Title:

Group F: Vege

Member:

Lin Zhao (lz2711@columbia.edu), Ming Gong (mg4122@columbia.edu)
Taketo Muroya (tm3068@columbia.edu), Wenjun Sun (ws2578@columbia.edu)

Abstract:

We try to visualize the spatial patterns of vegetarian restaurants across the United States using mapping methods, text analysis, and interactive functions. The primary data would be vegetarian and vegan restaurants data (the number of restaurants, locations, menus, reviews and so on) to examine regional differences about vegetarian food consumption. Besides, we are planning to combine these restaurant data with other spatial data related to the vegetarian attitudes, such as ideologies by State level (e.g. conservative or liberal), online search words related to vegetarian over years(e.g. Google Trend).

Data:

- Vegetarian Restaurants
  - [first choice] Restaurants that serve vegetarian or vegan food in the US
    https://www.kaggle.com/datafiniti/vegetarian-vegan-restaurants
    The primary data source would be "Vegetarian and Vegan Restaurants" on Kaggle, which includes U.S. vegetarian or vegan restaurants' address, city, state, business name, business categories, menu data, phone numbers, and more. The dataset size is over 200 restaurants and their 10,000 vegan or vegetarian menu items updated between January 2018 and September 2018. However, this version is a sample of a large data set. The full data set is available through the data company, Datafiniti's Business Database, which includes over 18,000 restaurants in the US that serve vegetarian or vegan food. Since this full dataset is not free and the size of Kaggle data may not be enough to be used for visualizations, we are now trying to figure out the way to get the full data (asking the Director Greg Eirich to see if QMSS can sponsor us or the Datafiniti to see if they can give us the dataset for academic purpose - updates: Datafiniti said no). Otherwise, we may be able to get the partial data with lower price by narrowing down the data such as focusing on the restaurants in NYC (according to Datafiniti database, the number of vegetarian restaurants in NYC would be around 1,300, which could be affordable data size and enough for visualizations).

    Data items for each restaurant (Kaggle data)

Id	dateAdded	dateUpdated	**address**	**categories**	primaryCategories	**city**	claimed	country	**cuisines**
descriptions.dateSeen	descriptions.sourceURLs	descriptions.value	facebookPageURL	features.key	features.value	hours.day	hours.dept	hours.hour	imageURLs
isClosed	keys	languagesSpoken	**latitude**	**longitude**	menuPageURL	menus.amountMax	menus.amountMin	**menus.category**	menus.currency
menus.dateSeen	menus.description	**menus.name**	menus.sourceURLs	Name	paymentTypes	Phones	postalCode	priceRangeCurrency	**priceRangeMin**
priceRangeMax	province	Sic	sourceURLs	Twitter	Websites	yearOpened			

- Happycow.net
  https://www.happycow.net/
  This website is like Yelp for vegetarian and vegan restaurants. They do not have API, so if we are going to use this source, then we probably need to do some web scraping.
  The difference between this dataset and the Datafiniti one is that, this dataset can focus on *vegetarian- and vegan- only restaurants*, but the Datafiniti dataset includes all restaurants that *serve* vegetarian- or vegan- food.

- Consumer

We plan to integrate some data about vegetarians. Although the data we found are aggregated ones, we think it could still be helpful to add this kind of information.

- Share of consumers who consider themselves vegan or vegetarian in the United States as of June 2018
  https://www.statista.com/statistics/738851/vegan-vegetarian-consumers-us/#statisticContainer
  https://news.gallup.com/poll/238328/snapshot-few-americans-vegetarian-vegan.aspx?g_source=link_NEWSV9&g_medium=NEWSFEED&g_campaign=item_&g_content=Snapshot%3a%2520Few%2520Americans%2520Vegetarian%2520or%2520Vegan
  This data is aggregated according to age group, income group, political ideology

  Another websites would be helpful too:
  https://veganbits.com/vegan-demographics/(data updated on 2020)
  It provides vegan demographics in the United States, such as how many people are vegans in the U.S., why are u vegan?, most vegan-friendly cities in America.

  Beyond looking at vegan distribution by age, income and political ideology, we can also consider gender, race, and religious affiliation.
  https://faunalytics.org/who-are-the-vegetarians/
  https://vegansofcolor.wordpress.com/2009/08/30/update-of-vegans-by-raceethnicity-survey/(2009)

- Consumer attitudes towards the vegetarian diet in the US in 2016
  https://www.statista.com/statistics/591963/us-consumer-attitudes-towards-the-vegetarian-diet/

- How many people are vegetarian or vegan? from the Vegetarian Resource Group
  https://www.vrg.org/nutshell/faq.htm#poll

- Recipes

We are also interested in doing text analysis on vegetarian recipes, to see what the most popular ingredients in the vegetarian and vegan world.

- [first choice] Spoonacular API
  https://spoonacular.com/
  They provide discounted API for students at $10, we applied to it and got it.
  This API has a parameter that classifies recipes according to diets, which includes vegetarian and vegan.

- [first choice] Early American Cookbooks
  https://babel.hathitrust.org/cgi/mb?a=listis;c=1934413200

1450 cookbooks published in the United States from 1800 to 1920.
All of the titles in the collection are in the public domain and are available in full-text on the HathiTrust Digital Library.
We plan to use this dataset to show how the terms "vegetarian" and "vegan" get popular over time.

- [alternative] Recipe collection from the Vegetarian Resource Group
  https://www.vrg.org/journal/CookingAndRecipes.htm
  A website collects vegetarian recipes. Probably need web scraping to get the data. But the web pages are not in a consistent format. So we probably won't go for this one.

- [alternative] BudgetByte.com
  https://www.budgetbytes.com/category/recipes/
  This is a recipe website; it has vegetarian and vegan categories. We can try web scraping it if other approaches don't work.

- Google trends
  - gtrendsR package
    https://cran.r-project.org/web/packages/gtrendsR/gtrendsR.pdf
    This R package allows us to request Google trends data in the US, in each state, city. We plan to use this to see how the popularity of vegetarian and vegan food gets popular.

- Nutrition
  - [first choice] Food and Nutrient Database for Dietary Studies, U.S. Department of Agriculture
    https://fdc.nal.usda.gov/help.html
    https://fdc.nal.usda.gov/download-datasets.html
    "FNDDS Nutrient Values.xlsx" at https://fdc.nal.usda.gov/data-documentation.html
    This data can be downloaded as a csv file or accessed using API.

- Political position of state/county
  - [first choice] David Leip's Election Atlas
    http://uselectionatlas.org/
    We plan to use the election data of each state/county in the 2016 election as an indicator of its liberal/conservative position. We have already used this data in week5's lecture.
    If we can get full access to the Datafiniti dataset, then we will do it on state level, if we can't, then we will focus on NY and do it on county level.

- City (optional)
  - Most Vegetarian- & Vegan-Friendly Cities
    https://wallethub.com/edu/best-cities-for-vegans-vegetarians/39706/
    Include 100 US cities, rank them based on affordability, diversity, and vegetarian lifestyle

Outlines:

- Map
  - Use maps to show the number of vegetarian/vegan restaurants in different state/county
  - Show the association between popularity of vegetarian/vegan food and political

position

- Word Cloud
  - Use word cloud to show the popularity of different ingredients in vegetarian/vegan recipe and show the key words to represent the popularities in vegetarian/vegan restaurants' reviews
  - We can also use word cloud to visualize most common words in customer reviews
- Scatterplot
  - How the association between the popularity of each ingredient and its nutrition (energy/protein)
- Line graph
  - Use data from Early American Cookbooks to show how vegetarian/vegan food get trending over time
- Ridgeline chart
  - Show Google Trend of "vegetarian/vegan" in each state over time
- Bar graph
  - Show the aggregated data about customers

Concerns:

- We may have data from different years, would that be a problem? For example, the share of vegetarian data is from 2018 but the attitude towards vegetarian diet is from 2016.
- Besides, the location information may vary across datasets. Some data may have exact restaurants' addresses, while other data may have longitude and latitude information. And some other data such as political ideology would be given by state/county level. So we have to think about the geocoding to integrate and match the location information.