Exploratory Data Analysis

Ashley Wang, Ellen Cheng

Dec 2023

The United States has the largest income inequality in the rich world (Council on Foreign Relations 2022). In fact, the top 10% have about 13.5 times more income than the bottom 10 percent. Between 2020 and 2021, income inequality increased 5%, most likely due to effects from the pandemic (US Census 2022). In a country that has built its foundation on "The American Dream," the country's inhabitants have long seen such a phenomenon as a dream more so than reality. For our project, we were very interested in seeing the causes and effects of income inequality. What drives disparities in income, and what are the results of such imbalances? To investigate this question, we compiled different sources of data and explored different research avenues. We first explored relationships between the poverty line, race, and region with income. After our initial round of exploration, we also decided to look into the relationship between education and income, which produced a very strong result and has now become our main topic of interest.

First, we started off with finding databases. As seen below, we compiled data with information on housing, income, demographics, and the poverty line from Social Explorer. We used data from Social Explorer because it had many variables that we could filter by and also had very relevant and recent data. Social Explorer contains data from the US Census Bureau and other prominent databases. Each variable was sorted by state (for example, we had income by state, sex by state, race by state, etc.), which made merging all the datasets quite easy. We put all the variables side by side. The variables we used included housing, gender, sex, percentage of

people below the poverty line, households in subsidized housing, income, monthly expenses, and region. We were interested to see how the variables of housing, gender, sex, and monthly expenses affect income. We were also interested to see how income affects the percentage of people below the poverty line and households in subsidized housing. We weren't sure whether we wanted to use income as an independent or dependent variable, as we didn't know which scenario would paint a better picture about income inequality, so we kept our exploration broad. We were also interested to see the differences in these relationships by region as well (does the West have a higher percentage of people below the property line vs. the East and etcetera?)

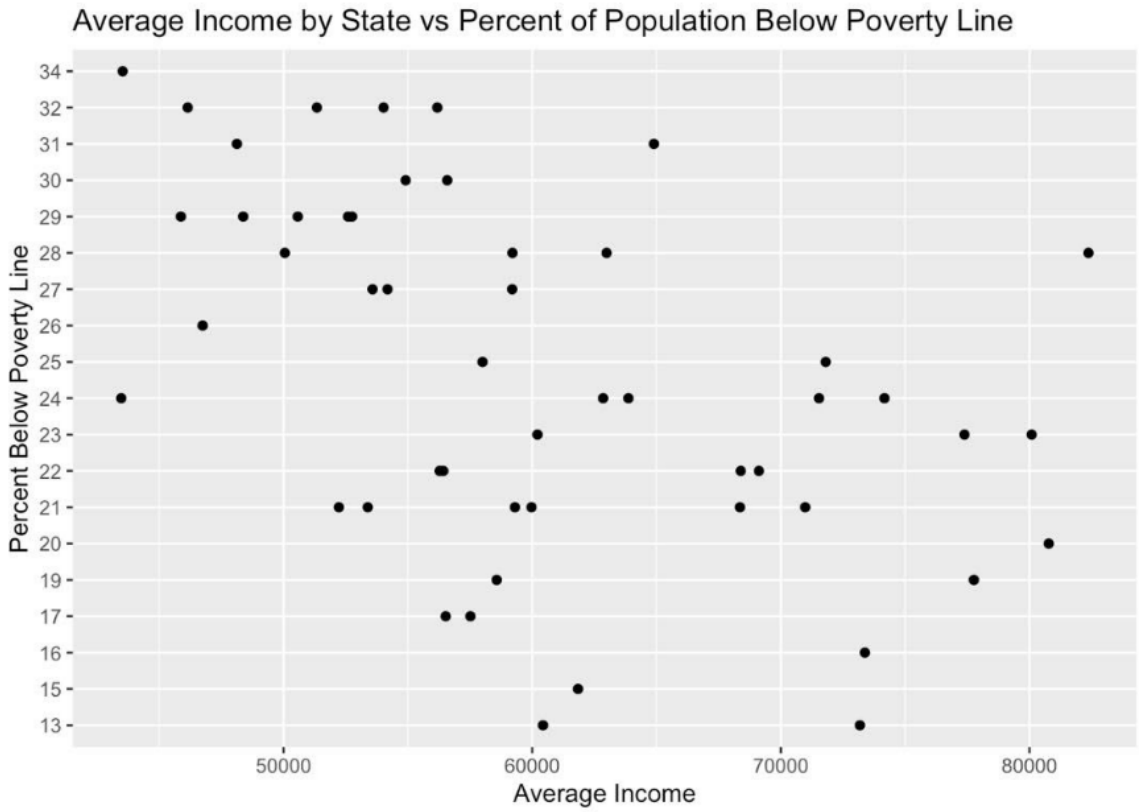
We cleaned the data by renaming the columns. For example, we renamed the number of people in subsidized housing into "number_indiv_subsidized" and the income for individuals in subsidized housing to "avg_indiv_income_subsidized". After renaming the columns to be more descriptive and simple, we were more equipped to explore our data. We decided to explore our data through investigating three different relationships: income level vs. percent of individuals below the poverty line and region vs. average income. These three relationships help us explore the poverty line, demographics, and location, which encompass significant drivers of the United States' appalling income inequality.

```
colnames(sub_housing)[colnames(sub_housing) == "Percentage.of.the.Population.
Below.Poverty.Level.in.the.Census.Tract.Where.HUD.Assisted.Families.Reside..
Summary.of.all.HUD.Programs."] <- "percent_poverty"
sub_housing <- sub_housing[,-1]
sub_housing <- sub_housing[-1,]
colnames(sub_housing)[
colnames(sub_housing) == "Name.of.Area"] <- "name" #we will join two datasets by "name"

sub_housing <- sub_housing[,-2:-11] # delete unnecessary columns

income_and_sub_housing <- full_join(income_by_race_2017, sub_housing, by = "name")
```

```
colnames(mydata)[colnames(mydata) == "People.in.Subsidized.Housing..Summary.of.all.HUD.Programs."
] <- "number_indiv_subsidized"
colnames(mydata)[colnames(mydata) == "Average.Individual.Income.Per.Year..All.Individuals.in.
Subsidized.Housing.Units...Summary.of.all.HUD.Programs."] <- "avg_indiv_income_subsidized"
colnames(mydata)[colnames(mydata) == "Aggregate.Household.Income..All.Households.in.Subsidized.
Housing.Units...Summary.of.all.HUD.Programs."] <- "household_income_subsidized"
colnames(mydata)[colnames(mydata) == "Average.Family.Expenditure.per.month..Payment.toward.
Rent.and.Utilities...Summary.of.all.HUD.Programs."] <- "family_expenditure"
colnames(mydata)[colnames(mydata) == "Average.HUD.Expenditure.per.month..Federal.Spending...
Summary.of.all.HUD.Programs."] <- "federal_spending"
colnames(mydata)[colnames(mydata) == "Households.in.Subsidized.Housing..Summary.of.all.HUD.
Programs."] <- "number_households_subsidized"
colnames(mydata)[colnames(mydata) == "Households.in.Subsidized.Housing..Households.Headed.by.
Female..Summary.of.all.HUD.Programs."] <- "household_female_head"
colnames(mydata)[colnames(mydata) == "Households.in.Subsidized.Housing..Households.Headed.by.
a.Female.With.Children..Summary.of.all.HUD.Programs."] <- "household_female_head_with_children"
```

First, we were very interested to see the relationship between income level and percentage of individuals below the poverty line. Since all states use the same metric for the poverty line, we were able to analyze the relationship between income and the percentage of people under the poverty line. The United States poverty line is defined as individuals having an annual income of $14,580, households of 2 people having a combined income of $19,720, and a household of 3 having a combined income of $24,860 (Healthcare.gov 2023). We explored this relationship by creating a scatterplot with the two variables by making a ggplot:

```
mydata %>%
ggplot(aes(x=Overall, y=percent_poverty))+
  geom_point() +
  xlab("Average Income") +
  ylab("Percent Below Poverty Line") +
  ggtitle("Average Income by State vs Percent of Population Below Poverty Line")
```

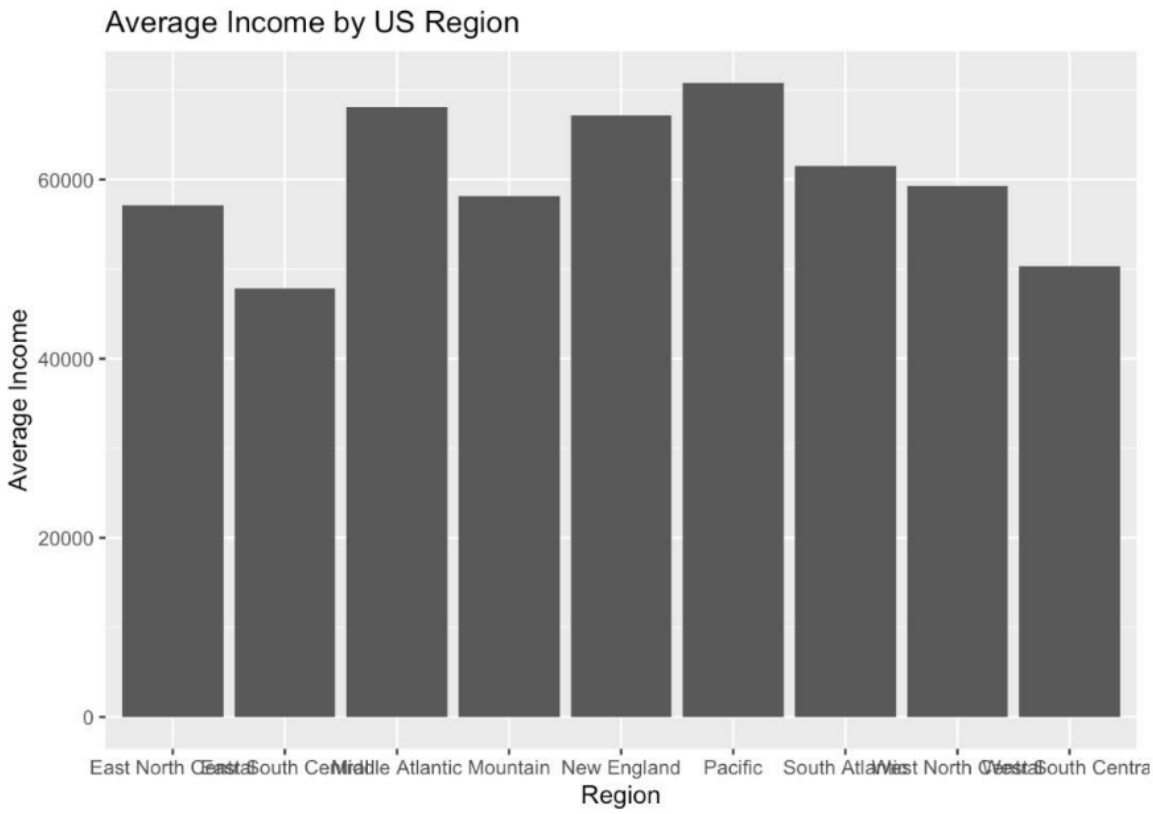


Since it is still an exploratory page, we didn't run a regression. However, the relationship seems to exhibit a downward sloping trend (negative correlation). This makes sense, as with an increase in income, one would expect there to be less individuals under the poverty line, as with a higher income, there would be a greater difference between you and the poverty line. This could be a very interesting relationship to continue exploring. If we were to continue exploring this relationship, we would like to learn more about what factors contribute to individuals having an income level being below the poverty level (demographic, region, occupation, etc.).

Second, we explored average income by region. We were interested to see whether certain areas of the country had higher or lower average incomes as compared to other areas. We had data on the average incomes of different regions of the United States, which were categorized by East South Central, Pacific, Mountain, West South Central, New England, South

Atlantic, East North Central, West North Central, and Middle Atlantic. As seen below, we renamed the columns for the different regions to be more descriptive, grouped the data by region, printed out the average for each region with the summarise function, and finally graphed a histogram with ggplot. As this was not a relationship between two variables, we visualized the data through a histogram:

```
mydata$region[mydata$region == "ESC"] <- "East South Central"
mydata$region[mydata$region == "P"] <- "Pacific"
mydata$region[mydata$region == "M"] <- "Mountain"
mydata$region[mydata$region == "WSC"] <- "West South Central"
mydata$region[mydata$region == "NE"] <- "New England"
mydata$region[mydata$region == "SA"] <- "South Atlantic"
mydata$region[mydata$region == "ENC"] <- "East North Central"
mydata$region[mydata$region == "WNC"] <- "West North Central"
mydata$region[mydata$region == "MA"] <- "Middle Atlantic"

avg_income_by_region <- mydata %>%
  group_by(region) %>%
  summarise(avg_income = mean(Overall))

avg_income_by_region %>%
  ggplot(aes(x= region, y = avg_income)) +
  geom_col() +
  labs(title = "Average Income by US Region",
       x = "Region",
       y = "Average Income")
```



From this visualization, it seems that the Pacific has the highest average income while East South Central has the lowest income. While the average income in the Pacific is around $80,000 per person, the average income in the East South Central region is around $50,000, which is about a $30,000 difference. This was very interesting data to examine as the Pacific region includes California, Oregon, Washington, and Alaska, while the East South Central region includes Alabama, Kentucky, Mississippi, and Tennessee. It would be quite interesting to explore how specific differences in the Pacific region and East South region - whether it be in education, types of occupations, access to banking, and more - drive this drastic difference in average income.

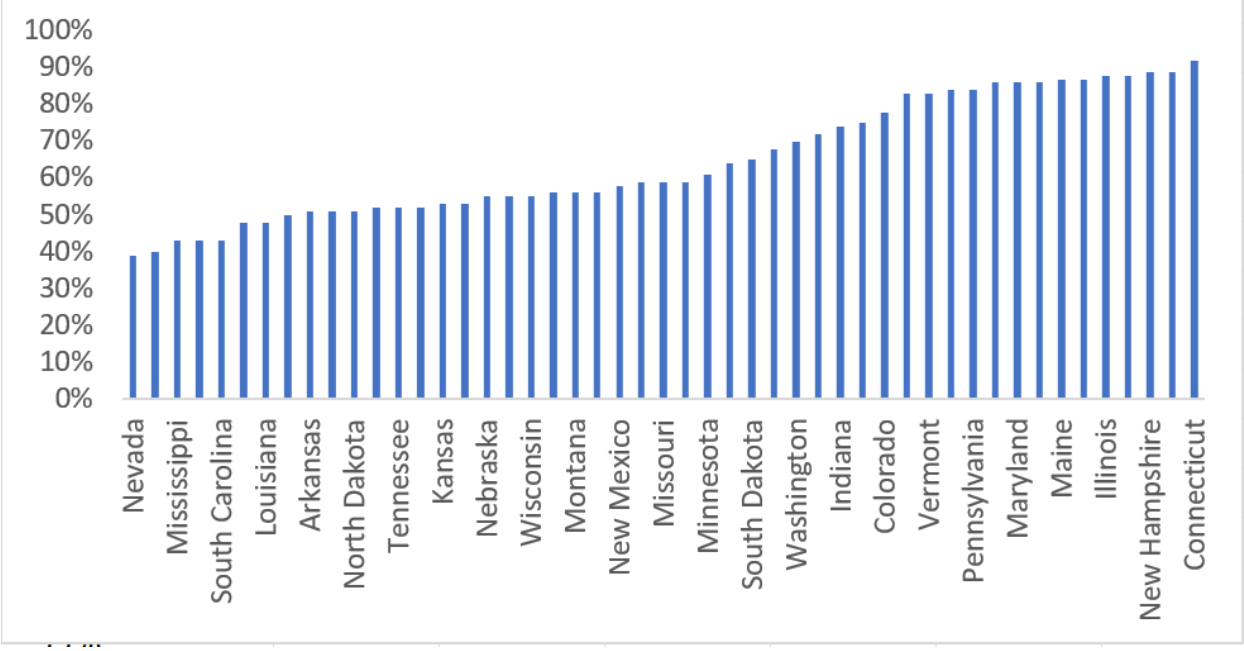
There are a couple things to note in our initial exploration. First, there were missing values for the average income for Black people in Wyoming. We were also unable to compile

multiple variables - for instance, something interesting to look at would be how sex and race by state impacts income, but with data on these separate variables and not in a combined form, we were unable to approach this issue. We also realize that some of our visualizations may not be the best and clearest way to analyze the data.

After our initial exploration, we also became interested in exploring the relationship between education and income inequality. We suspected that education by state could also be correlated with previous relationships we had explored (ex. Region and income). Indeed, the Economic Policy Institute found that between 1998 and 2010, large performance gaps existed between children in the lowest and highest socioeconomic groups, and they have remained largely unchanged over time (EPI 2017). The Brookings Institute also reports that "high school dropout rates are higher in cities and states with greater income inequality" (Brookings Institution 2016). Education is critical as studies show that education is highly correlated with social mobility (NYU 2021). We were able to find a dataset that explores this relationship. The dataset is from Wisevoter and lists the percentage of students by state meeting the ACT Math Benchmark and the ACT English Benchmark.

We first started by taking the average between the percentage of students who met the ACT Math and ACT English benchmark. By averaging these two percentages, we obtained a picture of a state's overall ACT performance, under the assumption that the average difference between each student's math and English performance is not drastic. Then, we plotted a histogram sorting the states from least % of students meeting the benchmark and most % of students meeting the benchmark:

```
mydata$english <- mydata$english / 100

mydata <- mydata %>%
  mutate(math_english_avg = (math + english
) / 2)
```

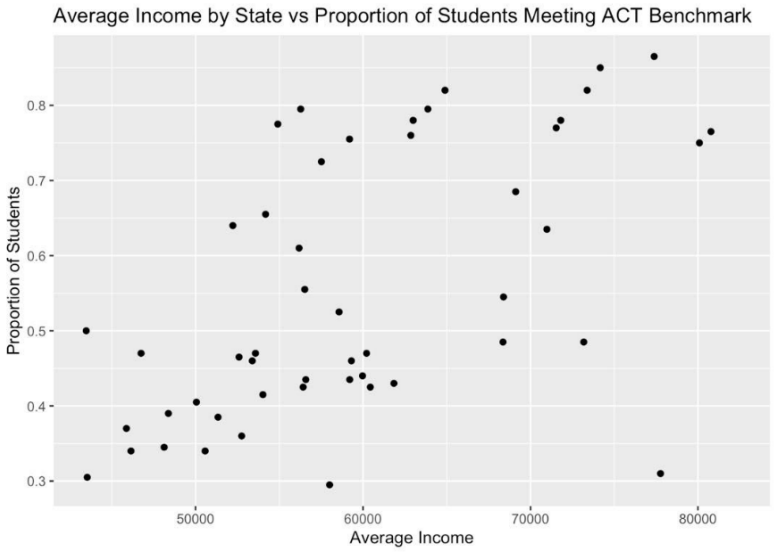


Interestly enough, many of the states residing in the East South region had the lowest percentage of students meeting the benchmark. This region also had the lowest average income, meaning that there is likely a relationship between the low average income level and the low percentage of students meeting the ACT benchmark. We would be very interested to see what kind of policies have resulted in the East South region lagging behind other regions in terms of income and education.

We explored the relationship between income and education further by producing a scatterplot in ggplot between average income by state and the percentage of students meeting the ACT math and English standards by the corresponding state:

```
mydata %>%
ggplot(aes(x=Overall, y=math_english_avg))+
  geom_point() +
  xlab("Average Income") +
  ylab("Proportion of Students") +
  ggtitle("Average Income by State vs Propo
rtion of Students Meeting ACT Benchmark")
```

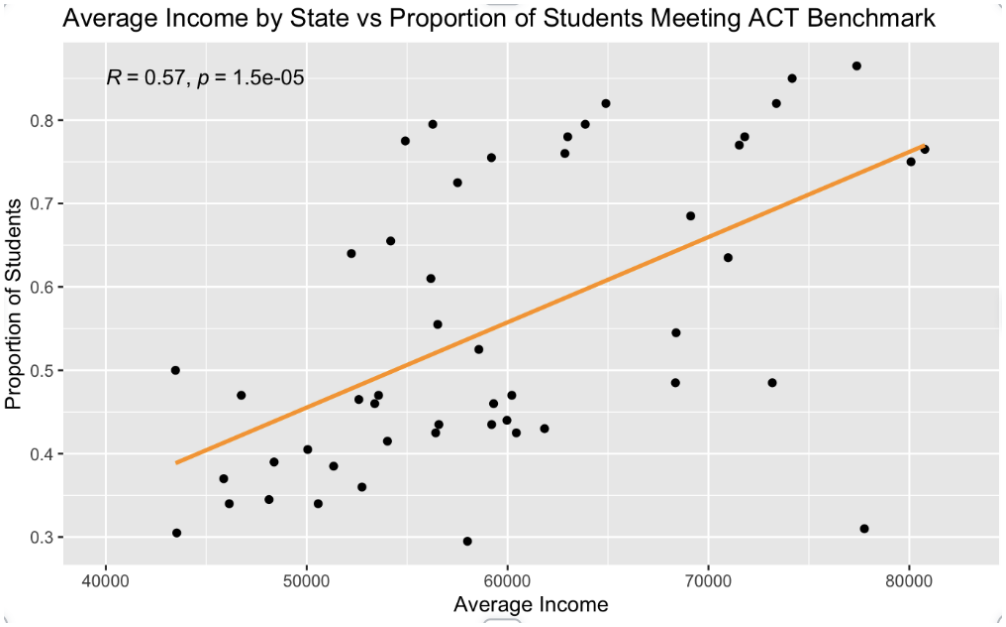
```
## Warning: Removed 1 rows containing missi
ng values (`geom_point()`).
```



Similarly, this scatterplot also shows a positive relationship between average income and percentage of students meeting the ACT benchmark. As the average income for the state increases, the proportion of students meeting the ACT benchmark also increases, showing that there is likely a relationship between the two. This means that as education level increases, average income increases by well, since social mobility is impacted by income.

Although we see a positive association between these two variables, we wanted to make sure that this association was significant, and we wanted to obtain the specific strength of that correlation. Therefore, we ran Pearson's *r* regression, which yielded the *r*-value of the association, as well as the *p*-value used to determine statistical significance.

```r
mydata %>%
ggplot(aes(x=Overall, y=math_english_avg))+
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "orange") +
  stat_cor(method = "pearson", label.x = 40000, label.y = 0.85) +
  xlab("Average Income") +
  ylab("Proportion of Students") +
  ggtitle("Average Income by State vs Proportion of Students Meeting ACT Benchmark")
```



We are very interested in exploring the education standards and prospects in states with lower income to help us understand America's glaring issue with economic inequality. We are also interested in tying all of our analyses together. As for next steps, we will explore how regions differ in their educational standards and whether the lowest socioeconomic groups have adequate access to education.

Works Cited

"Federal Poverty Level (FPL) - Glossary." *Federal Poverty Level (FPL) - Glossary | HealthCare.Gov*, www.healthcare.gov/glossary/federal-poverty-level-fpl. Accessed 3 Oct. 2023.

Garcia, Emma; Weiss, Elaine, September 27. "Education Inequalities at the School Starting Gate: Gaps, Trends, and Strategies to Address Them." *Economic Policy Institute*, www.epi.org/publication/education-inequalities-at-the-school-starting-gate/. Accessed 16 Oct. 2023.

Insight, 19 Mar. 2021, wp.nyu.edu/insight/2021/03/19/the-importance-of-social-mobility/#:~:text=Only%209%25%20of%20U.S.%20citizens,to%20social%20mobility%20is%20education.

Kearney, Melissa S; Levine, Philip, et al. "Inequality Undermines the Value of Education for the Poor." *Brookings*, 27 July 2020, www.brookings.edu/articles/inequality-undermines-the-value-of-education-for-the-poor/.

Kollar, Jessica Semega and Melissa. "Increase in Income Inequality Driven by Real Declines in Income at the Bottom." *Census.Gov*, 13 Sept. 2022, www.census.gov/library/stories/2022/09/income-inequality-increased.html#:~:text=The%20ratio%20of%20the%2090th,a%204.9%25%20increase%20from%202020.

"The U.S. Inequality Debate." *Council on Foreign Relations*, Council on Foreign Relations, www.cfr.org/backgrounder/us-inequality-debate. Accessed 3 Oct. 2023.

"The U.S. Inequality Debate." *Council on Foreign Relations*, Council on Foreign Relations, www.cfr.org/backgrounder/us-inequality-debate. Accessed 3 Oct. 2023.