quaint place on a hidden back street in Port Richmond. It would be easy to pass by this place, but it's atmosphere is charming and relaxing. But on a better note, the food is amazing. The best Italian
671e5d5cc	J5z4Bf3Nz	rUqWq9jO	4	Been
671e5d5cc	m_vnmiVP	rUqWq9jO	4	Tried to take my parents here last weekend and I can't believe it but seems like they CLOSED! The sign said they are closed for renovations and will open in the spring, but a local from the neighborhood said
671e5d5cc	63QlrQ4iU	rUqWq9jO	4	I love this
671e5d5cc	me2k734e	rUqWq9jO	4	Off the beaten path you'll find this place practically sitting under a highway overpass. Imagine if you gutted a run-of-the-mill pizzeria, maximized the space and tried to jazz it up by going just slightly more up

Low Rated reviews:



_id	review_id	business_i	stars	review_text
671e5d5fd	SZR7bJV-A	Wy6zB1iq	1	It is very apparent this company only cares about the money and not the customer. They just do not understand that quality products & service breeds a quality bottom line.
671e5d5fd	#NAME?	Wy6zB1iq	1	P & P
671e5d5fd	4S91pA7y	Wy6zB1iq	1	Everytime we go the mccafe machine is down, you can never get a shake at this location. The food is always sub-par. Why try and patronize a location where the franchise owners refuse to fix what is wrong. Mc
671e5d5fd	fNx6qDbc	Wy6zB1iq	1	Went to the McDonalds on Hampton & 66 What a bunch of (slow) lazy(rude) workers. Last visit I had to wait 30 minutes behind 1 car, only to have the lady walk up to window with my food, then start talking fo
671e5d5fd	pjTlXMGD	Wy6zB1iq	1	Whoever is in charge of the night staff at the Bayless McDonald's in Bella Villa should be fired! I just had a 20 minute wait in the drive thru and then got the worst customer service, only half of my order, and whe
671e5d5fd	qeX6eDMk	Wy6zB1iq	1	All of their locations are just horrible. I really do not understand how they are still open. I was at a location earlier and the drive thru person argued with me about how the screen said I ordered diet. Um no hone
671e5d5fd	XD1Dnz7n	Wy6zB1iq	1	Never go
671e5d5fd	9dCWNNY	Wy6zB1iq	1	All locations I've visited that are owned by this group absolutely suck. Speed is not a thing. Orders are wrong. Drinks are not filled correctly and then workers want to argue when asked to fill cups up with more dr
671e5d5fd	rTiSI2XW7	Wy6zB1iq	1	I went to this McDonald's about 9:30 on a Saturday night. I asked for a shake to which they replied that they're ice cream was down. So I asked what about a smoothie? They asked me to wait while she checke
671e5d5fd	mcYarUMi	Wy6zB1iq	1	I along