Introduction:

Greetings! I want to take a look at the interaction between mental health, more specifically mental health related to anxiety and social phobia, and video game preferences as well as other factors in a gamer profile. This would include employment, monetary gain from gaming, streaming habits, and so forth. I believe there is a significance in someone's level of anxiety and social phobias and why/what they choose to play. It's interesting because those with a similar anxiety score (derived from the GAD questionnaire) may still behave quite differently. As someone who is highly anxious, may play games to escape reality and in person interactions. So they may find comfort in online multiplayer. And another player with the same anxiety scores may absolutely despise any interaction with people and prefer single player. My hopes with this analysis are to discover underlying patterns in test scores that you can't see just looking at the numbers. Using these patterns, we can sort people into categories and see if those categories showcase unique gamer profiles. Essentially to see how significant the anxiety (GAD) and social phobia inventory (SPIN) scores are in defining an accurate gamer profile. To add more levels to this analysis, I also include satisfaction with life score (In the future, I will refer to the GAD, SPIN, and SWL scores as mental scores for short). This was to try and further differentiate people who may have similar GAD or SPIN scores. While they may be very anxious, they may have a very positive outlook and satisfaction with life. This could influence their likelihood to play for fun versus relaxing. One caveat is that while I can look at the mental scores and compare to gamer profiles. I have no way to say if mental scores influenced the gamer profile or if the gamer habits influenced a person's mental state. So for this, I am just looking at them in their current state. Not how they got to where they are. But currently; for example: a high GAD score cluster would have what kind of gamer profile in general.

The model being used for this project is clustering with further model expansions in the future. The data being used was collected by Marian Sauter and Dejan Draschkow for their own study, combining mental scores and a gamer profile. They received 13464 entries. Of these samples, I ended up using a total of 11,880 samples.

Once data was cleaned, it was prepped for clustering. I clustered only on the mental scores. Since each mental test is judged on a different scale, the values had to be scaled prior to the model. Once the cluster had been run, I reverted back to the original scale for analysis. I felt that to do well by each mental test, I needed to view it in its original scale. After cluster analysis was done on the mental scores, I moved into analyzing the gamer profiles of each cluster.

Related work:

The most related work to this project was done by Marian Sauter and Dejan Draschkow, the original collectors of this data. They used this data in their 2017 study, "Are Gamers Sad and Isolated? A database about the Anxiety, Life Satisfaction and Social Phobia of over 13000 participants." However, their focus with the data was quite different from mine. They looked at whether gaming impacted well being and psychological soundness. Not the other way around. They aimed to answer the well known question on if gaming has a negative psychological impact on users. A question posed by many of the public. I aimed to turn this on its head and see if a person's psychological state impacts their gaming habits, or is at least correlated with their habits. I want to cluster based on mental scores and see the differences in gamer profile. Which, in a way you could see if those that are more unhappy with life play more video games. But as referenced in my introduction, it would be difficult to say if the mental score impacted the habit or if the habit impacted the mental score. I am looking for correlation, not causation. The famous phrase, correlation does not equal causation, haunts my high school dreams. I would never wish to draw a false conclusion about which came first, the chicken or the egg. Instead, I want to look at what data is given, and simply examine the correlations and patterns presently.

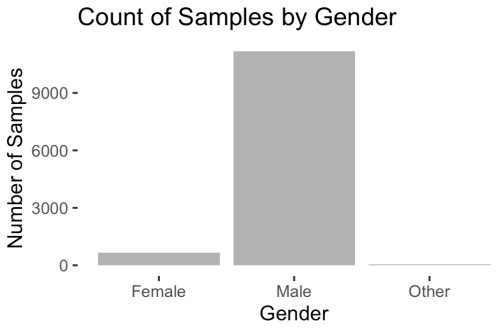
Data and Methods:

The data consisted of mental scores, gamer profiles, and demographics for each survey taker. Most survey takers found the survey via Reddit. Overall, the data started with 55 variables. 33 of those variables were mental scores, excluding the variables with total score for each mental test. The other variables were the demographics and gamer preferences. The demographics were not changed. However, the gamer profile answers required cleaning which I will go over later.

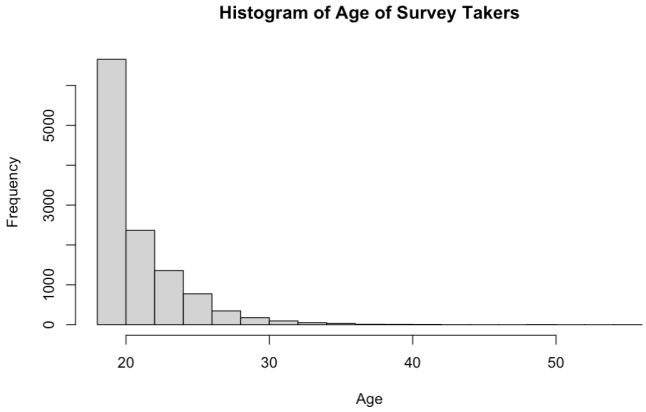
I used a total of 11,880 samples out of 13464. One reason for the change in samples was removal of NAs. I didn't want to use imputation, as I was worried the process of machine learning would defeat the purpose of clustering. As the computer is trying to decide what a value should be. This could lead to my results not being true to actual values. With how many samples the data started with, I felt ok to drop all samples with NA's. Looking at the data, there were many samples who answered hours played per week with values as far as 8000. There are not 8000 hours in a week, so I am not sure the reason for that response. Other ones around 260, and 215, also aren't possible. I chose to use samples that answered this question below 200. Upon further analysis, in the future I would use 168 as the cut off; since this is the actual amount of hours in a week, so anything above is impossible. I feel as if some people answered in minutes. As there were many samples with answers such as 120. While you are capable of playing 120 hours in a week, 2 hours may be more reasonable. However, that is not my place to judge. So no conversions were made to change possible minutes to hours. I am unable to tell a person's intention with their answer and I don't want to alter the true nature of the data.

Other data modifications were mainly just converting answers into dummy variables. This is when a categorical variable can be turned into binary variables. For example, a person's favorite color is red, orange, or blue would become 3 variables. If the variable red value is 1, the person's favorite color is red. In this case, orange and blue would both have the value of 0, since that is not the favorite color. Many of the questions offered free response, and this feature was used quite creatively (This is a kind way to phrase it) by survey takers. So data cleaning was applied to categorize answers so that we could properly form an analysis. The variables 'whyplay', 'Playstyle', and 'earnings' were converted. Playstyle was converted into multiplayer, singleplayer, friends, and all. Were I to repeat this, I don't know if I would have included all as a column. It would have been better for all dummy variables to have a '1' (which they already would have, so the all column is redundant and thus no analysis was done on it). Why a person plays was divided into fun, improving, winning, relaxing, and all. 'All' in this section was the combination of the first three options, which were the original choices given in the survey. Relaxing was added because it was a common fill in answer. However, due to the confusing nature of all, I would exclude it in the future and just have the first three variables with '1.' Lastly, earnings seems to be a spot of confusion for the survey takers. They seem to confuse earnings with why you play. Which is a fair connection, what do you earn from a game: fun, place to chill, etc. However, I believe the original survey was asking about monetary earnings, so my dummy variables were: earns a little, earns a living, and earns elo. Elo is associated with rank. As you play a game you can play ranked to place higher (or potentially lower) compared to others. While it isn't monetary, I thought it could be interesting to see in comparison with mental scores who plays for rank. In total, 12 dummy variables were created. Bringing the total number of variables for analysis after the clustering to 25 (combination of the demographics, gamer profile, and the mutated dummy variables).

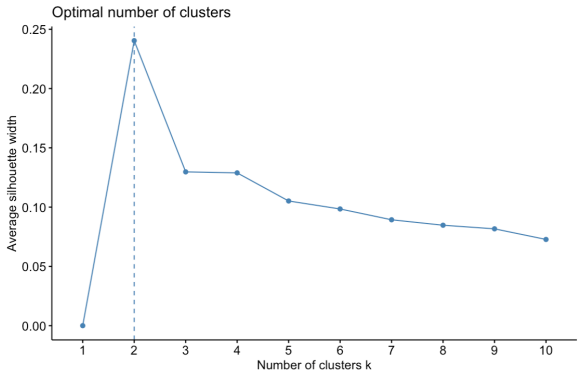
Once the data has been cleaned and oriented correctly, I wanted to look at potential biases that may still remain. Gender being a large one. The data set is vastly skewed towards males. This could be due to the skewed gaming community as well as reddit (where most people found and took the survey from).



Another factor would be the region of residence. It is skewed towards the USA; however, there is a decent proportion of entries outside of the United States. About 33% of the rows are from the US, but the rest are scattered across the globe, with the second most being the UK. There is potential bias in the questionnaire itself, because I believe it was only offered in English. This may account for why a large portion of the responses are from countries where the main language is English. I would provide a graph of proportion but there are too many regions for clarity. Age is also another area of high skew. This is reflected in the graph below. Most gamers are younger, as shown by the collected data. There is potential for bias because, again, the survey found the most takers on reddit, a forum mostly inhabited by those under the age of 30.



However, let's move into the calculations themselves! Once the mental scores were scaled (detailed above) they were perfect for a clustering model. However, for K means clustering (which works well for this model due to the non-hierarchical nature of the data) you need to specify the number of clusters prior to running. I used a silhouette plot to help with this.



It shows that I should use 2 clusters. While the data itself may be most happy in 2 clusters, I don't believe you can fit people into two boxes with mental scores. Especially 3 different tests. Had it been one test, you can cluster on whether the person scored high enough to be diagnosed. But with three individual tests, you have more patterns that can appear. For this reason, I went with the next uptick, 4.

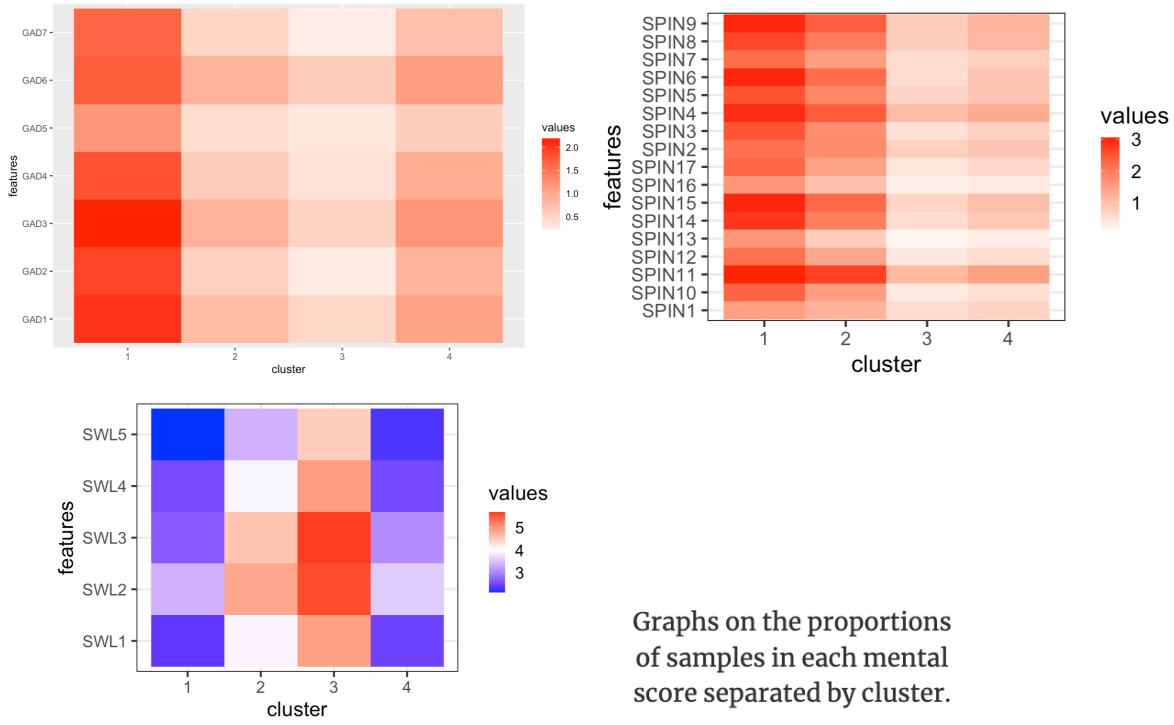
Once the model is prepared with 4 clusters, it proceeds to run an algorithm for each of the 4 clusters. It randomly assigns a center for 4 clusters. It then assigns the samples to their closest cluster centroid. Next, it calculates a new cluster centroid (center) for the mean value of the samples in the cluster. Once a center is calculated, you again will assign the samples (using euclidean distance) to their closest cluster. This process is iterated until there is no more movement of the centroids. At the completion of your model, each sample is assigned a cluster 1-4. I then add these cluster numbers to the original data so each sample is more easily identified and associated with their gamer profile.

Results and Discussion:

Due to clustering being an out of sample model, there are few metrics I can use to judge the accuracy of the clusters. Instead, I am looking at the characteristics associated with each cluster.

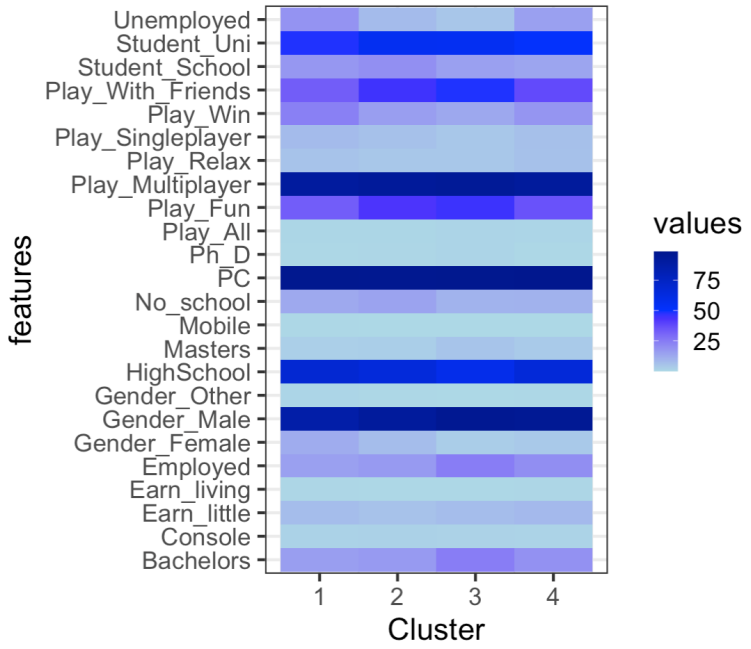
First, I will look at the mental scores for each cluster (reminder, this is the data they were clustered on). The mental scores and their scales are as follows. GAD, higher score equals more anxious, lower score equals low or minimal anxiety (score is measured in frequency: never to nearly every day (0-3)). SPIN, higher score equals more social phobia, low score means doesn't apply/No phobia (score is measured on How much a statement applies: Not at all to extremely (0-4)). Lastly, SWL, higher score equals high life satisfaction, low score equals low life satisfaction (score is measured from strongly disagree to strongly agree (1-7) based on a given statement).

Clusters 1 and 4 are higher in anxiety and phobia (Cluster 1 being more severe) as well as having a poor satisfaction with life. Clusters 2 and 3 have lower GAD and SPIN scores. They differ in SWL scores: Cluster 2 has a neutral satisfaction with life, Cluster 3 has a high SWL score.



Graphs on the proportions of samples in each mental score separated by cluster.

When the clusters are joined back with their respective gamer profiles (not used in the cluster model), we can see that some features vary highly and others not so much, between the clusters. To look at the features, since each cluster has a different amount of samples, I turned the features into proportions. So what proportion of the samples in cluster 1 play multiplayer. This way, it is a more equal playing field. Without further ado, the analysis...



A graph on the proportions of samples in each feature separated by cluster.

Playstyle was definitely varied for the clusters (however cluster 2 and 4 were similar on singleplayer and multiplayer preference proportions.) Cluster 1 had the lowest proportion for multiplayer and playing with friends (which is interesting given their streaming habits). Cluster 3 has the highest proportion for multiplayer and playing with friends. Clusters 2 and 4 differed mostly in paying with friends, in which cluster 4's proportion was 7% lower. It seems the clusters with higher GAD and SPIN scores have lower proportions in playing with friends.

Play reason was largely varied. Personally I would have guessed that those with more anxiety would play for fun and relaxing. It seems I was right with relaxing, but they also have a high proportion for playing to win. Something I was not expecting from them. Both cluster 1 and 4 had the highest playing to win and playing for relaxing. Usually those are not attributes we would pair together. Something else odd is Cluster 3, where they are the highest proportion for playing for fun and the lowest for relaxing. Typically you would pair fun and relax. It is unique to see that play reasons I would not have assumed to be clustered together, were clustered together.

There was no significant difference in platform for each cluster. So people with different scores don't seem to prefer one platform over the other. There are likely other attributes affecting this preference. For example, PC has a larger amount of games available.

Highest schooling wasn't significantly different for the clusters either. Any variation is most likely due to the large range of age for each cluster. As well as the large proportion of 18 year olds in each cluster.

Occupation was interesting for the features of unemployed and employed. The student of a university or of a standard school mostly equates with the age issues mentioned previously. It is interesting that the two clusters for which have the lower GAD and SPIN scores (Clusters 2 and 3) have a lower unemployment proportion. Between the two, Cluster 2 had a neutral SWL score, while Cluster 3 had a high SWL score. The cluster with the lowest proportion of unemployed gamers and highest proportion of employed gamers are the most satisfied with life. Cluster 1 has the highest proportion of unemployed gamers, and lowest proportion employed. Cluster 1 is also the cluster with the lowest SWL score and highest GAD and SPIN scores. Meaning they have anxiety, social phobias, and a poor outlook on life. It is interesting to see that interaction with their employment. I want to see if the clusters have an impact on the ability to profit off of games.

Another interesting note is that while cluster 4 has the second highest unemployed proportion, they also have the second highest employed proportion. This could be an interesting fact to dive in on. See if it may be impacted by other features outside of our GAD and SPIN scores.

For earning, there is some interesting variance. It seems the clusters with the most anxiety (Cluster 1 and 4) have the highest proportion for earning a living through gaming. It would be interesting to see why this is the case. Is it due to feeling safer behind a screen where you can control everything, there's less interaction with the public. But then there's also the aspect of social phobia, and how does that interact with the ability to have yourself on the internet playing games. It is important to specify, the questionnaire did not specify how money was earned. In the gaming industry there are many ways you can earn money; YouTube and Twitch are just the most popular. You can be a teacher, you can boost accounts (this is when someone of a high level or rank in a game, plays on other people's accounts to increase their rank), and many many more. We do have average hours streamed, so I will give those points. But I want to mention this may or may not include youtube or other forms of publication, in which the user uploads a video instead of actively streaming. This would be an interesting point to get data on. See if clusters are more likely to pre-record or stream.

Diving into hours streamed per week, it does appear that the most public-adverse cluster (Cluster 1) streams the most hours weekly. So the high anxiety levels don't appear to have a high impact on ability to stream. It may have the adverse effect, where you have the incentive to stream so that you don't have to get a job where you are in public with people. It could also be a replacement of time that more outgoing people are utilizing to go out with friends, or do activities that aren't online.

To help us look at streamed hours equally in terms of hours spent gaming, I converted them to a proportion (much like originally categorical variables in the cluster). It looks like Cluster 1 does spend the largest proportion of their play time streaming. Cluster 3 has the lowest proportion. I would be interested in why Cluster 3 plays for the least amount and streams for the least amount. Lesser play time could be due to partaking in more social activities or a busier life (they have a higher employment rate). But for the lower proportion of streaming time (streaming time / hours played), it would be interesting to find the reasoning. One reason could be that they don't need the potential financial income from streaming. But I'm sure each person has their reasons.

Conclusion and the Future:

For this project, clustering was used based off cleaned and transformed data found by Marian Sauter and Dejan Draschkow, in order to find a correlation between mental health and gamer attributes. Within this analysis, data was cleaned, transformed to dummy variables, potential biases were noted, mental scores were scaled and clustered using K-means clustering, and finally, clusters were analyzed by mental scores and gamer profiles. It does appear that people in each cluster have a varied profile. Higher variation was found in variables like play style and play reason. Lower variation between clusters were found in highest schooling, age, and platform played on.

In the future, I hope to run the same data with perhaps random forests after clustering. Just to see if the clusters can be accurately predicted with either the mental scores or the gamer profiles. I would hope that will highlight the significance of each predictor. Given a gamer profile, could I guess the mental score cluster. Another interesting approach would be to cluster just on total mental score for each test, or if someone scored high enough to be diagnosed (then you could decide if you wanted to include severity). However, I am quite happy that my model was off the raw scores, as there might have been patterns that I didn't see. The model didn't know that SPIN scores were from the same test, yet clustered them together anyways. It recognizes there is a pattern. I would hate to remove its ability to recognize patterns. I would also love to look closer into the interactions of psychology, gaming, and 'real world' habits (job industry, salary, etc).

Citations:

Sauter, Marian, and Dejan Draschkow. “Gaming Habits and Psychological Well-Being: An International Dataset about the Anxiety, Life Satisfaction and Social Phobia of over 13000 Gamers.” OSF, 18 Nov. 2017. Web.