

[티오더] 데이터 엔지니어 사전 과제

안녕하세요, 티오더 데이터 플랫폼팀입니다.

아래 데이터는 2022년 11월의 데이터 중 일부를 변환하여 랜덤 샘플링 데이터 입니다.

base knowledge

- 현재 매장의 주문 데이터는 7:00AM ~ 7:00AM(익일)을 기준으로 영업일로 보고 있다고 가정합니다.
- 각 오더에는 에러 정보를 저장하고 있으며, 편의를 위해 에러 텍스트가 있으면 True 없으면 False로 지정해 놓았습니다.

데이터의 명세서는 아래와 같습니다.

orders(주문 정보)

- id : pk
- store_code_enc : 주문이 일어난 상점 코드 암호화(null일 경우, Unknown 처리)
- time_order : 주문이 발생한 시각
- is_paid_order : 선결제가 일어난 주문여부
- is_first_order : 첫 주문이 일어난 주문여부
- is_error_exists : 해당 오더에 에러메시지가 있는지 여부

order_contents_information(주문 상품 정보)

- order_id : orders foreign key
- good_code : 상품 코드
- price : 상품 단위당 가격(null값 존재)
- quantity : 상품 주문 개수
- good_category : 상품의 카테고리(' ' 구분으로 여러가지가 나올 수 있으며 null 존재 가능)

데이터엔지니어 관련 수행해야 할 필수 과제는 아래와 같습니다.

- 조건 : Docker 로 서버 구성
<https://github.com/marclamberti/docker-airflow/blob/main/docker-compose.yml>

1. 첨부되어 있는 2개의 csv파일을 RDB(MySql 혹은 다른 RDB..)에 올립니다.
2. airflow를 구성하여 일일 통계와 월간 통계 생성을 합니다.
 - a. 일일통계
 - i. 전체 일일 에러수, 에러률, 주문수 통계
 - ii. 매장별 일일 에러수, 에러률, 주문수 통계
 - b. 월간통계
 - i. 전체 월별 에러수, 에러률, 주문수 통계
 - ii. 매장별 월별 에러수, 에러률, 주문수 통계
3. 전체 일일 통계가 완료되거나 전체 월간 통계가 완료 된 경우에는 알림을 보냅니다.
 - a. 알림은 특정 구글 시트에 올려 주시면 됩니다.

개인 Github Repository 생성하여 과제 수행후 URL을 과제 수령받은 메일에 작성하여 회신해주세요.

그리고 추가적으로 데이터 분석 관련 과제는 아래와 같습니다.

- 1) 1달 동안 회사 태블릿 주문 서비스를 이용한 주문추이를 보기 위해 시각화를 하고자 합니다.
11월 일별 주문 그래프를 시각화해 주시고, 추세가 어떠한지 설명해 주세요. (영업일 기준)
(추가적으로 그래프에 일별 에러율도 나오면 좋을 것 같습니다.)
- 2) 2022년 12월 1일 ~ 2022년 12월 3일에는 주문수가 어떻게 나올지 예측하고자 합니다. 주문수 예측을 해주세요.(영업일 기준)
- 3) 한 팀에서 회사 태블릿 주문 서비스의 하루 주문 중, 주류가 가장 많이 팔렸을 것이라고 주장하고 있습니다. 이를 뒷받침 하거나 반박하기 위한 데이터를 뽑아주세요. (단, 서비스 요청은 제외 부탁드립니다.)

위에서 진행한 내용에 대한 파일들은 추가로 메일로 첨부 주시길 바랍니다.

(e.g. python : jupyter or script, sql : sql file, excel : excel 첨부 등)

데이터 엔지니어 과제는 필수이고 추가적으로 데이터 분석 관련 과제까지 하시면 가산점이 있습니다.