

Predicting Youth Risk Behaviors:

Modeling the YRBS for Relevant Discussions with Youth

Becky Peters, Data Scientist; September 2021

Topics for Discussion

- **Part 1:** Background and Motivation
- **Part 2:** Exploring and Modeling the Data
- **Part 3:** Predictions and Dashboard
- **Part 4:** Future Work

Part 1:

Background and Motivation



“...kids learn to make good decisions by making decisions, not by following directions.”

- Educator / Author, Alfie Kohn



Youth Risk Behavior Survey

- CDC Youth Risk Behavior Surveillance System, 1997-2019
 - Violence, sexual behaviors, alcohol / drug / tobacco use, diet, physical activity
- [cdc.gov](https://www.cdc.gov/yrbs/index.html)

Overview

Results

Reports, Fact Sheets, and Publications

Participation Maps & History

Frequently Asked Questions

Methods

Questionnaires

Data & Documentation

Trends Report

Toolkit

Journal Articles

Data Request Form

NYPANS

Trends Report

The *Youth Risk Behavior Survey Data Summary & Trends Report* uses YRBS data from 2009 to 2019 to focus on four priority focus areas associated with STDs, including HIV, and unintended teen pregnancy:

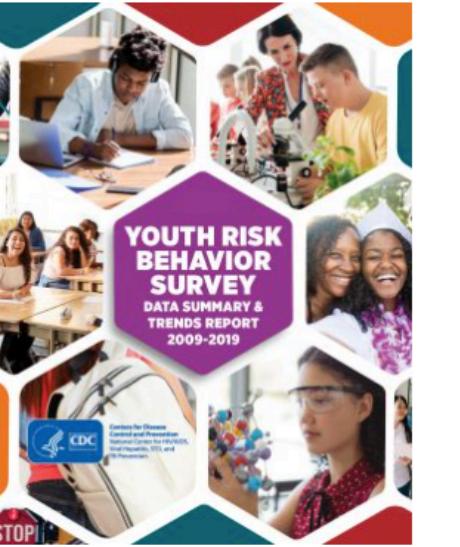
- Sexual Behavior
- High-Risk Substance Use
- Experiencing Violence, and
- Mental Health and Suicide.

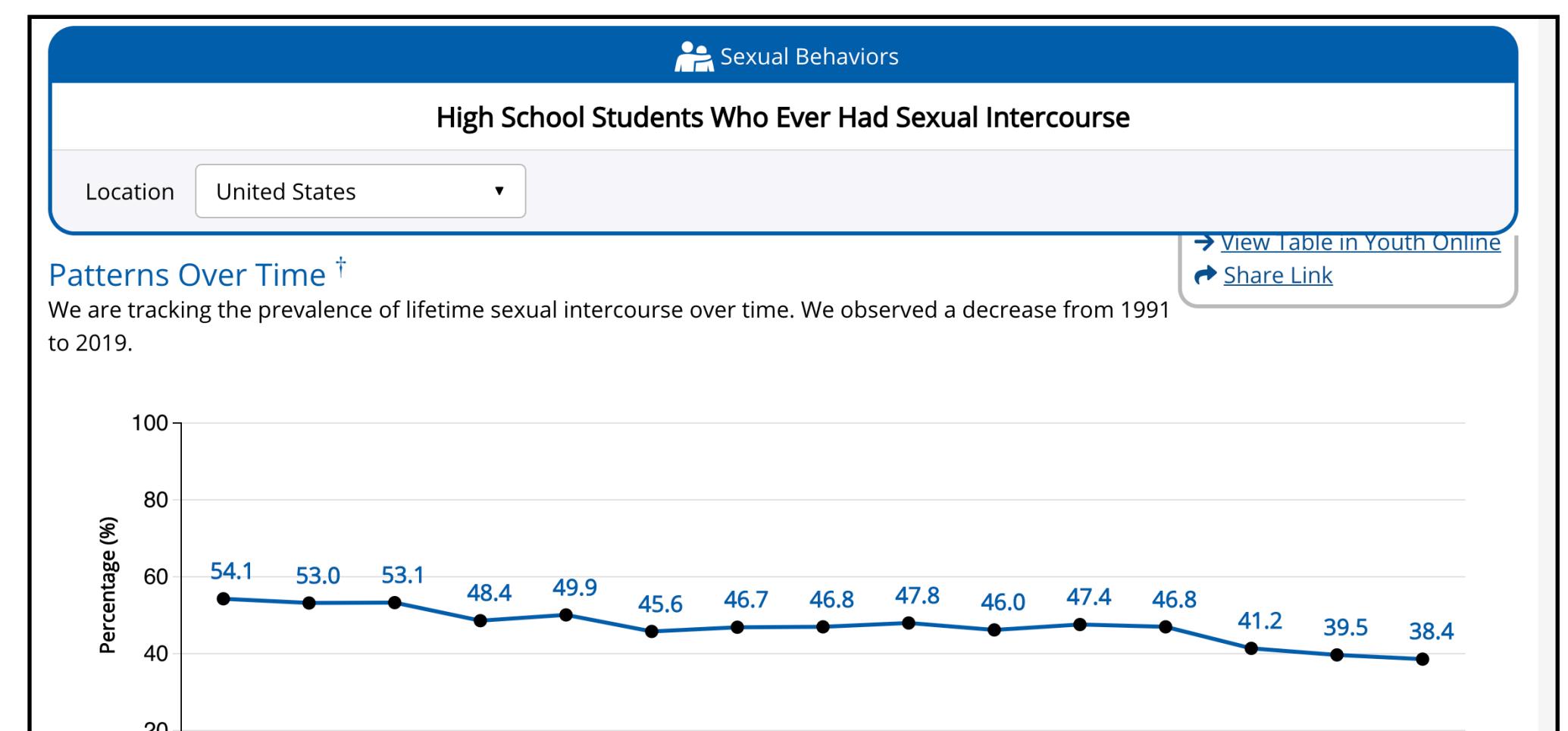
These health risk behaviors and experiences contribute to substantial health problems for adolescents.

To raise awareness and understanding of the prevention needs of adolescents, this report presents ten-year trends for 24 variables using YRBS data by sex, by race/ethnicity, and by sexual minority status.

[View Executive Summary](#)

[View Data Summary & Trends Report](#)





Project Inquiries

- Can we use 10 y of aggregated data to predict youth risk behaviors on an individual basis? (Supervised Multi-Label Classification Algorithms)
 - Q58: *Have you ever had sex?*
 - Q25: *Have you felt sad or hopeless for 2 weeks in a row?*
 - Q30: *Have you ever tried cigarettes?*
- Which negative behaviors are most related? (Graph Network)

Potential Users

- Health Care Providers
 - 2018 data from [Statista](#) reports that the majority of pediatricians spend **13-16 minutes** with each patient
- Parents / Youth
 - Provide more targeted information for **relevant conversations**
 - (CO is one of 26 states without a sex ed mandate)
- Education Agencies
 - Implement outreach programs; **individual counseling**

Other Important Considerations

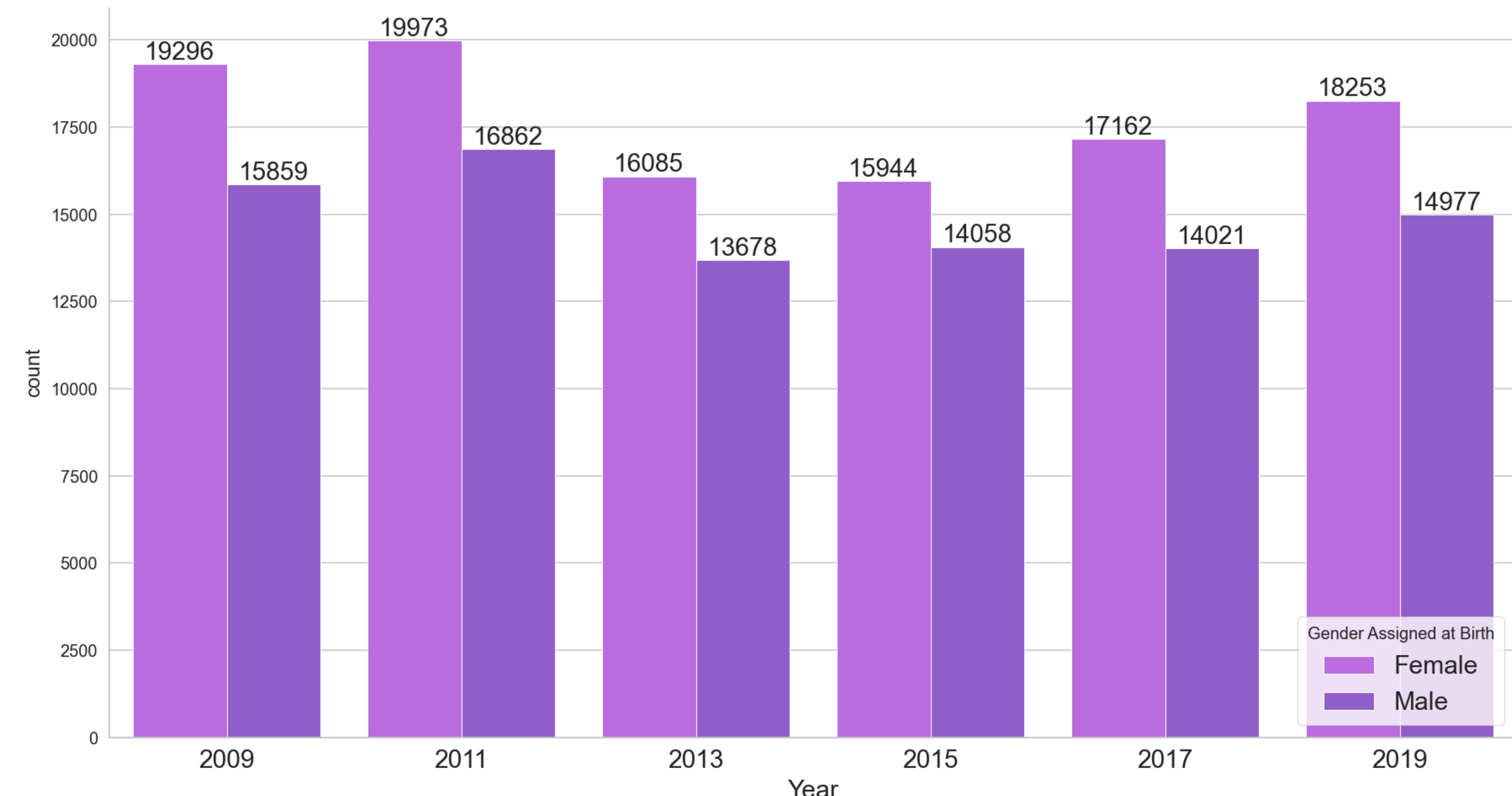
- Insights over Answers
- Discussions over Discipline
- Reality over Determinism

Part 2:

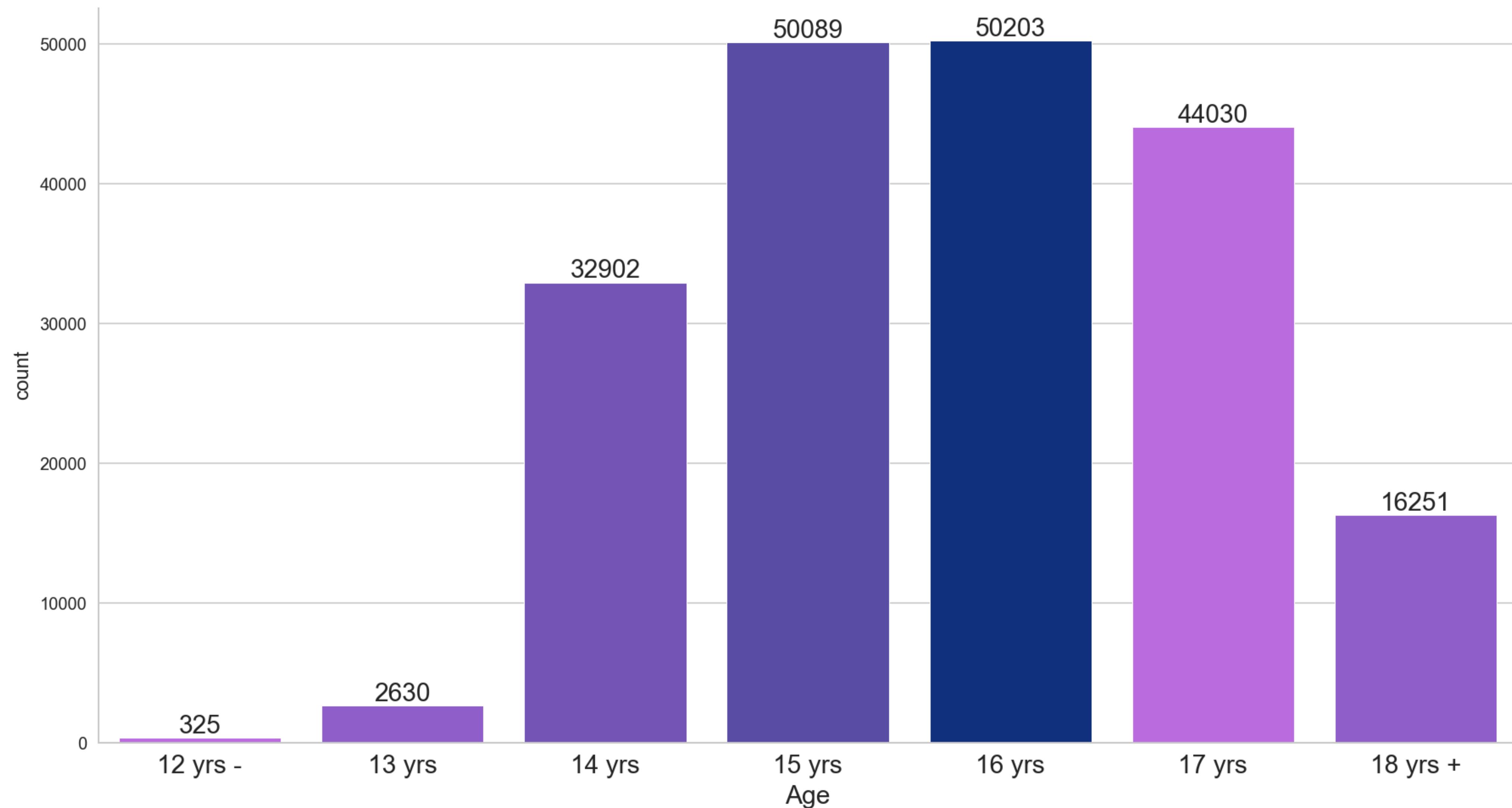
Exploring and Modeling the Data



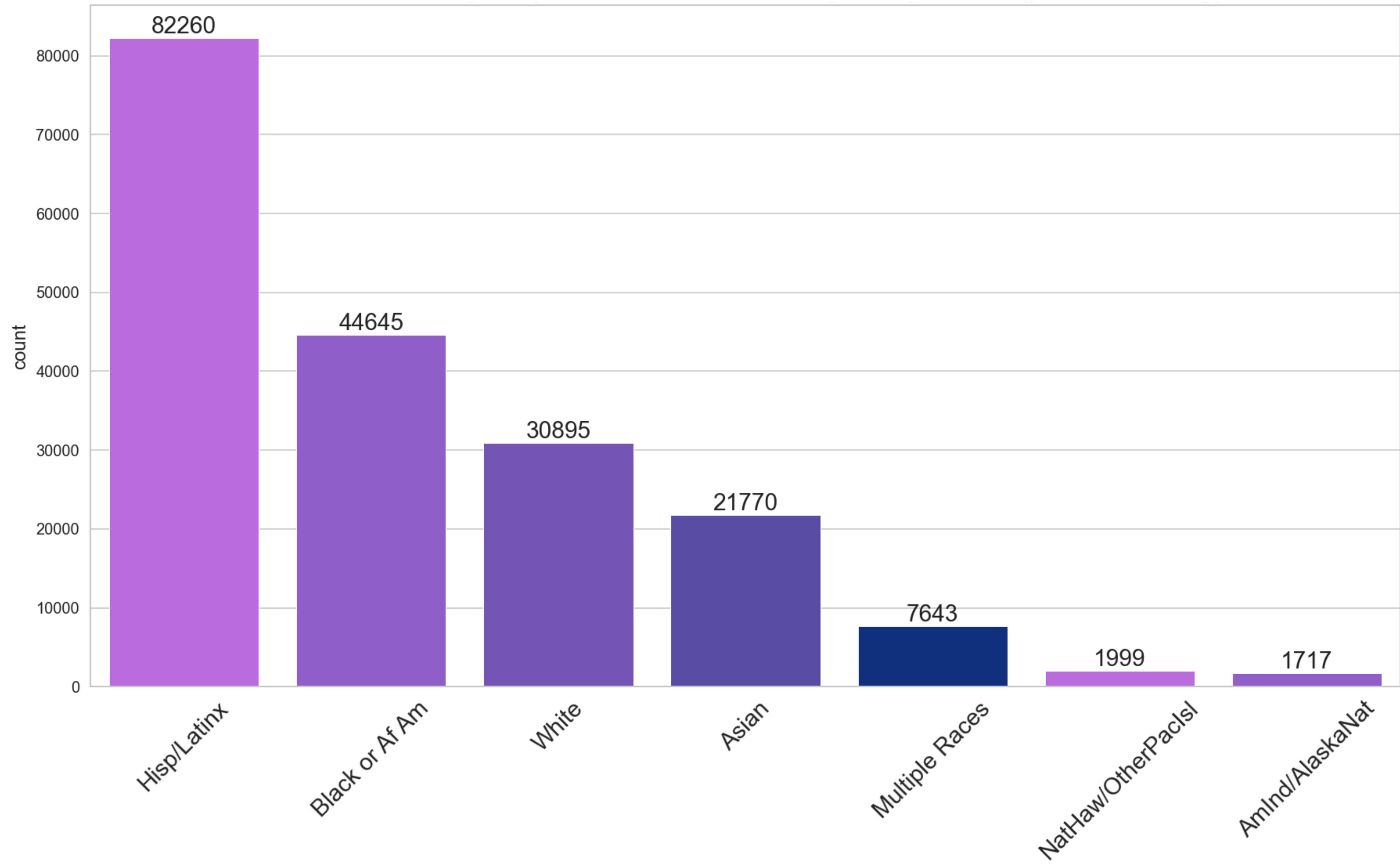
Dataset: Number of Surveys per Year



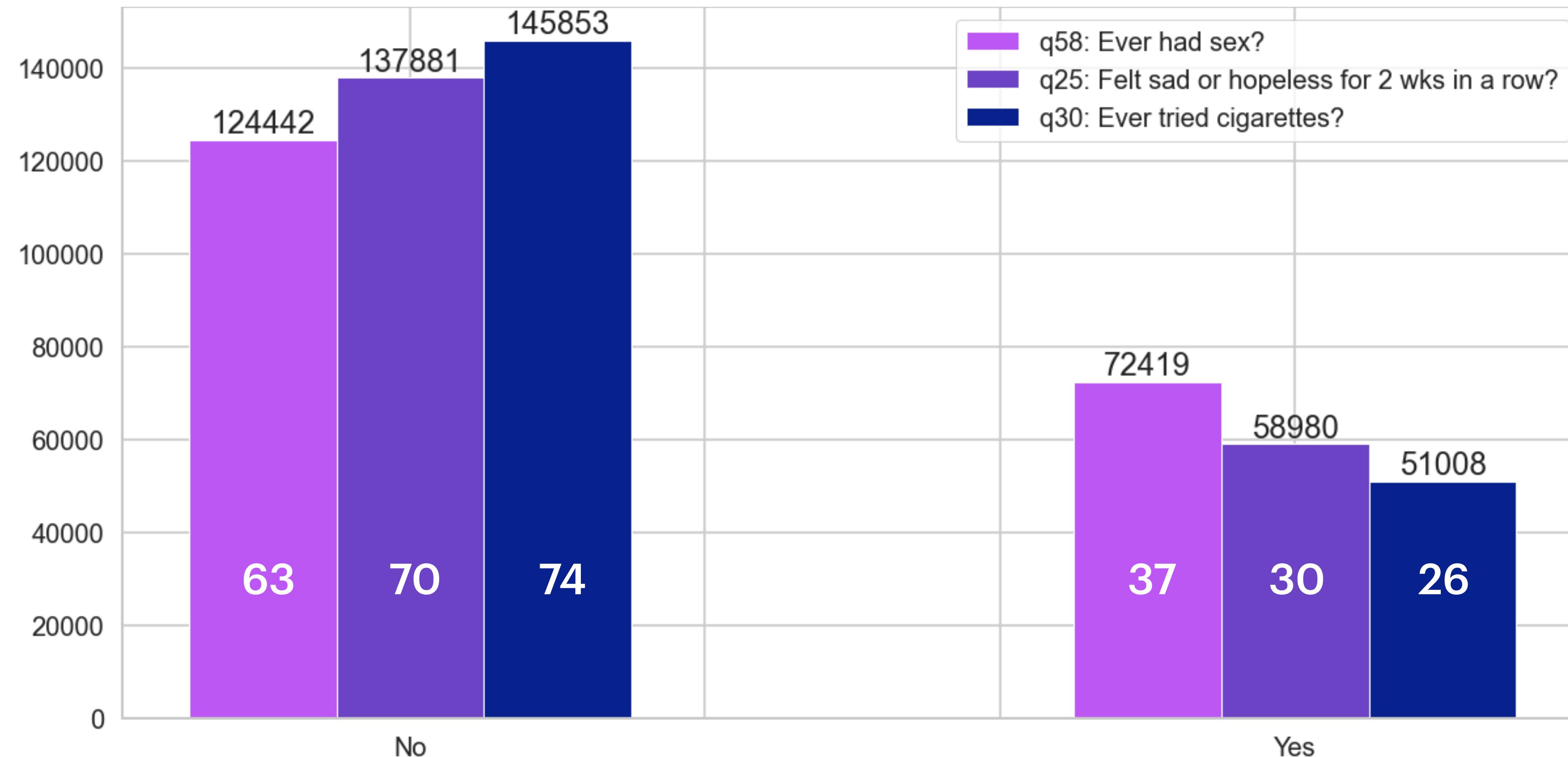
Dataset: Ages of Respondents



Dataset: Race / Ethnicity of Respondents



Distribution of Three Classification Targets



Baseline Metrics: Assume all YES Responses:
Accuracy and Precision = 37% | 30% | 26%
Hamming Loss: 0.69

Experimenting with Algorithms

scikit-multilearn

Trained and Compared
different Multi-Label
Classification Algorithms

- Logistic Regression
- Multi-label kNN
- Binary Relevance
- Classifier Chain
- Gaussian Naive Bayes
- Neural Network
- **Random Forest**

	Q58 (Sex)	Q25 (Sad)	Q30 (Cigarettes)
Baseline	37%	30%	26%
Precision (TP / TP + FP)	66.1%	70.6%	66.0%
Accuracy (TP + TN) / ALL	72.3%	74.2%	76.6%
Hamming Loss (TP + TN) / ALL	0.25 (<i>fraction of labels incorrectly predicted</i>)		

Part 3:

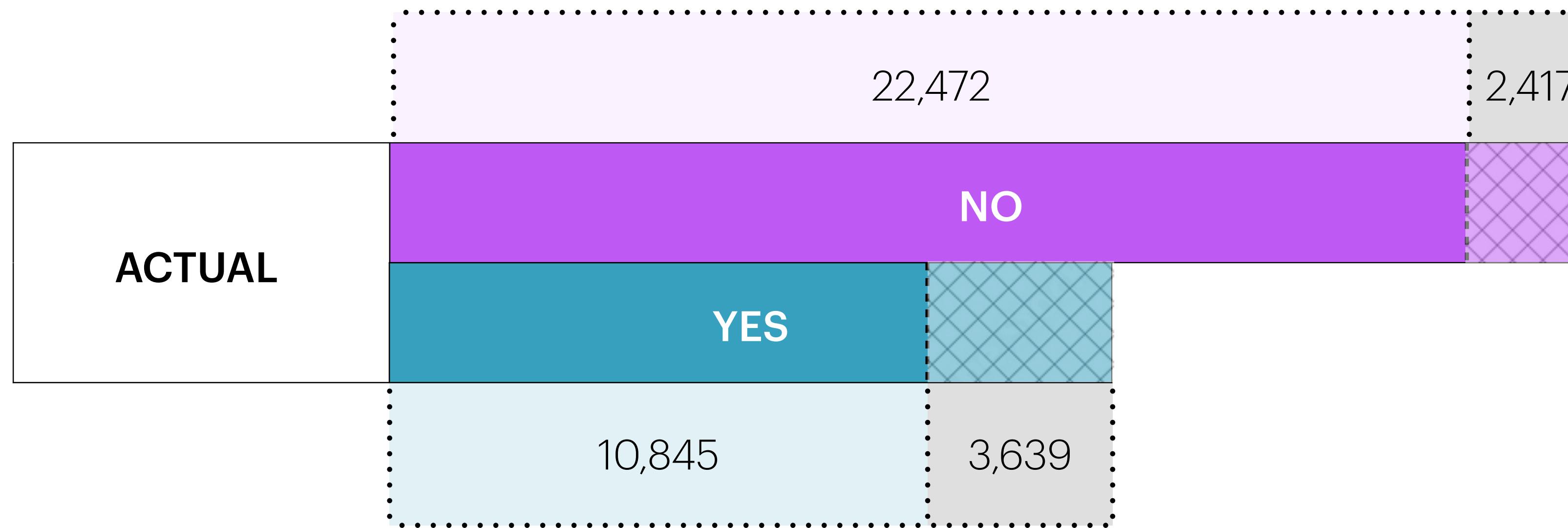
Predictions and Dashboard



How did we do?

Size of Test Set: 39,373

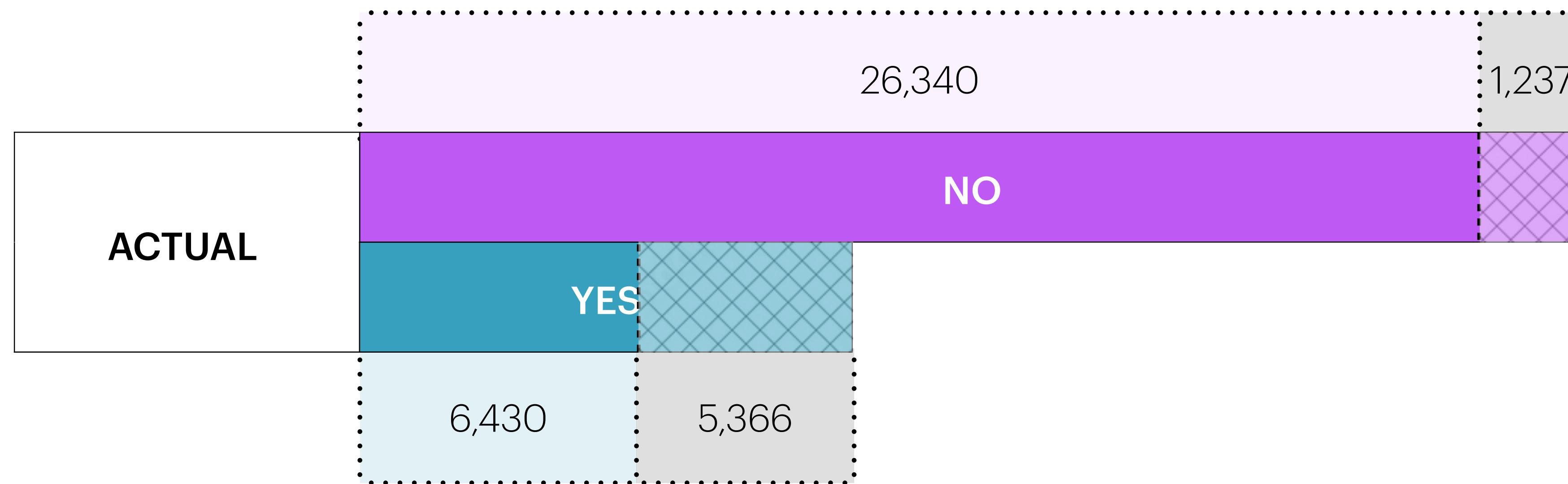
Q58: 'Have you ever had sexual intercourse?'
Baseline Precision = 0%
Model Precision = 81.2%



How did we do?

Size of Test Set: 39,373

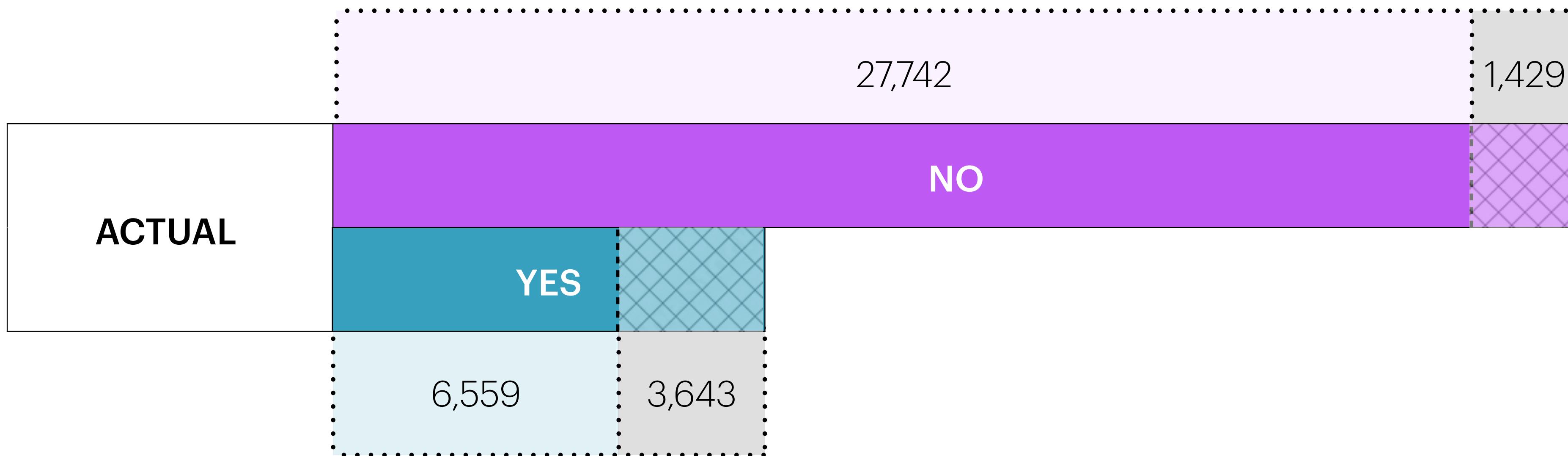
Q25: 'Felt sad or hopeless for 2 weeks in a row'
Baseline Precision = 0%
Model Precision = 83.9%



How did we do?

Size of Test Set: 39,373

Q30: 'Have you ever tried cigarettes?'
Baseline Precision = 0%
Model Precision = 82.1%



Dashboard Demonstration (Local Development)

Predicting Youth Risk Behavior

Helping parents and professionals have relevant discussions with youth.

About the Youth Risk Behavior Survey

About this Dashboard and Project

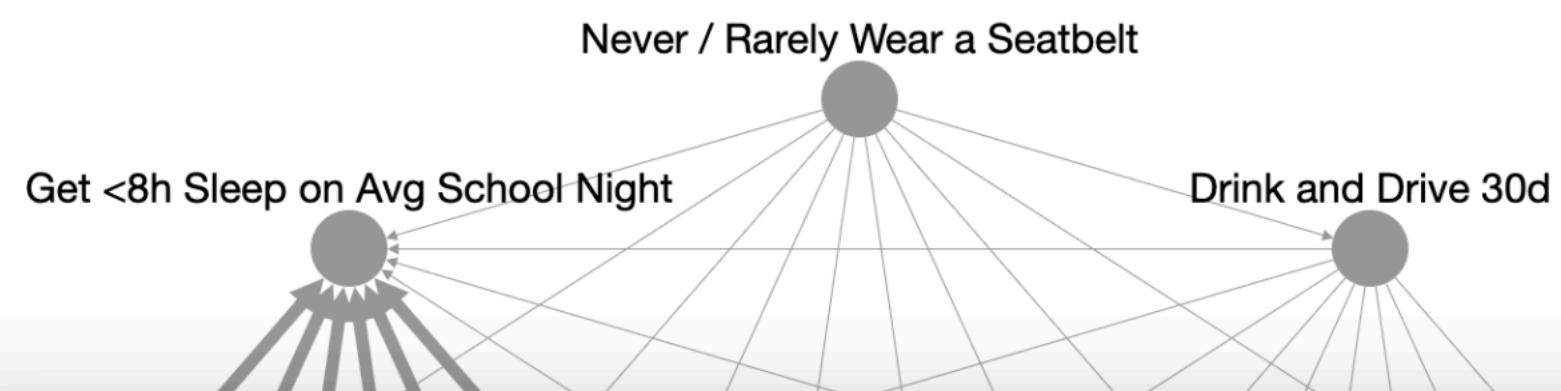
Important Considerations

Explore the Data

Make Predictions

Visualize the Connections

The graph network shown below is based on a subsample of 10,000 surveys from the original dataset, maintaining the proportions of target classifications. Each node is a negative risk behavior and the directional edges between the nodes represent the proportion of survey respondents who answered affirmatively to any of those questions. For example, 31.5 percent of respondents who reported feeling sad or hopeless for more than 2 weeks in a row also reported getting less than 8 hours of sleep on an average school night. This graph may provide some further direction to dashboard users about potentially relevant discussions with youth. Numbers listed are percentages.



Dashboard Screenshot 1: Explore the Data

Predicting Youth Risk Behavior

Helping parents and professionals have relevant discussions with youth.

About the Youth Risk Behavior Survey

About this Dashboard and Project

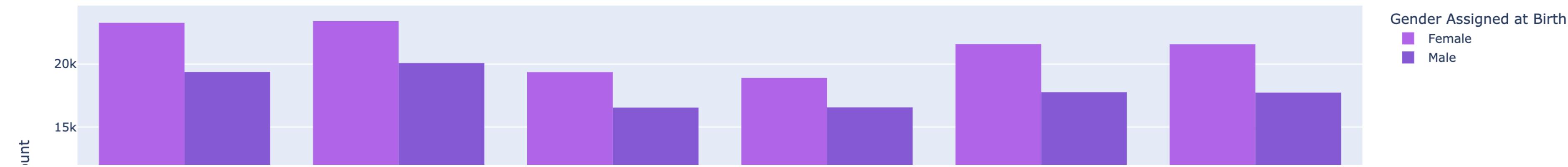
Important Considerations

Explore the Data

Make Predictions

Visualize the Connections

Figure 1: Number of Completed Surveys by Year



Dashboard Screenshot 2: Make Predictions

Explore the Data Make Predictions Visualize the Connections

Change the dropdown menus to predict youth risk behavior based on age, gender assigned at birth, race/ethnicity, age, grade, and other responses from the survey.

How old is this individual?

18 yrs + x ▾

Gender assigned at Birth?

Male x ▾

Race / Ethnicity?

Hisp/Latinx x ▾

What is this person's BMI?

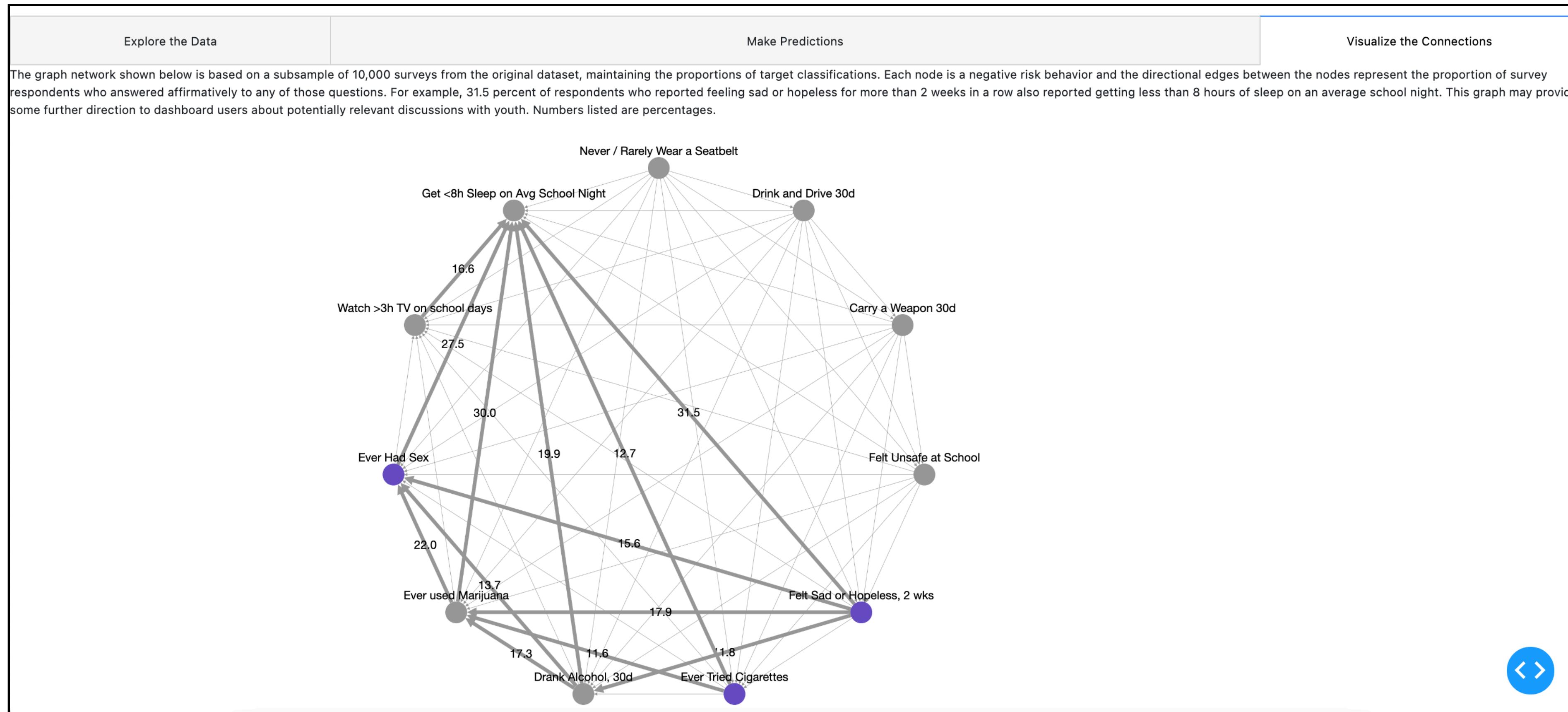
18.5-24.9 x ▾

Grade in School?

12th Grade x ▾

Classification results are: [[1. 0. 0.]]. The probability that this individual has had sex is 55.59 percent, that the individual has been sad / hopeless is 19.99 percent, and that the individual has tried smoking cigarettes 12.21 percent.

Dashboard Screenshot 3: Visualize the Connections



Part 4:

Future Work



Future Work

- Dashboard option to “Find a metro area like mine” (possibly match by size, population, political leanings, etc. to possibly refine predictions with geography)
- Consultation with experts in clinical settings for utility of predictions; dashboard creation for specific clinics based on their own survey data
- Tailor the inputs on the dashboard to different audiences
- Clean up prediction page of app / Live deployment through AWS

Tech Stack



XGBoost





Thank you



Becky Peters, Data Scientist

becky.e.peters@gmail.com

[linkedin.com/in/beckyepeters](https://www.linkedin.com/in/beckyepeters)

[GitHub.com/beckyepeters](https://github.com/beckyepeters)

twitter.com/beckyepeters