The background is a dark blue color. In the top-left and bottom-right corners, there are decorative geometric patterns made of light blue lines. These patterns consist of multiple nested, downward-pointing chevrons or 'V' shapes, creating a textured, crystalline effect. The main title is centered in the middle of the slide in a large, bold, light blue font.

# **Predicting Income Greater and Lesser than \$50k**

A Study of US Census Data (1994-95)

By: Terri John

## » Problem Statement and Background

- US Census Data results empower the USG to make data driven decisions regarding allocation of funding.
- Improved understanding of demographic characteristics of subpopulation.
- Income disparities.



# » Methodology

» 1

## Data Cleaning and EDA

Exploring the  
data and  
preparing it for  
modeling

Missing Data

» 2

## Feature Engineering

Categorical  
Features

» 3

## Modeling

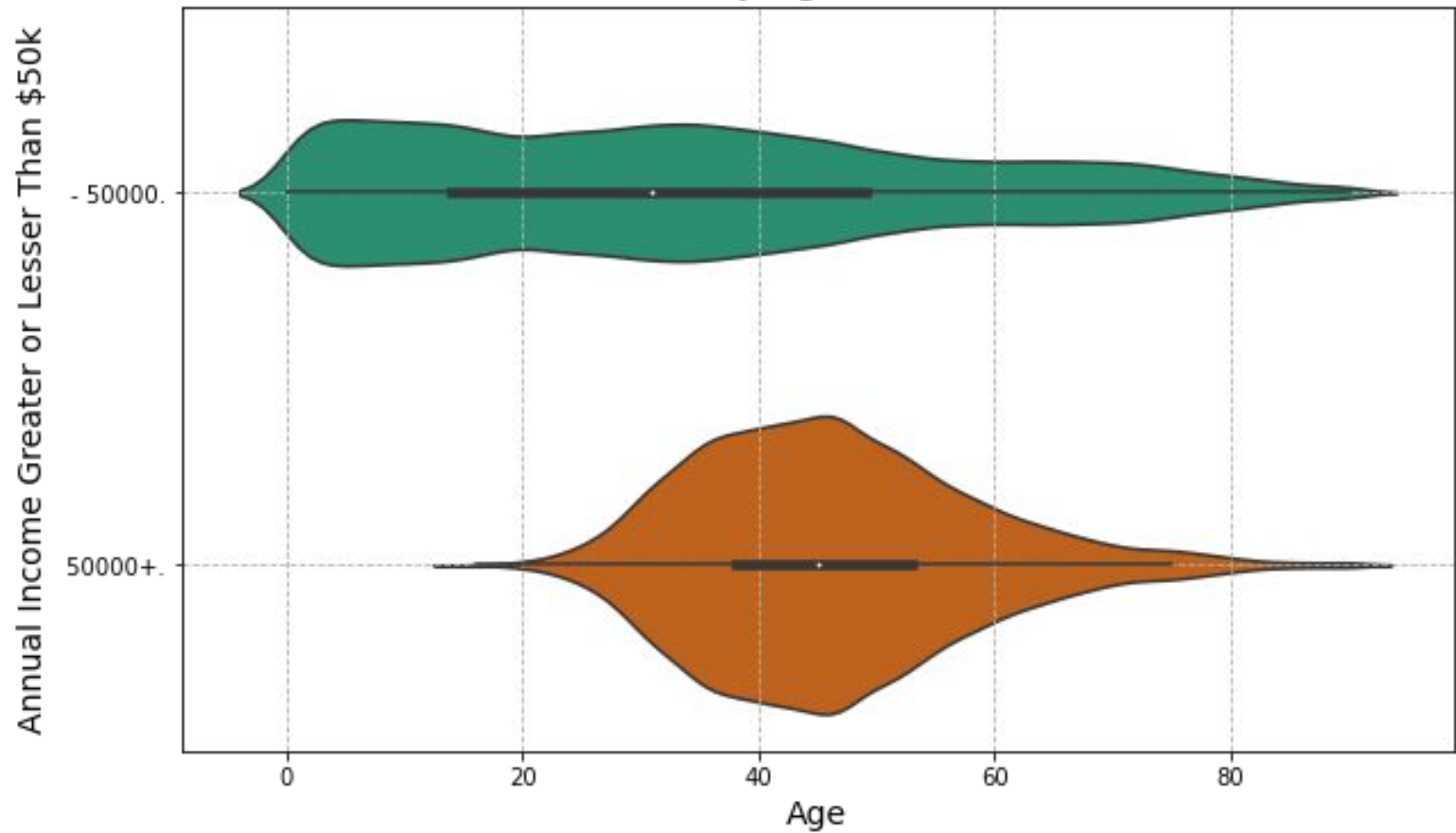
Binary  
Classification  
Model to  
predict income

» 4

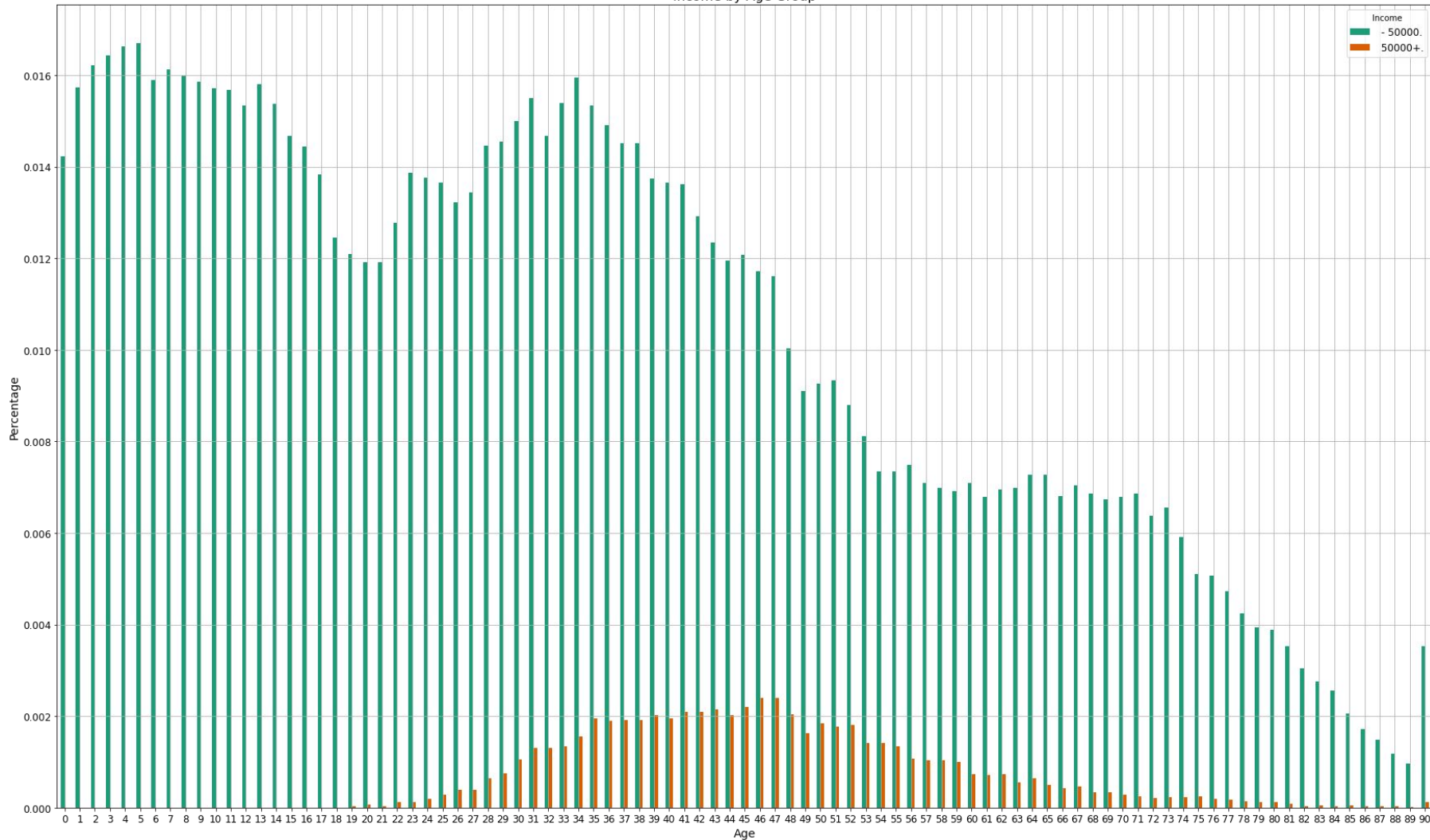
## Analysis

Model Performance  
  
Key Takeaways

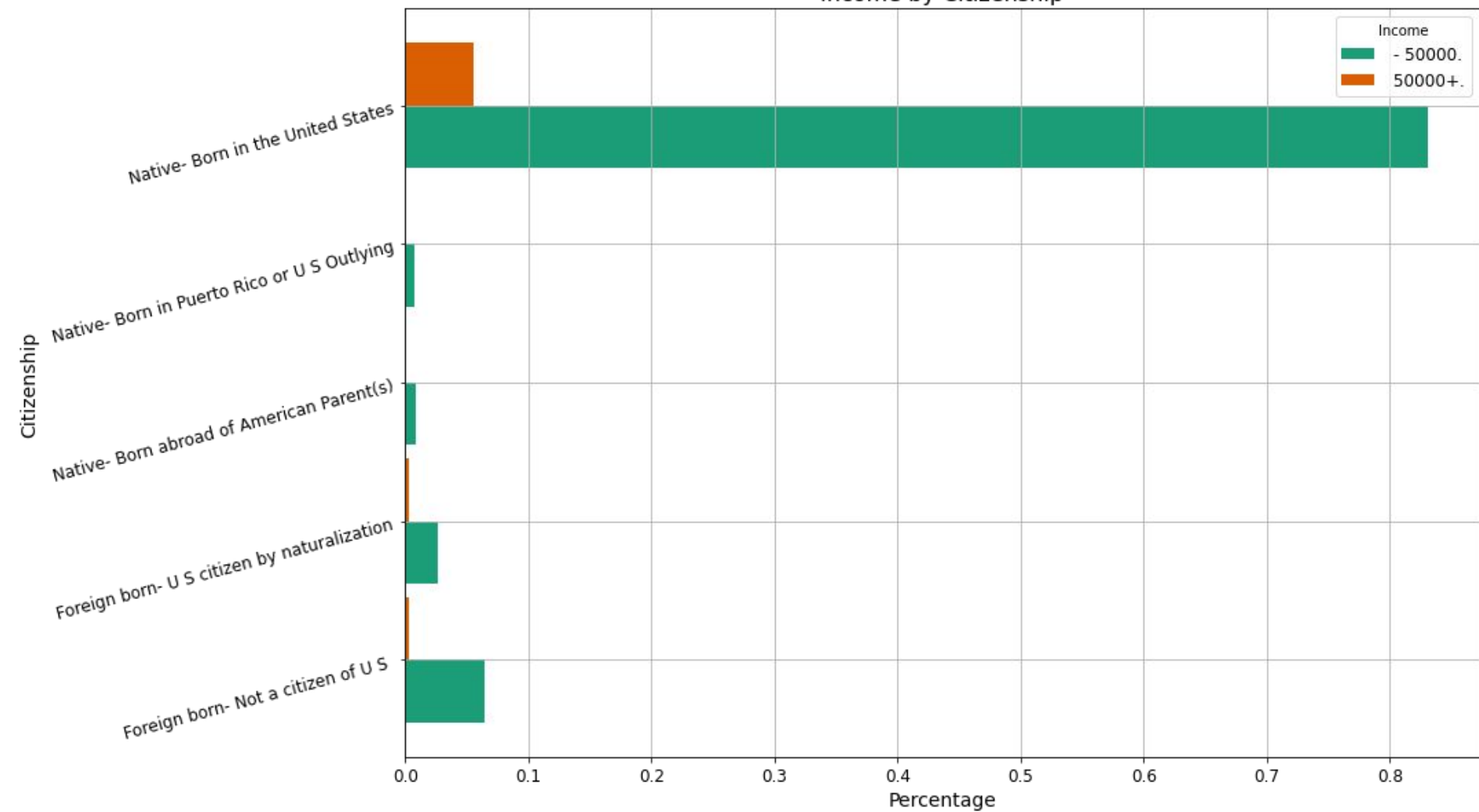
Income by Age Distribution



Income by Age Group



Income by Citizenship

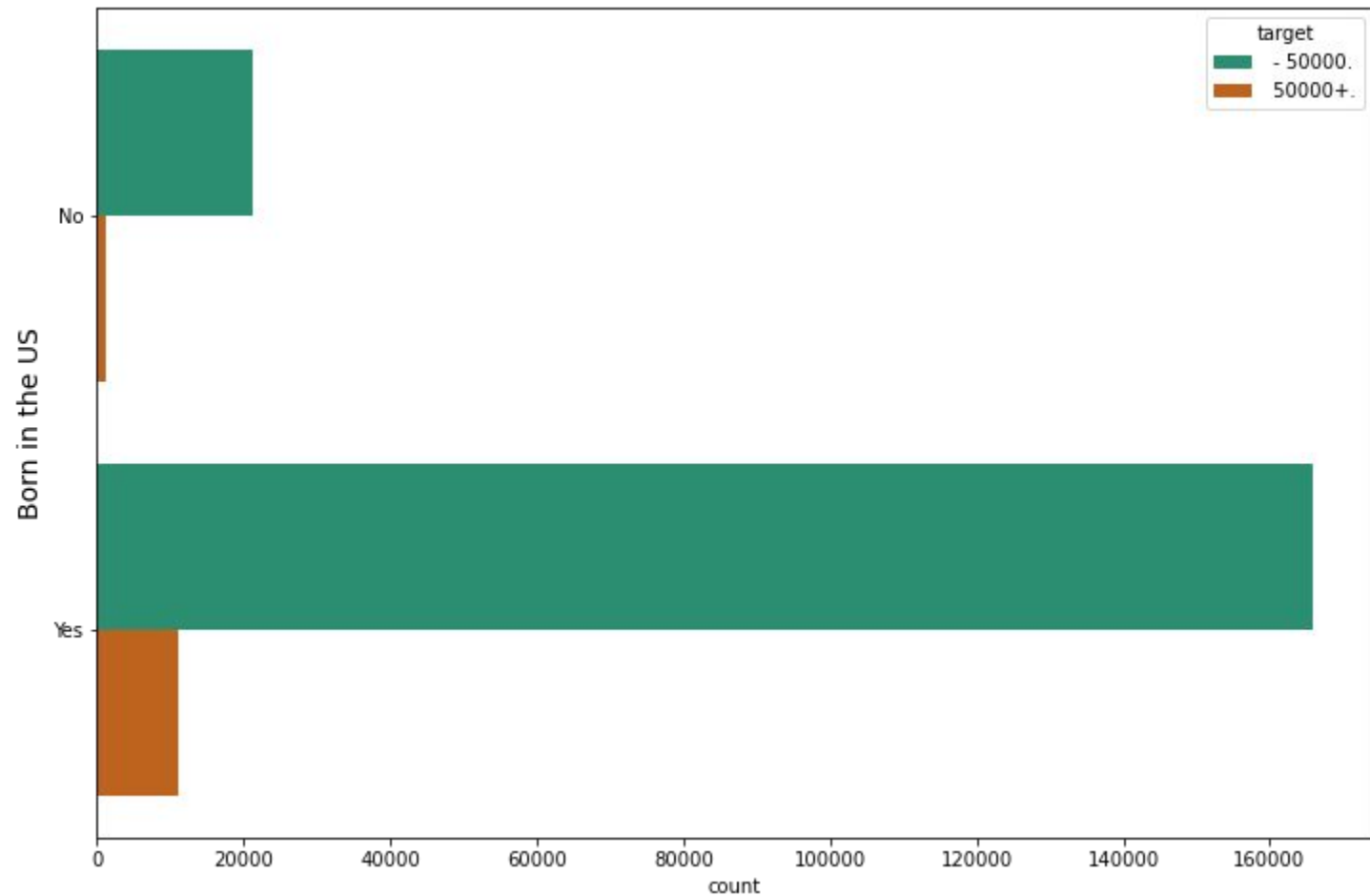




## » Citizenship

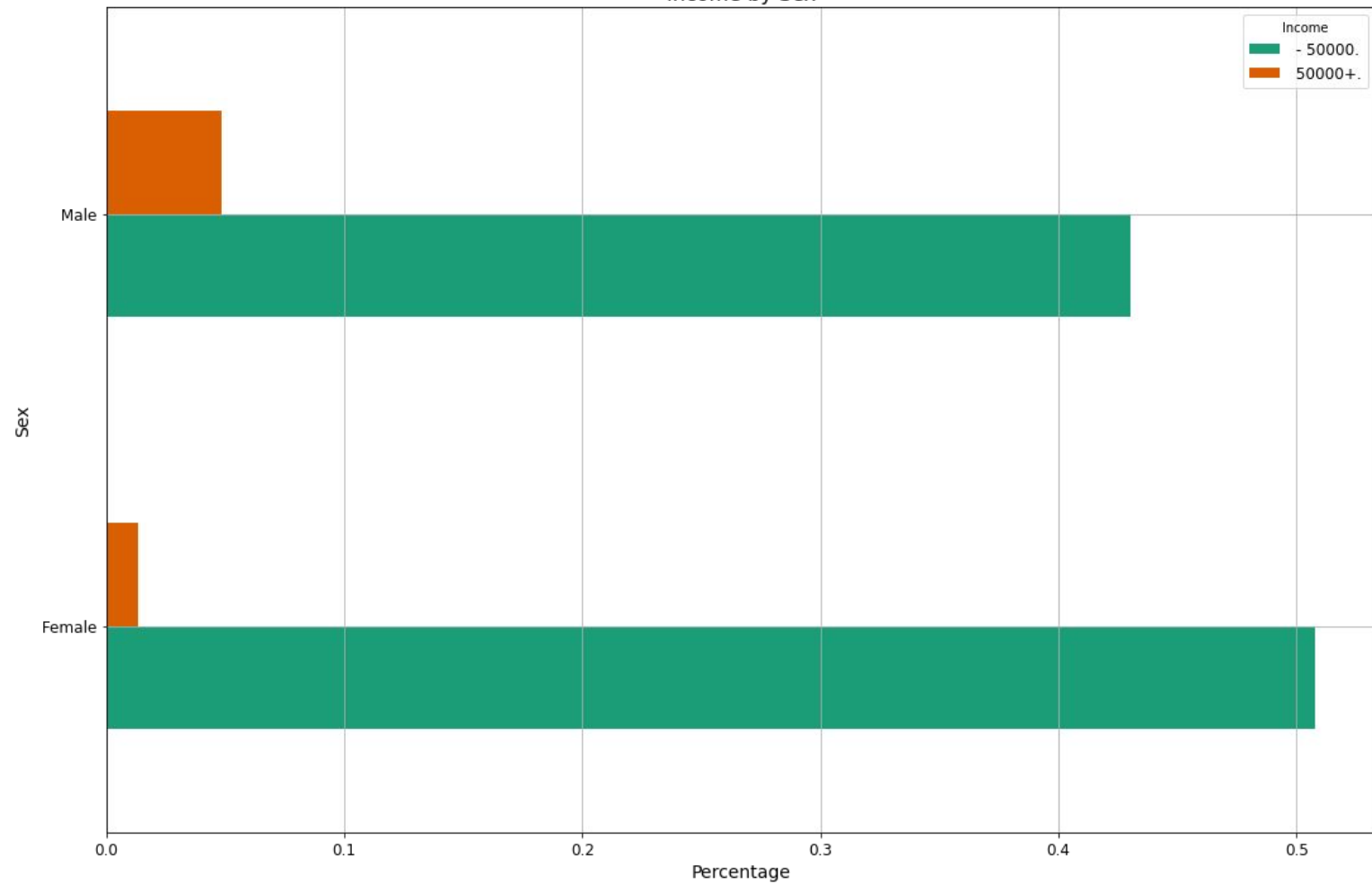
<b><u>Citizenship Status</u></b>	<b><u>Population %</u></b>	<b><u>% of Group +\$50k</u></b>
Native- Born in the United States	88.7%	6%
Foreign born- Not a citizen of U S	6.7%	3%
Foreign born- U S citizen by naturalization	2.9%	10%
Native- Born abroad of American Parent(s)	.8%	7%
Native- Born in Puerto Rico or U S Outlying	.7%	2%

Born in the USA



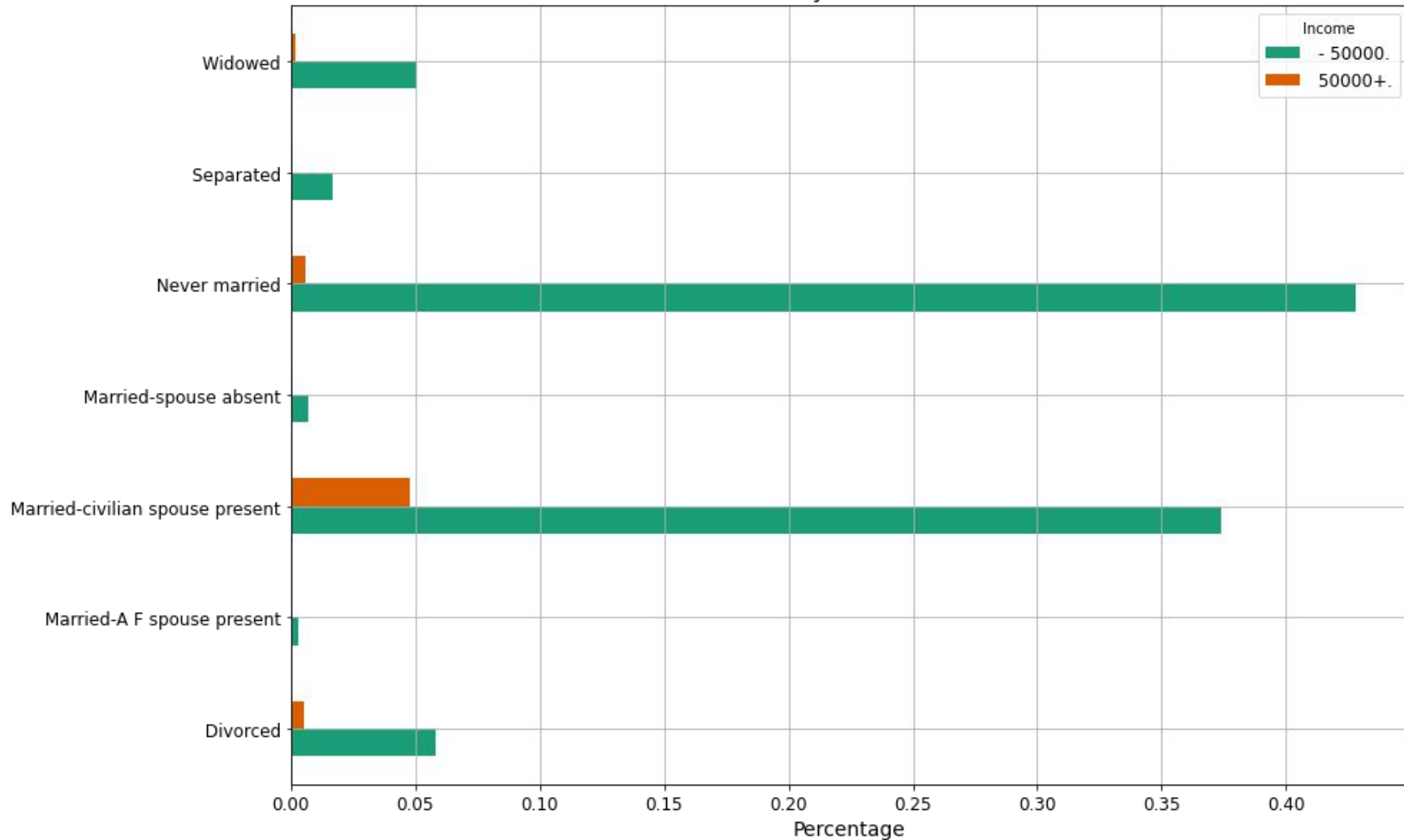


Income by Sex

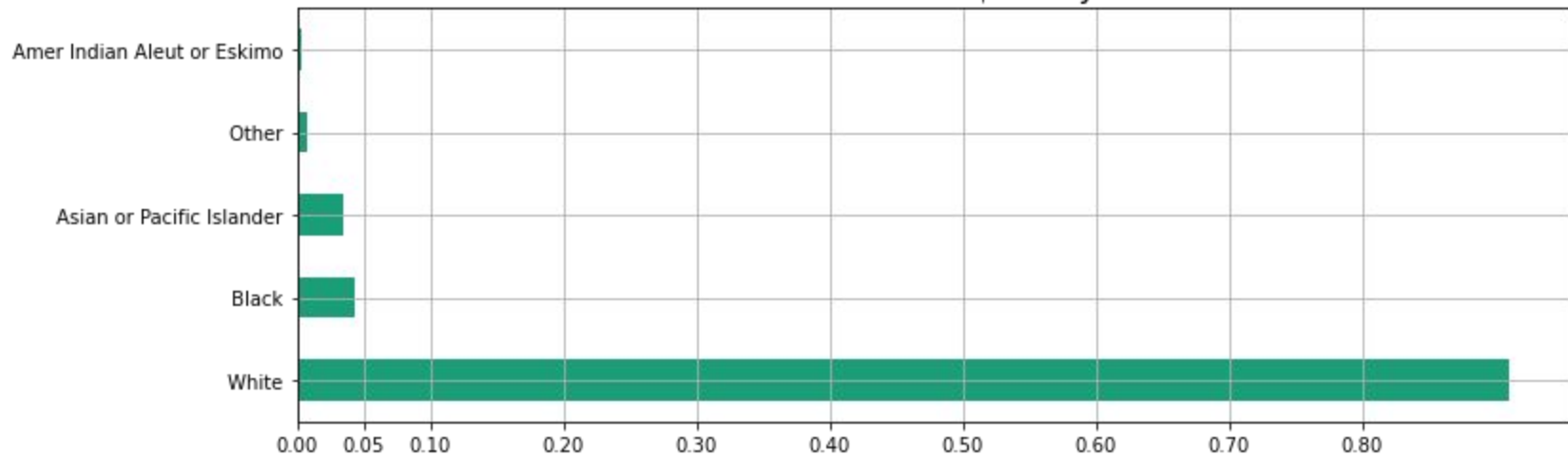


Income by Marital Status

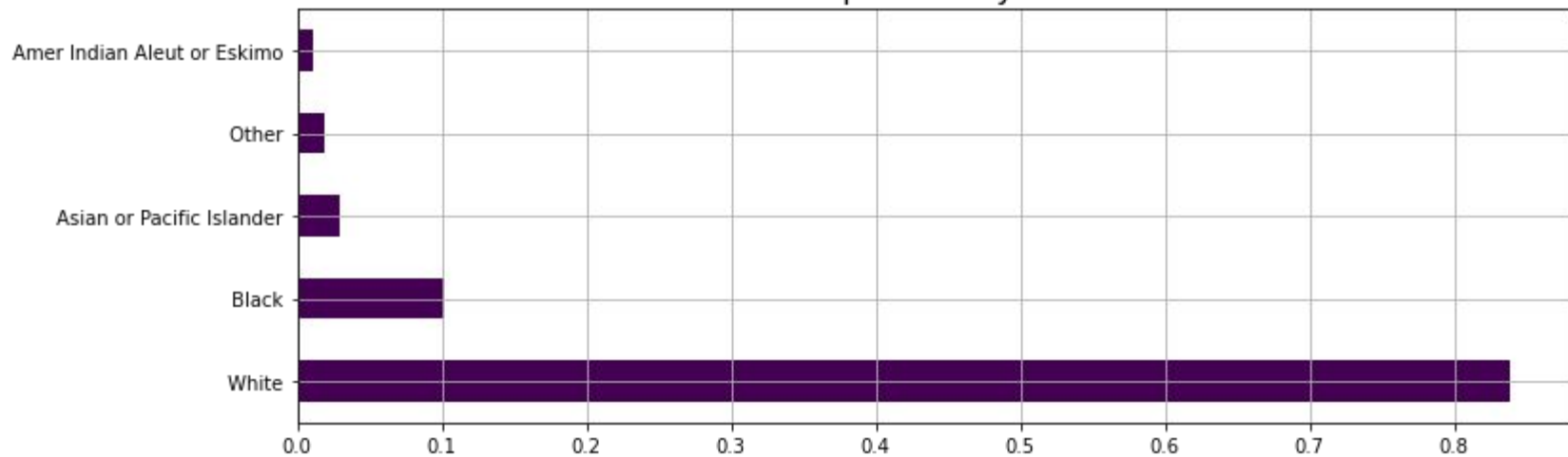
Marital Status



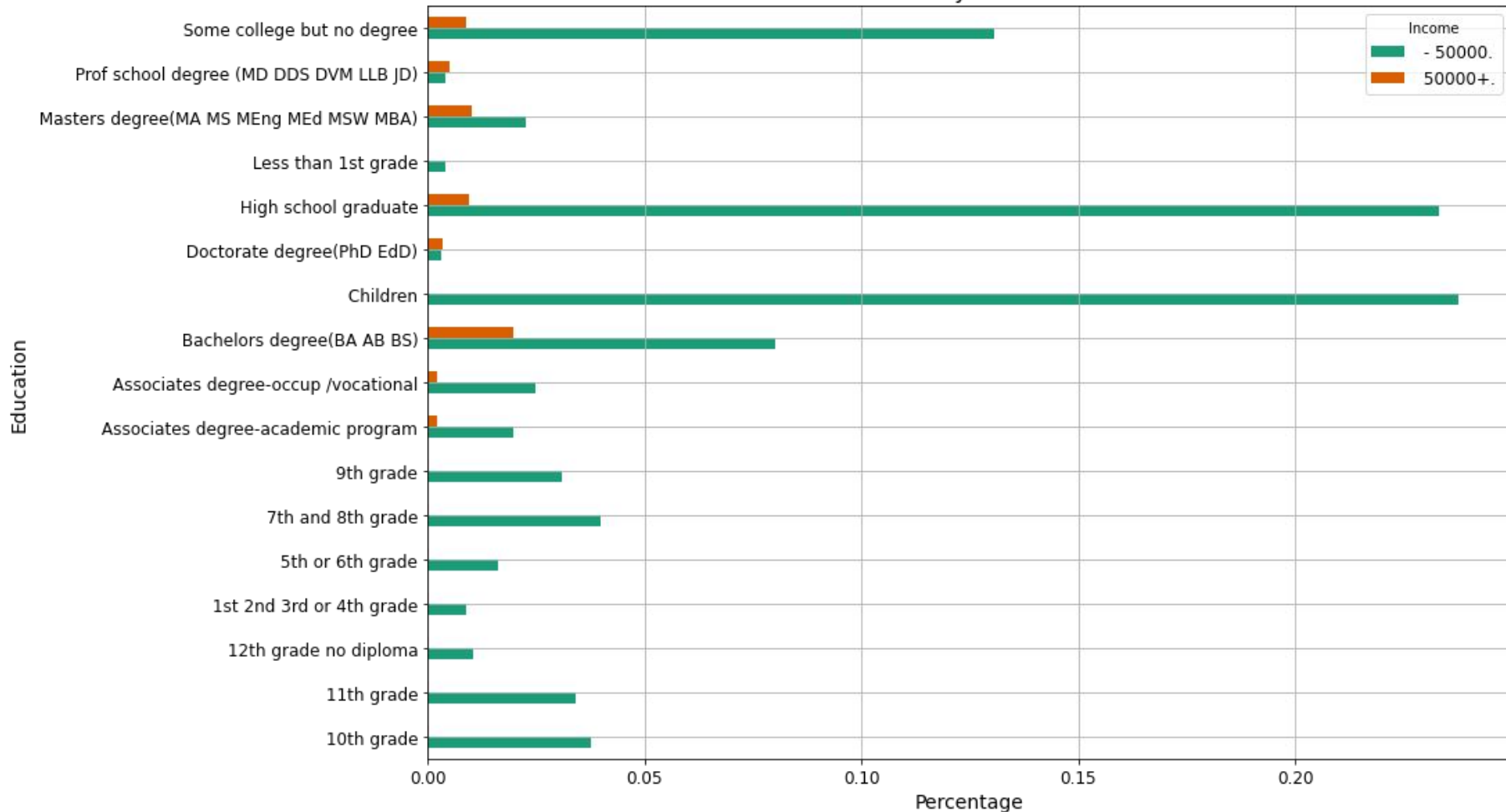
% Earners of Over \$50k by Race



% Population by Race



# Income by Education





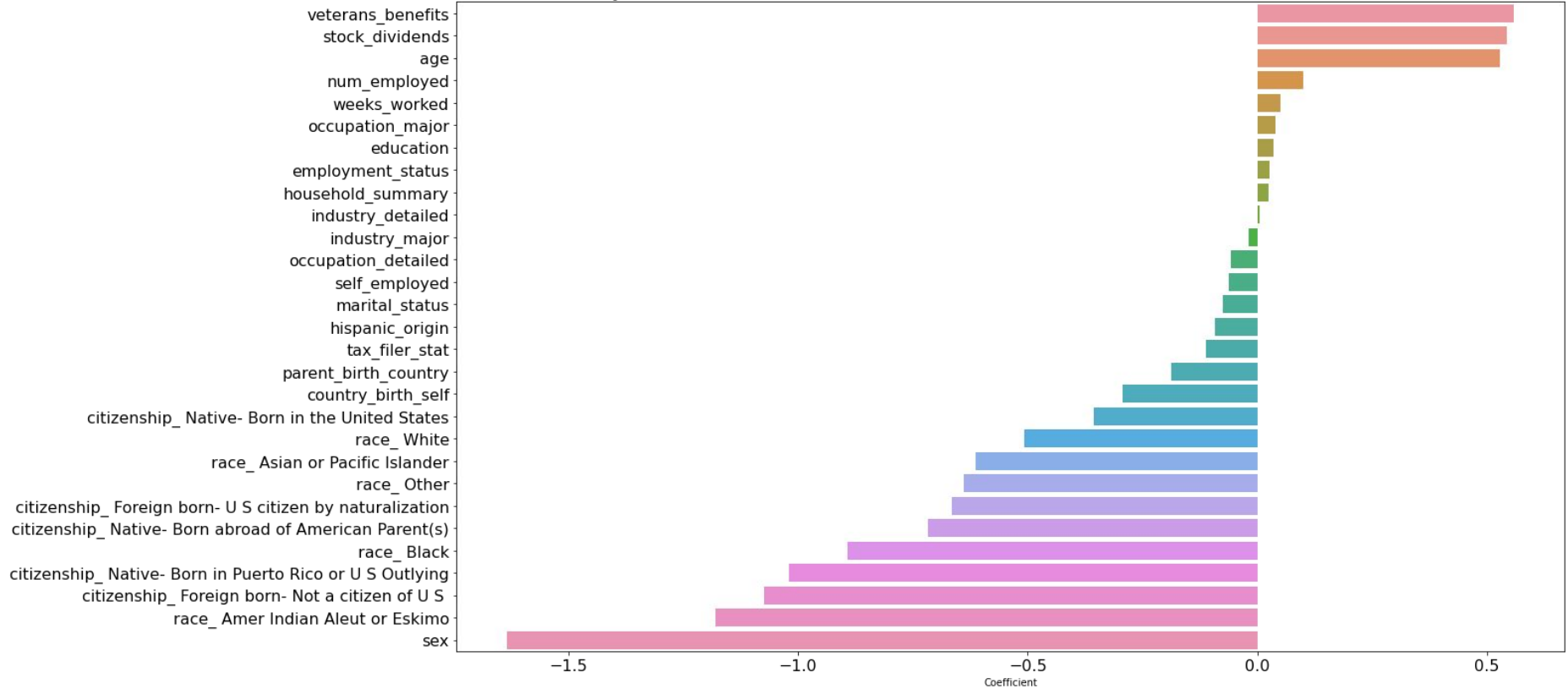
## » A Tale of Two Models

### Predicting Annual Income Greater/Lesser Than \$50k

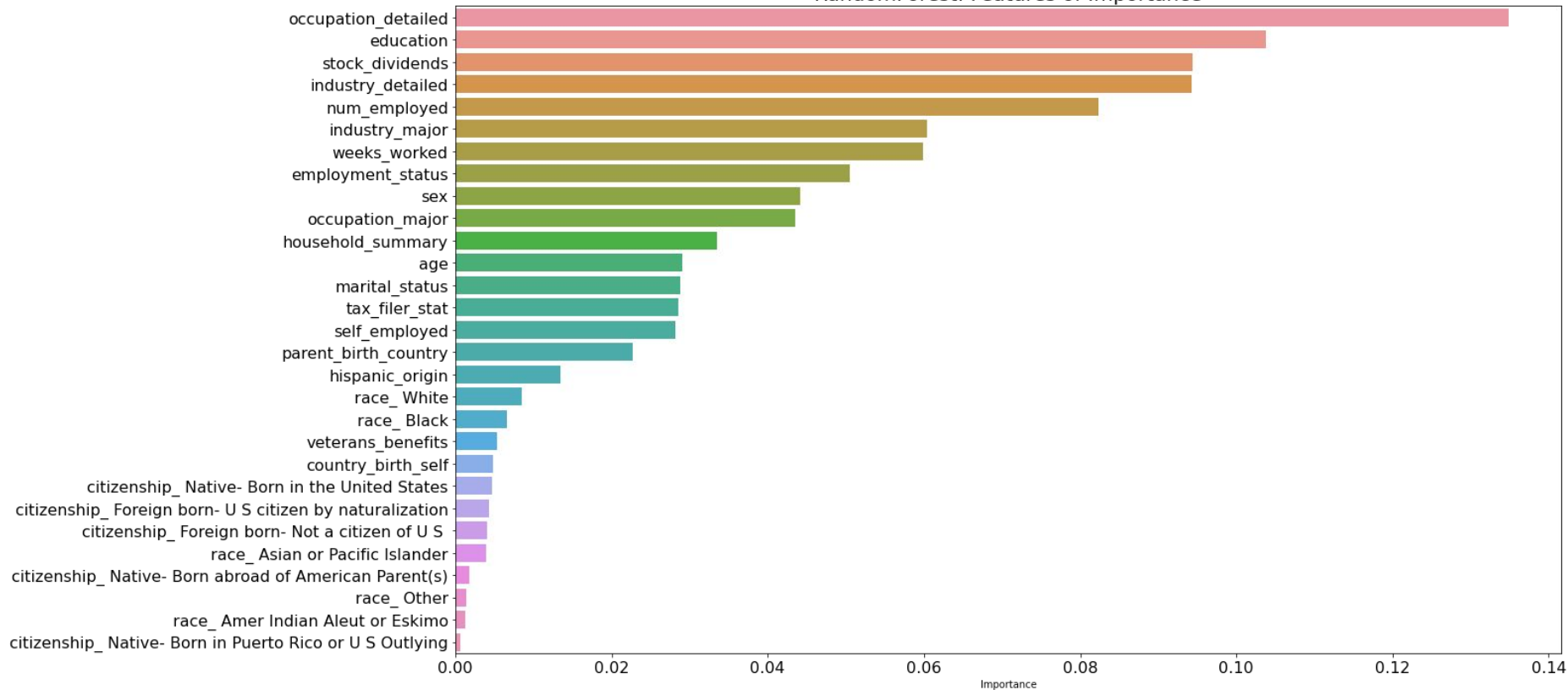
	Train Accuracy	Test Accuracy	Recall	F1 Score
<b>Logistic Regression</b>	94.5%	94.5%	0.24	0.35
<b>RandomForest</b>	99.2%	98.9%	0.86	0.90

\*The baseline model assumes a 93% accuracy score, based on the majority class.

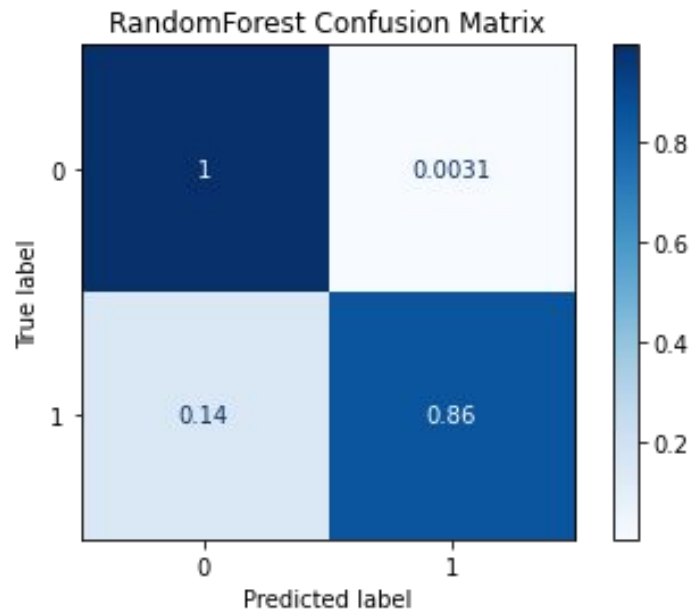
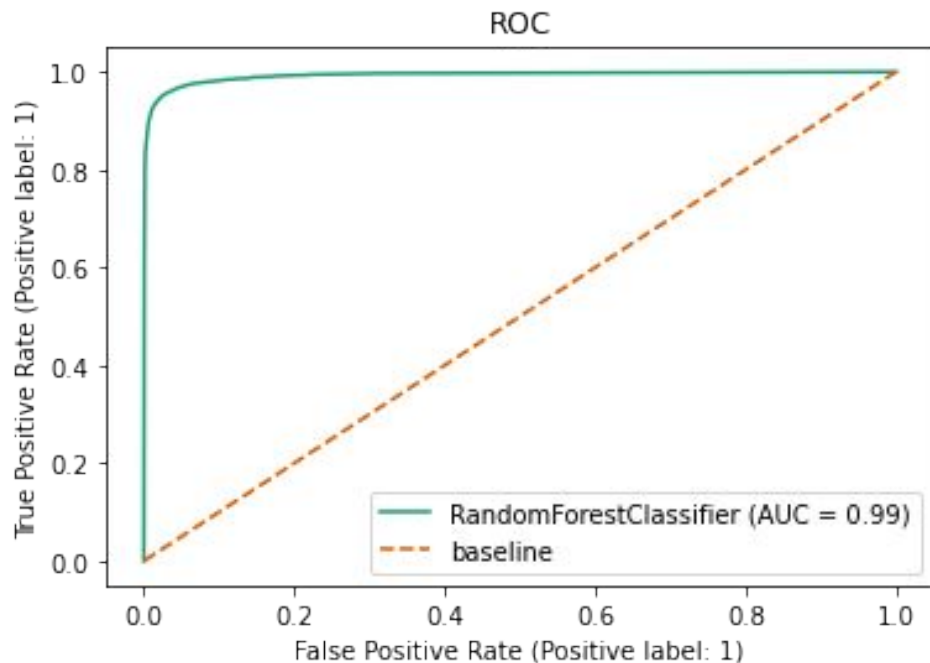
Binary Classification Model Coefficients: Predictors of Income Greater or Lesser Than \$50k



RandomForest: Features of Importance



# » RandomForest Model

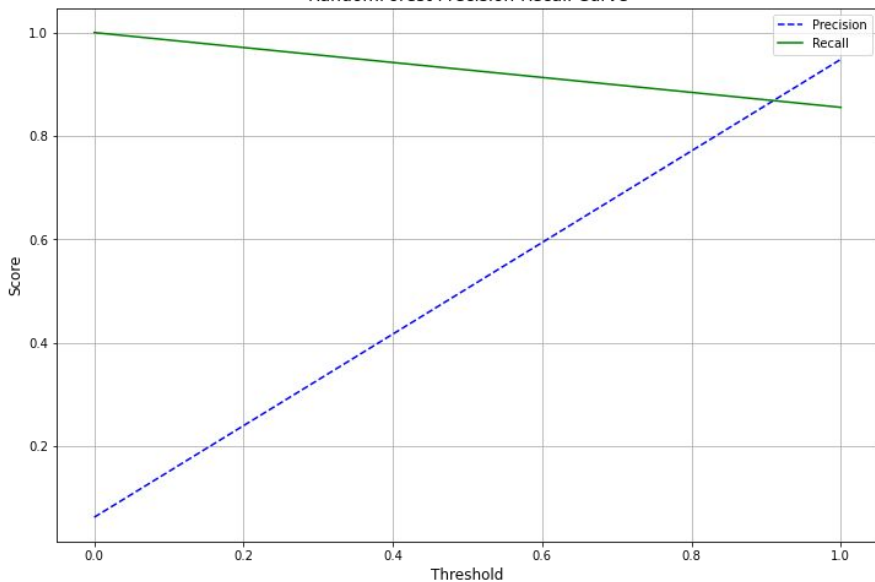




# » Precision - Recall

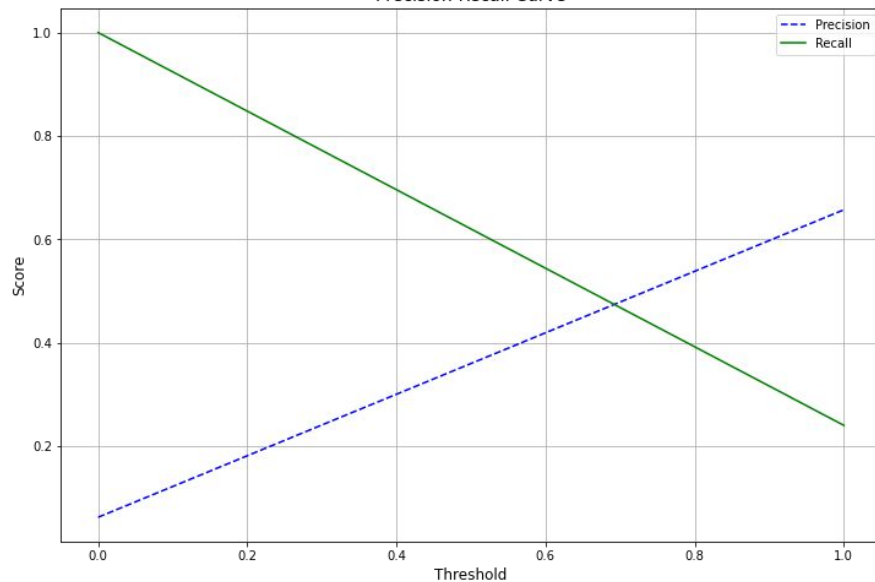
## RandomForest

RandomForest Precision-Recall Curve



## LogisticRegression

Precision-Recall Curve





# » Conclusion and Recommendations

## Key Takeaways

» RandomForest  
Model

» Influential  
Characteristics

» Why does it  
matter?

## Next Steps

» Bring in more  
data to fill null  
values.

» Cross compare  
features

» Pull in data for  
parent and self  
birth country



# Thanks!

**Do you have any  
questions?**


[tjohn07@gmail.com](mailto:tjohn07@gmail.com)

<https://www.linkedin.com/in/terri-john/>

<https://www.terrijohn.com>

**CREDITS:** This presentation  
template was created by **Slidesgo**,  
including icons by **Flaticon**, and  
infographics & images by **Freepik**

Please keep this slide for attribution





## » Sources

Data for this project was provided by Dataiku.

Original data is from the US Census Bureau: <http://www.census.gov/ftp/pub/DES/www/welcome.html>

### **Outside sources:**

- Forbes:  
<https://www.forbes.com/sites/marisadellatto/2021/10/05/single-adults-make-less-money-than-partnered-ones-study-says/?sh=360d12ad454f>
- 