Airbnb New York
데이터 분석

2021.12.27~2022.3.3

동아리 DNA 입부 지원 과제
수원대 3학년 20516096 차유진

Dataset 소개

- **New york Airbnb**(*Get the Data - Inside Airbnb. Adding data to the debate.*): 뉴욕시에 있는 에어비앤비 숙소 정보와 리뷰

- **Dataset1**: 호스트의 숙소 소개 글 정보 (게시글 넘버, 숙소 명, 숙소 설명, 전경, 호스트 정보(이름, 아이디, 사는 곳, 소개, 답장속도, 답장률, 슈퍼호스트 여부), 숙소 정보(이웃 지역, 위도, 경도, 인원, 방 타입, 가능 날짜, 가격, 평점(청결도, 정확도, 체크인, 소통, 위치, 가치)), 예약 수)

- **Dataset2**: 이용객의 숙소 리뷰 정보 (숙소 아이디, 이용객 아이디, 날짜, 작성자 이름, 내용)

분석 목적

1. What? 위치평점이 높은 지역을 지도 상에서 한눈에 파악하기
2. Why? 소비자는 지역 별 가격 비교를 편리하게 할 수 있고, 플랫폼은 인기 지역에서 집중적으로 호스트를 구하는 등 마케팅에 도움을 얻을 수 있다.
3. Who? 소비자, 에어비앤비 플랫폼
4. How? 데이터 마이닝, 시각화
5. Where? **Dataset1** '호스트의 숙소 소개 글 정보'

과정 요약

1. (전처리) 필요컬럼만 가져오고 결측값 제거
2. (전처리) 신뢰도가 낮은 데이터 제거
    2-1. 산점도 그려서' 리뷰 평점' 분포 파악
    2-2. Bad data 제거
    2-3. 일반적인 분포 구간만 추출
3. (시각화) Tableu 활용해서 지도 위 시각화 처리

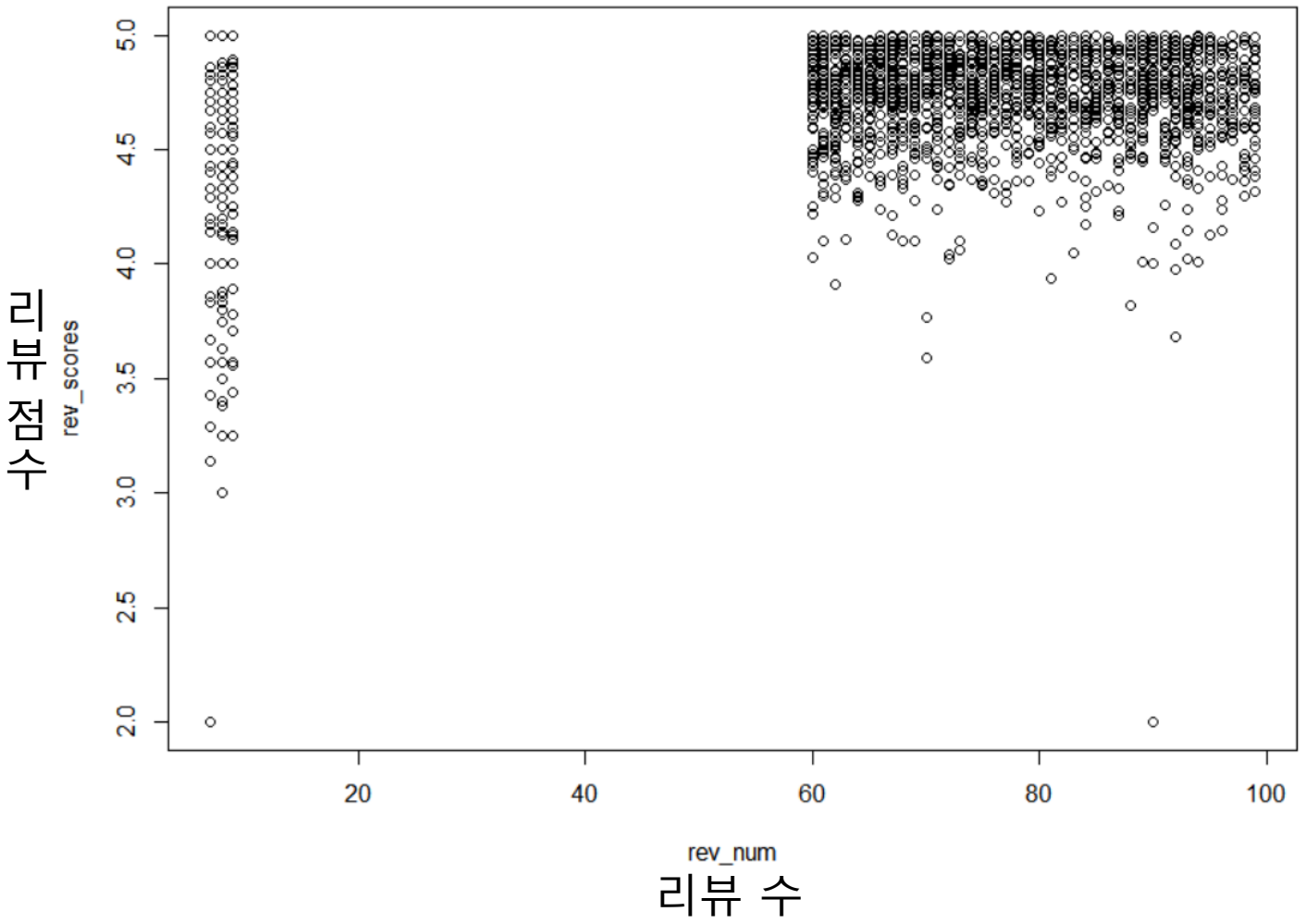
1. 필요한 컬럼만 가져오고 결측값 제거

열 추출, 결측값, 숫자형 변환

- data.frame()으로 리뷰 수, 리뷰 평점만 추출해서 새로운 데이터 셋 형성.
- na.omit()으로 리뷰 평점에 결측값이 있는 행을 모두 제거했다.
- summary()로 데이터형을 한눈에 알아보았다.
- as.numeric()을 각 열에 적용해서 숫자형으로 변형시켰다.

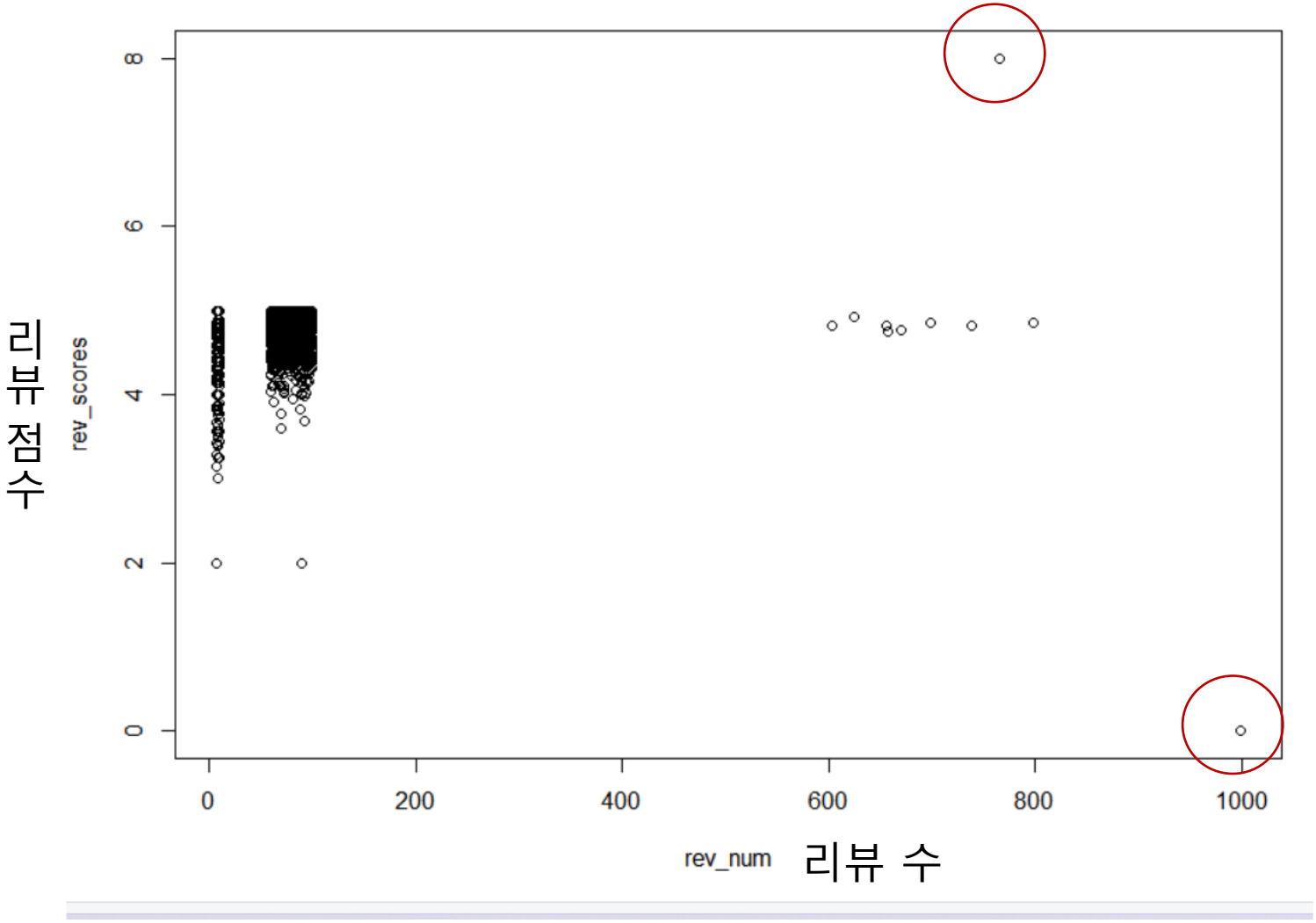
2. 신뢰도가 낮은 데이터 제거

구간 나누기(산점도)



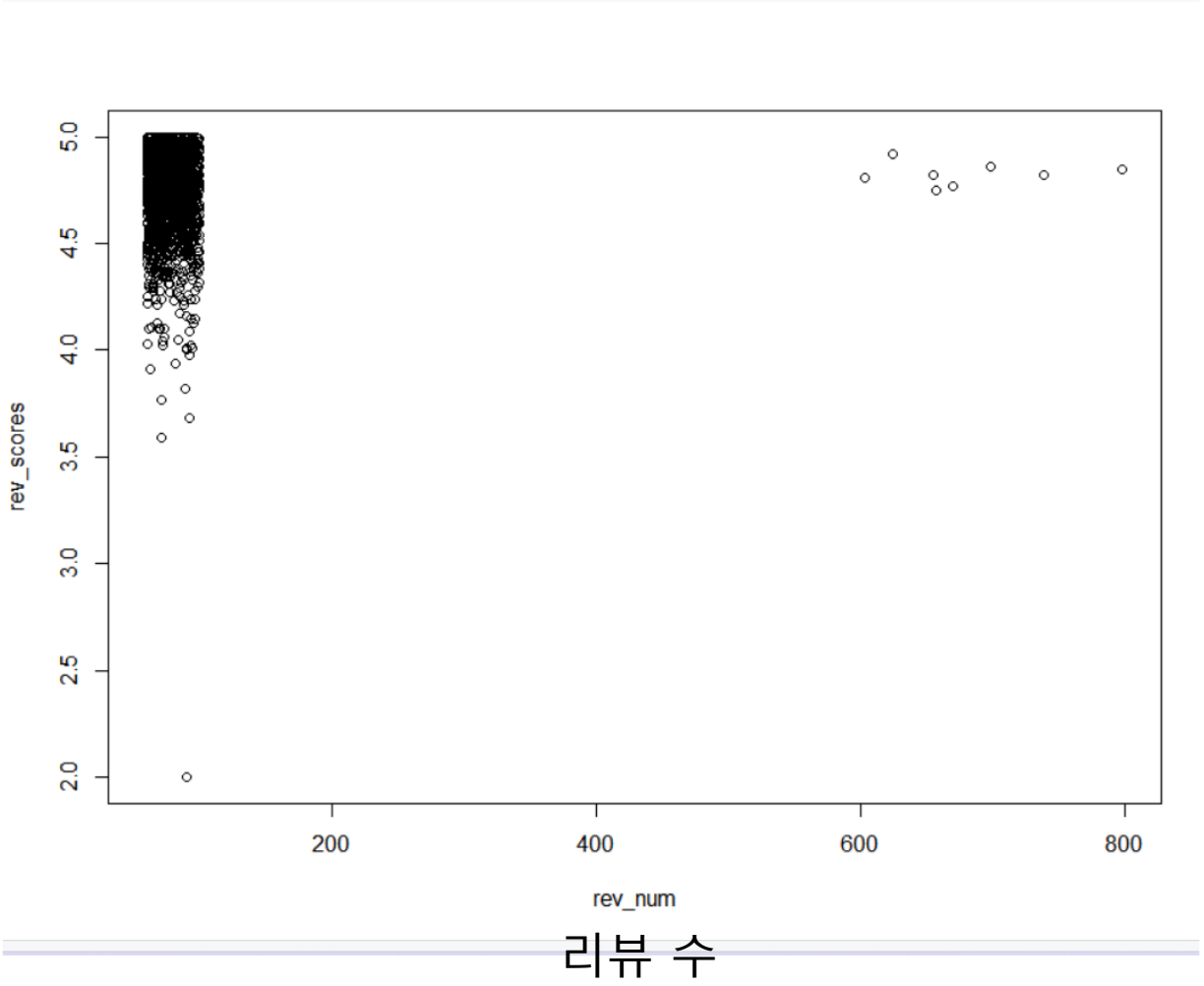
- Outliers를 제한 그래프이다.
- 리뷰 수를 x축, 리뷰 점수를 y축에 놓아 plot()으로 산점도를 그려 분포를 확인했다.
- 리뷰 수가 20 이하인 숙소들의 리뷰는 신뢰하지 못한다.
- 리뷰 수가 60 이하인 리뷰들의 분포와 많이 동떨어져있기 때문이다.

Bad Data 처리



- Bad data: 의도하지 않게 입력되었고 분석 목적에 부합되지 않아 제거해야 하는 경우
- 평점은 5점이 최댄데 저 8점은 뭐지?
- 1000명이 모두 0점을 줄 수가 있나?

쓸만한 데이터셋만 골라내기



- *평점은 5점이 최댄데 저 8점은 뭐지?*
- *1000명이 모두 0점을 줄 수가 있나?*

- subset()을 이용하여 0<rev_scores<=5인 데이터만 골라내주면 성공!

시각화용 데이터셋

neighbou	pst_locatic	host_id	longitude	latitude	num	scores
Midtown	New York,	2845	-73.9856	40.75356	48	4.86
Clinton H	New York,	4869	-73.9577	40.68494	408	4.72
Bedford-S	New York,	7356	-73.9551	40.68535	50	4.47
Hell's Kitc	New York,	8967	-73.9832	40.76457	505	4.87
Upper We	New York,	7490	-73.9675	40.8038	118	4.94
Park Slop	New York,	9744	-73.9878	40.66801	200	4.87

- host_neighborhood: 세부 주소
- host_location: 대략적인 주소
- host_id: 개체 구분(없어도 됨)
- latitude,longitude: 지도 상에 점 찍는 **위치**
- num(리뷰 수): 점의 **크기**
- scores(위치 평점): 점의 **색깔**

<Summary>

host_neighbourhood	host_location
Length:9194	Length:9194
Class :character	Class :character
Mode :character	Mode :character

host_id	longitude	latitude	num	scores
Length:9194	Min. :-74.21	Min. : 0.00	Min. : 21.00	Min. :1.000
Class :character	1st Qu.:-73.98	1st Qu.:40.68	1st Qu.: 33.00	1st Qu.:4.660
Mode :character	Median :-73.95	Median :40.72	Median : 56.00	Median :4.800
	Mean :-73.90	Mean :40.71	Mean : 82.83	Mean :4.759
	3rd Qu.:-73.92	3rd Qu.:40.76	3rd Qu.: 103.00	3rd Qu.:4.910
	Max. : 4.00	Max. :40.91	Max. :1125.00	Max. :5.000
	NA's :29	NA's :32		

3. Tableu 활용해서 지도 위 시각화 처리

Tableau - 문서1- 13일 후 Tableau 라이선스 만료

파일(F) 데이터(D) 워크시트(W) 대시보드(B) 스토리(T) 분석(A) 맵(M) 서식(O) 서버(S) 창(N) 도움말(H)

표현 방식

데이터 분석

Sheet1(file26e815d1525d)

검색

테이블

- Abc Host Id
- Abc Host Location
- Host Neighbourhood
- # Num
- Abc *측정값 이름*
- Latitude
- Longitude
- # Scores
- # *Sheet1(카운트)*
- *경도(생성됨)*
- *위도(생성됨)*
- # *측정값*

페이지

필터

마크

맵

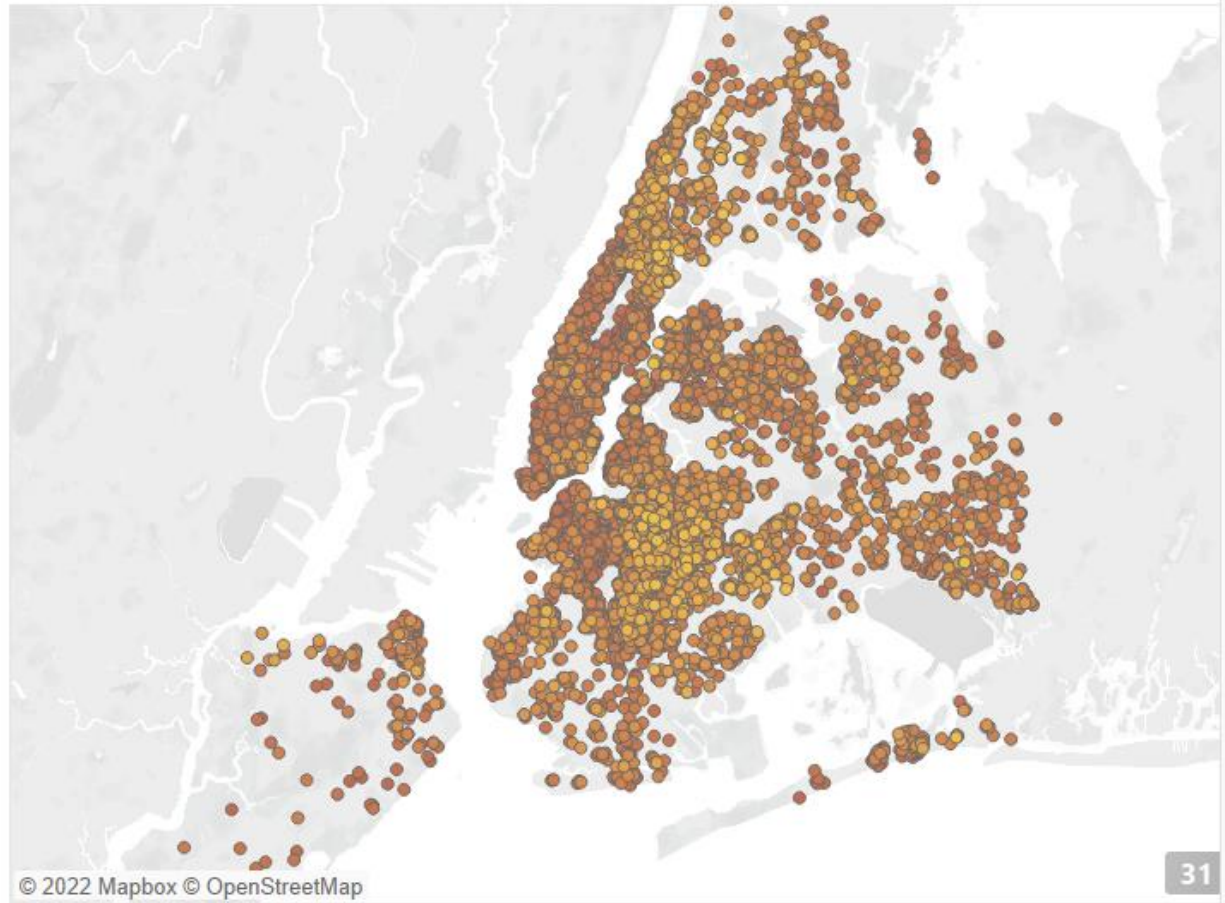
색상 크기 레이블 세부 정보 도구 설명

- Scores
- Num
- 특성(Host Locati..
- 특성(Host Neig..

iii 열 Longitude

행 Latitude

시트 3



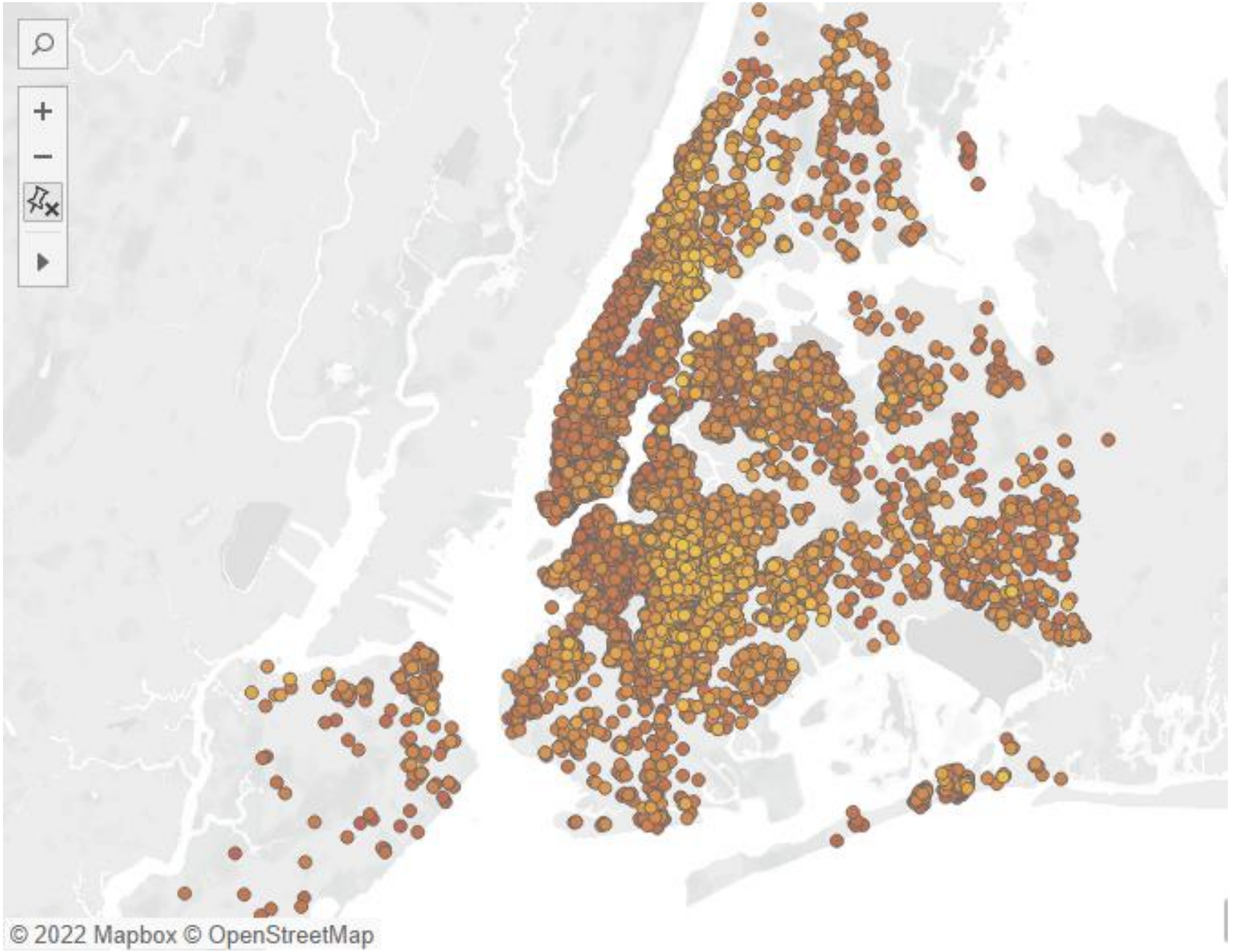
권장 사항 없음

권장 사항을 보려면 더 많은 데이터를 선택하거나 끄십시오.

데이터 원본 시트 3

마크 9162개 행 1개 x 열 1개

결과



- 해석: 붉은색>노란색
- 결론: 대체적으로 붉은색이 보이는 지역에 위치한 숙소는 가격을 더 높게 책정해도 좋을 것 같다.