# Sampling bias correction for TV viewership data

## Introduction

Consider a universe of television-watching individuals, denoted by the set $U$. We have access to data on the individual viewership behavior of a subset of $U$, denoted by $U_s$. The $U_s$ is known to be a biased sampling of $U$; specifically, certain demographic categories are over– or under–represented in the sampled dataset with respect to the national population. We would like to model and forecast the viewership activity of $U$ using our information on the members of $U_s$. Consequently, the bias in $U_s$ must be measured and compensated for.

## The Challenge

In this assignment you are required to tackle the problem of sampling bias correction in a realistic set of TV viewer data. For this we provide you with the following two files:

1. `'demographic_attributes.csv'` : This table is derived from a dataset on national TV viewers; it associates 400,000 `person_id`s with demographic categories to which the person belongs.

2. `'demo_ground_truth.csv'` : This table contains census-derived values for the number of individuals belonging to each of the demographic categories in `'demographic_attributes.csv'`. The categories are of three types: age, ethnicity, and education level.

We would like for you to devise an approach to balance the sampled dataset of people in demographic attributes.csv on each demographic category; i.e., the goal here is to compute a set of *person–level* weights (one single weight per person) that unbiases the dataset. Please submit *only* the following two files for your answer.

1. `'<APPLICANT_NAME>_VA_DS_challenge_weights.csv'` : A text file of comma-separated values consisting of your set of bias-correcting weights. There should be two columns: `person_id` and `weight`. The sum of the weights in each category should be consistent with the census data in `'demo_ground_truth.csv'`.

2. `'<APPLICANT_NAME>_VA_DS_challenge.ipynb'` : A Jupyter notebook file containing the code (in Python) you used to solve the problem, along with explanations of your methodology and results.

## How do we assess your work?

The Jupyter notebook you provide is the presentation of your work. The notebook has to be

- Clean, readable, and easy to follow.
- All nonempty cells should be executed in a sequential manner. Do not clear the output of the execution, however you can truncate the output if it is too long.

Your work will be judged by how clearly you justify, illustrate and explain your choices and the approach(es) you took to attempt this challenge. Please note that the precise mathematical formulation of this problem is up to you.

**Note: Please reply all when submitting your files to ensure we review it in timely manner.**