How would you visualise your data?

COMP5048/4448 Assignment 2

Yiwei Wang
Student ID: *530777511*

Group Number: prac 05
Tutor’s name:Freya Singh

I. Task 1

The dataset employed in this report comprises 70 attributes and 1429 observations, encompassing demographic and mental health information of individuals residing in approximately two hundred villages located near a lake in western Kenya. The data was collected and organized by the Busara Center, a research organization, primarily through the surveys and questionnaires. The dataset encompasses a wide range of personal details, including wealth, family environment, and health-related factors, alongside mental health indicators such as the presence of depression.

The primary objective of the Busara Center's research initiative is to conduct an in-depth analysis of the mental health status of local residents and to find out the key d affecting their mental well-being.

For the purpose of this report, a subset of five attributes from the dataset has been selected for visualization and analysis. These attributes include:

1. 'Village' (nominal data): Each unique number corresponds to a distinct village.
2. 'Survey Date' (ordinal data): In order to to protect respondent privacy, survey years are indicated as 1961 and 1960, while the actual data collection took place in 2015.
3. 'Gender' (nominal data): A binary classification with '0' means male and '1' means female.
4. 'Asset Savings' (ratio data): Recorded as floating-point values, representing financial assets.
5. 'Depression' (nominal data): Binary classification where '0' means individuals without depression, and '1' means individuals with depression."

II. Task 2

This is because unlike other factors affecting people's well-being, such as wealth, which can be reflected more intuitively through data on various aspects of daily life, the mental health of the population usually involves more specialized and careful research and statistics, especially in backward areas. The mental health status of a population usually involves more specialized and careful research and statistics, especially when it comes to people in backward areas, because these people pay less attention to psychological problems, so the mental health status of these residents is usually overlooked. Therefore the group of beneficiaries of this dataset is quite large.

Firstly, there are some government practitioners, who can visualize the personal details of some local residents and their actual mental health status through this dataset. Secondly, there are researchers in the field of mental health, who can analyze the impact of different aspects of life on the mental health status of the population through this dataset. They can also use the results to further screen the people in the area who are at higher risk of developing depression, so that healthcare resources can be better adjusted and allocated. This will allow for a more targeted approach to helping the local population when mental healthcare resources are in short supply.

III. Task 3

The following visualizations were derived from two separate studies related to this dataset.

Study 1

```
A variável target do treino possui 20.32% de positivos.
```

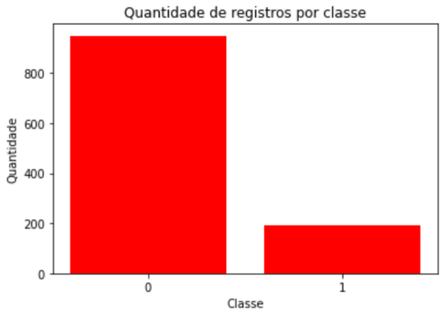


Figure 3.1

The histogram above counts the number of people in the dataset who are depressed or not. Since “depression” falls into the nominal category of data, the number of both is counted using the bar chart classification.

Study 2

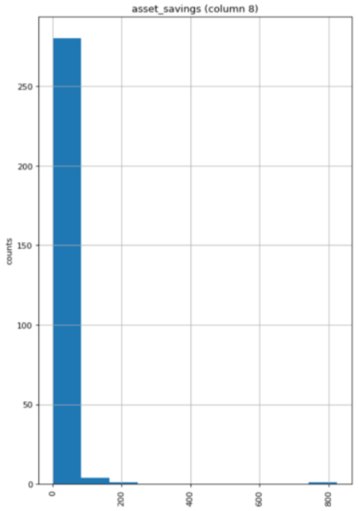


Figure 3.2

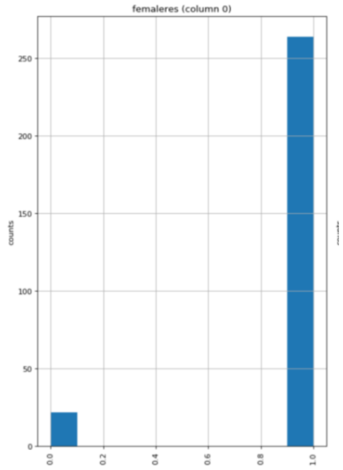


Figure 3.3

The two histograms above count the distribution of the number of properties and gender in the dataset, respectively.

The first histogram counts the properties of the respondents. Since the “asset savings” data belongs to the ratio category, it is also applicable to the histogram statistics.

The second histogram counts the gender ratio of the respondents. Since gender data belongs to the NOMINAL category, the histogram is used to count the number of different categories.

As can be seen from the visualization results of the above two related studies, histogram is the commonly used way to visualize this dataset. Researchers usually count and display the distribution of quantities for a single column.

IV. TASK 4

The histogram of the first study explains the question "What is the percentage of people with depression among those surveyed?" This question, as can be seen by the vertical coordinate of the histogram, shows that the number of people investigating this health is about 4.5-5 times higher than the number of people suffering from depression.

The two histograms of the second study explain the question "What is the gender ratio and property profile of the surveyed population?". The first histogram shows that the majority of the survey population is economically disadvantaged and that the gap between rich and poor is large (there is also the possibility of outliers in the data). The second histogram shows that the majority of the survey population is female, with only a small number of males participating in the survey.

V. TASK 5

The problem with the visualization of the first study was that it did not present the data comprehensively enough and did not show valuable information about the dataset, e.g. does property have an effect on a person's mental state? Are people of different genders in the area sick in the same proportion? And so on. The problem with the second study's visualization was that the information was presented in a fragmented manner and did not allow for easy comparisons to see the correlations between the various types of data.

VI. TASK 6

The designed SOG is shown as follow.

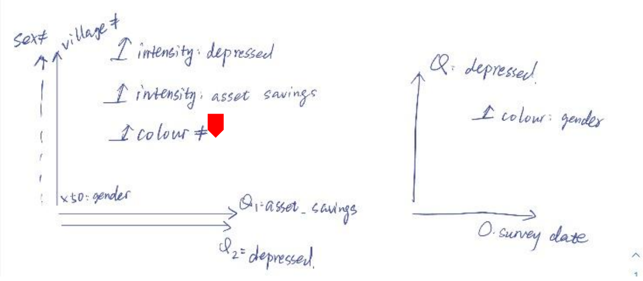
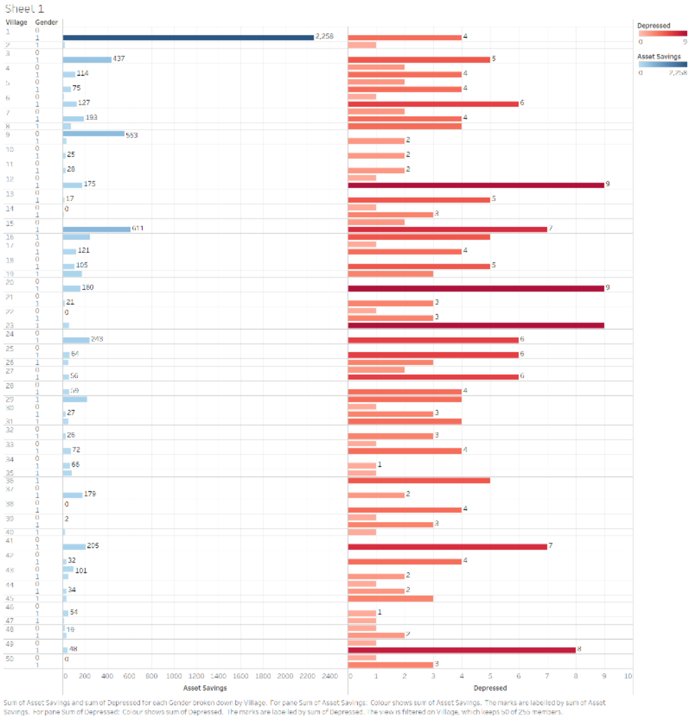


Figure 6.1

It is learned from other disciplines that the greater the wealth, the higher the person's ability to withstand mental illness. Therefore, I would like to compare the amount of wealth owned and the number of people who are sick in different villages through visualization images, so as to qualitatively observe whether wealth has an impact on people's mental health status. Also, I would like to observe if there is a significant difference in wealth by gender in the area. Therefore, "gender" and "village" are the variables I use to categorize my observations, and "depression" and "asset savings "depression" and "asset savings" are the quantitative variables I want to count. At the same time, I would like to distinguish the number of people with "depression" and the number of people with "asset savings" with different colors in the image, and I would like to use intensity to visualize the number of people of a certain gender in a certain place and the number of people with a certain disease. I want to use intensity to visualize how many people of a certain gender in a certain place are sick and how much wealth they have, so I got the above SOG design.

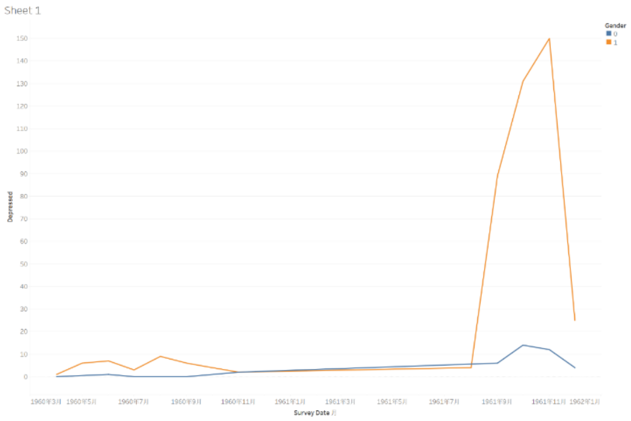
VII. TASK 7

The following visualizations are derived from task 6.



Sum of Asset Savings and sum of Depressed for each Gender broken down by Village. For pane Sum of Asset Savings: Colour shows sum of Asset Savings. The marks are labelled by sum of Asset Savings. For pane Sum of Depressed: Colour shows sum of Depressed. The marks are labelled by sum of Depressed. The view is filtered on Village, which keeps 50 of 255 members.

Figure 7.1



The trend of sum of Depressed for Survey Date Month. Colour shows details about Gender.

Figure 7.2

VIII. Task 8

According to my design, bar charts and line graphs are more appropriate to show the visualization options of the data. In this image, I use different colors to differentiate between depressed and asset saving data, and label the top of the bar chart with the specific quantities, which makes the comparison more intuitive and can make the data more easily accessible to the applicant.

IX. Task 9

The equivalent SOG of task 6 is shown as follow.

1

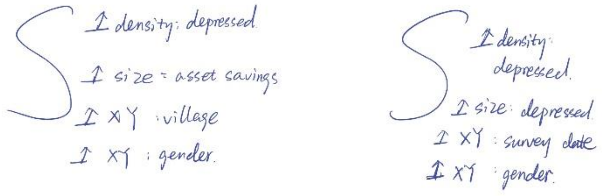
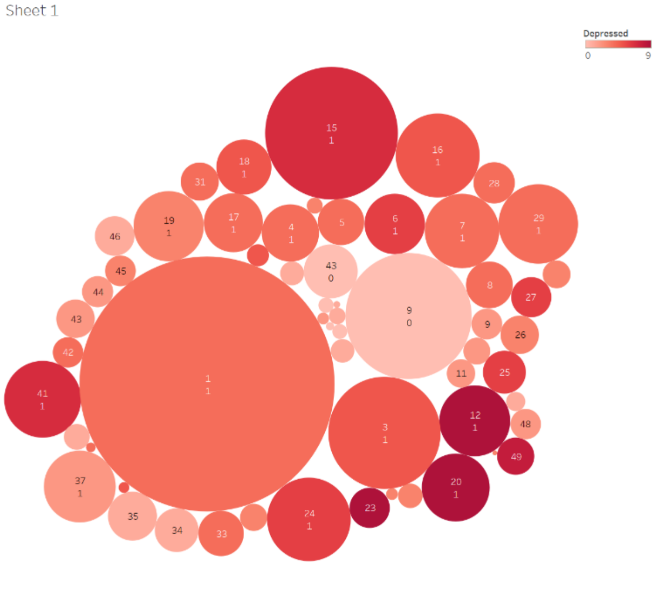


Figure 9.1

Since I used axes for visualization in the above image, in task9 I wanted to present the data in a more free-form image. So I differentiated the depressed and asset savings using intensity and size, and displayed the village and gender information as labels in the image, resulting in the above SOG design image.

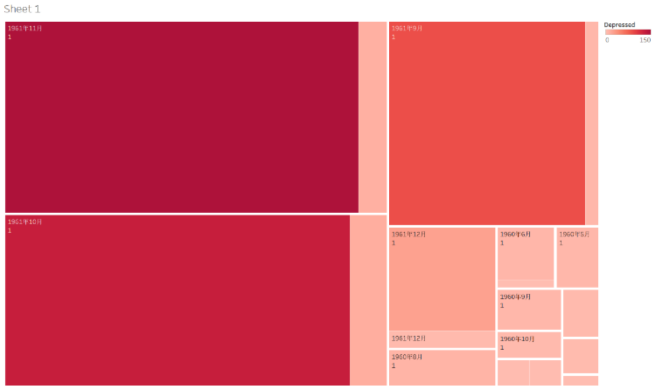
X. Task 10

The following visualizations are derived from task 9.



Village and Gender. Colour shows sum of Depressed. Size shows sum of Asset Savings. The marks are labelled by Village and Gender. Details are shown for Gender. The view is filtered on Village, which keeps 50 of 255 members.

Figure 10.1



Survey Date Month and Gender. Colour shows sum of Depressed. Size shows sum of Depressed. The marks are labelled by Survey Date Month and Gender.

Figure 10.2

In the two new visualization images, I use bubble charts and heat maps to show the original quantitative relationships. In the first image, the color and size of the bubbles represent the "depressed" and "asset saving" information respectively, which achieves the "simultaneous observation" that I wanted to achieve in Task 6. wealth and mental health at the same time" in task 6. At the same time, the information about village and gender is displayed in the form of labels, which is clear at a glance. In the second graph, I use a heat map to show the data. The color and size of the graph indicates the depressed information, and the time of the survey is shown as a label.

References

[1] https://www.kaggle.com/datasets/francispython/b-depression

[2] https://zindi.africa/competitions/busara-mental-health-prediction-challenge/data

[3] http s://www.kaggle.com/code/kerneler/starter-b-depression-00ffced3-4

[4] https://www.kaggle.com/code/patryckaug/depression-analysis

[5] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.