Data607 Project 3: Most Valued Data Science Skills

Group Members

The group members include:

- John Ferrara
- Alinzon Simon
- Akeem Lawrence
- Anthony Roman
- Ben Wolin

Introduction

The aim of this project is to find out what skills in data science are mostly demanded in the labour market that is now booming. Data science has become a highly important field in all sectors, where innovation and decisions are powered by insights drawn from available data. Increasingly, companies look up to data scientists to make sense of vast datasets, develop predictive models, and deliver actionable insights that would inform their business strategies.

With the growth of the data science field, the required skills are also changing; they now range from knowing programming languages like Python and R to machine learning, advanced techniques in the visualisation of data, cloud computing, and big data analytics. Understanding which skills are considered valuable helps aspiring data scientists and professionals currently working in the field orient their development efforts toward market needs.

The project will analyze current job postings to determine the key skills in demand today and how professionals and organizations can remain competitive within this exponentially growing industry. By studying trends across different regions and sectors, we learn how specific skills are valued differently depending on the industry or location.

Collaboration Tools

Communication

As a group, our main communication tools are iMessage and Slack. There may be other means of communication used, but so far these have been our main methods.

Code Sharing & Project Documentation

Other than collaborating on Slack, other tools to be leveraged for code sharing and project documentation are GitHub.

Database Tools

For this project, our data will live in a MySQL database hosted on CloudSQL. The languages used to analyze this data will be R and SQL.

The Data

Data Source

Our group has chosen to work with a Kaggle-sourced dataset that examines job postings on LinkedIn. This data contains information such as the locations of the entities hiring, the companies performing the hiring, the job titles for the open positions, along with additional information related to the position. Additional information, and the dataset itself can be found here. Lastly, the dataset files and their respective column names can be found in Table 1 below.

File Name	Columns
job_postings	job_link, last_processed_time, last_status, got_summary, got_ner, is_being_worked, job_title, company, job_location, first_seen, search_city, search_country, search_position, job_level, job_type
job_skills	job_link, job_skills
job_summary	job_link, job_summary

Table 1: Dataset Files and Columns

Sources

The links to the data sources are from here:

- Kaggle Job Postings
- Kaggle Data Science Skills
- Job Postings

Database Structuring

The proposed normalized tables for structuring the data within the MySQL database can be seen in Figure 1 below. The image also lives here with the acutal file here

Data Loading

Currently, the data loading process can be seen from the following file on GitHub within our shared public repo for this project.

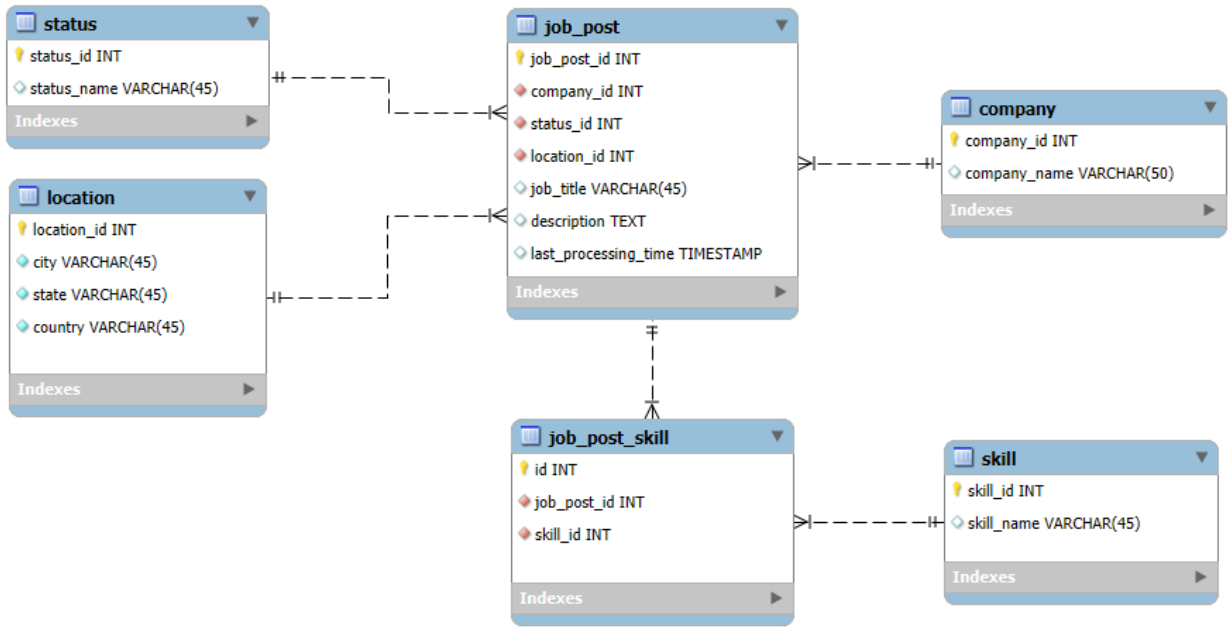

Figure 1: Proposed Database Table Structures