2. 필요한 이유

①데이터의 분포 및 값을 검토하여 수집한 데이터가 어떤 것을 나타내는지를 더 잘 이해하고,
수집한 데이터에 대한 잠재적인 문제를 발견하여 본격적인 분석에 들어가기 전에 수집의사를 결정하고 판단할 수 있어.

↑ 데이터를 통해, 주어진 데이터 셋에 대한 사용여부를 결정할 수 있다.

② 다양한 각도에서 살펴보는 과정을 통해,
문제 정의 단계에서 미처 발생하지 못했을 다양한 패턴을 발견하고, 이를 바탕으로 기존의 가설을 수정하거나 새로운 가설을 세울 수 있다.

3. 과정

: 기본적인 출발점은 문제 정의 단계에서 세웠던 연구 질문과 가설을 바탕으로 분석 계획을 세우는 것인데
해당 데이터에 대한 질문을 최대한 많이 만들어서 데이터를 표현하는 적절한 모형 등을 위한 데이터를 생성해야되.
분석 계획에는 어떤 속성 및 속성 간의 관계를 집중적으로 관찰해야할지,
이를 위한 최적의 방법은 무엇인지가 포함돼야해. 과정은 아래의 순서로 진행해

1) 분석의 목적과 변수가 무엇이 있는지 확인하고 개별 변수의 이름이나 설명을 갖는지 확인해야되.
이를 통해서 데이터가 어떤 용도로 만들어졌는지, 무엇을 설명하고자 하는지 확인하고
여러가지 질문이나 가설을 세울 수 있어.

2) 데이터를 전체적으로 살펴봐야해.
데이터에 문제가 있는지 없는지를 확인하는데 데이터의 첫부분이나 끝부분을 확인하고,
추가적으로 다양하게 데이터를 탐색할 필요가 있어.
① 수집한 데이터가 어떤 부분이 소실되어 결측값으로 나타나는지,
② 어떠한 변수에 이상치를 갖고 있는지 확인할 필요가 있어.

← 각 변수의 범위와 분포를 파악한다.

3) 데이터의 개별 속성값을 관찰해야해.
각 속성별로 갖고 있는 값이 예측한 범위와 분포를 갖는지를 확인하고
그렇지 않다면 어떠한 이유로 예측한 범위와 분포를 갖고 있지 않은지를 확인해야해

= 변수

4) "속성" 간의 관계에 초점을 맞추어 개별 속성 관찰에서 찾아내지 못했던 패턴을 찾아봐야해.
예를 들면 상관관계의 시각화를 통해서 각각의 변수가 어떤 관계를 띠는지 확인할 필요가 있어.

5. EDA의 형태

1) 그림으로 표현할지 여부에 따라서

: 데이터의 분포를 확인하고자 한다면 그림이 더 적합하고, 정확한 값을 필요하다면 수치로 표현하는 것이 더 적합해.

2) 단일변량 여부에 따라서

: 변수를 하나씩 확인할 것인지, 여러 변수를 동시에 확인할 것인지 여부에 대한 판단해야해.

이에 따라서 4가지 케이스로 EDA의 형태를 분류해봤어.

A. 단일변량이면서 수치로 표현하고자 하는 경우

** 범주형 자료 : 값의 범위, 빈도를 찾고자 할 때 빈도표를 작성하여 표현할 수 있어.

: 결측치가 존재한다면 비율이 100퍼센트가 안되기에 결측치의 유무를 판단할 수 있고, 구성비율을 파악할 수 있어.

** 양적 자료 : 샘플 데이터를 통해서 모집단의 개략적인 분포를 파악가능해.

- 통계량(sample statistics)

: 통계량의 중심위치(평균, 중위수 등)를 보고 자료의 치우친 정도를 판단하고,

산포(표준편차, IQR)를 통해서 퍼진 정도를 파악할 수 있어.

B. 단일변량이면서 그림으로 표현하고자 하는 경우

** 사용되는 방법

- 히스토그램 ← 양적 자료의 분포 파악

: 히스토그램을 통해서 자료의 분포가 어떻게 되는지, 어떤 모형을 가정할지, 이상치의 여부 등을 판단할 수 있어.

연속형 데이터에 사용하면 전체적인 분포를 파악할 수 있어.

- 막대그래프

: 범주형 자료의 분포 파악할 수 있으며, 연속형 데이터에 사용하면 전체적인 분포를 파악할 수 있어.

- 상자그림

: 이상치 판단, 자료의 치우침 여부 판단할 때 사용해.

- QQplot

: 표본이 정규분포에 근사시킬 수 있는지 여부를 판단할 때 사용해.

표본이 완전히 정규분포에 근사한다면 일직선이 그려져.

C. 다변량이면서 수치로 표현하는 경우

** 사용되는 방법

- 교차표

:범주형 자료에 사용되며, 각 행과 열에 변수를 지정하여 사용하며, 각 셀의 빈도나 비율을 파악할 때 사용할 수 있어.

- 공분산과 상관계수

: 두 변수가 양적 변수인 경우에 두 변수의 선형의 상관성에 대해서 판단할 때 사용함.

D. 다변량이면서 그림으로 표현하는 경우

** 사용되는 방법

- 산점도

: 변수가 양적 변수인 경우, 종속 변수를 y축에 두고 그리며 변수의 관계를 나타낼 수 있어.