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

*Write code that reads in the file and cleans it so it’s ready for analysis. Since this will be fairly long code for most datasets, it might be a good idea to have it in one or several R scripts. If that is the case, explain here briefly what kind of cleaning/processing you do, and provide more details and well documented code somewhere (e.g. as supplement in a paper). All materials, including files that contain code, should be commented well so everyone can follow along.*

# 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 <- read.csv(here("data", "raw-data", "LCA2020to2024.csv"))  
  
# get the overview of the data  
dim(LCAraw)

[1] 3564698 96

head(LCAraw)

CASE\_NUMBER CASE\_STATUS RECEIVED\_DATE DECISION\_DATE ORIGINAL\_CERT\_DATE  
1 I-200-19268-393467 Certified 2019-09-25 2019-10-01   
2 I-200-19268-638983 Certified 2019-09-25 2019-10-01   
3 I-200-19268-177184 Certified 2019-09-25 2019-10-01   
4 I-200-19268-936403 Certified 2019-09-25 2019-10-01   
5 I-200-19268-394079 Certified 2019-09-25 2019-10-01   
6 I-200-19268-495825 Certified 2019-09-25 2019-10-01   
 VISA\_CLASS JOB\_TITLE SOC\_CODE  
1 H-1B APPLICATION ENGINEER, OMS [15-1199.02] 15-1199  
2 H-1B BI DEVELOPER II 15-1132  
3 H-1B QUALITY ENGINEER 17-2141  
4 H-1B SOFTWARE DEVELOPER, APPLICATIONS 15-1132  
5 H-1B QUALITY ENGINEER LEVEL II 15-1199  
6 H-1B OPERATION RESEARCH ANALYSTS 15-2031  
 SOC\_TITLE FULL\_TIME\_POSITION BEGIN\_DATE END\_DATE  
1 COMPUTER OCCUPATIONS, ALL OTHER Y 2019-10-07 2022-10-07  
2 SOFTWARE DEVELOPERS, APPLICATIONS Y 2020-01-08 2023-01-07  
3 MECHANICAL ENGINEERS Y 2019-10-03 2022-10-02  
4 SOFTWARE DEVELOPERS, APPLICATIONS Y 2019-10-07 2022-10-01  
5 COMPUTER OCCUPATIONS, ALL OTHER Y 2019-10-09 2022-10-08  
6 OPERATIONS RESEARCH ANALYSTS Y 2019-10-07 2022-10-06  
 TOTAL\_WORKER\_POSITIONS NEW\_EMPLOYMENT CONTINUED\_EMPLOYMENT  
1 1 1 0  
2 1 0 1  
3 1 0 0  
4 1 0 1  
5 1 0 0  
6 1 0 0  
 CHANGE\_PREVIOUS\_EMPLOYMENT NEW\_CONCURRENT\_EMPLOYMENT CHANGE\_EMPLOYER  
1 0 0 0  
2 0 0 0  
3 0 0 1  
4 0 0 0  
5 1 0 0  
6 0 0 1  
 AMENDED\_PETITION EMPLOYER\_NAME  
1 0 JO-ANN STORES, INC.  
2 0 DENKEN SOLUTIONS INC.  
3 0 EPITEC, INC.  
4 0 SYSTEMS TECHNOLOGY GROUP, INC.  
5 0 E-GIANTS TECHNOLOGIES LLC  
6 0 BIZINTEX, INC.  
 TRADE\_NAME\_DBA EMPLOYER\_ADDRESS1  
1 JO-ANN FABRIC AND CRAFT STORES 5555 DARROW ROAD  
2 9170 IRVINE CENTER DRIVE  
3 24800 DENSO DRIVE  
4 3001 W. BIG BEAVER ROAD, SUITE 500  
5 8033 UNIVERSITY BLVD  
6 225 CREEKSTONE RIDGE,  
 EMPLOYER\_ADDRESS2 EMPLOYER\_CITY EMPLOYER\_STATE EMPLOYER\_POSTAL\_CODE  
1 HUDSON OH 44236  
2 SUITE # 200 IRVINE CA 92618  
3 STE. 150 SOUTHFIELD MI 48033  
4 TROY MI 48084  
5 SUITE A CLIVE IA 50325  
6 SUITE 11, WOODSTOCK GA 30188  
 EMPLOYER\_COUNTRY EMPLOYER\_PROVINCE EMPLOYER\_PHONE EMPLOYER\_PHONE\_EXT  
1 UNITED STATES OF AMERICA 3304636787 NA  
2 UNITED STATES OF AMERICA 8017923416 NA  
3 UNITED STATES OF AMERICA 2483536800 NA  
4 UNITED STATES OF AMERICA 2486439010 NA  
5 UNITED STATES OF AMERICA 5153091451 NA  
6 UNITED STATES OF AMERICA 6787854949 NA  
 NAICS\_CODE EMPLOYER\_POC\_LAST\_NAME EMPLOYER\_POC\_FIRST\_NAME  
1 451120 KLUSKA KRYSTA  
2 541512 MADDULA RAJ  
3 541511 WILDMAN TAYLOR  
4 541511 ARVAL SREESHA  
5 541511 BODDULURI KRISHNA  
6 541511 YERNENI RAMESH  
 EMPLOYER\_POC\_MIDDLE\_NAME EMPLOYER\_POC\_JOB\_TITLE  
1 MANAGER, TALENT ACQUISITION  
2 SOLUTION ARCHITECT  
3 IMMIGRATION LEAD  
4 IMMIGRATION PARALEGAL  
5 CEO  
6 PRESIDENT  
 EMPLOYER\_POC\_ADDRESS1 EMPLOYER\_POC\_ADDRESS2 EMPLOYER\_POC\_CITY  
1 5555 DARROW ROAD HUDSON  
2 9170 IRVINE CENTER DRIVE SUITE # 200 IRVINE  
3 24800 DENSO DRIVE STE. 150 SOUTHFIELD  
4 3001 W. BIG BEAVER ROAD, SUITE 500 TROY  
5 8033 UNIVERSITY BLVD SUITE A CLIVE  
6 225 CREEKSTONE RIDGE, SUITE 11, WOODSTOCK  
 EMPLOYER\_POC\_STATE EMPLOYER\_POC\_POSTAL\_CODE EMPLOYER\_POC\_COUNTRY  
1 OH 44236 UNITED STATES OF AMERICA  
2 CA 92618 UNITED STATES OF AMERICA  
3 MI 48033 UNITED STATES OF AMERICA  
4 MI 48084 UNITED STATES OF AMERICA  
5 IA 50325 UNITED STATES OF AMERICA  
6 GA 30188 UNITED STATES OF AMERICA  
 EMPLOYER\_POC\_PROVINCE EMPLOYER\_POC\_PHONE EMPLOYER\_POC\_PHONE\_EXT  
1 3304636787 NA  
2 8017923416 NA  
3 2483536800 NA  
4 2486439010 NA  
5 5153091451 NA  
6 6787854949 NA  
 EMPLOYER\_POC\_EMAIL AGENT\_REPRESENTING\_EMPLOYER  
1 KRYSTA.KLUSKA@JOANN.COM Y  
2 IMMIGRATION@DENKENSOLUTIONS.COM Y  
3 IMMIGRATION@EPITEC.COM Y  
4 SREESHA.ARVAL@STGIT.COM N  
5 LEGAL@THEEGIANTS.COM N  
6 RAMESH.YERNENI@BIZINTEX.COM N  
 AGENT\_ATTORNEY\_LAST\_NAME AGENT\_ATTORNEY\_FIRST\_NAME AGENT\_ATTORNEY\_MIDDLE\_NAME  
1 TZU-KAI TZU-KAI   
2 YELENA YELENA   
3 NATALIA NATALIA MADLEN  
4   
5   
6   
 AGENT\_ATTORNEY\_ADDRESS1 AGENT\_ATTORNEY\_ADDRESS2 AGENT\_ATTORNEY\_CITY  
1 333 WEST WACKER DRIVE 15TH FLOOR CHICAGO  
2 6622 SOUTHPOINT DR. S STE 330 JACKSONVILLE  
3 755 W. BIG BEAVER ROAD SUITE 1100 TROY  
4   
5   
6   
 AGENT\_ATTORNEY\_STATE AGENT\_ATTORNEY\_POSTAL\_CODE AGENT\_ATTORNEY\_COUNTRY  
1 IL 60606 UNITED STATES OF AMERICA  
2 FL 32216 UNITED STATES OF AMERICA  
3 MI 48084 UNITED STATES OF AMERICA  
4   
5   
6   
 AGENT\_ATTORNEY\_PROVINCE AGENT\_ATTORNEY\_PHONE AGENT\_ATTORNEY\_PHONE\_EXT  
1 3122636101 NA  
2 9043890055 NA  
3 2485199900 NA  
4 NA NA  
5 NA NA  
6 NA NA  
 AGENT\_ATTORNEY\_EMAIL\_ADDRESS LAWFIRM\_NAME\_BUSINESS\_NAME  
1 KLO@FRAGOMEN.COM FRAGOMEN, DEL REY, BERNSEN & LOEWY, LLP  
2 USCISNOTICES@SCARBOROUGHLAW.COM SCARBOROUGH LAW, L.L.C.  
3 NATALIA.SARRAF@ELLISPORTER.COM ELLIS PORTER, PLC  
4   
5   
6   
 STATE\_OF\_HIGHEST\_COURT NAME\_OF\_HIGHEST\_STATE\_COURT WORKSITE\_WORKERS  
1 IL SUPREME COURT NA  
2 FL FLORIDA SUPREME COURT NA  
3 MI SUPREME COURT NA  
4 NA  
5 NA  
6 NA  
 SECONDARY\_ENTITY SECONDARY\_ENTITY\_BUSINESS\_NAME WORKSITE\_ADDRESS1  
1 N 5555 Darrow Road  
2 Y NAVIHEALTH 210 WESTWOOD PL #400  
3 Y Ford Motor Company 21 Town Center Office (TCO)  
4 Y Ford Motor Company 21001 Van Born Road  
5 Y KROGER INC 11450, GROOMS ROAD  
6 Y Cisco Systems, Inc 170 West Tasman Dr  
 WORKSITE\_ADDRESS2 WORKSITE\_CITY WORKSITE\_COUNTY WORKSITE\_STATE  
1 Hudson Summit OH  
2 BRENTWOOD WILLIAMSON TN  
3 18900 Michigan Avenue Dearborn Wayne MI  
4 Taylor Wayne MI  
5 BLUE ASH HAMILTON OH  
6 San Jose SANTA CLARA CA  
 WORKSITE\_POSTAL\_CODE WAGE\_RATE\_OF\_PAY\_FROM WAGE\_RATE\_OF\_PAY\_TO  
1 44224 100000.00 NA  
2 37027 38.57 38.57  
3 48126 43.50 NA  
4 48180 57.69 57.69  
5 45242 75000.00 NA  
6 95134 73000.00 73000.00  
 WAGE\_UNIT\_OF\_PAY PREVAILING\_WAGE PW\_UNIT\_OF\_PAY PW\_TRACKING\_NUMBER  
1 Year 95118 Year   
2 Hour 39 Hour   
3 Hour 39 Hour   
4 Hour 53 Hour   
5 Year 65333 Year   
6 Year 72280 Year   
 PW\_WAGE\_LEVEL PW\_OES\_YEAR PW\_OTHER\_SOURCE PW\_OTHER\_YEAR PW\_SURVEY\_PUBLISHER  
1 IV 2018.0 OES NA   
2 II 2019.0 OES NA   
3 II 2019.0 OES NA   
4 IV 2019.0 OES NA   
5 II 2019.0 OES NA   
6 II 2019.0 OES NA   
 PW\_SURVEY\_NAME TOTAL\_WORKSITE\_LOCATIONS AGREE\_TO\_LC\_STATEMENT H\_1B\_DEPENDENT  
1 NA Y N  
2 NA Y Y  
3 NA Y Y  
4 NA Y Y  
5 NA Y Y  
6 NA Y Y  
 WILLFUL\_VIOLATOR SUPPORT\_H1B STATUTORY\_BASIS APPENDIX\_A\_ATTACHED  
1 N   
2 N Y BOTH   
3 N Y BOTH   
4 N Y BOTH   
5 N Y BOTH   
6 N Y BOTH   
 PUBLIC\_DISCLOSURE PREPARER\_LAST\_NAME PREPARER\_FIRST\_NAME  
1 Disclose Business   
2 Disclose Business   
3 Disclose Business   
4 Disclose Business   
5 Disclose Business   
6 Disclose Business   
 PREPARER\_MIDDLE\_INITIAL PREPARER\_BUSINESS\_NAME PREPARER\_EMAIL  
1   
2   
3   
4   
5 LEGAL@THEEGIANTS.COM  
6

# clean the data  
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  
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

# save as rds  
save\_data\_location <- here::here("data","processed-data","LCAdata.rds")  
saveRDS(LCAdata, file = save\_data\_location)

## 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.