Project 3 Part 1

Table of Contents

Group Members **2**
Collaboration Tools **2**
Data **2**
Data Sources 2
Where Data Can Be Found 2
Loading Data 2
Entity Relationship & Logical Model for Normalized Database **3**
Methods for Analysis **3**
Appendix A: Analysis of Robots.txt and Terms & Conditions **4**
Appendix B: E-R Diagram of Normalized Database **6**

Group Members

- Waheeb Algabri
- Keith Colella
- John Cruz
- Shoshana Farber
- Kayleah Griffen

Collaboration Tools

- Communication
  - Messaging: Slack
  - Meetings: Zoom
- Code Sharing
  - Github
- Project Documentation
  - Google drive

Data

Data Sources

We collected data from sources that we could download data directly from or web scrape from. To determine if we could scrape a webpage, we first learned about robots.txt files, here. Almost all websites have a robots.txt file and it clearly defines what links from the website can and cannot be scraped. To go one step further, we also checked the terms and conditions section. We analyzed multiple job boards robots.txt files and as well as terms and conditions sections (see Appendix A), based on this analysis we were able to select our data sources which are the following:

1) USA jobs
2) AI jobs
3) NYC jobs
4) DataAnalyst.com

Where Data Can Be Found

After data was collected, we posted the csv files to our github to make them easily accessible. The shortcoming of this is that we froze the data at a certain point in time, but we did this so that the results of this project could be replicated based on the specific job listings we accessed. There is also enough data from each csv to accurately assess the most important skills needed for data science.

Loading Data

The data can be loaded into R directly from the csv files on github.

Entity Relationship & Logical Model for Normalized Database

We wanted to take our extracted data and attempt to normalize it into a traditional database, via third-normal form (3NF). However, given the complexity of how data science job titles can vary, and storing the data source (website names), we decided to avoid adding unnecessary complexity on the queries. Job skills and location were top priorities to normalize and extract from the data frames into independent tables.

Creating the normalized database, we used two separate data frames. The first includes an observation for each job posting (i.e. a row with all of the relevant information from the website). In the second data frame we extracted each unique skills from the job description for each listing. There is a one to many relationship between the job postings data to the skills data. The primary key for creating the skills table is a composite key unique identifier of *job_id* and *skills*.

See Appendix B for the E-R diagram: Also available on **Lucidchart**

Methods for Analysis

For the analysis, we decided to focus on a few factors that we thought were important. After loading in the data frames from each job site, we created database tables that had the following:

job_posting

Column Name	**Description**
job_id	A unique identifier for each job listing
job_title	The job title for each listing
min_salary	The minimum salary amount (NA if not listed)
max_salary	The maximum salary amount (NA if not listed)
location_id	Unique identifier for each job location
website	The website the job listing was scraped/downloaded from

job_skills

Column Name	**Description**
job_id	Unique identifier for each job listing
skills	The individually extracted skills from the job description
job_description	A description of the job listing including all the skills required for the job

location

Column Name	**Description**
location_id	Unique identifier for each job location
country	Country the job was posted for

Appendix A: Analysis of Robots.txt and Terms & Conditions

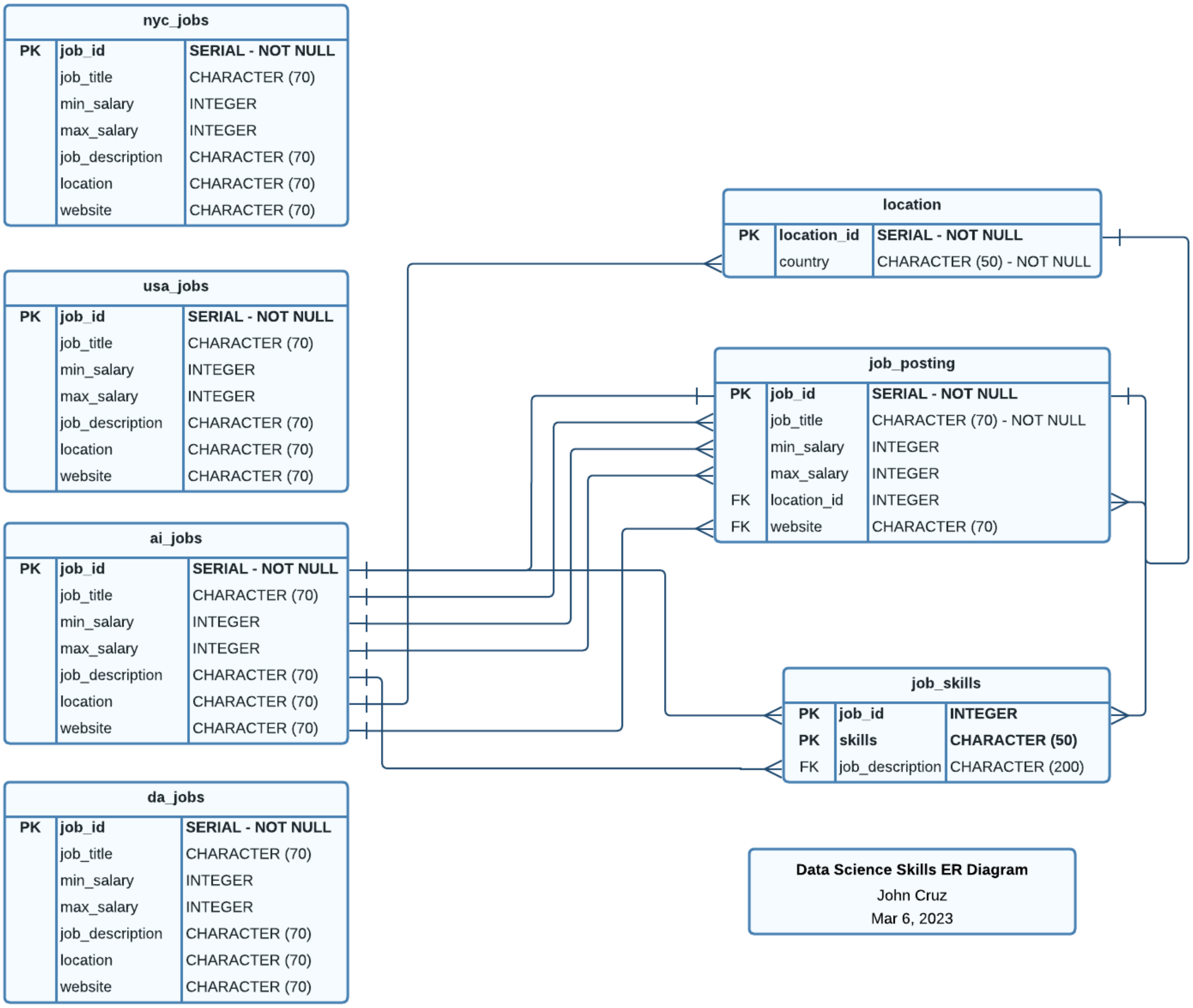
Service	Robots File	Relevant parts of robot.txt	Notes
AI jobs	https://ai-jobs.net/robots.txt	Sitemap: https://ai-jobs.net/sitemap.xml User-agent: * Disallow: /*/apply/ Disallow: /job/mark/* Disallow: /talent/* Disallow: /stats/*	robots.txt - Allows scraping of main page Terms of use - does not reference crawling or scraping
NYC Jobs	https://data.cityofnewyork.us/robots.txt	User-agent: * Crawl-delay: 1	robots.txt - Scraping of jobs page is allowed with a crawl delay of - you can just **export a csv** Terms of use - does not reference crawling or scraping
USA jobs	https://www.usajobs.gov/robots.txt	User-agent: * Disallow: /Content/ Disallow: /Scripts/ Disallow: /foresee/ Disallow: /Service References/	Scraping not explicitly banned in terms of use or in robots.txt
dataanalyst.com	Does not exist	Does not exist	No terms and conditions
Indeed	https://www.indeed.com/robots.txt	Disallow: /jobs/title	Although unclear in the robots.txt file, in the legal section it states "You **may not crawl**, scrape, data mine, extract data from, reproduce, duplicate, copy, sell, exploit, trade or resell any part of the Site, except as expressly permitted by Indeed beforehand, in writing."
Google	https://www.google.com/robots.txt	User-agent: * Disallow: /search	All scraping through search is banned by robots.txt.
Glassdoor	https://www.glassdoor.com/robots.txt	Based on the robots file, scraping jobs is allowed.	In terms of use you will not "Introduce software or automated agents to the services, or access the services so as to produce multiple accounts, generate automated messages, or to scrape, strip, or mine data from the services without our express written permission"

LinkedIn	https://www.linkedin.com/robots.txt	User-agent: * Disallow: /	Explicitly banned in robots.txt file.
Zip Recruiter	https://www.ziprecruiter.com/robots.txt	Based on the robots file, scraping jobs is allowed.	In terms of service “You agree not to engage in … ‘scraping’”
Monster	https://www.monster.com/robots.txt	User-agent: * Disallow: */jobs/search*	In terms of use “All Monster users will not… (d) use any data mining, robots or similar data gathering or extraction methods”
Simply Hired	https://www.simplyhired.com/robots.txt	Search is not explicitly banned in robots.txt.	In terms of service “you agree that you will not crawl, scrape…”
CareerBuilder	https://www.careerbuilder.com/robots.txt	Search is not explicitly banned in robots.txt.	In terms “prohibition includes: ...using or attempting to use… scripts, robots or other means, devices, …to navigate, search, access, ‘scrape,’ … any web pages or any Services provided on the Sites other than the search engine and search agents available from CareerBuilder on such CareerBuilder Sites and other than generally available third party web browsers…(c) aggregating, copying or duplicating in any manner any of the Content or information available”
Snag	https://www.snagajob.com/robots.txt	Disallow: /search?q=* Disallow: /search/?q=*	In terms “you will not use any manual or automated software, devices or other processes (including but not limited to spiders, robots, scrapers, crawlers, avatars, data mining tools or the like) ”
LinkUp	https://www.linkup.com/robots.txt	Search is not explicitly banned in robots.txt.	In terms “you shall not:... ‘scrape’”
Job.com	No robots file.	No robots file.	Terms and conditions - “Except to the extent expressly set out in these Terms, you are not allowed to:... scrape”

Appendix B: E-R Diagram of Normalized Database

Lucidchart

Note: The diagram does not fully represent the connections between our four data source tables and our analytical tables of job_posting, location, and job_skills. We used the “ai_jobs” table to describe these connections to avoid confusion about all the dependencies. However, the links are identical because the data source tables were formed similarly.



nyc_jobs

PK	**job_id**	**SERIAL - NOT NULL**
	job_title	CHARACTER (70)
	min_salary	INTEGER
	max_salary	INTEGER
	job_description	CHARACTER (70)
	location	CHARACTER (70)
	website	CHARACTER (70)

usa_jobs

PK	**job_id**	**SERIAL - NOT NULL**
	job_title	CHARACTER (70)
	min_salary	INTEGER
	max_salary	INTEGER
	job_description	CHARACTER (70)
	location	CHARACTER (70)
	website	CHARACTER (70)

ai_jobs

PK	**job_id**	**SERIAL - NOT NULL**
	job_title	CHARACTER (70)
	min_salary	INTEGER
	max_salary	INTEGER
	job_description	CHARACTER (70)
	location	CHARACTER (70)
	website	CHARACTER (70)

da_jobs

PK	**job_id**	**SERIAL - NOT NULL**
	job_title	CHARACTER (70)
	min_salary	INTEGER
	max_salary	INTEGER
	job_description	CHARACTER (70)
	location	CHARACTER (70)
	website	CHARACTER (70)

location

PK	**location_id**	**SERIAL - NOT NULL**
	country	CHARACTER (50) - NOT NULL

job_posting

PK	**job_id**	**SERIAL - NOT NULL**
	job_title	CHARACTER (70) - NOT NULL
	min_salary	INTEGER
	max_salary	INTEGER
FK	location_id	INTEGER
FK	website	CHARACTER (70)

job_skills

PK	**job_id**	**INTEGER**
PK	**skills**	**CHARACTER (50)**
FK	job_description	CHARACTER (200)

Data Science Skills ER Diagram
John Cruz
Mar 6, 2023