Homework Project #3

Yunxing Lu

ylu446@gatech.edu

***Abstract*—**In this homework assignment, I have classified the subgroups into 8 Protected Class Categories, which are listed in the first section. Following the requirements from homework steps 3 to 7, I obtained various statistical measures and plots from different datasets, all originally sourced from the raw dataset. The observations and conclusions are summarized in the later section of this report.

1 CALCULATING TOXICITY CORRELATIONS (MAINLY COVERED RESULTS FROM STEP 3)

1.1 Classify Subgroups into Protected Class Categories

Below is the my classification of each subgroup into its respective protected class category:

Category One: Race

african, african american, black, white, european, hispanic, latino, latina, latinx, mexican, canadian, american, asian, indian, middle eastern, chinese, japanese

Category Two: Religion

christian, muslim, jewish, buddhist, catholic, protestant, sikh, taoist

Category Three: Sex

male, female

Category Four: Gender Identity

transgender, trans, nonbinary

Category Five: Sexual Orientation

lesbian, gay, bisexual, queer, lgbt, lgbtq, homosexual, straight, heterosexual

Category Six: National Origin

chinese, japanese, indian, mexican, canadian, american, european, middle eastern, african, latino/latina/latinx, asian

Category Seven: Age

old, older, young, younger, teenage, millenial, middle aged, elderly

Category Eight: Disability Status

blind, deaf, paralyzed

1.2 Assigning a Unique Number to Each Subgroup Member

After running the Python code on the original datasets, I created a smaller dataset by removing rows where all the values were FALSE. Following the assignment instructions and FAQs, I gave each subgroup member a unique number. Then, I made a new dataset by putting the subgroups together into one column representing the protected class category. The results are listed below, and based on this data, a heat map showing the relationship between each of the protected classes is shown in Figure 1.

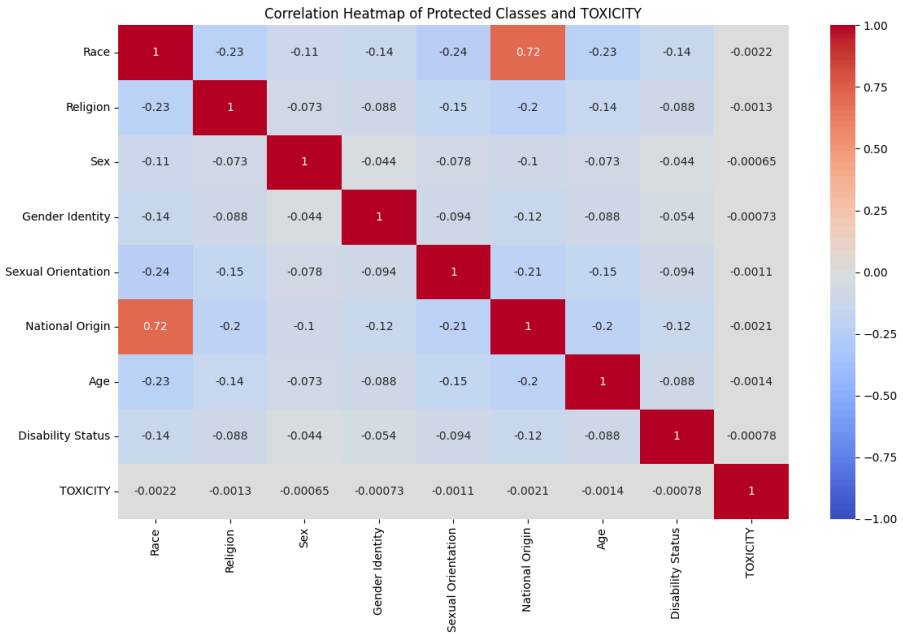


Figure 1—the heat map between each of the protected class

Based on the results above and the correlation table required for the homework, we created Table 1. This table shows the correlation coefficients for each protected class and how strong they are. As you can see in Table 1, all the correlation values are very low, which means their strengths are very weak.

Table 1 - Correlation Coefficients of Each Protected Class and Their Strength

Protected Class	Correlation Coefficient	Strength
Race	0.002239	Very weak
National Origin	0.002094	Very weak
Age	0.001372	Very weak
Religion	0.001319	Very weak
Sexual Orientation	0.001068	Very weak
Disability Status	0.000782	Very weak
Gender Identity	0.000729	Very weak

Sex	0.000648	Very weak

From Table 1, we can see that the top three protected class variables with the highest correlation coefficients are Race, National Origin, and Age. Per the homework requirements, the numerical Subgroup values vs. Toxicity for the three highest correlation coefficients are shown in Figures 2 through 5. Based on the results, I partially agree with the correlation values shown in the plot, but there are some aspects I would question. My analysis and thoughts are summrized as follows:

First of all, I agree with the correlation values for the age group. The plot shows a clear trend of increasing toxicity with age, which aligns with the given information. We can see that as we move from younger age groups (3 - younger, 5 - young) to older age groups (7 - elderly, 8 - teenage), there's a consistent increase in average toxicity scores. This trend makes sense and could be explained by various factors such as generational differences in communication styles or online behavior.

As for the National Origin and Race groups, I don't fully agree with the correlation values for the national origin and race groups. The first reason is that the plot shows some variation in toxicity across different national origins, but the trend is not as clear or consistent as with age. Some groups that we might expect to be culturally similar (e.g., Latino, Latina, Latinx, Hispanic) show quite different toxicity levels. This inconsistency raises questions about the reliability of these correlations.

Similar to national origin, the race category shows variations that are hard to explain solely based on racial categories. For instance, there's a significant difference between "African" and "African American" toxicity levels, which is surprising given the cultural connections between these groups.

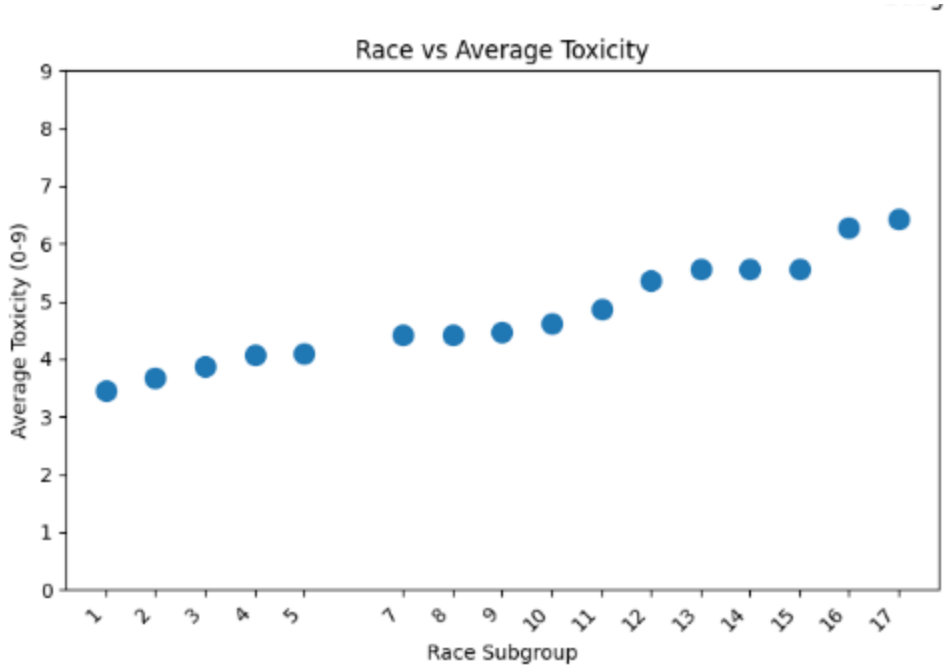


Figure 2—the numerical subgroup values versus the toxicity values plot for the race subgroup

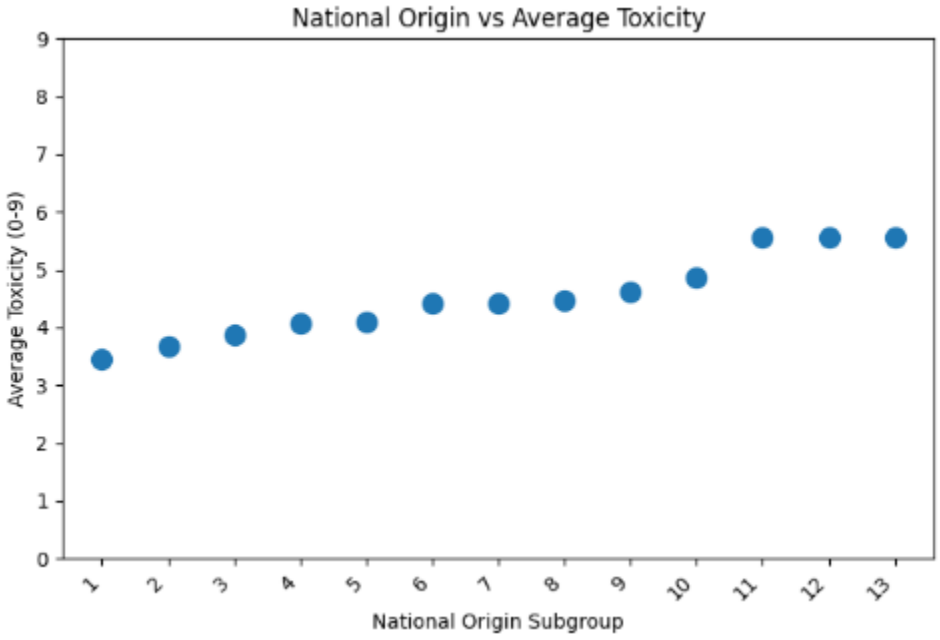


Figure 3—the numerical subgroup values versus the toxicity values plot for the national origin subgroup

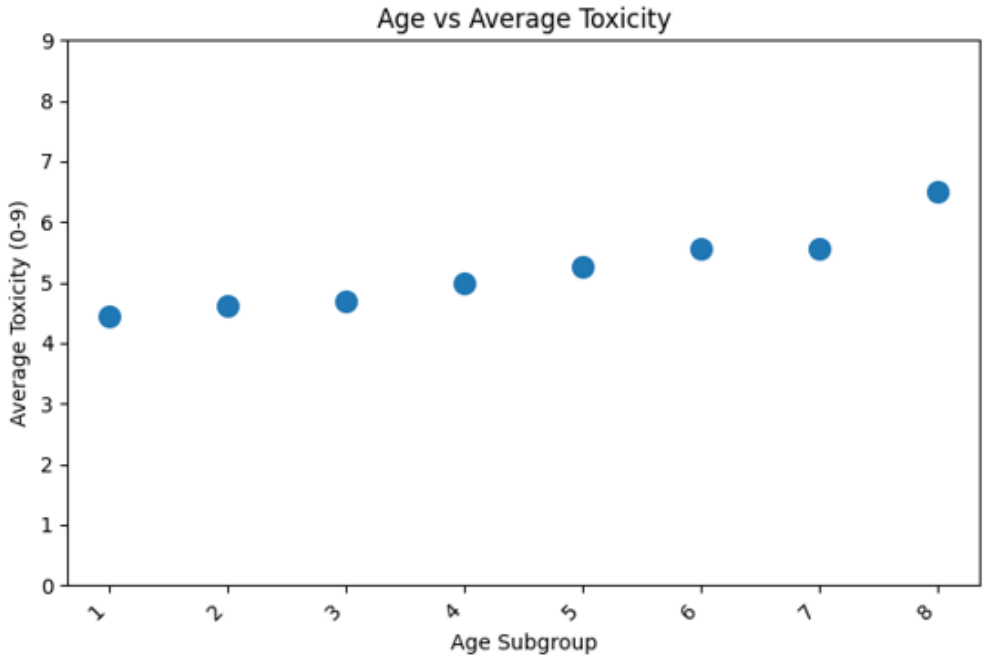


Figure 4—the numerical subgroup values versus the toxicity values plot for the age subgroup

2 ANALYZING TOXICITY ON THE REDUCED DATASET (MAINLY COVERED RESULTS FROM STEP 4)

Using the reduced data from Step 3, the population mean and population standard deviation of TOXICITY are reported as follows:

- Population Mean: 0.5514
- Population Standard Deviation: 0.3617
- 95% of TOXICITY values fall between: -0.1575 and 1.2603

For the random sampling method using 10% of the data, the population mean and population standard deviation of TOXICITY are reported as follows:

- Sample Mean: 0.5568
- Sample Standard Deviation: 0.3617
- Margin of Error: 0.0081

For the random sampling method using 10% of the data, the population mean and population standard deviation of TOXICITY are reported as follows:

- Sample Mean: 0.5505
- Sample Standard Deviation: 0.3619
- Margin of Error: 0.0033

3 ANALYZING TOXICITY FOR A CHOSEN PROTECTED CLASS (MAINLY COVERED RESULTS FROM STEP 5)

Using the reduced dataset from Step 3.1, I calculated the mean and standard deviation of TOXICITY associated with the protected class. I have chosen Age as the protected class, and its associated statistics are as follows:

- Mean TOXICITY: 0.5198
- Standard Deviation of TOXICITY: 0.3680

If use the 10% randomly sampled dataset, the associated statistics are

- Sample Mean: 0.5209
- Sample Standard Deviation: 0.3654
- Margin of Error: 0.0208

If use the 60% randomly sampled dataset, the associated statistics are

- Sample Mean: 0.5224
- Sample Standard Deviation: 0.3682
- Margin of Error: 0.0085

When we use the 10% smaller dataset, it shows that the average value of the protected class (from Step 5.2) fits within the expected range of error (from Step 4.2) for both the 10% and 60% sample sizes. In the case of the 10% sample, the average age from this smaller dataset falls within the expected range of error. This means that even though the sample is small, it still gives us a good picture of the age group in the whole population. However, because the sample size is smaller, there's a higher chance that this result happened by luck. On the other

hand, for the 60% sample, the average age also fits within the expected range. This is what we would expect because a larger sample size usually gives a better reflection of the whole population. This tells us that our way of picking samples is effective and that the age-related findings in our sample likely represent what's true in the general population.

4 ANALYZING TOXICITY FOR SUBGROUPS OF THE CHOSEN PROTECTED CLASS (MAINLY COVERED RESULTS FROM STEP 6)

Using the reduced dataset (Step 3.1), I calculated the mean and standard deviation of TOXICITY for each subgroup that belongs to the Age subgroup, which was selected as the protected class in Step 5. The results are as follows:

Table 2 - Population Statistics for each age subgroup

Subgroup	Mean	Std
older	0.443595	0.381772
middle aged	0.461126	0.377321
younger	0.468178	0.382677
old	0.498378	0.380707
young	0.527458	0.367179

millennial	0.555164	0.344547
elderly	0.555164	0.344547
teenage	0.649199	0.3174

Table 3 - Sample Means within Population Margin of Error (10% Sample)

Subgroup	Mean	Std
older	0.467092	0.381765
middle aged	0.502175	0.375713
younger	0.504959	0.388106
old	0.505501	0.38632
young	0.496685	0.36079

millennial	0.549573	0.343812
elderly	0.569294	0.339019
teenage	0.570506	0.340772

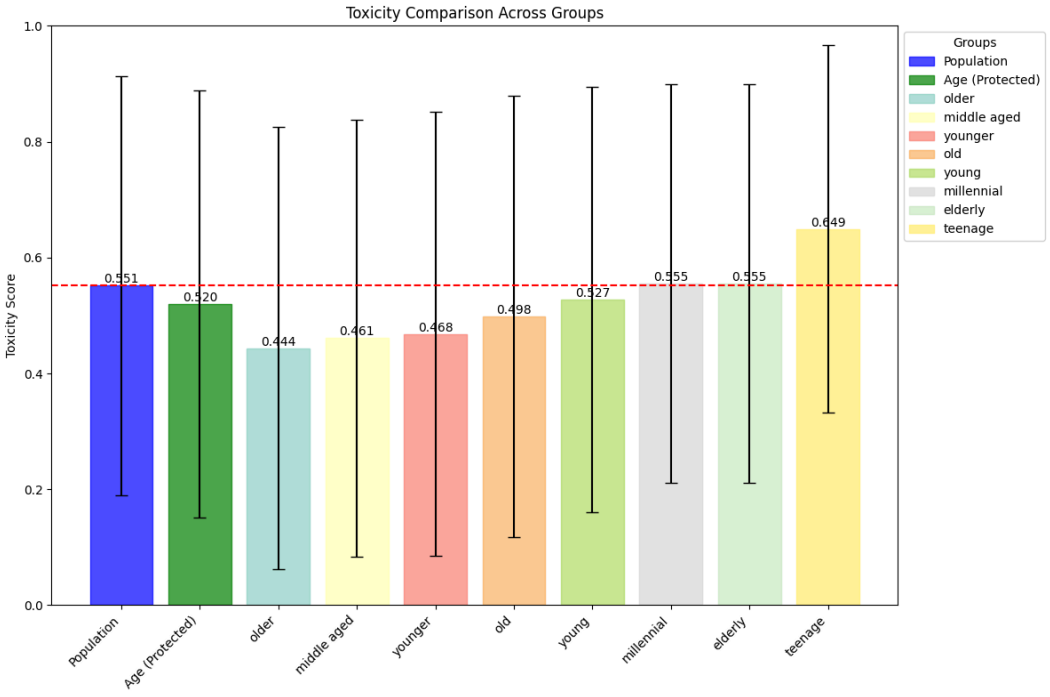
Table 4 - Sample Means within Population Margin of Error (60% Sample):

Subgroup	Mean	Std
older	0.448725	0.383006
middle aged	0.462432	0.376395
younger	0.468227	0.382476
old	0.488439	0.37918
young	0.533666	0.368177

millennial	0.56096	0.345969
elderly	0.563555	0.347367
teenage	0.648654	0.316329

In an analysis of sample representativeness, the 10% sample showed that the means for the subgroups of older, middle-aged, younger, young, and teenage deviated from the expected population means beyond the margin of error, indicating higher variability and less representativeness. In contrast, the subgroups of old, millennial, and elderly were within the margin of error, suggesting these samples accurately reflected the population means despite the smaller sample size. On the other hand, the 60% sample demonstrated improved representativeness, with all subgroups—older, middle-aged, younger, young, old, millennial, elderly, and teenage—falling within the population margin of error. This suggests that larger sample sizes typically provide more reliable estimates that are closer to the true population means, reducing variability across all subgroups.

5 PLOTS AND TOXICITY ANALYSIS (MAINLY COVERED RESULTS FROM STEP 6)



5. 1 Which subgroup has the highest TOXICITY value? Which subgroup has the lowest TOXICITY value? Explain your reasoning.

The subgroup called 'teenage' has the highest TOXICITY value, with an average of 0.649199. This is clear from the data and you can easily see it in the plot—it has the tallest bar compared to the other subgroups. The subgroup called 'older' has the lowest TOXICITY value, with an average of 0.443595. You can see this in the data too, and it shows up as the shortest bar in the plot.

5. 2 Which subgroup has the largest difference in TOXICITY value when compared to the population mean?

The 'teenage' subgroup shows the biggest difference from the overall population's average, which is 0.5514. The average for teenagers is

0.649199, which is about 0.0978 higher than the overall population's average. This is the largest difference compared to all other subgroups.

I think there are several possible reasons for the highest toxicity value in the teenage subgroup: First of all, it could be the developmental factors where teenagers are still developing emotionally and socially, which may lead to more impulsive or confrontational communication. Secondly, it may caused by the online culture: Teenagers may be more immersed in online cultures where toxic behavior is more prevalent or normalized. Thirdly, it could be due to the lack of real-world consequences: The anonymity of online interactions may encourage more toxic behavior among younger users who may not fully understand the impact of their words.

5. 3 What type of human bias is in the data?

The presenting data and results are suggesting an age-based bias, where different age groups exhibit varying levels of toxicity in their online communications. To quantify this bias, we have several potential methods. First of all, we could calculate the deviation of each age subgroup's mean toxicity from the overall population mean. And then compute the range of toxicity values across age groups (difference between highest and lowest subgroup means). Further more, we can use statistical tests like ANOVA to determine if the differences between age groups are statistically significant.

To minimize such bias, we could use several methods. For example, education is a really great option where we can implement age-appropriate digital literacy and online etiquette programs. Secondly, we can deduce the such bias from the platform design point of view. Design online platforms with features that encourage positive interactions and discourage toxic behavior across all age groups. Furthermore, we could also use the age-aware moderation strategies that take into account the tendencies of different age groups.