# **Project Title: Non-immigrant Data Science workforce in the US 2011-2016**

# Author: Julia Nguyen\_Individual project

# Project Description

**Objectives:** My application will use H1-B visa data from Department of Labor (DOI) to draw an overall picture of non-immigrant workforce in data science area from 2011 - 2016. H1-B visa is a temporary employment-based visa granted to non-US citizen individuals who are hired for specialty occupations that require special skillset and ability. Only 65,000 visas are granted each year to qualified professionals and this analysis will explore the relationship among different features in the dataset leading to case final decision.

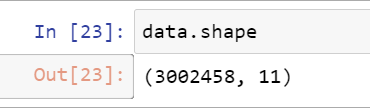
**What is this analysis used for?** Raw data is sourced from DOI and can be analysed in multiple ways to derive insights. This application is unique as my project will focus on the data science area only. This application can be used by:

* Individuals who are interested in applying for H1-B visa in data science can refer to this analysis as a frame of reference.
* Employers who want to determine case’s approval likelihood based on history records.
* Anyone else who is interested to see the trend in non-immigrant employment in data science field.

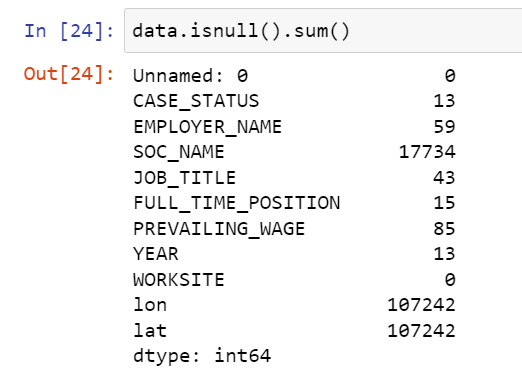
**About the dataset**

Raw data about H1-B visa application in this dataset was collected for the period of 2011 to 2016 by DOI. The dataset was created and uploaded by Sharan Naribole in 2016 as a csv file.

Link to dataset: <https://www.kaggle.com/datasets/nsharan/h-1b-visa>

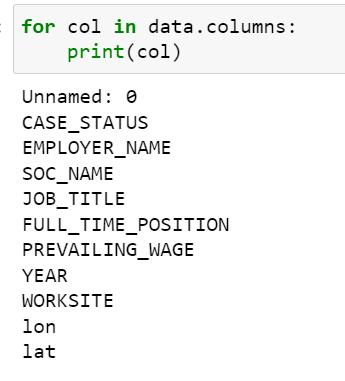


The dataset has 3,002,458 rows x 11 columns representing more than 3 million application records

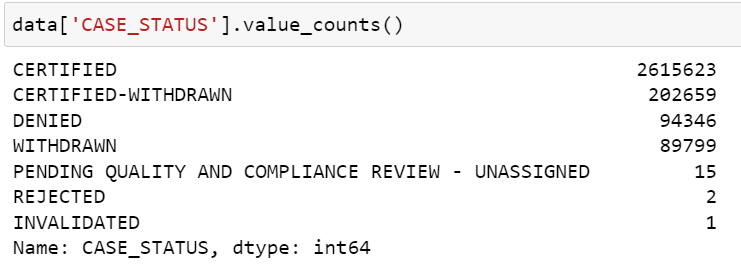


The dataset contains null values and will require cleaning

Independent variables or 11 columns in the dataset are labelled as follows:



* EMPLOYER\_NAME: Name of the company that sponsored the applicant
* SOC\_NAME: Occupational name associated with the SOC CODE, as classified by the Standard Occupational Classification (SOC) system
* JOB\_TITLE: the job that the applicant was hired for
* FULL\_TIME\_POSITION: Y = Full Time Position; N = Part Time Position.
* PREVAILING\_WAGE: annual compensation
* YEAR: the year of the application
* WORKSITE: location where the job was offered
* lon: we are not interested in this field for the purpose of this project
* lat: we are not interested in this field for the purpose of this project

‘CASE STATUS’ will be the target variable for purpose of building prediction model. This field has the following unique values indicating each application’s status or decision: 

**What is included in this project?**

Multiple visualization methods will be used to answer the following questions:

* Number of applications each year (bar chart)
* Top 10 average salary offered for Data Science (line chart)
* Is the number of petitions with Data Engineer job title increasing over time?
* Which part of the US has the most Data Science jobs?
* Which industry has the most number of Data Scientist roles? (bar chart)
* Which employers file the most petitions for Data Scientist each year?
* What are the most meaningful features (by ingesting data into Independent vs Dependent variables using heatmap)

I will use Logistic Regression to build prediction model on this dataset. An H-1B visa can be denied when the applicant does not meet the required knowledge or skills for the offered job. Another interesting angle to look from is the relationship between wage and case decision - Jobs with higher wage are more likely to require specialized skills which could lead to lower rejection rates.

[Placeholder]: Ideally, I’d like to be able to design my application so that one can enter relevant information about their potential case and the app will predict the likelihood of approval or denial. However, that will require some deep knowledge and skills which I’m not sure I can learn in the next few weeks.

This is my first experience using Github, see link below. A readme file is included.

Github repository: <https://github.com/jhnguyen2022/FA22-BL-DSCI-D590-11690>