Search Data and Geodemographics Determinants of the Avocado Sales in the US Markets

Di Wu [1,*]
Zhenning Xu [1]
Ji Li [1]

* Corresponding author, dwu2@csub.edu
[1] California State University, Bakersfield

Abstract

Purpose – *This work aims at providing insights into avocado sales and markets in the US by leveraging advanced data analytics and data visualization methods to model the industry data combined with geodemographics and search data.*

Method – *The dataset from the Hass Avocado Board (HAB), collected geodemographic data and search data across multiple metropolitan regions in the US are studied by employing quantitative methods and tools, such as principal component analysis (PCA), cluster analysis, data visualization, and regression models.*

Findings – *The results show that avocado sales and markets are highly affected by geodemographic factors and strongly correlated with the search data in the internet community. Income, Age, Ethnicity, and Education significantly correlate with avocado sales. Not only are Google Trends data correlated with avocado sales, but they also can further improve the explanatory power of regression models. Data analytics and visualization tools can be particularly useful for exploring market data in various areas, such as market segmentation, consumer preferences for different avocados (organic vs. conventional), seasonality, and market patterns. Geographically distant cities may have similar consumer markets for avocado sales.*

Limitations – *The study focuses on the limited panel data from 2015 to 2019 due to the COVID-19 disruption in the 2020 US Census data publication and hence the lack of the geodemographics data in and after 2020, which may not be used to study post-COVID markets. Other limitations include that the study may have possible contemporaneous correlations between conventional and organic avocado sales and only focus on the avocado markets with specific conditions. Therefore, the models developed in this research may need further expanded and studied for different food markets and retailing industries for future improvement and applicability.*

Implications – *The study offers insights into the regional trends and search trends of consumers for potentially better sales and market opportunities. The paper has the potential to advance the understanding of the role of geodemographics and public interest in commercial products. From descriptive and diagnostic perspectives, our work harnesses k-means clustering and hierarchical clustering methods as well as exploratory mapping methods to offer data-driven visualizations, which is one of few studies focusing on food markets. Specifically, the results provide useful information and knowledge of the avocado markets to policymakers, producers, distributors, retailers, and consumers as the trend for health-conscious consumption keeps growing in different regions. Furthermore, the study may have implications for practitioners and scholars interested in other agricultural products.*

Originality – *This paper is among the first to underscore heterogeneous characteristics of consumer appetite for avocados across different regions in the US and to provide evidence and examples on how search data and demographic information can be beneficial to further understanding the sales data, exploring markets in various aspects, and potentially developing effective and efficient sales and marketing strategies by region.*

Keywords: clustering and segmentation, data visualization, geodemographics, search data

Reference to this paper should be made as follows: Wu, D., Xu, Z.N., Li, J. (2023). Search data and geodemographics determinants of the avocado sales in the US markets. *Journal of Business and Management*, 29(1), July, 23-56. DOI: 10.6347/JBM.202307_29(1).0002.

Introduction

From sales and marketing perspectives, data analytics and visualization tools can help businesses gain additional information and insights into building brands, making optimal pricing decisions, increasing market shares, and improving customer loyalty. This process focuses on defining metrics, collecting data, employing analytics tools, developing business strategies and executable plans, making decisions, and finally auditing and verifying the results (Wedel & Kannan, 2016), which refers to data-driven decision-making.

Traditional economic research assumes that industry-level analysis is important for making good decisions. However, an increasing number of marketing and information system scholars suggest that adding geodemographic data and non-traditional data like Google Trends (gtrends) search data may help managers make better decisions (Iacobucci et al., 2019; Wedel & Kannan, 2016). Specifically, demographic data is vital to understanding customer behavior and consumption patterns in a specific market. Additionally, geodemographic analysis and time series analysis may help businesses take advantage of the power of data for predictive analysis or prescriptive

analysis, which are often more complex by nature (Wedel & Kannan, 2016). Recent studies suggest that combining industrial-level data with search data and time series data may provide insights into the longitudinal data available for marketers, uncovering subtle nuanced patterns that augment traditionally adopted marketing analytics in the industry (Liu et al., 2021; Skenderi et al., 2021).

A common approach in mainstream time series analysis studies is to estimate how an endogenous variable responds to a change in exogenous variables or predictors using industrial-level data (Zu, Wang, & Cui, 2022). Such studies often underperform since selecting the best predictors is often challenging (Rodriguez, Ozkul, & Marks, 2018). The use of new explanatory variables like Google Trends related to forecasting and prediction activities has gained prominence in the literature. Google Trends is considered a gold mine for business research and is becoming increasingly heavily used (Dos Santos, 2018; Einav & Levin, 2014; Silva et al., 2019). Google Trends data that tracks people's search interests over time can be used to identify consumer demands and preferences, market threats from business competitors, seasonal patterns, business opportunities, and market risks. Previous studies have used Google Trends data to forecast influenzas, fashion trends, and oil consumption (Dos Santos, 2018; Einav & Levin, 2014; Silva et al., 2019; Yu et al., 2019). However, using data analytics and visualizations to model the sales of agriculture markets effectively remains challenging and difficult (Yoo & Oh, 2020). For instance, the techniques as mentioned above may often become inaccessible to practitioners due to various challenges, such as collecting relevant and reliable data and handling advanced analytic methods. As a result, businesses may make suboptimal decisions if they are unable to address these technical challenges, and industrial applications of these methods may not have reached their full potential. There is still in need for more studies, methods, and insights to give practitioners and academics guidance on how to gain business insights via data analytics that goes beyond traditional industrial-level or firm-level variables (Ravenscraft, 1983). Furthermore, finding meaningful data sets for practitioners and academics to conduct comprehensive studies is always challenging.

As such, much less is known regarding the impact of search trends and other industrial-level or category-level factors (pricing, product attributes, etc.) on agriculture product sales. The availability of avocado sales and pricing data in the form of time series and the increasing popularity of avocados in different geographic regions of the US will likely make avocados a good fit for this empirical study. This research selects the avocado market for the planned study to close the gaps between marketing practices and theories. For instance, avocados are popular fruits due to their taste and nutritional value. Avocado consumption is critical in food coupling and cooking, especially among millennials (HAB, 2019). Avocados sold in the US Market are primarily the Hass type (Ambrozek, Saitone, & Sexton, 2018). Hass avocados imported from Mexico account for over 90% of avocados sold in the US (Ambrozek, Saitone, & Sexton, 2018). Hass avocados

are branded as a type of "superfood" that have significant health and nutritional benefits (Ambrozek, Saitone, & Sexton, 2018).

In particular, the Hass Avocado Board (HAB) is an industry association responsible for promoting, researching, and managing Hass avocados. HAB has released weekly Hass avocado sales data across different regions in the US since 2014. The importance of Hass avocados is growing in the US due to the impacts of globalization and marketing efforts (Kourgialas & Dokou, 2021). Another reason this study chooses Hass avocado markets is that there is scant research on comprehensive marketing analytics for the avocado industry. This study can potentially help different market participants (importers, retailers, etc.) develop a data-driven marketing analytics roadmap to understand Hass avocado sales and provide insights into allocating marketing resources and making customer-related decisions. Although some research studies are looking at avocado sales in the selected regional markets, very few studies have examined avocado sales using geodemographics and Google Trends search data. Additionally, there is an urgent need to understand what factors affect avocado sales as well as what market patterns can be useful in making important decisions in avocado businesses, such as branding and pricing (Migliore et al., 2017; Ambrozek, Saitone, & Sexton, 2018; Ambrozek, Saitone, & Sexton, 2019).

Specifically, this study focuses on integrating Google Trends search data into analysis for examining pricing, sales, and revenue forecasting. Understanding regional and global markets and building reliable forecasting solutions are often challenging, preventing businesses from developing effective pricing and marketing strategies and making meaningful budgets and plans for future business successes. On the one hand, unsuccessful forecasting and marketing analysis often lacks sufficient and critical factors in the constructed models. On the other hand, models with too much complexity and redundant factors may limit their applications and even mislead users in understanding relationships between factors and results. At a high level, the progress of technology, the continuous generation of data flows through a fast-growing online community (such as Google Trends), and advancements in data analytics and algorithms have enabled us to construct more robust models with more rigor. Encouraged by recent development in data analytics and recognizing the challenges and opportunities at hand, we have developed several research questions. By addressing these questions, we demonstrate the criticality of incorporating Google Trends search data into principal component analysis (PCA) and time series analysis and offer the possibility of helping businesses explore more resilient methods of decision-making related to pricing, marketing, budgeting, and planning endeavors.

The research inquiries include but are not limited to, 1) explain potential correlations between geodemographic factors, such as education, age, ethnicity, income, and avocado sales. Geodemographic factors have proven helpful in understanding markets, which motivates us to explore how these factors can contribute to the

explanation of constructed regression models in avocado markets. The research inquiry aims to investigate the potential inter-correlations among geodemographic factors, identify any redundant factors, and assess the impact of these factors on regional markets; and 2) examine how Google Trends can further strengthen the explanatory power of regression models. In addition to geodemographic factors, internet search data is believed to play an important role in understanding regional markets. In general, internet search data has demonstrated success in providing real-time insight into consumer interests and purchasing intentions. A few recent studies have inspired us to explore how Google Trends data can improve the potential applications of our constructed models in this study. By incorporating geodemographic factors and internet search data into this study, we can better understand how to cluster regional markets, including geographically distant regions with potentially highly similar markets, and vice versa. To address these research inquiries, we employ a combination of k-means clustering, hierarchical clustering, regression analysis, and exploratory mapping to study a comprehensive dataset that includes industrial-level variables, census data, and search data.

This study contributes to three areas of research. First, data analytics and visualization tools were used to analyze avocado pricing and sales based on industrial-level, geodemographic, and Google Trends data, offering insights beyond traditional longitudinal analysis. Second, the study shows combining Google Trends search data with industrial-level data in a time series analysis can improve model performance. Third, this study may assist decision-makers in the avocado industry in making general demand and supply assumptions about the avocado market and in positioning avocados based on market-driven factors like Google Trends search data. Although this study is done in the US as well as the data is for the US markets, the research tools, models, conclusions, and insights in this study can be particularly useful to business researchers, professionals, and regulatory bodies all over the world as this study provides a comprehensive analysis and a typical example. Hence, similar applications can be undertaken for products in other industries and regions.

On the other hand, understanding local, regional, and national sales opportunities for foods and fruits in the US presents a significant challenge. When it comes to making critical decisions like warehousing, managing cross-border trading activities, and projecting demand and supply, international fruit or food importers and exporters may find it beneficial to integrate geodemographic data and search data into industrial-level data like prices and sales as this study advocates for.

Literature Review

As a starting point, we selected thirteen seminal articles to clarify the theoretical context and the subject area of avocado sales. We used geodemographic information and

Google Trends for forecasting. These articles were summarized in Table 1, and the results were used to help us identify the research gap in this study. Please understand the weaknesses column may not give a complete overview of the scenarios due to space limitations. A detailed discussion of these articles is available in this literature review section.

Geodemographics refers to using both demographic and geographic information of consumers and target markets. Using geodemographic information in marketing has shown great potential for making complex decisions. For instance, geodemographic information can be used to help marketers monitor trade areas and open new locations (Baviera-Puig, Buitrago-Vera, & Escriba-Perez, 2016; Formánek & Sokol, 2022; Rains & Longley, 2021). Geodemographic variables, such as location, age, education, gender, income, and ethnicity, have been used to understand status consumption (Eastman & Liu, 2012), grocery retailing (Prasad & Aryasri, 2011), cigarette brand loyalty and purchase patterns (Dawes, 2014), and customer loyalty in retail banking (Kamath, Pai, & Prabhu, 2019). Notably, a distinction is observable between previous periods (pre-2021), when Google Search data was not included, and recent periods (post-2021), when almost all studies incorporated this data type in their analyses. However, the more recent studies have not fully considered geodemographic and other relevant factors, highlighting a gap. Our study may potentially help fill this gap.

Table 1. Overview of the reviewed sources

Author/year	Sample/design	Data/tools	Results	Weakness
Studies that did not use Google Trends search data				
Palmon & Sopranzetti (2004)	6385 observations of real estate transactions of properties	Transaction data/cluster analysis	Using cluster analysis can help marketers dive deeper into the transaction price and the turnover of houses.	Some potential to use Google Trends data
Calantone & Di, Benedetto (2007)	215 recent new product launches	New product data/cluster analysis	Cluster analysis is performed to help pinpoint possible reasons that lead to successful launches of new products.	Some potential to use geodemographic data and Google search data
Bezawada & Pauwels (2013)	75 stores that span 355 weeks across 56 categories	Persistence modeling	Marketing organic foods and products can be special and different.	Some potential to use Google search data
Nghiem et al. (2016)	Google search data from 2004 to 2013	Google Search data/regression analysis	The quantity of news articles is related to patterns in Google search volume.	Some potential to use geodemographic data
Ambrozek, Saitone, & Sexton (2018)	Hass Avocado sales data from 2013 to 2017	Regression analysis and trend analysis	The activities of the Hass Avocado Board and its member associations have played a fundamental role in this success.	Some potential to use Google search data
Ambrozek, Saitone, & Sexton (2019)	Hass Avocado sales data from 2013 to 2018	Elasticity analysis and regression analysis	Demand for fresh Hass avocados is highly inelastic at the shipper and retail levels.	Some potential to use Google search data
Kirby-Hawkins, Birkin, and Clarke (2019)	814000 unique customer data	E-commerce sales/quadrant analysis	Geodemographic information shows spatial patterns in online grocery sales	Some potential to use Google search data

Chou et al. (2020)	977 consumers	Survey data/CFA	Green consumption intention was significantly and indirectly driven by attitude to green products.	Some potential to use geodemographic variables and Google search data
Studies that used Google Trends search data				
France, Shi, & Kazandjian (2021)	Search data for 100 brands over ten years	Google Search data/regression analysis	Google Trends creates several brand equity series for 100 top-ranked brands.	Some potential to use geodemographic data
Glass & Jarett (2021)	25 search terms	Google Search data/text analysis	Google Trends data offers zero moments of truth about online digital records or search queries over time.	Ignores the use of industrial-level data
Higuchi & Maehara (2021)	381 respondents' quinoa consumption data	Survey data/factor-cluster analysis	A factor-cluster analysis is performed to identify reasons for consuming quinoa.	Some potential to use geodemographic data and Google search data
James et al. (2021)	250,000 expenditure and attitudinal data	Expenditure data and Survey data/Segmentation analysis	Geodemographic information may help meat processors understand temporal and spatial patterns of meat consumption and sales.	Some potential to use demographic data

In the food and grocery industry, geodemographic information is particularly useful for distributors and retailers to develop effective marketing plans and strategies. Geodemographic information may provide insights into the expansion opportunities of retail companies (Thompson et al., 2012). Geodemographic information shows spatial patterns in online grocery sales (Kirby-Hawkins, Birkin, and Clarke, 2019). A recent study shows that applying geodemographic information may help meat processors understand temporal and spatial patterns of meat consumption and sales (James et al., 2021). Furthermore, combining supermarket sales data and geodemographic data can help marketers gain new insights into dietary purchase patterns, not only from marketing and sales perspectives but also from public health perspectives (Clark et al, 2021).

As people become more concerned with food allergies, pesticides, preservatives, or other health issues, organic foods and products have gained more popularity over the years (Vukasovič, 2016). Literature shows organic foods and products have different characteristics from their conventional counterparts (Rana & Paul, 2017). Therefore, marketing organic foods and products can be special and different. For instance, organic products may have higher demand elasticities (Bezawada & Pauwels, 2013), which implies that different promotion strategies may be required. Factors like geodemographic information may have a nuanced relationship with product categories. For instance, age, education level, and income affect people's preference for organic products in both developed and emerging economies (Chou et al., 2020; Vukasovič, 2016).

On the other hand, the consumption of avocados is considered very healthy and may reduce the risk of heart disease (Dreher & Davenport, 2013). Among all types of avocados, Hass avocados rank highly in quality, nutritional value, and taste. Hass avocados can only grow in certain regions and soils and are harvested less frequently than other avocados and fruits (Kourgialas & Dokou, 2021). Due to weather changes

(such as drought), trade barriers, and natural disasters, the supply of Hass avocados is highly volatile. As a result, all these reasons make Hass avocados very expensive fruits.

As the internet evolves from Web 1.0 to 3.0, businesses with the tools necessary to navigate the complexity of consumer trends and geodemographic information obtain a competitive advantage that may lead to better sales (Wedel & Kannan, 2016). Using mobile devices or information technology, consumers nowadays can access product and sales information through tools or platforms as either users or reviewers. As a result, digital records that may be relevant to business operations have grown exponentially. Specifically, Google Trends data offers zero moments of truth about online digital records or search queries over time for different regions that may be due to various marketing efforts (campaign effectiveness, promotional events, etc.) (Glass & Jarett, 2021). Zero moments of truth are a term coined by Google, which means the moment that occurs after the consumers perform online searches but before they make a real-world purchase behavior. As such, Google Trends resembles Zero moments of truth and is an aggregated index based on online search volume data in different regions over time. Google Trends data have been applied to analyze different phenomena for now casting and forecasting purposes (Dos Santos, 2018; France, Shi, & Kazandjian, 2021). In a recent write-up, Glass and Jarett (2021) show that Google Trends allows marketers to investigate the choice of region, time, and geolocation. Glass and Jarett (2021) also recommend that combining Google Trends with other actionable data may be a solution for unlocking new marketing opportunities.

Moreover, Nghiem et al. (2016) suggest that Google Trends may provide useful time-series information on public interest changes and shifts and is particularly useful for topics that span weekly or spatially. This paper is among the first to underscore heterogenous characteristics of consumer appetite for avocados across different regions in the US. The study may have implications for practitioners and scholars interested in other agricultural products.

Clustering and segmentation of markets and customers have been proven to be useful tools for detecting patterns and identify similarities by considering different types of information. Cluster analysis has successfully revealed hidden traits of marketing data for both descriptive and diagnostic purposes. For instance, cluster analysis can help marketers dive deeper into the transaction price and the turnover of houses on the market associated with list price and agency characteristics (Palmon & Sopranzetti, 2004). Cluster analysis can also help pinpoint possible reasons that lead to successful launches of new products (Calantone & Di Benedetto, 2007). A recent study found that factor-cluster analysis proves useful in identifying reasons for consuming quinoa (Higuchi & Maehara, 2021). Although more than 100 clustering algorithms are available, several major methods are popular and widely used. Hierarchical clustering can be used to identify dissimilarities or distances among observations and construct trees of clusters (Scitovski et al, 2021). The algorithm starts at a single cluster and split/join the cluster(s) iteratively

until certain conditions are met. Another popular approach is k-means clustering which was first used in the 1960s (Scitovski et al, 2021). Given a predefined k number of clusters, the initial assignments of items to these clusters will start, and then centroids will be calculated. The algorithm runs iteratively until the summed distances between clusters are minimized.

This paper represents one of the first to study the U.S. Hass avocado market with industrial-level data, census data, and Google Trends search data. As a native fruit in the Americas, avocados became very popular in the US According to the USDA 2018 report, per capita consumption of avocados in the US has tripled since 2001 to 8 pounds per person in 2018 (USDA, 2022). Most imported avocados are the Hass type. Previous research studies on Hass avocados in the metropolitan and regional markets found that not all of the markets are elastic and, over the years, the promotion programs funded by the Hass Avocado Board (HAB) have affected the consumption and demand of fresh Hass avocados in the US positively (Ambrozek, Saitone, & Sexton, 2018; Ambrozek, Saitone, & Sexton, 2019). However, no research has been conducted to understand 1) how different these main regional markets are, 2) what main factors may lead to these regional differences, and 3) the similarities and differences between these primary target markets. The study collects data from various sources, including Hass avocado sales data across different regions from 2015 to 2019, geodemographic data, and Google Trends search data on ”avocado" from 2015 to 2019.

The purpose of this study is three-fold. Firstly, the study adopts cluster analysis and visualizations to uncover the differences and similarities between different regional markets. Secondly, a robust regression analysis is performed to understand different factors that may contribute to these regional differences in avocado consumption. Thirdly, the study aims to help different market participants (marketers, wholesalers, distributors, retailers, etc.) develop a data-driven marketing analytics roadmap to understand Hass avocado sales. This paper will contribute not only to the literature on marketing research in avocado or agriculture products, but also to the literature on marketing analytics.

Research Methodology and Data

In this research study, avocado markets are of primary concern. Avocados have become increasingly popular over the years due to their taste and nutritional value. Most avocados sold in the US market are primarily the Hass type (Ambrozek, Saitone, & Sexton, 2018). According to Hass Avocado Board reports (HAB, 2019), the annual consumption of Hass avocados per capita has increased steadily. HAB has released weekly Hass avocado sales data across different regions in the US since 2014. The importance of Hass avocados is growing in the US due to the impacts of globalization and marketing efforts (Kourgialas & Dokou, 2021).

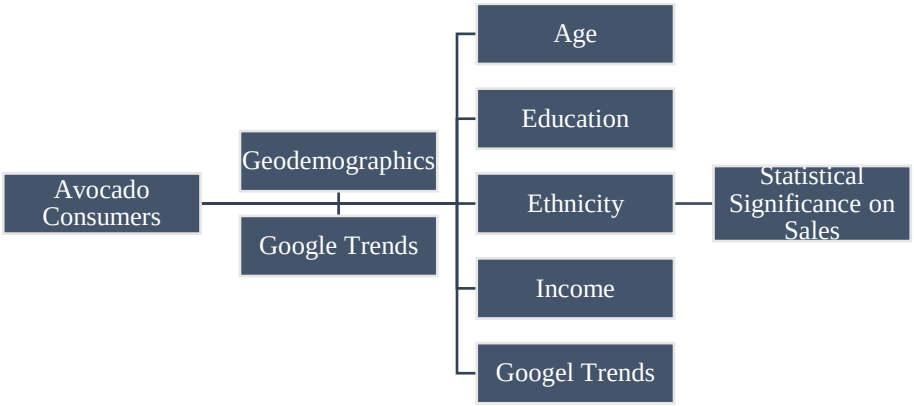
There are a few demographic factors that may affect the consumption of avocados. First, a high level of education may affect people's consumption of healthy food, such as avocados (Li & Powdthavee, 2015). Second, market reports also show that avocados are highly popular among young people (Lufkin, 2019). Third, as avocados originated in Latin America, it is not surprising that many studies support that Hispanic populations outpace other ethnic groups in consuming avocados. Fourth, the consumption of Hass avocados may be affected by income level as Hass avocados are relatively expensive fruits. Fifth, people have become highly aware of and concerned with antibiotics, pesticides, and preservatives, and organic Hass avocados have gained great popularity in the market (Chou et al., 2020). However, organic Hass avocados are usually more expensive than conventional ones. Finally, seasonality plays an important role in agricultural products; avocado is not an exception. Agriculture reports recommend that the avocado season usually starts from April to September, causing prices of avocados to be relatively lower in summer (The Produce News, 2021). Compared to the previous search studies (Ambrozek, Saitone, & Sexton, 2018, 2019), this study considers a broader set of geodemographic factors except population and focuses on more exploratory analysis, one of which studied how promotions affected avocado consumptions and sales. Therefore, the application of these studies may be limited, and the study may have limited practical applications and benefits for practitioners. Therefore, this research study is among the first to consider a broad set of geodemographic factors in the developed models.

As the internet community continues to grow, Google search, as one of the effective tools for revealing the zero moments of truth, shows a good track record in adding more predictive power to existing regression models (France, Shi, & Kazandjian, 2021). As a result of Google searches performed by millions of consumers across different regions, it makes sense to extract aggregated search insights from different regions on any topic, such as marketing, sales, health issues, and technology (France, Shi, & Kazandjian, 2021). A Google Trends dataset collected using certain search queries may indicate the popularity of these topics in the internet community. From sales and marketing perspectives, highly frequent searches may be correlated with more product sales. Following previous studies (France, Shi, & Kazandjian, 2021, Nghiem et al., 2016), Google Trends search interest data is included in the time series model as it may be a good indicator for the consumption of avocados across different regions. To the best of our knowledge, in studying possible factors affecting consumers' interests in avocado consumption, no paper has combined industry-level factors (organic, seasonal issues, etc.), geodemographic factors (education, income, ethnicity, age, etc.), and Google Trends search data that serve as a proxy of public interest in avocados in different regions (France, Shi, & Kazandjian, 2021).

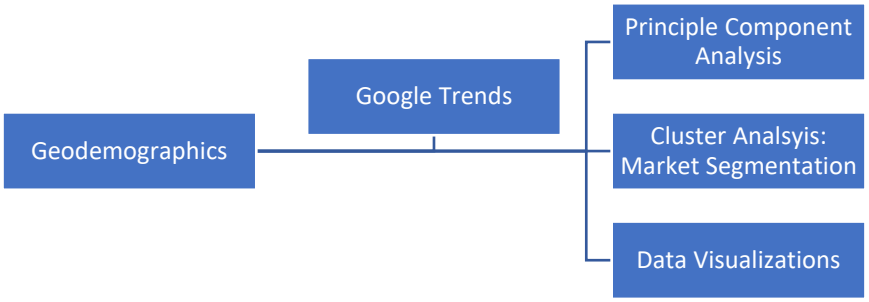
To help readers better understand our methods and procedures, a process that describes our research in these details is offered in Figure 1. The data used in this research are collected from several different sources. Weekly organic and conventional Hass avocados sales data from January 2015 to December 2019 for 27 major metropolitan regions are collected from the Hass Avocado Board. Please note that this dataset is from

2015 to 2019 due to the COVID-19 disruption in the 2020 US Census data publication and hence the lack of geodemographic data in and after 2020. The sales data and geodemographic data after 2020, when available, may provide additional insight into the avocado markets during COVID-19, which will be studied in the future. The geodemographic variables, including annual income, median age, population, education, and ethnicity, are collected from the US Census Bureau and the US Bureau of Labor Statistics. Weekly Google Trends data for these metropolitan markets are extracted from the Google Trends website by using the gtrends R package available in R. Since most people would use "avocado" instead of "Hass avocado" for Google search, it makes sense to use "avocado" as a generic search term to capture the popularity of "Hass avocado" searches. It is possible to include other languages in the search term. However, since Google Trends is a relative index, the popularity of avocados in a different language may not be directly comparable to that of avocados in English. A future study may be developed to offer insights into this issue. Data visualization and exploratory analysis are conducted using PCA, clustering, and several graphics algorithms available in several R packages. The detailed research process and proposed concept are summarized in Figure 1.

(1a) Data preparation



(1b) Regression model analysis and related research questions



(1c) Exploration analysis and data visualizations

Figure 1. Research process

This research study features panel data from the different metropolitan markets and weekly sales information from January 2015 to December 2019. The regression model is hypothesized and constructed below:

$$WeeklySalesPerCapita = \beta_1 Education + \beta_2 Ethnicity + \beta_3 Income + \beta_4 MedianAge + \beta_5 GoogleTrends + \alpha_{Type} Type + \alpha_{Month} Month + \alpha_{Year} Year + \varepsilon$$

The model description is as follows:

$WeeklySalesPerCapita$ is calculated by multiplying weekly average sales price by total sales volume of Hass avocados and then converted to sales per capita. It is the dependent variable;

$Education$ is the percentage of the population with a college degree or higher;

$Ethnicity$ gives the percentage of the Hispanic heritage population in each market;

$Income$ shows the average annual income in the region;

$MedianAge$ provides the median age in the region;

$GoogleTrends$ offers the weekly search popularity for the query "avocados" in the Google search engine for each region. Nghiem et al. (2016) suggest that Google Trends may contain time-series information on public interest shifts and is particularly useful for topics that span weekly or spatially;

$Type$ is a fixed effect variable for the type (organic or conventional) of Hass avocados sold on the market. This fixed effect variable distinguishes between these two types of avocados;

$Month$ is a fixed effect variable that handles seasonal variations that takes the values from January to December of each year. As such, the model takes one if the observation is drawn from that month and 0 otherwise;

$Year$ is a fixed effect variable that explains year-over-year variations that take the values from 2015 to 2019;

ε is the error term that assumes to follow the normal distribution.

A Hausman test (Dougherty, 2011) is applied to demine if fixed-effects or random-effects methods should be used (Dougherty, 2011; Hausman 1978) to model the panel data in the study. The Hausman test was performed using the PLM package available in

the R environment. In the Hausman test, the null hypothesis is developed as no systematic difference in fixed-effects coefficients and random-effects coefficients. The results of the Hausman test for both models show that the p-value is less than <.001 and is significant, suggesting that the fixed-effects model is preferred over the random-effects model (Hausman 1978). Therefore, this study will use the fixed-effects panel regression model as it controls for the time-invariant variables. The fixed-effects model is also essential to adjust the errors for potential heteroscedasticity. Table 2 summarizes the Hausman test, supporting fixed effects model.

Table 2. Hausman test for fixed effects, random effects, and pooled ordinary least squares (OLS) models

Step	Fixed Effects Output in R package	Random Effects Output in R Package
Step 1. Regression	Total Sum of Squares: 1039500 Residual Sum of Squares: 230560 R-Squared: 0.77821 Adj. R-Squared: 0.77717 F-statistic: 8045.8 on 6 and 13758 DF, p-value: < 2.22e-16	Total Sum of Squares: 1041800 Residual Sum of Squares: 231720 R-Squared: 0.77757 Adj. R-Squared: 0.77748 Chisq: 48287.5 on 6 DF, p-value: < 2.22e-16
Step 2. Hausman test	H_0: the preferred model is random effects H_1: the preferred model is fixed effects Conclusion: the result is significant to reject H_0, so the preferred model is fixed effects	Output data: psales1 ~ ethnicity + education + income1 + medianage + gtrends + ... chisq = 42.721, df = 6, p-value = 1.325e-07 alternative hypothesis: one model is inconsistent
Step 3. Random Effects VS Pooled OLS	This step is only conducted when a fixed effect model is not chosen. Since the results support the fixed effects model, this step is not necessary.	

Results and Discussions

The data analysis is carried out in the R environment except for mapping avocado sales per capita. The research process outlined in the previous section will start with a regression analysis of research questions and hypotheses, followed by exploratory analysis and data visualization.

Regression Analysis

A subsequent regression analysis is performed to analyze the impact of these factors on avocado sales per capita. Regression analysis shown in Table 3 indicates that multiple factors influence per capita sales of Hass avocados, as previously proposed and hypothesized. For instance, education level is statistically significant and is positively associated with per capita sales of Hass avocados. This suggests that other things being equal, regions with higher education levels tend to have better sales of avocados.

Moreover, the results show that the coefficients for ethnicity, income, and Google Trends search interest all appear positive and significant, while the coefficient for median age is negative and significant. Consumers with Hispanic ethnicity are inclined to consume more Hass avocados, which is consistent with the fact that, historically, avocado is part of their traditional foods. Income also plays an important role in predicting per capita sales of Hass avocados. Consumers with higher income would tend to purchase more avocados. Median age is an important proxy for future growth opportunities. In a recent study by HAB (2019), avocados are more popular among younger generations (Gen Z and millennials).

Google Trends search interest is a proxy or indicator for changes in public interest in different topics (France, Shi, & Kazandjian, 2021). In recent years, due to different promotions and marketing campaigns launched by health professionals on the rich nutrition of avocados, it has become trendy to eat avocados (Kerr, 2018). As such, Google Trends search interest is proven to be a statistically significant and positive predictor (i.e., coefficient value is .07) for per capita sales of Hass avocados. Table 3 also shows that adding Google Trends as a predictor improves the model's predictability (i.e., adjusted R squared value) from 0.775 to 0.782.

Table 3. Regression analysis of avocado sales per capita

Variables	Panel A: Without Google Trends				Panel B: With Google Trends			
	Coefficients	t value	P value		Coefficients	t value	P value	
(Intercept)	14.897	19.014	< 2e-16	***	10.339	12.947	< 2e-16	***
education	19.838	20.060	< 2e-16	***	21.387	21.931	< 2e-16	***
ethnicity	9.171	29.788	< 2e-16	***	10.517	34.030	< 2e-16	***
income	0.004	0.963	0.335	NS	0.022	4.793	0.000	***
median age	-0.246	-12.485	< 2e-16	***	-0.188	-9.607	< 2e-16	***
gtrends	NA	NA	NA	NA	0.074	21.731	< 2e-16	***
factor(organic)	-14.933	-211.584	< 2e-16	***	-14.933	-215.139	< 2e-16	***
factor (month) 2	0.238	1.374	0.169	NS	0.294	1.724	0.085	#
factor (month) 3	0.234	1.383	0.167	NS	0.268	1.612	0.107	NS
factor (month) 4	0.734	4.342	0.000	***	0.744	4.477	0.000	***
factor (month) 5	1.375	8.130	0.000	***	1.330	7.995	0.000	***
factor (month) 6	1.292	7.545	0.000	***	1.314	7.803	0.000	***
factor (month) 7	1.203	7.195	0.000	***	1.242	7.552	0.000	***
factor (month) 8	0.860	5.025	0.000	***	0.969	5.755	0.000	***
factor (month) 9	0.564	3.335	0.001	***	0.798	4.787	0.000	***
factor (month) 10	0.025	0.147	0.883	NS	0.303	1.818	0.069	#
factor (month) 11	-0.788	-4.600	0.000	***	-0.491	-2.908	0.004	**
factor (month) 12	-0.938	-5.319	0.000	***	-0.700	-4.025	0.000	***
factor (year) 2016	0.509	4.569	0.000	***	0.401	3.656	0.000	***

factor (year) 2017	1.590	14.235	< 2e-16	***	1.397	12.676	< 2e-16	***
factor (year) 2018	2.112	18.272	< 2e-16	***	1.712	14.866	< 2e-16	***
factor (year) 2019	2.501	21.657	< 2e-16	***	2.052	17.779	< 2e-16	***

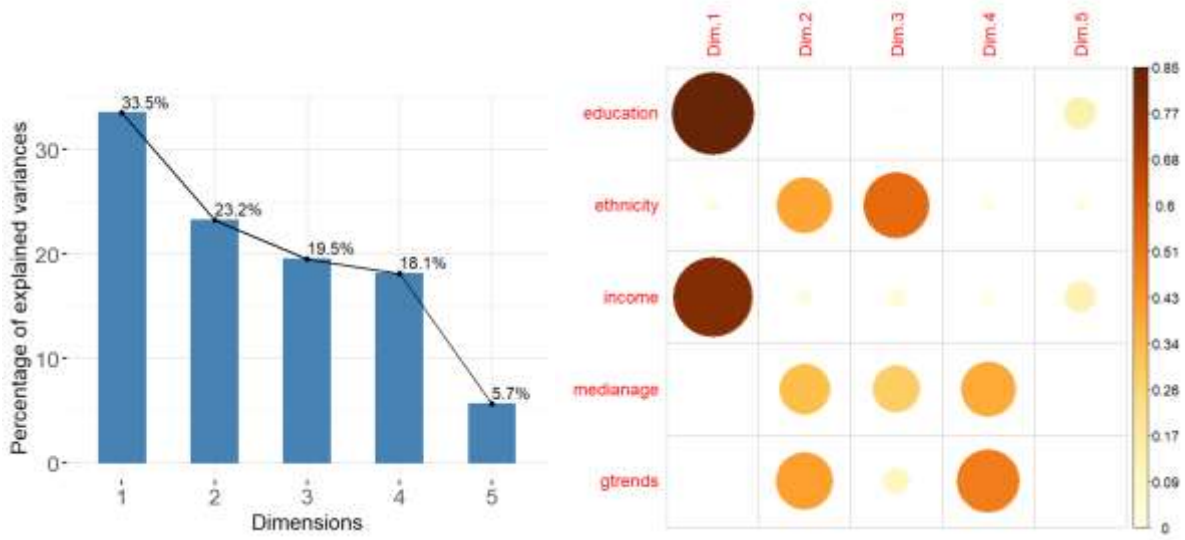
Notes: Adjusted R-squared: 0.775 for Panel A - Without Google Trends, and 0.782 for Panel B -With Google Trends. Significance levels: ***p<0.001; **p<0.01; *p<0.05; #p<0.10; NS=Not significant.

Li and Powdthavee (2015) found that more education is associated with health behaviors in Australia. In general, most previous studies focus on analyzing the causal relationship of education on risky behaviors like smoking, alcohol consumption, overeating, and drug abuse (Cutler & Lleras-Muney, 2010). Analyzing the causal relationship of education on the consumption of super-healthy fruits like avocados may provide ideas to address these risky behaviors. A coefficient of 2.14E-01 (or .214 in real number) means that an increase of one percent in the population with at least a college degree will lead to a surge of .214 dollars in avocado sales per capita. Interestingly, the positive effect of ethnicity on avocado sales is only 1.05E-01 (or 0.105 in real number) and is relatively smaller. In general, an increase of one percent in the Hispanic population tends to lead to a modest gain in avocado sales per capita at $.105.

Several fixed effect factors like types of avocados, year, and month are controlled in the regression model. The coefficient for type is negative and significant, which suggests that, compared to conventional Hass avocados, organic Hass avocados are less popular due to their high price. The results show the existing and significant seasonal effects. In particular, consumption of Hass avocados per capita starts to pick up from February, reaches its peak in summer, and slows down in winter, which coincides with the months when avocados are in season.

Principle component analysis (PCA) and data visualization

The model has five independent variables, including education level, ethnicity, income level, the median age of the population, and the Google Trends search interest for each region. PCA is critical when determining the importance of predictors before pursuing a standard regression analysis (Rodriguez, Ozkul, & Marks, 2018). Using PCA to study the variations of these five factors can reduce the complexity of the large dataset while summarizing the information efficiently in reduced dimensions (Rodriguez, Ozkul, & Marks, 2018). Table 4 summarizes the eigenvalues and cumulative variance percent of PCA analysis for these five factors. Table 4 and Figure 2a shows that the first four dimensions can explain about 94.35% of data variation. The results suggest that each one of these factors is critical, and they all have the potential to make better predictions. Please note that subspace dimensions calculated in PCA analysis generally do not have a specific meaning but provide insights into studying inter-relationships between model variables and reducing complex models to effective and simplified ones. Researchers can benefit from PCA analysis and gain more insights into the developed models.



(2a) (2b)

Figure 2. Histogram of cumulative variance percent (2a) and projection map (2b) of each factor

Figure 2b shows how variation changes of these factors are related, as well as the scale or magnitude of the projection of each factor on each PCA dimension. In the projection map of Figure 2b, education and income seem to be highly similar in the projections on PCA dimensions 1 and 5. According to social economics research, education and income are highly correlated. For instance, according to a recent article (Dickler, 2021), a person with at least a bachelor's degree may earn a median of $2.8 million in their lifetime, about 75% more than their counterparts with only high school diplomas. The PCA results suggest that these two factors may have similar effects on the consumption of Hass avocados in a regression analysis.

Moreover, median age (median age) and Google Trends (gtrends) may correlate as they are primarily expressed in the PCA dimensions 2, 3, and 4. However, Google Trends search data may be more related to young consumers and hence may be heading in the opposite direction of what the median age factor does. In a recent HAB (2019) study, avocados seem to be more popular among younger generations (e.g., Gen Z and Millennials). Interestingly, ethnicity and median age (median age) are correlated as they are represented in the PCA dimensions 2 and 3. However, they may affect Hass avocado consumption in the opposite direction as it is plausible that the high percentage of the Hispanic population, the greater the consumption of Hass avocados per capita in the region. According to the US Census Bureau (2021), the Hispanic ethnic group is the youngest in the US. Therefore, the higher the percentage of the Hispanic population in the region, the lower the median age. Lastly, education and income seem linearly independent from ethnicity, income, and median age (see Figure 2b and Figure 3).

Table 4. Principal component analysis

Dimension	Eigenvalue	Variance Percent	Cumulative Variance Percent
Dim.1	1.674909	33.49817	33.49817
Dim.2	1.162082	23.24165	56.73982
Dim.3	0.974451	19.48902	76.22884
Dim.4	0.906021	18.12042	94.34926
Dim.5	0.282537	5.650743	100

Furthermore, Figure 3 shows that when these five factors are projected on two dimensions, income and education move along the horizontal axis in the right or positive direction. Moreover, ethnicity and Google Trends search interest move along the vertical axis in the up or positive direction, with the median age being placed in the opposite direction of the vertical axis.

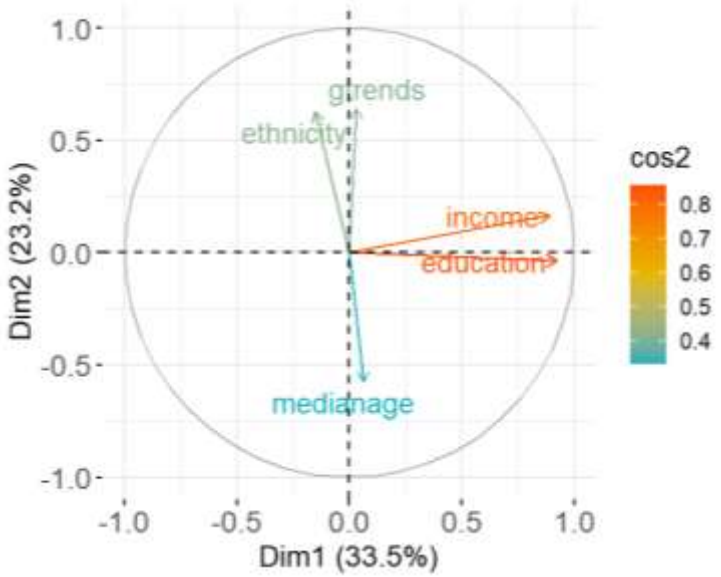


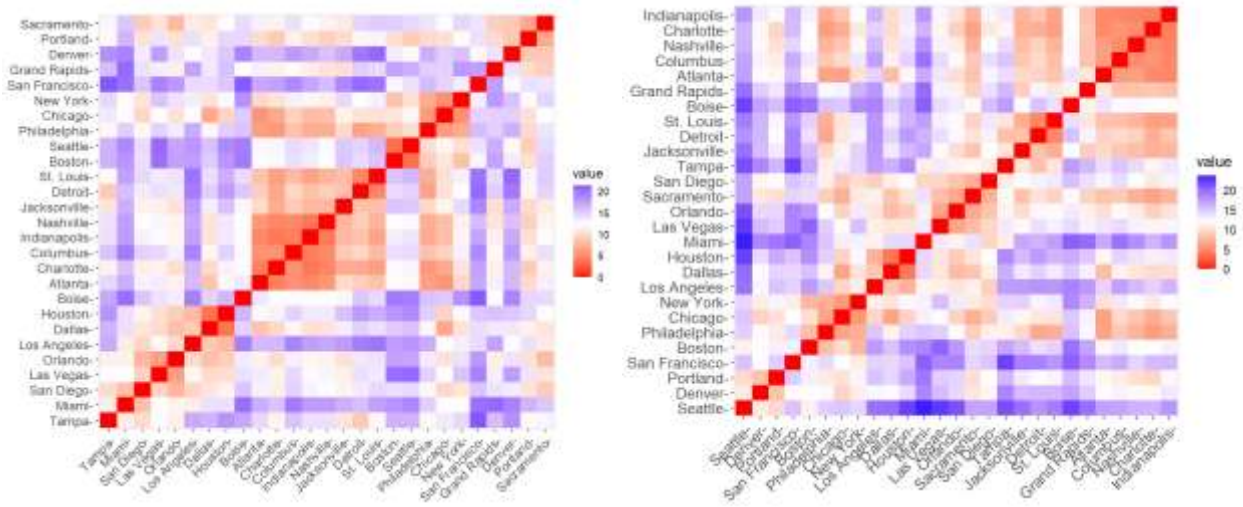
Figure 3. A 2D projection of factors on PCA dimensions

Clustering and exploratory analysis using maps

Cluster analysis is performed to find patterns and commonalities between the 27 regional markets used in this study. This analysis uses all the market information, such as geodemographics and Google Trends search data. The results may suggest that these regional markets can be aggregated into a few groups. In each group, these regions are generally similar in many ways but not necessarily geographically close.

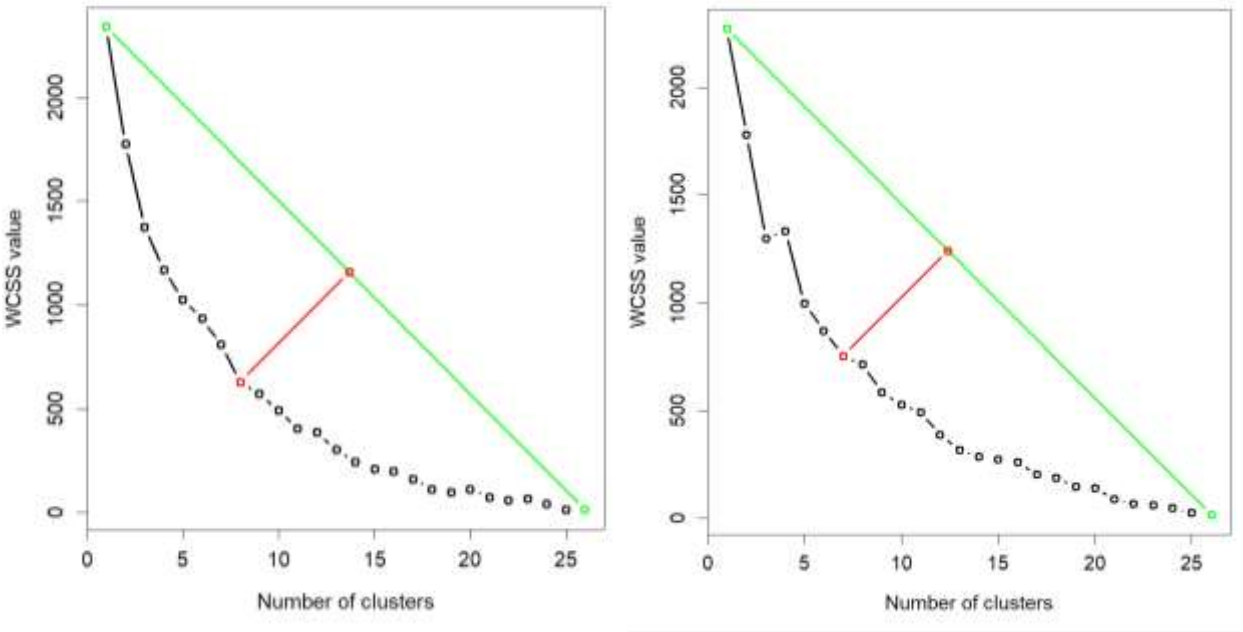
The first analysis presents the distance matrix between regions. In this analysis, the regional multi-dimension data, including geodemographics and Google Trends data,

are normalized, and then the Euclidean distance is calculated between pairwise markets. As shown in Figure 4a, similar markets have a short physical distance for the conventional Hass avocado markets, for example, Charlotte and Atlanta; for the organic Hass avocado markets in Figure 4b, New York and Chicago have similar markets. However, markets that are not geographically close may belong to the same cluster, such as Boston and Seattle in the distance map, as shown in Figure 4a, and Los Angeles and Dallas in the distance map, as shown in Figure 4b.



(4a) Conventional (4b) Organic

Figure 4. Distance matrix of regional markets



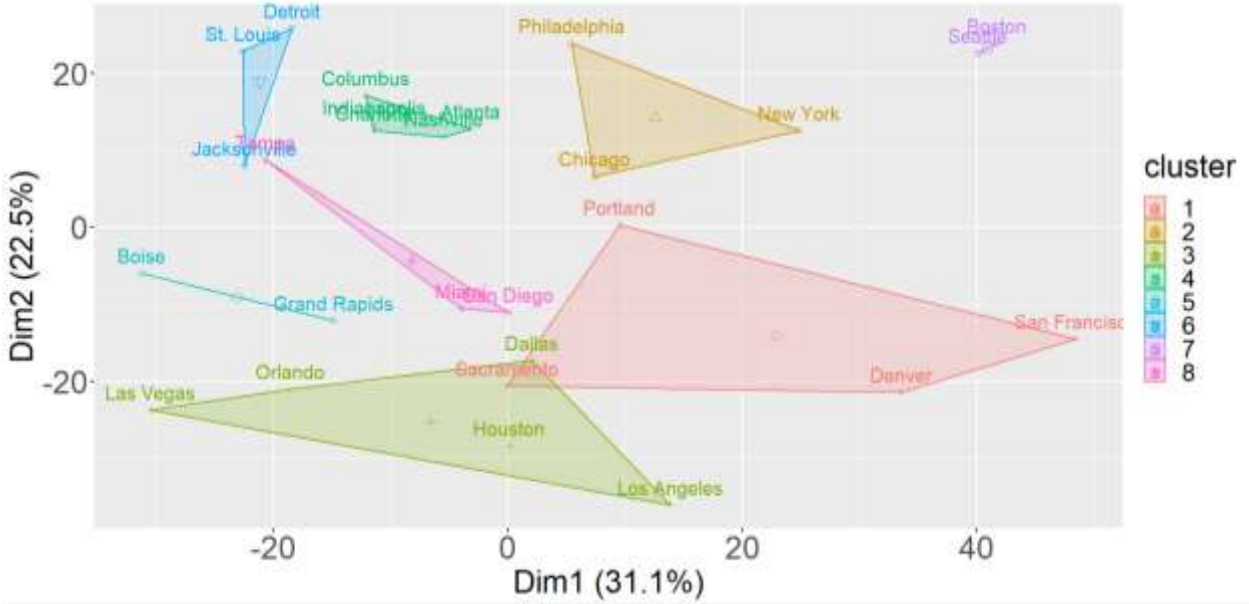
(5a) Conventional (5b) Organic

Figure 5. Elbow method for determining an optimal number of clusters

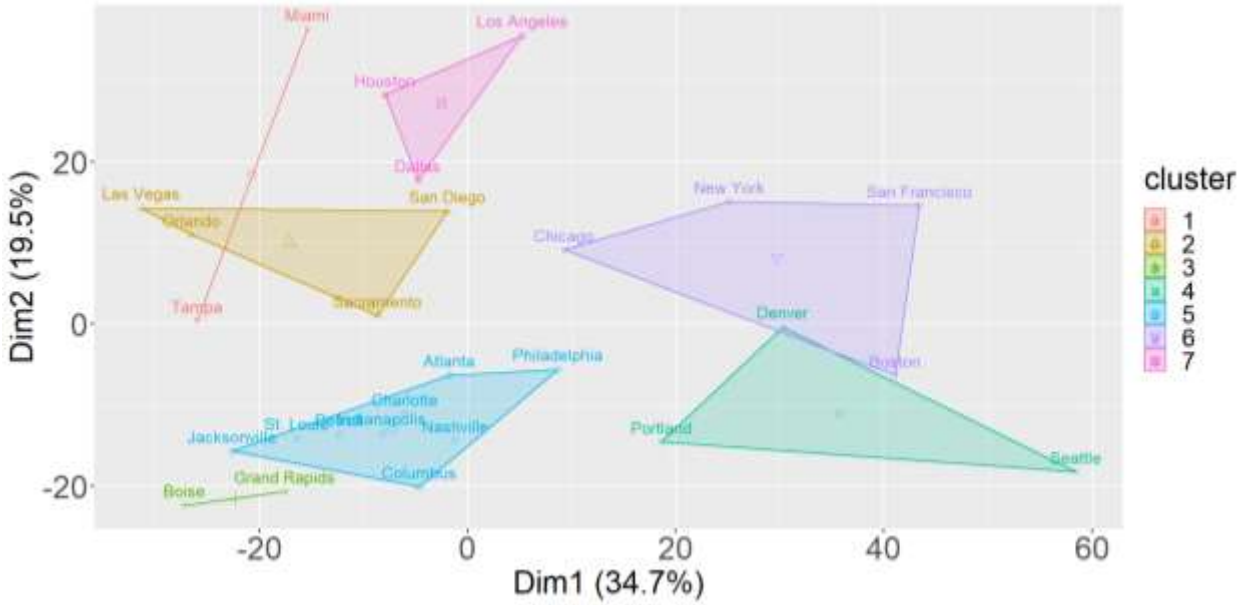
Two popular cluster analysis approaches, k-means clustering and hierarchical clustering methods, are adopted as they are widely used in data analytics and represent two major families of clustering methods (James et al, 2013). K-means clustering is centroid-based, in which the number of clusters is k, and each cluster is represented by a central vector (may not be an observation). K-means clustering minimizes the total squared distances from objects to the assigned cluster centers by finding the optimal centers and grouping the objects to the closest center (Hartigan &Wong, 1979; James et al., 2013). On the other hand, the hierarchical clustering method is connectivity-based, in which the objects are connected to form clusters based on distances. Hierarchical clustering does not partition the data set as k-means clustering does, instead using a dendrogram to represent merged clusters within certain distances (see Figure 7) (James et al., 2013). Considering the difference between organic and conventional Hass avocado markets, the analysis of each type of Hass avocado is studied separately. In the cluster analysis, selecting appropriate numbers of clusters these regions may be grouped into is important. The Elbow method efficiently and effectively determines the number of clusters in the analysis (Hardy, 1994). The idea behind this method is modeling and graphing the total within-cluster sum-of-squares (WCSS) based on the number of clusters. The optimal number of clusters is selected when the slope of the WCSS is flatter, which indicates adding more clusters will not add much to the variance, and the variance is then optimized (Hardy, 1994). In this study, employing the Elbow method suggests that for the conventional Hass avocado markets, an 8-cluster solution is optimal (see Figure 5a). In contrast, for the organic Hass avocado markets, a 7-cluster solution works better (see Figure 5b).

Applying k-means clustering to the conventional and organic Hass avocado markets allows us to see how these regional markets are grouped into different market segments. As summarized in Table 5 and Figure 6, regional markets that are geographically close are not necessarily in the same clusters and vice versa; clustering results are quite different between conventional and organic Hass avocado markets. For example, the San Diego market may be similar to the Tampa and Miami markets for conventional avocados; however, San Diego is not in the same cluster as Tampa and Miami for organic avocado markets. As a result, this indicates that when developing marketing and sales plans for conventional avocados, whatever plans or forecasting work for the San Diego market might also work for the Tampa and Miami markets, but it may not be the case for organic Hass avocados. One could argue that it is unsurprising to cluster the San Diego, Tampa, and Miami markets in one group as they all have high living expenses and are located in the south with large Hispanic populations. Another interesting observation is that for the conventional Hass avocado consumption, San Francisco, Denver, Portland, and Sacramento are in the same group, located in the west region. Still, the San Francisco market is grouped into the same cluster as the Boston, Chicago, and New York markets for organic Hass avocado consumption. This may indicate that as Hass organic avocados are more expensive and income may influence

more organic Hass avocados markets like Boston, Chicago, New York, and San Francisco, which are well known for their higher than usual living expenses and income level; for the conventional Hass avocado consumption, demographics may play a role as Denver, Portland, Sacramento, and San Francisco have similar levels of demographics and are all located in the west.

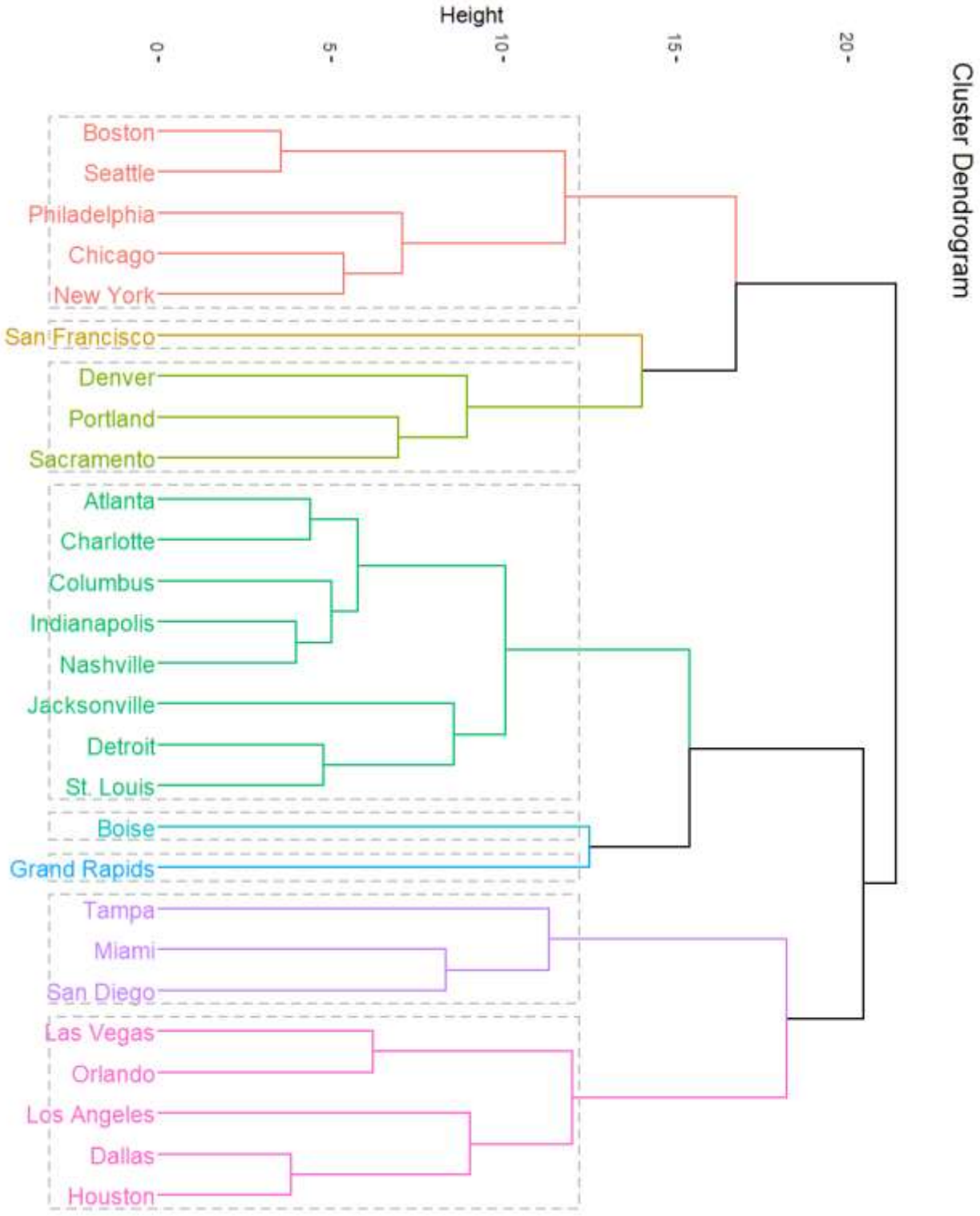


(6a) Conventional

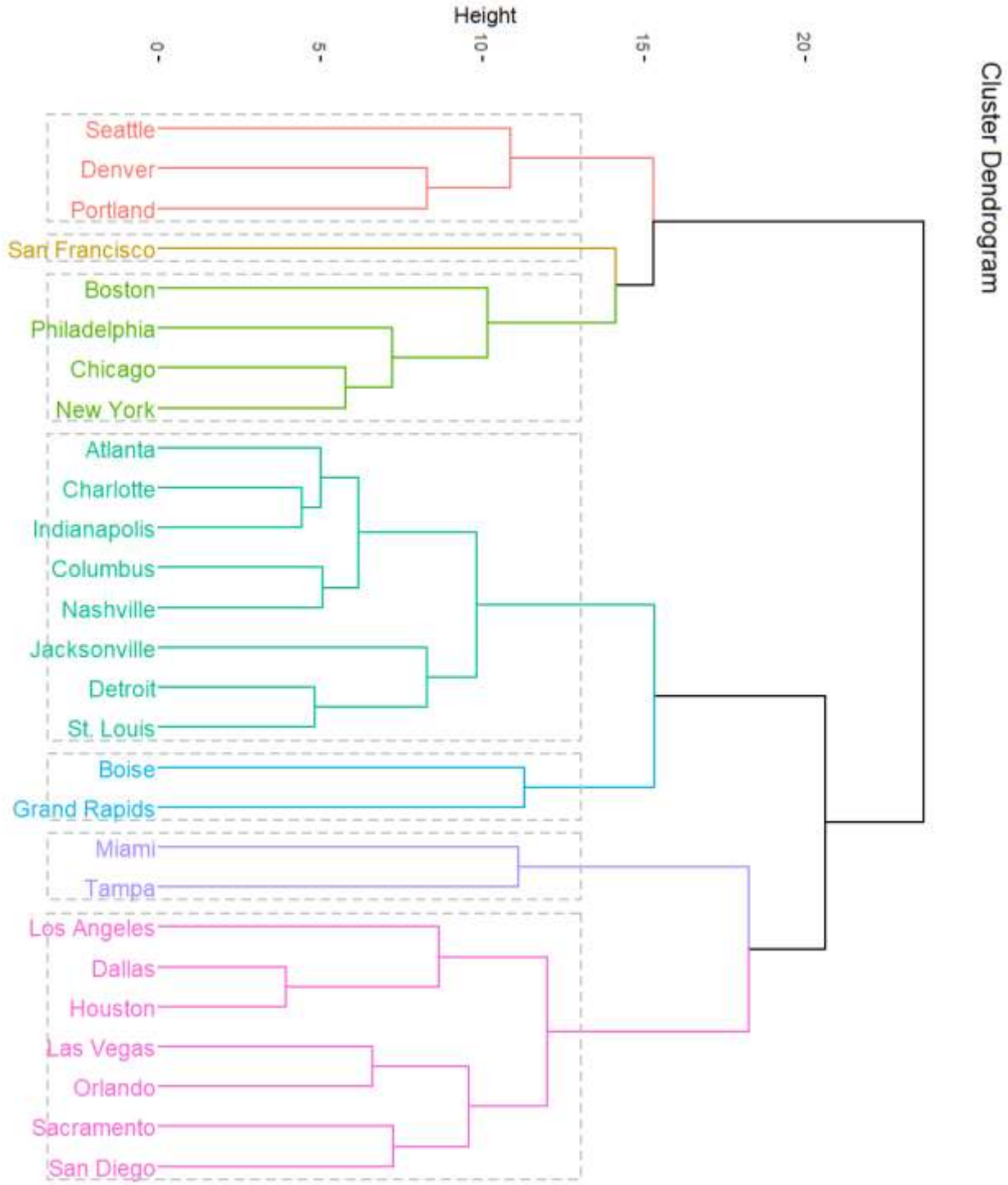


(6b) Organic

Figure 6. K-means clustering results projected on primary two dimensions



(7a) Conventional



(7b) Organic

Figure 7. Visualization of hierarchical clustering results

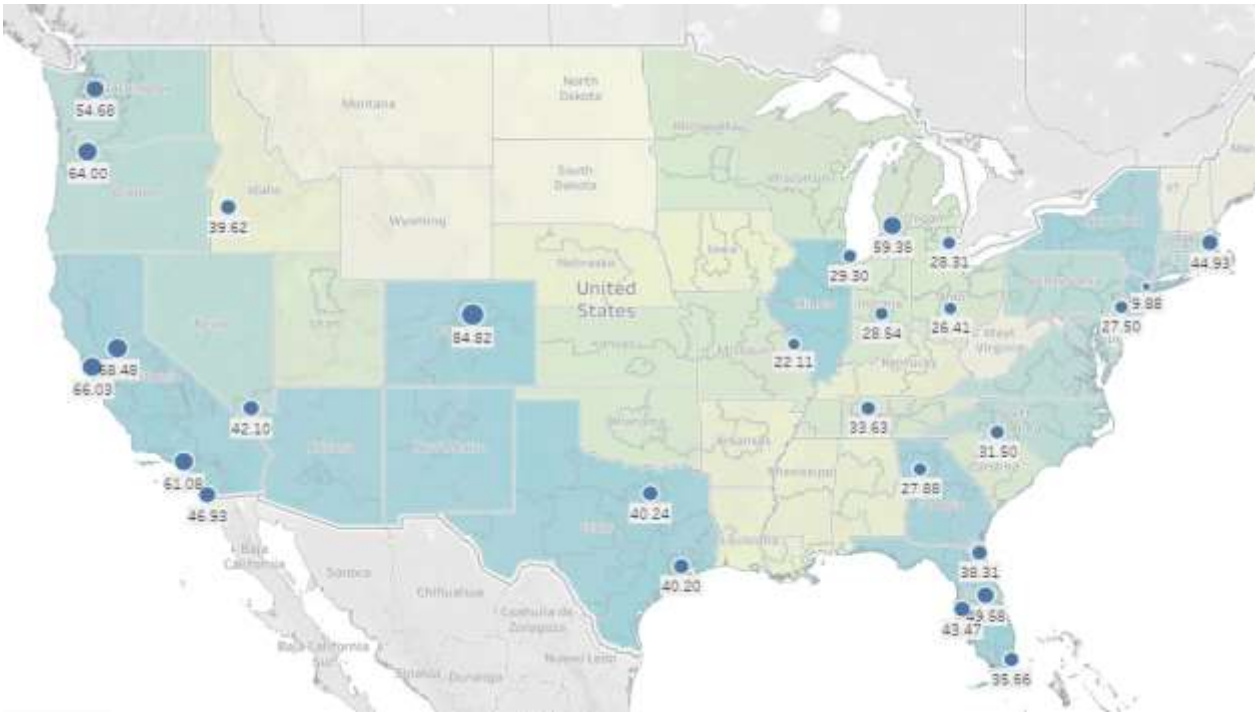
Hierarchical cluster analysis is also applied to analyze these regional markets from a different perspective. Compared with the k-means method, the results are similar, as

shown in Figures 7a-7b. The results indicate that although these two approaches use different algorithms and focus on different objectives, both methods give similar results in classifying these regional markets. Cluster analysis can help marketers discover critical geodemographic patterns between different regional markets. The results can help businesses develop effective and efficient plans to gain additional market share, be more strategic, cost-efficient, and agile, and become risk intelligent.

Table 5. K-means cluster analysis

Clusters	Conventional	Organic
1	Denver, Portland, Sacramento, San Francisco	Miami, Tampa
2	Chicago, New York, Philadelphia	Las Vegas, Orlando, Sacramento, San Diego
3	Dallas, Houston, Las Vegas, Los Angeles, Orlando	Boise, Grand Rapids
4	Atlanta, Charlotte, Columbus, Indianapolis, Nashville	Denver, Portland, Seattle
5	Boise, Grand Rapids	Atlanta, Charlotte, Columbus, Detroit, Indianapolis, Jacksonville, Nashville, Philadelphia, St. Louis
6	Detroit, Jacksonville, St. Louis	Boston, Chicago, New York, San Francisco
7	Boston, Seattle	Dallas, Houston, Los Angeles
8	Miami, San Diego, Tampa	

Mapping visualization is a good way to demonstrate how different regions behave regarding avocado sales per capita. Figure 8 shows sales per capita in each selected region. This visualization shows that Denver, Colorado, has the highest Hass avocado sales per capita, followed by Sacramento, California. In addition, cities in the western region have a higher range of sales per capita. For conventional Hass avocados, the same pattern can be observed. However, for organic Hass avocados, Seattle, Washington, has the highest per capita sale of 5.9, followed by Portland, Oregon, with per capita sales 4.57. The cluster analysis and the sales map suggest that businesses can gain insights into the similarities and differences of regional markets using exploratory analysis.

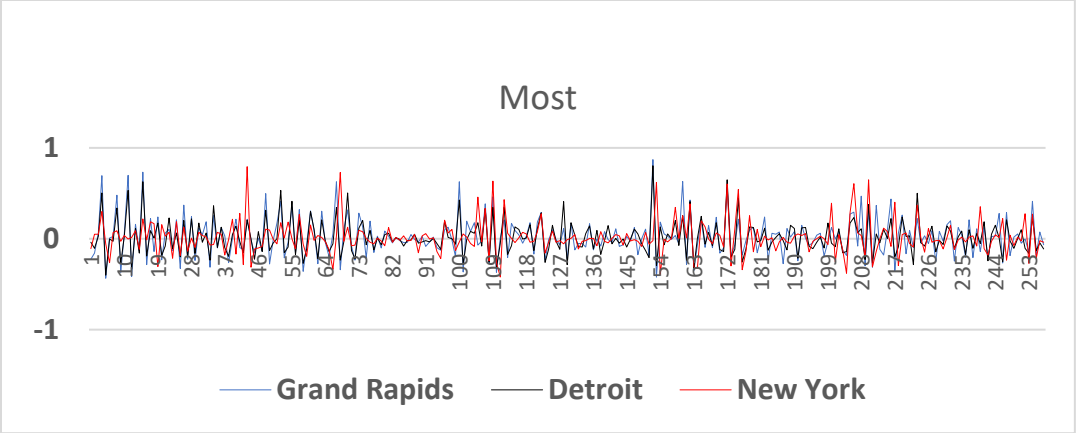


Legend: Blue regions indicate areas where Hispanics and Latino communities are highly concentrated.

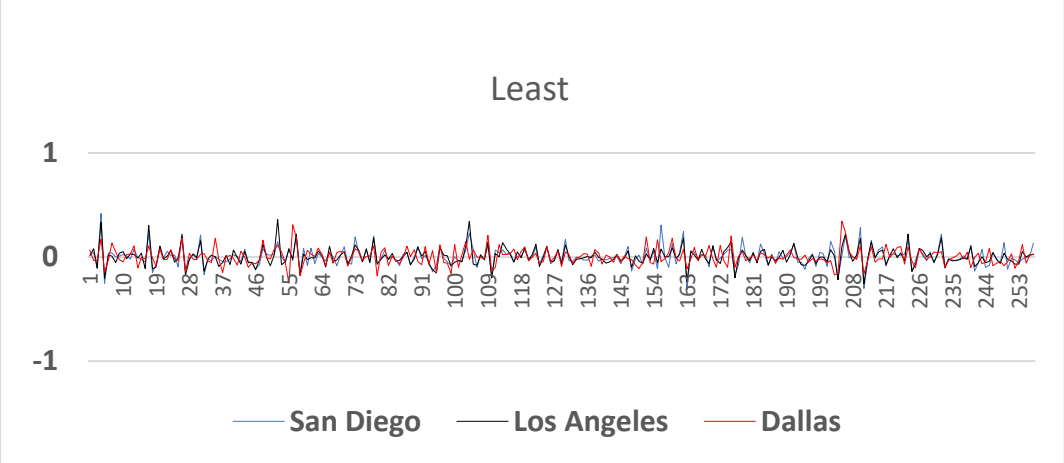
Figure 8. Avocado sales per capita across different regions

Sales Per Capita Change Volatility and Data Visualization

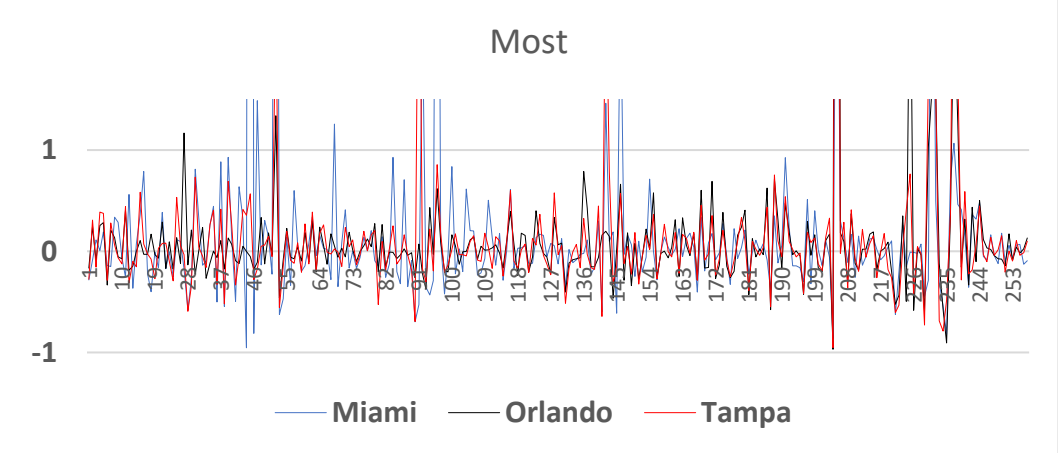
A series of figures are further explored to reveal the volatilities of avocado sales per capita (Figures 9a-9d). In general, there is some level of volatility in avocado sales per capita. For conventional Hass avocados, the following three cities, Grand Rapids, Detroit, and New York, experience the largest yearly volatility, as shown in Figure 9. These cities are in the northern area of the US, and they experience temperature and weather changes; thus Hass avocado supply is impacted accordingly. As such, avocado prices may also be affected and may experience the sales volatility per capita. Interestingly, cities that experience the least volatility change of sales per capita are all located in temperate climates of the US, like San Diego, Los Angeles, and Dallas (Figure 9b). These cities are also relatively closer to the places of product origin.



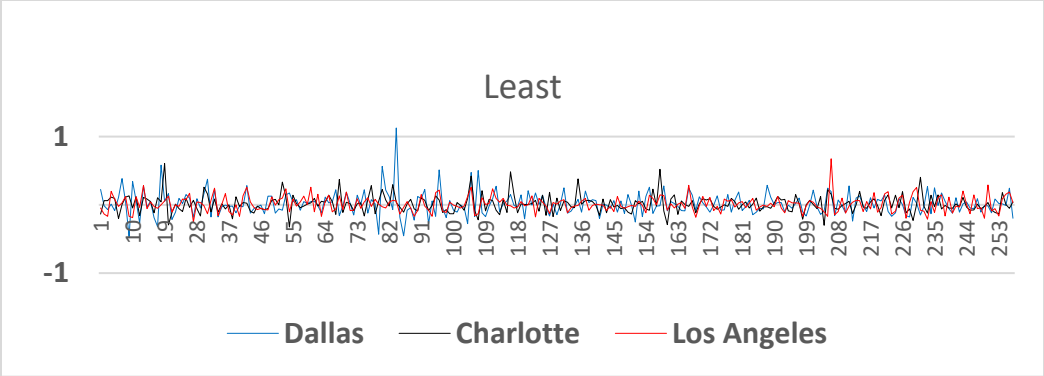
(9a) Most volatile markets of conventional avocados



(9b) Least volatile markets of conventional avocados



(9c) Most volatile markets of organic avocados



(9d) Least volatile markets of organic avocados

Figure 9. Volatility analysis

In terms of organic Hass avocados, Miami, Orlando, and Tampa experience the largest volatility in per capita sales, and we also find that these volatilities coincidentally occurred at almost the same time, as shown in Figure 9c. These three cities are all located in Florida and are relatively closer to the places of product origin. The abundant supply of Hass avocados and seasonal changes in Hass avocado supply throughout the year contribute to the lower prices of organic Hass avocados, thus stimulating the consumption in these regions. A stable trend of organic Hass, avocado sales per capita, is observed in the following cities: Dallas, Charlotte, and Los Angeles, as shown in Figure 9d. The variations in sales per capita among these cities are the lowest throughout the year. The results indicate that demographic factors, like race, eating habits, supply and demand, and the prices of organic Hass avocados in those cities, might be more similar over time.

In comparison, organic Hass avocado markets seem more volatile than conventional Hass avocado markets in general. Volatility analysis of these markets of different types of Hass avocado can help businesses develop more complex strategies in risk management, inventory, warehousing, and promotion activities.

Implications and Limitations

This study concludes with interesting findings that can be useful for importers and retailers. According to the results, the income level and the percentage of the Hispanic population, although significantly correlated, are not dominant factors for the growth of consumption of healthy fruits like avocados. In sum, well-educated consumers seem to be a more important category of customer segment that needs immediate attention from retailers. Based on our regression analysis, the study offers insights into the regional trends and search trends of consumers for potentially better sales and market opportunities. For instance, according to our volatility analysis, importers and retailers may further allocate resources in different regions to reduce volatility in certain regions. Furthermore, based on our classification of different cities and different types of factors,

importers and retailers may consider designing customized dashboards to integrate different types of demographic trends and search behaviors into their conventional managerial operations.

This work contributes to several critical aspects of data analytics and visualization development in the sales and market literature. By combining industry-level data with geodemographic data and Google Trends search data, the paper can advance our understanding of the role of geodemographics and public interest in commercial products. Using geodemographic and Google Trends search data to analyze product sales is an important endeavor that has not been systematically undertaken in business research. Through programmatic empirical studies, this paper applies marketing analytics to a practical business scenario (Iacobucci et al., 2019; Wedel & Kannan, 2016).

This study offers significant contributions to academia. From descriptive and diagnostic perspectives, our work harnesses k-means clustering and hierarchical clustering methods and exploratory mapping methods to offer data-driven visualizations. The cluster analysis and the mapping approach offer different types of information while complementing each other in the use of design patterns, trends, and algorithms. From predictive and prescriptive perspectives, this study shows how Google Trends search interest could complement traditional industry-level and geodemographic variables in marketing research. Furthermore, our predictive and prescriptive analyses shed light on the sales and market of other agriculture products.

Google Trends search data, a proxy of public interest in topics, offers a valuable tool through which suppliers, retailers, and wholesalers can now cast or even forecast consumers' purchasing patterns and preferences (organic vs. conventional avocados in our context) in different regions. In light of the post-COVID-19 era, such well-suited knowledge of consumers is particularly important as consumers would spend more time on the internet, and the influence of trendy topics would greatly impact consumers' decision-making. Furthermore, this analysis integrates descriptive, diagnostic, predictive, and prescriptive analyses into an easy-to-follow, practical approach that might inform future decision-making practices with data-driven marketing analytics.

Although the results are promising, this study does have a few limitations. First, the study performs empirical analyses using a panel data set describing avocado sales and pricing in multiple metropolitan regions from 2015 to 2019, the Google search index for ”avocado”, and the geodemographic data from the US Census Bureau. Due to the COVID-19 disruption in 2020, the geodemographic data in and after 2020 was unavailable when the study was performed. It may be interesting to study the post-COVID datasets to further explore the applicability of developed models in this study. Therefore, practitioners may use models and results from this study with caution. Secondly, the study may have possible contemporaneous correlations between conventional and organic avocado sales, which may have misleading results. For instance,

people who purchase conventional avocados may give up on organic avocados, and vice versa. This limitation can be further explored in the future. Third, this study only focuses on the avocado markets with specific conditions in the US regions. Similar results or conclusions may not be obtained when studying other markets or regions. This research may need to be further expanded and studied for different food markets and retailing industries for future improvement, the process's applicability, and insights. Fourth, the Google Trends index is consolidated and aggregated, which does not provide any information on consumers' preferences, likes or dislikes, and search purposes. For instance, we would not know if a search is performed to take a biology exam or look for avocado recipes. Therefore, these limitations may limit the applications of this study.

Conclusion

This paper is among the first to systematically study Hass avocado sales in the regional markets of the US with industry-level data, geodemographic data, and Google Trends data. Our empirical results potentially provide mechanisms for understanding and explaining causalities between industry-level variables, different geodemographic factors, public interest in avocados as measured by Google Trends, and avocado sales.

Specifically, major findings in this study include: 1) the regression analysis suggests that demographic factors, including education level (i.e., a bachelor' degree or higher) and ethnicity (percentage of Hispanic population in a region), have positive impacts on avocado sales per capita. Without Google Trends data, income is not significantly correlated with sales per capita in the model but significantly and positively correlated with sales per capita after adding Google Trends data. The results also imply that regions with a younger population tend to have lower avocado sales per capita; 2) most importantly, the impact of Google Trends search interest on avocados shows that the higher the active search of avocado as a keyword via the Google search engine in the region, the more avocado sales per capita; 3) through fixed effect factors, the models captures the seasonal patterns as Hass avocado sales normally pick up in spring, reach its peak season during summer and slows down through the rest of year, which is highly consistent with the seasons of growing and harvesting avocados. The models detect a strong growth of avocado consumption over the years; 4) furthermore, they also show that although organic avocados are heathier, their high prices may discourage sales significantly. Our additional volatility analysis shows that organic avocado sales are more volatile than conventional ones due to price, supply and demand, and other relevant issues; 5) the study employs clustering techniques and data visualization tools to explore regional markets further. The 27 regional markets in the study are grouped into a few clusters based on their market characteristics, such as geodemographic and Google Trends data. The clustering results show that geographically close markets are not always in the same market cluster, and markets that are distant geographically may be grouped into the same cluster. The cluster analysis is vital to gaining insights into market segmentation and developing marketing strategies and sales plans. With these

descriptive and diagnostic solutions, businesses can explore further the details and become risk intelligent.

Major findings 1) and 2) are critical for constructing reliable models for future applications. It is worth highlighting that geodemographic factors and Google Trends data are very helpful in understanding regional markets. These factors can be considered when businesses explore opportunities and develop mitigation plans for risks. Some factors show inter-correlation, such as education and income, while ethnicity, median age, and Google Trends provide different perspectives in marking analysis. These two major findings address the research questions in this study and provide foundational knowledge in developing regression models for studying avocado sales and regional markets. Major findings 3) and 4) provide important information for businesses to address seasonal effects and elasticity issues of organic products. Seasonal effects can be addressed by leveraging inventory management, high and low season planning, cash flow, new vendor sources, and cold chain warehouses, which may be further addressed in future research. Additionally, organic avocados have a price premium, and therefore, our study reveals that customers tend to be sensitive to price changes in organic avocado sales. Major finding 5) showcases how clustering tools can use these factors in the models to segment regional markets.

In summary, geodemographics and Google Trends data can be used to model avocado sales. Data analytics and data visualization tools used in this study are very effective and efficient in understanding the Hass avocado markets. Data-driven decision-making in marketing has been a hot topic in recent years. However, it is not easy to identify a practical approach. Although Wedel & Kannan (2016) proposed their data-driven decision-making model, there is scant research about this model, and we recommend that marketers and practitioners further test the data-driven decision-making model proposed by Wedel & Kannan (2016), namely how to support data-driven decision-making with descriptive, diagnostic, predictive and prescriptive analyses. The data-driven decision-making model and the research approach can also be extended to other fields, such as new product development, innovation, and sustainability.

As understanding consumer trends becomes more and more challenging year over year, importers and retailers may benefit from this study. In particular, the study offers a great example of how consumer trends in the food industry are affected by prices, geodemographic factors, and Google Trends. Specifically, the results indicate that importers may focus on regional differences to optimize sales and promotion efforts in different regions. In addition, since education and the percentage of Hispanic consumers emerge as significant predictors for per capita avocado sales, retailers may launch more targeted marketing efforts to reach out to Hispanic consumers and communities with better education levels.

References

Ambrozek, C., Saitone, T. L., & Sexton, R. J. (2018). *Five-Year Evaluation of the Hass Avocado Board's Promotion Programs: 2013–2017*. Hass Avocado Board: Mission Viejo, CA, USA. Retrived from https://hassavocadoboard.com/wp-content/uploads/2019/03/hab-latest-independent-economic-evaluation-2018.pdf

Ambrozek, C., Saitone, T. L., & Sexton, R. J. (2019). *Price Elasticities of Demand for Fresh Hass Avocados in the United States: Concepts, Estimation, and Applications*. Hass Avocado Board: Mission Viejo, CA, USA. Retrieved from https://hassavocadoboard.com/wp-content/uploads/2019/04/Hass-Avocado-Board-Price-Elasticities-of-Demand-or-Fresh-Hass-Avocados-2019.pdf

Baviera-Puig, A., Buitrago-Vera, J., & Escriba-Perez, C. (2016). Geomarketing models in supermarket location strategies. *Journal of Business Economics and Management*, *17*(6), 1205-1221.

Bezawada, R., & Pauwels, K. (2013). What is special about marketing organic products? How organic assortment, price, and promotions drive retailer performance. *Journal of Marketing*, *77*(1), 31-51.

Calantone, R. J., & Di Benedetto, C. A. (2007). Clustering product launches by price and launch strategy. *Journal of Business & Industrial Marketing,* 22(1), 4-19. DOI: 10.1108/08858620710722789.

Chou, S. F., Horng, J. S., Liu, C. H. S., & Lin, J. Y. (2020). Identifying the critical factors of customer behavior: an integration perspective of marketing strategy and components of attitudes. *Journal of Retailing and Consumer Services*, *55*, 102113.

Clark, S. D., Shute, B., Jenneson, V., Rains, T., Birkin, M., & Morris, M. A. (2021). Dietary patterns derived from UK supermarket transaction data with nutrient and socioeconomic profiles. *Nutrients*, *13*(5), 1481.

Cutler, D. M., & Lleras-Muney, A. (2010). Understanding differences in health behaviors by education. *Journal of Health Economics*, 29(1), 1-28.

Dawes, J. (2014). Cigarette brand loyalty and purchase patterns: an examination using US consumer panel data. *Journal of Business Research*, *67*(9), 1933-1943.

Dickler, J. (2021). More education doesn't always get you more money, report finds. *CNBC Report*. Retrieved from https://www.cnbc.com/2021/10/13/more-education-doesnt-always-get-you-more-money-report-finds.html

Dos Santos, M. J. P. L. (2018). Nowcasting and forecasting aquaponics by Google Trends in European countries. *Technological Forecasting and Social Change*, *134*, 178-185.

Dougherty, C. (2011). *Introduction to econometrics*. Oxford University Press, UK.

Dreher, M. L., & Davenport, A. J. (2013). Hass avocado composition and potential health effects. *Critical Reviews in Food Science and Nutrition*, *53*(7), 738-750.

Eastman, J. K., & Liu, J. (2012). The impact of generational cohorts on status consumption: an exploratory look at generational cohort and demographics on status consumption. *Journal of Consumer Marketing,* 29(2), 93-102. DOI: 10.1108/07363761211206348

Einav, L., & Levin, J. (2014). Economics in the age of big data. *Science*, *346*(6210), 1243089.

Formánek, T., & Sokol, O. (2022). Location effects: geo-spatial and socio-demographic determinants of sales dynamics in brick-and-mortar retail stores. *Journal of Retailing and Consumer Services*, 66, 102902.

France, S. L., Shi, Y., & Kazandjian, B. (2021). Web Trends: a valuable tool for business research. *Journal of Business Research*, *132*, 666-679.

Glass S., & Jarett, L. (2021). Make informed decisions with Google Trends data. *Google Cloud Blog - Developers & Practitioners, July 21.* Retrieved from https://cloud.google.com/blog/topics/developers-practitioners/make-informed-decisions-google-trends-data.

HAB (2019). Millennials more likely to buy avocados. *HAB Newsletter*, Hass Avocado Board: Mission Viejo, CA, USA. Retrieved from https://hassavocadoboard.com/2019/08/19/millennials-more-likely-to-buy-avocados/

Hardy, A. (1994). An examination of procedures for determining the number of clusters in a data set. In *New Approaches in Classification and Data Analysis* (pp. 178-185). Springer, Berlin, Heidelberg.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: a k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics),* 28(1), 100-108.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica: Journal of the Econometric Society*, 1251-1271.

Higuchi, A., & Maehara, R. (2021). A factor-cluster analysis profile of consumers. *Journal of Business Research*, *123*, 70-78.

Iacobucci, D., Petrescu, M., Krishen, A., & Bendixen, M. (2019). The state of marketing analytics in research and practice. *Journal of Marketing Analytics*, *7*(3), 152-181.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 112, p. 18). New York: Springer.

James, W. H., Lomax, N., Birkin, M., & Collins, L. M. (2021). Geodemographic Patterns of Meat Expenditure in Great Britain. *Applied Spatial Analysis and Policy*, *14*(3), 563-590.

Kamath, P. R., Pai, Y. P., & Prabhu, N. K. (2019). Building customer loyalty in retail banking: a serial-mediation approach. *International Journal of Bank Marketing,* 38(2), 456-484. DOI: 10.1108/IJBM-01-2019-0034.

Newsome, L. (2018). Avocado Creep is the Reason Why America is the Greatest Country in the World, *Nebo Agency Blog*, July 6. Retrieved from https://www.neboagency.com/blog/avocado-creep-is-the-reason-why-america-is-the-greatest-country-in-the-world/

Kirby-Hawkins, E., Birkin, M., & Clarke, G. (2019). An investigation into the geography of corporate e-commerce sales in the UK grocery market. *Environment and Planning B: Urban Analytics and City Science*, *46*(6), 1148-1164.

Kourgialas, N. N., & Dokou, Z. (2021). Water management and salinity adaptation approaches of Avocado trees: a review for hot-summer Mediterranean climate. *Agricultural Water Management*, *252*, 106923.

Li, J., & Powdthavee, N. (2015). Does more education lead to better health habits? Evidence from the school reforms in Australia. *Social Science & Medicine*, 127, 83-91.

Liu, R., An, E., & Zhou, W. (2021). The effect of online search volume on financial performance: marketing insight from Google Trends data of the top five US technology firms. *Journal of Marketing Theory and Practice*, *29*(4), 423-434.

Lufkin, B. (2019). How avocados and kale became so popular. *BBC Worklife*, https://www.bbc.com/worklife/article/20190304-how-avocados-and-kale-became-so-popular

Migliore, G., Farina, V., Tinervia, S., Matranga, G., & Schifani, G. (2017). Consumer interest towards tropical fruit: factors affecting avocado fruit consumption in Italy. *Agricultural and Food Economics*, *5*(1), 1-12.

Nghiem, L. T., Papworth, S. K., Lim, F. K., & Carrasco, L. R. (2016). Analysis of the capacity of Google Trends to measure interest in conservation topics and the role of online news. *PloS One*, 11(3), e0152802.

Palmon, O., Smith, B., & Sopranzetti, B. (2004). Clustering in real estate prices: determinants and consequences. *Journal of Real Estate Research*, *26*(2), 115-136.

Prasad, C. J., & Aryasri, A. R. (2011). Effect of shopper attributes on retail format choice behaviour for food and grocery retailing in India. *International Journal of Retail & Distribution Management,* 39(1), 68-86. DOI: 10.1108/ 09590551111104486

Rains, T., & Longley, P. (2021). The provenance of loyalty card data for urban and retail analytics. *Journal of Retailing and Consumer Services*, *63*, 102650.

Rana, J., & Paul, J. (2017). Consumer behavior and purchase intention for organic food: a review and research agenda. *Journal of Retailing and Consumer Services*, *38*, 157-165.

Ravenscraft, D. J. (1983). Structure-profit relationship at the line of business and industry level. *The Review of Economics and Statistics*, 65(1), 22-31. DOI: 10.2307/1924405 22-31.

Rodriguez, A. E., Ozkul, A. S., & Marks, B. A. (2018). Explaining impact of predictors in rankings: an illustrative case of states rankings. *Journal of Business Analytics*, *1*(2), 135-143.

Scitovski, R., Sabo, K., Martínez-Álvarez, F., & Ungar, Š. (2021). *Cluster Analysis and Applications*. New York: Springer.

Silva, E. S., Hassani, H., Madsen, D. Ø., & Gee, L. (2019). Googling fashion: forecasting fashion consumer behavior using Google Trends. *Social Sciences*, *8*(4), 111; DOI: 10.3390/socsci8040111

Skenderi, G., Joppi, C., Denitto, M., & Cristani, M. (2021). Well Googled is half done: multimodal forecasting of new fashion product sales with image-based Google Trends. *arXiv preprint arXiv:2109.09824*.

The Produce News. (2021). *California avocado forecast up over last season*. Retrieved from https://theproducenews.com/avocados/california-avocado-forecast-over-last-season.

Thompson, C., Clarke, G., Clarke, M., & Stillwell, J. (2012). Modelling the future opportunities for deep discount food retailing in the UK. *The International Review of Retail, Distribution and Consumer Research*, *22*(2), 143-170.

USDA (2022). *Fruit and Tree Nuts Yearbook Tables - Noncitrus Fruit.* US Department of Agriculture, Economic Research Service. Retrieved from https://www.ers.usda.gov/data-products/fruit-and-tree-nuts-data/fruit-and-tree-nuts-yearbook-tables/.

Vukasovič, T. (2016). Consumers' perceptions and behaviors regarding organic fruits and vegetables: marketing trends for organic food in the twenty-first century. *Journal of International Food & Agribusiness Marketing*, *28*(1), 59-73.

Wedel, M., & Kannan, P. K. (2016). Marketing analytics for data-rich environments. *Journal of Marketing*, *80*(6), 97-121.

Yoo, T. W., & Oh, I. S. (2020). Time series forecasting of agricultural product' sales volumes based on seasonal long short-term memory. *Applied Sciences*, *10*(22), 8169.

Yu, L., Zhao, Y., Tang, L., & Yang, Z. (2019). Online big data-driven oil consumption forecasting with Google trends. *International Journal of Forecasting*, *35*(1), 213-223.

Zu, X., Wang, X., & Cui, Y. (2022). Forecasting natural gas consumption in residential and commercial sectors in the US. *Journal of Business Analytics*, 6(1), 77-94. DOI: 10.1080/2573234X.2022.2064777.

About the Authors

Di Wu*

Department of Accounting and Finance
California State University, Bakersfield
E-mail: dwu2@csub.edu

Zhenning Xu

Department of Management and Marketing
California State University, Bakersfield
E-mail: zxu3@csub.edu

Ji Li

Department of Accounting and Finance
California State University, Bakersfield
E-mail: jli7@csub.edu

*Corresponding author

Di Wu is an Associate Professor of Accounting at California State University, Bakersfield. Dr. Wu' 's research interests include the application of data analytics in accounting, cost accounting, and information systems.

Zhenning Xu is an Assistant Professor of Marketing at California State University, Bakersfield. Dr. Xu' 's research interests include marketing analytics, marketing channels research, and marketing education.

Ji Li is a Professor of Accounting at California State University, Bakersfield. Dr. Li' 's research interests include financial accounting, executive compensation, and managerial accounting topics.