Term Project: Employment Projections (Pre-Post Pandemic)

The Bureau of Labor Statistics publishes 10-year projections about the US labor market: between 2008 and 2018, they were published every second year and since, 2019, they are being published every year.

Your goal in the project is to analyze the 2019-2029 projections data (pre-pandemic) and the post-pandemic projections for the years 2023-2033. The project is somewhat open-ended: feel free to explore interesting questions about the data, and especially try to understand how the forecast has changed in the intervening years. Please avoid the temptation of just framing statistical questions similar to those on the BLS website: e.g., what the fastest growing occupations are forecast to be etc. I would like to see more thought being given to thinking about trends from the point of view of educational attainment, skills requirements (data that BLS has started publishing from this latest projection), geographic distribution of employment categories, and so on.

Dataset

The provided data, once unzipped, is available to you in four directories:

- `2019-2029`: Employment Projections from 2019-2029
- `2023-33`: Employment Projections from 2023-33
- `oesm19nat`: Occupational Employment Statistics (2019)
- `oesm23nat`: Occupational Employment Statistics (2023)

To keep things in perspective, note that $100 in 2019 would have the same buying power as about $119.35 in 2023.

The `2023-33` directory contains a new file, `skills.xlsx`, that provides scores for specific skills like adaptability, speaking and listening etc. depending on how they apply to various jobs. If you want to be ambitious, you can also scrape or download additional data if you wish: for example, census data or climate data that could be used to predict geographic shifts in both population and employment opportunities across the country.

The actual datafiles are Excel spreadsheets with multiple sheets per file, so you will need to make sure that you install a suitable Excel reader (`openpyxl` is recommended) that can read the data into appropriate `DataFrame` objects for further analysis. The `2019-29` data has an additional sub-directory with crosswalks (e.g. relating occupations with their codes etc.) that you may or may not need in your analysis.

Code Organization

Treat this as a project that you would like to showcase: it should be well-organized with complete documentation of the codebase (using**Sphinx**), a suite of tests as needed for functions that you may develop, and version-controlled on Github.

You must practice modular organization of your codebase, and please make sure that you follow either Google style or numpy style in your documentation. Linting your code is always a good idea especially as is helps you avoid ugly coding practices.

For analysis and models, try to stick with the pure Python ecosystem (`pandas`, `sklearn` etc.) but if you wish to try to learn to use more powerful frameworks like `pyspark`, you could do that as well on the Google Cloud Platform (GCP) or on the Rutgers **Amarel** cluster. Choose whatever mining and ML techniques/models that you deem appropriate for the analysis, but there should be no need to use neural networks or LLMs for this project.

For those of you that want to use the project to learn some cloud skills, I am distributing certificates that can be used to deploy compute and storage resources on GCP.

Teamwork

Each group will be made up of 3-4 people: I have tried to ensure that each group contains roughly the same mix of undergraduate and graduate students. I will email one person in each group with the email addresses

of the others in the group so that you can start discussing the project among yourself.

As in any group project, please make an effort to contribute enthusiastically, while being inclusive and respectful of opinions: make sure that all voices are heard. The expectation is that the project responsibilities will be equitably divided.

Groups

1. Chowdhury, Dandu, Le, Pandya, Philip
2. Birla, Elango, Jackson-Connor, Makhijani, Naik
3. Ankola, Chen, Jariwala, Nazmee, Puthiya Kottal,
4. Champaneria, Granado, Lisa, Nikiforov, Singh,
5. Mathew, Koehn, Liona, Shetty, Siddhabathula
6. Garg, Kern, Kithani, Narayanan, Patwardhan

Timeline

I would like to see the project completed in 3 stages:

1. Brainstorm and get ideas for questions you want to ask about the data. By **Nov. 7**, each group must send me a short presentation video (5 minutes) about what they plan to accomplish. You should also share with me a Google Drive link (on ScarletApps) or a github site where you will be maintaining the code.
2. By **Nov 21** (before the Thanksgiving break), each group should demonstrate a prototype of the project's codebase. If you plan to create some sort of interactive dashboard, you should have a prototype of that ready as well. The demonstration will be done outside class, either during office hours in person or via a Zoom call at a pre-arranged time.
3. The final presentations for the project will be on **December 11**, the last day of classes. Each group will get 10 minutes to present followed by 5 minutes of questions.

To complete the project, you will be asked to submit a short 5-6 page report on the project. The grade will depend on the ideas, code development, presentations and the documentation for the code. Each group will also have to submit a short statement (signed by everyone) that indicates who did what on the project.

Good luck, and I hope you have fun playing with the data!