NYU

SPRING 2024 BIG DATA: CSGY 6513-D

Instructor: Mr. Amit Patel

JobLens: Leveraging Big Data for enhanced Job Market Analysis And skill-based Recommendations

Chenchen Guo
cg4421@nyu.edu
MS Computer Engineering
New York University

Jiaxin Dong
jd5226@nyu.edu
MS Computer Engineering
New York University

Mohammed Zakriah Ibrahim
mi2471@nyu.edu
MS Computer Engineering
New York University

Github Repo: https://github.com/JD5226/BigData/tree/main/Project

Agenda

- Abstract
- Objectives
- Technologies used and Data Source
- Data Cleaning and Transformation
- Key Findings & Insights
- Recommendation System implementation
- Conclusion and Lessons Learned

NYU

Abstract

This project proposes to utilize a comprehensive dataset containing 1.3 million job listings scraped from LinkedIn, augmented with detailed job skills information, to gain insights into the current job market trends, identify skill gaps, and develop a job recommendation system. The dataset, a rich source of information on job titles, industries, companies, and required skills, offers an unprecedented opportunity to analyze and address the needs of the modern workforce.

NYU

Objectives

- Perform Exploratory Data Analysis (EDA) on job market datasets to uncover patterns and insights.
- Identify emerging trends in job titles and industries to guide job seekers' focus.
- Map out in-demand skills across sectors to align job seekers' skill sets with market needs.
- Conduct a thorough analysis of the job market across industries, locations, and company types.
- Develop a job recommendation system to match job seekers with suitable listings.

Technologies used and Data Source

Data Cleaning, Statistical Analysis and Search Algorithms:

- Python 3.10+
- Pandas
- PySpark
- Jupyter Notebook

Visualization and GUI:

- Matplotlib
- Seaborn
- Streamlit/Gradio

Dataset Source:

- 1.3M LinkedIn Jobs & skills
- Dataset size: 2GB
- Number of Record: 1,296k
- Link: https://www.kaggle.com/datasets/asaniczka/1-3m-linkedin-jobs-and-skills-2024/data

NYU

Data cleaning and Transformation

1. Data Cleaning:

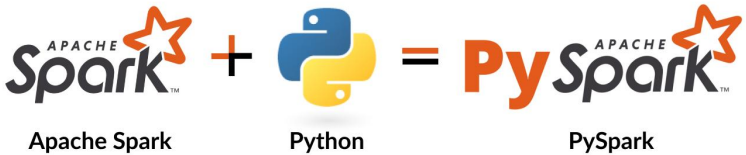
- Removing duplicate job listings to ensure data accuracy and consistency.
- Handling missing values in job titles, industries, and skills fields through imputation or deletion based on relevance and impact.
- Formatting data to ensure uniform consistency. Validating and cleansing job skills information to ensure relevance and accuracy in skill mapping.

2. Data Transformation:

- Aggregating job listings by industry, location, and company type to facilitate macro-level analysis of job market trends.
- Merged and harmonized multiple csv files and dropped unnecessary columns
- Normalizing numerical data to ensure consistency and comparability across different job listings.

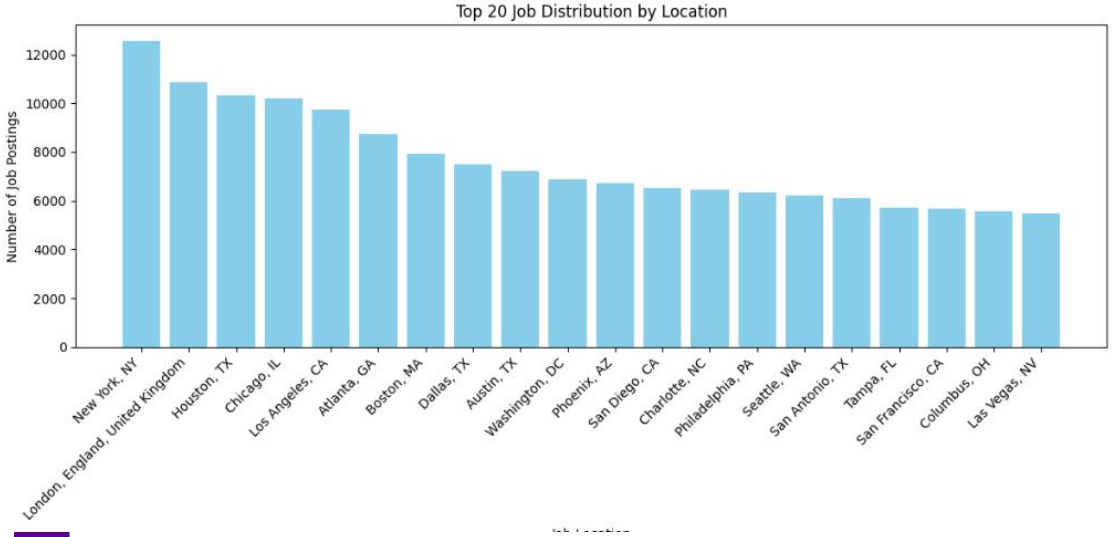
Data Analysis with PySpark

The PySpark module performs extensive data mining tasks. It includes analyzing job distribution across different locations, company recruitment activities, and job title frequencies. This component utilizes Spark sessions to load, process, and visualize data from CSV files, providing insights into job demand, industry trends, and the popularity of various job titles. It enables data-driven decision-making for job seekers and recruiters by identifying key trends and demands in the job market.



Key Findings & Insights:

Job Distribution by Location: Analyze the distribution of job postings across different locations to identify regions with high demand for specific job roles.



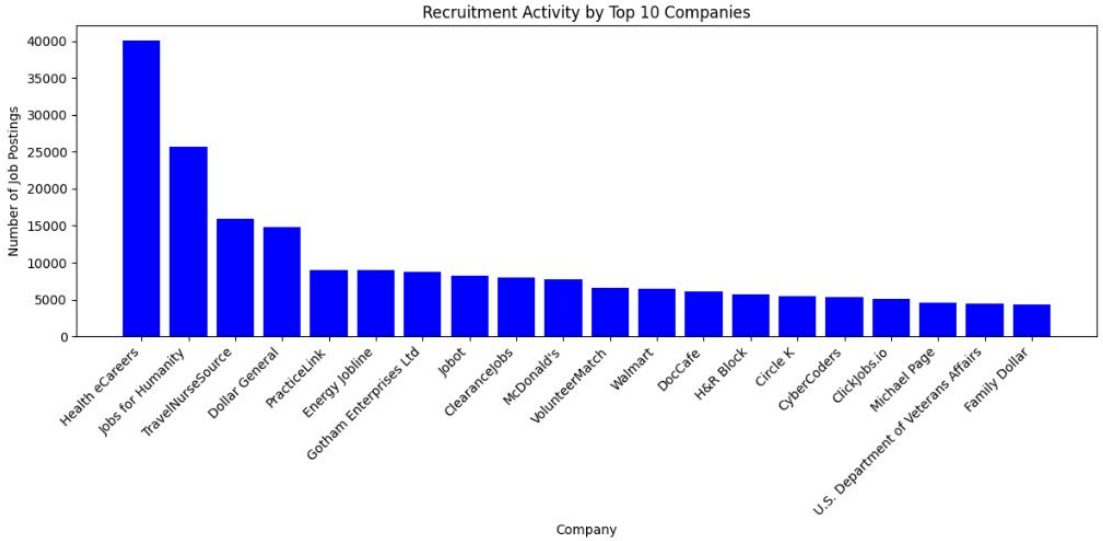
job_location	count
New York, NY	12579
London, England, ...	10878
Houston, TX	10332
Chicago, IL	10187
Los Angeles, CA	9736
Atlanta, GA	8738
Boston, MA	7924
Dallas, TX	7514
Austin, TX	7235
Washington, DC	6869
Phoenix, AZ	6722
San Diego, CA	6532
Charlotte, NC	6470
Philadelphia, PA	6326
Seattle, WA	6235
San Antonio, TX	6102
Tampa, FL	5701
San Francisco, CA	5684
Columbus, OH	5552
Las Vegas, NV	5468

only showing top 20 rows

NYU

Key Findings & Insights:

Company Recruitment Activity: Analyze the recruitment activity of different companies to identify organizations actively hiring for various positions.



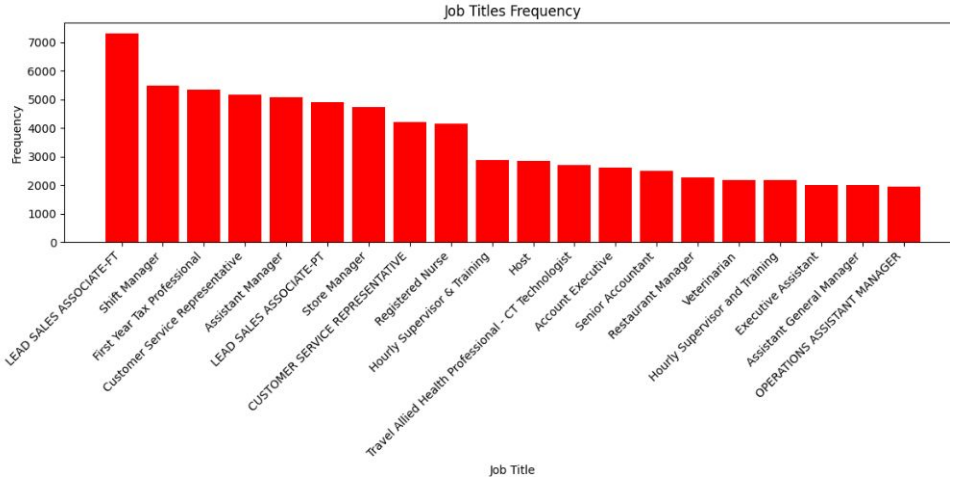
company	count
Health eCareers	40047
Jobs for Humanity	25629
TravelNurseSource	15997
Dollar General	14775
PracticeLink	9043
Energy Jobline	8987
Gotham Enterprises Ltd	8700
Jobot	8264
ClearanceJobs	8015
McDonald's	7742
VolunteerMatch	6653
Walmart	6455
DocCafe	6026
H&R Block	5668
Circle K	5493
CyberCoders	5273
ClickJobs.io	5105
Michael Page	4560
U.S. Department of Veterans Affairs	4473
Family Dollar	4349

only showing top 20 rows

NYU

Key Findings & Insights:

Job Titles Frequency: Determine the frequency of different job titles to understand popular job roles in the dataset.



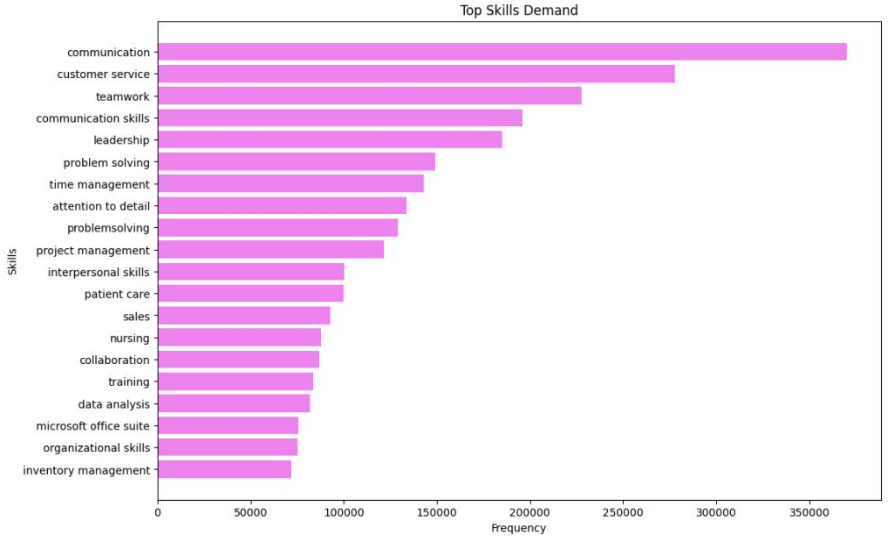
job_title	count
LEAD SALES ASSOCIATE-FT	7315
Shift Manager	5500
First Year Tax Professional	5351
Customer Service Representative	5165
Assistant Manager	5067
LEAD SALES ASSOCIATE-PT	4911
Store Manager	4739
CUSTOMER SERVICE REPRESENTATIVE	4214
Registered Nurse	4142
Hourly Supervisor & Training	2883
Host	2861
Travel Allied Health Professional - CT Technologist	2717
Account Executive	2614
Senior Accountant	2497
Restaurant Manager	2280
Veterinarian	2194
Hourly Supervisor and Training	2179
Executive Assistant	2021
Assistant General Manager	1998
OPERATIONS ASSISTANT MANAGER	1960

only showing top 20 rows

NYU

Key Findings & Insights:

Skill Demand Analysis: Identify the most in-demand skills across job postings to understand the skillset required for different roles..

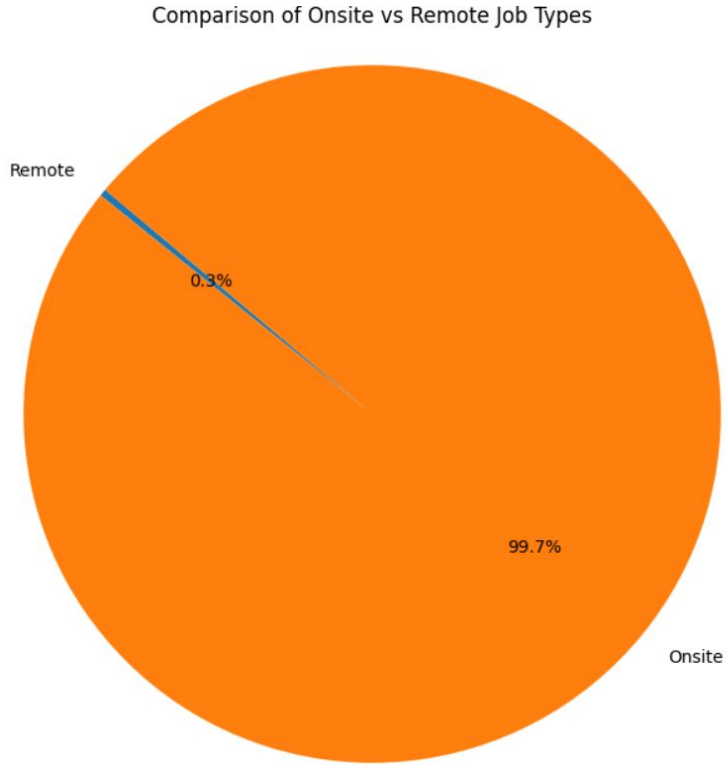


communication: 370020
customer service: 278012
teamwork: 227535
communication skills: 195820
leadership: 185134
problem solving: 148987
time management: 142861
attention to detail: 133916
problemsolving: 129293
project management: 121515
interpersonal skills: 100218
patient care: 99906
sales: 92977
nursing: 87945
collaboration: 87080
training: 83638
data analysis: 81945
microsoft office suite: 75507
organizational skills: 75245
inventory management: 71899

NYU

Key Findings & Insights:

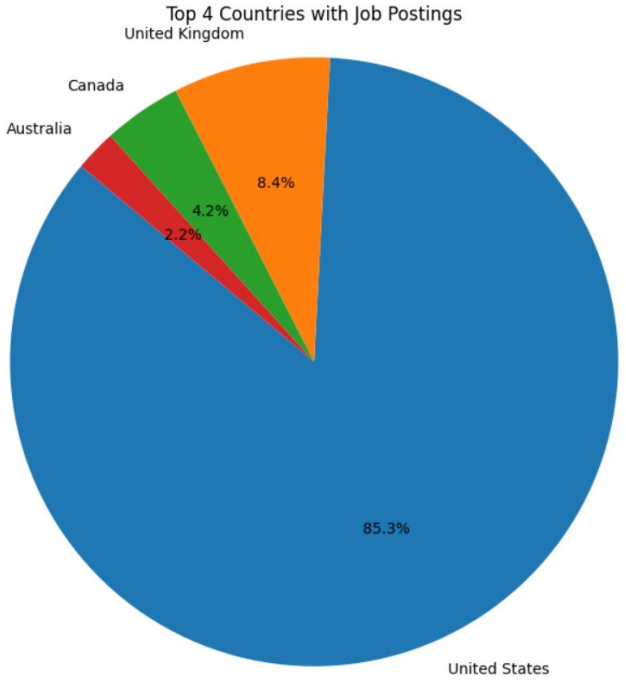
Job Type Distribution: Investigate the distribution of job types (e.g., onsite, remote) to understand the nature of job opportunities available.



NYU

Key Findings & Insights:

Geographical Distribution of Job Postings: Investigate the distribution of jobs across various countries to analyze which is a hit-spot for job availability



NYU

Recommendation System Implementation

Utility Functions:

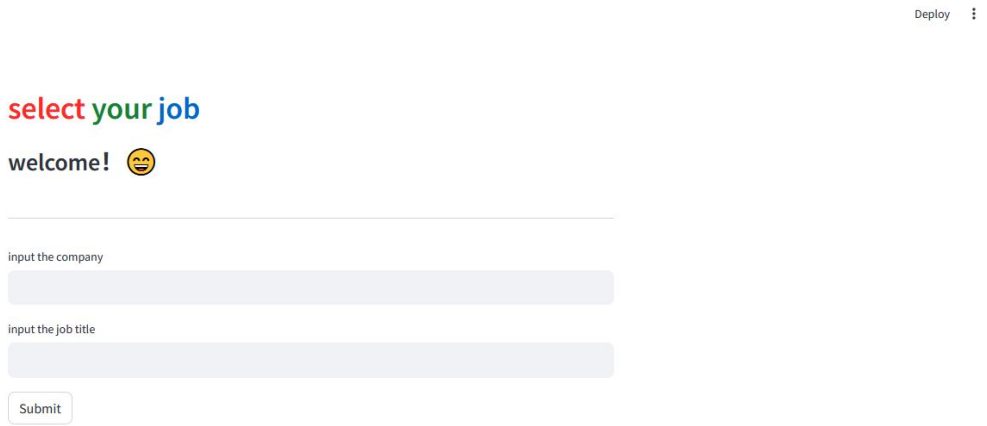
1. **Data Manipulation:** The Utils module offers essential functions for efficient data handling, including loading skill tags from JSON files, ensuring seamless integration of new skills into the system.
2. **Search Functionality**: It provides robust filtering capabilities, allowing users to specify search criteria such as job title, company, skills, and location, ensuring precise and relevant results.
3. **Performance Optimization**: The module incorporates pagination functionality, enabling smooth navigation through search results, enhancing user experience by minimizing load times and optimizing resource utilization.

Recommendation System Implementation

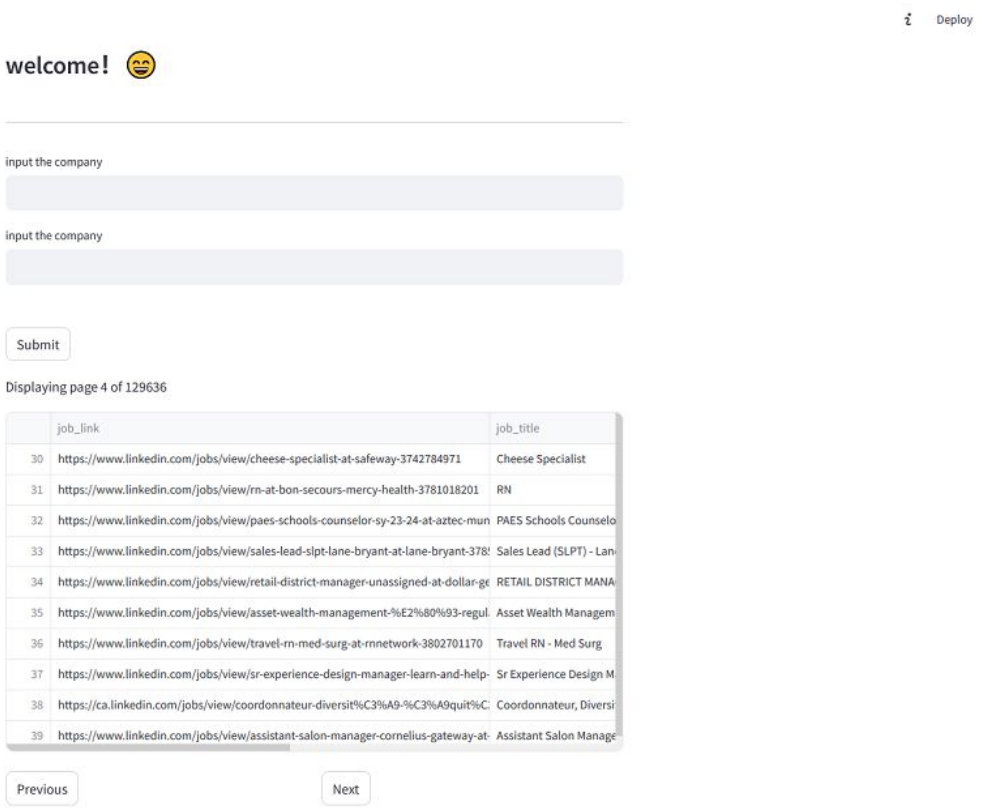
Web Application with Streamlit:

1. **Interactive User Interface**: Leveraging Streamlit, the web application offers an intuitive and user-friendly interface for job search, enhancing accessibility and usability for users of all skill levels.
2. **Dynamic Features:** Users can dynamically select job skills from a list, providing flexibility and customization options tailored to individual preferences and requirements.
3. **Customization Options:** The application allows users to customize the sidebar background, offering a personalized browsing experience and enhancing visual appeal.

Recommendation System Implementation



Recommendation System Implementation

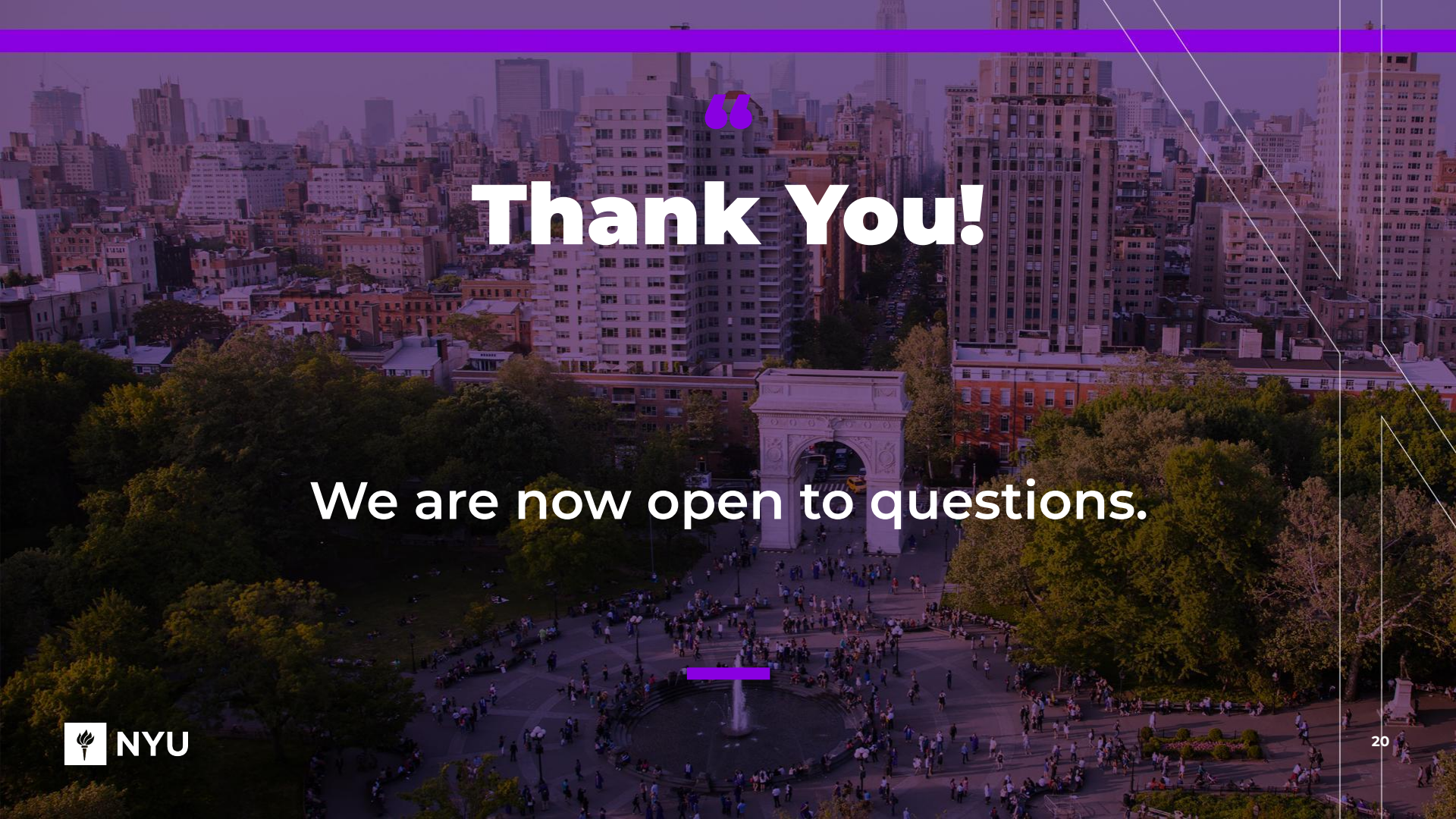


Conclusion

1. **Insights into Job Market Trends:** Our analysis uncovered key trends in job titles, industries, and required skills, providing valuable insights into the evolving job market landscape.
2. **Identification of Skill Gaps:** By identifying gaps between available skills and market demands, we highlighted areas where job seekers can focus their skill development efforts to enhance employability.
3. **Efficient Data Handling:** Leveraging robust data cleaning and transformation techniques, we ensured the accuracy and relevance of our analysis, facilitating informed decision-making.
4. **Empowering Job Seekers:** Through the implementation of a recommendation system and an intuitive web interface, we have provided job seekers with a powerful tool to navigate the job market effectively and receive personalized job recommendations tailored to their skills and preferences.

Lessons Learned

1. **Integration of Technologies:** Each technology plays a crucial role, with PySpark handling large-scale data operations, Python ensuring functional versatility, and Streamlit providing a user-friendly interface
2. **Data-Driven Insights:** Utilizing PySpark to analyze and visualize data has underscored the value of data-driven decision making in real-world applications.
3. **User-Centric Design:** The use of Streamlit to create an interactive web application emphasizes the importance of user-centric design.
4. **Modularity and Scalability:** By structuring the utility functions and data processing separately from the web interface, the project is made more modular, which enhances scalability and maintainability.



Thank You!

We are now open to questions.

NYU