Final Project

젠트리피케이션 예측

훈 련 과 정 명 : 데이터베이스(DB)활용을 위한 빅데이터분석가 양성과정

훈 련 기 간 : 2020.05.12 ~ 2020.10.30 (824시간/103일)

목차

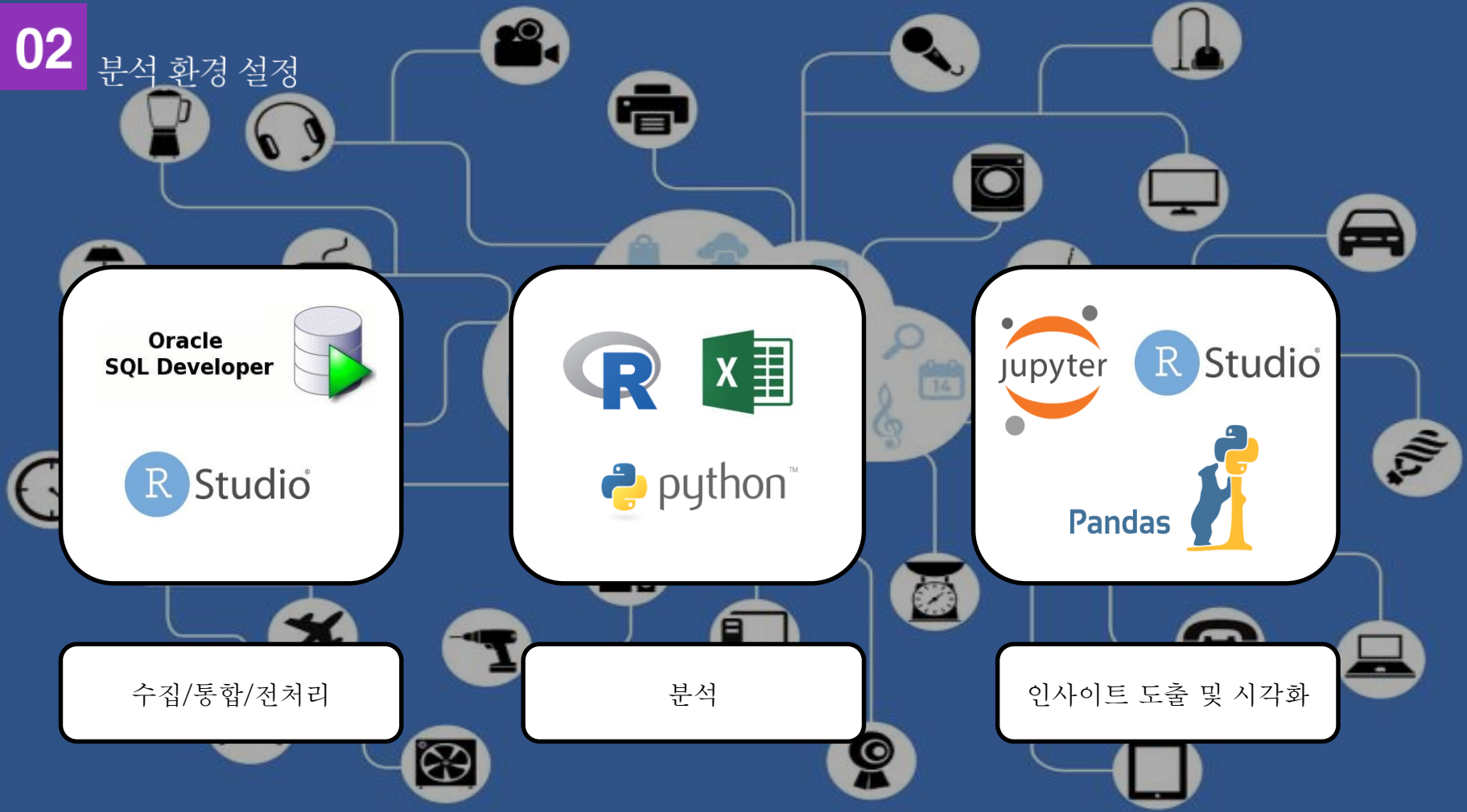
01 프로젝트 개요

1. 프로젝트 일정 계획
2. 분석 환경 설정
3. 역할 분담

01 프로젝트 일정 계획


1주
2주
3주
4주
5주
기획 및 설계
주제 선정/요구 사항 분석
Dataset 수집 계획 수립
업무 분할
Data 전처리
주요 변수 기술 통계
파생 변수 생성
변수 선택과 모델링
모델링
모델 생성/평가
인사이트 도출
결과물 구현(작성)
계획기간
완료기간
중요기간

02 분석 환경 설정


Oracle
SQL Developer
R Studio
python
jupyter
R Studio
Pandas
수집/통합/전처리
분석
인사이트 도출 및 시각화

03 역할분담

허석*(팀장)

- 수행역할
  - 데이터 탐색
  - 데이터 전처리 및 시각화
  - 인사이트 도출
  - 스크립트 제작 및 발표

김진*

- 수행역할
  - 데이터 탐색
  - 데이터 전처리 및 시각화
  - 인사이트 도출
  - ppt 제작 및 발표

신선*

- 수행역할
  - 데이터 탐색
  - 데이터 전처리 및 시각화
  - 머신러닝
  - 인사이트 도출
  - 스크립트 제작 및 발표

전용*

- 수행역할
  - 데이터 탐색
  - 데이터 모델링
  - 인사이트 도출

정연*

- 수행역할
  - 데이터 탐색
  - 데이터 모델링
  - 인사이트 도출

02 분석배경

1. 젠트리피케이션
2. 주제 선정
3. 선행연구
4. 분석 방법 구상
5. 지역 선정

01 젠트리피케이션

젠트리피케이션 (Gentrification)

신사 계급을 뜻하는 '젠트리(gentry)'에서 파생된 것으로
1964년 영국의 사회학자 루스 글라스(Glass)가 처음 도입한 개념.

낙후된 구도심이 활성화 되면서 중산층 이상의 계층이 유입됨으로써
기존의 노동자계층 또는 기존 원주민들이 내몰리는 현상을 말한다.

명동도 썰렁...젠트리피케이션에 코로나까지 자영업자 한계

이해리 기자 | 입력 2020.09.15 14:04 | 수정 2020.09.22 17:10 | 댓글 0

HOME > In-Depth > Point of View

연남동, 화려한 조명 뒤 밀려난 상인의 눈물

이지원 기자 | 호수 357 | 승인 2019.10.02 09:39 | 댓글 0

다 떠난 경리단길...'맛집' 39곳 중 5곳만 남았다[서울, 젠트리피케이션에 바래다]

01 젠트리피케이션

젠트리피케이션 과정

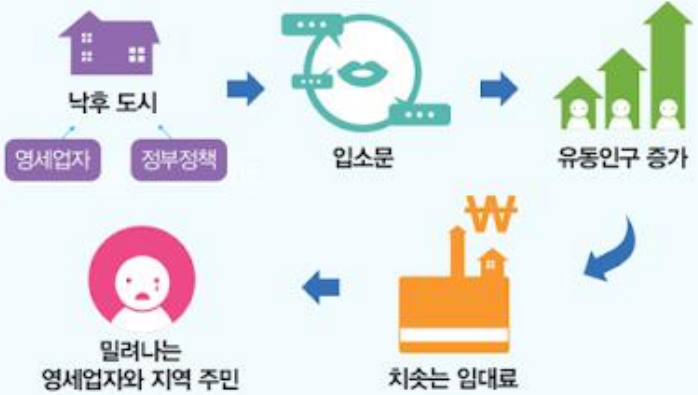


그림 출처
http://blog.daum.net/_blog/BlogTypeView.do?blogid=0A18G&articleno=17052195&_bloghome_menu=recenttext

02 주제 선정

젠트리피케이션 이슈의 쏠림

주거지역 젠트리피케이션

주거지역에서 발생하는 젠트리피케이션은
비단 상가 임차인만의 문제는 아니며,
근본적으로 지역주민들의 주거 문제와도 밀접하게 관련

주제 선정 이유

여러 지자체에서 정책적 노력

BUT 지금껏 상가 임대료 상승과 관련된 부분에만 초점을 맞추고 있음

➔ 주제

지역 소비 데이터, 부동산 시세를 활용한 주거지역 젠트리피케이션 예측

03 선행 연구

출처: 국토연구원

1. 젠트리피케이션 지표의 필요성

젠트리피케이션 현상의 발생 여부와 진행 단계를 종합점수(score)로 계량화해 지역의 변화 관찰 및 지역 간 비교 가능

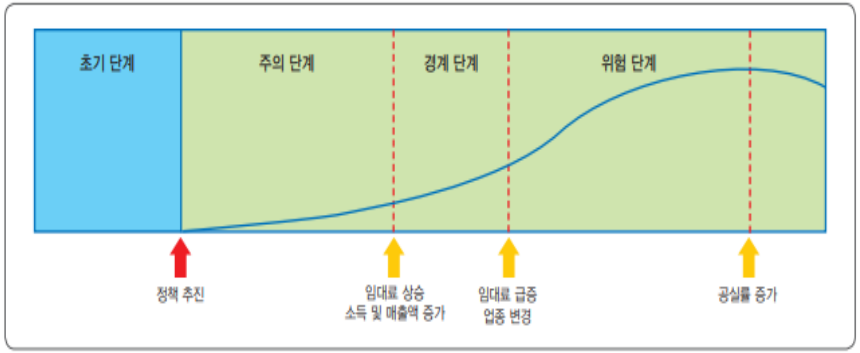
- 젠트리피케이션 지표는 현상을 설명하는 지역 내 인구학적 변화, 부동산 가치와 상업활동 변화 등의 자료를 종합적으로 판단해 문제를 분석할 수 있는 정량화된 수치로 표현
- 젠트리피케이션 지표값이 클수록 젠트리피케이션 현상이 심화된 것으로 해석

젠트리피케이션 지표는 지역의 변화를 신속히 파악해 정책적으로 적절히 대응하기 위해 필요

- 객관적인 자료 분석을 통해 산정된 정량화된 점수를 통해 젠트리피케이션 발생 여부의 진단과 문제의 심각성 파악이 가능
- 젠트리피케이션 발생 단계와 지역적 특성을 반영한 맞춤형 정책을 적용할 수 있으며, 조례 제정 및 구역 지정 등의 근거 자료로 활용
- 정부 정책 추진 시 수치화·도면화된 분석 결과를 지역 주민에게 제공해 젠트리피케이션 문제 대응을 위한 주민의 이해 증진에 기여

젠트리피케이션 지표값에 따라 젠트리피케이션은 현상 발생 이전부터 문제 심각화까지 4단계로 구분

- **(1단계: 초기)** 젠트리피케이션 발생 이전 또는 이후의 지역 쇠퇴 상태
- **(2단계: 주의)** 도시재생사업 추진으로 특정 지역에 자본이 유입, 개발사업 등이 시작되고 상업활동이 증가하면서 지역 활성화 진행
- **(3단계: 경계)** 자본의 지속적 유입에 따라 부동산 시세가 상승하고 유동인구와 매출액 증가
- **(4단계: 위험)** 언론 노출과 외부 자본의 지나친 유입으로 주거지 상업화와 대규모 프랜차이즈의 유입, 급격한 임대료 상승에 따른 비자발적 이주 등 부작용 발생



03 선행 연구

출처: 국토연구원

서울시 종로구

시간이 지날수록 경계·위험 단계로 분석된 블록 비율이 증가하고 있는 것으로 분석돼 젠트리피케이션 문제가 심화되고 있는 것으로 판단

- 2015년에는 주의·경계 단계로 진단된 블록 비율이 각각 11%, 1.3%에 불과했으나 2017년에 주의 단계 비율은 19%, 경계 단계는 약 4%로 상승
- 2015년과 2016년에 진단되지 않았던 위험 단계 역시 2017년에 처음으로 확인

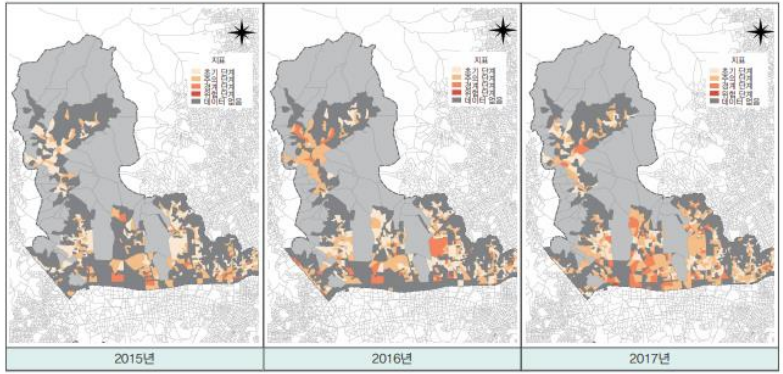
젠트리피케이션 지표값의 공간적 분포 확인 결과, 주의·경계 단계로 진단된 블록이 점차 증가, 특히 남측 상업지역에 집중돼 있음을 확인

- 2015년에는 남측 상업지역 일부에서 확인된 주의·경계 단계 블록이 2016년에 들어 대학로와 부암동 일대에서 증가
- 2017년에는 전반적으로 주의 단계가 증가했으나 삼청동에서 북촌·익선동에 이르는 경복궁과 창덕궁 사이 지역에 경계 단계가 밀집돼 있음을 확인

종로구 젠트리피케이션 발생에 관한 언론기사 보도횟수는 용산구에 비해 약 3배 이상으로 언론의 높은 관심을 받음

- 용산구는 98개 언론기사(전체 6.36%)에서만 젠트리피케이션이 언급된 것과 달리 종로구는 114개 언론기사(전체 17.47%)에서 언급
- 2015년 대비 2016년 줄어들었던 기사 비율이 2017년 1분기부터 상승세를 보였으며, 서촌·북촌·삼청동·광화문 지역에서 발생한 젠트리피케이션이 주로 논의돼 지표값과 일부 일치

그림 4 종로구 젠트리피케이션 지표 적용 결과



04 분석 방법 구상

1 *변수 탐색*	각 데이터에서 변수를 탐색하고 EDA 및 시각화
2 *지역 구분 및 선정*	변수별 지역 데이터 차이 조정 및 전처리 결과에 따른 지역 선정
3 *지표 변수 선정*	선행연구, 사례 검토, 타 변수와의 관계 등을 고려해 변수를 최종 선정
4 *점수 계량화*	각 지표에 따른 점수 계량화
5 *젠트리피케이션 여부 확인*	점수가 높을수록 젠트리피케이션 현상이 발생 확률이 높을 것으로 예상

05 지역 선정

종로구

강원도

03 데이터 탐색 및 전처리

1. EDA
2. 시각화

01 EDA

활용 데이터 및 변수 탐색

활용 데이터

데이터 출처 : BC카드 2020 금융 빅데이터 챌린지

1) 망고플레이트 인기 식당 정보
2) 비씨카드 거래내역
3) 세대, 부대시설 등 아파트 정보 데이터
4) KT 유동인구 집계 데이터

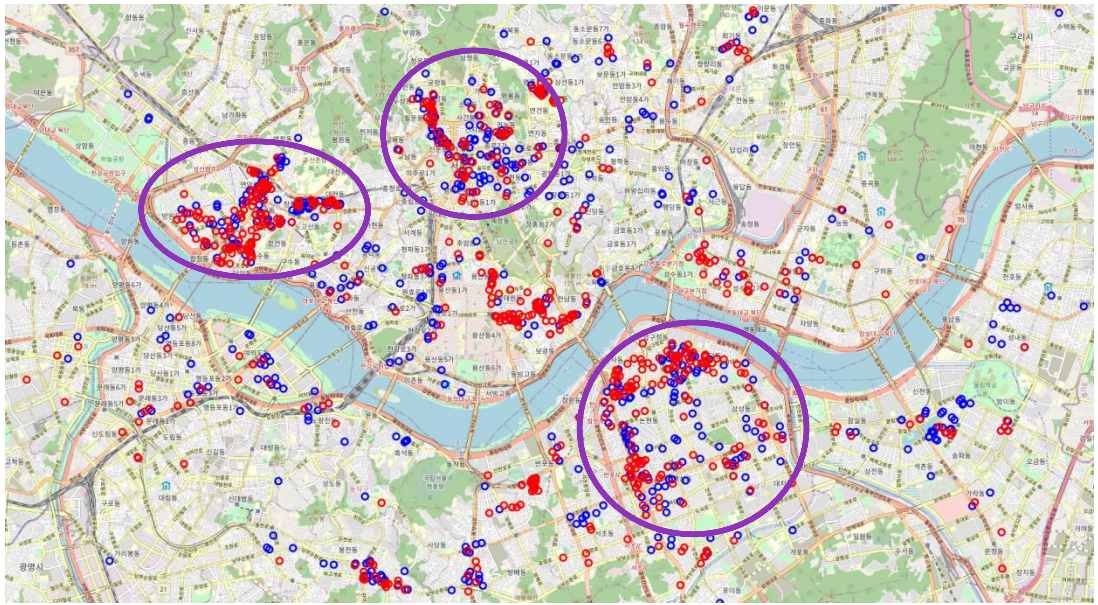
데이터 탐색

1) 데이터 분포 탐색하기
2) 통계값(중앙값, 평균, 최빈값) 활용 이상치 제거
3) 속성간 관계 분석 (상관분석)
4) 스케일링

02 시각화

망고플레이트 맛집 데이터
(서울,강원도)

외국인 관광객 인기식당
현지인 인기식당
분포 비교(folium)



서울시 맛집 지도

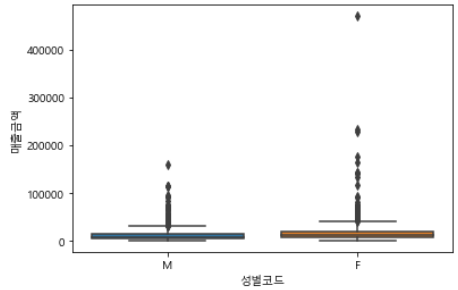
02 시각화

비씨카드 소비 데이터
(전국, 2019.04~05, 2020.04~05)

성별 연령대별 매출금액
이상치 제거 및 시각화

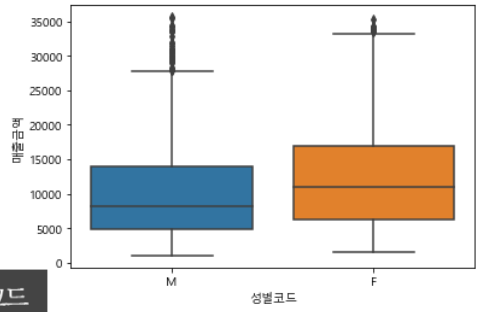
```
import seaborn as sns
sns.boxplot(x="성별코드", y="매출금액", data=raw)
```

```
<AxesSubplot:xlabel='성별코드', ylabel='매출금액'>
```




```
sns.boxplot(x="성별코드", y="매출금액", data=raw)
```

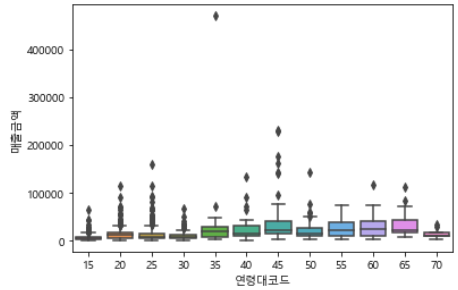
```
<AxesSubplot:xlabel='성별코드', ylabel='매출금액'>
```



성별코드

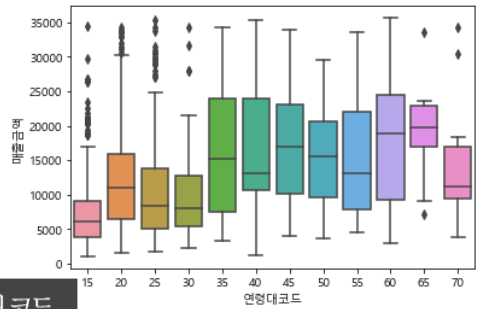
```
sns.boxplot(x="연령대코드", y="매출금액", data=raw)
```

```
<AxesSubplot:xlabel='연령대코드', ylabel='매출금액'>
```




```
sns.boxplot(x="연령대코드", y="매출금액", data=raw)
```

```
<AxesSubplot:xlabel='연령대코드', ylabel='매출금액'>
```



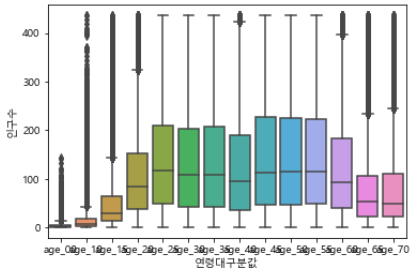
연령대코드

02 시각화

KT 유동인구 데이터
(종로구, 2019.04~05, 2020.04~05)

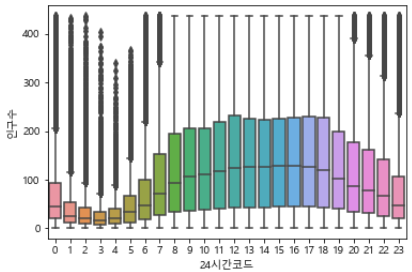
연령대 별 / 시간대 별
유동인구 추세 분석

성별 / 연령대 별
유동인구 추세 분석




```
sns.boxplot(x="24시간코드", y="인구수", data=df)
```

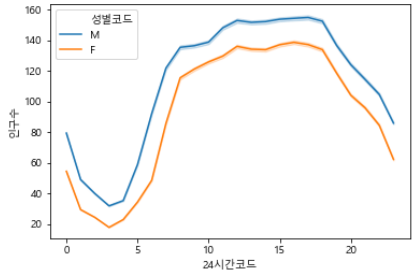
```
<matplotlib.axes._subplots.AxesSubplot at 0x1da0a248d30>
```



연령대별/시간대별

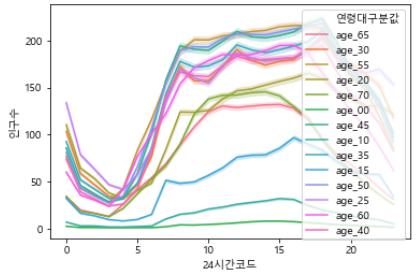
```
sns.lineplot(x="24시간코드", y="인구수", hue='성별코드',data=df)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1da0b578668>
```




```
sns.lineplot(x="24시간코드", y="인구수", hue='연령대구분값',data=df)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1da0b5676d8>
```



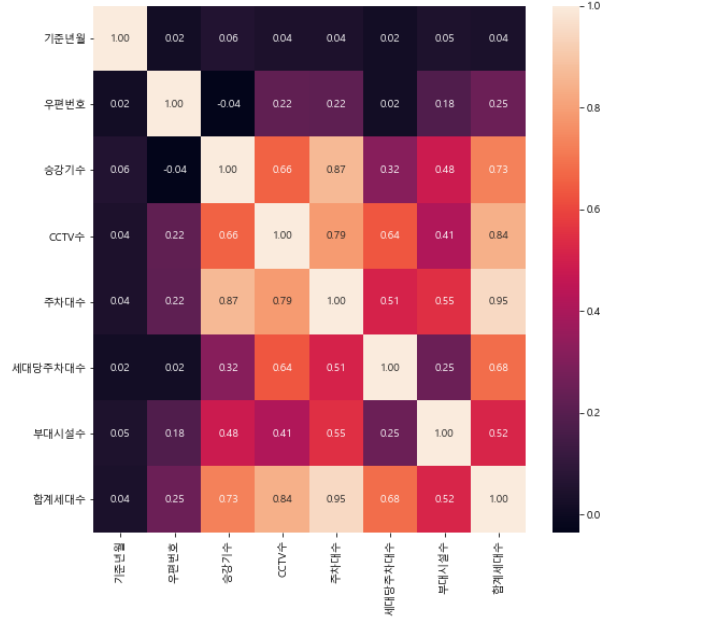
성별/시간대별

02 시각화

기웅 아파트 시설 정보 (서울,대구,강원도)

아파트 부대시설 간 상관관계 분석

```
import matplotlib.pyplot as plt
import seaborn as sns
plt.figure(figsize=(9, 9))
sns.heatmap(data=df.corr(), annot=True, fmt=".2f")
plt.show()
```



04 모델링

1. 모델링
2. ER Diagram

01 모델링

모델링 과정

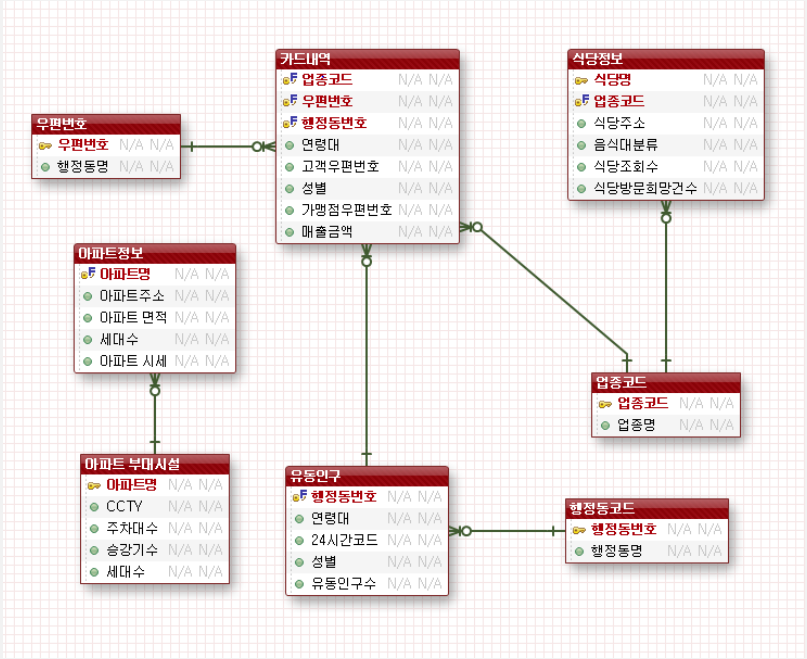
1) 핵심 entity 도출

- 카드 내역
- 유동인구
- 우편번호
- 업종코드
- 식당정보
- 행정동코드
- 아파트 정보
- 아파트 부대시설

2) 상세 속성 정의

3) 정규화

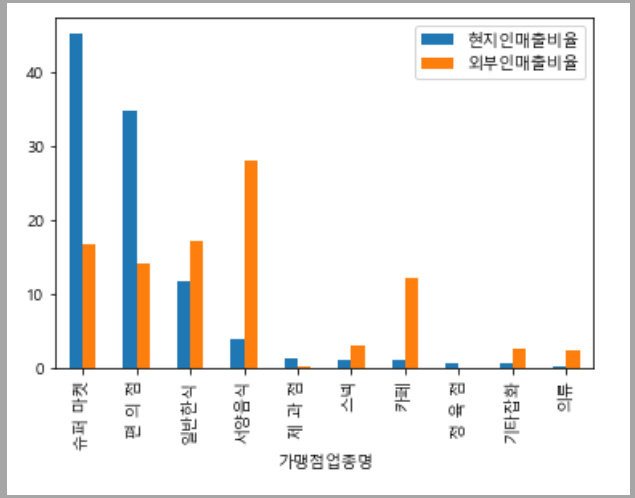
02 ER Diagram



05 인사이트 도출 및 예측분석

1. 시각화 및 인사이트 도출
2. 지표 선정
3. 분류 분석

01 시각화 & 인사이트



비씨카드 데이터(종로구, 2019.04~05,2020.04~05)

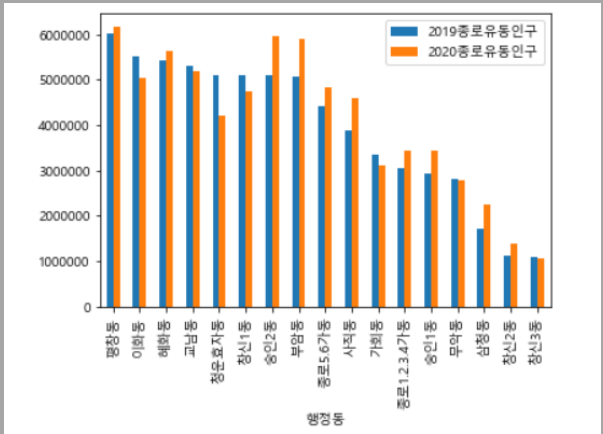
업종별 현지인/외부인 매출 비율 비교

➔슈퍼마켓, 편의점: 현지인 매출 비율이 현저히 높음

➔일반한식, 서양음식, 카페: 외부인 매출 비율이 높음

(현지인=종로구 거주자/ 외부인=종로구 외 거주자)

01 시각화 & 인사이트



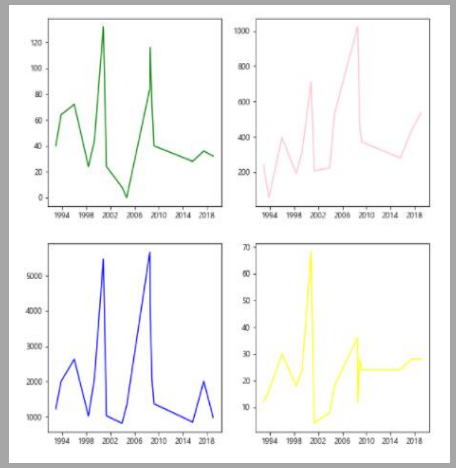
KT 유동인구 데이터(종로구,
2019.04~05,2020.04~05)

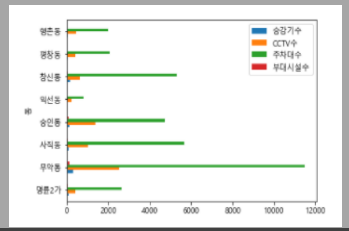
2019/2020 종로구 유동인구 비교

➔17개 동 중 10개 동 유동인구 증가, 7개 동 감소

(코로나 영향으로 감소했을 가능성도 배제하지 못함)

01 시각화 & 인사이트





기웅 아파트 부대시설 데이터
(종로구,2019.04~05,2020.04~05)

아파트 동별 부대시설 및 합계 세대수
아파트 준공일자 별 부대시설 및 합계 세대수

➔각 변수들이 같은 증감 추이를 보임
➔지역별 추이도 같음

02 지표 선정

선행연구

➔젠트리피케이션 요인 선정
(상업, 부동산 요인 등)

➔ 변수 간 상관계수 파악

➔각 요인에 영향을 미치는
주요 변수 선정

지표변수

1) 식당 조회수, 방문희망 건수
2) 성별 연령별 신용카드 매출
3) 카페 비율
4) 아파트 부대시설 수
5) 최근 전세 실거래가
6) 유동인구 수

03 분류분석(랜덤 포레스트)

서울시 종로구 데이터 분류분석
(랜덤포레스트)

1. 비씨카드 데이터

1) 매출에 영향을 미치는 변수 설정 :
성별, 연령

2) 매출을 기준으로 젠트리피케이션 발생할 지역과 그렇지 않을 지역을 구분하여 정답지를 만듦

3) 랜덤포레스트를 이용한 분류분석을 실시

Cross_val_score 결과:

```
print(np.mean(scores))
```

```
0.6219288174512055
```

feature_importances_를 바탕으로 주요 변수를 선정해본 결과:

```
print(model.feature_importances_)
```

```
[0.96218522 0.03781478]
```

정확도, F1 score, 정밀도, 재현률(민감도), ROC:

```
model.fit(X_train, y_train)
pred = model.predict(X_test)  # yhat = price

print(confusion_matrix(y_test, pred))
print("정확도:", accuracy_score(y_test, pred))
print("F1:", f1_score(y_test, pred))
print("정밀도:" , precision_score(y_test , pred))
print("재현률(민감도):" , recall_score(y_test , pred))
print("ROC:" , roc_auc_score(y_test, pred))
```

```
[[1035  496]
 [ 416  666]]
정확도: 0.6509758897818599
F1: 0.5935828877005348
정밀도: 0.5731497418244407
재현률(민감도): 0.6155268022181146
ROC: 0.6457777708020684
```

03 분류분석(랜덤 포레스트)

서울시 종로구 데이터 분류분석
(랜덤포레스트)

2. 망고플레이트 데이터

1) 맛집 선호도에 영향을 미치는 변수 설정 :
식당 조회수, 식당방문희망건수
2) 카페 비율 파생변수를 만들어 이를 기준으로 젠트리피케이션 발생할 지역과 그렇지 않을 지역을 구분하여 정답지를 만듦
3) 랜덤포레스트를 이용한 분류분석을 실시

Cross_val_score 결과:

```
print(np.mean(scores))
```

```
0.5432290157442287
```

feature_importances_를 바탕으로 주요 변수를 선정해본 결과:

```
print(model.feature_importances_)
```

```
[0.65199195 0.34800805]
```

정확도, F1 score, 정밀도, 재현률(민감도), ROC:

```
model.fit(X_train, y_train)
pred = model.predict(X_test)  # yhat = price

print(confusion_matrix(y_test, pred))
print("정확도:", accuracy_score(y_test, pred))
print("F1:", f1_score(y_test, pred))
print("정밀도:" , precision_score(y_test , pred))
print("재현률(민감도):" , recall_score(y_test , pred))
print("ROC:" , roc_auc_score(y_test, pred))
```

```
[[92 75]
 [71 68]]
정확도: 0.5228758169934641
F1: 0.4822695035460993
정밀도: 0.4755244755244755
재현률(민감도): 0.4892086330935252
ROC: 0.5200534183431698
```

03 분류분석(랜덤 포레스트)

서울시 종로구 데이터 분류분석
(랜덤포레스트)

3. 기웅 데이터

1) 최근 전세실거래 가격에 영향을 미치는 변수 설정 : CCTV수, 승강기수, 주차대수, 합계세대수

2) 전세가격을기준으로 젠트리피케이션 지역을 설정하여 정답지를 만듦

3) 랜덤포레스트를 이용한 분류분석을 실시

Cross_val_score 결과:

```
print(np.mean(scores))
```

```
0.6852216748768474
```

feature_importances_를 바탕으로 주요 변수를 선정해본 결과:

```
print(model.feature_importances_)
```

```
[0.08491659 0.27130963 0.3901187  0.25365509]
```

정확도, F1 score, 정밀도, 재현률(민감도), ROC:

```
model.fit(X_train, y_train)
pred = model.predict(X_test)  # yhat = price

print(confusion_matrix(y_test, pred))
print("정확도:", accuracy_score(y_test, pred))
print("F1:", f1_score(y_test, pred))
print("정밀도:" , precision_score(y_test , pred))
print("재현률(민감도):" , recall_score(y_test , pred))
print("ROC:" , roc_auc_score(y_test, pred))
```

```
[[574  68]
 [ 76 458]]
정확도: 0.8775510204081632
F1: 0.8641509433962263
정밀도: 0.870722433460076
재현률(민감도): 0.8576779026217228
ROC: 0.8758794497532291
```

06 결론 및 추후 과제

1. 결론
2. 한계점 및 추후 과제

01 결론

1) 주거지역 젠트리피케이션을 일으키는 요인:
성별& 연령별 유동인구, 식당 증가, 아파트 부대시설 수 등

2) 각 요인별 합산 결과, 이미 젠트리피케이션이 발생한 지역과 발생할 지역의 특성이 겹침

02 한계점 및 추후 과제

한계점: 1) 데이터의 지역 정보 불균형
2) 대부분의 데이터가 2019년 ~ 2020년
but 2020년 '코로나 바이러스'라는 변수로 증감량이 예상과 다르게 분석됨

추후과제: 1) 각 요인들에 가중치를 부여하여 젠트리피케이션 지표 생성
2) 지표에 따른 지역별 젠트리피케이션 단계 설정

Q & A

감사합니다.