1. Just put a number and a summary of what you changed along with your name – Will
2. <https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data/data> the link to a potential database (subject to change) – Will
3. <https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset>

We could also use this data set

1. Gift hub link here <https://github.com/SebasDtscience/DATA6550-Bias-/projects?query=is%3Aopen>
2. 5I set up out file structure on the github project, im just learning how to do this for the first time so if someone finds an error please feel free to fix it and post it here so we can all see
3. Set up week B discussion thread on d2l
4. Uploaded my part of the project and week a chatlogs
5. Set up space for us to put our analysis below here
6. Sorry if this looks weird, it crashed on me and i restored to an older file version before continuing on
7. Wrote up draft of presentation, including areas for you guys to fill in your part of the write up. This is currently below this part starting on page 2.
8. Uploaded my part of the project to github
9. I worked on the analyses for the presentation.

William Holt, Vaishnavi Paniki, and Sebastian Segura

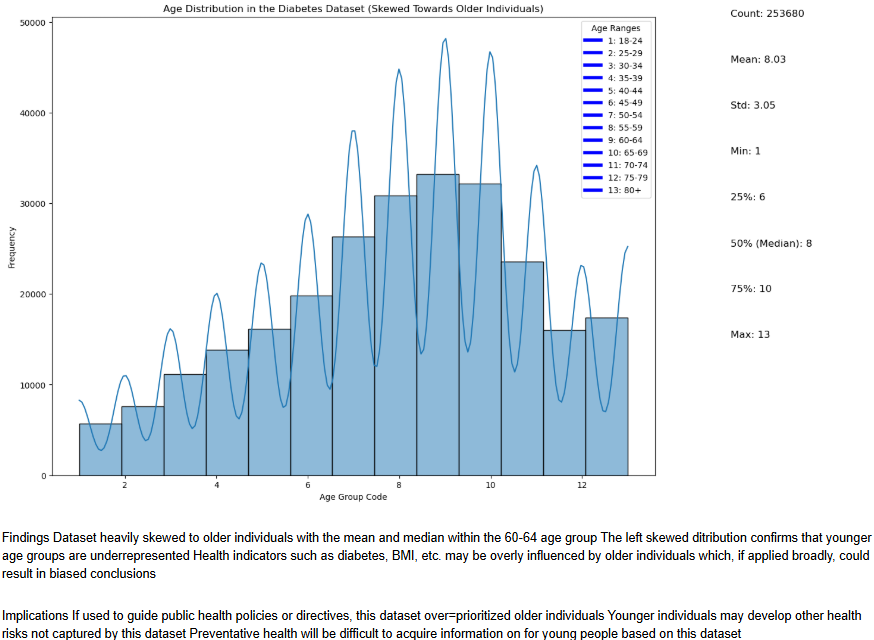
The purpose of this project is to identify, analyze, and report on biases found in the Diabetes Health Indicators Dataset from the Behavioral Risk Factor Surveillance System 2015. The dataset contains health factors for over 250,000 individuals. This dataset has been widely used for diabetes research, despite being known to have present biases. This project aims to explore these biases and better understand their implications for ethics in data science.

The dataset, contains 250,000+ records and several health indicator variables. The key variable in the dataset is Diabetes\_binary, which allows for our model training to study biases further. Some known biases in this dataset include age bias, gender bias, and various other types of sampling bias.

Bias Analysis

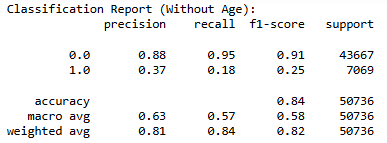
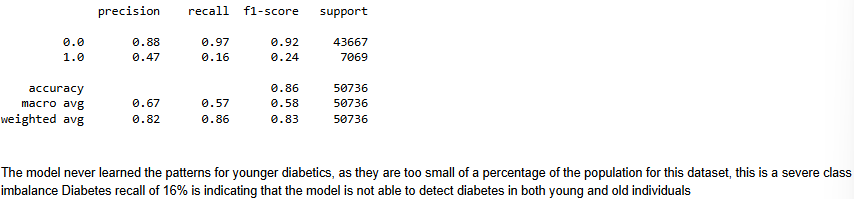
Age Bias

This dataset is very heavily skewed towards older people, which poses multiple issues for a dataset intended to explore public health issues. Through initial exploratory data analysis, we visualized the distribution of age to confirm that it is in fact showing a much higher average age than the general population.



Based on this initial analysis, the average age in the dataset is from 60-64, with only 10% under 45 years old. The left-skewed distribution confirms that younger age groups are underrepresented significantly. Should data biased in this way be used to shape public health policy, the results would limit the amount of preventative care available to younger individuals. It would be very difficult to acquire sufficient information on young people from this dataset.

To further investigate the impact of this age bias, two randomforest models were trained, one predicting diabetes with age, and one without age as a variable.



The first model, with age, was 85% accurate in predicting “no diabetes” but only 16% accurate in “diabetes” cases. Age was the second most important factor following BMI, indicating a strong influence in the model’s predictive power.

The second model, without age, maintained a high accuracy (82%) for “no diabetes” and a slight improvement (18%) in accuracy for “diabetes” cases. BMI remained most important, followed by various other demographic variables.

These results show that age is a significant predictor of diabetes, but less so than other factors in measuring these model’s performances. The dataset’s overrepresentation of older individuals skews the model’s predictive ability, lessening its accuracy among younger people. If a model was trained on similarly biased data and used to shape health policy, it would lead to the neglect of underrepresented groups. For this dataset, there would be a lack of information on young people leading to lessened attention to their preventative healthcare needs, particularly those at risk of diabetes.

[Vaishnavi analysis]

[Sebastian analysis]

**BIAS ANALYSIS**

**FOR UNDERREPRESENTED GROUPS**

While trying to analyze multiple factors linked with diabetes like sex, age, education, and income. Diabetes seems to be a growing concern in todays population, so with this research we wanted to show the relation between these factors and diabetes. The overall goal was to show the patterns that the underrepresented population may have when it comes to diabetes prevalence.

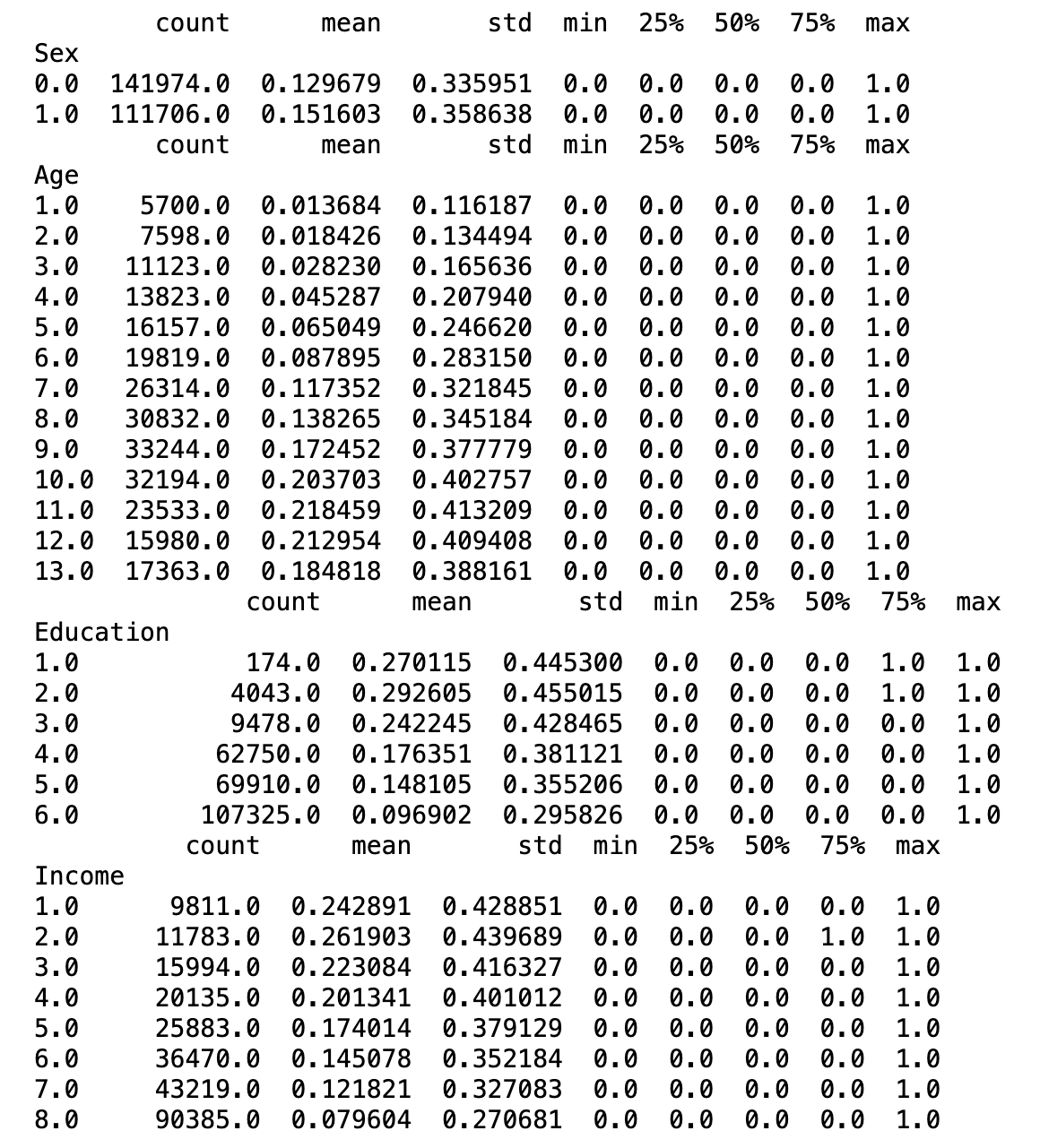
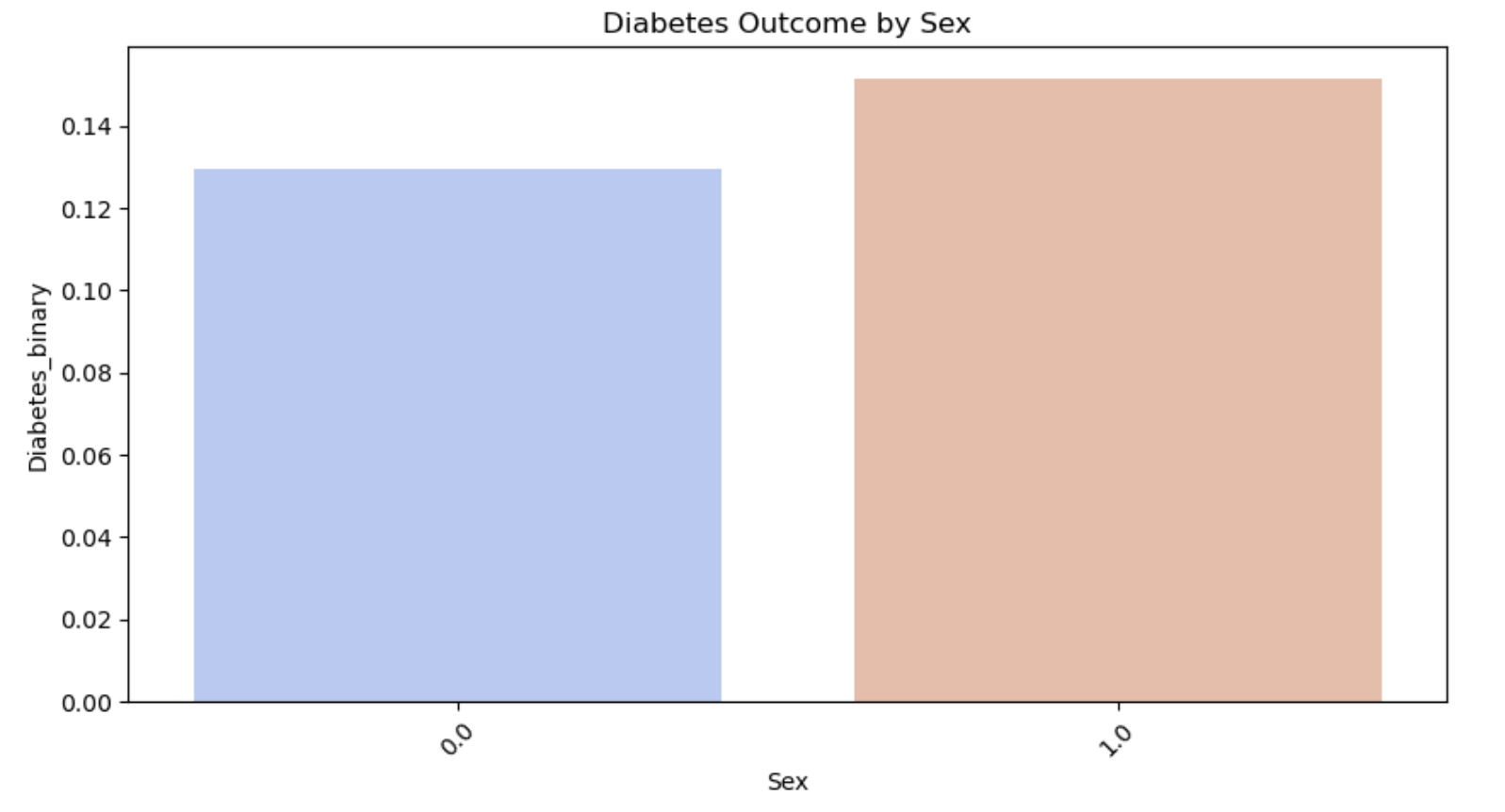


Figure 1: Some of the takeaways from this is that women seem to have a higher prevalence to diabetes than men. Men=0 and Women = 1. According to some research I did, this could be related to women being more susceptible to obesity-related implications which contribute to increased diabetes risk.

Figure 2: Enforces what was said previously on figure 1, which is how women are more prevalent to diabetes than men.

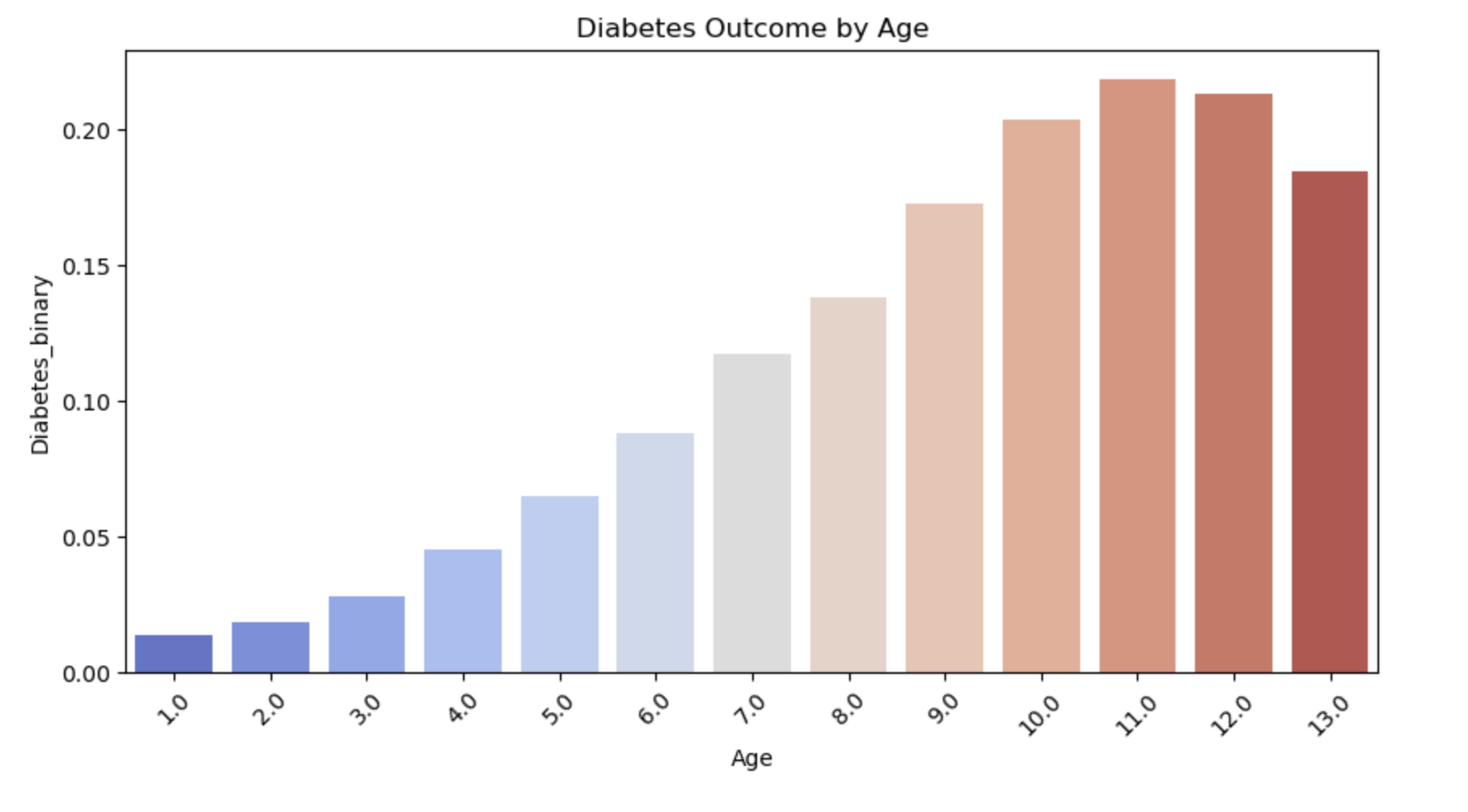
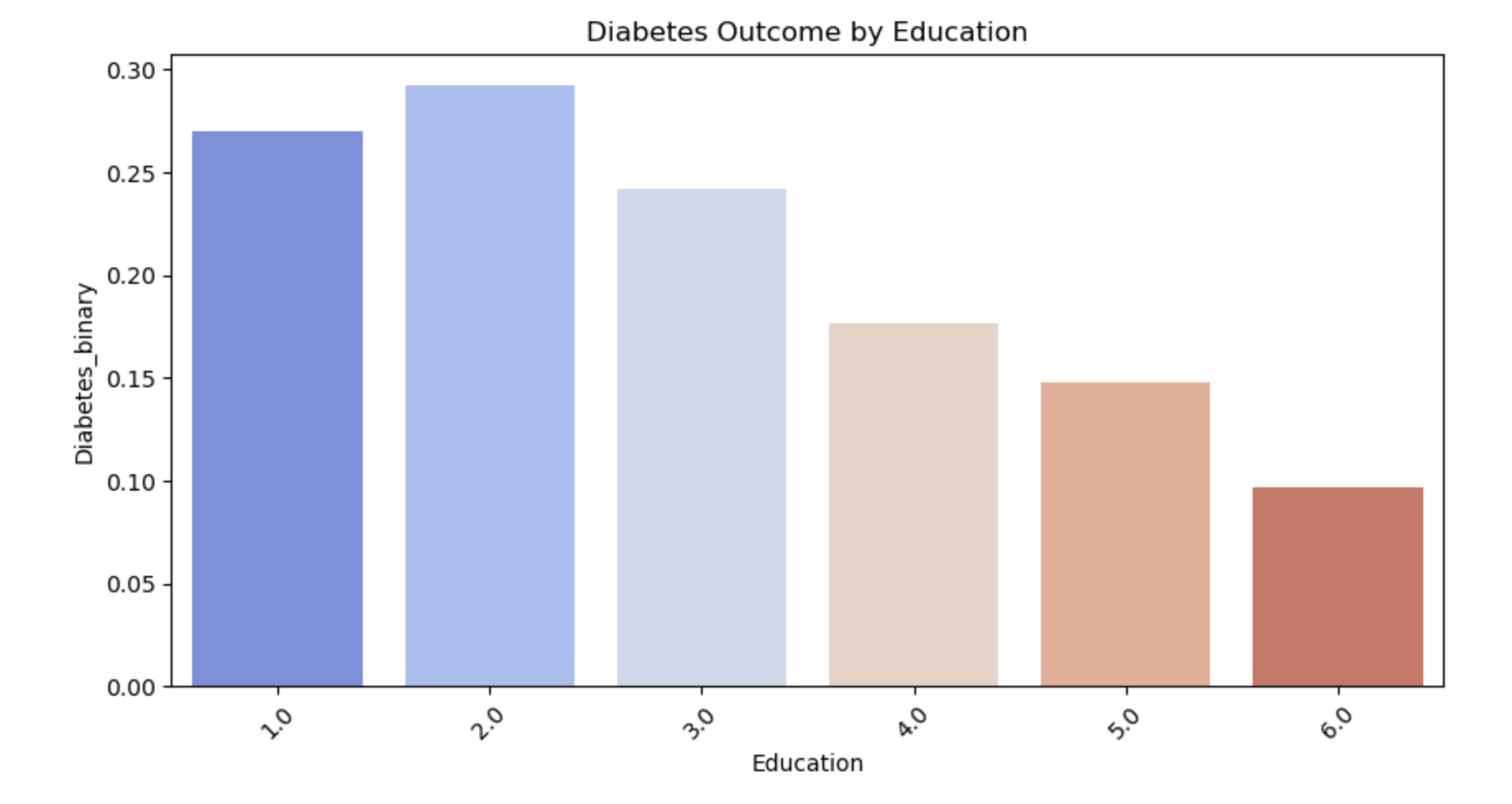
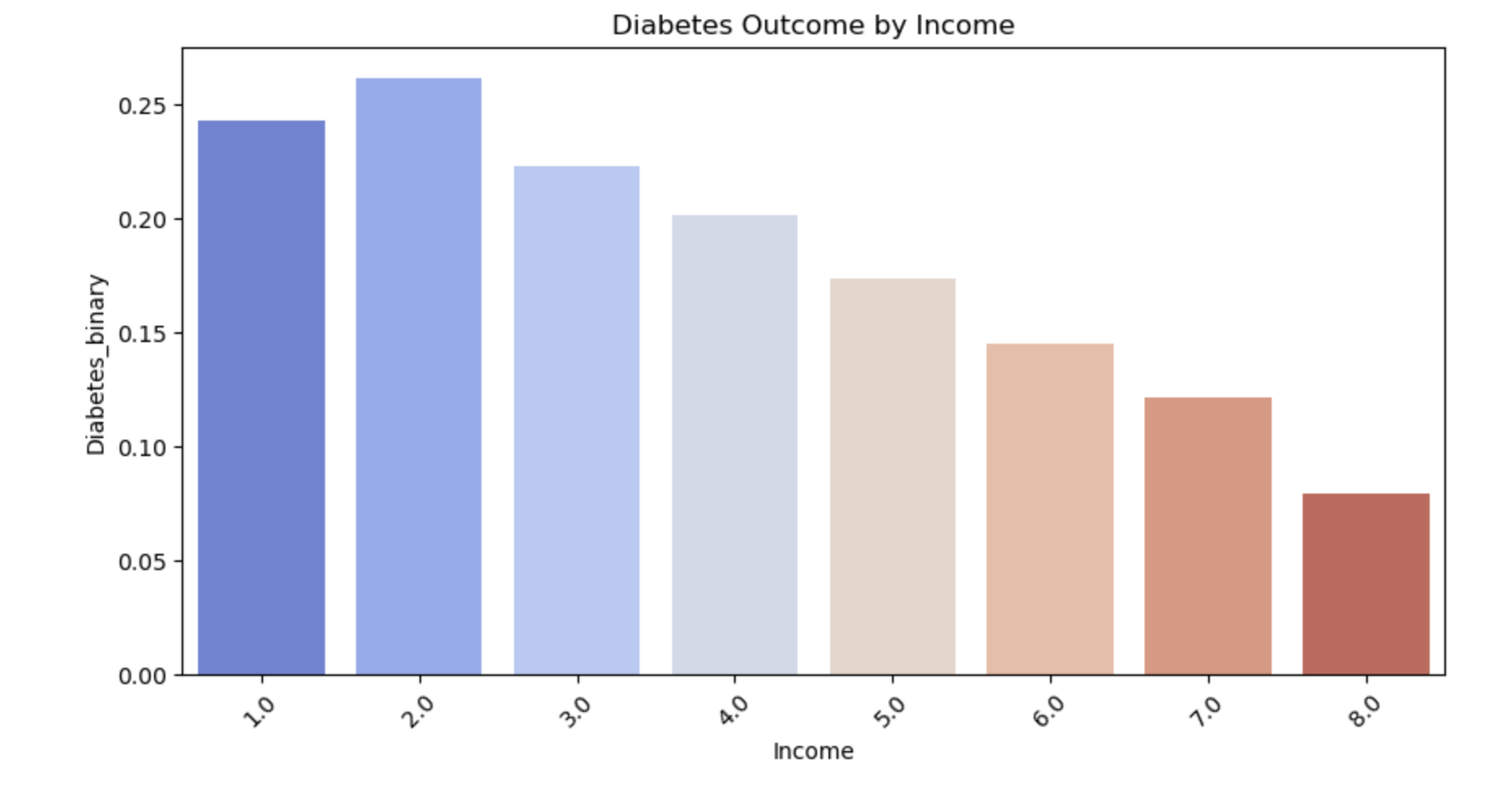


Figure 3: Young adults (1-3) have the smallest predictor for diabetes showing that age does play a role in diabetes. Middle age (6-10) have an increase in risk and older adults have the higher risk for diabetes. When looking online in some possible reasons, I found is that insulin decreases with age which generates an increased in metabolic issues.

Figure 4: Again, this keeps a pattern of showing the underrepresented people are the ones suffering a higher risk. For this graph it shows that the least educated (1-2) with what it seems to be between a percentage of 26% to 29% have a way higher prevalence than most educated people (5-6) with a very low percentage of 9%. This is definitely correlated to income which the next graph will explore. Figure 5: Explores the income and this goes in hand with figure 4 where economic status definetely has a correlation with health=diabetes. This is because more money can get you better foods, medical treatment, overall more resources that would better your health.

This project explored the presence and impact of different types of bias from the Diabetes Health Indicators Dataset. By comparing different levels of analysis and modeling we have been able to determine that age bias significantly skews the dataset, resulting in an overrepresentation of older populations. Our random forest models demonstrated that age is an important factor in diabetes prediction, but not as important as BMI, income, or other factors. Next, we [VAISHNAVI PUT SUMMARY HERE]. Finally, to tie it all together, We can assume based on the data that older age, lower education, and lower income are highly linked to a higher risk of diabetes on patients. Based on some of the graphs shown we can see how underrepresented communities the ones at risk are and more propense to diabetes and overall bad health issues.   
  
 Together, these findings explain the ethical implications behind using biased data to inform public health decisions and their potential negative effects on the underrepresented populations.

Potential mitigation techniques related to bias in healthcare data include collection of more representative data, using weighted sampling to adjust for overrepresented populations, or digging deeper to address different present biases. This project promotes the importance of deeply evaluating datasets for bias and ensuring that data-based decisions come from a place of equity and fairness for all. We could raise some awareness to lower income communities abouut health and maybe help with a more affordable healthcare which would help prevent diabetes with preventative screenings. Promoting a better diet will also help with this problem long term.