Designing Plots for Showing Causes of Strokes

Taoqi Yang, Adithya Reji AND Cyrus Li

Yang.4292, reji.3, li.10060

Abstract— We implemented a user interaction-focused design, providing a tool to help medical relevant researchers find potential indicators of strokes.

Index Terms— 3d visualization, parallel coordinates, pairwise scatter plot, medical science, Strokes

1. INTRODUCTION

We create a visual design for the dataset which is used by researchers to find relationships between several factors with stroke cases, select indicators for further prediction of stroke patients. For this design, we want to find the patterns of the distribution of stroke among the patients and which factors greatly contribute to the result of the stroke. Because we are not experts in the medical area, we want the user themselves to select which factors are important and should be shown in the graph.

1.1 Data Background

Link to data:
https://www.kaggle.com/fedesoriano/stroke-prediction-dataset
The data comes from Kaggle.com by an author named @fedesoriano. This dataset was compiled by this author for educational purposes. For the original purpose, it was used for machine learning training in which all other attributes were used to predict the result labels of the stroke. According to the World Health Organization, stroke is the 2nd leading cause of death globally which accounts for eleven percent of total death. The data would want to be used by doctors and medical students primarily to better understand strokes and be able to help their patients by giving them better advice and treatments. This dataset could also be used by the general public to understand what aspects of one's health can result in strokes. For this same reason, the reader could also benefit from understanding this dataset.

1.2 Data details

· Id: the identification of the patient which is the integer values

· Gender: the gender identifier of the patients which are text values of either 'male' or 'female'

· Age: the age of the patients which are integer values ranging from 0 to 100.

· Hypertension: the indicator variables about whether or not patients have hypertension with 1 as the answer of 'yes' and 0 as the answer of no.

· Heart disease: the indictor variable of whether or not patients have heart-related diseases. It has binary values with 1 as the answer of 'yes' and 0 as the answer of 'no'

- Ever_married: the indictor variable of whether or not patients have been married. It has two text values of ‘yes’ and ‘no’.
- Work_type: the indictor variable about the job types of the patients with three values of ‘self-employed’, ‘private’ and ‘Govt_job’/
- Urban: the indictor variable of the living area of patients with two text values of ‘urban’ and ‘rural’.
- avg_glucose_level: the indicator of the glucose level of the patients with digit values of range 0 to 300.
- Bmi: the body mass index of patients of digit values
- Smoking_status: the indictor variable of whether or not patients smoke. It has multiple text values of ‘formerly smoke’, ‘never smoke’, ‘smokes’, and ‘unknown’.
- Stroke: the indictor variable of whether or not patients get the stroke with 1 as the answer of ‘yes’ and 0 as the answer of ‘no’.

2 RELATED WORK

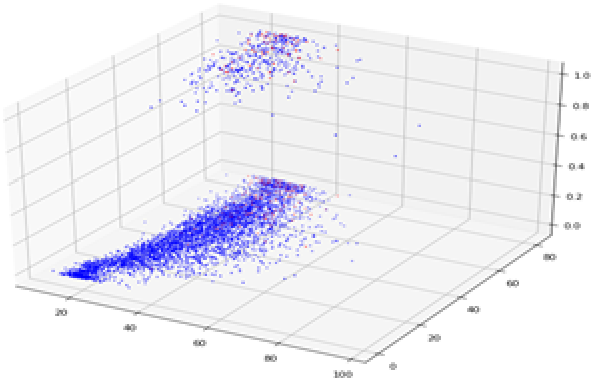
Here are the three design approaches:

The first design will be a 3-d plot and include six important attributes from the data.

The second design will cover all the information for the data and build an interface to let users choose whatever information they want to see from the data.

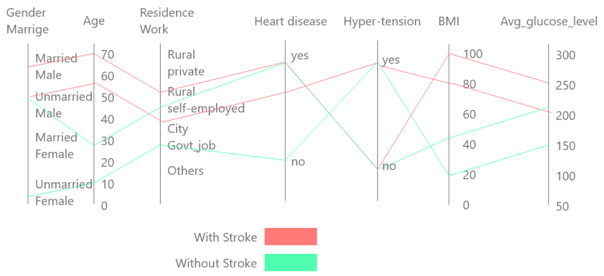
The third design will be based on parallel coordinates and will also cover all the information from the data.

2.1 Design solution 1



In this visualization, three axes represent the value of age, BMI, and hypertension. We do this because these three attributes are most important in predicting the result value of stroke. The shape of the points represents the gender: males are triangles and females are circles. Smokers are shaded. Patients with heart diseases are represented with a bigger circle.

2.2 Design solution 2

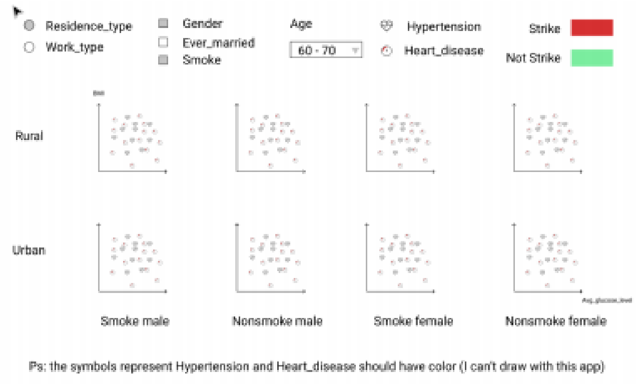


In this visualization, we use parallel coordinates. Every attribute is described as an axis with its values along the axis. Strike or not is presented by color for easy distinguishment. Users can focus on those causes they mind in this graph and find the result with the red/green density.

The demo graph only draws several lines but the final version will be complex and full of lines.

Users can specify the range of specific attributes, all the data lines with such attributes in the selected range will be colored and easy for users to differentiated.

2.3 Design solution 3



Ps: the symbols represent Hypertension and Heart_disease should have color (I can't draw with this app)

The dataset is composed of 11 components (not counting ID). What we want to present is an interactive graph where users can choose which they want to see from given groups.

From the visualization, users can see the distribution of data in whatever subgroups. Even the relationship between two attributes can also be seen.

We use colors to represent the stroke label so the result of the patients can be easily differentiated. Using Pairwise scatter plot. The matrix viewing will be useful in multiple dimensional data visualization

3 IMPLEMENTATIONS

For the 3d visualization, python Matplot lib is mainly used for graph generation.
The second and third visualization uses Javascript to implement the project. Plotly.js is used for graph generation and React.js is used to create the user interface.

4 TIMELINE

Week1: (3/29-4/2)

Discussed the designs.
Group members wrote the proposal together

Week 2: (4/5 - 4/9)
Cyrus completed the parallel coordinates graph.
Taoqi completed the first version of 3d visualization
Reji prepared the slide and demo for the first milestone.

Week 3: (4/12 - 4/16)
Cyrus completed the pairwise scatter plot and using HTML/js to implemented and deployed our works on Heroku.
Taoqi completed the second version of 3d visualization.
Reji prepared a delicate slide and presented it in class.

5 RESULT AND DISCUSSION

5.1 Result

We selected elements that are necessary for case prediction and which are optional. We choose interactive graphs for a better user experience. We decided all our figures are composed of basic simple graphs which will be helpful for easier understanding and usage. We chose to use color, shape for different elements' distinguishing.

5.2 Discussion

We have presented 3 visual implementations in class and received feedback from our professor Chen.

One drawback of our web page implementation is that we didn't combine all graphs together. But there are some technical issues with the sizing of the graph in Plotly, we need more time to figure it out.

Because we divided works to each member, not all graphs are implemented with the same tools, they don't work well when put together.

The python version 3d graph can't change the view as our expectation. It also has some overlay problems.

The pairwise scatter plot has problems in layout and white space in the page will influence the user experience.

6 CONCLUSION

This project can lead to a better summary of the contribution of different factors to the result of the stroke. It provides a tool to customize data visuals with the interest of the user. We believe this will improve direction for the biologists/students for finding the correlations between their interesting factors and the stroke occurrence, which could improve their effectiveness in stroke diagnosis and analysis in the future.