Rakamin Academy

JUPYTHON

Dokumen Laporan Final Project

STAGE 1 (PREPARATION)

3 April – 9 APRIL 2023

Descriptive Statistics

Apakah ada kolom dengan tipe data kurang sesuai, atau nama kolom dan isinya kurang sesuai?

Dari semua informasi di atas, terlihat bahwa data tersebut memiliki 26 feature/ kolom. Antara nama kolom dengan isinya juga sudah sesuai. namun ada beberapa kolom yang tipe datanya masih harus disesuaikan.

Apakah ada kolom yang memiliki nilai kosong? Jika ada, apa saja?

Terdapat 11 kolom yang memiliki nilai kosong yaitu Name, City, Zip, Bank, BankState, NewExist, RevLineCr, LowDoc, ChgOffDate, DisbursementDate, MIS_Status.

Descriptive Statistics

Apakah ada kolom yang memiliki nilai summary agak aneh

	LoanNr_ChkDgt	Zip	NAICS	Term	NoEmp	NewExist	CreateJob	RetainedJob	FranchiseCode	UrbanRural
count	8.991640e+05	899164.000000	899164.000000	899164.000000	899164.000000	899028.000000	899164.000000	899164.000000	899164.000000	899164.000000
mean	4.772612e+09	53804.391241	398660.950146	110.773078	11.411353	1.280404	8.430376	10.797257	2753.725933	0.757748
std	2.538175e+09	31184.159152	263318.312760	78.857305	74.108196	0.451750	236.688165	237.120600	12758.019136	0.646436
min	1.000014e+09	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	2.589758e+09	27587.000000	235210.000000	60.000000	2.000000	1.000000	0.000000	0.000000	1.000000	0.000000
50%	4.361439e+09	55410.000000	445310.000000	84.000000	4.000000	1.000000	0.000000	1.000000	1.000000	1.000000
75%	6.904627e+09	83704.000000	561730.000000	120.000000	10.000000	2.000000	1.000000	4.000000	1.000000	1.000000
max	9.996003e+09	99999.000000	928120.000000	569.000000	9999.000000	2.000000	8800.000000	9500.000000	99999.000000	2.000000

- NAICS, ada potensi Negatively skewed
- Term, ada potensi Positively skewed
- NoEmp, ada potensi Positively skewed
- CreateJob, ada potensi Positively skewed
- RetainedJob, ada potensi Positively skewed
- FranchiseCode, ada potensi Positively skewed

Descriptive Statistics

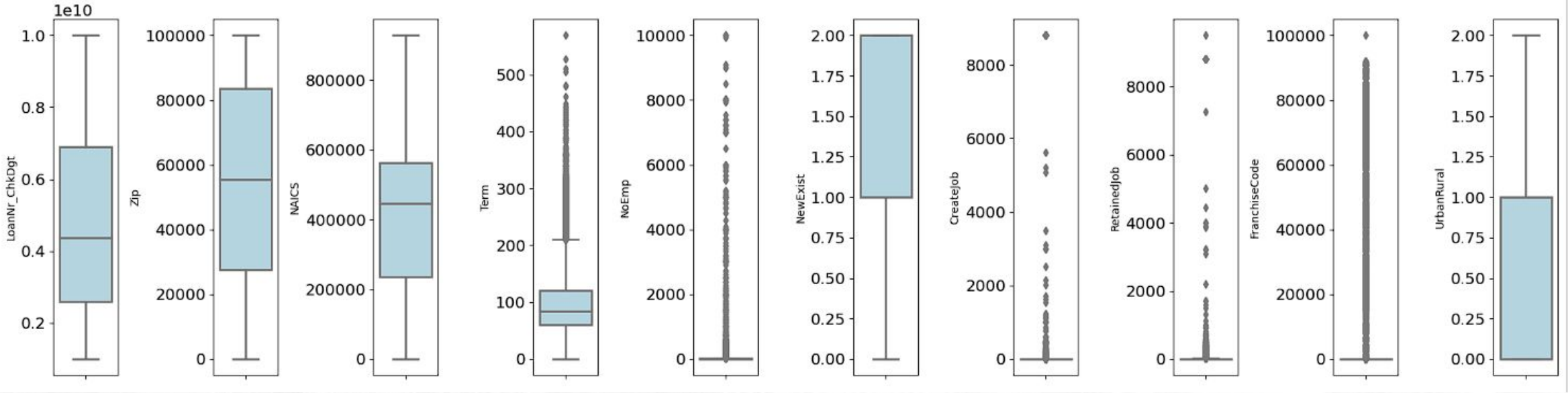
Apakah ada kolom dengan tipe data kurang sesuai, atau nama kolom dan isinya kurang sesuai?

1. Tipe data DisbursementGross, BalanceGross, ChgOffPrinGr, GrAppv, SBA_Appv seharusnya float karna merupakan nominal angka yang bisa dihitung.
2. Tipe data UrbanRural dan Zip seharusnya string, karna walaupun bentuknya angka. tetapi tidak dapat digunakan untuk perhitungan.
3. Tipe data NewExist seharusnya integer karna bentuknya binary.
4. Tipe data ApprovalFY, ApprovalDate, DisbursemenDate, ChgOffDate harusnya datetime.
5. RevlineCr dan LowDoc isinya kurang sesuai dimana seharusnya melambangkan 'Yes' dan 'No'.
6. UrbanRural terdapat isi yang kurang sesuai seharusnya Urban dan Rural tetapi ada satu kategori undefine.
7. ApprovalFY ada isi yang kurang sesuai dimana ada tahun '1976A'.
8. NAICS banyak isi yang unique dan tidak beraturan, sehingga kita bisa mengambil 2 angka didepan untuk mengkategorikannya.

Apakah ada kolom yang memiliki nilai kosong? Jika ada, apa saja?

Terdapat 11 kolom yang memiliki nilai kosong yaitu Name, City, Zip, Bank, BankState, NewExist, RevLineCr, LowDoc, ChgOffDate, DisbursementDate, dan MIS_Status.

Univariate Analysis



Numerical:

LoanNr_ChkDgt (Primary Key) :

1. tidak ditemukan outlier

ZIP (Kode Pos) :

1. tidak ditemukan outlier
2. persebaran cukup merata dimana 50 % nasabah tinggal dikota yg berkode pos 4xxxx - 8xxxx

Univariate Analysis

NAICS (Kode Klasifikasi Indusri) :

1. tidak ditemukan outlier
2. 50 persen industri yang menggunakan pinjaman berkode 2xxxx-6xxxxx

Term (jangka waktu pinjaman dalam bulan):

1. ditemukan high outlier
2. sebagian besar kurang dari 100 bulan waktu pinjaman

NoEmp (Jumlah Karyawan):

1. ditemukan high outlier
2. sebagian besar tidak memiliki karyawan

NewExist(bisnis baru atau existing):

- sebaran berbentuk bimodal
- seharusnya kategorical.

Univariate Analysis

CreateJob (jumlah loker yang tercipta):

1. ditemukan highoutlier
2. sebagian besar pekerjaan tercipta 0 / tidak ada

RetainedJob (jumlah pekerjaan yg bertahan):

1. ditemukan highoutlier
2. sebagaian besar pekerjaan yang bertahan 0/tidak ada.

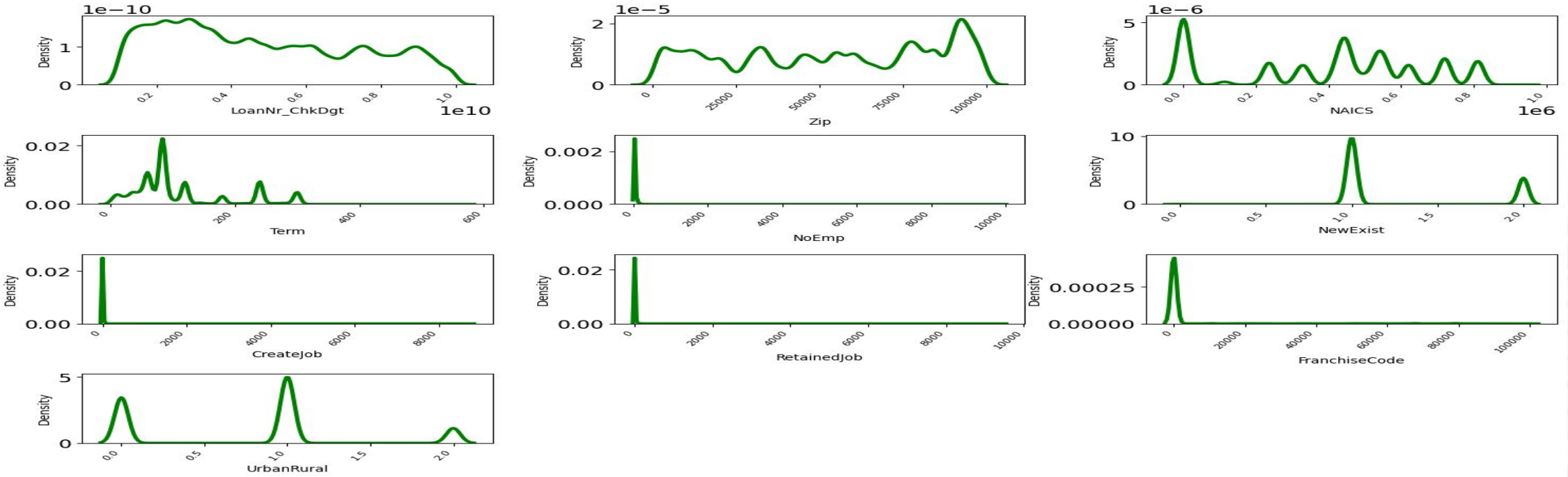
FranchiseCode (kode franchise):

1. ditemukan high outlier

UrbanRural (perkotaan/pedesaan):

1. sebagian besar bernilai 1 artinya nasabah tinggal di perkotaan
2. sebaran berbentuk multimodal

Univariate Analysis



Numerical:

Term :

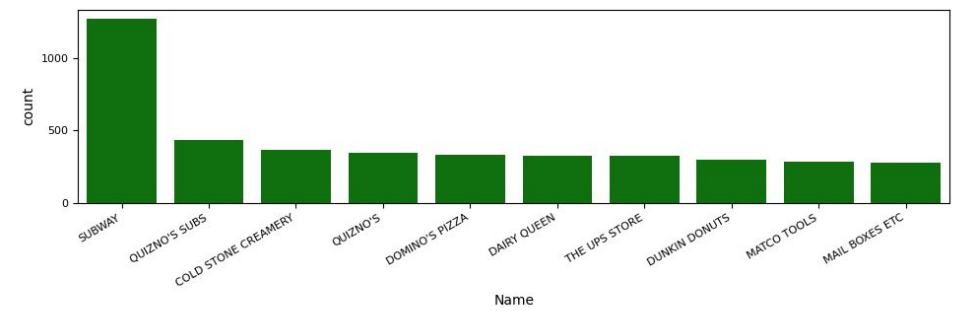
1. Jika dilihat dari visualisasinya Term memiliki distribusi skew kanan yang artinya mean > median.

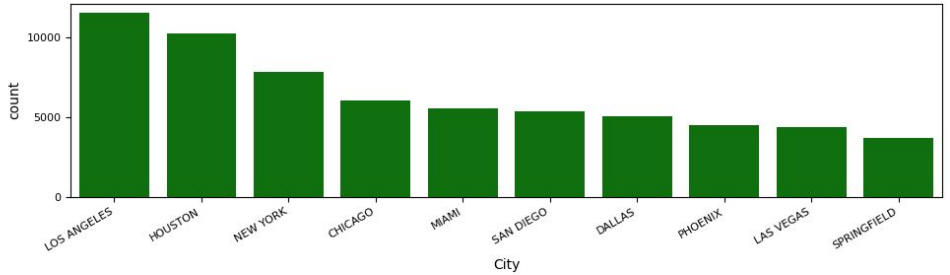
NoEmp, CreateJob, RetainedJob, FranchiseCode :

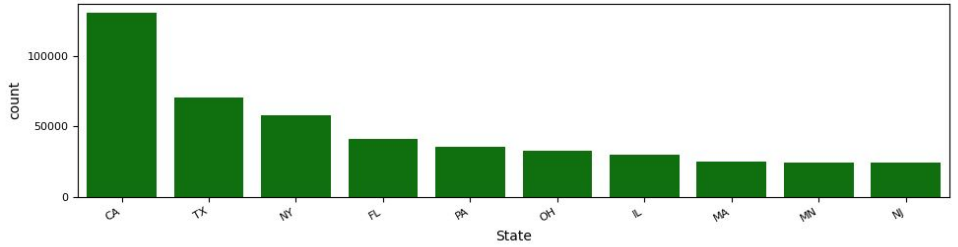
1. Jika dilihat dari visualisasinya kolom-kolom ini memiliki outliers karna distribusinya sangat dominan disebalah kiri.

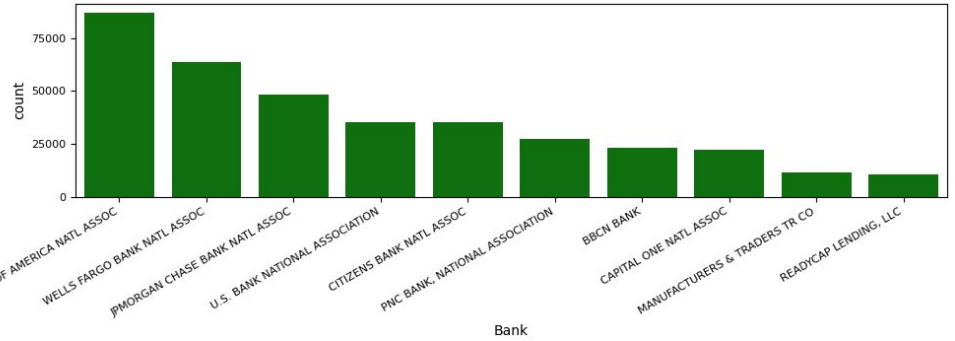
Univariate Analysis

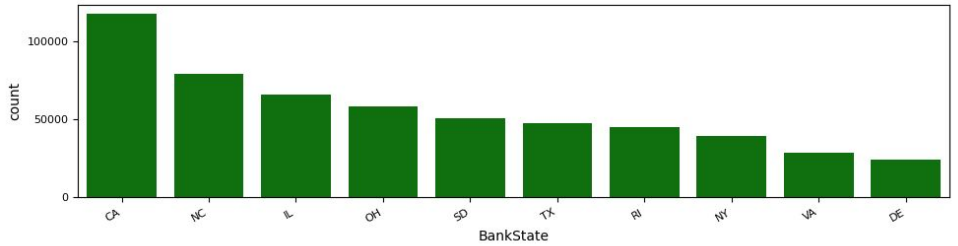
Categorical :

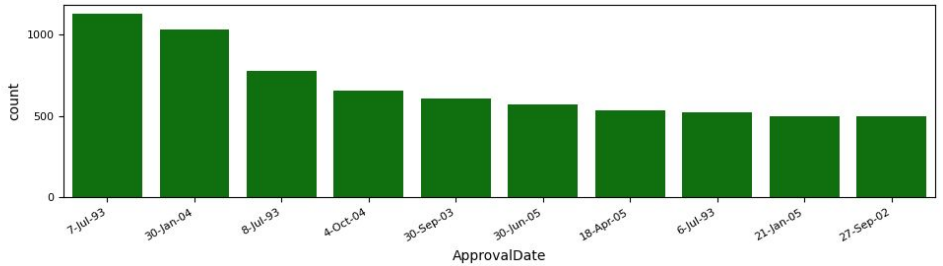






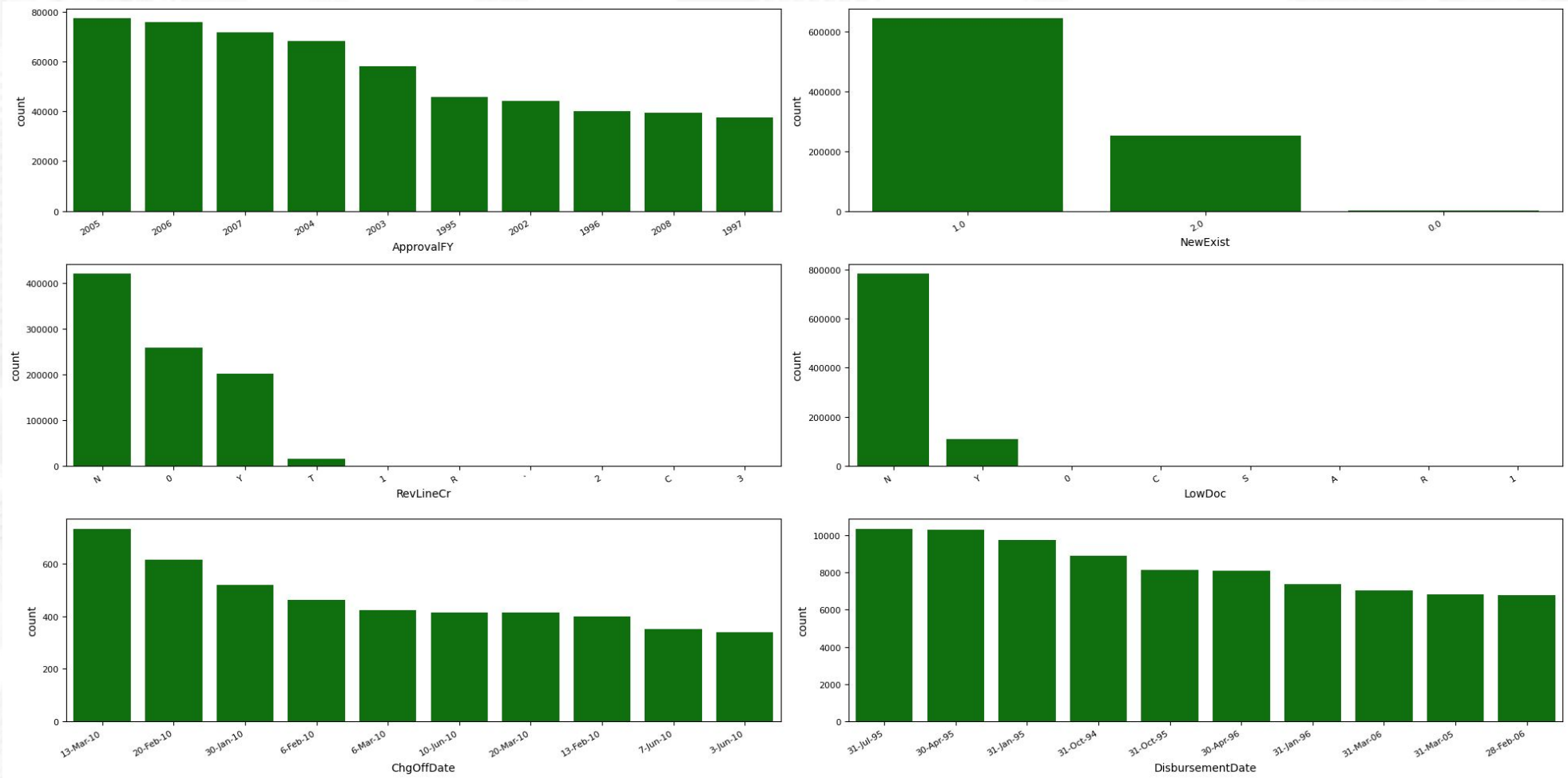






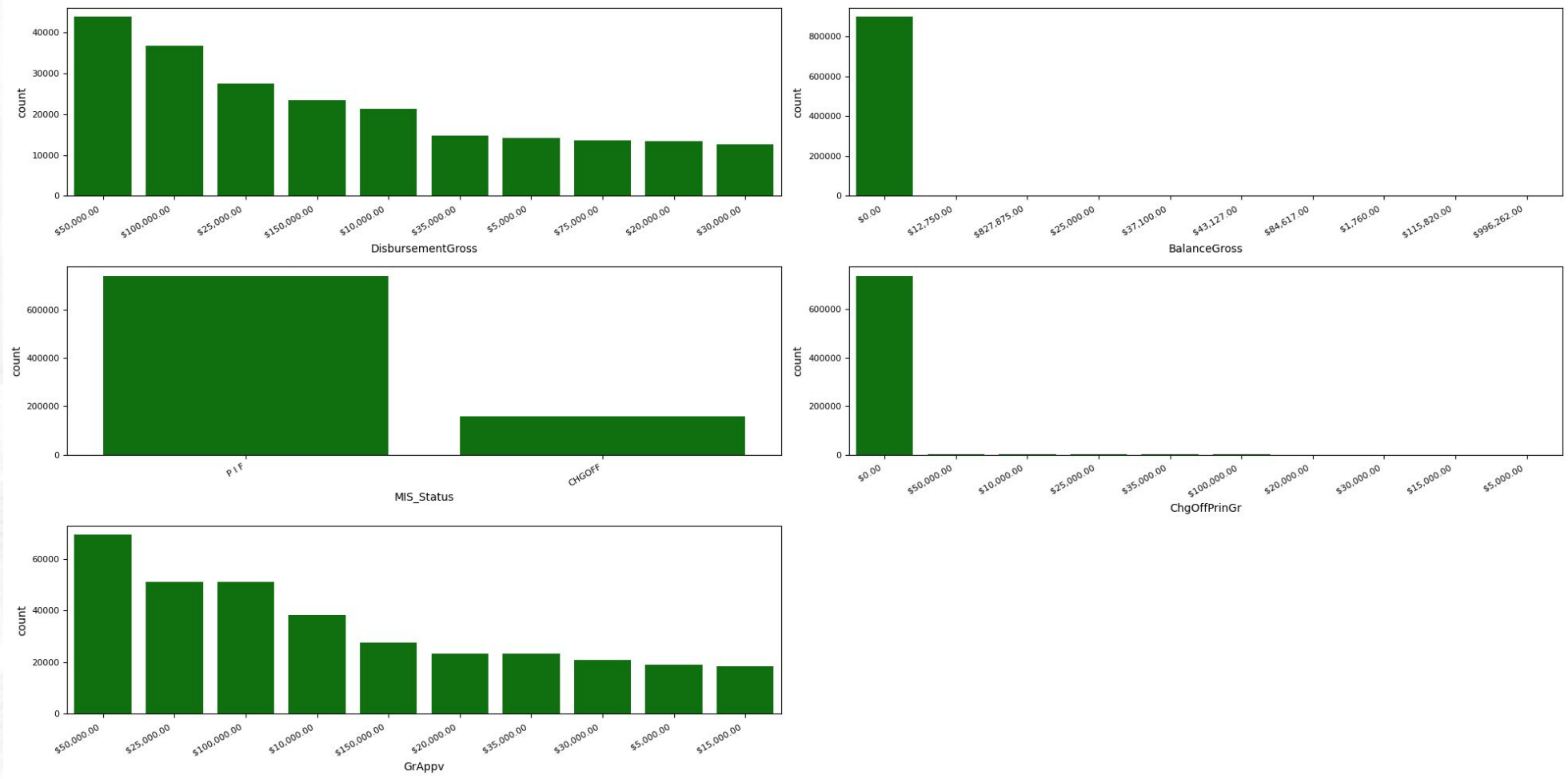
Univariate Analysis

Categorical :



Univariate Analysis

Categorical :

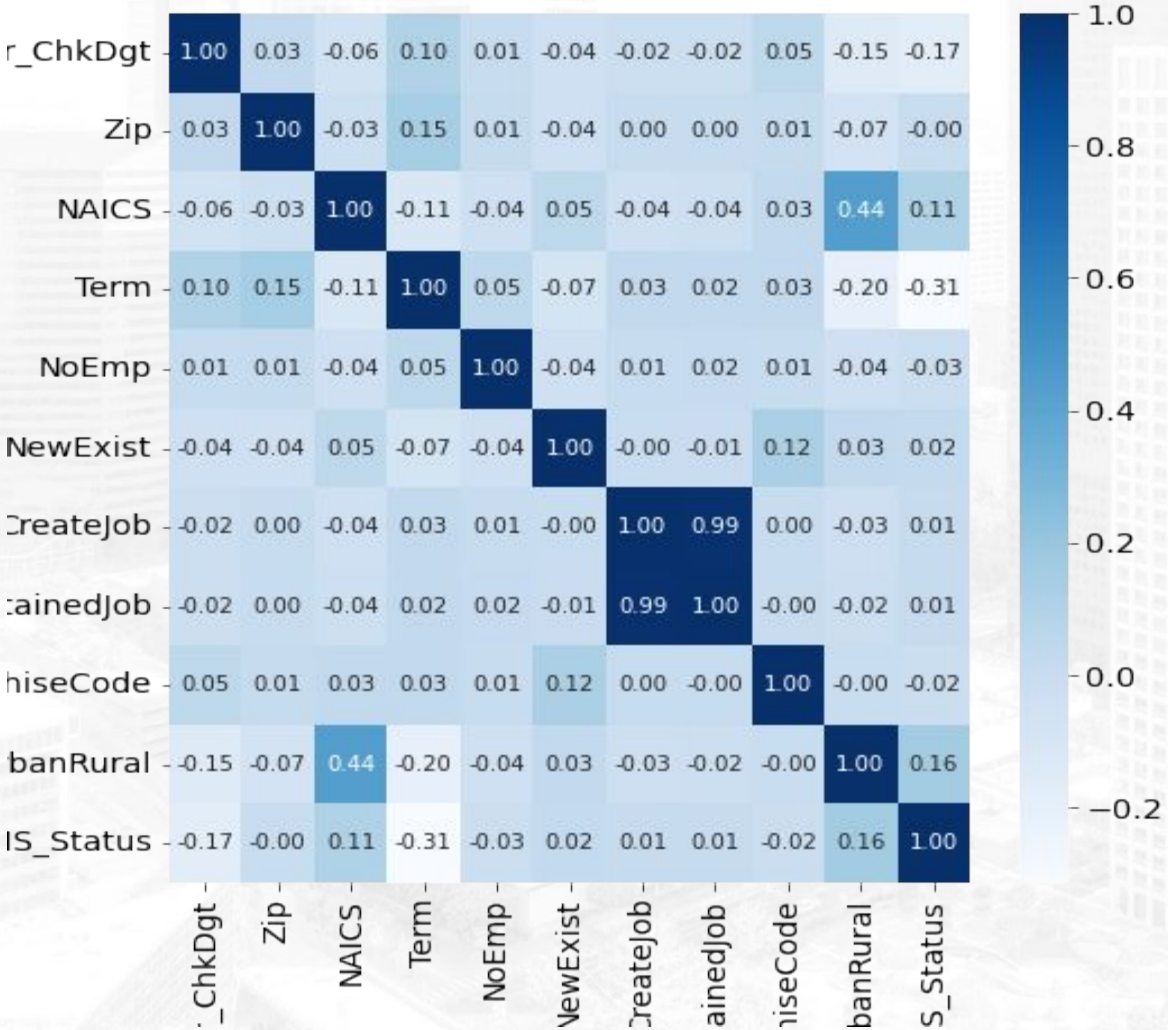


Univariate Analysis

Categorical :

1. Jika dilihat dari visualisasi jumlah peminjam paling banyak berada pada state **CA**
2. Kolom MIS_Status, LowDoc, dan NewExist memiliki 2 unique value.

Multivariate Analysis

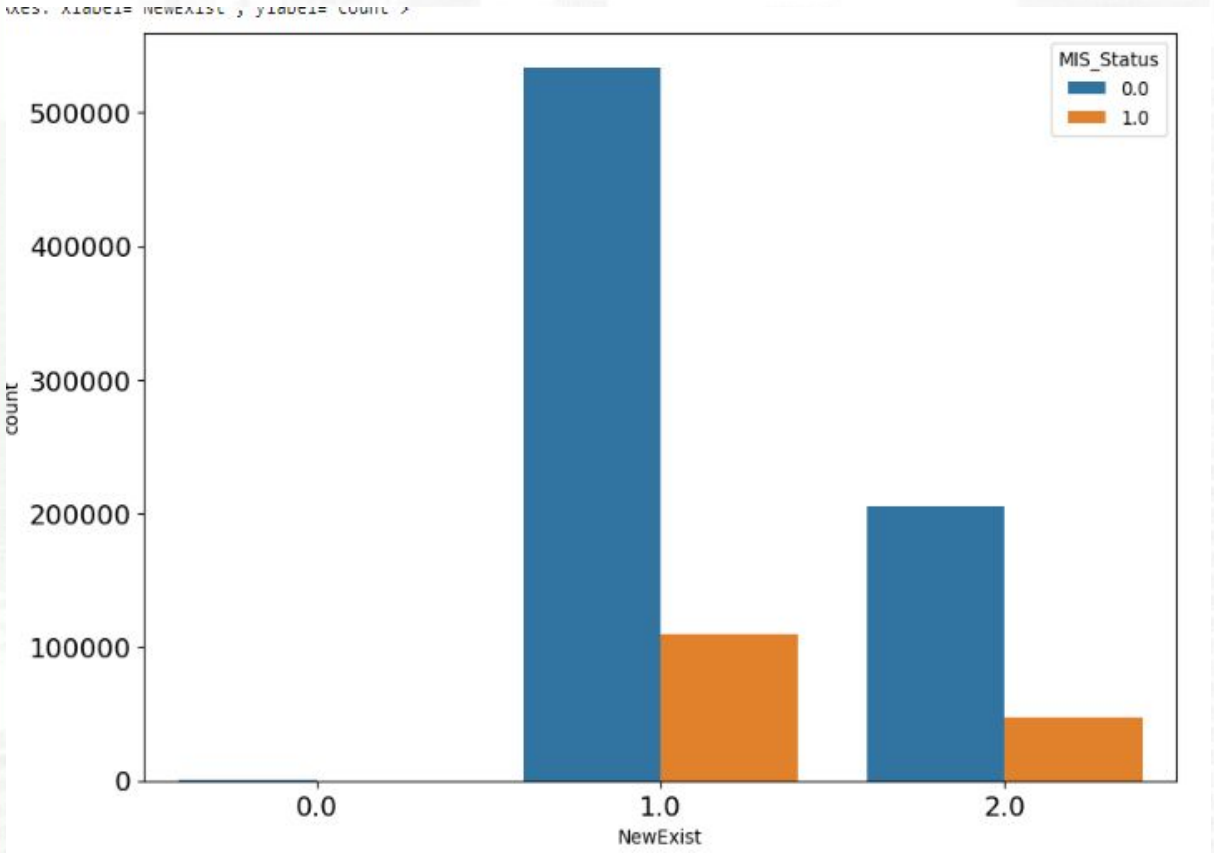


A. Bagaimana korelasi antara masing-masing feature dan label. Kira-kira feature mana saja yang paling relevan dan harus dipertahankan?

Korelasi antar feature numerik dengan target

- MIS_Status sebagai variabel target / variabel dependent / label, MIS_Status adalah variabel yang ingin diprediksi dari variabel bebas lainnya.
- Tidak ada korelasi yang kuat antara target MIS_Status dengan feature numerikal yang ada, mengindikasikan penggunaan model non-linear lebih tepat untuk dataset ini. Nilai korelasi tertinggi target-feature ada pada MIS_Status dengan urbanural, yaitu 0.16 (tidak cukup kuat).
- informasi yang didapatkan pada visualisasi heatmap yaitu kolom CreateJob dan RetainedJob memiliki korelasi positif yang sangat kuat yaitu sebesar(0.99). Ada kemungkinan kolom ini redundan, sehingga pada tahap selanjutnya dapat menghapus salah satu kolom ini.

Business Insight



dari grafik disamping dapat dilihat bahwa mayoritas nasabah yang gagal bayar merupakan yang memilik bisnis yang sudah ada (existing).

Link Git Hub

https://github.com/gustiayuseptiandani/Homework-Week-11.git

Link Google Colab

https://colab.research.google.com/drive/10kMUX5FlOVc77t2hKpm2YAn1BGQc9OwC?usp=sharing