

Final Report

Jesse Conlon

3/13/2020

Shiny App:

https://jesse-conlon.shinyapps.io/Jesse_Conlon_Final_Project/

The data sets used provide statistics for US income in 2015, with supplemental information for education, crime, and religion and can be located on [kaggle.com](https://www.kaggle.com) and data.world.com, respectively.

Below is the beginning of the R Markdown file used to generate the Static Visuals used in the project as well as in the R Shiny App. At the end of the static visuals is an executive summary as well as the code used in the R Shiny App.

Final Project Static Visuals Code

```
## Importing the Income Data

USincome <- read.csv(file = "/cloud/project/kaggle_income2.csv", header = TRUE,
  stringsAsFactors = TRUE, na.strings = "")

USdata <- read.csv(file = "/cloud/project/state_education.csv", header = TRUE,
  stringsAsFactors = TRUE, na.strings = "")

## Importing the shape file of USA States and Counties

states_map <- st_read("/cloud/project/cb_2018_us_state_500k.shp", quiet = TRUE)

counties_map <- st_read("/cloud/project/cb_2018_us_county_500k.shp", quiet = TRUE)

## Additional files that may be used

countries <- st_as_sf(maps::map('world', plot = FALSE, fill = TRUE))

states <- st_as_sf(maps::map('state', plot = FALSE, fill = TRUE))

counties <- st_as_sf(maps::map('county', plot = FALSE, fill = TRUE))
```

```

## Removing data from outside of the 48 continental states

state_filter <- states_map %>%
  filter(!NAME == "United States Virgin Islands") %>%
  filter(!NAME == "Commonwealth of the Northern Mariana Islands") %>%
  filter(!NAME == "District of Columbia") %>%
  filter(!NAME == "Guam") %>%
  filter(!NAME == "Alaska") %>%
  filter(!NAME == "Hawaii") %>%
  filter(!NAME == "American Samoa") %>%
  filter(!NAME == "Puerto Rico")

state_data_filter <- USdata %>%
  filter(!State == "Alaska") %>%
  filter(!State == "Hawaii")

## Grouping income by states and gaining a mean location for the areas

state_avg_income <- USincome %>%
  group_by(State_Name) %>%
  summarize(Income = median(Median), Long = mean(Lon), Lat = mean(Lat),
            Median = median(Median), households = sum(sum_w), Mean = median(Mean))

## Joining the data sets

avg_income_map_sd <- state_filter %>%
  left_join(state_avg_income, by = c("NAME" = "State_Name"))

## Warning: Column `NAME`/`State_Name` joining factors with different levels,
## coercing to character vector

## Create a new column for standard deviation

avg_income_map_sd$'States' <- rownames(avg_income_map_sd)

avg_income_map_sd$'States' <- factor(avg_income_map_sd$'States',
                                   levels = avg_income_map_sd$'States')

avg_income_map_sd$Income <- round((avg_income_map_sd$Income -
                                   mean(avg_income_map_sd$Income))
                                   /sd(avg_income_map_sd$Income), 2)

avg_income_map_sd$Income_Type <- ifelse(avg_income_map_sd$Income < 0, "below", "above")

avg_income_map_sd$'States' <- factor(avg_income_map_sd$'States',
                                   levels = avg_income_map_sd$'States')

avg_income_map_sd <- avg_income_map_sd %>%
  mutate(States = fct_reorder(NAME, Income))

avg_income_map_sd <- avg_income_map_sd %>%
  left_join(state_data_filter, by = c("NAME" = "State"))

## Warning: Column `NAME`/`State` joining character vector and factor, coercing
## into character vector

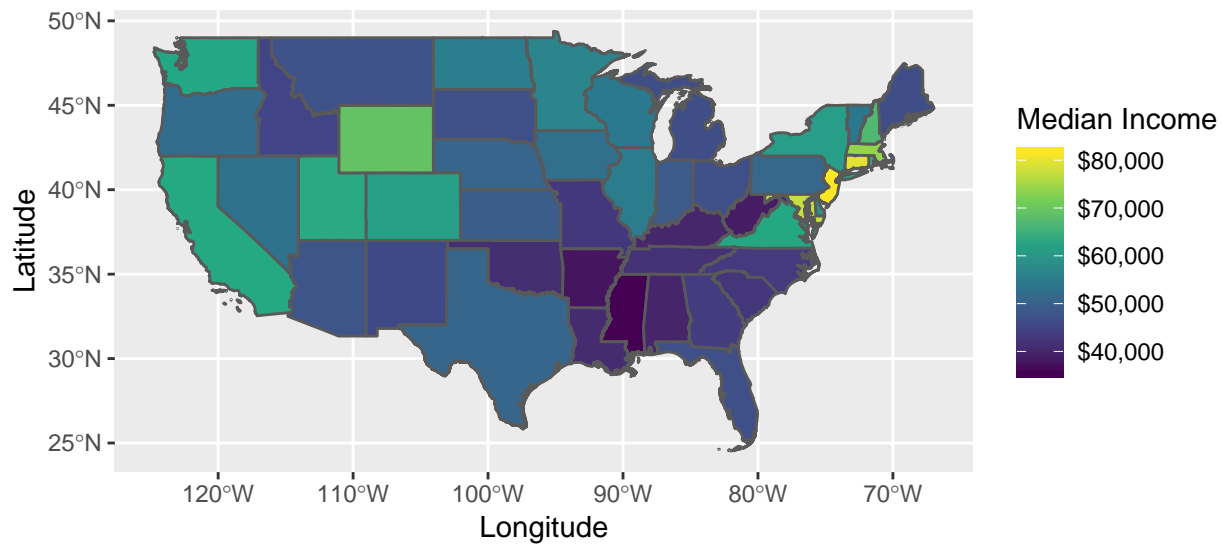
```

```
# Plotting the data to a Chloropleth Map
```

```
ggplot(avg_income_map_sd, aes(fill = Median)) +
  geom_sf() +
  scale_fill_viridis_c("Median Income", labels = dollar) +
  labs(x = "Longitude",
       y = "Latitude",
       title = "Median Income per State",
       subtitle = "Created for Data Visualization Class",
       caption = "Source: kaggle.com & data.world.com")
```

Median Income per State

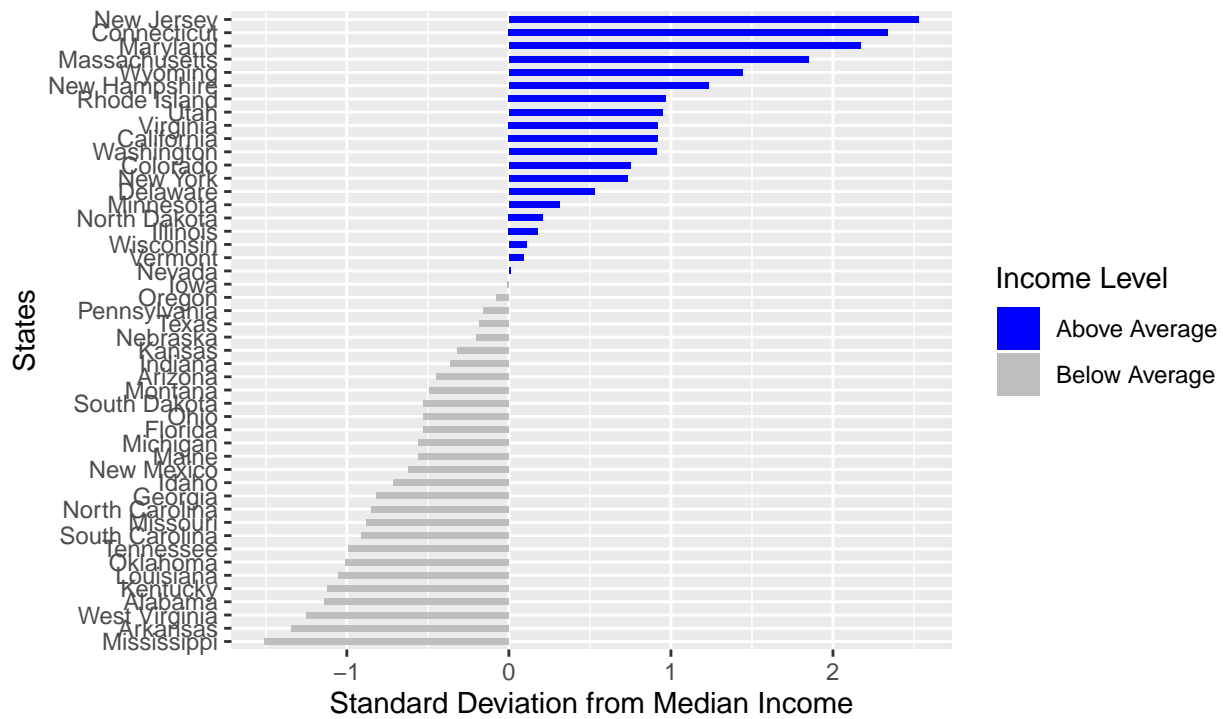
Created for Data Visualization Class



```
ggplot(avg_income_map_sd, aes(x = States, y = Income, label = Income)) +
  geom_bar(stat = 'identity', aes(fill = Income_Type), width = .5) +
  scale_fill_manual(name = "Income Level",
                    labels=c("Above Average", "Below Average"),
                    values = c("above"="blue", "below"="gray")) +
  coord_flip() +
  labs(x = "States",
       y = "Standard Deviation from Median Income",
       title = "Median Income per State",
       subtitle = "Created for Data Visualization Class",
       caption = "Source: kaggle.com & data.world.com",
       fill = "Median Income", labels = dollar)
```

Median Income per State

Created for Data Visualization Class

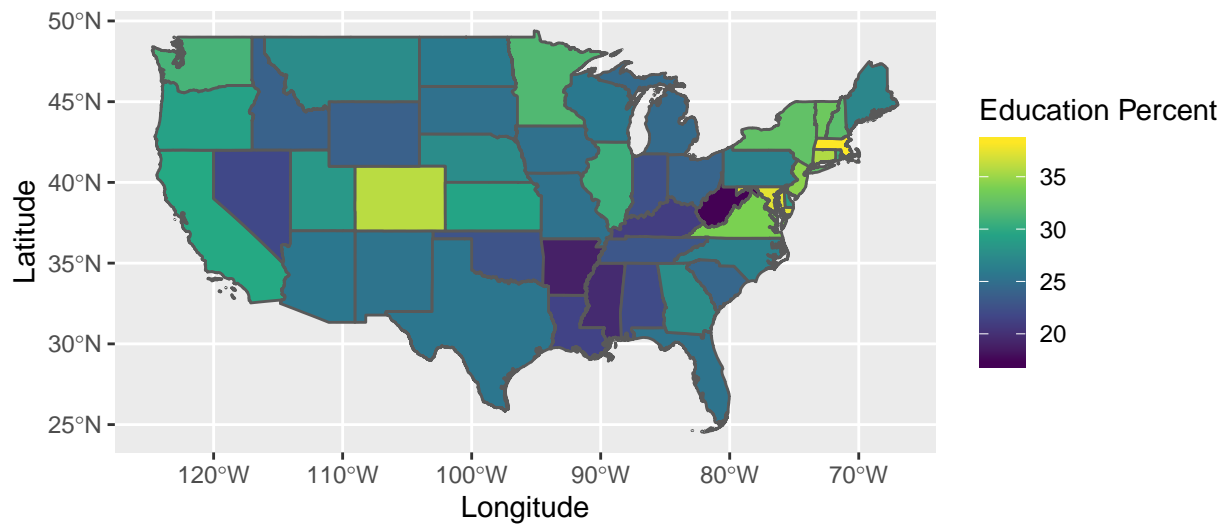


Source: kaggle.com & data.world.com

```
ggplot(avg_income_map_sd, aes(fill = Percent.Educational.Attainment)) + geom_sf() +
  scale_fill_viridis_c("Education Percent") +
  labs(x = "Longitude",
       y = "Latitude",
       title = "Percentage of Bachelor Degree or Higher",
       subtitle = "Created for Data Visualization Class",
       caption = "Source: kaggle.com & data.world.com")
```

Percentage of Bachelor Degree or Higher

Created for Data Visualization Class

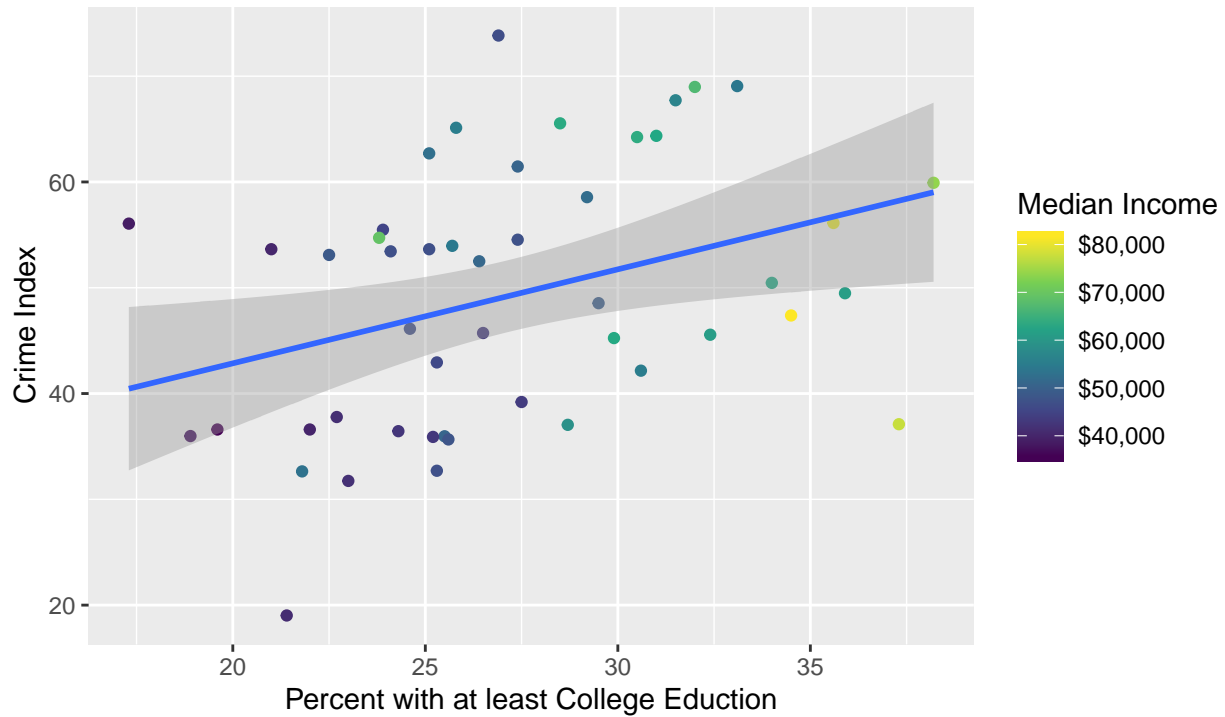


Source: kaggle.com & data.world.com

```
ggplot(data = avg_income_map_sd, aes(x = Percent.Educational.Attainment,  
                                     y = Percent.Peace.Index, color =  
                                     Median)) +  
  scale_colour_viridis_c("Median Income", labels = dollar) +  
  geom_point() + geom_smooth(method="lm") +  
  labs(title = "Education and Peace Coorelation",  
        x = "Percent with at least College Eduction",  
        y = "Crime Index",  
        subtitle = "Created for Data Visualization Class",  
        caption = "Source: kaggle.com & data.world.com")
```

Education and Peace Coorelation

Created for Data Visualization Class

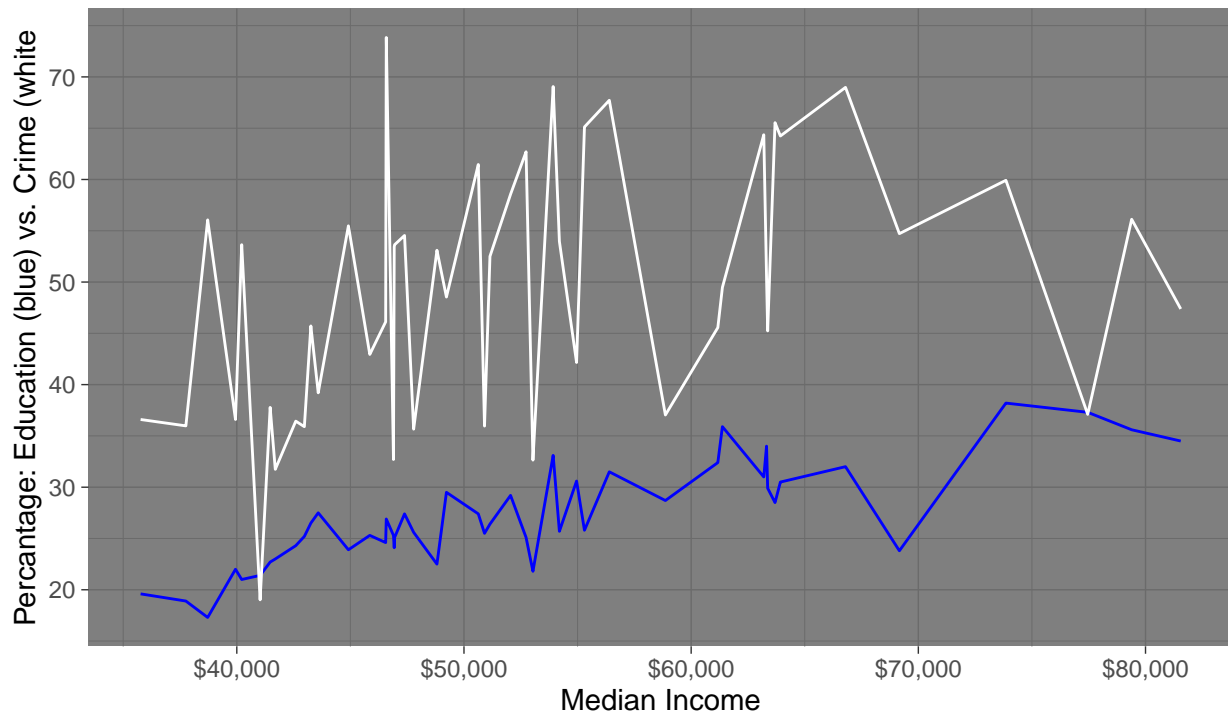


Source: kaggle.com & data.world.com

```
ggplot() +
  geom_line(data = avg_income_map_sd, aes(x = Median,
                                           y = Percent.Educational.Attainment),
            color = "blue", group = 1) +
  geom_line(data = avg_income_map_sd, aes(x = Median, y = Percent.Peace.Index,),
            color = "white", group = 1) +
  scale_x_continuous(labels = dollar) +
  scale_y_continuous(breaks = seq(10,80, by = 10)) +
  labs(title = "Mean vs. Median Income and Household Size",
        subtitle = "Created for Data Visualization Class",
        caption = "Source: kaggle.com & data.world.com",
        x = "Median Income",
        y = "Percentage: Education (blue) vs. Crime (white)") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  theme_dark()
```

Mean vs. Median Income and Household Size

Created for Data Visualization Class

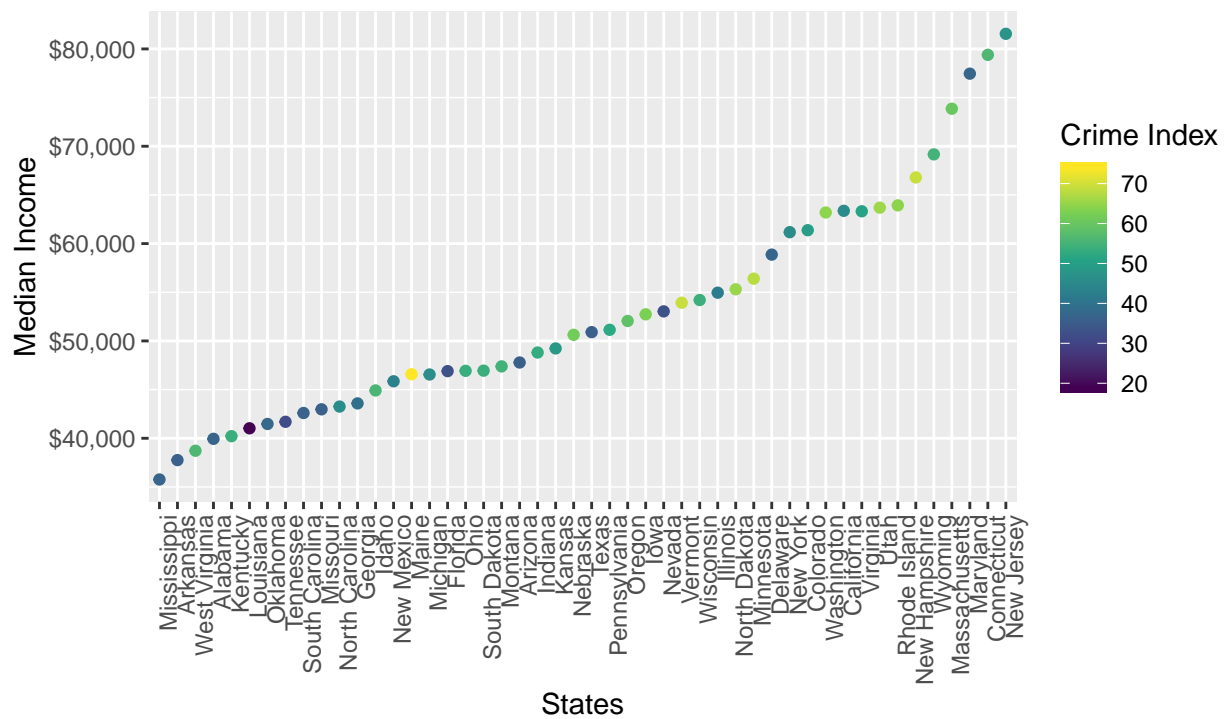


Source: kaggle.com & data.world.com

```
ggplot(data = avg_income_map_sd, aes(x = States, y = Median, color = Percent.Peace.Index)) +  
  geom_point() +  
  scale_colour_viridis_c("Crime Index") +  
  scale_y_continuous(labels = dollar) +  
  labs(title = "Median Income Per State",  
        subtitle = "Created for Data Visualization Class",  
        caption = "Source: kaggle.com & data.world.com",  
        x = "States", labels = dollar,  
        y = "Median Income") +  
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

Median Income Per State

Created for Data Visualization Class



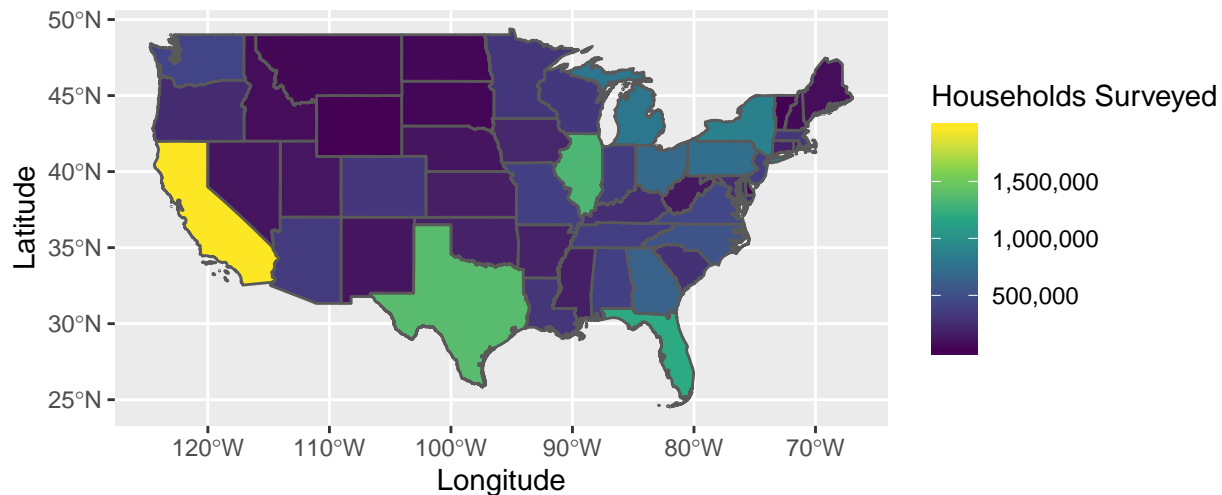
Source: kaggle.com & data.world.com

Households Surveyed

```
ggplot(avg_income_map_sd, aes(fill = households)) + geom_sf() +
  scale_fill_viridis_c("Households Surveyed", label = comma) +
  labs(x = "Longitude",
       y = "Latitude",
       title = "Number of Households Surveyed in Census",
       subtitle = "Created for Data Vis Class",
       caption = "Source: Kaggle.com - Income Census")
```


Number of Households Surveyed in Census

Created for Data Vis Class



Source: Kaggle.com – Income Census

Executive Summary

It was initially thought that there would be a relationship between the variables, such as education and income, and education and crime. A relationship between education and income is evident by adjusting the sliders where only states above ~29.7% college education or greater are above the national average.

However, there is not enough to indicate that there is a strong or consistent relationship between education, income, and/or crime. Initially it was thought that areas with a greater percent of education (also correlated with an increase in come) would mean less crime. However, with further thought and evidence, crime occurs at any income, though may come in different forms, as a result of different circumstances. Different methods were used to try and draw out a correlation and if anything there's a slight correlation in favor of the opposite of what was expected: crime slightly increases with income.

The conclusion is that there is a relationship between education and income but there is not substantial proof that crime is related to income or education.

R Shiny App Code

```
# ui

ui <- fluidPage(theme = shinytheme("slate"),
  titlePanel("US Income Statistics"),
  sidebarLayout(
    sidebarPanel(sliderInput("income", "Median Income",
      value = c(35766, 81550.5), min = 35766,
      max = 81550.5, pre = '$'),
      sliderInput("education", "College Education Rate",
        value = c(17.3, 38.2), min = 17.3,
        max = 38.2, post = '%')
    ),
    mainPanel(
```

```

    tabsetPanel(
      tabPanel("Income",
        plotOutput("Chloropleth_Income"),
        plotOutput("Diverging"),
      ),
      tabPanel("Education",
        plotOutput("Chloropleth_Education"),
        plotOutput("Crime")
      ),
      tabPanel("Crime",
        plotOutput("Line"),
        plotOutput("Polar")
      ),
      tabPanel("Households Surveyed",
        plotOutput("Chloropleth_Households")
      )
    ),
  ),
)
)

```

```
# server
```

```
## Loading all libraries that may have been used in working on or completing the report
```

```

library(dplyr)
library(forcats)
library(maps)
library(raster)
library(readr)
library(tibble)
library(scales)
library(sf)
library(shiny)
library(shinythemes)
library(stringr)
library(tidyverse)

```

```
## Loading data sets that have been used in the report
```

```
## Importing the Income Data
```

```

USIncome <- read.csv(file = "/cloud/project/kaggle_income2.csv",
  header = TRUE, stringsAsFactors = TRUE, na.strings = "")

```

```

USdata <- read.csv(file = "/cloud/project/state_education.csv",
  header = TRUE, stringsAsFactors = TRUE, na.strings = "")

```

```
## Importing the shape file of USA States and Counties
```

```
states_map <- st_read("/cloud/project/cb_2018_us_state_500k.shp", quiet = TRUE)
```

```

## Additional files that may be used

countries <- st_as_sf(maps::map('world', plot = FALSE, fill = TRUE))

states <- st_as_sf(maps::map('state', plot = FALSE, fill = TRUE))

counties <- st_as_sf(maps::map('county', plot = FALSE, fill = TRUE))

## Data Sets

## Removing data from outside of the 48 continental states

state_filter <- states_map %>%
  filter(!NAME == "United States Virgin Islands") %>%
  filter(!NAME == "Commonwealth of the Northern Mariana Islands") %>%
  filter(!NAME == "District of Columbia") %>%
  filter(!NAME == "Guam") %>%
  filter(!NAME == "Alaska") %>%
  filter(!NAME == "Hawaii") %>%
  filter(!NAME == "American Samoa") %>%
  filter(!NAME == "Puerto Rico")

state_data_filter <- USdata %>%
  filter(!State == "Alaska") %>%
  filter(!State == "Hawaii")

## Grouping income by states and gaining a mean location for the areas

state_avg_income <- USincome %>%
  group_by(State_Name) %>%
  summarize(Income = median(Median), Long = mean(Lon), Lat = mean(Lat),
    Median = median(Median), households = sum(sum_w), Mean = median(Mean))

## Joining the data sets

avg_income_map_sd <- state_filter %>%
  left_join(state_avg_income, by = c("NAME" = "State_Name"))

## Create a new column for standard deviation

avg_income_map_sd$'States' <- rownames(avg_income_map_sd)

avg_income_map_sd$'States' <- factor(avg_income_map_sd$'States',
  levels = avg_income_map_sd$'States')

avg_income_map_sd$Income <- round((avg_income_map_sd$Income -
  mean(avg_income_map_sd$Income))
  /sd(avg_income_map_sd$Income), 2)

avg_income_map_sd$Income_Type <- ifelse(avg_income_map_sd$Income < 0, "below", "above")

```

```

avg_income_map_sd$'States' <- factor(avg_income_map_sd$'States',
                                     levels = avg_income_map_sd$'States')

avg_income_map_sd <- avg_income_map_sd %>%
  mutate(States = fct_reorder(NAME, Income))

avg_income_map_sd <- avg_income_map_sd %>%
  left_join(state_data_filter, by = c("NAME" = "State"))

server <- function(input, output) {

  ## Creating a Education Slider

  Education <- reactive({
    avg_income_map_sd %>%
      filter(Percent.Educational.Attainment >= input$education [1] &
             Percent.Educational.Attainment <= input$education [2]) %>%
      filter(Median >= input$income [1] &
             Median <= input$income [2])
  })

  ## Creating a Income Slider

  Income1 <- reactive({
    avg_income_map_sd %>%
      filter(Median >= input$income [1] &
             Median <= input$income [2]) %>%
      filter(Percent.Educational.Attainment >= input$education [1] &
             Percent.Educational.Attainment <= input$education [2])
  })

  ## Plotting the data

  ## Diverging Bar Chart

  output$Diverging <- renderPlot({

    ggplot(Education(), aes(x = States, y = Income, label = Income)) +
      geom_bar(stat = 'identity', aes(fill = Income_Type), width = .5) +
      scale_fill_manual(name = "Income Level",
                        labels=c("Above Average", "Below Average"),
                        values = c("above"="blue", "below"="gray")) +
      coord_flip() +
      labs(x = "States",
           y = "Standard Deviation from Median Income",
           title = "Median Income per State",
           subtitle = "Created for Data Visualization Class",
           caption = "Source: kaggle.com & data.world.com",
           fill = "Median Income", labels = dollar)

  })

```

Census Maps

```
output$Chloropleth_Income <- renderPlot({

  ggplot(Income1(), aes(fill = Median)) + geom_sf() +
    scale_fill_continuous("Median Income", labels = dollar) +
    labs(x = "Longitude",
         y = "Latitude",
         title = "Median Income per State",
         subtitle = "Created for Data Visualization Class",
         caption = "Source: kaggle.com & data.world.com")
})

output$Chloropleth_Households <- renderPlot({

  ggplot(Income1(), aes(fill = households)) + geom_sf() +
    scale_colour_viridis_c("Median Income", labels = dollar) +
    labs(x = "Longitude",
         y = "Latitude",
         title = "Households Surveyed",
         subtitle = "Created for Data Visualization Class",
         caption = "Source: kaggle.com & data.world.com",
         fill = "Households", labels = dollar)
})

output$Chloropleth_Education <- renderPlot({

  ggplot(Income1(), aes(fill = Percent.Educational.Attainment)) + geom_sf() +
    scale_colour_viridis_c("Median Income", labels = dollar) +
    labs(x = "Longitude",
         y = "Latitude",
         title = "College Education or Higher Percentage",
         subtitle = "Created for Data Visualization Class",
         caption = "Source: kaggle.com & data.world.com",
         fill = "Education Percent", labels = percent)
})

## Line Graph with Income and States

output$Line <- renderPlot({

  ggplot(Education(), aes(x = States, y = Median, color = Percent.Peace.Index)) +
    geom_point() +
    scale_colour_viridis_c("Crime Index") +
    scale_y_continuous(labels = dollar) +
    labs(title = "Median Income Per State",
         subtitle = "Created for Data Visualization Class",
         caption = "Source: kaggle.com & data.world.com",
         x = "States",
         y = "Median Income") +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
})
```

```

## Crime

output$Crime <- renderPlot({

  ggplot(Education(), aes(x = Percent.Educational.Attainment,
                          y = Percent.Peace.Index, color =
                          Median)) + geom_point() + geom_smooth(method="lm") +
  labs(title = "Education and Peace Coorelation",
       x = "Percent with at least College Eduction",
       y = "Crime Index",
       subtitle = "Created for Data Visualization Class",
       caption = "Source: kaggle.com & data.world.com") +
  scale_colour_viridis_c("Median Income", labels = dollar)

})

output$Polar <- renderPlot({

  ggplot() +
    geom_line(data = Income1(), aes(x = Median, y = Percent.Educational.Attainment),
              color = "blue", group = 1) +
    geom_line(data = Income1(), aes(x = Median, y = Percent.Peace.Index,),
              color = "white", group = 1) +
    scale_x_continuous(labels = dollar) +
    scale_y_continuous(breaks = seq(10,80, by = 10)) +
    labs(title = "Mean vs. Median Income and Household Size",
         subtitle = "Created for Data Visualization Class",
         caption = "Source: kaggle.com & data.world.com",
         x = "Median Income",
         y = "Percantage: Education (blue) vs. Crime (white)" ) +
    theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
    theme_dark()

})

}

```