

## Data Collection and Preprocessing Phase

Date	10 July 2024
Team ID	740006
Project Title	Predictive Modelling for H1b Visa Approval Using Machine Learning
Maximum Marks	6 Marks

### Data Exploration and Preprocessing Report

Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modelling, and forming a strong foundation for insights and predictions.

Section	Description
Data Overview	<u>Dimension:</u> 923 rows × 49 columns
	<u>Descriptive statistics:</u>

```
df = df[df['PREVAILING_WAGE'] <= 500000]
by_emp_year = df[['EMPLOYER_NAME', 'YEAR', 'PREVAILING_WAGE']] [df['EMPLOYER_NAME'].isin(top_emp)]
# Group by the columns and reset the index to bring the grouping columns back as regular columns.
by_emp_year = by_emp_year.groupby(['EMPLOYER_NAME', 'YEAR']).mean().reset_index()
print(by_emp_year['EMPLOYER_NAME'])
```

## Data Preprocessing Code Screenshots

### Loading Data

```
df = pd.read_csv("h1b_kaggle.csv")
df.shape
df.head()
```

Unnamed: 0	CASE_STATUS	EMPLOYER_NAME	SOC_NAME	JOB_TITLE	FULL_TIME_POSITION	PREVAILING_WAGE	YEAR	WORKSITE
0	1	CERTIFIED-WITHDRAWN	UNIVERSITY OF MICHIGAN	BIOCHEMISTS AND BIOPHYSICISTS	POSTDOCTORAL RESEARCH FELLOW	N	36067.0	2016.0
1	2	CERTIFIED-WITHDRAWN	GOODMAN NETWORKS, INC.	CHIEF EXECUTIVES	CHIEF OPERATING OFFICER	Y	242674.0	2016.0
2	3	CERTIFIED-WITHDRAWN	PORTS AMERICA GROUP, INC.	CHIEF EXECUTIVES	CHIEF PROCESS OFFICER	Y	193066.0	2016.0
3	4	CERTIFIED-WITHDRAWN	GATES CORPORATION, A WHOLLY-OWNED SUBSIDIARY O...	CHIEF EXECUTIVES	REGIONAL PRESIDEN, AMERICAS	Y	220314.0	2016.0
4	5	WITHDRAWN	PEABODY INVESTMENTS CORP.	CHIEF EXECUTIVES	PRESIDENT MONGOLIA AND INDIA	Y	157518.4	2016.0

### Handling Missing Data

```
df.isnull().sum()
```

```
Unnamed: 0      0
CASE_STATUS      0
EMPLOYER_NAME    42
SOC_NAME      17698
JOB_TITLE        26
FULL_TIME_POSITION  0
PREVAILING_WAGE  0
YEAR            0
WORKSITE        0
lon      107089
lat      107089
dtype: int64
```

```
df['SOC_NAME'] = df['SOC_NAME'].fillna(df['SOC_NAME'].mode()[0])
```

```
df['CASE_STATUS'] = df['CASE_STATUS'].map({'CERTIFIED':0, 'CERTIFIED-WITHDRAWN': 1, 'DENIED': 2, 'WITHDRAWN': 3, 'PENDING QUALITY AND COMPLIANCE REVIEW': 4, 'REJECTED': 5, 'INVALIDATED': 6})
```

### Data Transformation

```
df['FULL_TIME_POSITION'] = df['FULL_TIME_POSITION'].map({'N': 0, 'Y': 1})
df.head()
```

Unnamed: 0	CASE_STATUS	EMPLOYER_NAME	SOC_NAME	JOB_TITLE	FULL_TIME_POSITION	PREVAILING_WAGE	YEAR	WORKSITE
0	1	1.0	UNIVERSITY OF MICHIGAN	BIOCHEMISTS AND BIOPHYSICISTS	POSTDOCTORAL RESEARCH FELLOW	0	36067.0	2016.0
1	2	1.0	GOODMAN NETWORKS, INC.	CHIEF EXECUTIVES	CHIEF OPERATING OFFICER	1	242674.0	2016.0
2	3	1.0	PORTS AMERICA GROUP, INC.	CHIEF EXECUTIVES	CHIEF PROCESS OFFICER	1	193066.0	2016.0
3	4	1.0	GATES CORPORATION, A WHOLLY-OWNED SUBSIDIARY O...	CHIEF EXECUTIVES	REGIONAL PRESIDEN, AMERICAS	1	220314.0	2016.0
4	5	3.0	PEABODY INVESTMENTS CORP.	CHIEF EXECUTIVES	PRESIDENT MONGOLIA AND INDIA	1	157518.4	2016.0

Feature Engineering	-
Save Processed Data	-

  

Bivariate Analysis	-
--------------------	---