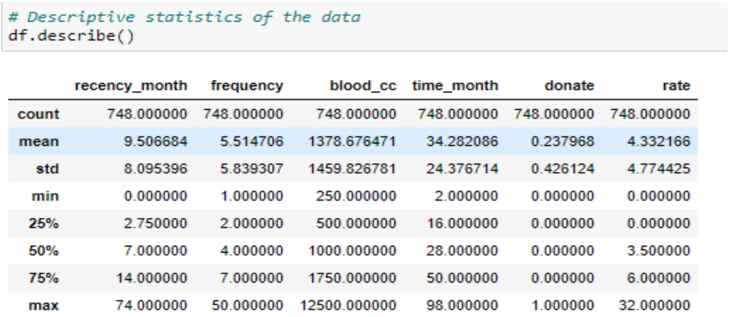
Section 1: Summary Statistic

Data Description & Summary

The dataset “Blood Transfusion Service Centre Dataset”, which was retrieved from UCI Machine Leaning Repository, includes the target variable of whether a donor donated blood in 2017. The aim of this assignment is to apply data analysis on pre-processed data and visualization skills in Python for the observation of the relationship between the attributes. The process of data pre-processing is stated in Section 1 (Data Pre-processing). *Figure 1.1* is the statistical summary of the finalized dataset.

Figure 1.1 DataFrame description



```
# Descriptive statistics of the data
df.describe()
```

```
       recency_month   frequency       blood_cc  time_month      donate        rate
count     748.000000  748.000000     748.000000  748.000000  748.000000  748.000000
 mean       9.506684    5.514706    1378.676471   34.282086    0.237968    4.332166
  std       8.095396    5.839307    1459.826781   24.376714    0.426124    4.774425
  min       0.000000    1.000000     250.000000    2.000000    0.000000    0.000000
  25%       2.750000    2.000000     500.000000   16.000000    0.000000    0.000000
  50%       7.000000    4.000000    1000.000000   28.000000    0.000000    3.500000
  75%      14.000000    7.000000    1750.000000   50.000000    0.000000    6.000000
  max      74.000000   50.000000   12500.000000   98.000000    1.000000   32.000000
```

Table 1.1 Attributes Description

Attribute	Description	Data type
Recency_month	Last donation in months	Int
frequency	Total number of times donation	Int
Blood_cc	Total volume of blood donated (c.c)	Int
Time_month	First donation in months	Int
Donate	Whether he/she donated blood in March 2007 {0,1}	Int (binary)
rate	Months per donate	Int
Rate_status	Whether donor is “often” or “seldom” to donate their blood	String

*red word indicated target variable

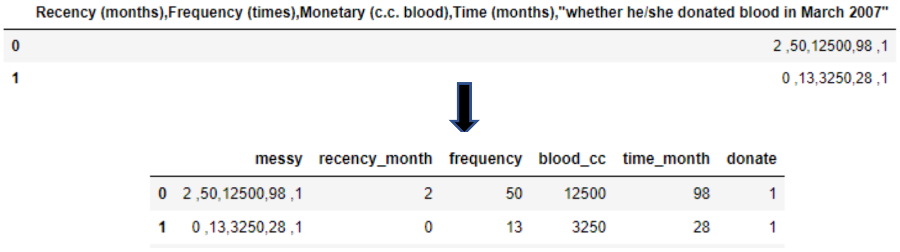
The dataset consist of 748x7 of raw data, it was defined as 748 rows and 7 columns with `df.shape()`. Based on *figure 1.1*, the mean in ascending order is: `donate`, `rate`, `frequency`, `recency_month`, `time_month` and `blood_cc`. The volume of blood donated (in cc.) has the highest standard deviation among all attributes, which means that there are great difference between the mean and total blood volume donated by each donor. The standard deviation in descending order following `blood_cc` are time in months since first donation, time in months since most recent donation, total donation times, rate of donation and the least is donate as the value of donate are binary. The minimum value of `recency_month`, `donate`, and rate are 0. The minimum total frequency of blood donations is one indicates that all 748 data all donated at least 1 time and the minimum volume of blood donated is 250 cc. The minimum durations since first donation was 2 months.

After arranging the data in ascending order, we observed that:

- `recency_month` – 25% of the 748 people had last donated blood at most 2.75 months ago, 50% had it at most 7 months ago, 75% at most a year and 2 months ago. The longest duration since last donation was 74 months.
- `frequency` – 25% of the 748 people had donated at most 2 times, 50% had at most 4 times, 75% at most 7 times. The highest frequency of donation was 50 times.
- `blood_cc` – 25% of the 748 people had donated at most a total volume of 500 cc., 50% had at most 1000 cc., 75% 1750 cc. at most. The largest volume of blood donated was 12500 cc.
- `time_month` – 25% of the 748 people had started blood donation at most in the last 16 months ago, 50% had started at most 28 months ago, 75% had started at most 50 months ago. The earliest starting period of time was 98 months ago, that is 8 years and 4 months ago.
- `donate` – Up to 75% of the 748 people did not donate blood.
- `rate` – 25% of 748 people would not donate blood, 50% would donate every 3 and a half months, 75% would donate every 6 and a half months. The maximum rate is that some people would donate once every 32 months.

Data Preprocessing

Data preprocessing is an important step to clean our dataset before we start to analyze our data. The original data was combined all the attributes and element into string format with dot symbol. First, rename the attribute to “messy”. Create split_dot function to split each of the elements and re-assign to the new data frame columns.



	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

⬇

	messy	recency_month	frequency	blood_cc	time_month	donate
0	2 ,50,12500,98 ,1	2	50	12500	98	1
1	0 ,13,3250,28 ,1	0	13	3250	28	1

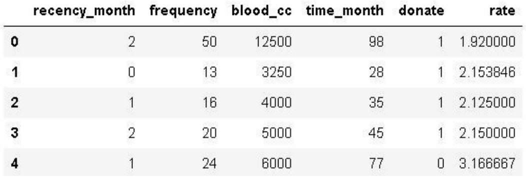
Next, we created logical test to check whether the data set is contained null value or not. If there exist null value, the dropna will be executed and returned to current rows and columns by using df.shape. After that, we removed useless “messy” column which may not be used for visualization.

Categorization of Feature

Now we add on a new feature to determine the rate of period takes for a particular donor to donate blood again by using the formula below.

$$\textbf{Rate of donating blood} = \frac{(First\ donate - Recency)}{Frequency}$$

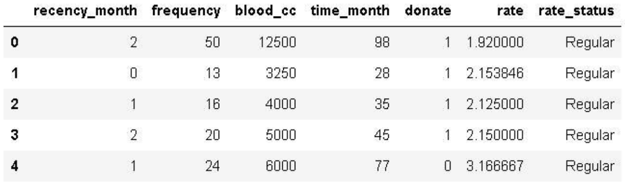
However, the new dataframe is updated and as shown below.



	recency_month	frequency	blood_cc	time_month	donate	rate
0	2	50	12500	98	1	1.920000
1	0	13	3250	28	1	2.153846
2	1	16	4000	35	1	2.125000
3	2	20	5000	45	1	2.150000
4	1	24	6000	77	0	3.166667

For instance, imagine a donor with the rate of donating blood is 1.92, which indicates the particular donor will donate once every 1.92 months. When the rate is zero, where the donor only donated once and it cannot be used for estimation of the expected months to donate again.

Since the rate has not been categorised yet, then we can assume the rate by categorising them into two different groups. By defining the new columns (rate_status), we declare "first-timer" donor and "regular" donor as the rate is zero and rate more than zero respectively. As a result, the updated dataframe is shown below.



	recency_month	frequency	blood_cc	time_month	donate	rate	rate_status
0	2	50	12500	98	1	1.920000	Regular
1	0	13	3250	28	1	2.153846	Regular
2	1	16	4000	35	1	2.125000	Regular
3	2	20	5000	45	1	2.150000	Regular
4	1	24	6000	77	0	3.166667	Regular

The reason that we have separated them is because of the difficulties of estimating the expected months for a donor to donate again given only donated once, while a regular donor is having an interval of the period that can be used for estimation of the next donation.

Section 2: Visualization

Radar Plot

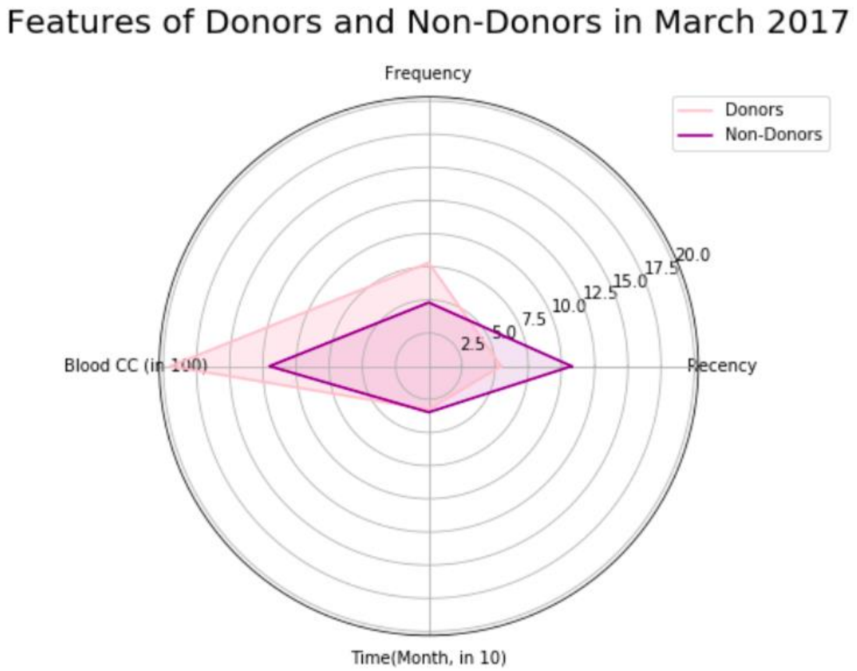


Figure 2.1 Radar Plot of the Features of Donors and Non-Donors in March 2017.

This radar chart compares the various mean values of donors and non-donors' features in March 2017. The pink and purple lines, including the regions covered respectively, each refers to blood donors and non-donors in March 2017. Based on *figure 2.1*, overall, we perceived that the donors have a higher average value in frequency a person has donated blood and the total volume of blood donated compared to non-donors. For donors in average, the total times and volume of blood donated is 8 times and 1950 cc respectively whereby 5 times and 8.5 cc for non-donors. This means that blood donors in March 2017 has donated blood many more times and donated a lot more blood than non-donors.

On the other hand, blood donors have averagely lower recency value than non-donors, which means that the time duration since their last donation to their donation in March 2017 is shorter. Averagely, we can see that each blood donors last donated blood was 3 months ago whereby non-donors in March 2017 last donated was 8 and a half months ago. Nevertheless, the time duration since first donation to their donation in March 2017, for donors and non-donors, does not differ much, each person has started their first blood donation since 30 and a half months ago. In other words, each person has at least 2 and a half years blood donation history.

Thus, we can infer that a person who has donated more blood, more times and more often is more likely to donate blood in March 2017.

Histogram

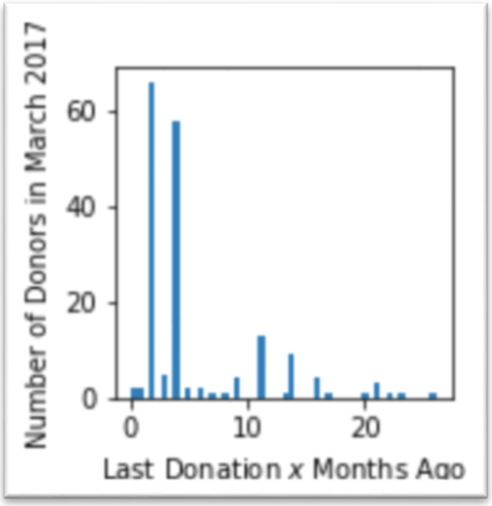


Figure 2.2.1 Histogram of the Numbers of Donors Last Donated Blood in x Months Ago.

According to the histogram, we observed that most of the donors donated blood 2 and 4 months ago, a total of approximately 123 donors out of 178 donors. This shows that the donors whose last donation was more recent in donating blood, is more likely to donate blood in March 2017.

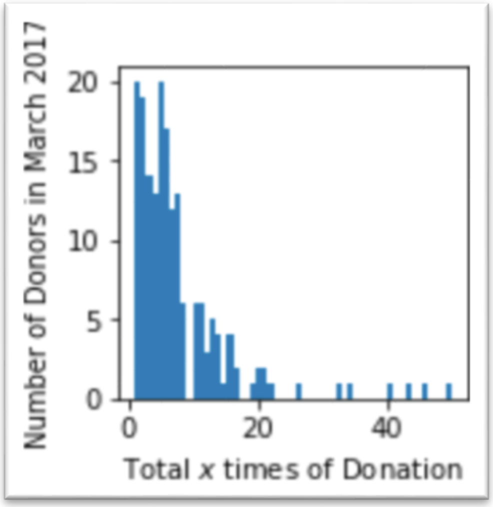


Figure 2.2.2 Histogram of the Numbers of Donors Had x times of Donation.

The histogram shows the number of donors in March 2017 for each total *x* times of donation. We can see that over 80% of the donors in March 2007 has donated blood less than 10 times. Which means than very little of the donors has done at least 10 times blood donation. In other words, we can say that blood donors in March 2017 are mostly new donors, if we assume less-than-10-times donors as new donors.

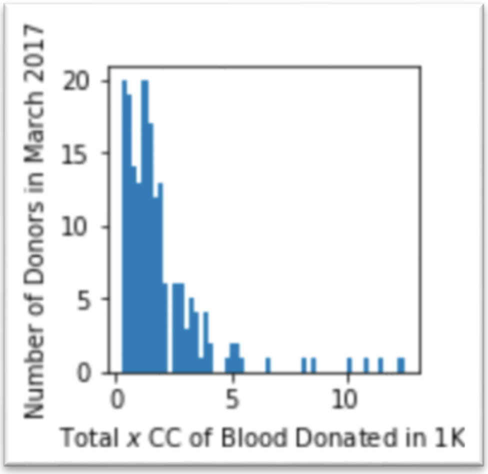


Figure 2.2.3 Histogram of the Numbers of Donors Donated x CC of Blood in 1K.

Figure 2.2.3 shows the total volume in cc of blood donated ever since first donation by *y* numbers of donors in March 2017. We can observed that there are 40 donors in total who have donated 330cc and 1167cc of blood each 20 persons. We can also see that less than half of the blood donors has donated more than 2500cc of blood ever since they started blood donation. Having an overview with *figure 2.2.2*, which shows most donors in March 2017 are new donors (with less than 10 times blood donation), we could infer that it is either new donors are relatively strong in physique that they could have donated more blood at a time or that old donors has long blood donation history.

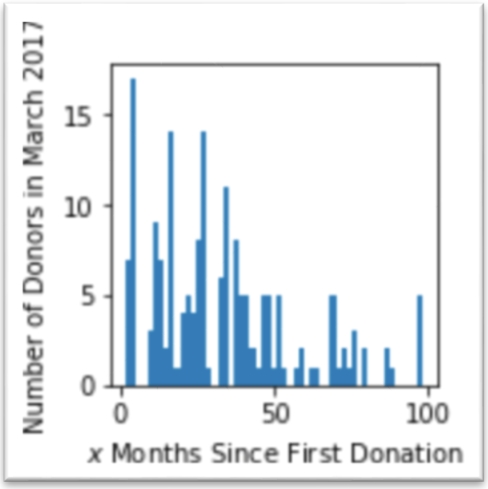


Figure 2.2.4 Histogram of the Numbers of Donors with Different Length of Blood Donation History.

Figure 2.2.4 gives us an overview of the number of donors having *x* months of blood donation history. We can see that the number of blood donors which is more than 10 are blood donors who has less than 50 months blood donation history. This is corresponding to the *figure 2.2.2* that shows most of the donors in March 2017 are new donors with shorter duration since first donation and donated less times than the minority (donated more than 10 times and more than 50 months). In short, majority of the donors have around 4 years blood donation history.

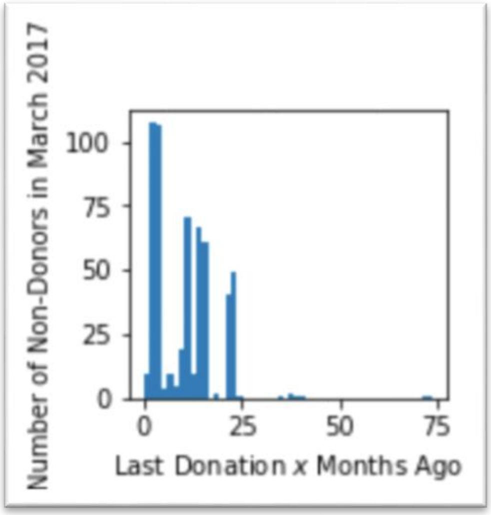


Figure 2.2.5 Histogram of the Numbers of Non-Donors Last Donated Blood in x Months Ago.

Figure 2.2.5 gives a quick overview of the numbers of non-donors last donated blood *x* months ago. Based on *figure 2.2.5*, we observed that the number of non-donors in March 2007 has most recent donation in 2 to 3 months ago. Overall, compared to *figure 2.21*, non-donors have greater difference of recency range (75 months) than donors (25 months). We also observed that most of the non-donors last donated was within 2 years ago.

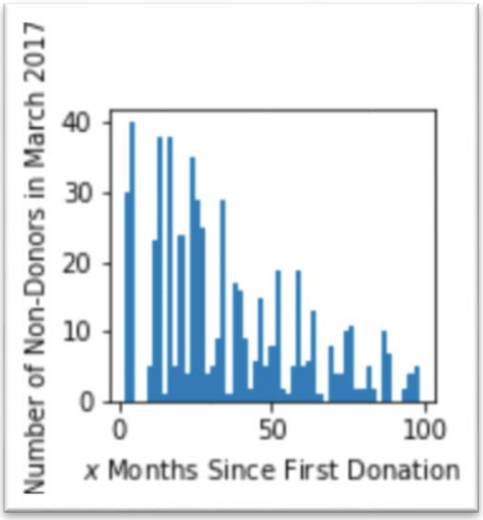


Figure 2.2.6 Histogram of the Numbers of Non-Donors Had x times of Donation.

This graph shows us the number of non-donors of a total of *x* times of blood donation. Overall, almost all non-donors in March 2017 had 2 to 20 times of blood donation before. The number of non-donors decreased as the total number of times of donation rises. We can also say that most of the non-donors are not active blood donors. Unlike donors, having the minority who has donated more than 20 times, non-donors have a relatively smaller mean value though both have positive skewed distributions. On the other hand, comparing *figure 2.2.6* to *2.2.5*, we can infer that almost all

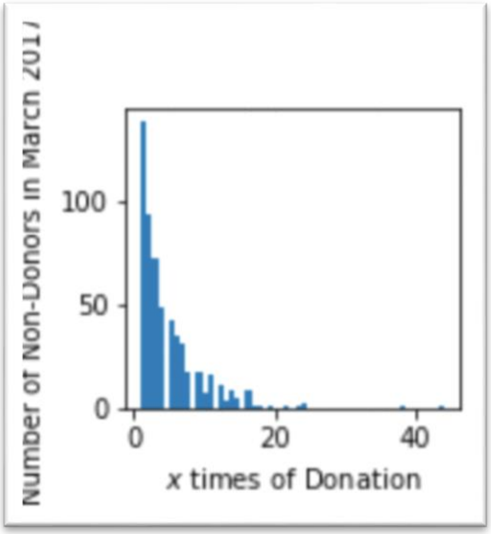


Figure 2.2.7 Histogram of the Numbers of Non-Donors Donated x CC of Blood in 1K.

Figure 2.2.7 shows us the numbers of non-donors of the total volume of *x* cc of blood donated. In all, 99% of non-donors in March 2017 had donated less than 5000cc of blood before. The distribution is also positively skewed which indicates that when the volume of donated blood rises, number of non-donors decreases.

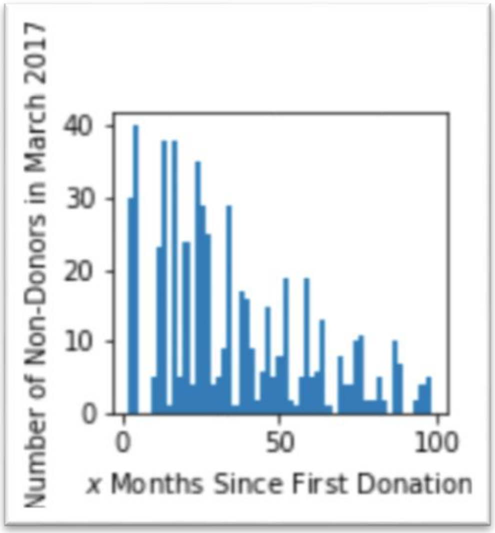


Figure 2.2.8 Histogram of the Numbers of Non-Donors with Different Length of Blood Donation History.

Based on the *figure 2.2.8*, we can see that more than half of the non-donors in March 2017 had a 4-year (48 months) blood donation history. Minority of the non-donors had been donating blood for 4 to 8 years. We could infer that those who has shorter blood donation history would be less likely to donate blood.

Box Plot

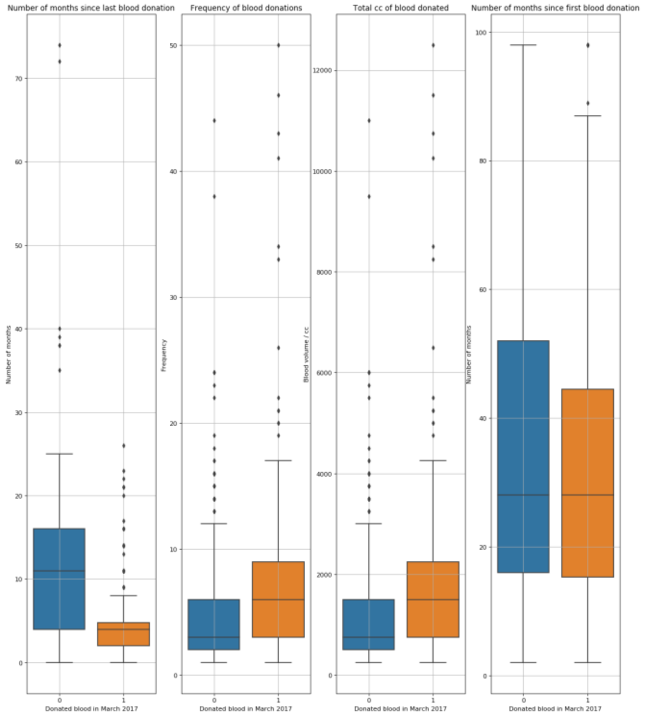


Figure 2.3.1 to Figure 2.3.4: Box Plots of Various Data

Figure 2.3.1 to figure 2.3.4 show 4 box plots of number of months passed since last blood donation, frequency of blood donation done, total volume of blood donated, and number of months passed since the first blood donation. In the box plots, the blue represents people who did not donate blood on March of 2017 and orange represents people who donated on said date. From the boxplots, we can see that all of the data are negatively skewed with various skewness, meaning only minority of the people have high numbers in the four categories. Based on figure 2.3.1. most people in the data donated relatively recently having the box at closer to zero. Two people who have not donated blood in the past over 70 months being the most outlying outliers among the seven outliers of the people who did not donate blood on March 2017. On the other hand, we can conclude that those who donated blood on March 2017 donated blood more recently relative to those who did not, even the most outlying outlier of it is around 2 years. From this we can make inference that people who did not donate blood for a longer time are less likely to donate blood on March 2017.

Twelve and eleven data points are lying outside of the frequency of blood donations box plot in figure 2.3.2. Overall, those donated blood on March 2017 have donated blood more times than those who did not. Hence, it is inferred that frequent donor of blood are more likely to donate blood. The boxplot of total volume of blood in figure 2.3.3 agrees with the previous inference as the frequent donor tend to donate more blood in total. People who have started to donate blood earlier tend not to donate blood on March of 2017. This may be because of certain chronic disease correlate to aging preventing them from donating blood because there are criteria of health condition for a person to qualify as blood donor. Hence, we can see in figure 2.3.4 data of people who donated blood on March of 2017, is more negatively skewed than those who did not despite having the same median in the boxplot of number of months since the first blood donation session.

As a conclusion from the four box plots, people who are more likely to donate blood on March 2017 are most likely people who donated blood recently around the past 5 months, has donated blood at least 5 times or more, has donated at least 1500 cubic centimeters of blood in total. And are relatively young.

Scatter Plot

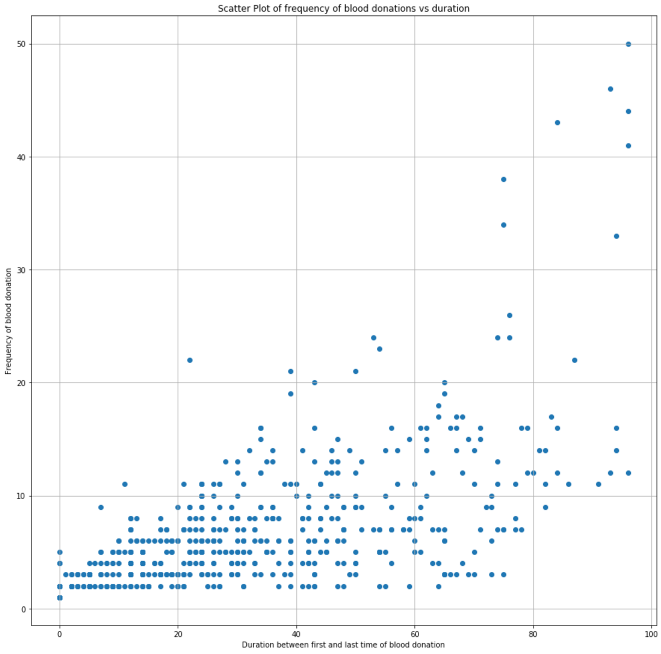


Figure 2.4: Scatter Plot of Frequency of Blood Donation against Duration

Figure 2.4 shows the scatter plot of frequency of blood donations against duration between first and last blood donation session of each person. We get the duration by taking the number of months passed after first blood donation from that of the most recent blood donation. The measurement of frequency is by the hospital so it is not counting our regular blood donating session. Hence, it is normal to see people with 3 people from the scatter plot having more than 1 in frequency despite donated blood only one time. From this scatter plot we can observe an upward trend of the data meaning people who have donated for a longer period of time, donated more times in that duration. Majority of the people have donated blood less than 30 times in period of 80 months; contrary, there are eight individuals who had donated blood over 30 times and upwards of 50 times in the sample data, they are the outliers of the public. Most people donated blood in less than 10 times over a few years of time, because the minimum interval of two blood donation is three months for the body to restore its iron stock. Besides, it is even better for people to wait for more than the minimum requirement before donating again might be contributing to people donated blood relatively less times throughout long duration of time. Generally, we can plot a best fit line as $y = \frac{1}{7}x$ in the scatter plot inferring on average a person donates blood once every seven months. As a conclusion, the pattern of the scatter plot suggests there is a moderately strong correlation of people who donated blood for a longer time tend to have higher frequency of donating blood.

Line Plot

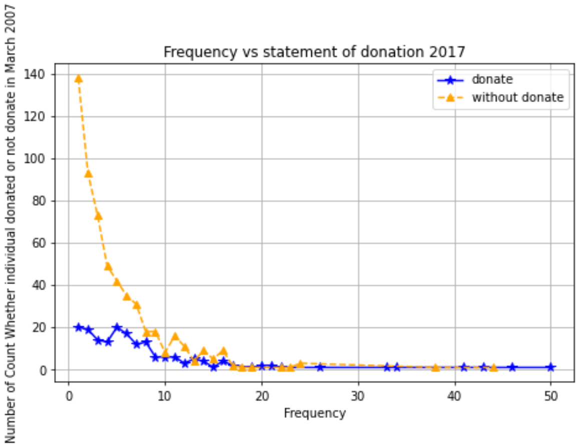


Figure 2.5.0 Frequency vs Donation {1,0} in 2017

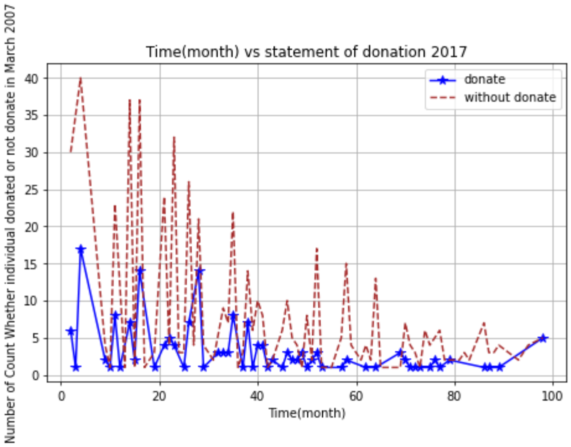


Figure 2.5.1 First Donation vs Donation {1,0} in 2017

From figure 2.5.0, the line graph shown that the data points were highly skewed data and blue line represented as individuals blood donated in 2017 and orange line was those individuals didn't donate their blood in 2017. X-axis represented as frequency and Y-axis was number of counts of representing whether the individuals donated or not donate in 2017. The highest count was 140 of the individuals without donate their blood in 2017 with the frequency n =1 (they did blood donation before target variable y). In comparison, the highest count of the individual with donated in 2017 was 20 and the frequency was n=1. Next, there was no individual donated in 2017 who had donated 50 times before 2017. *From figure 2.5.1*, the graph shown that there was high variance between the data points of individuals. X-axis "Time(month)" represented the duration of donor from first month until 2017 and Y-axis was number of counts of representing whether the individuals donated or not donate in 2017. Data was highly correlated where 18 of the individuals would donate their blood in 2017 with the interval of 4-5 months (first donation). There was 50% of the chances where 5 individuals would or wouldn't donate their blood in 2017 with 100 months of interval compared to first month.

Figure 2.5.2 Total Blood in CC vs Donation {1,0} in 2017

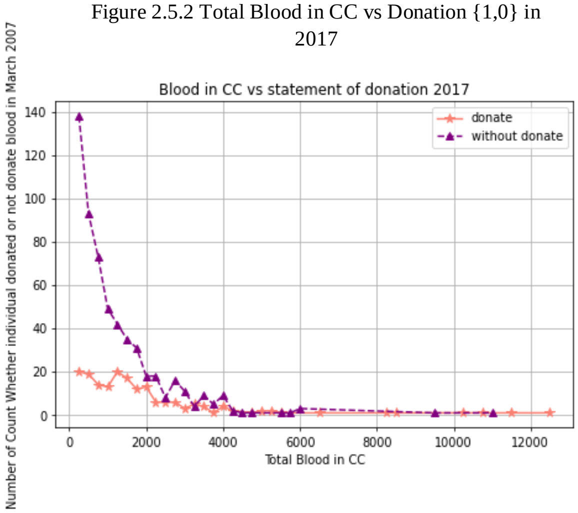
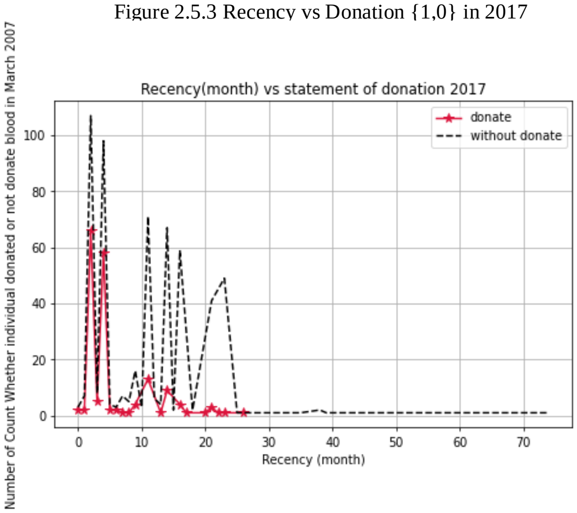


Figure 2.5.3 Recency vs Donation {1,0} in 2017



From figure 2.5.2, the line graph shown that the relation between total blood in cc with number individual whether donate or not donate in year 2017. X-axis represented as total blood in cc and Y-axis was number of counts of representing whether the individuals donated or not donate in 2017. The highest count was 140 of the individuals without donate their blood in 2017 with the recorded 250 cc donated before the target event. In comparison, the highest count of the individual with donated in 2017 was 20 and the donated blood in cc was 250 cc. However, there was no individual donate their blood in 2017 with previous record of 12500 cc in blood. *From figure 2.5.3*, the graph shown that last of the donation in month versus the donor donate or not donate on 2017.There were 70 of the donors would donate their blood in after 3-4 month of the recency time. More than 100 of the individuals wouldn’t donate after 3-4 months of the recency time. The result of donor donated in 2017 were slightly decreasing after 16-17 months of the recency time.

Strip plot

Figure 2.6.0 Total Blood in CC vs Donation {1,0} in 2017

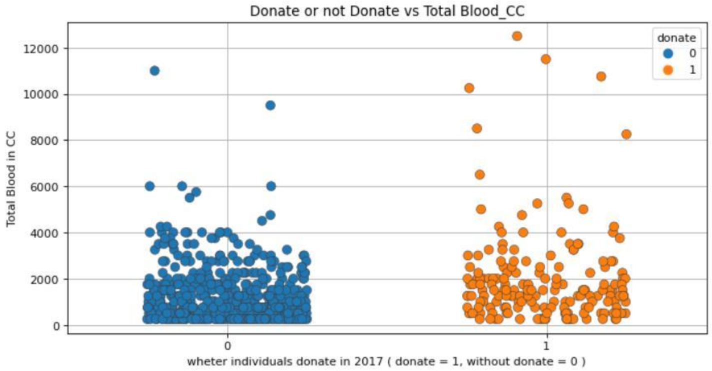
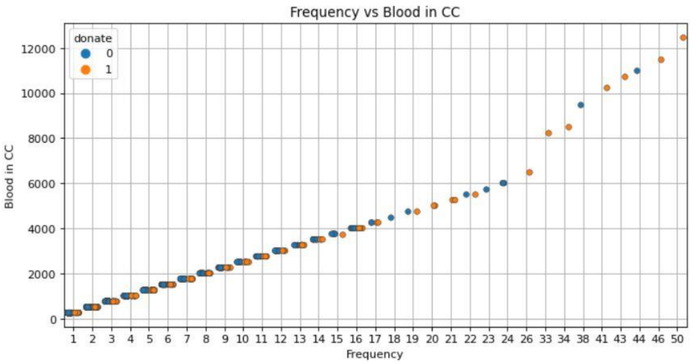


Figure 2.6.1 Rate vs Donation {1,0} in 2017



From figure 2.6.0, the Strip plot shown that the comparison between total blood in cc and whether the individual donate (1) or not donate (0) in 2017. There were two individuals who had donated 9000cc and 1100cc before the year 2017. These two individuals were considered as outliers who did not donate the blood in 2017. At 12500cc, the individual had donated in 2017 but we consider that is outlier for the data set since the data point was high distance error and not correlated to other points. Overall, the number of individuals without donated was more than the number of individuals donated in 2017.

From figure 2.6.1, the graph shown that the total blood in cc versus to the total number of donations with either the individual donates or without donate in 2017. At 12500cc, the individual had donated in 2017 with previous total recorded of 50 times. The lowest observation was frequency equal to 1 and the individual had donated around 250 cc in the previous records. The result shown that there was 50% of the individual either donated or not donated with frequency n =1 in year 2017.

Figure 2.6.2 Time (month) vs Donation {1,0} in 2017

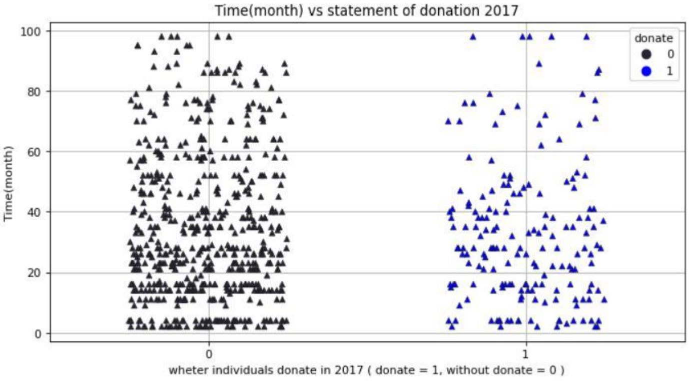
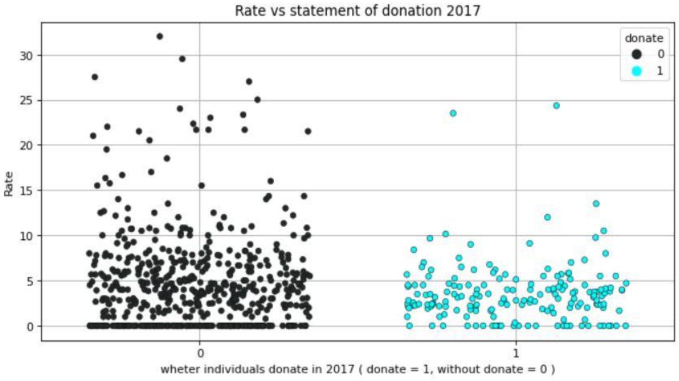


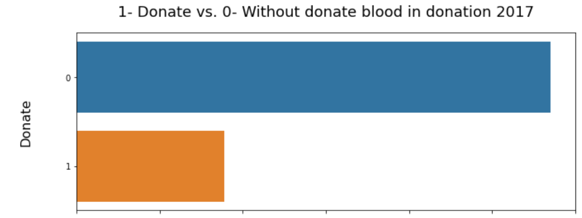
Figure 2.6.3 Rate vs Donation {1,0} in 2017



From *figure 2.6.2*, the graph shown that the relationship between whether the donor donate or not donate in 2017 versus the duration of interval from first donation until 2017 (time in month). Since the distance error were less than the *figure 2.5*.0. we can assume that there was no outlier between the time(month) and event of donate {0,1}. From figure 2.6.3, the graph shown that the rate of chances for the individual donate of not donate in 2017. The rate range were between 0 to 35. There were few outliers lied in rate 23-24 to compare the individual would donate in 2017. However, rate 33 was the outlier of the particular individual wouldn't donate in 2017. Overall, number of individuals who donated in 2017 were less than number of individuals who did not donate in 2017.

Countplot

Figure 2.7.1 – Decision of blood donation in 2017 against frequency of donors



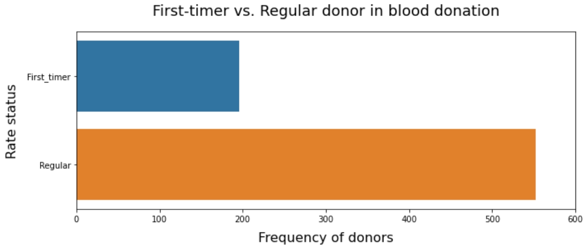


Figure 2.7.2 – Rate status of donors against frequency of donors

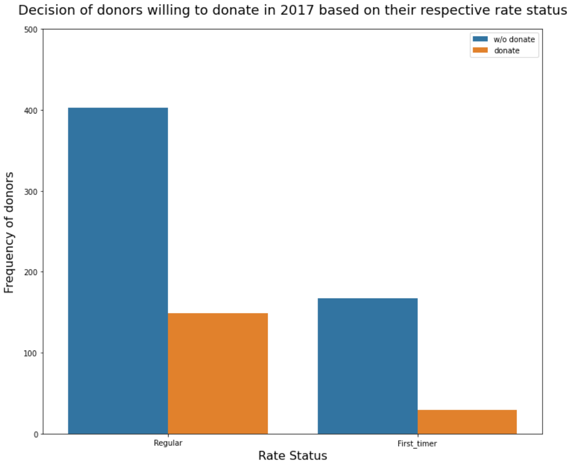


Figure 2.7.3 – Frequency of donors against rate status of donors about the decision of donation in 2017

From *figure 2.7.1*, the countplot is similar to bar chart and it simply means that the donors who donated blood in 2017 consists of 570 donors, while there are 178 donors who have not donated in 2017. People who are not donating in 2017 is about tripled the amount of the people who have donated. From *figure 2.7.2*, the countplot shows that the donors who are often or seldom donating their blood corresponding to the frequency of donors. There are 196 donors who are seldom donating their blood, where they only donated once. However, there are 552 donors who are often donating their blood, where they donated their blood more than twice regardless the interval of period. We see that the majority has at least twice blood donation before rather than only once. From *figure 2.7.3*, the countplot tells us that frequency of donors corresponds to their respective rate status by separating them into group with donated and without donate in 2017. For the donors who are regular donor, it consists of 403 people who have not donated in 2017 and 149 people who donated in 2017. On the other hand, for the donors who are first-timer donor, it consists of 167 people who have not donated in 2017 and 29 people who donated in 2017. When comes to the decision making of blood donation in 2017, we have the percentage of regular donors who are willing to donate their blood is 36.97%, while the percentage of first-timer who willing to donate their blood is only 19.46%. Nevertheless, there is majority who are regular donors decided not to donate in 2017, but there are 29 first-timer donors decided to donate in 2017.

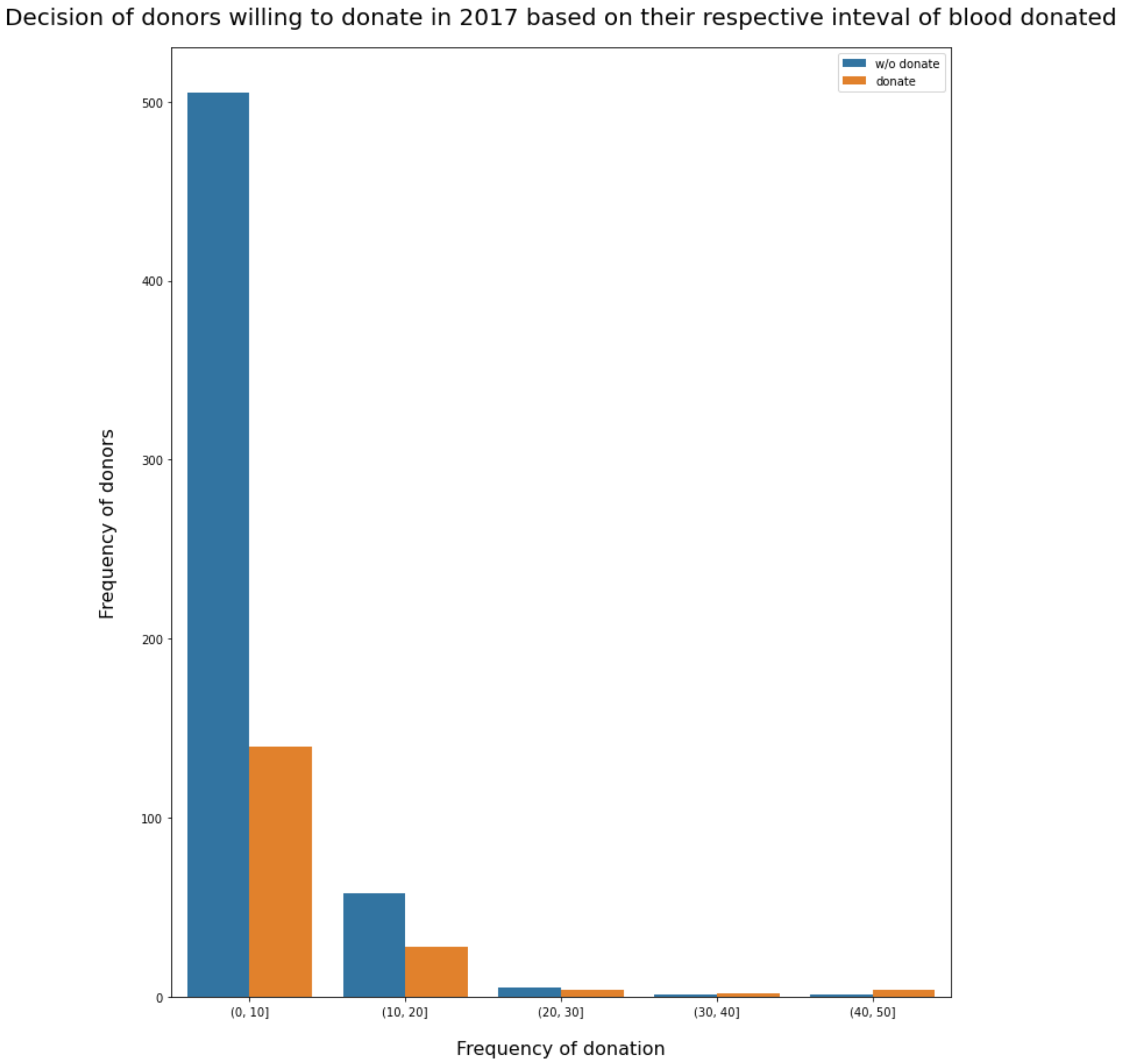


Figure 2.7.4 – Frequency of donors against interval of freqency of donation about the decision of donation in 2017

From *figure 2.7.4*, clearly it is a left skewed graph, majority of the donors have donated not more than 10 times before. Technically, there is 86.23% donors who have donate not more than 10 times in the past, where 78.29% of them decide not to donate in 2017 given they donate not more than 10 times. In each interval of frequency of donation from 0 to 30, the proportion of donors who have not donated in 2017 is more than the proportion of donors who have participated in donation 2017. Besides that, the proportion of donors who donated in 2017 is more than the proportion of donors who have not donated in 2017 for each interval above 30 times of donation. In addition, donors who have more times of blood donation are more likely to donate in 2017.

Heatmap

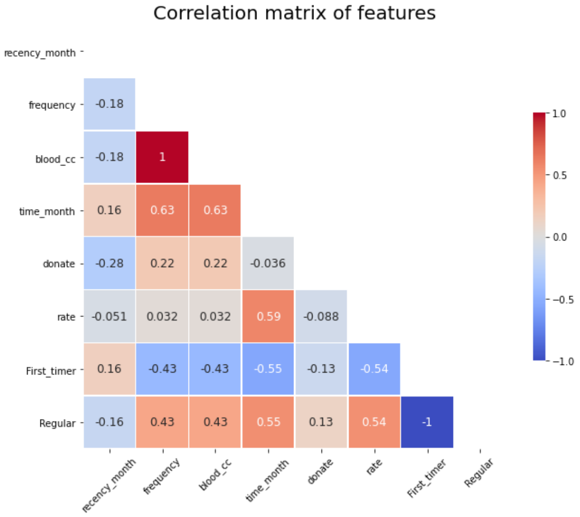


Figure 2.8 – Heatmap of features

From *figure 2.8*, the heatmap shows the correlation matrix about the interdependence of the features. In the label, red color represents the both features is related, while blue color represents the both features has inverse relationship. When the correlation is zero, it means both features has no relationship between each other. By inspection, it is obvious that the "blood_cc" and "frequency" of a donor donates blood is dependent to each other. With the reason of amount of blood they donated to be fixed at each time, which is 250 c.c.. We also see that the correlation of "First-timer" and "Regular" in "rate_status" to be negative related by logic. By giving the attention on the correlation of "rate" and "donate" with value of -0.088. The rate of period takes for a particular donor to donate seems to be slightly negative related or no related to their decision on donation in 2017. For example, the donor has a rate of donating blood once a year has no relationship on the decision of donation in 2017. We also have correlation of "donate" to "recency_month" and "time_month" to be -0.28 and -0.036 respectively. As the blood donation for each person must wait a minimum of 56 days to donate again, so the correlation for "recency_month" is more negative than "time_month" , which is first donation that seems to be less dependent on the decision on blood donation in 2017. For the correlation of "donate" and "frequency" is 0.22, from *figure 2.7.2*, it shows that people who donate more frequently are more likely to donate blood in 2017.