**Data acquisition**

Link to data source: <https://data.cdc.gov/NCHS/Conditions-Contributing-to-COVID-19-Deaths-by-Stat/hk9y-quqm>

Our group chose this data because it relates to the pandemic we are facing right now. This dataset has the potential to help people identify condition and age group information that could lead to COVID-19 deaths. In addition to that, the dataset is good for data mining because of its richness in data objects (286k rows) and also the moderate amount of features.

This chosen dataset records the number of deaths due to COVID-19 by time period, age group, and state. It has 286k rows and 14 columns. Most of the features that we can use are categorical features like “Condition” and “Age Group”. However, there are also some important numerical features like “COVID-19 Deaths”. More specific feature descriptions are listed below:

* Data As Of, Start Date, End Date: These three columns specify the time of analysis, the start of the data period, and the end of the data period, respectively;
* Group: What period of time a data object is recording: by total, by year, or by month;
* Year, Month: These two columns indicate the year or month in which death occured;
* State: The state in which death occured;
* Condition Group, Condition: These two columns describe the condition due to COVID-19;
* ICD10\_codes: Special code representing the condition;
* Age Group: Age group
* COVID-19 Deaths: The number of deaths related to COVID-19 that mention the condition listed;
* Number of Mentions: The number of total conditions mentioned in each age group. This column is necessary because there could be more than one condition reported for each individual.
* Flag: marked when the number of deaths is between 1 and 9.

Below is a description on features that I will study:

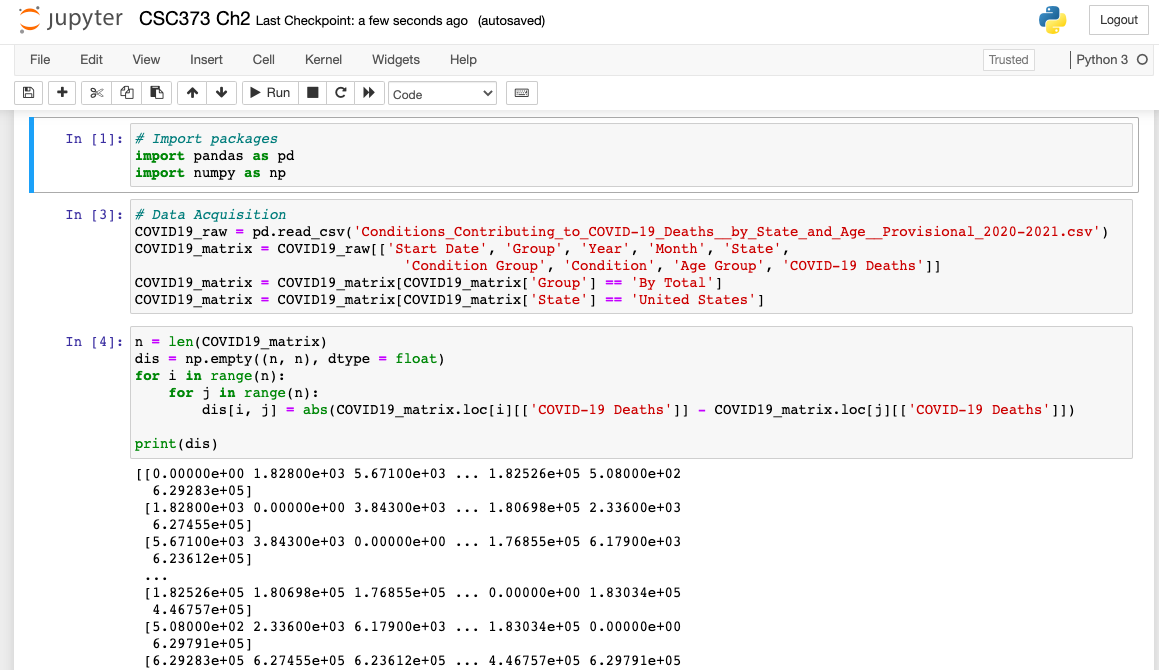
| **Important Features** | **Data Type** |
| --- | --- |
| State | Nominal |
| Condition Group | Nominal |
| Condition | Nominal |
| ICD10\_codes | Nominal |
| Age Group | Ordinal |
| COVID-19 Deaths | Numeric |

**Program development**

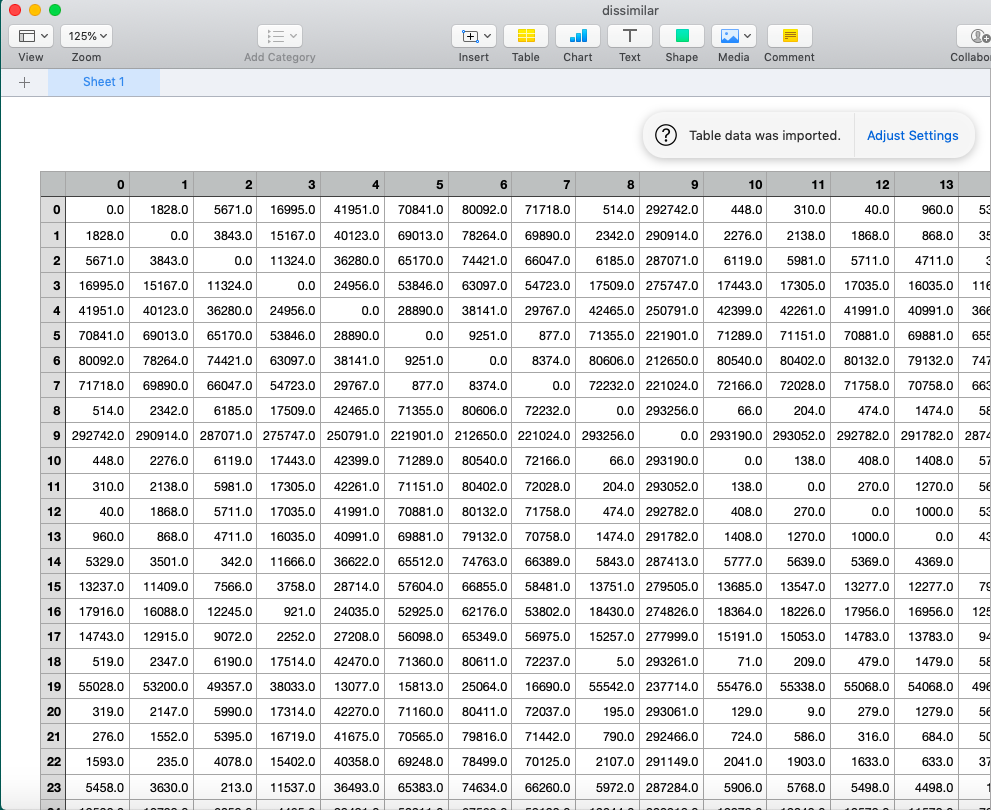
In order to better understand the structure of the dataset, I implemented several strategies learned in ch2 to analyze my dataset.

I want to understand the dataset better by using subsets of data that can give more general information about what I will study in the future. Therefore, I select data covering the entire period and also from the whole United States. Then I construct the data matrix using the data subsets. Finally I calculate the dissimilarity matrix based on how dissimilar two data objects’ “COVID-19 Deaths” is. The criteria I use is the absolute difference in “COVID-19 Deaths” between the two objects. Constructing a dissimilarity matrix is helpful because it can help me with classification based on “Condition” and “Age Group” in future steps.

Here below is my coding in the markdown file.



The output is a 230 \* 230 dissimilar matrix. I output the matrix to a csv file for further analysis. So here below is a more human readable version of the dissimilar matrix.



**Data analysis and package use**

I use python in Jupyter Notebook as my programming language and IDE. The packages I use are pandas and numpy, python packages for data processing.

I also use these packages to implement feature reduction:

There are 14 features in total in our dataset but only a few of them are features that we are interested in. Therefore, I need to do feature reduction before analyzing the data. I first exclude data with Irrelevant features:

1. Data As Of: This feature specifies the time this data object is analyzed but is irrelevant to our study.

Then I exclude the redundant features:

1. End Date: This feature can be calculated using the combination of Start Date, Group, Year/Month. Therefore this feature is redundant to our analysis.
2. Number of Mentions: This feature is really close to the feature COVID-19 Deaths in the sense that they have very similar distribution and also feature description. When analyzing, we can use either one of them to represent essentially the same thing. Therefore, we can delete this feature.
3. Flag: This part of information is contained in the feature COVID-19 whenever the value in this column is empty. Therefore, we can just delete this feature.

The remaining features are:

Start Date, Group, Year, Month, State, Condition Group, Condition, ICD10\_codes, Age Group, COVID-19 Deaths.

**Theory**

Exercise 21. Show that the set metric given by d(A, B) = size(A - B) + size(B - A) satisfies the metric axioms…

Ans:

1. Since (A - B) and (B - A) are matrices that have at least one element, therefore size(A - B) >= 0 and also size(B - A) >= 0. It follows that d(A, B) >= 0. Also, d(A, B) = 0 iff size(A - B) = 0 and also size(B - A) = 0 (since both are larger or equal to 0). That is, d(A, B) = 0 => ((A - B) = empty set, also (B - A) = empty set). This gives A = B since otherwise (A - B) will not be an empty set.
2. d(A, B) = size(A - B) + size(B - A) = size(B - A) + size(A - B) = d(B, A)
3. Since d(A, B) = size(A - B) + size(B - A) = size(A) - size(A and B) + size(B) - size(A and B) = size(A) + size(B) - 2 \* size(A and B), it follows that

d(A, C) + d(C, B)

= size(A) + size(C) - 2 \* size(A and C) + size(C) + size(B) - 2 \* size(C and B)

= size(A) + size(B) + 2 \* size(C) - 2 \* size(A and C) - 2 \* size(C and B)

That is, we need to compare:

=> (-2 \* size(A and B)) with (2 \* size(C) - 2 \* size(A and C) - 2 \* size(C and B))

=> size(A and B) with (size(A and C) + size(C and B) - size(C))

=> size(A and B) with (-(size(C) - size(A and C)) + size(C and B))

=> size(A and B) with (-(size(A^c and C) + size(C and B))

=> size(A and B) with (size(C and (B - A^c)))

=> size(A and B) with (size(C and (B and A)))

=> size(A and B) with (size(C and B and A)))

Therefore, we have size(A and B) is always larger than or equal to (size(A and C) + size(C and B) - size(C). Consequently d(A, B) is always smaller than or equal to d(A, C) + d(C, B). Proved.

Exercise 23. Given a similarity measure with values in the interval [0, 1], describe two ways to transform this similarity value into a dissimilarity value in the interval [0, inf].

Ans: Assume the similarity measure is s, then:

1. D = -log(s)
2. D = (1 - s) / s

Exercise 24a. Proximity is typically defined among a group of objects. Define two ways in which you might define proximity among a group of objects.

Ans: Assume that the group of objects contains n objects.

1. Let the proximity matrix be a n \* n matrix with entry (i, j), 1 <= i, j <= n, be defined as the proximity measure between object i and object j. Then the proximity among a group of objects is defined as the combination of the proximity between each pair of objects in the group.
2. Proximity can also be defined as a common attribute of this group of objects. One instance is to sum up the proximity measure for all possible pairs of objects. Then proximity can be interpreted as how close the objects are to each other.

This question can also be interpreted in another way. In that case,

1. Proximity can be defined as the similarity between objects in this group. In this case, when data points are closer to each other, they have larger similarity values, and vice versa.
2. Proximity can also be defined as the dissimilarity between objects in this group. Then when data points are closer to each other, they have smaller dissimilarity values, and vice versa.

Exercise 24b. How do you define the distance between two sets of points in Euclidean space?

Ans: One way is to define the distance between two sets of points as the maximum distance between every possible point x\_i and y\_j (x\_i is in set one and y\_j is in set two).

**Student learning summary and self-assessment**

What I Learn:

There are several aspects of data preprocessing that I found very helpful through learning this chapter. Moreover, I applied a few of those techniques that I think are crucial in my case to our dataset. For instance, I did attribute types analysis, dimensional reduction, proximity measure in our dataset. However, I’ve learned more techniques on this topic than listed above. For one, dealing with missing values that I learned in this chapter is extremely useful in the sense that almost all datasets have some missing values. In fact, I’ve already applied methods like “eliminate objects with missing values” and “estimate missing values” to another dataset I am using for one of my statistics classes. In addition, I also have a better understanding on how different types of proximity measures can be applied to actual datasets: cosine similarity is commonly used in asymmetric attributes; and correlation is commonly applied to time series, etc.. I believe I am now better equipped to do clustering and classification.

Discoveries:

I think I can do better in calculating the dissimilarity matrix for our dataset. The dissimilarity matrix is calculated based on the attribute COVID-19 Deaths: two objects are dissimilar if their values of COVID-19 Deaths differ by a lot. Since there is only one attribute I based my dissimilarity measure on, I only used the absolute difference between the attribute COVID-19 Deaths as my measure. However, I believe that there is more that I can do. For example, I can also evaluate the dissimilarity between different age groups instead of objects. In such a case, I will use all other attributes I have as constituting the dissimilarity measure. In short, I think there is more I can do when analyzing the dissimilarity.

Questions:

1. I think the discretization which uses entropy for partitioning is just like the branching for decision trees. Is my hypothesis correct?
2. I really want to consolidate my understanding of how to deal with missing values using more actual datasets. I don’t think this is necessarily a question but I think when I am practicing filling in missing data, there will be questions that I don’t yet know how to solve.

Self-assessment:

I think I have a very strong understanding of the data preprocessing part. If graded upon my understanding and my efforts in this group project, I think I may assess myself as A. However, I also realize that I didn’t finish most of this data portfolio until Sep.12. To give myself a more holistic evaluation, I rate myself as A- for this chapter. I will also make sure that I set the pace appropriately for the next chapter.