경기도

GBSA 경기도경제과학진흥원

2022년 데이터산업 인력양성 교육 프로젝트 데이

S_mile (Smart_mile) ☺

2022.08.30

팀명: 데굴데굴

CONTENTS

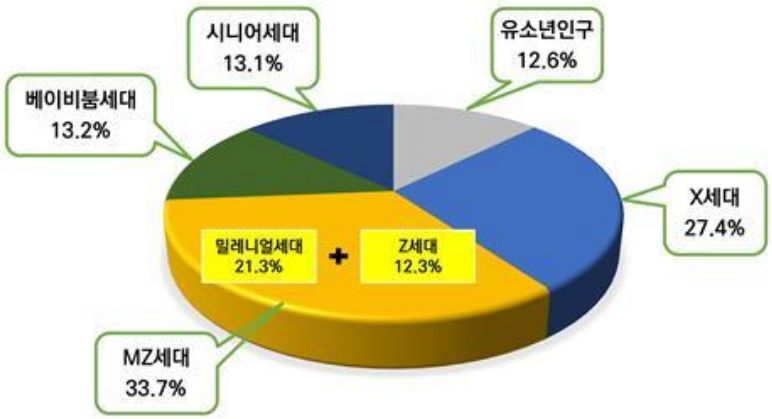
Note
- 목차는 변경 불가, 하단에 기업명 또는 로고 입력

아이디어 구상 및 개발배경

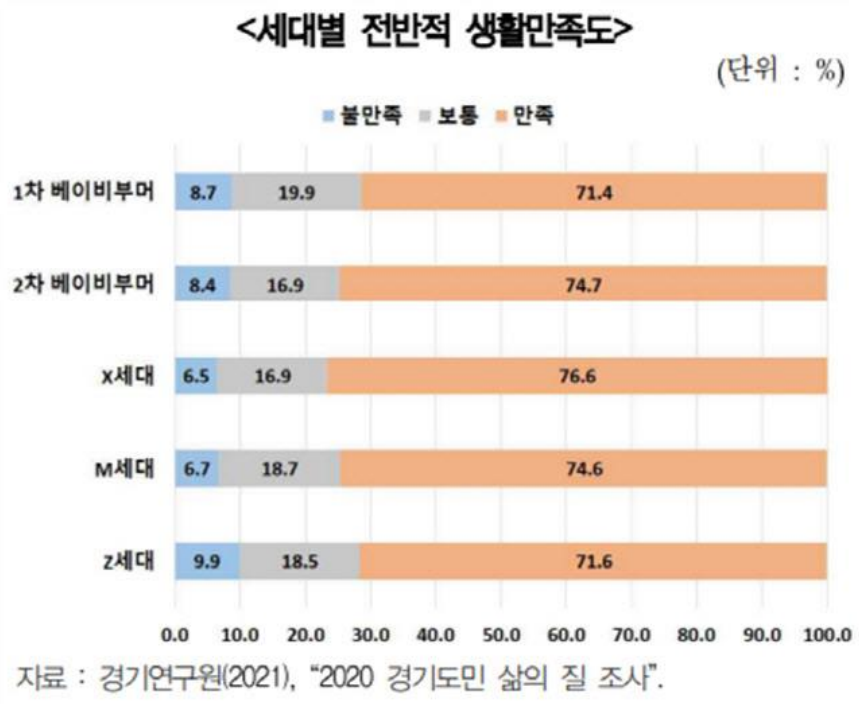
“ MZ세대, 전반적 생활만족도 낮고 우울감 커 ”

코로나19 이전과 비교해 삶의 질 수준이 가장 많이 하락한 집단
일자리에 대한 불안감과 우울감이 크다

=> 안정과 정신건강 회복을 위한 정책 필요



경기도 인구 1,331만5,000여 명 기준

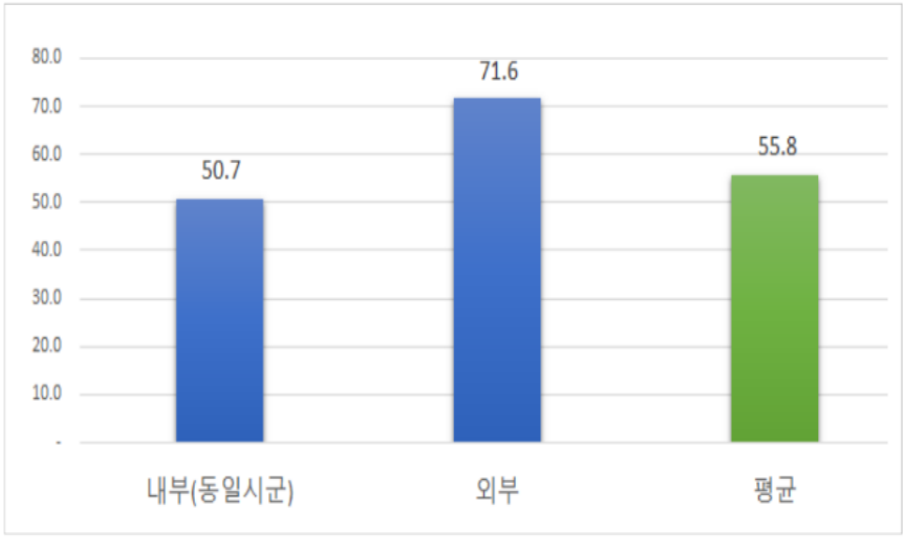


자료 : 경기연구원(2021), “2020 경기도민 삶의 질 조사”.

아이디어 구상 및 개발배경

경기도민 통근 삶의 질 특성

[통근지역별 스트레스 점수]



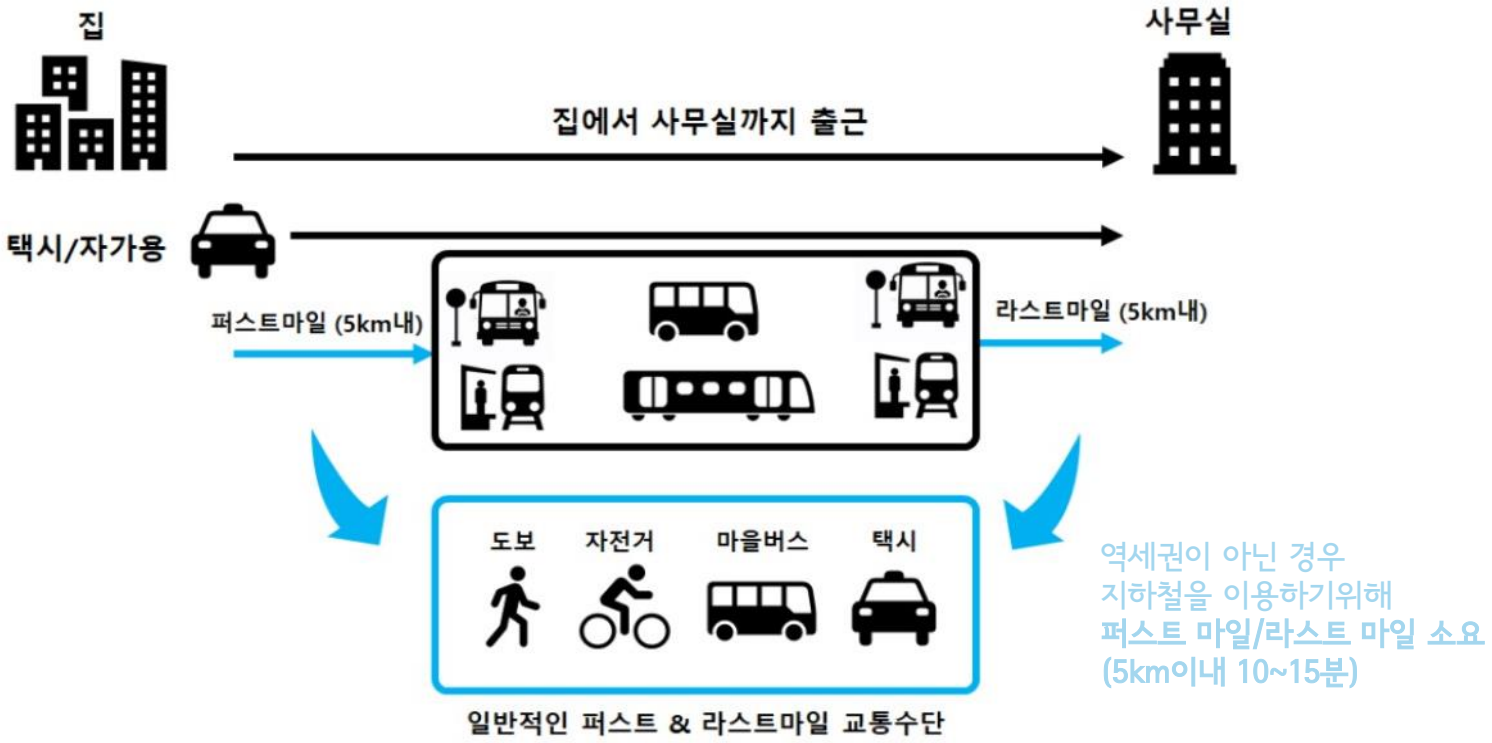
해당 연구보고서에 따르면

경기도 직장인 중 **26%**가 **지역 외부**로 통근,
지역 외부 통근자는 지역내 통근자에 비해 약 **1.4배**의 스트레스 받음

연령별 통근 스트레스 中 **청년층**이 가장 높음

아이디어 구상 및 개발배경

경기도민의 장거리 통근 살펴보기 (퍼스트/라스트 마일)



역세권이 아닌 경우
지하철을 이용하기위해
퍼스트 마일/라스트 마일 소요
(5km이내 10~15분)

퍼스트마일 & 라스트마일 교통수단이 열악한 경우 삶의 질 저하

아이디어 목표

서비스 대상 및 기대효과

최적화 입지에 환승 가능한 공공 자전거 도입

절약

간접 역세권의 확장
역세권 주거비 부담 감소

활성화된 대중교통으로
근거리 차량 이용 감소
(차량 유지비 절감)

교통

배차간격이 긴 대중교통 대신 first/last mile 모빌리티로 이용

대중교통 미운행 시간, 혼잡시간 대중교통 대체제

건강

일상적인 자전거 이용으로 신체적, 정신적 건강 관리

문화/여가시설 접근성이 좋아져 만족스러워진 여가생활

청년들의 경제적 부담을 덜어주고
코로나로 지쳐 있던 신체적, 정신적 건강을 되찾자 !

데이터 수집

1. 지표 선정 배경에 따른 데이터 수집

교통 데이터

활용 데이터	출처
버스 혼잡도	경기도교통정보센터
버스 정류장 수	경기도교통정보센터
지하철역 이용자 수	경기도교통정보센터
경기도 수원시 택시 승강장 현황	공공데이터포털

시설물 데이터

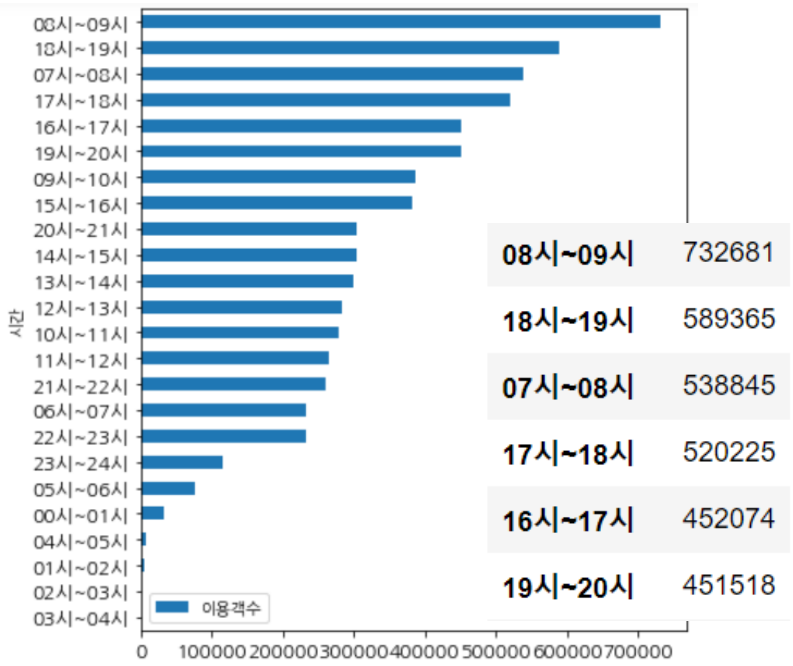
활용 데이터	출처
영화 상영 업체 현황	경기데이터드림
공공 시설 현황	공공데이터포털
도서관 현황	경기도교통정보센터
공원 및 자전거 도로	행정안전부
문화 복합 시설 및 마트	공공데이터포털

경제적 지표 데이터

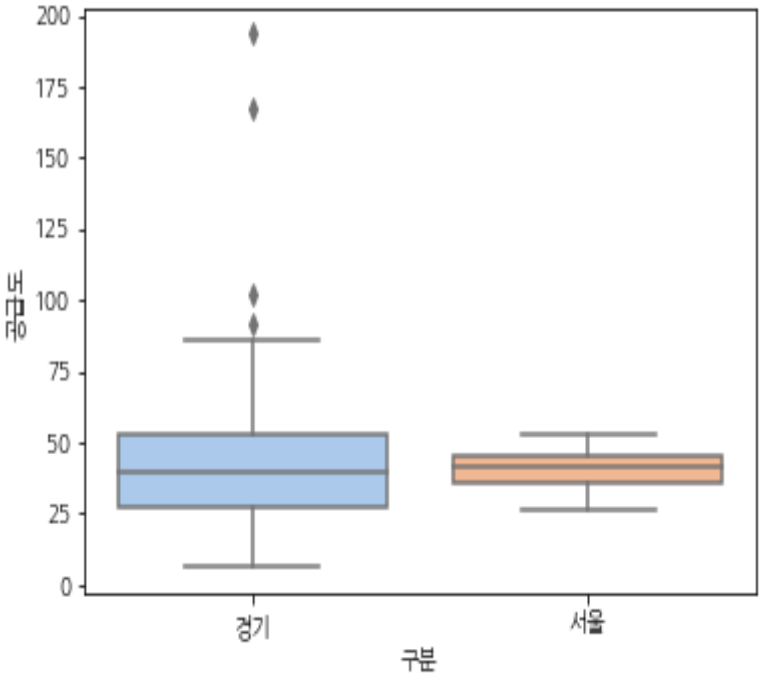
활용 데이터	출처
아파트/오피스텔/연립다세대 전월세 신고 현황	경기데이터드림
경기도 수원시 지역화폐 결제 정보	공공데이터포털

지표 선정 배경 (1/3)

1. 경기도 대중교통 현황



대중교통이 가장 혼잡한 시간대와 위치



버스 공급도

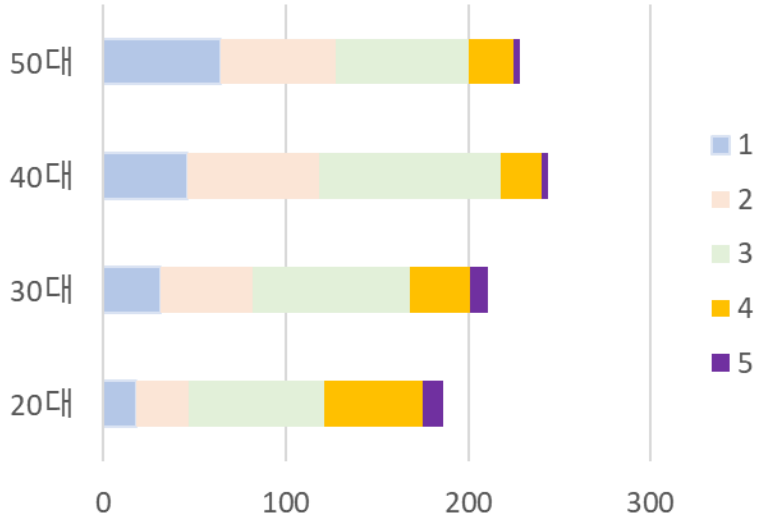
지표 선정 배경 (2/3)

2. 사회경제 지표: 경기도 청년들의 경제 수준과 지출 의향

경기도 청년들의 경제수준

차량소유	비율	연령	대출
O	20.4	28.5	890
x	79.6	24.9	2771

청년들의 여가비용 추가 지출 의향



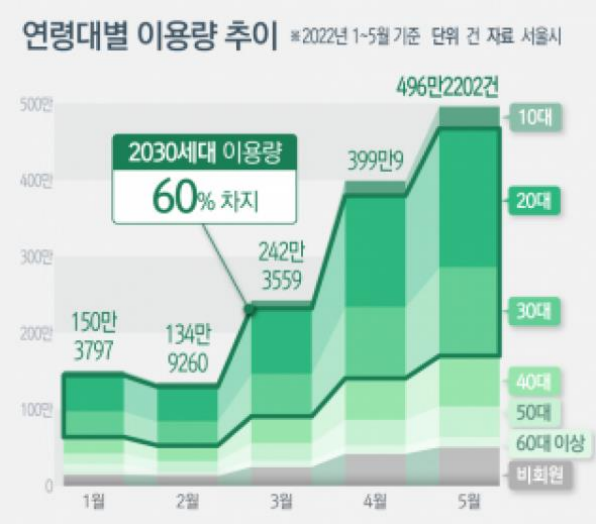
경기도 청년 통장 가입자 정보에 따르면 자차 소유율이 20.4%이며 자차를 소유한 청년들의 평균 대출 금액은 2700만원이다. (자차 미소유자에 비해 3배 많다)
20, 30대의 경우 다른 연령대보다 문화취미와 교통비 추가 지출 의향이 상대적으로 높다.

지표 선정 배경 (3/3)

3. 타 시도의 공공자전거 이용현황 (서울시 따릉이)

20, 30대 따릉이 이용량, 60%

30분 이내 이용자 비율 35%



따릉이 연령별 이용 형태는 **20, 30대가** 전체의 **60%** 차지
주말보다는 평일, 출퇴근 시간대에 가장 높게 나타남
출퇴근 시간 오전 7-9시, 오후 5-7시에 이용하는 비율이 전체의 56.4%
이들은 대중교통 이용 시 퍼스트·라스트 마일 모빌리티로 '따릉이'를 적극 이용!

서울시 <서울 교통이용통계보고서>

3. 분석 시나리오

분석 과정 flow chart

분석 개요

지표 선정 배경

경기도 대중교통 이용 현황

경기도 시별 공급도 대중교통 이용자수 많은 도시 선정 (수원)

청년 경제지표

청년들의 경제수준 항목별 지출 의향

따릉이 이용 현황

출퇴근 시간대 이용비율 (30분 이내) 반납지점에 따른 자전거 이용 용도

데이터수집

수원시 동별 인구수

버스 혼잡도, 버스 정류장수

지하철역 이용자수

택시 승강장 수

수원시 동별 지역화폐 결제건수

수원시 주거 형태별 전세, 월세 가격

수원시 동별 문화시설 / 여가시설 수

전처리

결과 및 해석

K-means 군집분석

군집별 특징

랜덤 포레스트

위치 정보 데이터

지도 시각화

우선설치 자전거 정류장 선정

1. 데이터 수집 후 전처리

 # 모델링 데이터

연립다세대/오피스텔/아파트전월세 - 법정동명별 보증금액과 월세금액의 평균.

(2차 수정시) 결측치는 중앙값으로 대체

버스환승시간: 1차: 평균, 최대값

(2차 수정시에 추가) 법정동명별 정류장별 환승시간 데이터의 평균값과 최대시간 결측치는 평균값로 대체

 # 시각화 데이터

위치정보가 누락된 부분은 geo-coding으로 위도 경도 추출

분석 내용

1단계 : 그룹으로 묶어 label별 특성 확인

첫번째 k-means 군집분석

```
result_df.groupby('label').mean()
```

label	게스트하우스	공공문화시설_cnt	공공문화시설이용자수	공공이용시설_cnt	공원(수)	도서관좌석수	도서관이용자수	동별세대	동별인구수	문화시설수	보증금_아파트	월세_아파트	보증금_연립다세대	월세_연립다세대	보증금_오피스텔	월세_오피스텔	지하철역이용객수
0	0.207547	0.169811	6901.792453	1.339623	5.943396	138.075472	5840.396226	8218.245283	19245.981132	0.679245	13322.57842	6.781127	6385.226402	11.297601	5981.35751	7.748301	2865.754717
1	1.000000	1.666667	290044.333333	4.333333	7.000000	211.000000	8868.000000	14972.333333	31664.666667	3.333333	21158.50414	12.298844	31137.719328	11.580159	7142.16460	14.807624	6529.666667
2	0.000000	1.000000	660869.000000	5.000000	36.000000	703.000000	33249.000000	30767.000000	73268.000000	7.000000	21839.96855	12.778012	9986.444906	7.654155	13263.18182	11.545455	0.000000

```
result_df['label'].value_counts()
```

```
0    53
1     3
2     1
```

최적의 k 값 (k=3)으로 모델링 후 중요한 열 외 drop

```
from sklearn.cluster import KMeans
model = KMeans(n_clusters = 3)
x2 = df.drop(['보증금_오피스텔','게스트하우스','공공문화시설_cnt','법정동명','동별세대','보증금_0
x2.sample(2)
```

	공공이용시설_cnt	공원(수)	도서관좌석수	도서관이용자수	동별인구수	문화시설수	지하철역이용객수
30	1	3	0	0	21254	1	0
56	2	22	354	15040	49357	1	11609

```
result_df['label'].value_counts()
```

```
0    39
4    12
1     3
3     2
2     1
```

재학습 -> label 별 개수 산출

분석 내용

1 단계: 첫번째 k-means 분석

새로운 변수추가 + 전처리

전처리 전

게스트하우스	공공문화시설_cnt	공공문화시설이용자수	공공이용시설_cnt	공원(수)	도서관좌석수	도서관이용자수	동별세대	동별인구수	문화시설수	보증금_아파트	월세_아파트	보증금_연립다세대	월세_연립다세대	보증금_오피스텔	월세_오피스텔	지하철역이용객수
0	0	0	1	4	0	0	9735	23268	1	10974.254740	10.108401	8764.064434	13.374063	7000.000000	30.000000	0
0	0	0	1	10	0	0	4140	10154	1	13631.532280	11.849307	8687.655172	14.074713	12000.000000	0.000000	0
0	0	0	1	6	0	0	8966	17966	0	20415.666780	10.091728	5894.298246	10.565789	0.000000	0.000000	0
0	1	4000	3	1	357	9826	1701	3410	2	17787.500000	3.750000	12728.511900	7.627976	23625.000000	9.500000	0
0	0	0	2	13	0	0	10835	24577	1	15688.960030	9.717424	6352.119883	13.638158	1000.000000	50.000000	0

전처리 후 + 새로운 변수 추가

	법정동명	지하철역	문화여가시설	공공이용시설	지역화폐결제건수합	택시승강장수	동별세대	동별인구수	보증금_아파트	월세_아파트	보증금_연립다세대	월세_연립다세대	평균버스환승시간	최대버스환승시간
0	고등동	0	5	1	58712	1	9735.0	23268	10974.25474	10.108401	8764.064434	13.374063	10.907861	11.567838
1	고색동	0	11	1	173166	1	4140.0	10154	13631.53228	11.849307	8687.655172	14.074713	9.708116	13.356923
2	곡반정동	0	6	1	163686	6	8966.0	17966	20415.66678	10.091728	5894.298246	10.565789	13.599917	24.542000
3	교동	0	4	4	34180	0	1701.0	3410	17787.50000	3.750000	12728.511900	7.627976	9.564046	11.170455
4	구운동	0	14	2	158672	2	10835.0	24577	15688.96003	9.717424	6352.119883	13.638158	13.022224	31.570000

2 단계: 두번째 k-means 분석

두번째 k-means 군집분석

1. 전체 열 기준

```
from sklearn.cluster import KMeans
model = KMeans(n_clusters = 5)
x = df.drop('법정동명', axis = 1)
```

```
model.fit(x)
result_df = x.copy()
result_df['label'] = model.labels_
result_df['label'].value_counts()
```

```
3    28
0    12
2     8
4     5
1     3
Name: label, dtype: int64
```

```
result_df['label'].value_counts()
```

0	39
4	12
1	3
3	2
2	1

변화

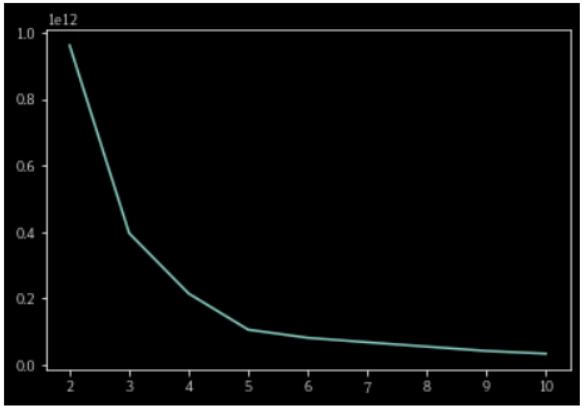
2	28
0	12
4	8
1	5
3	3

=> 이상값 수정 후 균일해진 군집 내 원소의 수

```
distance = []

for i in range(2,11):
    model = KMeans(n_clusters = i)
    model.fit(x)
    distance.append(model.inertia_)
sns.lineplot(x=list(range(2,11)), y=distance)
```

```
<AxesSubplot:>
```



=>K=5로 모델링 완료

2 단계: 랜덤포레스트

K = 5로 랜덤포레스트 실시

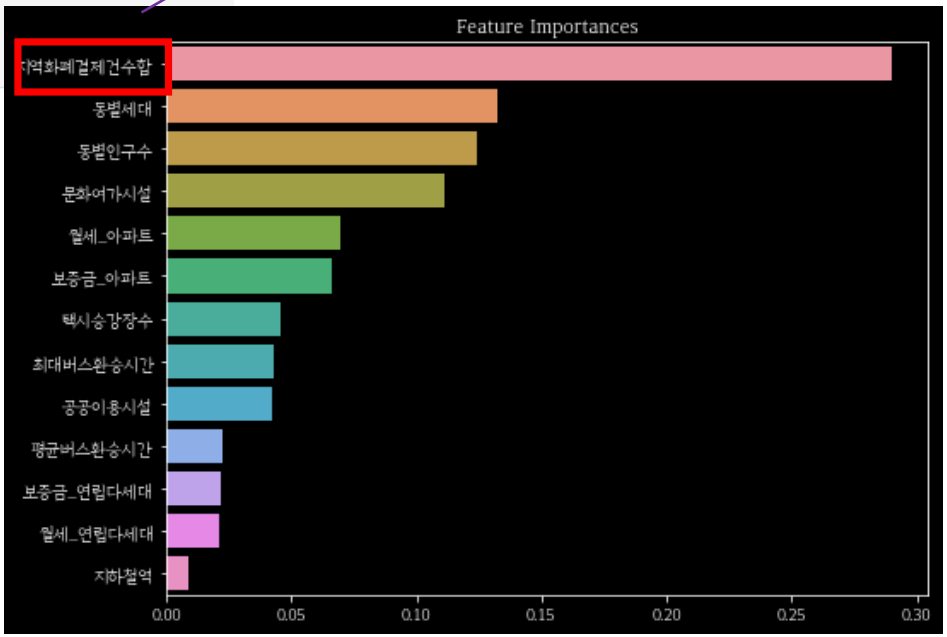
```
from sklearn.metrics import classification_report
# ▶ 학습
rfc = RandomForestClassifier(random_state = 123456, max_depth=6)
rfc.fit(X_train, y_train)
```

```
RandomForestClassifier(max_depth=6, random_state=123456)
```

```
# ▶ 예측
y_pred_train = rfc.predict(X_train)
y_pred_test = rfc.predict(X_test)
accuracy_score(y_test, y_pred_test)
```

```
0.75
```

지역화폐 결제건수 합



분석 내용

3 단계: 변수 제거 후 두번째 랜덤 포레스트

2. 지역화폐 제거 후

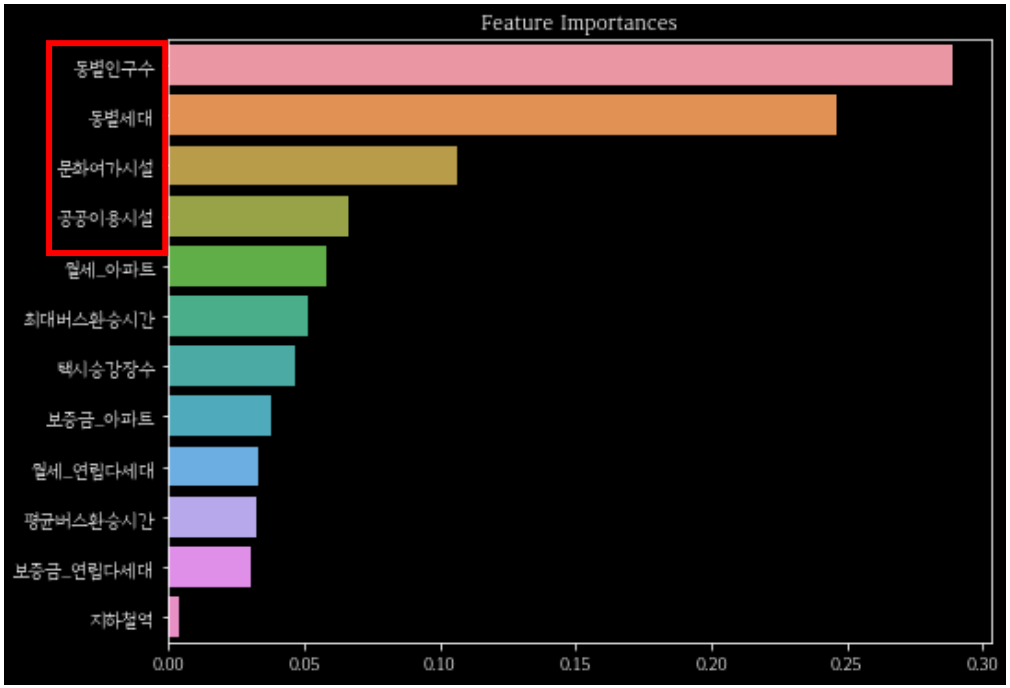
```
from sklearn.cluster import KMeans
model = KMeans(n_clusters = 5)
x_1 = df.drop(['법정동명','지역화폐결제건수합'], axis = 1)
```

```
from sklearn.metrics import classification_report
# ▶ 학습
rfc = RandomForestClassifier(random_state = 123456, max_depth=6)
rfc.fit(X_train, y_train)
```

```
RandomForestClassifier(max_depth=6, random_state=123456)
```

```
# ▶ 예측
y_pred_train = rfc.predict(X_train)
y_pred_test = rfc.predict(X_test)
accuracy_score(y_test, y_pred_test)
```

```
0.9166666666666666
```



→ 변수인 '지역 화폐 결제건수' 제거 후 예측도가 상승함

분석 내용

4 단계: 그룹 선정

label	지하철역	문화여가시설	공공이용시설	택시승강장수	동별세대	동별인구수	보증금_아파트	월세_아파트	보증금_연립다세대	월세_연립다세대	평균버스환승시간	최대버스환승시간
0	0.3	15.4	3.9	4.4	18603.1	42305.0	20167.6	12.8	7891.8	13.7	12.6	24.7
1	0.1	3.0	0.6	0.8	2307.3	5147.5	16427.7	7.7	7687.2	15.5	11.0	16.6
2	0.0	22.0	7.0	2.0	22147.0	53339.0	45867.6	24.5	74929.2	5.0	11.0	17.2
3	0.3	25.5	5.7	7.0	27905.7	70021.3	24887.0	11.0	8705.3	11.5	12.5	24.8
4	0.1	5.5	1.6	1.2	10416.3	22524.5	18638.1	10.0	7328.2	12.7	12.4	23.9

→ 가장 많은 변수의 최고 값을 가진 3 그룹이 시범 지역으로 적합

수 원 시
SUWON CITY

최종 선정 군집의 특징
1) 근방 다수의 문화시설 -> 시설 활용도 증가
2) 많은 동별 인구수 -> 잠재적 모빌리티 사용자
3) 긴 버스 환승 시간 -> 자전거 우선적 도입 필요

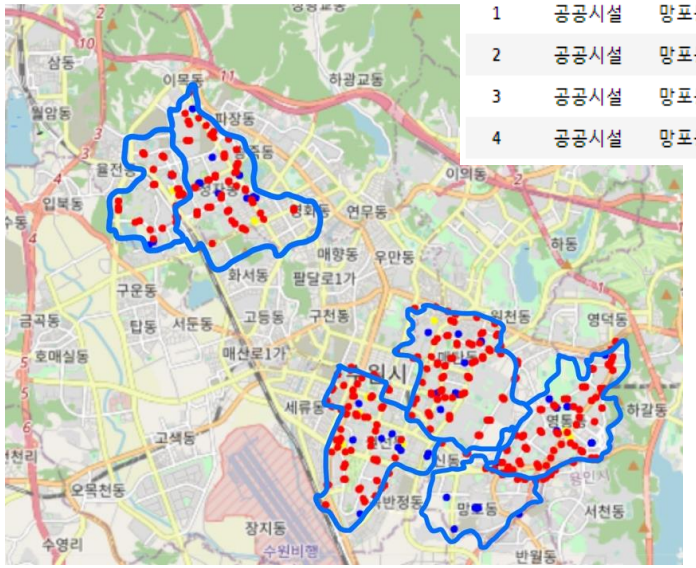
=> 6개동: 권산동 망포동, 매탄동, 영통동, 정자동, 천천동

분석 결과 및 시각화

5 단계: 시각화

수 원 시
SUWON CITY

최종 지도 시각화



	구분	법정동명	위도	경도
0	공공시설	망포동	37.238839	127.051247
1	공공시설	망포동	37.235025	127.046392
2	공공시설	망포동	37.237788	127.060843
3	공공시설	망포동	37.238849	127.051616
4	공공시설	망포동	37.240707	127.044466

우선우선 배치 지역
권산동, 망포동, 매탄동
영통동, 정자동, 천천동

우선 설치 자전거 정류장 선정

활용방안

사업화 방안
전광판, 홍보물로 운영비 충당

경기도 관광지 자전거 여행

러시아워 할인제도
출근 시간대 자전거로 환승 시 마일리지

기대효과

교통수단 간 연계 강화

지속 가능한 저탄소 이동수단

문화 여가 접근성 증가

QnA

경기도

GBSA
경기도경제과학진흥원

감사합니다.

Thank you