

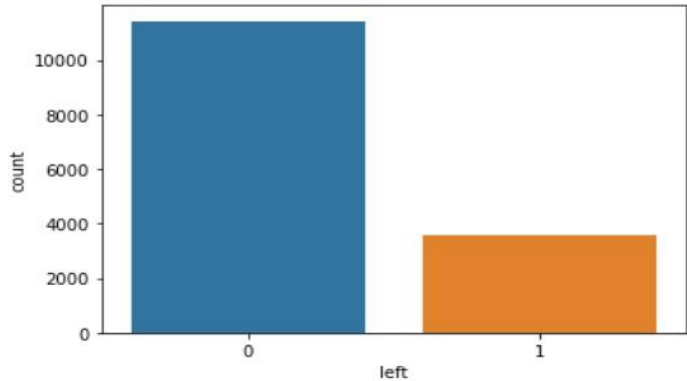
Data Collection and Preprocessing Phase

Date	15 March 2024
Team ID	739687
Project Title	SMS SPAM DETECTION
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	<div>Dimension: 14999 rows × 4 columns</div> <div>Descriptive statistics:</div> <div><pre># top 5 rows of the dataframe df.head()</pre></div> <div><table><thead><tr><th></th><th>label</th><th>text</th><th>label_num</th></tr></thead><tbody><tr><td>0</td><td>ham</td><td>Subject: enron methanol ; meter # : 988291\r\n...</td><td>0</td></tr><tr><td>1</td><td>ham</td><td>Subject: hpl nom for january 9 , 2001\r\n(see...</td><td>0</td></tr><tr><td>2</td><td>ham</td><td>Subject: neon retreat\r\nho ho ho , we ' re ar...</td><td>0</td></tr><tr><td>3</td><td>spam</td><td>Subject: photoshop , windows , office . cheap ...</td><td>1</td></tr><tr><td>4</td><td>ham</td><td>Subject: re : indian springs\r\nthis deal is t...</td><td>0</td></tr></tbody></table></div>		label	text	label_num	0	ham	Subject: enron methanol ; meter # : 988291\r\n...	0	1	ham	Subject: hpl nom for january 9 , 2001\r\n(see...	0	2	ham	Subject: neon retreat\r\nho ho ho , we ' re ar...	0	3	spam	Subject: photoshop , windows , office . cheap ...	1	4	ham	Subject: re : indian springs\r\nthis deal is t...	0
	label	text	label_num																						
0	ham	Subject: enron methanol ; meter # : 988291\r\n...	0																						
1	ham	Subject: hpl nom for january 9 , 2001\r\n(see...	0																						
2	ham	Subject: neon retreat\r\nho ho ho , we ' re ar...	0																						
3	spam	Subject: photoshop , windows , office . cheap ...	1																						
4	ham	Subject: re : indian springs\r\nthis deal is t...	0																						
Univariate Analysis																									
	<div><pre>sns.countplot(df['left'])</pre></div> <div><AxesSubplot:xlabel='left', ylabel='count'></div> <div><table><thead><tr><th>left</th><th>count</th></tr></thead><tbody><tr><td>0</td><td>11000</td></tr><tr><td>1</td><td>3500</td></tr></tbody></table></div>	left	count	0	11000	1	3500																		
left	count																								
0	11000																								
1	3500																								

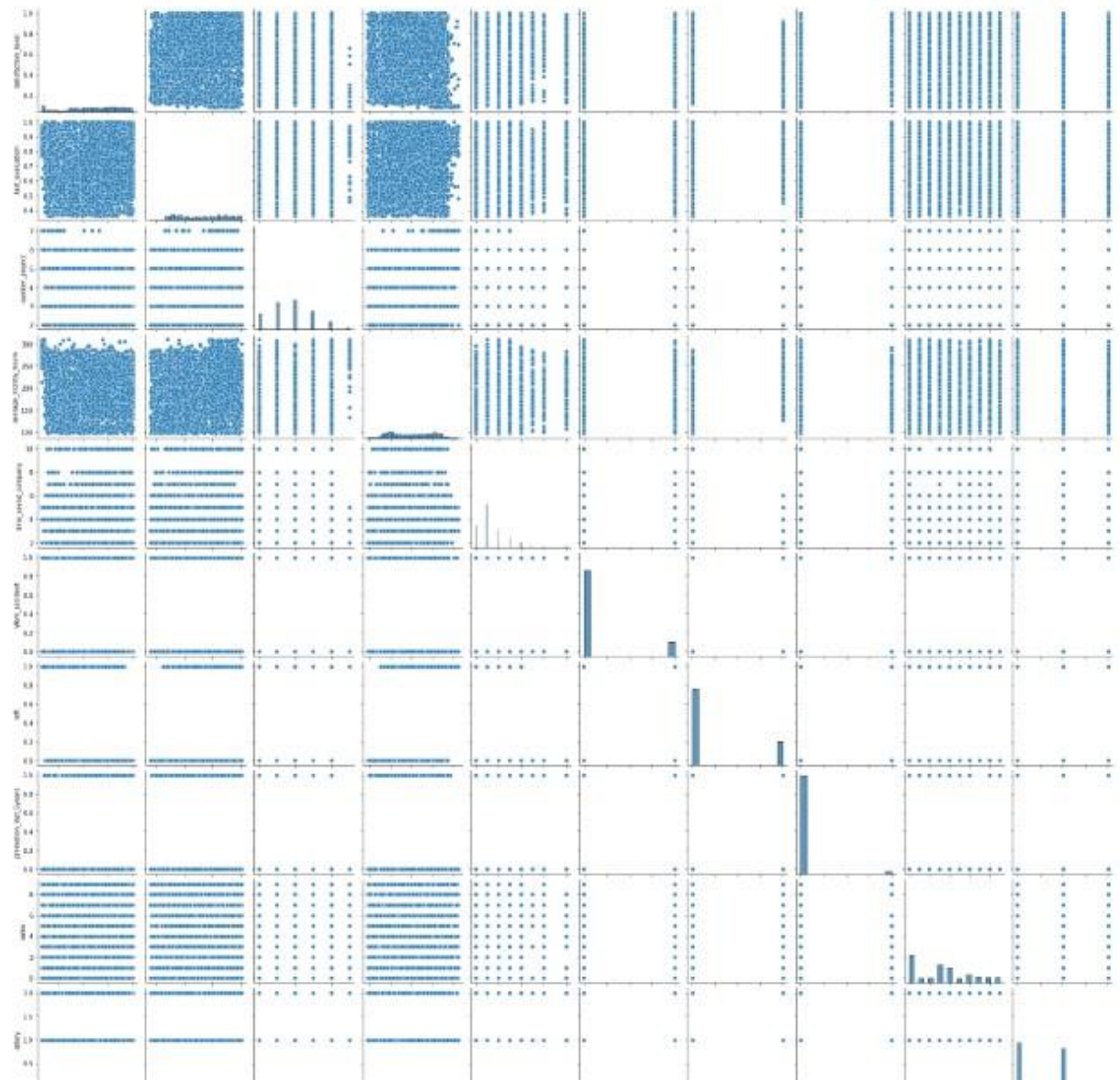
Bivariate Analysis

To find the relation between two features we use bivariate analysis. You can use seaborn package to plot visualisation using two variables of the dataset

Multivariate Analysis

```
sns.pairplot(df)
```

```
<seaborn.axisgrid.PairGrid at 0x2ab22b4e5b0>
```



Loading Data

Loading Data

```
: df = pd.read_csv('HR_comma_sep.csv')
```

: df

	satisfaction_level	last_evaluation	number_project	average_monthly_hours	time_spend_company	Work_accident	left	promotion_last_5years
0	0.38	0.53	2	157	3	0	1	
1	0.80	0.86	5	262	6	0	1	
2	0.11	0.88	7	272	4	0	1	
3	0.72	0.87	5	223	5	0	1	
4	0.37	0.52	2	159	3	0	1	
...	
14994	0.40	0.57	2	151	3	0	1	
14995	0.37	0.48	2	160	3	0	1	
14996	0.37	0.53	2	143	3	0	1	
14997	0.11	0.96	6	280	4	0	1	
14998	0.37	0.52	2	158	3	0	1	

14999 rows × 10 columns

Handling missing Values

```
df.shape
```

```
(14999, 10)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 14999 entries, 0 to 14998
```

```
Data columns (total 10 columns):
```

#	Column	Non-Null Count	Dtype
0	satisfaction_level	14999 non-null	float64
1	last_evaluation	14999 non-null	float64
2	number_project	14999 non-null	int64
3	average_monthly_hours	14999 non-null	int64
4	time_spend_company	14999 non-null	int64
5	Work_accident	14999 non-null	int64
6	left	14999 non-null	int64
7	promotion_last_5years	14999 non-null	int64
8	sales	14999 non-null	object
9	salary	14999 non-null	object

```
dtypes: float64(2), int64(6), object(2)
```

```
memory usage: 1.1+ MB
```

Feature Engineering Save processed data	Attached the codes in final submission -