Amazon Rainforest Degradation

Introduction:

The aim of this project is to analyze the effect the rise in temperature has had in the degradation of Brazil's Amazon Rainforest over time. This includes analyzing the temperature changes in the country of Brazil and the degradation area of the forest, to observe if these events are related to one another in some way. For this project, two datasets were used.

Data Sources:

Data Source 1: FAO Temperature Change
This dataset provides the mean surface temperature change for a large number of countries, but for this project, only the country of Brazil will be taken into account. The data is from FAO of the UN, thus the quality can be assured. This data source is of data type CVS, and structured with columns such as country name, value of the mean temperature, etc. The license is CC BY 3.0 IGO, which gives the user the right to redistribute the material and transform and build upon it, even commercially.

Data Source 2: INPE Deforestation Area
This dataset provides the deforestation area of the Amazon Rainforest. The data is from INPE program PRODES, which is dedicated to monitoring the rainforest via satellite. This data source is also of data type CVS, and structured with columns such as occurrence year and different states. The license is CC0 Public Domain, which gives the user the right to modify and redistribute the work, all without asking for permission.

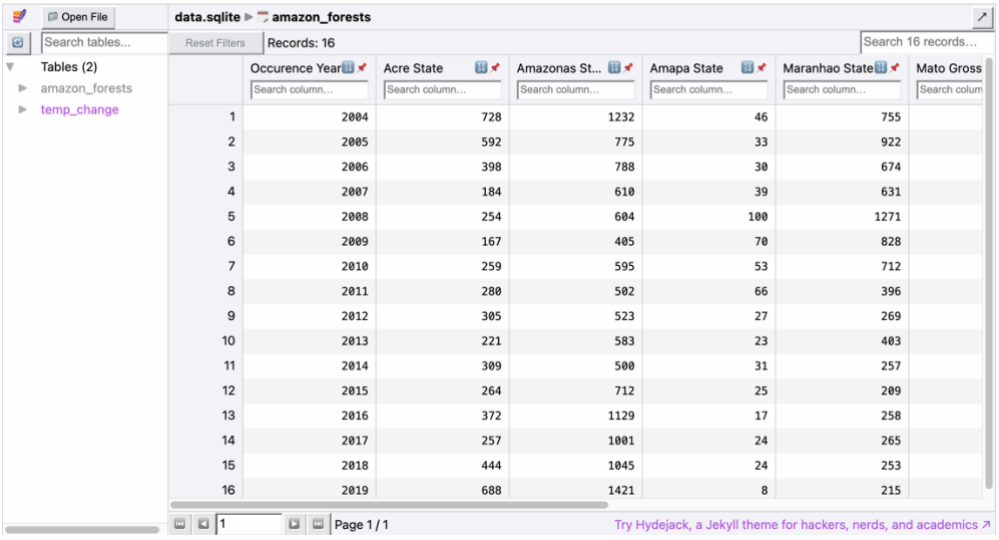
Data Pipeline:

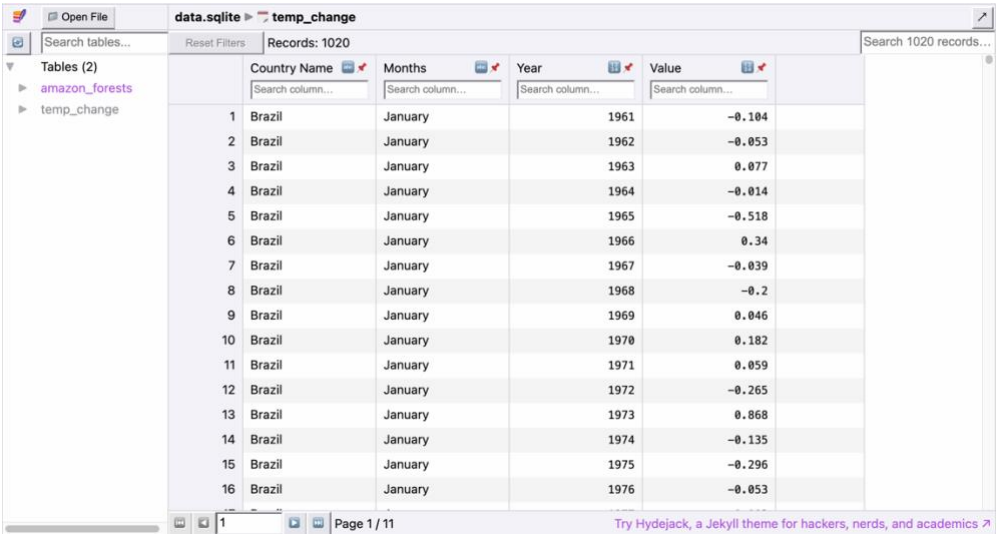
The data pipeline is automated using the ETL (extract, transform, load) process. For the pipeline, various Python libraries were used, such as Pandas (for modifying the datasets), Numpy (for mathematical purposes; in this project it is used to identify empty rows), and SQLite (for storing the modified datasets).

Because the datasets were found on Kaggle, and the website has its own API to extract the datasets, what was done was that the appropriate datasets linked above were put into Google Drive, and extracted from there. On the datasets, some basic transformation was made, where only certain columns were kept

and empty rows and columns were deleted. For the rainforest dataset, a filter was applied so that only the results for the country of Brazil were visible. Then the datasets were saved into an SQLite database as two tables: amazon_forests and temp_change.

One problem that was encountered had to do with the extraction of the datasets from Kaggle. A solution provided from many users online was with using the requests library, but unfortunately, that did not work for me, so to combat this problem, I simply downloaded the appropriate files and uploaded them to Google Drive, so that they could be extracted from there. Although it is not ideal, the use of APIs was prohibited, thus this was the only other solution I could think of.





These are the resulting tables after the transformation.

Results and Limitations:

The output of the following data pipeline is an SQLite database with two tables, that will be used throughout the course of this project to answer the primary question of the effects the temperature has had on Amazonian rainforests. Through certain steps of cleaning the data and eliminating unnecessary rows and columns, the data has been transformed and normalized into appropriate SQL tables that will help in analysis. Because the original datasets were from government sources, the quality and reliability is high.

One potential problem that may arise is the compatibility of these two datasets, due to them having different year range. This might cause inaccuracies when analyzing them and can impact how rigorous the overall analysis may be, considering the documentation of the Amazon rainforest deforestation does not spam across many years.