42046: Data processing using R.
Interactive Visualization System
Assessment 2

Prepared by: Syed Jazil Hussain
Student ID: 14369793
Date: 18/11/2022

Contents

1. Project Summary & Introduction:

All of the revenue received by all people or households collectively within a nation is referred to as personal income. The term "personal income" refers to compensation received from a variety of sources, such as salary, wages, and bonuses from jobs or self-employment. Consumer spending is significantly influenced by personal income, which is also an important metric used by governments to assess the state of the nation (Kagan, 2022). Personal income may rise or fall for a variety of causes, such as periods of economic expansion or a dip during recessions. When all these circumstances are held constant, additional factors, which we shall examine in detail in this paper, may also have an impact on people's income.

The report reveals whether a person's income is greater than 50,000 or less than 50,000 for each individual (value associated with each countries different currencies). Numerous continuous variables, such as age and the amount of hours worked each week, affect an individual's income. additionally, to category factors such sex, colour, education, and socioeconomic class. One country from each of the available continents in the dataset—the United States, India, Columbia, and England—will be examined for each of these characteristics.

The dataset for adult income was downloaded from Kaggle. 32,600 rows and 15 columns are present. Looking at the initial dataset, there are a few missing values, but they do not render the entire row unreliable because they are simply missing values when collecting data from individuals, or that row may not be applicable for that individual in the case of an individual's education or work class (if they are unemployed).

1.1 The Business Problem:

Analysing the dataset 'adult income' for the 4 countries, USA, India, Columbia, and England to see whether or not individuals from these countries have an income of > or < 50,000. Also determining the impact different continuous and categorical variables have on the personal income of an individual.

1.2 Business Questions:

The following inquiries are covered by the analysis:

1. Which nation's GDP is most likely to be the highest?
2. Do those who put in more hours earn greater money?
3. Does a person's age affect their income?
4. How do factors like sex, race, education level, and work class affect a person's personal income?

1.3 Dataset:

The dataset contains the predictions of individuals on whether or not their personal income is > or < 50,000. Each row represents one person, while the columns contain different attributes.

The dataset contains the following variables:

Age (Numeric): Age of the individual

Work class (Character): Job type, private/government/self-employed

Education (Character): Highest level of education attained

Education Number: Number of years spent in education

Marital Status (Character): If an individual is married/not married/divorced

Occupation (Character): Job role

Relationship (Character): Current relationship status

Race (Character): Ethnic race

Sex (Binary): Either Male or Female

Hours Per Week: Number of hours worked a week

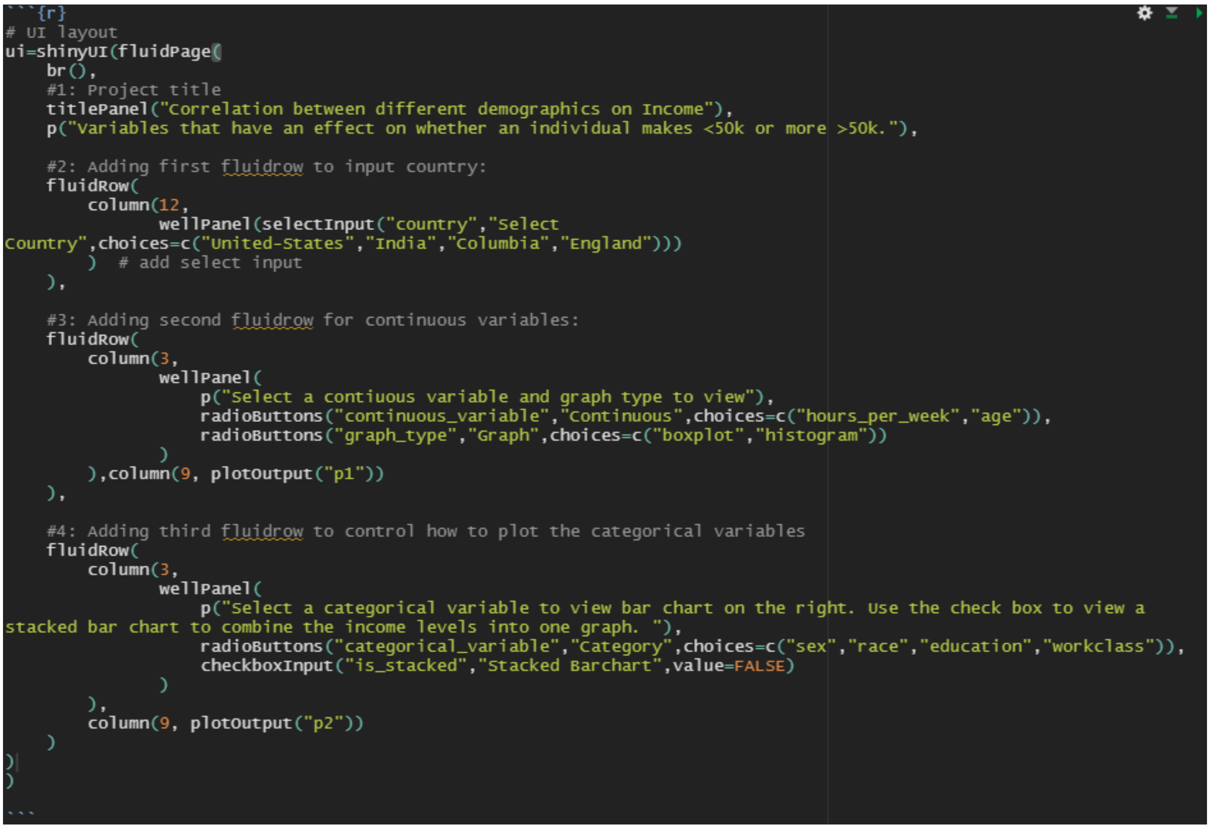
Native Country (Character): Country of nationality

Prediction (String): If an individual makes > < or = 50,000

2. Methodology & Discussion:

Creating a user interface and server.

2.1 User Interface or UI Interface:



~~~
```{r}
# UI layout
ui=shinyUI(fluidPage(
    br(),
    #1: Project title
    titlePanel("Correlation between different demographics on Income"),
    p("Variables that have an effect on whether an individual makes <50k or more >50k."),

    #2: Adding first fluidrow to input country:
    fluidRow(
        column(12,
               wellPanel(selectInput("country","Select
Country",choices=c("United-States","India","Columbia","England")))
        )  # add select input
    ),

    #3: Adding second fluidrow for continuous variables:
    fluidRow(
        column(3,
               wellPanel(
                   p("Select a contiuous variable and graph type to view"),
                   radioButtons("continuous_variable","Continuous",choices=c("hours_per_week","age")),
                   radioButtons("graph_type","Graph",choices=c("boxplot","histogram"))
               )
        ),column(9, plotOutput("p1"))
    ),

    #4: Adding third fluidrow to control how to plot the categorical variables
    fluidRow(
        column(3,
               wellPanel(
                   p("Select a categorical variable to view bar chart on the right. Use the check box to view a
stacked bar chart to combine the income levels into one graph. "),
                   radioButtons("categorical_variable","Category",choices=c("sex","race","education","workclass")),
                   checkboxInput("is_stacked","Stacked Barchart",value=FALSE)
               )
        ),
        column(9, plotOutput("p2"))
    )
)
)

```
~~~

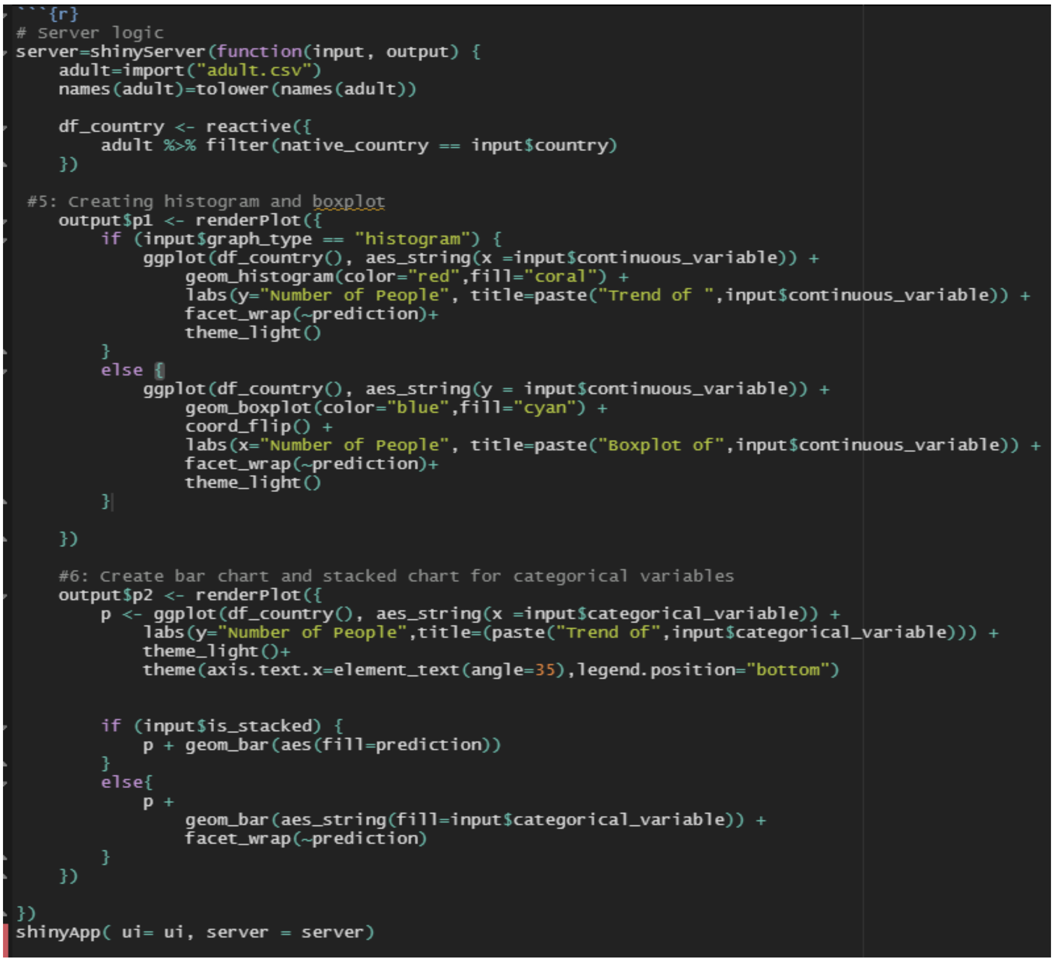
Figure 1: User interface code

The user interface (UI) consists of 2 types of panels. The first being a drop-down panel or a select panel where the user can pick between the 4 countries; USA, India, Columbia & England. The second being a check box where users can select between different continuous variables such as hours_per_week & age as well as different categorical variables such as sex, race, education & work class.

2.2 Main Panel:

The main panel is a check box used to change the output of the type of the chart being used; the first output option changes the chart type to boxplot, while the second output option changes the chart type to histogram. The main panel also includes the stacked chart which is used to compare each categorical variable with the distribution of income predicted.

2.3 The Server:



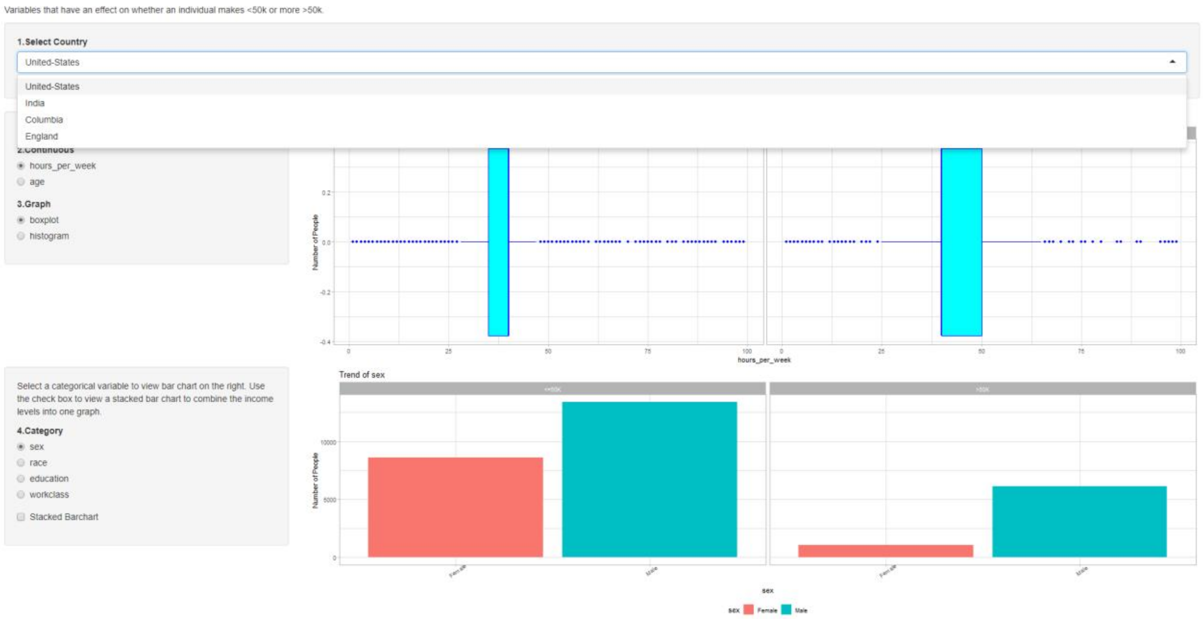
~~~
```{r}
# Server logic
server=shinyServer(function(input, output) {
    adult=import("adult.csv")
    names(adult)=tolower(names(adult))

    df_country <- reactive({
        adult %>% filter(native_country == input$country)
    })

 #5: Creating histogram and boxplot
    output$p1 <- renderPlot({
        if (input$graph_type == "histogram") {
            ggplot(df_country(), aes_string(x =input$continuous_variable)) +
                geom_histogram(color="red",fill="coral") +
                labs(y="Number of People", title=paste("Trend of ",input$continuous_variable)) +
                facet_wrap(~prediction)+
                theme_light()
        }
        else {
            ggplot(df_country(), aes_string(y = input$continuous_variable)) +
                geom_boxplot(color="blue",fill="cyan") +
                coord_flip() +
                labs(x="Number of People", title=paste("Boxplot of",input$continuous_variable)) +
                facet_wrap(~prediction)+
                theme_light()
        }

    })

    #6: Create bar chart and stacked chart for categorical variables
    output$p2 <- renderPlot({
        p <- ggplot(df_country(), aes_string(x =input$categorical_variable)) +
            labs(y="Number of People",title=(paste("Trend of",input$categorical_variable))) +
            theme_light()+
            theme(axis.text.x=element_text(angle=35),legend.position="bottom")


        if (input$is_stacked) {
            p + geom_bar(aes(fill=prediction))
        }
        else{
            p +
                geom_bar(aes_string(fill=input$categorical_variable)) +
                facet_wrap(~prediction)
        }
    })

})
shinyApp( ui= ui, server = server)
~~~

Figure 2: Server info code

The code above shows how the server connects the user input (country, hours_per_week, age, sex, race, education & work class) with the outputs in the main panel (boxplot, histogram & stacked chart).

2.4 The Interactive Tool:

The charts in the main panel are interactive and represent the data in either a histogram or boxplot as well as displaying the data alongside the predicted income in a stacked bar chart. The panel at the top displays the 4 regions that can be selected to highlight the data. Other check boxes present in the side panel can be selected to represent the affects of different variables on the predicted income for individuals.



Variables that have an effect on whether an individual makes <50k or more >50k.

1.Select Country

United-States

- United-States
- India
- Columbia
- England

2.Continuous

● hours_per_week
○ age

3.Graph

● boxplot
○ histogram

Select a categorical variable to view bar chart on the right. Use the check box to view a stacked bar chart to combine the income levels into one graph.

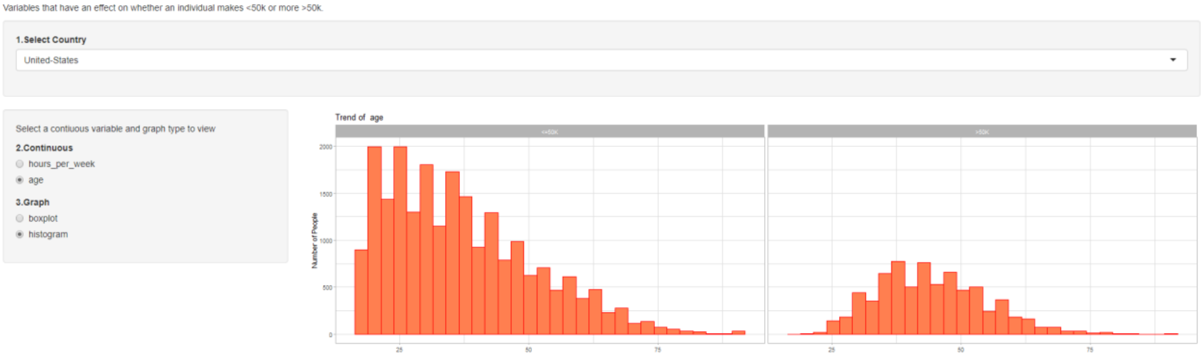
4.Category

● sex
○ race
○ education
○ workclass

□ Stacked Barchart

Figure 3: Regional selection

Figure 3 shows the 4 different regions that can be selected and cycled through to represent and display the data for that area, as we can see above, the USA has been selected.



Variables that have an effect on whether an individual makes <50k or more >50k.

1.Select Country

United-States

Select a contiuous variable and graph type to view

2.Continuous

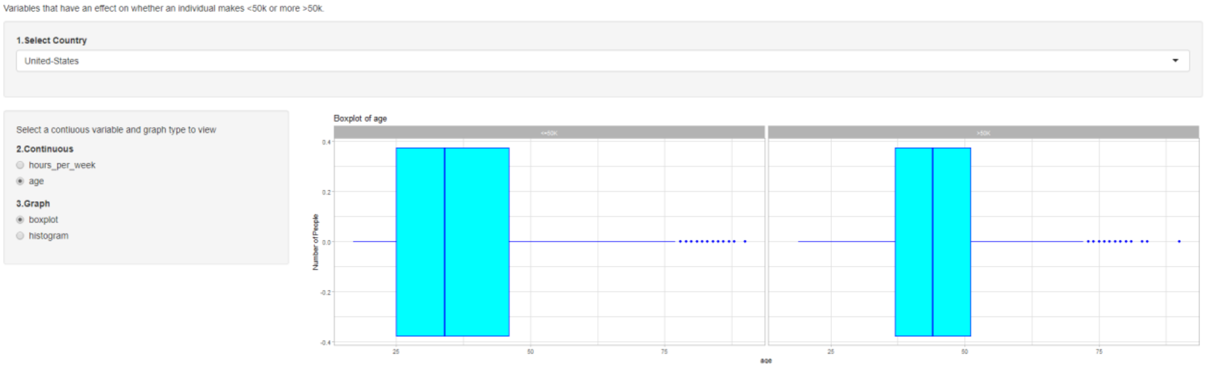
○ hours_per_week
● age

3.Graph

○ boxplot
● histogram

Figure 4: Age vs income (histogram) in USA

Figure 4 displays the affect age has on the personal income of an individual displayed as a histogram.



Variables that have an effect on whether an individual makes <50k or more >50k.

1.Select Country

United-States

Select a contiuous variable and graph type to view

2.Continuous

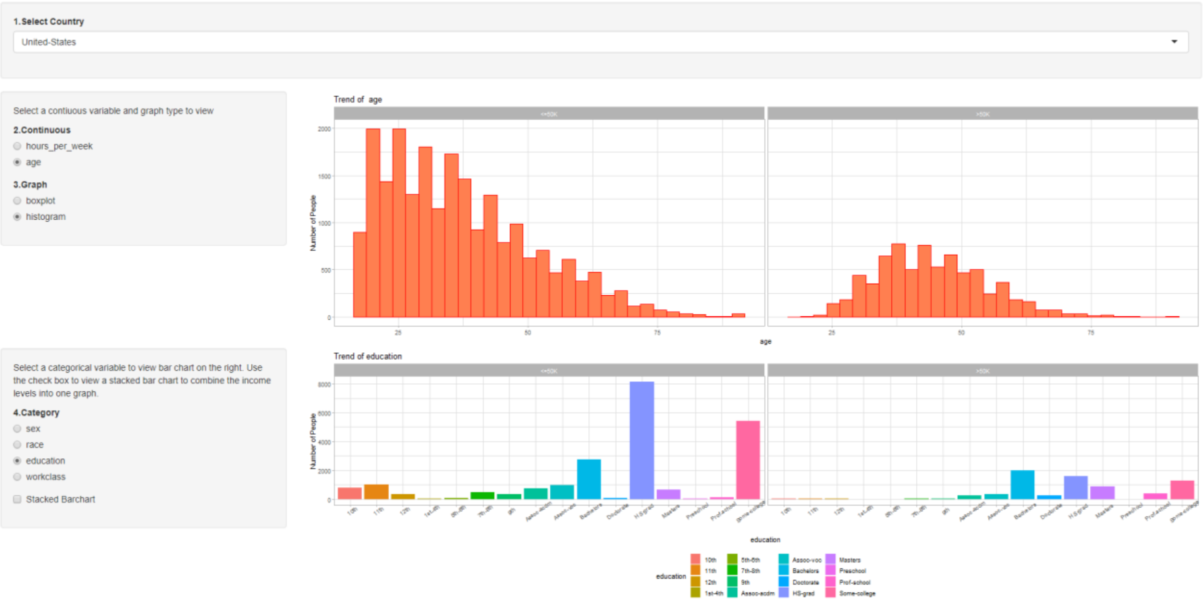
○ hours_per_week
● age

3.Graph

● boxplot
○ histogram

Figure 5: Age vs income (boxplot) in USA

By using the different check boxes in the main panel, we can represent the affect age has on personal income of an individual represented in a box plot.



1.Select Country

United-States

Select a contiuous variable and graph type to view

2.Continuous

○ hours_per_week

● age

3.Graph

○ boxplot

● histogram

Select a categorical variable to view bar chart on the right. Use the check box to view a stacked bar chart to combine the income levels into one graph.

4.Category

○ sex

○ race

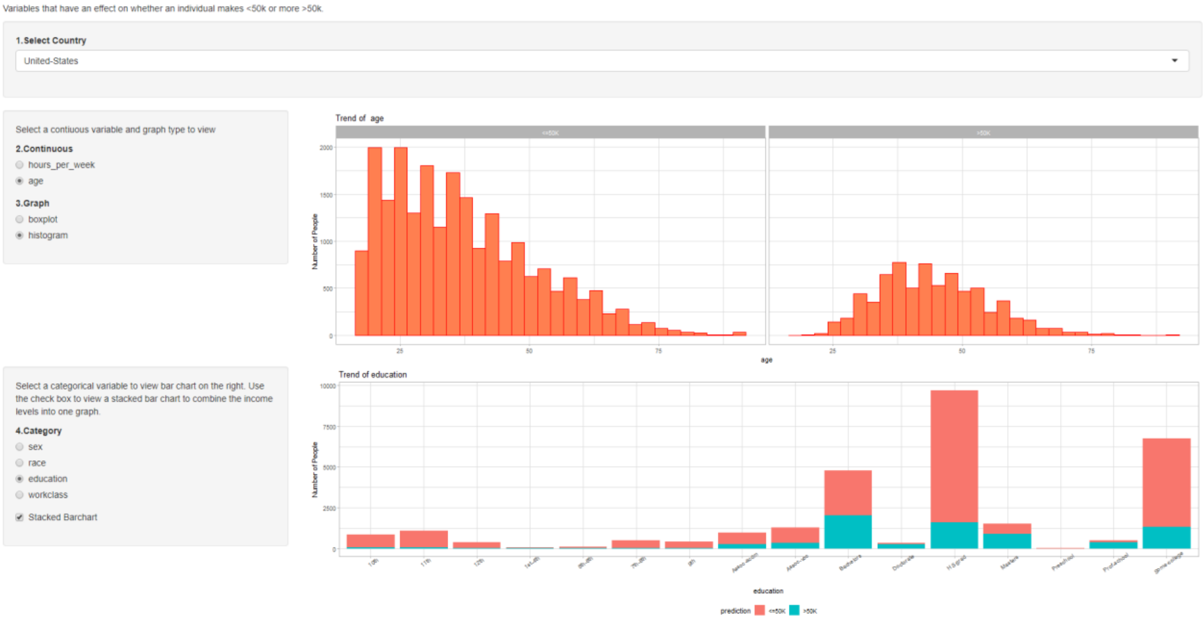
● education

○ workclass

□ Stacked Barchart

Figure 6: Age vs income (histogram) categorized by education in USA

After selecting the region, continuous variable, and graph type, we can categorize the data by selecting one of the 4 categorical variables, the one we've chosen above is education which represents the personal of income for individuals based on their highest level of education attained.



Variables that have an effect on whether an individual makes <50k or more >50k.

1.Select Country

United-States

Select a contiuous variable and graph type to view

2.Continuous

○ hours_per_week

● age

3.Graph

○ boxplot

● histogram

Select a categorical variable to view bar chart on the right. Use the check box to view a stacked bar chart to combine the income levels into one graph.

4.Category

○ sex

○ race

● education

○ workclass

☑ Stacked Barchart

Figure 7: Age vs income (histogram) categorized by education in USA compared to income prediction

The data can be further analysed and compared by selecting 'stacked bar chart' in the main side panel which helps compare the analysis the prediction of individuals of an income of less than or greater than 50,000.

3. Results & Conclusions:

After the data exploration and analysis were completed, the following conclusions can be made.

The USA has the highest number of individuals with a personal income of > 50,000. This may be because the USA is the only developed country besides England because developed countries tend to have a higher GDP compared to developing countries. The USA may have more individuals with a higher income because of the higher population it has to England which has a direct correlation with people with an income greater than 50,000. The number of hours worked by an individual has a direct positive correlation with the probability of them earning more than 50,000 which is because people are paid based off of how much work they put in. Furthermore, the age also has a direct correlation with the personal income of an individual until a certain point. This can be accustomed to someone older having more work experience. However, this declines after a certain point because people end up retiring at an older age. As shown in figure 7. Moreover, education has a direct correlation with the amount someone earns, people with a bachelors has the greatest number of individuals earning more than 50,000 while having a masters had the highest proportion of people earning more than 50,000.

Lastly, analysing the country, different continuous variables such as hours worked per week, age and categorical variables such as sex, race, education and work class help countries assess what they should improve on to help increase the GDP for the country, such as providing cheaper education so that individuals can easily access and complete a bachelors or a master’s degree, which will in theory increase their income and the GDP of the country.

4. References:

stanley888cy (2021) *Data visualization with R shiny*, *Kaggle*. Kaggle. Available at: https://www.kaggle.com/code/stanley888cy/data-visualization-with-r-shiny/data (Accessed: November 20, 2022).

Kagan, J. (2022) *Personal income definition & difference from disposable income*, *Investopedia*. Investopedia. Available at: https://www.investopedia.com/terms/p/personalincome.asp#:~:text=Personal%20income%20is%20the%20amount,and%20profit%20sharing%20from%20businesses (Accessed: November 20, 2022).