**Title (tentative):** Analysis of Cohort Default Rates in Post-Secondary Education in the United States:

**Subtitle (tentative):** Comparing Private vs. Public Institutions in Counties with Large Minority Populations

**Authors:** Dillon Orr, Marketne Noel, Prince Gyamfi, John Vivian, PhD., Ebrahim Tarshizi, PhD., and Michael Kirkpatrick

**Summary**

On September 25th, 2019, the US Department of Education announced that the national cohort default rate from the 2016 fiscal year decreased to 10.1% from 10.8% in 2015. (US Department of Education, 2019) A decrease in cohort default rate was seen for public, private, and proprietary institutions. However, the smallest decrease in default rate was seen in proprietary (private, for-profit) institutions. Out of 15 schools that were sanctioned by the US Department of Education due to having a cohort default rate higher than 30%, 13 were proprietary colleges. We propose that private, for-profit institutions operate more often in areas with larger minority (non-white) populations in order to generate larger enrollment figures.

**Objectives:**

1. Compare default rate of proprietary institutions in US Counties with large minority populations with default rate of public and private non-profit colleges. We expect to see a larger default rate in proprietary institutions in counties with larger minority populations.
2. Compare proprietary institution density in areas with larger minority populations. We expect to see more private colleges in areas with more non-white residents
3. Create a regression model that can predict default rate. (National Average (10.1%))
4. Classify school as private, proprietary, public

**Methods:**

1. Confirm average default rate between three classes for 2016 FY Cohort
2. Group census data by county
3. Binarize ethnicity data to white and non-white
4. Merge datasets on county
5. Binarize default rate on below or above national average (10.1%)
6. Compare county density of non-white population with default rate of institutions using grouping and heatmapping
7. Compare density of non-white population with density of colleges using grouping and heatmapping
8. Predictive model to determine whether a school is at high risk of having a high default rate based on minority population in school location

**Questions:**

1. We are considering limiting the scope of our project to consider a certain geographical area, as opposed to the entire US, i.e. California, the West Coast, or the South. How would this affect the conclusions drawn by the project?
2. What kind of predictive model would be the best fit for our analysis?

**Data Sources:** [Official Cohort Default Rates for Schools](https://www2.ed.gov/offices/OSFAP/defaultmanagement/cdr.html) (2019), [Census Race Data](https://www.census.gov/data/datasets/2010/demo/popest/modified-race-data-2010.html) (2010)

**Areas for Future Research:** 2020 Census Day is April 1, at which point we will look to compare 2010 figures regarding race data to the 2020 figures

**Team Contact Information**

Dillon Orr – [orr.dillon@gmail.com](mailto:orr.dillon@gmail.com)

Marketne Noel – [marketnenoel@gmail.com](mailto:marketnenoel@gmail.com)

Prince Gyamfi -

John Vivian, PhD. – [jtvivian@gmail.com](mailto:jtvivian@gmail.com)

Ebrahim Tarshizi, PhD. – [Ebrahim.KarimiTarshizi@natuniv.edu](mailto:Ebrahim.KarimiTarshizi@natuniv.edu)

Michael Kirkpatrick – [Michael.Kirkpatrick@natuniv.edu](mailto:Michael.Kirkpatrick@natuniv.edu)

**References**

Guzman (2019) Household Income: 2018. *American Community Survey Briefs*. Retrieved from <https://www.census.gov/content/dam/Census/library/publications/2019/acs/acsbr18-01.pdf>

Federal Student Aid (2019) Official Default Rates for Schools. *Federal Student Aid Default Management*. Retrieved from <https://www2.ed.gov/offices/OSFAP/defaultmanagement/cdr.html>

Federal Student Aid (2019) Official Cohort Default Rates for Schools. *Federal Student Aid Default Management*. Retrieved from <https://www2.ed.gov/offices/OSFAP/defaultmanagement/cdr.html>

US Census Bureau (2010) Modified Race Data.  *2010 US Census.* Retrieved from <https://www.census.gov/data/datasets/2010/demo/popest/modified-race-data-2010.html>

US Department of Education (2019) National Federal Student Loan Cohort Default Rate Continues to Decline. *US Department of Education Press Releases.* Retrieved from <https://www.ed.gov/news/press-releases/national-federal-student-loan-cohort-default-rate-continues-decline>

Notes from john

* Can’t prove intent, so we don’t know that the colleges are being malicious
* Perform linear regression with default rate being the y value
  + Multiple linear regression with dummy variables
* Continuous variable like minority percentage for regression and x value being default rate
* Bar chart with 3 groups, types of schools
* Hollowviews for package geographic heatmaps, counties
  + Heatmap for race data, college density
* Feature engineering (matrix) for school size, years active, etc.
* Bayesian linear regression to determine distribution of coefficient of feature
  + Do public/private/proprietary labels carry a lot of weight as coefficients?
* Divide control (average) by percentage that we get to determine significance, confidence interval cannot contain 1
* Government initiatives that schools have participated in
* Variance of default rate between proprietary schools, basic summary statistics are important in our case
* Sankey diagram
* Put datasets together, ensure data is tidy