

제2회 국민대학교AI빅데이터분석 경진대회

주요 5개 산업 품목에 대한 1개년의 무역량을 예측하는 알고리즘 개발

데이데이 팀

AI빅데이터융합경영학과 23 | 양시찬 황인성 오서영

CONTENTS

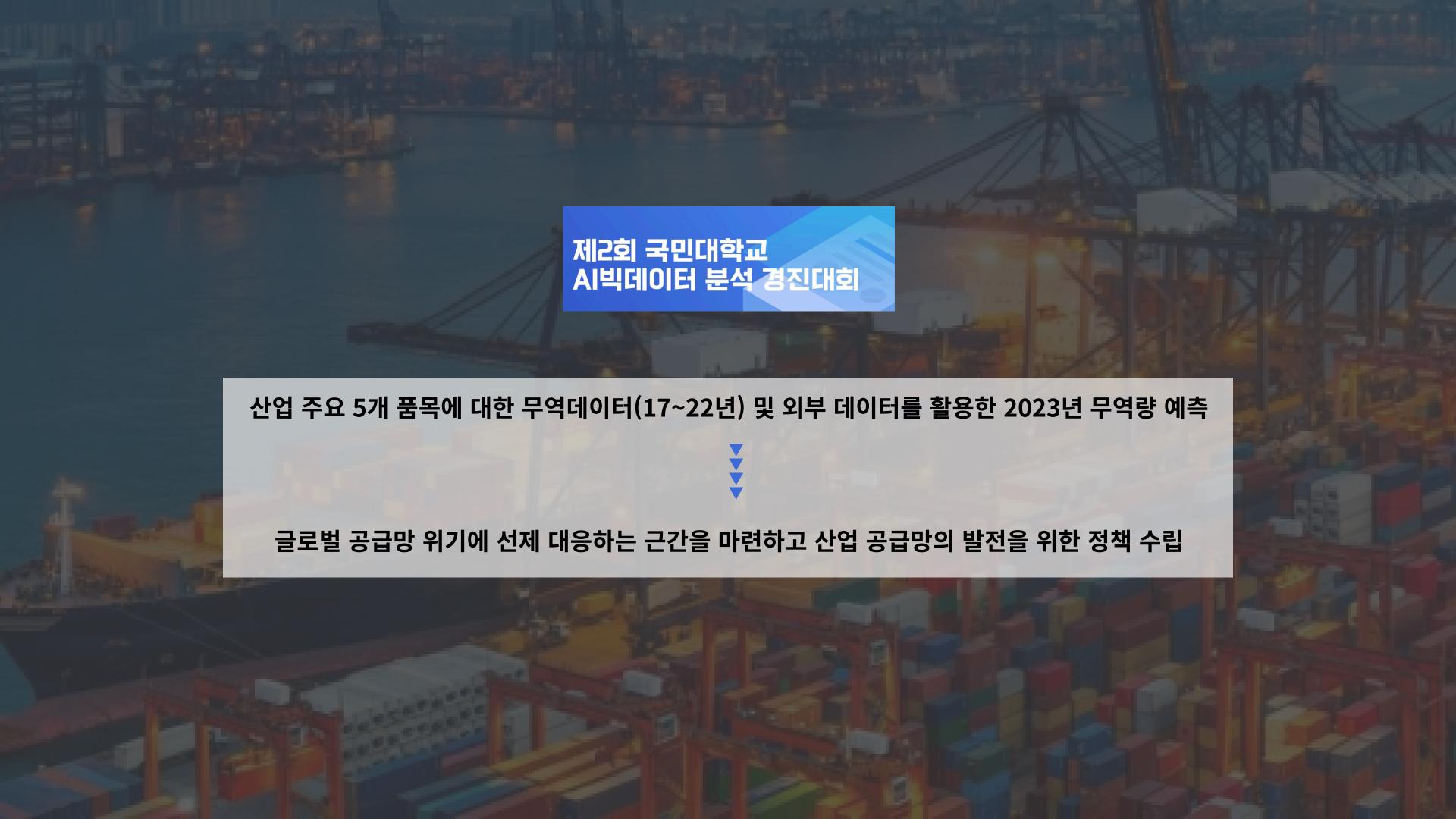
- **01** 데이터 설명
- **02** 아이디어 도출 과정
- ○3 코드 검증
- 모델의 유용성
- ○5 한계 및 개선 방안
- **06** 참고 자료

글로벌 주요국은 미·중 무역분쟁, 팬데믹 이후 국경봉쇄, 물류대란 등의 경험을 통해 공급망 자립화와 경제안보의 중요성에 대해 강조하기 시작

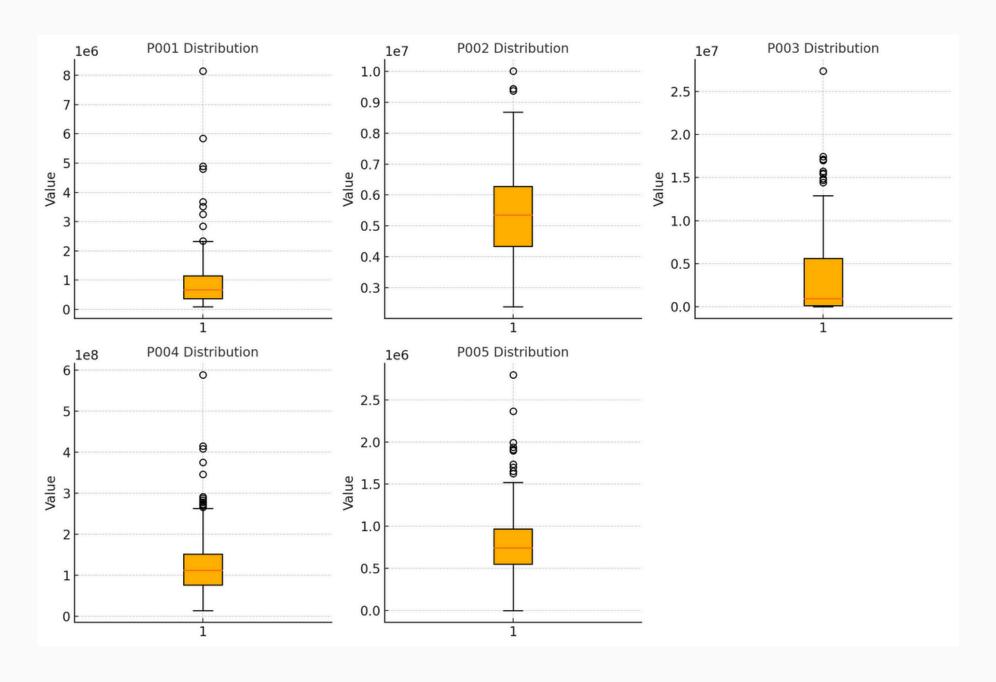
글로벌 공급망 위기를 경험한 주요국은 산업정책을 통해 경제안보를 달성하고자 노력 중

최근 주요국의 핵심 산업정책은 반도체, 이차전지, 전기차 등 첨단기술·친환경 분야의 산업 육성과이에 투입되는 원자재 확보에 집중되어 있음

원자재의 최대수입국 의존도는 전반적으로 감소하였으나 아직 상당수 품목의 리스크가 높은 편, 국내 기업의 해외투자는 주요국 산업정책으로 인해 중국에서 북미·유럽으로 이동

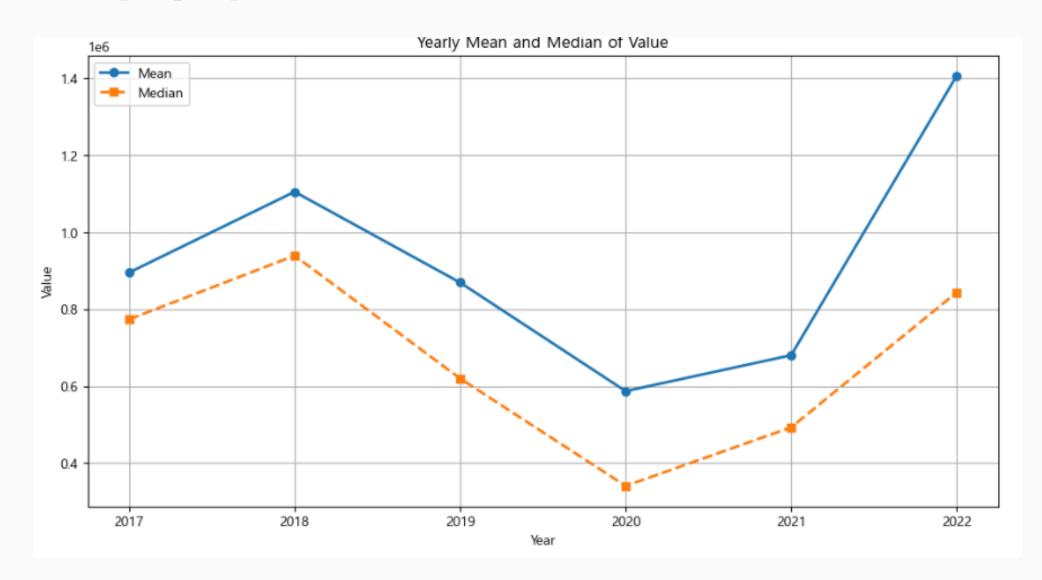


데이터 설명



전체적으로 분산이 크고 불규칙적인 경향을 보임

-> 불규칙한 데이터로부터 과적합을 방지하고 최대한 일반화된 결과를 도출하는 것이 중요하다고 판단



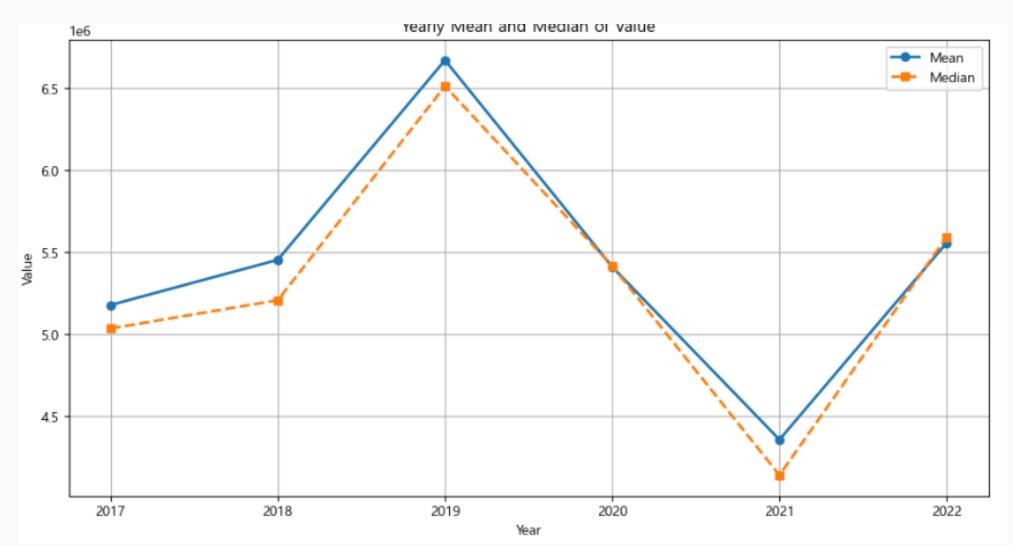
P001

산업: PAV (Personal Air Vehicle, 개인용 비행체) 품목명: 다중 수소연료전지 PEMFC(고분자전해질)

2017~2022년 데이터에서 평균과 중앙값이 유사한 추세를 보이며 상승 가능성이 관찰됨

-> 평균값과 중앙값의 평균을 계산한 후 25% 증가율을 적용해 예측값을 도출함

-> 증가된 값을 2023년도 수입금액 예측값으로 설정함



P002

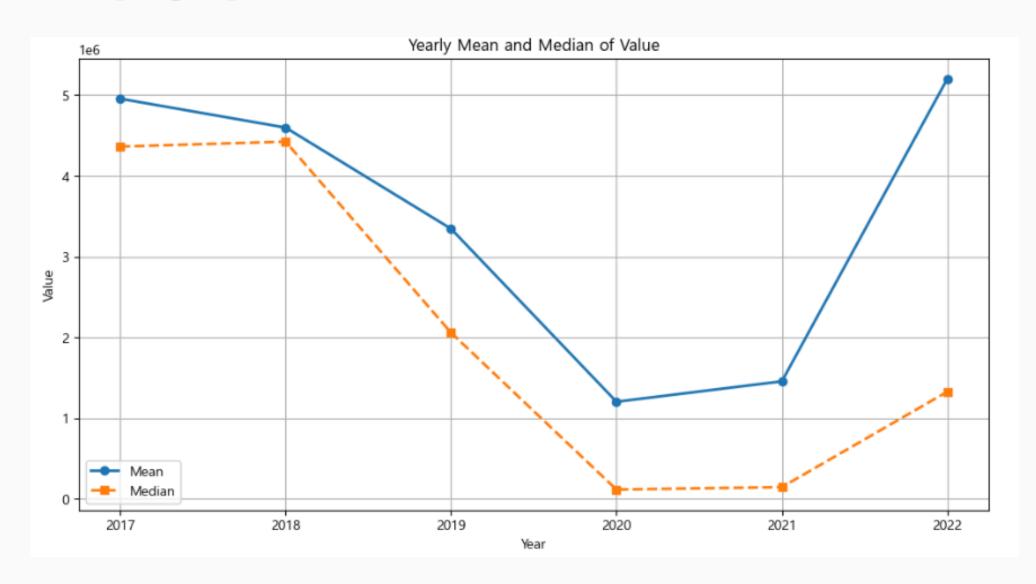
산업: 반도체

품목명: 티타늄 타겟

데이터 분석 및 중앙값 계산: 2017, 2020, 2021년의 중앙값을 계산하여 평균을 구함 중앙값의 평균 값을 만 단위에서 반올림하여 490만으로 설정

-> 조정값 계산: 490만의 1/1000의 1/3을 계산하고, 이를 1의 자리에서 반올림하여 1630으로 도출

-> 최종 예측값 설정: {(중앙값 평균) + 1630}에 1.25를 곱하여 최종 예측값을 도출함



P003

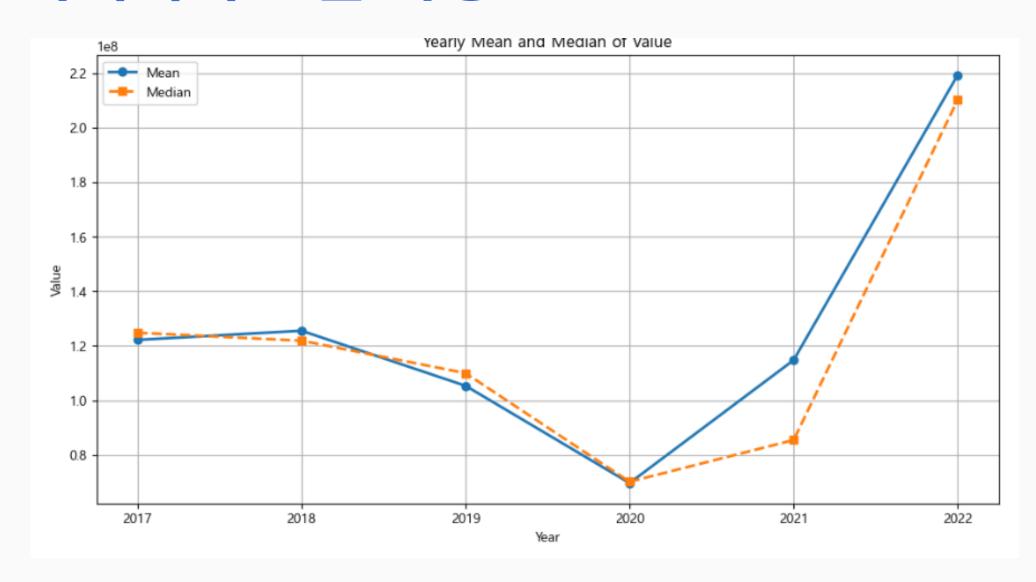
산업: 인조흑연

품목명:석유코크스

2022년 이후 평균과 중앙값 간 차이가 커지며 수입금액 변동이 불안정한 것으로 관찰됨

-> 2017~2022년 중앙값들의 평균을 계산하여 예측값을 도출함

-> 계산된 값을 2023년도 수입금액 예측값으로 설정함



P004

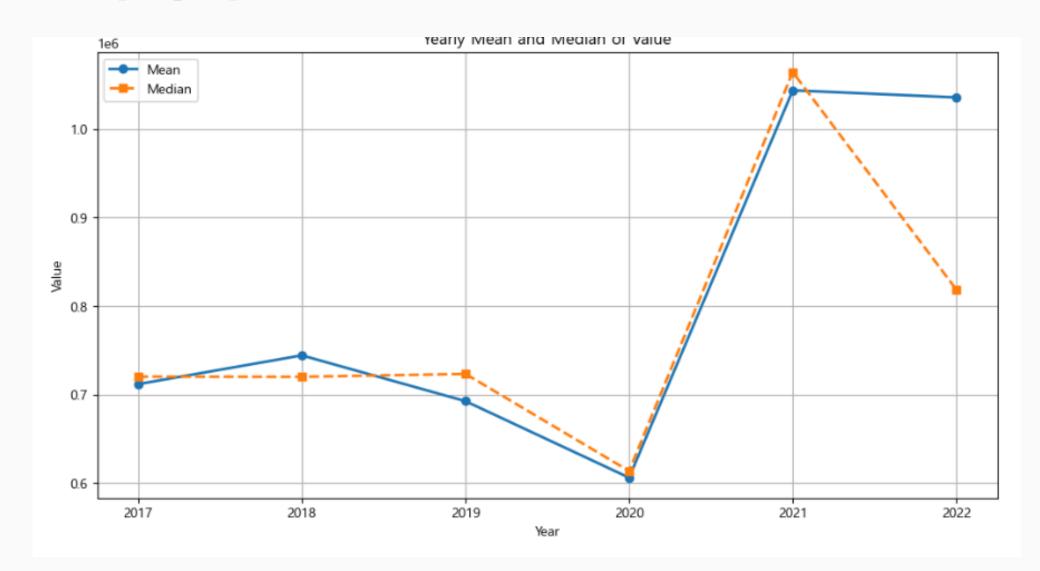
산업:철강

품목명:유연탄

평균과 중앙값이 유사한 추세를 보이나, 2022년에 수입금액이 급격히 상승하며 코로나의 영향을 받은 것으로 판단됨

-> 2021년과 2022년의 평균값과 중앙값의 평균간 평균을 계산하여 1억 6천만 원으로 설정

-> 상반기에는 500만 원을 추가한 1억 6천 500만 원, 하반기는 1억 6천만 원으로 예측



P005

산업:희소금속

품목명: 안티모니 화합물

1단계: 2021년까지 평균값과 중앙값이 비슷했으나, 2022년에 극단적인 값이 발생하여 차이가 커짐

-> 이상치 영향을 최소화하기 위해 2017~2022년 중앙값을 기준으로 예측값 도출

-> 계산된 중앙값을 2023년도 수입금액 예측값으로 설정

코드 검증 P001

```
P001_result = round(((P001_DF['value'].mean() + P001_DF['value'].median()) / 2) * 1.2502155)
P001_pred['value'] = [P001_result]*36
```

P002

```
### P002 pred ###
P002_MEDIAN_2021 = np.median(P002_DF[P002_DF['year']==2021]['value'])
P002_MEDIAN_2020 = np.median(P002_DF[P002_DF['year']==2020]['value'])
P002_MEDIAN_2017 = np.median(P002_DF[P002_DF['year']==2017]['value'])
# 중앙값의 평균값 정의
P002_MEDIAN_2017_20_21 = (P002_MEDIAN_2021 + P002_MEDIAN_2020 + P002_MEDIAN_2017)/3
# P002_MEDIAN_2017_20_21 값을 10만 자리에서 반올림
rounded to 100k = round(P002 MEDIAN 2017 20 21, -5)
# 추가 값 계산 (10만 자리 반올림 값 * 1/1000 * 1/3) 을 소수 첫째 자리에서 반올림
additional_value = round((rounded_to_100k * (1/1000) * (1/3)), 0)
# 최종 계산
P002_result = (P002_MEDIAN_2017_20_21 + additional_value) * 1.25
P002_result = math.floor(P002_result / 10) * 10
P002_pred['value'] = [P002_result]*36
### P003 pred ###
P003_result = round((P003_DF['value'].mean() + P003_DF['value'].median()) / 2)
P003_pred['value'] = [P003_result]*36
```

코드 검증

P003, P004, P005

```
### P003_result = round((P003_DF['value'].mean() + P003_DF['value'].median()) / 2)
P003_pred['value'] = [P003_result]*36

### P004 pred ###
P004_DF_2022_MM = (P004_DF[P004_DF['year']==2022]['value'].median() + P004_DF[P004_DF['year']==2022]['value'].mean())/2
P004_DF_2022_MM = (P004_DF[P004_DF['year']==2021]['value'].median() + P004_DF[P004_DF['year']==2021]['value'].mean())/2
P004_DF_2021_2022_MM = (P004_DF_2022_MM + P004_DF_2021_MM)/2
P004_DF_2021_2022_MM = round(P004_DF_2021_2022_MM, -7)
P004_2023_상반기 = P004_DF_2021_2022_MM + 5000000
P004_2023_상반기 = P004_DF_2021_2022_MM
P004_pred['value'] = [P004_2023_상반기]*18 + [P004_2023_하반기]*18

### P005_pred ###
P005_result = round(P005_DF['value'].median())
P005_pred['value'] = [P005_result]*36
```

모델의 유용성

주요 내용

1. 클래스 정의와 메서드:

- 클래스 이름: myModel .
- 주요 메서드: predict(df).
- 입력: df (데이터프레임, 제품 정보 포함).
- 출력: Total_pred (모든 제품의 예측값을 포함한 데이터프레임).

2. **pid** 별 데이터 분리:

- df 에서 제품 ID(pid)에 따라 데이터 분리:
 - P001_DF, P002_DF, P003_DF, P004_DF, P005_DF 생성.

모델의 유용성

각 제품의 예측 로직

P001:

● P001_pred['value'] 릂 고정된 값 ['1'] * 36 로 채움 (임의의 값으로 보임).

P002:

- P002_DF 에서 2021, 2020, 2017년의 중앙값 계산.
- 중앙값의 평균(P002_MEDIAN_2017_20_21)을 구한 후:
 - o 이 값을 10만 단위로 반올림(rounded_to_100k).
 - 추가 계산 값(additional_value)을 소수 첫째 자리로 반올림.
 - 최종 결과(P002_result)를 계산 후, 10의 자리로 내림(math.floor).
- 결과 값을 36개의 동일한 값으로 채움.

P003:

- P003_DF 의 평균과 중앙값의 평균(P003_result)을 계산.
- 결과 값을 36개의 동일한 값으로 채움.

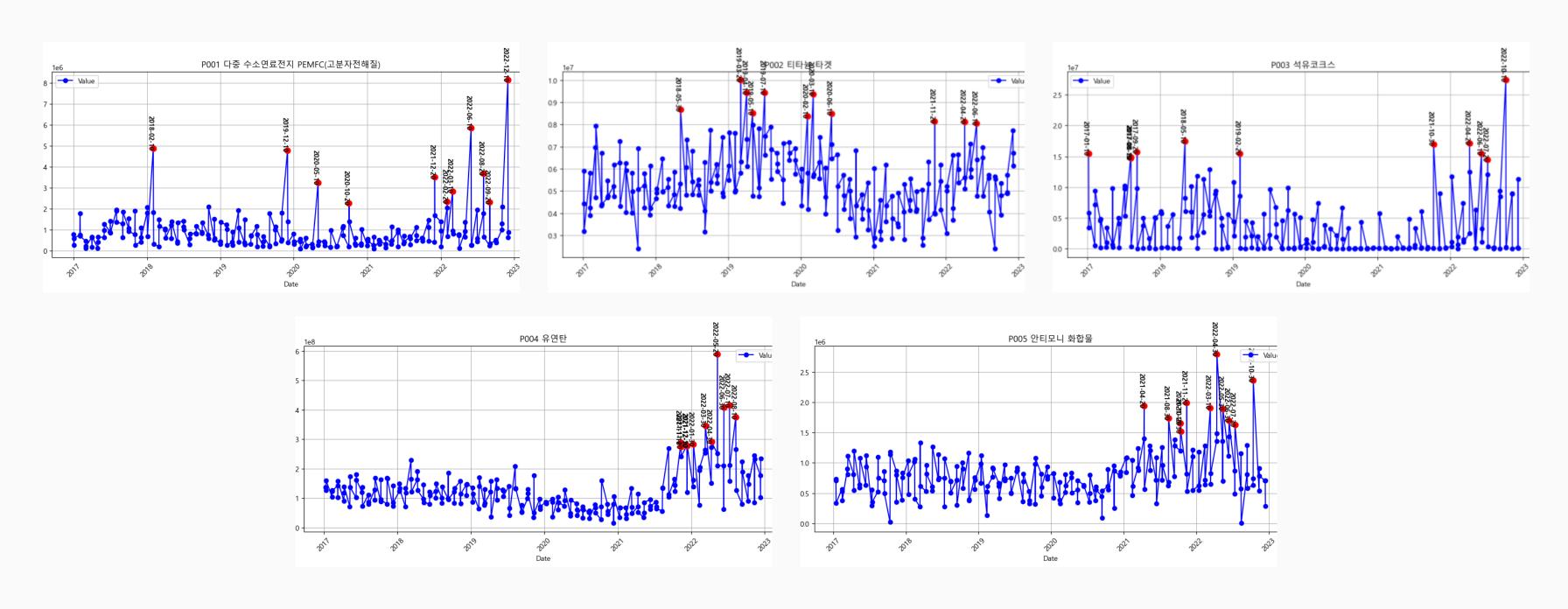
P004:

- 2022년, 2021년 데이터를 사용해 평균과 중앙값의 평균(MM) 계산.
- 두 해의 MM 값을 평균낸 후 1천만 단위로 반올림.
- 2023년 상반기와 하반기 값을 계산:
 - 상반기: 2021-2022 평균 + 5,000,000.
 - o 하반기: 2021-2022 평균.
- 상반기 값 18개 + 하반기 값 18개로 채움.

P005:

- P005_DF 의 중앙값(P005_result)을 계산.
- 결과 값을 36개의 동일한 값으로 채움.
- 결과 통합:
 - 모든 P00X_pred 데이터프레임을 pd.concat 으로 합침.
 - 결과 데이터프레임 Total_pred 를 반환.

한계 및 개선 방안



특정 날짜에서 값이 급격히 변동하여 날짜를 포함시키려고 했지만, 타당한 상관관계를 찾지 못해 날짜를 중요한 변수로 생각했음에도 불구하고 활용하지 못함

참고 자료

문화 주요국 산업정책과 산업별 국내 공급망 변화

(한국수출입은행에서 주요국 산업정책과 산업별 국내 공급망 변화를 분석한 보고서)

P001

다중 수소연료전지 PEMFC

P002

티타늄 타겟

티타늄 타겟은 반도체 제조 공정의 핵심 소재로, 보고서에서는 반도체 산업의 주요 원자재로 실리콘, 희토류, 텅스텐 등이 언급되며, 대부분 중국 의존도가 높은 리스크가 있음.

표> <수입편중도와 대체 후보국>
지로, 반도체 원자재 중 일부는 글로
E와 관 벌 수출편중도가 낮아 대체 후보
보 및 국 탐색이 가능하다고 판단되
대, 티타늄 타겟 역시 비슷한 접
근으로 공급망 다각화가 중요

P003

석유코크스

석유코크스는 인조흑연 제조의 필수 원료로, 이차전지 음극재 와 밀접한 관련이 있음. 보고서 에서는 이차전지 원자재의 수입 량과 공급망 리스크가 분석되어 있으며, 특히 중국 의존도가 낮 아지고 있는 추세가 언급됨.

<공급망 리스크>
생산량이 제한적이며 가격 변동
성이 높다는 석유코크스의 특징
은 보고서에서 리튬 및 흑연의
높은 대외 의존도와 유사한 리스
크를 공유

P004

유연탄

유연탄은 철강 제조에 필수적인 코크스 생산 원료로, 보고서에 서는 철강산업과 관련된 원자재 로 크롬과 마그네슘 등이 언급 됨. 특히 중국 의존도가 높은 원 자재들에 대한 대체 기술 개발과 공급망 다각화 필요성이 강조 됨.

<이산화탄소 배출> 보고서에서 탄소중립 및 친환경 정책과 맞물려, 유연탄 대체 기 술 개발은 필수적 P005

안티모니 화합물

안티모니 화합물의 원료는 중국 이 주요 생산국이며, 보고서에 서도 희토류와 같은 핵심 원자재 가 중국에 편중되어 있다는 리스 크가 제기됨.

< 독성 우려와 대체 기술> 안티모니 화합물의 독성 우려와 관련해 보고서에서는 친환경 및 탄소중립 목표와 연결하여 유해 물질의 대체 기술 개발 필요성이 언급

<만소중립 및 친환경 목표>
PEMFC는 친환경적 에너지로,
보고서에서 탄소중립 목표와 관 련된 원자재의 안정적 확보 및 정부 지원 정책이 중요

감사합니다

팀 데이데이