


## Data Collection and Preprocessing Phase

Date	14 July 2024
Team ID	739755
Project Title	Blood Donation Prediction
Maximum Marks	6 Marks

## Data Exploration and Preprocessing Template

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description																																																						
Data Overview	<div><div>EXPLORATORY DATA ANALYSIS</div><div>1.Descriptive statistical</div><div><div>df.describe()</div><table><thead><tr><th></th><th>Recency (months)</th><th>Frequency (times)</th><th>Monetary (c.c. blood)</th><th>Time (months)</th><th>whether he/she donated blood in March 2007</th></tr></thead><tbody><tr><td>count</td><td>703.000000</td><td>703.000000</td><td>703.000000</td><td>703.000000</td><td>703.000000</td></tr><tr><td>mean</td><td>8.645804</td><td>4.448080</td><td>1112.019815</td><td>31.783784</td><td>0.224751</td></tr><tr><td>std</td><td>8.163291</td><td>3.381765</td><td>847.841140</td><td>22.512287</td><td>0.417715</td></tr><tr><td>min</td><td>0.000000</td><td>1.000000</td><td>250.000000</td><td>2.000000</td><td>0.000000</td></tr><tr><td>25%</td><td>3.000000</td><td>2.000000</td><td>500.000000</td><td>14.000000</td><td>0.000000</td></tr><tr><td>50%</td><td>8.000000</td><td>3.000000</td><td>750.000000</td><td>28.000000</td><td>0.000000</td></tr><tr><td>75%</td><td>14.000000</td><td>6.000000</td><td>1500.000000</td><td>46.000000</td><td>0.000000</td></tr><tr><td>max</td><td>74.000000</td><td>14.000000</td><td>3500.000000</td><td>98.000000</td><td>1.000000</td></tr></tbody></table></div></div>		Recency (months)	Frequency (times)	Monetary (c.c. blood)	Time (months)	whether he/she donated blood in March 2007	count	703.000000	703.000000	703.000000	703.000000	703.000000	mean	8.645804	4.448080	1112.019815	31.783784	0.224751	std	8.163291	3.381765	847.841140	22.512287	0.417715	min	0.000000	1.000000	250.000000	2.000000	0.000000	25%	3.000000	2.000000	500.000000	14.000000	0.000000	50%	8.000000	3.000000	750.000000	28.000000	0.000000	75%	14.000000	6.000000	1500.000000	46.000000	0.000000	max	74.000000	14.000000	3500.000000	98.000000	1.000000
	Recency (months)	Frequency (times)	Monetary (c.c. blood)	Time (months)	whether he/she donated blood in March 2007																																																		
count	703.000000	703.000000	703.000000	703.000000	703.000000																																																		
mean	8.645804	4.448080	1112.019815	31.783784	0.224751																																																		
std	8.163291	3.381765	847.841140	22.512287	0.417715																																																		
min	0.000000	1.000000	250.000000	2.000000	0.000000																																																		
25%	3.000000	2.000000	500.000000	14.000000	0.000000																																																		
50%	8.000000	3.000000	750.000000	28.000000	0.000000																																																		
75%	14.000000	6.000000	1500.000000	46.000000	0.000000																																																		
max	74.000000	14.000000	3500.000000	98.000000	1.000000																																																		
Univariate Analysis	<div><div>Code + Text</div><div>UNIVARIATE ANALYSIS</div><div><div><div>sns.countplot(x='whether he/she donated blood in March 2007',data=df)</div><div>&lt;Axes: xlabel='whether he/she donated blood in March 2007', ylabel='count'&gt;</div></div><div></div></div></div>																																																						

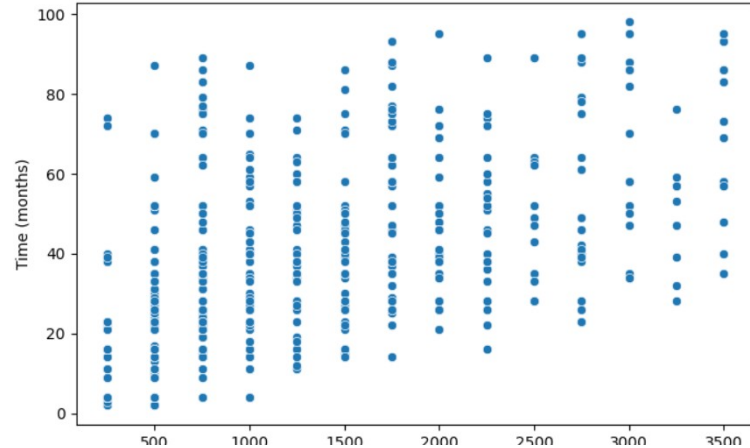
## Bivariate Analysis

CODE + TEXT

### BIVARIATE ANALYSIS

```
sns.scatterplot(x=df['Monetary (c.c. blood)'],y=df['Time (months)'])
```

<Axes: xlabel='Monetary (c.c. blood)', ylabel='Time (months)'

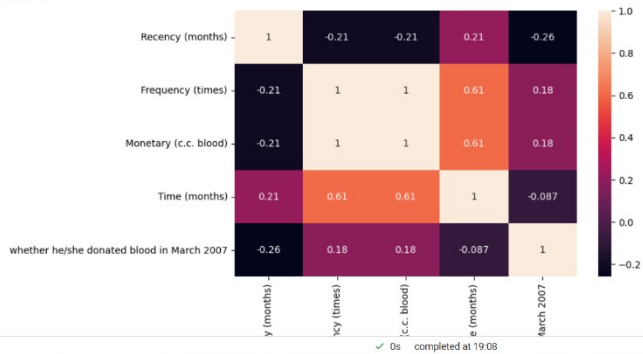


## Multivariate Analysis

### MULTIVARIATE ANALYSIS

```
[92] sns.heatmap(df.corr(),annot=True)
```

<Axes: >



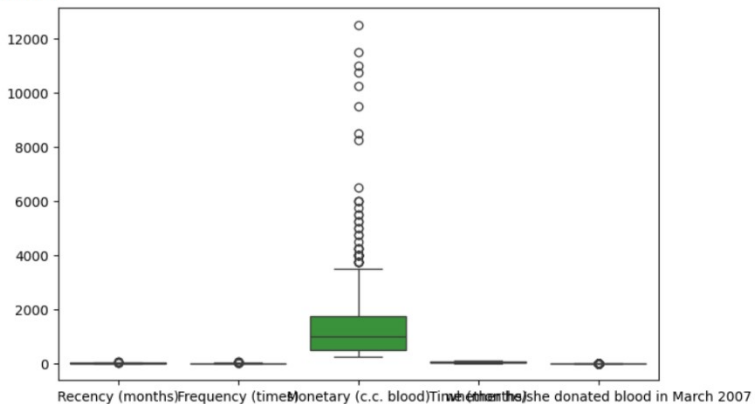
0s completed at 19:08

## Outliers and Anomalies

CODE + TEXT

```
sns.boxplot(df)
```

<Axes: >



## Data Preprocessing Code Screenshots

### Loading Data

```
[62]: df = pd.read_csv("/content/transfusion (2).csv")
      df.head()
```

	Recency (months)	Frequency (times)	Monetary (c.c. blood)	Time (months)	whether he/she donated blood in March 2007
0	2	50	12500	98	1
1	0	13	3250	28	1
2	1	16	4000	35	1
3	2	20	5000	45	1
4	1	24	6000	77	0

### Handling Missing Data

```
df.isnull().sum()
```

```
Recency      0
Frequency    0
Monetary     0
Time         0
Donated      0
dtype: int64
```

### Data Transformation

```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
scaled_features = scaler.fit_transform(df.drop(columns=["whether he/she donated blood in March 2007"]))
scaled_df = pd.DataFrame(scaled_features, columns=df.columns[:-1])

scaled_df["whether he/she donated blood in March 2007"] = df["whether he/she donated blood in March 2007"].values

# Display the transformed DataFrame
print("\nTransformed DataFrame:")
print(scaled_df.head())
```

```
Transformed DataFrame:
   Recency (months)  Frequency (times)  Monetary (c.c. blood)  Time (months) \
0      0.000000      0.923077      0.923077      0.270833
1      0.054954      0.230769      0.230769      0.020833
2      0.027027      0.461538      0.461538      0.125000
3      0.013514      0.846154      0.846154      0.343750
4      0.027027      0.615385      0.615385      0.208333

   whether he/she donated blood in March 2007
0      1
1      0
2      1
3      0
4      1
```

### Feature Engineering

Attached the codes in final submission.

### Save Processed Data

-