Data Collection and Preprocessing Phase

Date	14 JUNE 2025
Team ID	SWTID1749621188
Project Title	Anemia Sense Leveraging-Machine Learning-For-Precise Anemia Recognition
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template

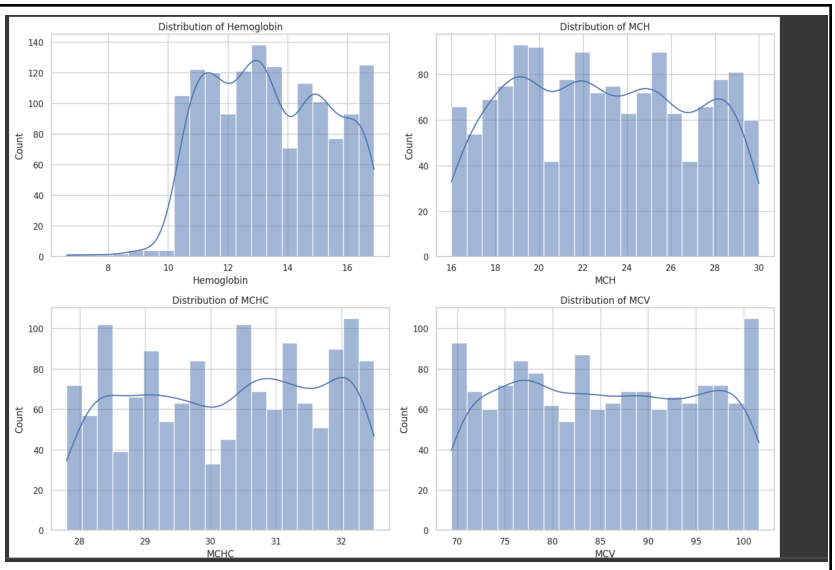
The dataset was obtained from a structured CSV file containing hematological parameters like Hemoglobin, MCH, MCHC, MCV, and a binary anemia result. Initial exploration involved checking the shape, data types, and value distributions. No missing values or duplicates were found. Numerical columns were analyzed through descriptive statistics and visualized using histograms, KDE plots, and box plots. Outliers were detected in Hemoglobin and MCH and reviewed for impact. Since all features were numerical, no encoding was required. The cleaned and processed data was then prepared for further modeling.

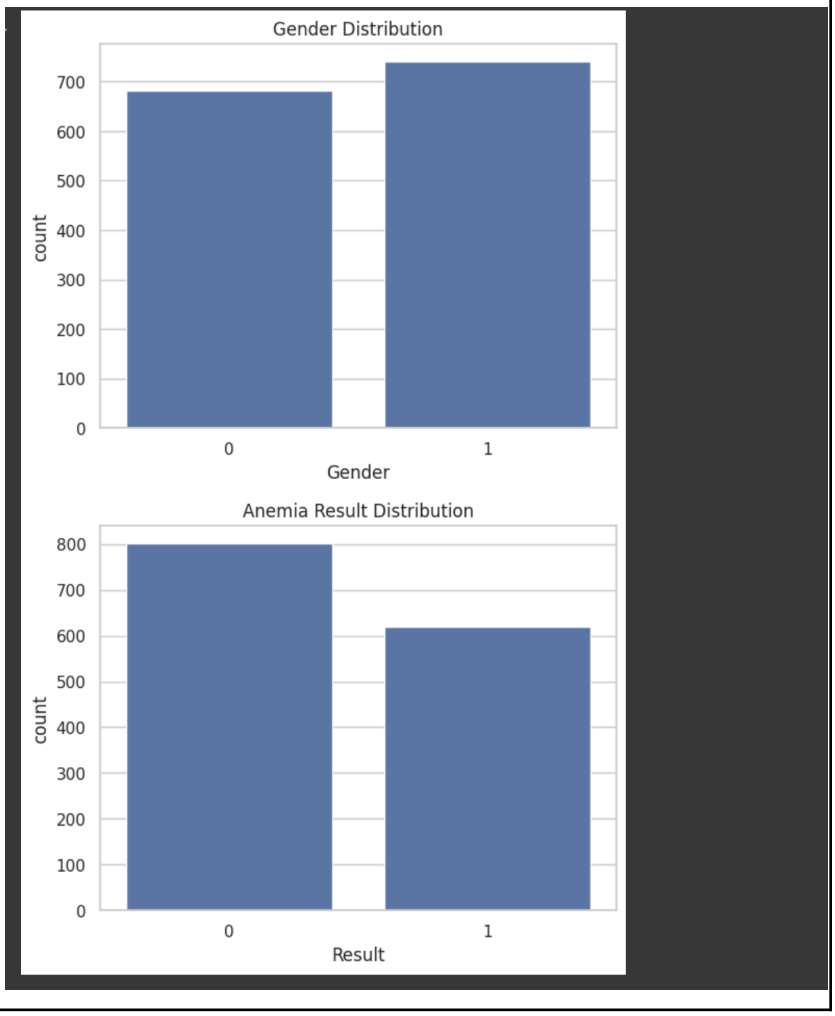
Section	Description
Data Overview	**DIMENSIONS:** `Shape of the dataset (rows, columns): (1421, 6)` **DESCRIPTIVE STATISTICS:**

```
Descriptive statistics for numerical columns:
              Gender   Hemoglobin          MCH         MCHC          MCV  \
count  1421.000000  1421.000000  1421.000000  1421.000000  1421.000000
mean      0.520760    13.412738    22.905630    30.251232    85.523786
std       0.499745     1.974546     3.969375     1.400898     9.636701
min       0.000000     6.600000    16.000000    27.800000    69.400000
25%       0.000000    11.700000    19.400000    29.000000    77.300000
50%       1.000000    13.200000    22.700000    30.400000    85.300000
75%       1.000000    15.000000    26.200000    31.400000    94.200000
max       1.000000    16.900000    30.000000    32.500000   101.600000

              Result
count  1421.000000
mean      0.436312
std       0.496102
min       0.000000
25%       0.000000
50%       0.000000
75%       1.000000
max       1.000000
```

Univariate Analysis



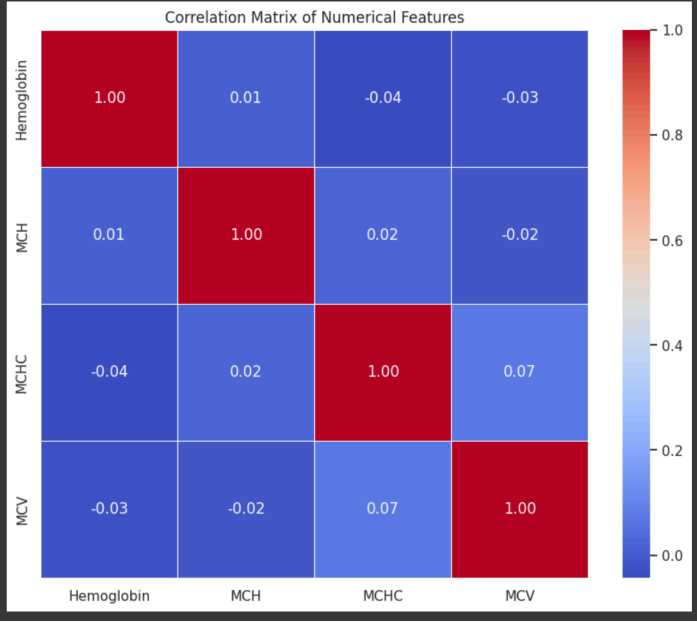


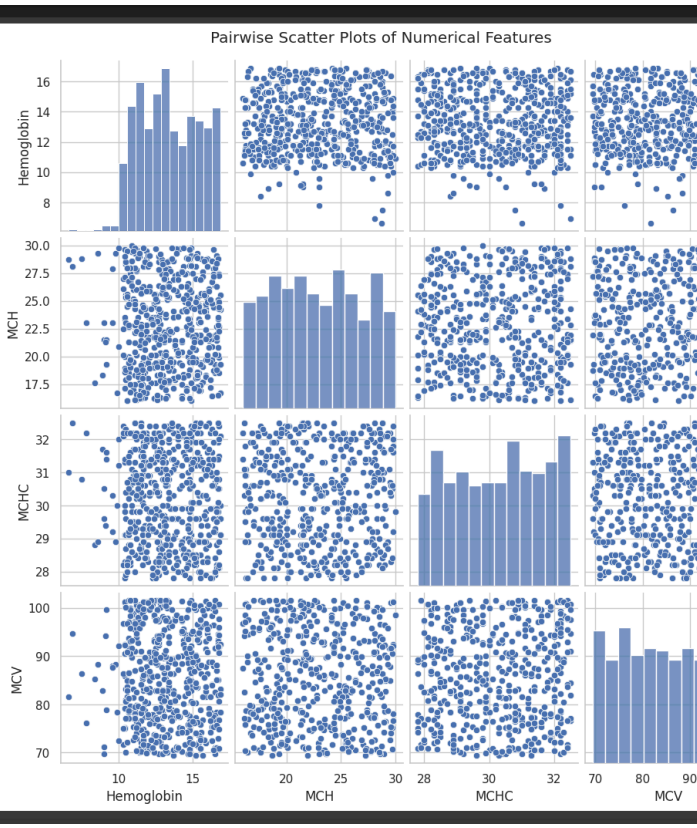
Bivariate Analysis

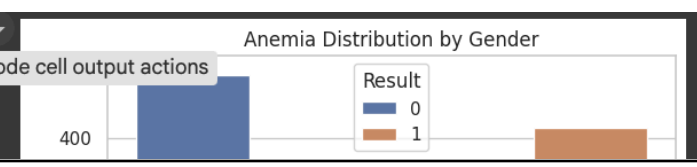
A. Numerical vs. Numerical:

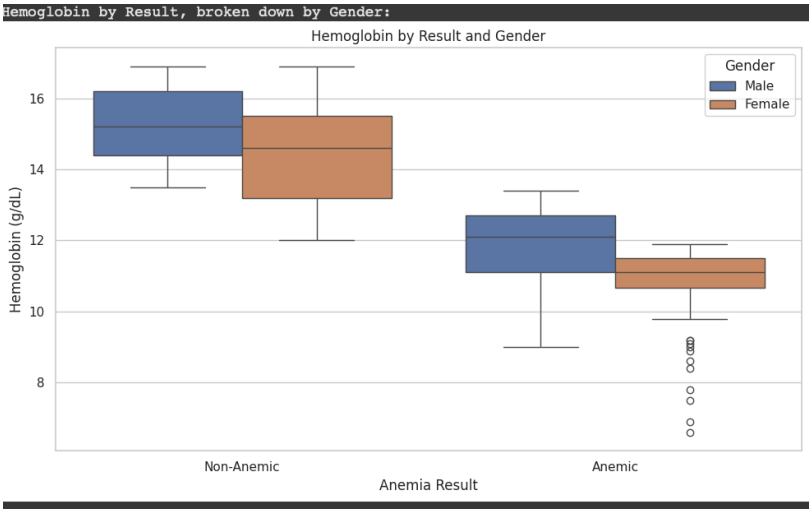
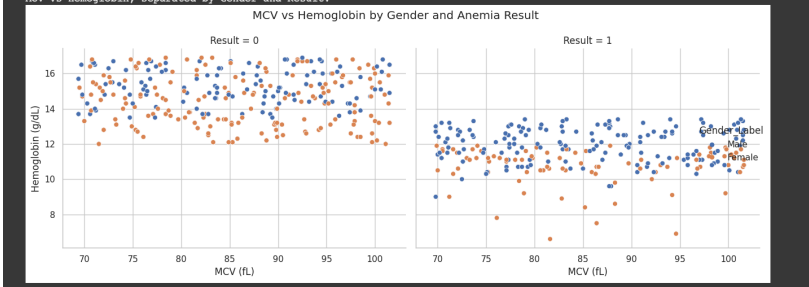
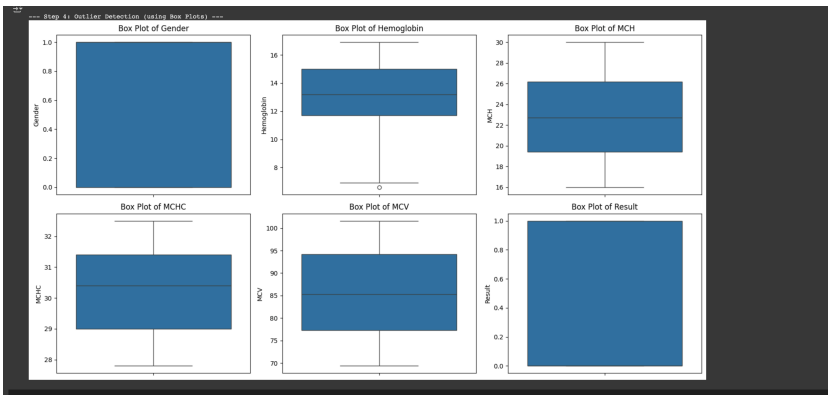
Correlation Matrix:

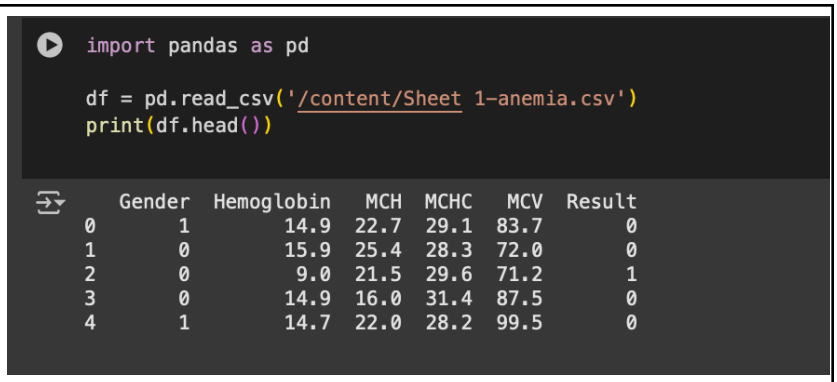
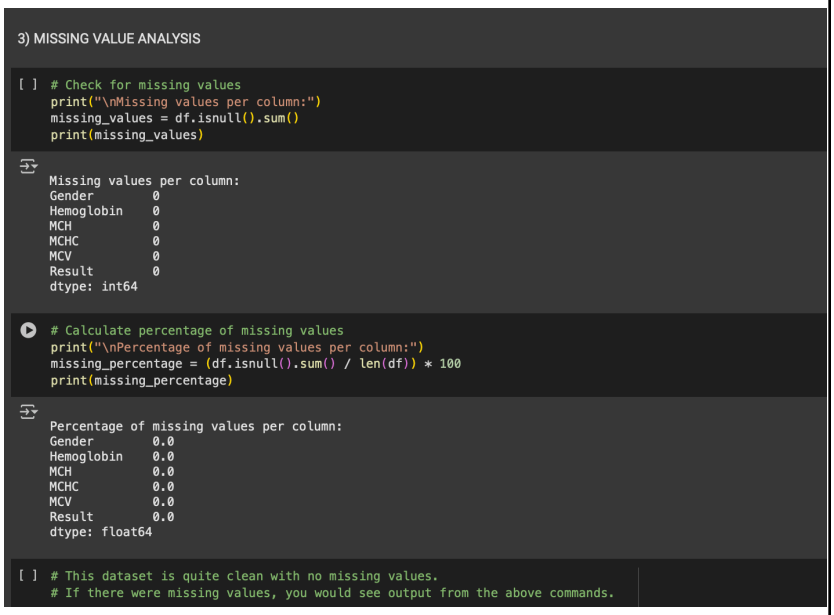
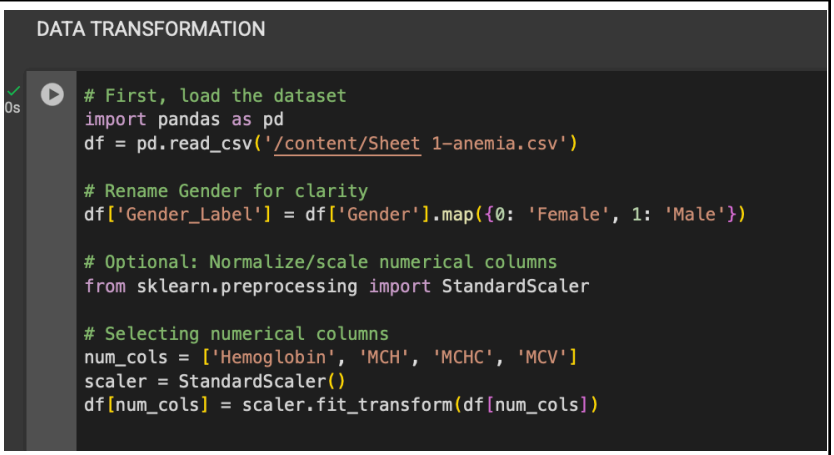
	Hemoglobin	MCH	MCHC	MCV
Hemoglobin	1.000000	0.014081	-0.042597	-0.025885
MCH	0.014081	1.000000	0.018795	-0.015948
MCHC	-0.042597	0.018795	1.000000	0.068450
MCV	-0.025885	-0.015948	0.068450	1.000000







Multivariate Analysis	Hemoglobin by Result, broken down by Gender:  MCV vs Hemoglobin, separated by Gender and Result: 
Outliers and Anomalies	
Data Preprocessing Code Screenshots	

Loading Data	
Handling Missing Data	
Data Transformation	

Feature Engineering	Attached the feature engineering code in the final submission
Save Processed Data