**Title: H1B-VISA Petitions Analysis for years 2011-2016**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **First Name** | **Last Name** | **Online Students? (Y or N)** | **Monday or Tuesday** | **Shared with ITMD 525? (Y or N)** |
| Abhishek | Dutta | N | Monday | N |

Contents

[**1.** **Introduction and Motivations** 3](#_Toc481534540)

[**2.** **Data Description** 4](#_Toc481534541)

[**3.** **Research Problems and Solutions:** 5](#_Toc481534542)

[**4.** **Model Learning** 6](#_Toc481534543)

[4.1. Data Processing 6](#_Toc481534544)

[4.2. Data Analytics Tasks and Processes 6](#_Toc481534545)

[**5.** **Evaluations and Results** 20](#_Toc481534546)

[5.1. Evaluation Methods 20](#_Toc481534547)

[5.2. Results and Findings 20](#_Toc481534548)

[**6.** **Conclusions and Future Work** 21](#_Toc481534549)

[6.1. Conclusions 21](#_Toc481534550)

[6.2. Limitations 21](#_Toc481534551)

[6.3. Potential Improvements or Future Work 21](#_Toc481534552)

# **Introduction and Motivations**

H1-B Visa is an employment based visa for temporary international workers in USA.

For an 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.

It 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, healthcare, education and business specialties etc.

Our motivations comes from analyzing about different factors such as which location files more number of H1-B petitions, get the list of top profiles that has maximum number of certified petition status and which profiles have maximum number of denied status. This shows us the demand for profiles/Job Titles based on location and other such analysis related to H1-B Visa petitions.

# **Data Description**

|  |  |
| --- | --- |
| Dataset: H1b visa petitions for 2011-2016 | |
| Data source link: | https://www.kaggle.com/nsharan/h-1b-visa |
| Number of files: | 1 |
| Total number of  columns including all  files: | 35 |
| Total number of rows  including all files: | 31000 |
| Potential columns for  consideration: | CASE\_STATUS, EMPLOYER\_NAME, SOC\_NAME, JOB\_TITLE, PREVAILING\_WAGES, WORKSITE, YEAR, FULL\_TIME\_POSITIONS, REGION, LON, LAT etc. |

In our dataset, there is one file which includes the h1b petitions data for the fiscal years starting from 2011 to 2016. The data contains information about the employer, the job titles for which it has applied h1b applications, the job title of the applicant and the region in which the employer has its origin. This will allow us to apply the logistic regression analysis on the data to know the status of the H1B visa application whether certified or denied.

The detailed column names and descriptions are as:

* **CASE STATUS**: Status associated with the last significant event or decision. Valid values include Certified, Certified-Withdrawn, Denied, and Withdrawn.
* **EMPLOYER NAME**: Name of employer submitting labor condition application.
* **SOC NAME**: Occupational name associated with the SOC CODE.
* **JOB TITLE**: Title of the job.

* **FULL TIME POSITION**: Y = Full Time Position; N = Part Time Position.
* **PREVAILING WAGE**: Prevailing Wage for the job being requested for temporary labor condition. The wage is listed at annual scale in USD. The prevailing wage for a job position is defined as the average wage paid to similarly employed workers in the requested occupation in the area of intended employment. The prevailing wage is based on the employer’s minimum requirements for the position.
* **YEAR**: Year in which the H-1B visa petition was filed.
* **WORKSITE**: City and State information of the foreign worker’s intended area of employment.
* **LON & LAT**: Longitude and latitude of the worksite

# **Research Problems and Solutions:**

From the data that we have extracted, we are targeting to perform multiple kinds of analysis and visually display their results in graphical formats.

**The research problems that we are targeting in our project are:**

1. To analyze the employers who file most H1B visa applications
2. To analyze percentage share out of 85000 applications for employers with most of applications
3. To analyze the maximum number of applications for the employers location wise
4. To analyze the trending jobs and related salaries for which H1B visas are applied
5. Comparison of wages offered by high-applicant employers to rest of the industry
6. To analyze the percentages of visa applications certified, withdrawn or denied.

**Solutions to the research problems:**

1. As the first step towards the solution, we have done pre processing on the data to filter out the noisy or missing values.
2. We have built the base model by taking status as the dependent variable and position type, wages, year, region and employer name as the independent variables.
3. We have used base model to use it in Forward, Backward and best subset model.
4. We have used AIC, Adjusted R2 and accuracy as the measures to compare these models and select the best amongst them to predict the visa application status.
5. We are using logistic regression analysis for which 0 represents application certified and 1 represents application denied state. To get the most of the output with equivalent probabilities, we are calculating accuracy of the above models with 60, 70, 80 & 90 percent of the train data.

# **Model Learning**

## Data Processing

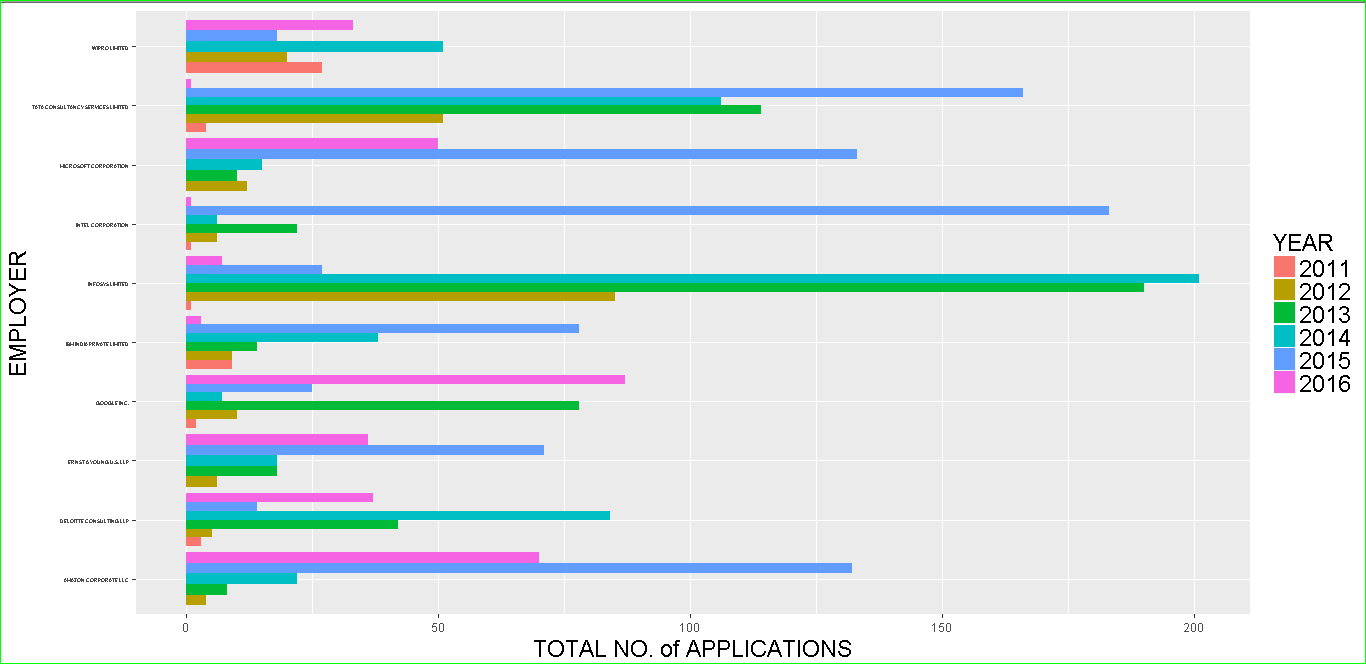
There are many entries containing N/A values, therefore we have removed those rows from our dataset in order to get a data helpful for our analysis by using **na.omit** in R.

We have categorized our data –Worksite containing city and state into Region South, East, West, North West etc. in order to simplify our analysis while creating a model.

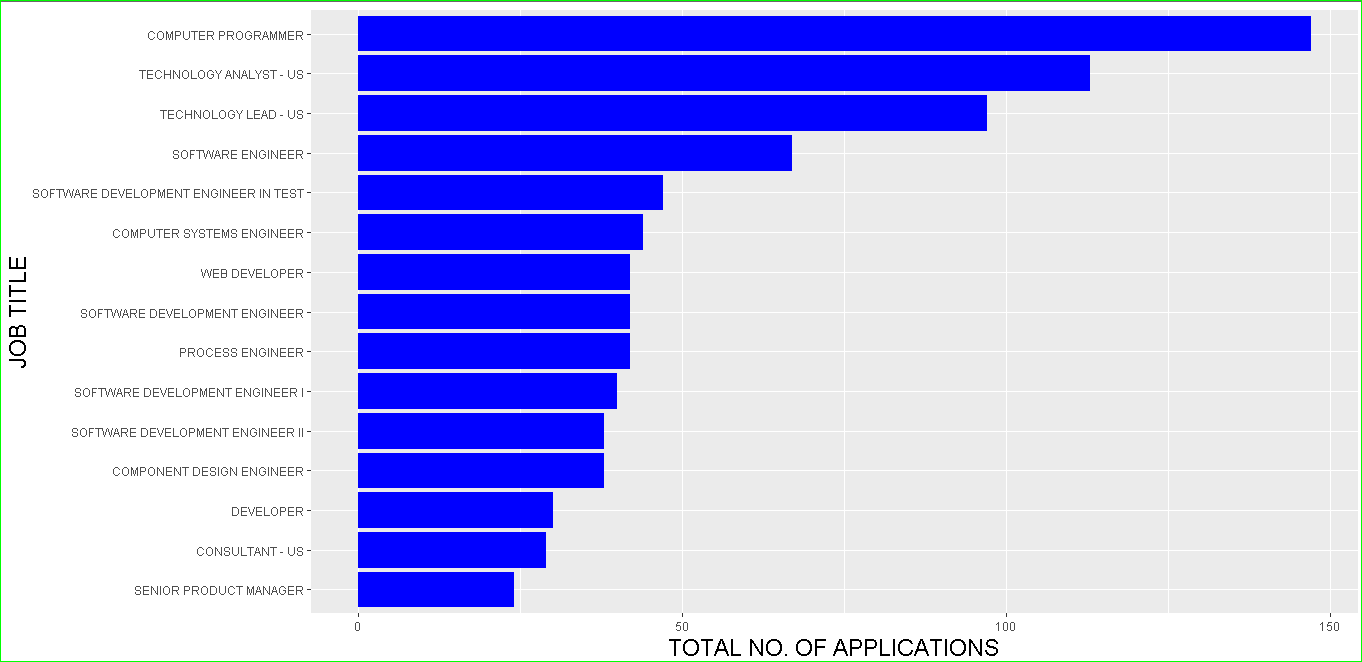
SOC\_NAME give us the department under which the Job\_Title is held we have bifurcated it to Technical and Non-technical while implementing it in our model for simplification.

## Data Analytics Tasks and Processes

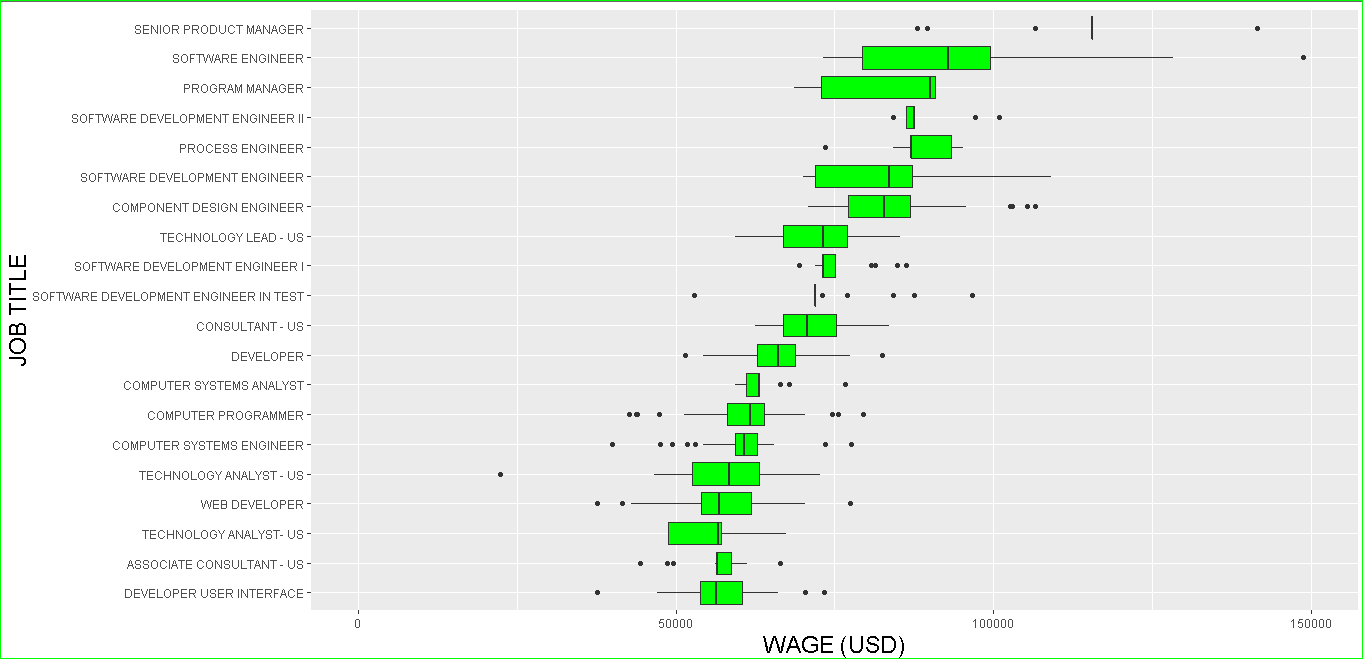
**#Plot for EMPLOYER NAME vs. NUMBER OF APPLICANTS:**



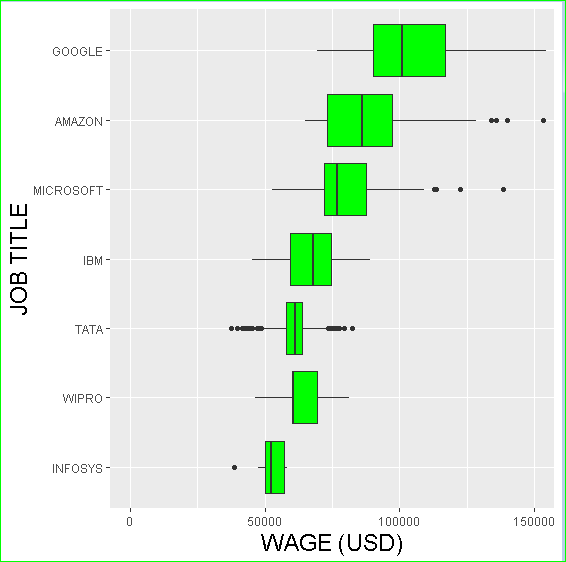
**#Plot for JOB TITLE vs. TOTAL NUMBER OF APPLICATIONS:**



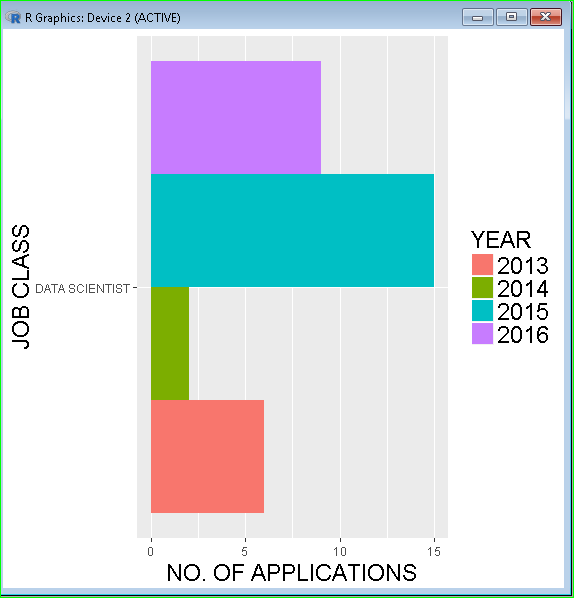
**# Box plot display for JOB TITLE and WAGES:**



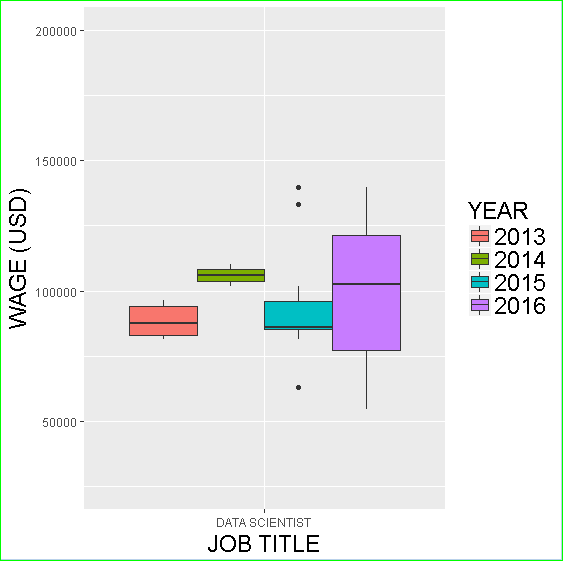
**# Box plot display for EMPLOYER and WAGES:**



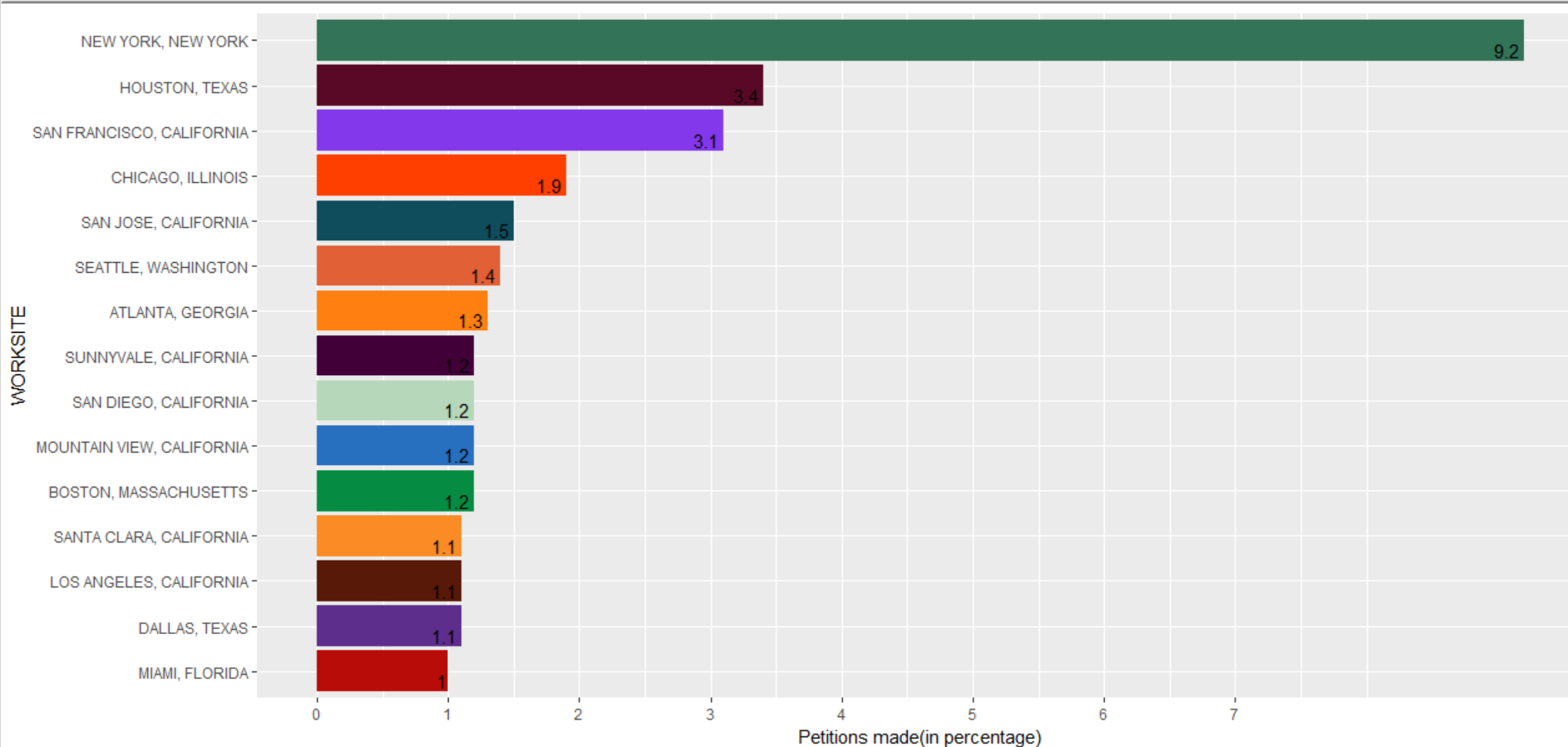
**# Plot for JOB CLASS vs. NUMBER OF APPLICATIONS:**

****

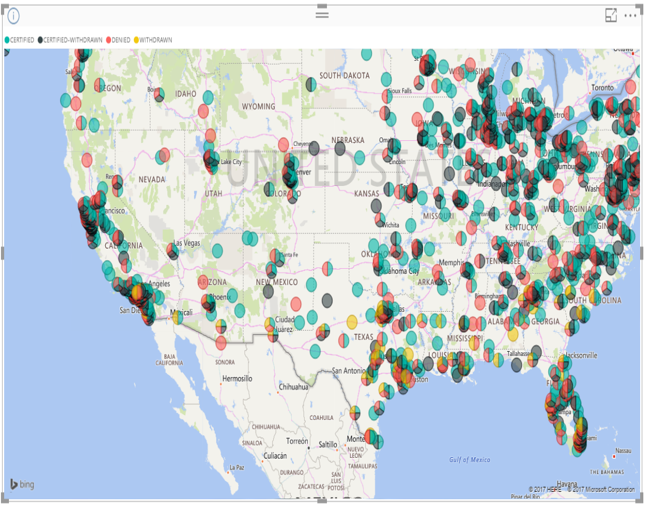
**# Plot for WAGES vs. JOB TITLES**

****

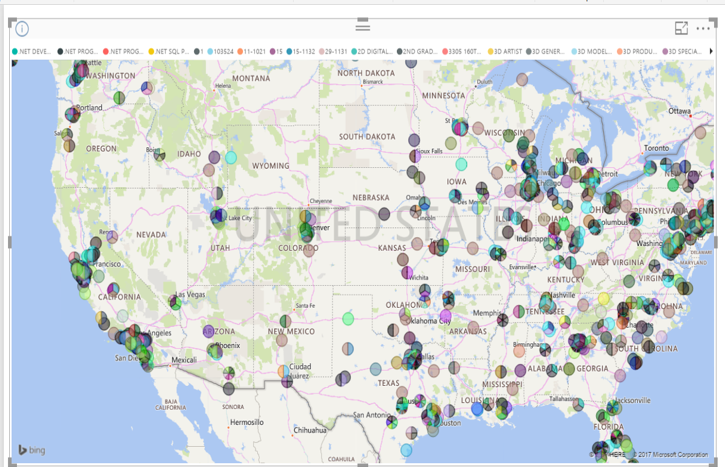
**# Plot for WORKSITE vs. PETITIONS MADE**

****

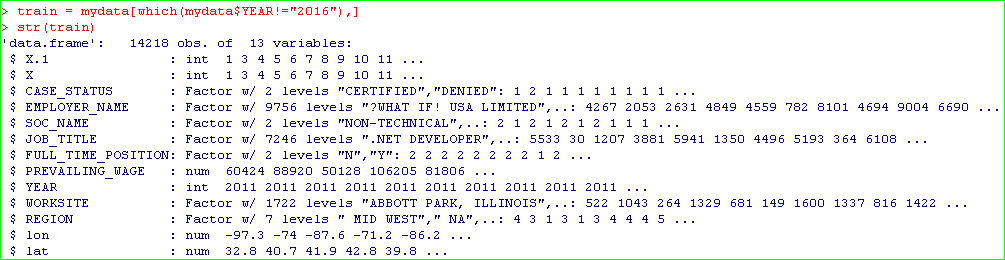
**# USA Map output for LOCATIONWISE CASE STATUS**

****

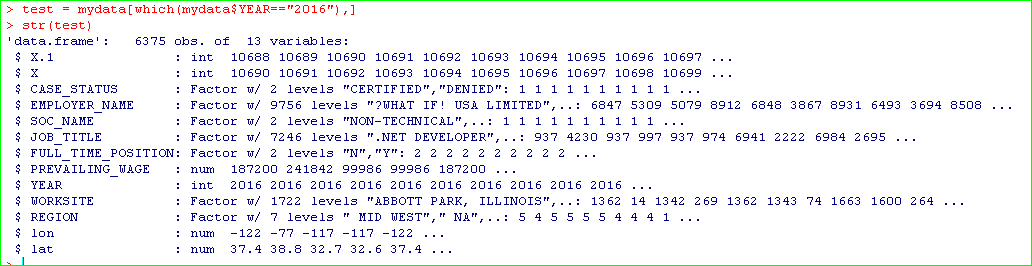
**# USA Map output for JOBS (nationwide):**

****

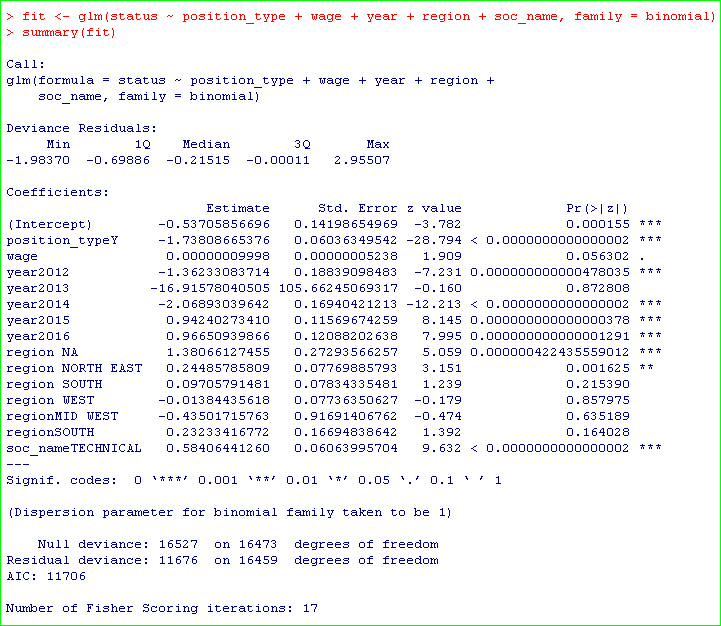
**# Snapshot of Train Data:**

****

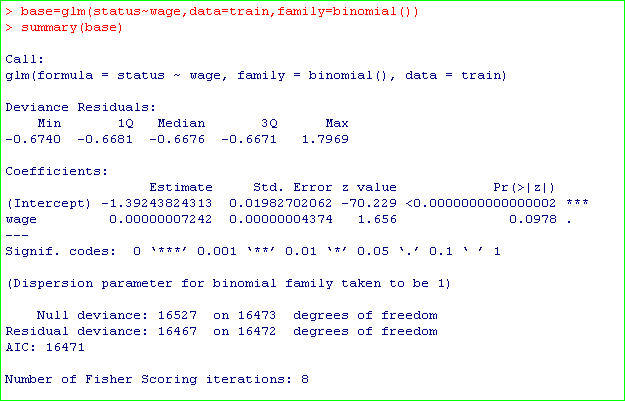
**# Snapshot of Test Data:**

****

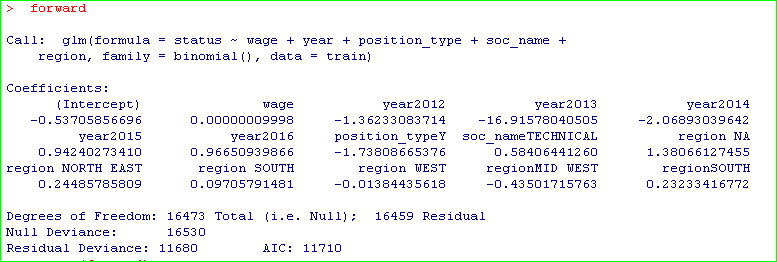
**# Snapshot of Initial Model with dependent and independent variables:**

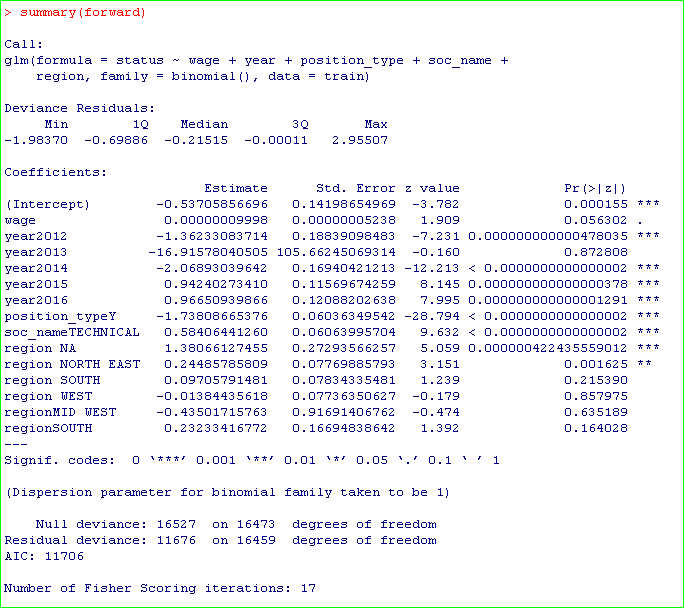
****

**# Snapshot of Base Model:**

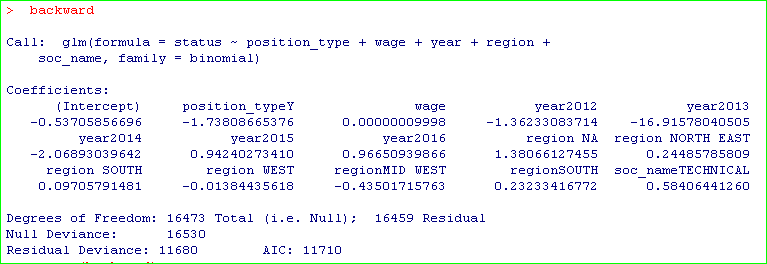
****

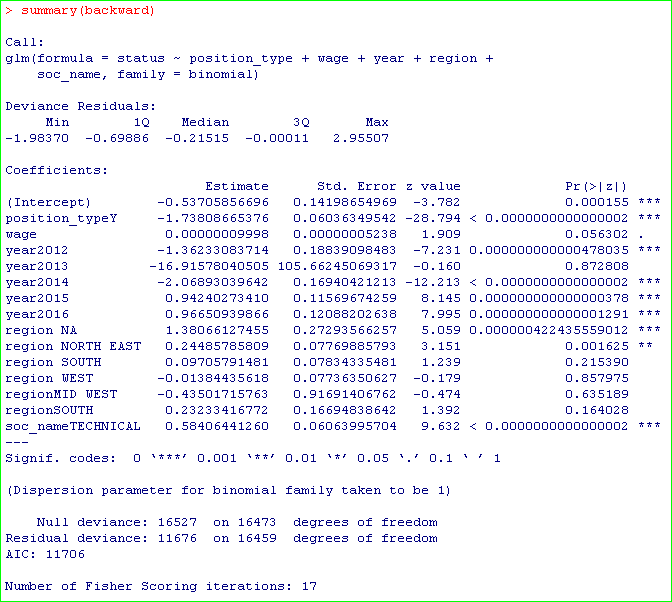
**# Snapshot of Forward Model:**

****

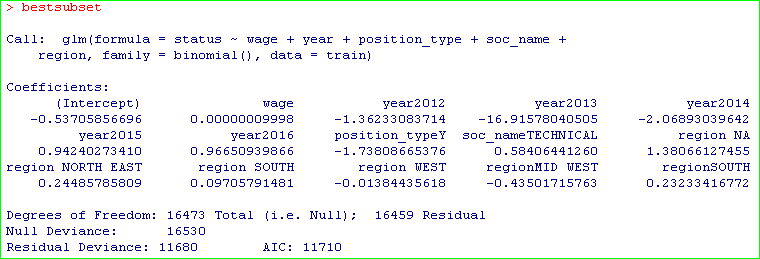
****

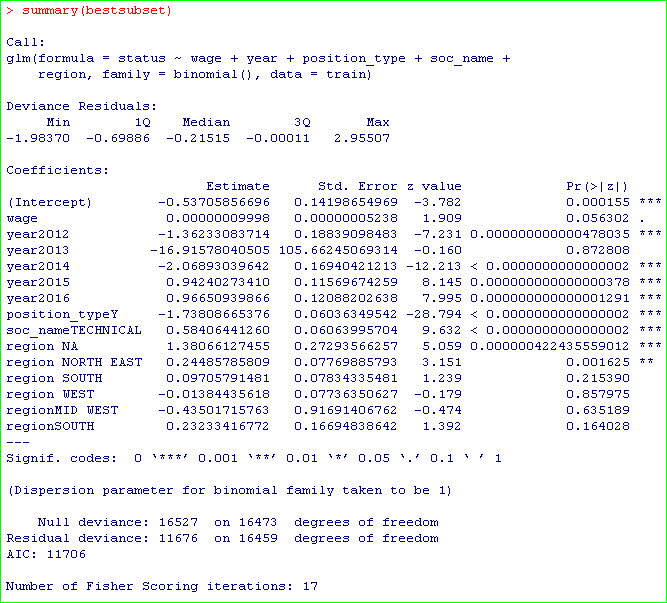
**# Snapshot of Backward Model:**

****

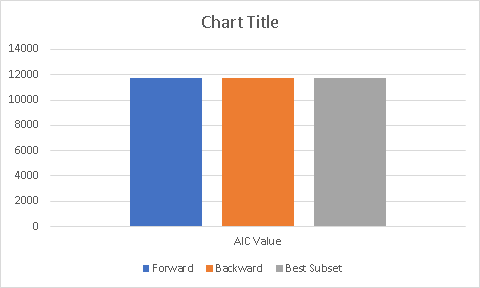
****

**# Snapshot of Best Subset Model:**

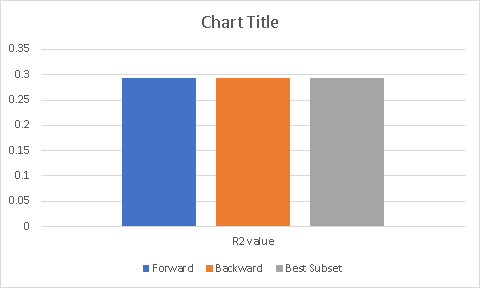
****

****

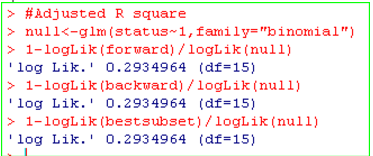
**# Plot for AIC value comparison between the models built:**

****

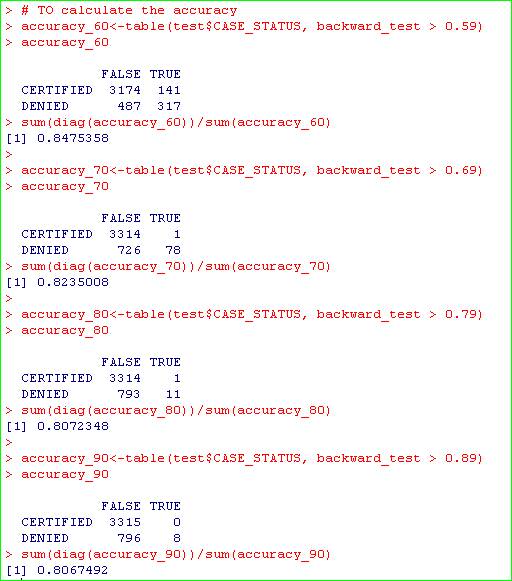
**# Plot for R2 value comparison between the models built:**

****

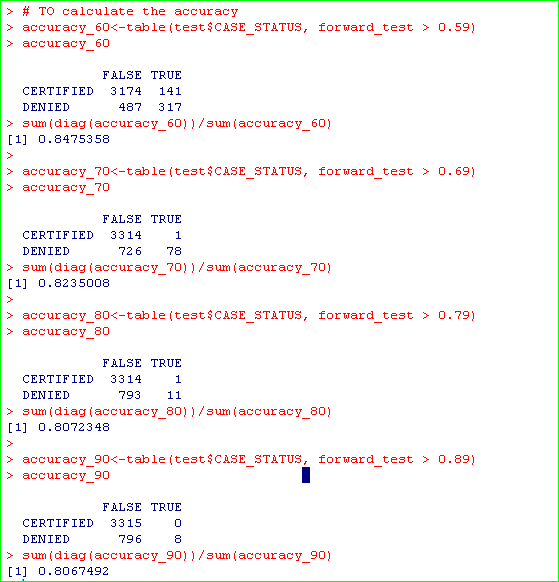
**# Snapshot for CP Mallow’s Adjusted R2 for forward, backward and best subset model:**

****

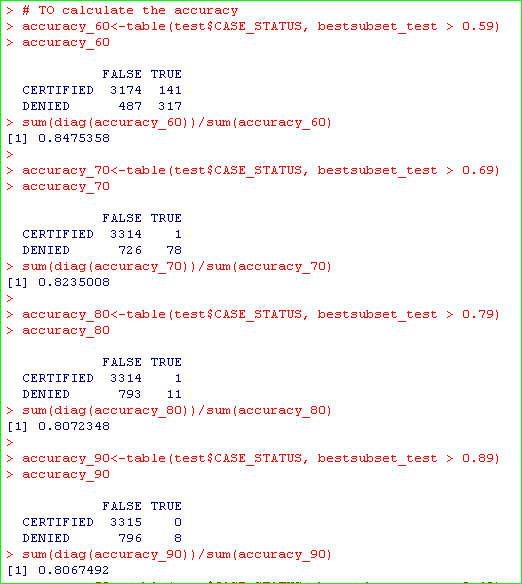
**# Snapshot for Backward Model Accuracy:**

****

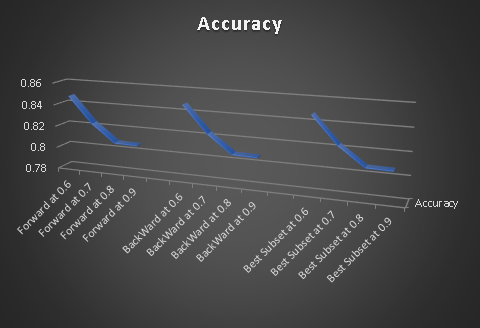
**# Snapshot for Forward Model Accuracy:**

****

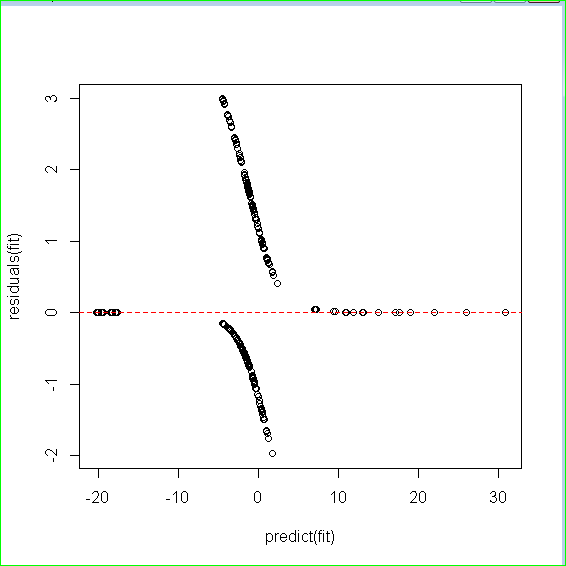
**# Snapshot for Best Subset Model Accuracy:**

****

**# Plot for Accuracy of Forward, backward and best subset model:**

****

**# Plot of Predicted vs. Residual values:**

****

# **Evaluations and Results**

## Evaluation Methods

## Results and Findings

# **Conclusions and Future Work**

## Conclusions

## Limitations

## Potential Improvements or Future Work