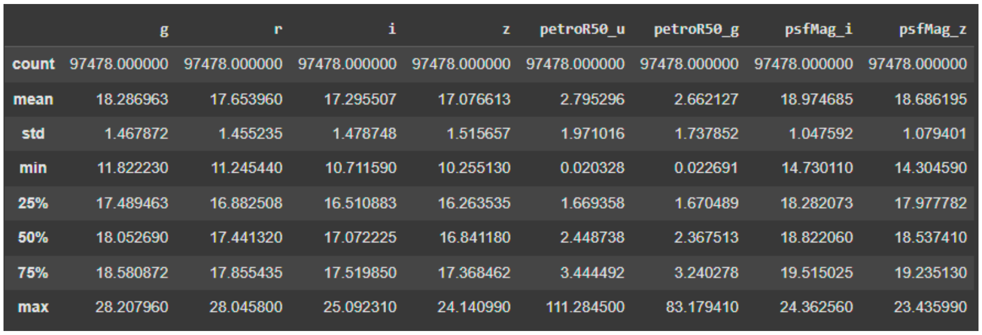
Data Collection and Preprocessing Phase

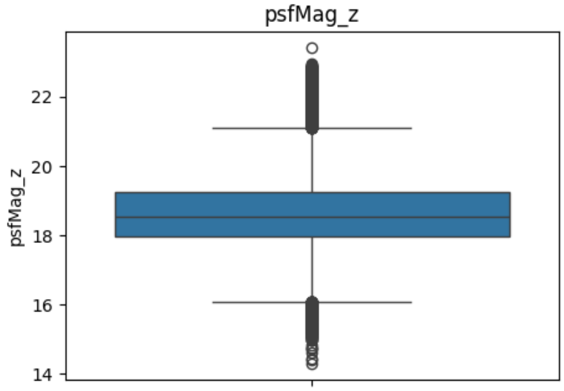
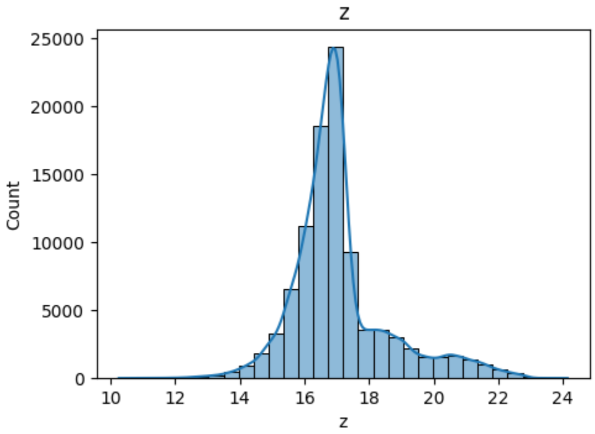
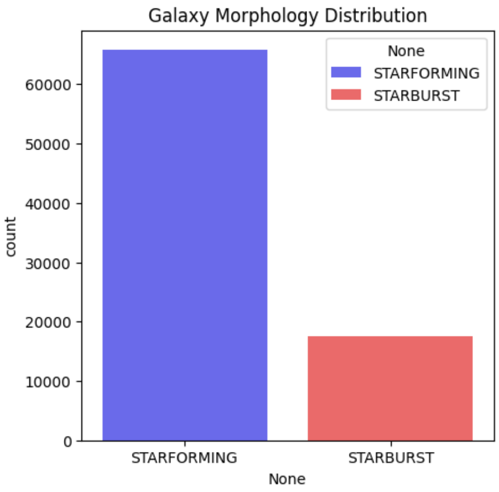
Date	18 June 2025
Team ID	xxxxxx
Project Title	sloan digital sky survey (sdss) galaxy classification using machine learning
Maximum Marks	6 Marks

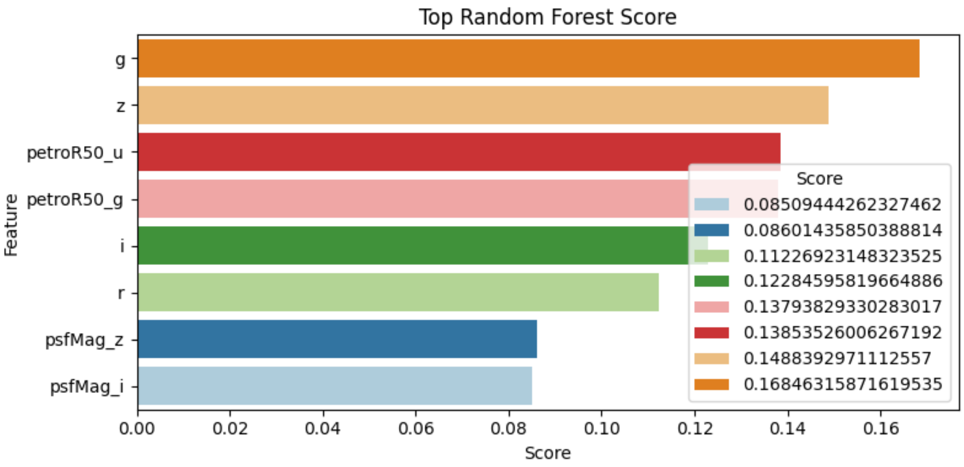
Data Exploration and Preprocessing

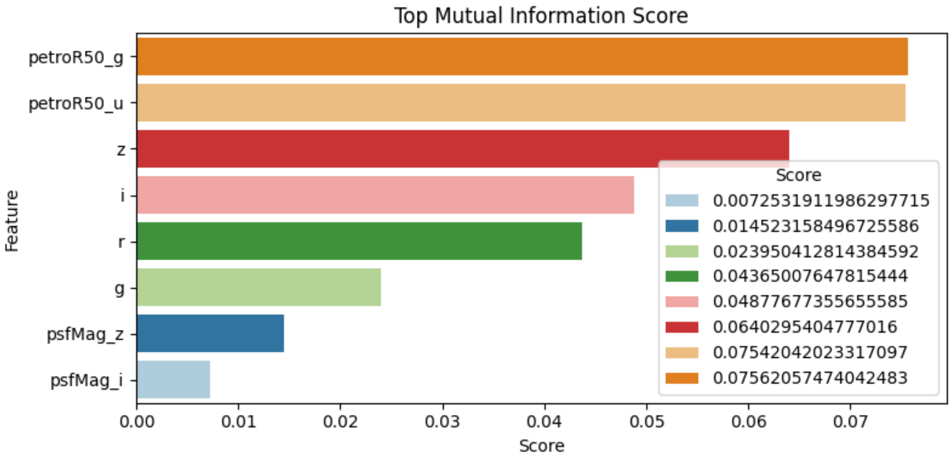
The dataset variables will undergo statistical analysis to detect patterns and outliers, while Python will be used for preprocessing steps such as normalization and feature engineering. Data cleaning will focus on handling missing values and outliers to ensure high-quality data for further analysis and modeling, establishing a solid basis for accurate insights and predictions.

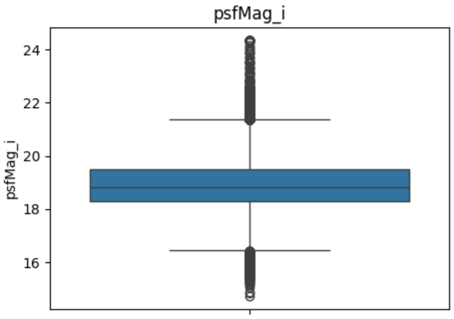
Section	Description
Data Overview	Shape: (97478, 8) 

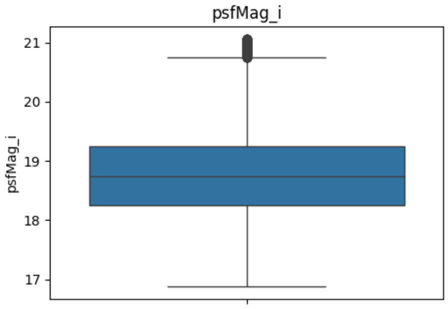
	g	r	i	z	petroR50_u	petroR50_g	psfMag_i	psfMag_z
count	97478.000000	97478.000000	97478.000000	97478.000000	97478.000000	97478.000000	97478.000000	97478.000000
mean	18.286963	17.653960	17.295507	17.076613	2.795296	2.662127	18.974685	18.686195
std	1.467872	1.455235	1.478748	1.515657	1.971016	1.737852	1.047592	1.079401
min	11.822230	11.245440	10.711590	10.255130	0.020328	0.022691	14.730110	14.304590
25%	17.489463	16.882508	16.510883	16.263535	1.669358	1.670489	18.282073	17.977782
50%	18.052690	17.441320	17.072225	16.841180	2.448738	2.367513	18.822060	18.537410
75%	18.580872	17.855435	17.519850	17.368462	3.444492	3.240278	19.515025	19.235130
max	28.207960	28.045800	25.092310	24.140990	111.284500	83.179410	24.362560	23.435990

Univariate Analysis	 
Bivariate Analysis	

<table>
<tr>
<td>Multivariate Analysis</td>
<td>


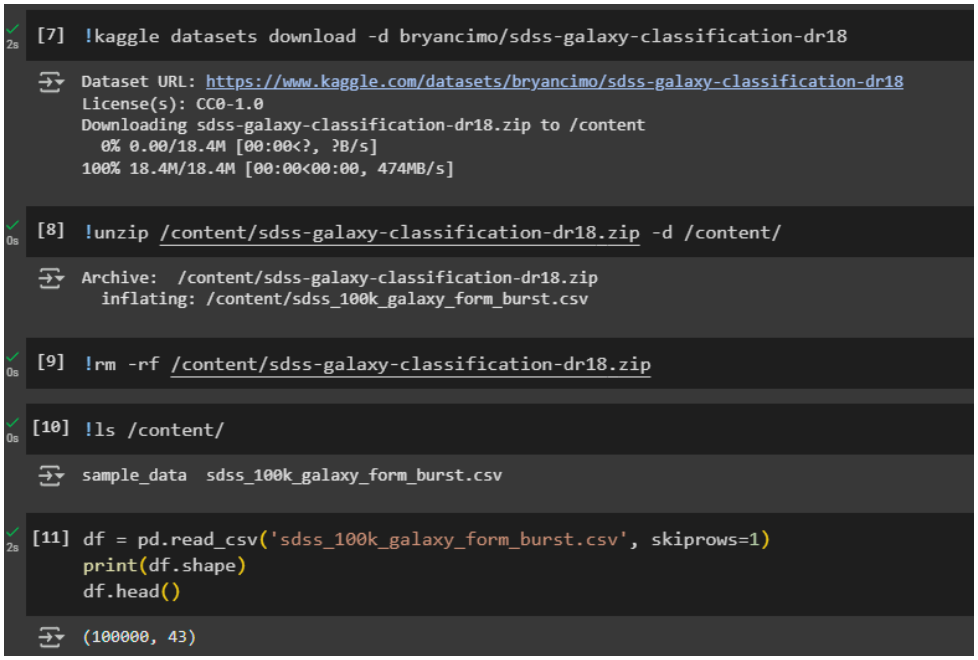


</td>
</tr>
<tr>
<td>Outliers and Anomalies</td>
<td>




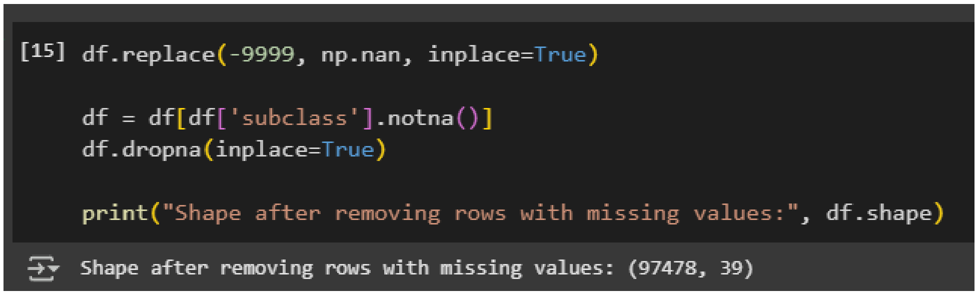
</td>
</tr>
<tr>
<td colspan="2"><b>Data Preprocessing Code Screenshots</b></td>
</tr>
</table>

Loading Data



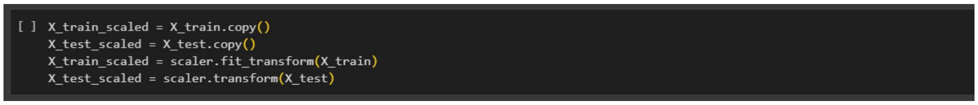
```
[7] !kaggle datasets download -d bryancimo/sdss-galaxy-classification-dr18

    Dataset URL: https://www.kaggle.com/datasets/bryancimo/sdss-galaxy-classification-dr18
    License(s): CC0-1.0
    Downloading sdss-galaxy-classification-dr18.zip to /content
      0% 0.00/18.4M [00:00<?, ?B/s]
    100% 18.4M/18.4M [00:00<00:00, 474MB/s]

[8] !unzip /content/sdss-galaxy-classification-dr18.zip -d /content/

    Archive:  /content/sdss-galaxy-classification-dr18.zip
      inflating: /content/sdss_100k_galaxy_form_burst.csv

[9] !rm -rf /content/sdss-galaxy-classification-dr18.zip

[10] !ls /content/

    sample_data  sdss_100k_galaxy_form_burst.csv

[11] df = pd.read_csv('sdss_100k_galaxy_form_burst.csv', skiprows=1)
     print(df.shape)
     df.head()

    (100000, 43)
```

Handling Missing Data



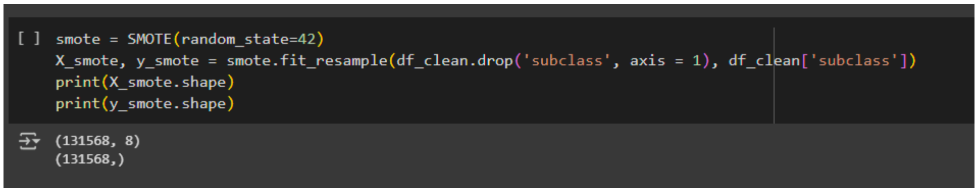
```
[15] df.replace(-9999, np.nan, inplace=True)

     df = df[df['subclass'].notna()]
     df.dropna(inplace=True)

     print("Shape after removing rows with missing values:", df.shape)

    Shape after removing rows with missing values: (97478, 39)
```

Data Transformation



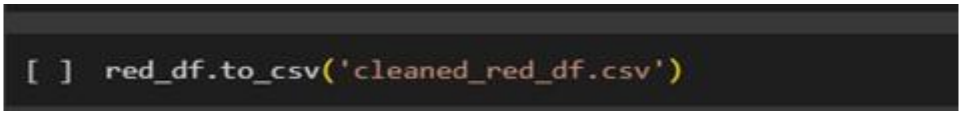
```
[ ] X_train_scaled = X_train.copy()
    X_test_scaled = X_test.copy()
    X_train_scaled = scaler.fit_transform(X_train)
    X_test_scaled = scaler.transform(X_test)
```

Feature Engineering



```
[ ] smote = SMOTE(random_state=42)
    X_smote, y_smote = smote.fit_resample(df_clean.drop('subclass', axis = 1), df_clean['subclass'])
    print(X_smote.shape)
    print(y_smote.shape)

    (131568, 8)
    (131568,)
```

Save Processed Data



```
[ ] red_df.to_csv('cleaned_red_df.csv')
```