

# Project title

```
library(tidyverse)
library(skimr)
```

## Data 1

### Problem or question

- Identify the Problem Can we predict the likelihood of H-1B visa approval (initial or continuing) based on factors such as employer characteristics, location, and industry sector?
- Importance of the Topic This question is important because understanding approval patterns can help employers and applicants better prepare and address potential issues in their applications. Especially for international students, this analysis can provide valuable guidance for seeking employment opportunities in the U.S. Knowing which factors impact approval rates enables both students and employers to prepare more effectively, addressing potential challenges and increasing the chances of a successful application. Additionally, these insights can inform policymakers, helping them refine visa approval guidelines by highlighting the impact of various factors such as industry, employer type, and regional tendencies.

- Types of Data / Variables Categorical Variables:

1. Employer: Identifies the employer, which can help capture differences in approval rates by employer.
2. NAICS (Industry Classification): Indicates the industry, which could correlate with approval likelihood.
3. State, City, ZIP: Represent the geographic location of employment, where approval rates may vary regionally.
4. Initial Approval/Denial: Provides the initial application status, useful for predicting Continuing Approval.

Quantitative Variables: 5. Fiscal Year: Helps analyze trends over time, where approval likelihood may change by year.

6. Tax ID: Can be treated as a unique identifier, though it might not directly impact predictions.
7. Initial and Continuing Approval/Denial Counts: Numerical counts that could indicate patterns in approval rates by employer or industry.

- Major Deliverables

1. An exploratory data analysis to display relationships between variables across industries, employers, and regions, providing a clear overview of Continuing Approval patterns.
  2. A machine learning model (e.g., logistic regression, random forest) designed to predict whether an H-1B visa renewal application will be approved.
- **Make the Model Accessible** To make it accessible, we can implement the model through a web interface or API, allowing employers or applicants to input relevant details and receive an approval prediction for continuing applications.

Additionally, we can create an interactive dashboard that visualizes approval patterns across variables like industry and location, making the model results easier to interpret and accessible to a broader audience.

## Introduction and data

- **Source of the Data** The dataset is sourced from the U.S. Citizenship and Immigration Services (USCIS), an official government agency responsible for processing immigration applications and issuing visas.
- **Data Collection Timeline and Method** The data was collected by USCIS between 2009 and 2023. USCIS gathers this information as part of its routine visa processing, recording details of each application, including employer information, industry classification, application approval status, and applicant location. This data is collected through mandatory application forms submitted by employers and applicants, which are then systematically documented for administrative and public reporting purposes.
- **Description of Observations** Each observation represents an H-1B visa application, either an initial application or a renewal (continuing) application. The dataset includes details on the application year, employer, approval or denial status, industry classification (NAICS), and the geographic location of employment.
- **Ethical Concerns** This dataset does not contain direct personally identifiable information (PII) such as names, gender, or race, which helps address some ethical concerns. However, certain ethical issues should still be considered, as details like employer name, industry classification, and geographic location could reveal patterns that may impact specific organizations or regions. For example, publishing approval or denial rates by employer or area might lead to misunderstandings or unfair stigmatization if interpreted without context.

## Glimpse of data

```
# add code here

# This dataset is only for the year 2023, and there are additional datasets for
other years (2009 - 2023) that share the same columns.
h1b_2023 <- read_csv("data/h1b_2023.csv")
```

Rows: 33332 Columns: 11

— Column specification

Delimiter: ","

chr (5): Employer, Tax ID, State, City, ZIP

dbl (6): Fiscal Year, Initial Approval, Initial Denial, Continuing Approval,...

i Use `spec()` to retrieve the full column specification for this data.

i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

`skim(h1b_2023)`







Name	h1b_2023
Number of rows	33332
Number of columns	11
_____	
Column type frequency:	
character	5
numeric	6
_____	
Group variables	None

Table 1: Data summary

### Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Employer	1	1.00	1	130	0	28061	0
Tax ID	94	1.00	4	4	0	9396	0
State	2475	0.93	2	2	0	57	0
City	2475	0.93	2	20	0	3245	0
ZIP	2479	0.93	5	5	0	6483	0

### Variable type: numeric

skim_ variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Fiscal Year	0	1	2023.00	0.00	2023	2023	2023	2023	2023	
Initial Approval	0	1	1.15	9.68	0	0	0	1	944	
Initial Denial	0	1	0.07	0.63	0	0	0	0	39	
Continuing Approval	0	1	4.16	46.54	0	1	1	2	4120	
Continuing Denial	0	1	0.12	1.45	0	0	0	0	149	
NAICS	0	1	51.58	10.95	11	51	54	54	99	

## Questions for reviewers

List specific questions for your peer reviewers and project mentor to answer in giving you feedback on this topic.

1. The dataset covers the years 2009 to 2023, which is a wide range. Would you recommend using all of this data, which would be quite extensive, or perhaps dividing it into 10-year intervals to observe trends more clearly?
2. Does the choice of variables and features in my model adequately capture the factors that could influence H-1B continuing approval, or are there additional variables I should consider?

## Data 2

### Problem or question

- Identify the problem you will solve or the question you will answer

Can we identify characteristics (e.g., age, time period) most associated with higher obesity rates among adults in the U.S.?

- Explain why you think this topic is important.

Identifying the key drivers of obesity is crucial because it allows us to understand the factors driving this major public health issue. By pinpointing specific age groups or time periods that see the greatest increase in obesity rates, we can develop more targeted interventions, tailor public health campaigns, and allocate resources more effectively. This approach not only aids in slowing

down the rising obesity trend but also helps prevent related chronic conditions, such as diabetes and heart disease, which can significantly impact individuals' quality of life.

- Identify the types of data/variables you will use.

Categorical Variables 1. Indicator: the BMI category (e.g., normal weight, overweight, obesity). 2. Panel: specific BMI ranges (e.g., "Normal weight (BMI from 18.5 to 24.9)"); help understand which BMI categories are being analyzed. 3. Stub Name / Stub Label: the population subgroup (e.g., Sex, Race), allowing for segment-specific analysis. 4. Age / Age Num: the age group of the population being studied. Numerical Variables 1. Year / Year Num: the time period when the data was collected. 2. Estimate: the percentage of the population within each BMI category, providing a direct measure of obesity rates to compare across groups.

- State the major deliverable(s) you will create to solve this problem/answer this question.
  1. An exploratory data analysis report with visualizations and statistical summaries that outline initial patterns, trends, and distributions of obesity rates across different age/race/sex groups and time periods.
  2. A machine learning model that identifies and ranks the characteristics most associated with higher obesity rates. This model might also include metrics like accuracy, precision, recall, and feature importance to evaluate its performance.
- How will you make the model usable?

To make the model usable, we plan to create an interactive web application using Shiny. The app will provide visualizations of obesity trends across different age groups, time periods, and demographic characteristics, etc., making the insights easy to understand and explore. Users will be able to adjust parameters to see how obesity rates vary, helping to identify high-risk groups and trends over time. This allows anyone, with or without technical expertise, to explore and interact with the model's predictions directly through a web browser.

## Introduction and data

If you are using a dataset:

- Identify the source of the data.

The data was derived from Data.gov, the U.S. government's open-data portal. Title: Normal weight, overweight, and obesity among adults aged 20 and over, by selected characteristics: United States Link: <https://catalog.data.gov/dataset/normal-weight-overweight-and-obesity-among-adults-aged-20-and-over-by-selected-characteris-8e2b1>

- State when and how it was originally collected (by the original data curator, not necessarily how you found the data).

The data for this project was originally collected through the National Health and Nutrition Examination Survey (NHANES), curated by the National Center for Health Statistics (NCHS). NHANES is a comprehensive survey designed to assess the health and nutritional status of adults and children in the United States. It has been conducted in various phases, starting with NHANES III from 1988 to 1994, followed by continuous data collection from 1999 onward. NHANES com-

bines in-home interviews and clinical examinations conducted in mobile centers, collecting data on chronic conditions, risk factors like obesity, and other health indicators. The survey employs a stratified, multistage probability sampling method to ensure national representation, with over-sampling of specific populations (e.g., low-income groups, racial/ethnic minorities) to enhance accuracy. This data is regularly updated to reflect the evolving health landscape, and its results are adjusted for response rates and population estimates, ensuring its reliability for public health research.

- Write a brief description of the observations.

The dataset comprises observations from the National Health and Nutrition Examination Survey (NHANES), focusing on adults aged 20 and over in the United States. It tracks BMI categories (normal weight, overweight, and obesity) across different time periods, highlighting changes in obesity rates over several decades. Each observation includes information on the proportion of the population within each BMI category, stratified by age, race and gender groups and survey years, along with measures like standard error to indicate estimate precision.

- Address ethical concerns about the data, if any.

NHANES uses stratified sampling to oversample specific populations, which helps ensure representation. However, there is still a risk of bias if certain groups (e.g., undocumented individuals or those unable to participate in surveys) are systematically excluded, potentially limiting the generalizability of the results.

## Glimpse of data

```
# add code here
weight_data <- read_csv("data/weight_20_over_US.csv")
```

Rows: 3360 Columns: 16

Column	specification
--------	---------------

Delimiter: ","

chr (8): INDICATOR, PANEL, UNIT, STUB\_NAME, STUB\_LABEL, YEAR, AGE, FLAG

dbl (8): PANEL\_NUM, UNIT\_NUM, STUB\_NAME\_NUM, STUB\_LABEL\_NUM, YEAR\_NUM, AGE\_N...

i Use `spec()` to retrieve the full column specification for this data.

i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

```
skim(weight_data)
```

Name	weight_data
Number of rows	3360
Number of columns	16
_____	
Column type frequency:	
character	8
numeric	8
_____	
Group variables	None

Table 4: Data summary

**Variable type: character**

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
INDICATOR	0	1.00	68	68	0	1	0
PANEL	0	1.00	37	55	0	6	0
UNIT	0	1.00	28	35	0	2	0
STUB_NAME	0	1.00	3	32	0	6	0
STUB_LABEL	0	1.00	4	62	0	34	0
YEAR	0	1.00	9	9	0	10	0
AGE	0	1.00	11	17	0	7	0
FLAG	2516	0.25	1	3	0	3	0

**Variable type: numeric**

skim_ variable	n_ missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
PANEL_NUM	0	1.00	3.50	1.71	1.0	2.00	3.50	5.00	6.00	
UNIT_NUM	0	1.00	1.61	0.49	1.0	1.00	2.00	2.00	2.00	
STUB_NAME_NUM	0	1.00	4.14	1.33	1.0	3.00	4.00	5.00	6.00	
STUB_LABEL_NUM	0	1.00	3.97	1.47	1.1	3.12	3.24	5.32	6.26	
YEAR_NUM	0	1.00	5.50	2.87	1.0	3.00	5.50	8.00	10.00	
AGE_NUM	0	1.00	1.07	0.16	1.0	1.00	1.00	1.00	1.60	
ESTIMATE	461	0.86	28.50	22.05	0.3	10.00	22.70	38.05	85.90	
SE	461	0.86	1.26	0.57	0.2	0.80	1.20	1.60	3.50	

## Questions for reviewers

List specific questions for your peer reviewers and project mentor to answer in giving you feedback on this topic.

1. Are there recommended methods for handling missing data in this dataset, especially given its stratified sampling design?
2. How should we handle the standard error (SE) column? Should it be used in weighting the observations or as part of the analysis for precision?
3. How should we treat time as a variable? Should we consider it for time series analysis, or would a static comparison across different years be more insightful?

## Data 3

### Problem or question

- Identify the problem you will solve or the question you will answer Problem: Analyze the trends and factors associated with adult arrests in New York State since 1970. Specifically, determine which types of offenses (e.g., drug related, violent, DWI) have seen significant changes over time and whether certain counties have disproportionately high arrest rates for specific types of offenses.
- Explain why you think this topic is important. Understanding arrest trends is important for making informed policy decisions, efficiently allocating resources, and proactively addressing public safety concerns. By recognizing patterns in crime data, law enforcement can better assess current policies' effectiveness and improve strategies for crime prevention.
- Identify the types of data/variables you will use. Categorical: County, Year (as a factor if modeling by category) Quantitative: Total Arrests, Felony Total, Drug Felony, Violent Felony, DWI Felony, and various Misdemeanor Types.



- State the major deliverable(s) you will create to solve this problem/answer this question.
1. EDA: Conduct an EDA to uncover trends and relationships between variables such as the year, county, and types of offenses (e.g., drug-related, violent, DWI, etc.). Use visualizations like time series plots to analyze trends over time, geographic heat maps to compare arrest rates across counties, and bar charts to highlight differences between offense categories.
  2. Develop a classification model (e.g., logistic regression, random forest) to predict the likelihood of a specific type of offense (such as a violent felony, drug felony, or DWI) based on factors like location (county), year, and other arrest characteristics.
  3. Evaluate the model's performance using metrics such as accuracy, F1 score, or AUC-ROC depends on the model chosen. Provide a summary of model insights, such as identifying counties with the highest probability of certain offenses or projecting changes in arrest trends. These insights could guide future policy decisions.
  4. To make this model accessible, it can be deployed as an interactive web application (via Shiny) where users select a county and year to view predictions and trends. Alternatively, the model could be accessible as a deployable API for integration into other systems.

## Introduction and data

If you are using a dataset:

- Identify the source of the data. The dataset is provided by the New York State Division of Criminal Justice Services, available through NY Open Data.
- State when and how it was originally collected (by the original data curator, not necessarily how you found the data). Data is collected by the New York State Division of Criminal Justice Services and is updated annually from law enforcement agencies for fingerprintable offenses, ensuring a comprehensive dataset on adult arrests across all New York counties.
- Write a brief description of the observations. Each row represents an annual count of adult arrests by county, categorized by offense type (e.g., violent felonies, drug related felonies, misdemeanors). The dataset spans from 1970 to the present, providing a robust timeline for trend analysis.
- Address ethical concerns about the data, if any. The data excludes sensitive information such as personal identifiers(race,job,etc), thereby respecting privacy concerns. However, it's essential to consider any potential biases in data reporting, as underreporting in certain areas could affect trend reliability.

## Glimpse of data

```
library(readr)
library(skimr)
# add code here
arrest_data <- read_csv("data/
Adult_Arrests_18_and_Older_by_County___Beginning_1970_20241030.csv")
```

Rows: 3348 Columns: 13

— Column specification

Delimiter: ","

chr (1): County

dbl (12): Year, Total, Felony Total, Drug Felony, Violent Felony, DWI Felony...

i Use `spec()` to retrieve the full column specification for this data.

i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

```
skim(arrest_data)
```

Name	arrest_data
Number of rows	3348
Number of columns	13
_____	
Column type frequency:	
character	1
numeric	12
_____	
Group variables	None

Table 7: Data summary

#### Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
County	0	1	4	12	0	62	0

#### Variable type: numeric

skim_ vari- able	n_ miss- ing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Year	0	1	1996.50	15.59	1970	1983.00	1996.5	2010.00	2023	
Total	0	1	6556.44	14679.17	22	868.75	1516.0	4198.00	107786	
Felony Total	0	1	2250.48	5674.83	6	192.00	371.0	1105.00	44632	
Drug Felony	0	1	488.33	1652.59	0	19.00	50.0	184.25	17442	
Vio- lent Felony	0	1	688.57	1936.50	0	39.00	78.0	265.00	16217	
DWIFelony	0	1	66.80	88.57	0	17.00	37.0	76.00	613	
Other Felony	0	1	1006.78	2295.38	1	105.00	201.0	584.00	15467	
Mis- demeanor Total	0	1	4305.96	9356.79	7	663.50	1176.0	3033.25	73365	
Drug Mis- demeanor	0	1	850.37	3030.88	0	26.00	73.0	252.25	29471	
DWI Mis- demeanor	0	1	611.81	839.82	0	176.00	326.0	659.25	8954	
Prop- erty Mis- demeanor	0	1	1314.10	3188.93	0	153.75	323.0	837.25	33334	
Other Mis- demeanor	0	1	1529.68	3136.90	1	241.75	451.0	1165.00	24875	

### Questions for Reviewers

1. Do the stated problem and objectives seem achievable and relevant to the dataset provided?
2. Are the chosen variables and deliverables appropriate for addressing the problem effectively?
3. Are there any other data aspects, trends, or comparisons you think would add value to this analysis?
4. Is there any additional information or data transformation needed to make the findings more impactful?

5. Are the ethical considerations adequately addressed, or is there anything else that should be mentioned?