Anticipating Gentrification Through Data Similarity Analysis

Sojung Noh

1. Introduction

Gentrification is a process where a stagnated region is activated by the influx of external capital, entailed by the exchange of the main inhabitant. Gentrification is ambivalent in that it has both positive and negative sides. On the negative perspective, as the real estate price increases so does the price of rent fees which gradually leads to displacement of original inhabitants. On the other side, as regional value increases, the commercial activity increases with expansion of additional business opportunity, employment and construction. Despite its drawbacks, this paper focuses on the positive side of gentrification, especially on new business opportunities.

Gentrification signifies 'unseen business opportunities'. Parties seeking to fetch the fruit out of these opportunities are evident; real estate developers who aim to gain revenue out of increasing land prices, entrepreneurs who plan regional business in the blue ocean market and municipalities craving for local revitalization. These stakeholders are the crucial drivers in assessing the commercial activeness analysis, which is becoming more important these days. The City of Seoul, SEMAS(Small Enterprises & Market Service), and KT(Korea Telecommunications) have constructed a public commercial analysis service online. And startups such 'Oasis Business', 'Changupin' and 'Market Stadium' have gained attention by obtaining significant amounts of investments, which implies the importance of commercial activeness analysis.

As the utilizable data is expanding in both variety and magnitude, data-driven approach extends further to the scope of anticipation than mere analysis. There is no doubt that anticipating gentrification will become an accelerating vehicle for the city-makers mentioned previously(real-estate investors, entrepreneurs, municipalities). This paper aims to propose an anticipation model of the commercial districts of Seoul under this intent.

2. Literature Review

The literature review consists of two parts; (i) the definition & feature of gentrification , and (ii) assessment & anticipation of commercial activity. The aspects of gentrification differ from country to country and city to city. Since the scope of this paper is the city of Seoul, literature review on the first part focuses on domestic papers. Second part was reviewed by focusing on the methodological and technical papers.

(i) Definition & Feature of Gentrification

Gentrification was first defined by Ruth Glass, 1964 as a middle class invasion of labor class residences. In other words, gentrification means replacement of the low-income class to the upper-middle class, which activates the urban environment (American Heritage, 1982). Gentrification in commercial districts facilitates investments, leading the commercial facilities and amenities to a much developed level, ameliorating the decrease of population in the region (Wang, 2011).

For the case of Seoul, South Korea, gentrification tends to occur in the vicinity of the center and sub-center of the metropolitan city (Leem 2017), where commercial stores are

densely distributed in the inner side of the boulevard. This specific form is termed as ‘golmok - commercial district’ (The City of Seoul, 2015).

(ii) Assessment & Anticipation of commercial activeness

In the case of Seoul, there have been several attempts to track the progress of gentrification since the 2000's to provide guidance for urban policy making. The first index utilized the influx of college graduates and high-income professionals of each administrative district in Seoul during 1990 - 2000 (Kim, 2007). Another approach added the change of land prices to develop gentrification during 2000 - 2010 (Oh & Kim, 2017). Most recent public research on the issue done by KRIHS(Korea Research Institute for Human Settlements) has taken into account; settled population, floating population, opening/closing enterprise ratio, operating duration, franchise counts and sales in building gentrification issue (Lee, 2019).

Meanwhile, literature on the method used to predict commercial activity was investigated comprehensively along with the keyword of ‘real-estate’. This was because there was a limited number of research on ‘commercial activity prediction’. Since commercial activity and real-estate share similar features in such; it is non-moveable, fixed concept in respect to the land, and it is non-elastic in supply and demand perspective.

Prediction of the change in the commercial district was undertaken with opening enterprise counts and sales with the Deep Neural Network (Kang, 2022, Lee, 2013). In the discipline of computer sciences, there were attempts to predict chronic commercial activity with satellite image data (Zhiyan H., 2018, Wang et al., 2018). In predicting commercial real-estate prices, machine learning methods; LSTM, ARIMA and Random Forest were used with sequential data of its prices (Bae et al., 2018).

Major findings through literature review are; First, gentrification in Seoul takes place in an inner-side street. Therefore, the investigation must take place in a street dimension. Second, settled population, floating population, opening/closing enterprise ratio, operating duration, franchise counts and sales data are the influential factors in gentrification. Finally, a deep learning framework may have a high learning rate, but its drawback is that the reason behind the result is illegible. Since the major objective of this paper is the provision of a business index to an investor, demonstrating the causal factors is crucial. To make up for this drawback, this paper uses the ‘similarity analysis’ in collaborative filtering framework of the recommendation system which has clear logic of prediction and is capable of handling multi-dimensional data.

3. Research Question

The objective of this paper is to develop a prediction model which supports tentative investor’s decision making in investing in a commercial district. The scope of the investigation is unit street of Seoul which embraces the commercial cluster, ‘Golmok - commercial district’. It aims to deduce rapidly growing commercial streets with comprehensive similarity analysis by incorporating influential factors of gentrification. Therefore, the research question of this paper is; In investor perspective, is similarity analysis a relevant methodology for predicting tentative commercial districts?

4. Method

This part of the paper will be explained in six steps; scope, data gathering, data preprocessing, similarity analysis, experiment and verification.

(1) Scope

The scope of the research is the city of Seoul. Since the conditions in Seoul, the capital city, represent the rest of the region in Korea and there are rich sources of data within the region.

(2) Data Gathering

Dataset used in this research was sourced from Golmok-Seoul, an e-government service providing map base commercial district analysis which is updated every quarter where it is provided in a street base unit. 5 datasets out of 9 were selected as an influential factor of gentrification; floating population, commuting population, commercial facility, estimated sales, open/closing enterprise ratio.

[Table. Extracted Data from SEMAS]

	LIST		DATA ITEM	PERIOD	AVAIL.
1	생활인구	상권	상권코드	2017 - 2021	O
		상권배후지	상권코드	2014 - 2021	O
2	상주인구	상권	상권코드	2014 - 2021	O
		상권배후지	상권코드	2014 - 2021	O
3	직장인구	상권	상권코드	2014 - 2021	O
		상권배후지	상권코드	2014 - 2021	O
4	점포	상권	상권코드	2014 - 2021	O
		상권배후지	상권코드	2014 - 2021	O
5	집객시설	상권	상권코드	2015 - 2021	O
		상권배후지	상권코드	2017 - 2021	O
6	아파트	상권	상권코드	2014 - 2021	O
		상권배후지	상권코드	2014 - 2021	O
7	추정매출	상권	상권코드	2017 - 2021	O
		~~상권배후지~~	-	-	-
8	소득소비	~~상권~~	-	-	-
		상권배후지	상권코드	2014 - 2021	O
9	상권변화지표	상권	상권코드	2014 - 2021	O
		자치구별	자치구코드	2014 - 2021	X
		행정동별	행정동코드	2014 - 2021	△
10	상권영역	-	SHP	-	△
DATASET TO BE			상권코드	**2017-2021**	

(3) Data Preprocessing

The datasets explained above have several subcategories. Therefore, necessary data were distilled out of it. Aggregating different sources into the same district code was

undertaken, excluding NaN values. Also, quarterly data were combined into a yearly average. In this way, ‘commercial street x feature’ (1009 x 9) matrix is constructed which is a target for conducting similarity analysis.

[Image. Sample of the Dataset]

	Unnamed: 0	T_remainpop	T_commuterpop	commercial_facility_count	subwayst_count	busst_count	분기당_매출_금액	2340대_매출금액	영업기간/서울평균	폐업기간/서울평균
0	1000001	1919.0	809.0	14.0	0.0	4.0	299.028533	0.071191	0.92	1.11
1	1000003	1150.0	1079.0	23.0	0.0	3.0	264.983656	0.043465	1.07	1.04
2	1000004	1497.0	20.0	10.0	0.0	5.0	168.833421	0.039035	0.84	1.10
3	1000005	1772.0	119.0	6.0	0.0	3.0	419.180139	0.088284	0.83	0.96
4	1000006	682.0	18.0	8.0	0.0	2.0	627.819759	0.135100	0.91	1.21
...	...	...	...	...	...	...	...	...	...	...
1205	1001492	1299.0	168933.0	402.0	2.0	27.0	14409.481200	2.985283	1.38	1.14
1206	1001493	2509.0	9790.0	118.0	3.0	14.0	2325.866252	0.594730	1.22	1.03
1207	1001494	2750.0	30331.0	241.0	5.0	34.0	3807.258200	0.881768	1.45	1.25
1208	1001495	8977.0	27304.0	168.0	0.0	20.0	4789.094944	1.023787	0.85	1.03
1209	1001496	19.0	22217.0	39.0	1.0	8.0	4653.147944	1.053172	0.83	1.21

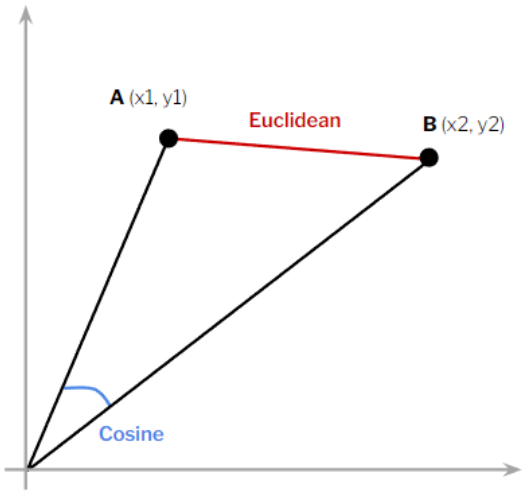
Commercial District Code

[계동길, 난계로2길, 돈화문로11가길, 명륜길, 백석동길 — 종로청계관광특구, 잠실관광특구, 강남마이스관광특구]

(4) Similarity Analysis

The ‘similarity analysis’ is the main body of the research. Similarity analysis mentioned here is derived from the recommendation system’s collaborative filtering framework. It is often used in predicting a user's rate on a single movie item. It is widely applied not only in movies, but also in music, books, news, images, webpages and many other items. In this case, it is a commercial street that we want to recommend.

[Image. Comparing Euclidean Distance & Cosine Similarity]



Example :

Distance Between [6.6, 6.2] ←→ [9.7, 9.9]

Euclidean Distance = 4.82

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

where $x = (x_1, x_2, \ldots, x_n), y = (y_1, y_2, \ldots, y_n)$

Cosine Similarity = 0.99

$$sim(x,y) = \frac{\langle x,y \rangle}{\|x\|\|y\|}$$

where $x = (x_1, x_2, \ldots, x_n), y = (y_1, y_2, \ldots, y_n)$

$$\|x\| = \sqrt{\sum_{i=1}^{n}(x_i)^2}, \langle x,y \rangle = \sum_{i=1}^{n} x_i y_i$$

To be specific, a user’s preference on a certain unexperienced item is deduced by averaging the neighbor users’ value on the item. It is the process of obtaining neighbors which requires similarity analysis. This experiment aims to calculate the similarity between a developed commercial street’s past data and multiple undiscovered commercial streets to deduce potential item.

Calculating the similarity between two parties can be done in several ways; Euclidean Similarity, Cosine Similarity, Pearson Similarity, Mean Squared Difference Similarity, Pearson-Baseline Similarity. In gaining a tentative commercial area, Cosine Similarity is deemed relevant, since it is the vector of data which matters more than absolute distance between data.

(5) Experiment

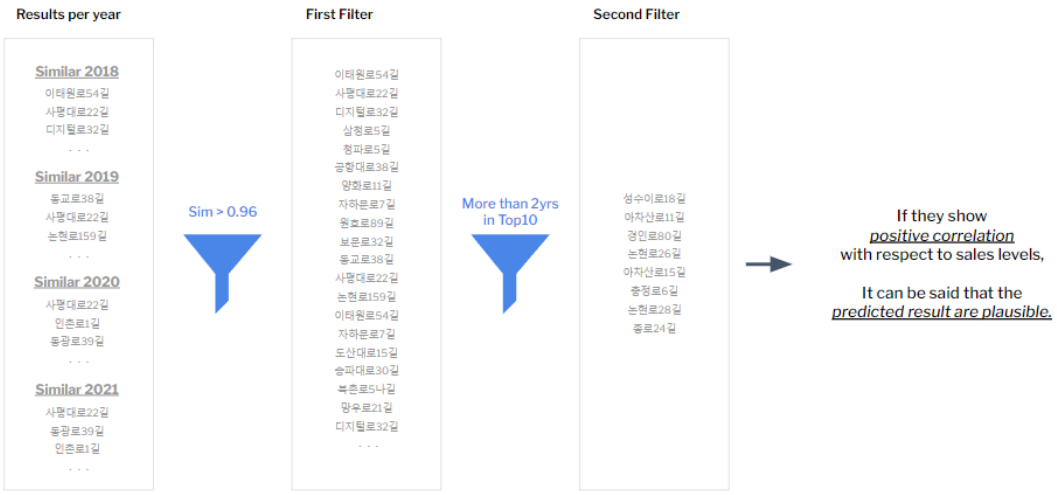
In conducting similarity analysis, a standard point should be set, since it is a relative concept. Two conditions were set to select a standard commercial street. First condition is that the estimated sales value should be bigger than 'Myeongdong-gil'. Myeong-dong had been a major commercial district in Seoul, but because of the outbreak of COVID-19 and the ban of Korean culture in China, the number of visitors drastically declined. Standard commercial street here should be the one with more awareness and sales. Second condition is that in five years(2017-2021) the growth rate should be over 10%. 10% may seem conservative, however, considering the lock-down situation during COVID-19, major commercial districts suffered downturns, which makes 10% meaningful value.

[Table 1. The standard commercial street to compare]

Type	Neighbor_name	Neighbor_code	Filter1. Estimated Sales	Filter2. Growth_rate
Type I	아차산로 15길 (성수동 북측)	1000114	0.107	82.4%
Type II	도봉로 114길 (쌍문역)	10000360	0.147	33.4%
Type III	녹사평대로 32길 (이태원 서측)	10000052	0.143	16.0%
Type IV	동교로 38길 (연남동)	1000470	0.260	10.5%

Distilled with two conditions mentioned above, four standard commercial streets were constructed; Achasan-ro 15 gil, Dobong-ro 114 gil, Noksapyeong-daero 32 gil, Donggyo-ro 28 gil. Each of these standard items and each of the 1009 items in the 'commercial x feature' matrix of five years 2017 - 2021 are compared with cosine similarity. Thus having five years of 1009 similarity values. Among those values, a street with cosine similarity value bigger than 0.95, and ones which appeared more than twice in the top 10 list of each year are filtered out as a final prediction of a tentative commercial street.

[Image 1. Process deducing the final prediction]



(6) Verification

To verify above results, results must go through verification.The hypothesis is that if predicted lists of streets exhibit positive correlation in terms of estimated sales and time, the result has enough clue to claim that they have potential for future growth in the future.

5. Result

As a result, lists of streets according to each type of standards. With these lists, analysis on correlation was conducted for every type, by analyzing 2030's sales in time.

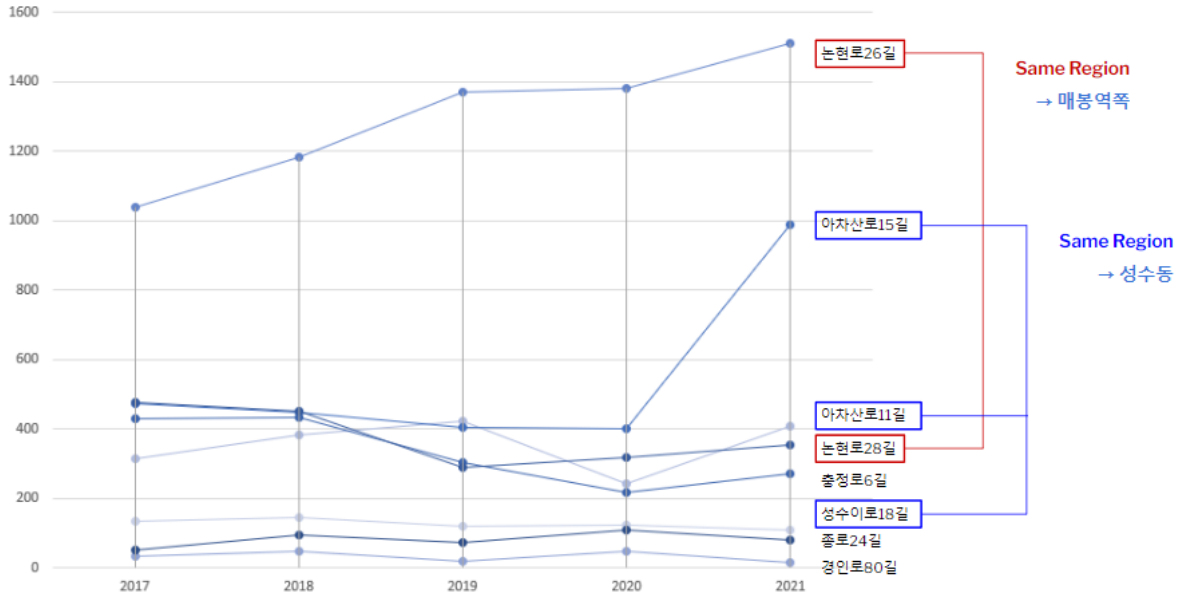
[Table 2. Prediction results]

Type	Neighbor_name	Neighbor_code
Type I	아차산로 15길 (성수동 북측)	논현로26길 / 아차산로15길 / 아차산로11길 / 논현로28길 / 충정로6길 / 성수이로18길 / 종로24길 / 경인로80길
Type II	도봉로 114길 (쌍문역)	이태원로54길 / 청파로47길 / 동교로38길 / 자하문로7길 / 녹사평대로32길 / 성지3길 / 북촌로5나길 / 인촌로24길 / 와우산로29길
Type III	녹사평대로 32길 (이태원 서측)	상도로62길 / 신흥로20길 / 서오릉로8길 / 청룡길 / 상도로61길 / 와우산로3길
Type IV	동교로 38길 (연남동)	강동대로52길 / 남현3길 / 경인로80길 / 화곡로4길 / 상암로51길 / 천호대로109길

(1) Type I, Achasan-ro 15 gil

Type I was the one which demonstrated the highest growth rate from 2017 to 2021, where the clearest sign of growth for the resulting streets were expected. The growth was apparent for most of the streets in this case. A noticeable point here is that some of the results are the streets in the vicinity, which can be grouped.

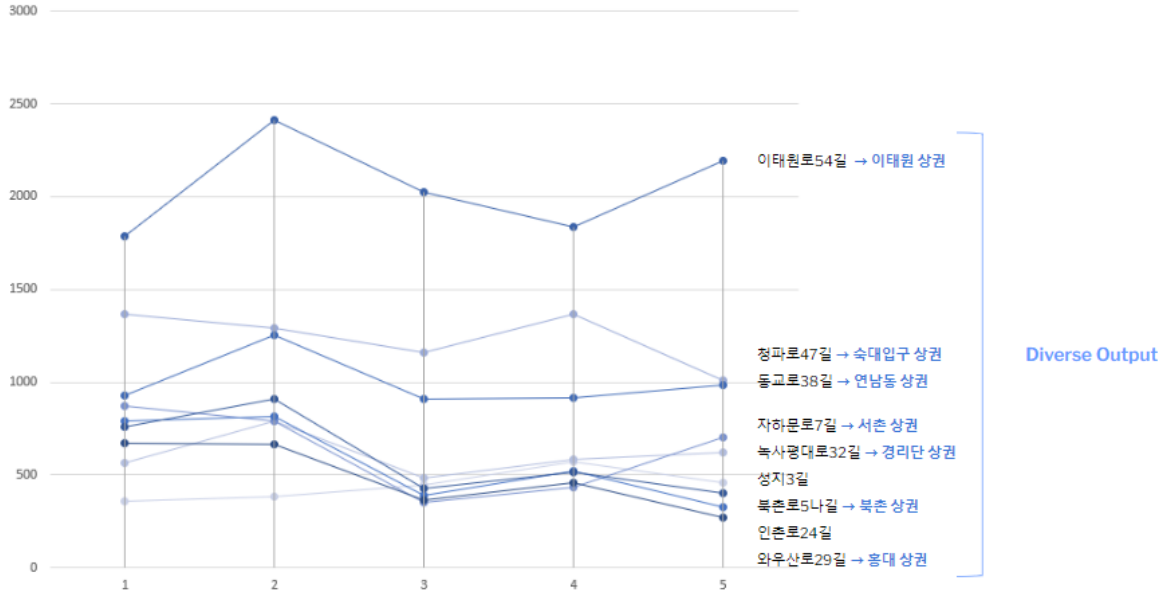
[Image. 2030's sales chart for Type I]



(2) Type II, Dobong-ro 114 gil

Second type has the second highest growth rate among four standards. The result of the second type exhibited a quasi-positive result. There are clear signs of depreciation during 2019-2020, of which COVID19 lockdown is the plausible reason behind its state. Unlike the previous result, the results are regionally dispersed, which does not show any groupings with respect to location.

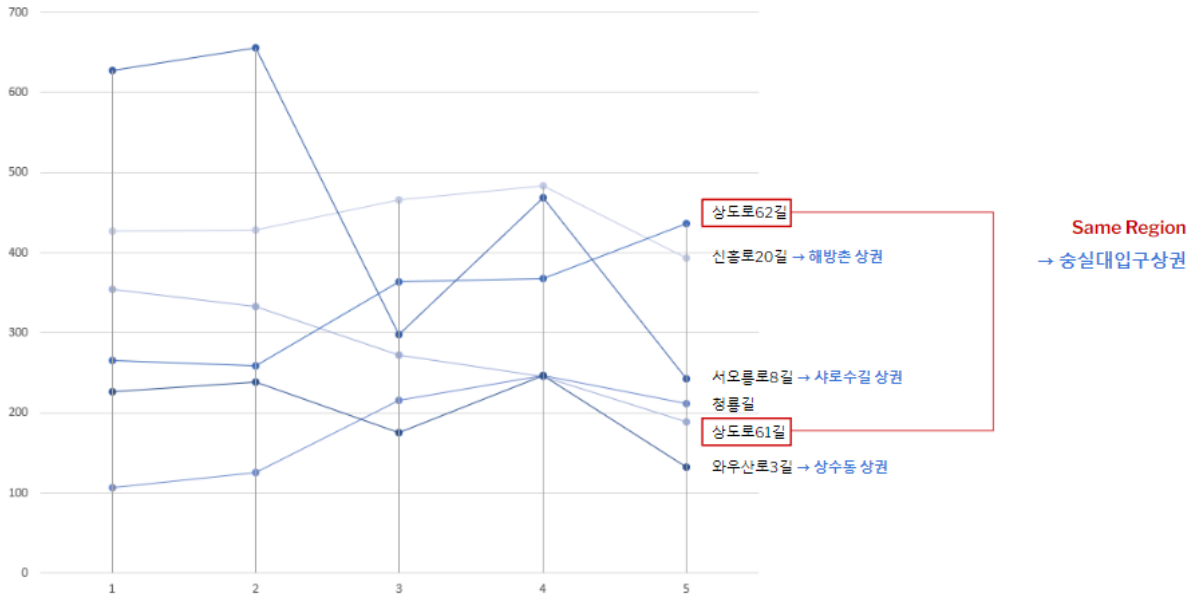
[Image. 2030's sales chart for Type II]



(3) Type III, Noksapyeong-daero 32 gil

Third type is the one with the second lowest growth rate. It shows a clear sign of negative correlation. Exhibiting high vulnerability during COVID19 lockdown period. And the results do not show much regional correlation.

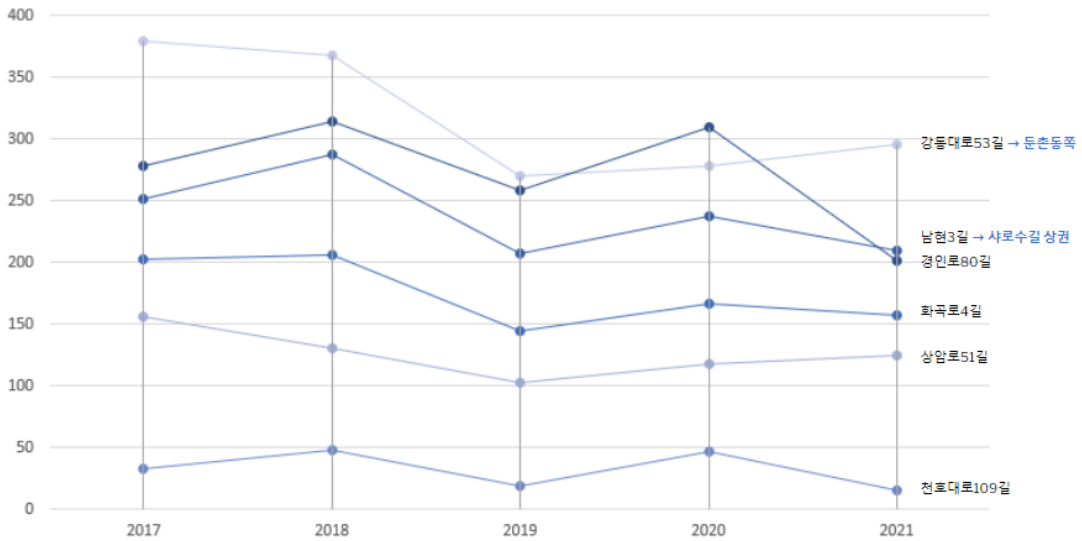
[Image. 2030's sales chart for Type III]



(4) Type IV, Donggyo-ro 38 gil

Last type has the lowest growth rate among four standards. It does not show any sign of growth during the whole period. And the results do not show much regional correlation.

[Image. 2030's sales chart for Type IV]



6. Conclusion

Even though not all results showed positive correlation, because of the first type, which exhibited an outperforming growth rate of itself as well as the result of the predicted streets, it is plausible to claim that this analysis is valid in deducing growing markets.

However, it lacks sophistication in many ways. And improvement of the problems will produce a much more apparent result than the current one. The improvements to be done are:

(a) Preprocessing Data

The data provided from SEMAS are based on 'streets' with different lengths. It needs to be calculated per meter to convey a precise analysis between items. Also data compared are currently a single time period. But it would be much better if the data were represented as 'difference'. In other words, a sequential data input would demonstrate better results.

(b) Street Types Classification.

The selected standard should be 'rising market'. In this case, Type I Itaewon is already a major commercial district, which demonstrated greatest growth as well. Standards should be refined to select baseline standards. In addition, qualitative methods such; locational, atmospheric, and emotional clustering would also help in deducing standard types.

(c) Verification Index

2030's sales plot was expected to show a clear sign of positive correlation, but the data were contaminated, showing big fluctuation between 2019-2020 because of COVID19 lockdown. Therefore another metric is necessary as a proof, i.e. SNS mention counts, floating population and many other indexes which may provide guidance in verifying the result.

Reference.

- Wang, S. W. H., 2011. "Commercial gentrification and entrepreneurial governance in Shanghai: A case study of Taikang road creative cluster", Urban Policy and Research, 29(4): 363-380.
- Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001, April). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web* (pp. 285-295).
- 오창화·김영호, 2017. "공간 회귀와 공간 필터링을 이용한 서울시 젠트리피 케이션의 발생 원인 및 특징 분석", 한국도시지리학회, 19(3): 71-86.
- 김걸, 2007. "서울시 젠트리피케이션의 발생원인과 설명요인", 한국도시지리 학회지, 10(1): 37-49.
- 박아름, 2016. "상업가로의 젠트리피케이션 과정 및 임대료 영향요인 분석: 경리단길을 사례로", 한양대학교 대학원 석사학위논문.
- 허자연, 2015. "서울시 상업가로 변천과정에 관한 연구", 서울대학교 대학원 박사학위논문.
- 이진희, 임상연, 박종순, & 이왕건. (2019). 젠트리피케이션 지표 개발과 활용 방안. 국토정책 *Brief*, 1-8.
- 강민경. (2022). 서울시 상권변화 유형도출 및 예측모델을 활용한 변화예측 (Doctoral dissertation, 한양대학교).
- 배성완, & 유정석. (2018). 머신 러닝 방법과 시계열 분석 모형을 이용한 부동산 가격지수 예측. 주택연구, *26*(1), 107-133.
- 서울시, 2016. 「젠트리피케이션 데이터 분석결과 보고」, 서울: 서울시 정보 기획관.
- 서울시, 2015. 「서울시 젠트리피케이션 종합대책」, 서울: 서울연구원