


Data Collection and Preprocessing Phase

Date	31 June 2024
Team ID	739853
Project Title	Software Salary Prediction
Maximum Marks	6 Marks

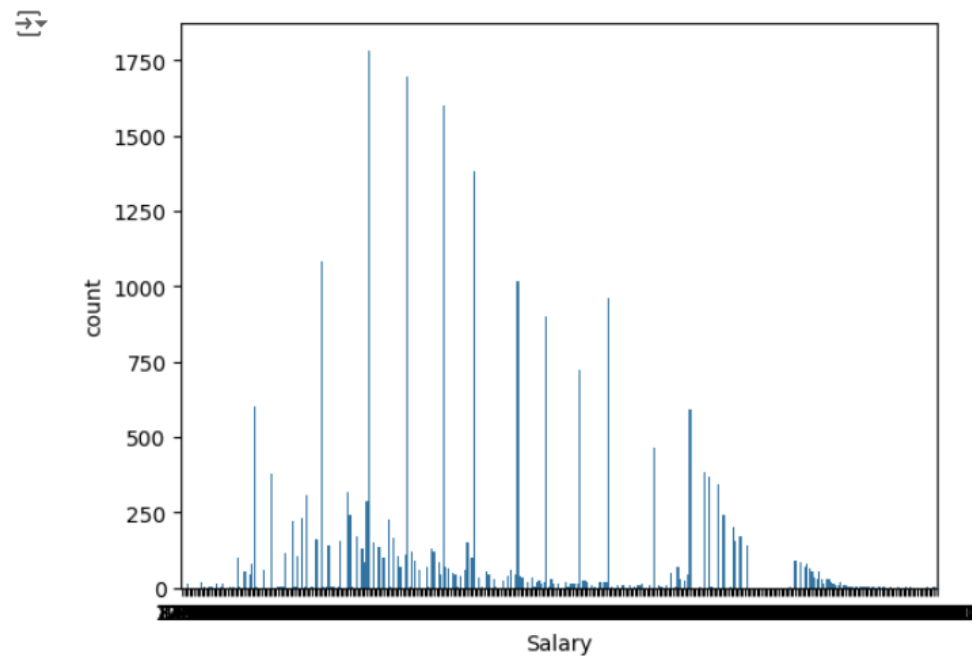
Data Exploration and Preprocessing

Dataset gets statistically analysed to identify patterns and outliers. Data preprocessing addresses missing values, improving data quality for further analysis and modelling, and forming a strong foundation for insights and predictions.

Section	Description																																																																																	
Data Overview	<p><u>Dimensions:</u> 64rows*8 Columns</p> <p><u>Descriptive Statistics:</u></p> <pre>[] df.describe()</pre> <div><div></div><table><tr><th></th><th>Rating</th><th>Company Name</th><th>Job Title</th><th>Salary</th><th>Salaries Reported</th><th>Location</th><th>Employment Status</th><th>Job Roles</th></tr><tr><td>count</td><td>22770.000000</td><td>22770.000000</td><td>22770.000000</td><td>2.277000e+04</td><td>22770.000000</td><td>22770.000000</td><td>22770.000000</td><td>22770.000000</td></tr><tr><td>mean</td><td>3.918213</td><td>5478.825209</td><td>597.435968</td><td>6.953872e+05</td><td>1.855775</td><td>3.150812</td><td>1.071322</td><td>5.465086</td></tr><tr><td>std</td><td>0.519675</td><td>3224.603280</td><td>348.305504</td><td>8.843990e+05</td><td>6.823668</td><td>3.529116</td><td>0.342450</td><td>3.221968</td></tr><tr><td>min</td><td>1.000000</td><td>0.000000</td><td>0.000000</td><td>2.112000e+03</td><td>1.000000</td><td>0.000000</td><td>0.000000</td><td>0.000000</td></tr><tr><td>25%</td><td>3.700000</td><td>2756.000000</td><td>237.000000</td><td>3.000000e+05</td><td>1.000000</td><td>0.000000</td><td>1.000000</td><td>3.000000</td></tr><tr><td>50%</td><td>3.900000</td><td>5317.500000</td><td>753.000000</td><td>5.000000e+05</td><td>1.000000</td><td>2.000000</td><td>1.000000</td><td>7.000000</td></tr><tr><td>75%</td><td>4.200000</td><td>8336.000000</td><td>850.000000</td><td>9.000000e+05</td><td>1.000000</td><td>8.000000</td><td>1.000000</td><td>8.000000</td></tr><tr><td>max</td><td>5.000000</td><td>11260.000000</td><td>1079.000000</td><td>9.000000e+07</td><td>361.000000</td><td>9.000000</td><td>3.000000</td><td>10.000000</td></tr></table></div>		Rating	Company Name	Job Title	Salary	Salaries Reported	Location	Employment Status	Job Roles	count	22770.000000	22770.000000	22770.000000	2.277000e+04	22770.000000	22770.000000	22770.000000	22770.000000	mean	3.918213	5478.825209	597.435968	6.953872e+05	1.855775	3.150812	1.071322	5.465086	std	0.519675	3224.603280	348.305504	8.843990e+05	6.823668	3.529116	0.342450	3.221968	min	1.000000	0.000000	0.000000	2.112000e+03	1.000000	0.000000	0.000000	0.000000	25%	3.700000	2756.000000	237.000000	3.000000e+05	1.000000	0.000000	1.000000	3.000000	50%	3.900000	5317.500000	753.000000	5.000000e+05	1.000000	2.000000	1.000000	7.000000	75%	4.200000	8336.000000	850.000000	9.000000e+05	1.000000	8.000000	1.000000	8.000000	max	5.000000	11260.000000	1079.000000	9.000000e+07	361.000000	9.000000	3.000000	10.000000
		Rating	Company Name	Job Title	Salary	Salaries Reported	Location	Employment Status	Job Roles																																																																									
	count	22770.000000	22770.000000	22770.000000	2.277000e+04	22770.000000	22770.000000	22770.000000	22770.000000																																																																									
	mean	3.918213	5478.825209	597.435968	6.953872e+05	1.855775	3.150812	1.071322	5.465086																																																																									
	std	0.519675	3224.603280	348.305504	8.843990e+05	6.823668	3.529116	0.342450	3.221968																																																																									
	min	1.000000	0.000000	0.000000	2.112000e+03	1.000000	0.000000	0.000000	0.000000																																																																									
	25%	3.700000	2756.000000	237.000000	3.000000e+05	1.000000	0.000000	1.000000	3.000000																																																																									
	50%	3.900000	5317.500000	753.000000	5.000000e+05	1.000000	2.000000	1.000000	7.000000																																																																									
	75%	4.200000	8336.000000	850.000000	9.000000e+05	1.000000	8.000000	1.000000	8.000000																																																																									
	max	5.000000	11260.000000	1079.000000	9.000000e+07	361.000000	9.000000	3.000000	10.000000																																																																									

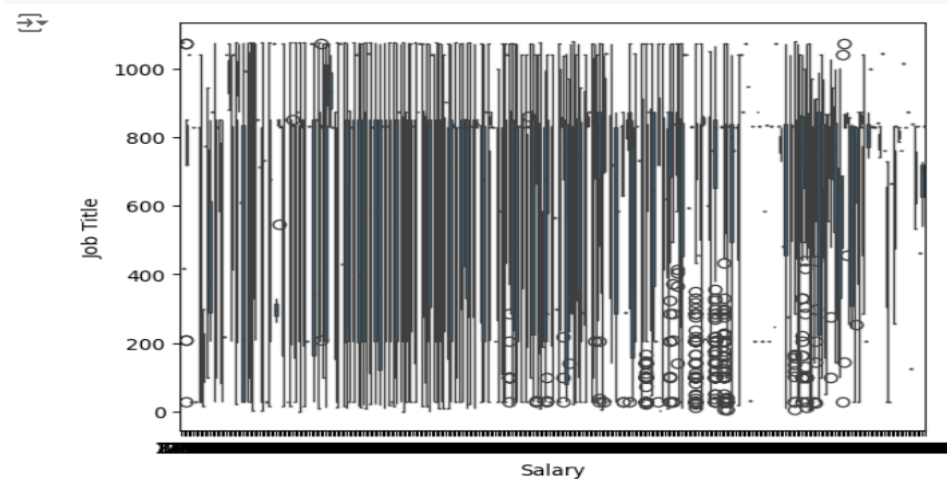
Univariate Analysis

```
[ ] sns.countplot(x='Salary',data=df)
plt.show()
```

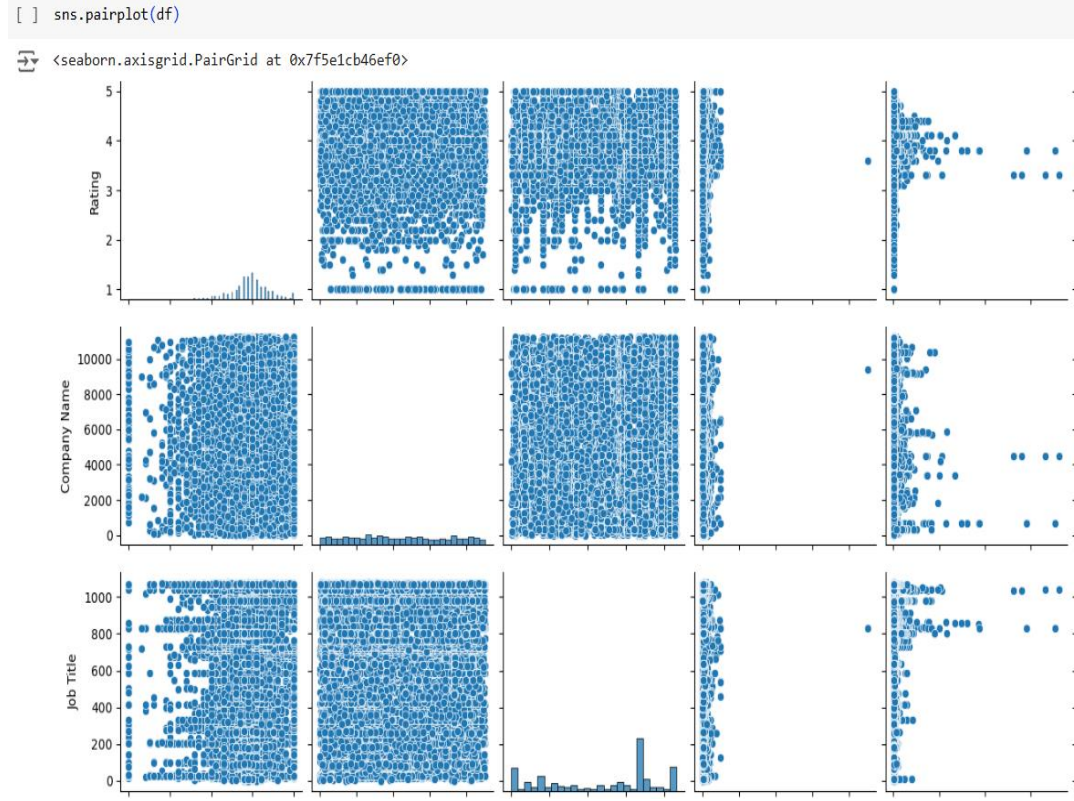


Bivariate Analysis

```
[ ] sns.boxplot(x='Salary',y='Job Title', data=df)
plt.show()
```



Multivariate Analysis



Outliers and Anomalies

Data Preprocessing Code Screenshots

Loading Data

```
df = pd.read_csv('Salary_Dataset_with_Extra_Features.csv')
```

```
df.head()
```

	Rating	Company Name	Job Title	Salary	Salaries Reported	Location	Employment Status	Job Roles
0	3.8	Sasken	Android Developer	400000	3	Bangalore	Full Time	Android
1	4.5	Advanced Millennium Technologies	Android Developer	400000	3	Bangalore	Full Time	Android
2	4.0	Unacademy	Android Developer	1000000	3	Bangalore	Full Time	Android
3	3.8	SnapBizz Cloudtech	Android Developer	300000	3	Bangalore	Full Time	Android
4	4.4	Appoids Tech Solutions	Android Developer	600000	3	Bangalore	Full Time	Android

Handling Missing Values	<pre>[] df.info()</pre> <pre><class 'pandas.core.frame.DataFrame'> RangeIndex: 22770 entries, 0 to 22769 Data columns (total 8 columns): # Column Non-Null Count Dtype --- --- 0 Rating 22770 non-null float64 1 Company Name 22769 non-null object 2 Job Title 22770 non-null object 3 Salary 22770 non-null int64 4 Salaries Reported 22770 non-null int64 5 Location 22770 non-null object 6 Employment Status 22770 non-null object 7 Job Roles 22770 non-null object dtypes: float64(1), int64(2), object(5) memory usage: 1.4+ MB</pre>
-------------------------	---

Data Transformation	--
Feature Engineering	--
Save Processed Data	--