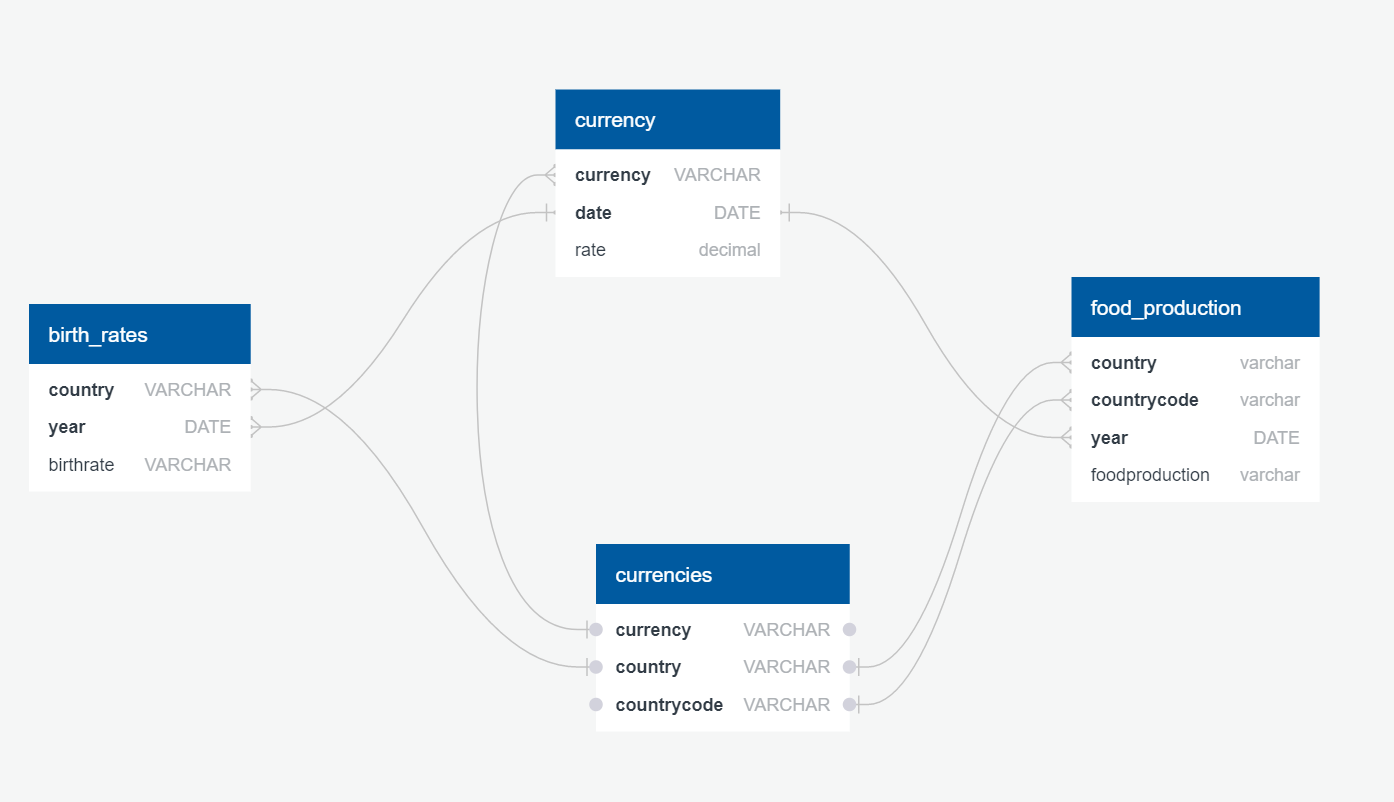
Vaishall Pradeepkumar

ETL Project

On The Subject of Our ETL Project

In the beginning, there was the Word, and the Word was with God, and the Word was God. The same was in the beginning with God. All things were made by him; and without him was not any thing made that was made. In him was life; and the life was the light of men. And the light shineth in darkness; and the darkness comprehended it not. There was a man sent from God, whose name was Youssef. Youssef brought together data from across the land, the food\_production of the land, the currency exchange rates of the sea. Even the birth\_rate data (by country and year) flocked to him. Then Youssef said: “The birth\_rate and food\_production csv’s we shall extract using pandas, and the currency data we shall web-scrape.” And so it was, and the jupyter notebooks were formed, and Youssef saw that it was good.

For transformation, we cleaned up the data and kept only the columns we need. We grouped our exchange rates into years to match the rest of our data. We also only kept countries and dates that we had complete data for. We then exported our dataframes as csvs. For loading, we put all our csvs into a postgres relational database, as described by the ERD below:



We chose data on food production by country and year from World Bank in a csv, birth rates by food and year also from World Bank, and currency exchange rates by month scraped from <https://www.ofx.com/en-us/forex-news/historical-exchange-rates/monthly-average-rates/>. We were curious as to how exchange rates might affect food production, and how food production could affect birth rates.

We took only the relevant columns, and decided that we were only interested in countries that had complete data. We also averaged the currency exchange rates over the course of the year to match our other data. We also needed to restructure some of our dataframes to get the format we wanted, we needed a single year column rather than our data in columns for each year.

As we had a commonality of country across all our data except currencies, we decided to create a relational database connecting the two, as shown in the ERD above. The final database thus contained 4 tables:

* currencies, containing a list of countries, currencies, and country codes
* birth\_rate, containing countries, country codes, dates, and birth rates
* food\_production, containing countries, country codes, dates, and food production
* currency, containing currencies, dates, and exchange rate

The most likely uses for our analysis are for seeing the potential impact of food production on birth rates, and how food production can affect currency exchange rate.