**FINA 375**

**Unsupervised Data Mining: Cluster Analysis**

**State: New York**

**Name: Romario Medel-Estrada**

**Introduction**

CVS, the second largest personal care and retail pharmacy chain in the United States, is interested in staffing retail associate positions. From multi-regression and LPM analytics findings, CVS is aware that female candidates and applicants with some disabilities are willing to accept significantly lower wages compared to male and non-disabled candidates. Using its database of applicants who have applied for jobs in the past, CVS wants to cluster or segment candidates into five groups based on characteristics provided by candidates on applications. These characteristics are sex (male or female), age category, marital status (married, not married), citizenship (American or not), education level attained (High School or less, Bachelor or less, or Master or higher), and self-identified previous employer category (public, private, or self/family), number of people in the household, wage income expected or earned, other income, and disability accommodation requested.

The goal of this project is to provide to the directors of each state CVS analytics team a summary of five clusters of potential job candidates. The business intelligence gathered by this project will guide the analytics directors in creating appropriate advertising outreach (medium) to encourage applicants to apply for these customer service positions.

Table 1 shows the summary of the analysis for each of the five clusters generated in the unsupervised data mining technique.

**Table 1: Clusters Summary for New York**



In this unsupervised cluster technique, we use the AGNES method (bottom up) of clustering, a sample of 1000 observations (drawn randomly without replacement) and calculate the distance between cases using the Gower formula because we have a mix of categories and quantitative variables. We then link up the clusters by the Ward method since our measures are quantitative.

The technique does a good job of generating homogenous clusters that are distinct from each other (the agglomeration coefficient is 0.9947 or almost perfectly 1). We generate 5 clusters.

**The Largest Cluster and Least Helpful Cluster**

The largest and least helpful cluster has 256 cases (cluster #4). This cluster represents 1.56% of individuals who earn low income (less than $35,000), 2.7% are female, 91.4% are American, 66% are married, with zero median other income, and 3 people (median) in the household. Most of them (199 out of 256 people) work in the public sector, have (some) college education (139), and are 56 years to 55 older (66). But age categories 26 to 35 (59), 36 to 45 (54), and 46 to 55 (58) are very closely behind. This indicates that this a very homogenous group of male individuals ranging from ages 26 to 55 and older. Most importantly, this cluster are made up of fairly healthy individuals, with only 0.7% to 2.3% who are disabled in some ways.

This cluster is comprised of mostly male individuals (97.3%) and individuals who earn more than $35000 (98.44%). They are mostly healthy and educated. This cluster is not a target audience for a low paying CVS job.

In Minnesota the largest cluster was #5 with 310 cases. With 85.8% individuals earning less that 35,000$. Compared to New York these clusters seem to at opposite ends of the wage income spectrum. About 62% are female in Minnesota where New York only had 2.7% female. Where they are similar is in education and employment categories. Minnesota had most working in the public sector and having some college education just as New York did. Age and disability metrics were also relatively similar with most individuals in Minnesota being of ages 56 or older and 6.5% disabled at most. New York also had the largest age category being 56 or older and a high of 2.3% for some sort of disability.

With these 2 clusters being so similar in age and education categories, it is interesting to see the major difference in percentage of those earning low wages in each state. 1.56% for New York compared to 85.8% in Minnesota.

The least unhelpful cluster in Minnesota is cluster #3 (168). This cluster is also like cluster #5 from New York. In Minnesota, 95.5% were male vs 97.3% in New York. All individuals in this cluster earned more than $35,000 in Minnesota vs 98.44% earning more than $35,000 in New York. Almost identical in these categories.

**The Second Largest Cluster**

Cluster#1 has 241 cases. This cluster is represented by almost all individuals who earn less than $35,000 (95.44%). The individuals in this cluster are mostly female (56.85%), 89.6% are American, 84.2% are married, with zero median other income, and 3 people in the household. Most of them (141) work in the public sector, most have no college education (128), and mostly represent working older adults 56 or older (108). As a working adult cluster, individuals in this group are decently healthy for their age, with only 2% to 8.2% who are disabled in some ways.

For Minnesota the second largest cluster was #1 which we have are going to compare to New York’s cluster #3 as they both provide little useful information (not helpful or least helpful).

**The Most Helpful Cluster**

The most helpful cluster #3 has 194 cases. This cluster is comprised of 62.89% earning less than $35,000. Most are female (90.72%) and 68.56% are American. Only 11.86% are married, with 2 as the median number of people in the household. They also have zero median other income, 166 work in the public sector. 52 are younger than 25, and 59 are between the ages of 26 to 35. Most of them have some college education (bachelor’s or less, 130). This group is very healthy; they have only .51% physical and .51% visual disability with no other disability, they are the healthiest cluster.

The most helpful cluster for Minnesota is cluster #4 with 178 cases. 63.5% earning less than $35,000, almost identical to the 63.89% for New York. Age category and education categories are the same; Minnesota has mostly younger adults and bachelor’s or less. New York also has mostly younger adults and most individuals with bachelor’s or less education. These being almost identical clusters makes complete sense for both to be the most useful.

**The Smallest Cluster**

The smallest cluster has 113 cases (cluster #5). This cluster represents 14.16% of individuals who earn low income (less than $35,000), 99.12% are female, All are American, most of them are married (84.9%), with zero median other income, and 3 people (median) in the household. Most of them (64) work in the public sector, the majority have some college education (63) and are mostly between the ages of 46 to 55 (39). Most importantly, this cluster is made up of fairly healthy individuals ranging from .088% to 1.7% suffer from some sort of disability. The smallest cluster in this case does not indicate to be our target demographic.

The smallest cluster for Minnesota is cluster #2 (125). 99.2% of individuals earn less than $35,000. Most work in the public sector and are younger than 25. For Minnesota this cluster did not seem to be very stable. Disability ranged from 1.6% for hearing to 21.6% for memory disability. This looks a lot like New York cluster #2 where 18.37% of individuals had memory disability yet only 2.5% were least disabled; and were also mostly younger in cluster #2. It seems like the cluster #2 for Minnesota needs to be broken up further as does cluster #2 for New York. Compared to cluster #5 for New York, Minnesota cluster #2 is not alike despite them both being the smallest.

**The Remaining Cluster**

Cluster#2 has 196 cases. This cluster represents 99.5% of individuals who earn less than $35,000. 32.14% individuals in this cluster are all female, 96.43% are American, mostly not married (married equals 3.06%), with zero median other income, and 2 people in the household. Most of them (192) work in the public sector, most have no college education (121 high school or less) and mostly are younger adults less than 25 (76). As a young adult cluster, individuals in this group are decently healthy individuals, with only 2.5% to 18.37% who are disabled in some ways. I believe that the 48 individuals ages 56 and make up majority of the 18.37% memory disability as it is uncommon for younger individuals to have this disability at high proportions. This cluster has the highest level of individuals earning less than $35,000 (99.44%). High memory disability is not a good target audience.

The remaining cluster for Minnesota was cluster #1. It had 219 individuals with 4.56% of individuals earning less than $35,000. About 76.2% were female and 66.2% were married. They were mostly comprised of ages 46 to 55 (69) and publicly employed (142). Most of them had at least some college education (bachelor’s or less, 152). They had zero other income (median) and the number of people in household is 2. This cluster was healthy with only 0.45% to .91% having some form of disability. 96.8% are American.

**Conclusion**

The results of this data mining exercise can assist CVS to tailor its job advertising campaigns to reach out to potential candidates who might accept lower wages for entry associate positions in New York. For instance, if CVS’s goal is to hire young (healthier) adults, it should target cluster 3 (mostly working young females). It should reach out to this cluster via text messages, YouTube, Twitch, and Discord ads. Ads should target public sector institutions (state and federal agencies) where most of these individuals work. If the goal is to increase bilingual workers who are willing to accept lower wages, it should target cluster 3 also for this reason, as only 68% are American. About .51% of the individuals in this cluster have some physical disability, which CVS could accommodate via remote work options. The females in this cluster are also young, with mostly some high school or some college education. Lastly, to reach out to these workers, CVS should advertise in Instagram, TikTok, and Snapchat. Ads should target public sector institutions (state and federal agencies) where most of these individuals work. CVS should ignore clusters 2 and 4 in New York as they show little promise and concentrate on 3 and 5 in that **order**. Cluster 2 should be avoided at all costs as that cluster has high memory disability (18%).

In Minnesota, CVS should concentrate on clusters 4 and 2 and ignore clusters 1 and 3. Avoid reaching out to cluster 2 with 21% memory disability. Individuals because smaller clusters’ results are very sensitive to sampling variability using the hierarchical technique

Minnesota’s cluster 2 and New York’s cluster 2 are almost identical despite not being identical in size. New Yorks smallest cluster was not as unstable as Minnesota’s.

**Screenshot of Script, Data Environment, Plots, and Output**

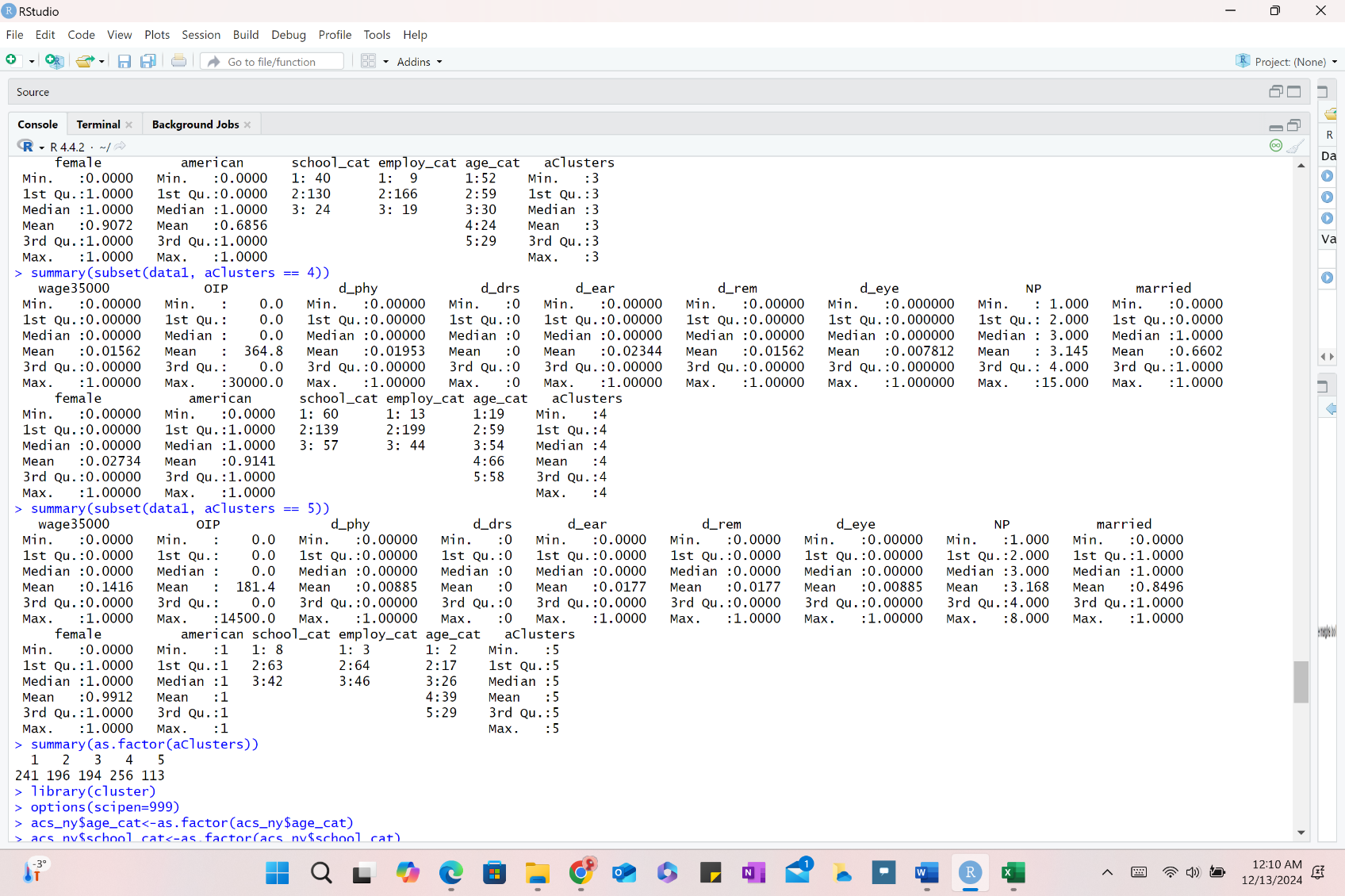
**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

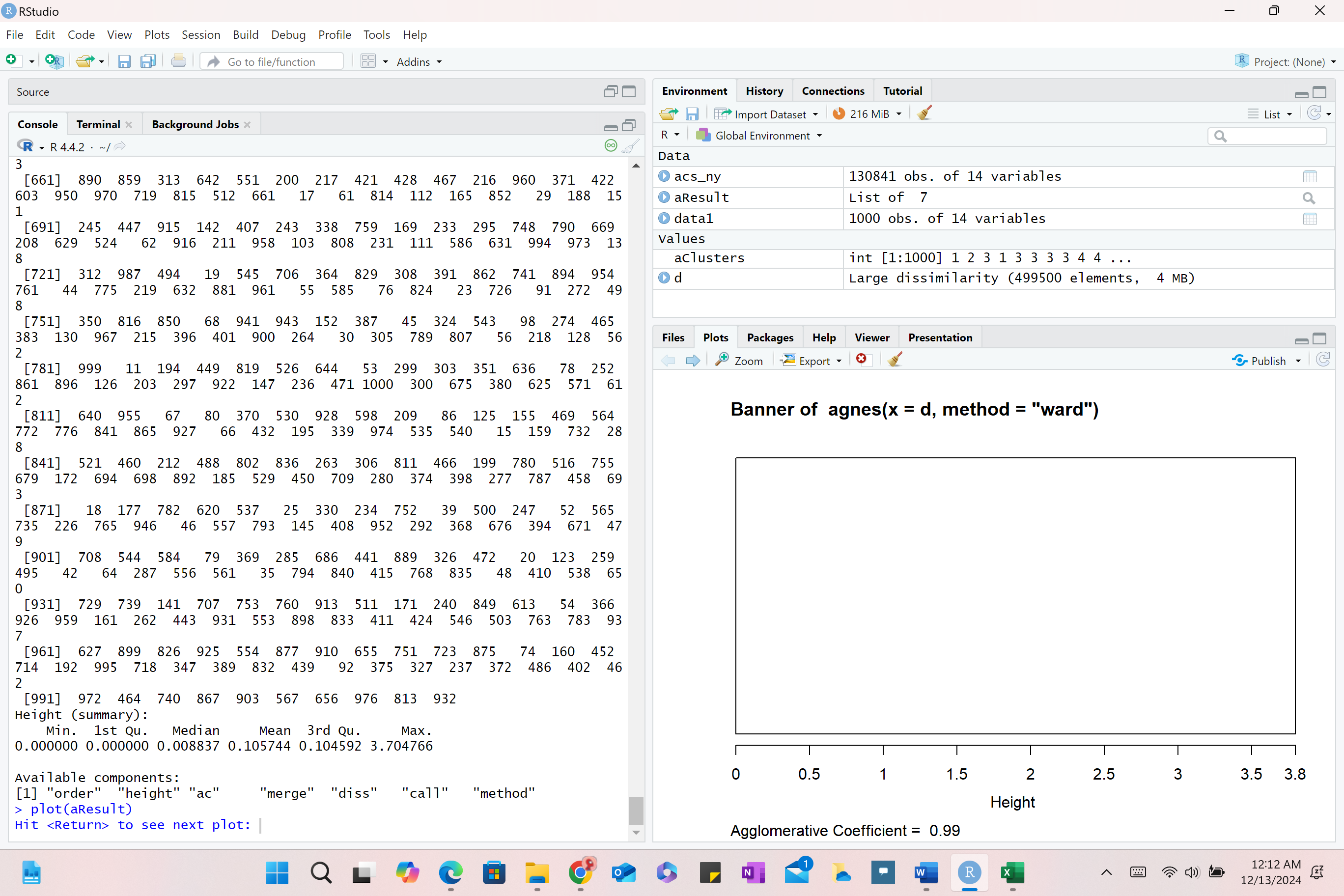
**A screenshot of a computer

Description automatically generated**

**A graph of a graph of a number of individuals

Description automatically generated with medium confidenceA graph of a graph of a person's equation

Description automatically generated with medium confidence**

****