Bayesian Analysis of H1B Visa Approval Patterns(2022-2024)

1. Introduction

The H1B visa program is crucial for U.S. employers hiring international talent in specialty occupations. While previous studies examined aggregate approval rates, limited research has applied advanced statistical methods to understand H1B petition outcomes. Understanding these patterns is essential for both employers and job seekers in the visa application process:

For Employers:

- Strategic planning for international hiring
- Understanding success factors in applications
- Optimizing approaches based on patterns

For Job Seekers:

- Informed decision-making in job search
- Better evaluation of potential employers
- Understanding impact of choices on approval likelihood

To address these needs, this study employs hierarchical Bayesian modeling to analyze approval patterns across multiple dimensions.

Our research objectives are to:

1. Quantify the relationship between employer application volume and approval rates
2. Identify industry sectors with consistently higher approval rates
3. Analyze geographic patterns in visa petition outcomes
4. Compare patterns between initial and continuing applications
5. Develop probabilistic insights for optimizing application strategies

2. Data Description and Preprocessing

2.1 Data Source and Structure

Our analysis utilizes USCIS H-1B Employer Data Hub (2022-2024) data, representing post-lottery petition outcomes. Key aspects include:

- Population: 102,865 companies
- Mean employer approval rate: 96.3%
- Mean employer denial rate: 0.037%
- Quarterly updates from USCIS
- Employer-based summarization

2.2 Variable Definitions

1. **Response Variables**:

- Let Y_{1i} represent the initial approval rate for employer i:

  Y_{1i} = Initial Approvals / (Initial Approvals + Initial Denials)
- Let Y_{2i} represent the continuing approval rate for employer i:

  Y_{2i} = Continuing Approvals / (Continuing Approvals + Continuing Denials)

2. **Predictor Variables**:
   - X_{1i}: Application volume category (1-10, 11-30, 31-100, 100+)
   - X_{2i}: Industry sector (NAICS code)
   - X_{3i}: Geographic location (State)
3. **Control Variables**:
   - Total application volume
   - Employer identification metrics

2.3 Data Cleaning and Preprocessing

The data preparation process involved several key steps to ensure analysis quality:

1. **Missing Value Treatment**:
   - Industry codes: Filled missing values with 'Unknown' category
   - Geographic information: Used city information to fill missing state data
2. **Volume Categorization**:
   - Four categories based on application counts (1-10, 11-30, 31-100, 100+)
3. **Rate Calculations**:
   - Standardized approval rate calculations across all employers
   - Separate calculations for initial and continuing applications
   - Treatment of zero denominator cases

2.4 Exploratory Data Analysis

2.4.1 Volume Distribution Analysis

Analysis of application volumes reveals the following distribution across categories:

- 91.39% of employers fall in the 1-10 applications category
- 5.89% in the 11-30 category
- 1.99% in the 31-100 category
- 0.73% in the 100+ category

2.4.2 Industry Sector Patterns

Top 3 Industries by Initial Approval Rate:

1. 61 - Educational Services (67.7% approval rate)
2. 54 - Professional, Scientific, and Technical Services (56.8%)
3. 81 - Other Services except Public Administration (53.7%)

Top 3 Industries by Continuing Approval Rate::

1. 52 - Finance and Insurance (77.9%)
2. 22 - Utilities (76.4%)
3. 51 - Information (74.2%)

2.4.3 Geographic Distribution

Initial Applications:

- Highest approval rates: New Jersey (~55%), Texas (~54%), Georgia (~53%)
- Highest volume states: California (68,409), Texas (58,744), New Jersey (34,684)

Continuing Applications:

- Highest approval rates: Michigan (~73%), New Jersey (~72%), Virginia (~71%)
- Highest volume states: California (189,869), Texas (94,452), New Jersey (69,981)

2.4.4 Initial vs. Continuing Application Comparison

1-10 applications:

- Initial: 48.39% approval rate
- Continuing: 64.98% approval rate

11-30 applications:

- Initial: 89.03% approval rate
- Continuing: 93.98% approval rate

31-100 applications:

- Initial: 95.76% approval rate
- Continuing: 96.88% approval rate

100+ applications:

- Initial: 98.30% approval rate
- Continuing: 98.09% approval rate

3. Methodology

3.1 Hierarchical Bayesian Framework

3.1.1 Model Structure

Our hierarchical Bayesian model is structured to capture the natural hierarchy in H1B visa application processes, with dependencies flowing from high-level factors to specific outcomes.

Hierarchical Structure:

1. **Top Level: Application Volume Data**
   - Serves as the primary conditioning factor
   - Influences both industry and geographic effects
   - Categories: 1-10, 11-30, 31-100, 100+ applications
2. **Second Level: Parallel Effects**
   - **Industry Effects**
     - Nested within volume categories
     - Captures sector-specific variation
     - Allows for industry-volume interactions
   - **Geographic Effects**
     - Also nested within volume categories

        - Accounts for state-level variations
        - Captures regional patterns
3. **Combined Effects Level**
    - Integrates volume, industry, and geographic influences
    - Accounts for interactions between factors
    - Forms the basis for success probability estimation
4. **Success Probability Level**
    - Separate parameters for initial and continuing applications
    - Transforms combined effects through logit function
    - Produces final probability estimates

The mathematical formulation for this structure is:

- Initial Applications Model(For employer n):

$$logit(p_{init,\, n}) = \beta_{vol}[vol[n]] + \beta_{ind,\, vol[n]}[ind[n]] + \beta_{state,\, vol[n]}[state[n]\}$$

$$initial\_success[n] \sim Binomial\_logit(initial\_total[n],\ logit(p_{init,\, n}))$$

- Continuing Applications Model(For employer n):

$$logit(p_{cont,\, n}) = \gamma_{vol}[vol[n]] + \gamma_{ind,\, vol[n]}[ind[n]] + \gamma_{state,\, vol[n]}[state[n]\}$$

$$continuing\_success[n] \sim Binomial\_logit(continuing\_total[n],\ logit(p_{cont,\, n}))$$

This hierarchical structure allows us to:

1. Model dependencies between different levels of factors
2. Account for varying effects across volume categories
3. Capture both direct and interaction effects
4. Maintain interpretability of results
5. Provide appropriate uncertainty quantification at each level

3.1.2 Prior Specifications and Justification

1. Standard Normal Prior for Raw Parameters

The choice of N(0,1) for raw parameters is justified by:

- Modeling on the logit scale, where N(0,1) provides:
    - Zero-centered distribution avoiding directional bias
    - Coverage of realistic approval rate variations within ±2 standard deviations
    - Natural regularization of effect sizes
    - Appropriate representation of the underlying approval rate distribution

For both initial and continuing applications:

$vol\_effect\_ * \_raw \sim N(0,\ 1)$

$ind\_effect\_ * \_raw[v] \sim N(0,\ 1)$

$state\_effect\_ * \_raw\ [v] \sim N(0,\ 1)$

2. Half-Normal Prior for Scale Parameters

The choice of $N^{+}(0,1)$ for scale parameters:

- $σ\_vol\_ * \sim N^{+}(0,1)$
- $σ\_ind\_ * \sim N^{+}(0,1)$
- $σ\_state\_ * \sim N^{+}(0,1)$

This specification is appropriate because:

- It allows natural adaptation to data through multiplicative structure
- Provides additional flexibility in posterior estimation
- Ensures computational stability in MCMC sampling
- Reflects the positive nature of variance parameters

3. Contextual Appropriateness

Our prior choices are particularly suitable for H1B visa analysis because:

- They appropriately cover typical H1B approval rates (40-95% range)
- The logit transformation effectively captures the rarity of extreme approval rates
- The hierarchical structure accounts for nested effects of volume, industry, and location
- The scale parameters allow for different levels of variation across factors

3.2 Statistical Inference Approach

3.2.1 MCMC Implementation

Our implementation uses Stan with the following specifications:

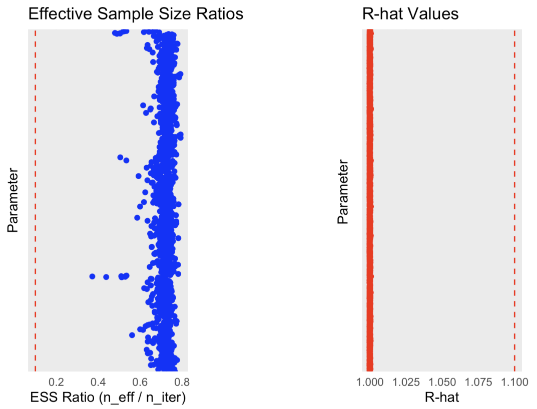
- Chains: 4
- Iterations: 15,000 per chain
- Warm-up: 7,500 iterations
- Thinning: 3
- Additional control parameters:
  - adapt_delta = 0.99999
  - max_treedepth = 20
  - stepsize = 0.001

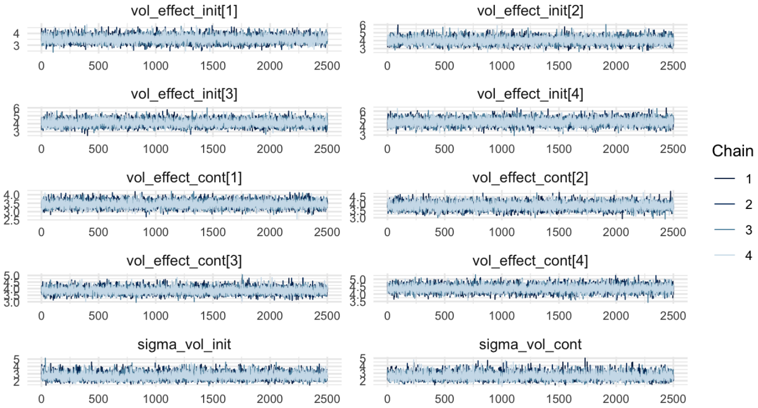
3.2.2 Convergence Diagnostics

We assessed convergence through multiple metrics:

- R-hat values (target < 1.1)
- Effective sample size ratios
- Trace plot examination
- Posterior predictive checks

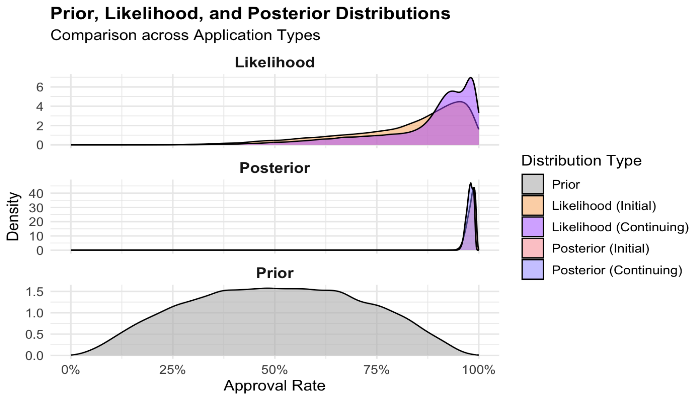
Convergence diagnostic (Trace plots for key parameters)





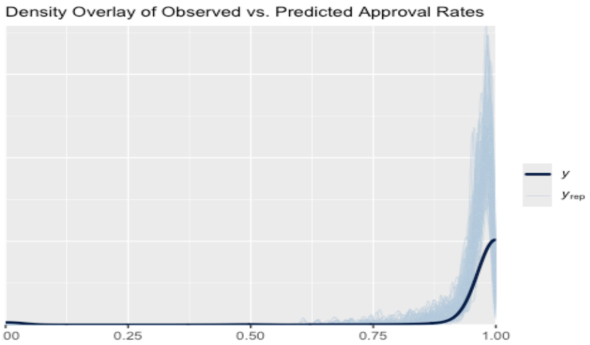
3.2.3 Posterior Predictive Checks

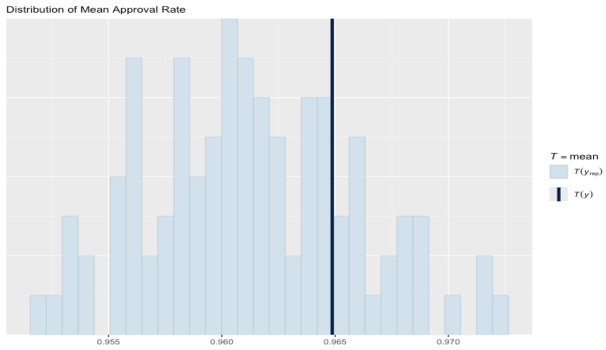
1. Distribution Comparisons:



- Prior distributions showing our initial beliefs
- Likelihood representing the data evidence
- Posterior distributions showing the updated beliefs
- Separate visualizations for initial and continuing applications

2. Posterior Predictive Checks:





Density Overlay of Observed vs. Predicted Approval Rates:

- The observed approval rates (solid dark line) closely align with the predicted approval rates (light blue density).
- This suggests that the model effectively captures the observed data's distribution, particularly the concentration of high approval rates near 1.

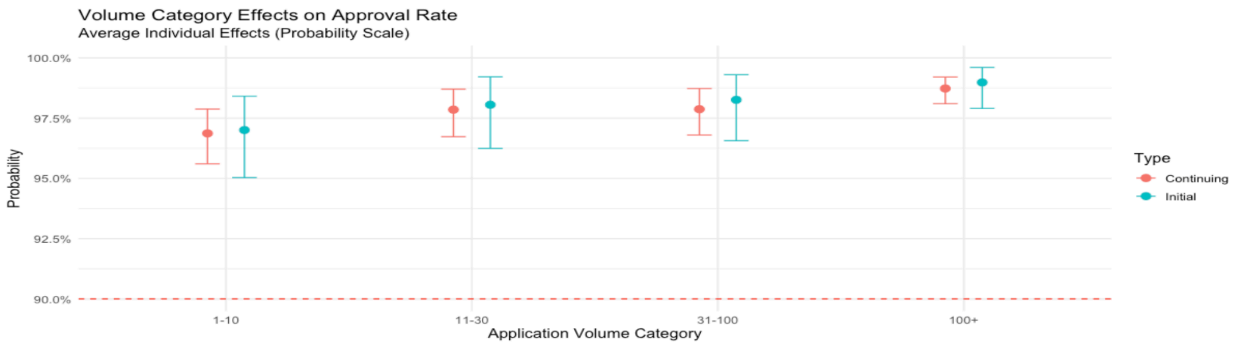
Distribution of Mean Approval Rate:

- The observed mean approval rate (dark vertical line) falls well within the range of predicted means from posterior samples (light blue histogram).
- This alignment confirms that the model's predictions are consistent with the observed mean approval rates.

4. Results and Analysis

4.1 Volume Effect Analysis (Volume Category Comparisons)

Our hierarchical Bayesian model reveals clear patterns in H1B approval rates across different application volume categories:


Volume Category Effects on Approval Rate
Average Individual Effects (Probability Scale)

Overall Trend: Strong positive correlation between application volume and approval rates

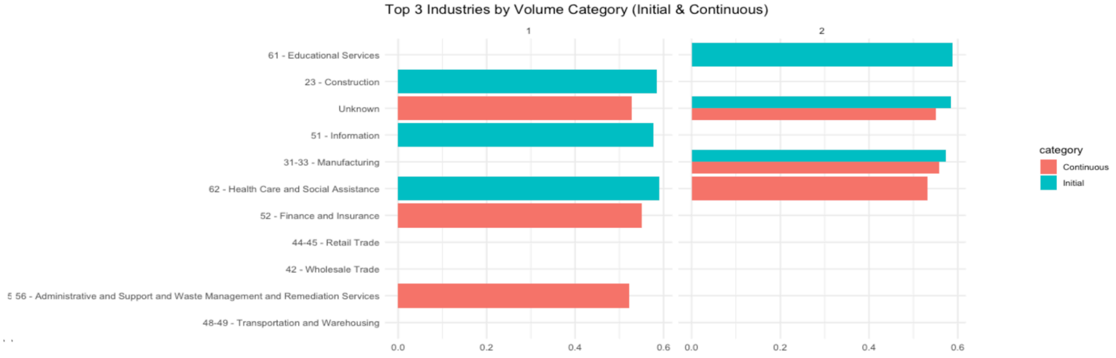
Initial vs Continuing Applications:

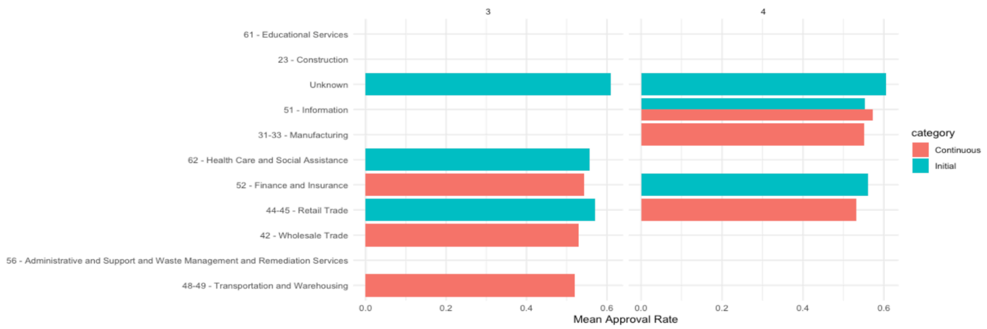
- 100+ category shows highest approval rates (Initial: 98.30% ± 0.5%, Continuing: 98.09% ± 0.4%)
- Largest gap in 1-10 category (Initial: 48.39%, Continuing: 64.98%)
- Differences diminish as volume increases

Credible Intervals: Narrower intervals for higher volume categories, indicating more reliable estimates

4.2 Industry Effect Analysis (Volume-Industry Interactions)

Industry effects show distinct patterns:


Top 3 Industries by Volume Category (Initial & Continuous)



Initial Applications by Volume Category:

- Volume Category 1 (1-10): Healthcare (62), Construction (23), and Information (51) lead
- Volume Category 2 (11-30): Educational Services (61) shows highest rate, followed by Manufacturing (31-33)
- Volume Category 3 (31-100): Retail Trade (44-45) and Healthcare (62) demonstrate strong performance
- Volume Category 4 (100+) : Finance (52) and Information (51) sectors dominate

Continuing Applications by Volume Category:

- Volume Category 1 (1-10): Finance (52) and Administrative Services (56) show highest rates
- Volume Category 2 (11-30): Manufacturing (31-33) and Healthcare (62) maintain strong performance
- Volume Category 3 (31-100): Finance (52), Wholesale Trade (42), and Transportation (48-49) lead
- Volume Category 4 (100+): Information (51), Manufacturing (31-33), and Retail (44-45) show highest rates

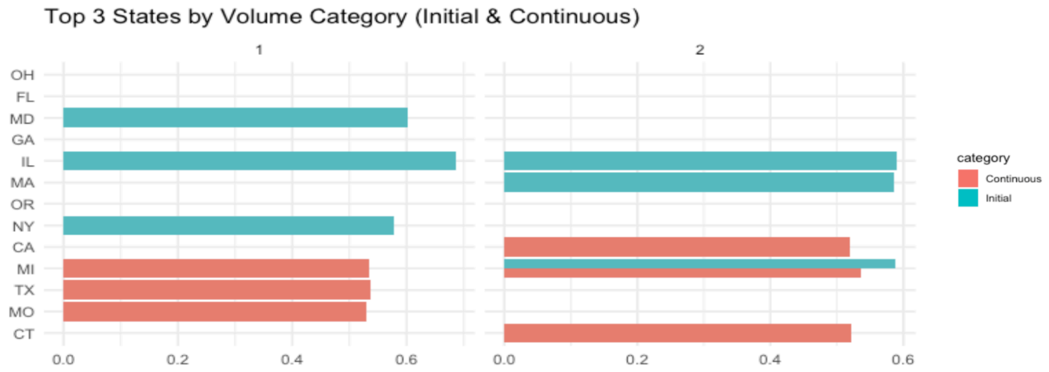
Key Industry Patterns:

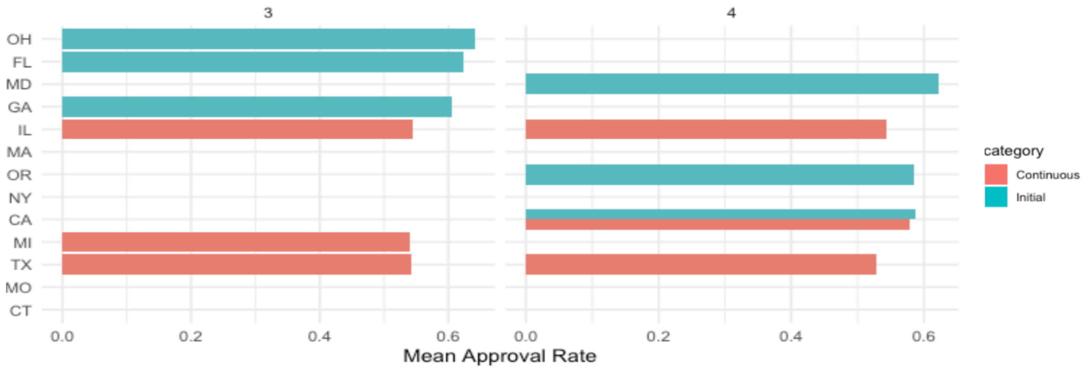
- Information and Manufacturing sectors show consistent performance across categories
- Healthcare strong in initial applications but varies in continuing
- Finance sector performs well in continuing applications
- Educational Services shows high approval rates in higher volume categories

Note: "Unknown" category shows significant presence across volume categories but excluded from main analysis for clarity.

4.3 Geographic Effect Analysis (Volume - Location Interactions)

Geographic analysis reveals significant state-level variations:


Top 3 States by Volume Category (Initial & Continuous)



Initial Applications (Mean Approval Rate):

- Volume Category 1 (1-10): IL leads (68.5%), followed by MD (60.2%) and NY (57.9%)
- Volume Category 2 (11-30): IL and MI dominate (58.9%), with MA close behind (58.6%)
- Volume Category 3 (31-100): OH shows highest (64.0%), followed by FL (62.3%) and GA (60.6%)
- Volume Category 4 (100+): MD leads (62.1%), with CA (58.8%) and OR (58.5%) following

Continuing Applications (Mean Approval Rate):

- Volume Category 1 (1-10): TX highest (53.7%), followed by MI (53.3%) and MO (53.1%)
- Volume Category 2 (11-30): MI leads (53.7%), with CT (52.1%) and CA (51.9%) following
- Volume Category 3 (31-100): IL shows highest (54.4%), followed by TX (54.1%) and MI (54.1%)
- Volume Category 4 (100+): CA leads (57.9%), with IL (54.5%) and TX (52.9%) following

Key Observations:

- Initial rates consistently higher than continuing rates
- Rate differences narrow with increasing volume
- MI and IL demonstrate consistent approval rates across categories

The interaction effects between volume, industry, and location suggest that volume is the dominant factor, with industry and geographic effects becoming less influential as volume increases.

5. Discussion and Implications

5.1 Key Findings

- Volume size significantly influences approval rates across all dimensions
- Geographic impact diminishes with increasing application volume
- Industry sector patterns show consistent trends across volume categories
- Initial applications generally show higher approval rates than continuing ones
- Tech hub states maintain stable performance regardless of volume

5.2 Practical Implications

5.2.1 For Employers

- Small Volume (1-10):

    - Focus on high-performing industries: Healthcare, Construction, Information
    - Consider locations like IL, MD, NY for higher success rates
    - Expect higher variability in outcomes
- Large Volume (100+):
    - Leverage stable approval rates in Finance and Information sectors
    - Tech hub locations (CA, MD, OR) offer consistent performance
    - Maintain standardized application processes for consistency

5.2.2 For Job Seekers

- Early Career:
    - Target large employers (100+ applications) for more predictable outcomes
    - Consider Information and Manufacturing sectors for consistent approval rates
    - Focus on tech hub states for higher success probability
- Experienced Professionals:
    - Finance sector shows strong continuing application performance
    - Consider established tech hubs for stable approval rates
    - Large volume employers offer more predictable outcomes

5.3 Model Limitations and Future Research

Our model, while providing valuable insights, has several limitations. These include use of aggregated data, state-level geographic restrictions, and broad industry categories. Future research should address these by investigating city-level patterns, sub-industry trends, and temporal policy impacts. Additionally, examining company size interactions and application quality factors would enhance understanding of H1B approval dynamics.

6. Conclusion

The analysis reveals complex interactions between volume categories, geographic locations, and industry sectors in H1B visa approvals. Large volume employers in established industries and tech hub states demonstrate the most consistent approval patterns. Initial applications show higher success rates across categories, while continuing applications show more stability in specific sectors like Finance and Information. The findings suggest that success in H1B applications depends on strategic alignment of volume, location, and industry factors, with larger organizations offering more predictable outcomes. These insights provide valuable guidance for both employers and job seekers in optimizing their H1B visa application strategies.

7. References

1. USCIS H-1B Employer Data Hub (2022-2024). Retrieved from [https://www.uscis.gov/data-reports/h-1b-employer-data-hub].
2. Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC Press.
3. Stan Development Team. (2023). *RStan: The R Interface to Stan*. Retrieved from [https://mc-stan.org/rstan/].
4. Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Academic Press