미리 생각해보면 도움이 될 만한 체크리스트

**0. 주제 정하기**

- 데이터에 알맞은 분석방법을 선택할 것인가?

- 아니면 분석방법을 정한 후 그에 알맞은 데이터를 선택할 것인가?

- 현 시점에서 화제가 되고 있는 주제를 선택할 것인가?

- 전통적으로 인기있는 주제를 선택할 것인가?

- 주제가 법적, 혹은 도덕적으로나 정치적으로나 접근에 문제가 없는 주제인가?

- 해당 주제를 선택해서 분석했을 때 후에 얻는 점은 무엇인가? 어떠한 결론을 얻을 수 있는가?

- 정말로 분석할 만한 주제인가?

- 그 주제에 대한 데이터가 많은가? 일일이 수집해야 하는가?

- 공공데이터포털 등 기관에서 미리 수집해 놓은 데이터가 있는가?

- API를 사용해서 데이터를 불러와야 하는가?

- 데이터가 존재한다 하더라도, 그 양은 얼마인가? 분석에 충분한 양인가?

- 신뢰할만한 데이터인가? 왜곡된 데이터인가? raw데이터 수집 과정은 어떠했는가?

- 데이터가 정적인가? 아니면 수시로 변동하는 데이터인가?

- 데이터를 모으기 쉬운가??????????????? \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*ㅠㅠㅠㅠ

- 해당 주제에 대해 비슷한 종류의 논문이나 분석이 존재하는가?

- 선행 연구가 존재한다면, 그 연구에서는 어떤 방법으로 분석을 진행했는가?

**1. 주제가 정해지고 난 후, 데이터 결정 단계**

- 어떤 분석 툴을 사용할 것인가? (통계프로그램 : 주피터노트북, R, Stata, SPSS등)

- 어떻게 데이터를 수집할 것인가? (다운로드? 드래그? 수작업? 크롤링?)

- 얼마나 많은 데이터를 수집할 것인가? (용량)

- 얼마나 많이, 어느 기간동안의 데이터를 수집할 것인가? (시간)

- 어떤 분석방법을 사용할 것인가?

- 선형회귀분석? 비선형? 로지스틱 회귀분석? 군집분석? 머신러닝?

- 예측을 위한 분석인가? 요인 영향 분석인가?

- Panel 데이터인가? Timeseries 데이터인가? Cross sectional 데이터인가?

- 데이터에 numeric 변수는 몇 개고, binary 변수는 몇 개며, categorical 변수는 몇 개인가?

- 데이터에 결측값이 분석에 영향을 줄 정도로 많은가?

- 데이터 랭글링을 얼마나 해야 하는가? 데이터 전처리 난이도는 어떠한가?