

# Loan Default Prediction: An EDA Approach

This presentation outlines an Exploratory Data Analysis (EDA) approach to predict loan defaults. The goal is to identify patterns that indicate a client's difficulty in repaying loans, enabling informed decisions like loan denial or adjusted interest rates. This ensures that creditworthy applicants are not unjustly rejected, optimizing the lending process.



by **Ramesh Sunkara**



# Business Understanding and Objectives

## Problem Statement

Loan companies struggle to assess credit risk due to insufficient credit history, leading to potential defaults and financial losses.

## Business Objective

Identify key indicators of loan default to improve risk assessment and portfolio management, ensuring deserving applicants are approved.

## Goal

Minimize financial losses by accurately predicting loan defaults and optimizing lending strategies.



# Data Overview and Understanding

## Application Data

Contains client information at the time of loan application, indicating payment difficulties.

## Previous Application Data

Includes data on previous loan applications, detailing outcomes like approval, cancellation, refusal, or unused offers.

## Columns Description

Provides a data dictionary explaining the meaning of each variable in the datasets.



# Handling Missing Data

Missing data is identified and addressed using appropriate methods. Columns with excessive missing values may be removed, while others are imputed with suitable values. The approach is clearly documented to maintain transparency and ensure data integrity.

Note that in EDA, since it is not necessary to replace the missing value, but if you have to replace the missing value, what should be the approach. Clearly mention the approach.

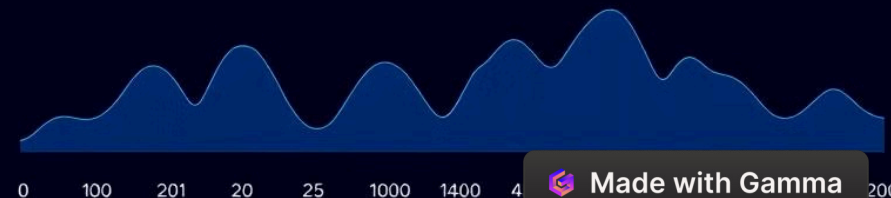
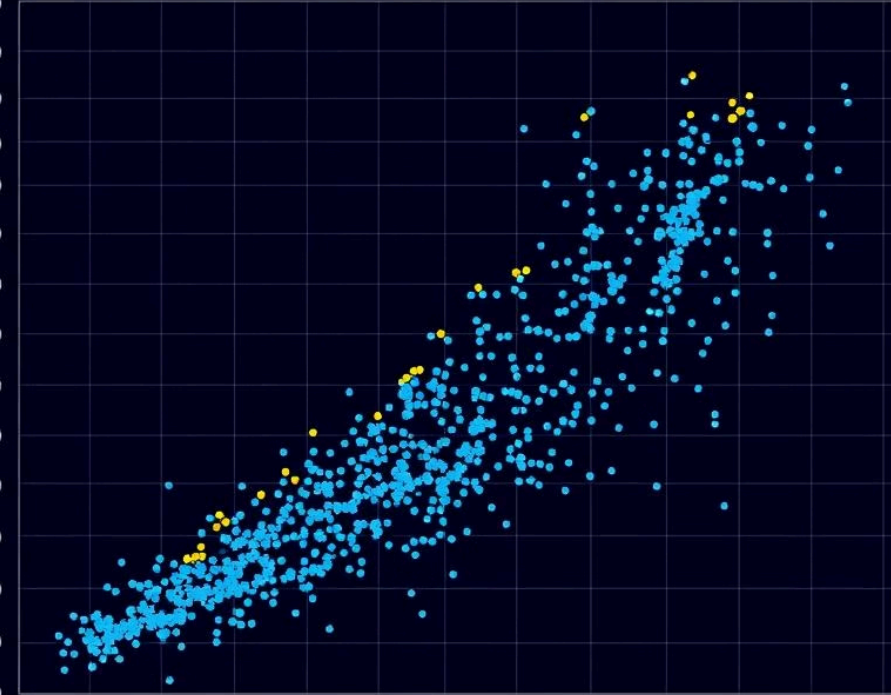


3.12571	264.0945	3995.45.953	2476.0090	-2172	:3395,406,3
5.19672	367,1756	8055.59.591	5246.3905	:2017	:5495,665,3
3.15577	568.0886	3854.35.456	2227,0993	-2011	:2355,399,3
5.15259	564.0938	5809.57.501	5324,39.33	:3778	:5961,566,1
5.12751	548.5532	3605.13.558	5333,13111	-3717	:5225,397,1
5.15355	967.5388	3364.03.571	5463.19.71	-1960	:2364,997,7
3.7339)	307,3985	2884.55.811	3443,00.68	-1593	:3415,299,7
4.15573	766.2007	3995.15.541	2334.0648	-2261	:4457,236,2
5.23471	566.3168	2351.19.507	2335.39.75	-2571	:4457,301,1
7.15591	323.3910	5457.35.681	2459.6025	-1519	:5345,555,2
5.715					
5.17151	706,2337	8625.20.11	5455,19.21	-1004	:2455,238,2
7.15431	536,6037	2055,31333,1331934,670		-4334	:4435,394,5
3.17341	716,9643	2554,34615,1743666,581		9635	:3347,399,2
5.(667)	164,3998	2455.19321,1047755,700		745	:4855,536,3
1.1737)	398,9993	2455.59113,17588723579		174	:4035,191,7
9.15979	185,3049	23251,1045,1301025,387		81	:3915,596,3
1.15531	174.8888			887	:4415,499,1
9.11531	446,2947			2548	:2319,138,2
9.15491	404,4245	995.36375,17473255,59		-3767	:3447,543,2
5.31951	335,0058	555,339,404		2277	:4435,565,1
4.15433	344,9790	2465,39,557	5357,197	9	:4935,395,3
9.15551	246,7344	2952,38,601	5234,0637		:3937,555,9
5.13371	514,9910	2499,17,593	5373,11,99	-	:3515,545,3
3.5794)	515,9542	3979,33,671	5075,1099	-39	:45,557,1
4.13577	455,0098	5475,45,563	3465,19111	-1373	:95,3
9.19831	455,5578	5454,45,551	3339,11011	-3355	
1.18875	455,3396	5575,35,512	5385,11661	-1761	:35
5.13751	242,0778	2985,55,584	5333,0595	-3355	:2975,3
5.17					
5.19881	475,7990	5057,35,464	5478,11011	-1787	:5193,596,2
1.17737	287,1277	2447,13,975	24		:6,1
3.18387	255,5735	3775,38,367	:3135,0337	-1333	:5335,306,2

# Outlier Identification

Outliers are identified within the dataset, with justifications for their classification. While removal isn't necessary for this exercise, understanding their impact is crucial. Outliers can skew analysis and affect model performance, so their presence is carefully considered.

OUTLIER IDENTIFICATION TO DATA

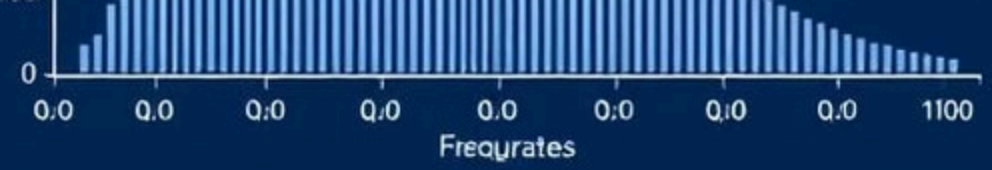
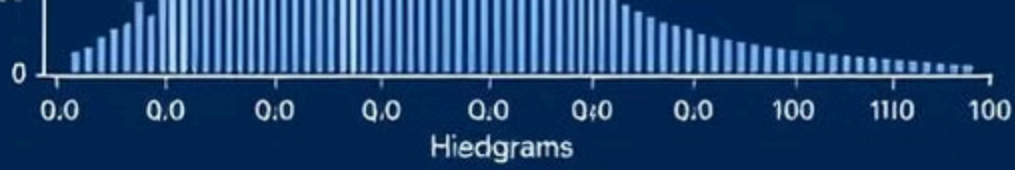




# Data Imbalance Analysis

Data imbalance is assessed by examining the ratio of clients with payment difficulties versus all other cases. Visualizations, such as plots with varying scales (percentage or absolute value), are used to analyze different aspects of this imbalance. This analysis focuses on the 'Target variable' to understand its distribution.





**Histograms**



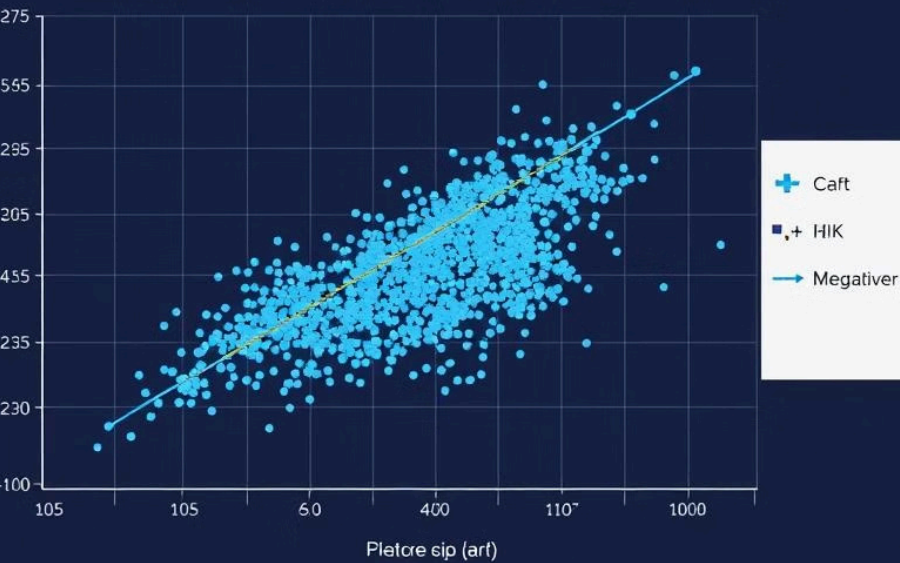
**Outliers**



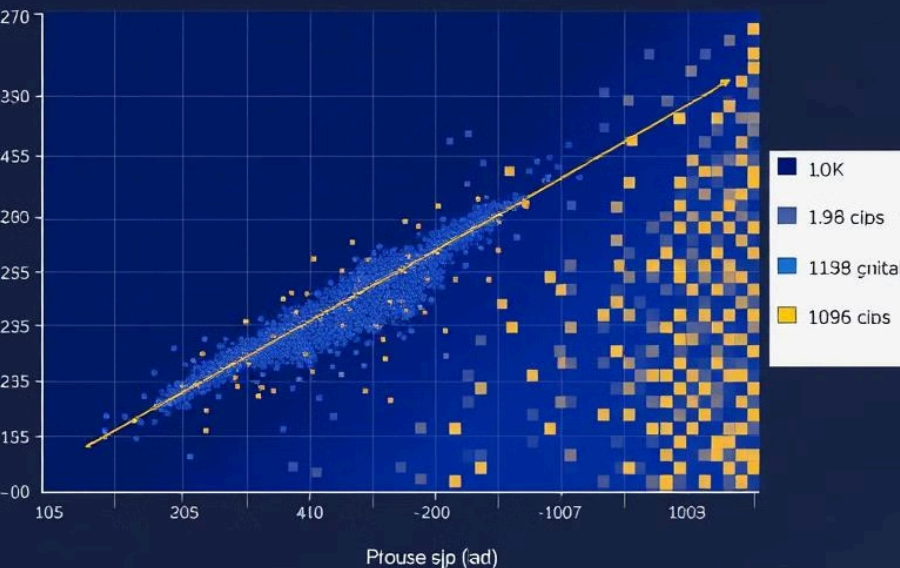
# Univariate Analysis Results

Univariate analysis results are explained in business terms, providing insights into individual variables. This includes understanding the distribution, central tendency, and spread of each variable. These insights help in identifying key characteristics of the client population and potential risk factors.

Meputtave, Corenats



Megut FayE Cirenats



# Bivariate Analysis Results

Bivariate analysis results are interpreted in business terms, highlighting relationships between variables. This includes segmented univariate analysis to understand how different groups behave. These analyses reveal potential drivers of loan default and inform targeted interventions.



# Top Correlations with Target Variable

The top 10 correlations for clients with payment difficulties and all other cases are identified. This involves segmenting the data frame with respect to the target variable and finding the top correlations for each segment. Insights are derived by comparing these correlations to understand key differences.



# Key Takeaways and Next Steps



## Insights

Identified key variables driving loan default, enabling targeted risk mitigation strategies.



## Actions

Implement stricter lending criteria, adjust interest rates, and refine portfolio management.



## Future Work

Develop predictive models to automate loan approval processes and minimize financial losses.