

EXAMINING INFLUENTIAL FACTORS ON POVERTY STATUS

BY :

BRIAN PINTO

NORA ADADUROVA

MAIA CLEMONS

REBECCA PIERRE



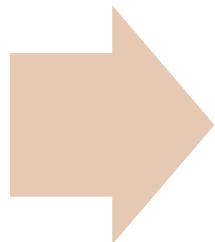


OVERVIEW

- Identify economic disparities in the United States.
- Apply data analysis techniques to detect at-risk groups.

METHODOLOGY

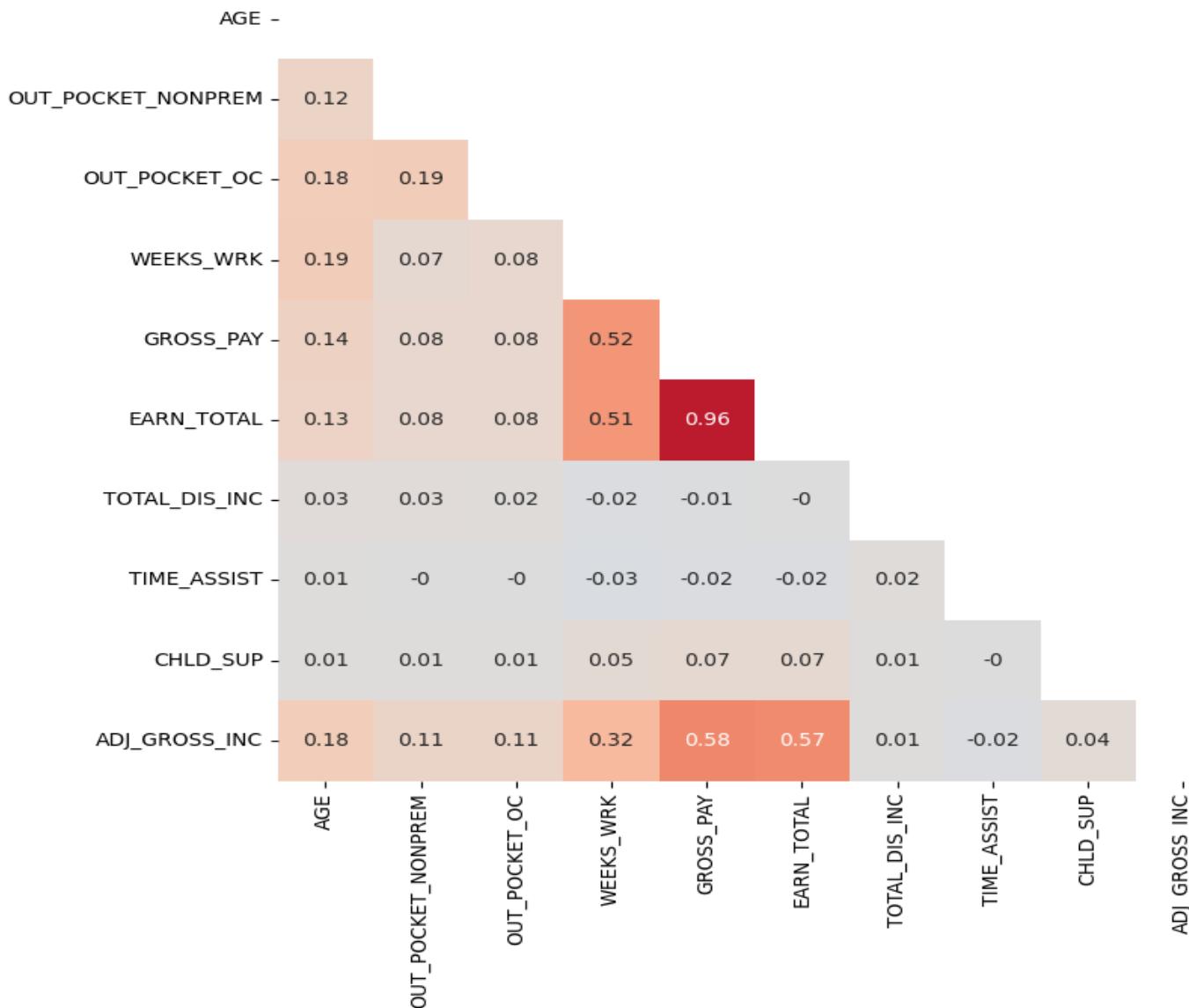
Utilize clustering and classification techniques for data analysis.



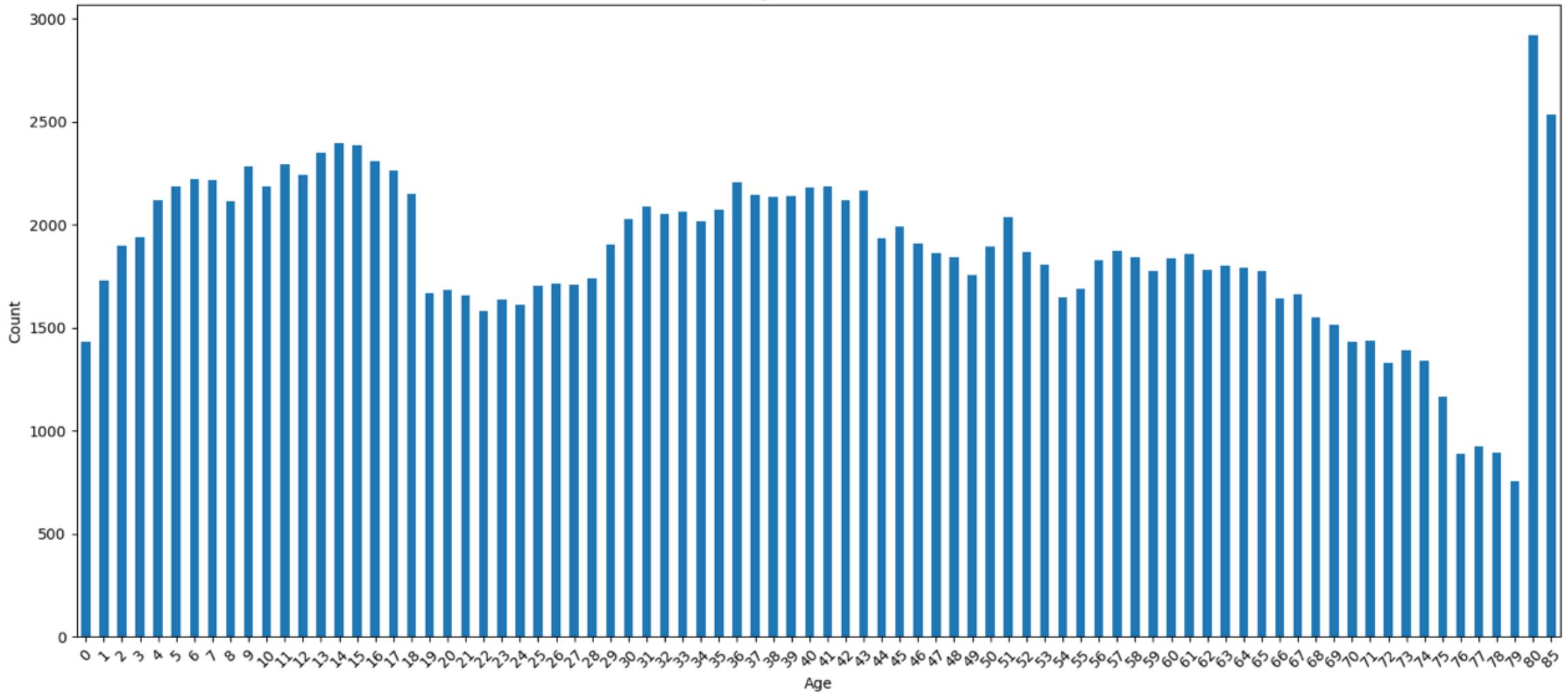
Group individuals based on Occupation, healthcare access, and other influential factors.

EDA

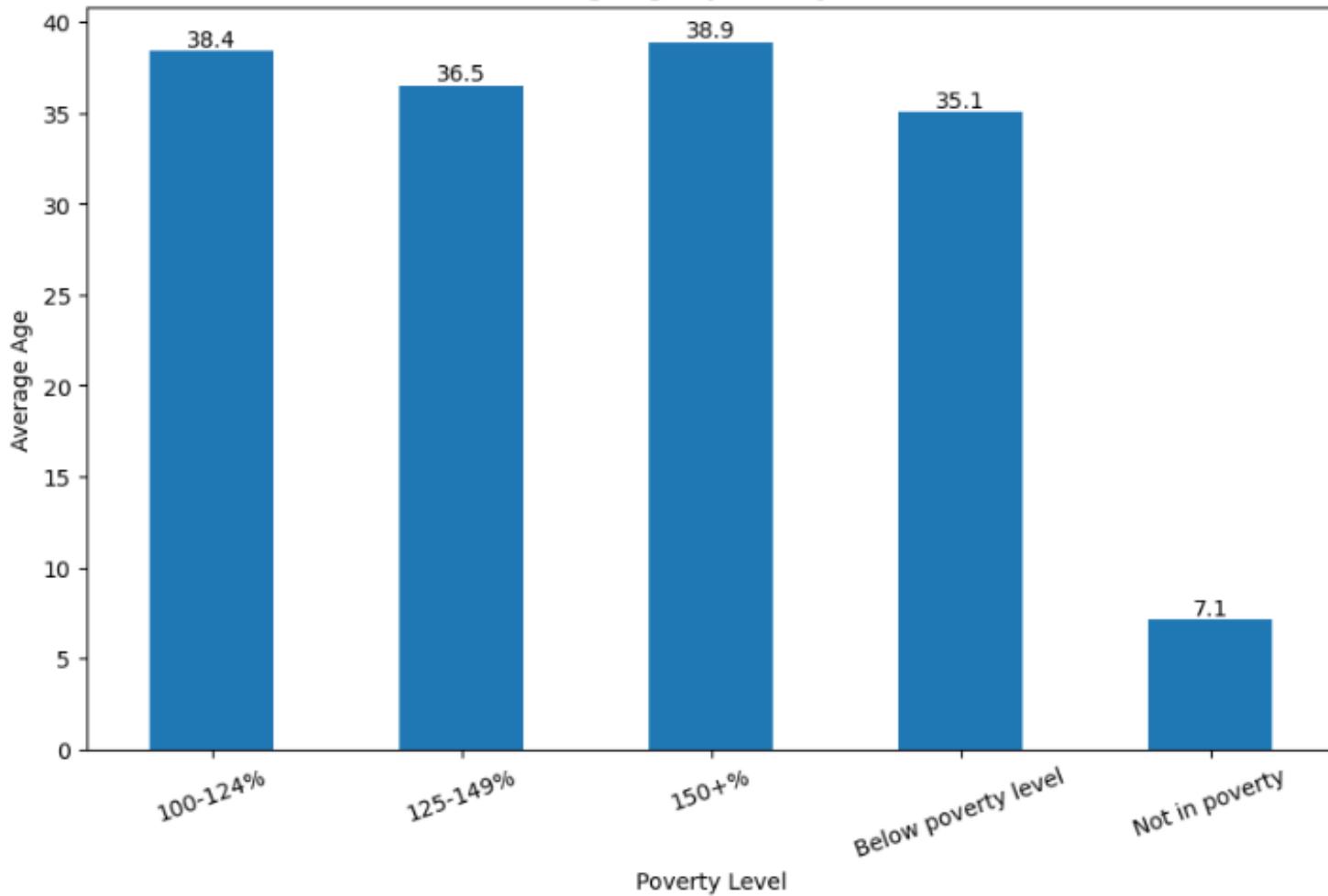
Correlation of Features



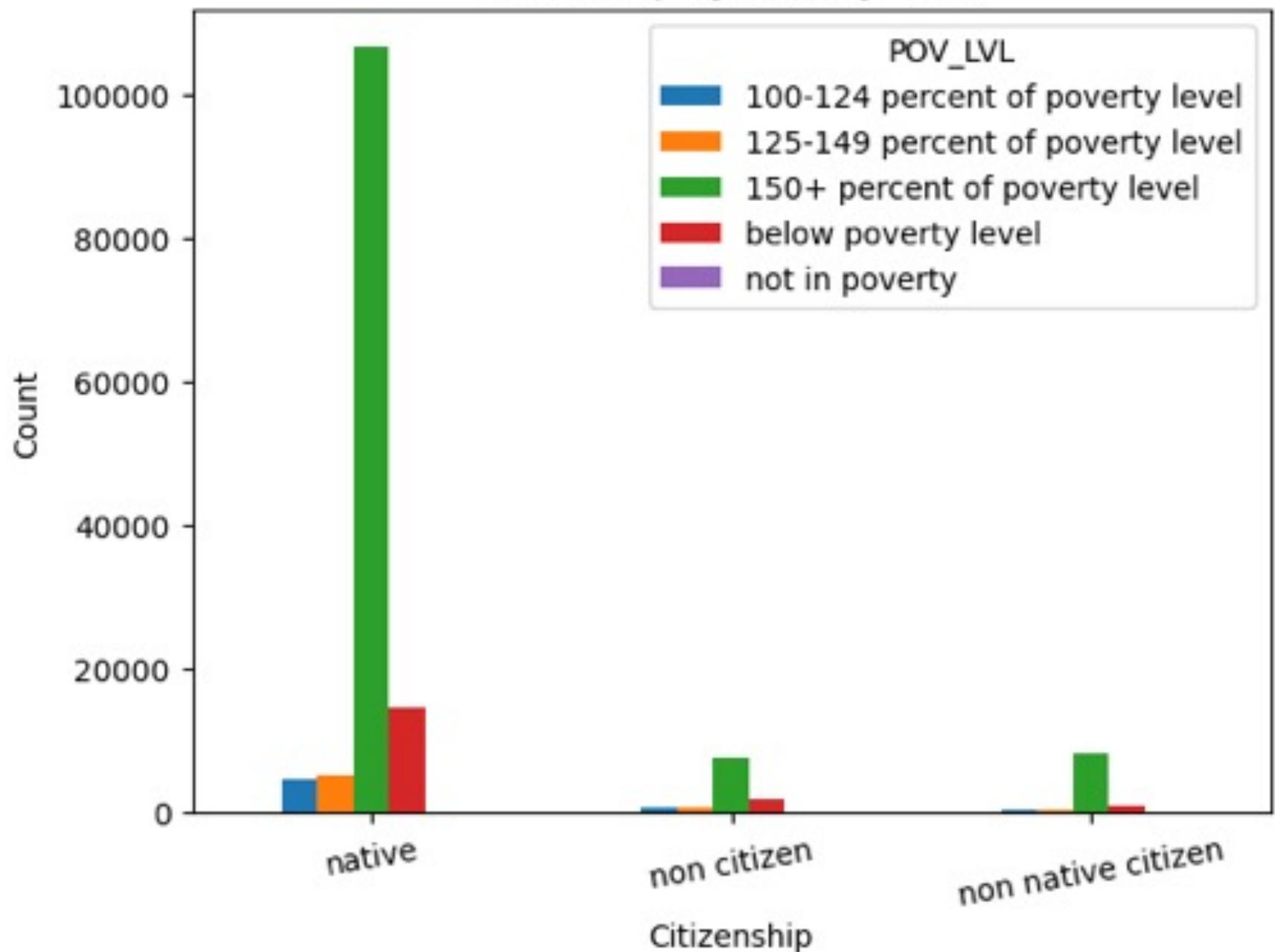
Age Distribution



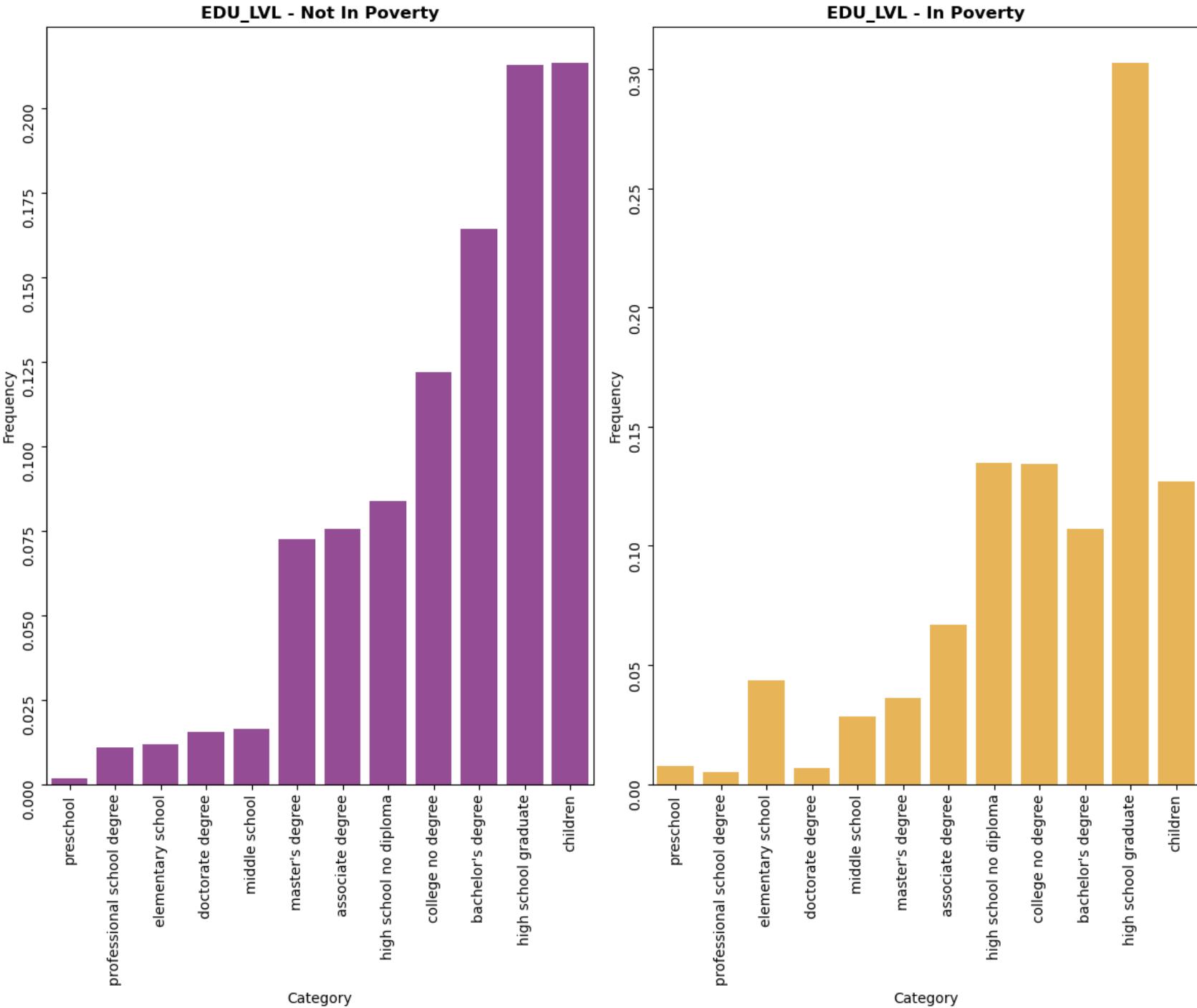
Average Age by Poverty Level



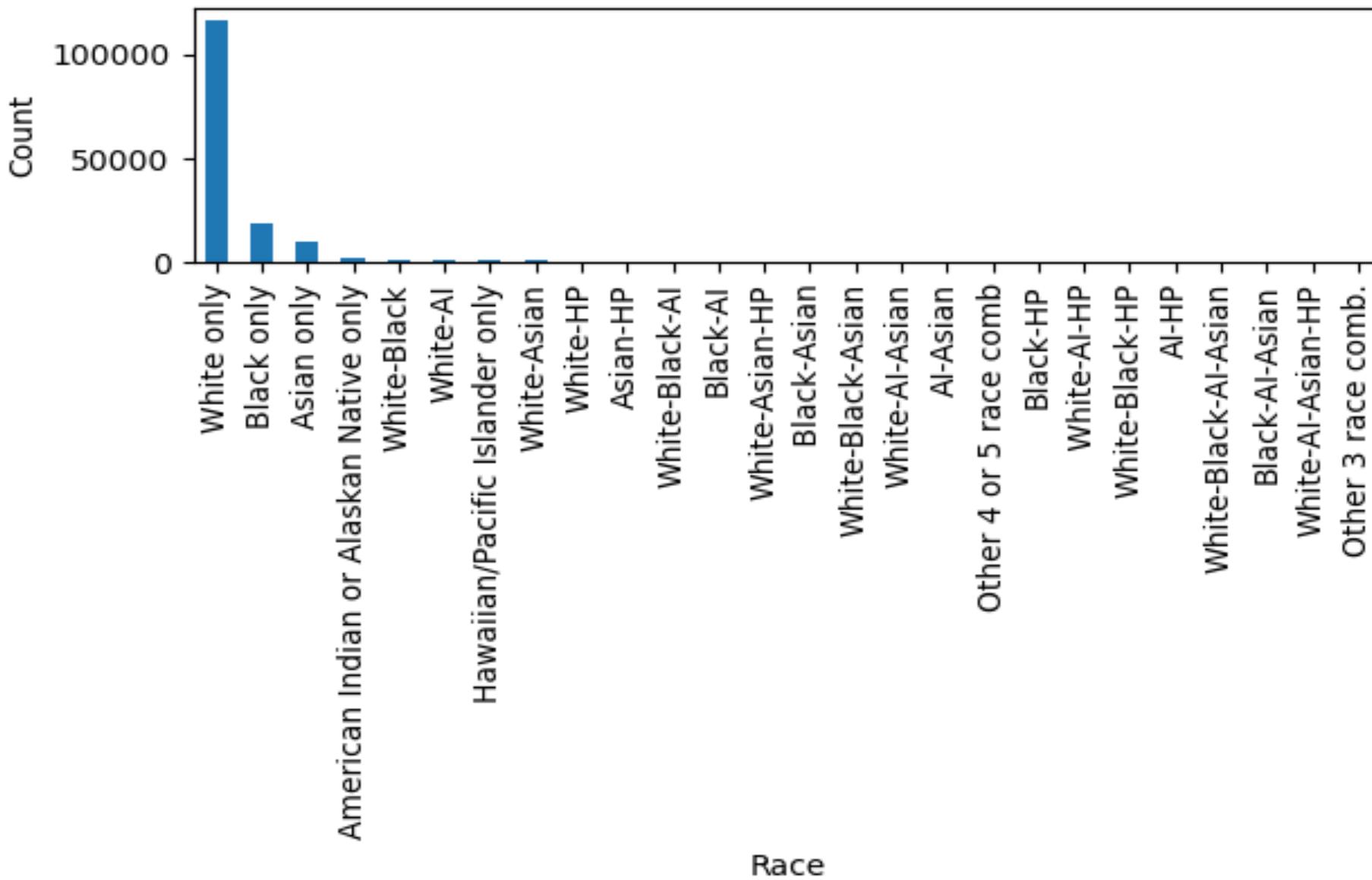
Citizenship by Poverty Level



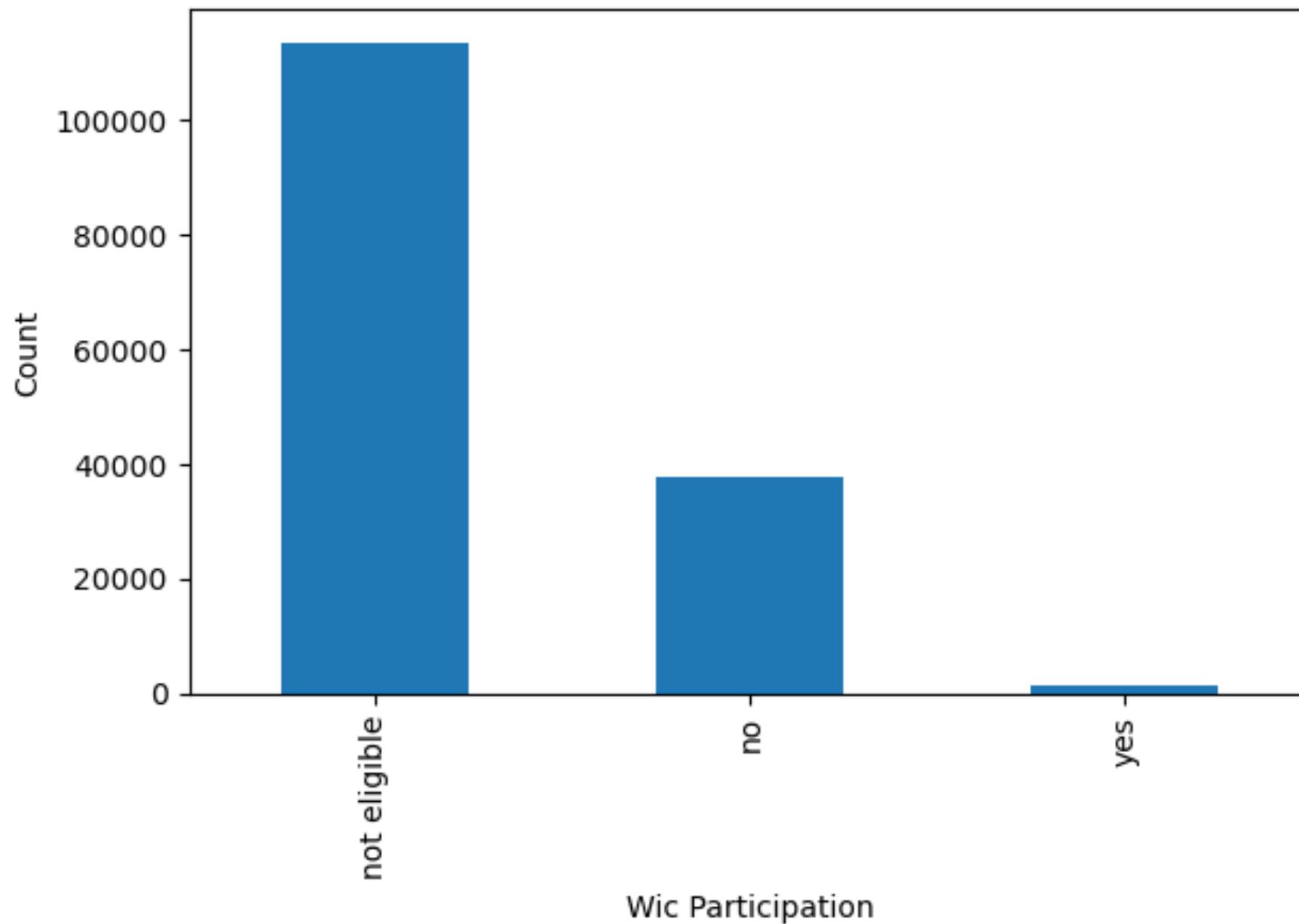
Education Levels



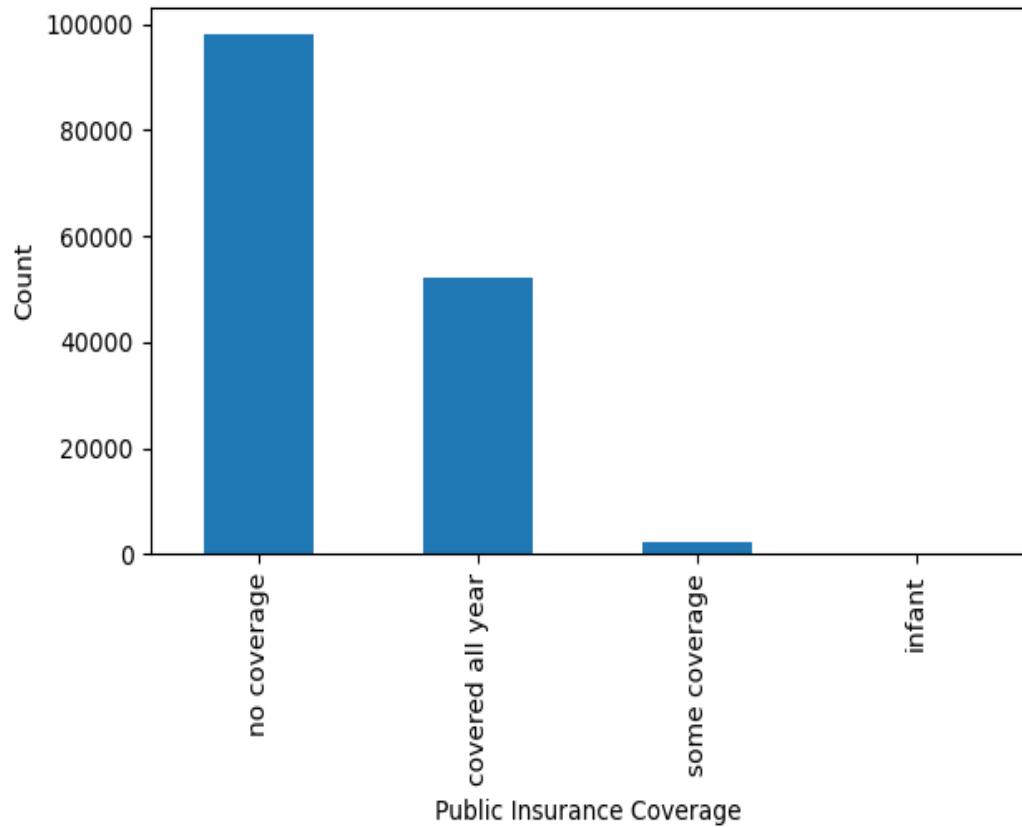
Distribution of Race



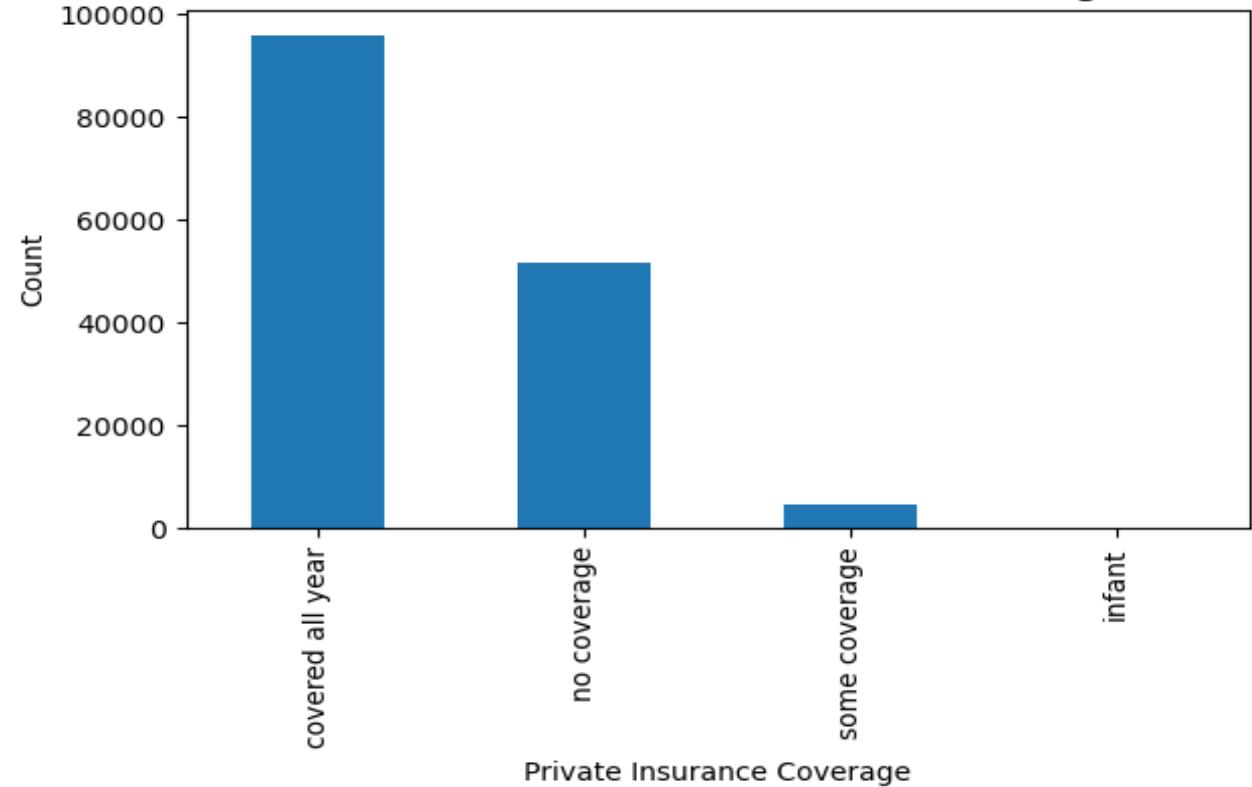
Distribution of Wic Participation



Distribution of Public Insurance Coverage



Distribution of Private Insurance Coverage



MODELS USED



Logistic Regression



Gradient Boosting

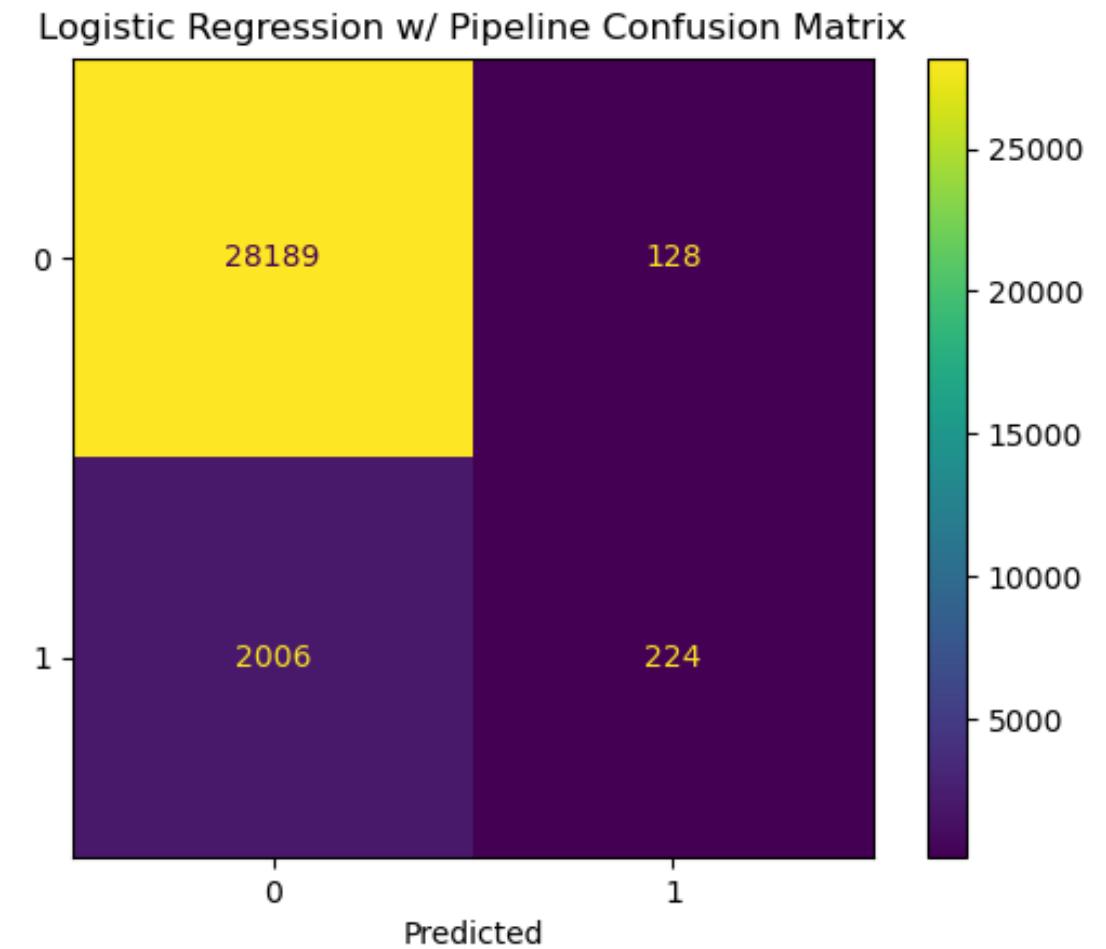
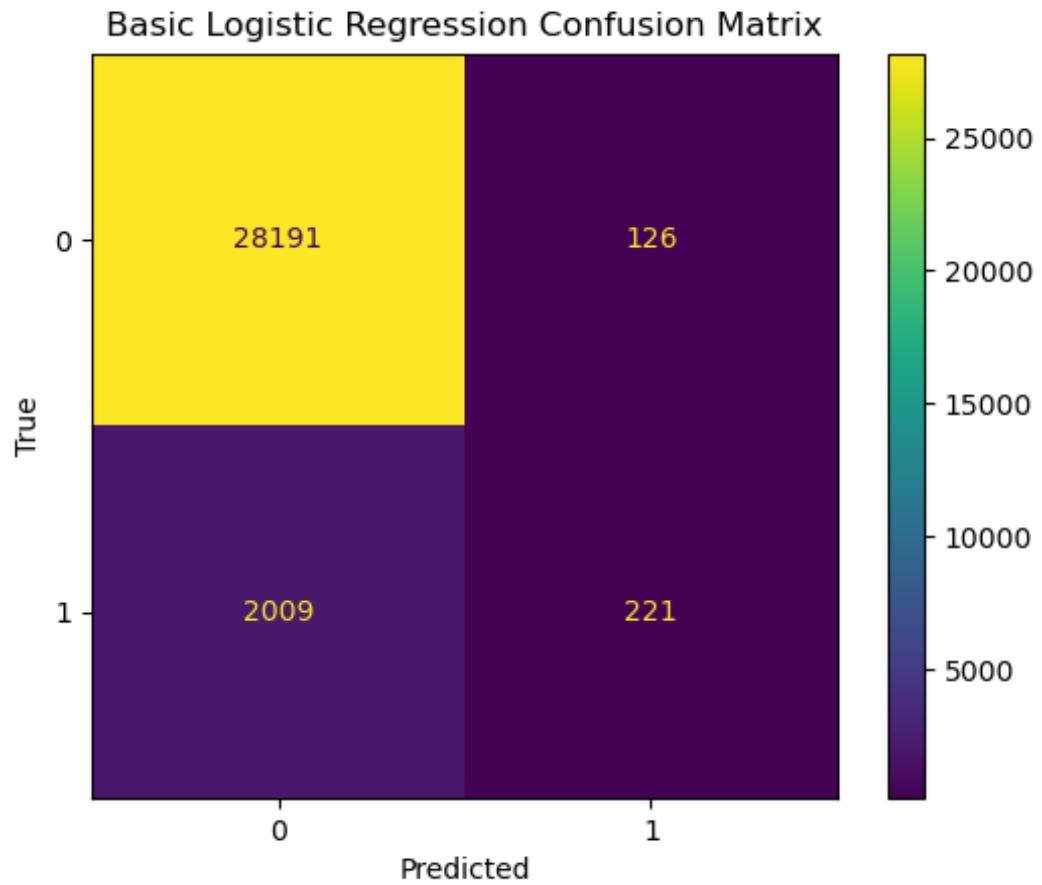


Random Forest

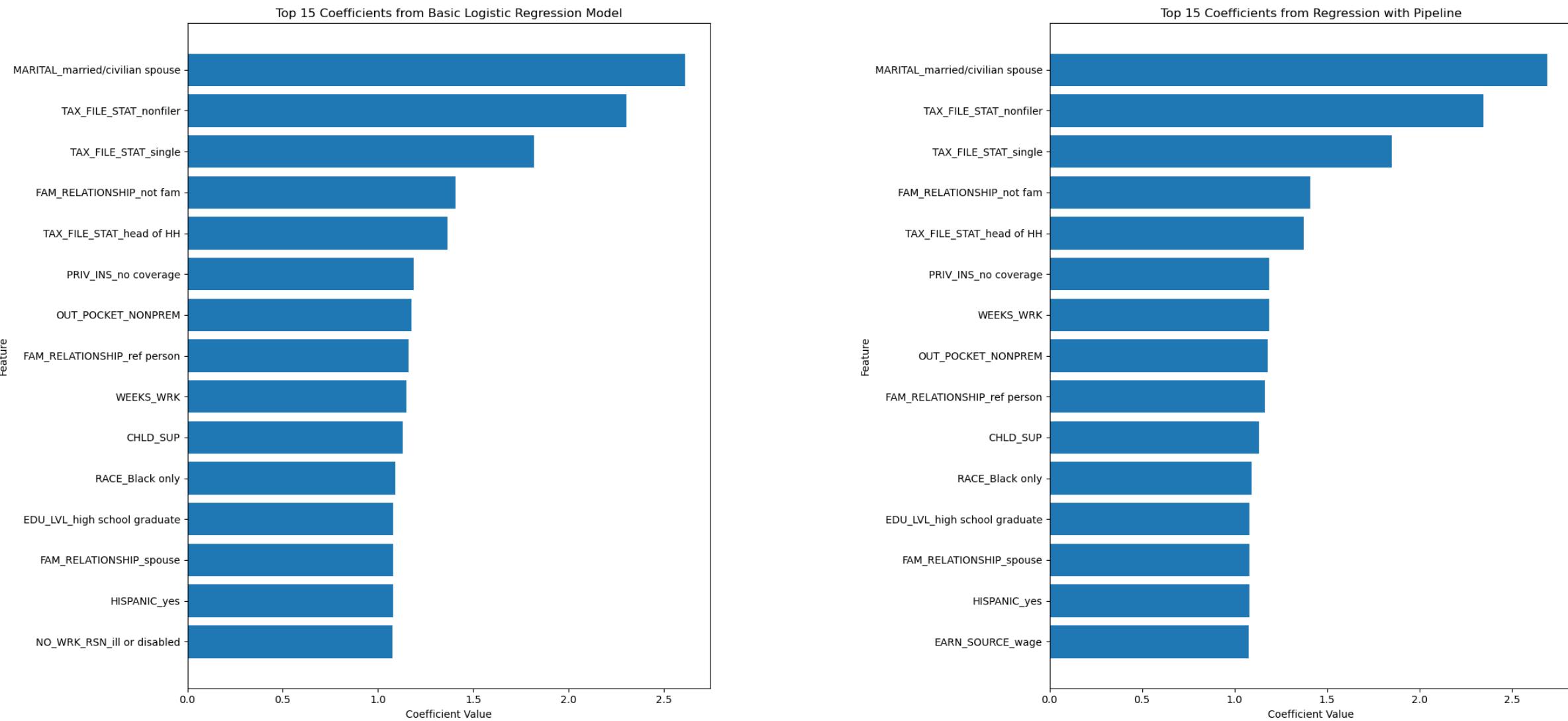


Clustering

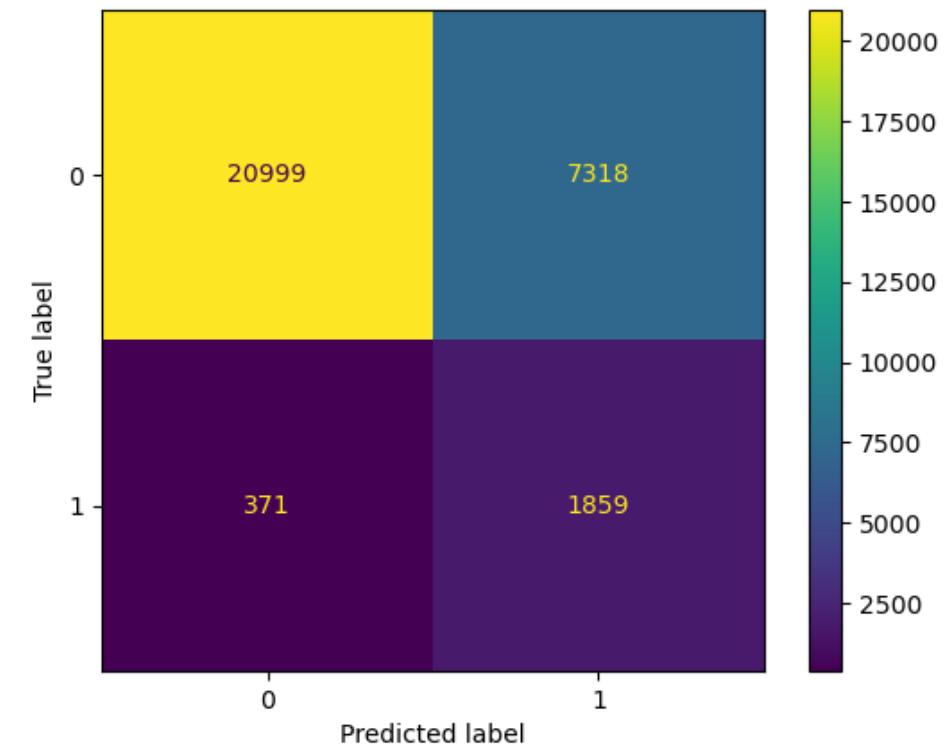
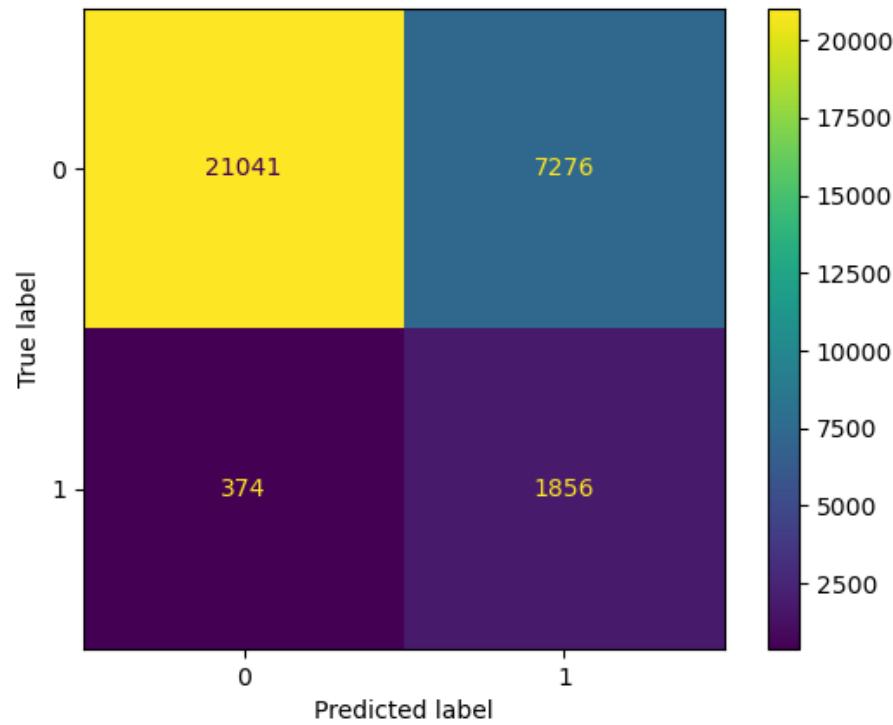
LOGISTIC REGRESSION



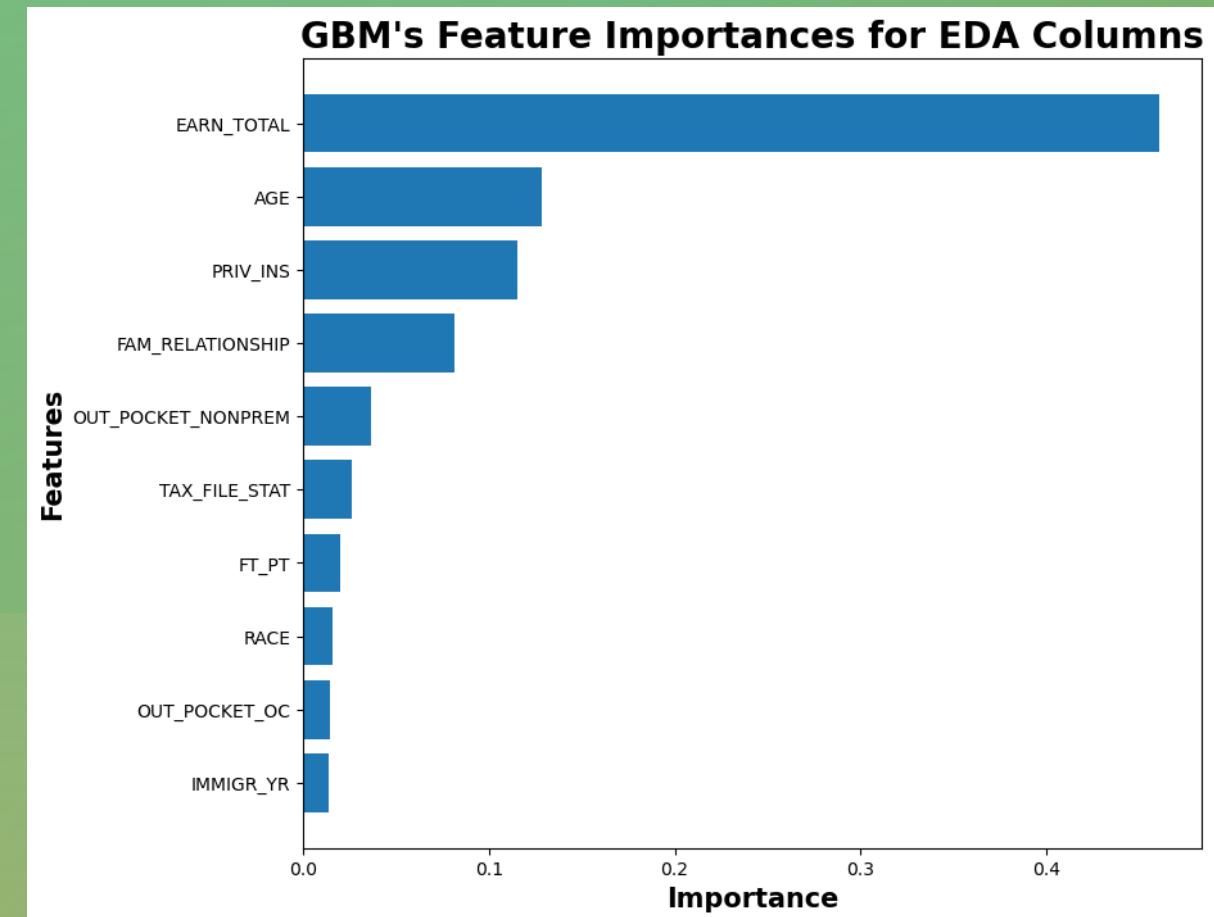
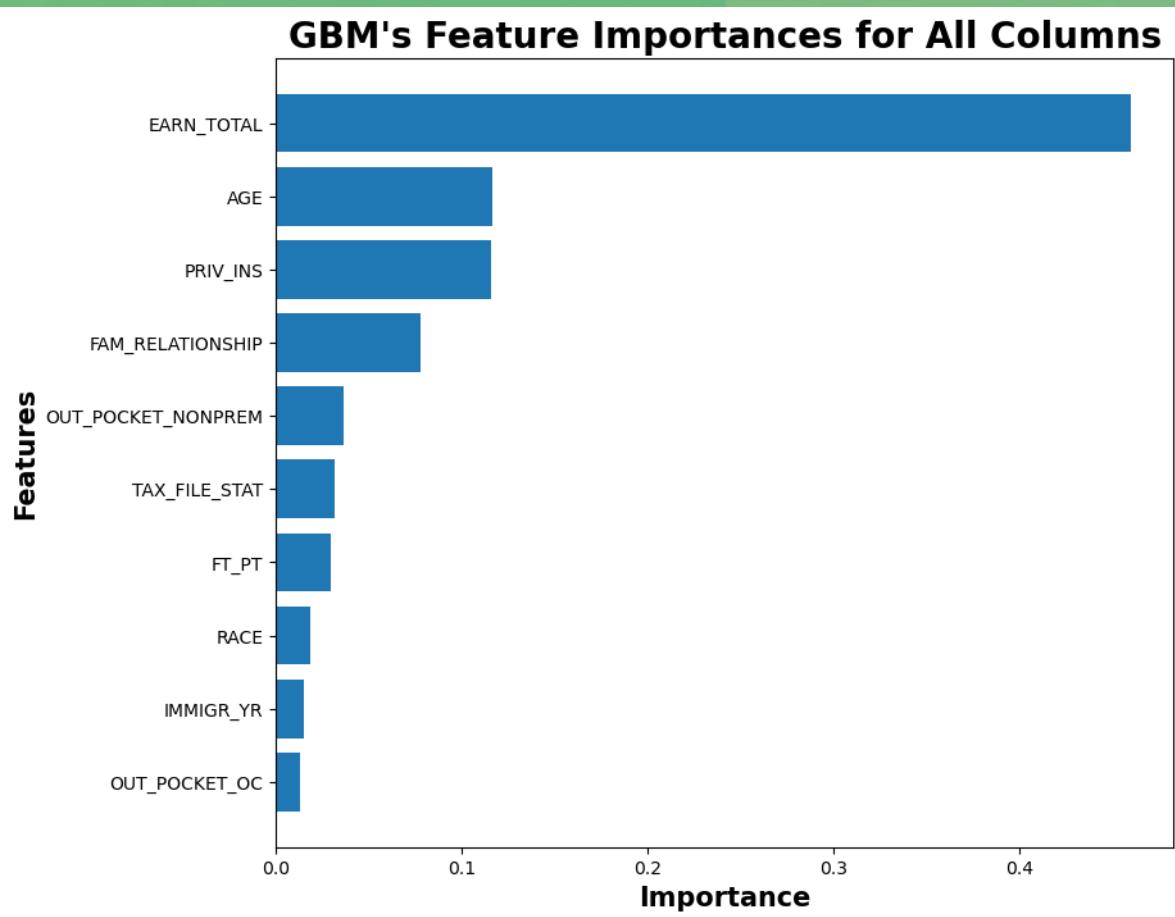
LOGISTIC REGRESSION



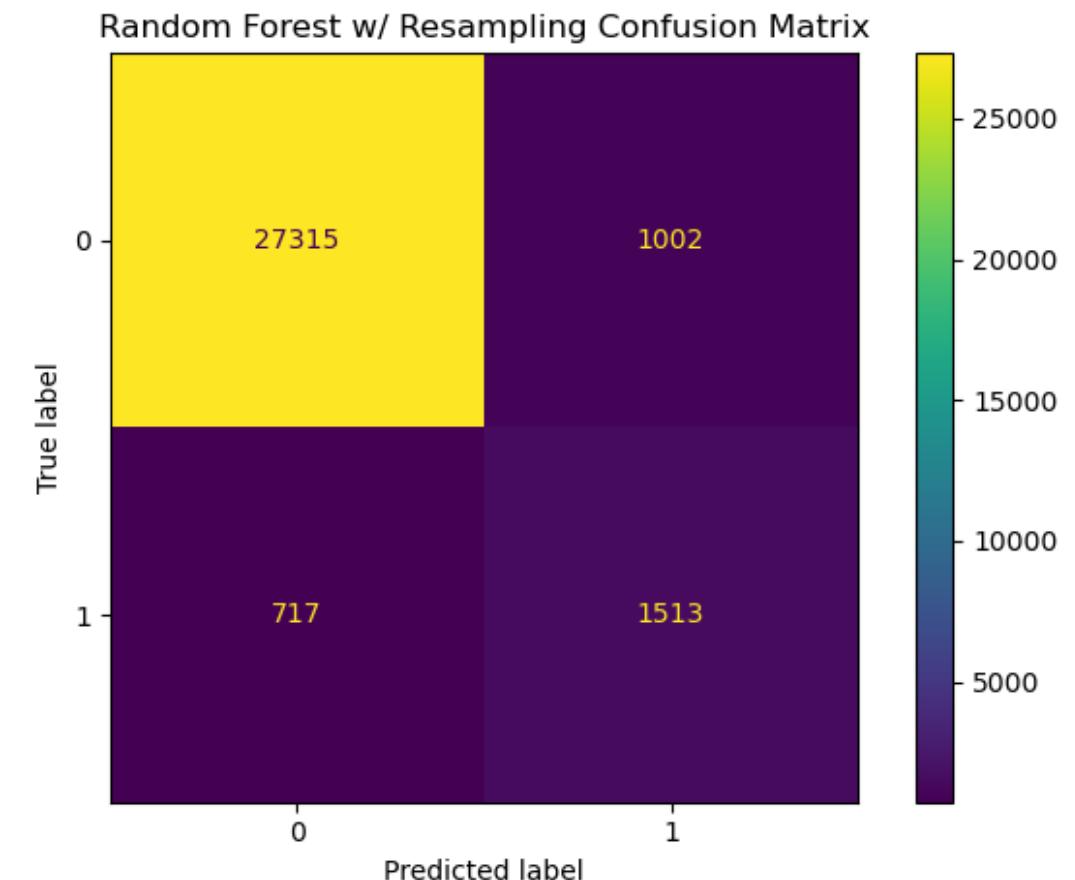
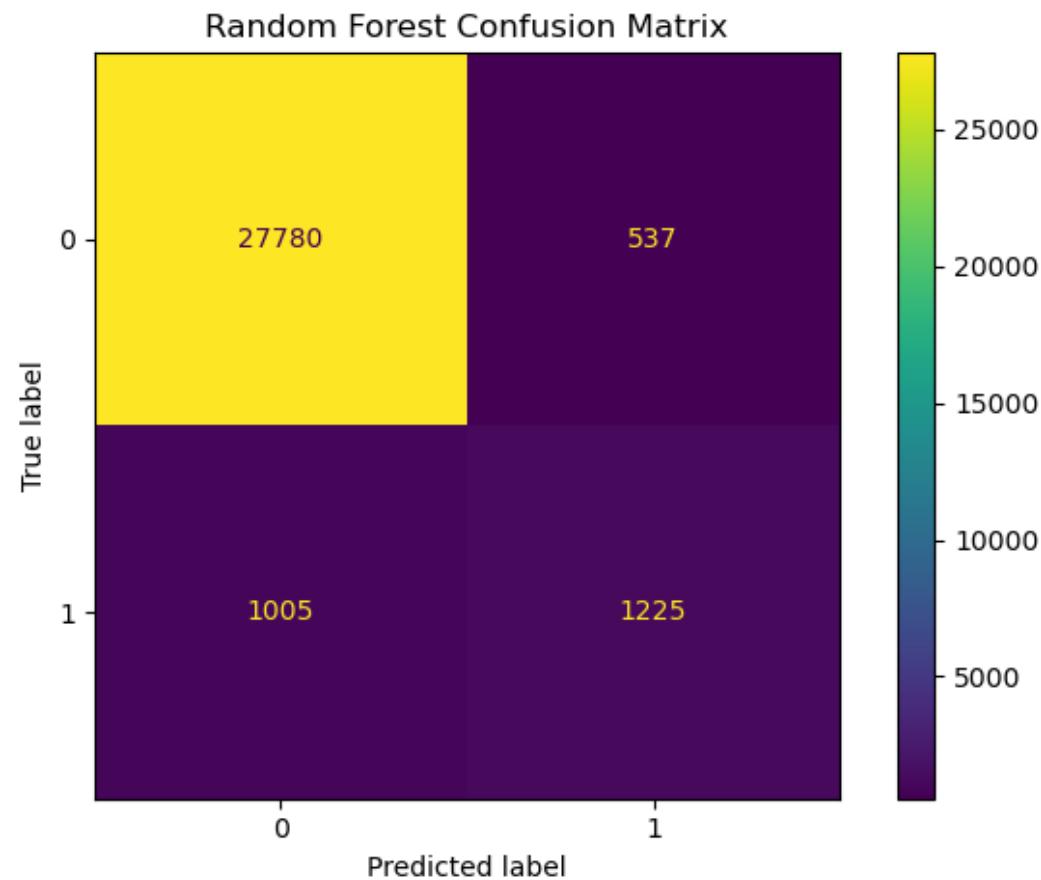
GRADIENT BOOSTING



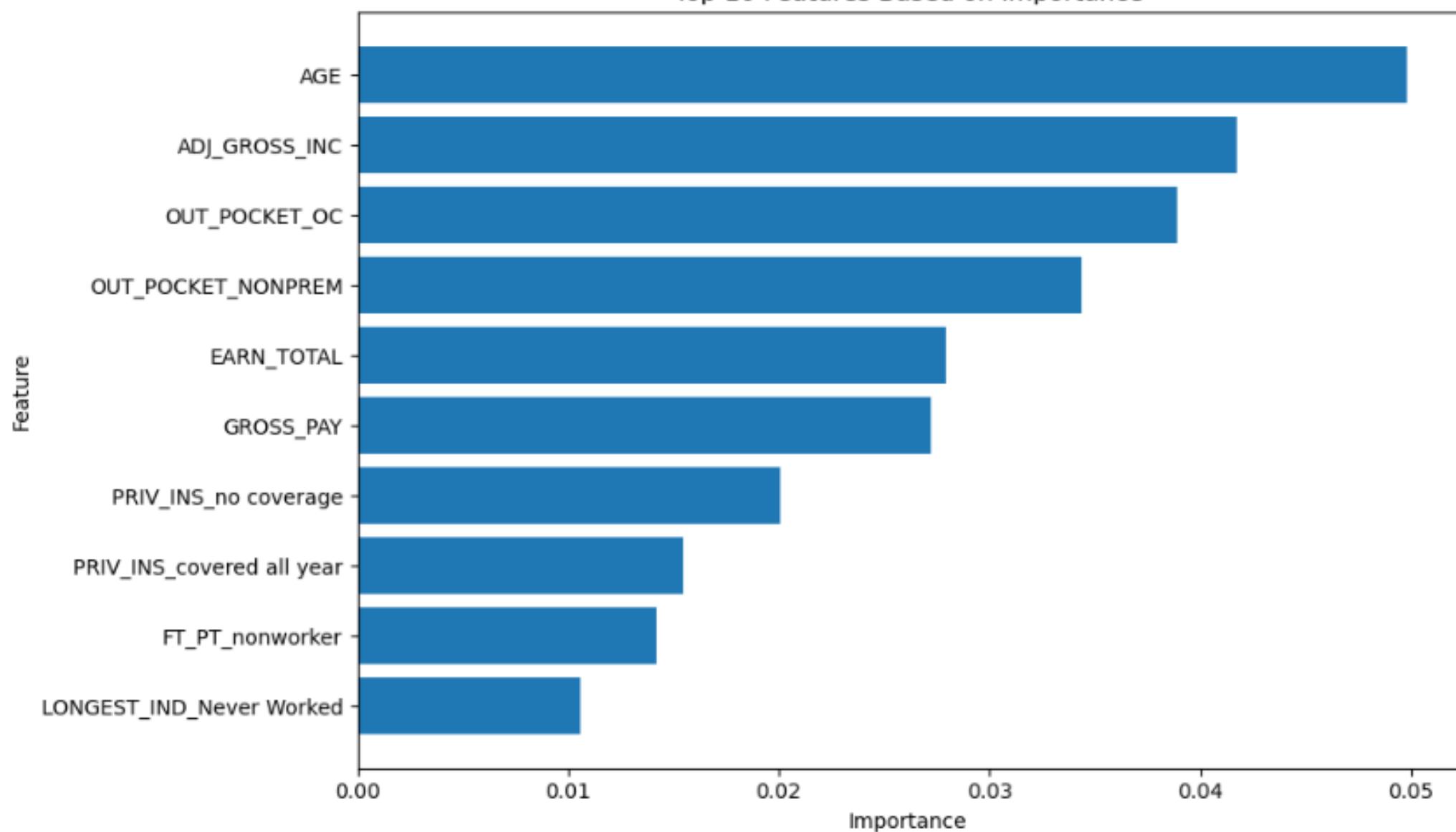
GRADIENT BOOSTING



R A N D O M F O R E S T



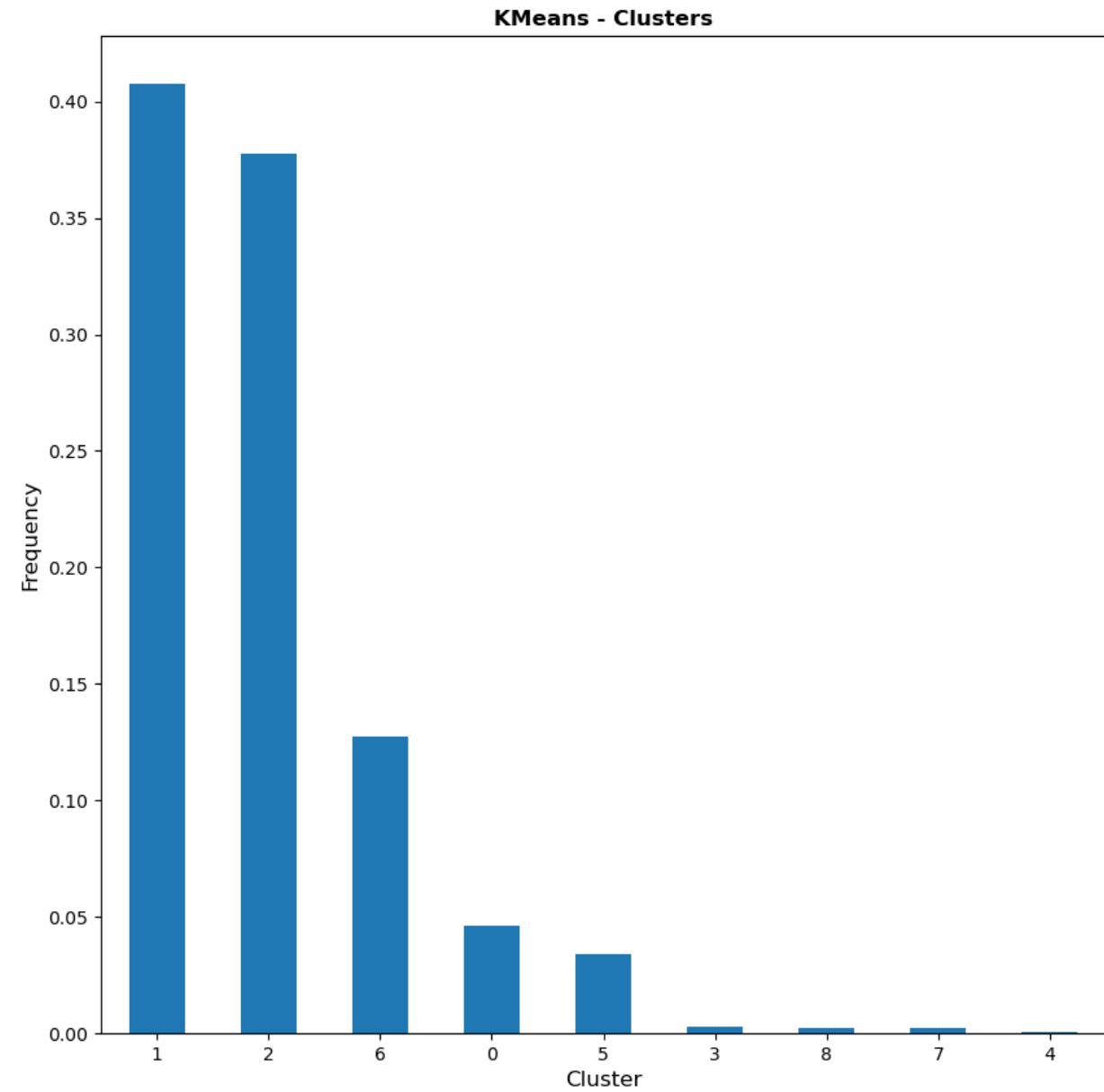
Top 10 Features Based on Importance



Model	Balanced Accuracy	F-1 Score
Logistic Regression	0.547	0.171
Logistic Regression with pipeline (no penalty)	0.547	0.173
GBM W/ Data resampling (all columns)	0.787	0.326
GBM W/ Data Resampling (with only columns selected at EDA process)	0.787	0.325
Random Forest	0.765	0.613
Random Forest w/ Resampling	0.815	0.631

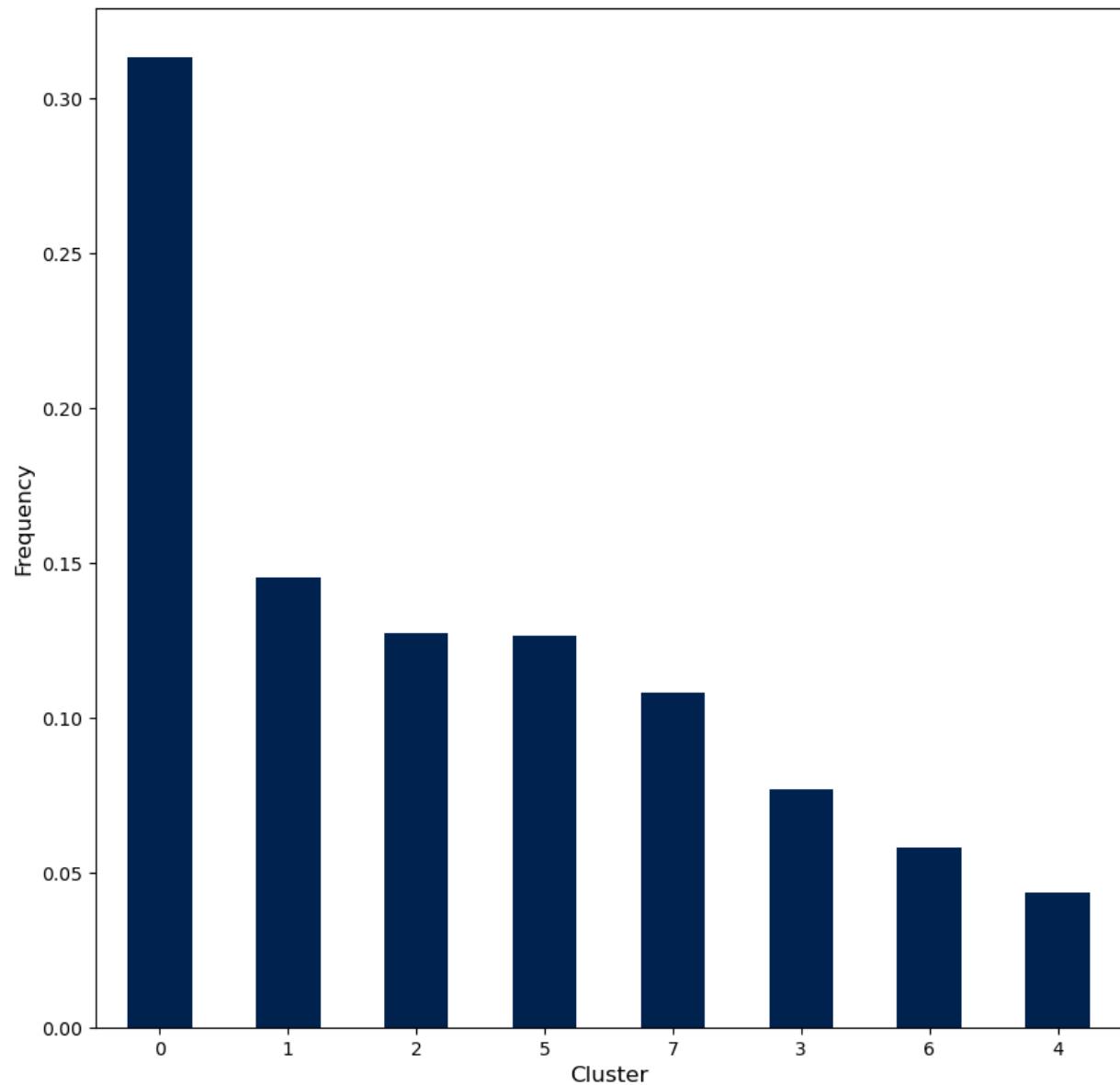
SUMMARY

CLUSTERING



K MODES CLUSTER

KModes - Clusters





CONCLUSION

- Demographic factors
- Education level
- Health insurance coverage
- Tax filing status
- Age