Milestone 1: Project Proposal and Data Selection/Preparation

A. Preparing for my proposal

1. Which client/dataset did you select and why?

   I have chosen CoffeKing dataset because I love coffee and would like to analyze what kind of features people love and provide insights to the company.

2. Describe the steps you took to import and clean the data.

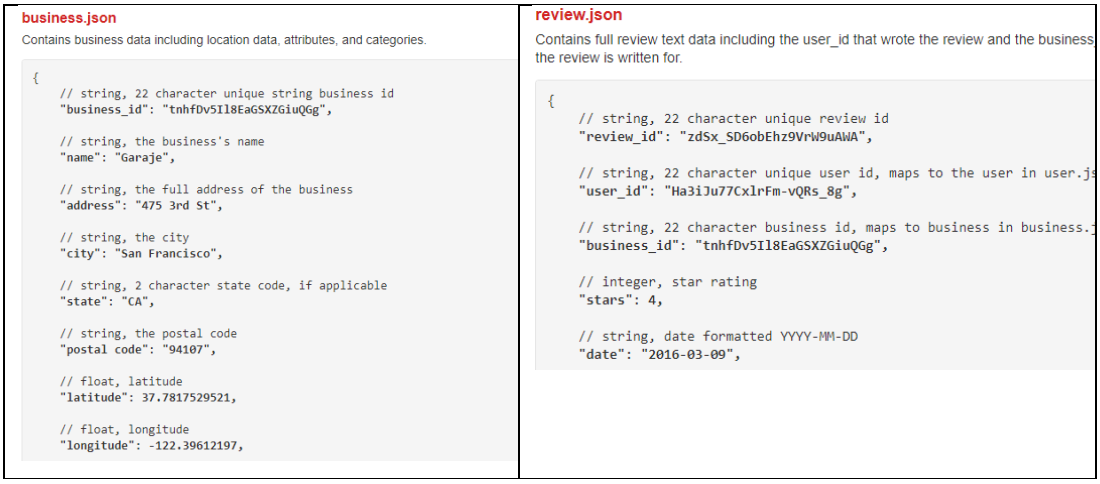
   I used Python programming language in Jupyter notebook to import and clean the dataset.

   For import, I used pandas library to read the business.json and review.json files.

   For cleaning the data, I only keep the business that are still open in the dataset, and find the business those related to coffee by filtering the categories.

3. Perform initial exploration of data and provide some screenshots or display some stats of the data you are looking at.

   i. Understand the data from the documentation. (https://www.yelp.com/dataset/documentation/main)



business.json

Contains business data including location data, attributes, and categories.

```
{
    // string, 22 character unique string business id
    "business_id": "tnhfDv5Il8EaGSXZGiuQGg",

    // string, the business's name
    "name": "Garaje",

    // string, the full address of the business
    "address": "475 3rd St",

    // string, the city
    "city": "San Francisco",

    // string, 2 character state code, if applicable
    "state": "CA",

    // string, the postal code
    "postal code": "94107",

    // float, latitude
    "latitude": 37.7817529521,

    // float, longitude
    "longitude": -122.39612197,
```

review.json

Contains full review text data including the user_id that wrote the review and the business the review is written for.

```
{
    // string, 22 character unique review id
    "review_id": "zdSx_SD6obEhz9VrW9uAWA",

    // string, 22 character unique user id, maps to the user in user.j
    "user_id": "Ha3iJu77CxlrFm-vQRs_8g",

    // string, 22 character business id, maps to business in business.
    "business_id": "tnhfDv5Il8EaGSXZGiuQGg",

    // integer, star rating
    "stars": 4,

    // string, date formatted YYYY-MM-DD
    "date": "2016-03-09",
```

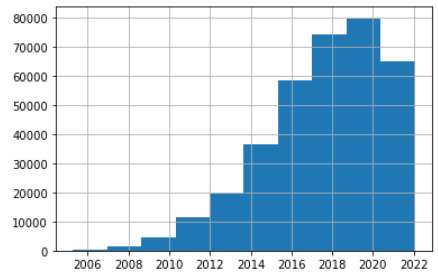
ii. Dataset profiling and understanding

```
# Find the earliest and latest review date
print('start_date: ', df.date.min())
print('end_date: ', df.date.max())
```

```
start_date:  2005-04-11 00:42:37
end_date:  2022-01-19 19:31:01
```

```
df.date.hist()
```

```
<AxesSubplot:>
```



iii. Find the relevant categories

```
print('Top 10 categories:')
df_explode.categories.value_counts()[:10]
```

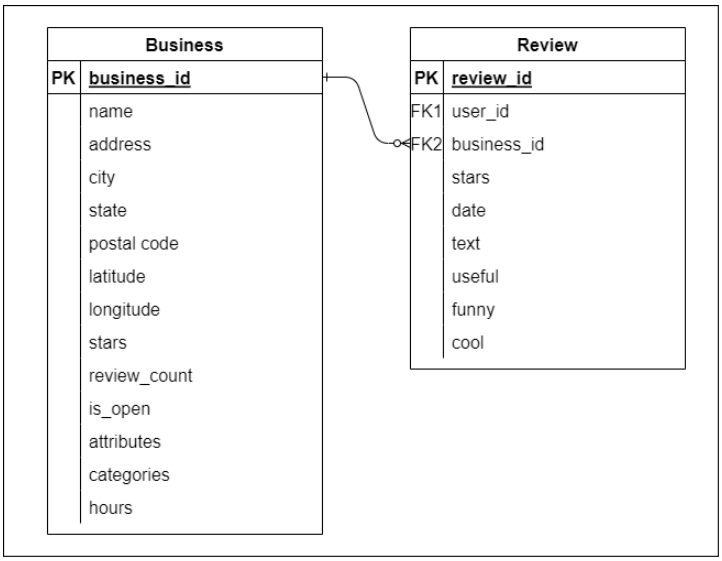
```
Top 10 categories:

Restaurants                 34987
Food                        20419
Shopping                    20186
Home Services               13322
Beauty & Spas               12263
Health & Medical            11046
Local Services              10138
Automotive                   9879
Nightlife                    8379
Event Planning & Services    8173
Name: categories, dtype: int64
```

```
df_explode[df_explode.categories.str.contains('Coffee',
                    case=True,na=False)].categories.value_counts()
```

```
Coffee & Tea             4954
Coffee Roasteries         238
Coffeeshops                 6
Coffee & Tea Supplies       5
Name: categories, dtype: int64
```

4. Create an ERD or proposed ERD to show the relationships of the data you are exploring.



B. Develop project proposal

Description

CoffeeKing is a new startup coffee company providing a unique and novel experience to their customers. They want to appeal to a wide variety of clientele. Hence, we utilize dataset from Yelp, a crowd-sourced reviews about businesses to provide insights to CoffeeKing.

Questions

1. Where should CoffeeKing open their shop?
2. What is the preferred operating hour for CoffeeKing?
3. What are the key features CoffeeKing should provide?

Hypothesis

1. Urban city will have higher reviews than rural city.
2. Coffee shop normally operating from morning until afternoon.
3. I think a coffee shop should have high quality coffee, foods, and pleasant environment.

Approach

1. I will be looking primarily at features such as review counts, star rating, location, and review context.
2. I would like to explore whether there is a relationship between operating hour and star rating.