README

1. Project Overview

   CareerMagnet is a product that provides comprehensive information about the job posting and link, salary, property/income ratio and H1B sponsorship. The information is collected from indeed.com (job posting), salary.com (salary), h1bgrader.com (immigration stats), and numbeo.com (living costs), which provides datasets covering the job description, salary level in a specific city, sponsorship information and living cost in the city.

   It allows users to obtain latest information on the job market through live web scraping, or take use of existing data that was pre-downloaded on Oct 10, 2023.

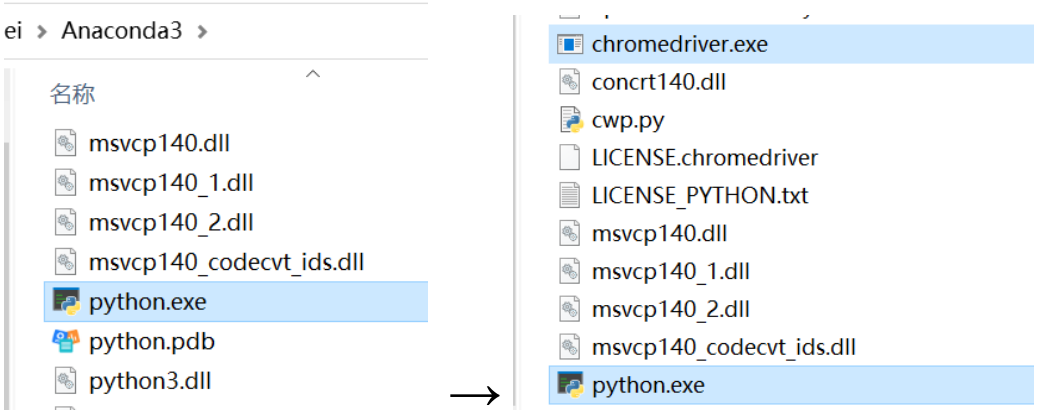
   Through menu choices, users may apply filters of selecting positions with H1B sponsorship, positions with remote option, etc. The top recommendations are provided based on minimum salary.

2. Directory Structure
   a. Main directory
      i. “main.py” is the main code file used to execute the application
      ii. “merged_all.py” is the file that merges the data from each website
      iii. “statistical_results.py” is the file that does graphical representation for the salary information
   b. “data” folder
      i. Contains all the web scraped and cleaned/merged data that has been converted to CSV
   c. “scrape_clean” folder
      i. “H1B_cleaned.py” contains the script used to scrape and clean H1B sponsorship data from dol.gov
      ii. “Indeed_scrape_and_clean.py” contains the script used to scrape and format indeed job postings
      iii. “salary.py” contains the script to scrape salary.com for the salary percentiles by city in the United States
      iv. “Property_prices_scrape_and_clean.py” contains the script used to scrape and clean property price information from numbeo.com
3. Installation
   a. Download Project folder to directory of your choice on local device
   b. Selenium Setup
      i. Type and run “pip install selenium” to install selenium
      ii. If the Chrome version is 117, win-64, you may directly use this link to download chromedriver-win64: https://edgedl.me.gvt1.com/edgedl/chrome/chrome-for-testing/117.0.5938.149/win64/chromedriver-win64.zip

iii. If the Chrome version is 117, mac-arm64, you may directly use this link to download chromedriver: https://edgedl.me.gvt1.com/edgedl/chrome/chrome-for-testing/117.0.5938.149/mac-arm64/chrome-mac-arm64.zip

iv. If using another version of chrome, visit the following link: chromedriver.storage.googleapis.com/index.html

v. Download, unzip, and place the chromedriver.exe file in the same folder as your python.exe on your local device (you may probably find it under the "Anaconda3" folder). Screenshots below:



vi. If you would like to run and test that selenium works properly, copy and run the following code in an IDE of your choosing:

```
from selenium import webdriver
driver = webdriver.Chrome()
driver.get("https://www.google.com/")
```

vii. Make sure to have Chrome browser installed on your device. The program uses Selenium to access Indeed.com through Chrome

4. Modules Import
   a. Primitive modules provided by Anaconda

   All the import statements are already included in the codes:

```
import numpy as np
import pandas as pd
import sys
import difflib
import warnings
import os
import requests
from bs4 import BeautifulSoup
import time
```

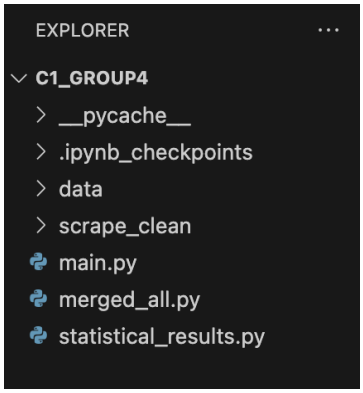
import matplotlib.pyplot as plt
import csv
from time import sleep
from random import randint
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.common.exceptions import NoSuchElementException
import re # Used for regular expression

b. Other modules

Here are the modules that are provided by us as .py files. All the import statements are already included in the codes:

import merged_all as ma
import statistical_results as sr
import scrape_clean.indeed_scrape_and_clean as isc
import scrape_clean.H1B_cleaned as h1b
import scrape_clean.property_prices_scrape_and_clean as prop
import scrape_clean.salary as sal

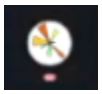
5. User Guide
   - Download zip folder to your machine and unzip to any location
   - Open the folder in your IDE of choice
   - Locate the main.py file (in the main directory) and open it



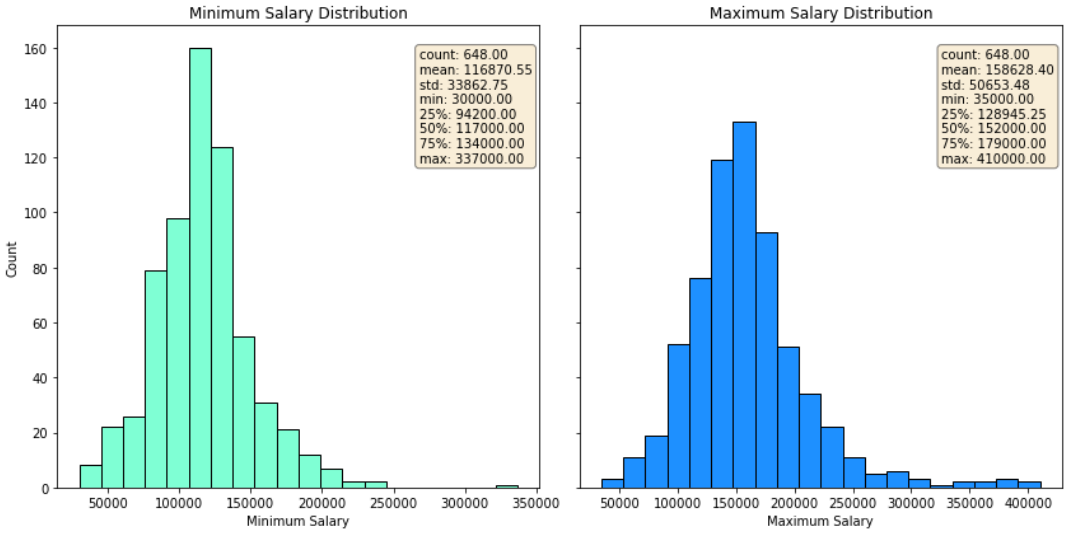
   - Click run in your IDE to start the “main.py” file
   - Choose whether you would like to use cached data (fastest option) or get live data (may take up to 3 hours to complete)

```
Welcome to CareerMagnet, a job finder for Data Science-related professions.
Our application gathers data from Indeed.com, numbeo.com, salary.com, & dol.gov
to provide a more complete job hunting experience for positions in Data Science.

NOTE: If you choose to download live data,
the processing could take around 2-3 hours, depending on the jobs available in the
location that you select. Doing a download of live data will also open and close
your browser. Additionally, The H1B sponsorship data is for last year
so information may have changed.

Would you like to download fresh job data or use the existing data? Select by typing 1, 2, or 3 below:

1. Download
2. Use existing data
3. Quit
> 1
```

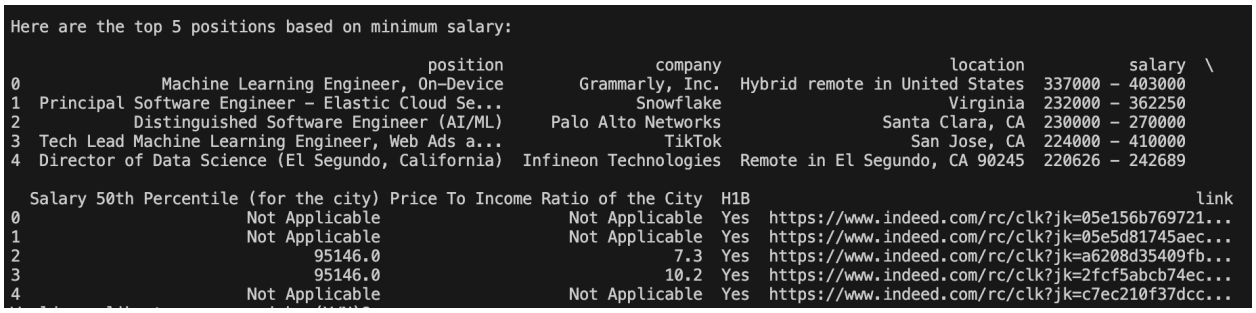
- If you chose to do get live data, your computer will open Chrome to get data from Indeed. Please DO NOT close the browser as this will interrupt the program
- Follow the on-screen prompts to select a filter that you would like to apply or show statistics of salaries
- The statistics will open a graph in a new window. Depending on your settings and Operating system, you may need to click the icon on your taskbar (pictured below)



- This will display the graph, which will look something like the one below (varies based on the data chosen for donwload)



- The program will return the 5 highest paying jobs, but will prompt if you would like to see more (type Y to show more results, N to end the output)



```
Here are the top 5 positions based on minimum salary:

                                      position              company                         location          salary  \
0              Machine Learning Engineer, On-Device       Grammarly, Inc.  Hybrid remote in United States  337000 - 403000
1  Principal Software Engineer - Elastic Cloud Se...           Snowflake                         Virginia  232000 - 362250
2           Distinguished Software Engineer (AI/ML)     Palo Alto Networks                Santa Clara, CA  230000 - 270000
3  Tech Lead Machine Learning Engineer, Web Ads a...              TikTok                   San Jose, CA  224000 - 410000
4  Director of Data Science (El Segundo, California)  Infineon Technologies  Remote in El Segundo, CA 90245  220626 - 242689

  Salary 50th Percentile (for the city) Price To Income Ratio of the City  H1B                                                link
0                        Not Applicable                   Not Applicable  Yes  https://www.indeed.com/rc/clk?jk=05e156b769721...
1                        Not Applicable                   Not Applicable  Yes  https://www.indeed.com/rc/clk?jk=05e5d81745aec...
2                               95146.0                              7.3  Yes  https://www.indeed.com/rc/clk?jk=a6208d35409fb...
3                               95146.0                             10.2  Yes  https://www.indeed.com/rc/clk?jk=2fcf5abcb74ec...
4                        Not Applicable                   Not Applicable  Yes  https://www.indeed.com/rc/clk?jk=c7ec210f37dcc...
```

- Re-run the program to try additional inquiries for jobs. The cached data we provided will remain, so you can try each of the filters

6. Demo Video
    a. You can find a video demonstrating how to use the application on Youtube, click here

7. Project Contributors

    Matt McMonagle (andrewID: mmcmonag)

    Yongbo Li (andrewID: yongbol)

    Paris Chen (andrewID: danyang2)

    Yufei Lei (andrewID: yufeilei)