# Data Collection and Preprocessing Phase

| | |
|---|---|
| Date | 22 June 2024 |
| Team ID | 739904 |
| Project Title | Income Activities Using Machine Learning |
| Maximum Marks | 6 Marks |

## Data Exploration and Preprocessing Report

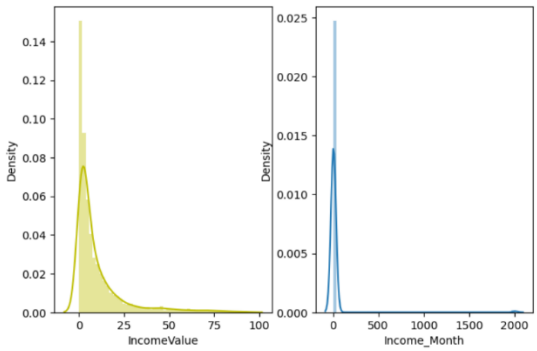
Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong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

| Section | Description |
|---|---|
| Data Overview | Descriptive statistics:<br>In [7]: `df.describe(include='all')`<br>Out[7]: |
| Univariate Analysis | |

| | ADM0_NAME | ADM1_NAME | ADM2_NAME | Income_Category | Income_Month | Income_Year | IncomeValue | Income_DataSource |
|---|---|---|---|---|---|---|---|---|
| count | 4370 | 4022 | 444 | 4349 | 4370.000000 | 4370.000000 | 4370 | 4345 |
| unique | 32 | 363 | 68 | 27 | NaN | NaN | 3744 | 32 |
| top | Senegal | $South/Amajyepfo | Lanao Del Norte | Labor - Salary/Regular | NaN | NaN | Baseline | WFP VAM Analyse Globale de la Vulnerabilite, d... |
| freq | 256 | 54 | 12 | 370 | NaN | NaN | 25 | 256 |
| mean | NaN | NaN | NaN | NaN | 18.535240 | 1998.965904 | NaN | NaN |
| std | NaN | NaN | NaN | NaN | 151.007077 | 150.455385 | NaN | NaN |
| min | NaN | NaN | NaN | NaN | 1.000000 | 0.350877 | NaN | NaN |
| 25% | NaN | NaN | NaN | NaN | 5.000000 | 2009.000000 | NaN | NaN |
| 50% | NaN | NaN | NaN | NaN | 7.000000 | 2010.000000 | NaN | NaN |
| 75% | NaN | NaN | NaN | NaN | 9.000000 | 2012.000000 | NaN | NaN |
| max | NaN | NaN | NaN | NaN | 2009.000000 | 2014.000000 | NaN | NaN |

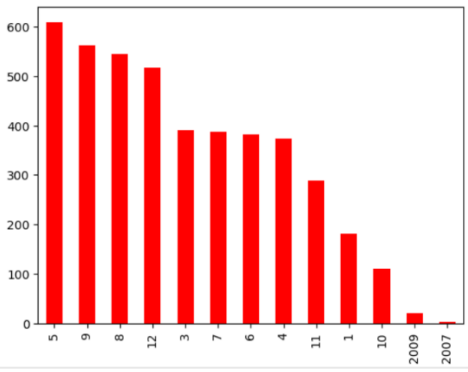
<table>
<tr>
<td></td>
<td>

```
In [21]: plt.figure(figsize=(12,5))
plt.subplot(131)
sns.distplot(df["IncomeValue"],color="y")
plt.subplot(132)
sns.distplot(df["Income_Month"])
plt.show()
```

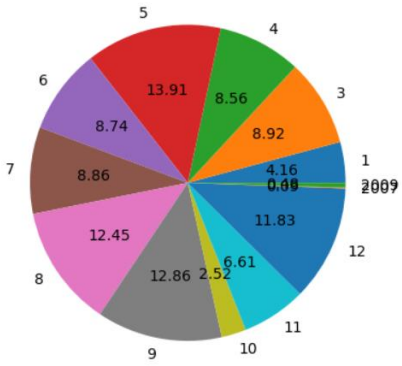



\
</td>
</tr>
<tr>
<td>Bivariate Analysis</td>
<td>

```
In [18]: df['Income_Month'].value_counts().plot(kind='bar',color='Red')
Out[18]: <Axes: xlabel='Income_Month'>
```



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

| | |
|---|---|
| | In [19]: `df.groupby('Income_Month').size().plot(kind='pie', autopct='%.2f')`<br>Out[19]: <Axes: ><br>  |
| Multivariate Analysis | In [30]:<br>`fig, ax = plt.subplots(figsize=(10, 8))`<br>`sns.heatmap(df.corr(),annot=True, fmt=".2f", ax=ax)`<br>Out[30]: <Axes: ><br>  |
| **Data Preprocessing Code Screenshots** | |

| | |
|---|---|
| Loading Data | In [7]: `df.describe(include='all')`<br>Out[7]: |

| | ADM0_NAME | ADM1_NAME | ADM2_NAME | Income_Category | Income_Month | Income_Year | IncomeValue | Income_DataSource |
|---|---|---|---|---|---|---|---|---|
| count | 4370 | 4022 | 444 | 4349 | 4370.000000 | 4370.000000 | 4370 | 4345 |
| unique | 32 | 363 | 68 | 27 | NaN | NaN | 3744 | 32 |
| top | Senegal | $South/Amajyepfo | Lanao Del Norte | Labor - Salary/Regular | NaN | NaN | Baseline | WFP VAM Analyse Globale de la Vulnerabilite, d... |
| freq | 256 | 54 | 12 | 370 | NaN | NaN | 25 | 256 |
| mean | NaN | NaN | NaN | NaN | 18.535240 | 1998.965904 | NaN | NaN |
| std | NaN | NaN | NaN | NaN | 151.007077 | 150.455385 | NaN | NaN |
| min | NaN | NaN | NaN | NaN | 1.000000 | 0.350877 | NaN | NaN |
| 25% | NaN | NaN | NaN | NaN | 5.000000 | 2009.000000 | NaN | NaN |
| 50% | NaN | NaN | NaN | NaN | 7.000000 | 2010.000000 | NaN | NaN |
| 75% | NaN | NaN | NaN | NaN | 9.000000 | 2012.000000 | NaN | NaN |
| max | NaN | NaN | NaN | NaN | 2009.000000 | 2014.000000 | NaN | NaN |

| | |
|---|---|
| Handling Missing Data | No missing values |
| Feature Engineering | Attached the codes in final submission. |
| Save Processed Data | - |