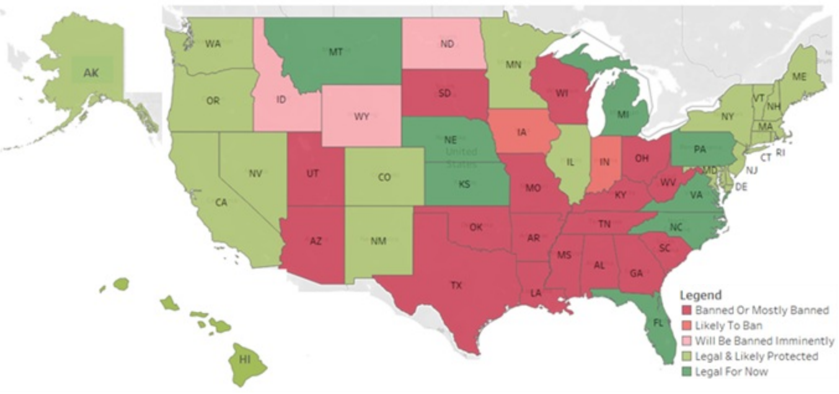
A Tweet Today, A Ban Tomorrow: Abortion Trends by US State

Background

Globally, **less than a third (29%)** of all pregnancies terminate in abortions. Abortions are undertaken for multiple reasons. If the procedure is performed by a **qualified medical professional**, it can be a **safe procedure** with manageable risks. In 2020, the World Health Organization affirmed the importance of safe abortion access through its inclusion in the list of **essential health services**.

When **abortion ban** is in place, there is an associated rise in unsafe abortions, obtained outside the safety of medical facilities, with a much **higher risk of death**. Majority of individuals who are likely to get an abortion in the United States are **women of color or from underrepresented communities**, and limited access to safe abortions has **long-term effects** on **mental health** and **socio-economic wellbeing**.



Data Source: Guttmacher Institute

Objectives

1. To **analyze determinants** of abortion bans in US states
2. To **analyze public opinion** of the recent reversal of the Roe vs Wade ruling

Methods

1. Data Sourcing
   - Guttmacher Institute, Wikipedia, US CDC and The Pew Research Center
   - SurgeAI
   - Webscraped Tweets (see GitHub for more information)
2. Data Wrangling
   - Interpolating Missing Data
   - Vectorizing tweets
3. Feature Selection
   - Correlation Matrix
   - Principal Component Analysis (PCA), Decision Tree
4. Model Technique
   - Logistic Regression, Decision Tree, K Mean w/ PCA for Abortion Classification
   - Logistic Regression, Support Vector Machine, Gradient Boost, Vader

Tools

Jupyter, Google CoLab, Google Docs, Kaggle, Tableau, Python, HackMD, GitHub, Slack

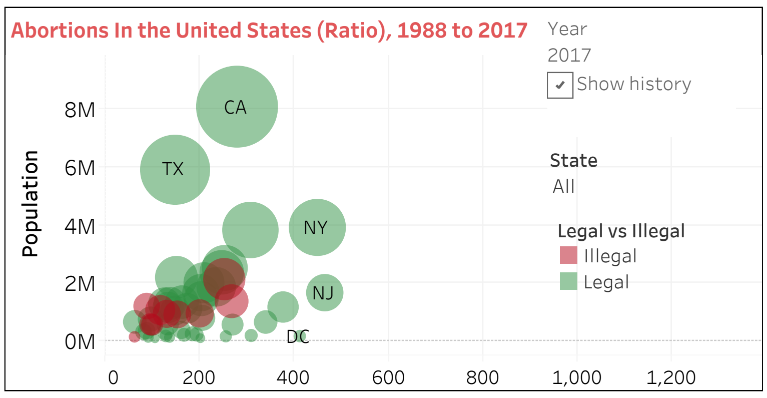
Data Exploration

I. Analyzing the Determinants of Abortion Bans

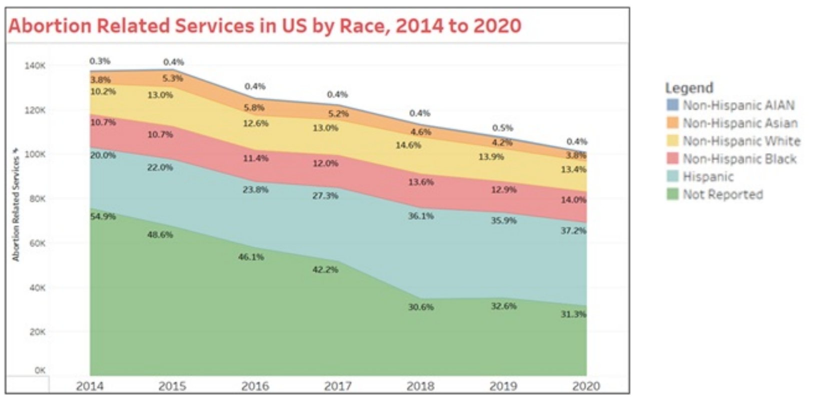
While relatively high in the late 1980s, in recent years, there has been a decline in the overall ratio of abortions to pregnancies. The population refers to the total number of reproductive individuals.

The trend analysis revealed that majority of **ethnicities** are **not reported. Hispanic** and **Non-Hispanic Black women** make up the **largest** number of services. Women **15 to 34 years** of age have the **greatest need** for abortion services in the US.

Abortions In the United States (Ratio), 1988 to 2017



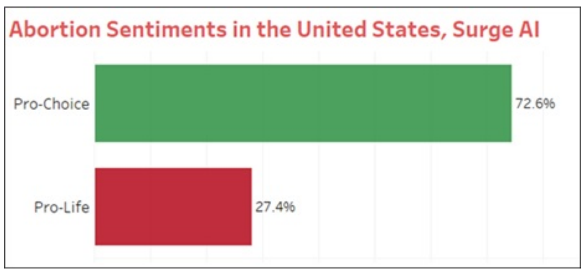
Abortion Related Services in US by Race, 2014 to 2020



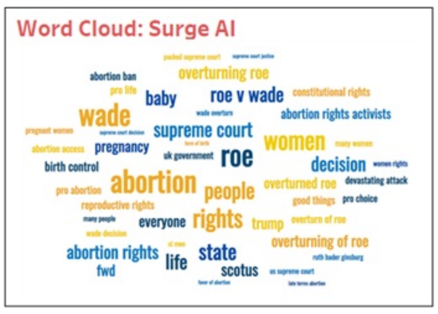
II. Analyzing Public Opinion on Twitter

In the Surge AI dataset, **72.6%** of the tweets were classified as **"pro-choice"**. Frequently occurring keywords in the **SurgeAI tweet** dataset include: **"abortion"** and **"Wade"**, while the **webscraped tweet** dataset had keywords such as: **"abortionrightsarehumanrights"** and **"Biden"**.

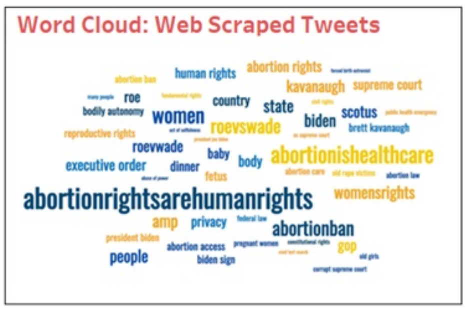
Abortion Sentiments in the United States, Surge AI



Word Cloud: Surge AI



Word Cloud: Web Scraped Tweets



Modeling

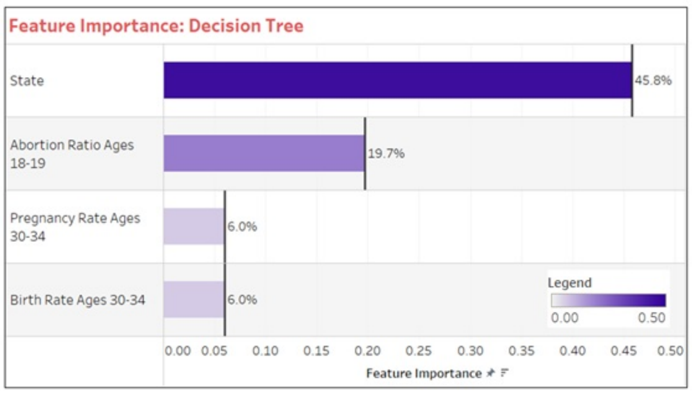
To analyze the determinants of abortion bans, we utilized **classification models** such as **Logistic Regression** and **Decision Tree** to identify features that impact legality, **Principal Component Analysis (PCA)** to reduce dimensionality of factors and **K-Means Clustering** to identify which feature groups are closely correlated. For the analysis of public opinion on Twitter, we aimed to classify tweets as "pro-life" and "pro-choice". Using **Natural Language Processing**, we preprocessed our data using **data wrangling**, assessed **text statistics** and identified **named entities** like **legislations, agencies and institutions**. We also used **Logistic Regression** to predict the probability of classifying tweets as "pro-life" vs "pro-choice", as well as modern machine learning techniques like **Support Vector Machine** and **Gradient Boost (XGBoost)** to accurately classify the tweets. The **VADER Rule** based scoring algorithm was leveraged to better capture sentiment, as "positive", "negative" and "neutral".

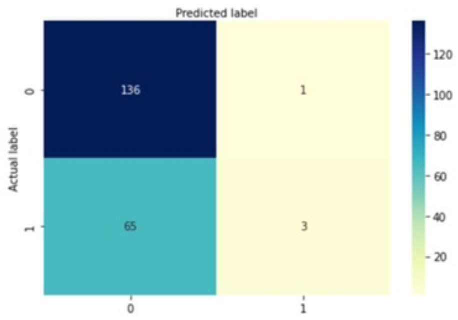
Findings

I. Analyzing the Determinants of Abortion Bans

The abortion dataset had **99 features** including pregnancy rate, birth rate, abortion ratio and miscarriages. The **PCA** and **K-Means** model was used to **reduce overfitting** and model training time by decreasing the number of features. For **Decision Tree** model, **cross validation** methodology was used to **improve accuracy** in prediction. This model correctly identified the likelihood of abortion bans based on selected features with **93% accuracy**. Of these features, the **state of residence** was a dominant factor. The second most important feature was the **abortion ratio** among women aged **18 to 19** years, followed by the **pregancy and birth rates** of women aged **30 to 34** years.

Feature Importance: Decision Tree





II. Analyzing Public Opinion on Twitter

The Surge AI training dataset contained **1025 observations**, while the web scraped dataset contained **50,000 observations**. Our **VADER Rule** based algorithm categorized **90%** of the tweets as **"pro-life"**, and **10%** as **"pro-choice"**. The model had a **72%** probability of **correctly classifying** tweets as **"pro-life"** and **"pro-choice"** on an unseen test data. The model actually identified each individual segment with **68%** accuracy. However, with a **precision score** of **75%** and a **recall** of **4%**, if a tweet was selected at random, there is a **75%** probability that it will be classified as pro-choice, and a **4%** probability that it would be misclassified.

We attribute these findings to a **high degree of variance, data quality and missing data** concerns. Model insights suggest the predominance of "pro-life" sentiment in our dataset, however, with the **high historical demand** for abortion-related services, we recommend **scaling up policy outreach efforts** specifically to **women 18 to 34 years of age**. In addition, **prior to implementation of large scale policy change**, we encourage a **deep analysis of the discrepancies** between **public sentiment** and the **demand for critical health services**.