Milestone Five

Ross L. Kim-Schreck

DSC630 Predictive Analytics

Professor Hua

2024.06.01

Exploring Policies to Stabilize Housing Prices in the ROK

Introduction

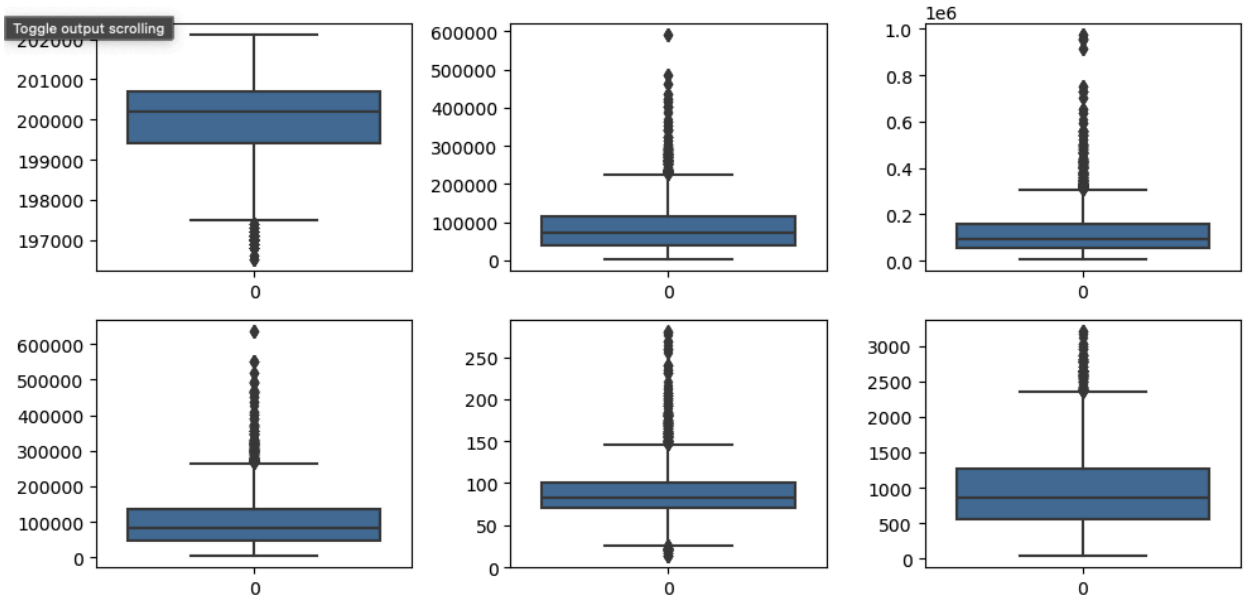
In the Republic of Korea, which I will refer to as ‘the ROK’ or ‘South Korea’ respectively, the real-estate market has been on the rise since the Korean War. Statistics have shown that housing prices have increased consistently while housing construction has been increasing at an even faster pace. Based on an article published in 2022, vacant housing has increased 40% since 2015 and roughly eight percent, which is 1.51 million houses remaining unoccupied. In this research, I set out to unveil demographic trends and how they coincide, revealing what implications, if any, there could be in the near and distant future.

Data Selection

In my second milestone, I chose two datasets relating to housing prices and other real-estate market trends, which I realized were inadequate for the research I needed to conduct. Thus, since then I have added multiple datasets relating to various demographics. The most significant datasets among the datasets used contain population data for the ROK since the 1970s. I acquired the majority of these datasets from the Korean government’s official website: KOSIS, which is an acronym for Korean Statistical Information Service. Much of the data is time series and covers periods from the Korean and Cold wars until the present day.

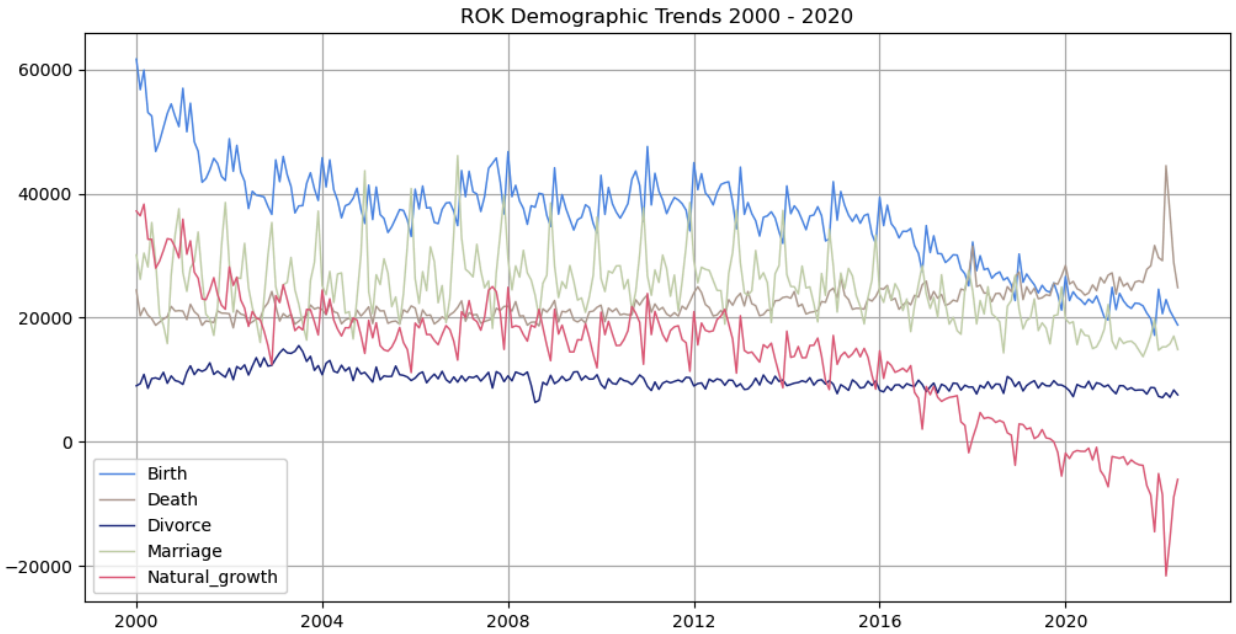
Methods and Results

I used linear-regression and logistic-regression models on a number of datasets. The first model I ran was a linear-regression model on a dataset containing housing prices over a period from 1970 until the present day. It returned a very high RMSE and low R2; this is due to the several outliers within the data.



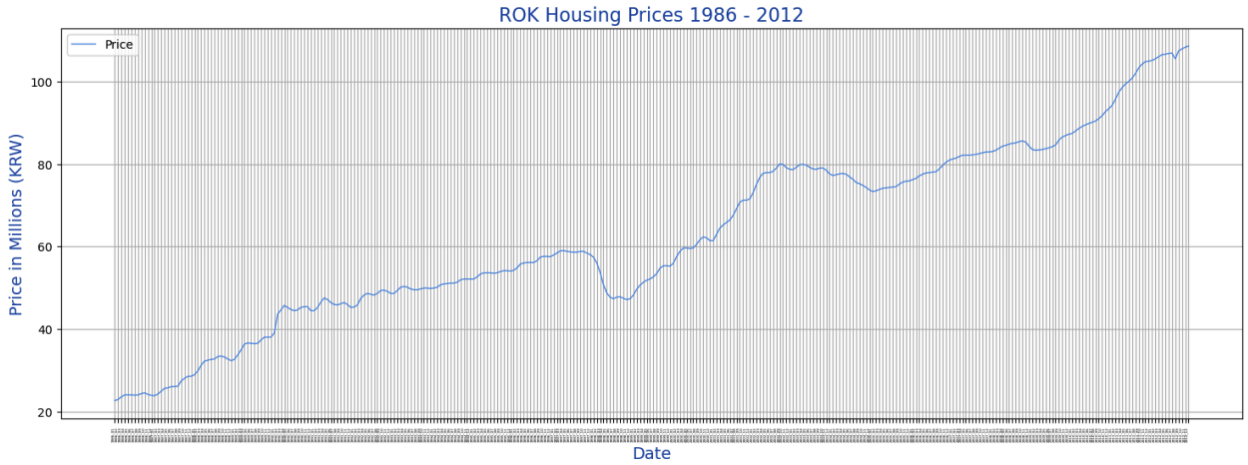
In the above graphs, build date, minimum sales, maximum sales, average sales, and price per area are all visualized. There are a significant number of outliers due to the contrast in land value between metropolitan and rural areas. Price per area in Seoul can be ten times higher than in rural areas.

I then focused on a dataset containing the following: birth rate, marriage rate, death rate, and natural-growth rate. I rendered a multivariate plot to compare each variable of this dataset:

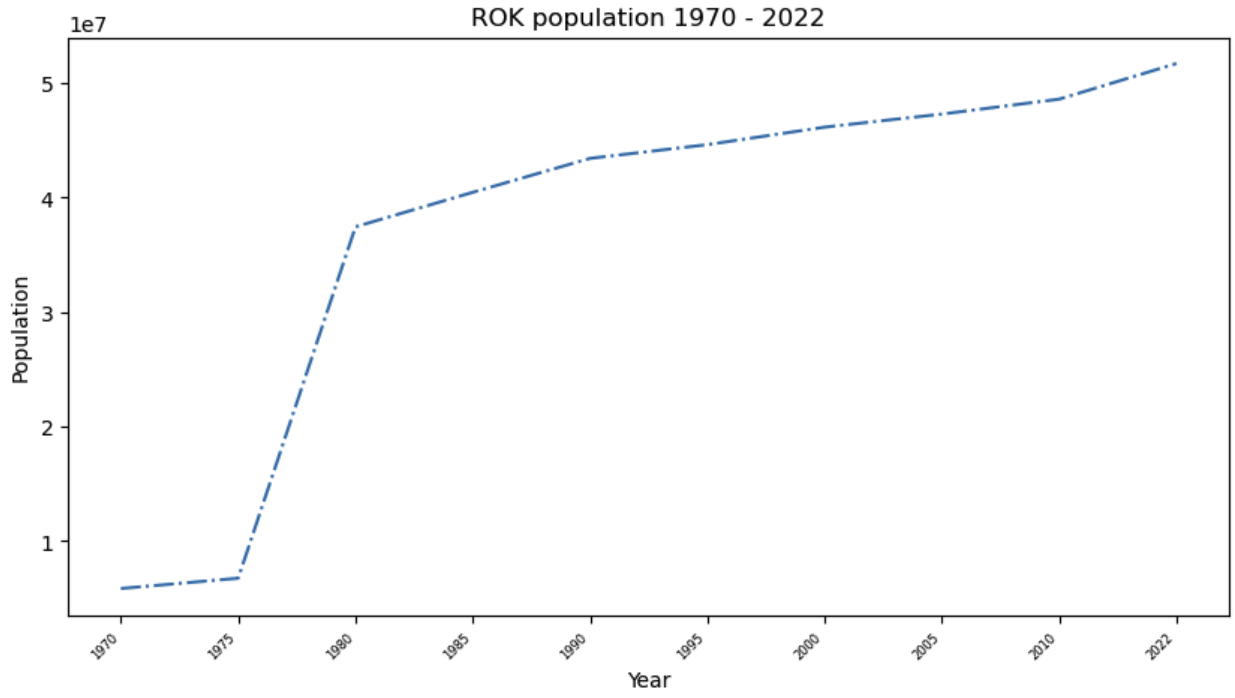


As you can see, the birth and natural-growth rates have declined significantly when compared to other variables. I should mention that natural-growth rate refers to the rate of births and deaths over time; the fewer of both results in an overall lower natural-growth rate.

I then ran a linear-regression model on a dataset containing only housing price data. Once again, RMSE was quite high, while R2 was low. The further back in time that we explore shows a higher contrast in prices between rural and urban areas. This also applies to the present day but to a lesser extent. According to the below graph, housing prices have steadily risen from 1986 to 2012:

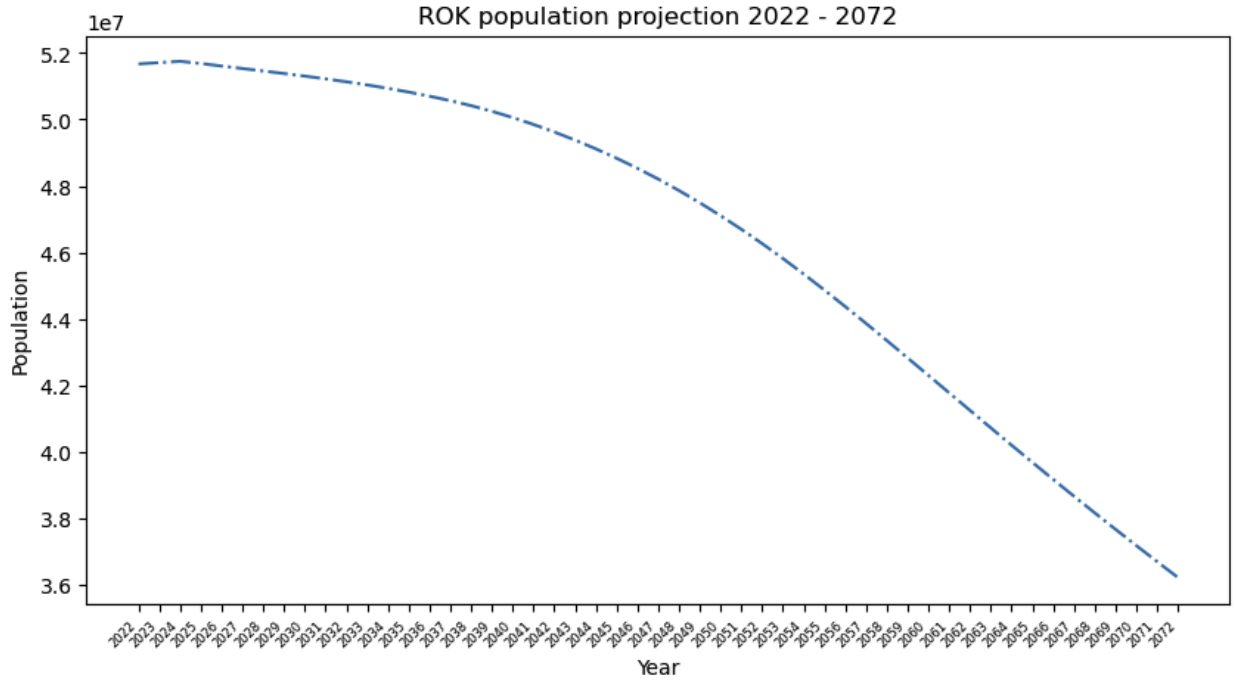


I compared historical data containing population trends with future trends in the following graph:



Despite the birth rate declining, the population has continued to increase since 1970 from under 10 million to over 50 million in 2022. This is due to the natural-growth rate, which as mentioned before, is the calculated difference between births and deaths. According to the graph below, the

overall population is predicted to level off in 2024 at over 50 million and then steadily decrease until at least 2072:



I ran a linear-regression model on the above dataset to conclude a significantly lower RMSE compared to other models' results but is still high, and an ideal R2 score of 1.0 indicating higher predictive accuracy compared to other models.

Conclusion

In conclusion, it is fairly obvious that South Korea is on the path to a demographic catastrophe. With the rise of vacant housing, especially in rural areas, urbanization, and the expected decline in population and natural-growth rate, in the near and distant future, South Korea will experience a real-estate crisis among many other crises.

I think a way to overcome this inevitable consequence would be to stop constructing houses and divert funds to channels that will assist married couples economically in order to encourage progeny, encourage marriage by providing perks to newly weds, and lowering housing prices to assist couples economically. Also importing migrant workers, both skilled and unskilled, to equally distribute the population throughout the entire peninsula would help to solve the regional-disparity issues and increase the birth and natural-growth rates. Over half of the total population lives in the Seoul-metropolitan area. This is the culprit to the disparity in land values from region to region. There are a number of new policies being proposed that will have a tremendous impact on the marriage rates, birth rates, and housing prices among other variables affecting the native population. Unfortunately, the politics of the ROK is multipolar. The particular policies cannot be anticipated as each party has their own agendas; therefore, it would be difficult to predict future housing prices, as they depend on a number of variables relating to future policies.

References

1. https://www.macrotrends.net/global-metrics/countries/KOR/south-korea/gdp-gross-domestic-product
2. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/
3. https://www.kaggle.com/datasets/jcy1996/seoul-real-estate-datasets

4. https://data.seoul.go.kr/dataList/801/S/2/datasetView.do
5. https://kosis.kr/statHtml/
6. https://www.koreatimes.co.kr/www/tech/2024/05/129_330929.html#:~:text=Data%20provided%20by%20the%20three,highest%20rate%20of%2013.8%20percent.
7. https://www.koreaherald.com/view.php?ud=20240214050617

Milestone Four

Ross L. Kim-Schreck

DSC630 Predictive Analytics

Professor Hua

2024.05.12

Exploring Policies to Stabilize Housing Prices in the ROK

Data Preparation

30 datasets were added that are all relevant and directly affect the subject:

1. dt01_real_estate_pr_01
    a. dt01
2. dt02_real_estate_se_01
    a. dt02
3. dt03_stats_vital_kr_01
    a. pd.read_csv('한국의_생명통계_20240427103929.csv')
4. dt04_housing_census_01
    a. pd.read_csv('요약__of_인구조사_주택별_행정_지구_별_종류_집_별_주거_가구수별__20240427102536.csv')
5. dt05_housing_constr_01
    a. pd.read_csv('착공_주택_건축_월_총계__20240427104335.csv')
6. dt06_household_empt_02
    a. pd.read_ csv ('미분양_주택__ 총계__ 2024 0427102612.csv')
7. dt07_housing_loss___01
    a. pd.read_csv ('주택_손실_20240427102442.csv')
8. dt08_housing_use_sv_01
    a. pd.read_csv('주택_이용_조사__월_총__20240427104424.csv')
9. dt09_jeonse_sales___01
    a. pd.read_csv('공급및_수요_판매_시장_별_지역_20240427102712.csv')

10. dt10_jeonse_market__02
    a. pd.read_csv('공급수요_of_전세_시장_별_지역_20240427102647.csv')
11. dt11_birthrate_age__01
    a. pd.read_csv('연령별_출산율_비율_한국__20240427103626.csv')
12. dt12_birthrate_age__02
    a. pd.read_csv('예상_인구_별_연령_그룹_한국__20240427103530.csv')
13. dt13_pop_houshold___01
    a. pd.read_csv('인구__가구_및_주택_단위_20240427102858.csv')
14. dt14_pop_houshold___02
    a. pd.read_csv('인구__가구_및_주택_단위_20240427103017.csv')
15. dt15_pop_houshold___03
    a. pd.read_csv('인구__가구_별_행정_지구_20240427103320.csv')
16. dt16_pop_houshold___04
    a. pd.read_csv('인구__가구_별_행정_구역_20240427103114.csv')
17. dt17_pop_houshold___05
    a. pd.read_csv('인구__가구_별_행정_구역_20240427103142.csv')
18. dt18_pop_houshold___06
    a. pd.read_csv('인구__가구_별_행정_지구_20240427103208.csv')
19. dt19_pop_houshold___07
    a. pd.read_csv('인구__가구_별_행정_지구_20240427103233.csv')

20. dt20_pop_houshold___08
    a. pd.read_csv('인구__가구_별_행정_지구_20240427103256.csv')
21. dt21_pop_houshold___09
    a. pd.read_csv('인구가구및주택단위별_행정_지구_20240427103044.csv')
22. dt22_pop_estimate___01
    a. pd.read_csv('예상_인구_별_연령_한국__20240427103433.csv')
23. dt23_pop_future_pro_01
    a. pd.read_csv('결정요인_미래_인구_변화_한국__20240427103555.csv')
24. dt24_pop_future_pro_02
    a. pd.read_csv('인구예측및요약지표별시나리오한국_20240427103708.csv')
25. dt25_pop_future_pro_02
    a. pd.read_csv('인구_추계_및_요약_지표_한국__20240427103505.csv')
26. dt26_salaries_entir_01
    a. pd.read_csv('임금근로자의_개인대출_20240427104104.csv')
27. dt27_income_job_typ_01
    a. pd.read_csv('직업유형_별_지역_및_소득_수준_20240427102358.csv')
28. dt28_employment_eff_01

   a. pd.read_csv('기본취업_이행현황_비율이후201020240427103804.csv')

29. dt29_income_median__01
   a. pd.read_csv('평균소득_중위소득_소득분배_20240427104029.csv')
30. dt30_ROK_demo_____01
   a. pd.read_csv('ROK_income_welfare.csv')

The first step in organizing the data was to translate using variable names in English. I then assigned the variables with names of all the same length to make it easier to edit the datasets in bulk. The variable names reflect the content of the datasets. I then changed the column names of all datasets so that it would be convenient to make transformations. I used a system of underscores so that a column name and dataset's variable can be double clicked for easy selection: eg:

1. Dt19_pop_houshold___07.rename
   a. (columns = {'1980.10':'1980_10'}, inplace = True)
2. Dt03_stats_vital_kr_01
   a. pd .read _csv ('한국의_생명통계_20240427103929.csv')

I determined the data type by plotting each dataset. I changed strings to floats where necessary.

The following are the steps taken for data preparation:

1. Translated CSV names using English for variable names.
   a.

```
# 04.02.01-01
# rename dt01, dt02
# read csv (27 added) dt03 - dt29
# relabeled datasets in english

dt01_real_estate_pr_01 = dt01
dt02_real_estate_se_01 = dt02
dt03_stats_vital_kr_01 = pd.read_csv('한국의_생명통계_20240427103929.csv')
dt04_housing_census_01 = pd.read_csv('요약__of_인구조사_주택별_행정_지구_별_종류_집_별_주거_가구수별__20240427102536.csv')
dt05_housing_constr_01 = pd.read_csv('착공_주택_건축_월_총계__20240427104335.csv')
dt06_household_empt_02 = pd.read_csv('미분양_주택__총계__20240427102612.csv')
dt07_housing_loss___01 = pd.read_csv('주택_손실_20240427102442.csv')
dt08_housing_use_sv_01 = pd.read_csv('주택_이용_조사__월_총__20240427104424.csv')
dt09_jeonse_sales___01 = pd.read_csv('공급__및_수요_판매_시장_별_지역_20240427102712.csv')
dt10_jeonse_market__02 = pd.read_csv('공급_수요_of_전세_시장_별_지역_20240427102647.csv')
dt11_birthrate_age__01 = pd.read_csv('연령별_출산율_비율_한국__20240427103626.csv')
dt12_birthrate_age__02 = pd.read_csv('예상_인구_별_연령_그룹_한국__20240427103530.csv')
dt13_pop_houshold___01 = pd.read_csv('인구__가구_및_주택_단위_20240427102858.csv')
dt14_pop_houshold___02 = pd.read_csv('인구__가구_및_주택_단위_20240427103017.csv')
dt15_pop_houshold___03 = pd.read_csv('인구__가구_별_행정_지구_20240427103320.csv')
dt16_pop_houshold___04 = pd.read_csv('인구__가구_별_행정_구역_20240427103114.csv')
dt17_pop_houshold___05 = pd.read_csv('인구__가구_별_행정_구역_20240427103142.csv')
dt18_pop_houshold___06 = pd.read_csv('인구__가구_별_행정_지구_20240427103208.csv')
dt19_pop_houshold___07 = pd.read_csv('인구__가구_별_행정_지구_20240427103233.csv')
dt20_pop_houshold___08 = pd.read_csv('인구__가구_별_행정_지구_20240427103256.csv')
dt21_pop_houshold___09 = pd.read_csv('인구__가구_및_주택_단위_별_행정_지구_20240427103044.csv')
dt22_pop_estimate___01 = pd.read_csv('예상_인구_별_연령_한국__20240427103433.csv')
dt23_pop_future_pro_01 = pd.read_csv('결정요인_미래_인구_변화_한국__20240427103555.csv')
dt24_pop_future_pro_02 = pd.read_csv('인구_예측_및_요약_지표_별_시나리오_한국__20240427103708.csv')
dt25_pop_future_pro_02 = pd.read_csv('인구_추계_및_요약_지표_한국__20240427103505.csv')
dt26_salaries_entir_01 = pd.read_csv('임금근로자의_개인대출_20240427104104.csv')
dt27_income_job_typ_01 = pd.read_csv('직업_유형_별_지역_및_소득_수준_20240427102358.csv')
dt28_employment_eff_01 = pd.read_csv('기본취업_이행현황_비율__이후_2010__20240427103804.csv')
dt29_income_median__01 = pd.read_csv('평균소득_중위소득_소득분배_20240427104029.csv')
dt30_ROK_demo_______01 = pd.read_csv('ROK_income_welfare.csv')
```

2. Changed column names for consistency and convenience.

a.

```
# 04.02.01-02-01
# change column names
# return new column names
# dt01

dt01_real_estate_pr_01.rename(columns = {'자치구별(1)':'region_01'}, inplace = True)
dt01_real_estate_pr_01.rename(columns = {'자치구별(2)':'region_02'}, inplace = True)
dt01_real_estate_pr_01.rename(columns = {'2023. 10':'2023_10_00'}, inplace = True)
dt01_real_estate_pr_01.rename(columns = {'2023. 10.1':'2023_10_01'}, inplace = True)
dt01_real_estate_pr_01.rename(columns = {'2023. 11':'2023_11_00'}, inplace = True)
dt01_real_estate_pr_01.rename(columns = {'2023. 11.1':'2023_11_01'}, inplace = True)
dt01_real_estate_pr_01.rename(columns = {'2023. 12':'2023_12_00'}, inplace = True)
dt01_real_estate_pr_01.rename(columns = {'2023. 12.1':'2023_12_01'}, inplace = True)
print('dt01 columns:',dt01_real_estate_pr_01.columns)
```

```
dt01 columns: Index(['region_01', 'region_02', '2023_10_00', '2023_10_01', '2023_11_00',
       '2023_11_01', '2023_12_00', '2023_12_01'],
      dtype='object')
```

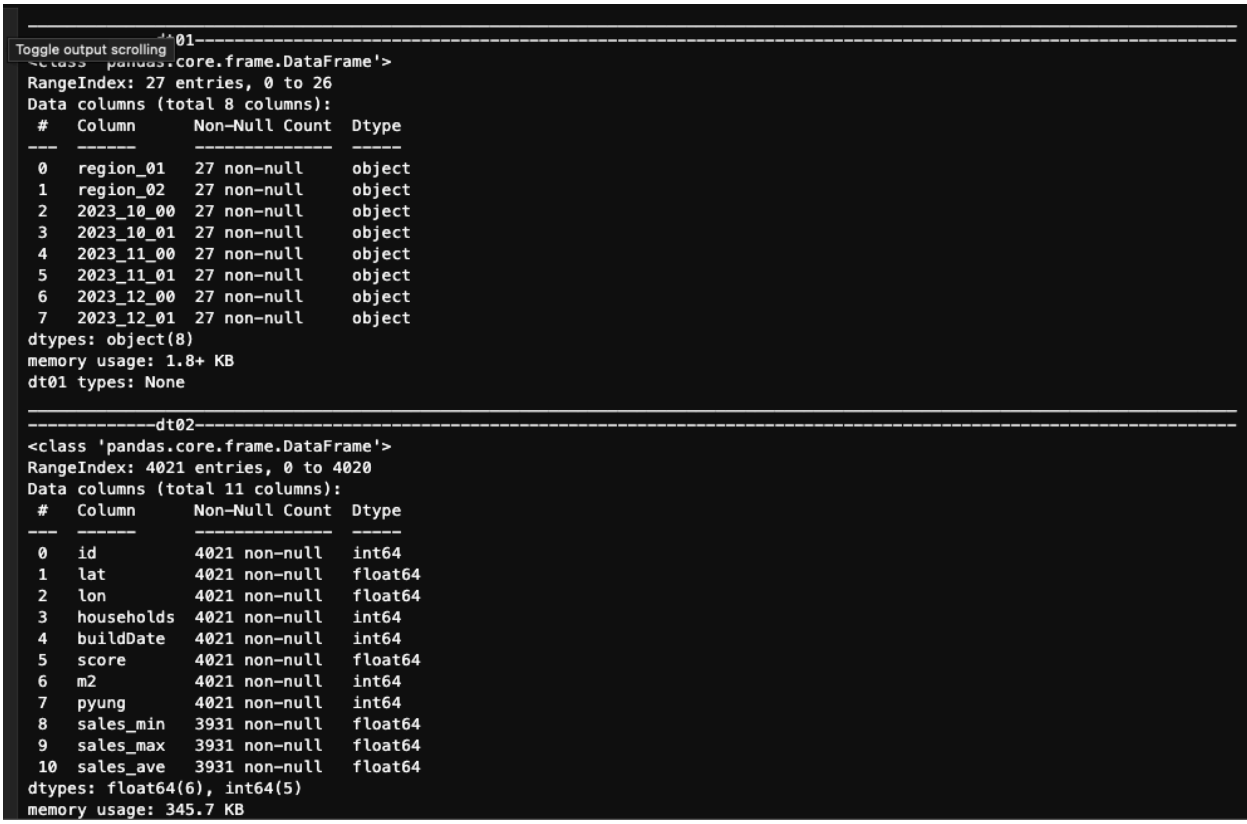
```
# 04.02.01-02-02
# change column names
# return new column names
# dt02

dt02_real_estate_se_01.rename(columns = {'id':'id'}, inplace = True)
dt02_real_estate_se_01.rename(columns = {'lat':'lat'}, inplace = True)
dt02_real_estate_se_01.rename(columns = {'lng':'lon'}, inplace = True)
dt02_real_estate_se_01.rename(columns = {'households':'households'}, inplace = True)
dt02_real_estate_se_01.rename(columns = {'build_date':'date_build'}, inplace = True)
dt02_real_estate_se_01.rename(columns = {'score':'score'}, inplace = True)
dt02_real_estate_se_01.rename(columns = {'m2':'m2'}, inplace = True)
dt02_real_estate_se_01.rename(columns = {'p':'pyung'}, inplace = True)
dt02_real_estate_se_01.rename(columns = {'min_sales':'sales_min'}, inplace = True)
dt02_real_estate_se_01.rename(columns = {'max_sales':'sales_max'}, inplace = True)
dt02_real_estate_se_01.rename(columns = {'avg_sales':'sales_ave'}, inplace = True)
print('dt02 columns:',dt02_real_estate_se_01.columns)
```

```
dt02 columns: Index(['id', 'lat', 'lon', 'households', 'buildDate', 'score', 'm2', 'pyung',
       'sales_min', 'sales_max', 'sales_ave'],
      dtype='object')
```

3. Determined data types and made proper adjustments.

a.



4. Split data into training and test sets.

a.

```
# 04.02.01-02-30-03
# focus on marriage rate as it correlates with housing
# split data
# target = marriage
# marriage = variable 01
# select columns
# for dataset dt30 containing income-level data
# dt30

dt30_x01 = dt30_dv.drop(['marriage'], axis=1)
dt30_y01 = dt30_dv['marriage']
```

```
# 04.02.01-02-30-04
# focus on marriage rate as it correlates with housing
# target = marriage
# marriage = variable 01
# split into train and test
# for dataset dt30 containing income-level data
# dt30

dt30_x01_trn, dt30_x01_tst, dt30_y01_trn, dt30_y01_tst = train_test_split(dt30_x01, dt30_y01, test_size=0.3, random_state=0)
```

b.

```
# 04.02.01-02-30-13
# focus on education level as it correlates with housing
# target = education level
# education level = variable 02
# select columns
# for dataset dt30 containing education-level data
# dt30

dt30_x02 = dt30_dv.drop(['education_level'], axis=1)
dt30_y02 = dt30_dv['education_level']
```

```
# 04.02.01-02-30-14
# focus on education level as it correlates with housing
# target = education level
# education level = variable 02
# split into train and test
# for dataset dt30 containing education-level data
# dt30

dt30_x02_trn, dt30_x02_tst, dt30_y02_trn, dt30_y02_tst = train_test_split(dt30_x02, dt30_y02, test_size=0.3, random_state=0)
```

Model

I used linear regression models on dataset 30. I used two target variables for the models: marriage rate and income. I chose these variables because it is likely that they both directly affect the subject the greatest.

1. The first model for marriage rate returned the following:
    a. RMSE of 0.8258674228682785 which is acceptable.
    b. R2 of 0.4301953970153479 which is slightly low.

i.

```
# 04.02.01-02-30-09
# focus on marriage rate as it correlates with housing
# target = marriage
# marriage = variable 01
# return rmse and r2 dt04
# rmse: 0.8258674228682785
# r2: 0.4301953970153479
# for dataset dt30 containing income-level data
# dt30

print(f'rmse: {dt30_rmse01}')
print(f'r2: {dt30_r201}')
```

```
rmse: 0.8258674228682678
r2: 0.4301953970153628
```

2. In the second model for income which returned the following:
    a. RMSE of 1.0642771383355127 which is acceptable
    b. R2 of 0.5929125769657074 which is slightly higher than the prior result.

i.

```
# 04.02.01-02-30-19
# focus on education level as it correlates with housing
# target = education level
# education level = variable 02
# return rmse and r2 dt04
# rmse: 1.0642771383355127
# r2: 0.5929125769657074
# rmse is above 1.0
# r2 is below 0.7
# for dataset dt30 containing education-level data
# dt30

print(f'rmse: {dt30_rmse02}')
print(f'r2: {dt30_r202}')
```

```
rmse: 1.064277138335494
r2: 0.5929125769657217
```

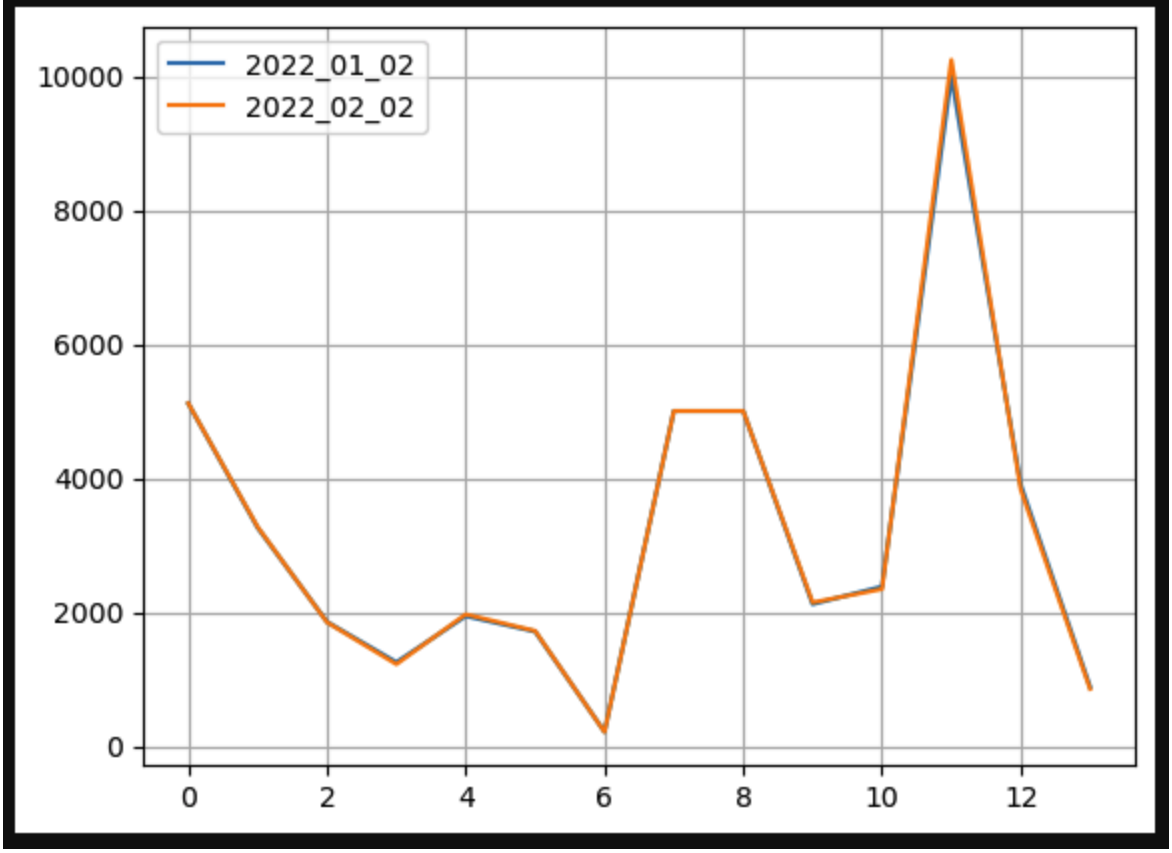
Interpretation

The marriage rate is on a steady decline and income is steadily rising. Despite this, housing sales are declining which is consistent with the marriage rate. The population rate is also sharply decreasing which is a variable I plan to explore in the next milestone. I think that inflation is the main culprit to the rise in housing prices despite there being an over abundance of dwellings for the population of the ROK.

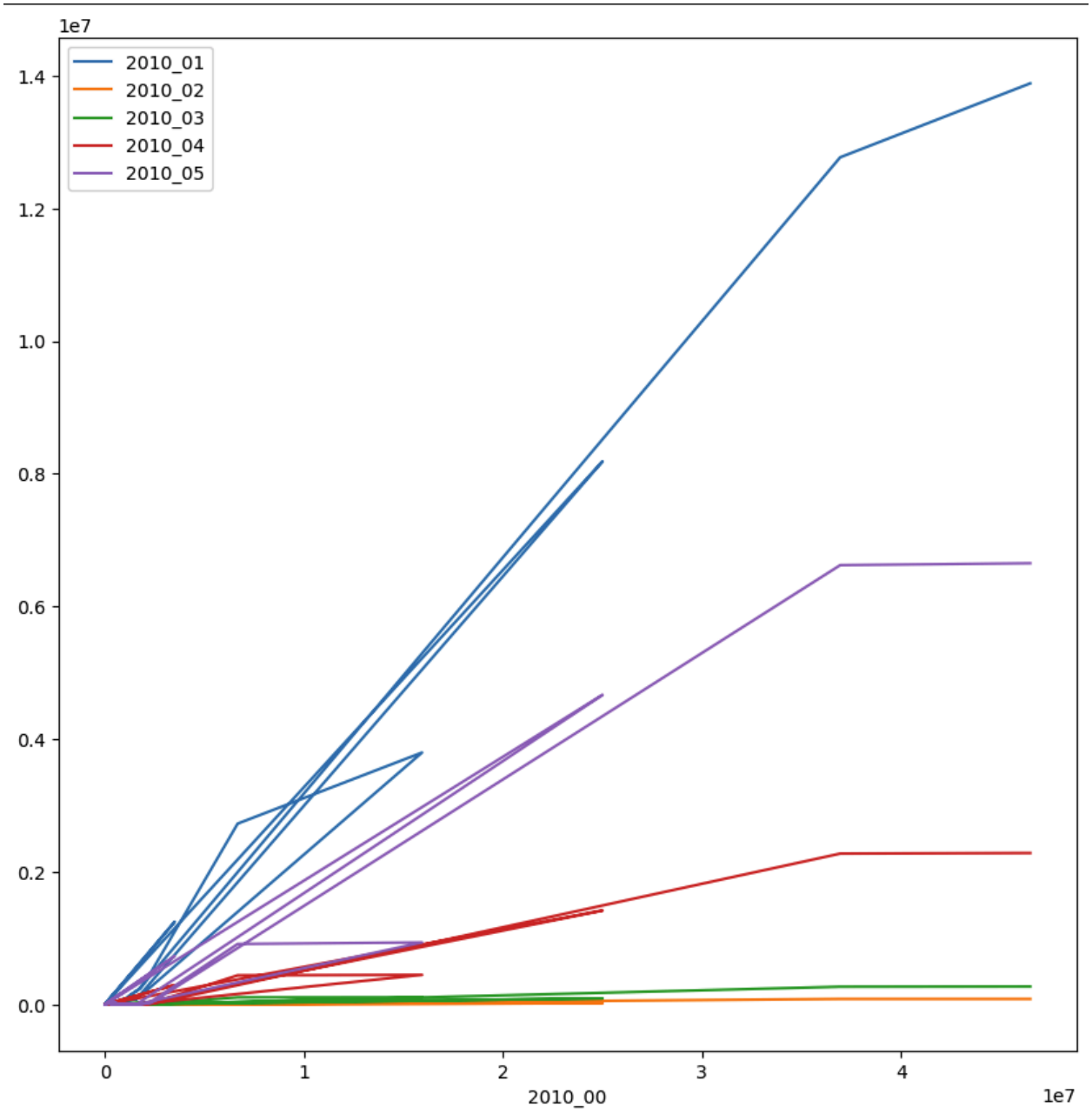
Below is a visualization of salary rates in the year 2022.

```
# 04.02.02-12
# return basic plot
# rendered basic plot to see vially and to determine if data is numeric.
# dt26

dt26_salaries_entir_01.plot()
plt.box(True)
plt.grid(True)
plt.title('', fontsize=16, color='#0047ab')
plt.xlabel('', fontsize=14, color='#0047ab')
plt.ylabel('', fontsize=14, color='#0047ab')
plt.show()
```



Due to inflation, annual incomes are also rising along with minimum wages which is consistent with any economy undergoing this change. Below is a visualization of salary rates for the year 2010.



Conclusion and Recommendations

I will need to explore population rate data sets in the final milestone to find any other correlations. Employment rates and annual construction rates would also be worth exploring. I will also run a model on population growth and compare it to forecasted growth rates provided by the Korean government. In order to gain a better understanding of the reason housing prices continue to rise, I will need to gain more insights into economically affecting metrics such as population, death rates, birth rates, natural growth rates, housing price rates, and other demographics.

Milestone Three

Ross L. Kim-Schreck

DSC630 Predictive Analytics

Professor Hua

2024.04.20

Exploring Policies to Stabilize Housing Prices in the ROK

Answering Questions

After analyzing the initial datasets, I can summarize that the datasets do not cover a long enough time period in order to explore trends over time. The first dataset is from the Korean government's official website. The dataset I'm using from that website is for the year 2023. I will retrieve every available dataset containing every year since the Korean war. This means that I will be adding an estimated 50 datasets with the same data as the 2023 dataset. It will be a lot to organize, but it is absolutely necessary in order for me to gain useful insights.

Useful Visualizations

Since this data as a whole represents trends over time, line graphs will especially be useful when it comes to visualization. I plan to also use bar graphs for categorical data; for example when comparing housing prices to average household income, employment rate, etc. I also plan to use multivariate line graphs for representation of categorical data and their transitions over periods of time.

Data adjustments and Driving Questions

I will need to make some major adjustments to the datasets. I will combine the government-provided datasets covering every year from 1975.

The datasets that I will be adding are the following from the Korean government's database:

1. 주택가격지수(매매)_20230320123223.csv
    a. 자치구별_01: name of city
    b. 자치구별_02: name of district
    c. 2023_01: housing prices (1월)
    d. 2023_01_01: housing prices (1월)
    e. 2023_02: housing prices (2월)
    f. 2023_02_01: housing prices (2월)
    g. 2023_03: housing prices (3월)
    h. 2023_03_01: housing prices (3월)
    i. 2023_04: housing prices (4월)
    j. 2023_04_01: housing prices (4월)
    k. 2023_05: housing prices (5월)
    l. 2023_05_01: housing prices (5월)
    m. 2023_06: housing prices (6월)
    n. 2023_06_01: housing prices (6월)
    o. 2023_07: housing prices (7월)
    p. 2023_07_01: housing prices (7월)
    q. 2023_08: housing prices (8월)
    r. 2023_08_01: housing prices (8월)
    s. 2023_09: housing prices (9월)

t. 2023_09_01: housing prices (9월)

u. 2023_10: housing prices (10월)

v. 2023_10_01: housing prices (10월)

w. 2023_11: housing prices (11월)

x. 2023_11_01: housing prices (11월)

y. 2023_12: housing prices (12월)

z. 2023_12_01: housing prices (12월)

The following datasets 02-49 contain the same categories as dataset one.

2. 주택가격지수(매매)_20220320133322.csv
3. 주택가격지수(매매)_20210320143421.csv
4. 주택가격지수(매매)_20200320153520.csv
5. 주택가격지수(매매)_20190320163619.csv
6. 주택가격지수(매매)_20180320173718.csv
7. 주택가격지수(매매)_20170320183817.csv
8. 주택가격지수(매매)_20160320193916.csv
9. 주택가격지수(매매)_20150320104015.csv
10. 주택가격지수(매매)_20140320114114.csv
11. 주택가격지수(매매)_20130320124213.csv
12. 주택가격지수(매매)_20120320134312.csv
13. 주택가격지수(매매)_20110320144411.csv
14. 주택가격지수(매매)_20100320154510.csv
15. 주택가격지수(매매)_20090320164609.csv

16. 주택가격지수(매매)_20080320174708.csv
17. 주택가격지수(매매)_20070320184807.csv
18. 주택가격지수(매매)_20060320194906.csv
19. 주택가격지수(매매)_20050320205005.csv
20. 주택가격지수(매매)_20040320215104.csv
21. 주택가격지수(매매)_20030320225203.csv
22. 주택가격지수(매매)_20020320235302.csv
23. 주택가격지수(매매)_20010320245401.csv
24. 주택가격지수(매매)_20000320255500.csv
25. 주택가격지수(매매)_19990320265699.csv
26. 주택가격지수(매매)_19980320275798.csv
27. 주택가격지수(매매)_19970320285897.csv
28. 주택가격지수(매매)_19960320295996.csv
29. 주택가격지수(매매)_19950320306095.csv
30. 주택가격지수(매매)_19940320316194.csv
31. 주택가격지수(매매)_19930320326293.csv
32. 주택가격지수(매매)_19920320336392.csv
33. 주택가격지수(매매)_19910320346491.csv
34. 주택가격지수(매매)_19900320356590.csv
35. 주택가격지수(매매)_19890320366689.csv
36. 주택가격지수(매매)_19880320376788.csv
37. 주택가격지수(매매)_19870320386887.csv

38. 주택가격지수(매매)_19860320396986.csv
39. 주택가격지수(매매)_19850320407085.csv
40. 주택가격지수(매매)_19840320417184.csv
41. 주택가격지수(매매)_19830320427283.csv
42. 주택가격지수(매매)_19820320437382.csv
43. 주택가격지수(매매)_19810320447481.csv
44. 주택가격지수(매매)_19800320457580.csv
45. 주택가격지수(매매)_19790320467679.csv
46. 주택가격지수(매매)_19780320477778.csv
47. 주택가격지수(매매)_19770320487877.csv
48. 주택가격지수(매매)_19760320497976.csv
49. 주택가격지수(매매)_19750320508075.csv
50. Housing_census_2020
    a. type
    b. region 01
    c. region 02
    d. region 03
    e. region 04
    f. YYYY/MM

The following datasets 51-54 contain the same categories as dataset 50.

51. housing_census_2015
52. housing_census_2010

53. housing_census_2005
54. housing_census_2000

I will be combining one metric from each 주택가격지수(매매) dataset: average housing price; I will combine all 49 variables into one matrix for time-series data analysis. I will use housing_census data for supply/demand statistics which will be compared and correlation determined based on a variety of metrics.

Model Adjustments and Evaluation Choices

I think that my initial plan to use various regression models to detect trends in historical data and to make future predictions is necessary, and there aren't any obvious adjustments to be made thus far. I will use moving-average models, exponential-smoothing models, and SARIMA which could all be used to determine patterns, detect anomalies, and forecast future trends.

Original Expectations

My expectations are reasonable as long as the necessary datasets are added, and the data is adjusted accordingly. I also plan to make comparisons and detect correlations in other categories such as average household income, employment rate, etc.

Milestone Two

Ross L. Kim-Schreck

DSC630 Predictive Analytics

Professor Hua

2024.04.13

Exploring Policies to Stabilize Housing Prices in the ROK

Introduction

For this project, I plan to take a deep dive into some datasets relating to real estate and housing prices in the Republic of Korea, which I will refer to as the 'ROK' in this research paper. The ROK was once a war-torn impoverished nation that was overcrowded, had a very high birth rate, and had a GDP of only 47.7 billion Korean won (KRW), which was lower than that of its counterpart, North Korea the (DPRK) at the time. However, the ROK now has a GDP of 1,485 trillion KRW, which is 13th of OECD nations, thus it boasts the 13th largest economy in the world and third largest economy in East Asia. The economic transition and stark contrast between past and present seem miraculous, in fact, this is known as 'the miracle of the Han'. Photos of post Korean-war-torn Seoul are bleak, revealing a leveled cityscape with few standing buildings.

After the Korean war, the nation underwent a spike in birth rate, not to be mistaken as a baby boom similar to that of post-world war two United States. This was a byproduct of being an impoverished nation, this 'baby boom' had a severe demographic implication: overcrowding. The nation's infrastructure simply could not accommodate such a high population, especially in the capital Seoul, where a third of the population is concentrated. Despite technically being a democracy, the ROK was a military dictatorship up until the late 1980s. The government took an economy-first initiative and since, has always first and foremost prioritized its economy and

its image abroad. BTS is government-funded manufactured entertainment exported abroad to promote the Korean nation. Any entity, whether it be corporate or social, would not be able to exist without the government's approval. The government hand picked families of privilege, much like oligarchs, and placed them in strategic economic positions. In Korean, these families are known as Chaebols; they are in fact remnants of Korea's monarchical past and function like mini autocracies to this day. In this democracy, autocracy and nepotism are still rampant. Companies such as Samsung, Hyundai, LG, and SK, account for the majority of Korea's economy and wield tremendous political power. And to this day, these monarchies are still run by the same hand-picked families. These conglomerates (Chaebols) were assigned specific industries by the government in which they could maneuver for the government. With the CEO as king, it is impossible for individuals outside of a Chaebol family to rise to high positions within the company; this is systematic nepotism and monarchism existing in a modern-day liberal democracy. In the 70s and 80s, this dictatorship implemented laws forbidding families from having more than two children, much like China's one-child policy. As a result the birth rate has plummeted and presently is the lowest in the world. There is a domino effect taking place in East Asia that began with Japan; now it's Korea's turn, and China will follow once the one-child policy aftermath ensues. These three nations and their demographics have direct and massive implications on East Asia,

which make up 20.66% of the world's population, and continental Asia, which make up 59.22% of the world's population.

Despite the plummeting birth rate, housing prices continue to steadily increase. There is now more housing available to individuals and families than ever before, and housing vacancies continue to rise. As the population growth rate is projected to level off in 2024, housing vacancies will likely increase at a rapid pace. And yet housing prices have increased steadily since the Korean war and continue this trend. I suspect that the Korean government, for the same reasons it keeps a firm grip on all other industries, is manipulating the housing market to its fiscal benefit. Despite having some of the cleanest water in the region, the government continues to persuade the public that tap water is undrinkable, prompting households to purchase water purifiers, which are very expensive. This is just one of countless examples of the government evading issues, omitting facts, and implementing policies for its own ulterior motive. I believe the main culprit to the low fertility level in this nation is the high prices of housing. The discouraging prices of real estate are deterring couples from marriage and progeny. The reason that, despite the natural growth rate is decreasing, the price of housing continues to increase, is an issue that I would like to unpack in this research.

Most individuals and families are being affected by the outrageously high prices of real estate and in turn is affecting the population's natural

growth rate. The plummeting growth rate is affecting many companies in every industry and will have catastrophic economic and demographic implications in the near future. I hope to persuade policymakers that the old economic-first initiative is no longer sustainable, and that we should take measures to stabilize the housing market to counter the ever so decreasing growth rate in the Republic of Korea.

Data Sources

I have chosen 56 datasets to start off my project:

1. 주택가격지수(매매)_20240320113224.csv
    a. 자치구별_01: name of city
    b. 자치구별_02: name of district
    c. 2023_10: housing prices (10월)
    d. 2023_10_01: housing prices (10월)
    e. 2023_11: housing prices (11월)
    f. 2023_11_01: housing prices (11월)
2. 주택가격지수(매매)_20230320123223.csv
    a. 자치구별_01: name of city
    b. 자치구별_02: name of district
    c. 2023_01: housing prices (1월)
    d. 2023_01_01: housing prices (1월)
    e. 2023_02: housing prices (2월)

f. 2023_02_01: housing prices (2월)
g. 2023_03: housing prices (3월)
h. 2023_03_01: housing prices (3월)
i. 2023_04: housing prices (4월)
j. 2023_04_01: housing prices (4월)
k. 2023_05: housing prices (5월)
l. 2023_05_01: housing prices (5월)
m. 2023_06: housing prices (6월)
n. 2023_06_01: housing prices (6월)
o. 2023_07: housing prices (7월)
p. 2023_07_01: housing prices (7월)
q. 2023_08: housing prices (8월)
r. 2023_08_01: housing prices (8월)
s. 2023_09: housing prices (9월)
t. 2023_09_01: housing prices (9월)
u. 2023_10: housing prices (10월)
v. 2023_10_01: housing prices (10월)
w. 2023_11: housing prices (11월)
x. 2023_11_01: housing prices (11월)
y. 2023_12: housing prices (12월)
z. 2023_12_01: housing prices (12월)

The following datasets 03-70 contain the same categories as dataset two.

3. 주택가격지수(매매)_20220320133322.csv
4. 주택가격지수(매매)_20210320143421.csv
5. 주택가격지수(매매)_20200320153520.csv
6. 주택가격지수(매매)_20190320163619.csv
7. 주택가격지수(매매)_20180320173718.csv
8. 주택가격지수(매매)_20170320183817.csv
9. 주택가격지수(매매)_20160320193916.csv
10. 주택가격지수(매매)_20150320104015.csv
11. 주택가격지수(매매)_20140320114114.csv
12. 주택가격지수(매매)_20130320124213.csv
13. 주택가격지수(매매)_20120320134312.csv
14. 주택가격지수(매매)_20110320144411.csv
15. 주택가격지수(매매)_20100320154510.csv
16. 주택가격지수(매매)_20090320164609.csv
17. 주택가격지수(매매)_20080320174708.csv
18. 주택가격지수(매매)_20070320184807.csv
19. 주택가격지수(매매)_20060320194906.csv
20. 주택가격지수(매매)_20050320205005.csv
21. 주택가격지수(매매)_20040320215104.csv
22. 주택가격지수(매매)_20030320225203.csv
23. 주택가격지수(매매)_20020320235302.csv
24. 주택가격지수(매매)_20010320245401.csv

25. 주택가격지수(매매)_20000320255500.csv
26. 주택가격지수(매매)_19990320265699.csv
27. 주택가격지수(매매)_19980320275798.csv
28. 주택가격지수(매매)_19970320285897.csv
29. 주택가격지수(매매)_19960320295996.csv
30. 주택가격지수(매매)_19950320306095.csv
31. 주택가격지수(매매)_19940320316194.csv
32. 주택가격지수(매매)_19930320326293.csv
33. 주택가격지수(매매)_19920320336392.csv
34. 주택가격지수(매매)_19910320346491.csv
35. 주택가격지수(매매)_19900320356590.csv
36. 주택가격지수(매매)_19890320366689.csv
37. 주택가격지수(매매)_19880320376788.csv
38. 주택가격지수(매매)_19870320386887.csv
39. 주택가격지수(매매)_19860320396986.csv
40. 주택가격지수(매매)_19850320407085.csv
41. 주택가격지수(매매)_19840320417184.csv
42. 주택가격지수(매매)_19830320427283.csv
43. 주택가격지수(매매)_19820320437382.csv
44. 주택가격지수(매매)_19810320447481.csv
45. 주택가격지수(매매)_19800320457580.csv
46. 주택가격지수(매매)_19790320467679.csv

47. 주택가격지수(매매)_19780320477778.csv
48. 주택가격지수(매매)_19770320487877.csv
49. 주택가격지수(매매)_19760320497976.csv
50. 주택가격지수(매매)_19750320508075.csv
51. seoul - SeoulRealEstate.csv
    a. id: unique identifier
    b. lat: latitude
    c. lon: longitude
    d. households: number of households
    e. date_build: date of construction
    f. score: land quality rating
    g. m2: size of area in square meters
    h. p: size of area in Korean units
    i. sales_min: minimum sales
    j. sales_max: maximum sales
    k. sales_ave: average sales
52. housing_census_2020
    a. type
    b. region 01
    c. region 02
    d. region 03
    e. region 04

f. YYYY/MM

The following datasets 53-56 contain the same categories as dataset 52.

53. Housing_census_2015
54. Housing_census_2010
55. Housing_census_2005
56. Housing_census_2000

These datasets can be found at the following Github link:

https://github.com/rlawnsdnjs706/DSC630

Models

I plan to use a variety of regression models to train based on past data, particularly regarding housing-price trends and to forecast future trends. I hope to determine which variables show consistency with past and future trends and which variables have the greatest impact.

Plan of Evaluation

Using the house price index of the ROK and other metrics, I plan to evaluate using root-mean-squared-error between observed sales prices and predicted values. I plan to create a series of clear and concise visualizations that could persuade and even motivate policy makers to avoid this crisis.

What I Hope to Learn

While the growth-rate increase is expected to level off in 2024, I hope to explore ways in which the Korean government could potentially avoid the coming demographic catastrophe through policies by accurately forecasting housing prices and land value and finding ways to stabilize them. I also hope to learn effective ways of storytelling through a series of visualizations.

Risks with the Proposal

My second source of data is from Kaggle: the second dataset which is for general international use thus is entirely in English, and the first dataset is from the Korean government's data resources: data.seoul.go.kr, which is in Korean. I will need to romanize the first dataset so that visualizations can be understood by an international audience. I will likely need to use similar functions to fuzzy matching for the process of romanization as there are three different methods of romanization of Korean script. I will make functions in Python to automate this process; however, this could potentially skew results if automated matching is ineffective or inaccurate. It is possible that if not done manually, this could create mismatches, missing data, and duplicates, so I will manually romanize accordingly.

The first dataset spans a much shorter time period; I may need to add datasets that can fill the missing gaps of historical data.

Contingency Plan

Depending on the accuracy of the results, I will likely need to find supplemental datasets to at least support my existing data. Also, if my results do not reveal a clear path to an initially sought-out solution and do not reveal a clear roadmap of action items for policy makers, I may need to add entirely new datasets containing additional relevant metrics such as demographics, cost of living, economic trends, suicide rates, marriage and divorce rates, individual income, etc.

References

8. https://www.macrotrends.net/global-metrics/countries/KOR/south-korea/gdp-gross-domestic-product
9. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/
10. https://www.kaggle.com/datasets/jcy1996/seoul-real-estate-datasets
11. https://data.seoul.go.kr/dataList/801/S/2/datasetView.do

Milestone One

Ross L. Kim-Schreck

DSC630 Predictive Analytics

Professor Hua

2024.03.16

Project Plan

I will be working individually as my time zone could affect collaboration with team members by making communication an obstacle. I plan to work on my project milestones every evening after work. I will decide on details once a data set and subject has been chosen. I hope to allocate at least 90 minutes per day for assignments and the term project.

I will be using week two to prepare for week three. In week three, I will treat milestone two as an introduction of the project. In this step I plan to select the data sets that I will use; I may include supplemental sets in case the amount of data doesn't meet the requirements in later stages. I will lay out a comprehensive plan on what I will do with the data and for which reasons.

In weeks four to six, I will be preparing for week seven's third milestone submission. I will update information from week two accordingly. I will adjust the data for analysis and visualization.

In weeks eight to nine, I will be preparing for week 10's fourth milestone submission. In this milestone, I hope to have made all the necessary data transformations needed after having considered supplemental data to be added prior to this step. I hope to build several models to validate specific outcomes.

In week 11, I will be preparing to submit the final paper and presentation in week 12.

Peer Review

I will be doing individual peer reviews with Kalyan Pothineni and Alexis Johnson for milestone two. We may decide to make a team if we're all able to coordinate our schedules.