Urban Polarization – Data Analysis

Contents

What is urban polarization?

I asked GPT!

"Occurs within or between cities

"It refers to the phenomenon of deepening

I. Project Overview – Defining Urban Polarization

Urban Polarization refers to the process by which economic, social, spatial, or demographic differences within a city become increasingly divided into contrasting extremes, rather than remaining evenly distributed across the urban population or territory.

In practice, urban polarization typically manifests as:

- **Economic polarization**: A growing gap between high-income and low-income residents, often accompanied by the expansion of high-wage professional jobs alongside low-wage service work, with a shrinking middle class.
- **Spatial polarization**: The physical separation of social groups within the city, such as affluent neighborhoods coexisting with areas of concentrated poverty, informal settlements, or marginalized districts.
- **Social and demographic polarization**: Segregation based on factors such as education, ethnicity, immigration status, or access to public services.
- **Access polarization**: Unequal access to housing, transportation, healthcare, education, and employment opportunities across different urban areas.

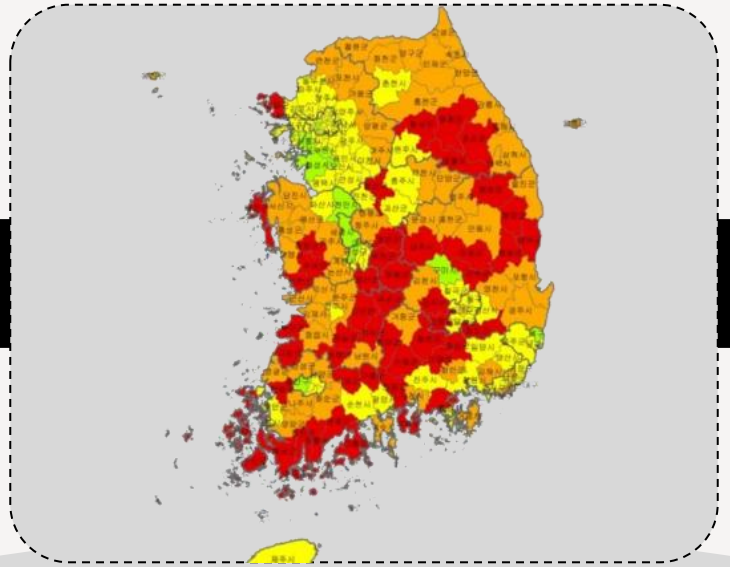
Urban polarization is commonly linked to processes such as globalization, deindustrialization, gentrification, housing market dynamics, and uneven urban policy interventions. Over time, it can intensify social inequality, reduce social cohesion, and create persistent disparities in life chances within the same city.

I. Project Overview – Why Polarization?

Why

Goo
d

p



?

Seri
ous

"The society is seriously aware of the problems of regional extinction and local extinction. Accordingly, we hope to help resolve

I. Project Overview – Analysis Plan – Target

- **Analysis subjects: Cities in the Chungcheong region (Sejong, Daejeon, North Chungcheong, South Chungcheong)**

Population influx

대전광역시

세종 특별자치시

Reason for selection

"It's not a metropolitan area,

but

with

cities in danger of extinction

it is considered appropriate to

Some extinction threats

충청남도

충청북도

How to analyze urban polarization?

As expected… I asked GPT!

economic gap

"Rich areas and poor areas

"Regional income disparities, employment

social gap

"Differences in access to social resources such as education, health,

spatial gap

"Imbalance in the distribution of residential, commercial, and

- **statistical data**

Proportion of the working-age population

crime rate

Apartment sales Actual

Average monthly income distribution

compare on

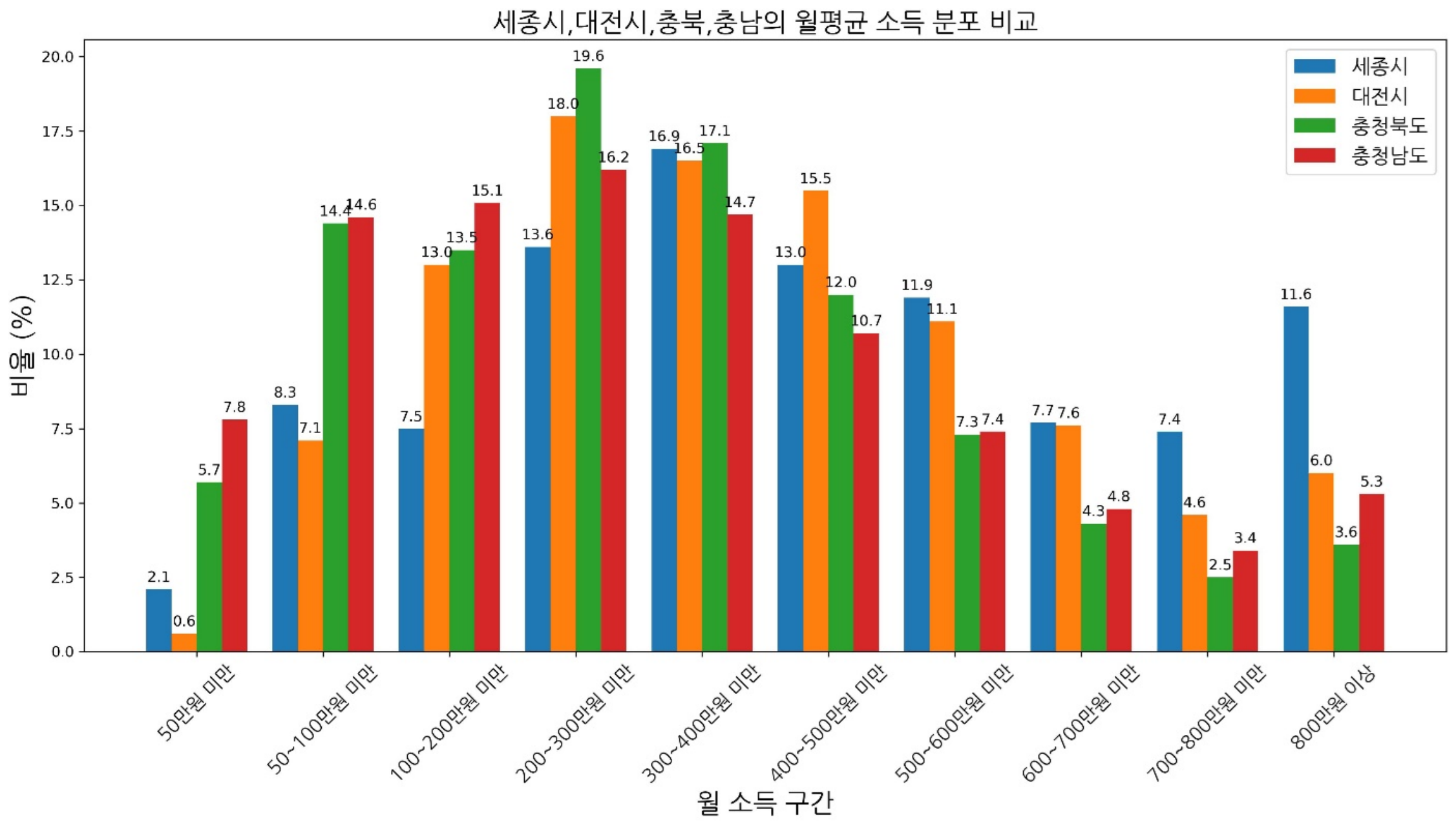
analyze

conclusion

the working-age population ratio, apartment transaction price index, and monthly average income distribution, we can see that polarization is clearly evident.

Among them, we will select **the apartment transaction price index as a dependent**

I. Project Overview – Analysis Plan – How?



Why did you choose apartment prices as a key variable?

Apartment prices are considered to be a variable that clearly shows the relative wealth and poverty (degree of backwardness) of a city .

who make up the majority of the region's average monthly income **, choose apartments that offer convenience or luxury apartments as their primary residences, while it is difficult for low-income people to enter apartments** .

I. Project Overview – Analysis Plan – Selection Variables

Variable selection method

Backward elimination method (Backward Elimination)

Initially there were 32 variables

I trained using it, but Overfitting occurs

① Dimension reduction to solve

(Variable removal,

Independent variables (final selection) ✔

- Regional population movement data (Total inflow, total transfer)
- Apartment sales supply and demand trends
- PIR index
- Presidential approval rating
- Comprehensive real estate tax
- Home mortgage interest rates
- Apartment sales volume
- consumer price index

dependent variable ✔

Apartments by region

Real transaction index data

I. Project Overview – Analysis Plan – Excluded Variables

Excluded variables and reasons

Number of single-person	◆ High correlation was confirmed, but **quick reflection was not possible due to lack of monthly data.**
Dominance in the number of National	◆ I tried inserting it using One-Hot Encoding, but **the prediction rate decreased.**
Housing Affordability Index	◆ Only quarterly data, not monthly data, is available, and **quick reflection is not possible.** ◆ Overlapping variables used with the PIR index
Base interes t rate	◆ Since they show a strong positive correlation of 0.86, **to resolve the multicollinearity problem** Adoption of mortgage loan interest rates with a high correlation with
LIR index	◆ The two assess the *affordability of the housing market from different perspectives, The PIR index, which has a very high correlation of 0.96 and is highly correlated with the real transaction index, was adopted. * Generally, the extent to which an individual or household can afford the costs required to purchase or rent a home.

team members and their roles

Data Analytics Professional Team

Seong

Data crawling
Data preprocessi

SHIN

DB design and construction
Data

Park

Planning and overall management
Building ML

Lee

Backend construction
Data visualization

Oh

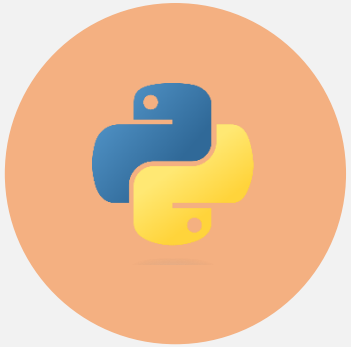
Backend construction
Data visualization

III. Development Process – 1. Data Collection

Data collection

Case ① Crawling (automatic collection)

API
(data URL)



crawling
(URL-based collection)

Case ② Research (direct collection)

API
(Unavailable)

Access the webpage (Direct collection)

..
.

III. Development Process – 1. Data Collection

Data collection

HF 한국주택금융공사 주택금융연구원

Gallup 한국갤럽

- Troubleshooting some variable crawl failures.

Failed to crawl data from the Korea Housing Finance Research Institute and Gallup Korea.

We plan to design and develop a crawling

Data preprocessing

Method 1: Data frame defect based on the 'time' column

Point in time (reference column)	population movement data	PIR index
January 2013	100000	100
February 2013	200000	110
March 2013	300000	120
…	…	…
December 2023	1000000	140

② Scaling

Min-Max scaling

$$x_{scaled} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Robust scaling

$$X_{scaled} = \frac{X_i - min(X)}{max(X) - min(X)}$$

III. Development Process – 3. Model Building (1)

Model building

(Existing model)

Mobile data

8 variables

Total transfer
Total transfer
Re-movement
Inter-city transfer
interprovincial transfer
Moving within the city/county
Moving within a city or province - moving between cities or counties
Moving within a city or province - moving between cities or counties

Dimension reduction

4 variables

Total transfer
Total transfer
Inter-city transfer
Intercity transfer

Overfitting occurs!

(result)

Bad !

Recognize the need for additional independent variables!

Prediction rate for the national apartment sales transaction index

Secure and **mix with regional data to improve performance**

III. Development Process – 3. Model Building (2)

① Min-Max

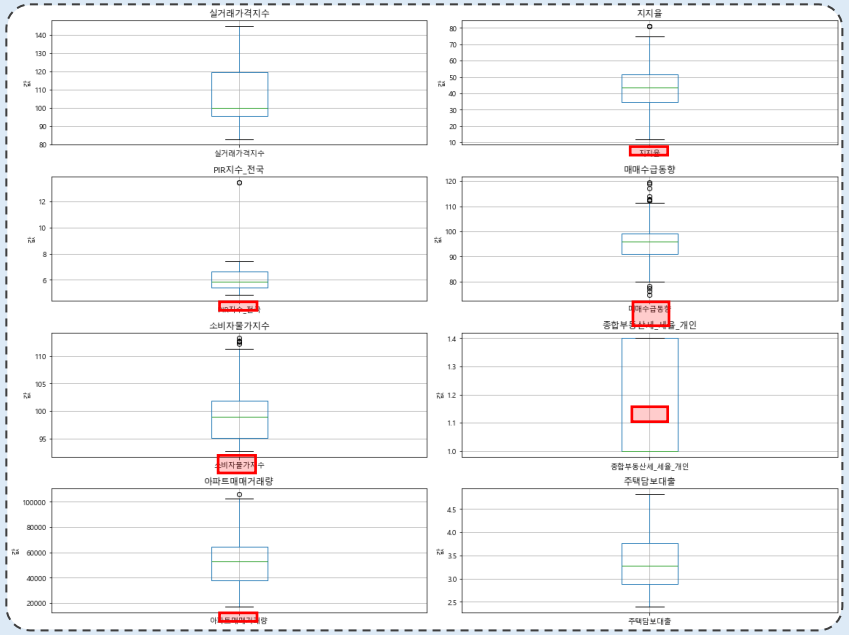
While maintaining data distribution

Select to

② Robust

Boxbeard drawing

Check variables where



III. Development Process – 3. Model Building (3)

result

Coefficient of determination

0.92~0.93

Polynomial transformation for performance improvement

Each applied

Degree 2

The coefficient of determination improves, but overfitting

Degree 3, Degree 4

Performance deterioration

Attempts to address overfitting

L1(Lasso)

1) Coefficient of determination: 0.983
2) Train MSE: 2.7359 / Test MSE: 3.9460
3) K-Fold average MSE: 5.2754

* **Slightly improved overfitting**

L2(Ridge)

1) Coefficient of determination: 0.978
2) Train MSE:3.6552 / Test MSE:5.1577
3) K-Fold average MSE: 6.3747

* **Reduced predictive power, increased average MSE = reduced performance**

ElasticNet (L1 + L2)

1) Coefficient of determination: 0.948
2)Train MSE:15.2791 / Test MSE:12.4829
3) K-Fold average MSE: 20.2519

* Decreased coefficient of determination, increased training/test MSE,

Increased average MSE = decreased performance

PCA (principal component analysis)

1) Coefficient of determination: 0.912
2) Train MSE:4.4733 / Test MSE:22.9941
3) K-Fold average MSE: 388.6948

* Decreased coefficient of determination, still overfitting,

K-Fold average MSE increases sharply -> performance deteriorates

Future development direction

Implementation of a linear regression model/deep learning model

Adoption

Select Robust + L1 (The best!!)

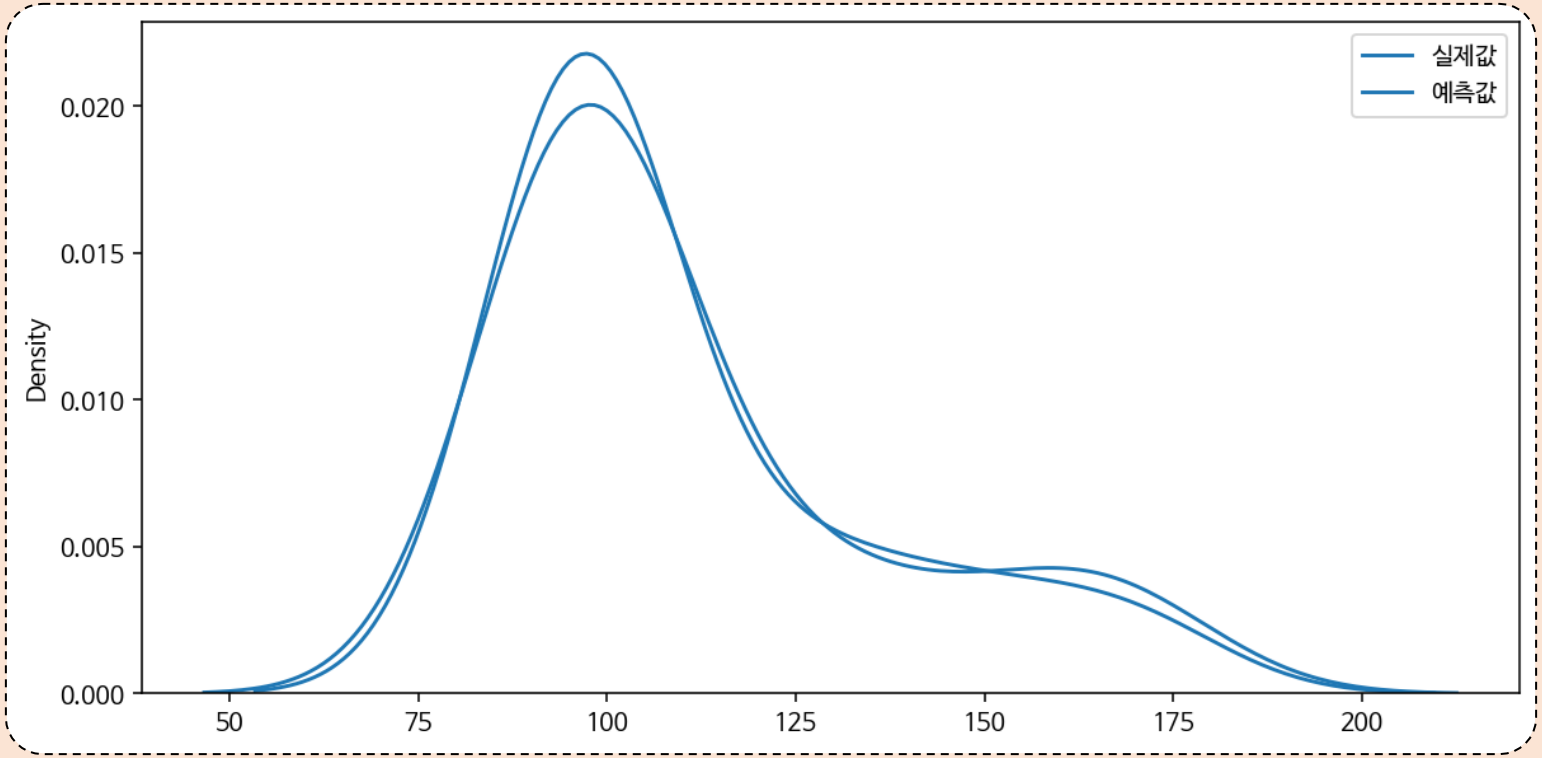
verification

Demonstration of Sejong City's actual transaction index forecast

III. Development Process – 3. Model Building (4)

Future Direction (still needs

We plan to improve performance by testing



Data Visualization & Service Building

Data Visualization & Service Building

① Net population migration: Implementation of a nationwide net population migration bar chart and map based on the combined data from 2013 to 2023.

② Real Transaction Price Index and Unemployment Rate: Insert the 2023 Daejeon/Sejong/Chungbuk/Chungnam Real Transaction Price Index and Unemployment Rate BAR CHART (image)

③ Crime Rate and Working-Age Population: Insert the 2023 Daejeon/Sejong/North Chungcheong/South Chungcheong crime rate and working-age population BAR CHART (image)

matplotlib

Streamlit

Data Visualization & Service Building

Network URL: http://192.168.71.220:8501

1. General
1) Map implementation: Daejeon, Sejong, South Chungcheong, and North Chungcheong regions are marked with colors.
2) Implementation of the Apartment Transaction Price Index LINE CHART: 4 regions: Daejeon, Sejong, Chungnam, and Chungbuk
3) Monthly average income bracket ratio implemented as a BAR CHART: 4 regions: Daejeon, Sejong, South Chungcheong, and North Chungcheong

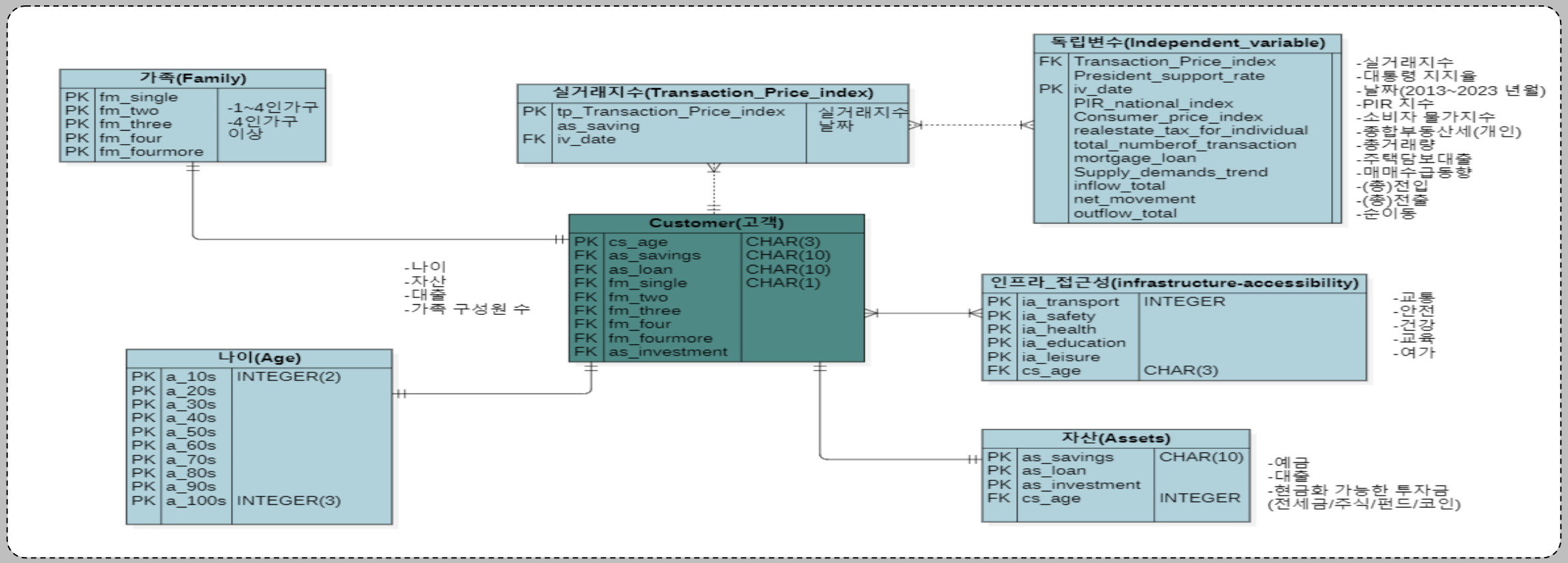
2. Real Transaction Index Prediction Model (Demo Version)
1) Select "Region," "Start Date," and "End Date" from the left menu.
2) Predicting future apartment transaction indices using linear regression machine learning algorithms.

3. Real Estate Listing Recommendation Model (Demo Version)

III. Development Process – 5. DB Design and Construction (1)

DB design and construction

All data loaded into DB



III. Development Process – 5. DB Design and Construction (2)

DB design and construction

Extracting data from a DB

① Writing a sample SQL query

```
# 머신러닝 데이터 선정
# 예시 : 2014.01 이전 데이터 + 인프라 수준 1등급 + 용산구

SELECT * FROM parameter where Pm_Date < 2014.01;
select inflow_total, outflow_total, net_movement from seoul where Io_Date < 2014.01;
select Subway, Primary_School, Middle_School, High_School, General_Hospital, Supermarket, Park
    from fancy where district = 'yongsan-gu';

SELECT
    p.*,
    s.inflow_total, s.outflow_total, s.net_movement,
    f.Subway, f.Primary_School, f.Middle_School, f.High_School, f.General_Hospital, f.Supermarket, f.Park
    FROM parameter p
    JOIN seoul s ON p.Pm_Date = s.Io_Date
    JOIN fancy f ON f.district = 'yongsan-gu'
    WHERE p.Pm_Date < '2014-01-01' AND  s.Io_Date < '2014-01-01';
```

② Data extraction

outflow_total	net_movement	Subway	Primary_School	Middle_School	High_School	General_Hospital
151771	-14152	981	1261	140	159	1200
131098	-11357	981	1261	140	159	1200
142141	-8773	981	1261	140	159	1200
109528	-6519	981	1261	140	159	1200
126101	-6930	981	1261	140	159	1200

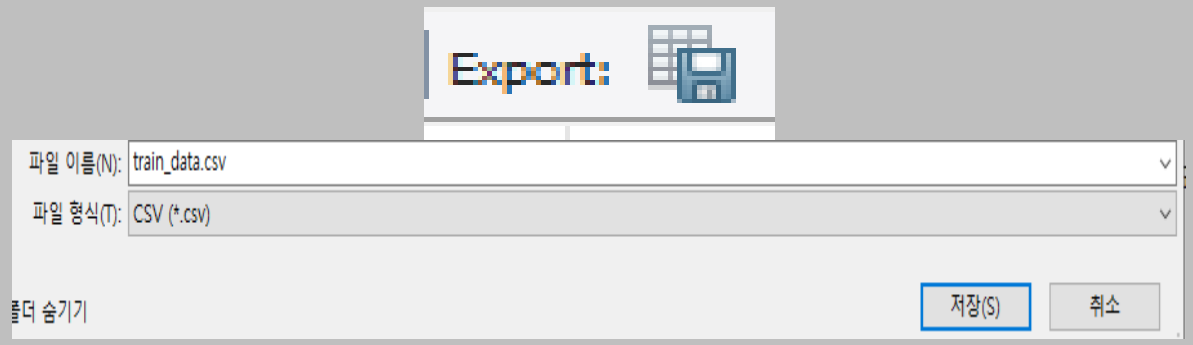
III. Development Process – 5. DB Design and Construction (3)

DB design and construction

Export and Model Training

※ plan to write a module to extract data from the DB in the future.

③ Export to CSV



④ Data for model

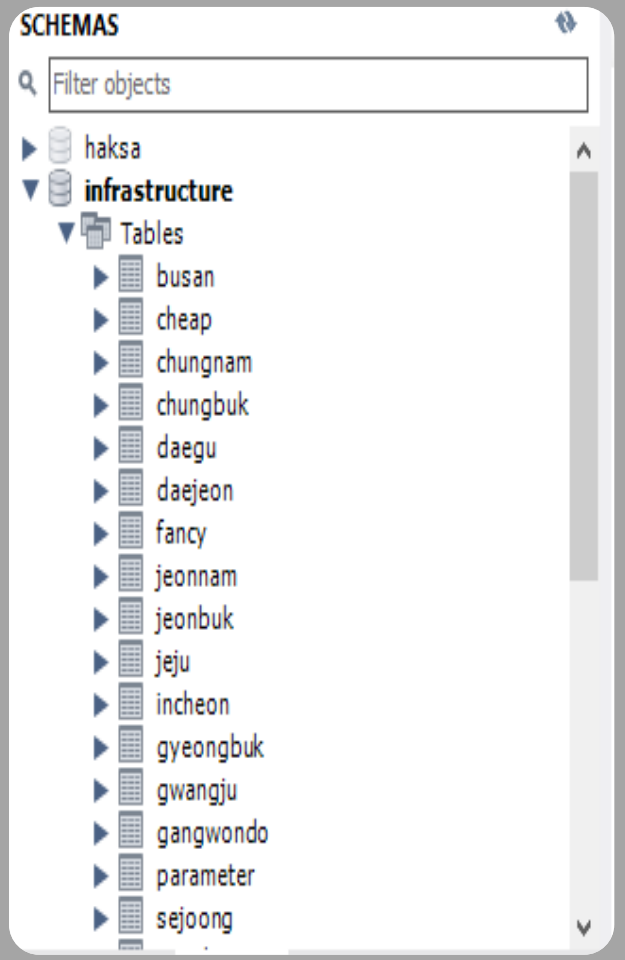
python

MACHINE LEARNING

```
Pm_Date,Transaction_Price_index,President_support_rate,PIR_national_index,Consumer_price_index,realestat
2013.12,85.7,54,5.1,93.229,1.0,67883,3.74,"94.53",137619,151771,-14152,981,1261,140,159,1200,1400,1600
2013.11,85.7,54,5.04,93.116,1.0,61844,3.77,"95.61",119741,131098,-11357,981,1261,140,159,1200,1400,1600
2013.10,85.7,54,5.09,93.134,1.0,65871,3.81,"95.66",133368,142141,-8773,981,1261,140,159,1200,1400,1600
2013.09,85.2,60,5.08,93.419,1.0,39801,3.82,"92.23",103009,109528,-6519,981,1261,140,159,1200,1400,1600
2013.08,84.6,60,5.09,93.238,1.0,30794,3.8,"89.56",119171,126101,-6930,981,1261,140,159,1200,1400,1600
2013.07,84.3,60,5.15,92.909,1.0,25079,3.77,"90.41",118420,127308,-8888,981,1261,140,159,1200,1400,1600
2013.06,83.7,51,5.52,92.71,1.0,94647,3.73,"92.1",115773,124523,-8750,981,1261,140,159,1200,1400,1600
2013.05,84.0,51,5.62,92.823,1.0,64538,3.77,"93.34",130619,138830,-8211,981,1261,140,159,1200,1400,1600
2013.04,83.8,51,5.65,92.823,1.0,55442,3.86,"92.34",134581,144586,-10005,981,1261,140,159,1200,1400,1600
2013.03,83.5,42,5.53,92.952,1.0,47375,3.97,"88.63",138507,146220,-7713,981,1261,140,159,1200,1400,1600
2013.02,83.1,42,5.49,93.038,1.0,34089,4.06,"87.21",150805,155528,-4723,981,1261,140,159,1200,1400,1600
2013.01,83.0,42,5.35,92.728,1.0,16968,4.17,"86.1",118477,123006,-4529,981,1261,140,159,1200,1400,1600
```

III. Development Process – 5. DB Design and Construction (4)

Why You Should Use a DBMS to Build a Machine Learning Model (1)



DBMS

Seoul City's High-Rise Apartment Infrastructure Data – fancy

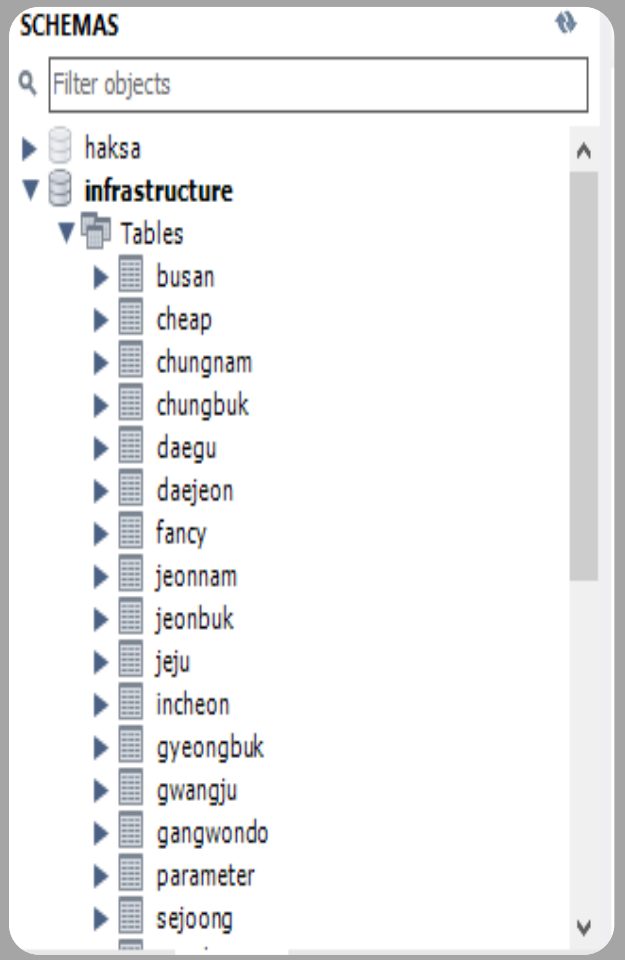
district	Subway	Primary_School	Middle_School	High_School	General_Hospital
yongsan-gu	302	769	772	741	1900
eunpyeong-gu	363	462	763	893	2600
jonglo-gu	499	756	1200	1200	1000
jung-gu	310	1417	286	224	1800
junglang-gu	845	601	856	491	860
seongdong-gu	288	428	840	336	1900

Seoul's low-cost apartments - infrastructure data – cheap

district	Subway	Primary_School	Middle_School	High_School	General_Hospital
yongsan-gu	981	1261	140	159	1200
eunpyeong-gu	413	367	344	929	3000
jonglo-gu	588	1000	382	365	430
jung-gu	514	524	139	183	1500
junglang-gu	575	733	983	857	342
seongdong-gu	528	706	619	572	2300

III. Development Process – 5. DB Design and Construction (5)

Why You Should Use a DBMS to Build a Machine Learning Model (2)



DBMS

Variables with a high contribution to the real transaction index – parameter

Pm_Date	Transaction_Price_in	President_support_ra	PIR_national_index	consumer_price_index
2013.01	83.0	42	5.35	92.728
2013.02	83.1	42	5.49	93.038
2013.03	83.5	42	5.53	92.952
2013.04	83.8	51	5.65	92.823
2013.05	84.0	51	5.62	92.823
2013.06	83.7	51	5.52	92.71

Population Migration Data – Busan

Io_Date	inflow_total	outflow_total	net_movement
2013.01	39348	40048	-700
2013.02	42857	42807	50
2013.03	39546	41004	-1458
2013.04	40603	41548	-945

Data extraction format (planning to use SQL)

Example of use)

1) 2013~2014 data + infrastructure level 1 (Seocho-gu) + Seocho-gu (population movement data)

The period with the most dramatic price fluctuations was 2019-2023.

2) 2019~2023 data + infrastructure level 1 (Yongsan-gu) + Yongsan-gu (population movement data)

3) 2013~2023 data + infrastructure level 5 (Seoul) + Seoul (population movement data)

Conclusion

1) Coefficient of determination: 0.968

2) Train MSE:6.3516 Test MSE:15.6219

3) Average MSE: 24.9379

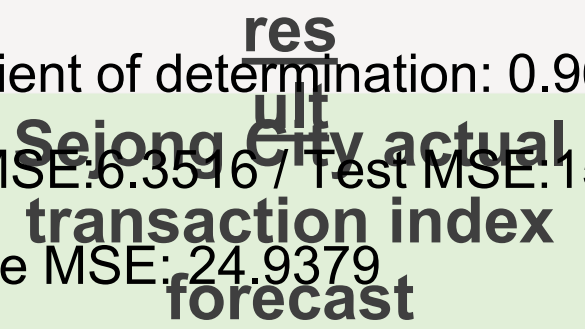
- After training by entering data from 13.01 to 23.11,

Predicting the actual transaction index for December 23

- Actual transaction index: 122.0

- Model predicted actual transaction index: 130.33

*** Many performance improvements are needed**



1. Apartment price prediction service

Hypothesis: Higher infrastructure levels lead to higher real estate prices.

insp

eati

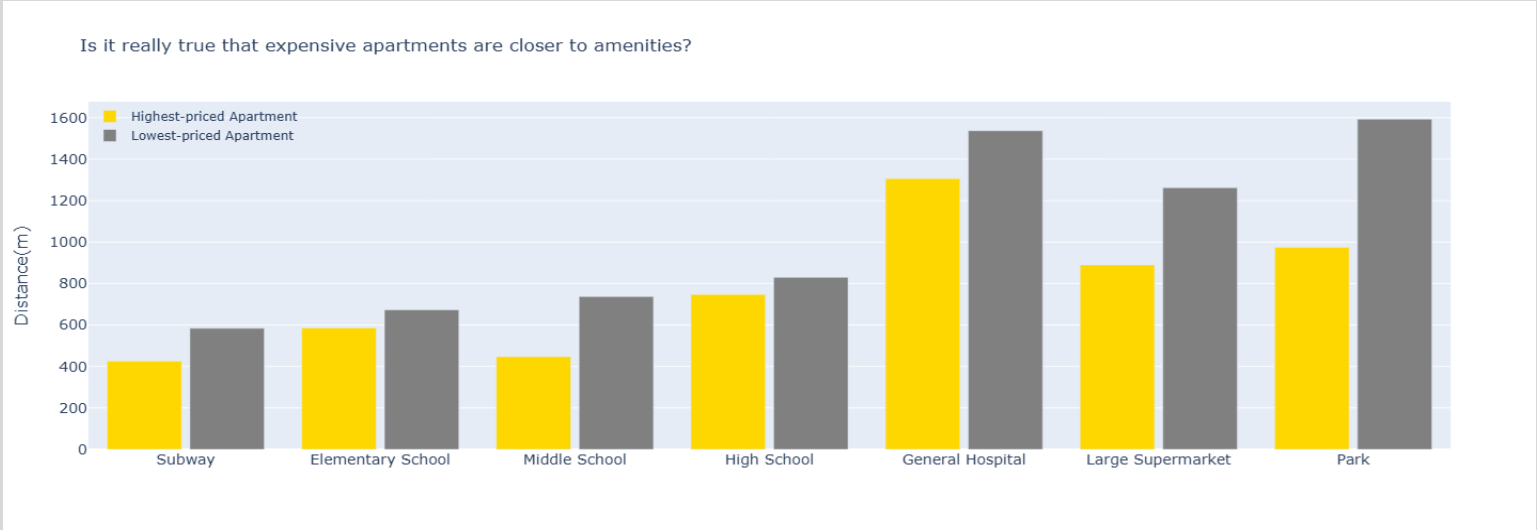
on!

(Scope) **Old apartments in Seoul**

(Target) **Highest price apartment, lowest price apartment**

(Confirmation) **Distance between apartment and infrastructure**

Results: The more expensive the apartment, the closer it is to



Calculate the average distance between apartments and

2. Future development direction

Scenario-Based Property Recommendation Solution (Case Study of Virtual Character Kim)

Standard 1

Preferred Infrastruct

ure

**“I am a working person and I live with my wife.
I am looking for a place close to the subway station.
I like to run and run around my house.
“There is a park and it looks nice.”**

Standard 2

Homeless, age, asset size, household size, income level

Non-homeowner, age: 20s-30s, asset size: 500 million won

Residential capacity: 2 people, Annual salary: 50 million won

2. Future development direction

Customer-specific data collection and infrastructure accessibility classification

infrastructure	1st grade	2nd grade	3rd grade	Grade 4	Grade 5
subway	425.08	464.81	504.54	544.27	584
elementary school	586.36	608.03	629.7	651.37	673.04
middle school	447.76	520.2	592.64	665.08	737.52
high school	747.2	767.99	788.78	809.57	830.36
general hospital	1306.96	1364.48	1422	1479.52	1537.04
hypermarket	890.44	983.55	1076.66	1169.77	1262.88
park	974.44	1129.12	1283.8	1438.48	1593.16

customer	age	Desired amount	Number of people in the household	Preferred Infrastructure - 1st Choice	Preferred Infrastructure - 2nd Choice
Shin Jeong-yoon	20	150,000,000	1	subway	hypermarket
Park Joo-chan	30	300,000,000	3	elementary school	general hospital
Mr. Oh Seung-pil	40	500,000,000	2	hypermarket	subway
Voice actors	50	1,000,000,000	2	high school	park
Mr. Lee Ui-jae	60	650,000,000	4	general hospital	hypermarket
Python	70	450,000,000	2	park	subway
Deep Learning	80	300,000,000	2	general hospital	hypermarket
Pandas	90	2,000,000,000	1	park	hypermarket

End of Page

- thank you-