EXPLORATORY DATA ANALYSIS ON BANK LOAN CASE STUDY

Description:-

This case study aims to give you an idea of applying EDA in a real business scenario. In this case study, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

Business Objectives:

It aims to **identify patterns** which indicate if a client has difficulty paying their installments 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 utilize this knowledge for its portfolio and risk assessment.

To develop your understanding of the domain, you are advised to independently research a little about **risk analytics** – understanding the types of variables and their significance should be enough)

Overall Approach of the analysis:-

- Data cleaning.
- Data imbalance.
- Univariate Analysis.
- outliers Analysis.
- Correlation.
- Top 10 correlation.

- For full solution click on the link below:-

https://docs.google.com/spreadsheets/d/1qW8_9l0_GphW-m-IP5ka57eaiYqKy-9Y/edit?usp=sharing&ouid=107932508938240092754&rtpof=true&sd=true

1.Data cleaning

Approach followed:-

- Count of missing data per column and missing Percentage.

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE
blank cells	0	0	0	0	0	0	0	0	0	12	278
% of blank cell	0	0	0	0	0	0	0	0	0	0.003902299	0.09040327

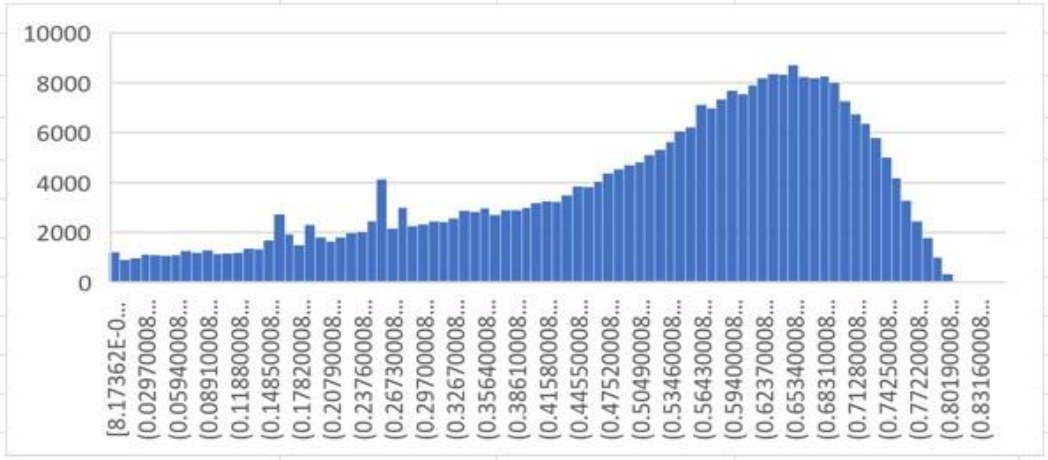
- Categorical Data
- Numerical Data

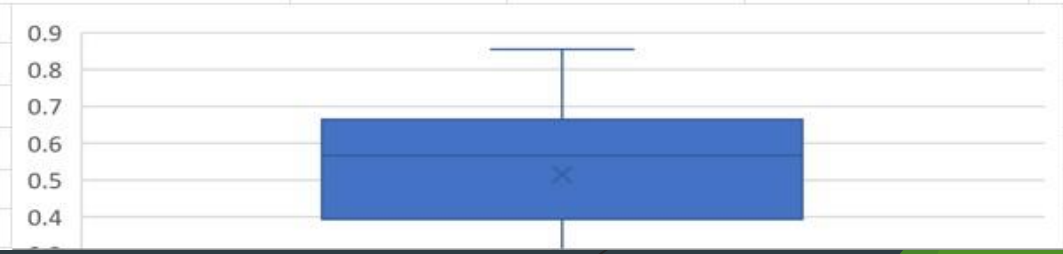
2.EXT_SOURCE_2 Column

Lets see the description of the column

count	307511
mean	0.514392674
std	0.191060155
min	8.17362E-08
25%	0.392457416
50%	0.565961426
75%	0.66361709
max	0.854999666

IQR	UL	LL
0.271159674	1.0703566	-0.0143





Data imbalance

imbalance percentage= 100* number of cases of 1 / number of cases of 0

TARGET	Perc_target
0	91.93%
1	8.07%
Grand Total	100.00%

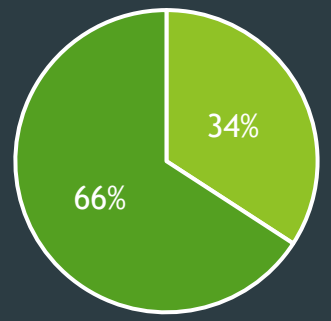
Percentage imbalance
8.7822

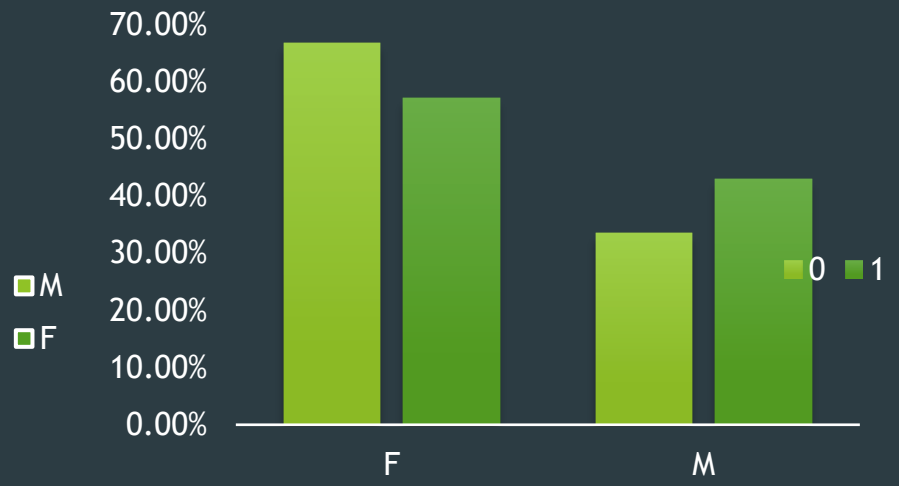
- Medium imbalance.

Univariate Analysis

Approach:-

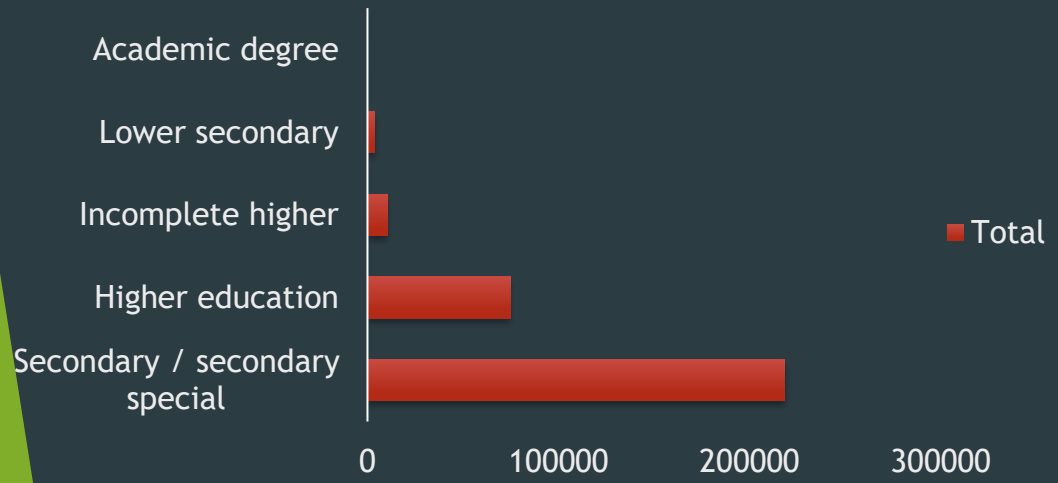
- Categorical columns:-
- Eg. Gender

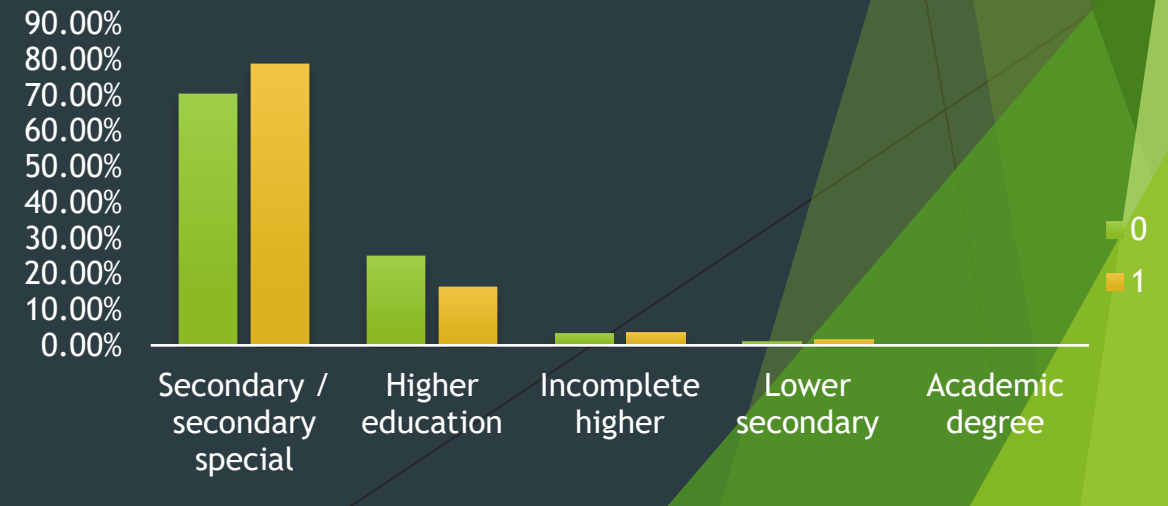




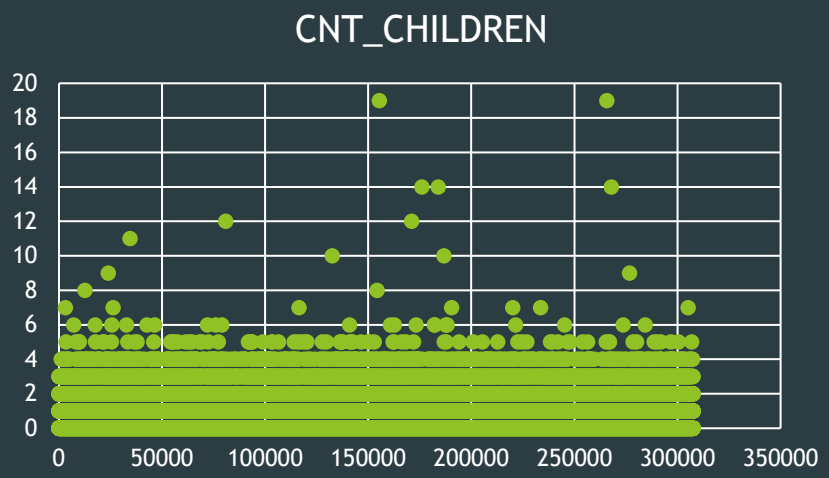
Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases)

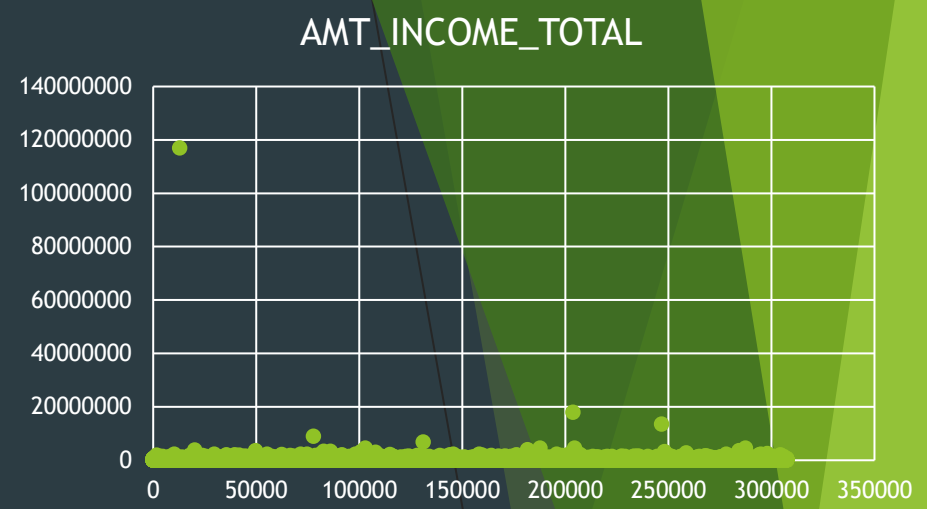
- loans taken by female are more than male but Also the defaulter rate is more in male.
- Eg2. Education -most of the people who took loan are from secondary education followed by higher education but the defaulters are more in secondary than higher education.

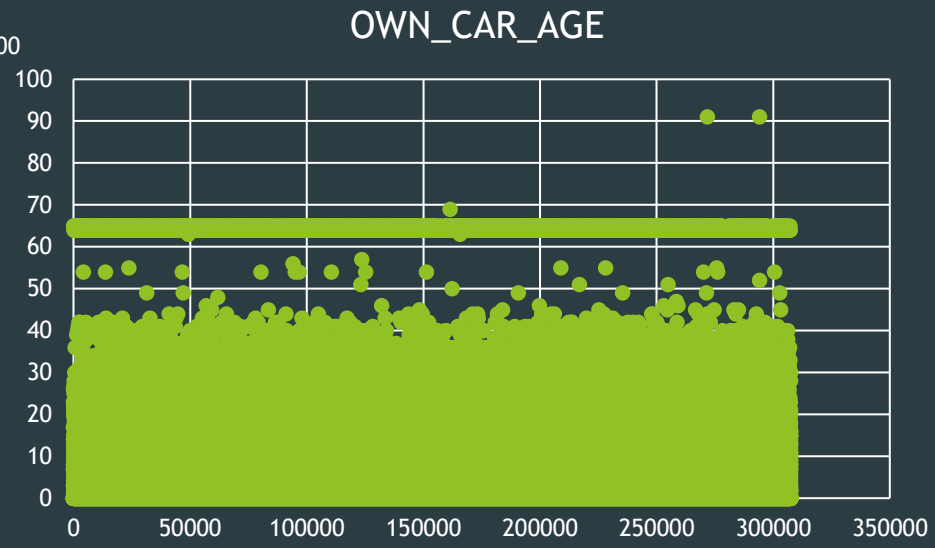




Outliers Analysis







- <u>Reports</u>:-
- CNT_CHILDREN - the count greater than 10 are extreme cases are considered as outliers
- AMT_INCOME_TOTAL - Income of the client - One value with value 117000000 seems to be outlier as the value is significantly far away form others and can be removed.
- OWN_CAR_AGE- The value with car age > 90 can be considered as outliers

Correlation matrix

	SK_ID_CURR	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_P
SK_ID_CURR	1.00							
TARGET	0.00	1.00						
CNT_CHILDREN	0.00	0.02	1.00					
AMT_INCOME_TOTAL	0.00	0.00	0.01	1.00				
AMT_CREDIT	0.00	-0.03	0.00	0.16	1.00			
AMT_ANNUITY	0.00	-0.03	0.02	0.97	0.77	1.00		
AMT_GOODS_PRICE	0.00	-0.04	0.00	0.16	0.99	0.78	1.00	
REGION_POPULATION_RELATIVE	0.00	-0.04	-0.03	0.07	0.10	0.12	0.10	
DAYS_BIRTH	0.00	0.08	0.33	0.03	-0.06	0.01	-0.05	
DAYS_EMPLOYED	0.00	-0.04	-0.24	-0.06	-0.07	-0.10	-0.06	
DAYS_REGISTRATION	0.00	0.04	0.18	0.03	0.01	0.04	0.01	
DAYS_ID_PUBLISH	0.00	0.05	-0.03	0.01	-0.01	0.01	-0.01	
OWN_CAR_AGE	0.00	0.00	0.07	0.02	0.03	0.04	0.03	
FLAG_MOBIL	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
FLAG_EMP_PHONE	0.00	0.05	0.24	0.06	0.07	0.10	0.06	
FLAG_WORK_PHONE	0.00	0.03	0.06	-0.02	-0.02	-0.02	0.00	
FLAG_CONT_MOBILE	0.00	0.00	0.00	-0.01	0.02	0.02	0.02	
FLAG_PHONE	0.00	-0.02	-0.03	0.00	0.03	0.01	0.04	
FLAG_EMAIL	0.00	0.00	0.02	0.04	0.02	0.07	0.02	
CNT_FAM_MEMBERS	0.00	0.01	0.88	0.02	0.06	0.08	0.06	

Chart Area

Top 10 Correlation

Column1	Column2	correlation coefficient
AMT_ANNUITY	AMT_INCOME_TOTAL	0.97
AMT_GOODS_PRICE	AMT_CREDIT	0.99
AMT_GOODS_PRICE	AMT_ANNUITY	0.78
CNT_FAM_MEMBERS	CNT_CHILDREN	0.88
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.95
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.86
LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.83
YEARS_BEGINEXPLUATATION_MODE	YEARS_BEGINEXPLUATATION_AVG	0.97
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.86

Takeaway from the Project:-

- Types of missing values and how to treat each type .
- How to replace missing values with mean /median/ mode depending upon type of column and distribution of data.
- How to approach stepwise like in univariate analysis first categorical column then numerical columns since each

Type has a different approach.

- How scatter plots can be used to find the outliers .
- How coefficient matrix helps us to identify which column has a effect on other .