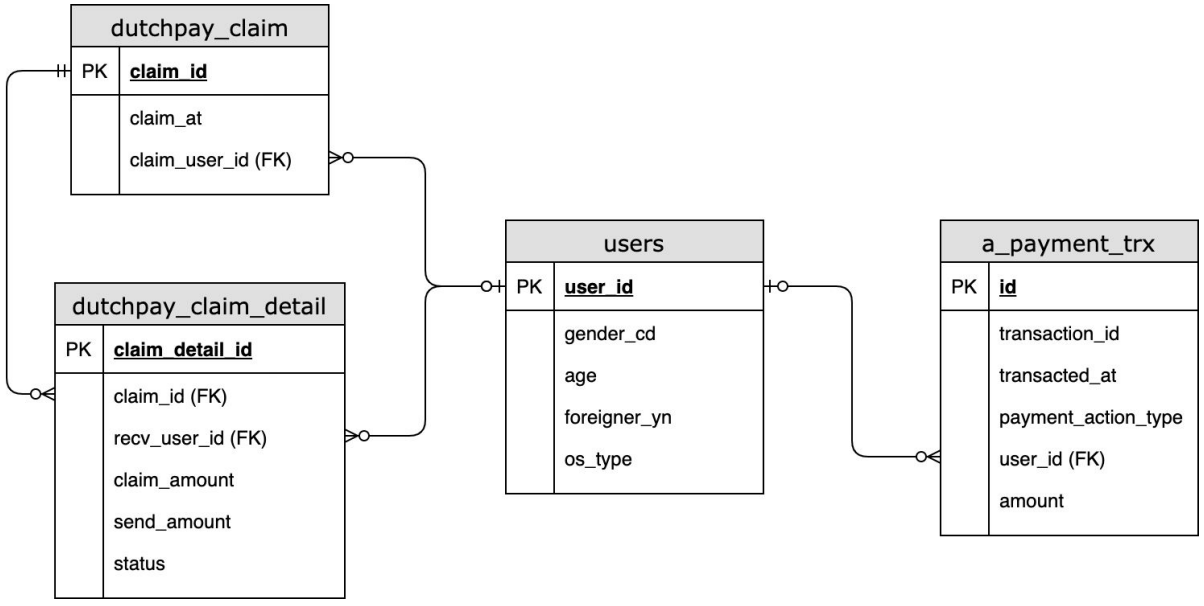
데이터 사이언티스트 (Data Scientist) 분석가 사전과제

아래 주어진 ERD와 Data를 기반으로 과제를 작성해주세요.
데이터는 csv 파일로 별도 첨부되어 있습니다. (**과제 3은 문제에 별도로 데이터셋 제시**)

※ 주의사항

- 원본 데이터 로드 단계부터 과제를 위해 진행한 과정 전체에 대한 스크립트를 첨부해주세요.
- 과제 2는 꼭 SQL 쿼리로 작성해주시고, 나머지 과제는 R 또는 Python 으로 작성해주세요.
- 과제 3에 대한 분석 리포트 제출 파일은 PDF, Markdown, HTML, PPT 등 제시하시는 인사이트를 보는 사람이 잘 이해할 수 있는 형태로 자유롭게 선택해주세요.
- 파일 전달은 Github 또는 메일로 전달해주세요.

1. ERD (과제 1, 2)



2. Data (과제 1, 2)

1) users.csv : 유저 테이블 (123,924건)

> user_id : 유저 ID

> gender_cd : 성별

> age : 연령

> foreign_yn : 외국인 여부

> os_type : 휴대폰 OS 유형

(sample)

user_id	gender_cd	age	foreigner_yn	os_type
01509a0865440e0	2	32	N	A
7d4697fbadb1c09	1	27	N	A
7b2a7724faf1400	1	26	N	A
b3876137175bec2	1	25	N	B

2) dutchpay_claim.csv : 더치페이 요청 테이블 (159,194건)

> claim_id : 더치페이 요청 ID

> claim_at : 더치페이 요청 일시

> claim_user_id : 더치페이를 요청한 유저 ID

(sample)

claim_id	claim_at	claim_user_id
4420721	2020-02-07 15:29:18	5cbd74112c55a0a
4420704	2020-02-07 15:26:54	5cbd74112c55a0a
4454342	2020-02-10 19:18:31	f077bc4ec8fd0ef
4453683	2020-02-10 18:15:11	f077bc4ec8fd0ef

3) dutchpay_claim_detail.csv : 더치페이 요청 상세 테이블 (557,644건)

> claim_detail_id : 더치페이 요청 상세 ID

> claim_id : 더치페이 요청 ID

> recv_user_id : 더치페이를 요청 받은 유저 ID

> claim_amount : 더치페이 요청한 금액

> send_amount : 송금한 금액

> status : 현재 상태

(sample)

claim_detail_id	claim_id	recv_user_id	claim_amount	send_amount	status
12918735	4075714	39476d42bd5f268	4		CLAIM
12918734	4075714	a84a2bf8ab324d3	4	4	CHECK
14666341	4577981	1d03cee8453eac1	12000	12000	SEND
14666340	4577981	bb21e5202ec9306	12000		CLAIM

4) a_payment_trx.csv : a 가맹점 결제 트랜잭션 테이블 (30,730건)

> id : 고유 트랜잭션 ID

> transaction_id : 트랜잭션 ID (결제 트랜잭션 발생시 채번되며, 취소 트랜잭션의 경우 취소되는 결제 트랜잭션과 같은 transaction_id를 사용함)

> transacted_at : 트랜잭션 발생 일시

> payment_action_type : 결제/취소 구분

> user_id : 결제/취소 유저 ID

> amount : 결제/취소 금액

(sample)

id	transaction_id	transacted_at	payment_action_type	user_id	amount
f7481b64e5b511e9bdcb5 09a4c8d0caf19080	3858754	2020-01-03 17:15:31	PAYMENT	789c51249d29be2	2000
811dc1bae89611e9bbe35 09a4c8d103b19080	4013773	2020-01-07 09:07:52	CANCEL	29930470b50a61a	74000
0166b4edeb5011e9aa115 09a4c8d0cf319080	4295919	2020-01-10 20:20:46	PAYMENT	ca85b790441d897	2500
0b274d35eaf611e991c05 09a4c8d0e6f19080	4260610	2020-01-10 09:36:48	PAYMENT	3c9c23defcaedf5	12000

3. 과제

1) '더치페이 요청에 대한 응답률이 높을수록 더치페이 서비스를 더 많이 사용한다.'라는 가설을 통계적으로 검정해주세요.
   - 해당 가설 검정 방법을 선택한 이유와 함께 전체 검정 과정을 기술해주세요.

2) 더치페이를 요청한 유저 중 a 가맹점에서 2019년 12월에 1만원 이상 결제한 유저를 대상으로 리워드를 지급하려고 합니다. 리워드 지급 대상자 user_id를 추출하는 SQL 쿼리를 작성해주세요.
   - 2019년 12월 결제분 중 취소를 반영한 순결제금액 1만원 이상인 유저만을 대상으로 함
   - 취소 반영기간은 2020년 2월까지로 함

3) 보스턴 지역의 지역별 집값에 영향을 미치는 요인을 정리한 데이터를 기반으로, 각 속성이 집 값에 미치는 영향을 분석해보려고 합니다. 아래 링크의 데이터 셋을 이용해서 보스턴의 집값에 영향을 미치는 Feature 들을 조합해서 각 지역의 집값(MEDV)을 예측하는 모델을 만들고, 해당 내용을 리포트로 작성해주세요.
   - 데이터 탐색 과정, Feature 선정 과정, 예측 모형 학습과정 및 Evaluation 결과, 고려사항 등을 리포트에 포함시켜 주세요.

   - 데이터 :

   https://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.data

   - 데이터 포맷 :

```
0.00632 18.00 2.310 0 0.5380 6.5750 65.20 4.0900 1 296.0 15.30 39 6.90 4.98 24.00
0.02731 0.00 7.070 0 0.4690 6.4210 78.90 4.9671 2 242.0 17.80 39 6.90 9.14 21.60
0.02729 0.00 7.070 0 0.4690 7.1850 61.10 4.9671 2 242.0 17.80 39 2.83 4.03 34.70
0.03237 0.00 2.180 0 0.4580 6.9980 45.80 6.0622 3 222.0 18.70 39 4.63 2.94 33.40
0.06905 0.00 2.180 0 0.4580 7.1470 54.20 6.0622 3 222.0 18.70 39 6.90 5.33 36.20
0.02985 0.00 2.180 0 0.4580 6.4300 58.70 6.0622 3 222.0 18.70 39 4.12 5.21
```

28.70
0.08829 12.50 7.870 0 0.5240 6.0120 66.60 5.5605 5 311.0 15.20 39 5.60 12.43 22.90
...

- 데이터 설명 :

1. Title: Boston Housing Data

2. Sources:
   (a) Origin: This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University.
   b) Creator: Harrison, D. and Rubinfeld, D.L. 'Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978.
   (c) Date: July 7, 1993

3. Past Usage:
   - Used in Belsley, Kuh & Welsch, 'Regression diagnostics ...', Wiley, 1980. N.B. Various transformations are used in the table on pages 244-261.
   - Quinlan,R. (1993). Combining Instance-Based and Model-Based Learning. In Proceedings on the Tenth International Conference of Machine Learning, 236-243, University of Massachusetts, Amherst. Morgan Kaufmann.

4. Relevant Information:
   Concerns housing values in suburbs of Boston.

5. Number of Instances: 506

6. Number of Attributes: 13
   continuous attributes (including "class" attribute "MEDV"), 1 binary-valued attribute.

7. Attribute Information:
   1) CRIM per capita crime rate by town
   2) ZN proportion of residential land zoned for lots over25,000 sq.ft.
   3) INDUS proportion of non-retail business acres per town
   4) CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
   5) NOX nitric oxides concentration (parts per 10 million)
   6) RM average number of rooms per dwelling
   7) AGE proportion of owner-occupied units built prior to 1940
   8) DIS weighted distances to five Boston employment centres
   9) RAD index of accessibility to radial highways
   10) TAX full-value property-tax rate per $10,000
   11) PTRATIO pupil-teacher ratio by town
   12) B 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
   13) LSTAT % lower status of the population
   14) MEDV Median value of owner-occupied homes in $1000's

8. Missing Attribute Values: None.