Group Project #1

Group 7 The group/team is comprised of:

- Richard Hoehn (**richardhoehn**)
- Hector Rogel (**Hrogel9007**)
- Isaiah Osborne(**IMOsbo**)
- Karson Woods (**Kwoods132**)

Overview of the Dataset

We chose the adult census income dataset to evaluate for potential biases [1]. There are 15 total columns and a total of 32,561 rows. During an initial look from Kaggle, we can already see there's some apparent biases in the dataset. For example, for race, it's 85% white and 15% black or other, which is quite a large disparity. As well, men comprise 67% of the gender category, whereas women only comprise 33% of it.

Biases Found

Gender Biases

When looking at our dataset, we found evidence of both sampling bias and also wider-reaching societal biases. When examining the proportion of men in the dataset, we found that the dataset was composed of almost 66% men. After looking at the original census data, we found that the true population proportion for men in 1994 was 48.81%, substantially different from our sample proportion [2]. A two sample Z-test was conducted to determine if the difference between the population and the sample was statistically significant; the Z test found the results were statistically different at a 0.01 confidence level.

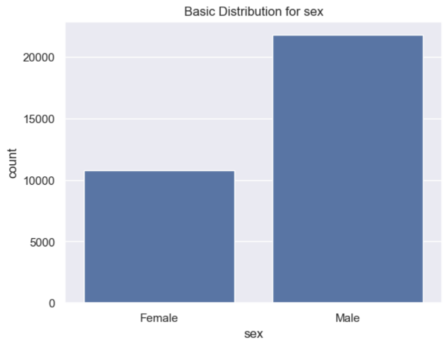

Figure 1: The Distribution of Men and Women across the Dataset

However, this overrepresentation of men is not as stark across races. While men are generally overrepresented, white men are substantially more overrepresented than their counterparts from other races.

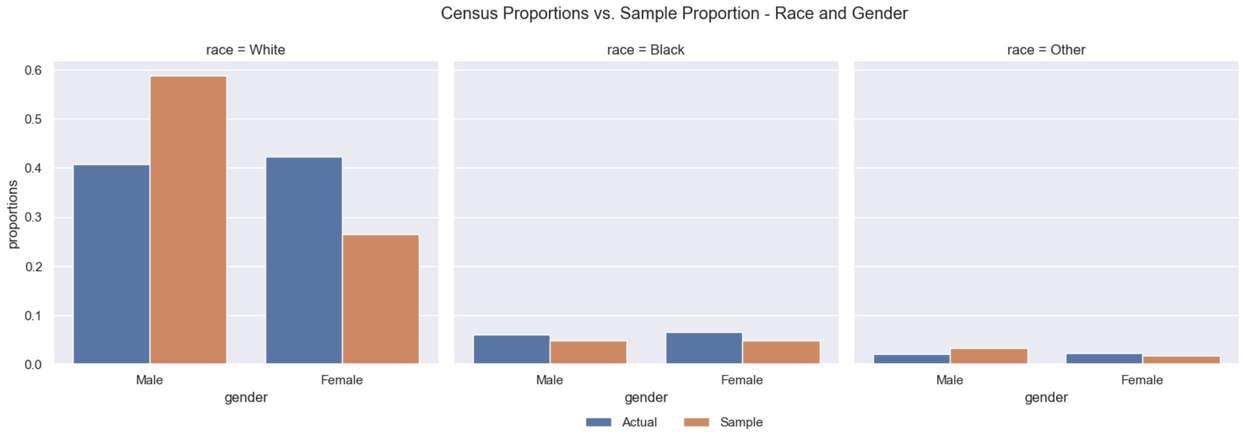

Figure 2: Census Gender Proportions vs. Sample Gender Proportions

As shown in the chart above, white men and men from other races have higher than expected proportions in the dataset, while black men and all categories of women have lower than expected proportions in the dataset. Whatever sampling method was used for collecting this dataset is definitely biased toward white men.

Pay Biases

In this dataset, pay is a categorical variable, which makes analysis difficult. However, we can look at the relative distribution of >50K and <50K across both genders, which should give us a reasonable idea of the salary ranges for men and women.

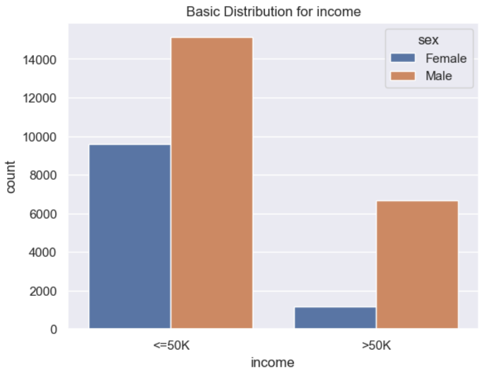

Figure 3: Distribution of >50K Salaries Across Men and Women

As shown in the chart above, about 30.57% of men are making more than >50K, while only 10.95% of women are making more than >50K. Using a z test, this difference was found to be significant at the 0.01 significance level.

After finding this result, we tried to discover whether this was associated with education; maybe the men in the sample happened more highly educated.

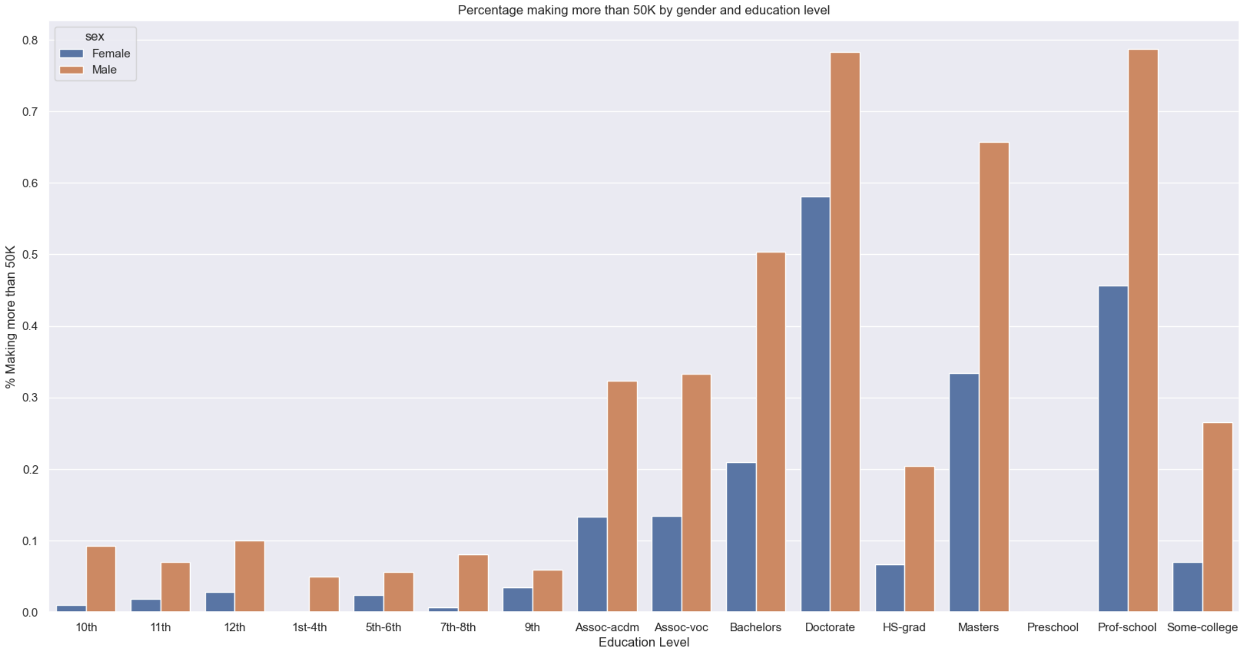

Figure 4: Distribution of High Salaries across Education Levels

Unfortunately, this pattern does not hold out. Even when looking at people with doctorate degrees, men are still making substantially more money than women; in fact, doctorate degrees are the only education category where more than half of the women surveyed make more than $50K. This pattern also holds across races and age ranges.

Racial Biases

In examining the dataset, racial biases also emerged. A noticeable bias was seen in the disproportionate representation of racial groups with 85% of individuals identified as white and only 15% from other racial backgrounds. This includes Black individuals. This disparity is a clear example of sampling bias. To further investigate, we analyzed the relationship between race and key variables, income and hours per week, and our findings suggested racial disparities in salary distributions.

The Chi-Square revealed a significant relationship between race and income categories (<=50K vs. >50K). White and Asian-Pac-Islander individuals were more likely to earn salaries above $50K, suggesting greater access to higher-paying jobs compared to Black and other racial groups. Additionally, an ANOVA test was conducted to see if certain racial groups have significantly different work hours. Seeing that race had an effect on number of hours worked per week, a Tukey's HSD test was conducted. The test revealed that the Black group, on average, works significantly more hours than the White and Asian-Pac-Islander groups. This finding could indicate that Black workers may be disproportionately employed in lower-wage or hourly positions that require longer hours, further contributing to income disparities.

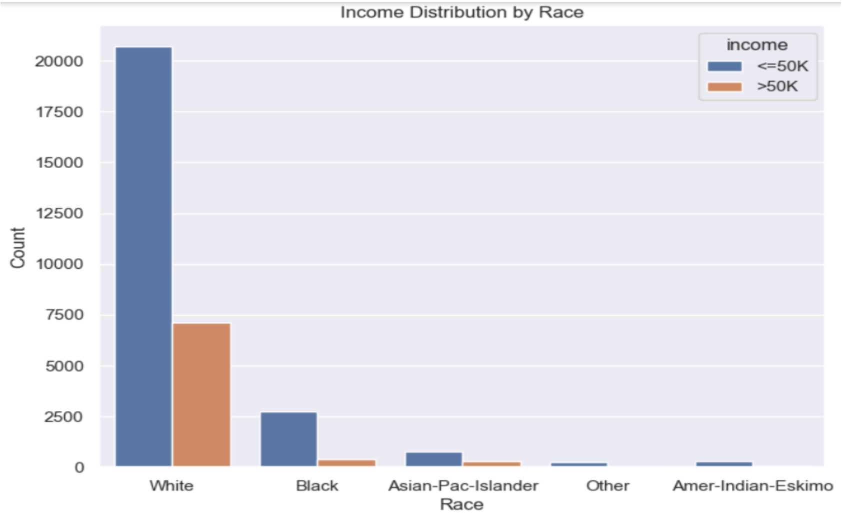

*Figure 5: Income Distribution by Race*Marital Status

This bar chart visualizes income distribution across different marital statuses in the USCI dataset, highlighting potential biases. The chart shows that individuals who are never married, divorced, or separated predominantly fall into the ≤50K income bracket, suggesting economic disadvantages for these groups.

Alternatively, married-civ-spouse individuals show a higher proportion in the >50K income bracket, indicating that marriage, particularly in a traditional civil setting, is associated with higher income levels.

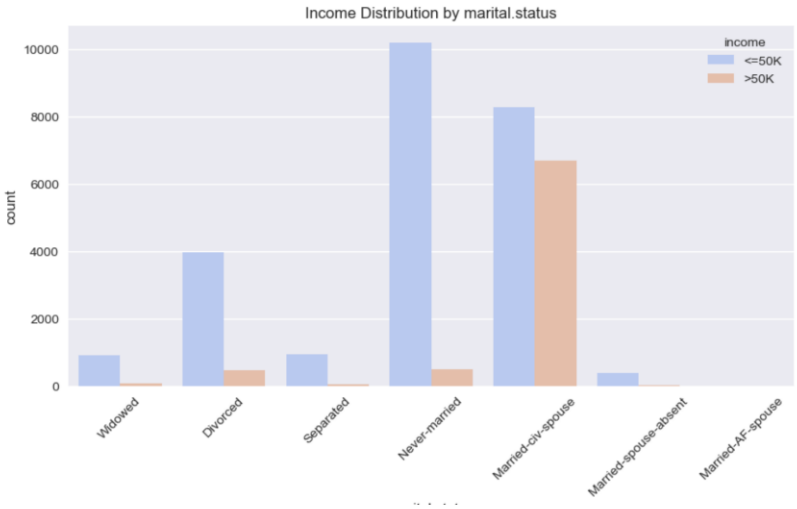

Figure 6: Income Distribution Divided Across Martial Status

Methods Used & Process

We utilized Python and the Seaborn library to analyze income distribution across different columns such as: age, work class, education, marital.status, occupation, relationship, race, and sex based on the USCI dataset. The data was first preprocessed using pandas, ensuring categorical variables like marital status and income were correctly formatted. We then employed Seaborn's countplot function to create a grouped bar chart, visually differentiating income categories (<=50K and >50K).

Results

The analysis of income distribution in the UCI dataset reveals distinct patterns based on demographic and economic factors. A significant portion of individuals fall within the ≤50K income category, suggesting that lower-income levels are more common in the dataset. However, a notable proportion of higher-income individuals (>50K) are concentrated in specific groups, potentially influenced by factors such as education level, marital status, and ethnicity.

Implications

There are some worrying biases we found after analyzing the adult census income dataset. Most notably were the gender and racial biases that we found.

Regarding the gender and racial biases we found, most notable was the gap in sampling of men and women. Considering the sample contained 67% men, whereas men only accounted for 48.81% of the population in 1994 [2]. Digging further into this, white men were found to be overrepresented compared to other races as well and women were found to be underrepresented in general. This could potentially be viewed as white men being thought of as the default in the US leading to their overrepresentation in datasets and is something that should be looked at further.

Looking more specifically at the racial biases we explored; we found that white and asian-pacific-islanders were more likely to earn greater than 50k compared to their black counterparts. This could be evidence of more overarching social biases from the 70s and 80s that carry on to today where black people were given less opportunities compared to their counterparts. We also found that black people worked more hours on average and earned less than their counterparts which potentially further reinforces the previous point.

References

[1] R. K. Barry Becker, “Adult.” UCI Machine Learning Repository, 1996. doi: 10.24432/C5XW20. Available: https://archive.ics.uci.edu/dataset/2/adult

[2] “Products - Vital Statistics of the US - Technical Appendices.” Accessed: Feb. 10, 2025. [Online]. Available: https://www.cdc.gov/nchs/products/vsus/ta.htm and https://www.cdc.gov/nchs/data/statab/techap94.pdf