

Data Collection and Preprocessing Phase

Date	22 June 2024
Team ID	739882
Project Title	Income Activities Using Machine Learning
Maximum Marks	6 Marks

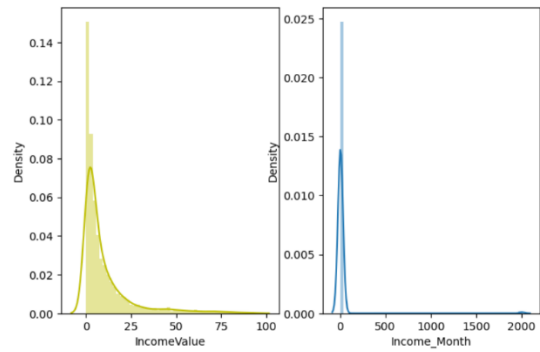
Data Exploration and Preprocessing Report

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	<u>Descriptive statistics:</u>																																																																																																												
	<pre>In [7]: df.describe(include='all')</pre>																																																																																																												
	<pre>Out[7]:</pre>																																																																																																												
	<table><thead><tr><th></th><th>ADMO_NAME</th><th>ADM1_NAME</th><th>ADM2_NAME</th><th>Income_Category</th><th>Income_Month</th><th>Income_Year</th><th>IncomeValue</th><th>Income_DataSource</th></tr></thead><tbody><tr><td>count</td><td>4370</td><td>4022</td><td>444</td><td>4349</td><td>4370.000000</td><td>4370.000000</td><td>4370</td><td>4345</td></tr><tr><td>unique</td><td>32</td><td>363</td><td>68</td><td>27</td><td>NaN</td><td>NaN</td><td>3744</td><td>32</td></tr><tr><td>top</td><td>Senegal</td><td>SSouth/Amajepfo</td><td>Lanao Del Norte</td><td>Labor - Salary/Regular</td><td>NaN</td><td>NaN</td><td>Baseline</td><td>WFP VAM Analyse Globale de la Vulnérabilité, d...</td></tr><tr><td>freq</td><td>256</td><td>54</td><td>12</td><td>370</td><td>NaN</td><td>NaN</td><td>25</td><td>256</td></tr><tr><td>mean</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>18.535240</td><td>1998.965904</td><td>NaN</td><td>NaN</td></tr><tr><td>std</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>151.007077</td><td>150.455385</td><td>NaN</td><td>NaN</td></tr><tr><td>min</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>1.000000</td><td>0.350877</td><td>NaN</td><td>NaN</td></tr><tr><td>25%</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>5.000000</td><td>2009.000000</td><td>NaN</td><td>NaN</td></tr><tr><td>50%</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>7.000000</td><td>2010.000000</td><td>NaN</td><td>NaN</td></tr><tr><td>75%</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>9.000000</td><td>2012.000000</td><td>NaN</td><td>NaN</td></tr><tr><td>max</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>2009.000000</td><td>2014.000000</td><td>NaN</td><td>NaN</td></tr></tbody></table>		ADMO_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	SSouth/Amajepfo	Lanao Del Norte	Labor - Salary/Regular	NaN	NaN	Baseline	WFP VAM Analyse Globale de la Vulnérabilité, 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
		ADMO_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	SSouth/Amajepfo	Lanao Del Norte	Labor - Salary/Regular	NaN	NaN	Baseline	WFP VAM Analyse Globale de la Vulnérabilité, 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																																																																																																					
Univariate Analysis																																																																																																													

```

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()
  
```

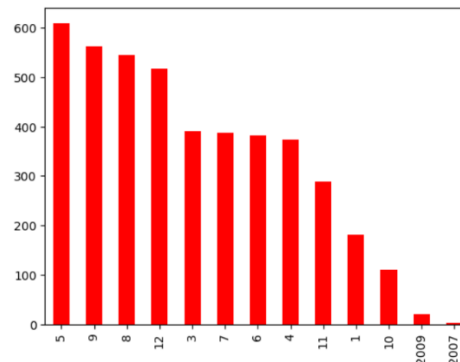


\

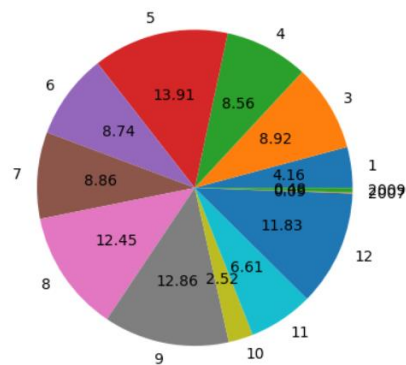
Bivariate Analysis

```

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

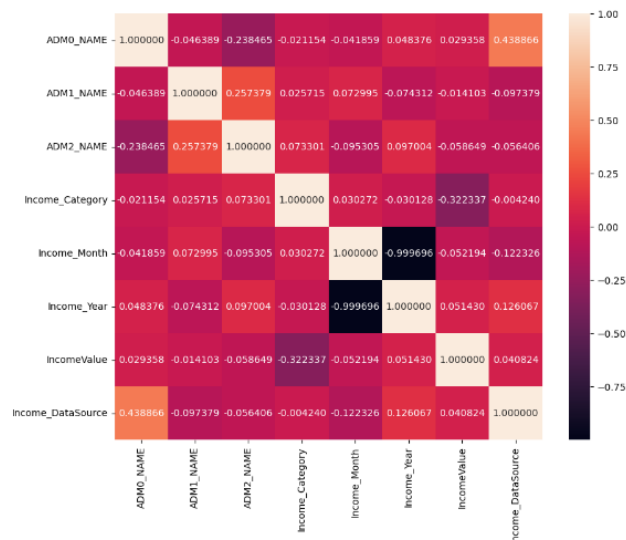


```
In [19]: df.groupby('Income_Month').size().plot(kind='pie', autopct='%2f')
Out[19]: <Axes: >
```



Multivariate Analysis

```
In [38]: fig, ax = plt.subplots(figsize=(10, 8))
sns.heatmap(df.corr(), annot=True, fmt='2f', ax=ax)
Out[38]: <Axes: >
```



Data Preprocessing Code Screenshots

Loading Data	<div>In [7]: df.describe(include='all')</div> <div>Out[7]:</div> <table><thead><tr><th></th><th>ADM0_NAME</th><th>ADM1_NAME</th><th>ADM2_NAME</th><th>Income_Category</th><th>Income_Month</th><th>Income_Year</th><th>IncomeValue</th><th>Income_DataSource</th></tr></thead><tbody><tr><td>count</td><td>4370</td><td>4022</td><td>444</td><td>4349</td><td>4370.000000</td><td>4370.000000</td><td>4370</td><td>4345</td></tr><tr><td>unique</td><td>32</td><td>363</td><td>68</td><td>27</td><td>NaN</td><td>NaN</td><td>3744</td><td>32</td></tr><tr><td>top</td><td>Senegal</td><td>\$South/Amajepfo</td><td>Lanao Del Norte</td><td>Labor - Salary/Regular</td><td>NaN</td><td>NaN</td><td>Baseline</td><td>WFP VAM Analyse Globale de la Vulnerabilite, d...</td></tr><tr><td>freq</td><td>256</td><td>54</td><td>12</td><td>370</td><td>NaN</td><td>NaN</td><td>25</td><td>256</td></tr><tr><td>mean</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>18.535240</td><td>1998.965904</td><td>NaN</td><td>NaN</td></tr><tr><td>std</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>151.007077</td><td>150.455385</td><td>NaN</td><td>NaN</td></tr><tr><td>min</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>1.000000</td><td>0.359877</td><td>NaN</td><td>NaN</td></tr><tr><td>25%</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>5.000000</td><td>2009.000000</td><td>NaN</td><td>NaN</td></tr><tr><td>50%</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>7.000000</td><td>2010.000000</td><td>NaN</td><td>NaN</td></tr><tr><td>75%</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>9.000000</td><td>2012.000000</td><td>NaN</td><td>NaN</td></tr><tr><td>max</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>2009.000000</td><td>2014.000000</td><td>NaN</td><td>NaN</td></tr></tbody></table>		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/Amajepfo	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	0.359877	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
	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/Amajepfo	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	0.359877	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	-																																																																																																												