This is my thoughts about each column that has missing data:

* Age: Can be imputed using iterative imputation: other features are used.
* Ethnicity: Even there are “other” values in this column, “nan” still exists, which may indicate a missing-not-at-random pattern (for example, “mixed” type). Thus, nan will be converted to “Unknown” (no imputation)
* Marital status: The missing value outnumbers other values (non-traditional and married) gives me a feeling that this is also a missing-not-at-random pattern. Usually, married and single take the majority part. However, the statistics from <https://statisticalatlas.com/place/Nevada/Las-Vegas/Marital-Status> shows me that the gap between the two status is not that far (married even outnumbers never married). Therefore, I decide to use the same technique as ethnicity’s
* Dwelling type: Possible candidate for missing values may be homelessness. However, from <https://www.usich.gov/homelessness-statistics/nv/>, its homeless population is only 6900 out of 3,000,000. I think imputation is too risky. Same technique as ethnicity’s.
* Education: Same technique as ethnicity’s.
* Cd: There are only 4 / 50k. Iterative imputation can work it out.
* Occupation industry: As the missing account for over 80% with the existing of “other”, same technique as ethnicity’s.
* Net\_worth: My gut tells me the missing value is highly because survey participants did not want to reveal their property, and it takes as much as half of all survey participants. So, same technique as ethnicity’s.
* petowner\_dog, intrst\_nascar\_in\_hh, intrst\_musical\_instruments\_in\_hh,: “nan” will become “no” as there are only two values
* donates\_to\_liberal\_causes, donates\_to\_conservative\_causes: In my opinion, they did not want to reveal their political affiliation while those answered “yes” are sure about which party they would like to vote for. Same technique as ethnicity’s.
* home\_owner\_or\_renter: “nan” makes up around 56% beside “Likely renter” and “Likely homeowner”. I have no idea why there is such “likely” but I will leave it as it is or “Unknown”
* g08, g10, g12, p08, p10, p12\_precinct\_turnout

Summary:

|  |  |  |  |
| --- | --- | --- | --- |
| **Column** | **Iterative Imputation** | **“nan” / np.nan to “Unknown” as one category** | **“nan” to “No”** |
| Age | ✓ |  |  |
| Ethnicity |  | ✓ |  |
| Marital status |  | ✓ |  |
| Dwelling type |  | ✓ |  |
| Education |  | ✓ |  |
| cd | ✓ |  |  |
| Occupation industry |  | ✓ |  |
| Net worth |  | ✓ |  |
| petowner |  |  | ✓ |
| intrst\_nascar\_in\_hh |  |  | ✓ |
| intrst\_musical\_instruments\_in\_hh |  |  | ✓ |
| donates\_to\_liberal\_causes |  | ✓ |  |
| donates\_to\_conservative\_causes |  | ✓ |  |
| home\_owner\_or\_renter |  | ✓ |  |
| precinct\_turnouts | ✓ |  |  |