Rakamin Academy

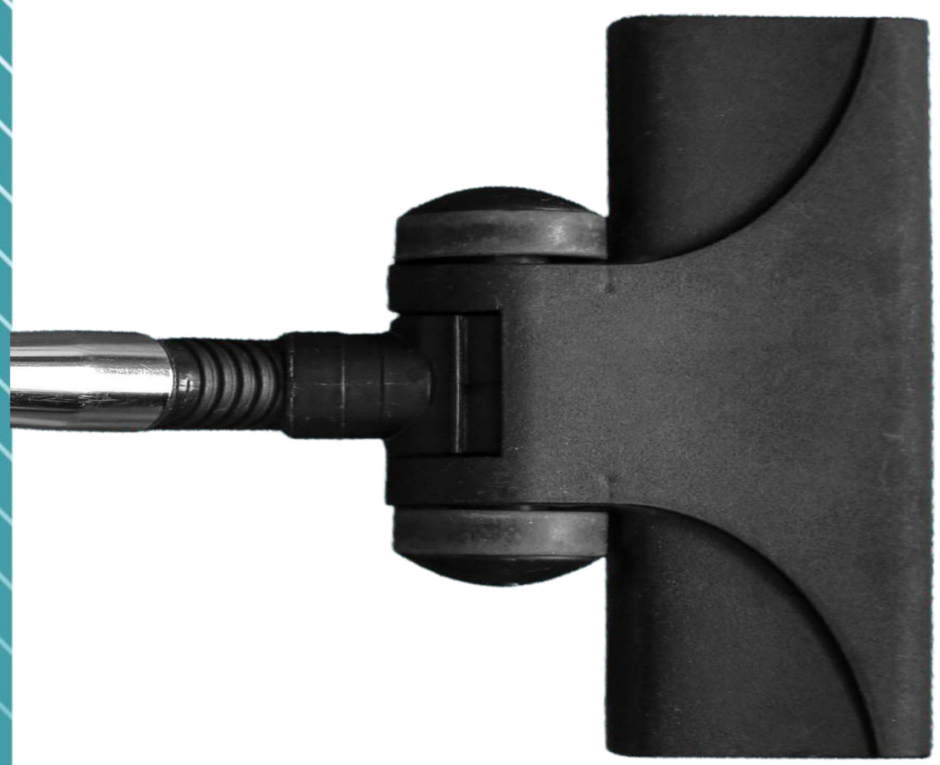
PT Heptad Data Collector, Tbk

Final Project Report - **Stage 1**

Aditya Fajri Melinianto
Apri Ansyah
Farah Fitria Sari
Oktafina Pingkan Purwanto
Pancaran Ratna Mustika
Ryan Fajar
Tiara Lailatul Nikmah

Rakamin Academy

Stage 1



Data Cleansing

Stage 1

Handle missing values

Missing values status: True

	Total Null Values	Percentage	Data Type
ChgOffDate	736465	81.905526	object
RevLineCr	4528	0.503579	object
LowDoc	2582	0.287156	object
DisbursementDate	2368	0.263356	object
MIS_Status	1997	0.222095	object
BankState	1566	0.174162	object
Bank	1559	0.173383	object
NewExist	136	0.015125	object
City	30	0.003336	object
State	14	0.001557	int64
Name	14	0.001557	object
LoanNr_ChkDgt	0	0.000000	int64
ChgOffPrinGr	0	0.000000	object
BalanceGross	0	0.000000	object
DisbursementGross	0	0.000000	object
UrbanRural	0	0.000000	int64
FranchiseCode	0	0.000000	object
RetainedJob	0	0.000000	object
GrAppv	0	0.000000	int64
NoEmp	0	0.000000	object
Term	0	0.000000	object
CreateJob	0	0.000000	int64
ApprovalDate	0	0.000000	int64
NAICS	0	0.000000	float64
Zip	0	0.000000	int64
ApprovalFY	0	0.000000	int64
SBA_Appv	0	0.000000	int64

Missing Value dihapus:

1. Kolom ChargeOffDate, Missing Value 80% maka kolom dihapus.
2. Kolom: MIS_Status memiliki xx% NaN dan akan dihapus terlebih dulu. Setelah itu, untuk kolom lain yang memiliki NaN seperti kolom Name, City, State, Bank, BankState, NewExist, RevLineCr, LowDoc, DisbursementDate, Missing Value hanya dibawah 0,5%, maka hanya dihapus baris yang mengandung NaN.

Stage 1

```
# cek duplikasi data pada semua kolom
data.duplicated().value_counts()
```

```
False    886240
Name: count, dtype: int64
```

Rakamin Academy

Stage 1

Feature Transformation

Stage 1

Feature Transformation

I. Mengubah kolom object ke numerik (DisbursementGross, BalanceGross, GrAppv, SBA_Appv) dan menghilangkan symbol [$,] pada kolom data currency ('DisbursementGross', 'BalanceGross', 'ChgOffPrinGr', 'GrAppv', 'SBA_Appv') diganti ke float untuk mempermudah analisis statistika.

	DisbursementGross	BalanceGross	ChgOffPrinGr	GrAppv	SBA_Appv
575912	$596,000.00	$0.00	$0.00	$596,000.00	$447,000.00
303046	$139,500.00	$0.00	$0.00	$139,500.00	$104,625.00

	DisbursementGross	BalanceGross	ChgOffPrinGr	GrAppv	SBA_Appv
106374	238278.0	0.0	14649.0	40000.0	20000.0
589945	750000.0	0.0	0.0	750000.0	637500.0

Stage 1

Feature Transformation

II. Menghilangkan nilai 0 pada: NAICS (memasukan ke kategori 81), Term & NoEmp (diganti dengan nilai median), NewExist & UrbanRural (diganti dengan nilai modus)

```
NAICS
81        198267
722110     27772
722211     19338
811111     14392
621210     13856
          ...
331411         1
336414         1
311351         1
316212         1
514190         1
Name: count, Length: 1311, dtype: int64
```

```
Term
84      226620
60       88507
240      84964
120      76712
300      44395
         ...
396          1
438          1
382          1
367          1
429          1
Name: count, Length: 410, dtype: int64
```

```
NoEmp
1       151454
2       136321
3        89355
4        79141
5        59520
         ...
660          1
4953         1
464          1
339          1
3713         1
Name: count, Length: 596, dtype: int64
```

```
data['NewExist'].mode()

0    1.0
Name: NewExist, dtype: float64
```

```
NewExist
1    637160
2    249080
Name: count, dtype: int64
```

```
data['UrbanRural'].mode()

0    1
Name: UrbanRural, dtype: int64
```

```
UrbanRural
1    782165
2    104075
Name: count, dtype: int64
```

Stage 1

Feature Transformation

III. Menghilangkan nilai bukan Y atau N pada kolom LowDoc dan RevLineCr (diganti dengan nilai modus (N))

```
#Merubah input LowDoc selain N dan Y dengan nilai modusnya
data['LowDoc'] = np.where((data['LowDoc'] != 'N') &
 (data['LowDoc'] != 'Y'), 'N', data.LowDoc)
data.LowDoc.value_counts()
```

```
LowDoc
N    778346
Y    107894
Name: count, dtype: int64
```

```
#Merubah input RevLineCr selain N dan Y dengan nilai modusnya
data['RevLineCr'] = np.where((data['RevLineCr'] != 'N') &
 (data['RevLineCr'] != 'Y'), 'N', data.RevLineCr)
data.RevLineCr.value_counts()
```

```
RevLineCr
N    687973
Y    198267
Name: count, dtype: int64
```

Stage 1

Feature Transformation

IV. Menghilangkan 3 karakter terakhir pada kolom NAICS

```
#Menangani kolom NAICS, kita akan merubahnya menjadi nama industrinya
#Berdasarkan guideline, dua  digit di awal adalah kode industrinya
naics_code = data['NAICS']

#Fungsi untuk mengambil ambil 2 digit awal dari kodenya
def get_code(naics_code):
    if naics_code <= 0:
        return 0
    return (naics_code // 10 ** (int(math.log(naics_code, 10)) - 1))
```

```
#Menerapkan fungsi yang dibuat ke data NAICS
data['NAICS'] = data.NAICS.apply(get_code)
```

```
data['NAICS'].value_counts()
```

```
NAICS
81    270021
44     83867
72     67084
54     66951
23     65635
62     54633
42     48148
45     41895
33     37740
56     32114
48     19955
32     17709
71     14460
53     13457
31     11660
51     11220
52      9378
11      8868
61      6313
49      2180
21      1820
22       654
55       256
92       222
Name: count, dtype: int64
```

Stage 1

Feature Transformation

V. Mengganti 0 dan 1 pada kolom FranchiseCode menjadi 'Not-Franchise' dan selain itu menjadi 'Franchise'.

```
#jika kolom FranchiseCode = 0 atau = 1 maka dia tidak ada frnachise, selain itu maka dia ada franchise
data['FranchiseCode'] = np.where((data.FranchiseCode != 0 ) & (data.FranchiseCode != 1 ),'Franchise',data.FranchiseCode)
data['FranchiseCode'] = data['FranchiseCode'].replace('0', 'Not-Franchise')
data['FranchiseCode'] = data['FranchiseCode'].replace('1', 'Not-Franchise')
data.FranchiseCode.value_counts()
```

```
FranchiseCode
Not-Franchise    835037
Franchise         51203
Name: count, dtype: int64
```

Stage 1

Feature Transformation

VI. Nilai 1976A pada kolom ApprovalFY diganti menjadi 1976.

VII. NoEmp, CreateJob, RetainedJob dilakukan robust scaler, untuk mempermudah dalam melakukan analisis statistic.

Output (VII)

	NoEmp	CreateJob	RetainedJob
count	886240.000000	886240.000000	886240.000000
mean	0.931244	8.463092	2.460602
std	9.273034	237.301746	59.434887
min	-0.375000	0.000000	-0.250000
25%	-0.250000	0.000000	-0.250000
50%	0.000000	0.000000	0.000000
75%	0.750000	1.000000	0.750000
max	1249.375000	8800.000000	2374.750000

Stage 1

Feature Transformation

VIII. Untuk Bank dengan count < 1500 akan dimasukkan ke kategori ‘Others’.

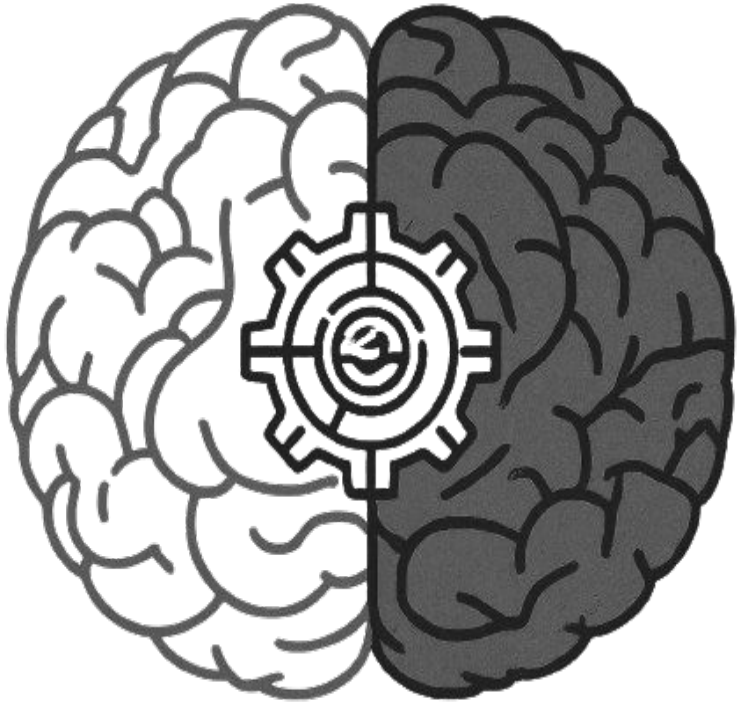
```
bank_counts
```

```
Bank
BANK OF AMERICA NATL ASSOC          86075
WELLS FARGO BANK NATL ASSOC         62934
JPMORGAN CHASE BANK NATL ASSOC      47460
U.S. BANK NATIONAL ASSOCIATION      34752
CITIZENS BANK NATL ASSOC            33569
                                      ...
AMER BK & TR WISCONSIN                  1
BANK OF IDAHO HOLDING COMPANY           1
APPLE CREEK BK. CO                      1
HERITAGE BK E. BAY A DIVISION           1
DEPCO                                   1
Name: count, Length: 5788, dtype: int64
```

```
Bank
Others                                317633
BANK OF AMERICA NATL ASSOC             86075
WELLS FARGO BANK NATL ASSOC            62934
JPMORGAN CHASE BANK NATL ASSOC         47460
U.S. BANK NATIONAL ASSOCIATION         34752
CITIZENS BANK NATL ASSOC               33569
PNC BANK, NATIONAL ASSOCIATION         27148
BBCN BANK                              22814
CAPITAL ONE NATL ASSOC                 22220
MANUFACTURERS & TRADERS TR CO          11150
READYCAP LENDING, LLC                  10616
THE HUNTINGTON NATIONAL BANK            9520
KEYBANK NATIONAL ASSOCIATION            9186
TD BANK, NATIONAL ASSOCIATION           8901
BRANCH BK. & TR CO                      8028
ZIONS FIRST NATIONAL BANK               7897
CALIFORNIA BANK & TRUST                 7476
CITIBANK, N.A.                          7402
REGIONS BANK                            7143
BANCO POPULAR NORTH AMERICA             7135
COMERICA BANK                           6991
BANK OF THE WEST                        6628
COMPASS BANK                            6384
BUSINESS LOAN CENTER, LLC               6262
GE CAP. SMALL BUS. FINAN CORP           6184
UMPQUA BANK                             6000
BMO HARRIS BK NATL ASSOC                5154
FIFTH THIRD BANK                        5144
HSBC BK USA NATL ASSOC                  4771
ASSOCIATED BANK NATL ASSOC              4625
SUPERIOR FINANCIAL GROUP, LLC           4272
```

Stage 1

Feature Encoding



I. Pengkategorian Term, dibuat kolom baru RealEstate dengan nilai >= 240 =1, dibawah 240=0
II. NAICS diubah menjadi nama-nama industri setiap kategori -> dilabelin menjadi angka numerik
III. MIS_status dilabelin chargeoff = 1, PIF=0
IV. Encoding 60 bank dengan count terbanyak

Stage 1

Feature Encoding

Input (II)

```
#Merubah 2 digit menjadi nama sektor
def industri(i):
    def_code = {11:'Agriculture, Forestry, Fishing & Hunting', 21:'Mining, Quarying, Oil & Gas',
                22:'Utilities', 23:'Constuction', 31:'Manufacturing', 32:'Manufacturing', 33:'Manufacturing',
                42:'Wholesale Trade', 44:'Retail Trade', 45:'Retail Trade', 48:'Transportation & Warehousing',
                49:'Transportation & Warehousing', 51:'Information', 52:'Finance & Insurance',
                53:'Real Estate, Rental & Leasing', 54:'Professional, Scientific & Technical Service',
                55:'Management of Companies & Enterprise',
                56:'Administrative, Support, Waste Management & Remediation Service',
                61:'Educational Service', 62:'Health Care & Social Assistance',
                71:'Arts, Entertainment & Recreation', 72:'Accomodation & Food Service',
                81:'Other Servieces (Ex: Public Administration)', 92:'Public Administration'
               }
    if i in def_code:
        return def_code[i]

df['Industri'] = df.ind_code.apply(industri)
df['Industri'].value_counts()
```

Stage 1

Feature Encoding

Output (II)

```
Industri
Other Servieces (Ex: Public Administration)                    270021
Retail Trade                                                   125762
Manufacturing                                                   67109
Accomodation & Food Service                                     67084
Professional, Scientific & Technical Service                    66951
Constuction                                                     65635
Health Care & Social Assistance                                 54633
Wholesale Trade                                                 48148
Administrative, Support, Waste Management & Remediation Service 32114
Transportation & Warehousing                                    22135
Arts, Entertainment & Recreation                                14460
Real Estate, Rental & Leasing                                   13457
Information                                                     11220
Finance & Insurance                                              9378
Agriculture, Forestry, Fishing & Hunting                         8868
Educational Service                                              6313
Mining, Quarying, Oil & Gas                                      1820
Utilities                                                         654
Management of Companies & Enterprise                              256
Public Administration                                             222
Name: count, dtype: int64
```

Hasil label encoding adalah sebagai berikut

```
df['Industri'].value_counts()
```

```
Industri
12    270021
16    125762
10     67109
0      67084
13     66951
4      65635
7      54633
19     48148
1      32114
17     22135
3      14460
15     13457
8      11220
6       9378
2       8868
5       6313
11      1820
18       654
9        256
14       222
Name: count, dtype: int64
```

Stage 1

Feature Encoding

(III)

```
# Memberi label pada kolom MIS_Status
data['MIS_Status'] = data['MIS_Status'].replace({'P I F': 0, 'CHGOFF':1}).astype(int)
data.MIS_Status.value_counts()
```

```
MIS_Status
0    730199
1    156041
Name: count, dtype: int64
```

(IV)

```
[ ] bank_counts = data['Bank'].nunique()
    bank_counts
```

```
60
```

```
[ ] bank_encode = ['Bank']
```

```
# Encode label pada kolom 'Bank'
data[bank_encode] =data[bank_encode].apply(LabelEncoder().fit_transform)
```

```
[ ] # Lihat data hasil encode
    data['Bank'].value_counts()
```

```
Bank
39    317633
3      86075
57     62934
32     47460
52     34752
13     33569
41     27148
5      22814
10     22220
35     11150
43     10616
50      9520
33      9186
49      8901
7       8028
59      7897
9       7476
12      7402
44      7143
1       7135
15      6991
```

Stage 1

Rakamin Academy

Handle Outlier

1. Fitur yang sudah di-robust scaler akan menggunakan metode Z Score.

Data setelah menghapus outliers dengan Z Score pada kolom yang dilakukan robust scaler:

	NoEmp	CreateJob	RetainedJob
count	884139.000000	884139.000000	884139.000000
mean	0.748505	1.987352	0.814611
std	2.152710	8.468232	2.828473
min	-0.375000	0.000000	-0.250000
25%	-0.250000	0.000000	-0.250000
50%	0.000000	0.000000	0.000000
75%	0.750000	1.000000	0.750000
max	28.750000	600.000000	149.750000

Rakamin Academy

Handle Outlier

2. Term menggunakan metode IQR

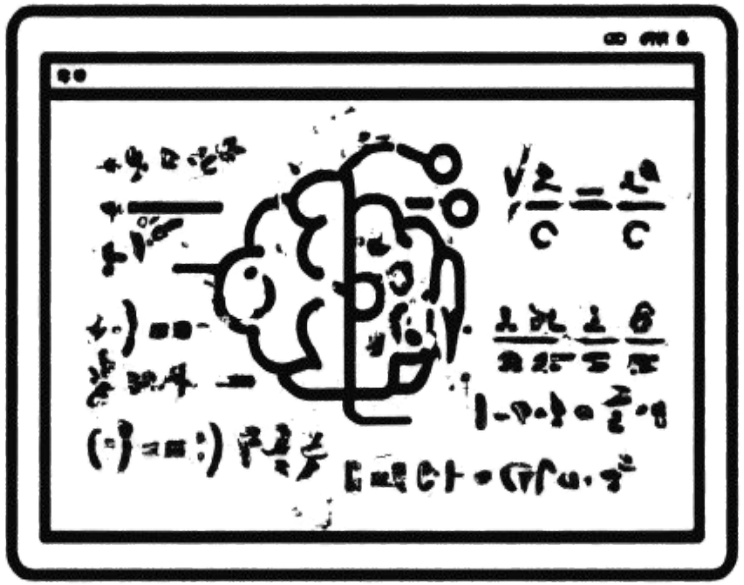
```
Q1 = df['Term'].quantile(0.25)
Q3 = df['Term'].quantile(0.75)
IQR = Q3 - Q1

df = df[~((df['Term'] < (Q1-1.5*IQR)) | (df['Term'] > (Q3+1.5*IQR)))]
```

```
print(df['Term'].describe())
```

```
count    727955.000000
mean         78.700653
std          38.740909
min           1.000000
25%          60.000000
50%          84.000000
75%          84.000000
max         210.000000
Name: Term, dtype: float64
```

Stage 1

Class Imbalance



- Class Imbalance pada MIS_Status di-handle dengan oversampling -> SMOTENC.
- Setelah dilakukan CI, dilakukan pengecekan terhadap duplikat dan missing values.

Class Imbalance

(I) Class Imbalance pada MIS_Status di-handle dengan oversampling SMOTE

```
sm = SMOTE(random_state=42)

target = df['MIS_Status']
oversampler = RandomOverSampler(sampling_strategy="minority")
oversampled_data, oversampled_target = oversampler.fit_resample(df, target)
```

(II) Jumlah data setelah dilakukan oversampling

```
oversampled_target.value_counts()
```

```
MIS_Status
0    576105
1    576105
Name: count, dtype: int64
```

(III) Melakukan pengecekan terhadap data duplikasi setelah dilakukannya oversampling

```
[ ] oversampled_data.duplicated().value_counts()
```

```
False    727955
True     424255
Name: count, dtype: int64
```

(IV) Menghilangkan duplikasi

```
[ ] oversampled_data_no_dup = oversampled_data.drop_duplicates()
```

```
[ ] oversampled_data_no_dup.duplicated().value_counts()
```

```
False    727955
Name: count, dtype: int64
```

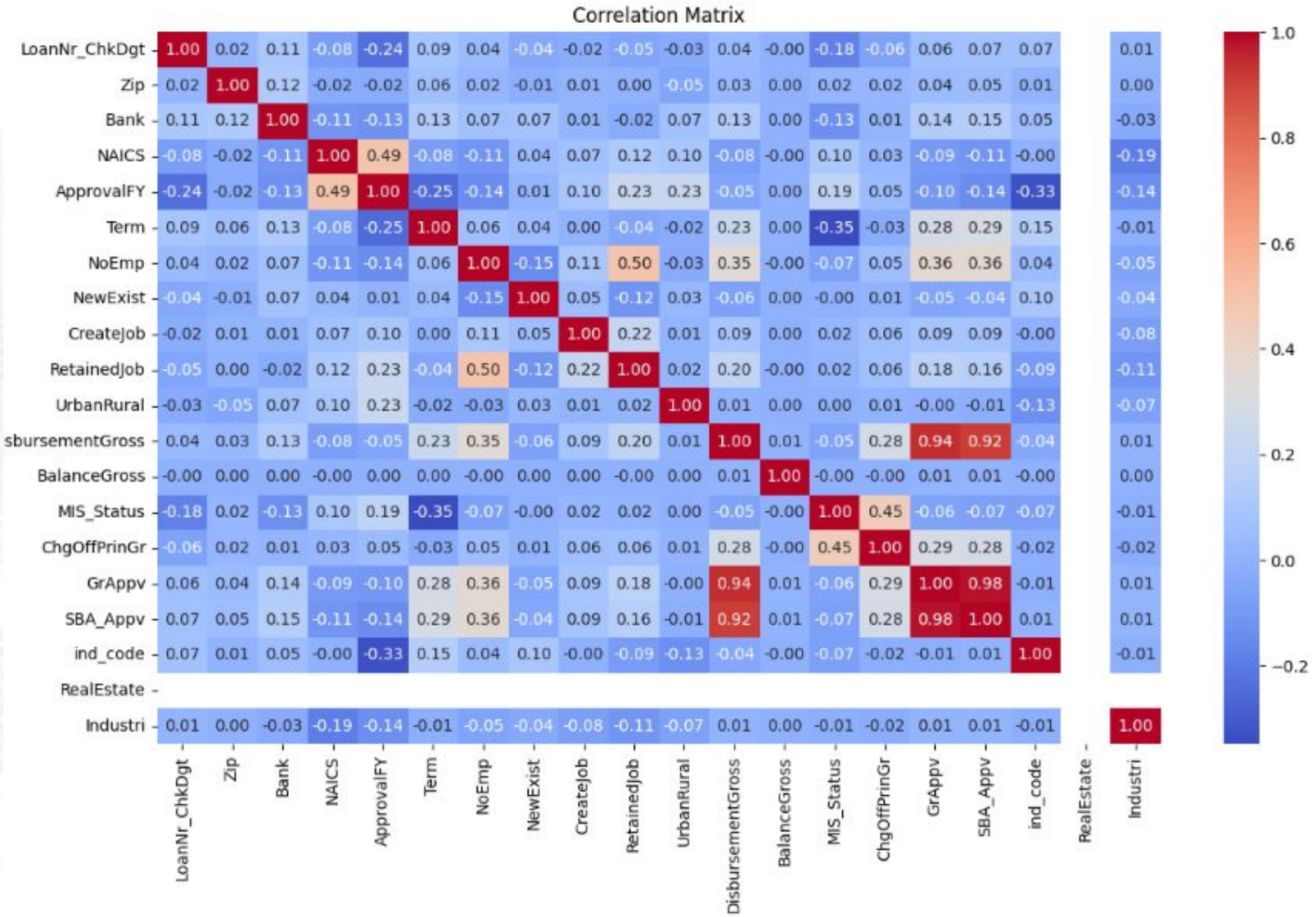
Stage 1

Feature Engineering

- Feature Selection
- Feature Extraction
- Additional Feature

Stage 1

Feature Selection



- Feature yang dihapus: ApprovalDate, DisbursementDate, ChgOffDate, Zip, City, LoanNr_ChkDgt, Name, SBA_Appv, GrAppv, ChgOffPrinGr, DisbursementGross, BalanceGros, ApprovalFY
- Feature yang stay: Term, NoEmp, CreateJob, RetainedJob, NewExist, UrbanRural, NAICS, FranchiseCode, LowDoc, RevLineCr, MIS_Status, Bank, State, BankState.

Stage 1

Feature Extraction

i. BankIsIn -> BankState = State, maka 1, jika tidak 0

ii. CompanyType -> Jika NewExist dan UrbanRural = 1, maka 1. Jika NewExist = 1 tapi UrbanRural = 2, maka 2. Jika NewExist = 2 tapi UrbanRural = 1, maka 3. Jika NewExist dan UrbanRural = 2, maka 4.

iii. Prod -> CreatedJob > RetainedJob maka 1, jika tidak 0

iv. Membuat kolom baru Recession = tahun terjadinya resesi di USA

```
 Data setelah feature selection:
<class 'pandas.core.frame.DataFrame'>
Index: 727955 entries, 0 to 899163
Data columns (total 18 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   State          727955 non-null  object
 1   Bank           727955 non-null  int32
 2   BankState      727955 non-null  object
 3   Term           727955 non-null  int64
 4   NoEmp          727955 non-null  float64
 5   NewExist       727955 non-null  int32
 6   CreateJob      727955 non-null  float64
 7   RetainedJob    727955 non-null  float64
 8   FranchiseCode  727955 non-null  object
 9   UrbanRural     727955 non-null  int32
 10  RevLineCr      727955 non-null  object
 11  LowDoc         727955 non-null  object
 12  MIS_Status     727955 non-null  int32
 13  Industri       727955 non-null  int32
 14  BankIsIn       727955 non-null  int64
 15  CompanyType    727955 non-null  int64
 16  Prod           727955 non-null  int64
 17  Recession      727955 non-null  int64
dtypes: float64(3), int32(5), int64(5), object(5)
memory usage: 91.6+ MB
None
```

Additional Feature

i. Pengkategorian bulan/hari Approval Date
ii. GDP/state
iii. Total export/import company
iv. Revenue company

Stage 1

PT Heptad Data Collector, Tbk

Link GitHub :

PT Heptad Data Collector, Tbk (PT. HDC)

Rakamin Academy

Stage 1

Exploratory Data Analysis (EDA)

Stage 1

PT Heptad Data Collector, Tbk

Fitur yang digunakan :

1. Term : Merupakan jumlah angsuran yang diberikan kepada peminjam (dalam bentuk bulan)
2. NoEmp : Jumlah karyawan yang terdapat pada UMKM peminjam
3. CreateJob : Jumlah pekerjaan yang tercipta dari UMKM tersebut
4. RetainedJob : Jumlah pekerjaan yang berhasil dipertahankan dari UMKM tersebut
5. NewExist : Mengklasifikasi apakah UMKM tersebut termasuk baru atau lama
6. UrbanRural : Letak dari UMKM tersebut apakah berada pada pedesaan / perkotaan
7. NAICS / Industri : Kode klasifikasi Industri yang ditetapkan oleh Amerika Utara
8. FranchiseCode : Apakah UMKM atau usaha tersebut termasuk franchise atau tidak
9. LowDoc : Apakah pinjaman yang diajukan tersebut support low doc
10. RevLineCr : Status Jalur kredit bergulir, Y = ya / N = tidak
11. MIS_Status : Kolom target yang menyatakan lunas atau gagal bayar
12. Bank : Nama bank yang mengeluarkan pinjaman
13. State : Negara bagian peminjam
14. BankState : Negara bagian bank yang mengeluarkan pinjaman

Stage 1

PT Heptad Data Collector, Tbk

Feature tambahan, semuanya kategorikal:

i. BankIsIn -> BankState = State, maka 1, jika tidak 0

ii. CompanyType -> Jika NewExist dan UrbanRural = 1, maka 1. Jika NewExist = 1 tapi UrbanRural = 2, maka 2. Jika NewExist = 2 tapi UrbanRural = 1, maka 3. Jika NewExist dan UrbanRural = 2, maka 4.

iii. Prod -> CreatedJob > RetainedJob maka 1, jika tidak 0

iv. Membuat kolom baru Recession = tahun terjadinya resesi di USA

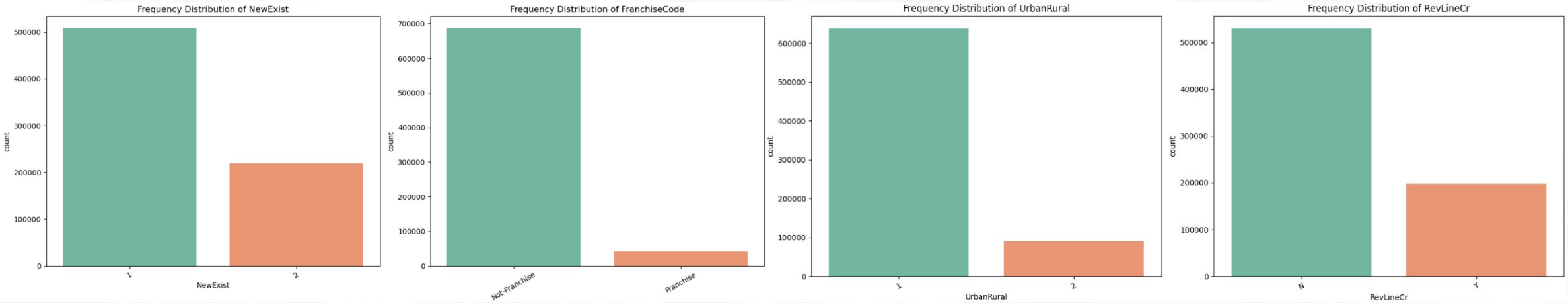
Rakamin Academy

Stage 1

Insight & Visualization

Univariate Analysis - Categorical

NewExist, FranchiseCode, UrbanRural, RevLineCr



NewExist: Jumlah bisnis yang menerima pinjaman dalam dataset didominasi oleh bisnis yang sudah existing (label 1) dibandingkan dengan bisnis baru (label 2). Kemungkinan penjelasan untuk hal ini adalah: bisnis yang sudah ada memiliki rekam jejak yang lebih panjang dan stabil, sehingga lebih mudah untuk mendapatkan pinjaman.

FranchiseCode: Sebagian besar bisnis yang menerima pinjaman adalah bisnis independen atau bukan bagian dari franchise. Ini bisa mencerminkan kenyataan bahwa bisnis independen mungkin lebih membutuhkan dukungan finansial dibandingkan dengan bisnis yang merupakan bagian dari sistem franchise

UrbanRural: Data menunjukkan bahwa sebagian besar pinjaman diberikan kepada bisnis di daerah perkotaan. Ini mungkin disebabkan oleh beberapa faktor:

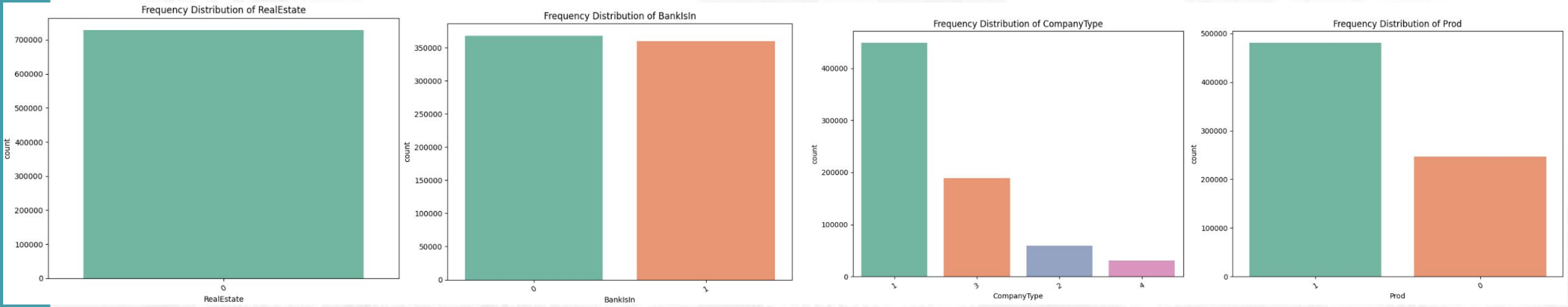
a. Konsentrasi bisnis yang lebih tinggi di daerah perkotaan.
b. Akses yang lebih mudah ke layanan perbankan dan keuangan di daerah perkotaan.
c. Potensi pasar yang lebih besar di daerah perkotaan.

Meskipun jumlahnya lebih sedikit, ada sejumlah signifikan pinjaman yang diberikan di daerah pedesaan. Ini menunjukkan bahwa ada kebutuhan akan dukungan finansial di daerah pedesaan, meskipun aksesnya mungkin lebih terbatas dibandingkan dengan daerah perkotaan.

RevLineCr: Data menunjukkan bahwa sebagian besar pinjaman diberikan tanpa revolving line of credit. Ini mungkin disebabkan oleh beberapa faktor: Bisnis mungkin lebih memilih pinjaman konvensional dengan struktur pembayaran tetap daripada revolving line of credit yang lebih fleksibel namun mungkin lebih kompleks. Kebijakan pemberian pinjaman dari lembaga keuangan yang mungkin lebih ketat untuk revolving line of credit.

Univariate Analysis - Categorical

RealEstate, BankIsIn, CompanyType, Prod



RealEstate: Setelah kolom Term dilakukan handle outlier dengan metode IQR, nilai Term pada dataset ada di bawah 240.

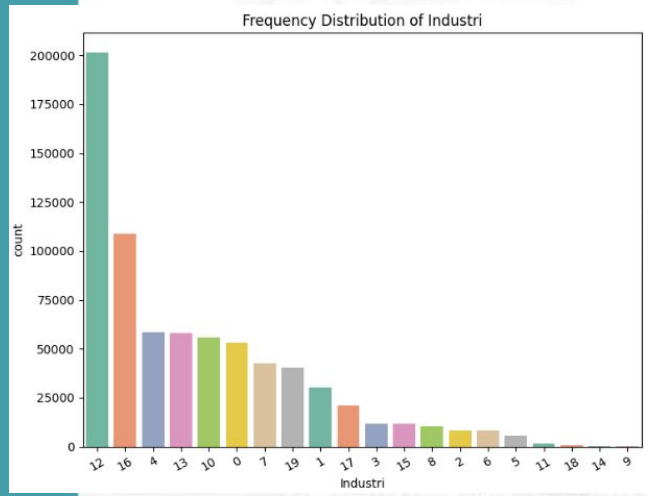
BankIsIn: Jumlah pinjaman, dimana bisnisnya berada pada state yang sama dengan Bank pemberi pinjamnannya terlihat tidak berbeda signifikan.

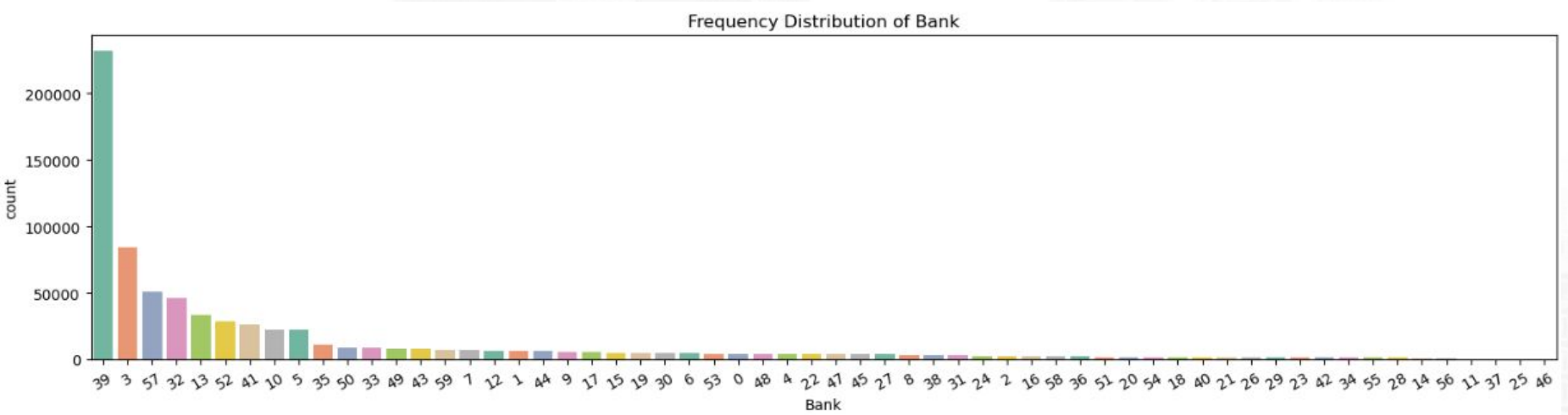
CompanyType: Jumlah bisnis yang sudah ada dan berada di perkotaan paling tinggi, diikuti oleh bisnis baru dan ada di perkotaan. Sisanya adalah bisnis yang ada di pedesaan, baik yang sudah ada maupun baru.

Prod: Sebagian besar data menunjukkan bahwa jumlah pekerjaan yang diciptakan (CreateJob) lebih besar daripada pekerjaan yang dipertahankan (RetainedJob). Ini bisa mengindikasikan bahwa banyak bisnis yang menerima pinjaman berhasil menciptakan lebih banyak lapangan pekerjaan baru daripada mempertahankan pekerjaan yang ada..

Univariate Analysis - Categorical

Industri, Bank



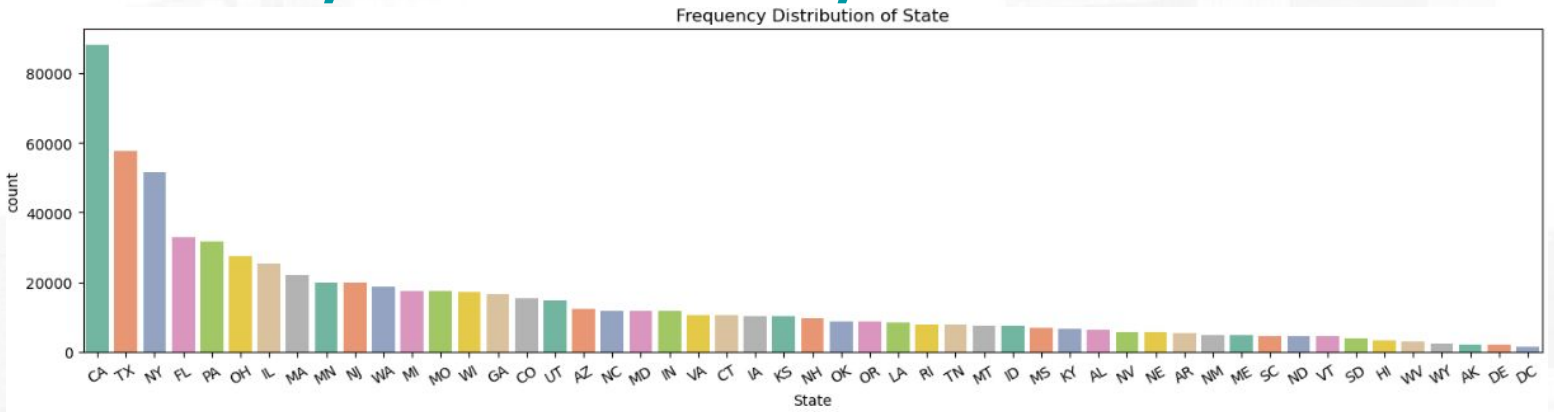


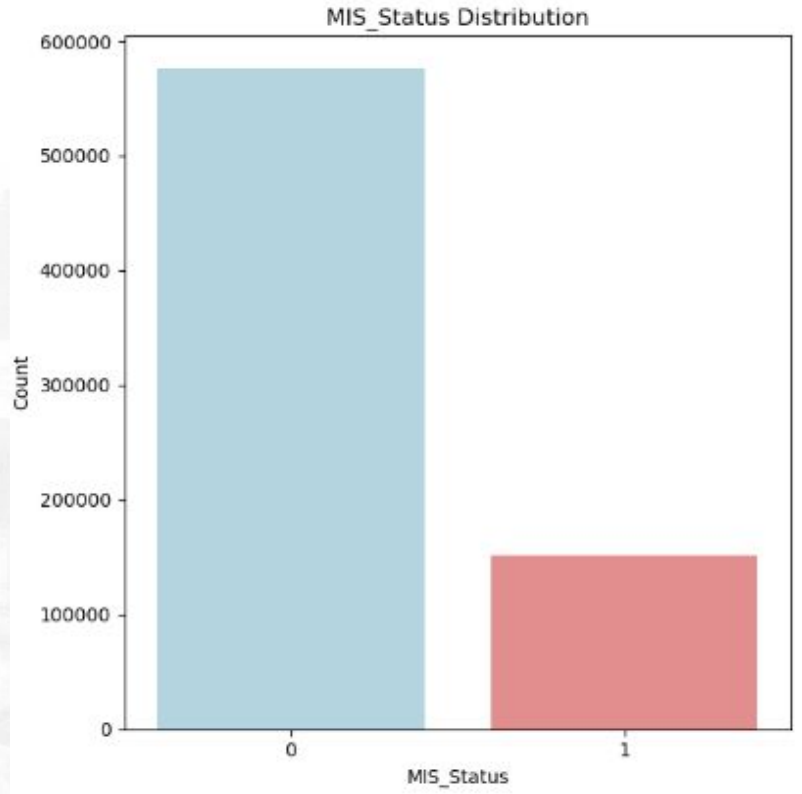
Industri: Frekuensi distribusi industri tertinggi di bidang Other Services karena banyak yang nilai awalnya 0 dimasukkan ke kategori ini. Nilai awal 0 menandakan perusahaan tidak diberikan label NAICS, yaitu tipikal pinjaman sebelum NAICS berdiri tahun 1997. Setelah itu, 3 sektor teratas adalah Retail, Manufacturing dan Accomodation & Food Services, mencerminkan pentingnya ketiga industri ini dalam perekonomian. Selanjutnya, sektor construction, healthcare dan social assistance juga menunjukan aktivitas yang signifikan dalam memperoleh pinjaman. Sektor lainnya: Wholesale Trade, Administrative Support, Transportation & Warehousing, dan sektor lainnya menunjukkan aktivitas yang lebih rendah tetapi tetap signifikan.

Bank: Kategori di kolom ini yang paling banyak kuantitasnya adalah Others. Ini karena bank yang hanya muncul < 1500 dimasukkan ke dalam kategori ini dan bank dengan karakteristik seperti ini ternyata banyak. Bank yang paling dipakai untuk meminjam ke SBA adalah Bank of America, Wells Fargo, JP Morgan, US Bank National of Association, dan Citizens Bank National Association.

Univariate Analysis - Categorical

State, BankState, MIS_Status





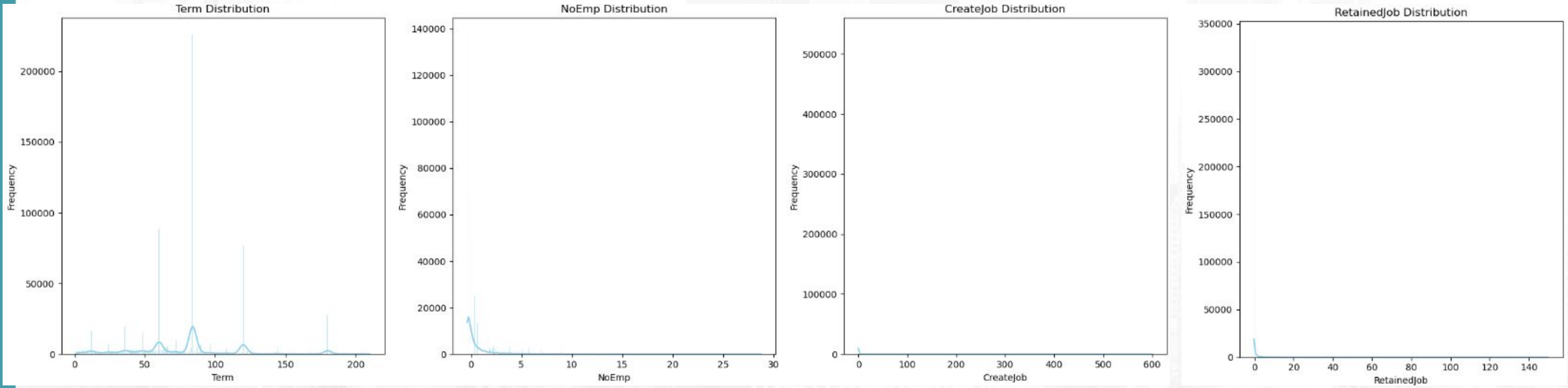
State: Dari total 50 state di USA, peminjam paling banyak ada pada state California, Texas, New York, Florida, dan Philadelphia. Untuk fokus implementasi program improvement kepada customer, SBA bisa fokus kepada customer di 5 state ini.

BankState: Dari total 50 state di USA, bank peminjam paling banyak ada pada state California, North Carolina, Illinois, Ohio, dan Rhode Island. Untuk fokus implementasi program improvement kepada bank, SBA bisa fokus kepada bank ke 5 state ini.

MIS_Status: Setelah dilakukan class imbalance, persentase rasio CHGOFF dan PIF tidak berubah, yakni masing-masing masih sekitar 18% dan 72% dari total dataset.

Univariate Analysis - Numerical

Term, NoEmp, CreateJob, RetainedJob



Term: Setelah di-handle outlier dengan metode IQR, limit atas Term berubah menjadi kurang dari 240. Ini tercermin di kolom RealEstate yang menandakan bahwa tidak ada Term di atas 240. Untuk distribusi dari Term sendiri bisa dilihat menyerupai distribusi normal.

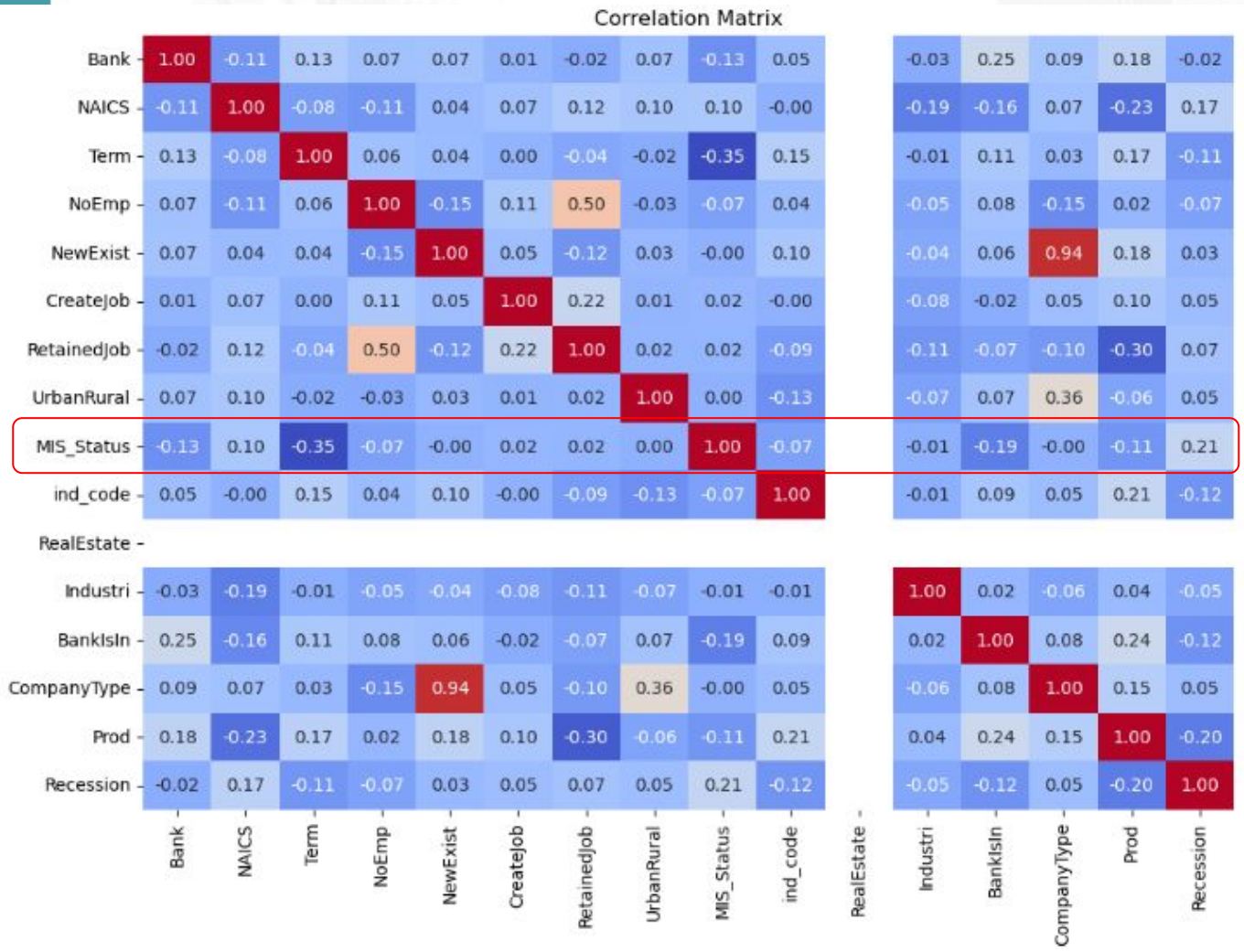
NoEmp: Setelah ditransform dengan robust scaler karena distribusi awal kolom ini positive skew dan di-handle outlier dengan metode Z-score, kolom ini masih memiliki distribusi yang kurang lebih positive skew. Hanya saja ada perubahan limit maksimumnya yang menjadi hanya 29. Perusahaan peminjam memiliki karyawan yang

CreateJob: Setelah ditransform dengan robust scaler karena distribusi awal kolom ini positive skew dan di-handle outlier dengan metode Z-score, kolom ini masih memiliki distribusi yang kurang lebih positive skew. Hanya saja ada perubahan limit maksimumnya yang menjadi hanya 600.

RetainedJob: Setelah ditransform dengan robust scaler karena distribusi awal kolom ini positive skew dan di-handle outlier dengan metode Z-score, kolom ini masih memiliki distribusi yang kurang lebih positive skew. Hanya saja ada perubahan limit maksimumnya yang menjadi hanya 150.

Multivariate Analysis - Numerical

Heatmap



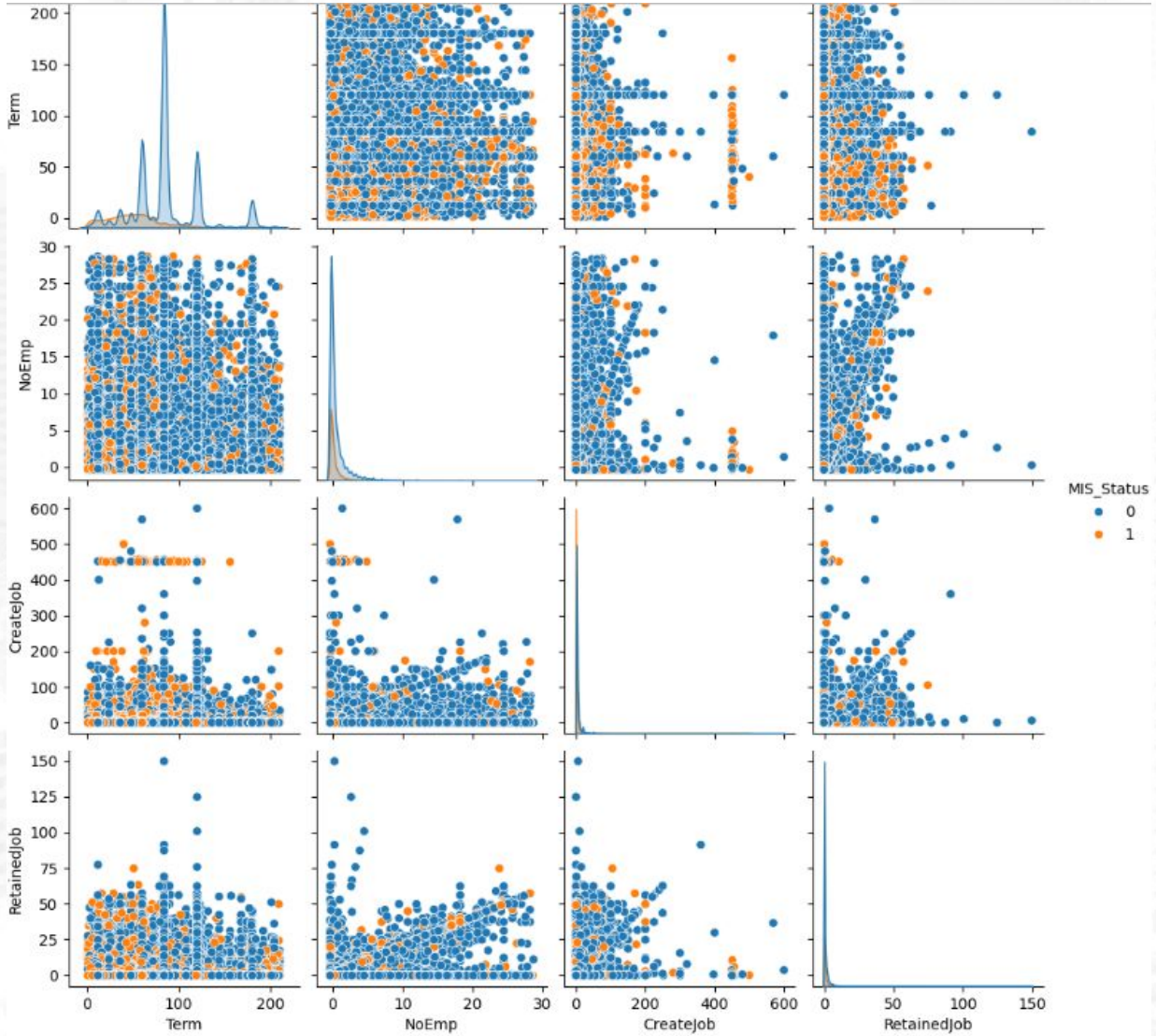
Pada matriks korelasi dapat dilihat bahwa urutan korelasi absolut paling besar dari 14 kolom lain yang bisa dikorelasi (1 kolom, RealEstate tidak bisa dikorelasi) jika diurutkan adalah:

1. Term (0.35)
2. Recession (0.21)
3. BankIsin (0.19)
4. Bank (0.13)
5. Prod (0.11)
6. NAICS (0.10)
7. NoEmp (0.07)
8. Ind_code (0.07)
9. CreateJob (0.02)
10. RetainedJob (0.02)
11. Industri (0.01)
12. NewExist (0)
13. UrbanRural (0)
14. CompanyType (0)

Dengan ini dapat dilihat bahwa 5 faktor teratas yang sangat berkaitan dengan MIS_Status adalah jangka waktu kredit, tahun resesi, apakah letak bank sama dengan letak peminjam, bank dari peminjam, dan apakah pembukaan pekerjaan lebih tinggi dibanding pekerjaan yang di-retain.

Multivariate Analysis - Numerical

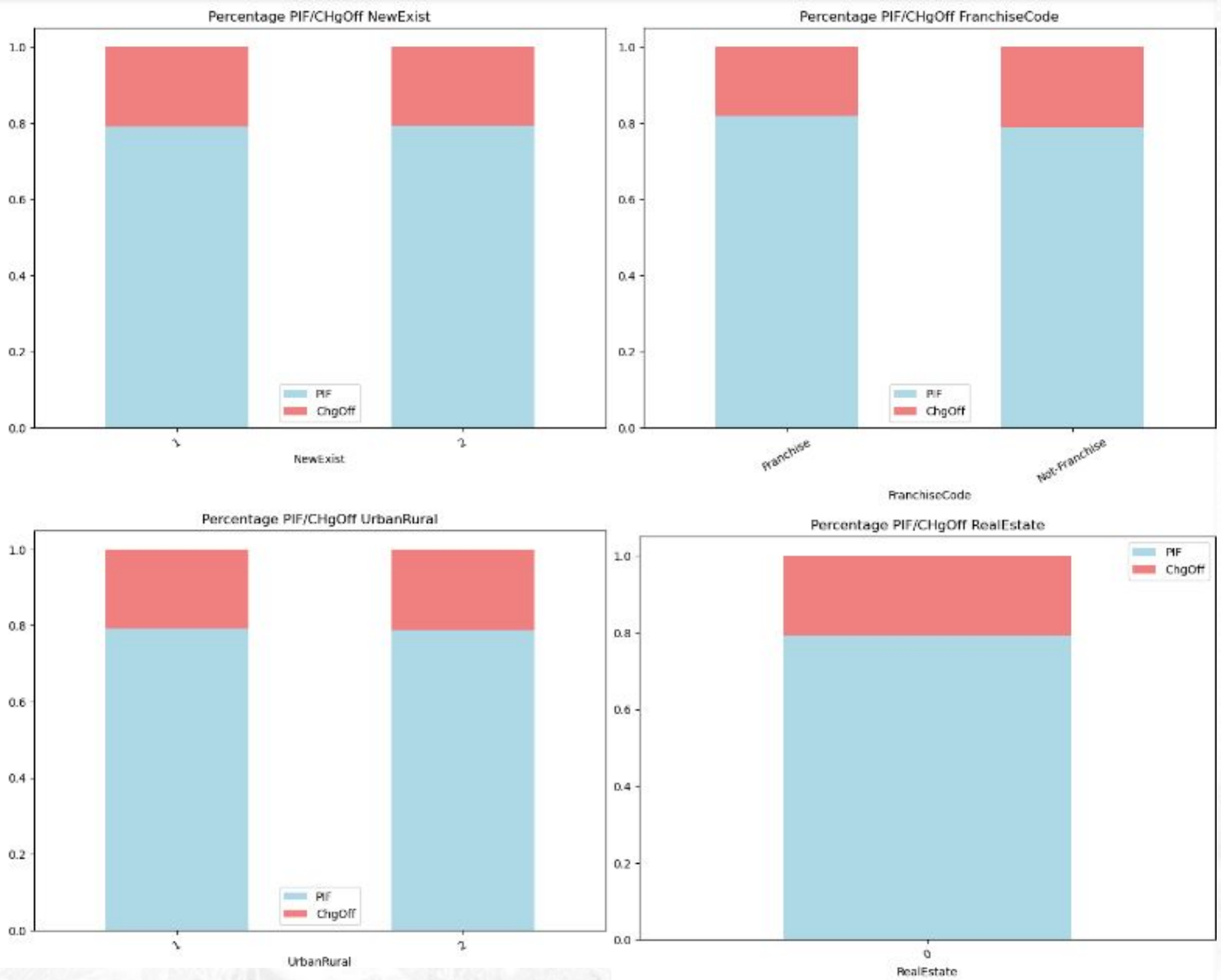
Pairplot



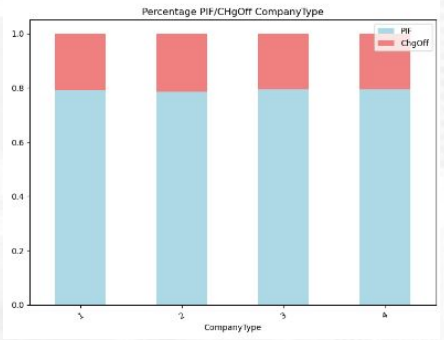
Pada pairplot untuk kolom numerikal, dapat dilihat bahwa rata-rata korelasi antara kolom numerikal tidak terlalu terlihat signifikan. Scatterplot terlihat random berserak. Untuk distribusi kategori MIS_Status pada kolom Term bisa terlihat bahwa rata-rata pinjaman gagal bayar ada di distribusi rendah, begitu pula dengan NoEmp, CreateJob, dan RetainedJob. SBA bisa mengacu pada grafik ini bahwa untuk menghindari pinjaman dengan risiko tinggi. SBA harus menghindari term yang rendah, peminjam dengan karyawan dan pekerjaan yang sedikit, serta peminjam yang sedikit membuka lapangan pekerjaan baru.

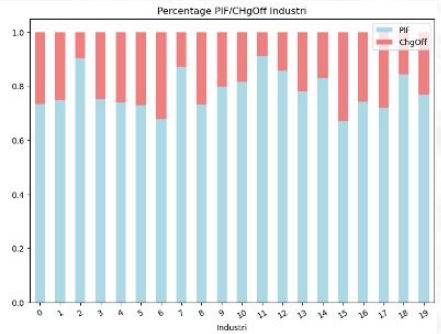
Multivariate Analysis - Categorical

Barchart with Hue - Low Correlation



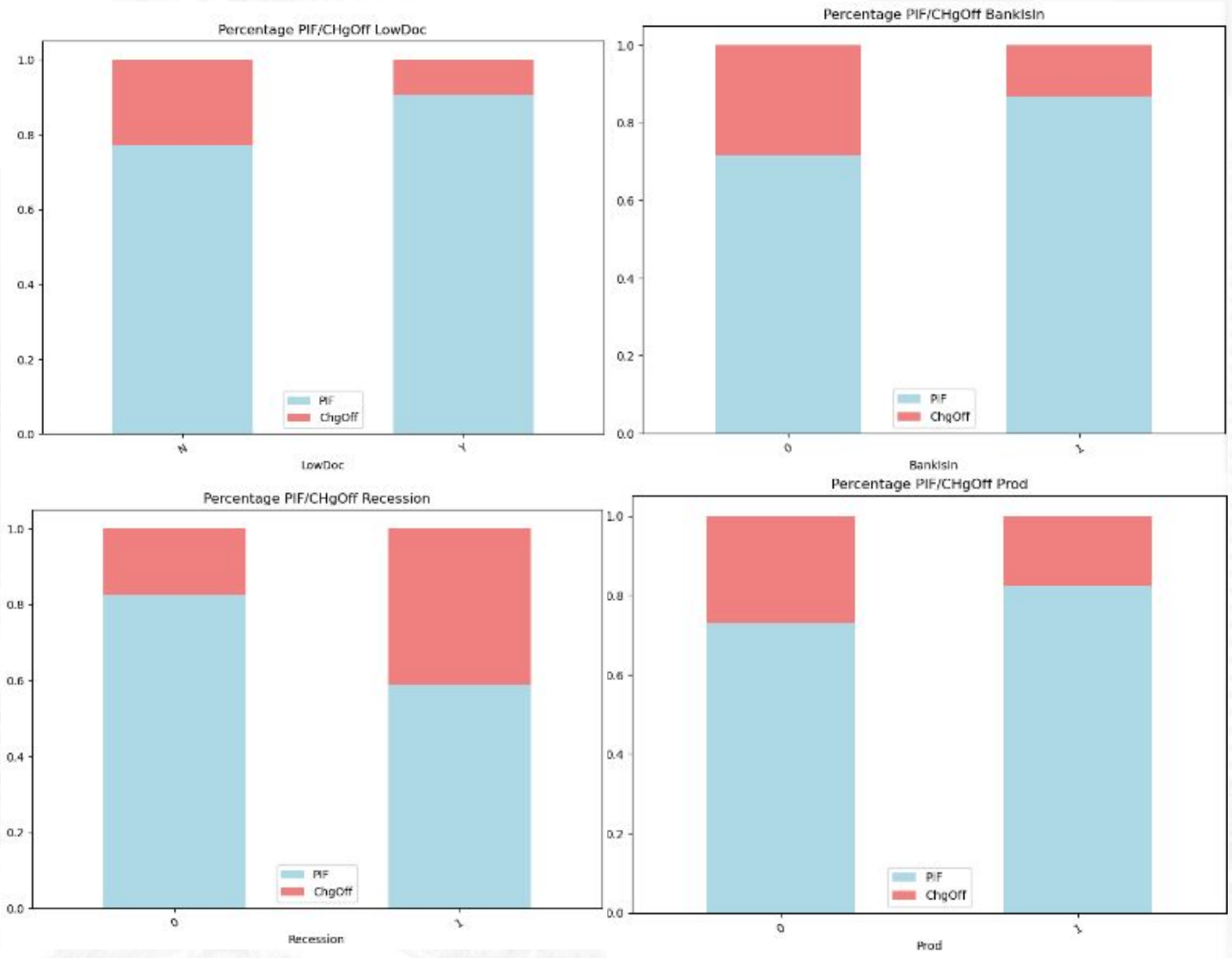
Tidak ada perbedaan rasio antara peminjam gagal bayar dengan tidak gagal bayar pada kolom NewExist, FranchiseCode, UrbanRural, CompanyType, Industri, dan RealEstate. Hal ini sejalan dengan korelasi kolom ini yang rendah terhadap kolom MIS_Status. SBA tidak disarankan melihat ke fitur-fitur ini untuk mengecek MIS_Status.



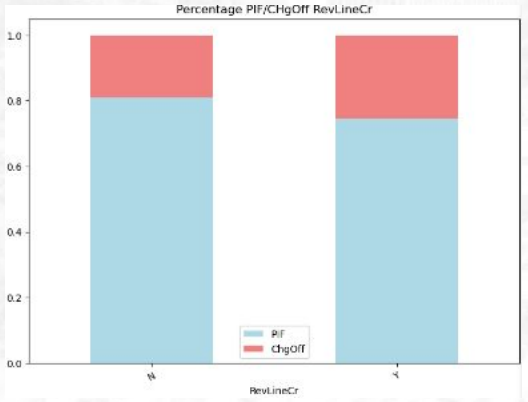


Multivariate Analysis - Categorical

Barchart with Hue - Intermediate Correlation

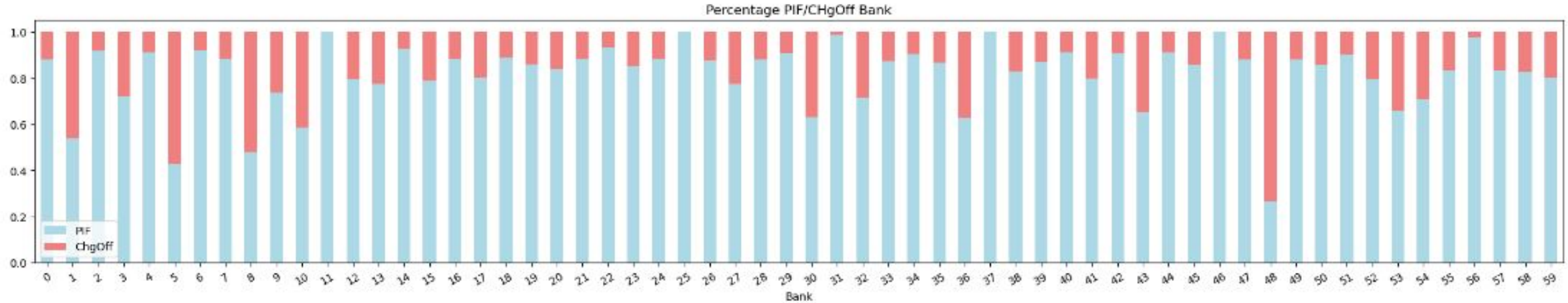


Ada perbedaan rasio yang cukup signifikan antara peminjam gagal bayar dengan tidak gagal bayar pada kolom BandIsIn, Recession, dan Prod. Hal ini sejalan dengan korelasi kolom ini yang tinggi terhadap kolom MIS_Status. Begitu juga pada kolom kategorikal LowDoc dan RevLineCr. Maka dari itu SBA bisa melihat ke kriteria-kriteria di mana CHGOFF rendah pada kolom-kolom ini: peminjam melakukan LowDoc yang berarti memiliki kredit skor yang bagus dan income stabil, peminjam memakai bank yang sama dengan state dia berada, peminjam membuka job lebih banyak, tidak memiliki kredit bergulir dan tahun tersebut USA tidak mengalami resesi.



Multivariate Analysis - Categorical

Barchart with Hue - Intermediate Correlation



Seperti yang sudah diperlihatkan di heatmap, korelasi bank terhadap MIS_Status termasuk yang paling besar. Ada beberapa bank yang bahkan tidak memiliki kadar kredit gagal bayar seperti pada bank SBA - EDF Enforcement Action, Florida Business Development Action, dan CDC Small Business Finance Corporation. Disarankan untuk SBA agar jika ingin menyalurkan pinjaman dengan risiko gagal bayar rendah bisa ke bank bebas CHGOFF atau rendah CHGOFF.