Data Wrangling

Jana Nassir

11/20/2021

library("tidyverse")

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.4 v dplyr 1.0.7  
## v tidyr 1.1.3 v stringr 1.4.0  
## v readr 2.0.1 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(readr)

income <- read\_csv("~/finalp/datasets/income.csv")

## Rows: 53 Columns: 38

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (38): State, 2020, 2019, 2018, 2017, 2016, 2015, 2014, 2013, 2012, 2011,...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

crime <- read\_csv("~/finalp/datasets/crime.csv")

## Rows: 2116 Columns: 15

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (3): state\_abbr, state\_name, caveats  
## dbl (12): year, population, violent\_crime, homicide, rape\_legacy, rape\_revis...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

crime\_dataset <- read\_csv("~/finalp/datasets/MODIFIED/crime\_dataset.csv")

## Rows: 1924 Columns: 14

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (2): state\_abbr, state  
## dbl (11): year, population, violent\_crime, homicide, rape\_legacy, robbery, a...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

crime\_dataset\_edu <- read\_csv("~/finalp/datasets/MODIFIED/crime\_dataset\_edu.csv")

## Rows: 688 Columns: 15

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (2): state\_abbr, state  
## dbl (12): year, population, violent\_crime, homicide, rape\_legacy, robbery, a...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

crimes <- read\_csv("~/finalp/datasets/MODIFIED/crimes.csv")

## Rows: 1861 Columns: 13

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (2): state\_abbr, state  
## dbl (11): year, population, violent\_crime, homicide, rape\_legacy, robbery, a...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

edu\_attainment <- read\_csv("~/finalp/datasets/edu\_attainment.csv")

## Rows: 636 Columns: 3

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (1): state  
## dbl (2): year, attainment rate

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

edu\_finance <- read\_csv("~/finalp/datasets/MODIFIED/edu\_finance.csv")

## Rows: 477 Columns: 18

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (2): state, state\_abbr  
## dbl (15): year, ENROLL, TOTAL\_REVENUE, TOTAL\_EXPENDITURE, population, violen...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

yearly\_income <- read\_csv("~/finalp/datasets/MODIFIED/yearly\_income.csv")

## Rows: 1924 Columns: 3

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (1): state  
## dbl (1): year

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

yearly\_unemployment <- read\_csv("~/finalp/datasets/MODIFIED/yearly\_unemployment.csv")

## Rows: 2028 Columns: 3

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (1): state  
## dbl (2): year, unemployment

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

education\_finance <- read\_csv("~/finalp/datasets/education\_finance.csv")

## Rows: 1275 Columns: 5

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (1): STATE  
## dbl (4): YEAR, ENROLL, TOTAL\_REVENUE, TOTAL\_EXPENDITURE

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

unemployment <- read\_csv("~/finalp/datasets/unemployment.csv")

## Rows: 52 Columns: 40

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (1): state  
## dbl (39): 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, ...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

income <- filter(income, !is.na(State))

Fil

crime <- rename(crime, state = state\_name)

crimes <- filter(crime, year>=1984)  
crimes <- subset (crimes, select = -c(rape\_revised, caveats))

yearly\_income <- gather(income, year, income, 2:38)   
yearly\_income <- rename(yearly\_income, state = State)  
yearly\_income$year <- as.numeric(yearly\_income$year)

yearly\_unemployment <- gather(unemployment, year, unemployment, 2:40)  
yearly\_unemployment <- filter(yearly\_unemployment, year>=1984)  
yearly\_unemployment$year <- as.numeric(yearly\_unemployment$year)

crime\_dataset <- full\_join (crimes, yearly\_income, by = c("state", "year"))  
crime\_dataset <- full\_join (crime\_dataset, yearly\_unemployment, by = c("state", "year"))  
crime\_dataset <- filter(crime\_dataset, year<=2019)

edu\_attainment[edu\_attainment == "United States (with PR)"] <- "United States"

crime\_dataset\_edu <- crime\_dataset  
crime\_dataset\_edu <- full\_join (crime\_dataset\_edu, edu\_attainment, by = c("state", "year"))  
crime\_dataset\_edu <- filter(crime\_dataset\_edu, year >= 2008, year <2020)

edu\_finance <- education\_finance  
edu\_finance <- rename(edu\_finance, year = YEAR, state = STATE)  
edu\_finance <- full\_join(edu\_finance, crime\_dataset\_edu, by = c("state", "year"))  
edu\_finance <- filter(edu\_finance, year >= 2008, year < 2017)

crime$state[is.na(crime$state)] <- "United States"  
crimes$state[is.na(crimes$state)] <- "United States"  
crime\_dataset\_edu$state[is.na(crime\_dataset\_edu$state)] <- "United States"  
crime\_dataset$state[is.na(crime\_dataset$state)] <- "United States"  
  
crime$state\_abbr[is.na(crime$state\_abbr)] <- "USA"  
crimes$state\_abbr[is.na(crimes$state\_abbr)] <- "USA"  
crime\_dataset\_edu$state\_abbr[is.na(crime\_dataset\_edu$state\_abbr)] <- "USA"  
crime\_dataset$state\_abbr[is.na(crime\_dataset$state\_abbr)] <- "USA"

write.table (crime\_dataset, file = “/Users/asus/Documents/finalp/datasets/MODIFIED/crime\_dataset.csv”, row.names = F, sep = “,”) write.table (crime\_dataset\_edu, file = “/Users/asus/Documents/finalp/datasets/MODIFIED/crime\_dataset\_edu.csv”, row.names = F, sep = “,”) write.table (crimes, file = “/Users/asus/Documents/finalp/datasets/MODIFIED/crimes.csv”, row.names = F, sep = “,”) write.table (yearly\_income, file = “/Users/asus/Documents/finalp/datasets/MODIFIED/yearly\_income.csv”, row.names = F, sep = “,”) write.table (yearly\_unemployment, file = “/Users/asus/Documents/finalp/datasets/MODIFIED/yearly\_unemployment.csv”, row.names = F, sep = “,”) write.table (edu\_finance, file = “/Users/asus/Documents/finalp/datasets/MODIFIED/edu\_finance.csv”, row.names = F, sep = “,”)