**Bias and Drift: Responsible Use of Synthetic Data in Financial Decision-Making for Credit Scores**

Race Groups:

* 1 - American Indian or Alaska Native
* 2 - Asian
* 21 - Asian Indian
* 22 - Chinese
* 23 - Filipino
* 24 - Japanese
* 25 - Korean
* 26 - Vietnamese
* 27 - Other Asian
* 3 - Black or African American
* 4 - Native Hawaiian or Other Pacific Islander
* 41 - Native Hawaiian
* 42 - Guamanian or Chamorro
* 43 - Samoan
* 44 - Other Pacific Islander
* 5 - White
* 6 - Information not provided by applicant in mail, internet, or telephone application
* 7 - Not applicable

Additional subset: Gender (Female, Male, Joint, NAN)

* 1 - Male
* 2 - Female
* 3 - Information not provided by applicant in mail, internet, or telephone application
* 4 - Not applicable
* 6 - Applicant selected both male and female

Hispanic or Latino subgroupding:

* 1 - Hispanic or Latino
* 11 - Mexican
* 12 - Puerto Rican
* 13 - Cuban
* 14 - Other Hispanic or Latino
* 2 - Not Hispanic or Latino
* 3 - Information not provided by applicant in mail, internet, or telephone application
* 4 - Not applicable

**Columns of Interest**

Income

Debt to income ratio

Interest\_rate

Denial reason

Loan amount (see if there are differences in the loan amounts people are applying for between groups)

Loan to equity ratio ( -||- )

Rate\_spread (are certai groups being offered worse spreads?)

**Model decay**

Regression to the mean: the toy model Alan built displays a general convergence onto a single output when data is refed into the algorithm. With credit, we have a lot of groups, each with their own means (% approved). Will we see a disproportionate decline (or increase) in means for certain groups over others?

**Next Steps**

Add Approved Column (simple binary 0,1) for easy mean calculations

^^Calculate Means for each group (decide which groups are of interest, and cut dataset down to that, as there may be too many to work with effectively)