CMPE -257: Machine learning

Project proposal

Project title- *“Business reviews and Data analysis using Machine learning on yelp.”*

Submitted to - Prof. Carlos Rojas (San Jose State University)

Submitted by- Team 12

Team members-

Name	Github Username	Github link	SJSUID
1.Sanika Vijaykumar Karwa	sanika-karwa	https://github.com/sanika-karwa	016688815
2.Harshika Shrivastava	harshika14	https://github.com/harshika14	016019120
3.Tirupati Venkata Sri Sai Rama Raju Penmatsa	rajuptvs	https://github.com/rajuptvs	016037047
4.Swapna Kotha	kothaswapna	https://github.com/kothaswapna	016007979

About our Dataset:

For this project, we are using the Yelp dataset which can be accessed from here -

Link to dataset: 1. https://www.kdnuggets.com/datasets/index.html

2. https://www.yelp.com/dataset

We are using this Yelp dataset and performing the preprocessing on it to get clean data for our project.

The motivation behind our Project:

Reviews are one of the most important assets for any business. They can be useful for any business for attracting new customers, increasing sales, and for understanding the scope of improvement for any business. From the customer's point of view, it helps customers to get a better user experience. We have decided to work on this project because utilize our machine learning models to help business owners as well as customers for a better experience.

For this project, we are using the yelp dataset. Yelp is an American company that provides an application and website where users can access the reviews given by other customers about any business. We are using an open dataset provided by Yelp.

Problem Statements and our solution:

One of the problems that any user on Yelp faces is that there are reviews but these reviews are not filtered. They are not providing a clear and overall conclusion or any graphical representation of the information to the user. We will try to provide some representation to the user using our project.

Another problem that we want to deal with in this project is - there can be fake reviews on yelp, we will try to identify a pattern in them and will try to detect fake reviews from our model.

Thirdly, it is often common to get negative reviews as well as positive reviews. With our project implementation, we will try to draw conclusions from the reviews of the user and will try to analyze the sentiments of the user about any business using machine learning.

On top of that, we will try to provide the user with a visual representation (more like a graph or word cloud) of the most frequent keywords from the data about a particular business, so they do not have to go through all the reviews instead they can look at the keywords and get an idea about the reviews.

Potential Methods:

For this project, we will be implementing supervised learning methods. We are trying to implement some functionality in our project like-

- Classification of reviews using classification methods such as- TF-IDF VECTORIZER,Gradient Boosting classifier, Naive Bayes, Decision Tree, Neural Networks (probably a transformers if time permits)
- Reviews and the text data will be preprocessed using the following techniques- NLTK library which includes tools for lemmitization, stemming and removing the unnecessary words
- Data visualization for users using- Interactive Plots using the Plotly library
- Data visualization for business owners by comparing various other businesses in the same city.

Some questions and conclusions

With our project, we will try to answer some questions like-

1. What are the top reviews about this shop?
2. How many other related businesses are there in the same city?
3. What is the feedback for any business, are the reviews positive or negative?

Preprocessing of data

In the Preprocessing, We have downloaded the datasets as a .json files and converted them to csv files respectively.

Null values have been removed from the detected columns, the detected number of null values are fairly insignificant in comparision to the size of the dataset, so they have been removed.

Using the sqldf library, we have gotten the top categories of business and their reviews have been merged.

Using the Business Id as a unique identifier key, these csv's have been merged to combine the business dataset with the reviews associated to them.

Data visualization have been plotted for checking the rate of reviews, no of reviews that are positive vs negative reviews- Interactive Plots using the Plotly library

Feature Engineering to gain some extra info on the polarity of the text using the reviews

Some of the Challenges which we might face:

Dataset looks fairly clean, but the dataset has some special attributes, which can be further extracted, but currently due to massive size of the dataset, we are still observing the data further.

Feature Engineering also would be my one of our challenge, as we would need a better understanding of the domain knowledge.

Distribution of the rating's seems to indicate that majority of the ratings are positive or neutral at the most, this can make the model biased, this is something that needs to be further looked into while creating a train/test split to have an even split.

Handling of the .ipynb notebooks would be difficult due to the vast size of the dataset -- Planning to split the .ipynb files for individual tasks

Challenges faced: As we are working with Yelp dataset, we have a lot of data to deal with. In order to deal with large data, we tried various environmental setups to increase the performance and to decrease the loading time. Another challenge was to understand the pattern in the dataset and to clean the data, as it was taking lot of time to load the data and draw plots from it. Also, Dataset has various missing values, we tried to clean the data to make it suitable for our models. Besides these challenges, we are happy that we come up with some solutions and we are excited to work with this dataset.

Data preprocessing on Yelp data set

In [4]:

```
import matplotlib.pyplot as plt # plotting
import numpy as np # linear algebra
import os # accessing directory structure
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import plotly.express as px
from pandasql import sqldf
```

Reading the data from business dataset

In [7]:

```
#reading json file-yelp_academic_dataset_business.json
yelp_academic_dataset_business_json_path = '/Users/swapnakotha/Desktop/CMPE-257/yelp
yelp_business_dataset_json = pd.read_json(yelp_academic_dataset_business_json_path,
#printing the overview of the business dataset
print(yelp_business_dataset_json.shape)
print('No of records in business dataset',yelp_business_dataset_json.shape[0])
print('No of features in business dataset',yelp_business_dataset_json.shape[1])
yelp_business_dataset_json.head()
```

```
(150346, 14)
No of records in business dataset 150346
No of features in business dataset 14
```

Out[7]:

	business_id	name	address	city	state	postal_code	latitude
0	Pns2l4eNsfO8kk83dixA6A	Abby Rappoport, LAC, CMQ	1616 Chapala St, Ste 2	Santa Barbara	CA	93101	34.426679
1	mpf3x-BjTdTEA3yCZrAYPw	The UPS Store	87 Grasso Plaza Shopping Center	Affton	MO	63123	38.551126
2	tUFrWirKiKi_TAnsVWINQQ	Target	5255 E Broadway Blvd	Tucson	AZ	85711	32.223236
3	MTSW4McQd7CbVtyjqoe9mw	St Honore Pastries	935 Race St	Philadelphia	PA	19107	39.955505
4	mWMc6_wTdE0EUBKIGXDVfA	Perkiomen Valley Brewery	101 Walnut St	Green Lane	PA	18054	40.338183

In [8]:

```
# 1 = open, 0 = closed
# There are significant amount of businesses that are not open anymore,
yelp_business_dataset_json.is_open.value_counts()
```

Out[8]:

```
1    119698
0     30648
Name: is_open, dtype: int64
```

In [11]:

```
df_categories =yelp_business_dataset_json.assign(categories = yelp_business_dataset_
df_categories.sample(3)
```

Out[11]:

	business_id	name	address	city	state	postal_code	latitude
27865	30QeVr3702wSurnf-kLHWA	Art Smart Coffee & Gallery	1275 Bayshore	Dunedin	FL	34698	28.021729
82812	ykwIZ4GWkCh2MazOdY_NOA	Blake's On Poydras	920 Poydras St	New Orleans	LA	70112	29.950040
68816	tELoGj2QJYYFQ1jo_XXbpA	Robert's Western World	416 Broadway, Ste B	Nashville	TN	37203	36.161041

Getting the categories

In [12]:

```
#showing the business categories
print('Total number of categories:', len(df_categories.categories.value_counts()))
#showing the top 20 business categories
print('Top 20 categories:')
df_categories.categories.value_counts()[:20]
```

```
Total number of categories: 1311
Top 20 categories:
```

Out[12]:

```
Restaurants                      52268
Food                             27781
Shopping                         24395
Home Services                    14356
Beauty & Spas                    14292
Nightlife                        12281
Health & Medical                 11890
Local Services                   11198
Bars                             11065
Automotive                       10773
Event Planning & Services         9895
Sandwiches                        8366
American (Traditional)            8139
Active Life                       7687
Pizza                             7093
Coffee & Tea                      6703
Fast Food                         6472
Breakfast & Brunch                6239
American (New)                    6097
Hotels & Travel                   5857
Name: categories, dtype: int64
```

In [13]:

```
df_categories
```

Out[13]:

	business_id	name	address	city	state	postal_code	latit
0	Pns2l4eNsfO8kk83dixA6A	Abby Rappoport, LAC, CMQ	1616 Chapala St, Ste 2	Santa Barbara	CA	93101	34.426
0	Pns2l4eNsfO8kk83dixA6A	Abby Rappoport, LAC, CMQ	1616 Chapala St, Ste 2	Santa Barbara	CA	93101	34.426
0	Pns2l4eNsfO8kk83dixA6A	Abby Rappoport, LAC, CMQ	1616 Chapala St, Ste 2	Santa Barbara	CA	93101	34.426
0	Pns2l4eNsfO8kk83dixA6A	Abby Rappoport, LAC, CMQ	1616 Chapala St, Ste 2	Santa Barbara	CA	93101	34.426
0	Pns2l4eNsfO8kk83dixA6A	Abby Rappoport, LAC, CMQ	1616 Chapala St, Ste 2	Santa Barbara	CA	93101	34.426
...	...	...	...	...	...	...	
150344	mtGm22y5c2UHNXDFAjaPNw	Cyclery & Fitness Center	2472 Troy Rd	Edwardsville	IL	62025	38.782
150345	jV_XOycEzSlTx-65W906pg	Sic Ink	238 Apollo Beach Blvd	Apollo beach	FL	33572	27.771
150345	jV_XOycEzSlTx-65W906pg	Sic Ink	238 Apollo Beach Blvd	Apollo beach	FL	33572	27.771
150345	jV_XOycEzSlTx-65W906pg	Sic Ink	238 Apollo Beach Blvd	Apollo beach	FL	33572	27.771
150345	jV_XOycEzSlTx-65W906pg	Sic Ink	238 Apollo Beach Blvd	Apollo beach	FL	33572	27.771

668695 rows × 14 columns

In [14]:

```
df_state = yelp_business_dataset_json.assign(categories = yelp_business_dataset_json
df_state.sample(3)
```

Out[14]:

	business_id	name	address	city	state	postal_code	latitude
44684	nE34vey307UqfUlfP8E9Tg	Twill Auto Service	9545 47th Ave N	Saint Petersburg	FL	33708	27.815214
133665	Qhl1JSjGUOnegpB1_QFLRA	Salon Ya Ya	2210 Crestmoor Rd, Ste 15, The Gallery of Gree...	Nashville	TN	37215	36.109259
144647	QwGeQLFByptdXLxqnNrcYg	My SEO Guys	1301 Seminole Blvd, Ste 136	Largo	FL	33770	27.919617

In [15]:

```
#printing the number of states in which the businesses are available
print('Total number of categories:', len(df_state.state.value_counts()))
```

```
Total number of categories: 27
```

Cleaning the dataset

In [16]:

```
#countof missing values for each attribute
df_categories.isna().sum()
```

Out[16]:

```
business_id         0
name                0
address             0
city                0
state               0
postal_code         0
latitude            0
longitude           0
stars               0
review_count        0
is_open             0
attributes      46607
categories        103
hours           73794
dtype: int64
```

In [17]:

```
#Eliminating the missing values
df_categories_clean=df_categories.dropna()
```

In [18]:

```
df_categories_clean.isna().sum()
```

Out[18]:

```
business_id     0
name            0
address         0
city            0
state           0
postal_code     0
latitude        0
longitude       0
stars           0
review_count    0
is_open         0
attributes      0
categories      0
hours           0
dtype: int64
```

Selecting a few categories from the larger dataset and merging them with reviews json

In [19]:

```
#Taking the business which are classified as 'Active Life' and creating a smaller su
active_life = df_categories_clean[df_categories_clean['categories'].str.contains("Ac
active_life
```

Out[19]:

	business_id	name	address	city	state	postal_code
30	fvWn8oXXwbj2l79cochZyw	Altitude Trampoline Park - Boise	1301 N Milwaukee St	Boise	ID	83704
38	LcAozWCMLGjwRbokaJAKMg	Edwardsville Children's Museum	722 Holyoake Rd	Edwardsville	IL	62025
50	Hwt3_mOEmU-t--ywcemnMg	Gold's Gym	203 - 38th Ave N	St. Petersburg	FL	33704
100	8KMIT0NXu30Jz5Ojo5uxaw	Cornerstone Physical Therapy Associates	1338 Bristol Pike, Ste 203	Bensalem	PA	19020
109	I6DCYks9lqZeoZiVzW7PmA	Its Sold Here	94 York Rd	Willow Grove	PA	19090
...	...	...	...	...	...	...
150278	jYd7okFv6JMjIXMDjZNCDQ	Ace Golf	820 S Kings Ave	Brandon	FL	33511
150285	fWeWzB9STxcX40AgSEQVcw	Arizona-Sonora Desert Museum	2021 N Kinney Rd	Tucson	AZ	85743
150303	JhSByBTYY1rGstRy76YmLA	Reiki with Darren		Santa Barbara	CA	93105
150334	LJ4GjQ1HL6kqvIPpNUNNaQ	Shanti Yoga and Ayurveda	1638 Pine St, Fl 1	Philadelphia	PA	19103

	business_id	name	address	city	state	postal_code
150338	fn3ybdsRSrIDpKZTsRuAWg	INSPcenter/Thai Clinical Massage	2625 N Meridian St, Unit 50	Indianapolis	IN	46208

5842 rows × 14 columns

Reading the data from Reviews dataset

In [22]:

```
#reading the data from yelp_academic_dataset_review
yelp_academic_dataset_review_json_path = '/Users/swapnakotha/Desktop/CMPE-257/yelp_c
```

In [23]:

```
#Date set which contain is very large(it contaings around 6 million reviews)
#Instead of reading it at once,read it in smaller part for
size = 1000000
review = pd.read_json(yelp_academic_dataset_review_json_path, lines=True,
                      dtype={'review_id':str,'user_id':str,
                             'business_id':str,'stars':int,
                             'date':str,'text':str,'useful':int,
                             'funny':int,'cool':int},
                      chunksize=size)

chunk_list = []

for chunk_review in review:
    chunk_review = chunk_review.drop(['review_id','useful','funny','cool'], axis=1)
    chunk_review = chunk_review.rename(columns={'stars': 'review_stars'})
    #Merge reviews which corresponds to the business that are categorized as 'Active
    chunk_merged = pd.merge(active_life, chunk_review, on='business_id', how='inner'
    print(f"{chunk_merged.shape[0]} out of {size:,} related reviews")
    chunk_list.append(chunk_merged)
active_life_review = pd.concat(chunk_list, ignore_index=True, join='outer', axis=0)
```

```
26386 out of 1,000,000 related reviews
27572 out of 1,000,000 related reviews
25482 out of 1,000,000 related reviews
22872 out of 1,000,000 related reviews
23464 out of 1,000,000 related reviews
26750 out of 1,000,000 related reviews
23804 out of 1,000,000 related reviews
```

In [24]:

```
#Converting into CSV file
active_life_review.to_csv("active_life_reviews.csv",index=False)
```

In [25]:

```
#Taking the business which are classified as 'Restaurants' and creating a smaller su
restaurants = df_categories_clean[df_categories_clean['categories'].str.contains("Re
```

In [26]:

```
restaurants
```

Out[26]:

	business_id	name	address	city	state	postal_code	l
3	MTSW4McQd7CbVtyjqoe9mw	St Honore Pastries	935 Race St	Philadelphia	PA	19107	39.
5	CF33F8-E6oudUQ46HnavjQ	Sonic Drive-In	615 S Main St	Ashland City	TN	37015	36.
9	bBDDEgkFA1Otx9Lfe7BZUQ	Sonic Drive-In	2312 Dickerson Pike	Nashville	TN	37207	36.
11	eEOYSgkmpB90uNA7lDOMRA	Vietnamese Food Truck		Tampa Bay	FL	33602	27.
12	il_Ro8jwPlHresjw9EGmBg	Denny's	8901 US 31 S	Indianapolis	IN	46227	39.
...	...	...	...	...	...	...	
150325	l9eLGG9ZKpLJzboZq-9LRQ	Wawa	19 N Bishop Ave	Clifton Heights	PA	19018	39.
150327	cM6V90ExQD6KMSU3rRB5ZA	Dutch Bros Coffee	1181 N Milwaukee St	Boise	ID	83704	43.
150336	WnT9NIzQgLlILjPT0kEcsQ	Adelita Taqueria & Restaurant	1108 S 9th St	Philadelphia	PA	19147	39.
150339	2O2K6SXPWv56amqxCECd4w	The Plum Pit	4405 Pennell Rd	Aston	DE	19014	39.
150340	hn9Toz3s-Ei3uZPt7esExA	West Side Kebab House	2470 Guardian Road NW	Edmonton	AB	T5T 1K8	53.

44736 rows × 14 columns

In [27]:

```
#Date set which contain is very large(it contaings around 6 million reviews)
#Instead of reading it at once,read it in smaller part for
size = 1000000
review = pd.read_json(yelp_academic_dataset_review_json_path, lines=True,
                      dtype={'review_id':str,'user_id':str,
                             'business_id':str,'stars':int,
                             'date':str,'text':str,'useful':int,
                             'funny':int,'cool':int},
                      chunksize=size)
```

In [28]:

```
restaurant_chunks = []
for chunk_review in review:
    chunk_review = chunk_review.drop(['review_id','useful','funny','cool'], axis=1)
    chunk_review = chunk_review.rename(columns={'stars': 'review_stars'})
    # Merge reviews which corresponds to the business that are categorized as 'Resta
    chunk_merged = pd.merge(restaurants, chunk_review, on='business_id', how='inner'
    print(f"{chunk_merged.shape[0]} out of {size:,} related reviews")
    restaurant_chunks.append(chunk_merged)
restaurant = pd.concat(restaurant_chunks, ignore_index=True, join='outer', axis=0)
```

```
666750 out of 1,000,000 related reviews
657273 out of 1,000,000 related reviews
646104 out of 1,000,000 related reviews
644296 out of 1,000,000 related reviews
646484 out of 1,000,000 related reviews
664130 out of 1,000,000 related reviews
636531 out of 1,000,000 related reviews
```

In [29]:

```
restaurant
```

Out[29]:

	business_id	name	address	city	state	postal_code	latitu
0	MTSW4McQd7CbVtyjqoe9mw	St Honore Pastries	935 Race St	Philadelphia	PA	19107	39.9555
1	MTSW4McQd7CbVtyjqoe9mw	St Honore Pastries	935 Race St	Philadelphia	PA	19107	39.9555
2	MTSW4McQd7CbVtyjqoe9mw	St Honore Pastries	935 Race St	Philadelphia	PA	19107	39.9555
3	MTSW4McQd7CbVtyjqoe9mw	St Honore Pastries	935 Race St	Philadelphia	PA	19107	39.9555
4	MTSW4McQd7CbVtyjqoe9mw	St Honore Pastries	935 Race St	Philadelphia	PA	19107	39.9555
...	...	...	...	...	...	...	
4561563	hn9Toz3s-Ei3uZPt7esExA	West Side Kebab House	2470 Guardian Road NW	Edmonton	AB	T5T 1K8	53.5096
4561564	hn9Toz3s-Ei3uZPt7esExA	West Side Kebab House	2470 Guardian Road NW	Edmonton	AB	T5T 1K8	53.5096
4561565	hn9Toz3s-Ei3uZPt7esExA	West Side Kebab House	2470 Guardian Road NW	Edmonton	AB	T5T 1K8	53.5096
4561566	hn9Toz3s-Ei3uZPt7esExA	West Side Kebab House	2470 Guardian Road NW	Edmonton	AB	T5T 1K8	53.5096
4561567	hn9Toz3s-Ei3uZPt7esExA	West Side Kebab House	2470 Guardian Road NW	Edmonton	AB	T5T 1K8	53.5096

4561568 rows × 18 columns

In [30]:

```
#converting it into
restaurant.to_csv("restaurants_reviews.csv",index=False)
```

In [31]:

```
#Taking the business which are classified as 'medical' and creating a smaller subset
medical = df_categories_clean[df_categories_clean['categories'].str.contains("Health
medical
```

Out[31]:

	business_id	name	address	city	state	postal_code
13	jaxMSoInw8Poo3XeMJt8lQ	Adams Dental	15 N Missouri Ave	Clearwater	FL	33755
43	Kq51_lGAgAigqmENITTr-A	Bala Better Health	2 Bala Plz, Ste PL-11	Bala Cynwyd	PA	19004
57	DQ7PyYlp2bX96WZa7TcaWQ	LensCrafters	1150 Plymouth Meeting Mall, Ste 2230	Plymouth Meeting	PA	19462
74	9Rww8yE6Dm4dSMEq09nwXg	Holly Nails & Spa	9101 Belcher Rd	Pinellas Park	FL	33782
100	8KMIT0NXu30Jz5Ojo5uxaw	Cornerstone Physical Therapy Associates	1338 Bristol Pike, Ste 203	Bensalem	PA	19020
...	...	...	...	...	...	...
150318	VKbQHHUu_cB7M6jQwA3n-w	Planned Parenthood - FifthStreet Health Center	455 W 5th St	Reno	NV	89503
150333	LTeBejee7jIpaYWWIl-Ubw	Town & Country Dental Care	2821 N Ballas Rd, Ste 163	Saint Louis	MO	63131
150334	LJ4GjQ1HL6kqvIPpNUNNaQ	Shanti Yoga and Ayurveda	1638 Pine St, Fl 1	Philadelphia	PA	19103
150335	Gi1QPLu_y8rLS3uTN9Z_VA	St. Vincent Heart Center of Indiana	10580 N Meridian St	Indianapolis	IN	46290
150338	fn3ybdsRSrIDpKZTsRuAWg	INSPcenter/Thai Clinical Massage	2625 N Meridian St, Unit 50	Indianapolis	IN	46208

9821 rows × 14 columns

In [32]:

```
#Date set which contain is very large(it contaings around 6 million reviews)
#Instead of reading it at once,read it in smaller part for
size = 1000000
review = pd.read_json(yelp_academic_dataset_review_json_path, lines=True,
                      dtype={'review_id':str,'user_id':str,
                             'business_id':str,'stars':int,
                             'date':str,'text':str,'useful':int,
                             'funny':int,'cool':int},
                      chunksize=size)
```

In [33]:

```
medical_chunks = []
for chunk_review in review:
    chunk_review = chunk_review.drop(['review_id','useful','funny','cool'], axis=1)
    chunk_review = chunk_review.rename(columns={'stars': 'review_stars'})
    #Merge reviews which corresponds to the business that are categorized as 'medica
    chunk_merged = pd.merge(medical, chunk_review, on='business_id', how='inner')
    print(f"{chunk_merged.shape[0]} out of {size:,} related reviews")
    medical_chunks.append(chunk_merged)
medical_reviews = pd.concat(medical_chunks, ignore_index=True, join='outer', axis=0)
```

```
21579 out of 1,000,000 related reviews
23982 out of 1,000,000 related reviews
24359 out of 1,000,000 related reviews
25272 out of 1,000,000 related reviews
26704 out of 1,000,000 related reviews
23178 out of 1,000,000 related reviews
28964 out of 1,000,000 related reviews
```

In [34]:

```
#converting into CSV file
medical_reviews.to_csv("medical_reviews.csv",index=False)
```

In [35]:

```
#Taking the business which are classified as 'Home Services' and creating a smaller
home_services = df_categories_clean[df_categories_clean['categories'].str.contains('
home_services
```

Out[35]:

	business_id	name	address	city	state	postal_code	
84	eMjnw_7wp-CscyNh6Lu0ZA	AM&PM Locksmith	8540 Bustleton Ave	Philadelphia	PA	19152	4
94	ZM46RDLXaFNo_z6t-j_L4w	Absolutely Perfect Inc	1153 Byberry Rd	Bensalem	PA	19020	4
107	2n9HHBxG7yAyAUwXXa49aw	Mighty Dustless	1256 Valley Hill Trl	Southampton	PA	18966	4
120	bYjnX_J1bHZob10DoSFkqQ	Tinkle Belle Diaper Service		Santa Barbara	CA	93101	3
124	Q3kQYhkYxSRyYyeBgtk--A	Cook's Glass & Mirror	5703 W Morris St	Indianapolis	IN	46241	3
...	...	...	...	...	...	...	
150266	vb7t5_4aZ9yDgOMmFGYKgw	Cobb Property Management	5650 E 22nd St	Tucson	AZ	85711	3
150286	Vl0oo3jjuGpgMWaCbN5r9w	Steve Bright Handyman		Blue Bell	PA	19422	4
150289	Fck8i0fNQCa22ERz5Fa21w	Thoughtful Moving	5004 E Fowler Ave	Tampa	FL	33617	2
150307	nMx7lAeMqy1-GfB84RnyhQ	Devonshire	1100 Devonshire E Dr	Greenwood	IN	46143	3

	business_id	name	address	city	state	postal_code
150314	_h9b34onQc_26F9mvmsNhw	J&M Gutter Pros		Voorhees	NJ	08043

11760 rows × 14 columns

In [36]:

```
#Date set which contain is very large(it contaings around 6 million reviews)
#Instead of reading it at once,read it in smaller part for
size = 1000000
review = pd.read_json(yelp_academic_dataset_review_json_path, lines=True,
                      dtype={'review_id':str,'user_id':str,
                             'business_id':str,'stars':int,
                             'date':str,'text':str,'useful':int,
                             'funny':int,'cool':int},
                      chunksize=size)
```

In [37]:

```
homeservices_chunks = []
for chunk_review in review:
    chunk_review = chunk_review.drop(['review_id','useful','funny','cool'], axis=1)
    chunk_review = chunk_review.rename(columns={'stars': 'review_stars'})
    #Merge reviews which corresponds to the business that are categorized as 'Home_s
    chunk_merged = pd.merge(home_services, chunk_review, on='business_id', how='inne
    print(f"{chunk_merged.shape[0]} out of {size:,} related reviews")
    homeservices_chunks.append(chunk_merged)
homeservices_reviews = pd.concat(homeservices_chunks, ignore_index=True, join='outer
```

```
27240 out of 1,000,000 related reviews
31639 out of 1,000,000 related reviews
30570 out of 1,000,000 related reviews
29977 out of 1,000,000 related reviews
31513 out of 1,000,000 related reviews
24739 out of 1,000,000 related reviews
35367 out of 1,000,000 related reviews
```

In [38]:

```
#converting into CSV file
homeservices_reviews.to_csv("homeservices_reviews.csv")
```

In [39]:

```
homeservices_reviews
```

Out[39]:

	business_id	name	address	city	state	postal_code	latitu
0	eMjnw_7wp-CscyNh6Lu0ZA	AM&PM Locksmith	8540 Bustleton Ave	Philadelphia	PA	19152	40.072
1	eMjnw_7wp-CscyNh6Lu0ZA	AM&PM Locksmith	8540 Bustleton Ave	Philadelphia	PA	19152	40.072
2	eMjnw_7wp-CscyNh6Lu0ZA	AM&PM Locksmith	8540 Bustleton Ave	Philadelphia	PA	19152	40.072
3	eMjnw_7wp-CscyNh6Lu0ZA	AM&PM Locksmith	8540 Bustleton Ave	Philadelphia	PA	19152	40.072
4	eMjnw_7wp-CscyNh6Lu0ZA	AM&PM Locksmith	8540 Bustleton Ave	Philadelphia	PA	19152	40.072
...	...	...	...	...	...	...	
211040	_h9b34onQc_26F9mvmsNhw	J&M Gutter Pros		Voorhees	NJ	08043	39.851
211041	_h9b34onQc_26F9mvmsNhw	J&M Gutter Pros		Voorhees	NJ	08043	39.851
211042	_h9b34onQc_26F9mvmsNhw	J&M Gutter Pros		Voorhees	NJ	08043	39.851
211043	_h9b34onQc_26F9mvmsNhw	J&M Gutter Pros		Voorhees	NJ	08043	39.851
211044	_h9b34onQc_26F9mvmsNhw	J&M Gutter Pros		Voorhees	NJ	08043	39.851

211045 rows × 18 columns

In [40]:

```
active_life
```

Out[40]:

	business_id	name	address	city	state	postal_code
30	fvWn8oXXwbj2l79cochZyw	Altitude Trampoline Park - Boise	1301 N Milwaukee St	Boise	ID	83704
38	LcAozWCMLGjwRbokaJAKMg	Edwardsville Children's Museum	722 Holyoake Rd	Edwardsville	IL	62025
50	Hwt3_mOEmU-t--ywcemnMg	Gold's Gym	203 - 38th Ave N	St. Petersburg	FL	33704
100	8KMIT0NXu30Jz5Ojo5uxaw	Cornerstone Physical Therapy Associates	1338 Bristol Pike, Ste 203	Bensalem	PA	19020
109	I6DCYks9lqZeoZiVzW7PmA	Its Sold Here	94 York Rd	Willow Grove	PA	19090
...	...	...	...	...	...	...
150278	jYd7okFv6JMjIXMDjZNCDQ	Ace Golf	820 S Kings Ave	Brandon	FL	33511
150285	fWeWzB9STxcX40AgSEQVcw	Arizona-Sonora Desert Museum	2021 N Kinney Rd	Tucson	AZ	85743
150303	JhSByBTYY1rGstRy76YmLA	Reiki with Darren		Santa Barbara	CA	93105
150334	LJ4GjQ1HL6kqvIPpNUNNaQ	Shanti Yoga and Ayurveda	1638 Pine St, Fl 1	Philadelphia	PA	19103
150338	fn3ybdsRSrIDpKZTsRuAWg	INSPcenter/Thai Clinical Massage	2625 N Meridian St, Unit 50	Indianapolis	IN	46208

5842 rows × 14 columns

In [41]:

```
active_life_review=pd.read_csv("active_life_reviews.csv")
```

In [42]:

```
active_life_review
```

Out[42]:

	business_id	name	address	city	state	postal_code	
0	fvWn8oXXwbj2l79cochZyw	Altitude Trampoline Park - Boise	1301 N Milwaukee St	Boise	ID	83704	43
1	fvWn8oXXwbj2l79cochZyw	Altitude Trampoline Park - Boise	1301 N Milwaukee St	Boise	ID	83704	43
2	fvWn8oXXwbj2l79cochZyw	Altitude Trampoline Park - Boise	1301 N Milwaukee St	Boise	ID	83704	43
3	fvWn8oXXwbj2l79cochZyw	Altitude Trampoline Park - Boise	1301 N Milwaukee St	Boise	ID	83704	43
4	fvWn8oXXwbj2l79cochZyw	Altitude Trampoline Park - Boise	1301 N Milwaukee St	Boise	ID	83704	43
...	...	...	...	...	...	...	
176325	fn3ybdsRSrIDpKZTsRuAWg	INSPcenter/Thai Clinical Massage	2625 N Meridian St, Unit 50	Indianapolis	IN	46208	39
176326	fn3ybdsRSrIDpKZTsRuAWg	INSPcenter/Thai Clinical Massage	2625 N Meridian St, Unit 50	Indianapolis	IN	46208	39
176327	fn3ybdsRSrIDpKZTsRuAWg	INSPcenter/Thai Clinical Massage	2625 N Meridian St, Unit 50	Indianapolis	IN	46208	39
176328	fn3ybdsRSrIDpKZTsRuAWg	INSPcenter/Thai Clinical Massage	2625 N Meridian St, Unit 50	Indianapolis	IN	46208	39

	business_id	name	address	city	state	postal_code	
176329	fn3ybdsRSrIDpKZTsRuAWg	INSPcenter/Thai Clinical Massage	2625 N Meridian St, Unit 50	Indianapolis	IN	46208	39

176330 rows × 18 columns

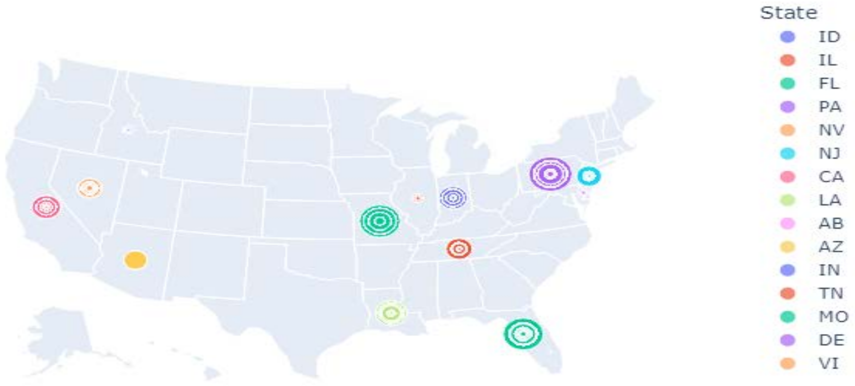
Data visualization

In [43]:

```
#visual representation of distribution of review count for the business categorized
fig1 = px.scatter_geo(active_life_review, locations='state', size='review_count', lo
```

In [44]:

```
fig1.show()
```



In [45]:

```
#reading the data of business categorized as restaurents
restaurant=pd.read_csv("restaurants_reviews.csv")
```

In [7]:

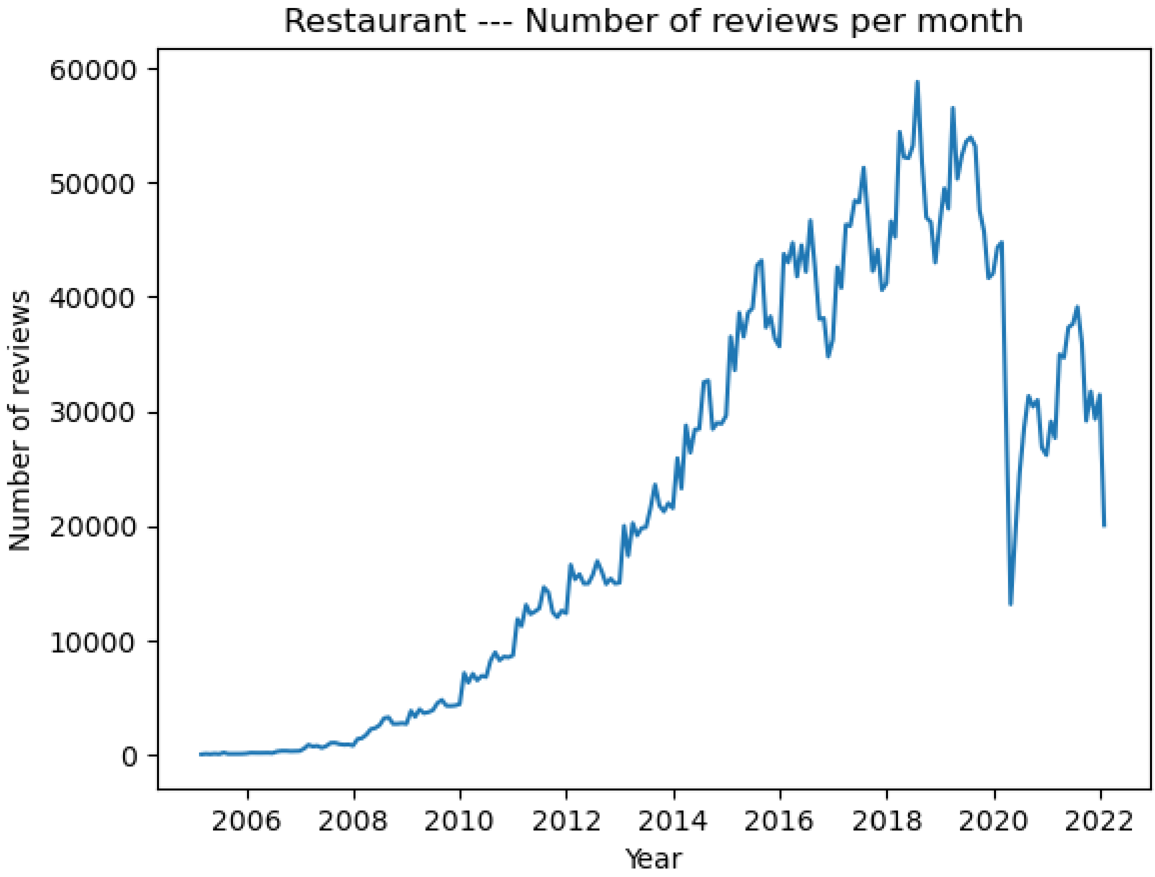
```
restaurant['date']=pd.to_datetime(restaurant['date'])
restaurant_testing=restaurant
```

In [8]:

```
restaurant_testing=restaurant_testing.set_index('date')
```

In [9]:

```
#plot to show variation of review count across the years(from 2006 to 2022)
plt.plot(restaurant_testing['text'].resample('M').count())
plt.xlabel('Year')
plt.ylabel('Number of reviews')
plt.title('Restaurant --- Number of reviews per month')
plt.show()
```



In [10]:

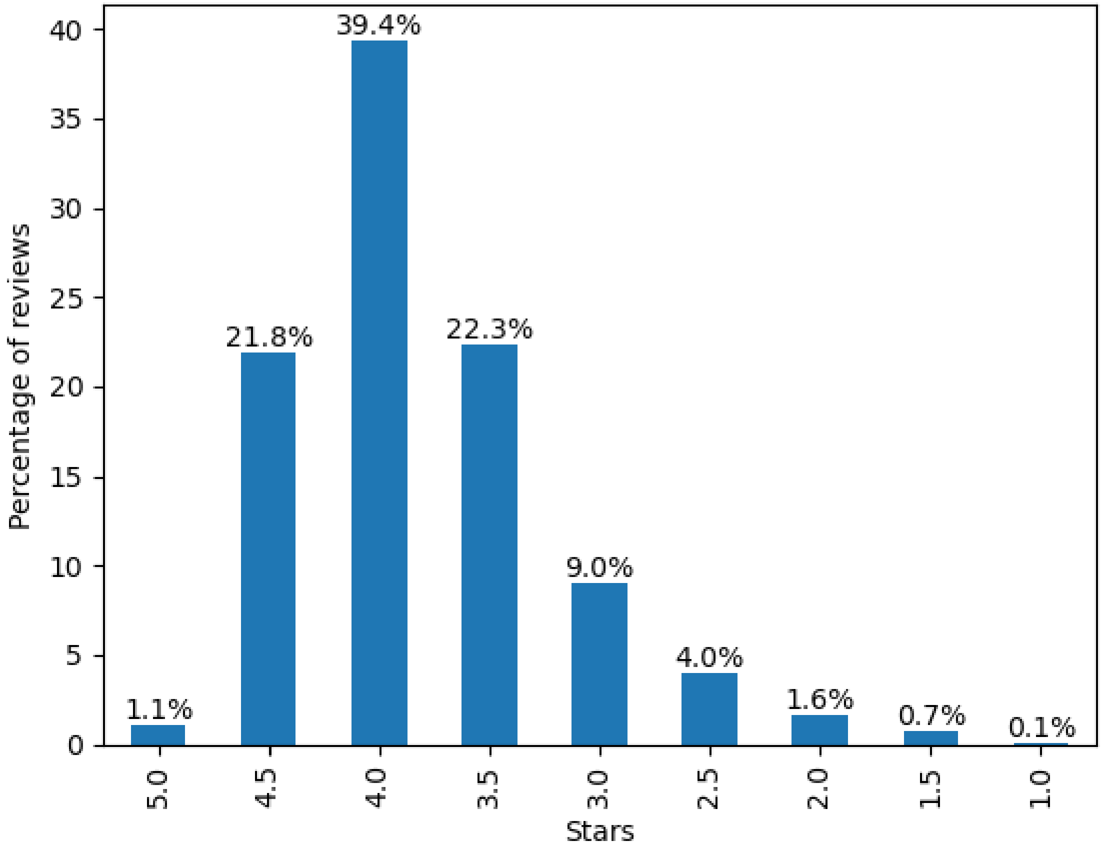
```
#calculating the review distribution based on rating
restaurant_testing_Stars = restaurant_testing['stars'].value_counts()
review_stars_percent = restaurant_testing_Stars.apply(lambda i : i / len(restaurant_
review_stars_percent
```

Out[10]:

```
4.0    39.404740
3.5    22.345299
4.5    21.844594
3.0     8.982920
2.5     3.956425
2.0     1.597762
5.0     1.074894
1.5     0.723940
1.0     0.069428
Name: stars, dtype: float64
```

In [12]:

```
#plot for review distribution
ax = review_stars_percent.sort_index(ascending=False).plot(kind='bar')
for patch in ax.patches:
    x = patch.get_bbox().get_points()[:, 0]
    y = patch.get_bbox().get_points()[1, 1]
    ax.annotate('{:.1f}%'.format(y), (x.mean(), y), ha='center', va='bottom')
plt.ylabel('Percentage of reviews')
plt.xlabel('Stars');
```



In [17]:

```
pysqldf = lambda q: sqldf(q, globals())
```

In [18]:

```
q = """SELECT state, count(*) as count
       FROM restaurant_testing
       group by state
       ;"""
```

In [19]:

```
missing = pysqldf(q)
missing
```

In [4]:

```
bus = pd.read_csv("business_updated.csv")
bus.head()

df_categories = bus.assign(categories = bus.categories.str.split(', ')).explode('cat
df_categories.head()

print('Total number of categories:', len(df_categories.categories.value_counts()))
print('Top 20 categories:')
abc = df_categories.categories.value_counts()[:20]
abc.shape


from pandasql import sqldf
pysqldf = lambda q: sqldf(q, globals())

q = """SELECT categories, count(*) as count
       FROM df_categories
       group by categories
       ;"""

missing = pysqldf(q)
missing
```

```
Total number of categories: 1311
Top 20 categories:
```

Out[4]:

	categories	count
0	None	103
1	& Probates	38
2	3D Printing	5
3	ATV Rentals/Tours	12
4	Acai Bowls	268
...	...	...
1307	Wraps	310
1308	Yelp Events	48
1309	Yoga	938
1310	Ziplining	12
1311	Zoos	52

1312 rows × 2 columns

In [5]:

```
restaurant_top = missing.sort_values('count')
Top20Cat = abc[-20:]
```

In [6]:

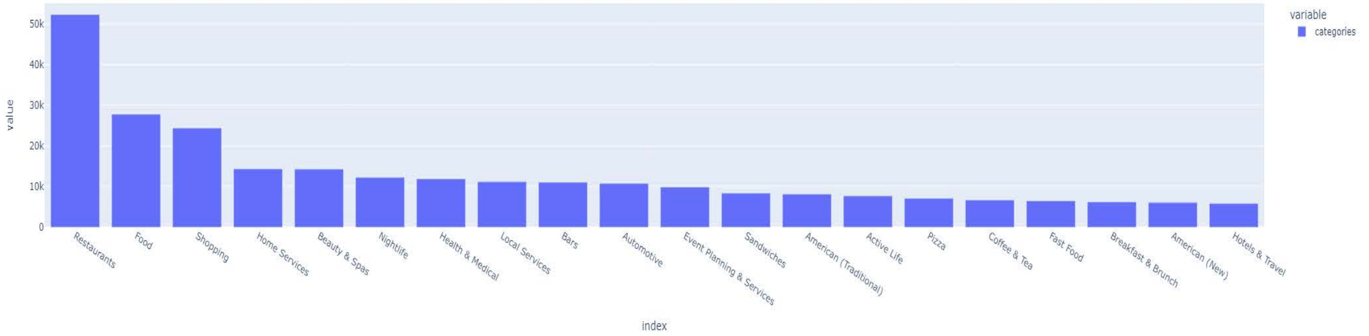
```
Top20Cat
```

Out[6]:

```
Restaurants                    52268
Food                           27781
Shopping                       24395
Home Services                  14356
Beauty & Spas                  14292
Nightlife                      12281
Health & Medical               11890
Local Services                 11198
Bars                           11065
Automotive                     10773
Event Planning & Services       9895
Sandwiches                      8366
American (Traditional)          8139
Active Life                     7687
Pizza                           7093
Coffee & Tea                    6703
Fast Food                       6472
Breakfast & Brunch              6239
American (New)                  6097
Hotels & Travel                 5857
Name: categories, dtype: int64
```

In [7]:

```
fig = px.bar(Top20Cat)
fig.show()
```



In [9]:

```
from pandasql import sqldf
pysqldf = lambda q: sqldf(q, globals())

q = """SELECT state, count(*) as count
       FROM df_categories
       group by state
       ;"""

missing = pysqldf(q)
missing
```

Out[9]:

	state	count
0	AB	22574
1	AZ	43173
2	CA	24296
3	CO	8
4	DE	9996
5	FL	119479
6	HI	15
7	ID	20136
8	IL	9326
9	IN	50181
10	LA	44315
11	MA	4
12	MI	6
13	MO	49839
14	MT	5
15	NC	8
16	NJ	36332
17	NV	34771
18	PA	149217
19	SD	7
20	TN	54963
21	TX	15
22	UT	6
23	VI	6
24	VT	7
25	WA	4
26	XMS	6

In [12]:

```
restaurant_top_states = missing.sort_values('count')
Top10State = restaurant_top_states[-10:]
```

In [15]:

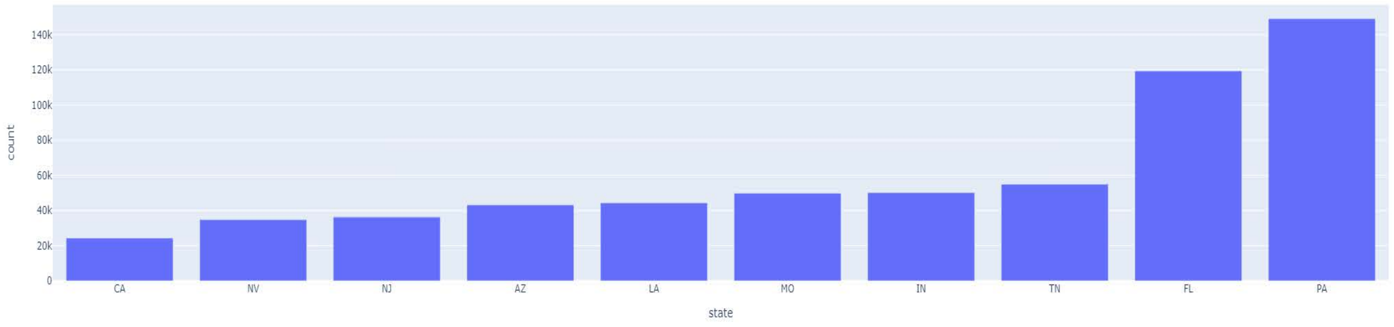
```
fig = px.bar(Top10State, x = 'state', y = 'count')
fig.show()
```

In [3]:

```
rest = pd.read_csv("restaurants_reviews.csv")
rest.head()
```

Out[3]:

	business_id	name	address	city	state	postal_code	latitude	lo
0	MTSW4McQd7CbVtyjqoe9mw	St Honore Pastries	935 Race St	Philadelphia	PA	19107	39.955505	-7
1	MTSW4McQd7CbVtyjqoe9mw	St Honore Pastries	935 Race St	Philadelphia	PA	19107	39.955505	-7
2	MTSW4McQd7CbVtyjqoe9mw	St Honore Pastries	935 Race St	Philadelphia	PA	19107	39.955505	-7
3	MTSW4McQd7CbVtyjqoe9mw	St Honore Pastries	935 Race St	Philadelphia	PA	19107	39.955505	-7
4	MTSW4McQd7CbVtyjqoe9mw	St Honore Pastries	935 Race St	Philadelphia	PA	19107	39.955505	-7



In [11]:

```
pysqldf = lambda q: sqldf(q, globals())

q = """SELECT name, count(*) as count
       FROM rest
       group by name
       ;"""

wordcloud1 = pysqldf(q)
wordcloud1
```

Out[11]:

	name	count
0	"Genuino" Italian Cafe'	69
1	#1 Mongolian BBQ - Best Stir Fried Noodles In ...	57
2	&pizza - UPenn	65
3	&pizza - Walnut	386
4	&pizza - Willow Grove	20
...	...	...
31404	¡CUATRO	30
31405	¡Juice!	9
31406	ÀLAVITA	320
31407	Àrdana Food & Drink	83
31408	ā café	50

31409 rows × 2 columns

In [12]:

```
abc = wordcloud1.sort_values('count')
abc[-20:]

df = abc[-50:]
abc = ''.join(df['name'])
```

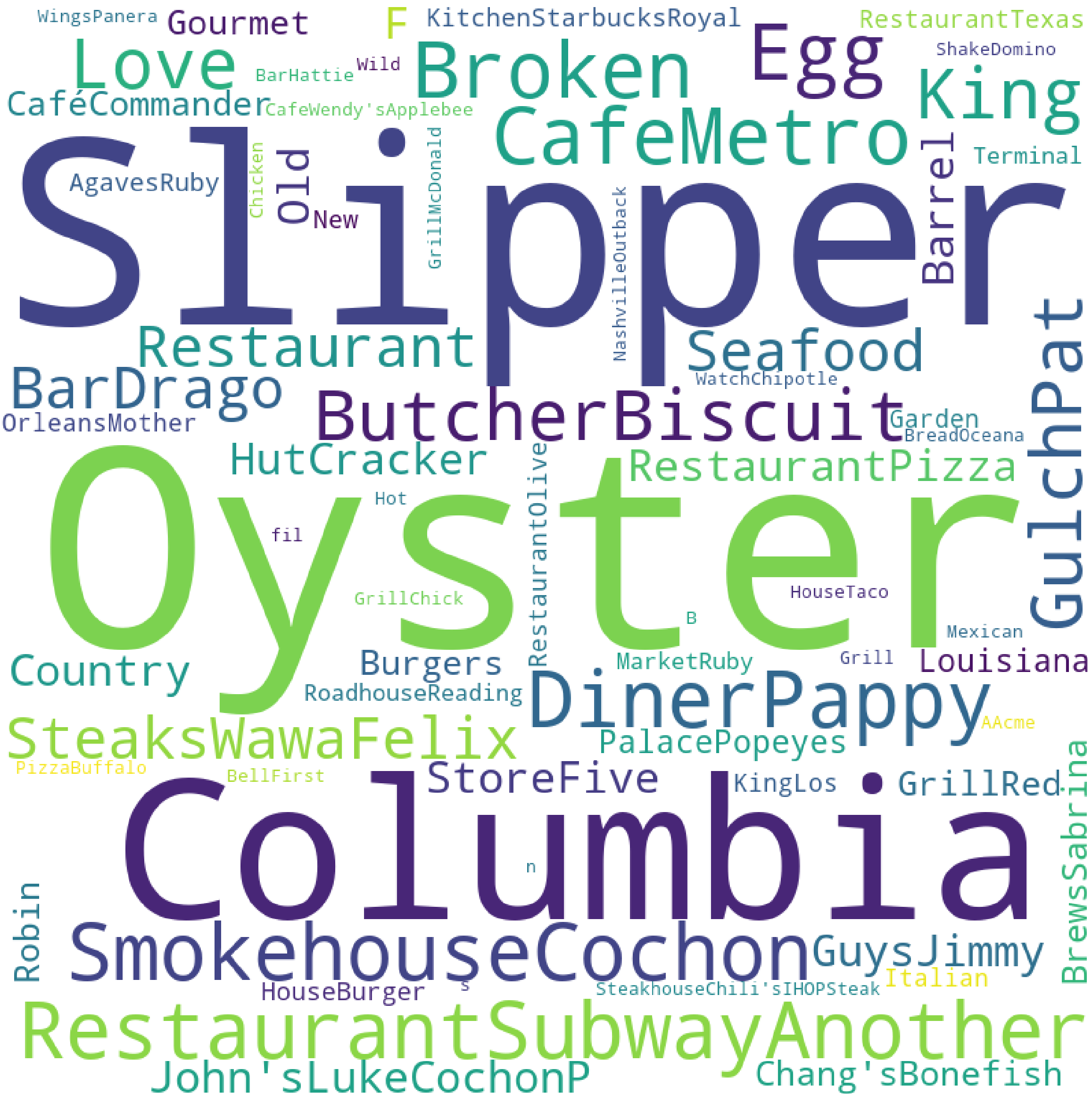
In [13]:

```
from wordcloud import WordCloud, STOPWORDS
import matplotlib.pyplot as plt

wordcloud = WordCloud(width = 800, height = 800,
                background_color ='white',
                stopwords = STOPWORDS,
                min_font_size = 10).generate(abc)

plt.figure(figsize = (8, 8), facecolor = None)
plt.imshow(wordcloud)
plt.axis("off")
plt.tight_layout(pad = 0)

plt.show()
```



In [ ]:

```
pysqldf = lambda q: sqldf(q, globals())

q = """SELECT name, count(*) as count
       FROM rest
       group by name
       ;"""

state_plot = pysqldf(q)
```