# CS 8803: AI, Ethics, and Society: Assignment 3

Sagarika Srishti (ssrishti3@gatech.edu)

March 2020

## 1 Answers to Step 2:

Given below are the protected class categories and their associated protected class numbers, with the numerical value assigned to each member. For each protected class, a FALSE corresponds to the value 0.

1. **Sexual Orientation**
   (a) Lesbian: 1
   (b) Gay: 2
   (c) Bisexual: 3
   (d) Queer: 4
   (e) Lgbt: 5
   (f) Lgbtq: 6
   (g) Homosexual: 7
   (h) Straight: 8
   (i) Heterosexual: 9
2. **Sex**
   (a) Transgender: 1
   (b) Trans: 2
   (c) Male: 3
   (d) Female: 4
   (e) Nonbinary: 5
3. **Ethnicity**
   (a) African: 1
   (b) European: 2
   (c) Hispanic: 3
   (d) Latino: 4

(e) Latina: 5

(f) Latinx: 6

(g) Middle Eastern: 7

4. **Race**

(a) African American: 1

(b) Black: 2

(c) White: 3

(d) Asian: 4

5. **Nationality**

(a) Mexican: 1

(b) Canadian: 2

(c) American: 3

(d) Indian: 4

(e) Chinese: 5

(f) Japanese: 6

6. **Religion**

(a) Christian: 1

(b) Muslim: 2

(c) Jewish: 3

(d) Buddhist: 4

(e) Catholic: 5

(f) Protestant: 6

(g) Sikh: 7

(h) Taoist: 8

7. **Age**

(a) Old: 1

(b) Older: 2

(c) Young: 3

(d) Younger: 4

(e) Teenage: 5

(f) Millenial: 6

(g) Middle-aged: 7

(h) Elderly: 8

8. **Disability**
   (a) Blind: 1
   (b) Deaf: 2
   (c) Paralyzed: 3

## 2 Answers to Step 3:

The table below lists the correlation between each protected class variable and the TOXICITY score of a comment, along with the correlation strength.

| Protected Class | Correlation Value | Correlation Strength |
|---|---|---|
| Sexual Orientation | 0.069 | Very weak correlation |
| Sex | 0.086 | Very weak correlation |
| Ethnicity | -0.018 | Very weak correlation |
| Race | -0.032 | Very weak correlation |
| Nationality | -0.118 | Very weak correlation |
| Religion | 0.014 | Very weak correlation |
| Age | -0.026 | Very weak correlation |
| Disability | 0.009 | Very weak correlation |

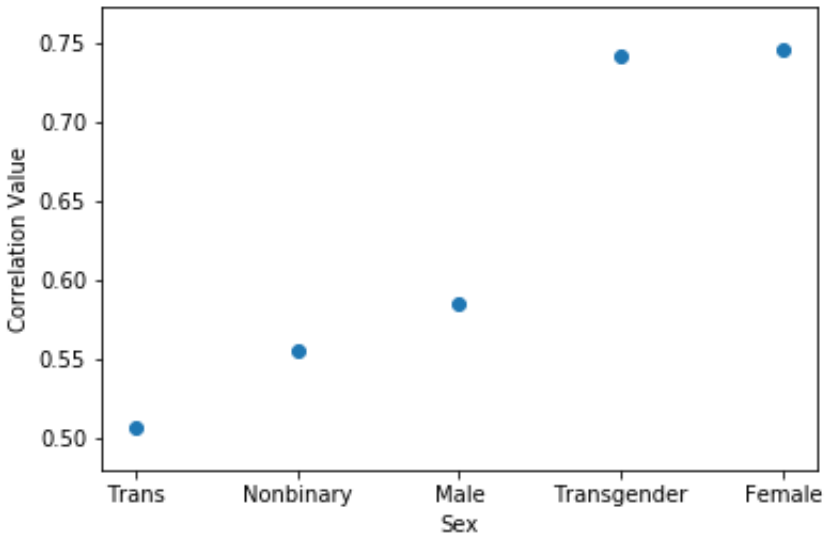


Figure 1: Correlation between TOXICITY and Sex

The protected variables having the three highest correlation values are:

1. Sex
2. Sexual Orientation
3. Religion

Given here are the plots of correlation with these protected classes. Plot between correlation and Religion is on the next page.

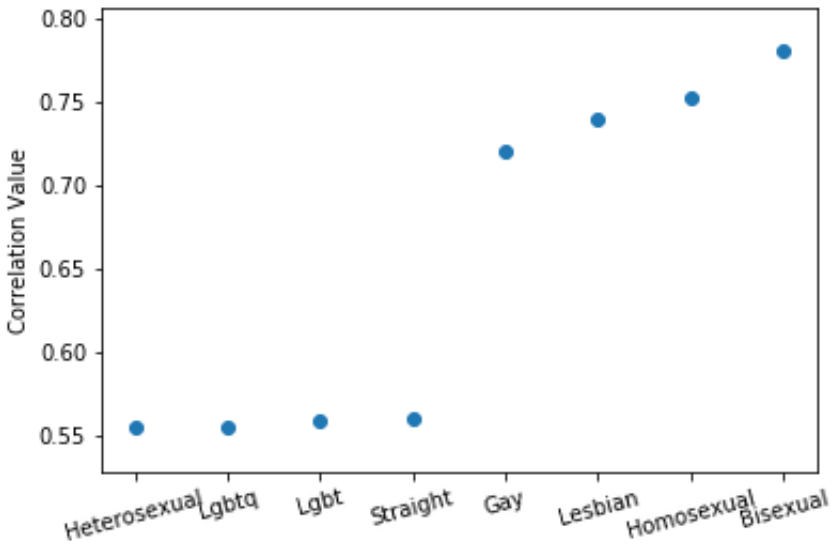


Figure 2: Correlation between TOXICITY and Sexual Orientation

# 3 Answers to Step 4:

- Population mean of TOXICITY: 0.554
- Population standard deviation of TOXICITY: 0.361
- 95% of the population lies within 2 standard deviations of the mean, so the range of values that include 95% of the TOXICITY is: -0.168 to 1.267
- Running the sampling for 25% of the original data:
  - Mean of TOXICITY: 0.555
  - Standard deviation of TOXICITY: 0.359
  - Margin of error: $\pm 0.005$

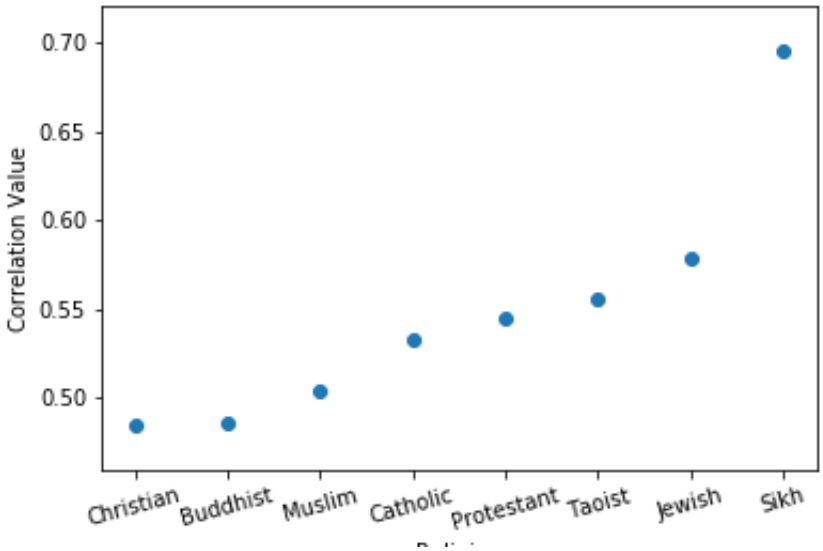


Figure 3: Correlation between TOXICITY and Religion

- Running the sampling for 50% of the original data:
  - Mean of TOXICITY: 0.555
  - Standard deviation of TOXICITY: 0.361
  - Margin of error: ±0.007

## 4 Answers to Step 5:

For this step, I've chosen the protected variable Sex.

- Population mean of TOXICITY: 0.647
- Population standard deviation of TOXICITY: 0.337
- Population margin of error: ±0.011
- Running the sampling for 25% of the original data:
  - Mean of TOXICITY: 0.656
  - Does mean lie within population margin or error? : Yes
  - Standard deviation of TOXICITY: 0.332
  - Does standard deviation lie within population margin or error? : Yes

- Running the sampling for 50% of the original data:
  - Mean of TOXICITY: 0.646
  - Does mean lie within population margin or error? : Yes
  - Standard deviation of TOXICITY: 0.337
  - Does standard deviation lie within population margin or error? : Yes

# 5 Answers to Step 6:

For this step, I've chosen the protected variable Sex.

1. Protected Member: Transgender
   - Population mean of TOXICITY: 0.741
   - Population standard deviation of TOXICITY: 0.257
   - Population Margin of error: $\pm 0.026$
   - Running the sampling for 25% of the original data:
     - Mean of TOXICITY: 0.746
     - Does mean lie within population margin or error? : Yes
     - Standard deviation of TOXICITY: 0.256
     - Does standard deviation lie within population margin or error? : Yes
   - Running the sampling for 50% of the original data:
     - Mean of TOXICITY: 0.745
     - Does mean lie within population margin or error? : Yes
     - Standard deviation of TOXICITY: 0.261
     - Does standard deviation lie within population margin or error? : Yes
2. Protected Member: Trans
   - Population mean of TOXICITY: 0.507
   - Population standard deviation of TOXICITY: 0.379
   - Population Margin of error: $\pm 0.026$
   - Running the sampling for 25% of the original data:
     - Mean of TOXICITY: 0.510
     - Does mean lie within population margin or error? : Yes
     - Standard deviation of TOXICITY: 0.374
     - Does standard deviation lie within population margin or error? : Yes
   - Running the sampling for 50% of the original data:

  - Mean of TOXICITY: 0.497
  - Does mean lie within population margin or error? : Yes
  - Standard deviation of TOXICITY: 0.379
  - Does standard deviation lie within population margin or error? : Yes

3. Protected Member: Male
   - Population mean of TOXICITY: 0.585
   - Population standard deviation of TOXICITY: 0.338
   - Population Margin of error: ±0.026
   - Running the sampling for 25% of the original data:
     - Mean of TOXICITY: 0.575
     - Does mean lie within population margin or error? : Yes
     - Standard deviation of TOXICITY: 0.343
     - Does standard deviation lie within population margin or error? : Yes
   - Running the sampling for 50% of the original data:
     - Mean of TOXICITY: 0.587
     - Does mean lie within population margin or error? : Yes
     - Standard deviation of TOXICITY: 0.339
     - Does standard deviation lie within population margin or error? : Yes
4. Protected Member: Female
   - Population mean of TOXICITY: 0.746
   - Population standard deviation of TOXICITY: 0.291
   - Population Margin of error: ±0.018
   - Running the sampling for 25% of the original data:
     - Mean of TOXICITY: 0.753
     - Does mean lie within population margin or error? : Yes
     - Standard deviation of TOXICITY: 0.291
     - Does standard deviation lie within population margin or error? : Yes
   - Running the sampling for 50% of the original data:
     - Mean of TOXICITY: 0.749
     - Does mean lie within population margin or error? : Yes
     - Standard deviation of TOXICITY: 0.291
     - Does standard deviation lie within population margin or error? : Yes

5. Protected Member: Nonbinary
    - Population mean of TOXICITY: 0.555
    - Population standard deviation of TOXICITY: 0.345
    - Population Margin of error: ±0.026
    - Running the sampling for 25% of the original data:
        - Mean of TOXICITY: 0.566
        - Does mean lie within population margin or error? : Yes
        - Standard deviation of TOXICITY: 0.347
        - Does standard deviation lie within population margin or error? : Yes
    - Running the sampling for 50% of the original data:
        - Mean of TOXICITY: 0.559
        - Does mean lie within population margin or error? : Yes
        - Standard deviation of TOXICITY: 0.343
        - Does standard deviation lie within population margin or error? : Yes

# 6 Answers to Step 7:

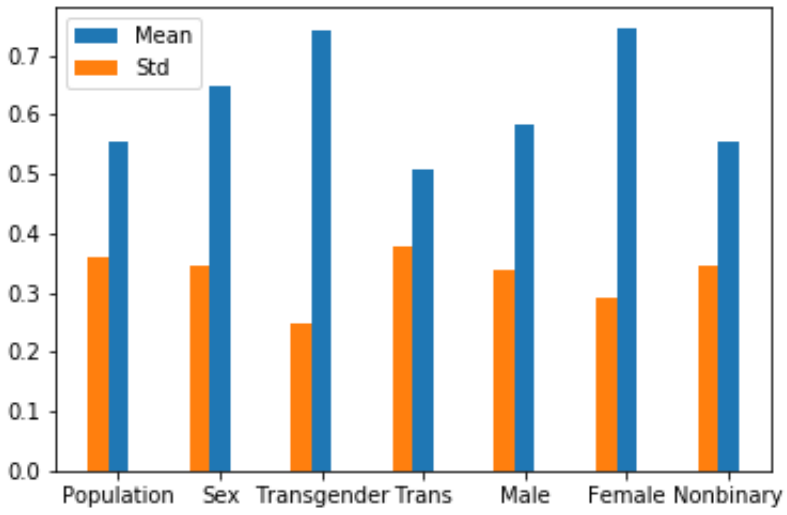


Figure 4: Mean and Standard Deviation values for various subsets of the dataset

- Subgroup having highest TOXICITY: Female
- Subgroup having lowest TOXICITY: Trans
- Subgroup having largest difference in TOXICITY as compared to the population mean: Female

Yes, there seems to be a lot of human bias in the data. There are various examples of this, for example the high correlation of toxicity with the Jewish religion, or the fact that toxicity values are significantly less correlated with the male gender, as compared to females or transgenders.