

Smart Credit Risk Modelling

Leveraging Machine Learning for Optimal Business Decisions

ID/X Partners - Data Scientist

Presented by
Bagus Cipta Pratama



Bagus Cipta Pratama
CS undergrad @UGM

A passionate computer science student with expertise in data science, machine learning, and app development, eager to innovate as a Data Scientist in the tech industry.



Yogyakarta



baguscp795@gmail.com



[Bagus Cipta Pratama](#)

Courses and Certification

Data Analysis with Python

January , 2025

Data Visualization with Python

January , 2025

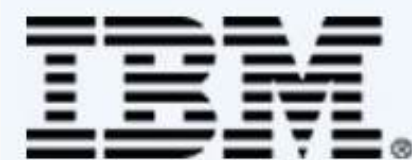
Database and SQL

December , 2024

Data Science Methodology

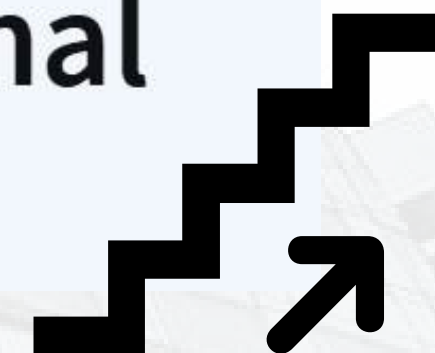
December , 2024

Etc .



IBM

**IBM Data Science Professional
Certificate**



About **Company**

id/x partners

id/x partners adalah perusahaan yang didirikan pada tahun 2002 oleh para profesional berpengalaman di bidang kredit, scoring, dan manajemen kinerja. Perusahaan ini telah melayani berbagai industri di Asia dan Australia, termasuk jasa keuangan, telekomunikasi, manufaktur, dan ritel.

Perusahaan ini mengkhususkan diri dalam konsultasi berbasis data analytic and decisioning (DAD) yang terintegrasi dengan manajemen risiko dan pemasaran untuk mengoptimalkan profitabilitas portofolio serta efisiensi proses bisnis. Dengan layanan yang komprehensif, id/x partners dikenal sebagai penyedia solusi terpadu.

Project **Portfolio**

Akurasi penilaian risiko kredit sangat penting dalam mengoptimalkan keputusan bisnis dan meminimalkan kerugian. saya mengembangkan model machine learning berbasis dataset pinjaman untuk menganalisis pola risiko, mengidentifikasi pinjaman berisiko tinggi, dan memberikan wawasan strategis guna meningkatkan efisiensi pengelolaan portofolio kredit serta mendukung pengambilan keputusan yang lebih cerdas.

Introduction



Latar Belakang



**Kebutuhan Akurasi
Penilaian Risiko Kredit**



**Peningkatan Efisiensi
Pengelolaan Portofolio Kredit**






**Pemanfaatan Teknologi
Machine Learning**

Challenges & Obstacle

Sebagai Data Scientist di ID/X Partners, Saya bertugas membantu perusahaan multifinance meningkatkan akurasi penilaian risiko kredit dan mengurangi kerugian dengan menggunakan dataset pinjaman yang mencakup informasi kredit, demografi, dan pembayaran. Data ini dieksplorasi untuk mengembangkan model prediksi risiko kredit yang baik .

Tantangan proyek ini adalah mengembangkan model machine learning untuk:

1. Mengklasifikasikan status kredit menjadi "High Risk" atau "Low Risk" 
2. Meningkatkan akurasi penilaian risiko kredit 
3. Mengurangi potensi kerugian akibat keputusan kredit yang tidak tepat 

Data Understanding



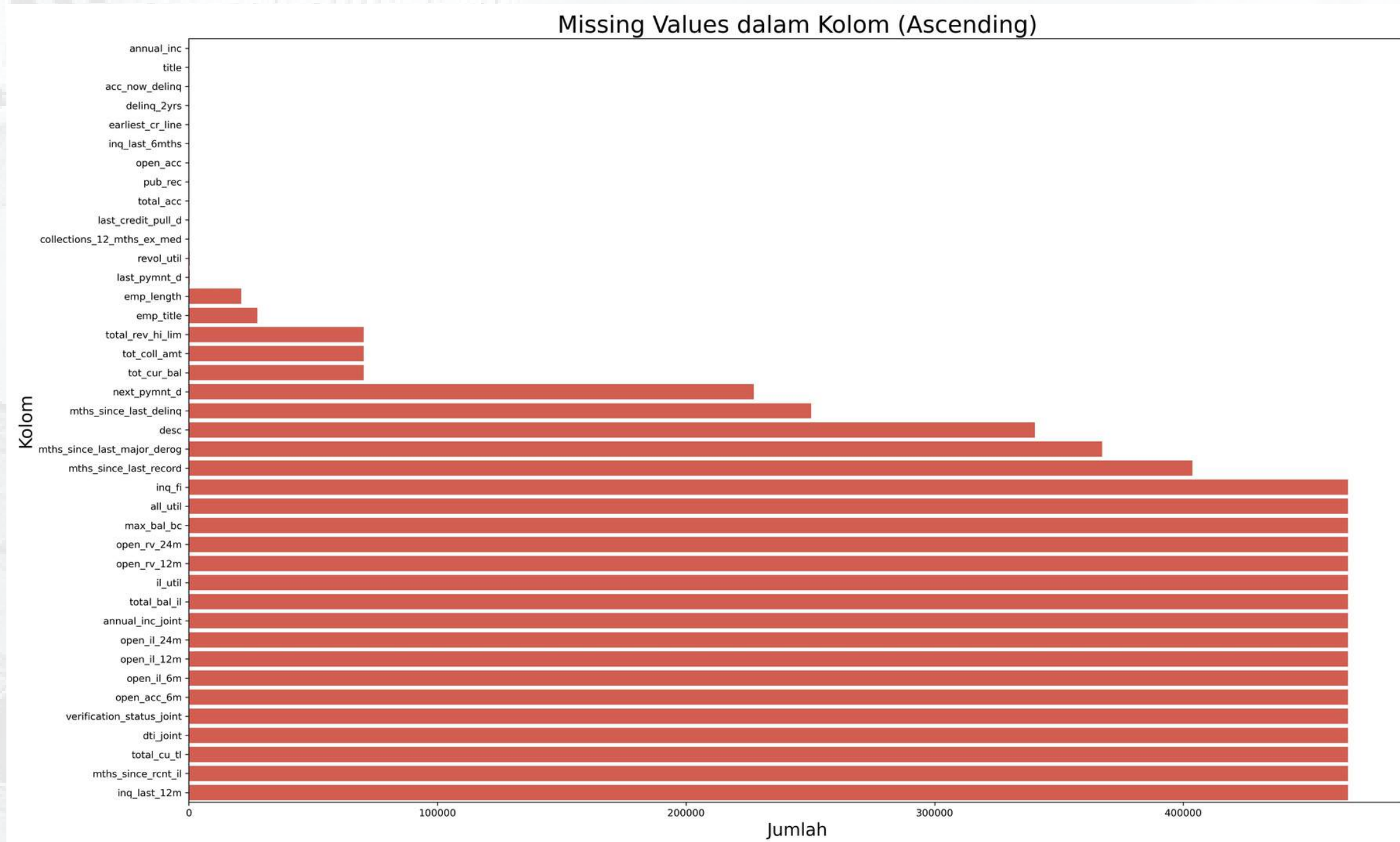
Rakamin
Academy



id/x partners

1. Data Understanding

Consist Of ➡ 466285 Rows and 75 Column



Dapat dilihat bahwa terdapat beberapa kolom yang kosong secara keseluruhan dan tidak memberikan informasi untuk analisis sehingga kolom dengan kondisi tersebut harus dihapus

```
: data.shape
```

```
: (466285, 75)
```

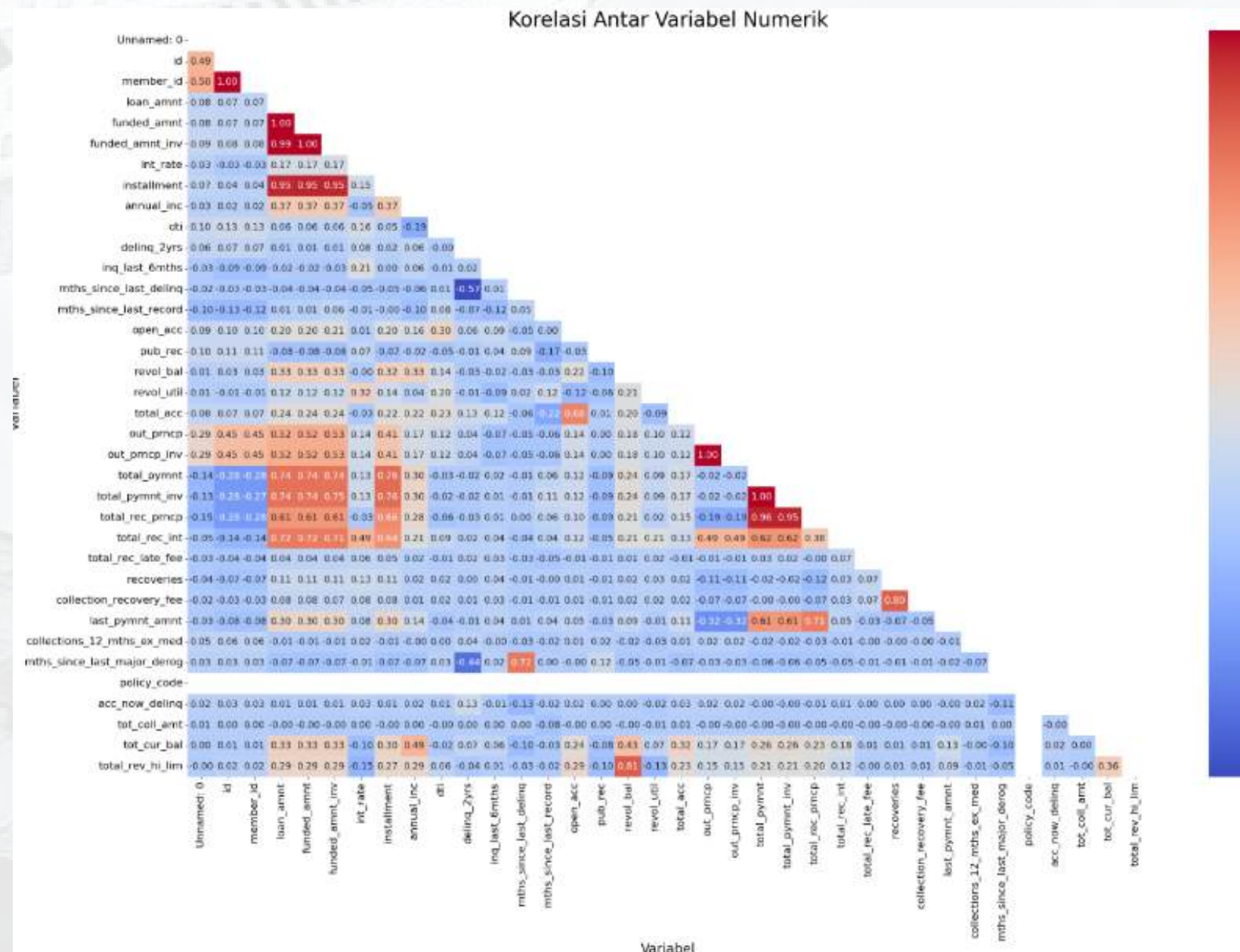

Exploratory Data Analysis



2. Exploratory Data Analysis

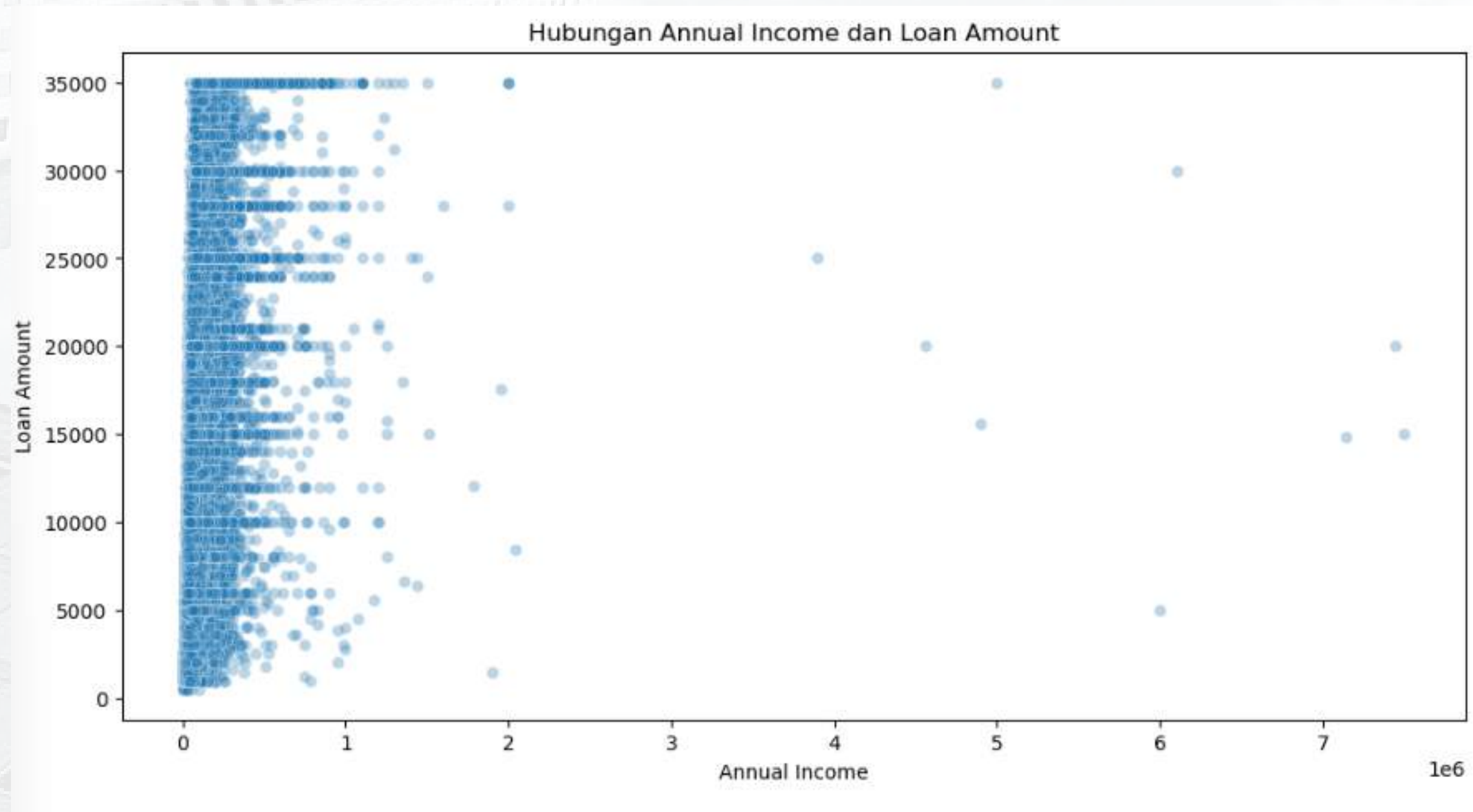
HeatMap

Pada awal sebelum proses data preparation , saya melakukan EDA dengan sebelumnya melakukan penghilangan kolom dengan nilai kosong total . dan ekstraksi bulan untuk pada kolom waktu kemudian memvisualisasikan heatmap untuk menunjukkan korelasi antara variabel numerik pada data



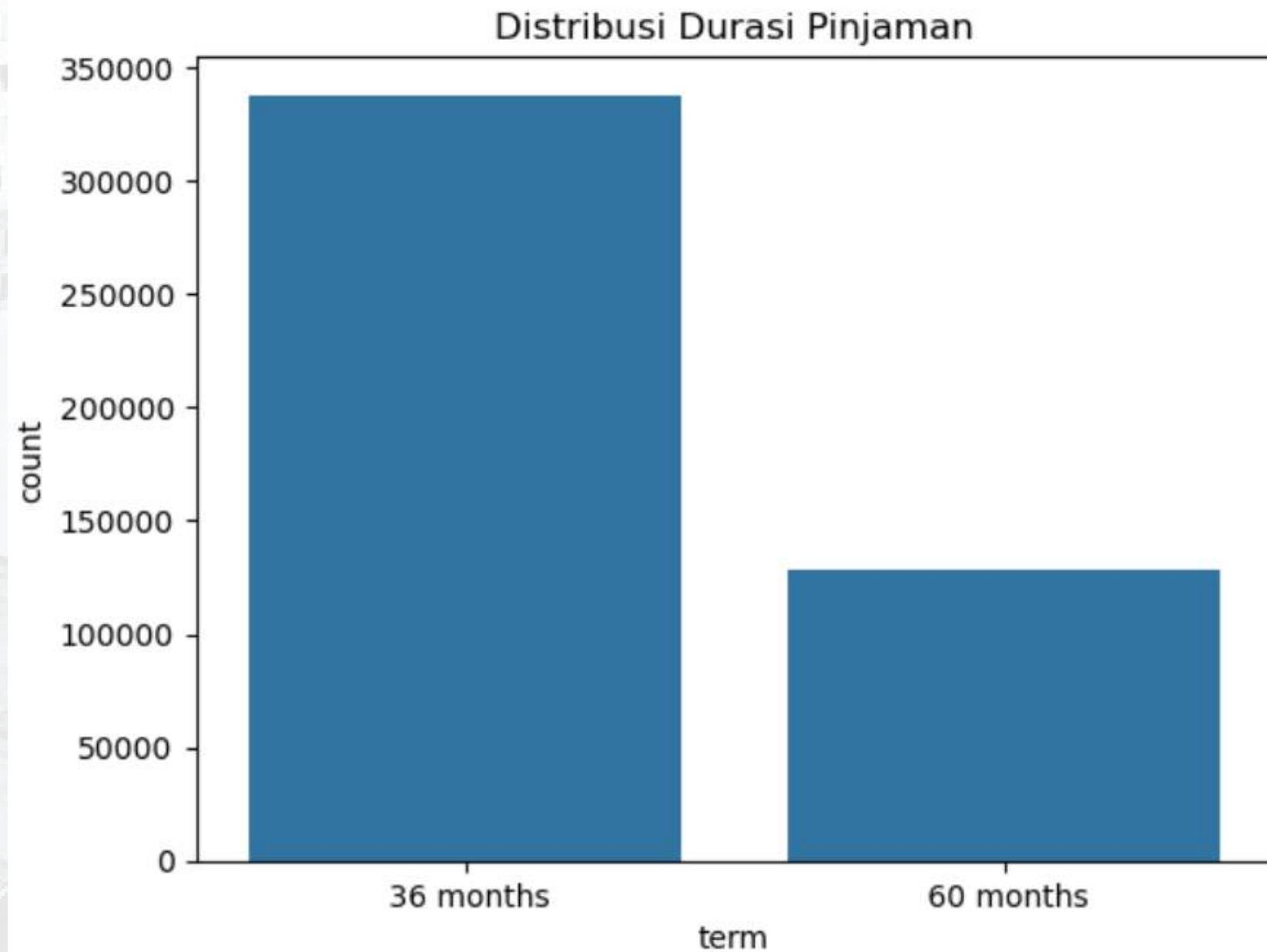
2. Exploratory Data Analysis

Hubungan Annual Income dan Loan Amount



Grafik tersebut menunjukkan distribusi status kredit, dengan mayoritas pinjaman berada dalam kategori "Current" dan "Fully Paid," mencerminkan tingginya pembayaran lancar. Sebaliknya, kategori bermasalah seperti "Charged Off" dan "Default" memiliki proporsi kecil, menyoroti risiko kredit yang relatif rendah dibandingkan total pinjaman.

2. Exploratory Data Analysis

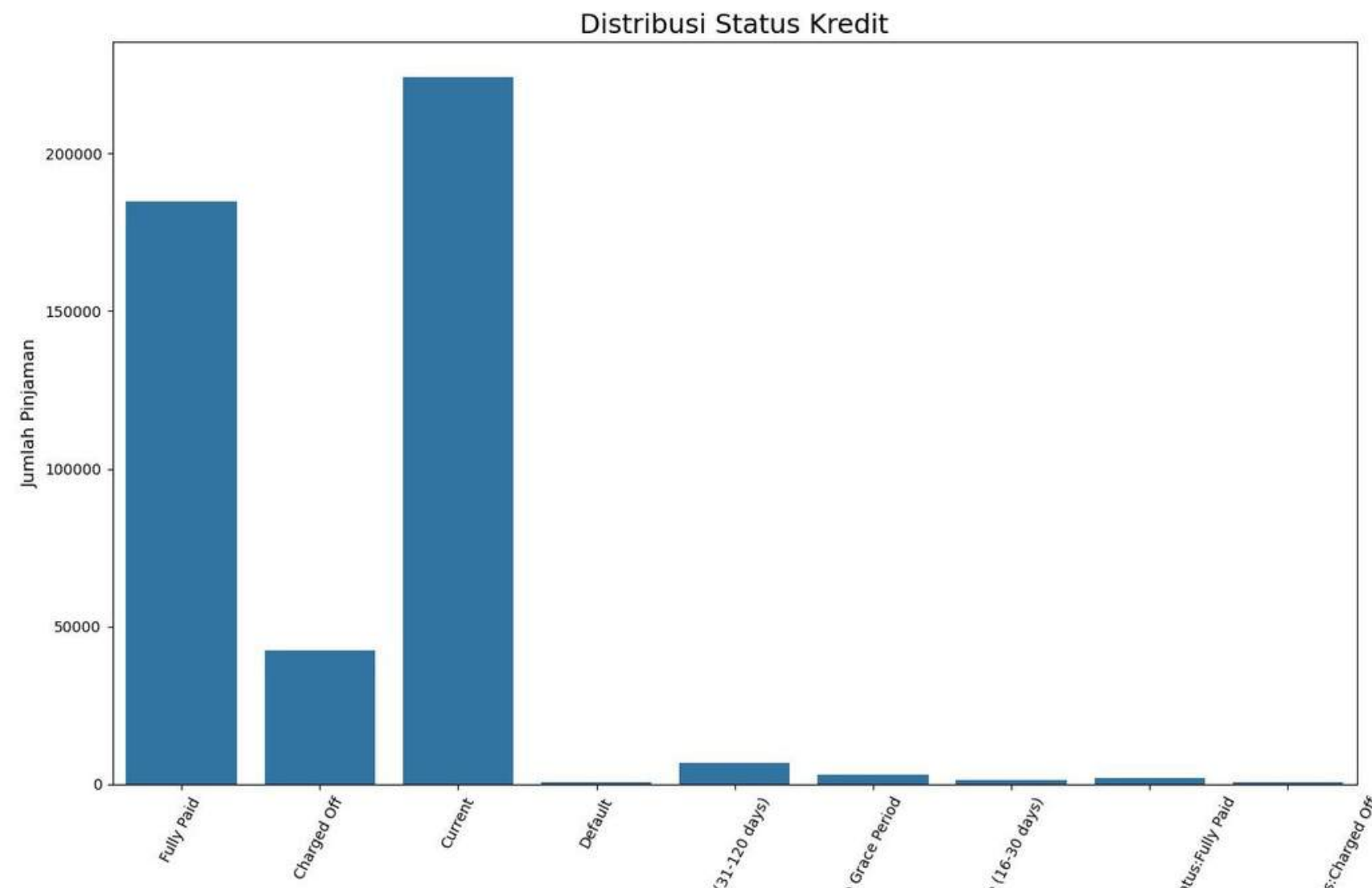


Distribusi Status Pinjaman

Grafik ini menunjukkan bahwa mayoritas pinjaman memiliki durasi 36 bulan, sementara pinjaman 60 bulan jauh lebih sedikit. Hal ini mengindikasikan preferensi peminjam untuk memilih durasi pinjaman yang lebih pendek.

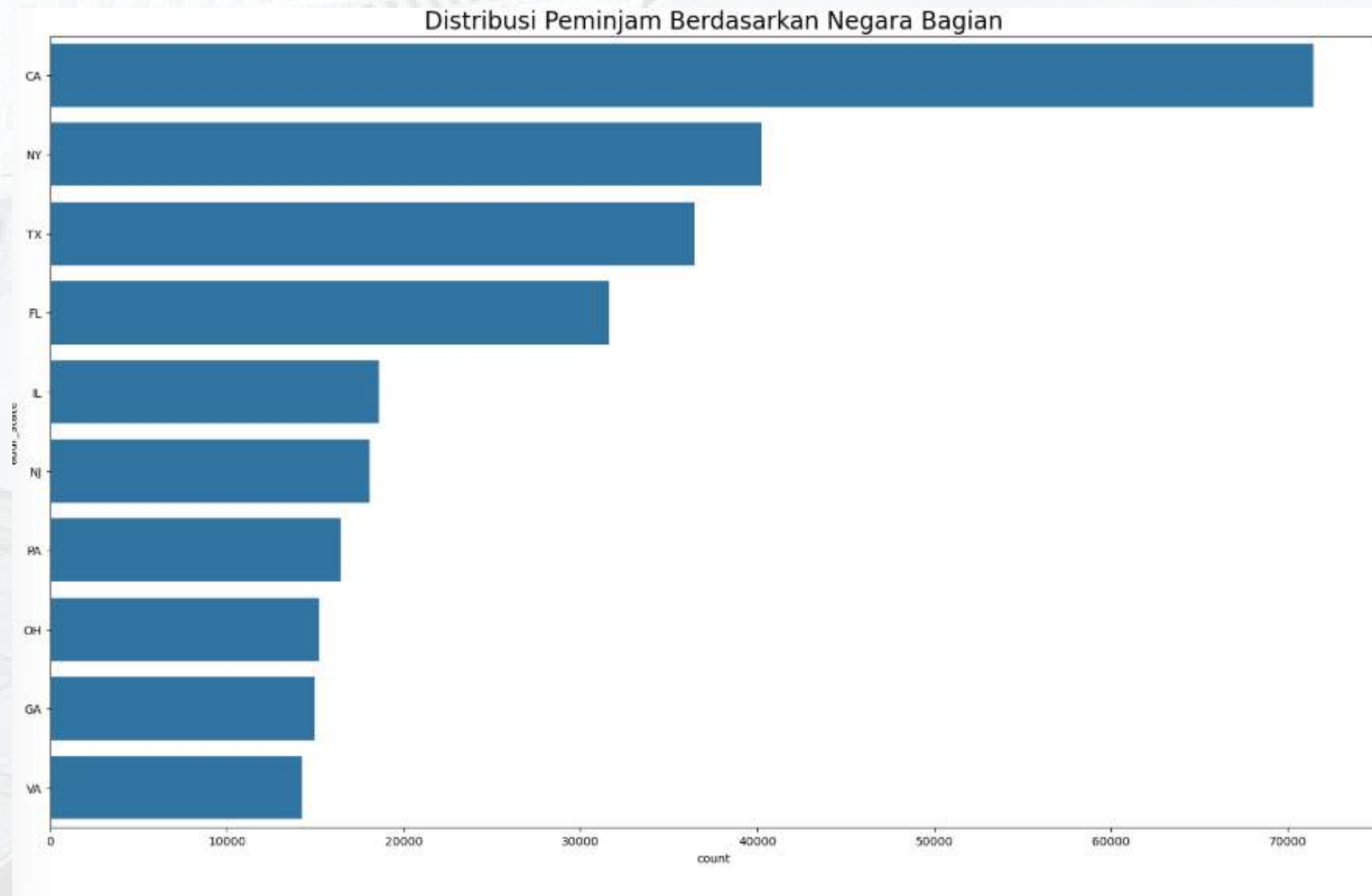
2. Exploratory Data Analysis

Distribusi status kredit



Grafik tersebut menunjukkan distribusi status kredit, dengan mayoritas pinjaman berada dalam kategori "Current" dan "Fully Paid," mencerminkan tingginya pembayaran lancar. Sebaliknya, kategori bermasalah seperti "Charged Off" dan "Default" memiliki proporsi kecil, menyoroti risiko kredit yang relatif rendah dibandingkan total pinjaman.

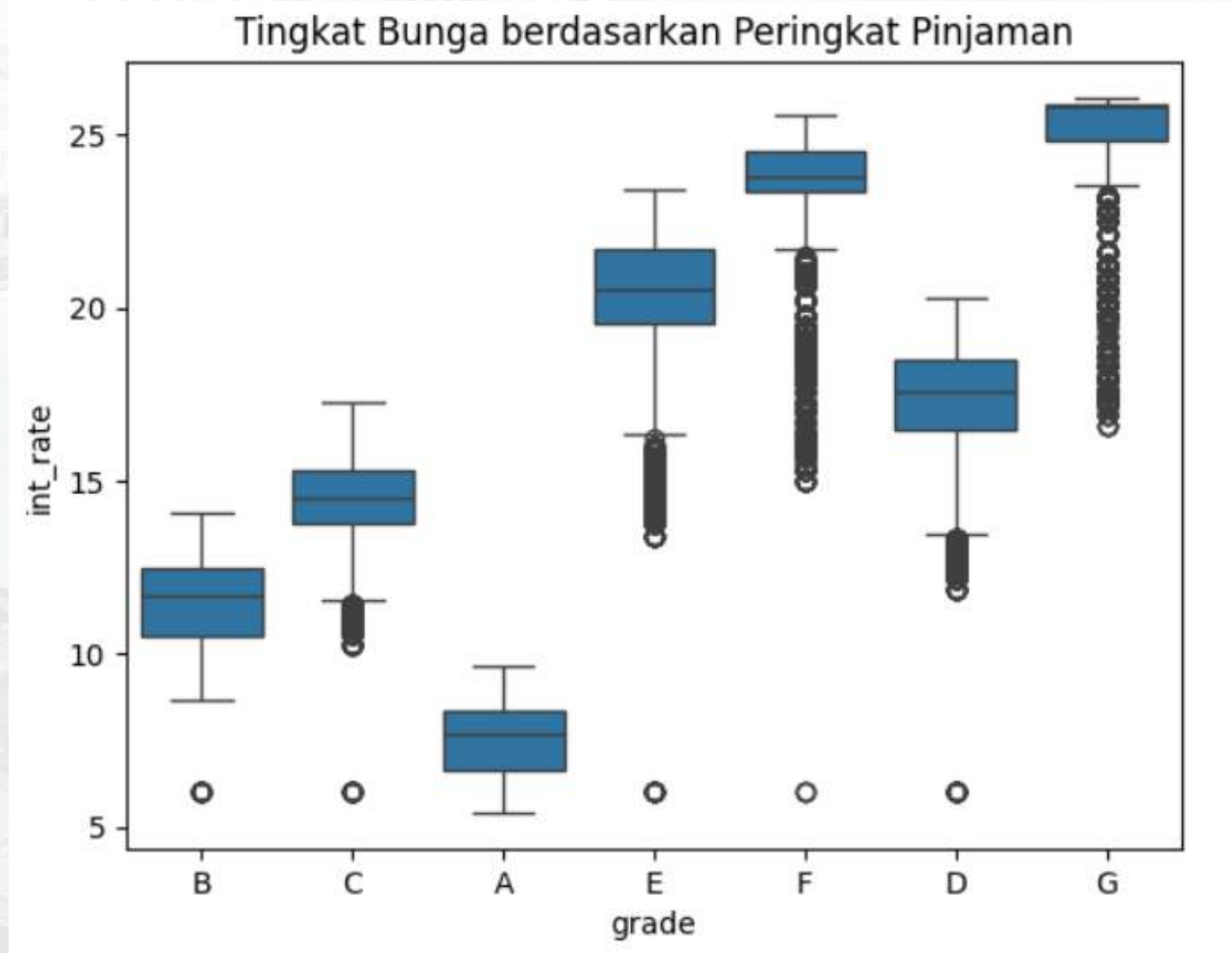
2. Exploratory Data Analysis



Top 10 Negara Bagian

Grafik ini menunjukkan distribusi peminjam terbesar di 10 negara bagian AS, dengan California (CA) menempati posisi teratas, diikuti oleh New York (NY) dan Texas (TX). Informasi ini membantu lembaga keuangan dalam menilai risiko berdasarkan wilayah, dan mengidentifikasi peluang pertumbuhan di pasar tertentu.

2. Exploratory Data Analysis



Tingkat Bunga Berdasarkan Peringkat Pinjaman

Grafik ini menunjukkan bahwa tingkat bunga meningkat seiring dengan penurunan peringkat pinjaman, dari A dengan bunga rendah hingga G dengan bunga tinggi. Peringkat rendah juga menunjukkan variasi bunga yang lebih besar dengan banyak outliers.

Data Preparation



3. Data Preparation

Menghapus Kolom

id & member_id

url

Unnamed: 0

adalah ketiga kolom dengan nilai unik sama dengan jumlah baris pada dataset sehingga kita perlu mengeliminasinya

3. Data Preparation

✦ Menghapus Kolom

zip_code

Fitur ini memiliki terlalu banyak kategori yang berbeda, sementara fitur addr_state sudah cukup mewakili informasi lokasi dengan cara yang lebih sederhana dan jelas.

application_type

Fitur ini hanya memiliki satu kategori

desc dan title

kedua fitur sudah cukup diwakilkan oleh satu kolom yaitu kolom purpose

3. Data Preparation

Menghapus Kolom

emp_title

saya memutuskan untuk mendrop kolom emp_title karena Jabatan pekerjaan peminjam memiliki terlalu banyak variasi (kategori unik) dan sering kali sulit untuk dinormalisasi. Stabilitas pekerjaan lebih baik diwakili oleh emp_length.

pyment_plan

Fitur sangat imbalanced (perbedaan kategori yang sangat besar)

Menghapus baris

saya melakukan penghapusan baris yang memiliki dengan kolom yang memiliki nilai missing value kurang dari 10000(2.5% total baris)

3. Data Preparation

Handling Missing value

Saya mengisi missing values dengan Iterative Imputer untuk kolom **tot_coll_amt**, **tot_cur_bal**, **total_rev_hi_lim**, dan **next_pymnt_d** (setelah ekstraksi bulan), serta menggunakan modus untuk kolom **mths_since_last_delinq**, **mths_since_last_record**, dan **mths_since_last_major_derog** agar dataset lebih lengkap untuk analisis.

Column	Missing Values \
emp_title	27507
emp_length	20969
desc	339789
mths_since_last_delinq	249915
mths_since_last_record	402934
next_pymnt_d	226579
mths_since_last_major_derog	366574
tot_coll_amt	69917
tot_cur_bal	69917
total_rev_hi_lim	69917

3. Data Preparation

About Iterative Imputer

IterativeImputer menggunakan pendekatan regresi linear (atau model lain seperti Bayesian Ridge) untuk mengestimasi nilai yang hilang dalam dataset. Misalnya, jika kita memiliki kolom X_1, X_2, \dots, X_n dan kolom X_k memiliki nilai hilang, maka iterasi pertama akan memperkirakan nilai X_k menggunakan model:

$$X_k = f(X_1, X_2, \dots, X_{k-1}, X_{k+1}, \dots, X_n)$$

Dimana f adalah fungsi regresi yang dipelajari dari data yang ada.

3. Data Preparation

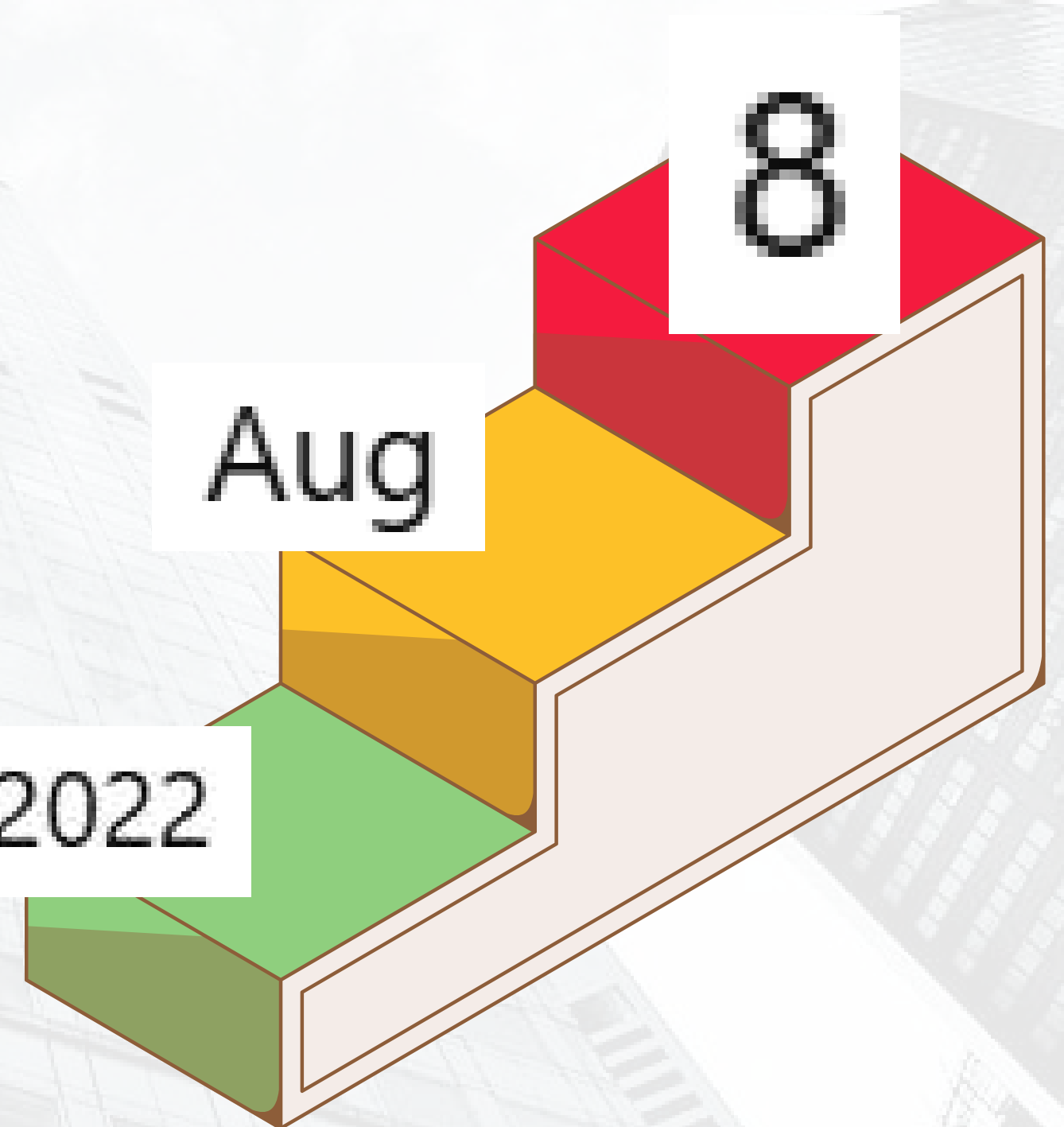
Converting Date Columns

1. Ekstraksi 3 karakter pertama
2. Pemetaan angka ke bulan
3. mendapatkan Hasil

Aug-2022

Aug

8



3. Data Preparation

Labelling

Non Risky

'Fully Paid' ,Status:Fully Paid',Current

High Risk

"In Grace Period", "Late (31-120 days)",
"Default"

Mapping value

```
'High Risk': 1,  
'Non Risky': 0
```


Feature Engineering



4. Feature Engineering

Encoding

Ordinal Encoding

- `term`
- `grade`
- `sub_grade`
- `emp_length`
- `verification_status`

Target Encoding

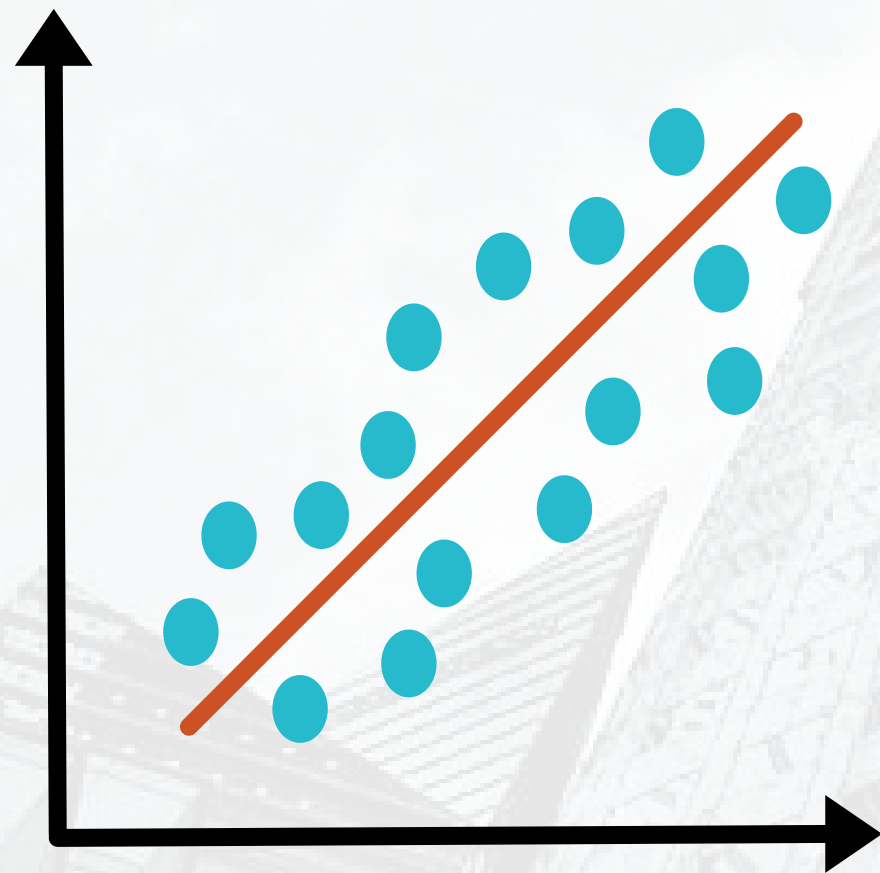
- `home_ownership`
- `purpose`
- `addr_state`
- `initial_list_status`

Data Modelling



1. Data Modeling

Saya Menggunakan Dua model untuk melakukan prediksi risiko pinjaman



Logistic Regression



XGBoost

1. Data Modeling

Data Siap Pakai

Training Set 70%

Dibagi

Test Set 30%

Dilatih

Best Parameter (GridSearch)

Evaluation & Conclusion



5. Evaluation

Best Parameters

Logistic Regression

$C = 10$ dan $\text{solver} = \text{'liblinear'}$

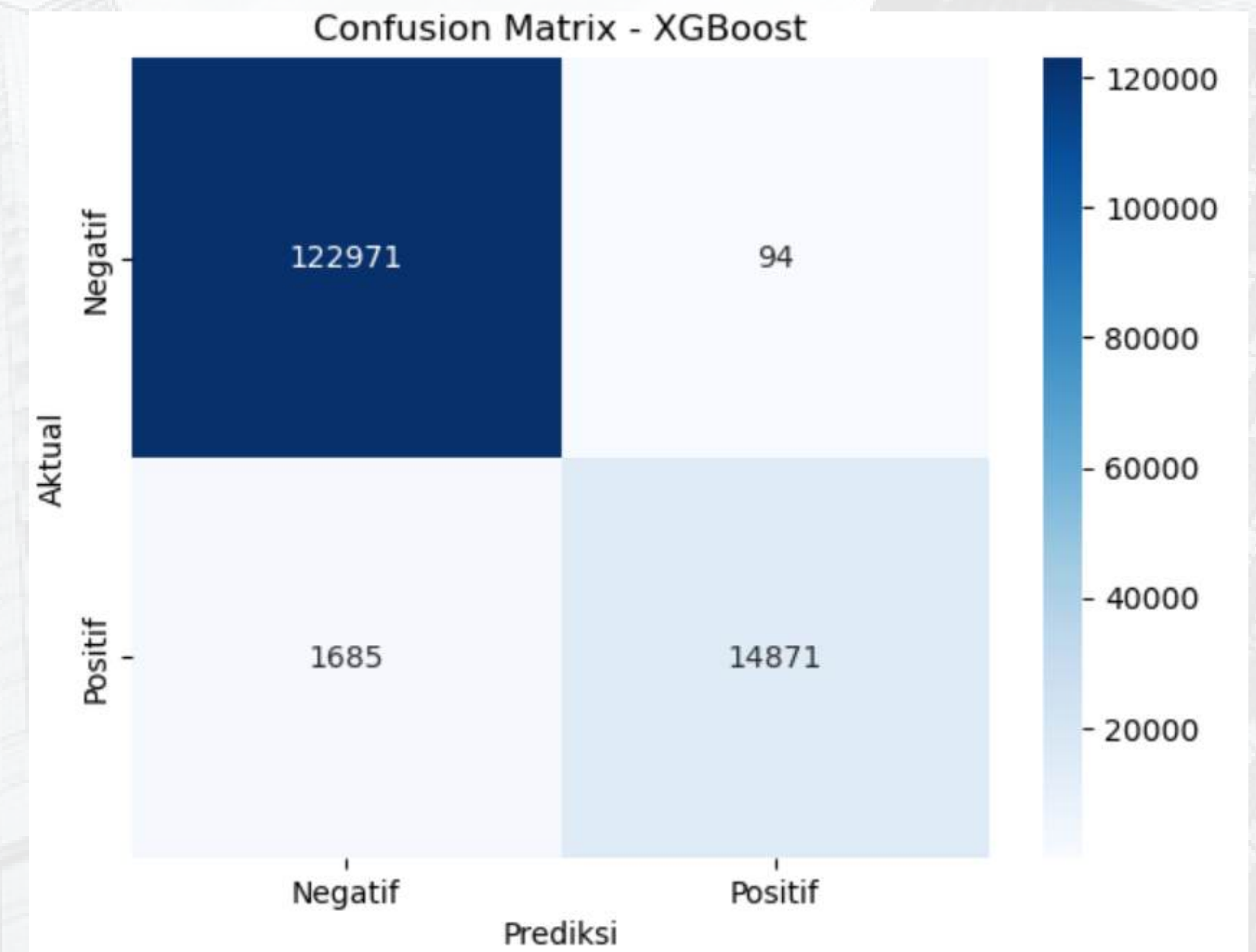
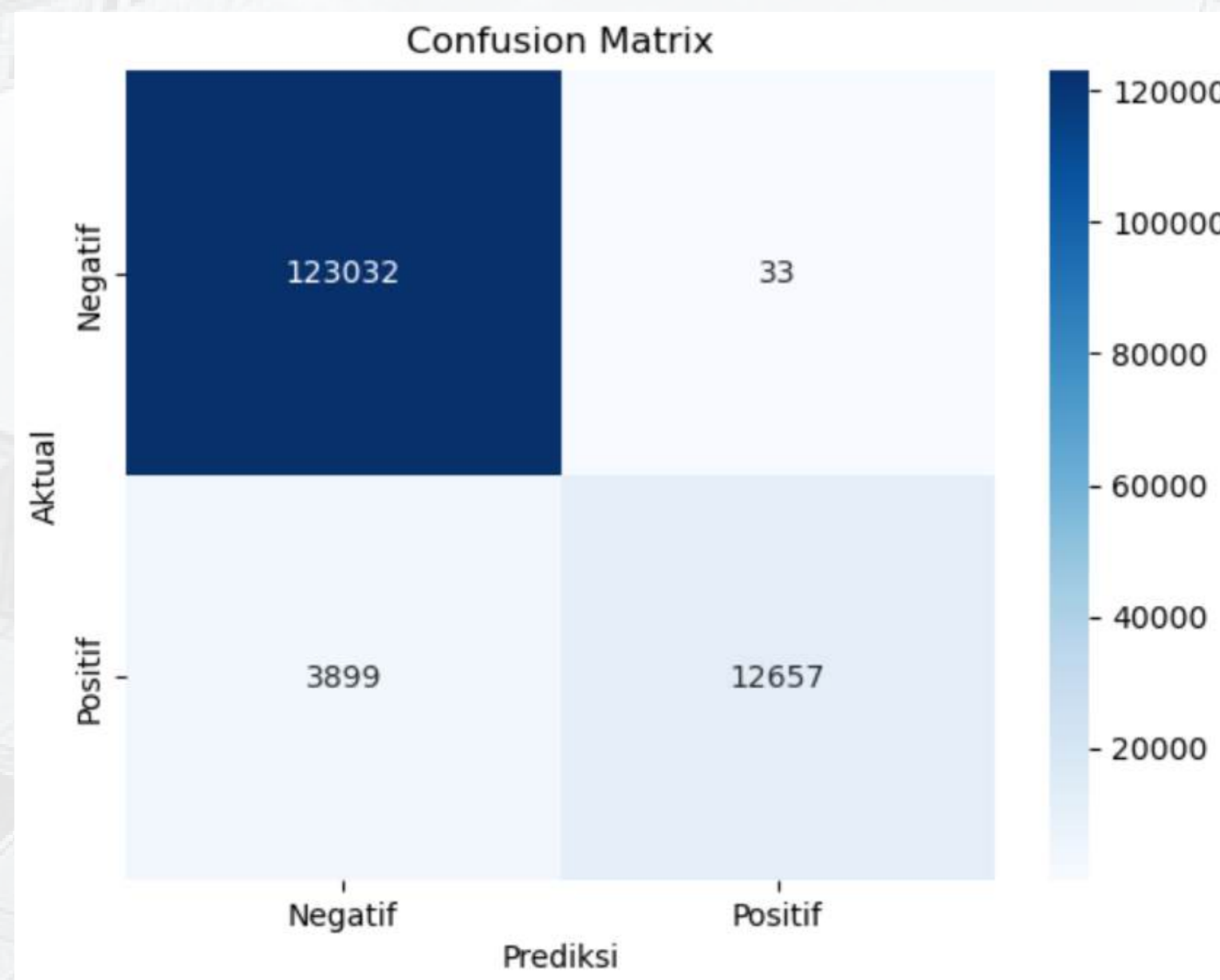
XGBoost

$\text{max_depth} = 5$, $\text{learning_rate} = 0.2$, $\text{n_estimators} = 200$,
 $\text{subsample} = 1.0$, dan $\text{colsample_bytree} = 0.8$



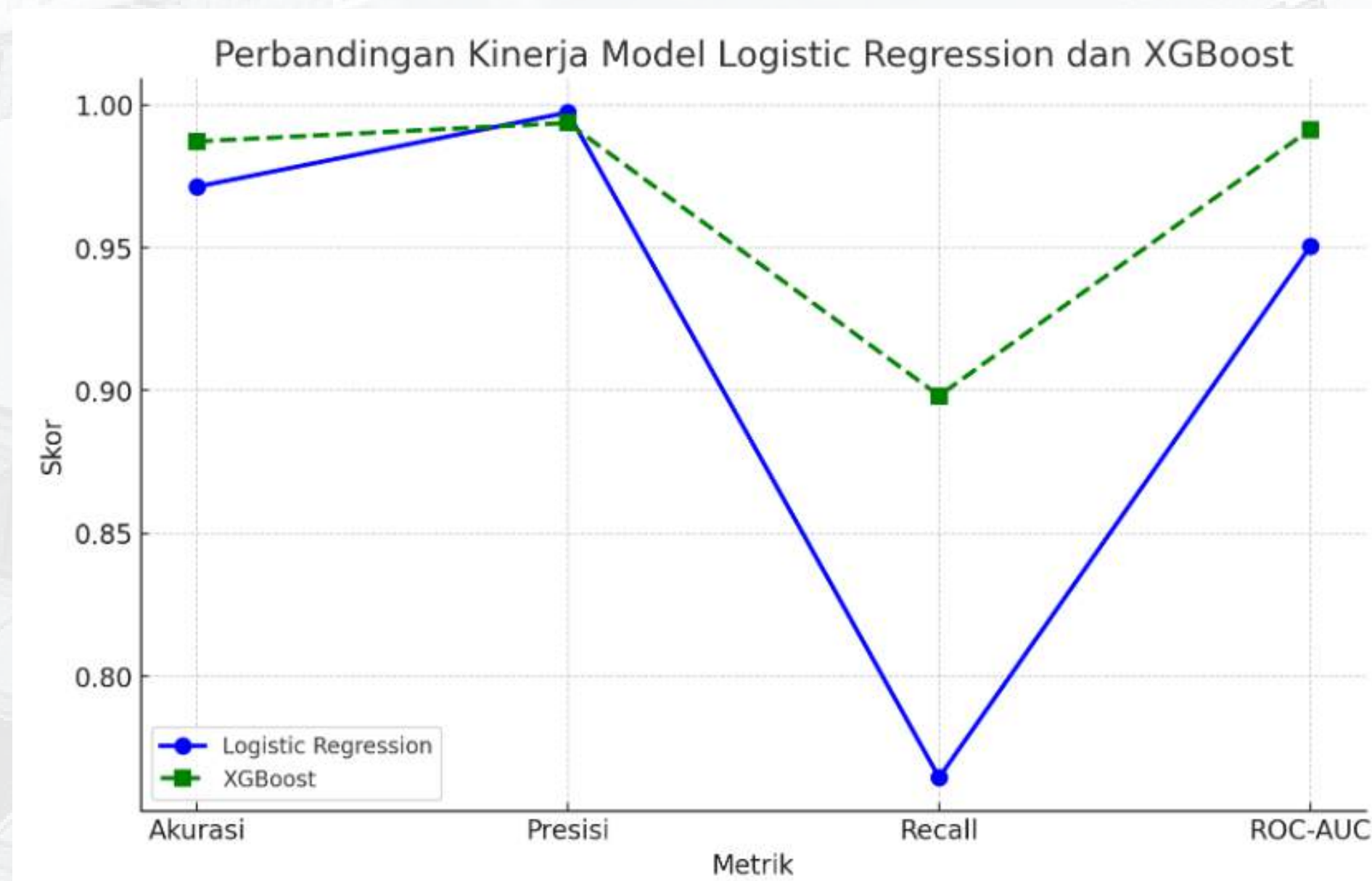
5. Evaluation

Confusion Matrix



5. Evaluation

Perbandingan Akurasi



Logistic Regression

Accuracy: 0.9718380472851506
Precision: 0.9973995271867613
Recall: 0.7644962551340904
ROC-AUC: 0.950458480216491
Train Accuracy: 0.9720240653201547
Test Accuracy: 0.9718380472851506

XGBoost

Accuracy: 0.987258363713195
Precision: 0.9937186769127965
Recall: 0.8982242087460739
ROC-AUC: 0.9914449939717711
Train Accuracy: 0.9889158327705814
Test Accuracy: 0.987258363713195

5. Conclusion

Perbandingan hasil kedua model menunjukkan bahwa XGBoost unggul dalam Akurasi (98.73%), Recall (89.82%), dan ROC-AUC (99.14%), dengan kinerja lebih baik dalam memprediksi kategori positif. Sementara itu, **Logistic Regression mencatatkan Presisi lebih tinggi (99.73%), meskipun Recall-nya lebih rendah (76.45%),** dengan Akurasi uji sebesar 97.13%. Kedua model memiliki kinerja yang seimbang antara data pelatihan dan pengujian, tetapi **XGBoost lebih unggul secara keseluruhan** dalam mengidentifikasi risiko pinjaman,

Thank You

