Preprocessing 보고서

데이터 파일 불러오기

- 크롤링한 대학교 데이터 'university_locations.csv'
- 서울시 상권 데이터 '소상공인시장진흥공단_상가(상권)정보_서울_202409.csv'
- 코드 파일과 데이터 파일이 같은 위치에 있어야 함

university_locations.csv 전처리 과정

- 서울시 46개의 대학 중 경희대학교의 위도, 경도 정보를 찾을 수 없어 직접 추가
- 면적 값 정리 및 결측값 처리
- 면적 별 반경 계산 및 정렬
  - 면적 100,000㎡ 이하 학교는 반경 500m
  - 면적 100,000㎡ 초과 시 반경 600m, 100,000㎡ 당 반경 100m 추가
  - 면적 1000,000㎡ 이상 학교의 반경은 2000m 고정
- 대학명에서 불필요한 문자 제거 (ex: 한국폴리텍대학[81])

소상공인시장진흥공단_상가(상권)정보_서울_202409.csv 전처리 과정

- 데이터의 column 이름 확인
- 필요한 column만 남기기

```
# 필요한 열 목록 정의
columns_to_keep = ['상호명', '상권업종대분류코드', '상권업종대분류명',
                   '상권업종중분류코드', '상권업종중분류명', '상권업종소분류코드',
                   '상권업종소분류명', '경도', '위도']
```

- 상권업종대분류 코드 'I2'를 제외한 모든 값 삭제 ('I2' = 음식점)

대학 반경 안에 있는 음식점 데이터

- pip install geopy : geopy 설치

- from geopy.distance import great_circle : 두 지점 간의 거리를 계산하기 위한 함수 great_circle을 제공

- 각 음식점에 대해 대학교 정보 추가

```
# '연세대학교' 자리에 다른 대학교 이름을 넣어서 확인 가능
matching_rows = df_restaurant_final_copy[df_restaurant_final_copy['대학교'].apply(lambda x: '연세대학교' in x)]

# 결과 출력
print(matching_rows)
```

```
                상호명 상권업종대분류코드 상권업종대분류명 상권업종중분류코드 상권업종중분류명 상권업종소분류코드  \
178          제이앤제이슨R         I2       음식       I211        주점     I21104
363           꼬뜨레스토랑         I2       음식       I204        서양식     I20401
750           쭈레쥬르신촌         I2       음식       I210     기타 간이     I21001
950            루트184        I2       음식      I210    기타 간이     I21001
1001           모미지식당         I2       음식       I201        한식     I20101
...              ...       ...      ...        ...      ...       ...
465040          우백장         I2       음식       I201        한식     I20107
465083         요거트퍼플         I2       음식       I212     비알코올      I21201
465195  컴포즈커피신촌자이엘라점         I2       음식       I212     비알코올      I21201
465238           신세계         I2       음식       I211        주점     I21101
465785          한사발포차         I2       음식       I201        한식     I20101

        상권업종소분류명          경도         위도                  대학교
178          요리 주점  126.924515  37.562797               연세대학교
363           경양식  126.945279  37.564840        연세대학교, 이화여자대학교
750          빵/도넛  126.935576  37.554661          연세대학교, 서강대학교
950          빵/도넛  126.946079  37.555476  연세대학교, 이화여자대학교, 서강대학교
1001       백반/한정식  126.944266  37.558839         연세대학교, 이화여자대학교
...            ...         ...        ...                 ...
465040  돼지고기 구이/찜  126.930443  37.555885               연세대학교
465083         카페  126.938264  37.555073          연세대학교, 서강대학교
465195         카페  126.942483  37.556859  연세대학교, 이화여자대학교, 서강대학교
465238   일반 유흥 주점  126.934638  37.555410               연세대학교
465785       백반/한정식  126.936190  37.558433               연세대학교

[2422 rows x 10 columns]
```

최종 결과 확인

- '연세대학교' 자리에 다른 대학교 이름을 넣어서 결과 확인 가능

```
# '연세대학교' 자리에 다른 대학교 이름을 넣어서 확인 가능
matching_rows = df_restaurant_final_copy[df_restaurant_final_copy['대학교'].apply(lambda x: '연세대학교' in x)]

# 결과 출력
print(matching_rows)
```

- 'restaurant_with_universities.csv' 파일로 저장