**HW2 — Project Progress Report**

**Hypothesis and Problem Statement**

My hypothesis is that **most jobs in the dataset require a college degree or higher education**. To validate this, I analyzed job title frequency and filtered specific columns (Job Title, Date of Birth, and Sex) to examine career types, age distributions, and gender patterns. This was initially introduced in HW1, and this report builds on it with additional data extraction.

**Repository Updates**

I continued working in my GitHub repository:  
[GitHub Repo – nnguyen79pine](https://github.com/nnguyen79pine/git-repo)

New files have been added, including:

* Java source files and compiled .class files for Hadoop jobs
* Manifest and executable .jar files
* Output datasets like job\_sex\_dob.csv and job\_title\_counts.txt

All work is committed and pushed regularly to maintain backups and track progress.

**Column Selection with head**

To view the dataset structure and choose relevant columns, I logged into the Hadoop VM and used:

A screenshot of a computer

AI-generated content may be incorrect.head -n 1 people-100000.csv

To support my hypothesis that most jobs in this dataset require formal education or a college degree, I extracted the following columns from the dataset:

* **Job Title**
* **Date of Birth**
* **Sex**

I used these columns to explore patterns such as the types of careers held by different individuals, age trends across job categories, and any notable gender distribution. The extracted file job\_sex\_dob.csv is included in the repository under the /data/ directory.

This file was generated to give a broader perspective on how personal attributes might correlate with professional roles, which extends the initial job title count analysis.

**Hadoop Extraction and Part File**

Using Hadoop and my custom JobTitleMapper, JobTitleReducer, and JobTitleDriver, I processed the full dataset to count the frequency of each job title. The job was executed with the following command:

bin/hadoop jar JobTitleCount.jar /inputjobs /outputjobs

The output was stored as part-00000, then copied to the local file system and renamed to job\_title\_counts.txt. I added a header and formatted it into a proper CSV file named job\_title\_counts.csv. Both files are available in the repository under the /data/ folder for further analysis and documentation.

A black screen with white text

AI-generated content may be incorrect.

Using Hadoop and my custom JobSexDOBMapper, I processed the dataset to output only the selected columns. The job was executed with the following:

bin/hadoop jar JobSexDOB.jar /jobsex\_input /jobsex\_output

The output was saved as job\_sex\_dob.txt and then converted to job\_sex\_dob.csv by adding a header and cleaning formatting issues. This file is now available in the repository under the /data/ folder.

A black screen with white text

AI-generated content may be incorrect.

**Updated Conclusions**

After further processing and reviewing the filtered data:

* The **most frequent job titles** are roles like **Engineer**, **Therapist**, **Psychologist**, **Clinical Scientist**, and **Teacher**
* These roles generally **require a college degree or formal training**, further supporting my original hypothesis
* The added columns helped me notice that many technical or scientific professions are concentrated in the 25–45 age range, with gender skewing slightly depending on the field

Overall, these findings reinforce my original conclusion that most job opportunities in this dataset require higher education, and filtering the columns helped visualize it more clearly.