코트라 데이터활용 빅데이터 분석 경진대회

기업 특성에 따른 잠재 파트너 매칭

팀장 장성민 jsm50660@gmail.com
팀원 한보혜 bohaehan@gmail.com
팀원 마민정 maminjeong3199@gmail.com

목차

1P

목차 1

대회주제
추가 데이터
데이터 시각화

목차 2

분석 방향성
파생변수 생성

목차 3

전처리
모델 선정 및 튜닝

목차 4

품목 기준 파트너 매칭
국가 기준 파트너 매칭

목차 5

기존 코트라 솔루션 개선
결론

대회주제

차년도 해당국가가 해당품목을 한국에서 **얼마나 수입**하는지 예측

중소 중견기업의 **해외진출 지원**을 위한 활용방안을 탐색

추가 데이터

UN Comtrade + World Bank

크롤링	경제력	소비력
TRADE_COUNTRYCD	GDP	구매력평가지수(PPP)
TRADE_HSCD_COUNTRYCD	GNI	연간 평균 소비자 구매지수
TRADE_HSCD		인구성장률
HSCD_name		수입가치

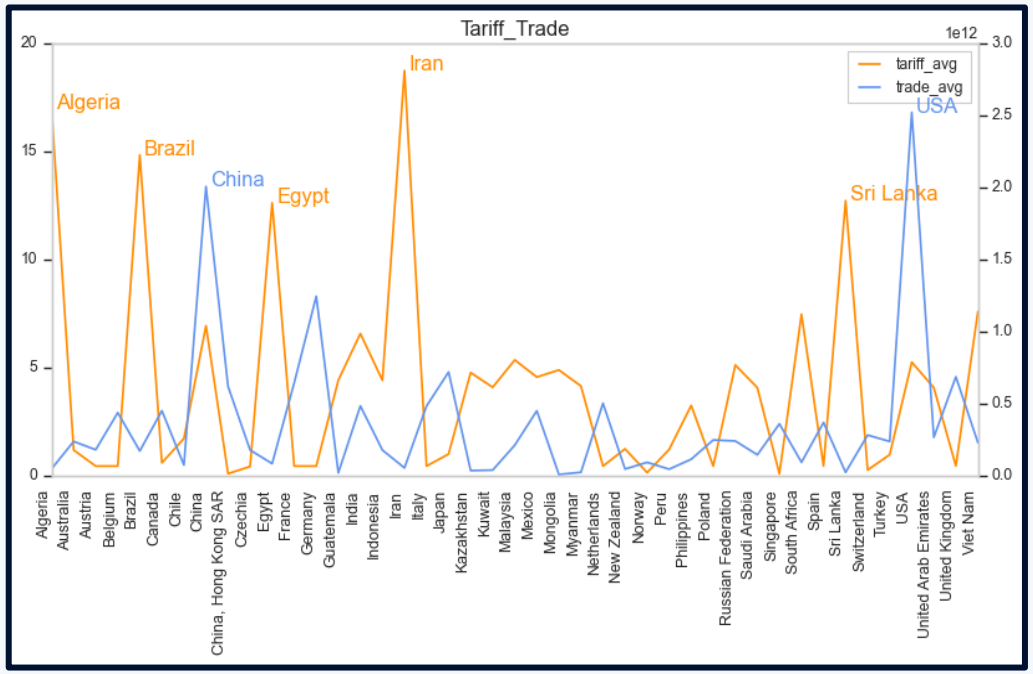
* 2012 ~ 2016년 총 5개년 데이터 수집
* 총 42개의 열 추가

제안배경

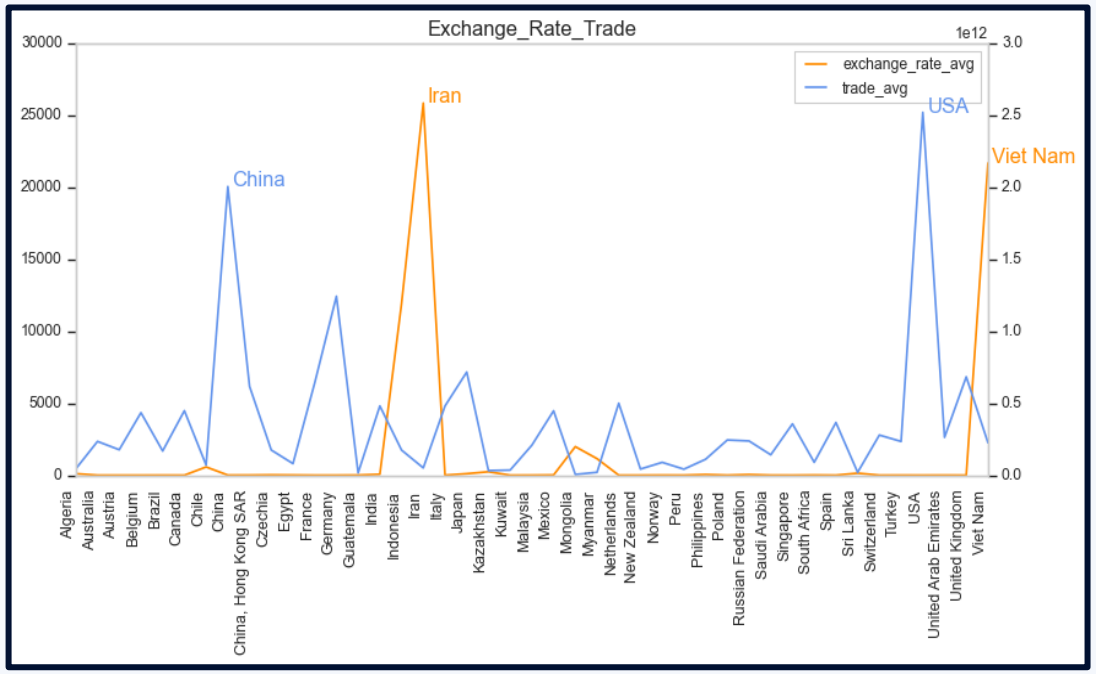
4P

데이터 시각화

[관세]



[환율]



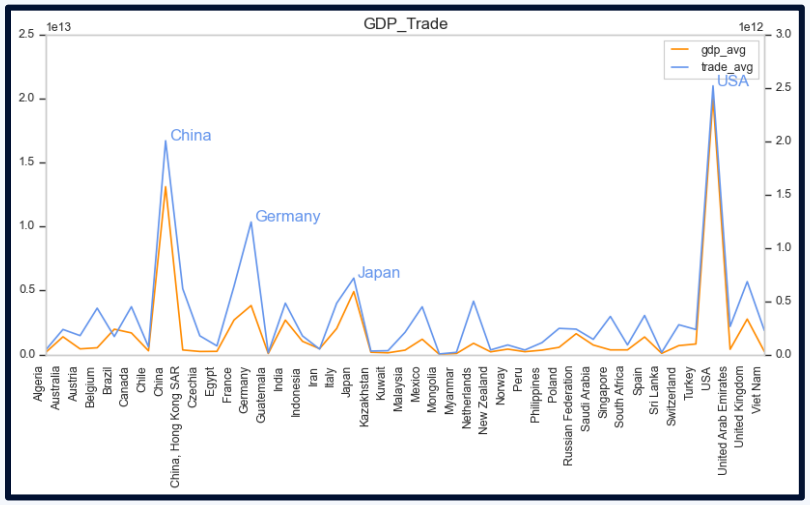
→ 관세와 환율은 수입금액과 대체로 **음의 상관관계**를 보임

제안배경

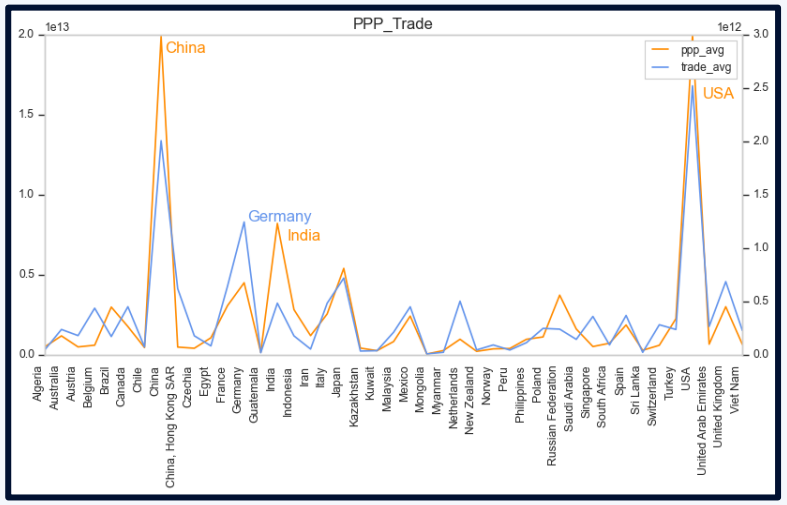
5P

데이터 시각화

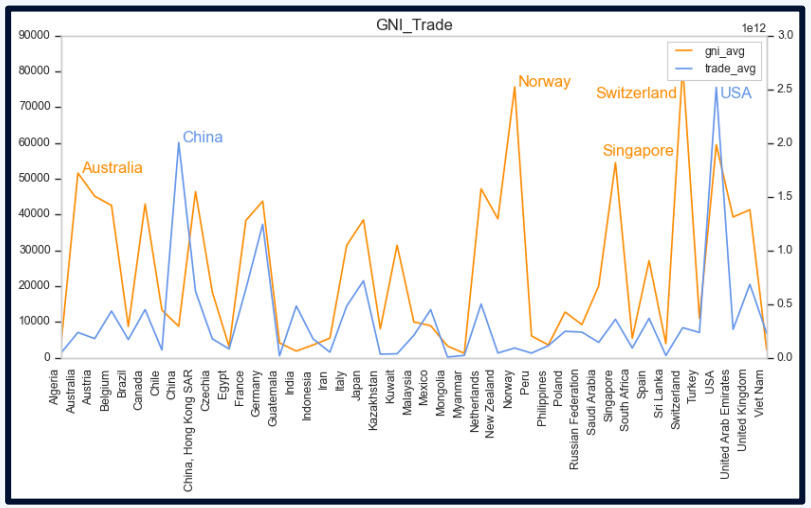
[GDP]



[PPP]

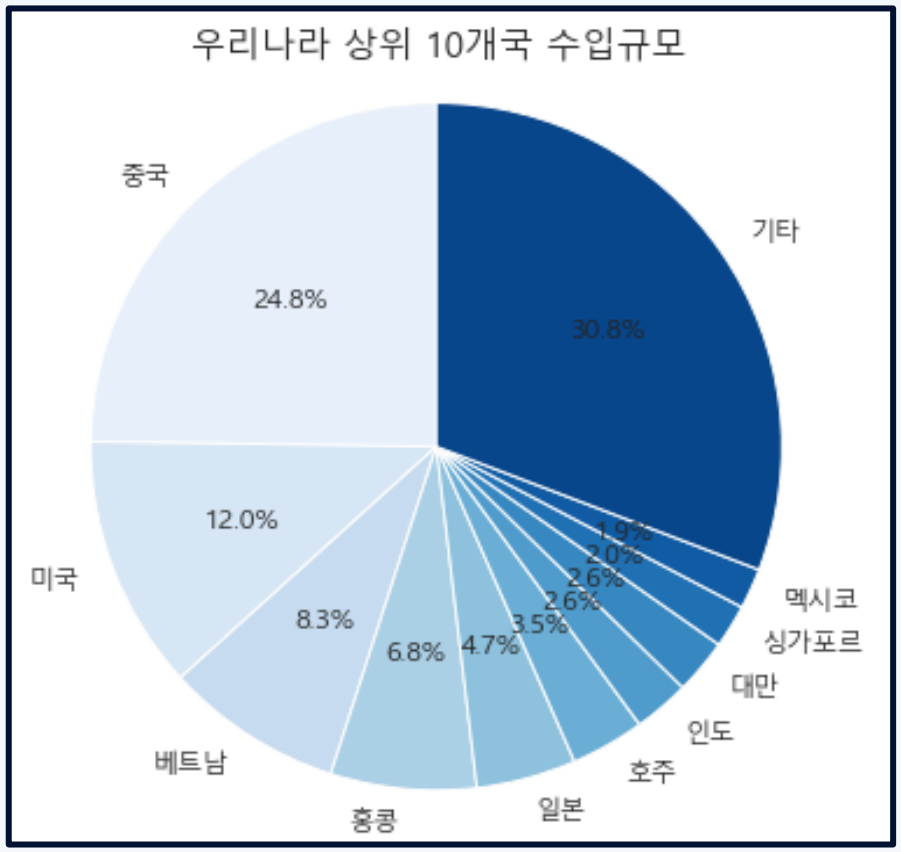


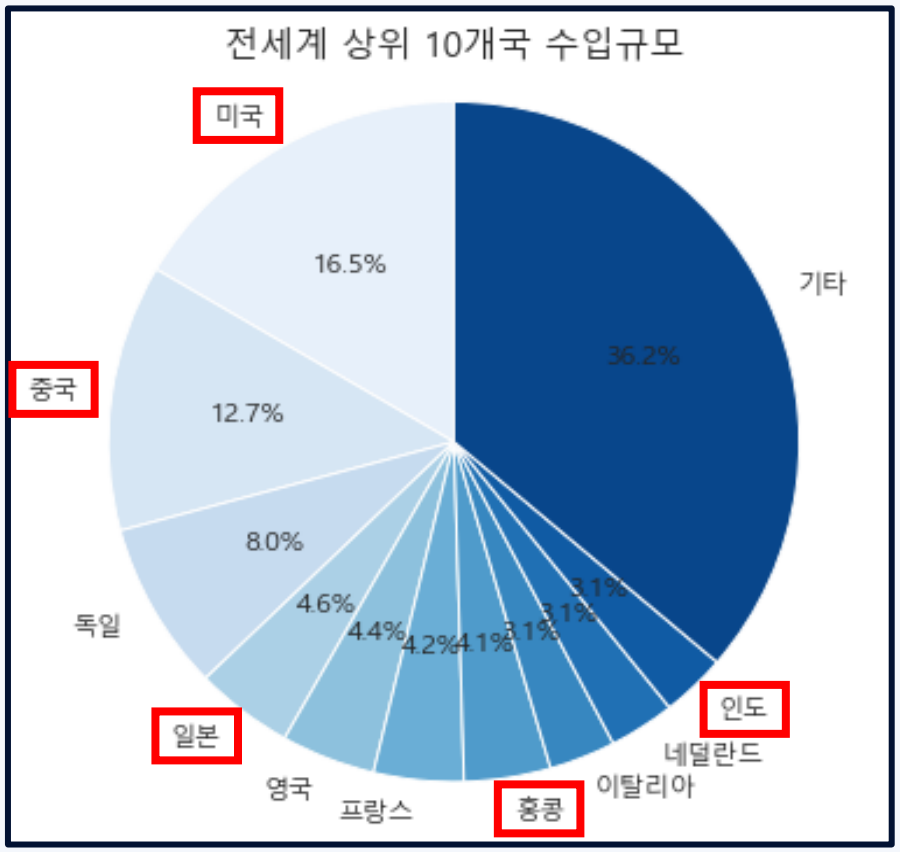
[GNI]



→ GDP, PPP, GNI는 수입금액과 대체로 양의 상관관계를 보임

분석 방향성





→ 국가의 특성을 잘 나타낼 수 있는 파생변수 필요!

분석 방향성

[상위 20개 품목]

	HSCD_name	전체수입비율
0	Plastics	4.8%
1	Rubber	4.2%
2	Iron or steel	3.6%
3	Vehicles	3.6%
4	Electrical apparatus	2.8%
5	Machinery	2.4%
6	Engines	1.6%
7	Pumps	1.6%
8	Instruments and apparatus	1.4%
9	Tools, hand	1.4%
10	Paper and paperboard	1.4%
11	Ignition or starting equipment	1.2%
12	Insulated electric conductors	1.0%
13	Bearings	1.0%
14	Lamps	1.0%
15	Cosmetic and toilet preparations	1.0%
16	Furniture	1.0%
17	Units of automatic data processing machines	0.8%
18	Signalling apparatus	0.8%
19	Valves	0.8%

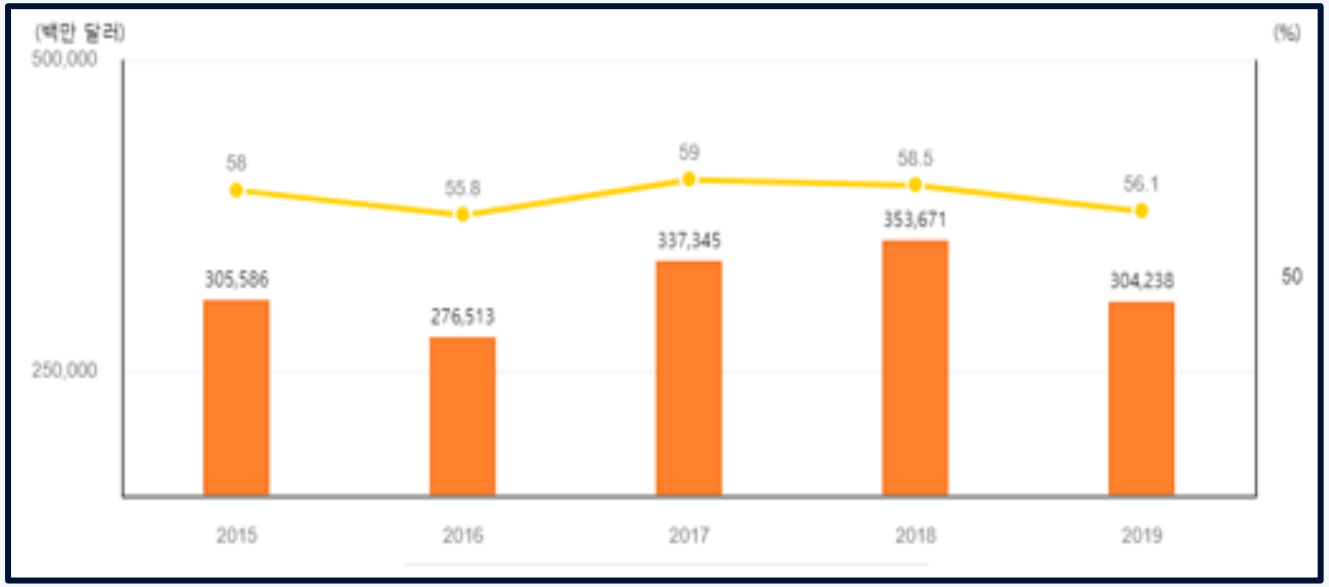
*상품과 비율은 상위 20위까지만 제시함.

→ 전체의 약 50% 차지

[우리나라 10대 수출품목]

2017	
품목명	금액
반도체	97,937
선박해양구조물 및 부품	42,182
자동차	41,690
석유제품	35,037
평판디스플레이 및 센서	27,543
자동차부품	23,134
무선통신기기	22,099
합성수지	20,436
철강판	18,111
컴퓨터	9,177

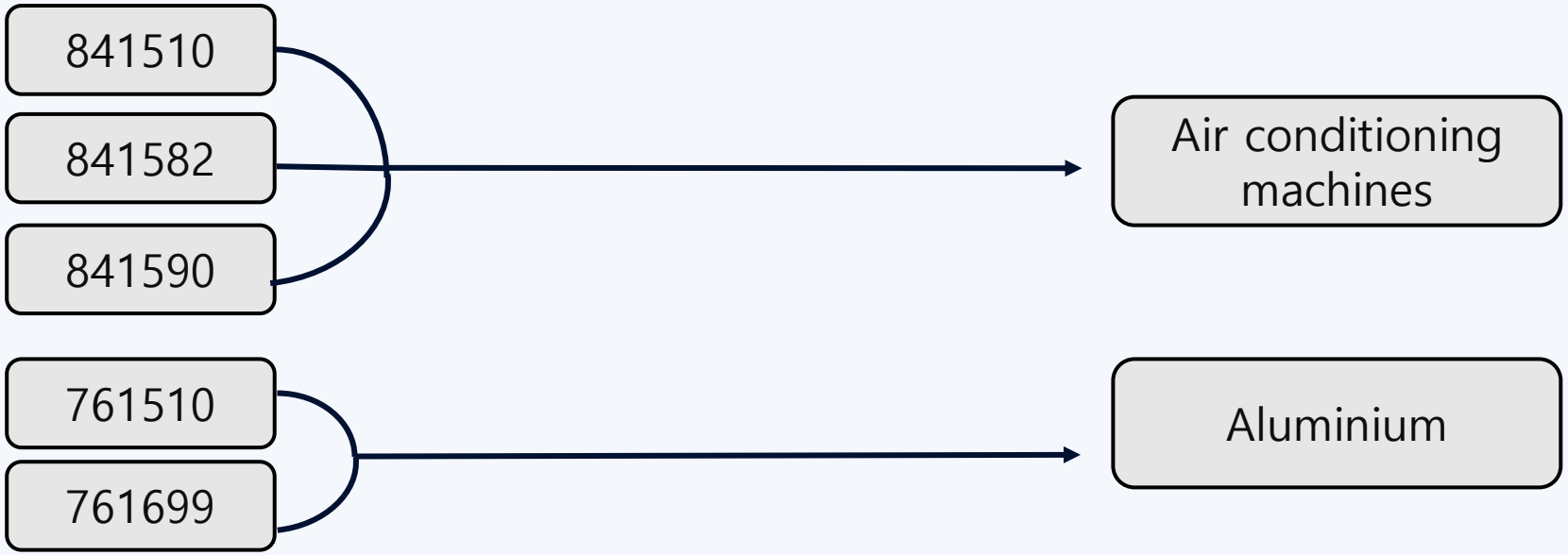
[우리나라 총 수출액 중
10대 수출품목 비중 변화추이]



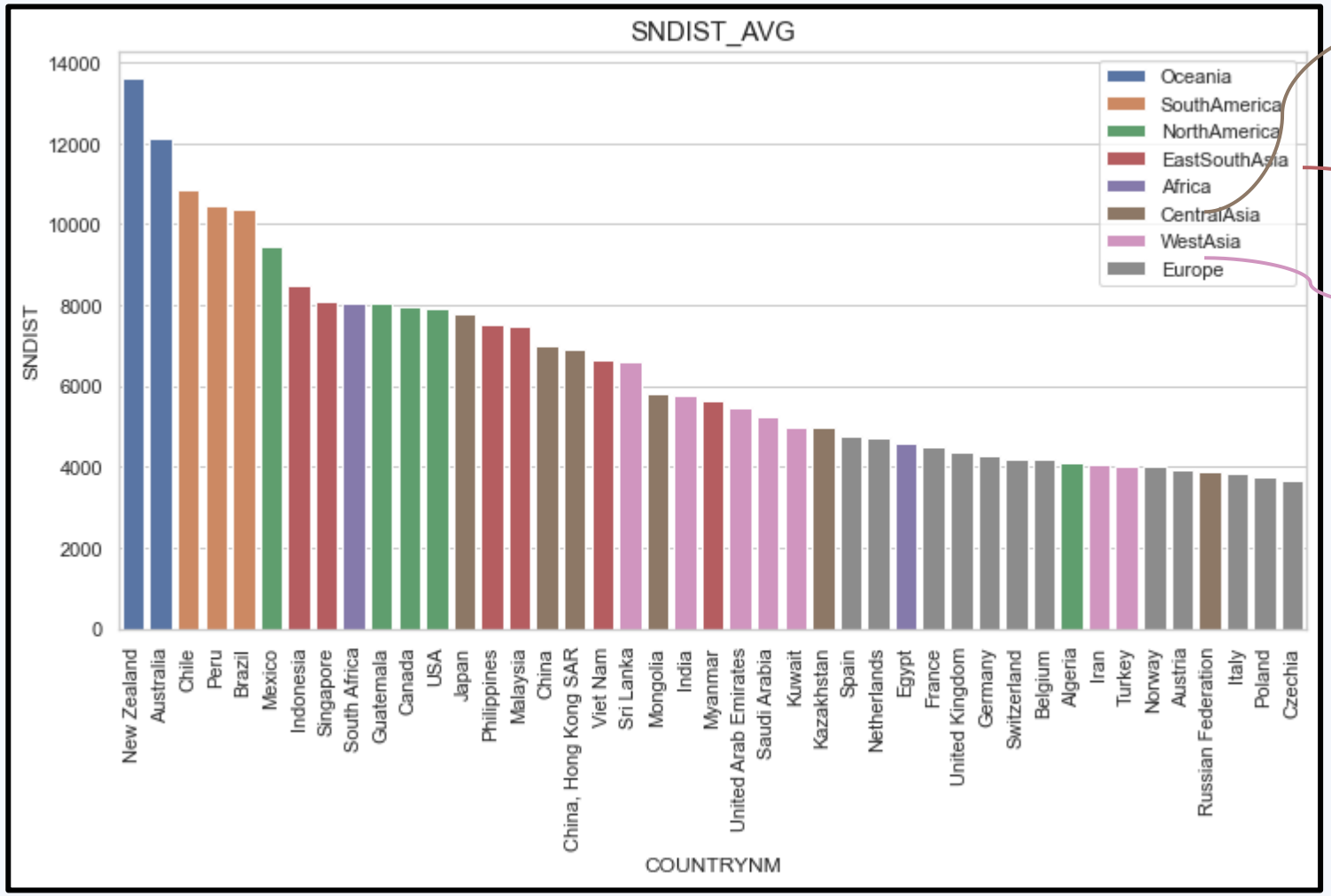
분석 방향성

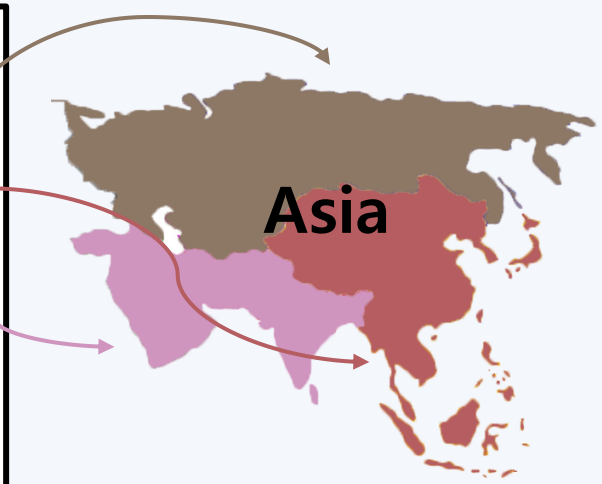
HSCD
500 종류

HSCD_name
239 종류



분석 방향성





→ 대륙별 수입특성의 존재 파악으로 관련 파생변수 필요!

파생변수 생성

기본	국가	품목	대륙	10대 품목	합계
52개	39개	40개	41개	25개	197개

One-Hot-Encoding		
HSCD, HSCD_name	COUNTRYNM	CONTINENT
739개	43개	8개

→ 총 987개 피처 생성

전처리

Missing Value

TARIFF_AVG	SNDIST	PA_NUS_PCRF	TRADE_COUNTRYCD TRADE_HSCD_COUNTRYCD
총 388개 중 90개의 수치가 90%를 차지 → 90개 수치들에 대한 **가중평균**으로 대체	해당 나라의 **평균값**으로 대체	7개의 소수의 국가에만 결측치 존재 → **2021기준 환율**로 대체	각 열의 **평균값**으로 대체 단, 모든 연도가 NaN값이라 파생변수를 만들 수 없을 경우 **0으로** 대체

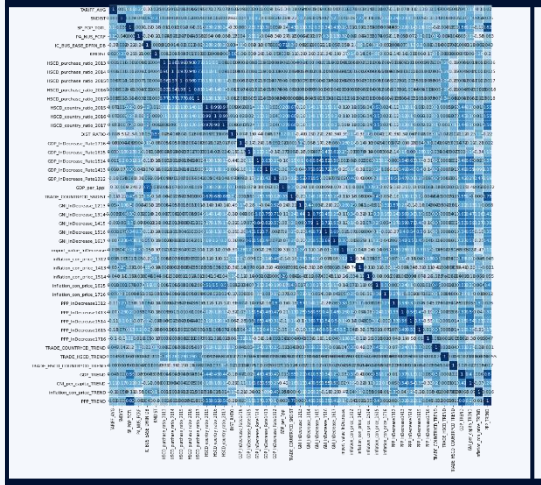
→ 각 열의 특성에 맞게 평균, 가중평균, 대체값 등으로 결측치를 대체함!

전처리

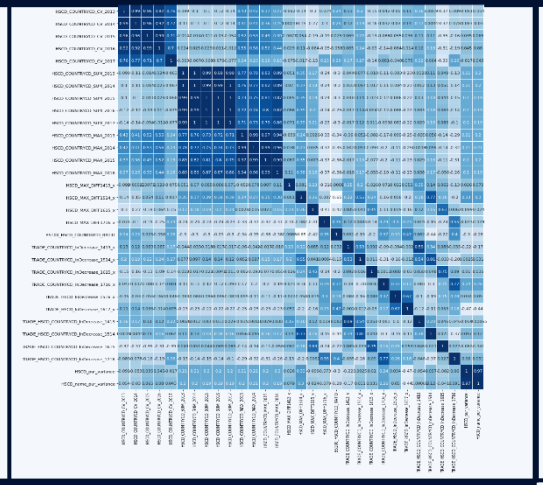
Correlation

기본 피처
0개 제거
국가기준 피처
6개 제거
대륙기준 피처
8개 제거
품목기준 피처
5개 제거
10대 품목 피처
0개 제거

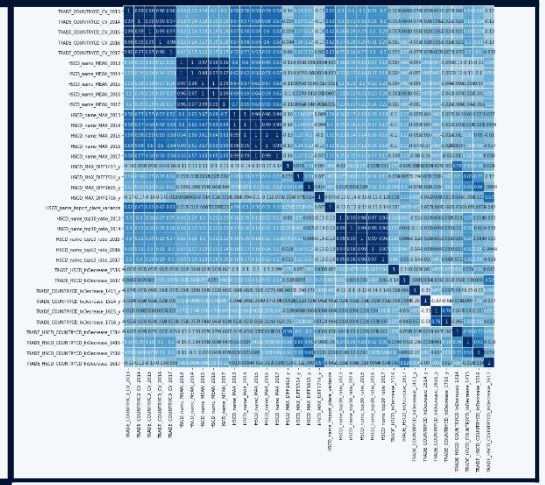
기본 피처



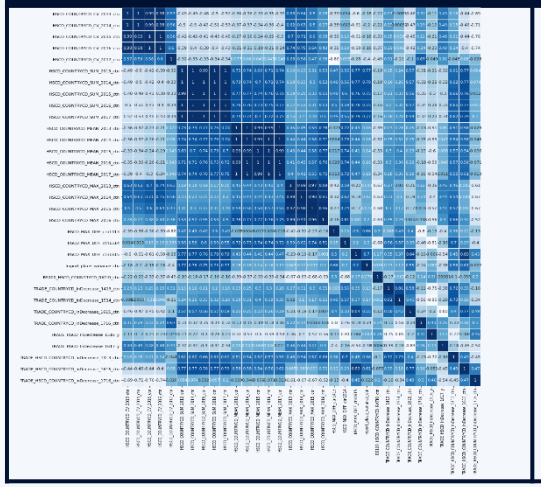
국가기준 피처



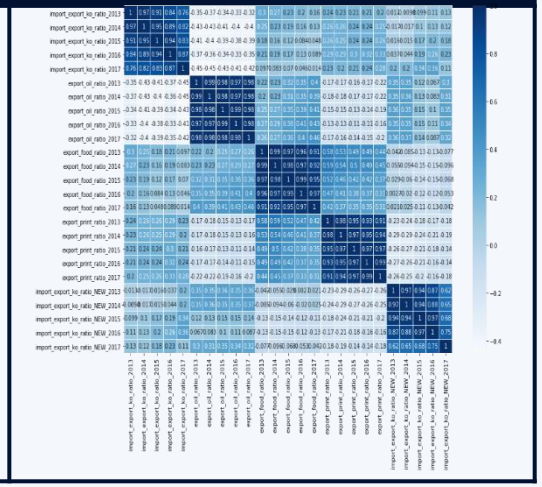
대륙기준 피처



품목기준 피처



10대 품목 피처

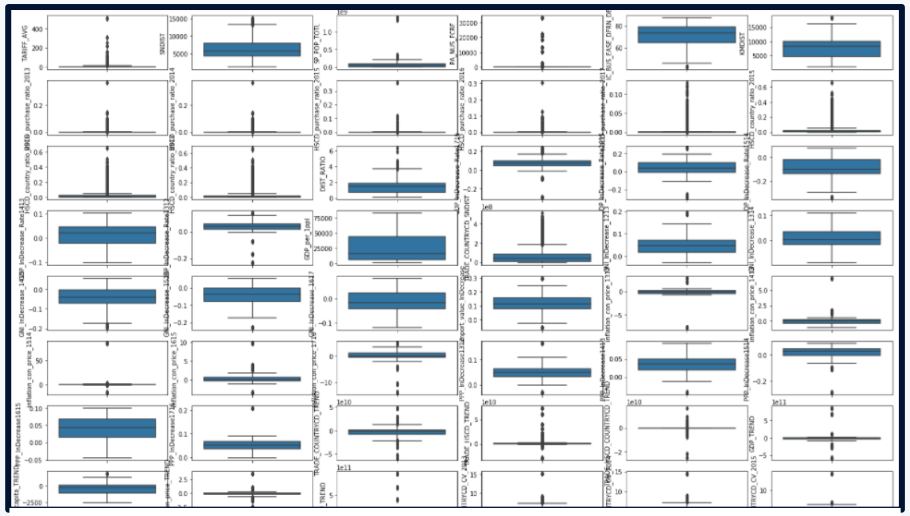


* 대상) one-hot-encoding 제외 197개 피처
* 기준)상관성 90% 이상
* 연도별로 생성된 피처의 상관성은 무시

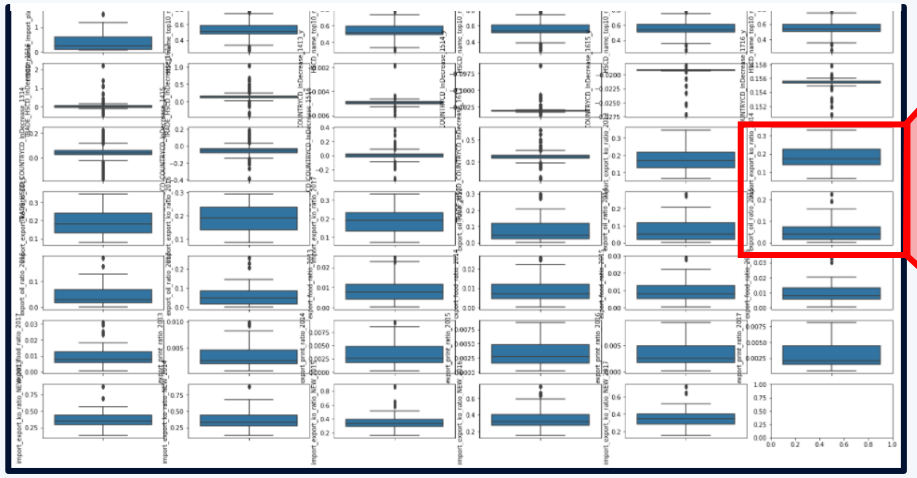
→ 총 19개 피처 제거

전처리

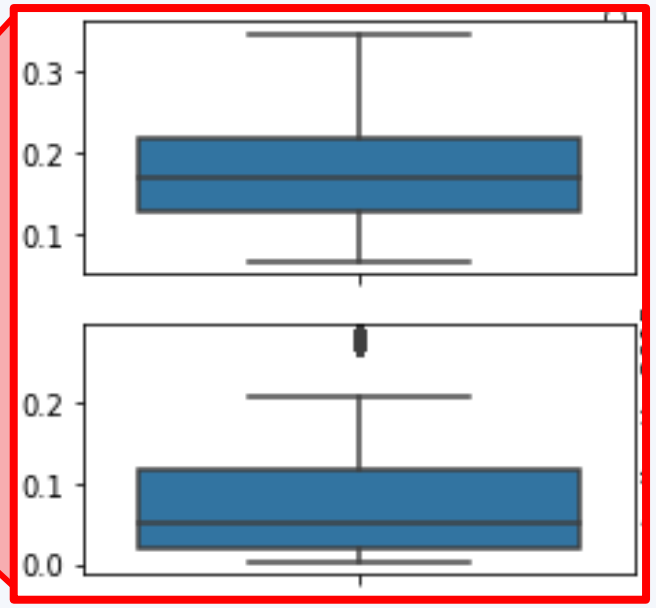
Outlier



(중략)



*** y축 스케일이 매우 작아 이상치로 간주하지 않음**



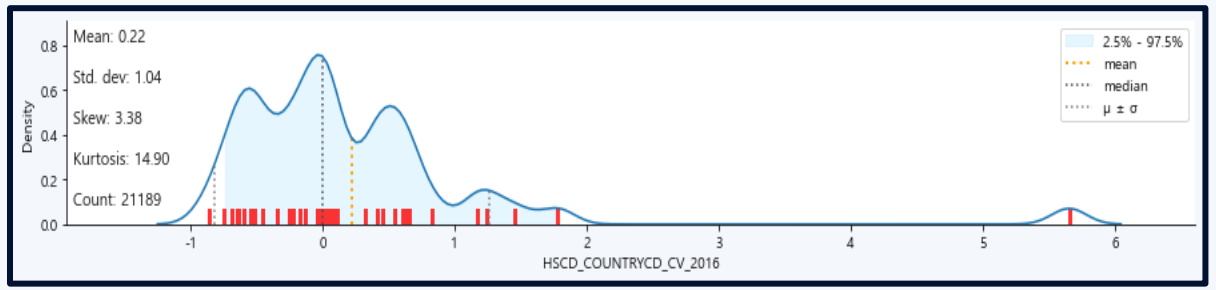
분석 및 모델링

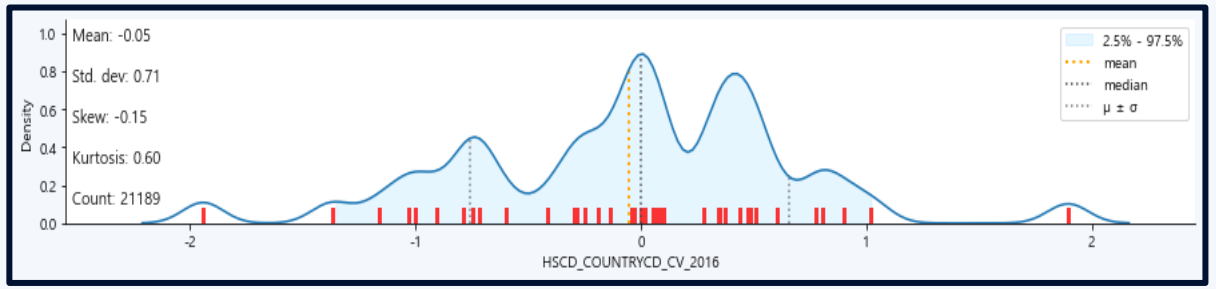
14P

전처리

Scaling

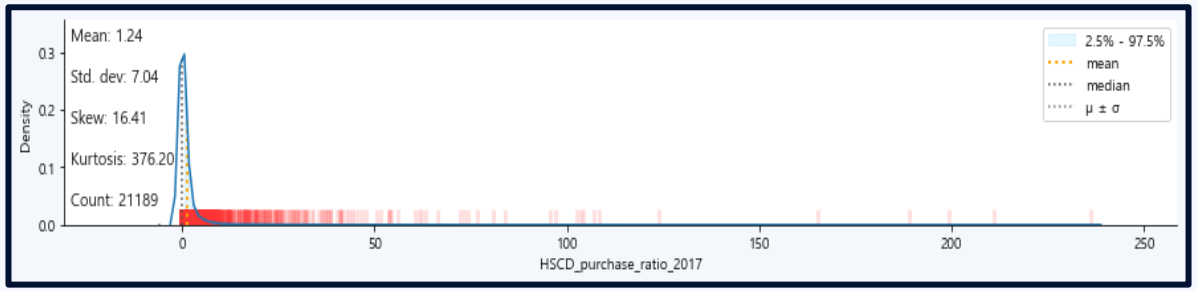
[1번의 log1p 적용 예]

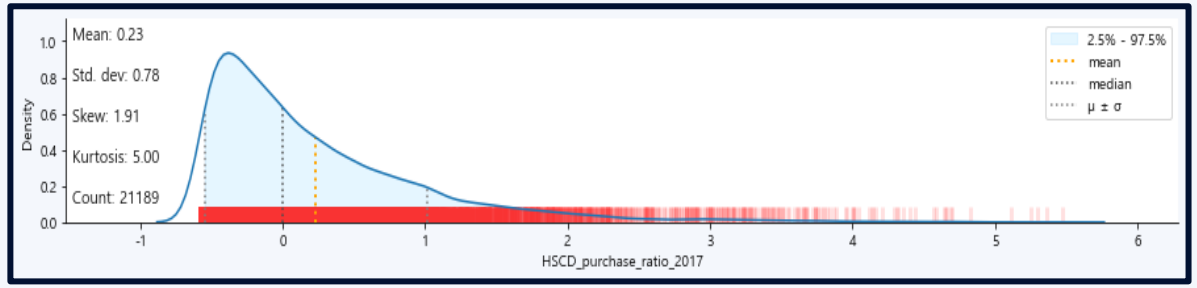


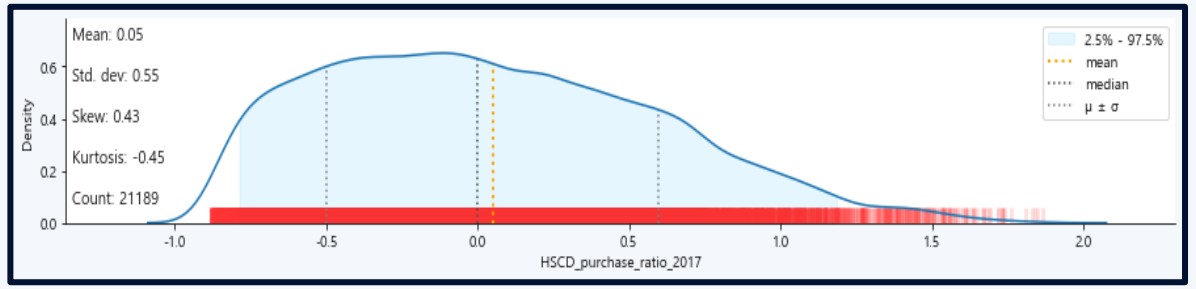


* 이상치에 덜 민감한 **RobustScaler**를 사용한 후 **np.log1p**를 사용해 치우침을 완화

[2번의 log1p 적용 예]







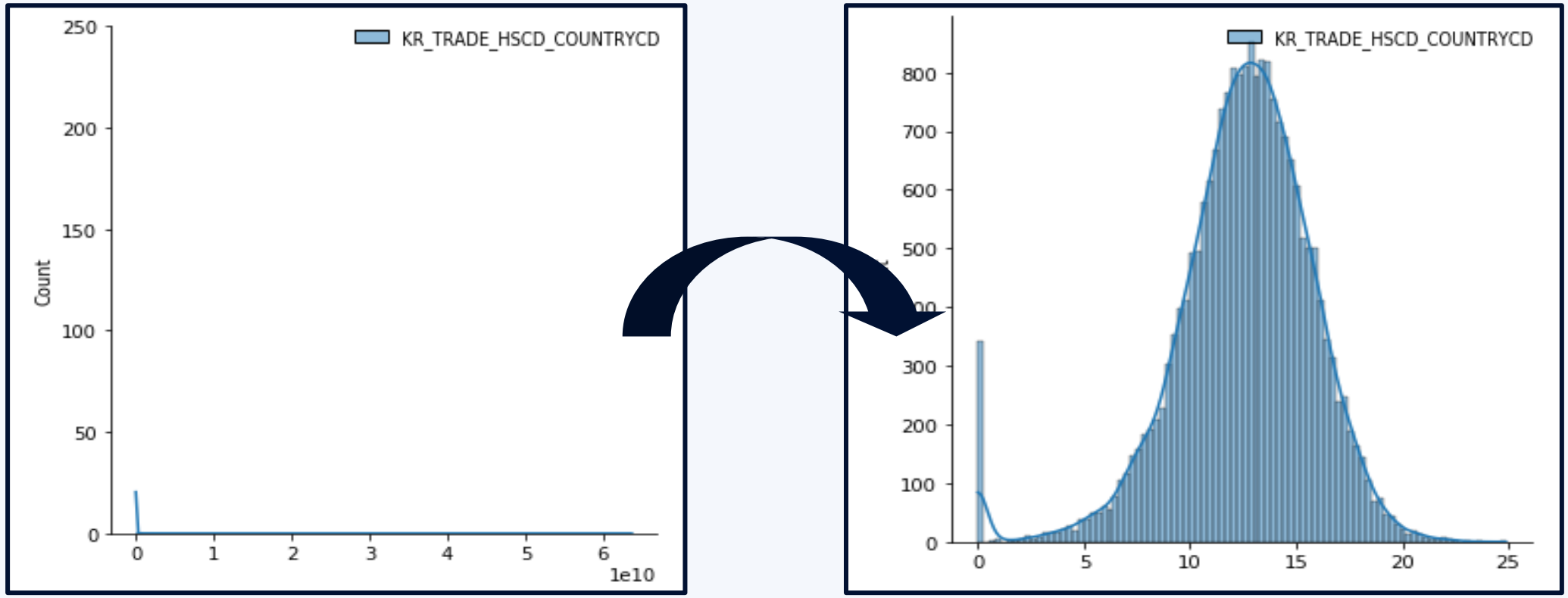
분석 및 모델링

15P

전처리

Log Scaling

[종속변수 로그화]



전처리

PCA

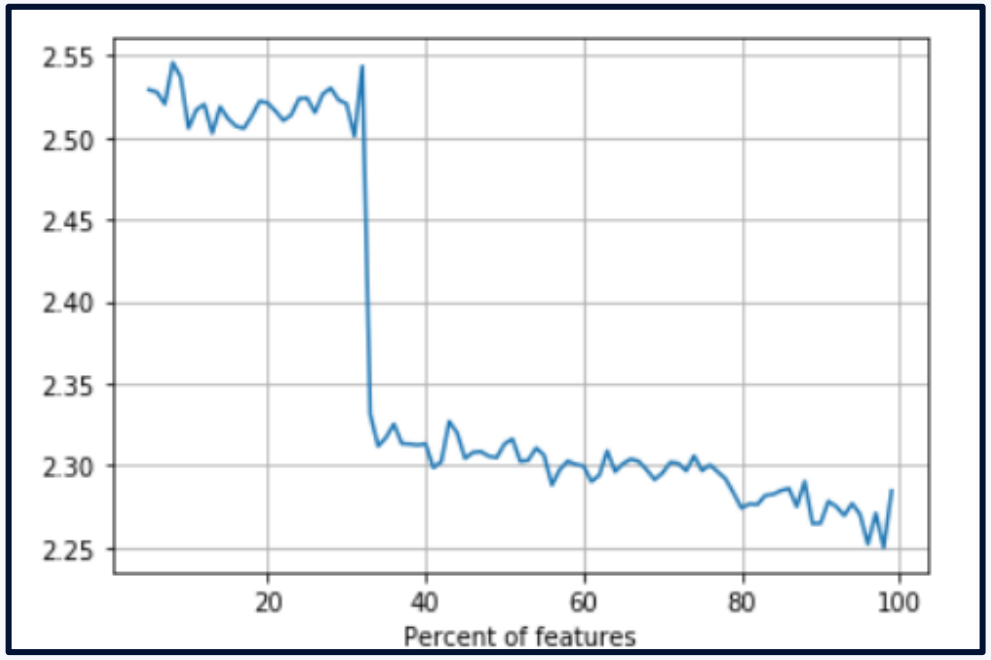
One-Hot-Encoding을 통해 생성된 790개의 피처에 대해 주성분 분석

→ 90%의 설명력을 가지는 351개의 피처로 차원축소

전처리

Feature Selection

전처리 Numeric **178개** + 주성분분석 **351개** = **총 529개**



FS 사용모델) **LGBM Regressor**

→ **98%**의 피처를 사용하여 모델링

모델 선정 및 튜닝

사용모델

- XGB
- LGBM
- CatBoost
- Extra tree

* 평가지표 : RMSE
* 성능 검증 :
  - Data Split
  - k-fold cross validation
* 모델 튜닝 : Bayesian Optimization

Average Ensemble

- LGBM
- CatBoost
- 멱평균

Stacking

- 4가지 모델 모두 사용
- Meta Model) voting, LGBM

Seed Ensemble

- LGBM Seed값을 3번 변경해 예측값 생성
- cross validation을 통해 앙상블 진행

모델 선정 및 튜닝

모델	XGB	LGBM	CAT	Extra	Avg.E	Stk(LGBM)	Stk(Voting)	Seed.E
기본성능	2.037	1.959	1.927	2.075	-	-	-	-
최종성능	1.958	1.917	1.927	2.067	2.003	2.017	2.012	1.934/1.919/1.925

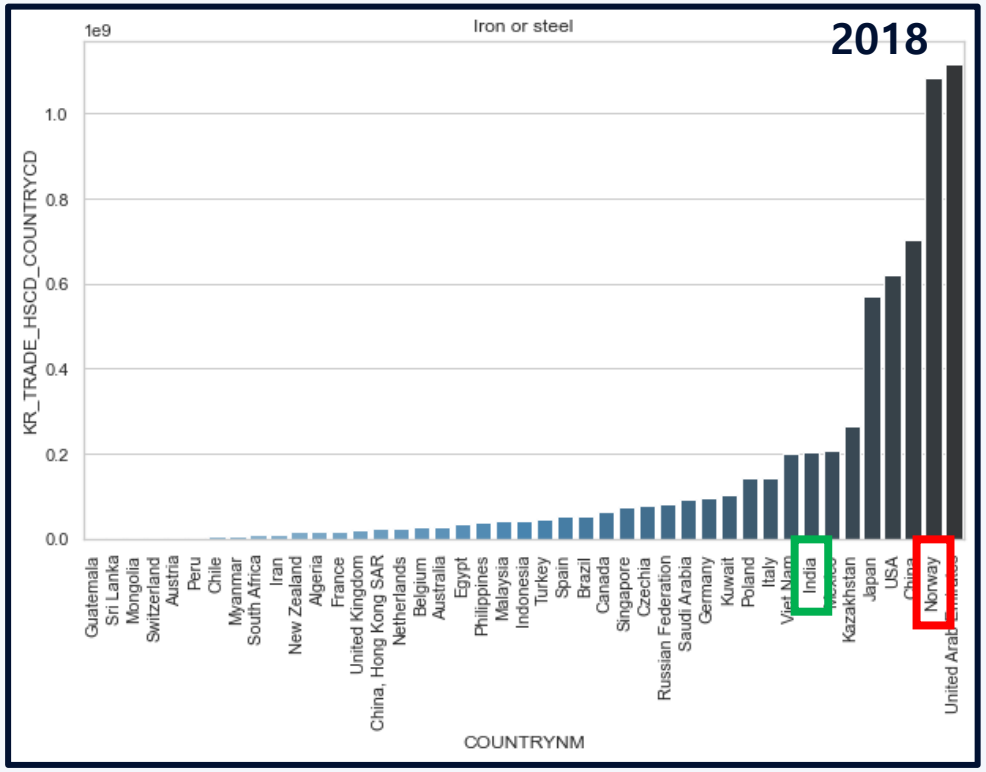
최종 submission

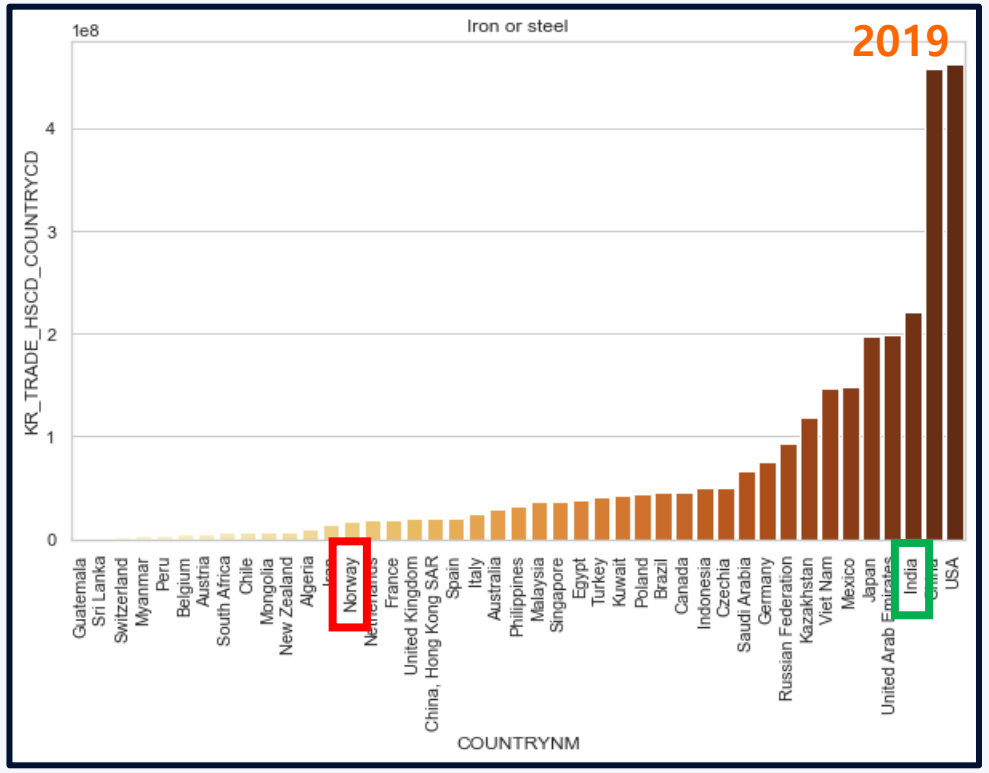
기대효과

20P

품목 기준 파트너 매칭

[Iron or Steel 기준 나라 매칭]

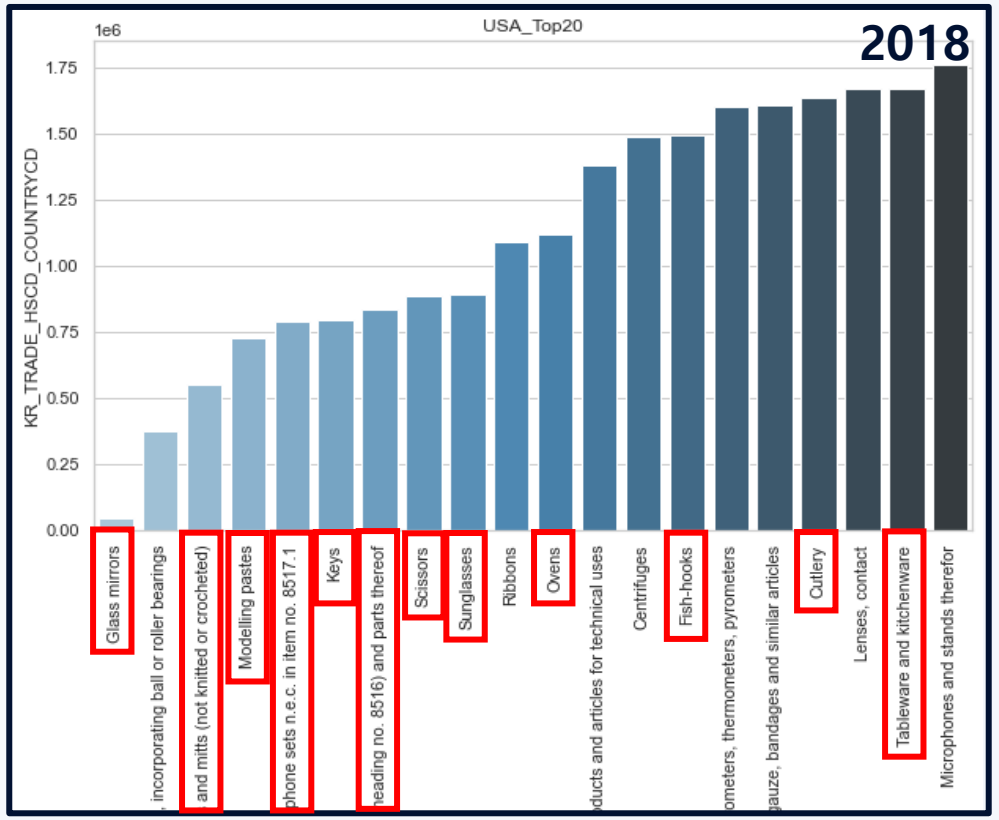


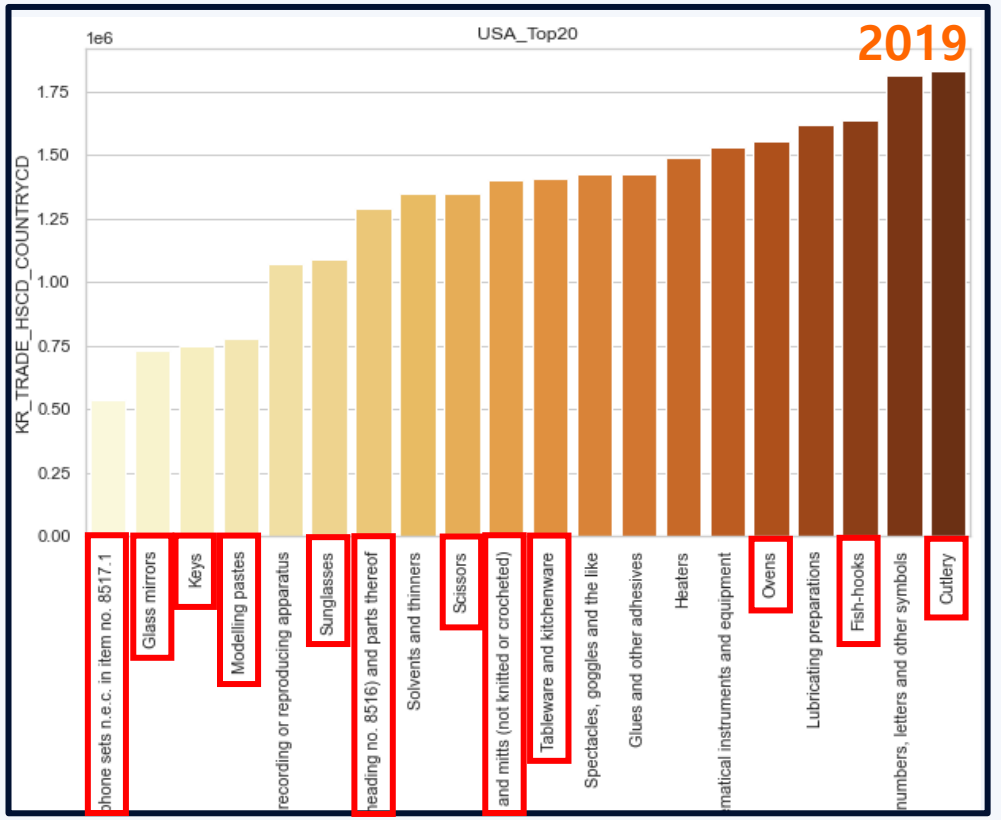


※ 기존 연도에 비해 예측한 연도의 수입 규모가 전체적으로 줄어든 것을 확인할 수 있음

국가 기준 파트너 매칭

[USA 기준 품목 매칭]

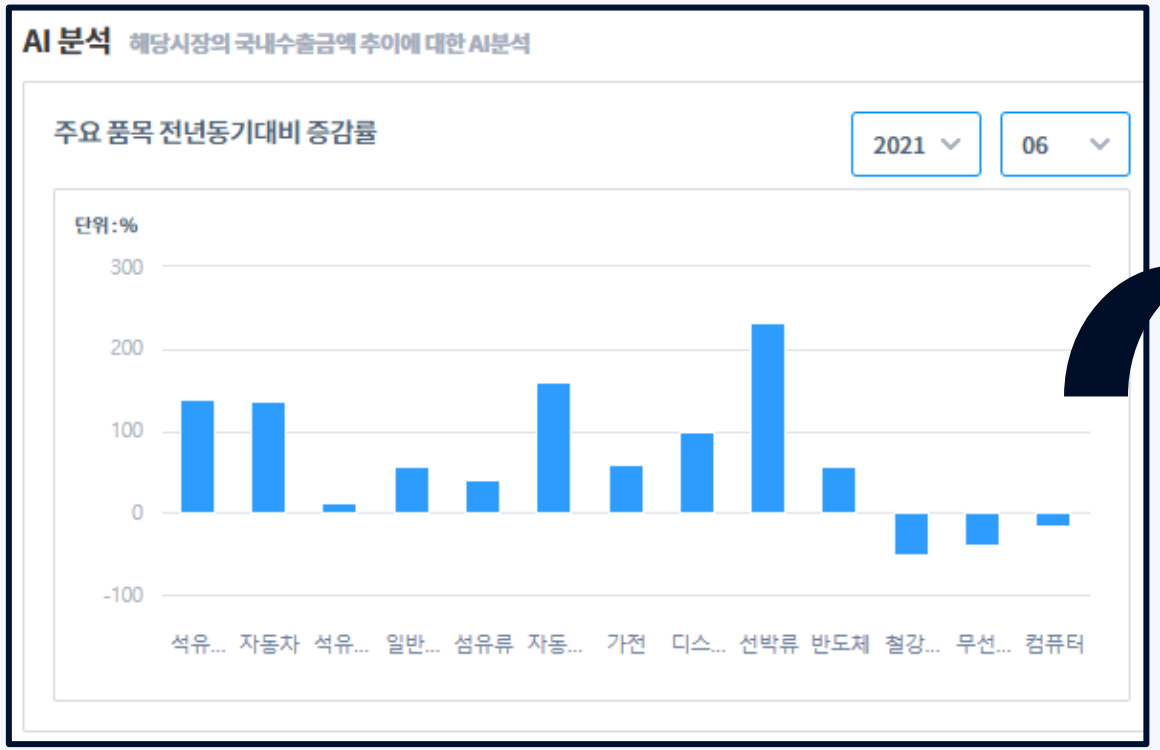


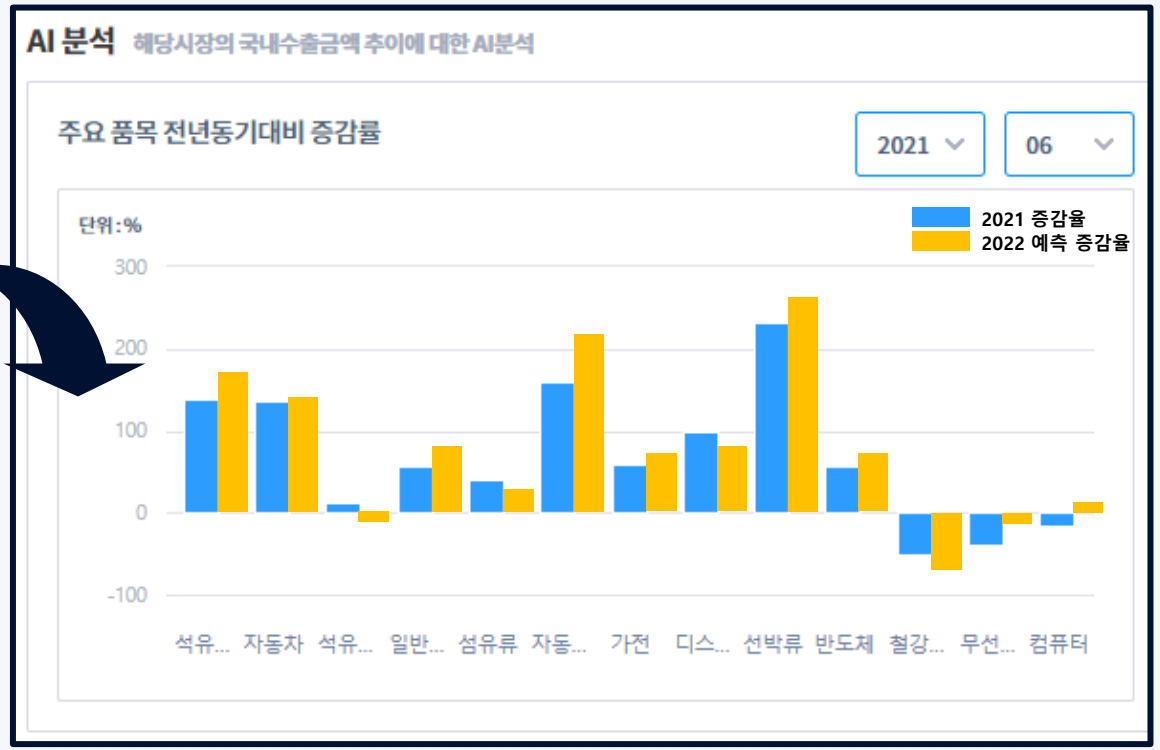


※ 전체 239개의 품목 중 상위 20개의 품목이 대부분 겹치는 것을 보아
USA에 비슷한 품목을 투자 유치하는 것이 안정성이 있음

기존 코트라 솔루션 개선

[AI 분석-01 예시]





활용방안

23P

기존 코트라 솔루션 개선

[AI 분석-02 예시]

한국과의 교역통계 (조회 품목 기준)

미국 수입액 현황 | 한국 수출액 현황 | 한국→미국 수출기업수 | 관세청 기준

단위: 천$

5,000,000 / 2,500,000 / 0

2016 2017 2018 2019 2020 2021

총 수입액 對한국 수입액

단위: 천$

300,000 / 200,000 / 100,000 / 0

2016 2017 2018 2019 2020 2021

총 수출액 對미국 수출액

250 / 225 / 200 / 175 / 150

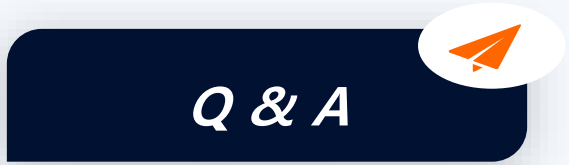
2018 2019 2020 2021

결론

예측 모델을 통해 미래의 경향을 포함하여 중소 중견기업에게
좀 더 안정적인 **잠재 파트너 매칭**을 제공할 수 있음!

KOTRA에서 제공하는 여러 **그래프에 예측된 미래의 분석을 추가**하여
고객의 의사결정에 좀 더 도움될 수 있음!

우리나라와 거래를 한 **구체적인 국내·해외 기업**들의 정보를 알 수 있다면
좀 더 정밀한 예측모델을 만들 수 있을 것이라 기대됨!



감사합니다

팀장 장성민 jsm50660@gmail.com
팀원 한보혜 bohaehan@gmail.com
팀원 마민정 maminjeong3199@gmail.com