**Professor Tingting Nian**

**MSBA 212 Data and Programming for Analytics**

**25 October 2021**

**Project Proposal:   
Scraping for Desired Skills from JobBoards**

**By: Team 23**

**Kate Zagrebneva   
Prashasti Sharma   
My Han Mai   
Noor Zia   
Samira Shafiei**

**Introduction**

Glancing at hundreds of lengthy job descriptions can be a daunting task for anyone on a job hunt. At the same time, it is essential to stay market relevant and possess a role-specific skill set. In our project, we aim to ease the tension by underlining the key skills desired by employers for various job roles/titles in the data science industry. Job portals (e.g.: indeed.com) lists countless job openings with information on the skill requirements. We intend to scrape one of the portals, sieve relevant skills required corresponding to the popular job titles/roles in top-10 desired work locations.

**Project goal/question**

What skills/combination of skills are most desired in job listings for specific data-focused roles?

**Plan to collect and analyze the data:**

1. **Data collection:**

We plan on using a web scraper, Apify.com, to collect job listings for various cities. We will compile 10 top cities with between 300 to 900 listings per city. The collection will be to extract the position name, company, location, and job description of each listing.

1. **Data Cleaning:**

Data cleaning is one of the most necessary steps of data analysis. If this most basic stage is not done properly, then the analysis presented in the following stages will not be reliable and there will be uncertainty in the accuracy of the resulting conclusions. In order to make sure that the data is ready we will follow these steps:

1. Delete senior level positions, since they are most likely not attainable for recent graduates

2. Fix structural errors

3. Organize data (e.g. making “location” values the same format for all postings)

4. Handle missing data

5. Filter unwanted outliers

6. Compile data into a csv or excel file for better readability and accessibility

1. **Analyze the data**

We will use NLP and n-grams to see which skills are the most popular in the job posting descriptions, using a set of keywords (skills and softwares that we are learning in this program) to figure out what we need to focus on more during this academic year. We will also do a cluster analysis to see the combinations of skills that employers ask the most for (e.g. some employers ask for Python and Tableau, but if they ask for R, they don’t usually include Tableau as a required skill).

Since we have data from 10 most desired locations in the US, we will also see which location has the most amount of entry-level positions suitable for recent graduates from the MSBA program. This may guide us towards relocating to the city after graduation where we would have the biggest chance to land our desired job.

1. **Interpreting the result**

Based on our results from the n-grams and NLP, we will have a list of most common skills that come up within job requirements. These will be the skills that we will deem most relevant in the preparation for finding and securing a data-focused position in a company. The skills will be ranked by the level of frequency in which they appear.