An geo-spatial exploration and analysis of the health food market in San Francisco

IBM Applied Data Science Capstone

Final Project Report

By

Anzar A. Syed

Table of Contents

Introduction

Background

Consumers across the world are becoming increasingly more conscious of their food and lifestyle choices. This is trend is even more dominant for younger consumers, particularly millennial and Gen Z consumers.[1] Key trends in this area are the incredible growth of the overall health and wellness market and well as specific areas of the market such as the organic and vegan food markets.

Health and Wellness Food market is expected to see growth rate of 5.7% and may see market size of USD 1,253 billion by 2024.[2]

The vegan and vegetarian food market is a high growth segment with the number of vegans in the United States jumping from 1% in 2014 to 6%in 2017, while an estimated one in three Americans now identify as "flexitarians." [3] This translates into an important industry trend as seen by the remarkable 20% growth in the dollar sales of plant-based food in the US versus a 2% growth in overall food sales. [4] While more dominant in the US market, this trend is on the rise globally as well, for instance Sales of organic food and drink in the UK rose by 4.5% in 2019 to a record GBP 2.45 billion. [5]

The organic food market is another high potential market segment. According to Market Research, it is estimated that the global organic food and beverages market will reach USD 323.56 Billion by 2024. [6]

San Francisco in amongst the leading major cities in the US and the world in the upsurge in this health-conscious food trend. It is considered one of the top 5 most vegetarian and vegan friendly cities in the US by Vitacost as well as the best city to live the organic foodie lifestyle by Organic Authority. [7][8] Lucy McDonald, a writer for the Guardian, calls San Francisco "the crucible of the world's health food movement." [9]

Problem and Interest

The trends discussed make San Francisco one of the most promising locations to venture into the health food market. However, as a consequence of the overall market's rapid growth and its massive popularity and success in San Francisco, any potential business venture in this space faces incredible competition. We turn to some data and analyse it to make sense of this competition and gain actionable insights such as market concentration, geographic saturation and other similar trends.

This analysis would be useful to individuals and businesses that are interested in venturing into the health food market in San Francisco and would like to explore such data to help their analysis.

Data Acquisition and Wrangling

For the Analysis, I will be using the Foursquare API to explore the features of neighbourhoods in San Francisco and to acquire data of various venue categories related to the health food market, namely Organic Grocery, Farmers Market and Vegetarian / Vegan Restaurant.

I will also scrape the following webpage for data of San Francisco neighbourhoods and their corresponding Zip codes: , http://www.healthysf.org/bdi/outcomes/zipmap.htm as well as the python database uszipcode to get the longitude and latitude values corresponding to each zip code.
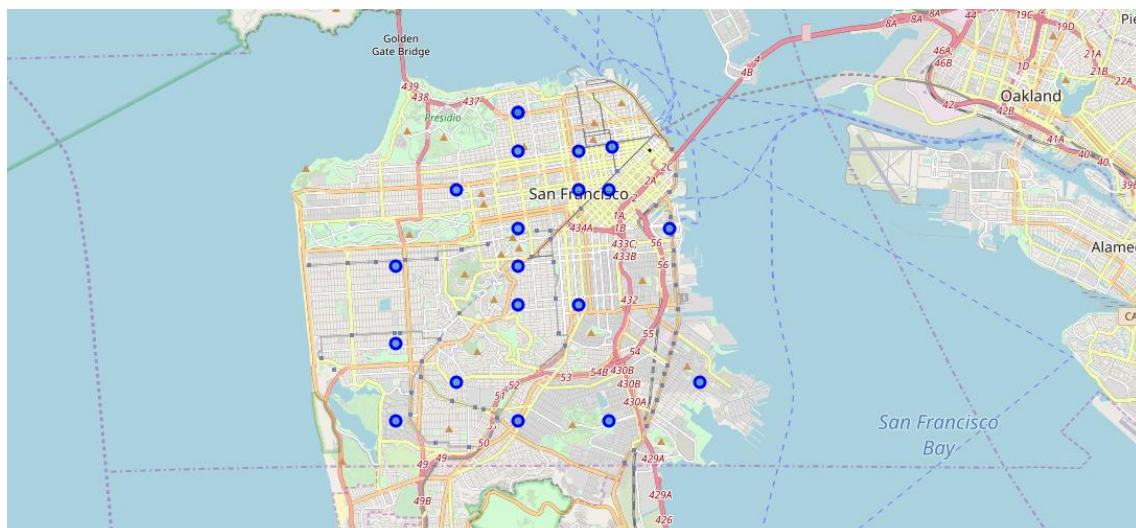
The geopy geocoder class Nominatum will also be used for geocoding certain fields of data into longitude and latitude values.

I will use segmentation and clustering to find similar neighbourhoods and analyse the prevalence of existing health food businesses across the neighbourhoods to find potential gaps in the market.

For visualization, I will use the Folium library to map the neighbourhoods in San Francisco and their emerging clusters.

Methodology and Results

I started by scraping the following Wikipedia page: http://www.healthysf.org/bdi/outcomes/zipmap.htm using the Beautiful Soup Python package for parsing HTML and XML documents and obtaining the data for neighborhoods in San Francisco and their corresponding Zip Codes in the table as a dataframe. In order to utilize Foursquare location data, I needed to get the latitude and the longitude coordinates of each neighbourhood; I used the Python database uszipcode and its function SearchEngine to acquire this and the geopy geocoder class Nominatum was also be used for geocoding the address San Francisco into longitude and latitude values. I then used the Folium Library to visualize the geo-spatial data of neighbourhoods of San Francisco on the following map:

After this I connected with the Foursquare API to begin the actual analysis. I began by getting data of the top 100 venues in a 500 metre radius of the centre of the neighbourhoods from the Foursquare API and organising it.

This was followed by one-hot encoding, which is process to turn the qualitative data into quantitative data for further analysis, and the grouping the resultant data by neighbourhood.

I then ran some code to see the top 10 most common venue categories for each neighbourhood. This gave us the following dataframe:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Bayview-Hunters Point	Food & Drink Shop	Motorcycle Shop	Park	Coffee Shop	Restaurant	Ethiopian Restaurant	Food	Fondue Restaurant	Flower Shop	Fish Market
1	Castro/Noe Valley	Gay Bar	Park	Yoga Studio	Grocery Store	Thai Restaurant	Clothing Store	Coffee Shop	Playground	Wine Bar	Pet Store
2	Chinatown	Hotel	Coffee Shop	American Restaurant	Boutique	Men's Store	Cocktail Bar	Church	Jewelry Store	Spa	Café
3	Haight-Ashbury	Coffee Shop	Tennis Court	Grocery Store	Yoga Studio	Dog Run	Boutique	Bookstore	Restaurant	Gastropub	Comic Shop
4	Hayes Valley/Tenderloin/North of Market	Coffee Shop	Cocktail Bar	Performing Arts Venue	Vietnamese Restaurant	Café	French Restaurant	Theater	Beer Bar	Mexican Restaurant	Southern / Soul Food Restaurant

I then wanted to segment and cluster the data gathered and analysed so far. To do so I first merged the dataframe above with the dataframe of geo-spatial data of Zip Codes, Longitudes and Latitudes that I had created earlier. The resultant dataframe was as follows:

	Zip Code	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	94102	Hayes Valley/Tenderloin/North of Market	37.780	-122.420	3	Coffee Shop	Cocktail Bar	Performing Arts Venue	Vietnamese Restaurant	Café	French Restaurant	Theater	Beer Bar	Mexican Restaurant	Southern / Soul Food Restaurant
1	94103	South of Market	37.780	-122.410	3	Coffee Shop	Sandwich Place	Theater	American Restaurant	Café	Bakery	Vietnamese Restaurant	Wine Bar	Restaurant	Plaza
2	94107	Potrero Hill	37.770	-122.390	2	Food Truck	Gym	Coffee Shop	Pharmacy	Performing Arts Venue	Pizza Place	Harbor / Marina	Park	Basketball Stadium	Café
3	94108	Chinatown	37.791	-122.409	3	Hotel	Coffee Shop	American Restaurant	Boutique	Men's Store	Cocktail Bar	Church	Jewelry Store	Spa	Café
4	94109	Polk/Russian Hill (Nob Hill)	37.790	-122.420	3	Grocery Store	Massage Studio	Sushi Restaurant	Bar	Thai Restaurant	Vietnamese Restaurant	Bakery	Diner	Coffee Shop	Pet Store

The next step was visualizing the clusters on a map using Folium:
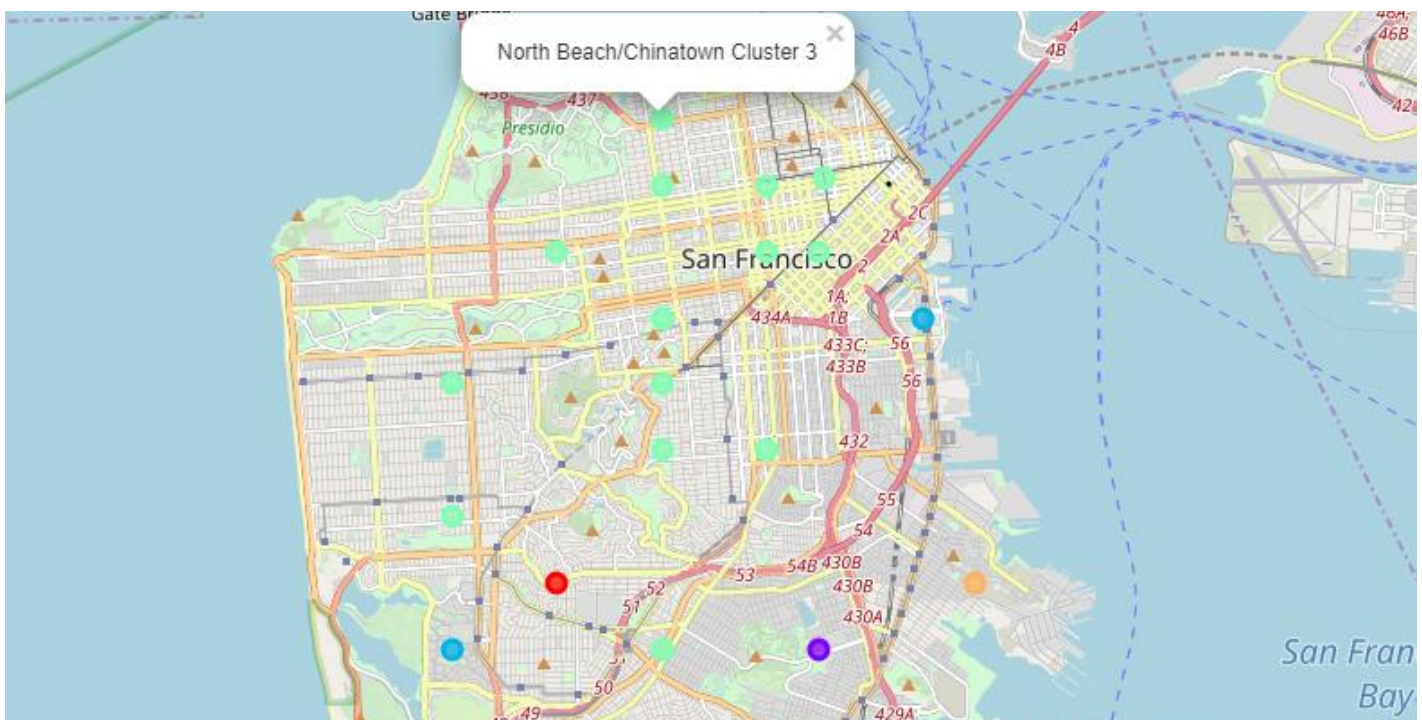
This was followed by examining each cluster in more detail as follows:

Cluster 0

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
15	St. Francis Wood/Miraloma/West Portal	Fountain	Bus Line	Park	Scenic Lookout	Event Space	Food & Drink Shop	Food	Fondue Restaurant	Flower Shop	Fish Market

Cluster 1

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
19	Visitacion Valley/Sunnydale	Garden	Park	Baseball Field	Yoga Studio	Event Space	Food & Drink Shop	Food	Fondue Restaurant	Flower Shop	Fish Market

Cluster 2

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
2	Potrero Hill	Food Truck	Gym	Coffee Shop	Pharmacy	Performing Arts Venue	Pizza Place	Harbor / Marina	Park	Basketball Stadium	Café
17	Lake Merced	Gym	Performing Arts Venue	Park	Sandwich Place	Café	Paper / Office Supplies Store	Coffee Shop	Rental Car Location	Cocktail Bar	Fish Market

Cluster 3

|	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Hayes Valley/Tenderloin/North of Market	Coffee Shop	Cocktail Bar	Performing Arts Venue	Vietnamese Restaurant	Café	French Restaurant	Theater	Beer Bar	Mexican Restaurant	Southern / Soul Food Restaurant
1	South of Market	Coffee Shop	Sandwich Place	Theater	American Restaurant	Café	Bakery	Vietnamese Restaurant	Wine Bar	Restaurant	Plaza
3	Chinatown	Hotel	Coffee Shop	American Restaurant	Boutique	Men's Store	Cocktail Bar	Church	Jewelry Store	Spa	Café
4	Polk/Russian Hill (Nob Hill)	Grocery Store	Massage Studio	Sushi Restaurant	Bar	Thai Restaurant	Vietnamese Restaurant	Bakery	Diner	Coffee Shop	Pet Store
5	Inner Mission/Bernal Heights	Mexican Restaurant	Grocery Store	Dive Bar	Italian Restaurant	Bakery	Cocktail Bar	Gym / Fitness Center	Pizza Place	Massage Studio	Coffee Shop
6	Ingelside-Excelsior/Crocker-Amazon	Pizza Place	Mexican Restaurant	Chinese Restaurant	Vietnamese Restaurant	Coffee Shop	Café	Sandwich Place	Bar	Pharmacy	Gas Station
7	Castro/Noe Valley	Gay Bar	Park	Yoga Studio	Grocery Store	Thai Restaurant	Clothing Store	Coffee Shop	Playground	Wine Bar	Pet Store
8	Western Addition/Japantown	Park	Yoga Studio	Sandwich Place	Sushi Restaurant	Chinese Restaurant	Furniture / Home Store	Spa	Historic Site	Salon / Barbershop	Playground
9	Parkside/Forest Hill	Chinese Restaurant	Park	Pizza Place	Bubble Tea Shop	Light Rail Station	Sandwich Place	Café	Shoe Repair	Breakfast Spot	Burrito Place
10	Haight-Ashbury	Coffee Shop	Tennis Court	Grocery Store	Yoga Studio	Dog Run	Boutique	Bookstore	Restaurant	Gastropub	Comic Shop
11	Inner Richmond	Sushi Restaurant	Japanese Restaurant	Pizza Place	Bakery	Burger Joint	Vietnamese Restaurant	Burmese Restaurant	Wine Shop	Italian Restaurant	Mobile Phone Shop
12	Sunset	Vietnamese Restaurant	Bubble Tea Shop	Bakery	Chinese Restaurant	Deli / Bodega	Grocery Store	Szechuan Restaurant	Thai Restaurant	Bank	Dim Sum Restaurant
13	Marina	Italian Restaurant	French Restaurant	Salad Place	Wine Bar	Gym / Fitness Center	Thai Restaurant	Diner	Mexican Restaurant	Bar	Burger Joint
16	Twin Peaks-Glen Park	Park	Yoga Studio	Café	Gift Shop	Library	Coffee Shop	Mexican Restaurant	North Indian Restaurant	Thai Restaurant	Bookstore
18	North Beach/Chinatown	Italian Restaurant	French Restaurant	Salad Place	Wine Bar	Gym / Fitness Center	Thai Restaurant	Diner	Mexican Restaurant	Bar	Burger Joint

Cluster 4

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
14	Bayview-Hunters Point	Food & Drink Shop	Motorcycle Shop	Park	Coffee Shop	Restaurant	Ethiopian Restaurant	Food	Fondue Restaurant	Flower Shop	Fish Market

As we can see from the map and the tables, Cluster 3 is the largest, followed by Cluster 2. Neighbourhoods in these clusters have similar venues and features. This will help in further analysis.

I then make a new dataframe with only the venue categories mentioned earlier that are of interest to us for our analysis and cluster this data into 5 clusters to compare with the overall neighbourhood clusters and try to gain some insights.

To do so, as shown earlier I merge this dataframe with the earlier dataframe of geo-spatial data of Zip Codes, longitudes and latitudes to create a new dataframe. This new dataframe is used to create a map of the clusters of neighbourhoods with similar health food venues based on the related venue categories we selected:
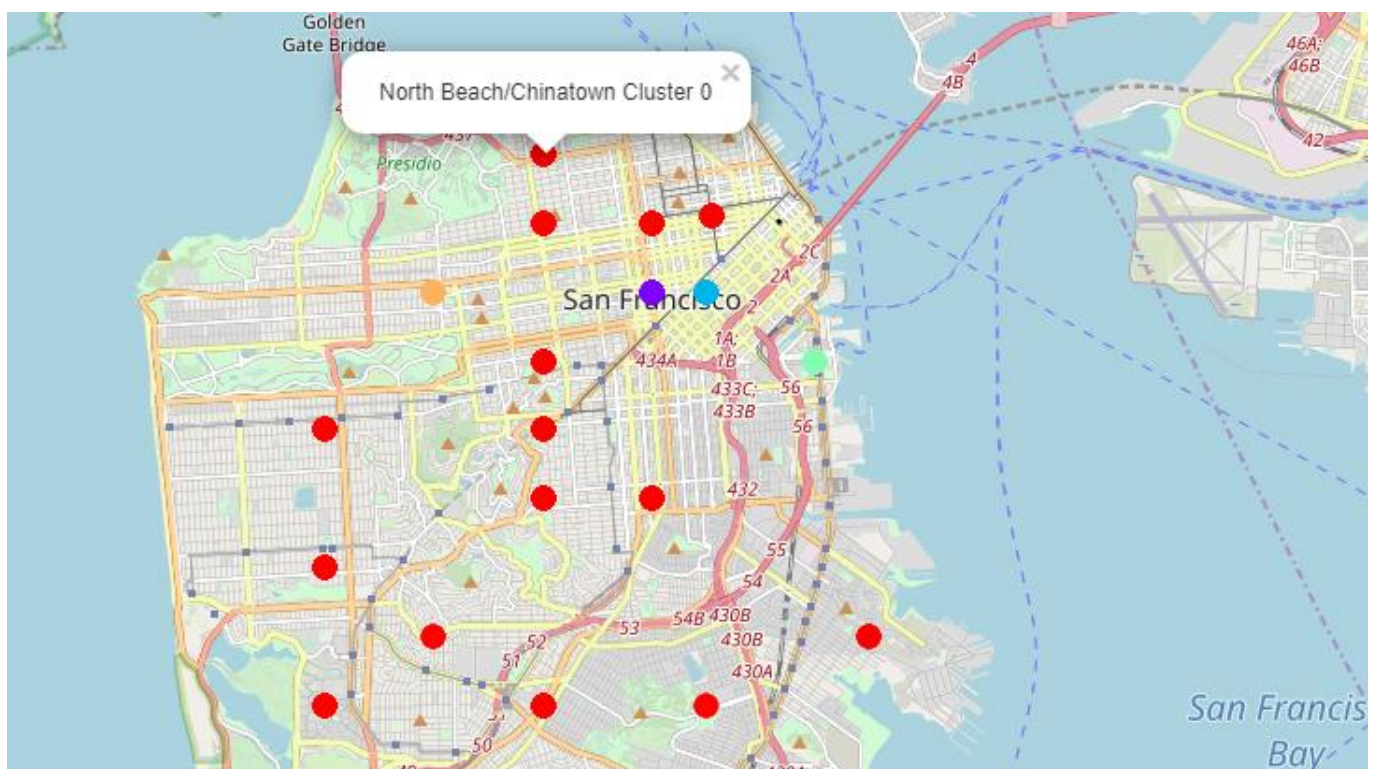
Discussion

As we can see from the map, there are several differences between the clusters generated of the overall neighbourhoods and those generated when focusing on the health food related categories we focused on

The neighbourhoods of Hayes Valley/Tenderloin/North of Market, South of Market and Inner Richmond are all part of the same biggest cluster, cluster 3 of the neighbourhoods overall, but when evaluated on health food venue categories, they all stand alone as distinct and dissimilar clusters

Similarly, Lake Merced, which overall is more similar to Potrero Hill, here stands as part of Cluster 0, the largest cluster of neighbourhoods similar in terms heath food venues

The neighbourhoods of Bayview-Hunters Point, Visitacion Valley/Sunnydale and St. Francis Wood/Miraloma/West Portal are all also part of Cluster 0 here and similar in terms of health food venues, whereas in terms of overall venues and similarity they are all unique.

These insights can prove to be very useful in determining a successful location for a new health food venture by a business or entrepreneur or to determine the potential for a new similar or related venue in a particular neighbourhood.

They uncover some interesting findings, such as neighbourhoods which may currently be underserved and where there exists a higher possibility of success for a new venture as well as neighbourhoods with similar overall traits and consumer behaviour but where the health food market has yet to maturely develop.

Conclusion

The results of out analysis hold important insights into the geographical concentration and saturation of the health food market in San Francisco at a neighbourhood level. As always however, this analysis is however limited by the limitation of the data and parameters used. A greater number of parameters such as a higher number of top venues, a larger radius, more categories of venues and other forms of data such as resident demographics and consumer behavior. A comparison with cities with similar health food trends to those of San Francisco could also be useful to improve the rigour of the analysis as well as inclusion of non-food venues related to health food such as gyms and yoga studios.

References

By. "Health and Wellness Food Market Is Expected to See Growth Rate of 5.7% and May See Market Size of USD1253 Billion by 2024." MarketWatch, September 6, 2019. https://www.marketwatch.com/press-release/health-and-wellness-food-market-is-expected-to-see-growth-rate-of-57-and-may-see-market-size-of-usd1253-billion-by-2024-2019-09-06.

Chokhani, Rishabh. "Increase in the Global Demand of Organic Food: Decoding This Trend." Entrepreneur, July 23, 2019. https://www.entrepreneur.com/article/337171.

Egan, John. "The 5 Most Vegetarian- and Vegan-Friendly Cities in the U.S." VitaVibes Blog | Vitacost.com, January 3, 2018. https://www.vitacost.com/blog/food-nutrition/the-5-most-vegetarian-and-vegan-friendly-cities-in-the-u-s.html.

Ettinger, Jill. "Top 9 Cities to Live the Organic Foodie Lifestyle." Organic Authority, January 31, 2012. https://www.organicauthority.com/buzz-news/top-9-us-cities-to-live-in-if-you-love-organic-food.

Gustafson, Timi. "Younger Consumers Are More Health Conscious Than Previous Generations." HuffPost Canada. HuffPost Canada, January 23, 2017. https://www.huffingtonpost.ca/timi-gustafson/younger-consumers-are-mor_b_14290774.html.

Kateman, Brian. "Vegan Restaurants Are On The Rise." Forbes. Forbes Magazine, August 21, 2019. https://www.forbes.com/sites/briankateman/2019/08/21/vegan-restaurants-are-on-the-rise/#461325d31e80.

McDonald, Lucy. "San Francisco's Food Scene Is Bursting with Health (and Kale)." The Telegraph. Telegraph Media Group, June 25, 2014. https://www.telegraph.co.uk/foodanddrink/healthyeating/10923116/San-Franciscos-food-scene-is-bursting-with-health-and-kale.html.

Orlando, Brian. "Retail Sales Data 2018." Plant Based Foods Association. Accessed February 29, 2020. https://plantbasedfoods.org/marketplace/retail-sales-data-2018/.

Smithers, Rebecca. "Organic Food and Drink Sales Rise to £2.45bn." The Guardian. Guardian News and Media, February 5, 2020. https://www.theguardian.com/business/2020/feb/05/organic-food-and-drink-sales-rise-to-245bn.