# **H-1B Visa petitions 2011-16**

|  |  |
| --- | --- |
| **Name** | **CWID** |
| Abhishek Dutta | A20379160 |

**Type of Your Projects:** Multiple linear regression.

1. **Introduction**

H-1B visa is considered to be the most common visa status and is held by international students once they complete their higher studies (graduation, post-graduation etc.) and work in a full-time position.

It is an employment-based visa for temporary international workers in USA. For in international / foreign worker to get an H-1B visa their employer must sponsor them by giving a job and filing a petition for them in the US immigration department.

H-1B visa is on basis of specialty occupations or as fashion models of distinguished merit and ability. A specialty occupation requires the theoretical and practical application of specialized knowledge and a bachelor's degree or the equivalent in the specific specialty like - science, medicine, health care, education and business specialties, etc).

Our motivation is to come out with answers that would really help us to know about employment areas that are doing well, Which part of USA has more opportunities for specific Job titles for international professionals/students, Which industry has more requirement for specific profiles and which employer files most number of petitions in a year.

1. **Data Sets**

 The Office of Foreign Labor Certification (OFLC), generates the data for different visa types this data is program data and we are planning to take this disclosure data and would analyze about the H-1B visa type from 2011-2016.

In our project, there will be around 3 million data records for H-1B petitions doing the above-mentioned period. However, the raw data available here is little messed up and therefore we would require some transformation to make it more accessible and understandable for better interpretation.

<https://www.foreignlaborcert.doleta.gov/performancedata.cfm#sthash.9k5WFDT3.dpuf>

1. **Research Problems**

* Since, the data that we have has been taken from 2011 till 2016 therefore there are few changes in the column name after 2015 that we need to consider. Therefore, we would have to bring the column name after 2015 to match the earlier year’s column name.
* Also, we need to perform few data transformations on various columns that may have missing data, data on different scales and we need to normalize it to bring it to a similar scale for our easy interpretations and understanding and accurate data outputs. Few examples are given below as per our observations: