Part 1: Data Summarization

Attribute Information:

Given is the variable name, variable type, the measurement unit and a brief description. The "Blood Transfusion Service Center" is a classification problem. The order of this listing corresponds to the order of numerals along the rows of the database.

1. R (Recency - months since last donation),
2. F (Frequency - total number of donation),
3. M (Monetary - total blood donated in c.c.),
4. T (Time - months since first donation), and
5. a binary variable representing whether he/she donated blood in March 2007 (1 stand for donating blood; 0 stands for not donating blood).

Below is the dataframe with a summary table of each feature.

	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
...	...	...	...	...	...
743	23	2	500	38	0
744	21	2	500	52	0
745	23	3	750	62	0
746	39	1	250	39	0
747	72	1	250	72	0

748 rows × 5 columns

Figure 1: Dataframe

	Recency (months)	Frequency (times)	Monetary (c.c. blood)	Time (months)	whether he/she donated blood in March 2007
count	748.000000	748.000000	748.000000	748.000000	748.000000
mean	9.506684	5.514706	1378.676471	34.282086	0.237968
std	8.095396	5.839307	1459.826781	24.376714	0.426124
min	0.000000	1.000000	250.000000	2.000000	0.000000
25%	2.750000	2.000000	500.000000	16.000000	0.000000
50%	7.000000	4.000000	1000.000000	28.000000	0.000000
75%	14.000000	7.000000	1750.000000	50.000000	0.000000
max	74.000000	50.000000	12500.000000	98.000000	1.000000

Figure 2: Summary Table

Part 2: Visualization

Part 2.1: Box Plots

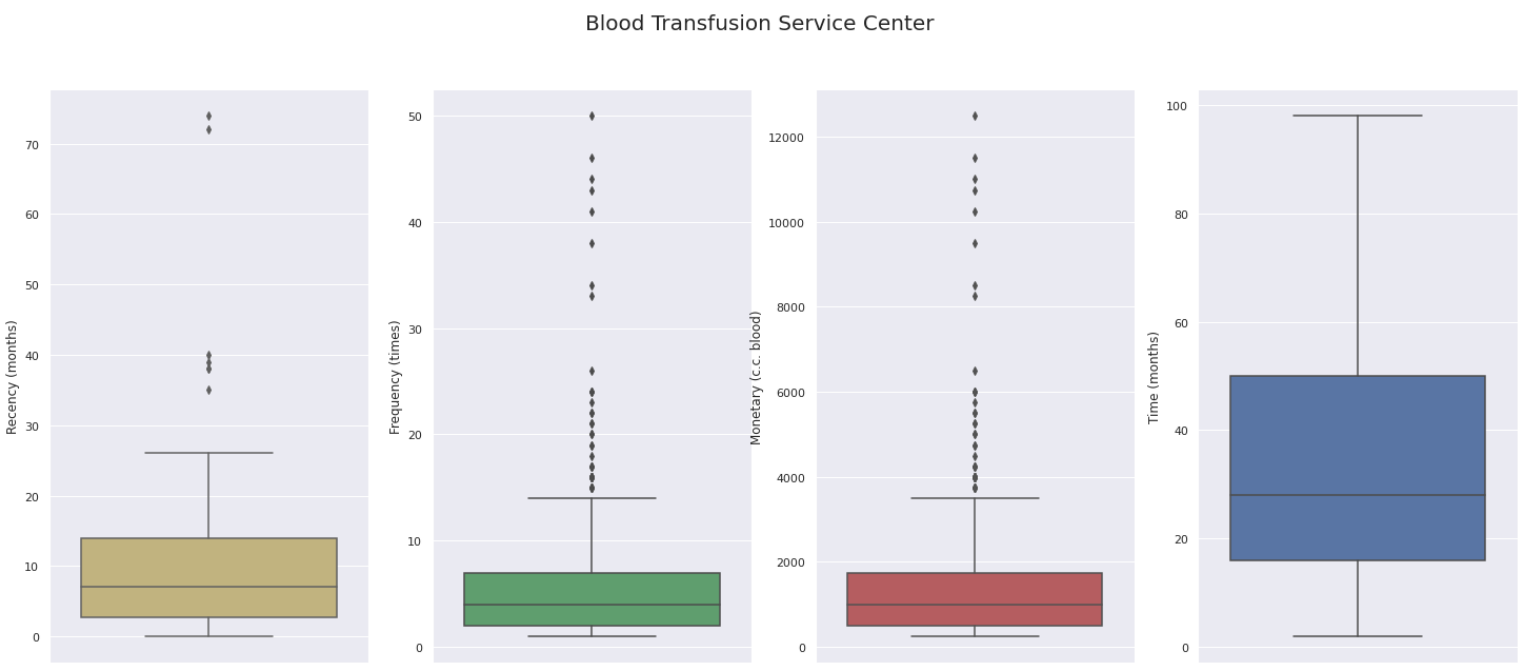

Figure 3: Box plots for each feature

Part 2.2: Scatter Plots

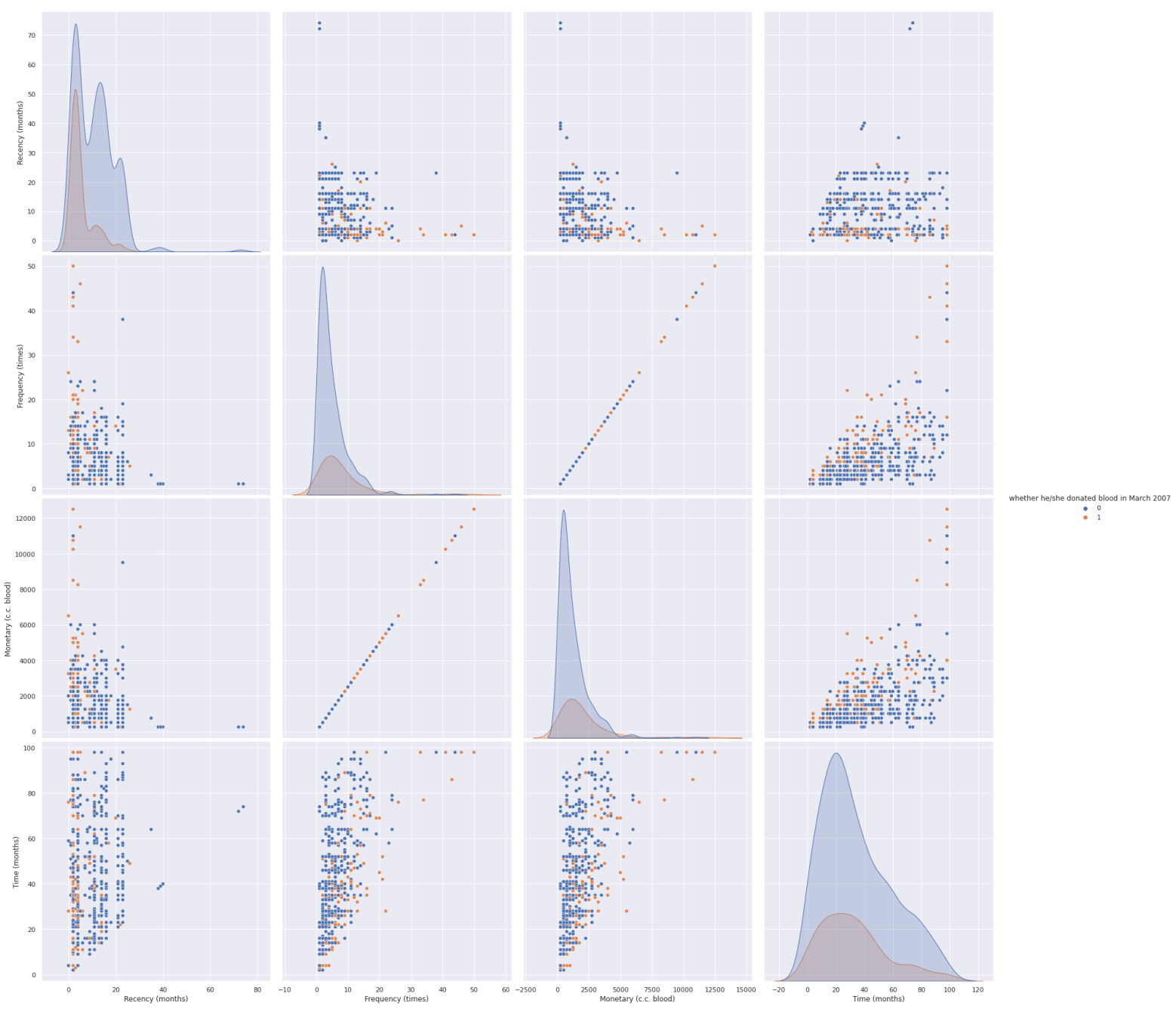

Figure 4: Pairwise scatter plots of features,

the orange dots show those who have donated blood in March 2007

the blue dots show those who have not donated blood in March 2007

Part 4: Interpretation

- From **boxplots**, we can see that the most spread out feature is **time**. This means when using this feature, the distance between data points is larger and hence they can probably be classified better. The next best option would be the **recency** feature.

- From the **pairwise scatterplots**, we can see that the **time** feature with **frequency** or **moneray** can cluster out the data points to some extent. Although they are not completely successful, they're the best we can get from using scatterplots. Also the **frequency** and **moneray** features are highly correlated and in fact, there's a linear relationship between these two.