Credit EDA Case Study

DRIVING FACTORS BEHIND LOAN DEFAULT

Contents

Problem Statement

A Consumer Finance company 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.

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 utilize this knowledge for its portfolio and risk assessment.

The aim of the Case Study is to identify the patterns which indicate if a client has difficulty paying their installments which may use for taking actions such as denying the loan, reducing the loan amount, lending at higher rate etc..

Data Sourcing

'application_data.csv' contains all the information of the client at the time of application.
The data is about whether a **client has payment difficulties.**

'previous_application.csv' contains information about the client's previous loan data. It contains the data whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**

'columns_description.csv' is data dictionary which describes the meaning of the variables.

Data Cleaning

Standardizing Columns

- Standardize the data in below columns by converting days to years in the data set.
  - DAYS_BIRTH
  - DAYS_EMPLOYED
  - DAYS_REGISTRATION
  - DAYS_ID_PUBLISH
- For the column DAYS_EMPLOYED, there are values which are greater than 0.
  - Since this column indicates number of days before application, the client has started the employment, we will replace the +ve values with the median of the remaining values.
- Create 4 new columns as below and impute the data:
  - Age
  - AGE_GROUP
  - YEARS_REGISTRATION
  - YEARS_EMPLOYED
  - YEARS_ID_PUBLISH

Data Cleaning

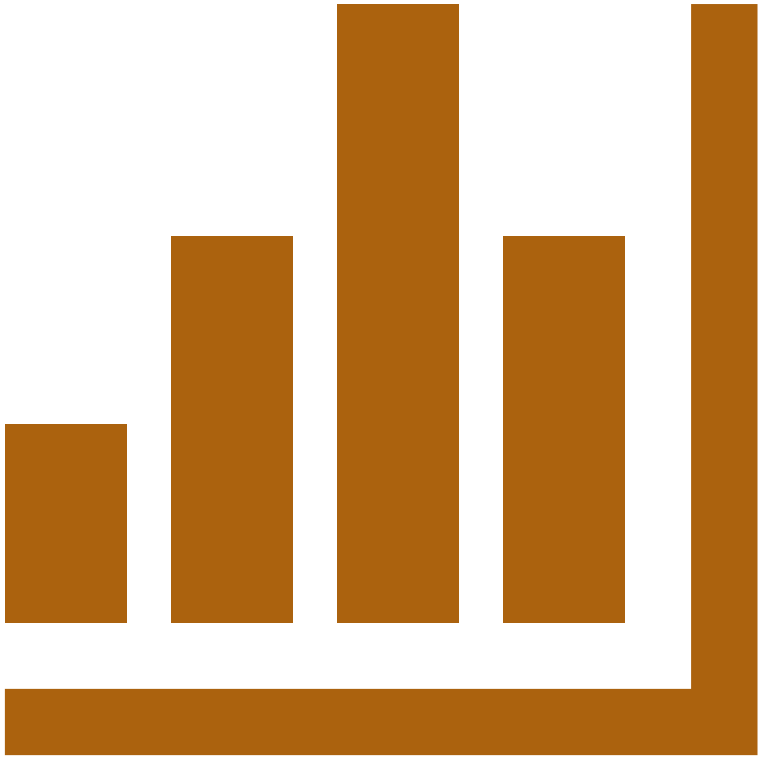
Missing Values

- Missing Values in AMT_ANNUITY, OWN_CAR_AGE, CNT_FAM_MEMBERS
- AMT_ANNUITY
  - There are only 12 null records for the same.
  - Of these 12 clients, 8 of them are aged between 30 and 60.
  - Since people of these age are more prone to have annuity funds, missing values should be replace.
  - Replace the missing values with median because mean might lower the variance of the data set.
- AMT_GOODS_PRICE
  - There are 278 records.
  - It is not necessary that always the company gives loan based on price of the goods.
  - Hence do not impute any data and leave as it is.

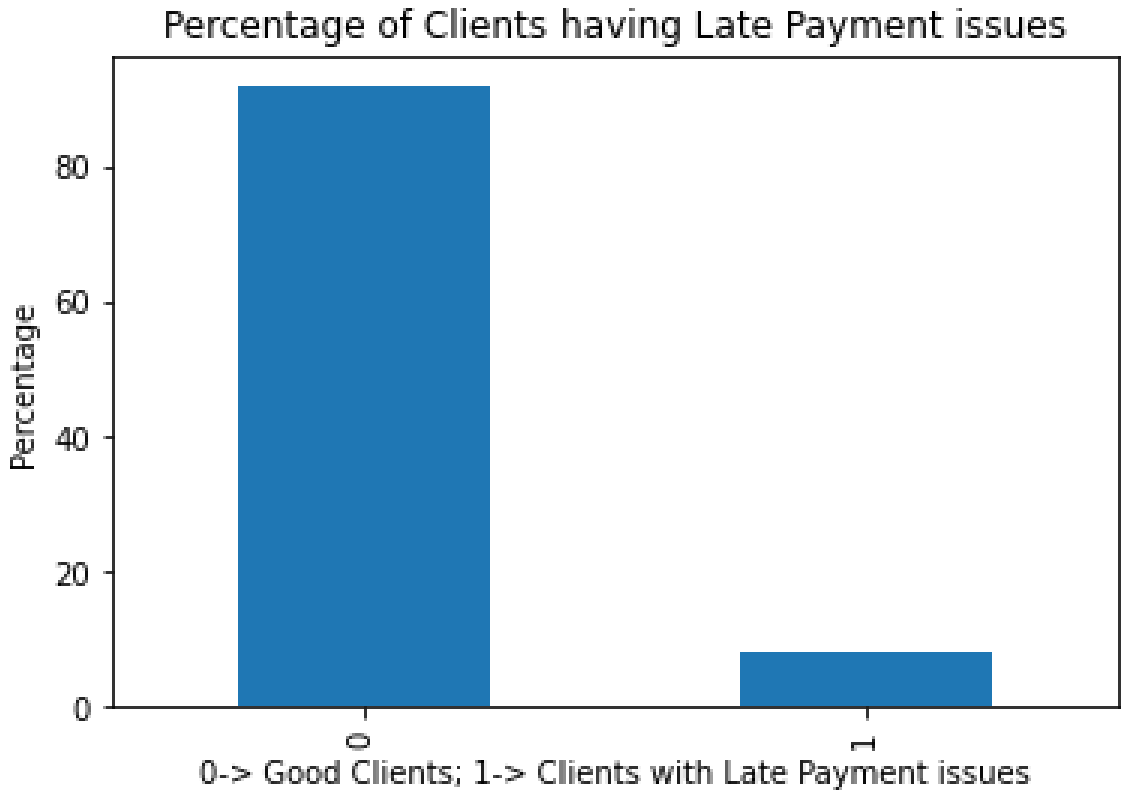
Data Cleaning

Missing Values

- OWN_CAR_AGE
  - There are 5 records which has FLAG_OWN_CAR as Y and OWN_CAR_AGE as null.
  - There are clearly many outliers for this columns. For now replacing the missing values(5) with mode(7).
- OCCUPATION_TYPE
  - We cannot infer the occupation type of a client based on his income/income type. Hence decided to leave this columns as is.
- CNT_FAM_MEMBERS
  - There are only 2 records which has CNT_FAM_MEMBERS missing.
  - As both mean and median are equal to 2, replacing the missing values with median.

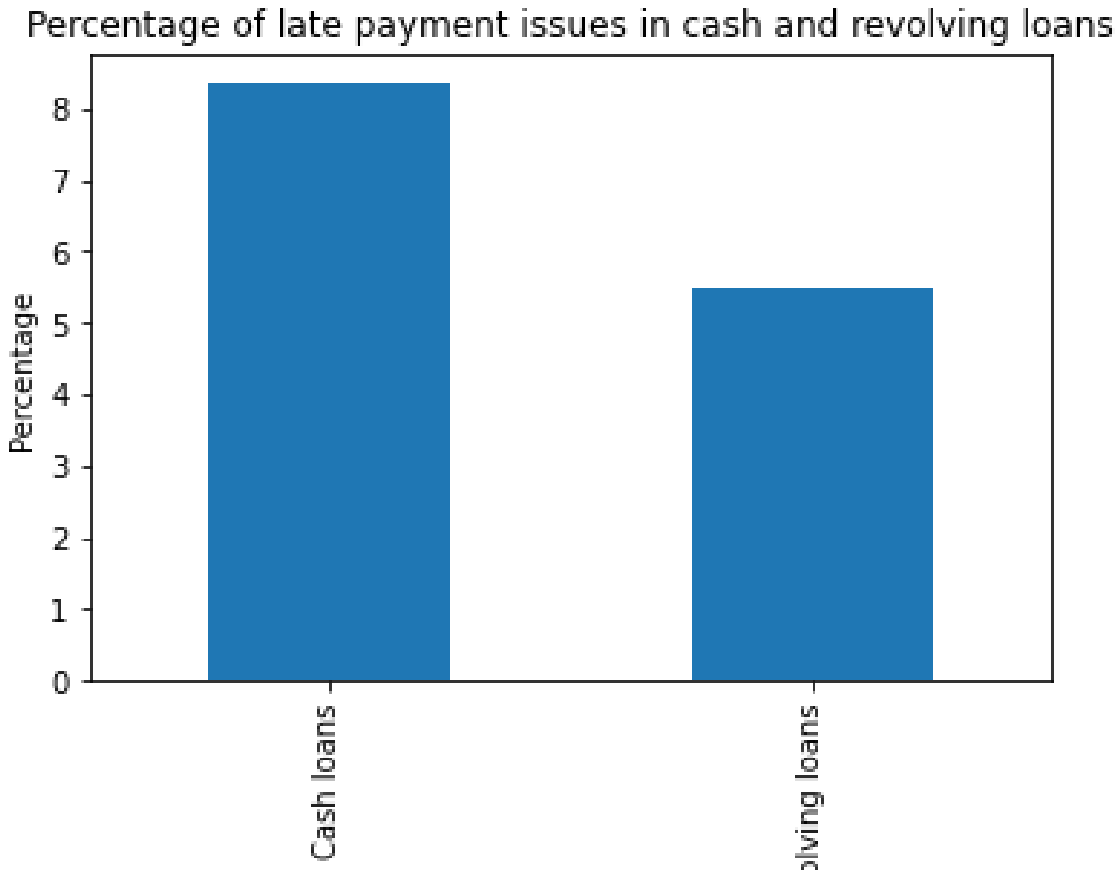


Data Analysis



Clients having Late Payment Issues

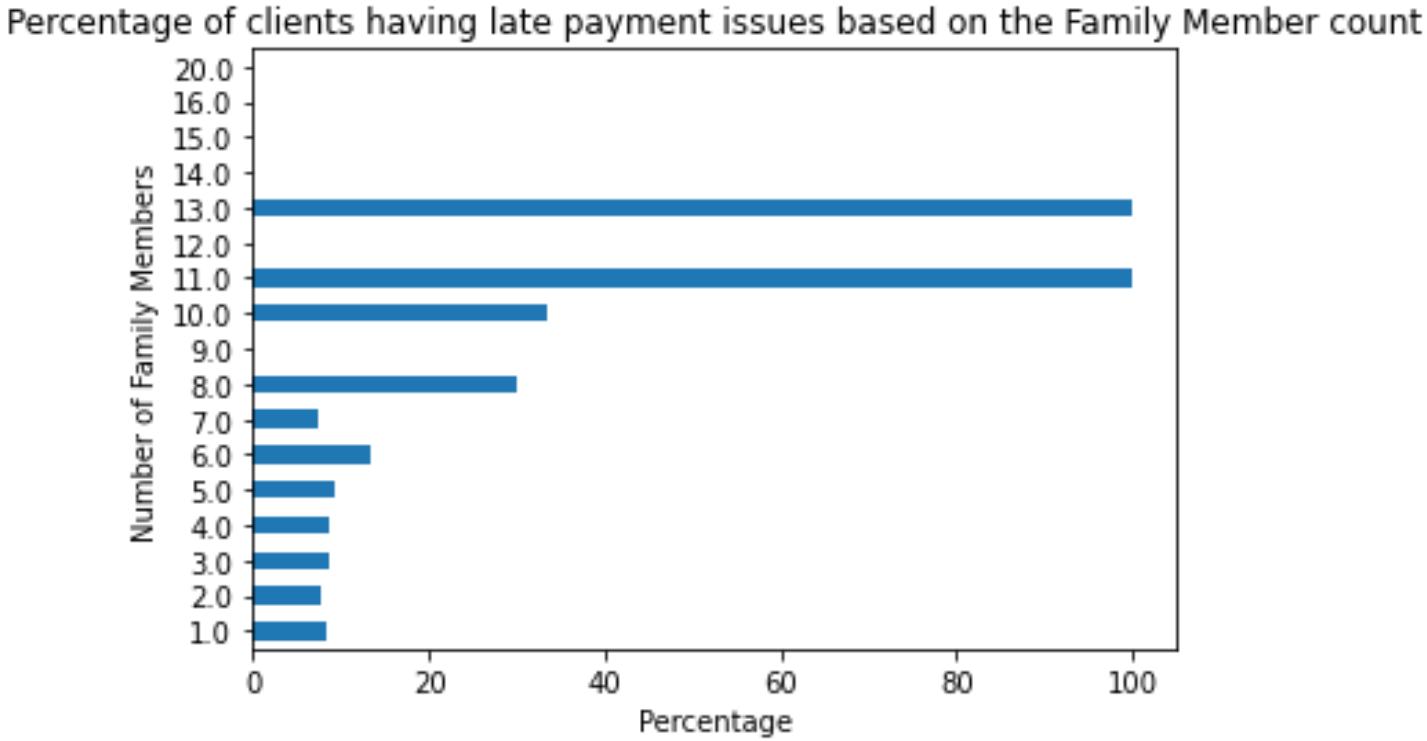
- Approximately 8% of the existing clients for the Company have Payment Difficulties(i.e. he/she had late payment more than X days on at least one of the first Y installments of the loan)



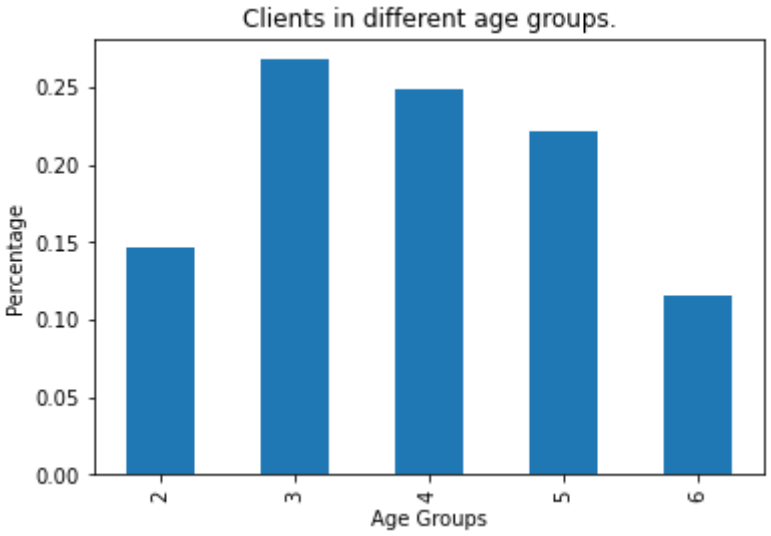
Late Payment Issues in different types of Loans

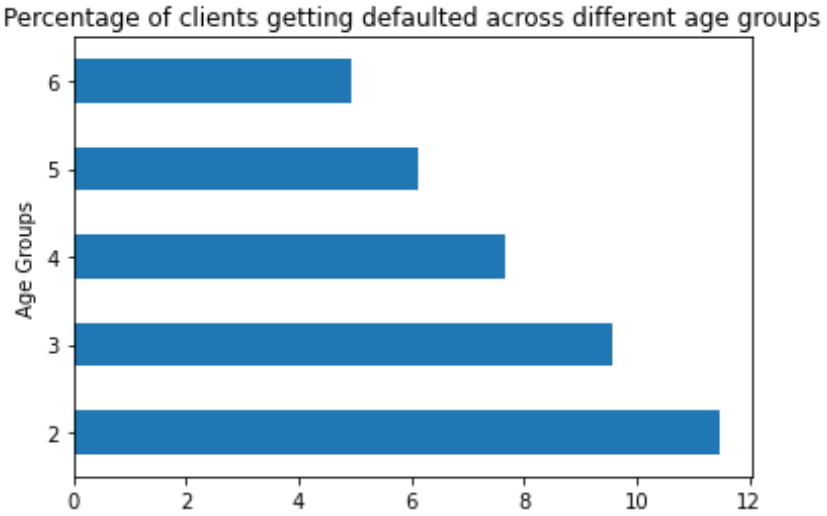
- ~85% of the Cash Loans are having Late Payment Issues.
- ~ 55% of the Revolving Loans are having Late Payment Issues.

Family Members Count



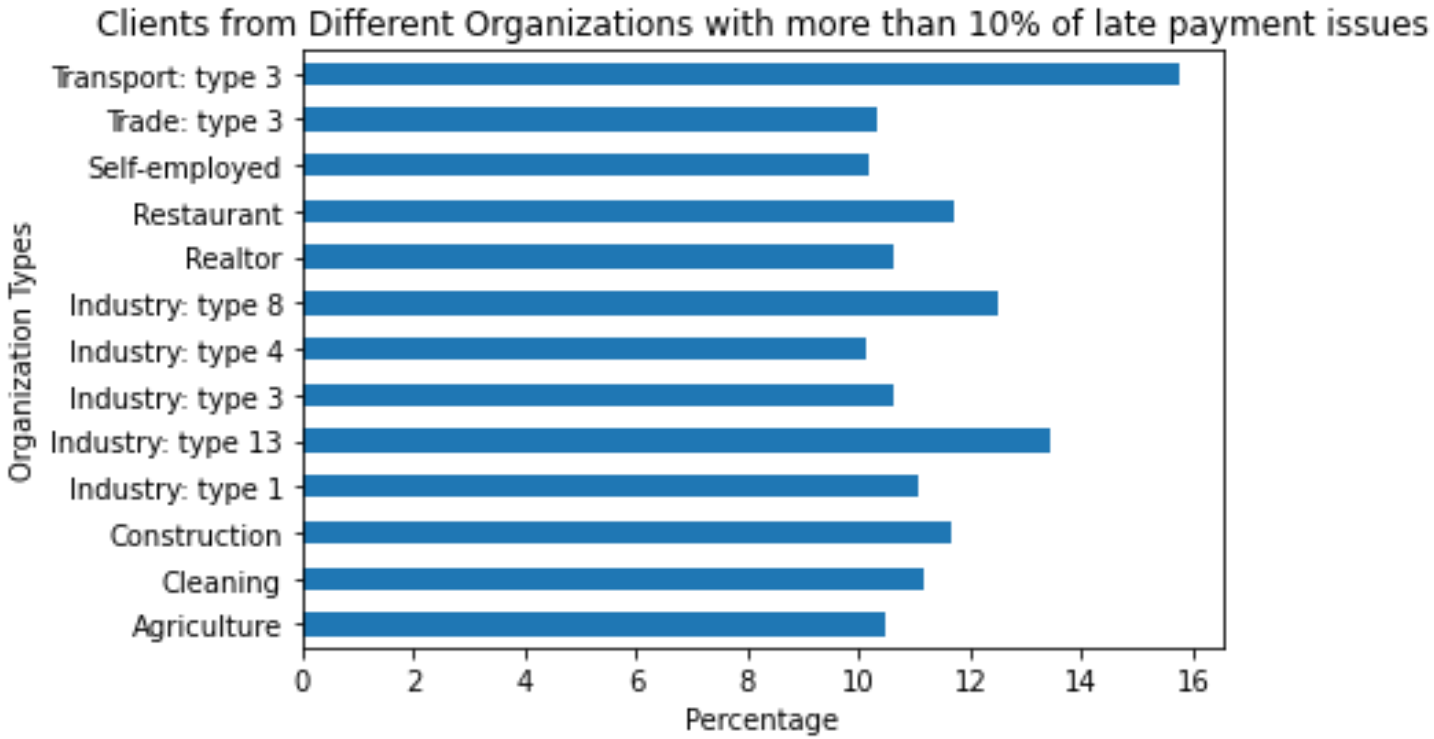
- A small inference can be drawn here that clients with more than 5 members in a family are more prone for late payment issues
- Also 100% of clients having 11 or 13 members in their family are defaulting at least once





Age Groups

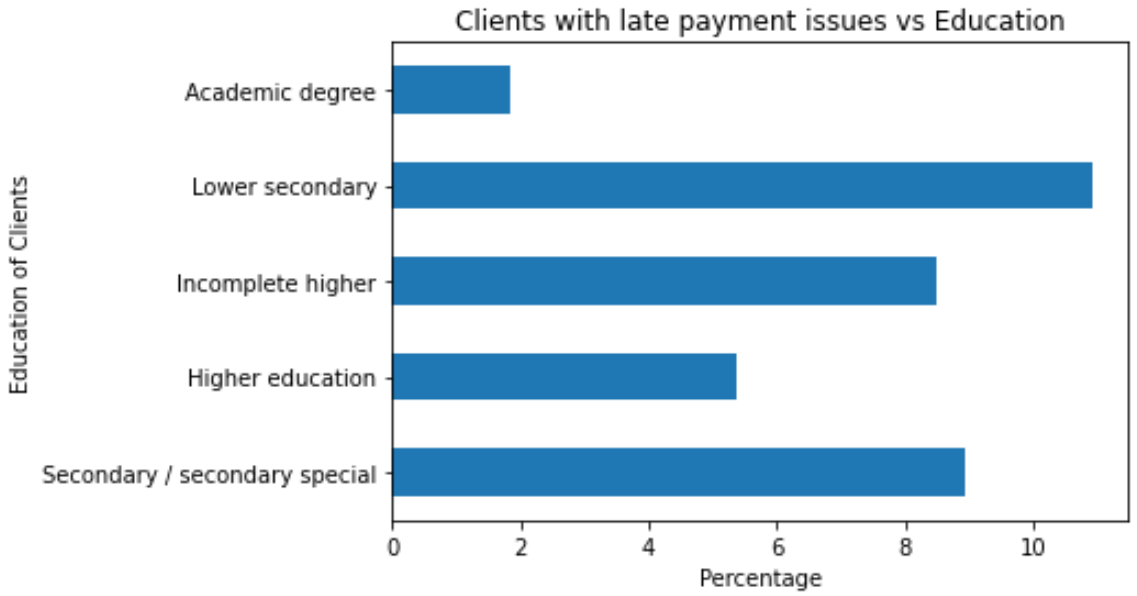
- Majority of the clients are from Age between 30-60.
- 11-12% of clients belonging to Age group 2(i.e. age between 20 and 30) are having high chances of having late payment more than X days on at least one of the first Y installments of the loan
- This trend gradually decreases with the increase in Age.



Organization Types

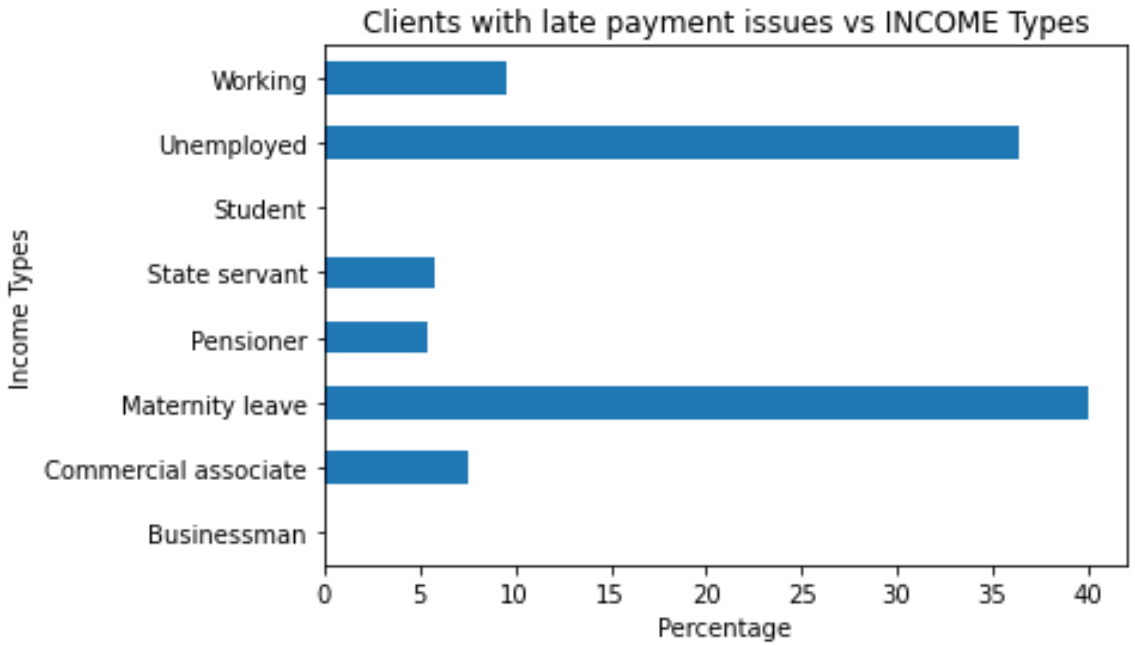
- 10% of more number of clients from the organizations listed beside are having late payment issues

Note: As there are many Organization Types given, the graph is plotting only the organizations that have more than 10% of late payment issues



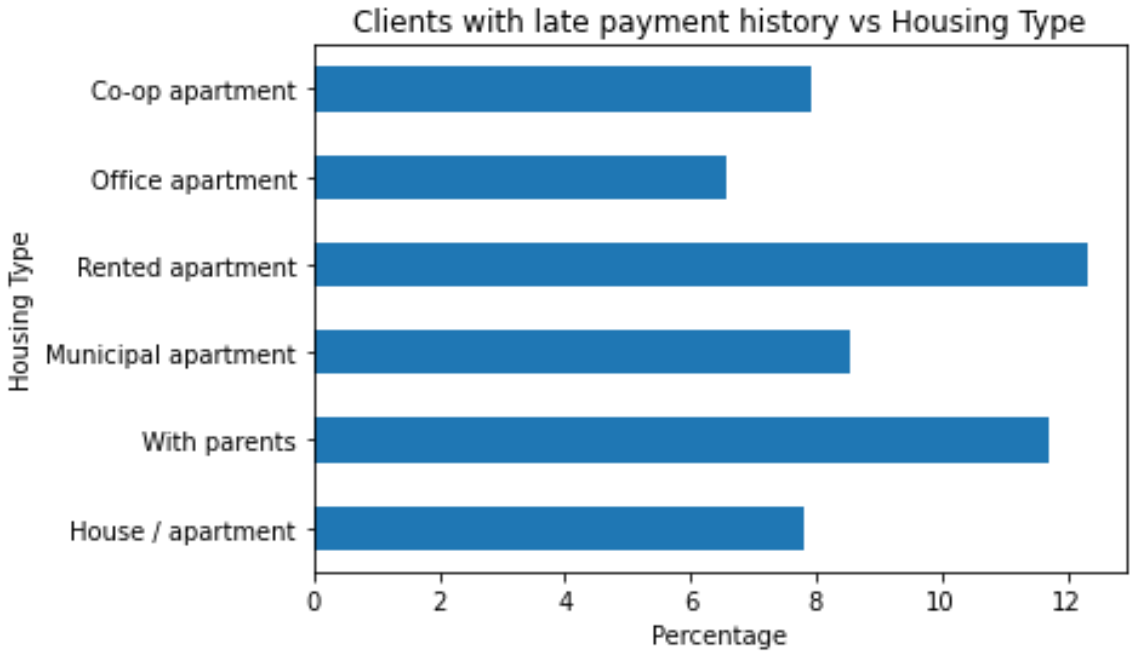
Education

- 8% or more clients who are under below categories having late payment issues.
  - Lower secondary.
  - Incomplete Higher.
  - Secondary/Secondary Special.



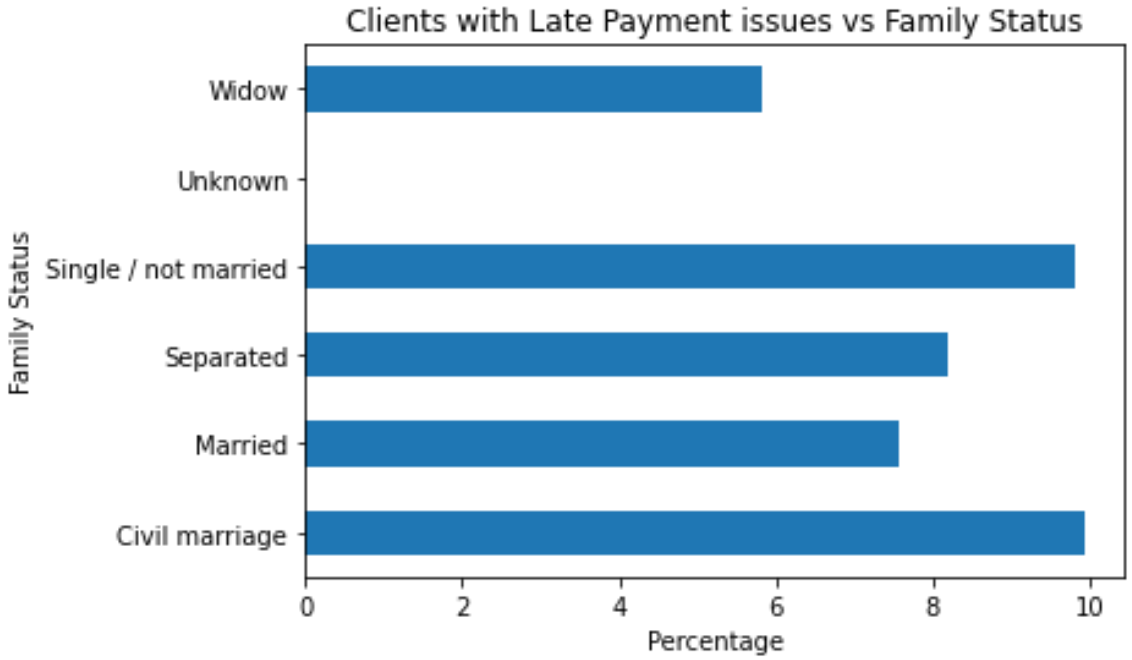
Income types

- It is clear that Clients belonging to the categories “Unemployed” and “Maternity Leave” are more prone towards late payment issues.
- 35% and 40% of Unemployed and Maternity Clients are already having late payment issues.
- Also it is good to note that 10% or more number of Clients belonging to the category “Working” are prone for late payment issues.
- Pensioners are having a low risk of 5%.
- It is good to note that Students and Businessman do not have any late payment history



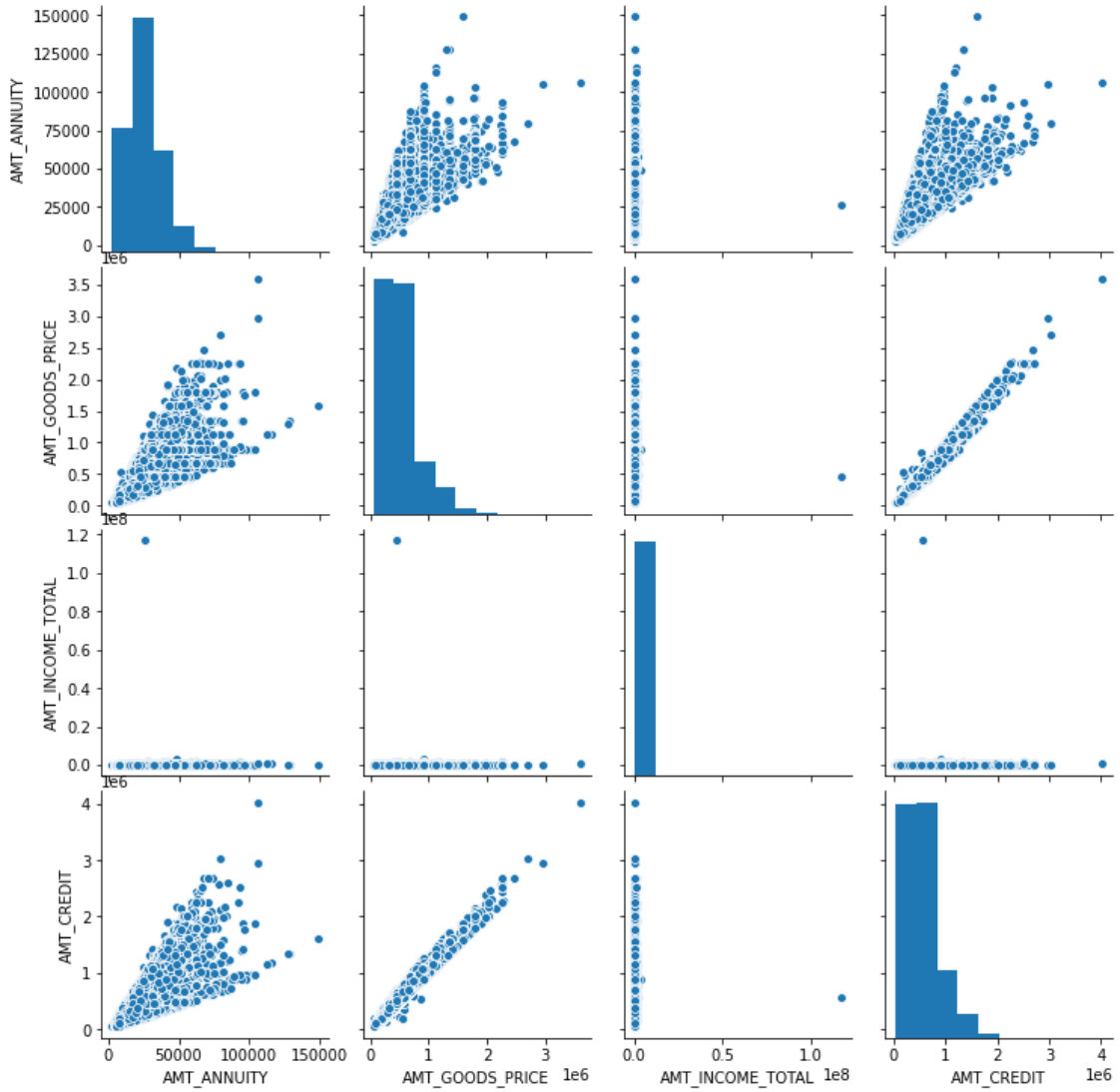
Housing types

- Approximately 12% of clients belong to the housing categories “With Parents” and “Rented Apartment” are having late payment history.
- Also it is important to notice at least 5% of clients belonging to all housing types are having late payment history.



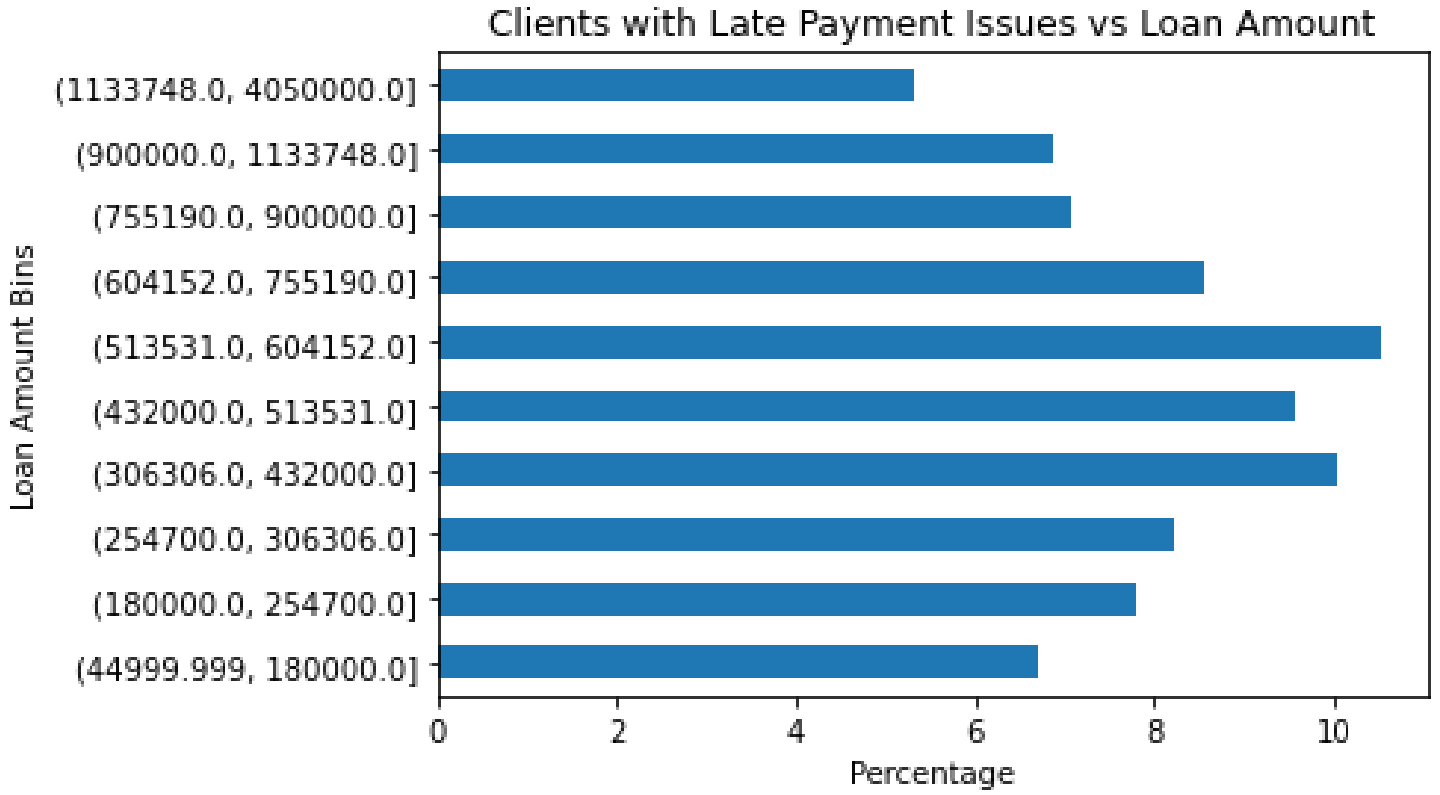
Family Status

- There are 2 clients with “Unknown” Family Status and they do not have late payment history.
- All other Family Statuses are having at least 5% late payment history.
- It is important to note clients having Family status as “Civil Marriage” and “Single” are more prone for late payment issues.
  - The risk in above categories can be mitigated by increasing the interest rates for the same.



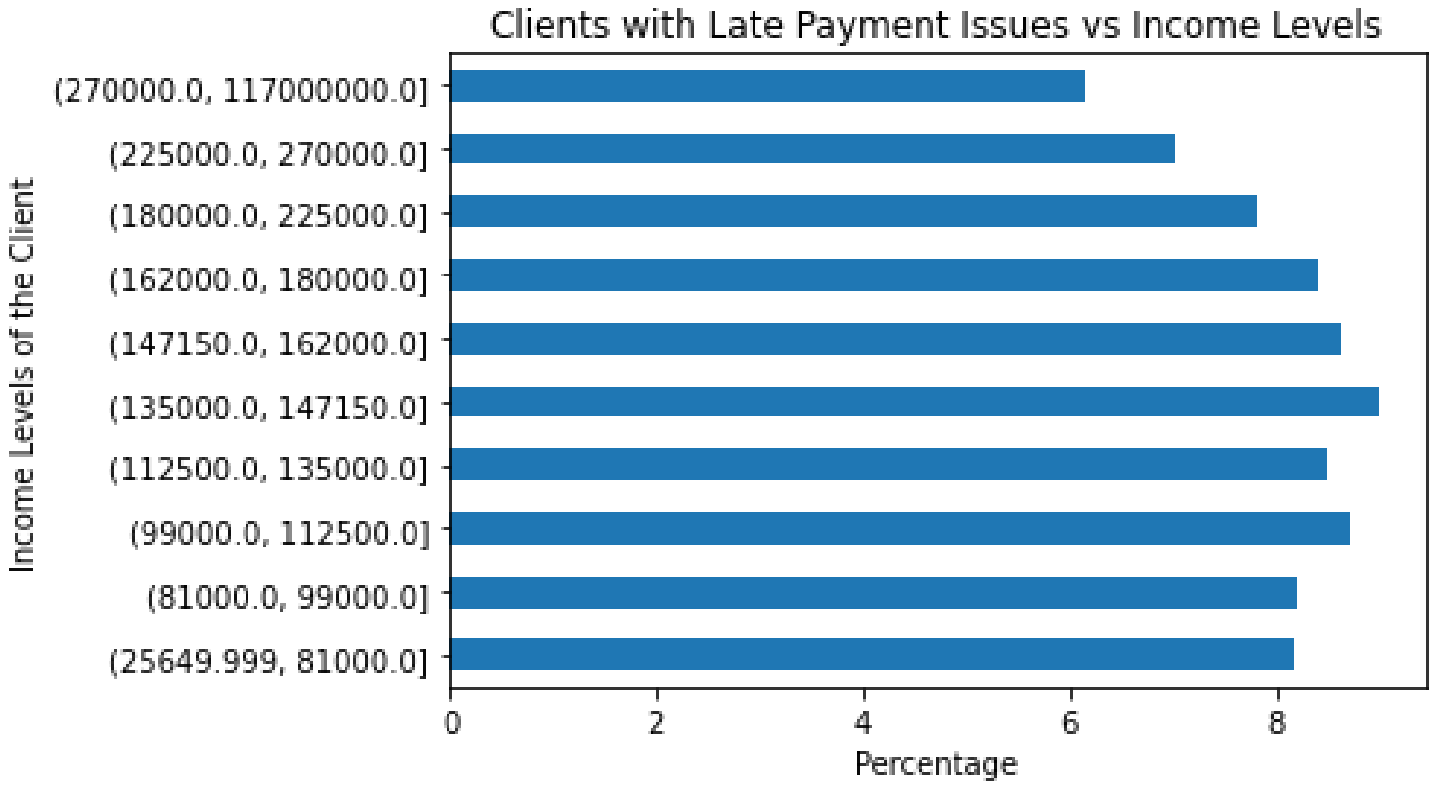
Correlation between different columns

- From the given graph we can identify that:
  - AMT_CREDIT, AMT_ANNUITY, AMT_INCOME and AMT_GOODS_PRICE are having a positive correlation.
  - We can infer that as the income increases, annuity amount increases and the loan amount for which the client applies will also increase



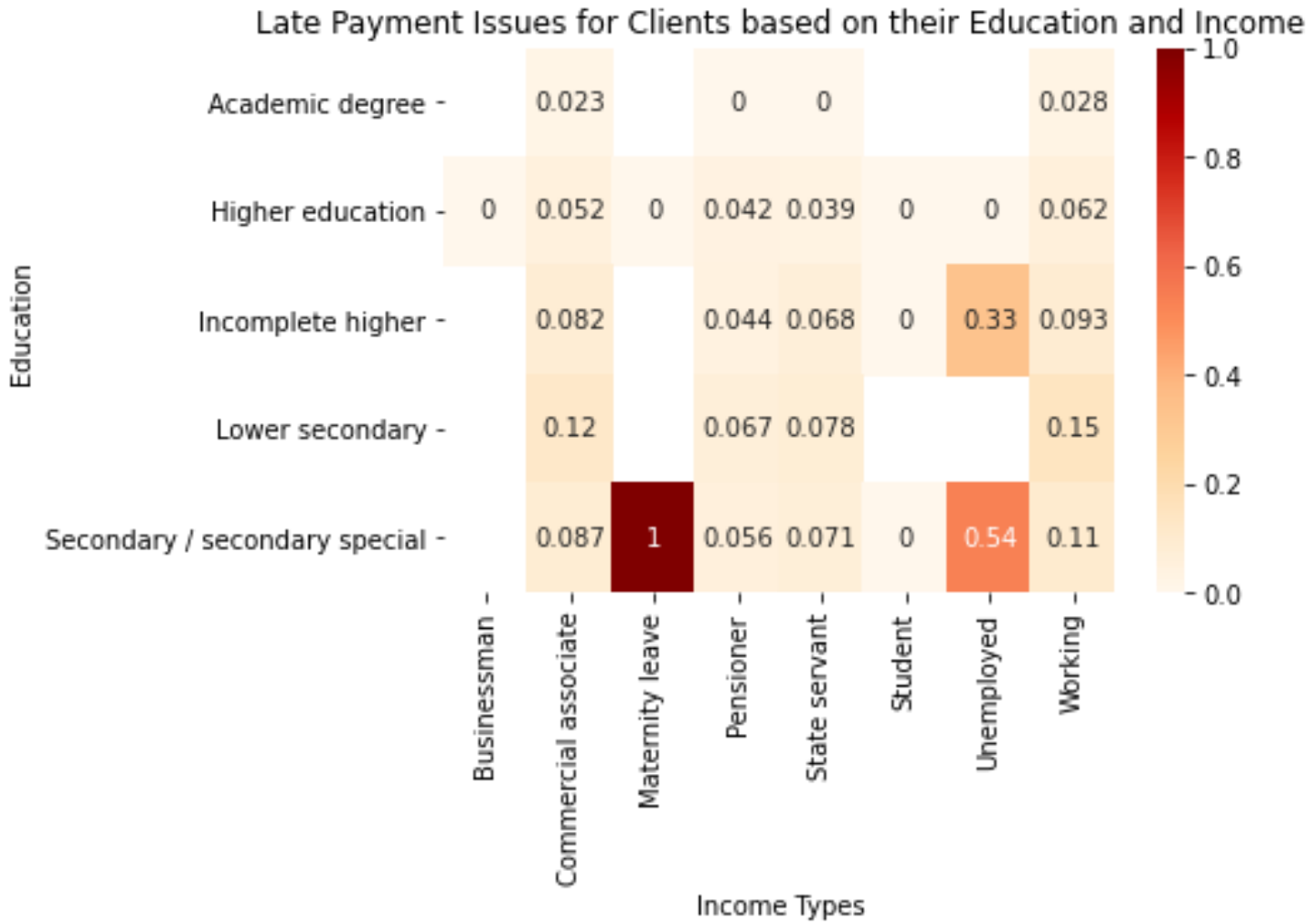
Loan Amount

- 6% of the Clients with loan amount less than or equal to 1133748.0 are having at least late payment issues.
- 10% or more clients with the loan amount ranging between 306306 to 604152 are having late payment issues.
- It is good to keep all the clients and adjust the interest rates according to the risk



Income Levels

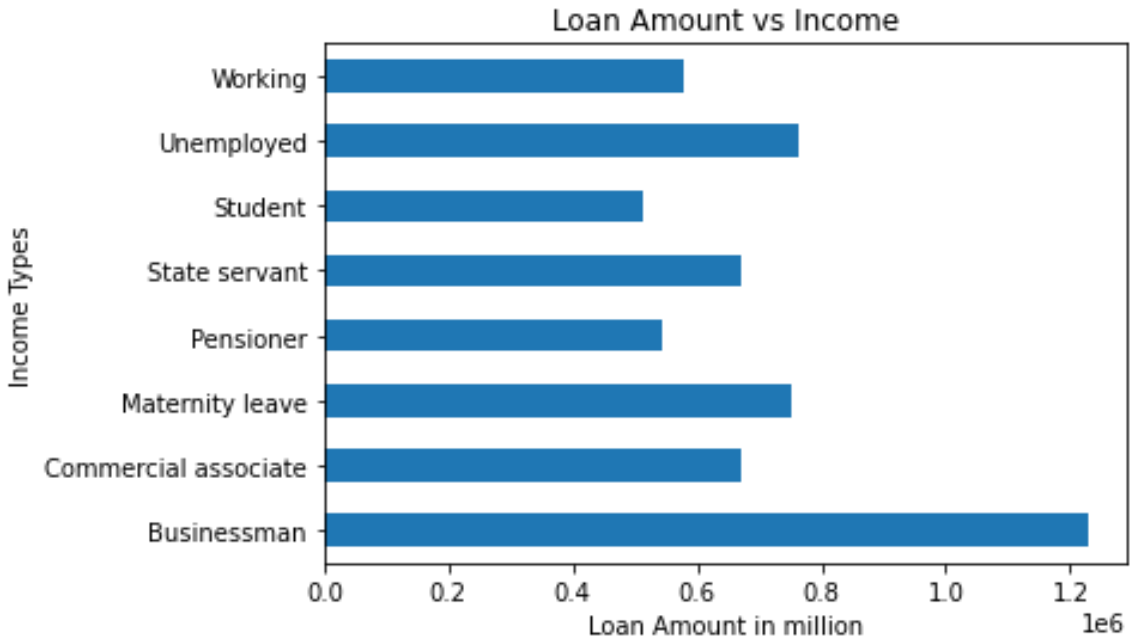
- 8% or more Clients having income ranging between 25000 to 180000 are having late payment history.
- Again, adjusting the interest rates for these clients would be a good option.



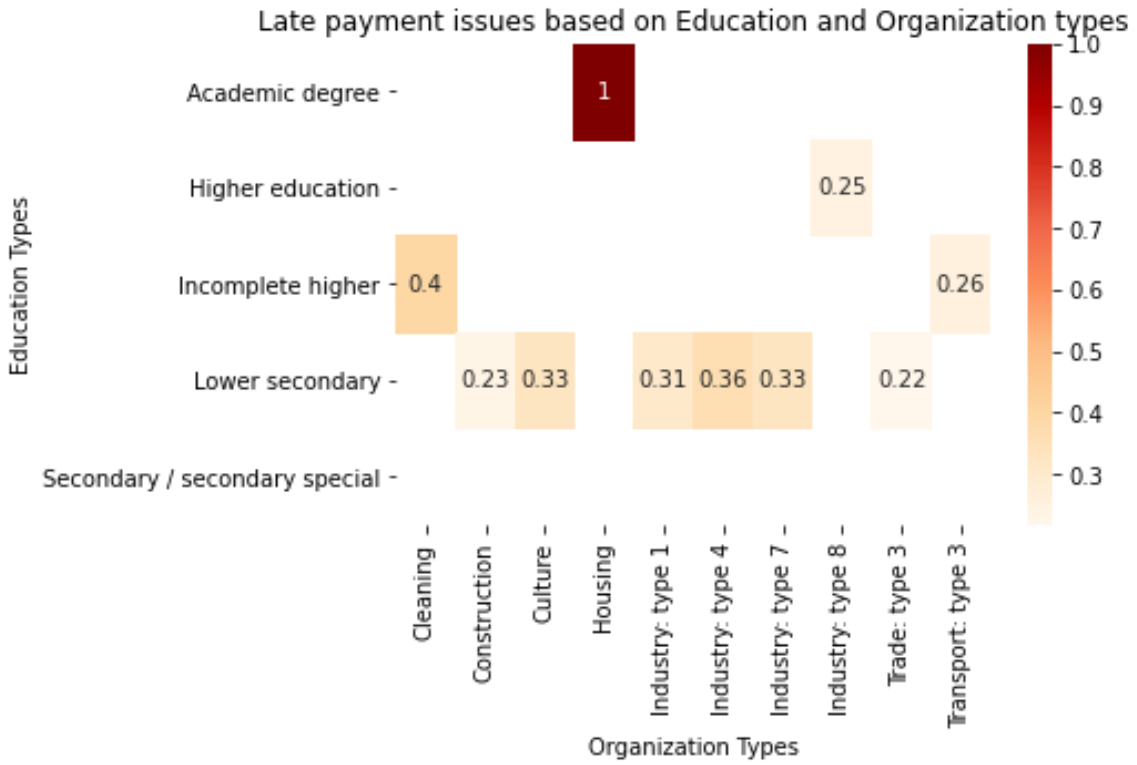
Education vs Income

- The heatmap shown in this slide is plotted late payment issues against Education and Income Types of the clients
- We can infer that Clients with Maternity Leave having Secondary education is the most prone area.
  - Company can take a decision to sanction a loan or not for these clients
- Also it is important to note that 30% or more Unemployed clients with Secondary Education and Incomplete Higher Education are prone for late payment.
  - Company can adjust the interest rates for these clients.

Income vs Loan Amount

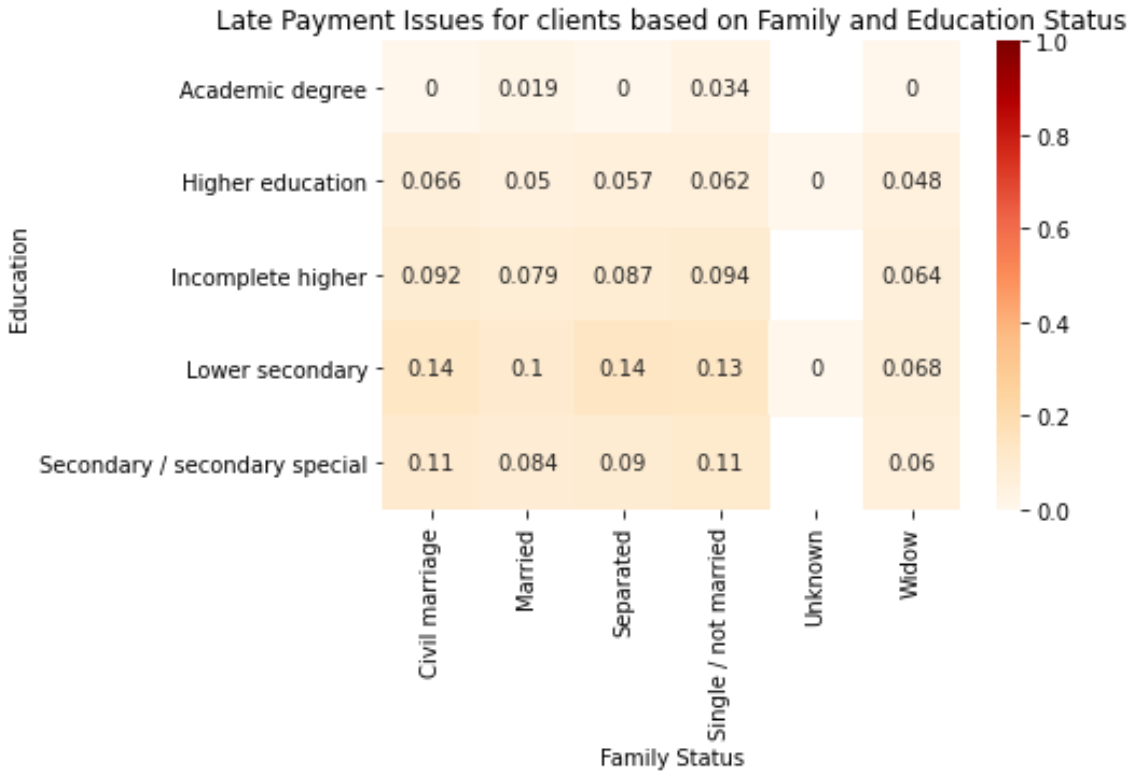


- We should observe that in spite of being a high risk category, Maternity and Unemployed clients are having the loan amounts of more than 0.5 million.
  - Company should have a maximum threshold amount of loan for these categories
- Though Businessman are having maximum loan amount this zone is risk free.



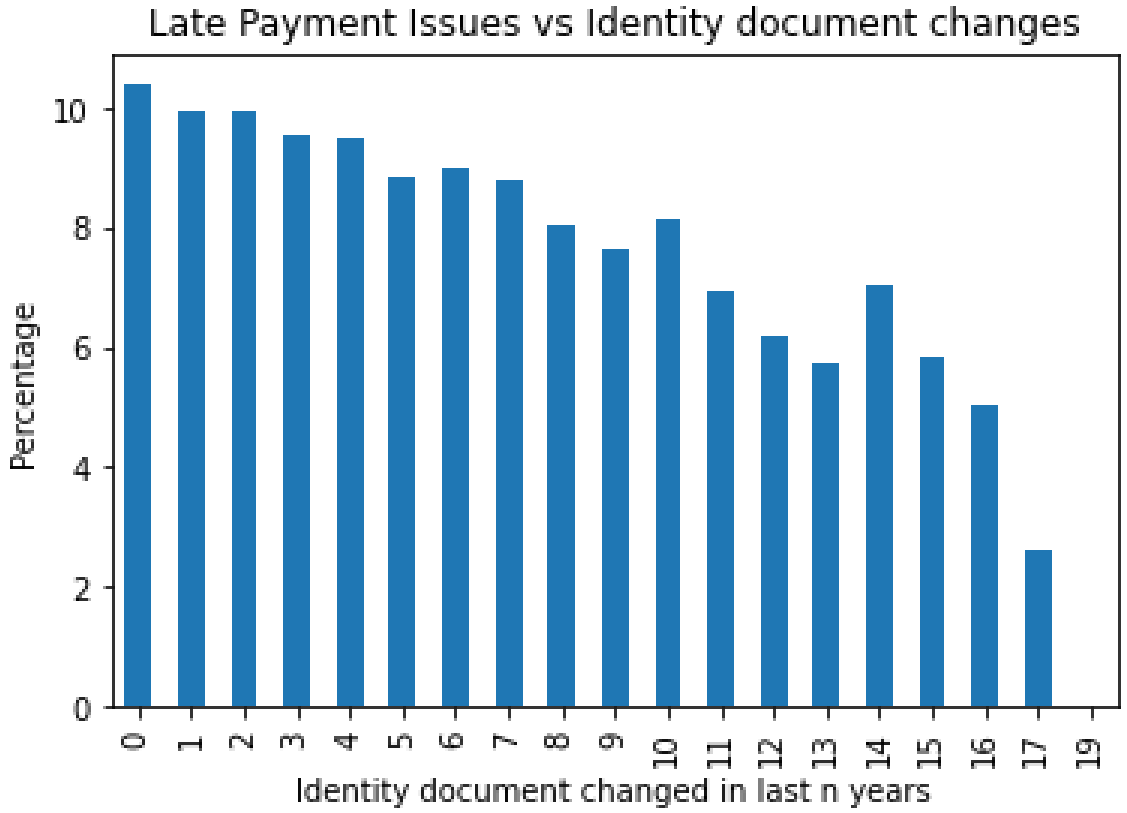
Education vs Organization Types

- Clients working in Housing Organization having Academic degrees are more prone(100%) for late payments.
  - Company can take a decision for granting loan for these clients.
- Also it is important to note that 30% or more clients working in Industry Type 1,4,7 and having Education Types Incomplete Higher and Lower secondary are having late payments.
  - Company can think on increasing interest rates for these clients.



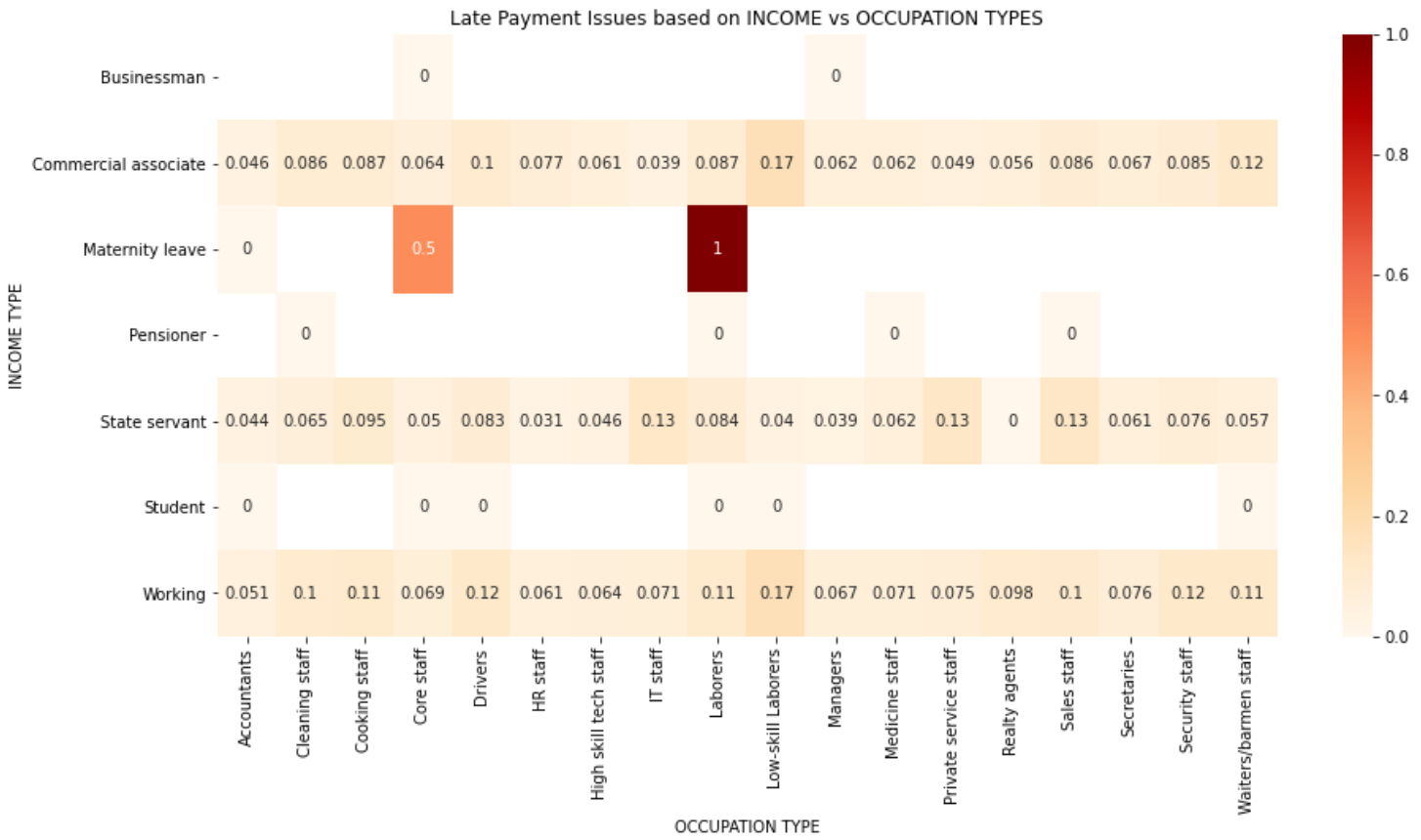
Education vs Family

- Very clearly 10% or more clients with Lower secondary education having below family status are having late payment issues.
  - Civil Marriage.
  - Married.
  - Separated.
  - Single
- Company should consider these categories while granting loans.



Identity Document Updates/Changes

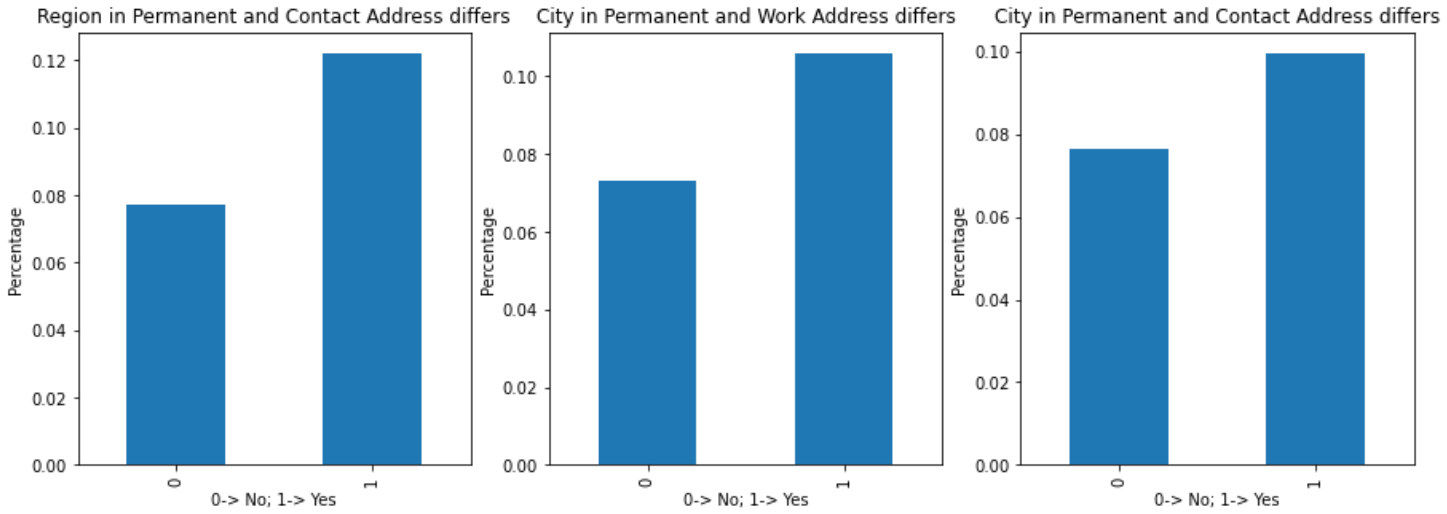
- More than 8% of Clients who have changed their identity document up to 10 years before are having late payment issues.
- Company can increase the interest rates based on how recent the client has changes his/her identity document.
- Also Clients who provided only email or home phone are having more than 10% of risk for late payments(please refer to Jupyter notebook for table)



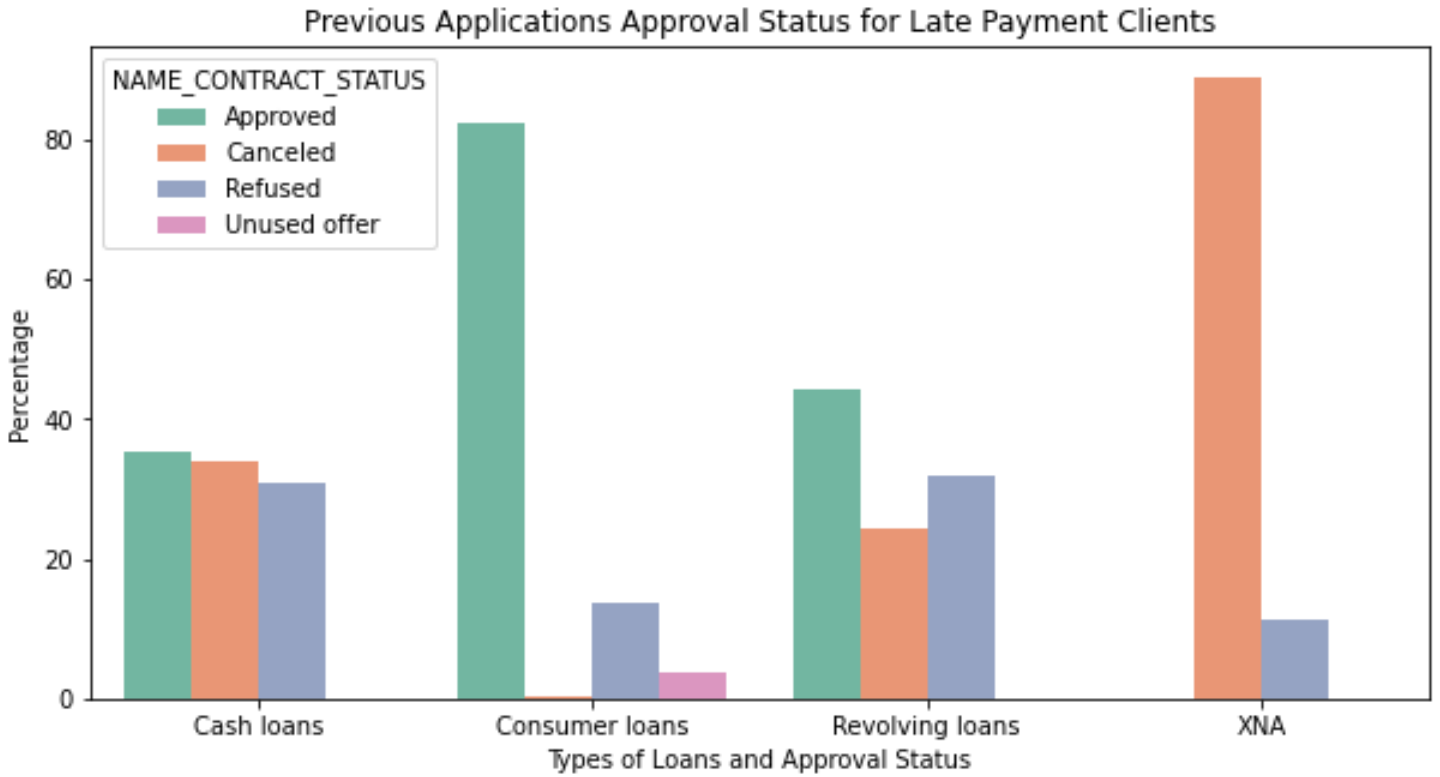
Income vs Occupation Types

- Laborers having Maternity leave are 100% prone for late payment.
- Core Staff with Maternity are 50% prone.
- Company should keenly consider the categories Maternity, Laborers and Core Staff while granting the loan.
- Below combinations are atleast 10% prone.
  - Drivers, Laborers, Cleaning Staff, Cooking Staff, Low-skill laborers, Sales staff, Security and Waiters under Working Category.
  - IT, Private Service and Sales Staffs under State Servants category.
  - Waiters, Low skill laborers and Drivers under Commercial Associate Category.

Address provided by Client

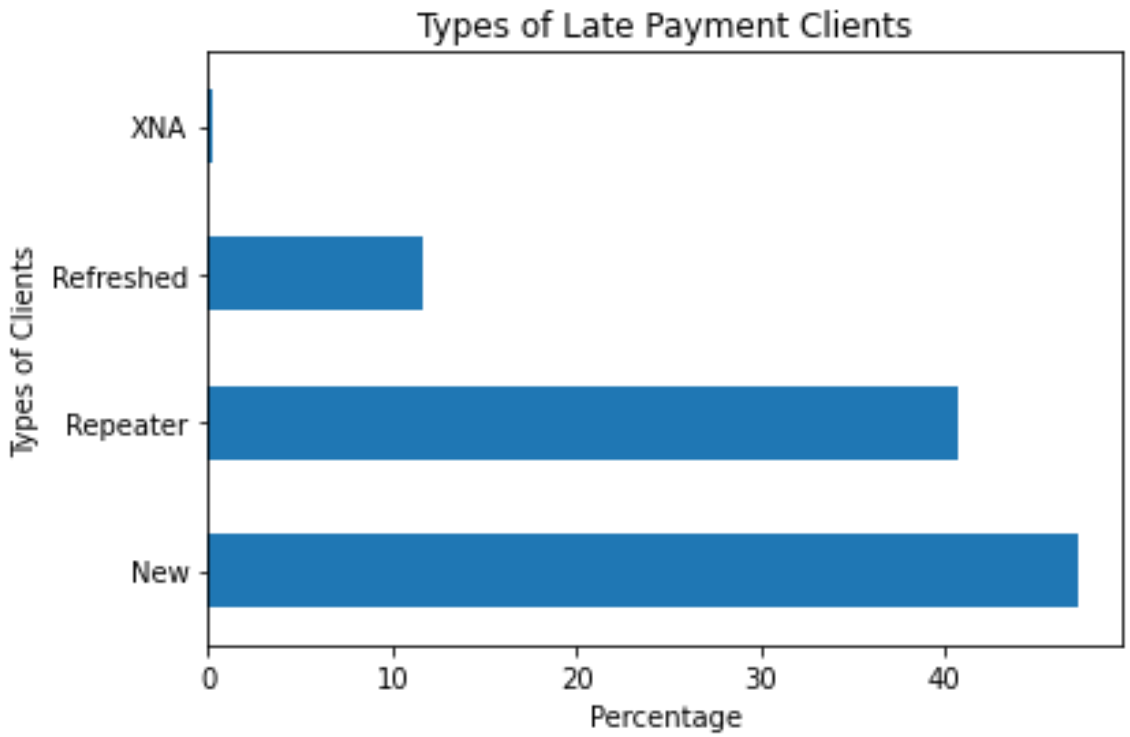


- Clients for whom the permanent address not match contact address(REGION), contact address not match work address(CITY) and permanent address not match contact address(CITY) are having 10% of late payment issues
- A thorough Back Ground Check for the clients could avoid the issue.



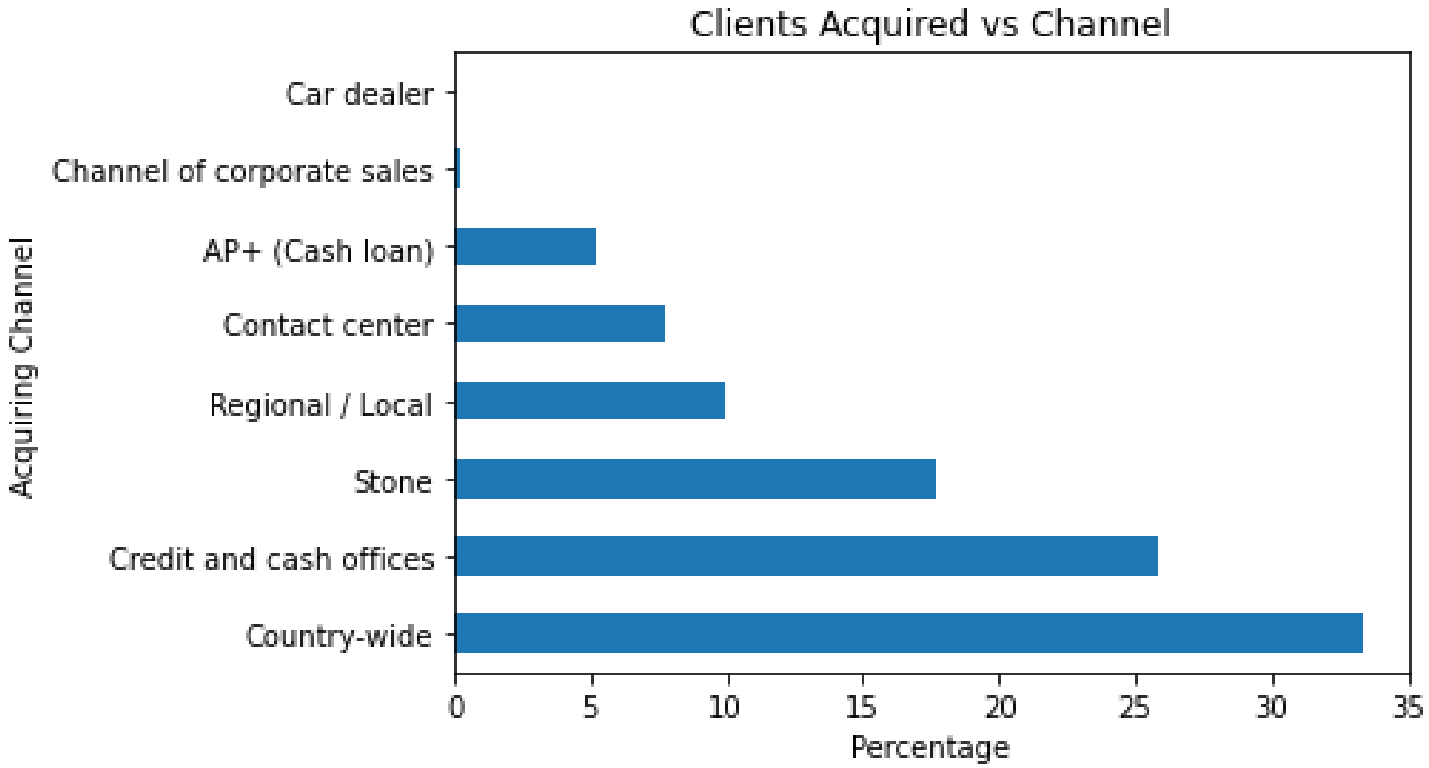
Approval status of previous application for Late Payment Clients

- More than 30% of Revolving loans and 35% of Cash Loans for the late payment clients were Refused by the Bank



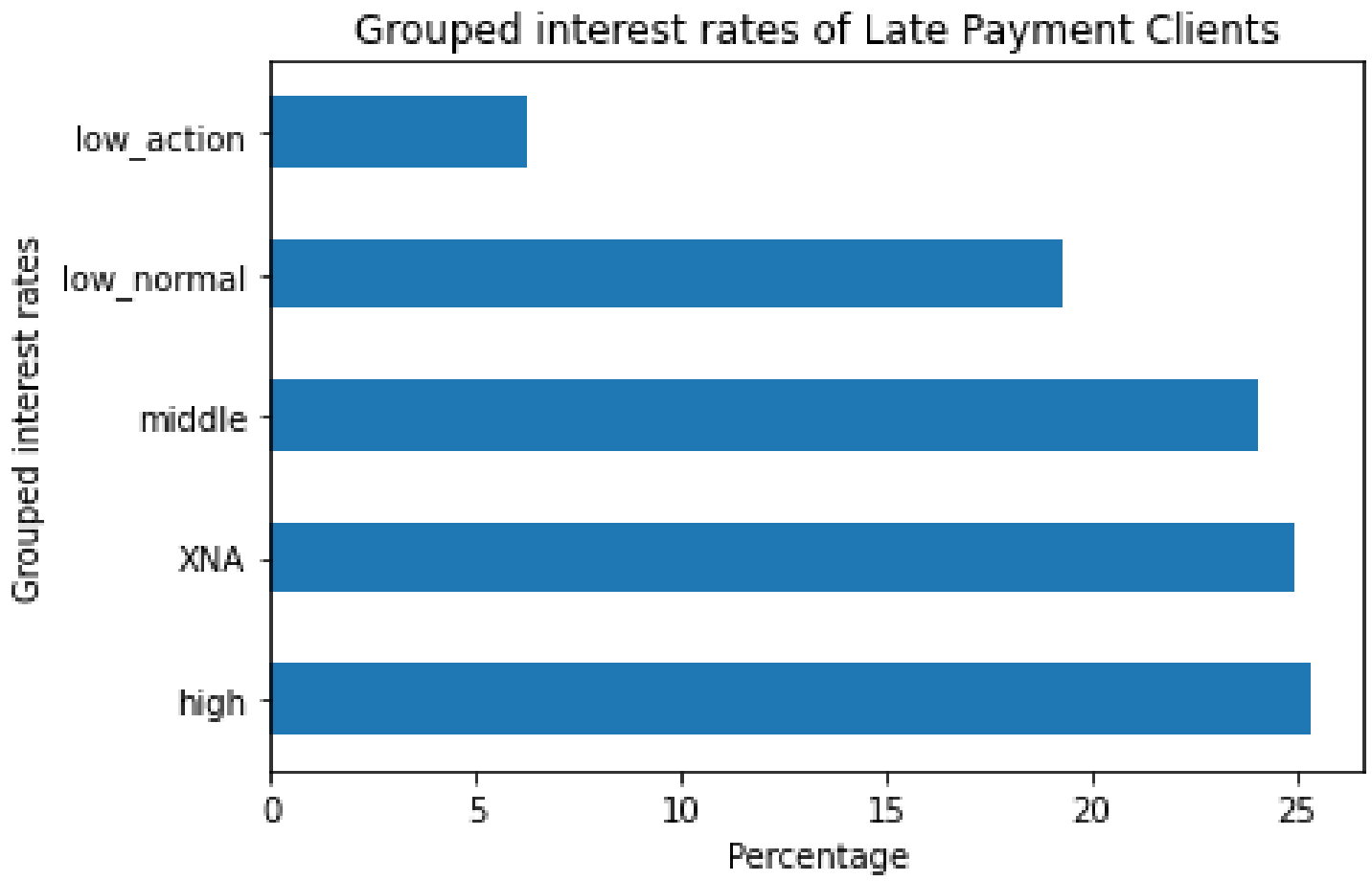
Types of Late Payment Clients

- More than 40% of clients with late payment history are either New Clients or Repeated to the Company.
- ~12% of Late Payment Clients are Refreshed and ~1% belong to XNA Category.



Late Payment Clients vs Acquiring Channel

- Majority of the Late Payment Clients (more than 30%) are acquired by Country wide.
- From Credit and Cash Offices ~30% of clients are acquired.



Grouped Interest rates of Late Payment Clients

- Majority of the late payment clients are from High, XNA and Middle interest rates.
- From low_normal there are ~20% of late payment clients

Thank You

- Team:
  - Srihari K S S.
  - Sandeep Rana
- Program:
  - Data Science March 2020.
- Email Ids:
  - srihariкss@gmail.com
  - sandeep.ece.111090@gmail.com