Matt Roberts
4/5/2023
Final Project Self-Assessment

We divided the load of the majority of our roles across multiple people. I directly handled cleaning and processing for the GDP data and the creation/maintenance of our database, but I also assisted in the cleaning work for consumption and production, and received input on my work as well. Tableau visualizations were shared by basically the entire team as well. Our team was very cohesive which helped quite a bit.

The greatest challenge that I faced actually came in the form of trying to take on some of the production data cleaning. There was a difficult issue with our country lists not quite lining up due to some historic country names being represented differently from different sources. Fuzzy merging is a concept that wasn't really covered in class so it was a unique experience to go figure out a library with no formal training on it. Satisfying, but it was my first encounter with a library that doesn't seem to have a website that is as helpful and easy to find as, say, pydata.org.

There's definitely a few things I'd change if I were doing this project again. We sunk a lot of time into data cleaning and organizing, when we probably would've been better off with a different dataset from the start. The production and consumption tables are really meant to exist as two-index tables, which just adds complexity in the wrong way. We simplified them easy enough, but there was a lot of time wasted in figuring out what to scrub from the data and how to use what was left.

Our team communication was almost exclusively carried out via Slack and Zoom, with prototyping our model being done via jupyter notebooks. On a broad level, everything for our ETL process was done in Python, mostly using the Pandas library. Our machine learning model was the linear regression model from scikitlearn. While we ended up achieving an accuracy score of 70% on our 2021 model, I still feel that we would need to exclude some of our top

producing countries to get a more relevant model for the other 190+ countries that are being dwarfed into the bottom right corner of our model. It's definitely not a significant enough value that I'd recommend anyone start swinging investment money around based on it.