DSC530 Term Project

Ross L. Kim (Schreck)

Bellevue University

DSC530 Data Exploration and Analysis

Professor Jim

2023.11.18

The Question

The answers I have set out to find for the question ‘What are the main causes of the decline in birth rate of the Republic of Korea has become more complex than I had initially thought. It is obvious now that It would require more extensive data and additional variables. Unfortunately the data used in this project was insufficient.

EDA

After having gone through all the required steps for exploratory data analysis, the results are inconclusive and require further data. The datasets were incomplete and cover limited time spans. I would have used a different dataset if I had known earlier, but due to time constraints, I completed this study with the initial set. According to the relevant data, there is a gradual positive or negative correlation with all variables examined. The most obvious of these were the correlations with birth rate and natural growth rate. This supports the evidence of the rapid population decrease that the Republic of Korea is experiencing; this is also likely the main culprit of this phenomenon.

Missed During Analysis

The available data only spanned several decades from 1970-2020 respectively. Some variables were even shorter time spans from 2000-2020. The variables from 1970 had several outliers and were quite erratic, indicating that there were inaccuracies and missing fields. Ideally, a study of this kind would require data that covers at least 100 years to gain any significant insight.

Variables that Would have Helped

There are definitely several additional variables that would have helped in this study. While the dataset used contained variables that were demographical such as 'birth rate', 'death rate', 'marriage rate', etc, other variables such as annual income, real estate value and prices, cost of living, suicide rate, education level, and happiness (if accurately quantifiable) as constant variables would have been of significance.

Incorrect Assumptions

Aside from the missing gaps of data, most results were assumed and consistent with reality. It should be mentioned that CDF seemed inaccurate as it showed an increase in natural growth rate over time.

Challenges

The biggest challenge was to come up with a valid result and persuasive argument from incomplete data. The chosen dataset contained relevant variables but were limited in time period. A study of this nature requires constant variables that cover significant time periods. As this was my first time doing EDA, there are many steps that I feel I did not fully understand. I hope that repetition will help me to better understand what each step is, why it's necessary, and to refine more effective ways of executing them. I had initially planned to use more than one dataset for this project; however, due to time constraints, I wasn't able to delve so far.

GIT link: https://github.com/rlawnsdnjs706/DSC530_term.git