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, and fund public schools, emergency services, welfare programs, and national security or defense. (Kagan, 2021) Every year, the US Internal Revenue Service (IRS) collects data about Individual income tax returns. This data contains relatively large numbers of columns, which makes the 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 create a model for prediction purposes. 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 152 attributes which shows the data contains every single detail about an individual tax return, but this large size is somehow difficult to do some analysis like creating a model or clustering.

Since the data has 152 columns it is difficult and visually less appealing to put the exact description of each column therefore this description was represented by some variables. The data along with the documentation guide available for free source.

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

1. Visualize Adjusted gross income at the state level
2. What is the correlation between itemized deductions and taxable income?
3. What is the correlation between refundable education credit and total income?
4. Calculate a correlation between Ordinary dividends amount and qualified dividends amount.

Requirements and Resources needed

In this project, we use r-studio and python for data analysis and data visualization. Since the data has 152 attributes and each attribute was assigned by a variable instead of descriptive column names, we had to select some of the attributes and assign column names in the data frame based on the requirements of our data analysis and visualization using panda’s package in python.

Descriptive Analysis

A picture containing graphical user interface

Description automatically generated

Correlation between itemized deductions and taxable income

Chart, scatter chart

Description automatically generated

The correlation between itemized deductions and taxable income is 0.88 which implies that their strong correlation as itemized deduction increases taxable income increases.

# References

Kagan, J. (2021, October 14). *Tax Definition*. Retrieved from www.investopia.com: https://www.investopedia.com/terms/t/taxes.asp#:~:text=27-,Why%20Do%20We%20Pay%20Taxes%3F,emergency%20services%2C%20and%20welfare%20programs.

Segal, T. (2021, December 23). *Internal Revenue Service (IRS)*. Retrieved from www.investopedia.com: https://www.investopedia.com/terms/i/irs.asp