Divakar Borra

Mac Owusu

CICS 397a

Final Project Write-Up

For this project, the work was evenly distributed between both of us. Divakar worked on clustering and predictive modeling as well as two visualizations of the data, while Mac worked on utilizing SQL databases to investigate the data as well as two of the visualizations for the project. The dataset we chose was the CFPB Financial Well-Being Survey dataset which goes over an individual's overall financial well-being. We chose to look at poverty status and education because schooling after K-12 is looked at in society as “the right thing to do” but it’s becoming a controversial topic. Many people like to claim that college is a scam so we will use variables to investigate if you’re better off with more education or not. Along with that we would like to see the correlation between the poverty rate level and satisfaction of life, the saying “money doesn’t buy happiness” is a popular saying but does money buy happiness?

From the dataset the first variable we wanted to look at was poverty status and how it influenced certain factors of an individual's life. We first decided to investigate the Financial Well-Being Survey dataset with DB browser by using SQL queries to gather certain information. We wanted to observe the correlation between one’s poverty status and how they rated their satisfaction with life. The poverty status was labeled fpl, and the numbers were represented as 1 = < 100% , 2 = 100-199% , and 3 = 200% , and for life satisfaction rate the scale was -4, -1, 1, 2, 3, 4, 5, 6, 7. The number -4 was the numerical value for “response not written to database” and -1 was the numerical value for “refusal”, after becoming aware of this we modified the life satisfaction rate data so that the scale was from 1-7; strongly disagree to strongly agree to make it more comprehensible. Another modification we made for this query was limiting the data to 400 responses which made it easier to comprehend. It was shown that there was a

correlation between the two, the satisfaction rate steadily increased the higher the poverty score increased. This determined that individuals who occupy the higher or “less poor” portion of the poverty level rate their satisfaction in life higher. The graph labeled “Poverty Level vs. Satisfaction with Life”, shows a visual representation of the data that was queried in the form of a barplot, with the y-axis representing Satisfaction with life and the x-axis representing the 3 poverty level rates, each bar has its own color distinguishing it from the other. In the plot you see the steady increase in satisfaction as the numbers in the poverty level become better.

In the second query of the SQL portion of the assignment, we looked at the relationship between education, race, and poverty status. Education was labeled PPEDUC and the numerical values were 1 = Less than high school, 2 = HighSchool/GED, 3 = Some College, 4 = Bachelors degree, and 5 = Graduate Professional Degree. Race was labeled PPETHM and the values were 1 = White, 2 = Black, 3 = Other, and 4 = Hispanic, and lastly the poverty status had the same numerical values from the first query the data was again limited to 400 responses for comprehensibility. What was observed was that individuals who had a higher education generally occupied the low poverty status area much less, while people with lower education were found more on the lower side of the poverty level. Analyzing and manipulating the data a bit more we also concluded that specifically black people who get a higher level of education are typically less poor. With at least some college or higher they generally stay in the second and third category of poverty status with a few outliers. From the observations we made from the SQL queries, we decided to make a visualization of this query. One of the visualizations we formed was a scatter plot labeled, “Poverty Status based on Education & Race” and it showed the relationship between these three variables. The x-axis represents education, the y-axis represents race, and lastly the plots on the graph are color coded by poverty status which represents the 3 levels within it, each dot has a color corresponding to one of the three levels that the given participant belongs to.

Next, through querying we decided to look into the relationship between an individual's financial well-being score and their understanding of credit card minimum payments based on their answer to a question about credit card minimums. In the code we selected FWBscore (Financial Well-Being Score) and KH7 (Question Answered Correctly) from the database. We then specifically grouped the data to answer the question of whether individuals with a financial well-being score higher than the mean score of 54 scored higher or lower on the credit card minimum survey question. This would conclude if they had a solid understanding of credit card minimum payments. There were two options, 0 which was no and 1 which was yes. In the first set we isolated individuals with financial well-being scores greater than 54 and it resulted in 1,758 correct answers and 1,735 incorrect answers. Next we grouped individuals with financial well-being scores lower than 54 and that resulted in 973 correct answers 1,741 incorrect answers. With that being said the data showed that although it is not by a huge percentage, people with a score higher than average on the financial well-being scale achieve correct answers more frequently than people with lower than average on the scale. As well as those with a score lower than average choose the incorrect answer more frequently. It can be concluded that based on the data from CFPB Financial Well-Being Survey individuals with a score higher than average on the financial well-being scale understand credit card minimum payments concept more than individuals with lower than average scores. An important visualization we made that was similar to the query just mentioned was the density curve with the people that got the question correct regarding credit card minimum payments and poverty rate. The individuals that are considered rich or have the highest poverty score were usually the people that got the question correct regarding credit card minimum payments. The y axis measures the density curve of the graph while the x axis is whether the individual got the question correct. The poverty rate is distinguished by the different color lines in the graph.

In the last SQL query we chose to look into the correlation between life expectancy and generation based on a question in which participants, who were one of four generations, were

asked. The question was “How likely do you believe it is that you will live beyond age 75?” and the responses ranged from -2 to 100, however -2 and -1 were excluded due to them being non-numerical results, so the modified scale was 0-100. A visual was made to help explain the results. We utilized a barplot to further emphasize these findings. Having generation on the x axis and life expectancy on the y axis while faceted by poverty rate, we wanted to see how poverty rate varied per generation. It was seen that the Boomer Generation had the overall highest life expectancy while also having a high score on poverty rate hinting at a correlation between the life expectancy and high score on poverty rate.

We also utilized clustering on poverty rate, ethnicity, and education. This was organized in 3 clusters in which one cluster was larger or more unevenly distributed than the other 2 clusters. The larger cluster had poverty rate 3 associated with higher scores of education which supports the idea that higher education allows less poverty for an individual. We calculated the clustering by average. We later then used a classification method on poverty rate to better understand the variable in which we used a 80% training and 20% testing split.

A majority of the code written was from libraries that needed to be imported beforehand. Some examples of the libraries used consisted of pandas, numpy, matplotlib, and seaborn. The dataset was read in using pandas and assigned with a variable named df_all, we would then isolate an x-axis and y-axis value from the dataset based on the plot we were trying to create. Different code was used for different plots such as scatterplots, barplots, and a density curve plot. One particular challenge faced during this assignment was creating a scatter plot with three variables. At first we were unaware that a variable could be portrayed by the color, but by using resources such as geeksforgeeks, youtube, and the lecture slides we were able to overcome this issue and successfully make the scatter plot. Apart from that there weren’t many other prominent obstacles that we faced that we didn’t overcome.

It can be concluded that overall life is better when an individual is wealthier, poverty status can have a negative impact on certain life factors of an individual who occupies the lower

level of the poverty status scale. However, there are certain factors that may determine what poverty level you will belong to as seen with the scatter plot displaying poverty based on education and race. This also provides some answer to the topic of how school is viewed in society, schooling after highschool is looked at as “the right thing to do”. Based on the information, it indicates that more individuals end up in better living conditions in terms of poverty when they have a higher education. Along with that, participants who had a higher education also had a higher financial well-being score, showing that higher education promotes multiple factors in an individual's life. It seems as if school is “the right thing to do” but maybe not in all instances, this taught us to appreciate school even though we can’t wait to get out of here!