final project step1

Sarah Qiang

2022-05-15

Introduction:
For foreigners to legally work in the US, we have to obtain a work visa, it is through a lottery system and there's approval process based on your education background and experience level once you are selected in the lottery. Once approved for work visa, the data will be available online at
https://www.flcdatacenter.com/download.aspx.

The data includes the job title, job location (very detailed to specific county) and wage levels (there are 4 wage levels, foreigners must meet level 1 wage requirement to obtain the work visa). The data provides a rough guideline for job hunters like me to get an idea of what are the acceptable salaries in a specific region for a specific job title.

As I am pursuing this master in data science degree, I want to see where the most data science jobs are offered and the pay range for the data science jobs. I want to compare my options if I want to look for a business analyst role or a data analyst role. This is a data science problem because I would need to analyze the data to find the answers. Data science jobs can have different job titles, and the data set is massive. Using R to analyze the problem will make the process more efficent.

Research questions:
1. What is the average pay for data anlyst role in Ohio?
2. Do most data analyst role require a master degree? Definition of most here is more than 75%. Master degree is coded under the DOL training column in the ONet_Occs.csv file.
3. Which state/ which county has the highest paid data analyst position?
4. Which state/ which county has the most petition filed for data analyst roles?
5. Which title earns more money? Business analyst or data analyst?

Approach:
First, I would need to consolidate and clean the data. I was able to obtain 4 cvs file, each contains different information. (See below for more detailed explanations). I would need to rename the columns in a way that's easy to understand.
Secondly, I would install all the necessary packages for the analysis. Thirdly, I would begin my analysis by filtering data, as the datasets are quite large, I believe filtering the data would be helpful in tackling each of my research questions.
Last but not least, I want to input a countermeasure to make sure what I gathered makes sense and is realistic.

To-dos:
1. Understand the data sets, see what variables are captured
2. Make necessary changes, add notes for explanation. For example, the education information, the file uses specific codes to represent each level, 4 means bachelor's degree and 5 means master's degree, so when making graphs/tables, certain translation is needed.
3. Perform the analysis
4. Search open roles on Linkedin/ Indeed (job searching website) to verify if the results from the analysis are valid.

Data:
The file geography.csv file has the BLS geography areas and codes.
The file ONet_Occs.csv contains O*Net 24.3 occupational information, including the Job Zone and the DOL Education and Training Code.*
*The file XWalk_Plus.csv contains records that match the O*Net 24.3 occupations to the corresponding OES/SOC occupations.
The oes_soc_occs.csv file has codes and definitions for the Standard Occupational Classification System as modified by BLS for use in the OES program.
The ALC_Export File includes data from all industries and has the most prevailing wage determinations.
These data are all downloadable from https://www.flcdatacenter.com/download.aspx.
The data sets are available in the Microsoft Excel (.xlsx) file format and organized by the federal fiscal year (October 1 through September 30). Because the data is published from an legitimate and authoritive source, the chances of having missing data and wrong information is very low.

Required packages:
dplyr (to work with data frames)
readxl (data import, to read excel file)
ggplot2 (data visualization, histograms)
plotly(data visualization)
purr (apply functions to the data frame)
rmarkdown (for data display purpose)

Plots and Table needs:
Histograms (easy to see the density of the data)
Scatter plots (easy to see the density of the data)
Box plots (easy to see the min, max, medium)

Questions for future steps:
I want to find the business/ data analyst roles that require the specific skillset I have and find out the salary information. The file doesn't have detailed job description, it only has some brief descriptions. So I might need to learn how to search specific keyword and filter the data that way. It is similiar to doing a vlookup in excel but I need to learn to to perform that function in R.