CS 6603: AI, Ethics, and Society
AI/ML - I Assignment

Cleo Zhang

yzhang3761@gatech.edu

***Abstract*—this assignment will continue exploring relationships in data by computing some basic inferential statistical measures on *toxity_per_attribute.csv*, a modified dataset from the comments from Wikipedia Talk Pages. The following steps will start from a classifier built to identify toxicity in comment items on the given dataset.**

1 PROTECTED CLASS CATEGORIES

See *Table 1* for the identified protected class categories and members associated with each protected class category.

Table 1 — Protected class categories and associated members.

Protected class	**Associated members**
Race	african, african, european, hispanic, latino, Latina, latinx, asian, middle eastern
Sex	male, female, nonbinary
Color	black, white
Sexual orientation	lesbian, gay, bisexual, transgender, trans, queer, lgbt, lgbtq, homosexual, straight, heterosexual
Religion	christian, muslin, jewish, buddhist, catholic, protestant, sikh, taoist
National Origin	american, mexican, chinese, japanese, indian, canadian
Age	old, older, young, younger, teenage, millennial, middle aged, elderly

Protected class	Associated members
Disability status	blind, deaf, paralyzed

2 CALCULATE CORRELATION COEFFICIENTS

To calculate the correlation coefficients, I first removed any rows with all *FALSE* values for every column, then identified an ordering scheme for each protected class and defined values for each protected class member. Finally, I combined the columns associated with the related protected class members into one column (using the *MAX* value) for each protected class. Then I used this compacted table to calculate the correlation between the protected class category and TOXICITY, as shown in *Table 2*.

Table 2 - Correlation coefficients for TOXICITY vs. Each Protected Classes.

Protected Class	Correlation coefficients (vs. TOXICITY)	Absolute values	Correlation strength
Race	*-0.070*	0.070	"Very weak" correlation
Sex	*0.022*	*0.022*	"Very weak" correlation
Color	0.025	0.025	"Very weak" correlation
Sexual orientation	0.202	0.202	weak" correlation
Religion	0.006	0.006	"Very weak" correlation
National Origin	-0.122	0.122	"Very weak" correlation
Age	-0.030	0.030	"Very weak" correlation
Disability status	0.024	0.024	"Very weak" correlation

Select the three highest correlation coefficients and plot subgroups on the X-axis and the Y-axis's average toxicity values. See below for details.

- TOXICITY and Sexual orientation have a "weak" correlation.

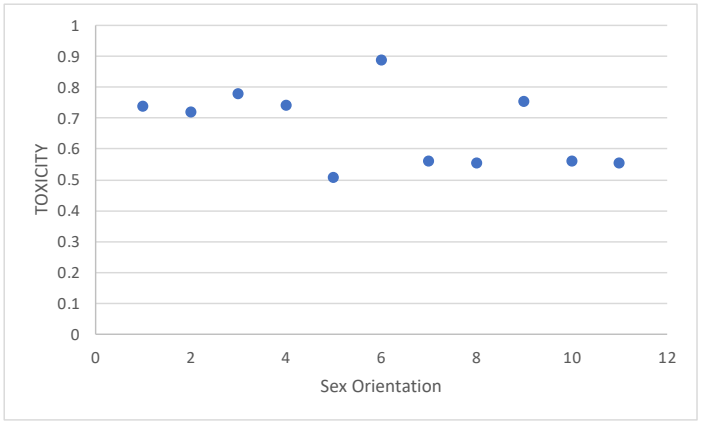

Figure 1 – TOXICITY vs. Sexual orientation.

- TOXICITY and National origin have a "very weak" correlation.

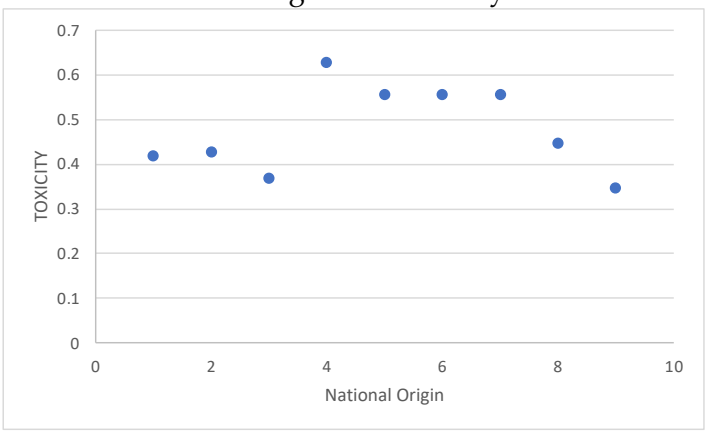

Figure 2 – TOXICITY vs. National Origin.

- TOXICITY and Age have a "very weak" correlation.

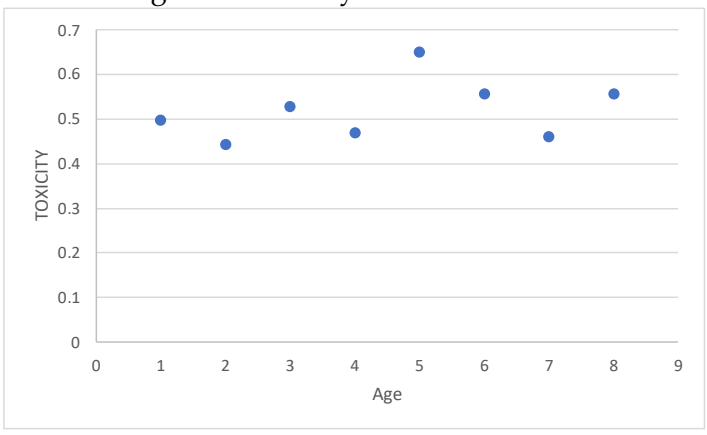

Figure 3 – TOXICITY vs. Age.

3 MEAN AND POPULATION STANDARD DEVIATION OF TOXICITY.

Table 3 shows the population mean and population standard deviation of TOXICITY. According to the Empirical Rule, 95% should fall in the range of $\mu \pm 2\sigma$, where μ is the population mean and σ is the population standard deviation. Therefore, 95% of the TOXICITY lies within 0.724 of the mean (between -0.174 and 1.274).

Table 3 - the population mean and population standard deviation of TOXICITY.

Population mean	Population standard deviation
0.550	*0.362*

Run the random sampling method using 10% and 60% of the records and calculate the mean, standard deviation, and margin of error. See *Table 4* and *Table 5* for details.

Table 4 - the mean, standard deviation and MoE of TOXICITY for the 10% sample.

Mean	Standard deviation	Margin of Error
0.552	*0.361*	*0.0115*

Table 5 - the mean, standard deviation and MoE of TOXICITY for the 60% sample.

Mean	Standard deviation	Margin of Error
0.550	*0.362*	*0.0047*

4 *FURTHER DISCUSSION ON THE SEX CLASS*

4.1 TOXICITY vs. Sex

Using the reduced data set from Step 3, calculate the mean and standard deviation of TOXICITY associated with Sex, and only include the values in the calculation when the associated protected class value is not FALSE. See *Table 6* for details.

Table 6 - the mean and standard deviation of TOXICITY for the chosen protected class - Sex.

Protected Class	Mean	Standard deviation
Sex	*0.581*	*0.340*

Run the random sampling method using 10% and 60% of the records and calculate the mean, standard deviation, and margin of error. See *Table 7* and *Table 8* for details.

Table 7 - the mean, standard deviation and MoE of TOXICITY for the 10% Sex sample.

Protected Class	Mean	Standard deviation	MoE	Mean lies within the associated population MoE?
Sex	*0.563*	*0.334*	*0.047*	*YES*

Table 8 - the mean, standard deviation and MoE of TOXICITY for the 60% Sex sample.

Protected Class	Mean	Standard deviation	MoE	Mean lies within the associated population MoE?
Sex	*0.572*	*0.343*	*0.019*	*YES*

4.2 TOXICITY vs. Male, Female, Nonbinary

Using the reduced data set from Step 3, calculate the mean and standard deviation of TOXICITY associated with Male, Female, and Nonbinary (members of the Sex class), FALSE values excluded.

Table 9 - the mean and standard deviation of TOXICITY for Female, Male and Nonbinary.

Subgroup	Mean	Standard deviation
Male	*0.585*	*0.338*
Female	*0.603*	*0.337*

Subgroup	Mean	Standard deviation
Nonbinary	*0.555*	*0.345*

Run the random sampling method using 10% and 60% of the records and calculate the mean, standard deviation, and margin of error as follows tables.

Table 10 - the mean, standard deviation and MoE of TOXICITY for the 10% samples.

Subgroup	Mean	Standard deviation	MoE	Mean lies within the associated population MoE?
Male	0.575	0.331	0.081	*YES*
Female	0.588	0.340	0.081	*YES*
Nonbinary	0.569	0.352	0.081	*YES*

Table 11 - the mean, standard deviation and MoE of TOXICITY for the 60% samples.

Subgroup	Mean	Standard deviation	MoE	Mean lies within the associated population MoE?
Male	*0.579*	*0.341*	*0.033*	*YES*
Female	*0.604*	*0.337*	*0.033*	*YES*
Nonbinary	*0.533*	*0.347*	*0.033*	*YES*

4.3 Plot on the graph

Figure 4 shows the computed population mean/standard deviation (from Step 4), the calculated mean/standard deviation for the protected class category (from Step 5) and the computed mean/standard deviation for each subgroup of the protected class category (From Step 6).

	Population	Sex	Male	Female	Non-binary
AVG	0.55	0.581	0.585	0.603	0.555
STD	0.362	0.34	0.338	0.337	0.345

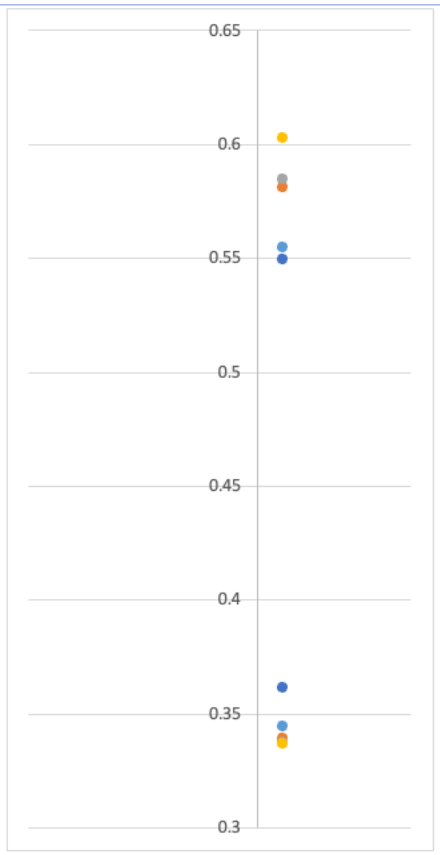

Figure 4 – Plot means and standard deviations.

Based on the scatter graph and the above tables, the Female has the highest TOXICITY value, and non-binary has the lowest TOXICITY value; the Female also has the largest difference in TOXICITY value when compared to the population mean.

There exists sampling bias in the data. As this dataset is modified from the comments from Wikipedia Talk Pages, it will likely only collect data from the people who have internet access and ignore those who don't. Therefore, this dataset only partially reflects the actual population, similar to the data analysis results.