Florida Airbnb 영업 데이터 기반 Host의 수익 향상을 위한 예약률 예측 모델 연구

캡스톤 디자인 최종발표회

엔프로(강수정, 김지민, 송현지)

2023.12.16.

This work was supported by Dr.Jinwon Kim who is an Associate Professor (Department of Tourism, Hospitality and Event Management, University of Florida, USA) and a Director of the Center for Sustainable Business and Community Analytics.

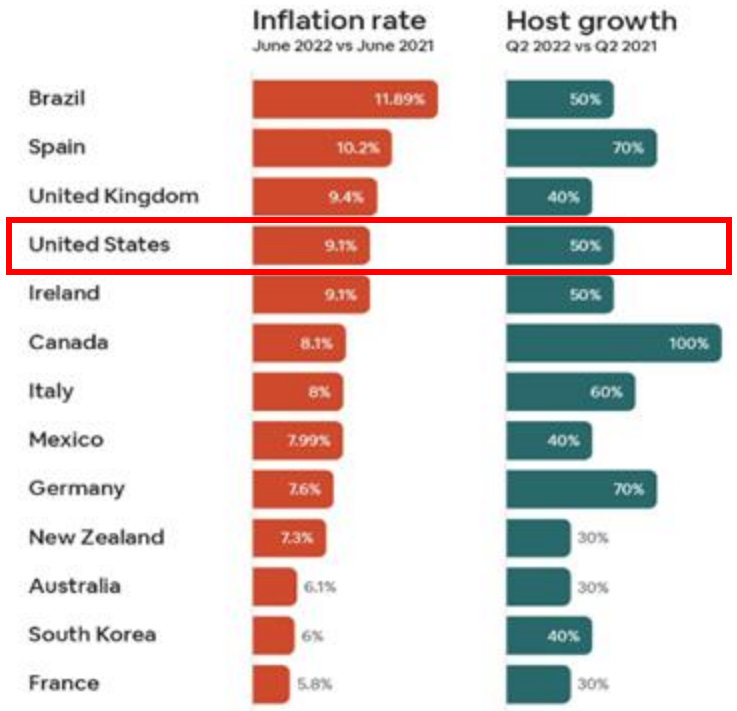
목차

1. 연구배경 및 목적

미국 인플레이션 증가에 따른 수익 창출의 기회로 Airbnb 신규 호스트 급증 및 플로리다 호스트 급여 격차는 지속 증가 예상

'22년 미국 인플레이션 상승 효과

- 미국 인플레이션 9.1% 증가에 따른 미국 내 Airbnb 신규 호스트 50% 이상 증가
- 경제적 압박에 따른 수익 창출의 수단으로 호스트 사업 시작(미국인의 약 41%)



[출처] Airbnb's Post(Linkedin)

호스트가 하나의 일자리로 대체

- '21년 기준 플로리다 키시미 활성 호스트: 약 46,600명

[출처] AllTheRooms Insights Articles

도시	액티브 에어비앤비 리스팅	에어비앤비 입주율 평균(2021)	에어비앤비 입주율 평균(2020년)
뉴욕시, 뉴욕	94,198	24.6%	13.5%
로스앤젤레스, 캘리포니아	59,278	24.2%	18.9%
키시미, 플로리다	46,569	25.1%	13.7%

- '23.12월 기준 플로리다 에어비앤비 호스트 Salary 통계

[출처] ZipRecruiter

	연봉	월급	주간 급여	시급
최고 소득자	$37,364	$3,113	$718	$18
75번째 백분위수	$29,900	$2,491	$575	$14
평균	$26,986	$2,248	$518	$13
25번째 백분위수	$22,400	$1,866	$430	$11

플로리다 Host들 간의 수익 양극화를 줄이고 예약률이 저조한 호스트들의 예약률 향상을 위한

숙소 정보 등록 및 운영 시점에

‘예약률 예측 + 지속 관리 feature’ 제안 서비스

2. 연구 가설

✓ 평균 1박 가격이 저렴하면 예약률이 높을 것이다.

✓ 유사한 특성을 가진 숙소별 군집 간에는 서로 다른 예약률 패턴과 연관 변수들을 가질 것이다.

3. 연구 방법

[분석 방안]

1) 데이터 전처리 및 상관관계 파악	2) SPCA를 통한 차원축소	3) 숙소별 군집화	4) 군집별 모델 학습
• 데이터 수: 43,219개(결측치 X) • 데이터 타입 변환	• Sparse PCA를 통한 주성분 파악 및 개수 선정	• 유사한 특성을 가진 숙소 군집을 위해 K-means clustering 진행	• 활용 모델 1) Multiple Regression 2) Random Forest 3) Support Vector Regression 4) Gradient Boosting
• 변수 간 Correlation 파악 • 로그변환을 통한 데이터 정규분포화	• 추출한 주성분 개수로 차원축소 진행	• 적절한 군집 수 선정 지표: 실루엣 계수, DBI score, 3D 산점도 활용	• 군집별 성능 평가 및 최종 모델 확정 • Shap value 기여도 분석 • 고객 리뷰 Wordcloud 분석
• OLS 회귀분석 실시 • 설명변수 간 다중공선성 여부 확인 • 예약률 예측에 활용할 설명변수 선정	• 주성분 10차원 축소	• 최종 군집 수: 3개 • 군집별 EDA 시각화	

4. 연구내용 및 결과

1-1) 데이터 전처리

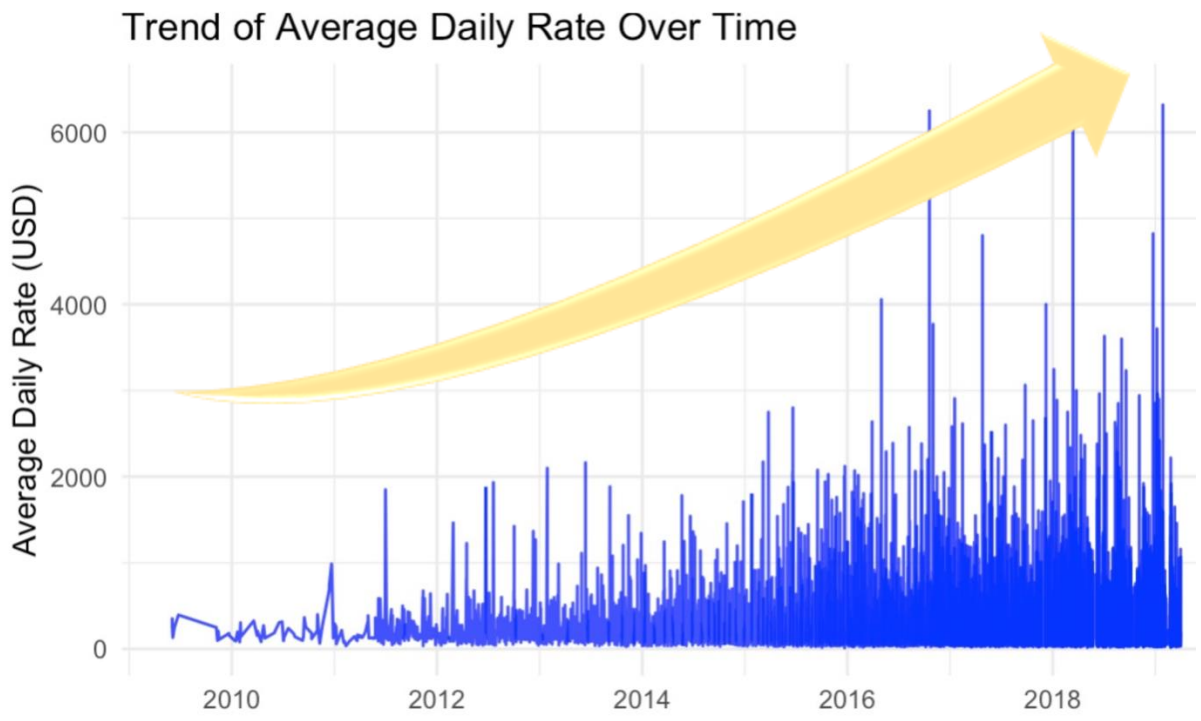
- 데이터 수: 43,219개
- 중복변수, 불필요한 변수 제거 및 Label Encoder를 통한 데이터 타입 변환 → 1차 설명변수: 21개

설명변수							
ADR	평균 1박 요금	**Guest**	예약 가능한 게스트 수	**Ministay**	최소 숙박 수	**Checkrating**	체크인 용이성(10점)
ARL	평균 연간 수익	**Response**	응답률	**photonum**	사진 수	**Locaterating**	위치 만족도 (10점)
Booking	예약 수	**Superhost**	슈퍼호스트 여부	**Overall**	전체 평점(%)	**Valuerating**	가치 만족도 (10점)
Reviewnum	리뷰 수	**Deposit**	보증금	**Commrating**	소통 평점 (10점)		
Bedrooms	침대 수	**Cleaningfee**	청소 요금	**Accuracyrating**	광고 대비 만족도(10점)		
Bathrooms	화장실 수	**Nightfee**	1일 숙박요금	**Cleanrating**	청결도(10점)		

반응변수
Occupancy (예약률)

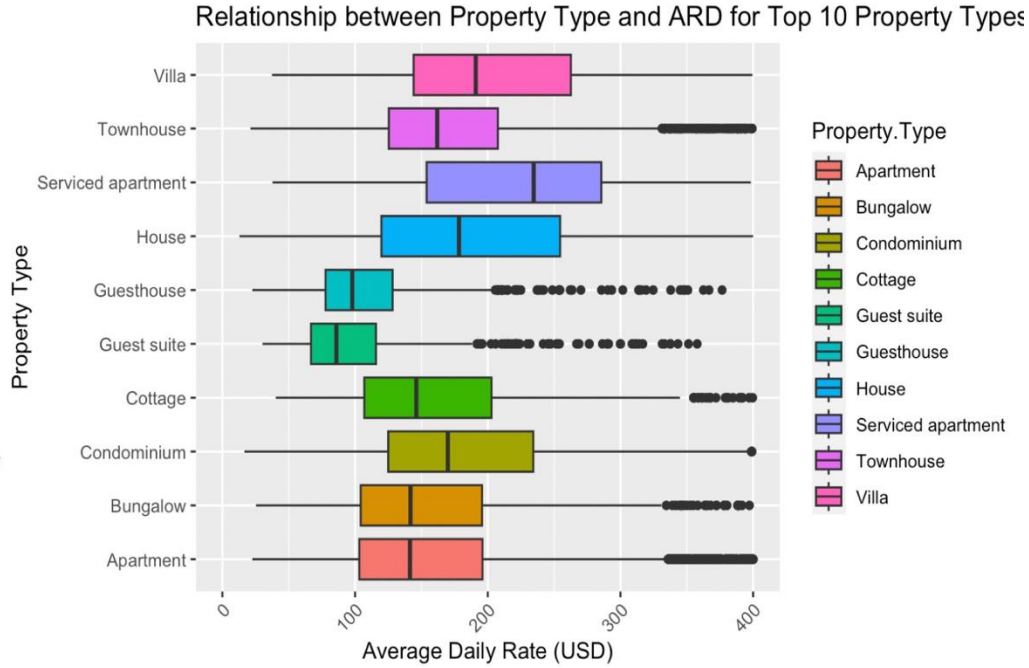
4. 연구내용 및 결과

1-2) EDA를 통한 전체 데이터 분포 확인



[숙소 유형별 TOP 10의 일 평균요금]

- Serviced Aprt > Villa > House > Condominium

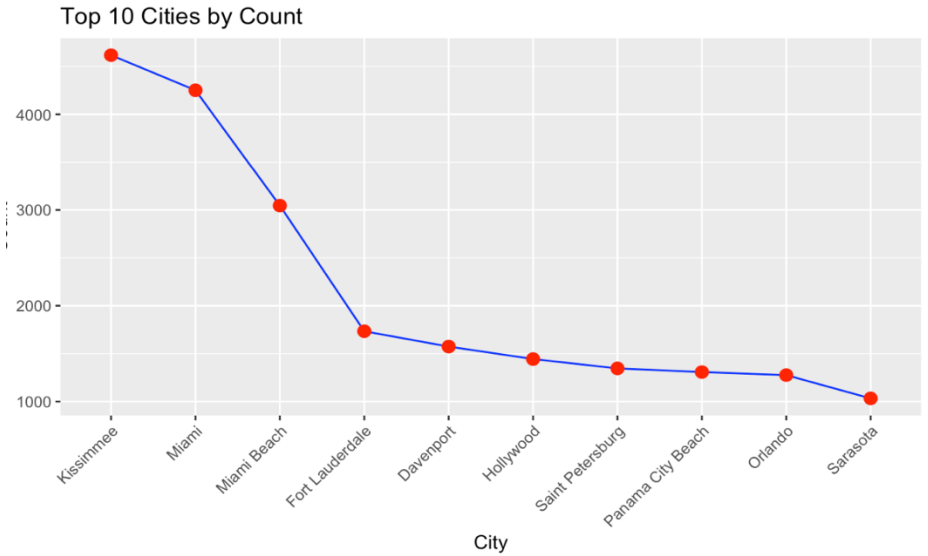


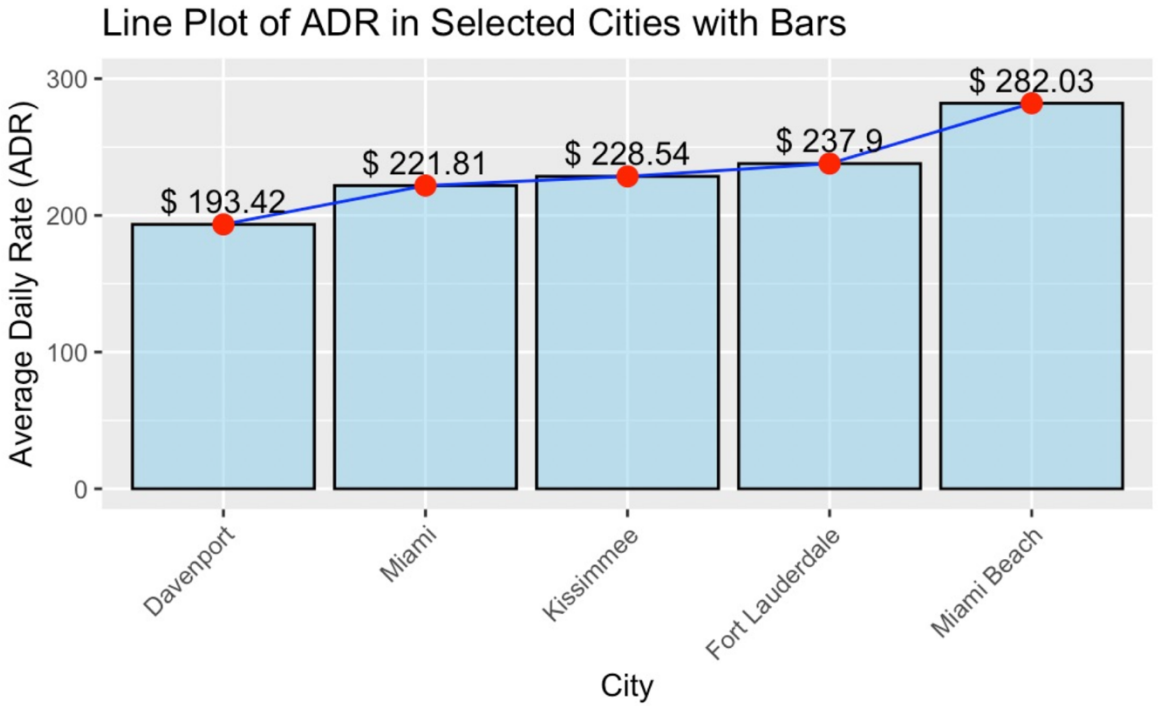
4. 연구내용 및 결과

1-2) EDA를 통한 전체 데이터 분포 확인

[도시별 숙소 수 vs 숙소 수가 많은 도시의 ADR]

Kissimmee >Miami > Miami Beach > Fort Lauderdale> Davenport





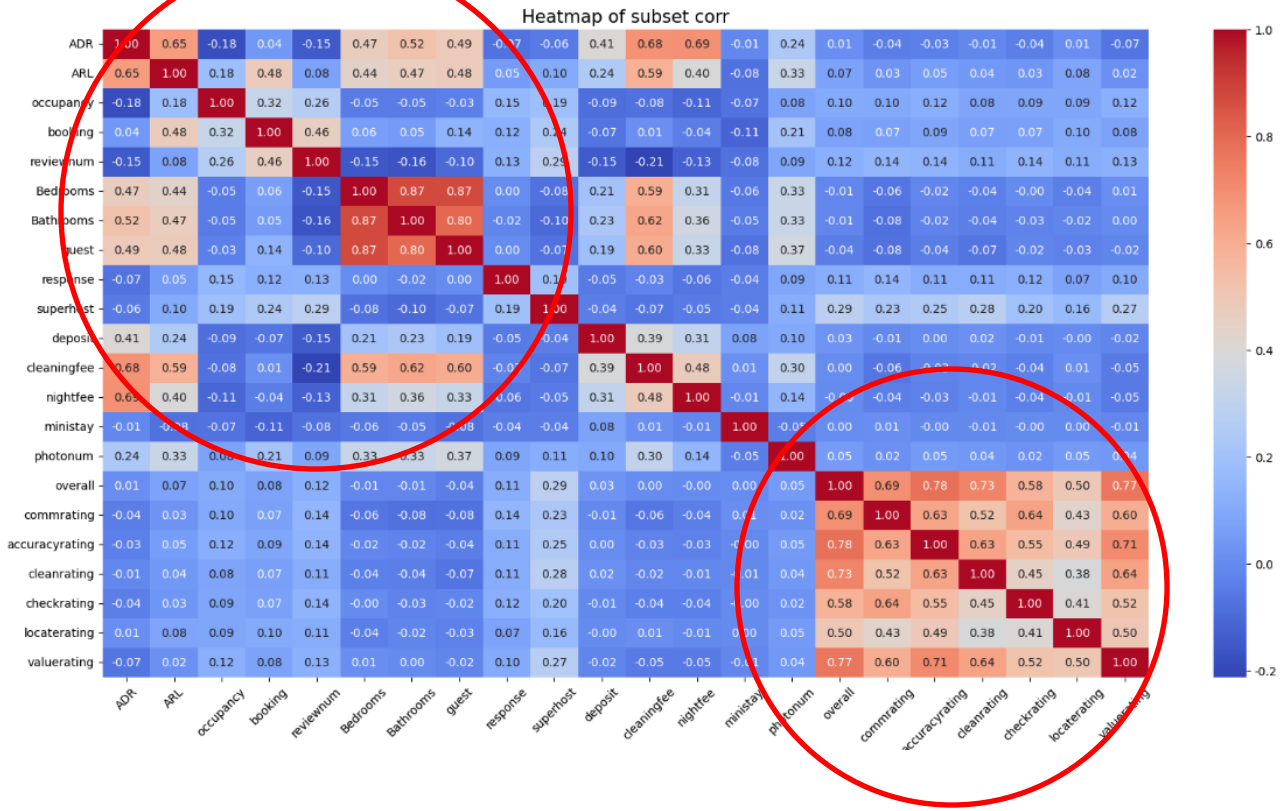
4. 연구내용 및 결과

1-3) 변수 간 상관관계 파악 및 데이터 정규분포화

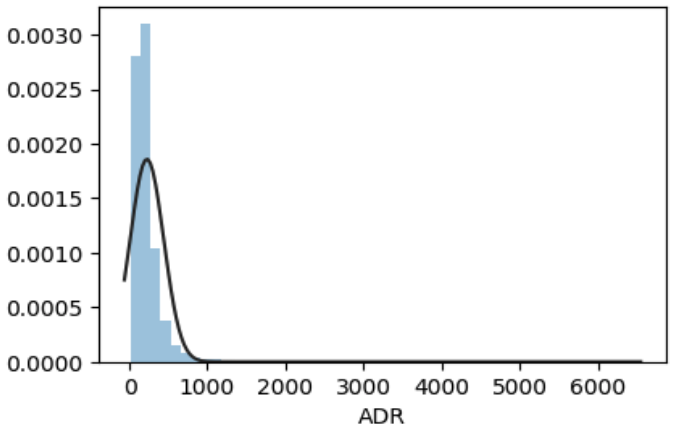
- Heatmap을 활용한 correlation 파악
- 로그변환을 통한 전체 데이터 정규분포화

✓ 1박 요금 ⇔ 연간 수익 / Bedrooms ⇔ Bathrooms ⇔ guest 간 높은 상관관계

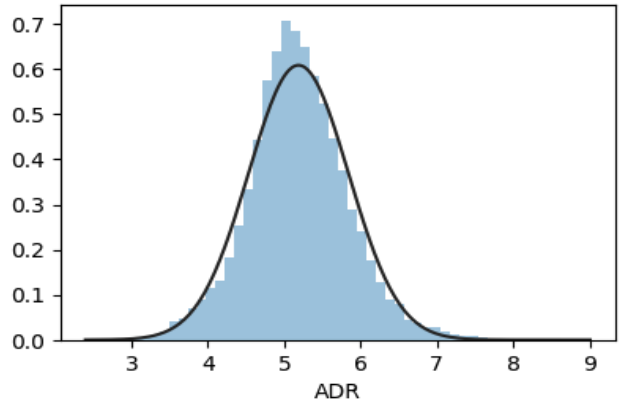
✓ Rating 변수들의 데이터 편중이 높음



[ADR 로그 변환 전]



[ADR 로그 변환 후]



4. 연구내용 및 결과

1-4) 예약률 예측에 활용할 설명변수 선정 과정

- OLS 검정을 통한 다중회귀 분석
- 설명변수 간 다중공선성 진단을 위한 VIF 지수 확인

✓ 예약률에 대한 설명변수의 설명력(R-suqared): 약 46%

OLS Regression Results

Dep. Variable:	occupancy	**R-squared:**	0.455
Model:	OLS	**Adj. R-squared:**	0.454
Method:	Least Squares	**F-statistic:**	1636.
Date:	Fri, 08 Dec 2023	**Prob (F-statistic):**	0.00

✓ ministay, 일부 rating의 p-value 값이 0.5를 넘어 귀무가설 수용

	coef	std err	t	P>|t|	[0.025	0.975]
ministay	0.0003	0.001	0.331	0.741	-0.002	0.002
photonum	0.0116	0.002	6.988	0.000	0.008	0.015
overall	0.0305	0.015	2.070	0.038	0.002	0.059
commrating	-0.0032	0.012	-0.258	0.797	-0.028	0.021
accuracyrating	0.0139	0.013	1.033	0.302	-0.012	0.040
cleanrating	-0.0201	0.011	-1.869	0.062	-0.041	0.001
checkrating	-0.0359	0.012	-2.902	0.004	-0.060	-0.012
locaterating	0.0143	0.014	1.054	0.292	-0.012	0.041
valuerating	0.0189	0.013	1.421	0.155	-0.007	0.045

✓ VIF 지수 10 이상 변수

- Bathrooms / guest / Bedrooms / rating 변수

```
#vip(분산팽창요인)을 통한 다중공선성 여부를 확인하는 지표
#지표 기준: 보통 10이 넘어가면 다중공선성이 있다고 판단하지만 다른 방법이랑 같이 확인하는 것이 필요함
#vip가 높은 지표를 한번에 지우지 말고 하나씩 지우면서 확인하는게 중요함
from statsmodels.stats.outliers_influence import variance_inflation_factor

vif = pd.DataFrame()
vif["vif factor"] = [variance_inflation_factor(
    df.values, i) for i in range(df.shape[1])]
vif["features"] = df.columns
vif = vif.sort_values("vif factor").reset_index(drop=True)
vif
```

	vif factor	features
12	18.449363	Bathrooms
13	21.047135	guest
14	21.058100	Bedrooms
15	31.975810	response
16	242.287160	cleanrating
17	269.601616	locaterating
18	328.446057	checkrating
19	334.254456	valuerating
20	357.938492	commrating
21	404.047496	accuracyrating
22	577.740770	overall

' 예측 성능을 저하시킬 수 있는 요인으로 판단됨 '

OLS 회귀분석, VIF 지수를 활용해 단순히 변수들을 삭제하지 않고

예약률 예측에 의미있다고 생각하는 설명변수

최종 17개 선정

4. 연구내용 및 결과

2-1) Sparse PCA를 통한 차원축소

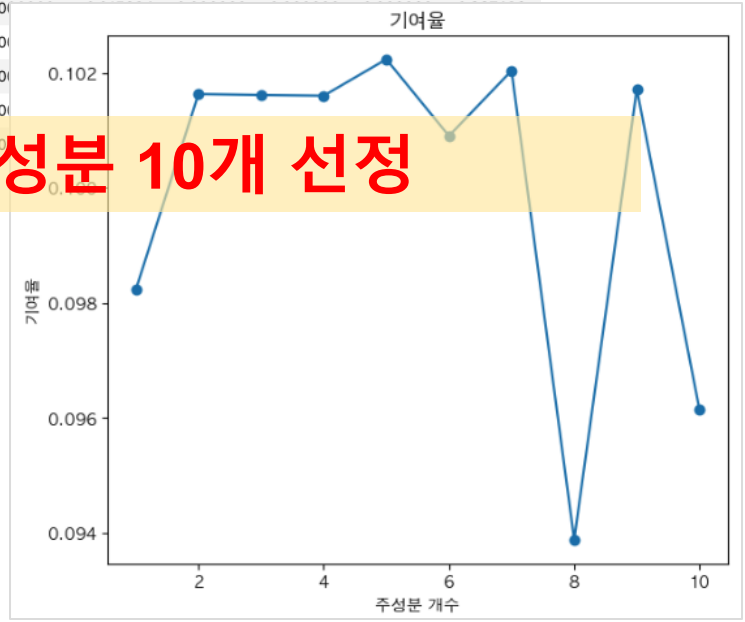
주성분 개수 별 기여율 및 누적기여율 확인

✓주성분 개수 별 가중치가 낮은 변수들은 0으로 수렴

✓데이터의 설명력을 높이기 위해 누적기여율 70-90% 구간의 주성분 개수 선정

	ADR	ARL	booking	reviewnum	Bathrooms	guest	response	superhost	deposit
0	-0.040017	-0.965205	-0.258405	-0.000000	-0.000000	-0.000000	-0.000000	-0.000000	-0.000000
1	0.605688	-0.000000	-0.000000	-0.302553	0.000223	-0.000000	-0.000000	-0.000000	-0.000000
2	0.006400	0.000000	0.000000	0.000000	0.011252	0.000000	0.000000	0.000000	0.000000
3	-0.000000	-0.000000	-0.000000	-0.999342	0.022021	-0.000000	-0.000000	-0.000000	-0.000000
4	0.012848	-0.000000	-0.000000	-0.000000	-0.000000	-0.000000	-0.999917	-0.000000	-0.000000
5	0.069736	0.000000	0.000000	0.0					
6	0.002727	0.000000	0.000000	0.0					
7	0.143485	0.000000	0.000000	0.0					
8	-0.062193	-0.000000	-0.997766	-0.0					
9	0.016688	-0.000000	-0.000000	0.0					

주성분 10개 선정



주성분 구성 변수 파악 및 10 차원 축소

```
주성분 8의 가중치가 큰 상위 변수 10개:
Index(['guest', 'Bathrooms', 'ADR', 'cleanrating', 'valuerating', 'photonum',
       'ministay', 'nightfee', 'locaterating', 'superhost'],
      dtype='object')

주성분 9의 가중치가 큰 상위 변수 10개:
Index(['booking', 'ADR', 'nightfee', 'Bathrooms', 'superhost', 'ARL',
       'reviewnum', 'guest', 'response', 'valuerating'],
      dtype='object')

주성분 10의 가중치가 큰 상위 변수 10개:
Index(['superhost', 'overall', 'cleanrating', 'valuerating', 'accuracyrating',
       'nightfee', 'locaterating', 'ADR', 'ministay', 'photonum'],
      dtype='object')
```

	spca1	spca2	spca3	spca4	spca5	spca6	spca7	spca8	spca9	spca10
0	-0.630625	-1.949110	-0.422810	1.092152	-0.130665	0.016239	0.215654	-0.384828	-0.476040	-0.437673
1	-0.576608	0.793897	0.550055	1.772874	-0.134954	0.505041	0.149226	0.812657	0.469913	0.214842
2	1.347145	1.130965	1.505871	-0.502430	-0.132222	0.517780	-0.069903	0.086839	1.519658	0.223406
3	0.047164	-0.333355	-0.134507	0.534735	-0.129117	-0.378056	-0.534126	-0.806092	-0.097042	-0.441651
4	-0.440512	-0.502138	-0.135944	0.624535	-0.130458	0.016568	0.552944	0.904038	0.059379	0.263383
...	...	...	...	...	...	...	...	...	...	...
43214	1.258832	0.419764	-0.421287	-0.776624	1.207823	-0.689454	0.408509	-0.825436	-0.399844	0.202766
43215	0.743428	-0.077754	-0.422841	-0.645333	-0.131363	-1.070040	0.083178	0.362535	0.077922	0.253665
43216	-1.139807	-1.165035	-0.141914	-0.134566	-0.075099	-0.005649	-0.113803	-0.245425	-0.702781	0.247832
43217	-0.682788	0.480011	-0.139782	0.661879	-0.133452	0.004640	-0.355409	-0.422099	0.016405	-0.446949
43218	-1.054518	0.176086	-0.138150	0.092164	-0.133564	0.508772	0.008177	-0.041149	-0.245964	-0.447846

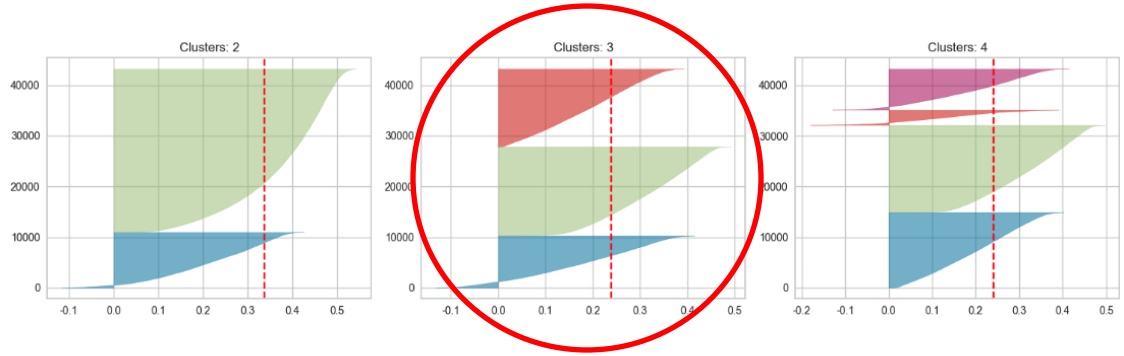
4. 연구내용 및 결과

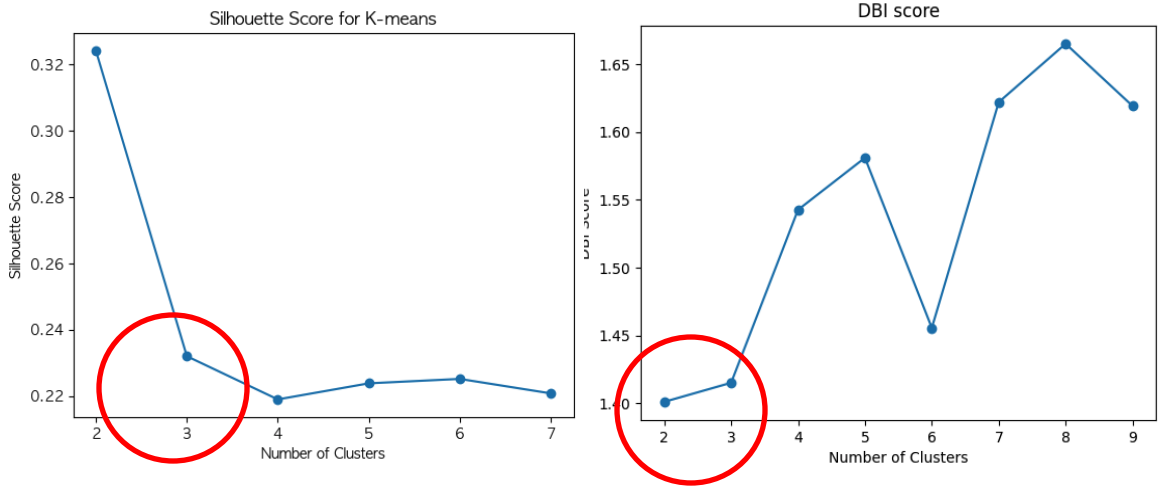
3-1) 주성분 데이터 기반 K-means 클러스터링

실루엣 계수, DBI 지표를 활용한 군집 수 선정

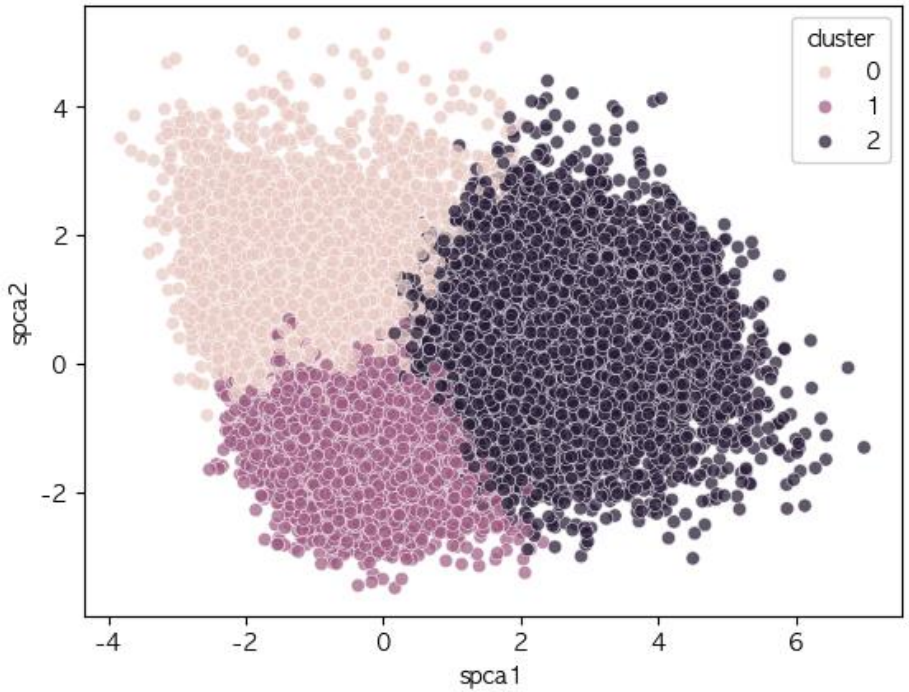
✓실루엣 계수: 군집 수 2~4의 **평균 20~35% 사이**

✓DBI 지표: 낮은 score 기준으로 cluster 3이 가장 적절하다고 판단



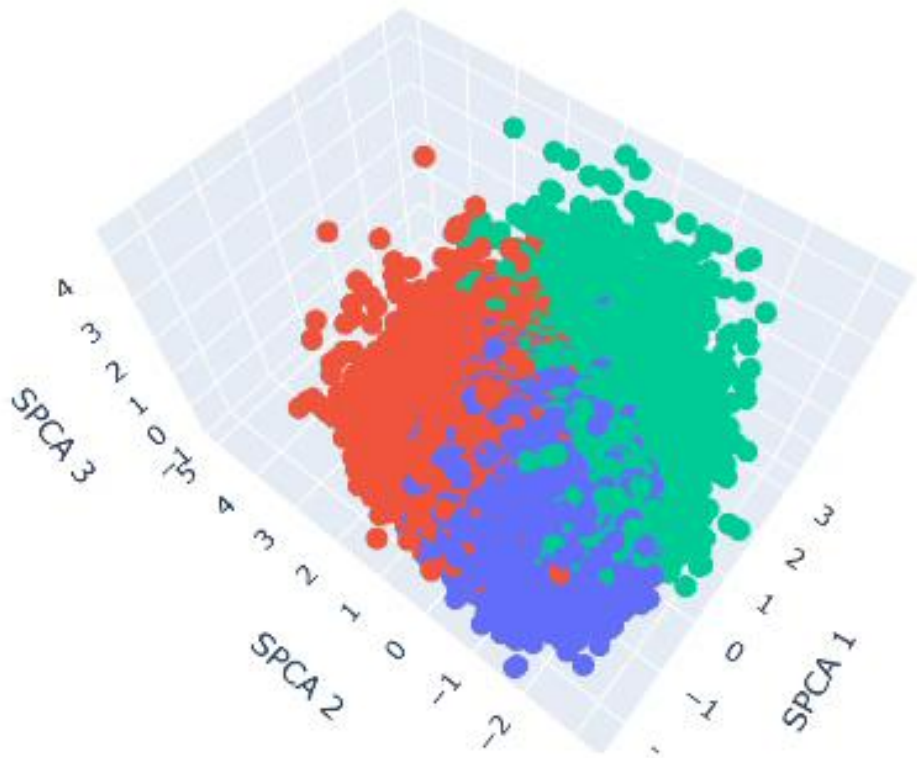


[3개의 군집화 결과]



4. 연구내용 및 결과

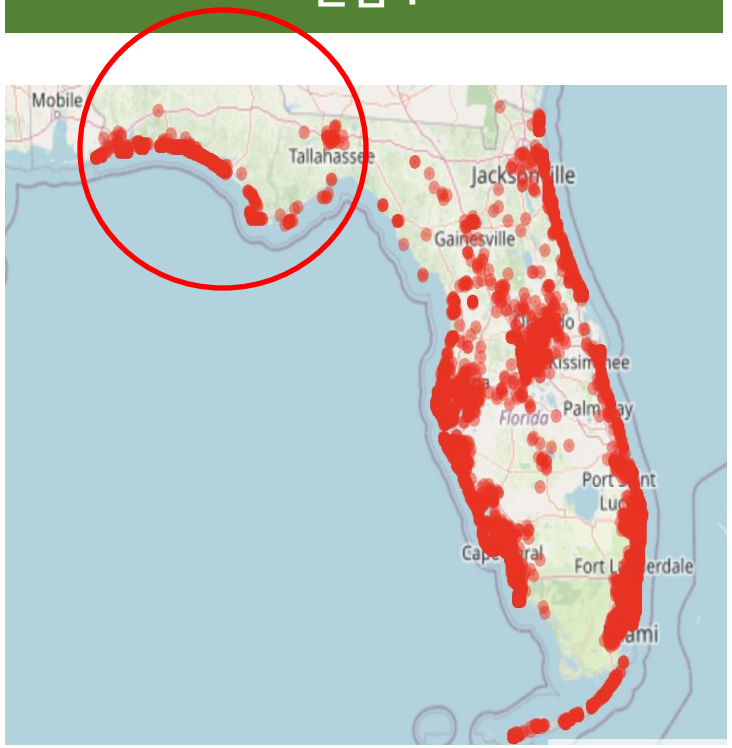
3-2) 클러스터링 3D 산점도 분포

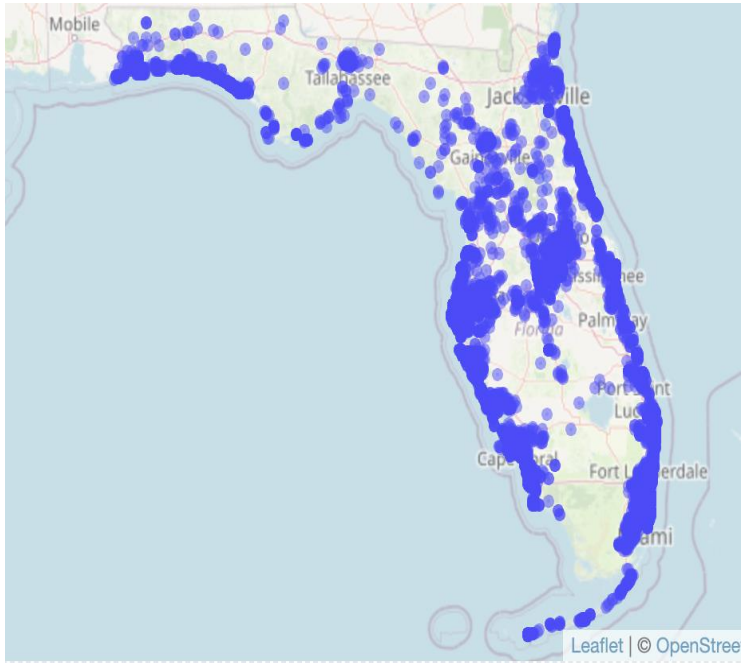


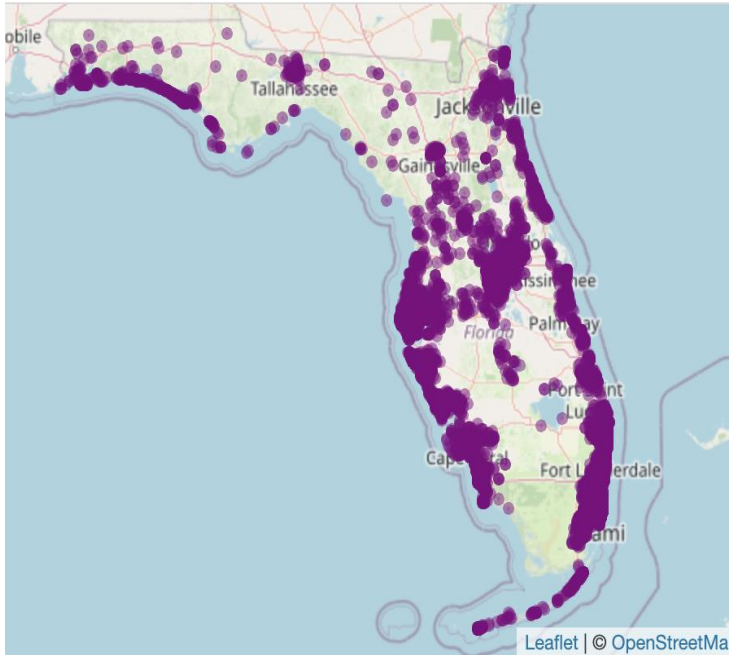
3-3) EDA를 통한 군집별 특징

✓ 위도, 경도로 숙소 분포 지역 확인

군집 1	군집 2	군집 3







'전체적으로 숙소들이 해변가에 몰려있고 지리적 위치는 비슷하나
cluster 1의 숙소들이 cluster 2,3보다 각 위치에 집중되어 있는 편 '

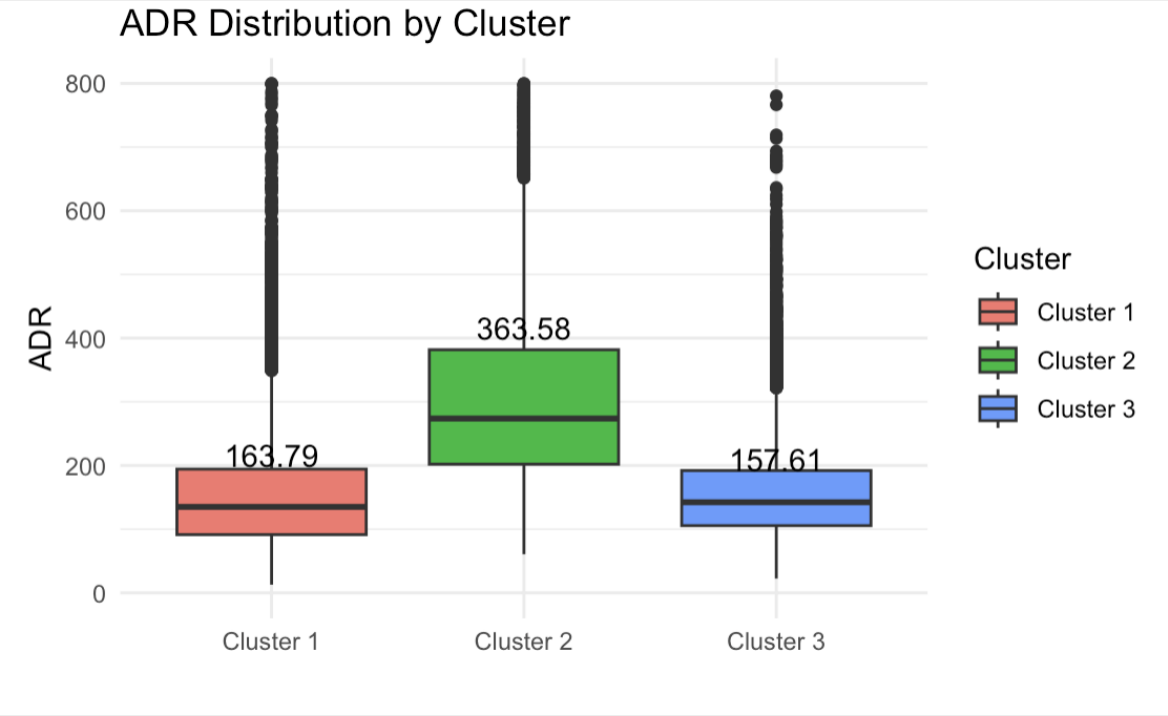
4. 연구내용 및 결과

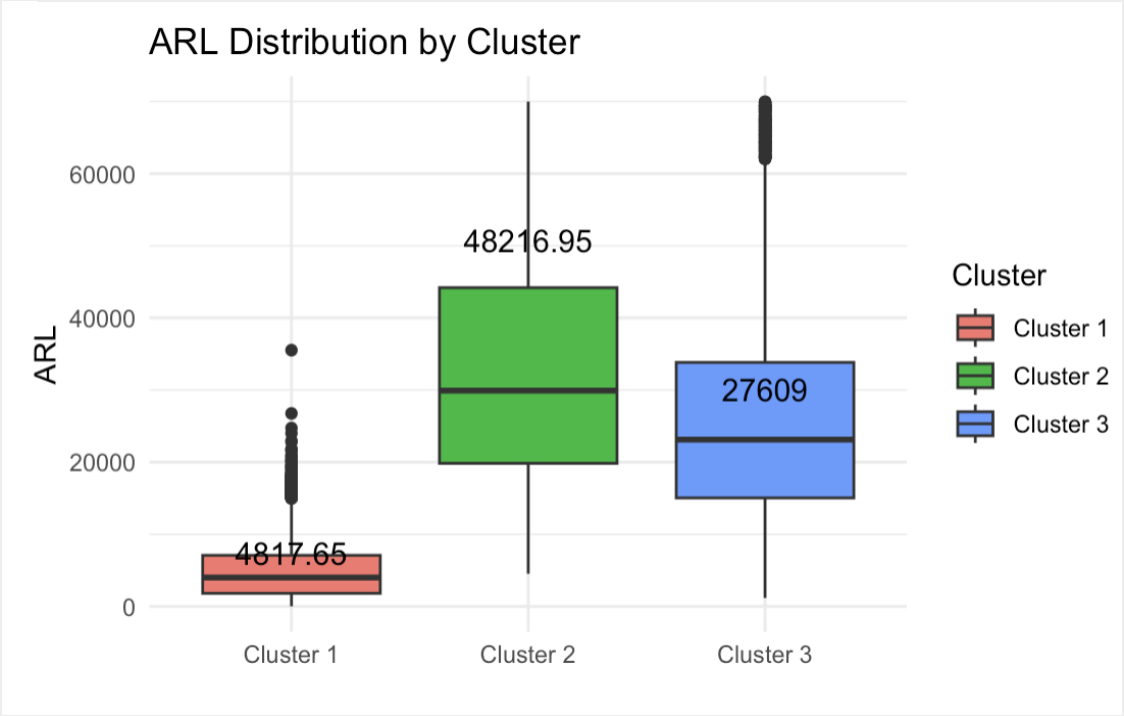
3-3) EDA를 통한 군집별 특징

✓ ADR(평균 1박 가격) 및 ARL(평균 연간 수익) 평균 비교

₩ 21만	₩ 47만	₩ 20만

₩ 628만	₩ 6287만	₩ 3600만





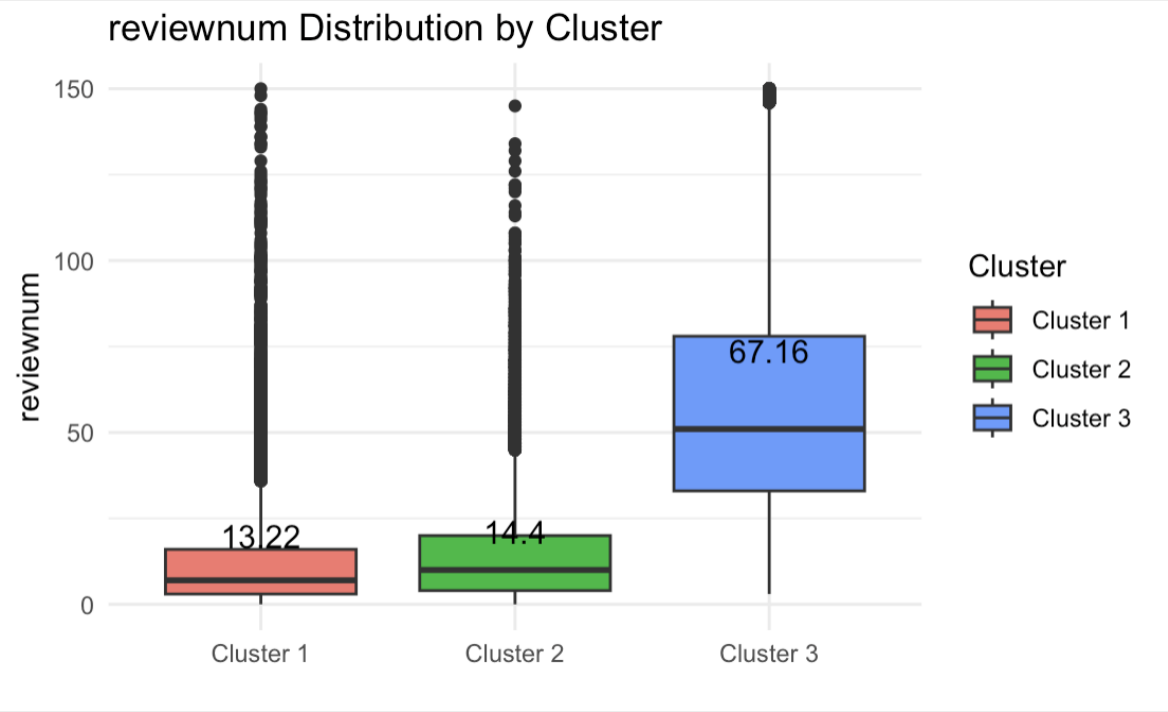
4. 연구내용 및 결과

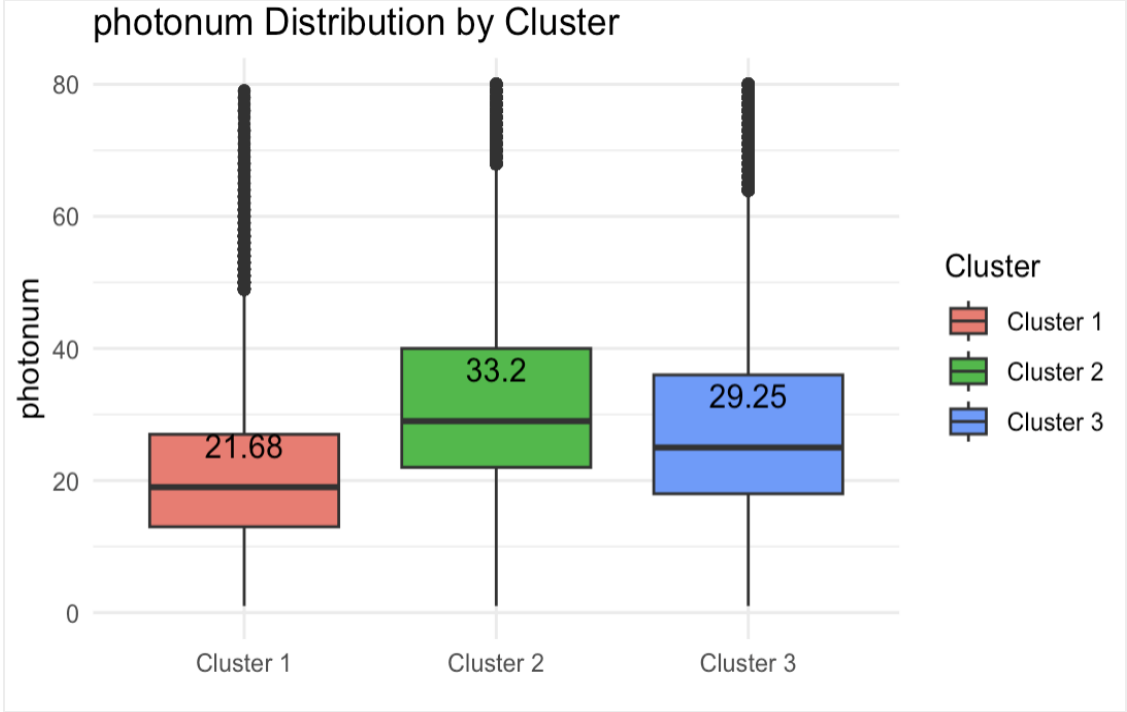
3-3) EDA를 통한 군집별 특징

✓ 숙소 리뷰 수 및 호스트가 등록하는 숙소 사진 수 평균 비교

13개	14개	67개

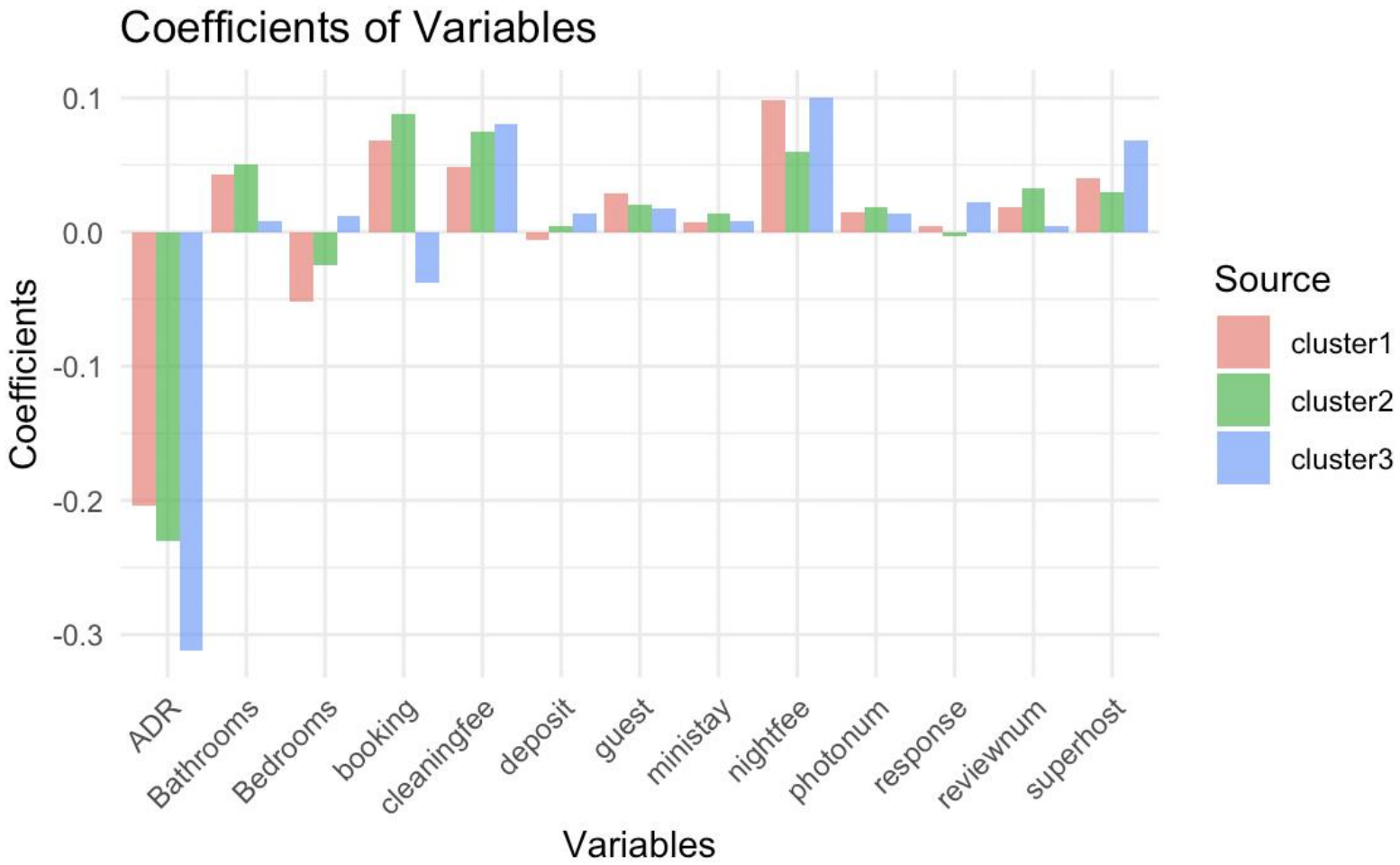
22개	33개	29개





3-3) EDA를 통한 군집별 특징

✓ 군집별 예약률과 각 변수 간의 상관관계



'3개의 군집 모두 평균 1박 가격이 저렴하면 예약률이 높아지는 경향'

'고객들은 침실보다는 화장실의 갯수나 청결도를 좀 더 고려하는 편'

4. 연구내용 및 결과

4-1) 군집별 예약률 예측 모형

- ✓ 군집별 예약률 예측 성능 평가 지표: RMSER, R-squared
- ✓ 군집 1의 설명력이 약 67%로 가장 높은 예측력을 보임

예측 과정

Train / Test data 분리: 70 : 30

최적의 Hyper parameter 설정(비선형 모델)

	군집 1	군집 2	군집 3
n_estimaters	800	1000	1000
min_sample_split	2	2	2
min_sample_leaf	2	2	2
max_features	Auto	Auto	Auto
max_depth	50	50	50

모델	군집	RMSE	R-squared
Multiple Regression	1	0.211	0.505
	2	0.138	0.416
	3	0.116	0.458
Random Forest (Regression)	1	0.171	0.674
	2	0.115	0.593
	3	0.100	0.595
Support Vector Regression	1	0.214	0.488
	2	0.139	0.409
	3	0.117	0.454
Gradient Boosting Regression	1	0.179	0.643
	2	0.119	0.564
	3	0.104	0.562

최종 모형 선택

4. 연구내용 및 결과

4-1) 군집별 예약률 예측 모형

✓ Test data 기준 **실제 예약률 – 예측 예약률** 비교

군집 1

차이

```
=== cluster0 ===
       real_occu  pred_occu  diff
35855       0.24       0.34  0.10
39077       0.64       0.63  0.01
28090       0.31       0.60  0.29
41216       0.69       0.58  0.11
33158       0.33       0.34  0.01
...          ...        ...   ...
23469       0.18       0.29  0.11
37743       0.48       0.44  0.04
13430       0.56       0.48  0.08
29809       1.00       0.37  0.63
43080       0.80       0.51  0.29

[3338 rows x 3 columns]
```

군집 2

차이

```
=== cluster1 ===
       real_occu  pred_occu  diff
30969       0.51       0.56  0.05
12659       0.63       0.68  0.05
40422       0.57       0.67  0.10
8269        0.63       0.54  0.09
1941        0.64       0.71  0.07
...          ...        ...   ...
31499       0.60       0.66  0.06
40411       0.63       0.55  0.08
22729       0.32       0.43  0.11
2045        0.42       0.51  0.09
35603       0.28       0.34  0.06

[4680 rows x 3 columns]
```

군집 3

차이

```
=== cluster2 ===
       real_occu  pred_occu  diff
24093       0.25       0.36  0.11
25688       0.27       0.36  0.09
27234       0.82       0.74  0.08
26304       0.47       0.49  0.02
4155        0.76       0.74  0.02
...          ...        ...   ...
5986        0.76       0.78  0.02
31345       0.81       0.71  0.10
18977       0.55       0.50  0.05
24162       0.82       0.77  0.05
17895       0.76       0.76  0.00

[4948 rows x 3 columns]
```

4-2) SHAP value를 통한 변수 기여도 분석

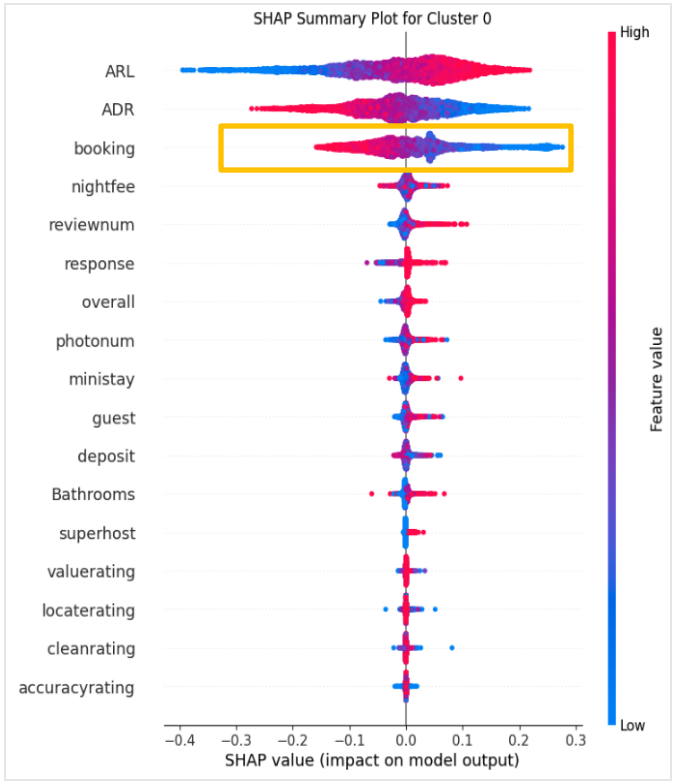
✓ 예약률은 공통적으로 ADR, ARL과 같은 가격적 요소에 가장 영향을 많이 받음

군집 1

✓ 실제 예약률 평균: 0.489

✓ 변수 영향도: 리뷰 수, 응답시간 등

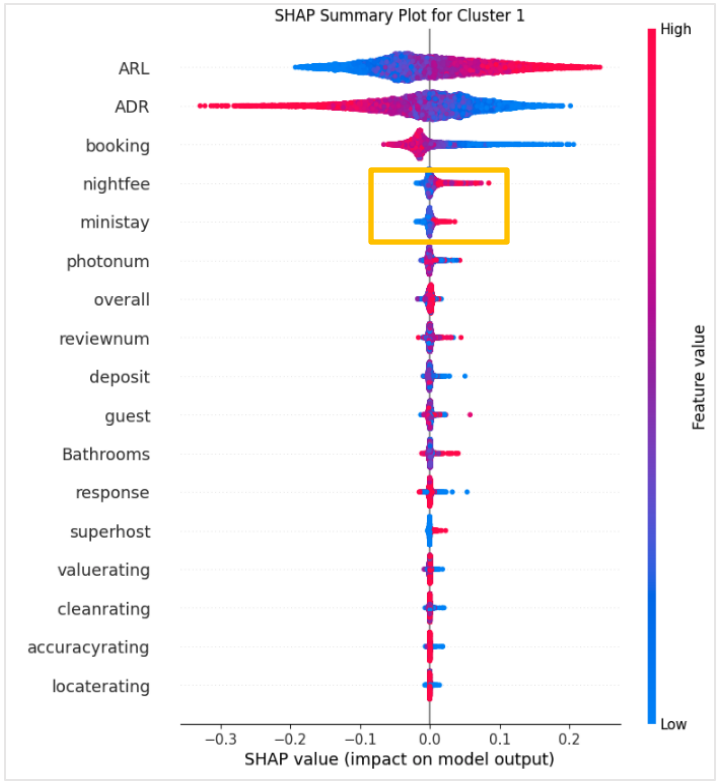
숙소 평가 위주



군집 2

✓ 실제 예약률 평균: 0.549

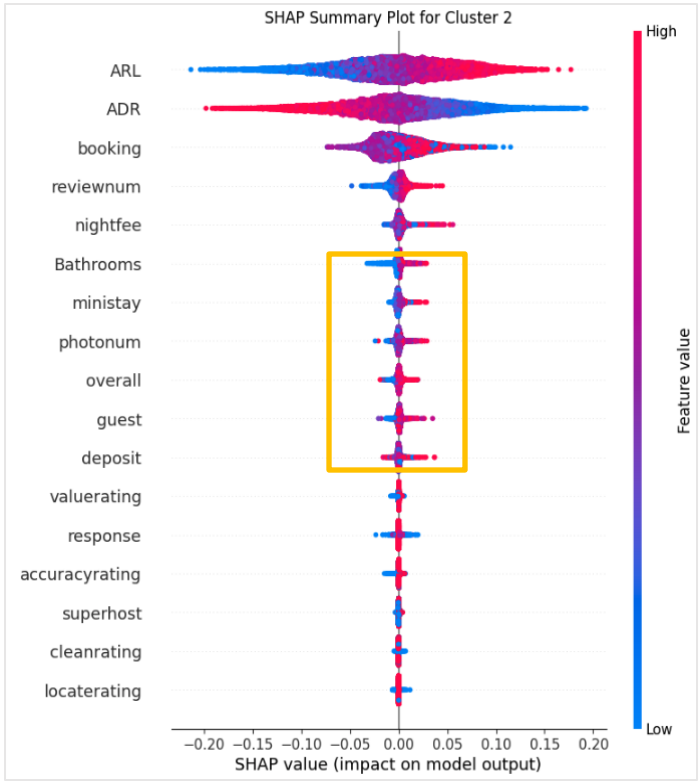
✓ 변수 영향도: 1박 가격, 최소 숙박일수, 보증금 등 경제적 요소 위주



군집 3

✓ 실제 예약률 평균: 0.646

✓ 변수 영향도: 리뷰 수, 게스트 수, 화장실 수 등 전반적인 숙소 상태 위주

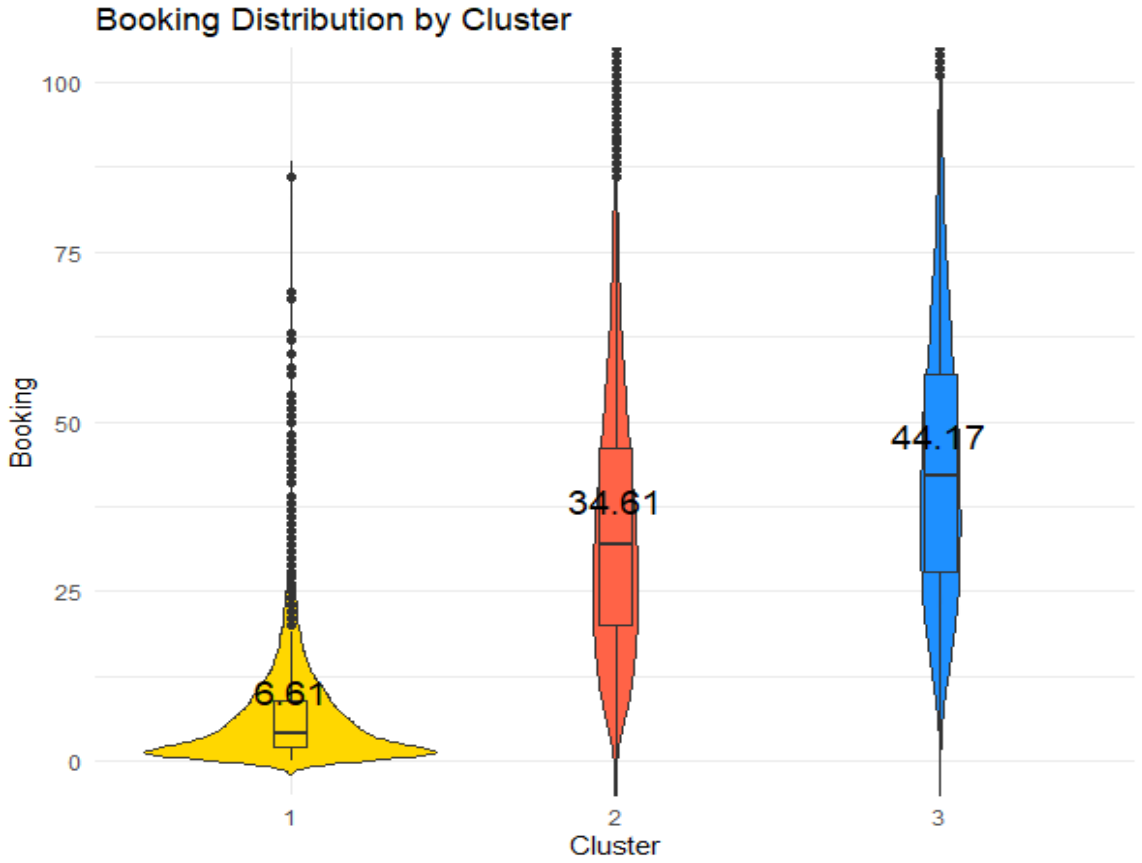


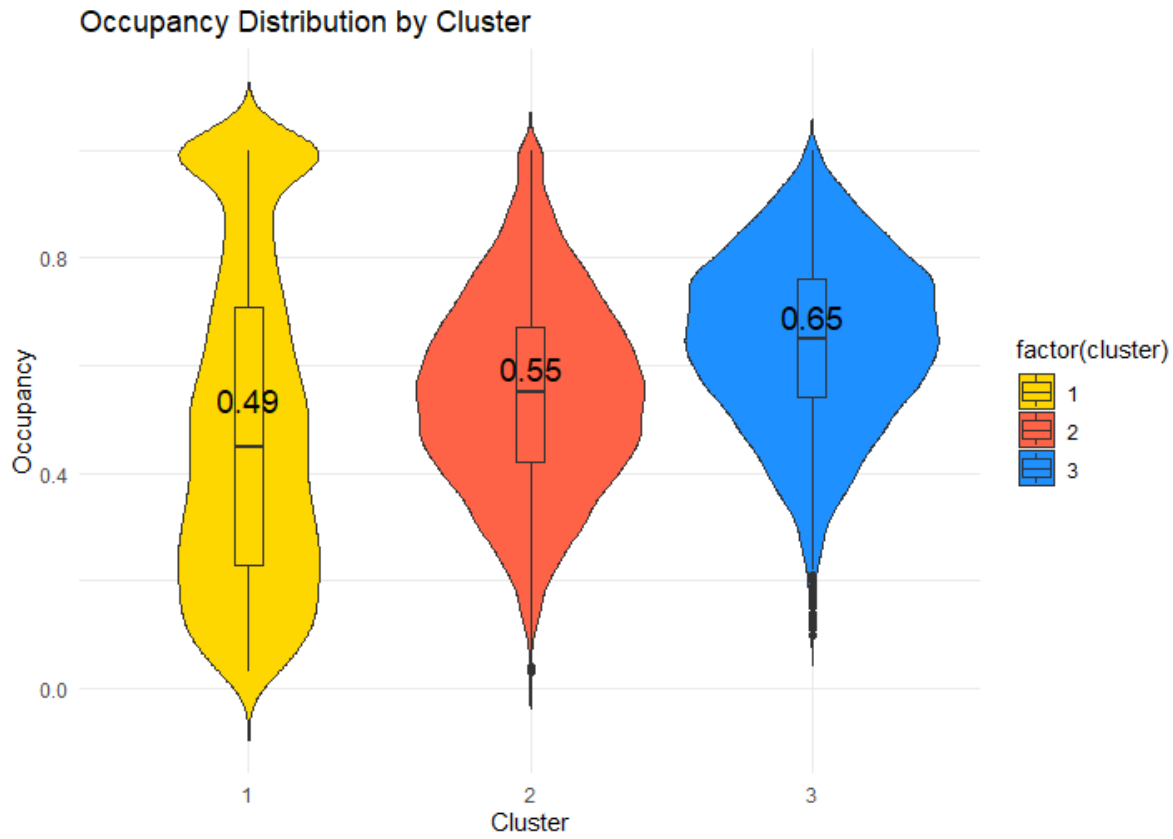
4. 연구내용 및 결과

4-3) 군집별 예약률(occupancy)과 예약일 수(booking) 관계

- ✓ 군집 1 예약일 수 평균 : 6.61
- ✓ 군집 2 예약일 수 평균 : 34.6
- ✓ 군집 3 예약일 수 평균 : 44.2

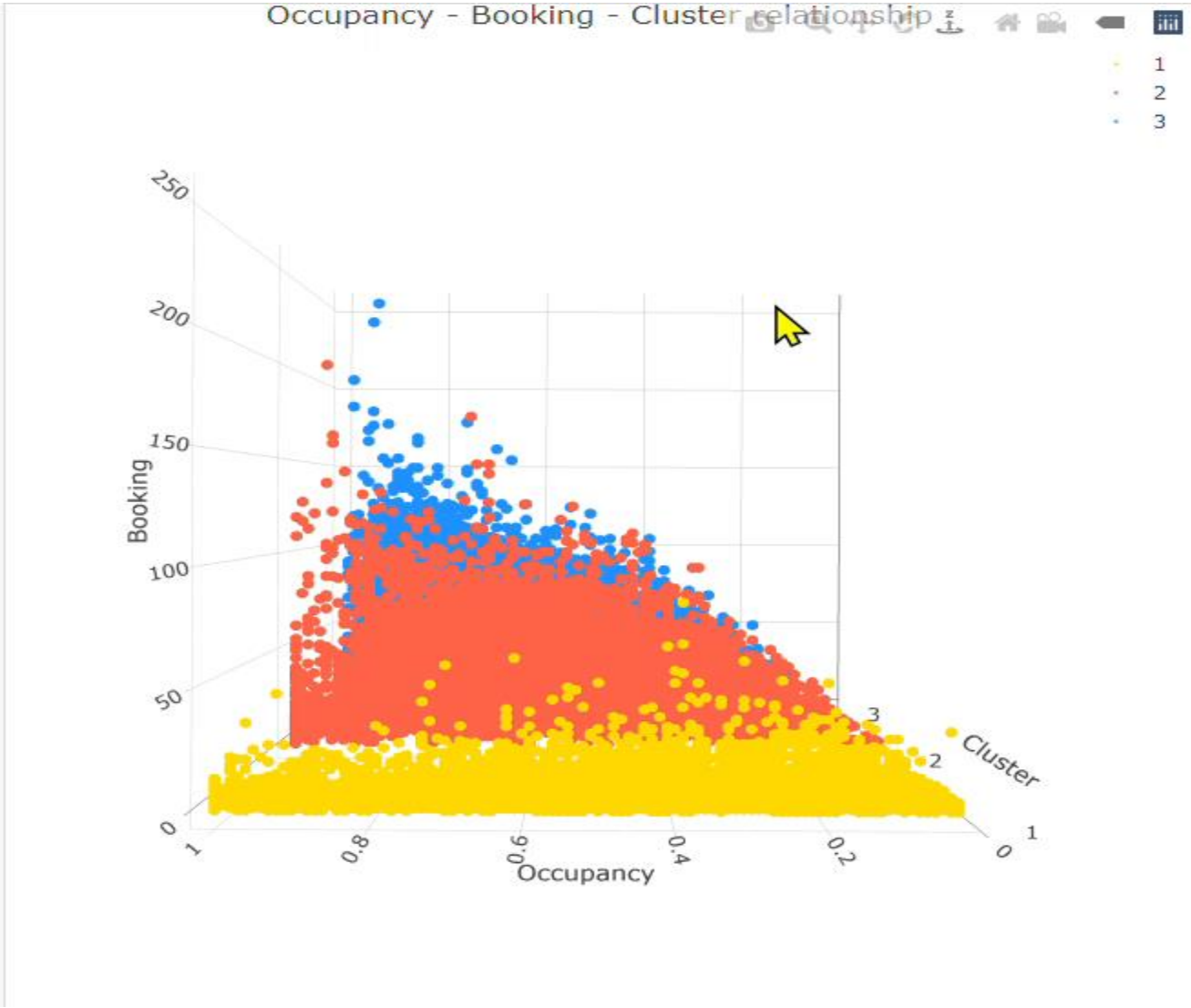
- ✓ 군집 1 예약률 평균 : 0.489
- ✓ 군집 2 예약률 평균 : 0.549
- ✓ 군집 3 예약률 평균 : 0.646





4. 연구내용 및 결과

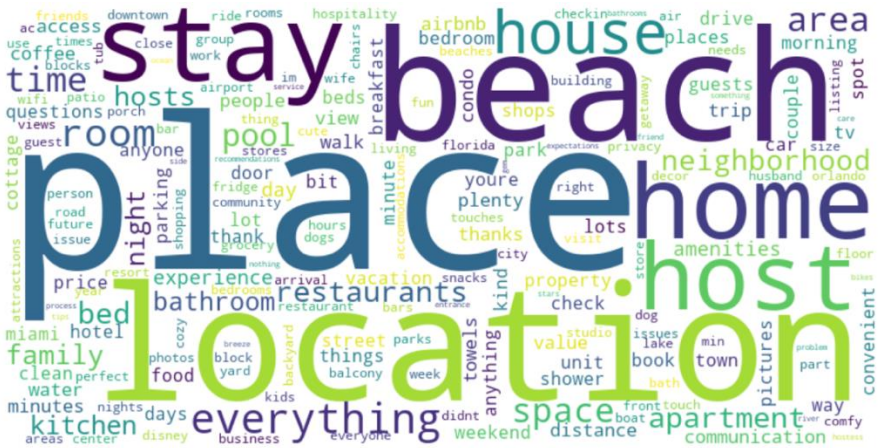
4-3) 군집별 예약률(occupancy)과 예약일 수(booking) 관계



4. 연구내용 및 결과

4-3) 텍스트 감성분석 데이터 기반 Wordcloud 분석

[긍정 리뷰 명사 및 형용사]



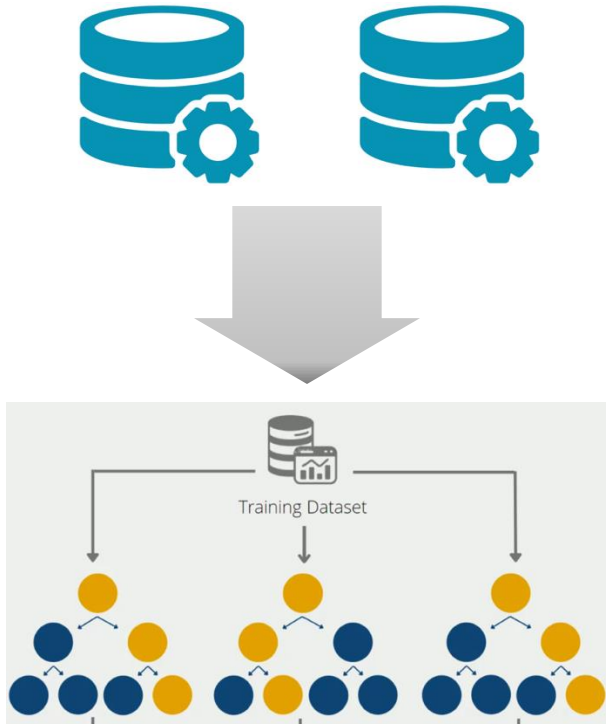
[부정 리뷰 명사 및 형용사]

실제 숙박 이용 후, 고객은 ADR, deposit 등 정량적인 요소보다는 place(위치), host(호스트 친절도), 숙소 청결도, Amenity 관리 등 정성적인 요소에 더 민감하게 반응함 '

5. 결론

5-1) 서비스 예상 시나리오

1) 호스트 숙소 정보 셋팅 → 2) 랜덤 포레스트 알고리즘 동작 → 3) 예약률 예측을 통한 대시보드 안내 > 호스트 숙소 정보 개선



등록하신 숙소 정보 기준
현재 예약 가능률은 25% 입니다.
1박 요금을 조금 낮게 책정해도 좋을 것 같아요.

5. 결론

5-2) 연구 결과

- 데이터 복잡성을 줄이고, 예측 성능을 높이기 위해 SPCA 차원축소. K-means 활용 3개 군집으로 분류
- 선형/비선형 모델을 적용하여 RMSE, R계수를 기준으로 예측력을 비교한 결과, 예약률 예측에 적합한 최종 모형은 **'Random Forest'로 결정**
- SHAP value를 통해 '평균 1박 가격이 저렴하면 예약률이 높아진다'는 연구가설 입증
- SHAP value를 통해 숙소별 예약률 예측에 기여한 변수들을 기반으로 **군집별 Host가 숙소 정보 등록 및 운영 시점에 지속적으로 관리해야 할 요소들을 제안**한다면 효과적인 호스팅 비즈니스 운영이 가능할 것으로 판단

숙소별 Shap 가중치 높은 변수 출력

```
=== cluster0 ===
Row 0 - Top Features: ['ARL', 'ADR', 'listype', 'reviewnum', 'booking', 'nightfee', 'l
ms', 'photonum', 'deposit', 'response', 'cleaningfee']
Row 1 - Top Features: ['ARL', 'ADR', 'listype', 'booking', 'reviewnum', 'nightfee', '
ooms', 'guest', 'ministay', 'response', 'cleaningfee']
Row 2 - Top Features: ['ARL', 'ADR', 'booking', 'reviewnum', 'listype', 'ministay', '
oms', 'superhost', 'Bedrooms', 'cleaningfee', 'guest']
Row 3 - Top Features: ['ARL', 'reviewnum', 'ADR', 'cleaningfee', 'listype', 'photonum
ing', 'superhost', 'Bedrooms', 'nightfee', 'response']
Row 4 - Top Features: ['ARL', 'booking', 'ADR', 'Bathrooms', 'cleaningfee', 'ministay
edrooms', 'listype', 'response', 'superhost', 'guest']
Row 5 - Top Features: ['booking', 'ARL', 'nightfee', 'photonum', 'reviewnum', 'ADR',
esponse', 'guest', 'ministay', 'superhost', 'deposit']
Row 6 - Top Features: ['ARL', 'booking', 'ADR', 'photonum', 'Bedrooms', 'reviewnum',
'ministay', 'listype', 'deposit', 'superhost', 'guest']
```

군집 1 Host 제안 변수

ADR | Nightfee | Photonum | Reviewnum | response

군집 2 Host 제안 변수

ADR | Nightfee | Minimum stay | deposit

군집 3 Host 제안 변수

ADR | Reviewnum | Nightfee | Photonum | guest | deposit

감사합니다