**Data Analysis**

**<programming>**

**Q1**

**문제0**

****

Pd.read\_csv로 파일 3개를 불러오고, 1,2열은 key가 index id와 중복된 데이터가 있으므로 drop 시킵니다.

**문제1**

(1-1) Pd.to\_datetime 함수와 dt.month 함수를 이용해서 새로운 month 열 데이터를 만들었습니다.이후 두개의 데이터셋(data\_2,data\_3)를 pd.concat 함수를 사용하여 행을 기준으로 합치고, ignore\_index=True 파라미터를 입력하여 인덱스 종속을 초기화 시켰습니다.



(1-2) pd.merge 함수를 사용하여 problem\_1\_1\_df와 data\_1을 joint 시킵니다. Pd.merge 함수는 inner merge를 디폴트 값으로 가지고 있기 때문에, 기준(‘url’)만 추가로 입력하여 완성시켰습니다.

**문제2**

(2-1) isnull 함수를 사용해서 null 여부의 Boolean 값을 데이터로 가지고 있는 데이터베이스를 만들고, 각 열을 기준으로 mean함수를 이용하여 평균을 구합니다

(2-2) 2-1 문제에서 확인한 null 비율에서 ‘fat\_content’와 ‘calcium\_content’ 칼럼의 결측치 비율이 60%가 넘는다는 사실을 발견 한 후, drop 함수를 이용하여 그 두 칼럼을 제거합니다.

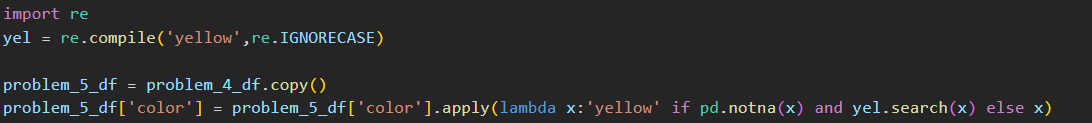
**문제3**

(3-1) if 문을 활용하여 함수를 생성합니다. Pd.isnull을 활용하여 문자열 또는 null값을 구분하고, 문자열이라면 split 함수를 활용하여 ‘,’를 기준으로 분리하는 함수를 생성합니다.

(3-2) 각 mixed, country\_mixed 열을 milk, country 열 바로 옆에 놓고 싶었고, 그래서 dataframe.insert 함수를 활용하고자 하였습니다. 중복참조 문제를 없애기 위해 .copy()함수를 사용하여 problem\_3\_fd 데이터를 만들고, apply함수를 활용하여 count\_mixed 함수를 적용시킨 뒤에 insert 함수를 활용하여 각 위치에 맞는 데이터를 삽입하였습니다.

**문제4** – problem\_3\_df의 두개의 열의 데이터가 모두 1이어야 하므로 비교연산자(&)를 활용하여 problem\_4\_df 를 생성하였습니다. 이 경우에서 논리연산자(and)를 쓰게 될 경우 series의 Boolean 값이 series로 산출되는지, Boolean 그 자체로 산출되는지 모호해지기 때문에 비트연산자인 비교연산자를 사용합니다.

**문제5** –

re 모듈을 통해 yellow 문자열을 찾는 모듈을 불러오고, ignorecase를 활용하여 대문자 식별은 하지 않습니다. 이후 중복참조 문제를 없애기 위해 .copy()함수를 사용하여 problem\_5\_fd 데이터를 선제적으로 만들고, apply 함수를 이용해서 yellow를 찾습니다. Apply 함수를 적용할 때 and가 사용되는데, null 값일 때 yel.search 함수가 적용되지 않는 부분을 방지하기 위해 pd.notna(x)를 먼저 판단합니다.

**문제6** – value\_counts() 함수는 series를 객체로 받기 때문에, country 열의 데이터를 시리즈로 불러오고, 이에 value\_count 함수를 적용하여 country 객체를 만듭니다. 이후 .to\_dict 함수를 사용하여 딕셔너리형태의 객체를 만듭니다.

**Q2**

아이디어: 일단 바닥에 구멍이 없다고 가정하고 문제를 풀고, 이후에 구멍을 고려했습니다. 구멍을 기준으로 input 리스트를 분리하고, 분리된 각각의 리스트에 구멍이 없다고 가정했을 때의 방법론을 똑같이 적용하여 더하면, 저류조에 고인 물의 총량을 구할 수 있습니다.

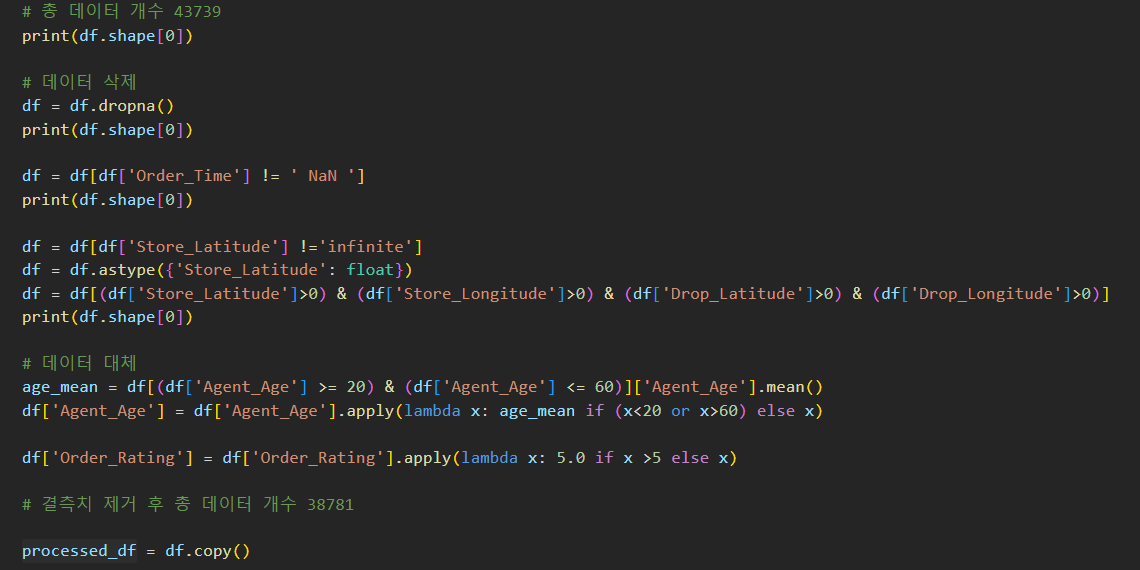
구현: 구멍이 없다고 가정했을 때, 저류조에 고인 물의 양을 구하는 것을 ‘trap\_water’ 함수로 구현하였습니다. Input 값은 저류조 리스트 데이터(list), output 값은 저류조에 고인 물의 양(int)입니다. 저류조 한 칸에 쌓일 수 있는 물의 양은 왼쪽, 오른쪽에서부터의 저류조 높이의 최댓값을 구한 뒤 그 두개의 최댓값중에서 더 작은 값을 의미합니다. 결국엔 파인 부분에 물이 고이게 되는데, 그 파인 부분에 담긴 물의 양을 결정하는 것은 좌측과 우측의 가장자리에 해당되고, 더 낮은 가장자리의 값일 것이기 때문입니다. 쌓일 수 있는 물의 높이를 구한 뒤, 만약 해당 칸의 저류조 자체의 높이가 물의 높이보다 크거나 같다면, 물이 고이지 않게 되고, 저류조 자체의 높이가 쌓일 수 있는 물의 높이보다 낮다면, 그 차이만큼 물이 차게 됩니다. 따라서, left\_max, right\_max 라는 두개의 리스트를 생성하고 (길이는 input 리스트의 길이) 각 칸에서 식별되는 왼쪽, 오른쪽의 최대 저류조 높이를 저장합니다. 이후 쌓일 수 있는 물의 높이를 water\_level 변수에 저장 한 뒤에, 실제 높이와 비교해서 쌓일 수 있다면 trapped\_water에 계속해서 더하여 전체 저류조에 고이는 물의 양을 계산합니다.

위의 구현된 함수를 가지고, segments 라는 리스트에, 구멍을 기준으로 분리된 새로운 리스트들을 저장합니다. 이후 segments에 저장된 리스트들 각각에 함수를 apply 시키고, 이를 더합니다.

**<EDA>**

**문제0**

이 데이터는 Order\_ID를 key로 가지고 있기 때문에 중복데이터는 존재하지 않습니다. 데이터들을 관찰한 결과 결측치를 가지고 있는 데이터, 배송출발지의 위도 값이 infinite인 데이터,배달원의 나이가 20~60세의 범위에 들어오지 않는 데이터, order\_rating이 5점을 초과하는 데이터 등이 존재한다는 문제점들을 발견하였고, 이들을 삭제, 대체 하였습니다.



결측치들을 제거하고, 위도 결측치 infinite를 제거하였습니다. 위도 결측치를 포함하고 있는 데이터의 개수가 약 4000개로 전체 데이터의 약 10%를 차지하였지만, 다른 대체 방법을 강구함에 어려움을 겪어서 일단 삭제하였습니다. 나이의 경우 정상범위(20~60세)를 넘어갈 때 평균 나이 값으로, 평점의 경우 이상치(5점을 넘는 경우)를 5점으로 대체하였습니다.

**문제1**

텍스트, 스크린샷, 디스플레이, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

* 시각화를 진행했습니다. Boxplot을 그렸고, 특징적인 부분은 Grocery 항목의 Delivery Time이 다른 카테고리와 비교하여 매우 낮다는 것입니다. 다른 항목들과 비교하여 식료품점의 접근성이 좋을 것이기 때문에 이러한 차이가 관찰되었다고 생각합니다.

**문제2**

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

* Delivery time과 order rating을 상관관계 분석을 살펴보면 약 -0.3으로 약한 음의 상관관계가 있는 것을 확인했습니다. 조금 더 시각적으로 파악하기 위해 넓은 범위에 흩뿌려져 있는 delivery time을 사분위수로 분류 한 뒤에 추가적인 시각화를 진행하겠습니다.

텍스트, 스크린샷, 소프트웨어, 멀티미디어 소프트웨어이(가) 표시된 사진

자동 생성된 설명

* Delivery time을 사분위수로 분류한 뒤, order rating에 따른 delivery time의 분포를 시각화했습니다. 4.5점 이전까지 배송시간 160분 이상에 해당되는 (빨간색) 4사분위수의 수가 많은 한편, 4.5 이상의 평점에 대해서는 오래 걸리지 않은 배송 건의 빈도가 급격하게 많아지는 것을 관찰할 수 있었습니다.

텍스트, 소프트웨어, 멀티미디어 소프트웨어, 스크린샷이(가) 표시된 사진

자동 생성된 설명

* 배송이 오래 걸린 건일수록 낮은 평점을 부여하는 것으로 나타났습니다. 이를 통해 delivery time과 order rating은 음의 상관관계를 가지고 있으며, 특히 4.5점 이하의 점수를 부여한 배송건에게서 더 극단적으로 나타났습니다.

**<modeling>**

Ss

**<전략도출>**

HD현대에너지솔루션은 HD현대의 계열사로서 세계수준의 태양광 셀과 모듈을 생산하고 있는 기업입니다. 이름값과는 달리

문제: 경쟁력이 없다

대안: 중동 등 새로운 시장에 대한 빠른 접근. 중공업 특화된 계열사 특성상 중동에 대한 어느정도의 영향력이 존재할 것.