Data Collection and Preprocessing Phase

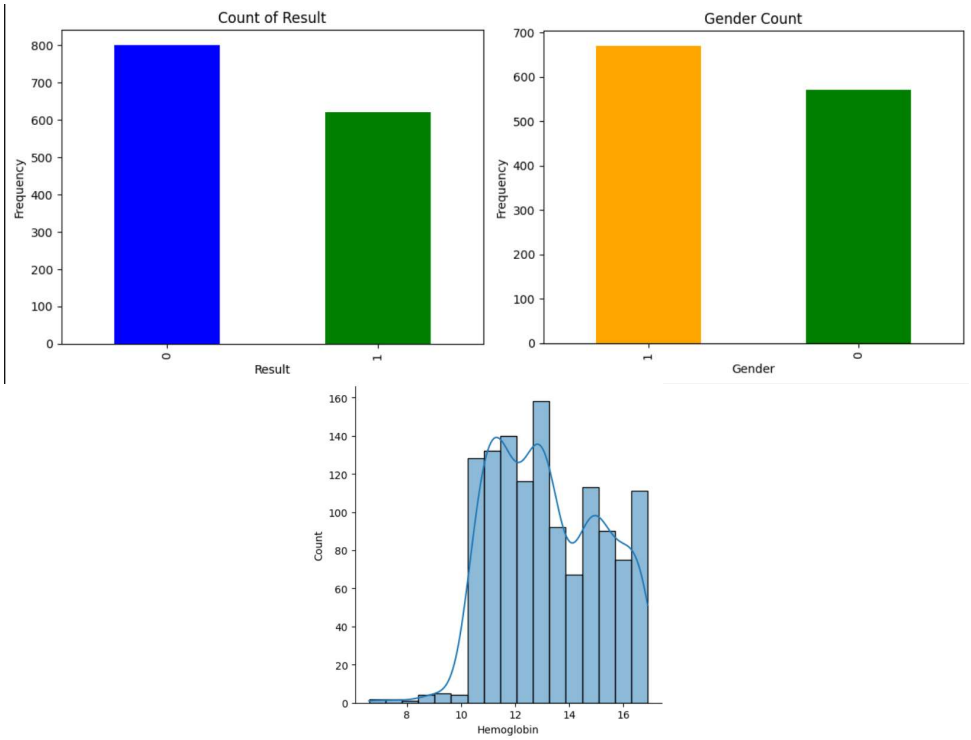
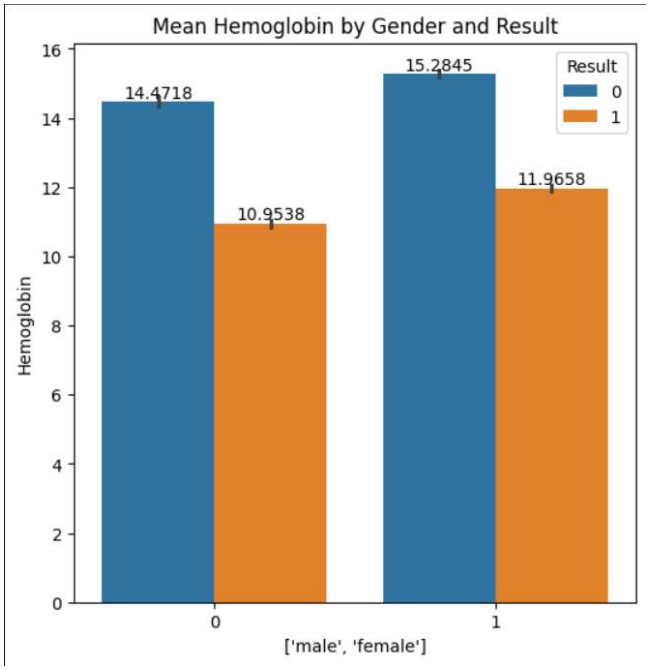
Date	6 August 2025
Team ID	SWUID20250185217
Project Name	Anemia Sense: Leveraging Machine Learning For Precise Anemia Recognitions
Maximum Marks	6 Marks

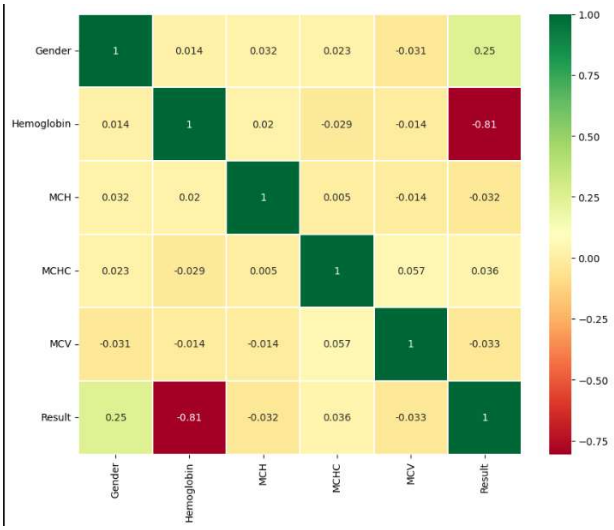
Data Exploration and Preprocessing Report

Dataset variables were statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and balancing. Data cleaning addressed missing values and ensured quality for subsequent analysis and modeling, forming a strong foundation for insights and predictions.

Section	Description
Data Overview	Dimension: 1421 rows × 6 columns

	Gender	Hemoglobin	MCH	MCHC	MCV	Result
0	1	14.9	22.7	29.1	83.7	0
1	0	15.9	25.4	28.3	72.0	0
2	0	9.0	21.5	29.6	71.2	1
3	0	14.9	16.0	31.4	87.5	0
4	1	14.7	22.0	28.2	99.5	0

Univariate Analysis	
Bivariate Analysis	

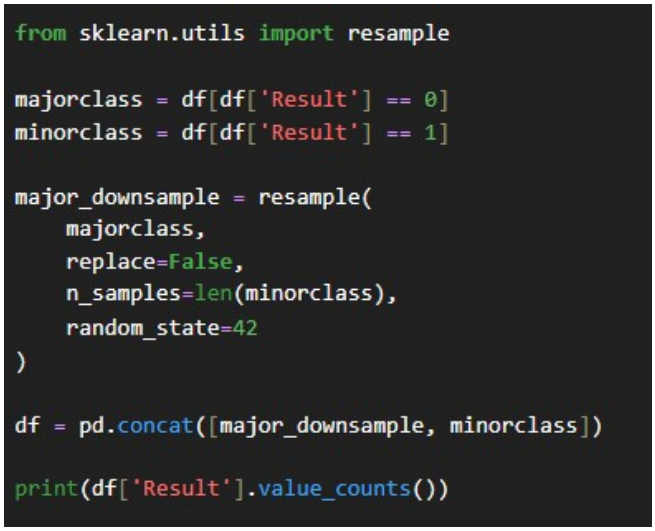
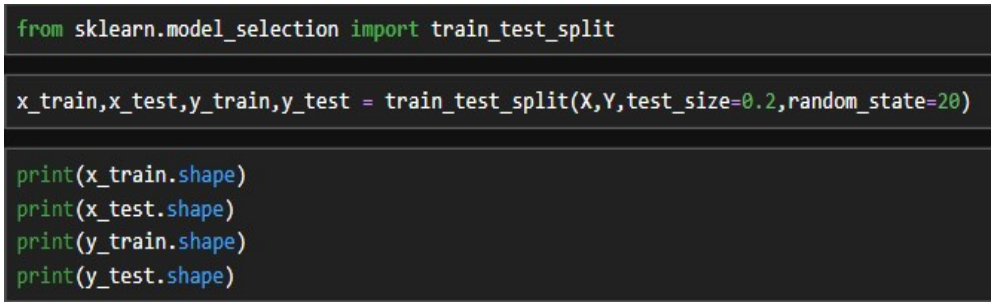
Multivariate Analysis	
Outliers and Anomalies	-

Data Preprocessing Code Screenshots

Loading Data

```
import pandas as pd
df = pd.read_csv('anemia.csv' )
df.head()
```

	Gender	Hemoglobin	MCH	MCHC	MCV	Result
0	1	14.9	22.7	29.1	83.7	0
1	0	15.9	25.4	28.3	72.0	0
2	0	9.0	21.5	29.6	71.2	1
3	0	14.9	16.0	31.4	87.5	0
4	1	14.7	22.0	28.2	99.5	0

Data Transformation & Balancing	
Feature Engineering	

```python
from sklearn.utils import resample

majorclass = df[df['Result'] == 0]
minorclass = df[df['Result'] == 1]

major_downsample = resample(
    majorclass,
    replace=False,
    n_samples=len(minorclass),
    random_state=42
)

df = pd.concat([major_downsample, minorclass])

print(df['Result'].value_counts())
```

```python
from sklearn.model_selection import train_test_split

x_train,x_test,y_train,y_test = train_test_split(X,Y,test_size=0.2,random_state=20)

print(x_train.shape)
print(x_test.shape)
print(y_train.shape)
print(y_test.shape)
```