Jenan Taibah
210 final project writeup

Racial disparities in policing

I found my data set on the American Civil Liberties Union (ACLU) “Data for Justice” program. The dataset that piqued my interest the most was the Stanford opening policing project. This project is very interesting to me because as an international student, I have experienced more racism compared to other countries I have visited. Thus, I wanted to see if there is evidence of racial discrimination with police operations specifically because the security/ police of a country tends to represent the country itself.
The data set I analyzed can be found here (https://openpolicing.stanford.edu/data/) - Rhode Island under “Data Downloads”

To run this project correctly the data set should be placed within the src file and then the path to the dataset should be specified in the variable “file_path” within the main function in main.rs.

- Cargo build → Cargo run → code should run correctly
- Cargo test → to run the tests

The project analyzes the relationship between police arrests/ searches and race/ gender. To answer this question of racial disparities in policing I wrote several functions. My first function is placed within my “reading_csv.rs” module. The goal of this set of functions is to take the data set provided and then create an instance of it that is then used for all the operations. To do this I created a struct that will be used to create instances of the table. I then wrote the function “create_readable” which creates a table from the CSV file provided, it is under “Impl Data” creating instances of the specified struct. The function opens the specified file (specified in the main function), reads it, and filters through it by creating a vector for the column labels, and then initializes an empty hashmap which is then filled by iterating over the keys in the data set. The resulting hashmap only has the specified columns, because we don't need the entire data set for this analysis only the following: "subject_race", "subject_age", "subject_sex", "arrest_made", "search_conducted" The function also makes sure to return N/A if the value was not found. The second function in the impl block is print_readable. This function simply takes an instance of the struct and the function above and then prints out a readable, clean version of the data set which will be used later on. This function only prints out the first 20 rows as a visualization, the length could be adjusted as needed.

The next part of my code is the “basic_analysis.rs” module. This module does all the basic analysis (aka, counting and ratios) for the provided data set. The first function “arrests_and_searches_by_race” counts the arrests and searches made by race. It initializes counters and variables for each race and then increments as it iterates through the table. The following function “print_arrests_and_searches_by_race” then prints out the results in a table. The “race_ratio” compares the results from “arrests_and_searches_by_race” and creates ratios for the minorities/whites (which is our reference race) and then the “print_race_ratios” prints out the result. The last function does the same but in regards to gender. It reads the CSV file and increments the counter based on how many males and females are recorded in the data set, then creates a percentage and ratio of Males to females. This function also computes and prints it out at once.

Jenan Taibah
210 final project writeup

“Scatterplot.rs” is the module that takes in the read CSV file and then creates a scatter plot visualization. The “generate_scatter_plot” function generates two scatterplots, it starts off by assigning a drawing area and splitting it into two for the two scatterplots. Minority search rates on the y-axis and the reference race search rates on the x-axis (black v white) (Hispanic v white) Based on the read data the function gets the search rates for each race and then plots the search rates as red circles. The diagonal line in the middle is simply a reference line with the slope of y=3x, to compare the minority search rate with the reference race to help visualize whether it's overall more or less than the reference race (below or above the line) The scatter plot then gets saved and the path to it is printed out by the main function.
Creating a scatter plot was a very big struggle; it took me several hours to not only figure out the metrics but to figure out how to use the “plotters” crate.

I did include another visualization, this time for the gender. The function “generate_pie_chart” in the piechart_gender.rs module simply takes in the gender ratio created previously in my basic_analysis.rs module and uses a crate I found on crate io that creates a visual pie chart. For this code, I mostly based it on the code found on the rust guide for this crate. The function prints out a circle with females as pink circles and males as blue squares with the percentages to the side.

The final module in my project is the chaisquared.rs module. The main analysis for my function comes from this test. I struggled a lot before settling on this test. At first, I wanted to create a linear regression function, I knew it was going to be hard because my data points are categorical but I wrote a function for one hot encoding, but my code did not make any sense, linear regression is used on a continuous set of points not categorical, so even though the comment I got on my project proposal said I should do regressions or predictions I discarded it after two weeks of working on the idea with no progress. I also attempted to make a decision tree where the options would traverse down the tree and ultimately say which minority is most likely to be searched/ arrested. However, I also faced an issue with this that I could not solve. My “.fit” method was not working, the error mentioned that the parameter bounds were not correct even though I explicitly wrote the types as the desired bounds. After working on that and not getting it to work even with the help of the TA, he recommended I do a chi-square test. While I did not know what this test was, I researched it and thought it was perfect for my data set. The chi-squared test tells you whether or not there is a relationship between your variables and if there is how significant it is. This is perfect for my data set because all my races are variables and I already had a reference race selected, so I created the function “chi_squared_test” to compare the minorities to the reference race and then print out whether or not a significant relationship existed.

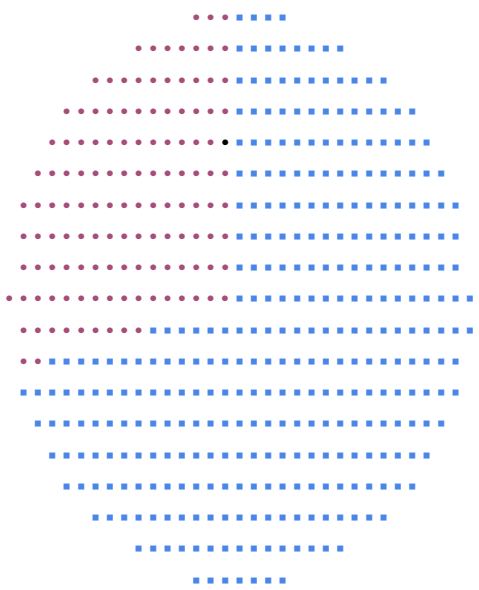
All my attempts for the linear regression and decision trees can be found in separate branches of my repository, however, that is not included in the submission of my project, rather I'm keeping it there for personal use, I would like to attempt to make them work on my own time.
(the comments in the code itself go into more detail about the steps and how each function works, this is more of an overview)

Jenan Taibah
210 final project writeup

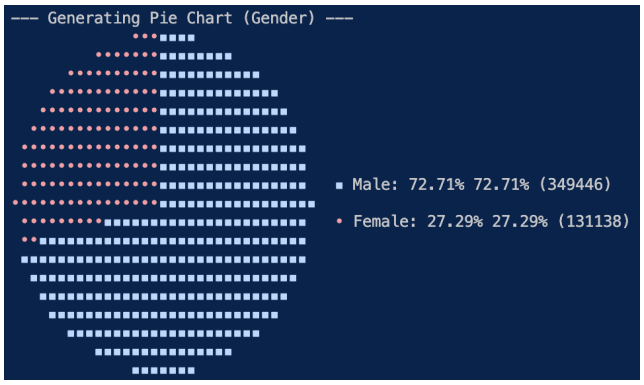
The output of my function should look like the following:

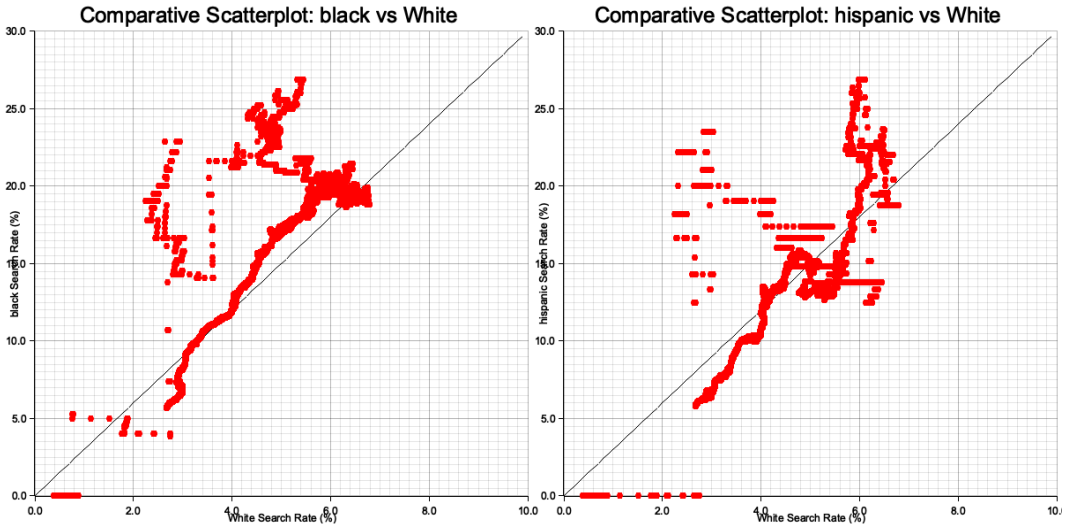
```
--- Printing Original Dataset Sample ---
 subject_race  |  arrest_made  |  subject_age  |search_conducted|  subject_sex  |
-------------------------------------------------------------------------------
     white     |     FALSE     |      N/A      |     FALSE     |     male      |
     white     |     FALSE     |      N/A      |     FALSE     |     male      |
     white     |     FALSE     |      N/A      |     FALSE     |    female     |
     white     |     FALSE     |      N/A      |     FALSE     |     male      |
     white     |     FALSE     |      N/A      |     FALSE     |    female     |
     white     |     FALSE     |      N/A      |     FALSE     |     male      |
     white     |     FALSE     |      N/A      |     FALSE     |     male      |
     white     |     FALSE     |      N/A      |     FALSE     |    female     |
   hispanic    |     FALSE     |      N/A      |     FALSE     |     male      |
     white     |     FALSE     |      N/A      |     FALSE     |     male      |
     white     |     FALSE     |      N/A      |     FALSE     |    female     |
     white     |     FALSE     |      N/A      |     FALSE     |     male      |
     white     |     FALSE     |      N/A      |     FALSE     |     male      |
     white     |     FALSE     |      N/A      |     FALSE     |     male      |
     white     |     FALSE     |      N/A      |     FALSE     |     male      |
     white     |     FALSE     |      N/A      |     FALSE     |    female     |
     white     |     FALSE     |      N/A      |     FALSE     |     male      |
     white     |     FALSE     |      N/A      |     FALSE     |    female     |
     white     |     FALSE     |      N/A      |     FALSE     |    female     |
     white     |     FALSE     |      N/A      |     FALSE     |    female     |

--- Basic Analysis on Original Dataset ---

Race                  Arrests         Searches
-------------------------------------------------
hispanic                  | 3156       | 3189
other                     | 14         | 15
asian/pacific islander    | 258        | 280
na                        | 0          | 0
white                     | 9237       | 9968
black                     | 3938       | 4310

Race                  Arrest Ratio    Search Ratio
-------------------------------------------------
hispanic                  | 0.34               | 0.32
other                     | 0.00               | 0.00
asian/pacific islander    | 0.03               | 0.03
na                        | 0.00               | 0.00
white                     | 1.00               | 1.00
black                     | 0.43               | 0.43

Males: 349446
Females: 131138
Male to Female Ratio: 2.66

--- Generating Scatter Plot (Race) ---
Scatter plot can be found in: scatterplot.png

--- Generating Pie Chart (Gender) —
```

Jenan Taibah
210 final project writeup




```
Pie chart successfully generated.

--- Performing Chi-Squared Test ---
Chi-Squared Statistic: 457805.40
Degrees of Freedom: 10
P-Value: 0
The relationship is significant!!
Chi-squared test completed successfully.
```





Jenan Taibah
210 final project writeup

The results of my analysis of the Stanford open policing project show that there is in fact a difference in how police officers treat different races. In general, officers tend to stop black drivers at higher rates than white drivers and stop Hispanic drivers at similar or lower rates than white drivers. This can clearly be seen from my scatterplot where most of the red circles representing black drivers are above the reference line. On the other hand, the red circles representing Hispanic drivers are under or along the reference line. In terms of gender, it can also be said that officers tend to stop male drivers more than female drivers as the ratio of male to female is 2.6. Overall I would say this analysis has been very informative but it should be taken into consideration that the data set is biased and contains mostly white drivers which could lead to a skewed analysis. The data set is also only based on the state of Rhode island, an analysis of other states could differ. However, the overall analysis for all states agrees with the analysis I have conducted in my project.