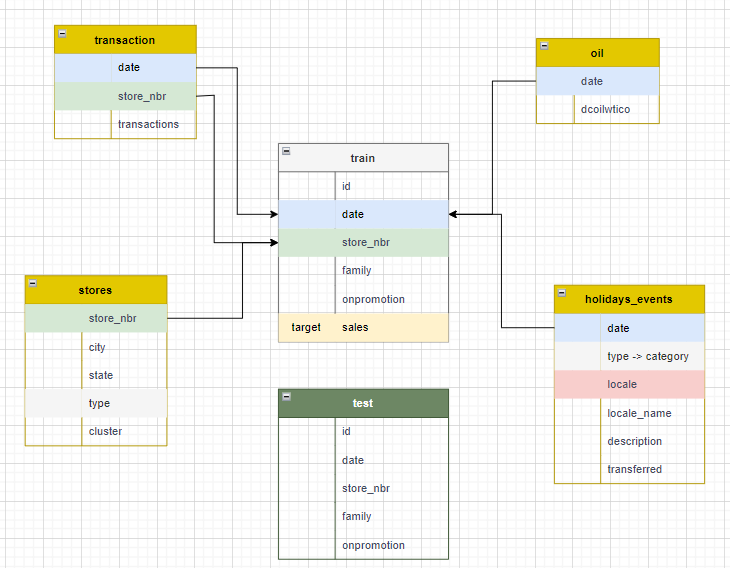
**[ERD]**

****

**a) train.csv**

- 학습용 데이터

- 데이터 건수: 300만개

- 데이터 시점: 2013-01-01 ~ 2017-08-15

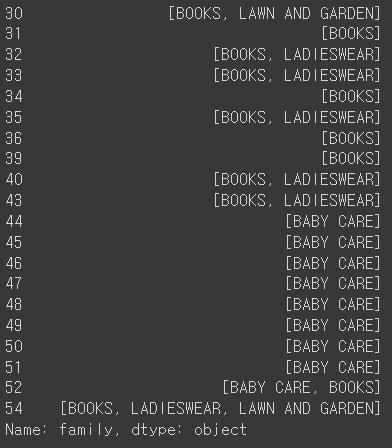
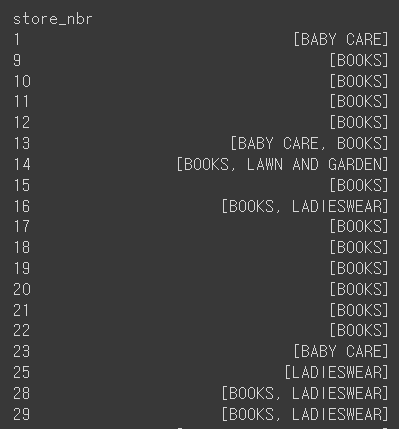
- 중복 데이터: x

**[Description]**

|  | **변수명** | **dtype** | **Null** | **설명** |
| --- | --- | --- | --- | --- |
| 1 | id(PK) | int64 | Not Null | - 각 데이터를 구분하기 위한 식별자 - 0 ~ 3000887(총 3000888건의 데이터)  - 빠진 날짜 4개를 제외하고 (날짜) \*(가게 수) \*(제품군 수) = 3000888건 |
| 2 | date(FK) | datetime | Not Null | - 판매일자 - 2013-01-01 ~ 2017-08-15 |
| 3 | store\_nbr(FK) | int64 | Not Null | - 매장 고유 식별번호 - 1 ~ 54 |
| 4 | family | object | Not Null | - 판매되는 제품 유형  - 33개의 제품군 |
| 5 | sales | float64 | Not Null | - 특정 날짜에 특정 매장의 특정 제품군의 총 매출액(금액) - 제품은 분수 단위로 판매될 수 있기 때문에 소수 값 가능  (예: 1.5 kg의 치즈, 1 봉지의 감자칩 등) - 0.00 ~ 124717.00 |
| 6 | onpromotion | int64 | Not Null | - 특정 날짜에 매장에서 프로모션되고 있는 제품군의 총 가짓수 - 0 ~ 741 |

**[family(제품군)에 대한 처리]**

| - 해당 매장에서 전 기간에 걸쳐 매출액 합계가 0인 제품군은 안 파는 제품군이 아닐까?  -> 확인 후 삭제  - test 데이터에는 16(일) \* 54(매장 수) \* 33(제품군) = 28512건의 데이터가 존재  -> 이후 예측 시 해당 제품군들에 대한 sales는 모두 0으로 예측해야 함 |
| --- |

****

**b) stores.csv**

- 매장 메타데이터

- 데이터 건수: 54개

- 데이터 시점: x

- 중복 데이터: x

**[Description]**

|  | **변수명** | **dtype** | **Null** | **설명** |
| --- | --- | --- | --- | --- |
| 1 | store\_nbr(PK) | int64 | Not Null | - 매장 고유 식별번호 - 1 ~ 54 |
| 2 | city | object | Not Null | - 매장이 위치한 도시  - 22개 도시 |
| 3 | state | object | Not Null | - 매장이 위치한 주  - 16개의 주 |
| 4 | type | object | Not Null | - 매장 유형  - A, B, C, D, E |
| 5 | cluster | object | Not Null | - 유사한 매장들의 그룹 - 1 ~ 17 |

**c) oil.csv**

- 일일 유가

- 데이터 건수: 1218개

- 데이터 시점: 2013-01-01 ~ 2017-08-31

- 중복 데이터: x

**[Description]**

|  | **변수명** | **dtype** | **Null** | **설명** |
| --- | --- | --- | --- | --- |
| 1 | date(PK) | datetime | Not Null | - 관측일자  - 2013-01-01 ~ 2017-08-31 |
| 2 | dcoilwtico | float64 | Have | - 유가  - 26.19 ~ 110.62  - 결측치 개수: 43개 |

**d) transaction.csv**

- 거래 데이터, 하루 동안 매장에 방문한 사람 수 또는 하루에 생성된 송장(영수증) 수

- 데이터 건수: 83488개

- 데이터 시점: 2013-01-01 ~ 2017-08-15

- 중복 데이터: x

**[Description]**

|  | **변수명** | **dtype** | **Null** | **설명** |
| --- | --- | --- | --- | --- |
| 1 | date(PK) | datetime | Not Null | - 거래일  - 2013-01-01 ~ 2017-08-15 |
| 2 | store\_nbr(PK) | int64 | Not Null | - 매장 고유 식별번호  - 1 ~ 54 |
| 3 | transaction | int64 | Not Null | - 거래량  - 5 ~ 8359 |

**e) holidays\_events.csv**

- 휴일 및 이벤트에 대한 메타 데이터

-> 공휴일의 규모가 가장 중요하다고 판단됨

ㄴ 국가 공휴일인 경우 전 지역, 전 매장이 영향을 받음

ㄴ 지역 공휴일인 경우 해당 지역(주, 도시)만 영향을 받음

- 데이터 건수: 350개

- 데이터 시점: 2012-03-02 ~ 2017-12-26

- 데이터 중복: O

**[Description]**

|  | **변수명** | **dtype** | **Null** | **설명** |
| --- | --- | --- | --- | --- |
| 1 | date(PK) | datetime | Not Null | - 공휴일 일자  - 2012-03-02 ~ 2017-12-26 |
| 2 | category  (변수명 수정) | object | Not Null | - 공휴일 속성  Additional: 추가 휴일 Bridge: 중간일, 휴일에 추가된 여분의 날 Event: 축제일(국가 이벤트) Holiday: 공휴일 Transfer: 변경된 휴일 Work Day: 안식일 |
| 3 | locale | object | Not Null | - 공휴일 지정 지역 범위  National: 국가 Regional: 지역(주)  Local: 지역(도시) |
| 4 | locale\_name(PK) | object | Not Null | - 공휴일 지정 지역명 |
| 5 | description | object | Not Null | - 공휴일 설명(이유) |
| 6 | transferred | bool | Not Null | - 공휴일 변경 유무  True: 이전 -> type이 Transfer임, 실제로는 다른 날에 축하됨 |

**[holiday에 대한 추가적인 정보 & 생각 정리]**

| 최종적으로 필요하다고 생각한 변수: **date, locale**  ㄴ 실제로 데이터 병합 후 분석에 영향을 줄만한 변수는 이 3개라고 생각했음  ㄴ 날짜는 당연히 중요함(시계열 분석이니)  ㄴ 국경일(National)이면 전국적으로 매출에 어떠한 영향이 있을 것이라 생각함  지역 기념일(Regional, Local)이면 해당 지역(주, 시)의 매출에만 영향이 있을 것이라 생각함 |
| --- |

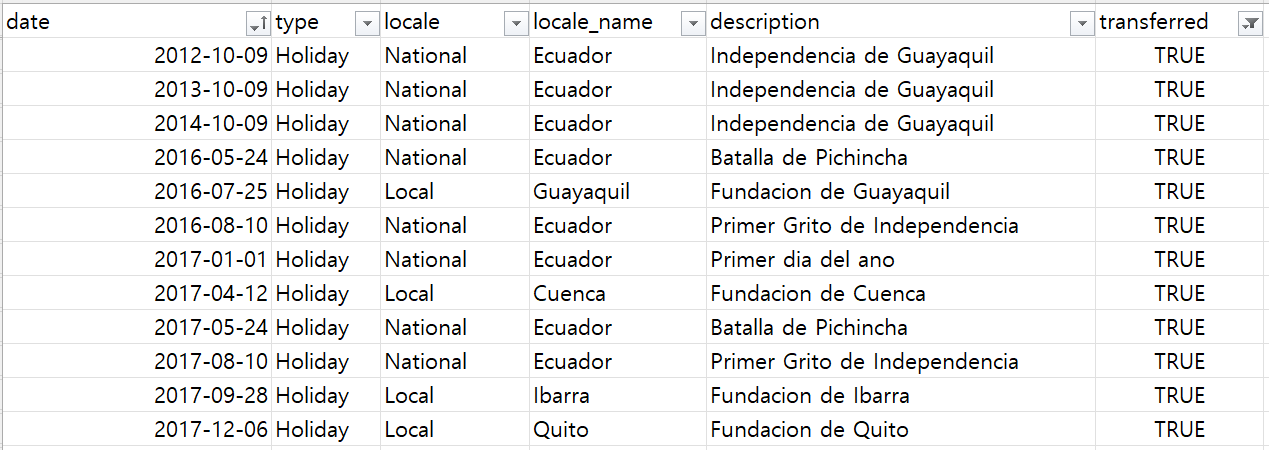
**1) category**

a. Transfer

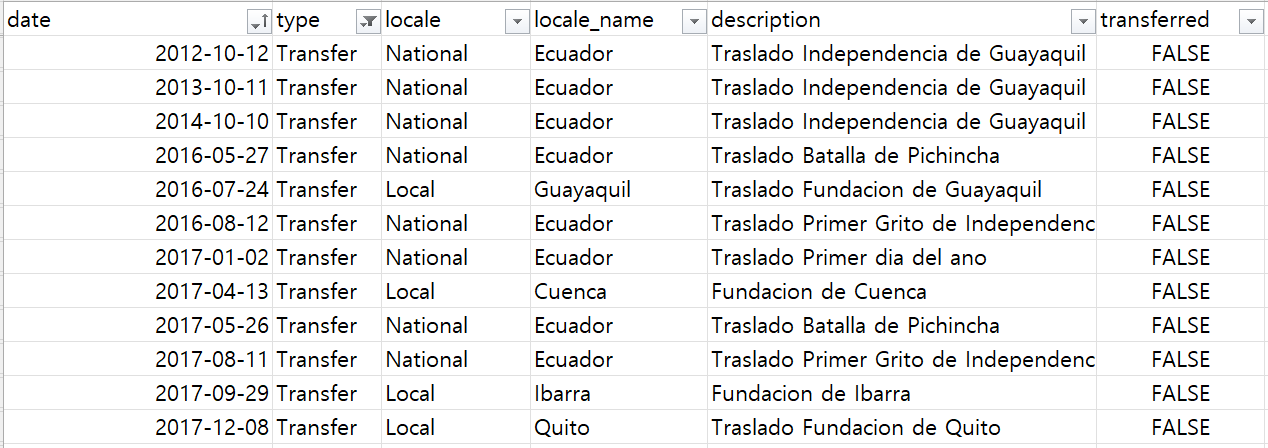
- 달력상 날짜는 유지하되, 정부에 의해 다른 날짜로 변경된 휴일

- 변경된 날은 휴일로 취급되지만, 실제로는 일반적인 날과 유사함(휴일 분위기보단 일반 분위기)

- 실제 축하된 날짜를 찾으려면 "type"이 "Transfer"인 해당 행을 확인하면 됨



▲ 사실 얘내는 휴일로 취급되지 않음.. -> holiday 데이터에서 삭제



▲ 얘내가 진짜 쉬는날임(위와 순서대로 일대일 매칭됨) -> 유지

| 예를 들어, Independencia de Guayaquil 휴일은 2012-10-09에서 2012-10-12로 이전되었으며, 이는 2012-10-12에 실제로 축하된 것을 의미 |
| --- |

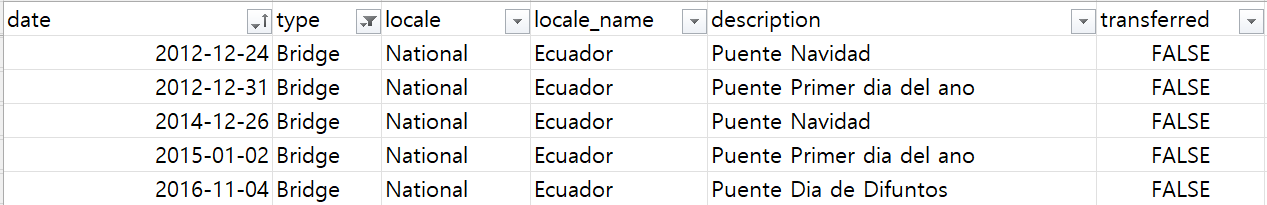
b. Bridge & Work Day

- Bridge: 휴일 기간을 연장하기 위해 추가된 날짜

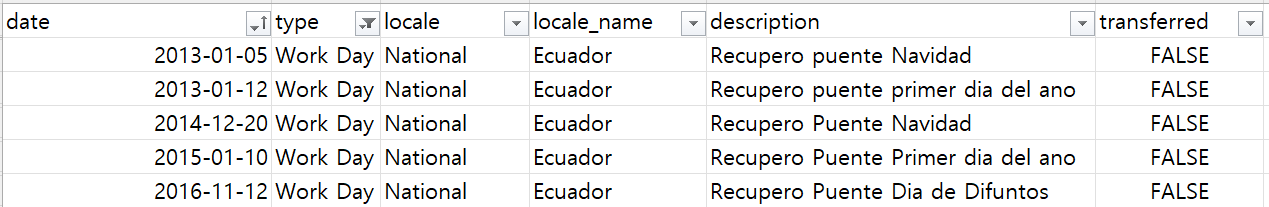
- 이러한 날은 원래는 근무일이지만 그냥 휴일로 취급

- 이를 Work Day로 보상

- 주말이지만 근무함(주로 주말이지만, 회사 출근함)



▲ 휴일임 -> 유지



▲ 주말인데 일함(휴일 아님) -> 데이터 병합 후 처리

c. Additional

- 보통 달력상 휴일에 추가된 날짜

- 일반적으로 크리스마스 주변에 발생하며, 이러한 날은 일반적인 휴일로 취급됨

-> 아무튼 노는날임(유지)

d. Event

- 국경일임(National)

e. Holiday

- 무튼 노는날?(왜 써놓은 건지는 모르게씀..)

**2) locale**

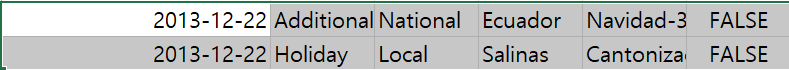
a. National

- 국가 공휴일

- locale\_name이 모두 ‘Ecuador’

- 해당 날짜에 Regional이나 Local이 겹치더라도 그냥 National 하나로 처리해도 될 듯함

(어짜피 전국적으로 다 노니까!)



▲ 이런 경우 위에만 남기겠다는 의미

- 이후 데이터 병합 시 **date**를 기준으로 병합

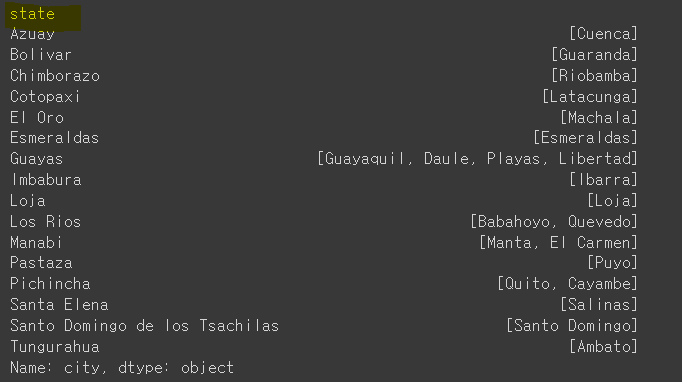
b. Regional

- 지역(주) 공휴일

- 주 > 도시

- 해당 날짜에 Local과 겹치면 그냥 Regional 하나로 처리해도 될 듯함

(지역별 확인은 필요)



▲ 각 state에 해당하는 city명

- 이후 데이터 병합 시 **date**와 **state**를 기준으로 병합(locale\_name이 state임)

c. Local

- 지역(도시) 공휴일

- 해당 도시의 매출만 영향을 받을 것임

- 이후 데이터 병합 시 **date**와 **city**를 기준으로 병합(locale\_name이 city)

**[최종 train 데이터]**

(일단 이렇게 생각했었습니다..)

|  | **변수명** | **dtype** | **Null** | **설명** |
| --- | --- | --- | --- | --- |
| 1 | id | int | Not Null | - 각 데이터를 구분하기 위한 식별자 - 0 ~ 3008015 |
| 2 | date | datetime | Not Null | - 판매일자 - 2013-01-01 ~ 2017-08-15 |
| 3 | store\_nbr | int | Not Null | - 매장 고유 식별번호 - 1 ~ 54 |
| 4 | family | object | Not Null | - 판매되는 제품 유형  - 33개의 제품군 |
| 5 | sales | float | Not Null | - 특정 날짜에 특정 매장의 특정 제품군의 총 매출액(금액) - 제품은 분수 단위로 판매될 수 있기 때문에 소수 값 가능  (예: 1.5 kg의 치즈, 1 봉지의 감자칩 등) - 0.00 ~ 124717.00 |
| 6 | onpromotion | int | Not Null | - 특정 날짜에 매장에서 프로모션되고 있는 제품군의 총 가짓수 - 0 ~ 741  - 모델링 시에는 제외해야 함 |
| 7 | city | object | Not Null | - 매장이 위치한 도시  - 22개 도시 |
| 8 | state | object | Not Null | - 매장이 위치한 주  - 16개의 주 |
| 9 | type | object | Not Null | - 매장 유형  - A, B, C, D, E |
| 10 | cluster | object | Not Null | - 유사한 매장들의 그룹 - 1 ~ 17 |
| 11 | dcoilwtico | float | Null | - 유가  - 26.19 ~ 110.62 |
| 12 | transactions | int | Not Null | - 거래량  - 5 ~ 8359 |
| 13 | holiday | object | Not Null | - 공휴일 종류  National:국경일  Regional: 지역(state) 휴일  Local: 지역(city) 휴일  —-----------------------------------  Weekend: 쉬는 주말  Weekday: 평일(+ 주말인데도 근무일, Work Day 유형 처리) |