데이터 분석을 통한 1인가구 거주지 추천

Change:ON

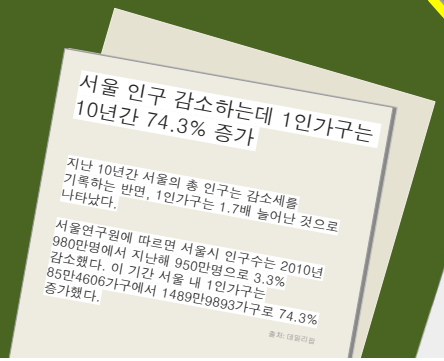
이상민, 이호용, 이혜연, 정형준, 홍영준

| 응용통계학과 공모전 |

1 분석 개요 연구 배경

☑ 전체 가구 유형 중 1인가구가 가장 큰 비중을 차지

☑ 앞으로도 1인가구의 수가 계속해서 증가할 전망



1인가구의 수가 계속 증가하는 추세에 따라

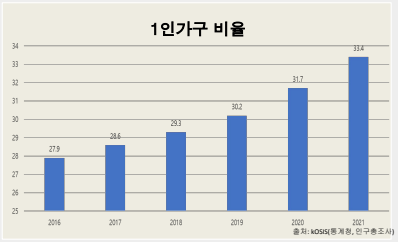
1인가구 생활에 입문하는 사람들을 위한 주거지 추천 모델을 만들어보자!

1 분석 개요 연구 배경

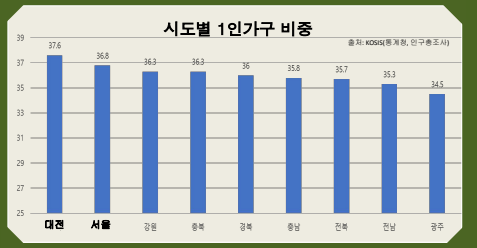
1인가구 증가율 OECD 국가 중 상위권

1인가구는 사회, 경제적으로 많은 현상을 야기함

※1인가구란?
1명이 단독으로 생계를 유지하고 있는 생활단위



| 해마다 증가세를 보이는 1인가구 비율 |



서울에 거주하는 1인가구의 비율은 36.8%로
시도별 1인가구 비중에서 대전(37.6%)
다음으로 높은 순위를 차지

1 분석 개요 연구 배경

서론 본론 결론

Where is the best?

2 분석 과정 분석 방향

서론 본론 결론

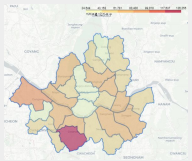
1. 데이터 확인 및 전처리

데이터 수집
데이터 전처리
데이터 결합

	자치구명	월세보증금	전세보증금	월세	역개수
0	용산구	3000.0	26000.0	55.0	18
1	성동구	4000.0	37800.0	52.0	19
2	광진구	2000.0	22000.0	45.0	11
3	성북구	3000.0	26000.0	45.0	9
4	강북구	1000.0	15750.0	40.0	3
5	중랑구	2000.0	18900.0	35.0	13

2. 시각화 & EDA

자치구별 비교
지도에 데이터 시각화
변수 간 상관분석



3. 가중치 산정

AHP 기법
회귀모델
가중치 산정

	경제	교통	범죄	여가시설	음식점	가중치
경제	0.282	0.307	0.288	0.286	0.284	0.291
교통	0.195	0.205	0.235	0.194	0.209	0.207
범죄	0.167	0.144	0.164	0.173	0.180	0.165
여가시설	0.170	0.177	0.158	0.167	0.157	0.166
음식점	0.175	0.167	0.155	0.181	0.170	0.170

4. 추천 자치구 선정

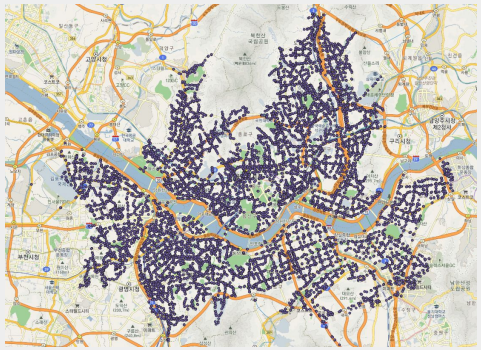
가중치를 이용하여 최종 자치구 선별
상위 5개
하위 5개

최종순위
마포구 1.282
강서구 1.094
노원구 0.610
영등포구 0.552
종로구 0.524
...

2 분석 과정 데이터 수집 및 전처리

서론 본론 결론

역geocoding



gps	주소
126.9740118, 37.5352699	용산구
127.0158952, 37.54530457	성동구
127.0797293, 37.56974605	광진구
127.050133, 37.623825	성북구
127.0080441, 37.60870425	성북구

버스정류소 위치정보 데이터는 위도, 경도로 구성되어 있어
역geocoding하여 자치구별로 정리

2 분석 과정 데이터 취합

서론 본론 결론

	자치구명	전월세구분	보증금(만원)	임대료(만원)
0	용산구	월세	500	33
1	용산구	전세	12000	0
2	마포구	월세	10000	55
3	마포구	전세	11000	0
4	금천구	월세	500	50

park

	자치구명	공원수
5	종로구	104
6	중구	76

```
df_4 = pd.merge(df_3, df_trans_satis)
df_5 = pd.merge(df_4, market)
df_6 = pd.merge(df_5, df_crime)
df = pd.merge(df, movie, on='자치구명')
df = pd.merge(df, ingu)
df = pd.merge(df, restaurant)
```

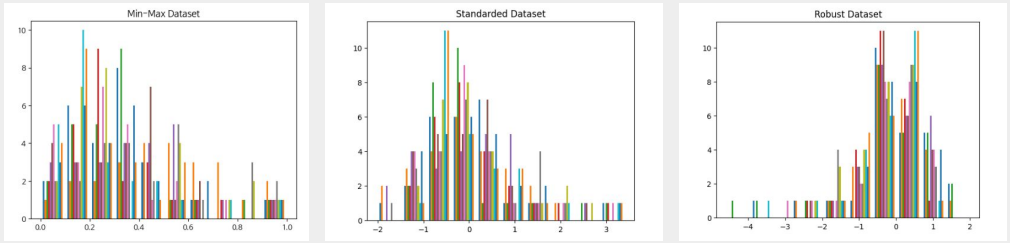
...

	자치구명	총1인가구수	버스정류장수	전세보증금	월세	환승체외	환승역	슈퍼개수	공원수	영화관수	커피-음료	범죄5대	선호가게수
0	용산구	39270	351	26000.0	55.0	12.25	5.75	598	108	2	847	2381	1321.255
1	성동구	44946	463	37800.0	52.0	15.00	4.00	482	88	2	759	2112	1284.579
2	광진구	66140	288	22000.0	45.0	9.00	2.00	538	68	5	823	3087	1504.932
3	성북구	64985	608	26000.0	45.0	9.00	0.00	632	122	3	725	2411	1373.904
4	강북구	48428	420	15750.0	40.0	3.00	0.00	500	84	3	480	2301	1245.598

2 분석 과정 정규화

서론 **본론** 결론

Z-score scaler



Min-Max scaler　Z-score scaler　Robust scaler

세 가지 scaler를 적용해본 결과, 가장 정규성을 띄며 극단 값을 보정해 줄 수 있는 **Z-score scaler** 선택

☑ Z-score scaling

: 데이터가 표준 정규 분포(gaussian distribution)에 해당하도록 값을 바꿔주며 이상치(outlier)를 잘 처리하는 특성이 있다.

$$z = \frac{x - \mu}{\sigma}$$

2 분석 과정 정규화

서론 **본론** 결론

```python
def normalize(df):
    result = df.copy()
    for feature_name in df.columns:
        std_value = df[feature_name].std()
        mean_value = df[feature_name].mean()
        result[feature_name] = (df[feature_name] - mean_value) / std_value
    return result
```

Column 단위로 정규화

```python
df_scaled = normalize(df1)
df_scaled.head()
```

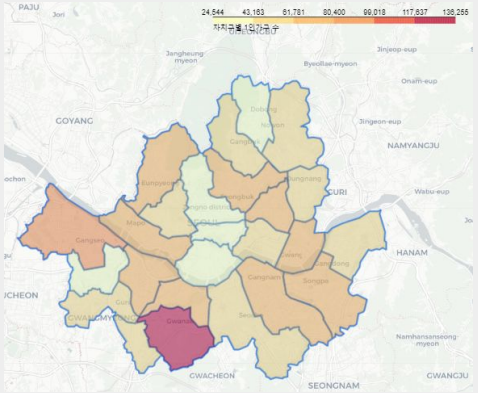
	총1인가구수	버스정류장수	전세보증금	월세	환승제외	환승역	슈퍼개수	공원수	영화관수	커피-음료	범죄5대	선호가게수
0	-0.902313	-0.938385	0.070274	0.815125	0.224459	1.221736	-0.317816	-0.279826	-0.746851	-0.151760	-0.878612	-0.625831
1	-0.650340	0.102200	1.568122	0.484669	0.770710	0.432792	-0.948661	-0.822123	-0.746851	-0.328599	-1.161054	-0.685312
2	0.290518	-1.523715	-0.437470	-0.286395	-0.421110	-0.468858	-0.644115	-1.364421	0.586812	-0.199989	-0.137336	-0.327940
3	0.239245	1.449387	0.070274	-0.286395	-0.421110	-1.370509	-0.132913	0.099783	-0.302297	-0.396923	-0.847113	-0.540444
4	-0.495765	-0.297310	-1.230822	-0.837155	-1.612929	-1.370509	-0.850771	-0.930583	-0.302297	-0.889258	-0.962610	-0.748533

2 분석 과정 EDA

서론 **본론** 결론

자치구별 1인가구 수 데이터 시각화

☑ EDA - 데이터



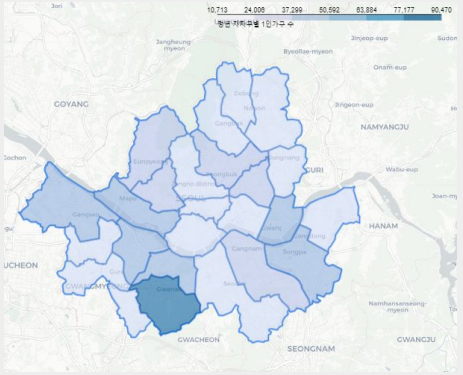
1인가구가 많이 사는 자치구는
차례로
관악구, 강서구, 송파구

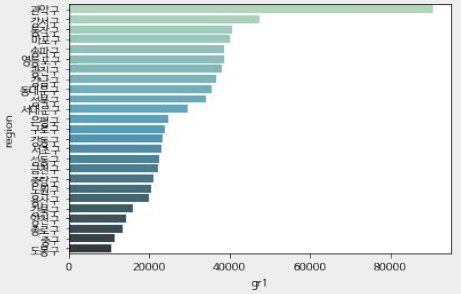
2 분석 과정 EDA

서론 **본론** 결론

청년층(0-30대) 자치구별 1인가구 수

☑ EDA - 데이터





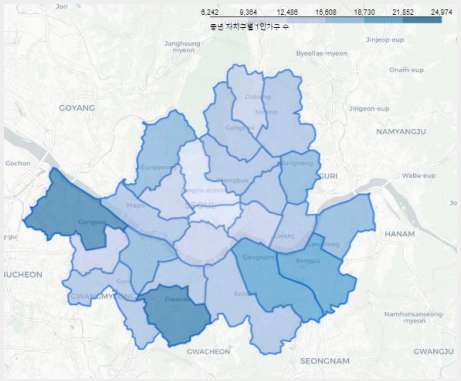
청년층 1인가구가 제일 많이 사는
자치구는 관악구

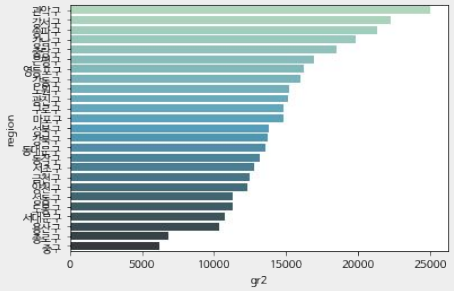
2 분석 과정 EDA

서론 **본론** 결론

중년층(40-50대) 자치구별 1인가구 수

☑ EDA - 데이터





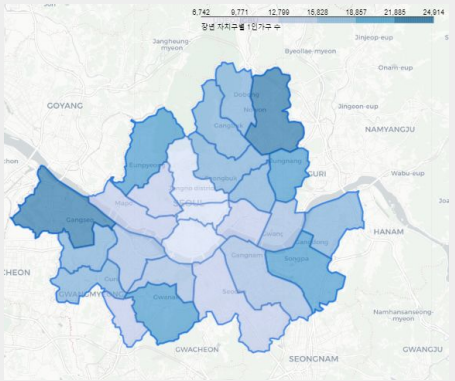
중년층 1인가구가 제일 많이 사는 자치구는 **관악구, 강서구**

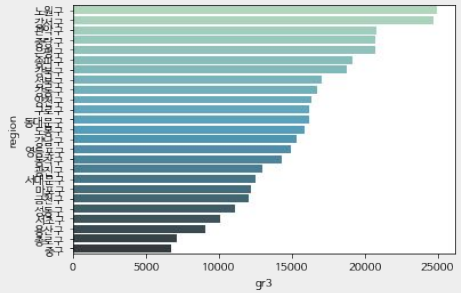
2 분석 과정 EDA

서론 **본론** 결론

장년층(60대 이상) 자치구별 1인가구 수

☑ EDA - 데이터





장년층 1인가구가 제일 많이 사는 자치구는 **노원구, 강서구**

2 분석 과정 AHP기법을 이용한 가중치 산정

서론 **본론** 결론

☑ **AHP 기법이란?**

다수의 분야들을 계층적으로 분류하여 각 분야의 중요도를 파악함으로써 최적 대안을 선정하는 기법

$$G = \sqrt[n]{a_1 a_1 \cdots a_n}$$

(여러 응답자들의 결과를 취합할 때 기하평균(Geometric Mean) 이용)

* 대상을 A_n, 각각의 가중치를 w_n로 정의하고 n개의 대상 간 이원비교행렬 표현

$$A = \begin{array}{c} \\ A_1 \\ A_2 \\ A_3 \\ A_4 \end{array} \begin{array}{c} \begin{array}{cccc} A_1 & A_2 & A_3 & A_4 \end{array} \\ \begin{pmatrix} w_1/w_1 & w_1/w_2 & w_1/w_3 & w_1/w_4 \\ w_2/w_1 & w_2/w_2 & w_2/w_3 & w_2/w_4 \\ w_3/w_1 & w_3/w_2 & w_3/w_3 & w_3/w_4 \\ w_4/w_1 & w_4/w_2 & w_4/w_3 & w_4/w_4 \end{pmatrix} \end{array}$$

2 분석 과정 AHP기법을 이용한 가중치 산정

서론 **본론** 결론

☑ 가중치 산정

분야별 평균을 계산해서 표로 작성하였고, 각 행의 평균을 계산하여 분야별 가중치를 산출하였다.

	교통	경제	범죄	여가시설	음식점	가중치
경제	0.292	0.307	0.288	0.286	0.284	0.291
교통	0.195	0.205	0.235	0.194	0.209	0.207
범죄	0.167	0.144	0.164	0.173	0.180	0.165
여가시설	0.170	0.177	0.158	0.167	0.157	0.166
음식점	0.175	0.167	0.155	0.181	0.170	0.170

2 분석 과정
AHP기법을 이용한 가중치 산정

서론 본론 결론

```
df_norm = normalize(df_num)
df_norm
```

	버스정류장수	전세보증금	월세	환승제외	환승역	슈퍼개수	공원수	영화관수	커피-음료	범죄5대	선호가게수
0	-0.938385	0.070274	0.815125	0.224459	1.221736	-0.317816	-0.279826	-0.746851	-0.151760	-0.879612	-0.625831
1	0.102200	1.568122	0.484669	0.770710	0.432792	-0.948661	-0.822123	-0.746851	-0.328599	-1.161054	-0.685312
2	-1.523715	-0.437470	-0.286395	-0.421110	-0.468858	-0.644115	-1.364421	0.566812	-0.199989	-0.137336	-0.327940
3	1.449387	0.070274	-0.286395	-0.421110	-1.370509	-0.132913	0.099783	-0.302297	-0.396923	-0.847113	-0.540444
4	-0.297310	-1.230822	-0.837155	-1.612929	-1.370509	-0.850771	-0.930583	-0.302297	-0.889258	-0.962610	-0.748533
5	-0.603911	-0.830973	-1.387915	-0.222473	-0.018033	-0.361322	-0.252711	-0.302297	-0.877200	-0.008190	-0.510717
6	-0.371638	-0.056662	0.264365	-0.023836	0.432792	0.628453	-0.388285	1.920474	0.865063	-0.531073	0.518028
7	-2.016135	0.070274	1.365885	0.522414	2.799625	3.043068	-1.147502	1.920474	0.549567	-0.374628	0.554832
8	-1.263568	-0.437470	-0.286395	-0.619746	-0.018033	0.329345	-0.713664	-1.191405	-0.519504	-0.271731	-0.148722
9	-0.715403	-0.812279	-0.506699	-1.017019	-0.468858	-1.329344	-1.039042	-1.191405	-1.116335	-1.425645	-1.229329
10	0.845476	-0.288193	-0.286395	0.373437	-0.018033	-0.323254	1.943594	0.566812	-0.509456	0.217553	-0.606870
11	0.492420	-0.569794	-0.837155	0.174800	0.432792	-0.203611	0.479391	-0.746851	-0.515485	0.027509	-0.468563
12	0.092909	-0.056662	0.264365	-1.017019	-0.919684	-0.910593	0.154013	0.566812	-0.437113	-0.986759	-0.750036

```
stand_df = pd.DataFrame()
```

```
stand_df["표준화교통합"] = df_norm["버스정류장수"] + df_norm["환승제외"] + df_norm["환승역"]
stand_df["표준화경제합"] = df_norm["전세보증금"] + df_norm["월세"]
stand_df["표준화여가합"] = df_norm["공원수"] + df_norm["영화관수"]
stand_df["표준화음식합"] = df_norm["커피-음료"] + df_norm["선호가게수"]
stand_df["표준화음식합"] = df_norm["커피-음료"] + df_norm["선호가게수"]
stand_df["표준화범죄합"] = df_norm["범죄5대"]
```

```
stand_df
```

	표준화교통합	표준화경제합	표준화여가합	표준화음식합	표준화범죄합
0	0.507811	0.885399	-1.026677	-0.777590	-0.878612
1	1.305703	2.052791	-1.568974	-1.013911	-1.161054
2	-2.413682	-0.723866	-0.777609	-0.527928	-0.137336
3	-0.342231	-0.216121	-0.202514	-0.937366	-0.847113
4	-3.280748	-2.067977	-1.232879	-1.637790	-0.962610
5	-0.844417	-2.218888	-0.555008	-1.387918	-0.008190
6	0.037318	0.207703	1.532189	1.383091	-0.531073
7	1.305905	1.436160	0.772973	1.104399	-0.374628
8	-1.901347	-0.723866	-1.905069	-0.668226	-0.271731
9	-2.201280	-1.324978	-2.230447	-2.345663	-1.425645
10	1.200880	-0.574589	2.530406	-1.116326	0.217553

2 분석 과정
AHP기법을 이용한 가중치 산정

서론 본론 결론

stand_df

	표준화교통합	표준화경제합	표준화여가합	표준화음식합	표준화범죄합
0	0.507811	0.885399	-1.026677	-0.777590	-0.878612
1	1.305703	2.052791	-1.568974	-1.013911	-1.161054
2	-2.413682	-0.723866	-0.777609	-0.527928	-0.137336
3	-0.342231	-0.216121	-0.202514	-0.937366	-0.847113
4	-3.280748	-2.067977	-1.232879	-1.637790	-0.962610
5	-0.844417	-2.218888	-0.555008	-1.387918	-0.008190
6	0.037318	0.207703	1.532189	1.383091	-0.531073
7	1.305905	1.436160	0.772973	1.104399	-0.374628
8	-1.901347	-0.723866	-1.905069	-0.668226	-0.271731
9	-2.201280	-1.324978	-2.230447	-2.345663	-1.425645
10	1.200880	-0.574589	2.530406	-1.116326	0.217553

```
weight = np.array([0.207, -0.291, 0.166, 0.170, -0.165])
weight
```

```
array([ 0.207, -0.291,  0.166,  0.17 , -0.165])
```

```
score_df = pd.DataFrame()
```

```
for i in range(len(stand_df)) :
    score_df.loc[i, ["Total_score"]] = np.dot(stand_df.iloc[i], weight.T)
```

$$(x_i, \alpha_i, \beta_i, \gamma_i, \delta_i) \bullet w^T$$

행렬의 내적을 이용하여 총점 Yi 계산

2 분석 과정
AHP기법을 이용한 가중치 산정

서론 본론 결론

☑ 가중치를 이용한 최종 점수 도출

$$Yi = 0.291xi + 0.207\alpha i + 0.165\beta i + 0.166\gamma i + 0.170\delta i$$

Yi : i번째 자치구의 최종점수 (i = 1, 2, 3, …, 25)

xi: i번째 자치구의 경제 점수

αi: i번째 자치구의 교통 점수

βi: i번째 자치구의 범죄 점수

γi: i번째 자치구의 여가시설 점수

δi: i번째 자치구의 음식점 점수

2 분석 과정
AHP기법을 이용한 가중치 산정

서론 본론 결론

	자치구명	표준화교통합	표준화경제합	표준화여가합	표준화음식합	표준화범죄합	Total_score
상위 5개	마포구	3.229357	0.775247	2.172230	3.147132	0.338299	1.282663
	강서구	2.710828	-0.723866	2.069453	0.566700	0.712088	1.094160
	노원구	1.200880	-0.574589	2.530406	-1.116326	0.217553	0.610163
	영등포구	1.955616	-0.317670	-0.582122	1.871042	1.009228	0.552177
	종로구	0.037318	0.207703	1.532189	1.383091	-0.531073	0.524379
하위 5개	동대문구	-1.901347	-0.723866	-1.905069	-0.668226	-0.271731	-0.567938
	성동구	1.305703	2.052791	-1.568974	-1.013911	-1.161054	-0.568322
	도봉구	-2.201280	-1.324978	-2.230447	-2.345663	-1.425645	-0.603882
	양천구	-3.465277	-0.513009	-0.619953	-1.604734	-0.212933	-0.908610
	금천구	-3.313052	-1.133552	-2.463765	-1.663279	-0.817714	-0.912758

2 분석 과정
AHP기법을 이용한 가중치 산정

	교통	경제	여가	음식	범죄	최종순위
1위	3.229	0.775	2.172	3.147	0.338	마포구 1.282
2위	2.710	−0.723	2.069	0.556	0.712	강서구 1.094
3위	1.200	−0.574	2.530	−1.116	0.217	노원구 0.610
4위	1.955	−0.317	−0.582	1.871	1.009	영등포구 0.552
5위	0.037	0.207	1.532	1.383	−0.531	종로구 0.524