DSC550 Term Project - Milestone One

Ross L. Kim-Schreck

Bellevue University

DSC550 Data Mining

Professor Werner

2024.03.01

Introduction

For this project, I took a deep dive into some datasets relating to the demographics of the Republic of Korea, which I will refer to as the 'ROK'. I could explore the culprits behind the demographic issues being experienced by the ROK. Despite being an impoverished and war-torn nation post Korean war era, South Korea is a flourishing democracy with the 13th largest economy in the world and third largest in Asia. It underwent a birth-rate spike in the 1970s and 1980s which had severe demographic implications: overcrowding; at the time, Seoul's infrastructure was not developed enough to accommodate such a high population especially in more densely-populated areas. The government, which was a military dictatorship, decided to rebuild and revive itself after the Korean war by taking an 'economy-first' initiative. While technically a democracy, the quasi dictatorship hand picked certain families to found conglomerates (Chaebols) which were then assigned specific industries in which they could maneuver, simply speaking, every South Korean conglomerate, to the present day, is constructed and functions like a monarchy inside this pseudo democracy – the ROK – each 'monarchy' has the CEO as king; this is nepotism and monarchism existing in a modern-day liberal democracy. In the 1980s, this dictatorship ordered its population to have no more than two children, much like China's one-child policy. Because of this, the rate of fertility has plummeted and is currently the lowest rate in the world. This will undoubtedly have an effect in all facets of society. This is the same byproduct of Japan's demographic issues in the late 1990s and two-thousands, and its economic issues today; furthermore, China, due to its one-child policy, is projected to follow the same path in a few decades to come. These three nations and their demographics have direct and massive implications on East Asia as a whole and continental Asia, which is roughly a fifth of the world's population. These three nations are the three largest economies in Asia.

All companies are being affected by the plummeting birth rate directly or indirectly. With labor shortages, companies will find it difficult to staff themselves. Companies that produce and sell baby products are being directly affected. Public schools are dwindling in enrollment and are being forced to consolidate as the younger population continues to rapidly decline, while millennials and older generations' populations are on a steadier trajectory.

After graphical analysis of the datasets that I have chosen, I can conclude that the birth rate spiked in the mid 1980s and has been declining rapidly until the present day. The birth rate is consistent with both the marriage rate and natural growth rate, so with further analysis, I could possibly find statistical significance and correlation.

After conducting analysis of the data, I discovered that despite having the lowest birth rate in the world, the overall population has been steadily increasing to the present day and is expected to continue this trajectory until 2026. South Korea is a hyper-developed democracy that is also incredibly capitalistic. Because of this development, the nation is able to accommodate such large populations in its urban and rural areas through the construction of countless large-scale high-rise apartment complexes centered in urban areas and scattered all around the countryside. Living space for its citizens is more than sufficient. However, it is projected that the population will steadily decrease from 2026 and by 2072, will have significantly dropped. This will be of catastrophic demographic consequences.

As mentioned above, with the birth rate decreasing, currently all industries are being affected in some capacity. If the overall population also begins to decline, this will have an irreversible effect on all facets of society and industries, thus negatively affecting the economy.

I think the best approach to pitching this problem would be to, assuming that my audience members are policy makers and heads of businesses, begin with detailed visuals

regarding the trends of birth rate and overall population growth from the past and detailed visuals regarding the future projections of both metrics.

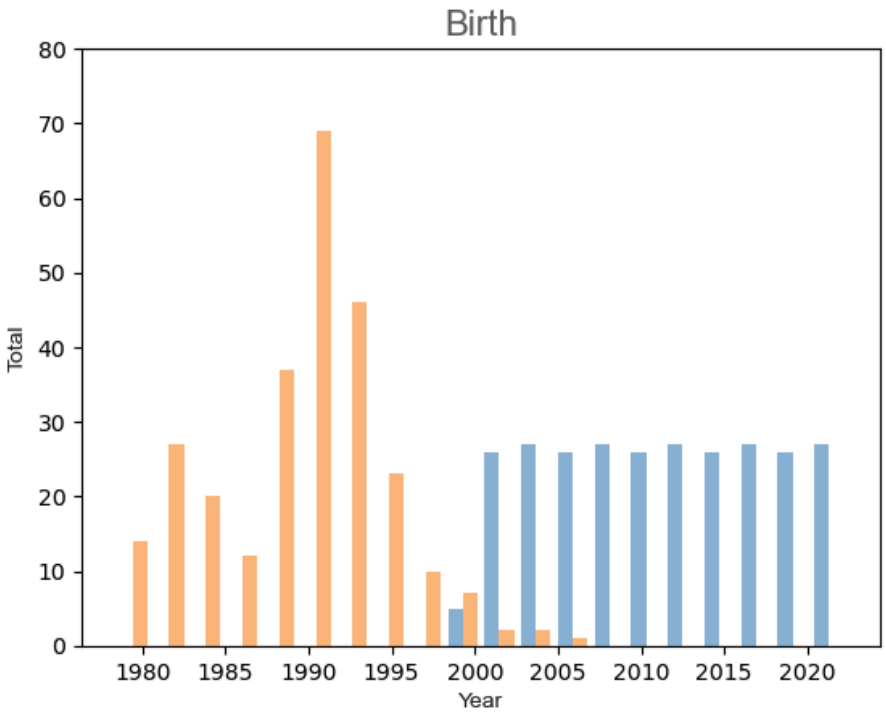
The data used to explore these issues are from various sources. These sources include the following: kosis.kr, which is a government database containing demographic data; datasets from Kaggle, which include population trends from 1970 to 2022; Korean demographics 2000–2022; ROK income and welfare; and Seoul real estate prices. All datasets used in this project can be found at the following link: https://github.com/rlawnsdnjs706/DSC550_term.

Summary of Milestones 1-3

Please note that any reference to a six-digit number refers to specific cells in the attached Jupyter notebook, eg. 00.00.00. The first digit refers to the milestone number, the second digit refers to the step, and the third digit refers to any parts of a step. For example, if the reference number is 03.02.12, then it would be milestone three, step two, and twelfth process of the step.

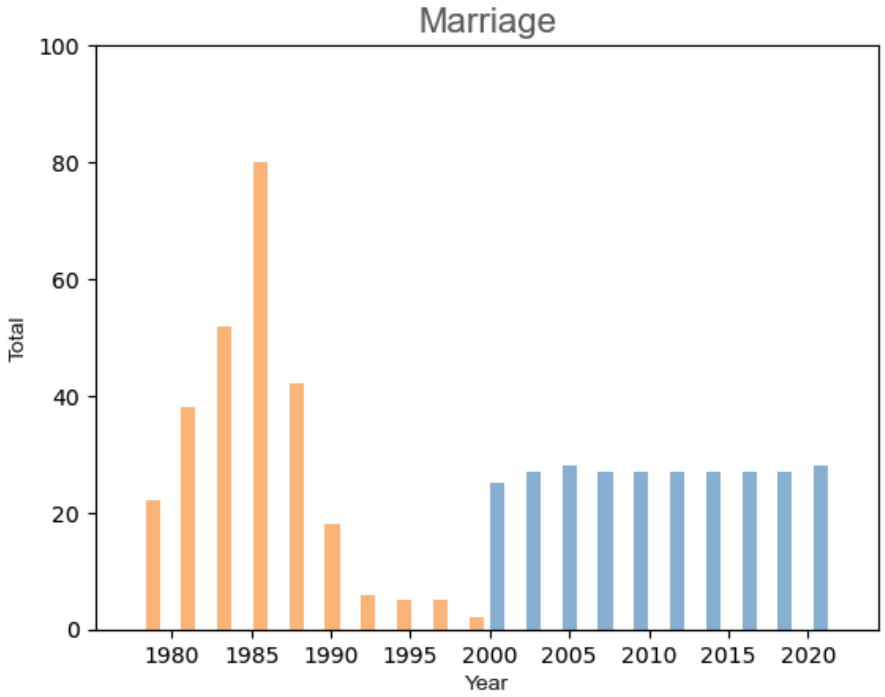
Milestone one is a pdf which was submitted in week six (DSC550 -WK06 -term -Kim -Schreck.pdf). It is a word file containing the criteria for milestone one.

Milestone two is composed of the chosen datasets transformed for visualization. (01.02.01) is the first dataset used (ROK_demographics_2000–2022). Fourteen visualizations were rendered using eight metrics. I was able to obtain birth rate, death rate, marriage rate, divorce rate, and natural growth rate data. For (01.02.03), I converted dates to international dates which are compatible with computer format. (01.03.01) is a selection of a specific region (Seoul) as nearly half of the overall population resides in the Seoul metropolitan area. The first ten visualizations are histograms using the basic plot function; the last four are histograms using the thinkstats2 package. The following are visualization that infer statistical significance:



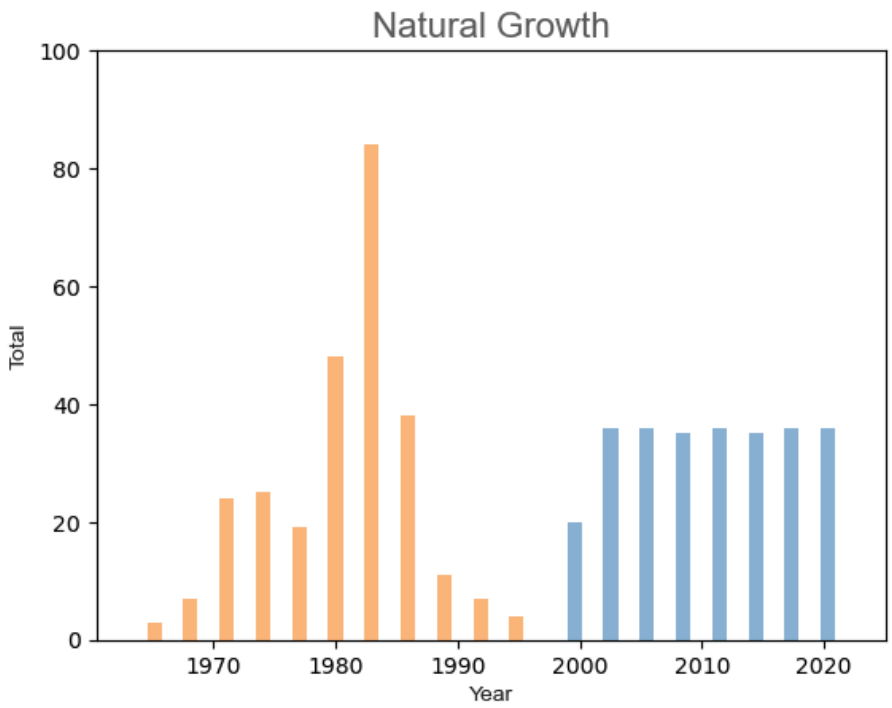
01.04.01

Histogram of Birth - total number of births in the year



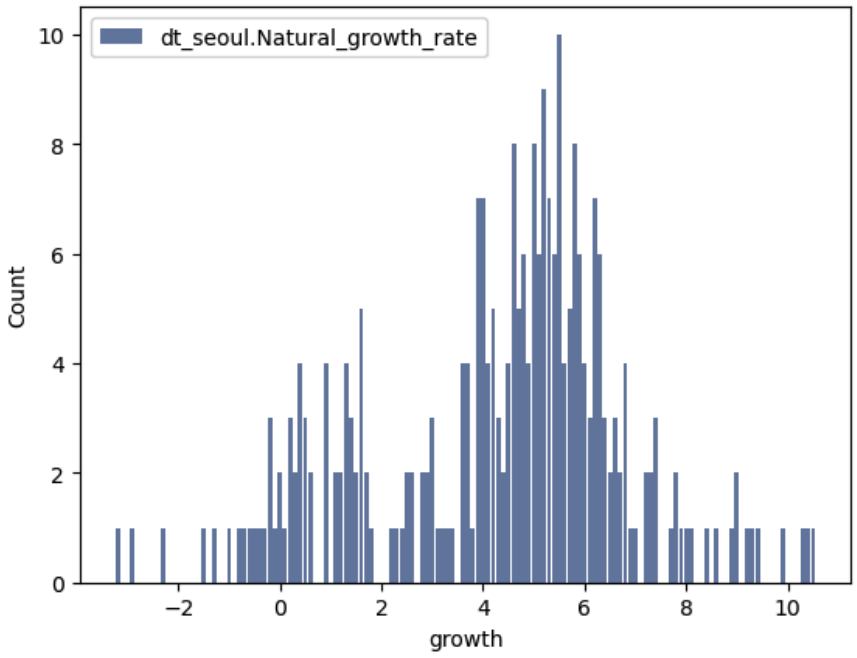
01.04.07

Histogram of Marriage - total number of marriages in the year



01.04.09

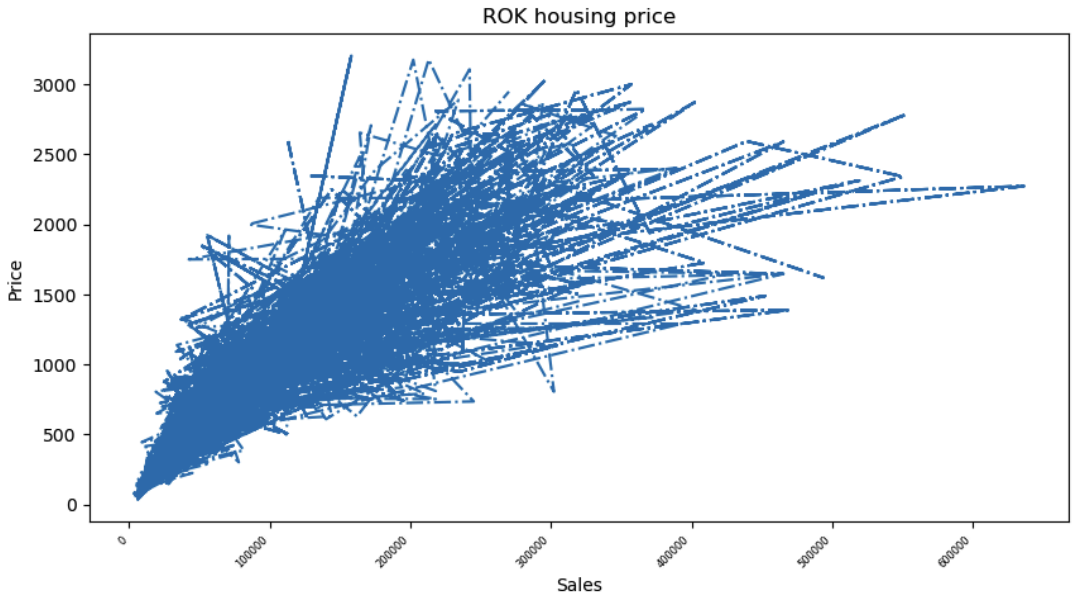
Histogram of Natural growth - total difference between births and deaths in the year



01.04.11

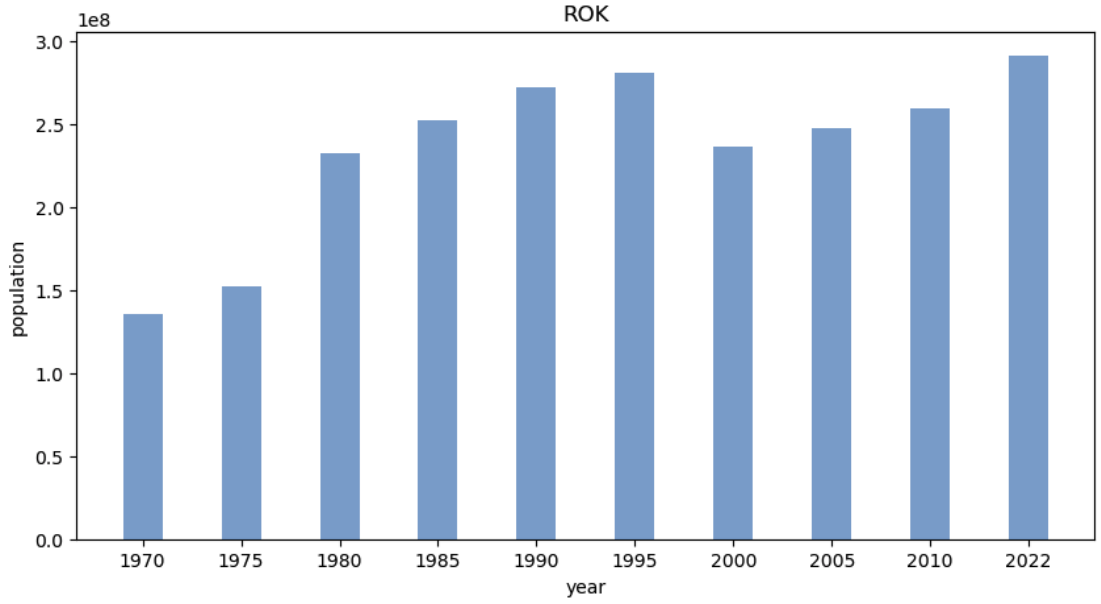
Histogram of Natural growth - total difference between births and deaths count

04.31.01



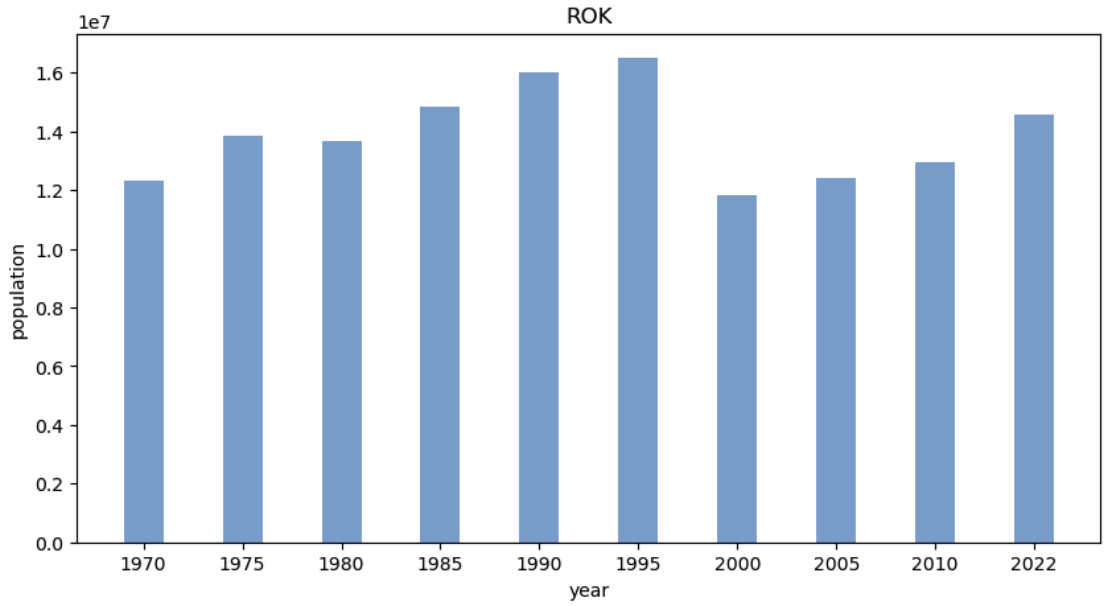
Plot of housing prices - real estate trends to date

04.32.02



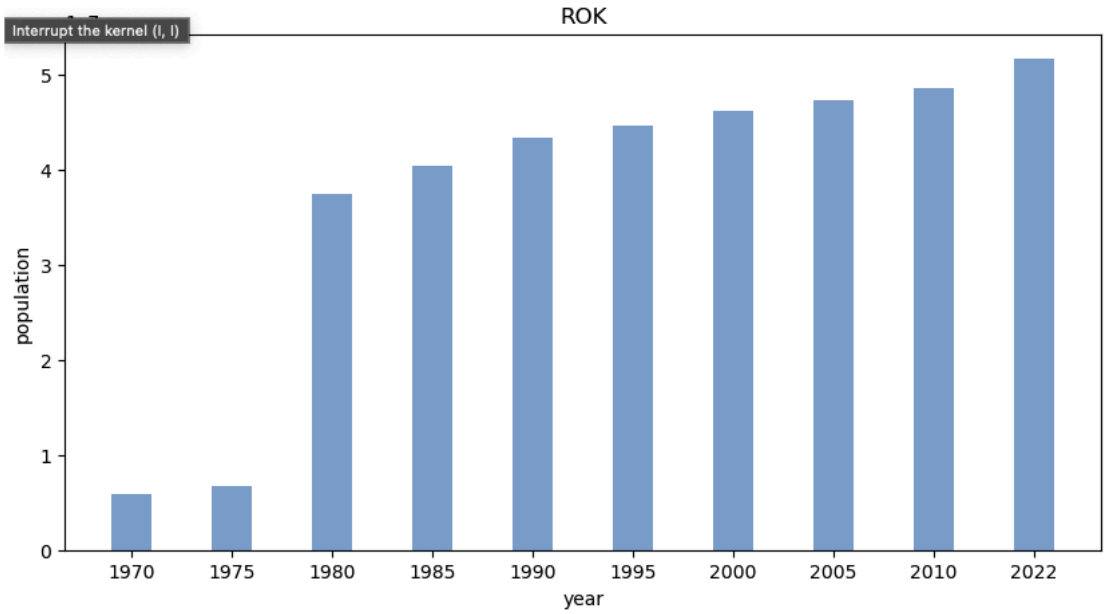
Bar graph of population - total populations by year

04.32.02

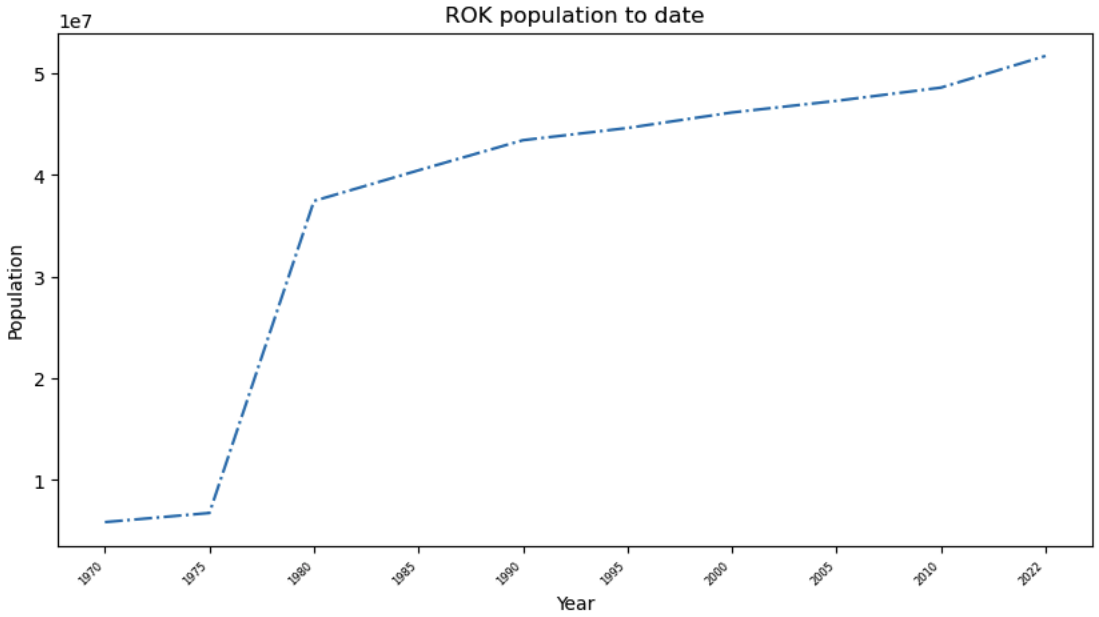


Bar graph of population - mean populations by year

04.34.02

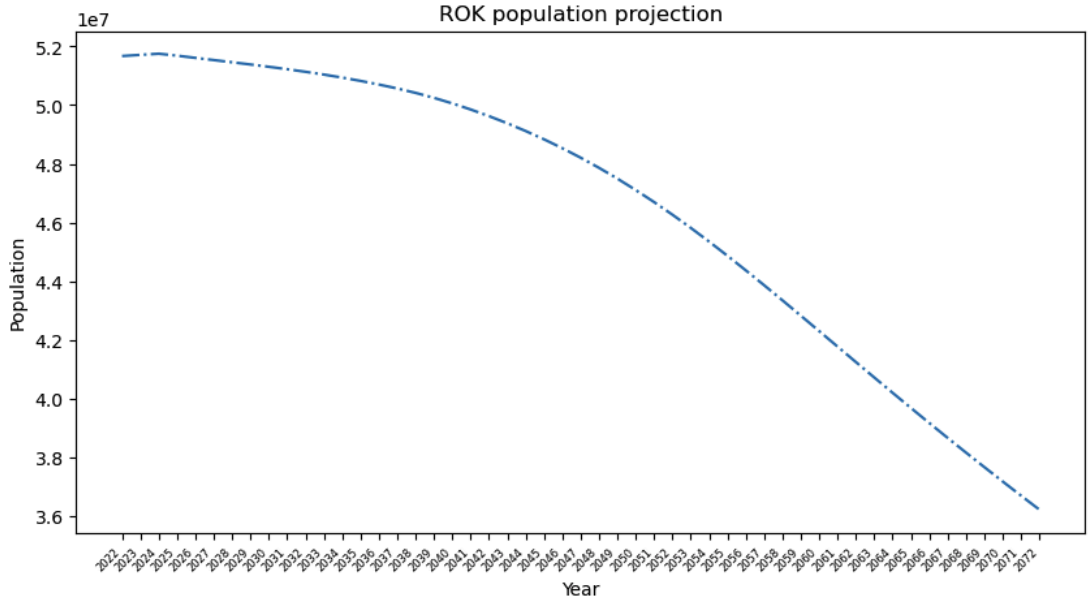


Bar graph of population - total populations by year



04.35.02

Line graph of population - total populations by year



04.36.01

Line graph of population - future projection of total population

In milestone one, (01.03.01) is the changed region of DT01 to Seoul as mentioned above. There are several noticeable visualizations that haven't statistical meaning without further data transformations. Additional transformations should be made before deployment.

The purpose of milestone two is to determine population trends and use various metrics to predict future trends. (02.01.02) is a renaming of columns for DT01 for better navigation. (02.02.01) is omitted categories that showed little statistical significance in milestone one: birth rate, death rate, divorce rate, marriage rate, and natural growth rate. In (02.03.03 -- 02.03.12), I added new datasets centered on population to supplement DT01; these are past population trends spanning from 1970 to 2022. The dimensions were confirmed for the new datasets in consideration of combining them in later steps (02.05.01–02.05.12). Nine datasets were combined into three (02.06.01–02.06.03) and NaN values were removed and replaced with zero. Entire columns were null after concatenating, so the logical solution was to replace them with zero values rather than the mean (02.08.02–02.08.06). I created dummy variables for the three combined datasets composed of the nine total datasets DT01–DT12 (02.08.09–02.08.12). This was done in preparation for milestone three.

In milestone three, I began by adding a new dataset DT02 - ROK income and welfare (03.02.01). I then indexed the column and chose a specific row from the new dataset (03.02.02–03.02.03). Columns were then renamed (03.02.05) for easier navigation. I renamed the total population column as this is the primary metric needed for this project (03.03.01–03.12.02). I initially mistakenly created dummy variables for the new dataset DT02 (03.13.01) as booleans. For the final, they were reproduced as integers. Dataset DT02 was split into training and test sets with the target being 'marriage' (03.14.01). The purpose of this step is to find correlation in marriage over time. Linear regression was conducted to return RMSE and

R^2; they indicated higher correlation than the previous model (03.14.02–03.14.07). The same linear regression model was then run with the targeting being ‘education level’ (03.16.01–03.16.07). It returned a relatively low correlation.

For data preparation in milestone three, I created a function for detection of duplicates (03.18.01). This function was applied to all datasets DT01–DT12 (03.19.01–03.19.12). I created dummy variables as a test to DT03 (2022 population) (03.20.01) and split the dataset DT03 into training and test sets with the target being ‘total population’ (03.21.01 -- 03.21.02). A linear regression model was run on DT03 (03.21.03–03.21.07), but numeric data seems to have been removed in the dummy dataset, thus RMSE and R^2 return is of no significance. I took a different approach in the final milestone and revised linear regression in milestone three by converting boolean dummy variables into integers.

There are obvious gaps to be filled in the data in order to return any meaningful insights. In the final, several modifications were made to the datasets in order to achieve this.

Conclusion

In the final, I begin by confirming the column names of all datasets DT01–DT12 (04.01.01 -- 04.01.12). DT13 and DT14 are also added (real estate prices and population projections) (04.01.13 -- 04.01.14). Specific columns and rows were chosen and indices were added to each dataset to simplify for concatenation (04.02.01, 04.03.01, 04.04.01, 04.05.01, 04.06.01, 04.07.01, 04.08.01, 04.09.01, 04.10.01, 04.11.01, 04.12.01, 04.13.01, 04.14.01). Variables were assigned to the selected columns by selecting corresponding indices and renaming them accordingly (04.02.03, 04.03.03, 04.04.03, 04.05.03, 04.06.03, 04.07.01, 04.08.03, 04.09.03, 04.10.03, 04.11.03, 04.12.03, 04.13.02, 04.14.02). All strings were converted

to floats in all datasets and the dates' formats were changed to international format (04.15.01–04.15.12). I used the variables to calculate the total population of each year (04.17.03–04.17.12) and then assigned each total with a new variable. I used the variables to calculate the mean of each year (04.18.03–04.18.12) and then assigned each mean a new variable. I combined the variables for total populations of each year into a new dataset (04.21.01) in order to plot a visualization (04.21.02). It seems that the columns are inconsistent in value, and there are many outliers. The same steps were taken for the mean (04.22.01–04.22.02) returning similar results. Total populations were plotted for each year (DT03–DT12) (04.21.01–04.30.02). Housing prices were plotted with x and y representing price and sales (04.31.01). This was reflective of the reality that people generally do not buy expensive housing. Total population was plotted by year based on the combined datasets (04.32.02–04.35.02). I created a new dataframe with population totals combined from nine different datasets (04.35.01). I created dummy variables with integers and ran a linear regression model with the target being the year 2022 (04.36.03–04.37.03). It returned an RMSE of 0.82 and R^2 of 0.0, suggesting that with a range of 0 to 1, RMSE is not small. It is also assumed that R^2 is zero because y values are the same for predicted and test. The new dataframe with population totals combined from nine different datasets returned statistically significant visualizations. For the final visualization, I plotted future population projections (2022–2072) (04.39.01).

Based on the models and visualizations, birth rates are correlated with population growth. As birth rates plummet, the population is projected to steadily decrease over time. At this rate, the population will halve by 2100. I should note that despite, for a short period, the birth rate decreases while the population continues an increase; this is due to the natural growth rate, which is likely a byproduct of the advancements in science and healthcare. People simply live longer in

developed nations. Finding correlation between population growth and housing prices was not possible due to the lack of data. This should be explored further for deployment. Also marriage, divorce, birth, death, and growth rates should be projected and compared to total populations and real estate price projections. This was not possible due to missing data, thus the model is not ready to be deployed.

It is a fact, based on my insights, that the ROK will face major demographic challenges. I would recommend making adjustments to the economy to make real estate more affordable and offer perks and incentives to newly weds and couples giving birth to multiple children. I would also recommend offering higher salaries and additional benefits to migrant workers.

If this problem is not addressed, and policies are not implemented within decades, the ROK's economy will likely take the same trajectory as that of Japan. We should consider the Japanese economic situation at the time that their population began to decline and consider measures taken by the Japanese government to counter this. We should consider the pros and cons of Japanese policies at the time. The next domino to fall will be China, and if the three East-Asian nations, which happen to be the three largest economies in Asia, cannot find a solution, it will have massive implications on the world economy.