

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/9231grwf8pw2sis5jhkvodavc6z21gt3>

Phase: 03, Data cleaning and processing

Steps I have taken to clean my data:

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

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

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

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

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

```
20 #Removing irrelevant columns (features)
21 intermediate_FBI_arrest_by_race_total <- select(intermediate_FBI_arrest_by_race_total, -c(14:19))
22 intermediate_FBI_arrest_by_race_under18 <- select(intermediate_FBI_arrest_by_race_under18, -c(14:19))
23 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
 - o I wanted to assign my first row to be column names (as that is how the data is meant 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.
 - o 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

```

25 #Changing column values so I can use them as column names later
26 indecies <- seq(8,13)
27 for(i in indecies){
28   "For each of the columns 8:13,
29   add a '%' sign in front of the values of the first row"
30   val <- intermediate_FBI_arrest_by_race_total[1,i]
31   intermediate_FBI_arrest_by_race_total[1,i] = paste("%",val)
32
33   val <- intermediate_FBI_arrest_by_race_under18[1,i]
34   intermediate_FBI_arrest_by_race_under18[1,i] = paste("%",val)
35
36   val <- intermediate_FBI_arrest_by_race_18_and_over[1,i]
37   intermediate_FBI_arrest_by_race_18_and_over[1,i] = paste("%",val)
38 }

```

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

```

40 #Assigning appropriate column names for ease of readability
41 names(intermediate_FBI_arrest_by_race_total) <- intermediate_FBI_arrest_by_race_total[1,]
42 names(intermediate_FBI_arrest_by_race_under18) <- intermediate_FBI_arrest_by_race_under18[1,]
43 names(intermediate_FBI_arrest_by_race_18_and_over) <- intermediate_FBI_arrest_by_race_18_and_over[1,]
44
45 #Dropping the first rows as they are no longer needed
46 intermediate_FBI_arrest_by_race_total <- intermediate_FBI_arrest_by_race_total[-c(1),]
47 intermediate_FBI_arrest_by_race_under18 <- intermediate_FBI_arrest_by_race_under18[-c(1),]
48 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
 - o Both of the last two changes were done so that I have an easier time analyzing the data in the EDA phase

```

50 #Changing the data values from character to numeric
51 intermediate_FBI_arrest_by_race_total[,2:13] <- lapply(2:13, function(x) as.numeric(intermediate_FBI_arrest_by_race_total[[x]]))
52 intermediate_FBI_arrest_by_race_under18[,2:13] <- lapply(2:13, function(x) as.numeric(intermediate_FBI_arrest_by_race_under18[[x]]))
53 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]]))
54
55 #Changing the first column into factors
56 intermediate_FBI_arrest_by_race_total[,1] <- lapply(1, function(x) as.factor(intermediate_FBI_arrest_by_race_total[[x]]))
57 intermediate_FBI_arrest_by_race_under18[,1] <- lapply(1, function(x) as.factor(intermediate_FBI_arrest_by_race_under18[[x]]))
58 intermediate_FBI_arrest_by_race_18_and_over[,1] <- lapply(1, function(x) as.factor(intermediate_FBI_arrest_by_race_18_and_over[[x]]))

```

Phase: 04, EDA and Data engineering

List of EDA steps I have taken:

- Used
 - Head()
 - Tail()
 - Summary()
 - Colnames()
 - View()
- In various places throughout this phase and throughout the previous data cleaning phase to get a better understanding of the data I'm dealing with and figure out what to do next. For example, head and tail were useful in figuring out quickly whether the top and bottom of the data were similar, whether there were any inconsistencies that needed to be dealt with. Summary() gave me a whole lot of useful information. To start, it would tell me quickly if the data I am dealing with numbers as it seems or characters. With my mortgage data, summary() told me that I have 51 NAs in my "cleaned" mortgage_data\$loan_amount_000s. It also told me that the minimum amount of loan requested was \$1000 and maximum was \$30,000,000, which I thought was interesting. Colnames() was needed because I realized that some of the column names were not what they seemed. For example, in my FBI arrests data, I see a column name as "Black or African American" when I look at it with view(), but the actual column name is "Black or\r\nAfrican\r\nAmerican". View() was used frequently not only to get a wholistic idea of the raw data, but also to see if the changes I was making while cleaning was behaving the way I expected them to.

```

19 "Initial EDA of financial_well_being_survey"
20 summary(financial_well_being_survey)
21 colnames(financial_well_being_survey)
22
23 "Initial EDA of mortgage_data"
24 colnames(mortgage_data)
25 head(mortgage_data)
26 tail(mortgage_data)
27 summary(mortgage_data)

```

```

19 #Initial look at the cleaned data sets
20 View(FBI_arrest_by_race_18_and_over)
21 View(FBI_arrest_by_race_under18)
22 View(FBI_arrest_by_race_total)
23 head(FBI_arrest_by_race_total)
24 colnames(FBI_arrest_by_race_total)

```

```

> summary(mortgage_data)
 loan_amount_000s  preapproval_name  preapproval  action_taken_name  action_taken  applicant_race_name_1  applicant_race_1
Min. : 1.0  Length:138236  Min. :1.000  Length:138236  Min. :1.000  Length:138236  Min. :1.000
1st Qu.: 100.0  Class :character 1st Qu.:3.000  Class :character 1st Qu.:1.000  Class :character 1st Qu.:5.000
Median : 177.0  Mode :character  Median :3.000  Mode :character  Median :2.000  Mode :character  Median :5.000
Mean : 221.3  Mean :2.786  Mean :2.416  Mean :2.416  Mean :4.521
3rd Qu.: 285.0  3rd Qu.:3.000  3rd Qu.:3.000  3rd Qu.:3.000  3rd Qu.:5.000
Max. : 30000.0  Max. :3.000  Max. :8.000  Max. :8.000  Max. :7.000
NA's :51

applicant_race_name_2  applicant_race_2  applicant_race_name_3  applicant_race_3  applicant_race_name_4  applicant_race_4
Length:138236  Min. :1.00  Length:138236  Min. :1.00  Mode:logical  Mode:logical
Class :character 1st Qu.:4.00  Class :character 1st Qu.:3.00  NA's:138236  NA's:138236
Mode :character  Median :5.00  Mode :character  Median :5.00
Mean :4.31  Mean :4.36
3rd Qu.:5.00  3rd Qu.:5.00
Max. :5.00  Max. :5.00
NA's :136643  NA's :138059

applicant_race_name_5  applicant_race_5  co_applicant_race_name_1  co_applicant_race_1  co_applicant_race_name_2  co_applicant_race_2
Mode:logical  Length:138236  Min. :1.000  Length:138236  Min. :1.00
NA's:138236  TRUE:3  Class :character 1st Qu.:5.000  Class :character 1st Qu.:4.00
NA's:138233  Mode :character  Median :8.000  Mode :character  Median :5.00
Mean :6.647  Mean :4.45
3rd Qu.:8.000  3rd Qu.:5.00
Max. :8.000  Max. :5.00
NA's :137732

co_applicant_race_name_3  co_applicant_race_3  co_applicant_race_name_4  co_applicant_race_4  co_applicant_race_name_5  co_applicant_race_5
Mode:logical  Mode:logical  Mode:logical  Mode:logical  Mode:logical  Mode:logical
NA's:138236  TRUE:1  NA's:138236  NA's:138236  NA's:138236  TRUE:1
NA's:138235  NA's:138235

applicant_sex_name  applicant_sex  co_applicant_sex_name  co_applicant_sex  applicant_income_000s  denial_reason_name_1  denial_reason_1
Length:138236  Min. :1.000  Length:138236  Min. :1.000  Min. : 1.0  Length:138236  Min. :1.00
Class :character 1st Qu.:1.000  Class :character 1st Qu.:2.000  1st Qu.: 49.0  Class :character 1st Qu.:1.00
Mode :character  Median :1.000  Mode :character  Median :5.000  Median : 75.0  Mode :character  Median :3.00
Mean :1.331  Mean :3.711  Mean :109.9  Mean :3.711  Mean :3.72
3rd Qu.:2.000  3rd Qu.:5.000  3rd Qu.:118.0  3rd Qu.:5.000  3rd Qu.:5.00
Max. :2.000  Max. :5.000  Max. :175000.0  Max. :5.000  Max. :9.00
NA's :  NA's :5992  NA's :111388

denial_reason_name_2  denial_reason_2  denial_reason_name_3  denial_reason_3  population  minority_population  hud_median_family_income
Length:138236  Min. :1.00  Length:138236  Min. :1.00  Min. : 0  Min. : 0.00  Min. : 20500
Class :character 1st Qu.:3.00  Class :character 1st Qu.:3.00  1st Qu.: 3702  1st Qu.: 10.02  1st Qu.: 62600
Mode :character  Median :3.00  Mode :character  Median :5.00  Median : 5009  Median : 23.04  Median : 72400
Mean :4.28  Mean :5.55  Mean : 5751  Mean : 33.58  Mean : 70592
3rd Qu.:6.00  3rd Qu.:9.00  3rd Qu.: 6631  3rd Qu.: 51.72  3rd Qu.: 77500
Max. :9.00  Max. :9.00  Max. :53812  Max. :100.00  Max. :131500
NA's :132297  NA's :137367  NA's :577  NA's :577  NA's :577

tract_to_msamd_income  number_of_owner_occupied_units  number_of_1_to_4_family_units

```

- Used dplyr techniques such as:
 - o Select()
 - Select was used in a few different places primarily to drop columns that were unnecessary or those that became obsolete
 - o Slice()
 - Slice() was used a couple of times to get rid of unwanted rows

- %>%
 - The pipe operator was used extensively throughout the last two phases for a multitude of reasons, including for simplicity and sake of readability
- Mutate()
 - Was used to add my own column with information about all non-white races in my FBI arrest datasets
- Relocate()
 - Was used to re-arrange the placement of my recently added column for readability purposes.

```

26 #Combine non-white races into a single column
27 #Combine %non-whites into a single column
28 #Drop first column because it is useless
29 #relocate the new rows to a better position for readability
30 alt_FBI_arrest_by_race_total <-FBI_arrest_by_race_total %>%
31   mutate(Non_white = .[[5]]+. [[6]]+. [[7]]+. [[8]]) %>%
32   mutate("% Non_white" = .[[11]]+. [[12]]+. [[13]]+. [[14]]) %>%
33   select(-c(1)) %>%
34   relocate(Non_white, .after = "White") %>%
35   relocate("% Non_white", .after = "% White")
36 alt_FBI_arrest_by_race_under18 <-FBI_arrest_by_race_under18 %>%
37   mutate(Non_white = .[[5]]+. [[6]]+. [[7]]+. [[8]]) %>%
38   mutate("% Non_white" = .[[11]]+. [[12]]+. [[13]]+. [[14]]) %>%
39   select(-c(1)) %>%
40   relocate(Non_white, .after = "White") %>%
41   relocate("% Non_white", .after = "% White")
42 alt_FBI_arrest_by_race_18_and_over <-FBI_arrest_by_race_18_and_over %>%
43   mutate(Non_white = .[[5]]+. [[6]]+. [[7]]+. [[8]]) %>%
44   mutate("% Non_white" = .[[11]]+. [[12]]+. [[13]]+. [[14]]) %>%
45   select(-c(1)) %>%
46   relocate(Non_white, .after = "White") %>%
47   relocate("% Non_white", .after = "% White")
48

```

- Used the following techniques to create graphs:
 - Boxplot()
 - I initially created the boxplot in hopes of learning some useful information about the loan amounts that were requested. I found none, so I moved on to geom_boxplot
 - Geom_boxplot()
 - After tinkering with this a little bit, I was able to graph something that actually showed me useful information
 - Geom_bar()

- The geom barplot was used to draw the number of loans that were accepted and denied and with colors to show much of it belonged to each race

```
60 mortgage_data <- select(mortgage_data, -c(1))
61
62 boxplot(mortgage_data$loan_amount_000s)
63
64 ggplot(mortgage_data, aes(y=loan_amount_000s, x=applicant_race_name_1,
65                           color=applicant_race_name_1)) + geom_boxplot()
66
67 ggplot(mortgage_data, aes(x=action_taken_name)) + geom_bar()
68
69 barplot <- ggplot(mortgage_data, aes(x=action_taken,
70                                     color=applicant_race_name_1)) + geom_bar()
71 barplot
72
```



