STAT220 Introduction to Data Science
Final Project Technical Report

Sunny Kim, Joshua Song

Research Question

There are many interesting statistics of the OECD countries around the world, but two dataset we focused on were suicide rate and fertility rate. We have questions we want answers to using the two datasets for further analysis. How does suicide rate relate to the continent of the OECD countries? How does fertility rate relate to the continent of the OECD countries?

Data Selection

- **Dataset**

For our final project, we plan to look at the dataset that contains the OECD countries' suicide and fertility rate and create an interactive graph that displays the data through a shiny app. According to the official OECD(Organization for Economic Co-operation and Development) website, the definition of suicide rate and fertility rate are the following:

Suicide rate: Deaths deliberately initiated and performed by a person in the full knowledge or expectation of its fatal outcome.

Fertility rate: Total number of children that would be born to each woman if she were to live to the end of her childbearing years and give birth to children in alignment with the prevailing age-specific fertility rates.

The following are the datasets we used for our final project.

OECD suicide rate

https://data.oecd.org/healthstat/suicide-rates.htm

Citation: OECD (2023), Suicide rates (indicator). doi: 10.1787/a82f3459-en (Accessed on 02 March 2023)

OECD fertility rate

https://data.oecd.org/pop/fertility-rates.htm \

Citation: OECD (2023), Fertility rates (indicator). doi: 10.1787/8272fb01-en (Accessed on 02 March 2023)

The original dataset we downloaded from the official OECD(Organization for Economic Co-operation and Development) website includes the following columns:

Column Name	Explanation
Location	Country name
Indicator	What the value indicates (Suicide/Fertility)
Subject	What group of people the data is based on (Total/Men/Women)
Measure	How the data is measured
Frequency	How frequent the values are in terms of levels
Time	The year in which the data is collected
Value	Value of indicator
Flag.Codes	Flag Code for each country

- **Method/Tools**

In our analysis, the columns we use will be LOCATION, INDICATOR, TIME, and Value. We also wanted to modify the dataset so that we can also add a panel for the user to change the input of the animated graph. Following are some of the steps we went through to make the dataset into a format that we could use for our graphics.

1. *Delete unnecessary columns*: We deleted the columns "*Flag.Codes*", and "*Frequency*", because they contained a lot of NA values as well as unnecessary values that we did not need for our research questions.

2. *Format Data*:
    a. *Delete rows with NA values*: We deleted rows that had *Location* values that are "EU" or "OAVG". This is because they do not belong to any continent codes of the *countrycode* package. When we access the continents for these countries, they return a NA value, which is why we decided to remove these rows. We did this by adding a filter function that only filters the *Location* values that are not "EU" or "OAVG", and leave only the other values.
    b. *Filter data*: We used the pipe functions to filter, change, and add data to the dataset. Most of this process was to use the three letter code country names to access full country names or their continents. This is explained in more detail below under the *library(countrycode)* section. After accessing the library and the saved country codes, we added the corresponding continents to each row by using the *mutate* function.
    c. *Change column names*: We changed the column names so that it is easier for us and the user to understand what each column is showing. For example, we changed the column "*Perspective*", which referred to one out of three categories Total, Men, and Women, to "*Gender*". We did this by creating a new vector with a new set of column names and directly set it as the column names of the dataset.
    d. *Change order of rows*: We arranged the countries in alphabetical order in order to make it easier for users to find them when choosing their inputs.
    e. *Rename values*: We renamed the values in the *Subject* column so that we have a better idea of what each category indicates. We did this by using the *mutate* function on the dataset.
    f. *Print datatable*: We print the datatable onto the dataset tab using the *renderDataTable* function and the *datatable* function from the *DT* package.

library(countrycode)

countrycode is a built-in package that contains standardized information about countries such as country names, country codes, conversion between them, and various other country information. We used this package to access the continents each country is located in so that we can add the information as another column to our dataset. This step was required to allow the users to select

input as continents so that the graph can show data from countries that are located in the selected continents. We also used this package to change initial country names, which were in an “iso3c” format, to full names, so that it becomes more user-friendly.

Shiny Application

- **Method/Tools**

library(shinythemes)

shinythemes is a built-in package to select a theme for the shiny website. We decided to use the *flatly* theme from this package.

Multiple selections

For our interactive graph, we decided to allow users to select multiple countries/continents so that they can look at the data for the countries they want to see. In order to do so, we added the argument “*multiple = TRUE*” in *selectInput()*.

User interactive graphs

This is done by using the *renderPlotly* and the *ggplotly* functions that we learned in class. In the part where we filter the data so that the graph shows selected input as data, we have an *ifelse* statement that checks whether the input is *NULL*. The input being *NULL* means that the user left the input for *Country* and *Continent* blank. If this is the case, the graph shows data for all countries. Otherwise, the graph only shows data for selected countries and continents.

Sidebar Layout

We decided to add a sidebar to the interactive graphs so that the user can look at the graph changing as they select their inputs. This was done by adding *sidebarLayout* onto our panel.

- **Description of Application**

In our application, there are four different tabs that the user can navigate through using the top navigation bar: *About, Dataset, Suicide Rate, and Fertility Rate*. Below is the information for each tab:

- *About*: The main page of the website has information about the website and each tab.
- *Dataset*: Explains the definition of suicide rate and fertility rate and the source of data.
- *Suicide Rate*:

  The tab is divided into two sub-tabs:

  - *View By Continent:*

  The user will be seeing an interactive *ggplotly* graph that has "*Year*" on the horizontal axis, "*Value*" on the vertical axis, and the suicide rate data of countries located in each selected continent as a line graph. The following are inputs that the user can choose on the sidepanel located on the leftside of the screen:

    - *Year*: The graph will show the data for the selected year range.
    - *Continents*: The graph will show the data for all countries that are in the selected continents.

  If the user leaves the input for *Continent* blank, the graph will automatically show data for all continents.

  - *View By Country:*

  The user will be seeing an interactive *ggplotly* graph that has "*Year*" on the horizontal axis, "*Value*" on the vertical axis, and the suicide rate data for all countries as a pale gray colored line graph. On the same graph, the suicide rate fata for the country that the user selected will be the only line shown in color, which allows the users to understand where the selected country is among other OECD countries in terms of suicide rate. The following are inputs that the user can choose on the sidepanel located on the leftside of the screen:

    - *Year*: The graph will show the data for the selected year range.
    - *Country*: The graph will show a colored line graph for a selected country, while the other countries' data stay gray in the background of the grid.

  By default, the graph will automatically show data for the first country in the list(Argentina).
- *Fertility Rate*:

  The tab is divided into two sub-tabs:

  - *View By Continent:*

The user will be seeing an interactive *ggplotly* graph that has "*Year*" on the horizontal axis, "*Value*" on the vertical axis, and the fertility rate data of countries located in each selected continent as a line graph. The following are inputs that the user can choose on the sidepanel located on the leftside of the screen:

- *Year*: The graph will show the data for the selected year range.
- *Continents*: The graph will show the data for all countries that are in the selected continents.

If the user leaves the input for *Continent* blank, the graph will automatically show data for all continents.

- *View By Country:*

The user will be seeing an interactive *ggplotly* graph that has "*Year*" on the horizontal axis, "*Value*" on the vertical axis, and the fertility rate data for all countries as a pale gray colored line graph. On the same graph, the fertility rate data for the country that the user selected will be the only line shown in color, which allows the users to understand where the selected country is among other OECD countries in terms of fertility rate. The following are inputs that the user can choose on the sidepanel located on the leftside of the screen:

- *Year*: The graph will show the data for the selected year range.
- *Country*: The graph will show a colored line graph for a selected country, while the other countries' data stay gray in the background of the grid.

By default, the graph will automatically show data for the first country in the list(Argentina).

We decided to use line graphs to show the trend over the years. Suicide rate and fertility rate are statistics that affect economy and also reflect life quality, and therefore these two statistics are considered important when looking at different countries' data. We wanted to see if there were any trends between factors in the dataset of these statistics. More specifically, we were aiming to determine whether there was a relationship between continent and suicide rate and between continent and fertility rate. In addition to using the line graph to emphasize the trends over the years, we made the line graph interactive so users can select certain factors and see how the

trends and data transform. The user will be more convinced of our analysis if user can subset the data by continent for suicide rate and fertility rate.

- **Graphics**

The graphics part of our application is located in the server section of the shiny application. Based on the user input, we utilize the *reactive* expression and *filter* the formatted data set multiple times via *pipe* to only select data rows. With filtered reactive expressions assigned to variables, we load them in our interactive line graph to only display the data user manually subset. We use the *reactive* expression from the *shiny* package, *filter* from the *tidyverse* package and *pipe* from the *dplyr* package learned in class.

Analysis

- **Suicide Rate**

From the interactive line graph, we see that the suicide rates generally is the highest for countries in Europe. Then follow countries in Asia, countries in Oceania, countries in Americas and a country in Africa (only one OECD country from Africa). Continent may not be the only factor that affects suicide rate, but is observed to have a correlation to the suicide rate of OECD countries.

- **Fertility Rate**

From the interactive line graph and the trends of countries, we see that fertility rates of countries in all the continents are decreasing over time. The rate of decrease may differ by each country, but we can make an observation by viewing the countries by the continent. We see that countries in Asia, especially Saudi Arabia, show a strong decrease in fertility rate over time. Then, countries in the Americas and Africa show a steady decrease in fertility rate over time. Lastly, countries in Europe and Oceania do show a general decrease in fertility rate, but not by much. Continent may not be the only factor that affects the decrease in fertility rate, but is observed to have a correlation to the fertility rate of OECD countries.

Link to Shiny app

https://songj2.shinyapps.io/Final_Project/

References

Package “Countrycode.” 2020.

“Shiny Themes.” Rstudio.github.io, rstudio.github.io/shinythemes/.

“Create a Select List Input Control — SelectInput.” Shiny.rstudio.com, shiny.rstudio.com/reference/shiny/latest/selectinput. Accessed 15 Mar. 2023.

“Shiny - Application Layout Guide.” Shiny.rstudio.com, shiny.rstudio.com/articles/layout-guide.html.

“Shiny - How to Use DataTables in a Shiny App.” Shiny.rstudio.com, shiny.rstudio.com/articles/datatables.html.

“[Solved] How to Center an Image in a Shiny App?” 9to5answer.com, 9to5answer.com/how-to-center-an-image-in-a-shiny-app. Accessed 15 Mar. 2023.

“How Do I Rename Values of a Variable in R?” Stack Overflow, stackoverflow.com/questions/66736264/how-do-i-rename-values-of-a-variable-in-r. Accessed 15 Mar. 2023.