EAS 345: Phase 01
Systemic racism in America
Aghose, 10/09/2020

Author: Akash Ghose.

Area of research: Social issues surrounding racial tension in America.

Title of project: Does systemic racism exist in America?

Potential clients: People who wish to view the data about social disparities in America

Potential sponsors: People who wish to inform others of the existence (or non-existence) of social disparities in America

Potential data sources:

FBI Crime/Arrest data:

https://ucr.fbi.gov/crime-in-the-u.s/2019/crime-in-the-u.s.-2019/tables/table-49

https://ucr.fbi.gov/crime-in-the-u.s/2018/crime-in-the-u.s.-2018/tables/table-49

**Replace 'YEAR' with actual year, dating back until 1995.

US Sentencing commission data:

https://www.ussc.gov/sites/default/files/pdf/research-and-publications/annual-reports-and-sourcebooks/2019/2019-Annual-Report-and-Sourcebook.pdf

^Of interest here, 56% of federal offenders were Hispanic. Even though Hispanics make up a very small percentage of the US population. How is this possible?

US census bureau population data for the last 10 years:

https://www.census.gov/data/tables/time-series/demo/popest/2010s-national-detail.html

Consumer Financial Protection Bureau data:

https://www.consumerfinance.gov/data-research/hmda/historic-data/

Goal:

In the Oct. 7, 2020 Vice Presidential debate, Mike Pence declared that systematic racism does not exist in America. My goal with this project is to aggregate data to prove or disprove that notion. I will try to do so by looking at the public data available and attempt to compare race vs arrests vs population size, race vs severity of crime vs incarceration rate/time, race vs mortgage loans denied/accepted, race vs income/job opportunities, race vs educational opportunities. In the end, I hope to be able to use this data to paint a very clear and coherent picture about social disparities in America, and aggregate it all in a very clean and concise place for all to view.

Phase 02: Data Collection

Akash Ghose, 10/23/20

Crime Related Data sources

Data source 01:

Name of files: 2019_FBI_arrests_by_race_total.csv
2019_FBI_arrests_by_race_under18.csv
2019_FBI_arrests_by_race_18_and_over.csv

Source: https://ucr.fbi.gov/crime-in-the-u.s/2019/crime-in-the-u.s.-2019/topic-pages/tables/table-49

Details: As the name suggests, this contains data about arrests in 2019. It contains details such as the race of the perpetrators and the type of crime they were arrested for.

**Of note: I currently only downloaded the data provided for 2019, because I am not sure I need more than one year's data. So, for the sake of cleanliness, I have limited the data. However, if in the future, I need/want to get more data, it can be obtained with little to no effort.

Finance Related Data sources

Data source 01:

Name of file: NFWBS_PUF_2016_data_readable.csv

Source: https://www.consumerfinance.gov/data-research/financial-well-being-survey-data/

Details: This is the National Financial Wellbeing Survey data from a survey that was conducted in 2016. This contains details about respondents and respondents' financial well-being, including characteristics like income, age, race, savings, past financial experiences, financial skills, behaviors, attitudes ect.

**Of note: The original file I downloaded was: NFWBS_PUF_2016_data.csv. I used NFWBS_PUF_2016_read_in_R.R to read the file and then write it into the more readable csv.

Data source 02:

Name of file: hmda_2017_nationwide_all-records_labels.csv

Source: https://www.consumerfinance.gov/data-research/hmda/historic-data/?geo=nationwide&records=all-records&field_descriptions=labels

Details: This contains all the mortgage applications filed in 2017. It contains data about the applications and applicants, including details such as applicants' demographics and whether the application was accepted or rejected.

US Population Data source

Data source 01:

Name of file: US_population_est_2010-2019.csv

Source: https://www.census.gov/data/tables/time-series/demo/popest/2010s-national-detail.html#par_textimage_1537638156

Details: Contains US population estimates from 2010-2019. Includes separation by race as well as the totals.

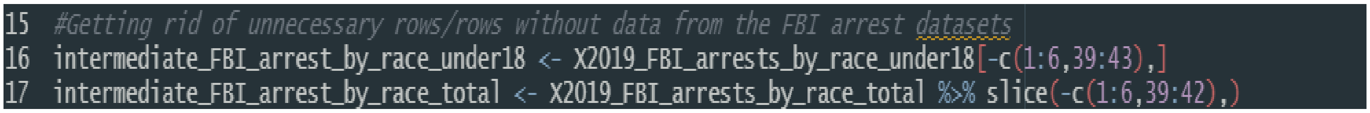
UB box that contains my data:

https://buffalo.box.com/s/9231grwf8pw2sjs5jhkvodavc6z21gt3

Phase: 03, Data cleaning and processing

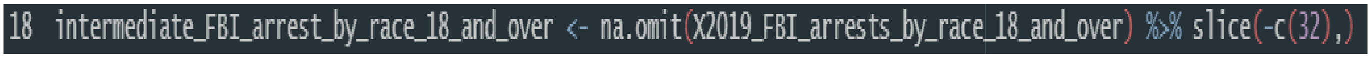
Steps I have taken to clean my data:

- Dropped irrelevant rows.
    - I have dropped rows using both the native “-“ operator and also the dplyr slice method.
    - I did this because these rows were unnecessary



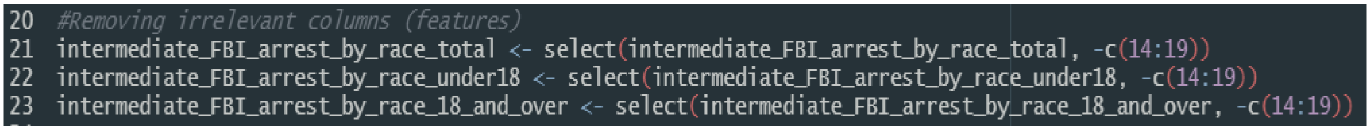
```
#Getting rid of unnecessary rows/rows without data from the FBI arrest datasets
intermediate_FBI_arrest_by_race_under18 <- X2019_FBI_arrests_by_race_under18[-c(1:6,39:43),]
intermediate_FBI_arrest_by_race_total <- X2019_FBI_arrests_by_race_total %>% slice(-c(1:6,39:42),)
```

- Dropped rows with NA values.
    - I did this because these rows were unnecessary



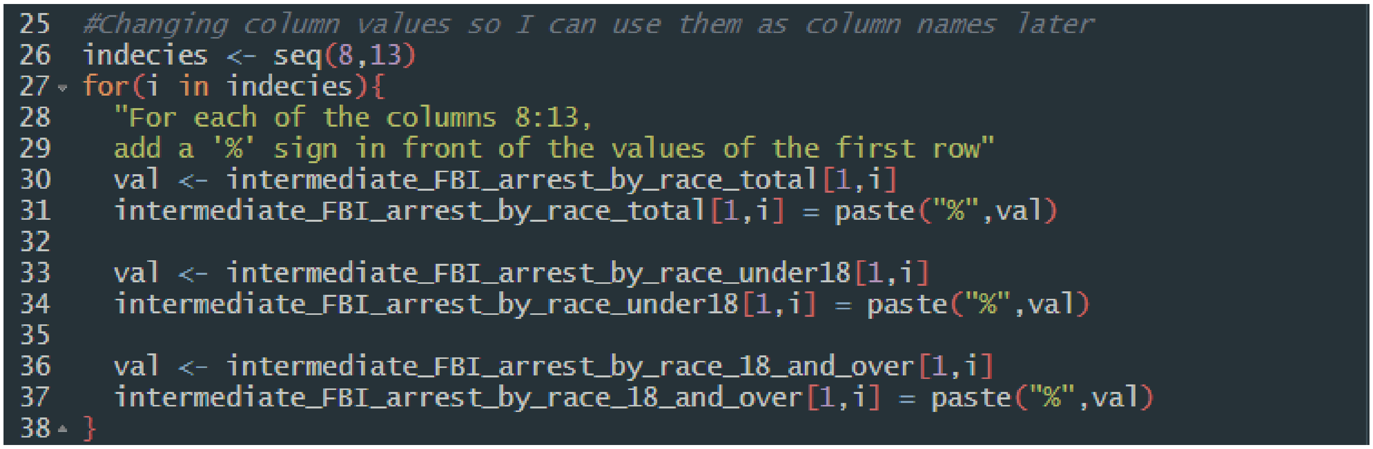
```
intermediate_FBI_arrest_by_race_18_and_over <- na.omit(X2019_FBI_arrests_by_race_18_and_over) %>% slice(-c(32),)
```

- Dropped irrelevant columns.
    - I dropped ethnicity data because I do not need them. Race data is sufficient for my intents and purposes.



```
#Removing irrelevant columns (features)
intermediate_FBI_arrest_by_race_total <- select(intermediate_FBI_arrest_by_race_total, -c(14:19))
intermediate_FBI_arrest_by_race_under18 <- select(intermediate_FBI_arrest_by_race_under18, -c(14:19))
intermediate_FBI_arrest_by_race_18_and_over <- select(intermediate_FBI_arrest_by_race_18_and_over, -c(14:19))
```

- Changed column values so I can use them as column names
    - I wanted to assign my first row to be column names (as that is how the data is mean to be read), however, as it stood, R wouldn’t let me do so because it wanted the column names to be unique, and the values in the first row were not unique.
    - So, I had to change the values (by adding “%” in front of values that need them) so that it can be read the way it was meant to be read

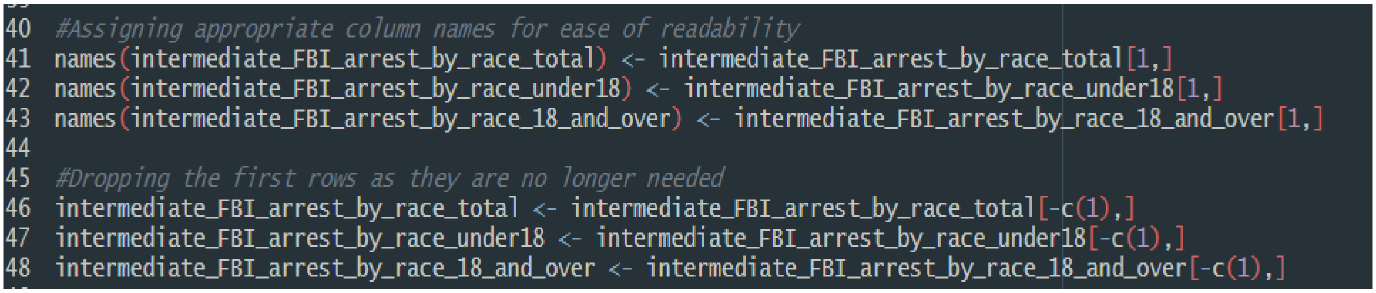


```r
#Changing column values so I can use them as column names later
indecies <- seq(8,13)
for(i in indecies){
  "For each of the columns 8:13,
  add a '%' sign in front of the values of the first row"
  val <- intermediate_FBI_arrest_by_race_total[1,i]
  intermediate_FBI_arrest_by_race_total[1,i] = paste("%",val)

  val <- intermediate_FBI_arrest_by_race_under18[1,i]
  intermediate_FBI_arrest_by_race_under18[1,i] = paste("%",val)

  val <- intermediate_FBI_arrest_by_race_18_and_over[1,i]
  intermediate_FBI_arrest_by_race_18_and_over[1,i] = paste("%",val)
}
```

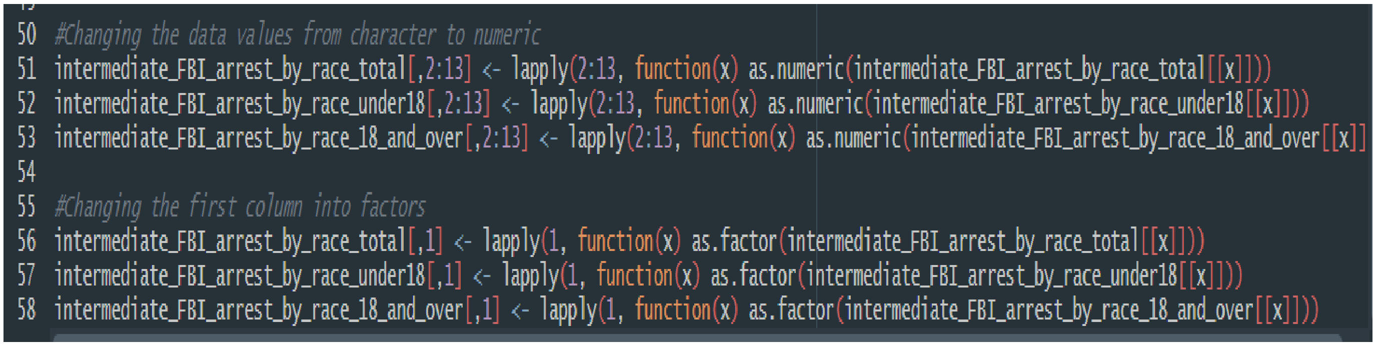
- Assigned first row to be column names
    - Instead of column names being just numbers, they are now properly labeled
    - Also, the first row (which contained what are now the column names) is dropped as it becomes redundant here.



```r
#Assigning appropriate column names for ease of readability
names(intermediate_FBI_arrest_by_race_total) <- intermediate_FBI_arrest_by_race_total[1,]
names(intermediate_FBI_arrest_by_race_under18) <- intermediate_FBI_arrest_by_race_under18[1,]
names(intermediate_FBI_arrest_by_race_18_and_over) <- intermediate_FBI_arrest_by_race_18_and_over[1,]

#Dropping the first rows as they are no longer needed
intermediate_FBI_arrest_by_race_total <- intermediate_FBI_arrest_by_race_total[-c(1),]
intermediate_FBI_arrest_by_race_under18 <- intermediate_FBI_arrest_by_race_under18[-c(1),]
intermediate_FBI_arrest_by_race_18_and_over <- intermediate_FBI_arrest_by_race_18_and_over[-c(1),]
```

- Changed data values from character to numeric
- Changed first column from characters to factors
    - Both of the last two changes were done so that I have an easier time analyzing the data in the EDA phase



```r
#Changing the data values from character to numeric
intermediate_FBI_arrest_by_race_total[,2:13] <- lapply(2:13, function(x) as.numeric(intermediate_FBI_arrest_by_race_total[[x]]))
intermediate_FBI_arrest_by_race_under18[,2:13] <- lapply(2:13, function(x) as.numeric(intermediate_FBI_arrest_by_race_under18[[x]]))
intermediate_FBI_arrest_by_race_18_and_over[,2:13] <- lapply(2:13, function(x) as.numeric(intermediate_FBI_arrest_by_race_18_and_over[[x]]

#Changing the first column into factors
intermediate_FBI_arrest_by_race_total[,1] <- lapply(1, function(x) as.factor(intermediate_FBI_arrest_by_race_total[[x]]))
intermediate_FBI_arrest_by_race_under18[,1] <- lapply(1, function(x) as.factor(intermediate_FBI_arrest_by_race_under18[[x]]))
intermediate_FBI_arrest_by_race_18_and_over[,1] <- lapply(1, function(x) as.factor(intermediate_FBI_arrest_by_race_18_and_over[[x]]))
```