Python Project – 1: Analyzing Naming Trends using Python

Project: Analyzing the naming trends using Python

Industry: General

Problem Statement:

The dataset is in Zipped format, we have to extract the dataset in the program, visualize the number of male and female babies born in a particular year, and find out popular baby names.

Description: This project not only focusses on implementing data manipulation and data visualization using Pandas library, but also tests your ability to deal with real word problem statements.

Dataset: Popular baby names data provided by Social Security Administration (SSA) of United States

How to download the dataset:

- Go to https://www.ssa.gov/oact/babynames/limits.html
- Click on 'National data'
- Get the zipped file

Here's what the zipped folder looks like,

NationalReadMe	PDF File	225 KB	No
yob1880	Text Document	9 KB	No
yob1881	Text Document	8 KB	No
yob1882	Text Document	9 KB	No
yob1883	Text Document	9 KB	No
yob1884	Text Document	10 KB	No
yob1885	Text Document	10 KB	No
yob1886	Text Document	10 KB	No
yob1887	Text Document	10 KB	No
yob1888	Text Document	11 KB	No
yob1889	Text Document	11 KB	No
yob1890	Text Document	11 KB	No
yob1891	Text Document	11 KB	No
yob1892	Text Document	12 KB	No
yob1893	Text Document	12 KB	No
yob1894	Text Document	12 KB	No
yob1895	Text Document	13 KB	No
yob1896	Text Document	13 KB	No
yob1897	Text Document	13 KB	No

Hints:

- First, use Pandas, zipfile, and BytesIO library to extract the data. Find out a way to extract only files that consists useful data.
- Hint: pd.read_csv(BytesIO(z.read(file_name)), encoding='utf-8', engine='python', header=None)
- Then, visualize the number of male and female babies born in a particular year with the help of pandas.DataFrame.plot, then Analyse baby names by sorting out all birth counts.
- Then, analyse baby names by sorting out top 100 birth counts and group them by names to find out popular baby names