Analysis of H-1B Visa Workers Using H-1B LCA Disclosure Data (2020–2024)

Yufei Wu

The structure below is one possible setup for a manuscript, or a general data analysis project (including the course project). Adjust as needed. You don’t need to have exactly these sections, but the content covering those sections should be addressed.

This uses MS Word as output format. [See here](https://quarto.org/docs/output-formats/ms-word.html) for more information. You can switch to other formats, like html or pdf. See [the Quarto documentation](https://quarto.org/) for other formats.

Warning: package 'here' was built under R version 4.3.3

Warning: package 'knitr' was built under R version 4.3.3

**Authors**

* Yufei Wu (ORCID: 0009-0003-5051-0623)

**Author affiliations**

1. Department of Chemistry, University of Georgia, Athens, GA, USA.

Corresponding author: yufeiwu@uga.edu

Disclaimer: The opinions expressed in this article are the author’s own and don’t reflect their employer.

# 1. Summary/Abstract

*Write a summary of your project.*

# 2. Introduction

## 2.1 General Background Information

H-1B is a nonimmigrant visa category that allows employers to hire foreign workers in specialty occupations, like methematics, engineering, technology, and medical sciences (1). It requires the employee to have a bachelor’s degree or equivalent in the specific specialty. Typically, the maximun duration of an H-1B visa is six years.It is the most common work visa in the US. There are 65,000 available H-1B visas each year, with 20,000 additional visas for candidates with a master’s or doctorate degree from a U.S. institution. If there are more than 65,000 applications, USCIS will run a lottery to decide who can file an H-1B petition. As USCIS is receiving more and more H-1B registration these years, it’s harder for a foreign worker to get an H-1B visa. The Labor Condition Application (LCA) is an application filed by employers to apply for work authorization on behalf of employees as a prerequisite for H-1B (2). The LCA contains essential details such as job title, wage, and location about the proposed H-1B employment. Therefore, analyzing the LCA data can provide some ideas on the H-1B application status.

## 2.2 Description of data and data source

The data is H1B LCA Disclosure Data (2020-2024) from [Kaggle](https://www.kaggle.com/datasets/zongaobian/h1b-lca-disclosure-data-2020-2024/data). It contains LCA disclosure datasets from U.S. Department of Labor, covering the period from 2020 to 2024. It includes information such as case status, job title, Standard Occupational Classification (SOC) title, location, and wages.

## 2.3 Questions/Hypotheses to be addressed

This analysis will primarily focus on certified cases. The key research questions to be examined include: Distribution Analysis: What is the distribution of certified cases across various features such as SOC titles, locations, wages, and employer names? Which SOC titles, states, and employers have the highest and lowest number of certified cases? Variable Relationships: How do wages vary across different states, SOC titles, and employers? Employment Trends and Remote Work Patterns: How have employment patterns evolved over time? What trends can be observed in remote work prevalence before and after the pandemic? Predictive Modeling: Can a predictive model be developed to estimate wages based on job-related and employer-specific features? Which factors contribute most significantly to wage determination in certified cases?

To cite other work (important everywhere, but likely happens first in introduction), make sure your references are in the bibtex file specified in the YAML header above and have the right bibtex key. Then you can include like this:

Examples of reproducible research projects can for instance be found in (3,4).

# 3. Methods

The data will be reduced to include obeservations with certified case status and visa class of H-1B. Variable will only include RECEIVED\_DATE, SOC\_TITLE, EMPLOYER\_NAME, EMPLOYER\_STATE, WORKSITE\_STATE, WAGE\_RATE\_OF\_PAY\_FROM, PREVAILING\_WAGE.

## 3.1 Schematic of workflow

Sometimes you might want to show a schematic diagram/figure that was not created with code (if you can do it with code, do it). [Figure 1](#fig-schematic) is an example of some - completely random/unrelated - schematic that was generated with Biorender. We store those figures in the assets folder.

|  |
| --- |
| Figure 1: A figure that is manually generated and shows some overview/schematic. This has nothing to do with the data, it’s just a random one from one of our projects I found and placed here. |

## 3.2 Data aquisition

The dataset used in this analysis was obtained from Kaggle. The data was downloaded in CSV format.

## 3.3 Data import and cleaning

Since the dataset for this project is too large to upload to GitHub, I reduced its size and saved the smaller version as an RDS file, treating it as the raw data. Here’s the code I used to reduce size:

# load packages library(here) library(dplyr) # load raw data LCAraw1 <- read.csv(here(“data”, “raw-data”, “LCA2020to2024.csv”)) # Shrink the data colnames(LCAraw1) LCAraw <- LCAraw1 %>% select(CASE\_STATUS, VISA\_CLASS, RECEIVED\_DATE, SOC\_TITLE, EMPLOYER\_NAME, EMPLOYER\_STATE, WORKSITE\_STATE, WAGE\_RATE\_OF\_PAY\_FROM, PREVAILING\_WAGE) # Save as rds save\_data\_location <- here::here(“data”,“raw-data”,“LCAraw.rds”) saveRDS(LCAraw, file = save\_data\_location)

# load packages  
library(here)  
library(dplyr)

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':  
  
 filter, lag

The following objects are masked from 'package:base':  
  
 intersect, setdiff, setequal, union

# load raw data  
LCAraw <- readRDS(here("data", "raw-data", "LCAraw.rds"))  
  
# get the overview of the data  
dim(LCAraw)

[1] 3564698 9

head(LCAraw)

CASE\_STATUS VISA\_CLASS RECEIVED\_DATE SOC\_TITLE  
1 Certified H-1B 2019-09-25 COMPUTER OCCUPATIONS, ALL OTHER  
2 Certified H-1B 2019-09-25 SOFTWARE DEVELOPERS, APPLICATIONS  
3 Certified H-1B 2019-09-25 MECHANICAL ENGINEERS  
4 Certified H-1B 2019-09-25 SOFTWARE DEVELOPERS, APPLICATIONS  
5 Certified H-1B 2019-09-25 COMPUTER OCCUPATIONS, ALL OTHER  
6 Certified H-1B 2019-09-25 OPERATIONS RESEARCH ANALYSTS  
 EMPLOYER\_NAME EMPLOYER\_STATE WORKSITE\_STATE  
1 JO-ANN STORES, INC. OH OH  
2 DENKEN SOLUTIONS INC. CA TN  
3 EPITEC, INC. MI MI  
4 SYSTEMS TECHNOLOGY GROUP, INC. MI MI  
5 E-GIANTS TECHNOLOGIES LLC IA OH  
6 BIZINTEX, INC. GA CA  
 WAGE\_RATE\_OF\_PAY\_FROM PREVAILING\_WAGE  
1 100000.00 95118  
2 38.57 39  
3 43.50 39  
4 57.69 53  
5 75000.00 65333  
6 73000.00 72280

# Keep only the certified case and H-1B visa, and remove these two column  
LCAdata <- LCAraw %>%  
 filter(CASE\_STATUS == "Certified", VISA\_CLASS == "H-1B")%>%  
 select(RECEIVED\_DATE, SOC\_TITLE, EMPLOYER\_NAME, EMPLOYER\_STATE, WORKSITE\_STATE, WAGE\_RATE\_OF\_PAY\_FROM, PREVAILING\_WAGE)  
  
# take a look at the processed data  
dim(LCAdata)

[1] 3208872 7

head(LCAdata)

RECEIVED\_DATE SOC\_TITLE  
1 2019-09-25 COMPUTER OCCUPATIONS, ALL OTHER  
2 2019-09-25 SOFTWARE DEVELOPERS, APPLICATIONS  
3 2019-09-25 MECHANICAL ENGINEERS  
4 2019-09-25 SOFTWARE DEVELOPERS, APPLICATIONS  
5 2019-09-25 COMPUTER OCCUPATIONS, ALL OTHER  
6 2019-09-25 OPERATIONS RESEARCH ANALYSTS  
 EMPLOYER\_NAME EMPLOYER\_STATE WORKSITE\_STATE  
1 JO-ANN STORES, INC. OH OH  
2 DENKEN SOLUTIONS INC. CA TN  
3 EPITEC, INC. MI MI  
4 SYSTEMS TECHNOLOGY GROUP, INC. MI MI  
5 E-GIANTS TECHNOLOGIES LLC IA OH  
6 BIZINTEX, INC. GA CA  
 WAGE\_RATE\_OF\_PAY\_FROM PREVAILING\_WAGE  
1 100000.00 95118  
2 38.57 39  
3 43.50 39  
4 57.69 53  
5 75000.00 65333  
6 73000.00 72280

## 3.4 Statistical analysis

*Explain anything related to your statistical analyses.* Various plots will be made to visualize the data. Linear regression will be used to analyze the relationship between variables. Machine learning methods will be used as modeling technique.

# 4. Results

## 4.1 Exploratory/Descriptive analysis

*Use a combination of text/tables/figures to explore and describe your data. Show the most important descriptive results here. Additional ones should go in the supplement. Even more can be in the R and Quarto files that are part of your project.*

[Table 1](#tbl-summarytable) shows a summary of the data.

Note the loading of the data providing a **relative** path using the ../../ notation. (Two dots means a folder up). You never want to specify an **absolute** path like C:\ahandel\myproject\results\ because if you share this with someone, it won’t work for them since they don’t have that path. You can also use the here R package to create paths. See examples of that below. I generally recommend the here package.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Data summary table.   | skim\_type | skim\_variable | n\_missing | complete\_rate | factor.ordered | factor.n\_unique | factor.top\_counts | numeric.mean | numeric.sd | numeric.p0 | numeric.p25 | numeric.p50 | numeric.p75 | numeric.p100 | numeric.hist | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | factor | Gender | 0 | 1 | FALSE | 3 | M: 4, F: 3, O: 2 | NA | NA | NA | NA | NA | NA | NA | NA | | numeric | Height | 0 | 1 | NA | NA | NA | 165.66667 | 15.97655 | 133 | 156 | 166 | 178 | 183 | ▂▁▃▃▇ | | numeric | Weight | 0 | 1 | NA | NA | NA | 70.11111 | 21.24526 | 45 | 55 | 70 | 80 | 110 | ▇▂▃▂▂ | |

## 4.2 Basic statistical analysis

*To get some further insight into your data, if reasonable you could compute simple statistics (e.g. simple models with 1 predictor) to look for associations between your outcome(s) and each individual predictor variable. Though note that unless you pre-specified the outcome and main exposure, any “p<0.05 means statistical significance” interpretation is not valid.*

[Figure 2](#fig-result) shows a scatterplot figure produced by one of the R scripts.

|  |
| --- |
| Figure 2: Height and weight stratified by gender. |

## 4.3 Full analysis

*Use one or several suitable statistical/machine learning methods to analyze your data and to produce meaningful figures, tables, etc. This might again be code that is best placed in one or several separate R scripts that need to be well documented. You want the code to produce figures and data ready for display as tables, and save those. Then you load them here.*

Example [Table 2](#tbl-resulttable2) shows a summary of a linear model fit.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 2: Linear model fit table.   | term | estimate | std.error | statistic | p.value | | --- | --- | --- | --- | --- | | (Intercept) | 149.2726967 | 23.3823360 | 6.3839942 | 0.0013962 | | Weight | 0.2623972 | 0.3512436 | 0.7470519 | 0.4886517 | | GenderM | -2.1244913 | 15.5488953 | -0.1366329 | 0.8966520 | | GenderO | -4.7644739 | 19.0114155 | -0.2506112 | 0.8120871 | |

# 5. Discussion

## 5.1 Summary and Interpretation

*Summarize what you did, what you found and what it means.*

## 5.2 Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

## 5.3 Conclusions

*What are the main take-home messages?*

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

This paper (5) discusses types of analyses.

These papers (3,4) are good examples of papers published using a fully reproducible setup similar to the one shown in this template.

Note that this cited reference will show up at the end of the document, the reference formatting is determined by the CSL file specified in the YAML header. Many more style files for almost any journal [are available](https://www.zotero.org/styles). You also specify the location of your bibtex reference file in the YAML. You can call your reference file anything you like.

# 6. References

1. The H-1B Visa Program and Its Impact on the U.S. Economy. *American Immigration Council*. 2016;

2. Labor Condition Application. *Wikipedia*. 2025;

3. McKay B, Ebell M, Billings WZ, et al. [Associations Between Relative Viral Load at Diagnosis and Influenza A Symptoms and Recovery.](https://doi.org/10.1093/ofid/ofaa494) *Open forum infectious diseases*. 2020;7(11):ofaa494.

4. McKay B, Ebell M, Dale AP, et al. [Virulence-mediated infectiousness and activity trade-offs and their impact on transmission potential of influenza patients.](https://doi.org/10.1098/rspb.2020.0496) *Proceedings. Biological sciences*. 2020;287(1927):20200496.

5. Leek JT, Peng RD. [Statistics. What is the question?](https://doi.org/10.1126/science.aaa6146) *Science (New York, N.Y.)*. 2015;347(6228):1314–1315.