

Chocolate

Chocolate Data Science & Analysis

Minsoo Kwak

Features

Dataset: Kaggle " Chocolate Bar Ratings" (초기 데이터셋: (1795,9)

- Company : 초콜릿 회사
- Bean Origin : bean 재배지
- REF : review를 db에 옮긴 연도
 - + g_count : 해당 company 개수
- Review year : review가 작성된 연도
 - + Bean_num : bean의 개수
- Cacao percent : 카카오가 함유된 퍼센트
 - + Blended : blended 여부 (0 / 1)
- Rating : 평점 (1~5)
- Bean type : bean 종류
- Broad Bean Origin : bean의 원래 원산지

EDA & Feature Engineering

Data Science & Analysis

기본 결측치 처리

\xa0 처리
(non breaking space)

High Cardinarity 고려
묶을 수 있는 부분 대체

Company

416 company

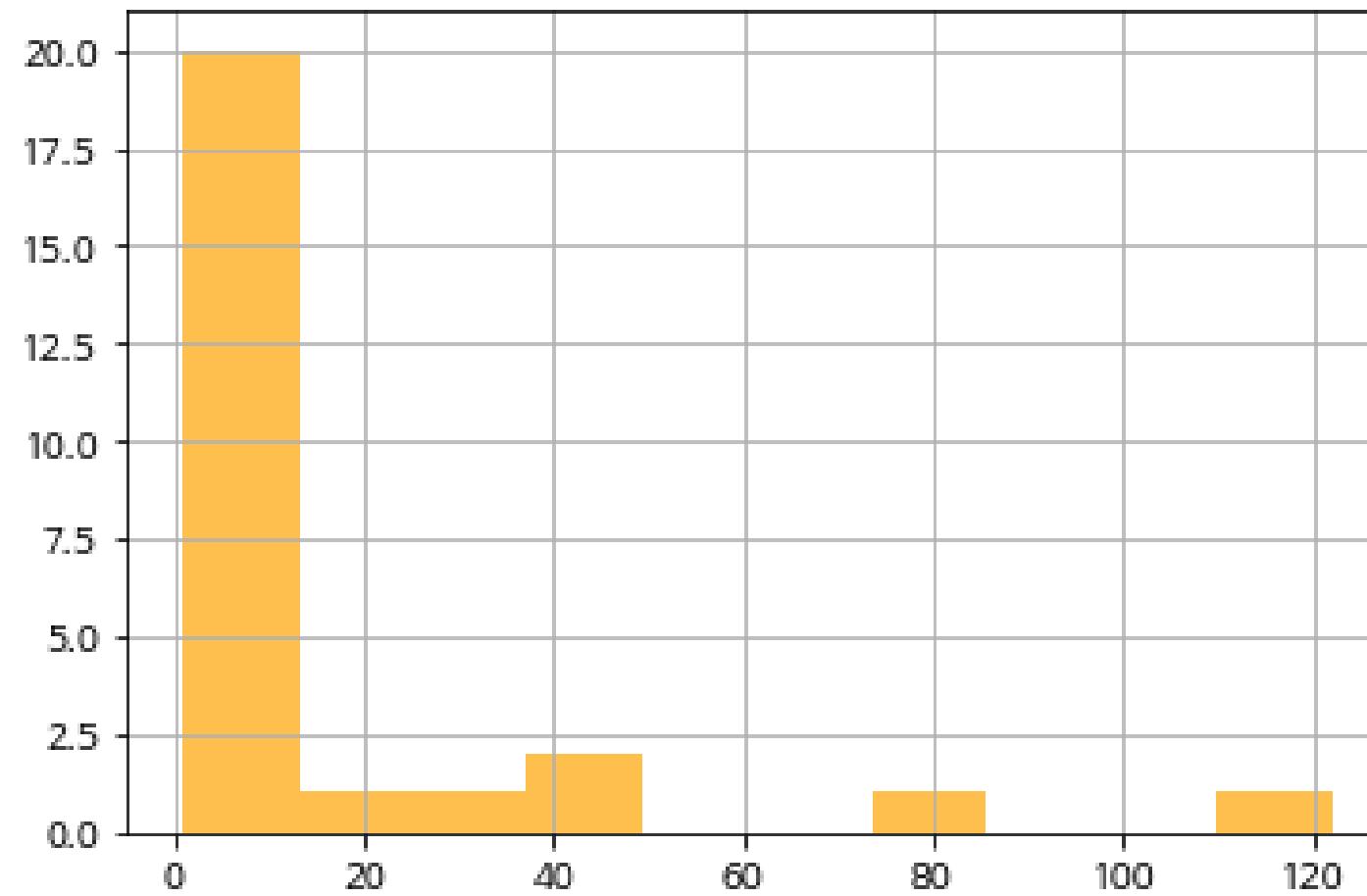
Chocolate Bar Ratings

Company Count 26개 범주

Company Location (60-2 ; nan, ecuador)

	Soma	Bonnat	Fresco	Pralus	A. Morin	Arete	Guittard	Domori	Valrhona	Hotel Chocolat	(Coppeneur)	Coppeneur	Scharffen Berger	Mast Brothers
count	46	27	26	25	23	22	22	22	21	19	18	17	17	
	Lilla	Lindt & Sprungli	Love Bar	Mayacama	Majani	Malagos	Malmö	Manifesto Cacao	Chocablog	Marigold's Finest	Chloé Chocolat	Chequessett	Ki' Xocolatl	
count	1	1	1	1	1	1	1	1	1	1	1	1	1	

등장 빈도 10개 단위로 묶었을 때 (bins=10) 회사 counting



x축 : counting (등장 빈도)

y축 : 회사 수

1개 회사가 가장 많음 (반비례)

416개의 회사는 count에 따라 26개 범주로 구분됩니다. 빈도에 따라 counting을 해보니 1개 회사가 가장 많은 비율을 차지하고, 회사와 빈도가 회사의 수와 반비례하게 나타납니다.

Species

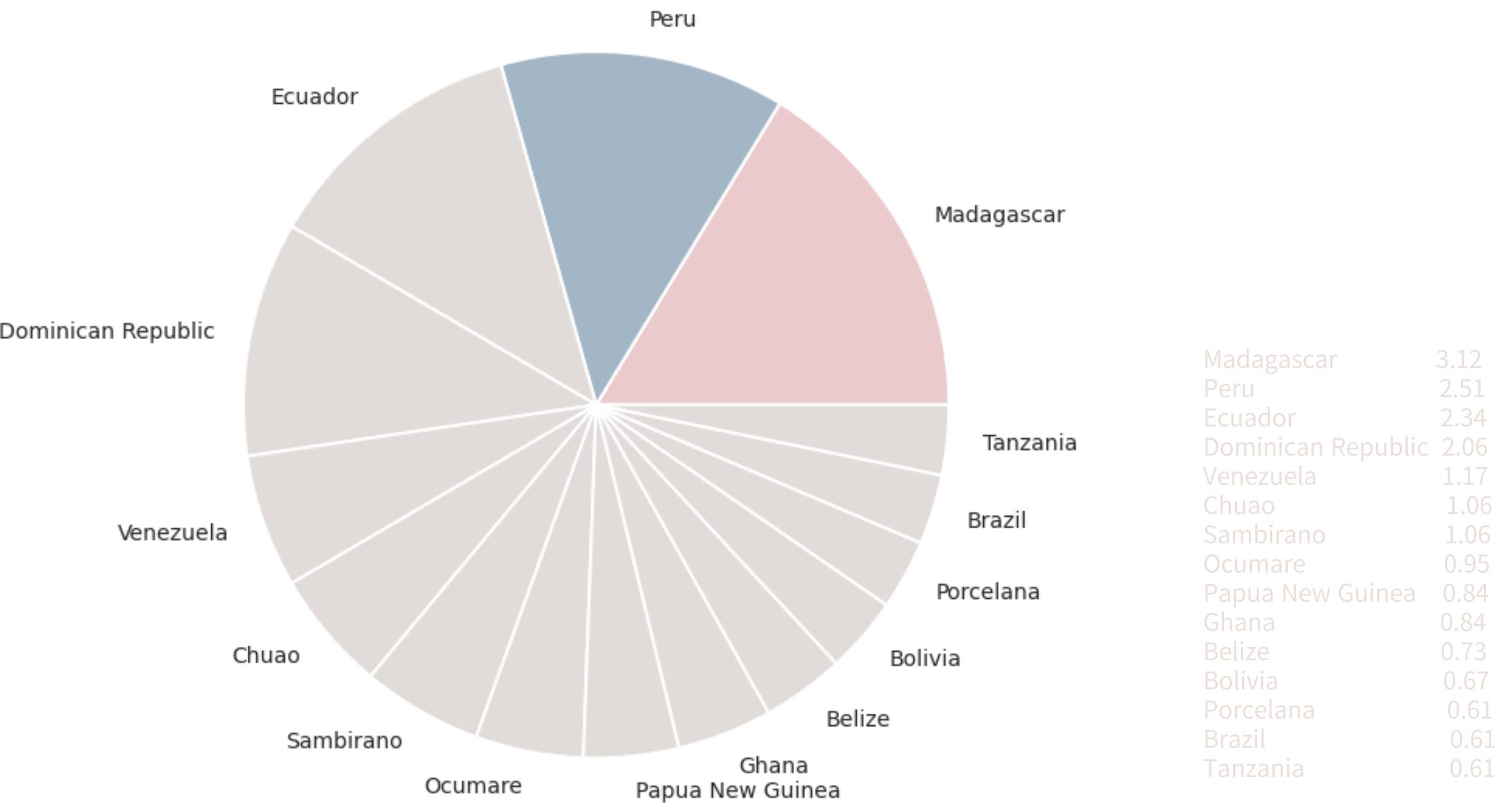
Bean origin (bean 재배지)
(1038)

Chocolate Bar Ratings

Species (재배지) 1038

전체에서 Madagascar가 차지하는 비율 : 5.49%
전체에서 Peru가 차지하는 비율: 4.34%

Madagascar, Peru : 전체의 10% 차지



Bean Type

41-->16

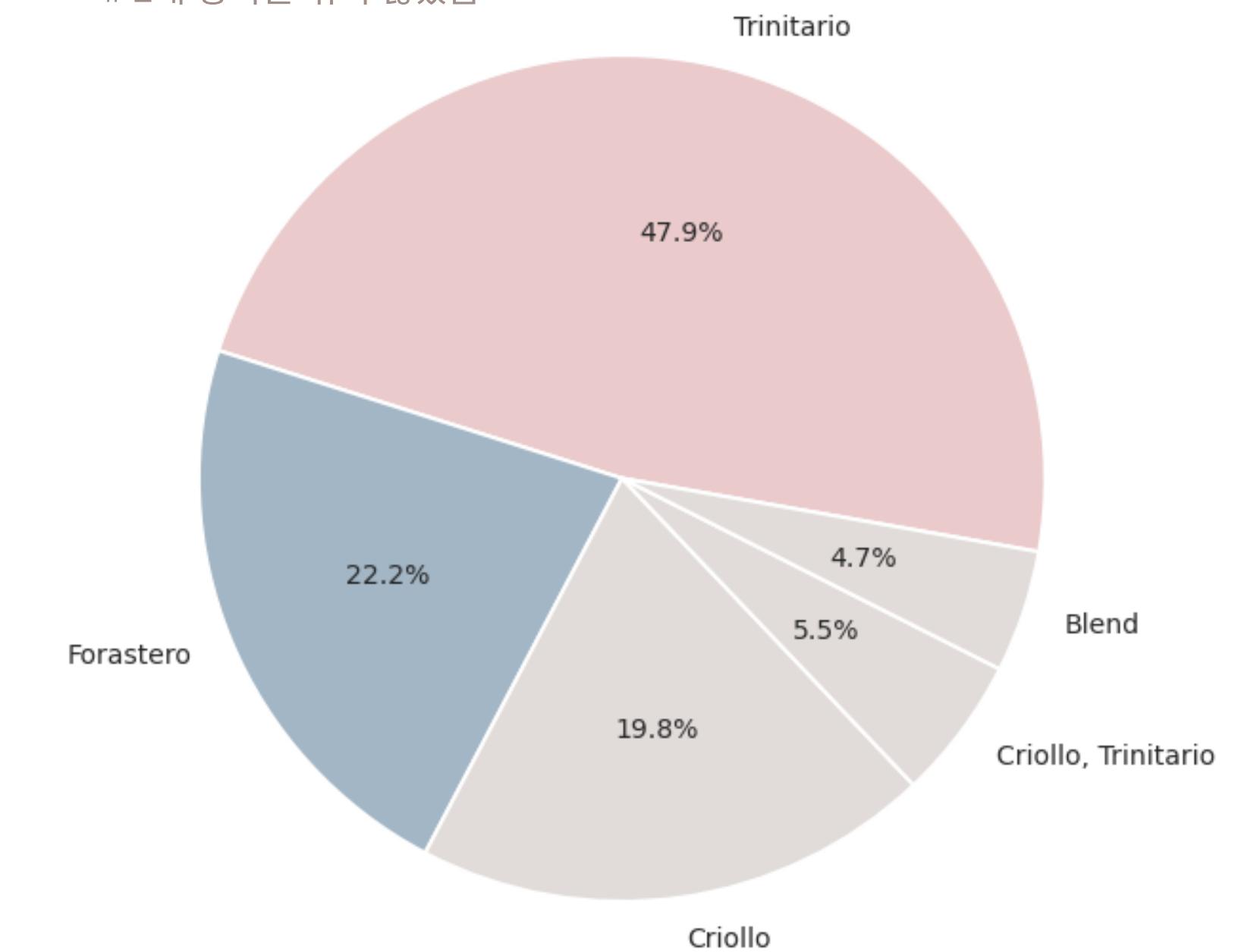
Chocolate Bar Ratings

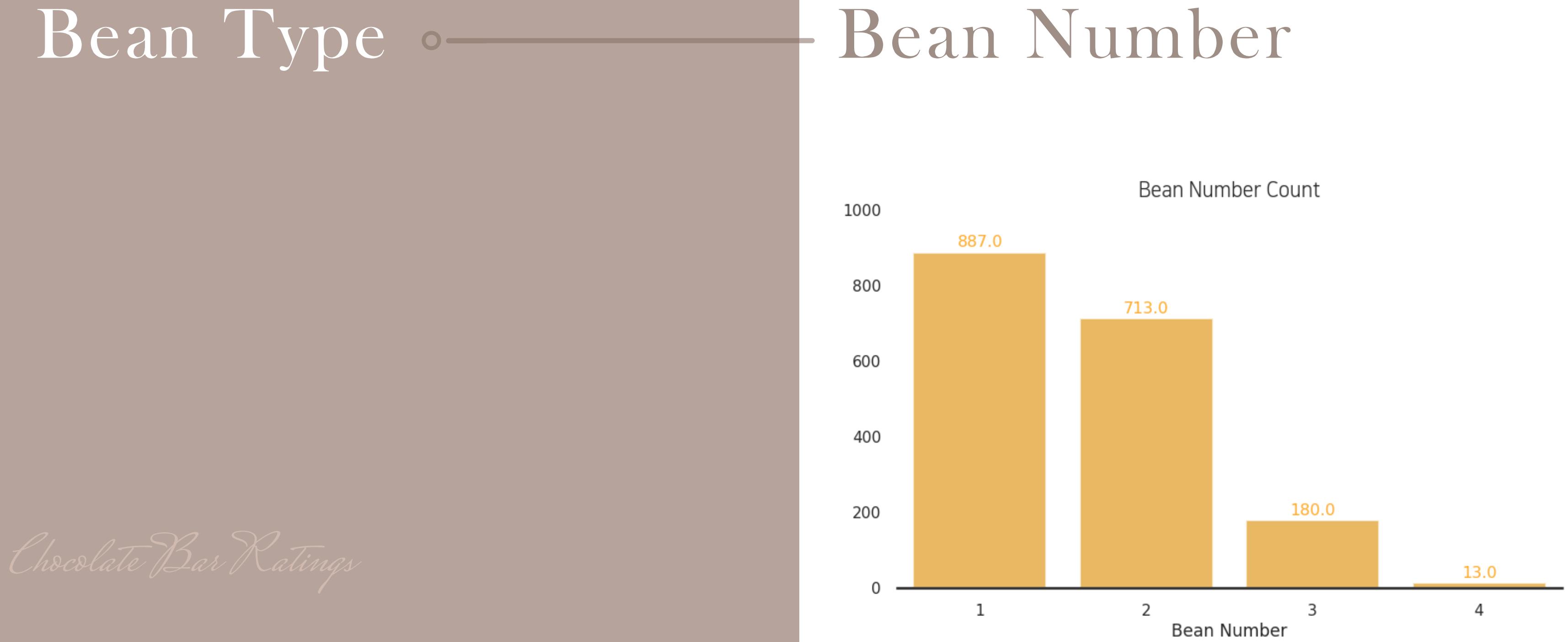
- Trinitario > Forastero > Criollo Bean 순

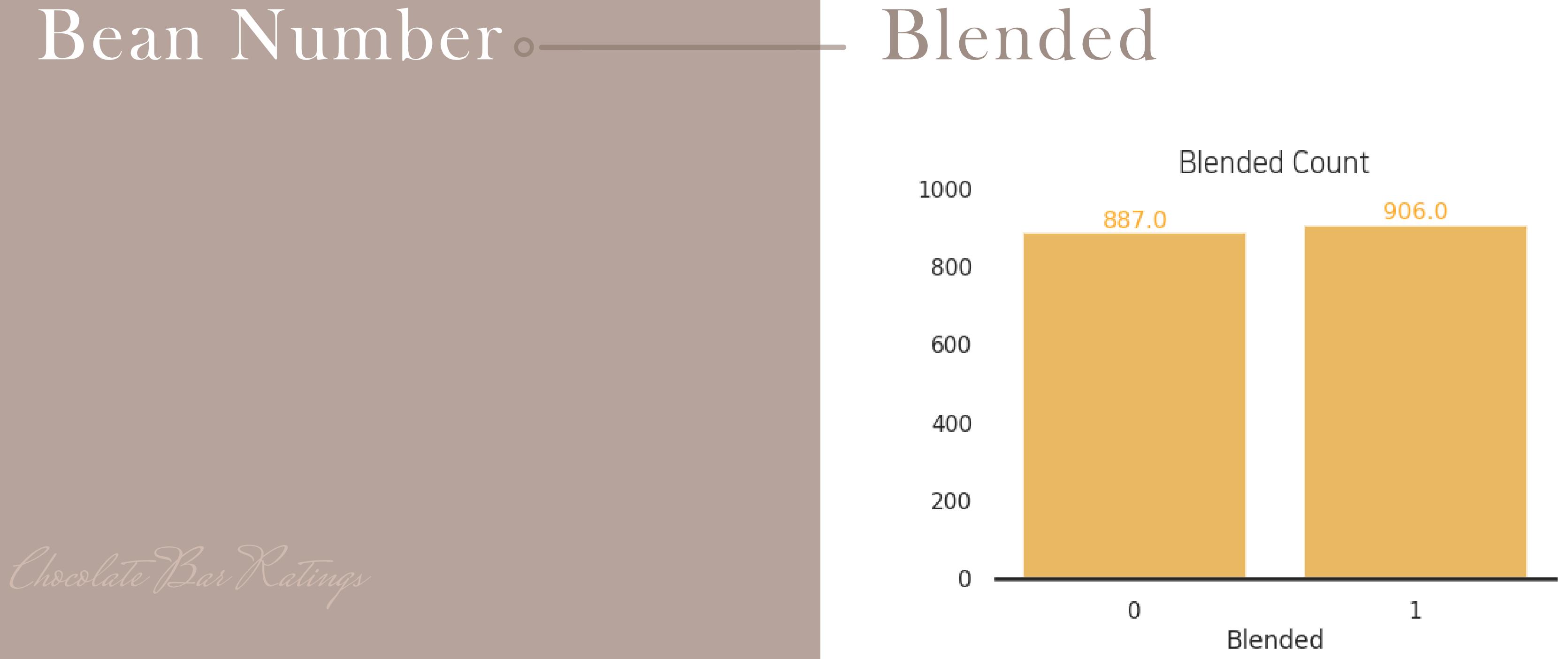
- Trinitario 가 약 48%의 비중 차지

- Forastero와 2배 이상 차이

2개 항목은 묶지 않았음







Numerical Type

- review year 연도에 따라 rating 달라지는가?



Hypothesis

Data Science & Analysis

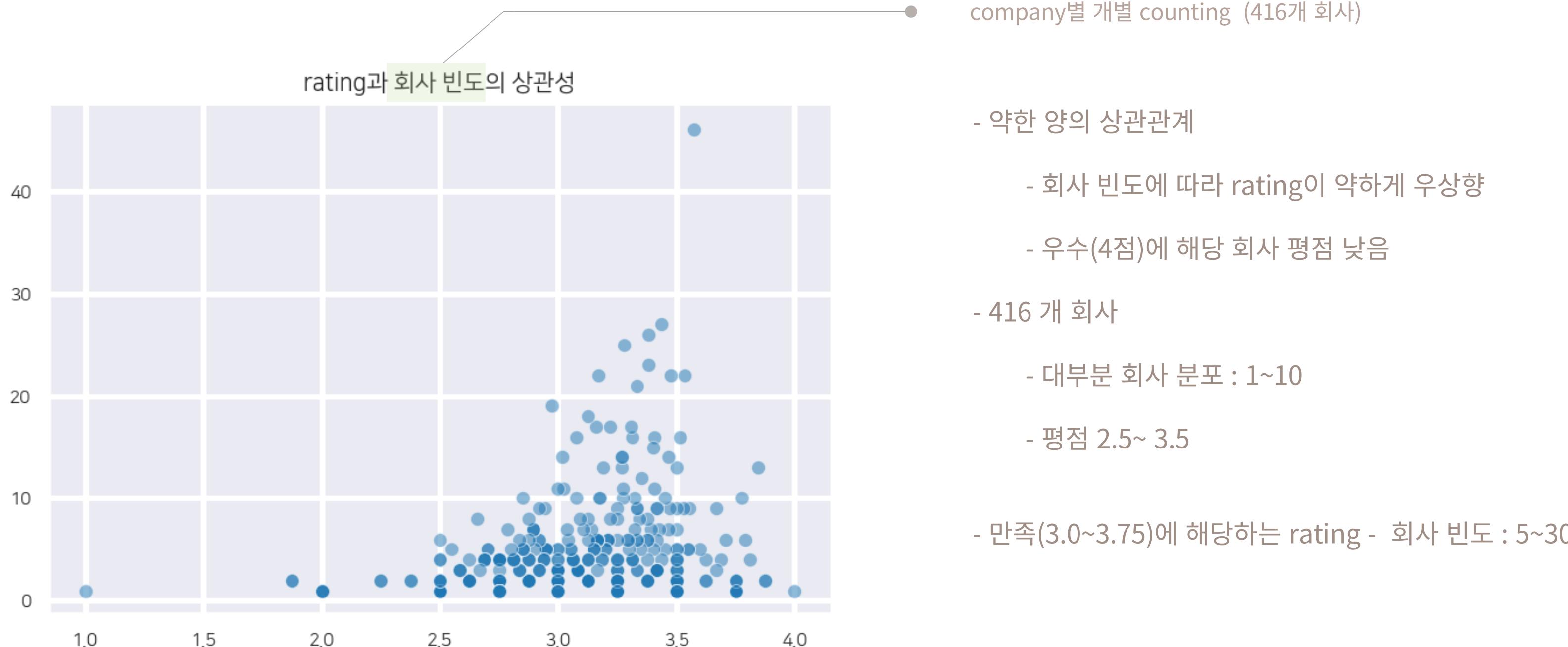
1. g_count(초콜릿 회사의 수)가 rating과 양의 상관관계를 가질 것이다.
2. Company Location(회사 위치 지역)에 따른 g_count(회사 수)가 rating과 상관관계를 가질 것이다.
3. 특정 지역의 원산지에서 나는 초콜릿이 rating이 더 높을 것이다.
4. 본래 원산지와 재배지가 같은 경우와 다른 경우 rating에서 차이를 보일 것이다.

Hypothesis

H. 01

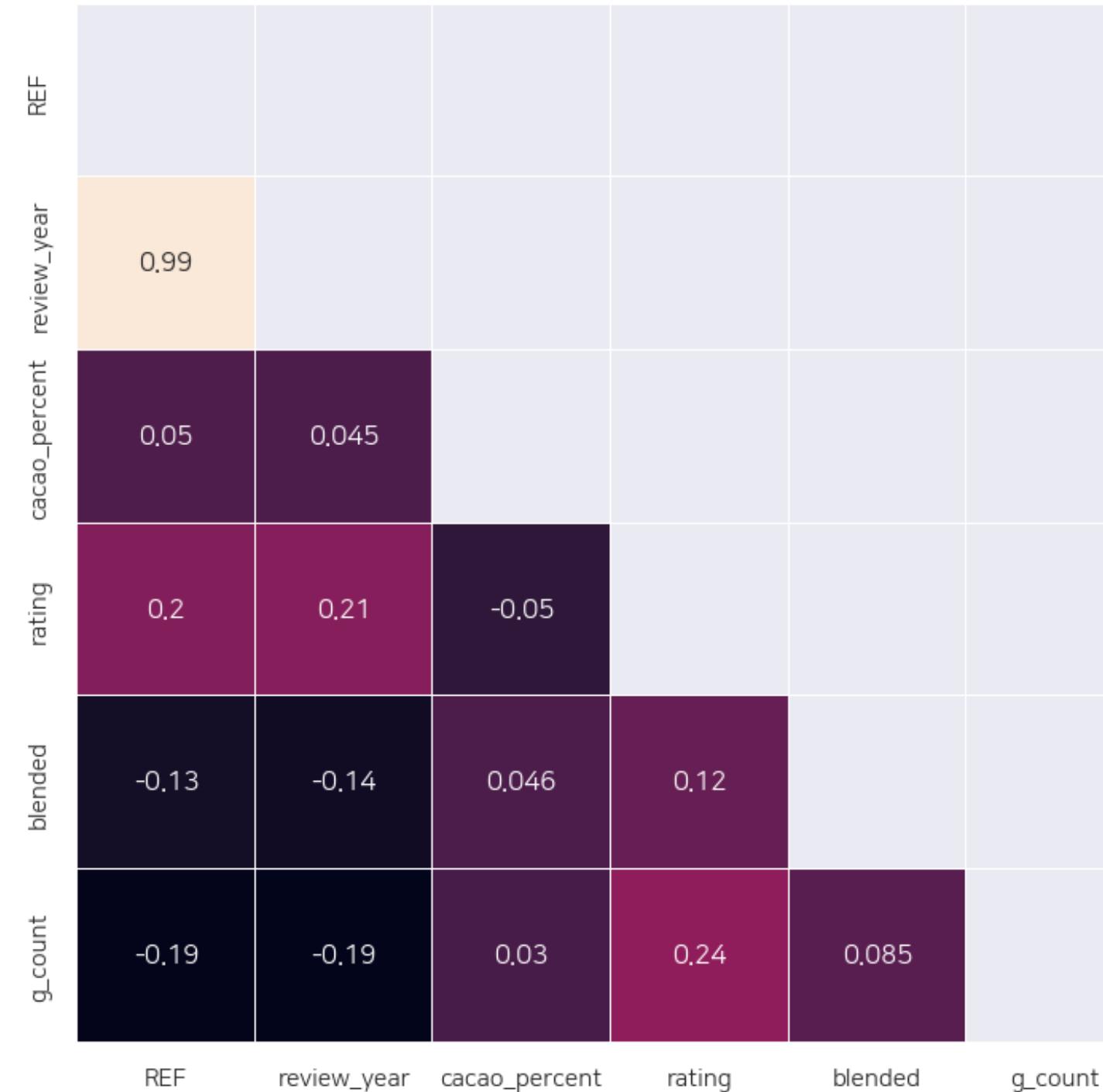
Data Science & Analysis

H0: g_count(회사빈도)가 rating과 양의 상관관계를 가질 것이다.



Company : Heatmap

Company 기준 상관관계 Heatmap



- Rating과 상관관계를 보이는 것:

	correlation	관계
blended	0.12	약한 양적 상관관계
REF	0.2	약한 양적 상관관계
review_year	0.21	약한 양적 상관관계
g_count	0.24	약한 양적 상관관계

Heatmap note:

Company 기준 상관관계 Heatmap 분류

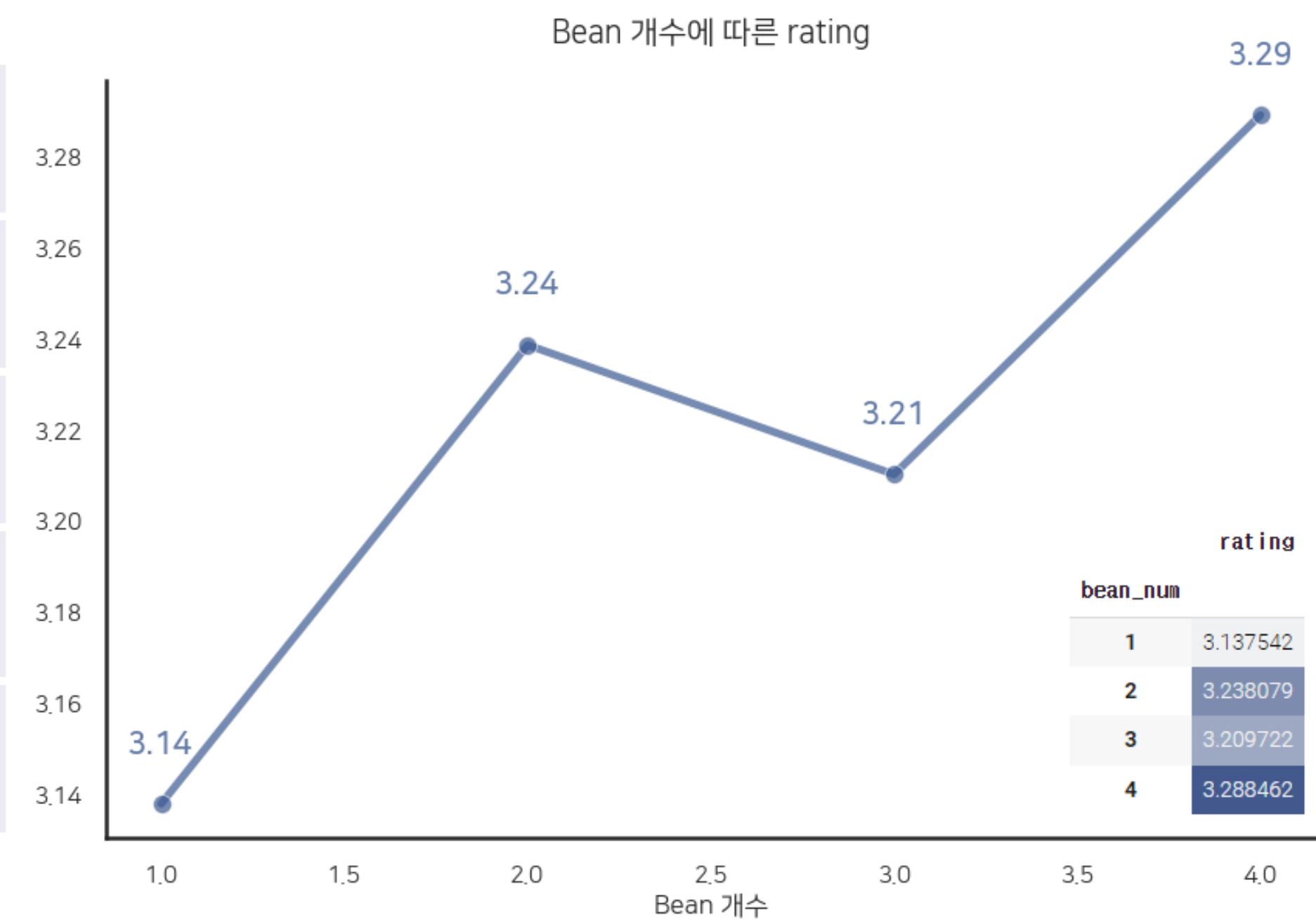
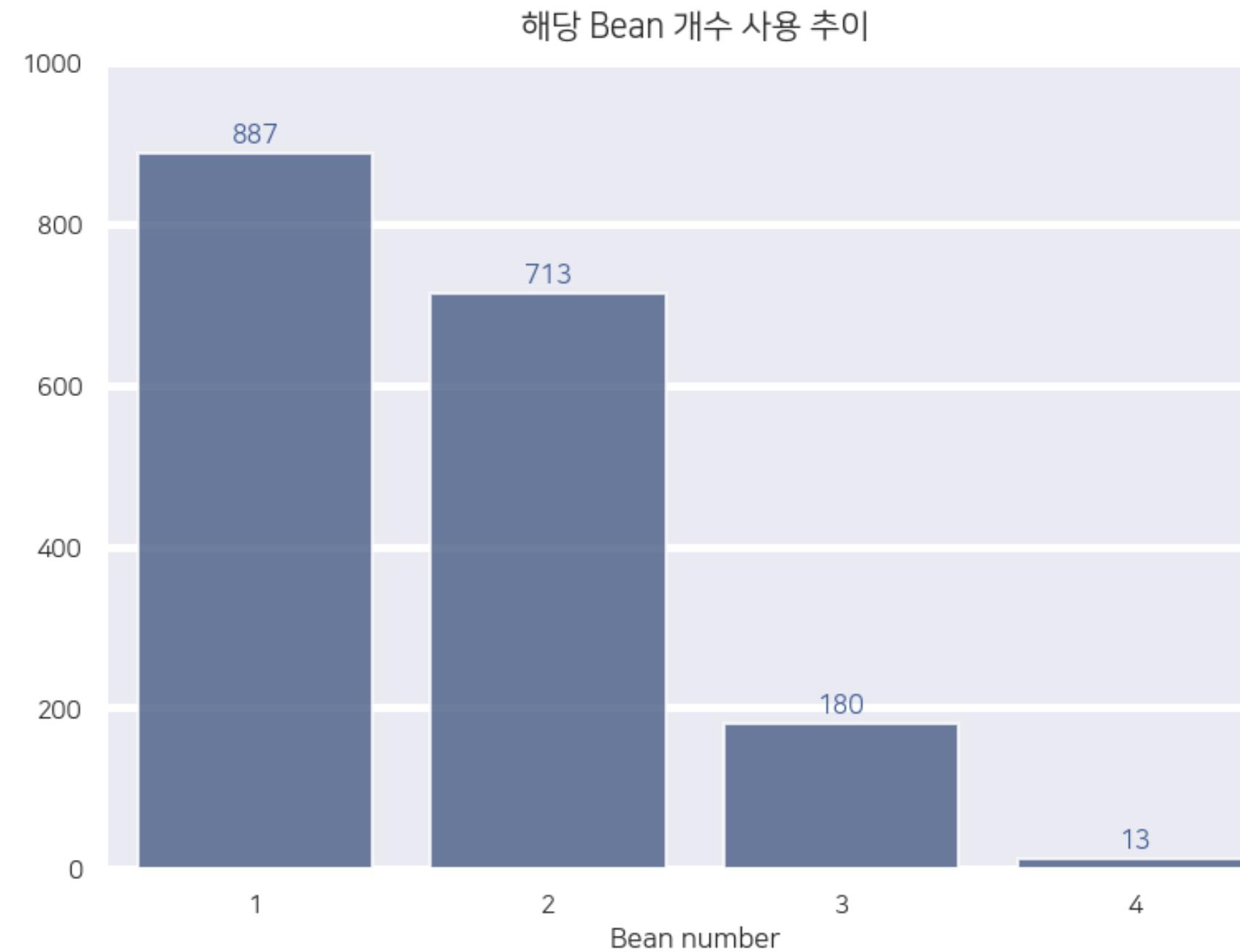
	강한 (0.7~1.0)	뚜렷한 (0.3~0.7)	약한 (0.1~0.3)
양적 상관관계	- Review year & REF		<ul style="list-style-type: none">- g_count & rating- review year & rating- REF & rating- blended & rating
음적 상관관계			<ul style="list-style-type: none">- g_count & review_year- g_count & REF- blended & REF- blended & review year
없다 할 수 있는 관계		<ul style="list-style-type: none">- cacao_percent & Blended- cacao_percent & rating- cacao_percent & g_count- cacao_percent & REF <ul style="list-style-type: none">- cacao_percent & review_year- g_count & Blended	

- rating을 포함해 cacao_percent와 상관관계를 가졌다 할 것이 없음
- cacao_percent는 rating과 무관하다.
- rating과 연관된 것: blended, g_count

Bean Number

rating이 blended와 상관관계를 보이므로, 추가적으로 blended한 bean의 개수와 rating이 상관성이 있는지 확인

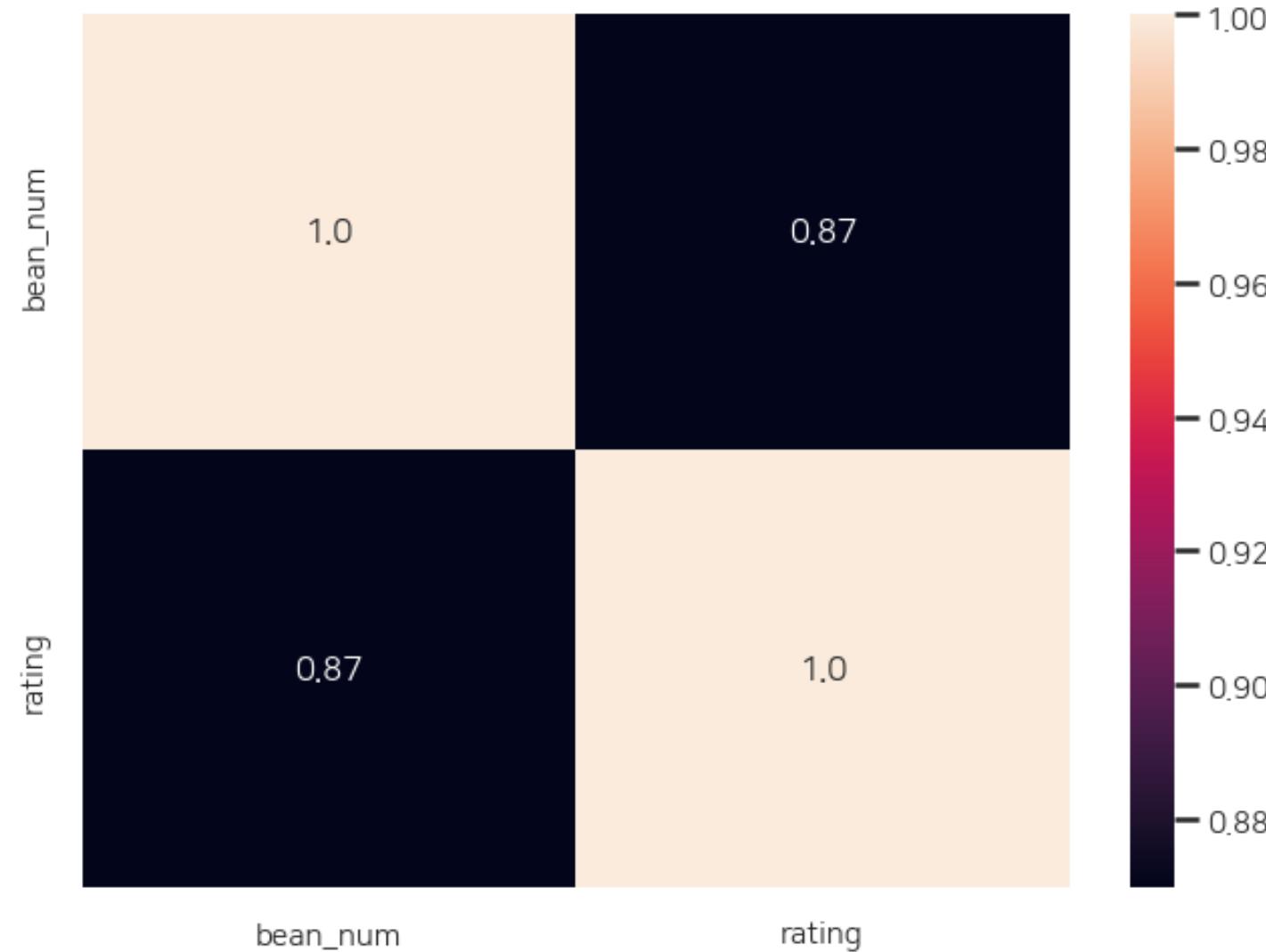
- Bean의 개수와 Bean 사용 추이는 반비례
- Bean 사용률은 Bean 개수에 따라 줄어드는 양상/ Bean의 개수가 늘어날 때 rating 증가하는 양상



Bean Number

rating이 blended와 상관관계를 보이므로, 추가적으로 blended한 bean의 개수와 rating과 상관성이 있는지 확인

Bean의 개수에 따른 rating 상관성

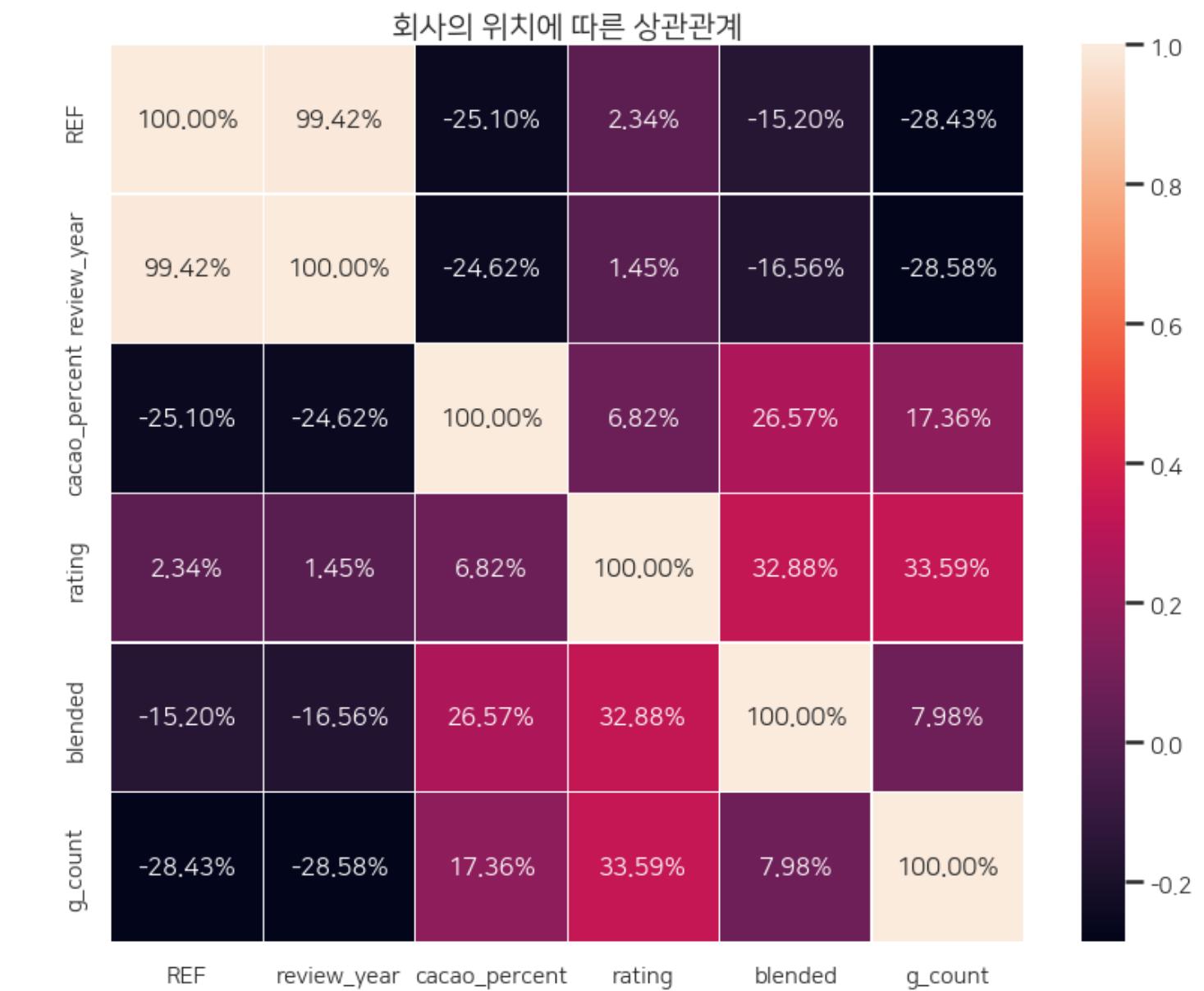
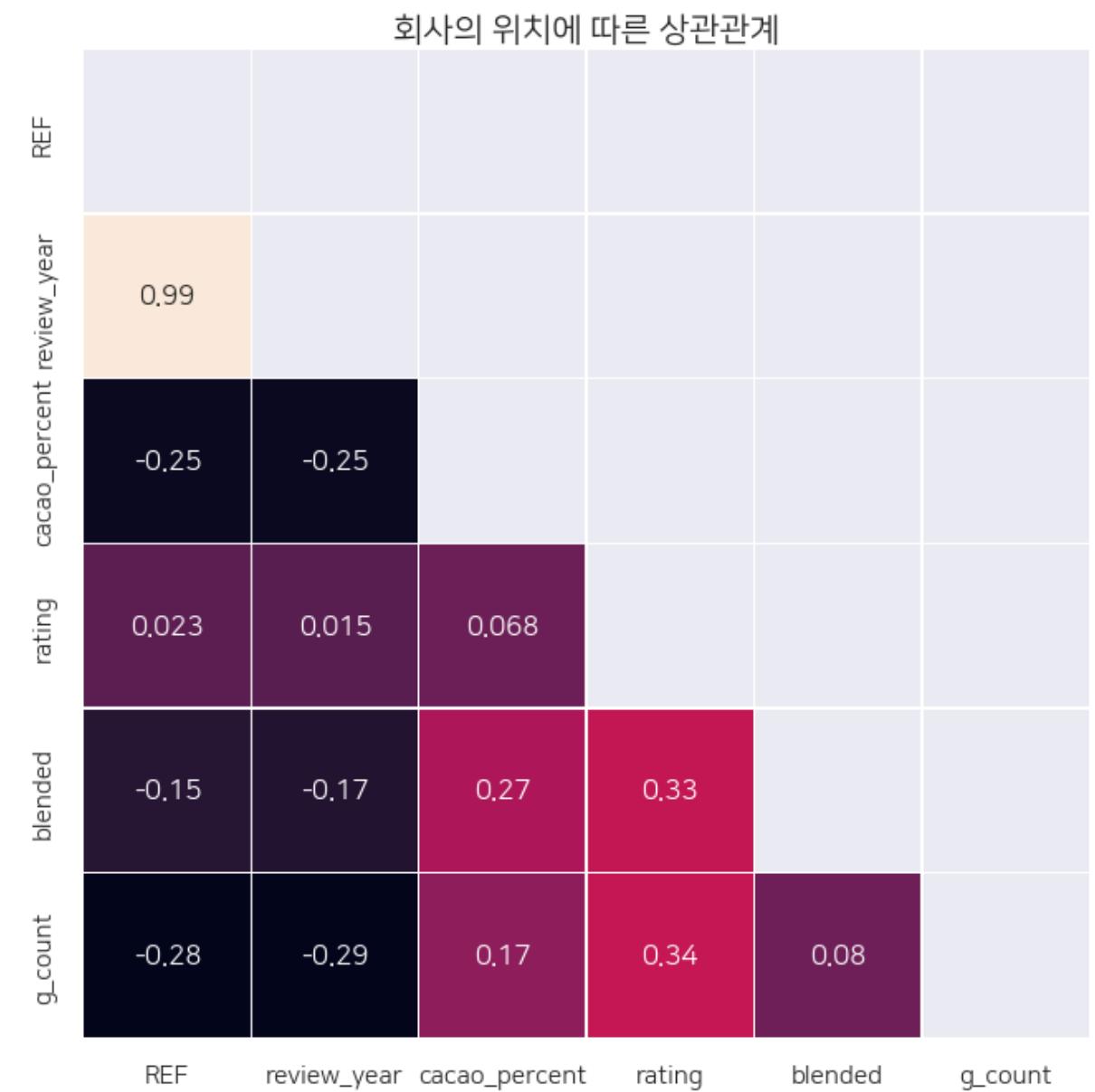


- Bean의 개수와 rating은 강한 양적 상관관계를 보임
- corr : 0.87
- > Bean의 개수를 늘리는 것이 좋을 것이다.

Hypothesis

H0: Company Location(회사 위치 지역)에 따른 g_count(회사 수)가 rating과 상관관계를 가질 것이다. (H0 만족)

- dataframe을 company_location에 따라 group화해서 구성
- company location (59)



Hypothesis

H0: Company Location(회사 위치 지역)에 따른 g_count(회사 수)가 rating과 상관관계를 가질 것이다.

	강한 (0.7~1.0)	뚜렷한 (0.3~0.7)	약한 (0.1~0.3)
양적 상관관계	- REF & Review year	- g_count & rating - blended & rating	- blended & cacao_percent - g_count & cacao_percent
음적 상관관계			- blended & REF - blended & review_year - cacao_percent & review_year - cacao_percent & REF - review_year & g_count - REF & g_count - REF & review_year
없다 할 수 있는 관계		- REF & rating - rating & review_year - rating & caco_percent - g_count & blended	

rating과 상관관계가 있는 것은 g_count, blended입니다.

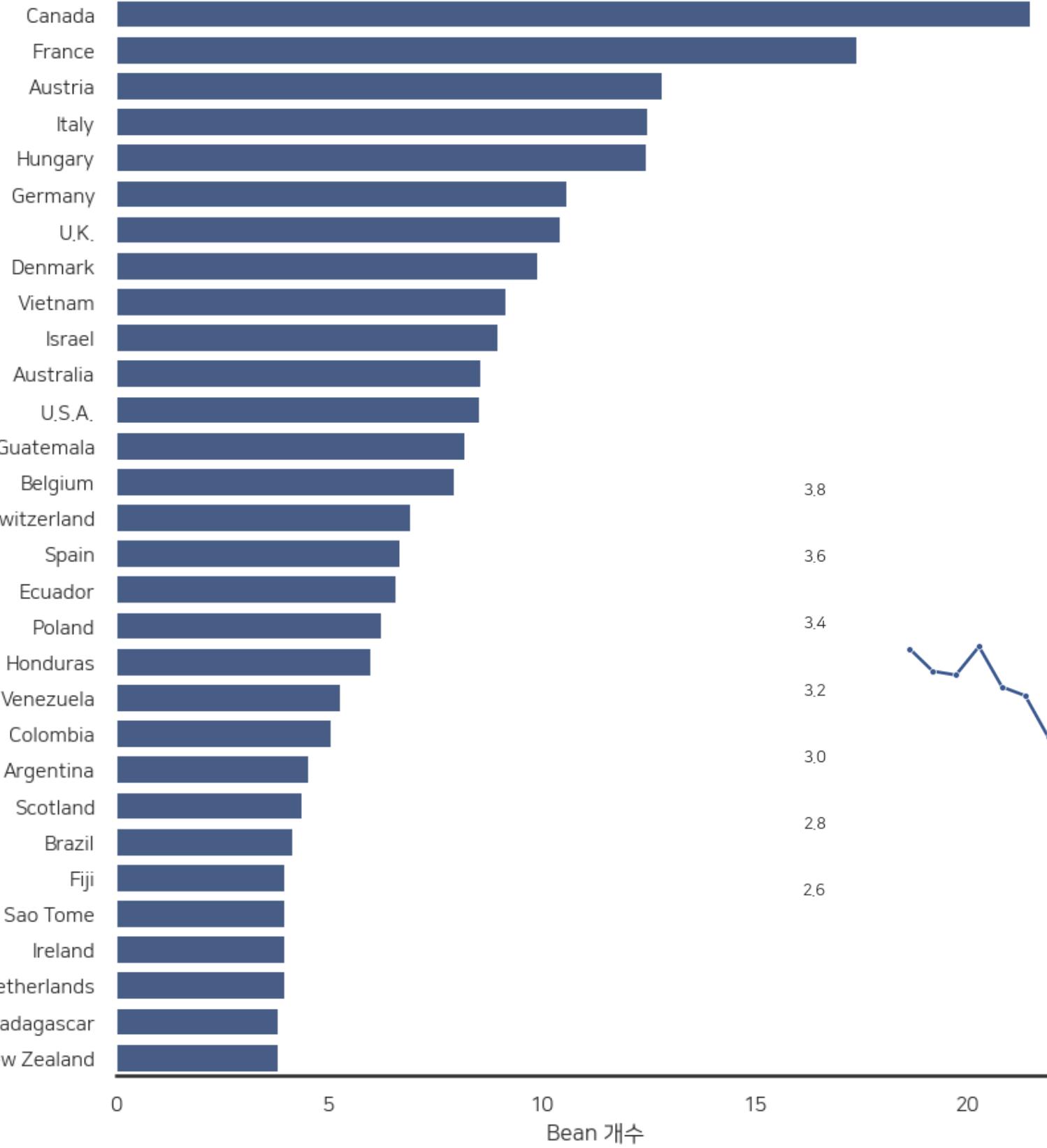
Company Location

H. 02

Data Science & Analysis

Location별 회사 분포 Top30

g_count & location



회사 분포 내림차순 --> Location별 회사 분포 Top30, rating 추이 확인

- 불규칙하지만 우하향 --> g_count & rating : 약한 상관관계(가설2 H0 만족)

회사 분포(내림차순)에 따른 rating 추이

회사 분포(corr=0.33)



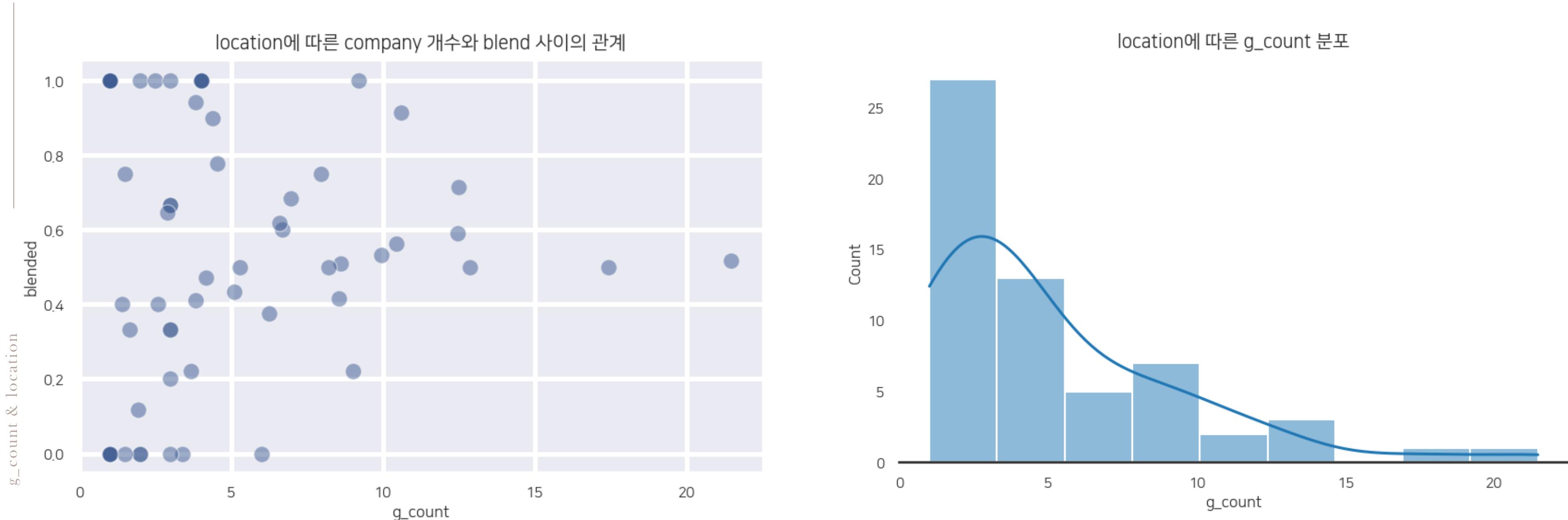
Company Location

H. 02

Data Science & Analysis

g_count와 rating이 약한 상관관계를 보였으므로 blended 추이를 확인해 각 회사의 blended 양상을 확인할 수 있을 것

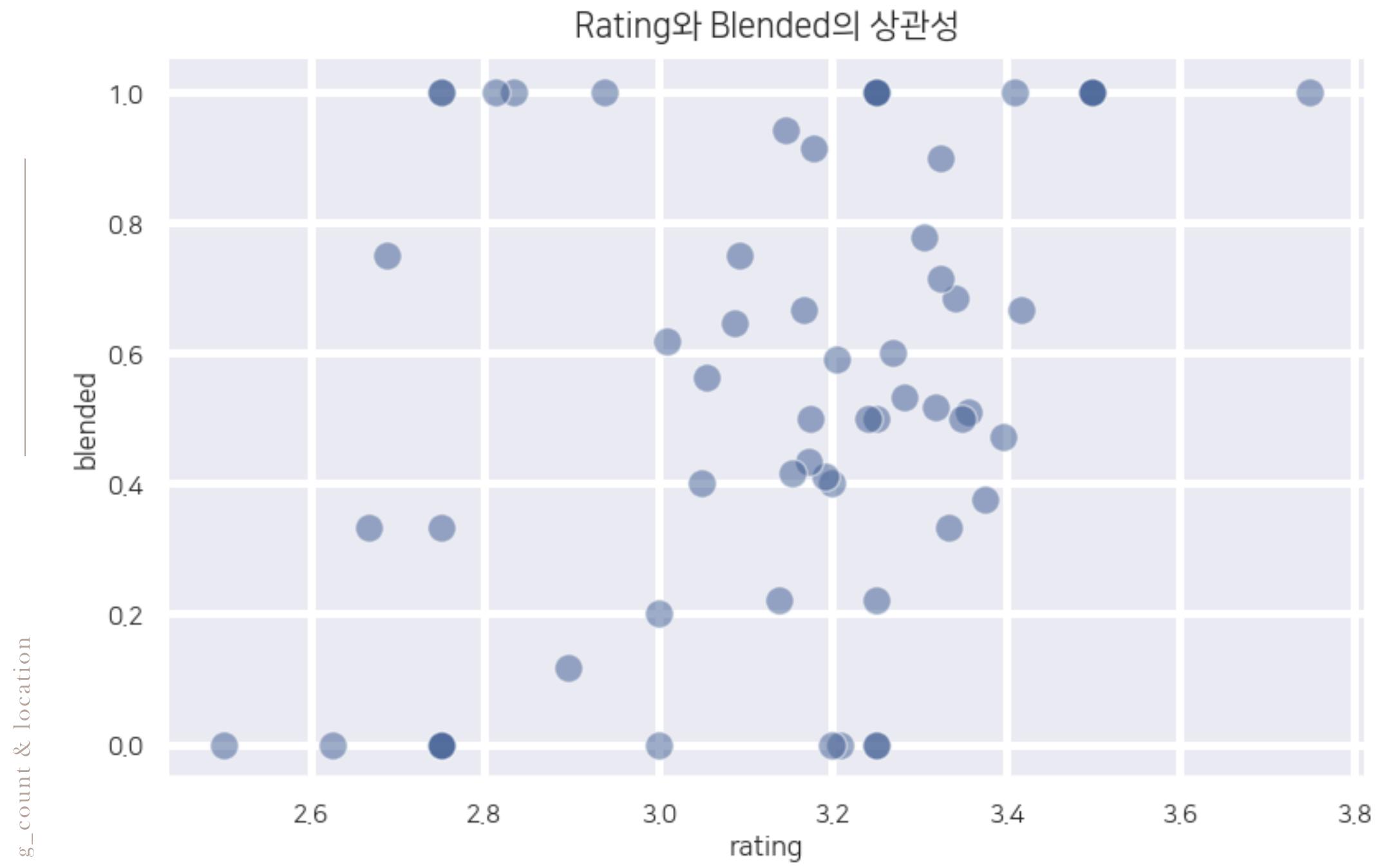
- 보통 지역에 따른 company 수가 0~ 10 사이에 위치함
- 0~ 5가 회사 분포대로 가장 많음을 확인할 수 있음 --> 좌측 그래프에 따르면 명확한 추이 없이 고르게 분포
- > location에 따른 company수와 blended 여부는 상관관계가 없다.



정리: 지역에 따라 회사의 분포가 영향을 가졌고, 회사분포에 따라 rating이 달라졌으나 회사의 분포와 Blended와는 상관이 없었다

-회사 분포가 blended 여부를 결정 x -> 지역에 따른 blended의 영향력이 없다 (지역과 blended는 무관하다)

Blended & Rating corr.



약한 우상향 (corr = 0.328)

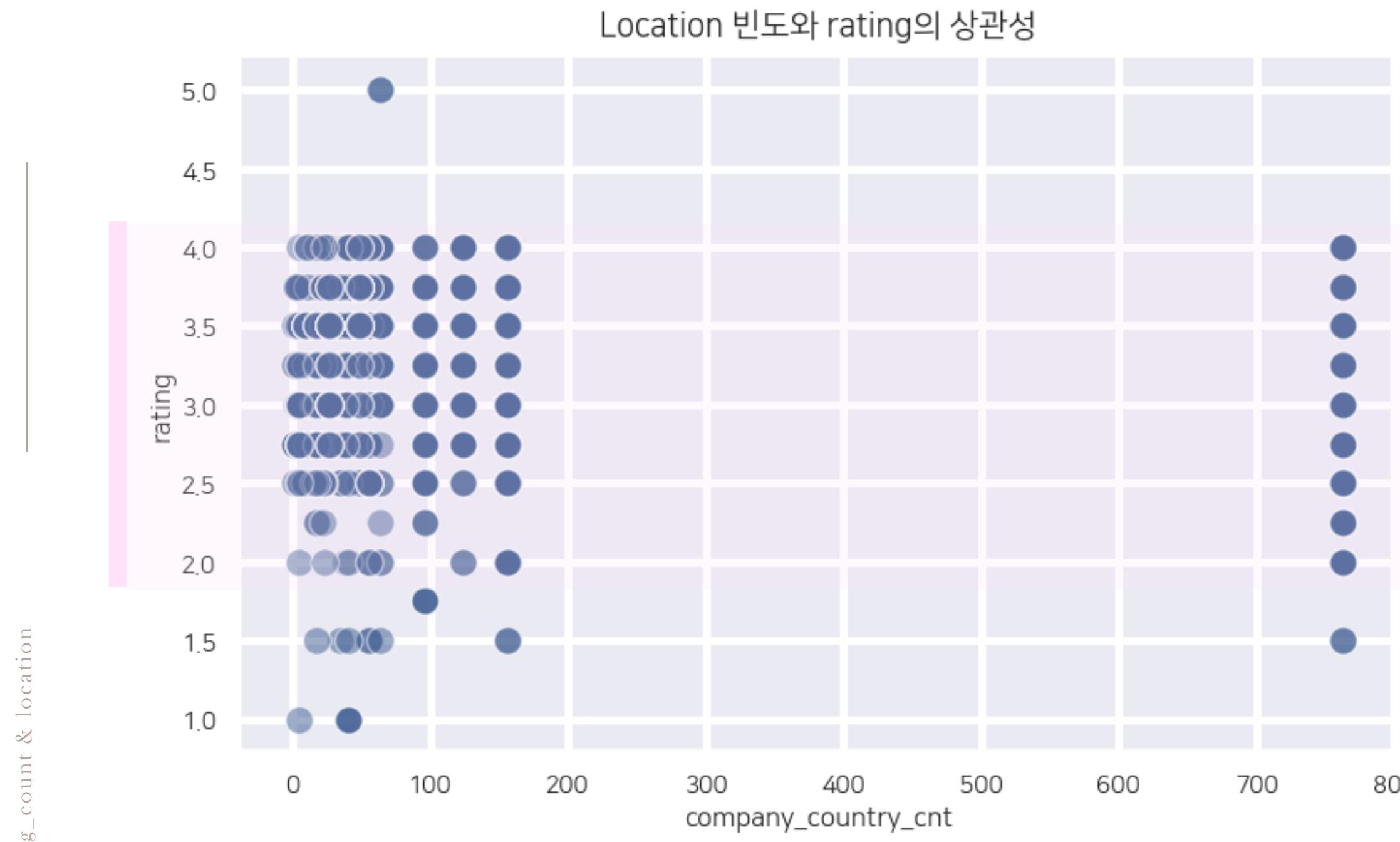
낮은 rating	blend 안한 경우 많음
중간 rating	섞여있음
높은 rating	blended

지역에 따른 blended에 대한 상관성은 없으나,

Blended와 Rating은 약한 양의 상관관계를 가짐

- Blended 하는 것이 좋다.

Company count & rating



- 같은 회사에서 출시한 제품일 경우 : 상-하 관계 추정
- 국가 별 추이와 무관하게 rating이 유사하게 분포
- 2점~4점대 빈번
- > country 빈도가 rating에 영향을 주지 않는다.

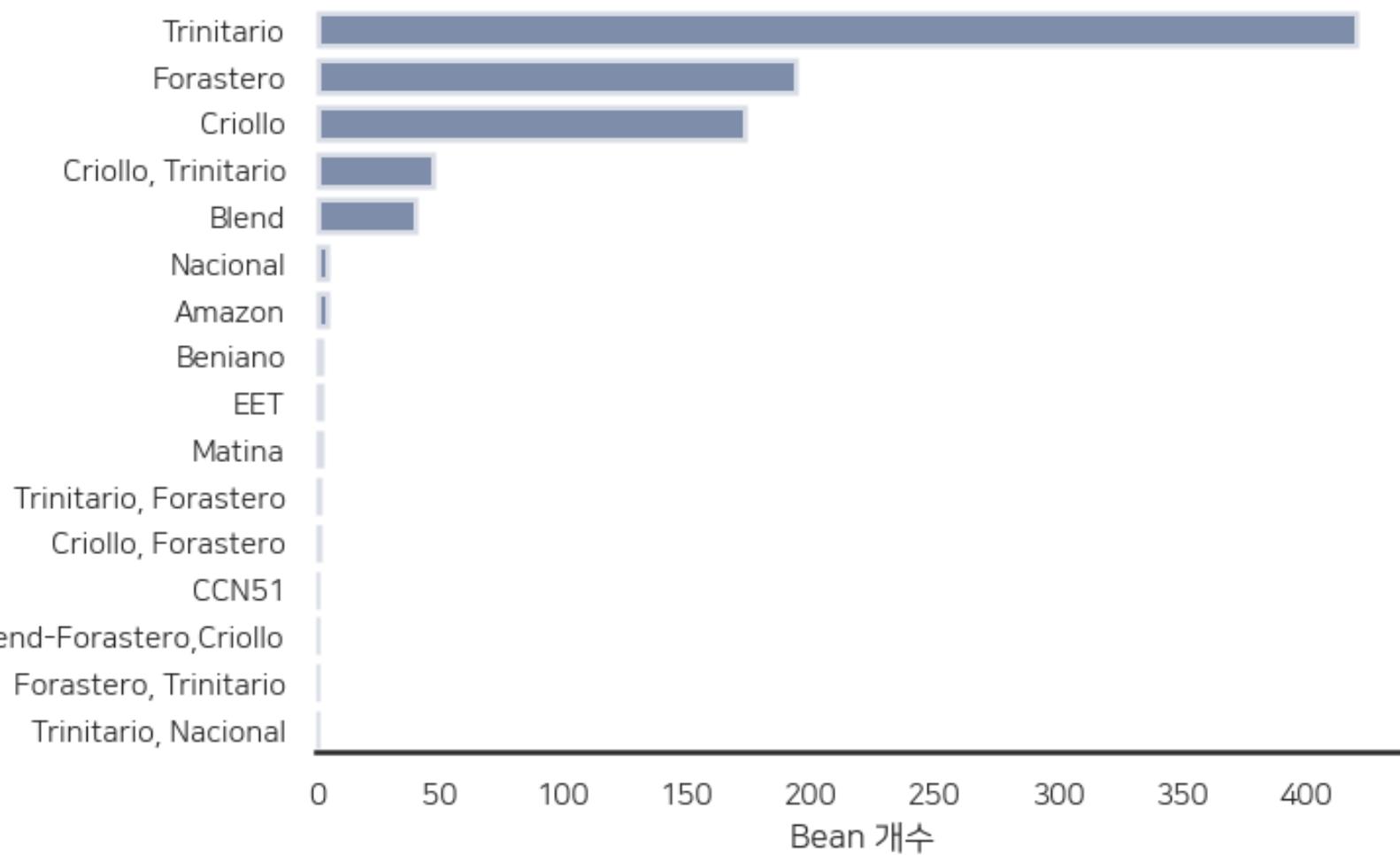
Bean Type

H0: 특정 지역의 원산지에서 나는 초콜릿이 rating이 더 높을 것이다.



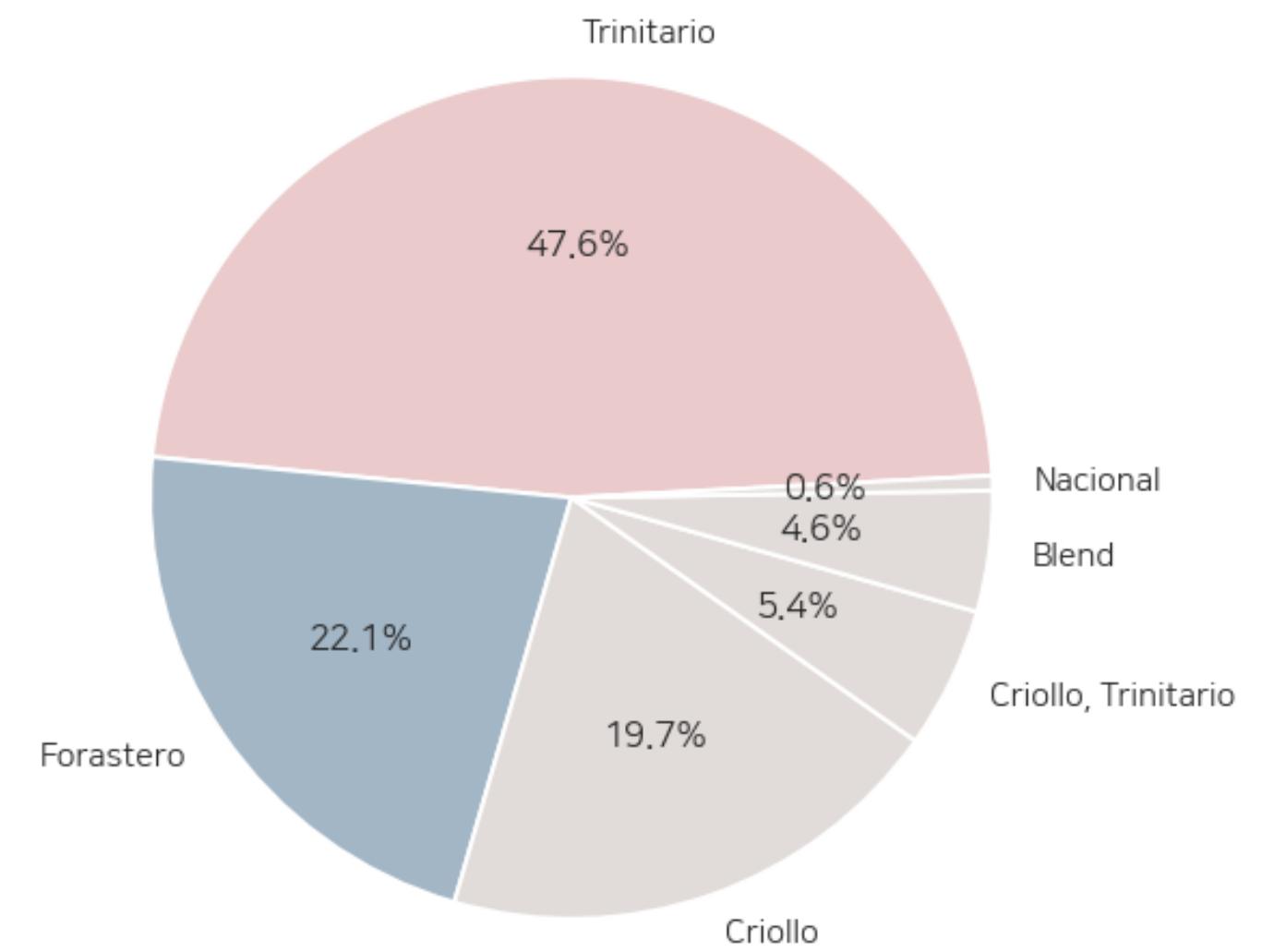
Bean Type 확인 (공백 제거)
1793 --> 906 dataset

Bean Type Count



Hypothesis ---

Bean Type percent Top5



Bean Type Count

Bean Type을 많이 쓰는 것이 rating에 영향을 주지 않는다

- Bean Type에 따라 group화
- rating 의 평균 유사한 것을 확인 할 수 있음
- Bean Type 종류 역시 rating과 무관하다.



	rating	type_count
bean_type		
Amazon	3.600000	5.000000
Beniano	3.583333	3.000000
Blend	3.353659	41.000000
Blend-Forastero,Criollo	3.750000	1.000000
CCN51	3.500000	1.000000
Criollo	3.267241	174.000000
Criollo, Forastero	3.625000	2.000000
Criollo, Trinitario	3.244792	48.000000
EET	3.583333	3.000000
Forastero	3.112821	195.000000
Forastero, Trinitario	3.000000	1.000000
Matina	3.416667	3.000000
Nacional	3.150000	5.000000
Trinitario	3.248812	421.000000
Trinitario, Forastero	3.000000	2.000000
Trinitario, Nacional	3.750000	1.000000

Species & Origin

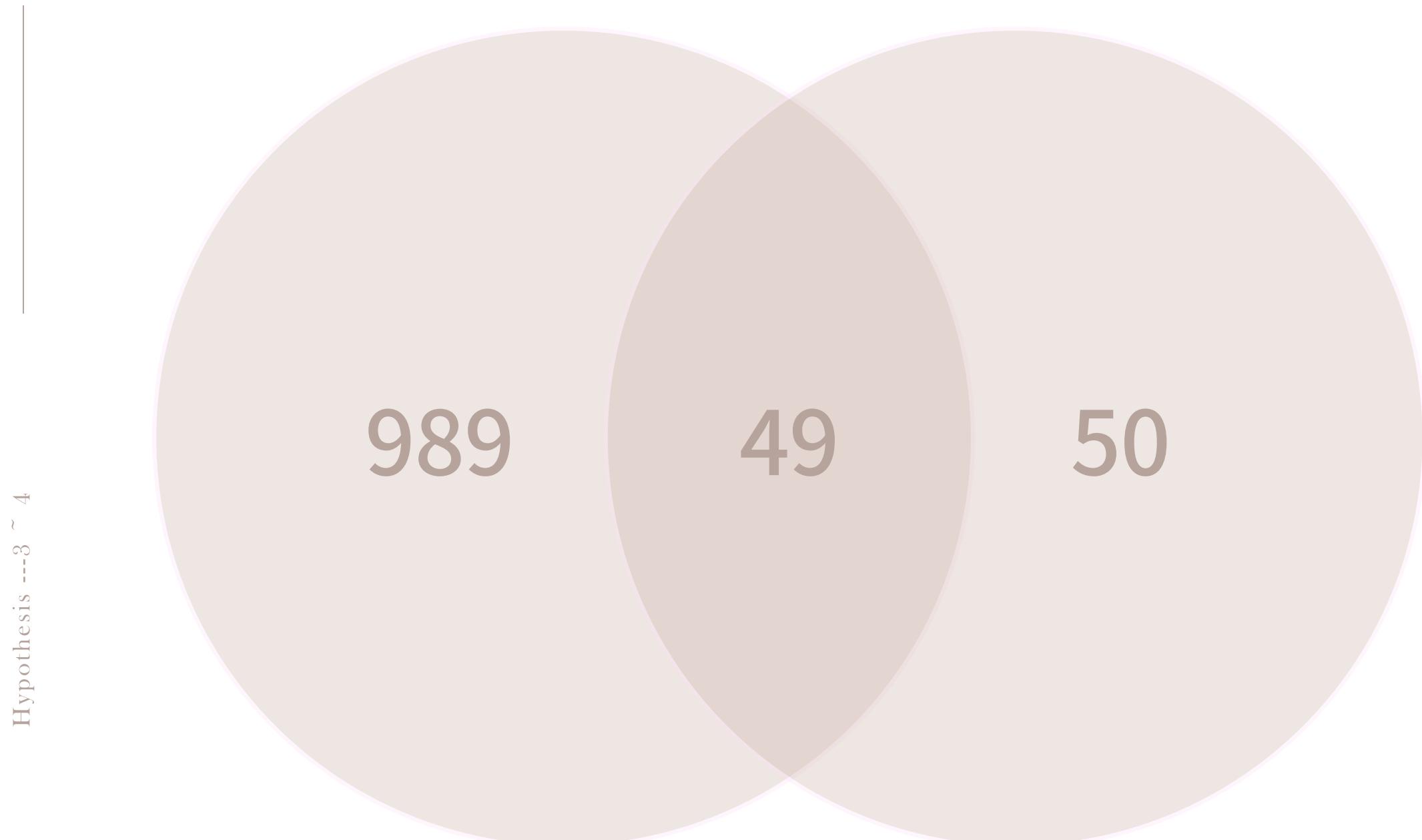
H. 03 - 04

Data Science & Analysis

H0: 특정 지역의 원산지에서 나는 초콜릿이 rating이 더 높을 것이다.

H0: 원산지와 재배지가 동일한 경우 rating이 더 높을 것이다.

Species (재배지) Origin (원산지)

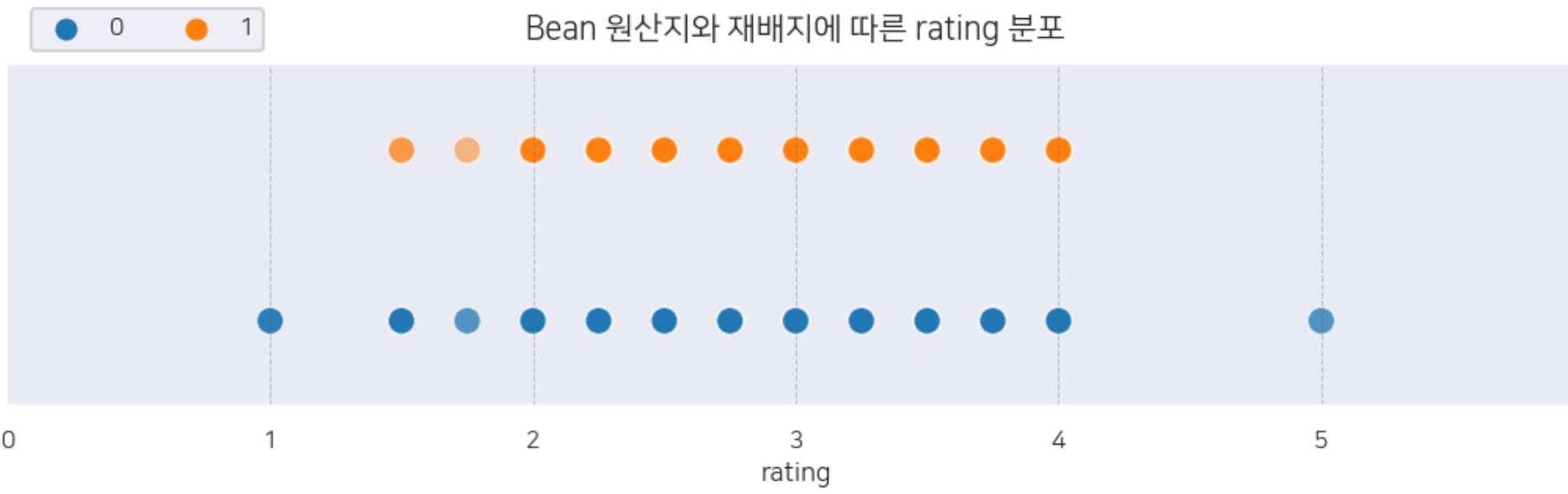


	전체	차집합	공통
원산지	1038	989	49
재배지	99	50	
품목 일자 데이터셋	401 (22%)		

Species & Origin

H. 03 - 04

Data Science & Analysis

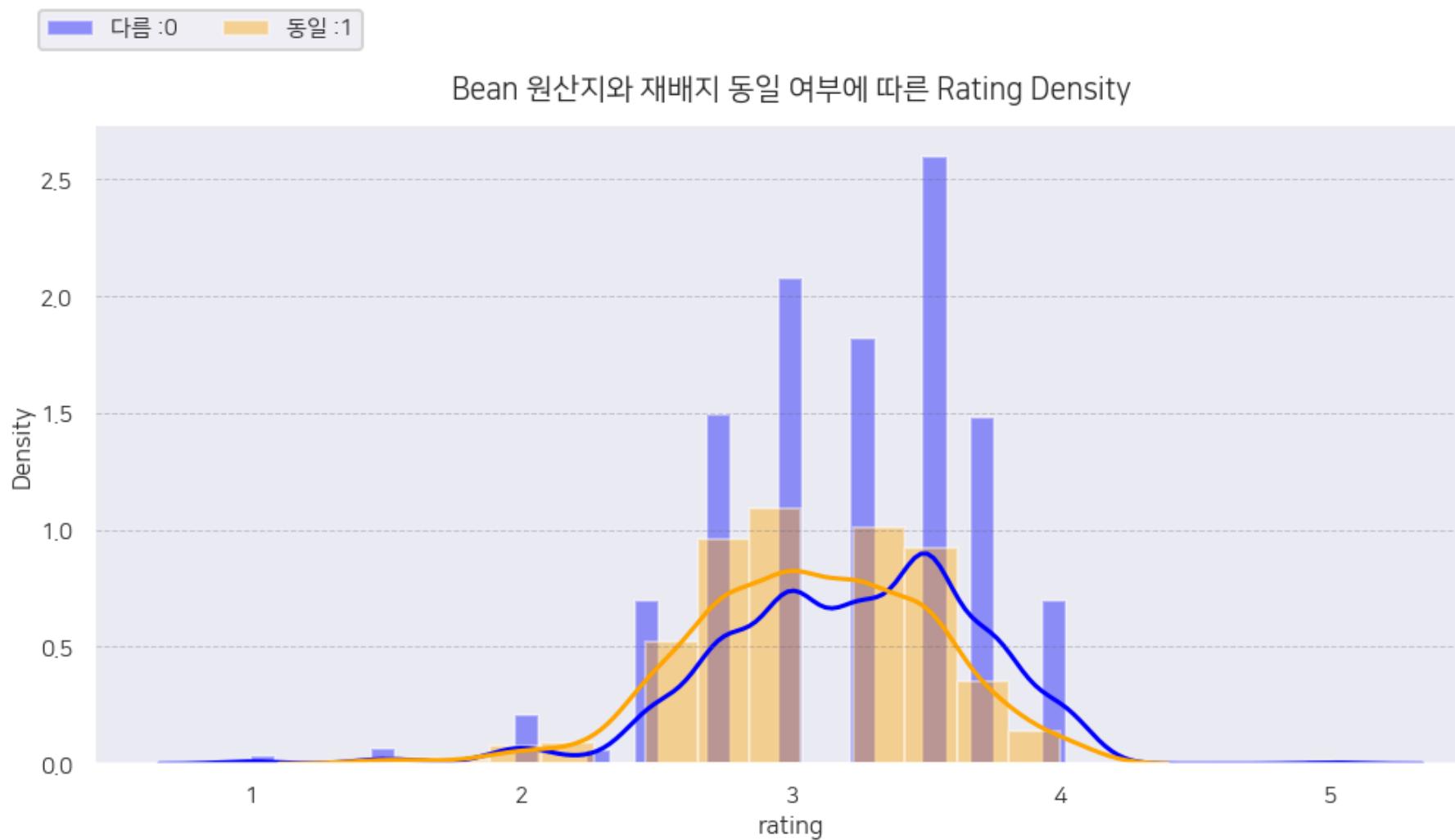


같은 경우 1, 다른 경우 0 (bean_born_same)

- 기존 데이터셋의 분포가 치우쳐져 있었기 때문에 정확한 파악은 어려웠음

동일: 1.5~4 (응집)

상이: 1~4// 5 (퍼져있음)

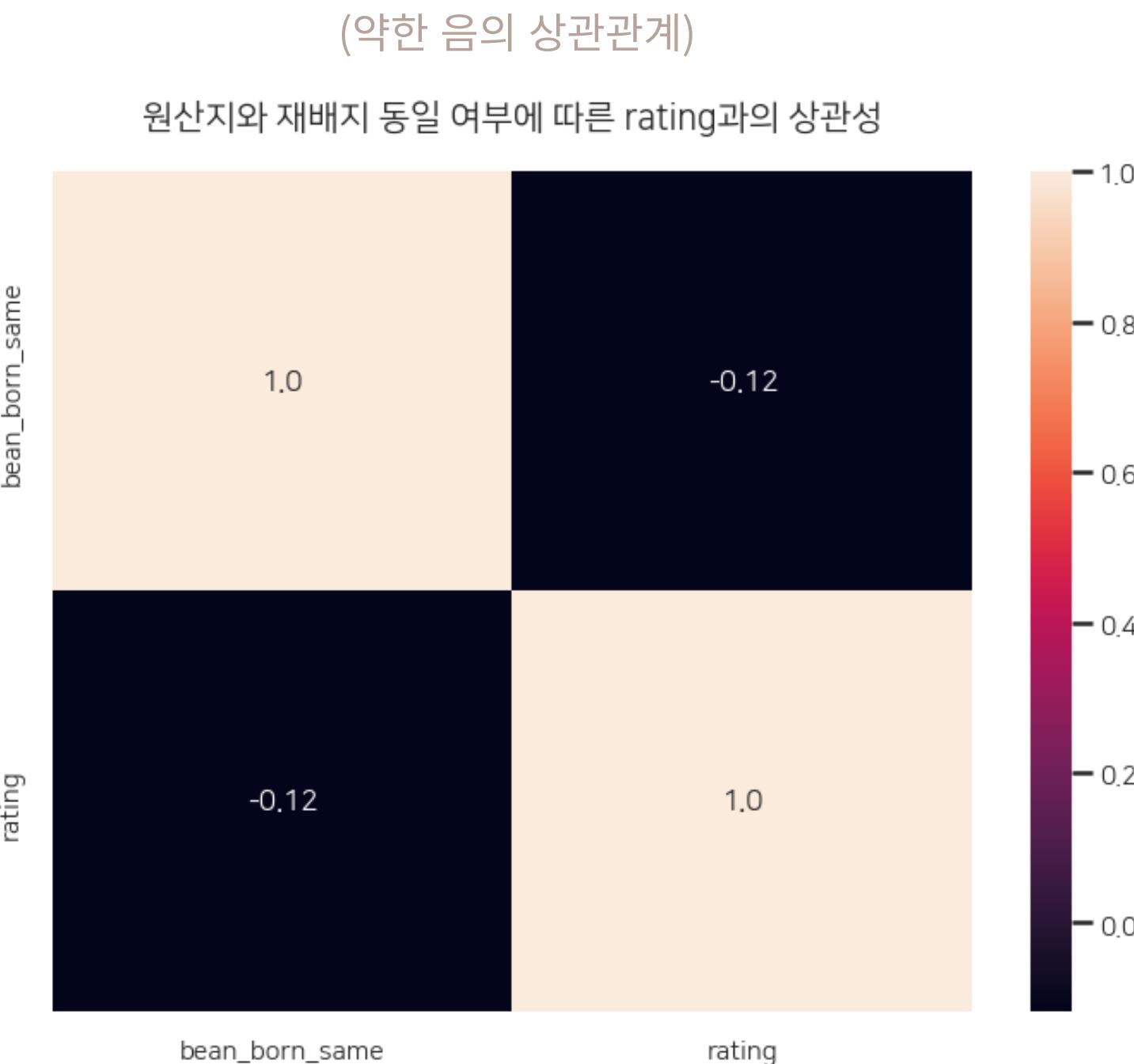


Species & Origin

H. 03 - 04

Data Science & Analysis

재배장소와 원산지가 동일한 경우 rating에 부정적인 영향을 미친다



Species & Origin

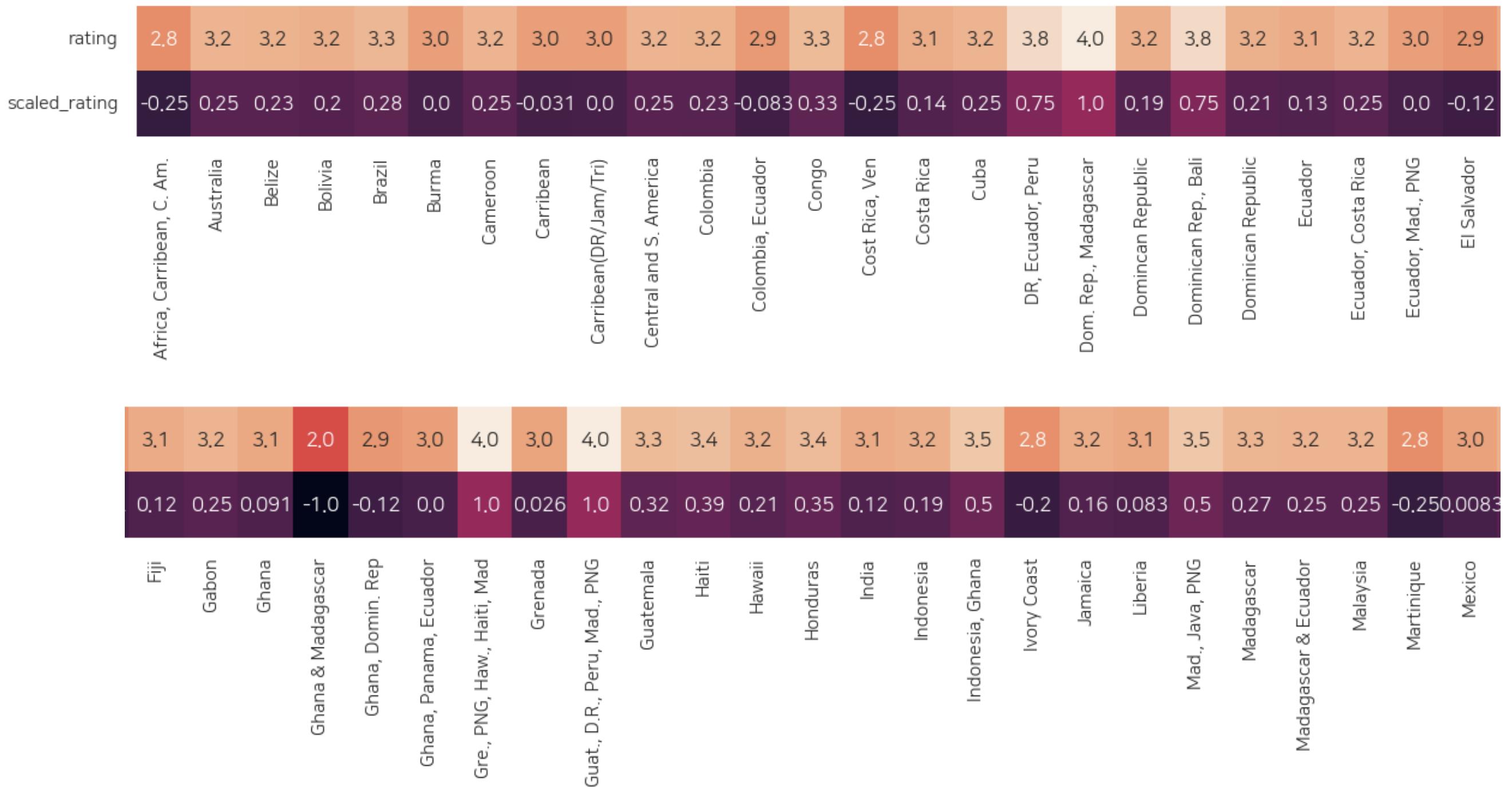
H. 03

H0: 특정 지역의 원산지에서 나는 초콜릿이 rating이 더 높을 것이다.

Data Science & Analysis

Bean Origin과 rating Dataset --> 원산지에 대한 rating 평균 구해줌

Heatmap을 구했는데 수치가 2~4로 나와 0~1 사이의 수치로 맞춰줌



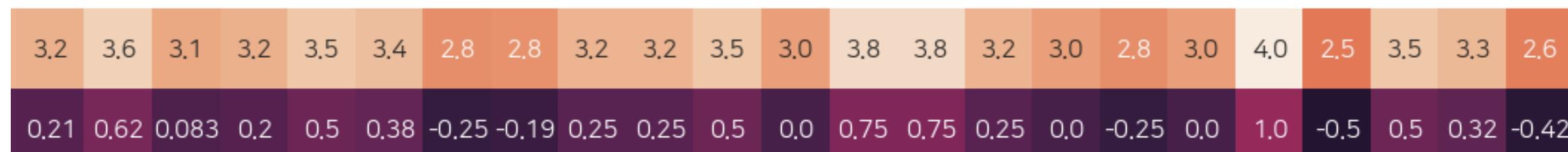
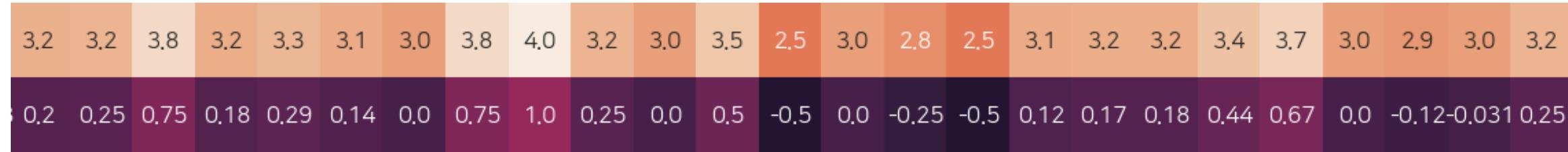
Species & Origin

Hypothesis --- 3 ~ 4

H0: 특정 지역의 원산지에서 나는 초콜릿이 rating이 더 높을 것이다.

H. 03

Data Science & Analysis



Species & Origin

H. 03

Data Science & Analysis

원산지 항목이 많으므로 강한 양적 상관관계와 뚜렷한 양적 상관관계만 정리

강한 양적 상관관계 (0.7 ~ 1)	뚜렷한 양적 상관관계 (0.3 ~ 0.7)
DR, Ecuador, Peru Dominican Rep., Bali Gre., RNG, Haw., Haiti, Mad Guat., DR., Peru, Mad., PNG PNG, Vamaiatu, Mad Peru, Belize Ven., Ecu., Peru, Nic. Venez,Africa,Brasil,Peru,Mex Venezuela, Java	Congo Guatemala Haiti Honduras Indonesia Ghana Mad, Java, PNG Tobago Trinidad, Ecuador Trinidad, Tobago

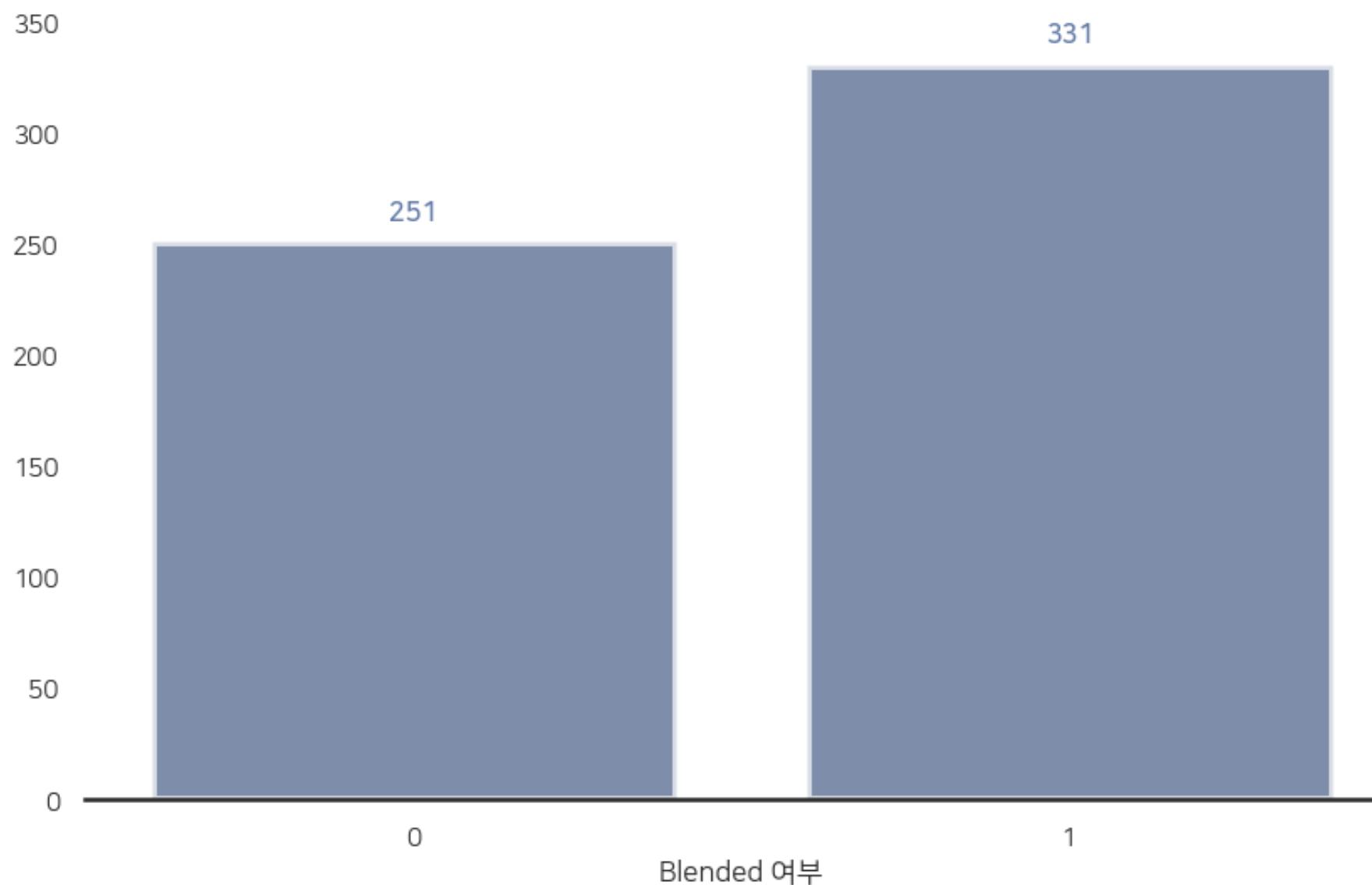
Species & Origin

H. 03 - 04

Data Science & Analysis

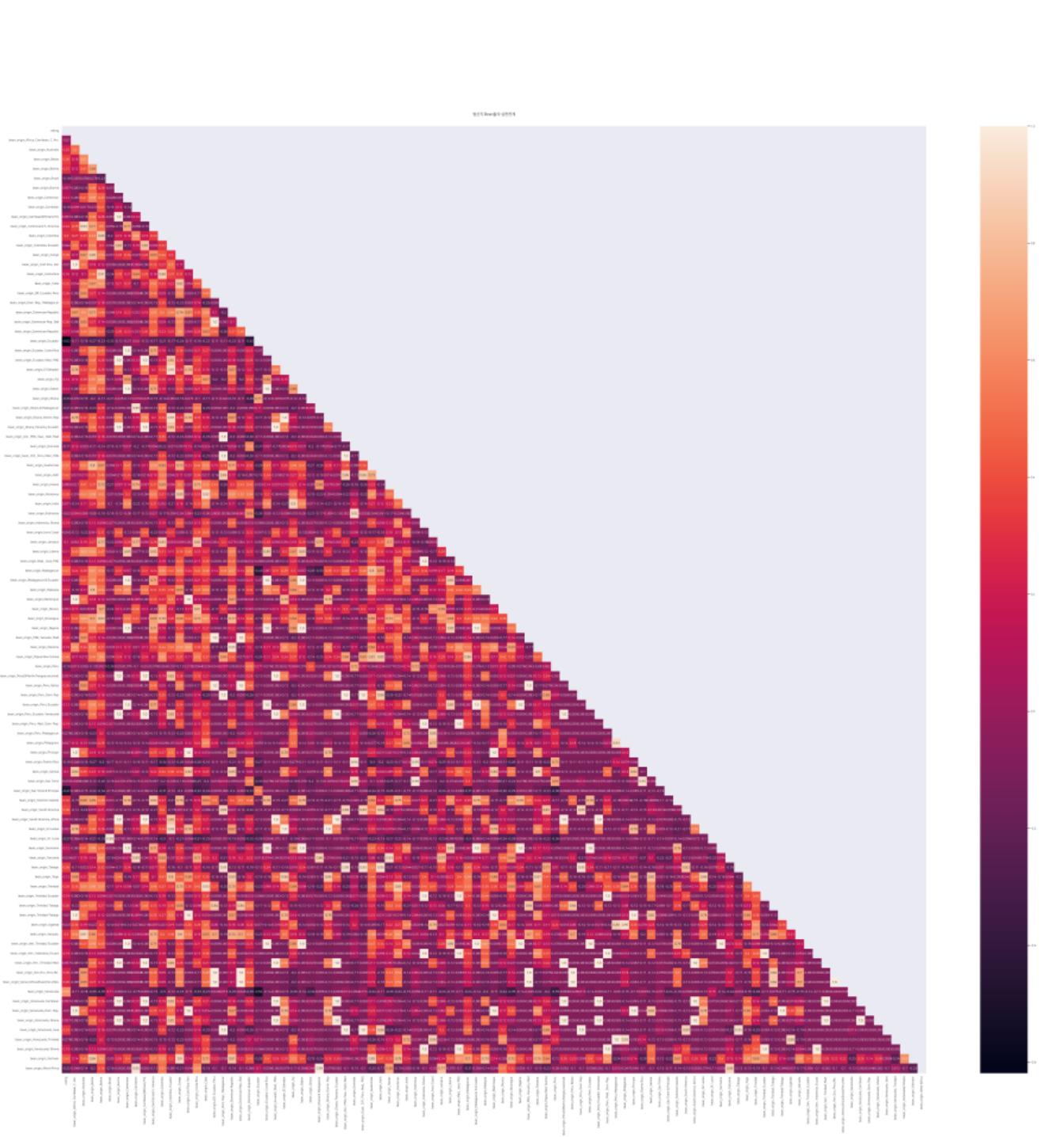
- 여러 원산지를 포함한 항목에 대해 Blended라 가정
- 그러나 여러 원산지를 가지고 있어도 Blended된 것과 Blended 안 된 항목 유사했음
- 원산지나 재배지가 여러 곳인 것은 Blended 여부와는 무관함

여러 Species 가진 경우 blended 분포 확인

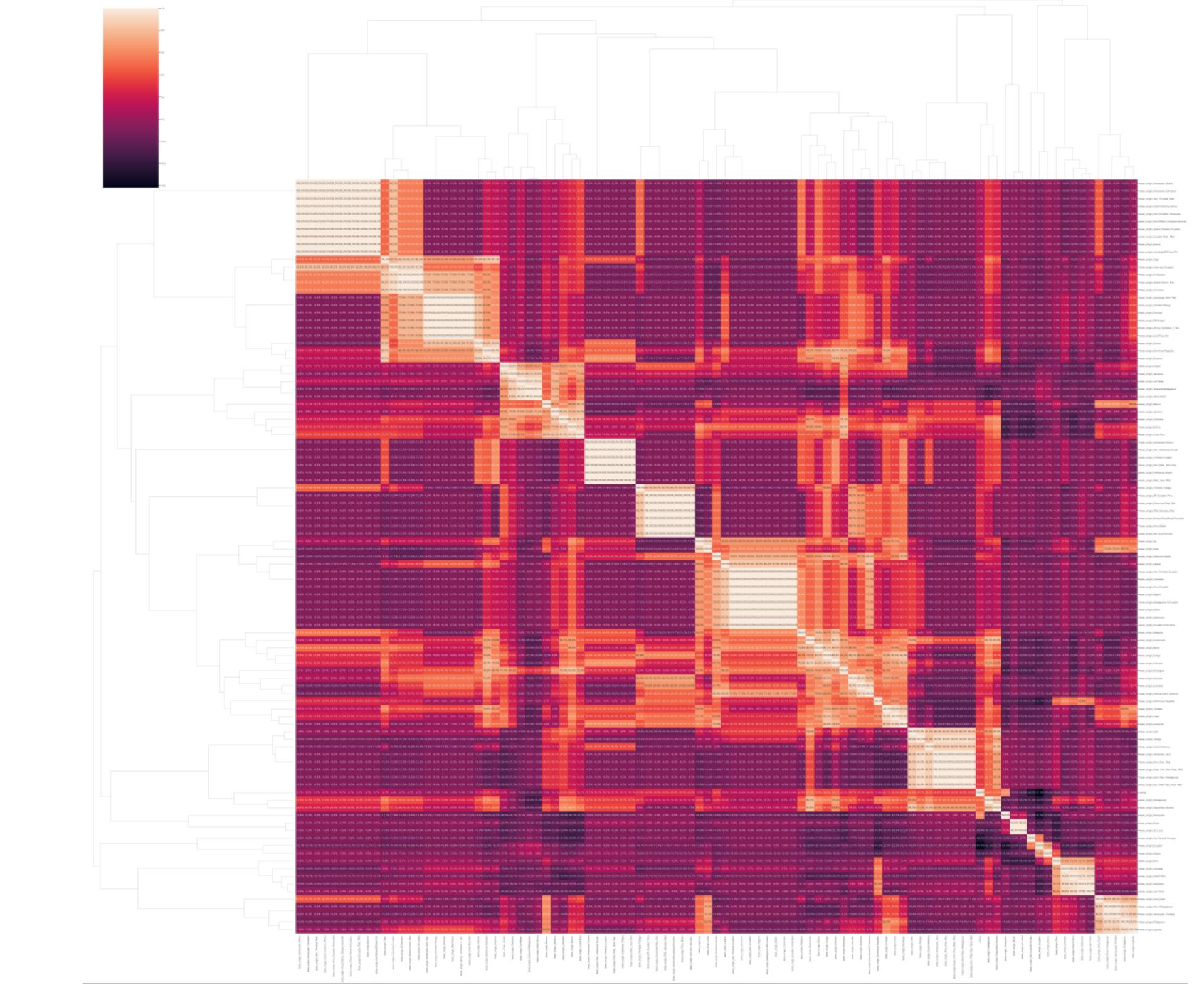


Bean corr.

Bean들의 상관관계 (Heatmap)



(Clustermap)



Hypothesis ---3

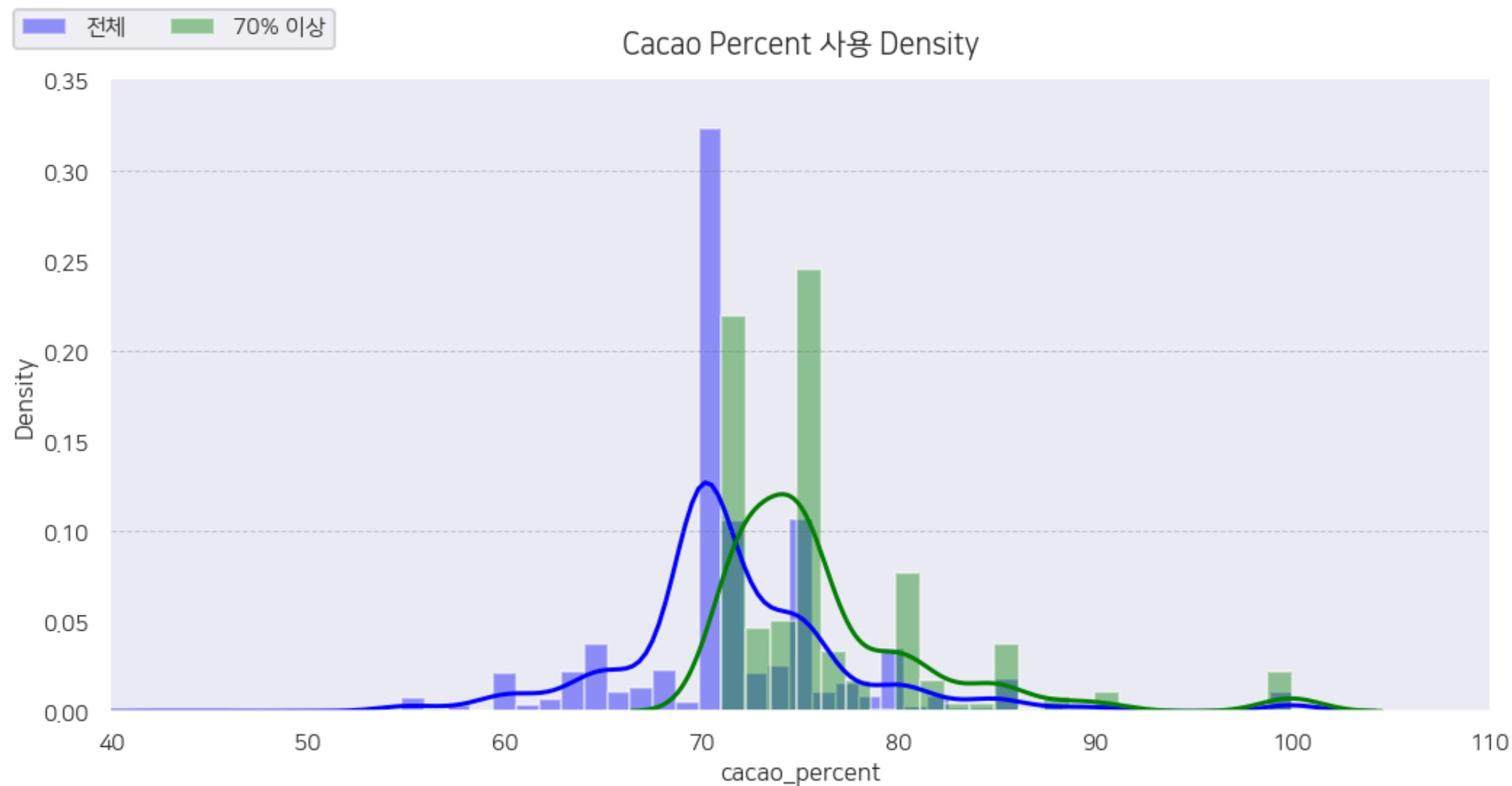
Cacao Percent

Data Science & Analysis

- cacao percent 70 > 75 > 72 > 80

- 고함량시 70 -75- 80 순 추천

- 함량 무관할 시 70 -75- 72 순 추천



cacao_percent	Count
70.0	37.423313
75.0	12.381484
72.0	10.485220
65.0	4.350251
80.0	4.015616
74.0	2.788622
68.0	2.621305
60.0	2.398215
73.0	2.230898
85.0	2.007808

x 3

Cacao Percent

usage & rating

Data Science & Analysis

usage & rating

corr = 0.157

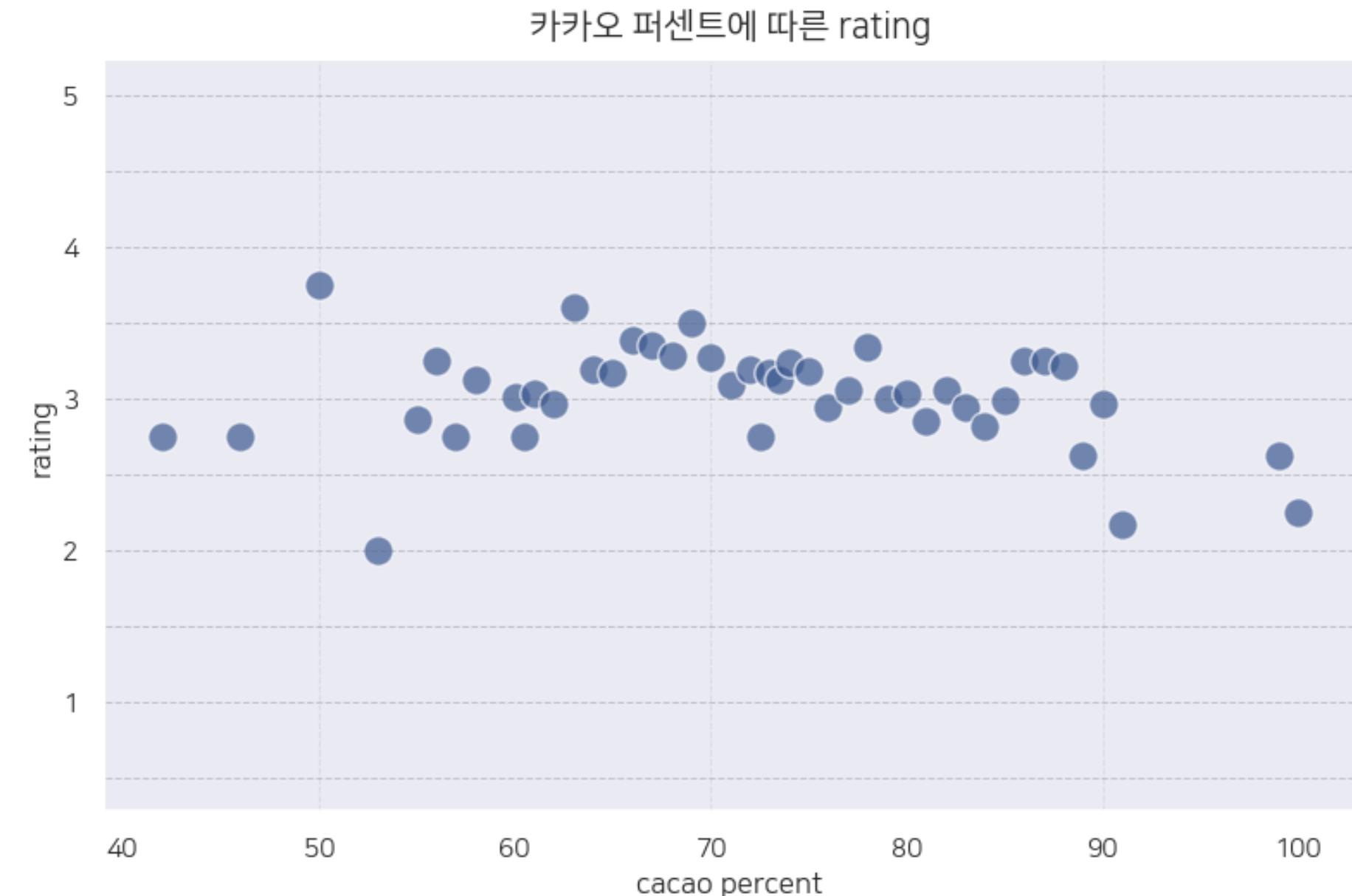
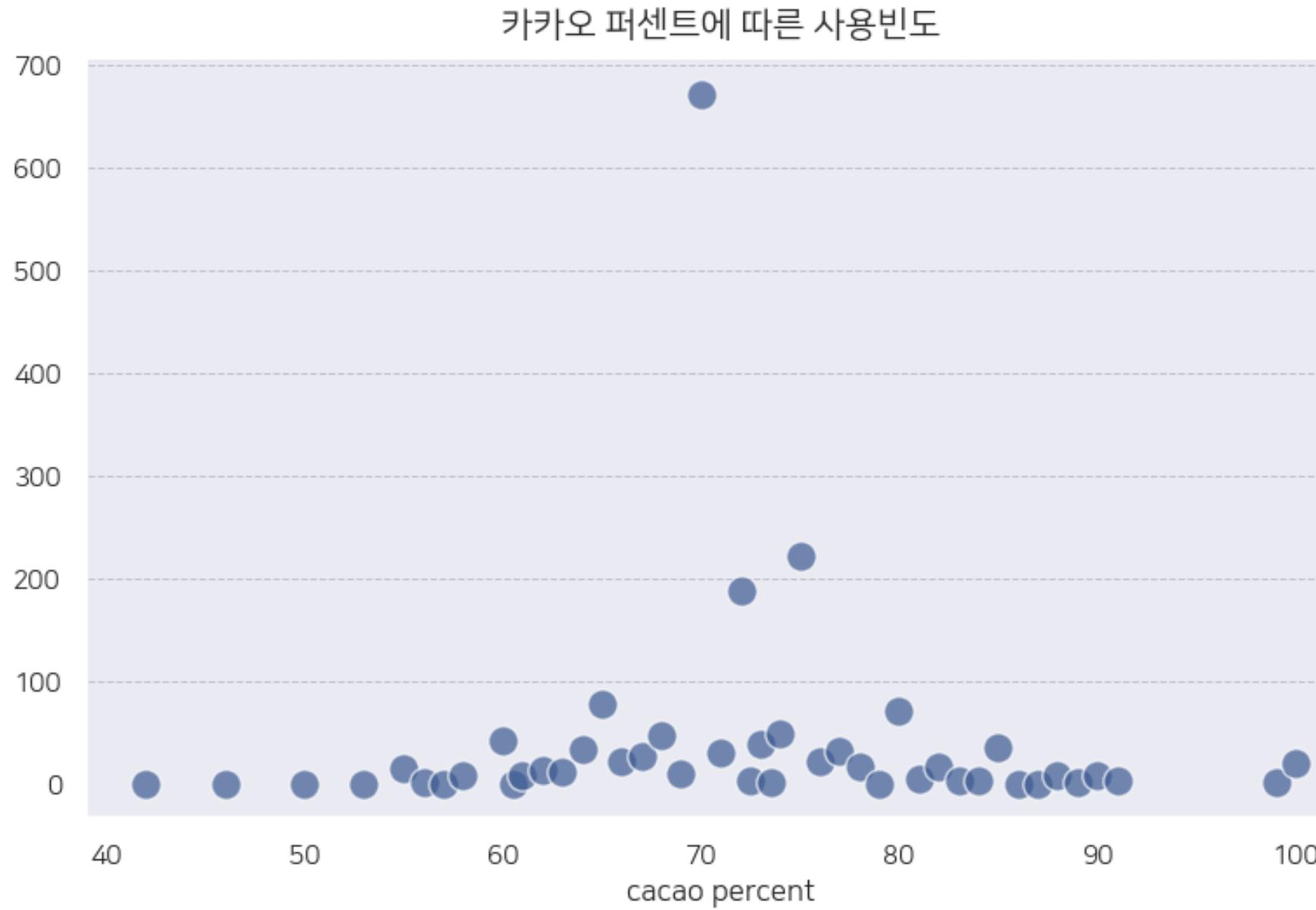
다음으로 카카오 사용률과 rating은 상관관계가 있는지
카카오 퍼센트와 rating의 상관성이 있는지 살펴보도록 하겠습니다.

- 카카오 퍼센트의 사용빈도를 나타내는 c_count feature
- 약한 양의 상관관계
- cacao percent 사용빈도가 높을수록 rating에 긍정적인 영향

Cacao Percent

- cacao percent 전반적으로 고르게 분포
- 보통 100 이하의 사용률
- cacao percent 70 --> 일반적 분포의 7배

- cacao_percent & rating corr = -0.19
- 카카오 함량을 높일수록 rating에 좋지 않은 영향

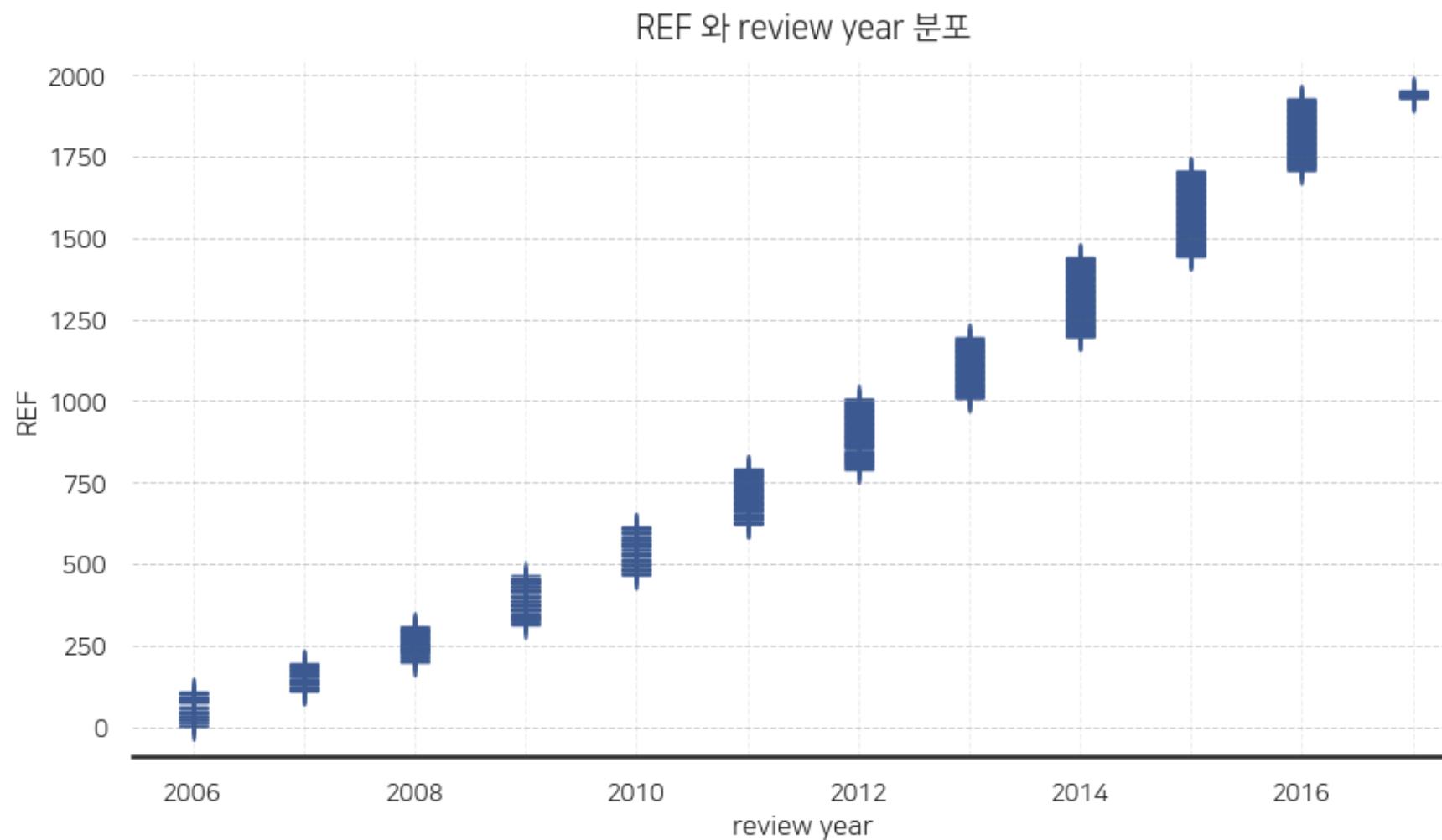


REF & Review Year

Data Science & Analysis

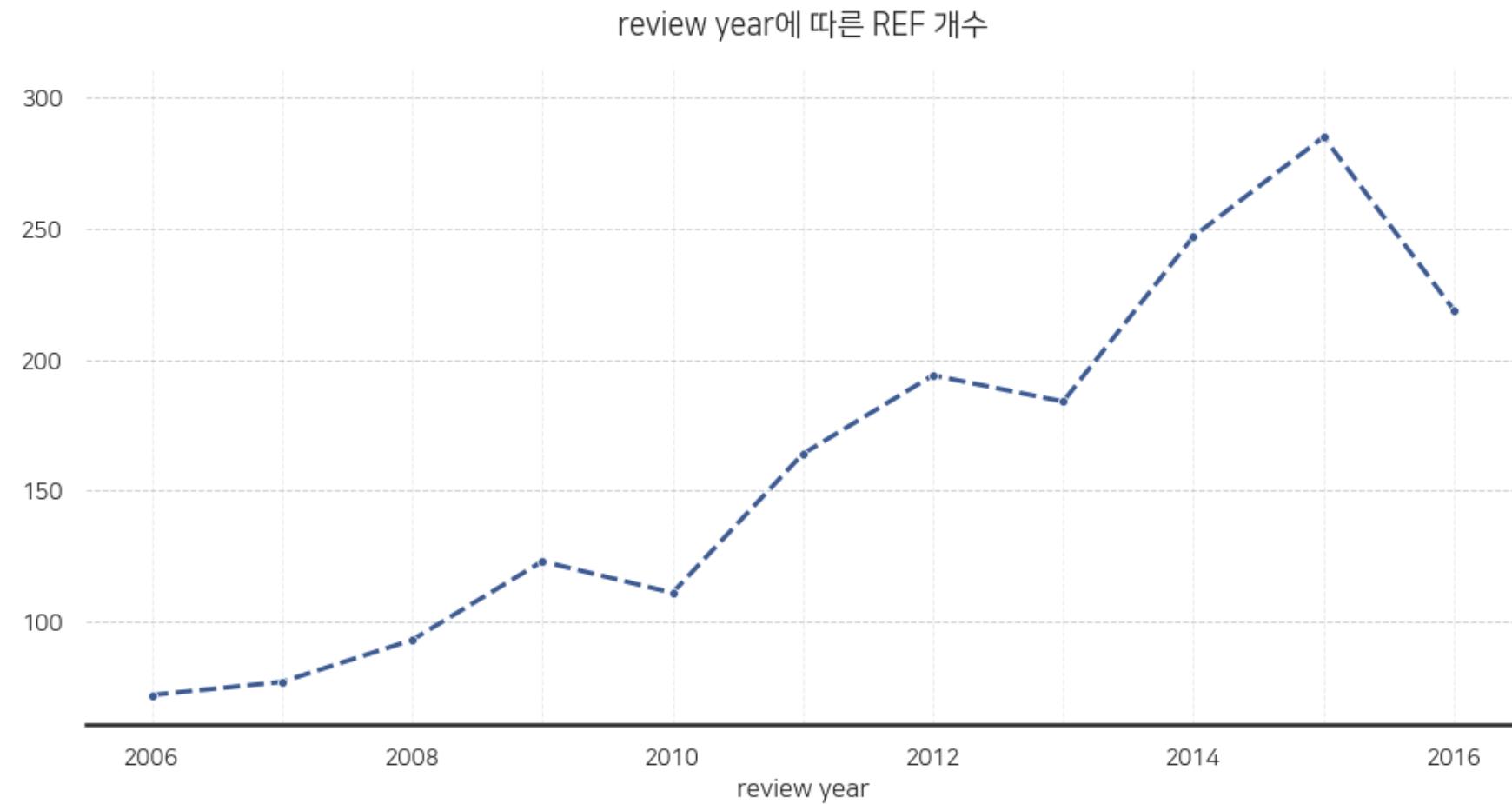
- review year에 따른 올림차순 정렬

- 리뷰가 작성 연도에 따라 REF가 증가하는 추세



- rating과 약한 양의 상관관계

	rating과 correlation
review year	0.100
REF	0.101

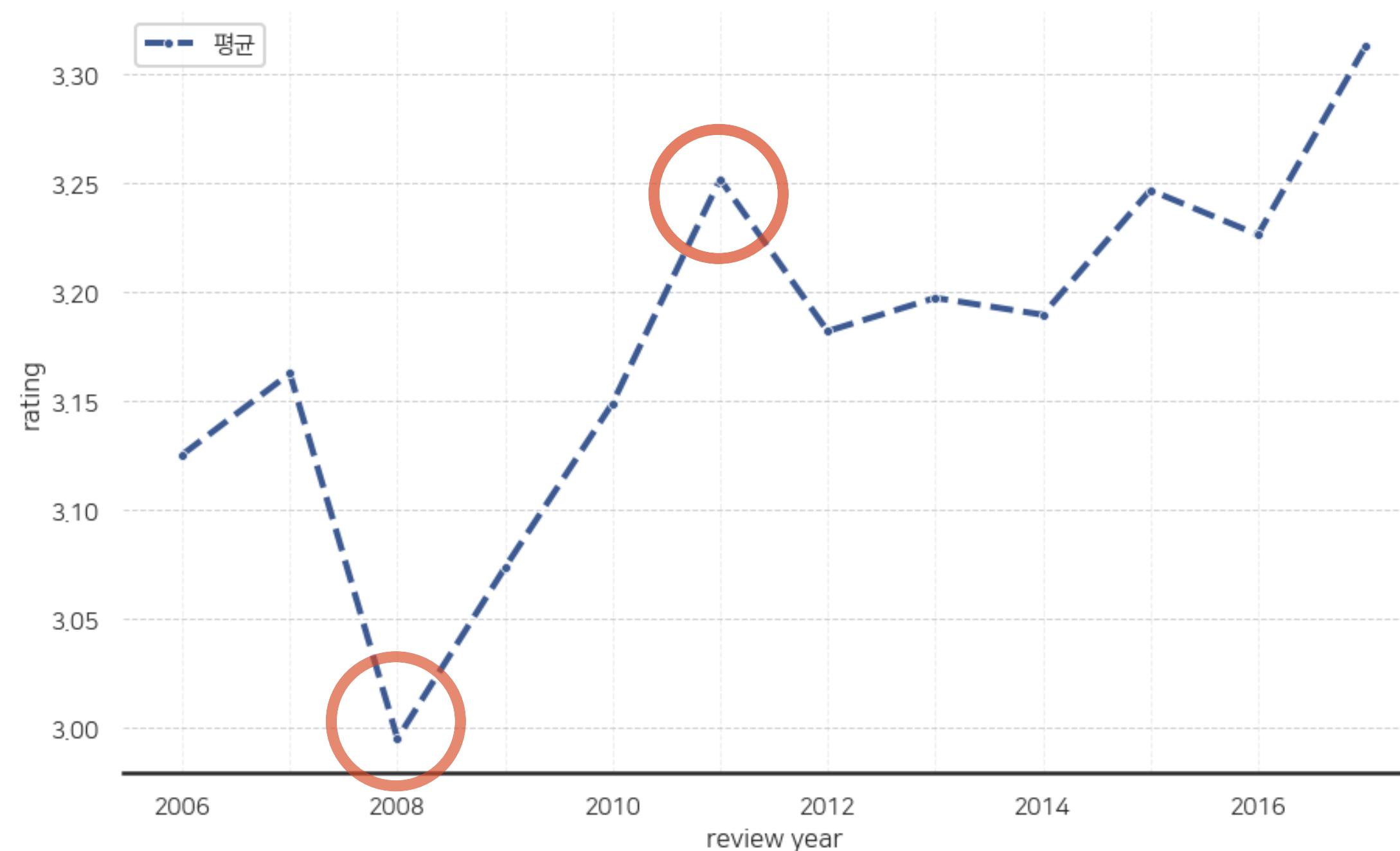


REF & Review Year

Data Science & Analysis

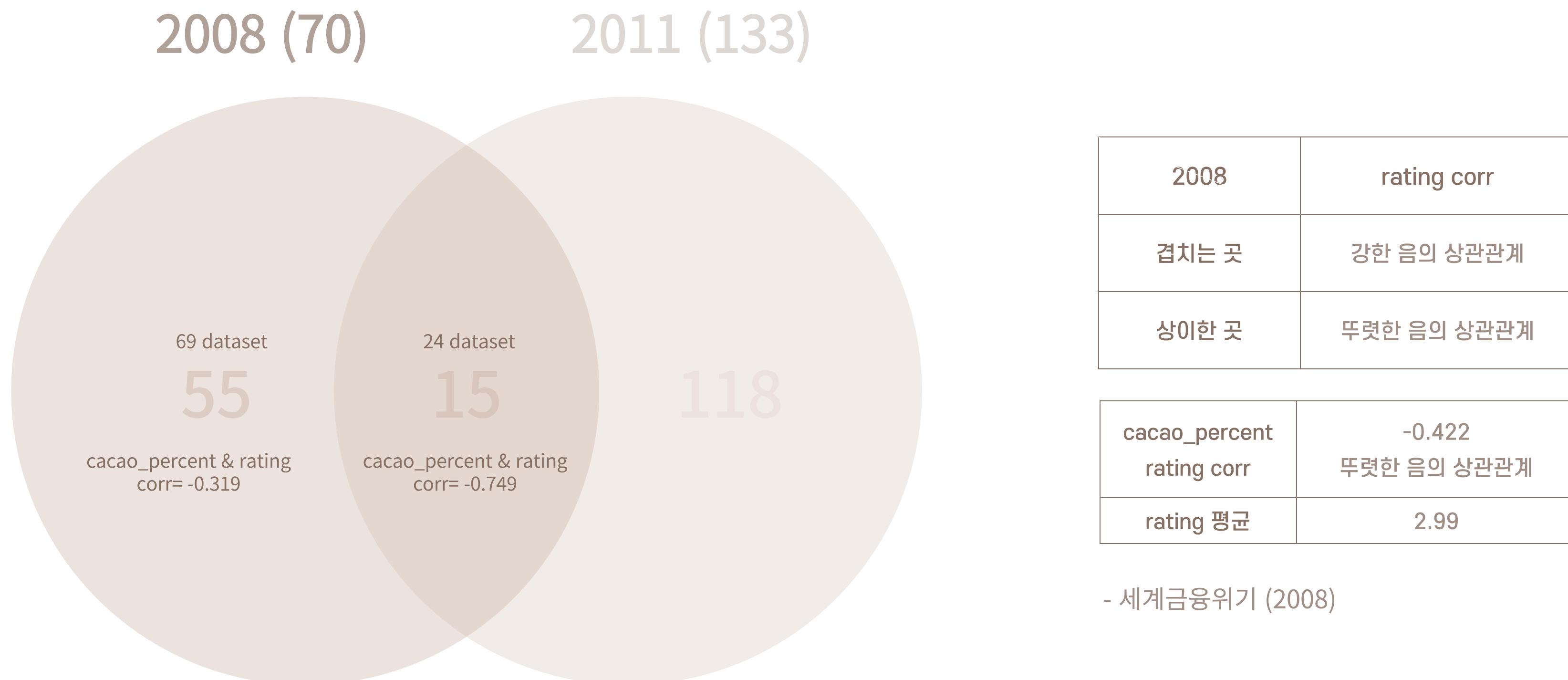
- 2008년 : 급격한 rating의 하락세
- 2011년 : rating 상승세의 정점

review year에 따른 rating 추이



2008 (98 dataset)

Data Science & Analysis

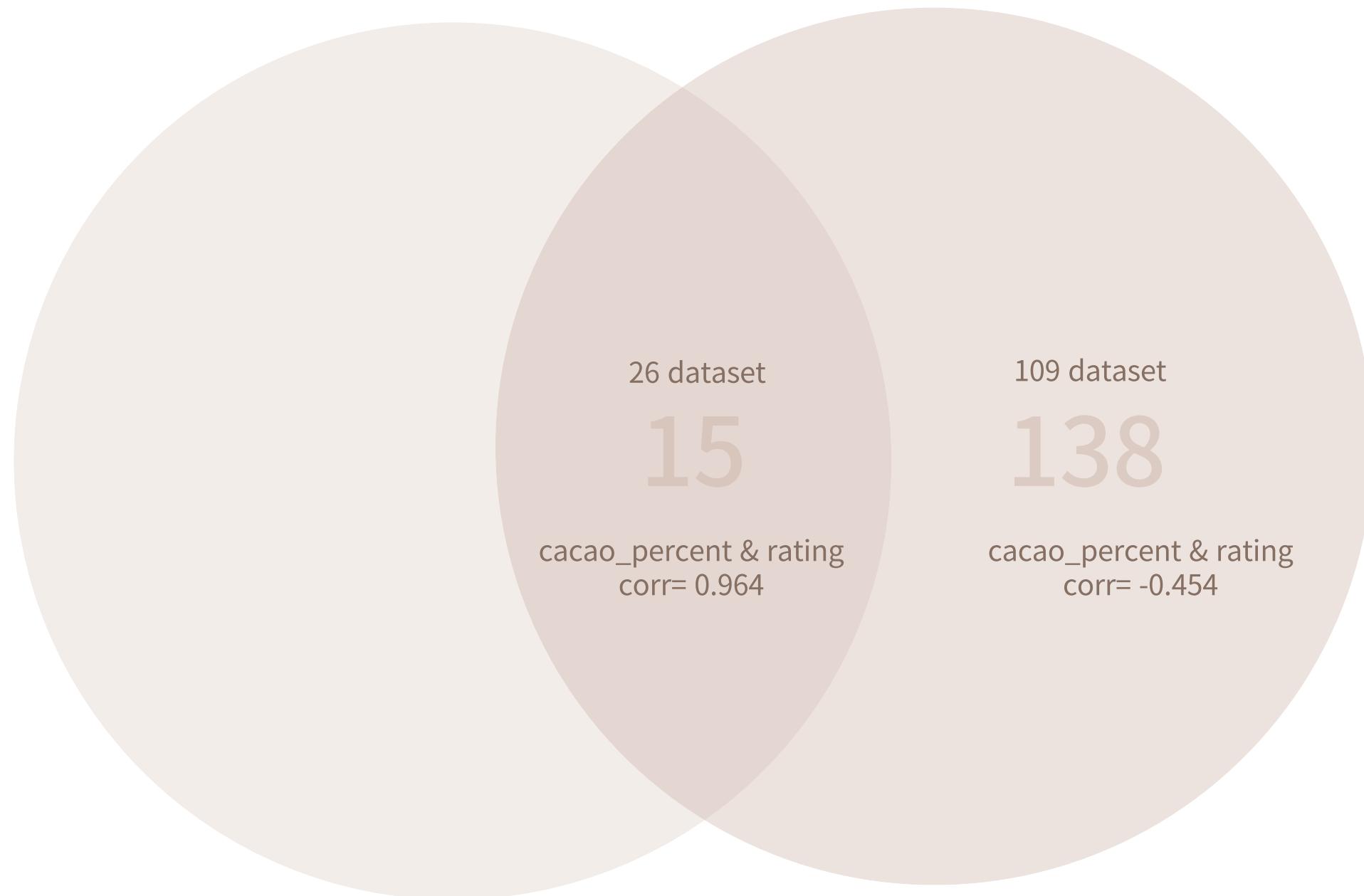


2011 (133 dataset)

Data Science & Analysis

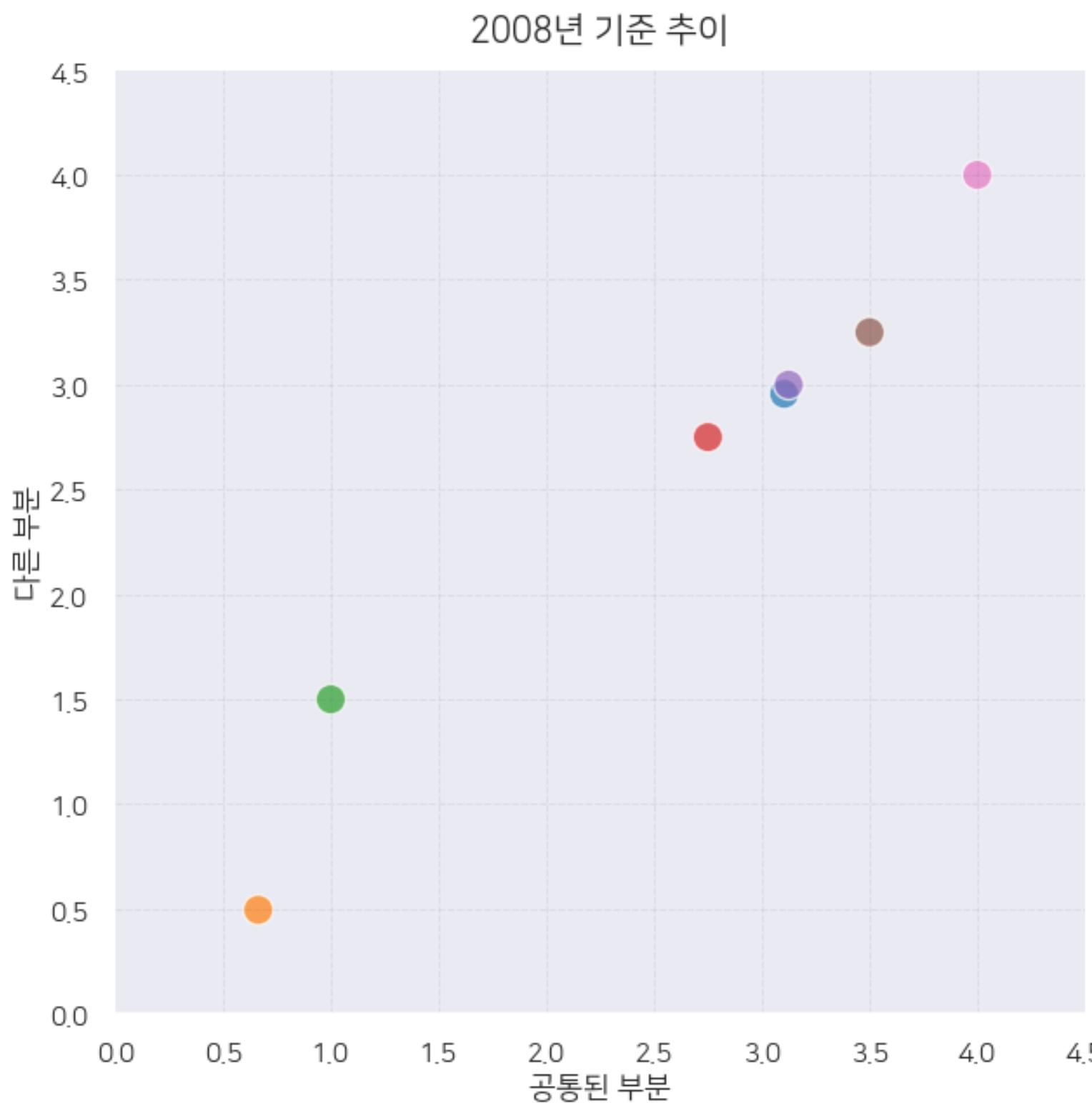
2008 (70)

2011 (133)



2011	rating corr
겹치는 곳	강한 양의 상관관계
상이한 곳	뚜렷한 음의 상관관계

cacao_percent	-0.02
rating corr	상관관계가 없다봐도 무방
rating 평균	3.25



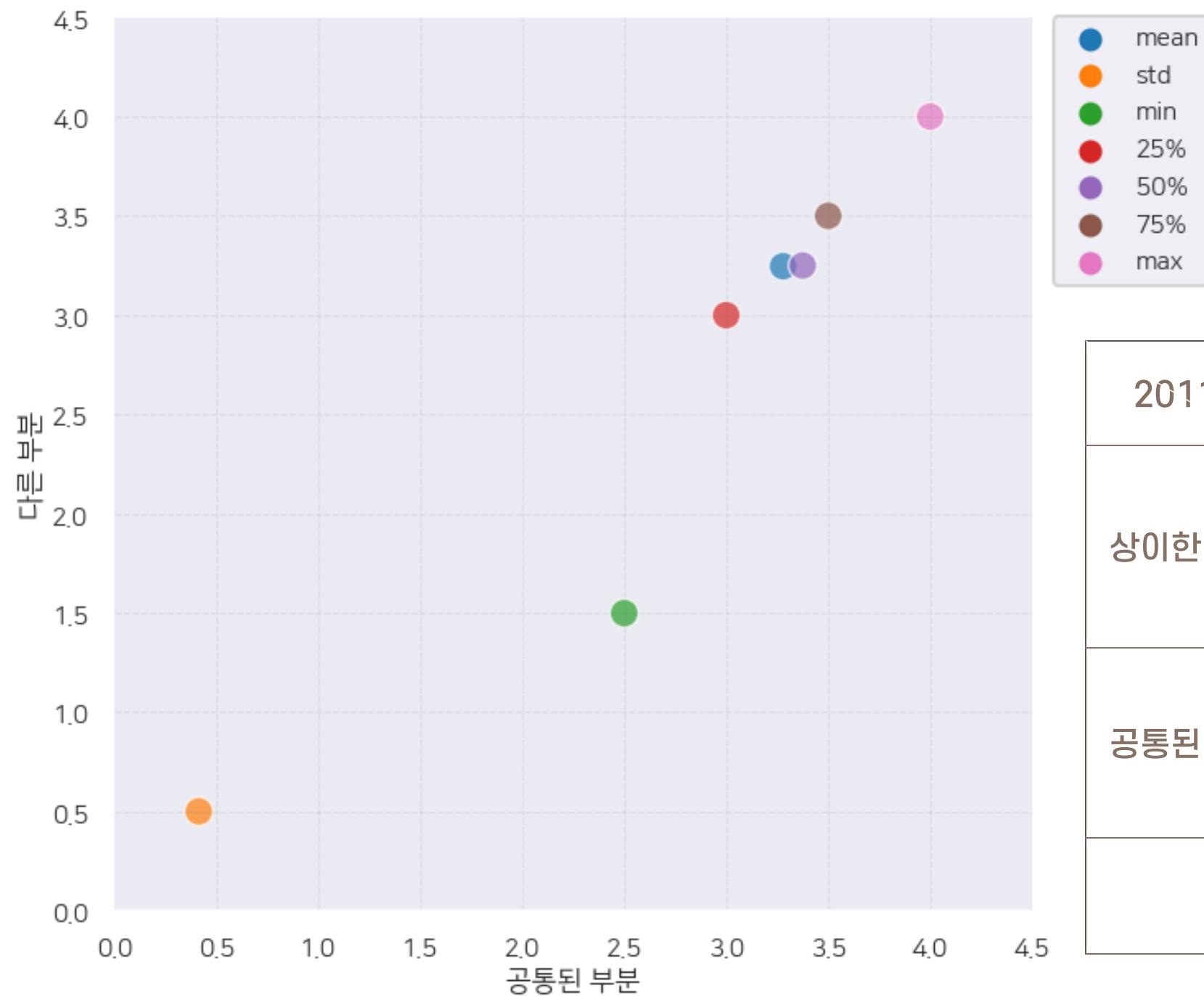
- mean
- std
- min
- 25%
- 50%
- 75%
- max

2008	평균	표준편차 (std)
상이한 곳	유사, 3.0 이상	0.49
공통된 곳	유사, 3.0 미만	0.66

- 최소값을 제외한 모든 영역에서 재배지가 공통된 부분이 rating 수치가 더 크게 나타남

--> bean 재배지가 rating 수치에 영향을 끼칠 수 있음을 의미

2011년 기준 추이

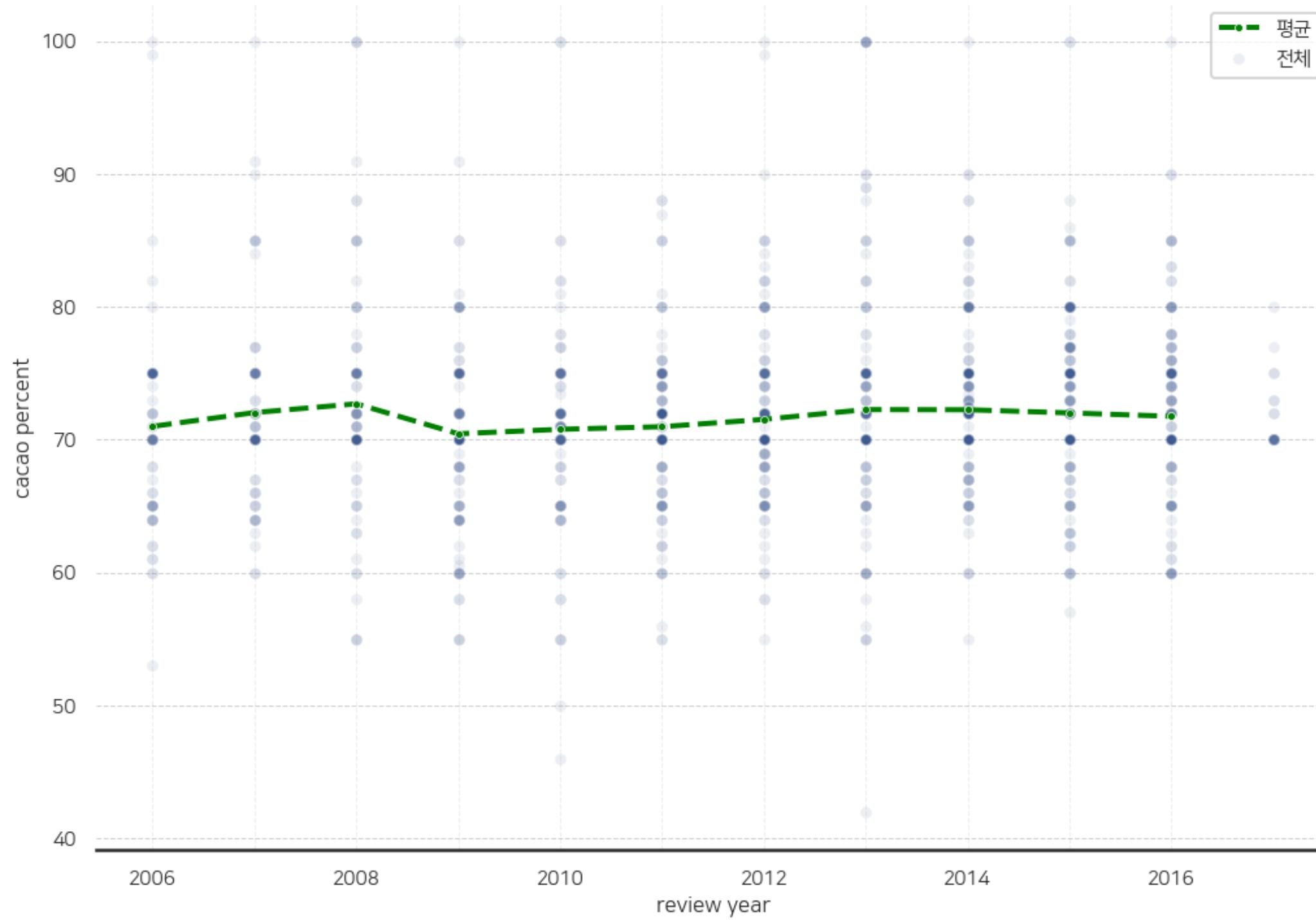


2011	평균	75%	50%	25%	최소	표준편차 (std)
상이한 곳	유사	동일	bigger	동일	1.5 rating	higher
공통된 곳			smaller		2.5 rating	

- 표준편차에 대해서는 상이한 부분이 조금 더 높은 수치를 보이고 있으나
전반적으로 재배지가 도일한 부분에서 다소 높은 수치를 보임

Review Year & Cacao_percent

review year에 따른 cacao percent 추이

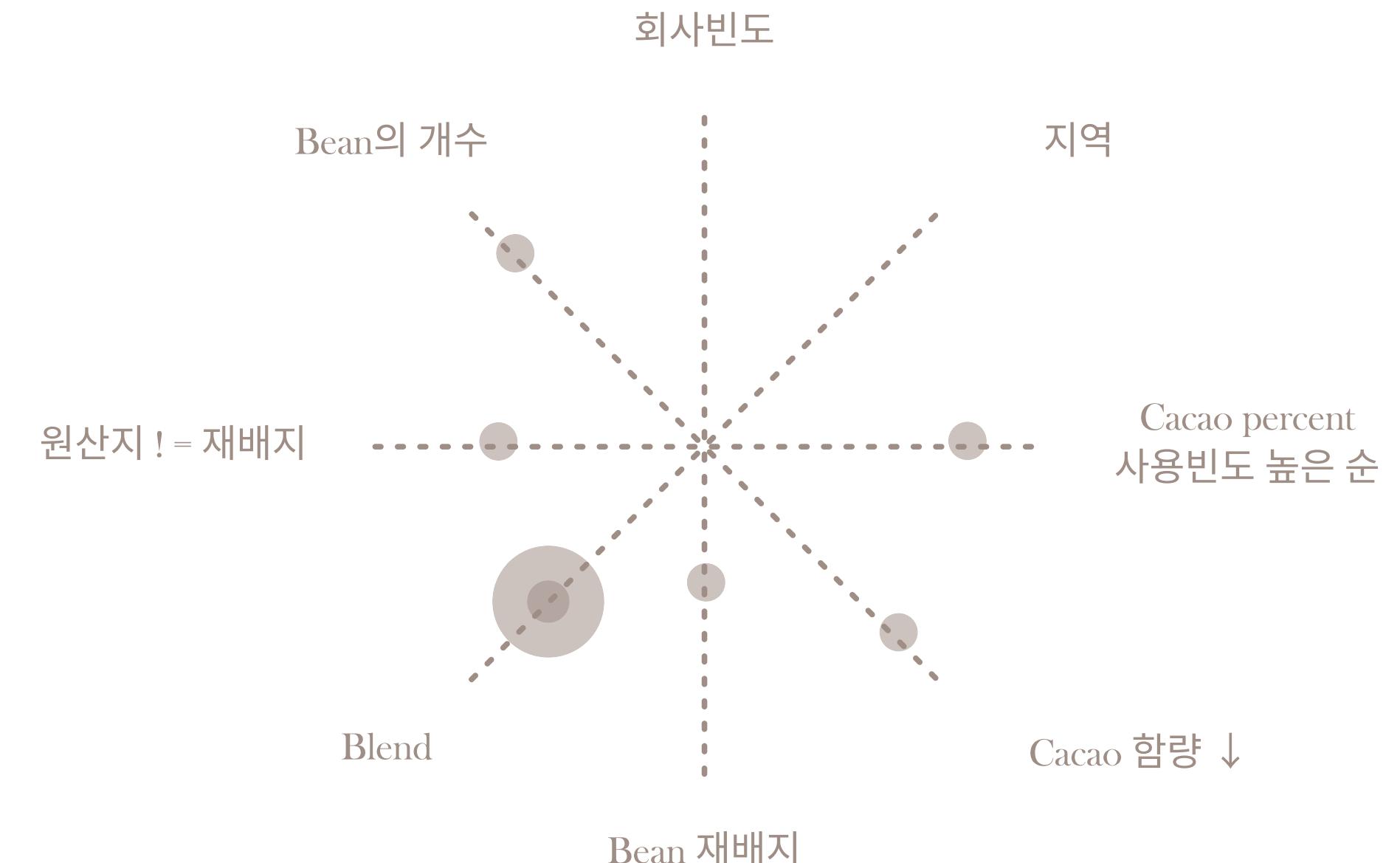


correlation	
전체 dataset	0.038 상관관계가 없다 봐도 무방
연도별 dataset	0.231 약한 양의 상관관계

Chocolate Data Strategy

- 회사 빈도 수는 rating에 영향을 끼친다 --> 많은 제품을 보유할 수록 rating에 긍정적인 영향을 끼칠 것
- Bean의 개수를 늘리는 것이 rating에 긍정적
- Blend 개수는 무관하나 Blend하는 것이 좋다.
- 출시한 지역 빈도가 rating에 영향을 주진 않는다.
- Bean Type 선호도는 rating에 영향을 주지 않는다. 어떠한 Bean 사용해도 무방하다.
- 특정 지역 재배지에서의 rating이 높을 것이므로, 앞서 살펴본 특정 지역의 재배지를 활용한다.
- 재배지와 원산지가 다른 것이 rating에 긍정적인 영향을 주었다.
- 카카오 함량 70%, 75%, 72% 순으로 많이 나왔으며, 카카오 퍼센트 사용 빈도가 높을 수록 rating에 긍정적이었다.
- 고함량 카카오를 타겟으로 한다면 70% - 75% - 72%- 80% 순을 저격한다.
- 카카오 함량이 높을수록 rating에 좋지 않은 영향을 끼치나, 매년 카카오 함량이 증가 추세이다.
사람들의 건강에 대한 관심이 증가하고 있으며, rating 또한 매년 증가 추세이다.

Data Strategy



Chocolate Strategy