**Team Exercise 2**

**Data integration**

The provided dataset for this exercise contains information about various job postings from LinkedIn. The datasets consist of 8 CSV files, each containing different types of job-related information.

실습을 위해서 주어진 데이터는 LinkedIn에서 Posting된 Job에 대한 여러가지 정보를 담고 있으며, 8개의 csv파일로 구성되어있다. 각 파일에 포함되어 있는 정보는 아래와 같다.

**Description**

This dataset contains information about job postings on LinkedIn. The data is divided into several files, each containing different aspects of the job postings:

1. job\_postings.csv: This file contains detailed information about each job posting, including the job title, description, salary, work type, location, and more.
2. companies.csv: This file contains detailed information about each company that posted a job, including the company name, website, description, size, location, and more.
3. company\_industries.csv: This file contains the industries associated with each company.
4. company\_specialities.csv: This file contains the specialties associated with each company.
5. employee\_counts.csv: This file contains the employee and follower counts for each company.
6. benefits.csv: This file contains the benefits associated with each job.
7. job\_industries.csv: This file contains the industries associated with each job.
8. job\_skills.csv: This file contains the skills associated with each job.

Task1-1

Import each of the CSV files into separate data frames. Study and explain the meaning, types, and values of the columns in each data frame.

모든 csv파일을 각각의 data frame으로 import하여 각 data frame의 column들의 의미와 type, value 등을 파악하여 보자.

Task1-2

Make a plan how to integrate the eight data frames into one data frame. Consider which data frames should be merged and how they should be joined (e.g., by company name, job ID, etc.). Describe the specific order and process of integration. Students are encouraged to make a visual diagram to explain your integration plan.

8개의 data frame을 어떻게 integration을 할지 계획을 세워보자. 가령 어떤 data frame과 어떤 data frame을 어떤 방법으로 합칠 것인지. 그 순서와 방법에 대해서 기술해보자. 가급적이면 도식(그림)을 사용해서 integration 방법에 대한 plan을 작성하는 것이 좋다.

Task1-3

Perform the actual data integration as planned above. Verify the integrated data frame to ensure no data is missing or errors occurred during the process. Create a checklist and perform verification to confirm that the integration is done correctly and properly. If errors occur, refine the process for error-free integration.

위 계획된 방법을 사용하여 실제 Integration을 수행하자. Integrated Data Frame을 검증하고 누락되는 데이터나 오류가 발생하지 않고 정상적으로 Integration이 되는지 확인하자. 어떤 check list를 제작해보고 verification을 수행하자. 검증 결과 integration 과정의 오류가 발생하였다면 위 과정을 다시 반복하여 오류가 최대한 없는 integration을 하도록 하자.

Task 2.

Use the integrated dataset to perform Exploratory Data Analysis (EDA) and analyze the job market. Answer the following questions through data summarization:

How does the salary range vary by industry?

How does the salary range vary based on required skills?

Do larger companies (in terms of employee count) post more job openings?

What differences are observed in jobs based on country or region?

What interesting differences exist between remote, on-site, and hybrid jobs?

Find any other interesting insights.

Integration된 결과를 활용하여 Job Market에 대해 파악하기 위한 EDA 및 데이터 분석을 수행해보자.

아래와 같이 궁금한 점들이 있다고 할 때, Data Summarization을 답을 찾아보자.

Industry별로 Salary range가 어떻게 다른지

요구하는 Skill에 따라 Salary range가 어떻게 다른지.

회사의 규모(직원의 수)가 많은 회사는 리쿠르팅도 많이 하는지.

국가나 지역에 따른 Job에서 어떤 차이점이 발견되는지

Remote job과 On-site job사이에 발견되는 특이한 차이점이 있는지? 혹은 Hybrid?

또 다른 흥미로운 insight를 찾을 수 있다면 찾아보자.

Task 3.

Assume that your team members are considering international job opportunities. Define the desired job requirements (e.g., job duties, salary, benefits, location, skills) and find job postings that match these preferences. Each team member should find and list their top 3 preferred jobs in order of preference.

조원들이 해외취업을 한다고 할 때, 원하는 직업의 요구 직무, salary 수준, 복지, 위치(국가 혹은 지역), skill등의 조건을 정리한 후 원하는 조건에 해당하는 job posting을 찾고, 선호하는 순서 top 3를 제시해보자.

(조원 수만큼 수행)