Bank Loan Case Study Final Project-2

Project

- The project involves applying EDA techniques to a dataset from a consumer finance company that specializes in lending various types of loans to urban customers.
- The objective is to identify patterns and relationships in the data that can be used to minimize the risk of financial loss for the lending company by identifying loan applicants who are likely to have difficulty paying back their loans.
- The analysis will involve data cleaning, data visualization, and statistical analysis to gain insights into the factors that may be driving loan defaults.

Description

- This project is about using Exploratory Data Analysis (EDA) to identify patterns in loan application data to minimize the risk of losing money while lending to customers. The aim is to understand the driving factors behind loan default and identify strong indicators of default to enable the company to take appropriate actions.

- Through the project, we will handle the loan application data using EDA techniques to identify patterns that indicate if a client has difficulty paying their installments. This will help in taking actions such as denying the loan, reducing the amount of loan, or lending to risky applicants at a higher interest rate. The ultimate goal is to ensure that the consumers capable of repaying the loan are not rejected.

Mention the problem statement and the analysis approach briefly

- Problem statement: Identify patterns indicating loan applicants likely to have difficulty paying back, minimizing risk of financial loss for lending company.
- Analysis approach: EDA techniques including data cleaning, data visualization, statistical analysis, and potentially using machine learning algorithms for predictive modeling.
- Goal: Develop better understanding of driving factors behind loan default and identify strategies to minimize risk of financial loss for lending company.

Identify the missing data and use appropriate method to deal with it.

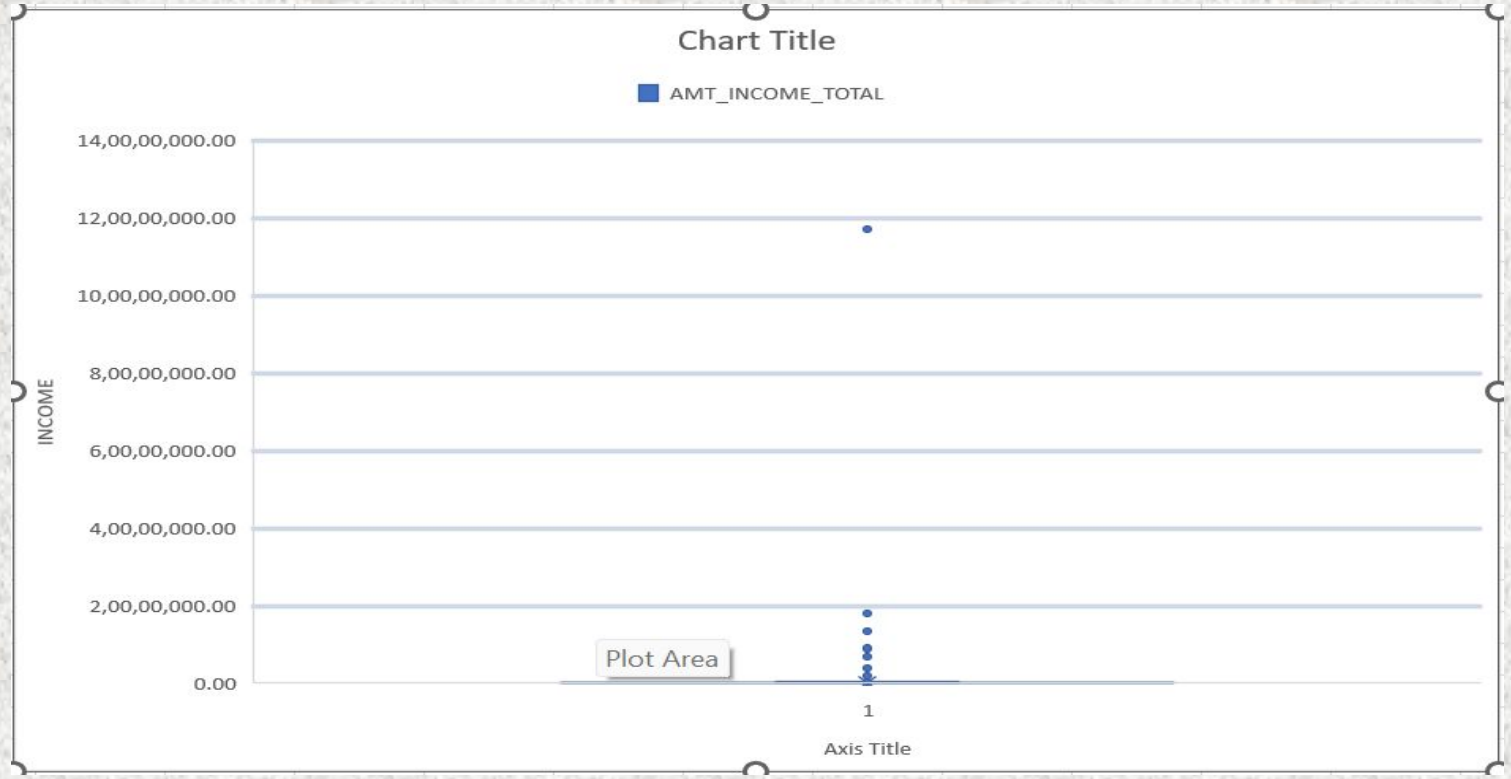
- In the given datasets there are two types of data
- Numerical columns in a dataset are typically called quantitative or continuous variables.
- Word columns in a dataset are typically called categorical or nominal variables.
- The blanks in categorical variables will be filled with NP.
- The blanks in continuous variables will be filled with 0 because each row is belongs too different records.

Identify if there are **outliers** in the dataset

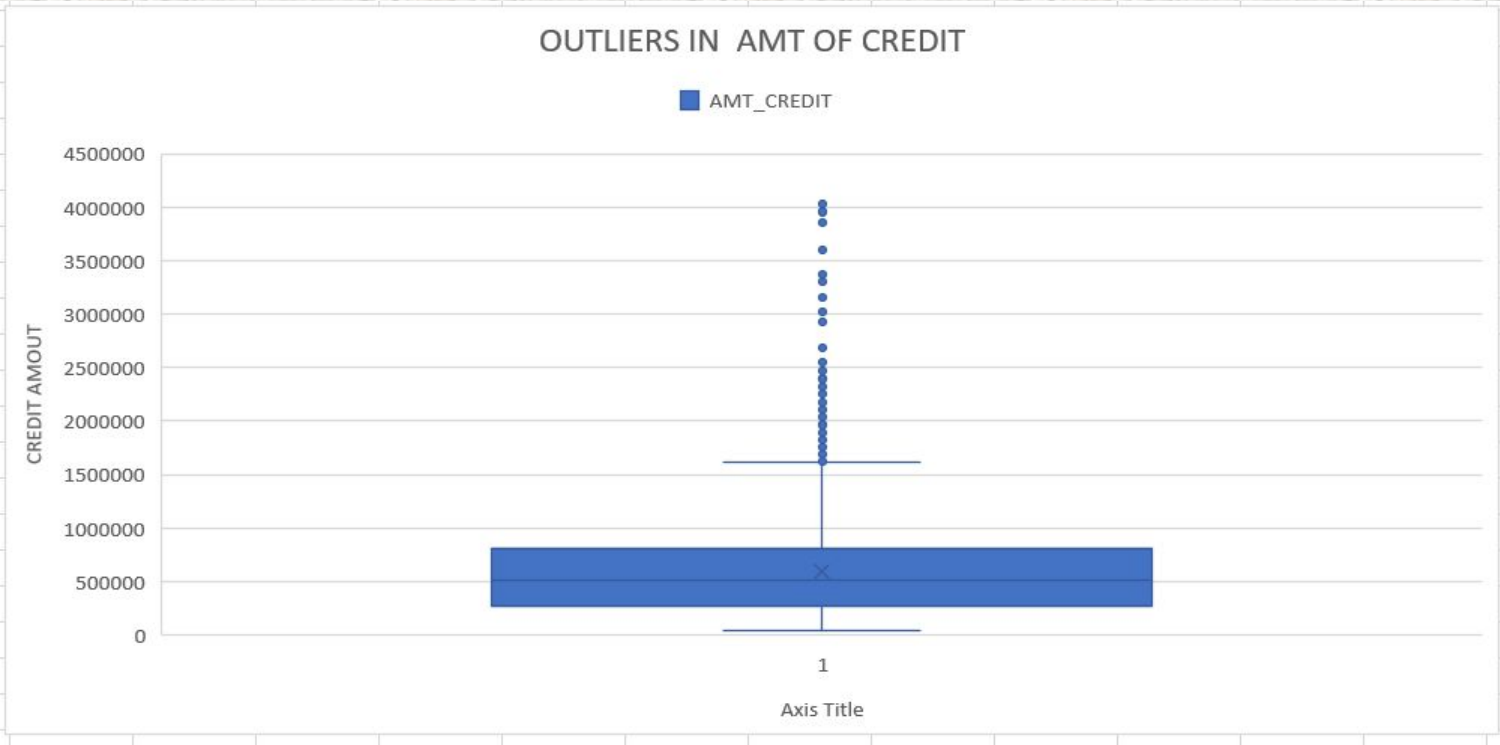
- an outlier is a value or point that differs significantly from the rest of the data in a dataset.
- Here you can see the outliers in amount of income and credit

MIN OF INCOME	25650
Q1 OF INCOME	112500
MEDIAN	147150
Q3 OF INCOME	202500
MAX OF INCOME	117000000
MEAN	168797.92
IQR	90000
LOWER LIMIT	-22500
UPPER LIMIT	337500

- The data points which are out of lower limit and upper limit are outliers



MIN OF CREDIT	45000
Q1 OF CREDIT	270000
MEDIAN	513531
Q3 OF CREDIT	808650
MAX OF CREDIT	4050000
MEAN	599026
IQR	538650
LOWER LIMIT	-537975
UPPER LIMIT	1616625



Identify if there is data imbalance in the data. Find the ratio of data imbalance.

Owning a house implies having the lones .

The ratio of having imbalance according to owning an house is 2:1

count of yes	213312
count of no	94199
total count	307511
yes/total count	0.693672747
no/total count	0.306327253
ratio	02:01

Explain the **results of univariate, segmented univariate, bivariate analysis, etc.** in business terms

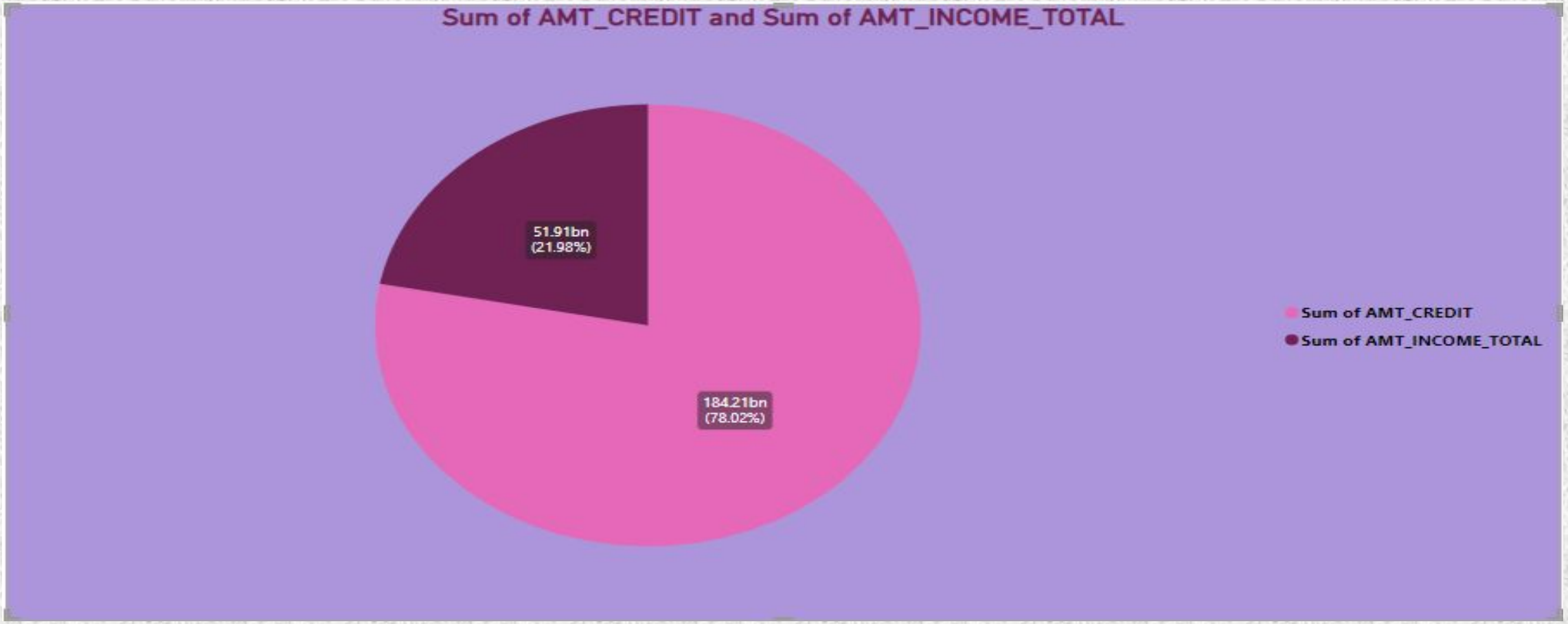
1. Univariate analysis is a data analysis technique that focuses on a single variable and aims to understand its distribution and patterns within a dataset.
2. Segmented univariate analysis involves analyzing a dataset by dividing it into segments based on some factor, to explore variations within different segments of the data.
3. Bivariate analysis involves analyzing two variables in relation to each other to understand any patterns or relationships between them.

1. These types of analyses can help businesses identify patterns and trends in their data.
2. The insights gained from data analysis can inform business decisions and lead to more effective outcomes.

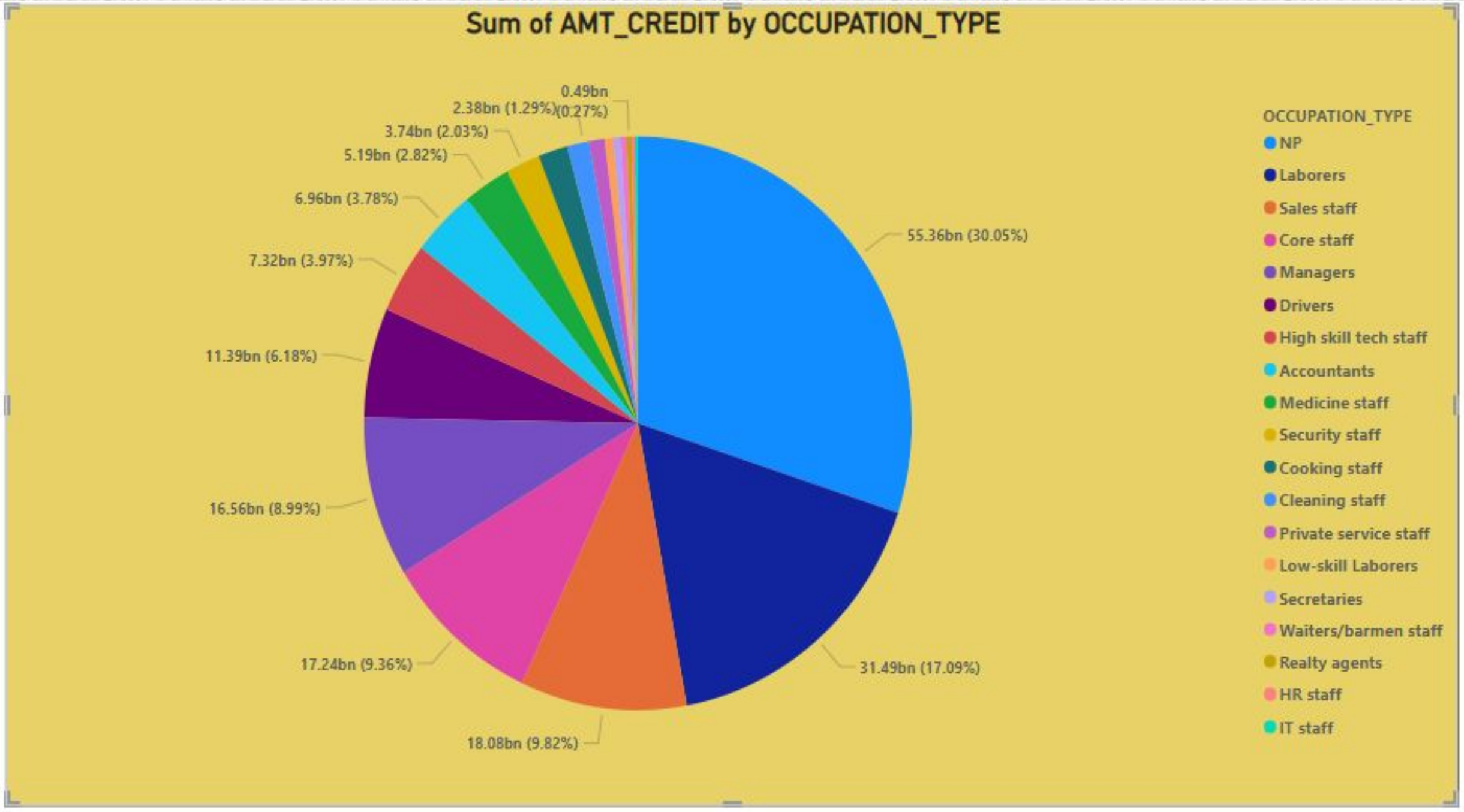
Find the top 10 **correlation** for the Client with payment difficulties and all other cases

- **correlation** of amount of income and credit 0.156870272
- **correlation** of amount of annuity and goods price 0.775063355
- **Correlation of extra income source 1 and 2** 0.114082504
- **Correlation of extra income source 2 and 3** 0.096665346
- **Correlation of extra income source 1 and 3** 0.040438059
- **Correlation of** OBS_30_CNT_SOCIAL_CIRCLE and DEF_30_CNT_SOCIAL_CIRCLE 0.329721112
- **Correlation of** OBS_60_CNT_SOCIAL_CIRCLE and DEF_60_CNT_SOCIAL_CIRCLE 0.255930967

Include visualizations and **summarize** the most important results in the presentation.



- In above pie chart you can see the percentage of income and credit how it depends in taking credit.


Sum of AMT_CREDIT by OCCUPATION_TYPE
55.36bn (30.05%)
31.49bn (17.09%)
18.08bn (9.82%)
17.24bn (9.36%)
16.56bn (8.99%)
11.39bn (6.18%)
7.32bn (3.97%)
6.96bn (3.78%)
5.19bn (2.82%)
3.74bn (2.03%)
2.38bn (1.29%)
0.49bn (0.27%)
OCCUPATION_TYPE
NP
Laborers
Sales staff
Core staff
Managers
Drivers
High skill tech staff
Accountants
Medicine staff
Security staff
Cooking staff
Cleaning staff
Private service staff
Low-skill Laborers
Secretaries
Waiters/barmen staff
Realty agents
HR staff
IT staff

- In above pie chart we can see the NP is larger is portion but NP is not participation.
- The second high occupation people take credit is labores
- So the bank need to more concentrate on this occupation people and offer credit and they are paying the debt on time.

Approach

I used excel to use the EDA analysis and get the use full insights to solve the problem of people difficulty repaying debt.

I used statistics ,like correlation interquartile etc to find outliers.

Tech-Stack

- Excel
- Microsoft powepoint

Insights and Result

- By the outlier we can see that how the credit is high then the top income by this we can say the people are taking more credit then income.
- We have normalized the data by fille appropriate data in blanks so the imbalance of data is less.
- By the correlation we observe in different parameter is more near to 0 so we can say the there is neutral correlation .