**About the Data**

The dataset contains information about US taxpayers. There are 10 independent columns and 1 dependent column. This dataset includes attributes like household income, household debt level, if the taxpayer is married or not, how many cars their household has, if they filed their taxes in the last three years or not. Some of the attribute informations are given below:

To load the training data in your jupyter notebook, use the below command:

import pandas as pd

tax\_data  = pd.read\_csv("https://raw.githubusercontent.com/dphi-official/Datasets/master/tax\_payers/train\_set\_label.csv" )

**Data Description**

* HHI: Household income
* HHDL: Household debt level
* Married: There are three categories for a taxpayer 0, 1, 2 with regards to marriage.
* PoliticalParty: Name of the political party
* CollegeGrads: Grade in College out of 5
* AHHAge: Average household age
* cars: number of cars in house
* Filed in YYYY: Tax filed in given year YYYY

**Saving Prediction File & Sample Submission**

You can find more details on how to save a prediction file here: <https://discuss.dphi.tech/t/how-to-submit-predictions/548>

**Sample submission:**

 You should submit a CSV file with a header row and the sample submission can be found below.

prediction

Democrat

Independent

Republican

Republican

Independent

Democrat

Republican

.

.

Etc.

Note that the header name should be prediction else it will throw an evaluation error. A sample submission file can be found

[here](https://github.com/dphi-official/Datasets/blob/master/tax_payers/sample_submission.csv" \t "_blank)

**Test Dataset**

Load the test data (name it as test\_data). You can load the data using the below command.

test\_data = pd.read\_csv('https://raw.githubusercontent.com/dphi-official/Datasets/master/tax\_payers/test\_set\_label.csv')

Here the target column is deliberately not there as you need to predict it