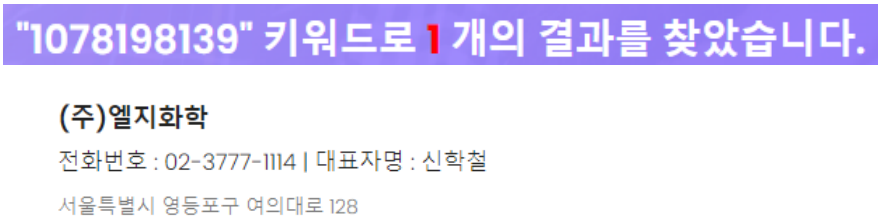
[기후금융] 기업명 전처리

✅ 할 일

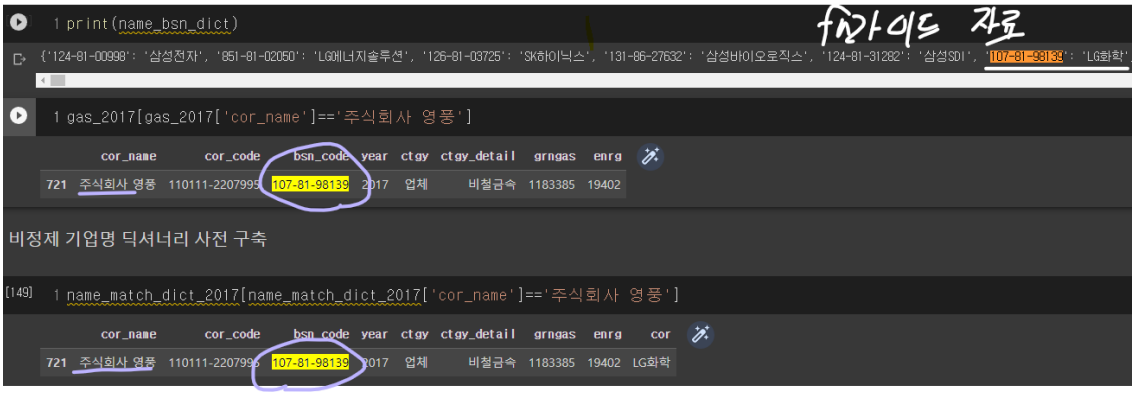
- 사업자등록번호 매칭 → 칼럼

▼ 사업자번호 매칭 데이터(유료데이터) 오류 다수 발견

- **(예시) 사업자번호 검색사이트에서 검색해보면 엘지화학 == 107-81-98139**



- **Fn가이드 및 유료 데이터에서 엘지화학**



  - (위) Fnguide에서 제공한 사업자번호데이터 → 107-81-98139 == LG화학
  - (아래) **유료 데이터**에서 제공받은 사업자번호데이터 → 107-81-98139 == **영풍**

- 오류 있는 기업명 리스트

```
[['LG화학', '영풍'],
 ['평화산업', '포스코스틸리온'],
 ['한일홀딩스', '한일시멘트'],
 ['LG화학', '영풍'],
 ['평화산업', '포스코스틸리온'],
 ['한일홀딩스', '한일시멘트'],
 ['LG화학', '영풍'],
 ['평화산업', '포스코스틸리온'],
 ['한일홀딩스', '한일시멘트'],
 ['한일홀딩스', '한일시멘트'],
 ['평화산업', '포스코스틸리온'],
 ['LG화학', '영풍'],
 ['LG화학', '영풍'],
 ['한일홀딩스', '한일시멘트'],
 ['평화산업', '포스코스틸리온'],
 ['LG화학', '영풍'],
 ['카프로', '진에어'],
 ['한일홀딩스', '한일시멘트'],
 ['한국조선해양', '현대중공업']]
```

▼ 전처리

- **목표**: **Fnguide 기업명으로 통일**시켜, 기업재무정보와 탄소배출 데이터를 매칭
  - 특히, **사업자번호無 연도의 비정제 기업명 변환이 핵심**
  - 기업재무정보가 있는 데이터 기업명 == **정제 데이터**
  - 변환해야 하는 탄소배출 데이터 자료 == **비정제 데이터**

	non_cor_name	year	ctgy	ctgy_detail	grngas	enrg	cor_name
0	(주)강원랜드	2011	업체	건물	70,829	1,301	(주)강원랜드
1	(주)경기고속	2011	업체	교통(여객)	167,188	2,540	(주)경기고속
2	(주)대명레저산업	2011	사업장	건물	49,618	955	(주)대명레저산업
3	(주)대원고속	2011	업체	교통(여객)	152,566	2,277	(주)대원고속
4	(주)무주덕유산리조트	2011	사업장	건물	27,636	465	(주)무주덕유산리조트
...	...	...	...	...	...	...	...
1072	효성중공업 주식회사	2021	업체	산업	54,101	1,132	효성중공업 주식회사
1073	효성첨단소재 주식회사	2021	업체	산업	202,512	4,210	효성첨단소재 주식회사
1074	효성티앤씨 주식회사	2021	업체	산업	365,131	8,403	효성티앤씨 주식회사
1075	효성화학 주식회사	2021	업체	산업	887,229	18,850	효성화학 주식회사
1076	휴비스	2021	업체	산업	465,851	7,089	휴비스

4487 rows × 7 columns

  - 변환할 기업명 데이터 개수: 4487 개

▼ [STEP 1] Fnguide 기업명(한글) 데이터 사용

- { **사업자번호無 비정제 데이터 기업명 : Fnguide에서 제공하는 기업명**} 매칭 → 정제 데이터 '**Name**'으로 변환
  - Fnguide에서 제공하는 'Name(=정제)', '기업명(한글)'과 '사업자등록번호' 자료

Symbol	Name	기업명 (한글)	사업자등록번호	법인등록번호
A005930	삼성전자	삼성전자(주)	124-81-00998	1.30E+12
A373220	LG에너지솔루션	(주)엘지에너지솔루션	851-81-02050	1.10E+12
A000660	SK하이닉스	에스케이하이닉스(주)	126-81-03725	1.34E+12
A207940	삼성바이오로직스	삼성바이오로직스(주)	131-86-27632	1.20E+12
A006400	삼성SDI	삼성SDI(주)	124-81-31282	1.10E+12
A051910	LG화학	(주)엘지화학	107-81-98139	1.10E+12

- **결과**
  - 변환되지 않고 남은 데이터 개수: 3756 개

	non_cor_name	year	ctgy	ctgy_detail	grngas	enrg	cor_name
0	(주)강원랜드	2011	업체	건물	70,829	1,301	강원랜드
1	(주)경기고속	2011	업체	교통(여객)	167,188	2,540	(주)경기고속
2	(주)대명레저산업	2011	사업장	건물	49,618	955	(주)대명레저산업
3	(주)대원고속	2011	업체	교통(여객)	152,566	2,277	(주)대원고속
4	(주)무주덕유산리조트	2011	사업장	건물	27,636	465	(주)무주덕유산리조트
...	...	...	...	...	...	...	...
4482	효성중공업 주식회사	2021	업체	산업	54,101	1,132	효성중공업 주식회사
4483	효성첨단소재 주식회사	2021	업체	산업	202,512	4,210	효성첨단소재 주식회사
4484	효성티앤씨 주식회사	2021	업체	산업	365,131	8,403	효성티앤씨 주식회사
4485	효성화학 주식회사	2021	업체	산업	887,229	18,850	효성화학 주식회사
4486	휴비스	2021	업체	산업	465,851	7,089	휴비스

비정제 기업명 / 정제된 기업명

▼ [STEP 2] 사업자등록번호 있는 연도(2016~2020) 데이터 사용

1. { **Fnguide에서 제공하는 사업자등록번호 : 사업자번호有 탄소배출 데이터**} 매칭 → **사업자번호有 연도의 비정제 기업명** → **정제 기업명** 변환
   - 즉, 사업자번호가 같으면 기업명을 정제할 수 있음
   - Fnguide에서 제공하는 'Name(=정제 기업명)'과 '**사업자등록번호**' 자료

Symbol	Name	기업명 (한글)	사업자등록번호	법인등록번호
A005930	삼성전자	삼성전자(주)	124-81-00998	1.30E+12
A373220	LG에너지솔루션	(주)엘지에너지솔루션	851-81-02050	1.10E+12
A000660	SK하이닉스	에스케이하이닉스(주)	126-81-03725	1.34E+12
A207940	삼성바이오로직스	삼성바이오로직스(주)	131-86-27632	1.20E+12
A006400	삼성SDI	삼성SDI(주)	124-81-31282	1.10E+12
A051910	LG화학	(주)엘지화학	107-81-98139	1.10E+12

   - 사업자번호有 연도의 탄소배출 데이터: '비정제 기업명'과 '**사업자번호**'

	non_cor_name	cor_code	bsn_code	year	ctgy	ctgy_detail	grngas	enrg
0	(유)에스케이씨에보닉페록사이드코리아	230114-0001992	610-81-82792	2016	사업장	석유화학	60242	1195
1	(주)MH에탄올	190111-0003563	608-81-03800	2016	사업장	음식료품	23971	287
2	(주)MSC	-	사업자번호 -	2016	사업장	음식료품	25332	506
3	(주)SIMPAC METALLOY	-	-	2016	업체	철강	334985	3726
4	(주)SPP조선	-	-	2016	사업장	조선	50797	898

사업자번호有 비정제 기업명

2. 위를 이용해 { **사업자번호有 연도 비정제 기업명** : **정제 기업명, 사업자번호** } dictionary 구축

   ⇒ **사업자번호有 연도 비정제 기업명**와 **사업자번호無 연도의 비정제 기업명**가 일치한다면, **사업자번호無 연도의 비정제 기업명** → **정제 기업명** 가능

   - 사업자번호有 연도 비정제 기업명 : 정제 기업명

```
{'MH에탄올': ['(주)MH에탄올'], 'SIMPAC Metal': ['(주)SIMPAC METAL'], '강원랜드': ['(주)강원랜드'],
```

- 정리

- 결과
  - 변환되지 않고 남은 데이터 개수: 2980 개

	non_cor_name	year	ctgy	ctgy_detail	grngas	enrg	cor_name	bsn_code 사업자번호로 변환
0	(주)강원랜드	2011	업체	건물	70,829	1,301	강원랜드	225-81-10770
1	(주)경기고속	2011	업체	교통(여객)	167,188	2,540	(주)경기고속	(주)경기고속
2	(주)대명레저산업	2011	사업장	건물	49,618	955	(주)대명레저산업	(주)대명레저산업
3	(주)대원고속	2011	업체	교통(여객)	152,566	2,277	(주)대원고속	(주)대원고속
4	(주)무주덕유산리조트	2011	사업장	건물	27,636	465	(주)무주덕유산리조트	(주)무주덕유산리조트
...	...	...	...	...	...	...	...	...
4482	효성중공업 주식회사	2021	업체	산업	54,101	1,132	효성중공업	578-87-00896
4483	효성첨단소재 주식회사	2021	업체	산업	202,512	4,210	효성첨단소재	198-87-00929
4484	효성티앤씨 주식회사	2021	업체	산업	365,131	8,403	효성티앤씨	880-87-01070
4485	효성화학 주식회사	2021	업체	산업	887,229	18,850	효성화학	175-88-01164
4486	휴비스	2021	업체	산업	465,851	7,089	휴비스	215-81-98804

비정제 기업명

▼ **[STEP 3] 정규표현식 (주), (유), 주식회사 제거**

1. **사업자번호無 연도의 비정제 기업명**에서 정규표현식으로 (주), (유), 주식회사, 유한회사, 유한책임회사 등 제거

   ex. (주) 대원고속 → 대원고속

2. **처리한 기업명**이 Fnguide의 "**Name**"과 일치하면, **사업자등록번호**로 변환
   - 단, Fnguide 재무정보가 존재하지 않으면, "Name"에 없으므로 변환X

     ex. 대원고속은 Fnguide 재무정보가 없으므로, 사업자등록번호로 변환X

Symbol	Name	기업명 (한글)	사업자등록번호	법인등록번호
A005930	삼성전자	삼성전자(주)	124-81-00998	1.30E+12
A373220	LG에너지솔루션	(주)엘지에너지솔루션	851-81-02050	1.10E+12
A000660	SK하이닉스	에스케이하이닉스(주)	126-81-03725	1.34E+12
A207940	삼성바이오로직스	삼성바이오로직스(주)	131-86-27632	1.20E+12
A006400	삼성SDI	삼성SDI(주)	124-81-31282	1.10E+12
A051910	LG화학	(주)엘지화학	107-81-98139	1.10E+12

- **결과**
  - 변환되지 않고 남은 데이터 개수: 2924 개

	non_cor_name	year	ctgy	ctgy_detail	grngas	enrg	cor_name	bsn_code	del_name
0	(주)강원랜드	2011	업체	건물	70,829	1,301	강원랜드	225-81-10770	225-81-10770
1	(주)경기고속	2011	업체	교통(여객)	167,188	2,540	(주)경기고속	경기고속	경기고속
2	(주)대명레저산업	2011	사업장	건물	49,618	955	(주)대명레저산업	대명레저산업	대명레저산업
3	(주)대원고속	2011	업체	교통(여객)	152,566	2,277	(주)대원고속	대원고속	대원고속
4	(주)무주덕유산리조트	2011	사업장	건물	27,636	465	(주)무주덕유산리조트	무주덕유산리조트	무주덕유산리조트
...	...	...	...	...	...	...	...	...	...
4482	효성중공업 주식회사	2021	업체	산업	54,101	1,132	효성중공업	578-87-00896	578-87-00896
4483	효성첨단소재 주식회사	2021	업체	산업	202,512	4,210	효성첨단소재	198-87-00929	198-87-00929
4484	효성티앤씨 주식회사	2021	업체	산업	365,131	8,403	효성티앤씨	880-87-01070	880-87-01070
4485	효성화학 주식회사	2021	업체	산업	887,229	18,850	효성화학	175-88-01164	175-88-01164
4486	휴비스	2021	업체	산업	465,851	7,089	휴비스	215-81-98804	215-81-98804

비정제 기업명

정규표현식으로 삭제 후, 저리된 기업명과 Fnguide 기업명이 일치하면, { Fnguide 기업명 : 사업자번호} 로 매칭

▼ **[STEP 4]: 정규표현식 한글 → 영어**

1. **사업자번호無 연도의 비정제 기업명**에서 정규표현식으로 한글 → 영어로 변환
   - 대기업 이름 영어 발음 (ex. 에스케이에너지 → SK에너지)

```
## 대기업 영어발음 사전
big_cor_eng_kor = {'SK': '에스케이', 'LG': '엘지', 'SH': '에스에이치', 'KT': '케이티',
       'DB':'디비', 'KG':'케이지', 'SPC': '에스피씨', 'SKC':'에스케이씨', 'OCI':'오씨아이'}
```

   - 모든 영어 발음 (ex. 에스케이에너지 → SK에너G)

```
## 한글-영어발음 사전 및 기타
eng_kor = {'A': '에이', 'B': '비', 'C': '씨', 'D': '디',
       'E':'이', 'F':'에프', 'G': '지', 'H': '에이치',
       'I': '아이', 'J': '제이', 'K': '케이', 'L': '엘',
       'M':'엠', 'N':'엔', 'O': '오', 'P': '피',
       'Q': '큐', 'R': '알', 'S': '에스', 'T': '티',
       'U': '유', 'V': '브이','W':'더블유', 'X': '엑스',
       'Y':'와이', 'Z':'지', '&':'앤'}
```

2. 대기업 이름 영어 발음으로 **처리한 기업명**이 Fnguide의 "**Name**"과 일치하면, **사업자등록번호**로 변환
3. 그리고 모든 영어 발음으로 **처리한 기업명**이 Fnguide의 "**Name**"과 일치하면, **사업자등록번호**로 변환
    - 단, Fnguide 재무정보가 존재하지 않으면, "Name"에 없으므로 변환X
    - 2번 과정 후 3번 과정을 거치면, '에스케이에너지' → 'SK에너지'로 변환되고 사업자등록번호와 매칭. 그러면 'SK에너G'로 잘못 변환되어 매칭이 안되는 경우를 방지

▼ [STEP 5]: 정규표현식으로 직접 수정 및 매칭

1. 행정구역이름 데이터를 가져와 두 글자 이상의 지역명만 남김

```
'도척', '북천', '달성', '대합', '파천'
```

2. **사업자번호無 연도의 비정제 기업명**에서 '지역명+공장' → 삭제 후, **처리한 기업명**이 Fnguide의 "**Name**"과 일치하면, **사업자등록번호**로 변환
    - ex. 콘프로덕츠코리아 부평공장 → 콘프로덕츠코리아
3. **사업자번호無 연도의 비정제 기업명**에서 '지역명+숫자+공장' → 삭제 후, **처리한 기업명**이 Fnguide의 "**Name**"과 일치하면, **사업자등록번호**로 변환
    - ex. 콘프로덕츠코리아 부평3공장 → 콘프로덕츠코리아
4. **사업자번호無 연도의 비정제 기업명**에서 '공장' → 삭제 후, **처리한 기업명**이 Fnguide의 "**Name**"과 일치하면, **사업자등록번호**로 변환
    - ex. 콘프로덕츠코리아 공장 → 콘프로덕츠코리아

- 사업자번호와 매칭이 안된 기업 개수확인(중복 제외): 872 개