Alexander Brown, Ruijie Rao, Ahmed Alsalim
DSCI551 Spring 2022
Project Proposal

Labor Statistics Dashboard for Data-Informed Career Path Decisions

Motivation

Right now, a record number of people are leaving their jobs and changing careers due to disruptions caused by the COVID-19 pandemic. According to Congressional Research Service, the unemployment rate in the US reached 14.8%—the highest rate observed since data collection began in 1948—in April 2020. Moreover, according to the latest CNBC jobs report, the number of people employed part-time for economic reasons, but would rather be working full-time, increased to 6.7 million in October after declining for five months. Meanwhile, choosing a new career is one of the most impactful decisions a person can make. It often takes years of careful planning and enormous costs. Unfortunately, many people often make very important decisions about which career they want to pursue or how to prepare for such a career, without a thorough knowledge of the current or future labor market. Researchers found out some factors on CDD (Career Decision Difficulty), such as employment status, lack of accurate information about careers, gender, and more. (Abdulfattah and Nizar). Much of the information for a given field, such as average salary, unemployment rate, worker hazards, turnover rate, and educational costs, are often hidden behind convoluted government or corporate datahubs and APIs.

For our project, we'd like to help ease this problem by creating a user-friendly dashboard to explore important labor statistics based on career interests, educational level, and geographic regions. We will start by first providing visualization and access to the U.S. Governments Bureau of Labor Statistics datasets, such as the [Occupational Employment and Wage Statistics](#), [Current Population Survey](#), [Modeled Wage Estimates](#), [Occupational Requirements Survey](#), and [Employment Projections](#). Time permitting we will also add additional datasets regarding career recommendations for the job posting website dice.com. We will focus on monthly data at the state level for 2019-2020 to also reflect the disruptions caused by the COVID-19 pandemic. More broadly, we feel that providing this information on a large scale could help ease some of the disruption caused by the pandemic and our changing labor markets. It could help address over and under-employment. We also feel that on a personal level, having access to important data about the labor market, could be hugely beneficial and stress relieving when trying to navigate a new career or make a career change.

Plan

The plan breaks down roughly into three phases. The first phase is the data acquisition, preparation, and database configuration. Phase 2 consists of the development of the local web app, user interface, and data visualization. The final stage will be the testing and online deployment of the web app. As with any project in the early stages, we are not 100% certain on every step and the exact tools we will need but the rough outline is as follows:

Phase 1: Data Acquisition/Preparation and Database Configuration
- Acquire static BLS data for 2019 and 2020
  - Monthly
  - State-level
  - Static datasets (xls)
  - Download from the website or use REST API
  - Put it into the Google Drive
  - Configure Google Colab Notebook for data prep
- Data preparation
  - Cleaning (Google Colab/Pandas/Excel)
    - Missing value
    - Remove duplicates
    - Make feature data is in the right format(data type)
    - Remove unnecessary features
    - Text reduction
  - Joining time-series data for 2019-2020 (potentially more)
- Database management
  - Load cleaned data into database
    - Cloud relational database or NoSQL database such as Firebase
  - Build queries scripts for retrieving data

Phase 2: Data visualization and Web app development
- Data visualization
  - Find tool for this (Tableu/Power BI/statsmodel/Python)
  - Develop visualizations
    - Which data to show?
    - Interactive or not?
    - How does a user select different visualizations?
    - Linking data to visualizations
- Webapp Development
  - Select tools
    - Python Backend, Flask, and Django framework
  - Backend routing
  - Frontend designing

Phase 3: Web app testing and deployment
- Webapp Development
  - Backend routing
  - Frontend designing
  - Deployment with AWS
  - User testing
- Demonstration and report

Data Sources

- US Bureau of Labor Statistics: https://www.bls.gov/oes/tables.htm
  - This government website includes a variety of labor market data including salary and employment rate for most positions, entry-level education requirements, retention rates, and other attributes. Moreover, the data is recorded annually from 1988 to 2021, which allows us to do some pattern analysis with huge time bounds. Since it is a government primary dataset, we believe that the data will be cleaner and more trustworthy than other public resources. As stated previously, the data is broken down into separate datasets based on the topic. Examples include Occupational Employment and Wage Statistics, Current Population Survey, Modeled Wage Estimates, Occupational Requirements Survey, and Employment Projections.
- Dice Job US: https://www.kaggle.com/rayjohnsoncomedy/job-skills
  - This is a public dataset crawled from Dice.com. It is a secondary dataset with a lot of unclean data pieces. Much of the information is gathered from different companies and the job titles are not in the same format. However, it provides a wide range of tech jobs as well as their required skills and job descriptions, which is absent in the former data source. Combining this dataset, after it is cleaned and filtered, with the BLS dataset above will help us recognize the pattern between required skills and raised salary/employment rate.
- We are still researching other useful data sources to be used if time permits.

Milestones/Timeline

We expect to be done with phase 1 and phase 2 by the week of March 28th and then phase 3 by the end of the semester. We further broke down the phases into weekly tasks. We don't have a weekly breakdown of tasks after the second phase because we are not sure exactly what will need to be done.

Phase 1: Weeks 5-8
Phase 2: Weeks 8-12
Phase 3: Weeks 12-16

Week 5 (2/7)	- Decide which data sets we want (Alexander) - Aquire data (Alexander) - Move data into Google Drive (Alexander) - Configure our data preparation Google Colab Notebook - Study for midterm (Alexander, Ruijie, Ahmed)
Week 6 (2/14)	- Clean data (Ahmed, Alexander) - Come up with a rough outline of user input/website flow (Ruijie, Alexander) - Join datasets (Ahmed)

Week 7 (2/21)	● Start web app frontend mock design (Ruijie) ● Load data into database (Ahmed, Alexander) ● Start building query scripts (Ahmed, Alexander, Ruijie)
Week 8 (2/28)	● Data visualization (Ruijie, Alexander) ● Continue building query scripts (Ahmed, Alexander)
Week 9 (3/7)	● Continue with frontend and backend development
Week 10 (3/14)	● Continue with frontend and backend development
Week 11 (3/21)	● Continue with frontend and backend development ● Start to plan any expansions features
Week 12 (3/28)	● Write progress report

Team Members and Responsibilities

See above

Ruijie Rao (front end, web deployment)

Alexander Brown (data filter/visualization)

Ahmed Alsalim (aquire data/data cleaning/database setup)

References

Falk, Gene., Romero, Paul D., Nicchitta, Isaac A., Nyhof, Emma C. (2021). "Unemployment Rates During the COVID-19 Pandemic." *Congressional Research Service.* R46554.
https://sgp.fas.org/crs/misc/R46554.pdf

Lui, Jennefer. (2020). "Millions of Workers Remain Underemployed as Visur Surges Prompt Closures." *CNBC.*
https://www.cnbc.com/2020/11/19/millions-of-workers-underemployed-as-virus-surges-prompt-closures.html

Yaghi, Abdulfattah., Alabed, Nizar. (2021)."Career decision-making difficulties among university students: does employment status matter?" *Higher Education, Skills and Work-Based Learning*
https://www.emerald.com/insight/content/doi/10.1108/HESWBL-07-2020-0149/full/html