Chris Iyer, 1/18/2023

Here, I tested each method in this document on the Georgia voter list file (which contains race information for each entry; merged 2020 and 2022).

Goal: predict race for entries in Texas voter list file. To do so, we want to test a variety of race prediction methods on Georgia voters (who we have race information on to assess accuracy) and select the best.

Links to code files:

- Main google folder
- R notebook (predictrace & wru libraries)
- Python notebook (ethnicolr library)

Results:
Tl;dr WRU with geocoding reaches pretty solid and balanced performance for AAPIs as well as other racial groups. Predictrace full-name averaging comes in second. These options, as well as numerous different combination schemes of different predictions, can be titrated to minimize false positives vs. false negatives. But, as a first pass, I recommend using WRU with geocoding if you have access to county-, tract-, or block-level information, and using predictrace with an averaging scheme like the one I use in the absence of geographic data.

I report sensitivity (% of people of a category that correctly get the label / coverage) labeled in that category) and positive predictive value (% of people of a label who are indeed of that category)

1 - sensitivity = false negative rate
1 - PPV = false positive rate

- Predictrace (R)
    - Surname only

```
[1] "Race: WH  Sensitivity: 0.8996  PPV: 0.6549"
[1] "Race: BH  Sensitivity: 0.1549  PPV: 0.7698"
[1] "Race: HP  Sensitivity: 0.6174  PPV: 0.7202"
[1] "Race: AP  Sensitivity: 0.678  PPV: 0.8"
[1] "Race: AI  Sensitivity: 0.0047  PPV: 0.1332"
```

    - First name only (uses this dataset)

```
[1] "Race: WH  Sensitivity: 0.9108  PPV: 0.7002"
[1] "Race: HP  Sensitivity: 0.3594  PPV: 0.5862"
[1] "Race: BH  Sensitivity: 0.0352  PPV: 0.895"
[1] "Race: AP  Sensitivity: 0.2186  PPV: 0.8089"
```

- Average probabilities from first and last name predictions; select max

```
[1] "Race: WH  Sensitivity: 0.9808  PPV: 0.6426"
[1] "Race: HP  Sensitivity: 0.6009  PPV: 0.7807"
[1] "Race: BH  Sensitivity: 0.0971  PPV: 0.9269"
[1] "Race: AP  Sensitivity: 0.6398  PPV: 0.8904"
[1] "Race: AI  Sensitivity: 0.0017  PPV: 0.1796"
```

- Ethnicolr (Python)
  - Census models (surname only)

```
2000 model:
Race: AP, Sensitivity: 0.5918, PPV: 0.7985
Race: BH, Sensitivity: 0.0455, PPV: 0.7334
Race: HP, Sensitivity: 0.7572, PPV: 0.6894
Race: WH, Sensitivity: 0.9703, PPV: 0.6238
2010 model:
Race: AP, Sensitivity: 0.5918, PPV: 0.7985
Race: BH, Sensitivity: 0.0455, PPV: 0.7334
Race: HP, Sensitivity: 0.7572, PPV: 0.6894
Race: WH, Sensitivity: 0.9703, PPV: 0.6238
```

  - Wiki models (surname and full name, respectively)

```
last name wiki model:
Race: AP, Sensitivity: 0.4366, PPV: 0.2752
Race: BH, Sensitivity: 0.0203, PPV: 0.3432
Race: HP, Sensitivity: 0.5234, PPV: 0.6019
Race: WH, Sensitivity: 0.9304, PPV: 0.6076
full name wiki model:
Race: AP, Sensitivity: 0.5492, PPV: 0.5097
Race: BH, Sensitivity: 0.0352, PPV: 0.5171
Race: HP, Sensitivity: 0.5114, PPV: 0.5661
Race: WH, Sensitivity: 0.9608, PPV: 0.6201
```

  - FL registered voter models (full name; 4 options, 4 options + other)

```
full name FL model:
Race: AP, Sensitivity: 0.5938, PPV: 0.8605
Race: BH, Sensitivity: 0.4169, PPV: 0.8451
Race: HP, Sensitivity: 0.8109, PPV: 0.6747
Race: WH, Sensitivity: 0.9423, PPV: 0.724
full name FL model with "other" category:
Race: AP, Sensitivity: 0.8115, PPV: 0.5517
Race: BH, Sensitivity: 0.7189, PPV: 0.6272
Race: HP, Sensitivity: 0.8139, PPV: 0.6584
Race: WH, Sensitivity: 0.7006, PPV: 0.8309
```

- NC registered voter model (full name; gives Hispanic Y/N + Race, so there are different ways of mapping onto the voter list race categories)

```
full name NC model (version 1):
Race: AI, Sensitivity: 0.0892, PPV: 0.0044
Race: AP, Sensitivity: 0.5643, PPV: 0.2864
Race: BH, Sensitivity: 0.3818, PPV: 0.5563
Race: HP, Sensitivity: 0.7988, PPV: 0.2366
Race: WH, Sensitivity: 0.3866, PPV: 0.7614
full name NC model (version 2):
Race: AI, Sensitivity: 0.0908, PPV: 0.0044
Race: AP, Sensitivity: 0.5664, PPV: 0.2837
Race: BH, Sensitivity: 0.4159, PPV: 0.5536
Race: WH, Sensitivity: 0.4442, PPV: 0.7319
```

- Wru (R)
  - Surname only (no geocoding)

```
[1] "Race: WH  Sensitivity: 0.9505  PPV: 0.6522"
[1] "Race: BH  Sensitivity: 0.1681  PPV: 0.7588"
[1] "Race: HP  Sensitivity: 0.7933  PPV: 0.7457"
[1] "Race: AP  Sensitivity: 0.6914  PPV: 0.7763"
```

  - County-level census geocoding

```
[1] "Race: WH  Sensitivity: 0.7594  PPV: 0.8146"
[1] "Race: HP  Sensitivity: 0.7805  PPV: 0.7635"
[1] "Race: BH  Sensitivity: 0.7309  PPV: 0.6371"
[1] "Race: AP  Sensitivity: 0.6746  PPV: 0.8617"
```

Top Scorers:

1. WRU (surname + geocoding)
   - AP: sens = 67%, PPV = 86%
   - Slightly better on other races
2. Predictrace (firstname + surname average)
   - AP: sens = 64%, PPV = 89%

- Combination of (1) and (2): mark as asian if one or the other
  - AP: sens = 71%, PPV = 83%
- Combination of (1) and (2): mark as asian if both
  - AP: sens = 60%, PPV = 93%

Choose various different combinations to maximize sensitivity or PPV

Untested methods:

- List of Asian last names from this paper (contact lauderdale@health.bsd.uchicago.edu to get these lists)

    - This would *only* match Asian names
    - Likely high PPV (low false positive rate) and low sensitivity (high false negative rate)
- This thing: https://rpubs.com/jwcb1025/est_ethnicity
- Calculating census tract and/or block from voter addresses; using WRU tract/block info