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

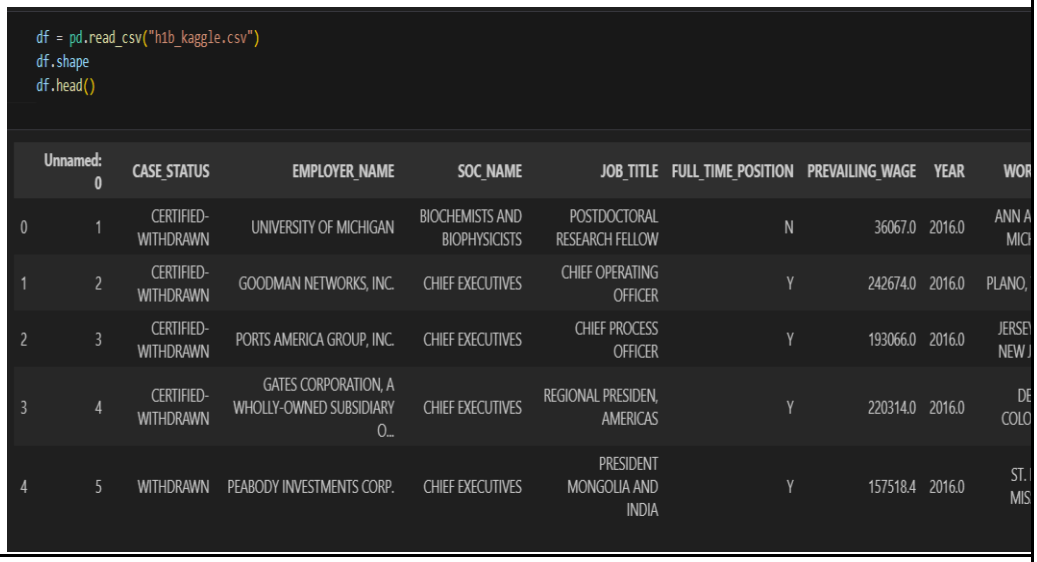
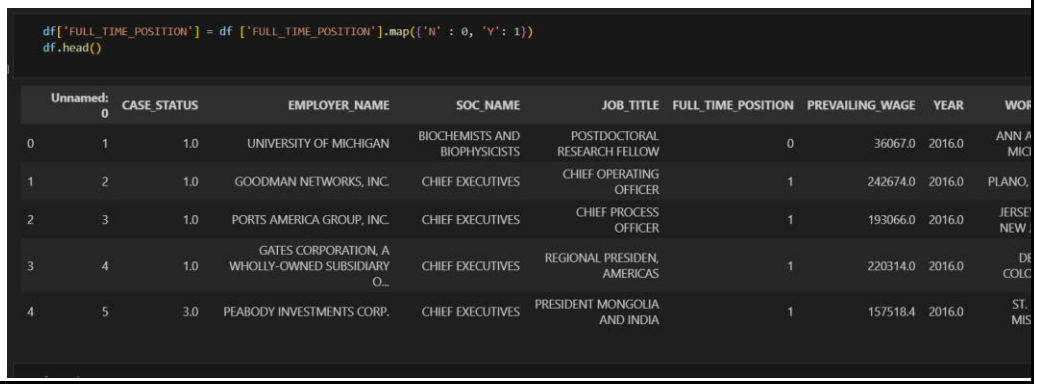
Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modelling, and forming a strong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

Section	Description
Data Overview	Dimension: 923 rows × 49 columns Descriptive statistics:

	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0	1	CERTIFIED-WITHDRAWN	UNIVERSITY OF MICHIGAN	BIOCHEMISTS AND BIOPHYSICISTS	POSTDOCTORAL RESEARCH FELLOW	N	36067.0	2016.0	ANN ARBOR, MICHIGAN
1	2	CERTIFIED-WITHDRAWN	GOODMAN NETWORKS, INC.	CHIEF EXECUTIVES	CHIEF OPERATING OFFICER	Y	242674.0	2016.0	PLANO, TEXAS
2	3	CERTIFIED-WITHDRAWN	PORTS AMERICA GROUP, INC.	CHIEF EXECUTIVES	CHIEF PROCESS OFFICER	Y	193066.0	2016.0	JERSEY CITY, NEW JERSEY
3	4	CERTIFIED-WITHDRAWN	GATES CORPORATION, A WHOLLY-OWNED SUBSIDIARY O...	CHIEF EXECUTIVES	REGIONAL PRESIDEN, AMERICAS	Y	220314.0	2016.0	DENVER, COLORADO
4	5	WITHDRAWN	PEABODY INVESTMENTS CORP.	CHIEF EXECUTIVES	PRESIDENT MONGOLIA AND INDIA	Y	157518.4	2016.0	ST. LOUIS, MISSOURI

Outliers and Anomalies	

Data Preprocessing Code Screenshots

Loading Data	 Unnamed: 0 | CASE_STATUS | EMPLOYER_NAME | SOC_NAME | JOB_TITLE | FULL_TIME_POSITION | PREVAILING_WAGE | YEAR | WOR 0 | 1 | CERTIFIED-WITHDRAWN | UNIVERSITY OF MICHIGAN | BIOCHEMISTS AND BIOPHYSICISTS | POSTDOCTORAL RESEARCH FELLOW | N | 36067.0 | 2016.0 | ANN A MICI 1 | 2 | CERTIFIED-WITHDRAWN | GOODMAN NETWORKS, INC. | CHIEF EXECUTIVES | CHIEF OPERATING OFFICER | Y | 242674.0 | 2016.0 | PLANO, 2 | 3 | CERTIFIED-WITHDRAWN | PORTS AMERICA GROUP, INC. | CHIEF EXECUTIVES | CHIEF PROCESS OFFICER | Y | 193066.0 | 2016.0 | JERSE NEW J 3 | 4 | CERTIFIED-WITHDRAWN | GATES CORPORATION, A WHOLLY-OWNED SUBSIDIARY O... | CHIEF EXECUTIVES | REGIONAL PRESIDEN, AMERICAS | Y | 220314.0 | 2016.0 | DE COLO 4 | 5 | WITHDRAWN | PEABODY INVESTMENTS CORP. | CHIEF EXECUTIVES | PRESIDENT MONGOLIA AND INDIA | Y | 157518.4 | 2016.0 | ST. MIS
Handling Missing Data	
Data Transformation	 Unnamed: 0 | CASE_STATUS | EMPLOYER_NAME | SOC_NAME | JOB_TITLE | FULL_TIME_POSITION | PREVAILING_WAGE | YEAR | WOR 0 | 1 | 1.0 | UNIVERSITY OF MICHIGAN | BIOCHEMISTS AND BIOPHYSICISTS | POSTDOCTORAL RESEARCH FELLOW | 0 | 36067.0 | 2016.0 | ANN A MICI 1 | 2 | 1.0 | GOODMAN NETWORKS, INC. | CHIEF EXECUTIVES | CHIEF OPERATING OFFICER | 1 | 242674.0 | 2016.0 | PLANO, 2 | 3 | 1.0 | PORTS AMERICA GROUP, INC. | CHIEF EXECUTIVES | CHIEF PROCESS OFFICER | 1 | 193066.0 | 2016.0 | JERSE NEW J 3 | 4 | 1.0 | GATES CORPORATION, A WHOLLY-OWNED SUBSIDIARY O... | CHIEF EXECUTIVES | REGIONAL PRESIDEN, AMERICAS | 1 | 220314.0 | 2016.0 | DE COLO 4 | 5 | 3.0 | PEABODY INVESTMENTS CORP. | CHIEF EXECUTIVES | PRESIDENT MONGOLIA AND INDIA | 1 | 157518.4 | 2016.0 | ST. MIS

Feature Engineering	-
Save Processed Data	-

Bivariate Analysis	-