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:

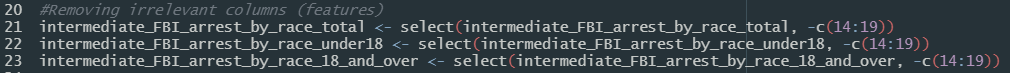
* Dropped irrelevant rows.
  + I have dropped rows using both the native “-“ operator and also the dplyr slice method.
  + I did this because these rows were unnecessary



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



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



* 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

Text

Description automatically generated

* 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.

Text

Description automatically generated

* 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

A close up of a screen

Description automatically generated

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.

Text

Description automatically generatedA close up of text on a black background

Description automatically generatedA close up of text on a black background

Description automatically generated

* Used dplyr techniques such as:
  + Select()
    - Select was used in a few different places primarily to drop columns that were unnecessary or those that became obsolete
  + 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.

Text

Description automatically generated

* 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 goem 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

Text

Description automatically generatedChart, box and whisker chart

Description automatically generatedChart

Description automatically generated

Chart, bar chart

Description automatically generated

Phase 05: Modeling and analysis

Modeling algorithms used:

***Linear regression***:

* + The intent while using this algorithm was to try and predict either the loan amount that would be requested or the applicant’s income, given the available information I have about the applicant
  + I made 4 different models.
  + The first one tried to model loan vs the applicant’s income and applicant’s race, where the amount of loan requested was the output/what the model would be predicting, and income and race were the inputs.
    - The p-value from this model was very good, less than 2.2e-16, which tells me that my variables had a very high degree of relatability.
    - However, the R-squared and Adjusted R-squared values were not that high, less than 0.3, which meant that my predictability power of my model was not that good.
  + I tried to increase the R-squared by adding more variables for my second and third model, i.e. co-applicant’s presence and co-applicant’s race.
    - Neither model crossed the 0.3 threshold for the R-squared, but I had the highest R-squared and Adjusted R-squared of all the models when using just income, race, and presence of co-applicant as my inputs.
  + For my fourth lm model, I used loan amount requested, applicant’s race and presence of co-applicant to try and predict the applicant’s income.
    - Again, I found that I had very low (good) p-values, but my R-squared did not cross 0.3, unfortunately.
  + I think the reason why my lm model does not have good predicting power is because, while there is significant correlation between all the variables (i.e. white people are more likely to request for larger loans than say black people, or people with higher incomes would ask for higher sums etc.), *most* people still just never ask for more than a few thousand dollars; they all cluster near the bottom.

Text

Description automatically generated

Text

Description automatically generatedText

Description automatically generatedText

Description automatically generatedText

Description automatically generated

***KNN Classification***:

* + I wanted to use the KNN classification algorithm to try and see if I could make a model which could predict whether a loan would be approved or denied, given the loan amount requested, the applicant’s income, race and presence of co-applicant
  + I took several steps to further clean my data to prepare for this process, including (but not limited to) normalizing my data so that they are all on the same scale (0 to 1).
  + I used roughly 90% of my data to train the model and 10% to test it
  + I made two models: one with a NN coefficient (k) of 357, which is roughly the square root of my total observations and one where k is 5 to see if one of them was better than the other
  + I used a confusion matrix to determine the accuracy of my models.
    - I found that my first model, where K=357, was more accurate than my other model
    - Model\_00 was accurate roughly 74.3% of the time, where as model\_01 was accurate only 68.6% of the time
    - I chose to stick with model\_00, but either way, they were both decently accurate.

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generatedText

Description automatically generated

***KMeans***:

* + I used this algorithm in hopes of organizing my data into similar clusters so that I may be able to extrapolate some useful insights from these clusters.
  + I made a few different models. The first one I built; I initially used the normalized dataset I used for my KNN classification.
    - While this model had was the best fit of the models I tried (had the highest between\_SS / total\_SS, 75.1 %), it did not allow for any useful interpretation of the data. At least none that I could find.
  + So then, I tried making a couple different models with non-normalized knn-data with 5 clusters (the number 5 was chosen arbitrarily).
    - These models were not as good of a fit as my first one (between\_SS / total\_SS = 66.4%), but in model\_02, I found that loan amounts clustered around 347,000 had the highest rate of approval, which I found was interesting.

Text

Description automatically generated

Graphical user interface, text

Description automatically generatedText

Description automatically generatedA close up of text on a black surface

Description automatically generated

Phase 06: Building the data product

***NOTE:*** In order to run my shiny app, make sure you manually run either the phase06.R script or the source() function in app.R. This will take some time due to the size of the data, so give it a minute.

My Rshiny code is fairly simple. It is organized into 4 files: ui.R, server.R, phase06.R and finally app.R to bring it all together.

The job of my ui file is to simply allocate to me room on the screen to insert input and to draw my outputs. The inserting input part of its job is achieved in the sidebarPanel via the two selectInput functions I have in it. The output(s) are displayed in the mainPanel via my uiOutput(‘tabs’) function. The inputs in this case are the number of nearest neighbors to consider for my knn models and the number of clusters for my k-means models.

The server.R has the primary function of reading the input and figuring out which model(s) to run and what to output. For the knn models, I am displaying the plot of the knn model, a confusion-matrix in table form and the analysis of the confusion matrix. For my k-means models, I am displaying the centers that the clusters have formed and the goodness of the model (as calculated by between\_ss/total\_ss) as these are the variables I’m most interested in.

The phase06.R is where most of the heavy lifting for this app is done; this is where the models are calculated. This was supposed to be done in the server.R file so as to make my app even more dynamic, however, R shiny was having issues performing these calculations due to the size of my data. Therefore, I had to make the calculations separately and beforehand, rather than dynamically via R shiny.

App.R is where I can execute all the necessary files with ease.

Graphical user interface, application

Description automatically generated