빅데이터 프로그래밍

11/15

2017101319 국제학과

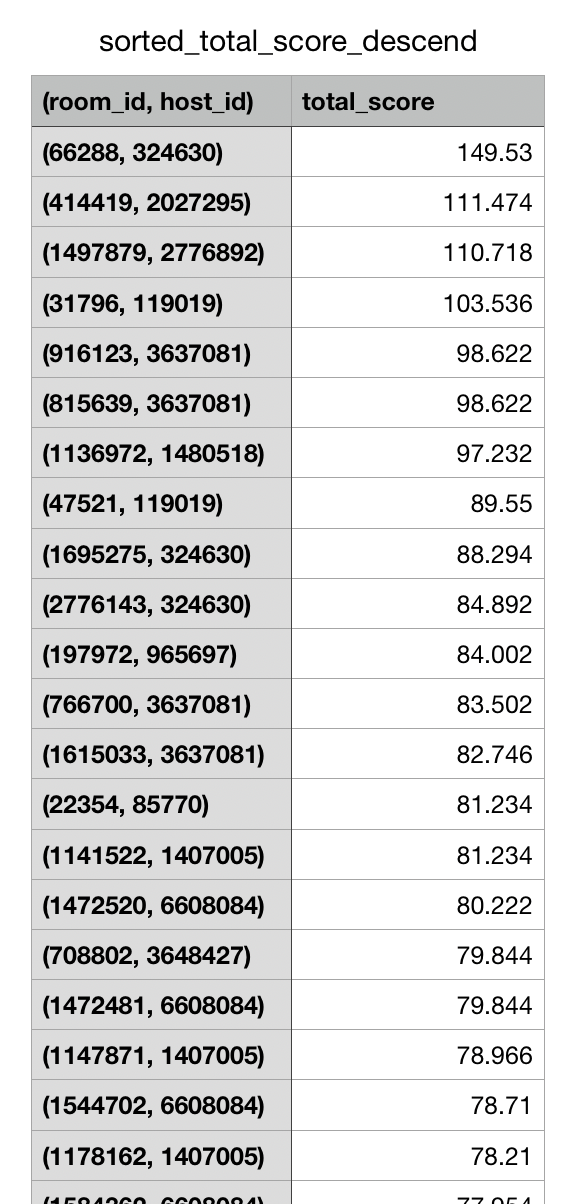
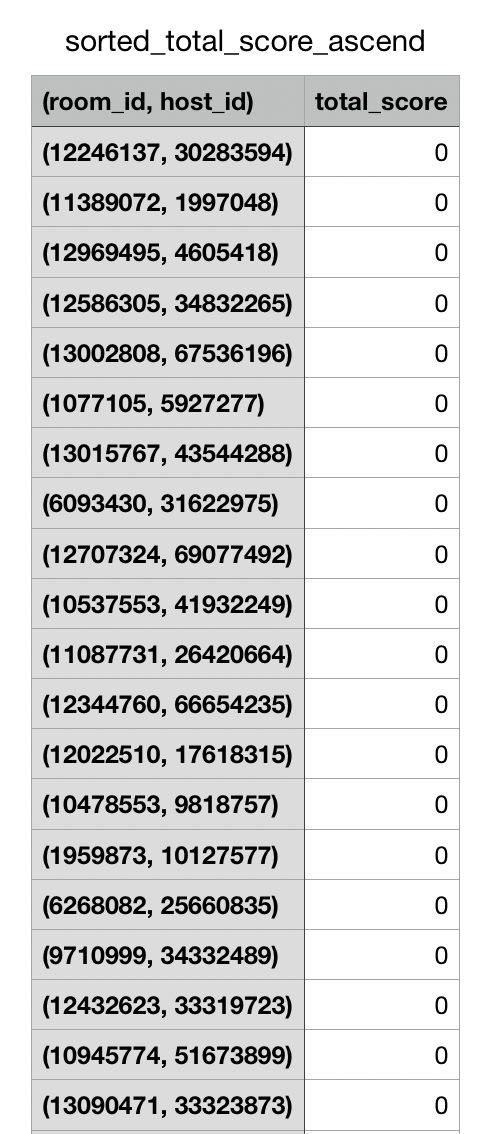
한동훈

1.

Index를 (room\_id, host\_id)로 하고, column은 total\_score로 정리

기존의 csv파일을 불러와서 room\_id, host\_id column 정보를 불러와서 하나의 변수로 합쳤다.

그리고 이를 index로 갖는 DataFrame을 만들었다. Total\_score에 해당하는 column 정보는



2. neighborhood를 index로 갖고, Column을 avg of reviews, avg of overall satisfaction, avg of price, max of Reviews, min of reviews, max of price, min of price을 가지도록 정리한다.

데이터를 정리하기에 앞서, NaN으로 되어있는 값은 0으로 바꿔주었다. Neighborhood를 기준으로 정리하기 위해서 기존 표를 neighborhood 단위로 정리하였다. 그후 각 수치의 max, min 값을 Dataframe의 내장 함수로 찾았다. 이 외의 avg of reviews, avg of overall satisfaction, avg of price를 구하기 위해서 평균을 구하는 함수를 정의 하였다. 이 때, 평균이 0 때문에 너무 낮게 치우치는 것을 방지하기 위해서 평균을 구할 때 0은 제외하고 평균을 구하였다.

텍스트이(가) 표시된 사진

자동 생성된 설명

3. Price를 범위에 따라서 나눈 값을 index로 적용하고, column 값으로 avg accommodation, med accommodation, avg bedrooms, med bedrooms, avg reviews, med reviews, neighborhood list, length을 갖는 데이터를 작성한다.

우선 csv파일을 읽어서. Price column을 읽어와서 미리 정의한 함수로 price columns의 정보들을 범위로 분류하여 price range 정보를 얻었다. 그리고 이를 price index라는 이름의 column으로 추가했다. 그리고 해당 column으로 그룹화하였다. 그 후 각 그룹 별로 필요한 정보(accommodation, bedrooms, reviews, neighbors)를 뽑아냈다. 그리고 평균(0을 제외하고 계산), 중간 값을 계산하는 함수를 정의하여 값을 얻어냈다. 그리고 neighbor list는 set 자료형을 가지고 중복된 neighborhood들을 없앴다. 또한 Length는 그룹별 columns의 길이를 가지고 구했다.

텍스트이(가) 표시된 사진

자동 생성된 설명

4. 해당 문제는 2번, 3번에서 얻은 csv파일을 불러온다.

3번의 csv를 가지고는 가로 축을 price range로 한 accommodation average, bedrooms average 그래프를 그렸다. 4번의. Csv를 가지고는. 가로축이 neighborhood 세로 축이 average of reviews, average of satisfaction, average of price를 각각 세로 축 값으로 가지는 subplot을 그렸다.