데이터 전처리와 EDA

1. Data Exploration
2. Data Cleaning
3. EDA
4. Data Transformation

의 순서로 데이터 전처리를 진행한다.

2. Data Cleaning
1)데이터 누락
(1)완전 무작위 결측
(2)결측된 이유가 결측된 열과 관련있음
--> 구매가를 안 알려주려고 의도적으로 구매가를 누락함
(3)결측된 이유가 결측되지 않은 열과 관련 있음
--> 어디에 사용goTsmsw 안 알려주려고 가격을 누락함

해결법
(1) 결측치 제거(행/열 제거)
(2) 평균값 넣기 - 대치
(3) k-nn이나 missforest(random forest)로 예측해 넣기 -대치
(4) 누락된 부분이 범주형 변수인 경우 최빈값도 사용 - 대치

2)데이터 중복
--> 제거

3) 데이터 오류
--> 수정, 제거도 가능하겠지

4) 튀는 데이터(아웃라이어)

해결법
(1) 제거하기
(2) 값 변경
(3) 가중치 조절

3. EDA

1) 데이터가 일변량/다변량

2) 시각화/비시각화 방법

3) Data Transformation

(1) encoding - 원핫, label(순서 고려x), ordinal(순서 고려)

(2) scaling - min-max, maximum absolute scaling, 표준화, robust scaling, log transform

(3) feature selection - 상관관계--> 낮으면 제거, 상호 정보량 --> 독립성이 강하면 제거