1. We First load all datasets(Cost of living, weather, population) in the notebook
2. For cost of living dataset, we find the number of adults and children for each datapoint.
3. We then merge the Cost of living dataset with the weather dataset based on common state and county values
4. We remove some columns like “ID”, “RANK”, “case\_id” etc from the merged dataset as they are not relevant to our research
5. For the population dataset, we only take the values for year 4 as it contains the latest data
6. We drop columns like “SUMLEV”, “YEAR” etc as these columns don’t have relevant data.
7. We Only select the features ‘POPEST\_MALE', 'POPEST\_FEM', 'MEDIAN\_AGE\_MALE', 'MEDIAN\_AGE\_FEM' as these are the most relevant features.
8. We then merge the Cost of living dataset with the population dataset based on common state and county values
9. We then use a pie chart to show the state classes are unbalanced and use undersampling to balance the dataset