경제자료분석 Term Project Final report			
팀명	경제자료분석 D조	팀원	응용통계학과 202111862 김서윤 응용통계학과 201911857 김한민 응용통계학과 202012089 이소민 스마트 ICT 융합공학과 201811567 주용한

1. 문제 설명

1.1 연구 배경

현대 사회에서 해외여행이나 교환학생 경험은 점차 증가하고 있으며, 많은 사람들이 에어비앤비와 같은 숙박 서비스를 활용하고 있다. 그러나 주변 지인들의 경험을 통해 발견된 것처럼, 모든 숙소에서 동일한 만족도를 느끼지 못하는 경우가 있다. 특히, 슈퍼호스트가 운영하는 숙소에서의 만족도가 높다는 인상을 받아, 슈퍼호스트 여부에 영향을 미치는 다양한 조건들이 무엇인지에 대한 궁금증이 생겼다. 이러한 경험을 바탕으로, 에어비앤비의 슈퍼호스트가 되기 위해서는 어떤 조건들이 중요한지 조사하는 연구가 필요하다고 생각이 들어 다음과 같은 주제를 정하게 되었다.

1.2 연구 목적

이 연구는 슈퍼호스트 여부와 접근성, 청결도 등 숙소 관련 변수들 간의 관계를 탐색하여 슈퍼호스트 여부에 영향을 미치는 변수들을 찾아 만족도가 낮은 숙소들의 개선 방안을 제시할 수 있으며, 에어비앤비 이용자들이 만족할 수 있는 더 나은 숙박 환경을 제공하기 위한 지침을 도출할 수 있을 것이다.

1.3 연구 질문

Airbnb, 어떻게 하면 고객을 만족시킬 수 있을까?

2. 데이터 출처 및 특징 설명

2.1 데이터 출처

Inside Airbnb는 Airbnb의 웹사이트에 등록되어 있는 숙소에 대한 정보를 추출하여 각 도시별로 종합한 자료를 대중에게 제공하고 있다. Airbnb의 공식 웹사이트에서 자료가 추출되었다는 점, 해당 자료에 가격 예측에 필요한 변수들인 가격, 숙소에 대한 위치 및 세부 정보, host에 대한 정보, 예약 관련 정보 등이 폭넓게 담겨 있다는 점, 실증적인 데이터를 제공한다는 점에서 본 프로젝트의 목적에 적절한 자료라고 판단된다. 74개의 columns중에 유의미하다고 판단되는 24개의 column들만 추출해 사용하였다.

2.2 데이터 특징

2.2.1 데이터 요약 및 변수 설명

```
> str(london)
'data.frame':   87946 obs. of  27 variables:
 $ description           : chr  "<b>The space</b><br />Hi everyone! I have 2 rooms with twin beds for people who wish to travel to London Englan"| __truncated__ "Gorgeous 2 bed grou
d floor apartment with period features and a real gas fire to keep you warm and cosy.  A gr"| __truncated__ "My bright double bedroom with a large window has a relaxed feeling! It comfor
ably fits one or two and is centr"| __truncated__ "Lots of windows and light.  St Luke's Gardens are at the end of the block, and the river not too far the other "| __truncated__ ...
 $ host_since            : chr  "2011-04-10" "2011-04-11" "2009-11-16" "2009-12-05" ...
 $ host_about            : chr  "I am employed at St Georges hospital in tooting London and my Partner Sid is a builder with a big construction "| __truncated__ "Been living in Lond
n for over 20 years and I love the city!" "I am a Multi-Media Visual Artist and Creative Practitioner in Education. I live in  London England with a Greek"| __truncated__ "English, grand
other,  I have travelled quite a lot.  I love being in different countries, as long as they are "| __truncated__ ...
 $ host_response_time    : chr  "N/A" "within a few hours" "within a few hours" "within a day" ...
 $ host_response_rate    : chr  "N/A" "100%" "100%" "100%" ...
 $ host_is_superhost     : chr  "f" "f" "f" "f" ...
 $ host_total_listings_count  : int  1 2 4 12 1 32 3 3 2 3 ...
 $ host_has_profile_pic  : chr  "t" "t" "t" "t" ...
 $ latitude              : num  51.4 51.5 51.6 51.5 51.5 ...
 $ longitude             : num  -0.1874 -0.2171 -0.1127 -0.1681 0.0144 ...
 $ room_type             : chr  "Private room" "Entire home/apt" "Private room" "Entire home/apt" ...
 $ accommodates          : int  2 5 1 2 2 6 4 3 4 2 ...
 $ bathrooms_text        : chr  "1.5 shared baths" "1 bath" "1 shared bath" "1 bath" ...
 $ bedrooms              : int  NA 2 NA 1 NA 3 2 1 1 NA ...
 $ beds                  : int  2 3 1 1 1 3 3 1 1 1 ...
 $ amenities             : chr  "[\"Heating\", \"TV with standard cable\", \"Wifi\", \"Smoke alarm\", \"Dryer\", \"Kitchen\", \"Washer\", \"Essentials\"]" "[\"Window guards\", \"Bat
tub\", \"Hot water kettle\", \"Laundromat nearby\", \"Indoor fireplace\", \"Microwave\"| __truncated__ "[\"Bathtub\", \"Hot water kettle\", \"Laundromat nearby\", \"Private patio or balco
y\", \"Paid street parking "| __truncated__ "[\"Shampoo\", \"Luggage dropoff allowed\", \"Dryer\", \"Microwave\", \"Coffee maker\", \"Hot water\", \"Iron\",\"| __truncated__ ...
 $ price                 : chr  "$42.00" "$175.00" "$79.00" "$150.00" ...
 $ minimum_nights        : int  2 5 1 7 4 3 5 2 2 10 ...
 $ maximum_nights        : int  730 240 29 30 365 365 365 1125 14 1120 ...
 $ number_of_reviews_ltm : int  9 2 11 5 25 4 3 0 0 0 ...
 $ review_scores_rating  : num  4.57 4.82 4.8 4.8 4.62 4.76 4.73 4.9 4.85 4.78 ...
 $ review_scores_accuracy : num  4.74 4.76 4.72 4.85 4.7 4.83 4.57 4.89 4.93 4.7 ...
 $ review_scores_cleanliness : num  4.86 4.62 4.72 4.88 4.59 4.71 4.7 4.91 4.71 4.94 ...
 $ review_scores_checkin  : num  4.71 4.85 4.74 4.88 4.63 4.71 5 4.9 4.93 4.91 ...
 $ review_scores_communication: num  4.67 4.88 4.82 4.83 4.81 4.71 4.96 4.93 5 4.89 ...
 $ review_scores_location : num  4.53 4.85 4.69 4.93 4.64 4.88 4.87 4.59 4.93 4.45 ...
 $ instant_bookable      : chr  "t" "f" "f" "f" ...
```

27개의 변수와 87846개의 관측치를 가지고 있으며 변수 설명은 아래와 같다.

변수명	변수 설명	변수명	변수 설명
description	호스트가 작성한 숙소에 대한 설명	beds	숙소 침대의 개수
host_since	숙소 처음 등록 시점	amenities	숙소에서 제공하는 편의시설, 물품 목록
host_about	호스트의 소개	price	현지 통화로 1일 가격
host_response_time	응답 시간(호스트가 게스트 요청에 응답하는데 걸리는 시간) 범주 : a few days or more | N/A | within a day within a few hours | within an hour	minimum_nights	숙소에 머무는 최소한의 박수 (캘린더 규칙은 다를 수 있음)
host_response_rate	응답	maximum_nights	숙소에 머무는 최대 박수 (캘린더 규칙은 다를 수 있음)
host_is_superhost	슈퍼호스트 여부	number_of_reviews_ltm	(지난 12개월 동안의) 숙소에 대한 리뷰의 수
host_total_listings_count	호스트가 보유한 숙소 수 (에어비앤비 계산 기준)	review_scores_rating	숙소 후기 점수
host_has_profile_pic	호스트 프로필에 사진 여부	review_scores_accuracy	
latitude	위도	review_scores_cleanliness	
longitude	경도	review_scores_checkin	

room_type	숙소 유형 범주 : Entire home/apt | Private room | Shared room | Hotel	review_scores_communication	
accommodates	숙소의 최대 수용 인원	review_scores_location	
bathrooms_text	화장실 종류	instant_bookable	게스트가 호스트의 승인을 요청하지 않고 숙소를 자동으로 예약할 수 있는지 여부, 상업적 숙소의 지표
bedrooms	숙소 침실의 개수		

2.3 데이터 전처리

```
> sum(is.na(london))
[1] 171591
> colSums(is.na(london)) # 컬럼별 결측치 개수
             description              host_since              host_about      host_response_time      host_response_rate        host_is_superhost
                       0                       0                       1                       0                       0                       0
  host_total_listings_count      host_has_profile_pic                latitude               longitude               room_type            accommodates
                       5                       0                       0                       0                       0                       0
           bathrooms_text                bedrooms                    beds               amenities                   price           minimum_nights
                       0                   32774                    1134                       0                       0                       0
           maximum_nights      number_of_reviews_ltm      review_scores_rating    review_scores_accuracy  review_scores_cleanliness  review_scores_checkin
                       0                       0                   22158                   23093                   23081                   23125
review_scores_communication   review_scores_location         instant_bookable
                     23095                   23125                       0
```

이때 전체 데이터 value에 171591개의 결측치가 존재한다.

컬럼별 결측치를 확인하였을 때 bedrooms, review_scores_accuracy review_scores_cleanlines, review_scores_checkin, review_scores_communication, review_scores_location 변수에 결측치가 대부분 분포했다. 따라서 결측치가 너무 많은 bedrooms 변수를 삭제했으며 review_scores_accuracy, review_scores_cleanlines, review_scores_checkin, review_scores_communication, review_scores_location 변수의 결측치들은 정보가 없다고 가정하고 결측치가 있는 행을 모두 삭제하였다.

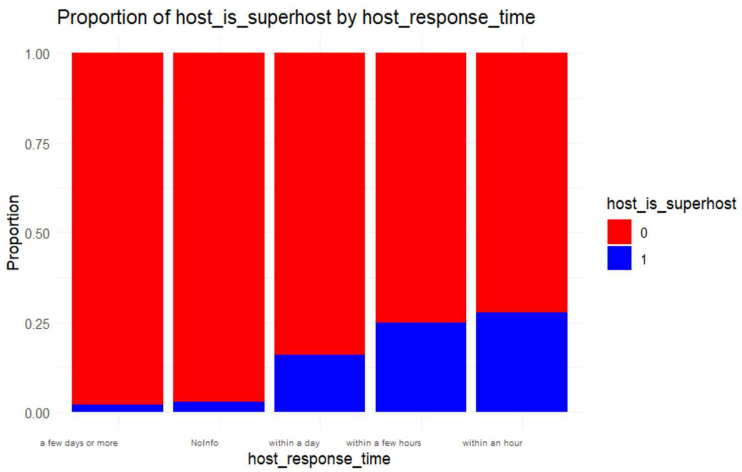
아래 표는 결측치 처리 외에 전처리를 한 방법들이다.

변수명	전처리 방법	변수명	전처리 방법
description	- 불용어를 처리 후 유의미한 단어의 개수를 셈	beds	
host_since	- 데이터 최종 업데이트 일자를 기준으로 숙소를 처음 등록한 시점	amenities	

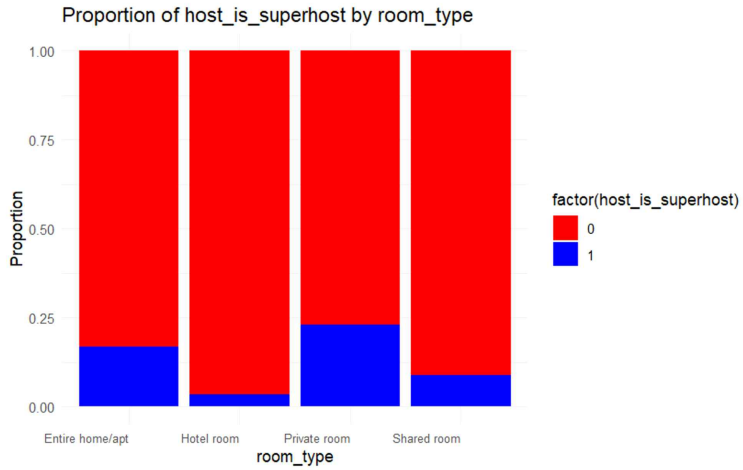
	부터 현재까지의 일(days) 수를 계산		
host_about	- 불용어를 처리 후 유의미한 단어의 개수를 셈	price	- $ 표시 빼고 수치형으로 바꿈 - 이후 결측치 행은 정보가 없다 생각해 삭제
host_response_time	- N/A를 Noinfo로 바꿈 - 범주형으로 바꿈	minimum_nights	
host_response_rate	- N/A를 0%로 바꿈 - 범주형으로 바꿈	maximum_nights	
host_is_superhost	- 범주형으로 바꿈	number_of_reviews_ltm	
host_total_listings_count		review_scores_rating	
host_has_profile_pic	- 범주형으로 바꿈	review_scores_accuracy	
latitude		review_scores_cleanliness	
longitude		review_scores_checkin	
room_type	- 범주형으로 바꿈	review_scores_communication	
accommodates	숙소의 최대 수용 인원	review_scores_location	
bathrooms_text	- barthroom type에 상관없이 bathroom개수만 셈 - 수치형으로 바꿈	instant_bookable	- 범주형으로 바꿈
bedrooms	- 앞에서 결측치 많아서 컬럼 자체 삭제		

2.4 EDA

latitude와 longitude를 런던 지도에 슈퍼호스트 여부에 따라 표시하였을 때 큰 차이가 없어 두 변수를 먼저 삭제하였다. <부록1, 그림1>



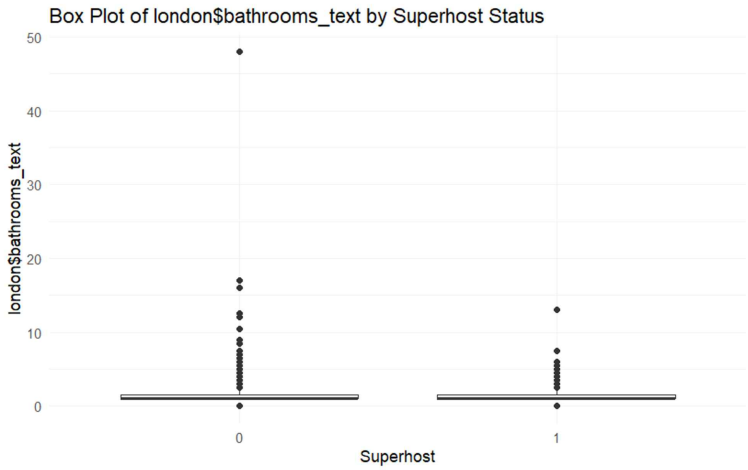
연속형 변수는 상자 그림, 범주형 변수는 범주별 수퍼호스트의 비율을 그래프로 나타냈다. hostresponse_time응답시간에 따라 슈퍼호스트 비율이 높아지는 모습을 보였고,



entire와 private에서 다른 숙소 유형보다 슈퍼호스트 비율이 높게 나타난 것으로 보아 room_type, response_time이 슈퍼호스트 여부에 영향을 미칠 것이라고 생각한다.

연속형 변수들의 상자 그림을 보았을 때 이상치 처리가 필요한 변수들은 description, host_since, host_total_listings_count, amenities, price, minimum_nights, maximum_nights로 나타났다. <부록1, 그림2>

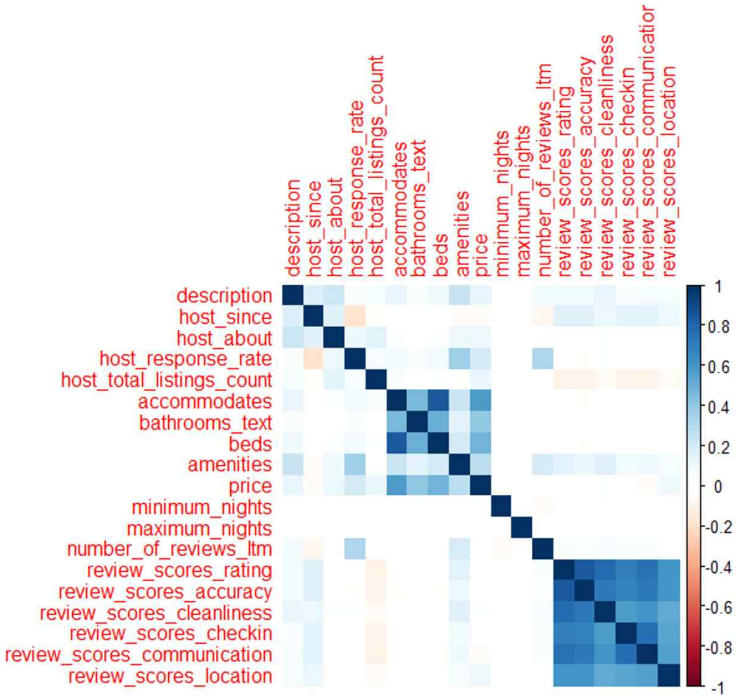
또한, 슈퍼호스트 여부에 영향을 미치는 변수만을 파악한 것이 아니라 이상치 등 일반적이지 않은 관측치들도 발견하였다.



숙소 총 평점을 나타내는 review_scores_rating 변수의 상자 그림에서는 슈퍼호스트(1)인데 4.8점 미만인 이상한 값들이 관측되었다.

더불어 세부적인 숙소 후기 점수를 나타내는 review_scores_accuracy, review_scores_cleanlines, review_scores_checkin, review_scores_communication, review_scores_location 변수의 상자 그림에서 0점인 값들이 발견되었다. <부록1, 그림3>

화장실 개수를 나타내는 bathrooms_text의 상자 그림에서는 화장실 개수가 0인 경우와 화징실이 48개인 경우도 있다.



<상관관계>

위 상관관계 그림을 보면 beds와 accommodate가 높은 상관관계를 보이고 있으며 bathrooms_text와 accomodate도 높은 상관관계를 보이고 있다.

2.5 추가 전처리

앞에서 EDA 시각화를 통해 발견된 일반적이지 않은 관측치들을 추가로 처리하였다.

이상치가 발견된 description, host_since, host_total_listings_count, amenities, price, minimum_nights, maximum_nights 변수들에서 이상치가 포함된 행들을 삭제했다. price는 해석 용이성을 위해 로그 변환까지 해주었다. 슈퍼호스트(host_is_superhost=1)인데 4.8점 미만인 이상한 값들을 삭제하였으며 bathrooms_text 변수의 최고 값 행을 삭제하였으며 화장실 개수가 0인 숙소는 일반적이지 않다고 판단하여 0개인 행들도 삭제하였다. review_scores_rating, review_scores_accuracy, review_scores_cleanlines, review_scores_checkin, review_scores_communication, review_scores_location 변수의 값이 0인 행들도 모두 삭제하였다.

상관관계 그림에서 accommodate 변수와 높은 상관관계를 보였던 beds 변수를 삭제하였다. 그리고 bathrooms_text 변수는 accommodates를 bathrooms_tex로 나눠 화장실 1개당 수용 인원을 나타

내는 변수로 변환해 다중공선성을 해결하고자 하였다.

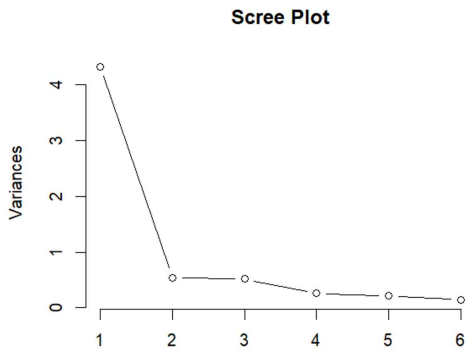
2.4 최종 데이터 요약

```
> str(london)
'data.frame':   45164 obs. of  23 variables:
 $ description                : int  86 108 104 113 106 118 105 98 114 14 ...
 $ host_since                 : num  4532 4531 5042 5091 4530 ...
 $ host_about                 : int  23 7 25 57 26 63 12 29 12 13 ...
 $ host_response_time         : Factor w/ 5 levels "a few days or more",..: 2 4 4 2 2 5 5
1 2 2 ...
 $ host_response_rate         : num  0 100 100 0 0 100 100 10 0 0 ...
 $ host_is_superhost          : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 1 1 1 1 ...
 $ host_total_listings_count  : int  1 2 4 3 2 11 2 1 2 2 ...
 $ host_has_profile_pic       : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
 $ room_type                  : Factor w/ 4 levels "Entire home/apt",..: 3 1 3 1 1 1 3 3 3
3 ...
 $ accommodates               : int  2 5 1 3 4 4 2 2 2 2 ...
 $ bathrooms_text             : num  1.33 5 1 2 4 ...
 $ amenities                  : int  8 48 55 60 29 11 49 15 12 21 ...
 $ price                      : num  3.74 5.16 4.37 5.52 4.32 ...
 $ minimum_nights             : int  2 5 1 2 2 3 3 2 2 1 ...
 $ maximum_nights             : int  730 240 29 1125 14 90 180 10 21 14 ...
 $ number_of_reviews_ltm      : int  9 2 11 0 0 0 4 0 0 0 ...
 $ review_scores_rating       : num  4.57 4.82 4.8 4.9 4.85 4.9 4.82 4.86 4.4 4.75 ...
 $ review_scores_accuracy     : num  4.74 4.76 4.72 4.89 4.93 4.87 4.71 4.81 4.47 5 ...
 $ review_scores_cleanliness  : num  4.86 4.62 4.72 4.91 4.71 4.79 4.74 4.84 4.58 5 ...
 $ review_scores_checkin      : num  4.71 4.85 4.74 4.9 4.93 4.9 4.93 4.88 4.58 4.75 ...
 $ review_scores_communication: num  4.67 4.88 4.82 4.93 5 4.94 4.9 4.91 4.53 5 ...
 $ review_scores_location     : num  4.53 4.85 4.69 4.59 4.93 4.56 4.46 4.47 4.68 4.5 ...
 $ instant_bookable           : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 1 1 ...
 - attr(*, "na.action")= 'omit' Named int [1:618] 98 1089 1758 2218 2438 2551 2678 2679 41
06 4116 ...
  ..- attr(*, "names")= chr [1:618] "108" "1278" "2034" "2556" ...
```

최종 데이터의 상관관계 그림을 확인했을 때 처음 데이터보다 설명 변수들끼리 높은 상관관계를 나타내는 다중공선성 문제가 많이 개선된 것처럼 보이지만 review_scores관련 변수들끼리 높은 상관관계가 나타내는 것을 보인다. 따라서 뒤 모델 생성 과정에서 기존 변수들의 선형조합으로 새로운 변수를 만드는 PCA(주성분 분석) 방법을 통해 해결하고자 한다.

3. 분석 모형 및 의미 설명

앞에서 언급한 것처럼




```
> summary(pca_result) # PCA 결과 요약 표시
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6
Standard deviation     2.080 0.73495 0.71845 0.50813 0.46771 0.37540
Proportion of Variance 0.721 0.09002 0.08603 0.04303 0.03646 0.02349
Cumulative Proportion  0.721 0.81099 0.89702 0.94005 0.97651 1.00000
```

review_scores_accuracy, review_scores_cleanlines, review_scores_checkin, review_scores_rating, review_scores_communication, review_scores_location의 여러 선형 조합에서 PCA 결과에 따라 누적 분산 비율이 72퍼센트인 제 1 주성분을 사용한다.

```
> print(coefficients)
                              Dim.1      Dim.2       Dim.3       Dim.4        Dim.5
review_scores_rating        0.9248577 -0.1699545  0.09213647 -0.09455008 -0.001017108
review_scores_accuracy      0.8988699 -0.1334117  0.07088195 -0.31803243 -0.203156751
review_scores_cleanliness   0.8242932 -0.3583345  0.30705769  0.27916560  0.090412935
review_scores_checkin       0.8370093  0.1070105 -0.41723993  0.24270868 -0.234327074
review_scores_communication 0.8651945  0.0583859 -0.34647331 -0.09981675  0.337754461
review_scores_location      0.7306668  0.5917775  0.33800384  0.03614823  0.017704441
```

이에 따라 제 1 주성분의 계수들을 곱해 각 변수에 곱해 새로운 변수 review_scores_rating를 만들었다.

최종적으로 비교할 모델0, 1, 2, 3을 생성했다.
모델 0은 이상치만 처리한 상태이다. 즉, price에서 로그 변환을 하지 않고, bathrooms_text를 화장실 당 수용인원으로 변수변환 하지 않았다. 모델 1에는 price 로그변환 bathrooms_text를 화장실 당 수용 인원으로 변수변환 하였으며 모델 2에는 모델 1 상태에서 기존 reviews_scores 변수들을 삭제하고 PCA로 만든 변수를 적용하였다. 이때 모델 3은 모델 2에 후진선택법을 사용해 변수를 선택하고자 하였는데 선택된 변수가 모델 1과 동일했다. 그래서 모델 3은 모델 2를 deviance 진단을 하였을 때 (<부록 2>, 그림1)유일하게 유의하지 않은 변수인 maximum_nights 변수를 제거해 보았다.

<모델 비교>

변수	모델 0	모델 1	모델 2	모델 3
description	4.209e-03***	4.116e-03***	3.309e-03***	3.348e-03***
host_since	1.319e-04***	1.343e-04***	1.240e-04***	1.254e-04***
host_about	5.187e-03***	5.247e-03***	4.868e-03***	4.890e-03***
host_response_timeNoInfo	7.006e-01*	7.376e-01**	6.843e-01*	6.883e-01*
host_response_time within a day	-7.047e-01.	-7.916e-01*	-7.471e-01*	-7.408e-01*
host_response_time within a few hours	-4.167e-01	-5.040e-01	-4.560e-01	-4.480e-01
host_response_time within an hour	-1.126e-02	-2.118e-01	-1.686e-01	-1.596e-01
host_response_rate	3.139e-02***	3.233e-02***	3.169e-02***	3.151e-02***
host_total_listings_count	8.158e-03.	7.690e-03.	-6.147e-03	-5.837e-03
host_has_profile_pic1	2.876e-01*	3.131e-01**	2.957e-01*	2.954e-01*
room_typeHotel room	-1.221e+00	-1.605e+00.	-1.689e+00*	-1.650e+00*

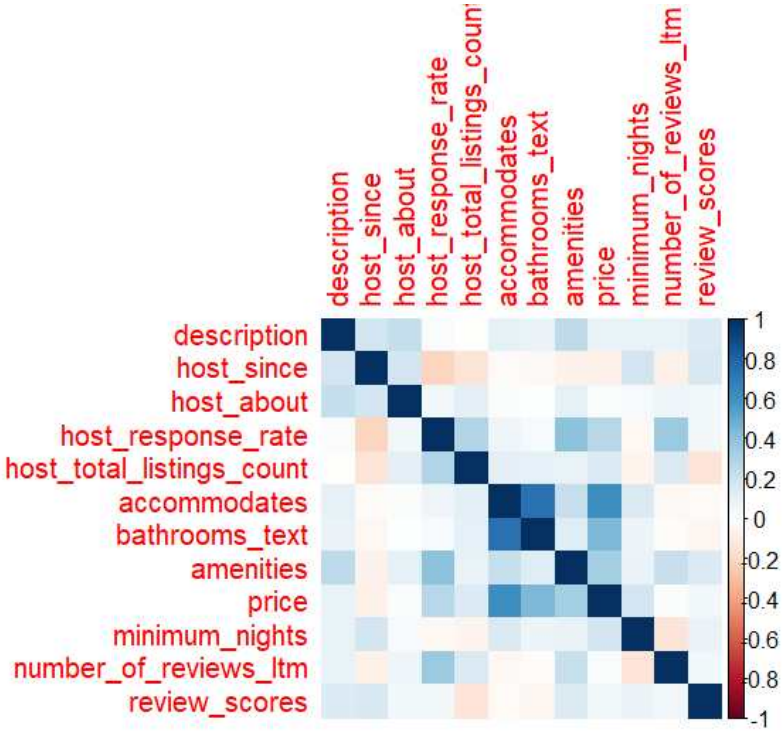
room_typePrivate room	5.012e-01***	5.548e-01***	5.043e-01***	5.023e-01***
room_typeShared room	-6.270e-01.	-5.341e-01	-6.612e-01.	-6.519e-01.
accommodates	-1.150e-01***	-1.513e-01***	-1.285e-01***	-1.283e-01***
bathrooms_text (화장실 당 수용인원)		6.288e-02**	5.092e-02**	5.205e-02**
bathrooms_text (화장실 개수)	-1.836e-01***			
amenities	2.013e-02***	1.976e-02***	2.247e-02***	2.253e-02***
log(price)		3.894e-01***	2.704e-01***	2.708e-01***
price	3.309e-03***			
minimum_nights	9.079e-02***	8.628e-02***	9.311e-02***	9.238e-02***
maximum_nights	1.523e-04***	1.754e-04***	1.191e-04***	
number_of_reviews_ltm	5.532e-02***	5.531e-02***	4.616e-02***	4.618e-02***
review_scores_rating	7.031e+00***	6.932e+00***		
review_scores_accuracy	5.311e-01***	5.201e-01***		
review_scores_cleanliness	6.185e-01***	6.302e-01***		
review_scores_checkin	2.318e-01.	2.627e-01*		
review_scores_communication	1.086e+00***	1.041e+00***		
review_scores_location	-4.506e-01***	-4.178e-01***		
instant_bookable1	-1.582e-01***	-1.490e-01***	-1.604e-01***	-1.599e-01***
review_scores			1.661e+00***	1.651e+00***
constant	-5.084e+01***	-5.215e+01***	-4.869e+01***	-4.840e+01***
AIC	23990	25910	26972	26984

자세한 모델 결과는 <부록2, 그림2>, <부록2, 그림3>,<부록2, 그림4>, <부록2, 그림5> 참고

최종적으로 선택한 모형은 모델 3이다.

$$pr(y(\text{슈퍼호스트여부})=1|x) = \frac{1}{1+e^{-(\beta_0+\beta_1 D_1(\text{응답 없음})+\beta_2 D_2(\text{하루이내응답})+\beta_3 D_3(\text{몇시간내로응답})+\beta_4 D_4(\text{한시간내로응답})+\beta_5 D_5(\text{프로필 사진 유무})+\beta_6 D_6(\text{호텔})+\beta_7 D_7(\text{개인방})+\beta_8 D_8(\text{공유방})+\beta_9 D_9(\text{즉시예약여부})+\beta_{10} x_1(\text{숙소 소개})+\beta_{11} x_2(\text{숙소등록기간})+\beta_{12} x_3(\text{호스트소개})+\beta_{13} x_4(\text{응답률})+\beta_{14} x_5(\text{호스트소유숙소개수})+\beta_{15} x_6(\text{최대수용인원})+\beta_{16} x_7(\text{화장실당수용인원})+\beta_{17} x_8(\text{호스트 소개})+\beta_{18} x_9(\text{편의시설및물품수})+\beta_{19} x_{10}(\text{가격})+\beta_{20} x_{11}(\text{최소숙박일수})+\beta_{21} x_{12}(1\text{년이내등록후기수})+\beta_{22} x_{13}(\text{리뷰 평점 } PCA))}}$$

<모델 3>



<모델 3 수치형 변수들의 상관관계>

최종적으로 모델 3을 선택하였다. 3개의 모델을 비교해보면 모델 3의 AIC가 제일 높지만 변수 제거로 인한 불가피한 AIC 증가라고 생각하였다. 따라서 변수 제거로 인해 모델을 더 간단하게 만들면서도 해당 변수들이 모델의 중요한 정보를 유지한다고 판단해 모델 3을 선택하였다.

최종적인 모델 3을 해석해보자.
먼저 유의미한 양의 계수를 가진 변수들을 해석하면,
description인 호스트가 설명하는 숙소 소개가 길수록
host_since인 숙소를 등록한 시점이 과거일수록
host_about인 호스트의 소개가 길수록
host_response_timeNoInfo인 응답 시간에 대한 정보가 없는 경우가
host_response_rate인 응답률이 높을수록
host_has_profile_pic1인 호스트의 프로필이 있는 경우가 없는 경우보다
room_typePrivate room인 숙소 유형이 개인 유형인 경우가 아닌 경우보다
bathrooms_text인 화장실당 수용 인원이 많을수록
amenities인 숙소에서 제공하는 물품이나 편의시설이 많을수록
price인 현지 통화 가격이 높을수록
minimum_nights인 최소 숙박 일수가 높을수록
number_of_reviews_ltm인 1년 이내의 숙소 리뷰의 수가 많을수록
reviews_scores가 높을수록 슈퍼호스트일 확률이 높아진다고 해석할 수 있다.

다음으로 유의한 음의 계수를 가진 변수들을 해석하면,

instant_bookable1인 즉시 예약할 수 있는 경우가 없는 경우보다
host_response_timewithin a day인 하루 안으로 답장하는 경우가 아닌 경우보다
room_typeHotel room인 숙소유형이 호텔인 경우가 아닌 경우보다
room_typeShared room인 숙소 유형이 여러 사람이 함께 쓰는 경우가 아닌 경우보다
accommodates인 최대 수용 인원 수가 많을수록 슈퍼호스트일 확률이 낮을 것이라고 해석할 수 있다.

4. 결론

초기 가설은 숙소 예약 가격이 낮을수록 슈퍼호스트일 가능성이 높을 것이며, 또한 신규 숙소일수록 슈퍼호스트일 가능성이 높을 것이었다. 그러나 실제 분석 결과에서는 숙소 예약 가격이 낮다고 해서 반드시 슈퍼호스트가 될 가능성이 높지 않았다. 오히려, 가격이 쌀수록 숙소의 퀄리티가 떨어질 수 있으며, 높은 가격의 숙소가 더 높은 퀄리티를 제공할 가능성이 있다.
또한, 신규 숙소라고 해서도 슈퍼호스트가 될 가능성이 높지 않았다. 오히려 호스트의 등록 기간이 오래되었을수록 슈퍼호스트일 확률이 높아진 것으로 나타났다. 이러한 결과는 신규 숙소라고해서 신축 건물일 보장이 없으며, 반면에 등록 기간이 오래될수록 호스트가 검증된 경험이 풍부해져서 슈퍼호스트가 될 가능성이 높아진다는 점에서 초기 가설과 다른 결과를 보여주고 있다.

슈퍼호스트 자격을 얻기 위해서는 적극적으로 숙소를 표현하는 것이 중요하다. 효과적인 방법으로는 숙소 및 호스트 소개 설명을 자세하게 작성하고, 호스트 프로필에 명확한 사진을 추가하는 것이 도움이 될 것이다. 또한, 투숙객에게 편안하고 개인적인 경험을 제공하기 위해 숙소에서 개인 공간을 충분히 제공하는 것이 핵심이다. 이를 위해 숙소 유형을 개인실로 설정하고, 화장실 수는 많고 수용 가능한 인원을 적게 설정하는 것이 좋다. 이러한 노력들은 슈퍼호스트 자격을 향한 긍정적인 경험을 조성하는데 도움이 될 것이다.

5. 분석에 대한 검토 및 한계점, 우수성 설명

5.1 우수성

분석에 사용된 데이터는 코로나 이후의 최신 자료이며, 에어비앤비와 같은 관심 플랫폼에 대한 전략을 이해하는 데 중요한 정보를 제공한다.

5.2 한계점

4.2.1 데이터 측면

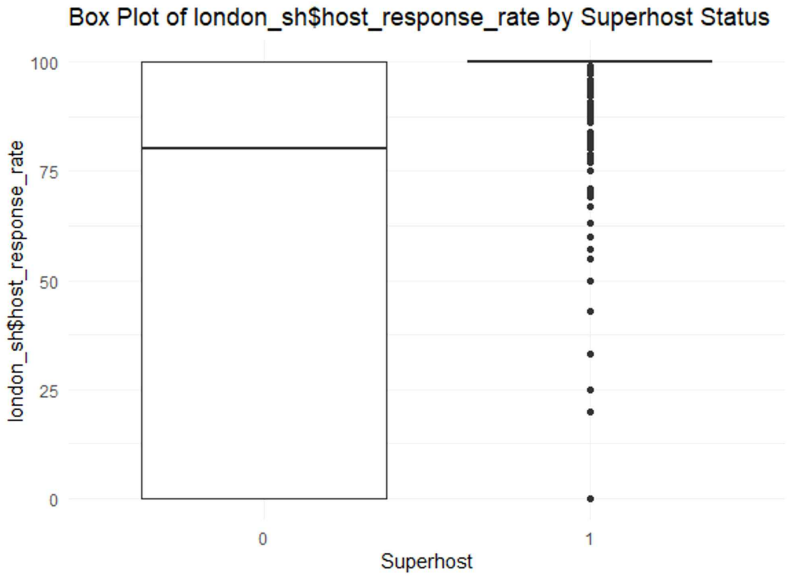
전체 데이터를 사용한 것은 특정 기간에 대한 표본이 아니라 모집단을 대표하는 것이 아니기 때문에, 결과의 일반화에 한계가 있을 수 있다.

5.3 모델 측면

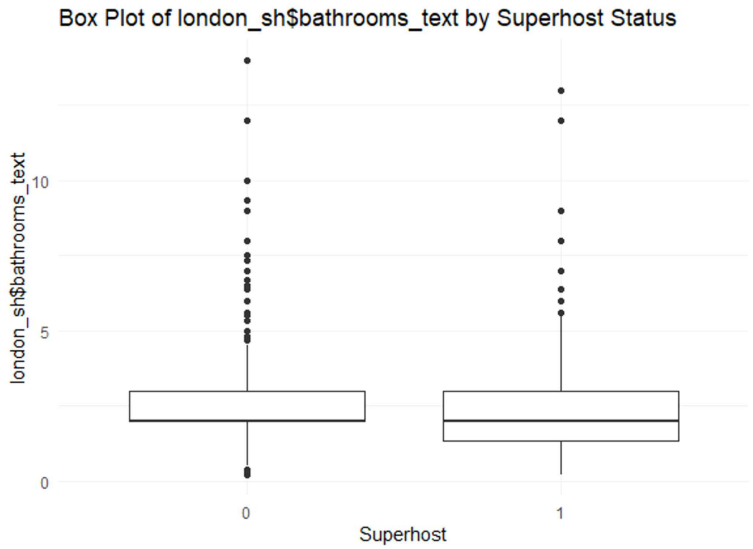
모델링 시 생길 수 있는 중요한 제약으로는 omitted variable 문제와 슈퍼호스트의 명확한 기준이 review로 제외되었을 때, 다른 변수들의 영향을 명확히 이해하기 어렵다는 것이 있다.

5.4 추가 분석

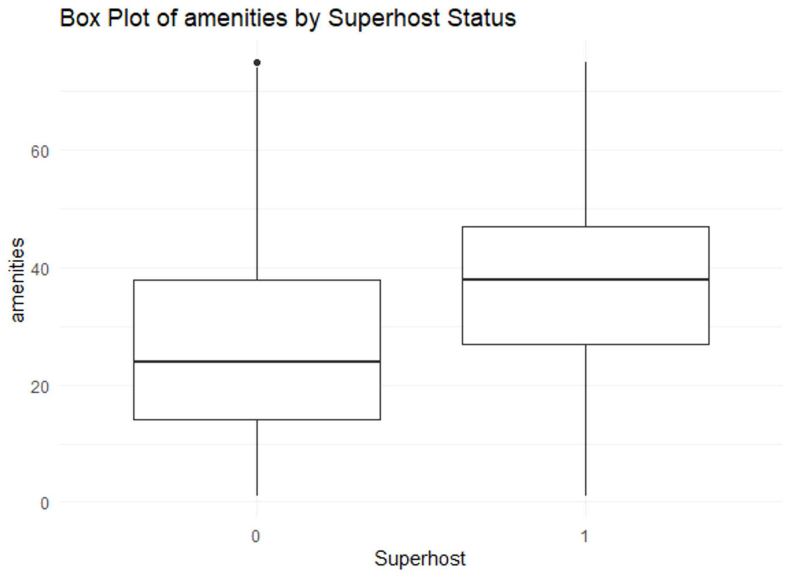
“숙소 후기 평점이 4.8 이상인데 왜 슈퍼호스트가 아닐까?”라는 추가 궁금증이 들어 숙소 후기 평점이 4.8 이상인 데이터만 추출해 추가 분석을 진행해보았다.



호스트의 응답률을 나타내는 host_response_rate가 슈퍼호스트가 아닌 경우 평균이 훨씬 낮은 것을 보아 응답률이 낮아서 슈퍼호스트가 안 된 것으로 추측할 수 있다.



더불어 숙소가 제공하는 편의시설이나 물품 개수를 나타내는 ammenities가 슈퍼호스트가 아닌 경우 보다 작은 값으로 분포되어 있는 것으로 보아 물품 개수가 적어서 슈퍼호스트가 안 된 것으로 추측할 수 있다.



화장실당 수용 인원 수를 나타내는 변수 bathrooms_text에서도 슈퍼호스트인 경우 좀 더 화장실당 인원이 적은 것을 볼 수 있다.

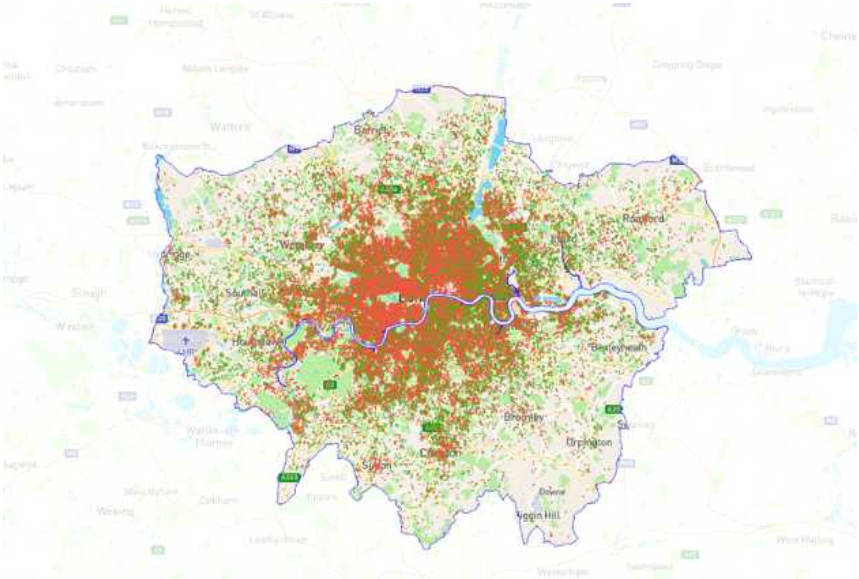
5. 참고문헌

AIRBNB, (2023), 도움말 센터, AIRBNB, https://www.airbnb.co.kr/help/article/829

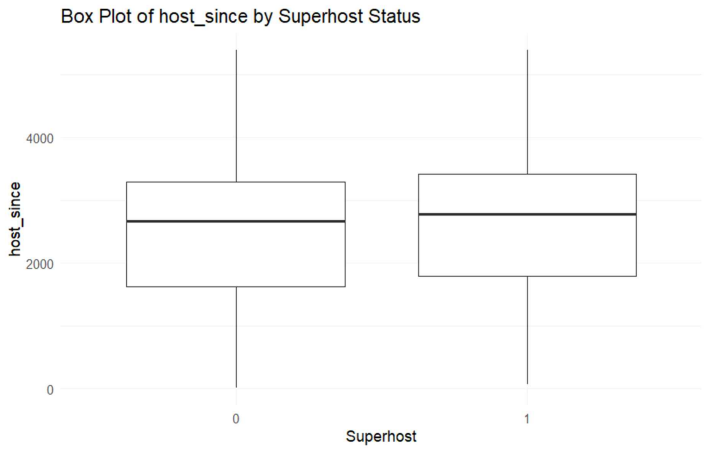
AIRBNB, (2023.09.06.), Get the Data, insideairbnb, http://insideairbnb.com/get-the-data/

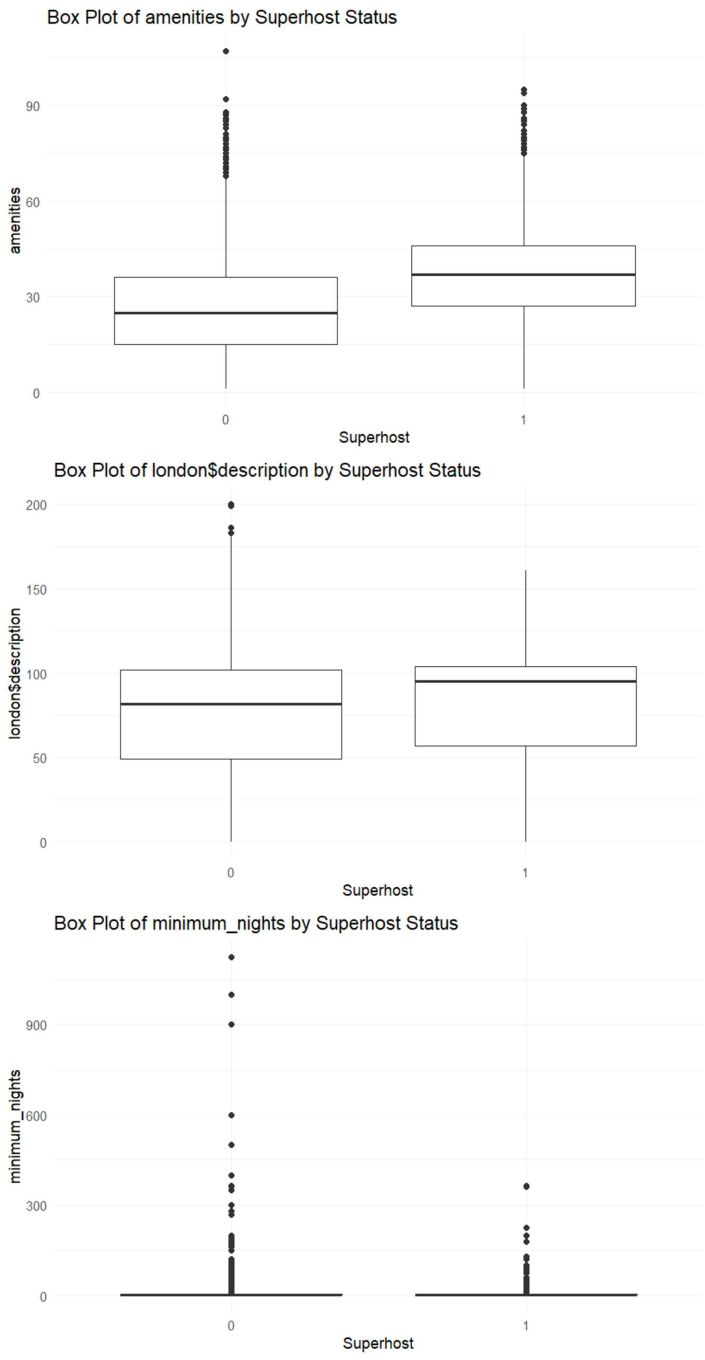
6. 부록

부록 1

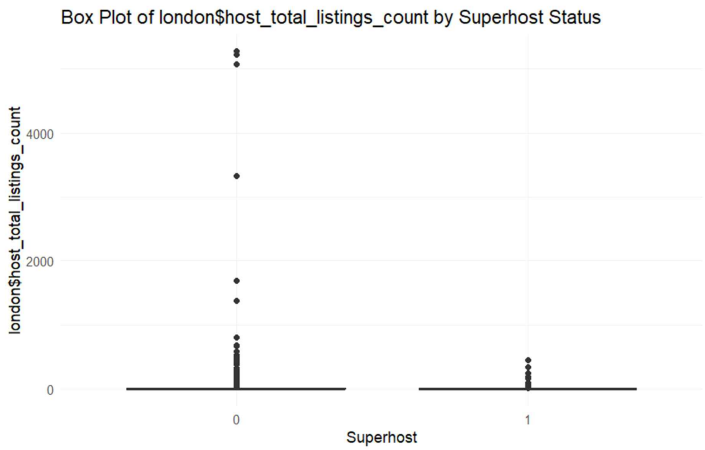


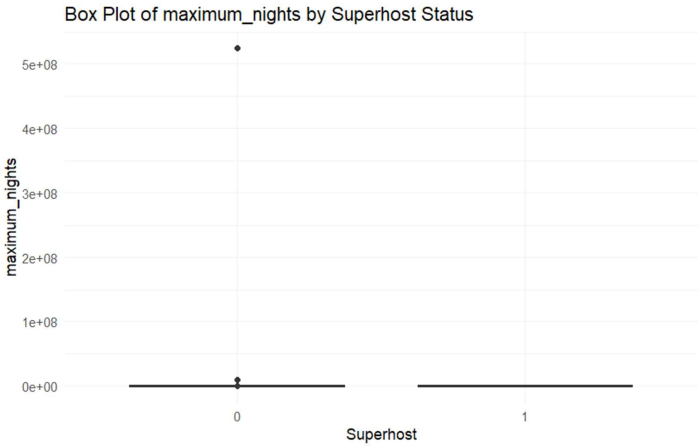
<그림 1>

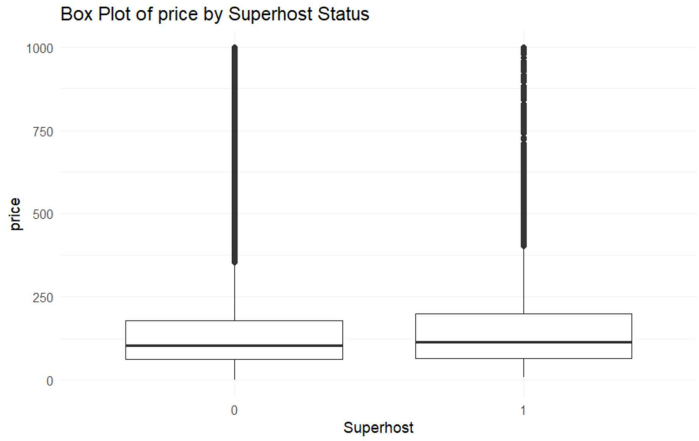

Box Plot of host_since by Superhost Status
host_since
4000
2000
0
0
1
Superhost

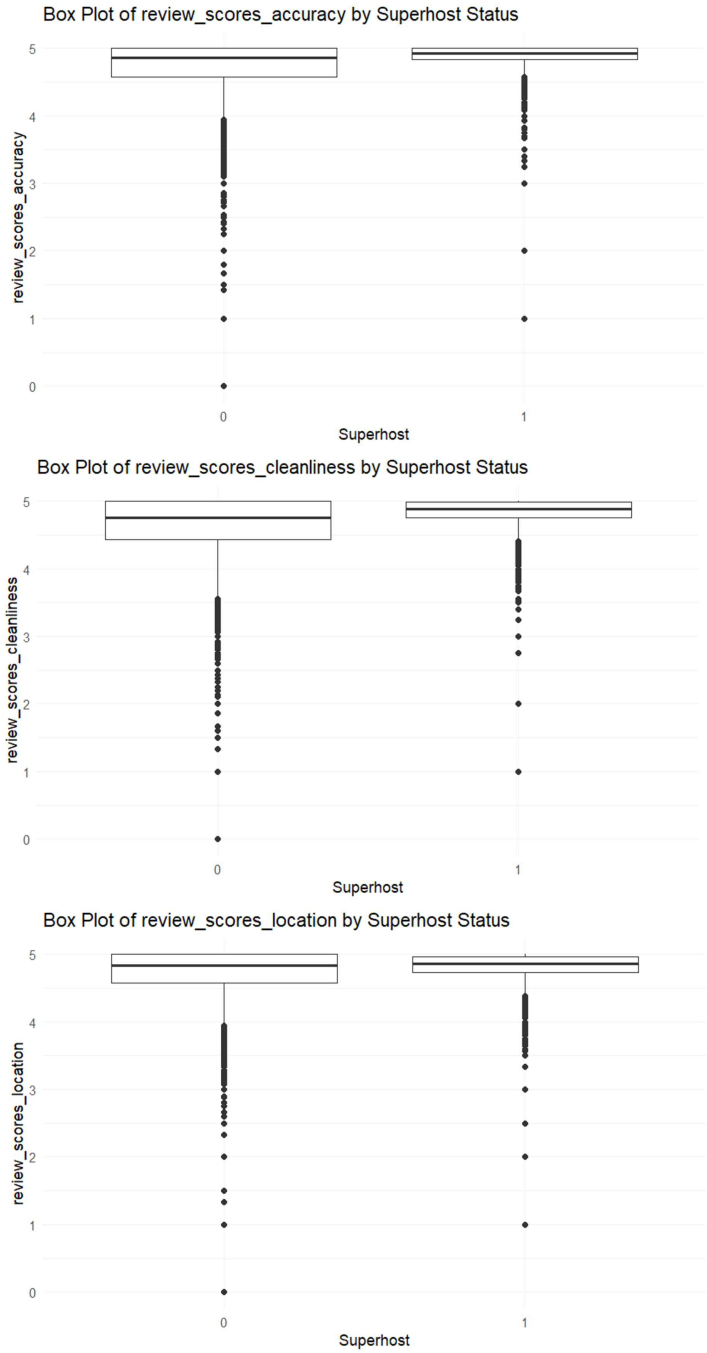

Box Plot of amenities by Superhost Status
amenities
90
60
30
0
0
1
Superhost
Box Plot of london$description by Superhost Status
london$description
200
150
100
50
0
0
1
Superhost
Box Plot of minimum_nights by Superhost Status
minimum_nights
900
600
300
0
0
1
Superhost

<그림2>

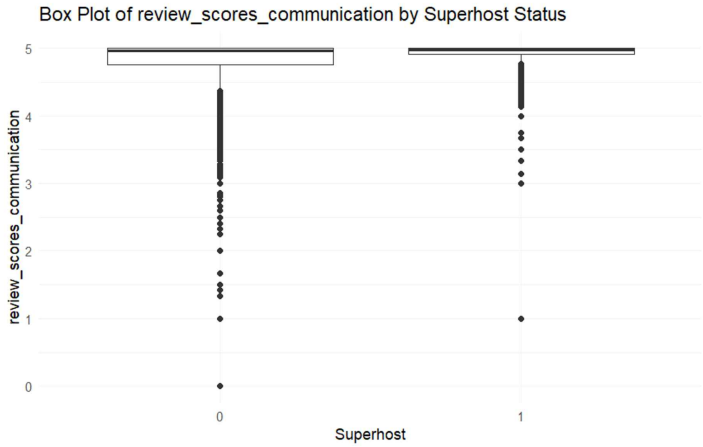

Box Plot of london$host_total_listings_count by Superhost Status
london$host_total_listings_count
4000
2000
0
0
1
Superhost


Box Plot of maximum_nights by Superhost Status
maximum_nights
5e+08
4e+08
3e+08
2e+08
1e+08
0e+00
0
1
Superhost


Box Plot of price by Superhost Status
price
1000
750
500
250
0
0
1
Superhost


Box Plot of review_scores_accuracy by Superhost Status
review_scores_accuracy
Superhost
Box Plot of review_scores_cleanliness by Superhost Status
review_scores_cleanliness
Superhost
Box Plot of review_scores_location by Superhost Status
review_scores_location
Superhost

<그림3>


Box Plot of review_scores_communication by Superhost Status
review_scores_communication
5
4
3
2
1
0
0
1
Superhost


Box Plot of review_scores_checkin by Superhost Status
review_scores_checkin
5
4
3
2
1
0
0
1
Superhost

부록 2

```
> deviance_value <- anova(mod2, test = "Chisq") # deviance 확인
> deviance_value
Analysis of Deviance Table

Model: binomial, link: logit

Response: host_is_superhost

Terms added sequentially (first to last)


                          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                      45163      41164
description                1    736.4     45162      40427 < 2.2e-16 ***
host_since                 1     21.2     45161      40406 4.093e-06 ***
host_about                 1    200.7     45160      40205 < 2.2e-16 ***
host_response_time         4   5596.1     45156      34609 < 2.2e-16 ***
host_response_rate         1    380.2     45155      34229 < 2.2e-16 ***
host_total_listings_count  1    265.7     45154      33963 < 2.2e-16 ***
host_has_profile_pic       1     23.3     45153      33940 1.353e-06 ***
room_type                  3    409.6     45150      33531 < 2.2e-16 ***
accommodates               1     25.0     45149      33505 5.606e-07 ***
bathrooms_text             1      5.0     45148      33500   0.02474 *
amenities                  1    918.9     45147      32582 < 2.2e-16 ***
price                      1    228.5     45146      32353 < 2.2e-16 ***
minimum_nights             1     21.1     45145      32332 4.423e-06 ***
maximum_nights             1      0.0     45144      32332   0.92866
number_of_reviews_ltm      1   1417.1     45143      30915 < 2.2e-16 ***
instant_bookable           1     85.5     45142      30829 < 2.2e-16 ***
review_scores              1   3903.1     45141      26926 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

<그림1>

```
> summary(mod0) # AIC: 24037

Call:
glm(formula = host_is_superhost ~ ., family = binomial(link = "logit"),
    data = london)

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -5.084e+01  1.165e+00 -43.625  < 2e-16 ***
description                      4.209e-03  5.714e-04   7.366 1.76e-13 ***
host_since                       1.319e-04  1.495e-05   8.821  < 2e-16 ***
host_about                       5.187e-03  6.001e-04   8.644  < 2e-16 ***
host_response_timeNoInfo         7.006e-01  2.881e-01   2.431 0.015038 *
host_response_timewithin a day  -7.047e-01  3.531e-01  -1.996 0.045958 *
host_response_timewithin a few hours -4.167e-01  3.605e-01  -1.156 0.247635
host_response_timewithin an hour -1.126e-01  3.619e-01  -0.311 0.755735
host_response_rate               3.139e-02  3.204e-03   9.797  < 2e-16 ***
host_total_listings_count        8.158e-03  4.317e-03   1.890 0.058793 .
host_has_profile_pic1            2.876e-01  1.198e-01   2.400 0.016409 *
room_typeHotel room             -1.221e+00  8.715e-01  -1.401 0.161299
room_typePrivate room            5.012e-01  4.688e-02  10.692  < 2e-16 ***
room_typeShared room            -6.270e-01  3.689e-01  -1.699 0.089254 .
accommodates                    -1.150e-01  1.470e-02  -7.823 5.16e-15 ***
bathrooms_text                  -1.836e-01  4.007e-02  -4.581 4.64e-06 ***
amenities                        2.013e-02  1.213e-03  16.598  < 2e-16 ***
price                            3.309e-03  3.581e-04   9.241  < 2e-16 ***
minimum_nights                   9.079e-02  1.368e-02   6.639 3.15e-11 ***
maximum_nights                   1.523e-04  3.460e-05   4.403 1.07e-05 ***
number_of_reviews_ltm            5.532e-02  1.322e-03  41.847  < 2e-16 ***
review_scores_rating             7.031e+00  2.221e-01  31.652  < 2e-16 ***
review_scores_accuracy           5.311e-01  1.582e-01   3.358 0.000785 ***
review_scores_cleanliness        6.185e-01  1.001e-01   6.182 6.33e-10 ***
review_scores_checkin            2.318e-01  1.300e-01   1.783 0.074555 .
review_scores_communication      1.086e+00  2.177e-01   4.988 6.10e-07 ***
review_scores_location          -4.506e-01  7.779e-02  -5.792 6.94e-09 ***
instant_bookable1               -1.582e-01  4.280e-02  -3.696 0.000219 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 38590  on 42879  degrees of freedom
Residual deviance: 23981  on 42852  degrees of freedom
AIC: 24037

Number of Fisher Scoring iterations: 8
```

<그림2>

```
> summary(mod1) # AIC: 25910

Call:
glm(formula = host_is_superhost ~ ., family = binomial(link = "logit"),
    data = london)

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -5.215e+01  1.124e+00 -46.380  < 2e-16 ***
description                      4.116e-03  5.502e-04   7.480 7.44e-14 ***
host_since                       1.343e-04  1.431e-05   9.382  < 2e-16 ***
host_about                       5.247e-03  5.719e-04   9.176  < 2e-16 ***
host_response_timeNoInfo         7.376e-01  2.787e-01   2.646 0.008139 **
host_response_timewithin a day  -7.916e-01  3.413e-01  -2.320 0.020358 *
host_response_timewithin a few hours -5.040e-01  3.484e-01  -1.447 0.148018
host_response_timewithin an hour -2.118e-01  3.498e-01  -0.606 0.544786
host_response_rate               3.233e-02  3.086e-03  10.475  < 2e-16 ***
host_total_listings_count        7.690e-03  4.066e-03   1.891 0.058610 .
host_has_profile_pic1            3.131e-01  1.169e-01   2.677 0.007423 **
room_typeHotel room             -1.605e+00  8.393e-01  -1.912 0.055905 .
room_typePrivate room            5.548e-01  4.787e-02  11.592  < 2e-16 ***
room_typeShared room            -5.341e-01  3.708e-01  -1.440 0.149760
accommodates                    -1.513e-01  1.524e-02  -9.928  < 2e-16 ***
bathrooms_text                   6.288e-02  1.919e-02   3.277 0.001049 **
amenities                        1.976e-02  1.157e-03  17.081  < 2e-16 ***
price                            3.894e-01  3.815e-02  10.208  < 2e-16 ***
minimum_nights                   8.628e-02  1.307e-02   6.603 4.02e-11 ***
maximum_nights                   1.754e-04  3.334e-05   5.261 1.44e-07 ***
number_of_reviews_ltm            5.531e-02  1.286e-03  42.999  < 2e-16 ***
review_scores_rating             6.932e+00  2.136e-01  32.455  < 2e-16 ***
review_scores_accuracy           5.201e-01  1.508e-01   3.448 0.000565 ***
review_scores_cleanliness        6.302e-01  9.705e-02   6.494 8.36e-11 ***
review_scores_checkin            2.627e-01  1.252e-01   2.099 0.035803 *
review_scores_communication      1.041e+00  2.075e-01   5.015 5.30e-07 ***
review_scores_location          -4.178e-01  7.652e-02  -5.460 4.77e-08 ***
instant_bookable1               -1.490e-01  4.099e-02  -3.634 0.000279 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41164  on 45163  degrees of freedom
Residual deviance: 25854  on 45136  degrees of freedom
AIC: 25910

Number of Fisher Scoring iterations: 8
```

<그림3>

```
> summary(mod2) # AIC: 26972

Call:
glm(formula = host_is_superhost ~ ., family = binomial(link = "logit"),
    data = london)

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -4.869e+01  9.317e-01 -52.261  < 2e-16 ***
description                      3.309e-03  5.402e-04   6.125 9.09e-10 ***
host_since                       1.240e-04  1.405e-05   8.824  < 2e-16 ***
host_about                       4.868e-03  5.504e-04   8.843  < 2e-16 ***
host_response_timeNoInfo         6.843e-01  2.773e-01   2.468 0.013595 *
host_response_timewithin a day  -7.471e-01  3.385e-01  -2.207 0.027318 *
host_response_timewithin a few hours -4.560e-01  3.456e-01  -1.319 0.187006
host_response_timewithin an hour -1.686e-01  3.469e-01  -0.486 0.626889
host_response_rate               3.169e-02  3.044e-03  10.410  < 2e-16 ***
host_total_listings_count       -6.147e-03  3.948e-03  -1.557 0.119541
host_has_profile_pic1            2.957e-01  1.159e-01   2.550 0.010758 *
room_typeHotel room             -1.689e+00  8.065e-01  -2.094 0.036227 *
room_typePrivate room            5.043e-01  4.682e-02  10.770  < 2e-16 ***
room_typeShared room            -6.612e-01  3.622e-01  -1.825 0.067947 .
accommodates                    -1.285e-01  1.494e-02  -8.604  < 2e-16 ***
bathrooms_text                   5.092e-02  1.883e-02   2.705 0.006839 **
amenities                        2.247e-02  1.140e-03  19.703  < 2e-16 ***
price                            2.704e-01  3.681e-02   7.345 2.06e-13 ***
minimum_nights                   9.311e-02  1.284e-02   7.253 4.08e-13 ***
maximum_nights                   1.191e-04  3.254e-05   3.659 0.000253 ***
number_of_reviews_ltm            4.616e-02  1.147e-03  40.256  < 2e-16 ***
instant_bookable1               -1.604e-01  4.011e-02  -3.999 6.37e-05 ***
review_scores                    1.661e+00  3.510e-02  47.320  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41164  on 45163  degrees of freedom
Residual deviance: 26926  on 45141  degrees of freedom
AIC: 26972

Number of Fisher Scoring iterations: 7
```

<그림4>

```
> summary(mod3) # AIC: 26984

Call:
glm(formula = host_is_superhost ~ description + host_since +
    host_about + host_response_time + host_response_rate + host_total_listings_count +
    host_has_profile_pic + room_type + accommodates + bathrooms_text +
    amenities + price + minimum_nights + number_of_reviews_ltm +
    instant_bookable + review_scores, family = binomial(link = "logit"),
    data = london)

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -4.840e+01  9.264e-01 -52.251  < 2e-16 ***
description                      3.348e-03  5.401e-04   6.199 5.67e-10 ***
host_since                       1.254e-04  1.404e-05   8.936  < 2e-16 ***
host_about                       4.890e-03  5.498e-04   8.895  < 2e-16 ***
host_response_timeNoInfo         6.883e-01  2.773e-01   2.482  0.01305 *
host_response_timewithin a day  -7.408e-01  3.384e-01  -2.189  0.02859 *
host_response_timewithin a few hours -4.480e-01  3.455e-01  -1.297  0.19474
host_response_timewithin an hour -1.596e-01  3.467e-01  -0.460  0.64539
host_response_rate               3.151e-02  3.043e-03  10.356  < 2e-16 ***
host_total_listings_count       -5.837e-03  3.946e-03  -1.479  0.13907
host_has_profile_pic1            2.954e-01  1.159e-01   2.548  0.01083 *
room_typeHotel room             -1.650e+00  8.053e-01  -2.049  0.04042 *
room_typePrivate room            5.023e-01  4.681e-02  10.730  < 2e-16 ***
room_typeShared room            -6.519e-01  3.622e-01  -1.800  0.07192 .
accommodates                    -1.283e-01  1.493e-02  -8.594  < 2e-16 ***
bathrooms_text                   5.205e-02  1.882e-02   2.766  0.00567 **
amenities                        2.253e-02  1.139e-03  19.768  < 2e-16 ***
price                            2.708e-01  3.681e-02   7.355 1.91e-13 ***
minimum_nights                   9.238e-02  1.283e-02   7.198 6.13e-13 ***
number_of_reviews_ltm            4.618e-02  1.146e-03  40.307  < 2e-16 ***
instant_bookable1               -1.599e-01  4.010e-02  -3.987 6.70e-05 ***
review_scores                    1.651e+00  3.493e-02  47.281  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41164  on 45163  degrees of freedom
Residual deviance: 26940  on 45142  degrees of freedom
AIC: 26984

Number of Fisher Scoring iterations: 7
```

<그림5>

분석에 사용된 R코드

```
install.packages("tidyverse")
install.packages("tm")
install.packages("SmowballC")
install.packages("tidytext")
install.packages("textdata")
install.packages("corrplot")
install.packages("car")
library(carData)
library(car)
library(tidyverse)
library(dplyr)
library(ggplot2)
library(NLP)
library(tm)
library(data.table)
library(SnowballC)
library(tidytext)
library(tidyverse)
library(textdata)
library(corrplot)
# 0. 데이터 불러오기
getwd()
london <- read.csv("listings.csv")
## 분석할 변수 선택
selected_columns = c("description","host_since","host_about",
                     "host_response_time", "host_response_rate", "host_is_superhost",
                     "host_total_listings_count", "host_has_profile_pic",
                     "latitude", "longitude","room_type","accommodates",
                     "bathrooms_text", "bedrooms", "beds", "amenities",
                     "price", "minimum_nights", "maximum_nights",
                     "number_of_reviews_ltm", "review_scores_rating",
                     "review_scores_accuracy", "review_scores_cleanliness",
                     "review_scores_checkin", "review_scores_communication",
                     "review_scores_location", "instant_bookable")
london = london %>% select(selected_columns)
# 1. 데이터 확인
str(london)
sapply(london, class)
# 2. 전처리
# 2-1. 결측치 확인
```

```
sum(is.na(london))
colSums(is.na(london)) # 컬럼별 결측치 개수
london <- select(london, -c(bedrooms, longitude, latitude)) # 결측치가 많은 bedrooms 삭제 / 위도, 경도 삭제
colSums(is.na(london)) # 컬럼별 결측치 개수
london = na.omit(london) # 결측치 존재 행 삭제
sum(is.na(london)) # 다시 결측치 확인 : 0개
# 2-2. 변수 전처리
########## [description, host_about] ##########
cleanFun <- function(htmlString) {
  htmlString <- gsub("<.*?>", "", htmlString)
  htmlString <- gsub("[^A-Za-z]", " ", htmlString)
  htmlString <- removeWords(htmlString, stopwords('en'))
  htmlString <- gsub(" +", " ", htmlString)
  return(sapply(strsplit(htmlString, " "), length))
}london$description <- cleanFun(london$description)
london$host_about <- cleanFun(london$host_about)
########## [host_since] ##########
dplyr::count(london, host_since, sort = TRUE)
london$host_since[which(london$host_since== '')]<- NA
sum(is.na(london$host_since))
london <- na.omit(london)
london$host_since <- as.Date("2023-09-06")-as.Date(london$host_since) # 기준 연도 바꿈
london$host_since <- as.numeric(london$host_since, units='days')
########## [host_response_time] ##########
table(london$host_response_time)
sum(is.na(london$host_response_time))
# 'N/A'를 'NoInfo'로 대체
london <- london %>%
  mutate(host_response_time = ifelse(host_response_time == 'N/A', 'NoInfo', host_response_time))
# change data type for dummy variable (character -> factor)
london$host_response_time=as.factor(london$host_response_time)
########## [host_response_rate] ##########
table(london$host_response_rate)
sum(is.na(london$host_response_rate))
# N/A 와 "" 값 : '0%' 로 변경 -> 응답 안했다고 간주
london <- london %>%
  mutate(host_response_rate = ifelse(host_response_rate == 'N/A', '0%',
```

```
host_response_rate))
# change data type (factor -> numeric)
london$host_response_rate <- as.numeric(gsub("\\D", "", london$host_response_rate))
str(london)
########## [host_is_superhost] ############
#dummy variable --> t=1, f=0
london$host_is_superhost <- ifelse(london$host_is_superhost=='t',1,0)
# 변수 유형 바꿈
london$host_is_superhost <- as.factor(london$host_is_superhost)
########## [room_type] ##########
table(london$room_type)
# replace outlier with most frequent variable : 'Entire home/apt'
# london %>% filter(room_type=='')
# which(london$room_type == '')
# london[66856,11] = 'Entire home/apt'
# check variable distribution
dplyr::count(london, room_type, sort = TRUE)
# change data type
london$room_type = as.factor(london$room_type)
########## [bathrooms_text] ##########
#check variable distribution
dplyr::count(london, bathrooms_text, sort = TRUE)
#total 41 types of baths and shared-baths --> Regardless of bathroom type, only count number of bathroom
which(london$bathrooms_text == '')
london$bathrooms_text[which(london$bathrooms_text== '')]<- '1 bath'
london$bathrooms_text[which(london$bathrooms_text== 'Shared half-bath')] <- '1 bath'
london$bathrooms_text[which(london$bathrooms_text== 'Half-bath')]<- '1 bath'
london$bathrooms_text[which(london$bathrooms_text== 'Private half-bath')]<- '1 bath'
london$bathrooms_text <- as.numeric(gsub("[A-Za-z]", "", london$bathrooms_text))
#london$bathrooms_text <- ifelse(london$bathrooms_text==NA, 0, next)
########## [amenities] ##########
# count number of amenities
london$amenities <- sapply(strsplit(london$amenities, ","), length)
amenities <- table(london$amenities)
barplot(amenities, beside = TRUE, legend=TRUE)
########## [price] ##########
#remove "$" sign
london$price = gsub("[\\$]","",london$price)
# change data type
```

```
london$price = as.numeric(london$price)
# 결측치 행 삭제
sum(is.na(london$price))
london = na.omit(london)
########## [instant_bookable] ##########
london$instant_bookable <- ifelse(london$instant_bookable=='t',1,0)
# change to factor variables
london$instant_bookable <- as.factor(london$instant_bookable)
#########[host_has_profile_pic]####################
# change to factor variables
london$host_has_profile_pic <- ifelse(london$host_has_profile_pic == "t", 1, 0)
london$host_has_profile_pic <- as.factor(london$host_has_profile_pic)
# 추가 전처리
#######################################
#                 이상치 제거                 #
#######################################
summary(london)
### beds 변수 삭제
london = select(london, -c("beds"))
### price 로그 변환
london$price = log(london$price)
which.max(london$price)
### 화장실 가장 많은 1개 삭제
max_bathroom_row <- which.max(london$bathrooms_text)
london$bathrooms_text[36802] # 화장실 최대 개수
london <- london[-max_bathroom_row, ] # 화장실 0인 컬럼 삭제
### 화장실 비율 변환
london <- london[london$bathrooms_text != 0, ]
london$bathrooms_text = london$accommodates / london$bathrooms_text
### 이상치들 제거
london <- london[!(london$host_is_superhost == 1 & london$review_scores_rating < 4.8), ]
london <- london[!(london$description %in% boxplot.stats(london$description)$out), ]
london <- london[!(london$host_since %in% boxplot.stats(london$host_since)$out), ]
london <- london[!(london$host_total_listings_count %in% boxplot.stats(london$host_total_listings_count)$out), ]
london <- london[!(london$amenities %in% boxplot.stats(london$amenities)$out), ]
london <- london[!(london$price %in% boxplot.stats(london$price)$out), ]
london <- london[!(london$minimum_nights %in% boxplot.stats(london$minimum_nights)$out), ]
```

```
london <- london[!(london$maximum_nights %in% boxplot.stats(london$maximum_nights)$out), ]
london$review_scores_accuracy <- ifelse(london$review_scores_accuracy == "", 0, as.numeric(london$review_scores_accuracy))
london$review_scores_rating <- ifelse(london$review_scores_rating== "", 0, as.numeric(london$review_scores_rating))
london$review_scores_checkin <- ifelse(london$review_scores_checkin == "", 0, as.numeric(london$review_scores_checkin))
london$review_scores_cleanliness <- ifelse(london$review_scores_cleanliness == "", 0, as.numeric(london$review_scores_cleanliness))
london$review_scores_communication <- ifelse(london$review_scores_communication == "", 0, as.numeric(london$review_scores_communication))
london$review_scores_location <- ifelse(london$review_scores_location == "", 0, as.numeric(london$review_scores_location))
### 리뷰 0점 제거 ###
london <- london[london$review_scores_rating != 0, ]
london <- london[london$review_scores_accuracy != 0, ]
london <- london[london$review_scores_checkin != 0, ]
london <- london[london$review_scores_cleanliness != 0, ]
london <- london[london$review_scores_communication != 0, ]
london <- london[london$review_scores_location != 0, ]
# 4.8 이상 데이터 만들기
london_sh <- london[london$review_scores_rating >= 4.8, , drop = FALSE]
# 2-3 최종 데이터 확인 # 이거 가지고 EDA
str(london)
numeric_ld <- select_if(london, is.numeric)
cor(numeric_ld)
corrplot(cor(numeric_ld),method = "color")
# 3. EDA (logtitude, latitude, bathrooms / beds 삭제 - accommodates랑 높은 상관관계인데 둘중에 beds를 삭제하는것이 좋다고 해석면에서 생각)
# description
summary(london$description)
ggplot(london, aes(x = london$host_is_superhost, y = london$description)) +
  geom_boxplot() +
  labs(title = "Box Plot of london$description by Superhost Status", x = "Superhost", y = "london$description") +
  theme_minimal()
ggplot(london, aes(x = london$description, fill = london$host_is_superhost)) +
  geom_bar(position = "dodge") +
  labs(title = "Distribution of london$description by Superhost Status",
```

```
        x = "london$description",
        y = "Count") +
  theme_minimal()
ggplot(london, aes(x = london$description, fill = london$host_is_superhost)) +
  geom_bar(position = "stack") +
  labs(title = "Distribution of london$description by Superhost Status",
        x = "london$description",
        y = "Count") +
  theme_minimal()
# host_since
## 박스 플롯
ggplot(london, aes(x = london$host_is_superhost, y = london$host_since)) +
  geom_boxplot() +
  labs(title = "Box Plot of host_since by Superhost Status", x = "Superhost", y =
"host_since") +
  theme_minimal()
ggplot(london, aes(x = london$host_since, fill = london$host_is_superhost)) +
  geom_bar(position = "dodge") +
  labs(title = "Distribution of host_since by Superhost Status",
        x = "host_since",
        y = "Count") +
  theme_minimal()
ggplot(london, aes(x = london$host_since, fill = london$host_is_superhost)) +
  geom_bar(position = "stack") +
  labs(title = "Distribution of host_since by Superhost Status",
        x = "host_since",
        y = "Count") +
  theme_minimal()
# host_about
## 박스 플롯
ggplot(london, aes(x = london$host_is_superhost, y = london$host_about)) +
  geom_boxplot() +
  labs(title = "Box Plot of london$host_about by Superhost Status", x = "Superhost", y
= "london$description") +
  theme_minimal()
ggplot(london, aes(x = host_about, fill = host_is_superhost)) +
  geom_bar(position = "dodge") +
  labs(title = "Distribution of host_about by Superhost Status",
        x = "host_about",
        y = "Count") +
```

```
  theme_minimal() +
  scale_x_discrete(limits = c(0,150)) +
  scale_y_continuous(limits = c(0,100))
ggplot(london, aes(x = london$host_about, fill = london$host_is_superhost)) +
  geom_bar(position = "stack") +
  labs(title = "Distribution of london$host_about by Superhost Status",
       x = "london$host_about",
       y = "Count") +
  theme_minimal() +
  scale_x_discrete(limits = c(0,150)) +
  coord_cartesian(ylim = c(0,300))
# host_response_time
summary(london$host_response_time)
## 수평 그래프
ggplot(london, aes(x = london$host_response_time, fill = factor(london$host_is_superhost))) +
  geom_bar(position = "dodge") +
  scale_fill_manual(values = c("0" = "red", "1" = "blue")) +  # 슈퍼호스트 여부에 따른 색 지정
  labs(title = "host_response_time & host_is_superhost", x = "host_response_time", y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle =45, hjust = 1, size = 8))
## 수직 그래프
ggplot(london, aes(x =london$host_response_time, fill = london$host_is_superhost)) +
  geom_bar(position = "stack") +
  labs(title = "Distribution of london$host_response_time by Superhost Status",
       x = "london$host_response_time",
       y = "Count") +
  theme_minimal()
## 비율로 보기
prop_data <- london %>%
  group_by(host_response_time, host_is_superhost) %>%
  summarise(count = n()) %>%
  group_by(host_response_time) %>%
  mutate(proportion = count / sum(count))
ggplot(prop_data, aes(x = host_response_time, y = proportion, fill = host_is_superhost)) +
  geom_col(position = "stack") +
  scale_fill_manual(values = c("0" = "red", "1" = "blue")) +
```

```
  labs(title = "Proportion of host_is_superhost by host_response_time",
       x = "host_response_time",
       y = "Proportion") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 0, hjust = 1, size = 6))
# host_response_rate
## 박스 플롯
ggplot(london, aes(x = london$host_is_superhost, y = london$host_response_rate)) +
  geom_boxplot() +
  labs(title = "Box Plot of london$host_response_rate by Superhost Status", x =
"Superhost", y = "london$host_response_rate") +
  theme_minimal()
summary(london$host_response_rate)
ggplot(london, aes(x = london$host_response_rate, fill =
factor(london$host_is_superhost))) +
  geom_bar(position = "dodge") +
  scale_fill_manual(values = c("0" = "red", "1" = "blue")) +  # 슈퍼호스트 여부에 따른 색
지정
  labs(title = "host_response_rate & host_is_superhost", x = "host_response_rate", y =
"Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 0, hjust = 1, size = 8))
ggplot(london, aes(x =london$host_response_rate, fill = london$host_is_superhost)) +
  geom_bar(position = "stack") +
  labs(title = "Distribution of london$host_response_rate by Superhost Status",
       x = "london$host_response_rate",
       y = "Count") +
  theme_minimal()
# host_is_superhost(종속변수)
london$host_is_superhost
table(london$host_is_superhost)
## 막대그래프
ggplot(london, aes(x = london$host_is_superhost, fill = host_is_superhost)) +
  geom_bar() +
  labs(title = "Distribution of host_is_superhost",
       x = "Superhost Status",
       y = "Count") +
  theme_minimal()
# host_total_listings_count
## 박스 플롯
```

```
ggplot(london, aes(x = london$host_is_superhost, y = london$host_total_listings_count)) +
  geom_boxplot() +
  labs(title = "Box Plot of london$host_total_listings_count by Superhost Status", x = "Superhost", y = "london$host_total_listings_count") +
  theme_minimal()
ggplot(london, aes(x = london$host_total_listings_count, fill = factor(london$host_is_superhost))) +
  geom_bar(position = "dodge") +
  scale_fill_manual(values = c("0" = "red", "1" = "blue")) +  # 슈퍼호스트 여부에 따른 색 지정
  labs(title = "host_total_listings_count & host_is_superhost", x = "host_total_listings_count", y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 0, hjust = 1, size = 8)) +
  coord_cartesian(xlim = c(0,10))  #이상치 날림
ggplot(london, aes(x = london$host_total_listings_count, fill = factor(london$host_is_superhost))) +
  geom_bar(position = "stack") +
  scale_fill_manual(values = c("0" = "red", "1" = "blue")) +  # 슈퍼호스트 여부에 따른 색 지정
  labs(title = "host_total_listings_count & host_is_superhost", x = "host_total_listings_count", y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 0, hjust = 1, size = 8)) +
  coord_cartesian(xlim = c(0,10))  #이상치 날림
# host_has_profile_pic
london_super <- subset(london,host_is_superhost==1)
london_host <- subset(london,host_is_superhost==0)
## 수평 그래프
ggplot(london, aes(x = london$host_has_profile_pic, fill = factor(london$host_is_superhost))) +
  geom_bar(position = "dodge") +
  scale_fill_manual(values = c("0" = "red", "1" = "blue")) +  # 슈퍼호스트 여부에 따른 색 지정
  labs(title = "host_has_profile_pic & host_is_superhost", x = "host_has_profile_pic", y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 0, hjust = 1, size = 8))
## 수직 그래프
```

```
ggplot(london, aes(x = london$host_has_profile_pic, fill = factor(london$host_is_superhost))) +
  geom_bar(position = "stack") +
  scale_fill_manual(values = c("0" = "red", "1" = "blue")) +  # 슈퍼호스트 여부에 따른 색 지정
  labs(title = "host_has_profile_pic & host_is_superhost", x = "host_has_profile_pic", y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 0, hjust = 1, size = 8))
# 비율로 보기
prop_data <- london %>%
  group_by(host_has_profile_pic, host_is_superhost) %>%
  summarise(count = n()) %>%
  group_by(host_has_profile_pic) %>%
  mutate(proportion = count / sum(count))
ggplot(prop_data, aes(x = factor(host_has_profile_pic), y = proportion, fill = factor(host_is_superhost))) +
  geom_col(position = "stack") +
  scale_fill_manual(values = c("0" = "red", "1" = "blue")) +
  labs(title = "Proportion of host_is_superhost by host_has_profile_pic",
       x = "Host_has_profile_pic",
       y = "Proportion") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 0, hjust = 1, size = 6))
# room_type
table(london_host$room_type)
table(london_super$room_type)
## 수평 그래프
ggplot(london, aes(x = london$room_type, fill = factor(london$host_is_superhost))) +
  geom_bar(position = "dodge") +
  scale_fill_manual(values = c("0" = "red", "1" = "blue")) +  # 슈퍼호스트 여부에 따른 색 지정
  labs(title = "room_type & host_is_superhost", x = "room_type", y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 0, hjust = 1, size = 6))
## 수직 그래프
ggplot(london, aes(x = london$room_type, fill = factor(london$host_is_superhost))) +
  geom_bar(position = "stack") +
  scale_fill_manual(values = c("0" = "red", "1" = "blue")) +  # 슈퍼호스트 여부에 따른 색 지정
```

```
  labs(title = "room_type & host_is_superhost", x = "room_type", y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 0, hjust = 1, size = 6))
## 비율로 보기
prop_data <- london %>%
  group_by(room_type, host_is_superhost) %>%
  summarise(count = n()) %>%
  group_by(room_type) %>%
  mutate(proportion = count / sum(count))
ggplot(prop_data, aes(x = factor(room_type), y = proportion, fill =
factor(host_is_superhost))) +
  geom_col(position = "stack") +
  scale_fill_manual(values = c("0" = "red", "1" = "blue")) +
  labs(title = "Proportion of host_is_superhost by room_type",
       x = "room_type",
       y = "Proportion") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 0, hjust = 1, size = 8))
# accommodates
## 박스 플롯
ggplot(london, aes(x = london$host_is_superhost, y = london$accommodates)) +
  geom_boxplot() +
  labs(title = "Box Plot of london$accommodates by Superhost Status", x =
"Superhost", y = "london$accommodates") +
  theme_minimal()
ggplot(london, aes(x = london$accommodates, fill = factor(london$host_is_superhost)))
+
  geom_bar(position = "dodge") +
  scale_fill_manual(values = c("0" = "red", "1" = "blue")) +  # 슈퍼호스트 여부에 따른 색
지정
  labs(title = "accommodates & host_is_superhost", x = "accommodates", y = "Count")
+
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 0, hjust = 1, size = 8))
ggplot(london, aes(x = london$accommodates, fill = factor(london$host_is_superhost)))
+
  geom_bar(position = "stack") +
  scale_fill_manual(values = c("0" = "red", "1" = "blue")) +  # 슈퍼호스트 여부에 따른 색
지정
  labs(title = "accommodates & host_is_superhost", x = "accommodates", y = "Count")
```

```
+
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 0, hjust = 1, size = 8))
# bathrooms_text
## 박스 플롯
ggplot(london, aes(x = london$host_is_superhost, y = london$bathrooms_text)) +
  geom_boxplot() +
  labs(title = "Box Plot of london$bathrooms_text by Superhost Status", x =
"Superhost", y = "london$bathrooms_text") +
  theme_minimal()
ggplot(london, aes(x = london$bathrooms_text, fill = factor(london$host_is_superhost)))
+
  geom_bar(position = "dodge") +
  scale_fill_manual(values = c("0" = "red", "1" = "blue")) +  # 슈퍼호스트 여부에 따른 색
지정
  labs(title = "bathrooms_text & host_is_superhost", x = "bathrooms_text", y = "Count")
+
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 0, hjust = 1, size = 8))  +  # x축 레이블 크
기 조절
  coord_cartesian(xlim = c(0, 5))  # 이상치날림
london_super <- subset(london,host_is_superhost==1)
london_host <- subset(london,host_is_superhost==0)
dplyr::count(london_host, bathrooms_text, sort = TRUE)
dplyr::count(london_super, bathrooms_text, sort = TRUE)
tbl_s <- london_super$bathrooms_text
tbl_h <- london_host$bathrooms_text
boxplot(list(tbl_h, tbl_s),
          names = c("Host", "Super"),
          main = "Boxplot of Host vs Super Bathrooms_text",
          ylab = "Values",
          ylim = c(0, max(c(tbl_h, tbl_s)))) +  # x축 레이블 크기 조절

  # bedrooms -> 결측치 많아서 삭제
  bedrooms_h <- dplyr::count(london_host, bedrooms,sort = TRUE)
bedrooms_s <- dplyr::count(london_super, bedrooms,sort = TRUE)
dplyr::count(london_super, bedrooms,sort = TRUE)
dplyr::count(london_host, bedrooms,sort = TRUE)
ggplot(london, aes(x = london$bedrooms, fill = factor(london$host_is_superhost))) +
  geom_bar(position = "dodge") +
```

```
  scale_fill_manual(values = c("0" = "red", "1" = "blue")) +  # 슈퍼호스트 여부에 따른 색 지정
  labs(title = "bedrooms & host_is_superhost", x = "bedrooms", y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 0, hjust = 1, size = 8)) +  # x축 레이블 크기 조절
  coord_cartesian(xlim = c(0, 22))  # x축 범위 지정
## 박스 플롯
ggplot(london, aes(x = factor(london$host_is_superhost), y = london$beds)) +
  geom_boxplot() +
  labs(title = "Box Plot of beds by Superhost Status", x = "Superhost", y = "beds") +
  theme_minimal()
ggplot(london, aes(x = london$beds, fill = factor(london$host_is_superhost))) +
  geom_bar(position = "stack") +
  scale_fill_manual(values = c("0" = "red", "1" = "blue")) +  # 슈퍼호스트 여부에 따른 색 지정
  labs(title = "bedrs & host_is_superhost", x = "beds", y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 0, hjust = 1, size = 8)) +  # x축 레이블 크기 조절
  coord_cartesian(xlim = c(0, 10)) #이상치 날림
# amenities
## 박스 플롯
ggplot(london, aes(x = factor(london$host_is_superhost), y = london$amenities)) +
  geom_boxplot() +
  labs(title = "Box Plot of amenities by Superhost Status", x = "Superhost", y = "amenities") +
  theme_minimal()
ggplot(london, aes(x = london$amenities, fill = london$host_is_superhost)) +
  geom_bar(position = "dodge") +
  labs(title = "Distribution of Amenities by Superhost Status",
       x = "Amenities",
       y = "Count") +
  theme_minimal()
ggplot(london, aes(x = london$amenities, fill = london$host_is_superhost)) +
  geom_bar(position = "stack") +
  labs(title = "Distribution of Amenities by Superhost Status",
       x = "Amenities",
       y = "Count") +
  theme_minimal()
```

```
# price
## 박스 플롯
ggplot(london, aes(x = london$host_is_superhost, y = london$price)) +
  geom_boxplot() +
  labs(title = "Box Plot of price by Superhost Status", x = "Superhost", y = "price") +
  theme_minimal()
ggplot(london, aes(x = london$price, fill = london$host_is_superhost)) +
  geom_bar(position = "dodge") +
  labs(title = "Distribution of price by Superhost Status",
       x = "price",
       y = "Count") +
  theme_minimal()
ggplot(london, aes(x = london$price, fill = london$host_is_superhost)) +
  geom_bar(position = "stack") +
  labs(title = "Distribution of price by Superhost Status",
       x = "price",
       y = "Count") +
  theme_minimal()
# minimum_nights
## 박스 플롯
ggplot(london, aes(x = london$host_is_superhost, y = london$minimum_nights)) +
  geom_boxplot() +
  labs(title = "Box Plot of minimum_nights by Superhost Status", x = "Superhost", y =
"minimum_nights") +
  theme_minimal()
ggplot(london, aes(x = london$minimum_nights, fill = london$host_is_superhost)) +
  geom_bar(position = "dodge") +
  labs(title = "Distribution of minimum_nights by Superhost Status",
       x = "minimum_nights",
       y = "Count") +
  theme_minimal()
# maximum_nights
## 박스플롯
ggplot(london, aes(x = london$host_is_superhost, y = london$maximum_nights)) +
  geom_boxplot() +
  labs(title = "Box Plot of maximum_nights by Superhost Status", x = "Superhost", y =
"maximum_nights") +
  theme_minimal()
ggplot(london, aes(x = maximum_nights, fill = host_is_superhost)) +
  geom_bar(position = "dodge") +
```

```
  labs(title = "Distribution of maximum_nights by Superhost Status",
       x = "maximum_nights",
       y = "Count") +
  theme_minimal()
# number_of_reviews_ltm
## 박스 플롯
ggplot(london, aes(x = london$host_is_superhost, y = london$number_of_reviews_ltm))
+
  geom_boxplot() +
  labs(title = "Box Plot of number_of_reviews_ltm by Superhost Status", x =
"Superhost", y = "number_of_reviews_ltm") +
  theme_minimal()
ggplot(london, aes(x = london$number_of_reviews_ltm, fill = london$host_is_superhost))
+
  geom_bar(position = "dodge") +
  labs(title = "Distribution of number_of_reviews_ltm by Superhost Status",
       x = "number_of_reviews_ltm",
       y = "Count") +
  theme_minimal()
# review_scores_rating
## 박스 플롯
ggplot(london, aes(x = london$host_is_superhost, y = london$review_scores_rating)) +
  geom_boxplot() +
  labs(title = "Box Plot of review_scores_rating by Superhost Status", x = "Superhost", y
= "review_scores_rating") +
  theme_minimal()
ggplot(london, aes(x = london$review_scores_rating, fill = london$host_is_superhost)) +
  geom_bar(position = "dodge") +
  labs(title = "Distribution of review_scores_rating by Superhost Status",
       x = "review_scores_rating",
       y = "Count") +
  theme_minimal()
# review_scores_accuracy
## 박스 플롯
ggplot(london, aes(x = london$host_is_superhost, y = london$review_scores_accuracy))
+
  geom_boxplot() +
  labs(title = "Box Plot of review_scores_accuracy by Superhost Status", x =
"Superhost", y = "review_scores_accuracy") +
  theme_minimal()
```

```
ggplot(london, aes(x = london$review_scores_accuracy, fill = london$host_is_superhost))
+
  geom_bar(position = "dodge") +
  labs(title = "Distribution of review_scores_accuracy by Superhost Status",
       x = "review_scores_accuracy",
       y = "Count") +
  theme_minimal()
# review_scores_cleanliness
## 박스 플롯
ggplot(london, aes(x = london$host_is_superhost, y =
london$review_scores_cleanliness)) +
  geom_boxplot() +
  labs(title = "Box Plot of review_scores_cleanliness by Superhost Status", x =
"Superhost", y = "review_scores_cleanliness") +
  theme_minimal()
ggplot(london, aes(x = london$review_scores_cleanliness, fill =
london$host_is_superhost)) +
  geom_bar(position = "dodge") +
  labs(title = "Distribution of review_scores_cleanliness by Superhost Status",
       x = "review_scores_cleanliness",
       y = "Count") +
  theme_minimal()
# review_scores_checkin
## 박스 플롯
ggplot(london, aes(x = london$host_is_superhost, y = london$review_scores_checkin)) +
  geom_boxplot() +
  labs(title = "Box Plot of review_scores_checkin by Superhost Status", x = "Superhost",
y = "review_scores_checkin") +
  theme_minimal()
ggplot(london, aes(x = london$review_scores_checkin, fill = london$host_is_superhost))
+
  geom_bar(position = "dodge") +
  labs(title = "Distribution of review_scores_checkin by Superhost Status",
       x = "review_scores_checkin",
       y = "Count") +
  theme_minimal()
# review_scores_communication
## 박스 플롯
ggplot(london, aes(x = london$host_is_superhost, y =
london$review_scores_communication)) +
```

```
  geom_boxplot() +
  labs(title = "Box Plot of review_scores_communication by Superhost Status", x = "Superhost", y = "review_scores_communication") +
  theme_minimal()
ggplot(london, aes(x = london$review_scores_communication, fill = london$host_is_superhost)) +
  geom_bar(position = "dodge") +
  labs(title = "Distribution of review_scores_communication by Superhost Status",
       x = "review_scores_communication",
       y = "Count") +
  theme_minimal()
# review_scores_location
## 박스 플롯
ggplot(london, aes(x = london$host_is_superhost, y = london$review_scores_location)) +
  geom_boxplot() +
  labs(title = "Box Plot of review_scores_location by Superhost Status", x = "Superhost", y = "review_scores_location") +
  theme_minimal()
ggplot(london, aes(x = london$review_scores_location, fill = london$host_is_superhost)) +
  geom_bar(position = "dodge") +
  labs(title = "Distribution of review_scores_location by Superhost Status",
       x = "review_scores_location",
       y = "Count") +
  theme_minimal()
# instant_bookable
table(london$host_is_superhost,london$instant_bookable)
## 수평 그래프
ggplot(london, aes(x = london$instant_bookable, fill = london$host_is_superhost)) +
  geom_bar(position = "dodge") +
  labs(title = "Distribution of instant_bookable by Superhost Status",
       x = "instant_bookable",
       y = "Count") +
  theme_minimal()
## 수직 그래프
ggplot(london, aes(x = london$instant_bookable, fill = london$host_is_superhost)) +
  geom_bar(position = "stack") +
  labs(title = "Distribution of instant_bookable by Superhost Status",
       x = "instant_bookable",
       y = "Count") +
```

```
  theme_minimal()
## 비율로 보기
prop_data <- london %>%
  group_by(instant_bookable, host_is_superhost) %>%
  summarise(count = n()) %>%
  group_by(instant_bookable) %>%
  mutate(proportion = count / sum(count))
ggplot(prop_data, aes(x = factor(instant_bookable), y = proportion, fill =
factor(host_is_superhost))) +
  geom_col(position = "stack") +
  scale_fill_manual(values = c("0" = "red", "1" = "blue")) +
  labs(title = "Proportion of host_is_superhost by instant_bookable",
       x = "instant_bookable",
       y = "Proportion") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 0, hjust = 1, size = 8))
# 3.1 기술 통계량
str(london)
dim(london)
# 3.2 시각화
# 4. model
# 4.1 모델 생성
# 모델 0 (로그변환x, 변수변환x)
mod0 = glm(host_is_superhost~.,data = london, family = binomial(link="logit"))
summary(mod0) # AIC: 24037
# 모델 1 (PCA X)
mod1 = glm(host_is_superhost~.,data = london, family = binomial(link="logit"))
summary(mod1) # AIC: 25910
# PCA
selected_columns_reviews = c("review_scores_rating",
                             "review_scores_accuracy", "review_scores_cleanliness",
                             "review_scores_checkin", "review_scores_communication",
                             "review_scores_location")
reviews = london %>% select(selected_columns_reviews)
##### PCA 시각화 #####
pca_result <- prcomp(reviews, scale. = TRUE) # PCA 수행
summary(pca_result) # PCA 결과 요약 표시
screeplot(pca_result, type = "line", main = "Scree Plot") # PCA 수행 결과를 이용하여
scree plot 그리기
screeplot(pca_result, type = "barplot", main = "Cumulative Proportion of Variance") #
```

```
주성분이 설명하는 분산의 누적 비율을 시각화하는 scree plot
##### PCA 계수 확인 #####
install.packages("FactoMineR")
library(FactoMineR)
pca = PCA(reviews)
coefficients <- pca$var$coord
print("주성분 계수:")
print(coefficients)
##### PCA 새로운 변수 추가 #####
london['review_scores'] =
  0.9248577*london['review_scores_rating'] +
  0.8988699*london['review_scores_accuracy'] +
  0.8242932*london['review_scores_cleanliness'] +
  0.8370093*london['review_scores_checkin'] +
  0.8651945*london['review_scores_communication'] +
  0.7306668*london['review_scores_location']
str(london)
selected_columns = c("review_scores_rating","review_scores_accuracy",
"review_scores_cleanliness",
"review_scores_checkin",
"review_scores_communication","review_scores_location")
london = london %>% select(-selected_columns)
# 모델 2 PCA
mod2 = glm(host_is_superhost~.,data = london, family = binomial(link="logit"))
summary(mod2) # AIC: 26972
# 모델 3 PCA + maximum_nights 삭제
deviance_value <- anova(mod2, test = "Chisq") # deviance 확인
deviance_value
mod3 = glm(host_is_superhost ~ description + host_since + host_about +
host_response_time +
             host_response_rate + host_total_listings_count + host_has_profile_pic +
             room_type + accommodates + bathrooms_text + amenities + price +
             minimum_nights + number_of_reviews_ltm +
             instant_bookable + review_scores,data = london, family =
binomial(link="logit"))
summary(mod3) # AIC: 26984
# 최종 모델 상관관계 확인
london_mod = select(london, -c("maximum_nights"))
numeric_ld <- select_if(london_mod, is.numeric)
cor(numeric_ld)
```

```
corrplot(cor(numeric_ld),method = "color")
######### 추가 분석
ggplot(london_sh, aes(x = london_sh$host_is_superhost, y = london_sh$reponse_rate)) +
  geom_boxplot() +
  labs(title = "Box Plot of london_sh$reponse_rate by Superhost Status", x = "Superhost", y = "london_sh$reponse_rate") +
  theme_minimal()
ggplot(london_sh, aes(x = london_sh$host_is_superhost, y = london_sh$amenities)) +
  geom_boxplot() +
  labs(title = "Box Plot of london_sh$amenities by Superhost Status", x = "Superhost", y = "london_sh$amenities") +
  theme_minimal()
ggplot(london_sh, aes(x = london_sh$host_is_superhost, y = london_sh$bathrooms_text)) +
  geom_boxplot() +
  labs(title = "Box Plot of london_sh$bathrooms_text by Superhost Status", x = "Superhost", y = "london_sh$bathrooms_text") +
  theme_minimal()
```