

2022 서울과학기술대학교 데이터마케팅
팀프로젝트

서울시 상권 매출액 예측

결과 발표 #22.05.26 Thu

232, Gongneung-ro, Nowon-gu, Seoul, Republic of Korea

7조

김동주, 김범수, 김석희, 박지혜

Agenda

1. 분석 배경 & 목적

2. 데이터 획득

3. 데이터 분석

4. 기대 효과와 의의 및 한계점, 추후 개선 방안

5. 참고 문헌 & github repo

1. 분석 배경 & 목적

1-1. 분석 배경

- “코로나 직격탄... 음식점업 폐업률 18.1% 달해”

(22.02, 문화일보)

- “코로나로 외식업 연매출 평균 683만원 줄어”

(22.05, 동아일보)

매년 높은 수치를 기록하는 폐업률,
예측하기 힘든 매출량으로
창업을 망설이는 예비창업자 多

서울시 자치구별 폐업률

출처 : 행정안전부 지방인허가 데이터
폐업률=폐업업체/(총 영업업체+폐업업체)

구별	2019	2020	전년 대비 증감(%P)
강남구	11.2%	11.1%	-0.1%
강동구	9.7%	9.2%	-0.5%
강북구	7.1%	7.2%	0.0%
강서구	8.8%	9.5%	0.7%
관악구	9.3%	10.2%	0.9%
광진구	9.3%	9.0%	-0.3%
구로구	9.6%	8.8%	-0.9%
금천구	8.3%	7.7%	-0.6%
노원구	10.8%	9.0%	-1.8%
도봉구	9.3%	8.1%	-1.2%
동대문구	10.0%	8.6%	-1.4%
동작구	8.3%	12.4%	4.1%
마포구	12.9%	10.9%	-2.0%
서대문구	13.0%	10.4%	-2.6%
서초구	10.0%	10.0%	0.0%
성동구	4.3%	4.8%	0.5%
성북구	5.8%	6.4%	0.6%
송파구	10.2%	9.9%	-0.3%
양천구	10.9%	10.1%	-0.8%
영등포구	7.2%	4.5%	-2.7%
용산구	5.5%	5.7%	0.2%
은평구	12.0%	10.7%	-1.3%
종로구	4.2%	5.2%	1.0%
중구	3.1%	3.0%	-0.1%
중랑구	7.6%	7.5%	-0.1%
총합	8.9%	8.5%	-0.4%

1. 분석 배경 & 목적

1-2. 분석 목적



회귀 모델 및 공공 상권 정보 데이터를 활용해,
상권 매출액 예측

2. 데이터 획득

2-1. 데이터 획득

우리 마을 가게 상권분석 서비스



우리가 자주 접하고 이용하는 골목상권

- (기존 존재 서비스) 서울시에서 제공하는 '우리마을 가게 상권분석 서비스' :
과밀화된 골목상권에 창업하고자 하는 소상공인들을 위해, 창업위험도를 알려주는 서비스.
- 과밀지수, 활성화지표 등 골목상권별 포화도와 활성화도, 안정성 등을 예측하여 창업 시 위험도를 쉽게 알 수 있도록 여러 비교분석 지표를 개발 및 도입하여 예비창업자의 위험 부담을 줄이고자 출범한 서비스.

2. 데이터 획득

2-1. 데이터 획득

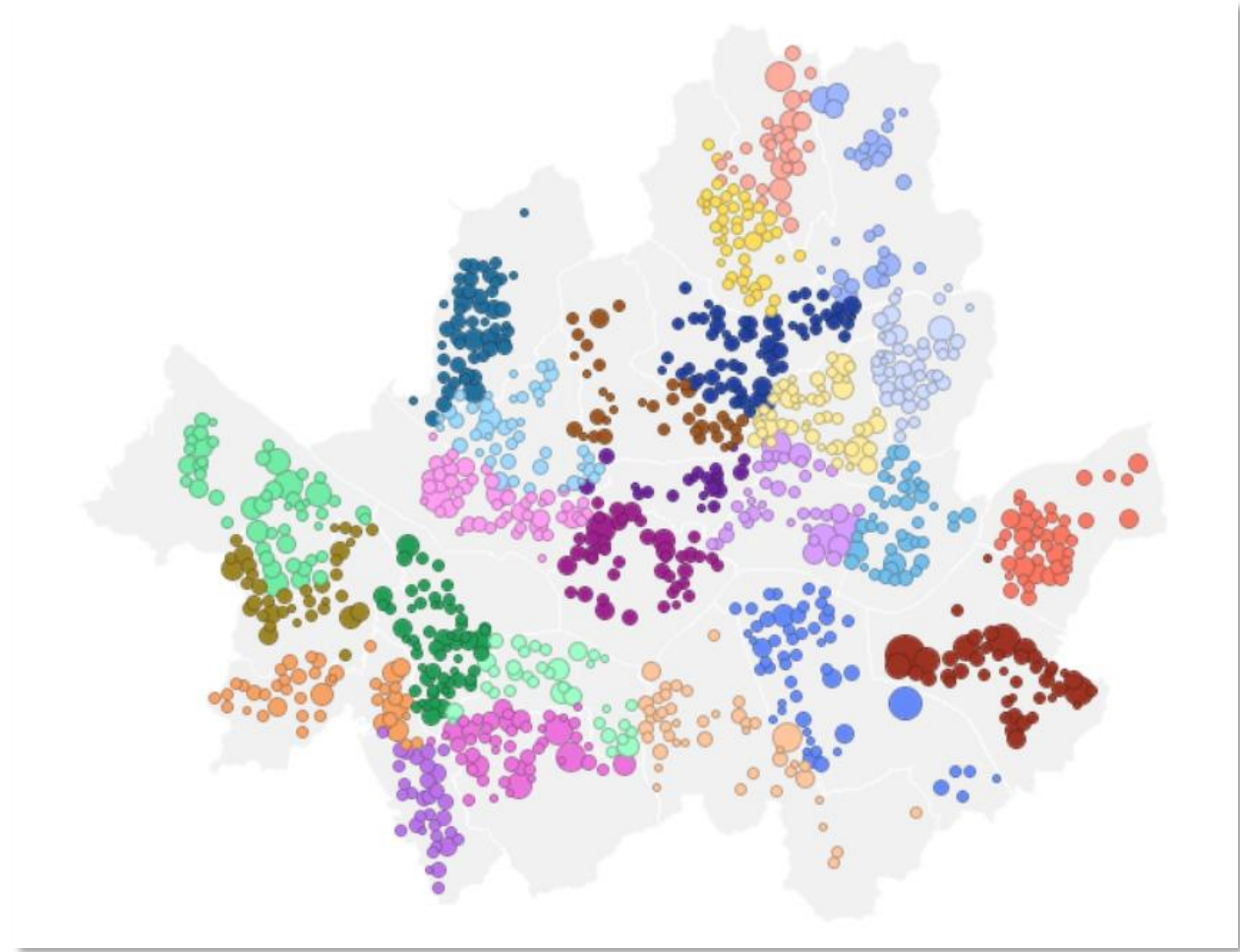
- 해당 서비스는 **공공데이터**로 개방되어 있는 **19개의 데이터셋**을 기반으로 분석 및 구축되었으며, **본 프로젝트에 해당 데이터셋들을 사용.**

일반행정	서울시 우리마을가게 상권분석서비스(상권영역)	공공데이터 ●
일반행정	서울시 우리마을가게 상권분석서비스(상권-아파트)	공공데이터 ●
일반행정	서울시 우리마을가게 상권분석서비스(자치구별 상권변화지표)	공공데이터 ●
일반행정	서울시 우리마을가게 상권분석서비스(상권-생활인구)	공공데이터 ●
일반행정	서울시 우리마을가게 상권분석서비스(상권-직장인구)	공공데이터 ●
일반행정	서울시 우리마을가게 상권분석서비스(상권배후지-직장인구)	공공데이터 ●
일반행정	서울시 우리마을가게 상권분석서비스(상권배후지-소득소비)	공공데이터 ●
일반행정	서울시 우리마을가게 상권분석서비스(상권배후지-추정매출)	공공데이터 ●
일반행정	서울시 우리마을가게 상권분석서비스(상권배후지-아파트)	공공데이터 ●
일반행정	서울시 우리마을가게 상권분석서비스(행정동별 상권변화지표)	공공데이터 ●
일반행정	서울시 우리마을가게 상권분석서비스(상권-상권변화지표)	공공데이터 ●
일반행정	서울시 우리마을가게 상권분석서비스(상권-점포)	공공데이터 ●
일반행정	서울시 우리마을가게 상권분석서비스(상권배후지-점포)	공공데이터 ●
일반행정	서울시 우리마을가게 상권분석서비스(상권-집객시설)	공공데이터 ●
일반행정	서울시 우리마을가게 상권분석서비스(상권배후지-집객시설)	공공데이터 ●
일반행정	서울시 우리마을가게 상권분석서비스(상권배후지-생활인구)	공공데이터 ●
일반행정	서울시 우리마을가게 상권분석서비스(상권배후지-상주인구)	공공데이터 ●
일반행정	서울시 우리마을가게 상권분석서비스(상권-상주인구)	공공데이터 ●
일반행정	서울시 우리마을가게 상권분석서비스(상권-소득소비)	공공데이터 ●

3. 데이터 분석

3-1. EDA

- 총 19개의 데이터셋 중, 프로젝트의 목표인
매출액 예측에 도움이 될 수 있는,
주요 데이터셋으로 생각되는
데이터셋의 분포를 시각화를 통해 선제 분석.



매출액 데이터의 지역구 별 분포 파악
(원의 크기는 매출액에 비례)

- 개방된 19개의 공공데이터셋 중, 15개의 데이터셋을 채택하여 사용하였다.
 - 상권의 매출액을 예측하는 데 있어, 활용할 수 있는 데이터셋을 선별하여 사용하였다.
- 각 데이터셋이 포함하고 있는 변수들을 병합하기 위해, 아래 3개의 행정정보 데이터를 통해 병합하였다.
 - 상권코드, 분기, 년도 (해당 데이터셋들은 대부분 17~21년도의 데이터로 이루어져 있다.)
 - 병합 결과, 38개의 feature 존재.

상권배후지-추정매출

상권구분코딩권구분코딩	상권코딩	상권코딩명	비사업종교/비사업종교등기대매출금/기대매출금/주중매출비율/주말매출비율/일별매출비/간대별매출비/상대매출비/월여성비/월경대별매출비/주중매출금/주말매출금/일/일별매출금/간대별매출금/상대매출금/여성성매출금/월경대별매출금/주중매출금/주말매출금/주/일별매출금/간대별매출금/상대매출금/여성성매출금	점포수
-------------	------	-------	---------------------------------------------------------------------------------------------------------------------------------------------------------------	-----

상권-추정매출

상권구분코드 상권구분코드	상권코드	상권코드명	비즈니스업종코드 비즈니스업종코드	기업매출금 기업매출금	기업매출건수 기업매출건수	중대매출비율 중대매출비율	일별매출비율 일별매출비율	연간매출금 연간매출금	연간매출건수 연간매출건수	주말매출건수 주말매출건수	월별매출건수 월별매출건수	간대매출건수 간대매출건수	성인형태별매출 성인형태별매출	점포수
------------------	------	-------	----------------------	----------------	------------------	------------------	------------------	----------------	------------------	------------------	------------------	------------------	--------------------	-----

상권배후지-소득소비

[illegible]

상권-소득소비

상권연구보고의 연구번호코드	상권코드	상권코드명	행정관소득금소득구간코드	지출총금액	리펀지출총액	신발지출총액	유동지출총액	비지출총액	교통	여가	문화	교육	기타지출총금액
----------------	------	-------	--------------	-------	--------	--------	--------	-------	----	----	----	----	---------

상권배후지-아파트

[illegible]

상권-아파트

상 권구분코드의 권구분코드 상 권코드 상 권코드명 아파트단지수 66제곱미터이하 66제곱미터 99제곱미터 132제곱미터 165제곱미터 가격1억미만 트 가격1억세 가격4억세 대 아파트평균 면적 아파트평균 시가

상권배후지-생활인구

[illegible]

상권-생활인구

[illegible]

상권배후지-직장인구

[illegible]

상권-직장인구

[illegible]

상권배후지-상주인구

상권구분코드	상권구분코드	상권구분코드	상권구분코드명	중앙주인구수	지방주인구수	지방주인구수	대별상주인	영대별상주인	영대발상주	영대발상주	중가구수	아파트가구수	아파트가구수	아파트가구수
--------	--------	--------	---------	--------	--------	--------	-------	--------	-------	-------	------	--------	--------	--------

상권_상주인구

[illegible]

상권배후지-집객시설

[illegible]

상권-집객시설

[illegible]

상권영역

[illegible]

3. 데이터 분석

3-2. 데이터 전처리

- 1) missing data 처리
 - 15개의 데이터셋들을 파악할 때에 한 데이터셋에서 변수 4개에 대해,
17년도 1분기 데이터가 없는 것을 발견.
 - 나머지 2~4분기 데이터셋을 비교해본 결과, 데이터의 분포가 대체로 비슷하여
2분기의 데이터를 1분기의 데이터로도 사용.

3. 데이터 분석

3-2. 데이터 전처리

Regression 진행 전 절차로, feature extraction 진행.

- 2) 상관성 분석 - correlation 계수
 - 15개의 데이터셋을 병합하기 전, 각 데이터셋에 대해 상관성 분석을 진행.
 - 한 데이터셋 내의 각 변수 간의 correlation 계수를 구하여 변수 선별을 일차적으로 진행.
 - 모든 데이터셋에 대해 correlation 값을 구해본 결과, 각 값들은 대부분 절댓값이 0에 가깝거나, 0.8 이상으로 나오는 두 가지의 경우로 나뉘었다.
 - 이에 후자의 경우, 대표되는 변수만을 선별하여 regression에 사용하고자 하였다.
 - Correlation의 절댓값이 높다면, 해당 변수가 타 변수를 잘 설명할 수 있다는 특성을 활용.

Ex) 상권_집객시설 데이터셋에서의 상관성 분석 – correlation 계수 0.8 이상 경우 존재 X

	관공서_수	은행_수	의료기관_수	교육기관_수	상점_수	극장_수	숙박_시설_수	교통기관_수
관공서_수	1	0.161663	0.105761	0.037052	0.105043	0.292254	-0.064351	0.145743
은행_수	0.161663	1	0.084998	-0.022163	0.051801	-0.090391	0.037799	0.031163
의료기관_수	0.105761	0.084998	1	0.174153	0.107557	-0.10071	0.051713	0.391193
교육기관_수	0.037052	-0.022163	0.174153	1	0.014804	0.048487	0.003037	0.175366
상점_수	0.105043	0.051801	0.107557	0.014804	1	-0.034796	-0.013142	0.072291
극장_수	0.292254	-0.090391	-0.10071	0.048487	-0.034796	1	0.533648	0.324992
숙박_시설_수	-0.064351	0.037799	0.051713	0.003037	-0.013142	0.533648	1	0.131955
교통기관_수	0.145743	0.031163	0.391193	0.175366	0.072291	0.324992	0.131955	1

Ex) 상권배후지_소득소비 데이터셋에서의 상관성 분석 – correlation 계수 0.8 이상 경우 존재 다수, 변수 선별 진행

	소득_구간_코드	지출_총금액	식료품_지출_총금액	의류_신발_지출_총금액
소득_구간_코드	1	-0.138888	-0.225858	-0.124707
지출_총금액	-0.138888	1	0.981693	0.997665
식료품_지출_총금액	-0.225858	0.981693	1	0.970945
의류_신발_지출_총금액	-0.124707	0.997665	0.970945	1

지출_총금액 변수가
식료품_지출_총금액 변수 및
의류_신발_지출_총금액 변수를 대표

1) 데이터셋 분할

- Training dataset
 - 17년도 1분기 ~ 21년도 2분기 데이터셋
- Validation dataset
 - 21년도 3분기 데이터셋
- Testing dataset
 - 21년도 4분기 데이터셋

3. 데이터 분석

3-3. 모델링

2) 앞서 전처리 및 import한 데이터셋을 통해 Linear regression 진행

MAE, RMSE

- 오차 계산 방법론, 대체로 값이 작을수록 좋은 성능

R_square

- 결정계수 (원자료에 대한 회귀선의 설명력)
- 값이 1에 가까울수록, 예측이 실제를 비슷하게 설명

$$R^2 = \frac{SSR}{SST} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

3. 데이터 분석

3-3. 모델링

2) 앞서 전처리 및 import한 데이터셋을 통해 Linear regression 진행

```
performance for TRAIN-----  
train MAE : 2191918759.671748  
train RMSE : 3554464401.714925  
train R_square : 0.5614943026057684  
performance for VAL-----  
val MAE : 2279884116.456788  
val RMSE : 3888471162.715774  
val R_square : 0.4730787755886081
```

- 그 결과, 약 **0.47**의 R_square 값을 validation dataset에 대한 값으로 획득.
- 성능을 높이기 위해, **dimension**을 줄이기 위한 후처리를 진행 후 다시 학습을 진행.

3. 데이터 분석

3-3. 모델링

3) vif, p-value, PCA를 통한 다중공산성 측정 및 feature extraction

VIF Factor		features
0	20.8	t_household
1	16.3	income_range
2	-0.0	t_spent
3	1.9	office
4	1.5	bank
5	2.9	medical
6	1.3	edu
7	1.2	store
8	4.0	traffic
9	4.8	office_b
10	6.4	bank_b

3-1) 다중공산성 측정 알고리즘인 VIF를 통해 도출한 계수가 10이상인 경우, 해당 feature들에 대해서 extraction을 진행해볼 가치가 있다.

앞서 전처리한 데이터셋에 대해 VIF를 적용한 결과,

그 계수가 10을 넘기는 변수들을 제거 및 linear regression 적용 후,

앞서 사용한 비교분석 지표인 RMSE, R_squared 값의 변화를 관찰.

3. 데이터 분석

3-3. 모델링

3) vif, p-value, PCA를 통한 다중공산성 측정 및 feature extraction

```
performance for VAL-----  
val MAE : 2279884116.456788  
val RMSE : 3888471162.715774  
val R_square : 0.4730787755886081
```



```
performance for vifVAL-----  
val MAE : 2346696084.4283543  
val RMSE : 4017940740.1925507  
val R_square : 0.43740614847863246
```

그 결과, 성능이 감소 (오차값 증가 및 R2 계수 감소) 하여, VIF를 적용해 변수를 제거하는 과정을 배제.

3. 데이터 분석

3-3. 모델링

3) vif, p-value, PCA를 통한 다중공산성 측정 및 feature extraction

3-2) 다른 다중공산성 측정 알고리즘인 p-value를 통한 dimension reduction 시도

- 비슷한 방식으로, P 값이 0.05 를 넘어가면 변수 제거를 고려해볼만 하다.

	coef	std err	t	P> t
t_household	3.435e+05	6.09e+04	5.640	0.000
income_range	1.262e+08	1.68e+07	7.493	0.000
t_spent	0.3484	0.045	7.783	0.000
office	1.948e+08	3.4e+07	5.729	0.000
bank	1.529e+09	5.1e+07	29.961	0.000
medical	9.572e+08	1.9e+07	50.381	0.000
edu	-9.496e+07	4.98e+07	-1.905	0.057

3. 데이터 분석

3-3. 모델링

3) vif, p-value, PCA를 통한 다중공산성 측정 및 feature extraction

마찬가지로, 변수 제거 후 학습을 진행해본 결과,

```
performance for VAL-----  
val MAE : 2279884116.456788  
val RMSE : 3888471162.715774  
val R_square : 0.4730787755886081
```



```
performance for pVAL-----  
val MAE : 2280845288.454243  
val RMSE : 3897053072.9069104  
val R_square : 0.47075036399211334
```

다시 한 번 성능이 감소 (오차값 증가 및 R2 계수 감소) 하여, p-value를 적용해 변수를 제거하는 과정을 배제.

3. 데이터 분석

3-3. 모델링

3) vif, p-value, PCA를 통한 다중공산성 측정 및 feature extraction

3-3) PCA를 통해 feature extraction 적용 후 학습 진행. (n_components = 0.95 로 설정)

```
Original shape: (19598, 38)
Reduced shape: (19598, 27)
```

38개에서 27개로, 총 11개의 변수 제거.

```
performance for VAL-----
val MAE : 2279884116.456788
val RMSE : 3888471162.715774
val R_square : 0.4730787755886081
```



```
performance for pcaVAL-----
p MAE : 2324603286.4577384
p RMSE : 3918378208.4358964
p R_square : 0.46494228283712646
```

다시 한 번 성능이 감소 (오차값 증가 및 R2 계수 감소) 하여, PCA를 적용해 변수를 제거하는 과정 또한 배제.

3. 데이터 분석

3-3. 모델링

4) Training dataset에 대한 모델 평가

앞서 3가지의 dimension reduction 방법론을 통해 성능을 증가시키지 못하여,
Correlation 계수를 구한 후 전처리 과정을 거친 **초반의 데이터셋을 평가에 사용.**

```
performance for VAL-----  
val MAE : 2279884116.456788  
val RMSE : 3888471162.715774  
val R_square : 0.4730787755886081
```

```
performance for TRAIN-----  
train MAE : 2191918759.671748  
train RMSE : 3554464401.714925  
train R_square : 0.5614943026057684  
performance for test-----  
test MAE : 2305539790.4288864  
test RMSE : 3755264948.9428883  
test R_square : 0.5111749733098415
```

Clustering 및 군집별 학습/예측

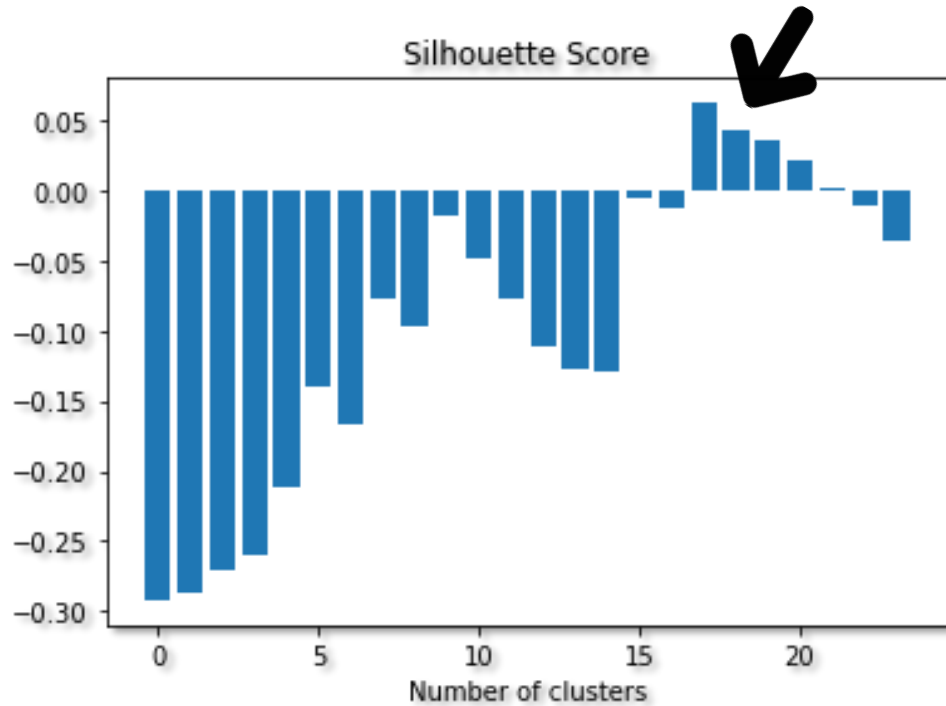
- K-means를 활용한 군집화 진행 후, 비슷한 특성을 가진 각 군집에 대해 linear regression을 별개로 진행하여 발전된 결과 도출을 도모.
- 17년도 1분기 데이터에 대한 클러스터링 진행, 타 시기의 데이터셋에 대해 같은 상권코드를 가진 데이터에 같은 label을 부여.
 - 상권 코드가 같은 상권은 같은 클러스터를 가질 것이라 판단

Clustering 및 군집별 학습/예측

- MinMax Scaling 진행
- 여러 k 값에 대한 K-Means clustering 후, 군집 평가 지표인 silhouette score를 통해 k 결정
- 학습 및 평가
- 클러스터링 라벨 별 데이터셋 각각에, 앞서 진행한 dimension reduction 시도 및 평가
 - VIF, p-value, PCA
- label 별 데이터셋 각각 매출 최종 예측

3. 데이터 분석

3-4. 후처리



- DBSCAN을 사용해 실루엣스코어 측정
- 그 값이 두 번째로 높으며 가장 적절히 나뉜 군집화 결과 채택
 - Agglomerative Clustering
(n_clusters=2, linkage="ward")
- 같은 상권코드를 가지는 데이터에 같은 label을 부여,
각 군집별로 (2개의 클러스터, label = 0 / label = 1)
dimension reduction 및 평가 등 이후 과정 진행

3. 데이터 분석

3-4. 후처리

- **Label == 0** 인 dataset (클러스터) 에 대해 학습 및 평가를 진행해본 결과,
MAE는 다소 증가하였으나 RMSE 감소 및 R2 증가 결과를 도출하였다.

```
performance for VAL-----  
val MAE : 2279884116.456788  
val RMSE : 3888471162.715774  
val R_square : 0.4730787755886081
```



```
performance for VAL-----  
val MAE : 2607754739.790089  
val RMSE : 3667465232.5989814  
val R_square : 0.589685229851524
```

3. 데이터 분석

3-4. 후처리

- 앞서 사용하였던 dimension reduction 방법론을 적용한 결과, 평가 결과에 있어 다시 한 번 더 나은 결과를 도출하지 못했다.
- Dimension reduction을 적용하지 않은 label == 0 의 데이터셋에 대한 최종 예측 결과,

```
performance for test-----  
test MAE : 2305539790.4288864  
test RMSE : 3755264948.9428883  
test R_square : 0.5111749733098415
```



```
performance for test-----  
test MAE : 2592758288.2099996  
test RMSE : 3693424832.8788104  
test R_square : 0.6048680126082132
```

- 오차값이 소폭 증가하였으나 R2가 상당히 큰 폭으로 증가한 결과물을 도출할 수 있었다.

3. 데이터 분석

3-4. 후처리

- Label == 1 인 데이터셋에 대해 같은 과정을 반복한 결과,

(p-value 이용 변수 제거 후 VIF 이용 변수 제거) 실험이 평가 성능을 증가시켜 이를 적용한 데이터셋에 대해 최종 예측을 진행.

```
performance for test-----  
test MAE : 2305539790.4288864  
test RMSE : 3755264948.9428883  
test R_square : 0.5111749733098415
```



```
performance for test-----  
test MAE : 2071024859.3063238  
test RMSE : 3664658386.4731975  
test R_square : 0.33699369737036255
```

- 그 결과, 오차값이 다소 감소하였으나 R2가 상당히 큰 폭으로 감소하는 결과를 도출하였다.

4. 기대 효과와 의의 및 한계점, 추후 개선 방안

4-1. 기대 효과와 의의 및 한계점

기대 효과

- 기존에 존재하는 위험도 예측 서비스를 넘어, 전반적인 예상 매출액을 알려줄 수 있는 서비스 제공

의의 및 한계점

- Label로 0을 가진 특성의 데이터들에 대해서, 예측을 준수하게 해낼 수 있는 모델 구축에 성공
- Label로 1을 가진 특성의 데이터들에 대해서, 동일 모델이 예측을 잘 해내지 못함

4. 기대 효과와 의의 및 한계점, 추후 개선 방안

4-2. 한계점, 추후 개선 방안

한계점 및 추후 개선 방안

- 사용한 데이터셋의 feature의 개수가 많아, **각 클러스터에 대한 특징 파악에 실패했다.**

- 시각화를 통해 직관적으로 이를 파악해본 결과,

- 푸른색 점으로 표시된 cluster 0의 경우**

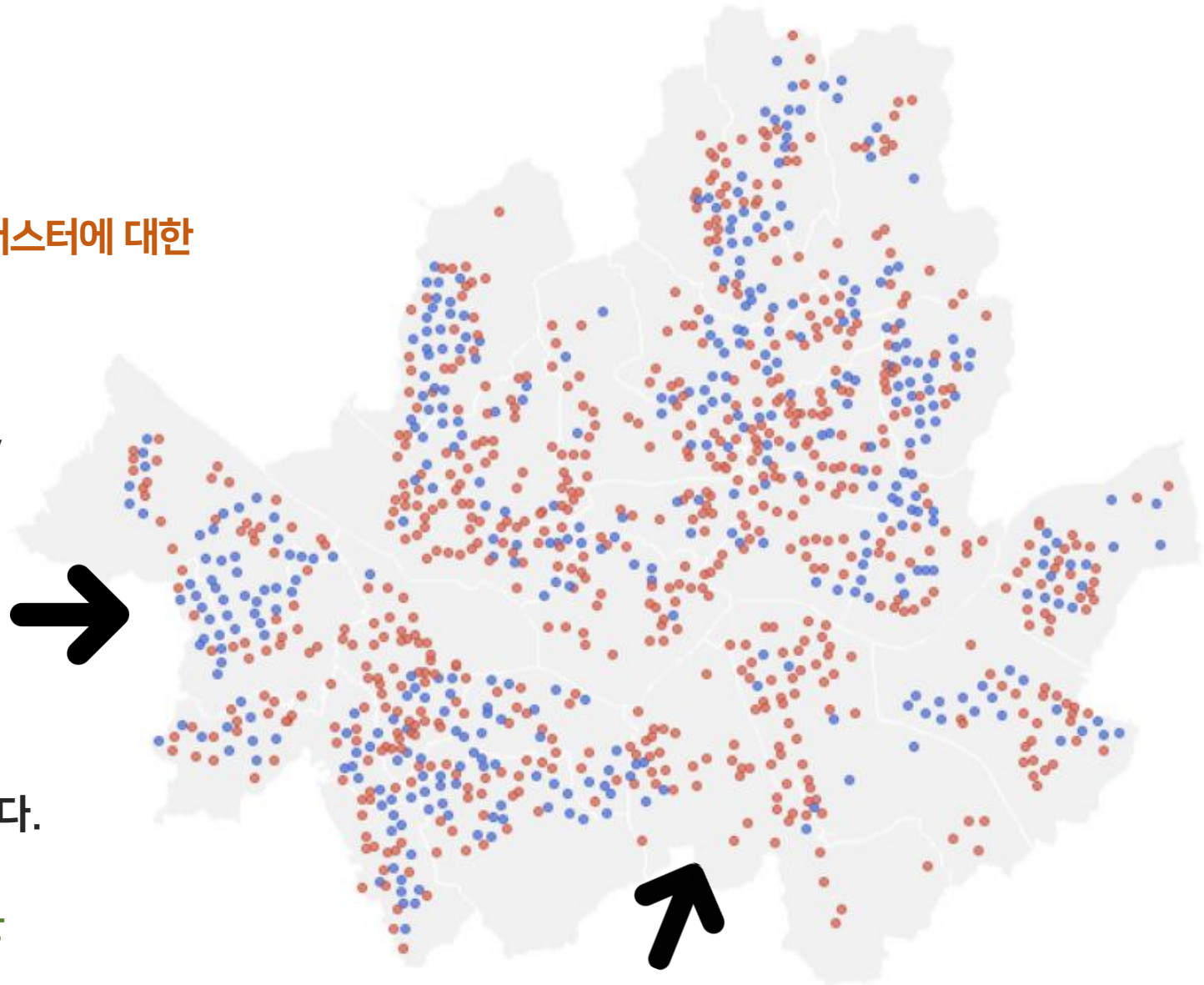
- 강서구 등을 포함한 주거단지의 경우가 많았고**

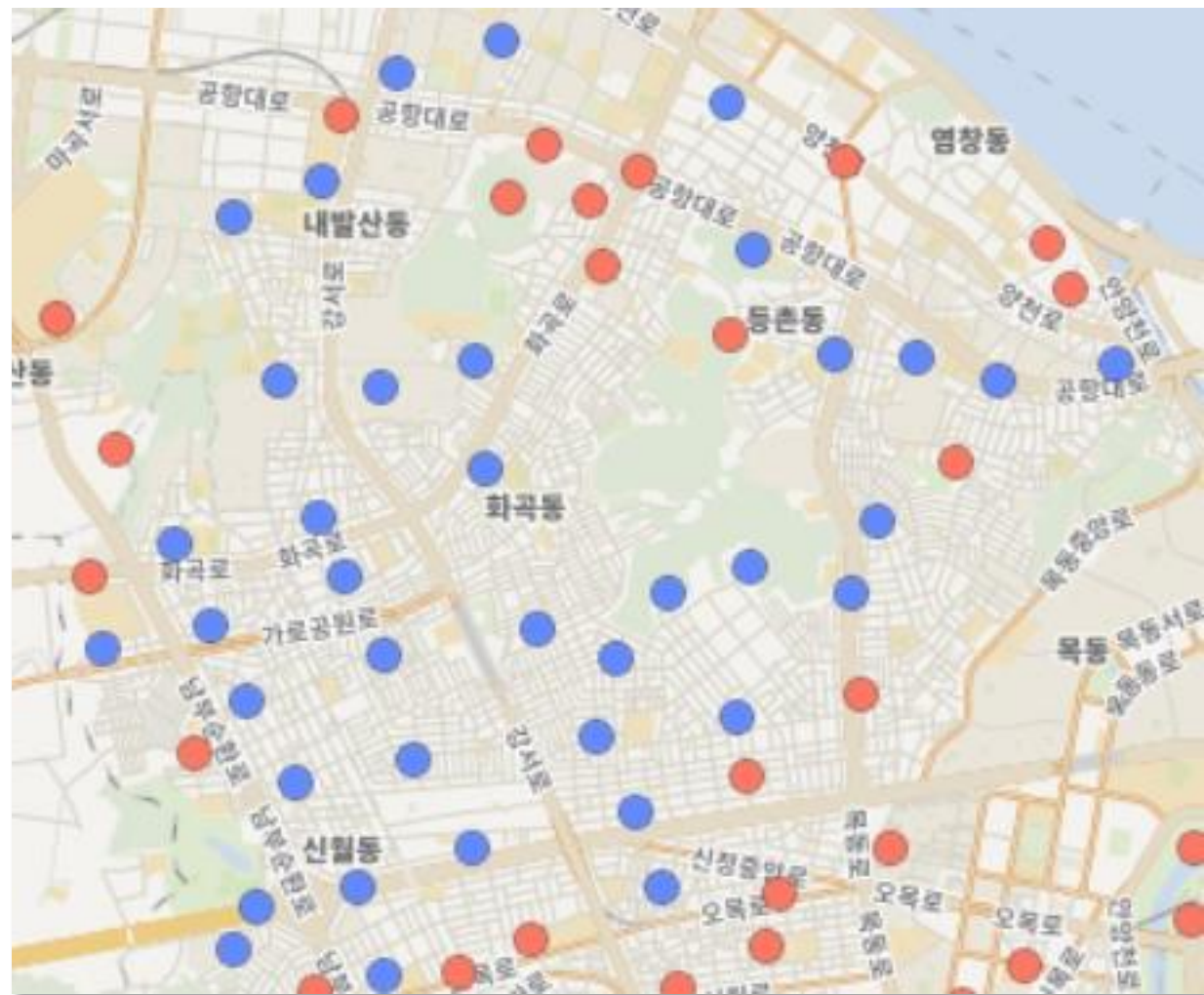
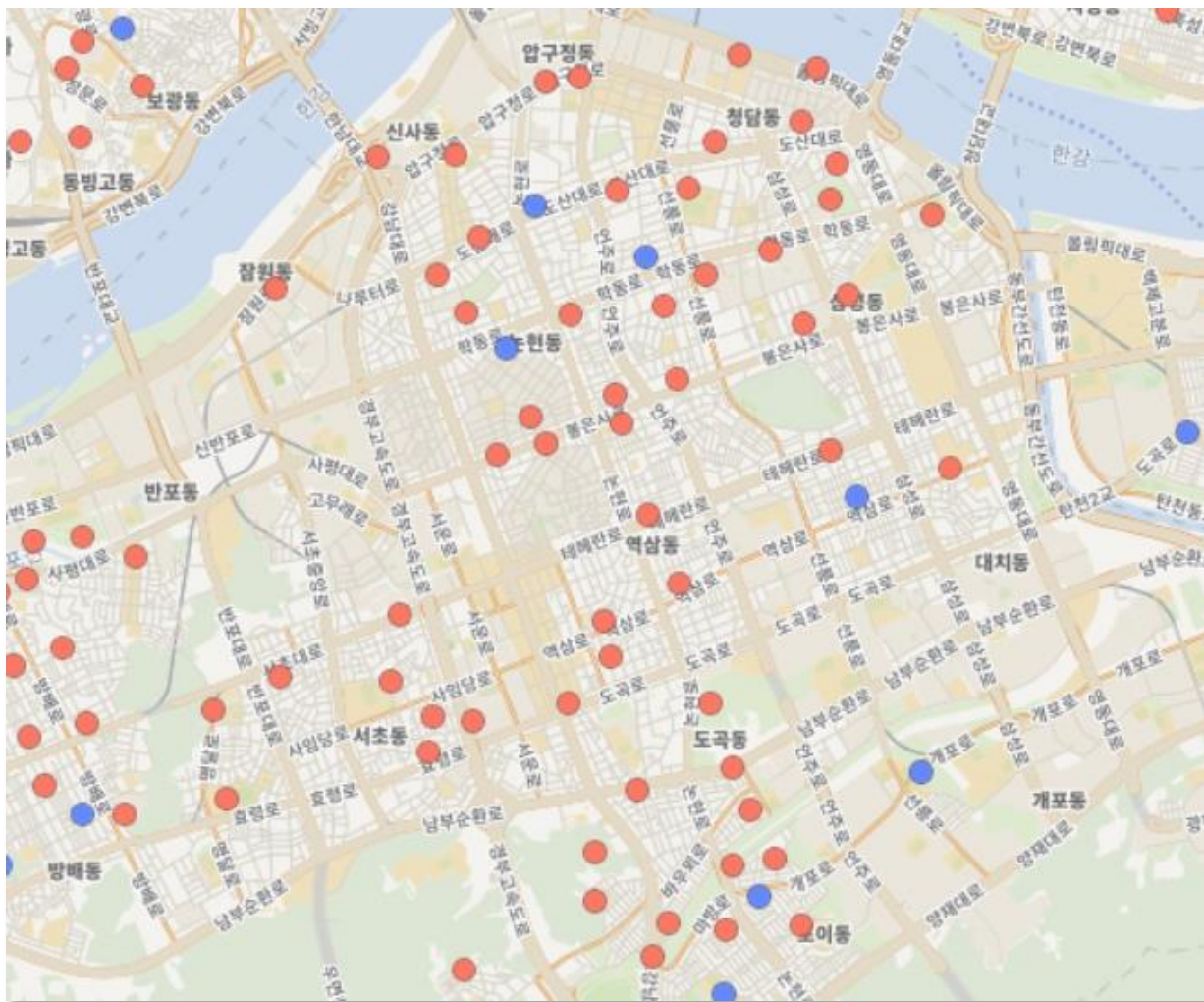
- 붉은색의 점으로 표시된 cluster 1의 경우**

- 강남구 등을 포함한 상가/직장단지의 경우가 많았다.**

- 추후 연구를 통해, **해당 클러스터들이 가지는 특성을**

- 더 자세히 파악할 수 있다면, 소비자 맞춤 예측 정보를 제공할 수 있으리라 생각된다.**





클러스터 0의 데이터가 밀집된 강남구 (좌) 와 클러스터 1의 데이터가 밀집된 강서구 (우)

5. 참고 문헌 & github repo

참고 문헌

“코로나 직격탄... 음식점업 폐업률 18.1% 달해” (22.02, 문화일보)
(<http://www.munhwa.com/news/view.html?no=2022022301072003355002>)
“코로나로 외식업 연매출 평균 683만원 줄어” (22.05, 동아일보)
(<https://www.donga.com/news/article/all/20220523/113565577/1>)

Github repo

- https://github.com/sukkykim/2022_Seoultech_DataMining_team7

THANK YOU