Zezula_Mateusz_Final

Mateusz Zezula

12/9/2018

Introduction

The U.S. Regional Dataset from the St. Louis Federal Reserve (link: https://fred.stlouisfed.org/categories/3008/downloaddata (https://fred.stlouisfed.org/categories/3008/downloaddata)) contains time-series economic data for each country in the United States. The dataset amounts to 330,000 .csv files and over 500 megabytes in uncompressed form.

Although the analytic possibilities of the dataset are endless, I limit myself to answering the following: How do changes in economic variables vary from state-to-state? Display those changes using a heatmap. The creation of functions that answer the above questions presupposes the existence of many supporting functions; these will be explained below.

```
knitr::opts_knit$set(root.dir = "/users/mateuszzezula/Box/1. First Semester/Programming for Data Science/final_pr
oject/")
library(pacman)
p_load(magrittr, ggplot2, stringr, dplyr, reshape2, tseries, zoo, maps, tidyverse, urbnmapr, plotly, Rd2md)
```

datasetIndex()

The dataset includes an index file outlining every .csv file contained in the dataset. Part of the index file is shown for illustrative purposes; the format and information is self-explanatory. In addition, I manually parsed the index file and identified 33 variables of analytic interest. Any variable included in the list can be analyzed with the create.us.heatmap() and graph.diverging.lollipop() functions.

```
datasetIndex <- function() {
  main <- read.table("../final_project/Final/Zezula/README_SERIES_ID_SORT.txt",
                     header = T, sep = ";", quote = "", row.names = NULL,
                     fill = TRUE, stringsAsFactors = F)
  subjects <- read.csv("../final_project/Final/Zezula/subjects.csv",
                       header = T, stringsAsFactors = F)
  return(list(main = main, subjects = subjects))
}
head(datasetIndex()$main)
```

```
##                                                 File
## 1 2\\0\\2\\0\\R\\A\\T\\I\\O\\0\\0\\2020RATIO001001.csv
## 2 2\\0\\2\\0\\R\\A\\T\\I\\O\\0\\0\\2020RATIO001003.csv
## 3 2\\0\\2\\0\\R\\A\\T\\I\\O\\0\\0\\2020RATIO001005.csv
## 4 2\\0\\2\\0\\R\\A\\T\\I\\O\\0\\0\\2020RATIO001007.csv
## 5 2\\0\\2\\0\\R\\A\\T\\I\\O\\0\\0\\2020RATIO001009.csv
## 6 2\\0\\2\\0\\R\\A\\T\\I\\O\\0\\0\\2020RATIO001011.csv
##                                    Title  Units Frequency
## 1 Income Inequality in Autauga County, AL  Ratio         A
## 2 Income Inequality in Baldwin County, AL  Ratio         A
## 3 Income Inequality in Barbour County, AL  Ratio         A
## 4    Income Inequality in Bibb County, AL  Ratio         A
## 5  Income Inequality in Blount County, AL  Ratio         A
## 6 Income Inequality in Bullock County, AL  Ratio         A
##   Seasonal.Adjustment Last.Updated
## 1                 NSA   2018-05-29
## 2                 NSA   2018-05-29
## 3                 NSA   2018-05-29
## 4                 NSA   2018-05-29
## 5                 NSA   2018-05-29
## 6                 NSA   2018-05-29
```

```
datasetIndex()$subjects
```

```
##                          Subject
## 1                 subprime_credit
## 2           private_establishments
## 3           private_establishments
## 4               homeownership_rate
## 5                income_inequality
## 6                per_capita_income
## 7                    snap_benefits
## 8                   per_capita_gdp
## 9                house_price_index
## 10             new_private_housing
## 11                             gdp
## 12                   poverty_all
## 13                  poverty_0_17
## 14               bachelors_degree
## 15                 tax_exemptions
## 16                premature_death
## 17 preventable_hospital_admissions
## 18                  migration_flow
## 19          median_population_age
## 20         violent_crime_incidents
## 21                           white
## 22                          latino
## 23                           asian
## 24                native_american
## 25                           black
## 26              unemployment_rate
## 27                  commute_time
## 28             disconnected_youth
## 29                    high_school
## 30            civilian_labor_force
## 31             resident_population
## 32        single_parent_households
## 33              patent_assignments
##                                                                          Title
## 1                                        Equifax Subprime Credit Population for
## 2                             Number of Private Establishments for All Industries in
## 3                             Number of Private Establishments for All Industries in
## 4                                                        Homeownership Rate for
## 5                                                          Income Inequality in
## 6                                                  Per Capita Personal Income in
## 7                                                   SNAP Benefits Recipients in,
## 8                                Total Per Capita Real Gross Domestic Product for,
## 9                                          All-Transactions House Price Index for,
## 10                New Private Housing Structures Authorized by Building Permits for,
## 11                                          Total Real Gross Domestic Product of,
## 12                                                Poverty Universe, All Ages for,
## 13                                                Poverty Universe, Age 0-17 for,
## 14                               Bachelor's Degree or Higher (5-year estimate) in,
## 15                                       Poverty Tax Exemptions Under Age 65 for)
## 16                                                    Premature Death Rate for
## 17                                         Rate of Preventable Hospital Admissions
## 18                                                     Net Migration Flow for
## 19                                                 Median Age of the Population in
## 20          Combined Violent and Property Crime Incidents Known to Law Enforcement in
## 21                              Population Estimate of Non-Hispanic White Persons in
## 22                              Population Estimate of Hispanic or Latino Persons in
## 23                              Population Estimate of Non-Hispanic Asian Persons in
## 24   Population Estimate of Non-Hispanic American Indian or Native Alaskan Persons in
## 25           Population Estimate of Non-Hispanic Black or African-American Persons in
## 26                                                       Unemployment Rate in
## 27                                             Mean Commuting Time for Workers in
## 28                                                     Disconnected Youth for
## 29                              High School Graduate or Higher (5-year estimate) in
## 30                                                   Civilian Labor Force in
## 31                                                    Resident Population in
## 32 Single-parent Households with Children as a Percentage of Households with Children
## 33                                                  New Patent Assignments in
```

get.subject()

Returns subject variable (that is, title), used as an input for other functions.

```
get.subject <- function(index, number) {
  subject <- index$subjects$Title[number]
  return(subject)
}
get.subject(datasetIndex(), 6)
```

```
## [1] "Per Capita Personal Income in"
```

get.csv()

Parses dataset and returns csv based on inputs. Assuming that inputs are correct (and given that some counties are mislabeled or inconsistent), returns NULL upon experiencing an error.

```
get.csv <- function(index, title, county, state) {
  search <- paste0(title, " ", county,", ", state)
  filter <- dplyr::filter(index$main, Title == search)
  clean <- stringr::str_replace_all(filter[, 1], "\\\\", "/")
  path <- str_trim(paste0("../final_project/raw_data/data/", clean), side = "right")

  output <- tryCatch(read.csv(path,
                              header = T,
                              stringsAsFactors = F), error = function(e) NULL)
  return(output)
}
head(get.csv(datasetIndex(), get.subject(datasetIndex(), 6), "Queens County", "NY"))
```

```
##         DATE VALUE
## 1 1969-01-01  5109
## 2 1970-01-01  5515
## 3 1971-01-01  5800
## 4 1972-01-01  6202
## 5 1973-01-01  6602
## 6 1974-01-01  7062
```

get.all.counties()

Returns all counties for a particular state, regardless of year.

```
get.all.counties <- function(index, state) {
  reference.title <- index$subjects$Title[5]
  filt.1 <- dplyr::filter(index$main, grepl(reference.title, Title))
  filt.2 <- dplyr::filter(filt.1, grepl(state, Title))
  counties <- filt.2$Title %>%
    gsub(reference.title, "", .) %>%
    gsub(state, "", .) %>%
    gsub(",", "", .) %>%
    str_trim(side = c("both"))
  return(counties)
}
get.all.counties(datasetIndex(), "AK")
```

```
##  [1] "Aleutians East Borough"
##  [2] "Aleutians West Census Area"
##  [3] "Anchorage Borough/municipality"
##  [4] "Bethel Census Area"
##  [5] "Bristol Bay Borough"
##  [6] "Denali Borough"
##  [7] "Dillingham Census Area"
##  [8] "Fairbanks North Star Borough"
##  [9] "Haines Borough"
## [10] "Hoonah-Angoon Census Area"
## [11] "Juneau City and Borough"
## [12] "Kenai Peninsula Borough"
## [13] "Ketchikan Gateway Borough"
## [14] "Kodiak Island Borough"
## [15] "Lake and Peninsula Borough"
## [16] "Matanuska-Susitna Borough"
## [17] "Nome Census Area"
## [18] "North Slope Borough"
## [19] "Northwest Arctic Borough"
## [20] "Petersburg Census Area"
## [21] "Prince of Wales-Hyder Census Area"
## [22] "Sitka City and Borough"
## [23] "Skagway Municipality"
## [24] "Southeast Fairbanks Census Area"
## [25] "Valdez-Cordova Census Area"
## [26] "Wade Hampton Census Area  (DISCONTINUED)"
## [27] "Wrangell City and Borough"
## [28] "Yakutat City and Borough"
## [29] "Yukon-Koyukuk Census Area"
```

```
length(get.all.counties(datasetIndex(), "AK"))
```

```
## [1] 29
```

get.filtered.counties()

Builds upon get.all.counties() by checking for county existence during a particular year (some counties only have data for particular years).

```
get.filtered.counties <- function(index, state, year) {
  counties.all <- get.all.counties(index, state)
  counties.filt <- rep(NA, length(counties.all))
  for (i in 1:length(counties.all)) {
    df <- get.csv(index, "Income Inequality in", counties.all[i], state)
    if (is.null(df)) {
      county <- NA
    } else {
      df$DATE <- as.numeric(gsub("-01-01", "", df$DATE))
      if (all(year %in% df$DATE)) {
        county <- counties.all[i]
      } else {
        county <- NA
      }
    }
    counties.filt[i] <- county
  }
  counties.filt <- counties.filt[!is.na(counties.filt)]
  return(counties.filt)
}
length(get.filtered.counties(datasetIndex(), "AK", 2015))
```

```
## [1] 28
```

Evidently, Alaska has 1 county that throws an error, and will therefore be excluded from other functions. Although non-trivial, the exclusions are not sufficiently large to affect results.

df.state.average()

Calculates state average for a particular variable for a particular state. Also instructive since it forms the foundation for other potential functions, such as one that calculates the sum of a particular variable. (Potential drawback beyond the scope of this project: counties are equal-weighted.)

```
df.state.average <- function(index, title, state, year) {
  counties <- get.filtered.counties(index, state, year)
  vector <- rep(NA, length(counties))
  for (i in 1:length(counties)) {
    df <- get.csv(index, title, counties[i], state)
    if (is.null(df)) {
      vector[i] <- 0
    } else {
      df$DATE <- as.numeric(gsub("-01-01", "", df$DATE))
      vector[i] <- dplyr::filter(df, DATE == year)[, 2]
    }
  }
  return(mean(vector))
}
df.state.average(datasetIndex(), get.subject(datasetIndex(), 6), "NY", 2015)
```

```
## [1] 45221.48
```

create.us.df()

Loops through df.state.average() function to create a 50-state data.frame for use in subsequent graphing functions.

```
create.us.df <- function(index, title, year) {
  df <- setNames(data.frame(matrix(ncol = 3, nrow = length(state.abb))), c("id", "value", "region"))
  for (i in 1:length(state.abb)) {
    df[i, "id"] <- state.abb[i]
    df[i, "region"] <- tolower(state.name[i])
    df[i, "value"] <- df.state.average(index, title, state.abb[i], year)
  }
  return(df)
}
head(create.us.df(datasetIndex(), get.subject(datasetIndex(), 6), 2015))
```

```
##   id    value     region
## 1 AL 33863.42    alabama
## 2 AK 48769.64     alaska
## 3 AZ 34010.93    arizona
## 4 AR 32782.71   arkansas
## 5 CA 47945.26 california
## 6 CO 44888.31   colorado
```

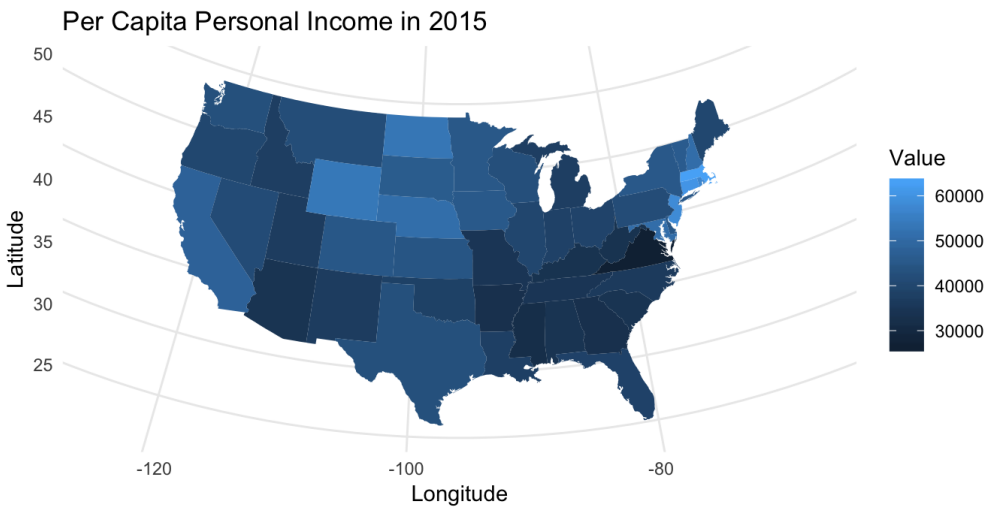
```
tail(create.us.df(datasetIndex(), get.subject(datasetIndex(), 6), 2015))
```

```
##    id    value        region
## 45 VT 46332.86       vermont
## 46 VA 26328.64      virginia
## 47 WA 42904.38    washington
## 48 WV 33489.62 west virginia
## 49 WI 42652.58     wisconsin
## 50 WY 53467.00       wyoming
```

create.us.heatmap()

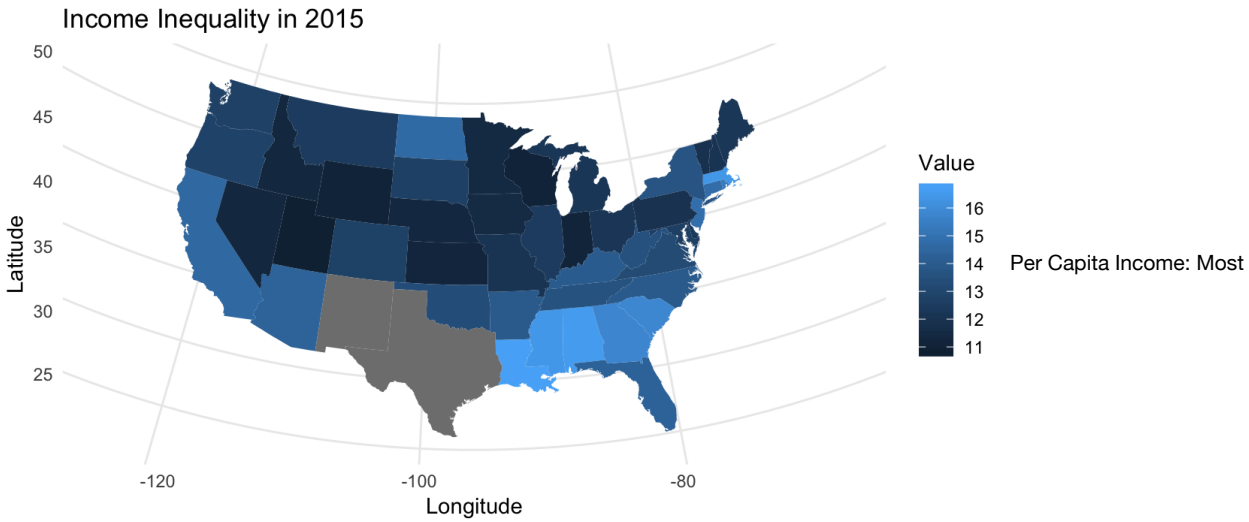
Creates a heatmap based on input variable. Provides an effective way of visualizing data. The power of create.us.heatmap() lies in the ability to analyze 31 additional variables. I limit myself to 2 variables for purposes of example. Note: Hawaii and Alaska are excluded on account of coding difficulties.

```
create.us.heatmap <- function(index, title, year) {
  df <- create.us.df(index, title, year)
  state.data <- map_data("state")
  df_map <- merge(state.data, df, sort = FALSE, by = "region")
  df_map <- df_map[order(df_map$order), ]

  map <- ggplot(data = df_map, mapping = aes(long, lat, group = group, fill = value)) +
    geom_polygon(color = NA) +
    coord_map(projection = "albers", lat0 = 39, lat1 = 45) +
    labs(title = paste0(title, " ", year), x = "Longitude", y = "Latitude", fill = "Value") +
    theme_minimal()
  return(map)
}
create.us.heatmap(datasetIndex(), get.subject(datasetIndex(), 6), 2015)
```




```
create.us.heatmap(datasetIndex(), get.subject(datasetIndex(), 5), 2015)
```

```
## Warning in mean.default(vector): argument is not numeric or logical:
## returning NA

## Warning in mean.default(vector): argument is not numeric or logical:
## returning NA
```



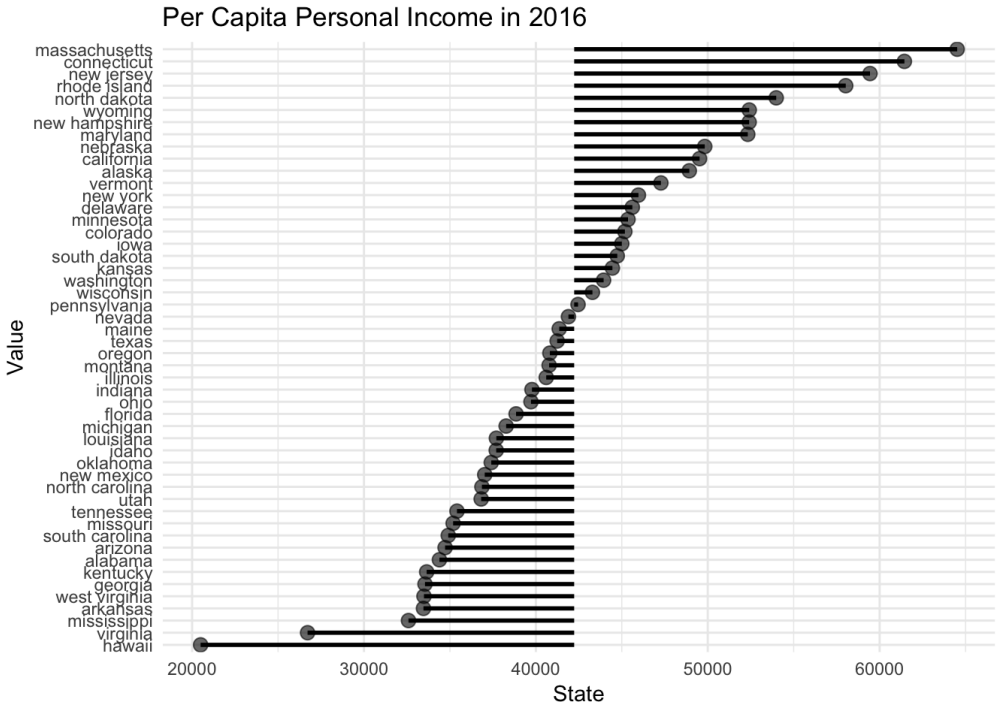
conclusions are relatively unsurprising. The southern part of the US is relatively poorer; Virginia appears to be the poorest state; the northeast, especially states such as Connecticut and Massachusetts, are relatively wealthier. Some unsurprising results include high relative incomes of Wyoming and North Dakota, which could be respectively explained by the prevalence of Jackson Hole and the shale boom.

Income Inequality: Again, results are unsurprising but instructive. The Midwestern states have the lowest income inequality, whereas the southern states, and especially Louisiana, have the highest levels of inequality. Graph suggests that a further avenue for more rigorous quantitative analysis would check whether per capita income levels have any causal relation to income inequality. Note: The grey states (New Mexico and Texas) suggest an error in csv retrieval; I would need to look into this issue further.

graph.diverging.lollipop()

Provides an alternative viewing of data. Graphically shows where a particular state is relative to the average.

```
graph.diverging.lollipop <- function(index, title, year) {
  df <- create.us.df(index, title, year)
  plot <- df %>%
    arrange(value) %>%
    mutate(region = factor(region, region)) %>%
    ggplot(aes(x = region, y = value)) +
    geom_segment( aes(x = region, xend = region, y = mean(df$value), yend = value), color = "black", size = 1) +
    geom_point(color = "black", size = 3, alpha = 0.6) +
    labs(title = paste0(title, " ", year), x = "Value", y = "State") +
    coord_flip() +
    theme_minimal()
  return(plot)
}
graph.diverging.lollipop(datasetIndex(), get.subject(datasetIndex(), 6), 2016)
```



Analysis: The same data in create.us.heatmap() but represented in a different way. Pennsylvania has the average per capita income level with, for instance, Massachusetts being the highest. Hawaii being the lowest is likely an error caused by expected discontinuities with its county-level data.