CS 6603: AI, Ethics, and Society
Stats 101 Assignment

Cleo Zhang

yzhang3761@gatech.edu

***Abstract*—This assignment will take the Mental Health in Tech Survey dataset as an example and start exploring relationships in data with basic statistical measurements, including identifying dependent and independent variables and creating graphs to support the "fairness" hypothesis and "bias" hypothesis. Furthermore, this report will run the random sampling method using 50% of the original data.**

1 OVERVIEW OF MENTAL-HEALTH-IN-TECH-SURVEY-2019.CSV

Dataset: Mental Health in Tech Survey

Number of Observations: 353

Number of Variables: 82

Regulated Domain in Law: HealthCare

Number of Protected Class Variables: 11

Table 1 — Variables associated with a protected class and the associated legal precedence

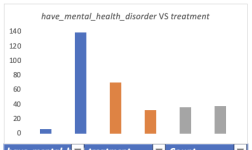
	Converted Variable Name	Protected Class	Law
Do you currently have a mental health disorder?	*have_mental_health_disorder*	Disability status	Rehabilitation Act of 1973; Americans with Disabilities Act
Have you ever been diagnosed with a mental health disorder?	*Diagnosed_with_mental_health*	Disability status	Rehabilitation Act of 1973; Americans with Disabilities Act
Have you had a mental health disorder in the past?	*had_mental_health_disorder*	Disability status	Rehabilitation Act of 1973; Americans with Disabilities Act

	Converted Variable Name	Protected Class	Law
What is your age?	*age*	Age	Age Discrimination in Employment Act of 1967
What is your gender?	*gender*	Sex	Equal Pay Act of 1963; Civil Rights Act of 1964, 1991
What is your race?	*race*	Race	Civil Rights Act of 1964, 1991

2 *RELATIONSHIPS BETWEEN DEPENDENT AND INDEPENDENT VARIABLES*

This section will compute the selected dependent variables (*treatment, mental_health_disclosure, mental_health_support*) as functions of the independent variables identified in *Table 1*.

- Frequency tables and histograms for *have_mental_health_disorder* VS. *treatment, mental_health_disclosure, mental_health_support*



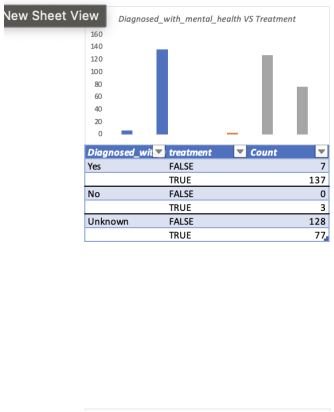
have_mental_h	treatment	Count
Yes	FALSE	7
	TRUE	140
No	FALSE	71
	TRUE	33
Possibly	FALSE	37
	TRUE	38

have_mental_h	mental_health	Count
Yes	Yes	48
	Maybe	47
	No	36
No	Yes	37
	Maybe	27
	No	24
Possibly	Yes	23
	Maybe	17
	No	24

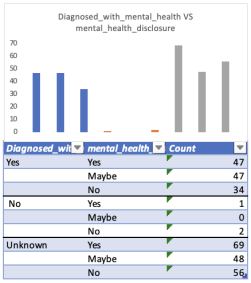
have_mental_h	mental_health	Count
yes	1	23
	2	50
	3	53
	4	19
	5	2
No	1	10
	2	24
	3	51
	4	15
	5	4
Possibly	1	12
	2	28
	3	28
	4	7
	5	0

Figure 2.1 *— have_mental_health_disorder VS. treatment, mental_health_disclosure, mental_health_support*

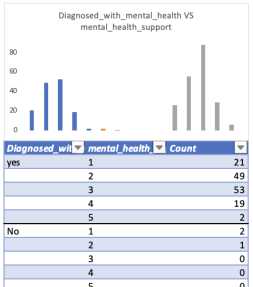
- Frequency tables and histograms for *diagnosed_with_mental_health* VS. *treatment, mental_health_disclosure, mental_health_support*



Diagnosed_wit	treatment	Count
Yes	FALSE	7
	TRUE	137
No	FALSE	0
	TRUE	3
Unknown	FALSE	128
	TRUE	77



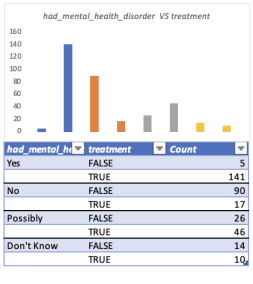
Diagnosed_wit	mental_health	Count
Yes	Yes	47
	Maybe	47
	No	34
No	Yes	1
	Maybe	0
	No	2
Unknown	Yes	69
	Maybe	48
	No	56



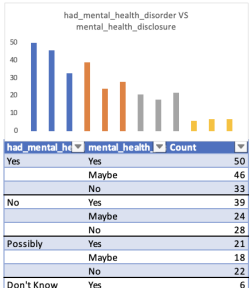
Diagnosed_wit	mental_health	Count
yes	1	21
	2	49
	3	53
	4	19
	5	2
No	1	2
	2	1
	3	0
	4	0
	5	0
Unknown	1	26
	2	56
	3	88
	4	29
	5	6

Figure 2.2 *— diagnosed_with_mental_health VS. treatment, mental_health_disclosure, mental_health_support*

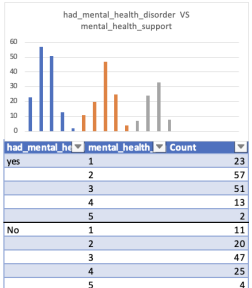
- Frequency tables and histograms for *had_mental_health_disorder* VS. *treatment, mental_health_disclosure, mental_health_support*



had_mental_hi	treatment	Count
Yes	FALSE	5
	TRUE	141
No	FALSE	90
	TRUE	17
Possibly	FALSE	26
	TRUE	46
Don't Know	FALSE	14
	TRUE	10



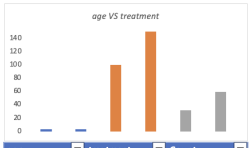
had_mental_he	mental_health	Count
Yes	Yes	50
	Maybe	46
	No	33
No	Yes	39
	Maybe	24
	No	28
Possibly	Yes	21
	Maybe	18
	No	22
Don't Know	Yes	6
	Maybe	7
	No	7



had_mental_he	mental_health	Count
yes	1	23
	2	57
	3	51
	4	13
	5	2
No	1	11
	2	20
	3	47
	4	25
	5	4
Possibly	1	7
	2	24
	3	33
	4	8
	5	0
Don't Know	1	0
	2	0
	3	0
	4	0
	5	0

Figure 2.3 *— had_mental_health_disorder VS. treatment, mental_health_disclosure, mental_health_support*

- Frequency tables and histograms for *age* VS. *treatment, mental_health_disclosure, mental_health_support*



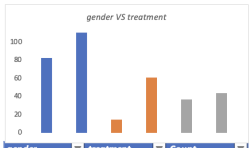
age	treatment	Count
0~20	FALSE	3
	TRUE	3
21~40	FALSE	100
	TRUE	154
41+	FALSE	32
	TRUE	60

age	mental_health	Count
0~20	Yes	1
	Maybe	1
	No	2
21~40	Yes	89
	Maybe	68
	No	69
41+	Yes	27
	Maybe	26
	No	21

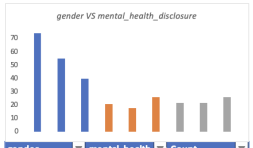
age	mental_health	Count
0~20	1	0
	2	0
	3	0
	4	0
	5	0
21~40	1	29
	2	80
	3	102
	4	37
	5	6
41+	1	19
	2	25
	3	37
	4	10
	5	1

Figure 1 — *age* VS. *treatment, mental_health_disclosure, mental_health_support*

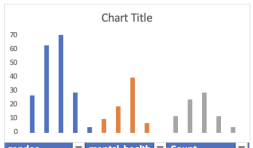
- Frequency tables and histograms for *gender* VS. *treatment, mental_health_disclosure, mental_health_support*



gender	treatment	Count
Male	FALSE	83
	TRUE	112
Female	FALSE	15
	TRUE	61
Other	FALSE	37
	TRUE	44



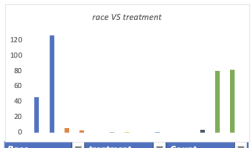
gender	mental_health	Count
Male	Yes	74
	Maybe	55
	No	40
Female	Yes	21
	Maybe	18
	No	26
Other	Yes	22
	Maybe	22
	No	26



gender	mental_health	Count
Male	1	27
	2	63
	3	72
	4	29
	5	4
Female	1	10
	2	19
	3	40
	4	7
	5	0
Other	1	12
	2	24
	3	29
	4	12
	5	4

Figure 2 — *gender* VS. *treatment, mental_health_disclosure, mental_health_support*

- Frequency tables and histograms for *race* VS. *treatment, mental_health_disclosure, mental_health_support*



Race	treatment	Count
White	FALSE	46
	TRUE	127
Asian	FALSE	6
	TRUE	3
Caucasian	FALSE	0
	TRUE	1
European Americar	FALSE	1
	TRUE	0
White Hispanic	FALSE	1
	TRUE	0
Black or African An	FALSE	0
	TRUE	4
I prefer not to ansv	FALSE	81
	TRUE	82

Race	mental_health	Count
White	Yes	69
	Maybe	48
	No	40
Asian	Yes	2
	Maybe	4
	No	3
Caucasian	Yes	0
	Maybe	0
	No	1
European Americar	Yes	0
	Maybe	0
	No	0
White Hispanic	Yes	0
	Maybe	1
	No	0
Black or African An	Yes	0
	Maybe	2
	No	0
I prefer not to ansv	Yes	46
	Maybe	40
	No	48

race	mental_health	Count
White	1	19
	2	49
	3	77
	4	26
	5	2
Asian	1	2
	2	1
	3	3
	4	3
	5	0
Caucasian	1	0
	2	0
	3	1
	4	0
	5	0
European Americar	1	0
	2	0
	3	0
	4	0
	5	1
White Hispanic	1	0
	2	0
	3	1
	4	0
	5	0
Black or African An	1	1
	2	0
	3	3
	4	0
	5	0
I prefer not to ansv	1	27
	2	56
	3	56
	4	19
	5	5

Figure 3 — *race* VS. *treatment, mental_health_disclosure, mental_health_support*

3 DATA MANIPULATION

This section will explain the manipulation of *gender* and *mental_health_disclosure* combination and create graphics to support Fair Hypothesis and Bias Hypothesis.

3.1 Create a graph to support the "fairness" hypothesis.

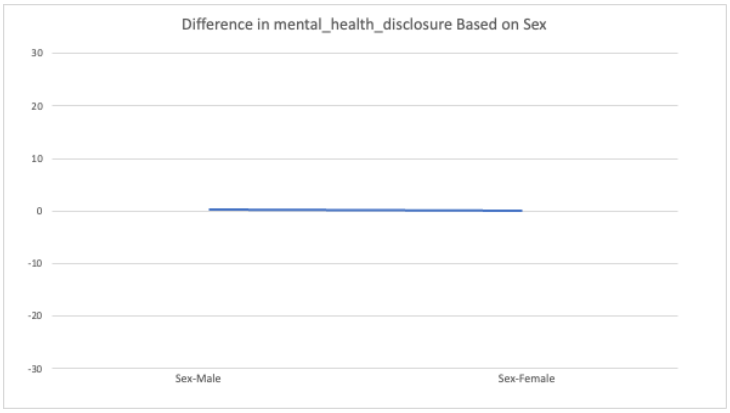


Figure 3.1 – Graphic to support "fairness" hypothesis

As seen from *Figure 3.1*, *mental_health_disclosure* decisions (Yes/No) are not dependent on the gender variable. [Manipulations: Used line graph; Increased the scale to +-30; Mapped the ratio of Yes decisions (74/304 versus 21/304)].

3.2 Create a graph to support the "bias" hypothesis.

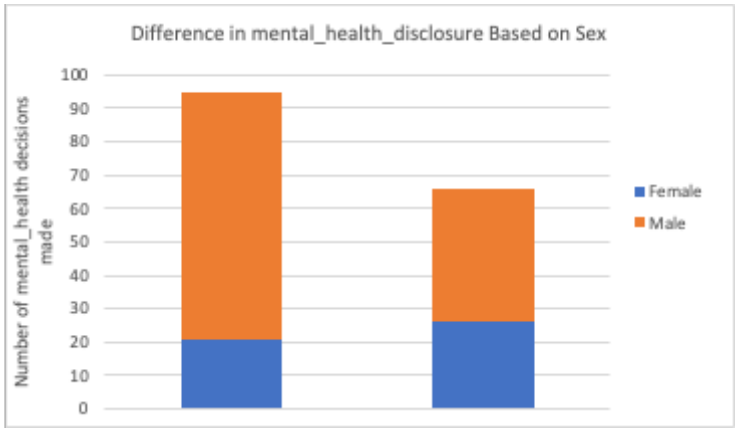


Figure 3.2 – Graphic to support "bias" hypothesis

As seen from *Figure 3.2*, *mental_health_disclosure* decisions (Yes/No) depend on the gender variable. [Manipulations: Used stacked bar graph; Used raw data from the frequency table in Section 2; Reworded labelled].

4 CALCULATE THE AVERAGE (USING MEAN, MEDIAN, AND MODE) OF THE PROTECTED CLASS GROUP.

Assign the answer *Yes, Maybe, No* to 1, 0, -1 for the *mental_health_disclosure* question; I can summarize my computation in *Table 2*.

Table 2 - Variables associated with a protected class and the associated legal precedence

Protected Class Variable (Sex)	**Mean**	**Median**	**Mode**
Original Data Set	0(Maybe)	0(Maybe)	1(Yes)
Reduced Data Set	0(Maybe)	0(Maybe)	1(Yes)
Difference	No Difference	No Difference	No Difference

5 EXPLAIN THE DIFFERENCES BETWEEN THE FREQUENCY TABLE AND HISTOGRAMS BETWEEN THE ORIGINAL AND REDUCED DATASET

Two of the most significant differences shown in *Figure 5* are:

- gender-male: the decisions of *Maybe* and *No* became the same, while the *Maybe* decision is much higher than *No* in the original dataset.
- gender-other: the Maybe decision (instead of No as in the original dataset) became the most popular one.

I used the Random Sampling Method to generate the reduced dataset, and I believe the members associated with the protected class variable are harmed by it. The reasons are:

- Even though the mean, median and mode values seemed the same, the data characteristics were not fully reflected in the reduced dataset, for example, standard deviation.
- The portion of each decision made by each gender became very different compared to the original data, which could be a result of small data samples.

gender	mental_health_disclosur Count	
Male	Yes	35
	Maybe	25
	No	25
Female	Yes	13
	Maybe	7
	No	12
Other	Yes	10
	Maybe	13
	No	12

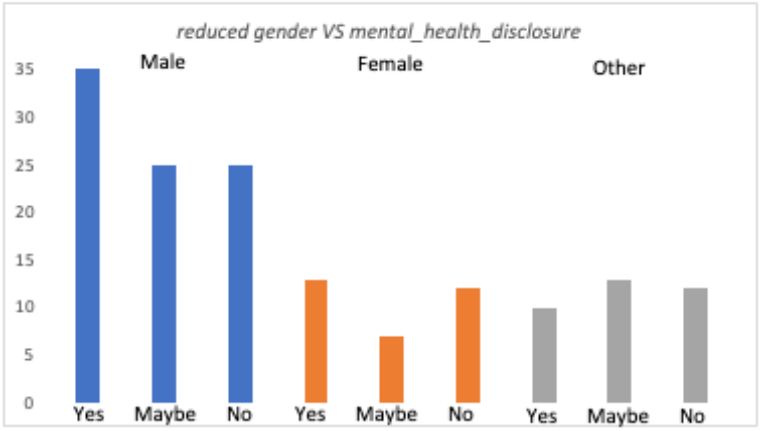


Figure 5 – Frequency table and histogram from the reduced dataset