Segmenting User Behavior on the Dating App Tinder with Unsupervised Learning Algorithms

Emma Taylor

CPT_S 575: Data Science

Washington State University

December 2022

Abstract

This paper leverages data from 925 individual Tinder users to cluster profiles of user behavior. K-means clustering was performed using features describing conversation metadata, user activity, and user metadata. Three clusters were compared on user activity and demographic makeup, revealing that men and women have substantially different behavior on Tinder. In addition, clusters demonstrated different levels of selectivity and took orthogonal approaches to find success on the app's dating market.

Introduction

Dating is the focus of an immense amount of cultural energy and interest, yet its mechanics are poorly understood and are currently undergoing a tremendous change in the digital revolution. 39% of partnered heterosexual Americans met their partners online as of 2019, up from 22% in 2009[1]. As dating moves online and is increasingly mediated by software that collects enormous amounts of data, can researchers utilize that data to uncover patterns in our collective dating behavior? Data that describes how people honestly acted represents a significant improvement over the results of self-reported surveys, and this paper attempts to capitalize on new data availability to generate knowledge on an impactful topic.

Using Tinder data submitted by users of the app, this paper clusters users by their behavior and compares the makeup and activity of each cluster. Clustering users by behavior is a common approach to user experience research for software companies because it allows them to understand several things about their client base, including who they are, what they engage within the software, and how they relate to other users[2]. In the case of network apps like Tinder, understanding these factors is key to the app's success. The clusters in this project can provide helpful information to social media industry professionals and social scientists alike.

In this project, I find that the primary separator between clusters is gender, i.e., men and women behave distinctly differently on Tinder. Women are much more discerning in who they "like" on the platform and tend to end conversations more frequently than men. Men vary more in behavior than women, with some men taking the "cast the net wide" approach of connecting

---

[1] Michael J. Rosenfeld, Reuben J. Thomas, and Sonia Hausen, "Disintermediating Your Friends: How Online Dating in the United States Displaces Other Ways of Meeting," *Proceedings of the National Academy of Sciences* 116, no. 36 (2019): pp. 17753-17758, https://doi.org/10.1073/pnas.1908630116.

[2] Carl Yang et al., "I Know You'll Be Back," *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery &Amp; Data Mining*, 2018, https://doi.org/10.1145/3219819.3219821.

with as many people as possible. In contrast, others are closer to the average woman's level of selectivity. Clusters demonstrate different patterns of opening the app and different levels of engagement with the app, and men are significantly overrepresented in the population of Tinder users. The information gleaned in this project highlights Tinder's challenge to attract more women to their app and design their features to appeal to women's desired dating approaches.

Problem Definition

Segmenting Tinder users is a problem of unsupervised learning, where many features describe each user, yet there is no "target" feature or initial label assigned to them. We use machine learning algorithms to uncover similarities in how users are described by their features and find further similarities until we have an interpretable number of "clusters.[3]" Each cluster represents a group of users who are sufficiently similar. Looking at one user solely in their group context should provide a reasonably good amount of information compared to individual-level analysis. Given that Tinder has 75 million monthly users, stakeholders need to have a way to reduce complexity in the user set, which clustering can provide[4].

For the platform's developers, cluster-level analysis could reduce the difficulty of designing new features and provide predefined experimental groups to test user interface or app functionality changes. Machine learning engineers also might be interested in cluster makeup to tune their algorithms that match users to each other. Clustering can also benefit Tinder's users, who could use their knowledge of the app's clusters to adjust their usage strategy to maximize their desired return, whether that be maximizing their number of matches, conversations, or some

[3] Jake VanderPlas, *Python Data Science Handbook: Essential Tools for Working with Data* (S.l.: O'REILLY MEDIA, 2023).
[4] Mansoor Iqbal, "Tinder Revenue and Usage Statistics (2022)," Business of Apps, September 6, 2022, https://www.businessofapps.com/data/tinder-statistics/.

other metric. Finally, clustering is of interest to social scientists, who are increasingly performing quantitative research on previously unquantifiable topics, like interpersonal relationships. Clustering compares group similarities and differences, a key concept for social researchers.

Models & Algorithms

This paper's methods are an exploration of unsupervised learning algorithms, including dimensionality reduction techniques and clustering approaches. I implemented principal component analysis (PCA) in my preprocessing stage for feature dimension reduction and noise reduction. PCA is a good strategy for reducing the number of features in this project because it is data-driven and objective, which is useful in a situation where there is little prior knowledge of the underlying data[5]. PCA finds linear combinations of the features that preserve the maximum amount of variance in the data, meaning the machine learning engineer has the same amount of information in a smaller package. PCA is especially beneficial as a preprocessing step in clustering projects because it prioritizes components with higher variance, which means they would be unlikely to have been victims of noise in the first place. Therefore, PCA effectively tosses out noise and keeps the essential signals in the data.

Clustering is the primary topic of this paper, and I carry it out in two ways. First, I implemented agglomerative hierarchical clustering to get a ballpark idea of the optimal number of clusters present in the data. Because the data is novel and because methods of finding the number of clusters are imprecise, I decided that getting a starting estimate from hierarchical clustering would improve the precision of the rest of the process. Following the hierarchical clustering, I implemented several instances of k-means clustering while varying *k* around the ballpark estimate established by hierarchical clustering and the "elbow method" for selecting *k*.

[5] VanderPlas. *Python Data Science Handbook.*

To evaluate each model, I calculated several external evaluation criteria, including inertias, silhouette scores, Calinski-Harabasz scores, and Davies-Bouldin scores. Post-evaluation, the best-performing model's labels were analyzed with the original dataset.

Data & Implementation

Data for this project was sourced from Swipestats.io, a database maintained by Norwegian developer and data scientist Kristian Bø[6]. Tinder users have a right to request their user data profile from Tinder, which will provide that user data in a JSON file promptly. Thousands of users have then submitted their data to Bø, who anonymizes it and maintains an organized database of every user's submitted data. Bø provides random samples of the dataset to parties who request to use it for academic purposes, and he provided me with data from 1209 profiles.

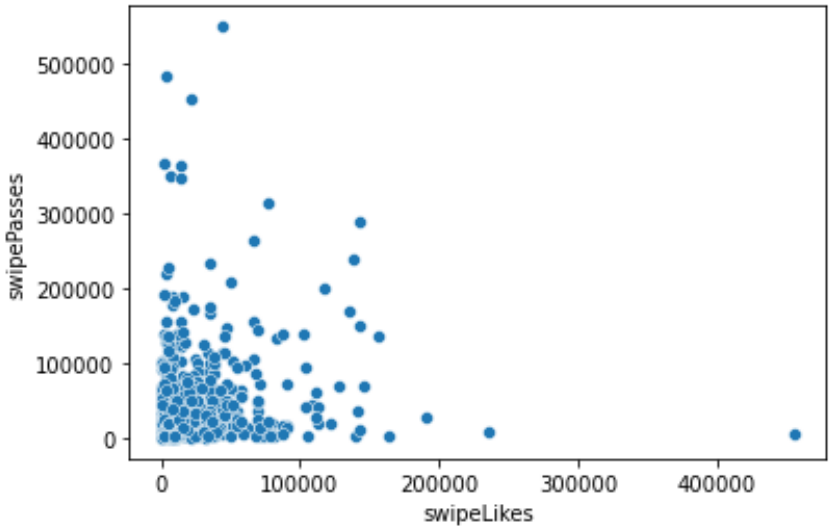
For each profile, I had access to the information contained in the user's public-facing profile page. I also had a record of every conversation they ever had on the app and their total number of app opens, swipes, and matches, broken out by day. The entire dataset was over 500 MB, so I read each profile into Python iteratively until I could remove the text conversation data and improve the working size of the data.

As with any data science project, the most significant use of my time was data processing. I explored some initial scatter plots between features, one of which appears below in Figure 1. These scatterplots enabled me to spot outliers like the data point that describes a profile with over 400,000 "likes" and almost no "passes." Bots are a significant problem on dating apps, and many of the outlying points in that plot are likely bots, so I removed them from the dataset. I

[6] Kristian Bø, "Visualize Your Tinder Data," (2022) https://www.swipestats.io/.

also removed 300 profiles whose app usage data was missing, as usage was of key interest in this analysis. In all, the final number of profiles analyzed was reduced from 1209 to 925.

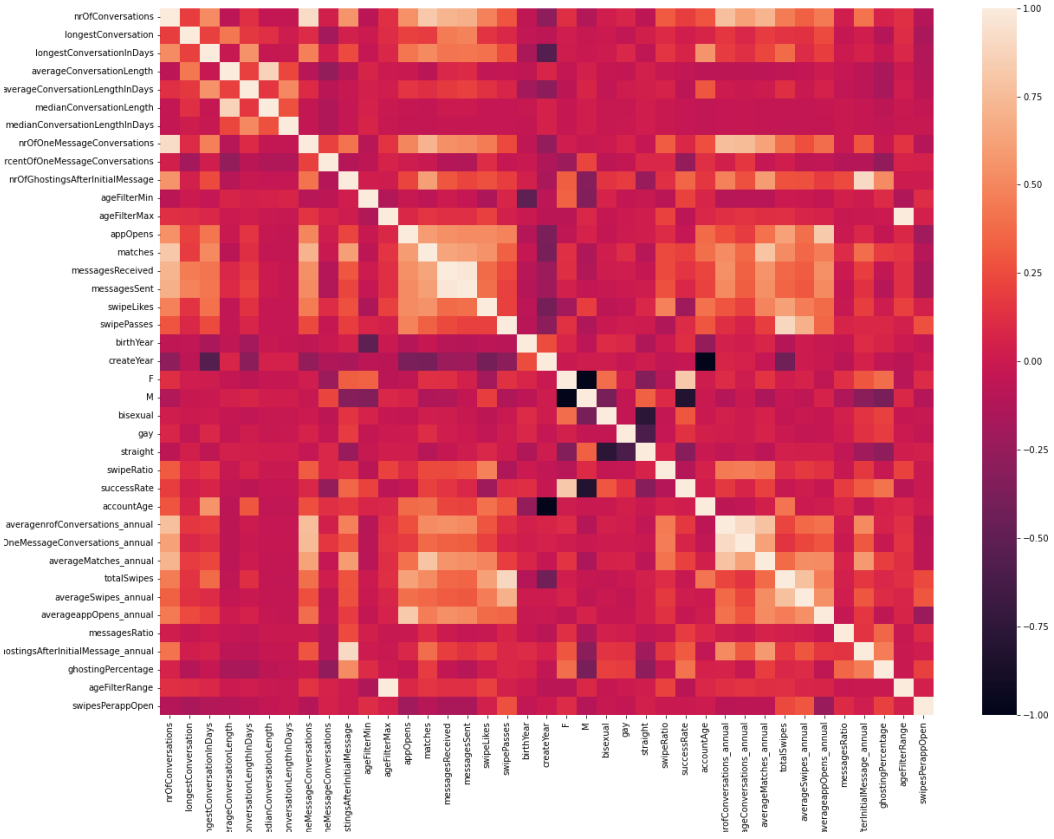
Figure 1: Scatter Plot of Users' Likes vs. Passes Behaviors



Though this dataset came with many initial features, not all these features came in a state that made sense for clustering, so I did a substantial amount of feature engineering. First, categorical variables needed to be encoded numerically, and DateTime variables also needed to be converted. I then standardized variables like "Number of Conversations" by averaging them by account age so that accounts of different ages would be more comparable. I also constructed features I heuristically understood to be vital to understanding a user's behavior. For example, while the original dataset contains how often users like and pass, I added a feature with a ratio of likes to passes as a metric of a user's selectivity. Another example of this was creating a feature measuring the percentage of conversations that a user "ghosted." In other words, how many times have they never responded to a match's last message?

After feature engineering, the data had over 40 features, many of which could have contained redundant information. As a result, I analyzed the correlation matrix heat map in Figure 2 and decided to implement principal component analysis.

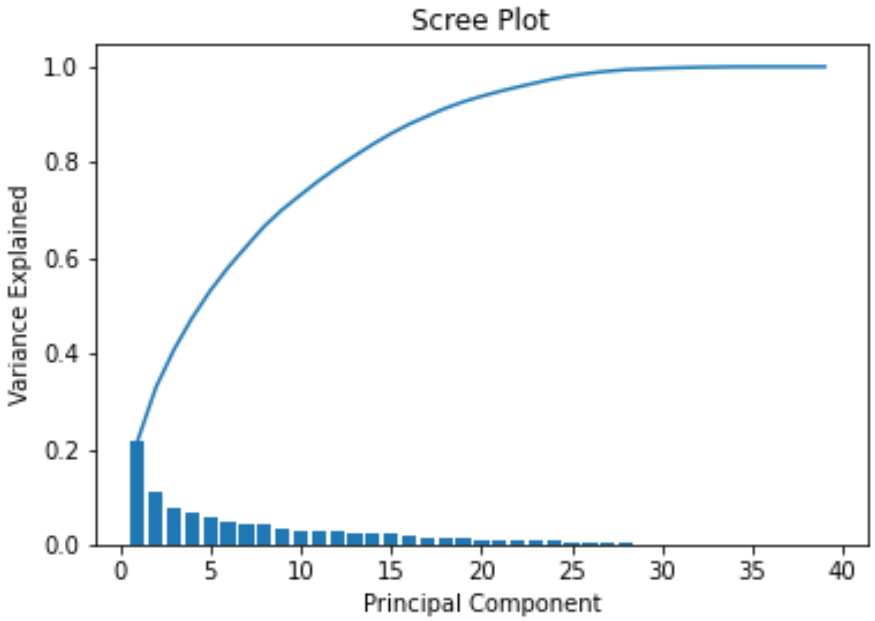
Figure 2: Correlation Matrix Heat Map of All Features



This heat map reveals some pockets of collinearity between metrics of user activity.

To implement PCA, I first standardized the data using sci-kit-learn's Standard Scaler. If I had not scaled the data, features with naturally large variances, like app opens, would have biased the selection of principal components. The Scree plot in Figure 3 visualizes the algorithm's ability to reduce the number of features while maintaining a large portion of the variance.

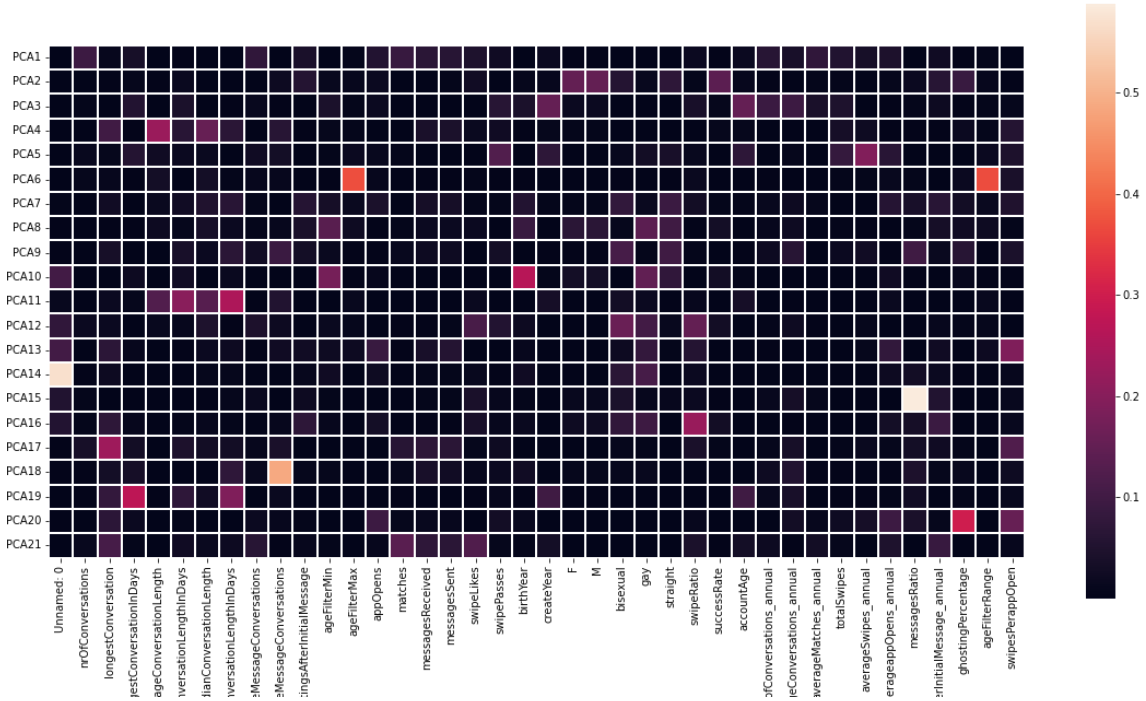
Figure 3: Scree Plot for Principal Component Analysis



95% of the dataset's variance can be maintained by using only 21 of the principal components.

Figure 4 displays a correlation matrix heat map that explains the makeup of each principal component. For each principal component, a correlation with an original feature means that that principal component is at least in part a linear combination of that original feature.

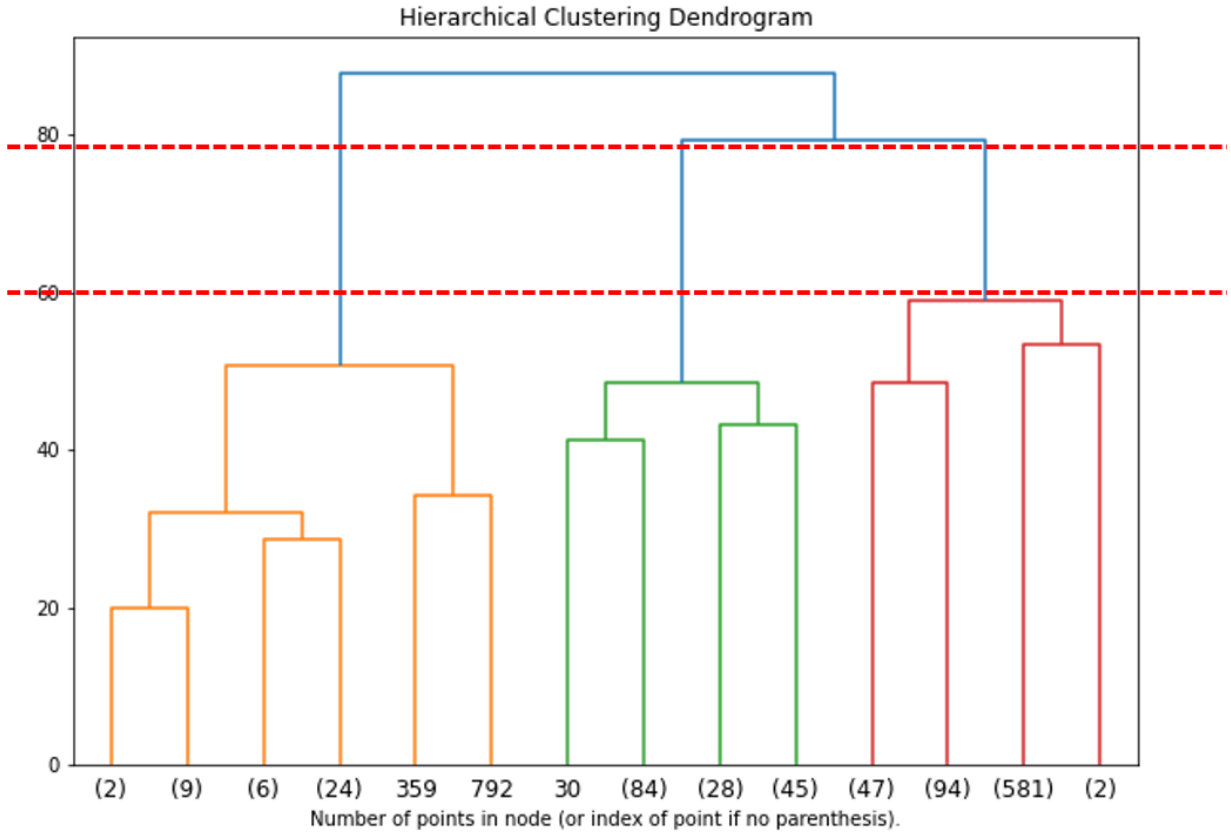
Figure 4: Correlation Matrix Heat Map of Principal Component Makeup



Message Ratio, Number of Ghostings, and Age Filter Range appear to be particularly dominant sources of variance.

Following my PCA application and the dataset's reduction to 21 features, I moved on to executing my clustering methodology. First, I grew a full hierarchical clustering tree using agglomerative clustering, the results of which are illustrated in the dendrogram below.

Figure 5: Hierarchical Clustering Dendrogram for Tinder User Segmentation

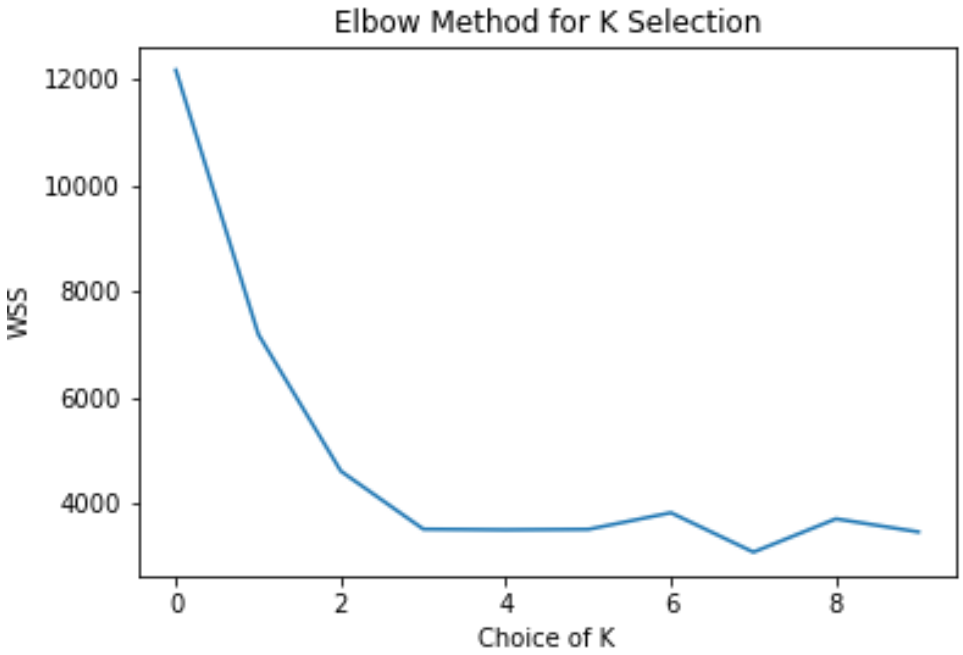


By drawing lines through the largest vertical distance in the dendrogram, I determined the optimal number of clusters was 3.

Hierarchical clustering implied that three clusters were a good number of clusters for this dataset, but I also tried the "elbow method" of finding the optimal number of clusters. The elbow method charts how total intra-cluster distance changes over different values of *k*. It is called the elbow method because it recommends choosing *k* at the chart's "elbow" or where additional clusters provide no further improvements in performance[7]. The elbow chart for this dataset is in Figure 6 below.

[7] Pratap Dangeti, *Statistics for Machine Learning* (Birmingham: Packt Publishing, 2017).

Figure 6: Elbow Method Chart for Selecting *k*



The "elbow" for this dataset appears where $k = 3$.

Given that hierarchical clustering and the elbow method both suggest choosing *k* as three, I took that information and decided to test out four iterations of k-means clustering with *k* as two, three, four, and five. Because there is no ground truth labeling available for this data, I needed to use evaluation metrics that were only functions of the data and the algorithm results themselves[8]. Therefore, I evaluated each level of *k* by the model's inertia, silhouette score, Calinski-Harabasz score, Davies-Bouldin score, and the number of iterations each model took to converge. I also tested two methods for initializing the centers of each cluster, random and k-means ++, which uses probability distributions to select a center. Figures 7 through 11 are a comparison of each evaluation metric for each model.

[8] Yufeng, "Three Performance Evaluation Metrics of Clustering When Ground Truth Labels Are Not Available," Medium (Towards Data Science, June 23, 2022), https://towardsdatascience.com/three-performance-evaluation-metrics-of-clustering-when-ground-truth-labels-are-not-available-ee08cb3ff4fb.

Figure 7: Inertia by Number of Clusters and Initialization Strategy

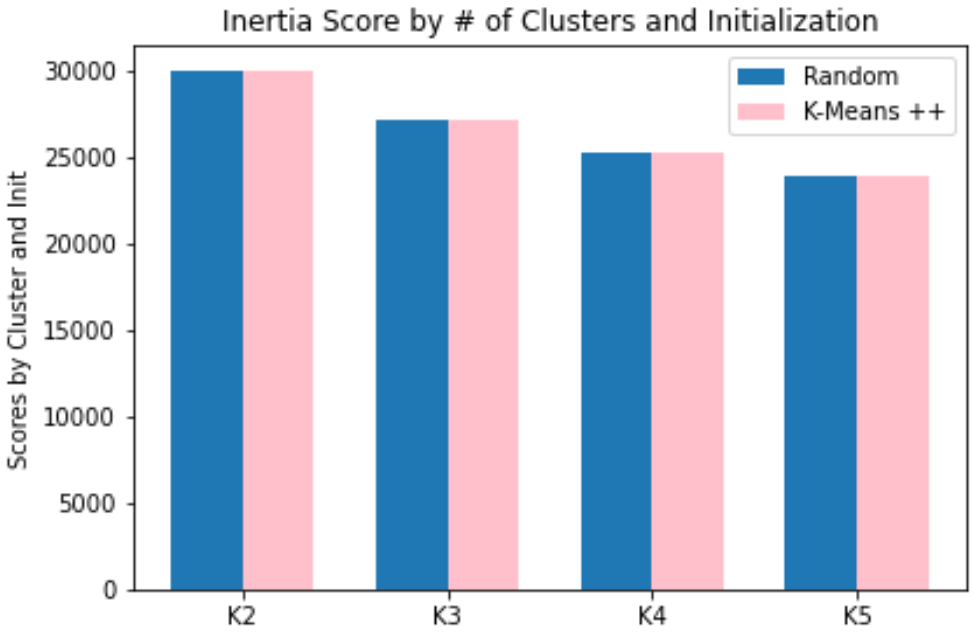


Figure 8: Silhouette Score by Number of Clusters and Initialization Strategy

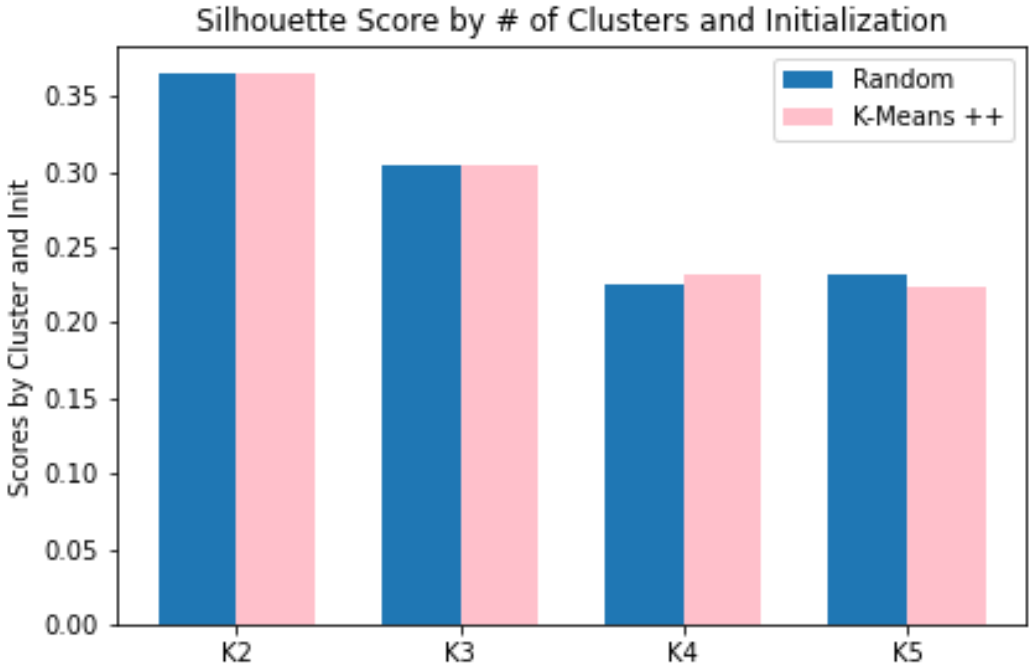


Figure 9: Calinski-Harabasz Score by Number of Clusters and Initialization Strategy

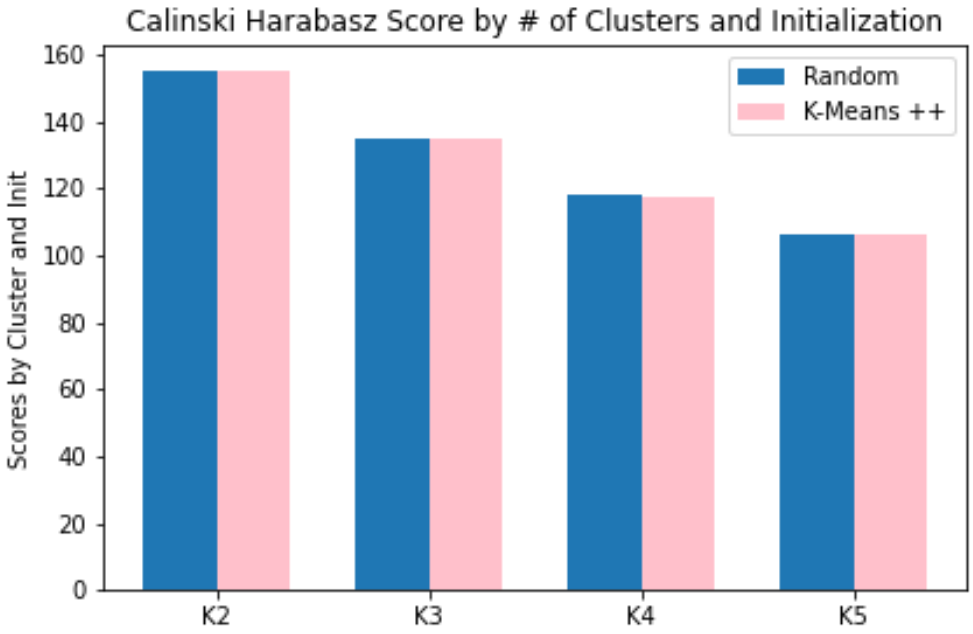


Figure 10: Davies-Bouldin Score by Number of Clusters and Initialization Strategy

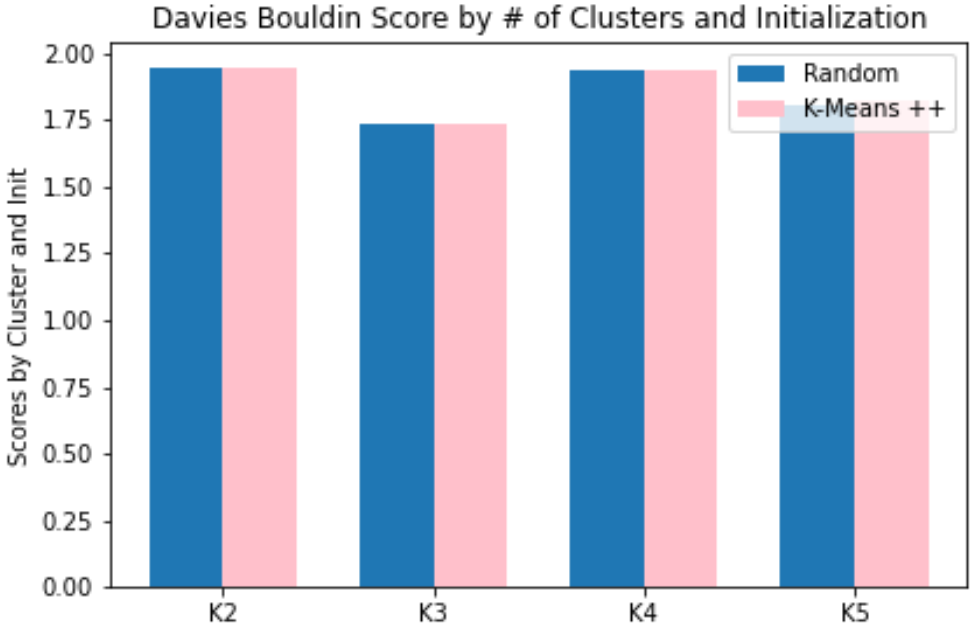
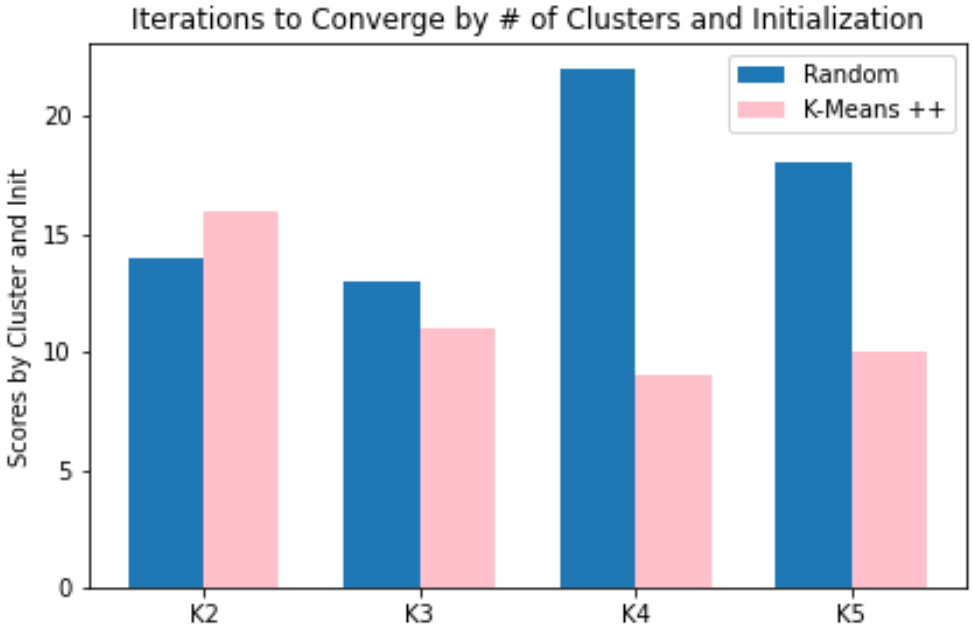


Figure 11: Iterations Until Convergence by Number of Clusters and Initialization Strategy



Each of these scores (except the number of iterations) represents how well the clustering algorithm could create distinct clusters that are very different from each other and very similar within themselves, given the number of clusters and initialization they are constrained to. Inertia is the summed squared distance between each data point and its centroid, meaning a well-defined and dense cluster would have low inertia. Therefore, we prefer a model with less inertia. The silhouette coefficient describes the relationship between inter- and intra- cluster distances and is equal to 1 when intra-cluster distance is minimized and inter-cluster space is maximized. -1 is the worst possible silhouette coefficient. We prefer a higher silhouette coefficient. The Calinski-Harabasz score is a ratio between the summed squared inter-cluster distance and the summed squared intra-cluster distance, adjusted by the degrees of freedom available. We prefer a higher Calinski-Harabasz score, which implies more separation. The Davies-Bouldin score measures

how similar the algorithm-defined clusters are, so we prefer a low Davies-Bouldin score to say that the clusters are sufficiently separate.

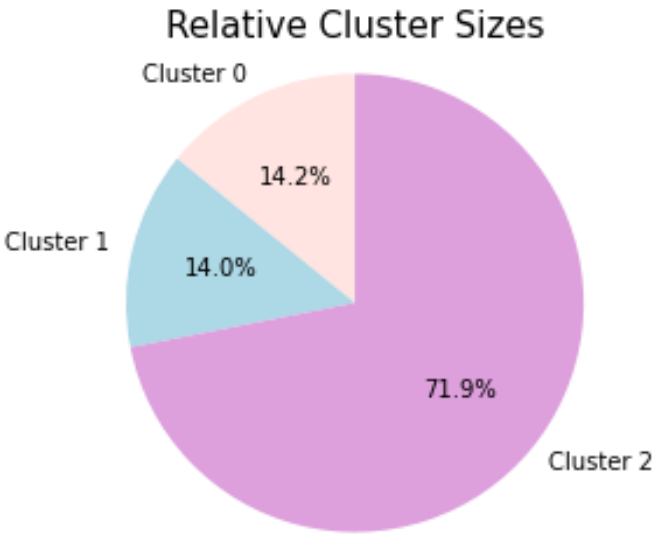
Picking k is notoriously tricky, and this project is no exception to that rule. While the hierarchical clustering result, elbow method, and Davies-Bouldin scores support a case for $k = 3$, the Calinski-Harabasz scores and silhouette coefficients support a case for $k = 2$. While convergence iterations are not essential to model evaluation, they are factors for efficiency concerns. Therefore, I decided to choose $k = 3$ as the optimal model. When $k = 3$, there were no significant differences in performance between random center initialization and k-means ++ initialization, so I left initialization as sci-kit-learn's default, which is random.

Given the lack of ground truth data available, it is difficult to objectively assess how accurate or precise this k-means clustering was. In the following results section, I analyze how user behavior differs by the cluster labels and create profiles for each cluster to describe the differences in their user behaviors.

Results & Discussion

Of prime interest to this project is finding out which users are in which clusters. Because we have access to user profile data, we have some demographic information about the clusters. As displayed in Figure 12, the clustering algorithm sorted Tinder users into three clusters, with a dominant cluster representing over two-thirds of Tinder users.

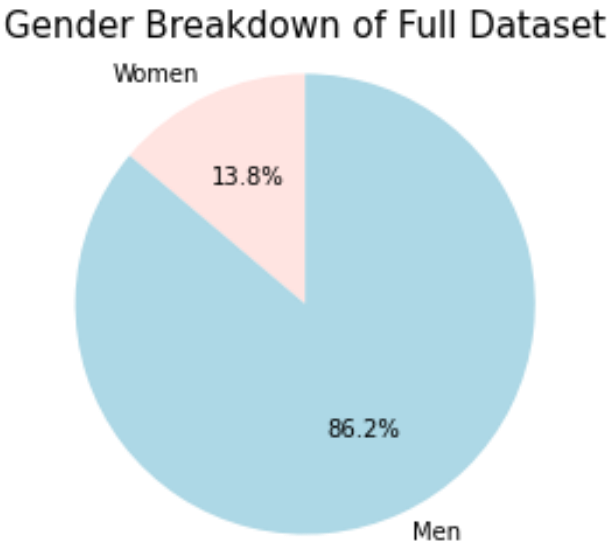
Figure 12: Relative Cluster Sizes by Percentage of Original Dataset



Though one would typically not prefer clusters to be of such different sizes, the constrained k-means algorithm to eliminate that problem is outside the scope of this paper[9]. Figure 13 shows the gender breakdown of the full dataset, while Figure 14 displays the distribution of user ages by each cluster.

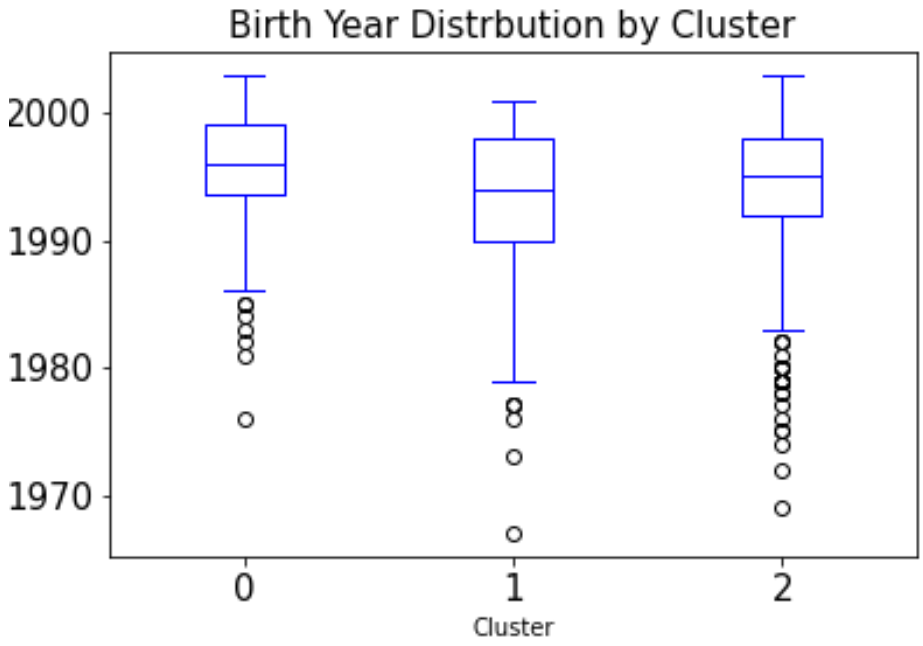
[9] Paul Bradley, Kristin Bennett, and Indrajit Bhattacharya, "Using Assignment Constraints to Avoid Empty Clusters in k-Means Clustering," *Chapman &Amp; Hall/CRC Data Mining and Knowledge Discovery Series*, 2008, pp. 201-220, https://doi.org/10.1201/9781584889977.ch9.

Figure 13: Gender Breakdown of Full Dataset



Compared to the general Tinder user ratio of 3:1 men:women, men are slightly overrepresented in this data.

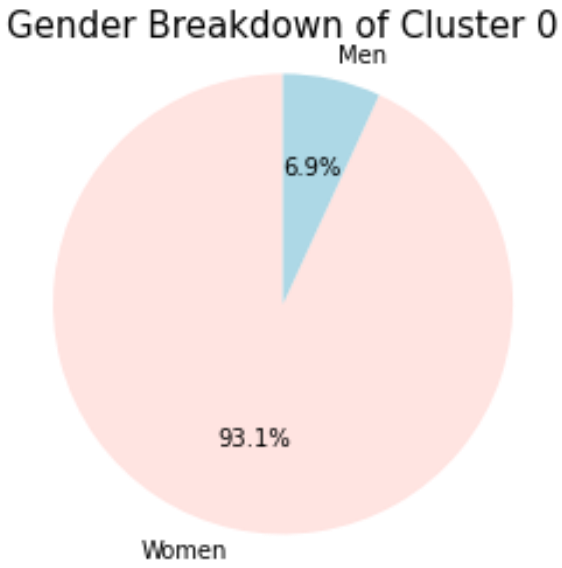
Figure 14: Distribution of Birth Years by Clusters



Cluster 0 is the youngest cluster, and Cluster 1 is the oldest on average, though Cluster 2 contains more older people. Generally, the average Tinder user in each cluster is in their mid-to-late 20s.

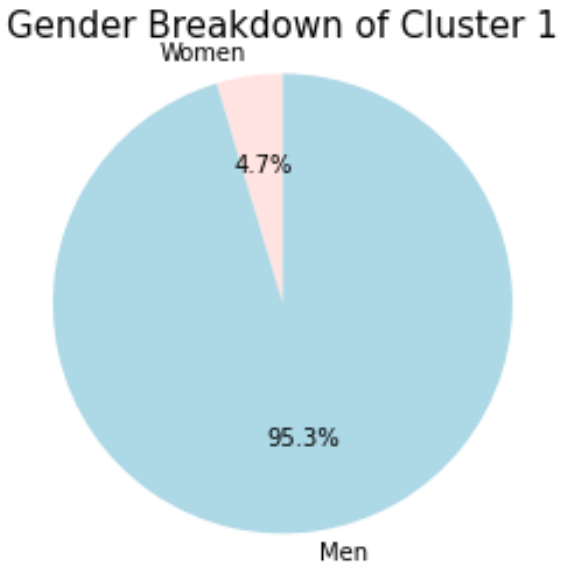
Within each cluster, user gender and sexuality demographics are quite homogeneous. Figures 15 through 17 visualize the gender makeup of each cluster, while Figure 18 shows the sexuality breakdown for Cluster 0. Clusters 1 and 2 did not receive sexuality pie charts because virtually all cluster members are straight.

Figure 15: Gender Breakdown of Cluster 0



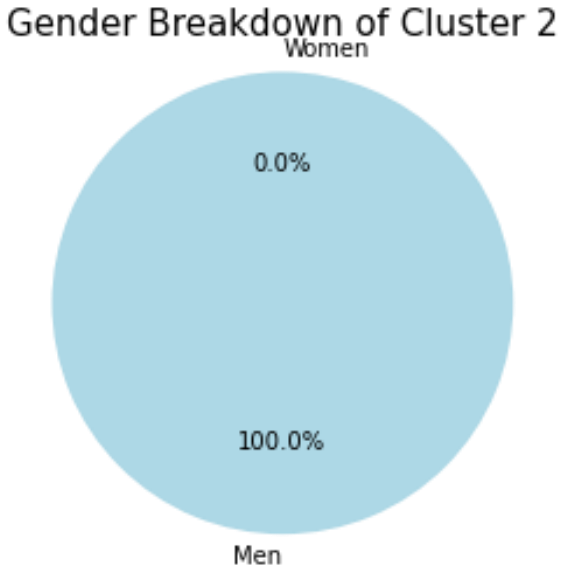
Cluster 0 is dominated by women, while the few men in this cluster are overwhelmingly LGBTQ+.

Figure 16: Gender Breakdown of Cluster 1



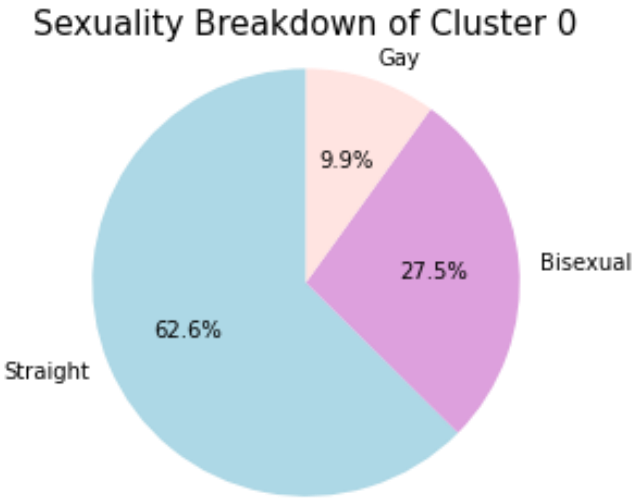
Cluster 1 is almost entirely men and almost entirely heterosexual people.

Figure 17: Gender Breakdown of Cluster 2



Cluster 2 is entirely made up of men and fewer than 2% of the cluster members are LGBTQ+.

Figure 18: Sexuality Breakdown of Cluster 0



Cluster 0 is the only cluster to exhibit diversity in sexuality, where over a third of the cluster members are not heterosexual. Though straight women still form the largest subgroup within Cluster 0, LGBTQ+ women and men are also very present. Cluster 0 follows a documented demographic pattern colloquially called "the girls and the gays," but that is also described in the literature, in which observers have noted the cultural affinity group that has emerged between women and gay men[10]. We can therefore understand that this cluster is a coherent, previously identified group.

Clusters 1 and 2 are not different at a demographic level, as they share similar gender ratios, sexualities, and age distributions. Therefore, we can expect that their group behavior is significantly different. In fact, there is a significant difference in the data between how Clusters 0, 1, and 2 approach their use of the Tinder app.

[10] Mairead Moloney and Tony Love, "#Thefappening: Virtual Manhood Acts in (Homo)Social Media," 2018, https://doi.org/10.31235/osf.io/tqsf7.

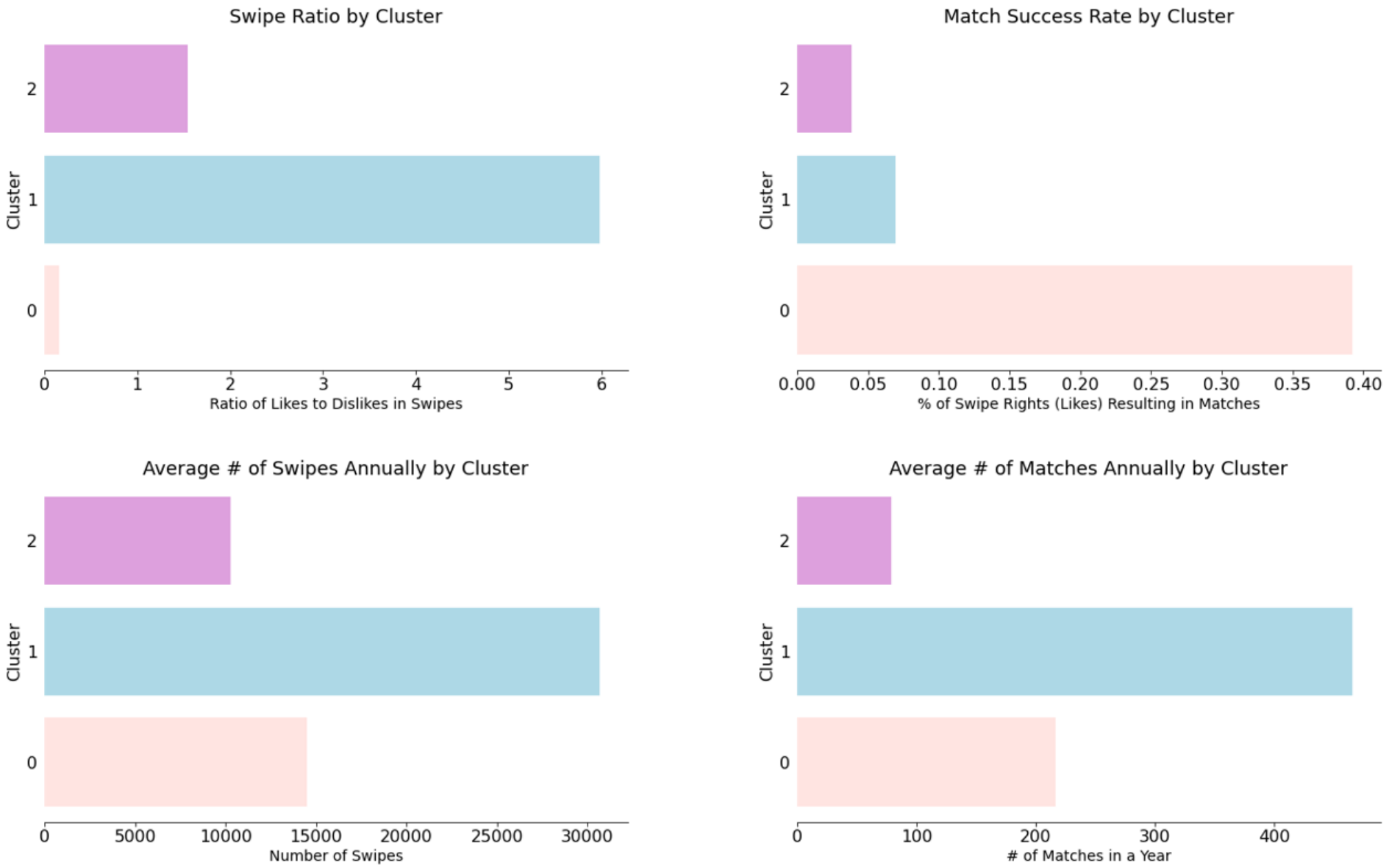


Figure 19: Swiping & Matching Behavior Differences by Cluster

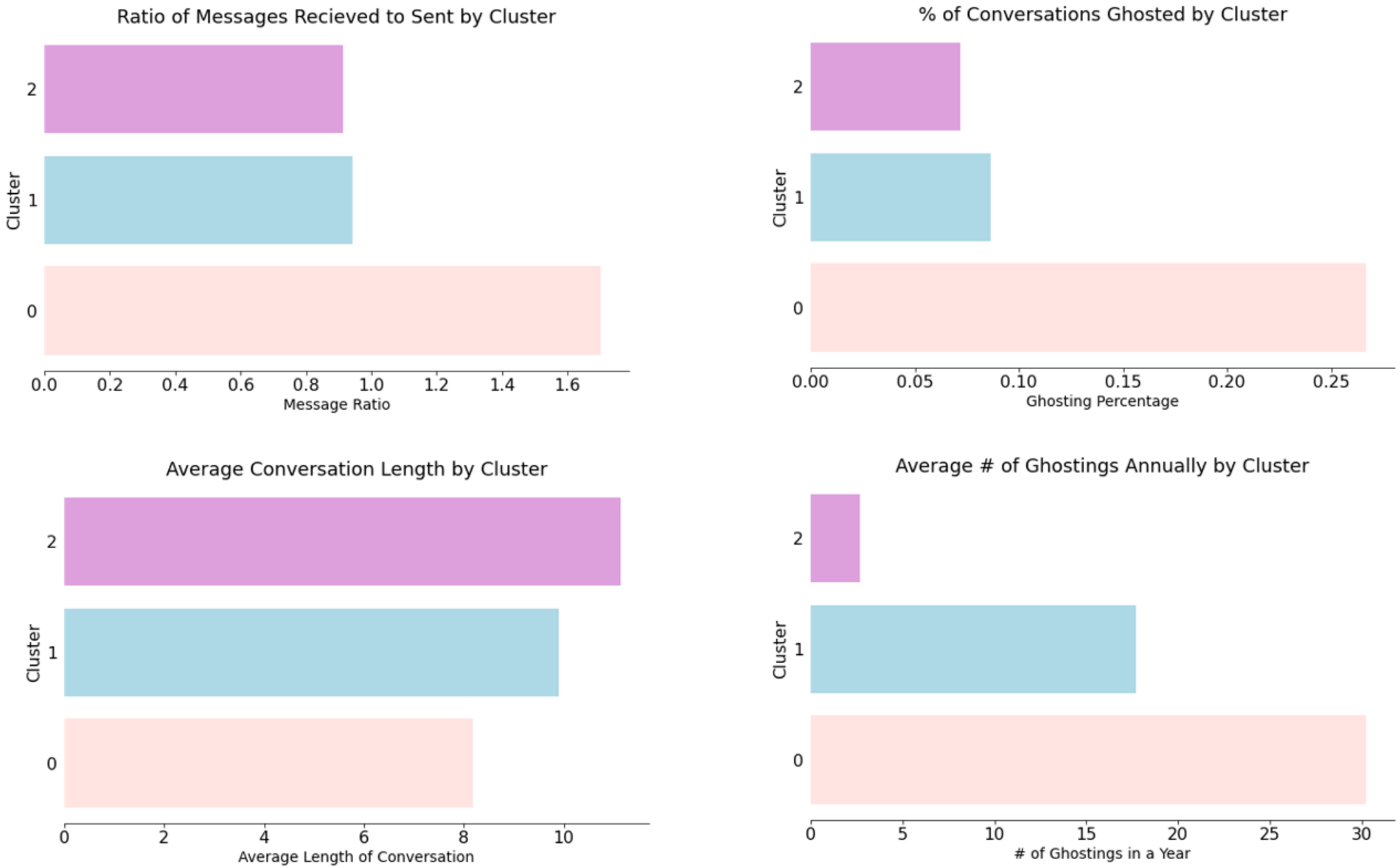


Figure 20: Conversation Behavior Differences by Cluster

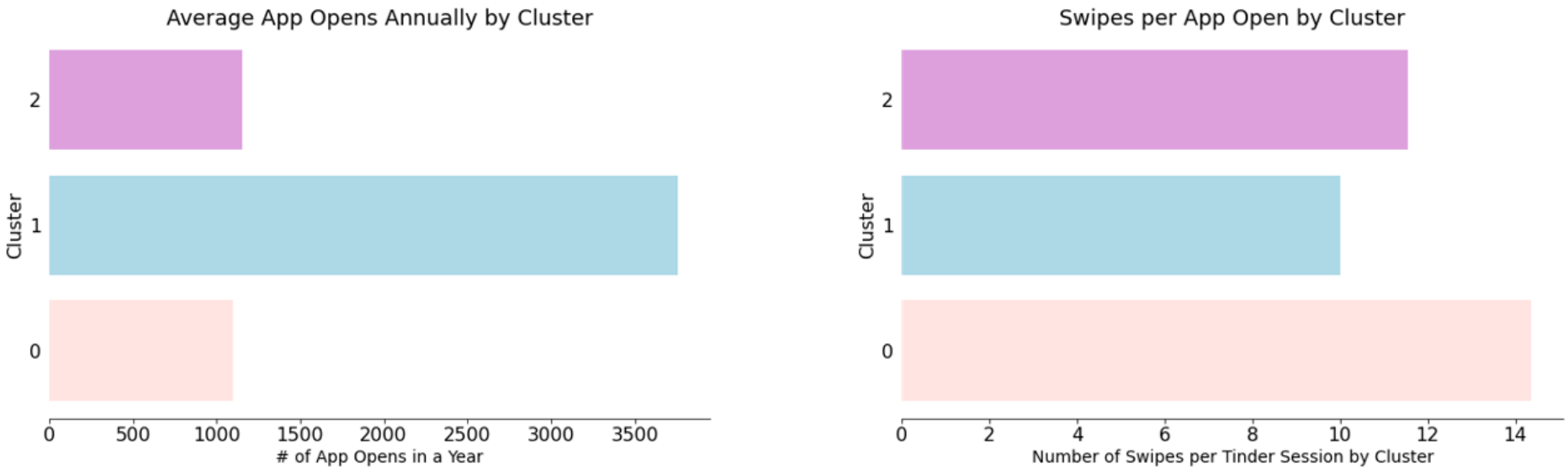


Figure 21: App Open Behavior Differences by Cluster

Composite Figures 19, 20, and 21 above summarize the behavior comparisons between each cluster. Figure 19 describes swiping and matching behaviors, where we notice an immediate difference between Cluster 1 and the others. Cluster 1 users like a disproportionate number of people compared to the number of people they pass on, and overall, they swipe a lot more than either of the other clusters. Though their match success rate is significantly lower than the success rate of Cluster 0, the sheer amount of swiping they do means that Cluster 1 users receive far more matches in a year than other clusters. Cluster 0 users swipe a moderate amount but are highly discerning in who they decide to like. They are very likely to be liked compared to Cluster 1 and 2, so they maintain a high success rate and a high number of matches, despite (or perhaps because of) their level of discernment. Within swiping behavior features, Cluster 2 tends to have moderate input and poor results. They are more discerning than Cluster 1, yet less than Cluster 0. They do the least swiping and have the lowest success rate and by far the lowest annual number of matches.

Within conversation behavior, Cluster 0 is the most unique cluster. Cluster 0 users receive more messages than they send, while Clusters 1 and 2 send more messages than they receive. Cluster 0 also seems to display a level of discernment in continuing conversations. In both measures of absolute “ghostings” (where a user chooses not to respond to a first message) and by a percentage of total conversations, Cluster 0 users dominate the other clusters. As a result, Cluster 0 users tend to have slightly shorter average conversations than Cluster 1 and 2 users. Clusters 1 and 2 are not as different in conversation measures as in swiping measures, but Cluster 2 users tend to have more extended conversations and do less ghosting overall than Cluster 1 users. Finally, we can analyze the differences between clusters by looking at their app-opening behaviors.

Cluster 1 users continue their theme of the most activity overall on the app with app opens, where they tend to open Tinder significantly more than the other clusters on average. The average Cluster 1 user opens the Tinder app nearly ten times daily. Their large number of daily opens also means that they swipe on fewer people every time they open the app than the other clusters. Conversely, Cluster 2 users swipe more times in each app session but open the app fewer times than the other clusters. Given the information about each cluster's demographics and behavior, how can the clusters be described in coherent profiles?

Cluster 0's profile is the easiest to describe, and I give it the name "Girls and Gays" (matching its demographic description). The Girls and Gays only like a small portion of people they come across on the app, and they only respond to messages from about 75% of the people they match with anyway. They receive more messages than they reply to by a factor of about 1.75. They expect that about 40% of people they like will like them back, which is much better than Cluster 1 and 2. They do not open the app as frequently as other clusters, but when they open it, they swipe a few more times than other clusters. Still, their total number of swipes annually is much lower than Cluster 1 and only slightly higher than Cluster 2. Though the Girls and Gays have more people showing interest in them on Tinder, their low levels of liking and lower response rates could mean that they are disappointed in the quality of other single people on Tinder.

Cluster 1's profile is defined by usage intensity. I call Cluster 1 the "Super Users." This group of mostly men is very engaged with the app and has the highest rates of swipes, matches, and app opens. They like six times more people than they dislike, three times higher than the next closest cluster. Though they are not particular about who they like, they do end up ghosting more people than the users of Cluster 2, though still far less than Cluster 0. They are similar to

Cluster 2 in the ratio of messages received to sent, the percentage of conversations ghosted, and the average conversation length. The Super Users cast their nets wide in the dating market, and they have some success getting more matches this way. However, it is unknown whether Super Users convert those matches into dates or relationships.

Cluster 2 seems to be the most nuanced. The entire cluster is men, with 98% being straight men. They have the widest range of possible ages and are by far the largest cluster. They are the least active on Tinder, with the fewest number of swipes and matches, while their app opens are about even with Cluster 0. Unfortunately, Cluster 2 has the lowest match success rate and the lowest received-to-sent messages ratio. Cluster 2 users are the least likely to ghost people and carry on the longest conversations. For those reasons, I have termed the Cluster 2 users "Average Joes." Average Joes prefer the quality of matches over quantity, and though they do not receive many matches, they engage deeply with the matches they receive and rarely ghost.

Most Tinder users fall into this category, though I am sure Tinder would prefer that more Average Joes were converted to Super Users. Tinder would also likely prefer more women join the app to reduce the high level of competition present in their dating market. It seems as if the average behaviors across each feature for the clusters are consistent and follow closely from one another. This consistency allowed me to define clusters that I think are coherent and meaningful and provide more information about the different types of Tinder users. Women seem to display homogeneity in their behavior on Tinder, while men diverge by their intensity and engagement with the app. In future work, adding an analysis of conversation text could be very informative as to the quality of conversations by cluster and external outcomes like dates or relationships.

Related Work

User clustering efforts on social media apps are common, but not as much work has been done to cluster users of closed-network apps like Tinder. A few types of papers inspired this project, including articles generally on social media and some about Tinder.

In "A Hierarchical Clustering Algorithm for Characterizing Social Media Users", the authors demonstrate clustering based on user behavior where a prior number of clusters is unknown[11]. The authors validate their clusters using external data, which was unavailable for the Tinder context. The authors of "Social media analysis using optimized K-Means clustering" present an improvement on standard k-means clustering by optimizing centroid initialization with the Genetic algorithm[12]. I compared initialization methods in my models because of this paper, but I found no difference in performance by initialization methods.

I also reviewed a paper that clustered Tinder users by their profile biographies instead of their user behavior[13]. The authors compared natural language processing methods that were tasked with determining the emotions that biographies were associated with. For the papers that I reviewed that clustered Tinder users more on their behaviors, the clustering example data came from surveys instead of app use data[14]. Interestingly, a paper on addictive usage of Tinder found that men were more likely to be in clusters associated with intense, problematic usage of the

---

[11] Priyanka Sinha et al., "A Hierarchical Clustering Algorithm for Characterizing Social Media Users," *Companion Proceedings of the Web Conference 2020*, 2020, https://doi.org/10.1145/3366424.3383296.
[12] K. Madhuri and Mr. K. Rao, "Social Media Analysis Using Optimized K-Means Clustering," *International Journal of Trend in Scientific Research and Development* Volume-3, no. Issue-2 (2019): pp. 953-957, https://doi.org/10.31142/ijtsrd21558.
[13] Esperanza Johnson et al., "'Matching Learning': Profiling and Clustering Users on Tinder Based on Emotion and Sentiment Analysis," *Proceedings of the International Conference on Ubiquitous Computing &Amp; Ambient Intelligence (UCAmI 2022)*, 2022, pp. 876-887, https://doi.org/10.1007/978-3-031-21333-5_87.
[14] Alexandru Mateizer and Eugen Avram, "Mobile Dating Applications and the Sexual Self: A Cluster Analysis of Users' Characteristics," *International Journal of Environmental Research and Public Health* 19, no. 3 (2022): p. 1535, https://doi.org/10.3390/ijerph19031535.

app[15]. The novel, direct data utilized in this paper means it has created a niche in the research of user behavior on dating apps and has created new knowledge in the field.

Conclusions

In this project, I aimed to explore a unique dataset and generate new knowledge about how people interact with each other on dating apps. When social science researchers study dating apps, they often base their conclusions on how study participants self-report their actions, while this paper utilizes users' actual behavior. This new data provides another level of confidence in our analysis of what occurs between individuals on dating apps. The large amount of numerical data available allowed me to implement an unsupervised clustering algorithm to separate users by their behavior beyond standard categories like gender and age.

I observed that there are three distinct clusters of Tinder users, which I classified as the following: "Girls and Gays," "Super Users," and "Average Joes." Girls and Gays are discerning in who they like on Tinder and yet still receive many matches, and they are not afraid to end conversations that they are not interested in. Super Users attempt to maximize their return by liking many other users and being very engaged with Tinder; these users have the highest number of matches. Finally, Average Joes are the largest cluster, and they tend to engage with the quality of their matches by rarely ghosting, yet they have the lowest number of matches.

However, the external performance of this model is difficult to validate, given that we have no ground truth information available for these subjective categories. Future researchers could focus on integrating text data to understand conversation quality by cluster better. They could also attempt to label users' data to provide ground truth and produce better models.

---

[15] Lucien Rochat et al., "The Psychology of 'Swiping': A Cluster Analysis of the Mobile Dating App Tinder," *Journal of Behavioral Addictions* 8, no. 4 (2019): pp. 804-813, https://doi.org/10.1556/2006.8.2019.58.

Bibliography

Bradley, Paul, Kristin Bennett, and Indrajit Bhattacharya. "Using Assignment Constraints to Avoid Empty Clusters in k-Means Clustering." *Chapman & Hall/CRC Data Mining and Knowledge Discovery Series*, 2008, 201–20. https://doi.org/10.1201/9781584889977.ch9.

Bø, Kristian. "Visualize Your Tinder Data." Swipestats. Accessed December 13, 2022. https://www.swipestats.io/.

Dangeti, Pratap. *Statistics for Machine Learning*. Birmingham: Packt Publishing, 2017.

Iqbal, Mansoor. "Tinder Revenue and Usage Statistics (2022)." Business of Apps, September 6, 2022. https://www.businessofapps.com/data/tinder-statistics/.

Johnson, Esperanza, Alfonso Barragan, Laura Villa, Jesus Fontecha, Ivan Gonzalez, and Ramon Hervas. "'Matching Learning': Profiling and Clustering Users on Tinder Based on Emotion and Sentiment Analysis." *Proceedings of the International Conference on Ubiquitous Computing & Ambient Intelligence (UCAmI 2022)*, 2022, 876–87. https://doi.org/10.1007/978-3-031-21333-5_87.

Madhuri, K., and Mr. K. Rao. "Social Media Analysis Using Optimized K-Means Clustering." *International Journal of Trend in Scientific Research and Development* Volume-3, no. Issue-2 (2019): 953–57. https://doi.org/10.31142/ijtsrd21558.

Mateizer, Alexandru, and Eugen Avram. "Mobile Dating Applications and the Sexual Self: A Cluster Analysis of Users' Characteristics." *International Journal of Environmental Research and Public Health* 19, no. 3 (2022): 1535. https://doi.org/10.3390/ijerph19031535.

Moloney, Mairead, and Tony Love. "#Thefappening: Virtual Manhood Acts in (Homo)Social Media," 2018. https://doi.org/10.31235/osf.io/tqsf7.

Rochat, Lucien, Francesco Bianchi-Demicheli, Elias Aboujaoude, and Yasser Khazaal. "The Psychology of 'Swiping': A Cluster Analysis of the Mobile Dating App Tinder." *Journal of Behavioral Addictions* 8, no. 4 (2019): 804–13. https://doi.org/10.1556/2006.8.2019.58.

Rosenfeld, Michael J., Reuben J. Thomas, and Sonia Hausen. "Disintermediating Your Friends: How Online Dating in the United States Displaces Other Ways of Meeting." *Proceedings of the National Academy of Sciences* 116, no. 36 (2019): 17753–58. https://doi.org/10.1073/pnas.1908630116.

Sinha, Priyanka, Lipika Dey, Pabitra Mitra, and Dilys Thomas. "A Hierarchical Clustering Algorithm for Characterizing Social Media Users." *Companion Proceedings of the Web Conference 2020*, 2020. https://doi.org/10.1145/3366424.3383296.

VanderPlas, Jake *Python Data Science Handbook: Essential Tools for Working with Data*. S.l.: O'REILLY MEDIA, 2023.

Yang, Carl, Xiaolin Shi, Luo Jie, and Jiawei Han. "I Know You'll Be Back." *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018. https://doi.org/10.1145/3219819.3219821.

Yufeng. "Three Performance Evaluation Metrics of Clustering When Ground Truth Labels Are Not Available." Medium. Towards Data Science, June 23, 2022. https://towardsdatascience.com/three-performance-evaluation-metrics-of-clustering-when-ground-truth-labels-are-not-available-ee08cb3ff4fb.

Appendix

The code utilized to carry out this project is housed on my GitHub at the following link:

https://github.com/emmataylor99/575datascience/blob/main/Tinder%20Profile%20Clustering.ipynb

A .csv of the data after the initial preprocessing level is also available in that repository, though the initial .json is not included because I am not sure about the data licensing permissions.