Using Clutter Reduction and Multi-Dimensional Visualization Methods to Create Exploratory Visualizations of High Dimensional Data

Jordan Atwell
University of Hawaii at Hilo
jatwell7@hawaii.edu

1. Abstract

High dimensional data is often difficult to visualize in a way that is easy to interpret due to the number of variables. By using methods of clutter reduction and applying mulit-dimensional visualization techniques, it is possible to create visuals that successfully interpret data into an image.

Keywords: dimension, visualization, data, parallel coordinates, scatterplot matrices, multi-dimensional visualization

2. Introduction

2.1. Clutter Reduction

Clutter reduction is a preprocessing method that is used to reduce the amount of noise between dimensions [1]. The methods implemented in this project consisted of clutter reduction via dimension reordering. Its main purpose is to assist in data exploration by establishing structure through a variety of visualizations. The two methods used to visualize and reorder dimensions in this project were parallel coordinates and scatterplot matrices.

2.1.1. Parallel Coordinates

Parallel coordinates is a method of visualization where each observation in a dataset is plotted between each attribute as a line segment [2]. It is used in clutter reduction to spot outliers in the data in order to reorder the dimensions and reduce the amount of outliers between attributes [1].

2.1.2. Scatterplot Matrices

Scatterplot matrices are used in clutter reduction to reorder the dimensions by cardinality. This helps to create structure within the matrix and assist data exploration by establishing patterns [1].

2.2. Multi-Dimensional Visualization

Creating visualizations of high dimensional data can be difficult due to the amount of attributes present. The goal is to create an understandable and interpretable image of our data. This can be done by applying multi-dimensional visualization methods, where features such as depth, color, and size can be used to create more complex yet visually appealing visuals of data [4].

3. FIFA19

The first dataset analyzed is a collection of player statistics from the Fédération Internationale de Football Association (FIFA). It contains player information in the 2019 season such as build, height, weight, as well as various ratings. It was found on Kaggle and it contains 18,209 records and 89 attributes [8].

3.1. Exploratory Questions

One item explored was what abilities and player qualities (i.e. agility, aggression, etc.) were desired in a player. This was done by analyzing a subset of the data with the overall rating of the player and ratings of various abilities and qualities.

Another was whether overall ranking determines wage and market value by investigating if there is a correlation between wage, value, and ranking. Overall ranking is an average rating based on rating of various player skills. Additionally, whether international reputation had an affect on the wage or value of a player was explored.

3.2. Preprocessing

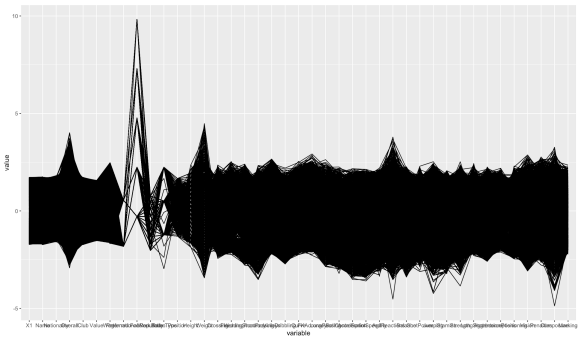
Methods of preprocessing used were dimension reduction by removing columns that were not useful for the analysis and removing observations that had any NA values in the columns that were being used. After performing this preprocessing, the dataset contained 35 attributes and 17,918 records.

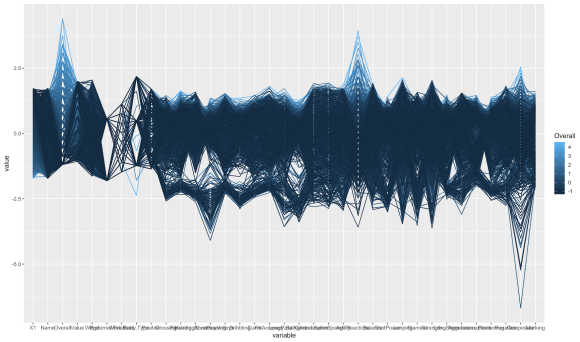
From this, a parallel coordinate plot was created to assess the number of outliers in the dataset. However, since the number of

records is large, the visualization created did not provide useful information as seen in Figure 1. The lines plotted could not be colored as RStudio would crash after 20 minutes of trying to perform the task.

In order to ensure that interpretable visuals could be made, the record size was reduced to the first 1,000 records, effectively the top 1,000 players by overall ranking. From this, a visual using the overall ranking for color was created as seen in Figure 2. When attempting to color by position, RStudio also crashed.

A further reduction in the number of records was implemented, limiting the dataset to the first 500 observations. From this, appropriate visuals were created. After preprocessing, the dataset had been reduced to 36 dimensions with 500 observations.



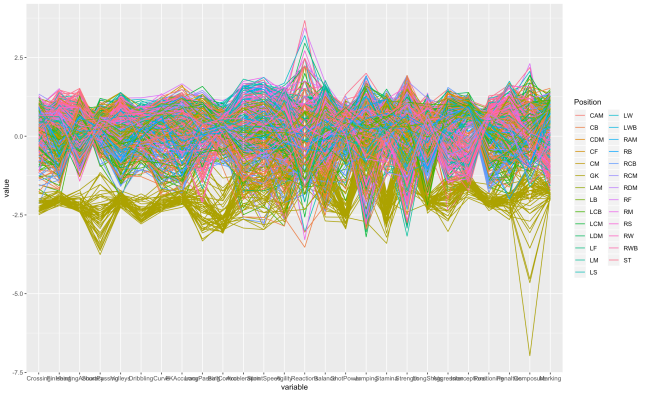


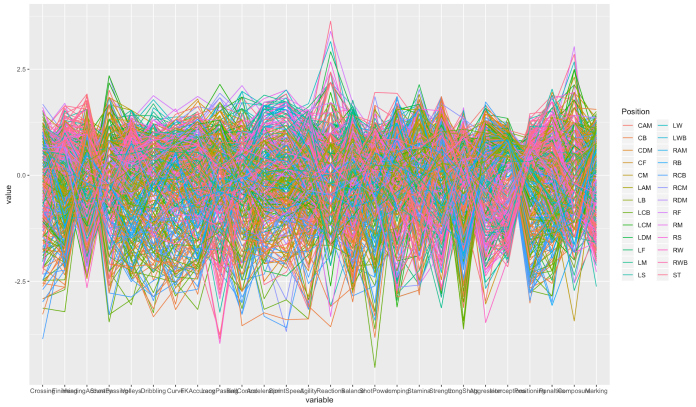
Figures 1 (left) & 2 (right)

3.3. Visualizations

3.3.1. Parallel Coordinates

An initial parallel coordinate visualization was created to analyze how the dimensions should be reordered to reduce outliers (Figure 3a). This was done using the ggparcoord function from the GGally library, an extension of ggplot [5]. The dimensions were in their default order for this. It was realized that goalkeepers are an outlier to player qualities and abilities, and were subsequently removed. Figure 3b is a revised parallel coordinates plot without the goalkeepers. The number of outliers in the data decreased greatly once goalkeepers were removed from observations. After a number of attempts to rearrange the dimensions to reduce outliers, it was decided that the default order was in fact the order i which there were the fewest number of clear outliers between dimensions.





Figures 3a(top) & 3b(bottom)

3.3.2. Scatterplot Matrices

In order to examine the cardinality of the data and determine whether there were any correlations between dimensions, a scatterplot matrix was created (Figure 4a). This matrix was created using the pairs function with the dimensions in their default order. The dimensions were then reordered by cardinality (Figure 4b).

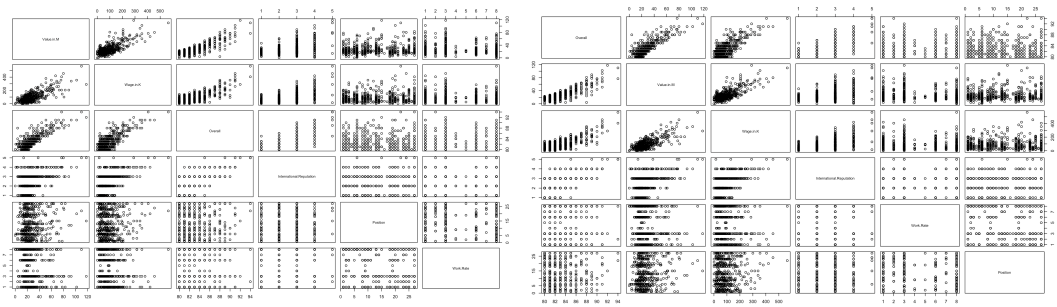
3.3.3. Correlograms

Correlograms are a powerful exploratory tool that create visualizations detailing where there may be relationships among data that are worth exploring [7]. Using the corrplot library, a basic visualization showing the correlation between different player qualities was created (Figure 5a). After reordering dimensions, the visualization was enhanced by adding a feature that crosses out insignificant correlations. This can be seen in Figure 5b. From this visual, it is clear that the three significant player qualities that contribute to overall ratings are jumping, reactions, and composure.

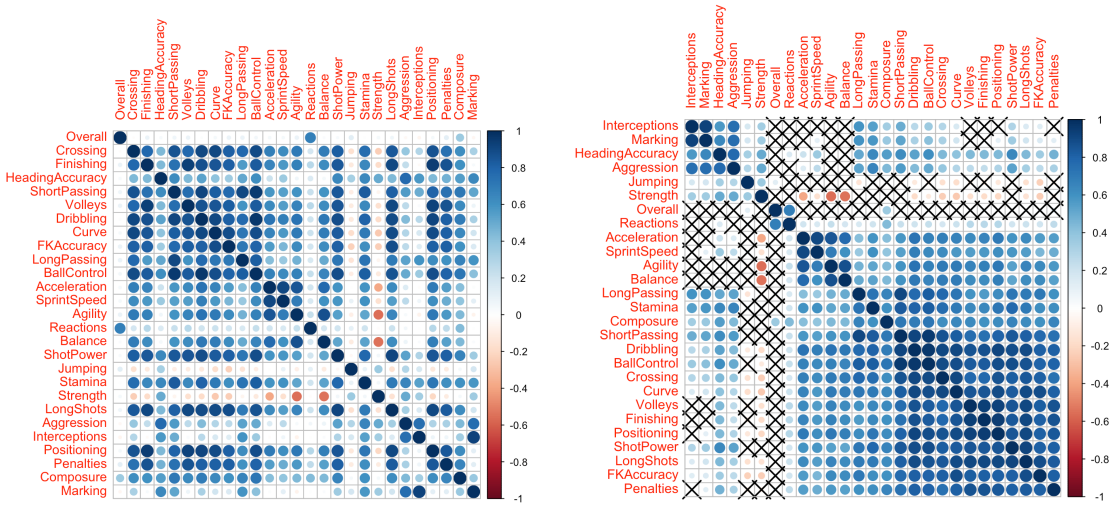
A second correlogram that examines the overall rating, value, wage, and position of a player was made. From this plot, it is apparent that the international reputation of players has the least influence on a player's value, wage, or overall rating.

3.3.4. Scatterplot

Scatterplots are useful for exploring data and analyzing where it clusters to find potential patterns to explore. A 4-dimensional scatterplot was created using the plot_ly function from the plotly library. The resulting plot is an interactive scatterplot that can be manipulated by clicking and dragging around the graph to change its orientation and therefore view the plotted points from various angles (Figure 6). Unfortunately, the interactive capability of the plot is not usable in image form. The data plotted was a subset of the FIFA19 data that consisted of the value, wage, overall rating, and position of each player.



Figures 4a(left) & 4b(right)



Figures 5a(left) & 5b(right)

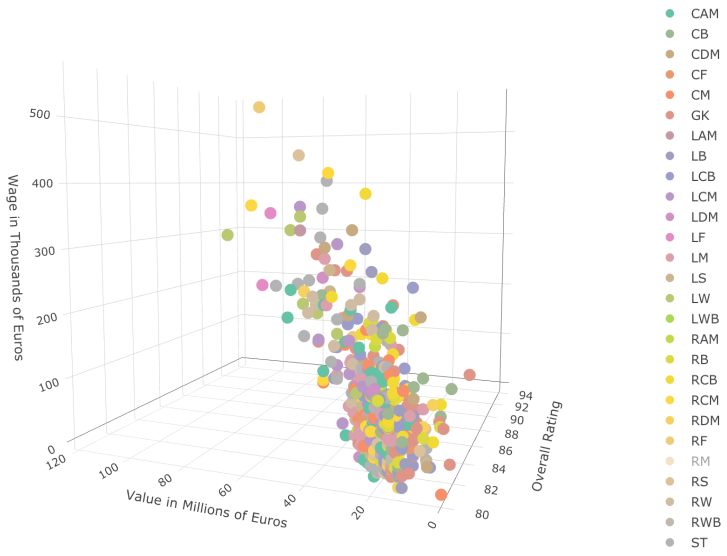


Figure 6

4. Medicare

This dataset was acquired from Kaggle and details patient demographics and other healthcare facility statistics for facilities that had Medicare beneficiaries in 2015. It is comprised of 41 dimensions and 15,026 records. Initially, a dataset on different gene expressions was going to be used for exploration, but it proved to be difficult to create visualizations of.

4.1. Exploration Questions

For this dataset, a more experimental approach was taken where visualizations were made in order to find items in the data to explore. This was done to test whether extracting further exploratory questions from plots is possible.

4.2. Preprocessing

Dimensional reduction was performed by removing irrelevant dimensions. In order to reduce the number of records to create interpretal visualizations, observations with

any NA values were removed. After this, the dataset was reduced to 21 dimensions and 48 records.

The columns detailing the percentage of patients with various pre-existing conditions was converted into the number of patients with the pre-existing condition to assist in visualization interpretability. Several columns were renamed as they were excessively long and would interfere with the readability of any visualizations created. This was done through piping and using the rename function from the tidyr library.

4.3. Visualizations

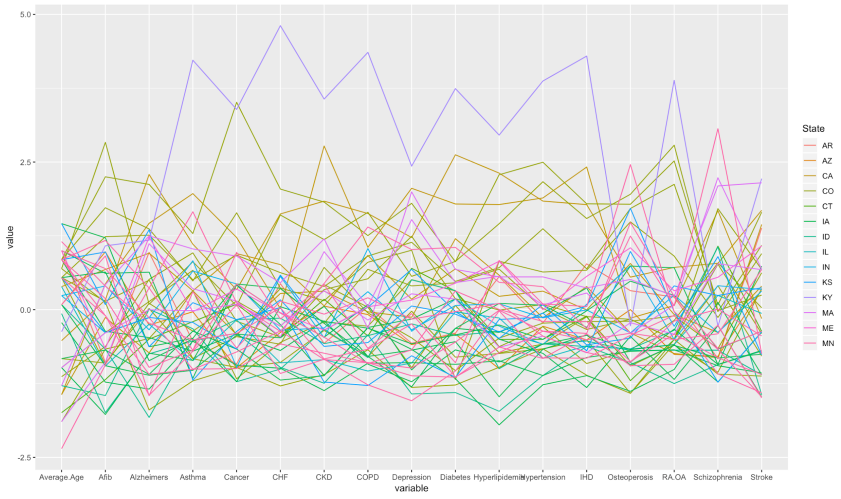
4.3.1. Parallel Coordinates

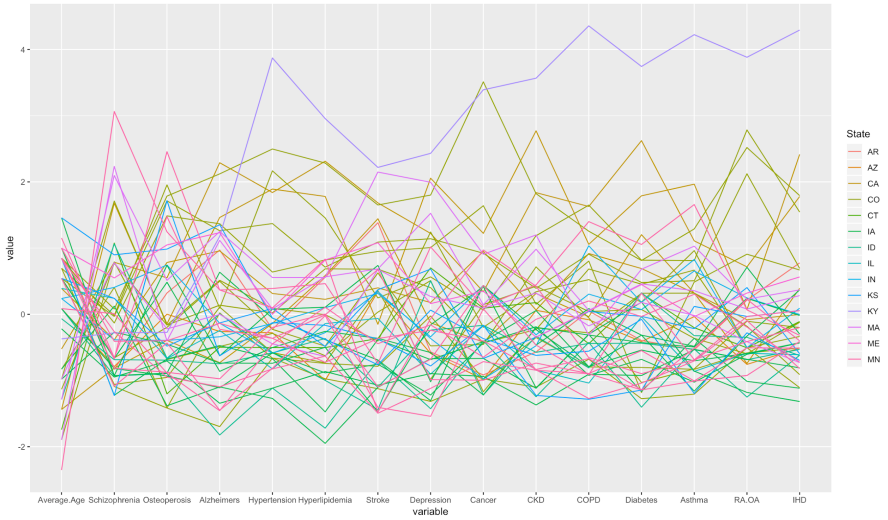
An initial parallel coordinate visualization was created using the ggparcoord function from the Ggally library, an extension of ggplot2 (Figure 7a). The dimensions were reordered to reduce the number of outliers between dimensions (Figure 7b). From this, it can be observed that an observation from the state of Kentucky is an outlier to the dataset. However, due to the fact that there are already so few records, it was not removed.

4.3.2. Correlograms

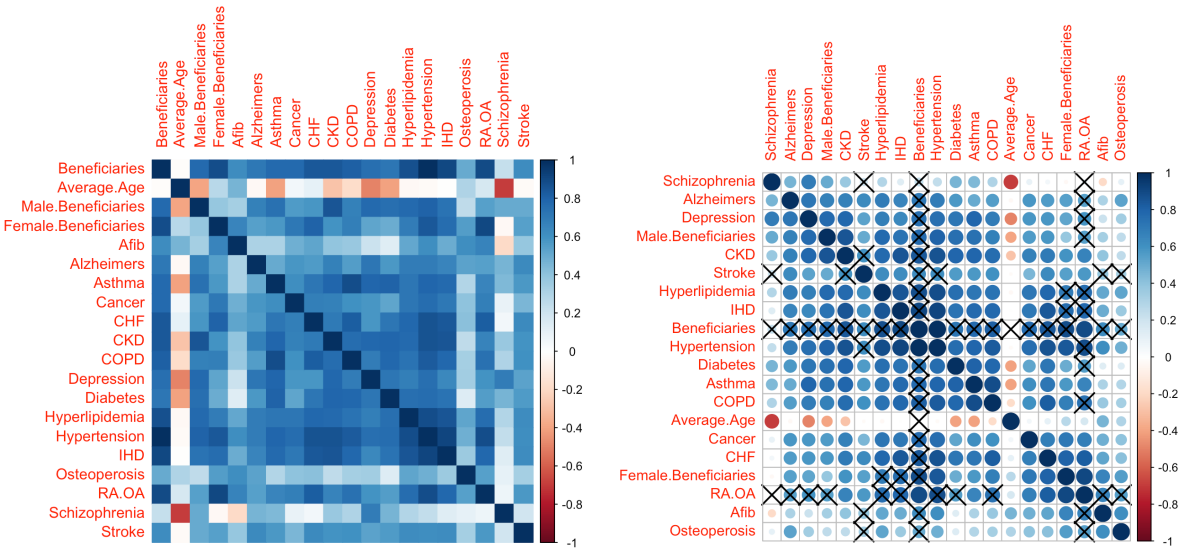
A correlogram of the data was created using the corrplot library to explore possible relationships between dimensions (Figure 8a). Interestingly, the average age of patients had a negative correlation with male patients but a positive correlation with female patients. Considering that the range of average age of patients for the analyzed facilities ranges from 69 to 92, the data could possibly show that healthcare facilities that have patients that use Medicare tend to care for older women. Of course, this requires further analysis.

A second correlogram showing the significance of correlations with the dimensions reordered was created using the same methods detailed in section 3.3.2 (Figure 8b. From this, it can be inferred that facilities with more patients on Medicare had larger numbers of Medicare beneficiaries that had hypertension. Additionally, there is a strong correlation between beneficiaries that have schizophrenia and beneficiaries that have depression. It is difficult to make an inference regarding any relationship from this observation alone. Further research and exploration will need to be conducted.





Figures 7a(top) & 7b(bottom)

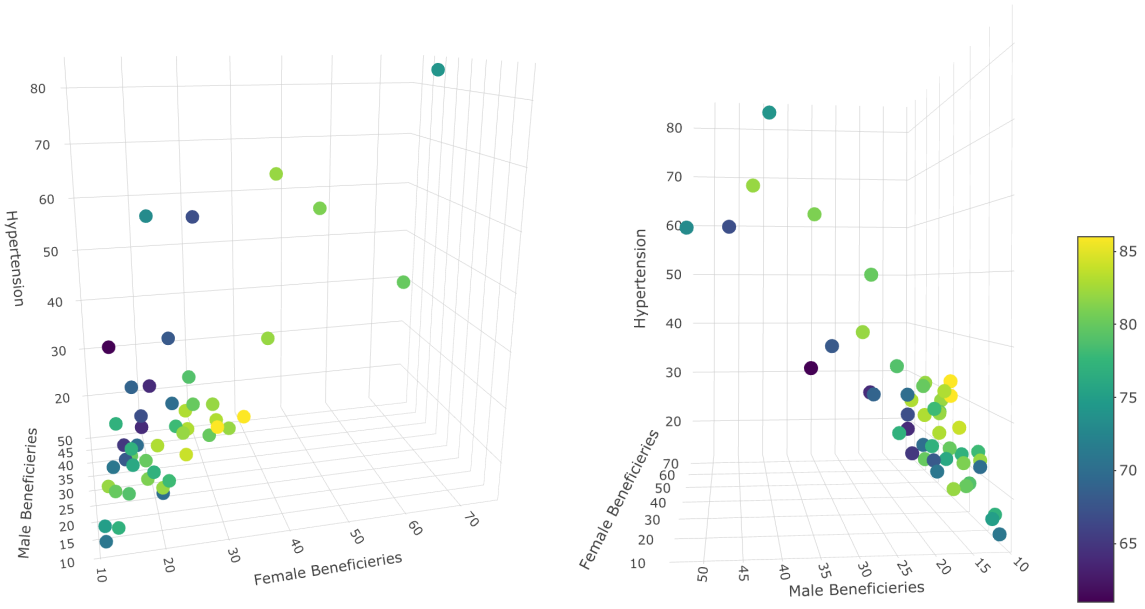


Figures 8a(left) & 8b(right)

4.3.3. Scatterplot

A 4-dimensional scatterplot exploring the relationship between male and female beneficiaries and the number of beneficiaries with hypertension (Figures 9a-b). This was done using the plotly library. Like the scatterplot generated for the FIFA19 data, this plot is interactive but cannot be viewed interactively unless it is compiled through R or viewed in a webpage. Multiple views of the plot have been included in order to observe the data. From this, it can be observed that the number of patients with hypertension increases as the proportion of female to male beneficiaries increases.

A second 4-dimensional scatterplot was created to further explore the relationship between schizophrenia and depression discovered in our exploratory correlograms from section 4.3.2 (Figures 10a-c). From this, it is still difficult to infer whether there is a clear correlation between schizophrenia and depression in Medicare beneficiaries. As mentioned earlier, further research regarding links between schizophrenia and depression will need to be conducted in order to determine whether there is a relationship between the two dimensions.



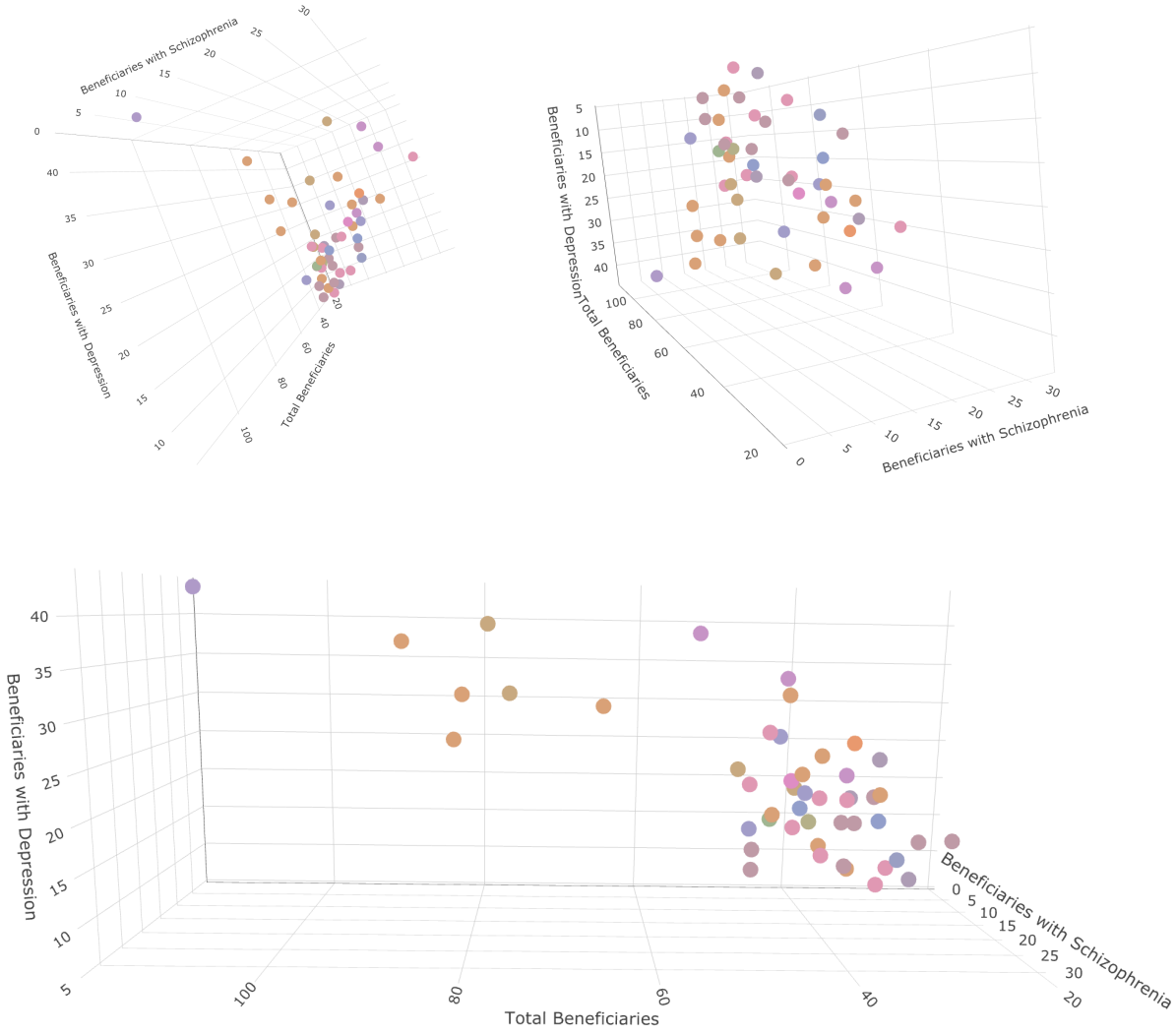
Figures 9a(left) & 9b(right)

5. Conclusion

When analyzing high dimensional data, using good preprocessing methods is key to unlocking patterns and relationships between dimensions in the data. Without this, it is nearly impossible to create visually appealing and easily understandable visualizations.

Using preprocessing methods such as clutter reduction and dimension reduction are vital to high-dimensional data successful exploration. As detailed in this paper, by using such methods, it is possible to extract questions for further analysis by exploring the data by creating visualizations of it such as parallel coordinates and correlograms.

Multi-dimensional visualizations are useful in successfully exploring high dimensional data and creating interpretable visualizations from it as shown with the use of 4-dimensional scatterplots to visualize relationships of interest between dimensions. By using these techniques, it is possible to create visualizations from high dimensional data that represent the data well and that can also be used for analysis of various exploratory questions.



Figures 10a(top left), 10b(topright), & 10c(bottom)

6. References

[1]W. Peng, M. Ward and E. Rundensteiner, "Clutter Reduction in Multi-Dimensional Data Visualization Using Dimension Reordering", *Digitalcommons.wpi.edu*, 2004. [Online]. Available: https://digitalcommons.wpi.edu/cgi/viewcontent.cgi?article=1071&context=computerscience-pubs. [Accessed: 23- Feb- 2019].

[2]A. Artero, M. de Oliveira and H. Levkowitz, "Enhanced High Dimensional Data Visualization through Dimension Reduction and Attribute Arrangement", *Tenth International Conference on Information Visualisation* (IV'06). Available: https://ieeexplore.ieee.org/abstract/document/1648337. [Accessed 13 February 2019].

[3]I. Fodor, "A survey of dimension reduction techniques", *E-reports-ext.llnl.gov*, 2002. [Online]. Available:

https://e-reports-ext.llnl.gov/pdf/240921.pdf. [Accessed: 02- Apr- 2019].

[4]D. Sarkar, "The Art of Effective Visualization of Multi-dimensional Data", *Towards Data Science*, 2018. [Online]. Available: https://towardsdatascience.com/the-art-of-effective-visualization-of-multi-dimensional-data-6c7202990c57. [Accessed: 26- Mar- 2019].

[5]"Parallel Coordinate Plots", *Homepage.stat.uiowa.edu*. [Online]. Available: http://homepage.stat.uiowa.edu/~luke/classes/STAT4580/parcor.html#australian-crabs-in-parallel-coordinates. [Accessed: 06- Apr- 2019].

[6]"Parallel Coordinates Plot", *Plot.ly*. [Online]. Available: https://plot.ly/r/parallel-coordinates-plot/. [Accessed: 02- Apr- 2019].

[7]"Visualize correlation matrix using correlogram - Easy Guides - Wiki - STHDA", *Sthda.com*. [Online]. Available: http://www.sthda.com/english/wiki/visualize-correlation-matrix-using-correlogram. [Accessed: 26- Apr- 2019].

[8]"FIFA 19 complete player dataset", *Kaggle.com*, 2019. [Online]. Available: https://www.kaggle.com/karangadiya/fifa19/version/4. [Accessed: 30- Mar- 2019].

[9]"Medicare Skilled Nursing Facility Provider Reports", *Kaggle.com*, 2019. [Online]. Available: https://www.kaggle.com/cms/medicare-skilled-nursing-facility-provider-reports. [Accessed: 10- Apr- 2019].