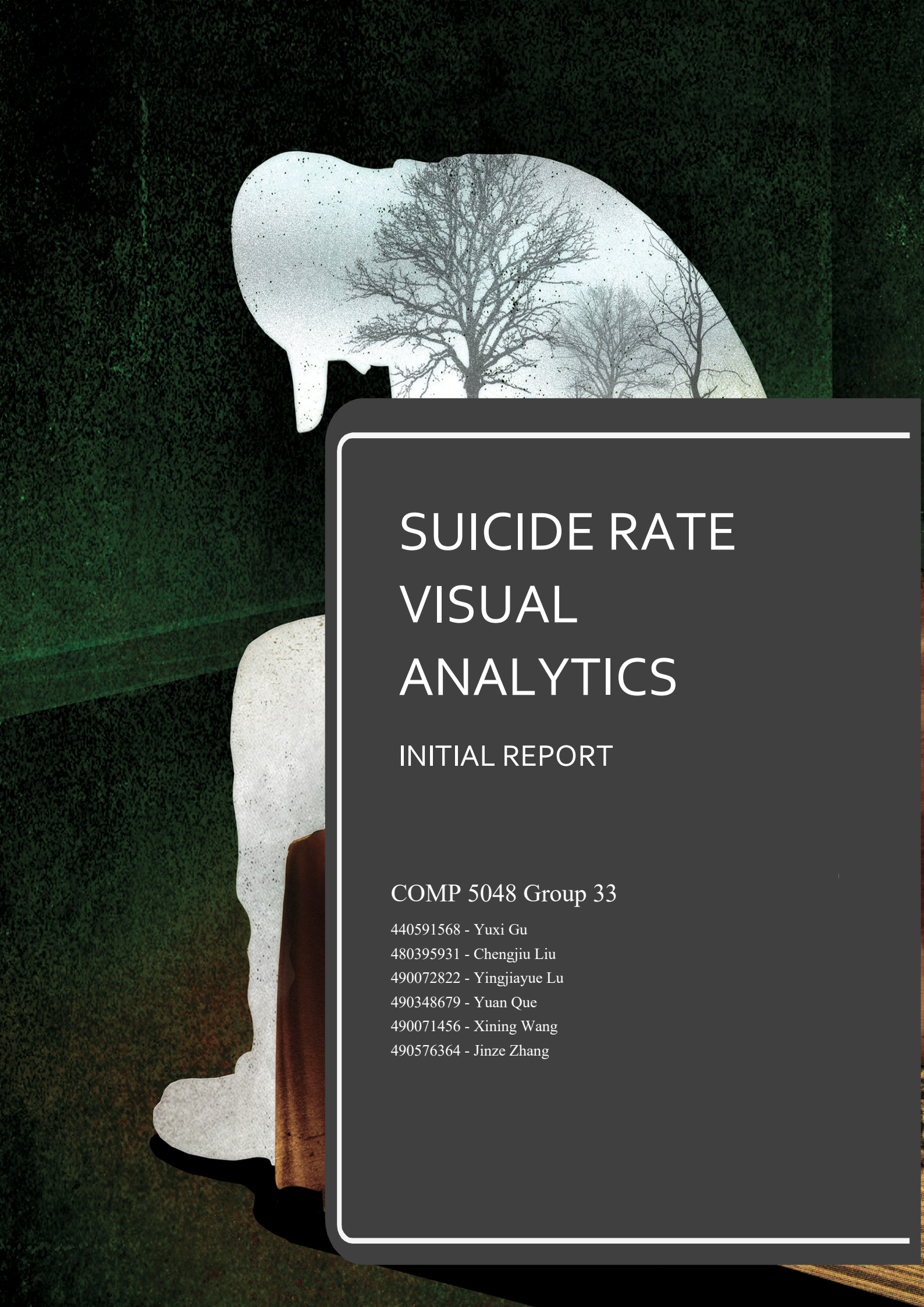

SUICIDE RATE VISUAL ANALYTICS

INITIAL REPORT

COMP 5048 Group 33

440591568 - Yuxi Gu
480395931 - Chengjiu Liu
490072822 - Yingjiayue Lu
490348679 - Yuan Que
490071456 - Xining Wang
490576364 - Jinze Zhang

1. Introduction

1.1 Dataset

1.1.1 Features in Dataset

This data visualization analysis suicide rate from 1985 to 2016 for 101 different countries in the world. This dataset is divided the group by 5 different age intervals and base on people's divided into 6 different generation. Each data also include a number of people contained in each country that year. This dataset also provides economical and development numerical index which GDP (Gross Domestic Product) for year, GDP per capita and HDI (Human Development Index)for the year that measures life expectancy, income and education.

Population size is number of people contained in each area that year. Number of Suicides is a number of suicides have the same gender, generation and area in the subsample. Suicides per 100k people are the number of suicides divided by the population size and multiplied by 100,000. This data is an index ratio of the number of suicide divide same subsample population which give normalized data in order to compare with another subsample. Data Cleaning and Transforming

1.1.2 Data Cleaning and Transforming

In order to improve the quality of data, data cleaning deals with detecting and removing errors and inconsistencies from data. This dataset also has incomplete or deficiency, such as the dataset not include all the countries in the world especially for many countries in Asia and Africa are not considered. Dataset fields do not have much deficiency, 70% of HDI for year data is missing. Most of the data are reliable, each instance includes features in the data set. Data cleaning need to modify or remove data according to requirements. One of the methods to solve this problem is using a mean value from all the available data to substitute the losing value.

1.2 Tasks

1.2.1 Map

For the first task, we want to know the suicide rate of male and female in different countries. Therefore, we can have a basic understanding of the condition of suicide among countries. We will use the data of countries, suicides per 100k, gender in the dataset.

1.2.2 Timeline

Timeline is to show the trends of suicide rate of different age intervals and generations.

1.2.3 Dynamic and Interactive Visualizations

We want to create a dynamic movement visualization which may display the changes in suicides per 100k population in different age stages in various countries. As we found that fluctuations in GDP per capital may have an impact on the suicide number, so we add it in the graph as x-coordinate. Because the suicides per 100k population will change as other indicators change, we use it as y-coordinate. As the dataset has lost a lot of HDI data, we ignore this index at this time.

1.2.4 Small Multiple

The fourth task is to gain an insight of which countries have the significant increasing trend of suicide rate in contrast with those have the significant decreasing trend between 1985 and 2016. Furthermore, we will look deep into these countries and analysis what is the primary reason or similar features for a significant increasing trend among these countries.

1.2.5 Heat Map

Heat map describes the correlations among few variables that impact suicide rates. Correlation is a statistical metric for showing the extent of relationship between contained index. Normally, heat map use python to implement the correlation matrix, utilize gradient color on the right side to implement the number of correlation.

For tasking, we need to identify the categorical data as country, year, gender, as well as numeric data as suicides /100k, age/generation, sex, population, GDP for year and GDP per

capita , suicide number, continent name. In this case, we choose location dataset (countries) to find the main Influencing factor that determine suicide rate in different countries, additionally provide two heat maps with male and female to distinguish the most effected factors leading to suicide of two genders.

1.2.6 Simpler Graph

We want to create a simpler graph to display two or three important indexes in a more painstaking way. It might be a dashboard contains two graphs, like a column chart and a line chart. From several complicated graphs showed before, there is no clear connection between two genders and other datasets. We can use line chart to display the relationship between gender and suicides per 100k population or gender and suicides number. After that, we may discover whether different sex affects people's suicide.

2. Design and Approaches

2.1 Analysis

In order to realize tasks, we separate dataset fields into difference visualize graphs and analysis relationship between columns data. We analysis dataset base on the different tasks we will implement later.

2.1.1 Map

A choropleth map displays divided geographical areas or regions that are colored in relation to a numeric variable. It allows studying how variable evaluates along with a territory which will give a more intuitive and visual display of the situation in different countries. We can more easily find countries with special features for further analysis. Since we would like to know the suicide rate by different gender, we may divide each country into two areas and use a different colors to indicate male and female.

2.1.2 Timeline

Visualize the trend is to show the degrees of different times, we create a time series of staked bars to show the trends of suicide rate of different ages. The staked bars display the evolution of the value of different ages on the same graphic. The values of each group are displayed on top of each other, which allows checking on the same figure the evolution of both the total of a numeric variable and the importance of each group. The generation will be shown with different colors together with the staked bars.

2.1.3 Dynamic and Interactive Visualizations

For the dynamic movement visualization itself, it could provide different trend of changes in suicides per 100k population through previous times. Compared with other kinds of figures, it is more intuitive to display over five related indicators. Also, we may find some obvious features when running the dynamic graph as the time periods are distinctive.

2.1.4 Small Multiple

Since there are 101 countries in this dataset, it is important to note that drawing all the trends in the single graph might not truly representative, which come across several problems such as a substantial amount of crossing lines, overly complex structure, increasing the risk of misleading the audience. In order to derive the countries with the similar pattern in terms of suicide rate trend, small multiple offer some valuable features to solve this problem. Small multiple use the same basic chart to display difference slice of a dataset. As a result, the audience can quickly learn from an individual chart and apply this knowledge as they scan the rest of the small charts. This reduce the audience's effort from understanding what the chart represents. In addition, it enables the comparison across countries and hence reveal the pattern of different trends. It conveys vast amounts of information in a small, well contained visualization.

2.1.5 Heat Map

The reason why we use the heat map to visualize suicide rate is that heat map generally means the point data analysis by calculating Kernel Density Estimation.

Kernel Density Analysis (KDE) visualizes features to achieve the translate from discrete object models to continuous-field models by calculating the density around the features, create a smooth surface for feature pattern detection and discovery in the end. Based on KDE, dot data analysis can be used to describe any type of incident data, because each event can be abstracted into a spatial location point. We can use point data to analyze the discipline behind the data, which is called 'point mode'. Point mode is ubiquitous in nature and economic society. Through analysis, we can make point data into point information, which can better understand the spatial point process and accurately find the discipline behind the space point.

2.2 Visualization

2.2.1 Maps

We visualize the data on a world map. In the map view， we divided each country into two areas with blue and red to show the different gender. Then the suicide rate in each country is expressed in shades of the color. The countries lack of data is shown in grey.

2.2.2 Timeline

We will implement a time series of staked bars. The x-axis will show the years from 1985 to 2016 and the y-axis will show the suicide rate. Each staked bars will show the rate of a different group of ages in a different color. The suicide rate of generations will also be shown as a line plot in the same picture.

2.2.3 Dynamic and Interactive Visualizations

We suppose that this dynamic movement visualization has three more outstanding indicators—suicides per 100k population as y-axis, GDP per capital as x-axis and the timeline from 1985 to 2016 at the bottom of the graph. We plotted the logarithmic GDP because it can reduce the absolute value of the data for easier calculation. And we use six different colors to represent different age groups of people as here are six different age stages from the dataset. The more population a country has, the larger radius its corresponding bubbles in the graph has. When we run the graph, we may discover most suicide cases

happened at which point and country, that is to see, we are able to analysis the background reasons that lead to this phenomenon.

2.2.4 Small Multiple

For our case, we decide to draw 12 small charts with respect to the trend of suicide rate, in which will be separated into two clusters. One of the clusters display the 6 steepest increasing trends, while the other display the steepest decreasing trends. The y-axis is suicide per 100k and the x-axis denote year from 1985 to 2016.

2.2.5 Heat Map

Heat map use projection coordinates to analysis, so it's a square graph contains (number of index)**2 little squares, with density above. Gradient color chart on the right side defines correlations in diversity extents, for example, deep color represent strong connection while light color express there is raw relationship among two factors.

2.2.6 Simpler Graph

The x-axis is suicide number and the y-axis are year, male and female are two lines in this graph. We can see the changes in the number of different sexes throughout different time period.

3. Implementation

We proceed analyzing dataset which helps us build relationships between multiple columns data. And into the visualization part, we introduce which type of visualization we will build. And next is detail methods we will use visualization software, open-source programming packages to achieve tasks.

3.1 Map

First, we need to process the data, combine data from different ages, then calculate the average suicide rate of all age of different years. Creating a table which contains the average suicide rate, country and gender. Then plot the data to the map to visualize. We would like to use a web-based tool called Flourish. Since this application accepted more geography choice and more feature options, it will make us easier to implement.

3.2 Timeline

In the stage of the process the data, we need to combine the data from different countries and gender by age groups and years, then use the combined suicide number and population to calculate the new suicide rate. We also need to combine the data by generations and calculate the new suicide rate of that year. We can simply use Excel to do the data process. For the visualization part, Tableau is a powerful data analytics tool to create the time series we need. it works well with Excel and can easily link with the data. Just after some Clicks and drags, we may create the layout we need.

3.3 Dynamic and Interactive Visualizations

We use python to draw this dynamic graph since we want to use some of the python libraries and write codes to present a better graph effect.

3.4 Small Multiple

In order to identify the significant trend, use R to generate hypothesis testing regarding of the linear relationship between variable years and suicide rate. Those p-values smaller than 0.05 are considered as significant linear trend. Then, the countries with significant trend should be included in the small multiples. R provide advance and decent graph drawing library, use ggplot to generate small multiple with smooth line across the scatter plots. The plot is extremely easy and fast.

3.5 Heat Map

Python has multi-functional libraries and functions to complete the drawing of heat map in a few line. Using NumPy package to input data set, and also can use Pandas package process dataset. Using color parameter ensure maximize and minimize range of value; matplotlib-colormap. Apply package function to build coordinate.

With python we can easily import Express to provide simply grammar for complex visualization graph, most drawings merely require to call one function. Additionally, changing currently parameter can be valuable to create extra heat map to analyze suicide rate multilayered.

3.6 Simpler Graph

We may use tableau to finish these graphs as it might be easier to drag the parameters we need. Also, the function of tableau is well-completed. After choosing the chart types we want, we could combine simpler graphs in a dashboard.

4. Evaluation

4.1 Evaluation of each graph:

1) Evaluation of map:
    a. Show the feature of data clearly
    b. Easy to tell the differences of suicide rates between countries
2) Evaluation of timeline
    a. Clarify key events and sequences
    b. Make connections between events
    c. No patterns emerge
3) Evaluation of dynamic & interactive visualizations:
    a. Provide different trend of changes in suicides rate in the past
    b. Show the trend clearly
4) Evaluation of heatmap:

a. Find the main Influencing factor of determine suicide rate in different countries
b. Smooth surface for feature pattern detection
c. Find out point mode and transfer it into information

5) Evaluation of small multiple:

a. Decrease crossings in each small graph
b. Easy to access information from whole graph
c. Easy to read
d. Identify significant trend

4.2 General evaluation rules:

a. Informative: usefulness, completeness, perceptibility, truthfulness, intuitiveness
b. Emotive: aesthetics, engagement
c. Easy to understand the structure and find patterns
d. Less crossings in the visualization
e. Reveal the hidden truth
f. Can predict the future

4.3 General evaluation method:

a. Survey
b. Analytic inspection
c. Empirical evaluation

5. Planning

Group member / ***Time***	**Gracie Gu**	**Chengjiu Liu**	**Yingjiayue Lu**	**Yuan Que**	**Xining Wang**	**Jinze Zhang**
Week7 (Initial Report)	Task, analysis, visualization and implementation of map and timeline	Task, analysis, visualization and implementation of small multiple	Task, analysis, visualization and implementation of dynamic &	Evolution and Weekly planning	Task, analysis, visualization and implementation of heatmap	Data set and integration report

			interactive visualization			
Week8	Data preprocessing of initial data	Calculate the PCA of the data set	Implement simple visualizations such as histogram and pie chart	Analyze the data set between age, generation, sex and suicide rate	Analyze the data set between suicide rates, gender and country	Analyze the data set between suicide rates, GDP, HDI, year
Week9	Draw heatmap using python	Draw small multiple	Draw dynamic & interactive visualization using plotly and python	Draw timeline using Tableau	Draw map using Gephi	Improvement of the above visualization and make a dashboard
Week10 (Presentation)	Make a PPT file and upload it	Evaluate heatmap	Evaluate small multiple	Evaluate dynamic & interactive visualization	Improve heatmap	Improve small multiple
Week11	Improve dynamic & interactive visualization	Evaluate timeline	Improve timeline	Evaluate Choropleth Map	Improve Choropleth Map	Integrations improvement
Week12 (Prepare Presentation**)**	Simple visualizations such as histogram and pie chart	Heatmap	Small multiple	Choropleth Map	Timeline	Dynamic & interactive visualization
Week13 (Final Report)	Data sets and Tasks, Aims and Contribution	Analysis and Visualization	Final Report Implementation	Final Report Evaluation	Results and Group meeting minutes	Discussion and integration report

6. References

N. Kerracher, J. Kennedy and K. Chalmers(2018) *Using a task classification in the visualisation design process for task understanding and abstraction: an empirical study*

Ruaty. (2018). Rates Overview 1985 to 2016 [dataset]. Retrieved from *https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016*

Samkian, A. and Greene, J. (2013, October 17). *Visualizing Process: How to Create a Stakeholder-friendly Graphic Timeline of Process Data*. Presented at the American Evaluation Society Evaluation 2013 Conference, Washington.

Williamson, T. and Long, A. (2005). *Qualitative data analysis using data displays' in Nurse Researcher*, 12(3): 7-19