

# Big Data Project

Team 7 - Should This Loan be Approved or Denied?

Name	Sec	BN	ID
بموا عريان عياد	1	17	9202391
مارك ياسر نبيل	2	14	9203106
بيتر عاطف فتحي	1	18	9202395
كريم محمود كمال	2	12	9203076

# Table of Contents

1. Problem statement
2. Pipeline
3. Preprocessing and cleaning
4. EDA
5. Association rules
6. ML
7. KNN
8. Results and Evaluation
9. Fully distributed mode
10. Business insights

# 1. Problem Statement

# 1. Problem Statement

## The Problem

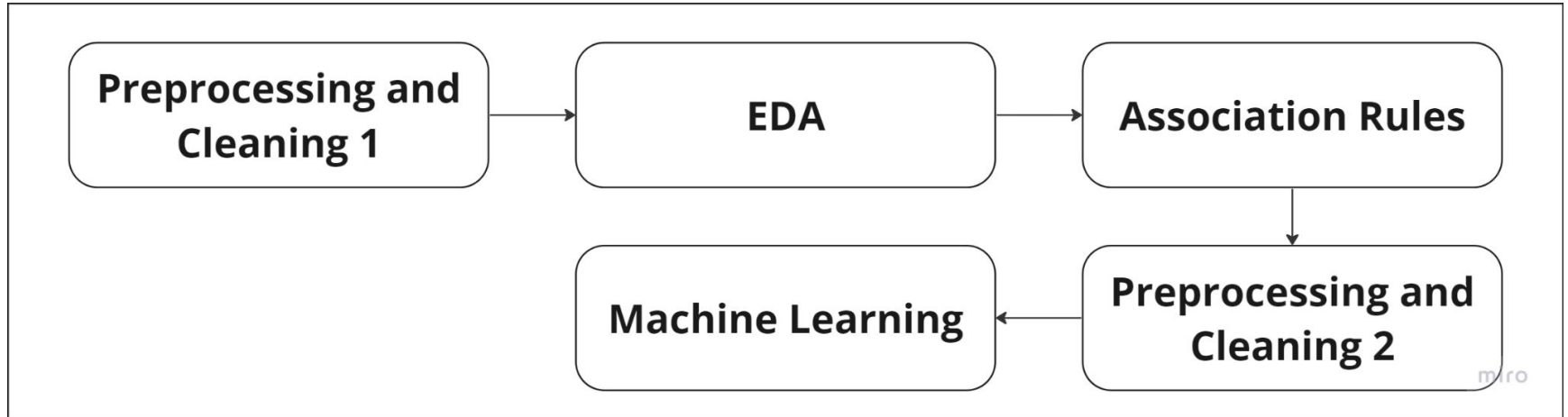
Given the dataset from the U.S. Small Business Administration (SBA) comprising loan information, the challenge is to develop a predictive model that evaluates whether the loan should be approved or denied.

## Motivation

This model should aid lending institutions in making informed decisions, ultimately contributing to the sustainability of small businesses and the broader economy.

## 2. Pipeline

## 2. Pipeline



# 3. Preprocessing and cleaning

# 3. Preprocessing and cleaning

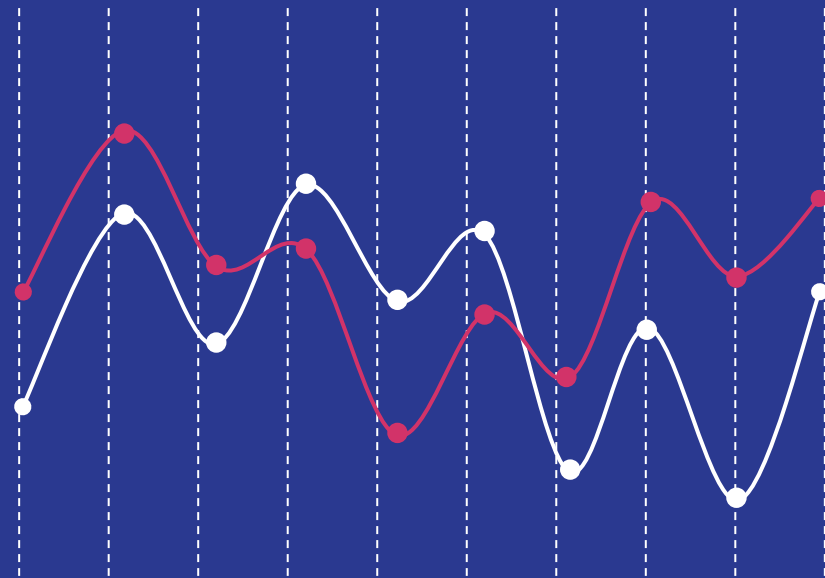
- **Done using PySpark Df**
- **Main steps before EDA**
  - Remove ID columns.
  - Try to fill/predict null values for each column based on other columns.
  - Remove null values that could not be predicted.
  - Remove wrong info (ex: non existing Zip codes).
  - Remove values that are not expected in binary columns.
  - Transform codes to their sector (5564 -> “retail”).
  - Drop unimportant dates.
  - Filter months/years of important dates.
  - Clean financial columns to be numbers instead of string (“\$50” -> 50)



# 3. Preprocessing and cleaning

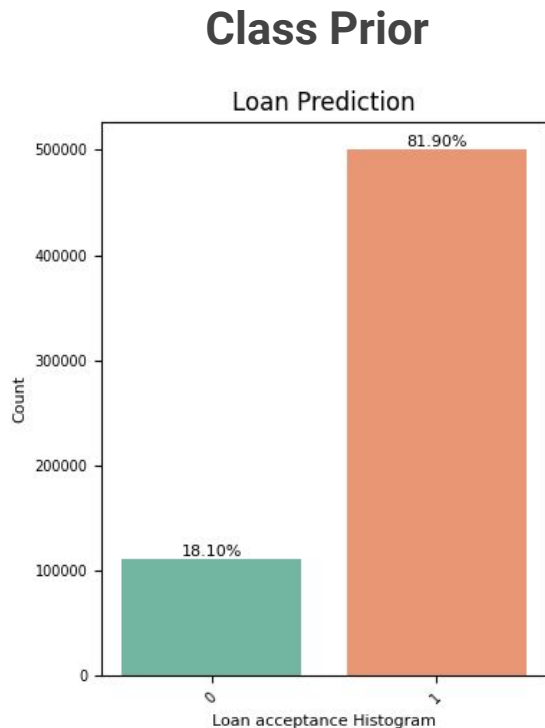
- **Main steps before ML (deduced from EDA and association rules)**
  - Drop highly correlated columns
  - Drop unimportant dates
  - Drop values that leak information to the model
  - Drop columns that contain too many unique values (that affects the model's generalization)

## 4. EDA

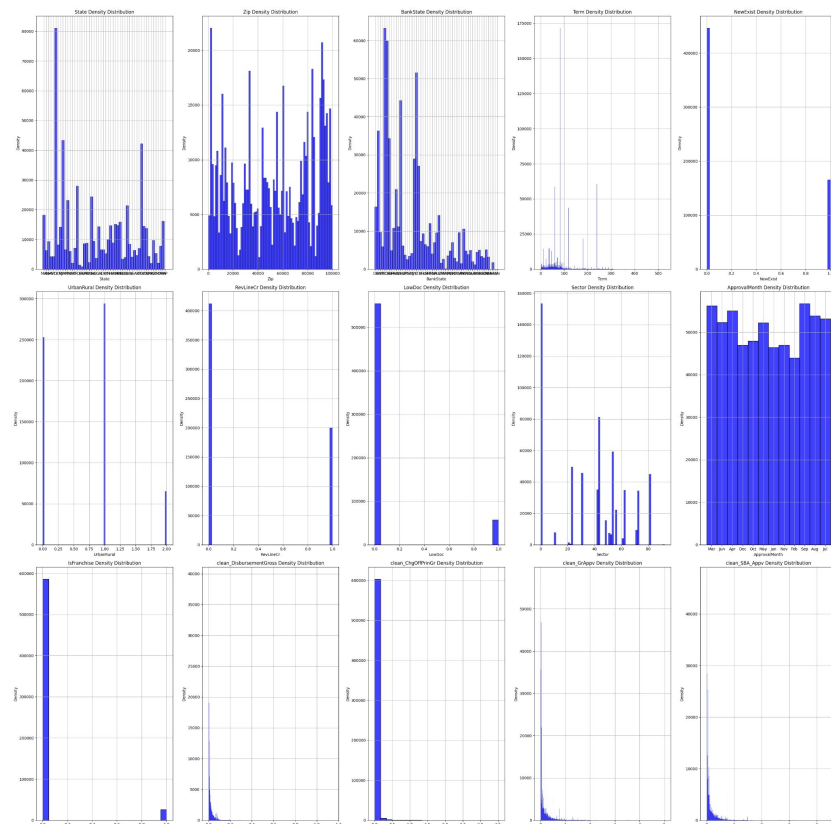


—

# Univariate Analysis



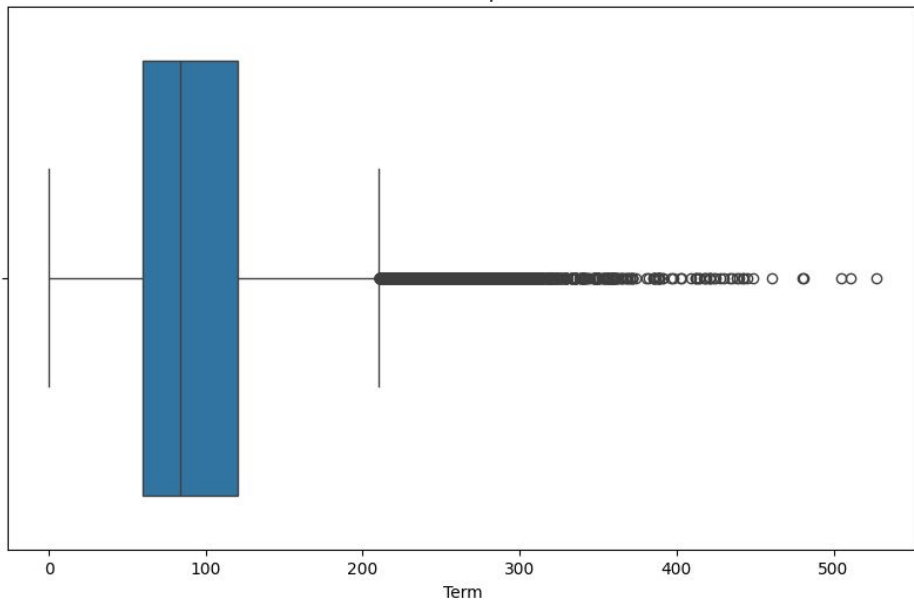
## Density function of each feature



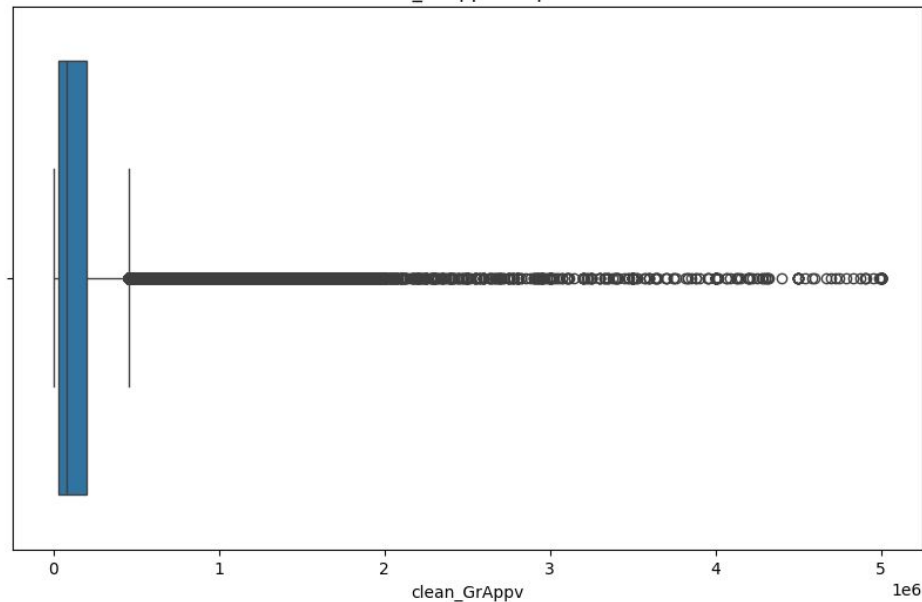
# Univariate Analysis

## Box plots

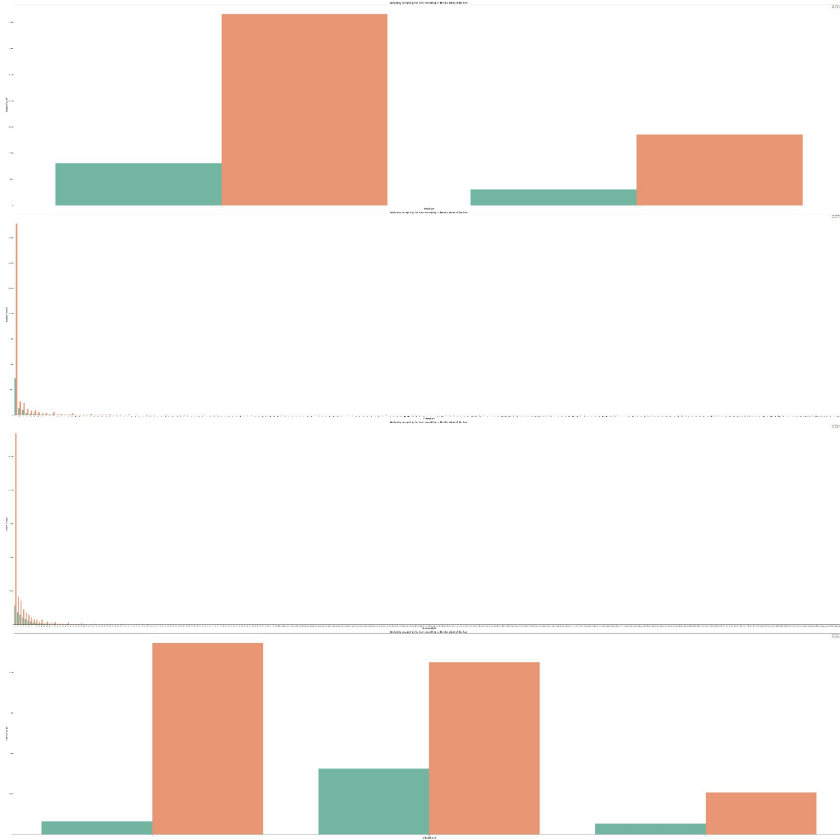
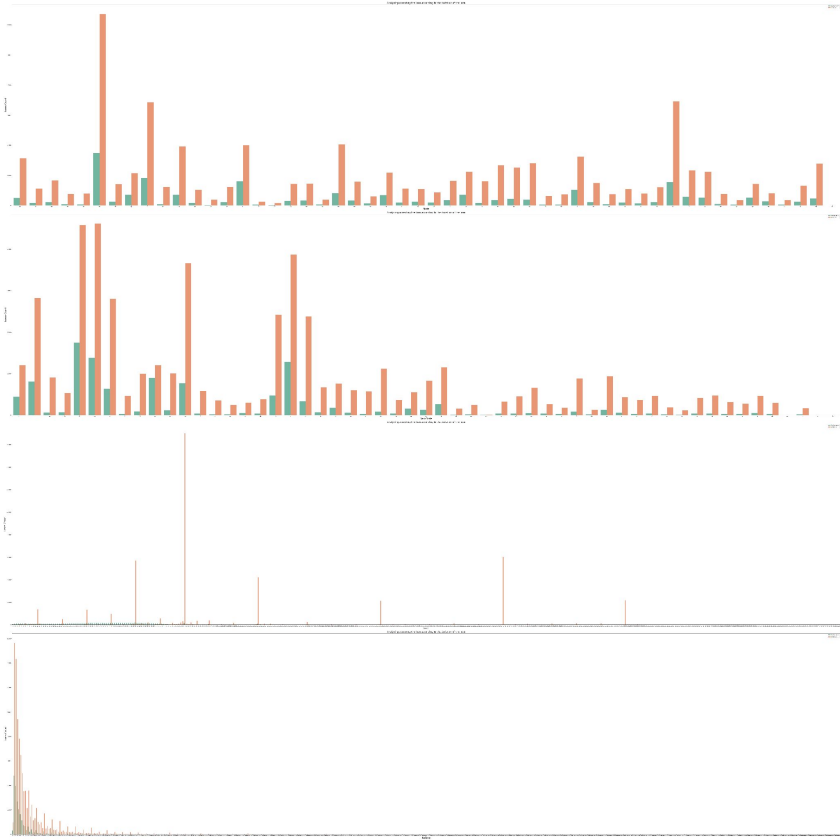
Term boxplot



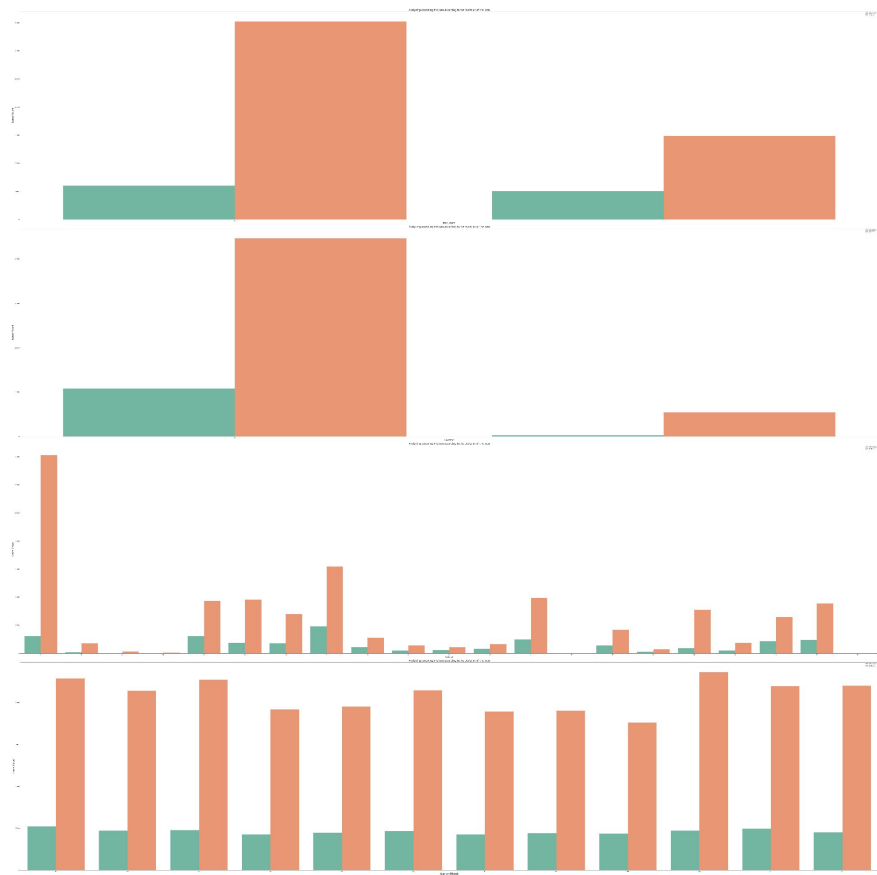
clean\_GrAppv boxplot



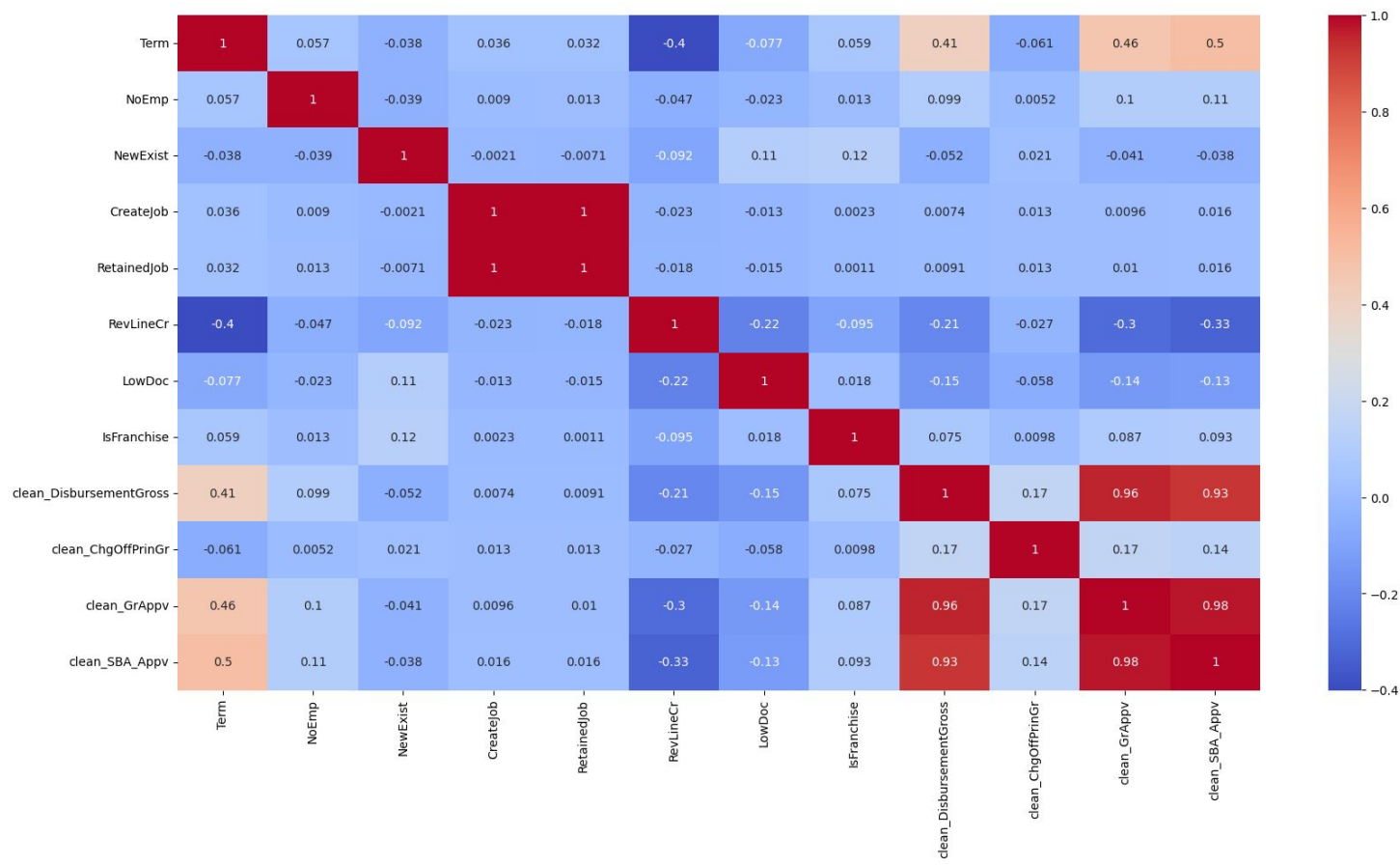
# Features vs Target Value



# Features vs Target Value



# Correlation



# 5. Association Rules



## 5. Association Rules

- Done using FPGrowth from PySpark (A parallel FP-growth algorithm to mine frequent itemsets).
- Example: Association rules sorted by Support

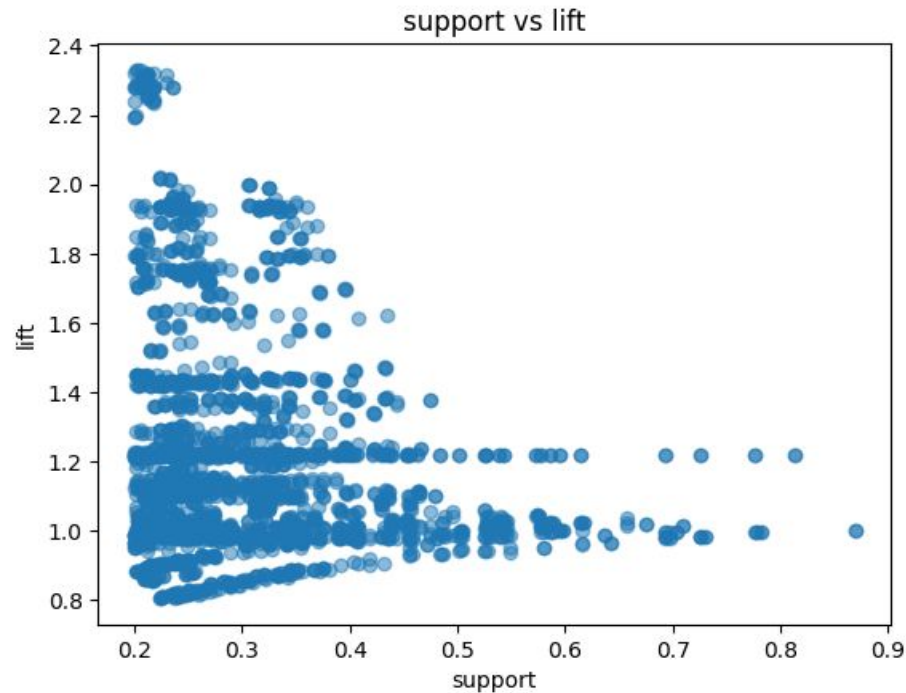
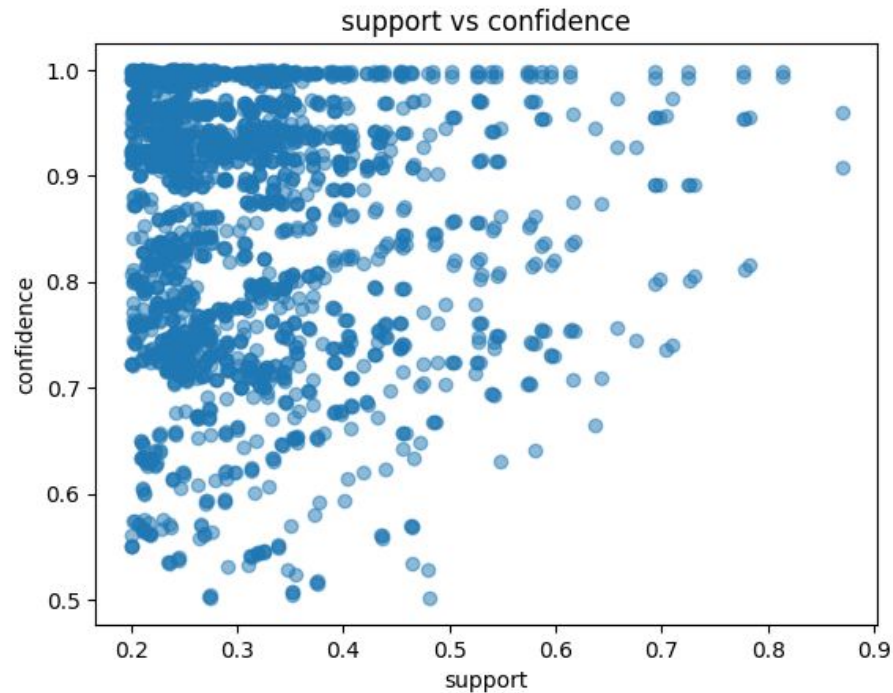
Association Rules sorted by Support:

antecedent	consequent	confidence	lift	support
[LowDoc_0]	[IsFranchise_0]	0.9591190774033764	1.001228723868855	0.8696665500795951
[IsFranchise_0]	[LowDoc_0]	0.9078488277857683	1.001228723868855	0.8696665500795951
[MIS_Status_1]	[clean_ChgOffPrinGr_0.0]	0.9936362928861776	1.21933110291366	0.8138224324421505
[clean_ChgOffPrinGr_0.0]	[MIS_Status_1]	0.9986742747119406	1.2193311029136598	0.8138224324421505
[MIS_Status_1]	[IsFranchise_0]	0.9546658205670066	0.9965799490041011	0.7819042700287328
[IsFranchise_0]	[MIS_Status_1]	0.8162333884421604	0.996579949004101	0.7819042700287328
[IsFranchise_0]	[clean_ChgOffPrinGr_0.0]	0.8119543500997248	0.996381875652817	0.7778051993475482
[clean_ChgOffPrinGr_0.0]	[IsFranchise_0]	0.9544760777785577	0.9963818756528168	0.7778051993475482
[MIS_Status_1, IsFranchise_0]	[clean_ChgOffPrinGr_0.0]	0.9934699679142149	1.2191269988436322	0.7767984100574328
[clean_ChgOffPrinGr_0.0, MIS_Status_1]	[IsFranchise_0]	0.9545060188699708	0.9964131312929693	0.7767984100574328

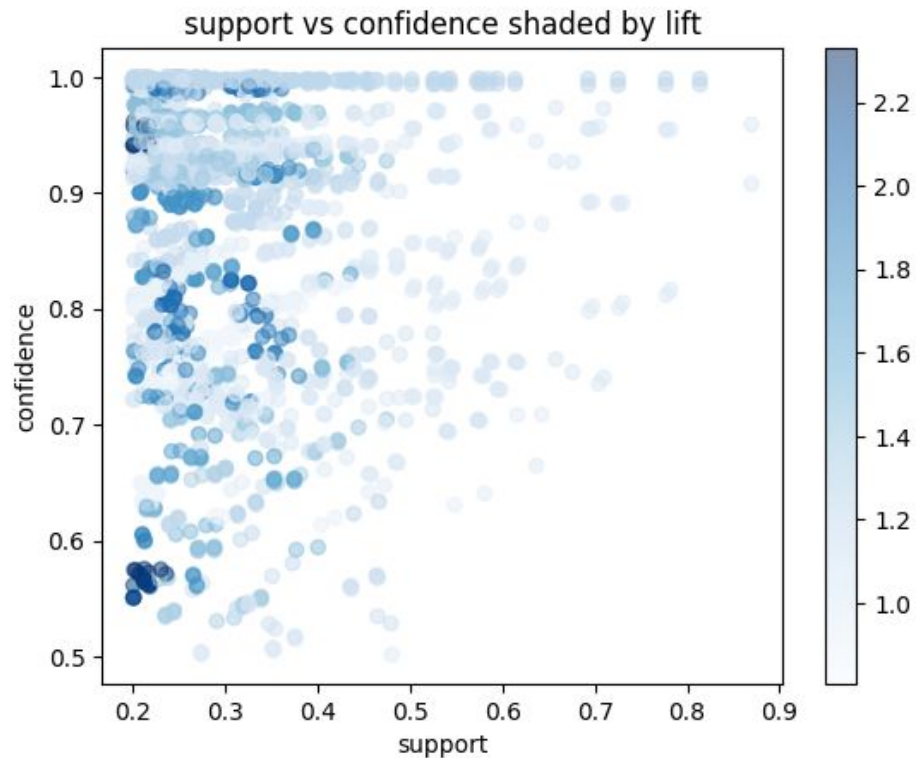
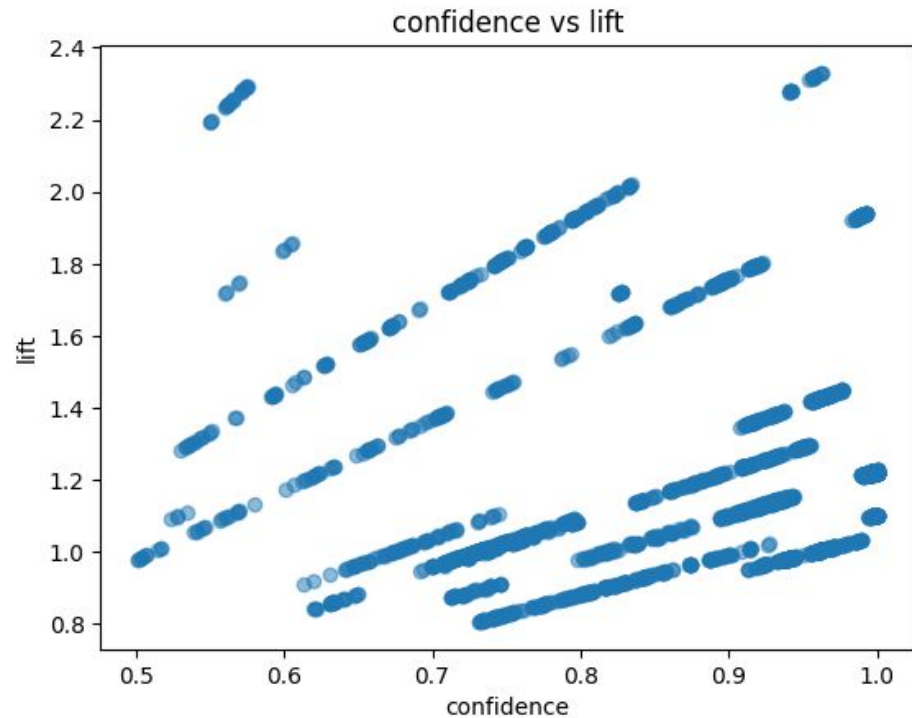
# Some Interesting Rules

- **LowDoc\_0 -> IsFranchise\_0**
  - Those who request loans that are not low doc are not franchises.
  - This rule has the highest support (0.86) and a very high confidence (0.96) and lift (1.001)
- **MIS\_Status\_1 -> clean\_ChgOffPrinGr\_0.0**
  - If a loan is paid in full, then there is no money to be charged off.
  - This rule has the 3rd highest support (0.81) and a very high confidence (0.99) and lift (1.22)
- **MIS\_Status\_1 -> IsFranchise\_0**
  - If a loan is paid in full, then the loan was requested by a non franchise
  - This rule has the 5th highest support (0.78) and a very high confidence (0.95) and lift (0.99)

# Exploring Relationships Between Support, Confidence, and Lift



# Exploring Relationships Between Support, Confidence, and Lift



# 6. ML

## 6. ML

- **Data Preparation**

- Split: Percentages for the training, validation, and test sets are 60-20-20.
- Categorical Features: Transform categorical features to one hot encoding.

- **Model Evaluation**

- After evaluating the models on the validation set, we chose the best-performing model based on the F1 score, which we will select for final training.

- **Model Training**

- Once the best model is selected, it is trained on the entire training dataset (combining training and validation sets). This step ensures the model learns from as much data as possible before final evaluation.

## 6. ML

- **Model Testing**

- The final step involves evaluating the selected model's performance on the test set, which is data the model has never seen before. This step provides a more accurate estimation of how the model will perform on unseen data in real-world scenarios.

- **Models Used**

- Logistic Regression
- Random Forest
- GBT
- SVM
- KNN

# 7. KNN



## 7. KNN

- **KNN Implementation**

- We implemented the KNN classifier using MapReduce.
- Similarity between neighbors is calculated using cosine similarity, which handles sparse vectors (due to one hot encoding of categorical features) better than euclidean distance.

- **Map**

- The map phase will determine the k-nearest neighbors in the different splits of the data.
- As a result of each map, the k nearest neighbors together with their computed distance values will be emitted to the reduce phase.
- Key-value pair: <None,{ 'similarity': dist, 'class': true\_class }>

## 7. KNN

- **Reduce**

- The reduce phase will compute the definitive neighbors from the list obtained in the map phase.
- The reduce phase will determine which are the final k nearest neighbors from the list provided by the maps.

- **Notes**

- There is no training error since there is no training at all for the KNN classifier, it just stores the training data.
- Results and evaluation were tested using only 50,000 rows for the full dataset, and 500 rows from the validation set after splitting. Evaluating each point requires calculating the distance between the validation point and each training point, which consumes a lot of time even when running in fully distributed mode on Azure.

## 8. Results and Evaluation

## 8. Results and Evaluation

	Logistic Regression (maxIter=10)		Random Forest		GBT (maxIter=100)		SVM (maxIter=100)		KNN (k=3)	
	Train	Validation	Train	Validation	Train	Validation	Train	Validation	Train	Validation
Accuracy	0.8831	0.8795	0.8191	0.8193	0.9336	0.9333	0.8885	0.8845	-	0.722
Precision	0.8751	0.8706	0.6710	0.6713	0.9317	0.9313	0.8815	0.8768	-	0.697
Recall	0.8831	0.8795	0.8191	0.8193	0.9336	0.9333	0.8885	0.8845	-	0.592
F1 Score	0.8734	0.8693	0.7377	0.7380	0.9318	0.9314	0.8818	0.8775	-	0.64

# Best Performing Model

As observed, the best performing model is GBT, so we will train on training+validation set and test it on test data

	GBT	
	Train	Test
Accuracy	0.9357	0.9345
Precision	0.9339	0.9326
Recall	0.9357	0.9345
F1 Score	0.9340	0.9328

## 9 . Fully distributed mode






## 9 . Fully distributed mode

- **Deployed a Spark cluster on Azure using HDInsight, with the following specifications:**
  - **Head node x2**
    - E8 V3 (8 Cores, 64 GB RAM)
  - **Zookeeper node x3**
    - A2 v2 (2 Cores, 4 GB RAM)
  - **Worker node x3**
    - A4m v2 (4 Cores, 32 GB RAM)
- **Ran each Jupyter notebook of the project on it in fully distributed mode.**

# Hosts Configuration


Microsoft Azure

Search resources, services, and docs (G+I)



14712019101648@stud...  
CAIRO UNIVERSITY - STUDENTS ...

Home > bd-team7-spark

 **bd-team7-spark** | Cluster size ☆ ...

HDInsight cluster

Search

« Refresh Feedback

Overview

Activity log

Access control (IAM)

Tags

Diagnose and solve problems

Settings

Cluster size

Quota limits

SSH + Cluster login

Data Lake Storage Gen1

Storage accounts

Applications

Script actions

External metastores

Properties

Locks

Monitoring

Summary

Automatically increase or decrease the number of worker nodes based on a schedule or specific performance metrics. [Learn More](#)

This estimate does not include subscription discounts or costs related to storage, networking, or data transfer.

Save Discard

Node type	Node size	Number of nodes	Estimated cost/hour
Head node	E8 V3 (8 Cores, 64 GB RAM), 0.76 USD/hour	2	1.52 USD
Zookeeper node	A2 v2 (2 Cores, 4 GB RAM), 0.00 USD/hour (FREE)	3	0.00 USD (FREE)
Worker node	A4m v2 (4 Cores, 32 GB RAM), 0.33 USD/hour	<input type="text" value="3"/> ✓	0.98 USD
<input type="checkbox"/> Enable autoscale			
Total estimated cost/hour			2.50 USD

Cluster size history



# Hosts Configuration

← ↻ 🔒 https://bd-team7-spark.azurehdinsight.net/#/main/hosts

Ambari

Dashboard

Services

- HDFS
- YARN
- MapReduce2
- Tez
- Hive
- Oozie
- ZooKeeper
- Ambari Metrics
- Zeppelin Note...
- Jupyter
- Spark3
- WebHCat

Hosts

Alerts

Cluster Admin

bd-team7-s... ⚙️ 0 🔔 0

admin

## Hosts

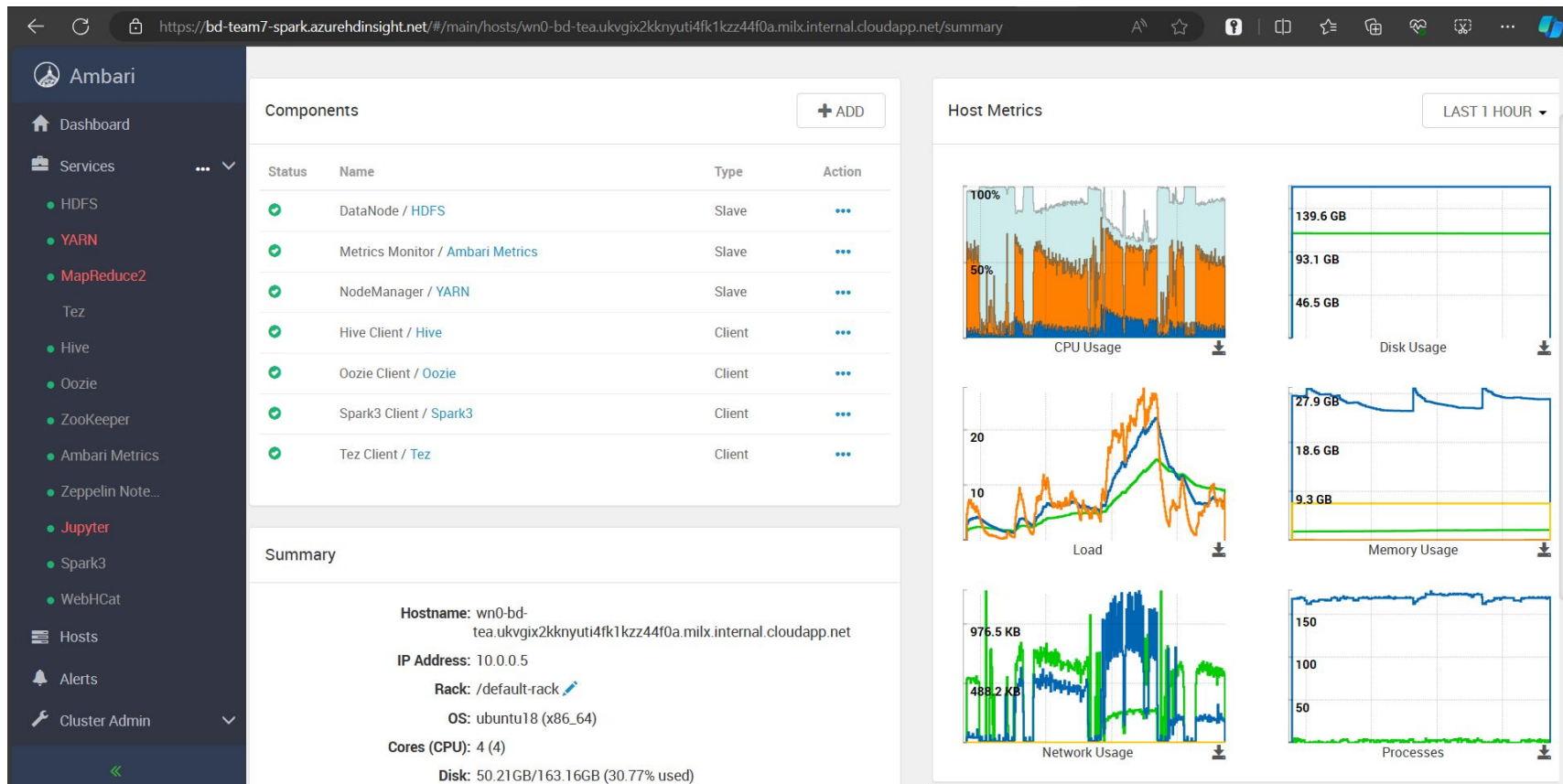
⌵ ACTIONS

<input type="checkbox"/>	Name	IP Address	Rack	Cores	RAM	Disk Usage	Load Avg	Versions	Components
<input type="checkbox"/>	hn0-bd-tea.ukvgix2kkny...	10.0.0.17	/default-rack	8 (8)	62.80GB			HDInsight-5.1.5.3	24 Components
<input type="checkbox"/>	hn1-bd-tea.ukvgix2kkny...	10.0.0.18	/default-rack	8 (8)	62.80GB			HDInsight-5.1.5.3	19 Components
<input type="checkbox"/>	wn0-bd-tea.ukvgix2kkny...	10.0.0.5	/default-rack	4 (4)	31.35GB			HDInsight-5.1.5.3	7 Components
<input type="checkbox"/>	wn1-bd-tea.ukvgix2kkny...	10.0.0.7	/default-rack	4 (4)	31.35GB			HDInsight-5.1.5.3	7 Components
<input type="checkbox"/>	wn3-bd-tea.ukvgix2kkny...	10.0.0.8	/default-rack	4 (4)	31.35GB		24.53	HDInsight-5.1.5.3	7 Components
<input type="checkbox"/>	zk0-bd-tea.ukvgix2kkny...	10.0.0.11	/default-rack	2 (2)	3.83GB			HDInsight-5.1.5.3	4 Components
<input type="checkbox"/>	zk1-bd-tea.ukvgix2kkny...	10.0.0.10	/default-rack	2 (2)	3.83GB			HDInsight-5.1.5.3	4 Components
<input type="checkbox"/>	zk2-bd-tea.ukvgix2kkny...	10.0.0.9	/default-rack	2 (2)	3.83GB			HDInsight-5.1.5.3	4 Components

Items per page: 10 1 - 8 of 8

Licensed under the Apache License, Version 2.0.  
See third-party tools/resources that Ambari uses and their respective authors

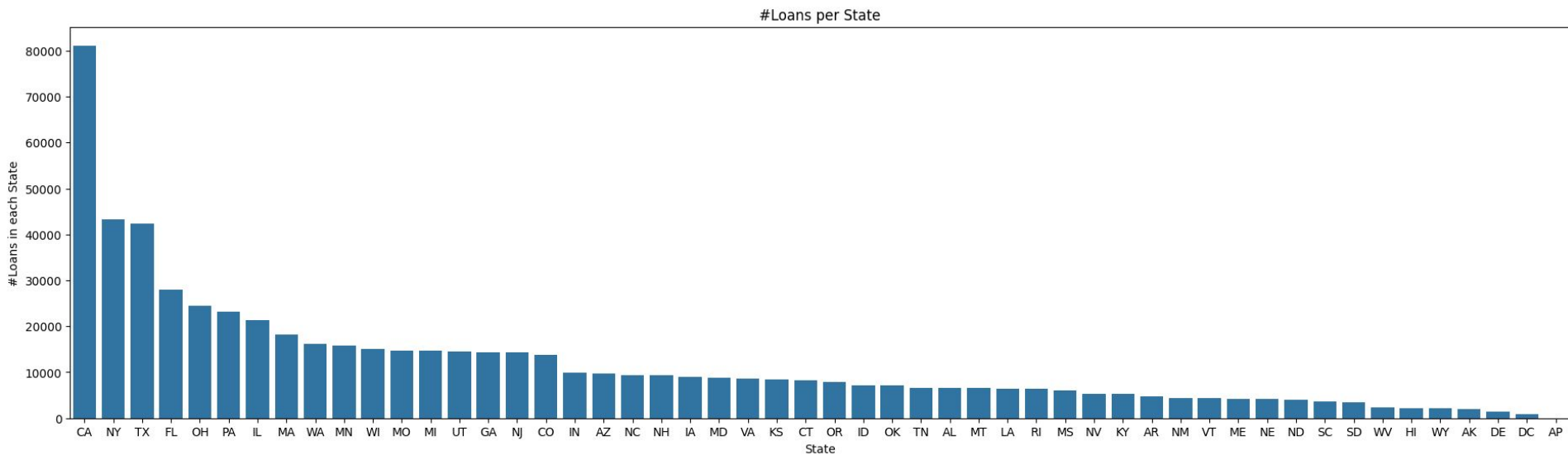
# Worker 0 metrics



# 10. Business Insights

# 10. Business Insights - EDA

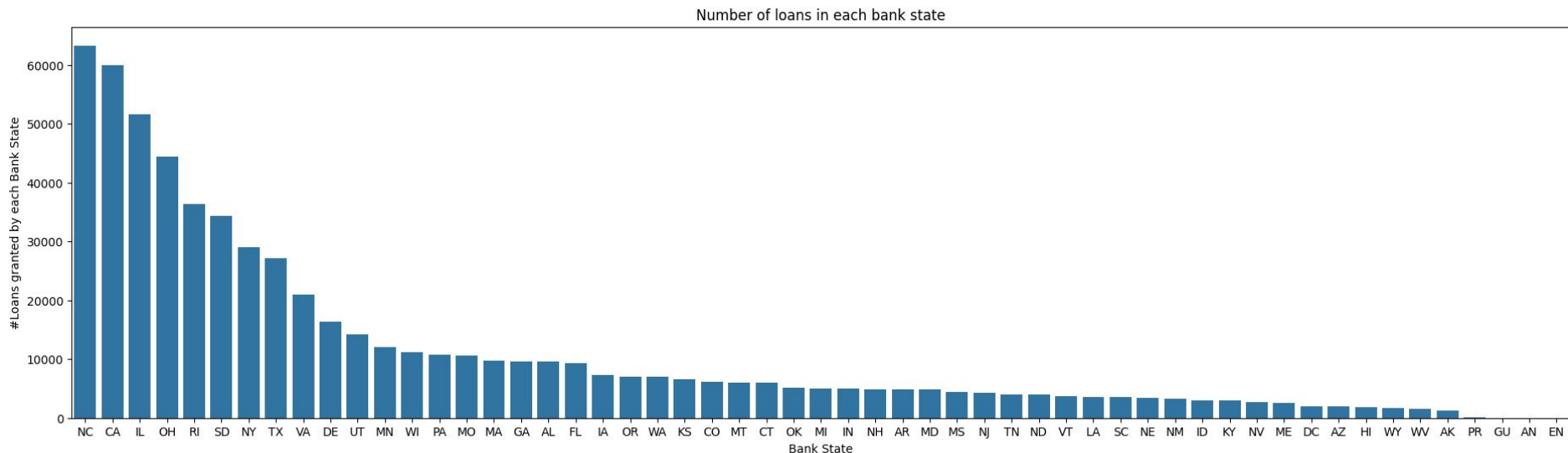
From EDA (Number of loans per state)



The vast majority of loans originate from California, due to factors like: high population density, high economic activity, active real estate market.

# 10. Business Insights - EDA

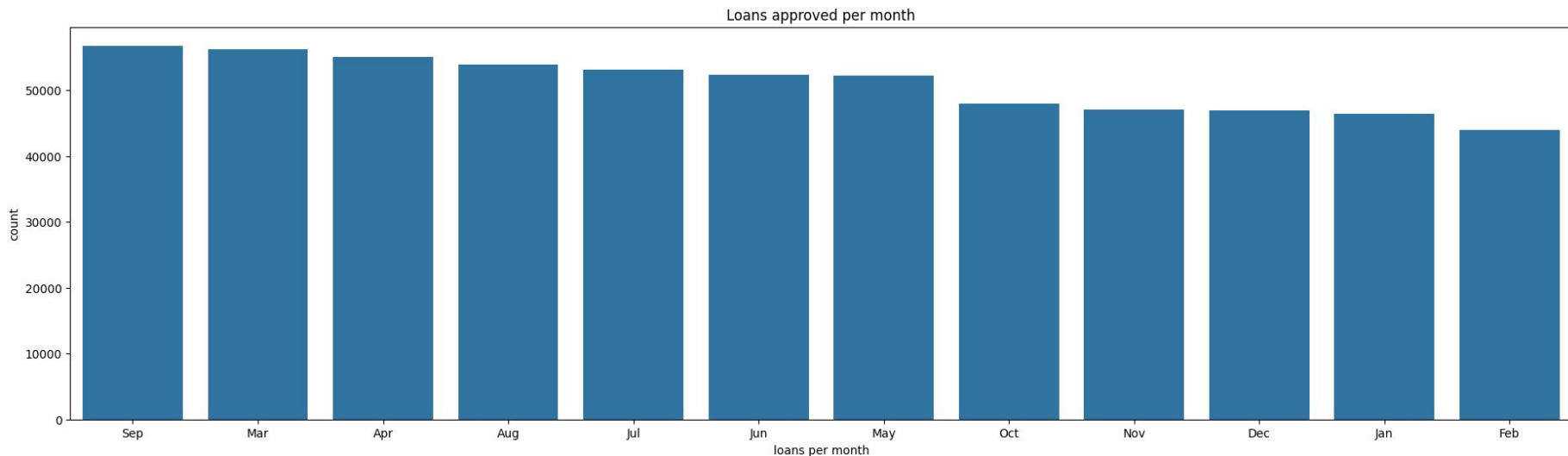
## From EDA (Number of loans per bank state)



Banks often extend loans to small businesses located in different states. For instance, while most loans are provided by banks headquartered in North Carolina, North Carolina doesn't rank among the top 15 states in terms of loan origination.

# 10. Business Insights - EDA

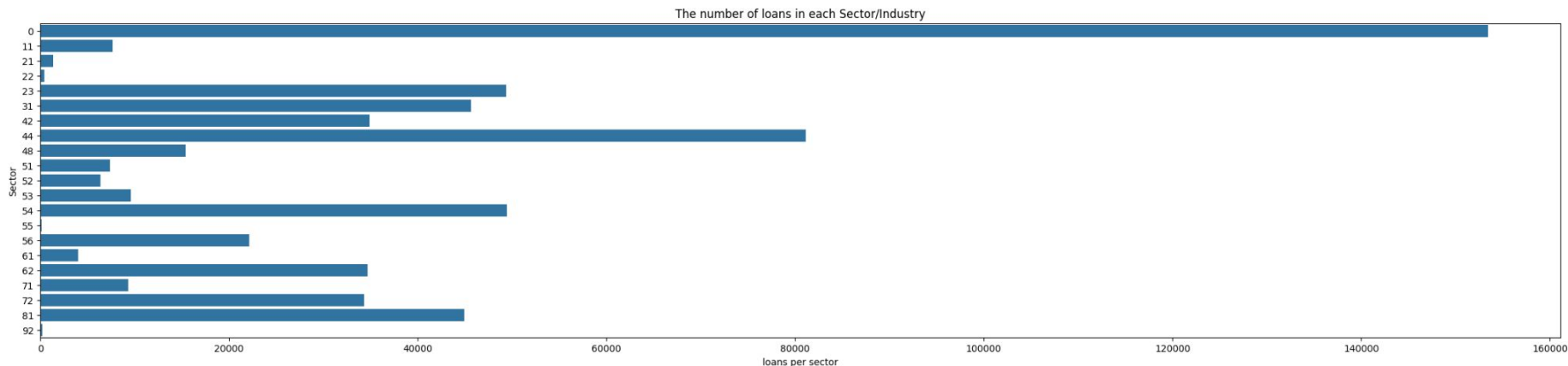
From EDA (Number of loans approved per month)



The highest number of loans approved in September, due to factors like the end of summer vacations, back-to-school expenses, or fiscal year-end considerations for businesses and organizations.

# 10. Business Insights - EDA

From EDA (Number of loans in each sector)



Most loans are lended to the “Retail Trade” sector (NAICS code 44).

# 10. Business Insights - Association Rules

From Association Rules, we can find some interesting rules after sorting them according to Support, Confidence and Lift:

- **LowDoc\_0 -> IsFranchise\_0**
  - Those who request loans that are not low doc are not franchises.
  - This rule has the highest support (0.86) and a very high confidence (0.96)
- **MIS\_Status\_1 -> clean\_ChgOffPrinGr\_0.0**
  - If a loan is paid in full, then there is no money to be charged off.
  - This rule has the 3rd highest support (0.81) and a very high confidence (0.99)
- **MIS\_Status\_1 -> IsFranchise\_0**
  - If a loan is paid in full, then the loan was requested by a non franchise
  - This rule has the 5th highest support (0.78) and a very high confidence (0.95)





Thank You.