CASE STUDY – Credit EDA

By **Syed Saifullah Tarique and Aman Jha**

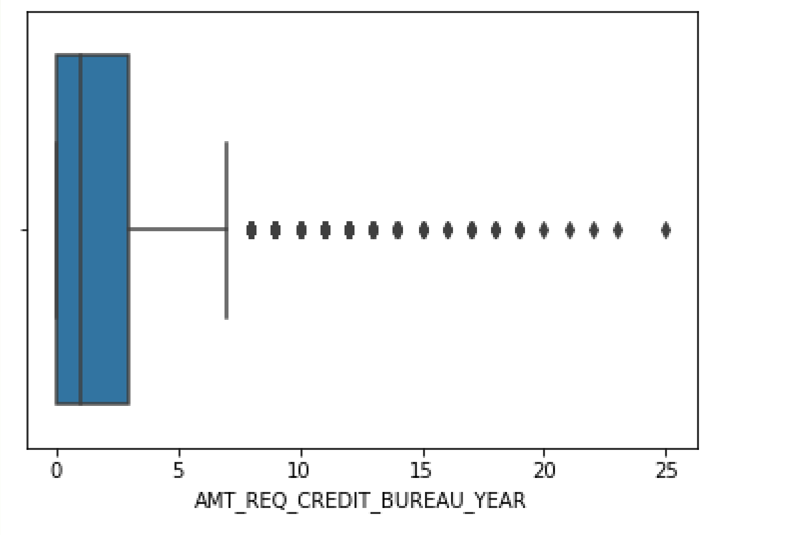
Business Understanding

- The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter.
- When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:
- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.
- The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:
- **The client with payment difficulties:** he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample,
- **All other cases:** All other cases when the payment is paid on time.
- When a client applies for a loan, there are four types of decisions that could be taken by the client/company):
- **Approved:** The Company has approved loan Application
- **Cancelled:** The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.
- **Refused:** The company had rejected the loan (because the client does not meet their requirements etc.).
- **Unused offer:** Loan has been cancelled by the client but on different stages of the process.

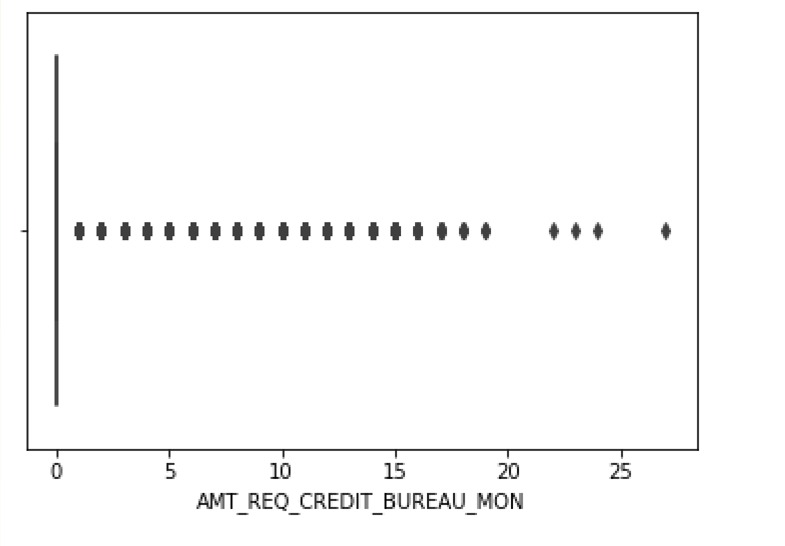
Analysis – Application Data

- Missing Values and approaches
- Median of the column is the 50th percentile value of that column in the above table
- In the case of missing categorical column values we ignore them mostly.
- For the 'EXT_SOURCE_3' column we should fill the missing values with the Median because both are close so we prefer the Median because it is much unbiased approach.
- For the 'AMT_REQ_CREDIT_BUREAU_YEAR' column we should fill the missing values with the Median because it clearly have an outlier by looking at the maximum value and the mean value. And since the mean value gets influenced by the outlier and median don't so we would use median value here.
- For the 'AMT_REQ_CREDIT_BUREAU_MON' column we should fill the missing values with the Median because it clearly have an outlier by looking at the maximum value and the mean value. And since the mean value gets influenced by the outlier and median don't so we would use median value here.
- For the 'AMT_REQ_CREDIT_BUREAU_WEEK' column we should fill the missing values with the Median because it clearly have an outlier by looking at the maximum value and the mean value. And since the mean value gets influenced by the outlier and median don't so we would use median value here.
- For the 'AMT_REQ_CREDIT_BUREAU_DAY' column we should fill the missing values with the Median because it clearly have an outlier by looking at the maximum value and the mean value. And since the mean value gets influenced by the outlier and median don't so we would use median value here.
- For the 'AMT_REQ_CREDIT_BUREAU_HOUR' column we should fill the missing values with the Median because it clearly have an outlier by looking at the maximum value and the mean value. And since the mean value gets influenced by the outlier and median don't so we would use median value here.
- For the 'AMT_REQ_CREDIT_BUREAU_QRT' column we should fill the missing values with the Median because it clearly have an outlier by looking at the maximum value and the mean value. And since the mean value gets influenced by the outlier and median don't so we would use median value here.
- For the 'OBS_30_CNT_SOCIAL_CIRCLE' column we should fill the missing values with the Median because it clearly have an outlier by looking at the maximum value and the mean value. And since the mean value gets influenced by the outlier and median don't so we would use median value here.
- - For the 'DEEF_30_CNT_SOCIAL_CIRCLE' column we should fill the missing values with the Median because it clearly have an outlier by looking at the maximum value and the mean value. And since the mean value gets influenced by the outlier and median don't so we would use median value here.

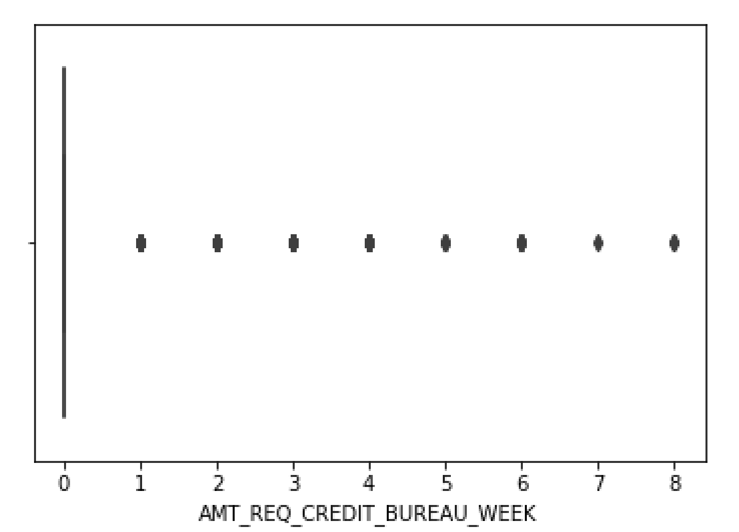
Various Plots



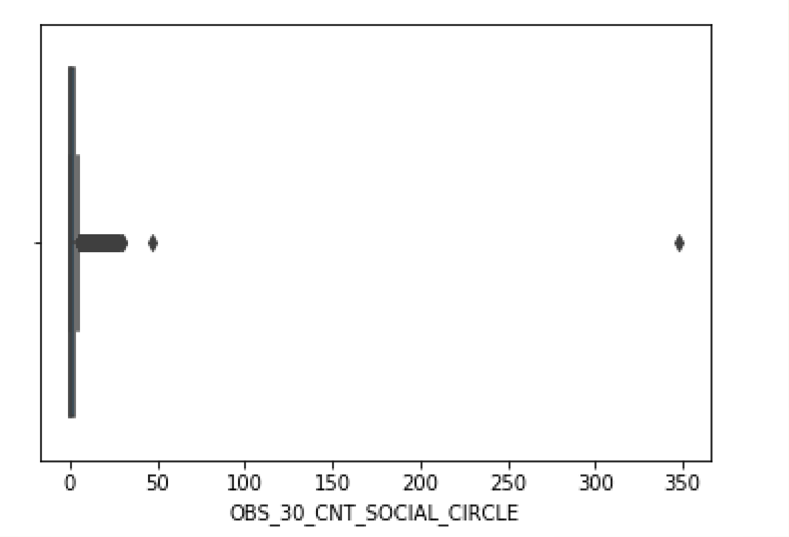
- Here we could easily see in the above plot that this column has outliers. Because those are the points which are way far away from the most of the data range as could be seen in the graph.
- This column represent the number of enquiries to credit bureau about the client one year before the application.
- So the outliers could be a mistake in which we have to rectify it(if we could) or remove it. In other case outlier could be a variance in the data which we could confirm by asking the company or by checking the other similar data for that employee.



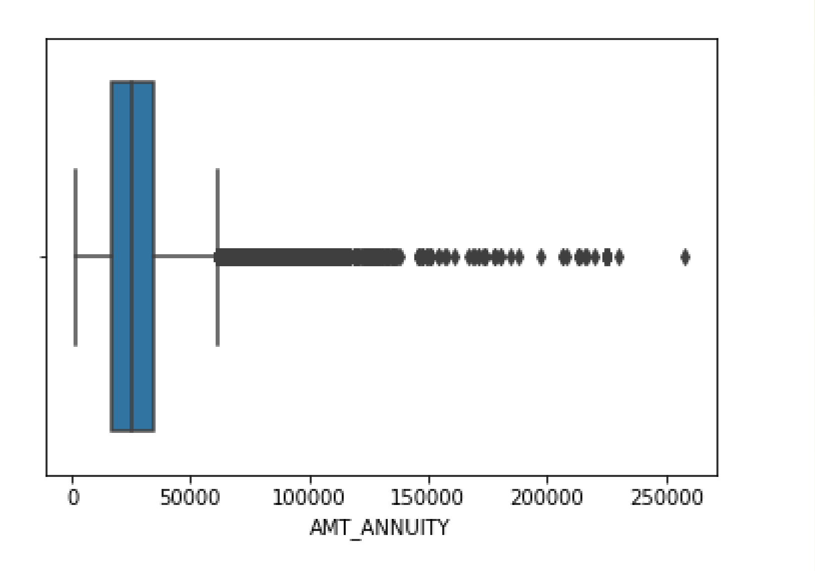
- Here we could easily see in the above plot that this column has outliers. Because those are the points which are way far away from the most of the data range as could be seen in the graph.
- This column represent the number of enquiries to credit bureau about the client a month before the application.
- So the outliers could be a mistake in which we have to rectify it(if we could) or remove it. In other case outlier could be a variance in the data which we could confirm by asking the company or by checking the other similar data for that employee.



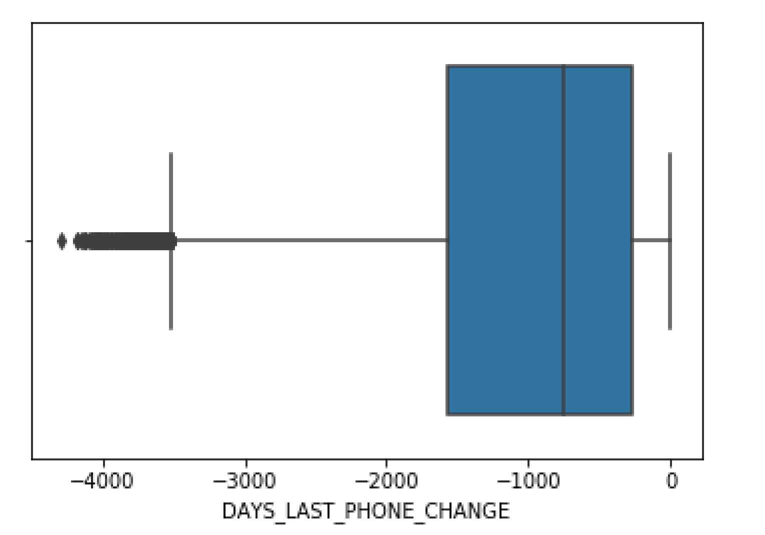
- Here we could easily see in the above plot that this column has outliers. Because those are the points which are way far away from the most of the data range as could be seen in the graph.
- - This column represent the number of enquiries to credit bureau about the client a week before the application.
- - So the outliers could be a mistake in which we have to rectify it(if we could) or remove it. In other case outlier could be a variance in the data which we could confirm by asking the company or by checking the other similar data for that employee.



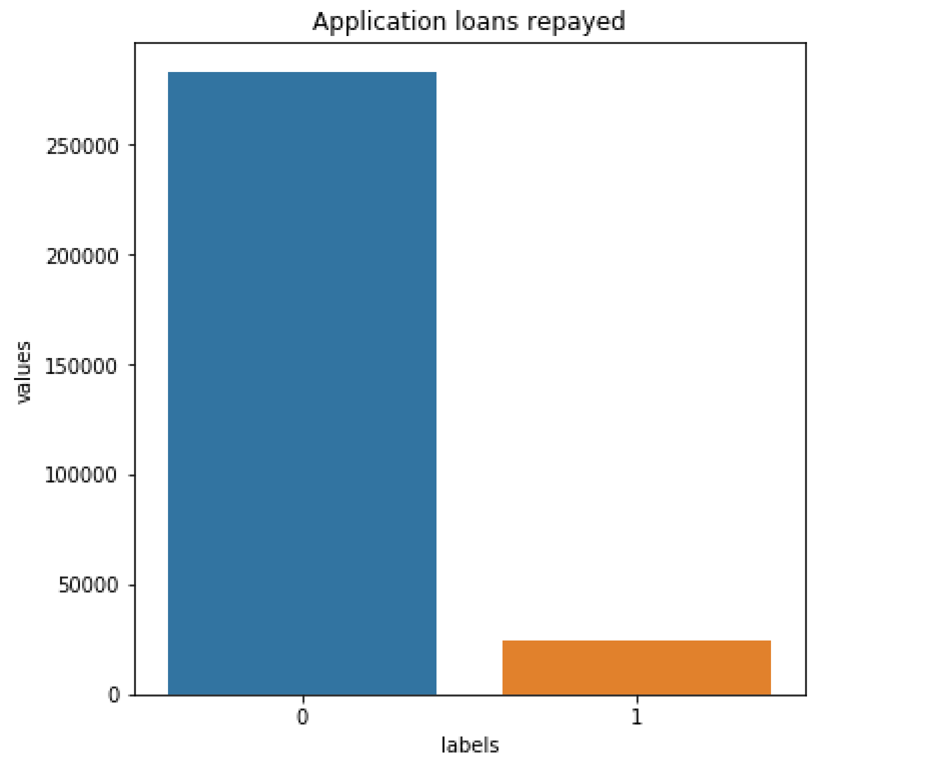
- Here we could easily see in the above plot that this column has outliers. Because those are the points which are way far away from the most of the data range as could be seen in the graph.
- - This column represent how many observation of the client social surrounding with observable 30 days past due.
- - Here the outliers is a mistake in which we have to rectify it (if we could) or remove it because the values can't be greater than 30 days.



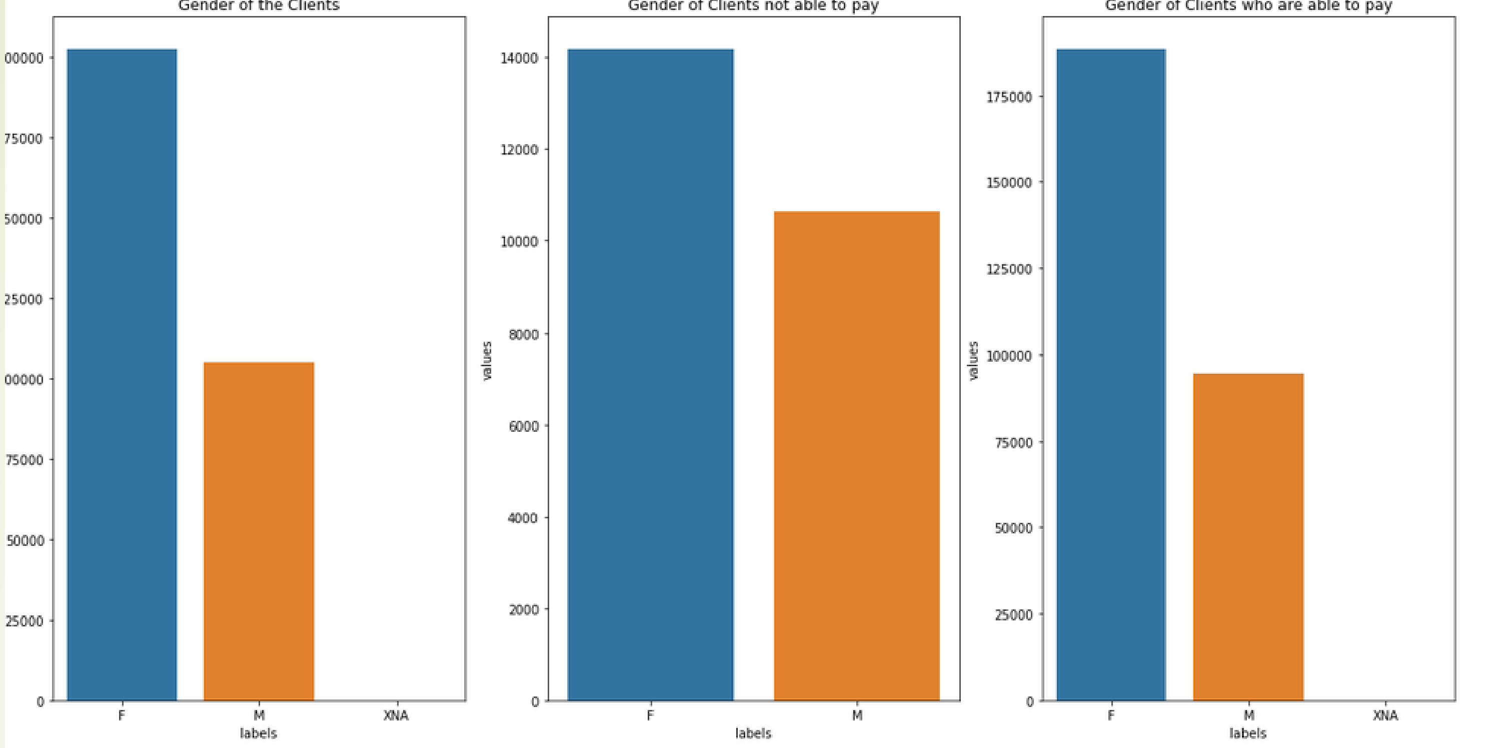
- Here we could easily see in the above plot that this column has outliers. Because those are the points which are way far away from the most of the data range as could be seen in the graph.
- This column represent the loan annuity.
- So the outliers could be a mistake in which we have to rectify it(if we could) or remove it. In other case outlier could be a variance in the data which we could confirm by asking the company or by checking the other similar data for that employee.



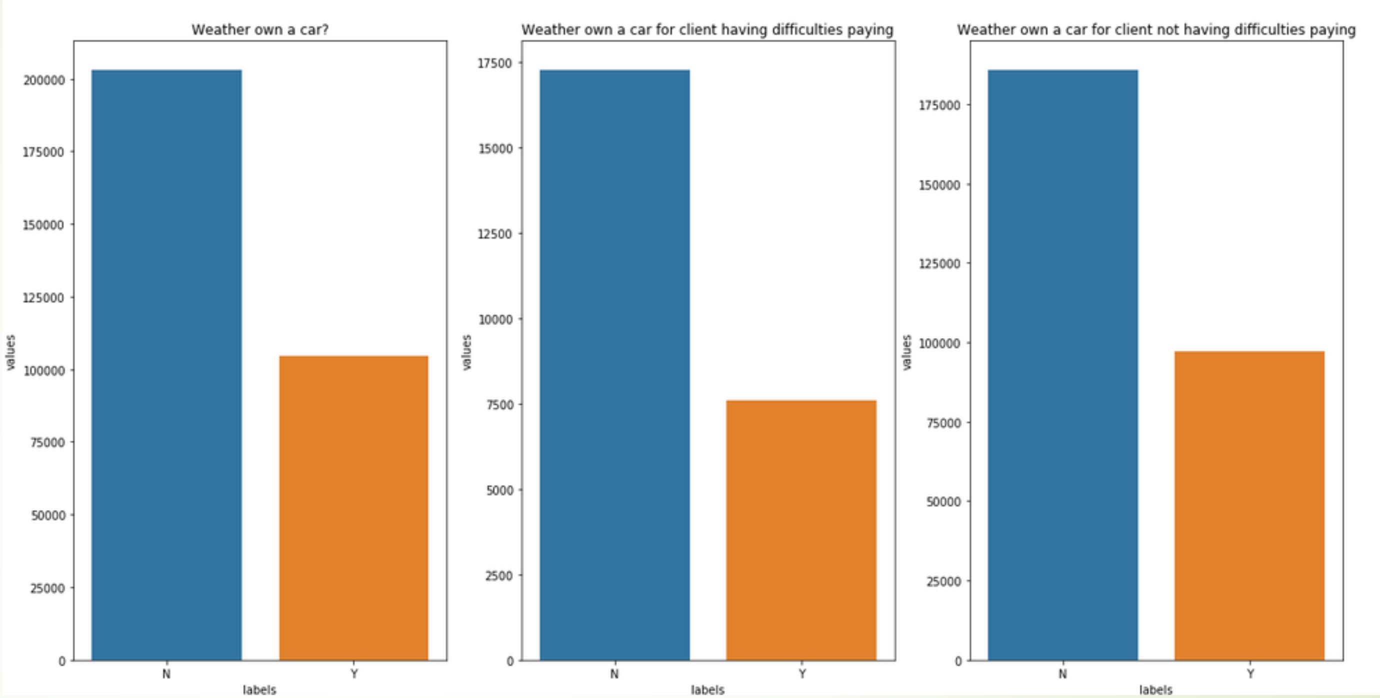
- - Here we could easly see in the above plot that this column has outliers. Because those are the points which are way far away from the most of the data range as could be seen in the graph.
- - This column represent how many days before the application did the client cahnge phone number.
- - Here the outliers is a mistake in which we have to rectify it (if we could) or remove it because he can't change the number very large number of days before because he won't be born to do so.



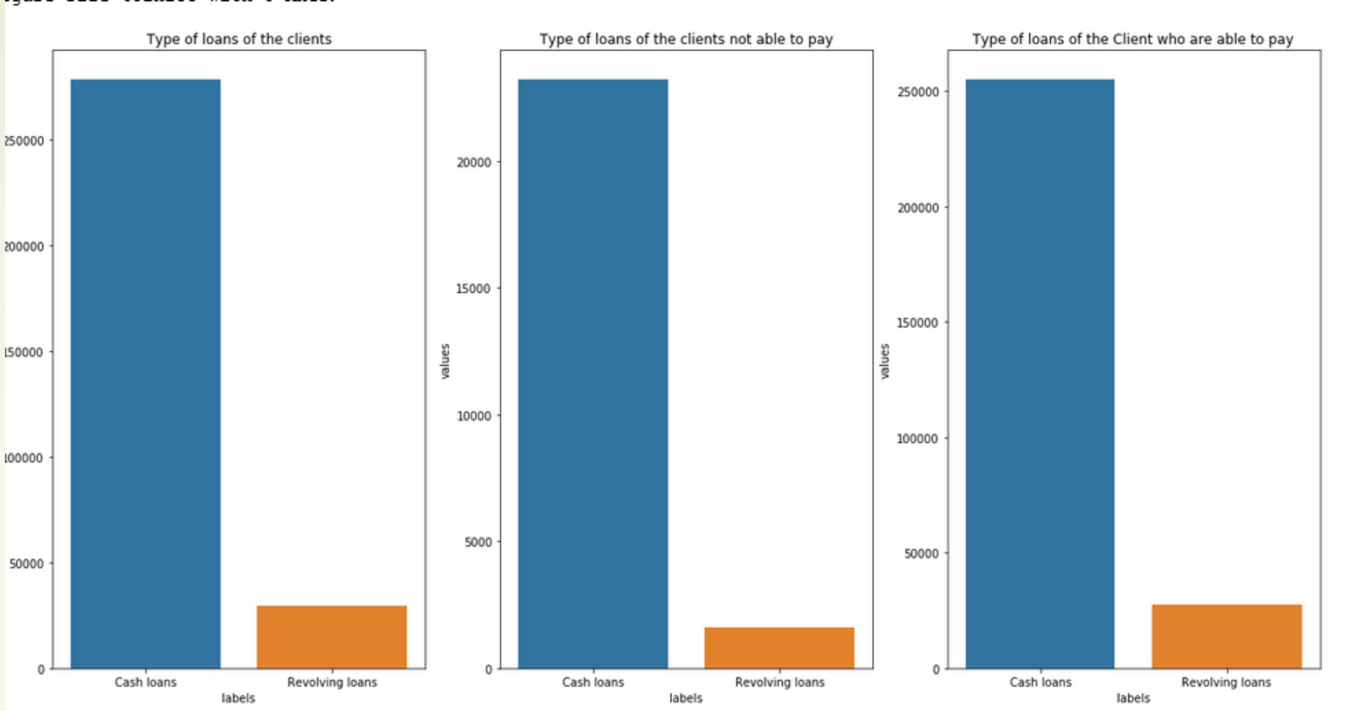
- In the above graph the 1 means Client has payment difficulties and 0 means he doesn't have payment difficulties



- We could see that there are more number of Female clients than Male Clients (Nearly twice as much).
- Here we could see in number wise there are more Female Client who has problem in paying the loan but If we see the above graph then we could see that there are much more number of Female who are taking loan. Therefore percentage wise there are more Male clients having problem in repaying loan than Female.



- Clients who don't own a car are nearly twice of those people who owns a car.
- Their are more client that don't have a car have difficulties paying than the ones who has a car.

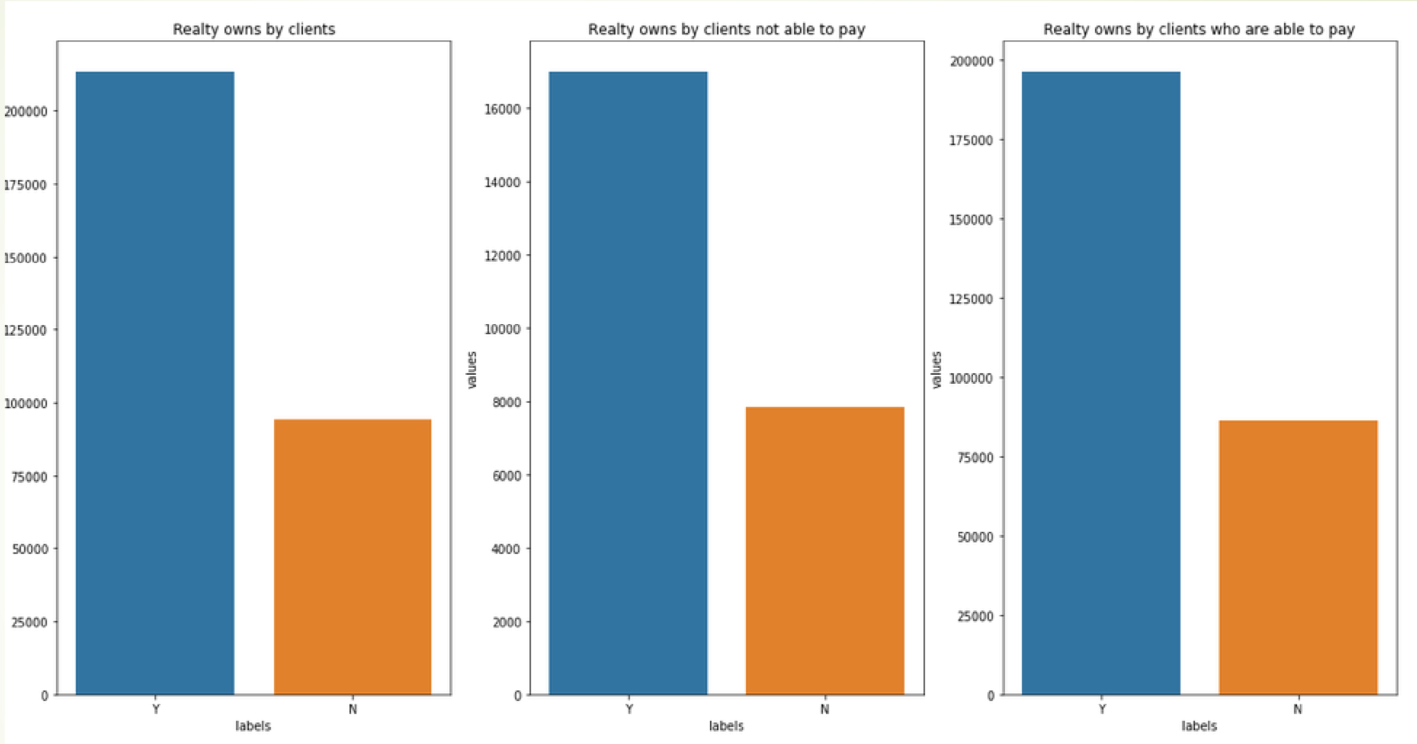


Type of loans of the clients

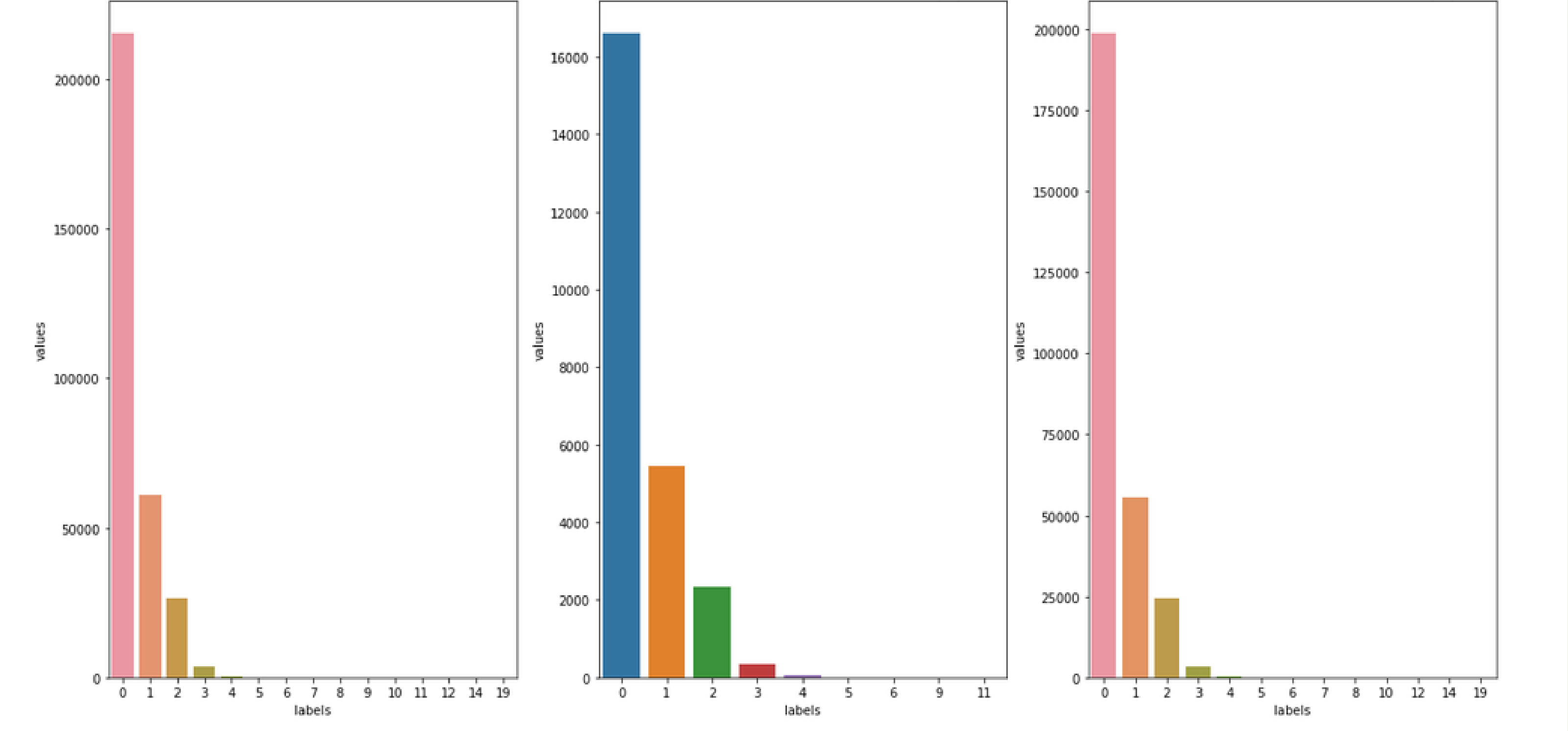
Type of loans of the clients not able to pay

Type of loans of the Client who are able to pay

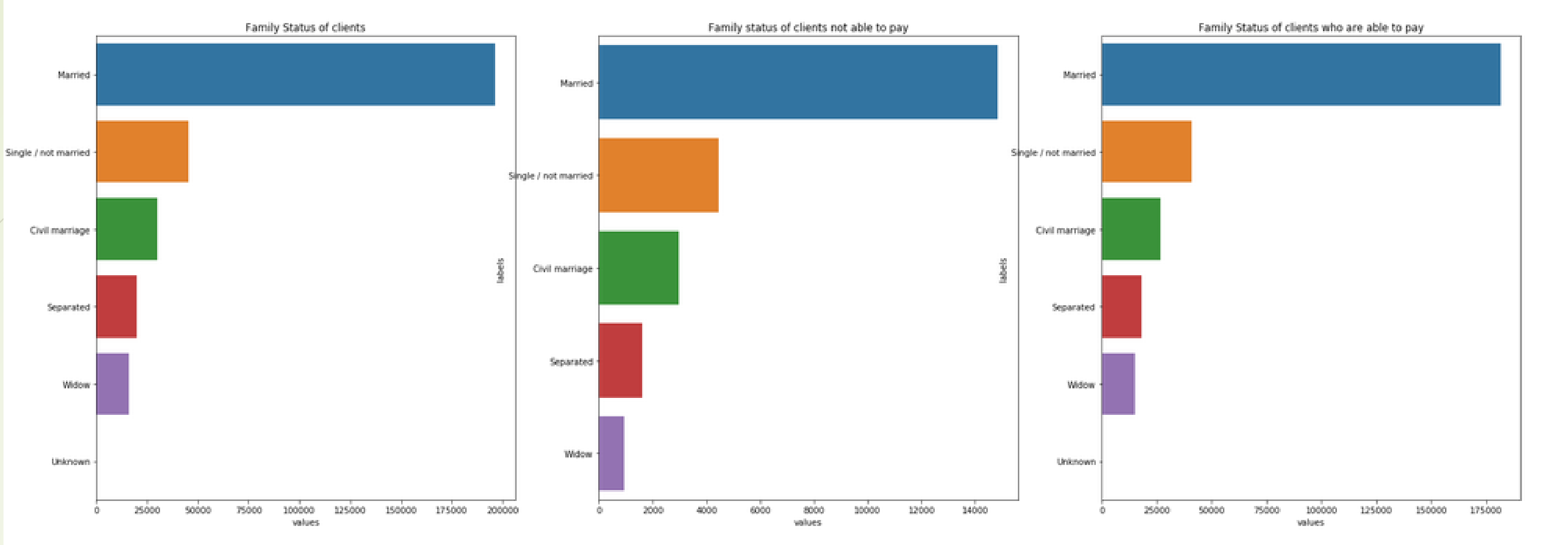
- Revolving type of loans are very small fraction nearly 10% of the loans.
- There are large amount of cash loans are there which are not being paid but there are also very large amount of revolving loans are not being paid compare to its frequency.



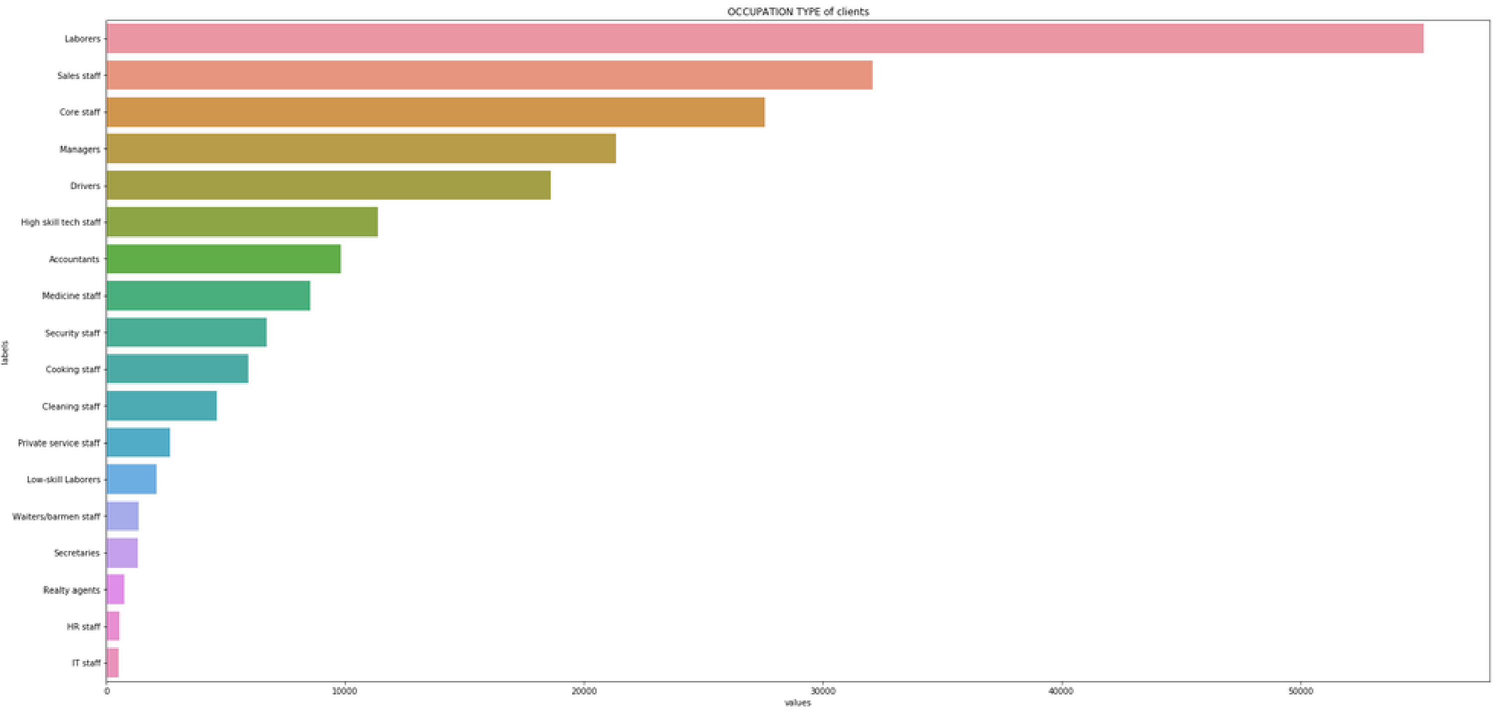
- Clients that owns realty are nearly twice that who doesn't.
- Here weather the client would pay the loan is not much effected by weather they own the realty or not.



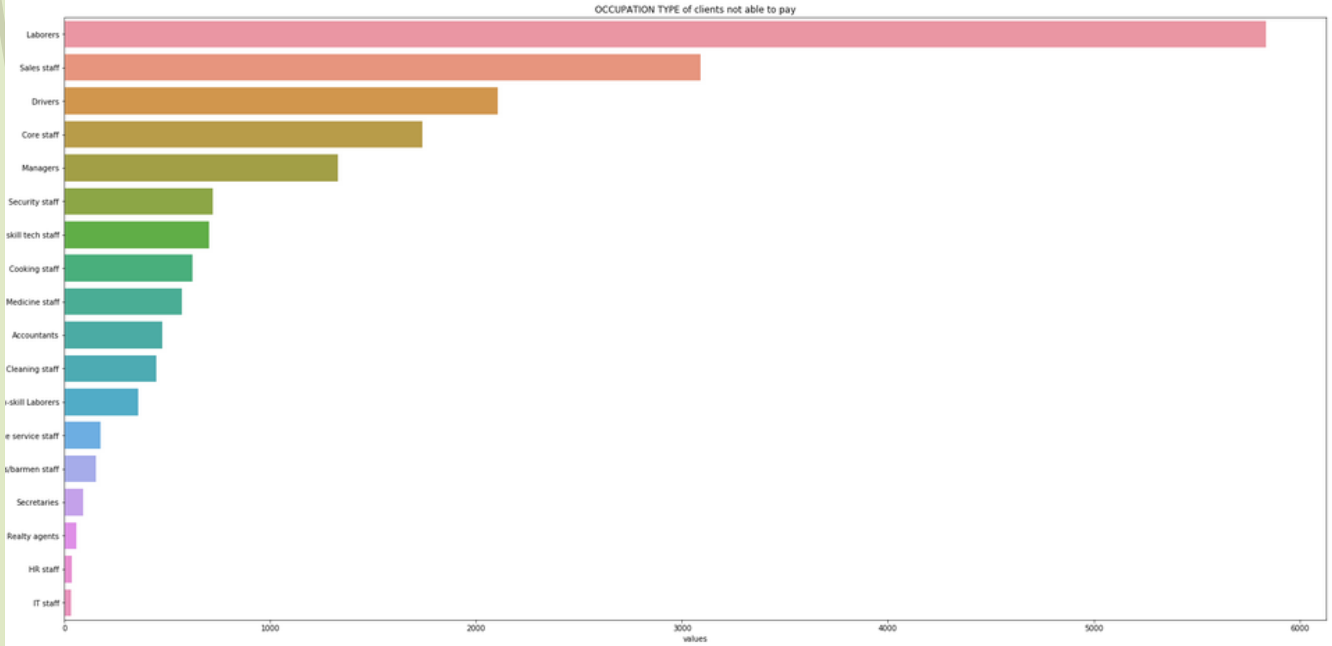
- Most of the clients don't have children.
- There are a large number of people who don't have children have problem paying the loan

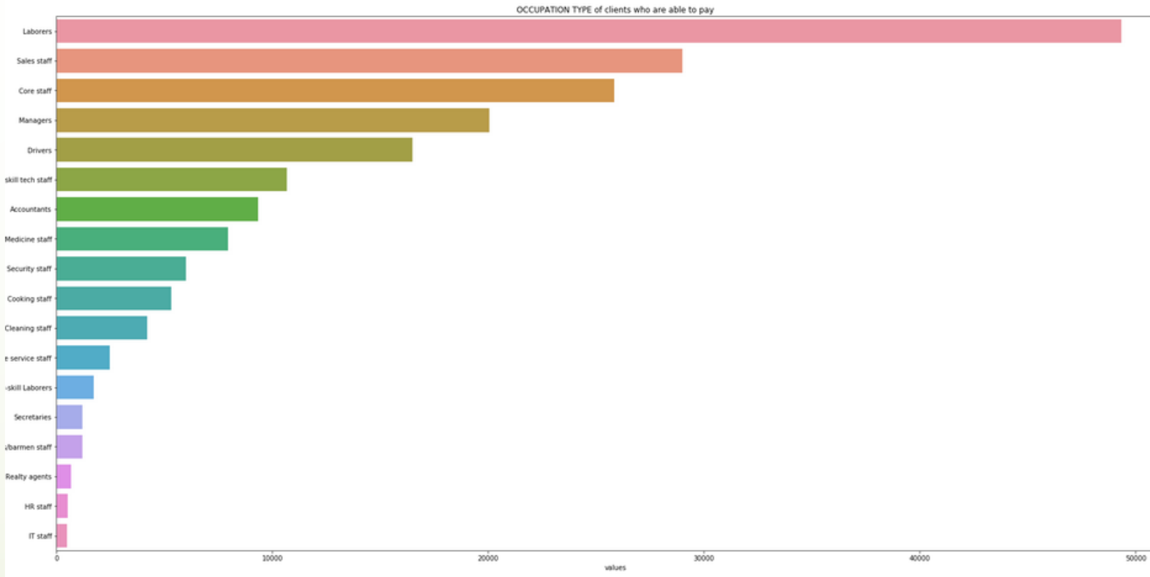


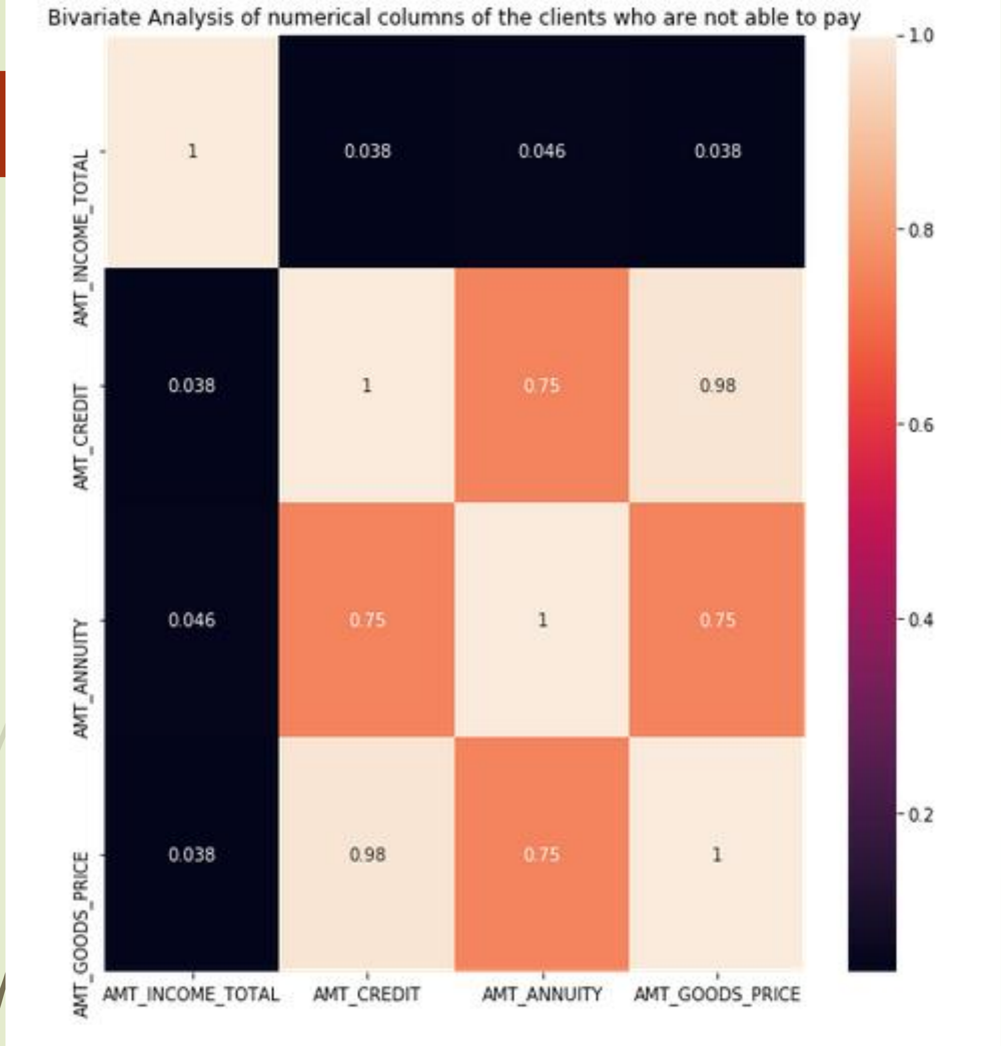
- Most of clients are married, followed by Single/not married and civil marriage.
- There are large number of client who are not able to pay are married and the least number of client who are unbale to pay are widowed.

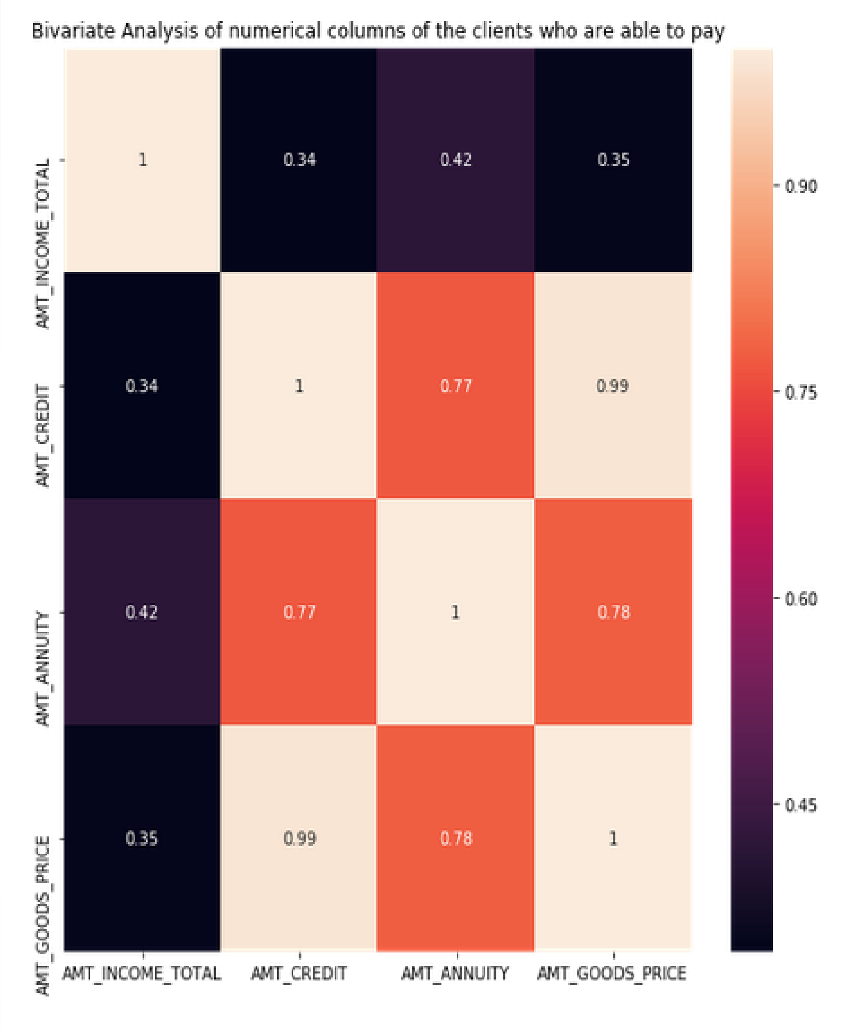


Most of the loans are taken by Laborers, followed by Sales staff.
IT staff take the lowest amount of loans.
Large numbers of unpaid loan by Laborers and Drivers followed by sales staff

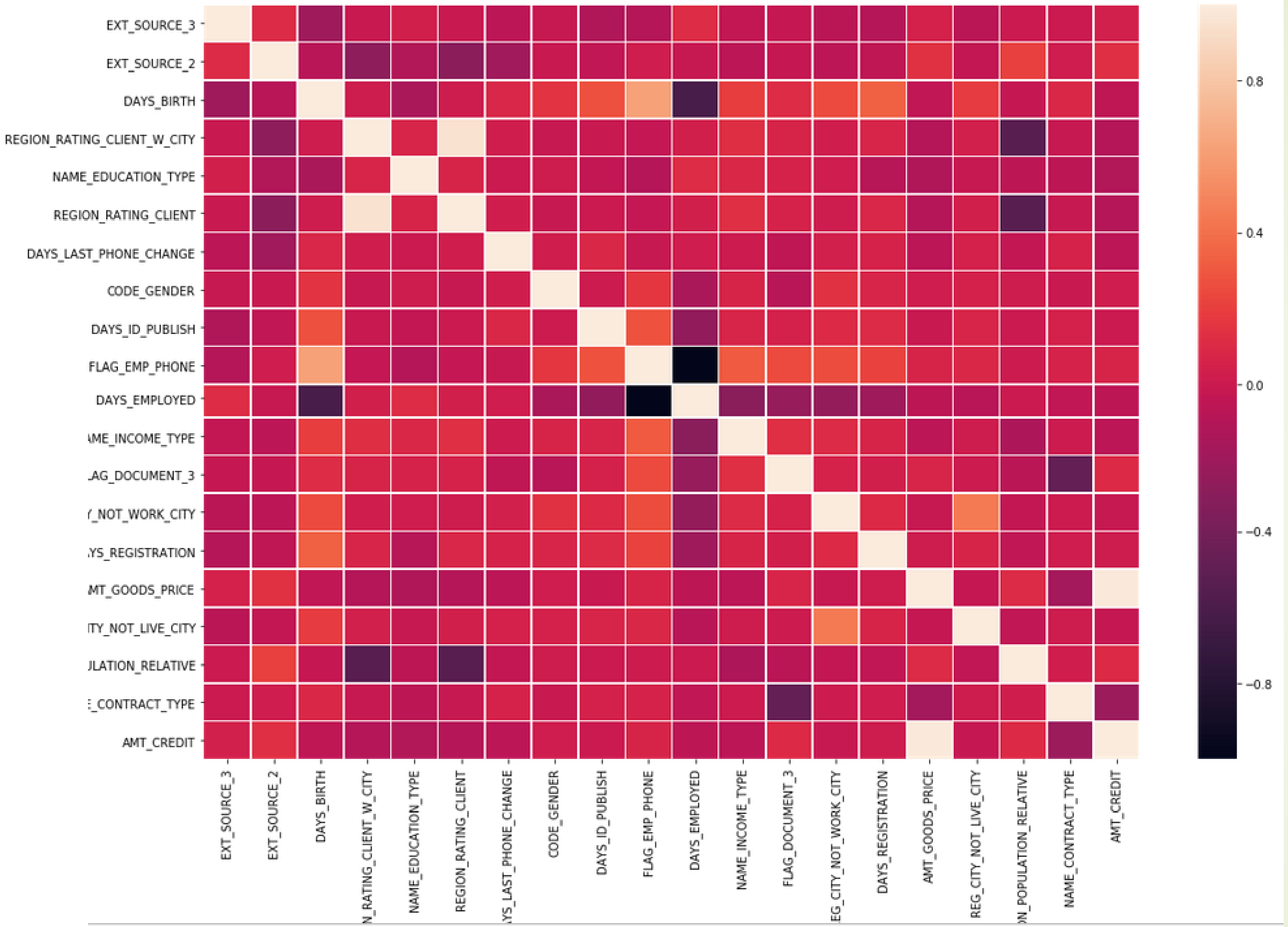





Bivariate Analysis of numerical columns of the clients who are not able to pay
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
1
0.038
0.046
0.038
0.038
1
0.75
0.98
0.046
0.75
1
0.75
0.038
0.98
0.75
1
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
1.0
0.8
0.6
0.4
0.2


Bivariate Analysis of numerical columns of the clients who are able to pay
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
1
0.34
0.42
0.35
0.34
1
0.77
0.99
0.42
0.77
1
0.78
0.35
0.99
0.78
1
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
0.90
0.75
0.60
0.45

- Here we could see that AMT_ANNUITY and AMT_CREDIT are highly correlated, if one goes up so s the other and vise versa. For both the sets but much more correlated for the Clients who are able to pay.
- - We could also see that AMT_CREDIT and AMT_GOODS_PRICE are also highly correlated, if one goes up so s the other and vise versa. For both the sets but are more correlated to the Clients who are are able to pay.
- - Similarly AMT_ANNUITY and AMT_GOODS_PRICE are also highly correlated, if one goes up so s the other and vise versa. For both the sets but are more correlated to the Clients who are able to pay.
- - For AMT_ANNUITY and AMT_INCOME_TOTAL the correlation is much higher for the Clients who are able to pay. Similar is the case with AMT_CREDIT and AMT_INCOME_TOAL.


EXT_SOURCE_3
EXT_SOURCE_2
DAYS_BIRTH
REGION_RATING_CLIENT_W_CITY
NAME_EDUCATION_TYPE
REGION_RATING_CLIENT
DAYS_LAST_PHONE_CHANGE
CODE_GENDER
DAYS_ID_PUBLISH
FLAG_EMP_PHONE
DAYS_EMPLOYED
AME_INCOME_TYPE
AG_DOCUMENT_3
_NOT_WORK_CITY
YS_REGISTRATION
MT_GOODS_PRICE
ITY_NOT_LIVE_CITY
JLATION_RELATIVE
_CONTRACT_TYPE
AMT_CREDIT
EXT_SOURCE_3
EXT_SOURCE_2
DAYS_BIRTH
N_RATING_CLIENT_W_CITY
NAME_EDUCATION_TYPE
REGION_RATING_CLIENT
YS_LAST_PHONE_CHANGE
CODE_GENDER
DAYS_ID_PUBLISH
FLAG_EMP_PHONE
DAYS_EMPLOYED
NAME_INCOME_TYPE
FLAG_DOCUMENT_3
EG_CITY_NOT_WORK_CITY
DAYS_REGISTRATION
AMT_GOODS_PRICE
REG_CITY_NOT_LIVE_CITY
N_POPULATION_RELATIVE
NAME_CONTRACT_TYPE
AMT_CREDIT
0.8
0.4
0.0
−0.4
−0.8

- We can have look on the values and there correction. Lighter the color more corrected values.
- - 'FLAG_EMP_PHONE' is highly correlated with 'DAYS_BIRTH' i.e if one goes up so is the other.
- Similarly 'REG_CITY_NOT_LIVE_CITY' is highly correlated with 'REG_CITY_NOT_WORK_CITY'.
- There are some which are highly negatively correlated that means that if one increases other decreases. e.g. 'DAYS_EMPLOYED' and 'FLAG_EMP_PHONE' etc.

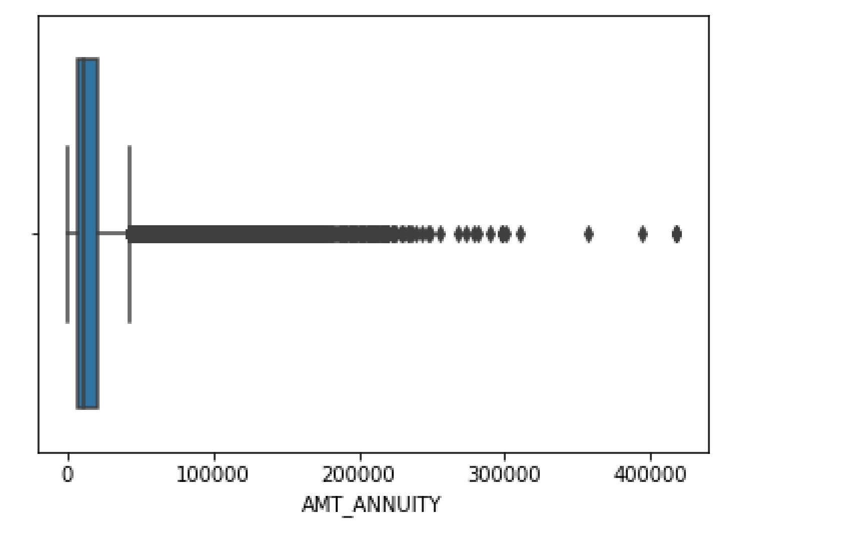
The top 10 correlation for the Client with payment difficulties and all other cases (Target variable)

'EXT_SOURCE_3', 'EXT_SOURCE_2', 'NAME_EDUCATION_TYPE', 'DAYS_LAST_PHONE_CHANGE', 'CODE_GENDER', 'DAYS_ID_PUBLISH', 'NAME_INCOME_TYPE', 'REGION_POPULATION_RELATIVE', 'NAME_CONTRACT_TYPE', 'AMT_CREDIT'

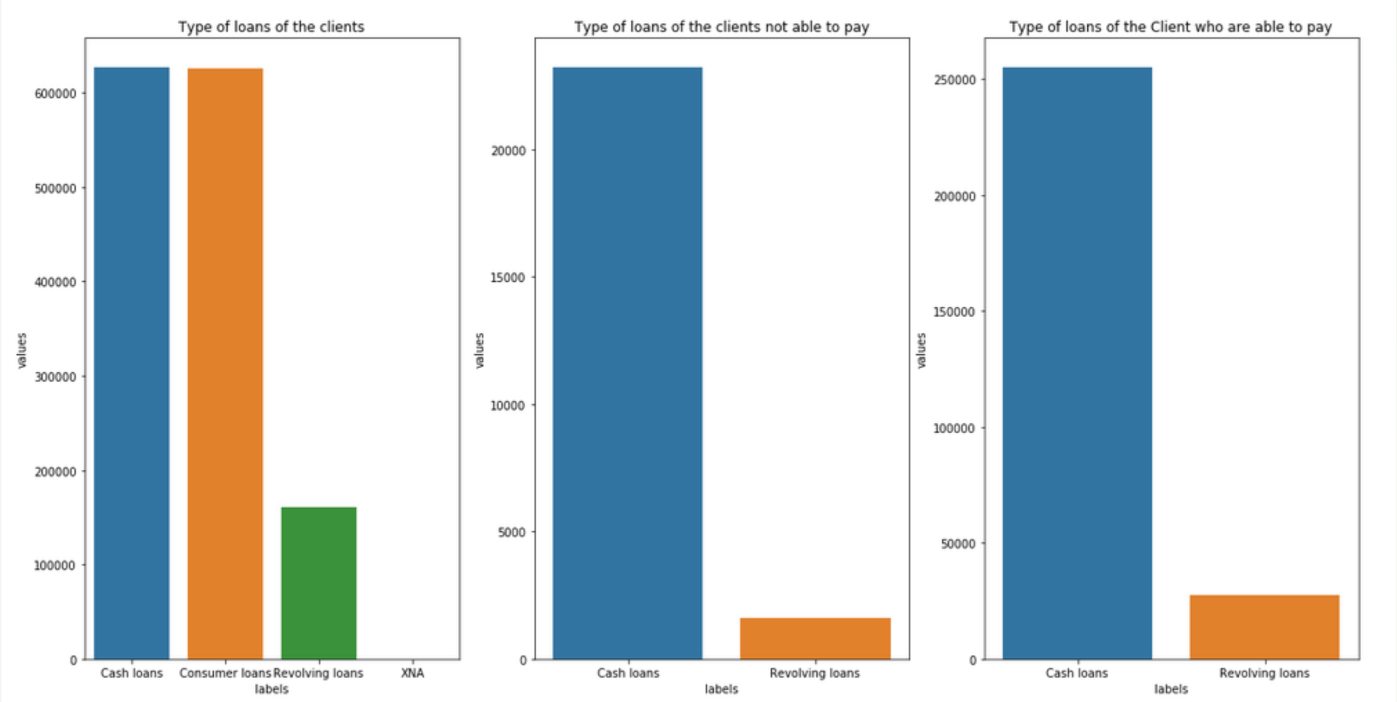
Loan defaulting is directly(positively) correlated to NAME_EDUCATION_TYPE,DAYS_LAST_PHONE_CHANGE,CODE_GENDER,DAYS_ID_PUBLISH,NAME_INCOME_TYPE

Loan defaulting is inverserly (negativly) correlated to EXT_SOURCE_3,EXT_SOURCE_2,REGION_POPULATION_RELATIVE,NAME_CONTRACT_TYPE,AMT_CREDIT

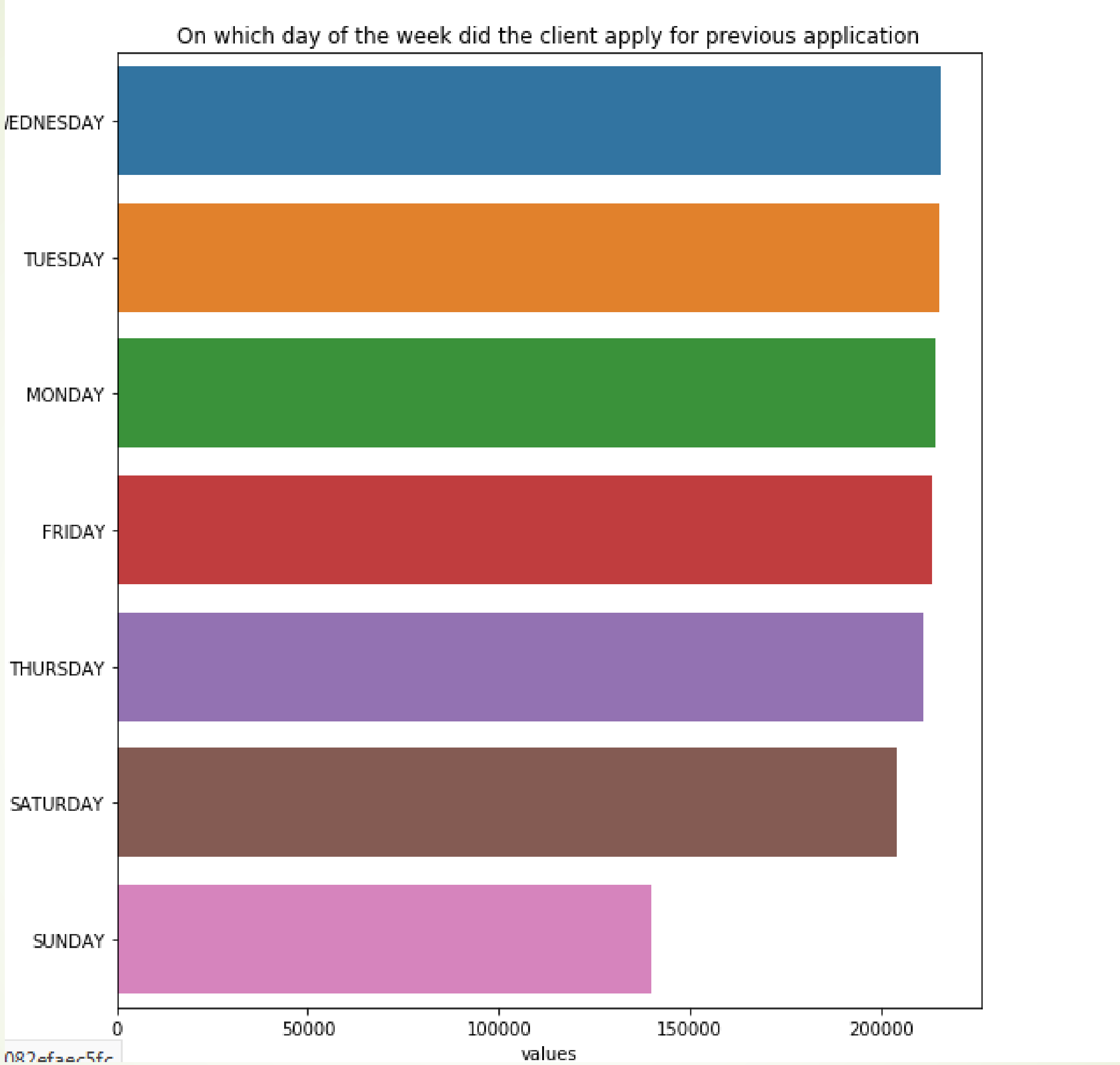
Analysis Previous Application



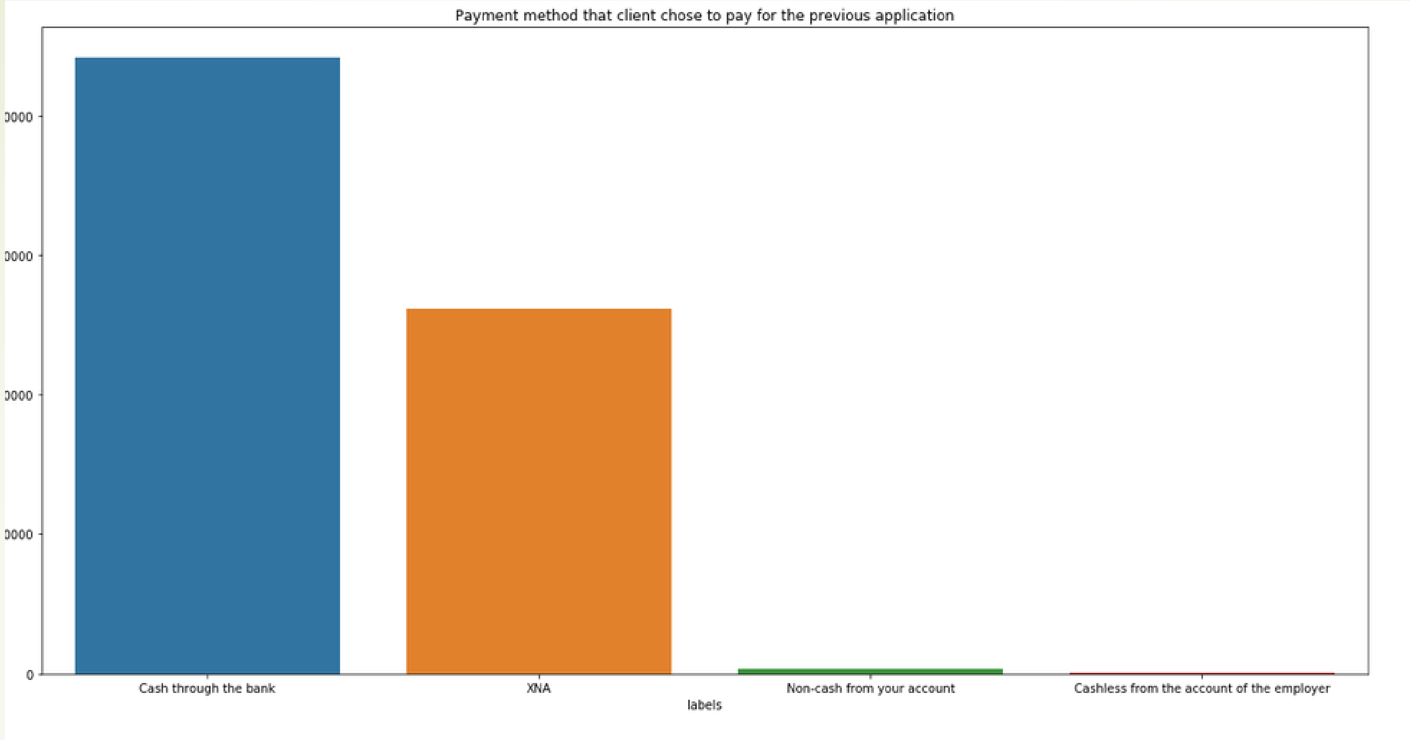
- ere we could easily see in the above plot that this column has outliers. Because those are the points which are way far away from the most of the data range as could be seen in the graph.
- - This column represent the annuity of the previous application.
- - So the outliers could be a mistake in which we have to rectify it(if we could) or remove it. In other case outlier could be a variance in the data which we could confirm by asking the company or by checking the other similar data for that employee.



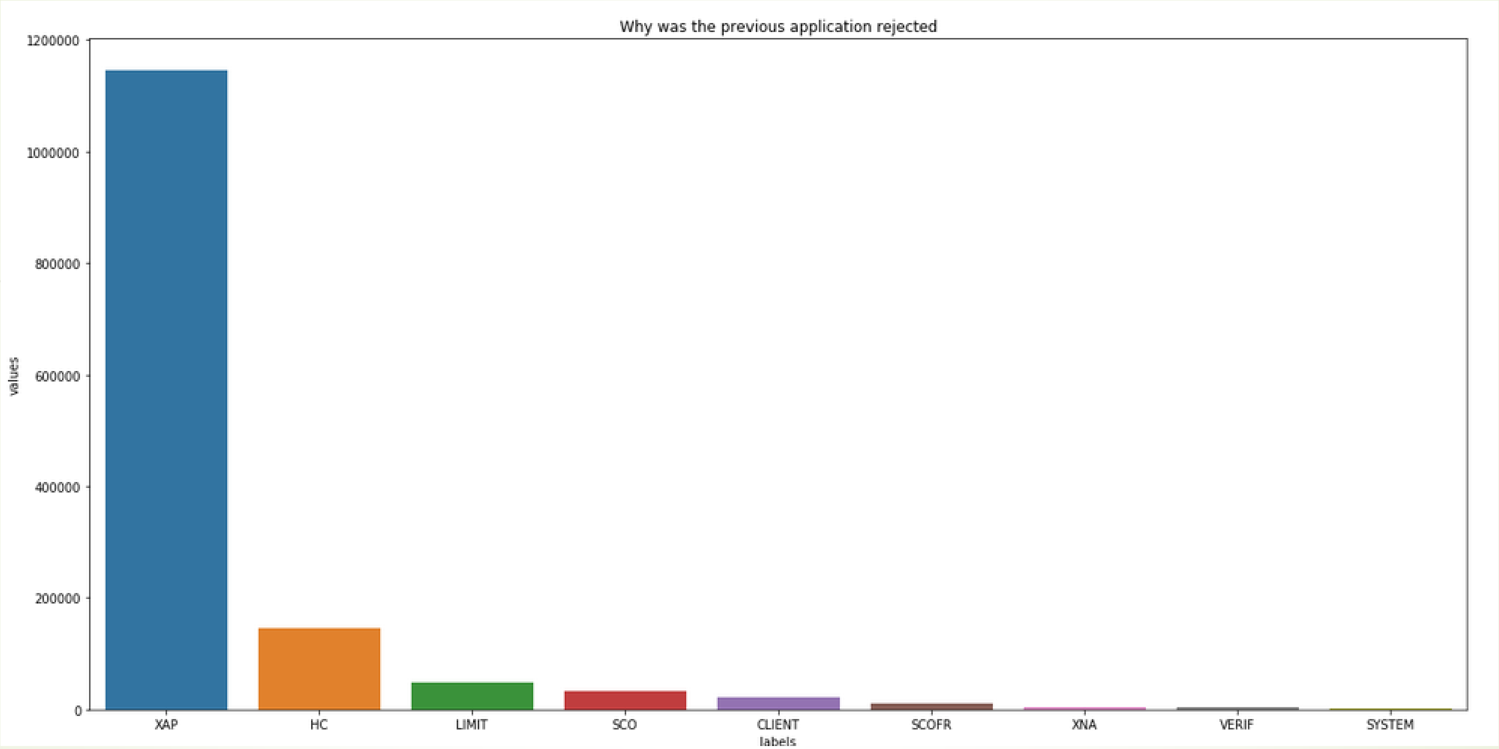
- We could see that the Cash loans are the large amount of which loans are unable to being paid


On which day of the week did the client apply for previous application

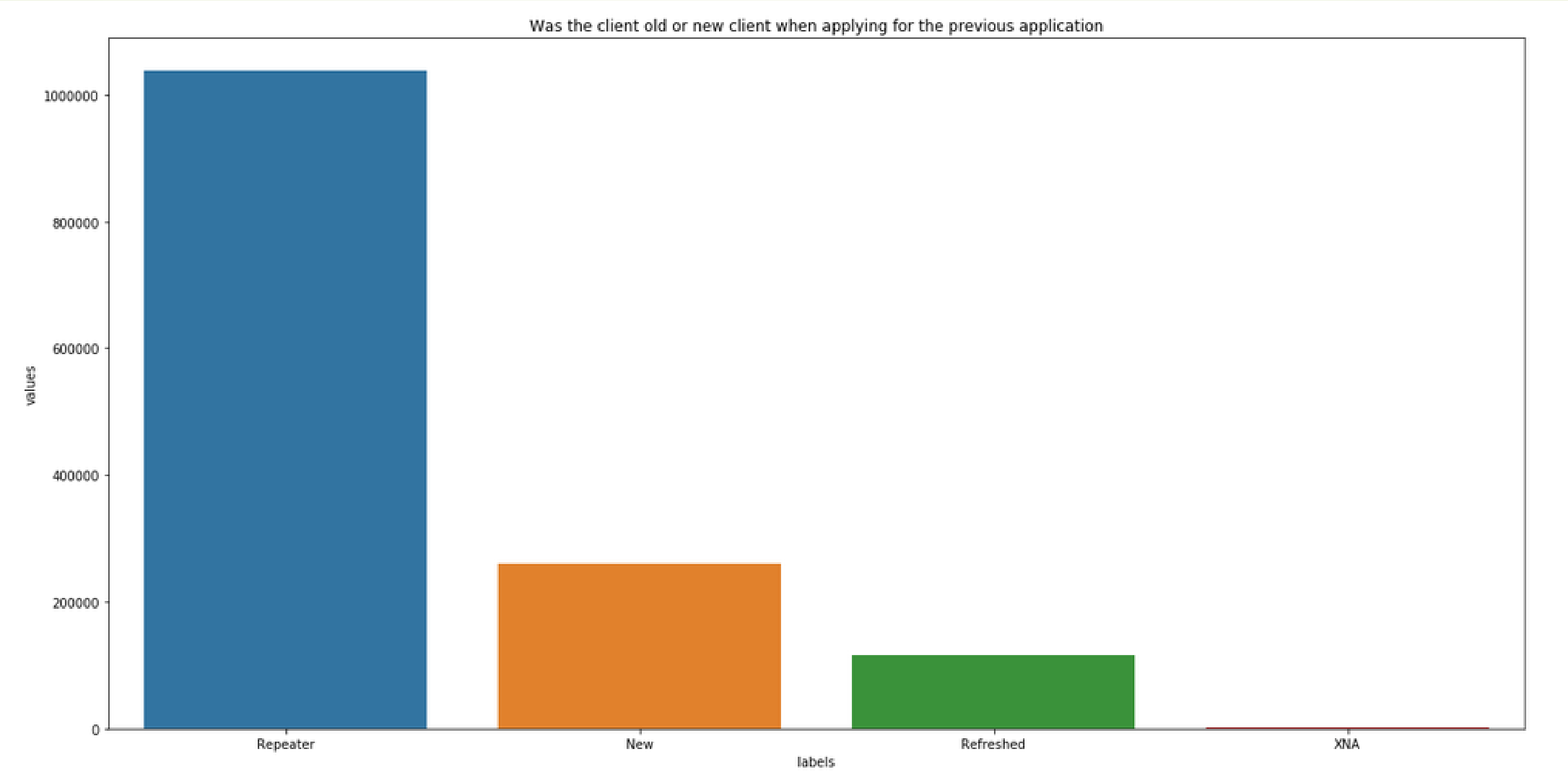
- Here we could see that except for Sunday loans are taken almost equally on every day of the week, but on Sunday there is quite decrease in the amount of loan taken. Which may be due the Weekend Holiday.



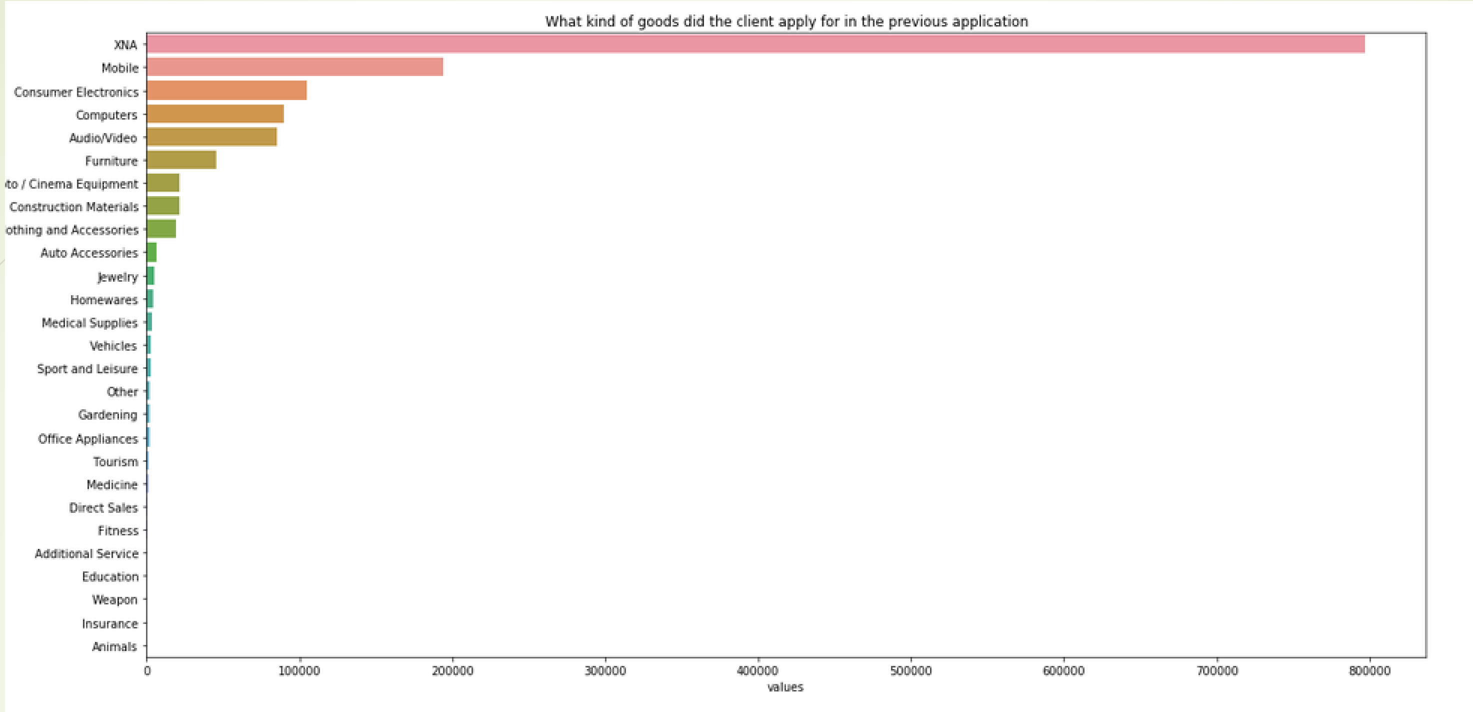
- We could see that most of the payment in previous application is done by cash through bank. Which tells us that people prefer the traditional method of paying by cash even the online method is available.



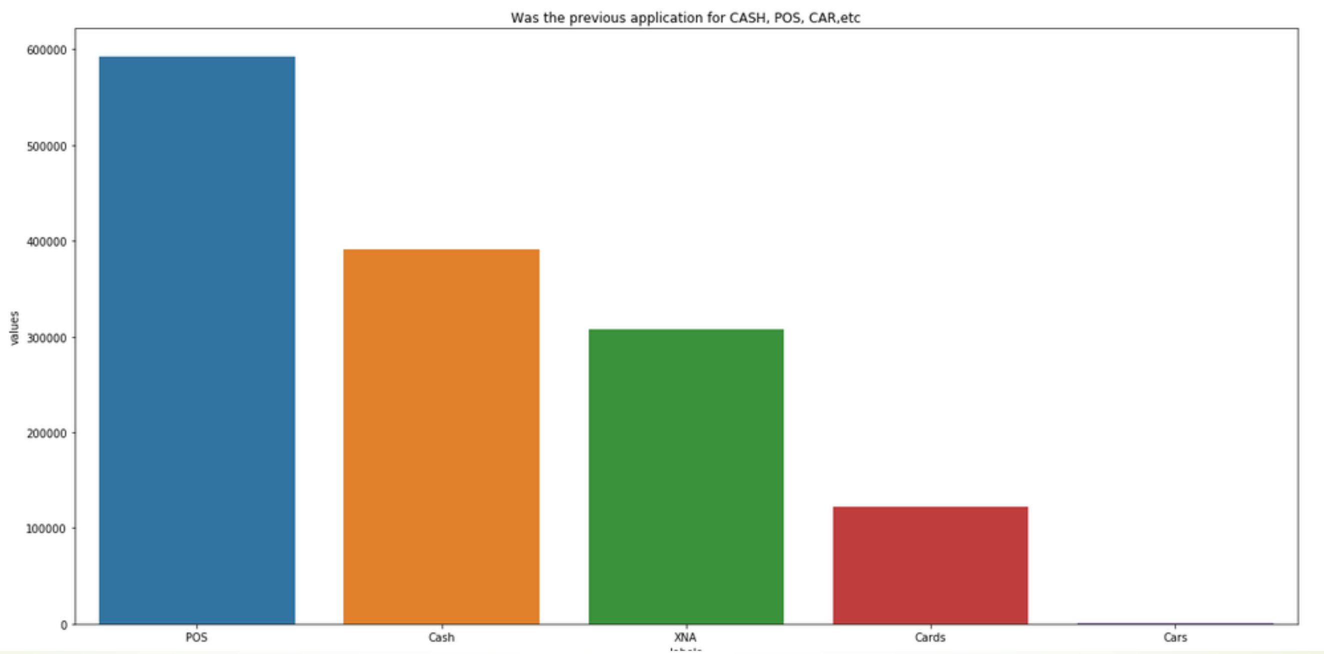
- The most of the previous application got rejected due in 'XAP' process, followed by 'HC' and the least rejection in 'SYSTEM'.



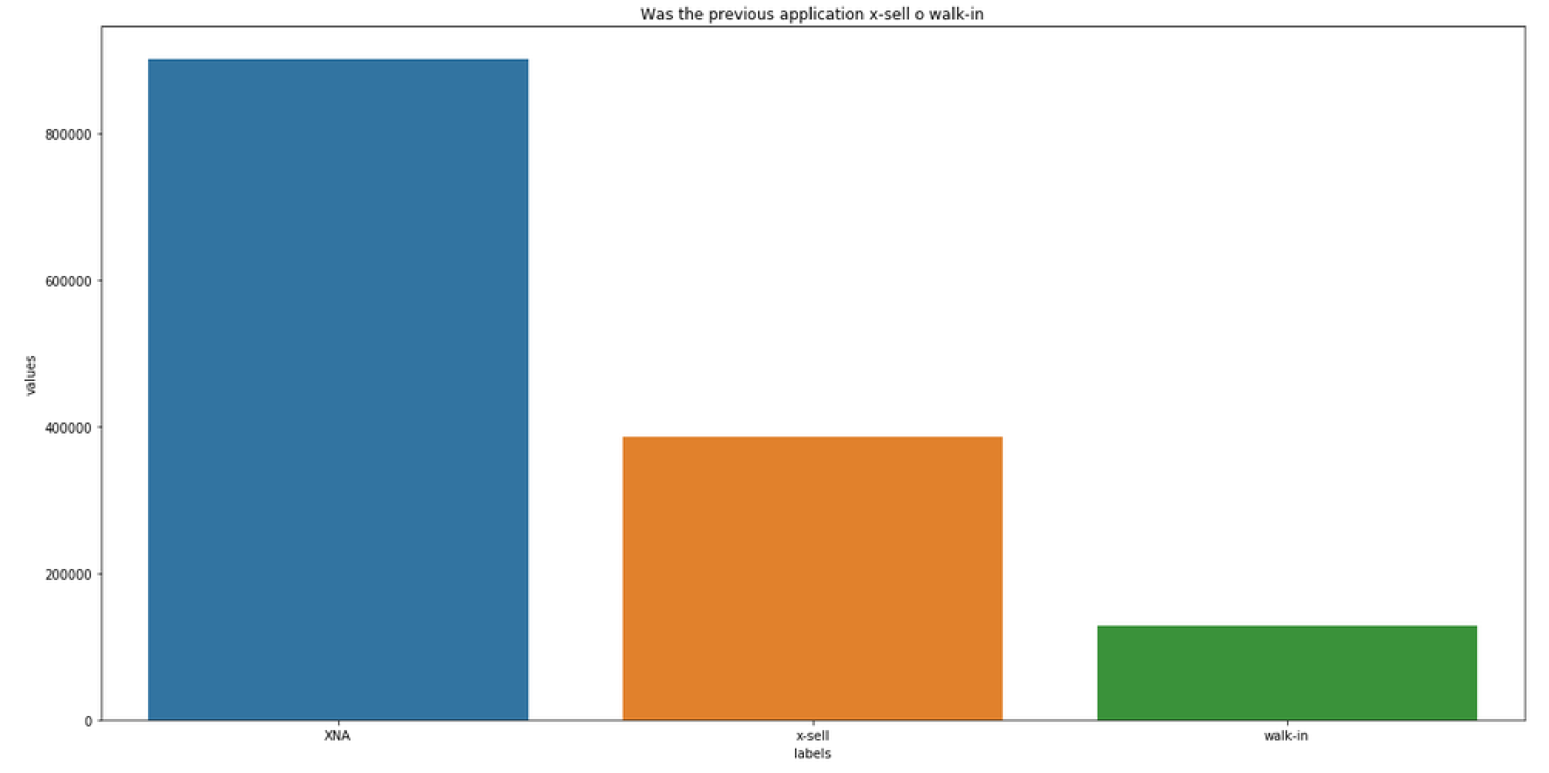
- Most of the client in previous application was Repeater, which told us that their trust on the company is good. Very few number of clients are nw compare to the repeater.



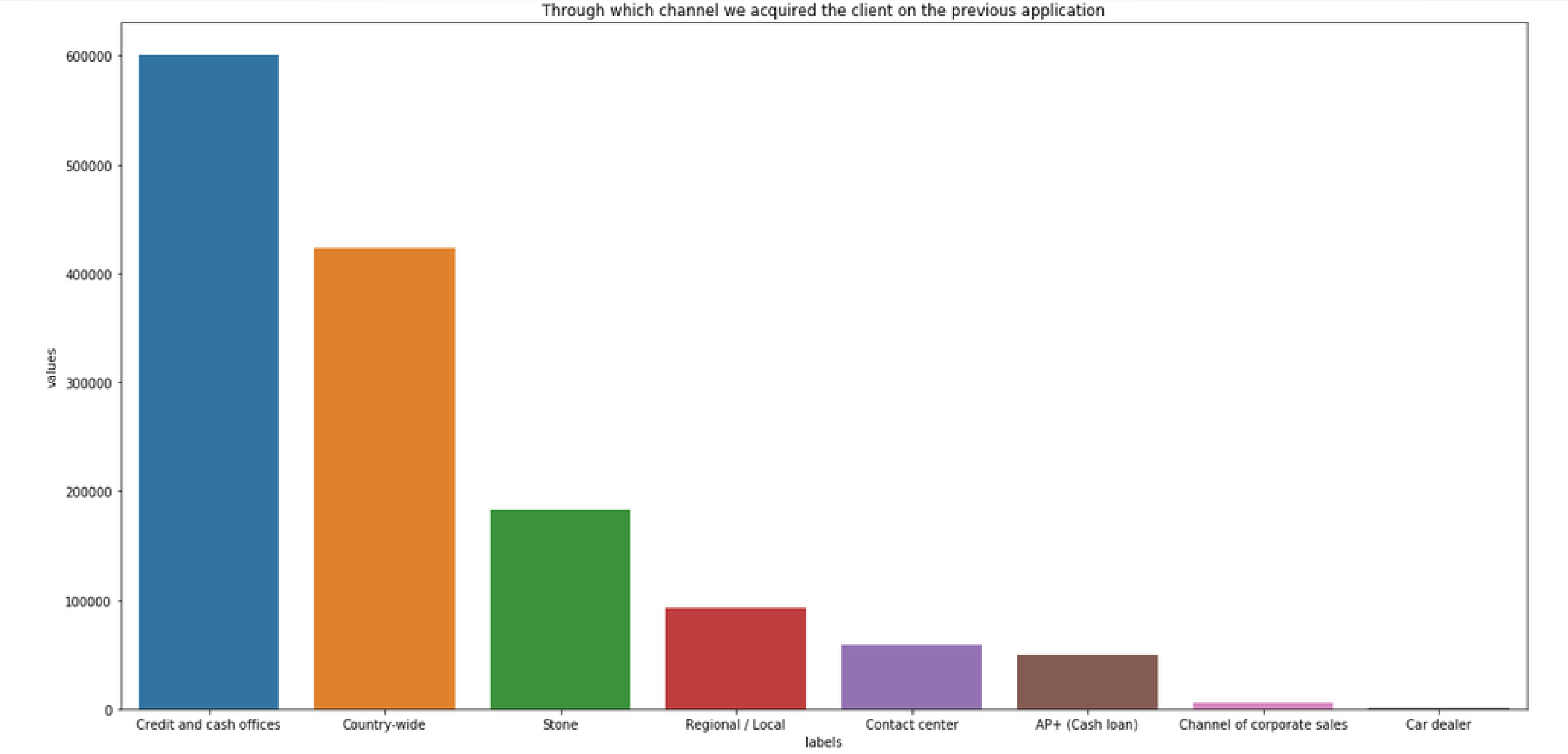
- Most of the goods for which client took loan are for Mobile followed by Consumer electronics and Computers.
- This shows trend that most no of loans are taken for the electronics products and which tells us where bank market could flourish.



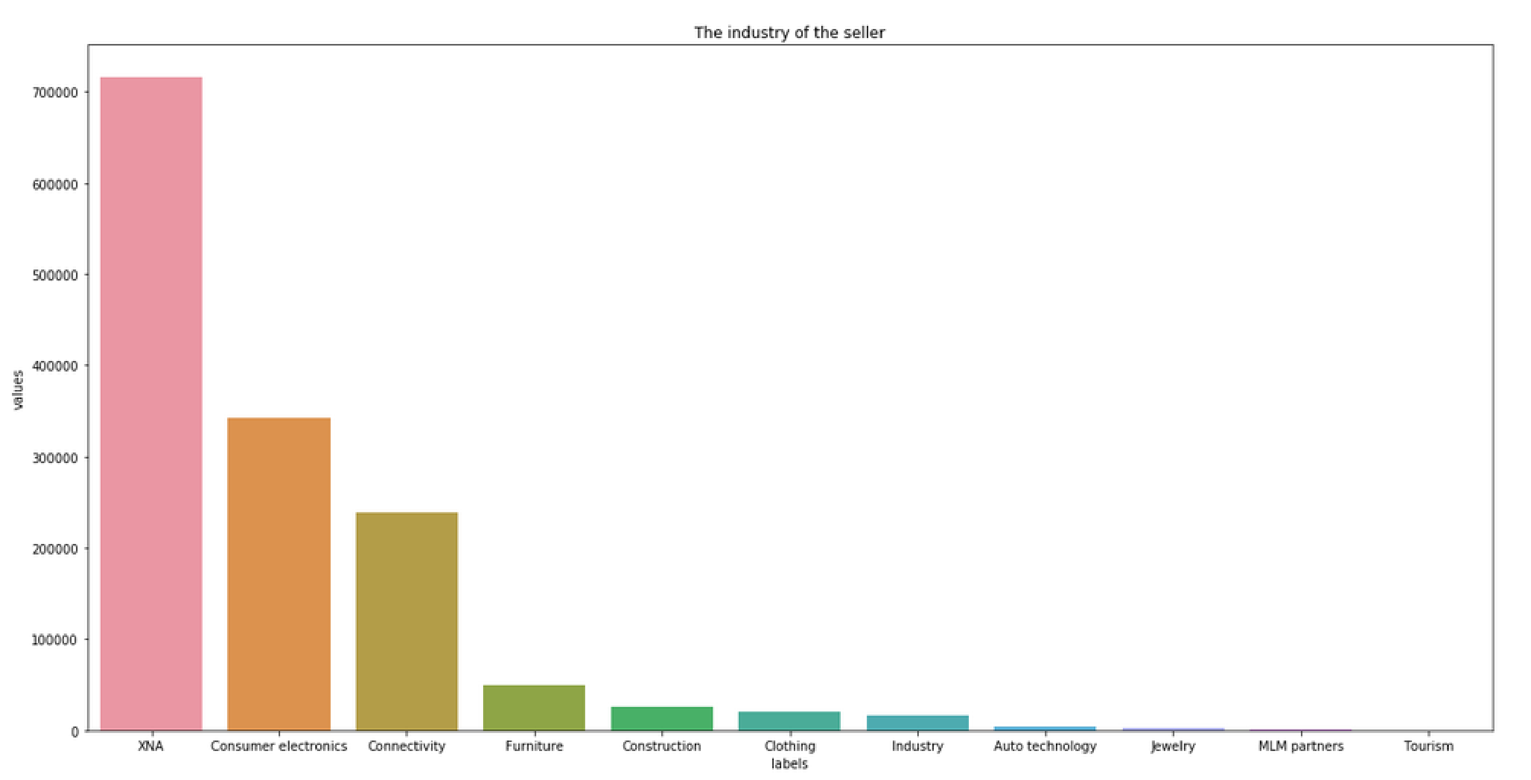
- This tells us that most of the previous application are for 'POS' followed by 'Cash'.



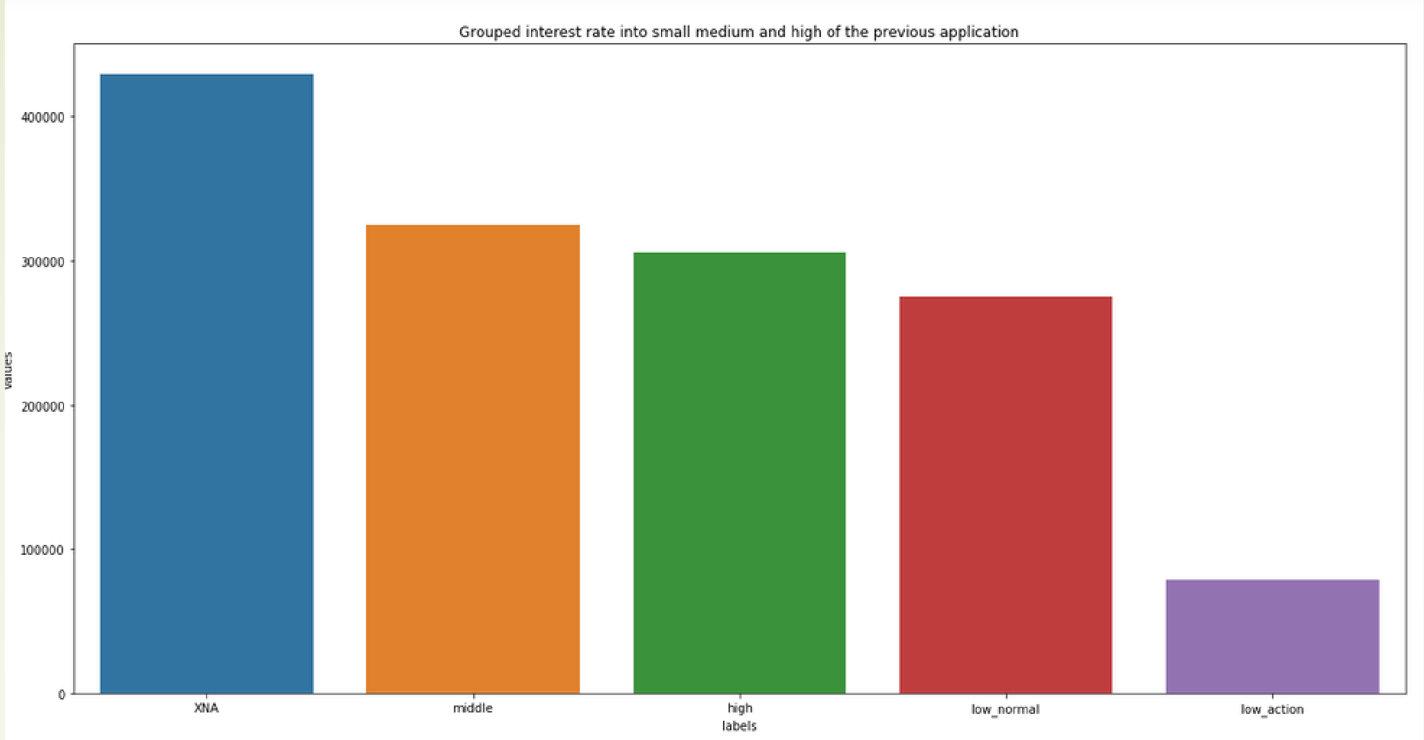
- This tells us nearly twice the application of previous application are for 'x-sell' than of 'walk-in'



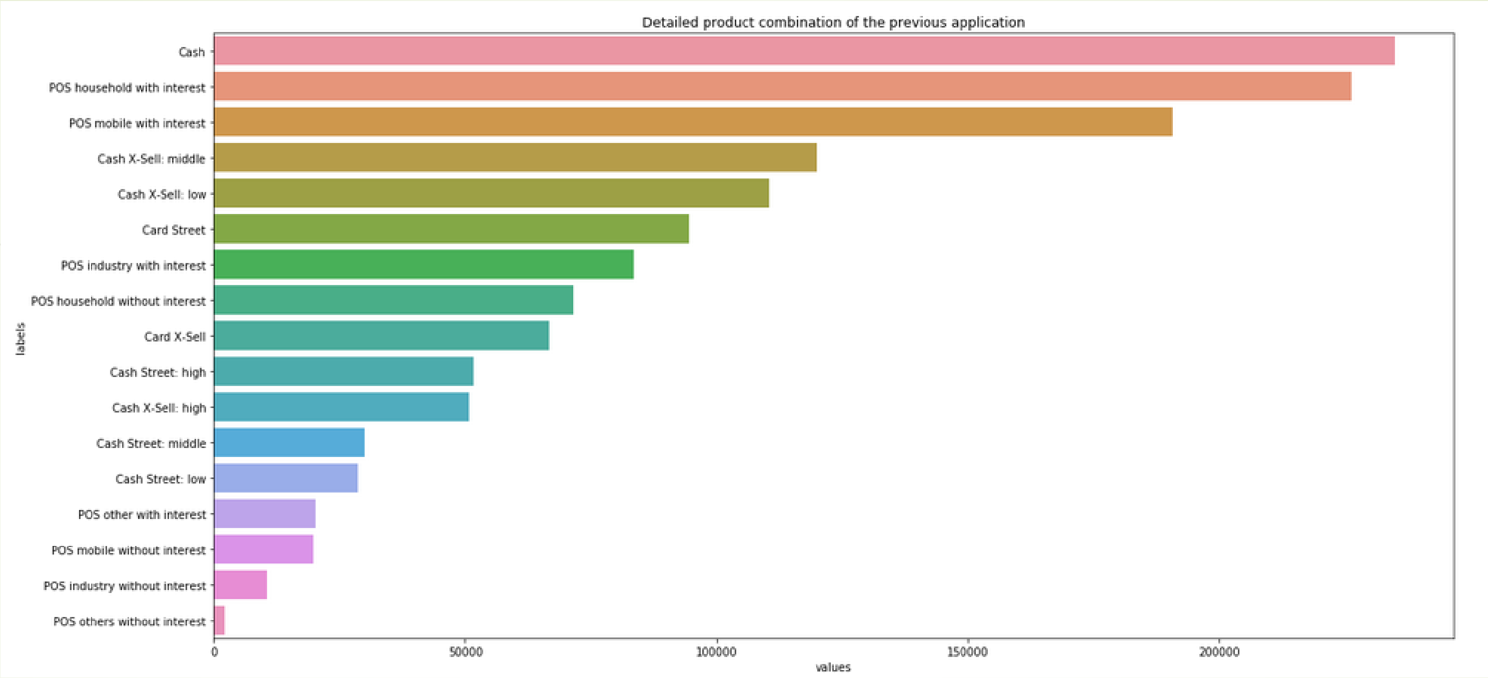
- Most of the client for previous application are acquired through 'Credit and Cash offices' followed by 'Country-wide' and least from 'Car dealer'.
- This tells us where to focus on in future for the market.



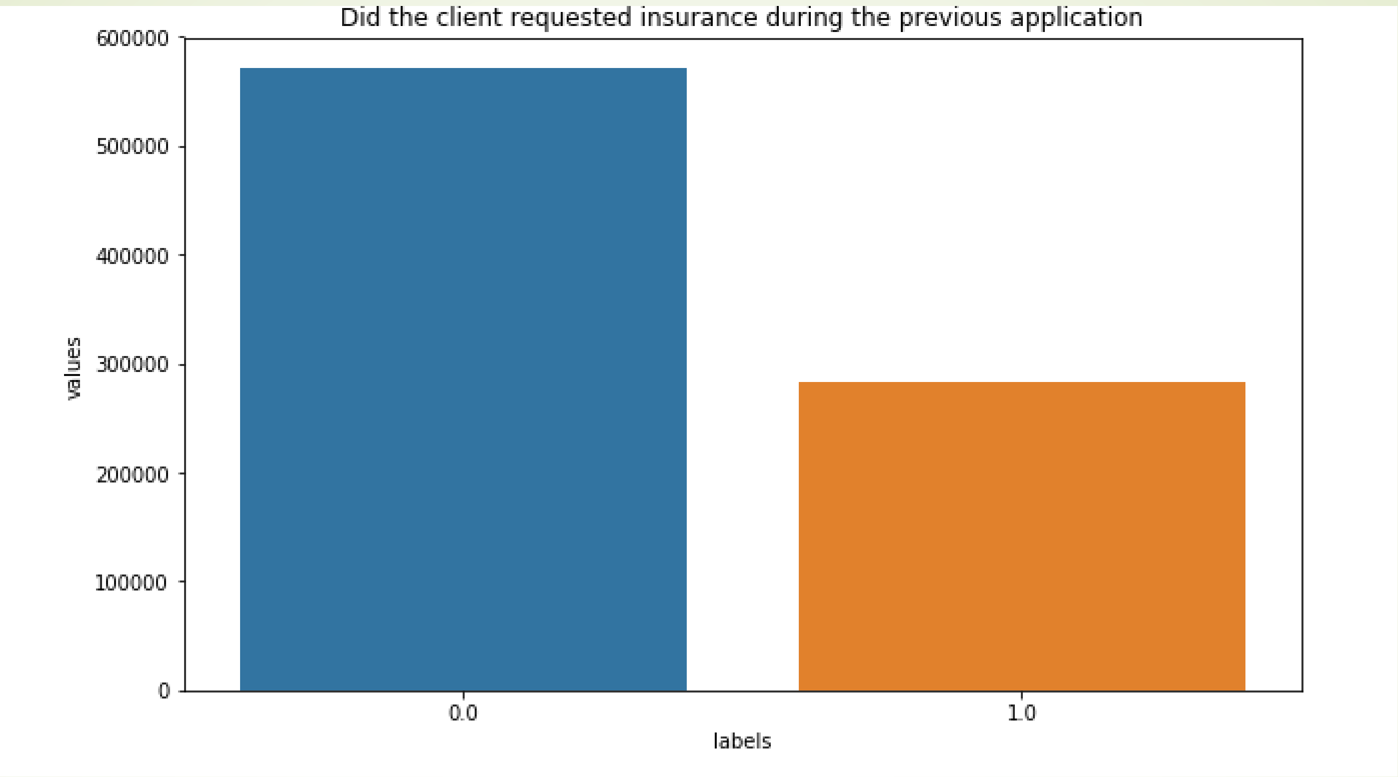
- In previous application most of the clients industry was 'Consumer electronics' followed by 'Connectivity' and the least one was 'Tourism'.
- This tells us in which industry we should focus more to get the more Clients.



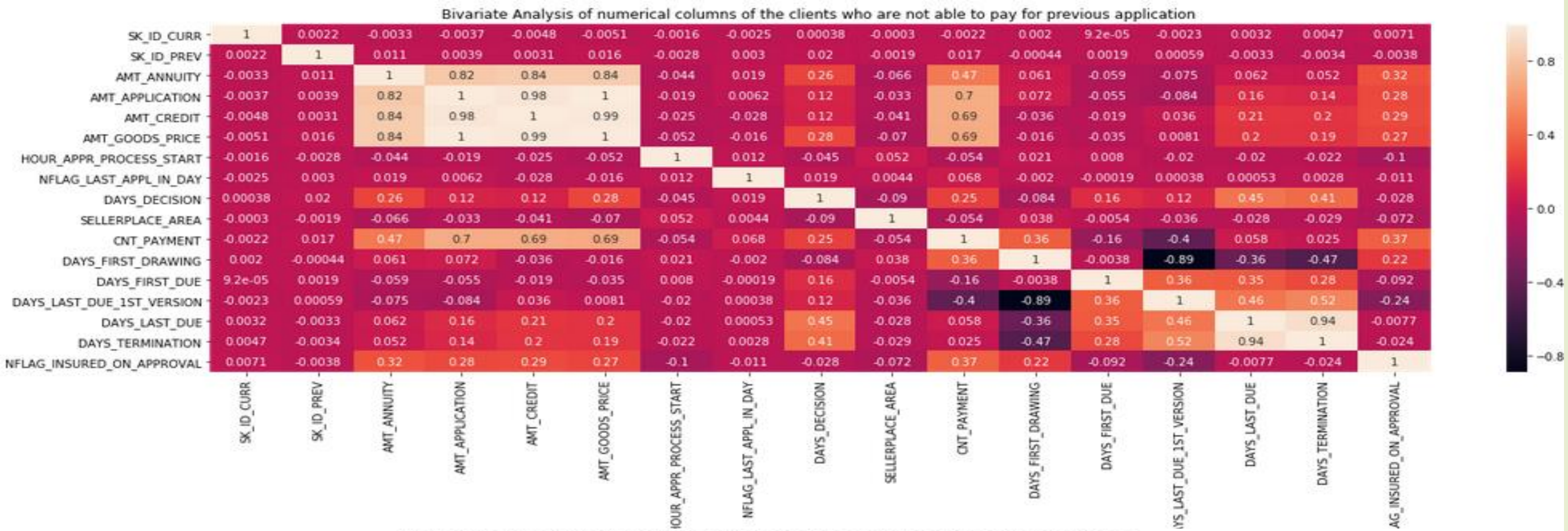
- Most the prefious applicaiton has medium interest rate followed by high interst rate, and the least one was low_action interest rate.



- For the Detailed product combination of the previous application the most of it is in the form of 'cash' followed by 'POS household with interest'.

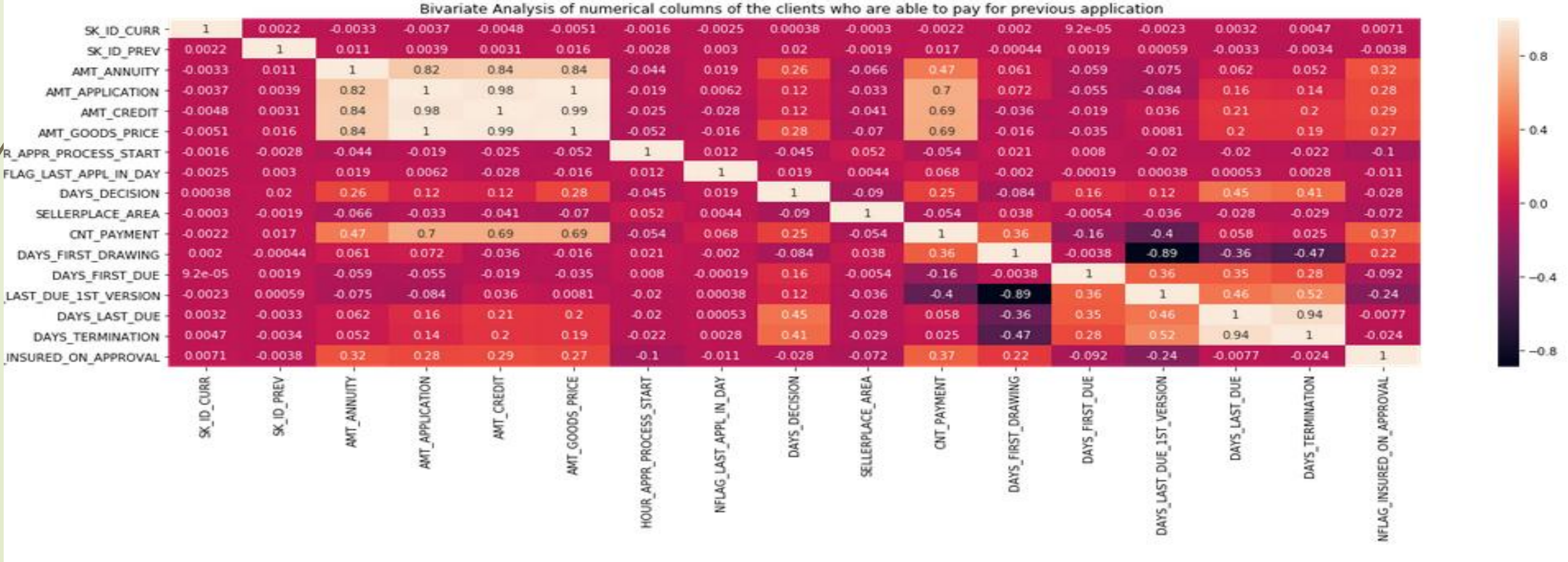


- Nearly twice the number of Client do not required insurance during the previous application.
- This tells us approximately how many client would required the insurance now.



Bivariate Analysis of numerical columns of the clients who are not able to pay for previous application

	SK_ID_CURR	SK_ID_PREV	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_GOODS_PRICE	IOUR_APPR_PROCESS_START	NFLAG_LAST_APPL_IN_DAY	DAYS_DECISION	SELLERPLACE_AREA	CNT_PAYMENT	DAYS_FIRST_DRAWING	DAYS_FIRST_DUE	YS_LAST_DUE_1ST_VERSION	DAYS_LAST_DUE	DAYS_TERMINATION	AG_INSURED_ON_APPROVAL
SK_ID_CURR	1	0.0022	-0.0033	-0.0037	-0.0048	-0.0051	-0.0016	-0.0025	0.00038	-0.0003	-0.0022	0.002	9.2e-05	-0.0023	0.0032	0.0047	0.0071
SK_ID_PREV	0.0022	1	0.011	0.0039	0.0031	0.016	-0.0028	0.003	0.02	-0.0019	0.017	-0.00044	0.0019	0.00059	-0.0033	-0.0034	-0.0038
AMT_ANNUITY	-0.0033	0.011	1	0.82	0.84	0.84	-0.044	0.019	0.26	-0.066	0.47	0.061	-0.059	-0.075	0.062	0.052	0.32
AMT_APPLICATION	-0.0037	0.0039	0.82	1	0.98	1	-0.019	0.0062	0.12	-0.033	0.7	0.072	-0.055	-0.084	0.16	0.14	0.28
AMT_CREDIT	-0.0048	0.0031	0.84	0.98	1	0.99	-0.025	-0.028	0.12	-0.041	0.69	-0.036	-0.019	0.036	0.21	0.2	0.29
AMT_GOODS_PRICE	-0.0051	0.016	0.84	1	0.99	1	-0.052	-0.016	0.28	-0.07	0.69	-0.016	-0.035	0.0081	0.2	0.19	0.27
HOUR_APPR_PROCESS_START	-0.0016	-0.0028	-0.044	-0.019	-0.025	-0.052	1	0.012	-0.045	0.052	-0.054	0.021	0.008	-0.02	-0.02	-0.022	-0.1
NFLAG_LAST_APPL_IN_DAY	-0.0025	0.003	0.019	0.0062	-0.028	-0.016	0.012	1	0.019	0.0044	0.068	-0.002	-0.00019	0.00038	0.00053	0.0028	-0.011
DAYS_DECISION	0.00038	0.02	0.26	0.12	0.12	0.28	-0.045	0.019	1	-0.09	0.25	-0.084	0.16	0.12	0.45	0.41	-0.028
SELLERPLACE_AREA	-0.0003	-0.0019	-0.066	-0.033	-0.041	-0.07	0.052	0.0044	-0.09	1	-0.054	0.038	-0.0054	-0.036	-0.028	-0.029	-0.072
CNT_PAYMENT	-0.0022	0.017	0.47	0.7	0.69	0.69	-0.054	0.068	0.25	-0.054	1	0.36	-0.16	-0.4	0.058	0.025	0.37
DAYS_FIRST_DRAWING	0.002	-0.00044	0.061	0.072	-0.036	-0.016	0.021	-0.002	-0.084	0.038	0.36	1	-0.0038	-0.89	-0.36	-0.47	0.22
DAYS_FIRST_DUE	9.2e-05	0.0019	-0.059	-0.055	-0.019	-0.035	0.008	-0.00019	0.16	-0.0054	-0.16	-0.0038	1	0.36	0.35	0.28	-0.092
DAYS_LAST_DUE_1ST_VERSION	-0.0023	0.00059	-0.075	-0.084	0.036	0.0081	-0.02	0.00038	0.12	-0.036	-0.4	-0.89	0.36	1	0.46	0.52	-0.24
DAYS_LAST_DUE	0.0032	-0.0033	0.062	0.16	0.21	0.2	-0.02	0.00053	0.45	-0.028	0.058	-0.36	0.35	0.46	1	0.94	-0.0077
DAYS_TERMINATION	0.0047	-0.0034	0.052	0.14	0.2	0.19	-0.022	0.0028	0.41	-0.029	0.025	-0.47	0.28	0.52	0.94	1	-0.024
NFLAG_INSURED_ON_APPROVAL	0.0071	-0.0038	0.32	0.28	0.29	0.27	-0.1	-0.011	-0.028	-0.072	0.37	0.22	-0.092	-0.24	-0.0077	-0.024	1



Bivariate Analysis of numerical columns of the clients who are able to pay for previous application

	SK_ID_CURR	SK_ID_PREV	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_GOODS_PRICE	HOUR_APPR_PROCESS_START	NFLAG_LAST_APPL_IN_DAY	DAYS_DECISION	SELLERPLACE_AREA	CNT_PAYMENT	DAYS_FIRST_DRAWING	DAYS_FIRST_DUE	DAYS_LAST_DUE_1ST_VERSION	DAYS_LAST_DUE	DAYS_TERMINATION	NFLAG_INSURED_ON_APPROVAL
SK_ID_CURR	1	0.0022	-0.0033	-0.0037	-0.0048	-0.0051	-0.0016	-0.0025	0.00038	-0.0003	-0.0022	0.002	9.2e-05	-0.0023	0.0032	0.0047	0.0071
SK_ID_PREV	0.0022	1	0.011	0.0039	0.0031	0.016	-0.0028	0.003	0.02	-0.0019	0.017	-0.00044	0.0019	0.00059	-0.0033	-0.0034	-0.0038
AMT_ANNUITY	-0.0033	0.011	1	0.82	0.84	0.84	-0.044	0.019	0.26	-0.066	0.47	0.061	-0.059	-0.075	0.062	0.052	0.32
AMT_APPLICATION	-0.0037	0.0039	0.82	1	0.98	1	-0.019	0.0062	0.12	-0.033	0.7	0.072	-0.055	-0.084	0.16	0.14	0.28
AMT_CREDIT	-0.0048	0.0031	0.84	0.98	1	0.99	-0.025	-0.028	0.12	-0.041	0.69	-0.036	-0.019	0.036	0.21	0.2	0.29
AMT_GOODS_PRICE	-0.0051	0.016	0.84	1	0.99	1	-0.052	-0.016	0.28	-0.07	0.69	-0.016	-0.035	0.0081	0.2	0.19	0.27
R_APPR_PROCESS_START	-0.0016	-0.0028	-0.044	-0.019	-0.025	-0.052	1	0.012	-0.045	0.052	-0.054	0.021	0.008	-0.02	-0.02	-0.022	-0.1
FLAG_LAST_APPL_IN_DAY	-0.0025	0.003	0.019	0.0062	-0.028	-0.016	0.012	1	0.019	0.0044	0.068	-0.002	-0.00019	0.00038	0.00053	0.0028	-0.011
DAYS_DECISION	0.00038	0.02	0.26	0.12	0.12	0.28	-0.045	0.019	1	-0.09	0.25	-0.084	0.16	0.12	0.45	0.41	-0.028
SELLERPLACE_AREA	-0.0003	-0.0019	-0.066	-0.033	-0.041	-0.07	0.052	0.0044	-0.09	1	-0.054	0.038	-0.0054	-0.036	-0.028	-0.029	-0.072
CNT_PAYMENT	-0.0022	0.017	0.47	0.7	0.69	0.69	-0.054	0.068	0.25	-0.054	1	0.36	-0.16	-0.4	0.058	0.025	0.37
DAYS_FIRST_DRAWING	0.002	-0.00044	0.061	0.072	-0.036	-0.016	0.021	-0.002	-0.084	0.038	0.36	1	-0.0038	-0.89	-0.36	-0.47	0.22
DAYS_FIRST_DUE	9.2e-05	0.0019	-0.059	-0.055	-0.019	-0.035	0.008	-0.00019	0.16	-0.0054	-0.16	-0.0038	1	0.36	0.35	0.28	-0.092
LAST_DUE_1ST_VERSION	-0.0023	0.00059	-0.075	-0.084	0.036	0.0081	-0.02	0.00038	0.12	-0.036	-0.4	-0.89	0.36	1	0.46	0.52	-0.24
DAYS_LAST_DUE	0.0032	-0.0033	0.062	0.16	0.21	0.2	-0.02	0.00053	0.45	-0.028	0.058	-0.36	0.35	0.46	1	0.94	-0.0077
DAYS_TERMINATION	0.0047	-0.0034	0.052	0.14	0.2	0.19	-0.022	0.0028	0.41	-0.029	0.025	-0.47	0.28	0.52	0.94	1	-0.024
_INSURED_ON_APPROVAL	0.0071	-0.0038	0.32	0.28	0.29	0.27	-0.1	-0.011	-0.028	-0.072	0.37	0.22	-0.092	-0.24	-0.0077	-0.024	1

- - Here we could see that AMT_ANNUITY and AMT_CREDIT are highly correlated, if one goes up so s the other and vise versa.
- - For both the sets but much more correlated for the Clients who are able to pay.
- - We could also see that AMT_CREDIT and AMT_GOODS_PRICE are also highly correlated, if one goes up so s the other and vise versa.
- - For both the sets but are more correlated to the Clients who are able to pay.
- - Similarly AMT_ANNUITY and AMT_GOODS_PRICE are also highly correlated, if one goes up so s the other and vise versa.
- - For both the sets but are more correlated to the Clients who are able to pay.
- - For AMT_ANNUITY and AMT_INCOME_TOTAL the correlation is much higher for the Clients who are able to pay.
- - Similar is the case with AMT_CREDIT and AMT_INCOME_TOAL.
- - We can see that there are group of value have a similar correlation AMT_ANNUITY,AMT_APPLICATION,AMT_CREDIT,AMT_GOODS_PRICE for a segment of values with highly dependent factors.
- - In the similar manner there are a lot for columns which are corrected for the case the Clients who are able to pay and the Clients who are not able to pay.
- - DAYS_LAST_DUE_1ST_VERSION and DAYS_FIRST_DRAWING are highly correlated negatively. If one goes up other goes down and vise versa.

THANK YOU