DSC540 Term Project - Milestone One

Ross L. Kim-Schreck

Bellevue University

DSC540 Data Preparation

Professor Williams

2024.03.01

Sources

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/DSC540_term.

Summary

Please note that any reference to a six-digit number refers to specific cells in the code, 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, twelfth process of the step. This research is regarding the Republic of Korea which I will refer to as the ROK throughout this report.

Milestone one is a pdf which was submitted in week six (DSC540_ WK04_ term_ Schreck-Kim.pdf). It is a word file containing criteria for milestone one.

Milestone two is composed of the chosen datasets transformed for visualization. The first dataset used is ROK_demographics_2000–2022 (02.01.01) (DT01). I renamed each column for easier navigation. (02.02.03) I converted dates to international dates which are compatible with computer format. (02.03.01–02.03.04) I conducted fuzzy matching in order to make romanized names more consistent. (02.04.01) I selected a specific region (Seoul) as nearly half of the overall population resided in the Seoul metropolitan area. For the first visualization, I used histograms using the thinkstats2 package (02.05.01). I finished milestone two by detecting duplicates (02.05.02–02.05.06).

In milestone three, I began by adding a new dataset DT02 - ROK income and welfare (03.02.01). I indexed the column and chose a specific row from the new dataset (03.02.02–03.02.03). I then renamed the columns (03.02.05) for easier navigation. I plotted demographic data and income (03.04.01–03.04.04). I modified DT02 by creating an index, removing duplicates, and omitting unnecessary columns (03.05.02–03.06.01).

The purpose of milestone four, I added new datasets centered on population to supplement DT01 (04.01.03–04.01.12); these are past population trends spanning 1970–2022. I confirmed the dimensions of the new datasets for consideration of combining in later steps (04.02.01–04.02.12). I imported the API from Kosis, which is a governmental database containing demographic data (0413.01–04.13.09). I went through data cleaning steps taken in previous milestones to further refine the transformations (04.14.02–04.18.01). For data preparation in milestone five, I created a function for detection of duplicates (04.19.01). I applied this function to all datasets DT01–DT12 (04.20.01–03.20.12).

For milestone five, I added a json file as DT13 (05.01.01), DT14 (population projections) (05.01.02), and DT15 (real estate prices) (05.01.03). I chose specific columns and rows and added indices to each dataset to simplify for concatenation (05.02.01, 05.03.01, 05.04.01, 05.05.01, 05.06.01, 05.07.01, 05.08.01, 05.09.01, 05.10.01, 05.11.01, 05.12.01, 05.13.01, 05.14.01). I assigned variables to the selected column by selecting indices and renaming accordingly (05.02.02, 05.03.03, 05.04.03, 05.05.03, 05.06.03, 05.07.01, 05.08.03, 05.09.03, 05.10.03, 05.11.03, 05.12.03). I converted all strings to float in all datasets and changed the date format to international (05.16.01–05.16.12). I used the variables to calculate the total population of each year (05.18.03–05.18.12) and then assigned each total a new variable. I used the variables to calculate the mean of each year (05.19.03–05.19.12) and then assigned each mean a

new variable. I combined the variables for total populations of each year into an SQL database (05.19.12–05.20.07). I plotted the total population by year based on the combined datasets (05.20.10–05.20.11) (05.34.02–05.37.02). This returned statistically significant visualizations. I plotted future population projections (2022–2072) (05.38.01).

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 would be unethical to manipulate and transform this type of data without careful consideration of demographics. An example of this would be simply replacing non values with the mean or mode. This can render the result inaccurate due to the lack of considerations of populations as byproducts of regional terrains, population densities due to lack of infrastructure (in the case of under-developed nations), and lack of careful analysis of metrics across various data sources. The metrics are seemingly countless with this type of research.