Final Project Proposal

I. Group members, email addresses, and GitHub usernames

Member	Email	GitHub Username
Jinjie Fan	jinjie@seas.upenn.edu	JJFWWL
Yusheng Ding	yding3@seas.UPenn.edu	yding3
Guihe Li	guiheli@seas.upenn.edu	lihe14569
Anqi Wang	wanganqi@seas.upenn.edu	angelaiscoding

II. Description of application/website idea

We plan to use Yelp datasets that contain restaurants, reviews, and users information, with the County Health Ranking datasets that contain health ranking of each county, to create a website that has the following functions:

1. Restaurant recommender

This function implements logics that filter restaurants by criteria (e.g. location–state and city, star, review count, food category, etc.) from user input or favorite restaurant extracted from past reviews to get restaurant candidates. We design a recommendation algorithm that measures how well the candidates match with user's preferences and use the algorithm to rank the candidates. Finally, top ranked candidates will be returned.

2. Find the network of people who like the same restaurant

Users can find out people who like the same restaurants using this function. We will implement the network up to 3-connection through the restaurants where the users gave good reviews. This will help the user find out the network of people who liked their favorite restaurant, and potentially hang out together and make friends.

We define the notion of an 'n-connection' between people based on the restaurants they gave high star reviews (a high star review is defined as a review with 4+ stars). The user has a 0-connection with himself/herself. The user has a 1-connection with a person if they both reviewed the same restaurant and gave 4+ stars. The user has a 2-connection with a person if they have never given the same restaurant 4+ stars, but each has a 1-connection with the same person (via a different restaurant reviewed with 4+ stars). The user has a 3-connection with a person if they do not have a 1- or 2-connection with each other, but each has a 1- or 2-connection with the same person (via different restaurants reviewed with 4+ stars). And so on.

3. Restaurant scientist

This function provides summary statistics for users to better understand the restaurants in a given area that he/she is interested in. We will provide summary statistics in the following aspects:

- Restaurant count and percentage by star rating (1-5)
- Restaurant count and percentage by price level ($-$$$$)
- Restaurant count and percentage by type (Chinese, Mexican, Fast food, etc.)

- Relationship between health ranking and price level
- Relationship between health ranking and restaurant type

III. Data sources, exploratory analysis, and proposed queries

Description of the datasets:

- **Yelp Data:** The Yelp datasets contain businesses, reviews, user and photo data. They are available as JSON files.
- **County Health Rankings Data:** The County Health Rankings datasets contain rankings of each county's health outcomes and factors, which provide a snapshot of a community's health. They are available as excel files and can be converted to csv files.
- **ZIP Code Tabulation Area to County Bridge Data:** The crosswalk contains the mapping between zipcode and county-state FIPS code and will be used to link the Yelp dataset and the County Health Rankings data. It is available as excel file and can be converted to csv file.

Link of datasets:

- **Yelp Data:** https://www.yelp.com/dataset
- **County Health Rankings Data:** https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation
- **ZIP Code Tabulation Area to County Bridge Data:** https://www.statsamerica.org/geography-tools.aspx

Exploratory data analysis:

We conducted exploratory analyses of the datasets we are going to use, and included the relevant size statistics (e.g. mb/gb of the dataset, number of rows, and number of attributes) and summary statistics of important attributes (e.g. report mean, standard deviation).

Table 1: Yelp Buisness	
Size	124.4MB
Number of rows	160585
Number of attributes	14
Attribute: business_id	varchar(22), e.g. '6iYb2HFDywm3zjuRg0shjw'
Attribute: name	varchar(40), e.g. 'Oskar Blues Taproom'
Attribute: address	varchar(50), e.g. '921 Pearl St'
Attribute: city	varchar(30), e.g. 'Boulder'
Attribute: state	varchar(3), e.g. 'CO'

Attribute: postal_code	varchar(7), e.g. 80302
Attribute: latitude	float, e.g. 40.017544 (mean=38.76, std=7.13, min=27.998972, med=42.177366, max=49.490000)
Attribute: longitude	float, e.g. -105.283348 (mean=-94.27, std=19.98, min=-123.393929, med=-84.383281, max=71.113271)
Attribute: stars	float, e.g. 4.0 (mean=3.66, std=0.94, min=1.0, med=4.0, max=5.0)
Attribute: review_count	int, e.g. 86 (mean=51.96, std=130.03, min=5, med=17, max=9185)
Attribute: is_open	int, e.g. 1 (value_count={1: 123248, 0: 37337})
Attribute: attributes	JSON, e.g. {'RestaurantsTableService': 'True', 'WiFi': "u'free'", 'BikeParking': 'True', 'BusinessParking': "{'garage': False, 'street': True, 'validated': False, 'lot': False, 'valet': False}", 'BusinessAcceptsCreditCards': 'True', 'RestaurantsReservations': 'False', 'WheelchairAccessible': 'True', 'Caters': 'True', 'OutdoorSeating': 'True', 'RestaurantsGoodForGroups': 'True', 'HappyHour': 'True', 'BusinessAcceptsBitcoin': 'False', 'RestaurantsPriceRange2': '2', 'Ambience': "{'touristy': False, 'hipster': False, 'romantic': False, 'divey': False, 'intimate': False, 'trendy': False, 'upscale': False, 'classy': False, 'casual': True}", 'HasTV': 'True', 'Alcohol': "'beer_and_wine'", 'GoodForMeal': "{'dessert': False, 'latenight': False, 'lunch': False, 'dinner': False, 'brunch': False, 'breakfast': False}", 'DogsAllowed': 'False', 'RestaurantsTakeOut': 'True', 'NoiseLevel': "u'average'", 'RestaurantsAttire': "'casual'", 'RestaurantsDelivery': 'None'} * Some records have empty attributes in this instance
Attribute: categories	varchar(200), e.g. 'Gastropubs, Food, Beer Gardens, Restaurants, Bars, American (Traditional), Beer Bar, Nightlife, Breweries' * Some records have empty categories in this instance
Attribute: hours	JSON, e.g. {'Monday': '11:0-23:0', 'Tuesday': '11:0-23:0', 'Wednesday': '11:0-23:0', 'Thursday': '11:0-23:0', 'Friday': '11:0-23:0', 'Saturday': '11:0-23:0', 'Sunday': '11:0-23:0'} * Some records have empty hours in this instance

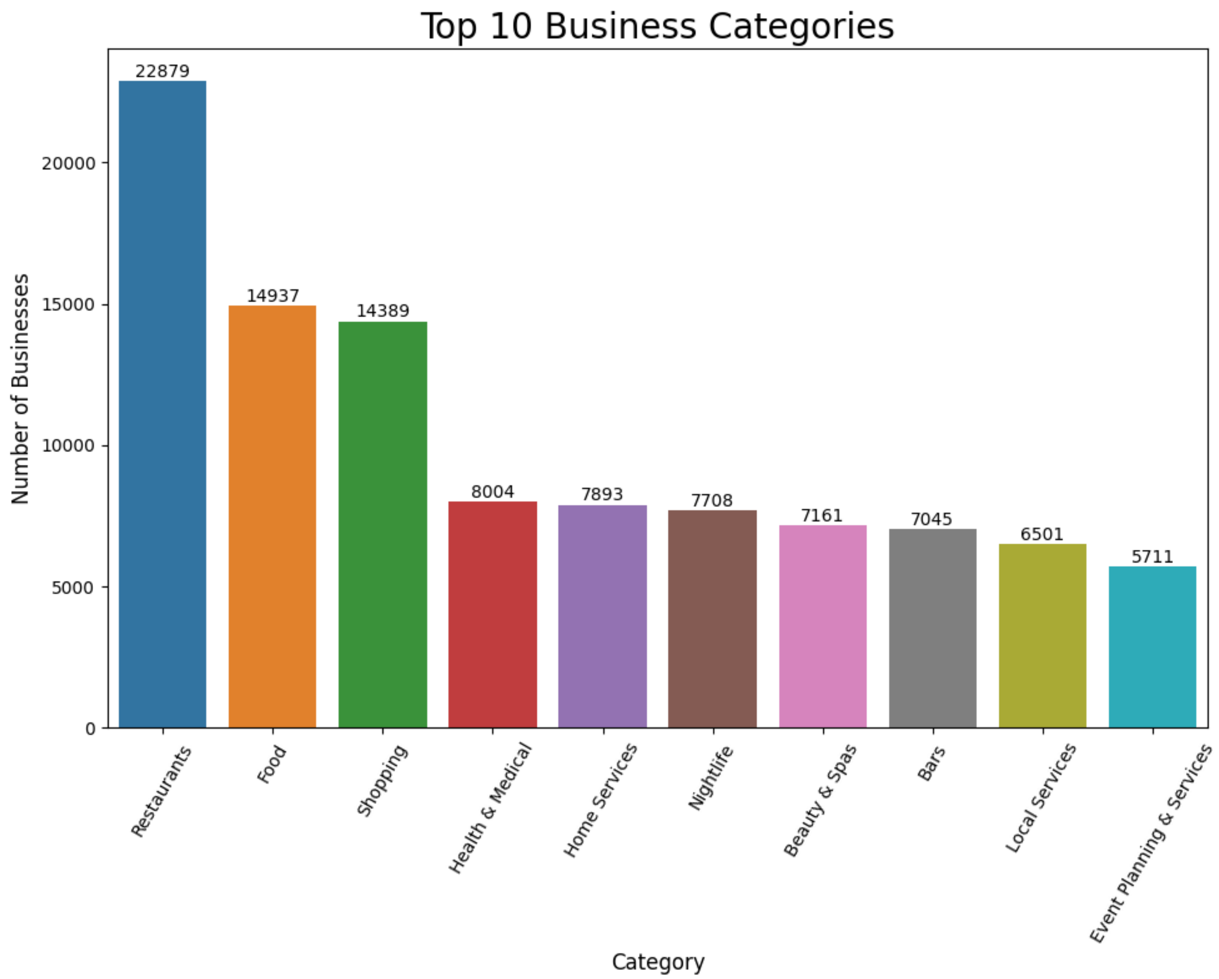

Table 2: Yelp Review	
Size	6.94GB
Number of rows	8635403
Number of attribute	9
Attribute: starts	mean=3.73, std=1.45, min=1.0, 25%, 3.0, 50%= 4.0, 75%=5.0, max= 5.0
Attribute: useful	mean=1.24, std=3.20, min=0, 25%=0, 50%=0, 75%=1,max=758
Attribute: funny	mean=0.42, std=1.86, min=0, 25%=0, 50%=0, 75%=0, max=610
Attribute: cool	mean=0.50, std=2.24, min=0, 25%=0, 50%=0, 75% =0, max=732

Table 3: Yelp User	
Size	3788.8MB
Number of rows	2189457
Number of attributes	22
Attribute: average_stars	mean=3.65, std=1.15, min=1, 25%=3, 50%=3.88, 75% =4.55, max=5
Attribute: review_count	mean=21.7, std=76, min=0, 25%=2, 50%=5, 75%=15, max=15686
Attribute: useful	mean=38, std=535, min=0, 25%=1, 50%=3, 75%=1.3, max=20438

Table 4: Yelp Photo	
Size	58.6MB
Number of rows	200000
Number of attributes	4
Attribute: photo_id	unique=199999, freq=2
Attribute: business_id	unique=39438, freq=493
Attribute: caption	unique=86333, freq=88090
Attribute: label	unique=5, freq=4000

Table 5: ZIP Code Tabulation Area to County Bridge	
Size	1.2MB
Number of rows	32989
Number of attribute	5 (useful attributes for our purpose: 3)
Attribute: ZIP	varchar(5), e.g. 27107, unique = 32989
Attribute: censusfips	varchar(5), e.g. 18055, unique=3137
Attribute: censusname	varchar(21), e.g. Santa Barbara, unique=1830

Table 6: County Health Rankings (Outcomes & Factors SubRankings)	
Size	182KB
Number of rows	3193
Number of attributes	16 (useful attributes for our purpose: 10)
Attribute: FIPS	varchar(5), e.g. 56045, unique = 39488
Attribute: state	varchar(20), e.g. New York, unique=51
Attribute: number_of_ranked_counties	int, e.g. 2 (mean=93.98, std=55.13, min=1, med=83, max=243)
Attribute: length_of_life_rank	int, e.g. 1 (mean=47.18, std=41.57, min=1, med=37, max=243)
Attribute: quality_of_life_rank	int, e.g. 2 (mean=47.82, std=41.73, min=1, med=38, max=243)
Attribute: health_behavior_rank	int, e.g. 3 (mean=47.82, std=41.73, min=1, med=38, max=243)
Attribute: clinical_care_rank	int, e.g. 4 (mean=47.82, std=41.73, min=1, med=38, max=243)
Attribute: social_and_economic_factors_rank	int, e.g. 5 (mean=47.82, std=41.73, min=1, med=38, max=243)
Attribute: physicial_environment_rank	int, e.g. 6 (mean=47.82, std=41.73, min=1, med=38, max=243)

IV. Proposed queries

1. Restaurant recommender

- A query that joins the datasets photo and business, which can show the required business information and related photos based on the user selected criteria (e.g. star, categories, location, etc.)
- A query that joins the datasets photo, business, user and review. This query can get the user's list of highest star reviewed businesses from a subquery using the datasets user and review, and then provide a list of recommended other businesses that are similar to the subquery result in terms of categories, stars, etc.

2. Find the network of people who like the same restaurant

- A query that joins the dataset photo, business, user and review. This query can get the user's list of high star (4+ stars) reviewed businesses from a subquery using the datasets user and review, and then find out the network (n-connection) of other users who gave high star reviews for the same or related businesses, to provide a friend recommendation list for the user.

3. Restaurant scientist

- A query that searches from dataset business to find the most reviewed highest star business
- A query that searches from dataset business to find the worst business, which has the lowest star value.
- A query that searches from dataset business to find the best business, which has the highest star value and sorted by number of reviews.
- A query that calculates the count and percentage of businesses in an area (aggregated by zipcode, or county, or state) by star level from the dataset business
- A query that calculates the count and percentage of businesses in an area (aggregated by zipcode, or county, or state) by price level from the dataset business
- A query that calculates the count and percentage of businesses in an area (aggregated by zipcode, or county, or state) by restaurant type from the dataset business
- A query that calculates the average price level and review count in an area (aggregated by zipcode, or county, or state) from the dataset business
- A query that joins the datasets business and county health rankings, provide each area's (county, or state) health ranking and star rating frequency distribution
- A query that joins the datasets business and county health rankings, provide each area's (county, or state) health ranking and price range frequency distribution
- A query that joins the datasets business and county health rankings, provide each area's (county, or state) health ranking and restaurant type frequency distribution