소프트웨어 응용
최종보고서

Airbnb 로 돈 벌기 : 어디가 좋을까?
클러스터링 및 시각화 엔진

2015920005 김대현
2015920060 현민지

1. 개발 과제 개요

에어비앤비 숙소들의 상관관계를 클러스터링 기법을 활용하여 분석하고, 숙소의 위치와 분석 결과를 시각화하여 지도에 나타낸다.

2. 개발 과정

- 에어비앤비 빅데이터 수집 (Data_collection.py)
  - 다운로드

Inside Airbnb(http://insideairbnb.com/get-the-data.html)에서 에어비앤비의 빅데이터를 쉽게 얻을 수 있다. 리스트에 있는 도시들 중 미국 28 개 도시의 에어비앤비 데이터(csv 파일)를 다운로드 받아 프로젝트에 활용한다. 데이터는 257252 개의 row 와 106 개의 column 로 이루어져있다. 각 row 는 하나의 숙소에 대한 정보를 나타내며, column 은 숙소 이름, 가격, 위치 등의 특징을 나타낸다.

- 수집한 데이터 정제 (Data_refining.py)
  - 필요 없는 row, column 제거
  - Mixed type column 처리
  - 문자열을 숫자로 변환
  - Categorical data 를 Numerical data 로 변환
  - 특수 기호 제거

- 차원 축소 (Scikit-learn 패키지 활용, dimension_reduction.py)

몇 백 차원의 데이터를 그대로 계산을 진행하면, 연산량이 상당히 많아서 계산이 복잡해지고, 차원의 저주로 인해 정확도가 오히려 떨어지게 된다. 또한, 연산 결과를 시각화하여 확인하기 위해서는 2~3 차원으로 축소를 시킬 필요가 있다.

  - 특징 선택

categorical data 를 numerical data 로 변환하면서, sparse 한 column 의 수가 급격히 증가했다. 무의미한 column 을 제거하기 위해 boolean 타입의 column 들에 대해 특징 선택을 실시하였다. 분산이 Var[X]=p(1-p) (p=0.8) 에 미치지 못하는 column 들을 제거하였다. 그 결과, 325 차원의 데이터가 89 차원으로 축소되었다.

  - 표준화

차원 축소를 진행할 때, 각 column 에 표준화(평균 0, 분산 1)를 적용시킴으로써, 각 column 의 크기에 대해 편향된 결과가 나타나지 않도록 하였다.

  - PCA

PCA 를 실시하여 89 차원의 데이터를 36 차원으로 축소하였다.

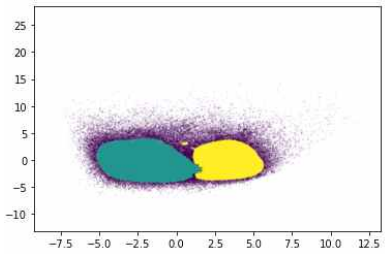
  - FAMD

categorical 데이터를 차원축소하는 MCA 와 PCA 를 결합한 차원 축소 방식인 FAMD 를 시도해보았으나, inverse 연산이 매우 어려워 복원을 통해 데이터 손실률을 확인하기가 힘들었다. 따라서, 이 과제에서는 FAMD 방식을 사용하지 않았다.

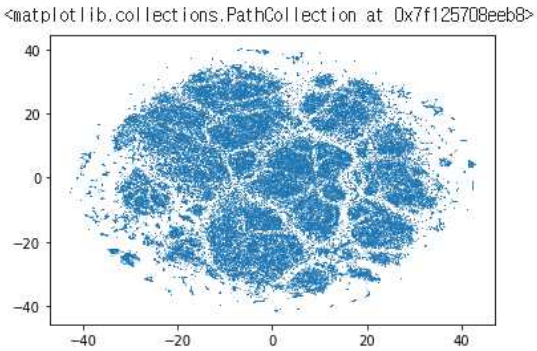
  - t-SNE

특징 선택과 PCA 를 거쳐 얻게 된 36 차원의 데이터는 클러스터링 연산 시에 시간적, 공간적 비용이 매우 크기 때문에 데이터를 더 축소해야 할 필요성이 컸다.

PCA 방식으로 데이터를 일정 수준 이하로 축소시킬 경우 원본 데이터의 손실이 커지기 때문에 데이터를 시각적으로 표현 가능한 정도로 축소시키기 어렵다는 단점이 있다. 다음은 PCA 방식으로 데이터를 2 차원으로 축소시킨 후 클러스터링 및 시각화한 결과이다. PCA 방식만을 사용해 축소한 데이터는 원본 데이터의 특성을 제대로 반영하고 있다고 보기 매우 어렵다.

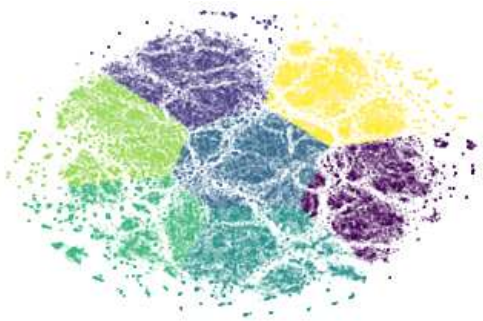


차원 축소시 데이터의 본래 특성을 잘 유지하는 것으로 알려진 t-SNE 를 활용함으로써, 36 차원의 데이터를 한번 더 축소하여 2 차원으로 만들고 이를 클러스터링에 사용하였다.



- 클러스터링 (Scikit-learn 패키지 활용, clustering.py)
  - K-means

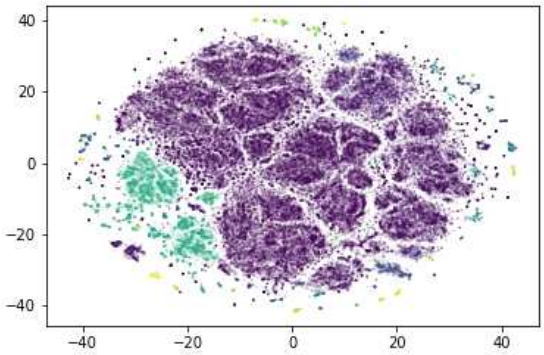
K-means 알고리즘은 저차원 데이터의 클러스터링에 좋은 성능을 보인다. 그러나 이 알고리즘은 데이터의 군집이 원형인 것으로 가정하기 때문에 우리가 분석하는 데이터에 적절하지 않았다. 클러스터 개수를 7 개로 하여 K-means 클러스터링을 실시한 결과는 다음과 같다.



원래의 데이터 군집 경계와 K-means 클러스터링 결과 군집 경계가 전혀 일치하지 않는다.

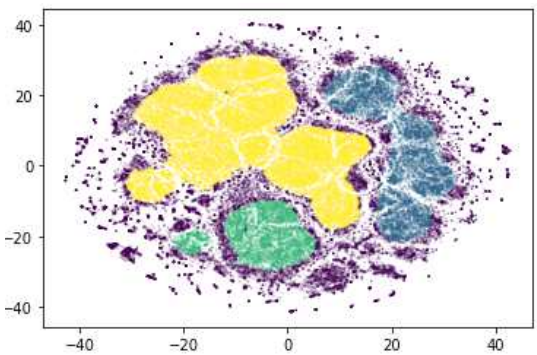
○ DBSCAN

DBSCAN 은 밀도 기반의 클러스터링 방식으로, 군집의 모양에 큰 영향을 받지 않는다. 클러스터의 개수를 지정하지 않는다는 점도 이 알고리즘의 특징인데, 이 특징은 오히려 과제 수행에 장애물이 되었다. 시각화 과정을 통해 데이터의 분포를 어느 정도 예상할 수 있음에도 불구하고, 클러스터의 개수를 지정하지 않고 파라미터들(epsilon, min_samples)의 값만을 세세히 조절하며 최적의 결과를 내놓는 파라미터 값을 찾아내는 것은 굉장히 어려운 작업이었다. 시각화된 데이터의 모습을 바탕으로 직관적으로 예상한 클러스터 개수, 모양과 비슷한 DBSCAN 결과를 얻을 수 없었다.



○ HDBSCAN

HDBSCAN 은 Hierarchical DBSCAN 로써, 기존의 DBSCAN 에 계층적 클러스터링 방식을 적용시킨 클러스터링 기법이다. 일반적인 클러스터링에 비해서, 계층적 클러스터링 방식은 거리 계산법에 따라서 차원의 저주를 덜 받게 된다. 그러나 차원의 저주를 덜 받는 형식으로 알려진 코사인 유사도 방식을 HDBSCAN 에 적용할 때, 많은 연산량으로 인해서 COLAB 에서 제공하는 RAM 을 다운시키게 된다. 그래서 코사인 유사도 방식을 적용시키지 못하고 t-SNE 로 차원 축소된 데이터에 유클리디언 거리 방식을 적용시켰다. 적용시킨 결과, 클러스터에 속하지 못한 garbage 클러스터가 상당히 많이 나왔으며, 클러스터링이 정확하게 진행되지 못했다.

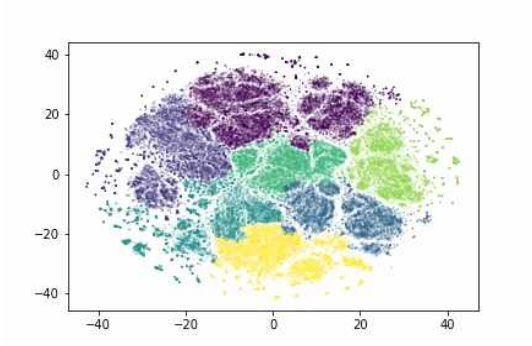


○ birch

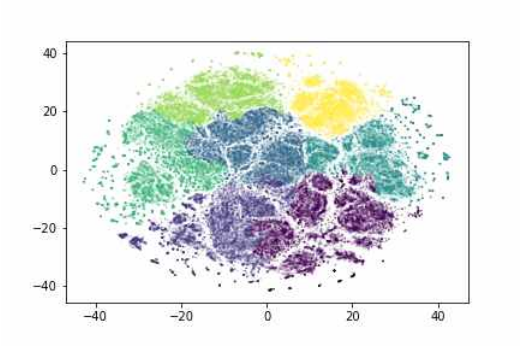
birch 는 계층적 클러스터링의 일환으로, 주어진 메모리를 활용하여 클러스터링을 가장

효율적으로 진행할 수 있는 클러스터링 기법이다. birch 는 데이터를 한 번만 스캔하여 그에 맞는 CF 트리(클러스터 데이터의 개수, 각 데이터 제곱의 합, 각 데이터의 합)를 만들고, CF 트리를 정제하며 원하는 형식의 클러스터링을 진행하는 방식이다. birch 클러스터링을 선정하게 된 이유는 크게 3 가지로, 어지럽혀진 클러스터의 모양을 효과적으로 표현할 수 있다는 점, 클러스터의 개수를 임의로 지정할 수 있다는 점, 마지막으로 큰 데이터에 대해 효율적으로 메모리를 사용할 수 있다는 점이다.

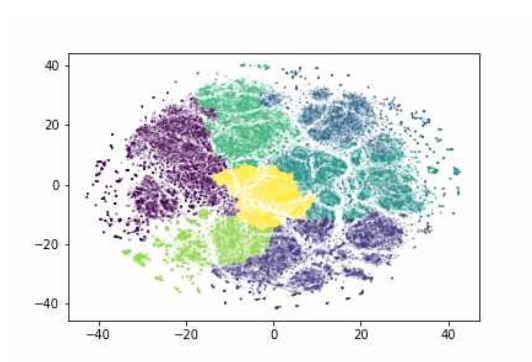
birch 에는 2 개의 파라미터를 사용하였다. n_clusters 로 클러스터의 개수를 지정하였고, threshold 로 서브 클러스터를 융합할지, 새로운 클러스터를 형성할지 결정한다. n_clusters 가 7 일 경우 가장 이상적인 수치가 나왔기 때문에, 클러스터의 수를 7 로 고정했다.



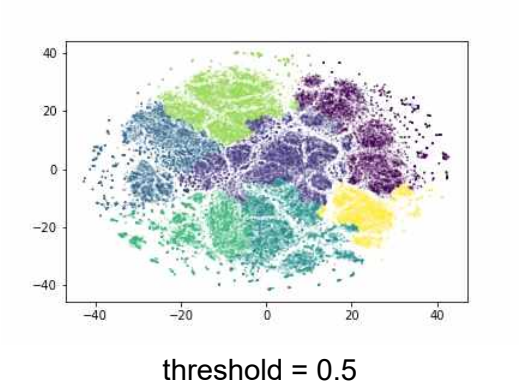
threshold = 0.2



threshold = 0.3



threshold = 0.4



threshold = 0.5

그림을 확인할 때, threshold 가 0.2 인 클러스터링이 가장 정교하게 되었다고 판단되었다. 따라서, threshold 가 0.2 인 birch 클러스터링을 최종 클러스터링으로 결정하였다.

- 클러스터링 결과 검증 (make_pairs.py, clustering.py)

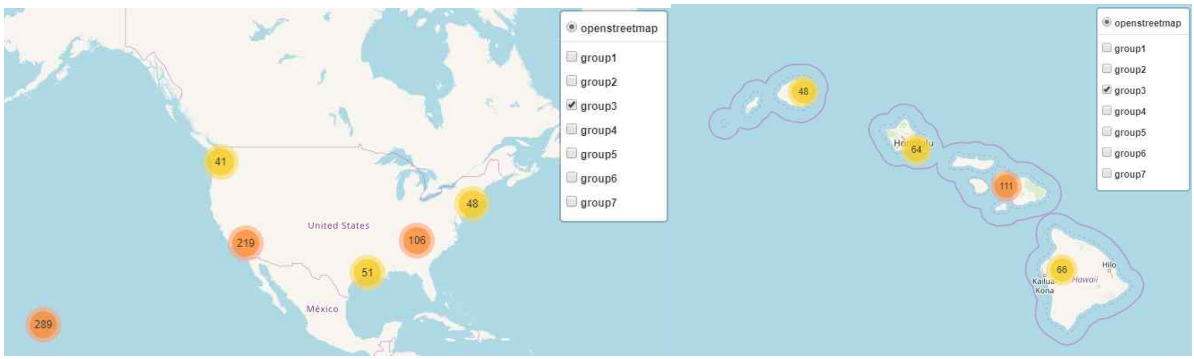
클러스터링이 잘 진행되었는지 검증하기 위해 선택한 방법은 다음과 같다.

1. 특징이 매우 비슷한 두 숙소의 쌍, 특징이 매우 다른 두 숙소의 쌍들을 찾는다. 숙소 사이의 유사도는 코사인 또는 피어슨 유사도로 계산한다.
3. 유사도가 0.999 보다 크면, 같은 클러스터에 있어야 하는 쌍이라고 판단하고, 두 숙소가 같은 클러스터에 속해있다면 옳게 클러스터링되었다고 판단한다.
4. 유사도가 -0.75 보다 작으면, 다른 클러스터에 있어야 하는 쌍이라고 판단하고, 두 숙소가 다른 클러스터에 속해 있다면 옳게 클러스터링되었다고 판단한다.
5. 정확도를 계산한다.

정확도 = (옳게 클러스터링된 숙소 쌍의 개수) / (검증에 사용한 전체 숙소 쌍의 개수)

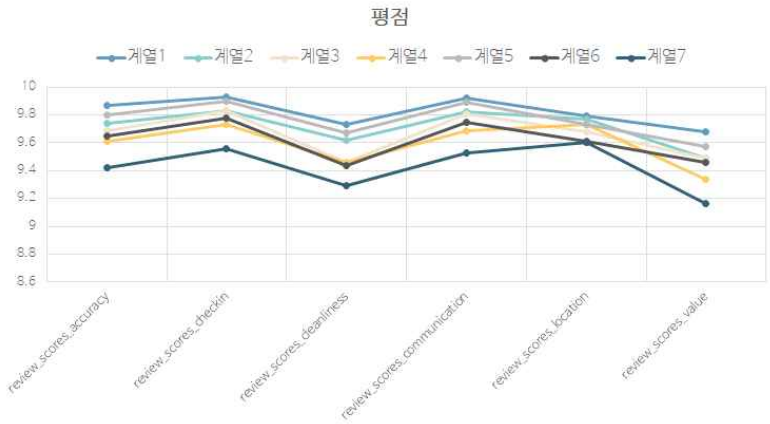
● 지도에 시각화 (folium 패키지 활용, CluMap.py, 영상 자료 참조)

클러스터링 정확도가 가장 높은 birch 클러스터링 데이터를 기반으로, 지도에 숙소의 위치를 표시하였다. 프로그램 최적화를 위해 분석에 활용한 숙소의 1/25 만을 지도에 표시했다. 각 클러스터 별로 숙소의 위치를 표시할 수 있으며 , 지도 확대시에 숙소의 구체적인 위치가 표시된다.



3. 클러스터링 분석 및 결론

(지면상의 이유로 생략/축소된 그래프는 발표 슬라이드 참고)

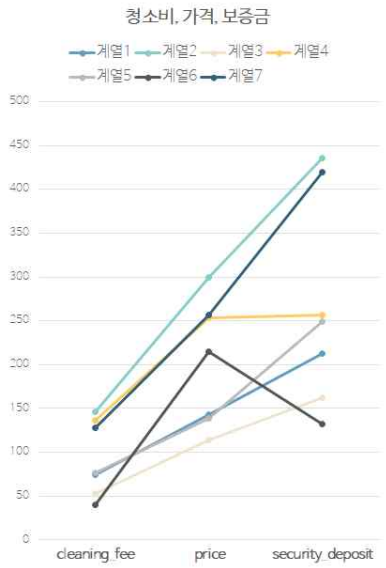


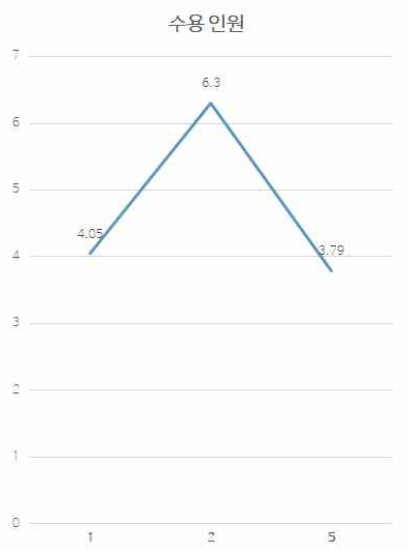
평점 상위 그룹인 1, 5, 2 번 그룹의 특징을 분석하여 성공적인 에어비앤비 호스트가 되기 위한 조건을 찾아냈다.

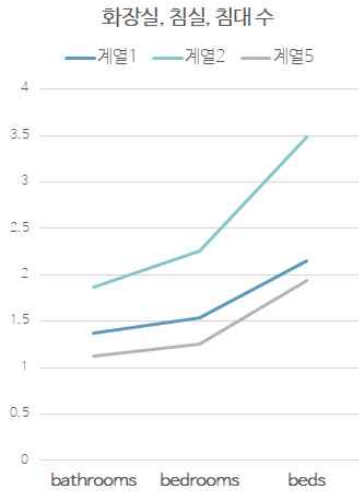
3-1. 타겟 결정

다음의 분석을 통해 에어비앤비 호스트가 어떤 고객을 타겟으로 하는 것이 합리적인지 알아냈다. 먼저, 다른 상위 그룹인 1, 5 번 그룹에 비해 특이한 경향을 보이는 2 번 그룹을 집중적으로 분석했다.

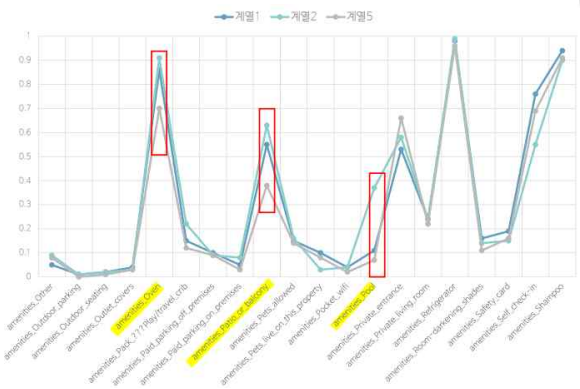
(1)

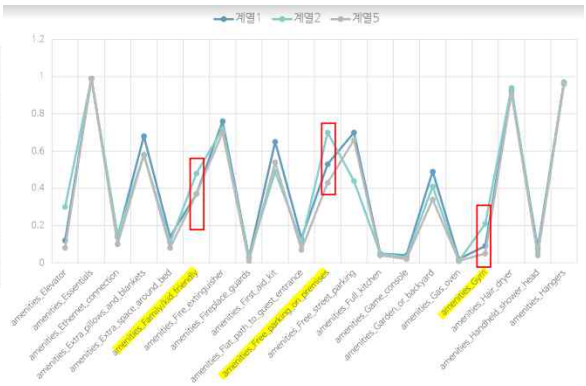






-> 2 번 그룹의 숙소들은 7 개 그룹 중 가격이 제일 비싸고, 수용 인원, 화장실 수, 침실 수, 침대 수가 많은 것을 보아 숙소의 크기가 크다.

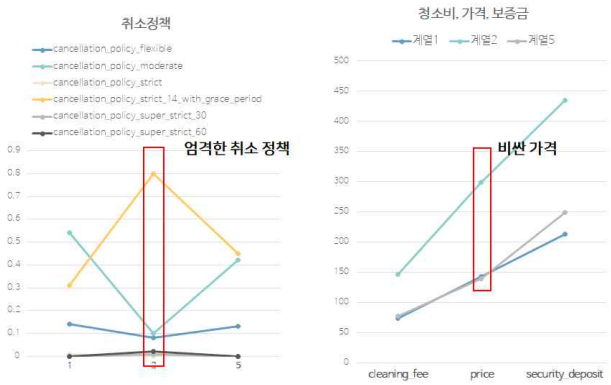




(어메니티 그래프 중략)

-> 2 번 그룹은 대체적으로 1, 5 번 그룹보다 어메니티 구비율이 낮지만, TV, 풀장, 오븐, 옥외 온수 욕조, family/kids friendly, 숙소 부지 내 무료 주차, 발코니 등의 어메니티 구비율이 특히 높다. (1), (2)를 보아 2 번 그룹은 가족 여행객을 대상으로 한 숙소 그룹임을 알 수 있다.

(3)



-> 2 번 그룹의 숙소들은 엄격한 취소 정책, 비싼 가격, 상대적으로 낮은 어메니티 구비율에도 불구하고 평가가 좋다.

(1), (2), (3)에 따르면, 가족 여행객은 개인 여행객들에 비해 돈을 더 쓰고, 숙소 서비스에 비해 만족도가 높다. 이는 숙소 제공자들로 하여금 숙소 유지를 위해 비용과 노력을 덜 들여도 괜찮도록 한다. 또한, 2 번 그룹 숙소 제공자는 1, 5 번 그룹에 비해 1.5 배 많은 인원을 수용해야 하지만, 2 배 더 높은 가격을 받을 수 있다. 따라서 에어비앤비의 호스트는 가족 여행객을 타겟으로 한 숙소를 제공하는 것이 합리적임을 알 수 있다.

3-2. 위치 결정

다음의 분석을 통해 에어비앤비 호스트가 어느 위치의 숙소를 제공하는 것이 합리적인지 알아냈다. 수용 인원 수가 비슷하고 가장 높은 평점을 받은 두 그룹(1, 5)를 비교, 분석하였다.

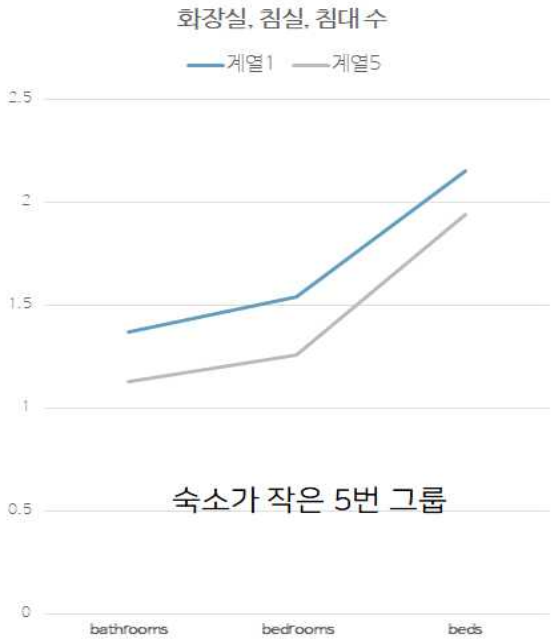
(1)

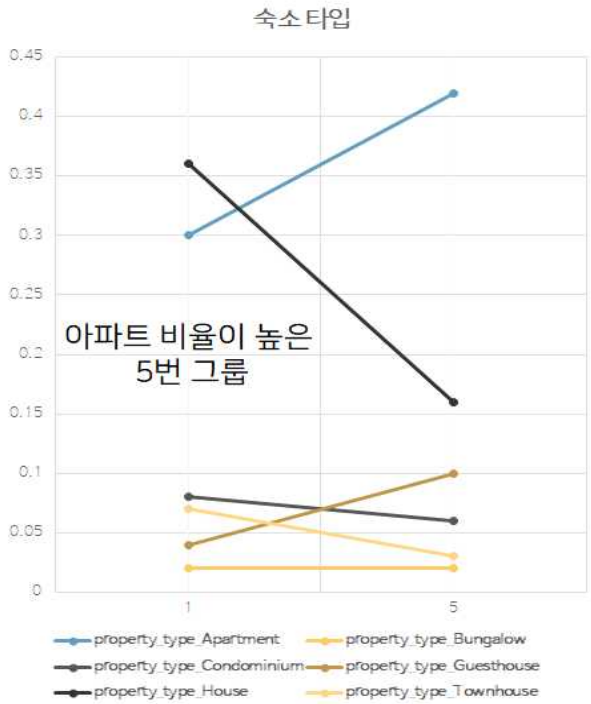
하와이에 위치한 숙소가 많은
1번 그룹

뉴욕에 위치한 숙소가 많은
5번 그룹

-> 1 번 그룹의 숙소들은 하와이, 캘리포니아에, 5 번 그룹의 숙소들은 뉴욕, 캘리포니아에 주로 위치해있다.

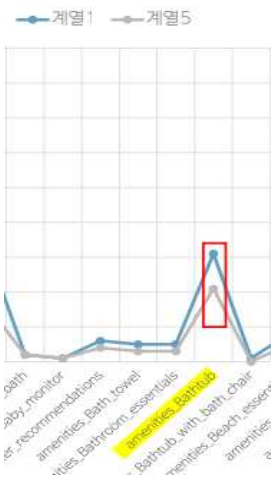
(2)

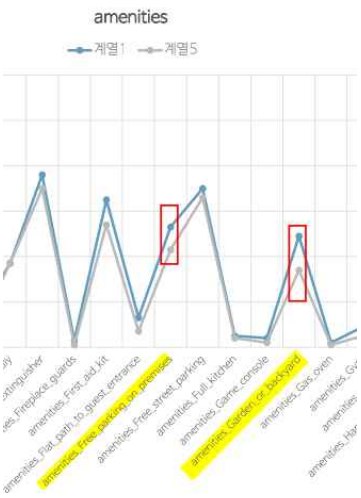




-> 화장실 수, 침실 수, 침대 수를 보아 5 번 그룹은 1 번 그룹에 비해 숙소가 작다. 또한, 1 번 그룹은 하우스 타입의 숙소가 많은 비중을 차지하지만 5 번 그룹은 아파트 타입의 숙소 비중이 매우 크다. 매우 높은 인구 밀도의 영향으로 뉴욕 시민들은 주로 아파트에 거주하고 있기 때문이다.

(3)





(어메니티 그래프 중략)
-> 1 번 그룹과 5 번 그룹의 어메니티 구비율은 매우 비슷한 경향성을 보이고 있으나, 욕조, 숙소 부지 내 무료 주차, 정원/뒤뜰 등의 어메니티 구비율에서 차이를 보이고 있다. 이는 5 번 그룹이 높은 인구 밀도에 의해 숙소 공간의 제한을 받고 있기 때문인 것으로 사료된다.

(4)

-> 5 번 그룹은 높은 월 리뷰 수(숙소 이용자 수)에 비해 평점이 낮으므로 지속적인 인기를 보장받기 어려울 것으로 예상된다.

(1), (2), (3), (4)에 따르면, 인구 밀도가 높은 도시의 숙소보다는 인구 밀도가 낮은 지역의 숙소가 고객들에게 쾌적한 경험과 더 높은 만족감을 줄 수 있다. 또한, 숙소 제공자의 입장에서는 현재 이용자 수에 비해 높은 평가를 받는 1 번 그룹의 숙소들을 벤치마킹할 필요성이 있다. 1 번과 5 번 그룹을 결정적으로 구분하는 요소는 위치이다.

종합적으로, 새로운 에어비앤비 호스트는 하와이에서 가족 여행객을 대상으로 숙박 서비스를 제공하는 것이 바람직하다는 결론을 내릴 수 있다. 평점 상위 그룹들을 분석한 결과 찾아낸 “높은 숙박 가격”, “서비스 퀄리티 대비 높은 고객 만족도”, “숙소 유지에 필요한 비용과 노력”, “전 세계에서 손꼽히는 하와이의 자연 경관” 등 여러 방면에서 긍정적인 지표들은 하와이의 부동산을 매입하는 데 필요한 높은 초기 투자비용을 상쇄하고도 남을 것이라고 생각된다.