Jacob Reed

5/5/2023

CS 1030-001

Final Project

Data Tools

This section of the program reviewed the basics of SQL and database management. This allows for the sorting of information into rows and columns, such as in Microsoft Excel. Furthermore, it covered the basics of storing data sets and commands to make quick calculations based on the data. It also highlighted some terminology regarding data set patterns, such as positive/negative trends and the use of linear regression to make predictions on where aggregate data might go in the future.

Big Data

There is an incredible amount of data in existence. A wide variety of industries utilize vast amounts of data, such as in scientific research, digital libraries, medical records, user-facing applications, and many others. Storing enough data for everything can be challenging. Data is stored by having many hard drives connected into a disk array or disk enclosure. These enclosures can hold up to hundreds of discs at a time, and hundreds of these enclosures can be housed in a data center. Data ending up in company hands isn’t necessarily a good thing, though, since data is also collected on users without explicit consent.

Bias in Machine Learning

Machine Learning is an algorithm which improves itself automatically based on experience. One form of implementing a machine learning algorithm is using a neural network. These networks can be trained to classify data based on weight it gives different characteristics. Because they’re designed to make these decisions based off data fed to them, machine learning has the potential for bias due to any biases in the data it receives. For example, risk assessment algorithms used in criminal justice sentencing will be utilizing data of past convictions and arrests, which has a significant racial disparity in the US. The same can be said for facial recognition, as software generally only works accurately consistently with white faces. Even in language translation, sexism exists in the system as gendered languages will often automatically associate given professions with a specific gender (i.e. men being engineers and women being cooks).

Unit Test

I found the unit test to be relatively easy and simple, where I only got the final question incorrect. Past that, the test was a pretty good summary of the content in the unit. It brought the most important concepts back to me, namely the ability to read and analyze data while being aware of any biases. This unit was helpful, despite not being entirely new information, and summarized important points and concepts we’ve read about in further detail in Hello World.

Data Analysis

While exploring the different datasets available online, I stumbled across one which provided the state and national provision count of live births, deaths, and infant deaths in the U.S. for the years 2021 and 2022. As COVID-19 was still a substantial issue in 2021 as compared to 2022, I hypothesized that the number of deaths in 2022 would be smaller than the number of deaths in 2021. Furthermore, I wanted to examine deaths against live births. Do we have more children being born month-to-month than we do people dying? To find out this information, I used a line chart to map the number of deaths in the U.S. for 2021 and for 2022 month by month. The chart only displayed a significant increase in deaths for the months following July, which suggests the basis of my hypothesis is incorrect, given that I would expect to see more significant differences in deaths at the start of the year, before vaccines were widely distributed.

To answer the question regarding the difference between live births and the deaths in the U.S., I charted their numbers month-to-month separately for the year 2021. Live births had a sharp incline from February to March, followed by a generally high count for the remainder of the year (likely due to couples being stuck in lockdown some 9 months prior). The deaths, however, showed a sharp decline around the same time- notably when vaccines were introduced, before increasing again towards the end of July. For the majority of the year, births were in excess of deaths by an average of 50,000 though my data indicates the opposite at the beginning and ends of the year. Without analyzing more data, this may indicate excessive deaths in the winter months but is too small of a sample size to make any conclusions.

The data used for this analysis was found online at catalog.data.gov. Downloading the data was little issue, though this and other large data sets had a tendency to crash Microsoft Excel when creating tables. To work around this, as well as Microsoft’s underwhelming UI, parts of the data table were copied to Google Sheets then utilized in the creation of the charts. From this point forward, general analysis became very simple. Any biases in the data may come from the politicization of COVID-19 in the U.S. and the subsequent underreporting of deaths that followed in its wake. Furthermore, the data doesn’t necessarily include any births that were off the record or any deaths of the same nature (which, I imagine, make up a statistically irrelevant subset of the overall data).