<데이터 전처리>

1. na값 처리 이슈

1) na값이 발생하는 컬럼: country, agent, company, children

- company : 회사 이름으로 예약하였을 경우 회사 id. na값이 너무 많고, 중요하지 않다고 파악하여 배제.

- agent : 여행사를 통해 예약하였을 경우, 여행사 id. na값이 약 2천개 정도 발생. 특정 여행사에 의한 예약 취소가 발생 가능하므로, 일단 na값을 -1로 대체하고, 예약 취소 모델에는 사용하고 방문자 수 예측 모델에는 사용하지 않음.

- children: 4개의 na값이 있는데, 그냥 0으로 대체.

Hotel\_df[‘children’].fillna(0)

- country : 방문자의 국적. na값은 not\_defined 로 대체.

2. drop 할 컬럼 선정

Arrival\_date\_week\_number : 삭제한 이유> 이게 뭔지 모르겠어서 삭제. 월까지만 있어도 충분할거같음.

Hotel : 삭제한 이유 > 2개 밖에 없는데다, 그냥 포르투갈에 있는 대표적인 호텔의 이름임.

3. 범주형 데이터의 처리

날짜 데이터 – 연, 월 데이터 합친 후 연-월 날짜형 데이터로 변환.

*https://www.kaggle.com/atasaygin/hotel-booking-demand-eda-and-of-guest-prediction#Data-Cleaning*