Dataset:

This dataset represents attributes that describe credit card holders, along with a label that will define them as a defaulter or non-defaulter (loan paid or not).

Preprocessing:

For preprocessing, we first identified duplicate rows of individuals in the dataset and removed them. We identified 9920 duplicate rows of values in the dataset.

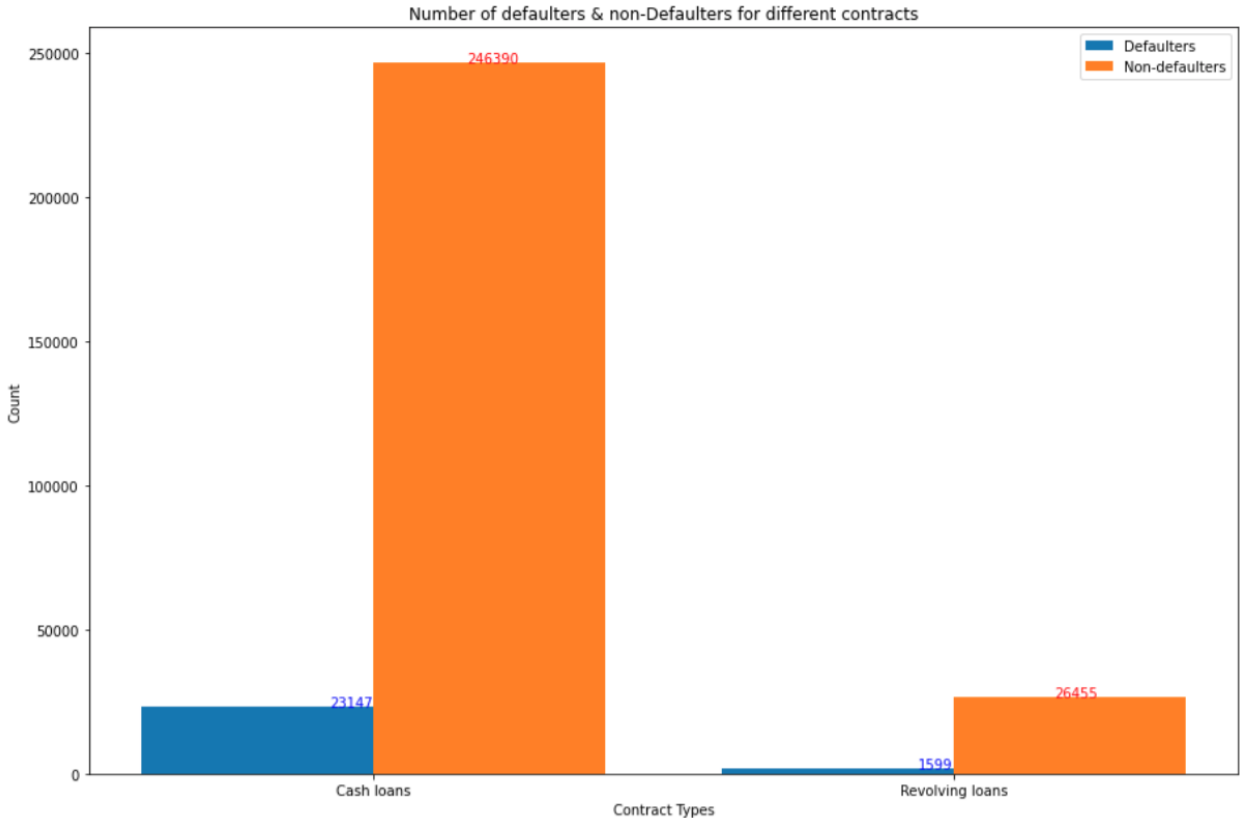
Next, we analyzed the dataset for missing values (NaN values). We were able to find a total of 88215 missing values. To cater to missing values, we decided to fill them. For the values belonging to numerical column types like credit and total income, we calculated the respective mean of the numerical columns and substituted the missing values with it. Similarly, for categorical column types like education type, family status, occupation type, etc; we fill missing values using the mode value of respective columns.

After that, we normalized our data. In particular, we normalized credit and total income columns, since they were the only continuous and numerical data we had. Normalization allows our numerical data to be given equal weightage/importance and to also ensure that a certain variable does not affect our analysis methods due it having large values.

Bar Graphs:

Bar graphs were added to get better visual insight to our data. The bar graphs were constructed between different attributes to gain some meaning

Number of defaulters and non defaulters for different contract types:



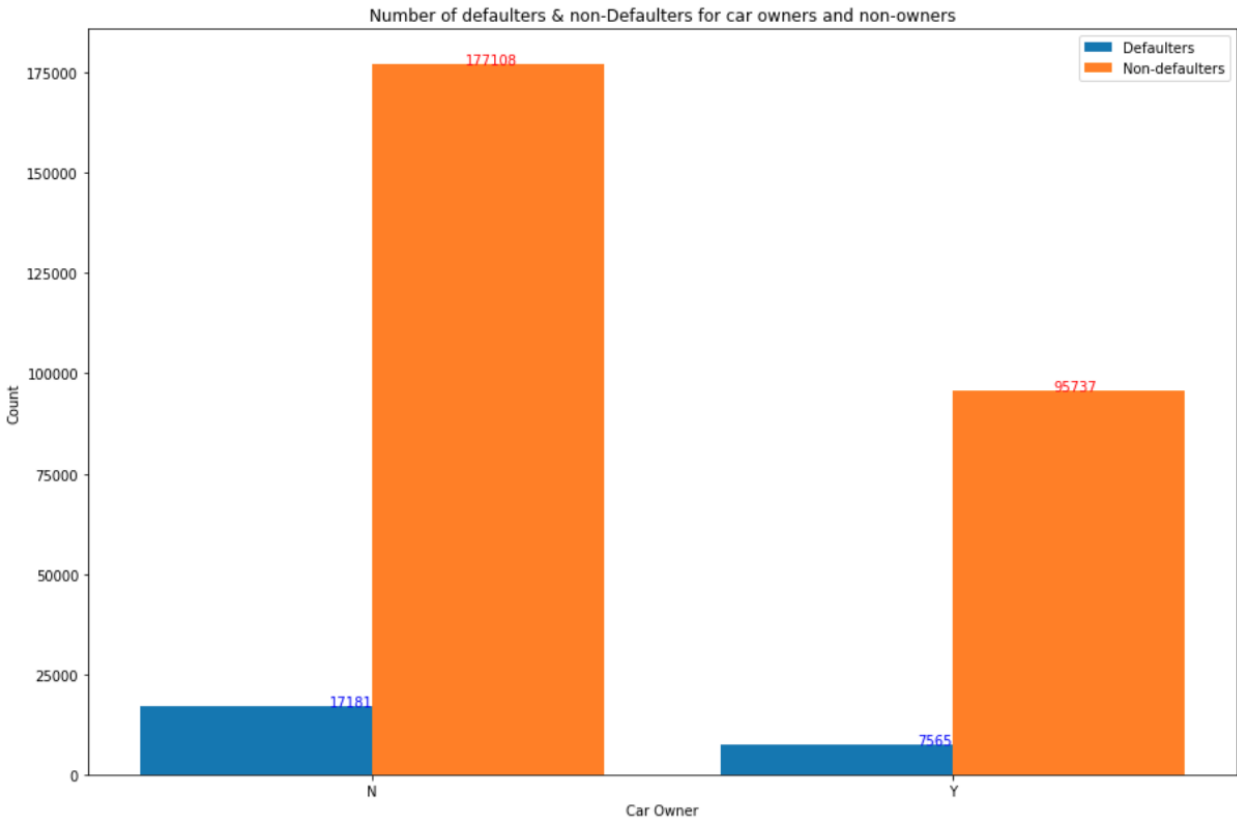
The above bar graph shows the number of defaulters and non-defaulters for the different contract types.

If we calculate the defaulter & non-defaulter ratios:
Cash loans: 23147/246390 = 0.094
Revolving loans: 1599/26455 = 0.060

We can assume that people who took cash loans had a higher ratio than those who took revolving loans. We can also assume that most of the defaulters and non defaulters both chose the option of cash loans as compared to revolving loans.

Suggestions:
Since people with revolving loans have a lesser number of defaulters, the bank could keep a check on people who took cash loans. With cash loans there would be a greater chance that a person is a defaulter.

Number of defaulters and non defaulters for car owners and non owners:



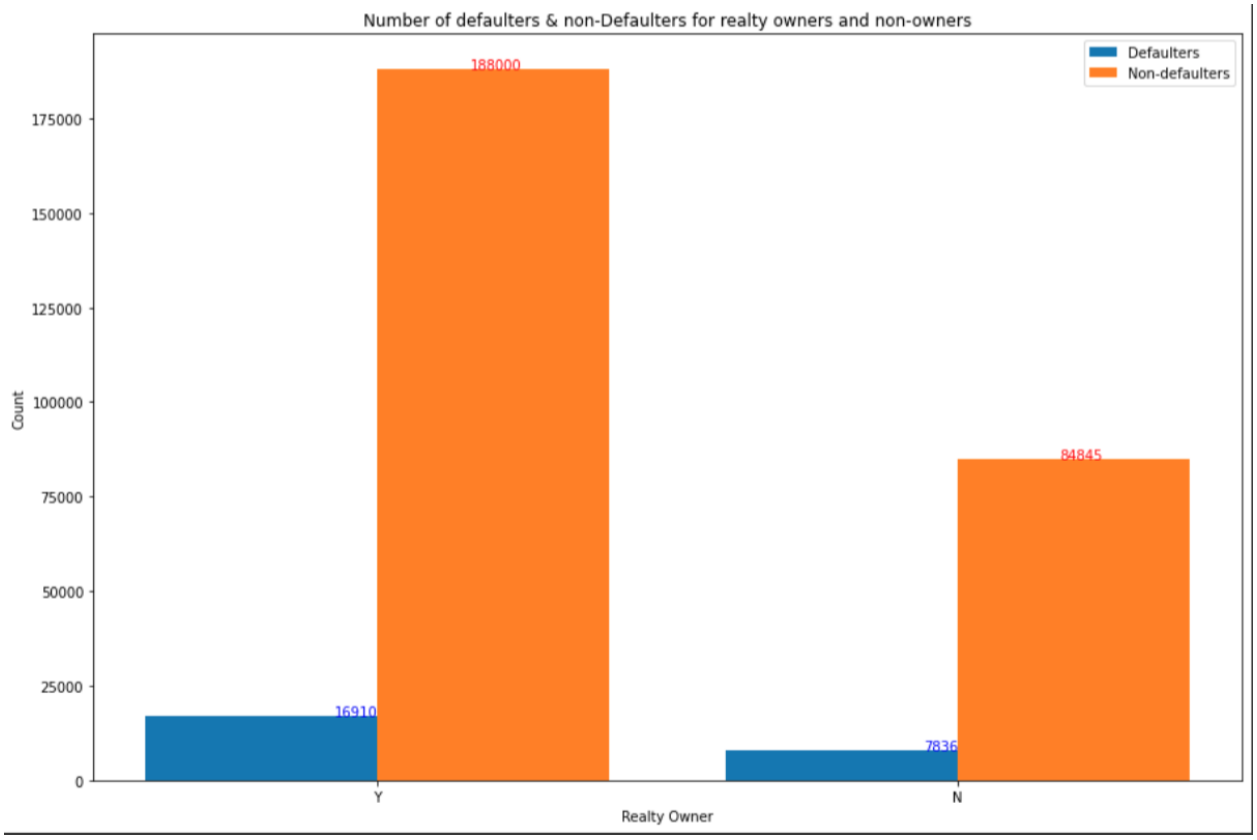
The above bar graph shows how many car owners and non-car owners were defaulters and non-defaulters.

If we calculate the defaulter & non-defaulter ratios:
Car owner: 17181/177108 = 0.097
Non-car owners: 7565/95737 = 0.079

We can assume that for both defaulter and non defaulters, most of them did not own a car while few did own a car.

Suggestions:
The defaulters are twice likely to not own a car hence the bank could keep a check on those who do not own a car since a small percentage of these would be more likely to be a defaulter.

Number of defaulters and non defaulters for realty owners and non realty owners:



The above bar graph shows how many realty owners and non-realty owners were defaulters and non-defaulters.

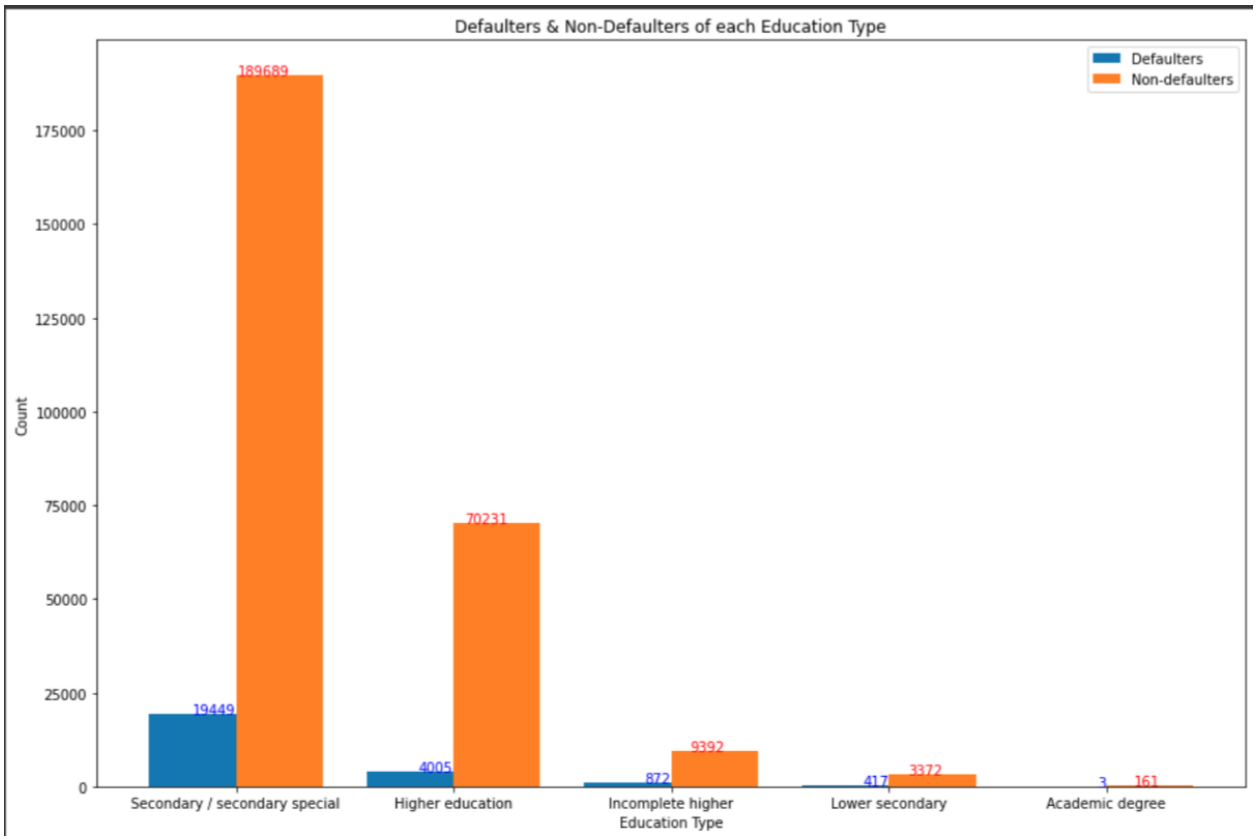
If we calculate the defaulter & non-defaulter ratio:
Realty owner: 16910/188000 = 0.0899
Non-Realty owners: 7836/84845 = 0.0924

The figure shows that a large number of non defaulters do own realty, however approximately double the amount of people own realty than those who do not.

We can determine that owning or not owning realty will not make a big difference whether someone is a defaulter or non-defaulter.

Suggestions:
The defaulters are twice likely to not own realty hence the bank could keep a check on those who do not own realty since a small percentage of these would be more likely to be a defaulter.

Number of defaulters and non defaulters of each education type:



The above graph shows the number of defaulters and non-defaulters with different educational backgrounds.

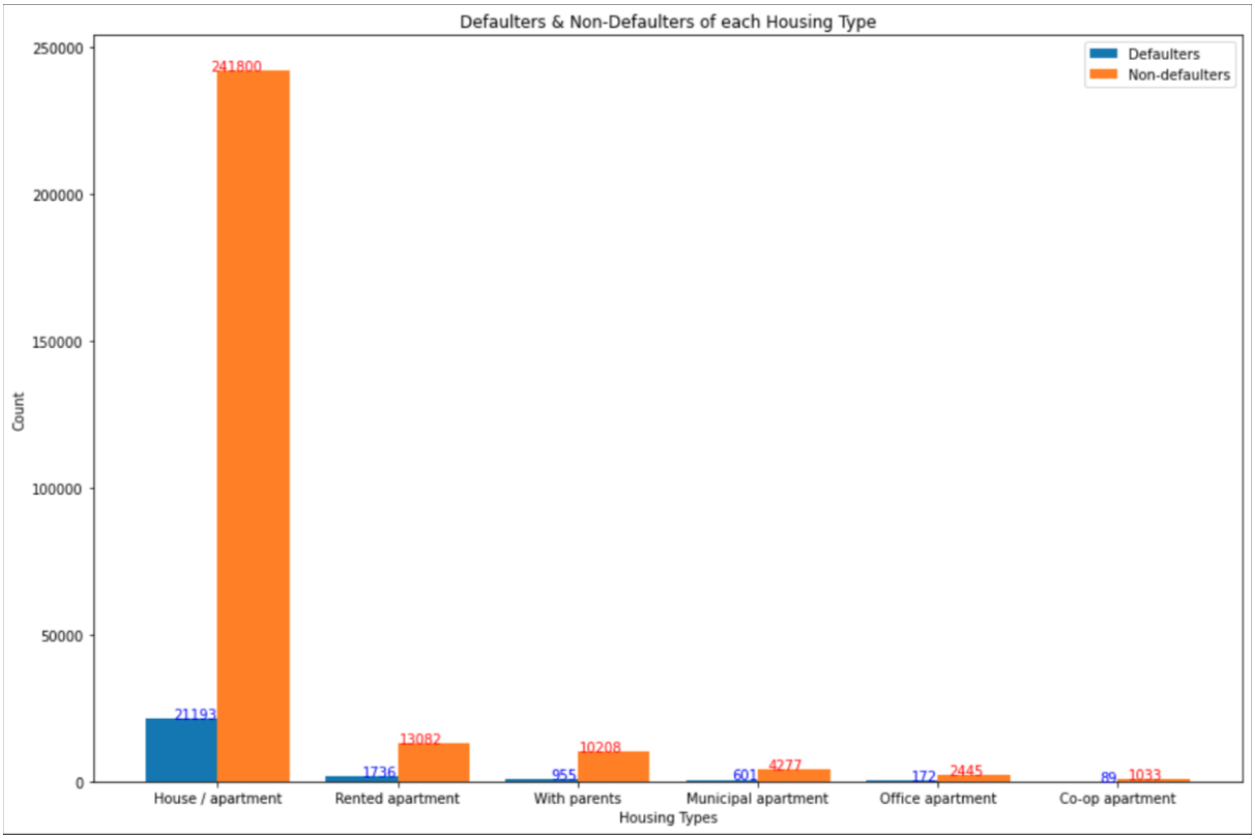
If we calculate the defaulter to non-defaulter ratios for each education type:
Secondary: 19449/189689 = 0.102
Higher education: 4005/70231 = 0.057
Incomplete higher education: 872/9392 = 0.093
Lower secondary: 417/3372 = 417/3372 = 0.123
Academic degree: 3/161 = 0.019

We can see that for both defaulters and non defaulters a large number of people have secondary education as an education type. We can see that Academic degree holders have the lowest defaulter ratio meaning they are least likely to default.

Suggestions:

The defaulters are more likely to gave education type as secondary or higher education as compare to an academic degree hence the bank could regulate the customers with these as education type.

Number of defaulters and non defaulters of each housing type:



The above graphs shows the number of defaulters and non-defaulters with different housing types.

If we calculate the defaulter to non-defaulter ratios for each housing type:
House/apartment: 21193/241800 = 0.088
Rented apartment: 1736/13082 = 0.133
With parents: 955/10208 = 0.094
Municipal apartment: 601/4277 = 0.141
Office apartment: 172/2445 = 0.070
Co-op apartment: 89/1033 = 0.086

We can see that people living in rented and municipal departments have the respective 0.133 and 0.141 ratios and tend to default more.

Suggestions:
The bank can keep a check on people who have a house/apartment as housing type. From this category there is a larger possibility of a customer being a defaulter than having a rented apartment, living with parents, having a municipal, office or co-op apartment.

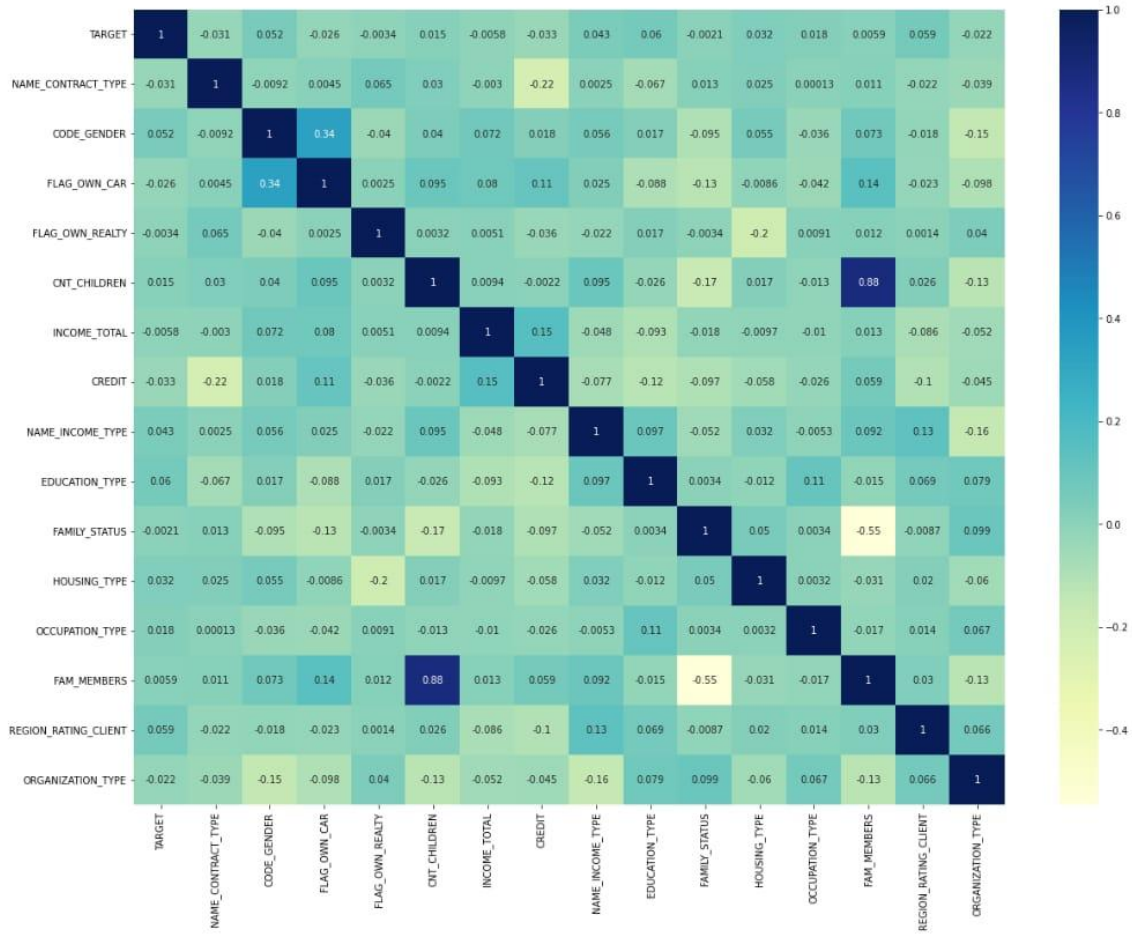
Correlation and Covariance:

Correlation tells us about the extent to which two variables are related. There are three possible results of a correlational study: a positive correlation, a negative correlation, and no correlation.

A **positive correlation** is a relationship between two variables in which both variables move in the same direction. Therefore, when one variable increases as the other variable increases, or one variable decreases while the other decreases. A **negative correlation** is a relationship between two variables in which an increase in one variable is associated with a decrease in the other. A **zero correlation** exists when there is no relationship between two variables.

Covariance also measures the linear relationship between two variables. The covariance is similar to the correlation between two variables, but there is a slight difference. Correlation coefficients are standardized while covariance coefficients are not standardized.

In our case some features were positively correlated while some were negatively correlated. E.g. car owner and gender code were highly positively correlated , number of family members and children count were highly positively correlated, number of family members and family status were highly negatively correlated. Same was the case with covariance. Some had negative covariance while some had positive covariance. Pictures are attached for reference.



The above figure shows a heatmap of the correlation between attributes. Darker shades of colors represent a highly positive correlation whereas a very light shade represents a highly negative correlation. And an adequate shade of color shows poor correlation. This figure gives us valuable information to predict future defaulters as explained below.

Suggestions:

The **covariance** was a useful method to detect defaulters and non-defaulters. For example, total income and credit had negative covariance which meant that higher a higher income and higher credit in the bank means that the person is non-defaulter. If the person has lower total income and higher credit in the bank, it means that the person is a defaulter. Using this way, a bank can easily detect defaulters. Therefore, we suggest that the banks avoid giving loans to people with low total incomes and high credits since they are likely to default.

Similarly, **Correlation** also helps to identify defaulters and non-defaulters. For example, Family members and family status of a person are highly negatively correlated. If a person has higher family members his family status must be low. If a person has less family members his

family status will be high. So, if a person with larger family members has high family status, he is most likely to be a defaulter. In this way, banks can easily find defaulters.

Chi Square Test:

```
Chi square value and p value between Target and region rating:
1042.637960827036 3.9268426110928483e-227

Chi square value and p value between Target and family members:
124.08121678220304 8.98332116699232e-19

Chi square value and p value between Target and number of children:
128.92149775473266 1.1096559879703614e-20

Chi square value and p value between Target and type of occupation:
1188.490639393036 3.851169773576799e-242
```

A chi square test was carried out to determine whether all the attributes had a meaningful relationship with the target value of either 1 or 0 (defaulter or non-defaulter). A large chi square value would imply there is a significant relationship between the two variables. For the above results since the chi square value is pretty large and the value of p is greater than the assumed significant value we can assume there is a significant relationship between Target and region rating, target and family members, target and number of children, target and the type of occupation. The attribute target(defaulter or non defaulter) depends on the region rating of the client, number of family members, number of children and type of occupation.

Suggestion:

The chi square test is only giving us an idea about what kind of attributes the target variable is depending on. It does not exactly allow us to predict what kind of individual with specific attribute values will default or not. But, it can let banks determine which attributes are significant in determining if an individual will default or not. For example, region rating of a client, number of family members, number of children, and type of occupation are the attributes that the target variable depends on. This means that banks can pay attention to these attributes the most.

Frequent Itemset Mining:

For frequent itemset mining the apriori algorithm was used. The goal was to find out associations between different attributes. We check different combinations of attributes that result in either a defaulter or non-defaulter.

Apriori algorithm uses frequent itemsets to generate association rules which can be filtered out for better analysis of the results. Some of the continuous data was discretized into categories for the apriori algorithm. These included ‘Credit’, ‘Total Income’, ‘Number of Children’, ‘Number of family members’. The apriori algorithm uses support, confidence and lift to make association rules. **Support** is the relative frequency that a rule shows up (or in simple terms a value that allows us to exclude a certain combination of attributes based on their appearance in our dataset). When we say **rule** we mean itemsets or more simply combinations of different attribute values. **Confidence** gives the reliability of a rule (or in simple terms the probability of a certain combination of attributes occurring one after the other). **Lift** is the ratio of observed support to the expected support. Running the algorithm gives a table of association rules. From these association rules we can determine the strong rules by checking the lift and support value for each rule.

First we check which rules apply to non defaulters. The rules were made with target type as a target hence only the rules which had ‘TARGET’ as consequent were added. This made it easier to differentiate the rules for a defaulter and non defaulter. Moreover the rules with a high confidence and support value were considered.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
468	(HOUSING_TYPE_House / apartment, EDUCATION_TYP...	(TARGET_0)	0.222581	0.916846	0.210897	0.947507	1.033443	0.006825	1.584119
17	(EDUCATION_TYPE_Higher education)	(TARGET_0)	0.249456	0.916846	0.235998	0.946050	1.031854	0.007285	1.541336
399	(EDUCATION_TYPE_Higher education, CNT_CHILDREN...	(TARGET_0)	0.222140	0.916846	0.209990	0.945301	1.031036	0.006321	1.520212
212	(NAME_CONTRACT_TYPE_Cash loans, EDUCATION_TYPE...	(TARGET_0)	0.217839	0.916846	0.205440	0.943079	1.028613	0.005715	1.460881
450	(HOUSING_TYPE_House / apartment, CREDIT_RANGE_...	(TARGET_0)	0.224983	0.916846	0.211085	0.938225	1.023319	0.004810	1.346092
1973	(HOUSING_TYPE_House / apartment, CREDIT_RANGE_...	(TARGET_0)	0.220763	0.916846	0.206925	0.937318	1.022329	0.004520	1.326613
395	(CREDIT_RANGE_810,000 < Credit <=4,050,000, CN...	(TARGET_0)	0.222231	0.916846	0.208286	0.937249	1.022253	0.004534	1.325140
13	(CREDIT_RANGE_810,000 < Credit <=4,050,000)	(TARGET_0)	0.248600	0.916846	0.232887	0.936795	1.021758	0.004959	1.315621
1882	(NAME_CONTRACT_TYPE_Cash loans, CREDIT_RANGE_8...	(TARGET_0)	0.218091	0.916846	0.204203	0.936319	1.021240	0.004247	1.305804
208	(NAME_CONTRACT_TYPE_Cash loans, CREDIT_RANGE_8...	(TARGET_0)	0.244036	0.916846	0.228391	0.935888	1.020770	0.004647	1.297019
2542	(CODE_GENDER_F, FAMILY_STATUS_Married, FAM_MEM...	(TARGET_0)	0.257041	0.916846	0.240068	0.933968	1.018675	0.004401	1.259301
2565	(CODE_GENDER_F, HOUSING_TYPE_House / apartment...	(TARGET_0)	0.301918	0.916846	0.281524	0.932453	1.017022	0.004712	1.231052
2887	(FLAG_OWN_CAR_Y, FAMILY_STATUS_Married, HOUSIN...	(TARGET_0)	0.232568	0.916846	0.216858	0.932452	1.017022	0.003630	1.231041
2597	(CODE_GENDER_F, FAM_MEMBERS_2.0, REGION_RATING...	(TARGET_0)	0.245747	0.916846	0.229049	0.932054	1.016588	0.003738	1.223837
2359	(CODE_GENDER_F, FAM_MEMBERS_2.0, FLAG_OWN_REAL...	(TARGET_0)	0.234096	0.916846	0.218128	0.931788	1.016297	0.003498	1.219055
2425	(CODE_GENDER_F, OCCUPATION_TYPE_Laborers, CNT_...	(TARGET_0)	0.277072	0.916846	0.258103	0.931538	1.016024	0.004071	1.214599
2557	(CODE_GENDER_F, HOUSING_TYPE_House / apartment...	(TARGET_0)	0.269397	0.916846	0.250938	0.931483	1.015965	0.003943	1.213633
2413	(CODE_GENDER_F, HOUSING_TYPE_House / apartment...	(TARGET_0)	0.525372	0.916846	0.489212	0.931172	1.015626	0.007527	1.208146
2408	(CODE_GENDER_F, FAMILY_STATUS_Married, CNT_CHI...	(TARGET_0)	0.340995	0.916846	0.317496	0.931088	1.015534	0.004857	1.206672
322	(FLAG_OWN_CAR_Y, FAMILY_STATUS_Married)	(TARGET_0)	0.255522	0.916846	0.237884	0.930971	1.015407	0.003609	1.204638

From the figures we can infer that **non defaulter** is most likely to have the following properties:

- Having only one child or none at all
- Have an apartment or a house for housing
- Own realty
- Be a female
- Higher education as education type
- Cash loans as contract type
- Credit Range from 810,000 to 4,050,000
- Family status as married
- Have upto 2 family members
- Own a car
- Have occupation as laborers

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
46	(NAME_INCOME_TYPE_Working)	(TARGET_1)	0.527892	0.083154	0.051053	0.096712	1.163036	0.007157	1.015009
1972	(NAME_CONTRACT_TYPE_Cash loans, EDUCATION_TYPE...	(TARGET_1)	0.645043	0.083154	0.061343	0.095098	1.143637	0.007704	1.013199
24624	(NAME_CONTRACT_TYPE_Cash loans, EDUCATION_TYPE...	(TARGET_1)	0.580014	0.083154	0.054309	0.093635	1.126034	0.006079	1.011563
48	(EDUCATION_TYPE_Secondary / secondary special)	(TARGET_1)	0.702770	0.083154	0.065355	0.092996	1.118354	0.006916	1.010851
24640	(NAME_CONTRACT_TYPE_Cash loans, EDUCATION_TYPE...	(TARGET_1)	0.572346	0.083154	0.052794	0.092241	1.109278	0.005201	1.010010
1996	(EDUCATION_TYPE_Secondary / secondary special,...	(TARGET_1)	0.630866	0.083154	0.057834	0.091675	1.102463	0.005375	1.009380
1958	(NAME_CONTRACT_TYPE_Cash loans, FLAG_OWN_CAR_N)	(TARGET_1)	0.591953	0.083154	0.053933	0.091110	1.095677	0.004710	1.008753
2000	(HOUSING_TYPE_House / apartment, EDUCATION_TYP...	(TARGET_1)	0.621968	0.083154	0.056111	0.090215	1.084905	0.004391	1.007760
44	(FLAG_OWN_CAR_N)	(TARGET_1)	0.652873	0.083154	0.057734	0.088430	1.063445	0.003444	1.005788
1986	(FLAG_OWN_CAR_N, CNT_CHILDREN_RANGE_<=1)	(TARGET_1)	0.595025	0.083154	0.051782	0.087026	1.046556	0.002304	1.004240
42	(NAME_CONTRACT_TYPE_Cash loans)	(TARGET_1)	0.905730	0.083154	0.077781	0.085877	1.032740	0.002466	1.002978
1964	(FLAG_OWN_REALTY_Y, NAME_CONTRACT_TYPE_Cash lo...	(TARGET_1)	0.614857	0.083154	0.052532	0.085437	1.027455	0.001404	1.002496
1968	(NAME_CONTRACT_TYPE_Cash loans, CNT_CHILDREN_R...	(TARGET_1)	0.812793	0.083154	0.069122	0.085042	1.022702	0.001534	1.002063
1982	(NAME_CONTRACT_TYPE_Cash loans, REGION_RATING_...	(TARGET_1)	0.664419	0.083154	0.055926	0.084172	1.012243	0.000676	1.001112
1978	(NAME_CONTRACT_TYPE_Cash loans, HOUSING_TYPE_H...	(TARGET_1)	0.802198	0.083154	0.066759	0.083221	1.000796	0.000053	1.000072

From the figures we can infer that **a defaulter** is most likely to have the following properties:

- Income type as working
- Cash loans as contract type
- Secondary education as education type
- Do not own a car
- Have an apartment or a house for housing
- Owns realty

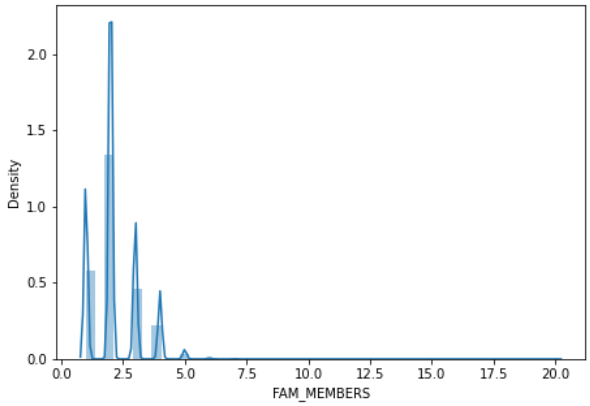
From the above properties we can make assumptions about whether a loan would turn out to be a defaulter or non defaulter. Most of the properties are common to both, such as owning an apartment or house, owning realty.

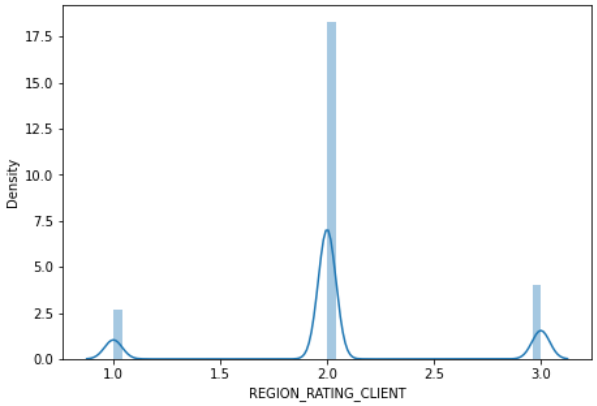
Suggestions:

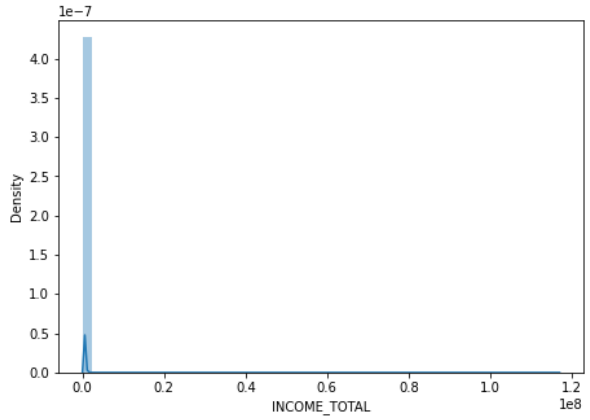
The bank can keep track of the above properties for each credit card holder. If a credit card holder has the common properties of a non defaulter such as their gender is female, they have a credit range from 810,000 to 4,050,000, own a car, have higher education and have occupation as laborers then it is safe to assume that they would be non defaulters. However if there is any deviation from these properties or they have an income type of working, secondary education, do not own a car then the bank should be more cautious when giving out loans as the holders might turn out to be a defaulter.

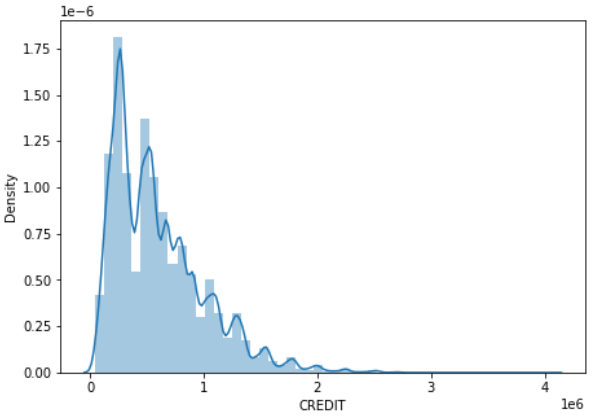
Outlier Analysis:

An **outlier** is an element of a data set that distinctly stands out from the rest of the data. In other words, outliers are those data points that lie outside the overall pattern of distribution. Outlier analysis is a process of analyzing outliers in the dataset. The easiest way to see the outliers in the dataset is by drawing graphs. Attached below are a few graphs from out dataset to understand this technique.









We used the z-score approach to analyze outliers in the dataset. In this approach, we find the mean and standard deviation of data. Anything below the three standard deviations from mean or above three standard deviations of mean is considered an outlier. We analyzed outliers of numerical data only because there is no way of finding outliers in categorical data.

Suggestions:

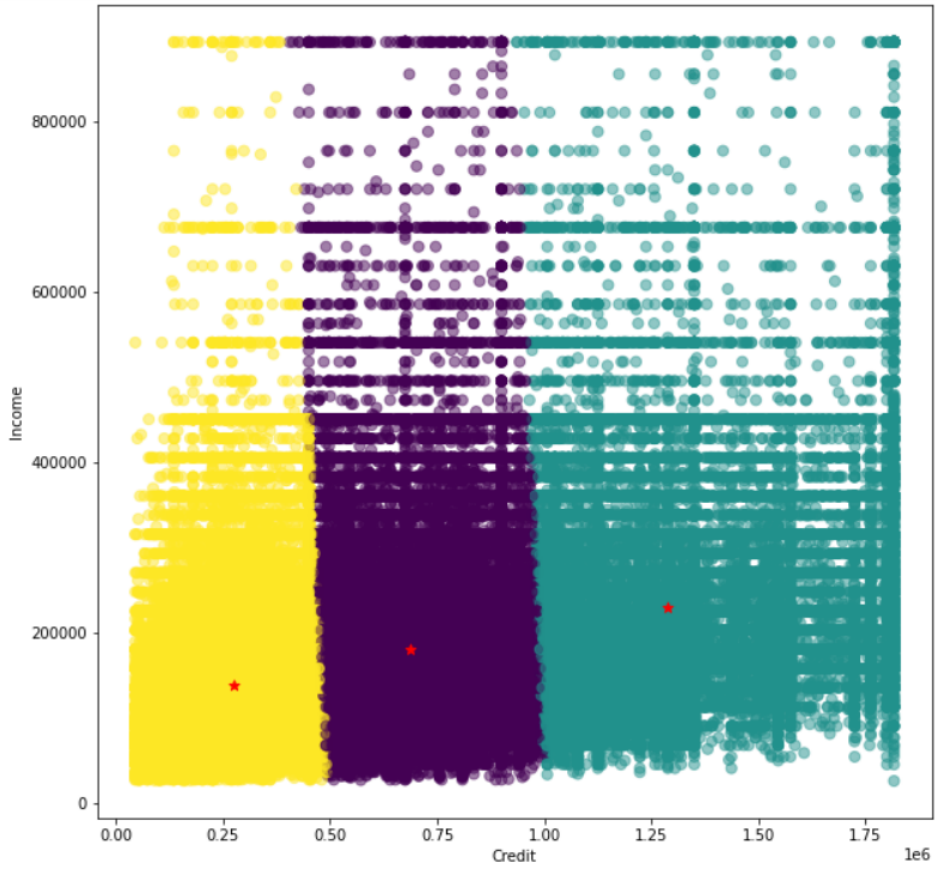
In this dataset, this technique can be used to detect non-defaulters in a bank. If a person is non-defaulter, he(his features) must exist in the normal range of the dataset. If a feature for a person lies in the outliers he is more likely to be a defaulter. If more features for the same person lie in outlier he must be a defaulter. A bank can easily use this technique to identify defaulters.

Clustering Analysis:

We also conducted cluster analysis of our data to observe if we could extract any valuable information. Clustering is a method of identifying and grouping similar data points in larger datasets without concern for the specific outcome. By conducting clustering, we can classify data into structures that are more easily understood.

Credit or credit score is a universal way used to determine the chances of an individual paying back a loan, by banks and financing institutions. Therefore, we are following the same idea. We will also compare credit against different attributes to see what kind of attribute values show promising credit values. So high credit values will suggest that those particular values of attributes will pay back their future loans on time.

Total Income vs. Credit:

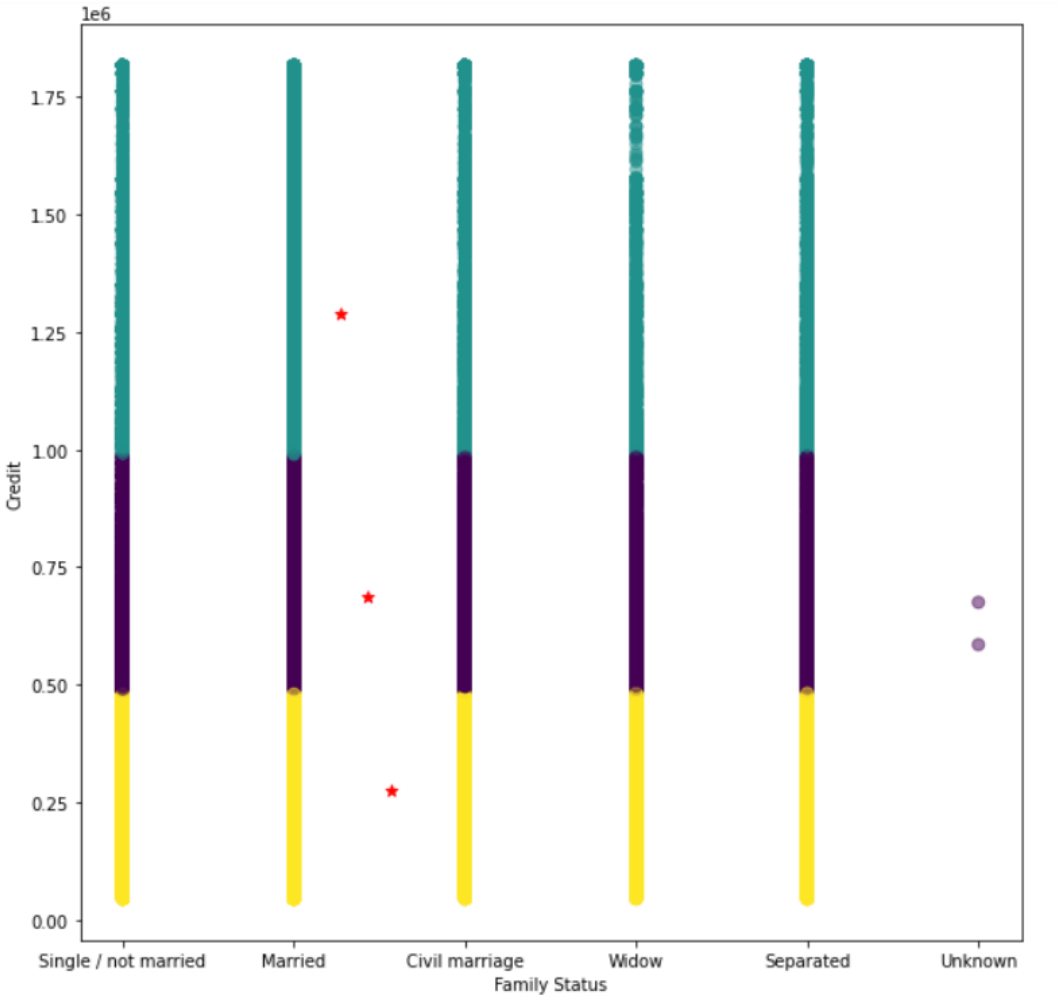


The above cluster diagram represents a scatter plot between the attributes total income and credit. We can clearly observe that there are three clusters. Out of these three clusters, the rightmost cluster appears to be the largest (blue one).

This suggests that people with high credit were usually the ones that had relatively low income as compared to others. This visual can have many explanations. The one that we were able to figure out was that, low income require loans in order to maintain their expenses that can't be fulfilled with their normal salary. Therefore, they have to take loans frequently. And to take frequent loans, they need to have good credit because banks give loans to people with high credit scores. So to ensure they get loans in the future, they try to maintain a high credit score.

From this observation, we would suggest the bank give loans to people that show a relatively low income, because they are dependent on loans to take care of heavy expenses. Since they are dependent on loans, and need them from banks, they will repay them on time and not default because they need to maintain a good credit score.

Credit vs. Family Status:



The above cluster diagram represents a scatter plot between the attributes credit and family status. We can clearly observe that there are three clusters. Out of these three clusters, the topmost cluster appears to be the largest (blue one).

We can notice that the blue cluster has the same size across each category. Plus, this cluster almost has the same density as well. This suggests that an individual's credit score does not depend on what their current family status is.

So we suggest to the bank that they do not need to worry about an individual's family status when deciding if they are eligible for the loan. Because we can see that regardless of their status, they maintain a high credit score, which means that they do not default on their loans and pay them back on time.

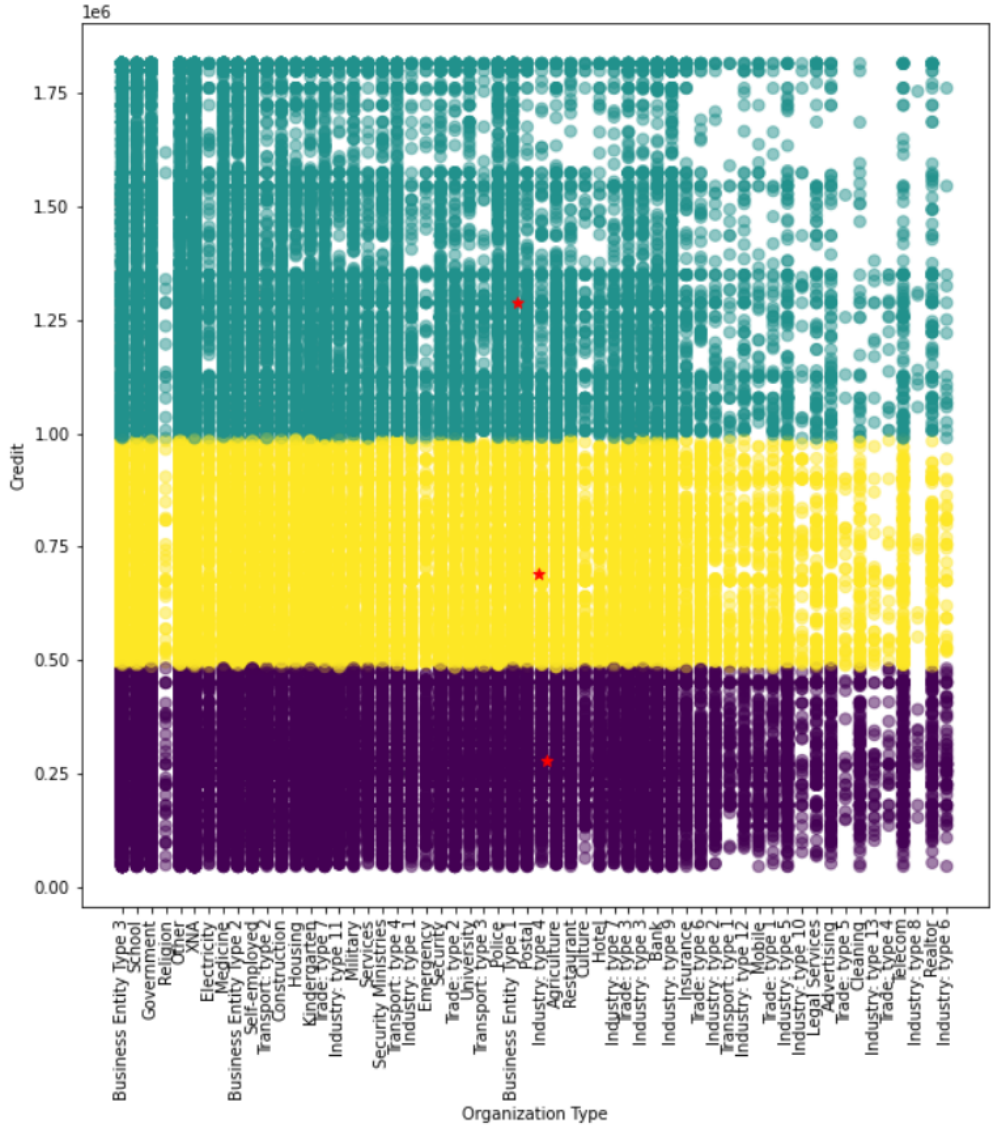
Credit vs. Occupation Type:

The above cluster diagram represents a scatter plot between the attributes credit and occupation type. We can clearly observe that there are three clusters. Out of these three clusters, the topmost cluster appears to be the largest (blue one). The topmost cluster indicates a group of occupations with high credit.

This cluster diagram shows us that most of the occupation types belong to the blue cluster, with decent density, for example: Laborers, Core staff, Accountants, Managers, Drivers, Sales staff,, Medicine staff, and etc. Although there are occupation types like: low-skill laborers, realty agents, waiters/barmen staff, etc that belong to the blue cluster but they do not have a presence in that group. This means that these occupations are the ones that have a considerable number of individuals that were defaulters and did not pay back loans on time.

Therefore, it is our suggestion to the bank that they should give loan preferences to individuals that belong to the occupations that show a high credit score. For example, they should give preference to the following occupations: Laborers, Core staff, Accountants, Managers, Drivers, Sales staff, Medicine staff, and etc.

Credit vs. Organization Type:



The above cluster diagram represents a scatter plot between the attributes credit and organization type. We can clearly observe that there are three clusters. Out of these three clusters, the topmost cluster appears to be the largest (blue one). The topmost cluster indicates a group of organizations with high credit.

The above cluster diagram shows us that the majority of organizations do not show high credit scores. Because the blue cluster is sparse for many organization types. For example: realtor, industry: type 6, cleaning, mobile and etc. The sparse number of values means that they default often and do not pay back loans on time. There are only a few organizations with high credit because they have denser regions on them like: business entity: type 3, school, government, medicine, and etc. Their dense number of values suggest that they are non-defaulters mostly and that they pay back loans on time.

So, we suggest the bank to prefer organizations with high credit scores for loans like: business entity: type 3, school, government, medicine, and etc. This is due to their good track record with loans according to the dataset.