Owen Huang CSE 332 Graph Report

What is data first of all?

My data consist off suicide rate in the United States based on age group from year 2000-2010. So for every year there is data for number of suicides for ages 5-14, 15-24, 25-34, 35-54, 55-74, and 75+. Additionally, there will be corresponding population for that age group during that year. And it will show the GdpPerCapital and GdpPerYear of the United States during that year. So basically, my data shows suicides rate base on age group but also include the year, GdpPerCapital, and GdpPerYear.

**Bivariate Scatter Plot:**

So, the bivariate scatter plot I believe does a good job at portraying the data for what determines suicides rate based on ages. For example, I learned that over the years suicide rates increased. While not much, they still increased. Similarly, as gdpperyear and gdppercapital increased, the suicide rates from all age group also increased. I also learned that suicides rates are highest among age groups from 35-54-year-old. An interesting finding, I saw was that older age group who had higher population had higher suicide rates. I thought population didn’t matter as much compared to age group, but it does seem like higher population equal more suicide. Or maybe it has to do with the fact that we have much more older people in our society compared to younger people. While the bivariate scatter plot did a good job at showing suicide rate vs. other factors, it didn’t to much in comparing other factors such as age vs. year or year vs. population. It didn’t tell me much since my data is mostly based on suicide rate and the other attributes are there to back up suicide rates and not the other attributes.

**Correlation Matrix:**

Similar to the bivariate scatter plot, the correlation matrix showed my what correlated really well with suicide numbers vs everything else. I learned that suicide number and population have the most positive correlation. Then age and suicide number have had the second-best correlation. Year, gdpperyear, and gdppercapital all had around the same correlation since they don’t change per year. For example, if it 2000 and for every age group the gdpperyear and gdppercapital would be the same for every age group since it doesn’t change base on age but based on age. The correlation matrix doesn’t work very well with my data since my data has lot of same data for like gdpperyear and gdppercapital which are the same for every age group until it changes to different year. So, as you can see, they are very correlated since they always are the same within the same year and changes every year. So, its no surprised that they are very correlated.

**Scatter Plot Matrix:**

The scatter plot matrix is similar to the bivariate scatter plot, but does a good job at showing relationship between the attribute and making it easier to look it. For example, I see no matter what people who are 75+ have the highest suicide number. Either it be population, year, or gdppercapital they are always at the top of the graphs indicating they have the highest suicide numbers. On the other hand, ages 5-14 has the lowest suicide rates which has no surprise there. While it does a good job at showing patterns and such for suicide numbers and the other attributes, there are a lot of useless information that is on the graph. Like I said of the bivariate scatter plot, there is no need for implement attributes vs other attributes if one of them is no suicide since it doesn’t really tell us anything. So, a majority of the scatter plot matrix is filled with data that has nothing to do with what we want to learn about.

**Parallel Coordinate Display:**

The parallel coordinate does show me a good representation of my data. I believe that it not the best graph to show what correlates with suicide number since it hard to look it and that some correlation don’t matter. For example, my highest correlated attributes are gdpperyear and gdppercapital. And this is because gdppercapital and gdpperyear don’t for difference age groups but only change base on year. So it’s no surprised that they are the highest correlated. But what did pop out to me was the how year and suicide number were most less than 6000. It surprising to me to see that a majority of suicides be year are less than 6000. I thought they would be more spread out. But seeing the graph I see now that they are mostly at a certain range with some outliers being in 10,000s. Other things that pop out was that over the course of the 10 years age group population didn’t change as much. I thought maybe in certain years different age group would have different population, but they are usually consistent with a certain range.

**PCA Plot and Scree Plot:**

The PCA plot does a good job at showing at showing clustering of samples base on similarity. Seeing how there are 10 distinct columns on my PCA plot it must being showing the similarities between the 10 years of data that I’m representing. I see now that my data base on certain attributes must have certain similarities seeing how most clusters are at certain horizontal levels that are the same compared to the other ones. The different levels must be representing the different age groups that I have in my data. Nevertheless, it does show me that certain age groups have similar similarities I already thought they would have based on the other graphs. However, a con about this graph is that it doesn’t really tell what’s similar. Base on the things I said, they were most what I think. Maybe there could have been some color that I could have added or features to tell me what group of things has similar traits.

The scree plot does a good job of representing the principle components of my data. I thought that it would be one or two attributes that would dominate my data and showing me that really matters when it comes to suicide numbers, but looking at my data now, it can be safe to say that everything plays a little part in it. There is no like one component that matters more than others. There weren’t any outliers where one component super dominated the others or one that didn’t play a part at all. While, the components were at different levels, it wasn’t skewed to one side. A con would be that base on the scree plot it really hard to tell when there is a elbow or bend in the data so we don’t know the correct number of factors/components to retain so it might be unreliable.

**MDS Euclidian:**

From the MDS graph of my two components, I could say that there I not certain point in telling what is similar or not. All the data points are very spread out, so I can’t really tell if there is a similarity or not. However, at the same time I can say that as component one increases in value component two decreases in value which I can maybe say have a opposite affect on one another. That was a surprising feature that I noticed.

**MDS Cosine:**

From the MDS graph of my cosine similarity, I can say that is a positive correlation between component 1 and component 2 seeing how one component increases as the other one increases. There is a direct affect of one thing onto the other. A con would be similar to that of the MDs Euclidian would be that we don’t know what is the attribute that is similar within the components. I wish there would be a way to tell what makes these two components similar in the way that they do.