

|  |  |
| --- | --- |
| 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 as a whole.  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 in regard to the following questions.   1. What impact does GDP have on Population Growth? 2. 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? 3. What impact does the Gender Ratio of a country have on the Population Growth? 4. 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 notebook “Pandas” was imported as pd and from “sqlalchemy” create\_engine and inspect were imported as our dependencies. All 3 csv files were read into the notebook and dataframes created to hold the relevant information.  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. 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 columns were renamed from just the year to year percentage change for clarity of the information being presented before being 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. 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. This updated data frame was then converted back to an updated csv file.  Load | | |
|  | | |