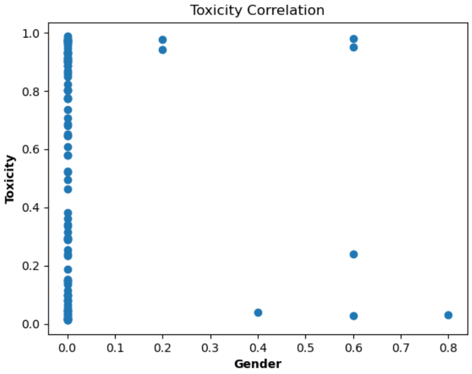
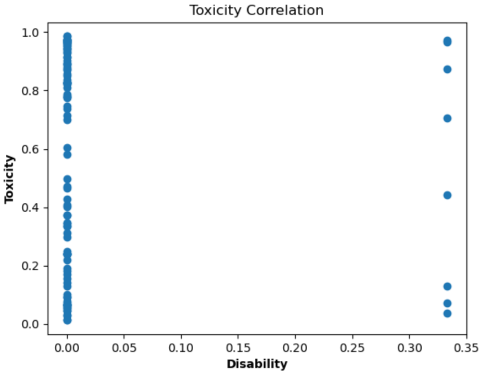
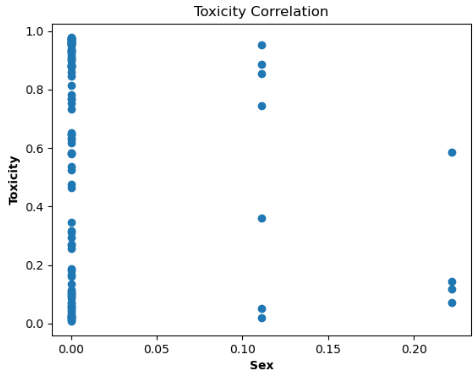
Machine Learning Part 1

Wesley Tomjack

1) **Identify Protected Classes**
    a. Sex (After most recent Supreme Court Ruling regarding sexuality)
        i. Lesbian, Gay, Bisexual, Transgender, Trans, Queer, LGBT, LGBTQ, Homosexual, Straight, Heterosexual, Male, Female, Nonbinary
    b. Race
        i. African, African American, European, Latino, Latina, Latinx, Asian, Indian, Middle Eastern, Hispanic
    c. Country of Origin
        i. Mexican, Canadian, American, Asian, Chinese, Japanese
    d. Color
        i. Black, White
    e. Religion
        i. Christian, Muslim, Jewish, Buddhist, Catholic, Protestant, Sikh, Taoist
    f. Age
        i. Old, Older, Young, Younger, Teenage, Millennial, Middle Age, Elderly
    g. Disability
        i. Blind, Deaf, Paralyzed
2) **Reduced Data Set Correlation**
    a. Features within the dataset were reduced down primarily to their protected classes, except for Sex which was broken out into two groups. ‘Sex’ which is used to classify an individuals sexual preference, and ‘Gender’ (Male, Female, Trans, etc.)
    b. Of these groups the 3 with the highest correlation to toxicity was that of Sex, Gender, and Disability, as shown in the table below.







c. Discrete values were used to replace the True/False categorization of each subgroup. The values for the 3 randomly sampled graphs (n = 100) shown above will be outlined as follows:

   i. Sex

      1. Lesbian/Gay/Bisexual/Queer/LGBT/LGBTQ/Homosexual = 1
      2. Straight/Heterosexual = 2

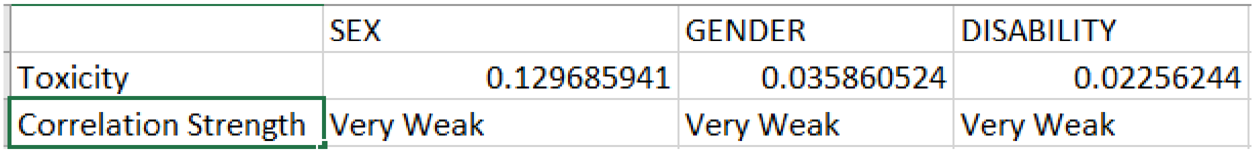
   ii. Gender

      1. Male = 1
      2. Transgender/Trans = 2
      3. Female = 3
      4. Non-binary = 4

   iii. Disability

      1. Blind/Deaf/Paralyzed = 1

d. To determine protected class values (Sex/Gender/Disability) values contained within the subgroups were first aggregated, and the sum of which was divided by the total number of subgroups within that category. Essentially just finding the mean within each class.



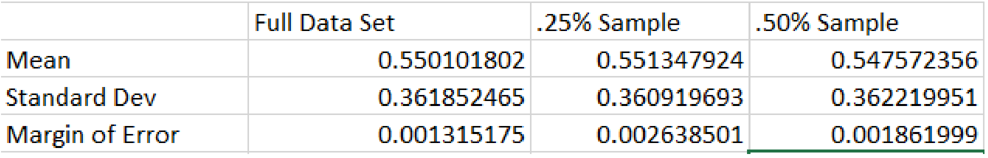
	SEX	GENDER	DISABILITY
Toxicity	0.129685941	0.035860524	0.02256244
Correlation Strength	Very Weak	Very Weak	Very Weak

3) Step 4

a. The toxicity values were computed below for a full population, and 25% and 50% sampling respectively.

b. The ranges surrounding 95% of toxicity based on a normal distribution would be approximately 2 standard deviations away from the mean in either direction, giving us an upper bound of 1.27381295 and lower range of -.173603128

c.

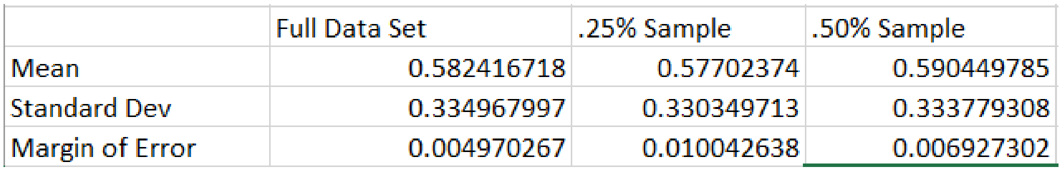


	Full Data Set	.25% Sample	.50% Sample
Mean	0.550101802	0.551347924	0.547572356
Standard Dev	0.361852465	0.360919693	0.362219951
Margin of Error	0.001315175	0.002638501	0.001861999

4) Step 5

a. For this step the protected class of **'Disability'** was going to be used in association with Toxicity measurement

b.



	Full Data Set	.25% Sample	.50% Sample
Mean	0.582416718	0.57702374	0.590449785
Standard Dev	0.334967997	0.330349713	0.333779308
Margin of Error	0.004970267	0.010042638	0.006927302

c. **25% Sample**
   i. The upper limit for the total population is .917384715, and for the .25% sample is .907373453 which is outside the margin of error for the total population
   ii. The lower limit for the total population is 0.247448721, and for the 25% sample is .246674027 which is within the margin of error.

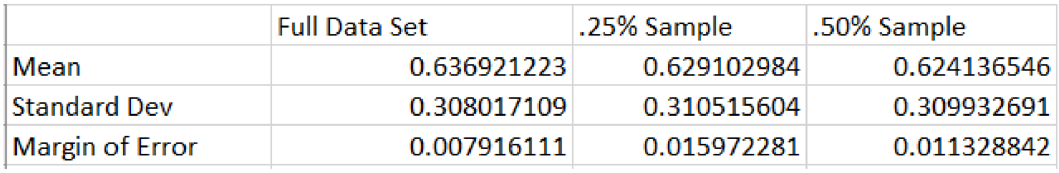
d. **50% Sample**
   i. The upper limit for the total population is .917384715, and for the 50% sample is 0.924229093 which is outside the margin of error for the total population
   ii. The lower limit for the total population is 0.247448721, and for the 50% sample is 0.256670477 which is outside of the margin of error.

5) **Step 6**

a. Similar to the above, we again calculated these measurements but this time within the subgroups of our 'Disability' categorization. The features included under which would be 'Blind', 'Deaf', 'Paralyzed'

b. **Blind Subgroup**

   i.



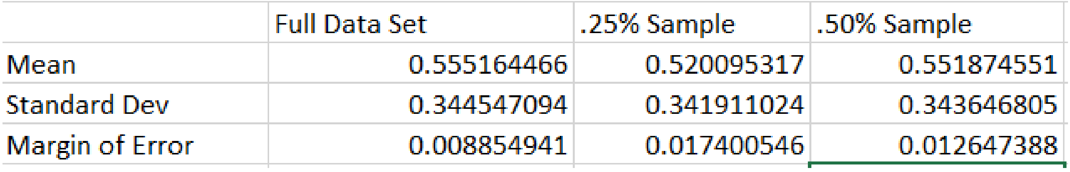
	Full Data Set	.25% Sample	.50% Sample
Mean	0.636921223	0.629102984	0.624136546
Standard Dev	0.308017109	0.310515604	0.309932691
Margin of Error	0.007916111	0.015972281	0.011328842

c. **25% Sample**
   i. The upper limit for the total population is0.944938332, and for the .25% sample is 0.939618588 which is with the margin of error for the total population of 0.007916111
   ii. The lower limit for the total population is 0.328904114, and for the 25% sample is 0.31858738 which is outside the margin of error.

d. **50% Sample**
   i. The upper limit for the total population is0.944938332, and for the .50% sample is 0.934069237 which is outside the margin of error for the total population of 0.007916111
   ii. The lower limit for the total population is 0.328904114, and for the 50% sample is 0.314203855 which is outside the margin of error.

e. **Deaf Subgroup**

   i.



	Full Data Set	.25% Sample	.50% Sample
Mean	0.555164466	0.520095317	0.551874551
Standard Dev	0.344547094	0.341911024	0.343646805
Margin of Error	0.008854941	0.017400546	0.012647388

f. **25% Sample**

i. The upper limit for the total population is 0.89971156, and for the .25% sample is 0.862006341 which is outside the margin of error for the total population

ii. The lower limit for the total population is 0.210617372, and for the 25% sample is 0.178184293 which is outside the margin of error.

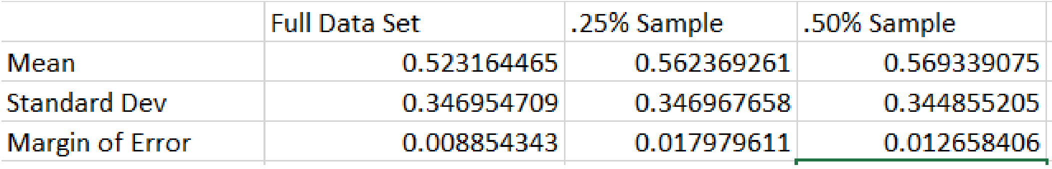
g. 50% Sample

i. The upper limit for the total population is 0.89971156, and for the 50% sample is 0.895521356 which is within the margin of error for the total population

ii. The lower limit for the total population is 0.210617372, and for the 50% sample is 0.208227746 which is within of the margin of error.

h. Paralyzed Subgroup

i.



	Full Data Set	.25% Sample	.50% Sample
Mean	0.523164465	0.562369261	0.569339075
Standard Dev	0.346954709	0.346967658	0.344855205
Margin of Error	0.008854343	0.017979611	0.012658406

i. 25% Sample

i. The upper limit for the total population is 0.870119174, and for the .25% sample is 0.909336919 which is outside the margin of error for the total population

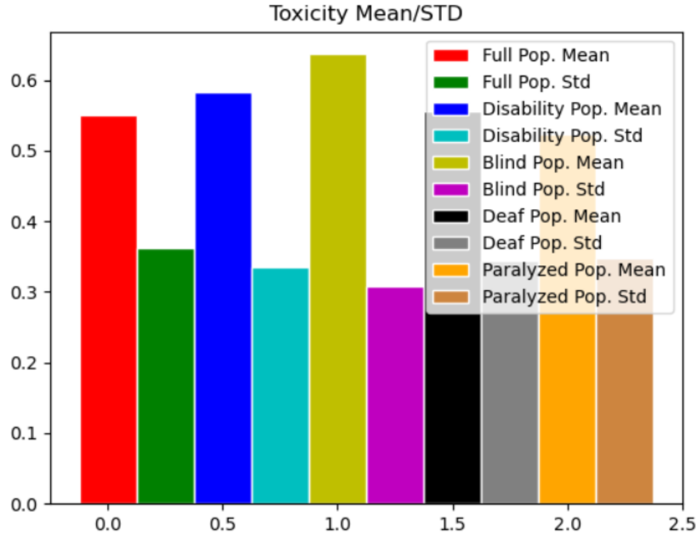
ii. The lower limit for the total population is 0.176209756, and for the 25% sample is 0.215401603 which is outside the margin of error.

j. 50% Sample

i. The upper limit for the total population is 0.870119174, and for the 50% sample is 0.91419428 which is outside the margin of error for the total population

ii. The lower limit for the total population is 0.176209756, and for the 50% sample is 0.22448387 which is outside of the margin of error.

6) Step 7



a.

b. Of the subgroups the blind population had the highest toxicity value, this was determined as the greatest mean, as well as the smallest standard deviation. Meaning the variance of toxicity values were the closest to the mean in comparison to the others.

c. The lowest toxicity of the subgroups would be the paralyzed population, although very close to the deaf population standard deviation the overall mean is lower.

d. The blind subgroup has the greatest difference in toxicity values when compared to the total population, by a value of about .087

e. There is absolutely human bias that went into this dataset, as the rating of each comments toxicity is rated by a human, who has their own inherent implicit bias. These biases would impact the quality of this data as some comments may be rated more toxic by one rater, while rated less by another. Also the subjectivity of being "toxic" opens the door to bias.