BANK LOAN CASE STUDY

Dendi Brundha

Project Description

- The project aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan ,reducing the amount of loan, lending (to risky applicants) at a higher interest rates,etc.
- This will ensure the consumers capable of repaying the loan are not rejected.
- The main purpose of this case study is to identify such applicants using EDA

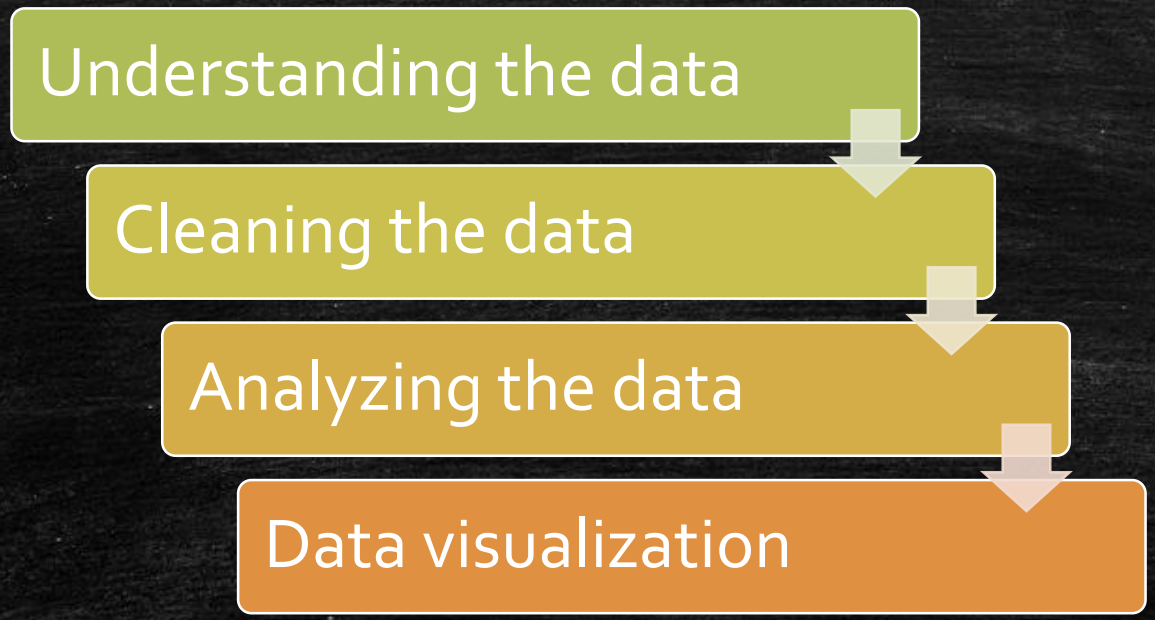
Approach

This case study is provided with three data sets :

- **application_data** :which contains data of client information at the time of application
- **previous_application** : which contains data of clients previous information i.e., whether the loan is approved , declined or cancelled etc.
- **column_description** file contains information about all the columns in both the files.
- These 3data sets are used for risk analysis and use results so that the loan is approved to correct candidates.

Tech-stack used: Jupyter , Excel, Tableau , PPt

Steps involved:



Understanding the data

We first go through all the data files and find out the columns that will be useful for the analysis and divide them as categorical and numerical variables(done in tableau).

- Categorical variables(mathematical operations cant be done): Gender, Education, Family_status, loan_purpose, etc.
- Numerical variables : Total income, credit amount, count of children ,Age , etc.

Cleaning the data

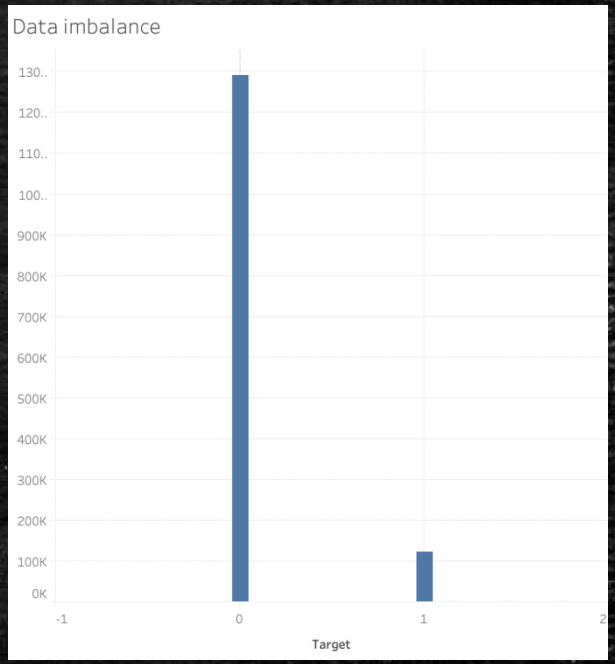
- The data sets that we are provided with have huge amount of data. So the data should be cleaned and handled to make it more efficient for data analysis.
- To clean the data, find the columns with null values greater than 50% and delete those columns. Now, in the columns remaining, we drop the columns that are not so useful for the analysis.
- After dropping the columns that are not useful, handling of data with null values need to be done.
- This can be done by doing statistical analysis on the data and fill the null values with appropriate values(i.e., mean , median etc.)

Analyzing the data

- The data after cleaning is now ready for analysis. We use the dataset to find relationship between variables and find the columns that have more impact while approving loan.
- We use univariate and bivariate analysis to find the patterns in the variables involved in loan application(Ex: Income of client , education status Vs application status.
- Data analysis is done to find the answers for the questions asked like:
  i. Data imbalance ratio
  ii. Outliers
  iii. Correlation

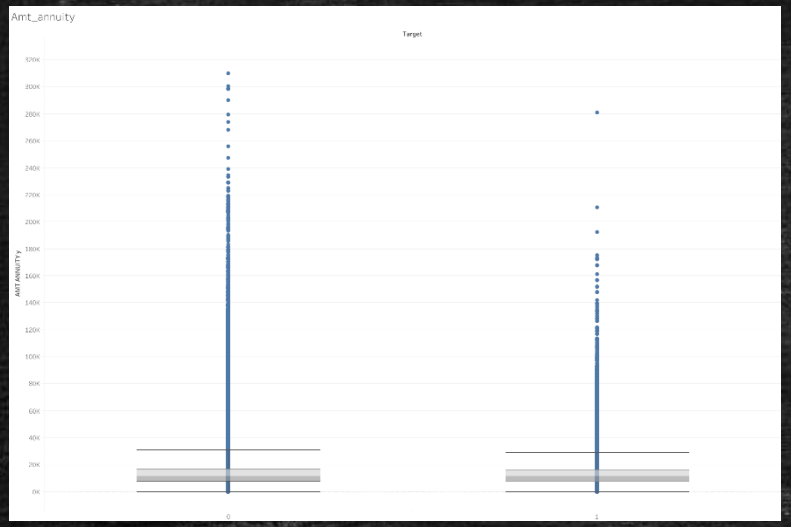
Data imbalance

- Data imbalance is generally checked for binary variables(i.e. when they have only two values [0,1]).
- Data is said to be imbalanced when it is not distributed equally on given range of values . Data imbalance ratio for balanced data is 0.5. All data sets other that 0.5 ratio are said to be imbalanced.
- When we plot a graph with count of people having difficulties in paying the loan and people who can pay the loan without any difficulties, we can see that maximum people pay loan on time, but there are few people who can't pay loan on time causing data imbalance. Giving loan to such people may cause risk to the bank.

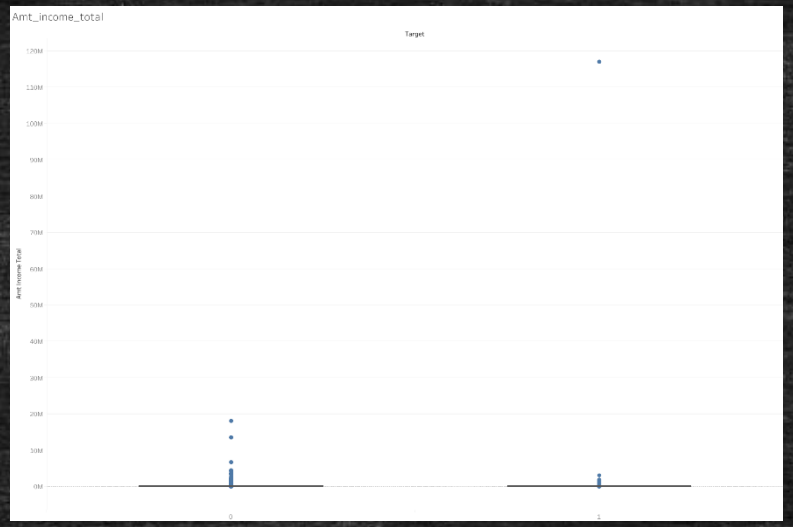


Outliers

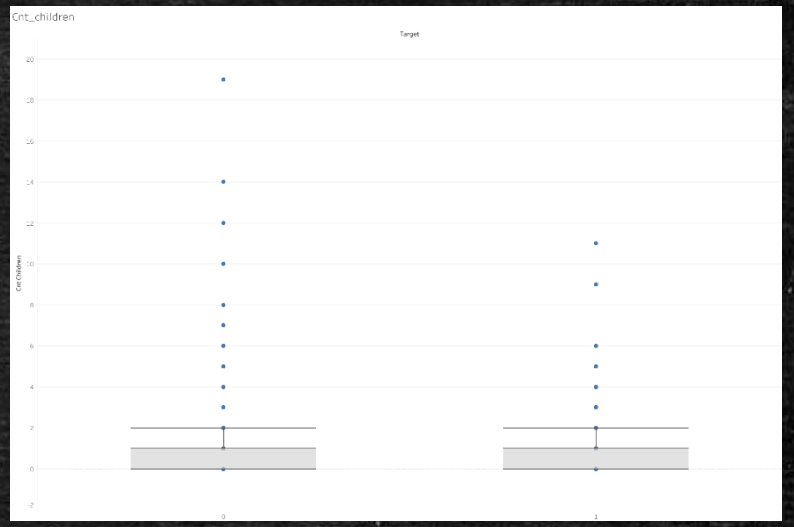
- Outlier is a data that is far away from the rest of the data.
- Outliers are generally found only for numerical data and can't be found for categorical data.
- These can be found by using box-plots.
- A statistical analysis is done on all the numerical columns and observed the columns that have high values of standard deviation.
- Box plot is then plotted for those columns which have high standard deviation as outliers may be one of the reason for higher deviation from values.
- Outliers are calculated for Annuity amount, total_income,children count, days employed,etc.



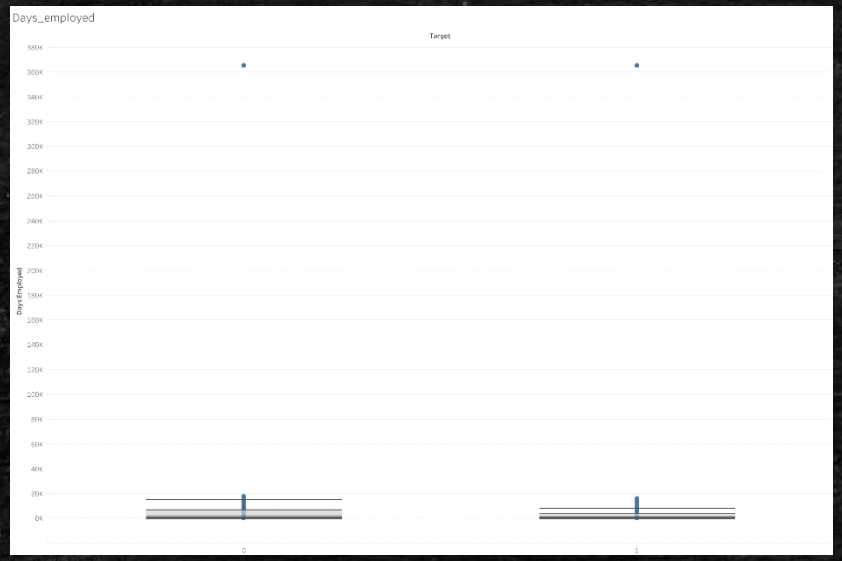
Amt_Annuity



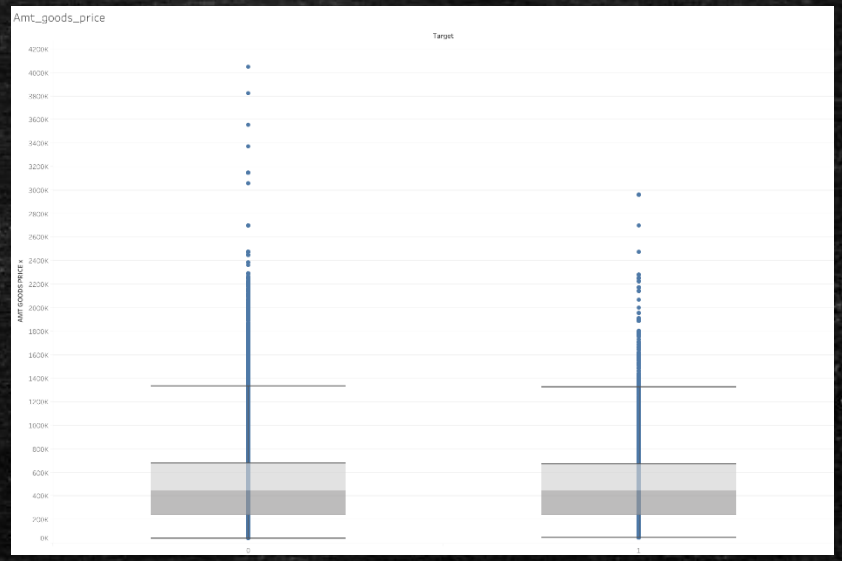
Amt_income_total



Cnt_children



Days_Employed



Amt_goods_Price

- From the above plots we can observe that annuity have high amount of outliers as most of the data points are outside the box. This may be because of the type of product the client took loan for, or the income of the client.
- Outliers for all the columns with payment difficulties have less outliers compared to those who don't have any difficulty as there are only 10% people with difficulties.
- We can see that count of children has an outlier value of 14 which is highly unusual in present days. Even if the client has 14 children for real , then the chances of him paying loan after providing basic amenities to all of them will be very less unless the client earn very high.
- Days employed has an outlier of 350k days which is impossible. So approving loan to such clients may cause loss to the bank.
- We can use these outliers to make sure that the bank don't approve loan to wrong customers.

Correlation

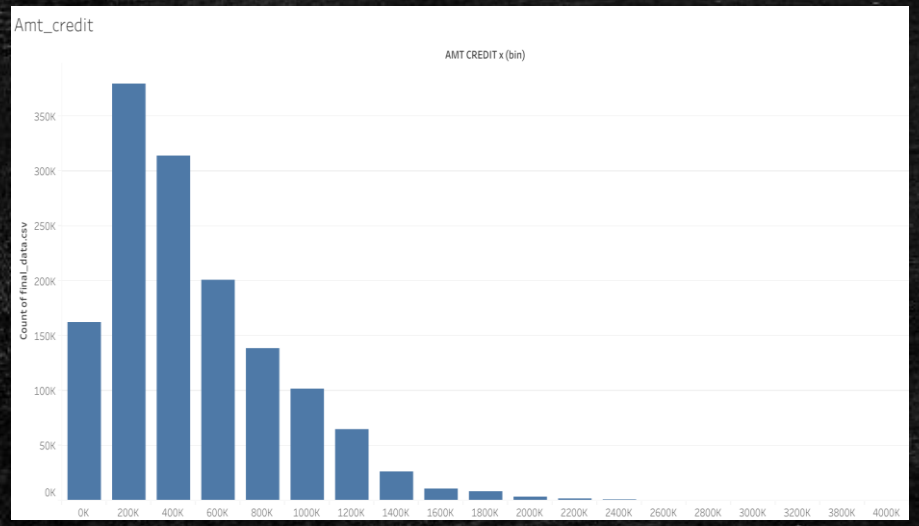
- Correlation is the relation or connection between two variables. It is generally found for numerical variables.
- Variables are said to be highly related when the value of their correlation is close to 1
- From the given data, the important columns for which correlation can be found are:
  - CNT_CHILDREN
  - CNT_FAM_MEMBERS
  - CNT_PAYMENT
  - AMT_ANNUITY
  - AMT_APPLICATION
  - AMT_CREDIT
  - AMT_GOODS_PRICE
  - DAYS_BIRTH
  - DAYS_EMPLOYED
  - AMT_INCOME_TOTAL

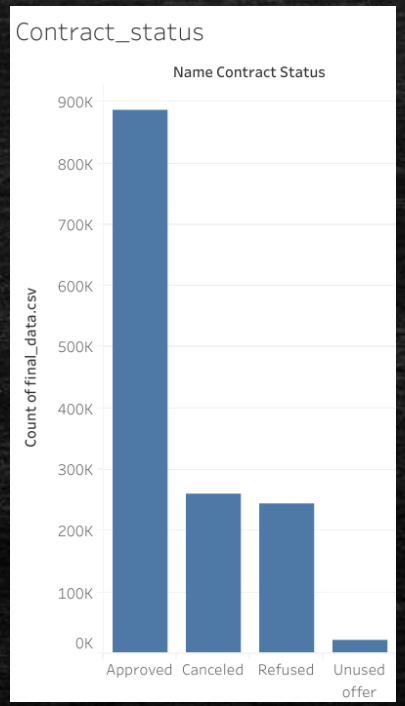
Columns	CNT_CHILI	CNT_FAMIL	CNT_PA	AMT_ANN	AMT_APP	AMT_CR	AMT_GO	DAYS_BI	DAYS_EM	AMT_IN
CNT_CHILDREN	1	0.879224	-0.049161	-0.032608	-0.034168	-0.03489	-0.034768	-0.36034	-0.249912	0.011661
CNT_FAMILY_MEMBERS	0.879224	1	-0.030302	-0.004577	-0.005809	-0.005233	-0.006154	-0.326241	-0.250514	0.014119
CNT_PAYMENT	-0.049161	-0.030302	1	0.39149	0.65092	0.641663	0.652509	0.110095	0.065006	0.017729
AMT_ANNUITY	-0.032608	-0.004577	0.39419	1	0.806458	0.812972	0.80843	0.071977	-0.006126	0.099077
AMT_APPLCATION	-0.34168	-0.005809	0.65092	0.806458	1	0.975683	0.999768	0.079786	0.010754	0.071491
AMT_CREDIT	-0.03486	-0.005233	0.641663	0.812972	0.957683	1	0.976391	0.078078	0.005046	0.070651
AMT_GOODS_PRICE	-0.034768	-0.006154	0.652509	0.80843	0.999768	0.997639	1	0.080834	0.01114	0.072037
DAYS_BIRTH	0.36034	-0.326241	0.110095	0.071977	0.079786	0.078078	0.080834	1	0.632509	-0.025717
DAYS_EMPLOYED	-0.249912	-0.250514	0.065006	-0.006126	0.010754	0.005046	0.01114	0.632509	1	-0.067046
AMT_INCOME_TOTAL	0.011661	0.014119	0.017729	0.099077	0.071491	0.070651	0.072037	-0.025717	-0.067046	1

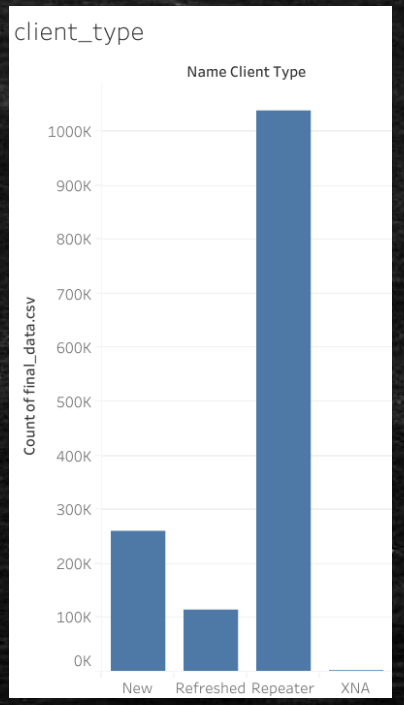
- Cells that are marked green in color are correlated positively, whereas cells those in red are negatively correlated.
- We can see that AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE have high correlation.

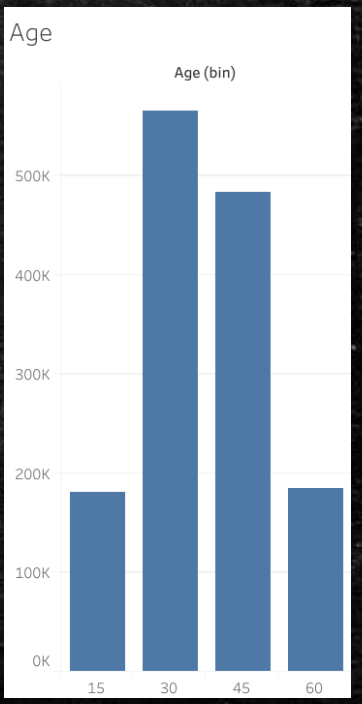
Univariate Analysis

- Univariate analysis is done using only one column. This can be done generally by using histograms or box-plots.
- Below are the graphs plotted using single variables to understand each variable in given data.

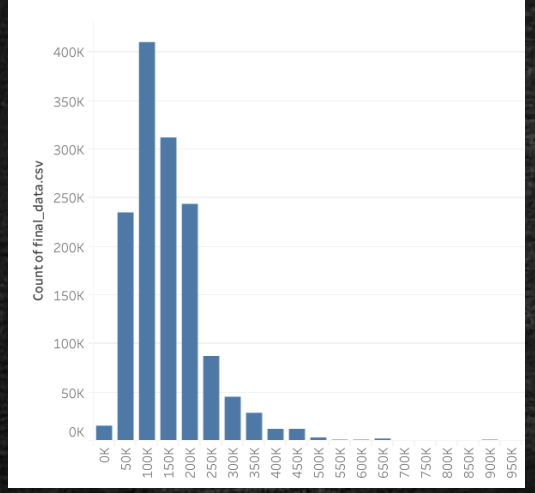




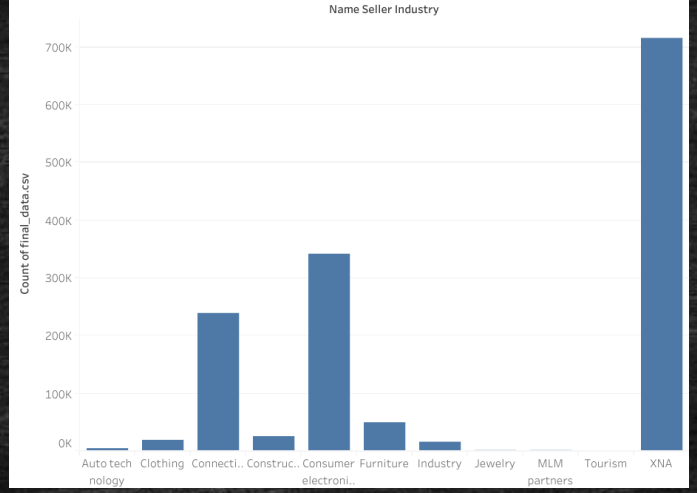




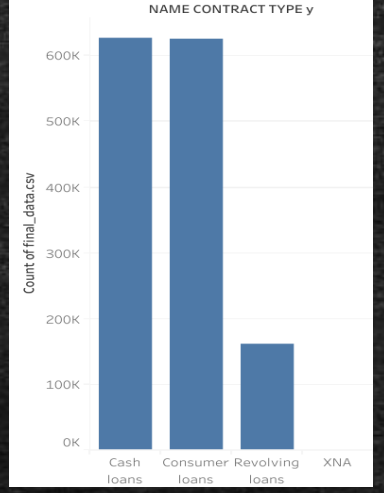
Amt_total_income



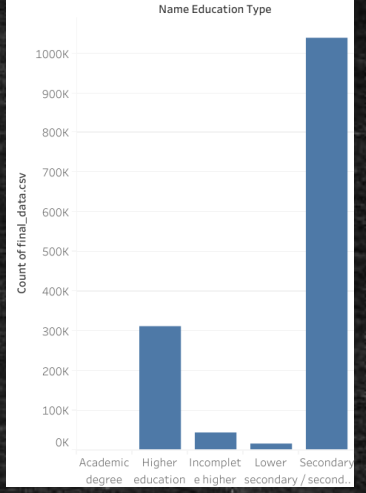
seller_industry



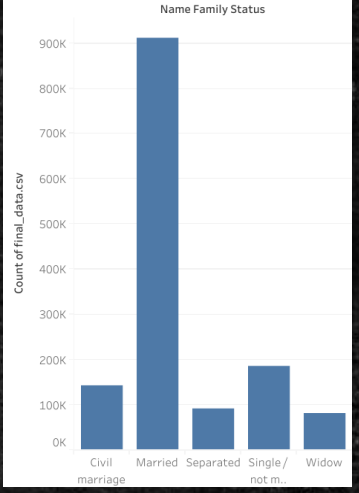
contract_type



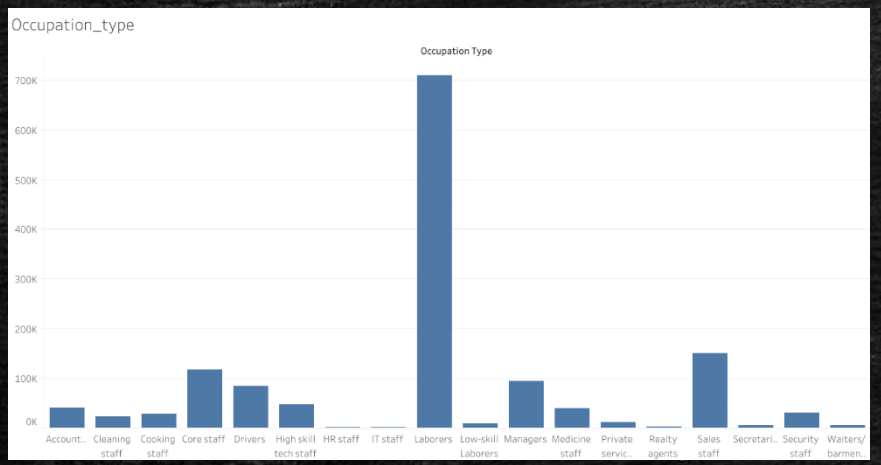
Education_type

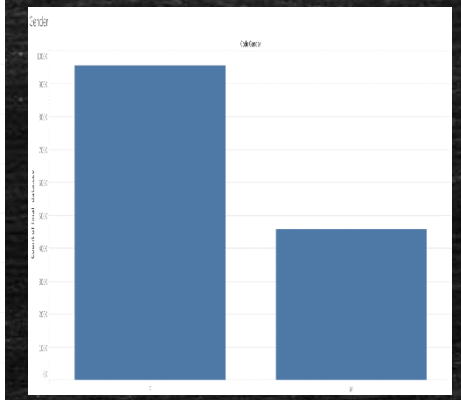


Family_status

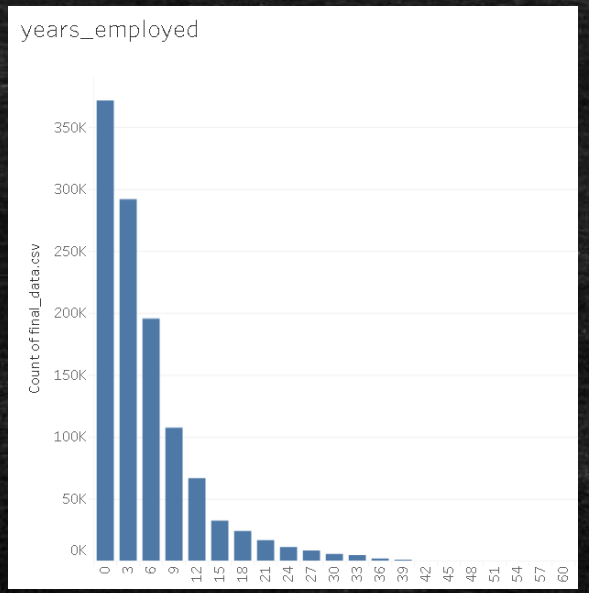


Occupation_type

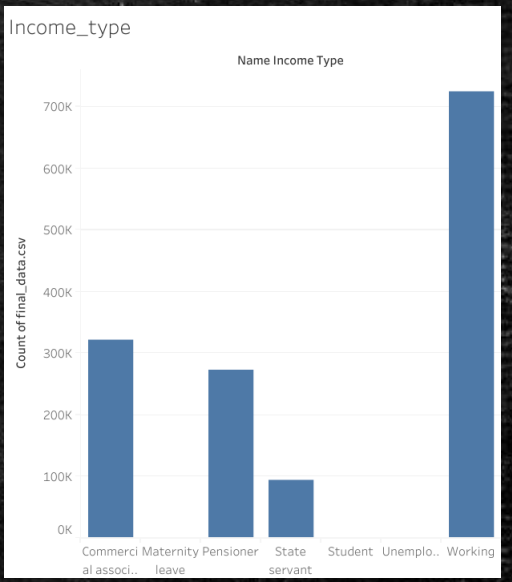




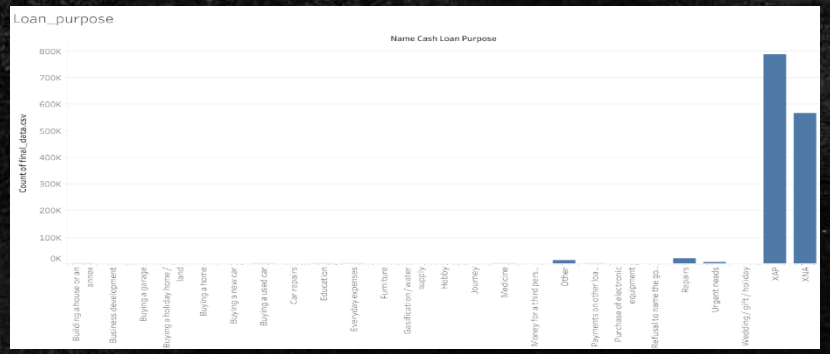
years_employed



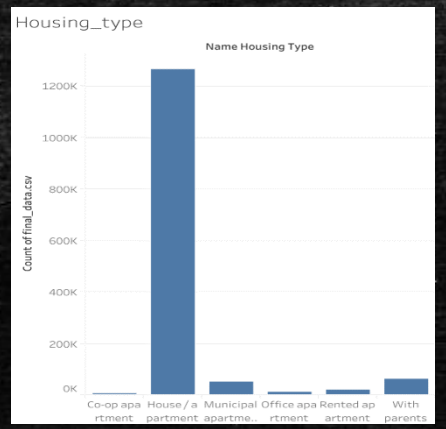
Income_type



Loan_purpose

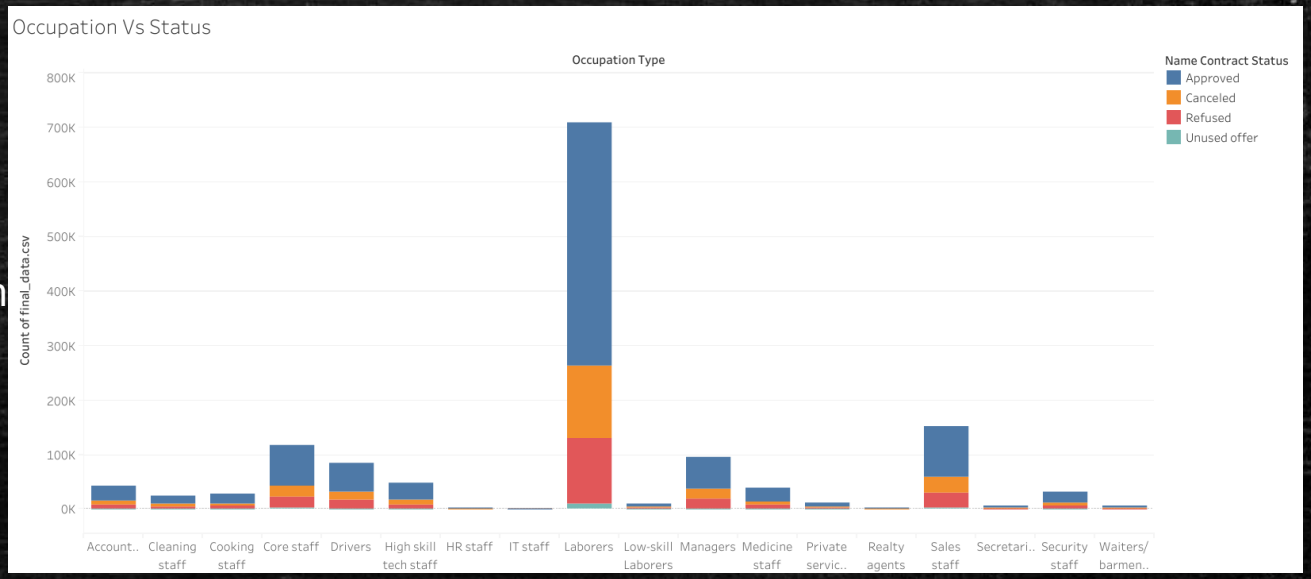


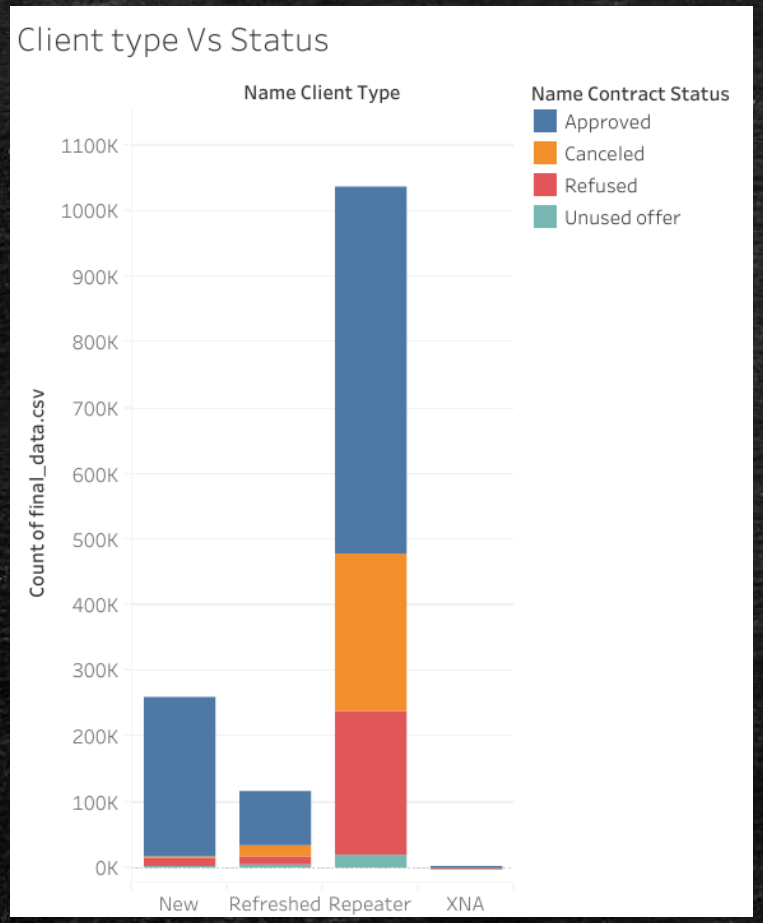
Housing_type

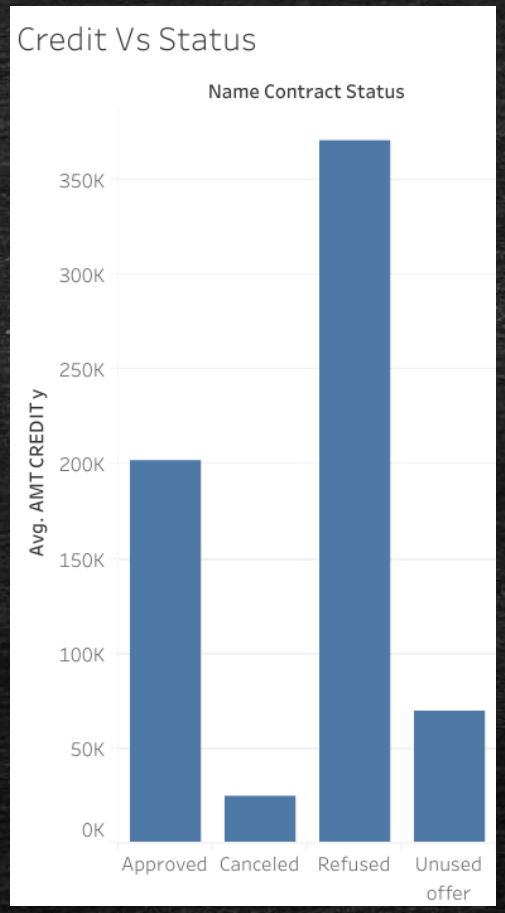


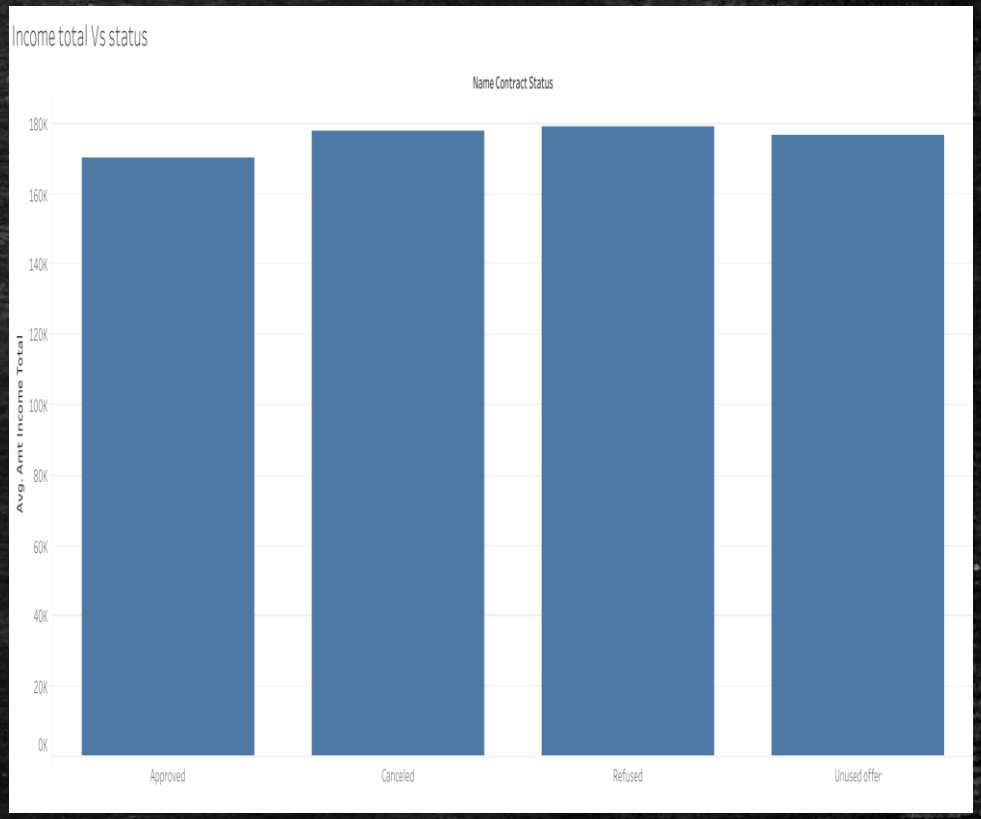
Bivariate Analysis

- This is a method of analyzing how two variables are related to each other.
- In this one is dependent variable and other is independent variable. Bivariate analysis can be done for both categorical and numerical values.
- Analysis can be done for more than two variables which is called as multivariate analysis.
- The graph below indicates relation between two variables Occupation and this graph we can observe loan approval but they have more the loan. Almost all occupation same kind of problem . So we can't whether to approve loan or not based on

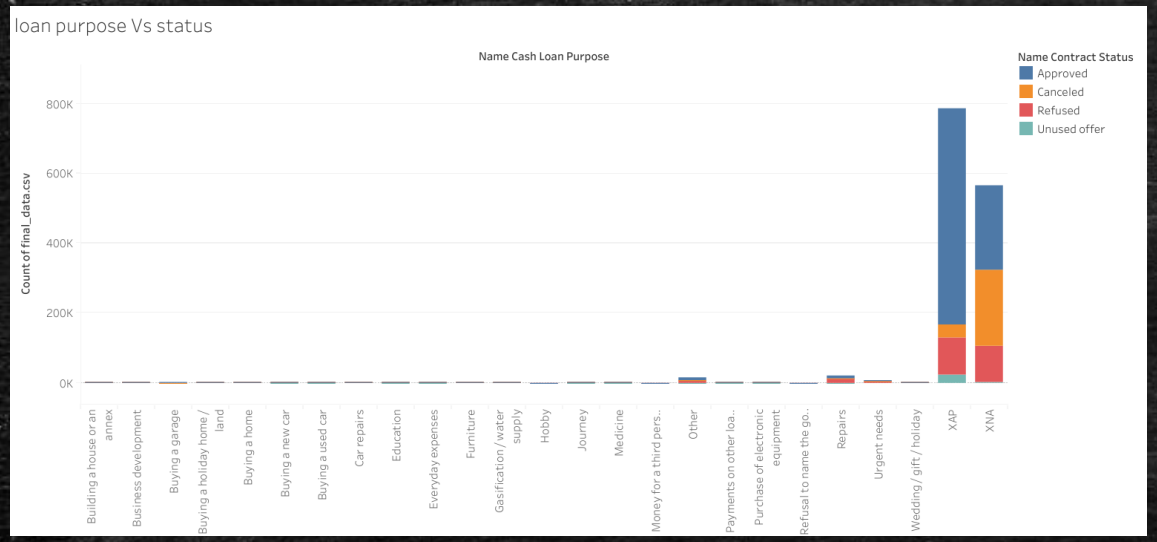


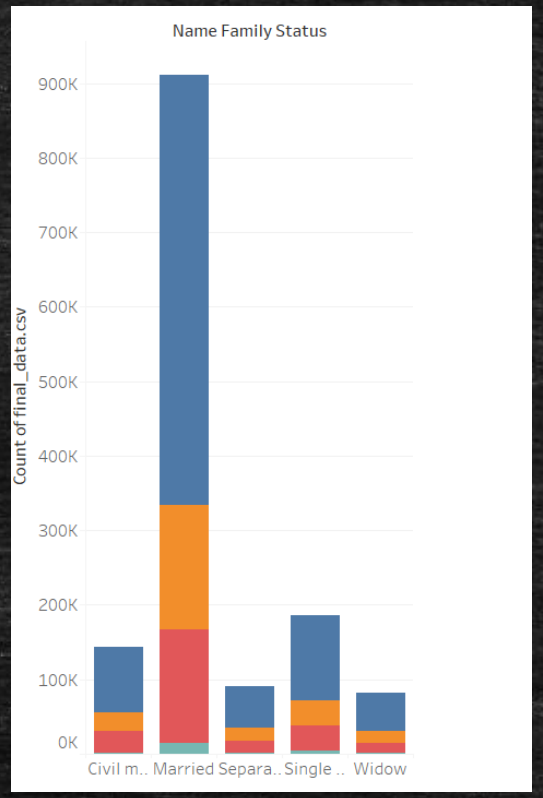


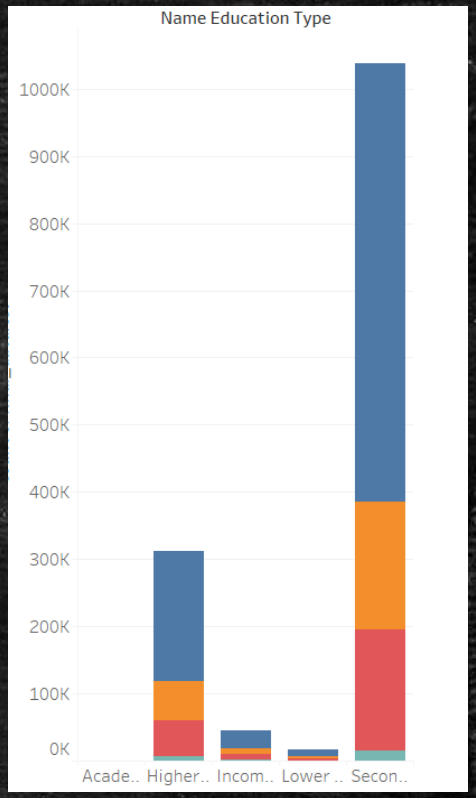




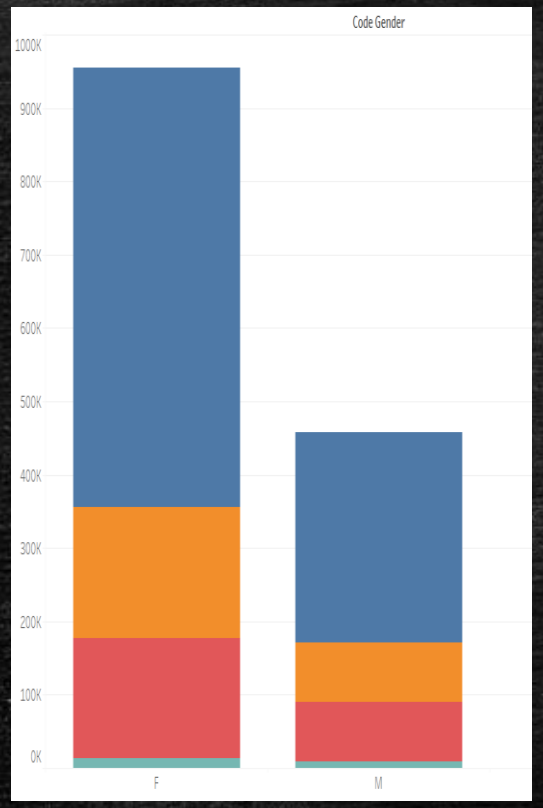
We can observe that the loans are accepted mostly of the repeaters . So it is safe to give loans to clients who have already took before. The clients who have high credit and income have more chances to pay loan as they have a source of income , while others with low income have difficulties paying the loan. Client with too high income refused the loan as they don’t need it.

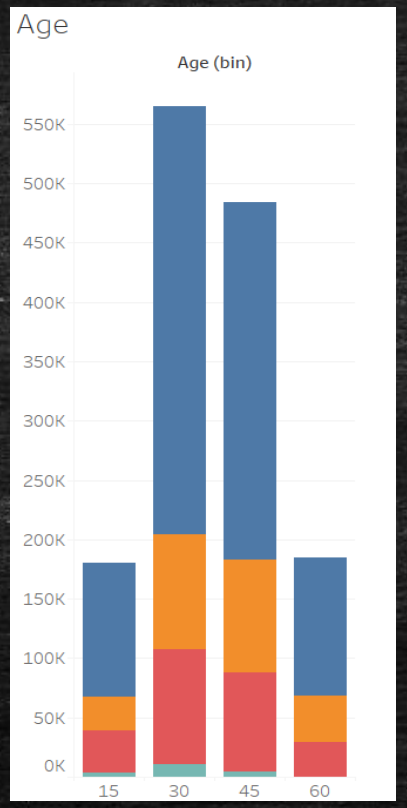


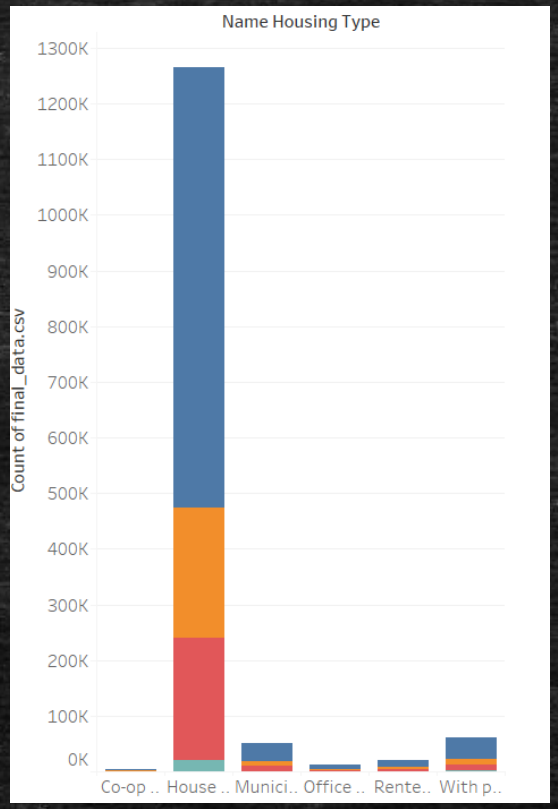


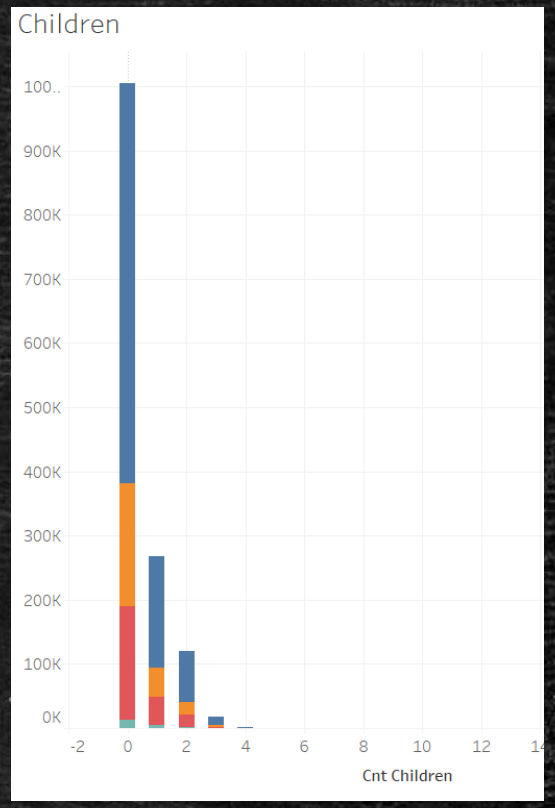


- Loan purpose with repairs have high difficulty paying the loan.
- Married clients are more better in paying loan on time than compared to single clients.
- Clients with more education are a better option for approving loan as they pay the loan back on time without loss.









- Approving loan of female is more better option and clients between age of 30 and 45 pay loan on time.
- Giving loan of clients living in a co-op apartment may cause loss to bank . So loan of clients staying in their house is a better option and clients with no kids tend to pay their loan back on time.

Results

Approving loan to the Clients with properties mentioned below may help the bank in making profits

- Staying in house and not in co-op apartment
- Age in between 30 and 45
- Female
- Have no children or 1 child
- Better education(finished secondary education)
- Married
- Not taking loan for repairs
- High source of income