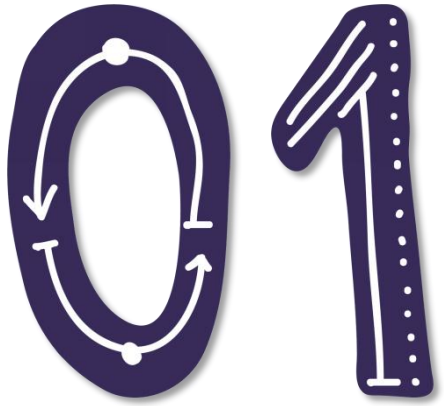
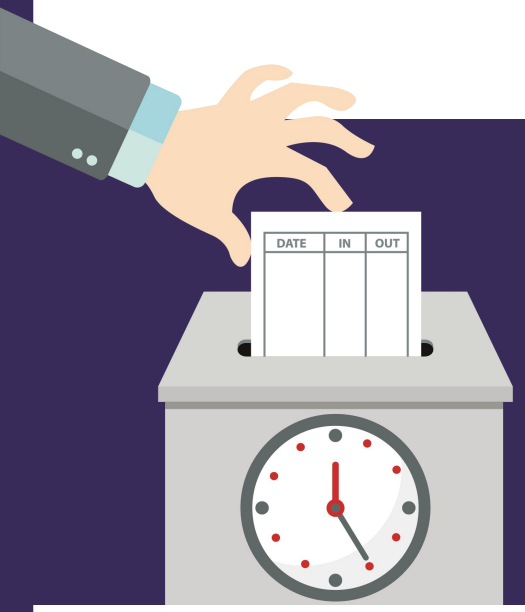
Bank Loan Case Study

Presented by Harsh Chauhan

CONTENTS



Project Description



Problem Statement:

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.

Approach

We have applied basic understanding of EDA and understand it's driving factors to improve the Analysis along with having an independent research of risk anaylsis to understand the problem.

Given Data are as follows:

> `application_data.csv` contains all the information of the client at the time of application.The data is about wheather a client has payment difficulties.

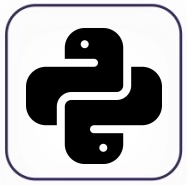
> `previous_application.csv` contains information about the client's previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.

> `columns_descrption.csv` is data dictionary which describes the meaning of the variables.

Tech Stack used

Visual Studio Code

Used Visual Studio code as a base for Jupyter notebook



Python

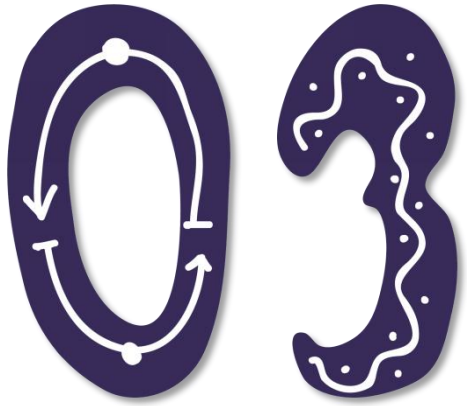
Python is used as the main programming language for datascience to code programs for running the tests and gain insights

Microsoft Excel

For rough estimation of how the dataset looks like and how much we can work on it

Jupyter Notebook

To run python and agian insights on the test



Insights

Detailed Approach

IDENTIFICATION

We have identified how we will approach the data,finding missing dataset and working on it accordingly to gain the required results

1

OUTLIERS

Identify Outliers and show how they play any role in our dataset

2

IMBALANCE

Understanding the ratio of imbalance in our data

3

Correlation Analysis

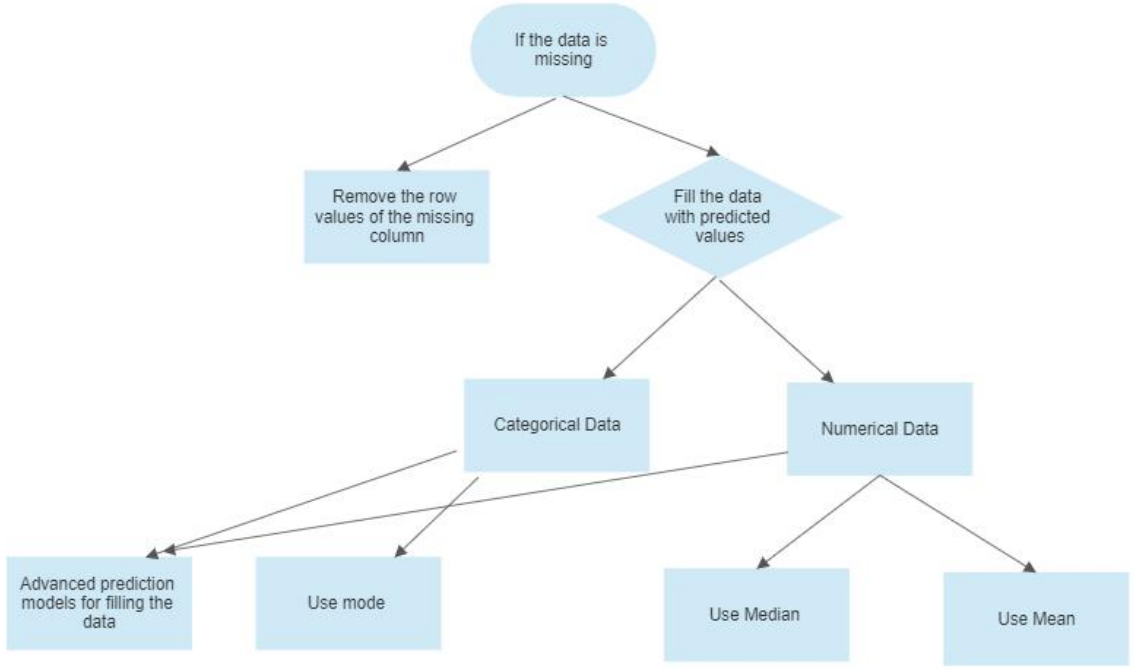
Finding the correlation between the 5 variables with respect to the target variables and find the top three correlation

4

VISUALISATION

Visualize data with the help of charts and graphs

5



Missing data approach

This is how we can deal with our dataset, we have to find whether the value which is missing has any impact on the dataset, if it doesn't, then we can remove it. If it does impact then we have to find the type of variable it is, if it's categorical data then mode, if not then we can use mean or median. Advance Machine learning models are also there to identify the missing values

Outliers

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations.

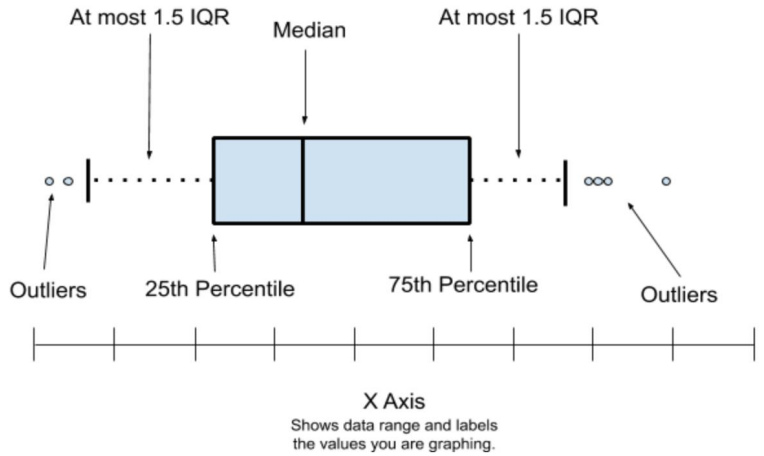
We have found that in previous_application.csv dataset, only SK_ID_CURR , SK_ID_PREV and NFLAG_INSURED_ON_APPROVAL have no outliers, the rest of all of them have outliers.

We have found that in applications.csv dataset, FLAG_DOCUMENT_3,EXT_SOURCE_3, EXT_SOURCE_2, EXT_SOURCE_1, FLAG_PHONE, DATE_ID_PUBLISH, DATE_BIRTH, and SK_ID_CURR have no outliers, others have outliers.

How it’s an Outlier?

To put it simply, you can identify such that, in this graph,

the values above 1.5IQR are considered outliers



Here is the python file where we conducted tests and identified the

outliers: Collab

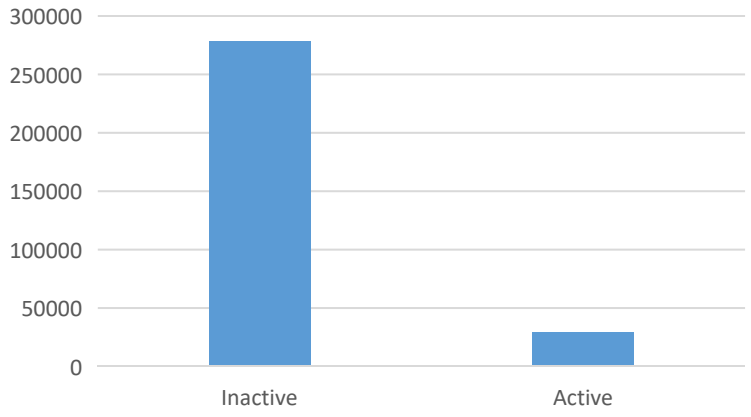
It has all the graphs of all the values of dataset which helps in identifying

which column values have outliers.

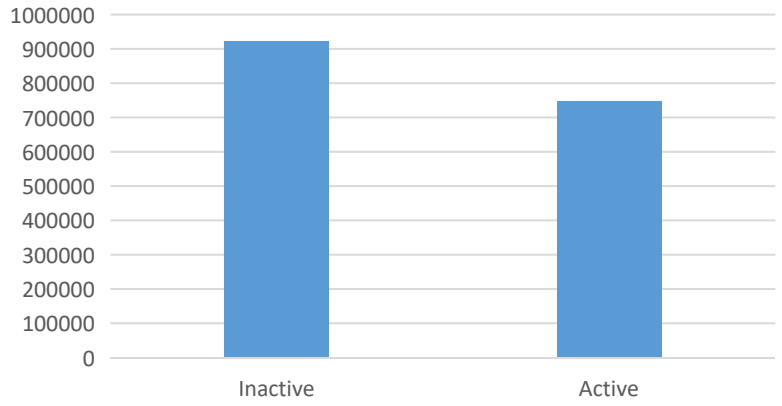
Imbalance in Data

On conducting many tests, we can say that:

In application.csv data, the percentage of data imbalance is 10.523% , where the number of active and inactive variables are 278232 and 29279



In previous_application.csv data, the percentage of data imabalance is 81.02%, where the number of active and inactive variables are 922661 and 747553



Univariate and Segmented univariate

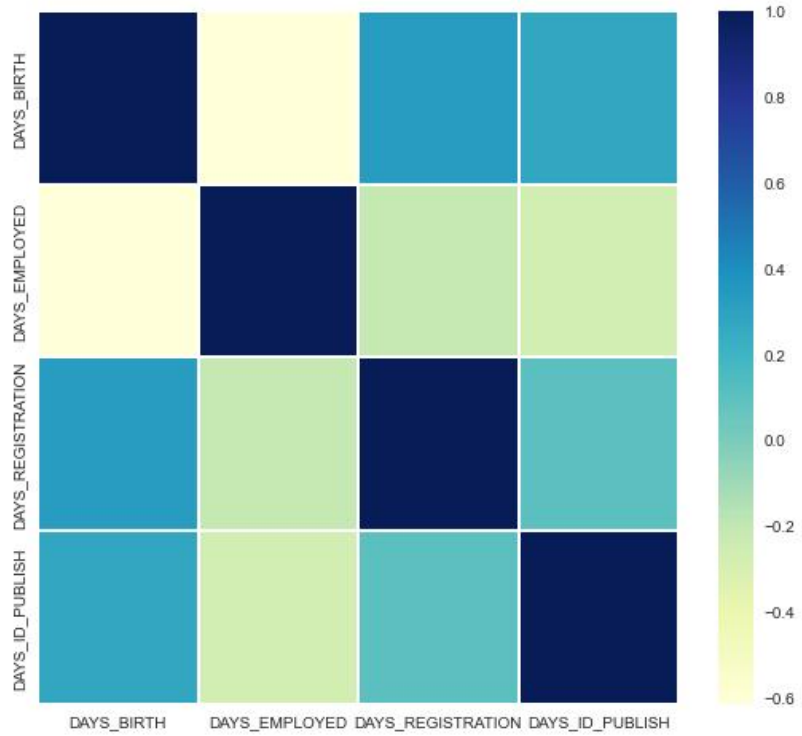
Univariate analysis is the simplest kind of data analysis in the field of statistics. This could be either descriptive or inferential in nature as is the case in any data analysis in statistics. The key thing about the univariate analysis to remember is that there is only one data involved here, since there are more variables involved in this dataset. So we will conduct bivariate analysis on the following dataset.

Bivariate analysis

Bivariate analysis is stated to be an analysis of any concurrent relation between two variables or attributes. This study explores the relationship of two variables as well as the depth of this relationship to figure out if there are any discrepancies between two variables and any causes of this difference. We have conducted correlation anaylsis on the basis of our dataset.

Correlation Analysis: 1

The following results of correlation of analysis for the applications.csv are as follows:



Correlation Matrix for applications.csv

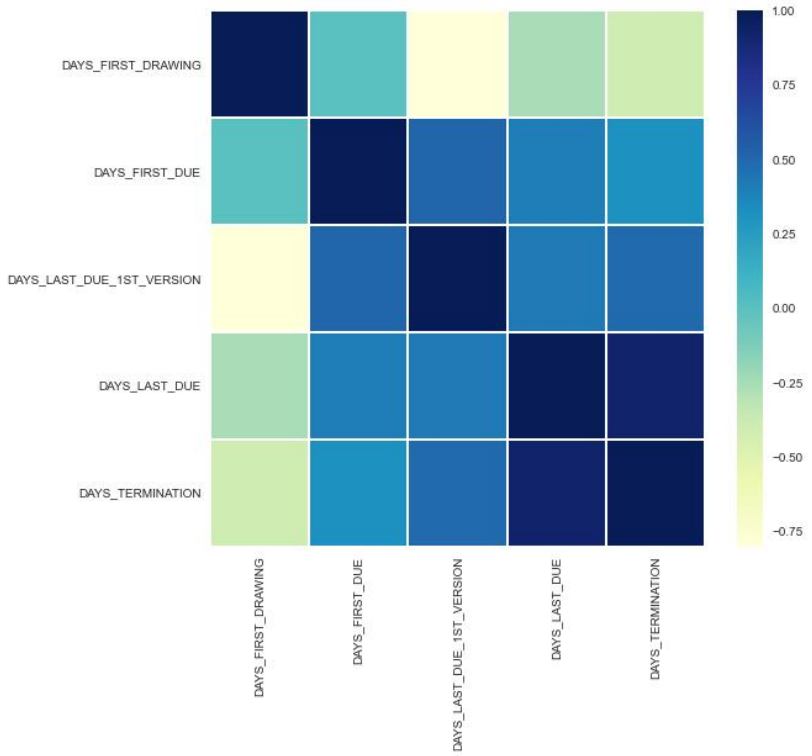
We have taken the following dataset values:DAYS_BIRTH,DAYS_EMPLOYED,DAYS_REGISTRATION,DAYS_ID_PUBLISH, and the target variable is contract_type.

The following top 3 correlations are:

DAYS_LAST_DUE DAYS_TERMINATION 0.927990
DAYS_FIRST_DRAWING DAYS_LAST_DUE_1ST_VERSION 0.803494
DAYS_FIRST_DUE DAYS_LAST_DUE_1ST_VERSION 0.513949

Correlation Analysis: 2

The following results of correlation of analysis for the previous_applications.csv are as follows:

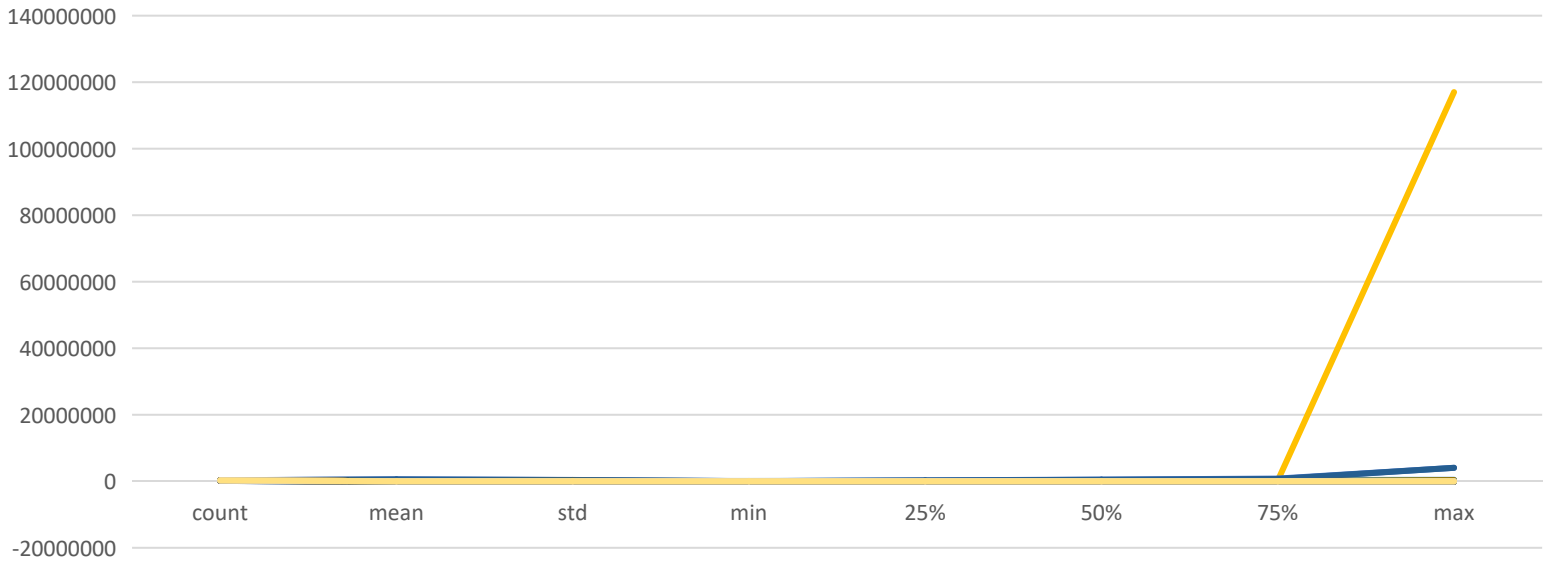


Correlation Matrix for applications.csv

We have taken the following dataset values: DAYS_FIRST_DRAWING,DAYS_FIRST_DUE,DAYS_LAST_DUE_1ST_VERSION,DAYS_LAST_DUE,DAYS_TERMINATION with respect to contact type, The top 3 correlations are:

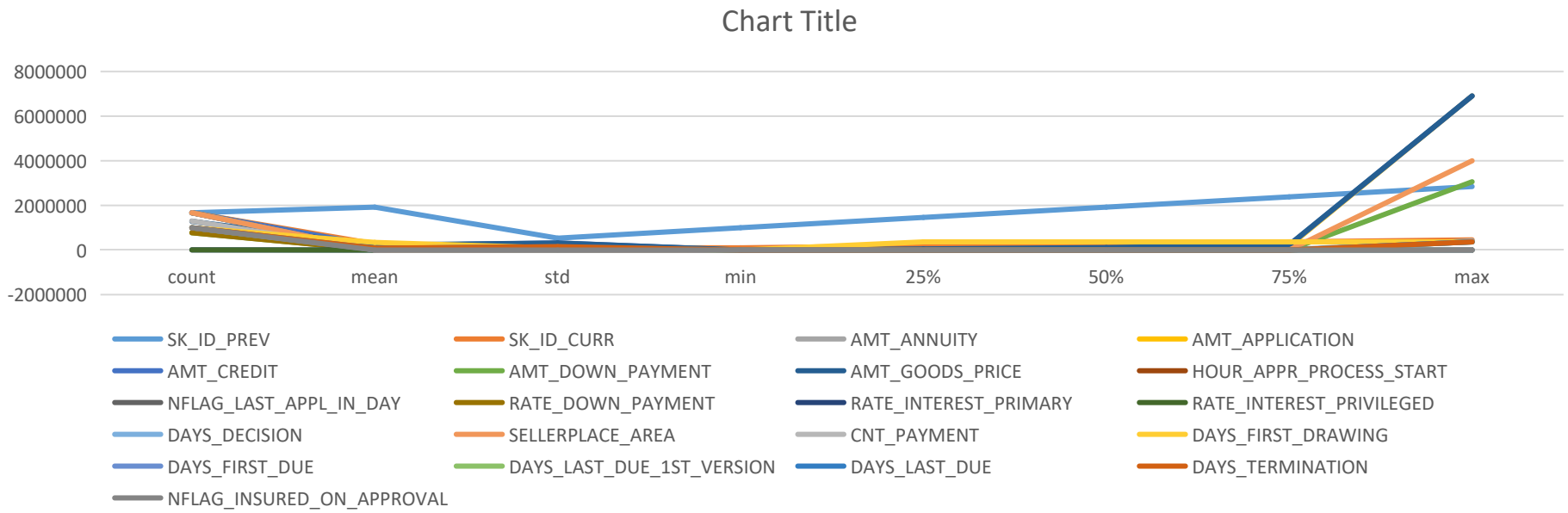
DAYS_LAST_DUE DAYS_TERMINATION 0.927990
DAYS_FIRST_DRAWING DAYS_LAST_DUE_1ST_VERSION 0.803494
DAYS_FIRST_DUE DAYS_LAST_DUE_1ST_VERSION 0.513949

Summary of the dataset,application.csv:



Summary is present in the following spreadsheet: summary_application.csv

Summary of the dataset,previous_application.csv:



Summary is present in the following spreadsheet: previous_application.csv

Result

Result:

The insights in the above slides are the required results. We have learnt a lot about what is EDA, what are outlier, what are univariate, bivariate and segmented univariate analysis. We also learnt about how to work on the data and created correlation between the variables with respect to target variable. Thank you for the opportunity to let us work on this project.

Thank You!