Data tools

Many applications that we use today are made using databases that use SQL (structured query language). If a single variable is being stored the data file is likely to be a text (txt) file, if there is more than one value being stored then the data may be stored in a comma separated values (csv) file. When using a spreadsheets application like google spreadsheets or a SQL database, we can use functions like count, average, and sum on these rows and columns to get answers that summarize our data. Datasets can be used to find patterns, spotting trends, and even finding a specific case that we may be looking for. Overall, databases and datasets have become extremely relevant and useful for many things we do today. Today, it is almost impossible to use an application on your computer that does not use a dataset or a database. We are surrounded by databases and will continue to be for the foreseeable future with technology.

Big Data

Given the state of technology today, it's critical to evaluate how data will be stored and processed effectively when a computing system needs to store enormous amounts of data. Effective storage and processing of big data become increasingly important as technology develops and the amount of data generated and collected grows tremendously. To ensure data quality, organizations must carefully choose scalable storage options, use effective algorithms and parallel computing approaches, and apply appropriate data management. To ensure data integrity and user trust,

security, privacy, and ethical factors such data encryption, access controls, and transparency should also be taken into account. By attending to these issues, businesses may use big data to spur innovation, improve decision-making, and create new opportunities while still following the moral principles and promising data security and privacy.

Bias in Machine Learning

This section provides a thorough introduction to machine learning, outlining the fundamentals of several methodologies like reinforcement learning, unsupervised learning, and supervised learning. It goes into great detail about neural networks, a well-liked supervised learning technique, outlining their structure and training process. Emphasizing the need to overcome potential biases in the data in order to provide fair and accurate findings, the need of high-quality and diverse training data is emphasized. We are capable of many absurd things, including identity detection, race detection, and even emotion detection, using artificial intelligence. But at the moment, these programs struggle with bias and are unable to operate equally for various types of people as accuracy in the criminal justice system is a major issue for AI. Many people of color have been falsely identified and arrested due to the inaccuracies of artificial facial detection. Artificial intelligence also has a problem with translation from a non gendered language to a gendered language and especially when trying to talk about one's profession.

Unit Test

The first question asked what methods were most likely to decrease the amount of time needed to speed up the data processing of a person's dataset. The solution is usually to split up the work into multiple CPUs or machines. The second question was about summarizing a dataset and what information the data tells us. Most of the questions in this unit test are about what we can understand from a dataset. The other questions were about optimization and how we can benefit from the data. These questions were fairly straight forward and not too difficult due to the fact that we had to answer very similar questions to these in the modules. Overall, I learned some interesting things about databases that I previously did not know but I was already pretty knowledgeable about most of the information in this module.

Part 2

For this project I used Kaggle to find my datasets. My process was fairly brief browsing through different datasets. I actually decided to pick two datasets for this project since I found it somewhat interesting to create these data charts and learn about these topics. For my formatting, I used google spreadsheets which allowed me to automatically create tables and charts using the information posted on Kaggle. I was able to download this zip file of the dataset to my computer to then unzip and upload to spreadsheets with little to no issues. The main struggle was getting the charts to look the way I wanted them to, it took some time but I was able to present the information in the way I had pictured when deciding to use these datasets. Uploading to github

afterwards was also fairly easy, I was able to create my own repository and simply drag and drop these files into a code editor to then push to github.

My first topic was based on Minecraft piracy. The Kaggle website that hosts this data explains that the data was extracted from a registration form of a major Minecraft event organized on discord. This was a problem determined by this discord group as pirating games is a fairly common thing in today's world. My hypothesis about this dataset is that children are largely the people behind this pirating due to the fact that their parents are unwilling to buy the game for them. After analyzing the data, we can see in the charts that a majority of the people are of the ages 14,15,16,17, and 18 with respective percentages of 9.6, 17.1, 18.4, 17.7, and 10.9 totaling up to 73.7 percent of the entire pie chart. This pile of information backs up my initial claim of children being the main culprit of Minecraft piracy, although there was one 43 year old person that was involved in this dataset which I found to be hilarious.

The second dataset that I had used was based on the salaries of Data Scientists throughout the years 2020-2023. My question on this dataset was what is the average salary per year in the data science field? This one took me much longer than the minecraft dataset to visualize. My initial thought was that salaries are bound to be raising throughout these four years but I wasn’t sure by how much. Based on the data retrieved for visualization, we can see that the average salary in USD had gone up significantly from just 2021 to 2022 as it jumped up a whopping $40,000 in between just those two years. This information backs up my initial thoughts but I was very shocked at

the amount that it had risen. In this document I included a bar chart and a scatter plot to show the averages and the distributions of those averages to show that there are some positions that exceed those averages by up to 5 times the amount. I also included a pie chart just because I like pie charts.