Team- STAR

(Phase 04: Data Exploration.)

Team Member 1**: Sridevi Jaidi** *[Head of the team]*
Email: sjaid1@unh.newhaven.edu

Team Member 2: **Shyam sunder Reddy Beeram**
Email: sbeer4@unh.newhaven.edu

Team Member 3: **Priyanka Nandigam**
Email: pnand3@unh.newhaven.edu

Content:

1.Introduction

About the dataset:

Dataset link: https://www.kaggle.com/datasets/nsharan/h-1b-visa

This dataset contains five years' worth of H-1B petition data, with approximately 3 million records overall. The columns in the dataset include case status, employer name, worksite coordinates, job title, prevailing wage, occupation code, and year filed.

In detail:

H-1B visas are a category of employment-based, non-immigrant visas for temporary foreign workers in the United States. For a foreign national to apply for H1-B visa, a US employer must offer them a job and submit a petition for a H-1B visa to the US immigration department. This is also the most common visa status applied for and held by international students once they complete college or higher education and begin working in a full-time position.

Research question:

Can we predict visa status of the applicants, by feeding the model with the dataset which contains the required fields by which the machine can classify the visa status as certified or denied.?

Data Exploration:

The list of exploration techniques (statistical and visualization techniques) used in this project are:

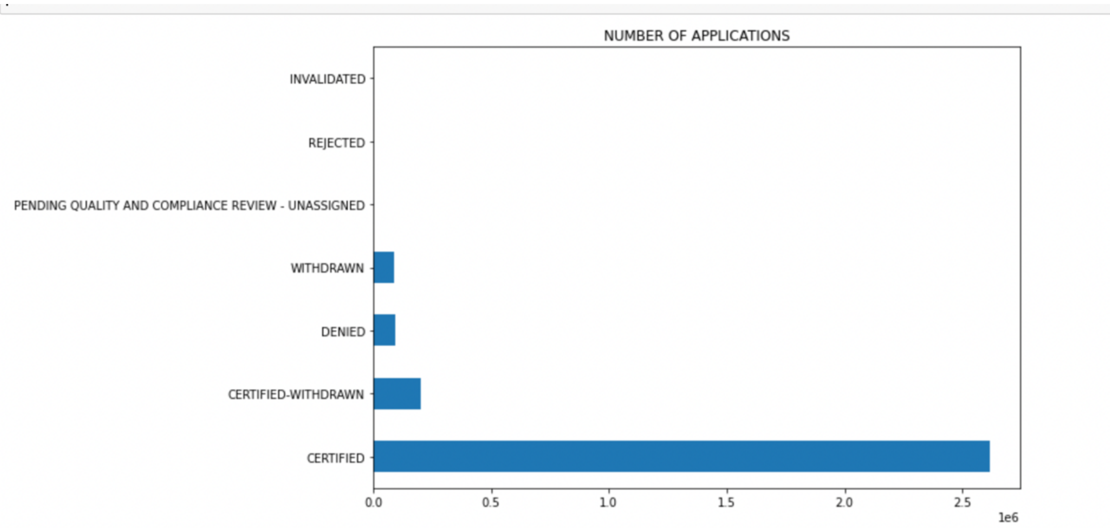
- Bar graph
- Histogram
- White grid graph
- Drawing plot
- Co-relation & Heatmap

After displaying five rows in the dataset, we understand how the dataset and its attributes. And also, we print information about data frame including the index type and columns non null values and memory usage. We also code to get to get a Series containing counts of unique values.

Then after cleaning the data, we start visualizing data in different perspectives.

Bar graph(i):

This is for the analysis of the case status of the applicants.

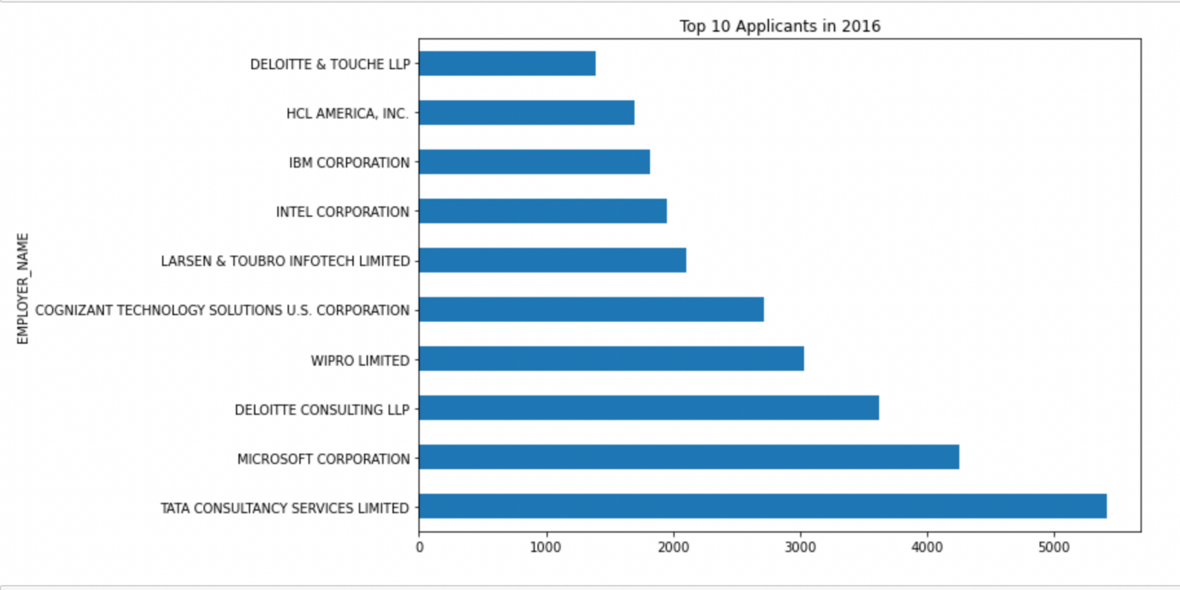


```
Number of Applications per year
```

The graph which we plotted here is a bar graph to know the case status of the applicants per year.

Bar graph (ii)

This is for the analysis of applicants who belong to top 10 companies in 2016.

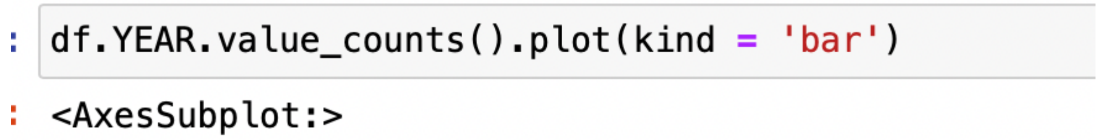


The graph which we plotted here is a bar graph to know the employer who has highest no of applicants in 2016.

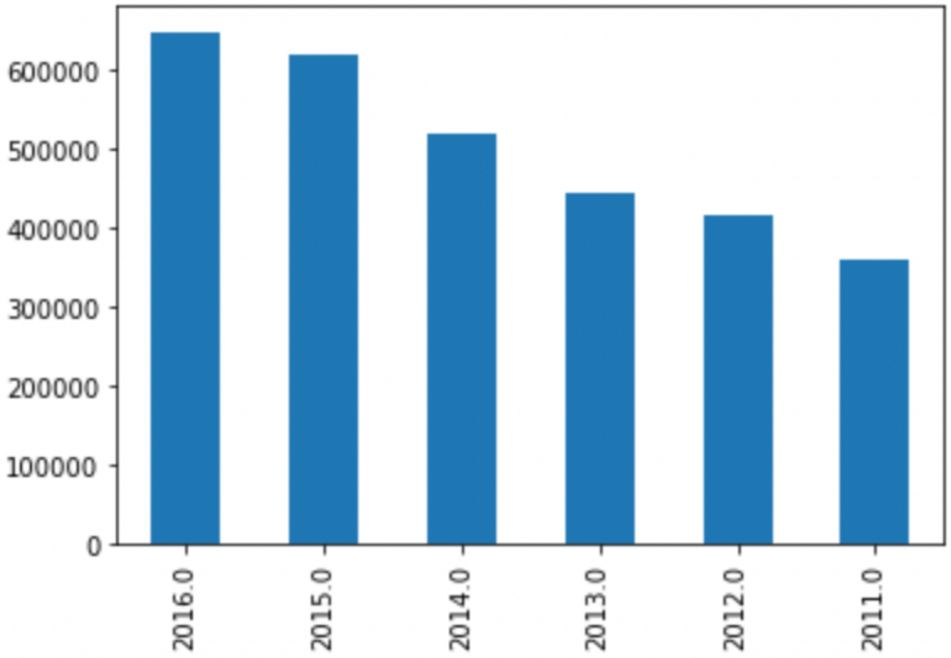
In the project there were few other bar graphs to know the status and employer who has good number of applicants.

Histogram:

This is for the analysis of no of applicants per year.



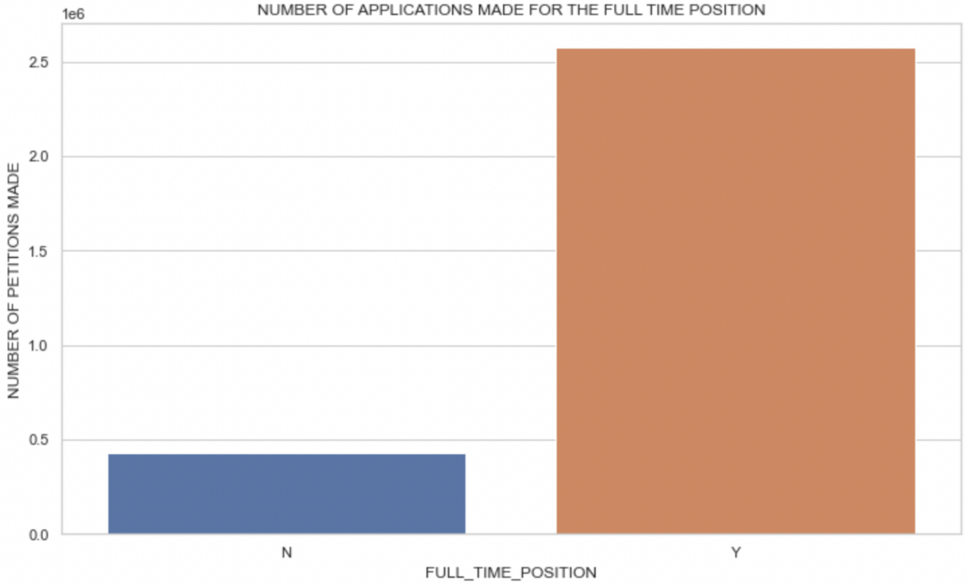
```
: df.YEAR.value_counts().plot(kind = 'bar')

: <AxesSubplot:>
```



The histogram shows the data of no of applicants filed visa in that respective year.

White grid graph:

This is for the analysis of Number of applicants made for the full time position.

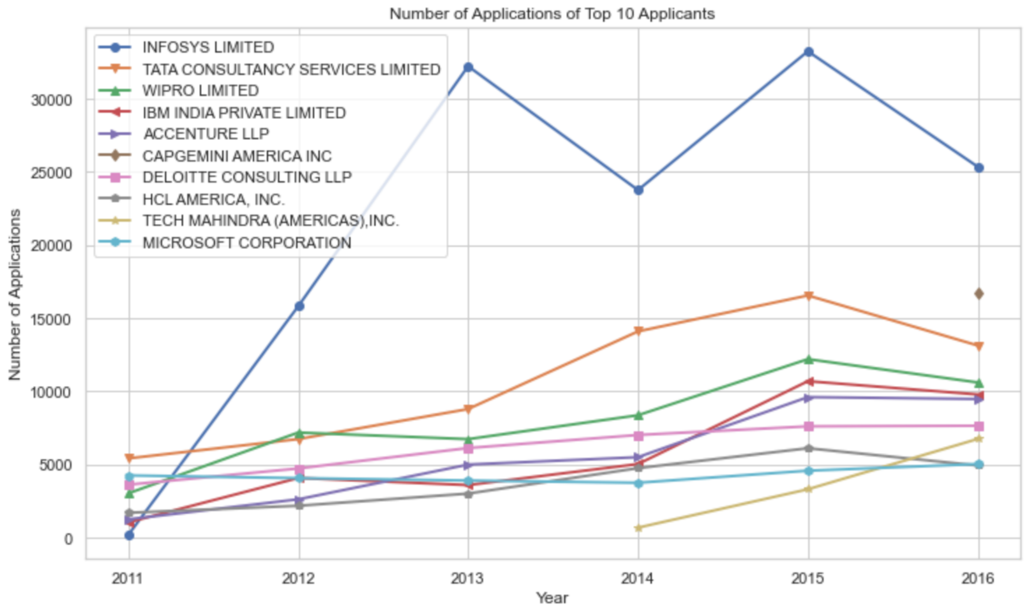


g = sns.countplot(x = 'FULL_TIME_POSITION', data = df)

Here, by using sns we are counting the full time positions and here X-axis is denoted for full time positions.

Drawing Plot:

This is for the analysis of number of applicants of the top 10 companies.



Here the graph tells about the total number of applications of the top 10 companies.
The plot of different color is for the respective company showed in menu graph.

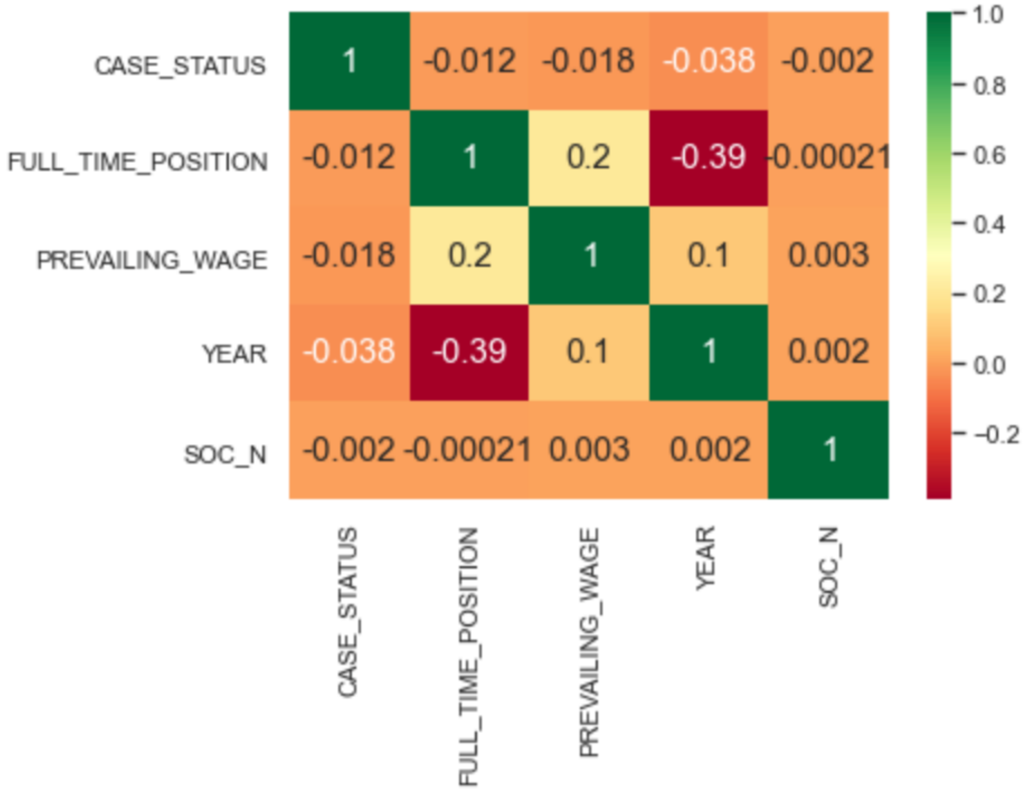
Co-relation & Heatmap:

Correlation is used for measuring the strength and direction of the linear relationship between two continuous random variables x and y. A positive correlation means the variables increase or decrease together. A negative correlation means if one variable increases then the other decrease.

Correlation values can be computed using the ‘corr()’ method of the Data Frame and rendered using heatmap.

According to our prediction we can drop columns which are not required or preferred.

```
<AxesSubplot:>
```



	CASE_STATUS	FULL_TIME_POSITION	PREVAILING_WAGE	YEAR	SOC_N
CASE_STATUS	1	-0.012	-0.018	-0.038	-0.002
FULL_TIME_POSITION	-0.012	1	0.2	-0.39	-0.00021
PREVAILING_WAGE	-0.018	0.2	1	0.1	0.003
YEAR	-0.038	-0.39	0.1	1	0.002
SOC_N	-0.002	-0.00021	0.003	0.002	1

Conclusion:

The objective this analysis is to observe of data using summarization basic statistical measures and visualization. Matplotlib and seaborn are the two most widely used libraries for creating a visualization. Plots like drawing plot, histograms, barographs, heatmap, can be created to find insights during exploratory analysis. Now we are able to find on what parameters machine can decide whether the status is certified or denied. Finally, after exploring our dataset, we have a brief idea on companies that files visa, applicants, and their status (fulltime or parttime), no of applications made per year.

We already done with Checking for outliers and removing them. Our next steps include building the model for our trained data.

Github:

https://github.com/shyambeeram/STAR