Analysis of US Job Advertisement Data

1. Dataset

The given dataset is an XML file, and contains details about various job postings in the US.
The XML tree structure is shown in Figure 1.

Figure 1 : XML Tree Structure

There are a total of 5,38,385 data points (job postings) in the given dataset, with each data point having 57 features (details about the job).

2. Data Processing

The XML file has been parsed using the 'Element Tree' library, and the required features (as mentioned in the assignment instructions document) have been stored separately as pickle files.

The SOC codes and Sector Codes are not available in the dataset readily. The required digits have been extracted from the available features i.e., SOC code has been taken from the first 6 digits of 'ConsolidatedONET' and the Sector code has been taken from the first two digits of 'ConsolidatedInferredNAICS'.

Further, for the 'SOC Names' have been downloaded from the website of US Bureau of Labor Statistics (link) and 'Sector Names' have been downloaded from the NAICS Association website (link). The csv files containing this data have been included in the zip folder submitted.

The data has been merged, to get the final dataset, with 5,38,385 rows and 10 columns. A snapshot of the final dataset is shown in figure 2 below.

	JobID	CleanJobTitle	JobDate	CanonEmployer	CanonCity	CanonCountry	CanonState	MSA	Sector	SectorName	SOC	SOCName
0	37983417185	Mental Health Specialist - Center	2015-12-03	Cape Fear Valley Health System	Fayetteville	USA	NC	22180: Metropolitan Statistical Area	62.0	Health Care and Social Assistance	211014	Mental Health Counselors
1	37983417189	Software Development Engineer - Mobile Excelle...	2015-12-03	Amazon	Seattle	USA	WA	42660: Metropolitan Statistical Area|500: Comb...	45.0	Retail Trade	151132	NaN
2	37983417190	Studio Engineer	2015-12-03	Ciber Incorporated	Warren	USA	MI	19820: Metropolitan Statistical Area|220: Comb...	54.0	Professional, Scientific, and Technical Services	172141	Mechanical Engineers
3	37983417197	It Administrator	2015-12-03	Texas Association Counties	Lubbock	USA	TX	31180: Metropolitan Statistical Area	NaN	NaN	151142	NaN

Figure 2 : Dataframe extracted from the XML data

This dataset has been stored as a CSV file, with filename - 'MasterJobData.csv' (present in the zip folder submitted).

Note : The final dataset has few null values, and has to be further processed as per the task at hand.

3. Heat Map of Job postings

A heat map has been created using the plotly choropleth maps functionality. The map in figure 3 signifies the number of job postings, for each state in the US, as per the given dataset.

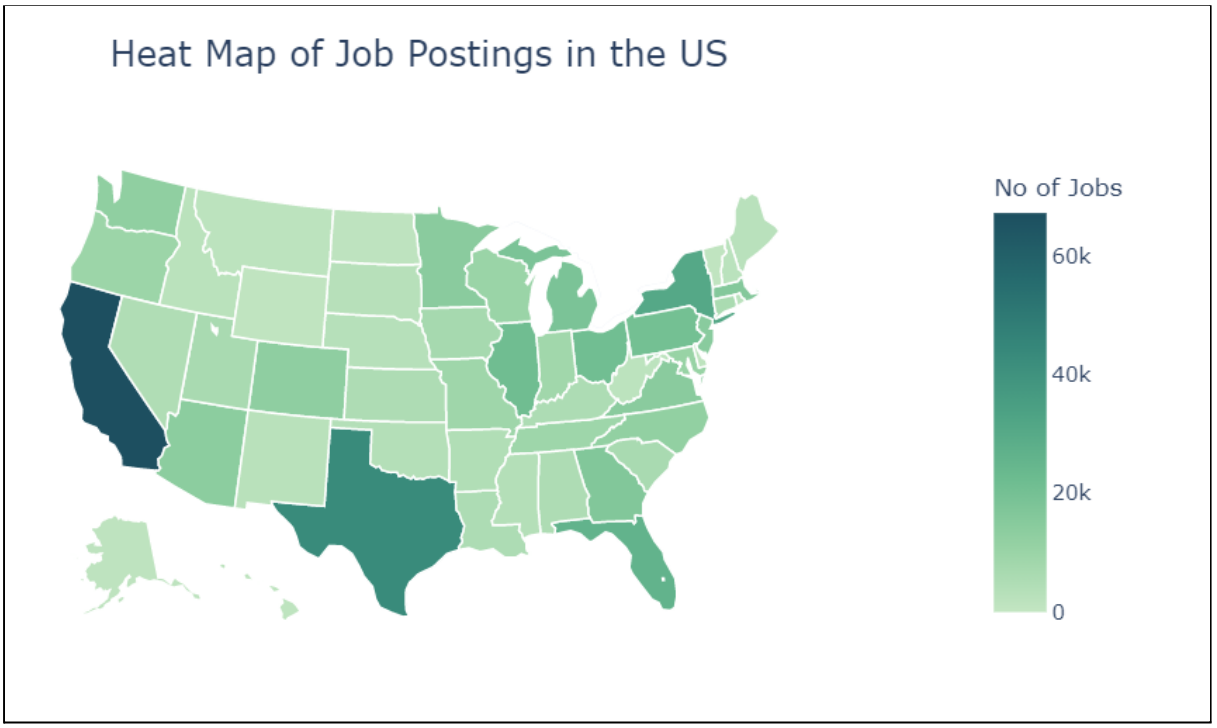

Figure 3 : Heat Map of number of job postings in the US (statewise)