



# Capstone: Women, Labor, & Education

Brandie Hatch, Data Scientist  
U.S. Census

# Overview

The U.S. Census Bureau definition of sex is based on the biological attributes of men and women:

- chromosomes
- anatomy
- hormones

Females account for **50.5%** of the U.S. 331,449,281 total population estimates as collected in the April 1, 2020, Census.



Women working at the U.S. Capitol switchboard, Washington, D.C. (Library of Congress)

# Introduction

## College-educated Workforce:

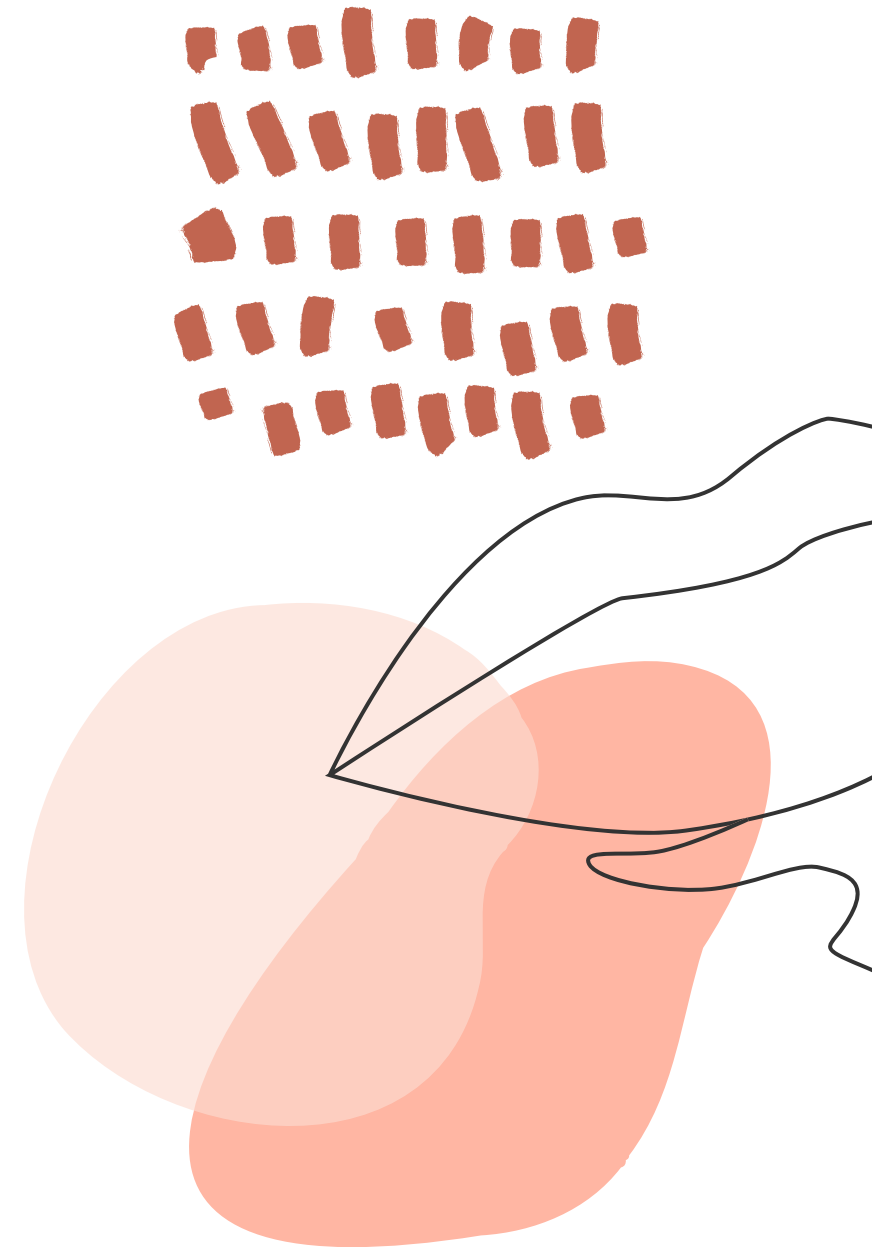
- According to a 2019 study from Pew Research, which analyzed data from the U.S. Bureau of Labor Statistics, women 25 and older now make up 50.2% of the college-educated work force.
- Women earn about 57% of bachelor's degrees (2019)

## In STEM Occupations → Women account for:

- 25% of college-educated workers in computer jobs
- 15% of college-educated workers in engineering jobs
- Majority of college-educated workers in office, administrative support, and health care practitioners or technicians

## And yet....

Women make 84% of what men earn

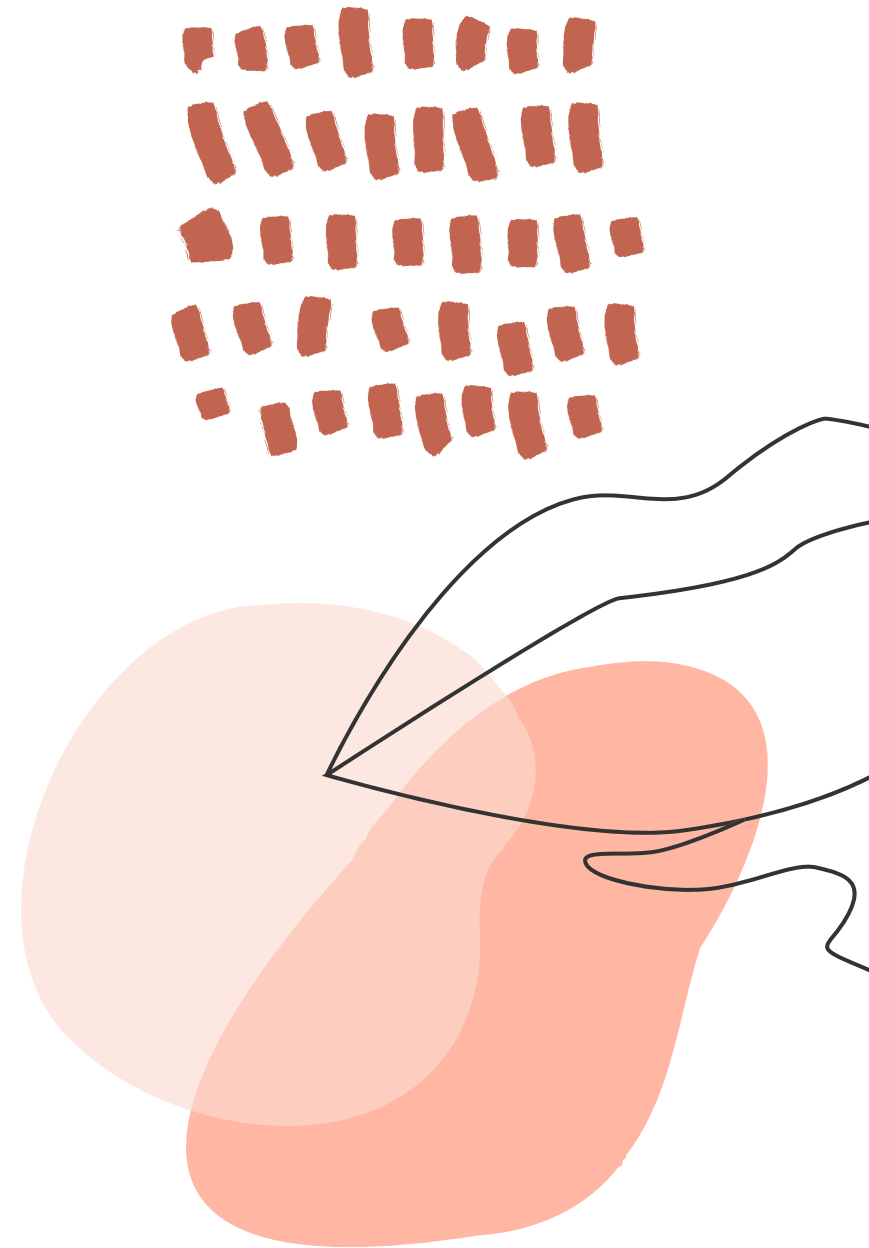


# Introduction

## **Significance of Women and Girl's Participation in Education**

Regain \$15-30 trillion in lost lifetime productivity and earnings

- Child marriage rates decline
- Child mortality rates fall
- Maternal mortality rates fall
- National growth rates rise
- Female earnings dramatically increase



# Question, Problem, and Goal

**Topic:** Gender Gaps in Labor & Education, United States

**Question:** As females attain higher levels of education (literacy), how does women's participation in labor change?

**Problem Statement (Hypothesis):**

The average difference in the years of education attained between gender=1, gender=2 is zero.

The average difference in the years of education attained between gender=1, gender=2 is not zero.

Gender 1 = male | Gender 2 = female

**Goal:** Create a model that predicts whether the survey respondent is male, or female based on labor market, income, and education features.



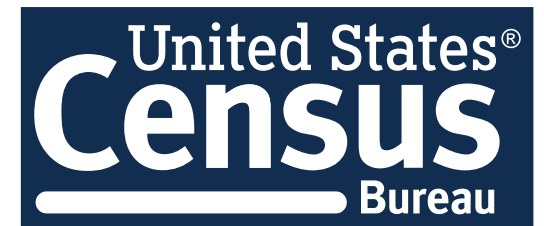
# Data Introduction

Collected from the U.S. Census' American Community Survey (ACS) Public Use Microdata Sample (PUMS) files.

"The PUMS files allow users to create estimates for user-defined characteristics. The files contain a sample of responses to the ACS. **The PUMS files include variables for nearly every question on the ACS survey.** Additional variables are also created from other recoded PUMS variables to provide data users with useful derived variables (such as poverty status) while protecting confidentiality and providing consistency within the PUMS files."

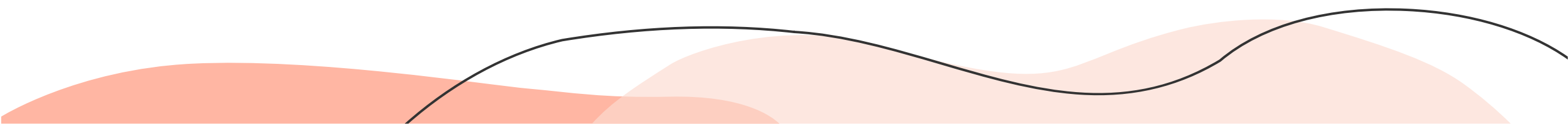
AMERICAN COMMUNITY SURVEY 2016-2020 5-YEAR PUMS

Prepared by American Community Survey Office, U.S. Census Bureau March 31, 2022



# Data Collection

The data came from using the Census API in a Jupyter Notebook to pull a json file. The initial data set included 3,239,553 observations and 26 variables.

- Age: 1 to 99 years
  - Citizenship status
  - Class of worker
  - Ability to speak English
  - Employment status recode
  - Yearly food stamp/Supplemental Nutrition Assistance Program (SNAP) recipient
  - Language other than English spoken at home
  - Marital Status
  - Multigenerational household
  - Number of own children in household (unweighted)
  - Occupation recode for 2018 and later based on 2018 OCC codes
  - Total person's income
  - Income-to-poverty ratio recode
  - Recoded detailed race code
  - School Enrollment
  - Grade Level Attending
  - Educational Attainment
  - Field of degree science and engineering flag - NSF definition
  - Field of degree science and engineering related flag - NSF definition
  - Self-employment income past 12 months
  - Sex
  - Wages or salary income past 12 months (use ADJINC to adjust WAGP to constant dollars)
  - Work experience of householder and spouse
  - When last worked
  - Worked last week
  - State code based on 2010 Census Definition
- 



# Cleaning & Data Exploration

Males	Females
49.0294%	50.9706%

- Dataset was balanced in that it had almost the same number of male responses as female responses.
- The Male to Female ratio was very close to the U.S. Census estimate as of April 1, 2020, where females accounted for 50.5% of the population.
- Males in this data set are approximately two years younger than females.

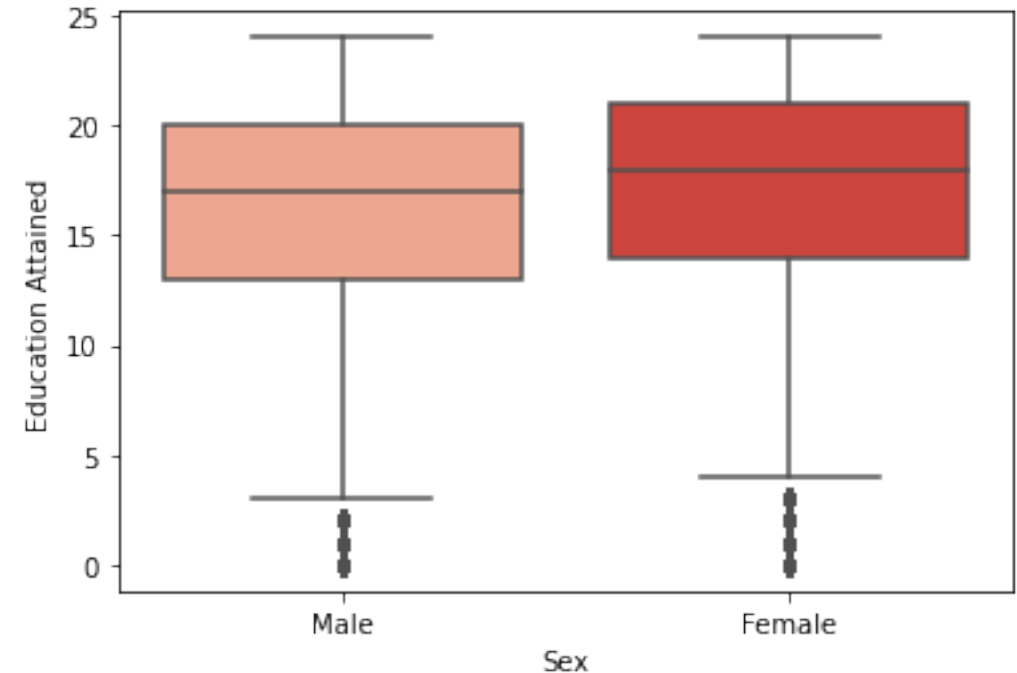


# Cleaning & Data Exploration

## Interesting Findings

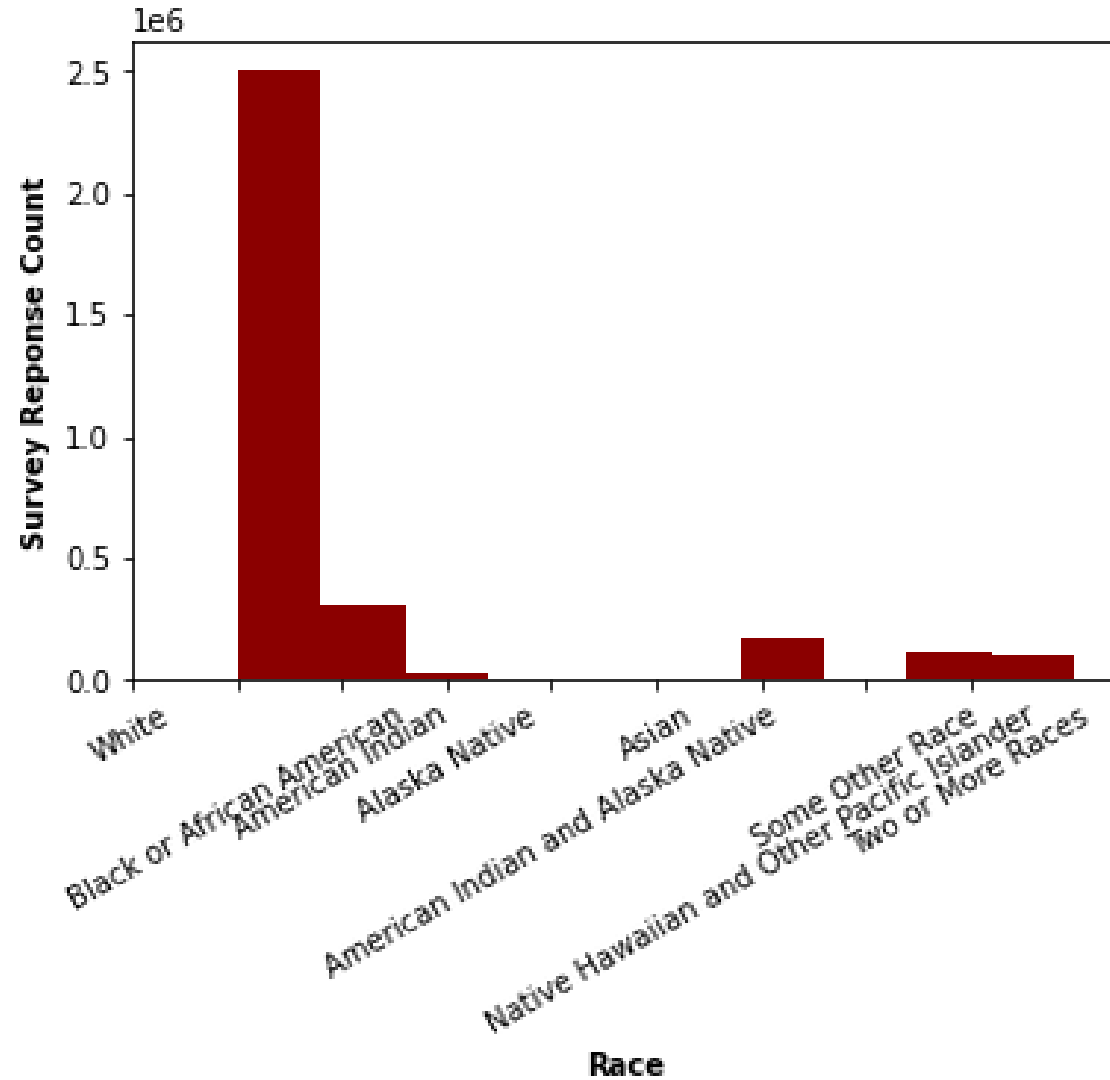
- Gradual increase of grade levels
- Outliers below grade five for both sexes
- Income belongs mostly to people above 18 years old
- Income increases during young to midlife and tapers off late life
- People over 18 have a balanced overall range of grade level
- Many adults never attain more than a high school degree

**Boxplot: Level of Education Attained, Grouped by Sex**

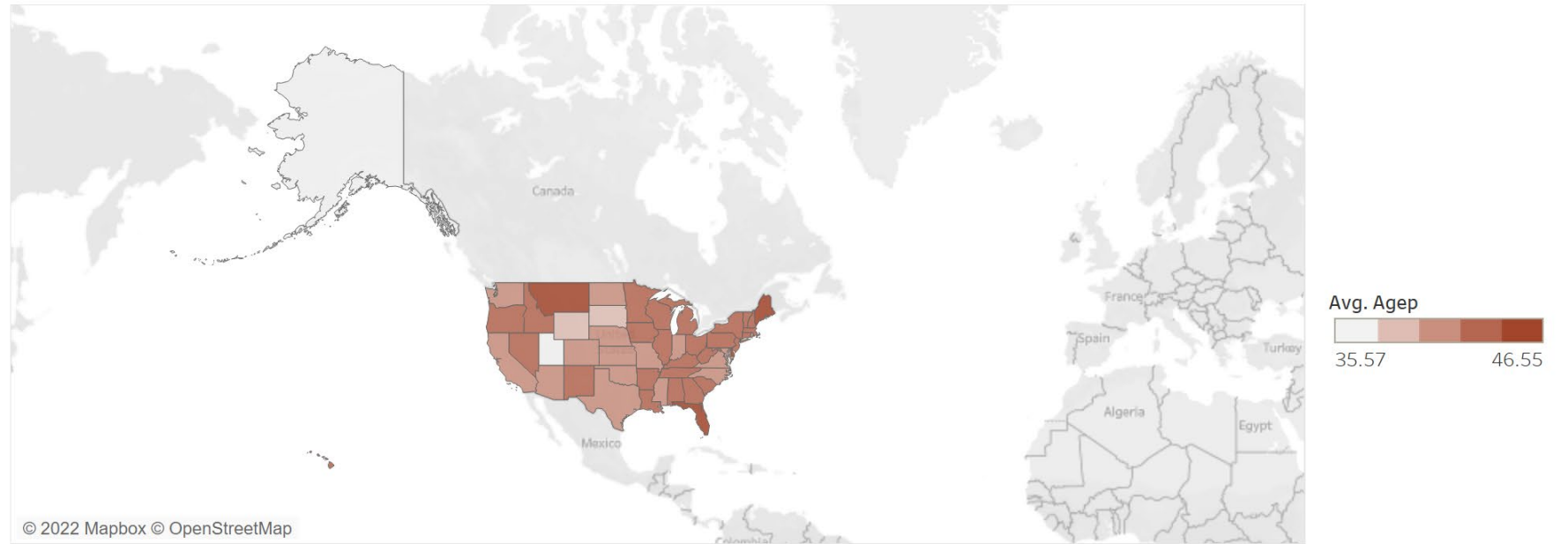


# Submissions Based on Race

The Race Counts for People Responding to the PUMS Survey



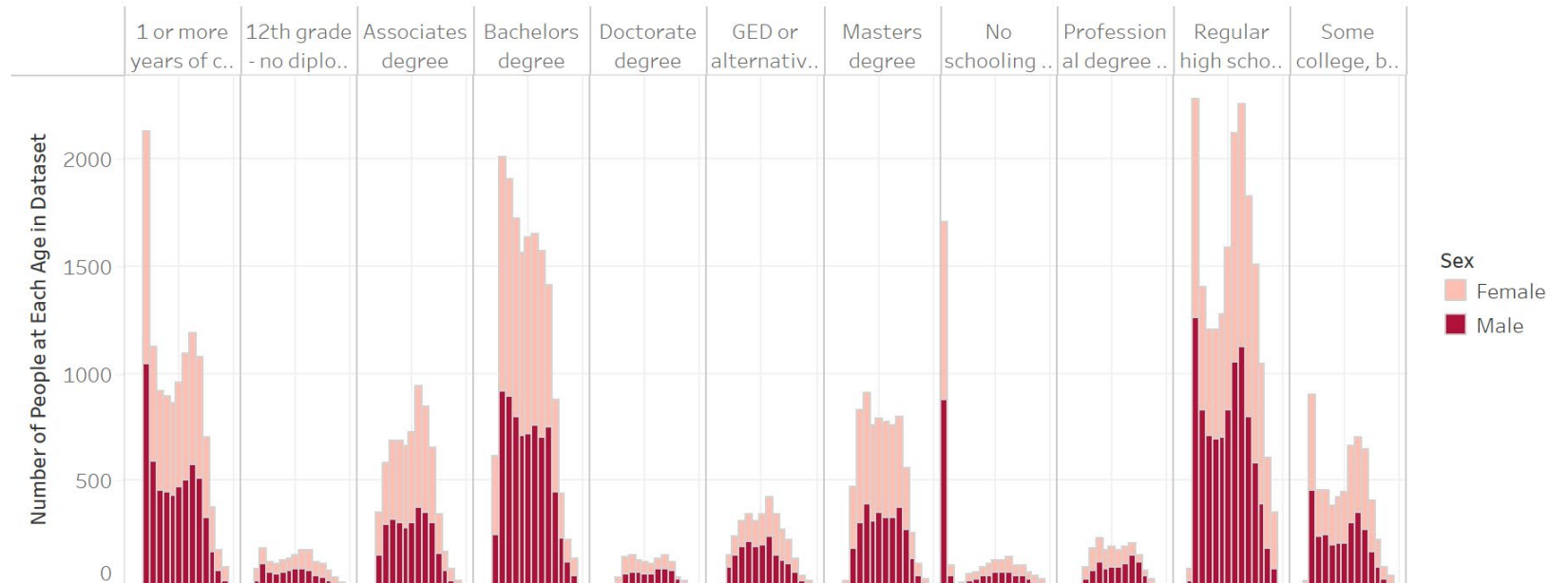
Mean Age Per State



Alaska has the youngest population

Maine has the oldest population

Number of People: Education Attainment, Ages, and Sex of People

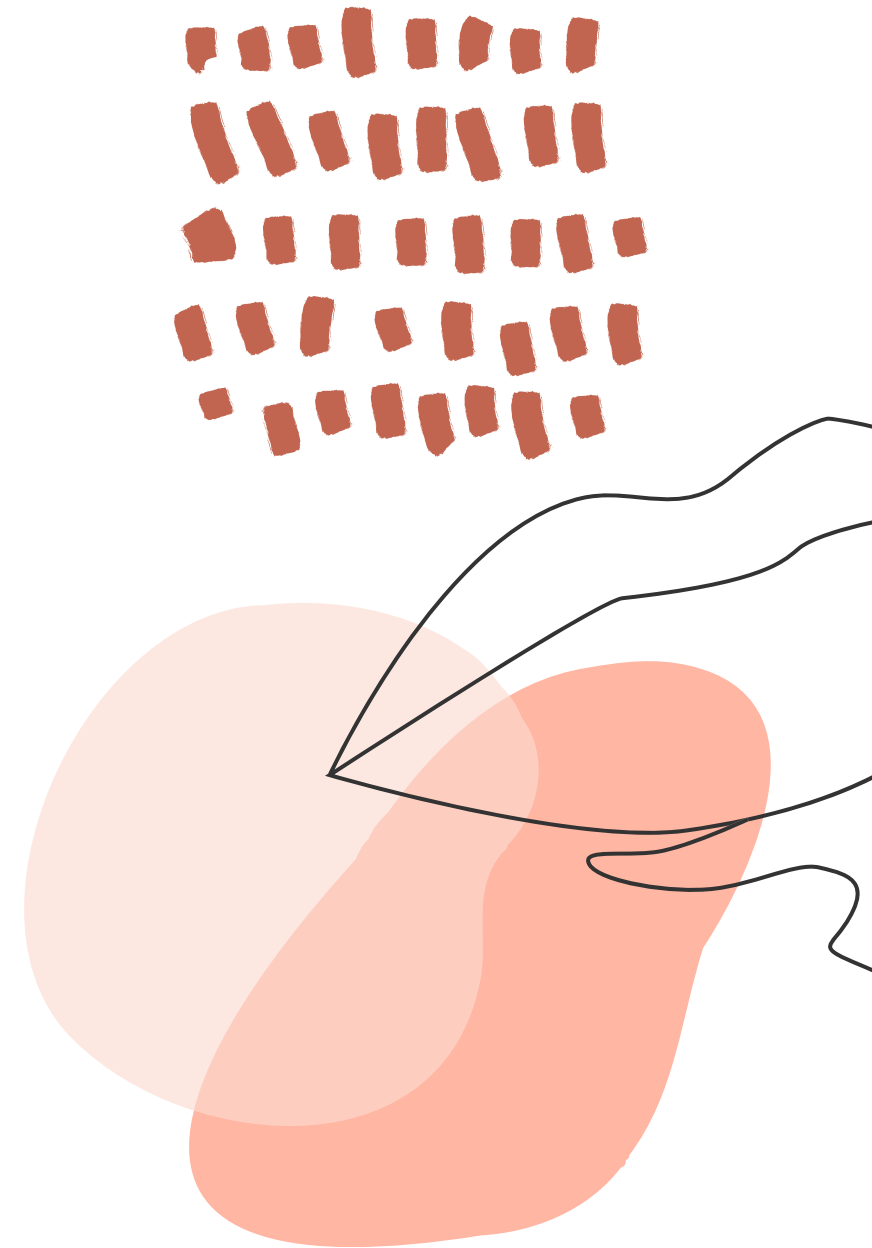


Females over 18 have attained a slightly higher level of education

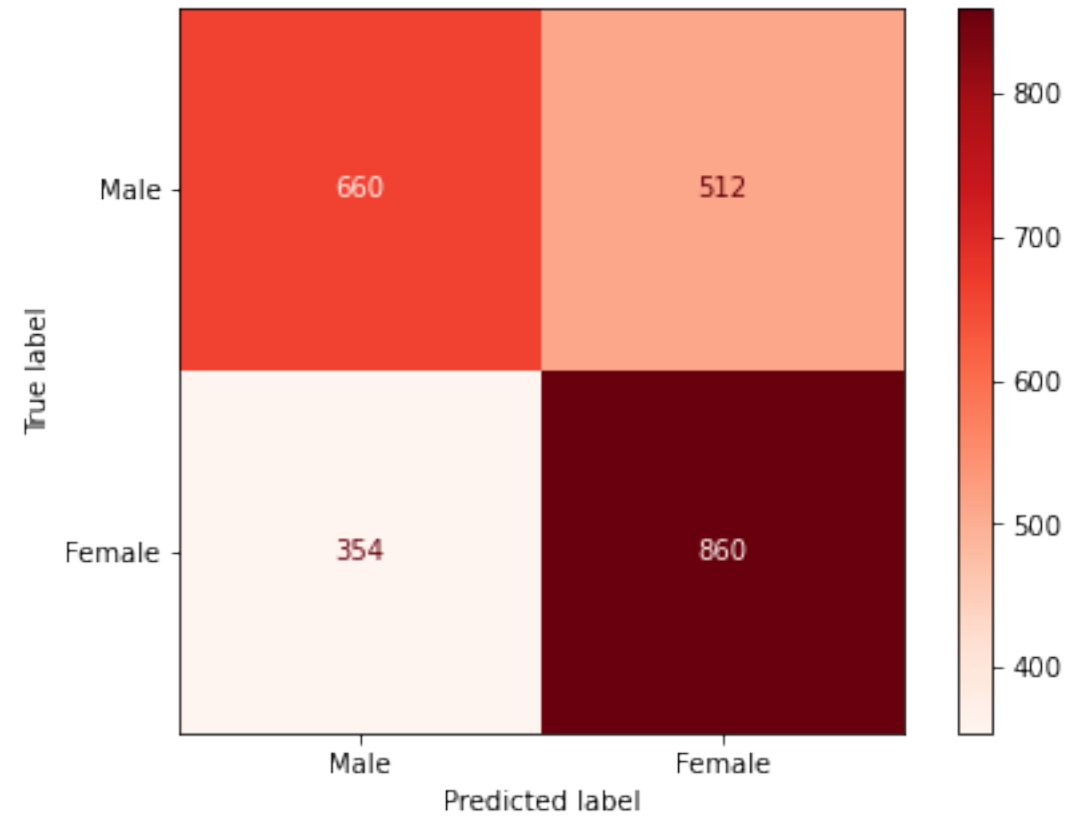
# Modeling

## Supervised Classification:

- Explored quick models to see which one performed the best at predicting male or female based on features.
  - ✓ Logistic Regression
  - ✓ Naive Bayes
  - ✓ Random Forest Classifier
- Logistic Regression had the best score, so I used it to model further.



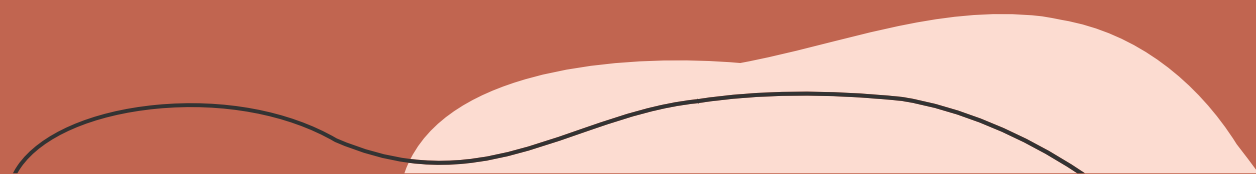
# Logistic Regression





# Best Model Findings

Best Score	Train Score	Test Score	Accuracy	Recall	Precision	Specificity	F1
0.6255166 217430368	0.6278526 5049416	0.6370494 551550713	0.6357709 543381819	0.5631399 317406144	0.6508875 73964497	0.5631399 317406144	0.6038426 349496798



# Implications

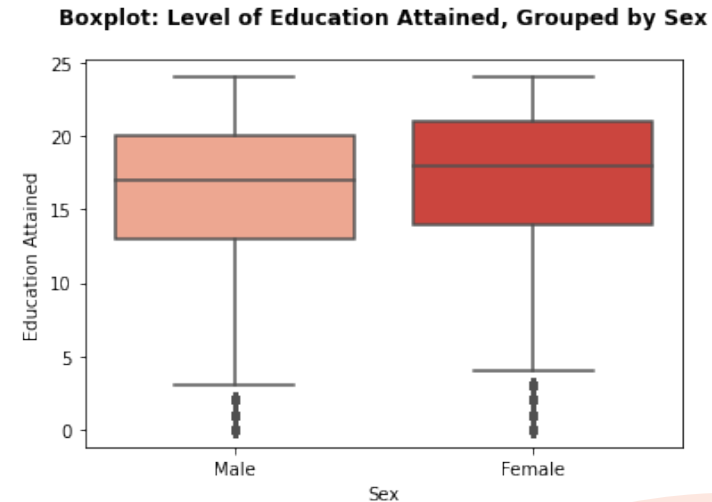
## Problem Statement (Hypothesis):

The average difference in the years of education attained between gender=1, gender=2 is zero.

The average difference in the years of education attained between gender=1, gender=2 is not zero.

Gender 1 = male | Gender 2 = female

The Null Hypothesis is accurate because the average difference in years of education between males and females is not zero. The difference is **0.186681**





# Conclusion

In conclusion, women are getting more degrees than men.

They are still paid less than men.





# Next Steps

1. Make sure that though females are getting more degrees than males, they can make the same amount of money.
2. Dive deeper into different forms of classification models to continue to explore the differences between males & females.





# Questions

Brandie Hatch  
[in/brandiehatch/](https://www.linkedin.com/in/brandiehatch/)

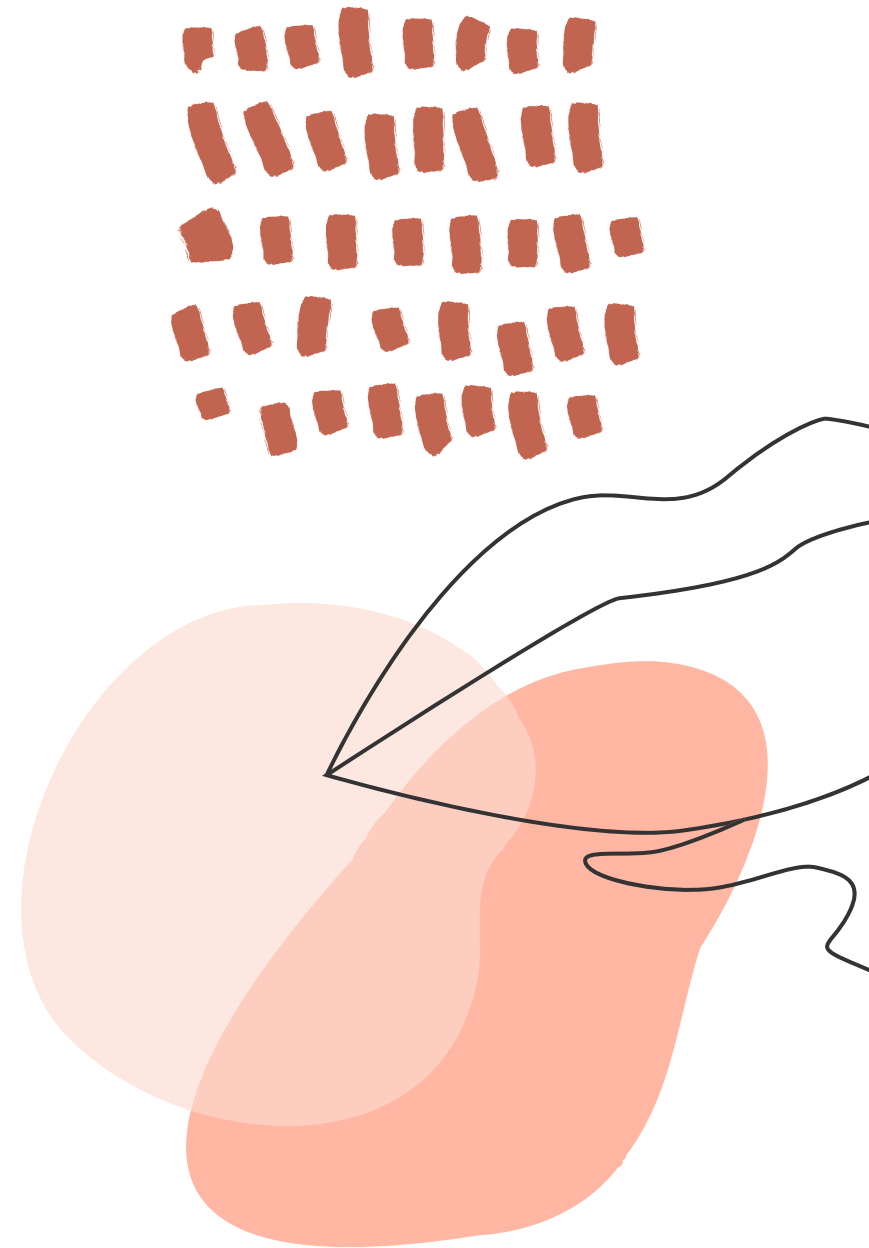


# Sources

## List the resources you used for your research:

- U.S. Census
- <https://www.census.gov/topics/population/age-and-sex/about.html>
- <https://ask.census.gov/prweb/PRServletCustom?pyActivity=pyMobileSnapStart&ArticleID=KCP-2950>
- [https://www2.census.gov/programs-surveys/acs/tech\\_docs/pums/ACS2016\\_2020\\_PUMS\\_README.pdf](https://www2.census.gov/programs-surveys/acs/tech_docs/pums/ACS2016_2020_PUMS_README.pdf)
- [https://www.census.gov/programs-surveys/economic-census/guidance-geographies/levels.html#par\\_textimage\\_34](https://www.census.gov/programs-surveys/economic-census/guidance-geographies/levels.html#par_textimage_34)
- Code Sources:
- Pull a smaller random sample from a large dataset:  
<https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.sample.html>
- Why did the initial Gridsearch return nan:  
<https://datascience.stackexchange.com/questions/91225/why-gridsearchcv-returns-nan>
- Logistic Regression scoring issues in Gridsearch: [https://scikit-learn.org/stable/modules/model\\_evaluation.html](https://scikit-learn.org/stable/modules/model_evaluation.html)
- <https://towardsdatascience.com/20-popular-machine-learning-metrics-part-1-classification-regression-evaluation-metrics-1ca3e282a2ce>

# Appendix



# Polynomial Features Attempt

```
[48] print(log_gs.best_score_)
      log_gs.best_params_
... 0.5062857142857143

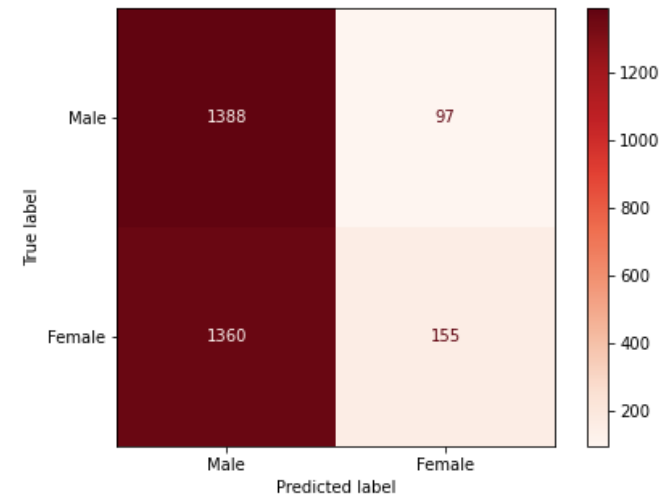
{'log_reg__C': 0.01, 'log_reg__penalty': 'l2'}
```

```
[49] log_gs.score(X_train, y_train), log_gs.score(X_test, y_test)
... (0.5064285714285715, 0.5143333333333333)
```

```
[50] pd.Series(log_gs.predict(X_test)).value_counts()
... 1    2748
     2    252
     dtype: int64
```

## Confusion Matrix Displays

```
[51] # Logistic Regression
      fig, ax = plt.subplots(figsize=(7, 5))
      ConfusionMatrixDisplay.from_estimator(log_gs,
      X_test,
      y_test,
      display_labels=['Male', 'Female'],
      cmap='Reds',
      ax=ax);
```



# Initial StandardScaler Scores

```
scoring=[accuracy, precision, recall, f1]
```

```
[119] print(log_gs.best_score_)
      log_gs.best_params_
...  0.5285714285714286
      {'log_reg__C': 1, 'log_reg__penalty': 'l2'}
```

```
[120] log_gs.score(X_train, y_train), log_gs.score(X_test, y_test)
...  (0.5275714285714286, 0.5293333333333333)
```

```
[121] pd.Series(log_gs.predict(X_test)).value_counts()
...  2    1621
      1    1379
      dtype: int64
```

## Model Evaluation & Comparison

Confusion Matrix Displays