DOI: https://doi.org/10.48009/2_iis_2022_118

Using unsupervised machine learning to determine social networking user groups

Alan Peslak, *Penn State University, arp14@psu.edu*
Wendy Ceccucci, *Quinnipiac University, Wendy.Ceccucci@quinnipiac.edu*
Scott Hunsinger, *Appalachian State University, hunsingerds@appstate.edu*

Abstract

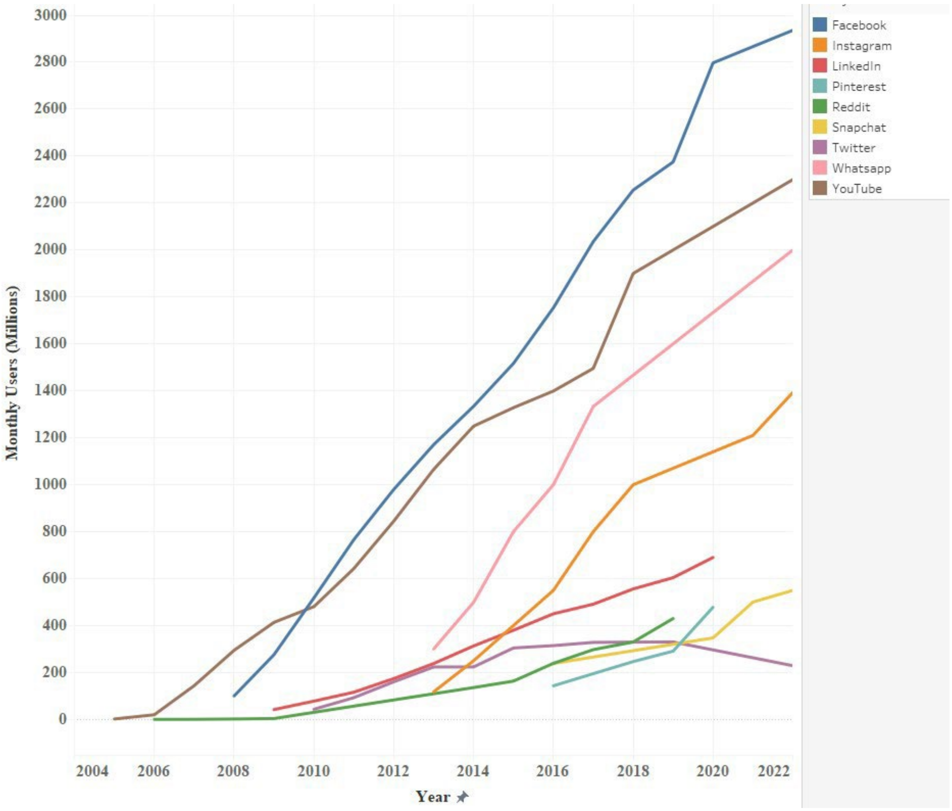
The growth of social media and networking has been exponential. From its humble beginnings in 1995 with Classmates.com through the founding of Friendster in 2002, LinkedIn and MySpace in 2003 and Facebook in 2004, social networking has grown to a worldwide phenomenon with nearly 2.89 billion worldwide active users of Facebook alone (Statista, 2022). The number of major social media sites has also grown, though the top active sites in the United States represent most of the social media activity. We examine a 2019 Pew Internet dataset via Unsupervised Machine Learning with the goal of finding Social Networking User Groups. Usage of the top social media websites is combined with relevant demographic and sociographic data to develop three specific clusters of users of social media in the US. Implications for marketers, researchers and society are discussed.

Keywords: Social media**,** Cluster analysis, Facebook, Social Networking

Introduction

Social Networking is the use of internet-based platforms to interact with other users, or to make new connections with others who have similar interests. Since the mid-1990s, the popularity and number of social networking platforms has grown greatly. Figure 1 shows the growing monthly use of each of these applications. In the United States at least 72% of adults use some type of social media (Pew Research, 2021). Some of the more popular social networking sites and applications include Facebook, Instagram, Twitter, Snapchat, and YouTube. A brief description of each of these applications is given in the next section.

The users of each of these sites may have similar characteristics. These groupings of similar social networking users would be useful to a variety of audiences. To determine social media networking groups, we first conduct a literature review, examining social media usage in the US through several categories including age and gender, education level, and income level. We also provide an overview of cluster analysis, the technique used to determine the groupings. Next, we describe the methodology used for our study along with the predictor importance. Three specific clusters of social media users in the US are identified. The discussion and conclusions section follows, in which we explore the implications of our findings and suggest ideas for future research.



Literature review

Social Media Platforms

A brief description of each of the different platforms are given here**.**

- Facebook, created in 2004 is the largest social network. It is a social networking site that allows users to create profiles to stay connected with colleagues, friends, and family. It allows users to share contents such as status posts, images, and videos.
- Instagram is a platform that emphasizes photo and video sharing via its mobile app. Users can take, edit, and publish visual content for both followers and non-followers, if the account is public. Other users can interact with the content via likes, comments, shares, and saves.
- LinkedIn is the world's largest professional network on the Internet. It is used to find jobs or internship, connect and strengthen professional relationships, and to learn the skills needed for different careers.
- Pinterest is a visual sharing and search platform. It is used to search for new ideas like recipes, home decorating ideas, and more.
- Snapchat is an app that is used on mobile phones that allows users to send texts, pictures or videos to friends. The one feature that makes Snapchat different from other forms of texting and photo sharing: the messages disappear from the recipient’s phone after a few seconds.
- Twitter is a platform that allow users to share their thoughts and news with a large audience. Users create and post tweets that are 280 characters or less. People who follow the user and potentially others will see the tweet.

- WhatsApp is a messaging application that uses the internet to send messages, images, audio or video.
- YouTube is a video sharing website that allows users to create, post, and watch online videos online.
- Reddit is a social news website and forum where content is socially curated and promoted by site members voting (Stafford, 2022). It is organized into a network of communities, where users can explore items of interest.

Social Media Usage & Demographics

There are now approximately four billion users of social media worldwide (Kemp, 2020). Each individual user, on average, spends over two hours daily on social media and uses eight different social media accounts (Omnicore, 2020).

The United States has one of the highest social network penetration rates in the world with over 70% of the US having a social media account (Statista, 2022). Facebook is the most popular social network in the US based on monthly active users. Other social media platforms have been gaining popularity. One such network is Instagram which has been gaining popularity amongst the Millennials and Gen Z users. In this section, we will look further into the usage and demographics of several of the more popular social media networks including Twitter, Instagram, Facebook, Snapchat, YouTube, WhatsApp, Pinterest, LinkedIn, and Reddit.

Age & Gender

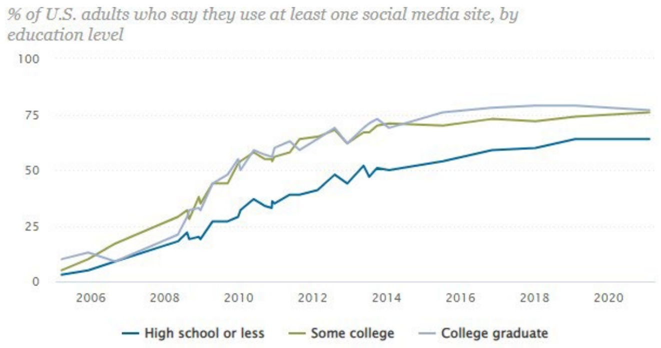
Table 1 shows data gathered on the age and gender of social media users. The data was taken from Statista and from Brent Barnhart, SproutSocial (2022). Except for Pinterest, young adults are the primary users of social media networking sites. The distribution of users by gender varies amongst each application. Twitter and Reddit have a larger percentage of male users compared to other applications, while Pinterest and Snapchat have a larger female base.

Table 1 Age and Gender of Social Media Users

	Largest Age Group	Gender	
		%Female	%Male
Twitter	18-29 (42%)	38.4	61.6
Instagram	25-34 (31.2%)	48.4	51.8
Facebook	25-34 (31.5%)	43	57
Snapchat	15-25 (48%)	54.4	44.6
YouTube	15-35 (N/A)	46	54
WhatsApp	26-35 (27%)	46	54
Pinterest	50-64 (38%)	78	22
LinkedIn	25-34 (58.4%)	48	52
Reddit	18-29 (36%)	36	64

Education Level and Marital Status

Figure 1 shows the social media usage by education level (Pew Research, 2021). US users with at least some college education are more likely to use at least one social media site than those who indicate they have a high school education or less. Data from Statista and Pew Research on US adults who were using each social media platform based on education level is shown in Table 2 (Statista, 2022 and Pew Research, 2021). This table indicates the percentage of US adults in each demographic group who indicated they have ever used each social media application. For example, 44% of respondents who had attained a college degree have used Instagram. According to Pew Research, users with higher levels of education are more likely to report being LinkedIn users. Roughly half of adults who have a bachelor's or advanced degree (51%) say they use LinkedIn, compared with smaller shares of those with some college experience (28%) and those with a high school diploma or less (10%). Additionally, most Reddit users have either some college education or a degree, with the smallest group of users having only a high school degree. As far as marital status, 39% of Facebook users report being married, while another 39% report being single. LinkedIn on the other hand removed their relationship status from their website in 2016.



% of U.S. adults who say they use at least one social media site, by education level

Note: Respondents who did not give an answer are not shown. 'High school or less' includes both 'Less than high school graduate' and 'High school graduate'. The 2021 sample is too small to separate these two categories.
Source: Surveys of U.S. adults conducted 2005-2021.

PEW RESEARCH CENTER

Figure 1: % of US Adults who say they use at least one Social Media site (by education Level) (Pew Research, 2021)

Table 2: 2021- 22 Education Level of Social Media Users

	High School or Less	Some College	College +
Twitter	14%	26%	33%
Instagram	30%	44%	49%
Facebook	64%	71%	73%
Snapchat	N/A	N/A	N/A
YouTube	70%	86%	89%
WhatsApp	20%	20%	33%
Pinterest	22%	36%	37%
LinkedIn	10%	28%	51%
Reddit	21%	42%	36%

Income Level

According to Statista (2022), As of February 2019, Americans in the income bracket from $50,000 to $74,999 were the largest users of social networks. 83% of all adults making an income within that bracket use social networks. High income users were next with 78%, followed by users in the $30,000 to $49,999 bracket (Figure 2). Lastly, 68% of adults with income less than $30,000 used social networks. In the US, YouTube is most popular social media platform among high earners. 90% of those earning $75,000 or more say that they sometimes visit YouTube. Among Americans with an income of $30,000-$49,999 income, 83% are YouTube users, and dropping to 79% among US adults with an income of between $50,000 and $74,999. It is even lower for Americans earning under $30,000, with 75% of this group using YouTube (Pew Research, 2021). Seventy percent of people earning more than $75,000 are on Facebook (SproutSocial).

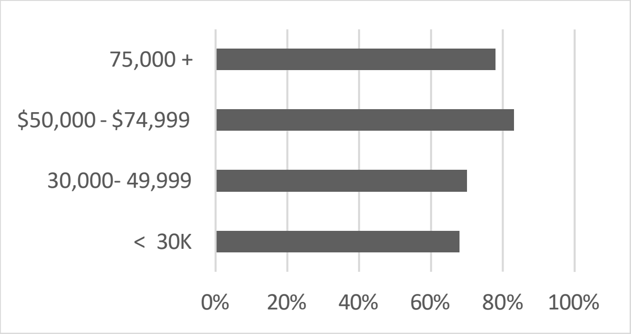


Figure 2: % of US Adults who say they use at least one Social Media site (by Marital Status)

Methodology

The dataset used in this exploratory analysis was from Pew Research and a report written by Monica Anderson (2019) with the Pew Research Center in June 2019. According to the report, the data gathered were based on telephone interviews conducted between Jan. 8 to Feb. 7, 2019. They sampled 1,502 adults, 18 years of age or older. A combination of landline and cellphone random-digit-dial samples were used. The interviews were conducted in English and Spanish. Respondents in the landline sample were randomly selected and were asked for the youngest adult male or female who is now at home. Interviews in the cell sample were conducted with the person who answered the phone, if that person was an adult 18 years of age or older.

According to the article the data was further adjusted as described:

> “The combined landline and cellphone sample is weighted using an iterative technique that matches gender, age, education, race, Hispanic origin, and nativity and region to parameters from the 2017 Census Bureau’s American Community Survey one-year estimates and population density to parameters from the Decennial Census. The sample is also weighted to match current patterns of telephone status (landline only, cellphone only, or both landline and cellphone), based on extrapolations from the 2018 National Health Interview Survey. The weighting procedure also accounts for the fact that respondents with both landline and cellphones have a greater probability of being included in the combined sample and adjusts for household size among respondents with a landline phone. The margins of error reported, and statistical tests of significance are adjusted to account for the survey’s design effect, a measure of how much efficiency is lost from the weighting procedures.”

From this dataset, relevant social media data and basic demographic information were extracted to understand whether there were specific discrete clusters within the sample population and then to further analyze these groups (Table 3).

Table 3 Pew Survey Questions Used

Category	Variable	Question
Social Media		Please tell me if you ever use any of the following social media sites. Do you ever use
	Web1a	Twitter?
	Web1b	Instagram?
	Web1c	Facebook?
	Web1d	Snapchat?
	Web1e	YouTube?
	Web1f	WhatsApp?
	Web1g	Pinterest?
	Web1h	LinkedIn?
	Web1i	Reddit?
Demographic	Race	
	Location	Urban-Suburban-Rural distinction
	Age	What is your age? [YEARS]
	Sex	
	Marital	Are you currently married, living with a partner, divorced, separated, widowed, or have you never been married?
	Educ2	What is the highest level of school you have completed or the highest degree you have received?
	Inc	Last year -- that is in 2018 -- what was your total family income from all sources, before taxes? Just stop me when I get to the right category
	Party	In politics TODAY, do you consider yourself a Republican, Democrat, or independent?
	L1	Now thinking about your telephone use. Do you have a working cell phone?

Cluster Analysis

Clustering is a technique of grouping similar observations into smaller groups within the larger population. The resulting groups should be homogeneous, with each member of the cluster having more in common with members of the same cluster than with members of the other clusters. Cluster analysis is an unsupervised machine learning technique which aims at sorting. different objects into groups in a way to maximize the degree of association between objects in the same cluster. In this research, the silhouette method is used to create groups of clusters of social media users. This method first described by Kaufman and Rousseuw (1990) is often used for validating clusters of data and can also be used to determine the appropriate number of clusters. The method measures how similar a variable is to its assigned cluster (*cohesion*) compared to other clusters (*separation*). It determines how well each variable lies within its cluster. The silhouette coefficient values range from -1 to 1. The higher the average silhouette score the better the clustering. Silhouette coefficients near positive 1 indicate that the variable is far away from the neighboring clusters. A value of 0 indicates that the variable is near or in between the two neighboring clusters and it is indifferent between the two clusters. A negative value indicates that the variable may have been assigned to the wrong cluster. The number of clusters that maximizes the average silhouette is the optimal number of clusters. The formula to calculate the silhouette score for each data value, i is

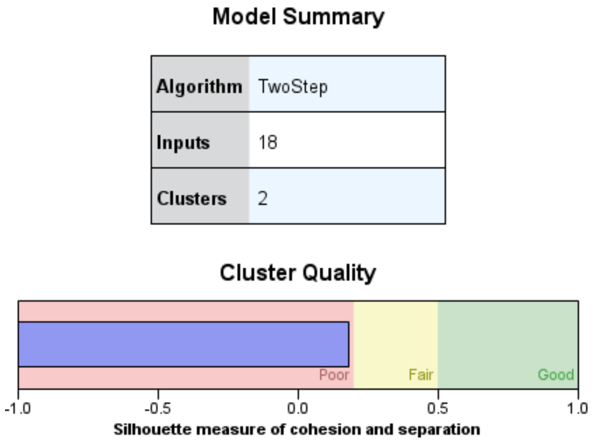
$$ss_{ii} = \frac{bb_{ii} - aa_{i \quad i}}{\max(bbii, aaii)}$$

where a_i measures the distance between the observation and all other observations within the cluster and b_i measures the separation, the distance between the observations and the next closest cluster center.

The technique used was cluster analysis, or clustering, and is an unsupervised machine learning task. It involves automatically discovering natural grouping in data. Unlike supervised learning (like predictive modeling), clustering algorithms only interpret the input data and find natural groups or clusters in feature space. (Wilson, 2020). Our goal was to obtain clusters with a minimum silhouette score of .3 or above. Cluster results are considered appropriate when silhouette width is >0.2. Though .2 is regarded as a fair score (Boos et al., 2021), we wished to provide a more robust clustering and through an iterative process of eliminating variables we worked to achieve this outcome. The silhouette score of 1 means that the clusters are very dense and nicely separated. The score of 0 means that clusters are overlapping. The score of less than 0 means that data belonging to clusters may be wrong/incorrect. (Bhardwaj, 2020).

To perform the cluster analysis, we used SPSS 26 and applied two-step cluster analysis. After review of the dataset, the following variables were included to determine the first pass at identifying relevant clusters within the dataset.

The first pass through the two-step cluster analysis algorithm used all the eighteen variables given in Table 3. This first run of the cluster analysis yielded a poor silhouette value, below .2 as shown in Figure 2.



Model Summary

Algorithm	TwoStep
Inputs	18
Clusters	2

Figure 2. Cluster Analysis Results First Attempt - All Variables

The Predictor Importance table, shown in Figure 3 showed that though there are both significant and insignificant variables. The insignificant variables provide little contribution to the cluster model. As a result, the low-ranking variables were purged and a new analysis was run without the following variables: sex, race, political party, Snapchat, urban-suburban, WhatsApp, and age.

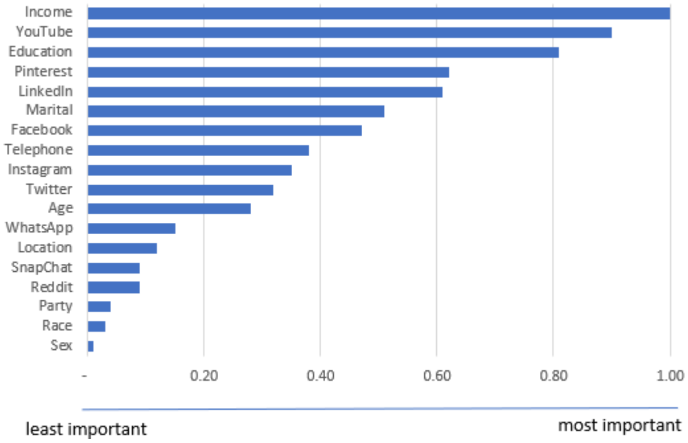
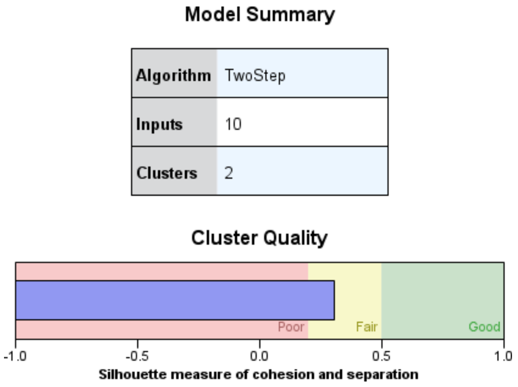


Figure 3 Predictor Importance - First Pass

The results are shown in Figures 4 and 5. This second pass resulted in an increased Silhouette score. However, the Predictor Importance table still revealed two variables with below a .25 influence score. Table 4 shows the two clusters that resulted from this second run and show the differences between the cluster. A third and final pass was then performed excluding the two variables marital status and cell phone ownership.



Model Summary

Algorithm	TwoStep
Inputs	10
Clusters	2

Figure 4. Cluster Analysis Results – Second Attempt

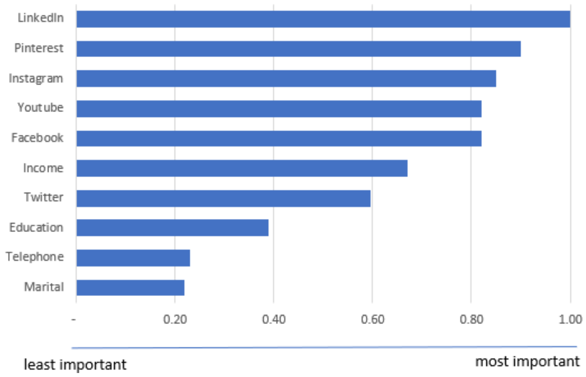
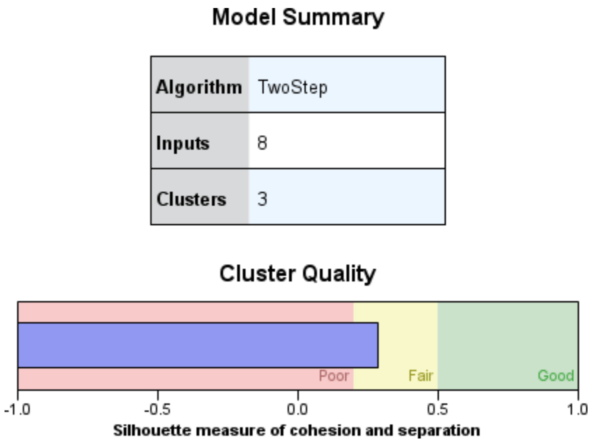


Figure 5. Predictor Importance – Second Pass

Table 4. Cluster Results – Second Pass

Variable	Cluster 1	Cluster 2
LinkedIn	No	Yes
Pinterest	No	Yes
Instagram	No	No
YouTube	No	Yes
Facebook	No	Yes
Income	Low	High
Twitter	No	No
Education	High School	4-year college
Cell Phone	Yes	Yes
Married	Yes	Yes

Figures 7 & 8 show the results of the third run. The cluster analysis yielded three specific clusters with a Silhouette value of .3, well above the minimum acceptable of .2. The characteristic of the three clusters of social media users in the United States are given in Table 5.



Model Summary

Algorithm	TwoStep
Inputs	8
Clusters	3

Figure 7. Final Cluster Model Results

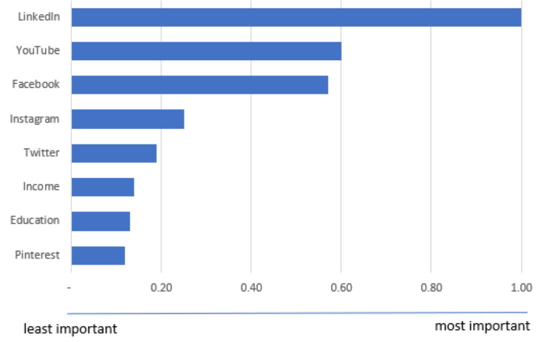


Figure 8 Predictor Importance – Final Pass

Table 5. Final Cluster Characteristics

Cluster 1	Cluster 2	Cluster 3
Middle Income	High Income	High Income
High School Education	High School Education	4-Year College
YouTube	No Social Media	YouTube
Facebook		LinkedIn
		Facebook
		Instagram

The characteristics of the three clusters are shown in Table 6. The clusters can be characterized as cluster 1 as basic users, cluster 2 as never use, and cluster 3 as all-in users. The cluster sizes are all relatively equal as shown in Figure 9. The largest group are the Never social media group. This suggests to marketers or researchers that reaching a full cross section of society will not be able to be accomplished solely through social media since this large group of society does not use any form of popular social media.

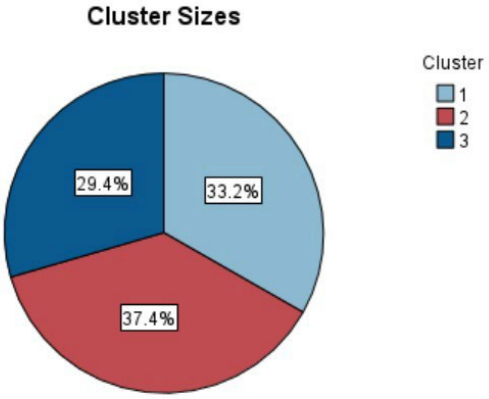


Figure 9. Final Cluster Proportions

Table 6. Final Cluster Distribution

	Cluster 1	Cluster 2	Cluster 3
Social Media	Basic	Never	All-In
LinkedIn	No 99.8%	No 93.1%	Yes 95.6 %
Pinterest	No 65.6%	No 90.9%	No 55.0%
Instagram	No 54.9%	No 96.8%	Yes 54.6%
YouTube	Yes 95.4%	No 70.4%	Yes 95.8%
Facebook	Yes 95.8%	No 74.9%	Yes 85.3%
Income	Middle	High	High
Twitter	No 77.2%	No 96.3%	No 53.6%
Education	High School	High School	4 year college

Looking at Cluster 1, it shows that 99.8% of the respondents do not use LinkedIn, but 95.8% use Facebook and 95.4% use YouTube. The distribution of education for Cluster 1 is given in Figure 10. From the chart we can see that the largest group was a high school degree but there were the next group was also somewhat large and were those with a 4-year degree. The income distribution, given in Figure 11 shows that the income distribution is relatively evenly distributed within the cluster.

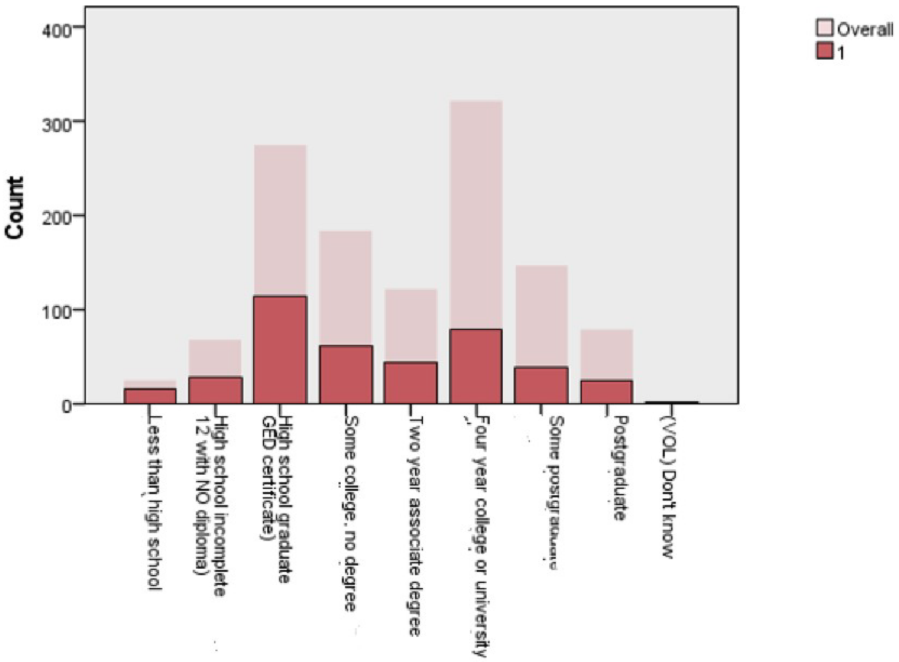


Figure 10. Cluster 1. Basic Users - Education Distribution

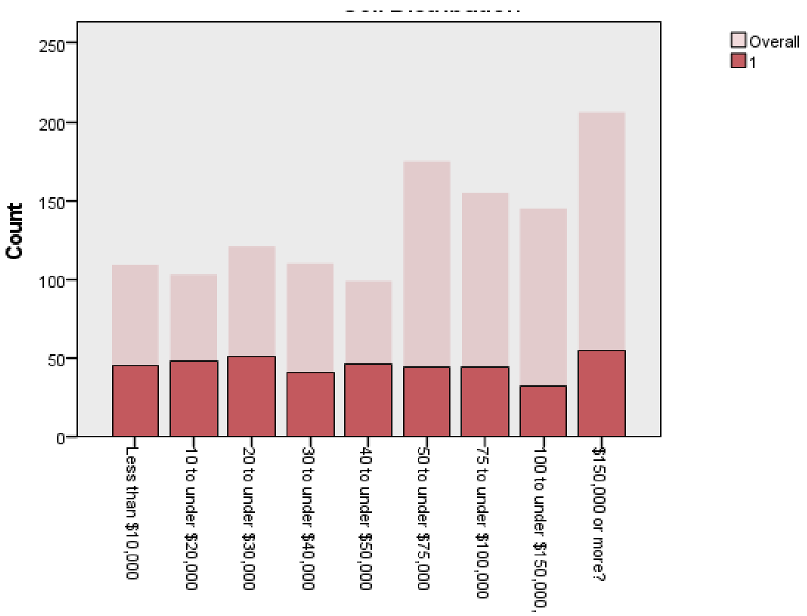


Figure 11. Cluster 1 – Basic Users Income Distribution

Cluster 2, described as the never use cluster are made of respondents who indicate that the do not use most of the different social media. Cluster 2 consists of 98.8% of users who do not use Instagram, 96.3% do not use Twitter and 93.1% do not use LinkedIn. The education level distribution, shown in Figure 12 indicates that the cluster is characterized by mostly respondents with a high- school level education. The income distribution is given in Figure 13. The highest percentage of respondents consist of those with an income between $50,000 to $75,000.

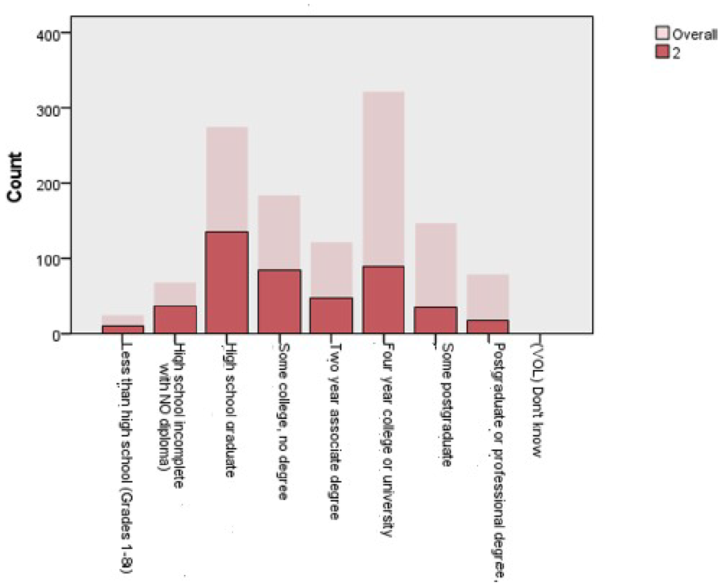


Figure 12. Cluster 2. Never Users - Education Distribution

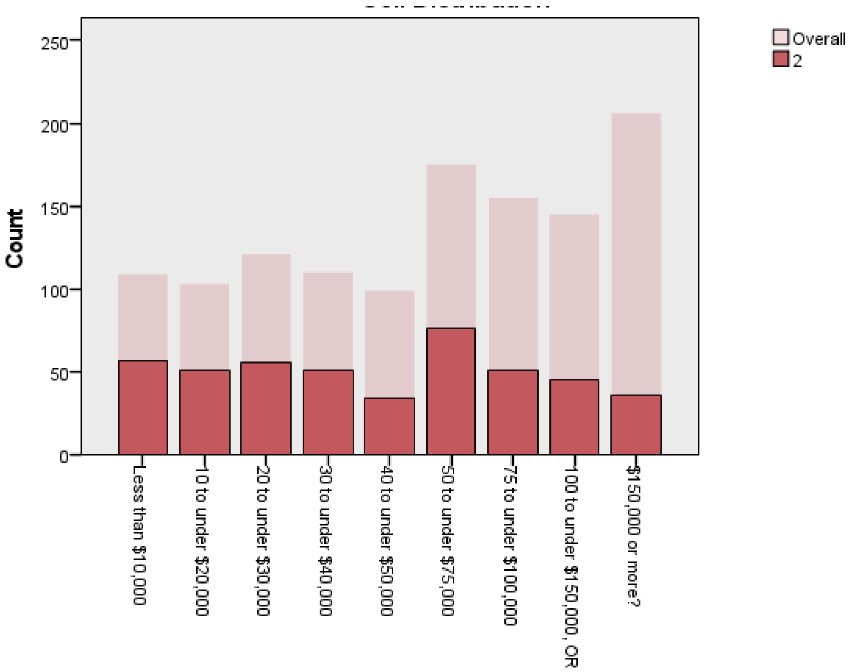


Figure 13. Cluster 2 Never Users - Income Distribution

Cluster 3, described as the All-In cluster are made of respondents who indicate that they use most of the different social media applications, except for Pinterest. Cluster 3 consists of 95.6% of users who use LinkedIn, 95.8% use YouTube and 85.3.% use Facebook. The education level distribution, shown in Figure 14 indicates that the cluster is characterized by mostly respondents with a 4-year college education. The income distribution is given in Figure 15. The highest percentage of respondents consist of those with an income between $50,000 to $75,000.

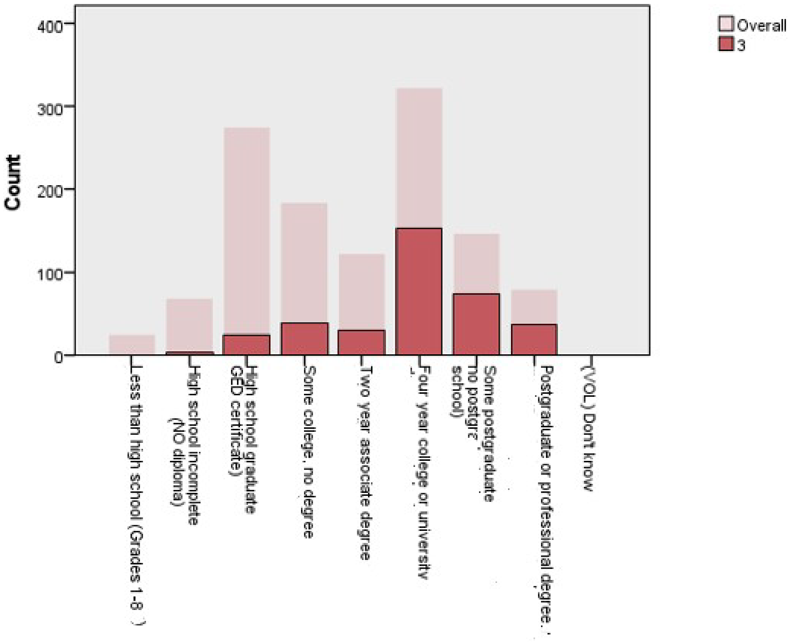


Figure 14. Cluster 3. All In Users - Education Distribution

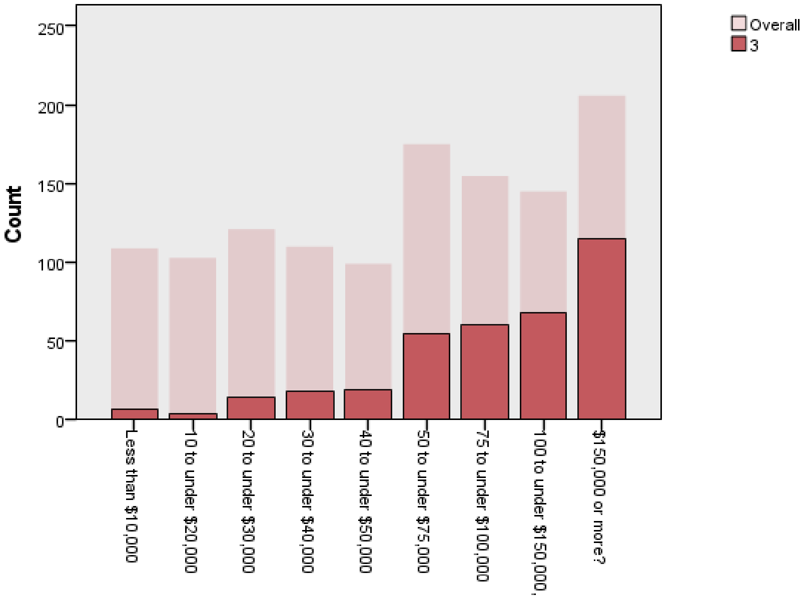


Figure 15. Cluster 3 All In Users - Income Distribution

Discussion and Conclusions

The unsupervised machine learning to determine social networking user groups cluster analysis found three distinct groups of social media users. We are dubbing these groups the Basic users, the Never users and the All-In Users. The determination of these groups holds high impact for researchers, marketers, and communicators alike.

The first group is the Basic group. These individuals are generally low to middle income individuals primarily with a high school education. They are almost totally absent from LinkedIn and show a low interest in Pinterest. They are primarily high users of both YouTube and Facebook and only are about 25% participants in Twitter and 50% in Instagram. Their specific interests seem to revolve around only entertainment (YouTube) and a social network (Facebook).

The Never users generally shun all forms of social media. Over 90% do not use LinkedIn, Pinterest, Instagram and Twitter. Their use of YouTube and Twitter is also low at only 20 to 30 percent. This is an interesting income group as they generally are about equal in all categories of income. The most prevalent educational category for this group is high school education but there are significant members with 4-year degrees as well as graduate and post graduate work. This group appears to be the most diverse in education dispersion. We speculate this may be a result of a combination of factors such as political philosophy, personal privacy concerns, and lack of interest in the related content. This group requires further study.

The final group is the All-In group. These individuals are very active users of YouTube, LinkedIn and Facebook and also have the highest participation in Pinterest, Instagram, and Twitter. An interest highlight is that though 85% use Facebook, 95% use LinkedIn. They have high income and are generally college graduates or above. Another interesting distinction is that even though they have high Facebook participation, they are not nearly as high as the Basic group (85% versus 95%).

Our cluster groupings should hold high interest for researchers, educators, and marketers as they identify specific groups that can be used for further research, directed curriculum, and targeted marketing campaigns.

For future study, it would be interesting to investigate whether there have been any changes in demographics since the data were collected in 2019. As cell phones and tablets are more heavily adopted by various age and income groups, we may see different trends in usage of social applications. In addition, newer social networking applications such as TikTok have increased in popularity in the US since 2019. Follow-up research should examine newer clusters of social media sites and applications.

References

Anderson, Monica (2019) Mobile Technology and Home Broadband, Pew Research Center, June Retrieved from https://www.pewresearch.org/internet/2019/06/13/mobile-technology-and-home-broadband-2019/ on April 19, 2022.

Bardwaj, A. (2020) Silhouette Coefficient https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c#:~:text=Its%20value%20ranges%20from%20%2D1,assigned%20in%20the%20wrong%20way.

Barnhart, Brent (2022) Social media demographics to inform your brand's strategy in 2022. Retrieved from https://sproutsocial.com/insights/new-social-media-demographics/#twitter-demographics on April 19, 2022

Boos, S. C., Wang, M., Karst, W. A., Hymel, K. P., & Pediatric Brain Injury Research Network (PediBIRN) Investigators. (2022). Traumatic head injury and the diagnosis of abuse: a cluster analysis. Pediatrics, 149(1).

Kaufman, L., and Rousseuw, P. (1990). Finding Groups in Data: An Introduction to Cluster Analysis, Wiley Publishing.

Kemp, S. (2020, November). Digital 2020: 3.8 billion people use social media. https://www.wearesocial.com/blog/2020/01/digital-2020-3-8-billion-people-use-social-media

Omnicore . (2020, November). Social media benchmark report 2020: Statistics, trends, and interesting facts. http://www.omnicoreagency.com/social-media-statistics/

Pew Research (2021) Social Media Fact sheet. Retrieved from https://www.pewresearch.org/internet/fact-sheet/social-media/?menuItem=f13a8cb6-8e2c-480d-9935-5f4d9138d5c4 on April 19, 2022

Stafford, C. (2022) Definition Reddit posted on TechTarget.com. Retrieved from https://www.techtarget.com/searchcio/definition/Reddit on July 1, 2022

Statista, (2022) Most popular social networks worldwide as of January 2022, ranked by number of monthly active users, retrieved from **https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/** on April 19, 2022.

Statista, (2022) Social media usage in the United States - Statistics & Facts, retrieved from https://www.statista.com/topics/3196/social-media-usage-in-the-united-states/#dossierKeyfigures

Wilson (2020) How are clustering algorithms different from supervised learning? https://it-qa.com/how-are-clustering-algorithms-different-from-supervised-learning/