**Priya Rao, Pratik Khedekar, Namrata Patel, Stuti Garg**

**Joshua Introne**

**IST 707 Applied Machine Learning**

**08 December 2022**

**H-1B Data Analytics**

***1.1 Introduction:***

**Non-Immigrant Visa:**

A nonimmigrant visa is a visa issued to individuals who are temporarily entering a country for a specific purpose, such as tourism, business, or study. Nonimmigrant visas are typically issued for a specific period, after which the individual must leave the country. In contrast, an immigrant visa is issued to individuals who are moving to a country permanently.

**Most popular Non-Immigrant Visa Categories:**

B-1 - Business visitors.​

B-2 - Visitors for pleasure or medical treatment.​

F-1 - Academic or language students.​

F-2 - Spouses and children of F-1 visa holders.​

**H-1B** - Persons working in specialty occupations requiring at least a bachelor's degree or its equivalent​

J-1 - Exchange visitors coming to the U.S. to study, work, or train as part of an exchange program officially recognized by the U.S. Department of State.​

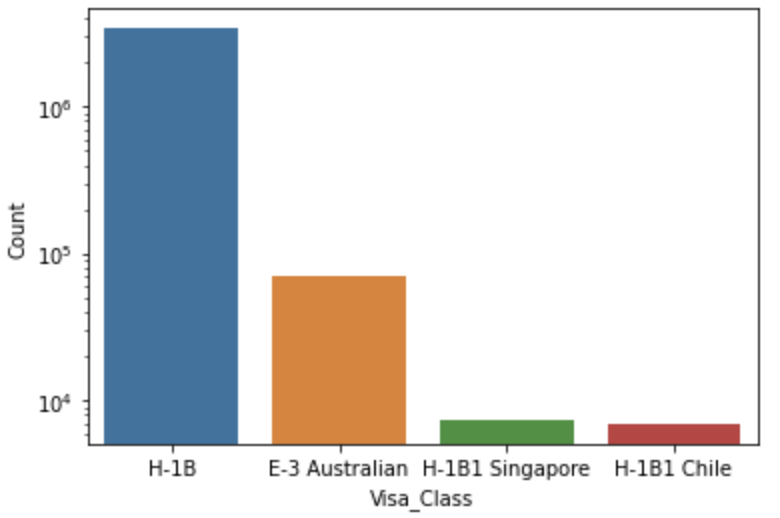
J-2 - Spouses and children of J-1 visa holders.​

**H-1B Data:**

* 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 an H1-B visa, a US employer must offer them a job and submit a petition for an 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 full-time.​
* Every year, around 65,000 H-1B visas are given, with an additional 20,000 H-1B visas held as an extra cap. ​

There are 4 types of H-1 B visas:

**H-1B, H-1B1(Chile & Singapore), E-3 Australian, and E3(Australia)**



**Goal:**

The H-1B visa program allows companies in the United States to employ foreign workers in specialty occupations temporarily. H-1B data analytics refers to analyzing data related to the H-1B visa program, such as the number of visas granted, the industries and companies that sponsor the most visas, and the countries of origin of H-1B visa holders. This analysis can provide valuable insights into the use and impact of the H-1B program on the U.S. workforce and economy.

**Motivation**: Determining the factors that affect the H1-B grant and creating a prediction model as H-1B is a lottery system it is assumed that there are no external factors that affect the process, our goal is to find those factors that play a vital role in the H-1B visa selection.

***1.2 Methods:***

**1.2.1 Data Description:**

Data collected from 2017 to 2022 from the United States Citizenship and Immigration Services. ​

There are 3.5 million lakh observations and 13 attributes. ​

* **Visa Class:** Indicates the type of temporary application submitted for processing. It includes H-1B, E-3 Australian, H-1B1 Chile, and H-1B1 Singapore. ​
* **Employer Name:** Legal business name of the employer submitting the Labor Condition Application ​
* **SOC title**: Occupational title associated with the SOC/O\*NET Code ​
* **Job title:** Title/ position of the job.​
* **Full-time position:** Y = Full-Time Position; N = Part-Time Position ​
* **Worksite:** Geographic Information for First Worksite Location.​
* **Wages:** Wage paid to nonimmigrant workers at the First Worksite Location.​
* **Unit of Pay:** Unit of pay values include “Hour”, “Week”, “Bi-Weekly”, “Month”, and “Year” ​
* **Employer Location:** Contact information for the Employer Point of Contact submitting a Labor Condition Application ​
* **Employer Country:** Location of employer ​
* **Case Status:** Status associated with the last significant event or decision. Valid values include “Certified”, “Certified-Withdrawn”, Denied”, and “Withdrawn”. ​
* **Wage rate of pay:** Wage paid to nonimmigrant workers. ​
* **Quarter:** There are 4 quarters in each year​
* **Year:** Application year

**1.2.2 Data Wrangling:**

* Cleaning, manipulating, and organizing data so that it can be used for analysis or other purposes are known as data wrangling.
* Firstly as part of Data Cleaning, we made the Unit of pay column to one scale i.e. Yearly Income.
* We divided each year into quarters to have a better understanding of how H-1B visas are picked in each quarter.

**Handling Missing Values:**

* We dropped the rows which have NULL values.
* Imputed the prevailing wage column with a median.
* We imputed the nan values in Employer\_country with the United States of America as the H-1B visa is specific to the US.
* To avoid redundancy we changed variables to one text format(Uppercase).
* As we are considering only H1-B for the USA. We dropped the other values in Visa Class.
* We maintained the Data Type consistent

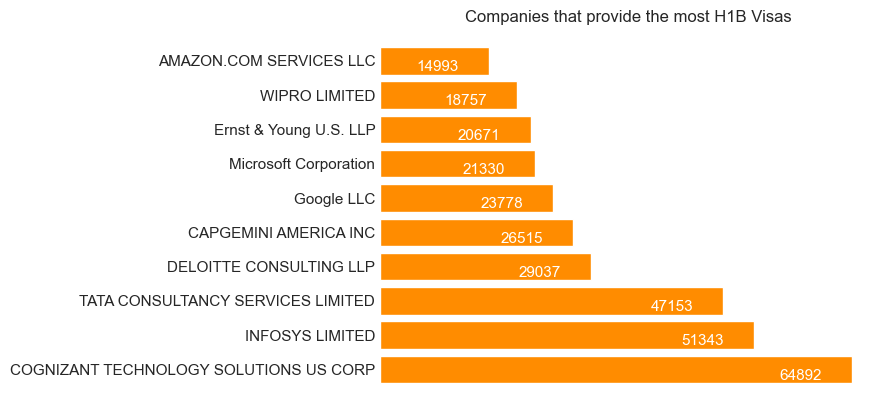
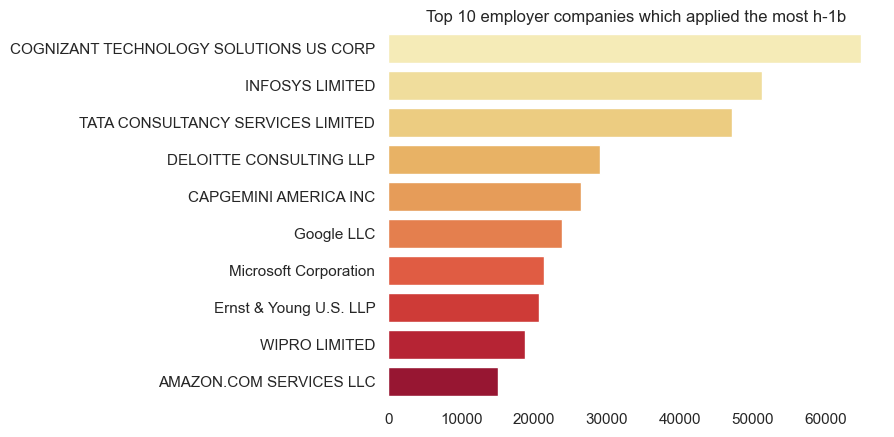
We downsampled the data to perform better analysis as our dataset is very larger, but downsampling resulted in a biased proportion of our target variable i.e case\_status.

**1.2.3 Exploratory Data Analysis:**

Exploratory data analytics is a type of data analysis that focuses on exploring and understanding data sets. The goal of exploratory data analytics is to discover patterns and relationships in the data that may not be immediately obvious.

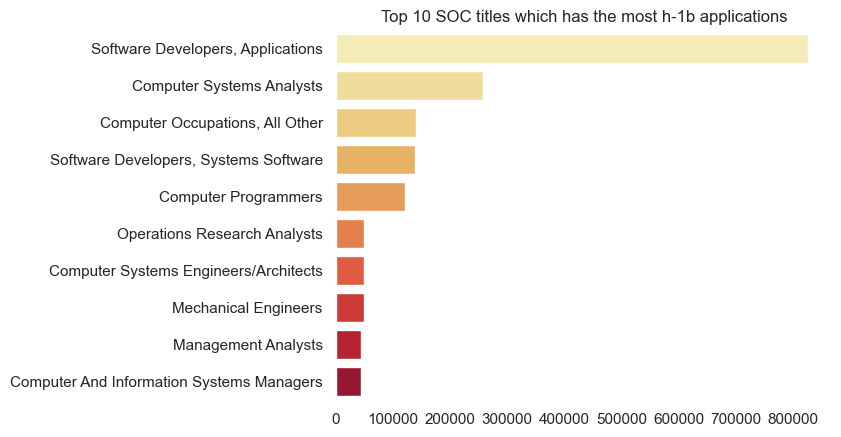
Exploring various factors that could affect getting an H-1B

**Top employers which filed the most h-1b applications over the years:**

****

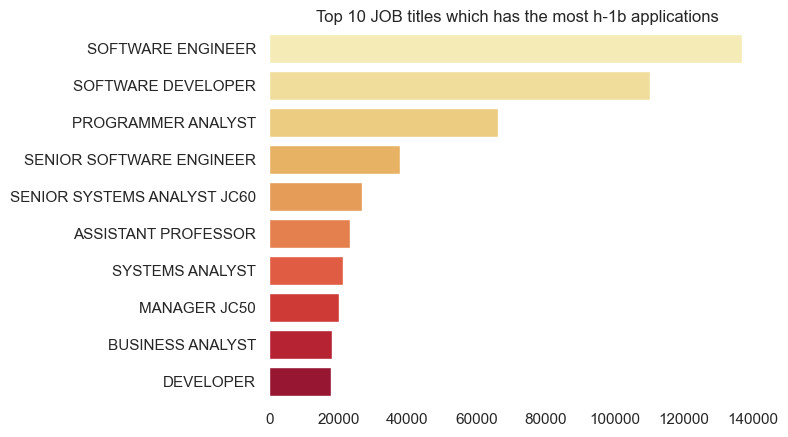
* It can be seen from the plot above that Cognizant, Infosys, and TCS filed the most H-1B.
* Cognizant files around 64,892 applications for H-1B while amazon filed 14993

**Top SOC titles which had the most h-1b applications over the years:**

****

* It can be seen from the plot above that Software Developers, Computer System Analysts, and Computer Occupations had the most H-1B.

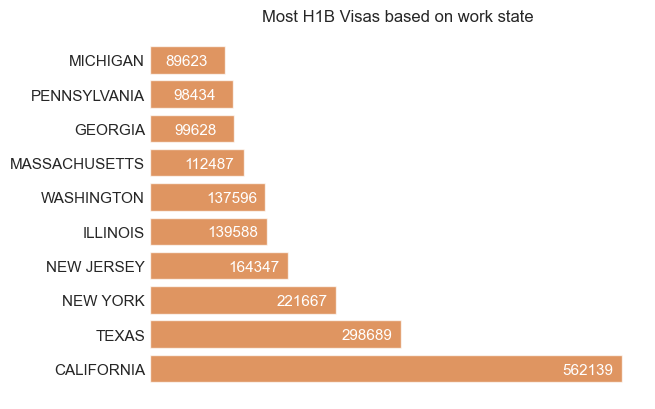
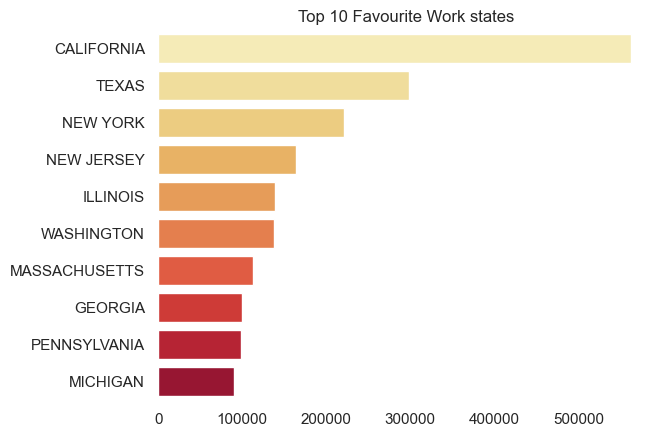
**Top JOB titles which had the most h-1b applications over the years:**

****

* It can be seen from the plot above that the Software Engineer, Software Developer, and Programme Analyst had the most H-1B applications.

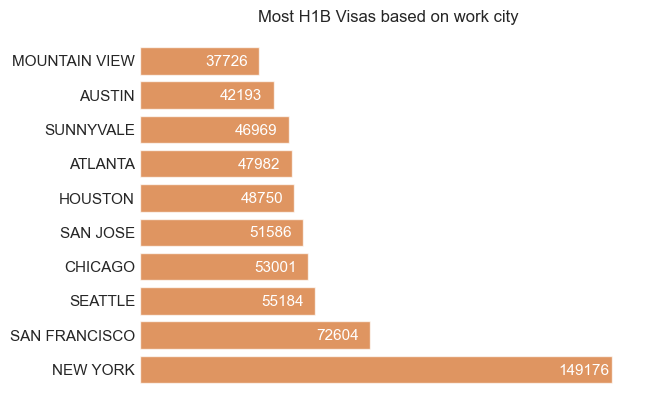
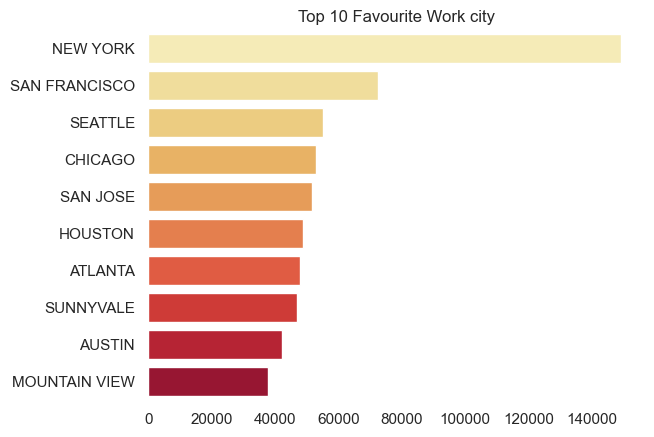
**Top 10 Favourite Work cityTop 10 Favourite Work ci**

**Top 10 Favourite Work State:**



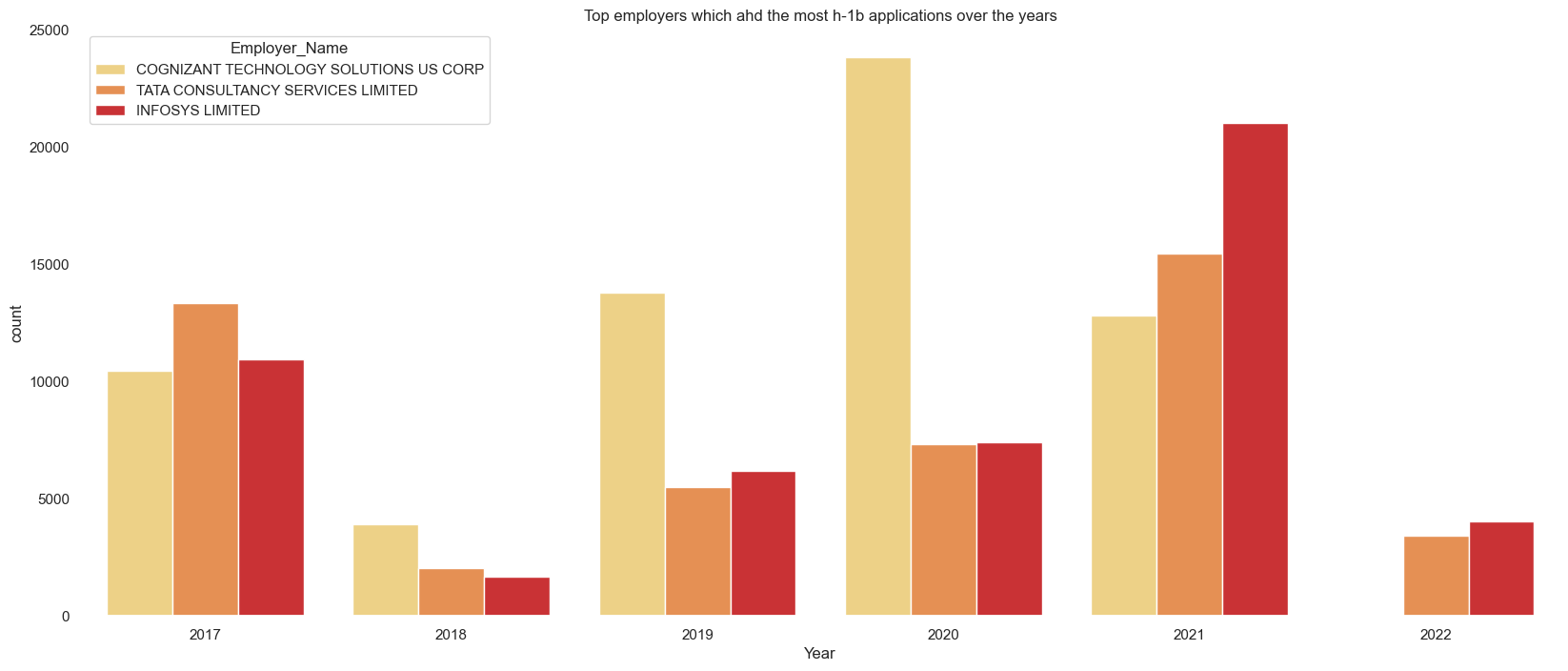
* California and Texas have the most H-1B visa application.
* California has 562139 applications which is the highest, followed by Texas which is 298689.

**Top 10 Favourite Work City:**

****

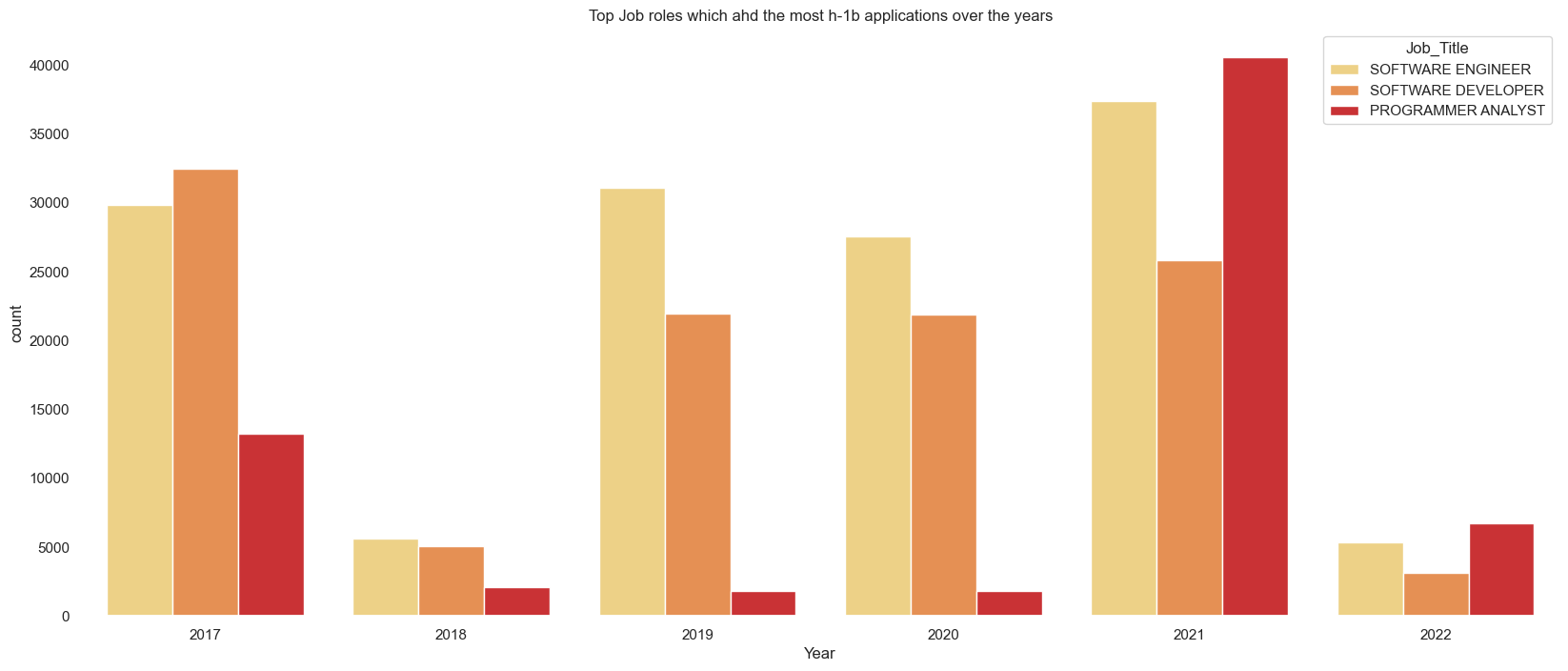
* New york city and San Francisco had the most H1-B applications, while if you see based on the states Texas had one of the most applications along with CA.
* In CA, Mountain View ( silicon valley) had the most H-1B applications being the tech hub but surprisingly it is not the city with the highest H-1b application, NY has the most H-1B application with 149176 applications.

**Top 3 employers which had the most H-1B applications trend over the years:**

****

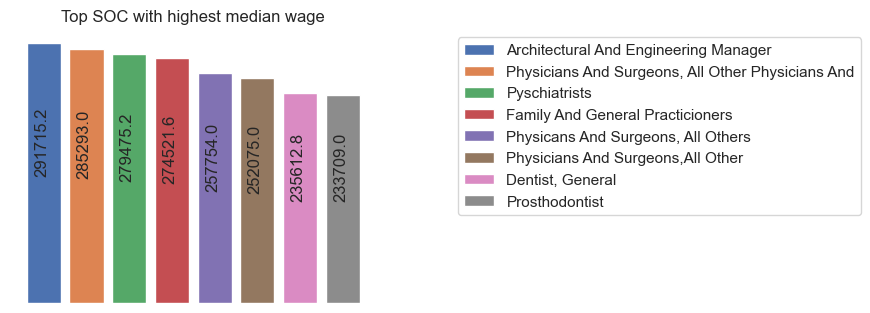
* The top three companies which gave the most H-1B it’s the trend over the years.
* Cognizant gave the most H-1B during covid i.e. 2020, TCS also gave the most visas during covid i.e. 2021, and Infosys too gave the most visa application in 2021.

**Top 3 Job-Title which had the most H-1B applications trend over the years:**

****

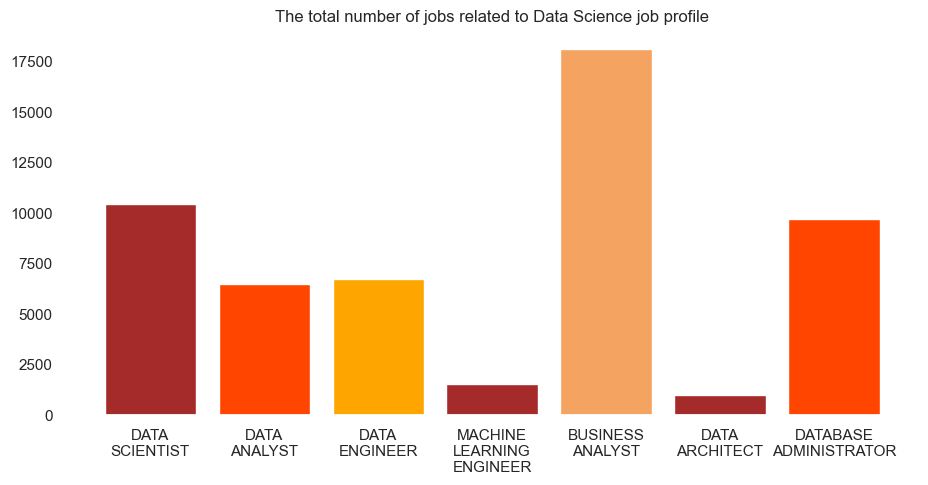
* Software Engineer job roles were given the most H-1B applications in 2021.
* Software Developer job roles were given most H-1B applications in 2022.
* Programmer Analyst job roles got the most H-1B in 2021, while had the least application in 2019 and 2020.

**Top Profession with highest median wage:**



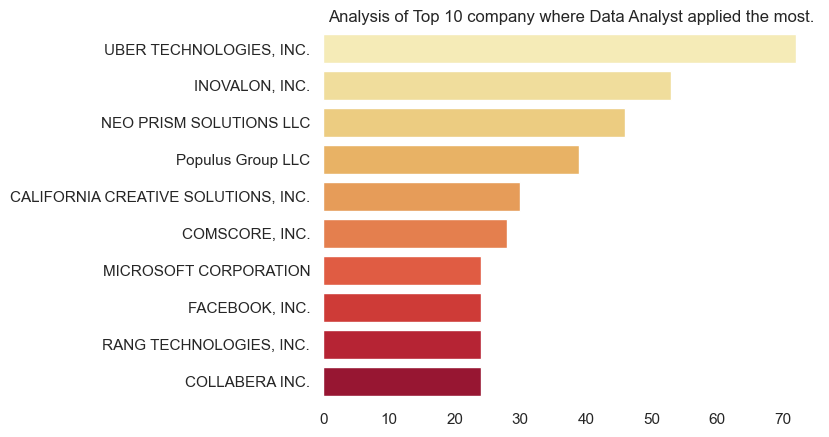
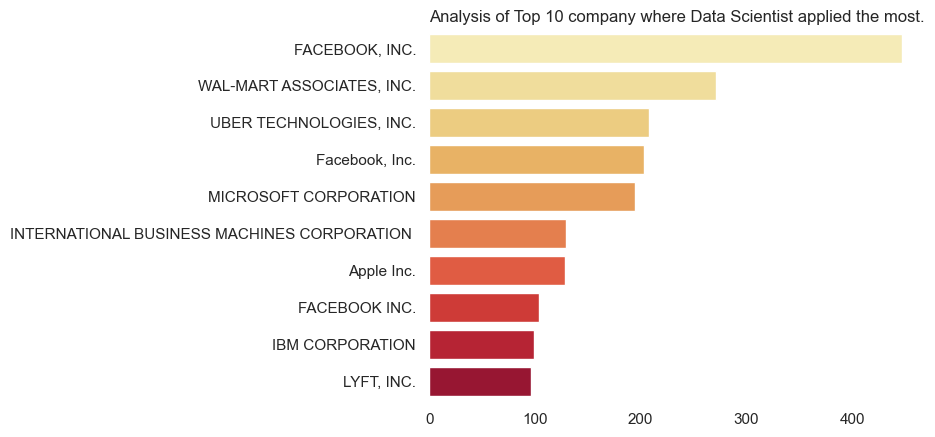
* It could be seen from the plot above that architecture and physician were the highest paid profession.

**Top Data Science job roles:**

****

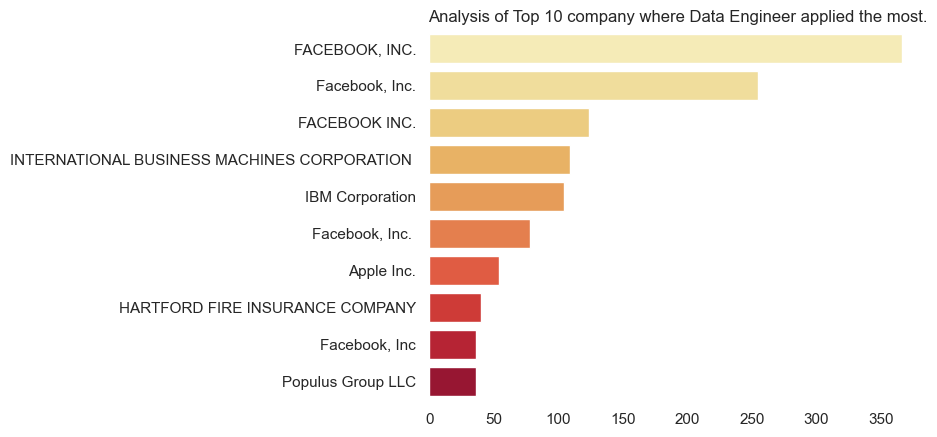
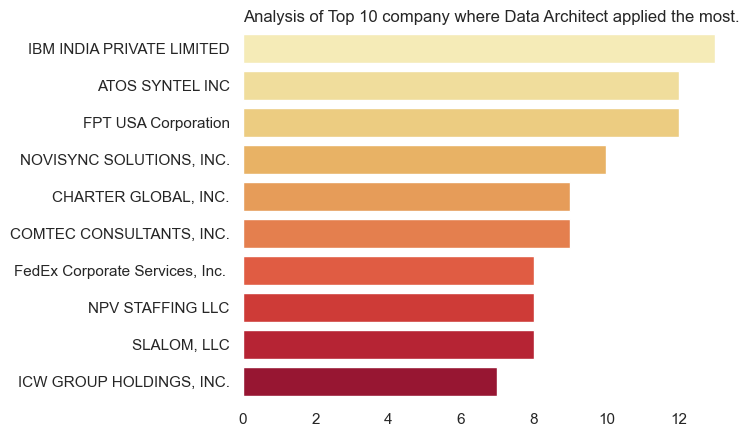
* Data-related job roles came into the picture in 2018, so it is a new domain in comparison to software engineering and development roles.
* Business analyst job roles were among the most common in H-1B applications.
* The total number of jobs related to the data profile is 54125

**Analysis of each job profile related to the Data domain to understand the trends in the Data job profile and its H-1B application:**

**Data Analyst & Data Scientist Job roles and companies that applied most H-1 B :******

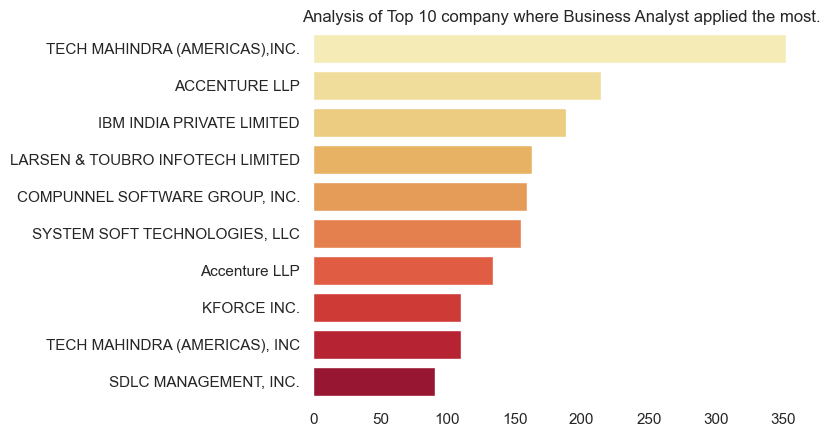
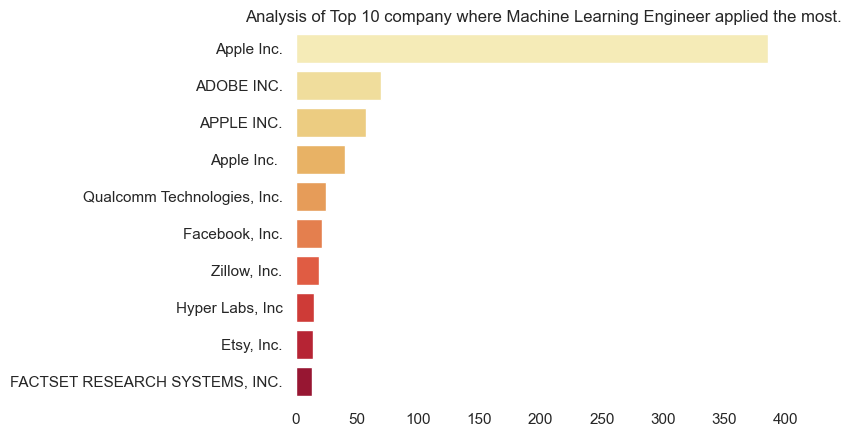
* Uber is the top most employer that filed most H-1B for Dana Analyst roles.
* Facebook(Meta is the top most employer that filed most H-1B for Data Scientist roles.

**Data Architect &Data Engineer Job roles and companies that applied most H-1 B for that roles:**

****

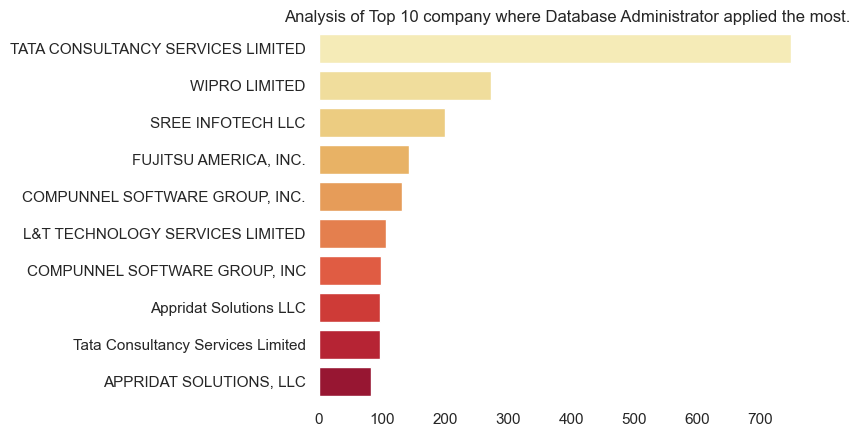
* IBM is the top most employer that filed most H-1B for Data Architect roles.
* Facebook(Meta) is the top most employer that filed most H-1B for Data Engineer roles.

**Machine Learning Engineer & Business Analyst Job roles and companies that applied most H-1 B for that roles:**



* Apple hired the most ML engineers and applied for H-1B.
* Tech Mahindra filed the most H-1B for BA job roles.

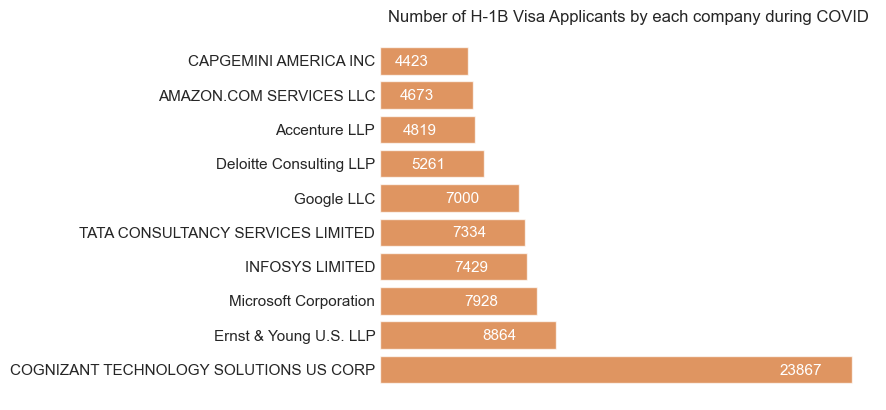
**Database Administrator Job roles and companies that applied most H-1 B for that roles:**



* TCS filed most H-1B for Data Administrator roles.

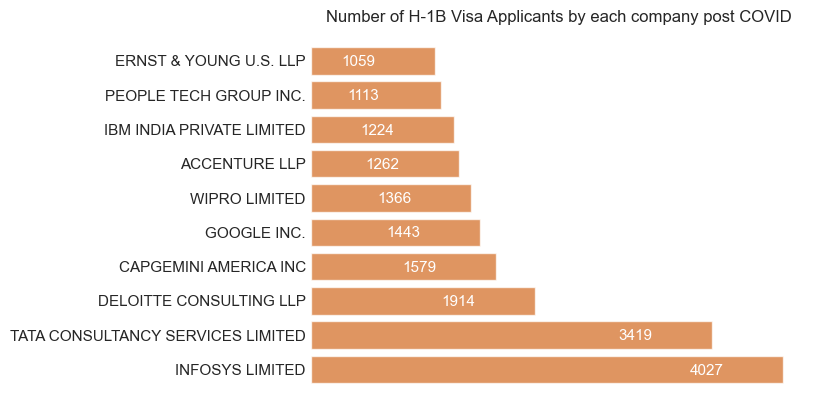
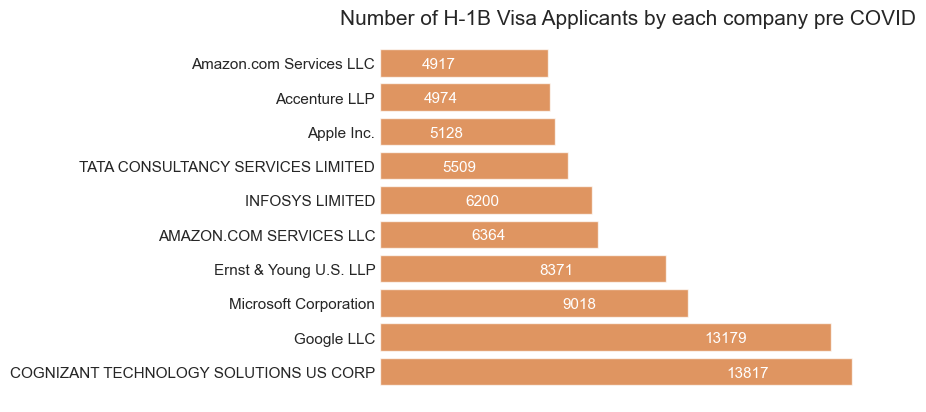
**Analysis of how COVID-19 affected the H-1 application:**

**Companies that filed most H-1b During Covid-19**

****

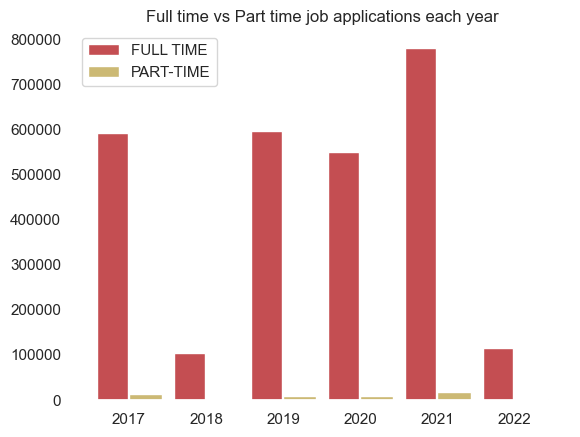
* As we can see from the plot above Cognizant had filed the most H-1B during Covid as well followed by Ernst & Young.
* After covid google’s h1-b application decreased to what it was in 2019

**Companies that filed most H-1b pre-Covid-19 & Post-Covid-19**

****

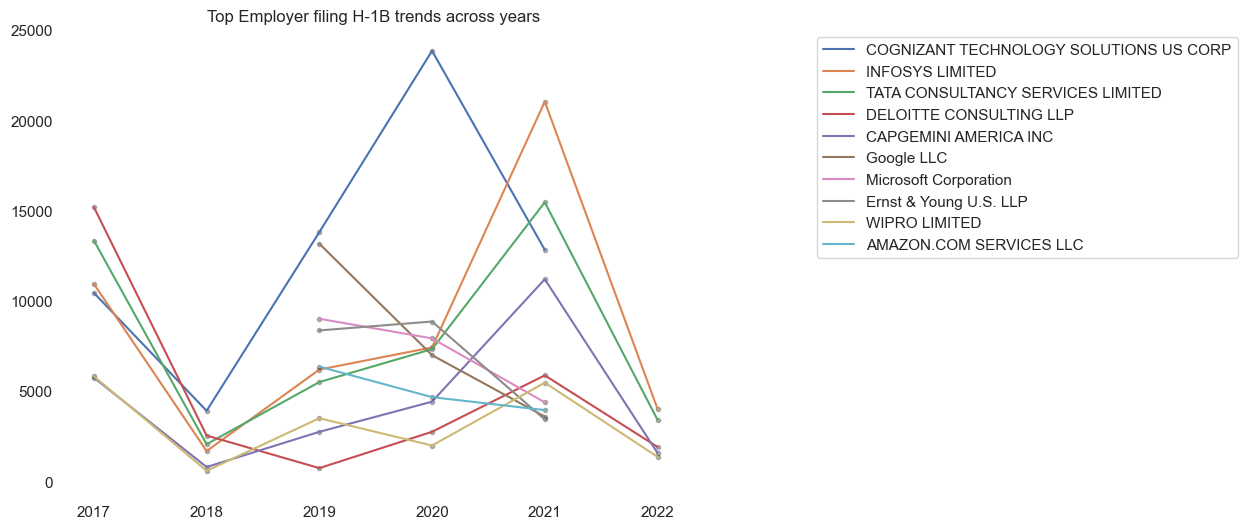
* Before covid Cognizant and Google filed the most H-1B applications
* After Covid, the trend completely changed Infosys was the top company that filed the most H-1b, followed by TCS.
* Surprisingly cognizant was not among the top companies which filed most H-1b post-Covid-19.

**H-1B application based on the type of job:**

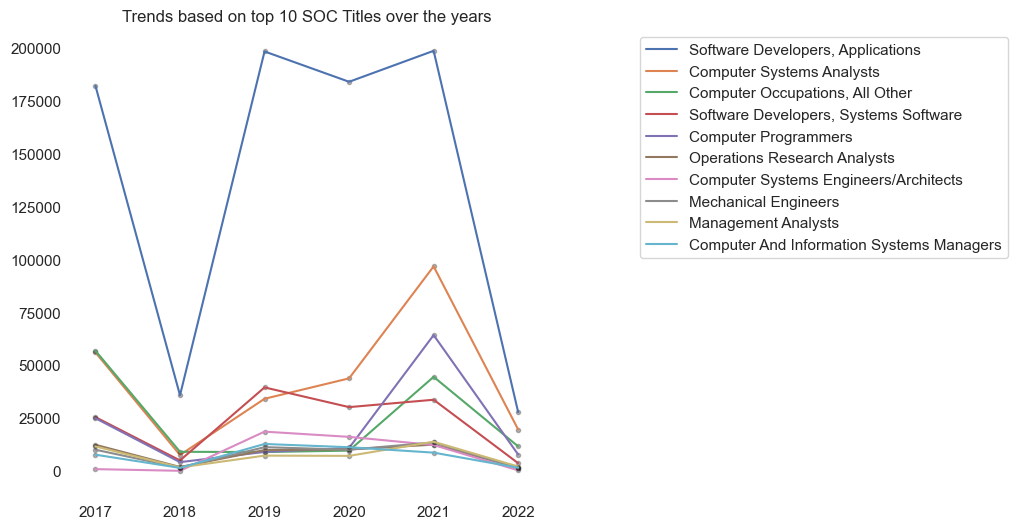


* We can see from the plot above H-1 applications are filed most for Full-time job roles and it’s extremely low for part-time job roles.

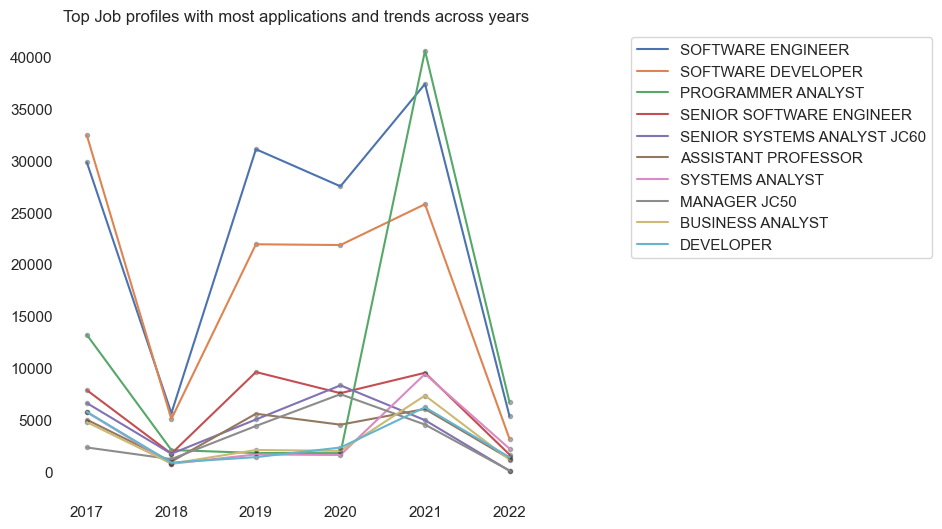
**Understanding trends in H-1 application over the years based on SOC title, Job title, and Employers.**

****

* The most variation in H-1B application made was seen for Cognizant which had its peak application in 2020, with a downfall after Covid i.e 2021.
* Infosys had its peak in application after COVID, which was quite low pre and during COVID.

****

* Software development had the most applications for H-1 over the years, only has a dreadful fall in 2018 which could be when data science roles were introduced and were on trend.
* After covid Computer analysis domain had its peak in H-1B applications.



* Programmer analysts had the most application in 2021, which had a drastic rise from what it was before 2020.

**1.2.4 Feature Engineering:**

Feature engineering is the process of using domain knowledge to extract features from raw data that can be used in machine learning algorithms.

* **Performed Bucketing**, also known as discretization, is the process of dividing continuous data into intervals or "buckets" based on the values of the data. This can be useful when working with data that has a large range of values, as it can make it easier to analyze and interpret the data. Constructed new features for wage \_categorization into 5 categories: very low, low, medium, high, and very high, and applied the new features to the prevailing\_annual\_wage.
* Performed **One-hot encoding**, a technique used in machine learning to encode categorical variables as numerical data.

Performed encoding on Wage\_Category ,Year, Work\_State, Full\_Time\_Position, SOC, Job\_tile, and Employer.

* **Recursive feature extraction** was used in machine learning to identify the most important features in a dataset. We extracted the 40 most correlated features to our target variable out of 75.

We performed SMOTE Analysis to upsample our data in order to balance our target variable category i.e case\_status(Accepted& Denied).

* **SMOTE (Synthetic Minority Oversampling Technique)** is an oversampling method used in imbalanced classification to increase the number of samples from the minority class. This is done by creating synthetic samples from the minority class, rather than simply duplicating existing samples. SMOTE works by selecting a sample from the minority class and finding its k nearest neighbors. It then generates new samples along the line segments connecting the selected sample and its neighbors, effectively increasing the number of samples in the minority class. This can help improve the performance of machine learning algorithms on imbalanced data sets
* After Performing SMOTE analysis our Data increased as we upsampled the data to get rid of the biasness in the minority class.

***1.3 RESULTS***

**1.3.1 (Data Mining )**

* Data mining is used to uncover hidden trends, predict future events, and identify relationships among different data elements using machine learning.
* Created a training and test set for a machine learning model by splitting the data into two subsets: a training set and a test set.

**Classification:**

**Decision Tree Classifier:**

* The decision Tree algorithm works by creating a tree based on the training data, with each node in the tree representing a decision that splits the data into smaller and smaller subsets based on the values of the input features.

**Logistic Regression:**

* A logistic regression classifier is a type of binary classification algorithm that uses a logistic function to model the probability of a binary outcome. The logistic function takes in a continuous input and outputs a value between 0 and 1, which can be interpreted as the probability of the input belonging to one of the two classes.

**Random Forest Classification:**

* A random forest classifier is a type of ensemble machine learning algorithm that uses multiple decision trees to make predictions. The multiple decision trees are trained on different subsets of the training data, using a random subset of the input features at each split. The predictions from all of the trees are then combined to make a final prediction.

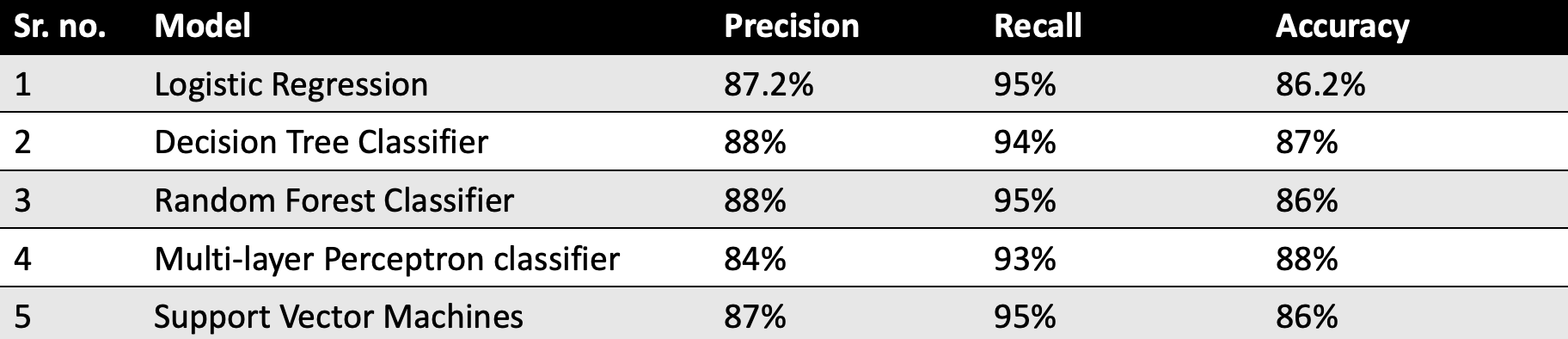
**Multi-layer Perceptron classifier:**

* A multi-layer perceptron (MLP) classifier is a type of neural network that uses a series of fully-connected layers to make predictions

**Linear Support Vector Classifier:**

* A support vector machine is a supervised learning algorithm that finds the hyperplane in a high-dimensional space that maximally separates the two classes. In the case of a linear SVC, the hyperplane is a linear boundary that separates the two classes.

**The precision, recall, and accuracy of all the classification models are below:**



The Multi-Layer Perceptron classifier had the best accuracy which is 88%.

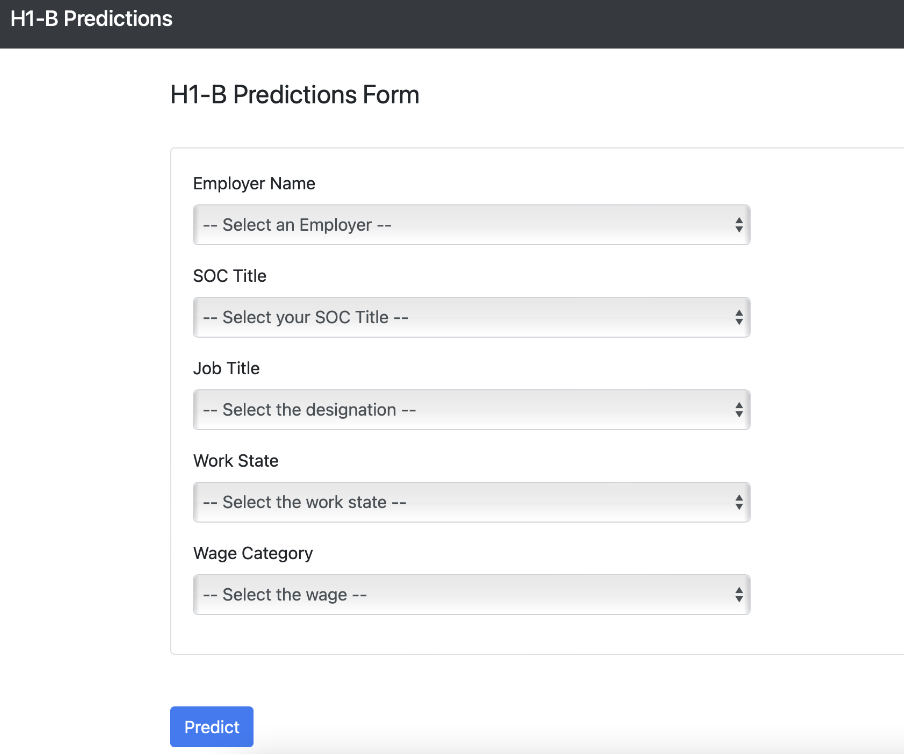
**1.3.2 Hyperparameter Tuning**

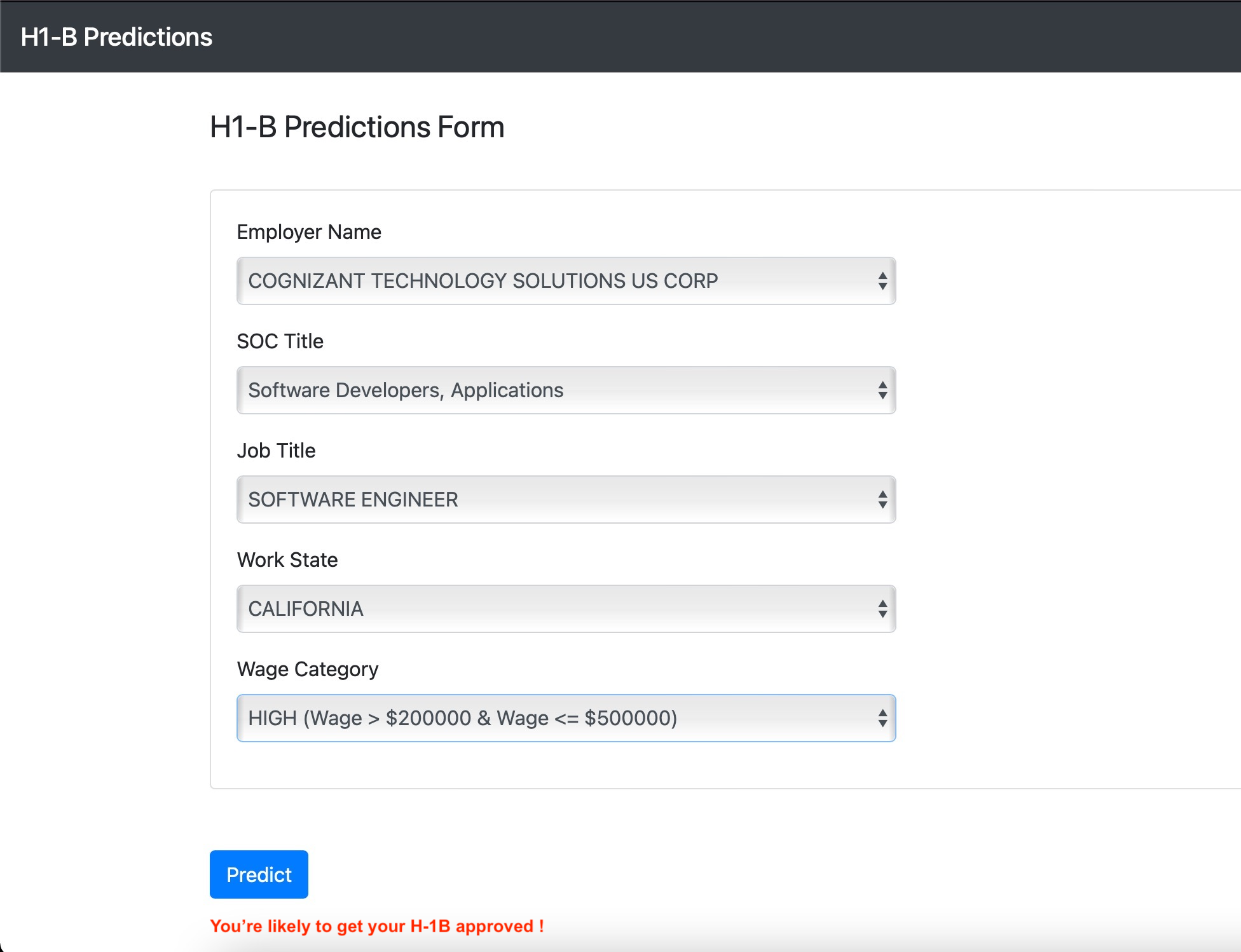
* Hyperparameter tuning is the process of selecting the optimal values for the hyperparameters of a machine-learning model.
* We Dropped the COVId year’s data i.e 2020 and 2021 to understand if Covid affected the H-1B allocation process.
* We performed a Decision Tree, Logistic Regression, and Random Forest Classifier with no covid years
* We used hyperparameter tuning to on decision tree using grid search cross validation by creating hyperparameters to test in grid search. The model has an average precision of 0.80, recall of 0.79, and f1-score of 0.79. These scores suggest that the model is making relatively accurate predictions and has a good balance between precision and recall.

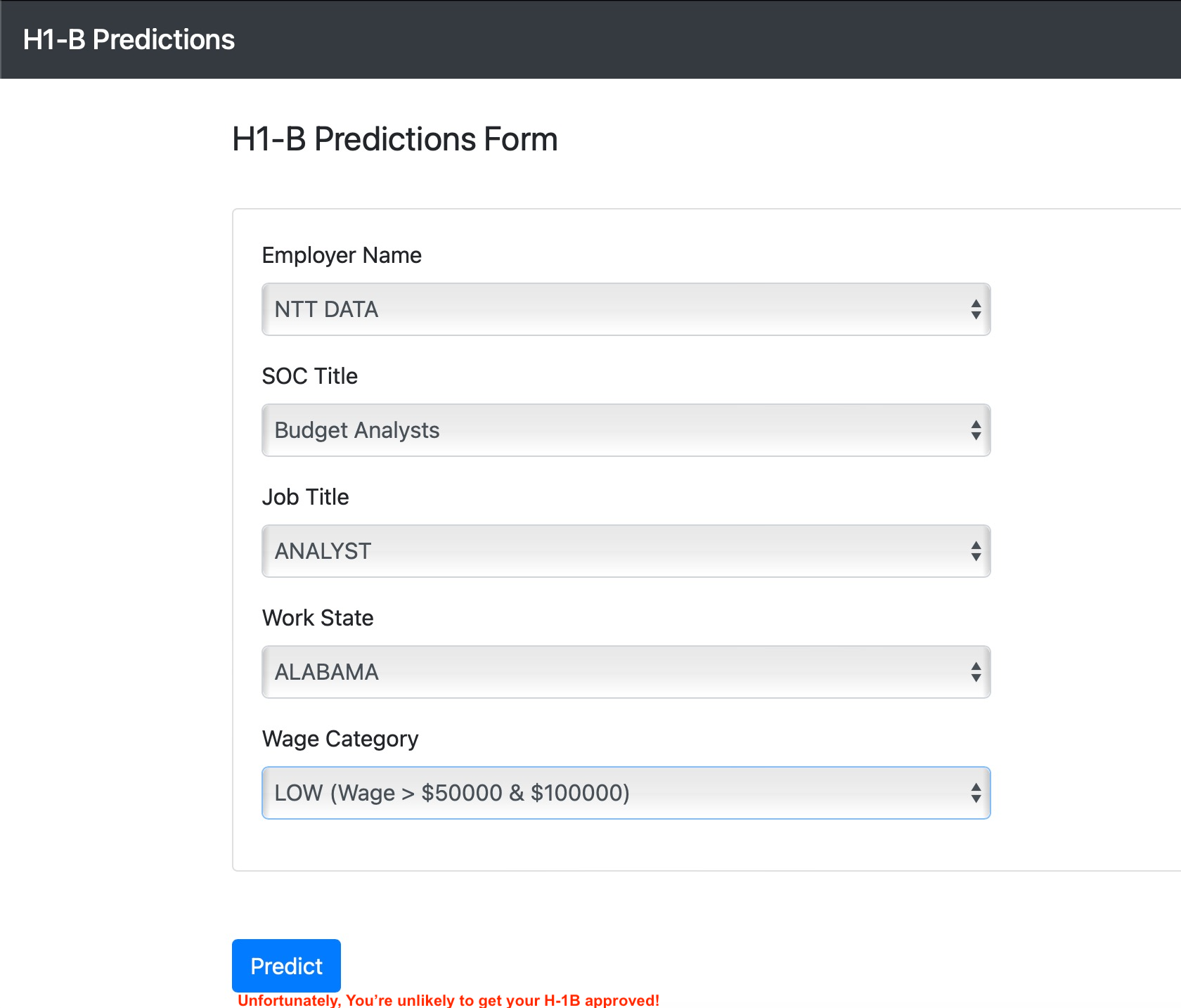
***1.4 Discussion:***

H-1B visa is a lottery system, but our analysis showed various factors play an important role in the prediction of whether an individual getting an H-1B visa or not.

To prove our finding we created a user interface that would predict if an individual will get H-1b or not based on the various factors in Python using Flask.







Above is a Demonstration of the H1-B prediction UI created using Flask.

Our analysis showed that Employer, SOC title, job title, work state, and wage category played an important role in predicting if an individual would get an H-1B certification or denial.

Our business problem was to predict if there were external factors that could determine if an individual can get an H-1B certified or denied, we were able to predict that there were several factors that affected in certification or denial of an H-1 visa.

We can see the demonstration of prediction for an individual to get their H-1B approved or not in the example above.

Software Engineer at Cognizant based in California with High-Income wage is likely to get their H-1B approved.

Analyst at NTT Data based in Alabama with Low-income wage is unlikely to get their H-1B visa approved.

***1.5 Conclusion:***

H-1B visa certification and denial can be affected by factors such as Employer, soc title, job title, work state, and wage category even though it is a lottery-based system.

**Limitation:**

* The data could have more additional information like the administration to understand if a change in government administration affected the H-1B visa policies and allocation.
* Most H-1B visas are filed by international non-immigrant students i.e F-1 visa holders, so the academic background could be helpful to understand the relationship between F-1 visa holders filing for H-1B visas.
* The academic and professional background of the applicant would be helpful to understand the employer's preference in terms of academic background.

**Future:**

* We would like to create a prediction probability of a person getting an H-1B on our User Interface using Flask.
* Merge F-1 Visa Data with H-1B to find a correlation between the two visa types and the probability of an F-1 visa holder getting an H-1B certification.
* ​​Applying the model to new domains, like deep learning.
* Developing new methods for interpreting and visualizing the model's predictions, to facilitate its use stakeholders.

***1.6 Appendix:***

All the team members has given equal contribution and involvement in every step of project. Each team member was responsible for completing specific tasks and collaborating on others to ensure the success of the project. We are grateful for the hard work and dedication of each team member, and we would like to acknowledge their contributions.

**Specific tasks led by each team member :**

* Data Wrangling, Feature Engineering : Priya Rao
* Exploratory Data Analysis and Visualization : Namrata Patel, Stuti Garg
* Data Mining : Pratik Khedekar
* Flask Tutorial : Priya Rao, Stuti Grag
* Report : Namrata Patel
* Presentation : Pratik, Namrata, Priya , and Stuti