Marriageability

Maria Hernandez | Makafui Kalefe | Stephane R. Miller |
Molly Morrison | Lourdes Pena

September 2019
Georgetown University
School of Continuing Studies
Data Science Certificate Capstone

Presentation Agenda

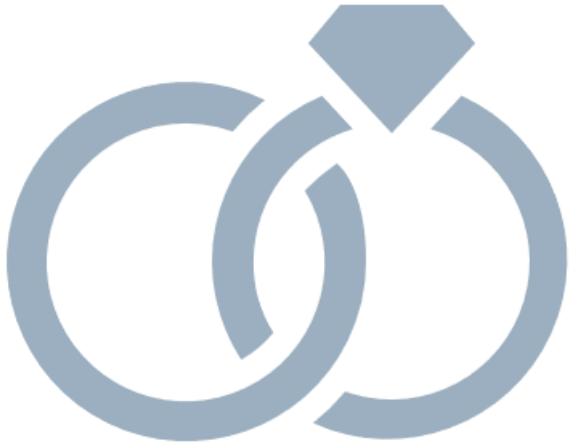
- Introduction
- Project Pipeline
- Ingestion & Wrangling
- Modeling
- Insights

Introduction

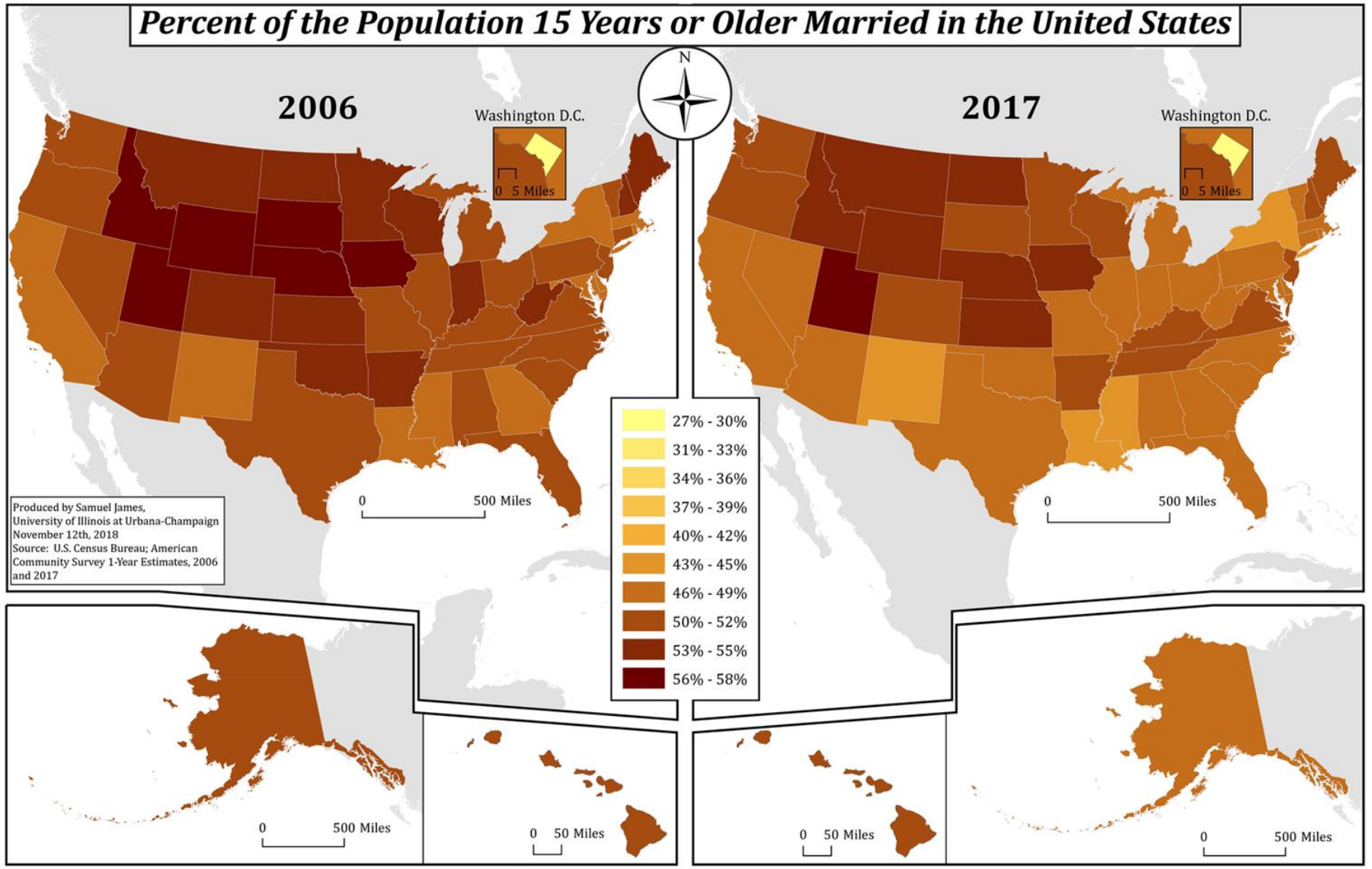
Project Overview

- **Objective:** To predict *marriageability* or whether a person will get married
- **Research Question:**
  - Can we predict *marriageability* based on key demographics, economic factors, and geographic location?

What is *Marriageability?*



- The likelihood that a person will be married, considering demographics, economic, and socio-cultural factors.



In the US, the proportion of the married population has decreased over time...

48.2% of the U.S. population is married

36.3% of men have never been married

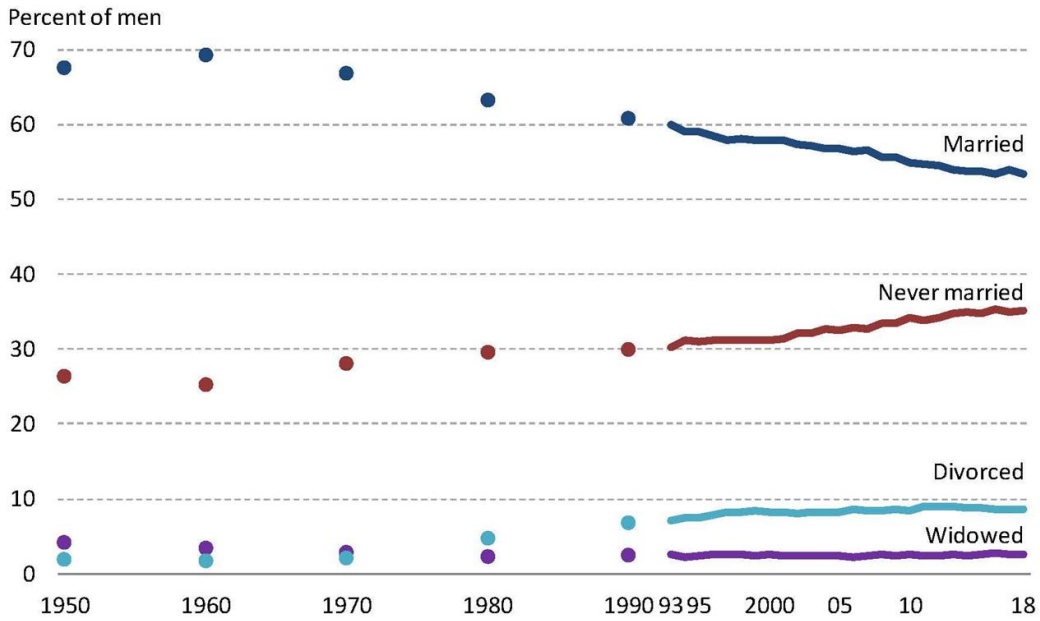
30.1% of women have never been married

Source: Latest ACS 5-Year Estimates Data Profiles/Social Characteristics

Marriage in the United States

Marriage is declining for men and women

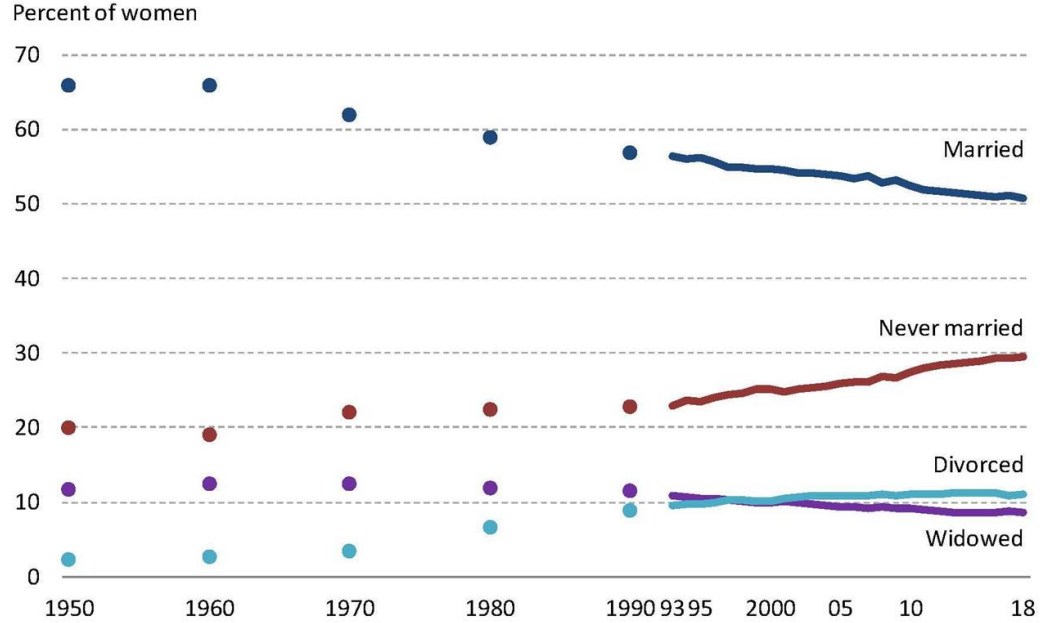
Men's marital status



United States Census Bureau | U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Source: U.S. Census Bureau, Decennial Censuses, 1950 to 1990, and Current Population Survey, Annual Social and Economic Supplements, 1993 to 2018.
Note: Married includes separated and married spouse absent.

Women's marital status



United States Census Bureau | U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

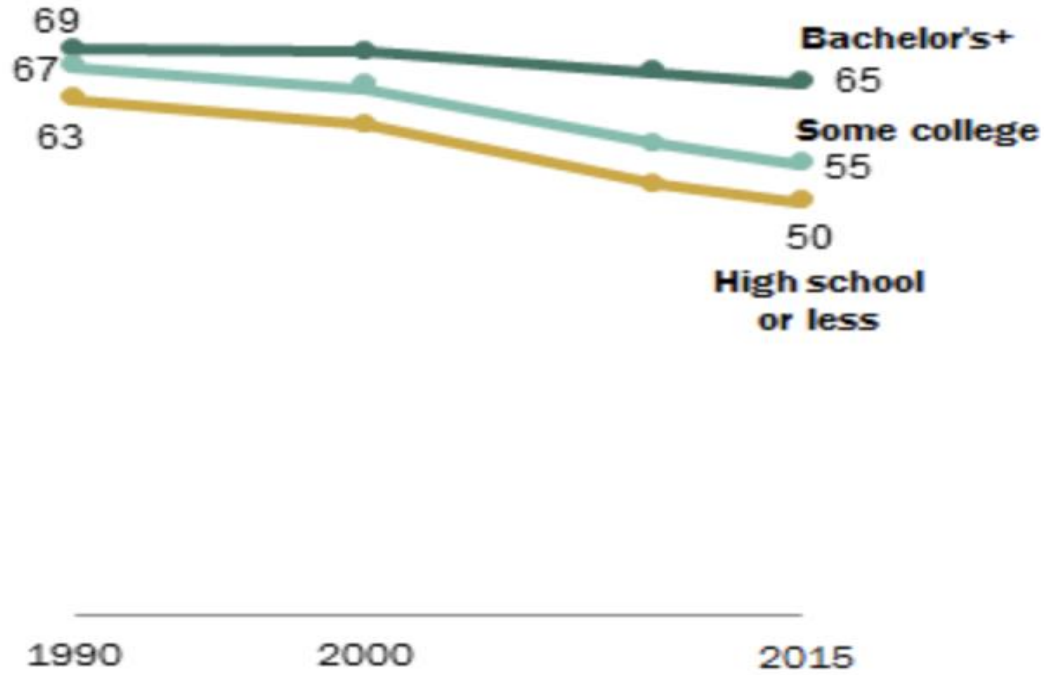
Source: U.S. Census Bureau, Decennial Censuses, 1950 to 1990, and Current Population Survey, Annual Social and Economic Supplements, 1993 to 2018.
Note: Married includes separated and married spouse absent.

The median age at first marriage is increasing

Men: 29.8 years
Women: 27.8 years

The education gap in marriage continues to grow

% of U.S. adults ages 25 and older who are married, by education

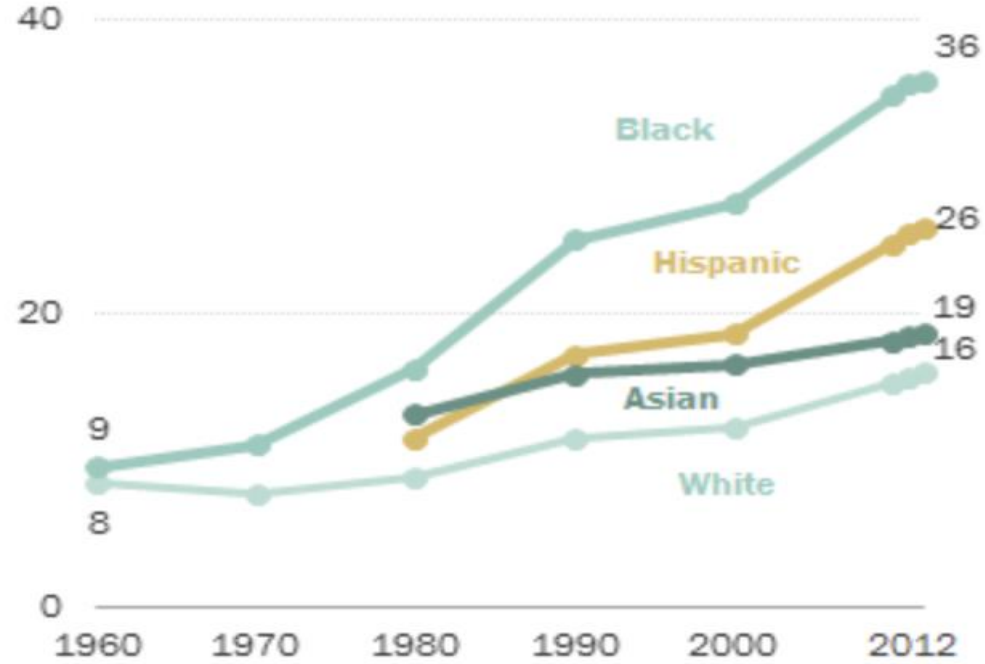


Note: "Some college" includes those with an associate degree and those who attended college but did not obtain a degree. Adults who are separated are not classified as married.
Source: Pew Research Center analysis of 1990-2000 decennial censuses and the 2010 and 2015 American Community Surveys (IPUMS).

PEW RESEARCH CENTER

Rising Share of Never-Married Adults, Growing Race Gap

% of adults ages 25 and older who have never been married

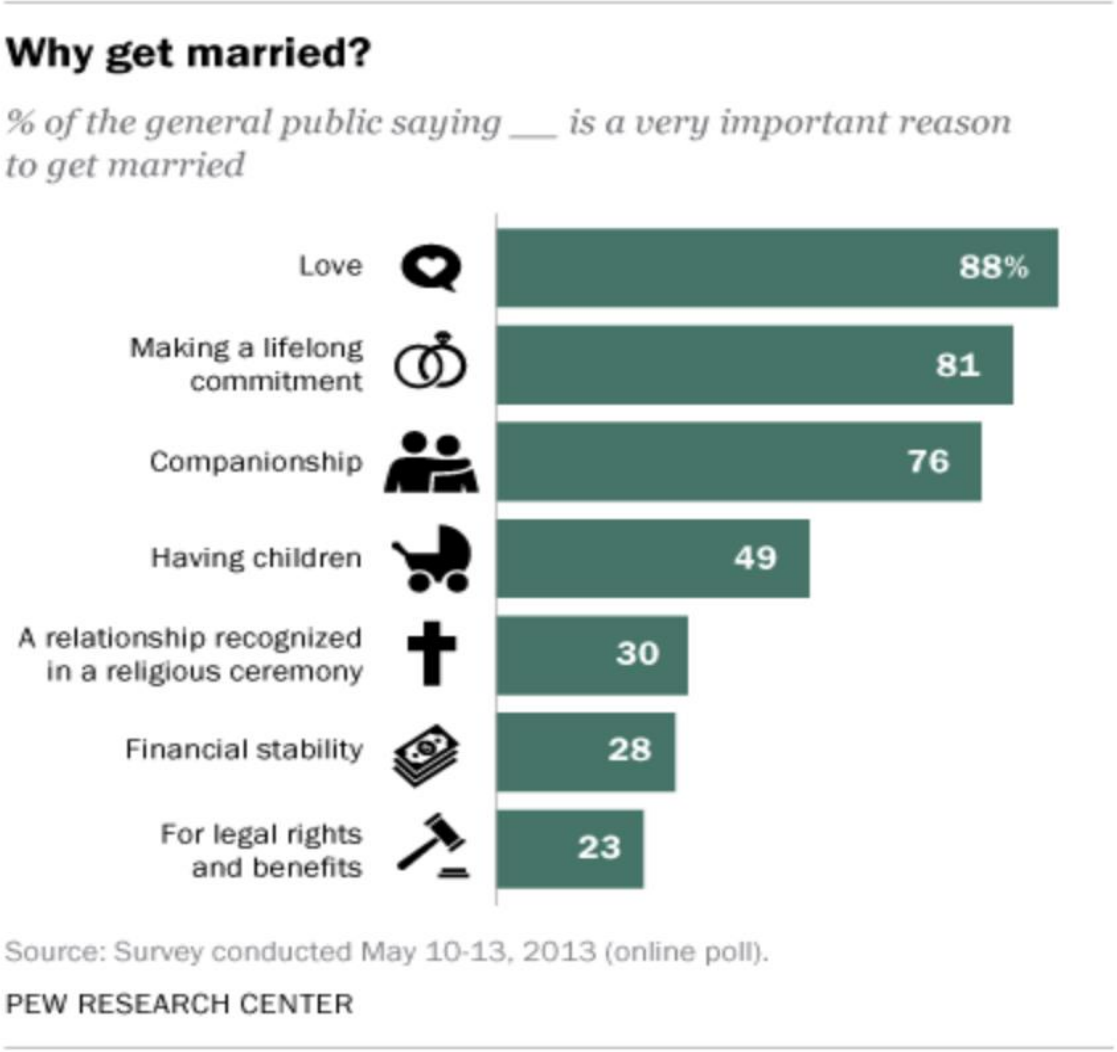


Note: Data on Hispanics and Asians prior to 1980 are not plotted given the small sample sizes.

Source: Pew Research Center analysis of the 1960-2000 decennial census and 2010-2012 American Community Survey, Integrated Public Use Microdata Series (IPUMS)

PEW RESEARCH CENTER

Reasons to get married?



Economic

Location

Social-ability

Education

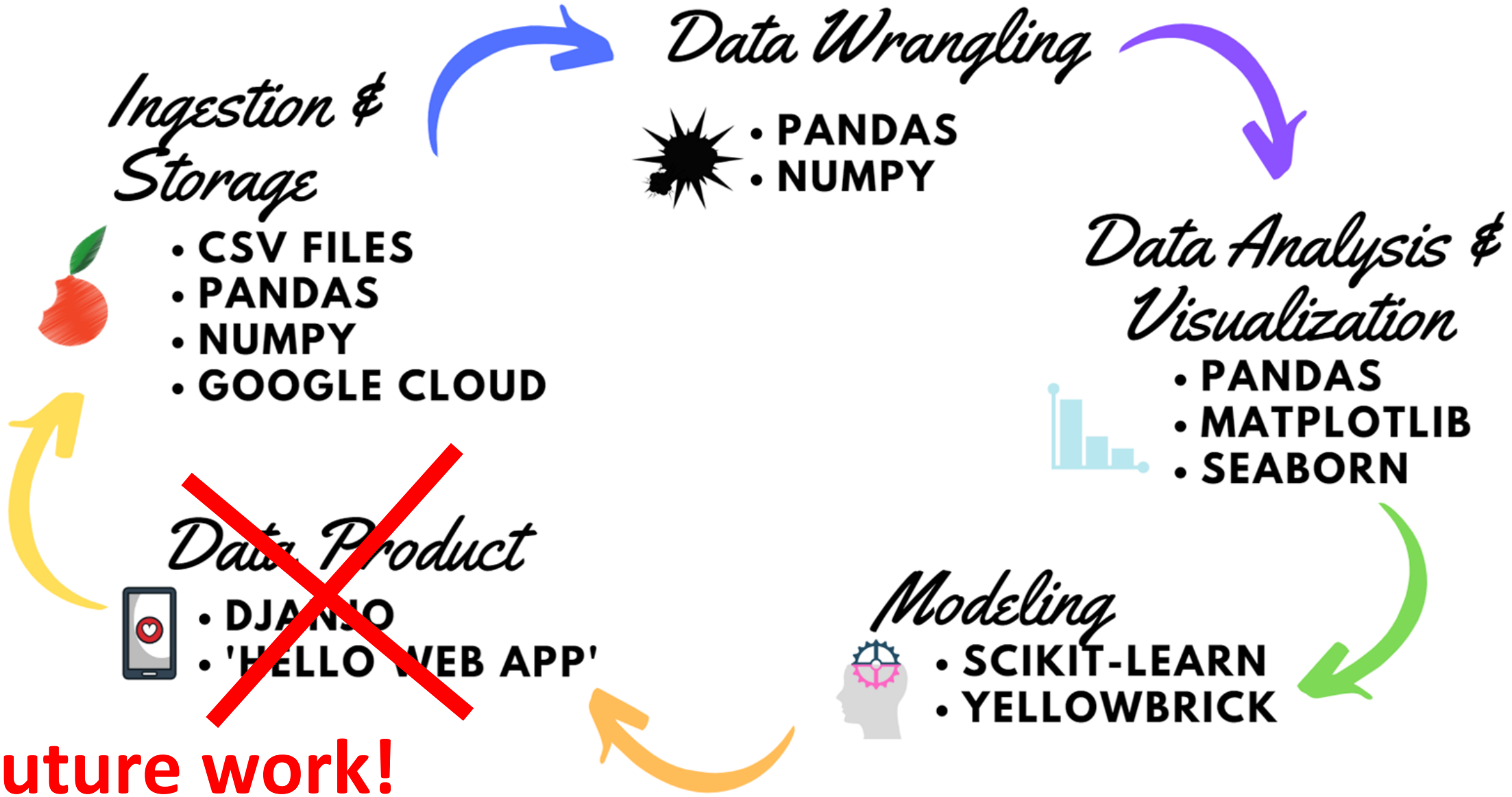
Demographics

Project Limitations & Challenges

- **DATA**: Racial bias (70% of instances are White)
- **TECHNICAL**: Prohibitive costs (processing time, hardware)
- **CONCEPTUAL**: No info on sexual orientation, common law marriage, long-term cohabitating relationships

Project Pipeline

Project Pipeline



Ingestion & Wrangling

Ingestion

- **Source**: U.S. Census American Community Survey (2017 1-year)
- ACS 1-year Public Use Microdata Samples (PUMS) files contain data on ~1% of the U.S. population
- Data downloaded as .csv files from Census American Fact Finder
- **Person Files**: each record represents a single person, organized into households
- **Household Files**: each record represents a single housing unit
- **Topics**: marital status, income, internet access, occupation, health insurance, type of commute, hours worked, etc.

Wrangling

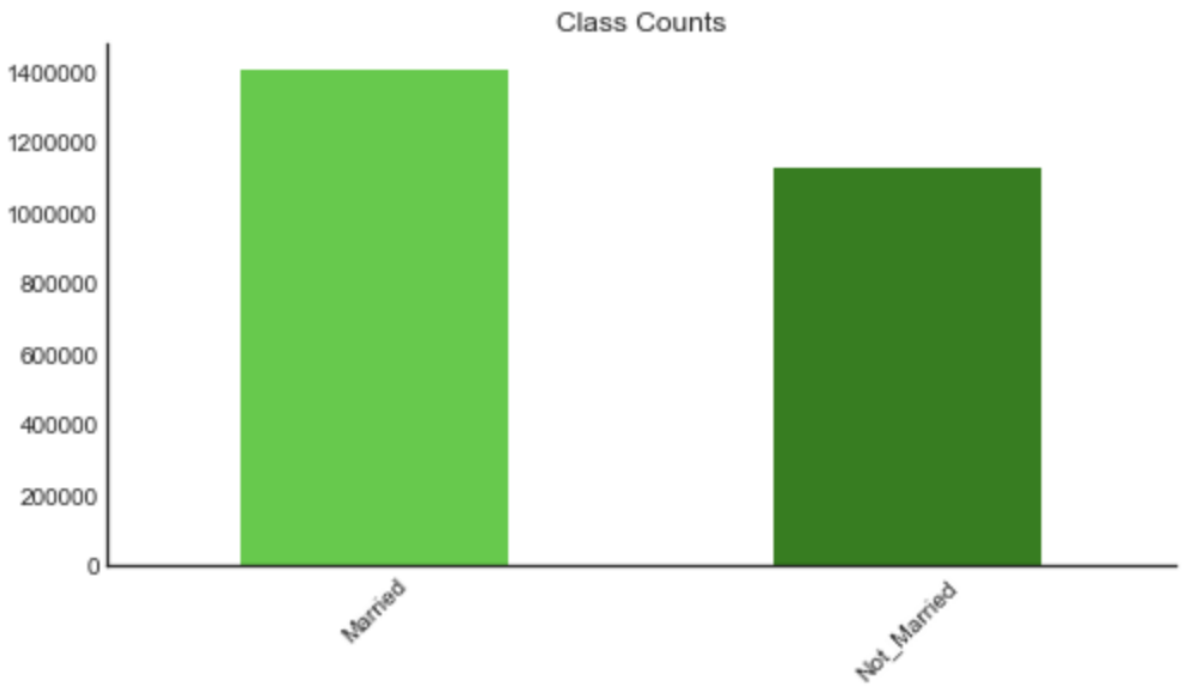
- Data required minimal wrangling
- Missing data was not an issue
- Converted N/A into 0s
- Dropped individuals who were less than 18 years of age

Target: *Marital Status*

- Binary classifier (0/1)
- Married (1) = married, separated
- Not Married (0) = single, never married, divorced and widowed

Not concerned with target class imbalance...

Married: 1404644
Not_Married: 1126082
Proportion: 1.25 : 1

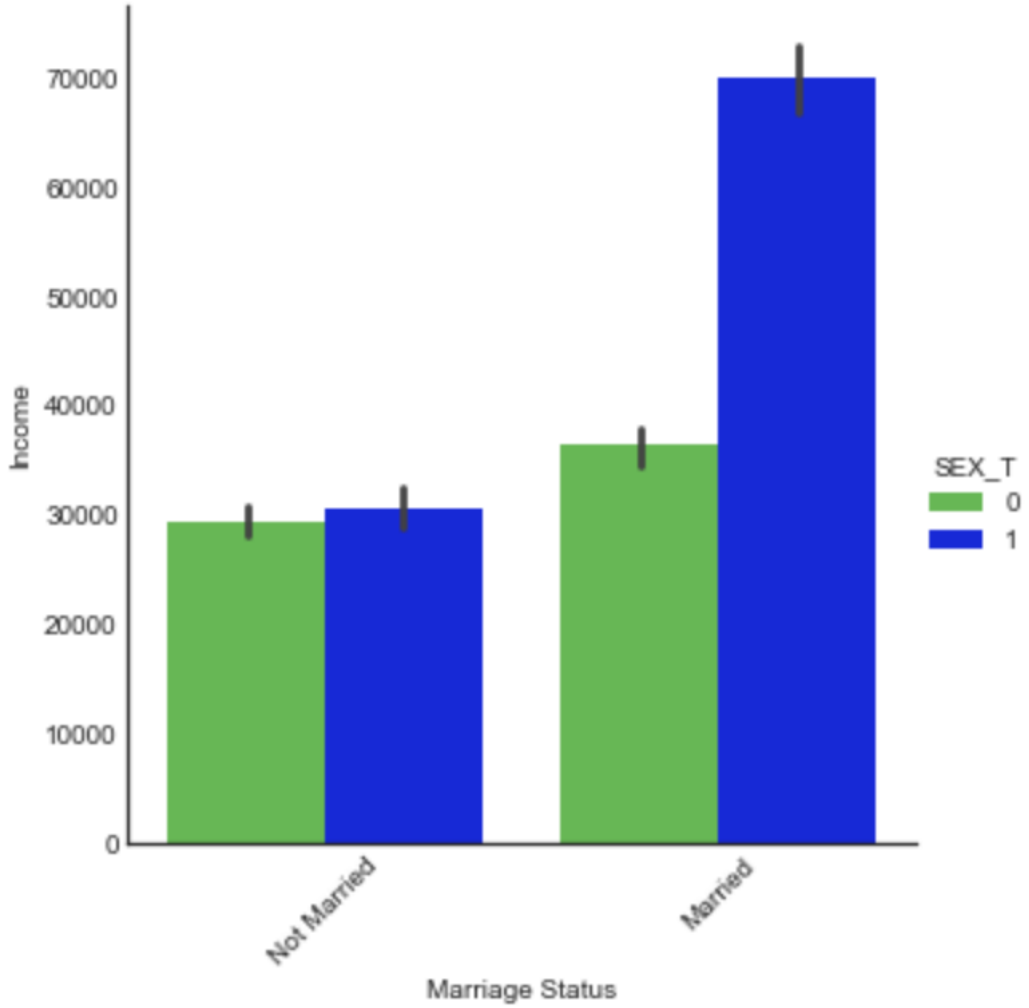


Selected features with expert knowledge

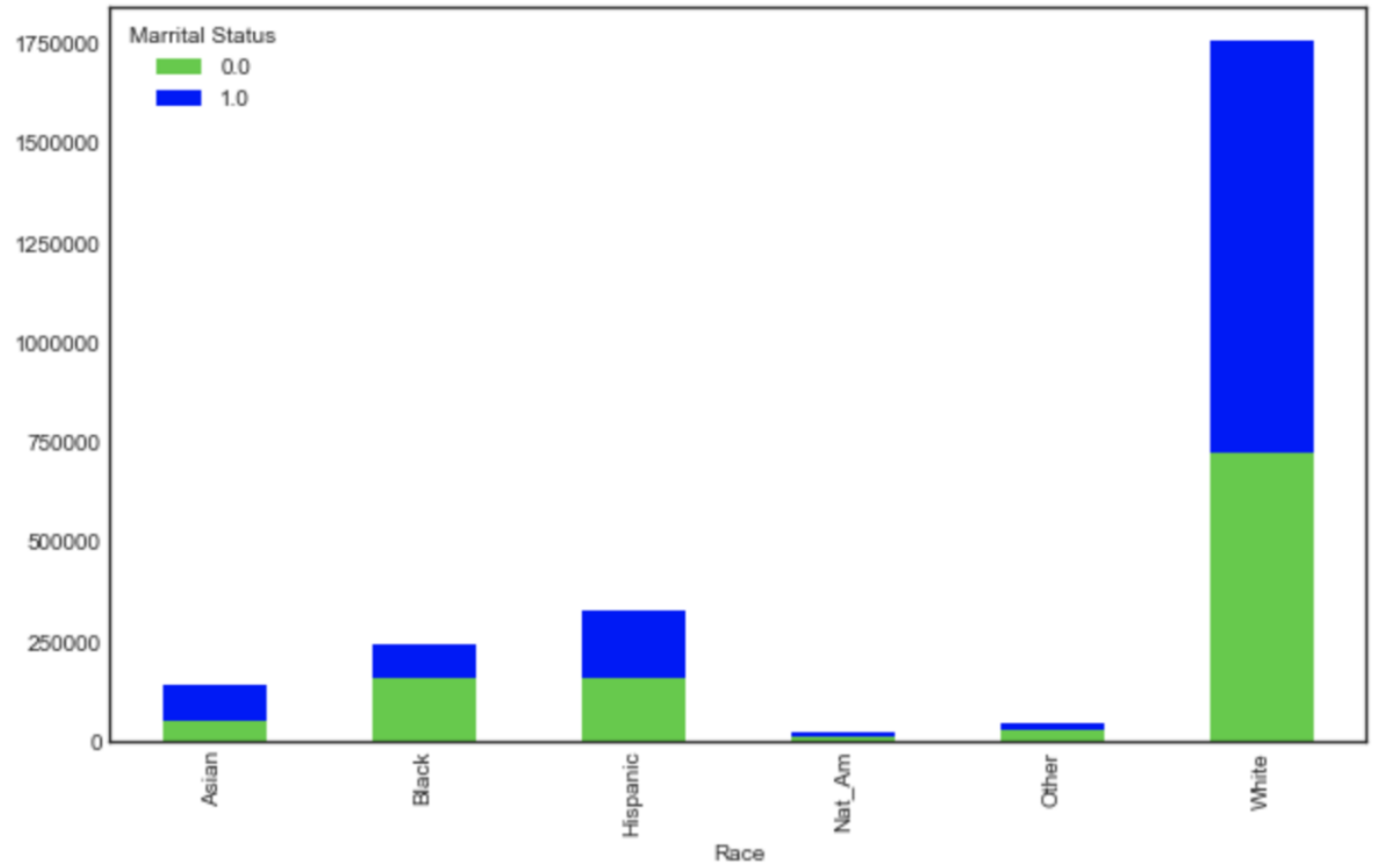
ONE-HOT ENCODING	INTEGER ENCODING
Citizenship	Educational Attainment
Transportation to Work	Age bins (6 categories)
Mobility Status	Income bins (8 categories)
Sex	
Race (6 categories)	
Occupation Sector	
Disability	
Presence of Children	
English speaker	
Location (State and Tri_State)	

Distribution and bias...

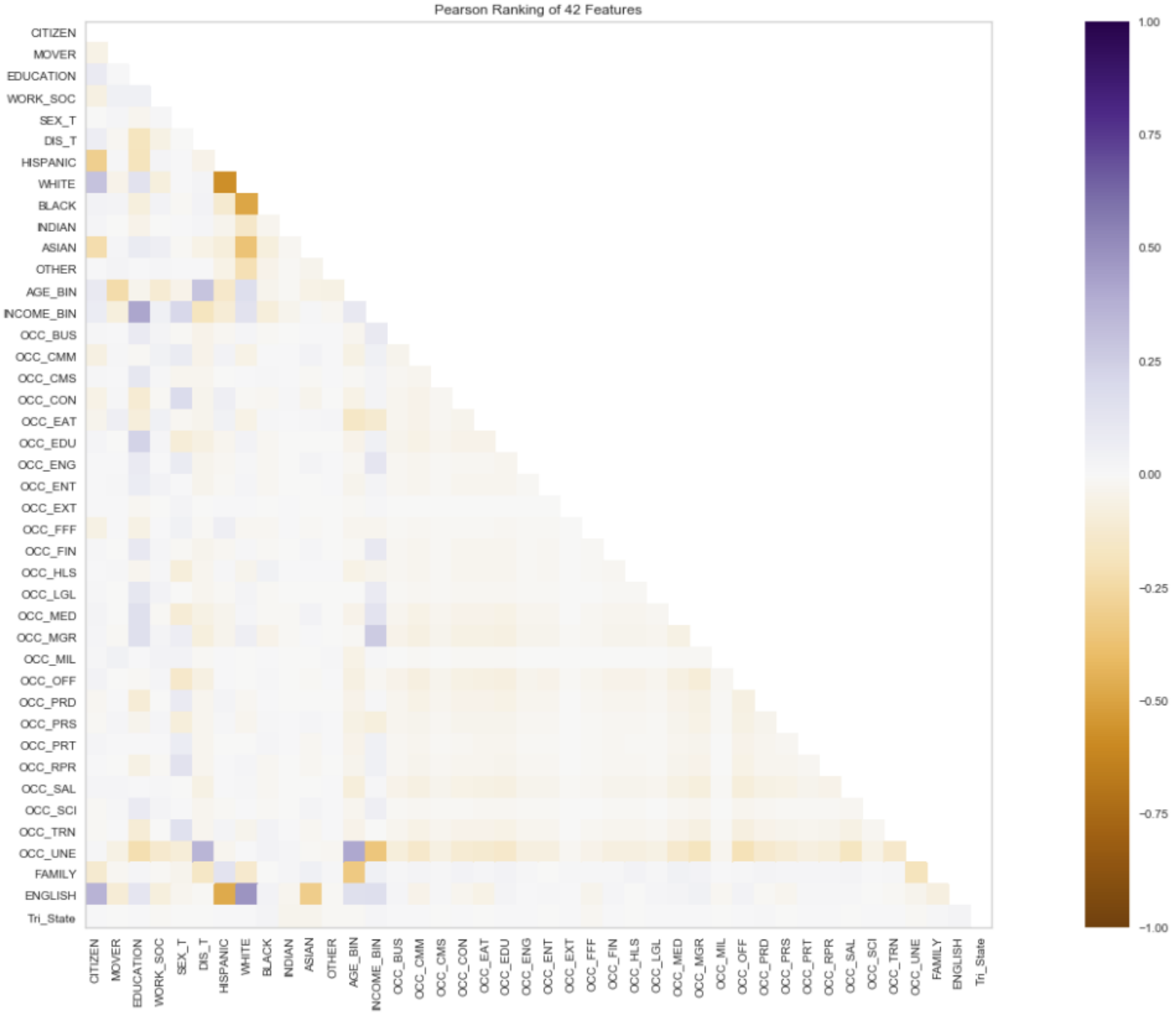
Marriage by Income and Sex



Marriage by Race



No concerns of collinear relationships...

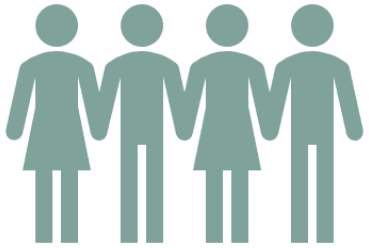


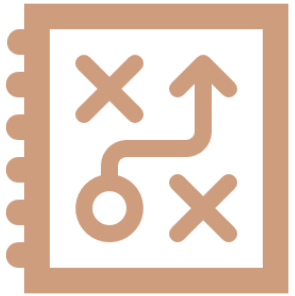
Location Features

- Originally intended to use Census PUMAs (geographic areas), but there are 2,378. Costly!
- One-hot encoded State
- Derived location feature
  - *Tri_State* – denote people who live in states connected by economy and geography

    Includes: CT, DE, DC, IL, IN, KY, MD, NJ, NY, OH, PA, VA, WV

Final Analytic Data Set





2,530,726 instances

93 features

Modeling

Supervised Machine Learning: Classification

- Logistic Regression
- K-Nearest Neighbors
- Random Forest
- Gaussian Naïve Bayes
- Gradient Boost

Used default parameters.

Preliminary Results

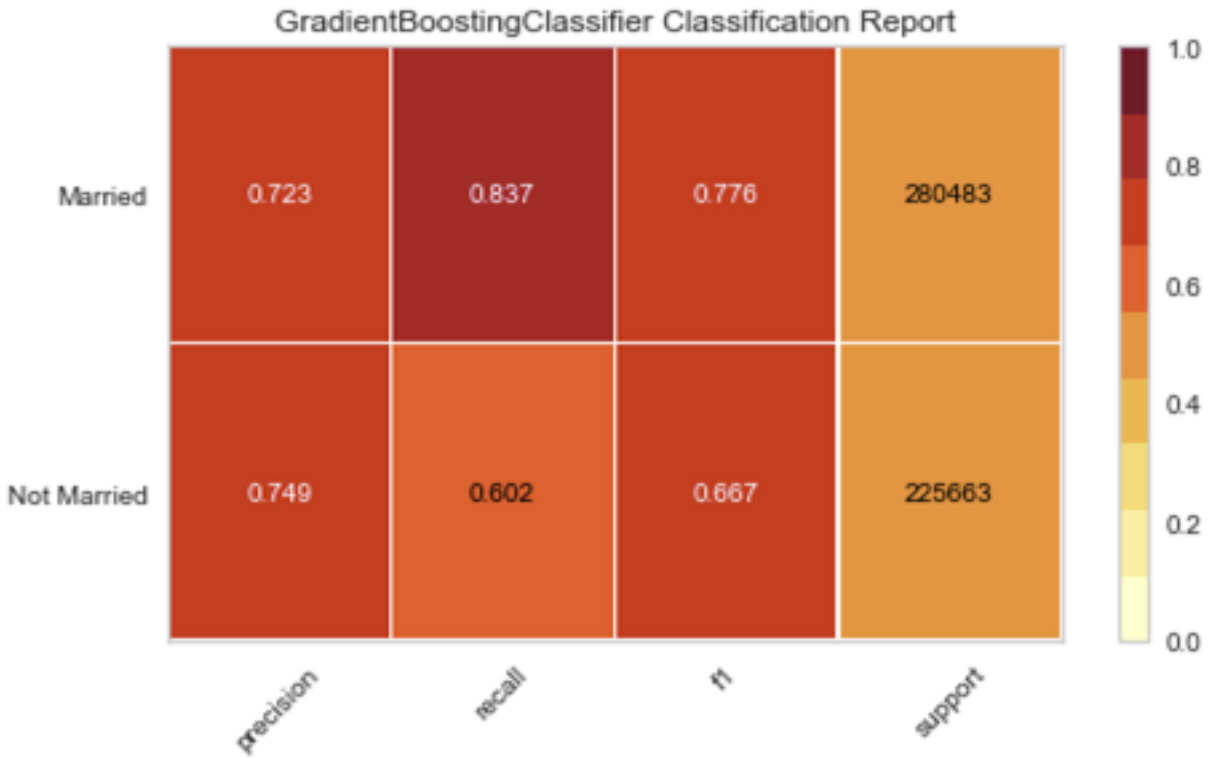
Classification Algorithm	f1	Precision*	Recall*	Average CV Score
Logistic Regression	0.728	0.682	0.781	0.677
K-Nearest Neighbor	0.000	0.000	0.000	0**
Gaussian Naïve Bayes	0.650	0.660	0.640	0.618
Gradient Boost	0.766	0.696	0.852	0.713
Random Forest	0.725	0.624	0.956	0.656***

*Precision and Recall are reported for "Married = 1"

**For the cross-validation step, k-folds = 6 for all algorithms except for K-Nearest Neighbor, k-folds = 3.

***The observed scores for Random Forest showed a relatively high variance in the accuracy between folds, ranging from 0.628 to 0.678

Classification reports for top performing models



Of those labelled married, ~68% actually were.

Model found ~78% of the married instances in the dataset.

Model Selection and Evaluation

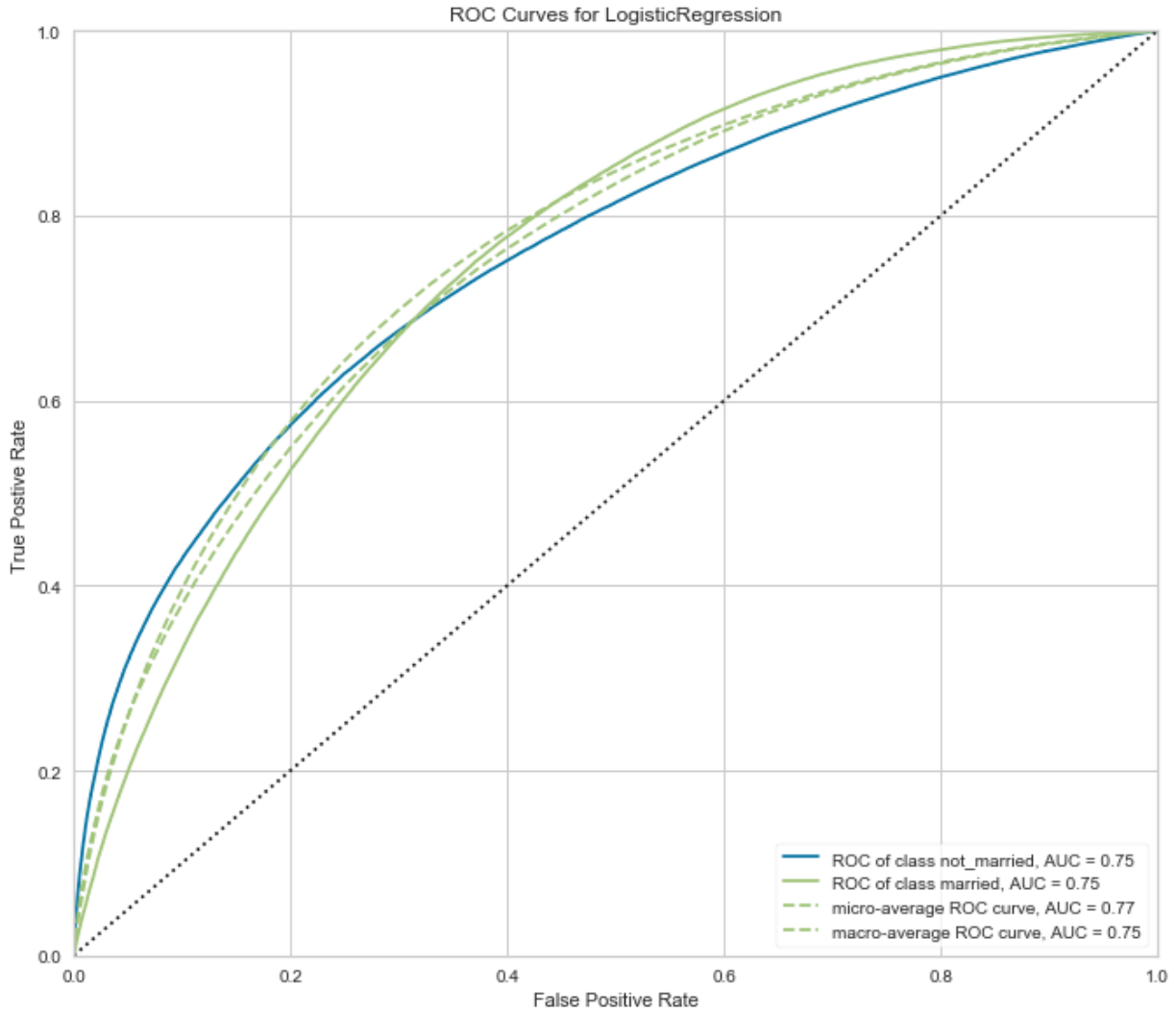
- Five Classifier Models
  - KNN did not run
  - SVM with NuSVC Classifier
    - ***Decided to move on!***
- Model Evaluation & Hyperparameter Tuning
  - Test/Train Split
  - 6-fold cross-validation
  - GridSearchCV
- The highest F1 Score → 0.766 (Gradient Boost)
- But Gradient Boost was **too costly** to run with as large of a dataset
- **So the team focused on Logistic Regression**:
  - F1 Score → 0.728

Final Selection: Logistic Regression

Features: 93

C: .001
penalty: l2

f1 score: 0.730



Application: Using LR Model to Make Predictions

Probability output from LR and GBC

Base Scenario

	LR		GBC	
	Not Married	Married	Not Married	Married
Lulu	[0.38254617,	0.61745383]	[0.35778459,	0.64221541]
Stephanie	[0.53220923,	0.46779077]	[0.58407321,	0.41592679]
Maria	[0.27268135,	0.72731865]	[0.34414361,	0.65585639]
Molly	[0.50298999,	0.49701001]	[0.78000214,	0.21999786]
Makafui	[0.56450502,	0.43549498]	[0.47930086,	0.52069914]

PREDICTIONS	Baseline	Changing Race		10 Years Older		Location	
Lulu	Married	Indian	Married	40 - 49	Married	TN	Married
Stephanie	Not Married	White	Married	40 - 49	Married	NY	Not Married
Maria	Married	White	Married	40 - 49	Married	TX	Married
Molly	Not Married	Black	Not Married	30 -39	Married	CO	Not Married
Makafui	Not Married	Asian	Married	40 - 49	Married	CA	Not Married

Insights & Future Work

Areas of improvement

Data:

- More representative and inclusive dataset
- Broader definition of relationships
- Explore additional features - use of dating apps, urban vs. rural, social-ness of a city
- 5-year data file to reduce bias
- Use PUMAs instead of state for location; finer level of analysis

Data product: develop app using Django

- make recommendations of alternative cities
- allow users to input preferences/rank features based on importance

Questions