Project 2 Proposal

LinkedIn Job Postings

Team Members:
Cameron Dyal, Chirag Agarwal, Ricky Pang, Natasha Waliany

Team's GitHub repository:
https://github.com/UC-Berkeley-I-School/Project2_Pang_Dyal_Agarwal_Waliany

- Includes the data with the largest csv uploaded to github lfs
- ^ Update - no longer includes data file since it is a best practice to store the data elsewhere, outside of the repository
- Includes a starter notebook that left joins the dataset with the company data and also with the job information

Primary Dataset:
LinkedIn Job Postings (https://www.kaggle.com/datasets/arshkon/linkedin-job-postings)

LinkedIn is the world's leading professional network with a billion users in over 200 countries. LinkedIn was founded in 2002, prior to most of the other social networking sites prevalent today and this gave it a head start in establishing its vast user base. On this platform, businesses and individuals can connect, share information and expand their network. It is also a prominent platform for job listings with up to 61 million people searching for jobs every week on LinkedIn. Given its dominance in the professional networking space and ability to aid job seekers by posting their resumes, specific skills and connecting them with other professionals, we decided it would be an interesting opportunity to explore the job listings dataset on LinkedIn. This dataset contains a nearly comprehensive record of 33,000+ job postings listed over the course of 2 days, months apart.

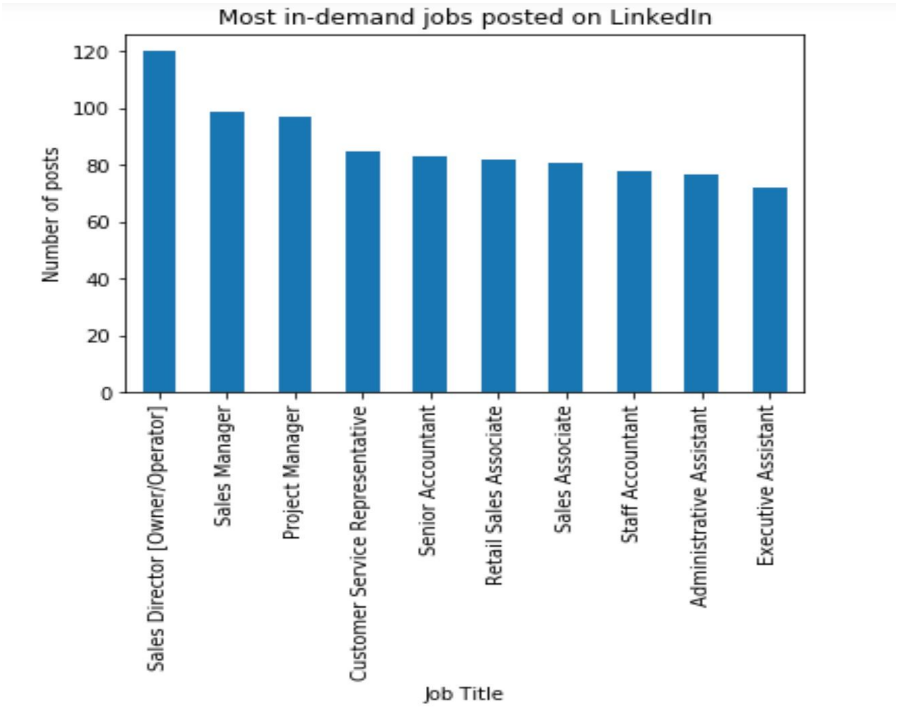
Initial plots, figures, or tables:

- **job_posting.csv** — This table will be useful for exploring a vast amount of job posting information, including average salary, remote work availability, and more.
- Company details:
    - **companies.csv** — This table maps each company ID listed in the job_posting.csv's "company_id" column with the company's name.
      company_industries.csv — This table maps each company ID listed in the job_posting.csv's "company_id" column with the company's industry.
      company_specialties.csv — This table maps each company ID listed in the job_posting.csv's "company_id" column with the company's industry.
      employee_counts.csv — This table maps each company ID listed in the job_posting.csv's "company_id" column with the amount of employees and followers the company has listed on LinkedIn.
- Job details:

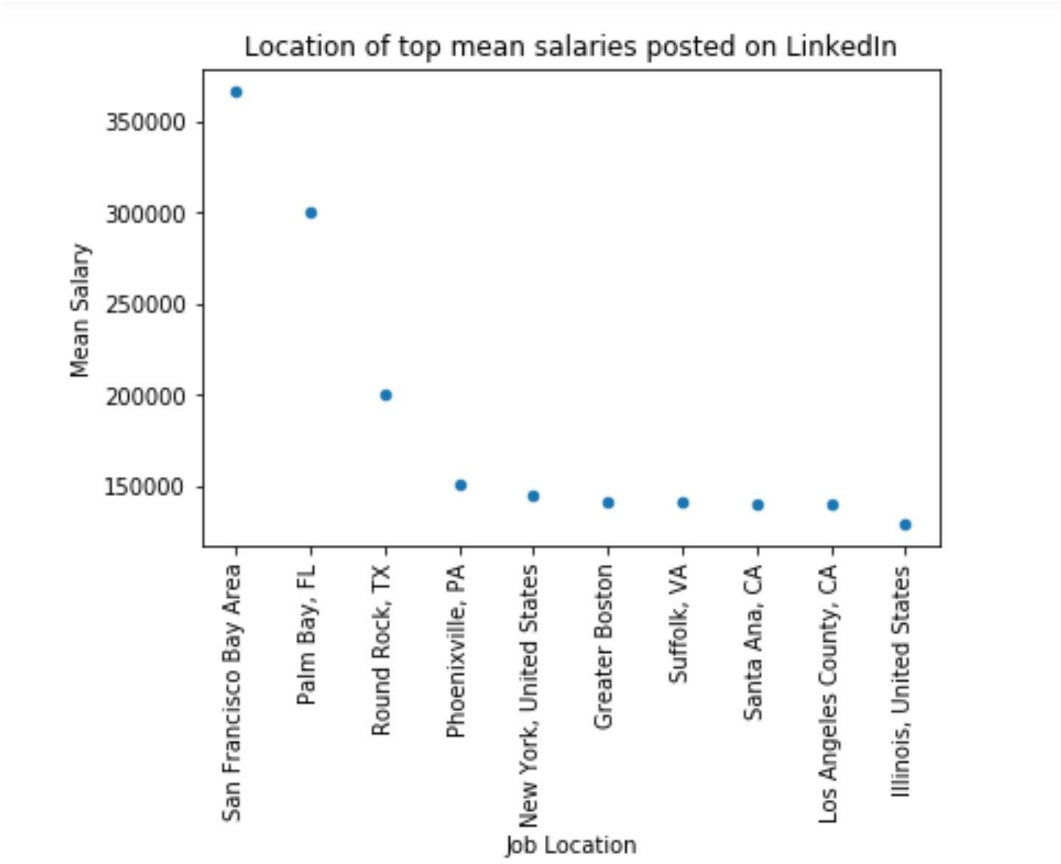
    - **benefits.csv** — This table maps each job ID listed in the job_posting.csv's "job_id" column with the role's benefits. The table will be useful for exploring which benefits were most commonly offered for job's posted on LinkedIn during 2023.
    - **job_industries.csv** — This table maps each job ID listed in the job_posting.csv's "job_id" column with each role's industry.
    - **job_skills.csv** — This table maps each job ID listed in the job_posting.csv's "job_id" column with each role's listed skills.
    - **salaries.csv** — This table maps each job ID listed in the job_posting.csv's "job_id" column with each role's listed min, med, and max salary.
- Maps:
    - **industries.csv** — This table maps each industry ID listed in the job_industries.csv's "industry_id" column with the name of each industry.
    - **skills.csv** — This table maps each skill ID listed in the job_skills.csv's "skill_id" column with the name of each skill.

Plots:

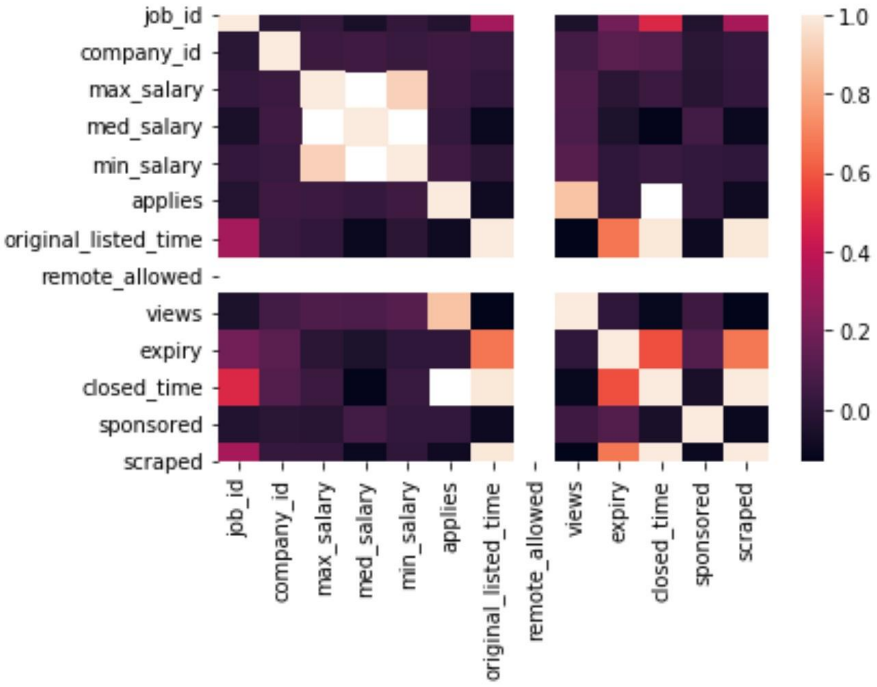
We created some initial plots to explore our dataset. This bar plot shows the 10 most in-demand jobs in the dataset. Perhaps after cleaning the data further, this might reveal different insights.



We grouped the dataset by location and median salary and plotted the top mean salaries by location.



We also created a heatmap of the correlation matrix to identify relationships between variables that we can further investigate.



<u>Some of the variables (column names) you intend to explore and what kind of insights you expect to glean:</u>

- **title** — This variable describes the title of each job posting's role. This table will be useful for exploring:
    - *Most common job titles sought after in job postings.*
    - *Correlation between title and other job ID variables.*
- **{ }_salary** {min, med, max} — This variable describes the minimum, median, and maximum salary of each job posting. This table will be useful for exploring:
    - *The average minimum salary offered.*
    - *The average maximum salary offered.*
    - *Percentage of job postings that offer an hourly salary vs an yearly salary.*
    - *Correlation between salary offered and job title/experience level.*
- **formatted_experience_level** — This variable describes the work type (entry-level, internship, executive, etc.) of each job posting. This table will be useful for exploring:
    - *The most common work type offered for job postings.*
    - *Correlation between work type and benefits offered.*
- **location** — This variable describes the location of each job posting. This table will be useful for exploring:
    - Locations with the highest number of job postings.
    - Correlations between location and industry.

- **company_id** → {name, specialty, industry, employee_count} — This variable describes the company ID of each job posting, which is utilized in other data to represent each company's name, specialty, industry, and employee count. This table will be useful for exploring:
    - *Companies with the highest number of jobs posted.*
    - *Industries with the highest number of jobs posted (and therefore industry growth).*
    - *Specialties with the highest usage.*
    - *Average company size of hiring companies.*
- **job_id** → {type (benefit), industry_id, skill_abr, salary_id} — This variable describes the job ID of each job posting, which is utilized in other data to represent each job posting's benefits, industry, skills, and salary details. This table will be useful for exploring:
    - *The most common benefits offered by jobs posted.*
    - *The most sought after skills for jobs posed.*
    - *The industry with the highest number of jobs posted.*
    - *Average salary for jobs posted.*

Supplemental datasets, if any, to complement your primary dataset - this means links, columns that you'll join on, etc.,

- The primary dataset will be joined with company details and job details on company_id and job_id.
- By joining these datasets we will gain insights into the industry the job belongs to, the size of the company, the skill set required, and the salary range belonging to the job posting.

What you plan to cover in the final report and how you plan to organize it.

- We will organize the final report as follows:
    - Business problem (background on LinkedIn)
    - Data Extraction and Manipulation (introduction to the dataset)
    - Exploratory Data Analysis (summarizing data, missing values, hypotheses)
    - Data augmentation
    - Visualizations to support findings
    - Conclusion
- Within the final report we will primarily be focused on job titles and the number of them as they are an indication of which roles are available on the platform for hire.

Plan of Action

- ☐ First we will need to join the datasets together into one main pandas dataframe and also create new columns where we would like to see grouped and aggregated statistics for data points.
- ☐ We will then explore the data features and their unique values and employ visualization as needed.

- [ ] Next we will need to clean the data and get rid of any major outliers. We will standardize and reformat the data as required. We will also identify and analyze missing values and duplicates and decide how to resolve them.
- [ ] We will also check the correlation between the attributes and explore the relationship between attributes that have a high correlation to find patterns and trends.
- [ ] We will form questions regarding the dataset and investigate these questions further using the clean dataset.
- [ ] Lastly we will need to create unique tables with grouped rows and aggregated metrics and then plot visualizations for each of the listed characteristics we would like to observe from above.