D212: Data Mining II: Task 2

Performance assessment

Part A: Research Question

1. One question that could be asked from this dataset is to ask: Do patients with higher income have more pre-existing conditions?
2. My goal of this analysis is to perform a PCA on a lot of the variables within this dataset to determine which ones have negative and positive effects on each other. The focus on this analysis is going to be income. I hope to use a scree plot in order to determine the effects different variables will have on the affect/being affected by income and whether it’s a positive or negative effect. Since PCA isn’t a prescriptive analysis type, this analysis will help us determine components and their weights between other components.

Part B: Expected Outcomes

1. PCA analyzes the dataset and finds the relationships between one variable and another. It checks how the altering of one principal component influences another principal component. By understanding which factors effect which, we will be able to isolate the most meaningful variables for our analysis to best the best results with our data analysis.
2. One of the main assumptions with principal component analysis, is to assume that the dataset has larger than 3 components. This is because the main problem PCA is meant to address is solving for data than can’t be fit nice and easily into a 3D plot or analysis which is why PCA checks multiple variables and finds which ones are the most significant.

Part C: Data Preparation

1. For my analysis, I’m using 2 continuous pieces of data to help with my analysis. The pre-existing conditions are binary so those are simple Yes/No, 1/0 for the sake of our calculations. The continuous variables that I’ll be testing will be Income and VitD\_levels. Both variables are continuous and will help in the determination of how income could affect pre-existing conditions as well as vitamin D levels.
2. The standardized dataset is attached

Part D: Analysis

1. See attached for matrix of principal components
2. By using the elbow rule for my scree plot, it seems that almost all components in my analysis have an effect. There are 2 points on the scree plot that show a slight bend, but nothing quite to the effect of most scree plots. It appears that there is a slight ‘bend’ at 2 principal components as well as at 12. This would indicate that either 2 principal components are needed, or almost all of them, which doesn’t really provide an incredibly useful insight into the analysis. The scree plot is attached.
3. The list below is the variance of each principal component in this analysis:

PC1: 0.08178216061897814 - Income

PC2: 0.16203381565930688 – VitD\_levels

PC3: 0.24166941212356716 – HighBlood

PC4: 0.3208133797227711 – Stroke

PC5: 0.39895685065519876 – Overweight

PC6: 0.476597308109013 – Arthritis

PC7: 0.5533973957615894 – Diabetes

PC8: 0.6295754306950528 – Hyperlipidemia

PC9: 0.7052162183054737 – BackPaid

PC10: 0.7801168455252614 – Anxiety

PC11: 0.8542102757495849 – Allergic\_rhinitis

PC12: 0.9272710379103934 – Reflux\_esophagitis

PC13: 0.9999999999999999 – Asthma

1. Total variance was calculated by adding up all the variables of all the individual components. Below is the calculated values for each principal component.

PC1 -0.888536

PC2 0.878456

PC3 -1.795204

PC4 0.737709

PC5 -0.405254

PC6 0.217091

PC7 0.554436

PC8 1.224028

PC9 0.079169

PC10 1.848499

PC11 1.275822

PC12 -0.135071

PC13 0.765284

1. For this analysis, the goal is to find how many components are needed to fully analyze our data. The original question was to find the relationships that pre-existing conditions certain Incomes had. From looking at the PCA matrix, we can see how all components are affected by each other component. If we solely look at Income, we can see that the conditions of Stroke, Arthritis, and BackPain are all 0.1>x>-0.1. This means that they have the largest relationship to Income. If we go across the columns, we can see that principal components 5 and 6 are within the 0.1>x>-0.1 range so we know that 5 and 6 components have the largest relationship to Income. This seems different than what our scree plot showed, but our scree plot wasn’t very good and looked almost linear as opposed to having an elbow bend. The biggest short coming of this analysis is probably the fact that our scree plot is so poor that it doesn’t do a great job of telling us how many components would optimally determine the relationship of Income. My only real recommended course of action with this analysis is that patients who have symptoms of strokes, arthritis, or back pain seem to have higher incomes than others. It is recommended that another analysis checking on a few more components might be able to add some clarification to the analysis by adding more points to help expand the scree plot and find a proper elbow bend in order to get a better accounting of how many components are necessary when determining what factors are affected by income. Since PCA is not a predictive analysis type, it would help to perform an analysis that could help with the prescriptive nature of pre-existing conditions and income

Part E:

Panopto Video: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=8ebb3be7-6b46-4160-a2b5-ae290183ea37>

Sources:

<https://docs.displayr.com/wiki/Kaiser_Rule> Kaiser Rule, Nunnally, J. C. and I. H. Bernstein

<https://jakevdp.github.io/PythonDataScienceHandbook/05.09-principal-component-analysis.html> In Depth: Principal Component Analysis, Jake VanderPlas

<https://ro-che.info/articles/2017-12-11-pca-explained-variance#:~:text=The%20total%20variance%20is%20the,divide%20by%20the%20total%20variance>. Explained variance in PCA, Roman Cheplyaka

<https://www.geeksforgeeks.org/python-pandas-dataframe-sum/#:~:text=sum()%20function%20return%20the,the%20values%20in%20each%20column>. Python Pandas dataframe.sum(), Shubham\_\_Ranjan

<https://stackoverflow.com/questions/24729447/what-does-it-mean-to-have-zero-mean-in-the-data> What does it mean to have zero mean in the data? Jamie Ballingall

<https://www.youtube.com/watch?v=FgakZw6K1QQ> StatQuest: Principal Component Analysis (PCA), Step-by-Step, StatQuest with Josh Starmer