Zewail City of Science and Technology

Machine learning course

Instructor: Mohamed Elshenawy

Omar El-Sakka 201900773

Ahmed Adel 201901464

Abdulla Sabry 201701484

University of Science and Technology January,

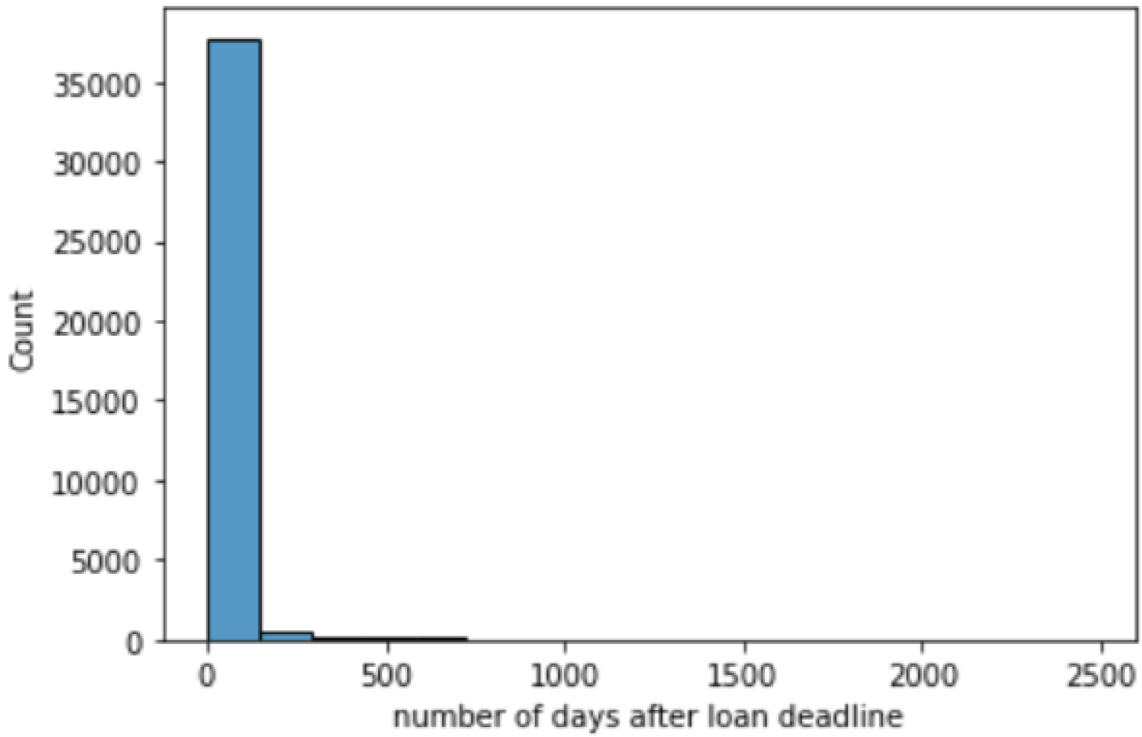
Feature Engineering:

1) Dropping the BOOKING_DATE: because the Entering date of the loan is not important and carries no knowledge about data. As it is completely variable and completely carries no useful data.
2) Dropping the MATURITY_DATE: maturity date is completely useless and it is in the future, and not in the past so it carries no previous knowledge about the client
3) Meanwhile, the same knowledge of both the Booking date and the maturity date is implemented in just a one-column called "Loan term" which carries only the whole time of the loan.
4) Dropping the TENOR_@BOOKING: because it carries the same knowledge of the "Loan Term" but in the number of months form.
5) DOB: if the date of birth is put as it is in the model would be useless but if we replace the whole date of birth with a month "MOB" in which the client was born, it would be somehow

useful. (“MOB” column that created to replace “DOB”)

6) MOB: this column as a number is meaningless, so we discretized the columns into 4 groups(“1 to 3”, “4 to 6”, “7 to 9”, “10 to 12”). For example, “1 to 3” means that the client is born in one of the first three months.(We going to do one-hot encoding to these groups)

7) DPD: most of the entered data is for clients that return the loan at a time. So, it would make data unbalanced.



8) DPD: the number of days that clients miss returning loans in the past. Is useful but not on that scale as we want to make classification with numbers to be more useful. So, we divide the DPD into 6 subgroups. ("Ontime", "within 1month", "within 2 months", "within 3 months", "within 4 months", "More than 4 months")

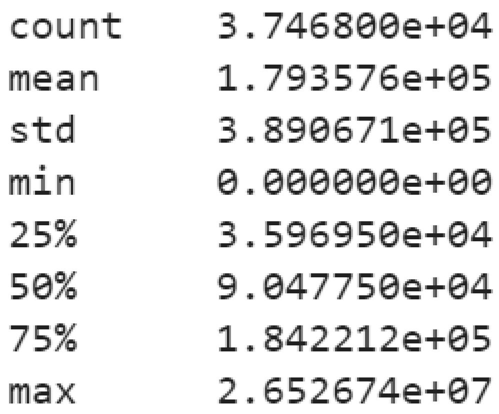
9) AGE: representing the age as a number would be useless and make no sense with the model. So, we replaced that "AGE" columns with a discretized age. For example, if the client has 32 years old, we replaced that with a string "30 to 40". And if the client has 48 years old, it would be "40 to 50". We going to do one-hot encoding to "30 to 40", "40 to 50", etc.

10) AGE AT MATURITY: carry the same knowledge of "AGE" and "Loan Term" so we drop that column.

11) Gender: It has 4 types (Male, MALE, Female, FEMALE), So we correct Each Male to be MALE, and the same for Female. To be just two types.

12) Customer Segment: it would be either salaried, or Self employed and Professional.

13) Total O/S: Total number of loans is also discretized into 4 subgroups(“Less than 36000”, “Less than 90000”, “Less than 184000”, “Less than 27000000”). And the reason beyond these numbers is that these number is approximately the first quartile(3.596950e+04), the second quartile(9.047750e+04), the third quartile(1.842212e+05), and maximum (2.652674e+07)



```
count    3.746800e+04
mean     1.793576e+05
std      3.890671e+05
min      0.000000e+00
25%      3.596950e+04
50%      9.047750e+04
75%      1.842212e+05
max      2.652674e+07
```

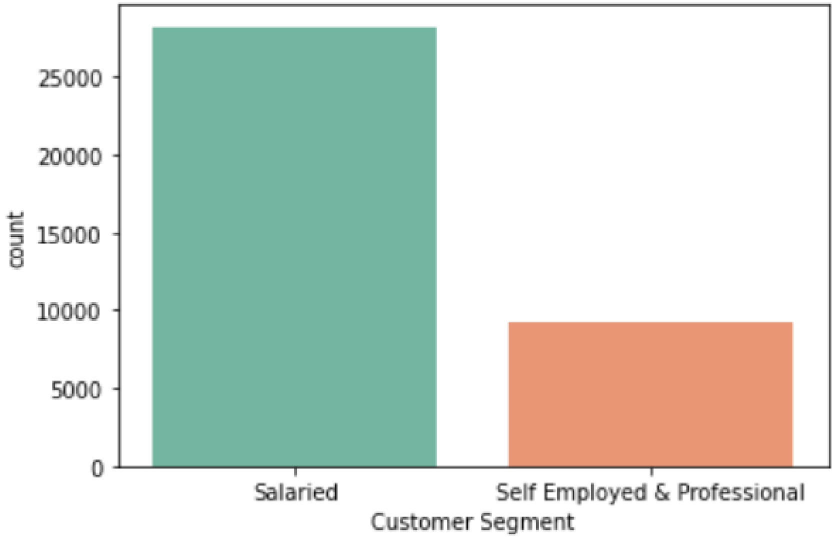
14) Dropping duplicated data

15) Dropping all nulls: it was so many small amounts in the whole data.

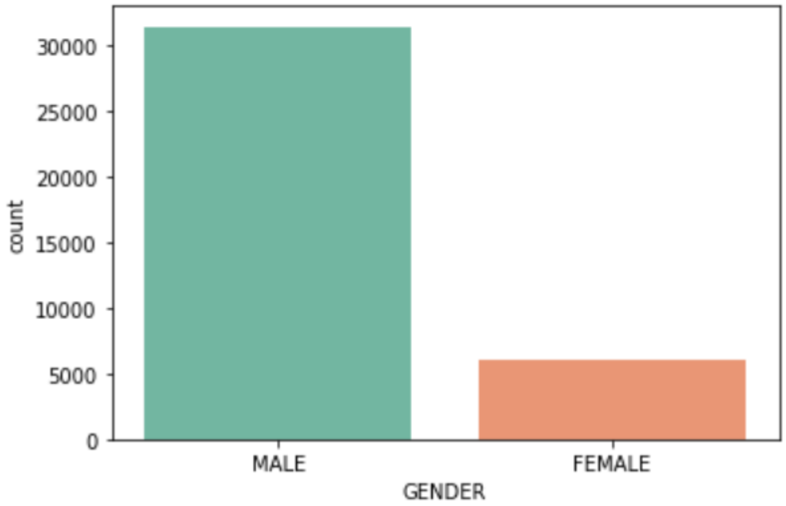
16) If the data was containing the salary of each client, it would affect the results positively.

Exploring Data

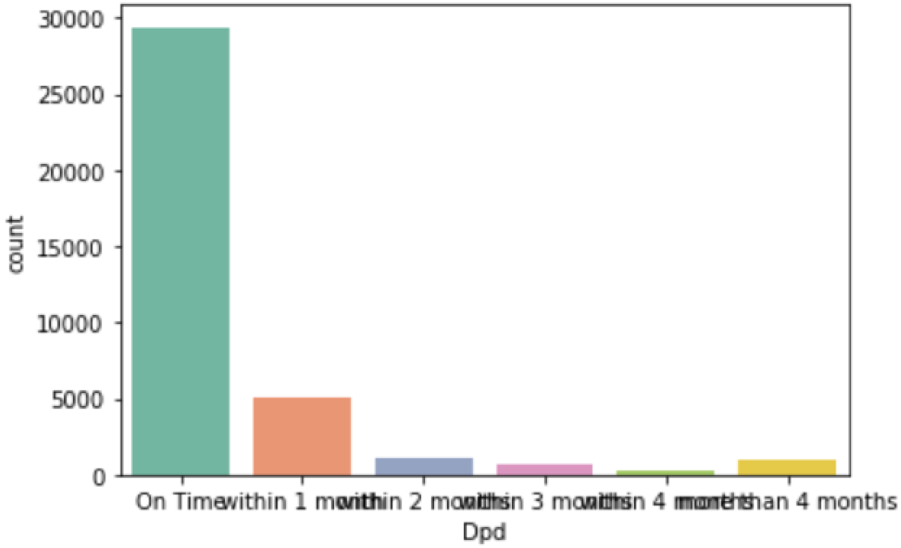
1) Customer Segment: We have about 28228 Salaried and 9240 Self-employed & Professionals. So, We can say that more than 75% of people interested in taking a loan are Salaried.



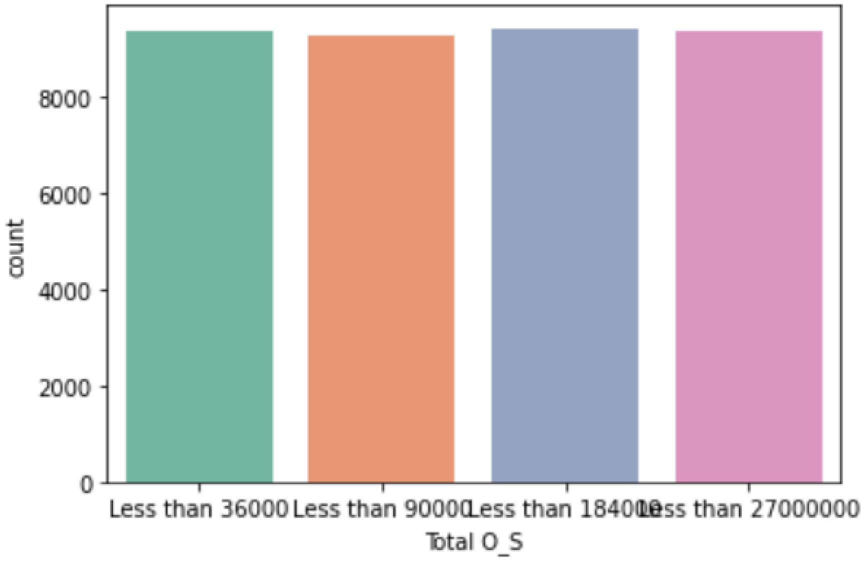
2) Gender: We have 31435 Males and 6033 Females in the dataset. So, More than 83% of loan takers are males.



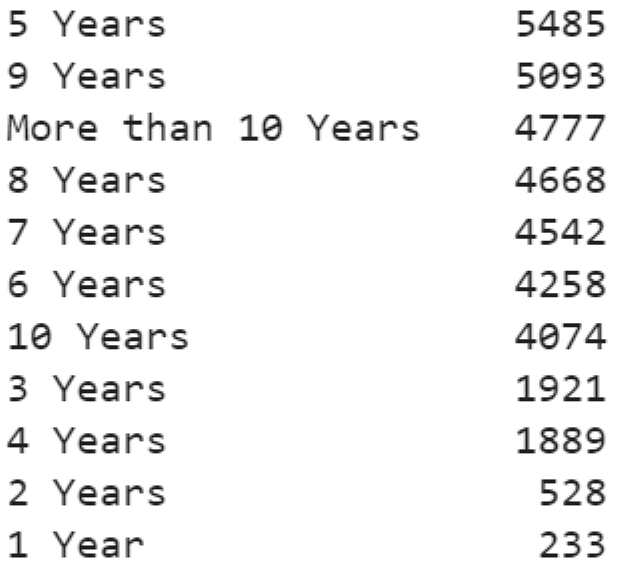
3) DPD: We have in the data set
29387 customers returned the loan on time,
5066 returned the loan within one month,
1053 returned the loan within two months,
667 returned the loan within three months,
325 returned the loan within four months, and
finally 970 retuned the loan after four months.
So, more than 78% of loan takers return the loan on time.



4) Total O_S: we have 9384 clients took a loan less than 36000, 9269 clients toke a loan less than 90000, 9437 clients took a loan less than 184000, and 9378 clients took a loan less than 27000000. And this is logical somehow because that we divide the "Total O_S" at its quatriles.

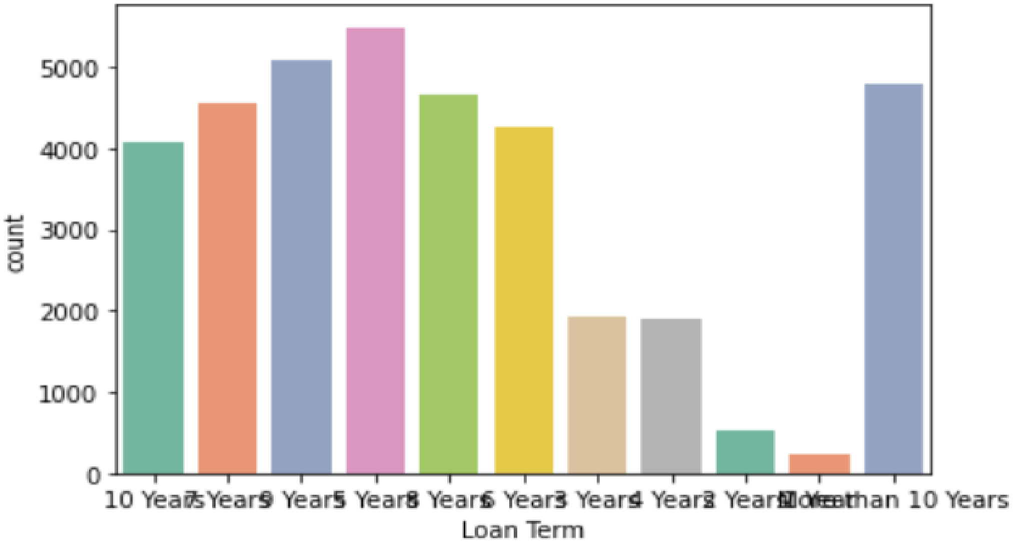


5) Loan Term: we have those numbers

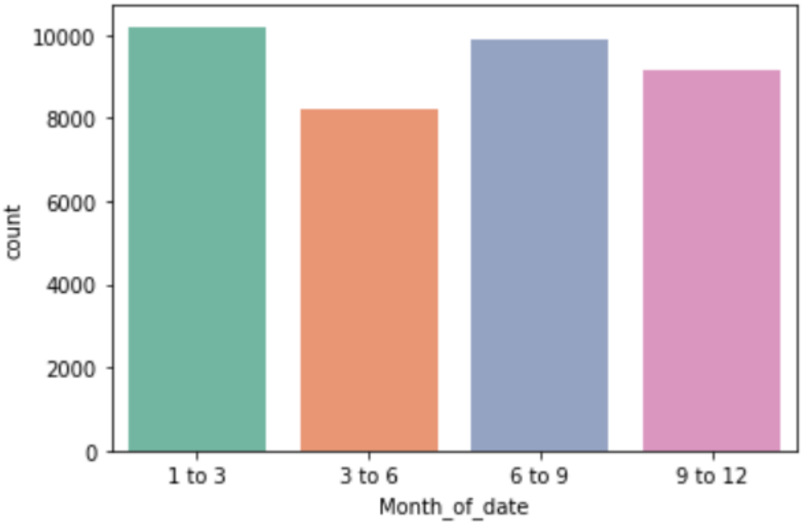


```
5 Years                 5485
9 Years                 5093
More than 10 Years      4777
8 Years                 4668
7 Years                 4542
6 Years                 4258
10 Years                4074
3 Years                 1921
4 Years                 1889
2 Years                  528
1 Year                   233
```

So, we can notice that most loans were for not less than 4 years.



6) Month of date: we have 10199 were born in the first three months(Jan-Mar), 9893 were born in the second three months(Apr-Jun), 9162 were born in the third three months(Jul-Nov), and finally 8214 were born in last three months(Oct-Dec).



7) Age: the is normally distributed

Number of People with age (10-20): 117

Number of People with age (20-30): 1773

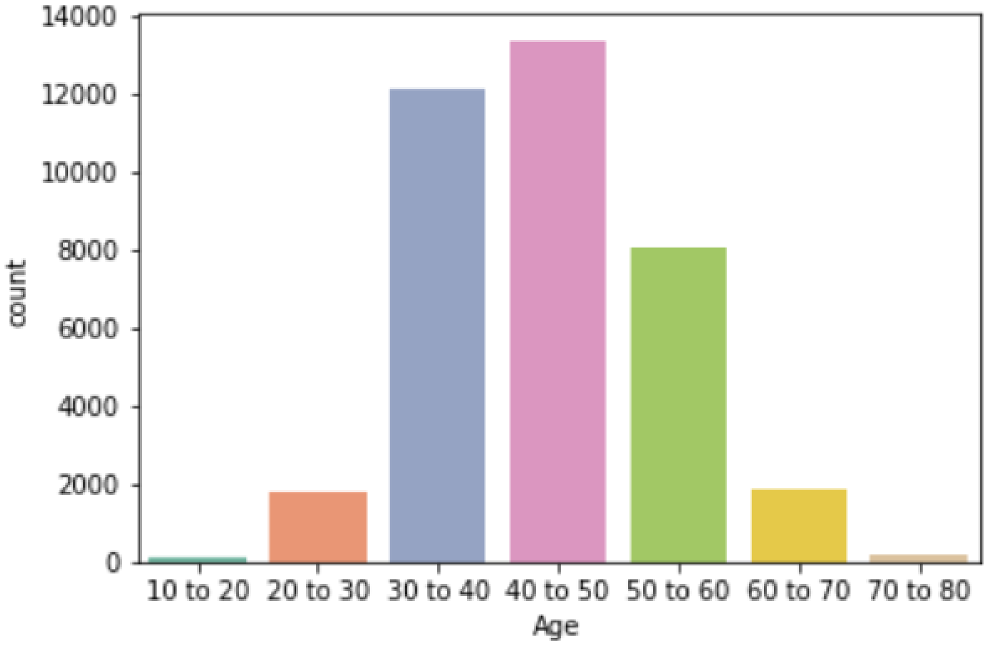
Number of people with age (30-40): 12131

Number of people with age (40-50): 13354

Number of people with age (50-60): 8069

Number of people with age (60 -70): 1870

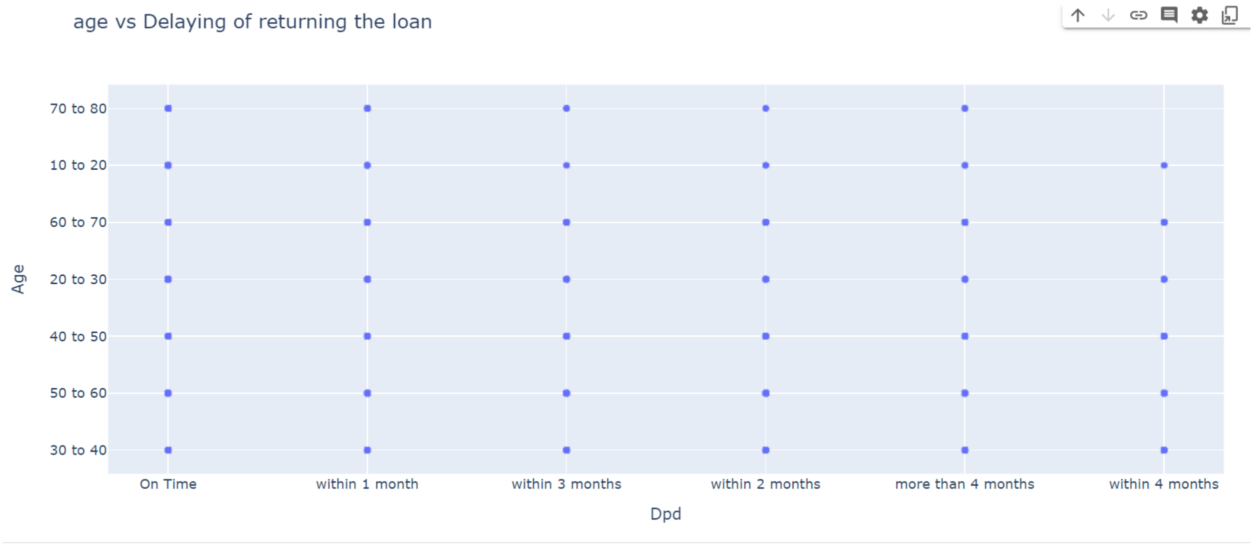
Number of people with age (70-80): 154



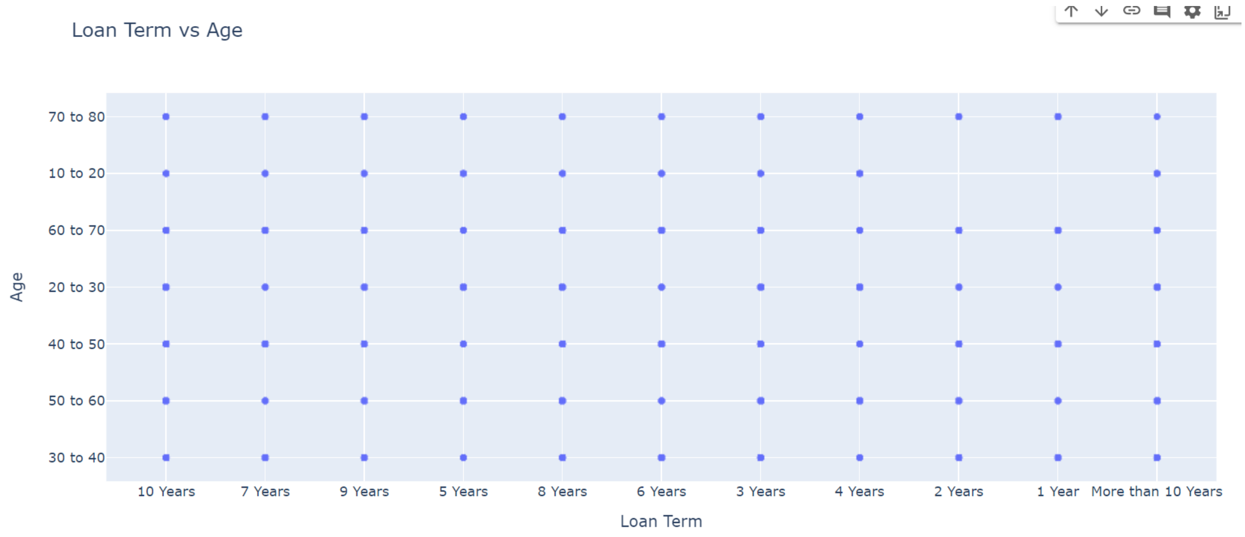
So, we can notice that the largest segment is the people with age from 40 to 50. And the data is normally distributed around that age.

Now, we studied the features all alone. Let’s study them together to get different relationships between columns.

1) A Scatter plot between age and the period of days of delaying the loan return is not so important. But we can notice that there is no case of delaying 3-4 months for clients with ages between 70 - 80 years old.



2) A scatter plot between age and Dpd shows that there is no case for a client with age between 10 to 20 years old that intends to take a loan for 1 or 2 years.

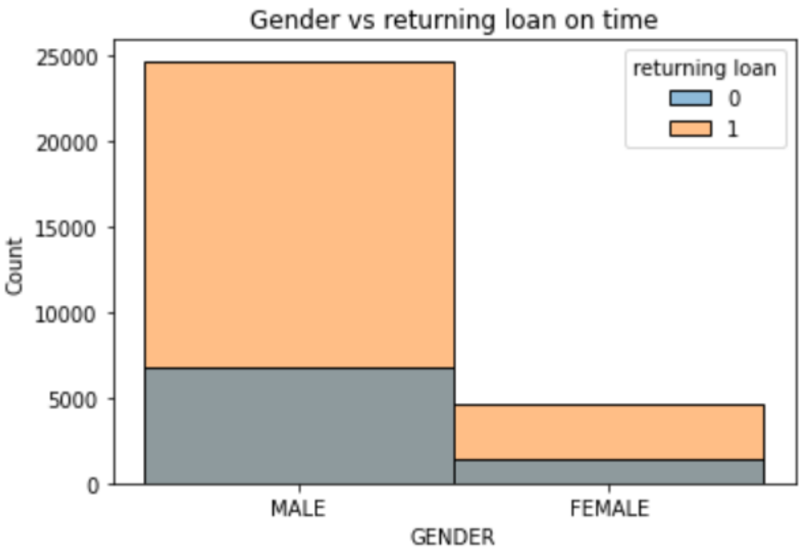


We did other scatter plots but it contains no information.

1) <u>Studying the relationships between gender and returning the loan period.</u>

1) Curve with Gender vs returning loan on

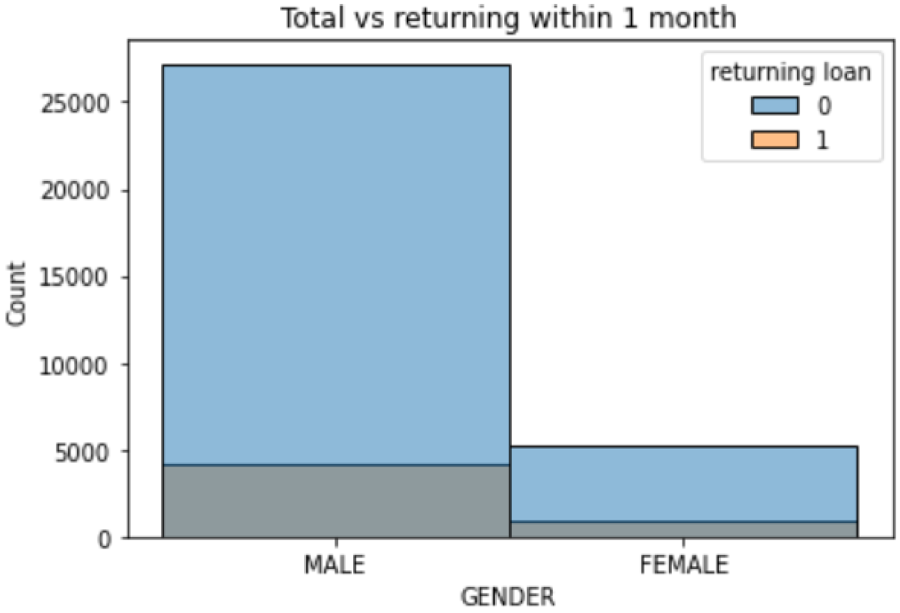
time



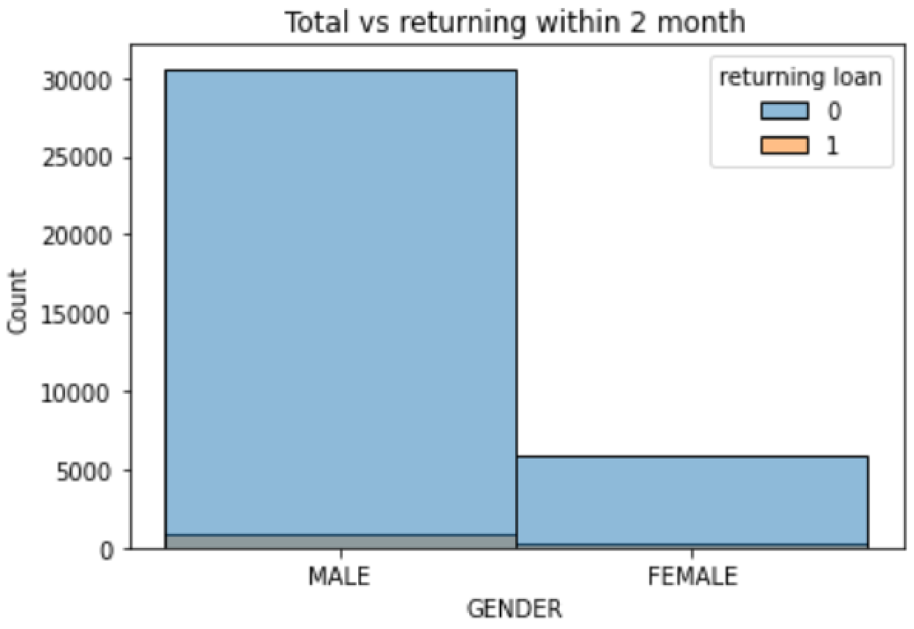
We have 24709 males returning loans on time, and 4678 females. Meanwhile, all males are31435 and all females are 6033. Then more than 78% from all males and more than 77.5% from all females returned the loan on time.

2) Curve with Gender vs returning loan within one month late

We have 4230 males returning loans within one month, and 836. Meanwhile, all males are31435 and all females are 6033. Then more than 13.4% from all males and more than 13.8% from all females returned the loan within one month.

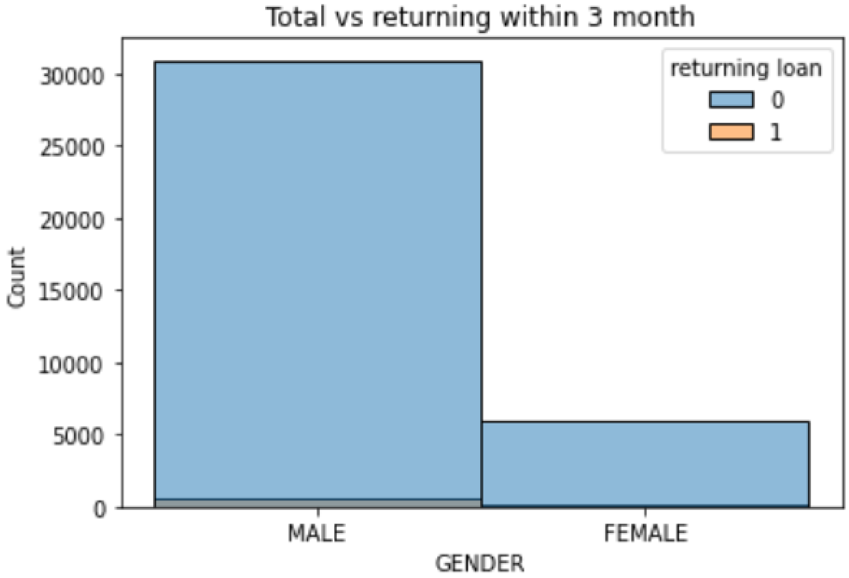


3) Curve with Gender vs returning loan within two months late



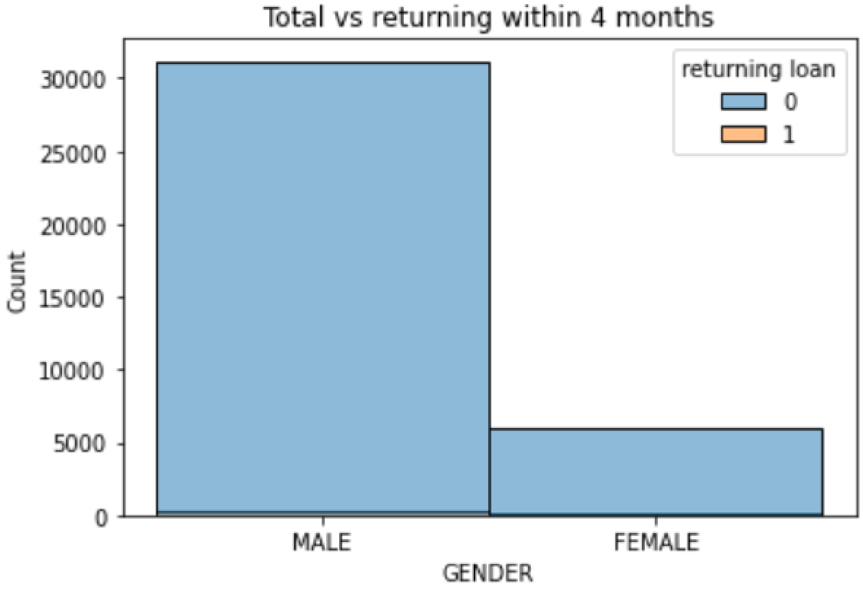
We have 840 males returning loans within two months, and 213. Meanwhile, all males are31435 and all females are 6033. Then more than 2.6% from all males and more than 3.5% from all females returned the loan within two months.

4) Curve with Gender vs returning loan within three months late



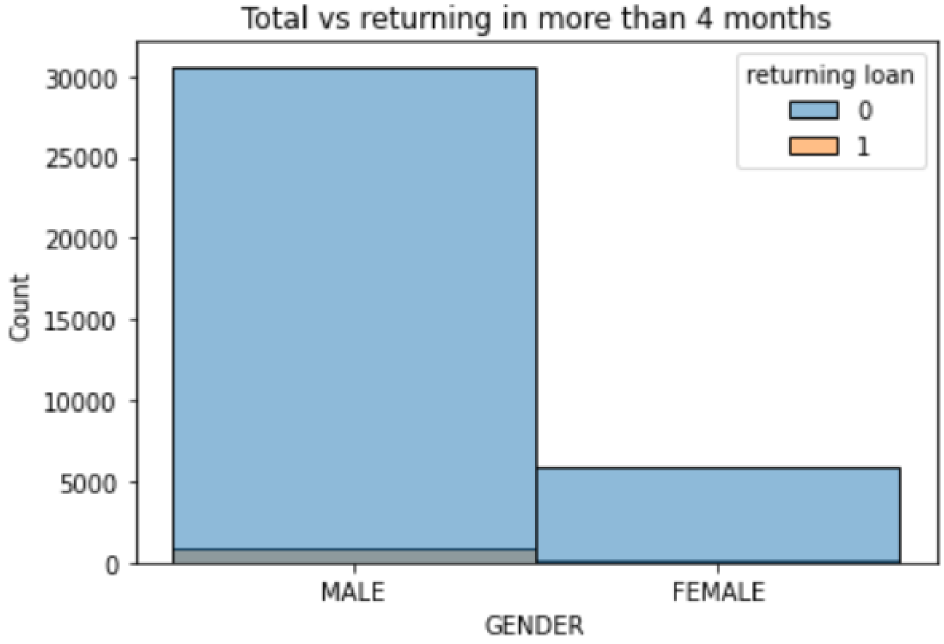
We have 538males returning loans within three months, and 129. Meanwhile, all males are31435 and all females are 6033. Then more than 1.7% from all males and more than 2.1% from all females returned the loan within three months.

5) Curve with Gender vs returning loan within four months late



We have 266 males returning loans within four months, and 59. Meanwhile, all males are31435 and all females are 6033. Then more than 0.8% from all males and more than 0.97% from all females returned the loan within four months.

6) Curve with Gender vs returning loan in more than four months late



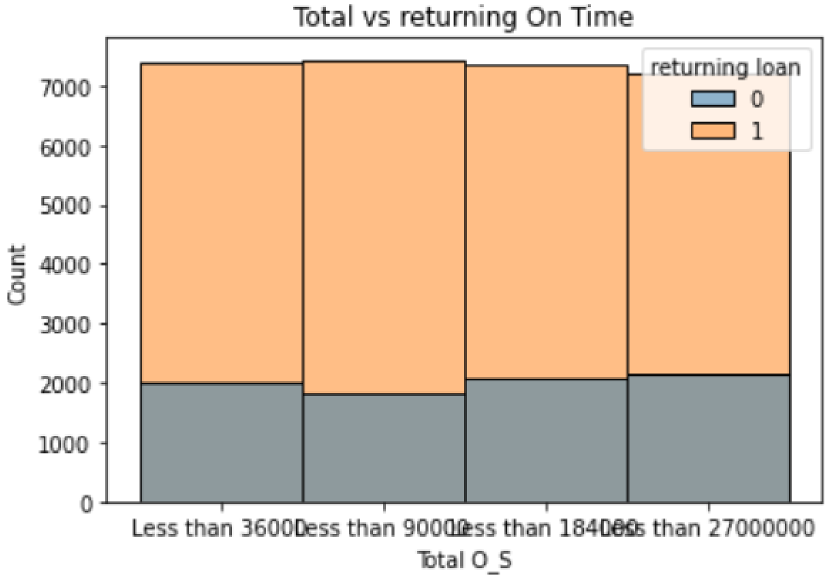
We have 852 males returning loans in more than four months, and 118. Meanwhile, all males are31435 and all females are 6033. Then more than 2.7% from all males and more than 1.95% from all females returned the loan in more than four months.

So, to summarize all of that this is a table that contains the probability that the clients are belonging to a certain gender and return the loan after the due with some delays.

	Males	Females
On-Time	78%	77.5%
One Month	13.4%	13.8%
Two Months	2.6%	3.5%
Three Months	1.7%	2.1%
Four Months	0.8%	0.97%
More than Four Months	2.7%	1.95%

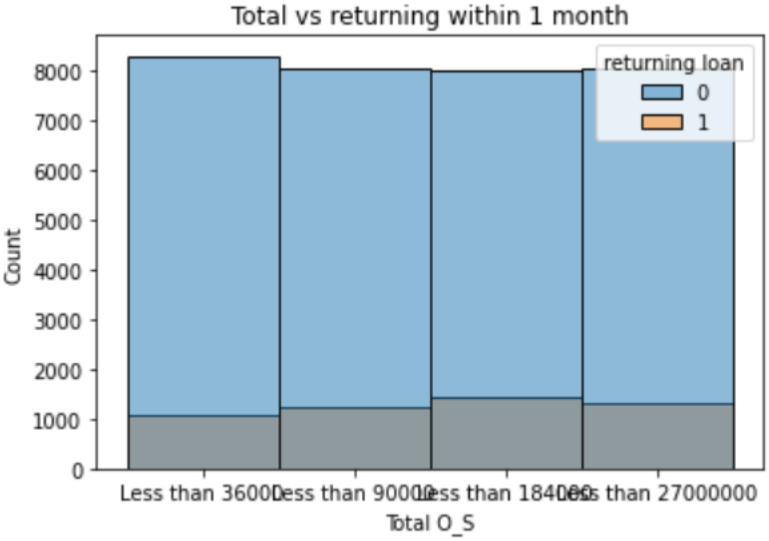
2) Studying the relationships between Total O-S and returning the loan period.

1) Curve with Total O-S vs returning loan on time



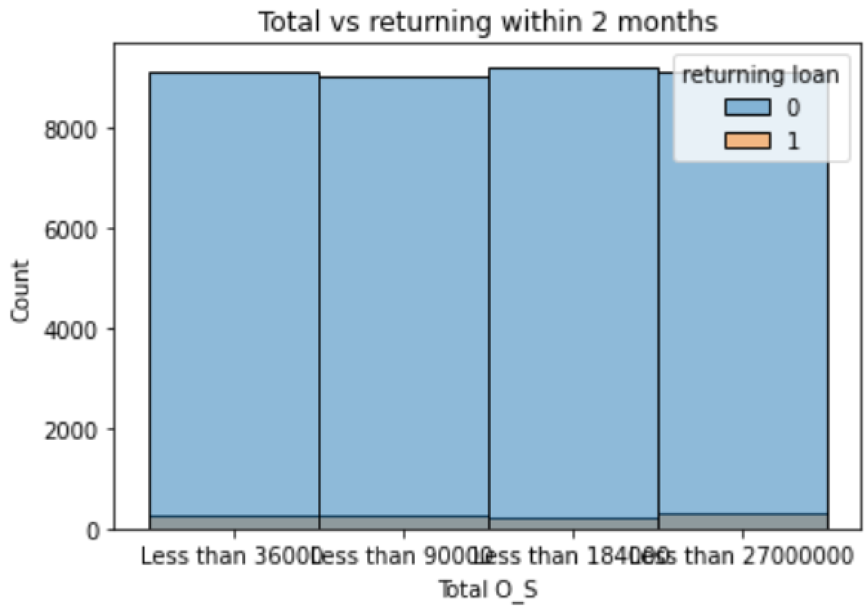
We have 7388 from all clients who return the loan on time were taking a loan with less than 36000, 7436 from all clients who return the loan on time were taking a loan with less than 90000, and 7350 from all clients who returned the loan on time were taking a loan with less than 184000, and finally, 7213 from all clients who return the loan on time were taking a loan with less than 27000000. Almost the same ratio and that is logical as it is

2) Curve with Total O-S vs returning loan within one month late



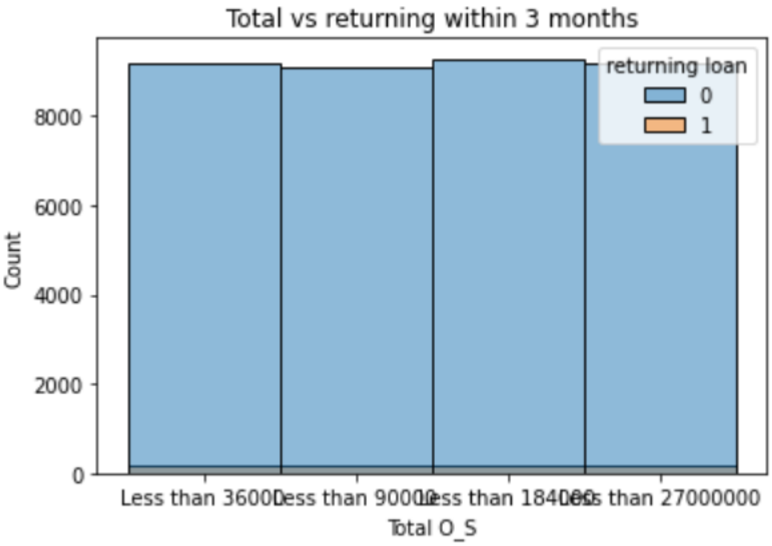
We have 1095 from all clients who return the loan with one month were taking a loan with less than 36000, 1123 from all clients who return the loan with one month were taking a loan with less than 90000, and 1431 from all clients who returned the loan with one month were taking a loan with less than 184000, and finally, 1317from all clients who return the loan with one month were taking a loan with less than 27000000.

3) Curve with Total O-S vs returning loan within two months late



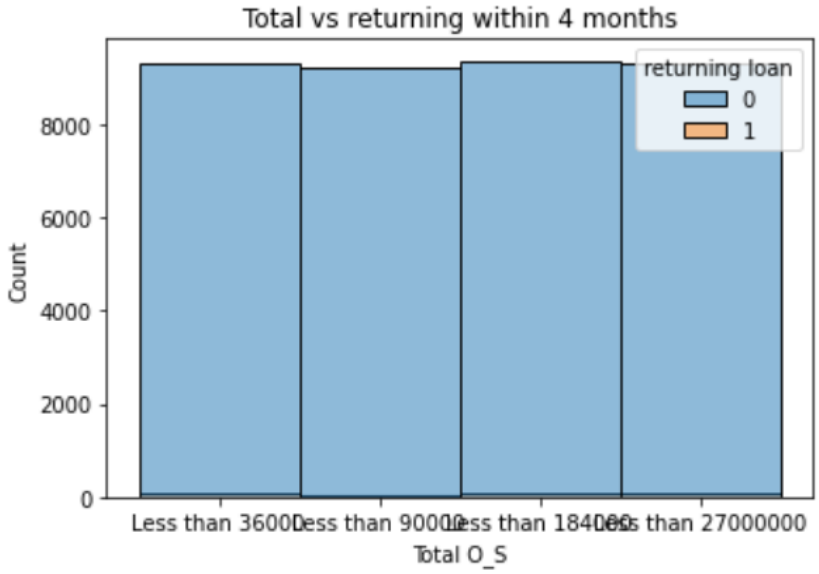
We have 281 from all clients who returned the loan with two months were taking a loan with less than 36000, 245 from all clients who returned the loan with two months were taking a loan with less than 90000, and 239 from all clients who returned the loan with two months were taking a loan with less than 184000, and finally, 288 from all clients who returned the loan with two months were taking a loan with less than 27000000.

4) Curve with Total O-S vs returning loan within three months late



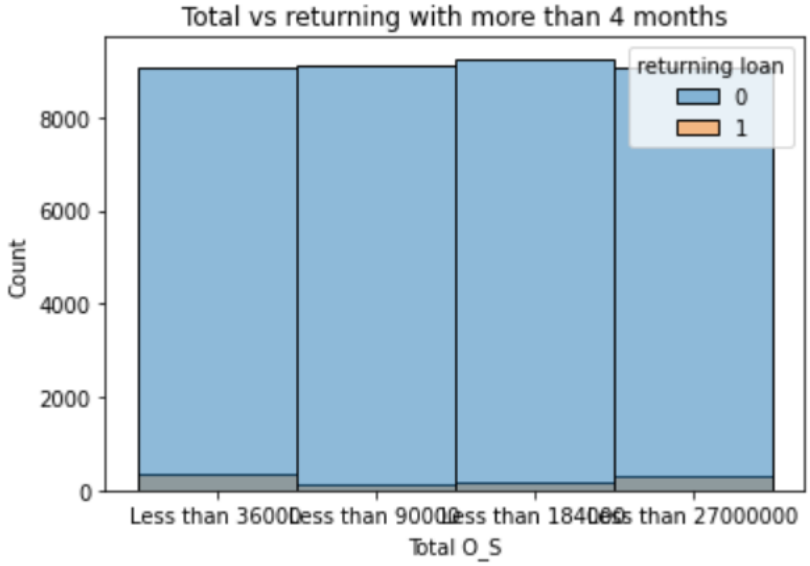
We have 181 from all clients who returned the loan with three months were taking a loan with less than 36000, 162 from all clients who returned the loan with three months were taking a loan with less than 90000, and 153 from all clients who returned the loan with three months were taking a loan with less than 184000, and finally, 171 from all clients who returned the loan with three months were taking a loan with less than 27000000.

5) Curve with Total O-S vs returning loan within four months late



We have 102 from all clients who returned the loan with four months were taking a loan with less than 36000, 53 from all clients who returned the loan with four months were taking a loan with less than 90000, and 82 from all clients who returned the loan with four months were taking a loan with less than 184000, and finally, 88 from all clients who returned the loan with four months were taking a loan with less than 27000000.

6) Curve with Total O-S vs returning loan in more than four months late



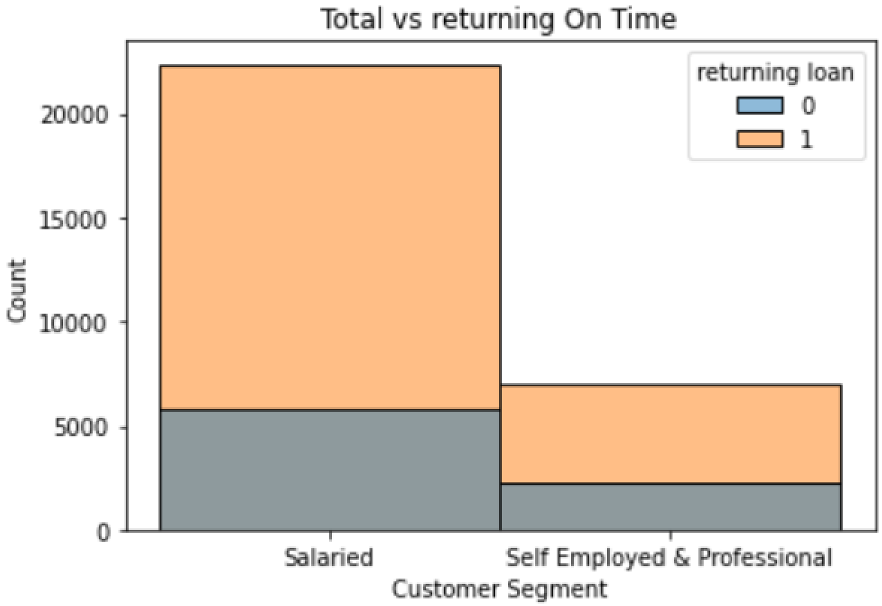
We have 337 from all clients who returned the loan with more than four months were taking a loan with less than 36000, 150 from all clients who returned the loan with more than four months were taking a loan with less than 90000, and 182 from all clients who returned the loan with more than four months were taking a loan with less than 184000, and finally, 301 from all clients who returned the loan with more than four months were taking a loan with less than 27000000.

So, to summarize all of that this is a table that contains the probability that the clients would a loan with a certain amount and return the loan after the due with some delays.

	<36000	**<90000**	**<184000**	**<27000000**
On-Time	78.72%	80.22%	77.88%	76.91%
One Month	11.67%	12.11%	15.16%	14.04%
Two Months	2.99%	2.64%	2.53%	3.07%
Three Months	1.92%	1.74%	1.62%	1.82%
Four Months	1.08%	0.57%	0.86%	0.93%
More than Four Months	3.59%	1.61%	1.92%	3.2%

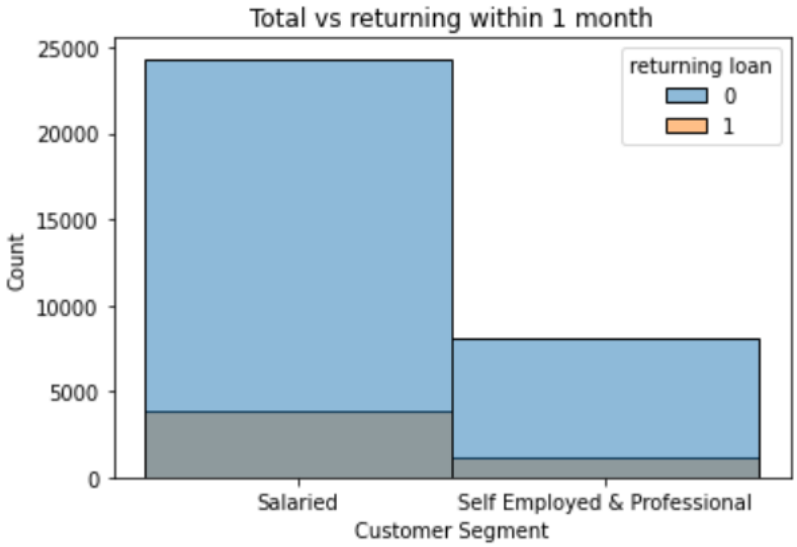
3) Studying the relationships between Customer Segment and returning the loan period.

1) Curve with Customer Segment vs returning loan On Time



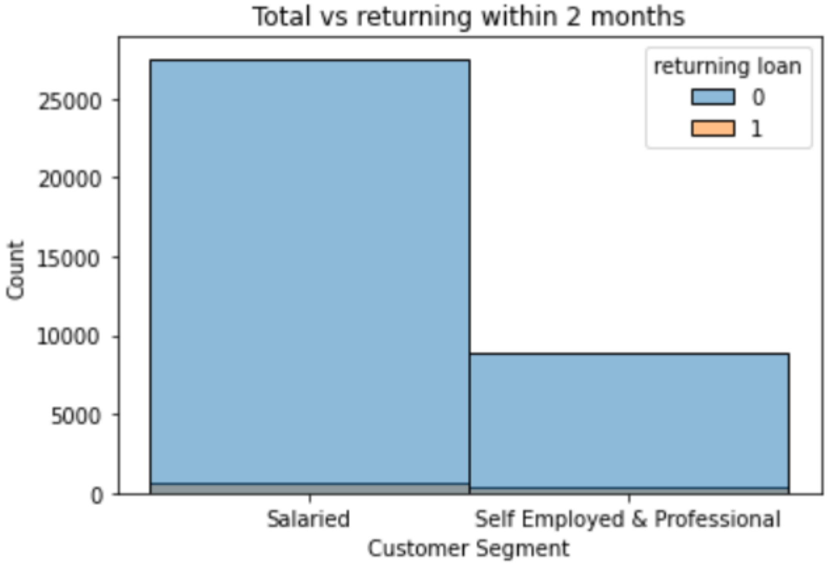
We have 22368 Salaried clients who returned the loan with no late out of all 28228 Salaried people, and 7019 Self Employed & Professionals out of all 9240 Self Employed & Professionals. So, we have 75.96% from all Salaried clients return the loan with no late, and 79.24% from all Self Employed & Professional clients return the loan with no late.

2) Curve with Customer Segment vs returning loan within one month



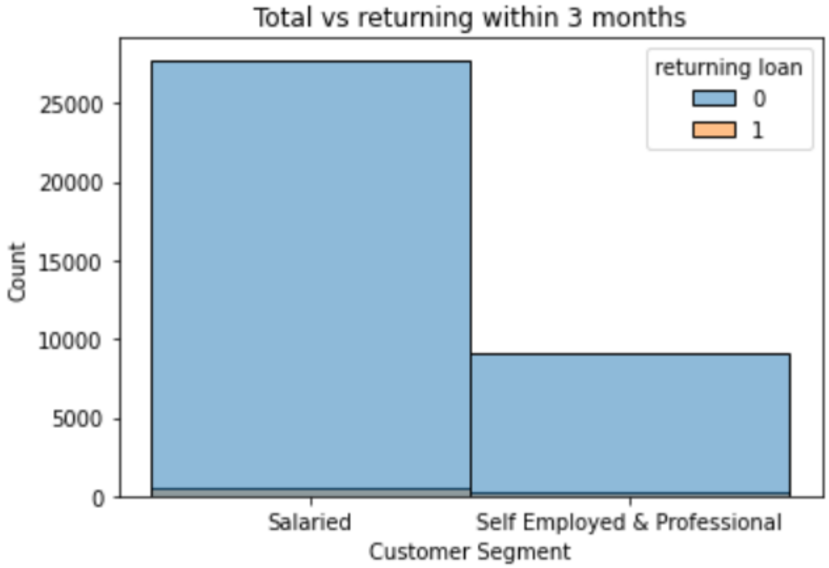
We have 3898 salaried clients who returned the loan within one month late out of all 28228 Salaried people, and 1168 Self Employed & Professionals out of all 9240 Self Employed & Professionals. So, we have 13.80% from all Salaried clients return the loan within one month late, and 12.64% from all Self Employed & Professional clients return the loan within one month late.

3) Curve with Customer Segment vs returning loan within two months



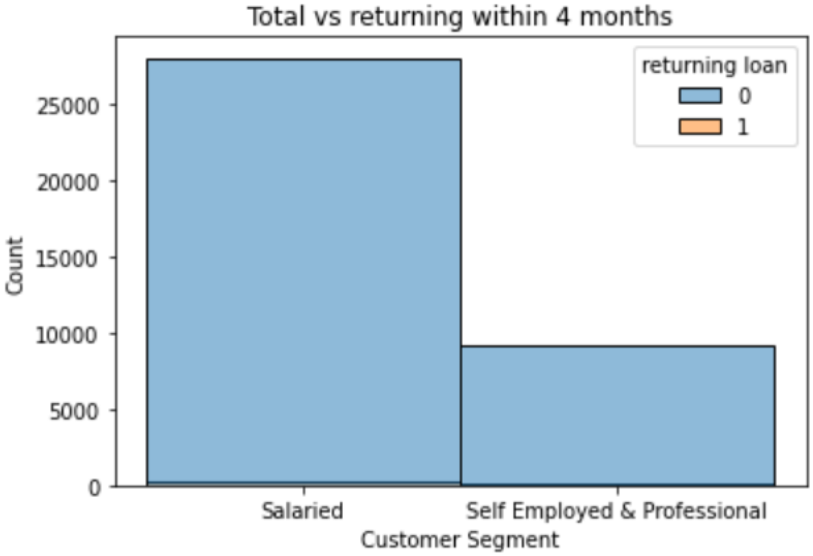
We have 672 salaried clients who returned the loan within two months late out of all 28228 Salaried people, and 381 Self Employed & Professionals out of all 9240 Self Employed & Professionals. So, we have 2.38% from all Salaried clients return the loan within two months late, and 4.12% from all Self Employed & Professional clients return the loan within two months late.

4) Curve with Customer Segment vs returning loan within three months



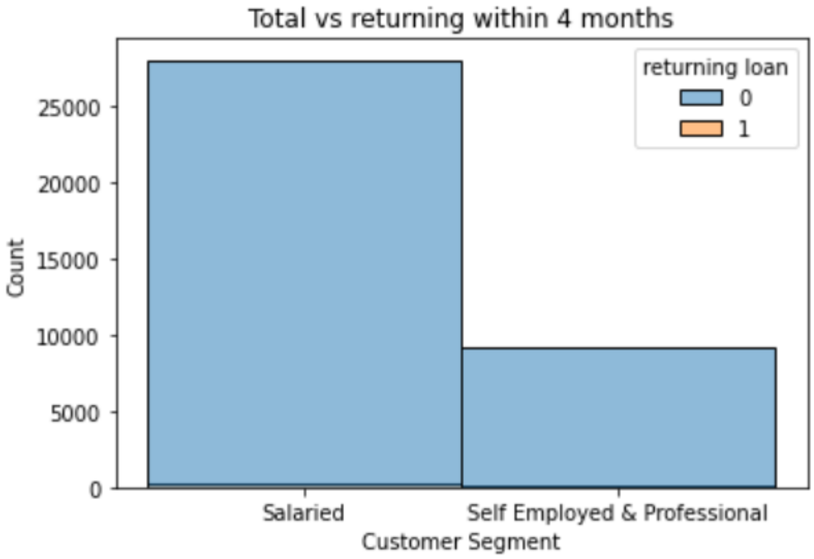
We have 458 salaried clients who returned the loan within three months late out of all 28228 Salaried people, and 209 Self Employed & Professionals out of all 9240 Self Employed & Professionals. So, we have 1.62% from all Salaried clients return the loan within three months late, and 2.26% from all Self Employed & Professional clients return the loan within three months late.

5) Curve with Customer Segment vs returning loan within four months



We have 219 salaried clients who returned the loan within four months late out of all 28228 Salaried people, and 106 Self Employed & Professionals out of all 9240 Self Employed & Professionals. So, we have 0.77% from all Salaried clients return the loan within four months late, and 1.14% from all Self Employed & Professional clients return the loan within four months late.

6) Curve with Customer Segment vs returning loan with more than four months



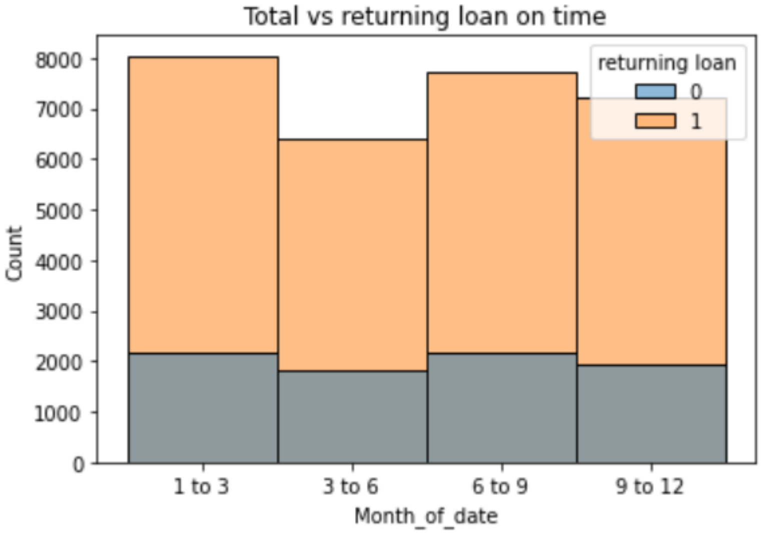
We have 613 salaried clients who returned the loan more than four months late out of all 28228 Salaried people, and 357 Self Employed & Professionals out of all 9240 Self Employed & Professionals. So, we have 2.17% from all Salaried clients return the loan more than four months late, and 3.86% from all Self Employed & Professional clients return the loan more than four months late.

So, to summarize all of that this is a table that contains the probability that the customer segment of clients and return the loan after the due with some delays.

	Salaried	**Self Employed & Professional**
On-Time	75.96%	79.24%
One Month	13.80%	12.64%
Two Months	2.38%	4.12%
Three Months	1.62%	2.26%
Four Months	0.77%	1.14%
More than Four Months	2.17%	3.86%

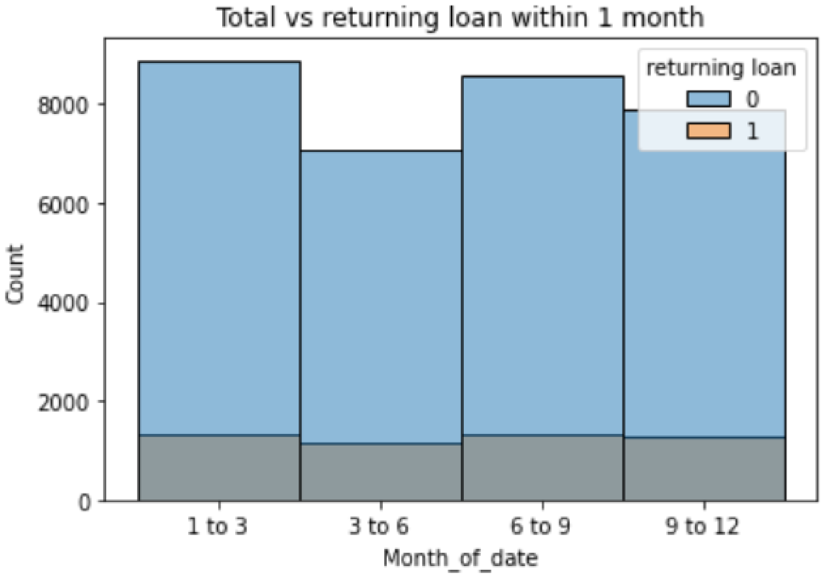
4) Studying the relationships between the Month of date of clients and returning the loan period.

1) Curve with Month of date of Clients vs returning loan On Time



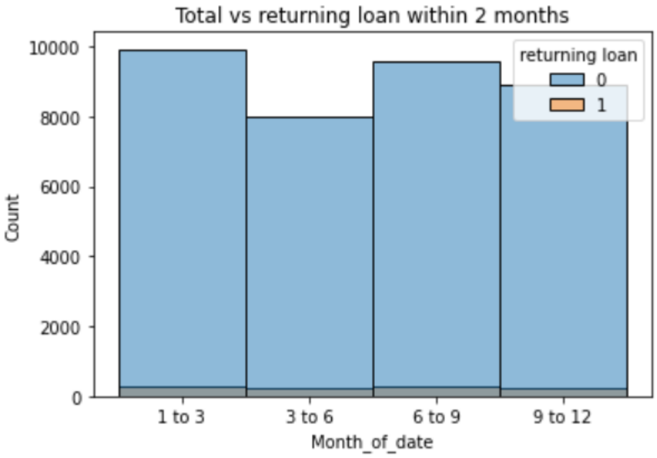
We have 8043 clients who were born in the first three months out of 10199 who were born in the first three months representing 78.86% from all who were born in (Jan-Mar), 6405 clients who were born in the second three months out of 8214 who were born in the second three months representing 77.97% from all who were born in (Apr-Jun), 7731 clients who were born in the third three months out of 9893 who were born in the third three months representing 78.14% from all who were born in (Jul-Nov),7208 clients who were born in the final three months out of 9162 who were born in the final three months representing 78.67% from all who were born in (Oct-Dec).

2) Curve with Month of date of Clients vs returning loan with one month late



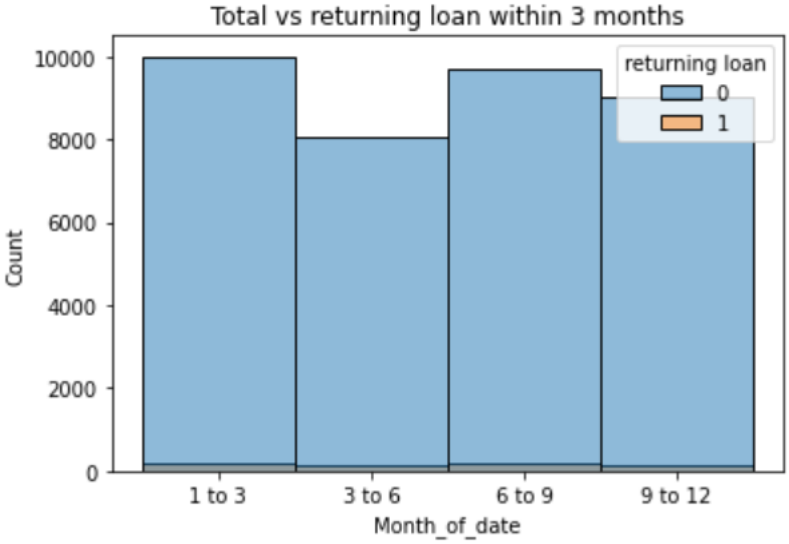
We have 1323 clients who were born in the first three months out of 10199 who were born in the first three months representing 12.97% from all who were born in (Jan-Mar), 1156 clients who were born in the second three months out of 8214 who were born in the second three months representing 14.07% from all who were born in (Apr-Jun), 1326 clients who were born in the third three months out of 9893 who were born in the third three months representing 13.40% from all who were born in (Jul-Nov),1261 clients who were born in the final three months out of 9162 who were born in the final three months representing 13.76% from all who were born in (Oct-Dec).

3) Curve with Month of date of Clients vs returning loan with two months late



We have 275 clients who were born in the first three months out of 10199 who were born in the first three months representing 2.69% from all who were born in (Jan-Mar), 230 clients who were born in the second three months out of 8214 who were born in the second three months representing 2.80% from all who were born in (Apr-Jun), 294 clients who were born in the third three months out of 9893 who were born in the third three months representing 2.97% from all who were born in (Jul-Nov),254 clients who were born in the final three months out of 9162 who were born in the final three months representing 2.77% from all who were born in (Oct-Dec).

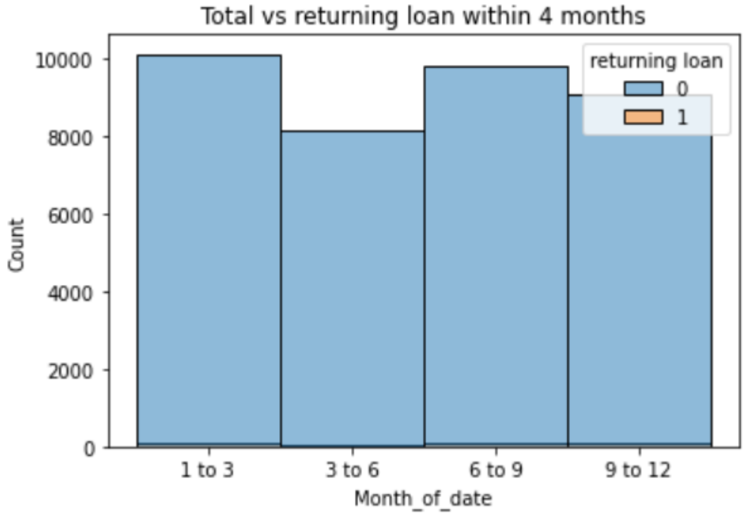
4) Curve with Month of date of Clients vs returning loan with three months late



We have 197 clients who were born in the first three months out of 10199 who were born in the first three months representing 1.93% from all who were born in (Jan-Mar), 155 clients who were born in the second three months out of 8214 who were born in the second three months representing 1.88% from all who were born in

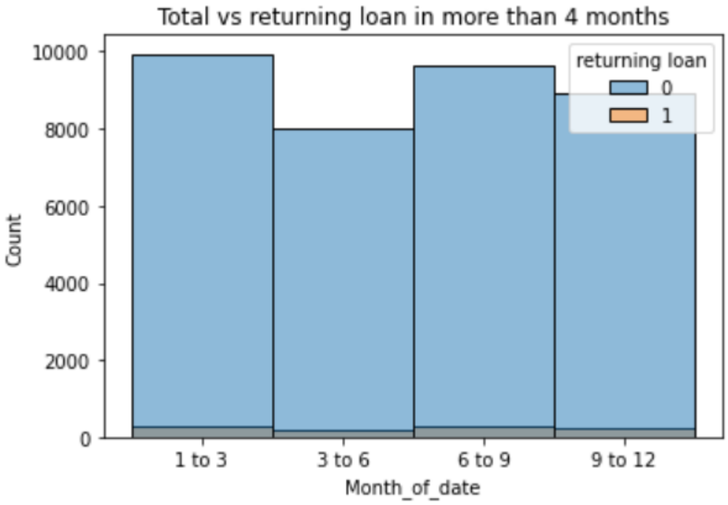
(Apr-Jun), 177 clients who were born in the third three months out of 9893 who were born in the third three months representing 1.78% from all who were born in (Jul-Nov),138 clients who were born in the final three months out of 9162 who were born in the final three months representing 1.50% from all who were born in (Oct-Dec).

5) Curve with Month of date of Clients vs returning loan with four months late



We have 96 clients who were born in the first three months out of 10199 who were born in the first three months representing 0.94% from all who were born in (Jan-Mar), 63 clients who were born in the second three months out of 8214 who were born in the second three months representing 0.76% from all who were born in (Apr-Jun), 92 clients who were born in the third three months out of 9893 who were born in the third three months representing 0.92% from all who were born in (Jul-Nov),74 clients who were born in the final three months out of 9162 who were born in the final three months representing 0.80% from all who were born in (Oct-Dec).

6) Curve with Month of date of Clients vs returning loan in more than four months late



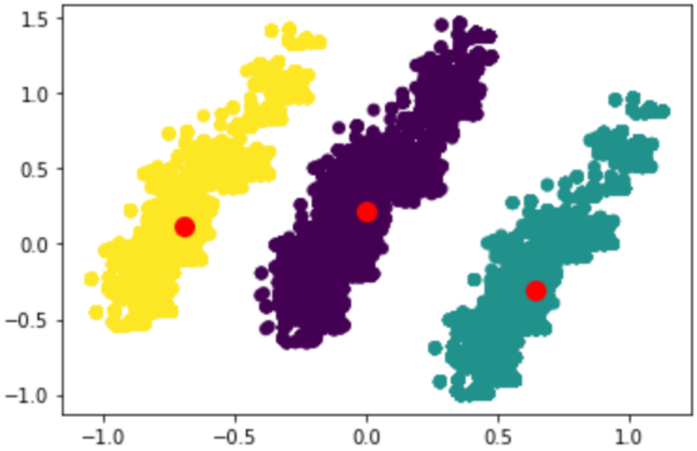
We have 265 clients who were born in the first three months out of 10199 who were born in the first three months representing 2.59% from all who were born in (Jan-Mar), 205 clients who were born in the second three months out of 8214 who were born in the second three months representing 2.49% from all who were born in (Apr-Jun), 273 clients who were born in the third three months out of 9893 who were born in the third three months representing 2.75% from all who were born in (Jul-Nov), 227 clients who were born in the final three months out of 9162 who were born in the final three months representing 2.47% from all who were born in (Oct-Dec).

So, to summarize all of that this is a table that contains the probability that the clients were born in a certain quarter year and return the loan after some days late.

	(Jan-Mar)	(Apr-Jun)	(Jul-Nov)	(Oct-Dec)
On-Time	78.86%	77.97%	78.14%	78.67%
One Month	12.97%	14.07%	13.40%	13.76%
Two Months	2.69%	2.80%	2.97%	2.77%
Three Months	1.93%	1.88%	1.78%	1.50%
Four Months	0.94%	0.76%	0.92%	0.80%
More than Four Months	2.59%	2.49%	2.75%	2.47%

Now, Generally, we can say that about 78% of all clients return the loan with no late, about 13% return the loan within one month late, about 2% return the loan within two months late, about 3% return the loan within three months late, about 1% return the loan within four months late, about 3% return the loan after more than four months late.

Kmeans can investigate the distribution of the clients into 3 groups, and this using (PCA = 7)



Models

Now, Columns are (Loan Term, Gender, Customer Segment, Age, Month of date, Total O_S, DPD)

We divided the data into train and test data and made one-hot encoding

1) RandomForest Classifier:

Accuracy 77.10%

And after applying the GridSearch

Cross-Validation, the score is 77.79%

2) KNN:

Score 77.16%

And After applying leave one out
Cross-Validation

3) Decision Tree Classifier:

Score 77.74%

Finally, if we got more data like clients jobs and salaries. It would highly affect our results.