오늘 세션 내용을 바탕으로, 본인 팀 데이터에 맞는 EDA를 어떻게 진행할 지, 자유 양식의 형태로 서술하여 제출해주세요.

1. 데이터 수집

* 데이터 정보: 스타트업 회사에 대한 펀딩, 투자, 인수합병 데이터셋
* funding\_rounds: 투자 종류, 투자 받은 회사, 투자금, 투자자 수 등 (투자 받은 회사 위주)
* investments: 투자 회사, 투자금, 투자 종류 (투자 한 회사 위주)
* organizations: company/ investor정보, employee, country, state 등 (투자 받은/한 회사 모두)

1. 데이터 전처리

* 결측치 처리, amount of fund raised 나 number of funding rounds/ number of investments 는 결측치 대신 0으로 표기되어있는 경우도 다수 있음. 이런 경우도 처리.
* 중복된 내용의 열, 분석에 불필요한 데이터(company description, url, update date) 정리
* Amount of fund raised 중심으로 status, number of employees, area, funding type, industry 등 데이터 정리

1. 데이터 Scaling:

* Funding type, category group 등의 자료에서는 카테고리가 불필요하게 많기 때문에 가장 빈도수가 높은 상위 카테고리 일부만 사용.
* Funding\_rounds, investments, organization 세 파일 모두 org\_uuid에 대해 분류될 수 있도록 처리 🡪 세 파일 병합 및 비교가 더 쉬울 것 같음.

1. 데이터 시각화

* 투자 종류나 투자 받는 기관을 계열별로 clustering 🡪 투자 동향 파악
* 투자 금액과 투자 종류의 관계 등 투자 규모 차이를 만드는 요인
* 투자 라운드는 회사의 성숙도를 나타내는 것이기도 함. 따라서 투자 라운드에 따라 회사를 구분하고, 특징 파악; 투자자 입장에서 투자 가치 판단에 필요한 요인 분석
* Status, number of employees, area, funding type, industry 등의 요소에 대해서 amount of funding 의 평균이 어떻게 차이가 나는지 분포 그려보기. 일변량 🡪 히스토그램으로 시각화, 다변량 🡪 산점도로 시각화.

1. 사후처리