Mitchell Kolb
CPTS 315
Course Project Report
12/12/2020

Table Of Contents

Introduction

The purpose of this project is to analyze the UC Irvine's Machine Learning Repository Divorce Predictors Data Set to see if there are any similarities to research done by professionals. There are two real world applications that I can see this directly applying to. The first being marriage counselors, they can use this information to help give better advice/counseling to their patients. The second being to me whenever I decide to get married. After completing this project hopefully I will see some of the main reasons why couples get divorced and hopefully that will not make the same mistakes as they have, if it's something I can avoid. Getting divorced is a long and complicated matter that is rough for both people.

Some of the questions that I want to answer in this project is, what are the most troubling parts of the relationship? What are the highest/lowest rated answers and do they match what researchers have gotten in their results. Can clustering be useful for me to help sort and find connections in the data. Is the agglomerative clustering method that I have looked into going to be effective with this dataset.

The personal motivation I have to select this dataset compared to doing others is that I have relatives who I was very close to for a large majority of my life get divorced and I wanted a reason to learn more about why people separate in the first place. Some of the separations I have witnessed have been very messy and I want to use the information from this dataset to if possible help me personally connect the dots on maybe why the people I once knew separated. My goal with this project is to classify the questions into groups based on categories like, compatibility errors, infidelity, financial issues, timing to try and pinpoint for at least the 170 people who took the survey for this dataset what has the highest rate of indication for divorce.

Some of the challenges I have to face in this project is that after doing a more in depth look into the dataset there are three issues with it that I will have to spend time to solve. The first being that the last column is labeled "CLASS" and it is not mentioned at all in the description of the dataset. Its value for every record is 0 or 1 and its 1 for the first 85 records and 0 for the last 85 records. I have a gut feeling this is the category that says if the man or woman answered but I don't know for sure. The second issue is that I have to go through all the questions and join them up into similar grouping based on subject matter. This will make it easier to draw conclusions from when I do my analysis. The last issue with the dataset is that I have only 170 entries so my data could be really lopsided and not accurate which will make it hard to see if my code implementation is correct. My approach to this task is going to be doing small initial calculations to get a feel for the data and see what the general consensus is then to implement the clustering method.

My results came out to be that

Data Mining Task

My input data is a .csv file with 170 entries which are the rows and the 54 questions each person was asked which are the columns.The 55th column is the Class attribute.

This is the information for every single question asked in the data set. Each column matches up with the question asked. (Ex. "Atr 2" in the dataset == Question 2 on this document)

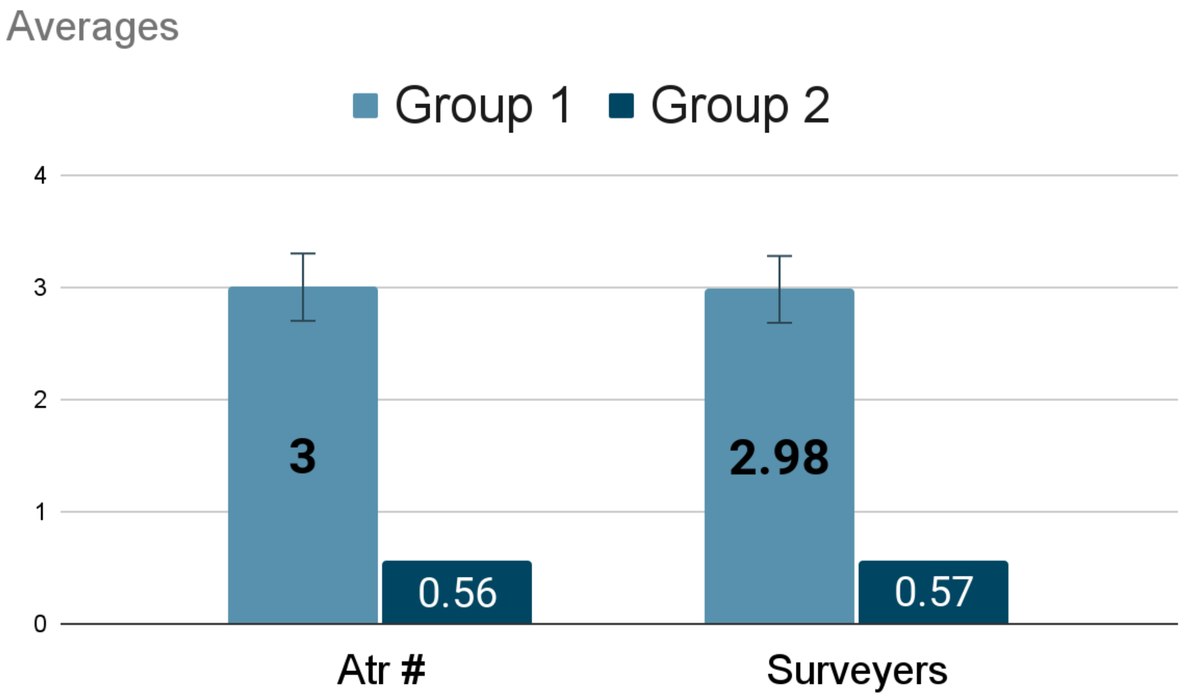
1. If one of us apologizes when our discussion deteriorates, the discussion ends.
2. I know we can ignore our differences, even if things get hard sometimes.
3. When we need it, we can take our discussions with my spouse from the beginning and correct it.
4. When I discuss with my spouse, contacting him will eventually work.
5. The time I spent with my wife is special for us.
6. We don't have time at home as partners.
7. We are like two strangers who share the same environment at home rather than family.
8. I enjoy our holidays with my wife.
9. I enjoy traveling with my wife.
10. Most of our goals are common to my spouse.
11. I think that one day in the future, when I look back, I see that my spouse and I have been in harmony with each other.
12. My spouse and I have similar values in terms of personal freedom.
13. My spouse and I have a similar sense of entertainment.
14. Most of our goals for people (children, friends, etc.) are the same.
15. Our dreams with my spouse are similar and harmonious.
16. We're compatible with my spouse about what love should be.
17. We share the same views about being happy in our life with my spouse, who is
18. My spouse and I have similar ideas about how marriage should be at
19. My spouse and I have similar ideas about how roles should be in marriage
20. My spouse and I have similar values in trust.
21. I know exactly what my wife likes.
22. I know how my spouse wants to be taken care of when she/he is sick.
23. I know my spouse's favorite food.

24. I can tell you what kind of stress my spouse is facing in her/his life.
25. I have knowledge of my spouse's inner world.
26. I know my spouse's basic anxieties.
27. I know what my spouse's current sources of stress are.
28. I know my spouse's hopes and wishes.
29. I know my spouse very well.
30. I know my spouse's friends and their social relationships.
31. I feel aggressive when I argue with my spouse.
32. When discussing things with my spouse, I usually use expressions such as ‘you always’ or ‘you never'.
33. I can use negative statements about my spouse's personality during our discussions.
34. I can use offensive expressions during our discussions.
35. I can insult my spouse during our discussions.
36. I can be humiliating when we have discussions.
37. My discussion with my spouse is not calm.
38. I hate my spouse's way of opening a subject.
39. Our discussions often occur suddenly.
40. We're just starting a discussion before I know what's going on.
41. When I talk to my spouse about something, my calm suddenly breaks.
42. When I argue with my spouse, ı only go out and I don't say a word.
43. I mostly stay silent to calm the environment a little bit.
44. Sometimes I think it's good for me to leave home for a while.
45. I'd rather stay silent than discuss it with my spouse.
46. Even if I'm right in the discussion, I stay silent to hurt my spouse.
47. When I talk with my spouse, I stay silent because I am afraid of not being able to control my anger.
48. I feel right in our discussions.
49. I have nothing to do with what I've been accused of.
50. I'm not actually the one who's guilty about what I'm accused of.
51. I'm not the one who's wrong about problems at home.
52. I wouldn't hesitate to tell my spouse about her/his inadequacy.
53. When I discuss, I remind my spouse of her/his inadequacy.
54. I'm not afraid to tell my spouse about her/his incompetence.
CLASS. Section was not used in dataset

Some of the details that I should add clarity to is the grouping system I determined based on the “Class” column in the dataset. I have put the people who answered 1 into group 1 and the people who answered 0 in group 2.

The Output data is the .txt file that I have python put all the calculations into with labels. Some of the things I calculate is the maximum, minimum, and mean of the results of every question for the entire group of people. The same thing is calculated with each individual who took the survey so I can see if they are an agreeable or disagreeable person to the questions being asked. I also used the sklearn.cluster model on a scatter plot to actually view the type of data and the volume of similar answers.

Figure 1. Calculated mean of Group 1 and Group 2 in both sections



The Data Mining questions I set out to answer:

- What are the most troubling parts of the relationship?
- What are the highest/lowest rated answers and do they match what researchers have gotten in their results.
- Can clustering be useful for me to help sort and find connections in the data.
- Is the agglomerative clustering method that I have looked into going to be effective with this dataset.

The key challenge that I plan to solve in this task is organizing all the data to make sure that I'm only dealing with the correct row or column of values. The class column tends to offset the indices of my for loops in my code. If I can keep the dataframe on the right section that I need it to be on I can comfortably rely on the scikit-learn library to display and plot my data correctly.

Technical Approach

- What are the most troubling parts of the relationship? (Addressing my challenge form above)
  - Answer: The most troubling part of the relationship between the groups is making sure that the dataframe is looping correctly. To ensure that this is correct and I'm not being set off by 1 every loop I have a check function that verifies with the pandas loop incrementer. This ensures that I stay within the same grouping and the start and end of the row/column.
- What are the highest/lowest rated answers and do they match what researchers have gotten in their results.
  - Answer: The Highest rated answers were the 4's in group 1 and they strongly agreed with a majority of the communication, time spent, and intimacy questions. This tracks with the research that professionals have done because the studies say that most relationships are at risk when one of the two people are displeased with the why they are being treated.
- Can clustering be useful for me to help sort and find connections in the data.

    - Clustering was useful with this dataset because it allowed me to view two versions of the data. One of the graphs was the data plainly put into an x, y scatter plot where I can see the clusters of answers (Figure 4). The other graph will then scale the points to the volume of them within the dataset range that I set (Figure 5). This would mean that I could be sure that answers 1 and 4 existed but if 4 was answered 90% of the time it would take up most of the space in the second graph.
- Is the agglomerative clustering method that I have looked into going to be effective with this dataset.
    - Agglomerative clustering was effective in this project because it can produce clusters of different sizes and shapes to show the amount of data in a certain field along with highlighting the clusters of data on the graph.

Figure 2: Visualization of the most agreeable answers in green

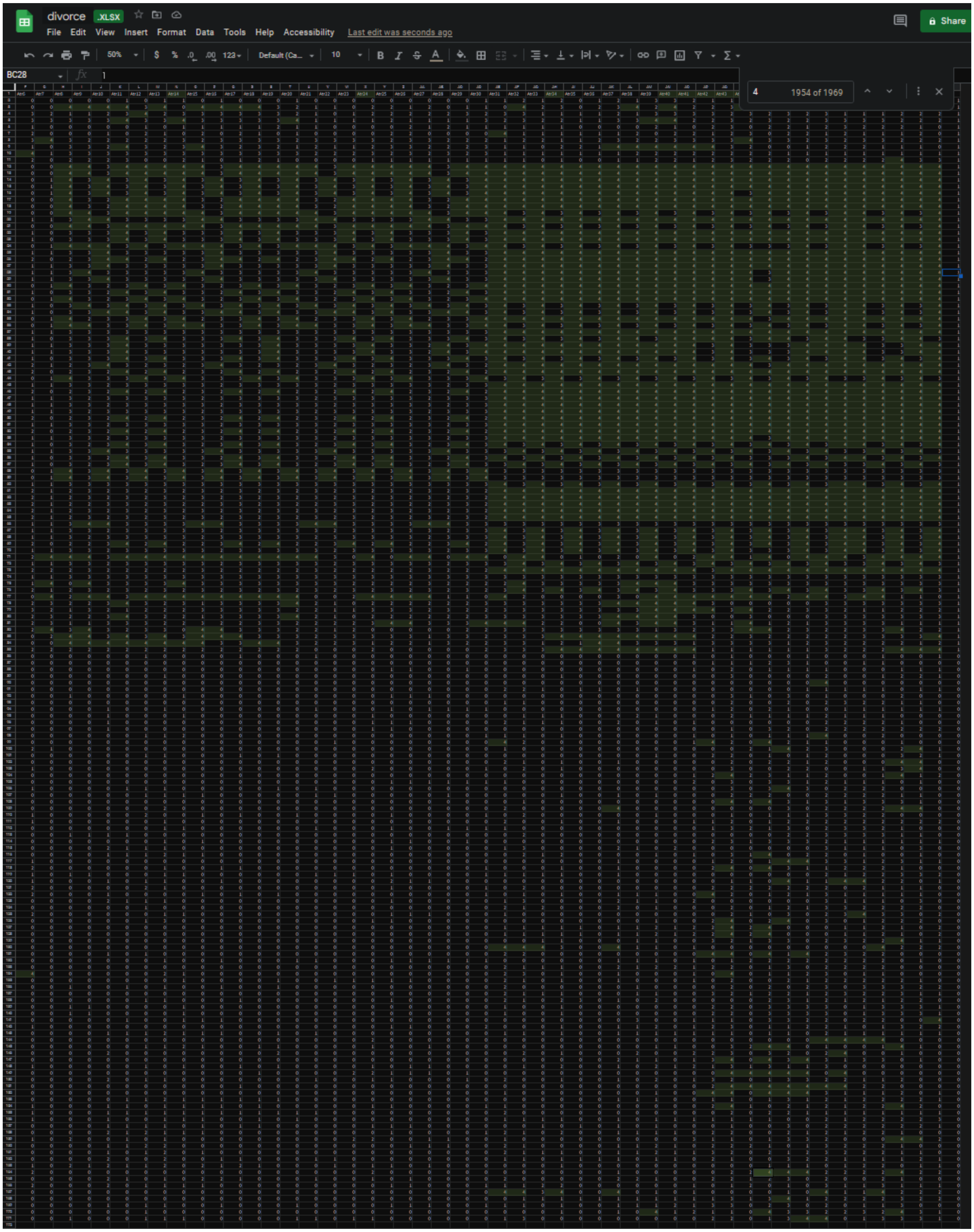


Figure 3: Visualization of the most disagreeable answers in green

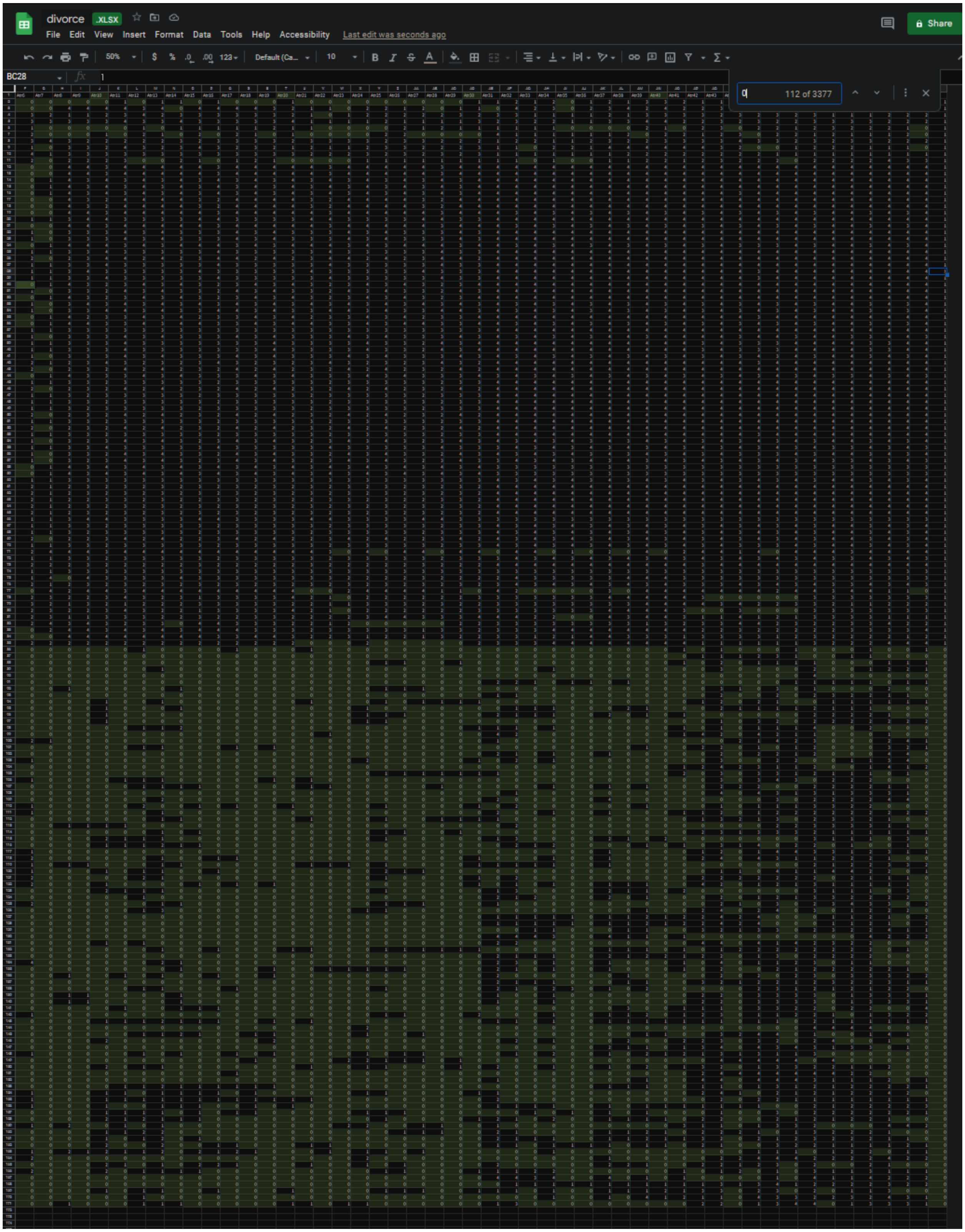


Figure 4. Scatter plot of the clustering model without logarithmic volume scaling

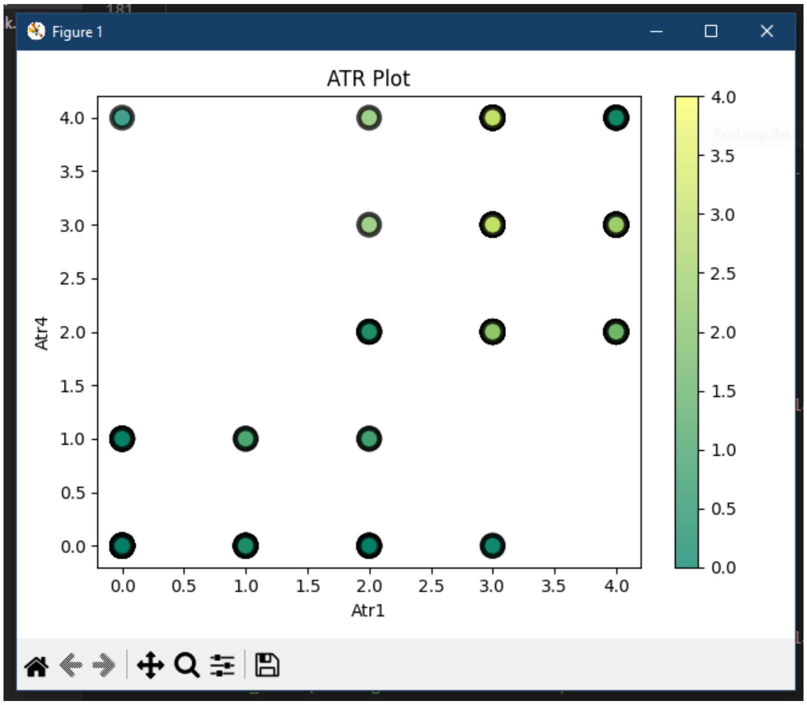
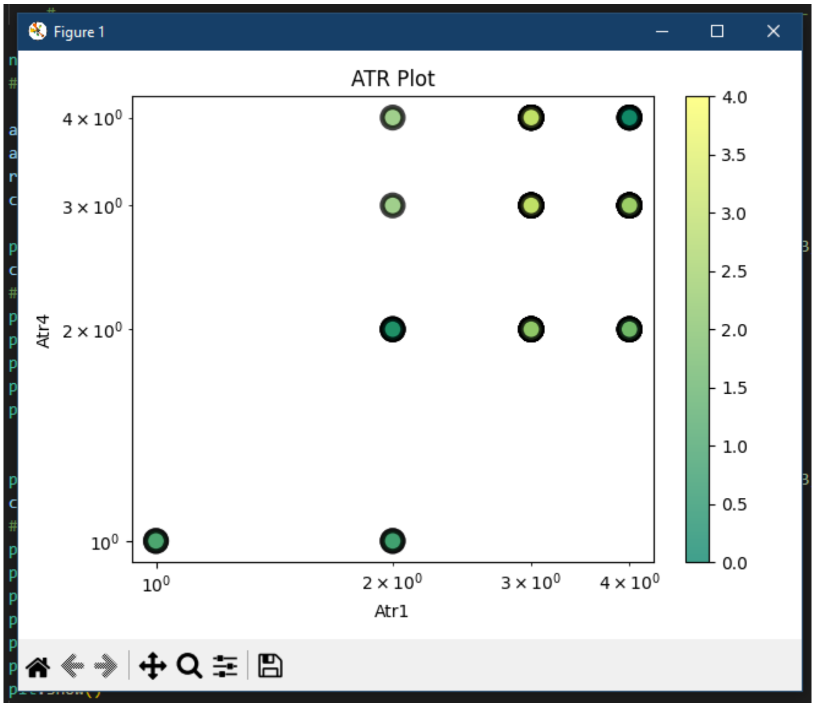


Figure 5: Scatter plot of the clustering model with logarithmic volume scaling



Evaluation Methodology

4. Evaluation Methodology
•Explain the dataset and its source that you employed to study this task. Any specific challenges to use this data for your task.
•List the metrics you employed to evaluate the output of data mining task and/or questions investigated. Justify their choice from real-world applications perspective.

Beginning Analysis:
This study took in the data of 170 people that were asked 54 questions. I put these people into two groups. The groups are separated by the class column. I did a majority of my calculations with the groups and compared the groups together. For the beginning of my analysis I took the maximum, minimum, and mean of results from each and every question. Along with this I also averaged the scores from every surveyor to determine if they were an agreeable/disagreeable person. Getting these results would allow me to determine what people have trouble with in relationships.

Clustering Model:
I included the agglomerative clustering model because it allows us to take the summations or averages of the questions or surveyors and plot them on a scatter plot. The agglomerative clustering is particularly useful because it allows me to specify the amount of clusters I want and merge the data points into that many clusters. In this project I do this twice to first show the trend line of the data placement for the questions I'm comparing (figure 4) and then the logarithmically scaled graph that will show the volume of points on the graph (figure 5). All points show up on the first graph while only the high amount of points show up on the second graph. This is how I can determine that the answer 4 is the most answered result for a given question.

Challenges:
One of the challenges I faced was using pandas to select the correct row/column. I forgot from last time I used it but the method I used had me using .iloc to the rows and .loc for the columns. Another thing I had a hard time understanding was the fact that I needed to use another graph when using the clustering method.

When dealing with two sets of 2d arrays it's impossible to display that on a graph, so what I did was to condense the data in one graph then just display the other array with the condensed graph to display both sets of data.

Results and Discussion

I had three sets of results that I ended up tracking.

The first one being whether the groups are agreeable or not. Group 1 scored an average of 3.0 for agreeing with the questions. Group 2 scored an average of 0.56 for agreeing with the questions (figure 1). This tells me that group 2 on average didn't know that their partner wasn't having a good time or that they gave up and stopped trying in their relationship.

The second thing I measured that I can draw conclusions on is whether people who scored highly actually answered highly and not just the full agreement 4 or full disagreement 0. Among the people who scored highly they would answer with a score of 2.98 which means that they were always hovering in the 3 range which is good. The people who scored low tended to answer low scores very frequently. Those people had an average score of 0.56.

The last thing I measured and could draw conclusions on is that when I grouped the questions together into related subject matter like finances, infidelity, effort, and communication. I could see based on the scores that the most troubling issues were, infidelity or lack of intimacy and communication. When questions relating to those two subjects were asked the high scorers in group 1 answered in agreement and the low scorers in group 2 broke away from disagreeing those few times to agree that there were troubles with them.

Lessons Learned

I learned two important things in this project.

1. Looking through your dataset before you pick it is extremely important. I liked this because of the subject matter and it made me gloss over the specifics on the data that is actually contained in it. This caused me a lot of trouble looking back because I had to do a lot of formatting and cleaning to make the data usable.

2. I should go with datasets that have more diverse numbers so the deviation between numbers is easier to understand. In this dataset I dealt with numbers 0-4, I had to rely on decimals to show me the changes in value. It is hard to know if a change in 0.5 is important or not. Working with bigger numbers would show the change a lot more clearly.

Acknowledgments

https://archive.ics.uci.edu/ml/datasets/Divorce+Predictors+data+set#

https://www.kaggle.com/datasets/adisak/divorce-predictors-data-set

https://datagy.io/pandas-groupby/

https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.scatter.html

A lot of youtube videos were watched in the making of this project