

|  |  |
| --- | --- |
| ETL Project Analysis of World Bank Data relating to Population, GDP & Gender Ratio’s | |
|  |  |
| Project Group 1September 2022Danny Bruzzese Lin Huan Jhe  Angela Alexander Smith |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Background With the growth of the global human population from 2.5 billion in the 1950's to an estimated 8 billion now, the viable livable areas and resources required are diminishing putting pressure on society.  We saw an experiment “Universe 25, 1968-1973” which is a series of rodent experiments that showed that even with abundant food and water, personal space is essential to prevent societal collapse. Although some people think the experiment was rigorous and human social networks are not like rodent animals’, we still do notice some similar phenomenon found in Universe 25 happening in our society.  Objective  The objective of this project was to extract the relevant data from <https://data.worldbank.org/>, transform the datasets to hold only relevant information from the years of 2000 through to 2020 allowing the following data to be analyzed regarding the following questions.  What impact does GDP have on Population Growth?  As country borders have generally remained static for the past 30 years, does the population increase rate slow down as time passes? How about GDP?  What impact does the Gender Ratio of a country have on the Population Growth? Does this also have an impact on the GDP of a country?  Extract  Three data sets were sourced from the following site <https://data.worldbank.org/>.  GDP.csv  Gender\_StatsData.csv  Population.csv  Utilizing Jupyter notebooks the following dependencies were imported  **import** pandas **as** pd  **import** matplotlib.pyplot **as** plt  **import** numpy **as** np  All 3 csv files were read into the notebook and dataframes created to hold the relevant information.  "InputData/GDP.csv"  "InputData/Population.csv"  "InputData/Gender\_StatsData .csv"  Transformation  After reading in the population.csv the irrelevant columns were dropped to only include the data relating to the Country Name and the years of interest (2000 – 2020) this was done by column number rather than column name in the interest of simplicity. The Country Name column was renamed to countryName for ease of the load process. All year columns were then converted to an integer using the astype method.    This updated data frame was then converted back to an updated csv file.  After reading in the GDP.csv the irrelevant columns were dropped to only include the data relating to the Country Name and the years of interest (2000 – 2020) this was done by column number rather than column name in the interest of simplicity. The Country Name column was renamed to countryName for ease of the load process using MongoDB. Year columns were then rounded to two decimal places.    This updated data frame was then converted back to an updated csv file.  After reading in the Gender\_StatsData.csv the irrelevant columns were dropped to only include the data relating to the Country Name, Indicator Name and the years of interest (2000 – 2020) this was done by column number rather than column name in the interest of simplicity. The Country Name column was renamed to countryName for ease of the load process using MongoDB. This was then further filtered to only include the Indicator Name referencing the population by gender (either Male or Female). The data type was then ascertained and then converted to an integer using the astype method. The reset\_index method was also used to further clean up the final data set.    This updated data frame was then converted back to an updated csv file.  Load  The data was then loaded by two different methods using both relational and non -relational methods.  Using PostgreSQL    The required dependencies were imported  **import** psycopg2  **from** psycopg2 **import** Error  **from** psycopg2.extensions **import** ISOLATION\_LEVEL\_AUTOCOMMIT  **from** sqlalchemy **import** create\_engine  **from** sqlalchemy **import** inspect  **from** SQLkeys **import** password  A host port was then created. We then connected to PostgresSQL to create a Database and then inputted the data using the create engine method. Tables were then created for the three datasets of GPD, Population & Gender.        Following this the tables were then read out into updated CSV files prior to closing the connection.  Using MongoDB  The required dependencies were imported  **import** pymongo  and a connection made to the client. A list of the Country names was created and the years were appended. Using a for loop a collection of the data was created of the key value pairs.    Summary  Using the ETL data integration process we were able to combine the three large csv files from <https://data.worldbank.org/> into a consistent data store that will allow the data analysis and visualization for users to be able to answer the following questions as per our initial objective.  *What impact does GDP have on Population Growth?*  *As country borders have generally remained static for the past 30 years, does the population increase rate slow down as time passes? How about GDP?*  *What impact does the Gender Ratio of a country have on the Population Growth? Does this also have an impact on the GDP of a country?* | | |
|  | | |
|  | | |