1. Understand the dataset
2. Merge Female & Male datasets
3. Delete unnecessary column based on the understanding of the data [which I **strongly** believe will not help to predict the target]
4. Delete classes of the target with less than 500 observations [as the question stated]
5. Map the remaining classes into their real names instead of the encoding [The mapping is from the PDF file of the dataset]
6. Check null and duplicates values
7. Examine the correlation [it is difficult due to the high number of variables]
8. I tried once to remove features with 5 or more correlations >0.9 or <-0.9, But it did not workout well [I excluded it from this file]
9. We built 4 models: [Logistic Regression, SVM, Random Forest, XGBoost] all with the same # of CV's and Random\_state using all features where maximizing Hispanic recall was the objective
10. We also tried an attempt with XGBoost where the target is F1 of the Hispanic since the recall Hispanic was too much going to it which lead to bad precision of Hispanic and bad result for thew other classes]
11. In Random Forest & XGBoost, **ordinal encoding** was done even for ordinal categorical variables [ Recommended based on experiments]
12. In Both, scaling was ignored
13. Since all the results were not satisfactory, we did:
14. SMOTE which is over-sampling technique for the classes, we applied it along with just partial features, we had chosen the features after looking in the box plots of each feature and chose the feature where there is variability across the classes distribution. [66 feature]
15. Examine feature by feature is not efficient when the number is high, However, we tried it.
16. The results of **All** the models are not satisfactory which means extra feature engineering is required, such as group the birth location [to avoid high cordinality] & Deal with multicollinearity [Which expected to raise the performance of LR & SVM].
17. The summary result is at the end of the notebook
18. However, we will ignore some overfitting for the sake of the project and use the best worse one: Which was logistic Regression with smote and parital features
19. LogisticRegression(C=1, class\_weight='balanced', penalty='l1', solver='saga')
    1. Fit it on the whole original dataset
       1. After transformation [scaling, ecndoing, etc]
       2. Using the hyperparameter of the model with SMOTE