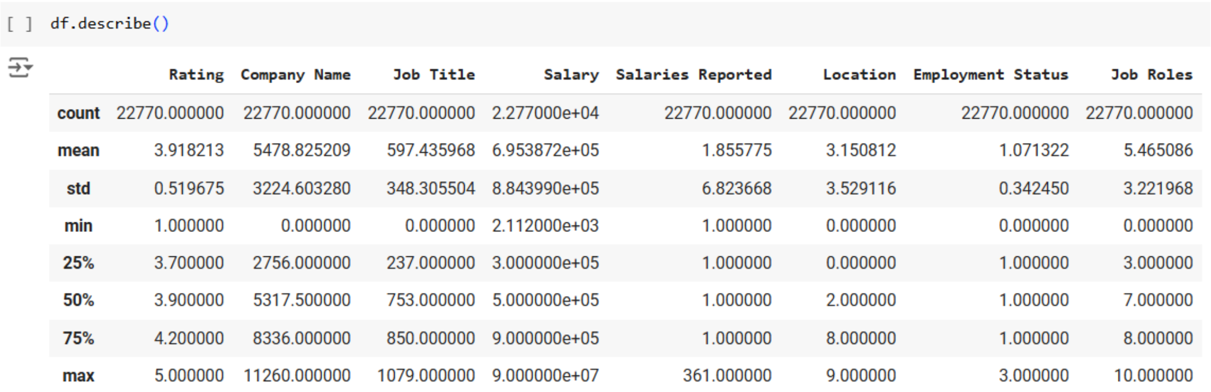
# Data Collection and Preprocessing Phase

| | |
|---|---|
| Date | 31 June 2024 |
| Team ID | 739854 |
| Project Title | Software Salary Prediction |
| Maximum Marks | 6 Marks |

## Data Exploration and Preprocessing

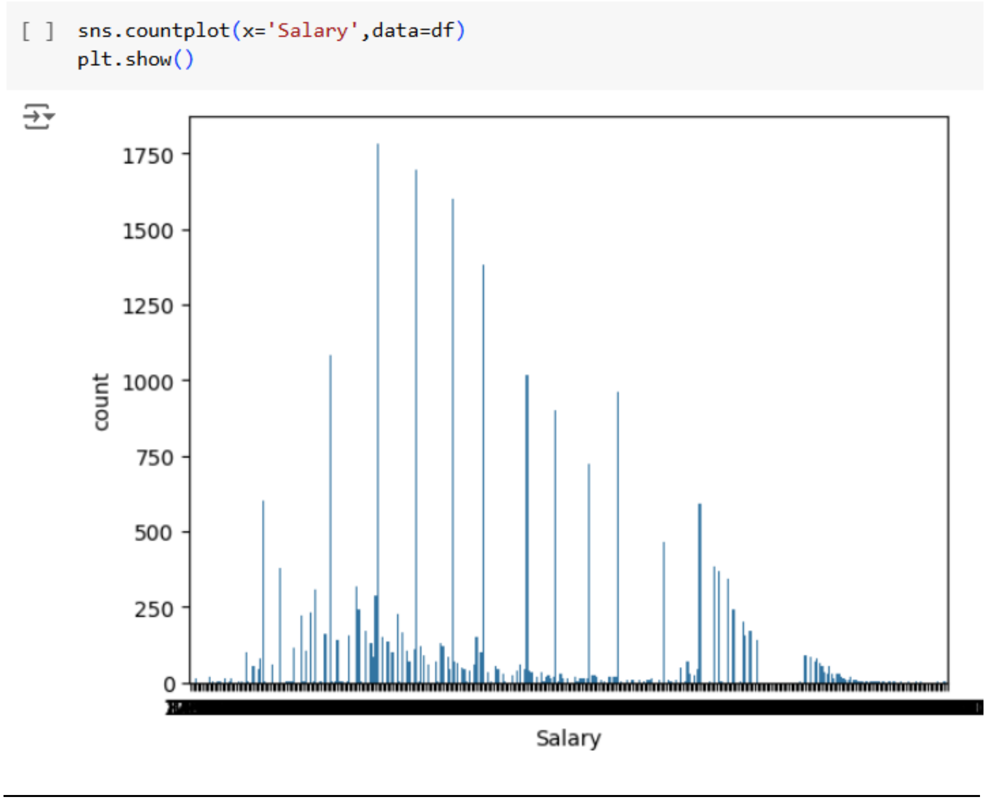
Dataset gets statistically analysed to identify patterns and outliers. Data preprocessing addresses missing values, improving data quality for further analysis and modelling, and forming a strong foundation for insights and predictions.

| **Section** | **Description** |
|---|---|
| Data Overview | Dimensions: 64rows*8 Columns<br>Descriptive Statistics:<br> |

```
[ ]  df.describe()
```

```
             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
```

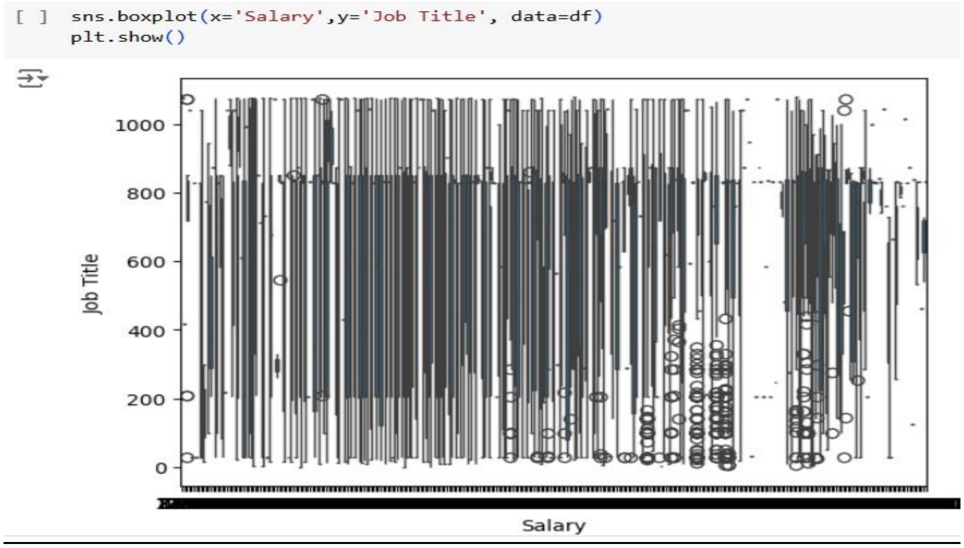
Univariate Analysis



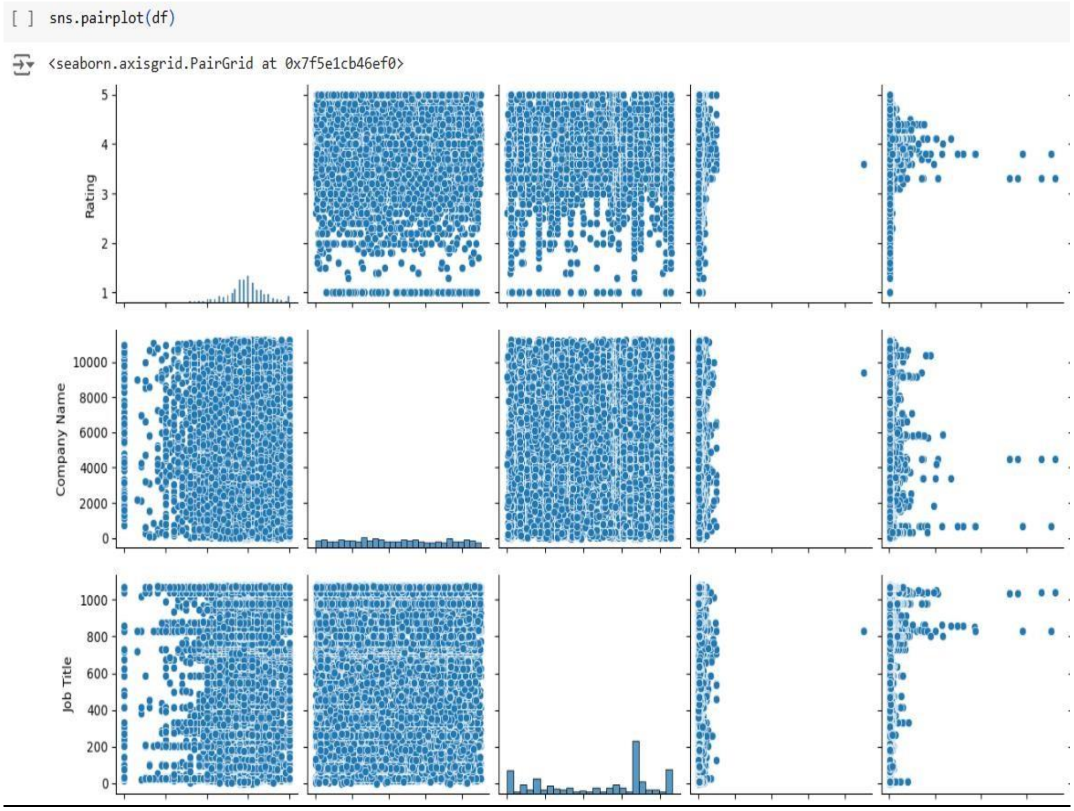
```
[ ]  sns.countplot(x='Salary',data=df)
     plt.show()
```

Bivariate

Analysis

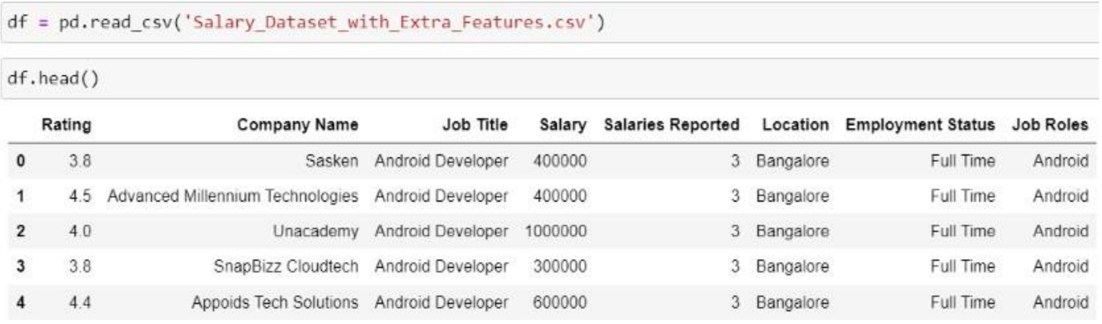


```
[ ]  sns.boxplot(x='Salary',y='Job Title', data=df)
     plt.show()
```

<table>
<tr>
<td>Multivariate Analysis</td>
<td>
<pre>[ ] sns.pairplot(df)

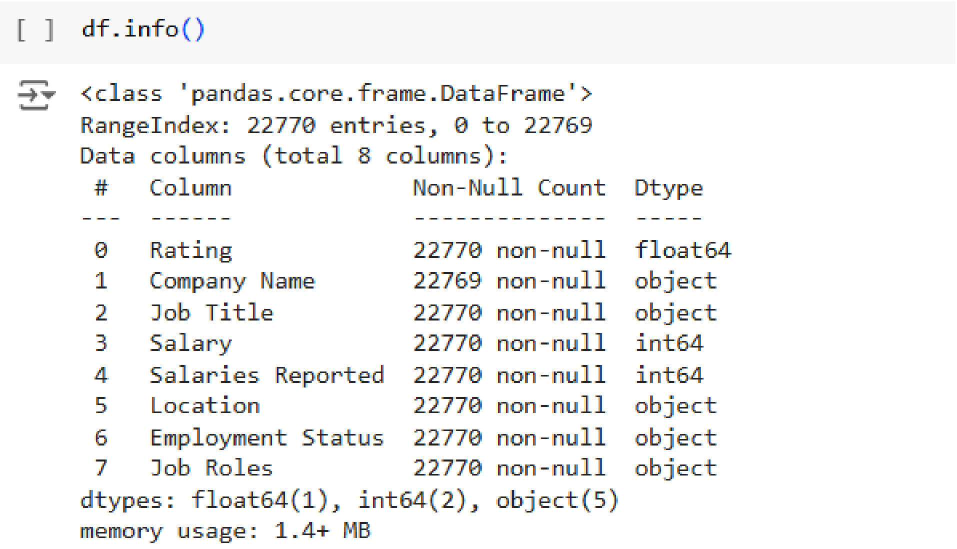
<seaborn.axisgrid.PairGrid at 0x7f5e1cb46ef0></pre>
</td>
</tr>
<tr>
<td>Outliers and Anomalies</td>
<td>--------------</td>
</tr>
</table>

## Data Preprocessing Code Screenshots

<table>
<tr>
<td>Loading Data</td>
<td>
<pre>df = pd.read_csv('Salary_Dataset_with_Extra_Features.csv')

df.head()</pre>
<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>0</td><td>3.8</td><td>Sasken</td><td>Android Developer</td><td>400000</td><td>3</td><td>Bangalore</td><td>Full Time</td><td>Android</td></tr>
<tr><td>1</td><td>4.5</td><td>Advanced Millennium Technologies</td><td>Android Developer</td><td>400000</td><td>3</td><td>Bangalore</td><td>Full Time</td><td>Android</td></tr>
<tr><td>2</td><td>4.0</td><td>Unacademy</td><td>Android Developer</td><td>1000000</td><td>3</td><td>Bangalore</td><td>Full Time</td><td>Android</td></tr>
<tr><td>3</td><td>3.8</td><td>SnapBizz Cloudtech</td><td>Android Developer</td><td>300000</td><td>3</td><td>Bangalore</td><td>Full Time</td><td>Android</td></tr>
<tr><td>4</td><td>4.4</td><td>Appoids Tech Solutions</td><td>Android Developer</td><td>600000</td><td>3</td><td>Bangalore</td><td>Full Time</td><td>Android</td></tr>
</table></td>
</tr>
</table>

<table>
<tr>
<td>Handling Missing Values</td>
<td>



```
df.info()
```

```
<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
```

</td>
</tr>
</table>

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