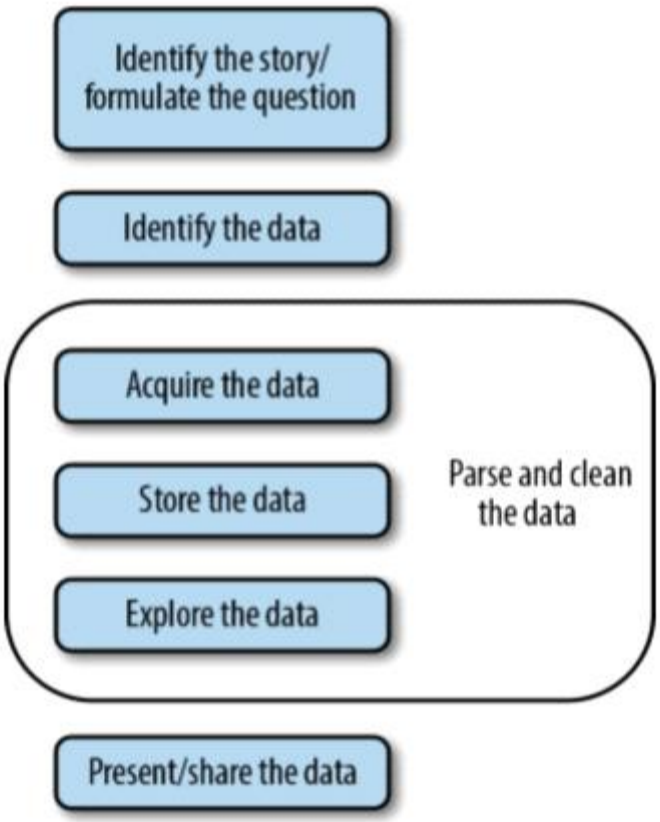
데이터 분석 이해하기

전형적인 데이터 분석과정은 일반적인 과학적 발견의 절차와 같다. 데이터 과학에서는 대답해야할 질문과 적용해야 할 분석 방법에서 발견이 시작된다. 가장 간단한 형태의 분석 방법은 기술(descriptive) 통계로서 데이터를 취합해 시각화한 형태로 표현한다. 주어진 데이터의 샘플의 크기가 작고 통계에 기반 한 추정에 적합하다. 예측(prediction)을 위한 데이터 분석은 과거의 자료를 사용하여 미래를 예측하며 한다. 인과(causal) 분석은 서로에게 영향을 미치는 변수를 식별한다.



데이터 분석을 위해서는 사전에 데이터 수집과 입력데이터 준비과정이 필요하다.

1. 데이터 수집: 웹 사이트에서 필요한 정보를 수집하거나 데이터를 추출함으로써 필요한 자료를 얻을 수 있으며, 데이터베이스, 웹 및 API를 이용하여 정보를 얻는다. 필요시 온도, 풍속 측정, 혈당량 또는 측정 가능한 것은 기기를 설치하여 얻을 수 있다. 시간과 노력을 절약하기 위해 공개되어진 자료를 사용하는 방법을 활용할 수 있다.

2. 분석 데이터 준비(전처리): 수집되어진 데이터를 유용한 형식으로 만드는 과정이 필요하다. 유용한 형식이란 분석에 필요한 도구에서 활용할 수 있는 자료형식 및 자료 구조[1]이다.

---

1) 파이썬을 사용한다면 리스트, numpy의 ndarray, pandas의 Dataframe 등이다.

분석에서 사용되어지는 자료의 속성 등에 따라 기존자료를 변환하는 과정을 수행한다. 데이터를 주의 깊게 살펴보는 과정으로 다수의 빈 값(empty values)이 없도록 하는 작업도 필요하다. 데이터 랭글링(Data Wrangling) 혹은 데이터 먼징(Data Munging) 과 같이 원자료(raw data)를 또 다른 형태로 수작업으로 전환하거나 매핑하는 과정을 수행한다.

데이터 분석 품질은 얼마나 좋은 데이터를 사용하느냐에 의해 결정된다. 데이터분석을 위해 필요한 자료인 데이터셋(data set)은 웹 또는 공개 API 을 통해 획득하고 데이터 정제 도구와 통계적 지식을 활용해서 데이터셋을 정규화해야 한다.

데이터를 정제한 후에는 기술 통계분석과 탐색적 분석을 수행한다. 이 단계에서 결과물은 산포도(scatter plot), 히스토그램, 통계적 요약이다. 정제된 데이터를 학습하여 모델을 구성하여 예측을 할 수 있다. 이 과정에서 예측 정확도를 평가하여 정확도가 낮은 경우 데이터 수집, 정재 및 재학습을 반복한다. 데이터셋의 품질은 통계학자와 프로그래머에 의해 결정하지 않으며 도메인 전문가에 의해 판단하게 된다.

여기서는 파이썬 언어를 사용한 데이터 분석의 준비 단계를 중심으로 설명하며 준비 단계에서는 데이터 수집, 전처리, 정리과정을 수행한다. 다른 단계에 비해 정형화되지 않아 다양한 창의적 접근이 가능하다.