ANA 515 Assignment 2, Loading, Saving, and Describing Data

Christopher Spann

11/10/2021

```
knitr::opts_chunk$set(echo = TRUE)
```

```
#This next chunk is a description of the data.
```

The police killings data set can be found on the Github data repository. The data contains information on police killings in the United States during the first 5 months of 2015. Records come from the Guardian’s database on police killings as well as census data from the American Community Survey that was combined to allow users of the data to look at demographic and economic information about the people and neighborhoods involved. The data to be analyzed comes in the form of a csv as listed on Github, this csv makes the data comma delimited. The data is reporting relevant information about individuals killed by police or other law enforcement agencies in the United States, and the data was collected by the Guardian as part of The Counted project. As mentioned in a FiveThirtyEight article describing the project, “Where Police Have Killed Americans in 2015,” official statistics on police killings are often inaccurate or flawed, so the purpose of The Counted is to build a data set by combining media coverage, reader submissions, and other open source efforts that are verified in order to bring better transparency to these types of problems in the United States. More specifically, users of the data can look into research questions such as:

Do police killings happen more often in poor or wealthy neighborhoods?

Is a particular race more likely to be involved in a police killing?

Is being armed a significant factor that results in an individual being killed by a law enforcement officer?

```
#This next chunk is to read the data into R. The data is stored in csv format on the Github site, so I will be using
```

```
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
url <- "https://raw.githubusercontent.com/fivethirtyeight/data/master/police-killings/police_killings.csv"
police_killings <- read_csv(url)
```

```
## Rows: 467 Columns: 34
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (17): name, age, gender, raceethnicity, month, streetaddress, city, stat...
## dbl (17): day, year, latitude, longitude, state_fp, county_fp, tract_ce, geo...
```

```
##
## 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.
```

```
#This next chunk is to clean some of the data in R. We will re-name some field names to be clearer. Then we will conc
```

```
names(police_killings)[names(police_killings) == 'cause'] <- 'cause_of_death'
names(police_killings)[names(police_killings) == 'pop'] <- 'tract_population'
names(police_killings)[names(police_killings) == 'share_white'] <- 'pop_percentage_white'
names(police_killings)[names(police_killings) == 'share_black'] <- 'pop_percentage_black'
names(police_killings)[names(police_killings) == 'share_hispanic'] <- 'pop_percentage_hispanic'
names(police_killings)[names(police_killings) == 'p_income'] <- 'median_personal_income_tract'
names(police_killings)[names(police_killings) == 'h_income'] <- 'median_household_income_tract'
names(police_killings)[names(police_killings) == 'urate'] <- 'unemployment_rate_tract'
names(police_killings)[names(police_killings) == 'county_bucket'] <- 'county_household_income_quintile'
names(police_killings)[names(police_killings) == 'nat_bucket'] <- 'nation_household_income_quintile'
names(police_killings)[names(police_killings) == 'county_income'] <- 'county_median_income'
names(police_killings)[names(police_killings) == 'comp_income'] <- 'household_income/county_median_income'
names(police_killings)[names(police_killings) == 'pov'] <- 'poverty_rate_tract'
library(tidyverse)
```

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

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v stringr 1.4.0
## v tidyr   1.1.4     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
police_killings$date <- paste(police_killings$year, police_killings$month, police_killings$day, sep="-") %>% ymd() %>
```

```
#This next chunk is to describe some characteristics of the data. I will be using inline code to describe the number
```

This dataframe has 467 rows and 35 columns. The original data set has 34 columns, but when cleaning the data, I added another column to concatenate all of the date fields into one column. The names of the columns and a brief description of each are in the table below:

```
column_names <- c(ls(police_killings))
column_description <- c("Age of deceased individual", "Whether deceased individual was armed", "Cause of death for de
table.df <- data.frame(column_names, column_description)
knitr::kable(table.df, "simple", col.names = c("Column Name", "Column Description"), align = c("c", "c"))
```

Column Name	Column Description
age	Age of deceased individual
armed	Whether deceased individual was armed
cause_of_death	Cause of death for deceased individual
city	City where incident occured

Column Name	Column Description
college	Share of 25+ population with BA or higher
county_fp	County FIPS code
county_household_income_quintile	Household income, quintile within county
county_id	Combined county ID code
county_median_income	County-level median household income
date	Date incident occured
day	Day of incident
gender	Gender of deceased individual
geo_id	Combined tract ID code
household_income/county_median_income	Tract-level median household income divided by county-level median household income
latitude	Latitude, geocoded from address
lawenforcementagency	Agency involved in incident
longitude	Longitude, geocoded from address
median_household_income_tract	Tract-level median household income
median_personal_income_tract	Tract-level median personal income
month	Month of killing
name	Name of deceased individual
namelsad	Tract description
nation_household_income_quintile	Household income, quintile nationally
pop_percentage_black	Share of population that is black (alone, not in combination)
pop_percentage_hispanic	Share of population that is Hispanic/Latino (any race)
pop_percentage_white	Share of population that is non-Hispanic white
poverty_rate_tract	Tract-level poverty rate (official)
raceethnicity	Race/ethnicity of deceased individual
state	State where incident occurred
state_fp	State FIPS code
streetaddress	Address/intersection where incident occurred
tract_ce	Tract ID code
tract_population	Population of Tract where incident occured

Column Name	Column Description
unemployment_rate_tract	Tract-level unemployment rate
year	Year of incident

```
#I have decided to show summaries of county_median_income, median_household_income_tract, and unemployment_rate_tract
```

```
subset_police_killings <- police_killings[,c("county_median_income", "median_household_income_tract", "unemployment_r
```

```
#This next chunk is to provide some summary statistics of 3 columns in the dataframe.
```

```
#Summary of all variables
summary_subset_police_killings <- summary(subset_police_killings)
summary_subset_police_killings
```

```
##  county_median_income median_household_income_tract unemployment_rate_tract
##  Min.   : 22545       Min.   : 10290                Min.   :0.01133
##  1st Qu.: 43804       1st Qu.: 32625                1st Qu.:0.06859
##  Median : 50856       Median : 42759                Median :0.10518
##  Mean   : 52527       Mean   : 46627                Mean   :0.11740
##  3rd Qu.: 56832       3rd Qu.: 56190                3rd Qu.:0.14083
##  Max.   :110292       Max.   :142500                Max.   :0.50761
##                       NA's   :2                     NA's   :2
```

```
#Calculations for summary statistics
mean_county_median_income <- mean(subset_police_killings$`county_median_income`, na.rm = TRUE)
mean_median_household_income_tract <- mean(subset_police_killings$`median_household_income_tract`, na.rm = TRUE)
mean_unemployment_rate_tract <- mean(subset_police_killings$`unemployment_rate_tract`, na.rm = TRUE)
min_county_median_income <- min(subset_police_killings$`county_median_income`, na.rm = TRUE)
min_median_household_income_tract <- min(subset_police_killings$`median_household_income_tract`, na.rm = TRUE)
min_unemployment_rate_tract <- min(subset_police_killings$`unemployment_rate_tract`, na.rm = TRUE)
max_county_median_income <- max(subset_police_killings$`county_median_income`, na.rm = TRUE)
max_median_household_income_tract <- max(subset_police_killings$`median_household_income_tract`, na.rm = TRUE)
max_unemployment_rate_tract <- max(subset_police_killings$`unemployment_rate_tract`, na.rm = TRUE)
missing_values_county_median_income <- sum(is.na(subset_police_killings$`county_median_income`))
missing_values_median_household_income_tract <- sum(is.na(subset_police_killings$`median_household_income_tract`))
missing_values_unemployment_rate_tract <- sum(is.na(subset_police_killings$`unemployment_rate_tract`))

#Mean Values of the Columns
mean_county_median_income
```

```
## [1] 52527.33
```

```
mean_median_household_income_tract
```

```
## [1] 46627.18
```

```
mean_unemployment_rate_tract
```

```
## [1] 0.1173994
```

```
#Minimum Values of the Columns
min_county_median_income
```

```
## [1] 22545
```

```
min_median_household_income_tract
```

```
## [1] 10290
```

```
min_unemployment_rate_tract
```

```
## [1] 0.01133501
```

```
#Maximum Values of the Columns
max_county_median_income
```

```
## [1] 110292
```

```
max_median_household_income_tract
```

```
## [1] 142500
```

```
max_unemployment_rate_tract
```

```
## [1] 0.5076142
```

```
#Number of Missing Values
missing_values_county_median_income
```

```
## [1] 0
```

```
missing_values_median_household_income_tract
```

```
## [1] 2
```

```
missing_values_unemployment_rate_tract
```

```
## [1] 2
```