ابتدا دیتاها را از ورودی می‌خوانیم. خلاصه‌ای از دیتافریم بدین صورت است:

```
In [2]: data_frame = pd.read_csv("transfusion.data")
        data_frame
```

Out[2]:

	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

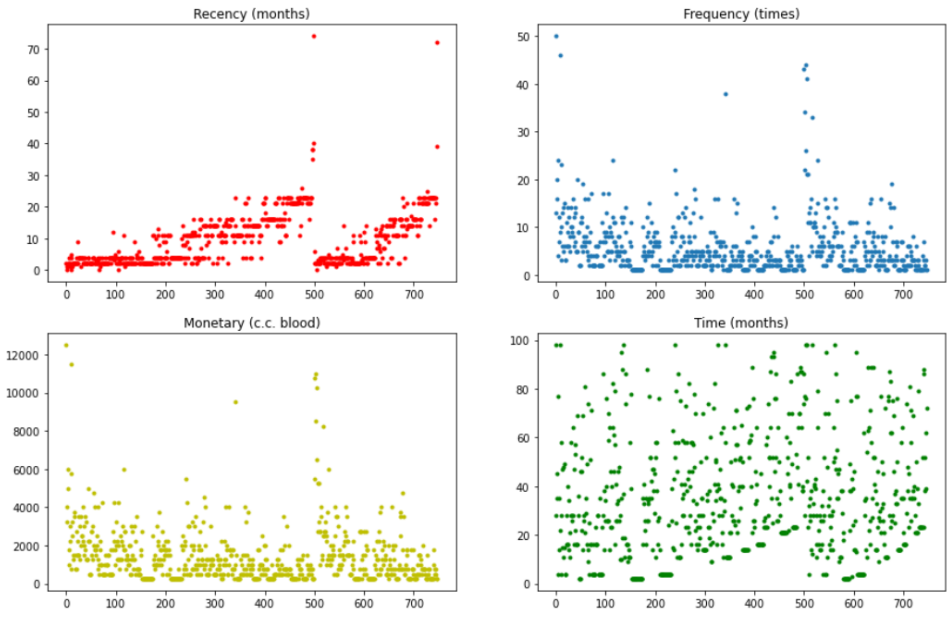
دیتا دارای ۷۴۸ سطر و ۵ ستون است. همچنین خلاصه‌ای از مشخصات دیتا به صورت زیر است:

```
In [3]: data_frame.describe()
```

Out[3]:

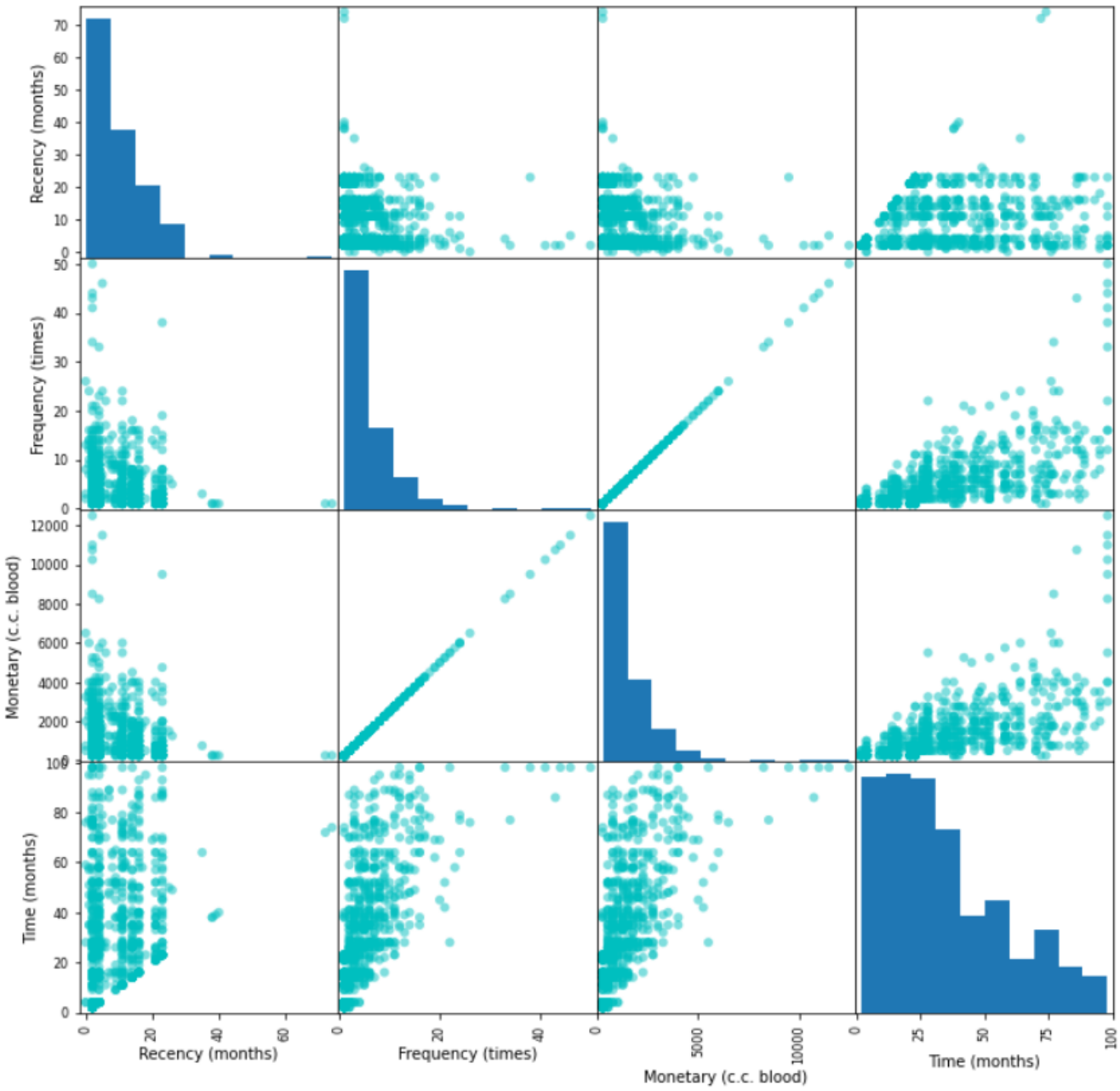
	Recency (months)	Frequency (times)	Monetary (c.c. blood)	Time (months)	whether he/she donated blood in March 2007
count	748.000000	748.000000	748.000000	748.000000	748.000000
mean	9.506684	5.514706	1378.676471	34.282086	0.237968
std	8.095396	5.839307	1459.826781	24.376714	0.426124
min	0.000000	1.000000	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

ابتدا نمودارهای مربوط به هر ستون را رسم کرده‌ایم. نتایج بدین صورت است:



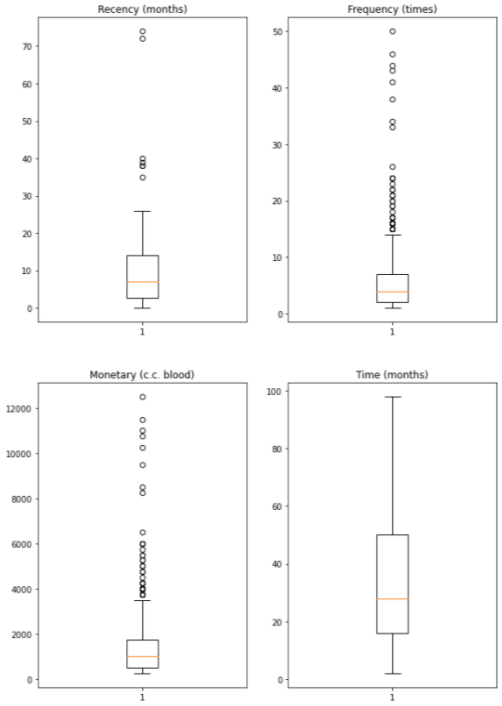
همان‌طور که مشخص است، اطلاعات time از پراکندگی بیشتری برخوردار هستند.

سپس نمودار همسبتگی دیتا را رسم می‌کنیم.



با توجه به نمودارهای بالا، Ferquency با Monetary ارتباط مستقیم و خطی دارد. در مابقی نمودارها چنین ارتباط مستقیمی دیده نمی‌شود.

حال باکس پلات‌های دیتا را رسم می‌کنیم:



طبق مباحث درس، آیتمی بهتر است که نمودار باکس پلات آن کشیده‌تر باشد. بنابراین بر این اساس، آیتم Time از مابقی آیتم‌ها بهتر است.

اما نمودار باکس پلات دیگری بر اساس ستون آخر دیتا ("whether he/she donated blood in March 2007") رسم می‌کنیم. آیتمی که باکس پلات‌های همپوشانی کمتری دارند می‌بایست عملکرد بهتری نیز داشته باشند. بر این اساس آیتم Recency مناسب‌تر از بقیه به نظر می‌رسد.