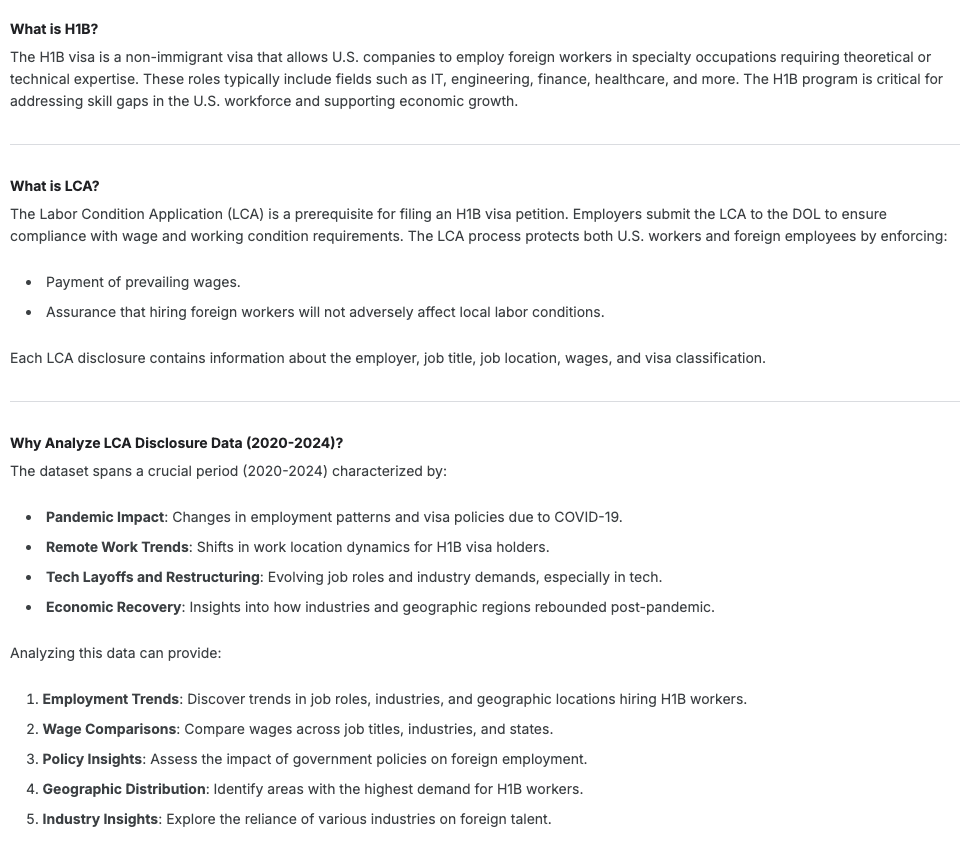
**Data exploration log for H1B LCA Data Disclosure**

* Source: <https://www.kaggle.com/datasets/zongaobian/h1b-lca-disclosure-data-2020-2024/data>
* Overview



1. Cleanness/structure check
   1. We set the data types as below to import 96 columns/5 files of each year into Azure data studio (see sql create table clause)

CREATE TABLE [dbo].[LCA\_20] (

[CASE\_NUMBER] NVARCHAR (50) NOT NULL,

[CASE\_STATUS] NVARCHAR (50) NULL,

[RECEIVED\_DATE] DATE NULL,

[DECISION\_DATE] DATE NULL,

[ORIGINAL\_CERT\_DATE] DATE NULL,

[VISA\_CLASS] NVARCHAR (50) NULL,

[JOB\_TITLE] NVARCHAR (150) NULL,

[SOC\_CODE] NVARCHAR (50) NULL,

[SOC\_TITLE] NVARCHAR (150) NULL,

[FULL\_TIME\_POSITION] NVARCHAR (50) NULL,

[BEGIN\_DATE] DATE NULL,

[END\_DATE] DATE NULL,

[TOTAL\_WORKER\_POSITIONS] INT NULL,

[NEW\_EMPLOYMENT] INT NULL,

[CONTINUED\_EMPLOYMENT] INT NULL,

[CHANGE\_PREVIOUS\_EMPLOYMENT] INT NULL,

[NEW\_CONCURRENT\_EMPLOYMENT] INT NULL,

[CHANGE\_EMPLOYER] INT NULL,

[AMENDED\_PETITION] INT NULL,

[EMPLOYER\_NAME] NVARCHAR (150) NULL,

[TRADE\_NAME\_DBA] NVARCHAR (150) NULL,

[EMPLOYER\_ADDRESS1] NVARCHAR (150) NULL,

[EMPLOYER\_ADDRESS2] NVARCHAR (150) NULL,

[EMPLOYER\_CITY] NVARCHAR (100) NULL,

[EMPLOYER\_STATE] NVARCHAR (50) NULL,

[EMPLOYER\_POSTAL\_CODE] NVARCHAR (50) NULL,

[EMPLOYER\_COUNTRY] NVARCHAR (100) NULL,

[EMPLOYER\_PROVINCE] NVARCHAR (50) NULL,

[EMPLOYER\_PHONE] NVARCHAR (50) NULL,

[EMPLOYER\_PHONE\_EXT] NVARCHAR (50) NULL,

[NAICS\_CODE] NVARCHAR (50) NULL,

[EMPLOYER\_POC\_LAST\_NAME] NVARCHAR (50) NULL,

[EMPLOYER\_POC\_FIRST\_NAME] NVARCHAR (50) NULL,

[EMPLOYER\_POC\_MIDDLE\_NAME] NVARCHAR (50) NULL,

[EMPLOYER\_POC\_JOB\_TITLE] NVARCHAR (100) NULL,

[EMPLOYER\_POC\_ADDRESS1] NVARCHAR (150) NULL,

[EMPLOYER\_POC\_ADDRESS2] NVARCHAR (150) NULL,

[EMPLOYER\_POC\_CITY] NVARCHAR (100) NULL,

[EMPLOYER\_POC\_STATE] NVARCHAR (50) NULL,

[EMPLOYER\_POC\_POSTAL\_CODE] NVARCHAR (50) NULL,

[EMPLOYER\_POC\_COUNTRY] NVARCHAR (100) NULL,

[EMPLOYER\_POC\_PROVINCE] NVARCHAR (50) NULL,

[EMPLOYER\_POC\_PHONE] NVARCHAR (50) NULL,

[EMPLOYER\_POC\_PHONE\_EXT] NVARCHAR (50) NULL,

[EMPLOYER\_POC\_EMAIL] NVARCHAR (100) NULL,

[AGENT\_REPRESENTING\_EMPLOYER] NVARCHAR (50) NULL,

[AGENT\_ATTORNEY\_LAST\_NAME] NVARCHAR (50) NULL,

[AGENT\_ATTORNEY\_FIRST\_NAME] NVARCHAR (50) NULL,

[AGENT\_ATTORNEY\_MIDDLE\_NAME] NVARCHAR (50) NULL,

[AGENT\_ATTORNEY\_ADDRESS1] NVARCHAR (150) NULL,

[AGENT\_ATTORNEY\_ADDRESS2] NVARCHAR (150) NULL,

[AGENT\_ATTORNEY\_CITY] NVARCHAR (50) NULL,

[AGENT\_ATTORNEY\_STATE] NVARCHAR (50) NULL,

[AGENT\_ATTORNEY\_POSTAL\_CODE] NVARCHAR (50) NULL,

[AGENT\_ATTORNEY\_COUNTRY] NVARCHAR (50) NULL,

[AGENT\_ATTORNEY\_PROVINCE] NVARCHAR (50) NULL,

[AGENT\_ATTORNEY\_PHONE] NVARCHAR (50) NULL,

[AGENT\_ATTORNEY\_PHONE\_EXT] NVARCHAR (50) NULL,

[AGENT\_ATTORNEY\_EMAIL\_ADDRESS] NVARCHAR (100) NULL,

[LAWFIRM\_NAME\_BUSINESS\_NAME] NVARCHAR (100) NULL,

[STATE\_OF\_HIGHEST\_COURT] NVARCHAR (100) NULL,

[NAME\_OF\_HIGHEST\_STATE\_COURT] NVARCHAR (100) NULL,

[WORKSITE\_WORKERS] NVARCHAR (50) NULL,

[SECONDARY\_ENTITY] NVARCHAR (50) NULL,

[SECONDARY\_ENTITY\_BUSINESS\_NAME] NVARCHAR (150) NULL,

[WORKSITE\_ADDRESS1] NVARCHAR (100) NULL,

[WORKSITE\_ADDRESS2] NVARCHAR (100) NULL,

[WORKSITE\_CITY] NVARCHAR (50) NULL,

[WORKSITE\_COUNTY] NVARCHAR (50) NULL,

[WORKSITE\_STATE] NVARCHAR (50) NULL,

[WORKSITE\_POSTAL\_CODE] NVARCHAR (50) NULL,

[WAGE\_RATE\_OF\_PAY\_FROM] FLOAT (53) NULL,

[WAGE\_RATE\_OF\_PAY\_TO] FLOAT (53) NULL,

[WAGE\_UNIT\_OF\_PAY] NVARCHAR (50) NULL,

[PREVAILING\_WAGE] FLOAT (53) NULL,

[PW\_UNIT\_OF\_PAY] NVARCHAR (50) NULL,

[PW\_TRACKING\_NUMBER] NVARCHAR (50) NULL,

[PW\_WAGE\_LEVEL] NVARCHAR (50) NULL,

[PW\_OES\_YEAR] NVARCHAR (50) NULL,

[PW\_OTHER\_SOURCE] NVARCHAR (50) NULL,

[PW\_OTHER\_YEAR] NVARCHAR (50) NULL,

[PW\_SURVEY\_PUBLISHER] NVARCHAR (150) NULL,

[PW\_SURVEY\_NAME] NVARCHAR (150) NULL,

[TOTAL\_WORKSITE\_LOCATIONS] NVARCHAR (50) NULL,

[AGREE\_TO\_LC\_STATEMENT] NVARCHAR (50) NULL,

[H\_1B\_DEPENDENT] NVARCHAR (50) NULL,

[WILLFUL\_VIOLATOR] NVARCHAR (50) NULL,

[SUPPORT\_H1B] NVARCHAR (50) NULL,

[STATUTORY\_BASIS] NVARCHAR (200) NULL,

[APPENDIX\_A\_ATTACHED] NVARCHAR (50) NULL,

[PUBLIC\_DISCLOSURE] NVARCHAR (150) NULL,

[PREPARER\_LAST\_NAME] NVARCHAR (100) NULL,

[PREPARER\_FIRST\_NAME] NVARCHAR (150) NULL,

[PREPARER\_MIDDLE\_INITIAL] NVARCHAR (150) NULL,

[PREPARER\_BUSINESS\_NAME] NVARCHAR (150) NULL,

[PREPARER\_EMAIL] NVARCHAR (150) NULL,

CONSTRAINT [PK\_LCA\_20] PRIMARY KEY CLUSTERED ([CASE\_NUMBER] ASC)

);

* 1. **The issue of duplicated case numbers**—primary keys couldn’t be set for the year of 2022 and 2024. So, ideally, the case number for each H1B petition should be treated as the unique key for data analysis purposes; however, we noticed that there are duplicated case numbers when importing the data. (see screen shot below)

A screenshot of a computer error

AI-generated content may be incorrect.

* 1. We want to check if the duplicated data is of rare case or prevalent—looks like there are a lot.

A screenshot of a computer

AI-generated content may be incorrect.

* 1. Let’s further dive into the patterns of duplicates (we tested and learned that the max of duplicated case number is 3)

A white screen with black text

AI-generated content may be incorrect.

A white background with red text

AI-generated content may be incorrect.

* 1. We then discovered that there is total 3.5M rows, with only around 2.2M rows with unique case numbers. We will have to address this table structure when selecting the data we want to analyze with.

A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

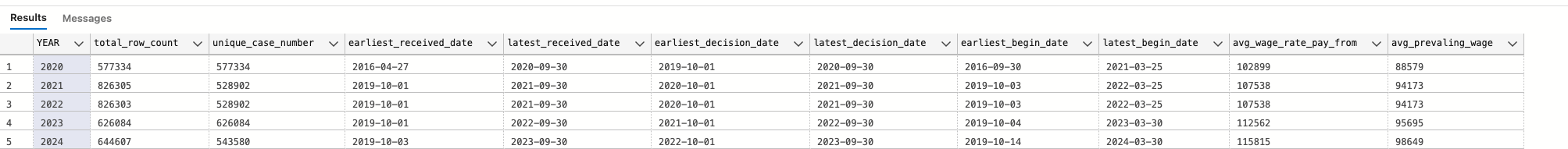
* 1. We also learned that case numbers are generated by job position+company. For example, Samsung filed a case for 25 SDEs in on case, which consisted of 5 new employment, 5 continued employment, 5 change of previously employment, 5 change of employer, and 5 amended petitions.

A screenshot of a computer

AI-generated content may be incorrect.

Now we are clear on the whole picture of how this data was modeled, let’s dive into some high level stats for us to understand the data before we decide our goals of analysis and the approaches to sift/clean the data.

1. Numerical stats exploration (2020~2024)

* SQL snippet
* SELECT
* COUNT(CASE\_NUMBER) as total\_row\_count
* ,COUNT(DISTINCT CASE\_NUMBER) as unique\_case\_number
* ,MIN(RECEIVED\_DATE) as earliest\_received\_date
* ,MAX(RECEIVED\_DATE) as latest\_received\_date
* ,MIN(DECISION\_DATE) as earliest\_decision\_date
* ,MAX(DECISION\_DATE) as latest\_decision\_date
* ,MIN(BEGIN\_DATE) as earliest\_begin\_date
* ,MAX(BEGIN\_DATE) as latest\_begin\_date
* ,ROUND(AVG(WAGE\_RATE\_OF\_PAY\_FROM),0) as avg\_wage\_rate\_pay\_from
* ,ROUND(AVG(PREVAILING\_WAGE),0) as avg\_prevaling\_wage
* FROM LCA\_20 # change for each year
* Result for each year

1. Categorical stats exploration

* SQL snippet (change tables for each year)
* select case\_status, COUNT(case\_status) as case\_count, SUM(TOTAL\_WORKER\_POSITIONS) as head\_count, ROUND(cast(SUM(TOTAL\_WORKER\_POSITIONS) as float) / COUNT(case\_status),2) AS avg\_hc\_per\_case from LCA\_20 group by case\_status order by case\_count desc;
* select VISA\_CLASS, count(VISA\_CLASS) as case\_count, sum(TOTAL\_WORKER\_POSITIONS) as head\_count, ROUND(cast(SUM(TOTAL\_WORKER\_POSITIONS) as float) / COUNT(case\_status),2) AS avg\_hc\_per\_case from LCA\_20 group by VISA\_CLASS order by case\_count desc;
* select TOP 10 SOC\_TITLE, count(SOC\_TITLE) as case\_count, sum(TOTAL\_WORKER\_POSITIONS) as head\_count, ROUND(cast(SUM(TOTAL\_WORKER\_POSITIONS) as float) / COUNT(case\_status),2) AS avg\_hc\_per\_case from LCA\_20 group by SOC\_TITLE order by case\_count desc ;
* select TOP 10 EMPLOYER\_NAME, count(EMPLOYER\_NAME) as case\_count, sum(TOTAL\_WORKER\_POSITIONS) as head\_count, ROUND(cast(SUM(TOTAL\_WORKER\_POSITIONS) as float) / COUNT(case\_status),2) AS avg\_hc\_per\_case from LCA\_20 group by EMPLOYER\_NAME order by case\_count desc ;
* select TOP 10 EMPLOYER\_STATE, count(EMPLOYER\_STATE) as case\_count, sum(TOTAL\_WORKER\_POSITIONS) as head\_count, ROUND(cast(SUM(TOTAL\_WORKER\_POSITIONS) as float) / COUNT(case\_status),2) AS avg\_hc\_per\_case from LCA\_20 group by EMPLOYER\_STATE order by case\_count desc ;
* Result of each year
  + 2020

A screenshot of a computer

AI-generated content may be incorrect.

* + 2021

A screenshot of a computer

AI-generated content may be incorrect.

* + 2022

A screenshot of a computer

AI-generated content may be incorrect.

* + 2023

A screenshot of a computer

AI-generated content may be incorrect.

* + 2024A screenshot of a computer

    AI-generated content may be incorrect.

2020~2024

