Genesis Grant

Ethics & Bias in Data Science

**2.2.6 Cognitive Bias**

Confirmation Bias:

Hailee is doing a project on global warming and its effects on the planet. Instead of looking for

sources from both perspectives, she researches only articles and papers that agree with her stance

and implement only those sources into her project.

Sampling Bias:

A survey of student food options and preferences is given only to cafetoria employees instead of

any other population within the school.

Contextual Bias: Depending on your credit history, residential area, age, etc. banks and credit

companies can decide whether they want to lend you money. Certain individuals will have better

or worse chances based on the context of themselves.

**2.2.8 Decomposing the Problem**

1. Columns needed to answer statistical questions would include defining facts or characteristics.

Maybe the financial earnings of a family, region they live, credit score (if applicable), etc. These

data columns can be compared with other varying stats to come up with conclusions. Columns

not needed would include specific information about individuals like names, id numbers,

address, etc. This information is exclusive to each person, so this data would irrelevant.

2. The calculations that could be used in this project include ranges of data to highlight changes

throughout time, we could use averages to compare different numerical data, or ttransfer number

data to charts to visualize significant changes.

3. In the data analysis process, the first step is data extraction, where relevant columns are

identified and extracted from the dataset. Next, data cleaning is performed to handle missing

values and outliers, ensuring that the dataset is accurate and complete. Following this, statistical

analysis is conducted to calculate percentages, averages, and correlations, providing insights into

the relationships within the data. Finally, visualization techniques such as charts and graphs are

used to present the findings in a clear and understandable manner, helping to communicate the

results of the analysis effective.

**4.6.3. Avoiding the Bias**

1. Specifically, the demographics represented in my data only accounts for age and occupation rather than specific demographic characteristics like race, gender, etc. Most demographics (that were given) are represented equally, as there is data for each column, as well as the columns being separated equally (ie age – each age range similar in size, each occupation choice includes different jobs but still relating to overall topic).

2. I attempted to find only unbiased data but there may be biases present, such as sampling bias or response bias, which could affect the accuracy and overall generalization of the findings.

3. Yes, after removing certain unnecessary columns and information, I believe the data found includes complete and good quality data. After fine tuning and filtering, I got back information that encompassed a large variety of occupational and educational data.

4. When left with missing data or outliers specifically related to education, unemployment, and automation, we can first identify any data points that are missing or fall outside expected ranges. For missing data, we can use imputation techniques to estimate values based on other available information.

5. Yes, when first choosing this topic I tried to maintain an unbiased point of view, but because I am the only researcher I realize that I am prone to analyzing more data that fits my argument. To combat this, I did research into arguments against my stance. I did find data against my point of view, and will be sure to implement.

6. To avoid interpreting data, I will focus on presenting the facts and statistics without injecting personal opinions or biases. I will use neutral language and avoid words that may imply a positive or negative bias. For example, instead of saying "Automation has significantly reduced unemployment rates," I would say "There is a correlation between automation and changes in unemployment rates." This way, I can present the data objectively and leave interpretation to the audience

7. My roommates found my topic interesting as they can see how structural unemployment affects people they may know in real life. They added that I should add more emphasis to the opposing side, just to strengthen my argument more.