Department of the Treasury Income Tax Return

**Introduction**

Taxes are the primary source of revenue for governments. Among other things, this money is spent to improve and maintain public infrastructure, fund public school, emergency services and welfare programs and national security or defense. (Kagan, 2021) Every year US Internal Revenue Service (IRS) collects data about Individual income tax return and this data contains relatively large numbers of columns which makes data collection process very expensive and difficult to do some data analysis like creating models. In this project we will visualize the data, figure out the correlation and relationship between some attributes and, we will create a model for prediction purpose. This data analysis benefits government officials to report and make projections and researchers in making policies and research taxpayer compliance and administration.

**Nature of the Data Curation**

Individual income tax return data at the state and ZIP code level is collected by US Internal Revenue Service (IRS). The Internal Revenue Service (IRS) is a U.S. government agency responsible for the collection of taxes and enforcement of tax laws (such as the wash sale rule).1 Established in 1862 by then-President Abraham Lincoln, the agency operates under the authority of the U.S. Department of the Treasury, and its primary purpose is the collection of individual income taxes and employment taxes. The IRS also handles corporate, gift, excise, and estate taxes. (Segal, 2021)

• Why did they collect the data (purpose)?

* To make projections
* To prepare reports
* To make estimates of frequencies of taxpayer entries recorded on the applicable lines of the forms and schedules filed with corporation tax returns
* To research taxpayer compliance and administration.
* To estimate gross domestic product
* To help in the development of national income accounts.
* For tax policy research.

The size of Individual income tax return data at the state and ZIP code level contains over 165k records and 126 attributes which shows the data contains every single detail about individual tax return, but this large size is somehow difficult to do some analysis like creatin a model or clustering.

Since the data has 126 columns it is difficult and visually less appealing to put the exact description of each column there fore this description was represented by some variables. The data along with documentation guide available free source.

In this project we are trying to answer the following questions.

1. Visualize Adjusted gross income at state level
2. Calculate the correlation between Total standard deduction and Total itemized deduction amounts.
3. Calculate a correlation between Ordinary dividends amount and qualified dividends amount

**Requirements and Resources needed**