EDA CASE STUDY

PREPARED BY –

ANUSHA KOGTA (anukogta008@gmail.com)

INTRODUCTION

- The term banking can be defined as receiving and protecting money that is deposited by the individual or the entities. This also includes lending money to the people which will be repaid within the given time. The primary objective of the bank is to provide their wealth in the safer hands.

PROBLEM STATEMENT

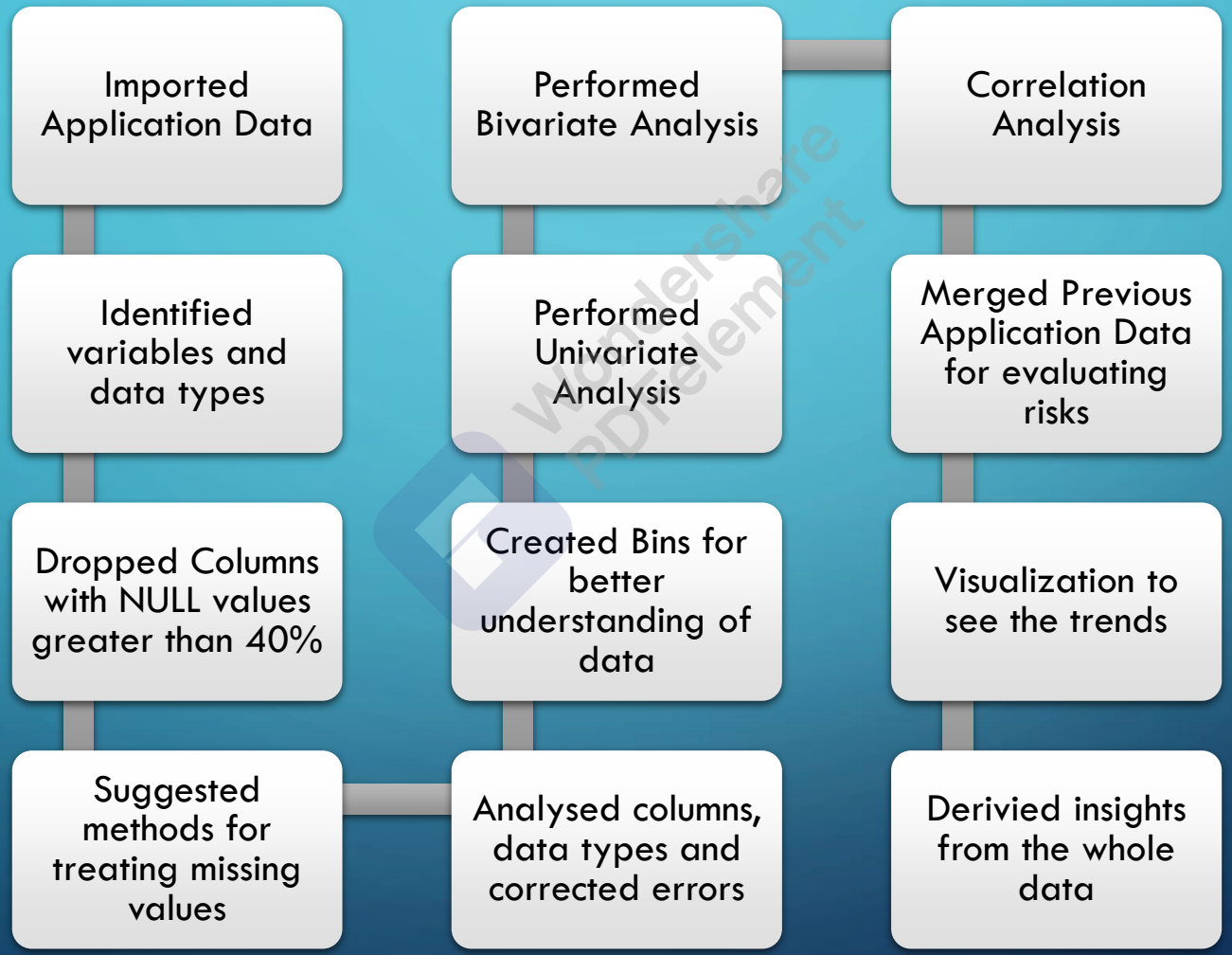
- In recent times, banks approve loan after verifying and validating the documents provided by the customer. Yet there is no guarantee whether the applicant is deserving or not.
- The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter.

BUSINESS OBJECTIVE

- This case study 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 rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

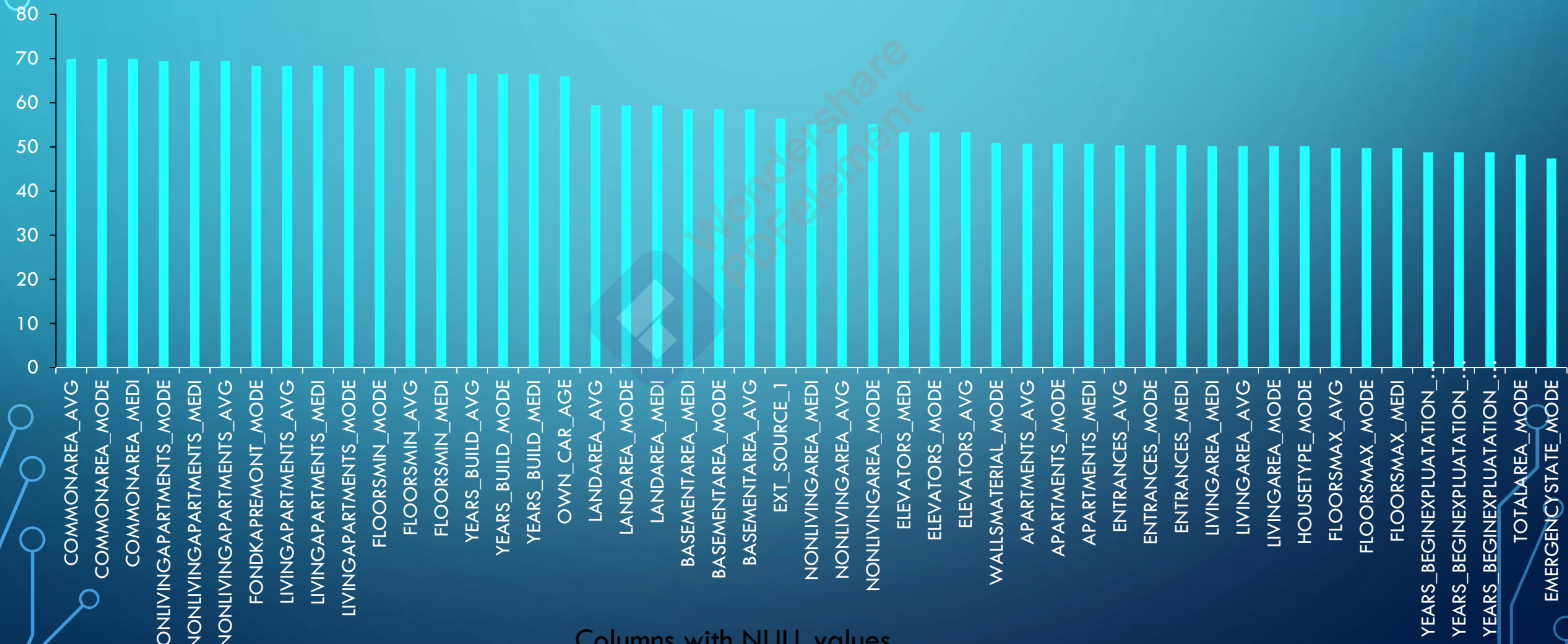
- In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

EDA PROCESS FLOW



HANDLING NULL VALUES

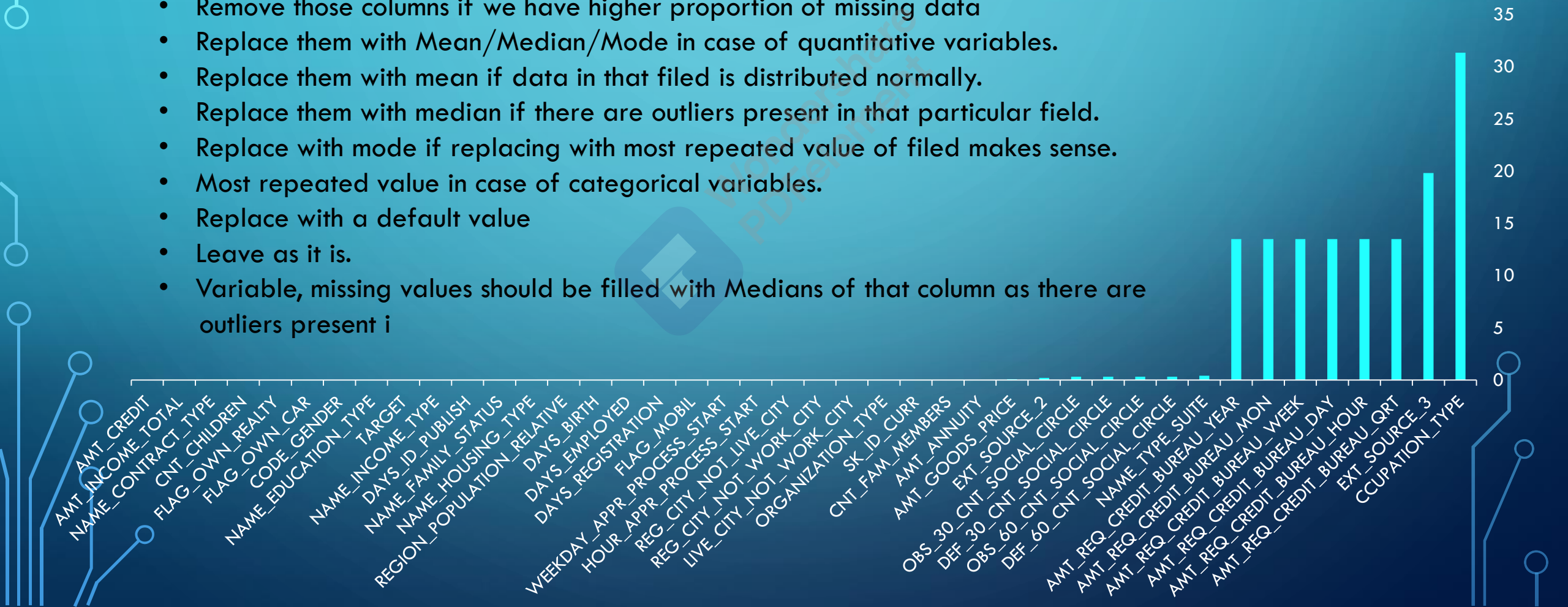
As observed from the Data , There are more than 40 columns which have Null values greater than 40%. We decided to remove these columns to get correct insights.



IMPUTATION ON COLUMNS WITH NULL VALUES LESS THAN 40 %

There are various ways to deal with missing values. Out of which, most common methods are as below:

- Remove those columns if we have higher proportion of missing data
- Replace them with Mean/Median/Mode in case of quantitative variables.
- Replace them with mean if data in that filed is distributed normally.
- Replace them with median if there are outliers present in that particular field.
- Replace with mode if replacing with most repeated value of filed makes sense.
- Most repeated value in case of categorical variables.
- Replace with a default value
- Leave as it is.
- Variable, missing values should be filled with Medians of that column as there are outliers present i



OUTLIER DETECTION

By observing the data set , We have identified these 3 variables for which the outliers are not justified. It seems that there might have been incorrect data entry for these .

1. CNT_CHILDREN

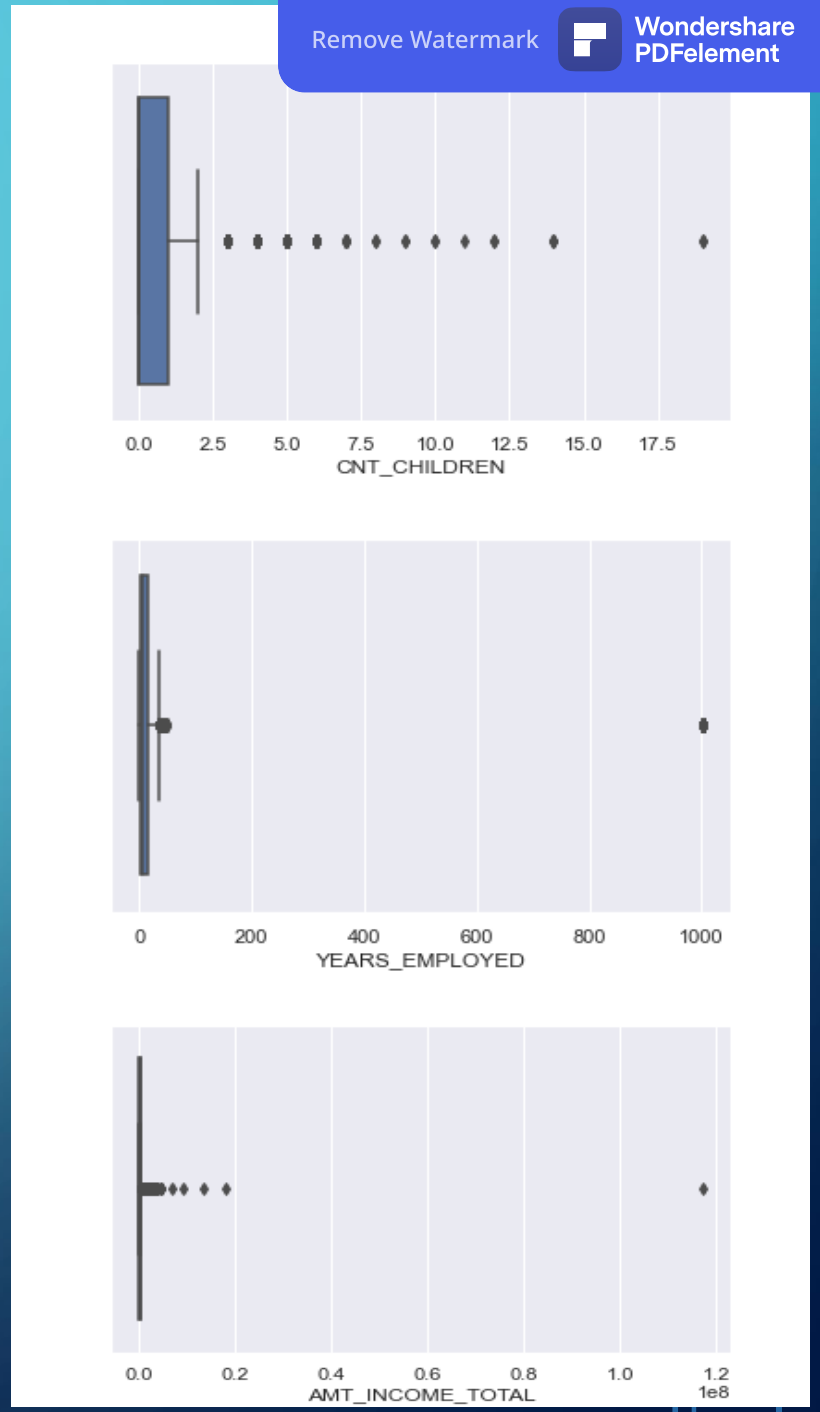
   It tells number of children a client has. From the Boxplot , we can see that there are a few clients which have children greater than 10 which is highly unlikely. Hence it can be considered as incorrect data point.

2. YEARS_EMPLOYED

   How many days before the application the person started current employment ? This is definitely incorrect as it shows years employed greater than 1000 years.

3. AMT_INCOME_TOTAL

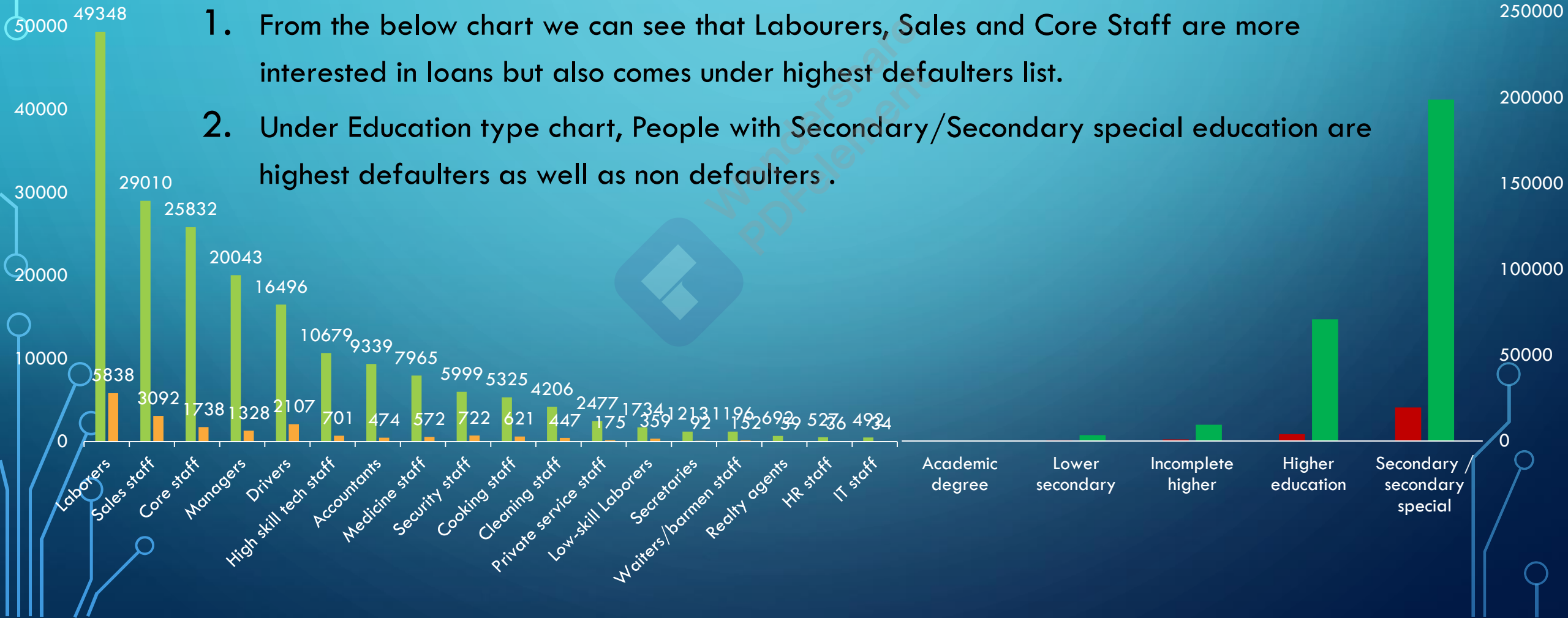
   It is the Income of the client. There is 1 instance where income is around 12 Crore which is possible if the client is at high level management but as checked this particular client is a labourer .Hence this data is incorrect as well.



DEFAULTERS AND NON DEFAULTERS COMPARISON

- This section talks about the different categories based on their Occupation and Education type and No of defaulters Vs Non Defaulters

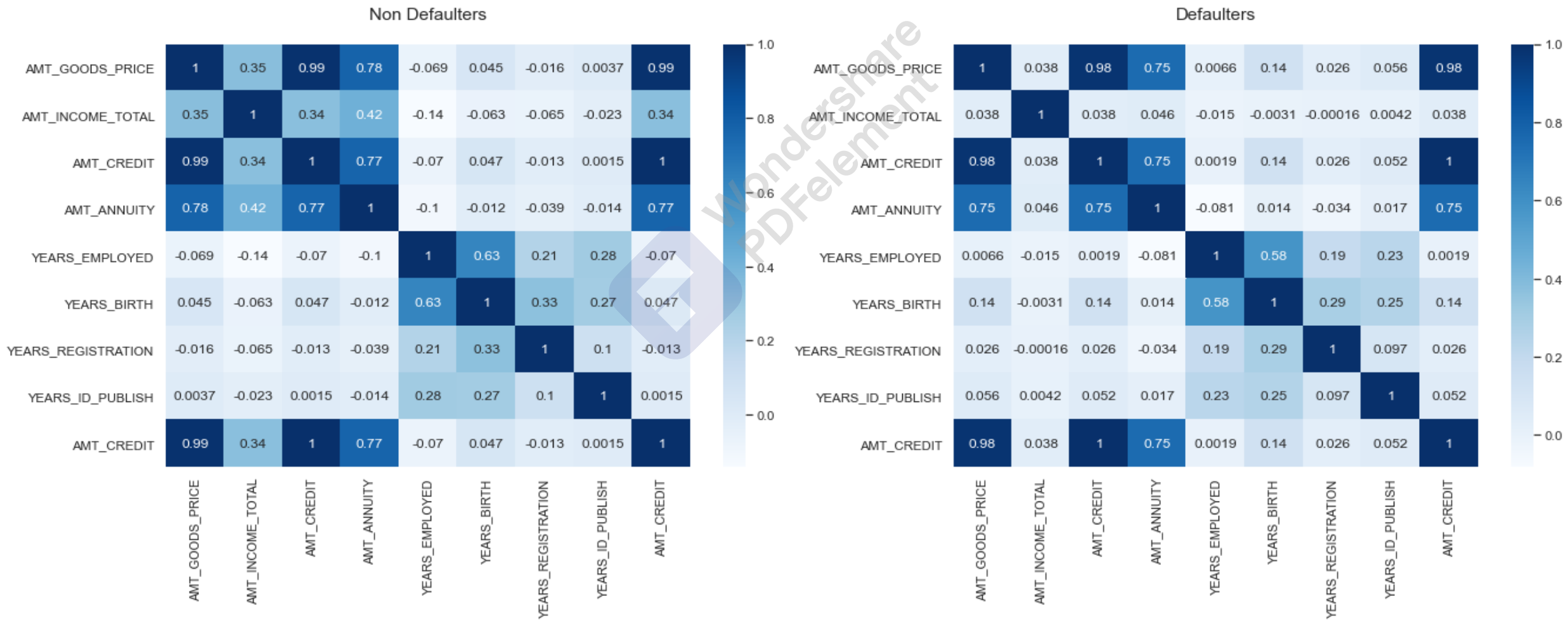
1. From the below chart we can see that Labourers, Sales and Core Staff are more interested in loans but also comes under highest defaulters list.
2. Under Education type chart, People with Secondary/Secondary special education are highest defaulters as well as non defaulters .



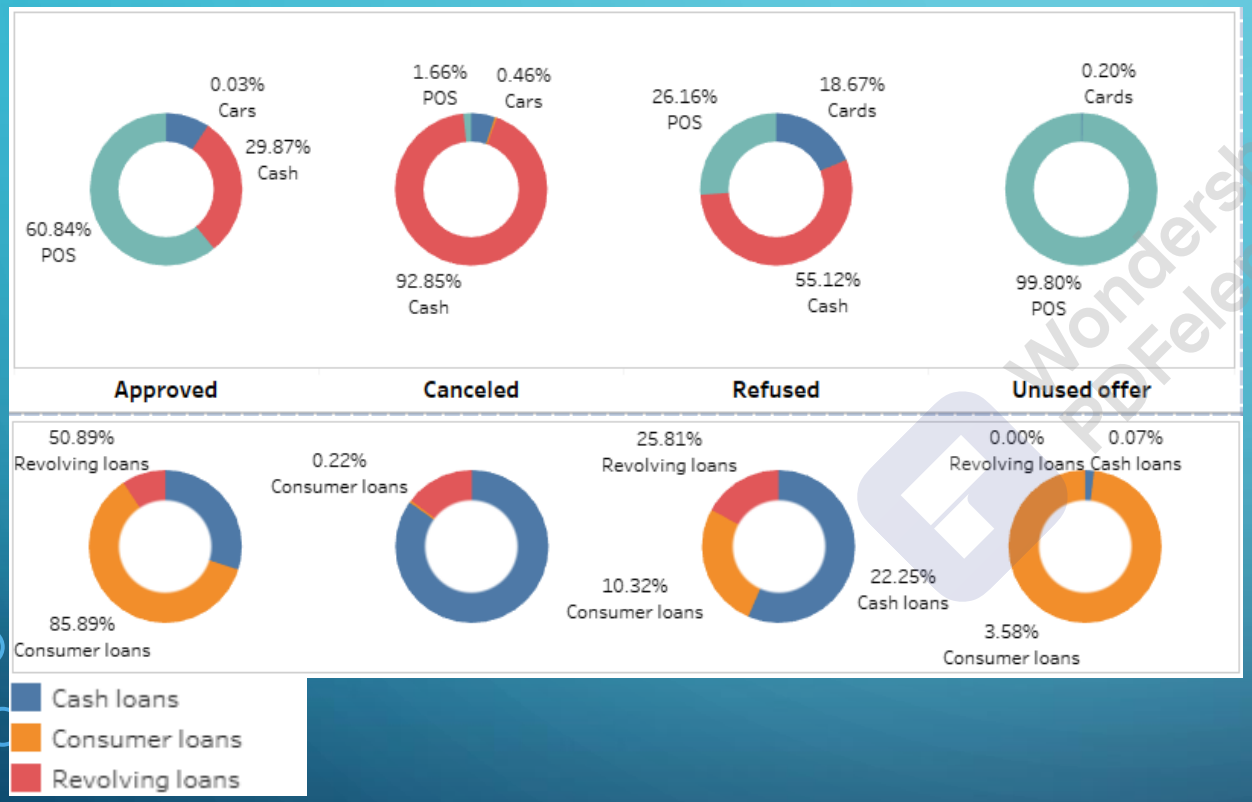
CORRELATION MATRIX OF MULTIPLE VARIABLES

Top 3 Correlation pairs for Defaulters as well as non Defaulters are almost similar

1. AMT_CREDIT - AMT_GOODS_PRICE
2. AMT_ANNUITY - AMT_GOODS_PRICE
3. AMT_CREDIT - AMT_ANNUITY



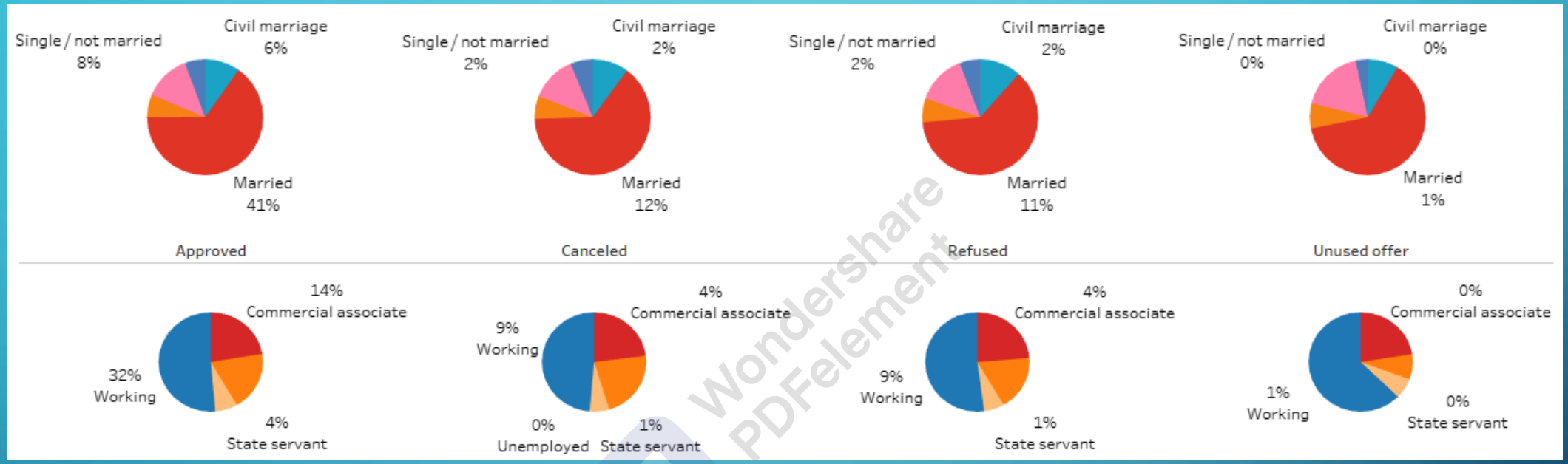
CONTRACT STATUS TYPE COMPARISON



Top doughnut chart shows comparison between different portfolios .It can be seen that Cash category is most cancelled & refused whereas POS is highly approved .

Bottom Doughnut chart shows comparison between different contract types . Cash loans are cancelled & refused the most whereas consumer loans are most approved .

CONTRACT STATUS TYPE COMPARISON

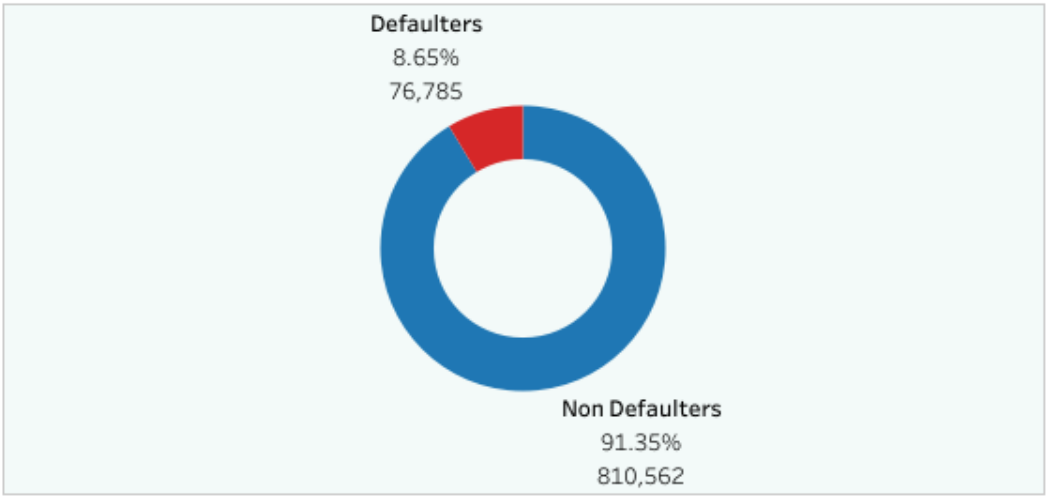


1.Top Pie chart displays Family Status type by Contract type .It can be seen that Married Status is highest amongst all the contract types.

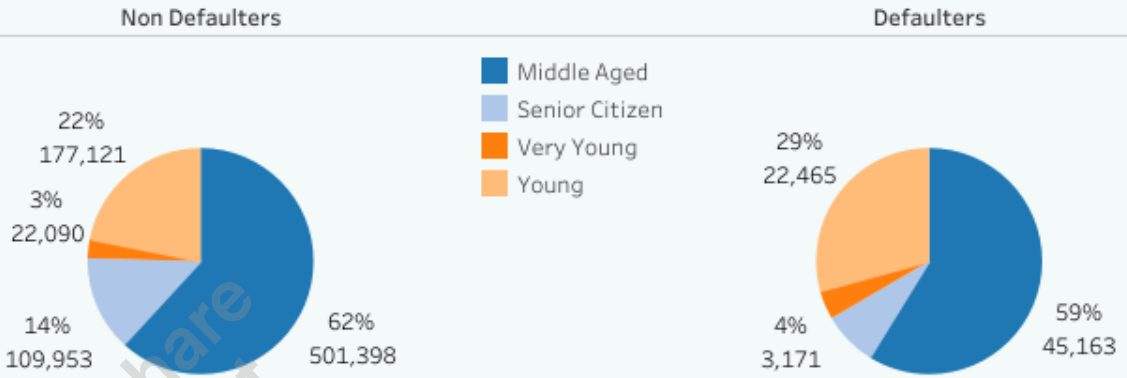
2. Bottom Pie chart displays Income Type by contract type. Working Type people are applying for more loans as compared to others and commercial associates are also taking more loans.

NON DEFAULTERS VS DEFAULTERS ANALYSIS

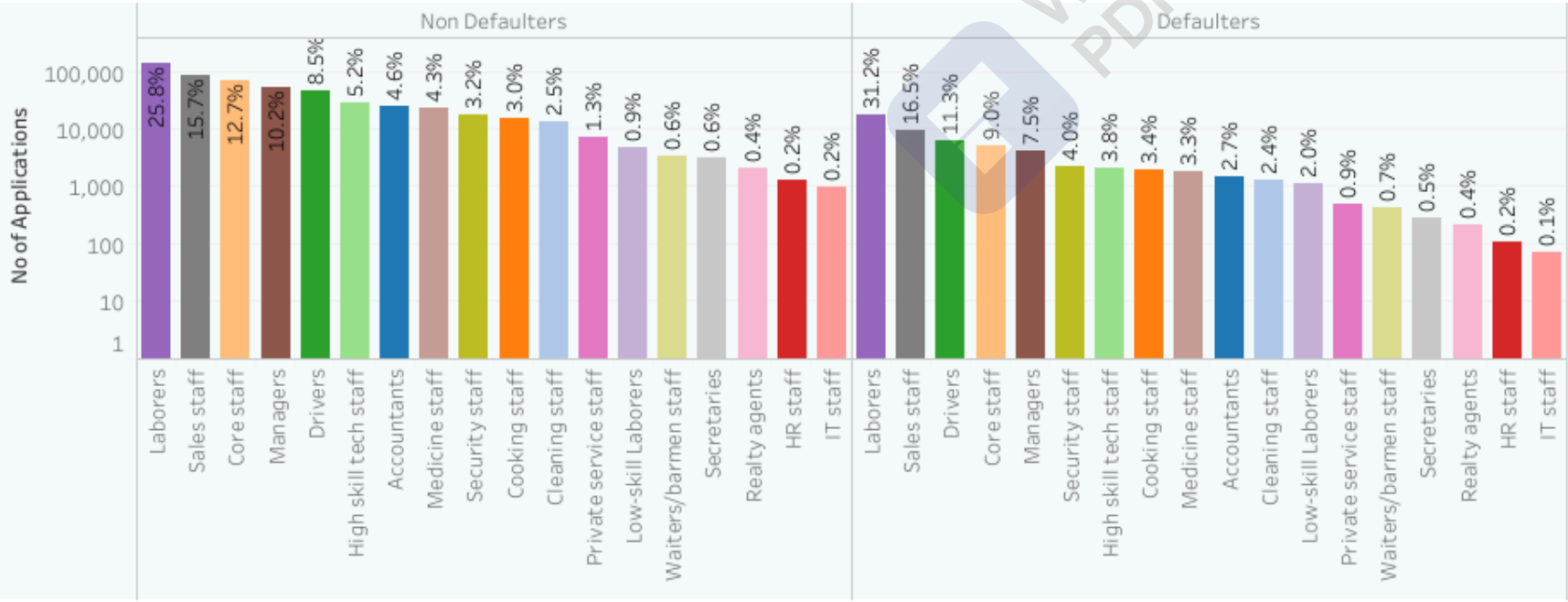
No of Defaulters vs Non defaulters



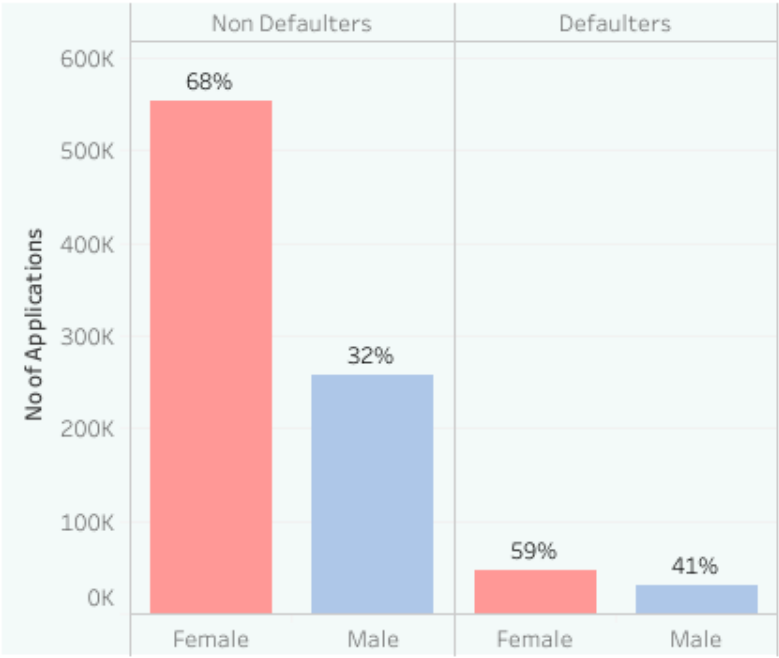
Non Defaulters Vs Defaulters by Age Type



Non Defaulters Vs Defaulters by Occupation Type



Male Vs Female Defaulters



Correlation Analysis

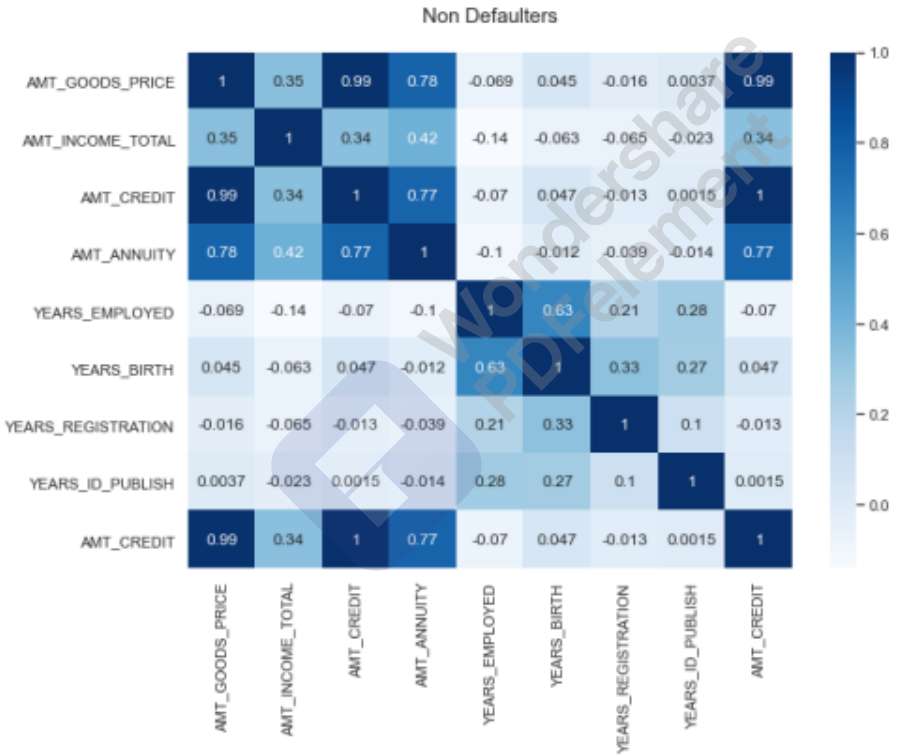
Non Defaulters Top 10 Correlation

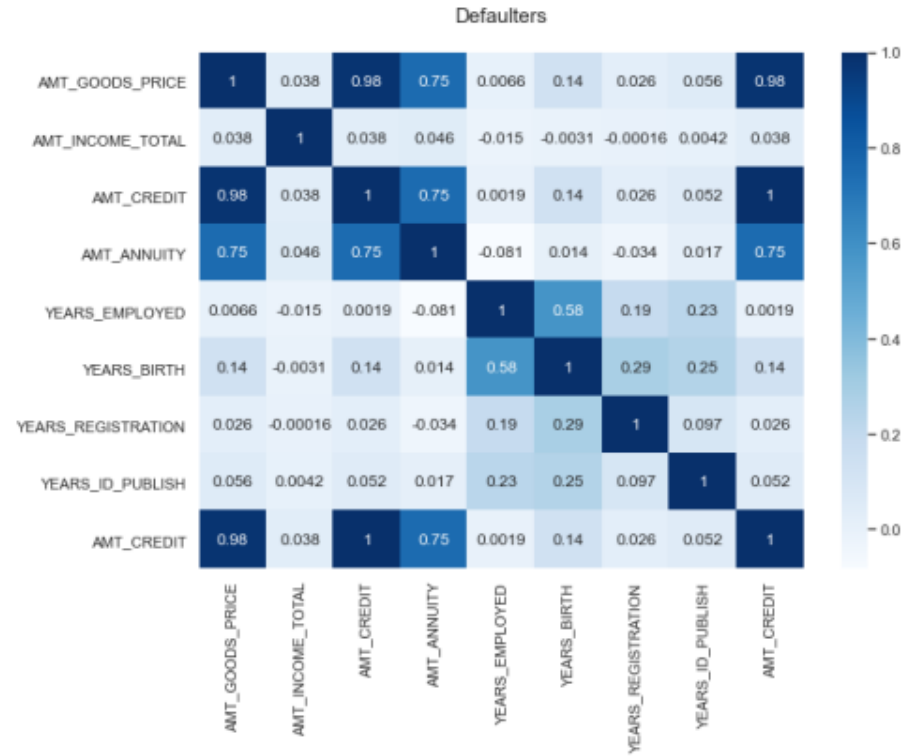
Variable 1	Variable 2	Correlation
AMT_CREDIT	AMT_GOODS_PRICE	0.99
AMT_ANNUITY	AMT_GOODS_PRICE	0.78
AMT_CREDIT	AMT_ANNUITY	0.77
YEARS_BIRTH	YEARS_EMPLOYED	0.63
AMT_ANNUITY	AMT_INCOME_TOTAL	0.42
AMT_INCOME_TOTAL	AMT_GOODS_PRICE	0.35
AMT_CREDIT	AMT_INCOME_TOTAL	0.34
YEARS_REGISTRATION	YEARS_BIRTH	0.33
YEARS_ID_PUBLISH	YEARS_EMPLOYED	0.28
YEARS_ID_PUBLISH	YEARS_BIRTH	0.27

Defaulters Top 10 Correlation

Variable 1	Variable 2	Correlation
AMT_CREDIT	AMT_GOODS_PRICE	0.98
AMT_ANNUITY	AMT_GOODS_PRICE	0.75
AMT_CREDIT	AMT_ANNUITY	0.75
YEARS_BIRTH	YEARS_EMPLOYED	0.58
YEARS_REGISTRATION	YEARS_BIRTH	0.29
YEARS_ID_PUBLISH	YEARS_BIRTH	0.25
YEARS_ID_PUBLISH	YEARS_EMPLOYED	0.23
YEARS_REGISTRATION	YEARS_EMPLOYED	0.19
YEARS_BIRTH	AMT_GOODS_PRICE	0.14
AMT_CREDIT	YEARS_BIRTH	0.14

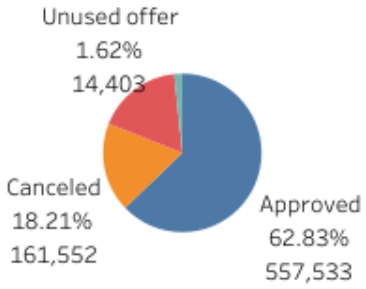
Correlation Matrix





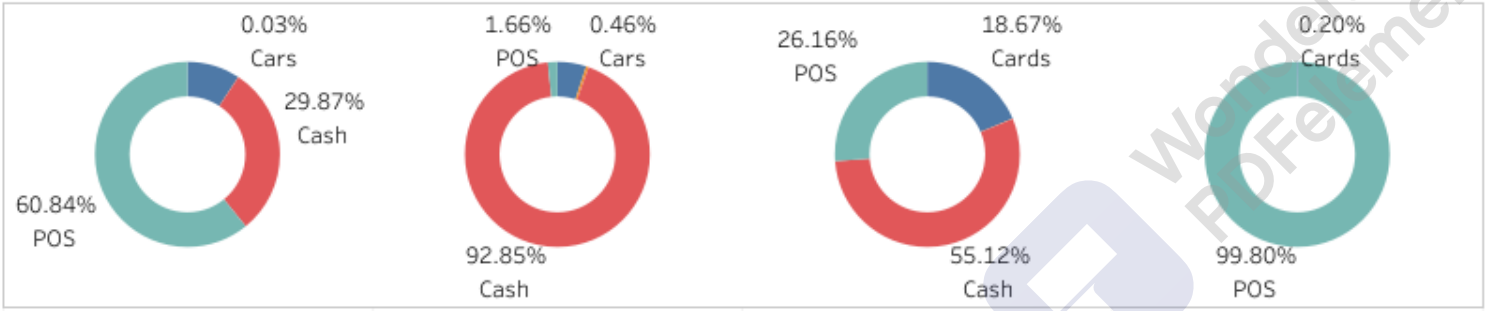
Previous Application Analysis

Status type

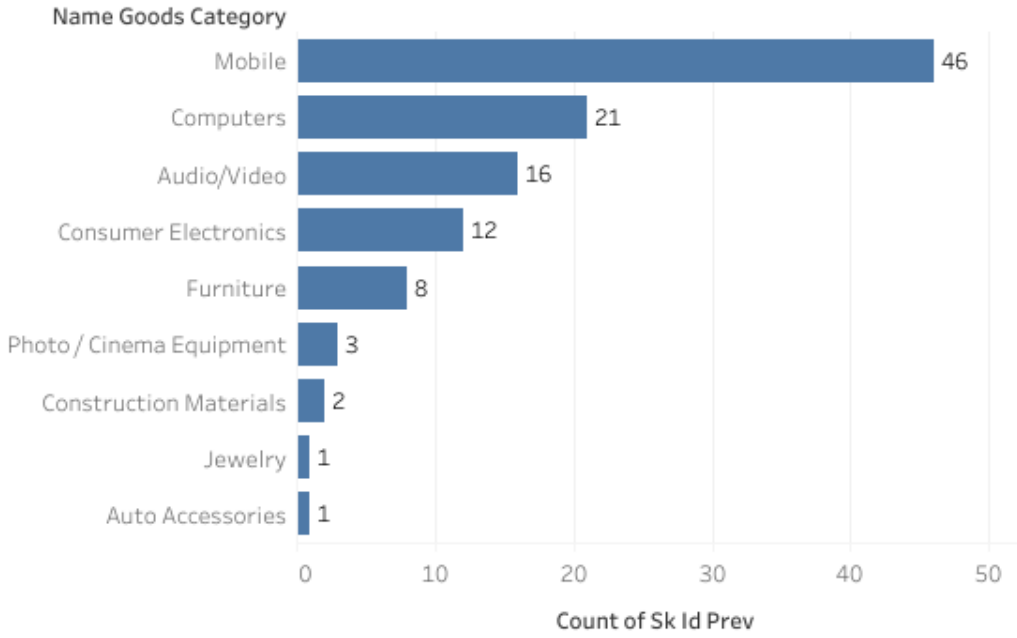


Top Cancelled Status Organization Types

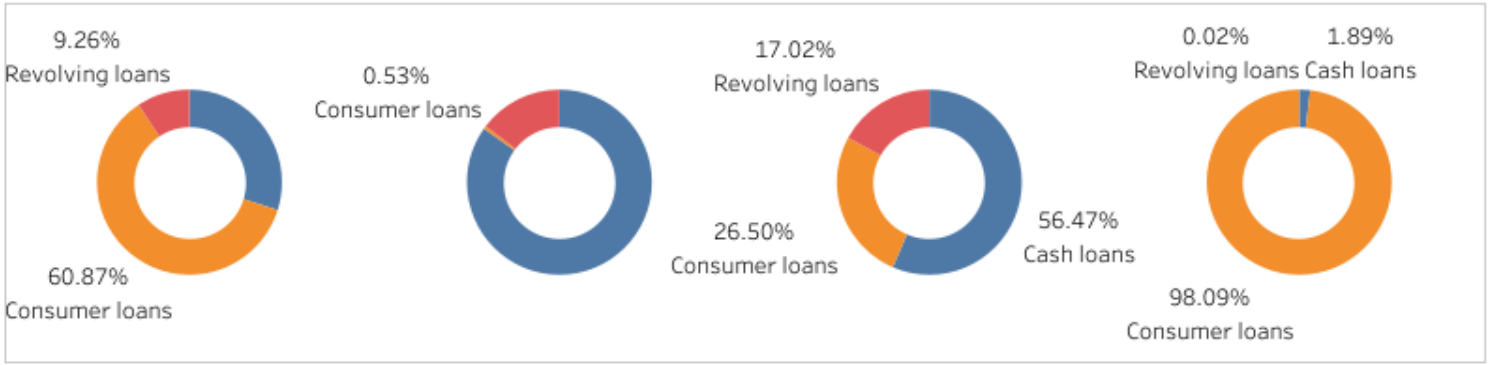
Name Portfolio



Top Cancelled Goods Category



Contract type



RECOMMENDATIONS

- Some key recommendations that could help bank minimize the risk factor of people avoiding payments so that bank could save money and avoid getting defaulted on payments.
  - Secondary / Secondary special as highest education should be avoided while approving a loan .
  - Banks should focus on Working class as they have most number of successful payments.
  - Most cancelled Loans are Cash type and POS mobile with interest is most unused.
  - Married people are more interested in taking loans .
  - Females tend to pay regularly as compared to Males.
  - Banks should focus on Middle aged clients as they are taking more loans and paying regularly.
  - Core Staff and Managers are regular payers and don’t come under defaulters.