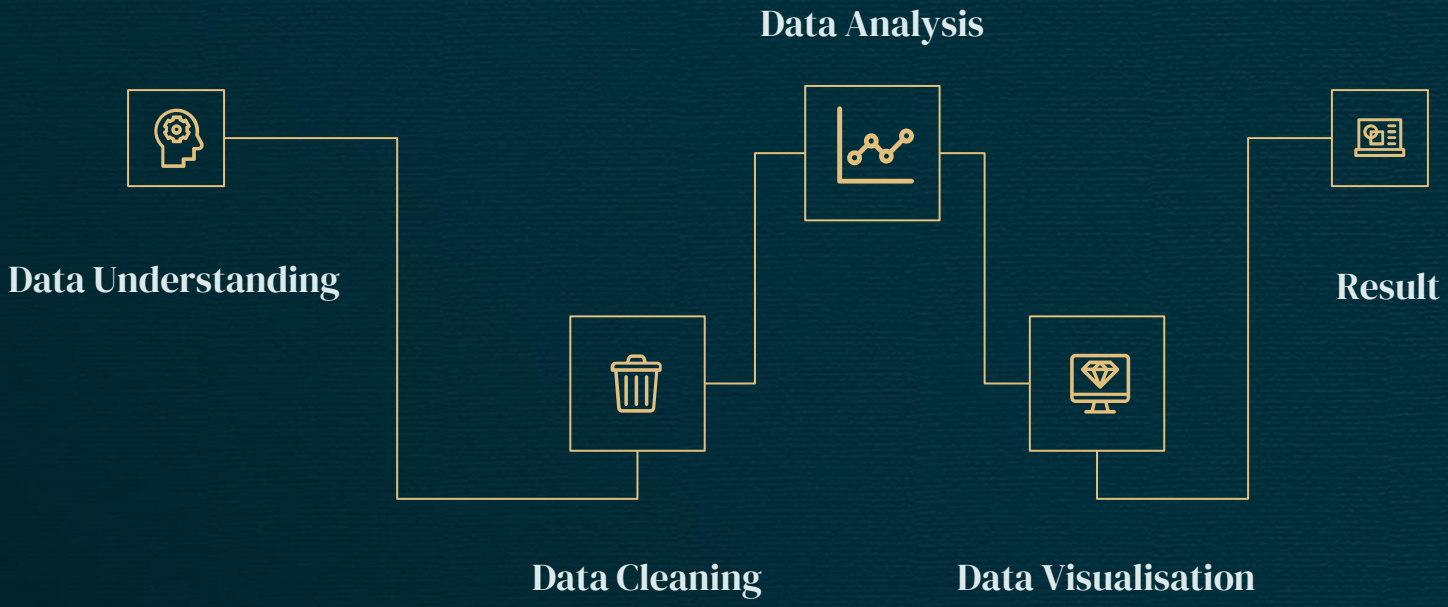
Bank Loan Case Study

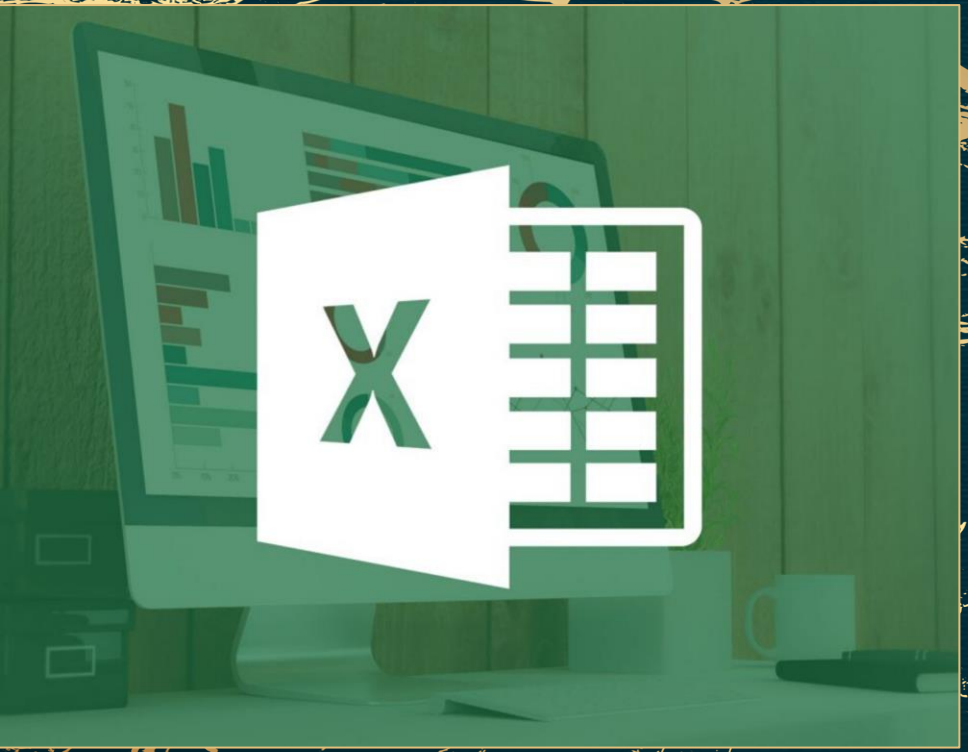
By Ayush Mahanta

PROJECT DESCRIPTION

In this project, I will figure out signs that show if a client might struggle to pay back their installments. I will use this information to decide whether to deny a loan, give less money, or offer a loan with higher interest rates to risky applicants. This helps ensure that people who can pay back the loan don't get turned down. I will study the data using Exploratory Data Analysis (EDA), which helps summarize and understand the characteristics of the dataset by looking for patterns and relationships to help with decision-making and analysis.

APPROACH


Data Analysis
Data Understanding
Result
Data Cleaning
Data Visualisation



Tech Stack

All the analysis has been performed in excel. This tool is also used to create graphical representation of the results and to understand the result set better.

Excel Link

:https://docs.google.com/spreadsheets/d/1HmMMXhynS3GySzA2xJIFXFksXyN6_SiU/edit?usp=drive_link&ouid=108396890359637253084&rtpof=true&sd=true

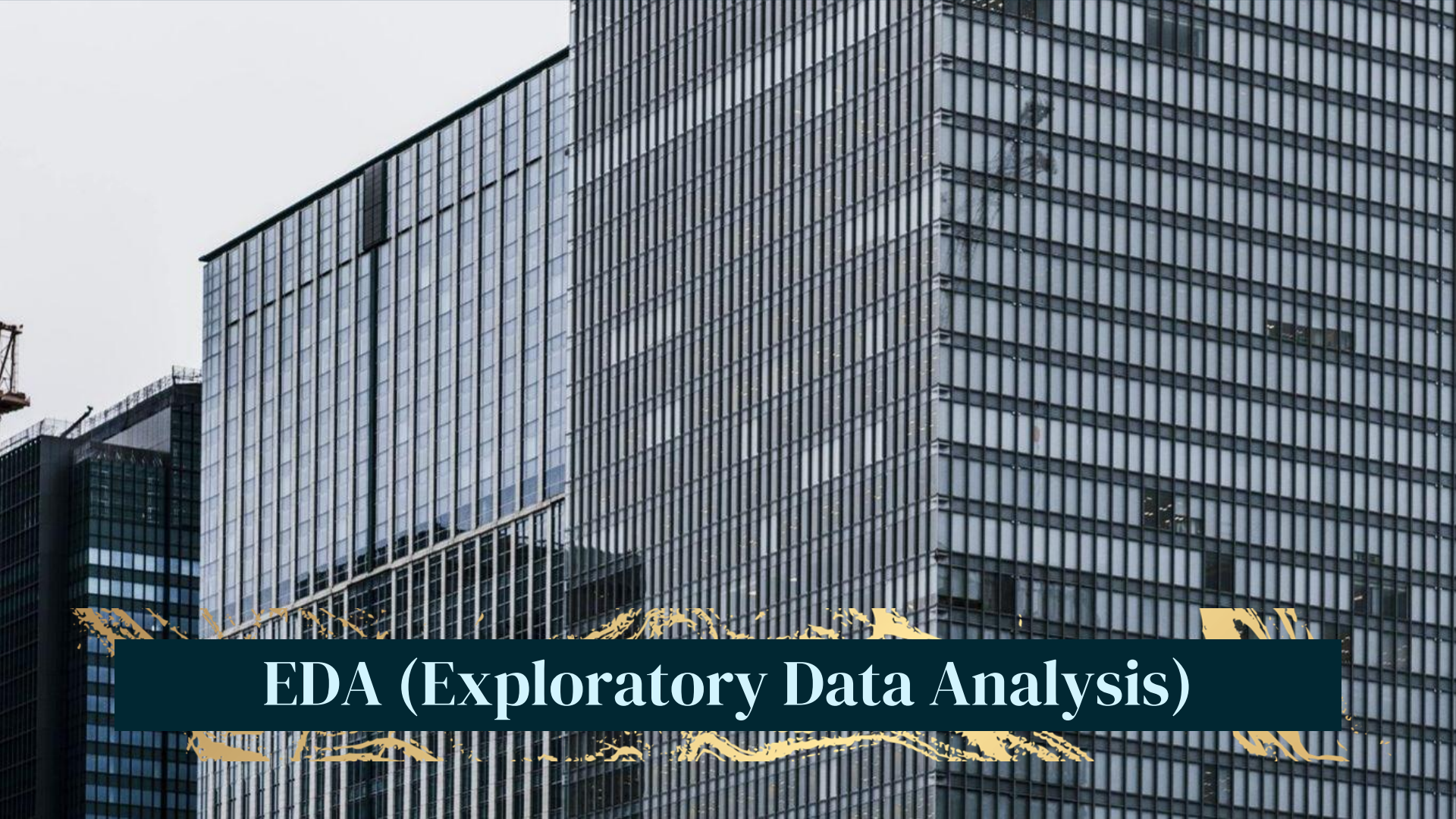
01

Data Cleaning

Columns to Drop

OWN_CAR_AGE	OCCUPATION_TYPE	EXT_SOURCE_1	APARTMENTS_AVG	BASEMENTAREA_AVG
YEARS_BEGINEXPLUATATION_AVG	YEARS_BUILD_AVG	COMMONAREA_AVG	ELEVATORS_AVG	ENTRANCES_AVG
FLOORSMAX_AVG	FLOORSMIN_AVG	LANDAREA_AVG	LIVINGAPARTMENTS_AVG	LIVINGAREA_AVG
NONLIVINGAPARTMENTS_AVG	NONLIVINGAREA_AVG	APARTMENTS_MODE	BASEMENTAREA_MODE	YEARS_BEGINEXPLUATATION_MODE
YEARS_BUILD_MODE	COMMONAREA_MODE	ELEVATORS_MODE	ENTRANCES_MODE	FLOORSMAX_MODE
FLOORSMIN_MODE	LANDAREA_MODE	LIVINGAPARTMENTS_MODE	LIVINGAREA_MODE	NONLIVINGAPARTMENTS_MODE
NONLIVINGAREA_MODE	APARTMENTS_MEDI	BASEMENTAREA_MEDI	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BUILD_MEDI
COMMONAREA_MEDI	ELEVATORS_MEDI	ENTRANCES_MEDI	FLOORSMAX_MEDI	FLOORSMIN_MEDI
LANDAREA_MEDI	LIVINGAPARTMENTS_MEDI	LIVINGAREA_MEDI	NONLIVINGAPARTMENTS_MEDI	NONLIVINGAREA_MEDI
FONDKAPREMONT_MODE	HOUSETYPE_MODE	TOTALAREA_MODE	WALLSMATERIAL_MODE	EMERGENCYSTATE_MODE

We see there are 50 columns with missing values greater than 30% we will drop those columns. These are the columns which mainly contain the residential details of the client



EDA (Exploratory Data Analysis)

Representation of columns having >30% NULL values.

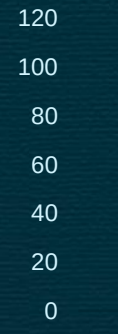
Comparison

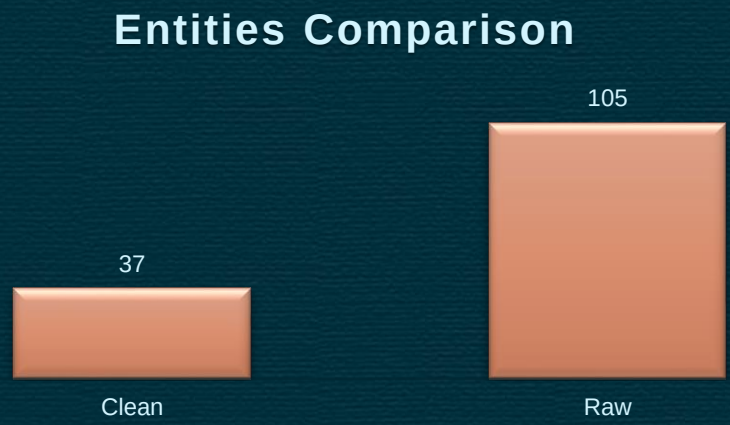
Updated Data

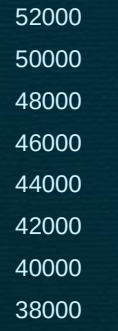
Clean	Clean Records
37	42857

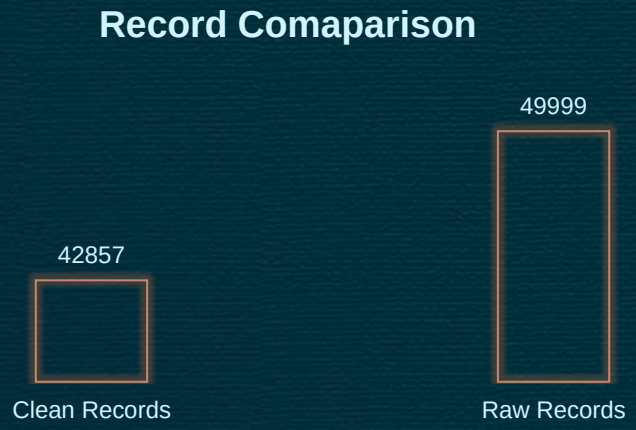
Raw Data

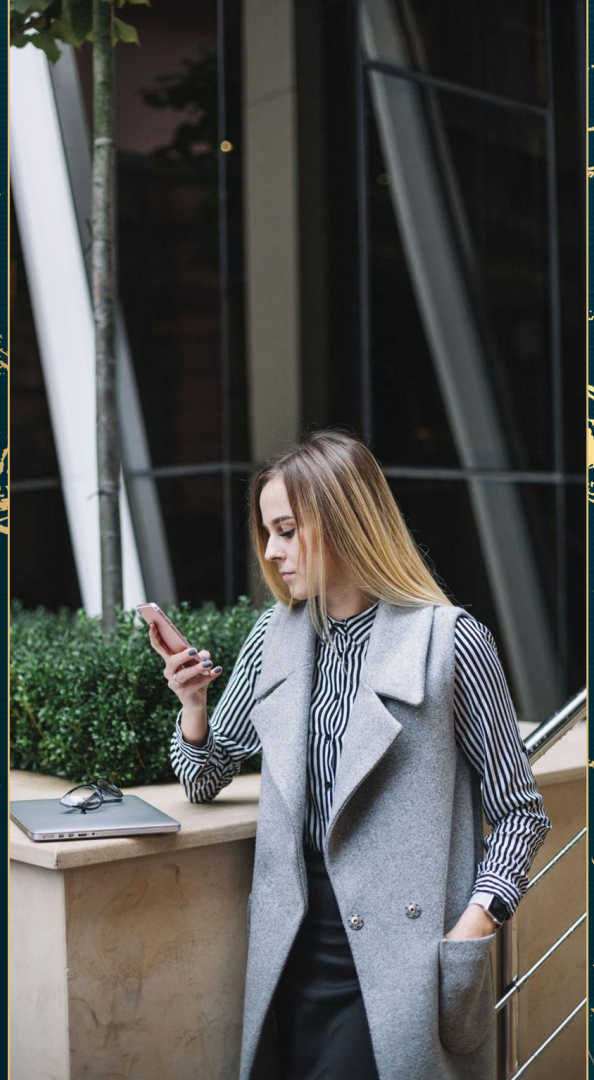
Raw	Raw Records
105	49999





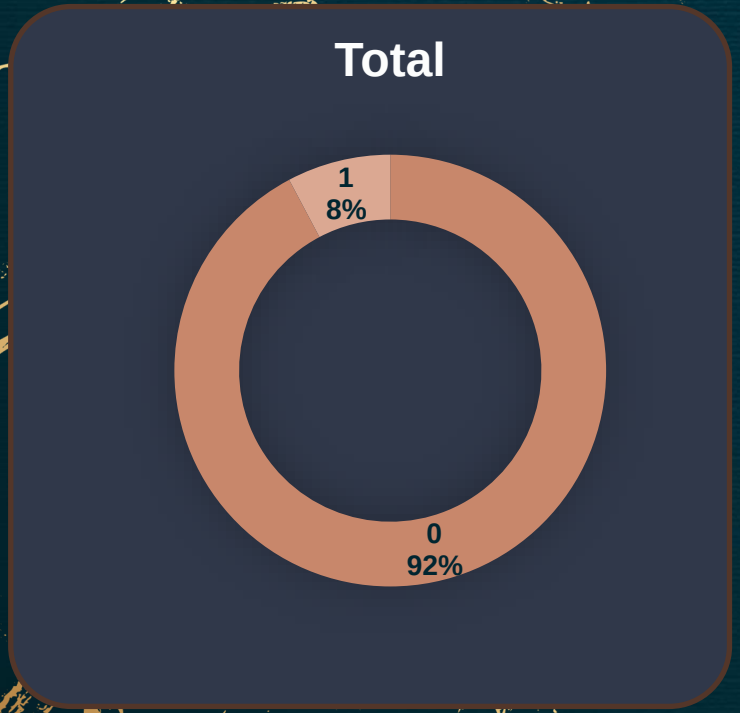






02

Data Imbalance



Data Imbalance

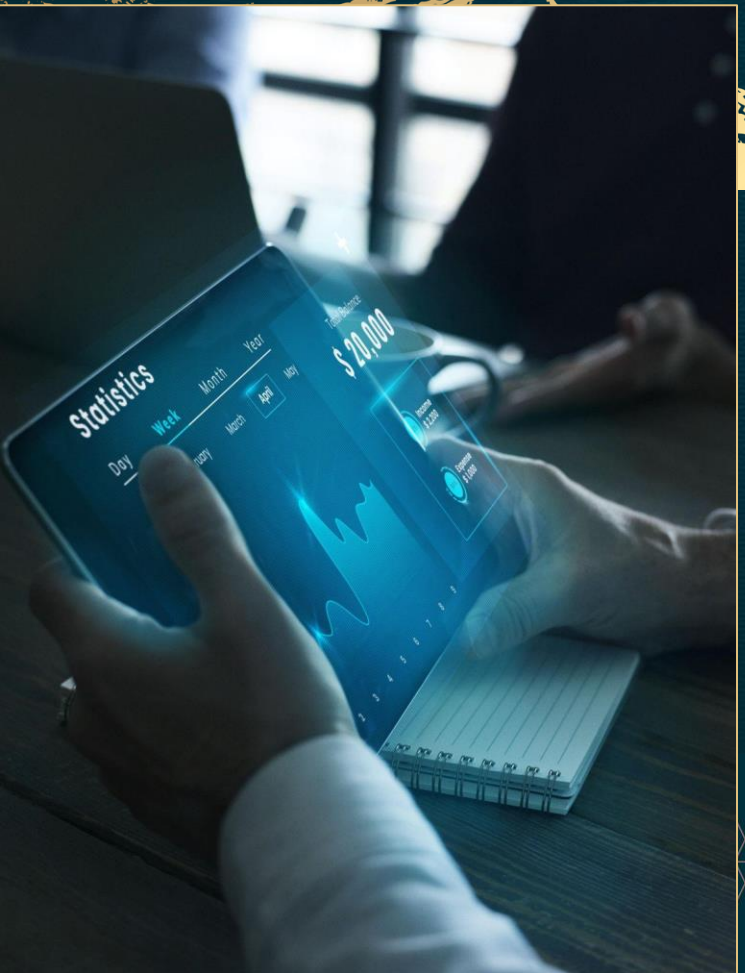
We see that we have :

- 92% as loan re-payers
- 8% as De faulters

Which gives us a clear indication that the data is highly imbalanced

03

Outliers



IDENTIFYING OUTLIERS

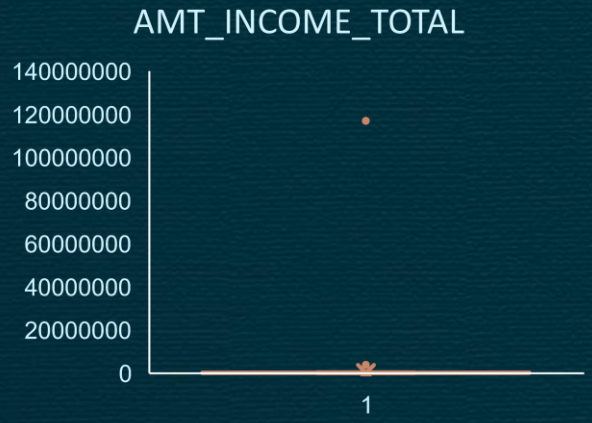
Steps to find Outliers using Tukey's method:

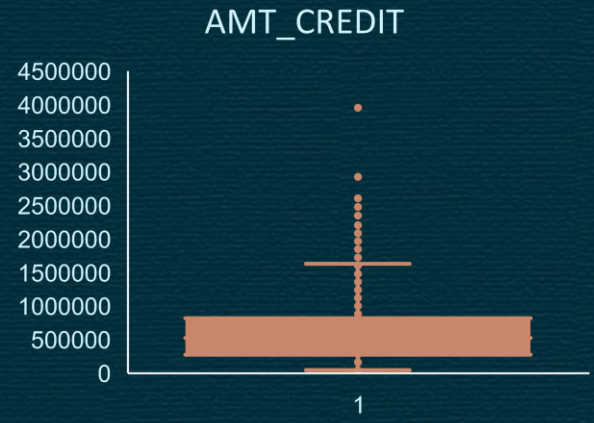
1. Finding 1st Quartile Q1 and 3rd Quartile Q3
2. Finding Inner Quarter Range(IQR)
3. Finding Upper Bound(Q3+(1.5*IQR)
4. Finding Lower Bound(Q1-(1.5*IQR)
5. Now, any data point above Upper Bound or Below Lower Bound Considered as Outlier

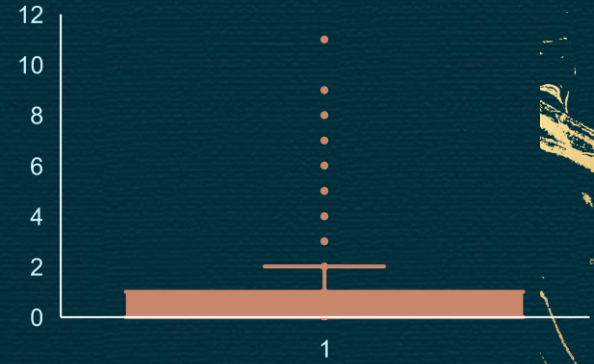
Below , I plotted some Box-Whisker Plots to find there are any outliers present in the data.

Outliers

- The first quartile almost missing for CNT_CHILDREN that means most of the data are present in the first quartile.
- There is single high value data point as outlier present in AMT_INCOME_TOTAL and DAYS_EMPLOYED. Removal this point will drastically impact the box plot for further analysis.
- The first quartiles is slim compare to third quartile for AMT_CREDIT, AMT_ANNUITY, DAYS_REGISTRATION. This mean data are skewed towards first quartile.







Quartile

AMT INCOME TOTAL

Quartile - 1
112500
Quartile - 3
202500
Inter Quartile Range
90000
UPPER LIMIT
337500
Lower Limit
-22500

CNT CHILDREN

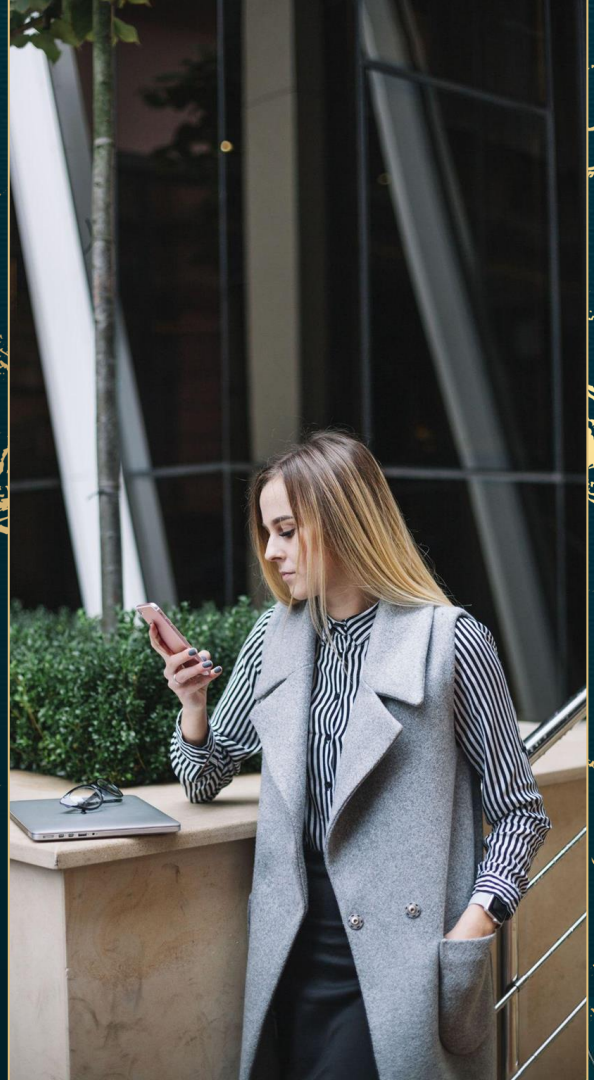
Quartile - 1
0
Quartile - 3
1
Inter Quartile Range
1
UPPER LIMIT
2.5
Lower Limit
-1.5

Days Employed (YRS)

Quartile - 1
2.652054795
Quartile - 3
15.89863014
Inter Quartile Range
13.24657534
UPPER LIMIT
35.76849315
Lower Limit
-17.21780822

AMT CREDIT

Quartile - 1
273636
Quartile - 3
816660
Inter Quartile Range
543024
UPPER LIMIT
1631196
Lower Limit
-540900



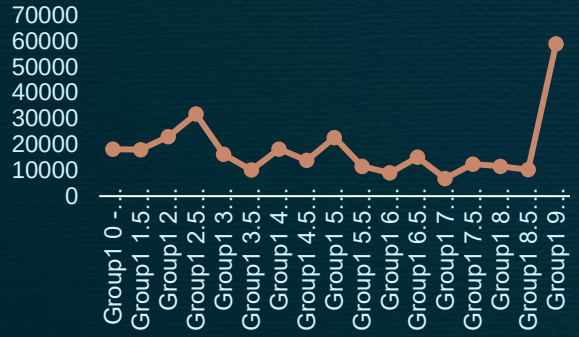
04

Univariate

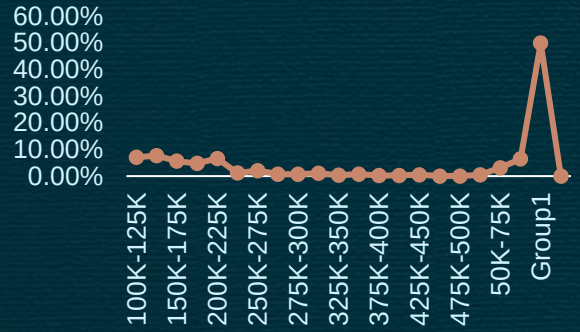
UNIVARIATE ANALYSIS

- Univariate analysis is statistical method to analyze the data with one variable.
- It involves the examining the distribution of single variable and deriving insights from it.
- Univariate analysis of categorical variables involves summarizing and examining the frequency or proportion of each category to gain better understanding of distribution and relationship between variables

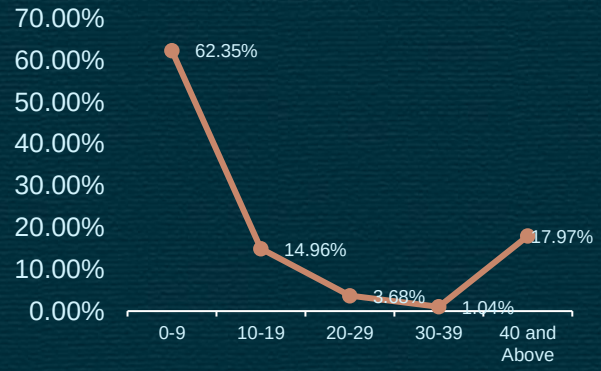
Segmented Applicants Per Credit Bins



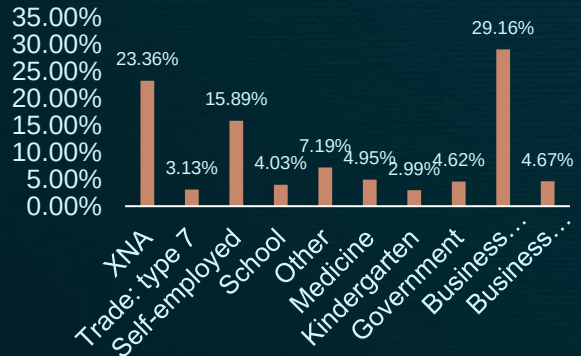
Segmented Target Applicants Per Income Bins



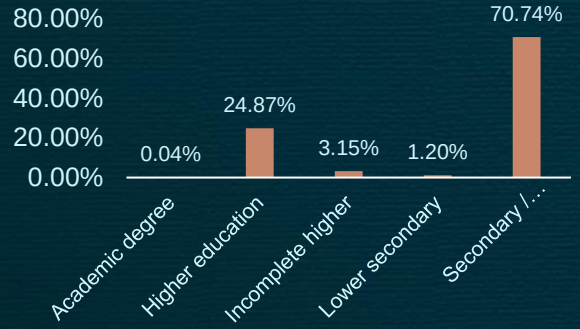
Segmented employment years



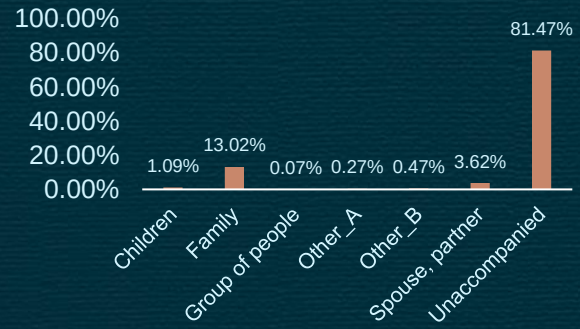
Segmented organization type



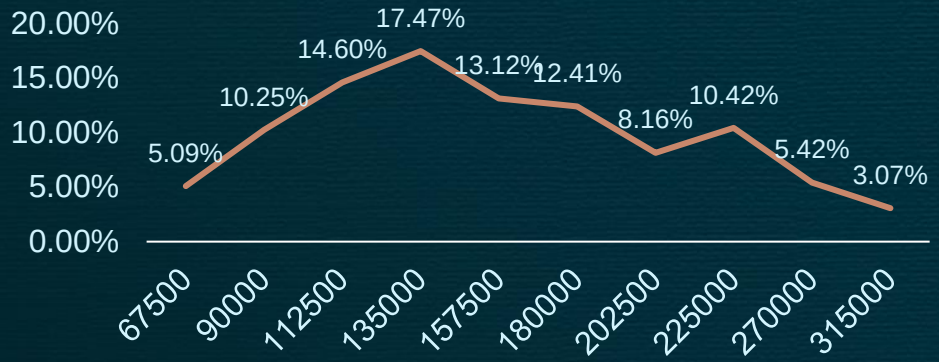
Segmented Education Type distribution



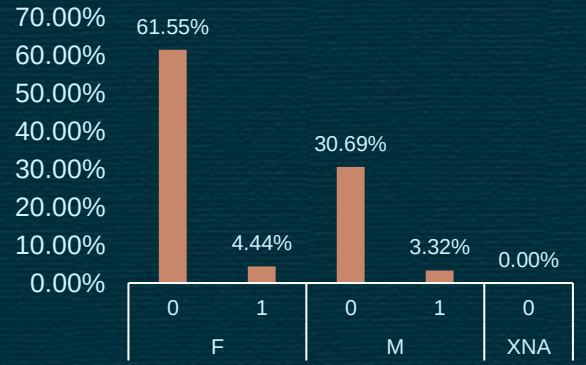
Segmented NAME_TYPE_SUITE



Amount Income



Applicants Per Credit Bins



Amount Credit

05

Bivariate Analysis

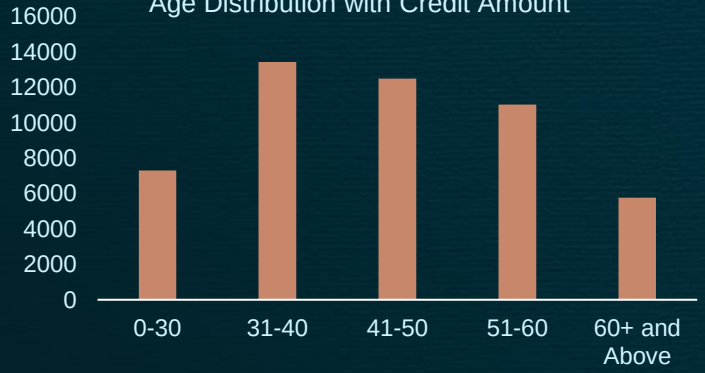
Bivariate Analysis

Bivariate analysis is a statistical method used to analyze data involving two variables. It's about looking at how two variables relate to each other. When we do bivariate analysis with categorical variables, we are interested in understanding the relationship between two categories or groups. This could mean comparing the frequency or proportion of each category across the two variables to see if there's any connection or pattern between them.

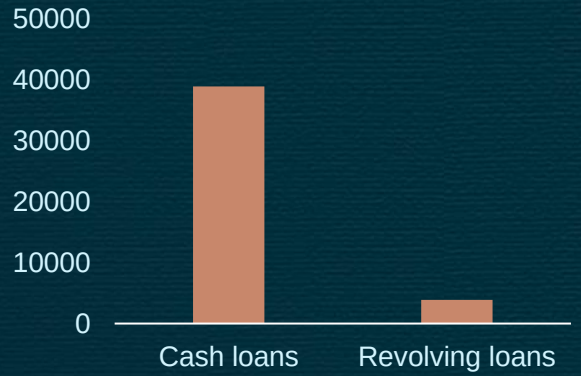
Average Credit Amount Per Credits Bins

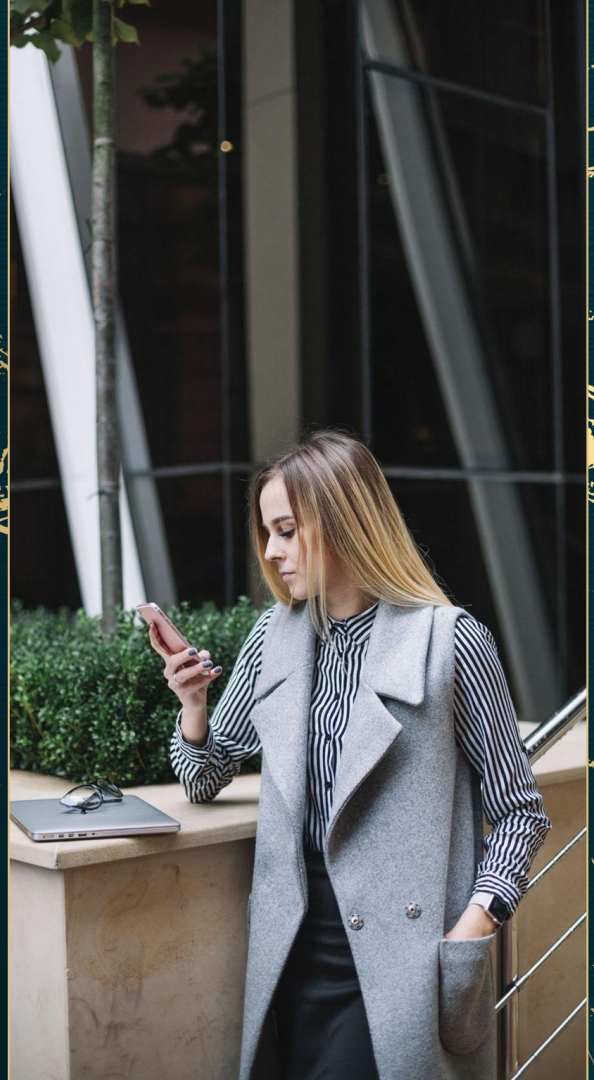
Average Amount Credited Per Name Income Type Segmented

Age Distribution with Credit Amount



NAME_CONTRACT_TYPE





06

Correlation

Correlation For Timely Payments

TARGET 0								
CNT Of Children	1	0.009300239	0.004117726	-0.025043214	-9.7283E-07	0.002368016	0.003949962	0.026733198
AMT_Income_Total	0.009300239	1	0.063511987	0.026543706	-0.000502928	-0.003429977	0.003573112	-0.035127419
AMT_Credit	0.004117726	0.063511987	1	0.098803701	-0.004437401	-0.000222036	-0.004349401	-0.102897278
Region Population Relative	-0.025043214	0.026543706	0.098803701	1	-0.003303866	-0.006023773	0.000844092	-0.527253308
Days_Birth(Yrs)	-9.7283E-07	-0.000502928	-0.004437401	-0.003303866	1	0.523388736	0.50850239	0.003170646
Days_Employed(YRS)	0.002368016	-0.003429977	-0.000222036	-0.006023773	0.523388736	1	0.296604311	0.001543287
Days_ID_Publish(YRS)	0.003949962	0.003573112	-0.004349401	0.000844092	0.50850239	0.296604311	1	0.003088812
Region_Rating_Client	0.026733198	-0.035127419	-0.102897278	-0.527253308	0.003170646	0.001543287	0.003088812	1
	CNT Of Chindren	AMT_Income_Total	AMT_Credit	Region_Population_Relative	Days_Birth(YRS)	Days_Employed(YRS)	Days_ID_Publish(YRS)	Region_Rating_Client

TARGET 1								
CNT Of Children	1	0.009300239	0.004117726	-0.025043214	-9.7283E-07	0.002368016	0.003949962	0.026733198
AMT_Income_Total	0.009300239	1	0.063511987	0.026543706	-0.000502928	-0.003429977	0.003573112	-0.035127419
AMT_Credit	0.004117726	0.063511987	1	0.098803701	-0.004437401	-0.000222036	-0.004349401	-0.102897278
Region Population Relative	-0.025043214	0.026543706	0.098803701	1	-0.003303866	-0.006023773	0.000844092	-0.527253308
Days_Birth(Yrs)	-9.7283E-07	-0.000502928	-0.004437401	-0.003303866	1	0.523388736	0.50850239	0.003170646
Days_Employed(YRS)	0.002368016	-0.003429977	-0.000222036	-0.006023773	0.523388736	1	0.296604311	0.001543287
Days_ID_Publish(YRS)	0.003949962	0.003573112	-0.004349401	0.000844092	0.50850239	0.296604311	1	0.003088812
Region_Rating_Client	0.026733198	-0.035127419	-0.102897278	-0.527253308	0.003170646	0.001543287	0.003088812	1
	CNT_Of_Children	AMT_Income_Total	AMT_Credit	Region_Population_Relative	Days_Birth(YRS)	Days_Employed(YRS)	Days_ID_Publish(YRS)	Region_Rating_Client

Result

- As people get older and gain more experience, they're less likely to miss loan payments. So, banks should give more attention to older and experienced clients.
- Clients with higher education tend to miss fewer payments compared to those with lower education levels like high school or lower secondary.
- Men tend to miss loan payments more often than women.
- Corporate clients are safer bets compared to labor class clients.
- People from Region Rating 3 have the highest percentage of defaulters, so banks could make stricter loan policies for clients from this region. Clients from Region 1 are the safest bet.
- As clients get older, they tend to take larger loan amounts, and since older clients have lower default rates, they are less risky and more profitable for the bank.
- Banks should focus more on clients with contract types like 'Student,' 'Pensioner,' and 'Businessman' who have housing types other than 'Co-op apartment' for successful payments.
- Banks should be cautious with clients whose income type is 'Working' as they have the highest number of missed payments.
- For loans for the purpose of 'Repairs,' although there are more rejections, there are also difficulties in paying on time.
- There are some areas where loan payments are significantly delayed. Banks should be cautious when giving loans for these purposes.
- Banks should avoid giving loans for co-op apartments as they have difficulties in payment.
- Banks can focus more on housing types like 'with parents,' 'House/apartment,' and 'municipal apartment' for successful payments.

Conclusion

This project helps in handling the large datasets. How exploratory data analysis can be applied to large datasets. When dealing with the large datasets it is also important to select only those columns which are extremely useful to our analysis. Finding correlations columns can become very convenient while dealing with large datasets as it saves time selecting which columns should be considered for analysis. The project also helps in understanding the various terminologies used in the banking domain.

Thanks!

Do you have any questions?

ayush17mahanta@gmail.com
https://www.linkedin.com/in/ayush17mahanta
https://www.instagram.com/ayush17mahanta