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 5o% of the original data.

# 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: Health Care

Number of Protected Class Variables: 11

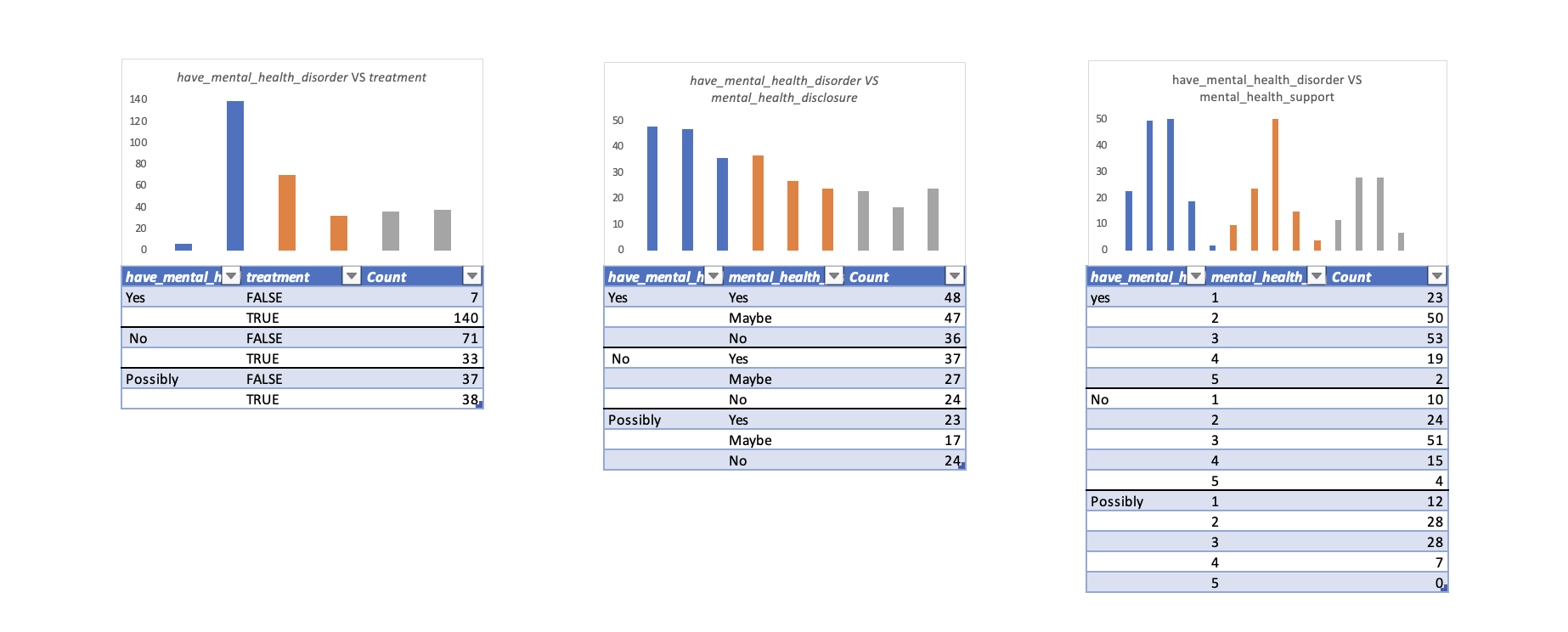
1. Variables associated with a protected class and the associated legal precedence

|  |  |  |  |
| --- | --- | --- | --- |
|  | 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 |
| 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 |

# *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*



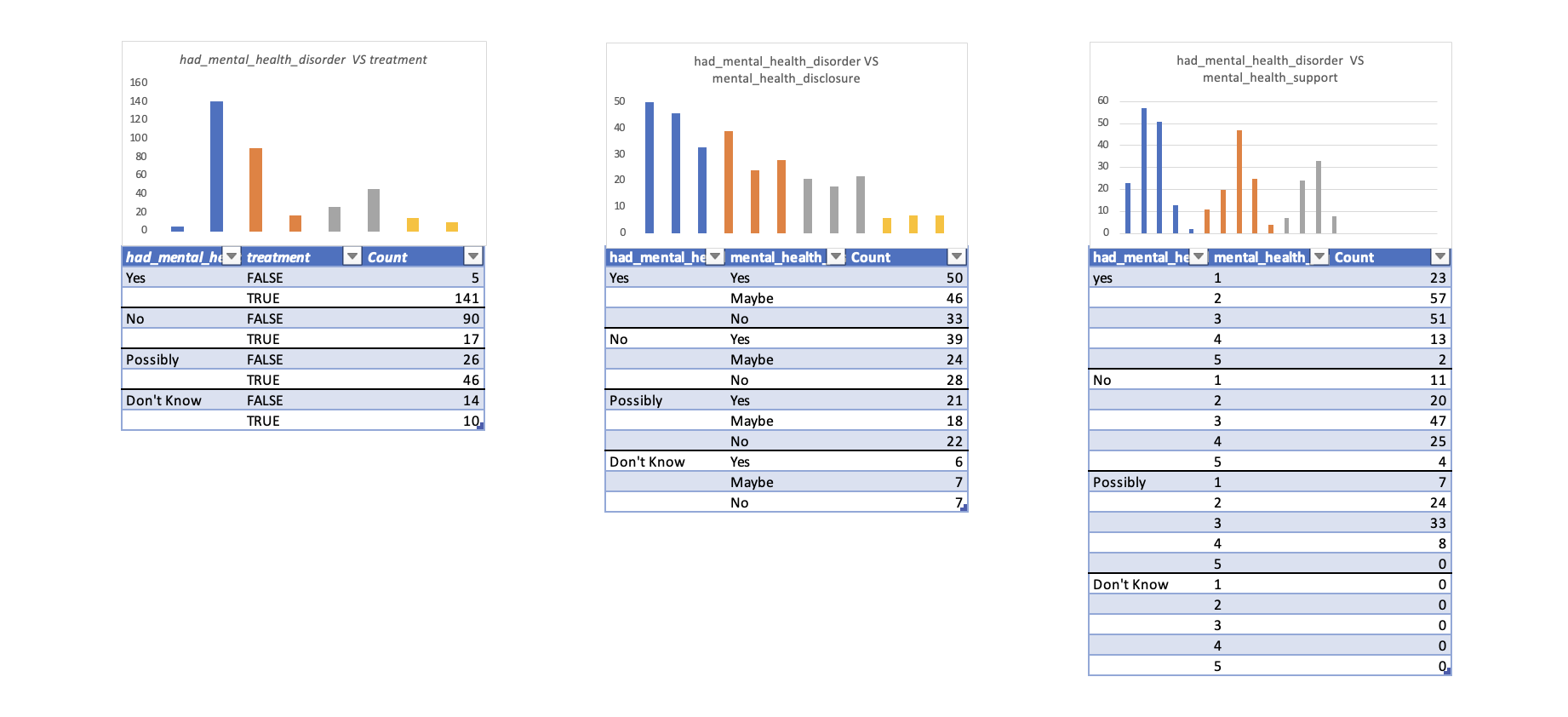
1. *have\_mental\_health\_disorder* VS. *treatment, mental\_health\_disclosure, mental\_health\_support*

* Frequency tables and histograms for *diagnosed\_with\_mental\_health* VS. *treatment, mental\_health\_disclosure, mental\_health\_support*



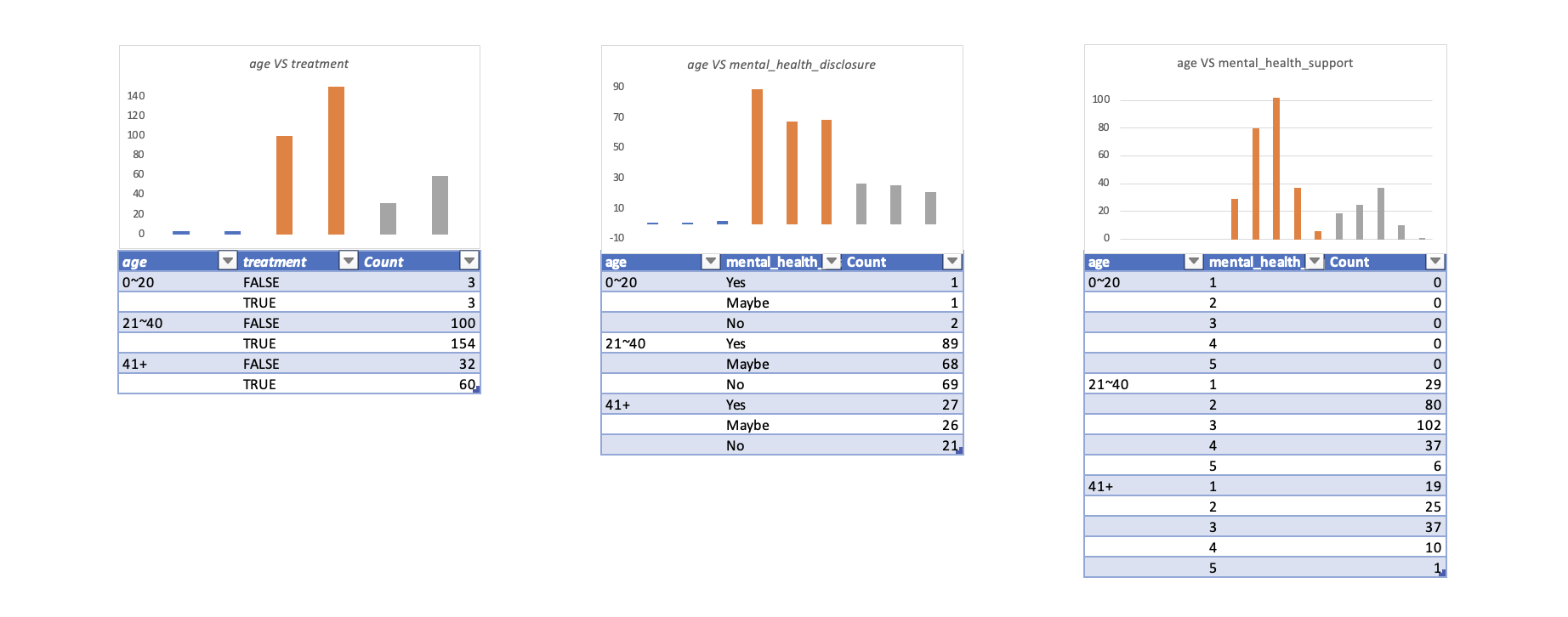
1. *diagnosed\_with\_mental\_health* VS. *treatment, mental\_health\_disclosure, mental\_health\_support*

* Frequency tables and histograms for *had\_mental\_health\_disorder* VS. *treatment, mental\_health\_disclosure, mental\_health\_support*



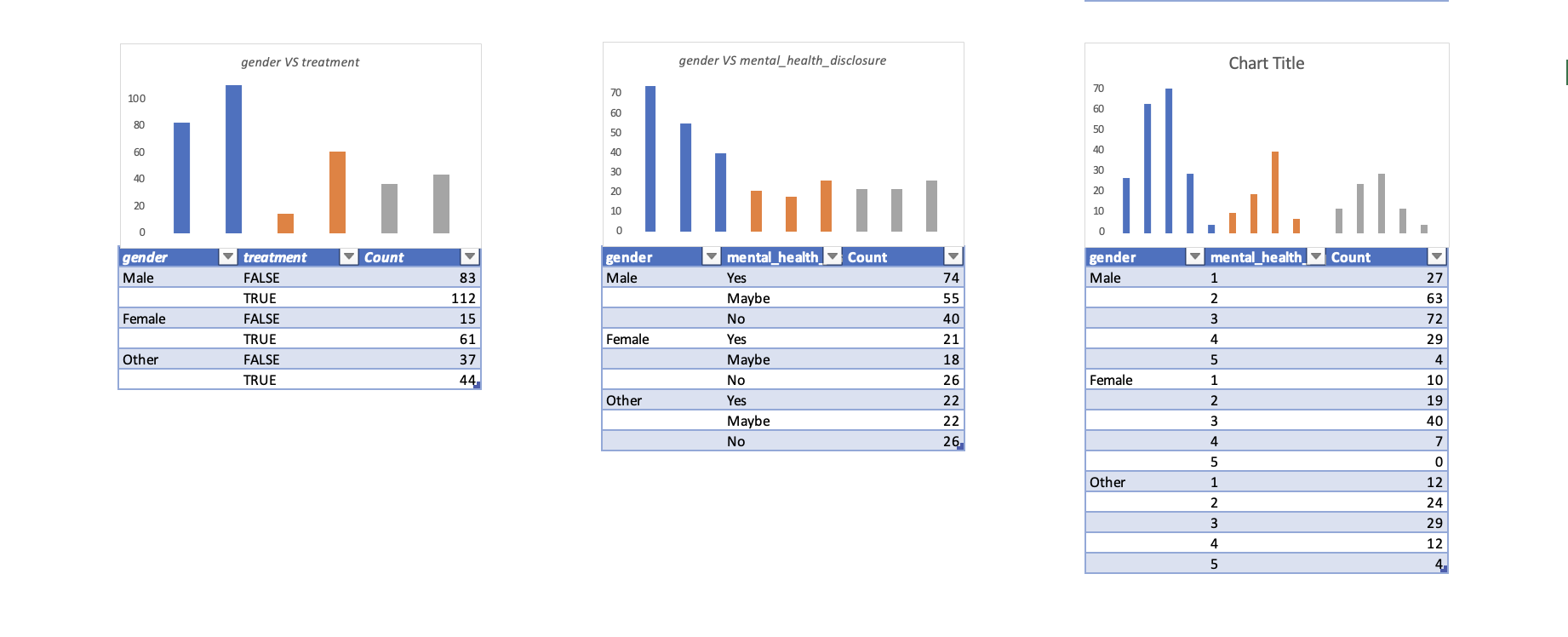
1. *had\_mental\_health\_disorder* VS. *treatment, mental\_health\_disclosure, mental\_health\_support*

* Frequency tables and histograms for *age* VS. *treatment, mental\_health\_disclosure, mental\_health\_support*



1. *age* VS. *treatment, mental\_health\_disclosure, mental\_health\_support*

* Frequency tables and histograms for *gender* VS. *treatment, mental\_health\_disclosure, mental\_health\_support*



1. *gender* VS. *treatment, mental\_health\_disclosure, mental\_health\_support*

* Frequency tables and histograms for *race* VS. *treatment, mental\_health\_disclosure, mental\_health\_support*



1. *race* VS. *treatment, mental\_health\_disclosure, mental\_health\_support*

# *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.

## Create a graph to support the “fairness” hypothesis

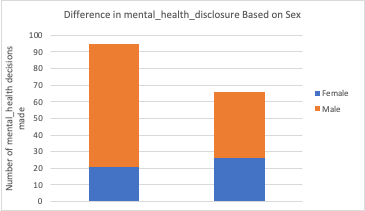
Chart, line chart

Description automatically generated

*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)].

## 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) are dependent on the gender variable. [Manipulations: Used stacked bar graph; Used raw data from the frequency table in Section 2; Reworded labeled].

# 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 to *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 |

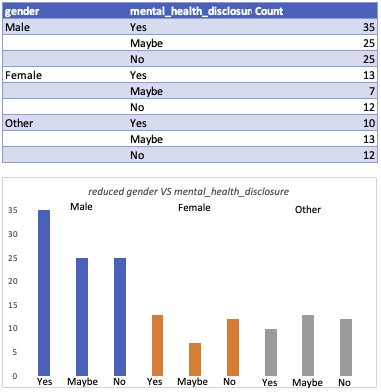
# Explain The differences between frequency table and histograms between the original and reduced dataset

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 from it. The reasons are:

* Even though the values of mean, median and mode 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.



*Figure 5 –* Frequency table and histogram for from the reduced dataset