ETL Project

**H1B VISA Prediction**

Han Yang and Monica Ramos

**Summary**

Datasets are available for users to predict the approval rate of obtaining an H1B VISA in United States of America based on job industry sectors.

**Data sources**

Data were extracted from two sources: Kaggle and NAICS Association websites.

Kaggle.com

<https://www.kaggle.com/abishekanbarasan1995/h1b-case-status-prediction>

H1B VISA information is presented in 4 separate datasets of fiscal years from 2015 to 2018 respectively. Each dataset contains information including employee application case number, case status, decision date, employer information and NAICS codes.

• H-1B\_Disclosure\_RAW\_Data\_FY15.csv (224.43 MB)

• H-1B\_Disclosure\_RAW\_Data\_FY16.csv (233.67 MB)

• H-1B\_Disclosure\_RAW\_Data\_FY17.csv (239.54 MB)

• H-1B\_Disclosure\_RAW\_Data\_FY18.csv (258.86 MB)

NAICS Association

<https://www.naics.com/business-lists/counts-by-naics-code/#countsByNAICS>

NAICS Association, LLC (North American Industry Classification System) provides valuable resources to assist companies in determining all things pertaining to NAICS and SIC Codes by industry. NAICS codes are presented in tables listed on its website.

**Cleaning and Preparation**

NAICS codes (NAICS\_CODE.csv) was obtained by web scrapping through Python libraries including Pandas, Splinter and Beautiful Soup. H1B VISA datasets, through Pandas, were processed individually to remove unnecessary columns and combined into a single file, which was then split into 2 separated files (EMPLOYER.csv and EMPLOYEE.csv) for the purpose of data normalization.

**Final Database Set and Relationship**

All datasets were stored into PostgresSQL database for further analysis.

[https://app.quickdatabasediagrams.com](https://app.quickdatabasediagrams.com/)

