프로그램명 : combine.py

제작자 : 이문형 (캘커타 랩 인턴)

연락처 : moonstar114@naver.com

**프로그램의 목적**

프로그램은 CMS로부터 받은 파일을 각각 오픈하며 앞에 6개의 column을 해당 내용에 알맞게 생성하는 프로그램이다.

프로그램 환경

1. 리눅스

2. python3.5

3. pandas 패키지 사용

프로그램 설명

1. 프로그램의 흐름

1) 현재 working 디렉토리의 subdirectory 리스트를 얻음

2) 각 subdirectory 순회

3) subdirectory에 있는 csv파일리스트를 얻음

4) 3)에서 얻은 파일리스트를 통해 모든 csv파일을 오픈한다.

5) 오픈한 파일명을 parsing함

ex ) file\_name\_parse = ['AN','M','KR','FREE', 'TRAVEL', '201901','csv']

6) add\_column

7) outputfile 새로 생성하여 수정한 CSV 파일을 저장한다.

8) 파일리스트의 나머지 파일들도 3)~7)과정을 반복한다.

9) 나머지 subdirectory에서도 위 과정을 반복한다.

2. 설명

이 프로그램은 download\_csv\_file (csv파일 다운로드 자동화툴)의 연장선 상에 있습니다.

download\_csv\_file 로부터 calcutta CMS (<http://203.245.4.81:8888/>)의 csv파일을 받은 디렉토리가 있는 가정 하에 진행됩니다.

다운로드를 받고 나면, 디렉토리 내에는 파일 이름이 정리가 되어 있지 않습니다.

파일을 다음과 같이 모두 바꿔주셔야 합니다.

(os\_type)\_(date\_type)\_(country\_code)\_(type)\_(category)\_(date).csv (\*type은 FREE/PAID/GROSS를 의미)

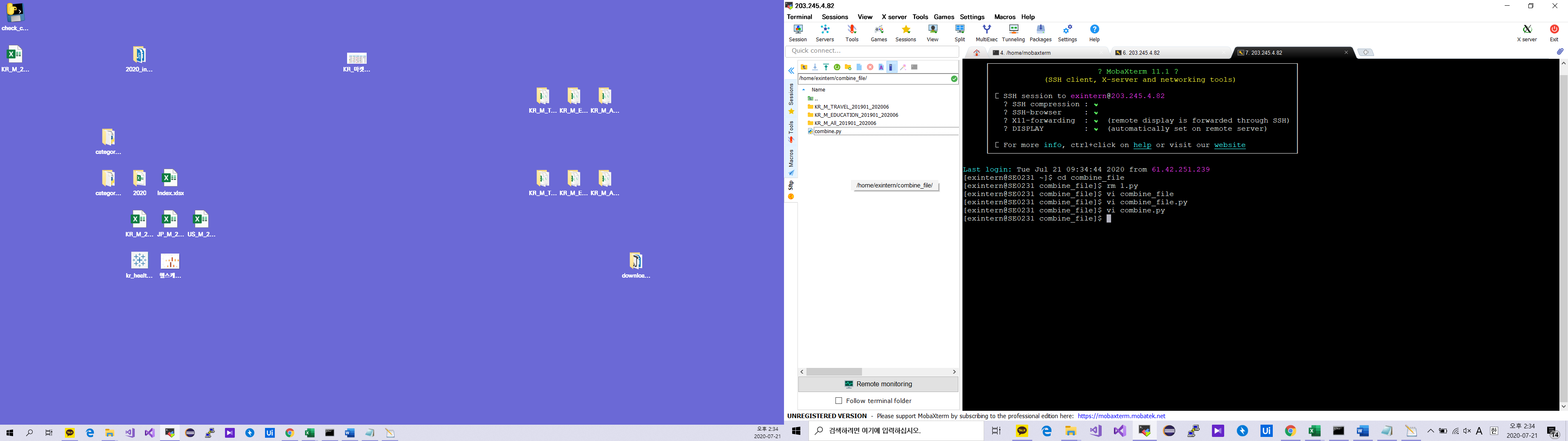
ex) 한국 안드로이드의 TRAVEL 카테고리의 FREE 타입 2019년 1월 데이터

AN\_M\_KR\_FREE\_TRAVEL\_201901.csv

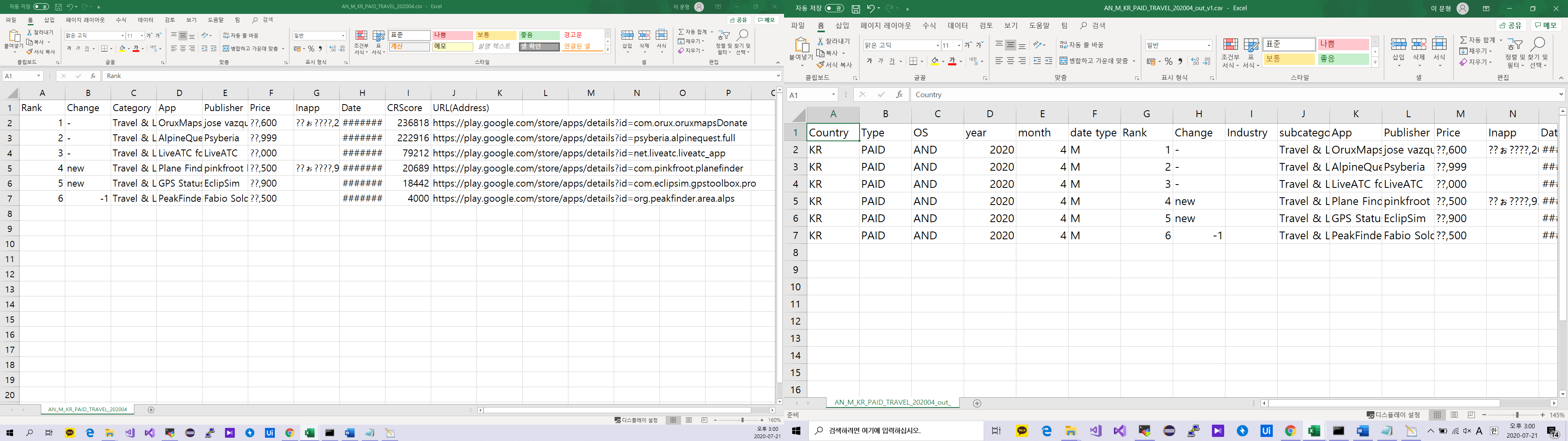
파일 이름을 parsing 해서 프로그램 내에서 사용하기 때문에 반드시 형식을 맞춰주세요.

코드 내에서 현재 working하고 있는 디렉토리의 subdirectory를 돌며 파일을 순회하는 방식을 택하였습니다.

따라서 이름을 바꾼 후에는 다음과 같이 폴더를 위치시켜야 합니다.

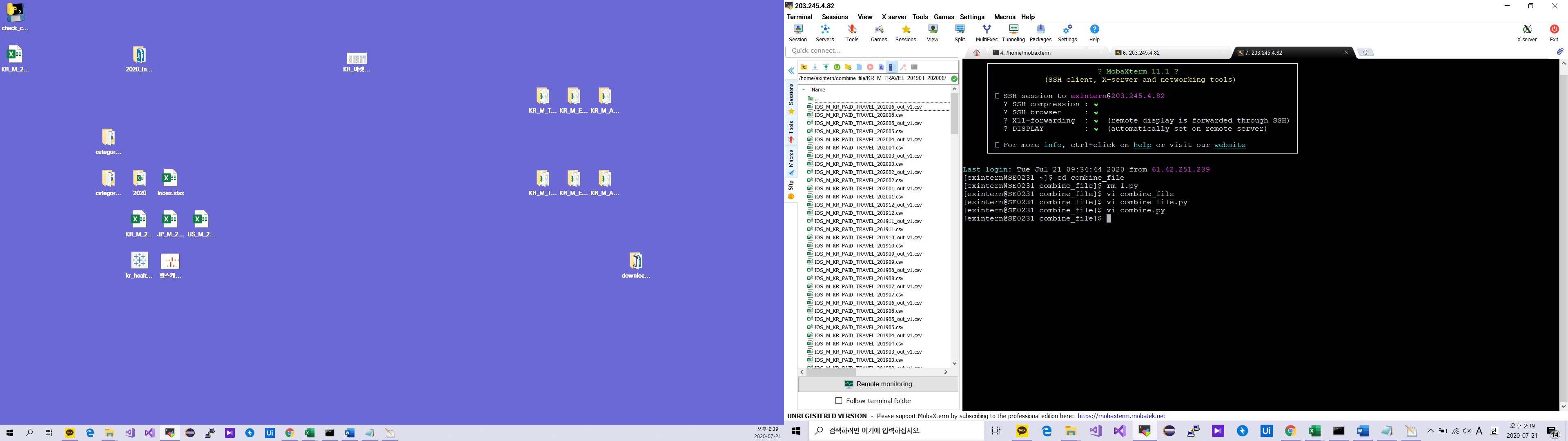


위치를 시킨 후에 shell에서 ~$ python3.5 combine.py 를 통해 실행을 하게 되면 디렉토리를 순회하며 안에 있는 파일을 각각 오픈하여 (국가, 운영체제 타입, 타입 (Free, Paid, Gross) , 연, 월, 데이트 타입) 의 필드를 채운 후 Industry의 빈 열을 생성합니다.

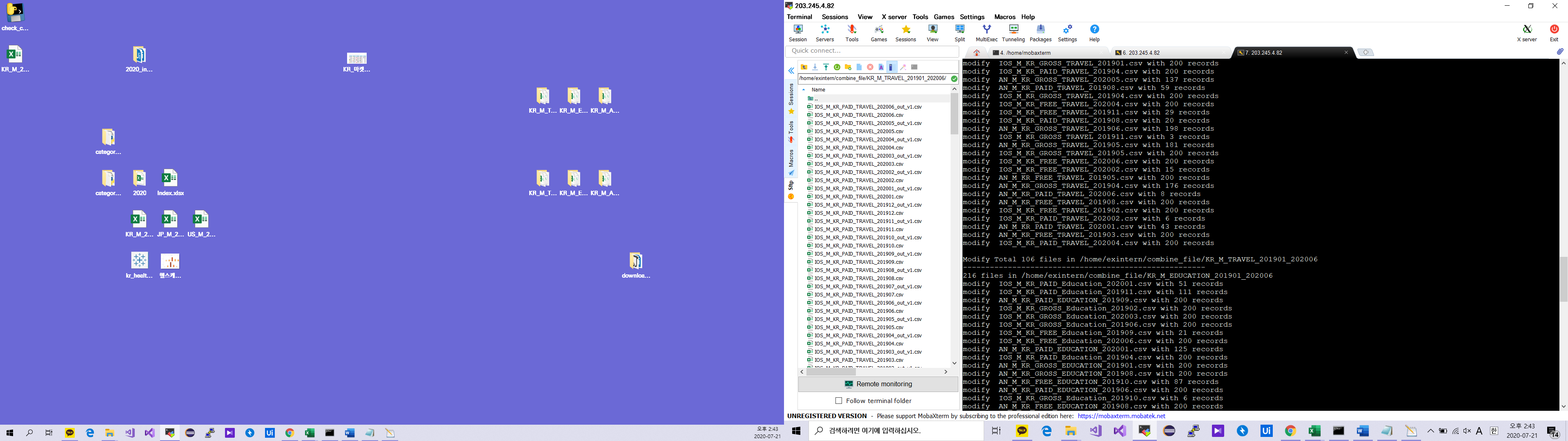


왼쪽은 원본, 오른쪽은 output파일입니다. Country, Type, Os, year, month, date\_type, Industry 필드가 생성됨을 확인할 수 있습니다. 기존 파일명\_out\_v1.csv 라는 파일을 기존 파일이 있었던 디렉토리에 생성합니다.

결과는 다음과 같습니다



각 파일을 돌며 생성할 때, 몇 개의 레코드가 있는지 출력하고, 해당 디렉토리에 몇 개 파일을 수정했는지 출력해줍니다.



위 과정을 마친 후에 output 파일들을 한군데에 합친 파일을 만들고 싶다면 다음과 같은 방법을 통해 진행할 수 있습니다.

첨부한 combine\_all\_csv\_file.bat 파일을 다음과 같이 수정하여 실행하면 됩니다. (수정법 : 우클릭->편집 )

copy \*out\_v1.csv (원하는 합본 파일 이름).csv

실행은 out\_v1.csv 파일이 있는 디렉토리에 옮겨놓고 사용하시면 됩니다.

합본을 얻으면 추가적으로 수정해야할 부분이 있습니다.

1) 한국어로 인코딩되어 파일이 깨지는 점

2) Country / type / OS / … 등 필드명을 나타내는 행이 여러 번 나타나는 점

1)번은 액셀 실행 -> 데이터 -> 데이터 가져오기 -> 파일에서 -> 텍스트/CSV -> (합본 파일 클릭) -> 한국어를 UTF-8로 변경 -> 데이터 변환 클릭

을 통해 해결할 수 있습니다.

2)번은 다음과 같은 순서를 따라 해결하실수 있습니다

(1) 필드명 행을 클릭하고 필터 적용

(2) 국가의 필터를 통해 'Country' 클릭

(3) 나온 결과 모두 삭제

(4) 첫번째 열 전체 클릭

(5) 홈 ->찾기 및 선택 -> 이동 옵션 -> 빈셀 체크 -> 삭제 탭 -> 시트 행 삭제 클릭