

STAGE 1



Conexus Group

I. EXPLORATORY DATA ANALYSIS

1.1. Descriptive Statistics

1.1.1. HC_application_train | test

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

- Seluruh kolom memiliki tipe data yang sesuai.

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

- Ada 3 kolom yang memiliki data yang kosong: AMT_ANNUITY, AMT_GOODS_PRICE, NAME_TYPE_SUITE.

Apakah ada kolom yang memiliki nilai summary agak aneh?

- max(outlier): CNT_CHILDREN, AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE.
- top(mode): NAME_CONTRACT_TYPE(Cash loans), CODE_GENDER(F), FLAG_OWN_CAR(N), FLAG_OWN_REALTY(Y), NAME_TYPE_SUITE(Unaccompanied), NAME_EDUCATION_TYPE(Secondary / secondary special), NAME_FAMILY_STATUS(Married), NAME_HOUSING_TYPE(House / apartment).

1.1.2. HC_bureau

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

- Bureau Dataset berisikan 1.716.428 baris data, terdiri dari 17 features/kolom.

- Pada Bureau Dataset **semua kolom memiliki tipe data yang sesuai**
- Pada feature "**CREDIT_CURRENCY**" berisikan : currency 1-3, hal ini belum diketahui dengan pasti maksud dari masing-masing kategori.
- Beberapa features Bureau Dataset merupakan informasi *days* dan bernilai negatif (-). Hal ini perlu dipertimbangkan secara khusus, karena model Machine Learning harus dipersiapkan agar dapat mengenali pola nilai tersebut.

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

- Pada Bureau Dataset **tidak ditemukan duplikat data antar setiap baris.**
- Pada Bureau Dataset terdapat kolom yang memiliki nilai kosong atau missing values, antara lain :

1. DAYS_CREDIT_ENDDATE sebanyak 105.553 baris	= 6,15% data
2. DAYS_ENDDATE_FACT sebanyak 633.653 baris	= 36,91% data
3. AMT_CREDIT_MAX_OVERDUE sebanyak 1.124.488 baris	= 65,51% data
4. AMT_CREDIT_SUM sebanyak 13 baris	= 0,0007% data
5. AMT_CREDIT_SUM_DEBT sebanyak 257.669 baris	= 15,01% data
6. AMT_CREDIT_SUM_LIMIT sebanyak 591.780 baris	= 34,47% data
7. AMT_ANNUITY sebanyak 1.226.791 baris	= 71,47% data

Apakah ada kolom yang memiliki nilai summary agak aneh?

Pada Bureau Dataset terdapat kolom dengan distribusi data dan statistic yang tidak normal, antara lain :

1. DAYS_CREDIT

Mean < median (negatively skewed distribution)

2. CREDIT_DAY_OVERDUE

Distribusi timpang karena nilai max terlalu jauh, potensi besar untuk skewness dan outliers

Min, 25%, 50% = 0

Max = 2.792

3. DAYS_CREDIT_ENDDATE

Distribusi timpang karena nilai mean, median dan max terlalu jauh, potensi besar untuk skewness dan outliers

Mean = 510.52

Median = -330.00

Max = 31.199

4. DAYS_ENDDATE_FACT

Mean < median (negatively skewed distribution)

5. AMT_CREDIT_MAX_OVERDUE

Distribusi timpang karena nilai max terlalu jauh, potensi besar untuk skewness dan outliers

Min, 25%, 50% = 0

Mean = 3.825

Max = 115.987.185

6. AMT_CREDIT_SUM

Mean > median (positively skewed distribution)

Distribusi timpang karena nilai max terlalu jauh, potensi besar untuk skewness dan outliers

Median = 125.518

Max = 585.000.000

7. AMT_CREDIT_SUM_DEBT

Mean > median (positively skewed distribution)

Distribusi timpang karena nilai max terlalu jauh, potensi besar untuk skewness dan outliers

Median = 0

Max = 170.100.000

8. AMT_CREDIT_SUM_LIMIT

Mean > median (positively skewed distribution)

Distribusi timpang karena nilai max terlalu jauh, potensi besar untuk skewness dan outliers

Median = 0

Max = 4.705.600

9. AMT_CREDIT_SUM_OVERDUE

Mean > median (positively skewed distribution)

Distribusi timpang karena nilai max terlalu jauh, potensi besar untuk skewness dan outliers

Min, 25%, 50% = 0

Max = 3.756.681

10. AMT_ANNUITY

Distribusi timpang karena nilai max terlalu jauh, potensi besar untuk skewness dan outliers

Min, 25%, 50% = 0

max = 118.453.423

Mean < median (negatively skewed distribution)

1.1.3. HC_credit_card_balance

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

Tipe Data sudah sesuai dengan Valuenya. Jadi tidak perlu mengubah tipe data

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

Ada beberapa kolom yang memiliki nilai kosong. Karena jumlah baris kolomnya tidak sesuai dengan jumlah bari keseluruhan, 'SK_ID_PREV', 'SK_ID_CURR',

'MONTHS_BALANCE', 'AMT_BALANCE', 'AMT_CREDIT_LIMIT_ACTUAL',
'AMT_DRAWINGS_ATM_CURRENT', 'AMT_DRAWINGS_CURRENT',
'AMT_DRAWINGS_OTHER_CURRENT', 'AMT_DRAWINGS_POS_CURRENT',
'AMT_INST_MIN_REGULARITY', 'AMT_PAYMENT_CURRENT',
'AMT_PAYMENT_TOTAL_CURRENT', 'AMT_RECEIVABLE_PRINCIPAL',
'AMT_RECVABLE', 'AMT_TOTAL_RECEIVABLE',
'CNT_DRAWINGS_ATM_CURRENT', 'CNT_DRAWINGS_CURRENT',
'CNT_DRAWINGS_OTHER_CURRENT', 'CNT_DRAWINGS_POS_CURRENT',
'CNT_INSTALMENT_MATURE_CUM', 'SK_DPD', 'SK_DPD_DEF'

**Apakah ada kolom yang memiliki nilai summary agak aneh?
(min/mean/median/max/unique/top/freq)**

Berikut beberapa kolom yang nilai mean nya berbeda jauh dengan nilai maximal :

'AMT_DRAWINGS_ATM_CURRENT', 'AMT_DRAWINGS_CURRENT',
'AMT_DRAWINGS_OTHER_CURRENT', 'AMT_DRAWINGS_POS_CURRENT',
'AMT_PAYMENT_CURRENT', 'AMT_PAYMENT_TOTAL_CURRENT'.

Berikut beberapa kolom yang nilai mean nya berbeda jauh dengan nilai maximal seperti kolom serta memiliki nilai Q1(kuartil-1), Q2(kuartil-2), Q3(kuartil-3) yang bernilai 0 :

'AMT_DRAWINGS_ATM_CURRENT', 'AMT_DRAWINGS_CURRENT',
'AMT_DRAWINGS_OTHER_CURRENT',
'AMT_DRAWINGS_POS_CURRENT', 'AMT_INST_MIN_REGULARITY',
'AMT_PAYMENT_TOTAL_CURRENT', 'AMT_RECEIVABLE_PRINCIPAL',
'AMT_RECVABLE', 'AMT_TOTAL_RECEIVABLE',
'CNT_DRAWINGS_ATM_CURRENT', 'CNT_DRAWINGS_CURRENT',
'CNT_DRAWINGS_OTHER_CURRENT', 'CNT_DRAWINGS_POS_CURRENT',
'NAME_CONTRACT_STATUS', 'SK_DPD', 'SK_DPD_DEF'.

Berikut kolom-kolom yang nilai meannya berbeda jauh dengan nilai median

'AMT_DRAWINGS_ATM_CURRENT', 'AMT_DRAWINGS_CURRENT',
'AMT_DRAWINGS_POS_CURRENT', 'AMT_INST_MIN_REGULARITY',
'AMT_PAYMENT_CURRENT', 'AMT_PAYMENT_TOTAL_CURRENT',
'AMT_RECEIVABLE_PRINCIPAL', 'AMT_RECVABLE',
'AMT_TOTAL_RECEIVABLE', 'CNT_DRAWINGS_ATM_CURRENT',
'CNT_DRAWINGS_CURRENT', 'CNT_DRAWINGS_OTHER_CURRENT',
'CNT_DRAWINGS_POS_CURRENT', 'CNT_INSTALMENT_MATURE_CUM',
'NAME_CONTRACT_STATUS', 'SK_DPD', 'SK_DPD_DEF'.

Maka perlu diteliti lebih lanjut apakah nilai 0 dipengaruhi karena ada data yang null, NaN atau emang bernilai 0.

Dan juga harus diteliti lebih lanjut mengenai nilai max yang ada apakah masih dikatakan wajar atau tidak, mengingat bahwa tidak semua orang memiliki karakteristik atau perilaku yang sama.

Jadi ada beberapa nilai median bernilai 0. Hal ini menunjukkan berarti nilai dibawah 50% bernilai 0. Kemungkinan data yang bernilai NaN atau kosong dianggap 0. **Maka perlu dianalisa lebih lanjut mengenai penyebabnya serta mengatasi permasalahan ini.**

Berdasarkan tabel describe pada data kategori dapat diketahui bahwa jumlah unique valuesnya masih masuk akal, dan frekuensi dari nilai yang paling umum tidak timpang.

1.1.4. HC_installments_payments

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

Semua kolom mempunyai tipe data yang sudah sesuai

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

Ada 2 kolom yang mempunyai nilai null, kolom DAYS_ENTRY_PAYMENT dan AMT_PAYMENT mempunyai 2905 nilai null atau sebesar 0.02%. Jika dilakukan left join dengan target, maka target mempunyai 2,013,809 nilai null atau sebesar 14.8%

Apakah ada kolom yang memiliki nilai summary agak aneh? (min/mean/median/max/unique/top/freq)

- Kolom NUM_INSTALMENT_VERSION dan NUM_INSTALMENT_NUMBER mempunyai nilai max yang sangat tinggi sehingga ada potensi outlier dan terlihat juga ada potensi skew yang terlihat dari perbedaan mean dengan median.
- Kolom DAYS_INSTALMENT dan DAYS_ENTRY_PAYMENT mempunyai perbedaan nilai median dengan mean sehingga berpotensi skew, dan kolom DAYS_ENTRY_PAYMENT mempunyai nilai min yang sangat tinggi sehingga ada potensi outlier.
- Kolom AMT_INSTALMENT dan AMT_PAYMENT mempunyai nilai max yang sangat tinggi sehingga ada potensi outlier dan mempunyai perbedaan mean dengan median sehingga berpotensi skew.

1.1.5. HC_POS_CASH_balance

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

Tidak ditemukan adanya tipe data yang kurang sesuai dengan isi dari kolom. Semua kolom memiliki tipe data yang sudah sesuai, sehingga tidak ada kebutuhan untuk mengubah tipe data. Nama kolom sesuai dengan isi yang direpresentasikan.

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

- Kolom **CNT_INSTALMENT** memiliki 26,071 nilai kosong (sekitar 26.07% dari total data).
- Kolom **CNT_INSTALMENT_FUTURE** memiliki 26,087 nilai kosong (sekitar 26.08% dari total data).
- Kolom lainnya (SK_ID_PREV, SK_ID_CURR, MONTHS_BALANCE, NAME_CONTRACT_STATUS, SK_DPD, dan SK_DPD_DEF) tidak memiliki nilai kosong.

**Apakah ada kolom yang memiliki nilai summary agak aneh?
(min/mean/median/max/unique/top/freq)**

- **Kolom CNT_INSTALMENT dan CNT_INSTALMENT_FUTURE:**

Terdapat perbedaan signifikan antara nilai maksimum dengan nilai mean dan median pada kedua kolom ini, yang menunjukkan kemungkinan outlier atau distribusi data yang sangat skewed .

- **Kolom SK_DPD dan SK_DPD_DEF:**

Nilai maksimum pada kedua kolom ini sangat jauh lebih besar dibandingkan dengan nilai mean, yang menunjukkan adanya ketidakseimbangan dalam distribusi data atau terdapat beberapa nilai yang sangat tinggi.

- **Kolom MONTHS_BALANCE:**

Rentang nilai dari -96 hingga -1 menunjukkan data historis dari bulan ke bulan. Hal ini tampak sesuai, tetapi distribusi ini perlu diperhatikan dalam analisis lebih lanjut untuk menghindari kesalahan interpretasi waktu kontrak.

Secara keseluruhan, ada beberapa indikasi ketidakseimbangan data di kolom numerik, yang mungkin memerlukan penanganan seperti normalisasi atau penghapusan outlier.

1.1.6. HC_previous_application

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

Semua tipe data pada previous_application.csv sudah sesuai

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

Terdapat nilai kosong pada kolom:

AMT_ANNUITY, AMT_CREDIT, AMT_DOWN_PAYMENT, AMT_GOODS_PRICE,
RATE_DOWN_PAYMENT, RATE_INTEREST_PRIMARY,
RATE_INTEREST_PRIVILEGED, NAME_TYPE_SUITE, CNT_PAYMENT,
PRODUCT_COMBINATION, DAYS_FIRST_DRAWING, DAYS_FIRST_DUE,
DAYS_LAST_DUE_1ST_VERSION, DAYS_LAST_DUE, DAYS_TERMINATION,
NFLAG_INSURED_ON_APPROVAL.

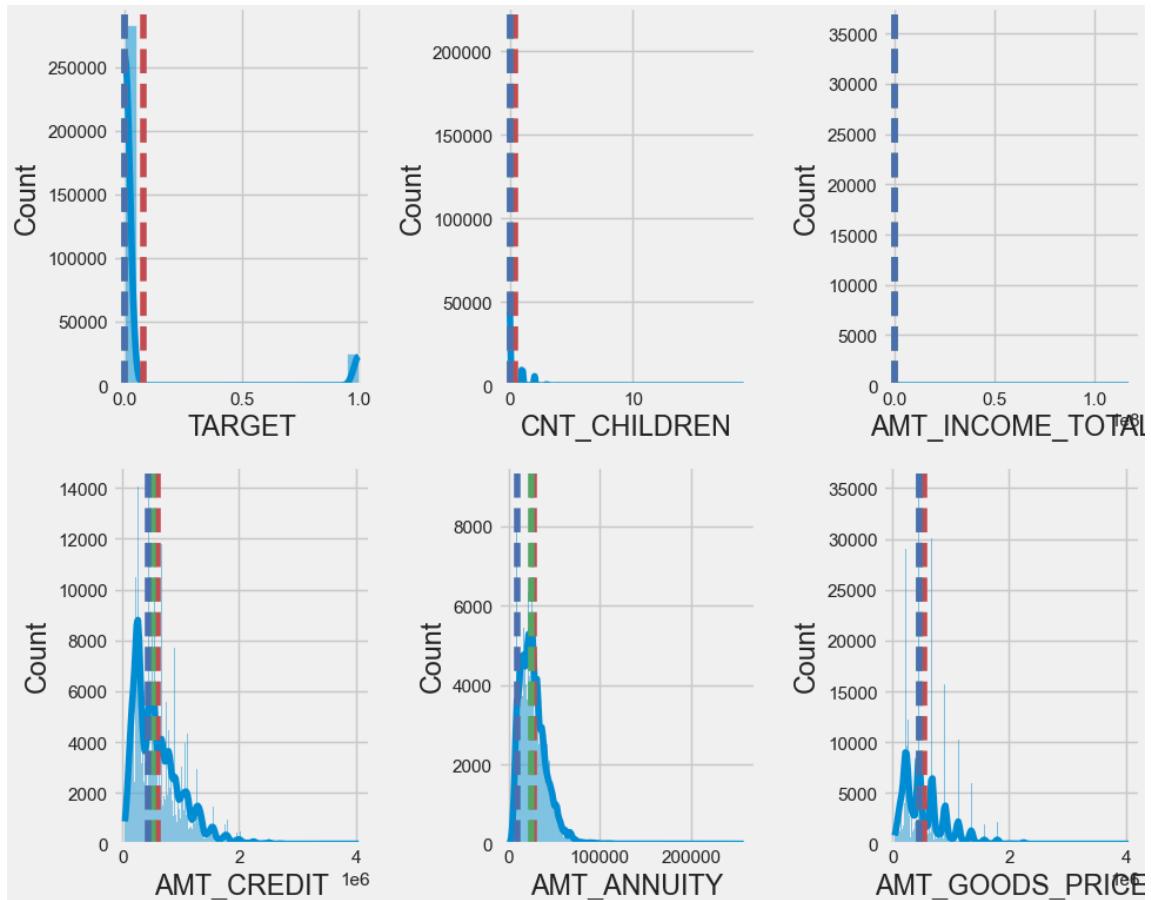
Apakah ada kolom yang memiliki nilai summary agak aneh?
(min/mean/median/max/unique/top/freq)

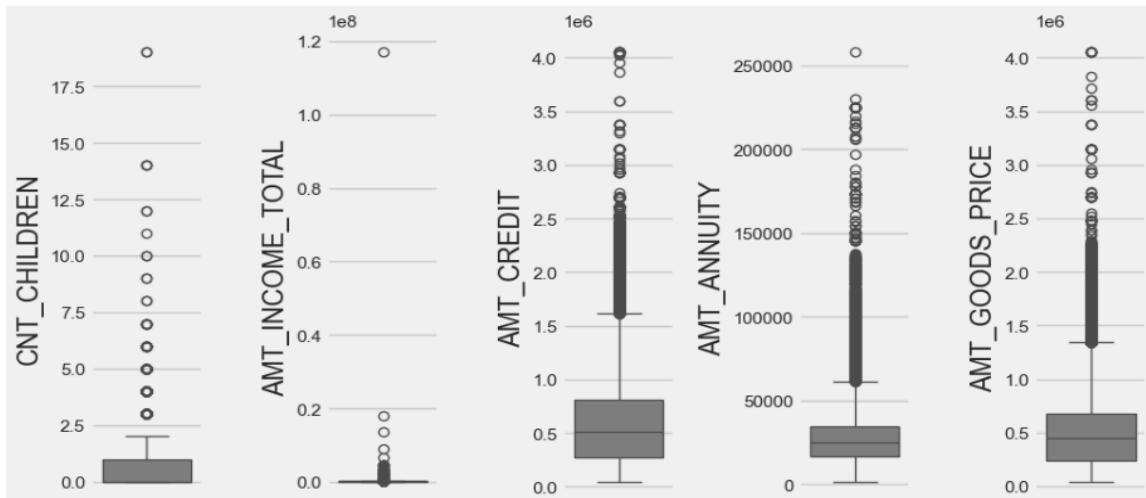
DAY_S_FIRST_DRAWING, DAY_S_FIRST_DUE, DAY_S_LAST_DUE_1ST_VERSION, DAY_S_LAST_DUE, dan DAY_S_TERMINATION memiliki nilai min yang jauh sekali dari mean, memiliki nilai sampai minus.

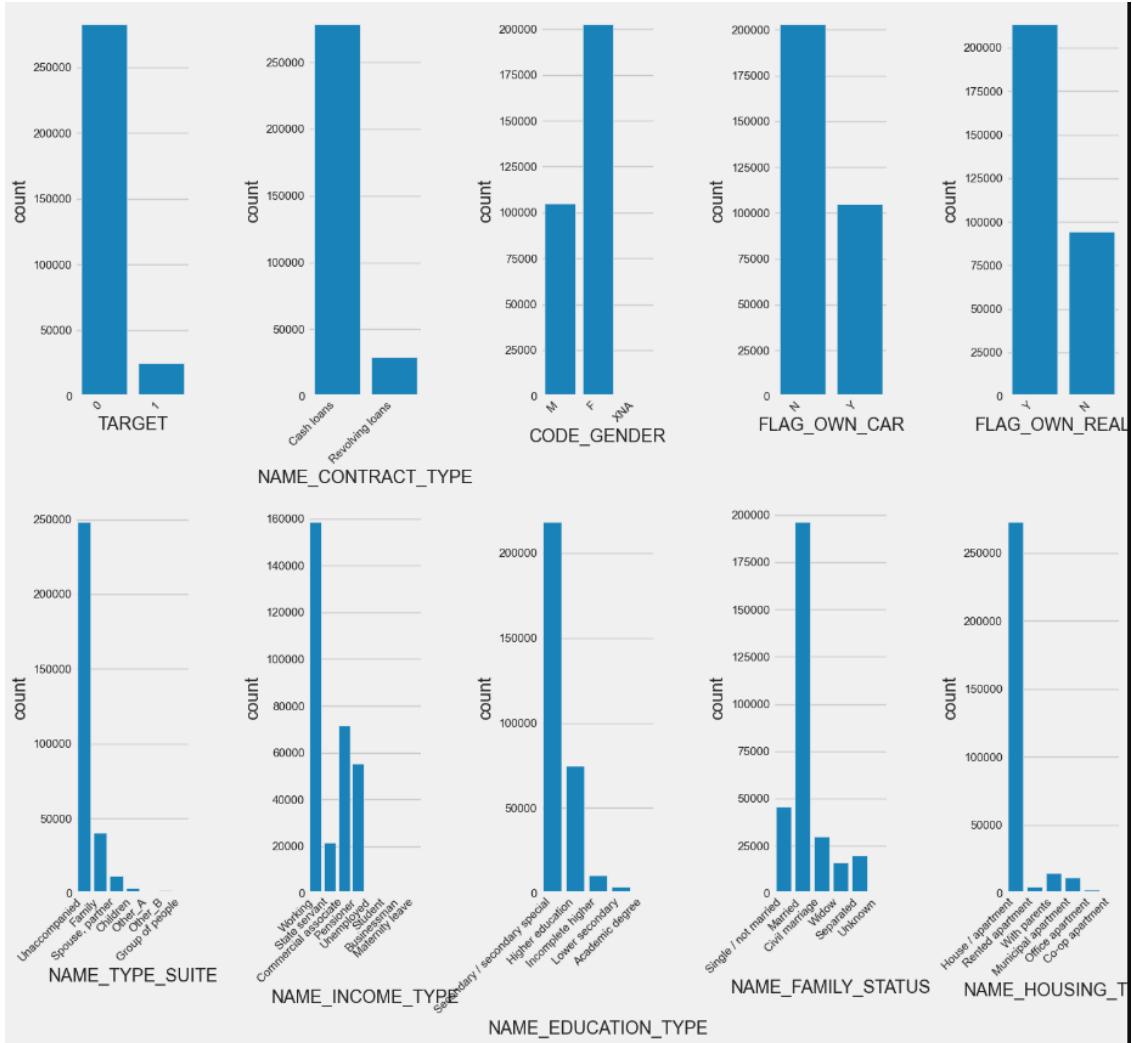
II. BUSINESS INSIGHT AND RECOMMENDATIONS

2.1. Univariate Analysis

2.1.1. HC_application_train | test







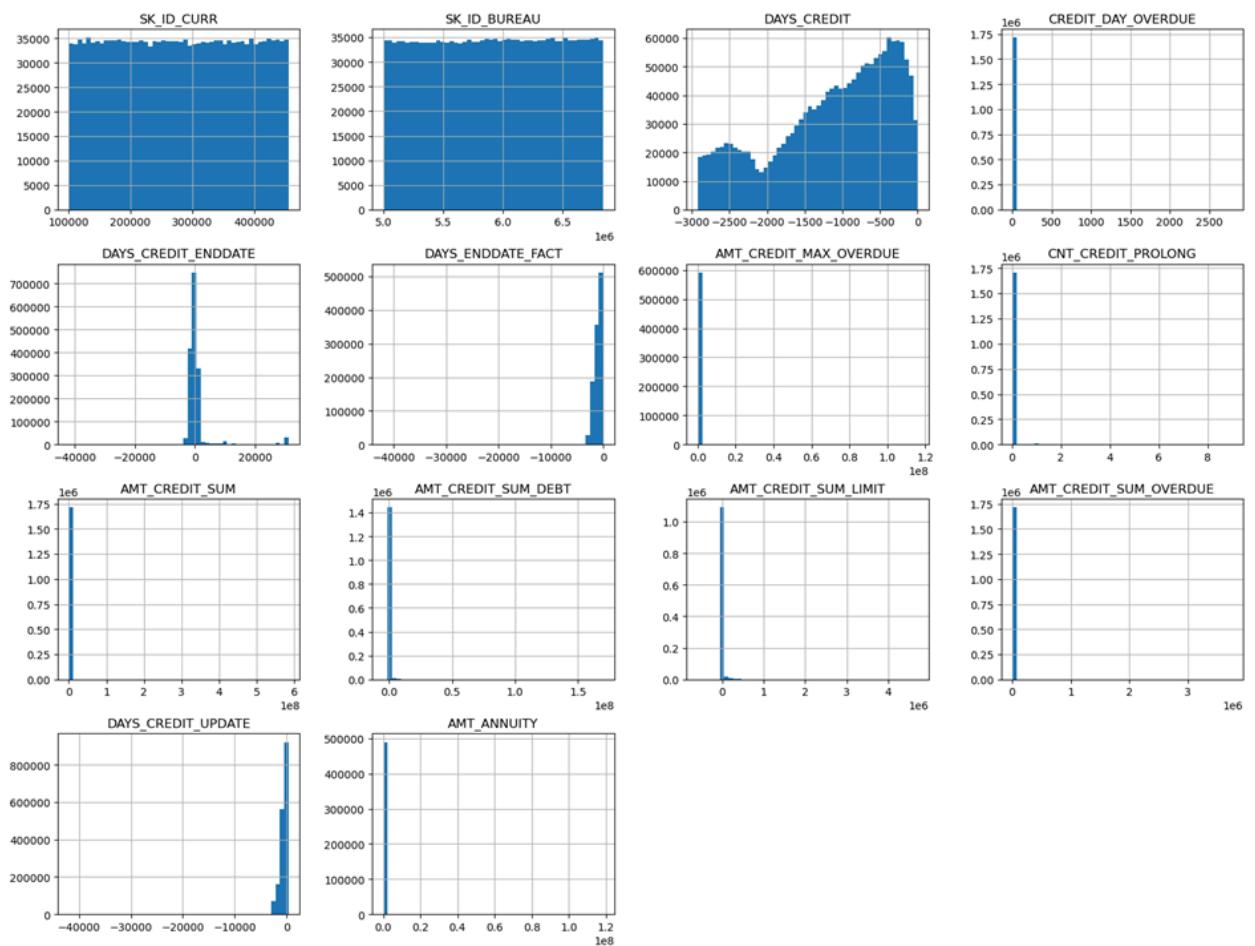
1. **Ketidakseimbangan Data:** Variabel numerik cenderung memiliki distribusi yang skewed ke kanan, artinya sebagian besar data terkonsentrasi pada nilai yang lebih rendah, sementara sedikit data memiliki nilai yang sangat tinggi.
2. **Adanya Outliers:** Hampir semua variabel memiliki outliers pada nilai yang sangat tinggi. Ini bisa mengindikasikan adanya data ekstrem atau kesalahan dalam data.
3. Untuk tipe data int juga memiliki nilai yang mendominasi.
4. kategori yang mendominasi pada tipe object(mode):
 - NAME_CONTRACT_TYPE(Cash loans), CODE_GENDER(F), FLAG_OWN_CAR(N), FLAG_OWN_REALTY(Y),
 - NAME_TYPE_SUITE(Unaccompanied), NAME_EDUCATION_TYPE(Secondary / secondary special), NAME_FAMILY_STATUS(Married),
 - NAME_HOUSING_TYPE(House / apartment).
5. **Dataset tidak seimbang:** Beberapa kategori memiliki frekuensi yang jauh lebih tinggi daripada kategori lainnya.

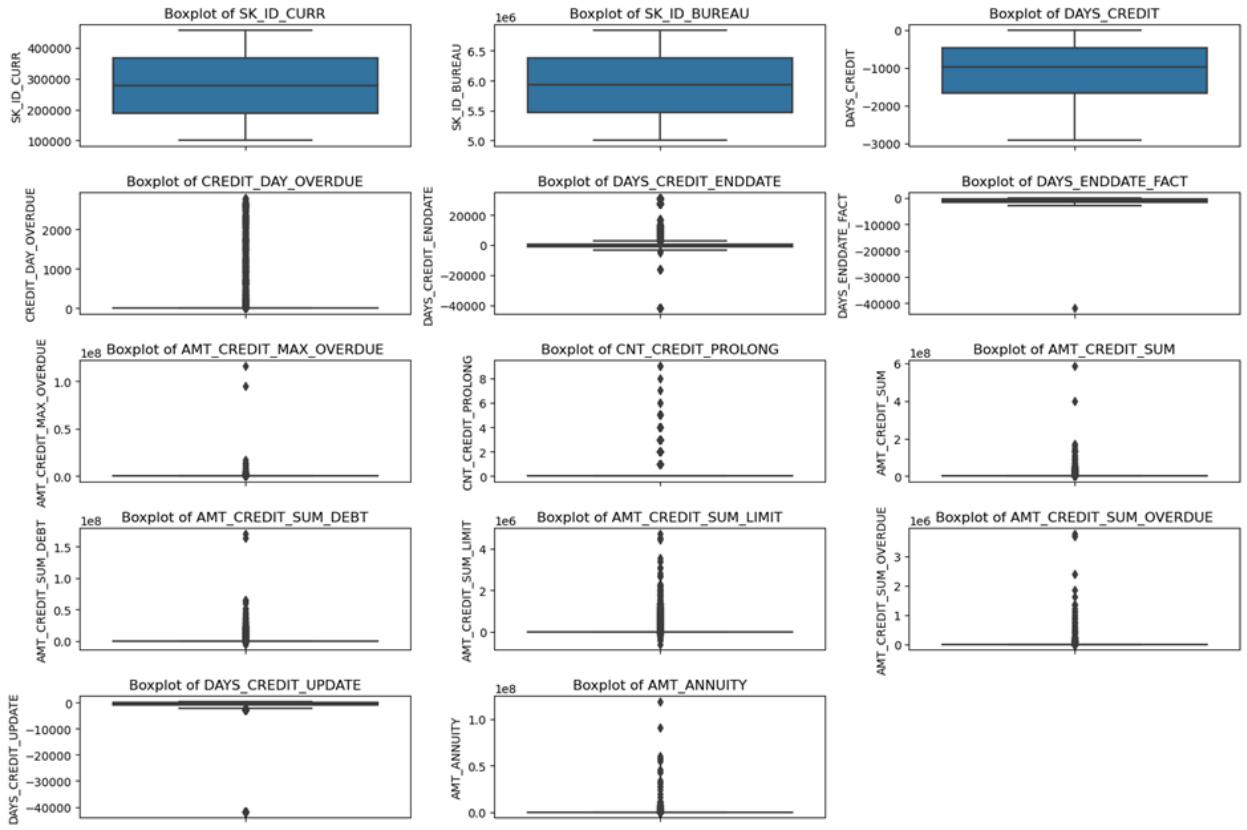
Hal-hal yang harus di-*follow up*:

- Analisis Statistik Deskriptif:** Menghitung mean, median, mode, standar deviasi, dan ukuran dispersi lainnya untuk setiap variabel.
- Visualisasi Data Lebih Lanjut:** Membuat visualisasi yang lebih kompleks, seperti heatmap dan lainnya untuk mendapatkan pemahaman yang lebih baik tentang data.
- Pembersihan Data:** Mengidentifikasi dan menangani outlier serta missing values.
- Analisis Korelasi:** Menganalisis hubungan antara variabel-variabel numerik.

2.1.2. HC_bureau

Histograms of Numeric Features in HC_bureau





1. SK_ID_CURR dan SK_ID_BUREAU

Distribusi merata (uniform) untuk kedua fitur ini. Ini menunjukkan bahwa ada banyak variasi ID yang berbeda dalam dataset, tanpa pola khusus. Ini merupakan fitur identifikasi unik, jadi distribusi merata seperti ini wajar dan diharapkan.

2. DAYS_CREDIT

Data ini memperlihatkan puncak distribusi pada nilai negatif. Nilai negatif ini menunjukkan seberapa lama sebelum aplikasi saat ini klien memiliki catatan kredit. Semakin mendekati 0, semakin baru catatan kredit tersebut. Distribusi menunjukkan ada lebih banyak klien dengan riwayat kredit baru (dengan nilai mendekati -3000 hingga 0).

3. CREDIT_DAY_OVERDUE

Kebanyakan nilai dalam fitur ini adalah nol, menunjukkan bahwa sebagian besar klien tidak memiliki kredit yang terlambat. Ada sedikit data di bagian atas yang menunjukkan beberapa kasus dengan nilai lebih besar dari 0, yang mungkin mewakili klien yang terlambat membayar.

4. DAYS_CREDIT_ENDDATE dan DAYS_ENDDATE_FACT

Keduanya memiliki distribusi yang sangat terkonsentrasi di sekitar nilai rendah.

DAY_S_CREDIT_ENDDATE menampilkan catatan kredit yang jatuh tempo di masa depan (sebagian besar nilai negatif), sedangkan

DAY_S_ENDDATE_FACT memperlihatkan kapan kredit tersebut benar-benar berakhir. Distribusi ini menunjukkan mayoritas data berada di nilai negatif atau sekitar nol, menandakan bahwa sebagian besar pelunasan kredit dilakukan sebelum jatuh tempo.

5. AMT_CREDIT_MAX_OVERDUE

Mayoritas nilai berada pada nol, menunjukkan bahwa sebagian besar klien tidak memiliki jumlah kredit yang tertunda. Hanya beberapa klien yang memiliki kredit tertunda.

6. CNT_CREDIT_PROLONG

Hampir semua nilai berada di nol, yang berarti bahwa sebagian besar klien tidak memperpanjang kredit mereka. Hanya sedikit kasus di mana kredit diperpanjang.

7. AMT_CREDIT_SUM, AMT_CREDIT_SUM_DEBT, AMT_ANNUITY, AMT_CREDIT_SUM_LIMIT, AMT_CREDIT_SUM_OVERDUE, dan

Nilai-nilai untuk fitur ini sangat terkonsentrasi di sisi kiri. Beberapa distribusi bahkan menunjukkan skewness (kemiringan) yang cukup tinggi, terutama untuk kolom seperti **AMT_CREDIT_SUM_DEBT** dan **AMT_CREDIT_SUM_LIMIT**, yang menunjukkan bahwa sebagian besar klien memiliki jumlah kredit, utang, atau batas kredit yang relatif kecil.

8. DAYS_CREDIT_UPDATE

Sebagian besar nilai berada pada rentang yang sangat sempit, menandakan bahwa catatan kredit sering diperbarui baru-baru ini (nilai negatif kecil), sementara hanya sedikit kasus dengan nilai negatif yang besar (artinya catatan lama tidak diperbarui dalam waktu lama).

Kesimpulan :

- Sebagian besar variabel yang berkaitan dengan waktu (seperti **DAY_S_CREDIT**, **DAY_S_CREDIT_ENDDATE**, dll.) terkonsentrasi pada nilai negatif, karena menggambarkan peristiwa di masa lalu (sebelum aplikasi).
- Variabel yang berkaitan dengan jumlah uang (seperti **AMT_CREDIT_SUM**, **AMT_CREDIT_SUM_DEBT**, dll.) menunjukkan skewness, di mana sebagian besar klien memiliki jumlah kredit atau utang yang relatif kecil.
- Banyak fitur seperti **CREDIT_DAY_OVERDUE** dan **CNT_CREDIT_PROLONG** memiliki sebagian besar nilai nol, yang berarti banyak klien yang tidak memiliki kredit terlambat atau memperpanjang kredit.

- Banyak dari fitur ini menunjukkan distribusi yang sangat terpusat dengan beberapa outliers, yang mungkin memerlukan penanganan lebih lanjut (misalnya, normalisasi atau pemangkasan outliers).
- Beberapa fitur seperti CREDIT_DAY_OVERDUE, CNT_CREDIT_PROLONG, dan AMT_CREDIT_SUM_OVERDUE memiliki banyak nilai nol, yang bisa jadi penting dalam analisis lanjutan.
- Dapat dilakukan transformasi data skewed untuk membuat distribusi lebih normal (misalnya menggunakan log transformation). Handling outliers untuk fitur-fitur dengan outlier yang signifikan seperti AMT_CREDIT_SUM dan AMT_CREDIT_SUM_DEBT. Feature engineering untuk menambahkan lebih banyak wawasan dari fitur-fitur yang terkait dengan waktu.

2.1.3. HC_credit_card_balance

Dari grafik boxplot terdapat beberapa feature yg memiliki outliers
 'MONTHS_BALANCE', 'AMT_BALANCE', 'AMT_CREDIT_LIMIT_ACTUAL',
 'AMT_DRAWINGS_ATM_CURRENT', 'AMT_DRAWINGS_CURRENT',
 'AMT_DRAWINGS_OTHER_CURRENT', 'AMT_DRAWINGS_POS_CURRENT',
 'AMT_INST_MIN_REGULARITY', 'AMT_PAYMENT_CURRENT',
 'AMT_PAYMENT_TOTAL_CURRENT', 'AMT_RECEIVABLE_PRINCIPAL',
 'AMT_RECVABLE', 'AMT_TOTAL_RECEIVABLE',
 'CNT_DRAWINGS_ATM_CURRENT', 'CNT_DRAWINGS_CURRENT',
 'CNT_DRAWINGS_OTHER_CURRENT', 'CNT_DRAWINGS_POS_CURRENT',
 'CNT_INSTALMENT_MATURE_CUM', 'SK_DPD', 'SK_DPD_DEF'

Feature-feature tersebut perlu dilakukan handling outliers lalu mengganti nilai-nilai outlier tersebut dengan mengisi median.

2.1.4. HC_installments_payments

- Kolom NUM_INSTALMENT_VERSION menunjukkan adanya positive skew yang sangat tajam dan terlihat ada data max yang sangat tinggi sehingga bisa dikatakan ada outlier. Pada pre-processing harus dilakukan outlier handling dan transformasi fitur.
- Kolom NUM_INSTALMENT_NUMBER menunjukkan adanya positive skew dan terlihat ada data max yang sangat tinggi sehingga bisa dikatakan ada outlier. Pada pre-processing harus dilakukan outlier handling dan transformasi fitur.
- Kolom DAYS_INSTALMENT menunjukkan adanya negative skew dan tidak terlalu ada nilai min/max yang sangat timpang. Pada pre-processing harus dilakukan transformasi fitur.
- Kolom DAYS_ENTRY_PAYMENT menunjukkan adanya negative skew juga tetapi pada kolom ini terlihat ada data dengan min yang sangat rendah sehingga bisa dikatakan ada outlier. Pada pre-processing harus dilakukan pengisian missing values, outlier handling dan transformasi fitur.

- Kolom AMT_INSTALMENT dan AMT_PAYMENT menunjukkan adanya positive skew yang sangat tajam dan terlihat ada data max yang sangat tinggi sehingga bisa dikatakan ada outlier. Pada pre-processing harus dilakukan outlier handling dan transformasi fitur dan untuk kolom AMT_PAYMENT dilakukan pengisian missing values.
- Kolom TARGET yang merupakan label menunjukkan distribusi bimodal dan terlihat adanya class imbalance kategori sedang karena masih diantara 1%-20%. Pada pre-processing harus dilakukan class imbalance handling dan pengisian missing values.

2.1.5. HC_POS_CASH_balance

1. CNT_INSTALMENT dan CNT_INSTALMENT_FUTURE:

Distribusi kedua kolom ini menunjukkan skewness ke kanan (positively skewed), yang berarti ada sejumlah besar nilai yang rendah dan beberapa nilai yang sangat tinggi. Hal ini mengindikasikan adanya outlier yang perlu ditangani saat pre-processing untuk menghindari bias pada model.

Follow-up: Lakukan analisis terhadap outlier dan pertimbangkan transformasi log atau trimming untuk mengurangi efek outlier pada model.

2. SK_DPD dan SK_DPD_DEF:

Kolom ini mengandung nilai tinggi yang menunjukkan keterlambatan dalam pembayaran. Distribusi ini juga sangat skewed ke kanan dengan sebagian besar nilai mendekati nol dan beberapa yang sangat tinggi, mengindikasikan beberapa nasabah yang menunggak pembayaran sangat lama.

Follow-up: Perlu ditentukan apakah perlu menangani nasabah yang memiliki keterlambatan ekstrim, misalnya dengan melakukan normalisasi atau mengelompokkan keterlambatan dalam kategori.

3. NAME_CONTRACT_STATUS:

Kolom ini memiliki distribusi yang cukup didominasi oleh beberapa kategori seperti "Active" dan "Completed". Ini berarti bahwa sebagian besar kontrak dalam dataset adalah kontrak aktif atau telah selesai, sementara kategori lainnya jarang muncul.

Follow-up: Pertimbangkan untuk mengelompokkan kategori yang jarang muncul ke dalam satu kategori "Lainnya" agar model tidak terlalu terbebani dengan kategori yang kurang representatif.

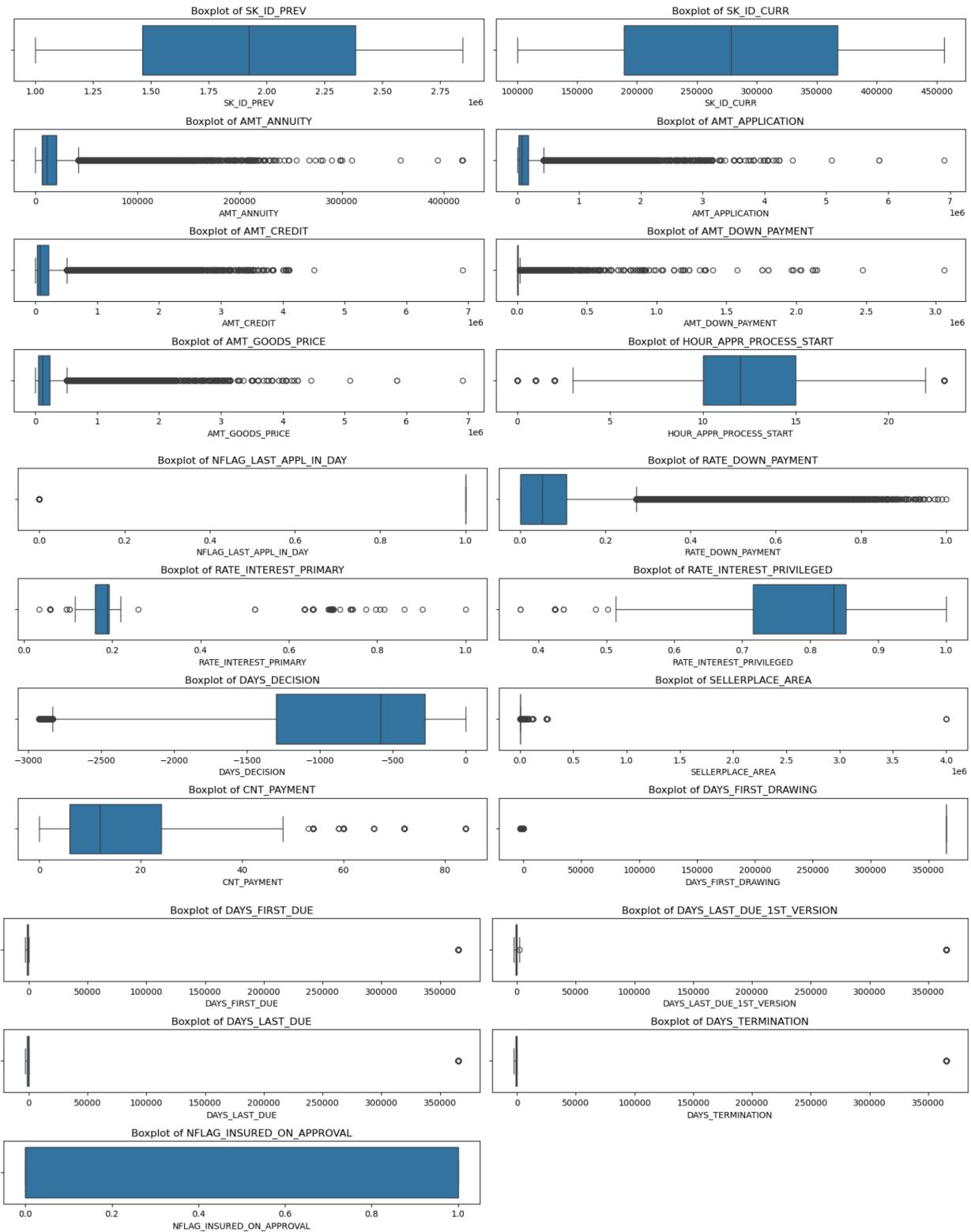
4. MONTHS_BALANCE:

Kolom ini menunjukkan bimodal distribution karena adanya data historis dalam jangka waktu yang panjang. Nilai pada kolom ini mewakili berbagai bulan dari -96 hingga -1, yang mencakup periode waktu yang luas.

Follow-up: Mempertimbangkan penggunaan lag features atau fitur agregasi (misalnya rata-rata per periode) untuk menangkap pola perubahan dari waktu ke waktu.

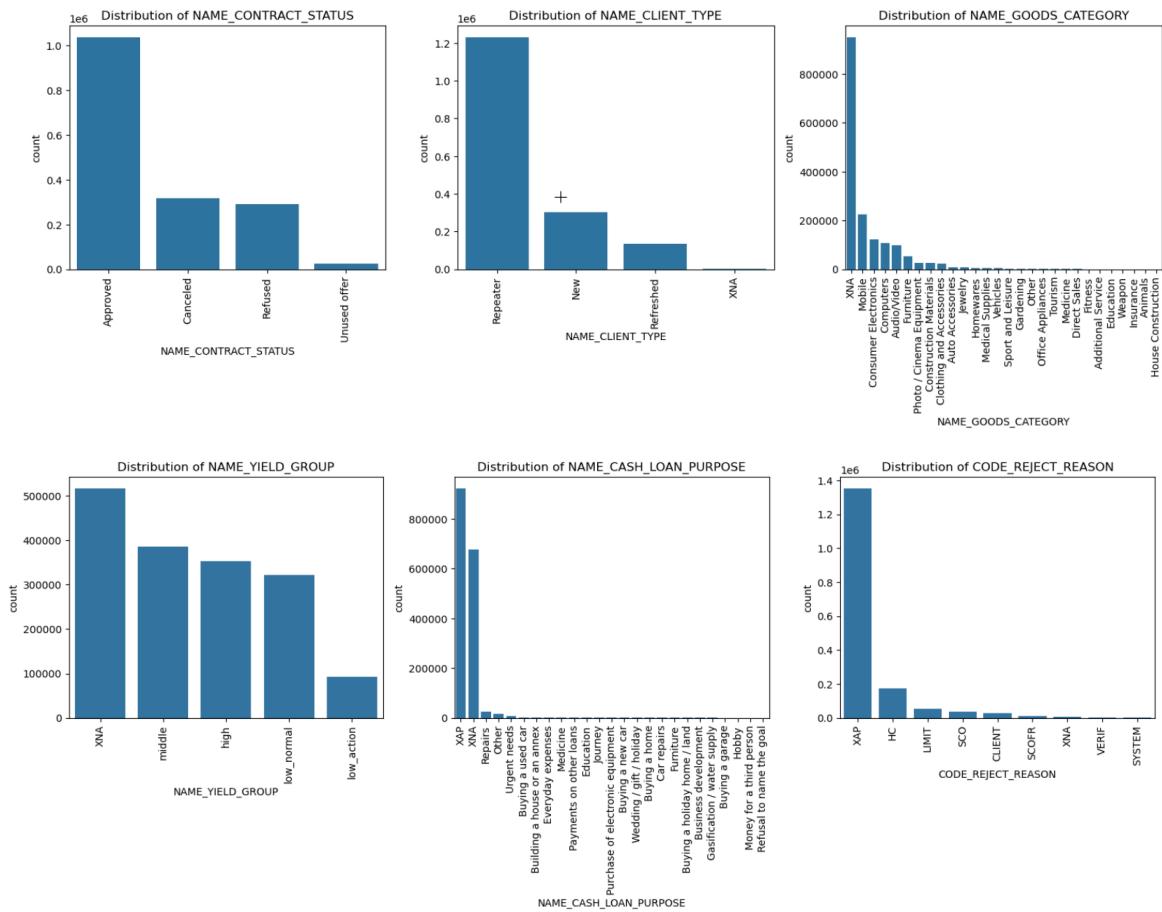
2.1.6. HC_previous_application

Univariate analysis pada data numerikal:



- Boxplot dari AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_DOWN_PAYMENT, AMT_GOODS_PRICE, RATE_DOWN_PAYMENT memiliki outlier yang sangat banyak.
- Boxplot dari SELLERPLACE_AREA, DAYS_FIRST_DRAWING, DAYS_FIRST_DUE, DAYS_LAST_DUE_1ST_VERSION, DAYS_LAST_DUE, DAYS_TERMINATION hampir tidak ada penggambaran apapun. Hal ini menunjukkan bahwa data pada kolom tersebut banyak bernilai kosong.

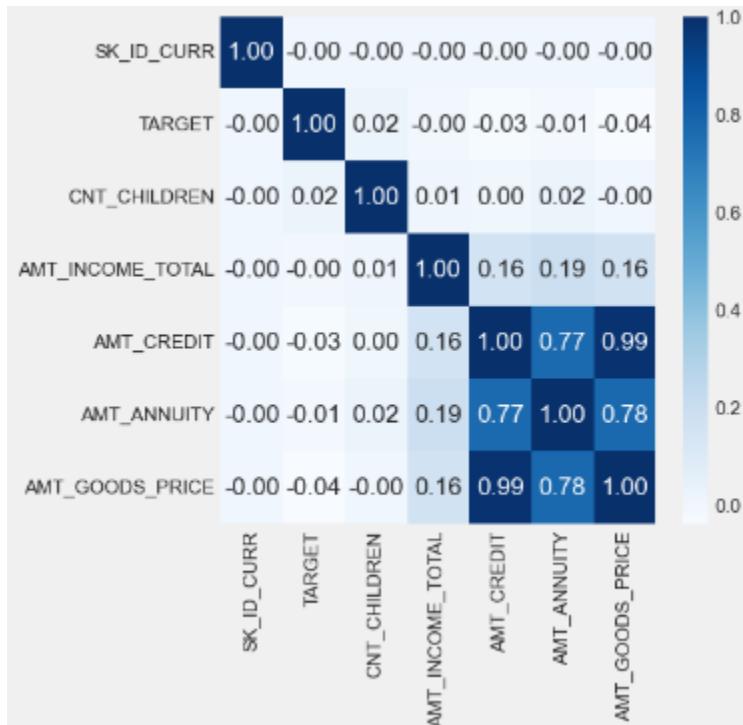
Univariate analysis dari data kategori:



- Data kategori yang dianalisis merupakan feature yang dipilih sekiranya akan memiliki pengaruh terhadap TARGET, yaitu feature NAME_CONTRACT_STATUS, NAME_CLIENT_TYPE, NAME_GOODS_CATEGORY, NAME_YIELD_GROUP, NAME_CASH_LOAN_PURPOSE, CODE_REJECT_REASON.
- Beberapa kode kategori belum diketahui artinya secara detail, seperti SCO, HC, SCOFR, dan sebagainya.
- Feature NAME_GOODS_CATEGORY dan NAME_CASH_LOAN_PURPOSE memiliki kategori yang sangat banyak.

2.2. Multivariate Analysis

2.2.1. HC_application_train | test



1. Bagaimana korelasi antara masing-masing feature dan label. Kira-kira feature mana saja yang paling relevan dan harus dipertahankan?
 - **Tidak ada korelasi yang sangat kuat antara TARGET dengan fitur-fitur lainnya.** Nilai koefisien korelasi antara TARGET dengan fitur-fitur lainnya cenderung sangat kecil, mendekati nol.
 - **Beberapa fitur menunjukkan korelasi yang sedikit negatif:** Ini berarti ketika nilai fitur tersebut meningkat, nilai TARGET cenderung menurun (sedikit). Namun, korelasinya sangat lemah.
2. Bagaimana korelasi antar-feature, apakah ada pola yang menarik? Apa yang perlu dilakukan terhadap feature itu?
 - A. AMT_CREDIT, AMT_ANNUITY, dan AMT_GOODS_PRICE: Ketiga variabel ini memiliki korelasi positif yang sangat kuat satu sama lain.
 - B. Sisanya, selain 3 variabel yang memiliki korelasi positif, memiliki koefisien korelasi yang sangat kecil, artinya tidak terdapat hubungan linier yang kuat antara dua variabel.
 - C. Apa yang perlu dilakukan terhadap fitur?
 - Korelasi yang sangat kuat antara AMT_CREDIT, AMT_ANNUITY, dan AMT_GOODS_PRICE dapat menyebabkan masalah multikolinearitas dalam

model regresi. Multikolinearitas dapat membuat model menjadi tidak stabil dan sulit diinterpretasikan.

- **Menghapus Salah Satu Fitur:** Anda bisa menghapus salah satu fitur yang sangat berkorelasi, misalnya AMT_GOODS_PRICE, karena informasi yang dikandungnya mungkin sudah terwakili oleh AMT_CREDIT dan AMT_ANNUITY.

2.2.2. HC_bureau

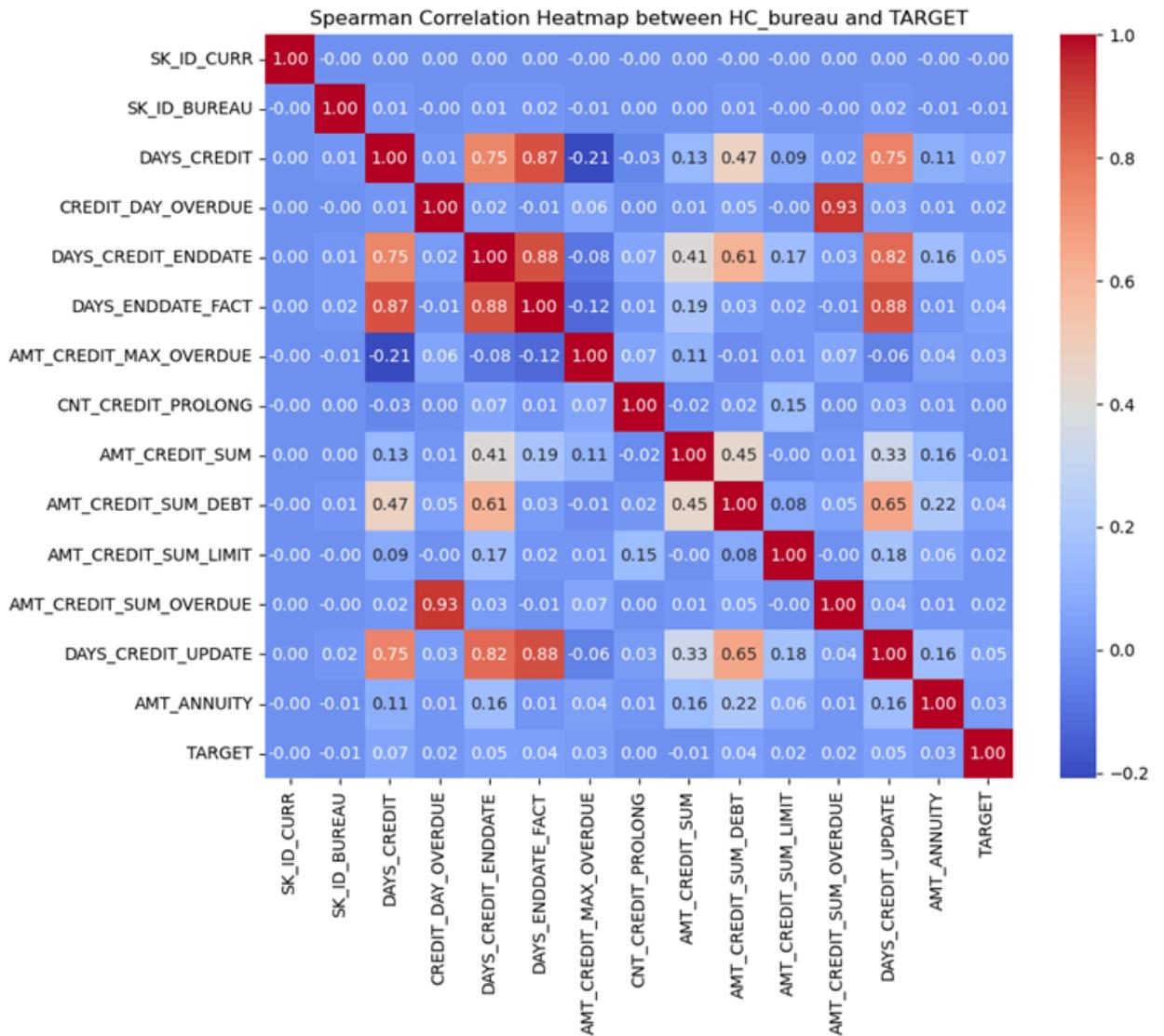
Multivariate Analysis ini berfungsi untuk menganalisa hubungan antara features/kolom pada Bureau Dataset sekaligus menjadi tahapan awal dalam proses features selection. Hasil korelasi pada analisis multivariate dapat menjadi informasi fitur mana yang paling berhubungan dengan tingkat default (gagal bayar). Telah dilakukan beberapa metode untuk mendapatkan analisis multivariatanya, antara lain :

- 1. Spearman Correlation Heatmap**
- 2. Recursive Feature Elimination (RFE)**
- 3. Tree-based Feature Importance**

Diuji coba metode Recursive Feature Elimination (RFE), namun tidak bisa di aplikasikan pada dataset Bureau sebelum dilakukan handling missing values. Hal ini dikarenakan, pada Bureau Dataset, missing values berkisar 0,0007% - 71,47% data, sedangkan metode RFE melakukan pemilihan fitur dengan melatih model berulang kali dan menghapus fitur berdasarkan kepentingannya. Model ini tidak dapat dilatih karena nilai yang hilang (missing values) hampir seluruh data.

Berdasarkan prinsip kerjanya, Tree-based Feature Importance juga tidak dapat diaplikasikan pada dataset Bureau sebelum dilakukan handling missing values. Tree-based Feature Importance (seperti Decision Trees, Random Forest, dan Gradient Boosting) mengabaikan NaN saat membuat keputusan, namun mereka tetap membutuhkan input yang lengkap untuk menghasilkan model yang valid. Mereka juga tidak dapat menghitung pentingnya fitur dengan benar jika ada nilai yang hilang.

Berikutnya dilakukan analisis multivariate menggunakan metode Spearman Correlation Heatmap. Analisis multivariate dilakukan untuk mendapatkan informasi hubungan antar fitur dalam dataset Bureau, dan menemukan fitur dataset Bureau yang paling berkaitan dengan TARGET (0 = tidak gagal bayar, 1 = gagal bayar) di dataset Application Train.



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

Berikut ini hasil korelasi antara features/kolom pada Bureau Dataset :

Very Strong Correlation

The range of a strong correlation value is between **0.80 and 1.00**

1. CREDIT_DAY_OVERDUE Vs AMT_CREDIT_SUM_OVERDUE (0.93)
2. DAYS_CREDIT_ENDDATE Vs DAYS_ENDDATE_FACT (0.88)
3. DAYS_CREDIT Vs DAYS_ENDDATE_FACT (0.87)
4. DAYS_ENDDATE_FACT Vs DAYS_CREDIT_UPDATE (0.87)
5. DAYS_CREDIT_ENDDATE Vs DAYS_CREDIT_UPDATE (0.81)

Strong Correlation

The range of a strong correlation value is between **0.60 and 0.79**.

1. DAYS_CREDIT Vs DAYS_CREDIT_ENDDATE (0.74)
2. DAYS_CREDIT Vs DAYS_CREDIT_UPDATE (0.74)
3. DAYS_CREDIT Vs DAYS_CREDIT_ENDDATE (0.74)
4. DAYS_CREDIT_ENDDATE Vs DAYS_CREDIT (0.74)
5. AMT_CREDIT_SUM_DEBT Vs DAYS_CREDIT_UPDATE (0.65)
6. DAYS_CREDIT_ENDDATE Vs AMT_CREDIT_SUM_DEBT (0.61)

Moderate Correlation

The range of a moderate correlation value is between **0.40 and 0.59**

1. DAYS_CREDIT Vs AMT_CREDIT_SUM_DEBT (0.46)
2. AMT_CREDIT_SUM Vs AMT_CREDIT_SUM_DEBT (0.45)

Weak Correlation

The range of a weak correlation value is between **0.20 and 0.39**

1. AMT_CREDIT_SUM_DEBT Vs AMT_ANNUITY (0.38)
2. AMT_CREDIT_SUM Vs DAYS_CREDIT_UPDATE (0.31)
3. DAYS_CREDIT Vs AMT_CREDIT_MAX_OVERDUE (-0.20)
4. DAYS_CREDIT_UPDATE Vs AMT_ANNUITY (0.28)
5. DAYS_CREDIT_ENDDATE Vs AMT_ANNUITY (0.27)
6. AMT_CREDIT_SUM Vs AMT_ANNUITY (0.25)

* The correlation with a value below the weak correlation threshold will be ignored.

Berikut hasil observasi terkait korelasi antara fitur di HC_bureau dan TARGET:

- DAYS_CREDIT memiliki korelasi 0.07 dengan TARGET, yang berarti ada sedikit korelasi positif antara waktu kredit terakhir dan kemungkinan gagal bayar (TARGET = 1). Artinya, semakin dekat seseorang mendapatkan kredit (semakin besar nilai DAYS_CREDIT), ada sedikit peningkatan kemungkinan gagal bayar.
- AMT_CREDIT_SUM_DEBT memiliki korelasi 0.22 dengan TARGET, menunjukkan hubungan yang lebih kuat. Ini artinya, semakin tinggi jumlah utang (AMT_CREDIT_SUM_DEBT), semakin tinggi kemungkinan gagal bayar. Fitur ini bisa sangat penting untuk model prediksi gagal bayar.
- AMT_CREDIT_SUM_OVERDUE juga memiliki korelasi positif sebesar 0.16, artinya semakin banyak jumlah kredit yang tertunggak, semakin besar peluang untuk gagal bayar.
- AMT_CREDIT_SUM memiliki korelasi 0.16 dengan TARGET, yang menunjukkan bahwa semakin besar jumlah kredit total yang dimiliki nasabah, semakin tinggi kemungkinan gagal bayar.

- **DAY_CREDIT_UPDATE** memiliki korelasi positif sebesar 0.16 dengan TARGET, menunjukkan bahwa nasabah yang baru saja memperbarui informasi kredit mereka sedikit lebih mungkin untuk gagal bayar.

Berdasarkan analisa multivariate menggunakan metode Spearman Correlation Heatmap, maka dipilih features yang akan dipertahankan berdasarkan nilai korelasi cukup tinggi antara sesama features, atau berkorelasi tinggi dengan TARGET (tujuan bisnis). Features tersebut antara lain :

- DAY_CREDIT** : Jumlah hari sejak kredit terakhir diberikan atau diperbarui
- AMT_CREDIT_SUM_OVERDUE** : Jumlah kredit yang tertunggak atau belum dibayar tepat waktu.
- AMT_CREDIT_SUM_DEBT** : Jumlah kredit yang masih terhutang (belum dibayar) oleh nasabah untuk kredit yang tercatat.
- AMT_CREDIT_SUM** : Total jumlah kredit yang dipinjamkan ke nasabah untuk kredit yang tercatat.
- DAY_CREDIT_ENDDATE** : Jumlah hari hingga kredit seharusnya lunas.
- DAY_ENDDATE_FACT** : Jumlah hari dari tanggal aplikasi hingga kredit tersebut sebenarnya dilunasi.
- CREDIT_DAY_OVERDUE** : Jumlah hari keterlambatan pembayaran kredit saat ini.

2. Bagaimana korelasi antar-feature, apakah ada pola yang menarik? Apa yang perlu dilakukan terhadap feature itu?

- Masih perlu dilakukan proses pre-processing, dikarenakan terdapat banyak missing values mencapai 0,0007% - 71,47% data. Hasil korelasi menggunakan metode Spearman Correlation Heatmap ini hanya sebagai gambaran awal, karena hasil bisa saja berubah apabila telah dilakukan cleansing process atau proses pre-processing.
- Beberapa metode seperti Recursive Feature Elimination (RFE), Tree-based Feature Importance (seperti Decision Trees, Random Forest, dan Gradient Boosting) dapat dilakukan kembali menggunakan data yang sudah dibersihkan.
- Tidak ditemukan korelasi yang kuat antara features pada Bureau Dataset dan TARGET pada Application_Train
- Perlu dilakukan pengecekan kembali apakah ada kesalahan dalam data, dsb. Hal ini dikarenakan masih banyak ketimpangan data pada sample feature dari Application Train, seperti masih terdapat outliers, serta missing values.

2.2.3. HC_credit_card_balance

Berdasarkan heatmap, didapatkan nilai korelasi antar feature maka ada beberapa feature yang cukup menarik dan diperlukan untuk dijadikan insight yaitu feature-feature berikut ini :

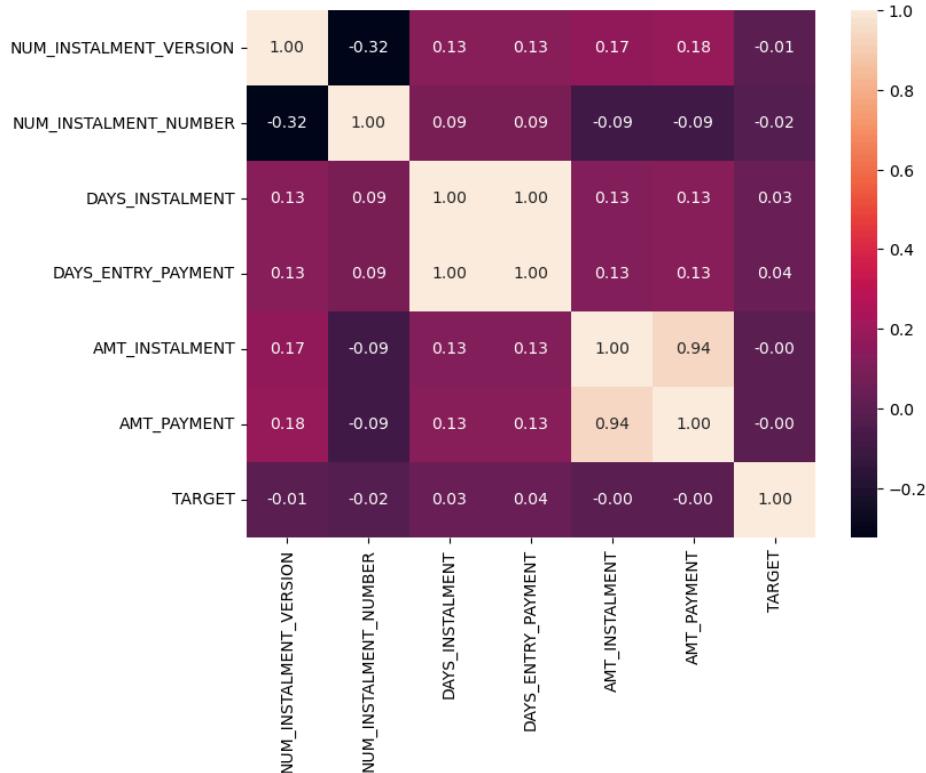
'MONTHS_BALANCE', 'AMT_BALANCE', 'AMT_CREDIT_LIMIT_ACTUAL',
'AMT_DRAWINGS_ATM_CURRENT', 'AMT_DRAWINGS_CURRENT',
'AMT_DRAWINGS_OTHER_CURRENT', 'AMT_PAYMENT_TOTAL_CURRENT',
'AMT_RECEIVABLE', 'AMT_TOTAL_RECEIVABLE',
'CNT_DRAWINGS_ATM_CURRENT', 'CNT_DRAWINGS_CURRENT',
'CNT_DRAWINGS_OTHER_CURRENT', 'CNT_DRAWINGS_POS_CURRENT',
'NAME_CONTRACT_STATUS', 'SK_DPD', 'SK_DPD_DEF'.

Ada juga beberapa feature yang perlu dihilangkan seperti CNT_INSTALMENT_MATURE_CUM, AMT_RECEIVABLE_PRINCIPAL, AMT_RECEIVABLE, AMT_PAYMENT_CURRENT

Ada beberapa alasan menghilangkan feature-feature tersebut :

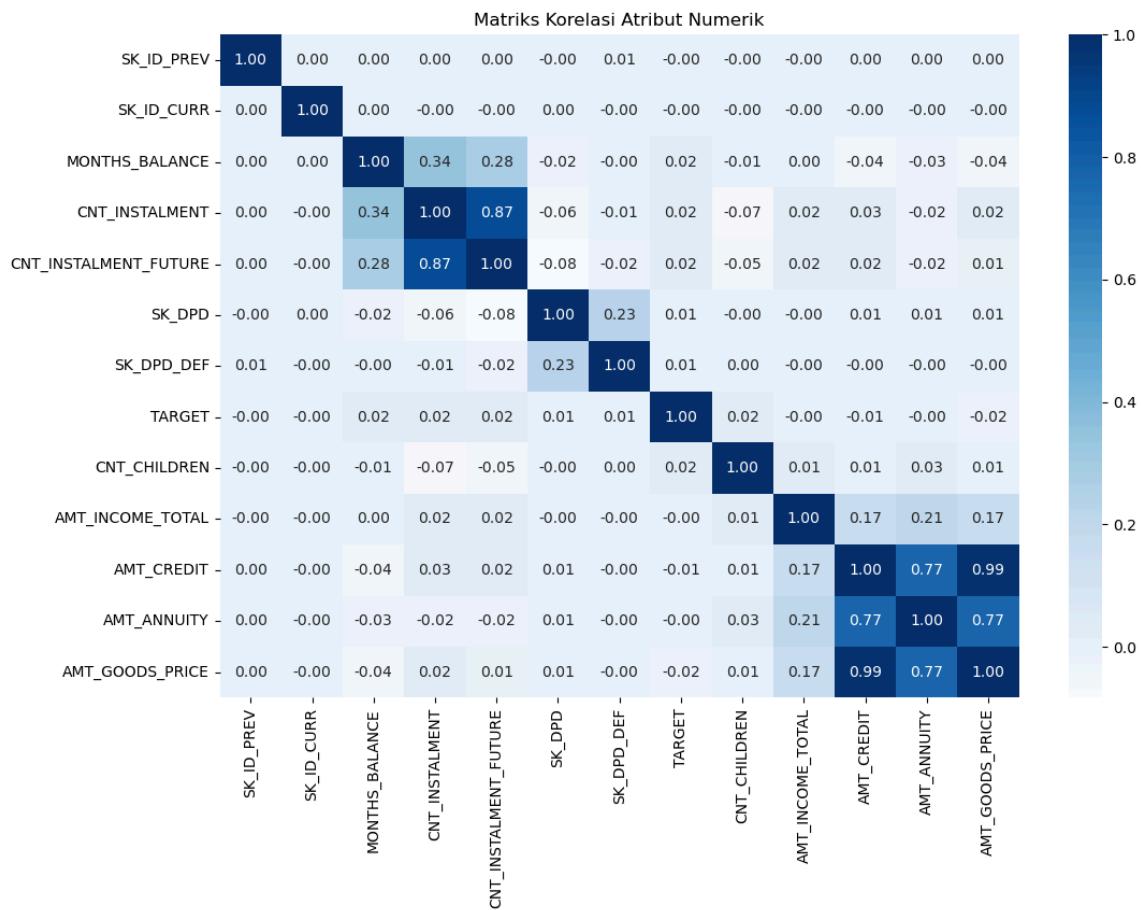
Dikarenakan tidak memiliki nilai korelasi dengan feature lainnya(Nilai korelasi dibawah 0.20) dan memiliki sifat antara feature yang artinya sama dengan feature lainnya (Redundant)

2.2.4. HC_installments_payments



1. Tidak ada korelasi antara fitur dengan target, fitur yang dapat dipertahankan adalah:
 - NUM_INSTALMENT_VERSION
 - NUM_INSTALMENT_NUMBER
 - DAYS_ENTRY_PAYMENT
 - AMT_PAYMENT
 - TARGET
2. Ada korelasi antara fitur yang kuat (>0.7) yaitu DAYS_INSTALMENT dengan DAYS_ENTRY_PAYMENT dan AMT_INSTALMENT dengan AMT_PAYMENT. Karena ada fitur yang redundan, maka akan ada fitur yang tidak dipakai.

2.2.5. HC_POS_CASH_balance



1. CNT_INSTALMENT dan CNT_INSTALMENT_FUTURE:

- Fitur seperti CNT_INSTALMENT dan CNT_INSTALMENT_FUTURE menunjukkan korelasi positif yang cukup tinggi. Ini mungkin karena kedua fitur ini menggambarkan jumlah angsuran yang berhubungan langsung, baik yang sudah dilakukan maupun yang akan datang.

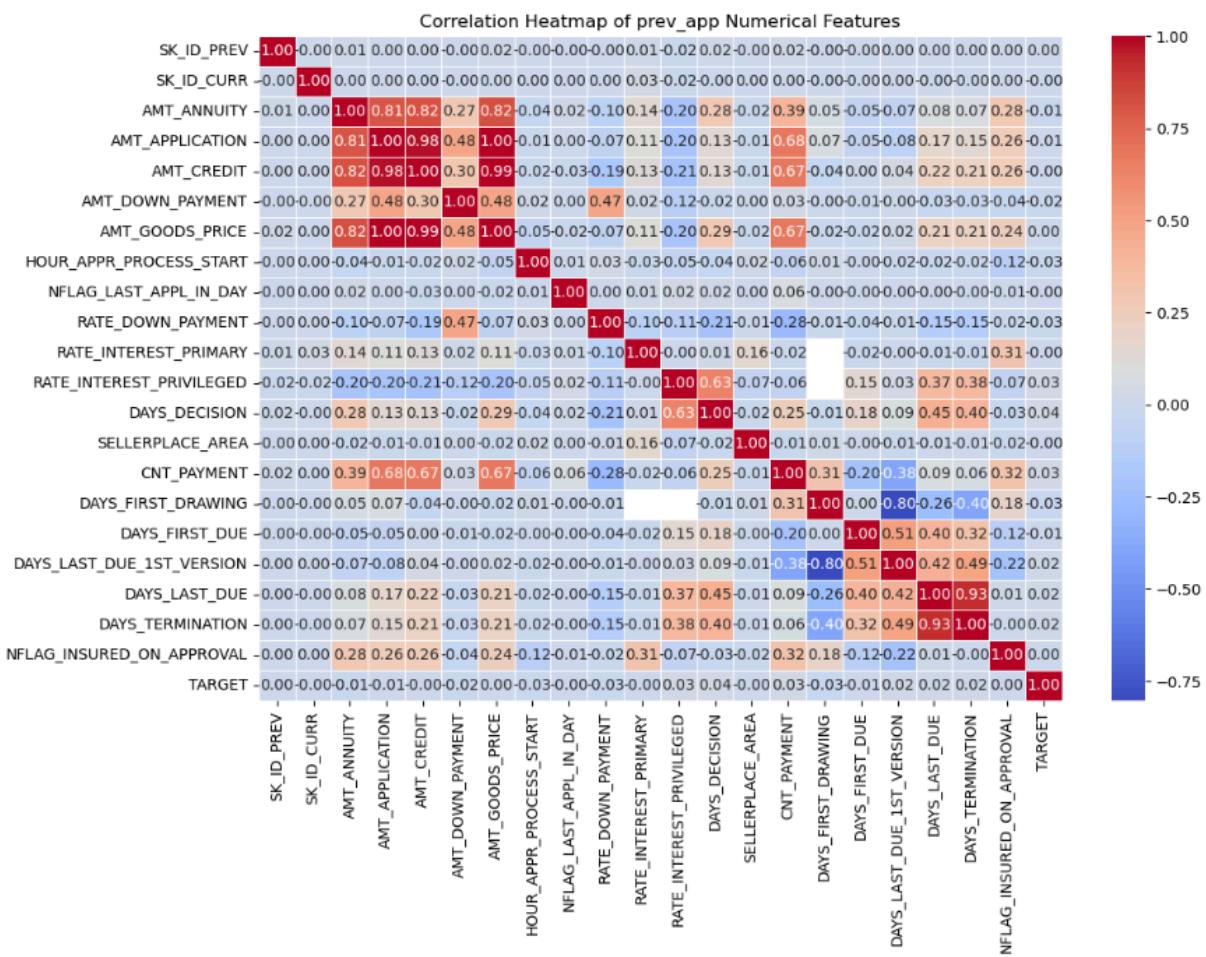
- Follow-up: Karena kedua fitur tersebut berkaitan langsung dan menunjukkan korelasi tinggi, kita bisa mempertimbangkan untuk membuat fitur gabungan yang merepresentasikan total angsuran, atau memilih salah satu dari fitur ini untuk dimasukkan dalam model.

2. NAME_CONTRACT_STATUS:

- Melakukan visualisasi menggunakan count plots atau bar plots menunjukkan bahwa kategori "Active" dan "Completed" mendominasi, sementara kategori lainnya muncul dalam jumlah yang jauh lebih kecil.
- Follow-up: Mengelompokkan kategori-kategori yang jarang muncul ke dalam satu kategori "Lainnya" dapat mengurangi jumlah kategori tanpa kehilangan banyak informasi. Ini juga dapat membantu mengurangi risiko overfitting pada model.

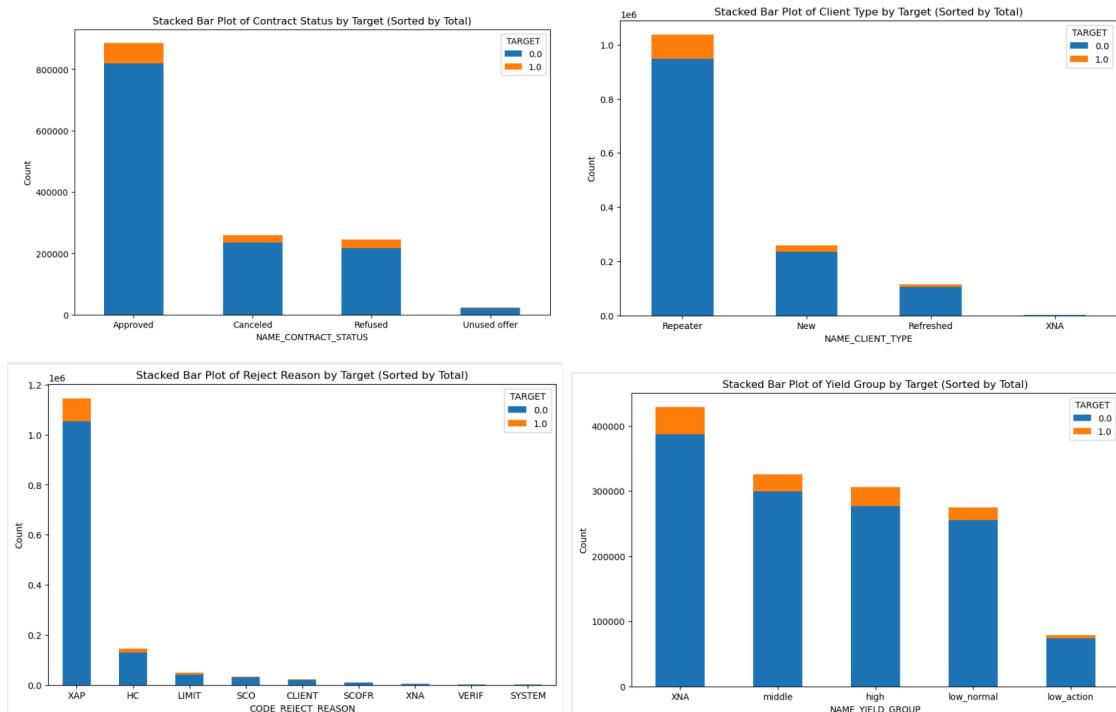
2.2.6. HC_previous_application

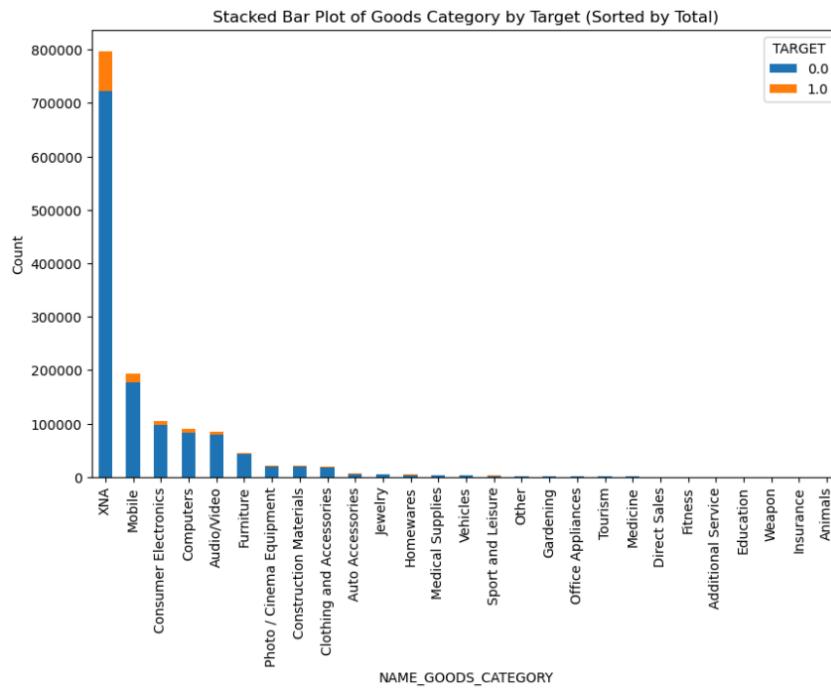
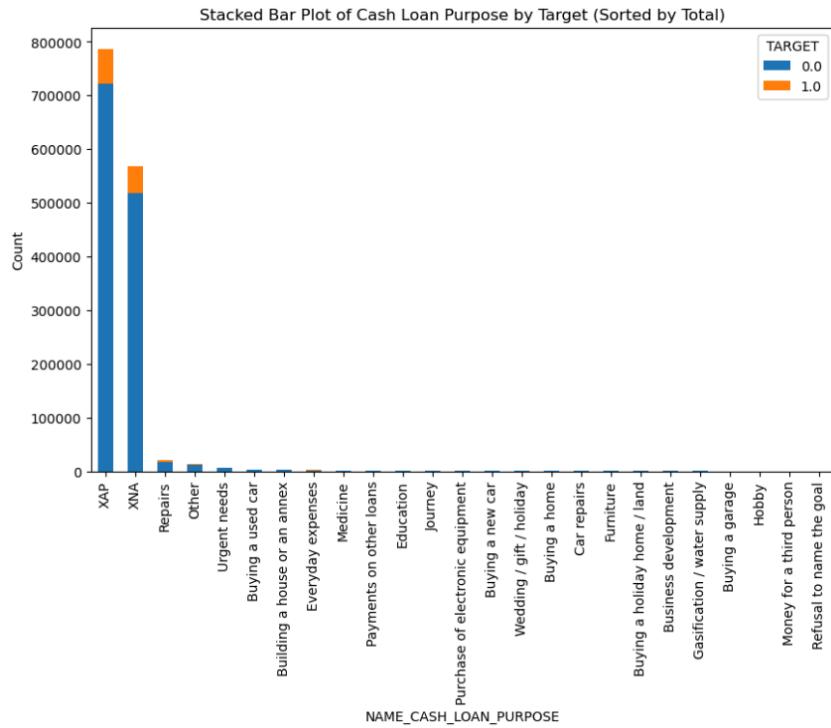
Multivariate analysis pada data numerik:



1. Berdasarkan korelasi dengan heatmap, semua feature numerical pada previous_application tidak ada yang memiliki hubungan dengan TARGET.
2. Terdapat kotak kosong putih, yang menunjukkan bahwa terlalu banyak nilai kosong sehingga tidak dapat dianalisis korelasinya yaitu pada feature:
 - DAYS_FIRST_DRAWING dan RATE_INTEREST_PRIMARY
 - DAYS_FIRST_DRAWING dan RATE_INTEREST_PRIVILEGED
3. Terdapat 2 feature yang berkorelasi kuat (>0.7) yaitu:
 - AMT_APPLICATION dan AMT_CREDIT
 - AMT_APPLICATION dan AMT_GOODS_PRICE
 - AMT_APPLICATION dan AMT_ANNUITY
 - AMT_ANNUITY dan AMT_CREDIT
 - AMT_ANNUITY dan AMT_GOODS_PRICE
 - AMT_CREDIT dan AMT_GOODS_PRICE
 - DAYS_LAST_DUE dan DAYS_TERMINATION
4. Kemungkinan besar pasangan feature yang memiliki korelasi kuat merupakan data redundant.

Multivariate pada data kategori:





- Berdasarkan hubungan feature categorical menggunakan stacked bar plot yang relevan pada previous_application dengan TARGET, bahwa:
 - NAME_CONTRACT_STATUS : approved ataupun tidak di-approved, tetap tidak mempengaruhi TARGET karena banyak kontrak yang di-approve tetapi ditolak

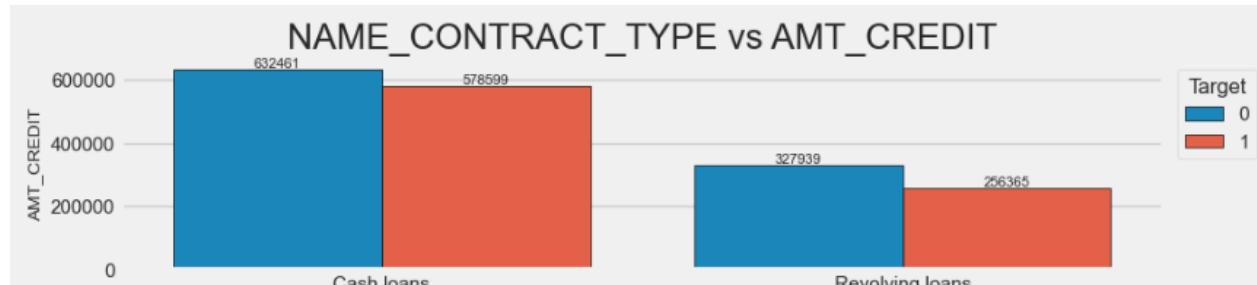
- pada target sedangkan pada status kontrak yang ditolak terdapat yang diterima pada TARGET.
- NAME_CLIENT_TYPE: kemungkinan besar tipe client yang telah repeat akan diterima pada TARGET.
 - CODE_REJECT_REASON: kategori XAP jauh mendominasi daripada kategori lain dan terlihat paling banyak diterima pada TARGET.
 - NAME_YIELD_GROUP: kategori high dan XNA merupakan kategori yang banyak diterima pada TARGET.
 - NAME_CASH_LOAN_PURPOSE: kategori XAP dan XNA sangat jauh mendominasi daripada kategori lain, walaupun banyak sekali kategori.
 - NAME_GOODS_CATEGORY: kategori yang paling banyak adalah XNA, sangat jauh jumlahnya daripada kategori lainnya. Kemungkinan tidak berpengaruh terhadap TARGET.

2.3. Business Insights and Visualizations

2.3.1. HC_application_train | test

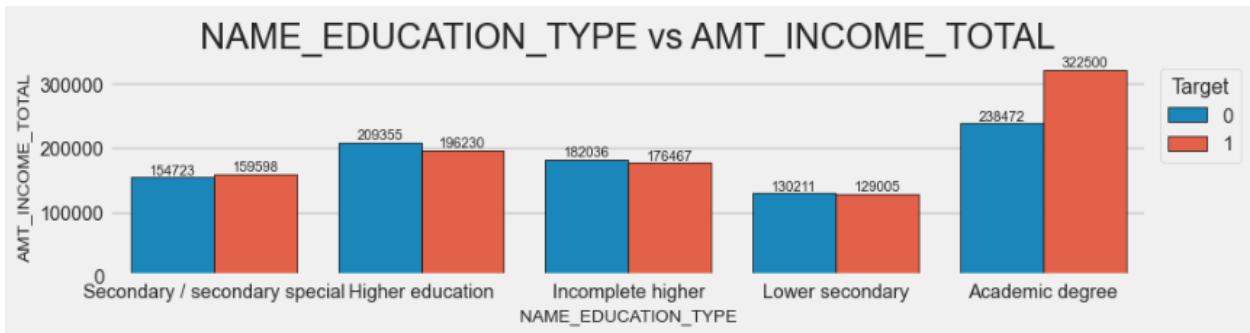
Business Insight:

1. Name Contract and Amount Credit



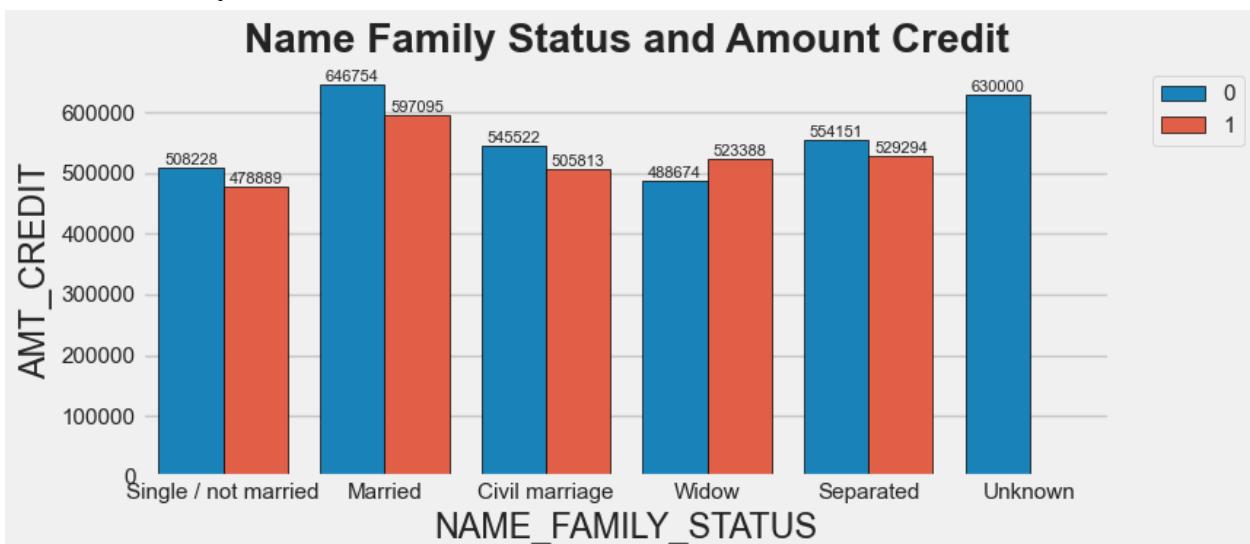
- Cash loans* Dominan: Jumlah total kredit yang diberikan untuk jenis *cash loans* jauh lebih besar dibandingkan dengan *revolving loans*. Ini menunjukkan bahwa mayoritas pelanggan cenderung memilih *cash loans*.
- Tingkat default:
 - ***Cash loans***: Baik untuk pelanggan yang pinjamannya tidak macet maupun yang macet, *cash loans* memiliki jumlah total kredit yang lebih besar.
 - ***Revolving loans***: Jumlah total kredit untuk *revolving loans* yang macet lebih kecil dibandingkan dengan *cash loans* yang macet. Ini bisa mengindikasikan bahwa *revolving loans* memiliki tingkat kemacetan yang relatif lebih rendah dibandingkan dengan *cash loans*.

2. Name Education Type and Amount Income Total



- A. **Pendidikan dan Pendapatan:** Secara umum, grafik menunjukkan korelasi positif antara tingkat pendidikan dan total pendapatan. Individu dengan tingkat pendidikan yang lebih tinggi cenderung memiliki pendapatan yang lebih tinggi.
- B. Perbedaan berdasarkan target:
 - **Nasabah dengan Target 0:** Nasabah dengan target 0 (menunjukkan nasabah yang baik) cenderung memiliki pendapatan yang lebih tinggi, terutama pada kelompok dengan tingkat pendidikan "Higher education" dan "Academic degree".
 - **Nasabah dengan Target 1:** Nasabah dengan target 1 (menunjukkan nasabah yang berisiko) juga menunjukkan tren yang serupa, namun dengan nilai yang lebih rendah, akan tetapi pada nasabah dengan tingkat pendidikan "Academic degree" dengan pendapatan yang lebih tinggi memiliki resiko gagal bayar yang lebih besar.

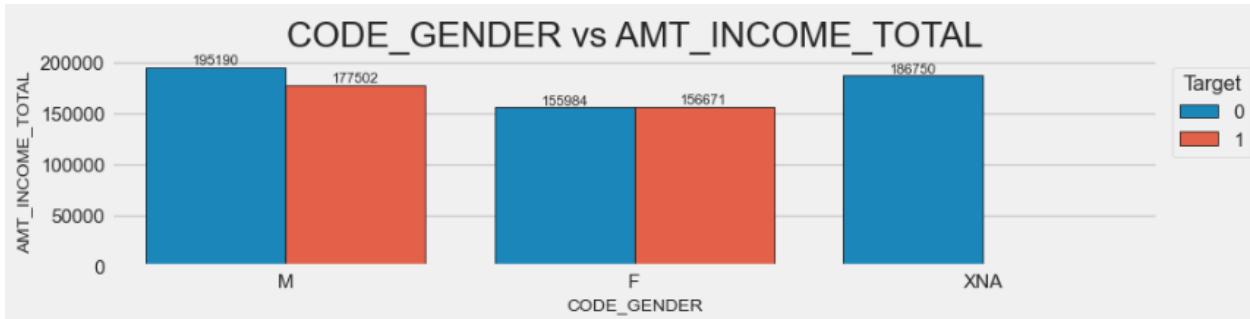
3. Name Family Status and Amount Credit



- A. Terdapat perbedaan yang cukup signifikan pada jumlah kredit yang diberikan untuk setiap status perkawinan. Misalnya, nasabah yang sudah menikah cenderung memiliki jumlah kredit yang lebih tinggi dibandingkan dengan nasabah yang belum menikah atau janda/duda.
- B. Perbedaan target:

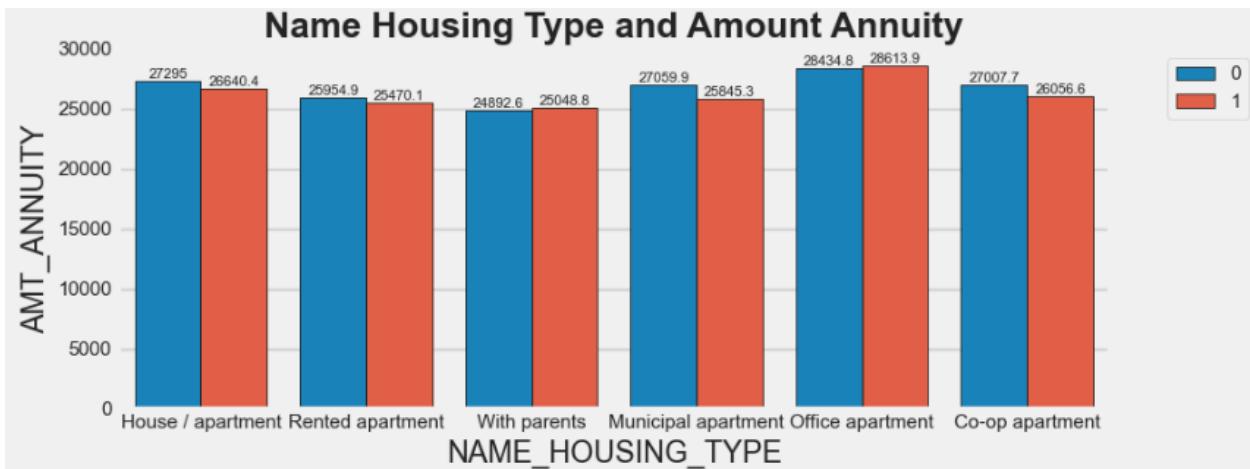
- **Nasabah dengan Target 0:** Nasabah dengan target 0 (menunjukkan nasabah yang baik) umumnya memiliki jumlah kredit yang lebih tinggi dibandingkan dengan nasabah dengan target 1, kecuali nasabah Janda.
- **Interaksi dengan Status Perkawinan:** Ada interaksi antara status perkawinan dan target. Misalnya, pada kelompok "Married", terlihat perbedaan yang cukup signifikan antara jumlah kredit untuk target 0 dan target 1.

4. Code Gender and Amount Income Total



- A. **Perbedaan Pendapatan Berdasarkan Jenis Kelamin:** Terdapat perbedaan pendapatan antara laki-laki (M), perempuan (F), dan kategori "XNA" (yang mungkin mewakili data yang tidak diketahui atau kategori lain). Secara umum, laki-laki memiliki pendapatan yang sedikit lebih tinggi dibandingkan dengan perempuan. Namun, kategori "XNA" juga memiliki pendapatan yang cukup tinggi.
- B. **Interaksi dengan Jenis Kelamin:** Ada interaksi antara jenis kelamin dan target. Misalnya, pada kelompok laki-laki, terlihat perbedaan yang cukup signifikan antara jumlah pendapatan untuk target 0 (baik) dan target 1 (beresiko).

5. Name Housing Type and Amount Annuity



- A. Terdapat perbedaan yang cukup signifikan pada jumlah angsuran bulanan untuk setiap jenis tempat tinggal. Misalnya, nasabah yang tinggal di rumah/apartemen sendiri cenderung memiliki jumlah angsuran yang lebih tinggi dibandingkan dengan nasabah yang tinggal bersama orang tua. Dan untuk nasabah yang memiliki jumlah angsuran bulanan terbesar terdapat pada nasabah yang tinggal di apartemen kantor.
- B. **Nasabah dengan Target 0:** Nasabah dengan target 0 (menunjukkan nasabah yang baik) umumnya memiliki jumlah angsuran yang lebih tinggi dibandingkan dengan nasabah dengan target 1. Kecuali pada nasabah yang tinggal dengan orang tua dan

tinggal di apartemen kantor, target 0 memiliki sedikit lebih rendah dibanding target 1(beresiko).

- C. Ada interaksi antara jenis tempat tinggal dan target. Misalnya, pada kelompok "municipal apartment"(apartemen yang dikelola pemerintah/rusun), terlihat perbedaan yang cukup signifikan antara jumlah angsuran untuk target 0 dan target 1.

Business Recommendation:

1. Fokus pada *cash loans*:

- **Optimasi Produk:** Karena *cash loans* merupakan produk unggulan, perusahaan dapat fokus pada optimasi produk ini, seperti memperluas jangkauan, meningkatkan fitur, atau menyesuaikan suku bunga.
- **Manajemen Risiko:** Meskipun dominan, perusahaan perlu tetap memperhatikan manajemen risiko pada pinjaman tunai, terutama terkait dengan tingkat kemacetan.

2. Analisis Lebih Dalam pada Pinjaman Berputar:

- **Potensi Pertumbuhan:** Meskipun jumlah total kredit lebih kecil, potensi pertumbuhan untuk pinjaman berputar mungkin lebih besar. Perusahaan dapat mempertimbangkan strategi untuk meningkatkan pangsa pasar pada jenis pinjaman ini.

3. Program Pengembangan Pelanggan:

- **Pendidikan Keuangan:** Menyediakan program pendidikan keuangan yang disesuaikan dengan tingkat pendidikan nasabah. Hal ini dapat membantu meningkatkan literasi keuangan nasabah dan mengurangi risiko kredit.
- **Program Loyalitas:** Menawarkan program loyalitas yang menarik bagi nasabah dengan tingkat pendidikan tinggi untuk meningkatkan retensi pelanggan.

4. Analisis Resiko Kredit:

- Kebijakan Kredit: Menyesuaikan kebijakan kredit berdasarkan tingkat pendidikan. Misalnya, memberikan persyaratan kredit yang lebih fleksibel untuk nasabah dengan tingkat pendidikan yang lebih tinggi.

5. Penawaran Produk:

- Perusahaan dapat menawarkan produk dan layanan yang lebih disesuaikan dengan status perkawinan nasabah. Misalnya, nasabah yang sudah menikah mungkin lebih tertarik pada produk pinjaman untuk perumahan.

6. Pengembangan Produk Baru:

- **Produk Keluarga:** Mengembangkan produk atau layanan yang ditujukan khusus untuk keluarga, seperti pinjaman untuk renovasi rumah atau pendidikan anak.

7. Strategi pemasaran dapat disesuaikan berdasarkan jenis kelamin. Misalnya, menggunakan pesan yang menekankan pada status sosial untuk laki-laki,

- sedangkan menggunakan pesan yang menekankan pada keamanan finansial untuk perempuan.
8. Menawarkan program loyalitas yang berbeda untuk setiap jenis kelamin. Misalnya, program yang memberikan diskon khusus untuk produk tertentu bagi perempuan.
 9. **Produk Terkait Tempat Tinggal:** Mengembangkan produk atau layanan yang terkait dengan tempat tinggal, seperti pinjaman untuk renovasi rumah atau asuransi properti.
 10. **Program Loyalitas:** Menawarkan program loyalitas yang berbeda untuk setiap jenis tempat tinggal. Misalnya, program yang memberikan diskon khusus untuk produk tertentu bagi nasabah yang memiliki rumah.

2.3.2. HC_bureau

1. Analisa berdasarkan *business metrics*

Salah satu analisa bisnis insight yang dilakukan yaitu menghitung *default rate* menggunakan features pada Bureau Dataset. Default Rate ini adalah business metrics yang dipilih untuk mengukur kesuksesan permodelan sistem (Machine Learning) Home Credit. Adapun Default Rate didapat menggunakan rumus sebagai berikut :

$$\text{Default Rate} = \frac{\text{Jumlah Pinjaman yang Gagal Bayar}}{\text{Total Pinjaman yang Diberikan}}$$

Sample features/kolom yang dipilih yaitu :

AMT_CREDIT_SUM (TOTAL PINJAMAN)

Total jumlah kredit yang dipinjamkan ke nasabah untuk kredit yang tercatat.

AMT_CREDIT_SUM_OVERDUE (PINJAMAN GAGAL BAYAR)

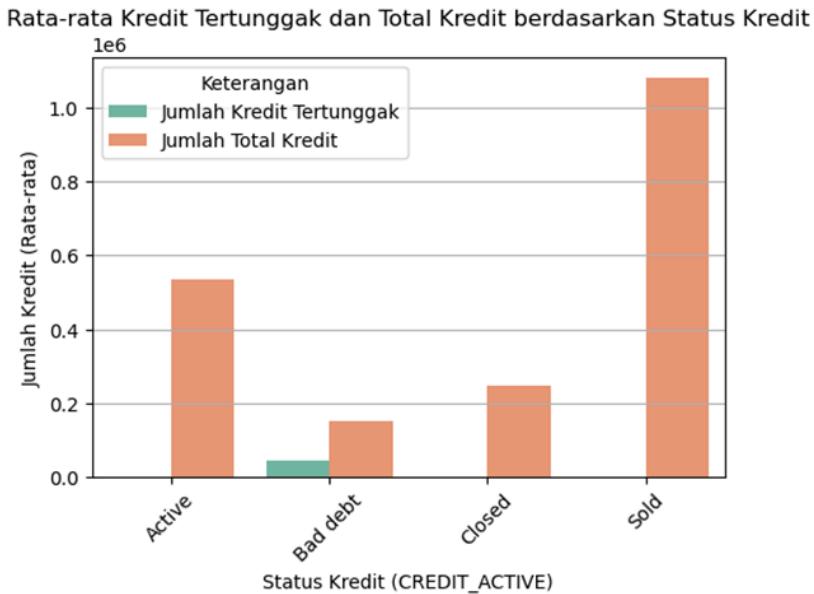
Jumlah kredit yang tertunggak atau belum dibayar tepat waktu.

Setelah dilakukan kalkulasi, didapatkan nilai akhir *default rate* memiliki digit desimal cukup banyak yaitu 6 angka dibelakang koma (,) hal ini akan berpengaruh terhadap tindakan permodelan sistem atau Machine Learning. Selain itu, kolom AMT_CREDIT_SUM memiliki missing values sebanyak 13 baris data, sehingga hasil kalkulasi menjadi NaN atau Not a Number.

Oleh karena itu perlu dipertimbangkan untuk melakukan pembulatan desimal atau melakukan proses Normalisasi/Standarisasi sehingga semua fitur berada dalam rentang yang sama tanpa mempengaruhi akurasi. Serta perlu dilakukan handling missing value untuk mendapatkan hasil kalkulasi yang lebih tepat, serta feature engineering untuk penetuan features yang paling relevant dengan kalkulasi *default rate*.

2. Analisa berdasarkan *features*.

Correlation between Overdue Credit and Total Credit by Credit Status



Grafik ini membandingkan jumlah kredit tertunggak dan jumlah total kredit berdasarkan status kredit :

- Active :

Jumlah Total Kredit: Sekitar 600.000 (0.6 juta).

Jumlah Kredit Tertunggak: Hampir tidak ada kredit tertunggak yang terlihat signifikan di kategori ini.

- Bad debt :

Jumlah Total Kredit: Sekitar 200.000.

Jumlah Kredit Tertunggak: Terlihat sekitar 50.000 kredit tertunggak, lebih besar daripada kategori lainnya, namun jumlah kredit total masih rendah.

- Closed :

Jumlah Total Kredit: Sekitar 300.000.

Jumlah Kredit Tertunggak: Hampir tidak ada kredit tertunggak.

- Sold:

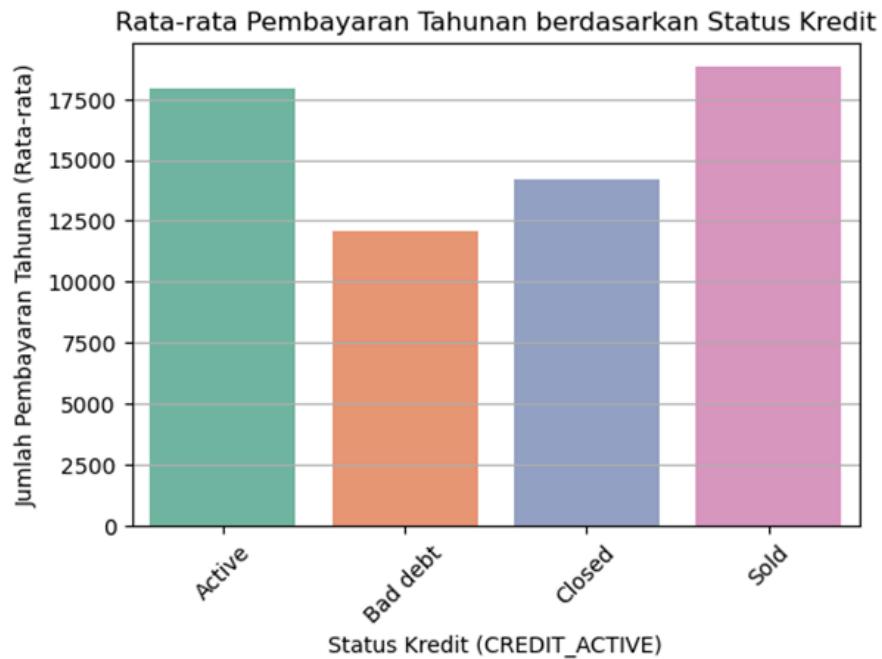
Jumlah Total Kredit: Sekitar 1 juta (kategori tertinggi).

Jumlah Kredit Tertunggak: Tidak ada kredit tertunggak.

Kesimpulan:

Status sold memiliki jumlah total kredit tertinggi, tetapi tidak ada kredit tertunggak. Status bad debt menunjukkan ada beberapa kredit tertunggak meskipun jumlah total kreditnya lebih rendah. Status active memiliki jumlah kredit total cukup besar, namun kredit tertunggaknya sangat kecil.

Yearly Payments by Credit Type



Grafik ini memberikan gambaran tentang jumlah pembayaran tahunan berdasarkan status kredit dari beberapa kategori :

- A. Active: Status kredit aktif memiliki rata-rata pembayaran tahunan tertinggi di grafik ini, sekitar 17.500.
- B. Bad debt: Kredit dengan status "utang buruk" atau gagal bayar memiliki rata-rata pembayaran tahunan yang lebih rendah, yaitu sekitar 11.000.
- C. Closed: Status kredit tertutup memiliki rata-rata pembayaran tahunan di bawah kategori "Active" dan di atas "Bad debt", sekitar 13.000.
- D. Sold: Status kredit yang sudah dijual kembali (mungkin ke pihak ketiga) memiliki rata-rata pembayaran yang hampir setara dengan "Active", yaitu sekitar 18.000.

Grafik ini menunjukkan bahwa kredit dengan status aktif dan sold (dijual) memiliki pembayaran tahunan yang lebih tinggi dibandingkan dengan status closed (ditutup) atau bad debt (utang buruk). Ini dapat menggambarkan bahwa akun-akun dengan status kredit yang lebih sehat (aktif atau dijual) cenderung memiliki kontribusi pembayaran lebih tinggi.

Kesimpulan :

- Klien yang kreditnya sudah dijual kembali (Sold) memberikan pendapatan tahunan yang tinggi, melebihi klien yang memiliki status kredit masih aktif. Jumlah kredit klien bukan faktor utama tertunggaknya pembayaran kredit. Dari dataset dapat diketahui ada beberapa kredit tertunggak meskipun jumlah total

kreditnya lebih rendah. Sedangkan klien dengan status active memiliki jumlah kredit total cukup besar, namun kredit tertunggaknya sangat kecil.

- Seluruh hasil analisis ini adalah sample atau gambaran awal, karena masih banyak terdapat nilai missing values yang menyebabkan data tidak menunjukkan hasil akurat.

Berdasarkan analisa dasar yang telah dilakukan terhadap Bureau Dataset, didapatkan beberapa business insight yang bisa menjadi pengetahuan untuk tindakan berikutnya. Namun untuk mencapai hasil yang akurat dan maksimal, dataset perlu dilakukan pembersihan terlebih dahulu atau cleansing data, karena besar kemungkinan insight bisnis akan berubah sebagaimana perubahan struktural data.

1. *Clustering* nasabah

Rekomendasi :

Utamakan nasabah/client dengan status kredit (CREDIT_ACTIVE) aktif atau sold. Ini dikarenakan nasabah dengan status aktif atau kredit telah dijual memiliki pola pembayaran kredit lebih baik dibandingkan dengan status kredit nasabah yang close dan bad debt. Jika seorang nasabah memiliki catatan pembayaran yang baik atau tidak pernah terlambat, bisa ditawarkan kredit dengan bunga rendah.

2. Pengelolaan nasabah bermasalah

Rekomendasi

Menganalisis nasabah yang sering mengalami keterlambatan atau menunggak dalam pembayaran (DAYS_CREDIT_ENDDATE,CREDIT_DAY_OVERDUE). Untuk nasabah dengan riwayat masalah ini, bisa dilakukan pengawasan lebih ketat atau ditawarkan program restrukturisasi utang.

3. Pembuatan skor kredit

Rekomendasi

Berdasarkan data seperti DAYS_CREDIT, AMT_CREDIT_SUM_DEBT, dan CNT_CREDIT_PROLONG, perusahaan bisa mengembangkan *credit scoring model* untuk otomatisasi keputusan kredit. Nasabah dengan skor kredit baik bisa langsung mendapatkan persetujuan kredit tanpa proses manual yang panjang.

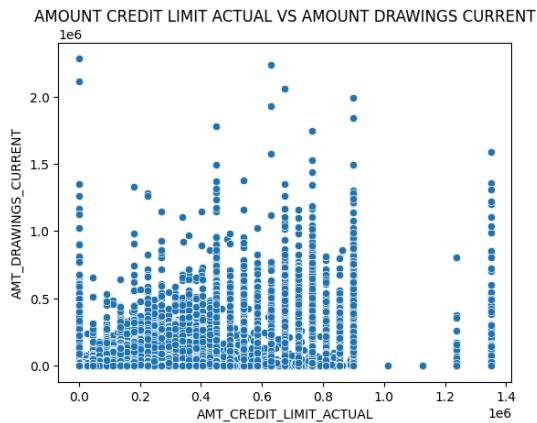
4. Pemantauan kredit secara real-time

Rekomendasi

Menggunakan data DAYS_CREDIT_UPDATE untuk memantau aktivitas kredit nasabah secara real-time. Jika terdapat sinyal nasabah yang mulai menunggak, perusahaan dapat mengambil tindakan cepat seperti pengingat pembayaran atau menawarkan solusi pembayaran alternatif.

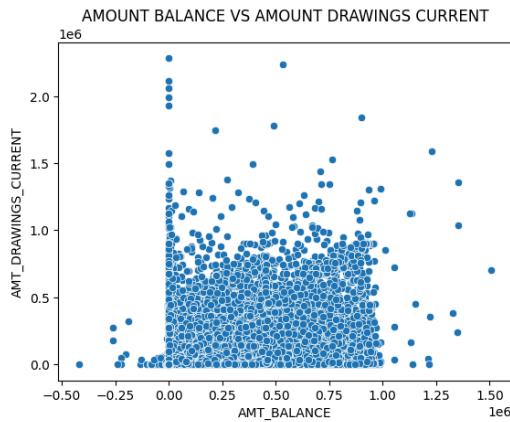
2.3.3. HC_credit_card_balance

1. BUSSINESS INSIGHT :



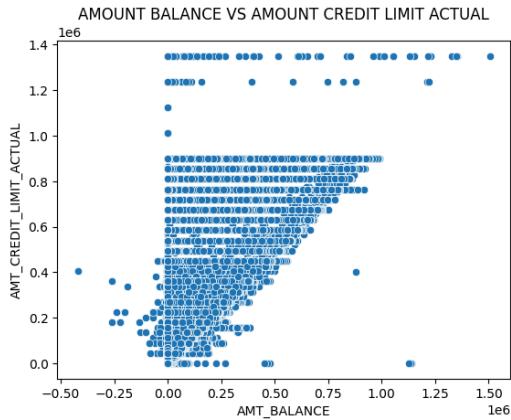
2. AMOUNT CREDIT LIMIT ACTUAL VS AMOUNT DRAWINGS CURRENT

Korelasi ini menunjukkan bahwa semakin banyak jumlah batas kredit yang diberikan maka jumlah penarikan yang diambil pengguna semakin banyak. Ini disebabkan bahwa pengguna semakin tinggi peluang untuk mendapatkan pinjaman yang lebih banyak



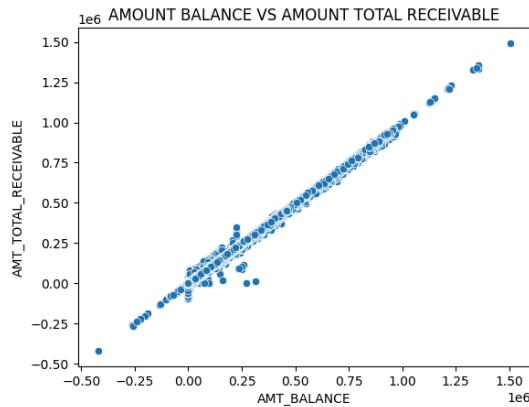
3. AMOUNT BALANCE VS AMOUNT DRAWINGS CURRENT

Korelasi ini menunjukkan bahwa semakin banyak jumlah saldo yang tersisa untuk maka jumlah penarikan yang diambil pengguna semakin banyak.



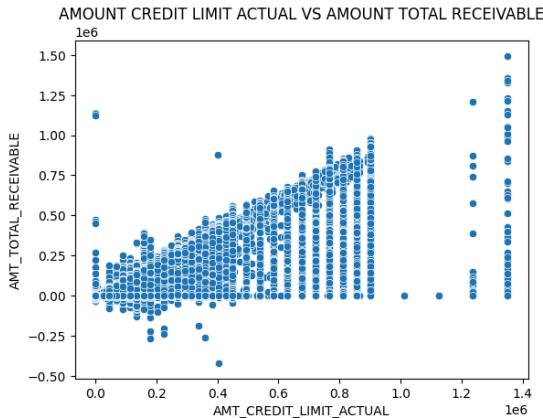
4. AMOUNT BALANCE VS AMOUNT CREDIT LIMIT ACTUAL

Korelasi ini menunjukkan bahwa jumlah saldo yang tersisa semakin tinggi maka pengguna mendapatkan jumlah batas kredit yang diberikan semakin banyak. Hal ini dikarenakan saldo yang banyak dapat dikatakan dia mampu untuk memenuhi kewajibannya. Hal ini akan mendorong pemberi pinjaman untuk memberikan batas kredit yang tinggi.



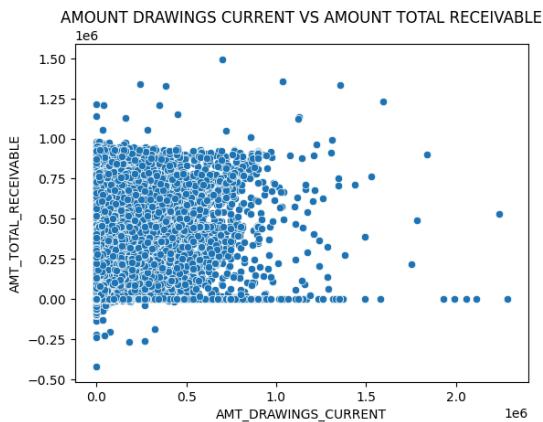
5. AMOUNT BALANCE VS AMOUNT TOTAL RECEIVABLE

Korelasi ini menunjukkan bahwa jumlah saldo yang tersisa pleh pengguna maka semakin banyak penerimaan yang diterima oleh lembaga peminjam uang. Hal ini dikarenakan jumlah saldo yng tinggi mengindikasikan dia mampu bayar atau melakukan kewajiban sehingga total penerimaan bank semakin banyak.



6. AMOUNT CREDIT LIMIT ACTUAL VS AMOUNT TOTAL RECEIVABLE

Korelasi ini menunjukkan bahwa jumlah batas kredit yang diberikan pengguna semakin tinggi maka semakin banyak penerimaan yang diterima oleh lembaga peminjam uang. Hal ini dikarenakan pengguna lebih leluasa dalam mendapatkan pinjaman, sehingga uang yang harus dikembalikan termasuk bunga dll semakin banyak.



7. AMOUNT DRAWINGS CURRENT VS AMOUNT TOTAL RECEIVABLE

Korelasi ini menunjukkan bahwa jumlah penarikan (dalam uang) yang dilakukan pengguna semakin tinggi maka semakin banyak penerimaan yang diterima oleh lembaga peminjam uang.

8. AMOUNT CREDIT LIMIT ACTUAL VS AMOUNT PAYMENT TOTAL CURRENT

Korelasi ini menunjukkan bahwa jumlah batas kredit yang diberikan pengguna semakin tinggi maka semakin banyak pembayaran yang harus dibayar oleh pengguna. Ini berkorelasi karena pengguna yang memiliki kredit limit yang tinggi akan semakin berpeluang untuk lebih banyak meminjam dan akhirnya total uang yang harus dibayar juga semakin banyak.

BUSSINESS RECOMENDATION :

1. Penawaran Batas Kredit yang Fleksibel:

Mengingat bahwa peningkatan batas kredit berhubungan dengan peningkatan jumlah penarikan dan pembayaran, perusahaan dapat mempertimbangkan untuk menawarkan batas kredit yang lebih fleksibel atau bertahap bagi pengguna yang menunjukkan kemampuan untuk membayar dengan baik.

2. Pengelolaan arus kas:

Amount Drawings Current yang tinggi bisa mengindikasikan bahwa pengguna menarik uang dari bisnis secara signifikan, yang dapat mengganggu arus kas jika tidak dikelola dengan baik.

3. Program Edukasi Keuangan:

Mengedukasi pengguna tentang pengelolaan kredit dan pinjaman bisa membantu mereka memahami risiko dan manfaat dari batas kredit yang lebih tinggi. Hal ini juga dapat meningkatkan loyalitas dan kepuasan pelanggan.

4. Strategi Pemasaran untuk Meningkatkan Penggunaan Kredit:

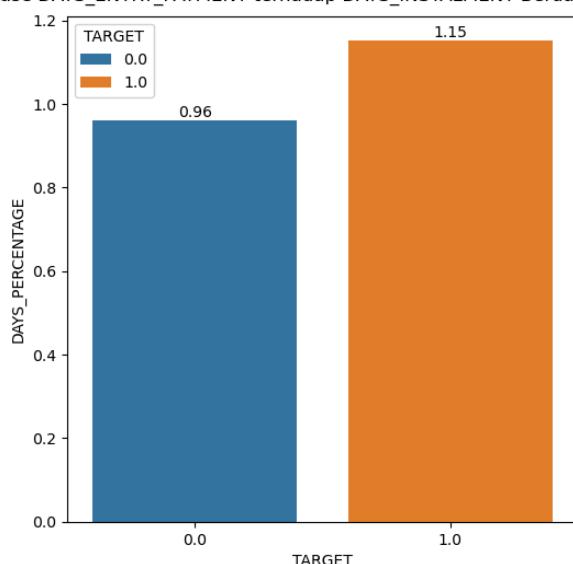
Mengembangkan kampanye pemasaran yang menonjolkan manfaat dari batas kredit yang lebih tinggi, serta cara-cara aman untuk memanfaatkan pinjaman, dapat menarik lebih banyak pengguna untuk melakukan penarikan.

5. Analisis Risiko yang Lebih Mendalam:

Menggunakan data untuk melakukan analisis risiko yang lebih baik dapat membantu dalam menentukan batas kredit yang sesuai bagi setiap pengguna, mengurangi risiko default, dan meningkatkan pendapatan dari pembayaran bunga.

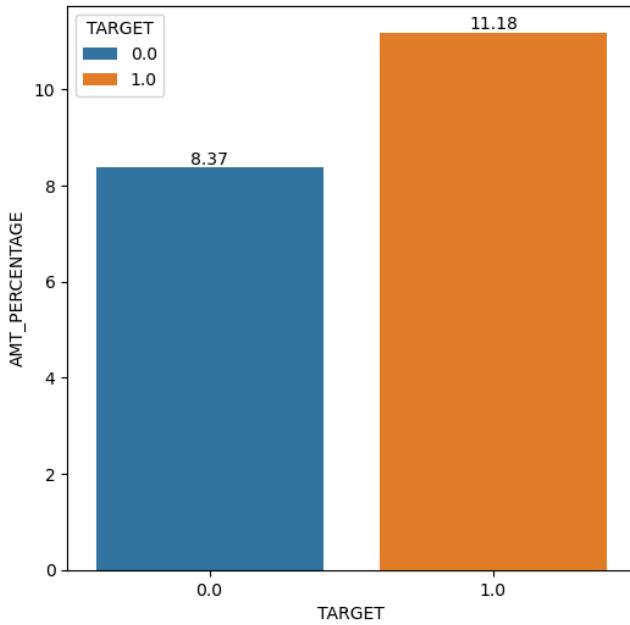
2.3.4. HC_installments_payments

Percentase DAYS_ENTRY_PAYMENT terhadap DAYS_INSTALMENT Berdasarkan TARGET

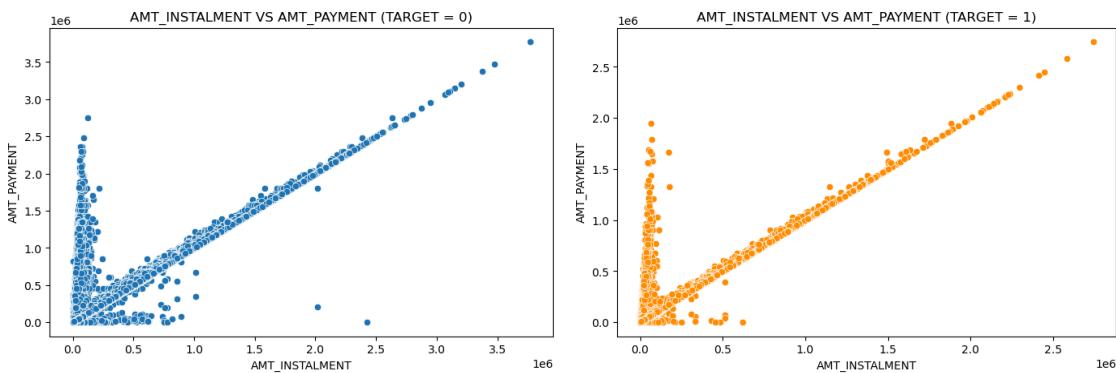


Fitur DAYS_INSTALMENT dan DAYS_ENTRY_PAYMENT pada target 0 mempunyai perbedaan persentase median yang lebih kecil dibandingkan target 1. Hal ini menunjukkan bahwa target 1 cenderung membayar sedikit lebih telat dibandingkan target 0.

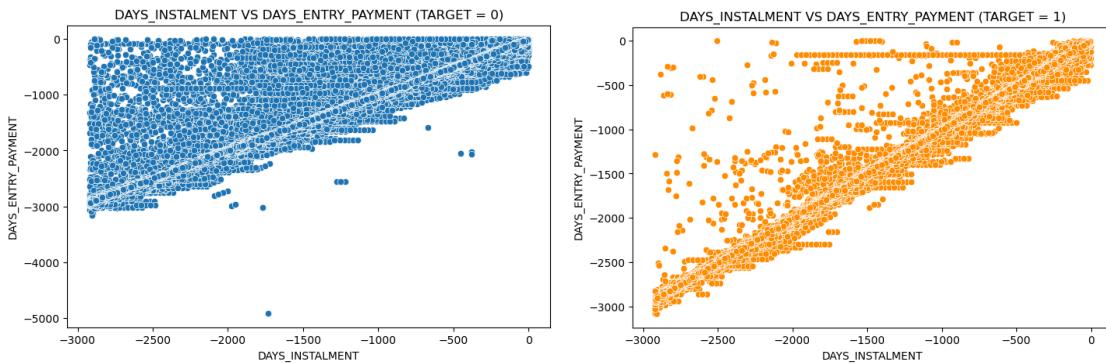
Persentase AMT_PAYMENT terhadap AMT_INSTALMENT Berdasarkan TARGET



Persentase perbedaan median dari AMT_INSTALMENT dan AMT_ENTRY_PAYMENT pada target 0 lebih kecil dibandingkan target 1. Hal ini menunjukkan target 1 lebih kesusahan untuk membayar nilai installment yang sudah ditentukan dibandingkan target 0.



Perbandingan dari AMT_INSTALMENT dengan AMT_PAYMENT menunjukkan kemiripan pola antara target 0 dengan 1, sehingga tidak ada perbedaan jumlah yang dibayarkan dengan tagihan yang diminta secara signifikan antara target 0 dengan 1..

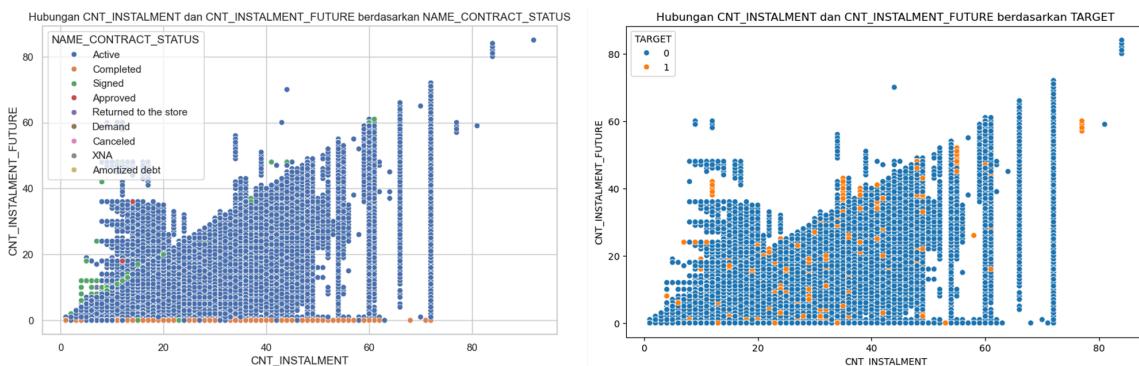


Perbandingan DAYS_INSTALMENT dengan DAYS_ENTRY_PAYMENT menunjukkan perbedaan pola. Target 1 menunjukkan pola korelasi yang lebih linear dan beraturan dibanding dengan target 0. Pola target 1 menunjukkan bahwa pengguna cenderung membayar tepat pada waktu yang ditentukan, sedangkan pola dari target 0 menunjukkan bahwa pengguna cenderung membayar sebelum batas waktu yang ditentukan.

REKOMENDASI BISNIS:

- Untuk klien yang rawan gagal bayar, berikan insentif untuk pembayaran tepat waktu, seperti diskon bunga jika mereka membayar sebelum jatuh tempo.
- Membuat program reward bagi nasabah yang melakukan pembayaran tepat waktu secara konsisten.
- Memberikan notifikasi jadwal yang terstruktur dan memberikan tools manajemen keuangan agar memudahkan klien mengatur keuangan dan bisa membayar tepat waktu.

2.3.5. HC_POS_CASH_balance



1. CNT_INSTALMENT dan CNT_INSTALMENT_FUTURE Berdasarkan NAME_CONTRACT_STATUS:

- Pada plot sebelah kiri, terlihat bahwa kontrak dengan status "Active" mendominasi distribusi data dengan jumlah installment (CNT_INSTALMENT) yang tinggi dan future installment (CNT_INSTALMENT_FUTURE) yang variatif.

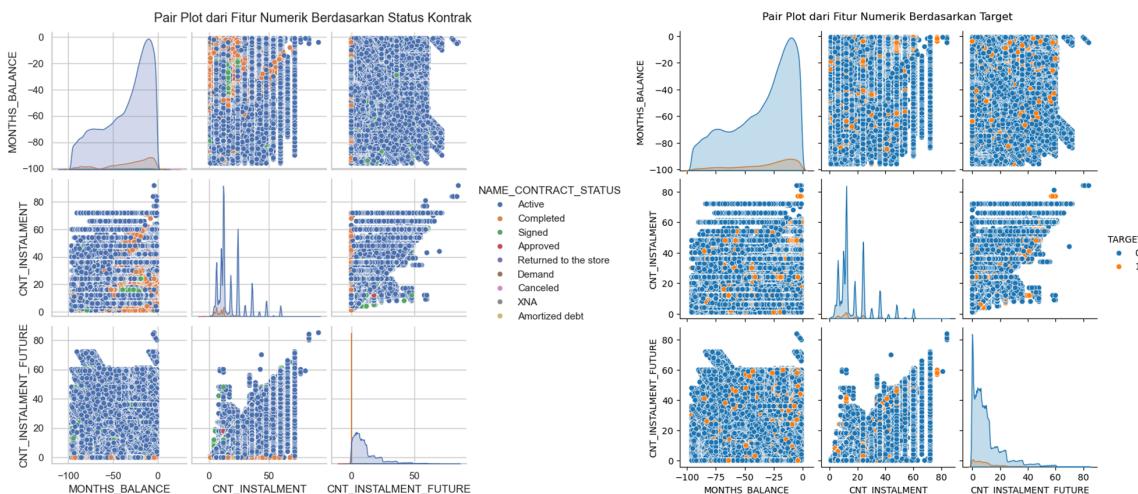
- Status seperti "Completed", "Signed", dan "Canceled" tampak berada pada rentang jumlah installment yang lebih rendah (sekitar 0-20).
- Hal ini menunjukkan bahwa kontrak "Active" umumnya memiliki sisa cicilan yang lebih besar, sementara kontrak "Completed" dan "Signed" cenderung memiliki jumlah installment yang lebih sedikit.

2. Pola Hubungan CNT_INSTALMENT dan CNT_INSTALMENT_FUTURE Berdasarkan TARGET:

- Pada plot sebelah kanan, terdapat korelasi antara jumlah installment (CNT_INSTALMENT) dan future installment (CNT_INSTALMENT_FUTURE) dengan TARGET.
- Klien dengan TARGET = 1 (berisiko gagal bayar, ditunjukkan dengan warna oranye) cenderung memiliki future installment (CNT_INSTALMENT_FUTURE) yang lebih tinggi, terutama di sekitar 20-40 installment.
- Ini menunjukkan bahwa klien yang berisiko lebih tinggi sering kali memiliki jadwal pembayaran yang lebih panjang atau sisa cicilan yang lebih besar. Oleh karena itu, jumlah installment di masa depan tampaknya berkorelasi dengan risiko gagal bayar yang lebih tinggi.

3. Variasi CNT_INSTALMENT dan CNT_INSTALMENT_FUTURE:

- Terdapat hubungan positif yang kuat antara CNT_INSTALMENT dan CNT_INSTALMENT_FUTURE di kedua plot, yang mengindikasikan bahwa semakin banyak jumlah cicilan yang telah dilakukan (CNT_INSTALMENT), semakin besar kemungkinan terdapat lebih banyak cicilan di masa depan (CNT_INSTALMENT_FUTURE).
- Kontrak dengan TARGET = 1 sering kali menunjukkan jumlah cicilan yang lebih besar di masa depan, yang berarti mereka memiliki kewajiban yang lebih tinggi, meningkatkan risiko gagal bayar.



1. Distribusi MONTHS_BALANCE:

- Pada kedua plot, distribusi MONTHS_BALANCE menunjukkan bahwa kontrak yang aktif (Active) dan TARGET = 0 (tidak berisiko gagal bayar) cenderung memiliki nilai MONTHS_BALANCE yang lebih baru, yang berarti aktivitas pembayaran cicilan terjadi lebih mendekati saat ini.
- Pada kontrak dengan TARGET = 1 (berisiko gagal bayar), distribusi MONTHS_BALANCE cenderung memiliki lebih sedikit kontrak di masa mendekati sekarang, menunjukkan bahwa mungkin klien yang berisiko mengalami kesulitan di periode-periode akhir pembayaran.

2. Pola CNT_INSTALMENT dan CNT_INSTALMENT_FUTURE Berdasarkan NAME_CONTRACT_STATUS dan TARGET:

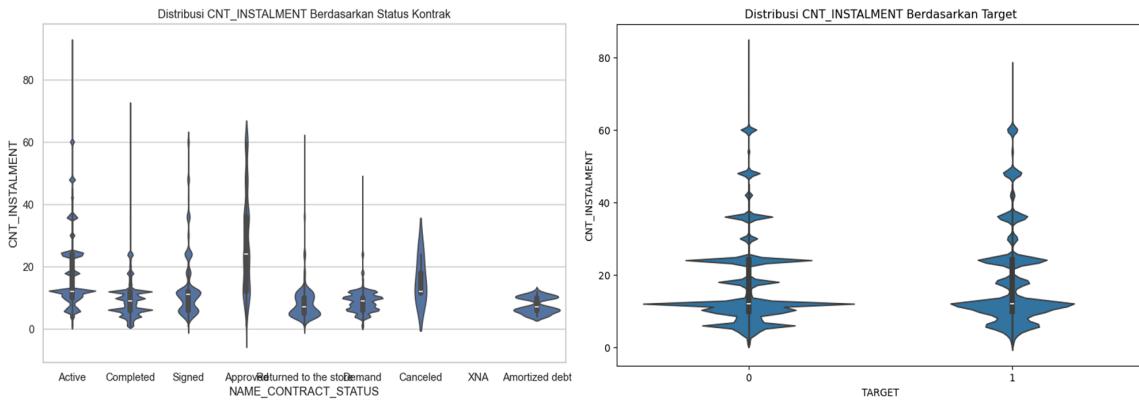
- Pada kedua plot, terlihat adanya hubungan positif yang kuat antara CNT_INSTALMENT dan CNT_INSTALMENT_FUTURE. Pada kontrak dengan status "Active" dan TARGET = 0, terdapat jumlah installment (CNT_INSTALMENT) yang tinggi bersamaan dengan jumlah future installment (CNT_INSTALMENT_FUTURE) yang cukup besar.
- Klien dengan TARGET = 1 cenderung memiliki jumlah CNT_INSTALMENT_FUTURE yang lebih tinggi, yang berarti bahwa mereka masih memiliki banyak cicilan yang harus dibayar, dan hal ini berkaitan dengan peningkatan risiko gagal bayar. Ini menunjukkan hubungan antara jumlah cicilan di masa depan dan kemungkinan risiko default.

3. Variasi Berdasarkan NAME_CONTRACT_STATUS:

- Pada pair plot yang mengelompokkan data berdasarkan NAME_CONTRACT_STATUS, kontrak "Active" memiliki variasi yang paling luas dalam hal CNT_INSTALMENT dan CNT_INSTALMENT_FUTURE, sementara kontrak dengan status seperti "Completed", "Signed", dan "Canceled" cenderung memiliki jumlah installment yang lebih sedikit.
- Hal ini menunjukkan bahwa kontrak yang masih aktif memiliki lebih banyak variabilitas dalam pembayaran cicilan yang sedang berlangsung dibandingkan dengan kontrak yang sudah selesai atau dibatalkan.

4. Tren dalam MONTHS_BALANCE:

- Klien dengan TARGET = 1 cenderung menunjukkan pola peningkatan installment ketika mendekati waktu terkini (MONTHS_BALANCE mendekati 0), yang menunjukkan adanya risiko yang lebih tinggi di periode-periode terakhir cicilan.
- Pada kontrak "Completed", nilai MONTHS_BALANCE menunjukkan angka yang lebih negatif, yang berarti kontrak tersebut sudah selesai sejak lama.

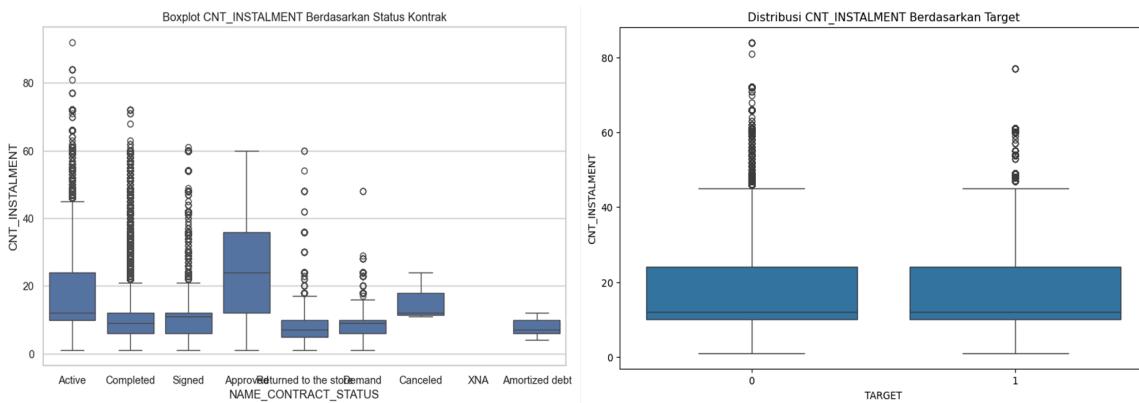


1. Distribusi Berdasarkan Status Kontrak (Grafik Kiri):

- Status kontrak "Active" dan "Approved" memiliki rentang CNT_INSTALMENT yang lebih besar dibandingkan dengan status lainnya, mencapai lebih dari 60 cicilan.
- Status kontrak "Completed" menunjukkan distribusi yang lebih terkonsentrasi di kisaran cicilan yang lebih rendah, dengan jumlah cicilan mayoritas di bawah 20.
- Kontrak dengan status "Amortized debt" dan "XNA" memiliki rentang cicilan yang relatif rendah dan distribusi yang sempit.
- Status "Canceled" dan "Demand" juga memiliki distribusi cicilan yang lebih rendah dibandingkan dengan status lainnya.

2. Distribusi Berdasarkan Target (Grafik Kanan):

- Baik pada target 0 (tidak gagal bayar) maupun target 1 (gagal bayar), rentang distribusi CNT_INSTALMENT cukup lebar, tetapi target 1 tampaknya memiliki lebih banyak outlier di jumlah cicilan yang tinggi.
- Distribusi CNT_INSTALMENT pada target 0 sedikit lebih terkonsentrasi di kisaran cicilan rendah hingga menengah (di bawah 30 cicilan), sedangkan pada target 1 terdapat kecenderungan cicilan yang lebih tinggi namun juga lebih merata.

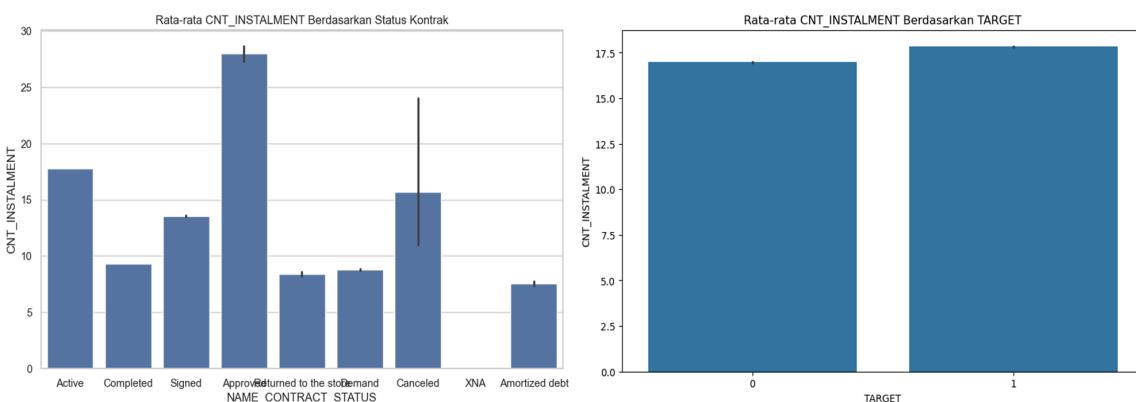


1. Distribusi CNT_INSTALMENT Berdasarkan Status Kontrak (Grafik Kiri):

- Status "Active": Memiliki median yang lebih tinggi dibandingkan status lainnya, dengan banyak outlier hingga mencapai lebih dari 80 cicilan. Ini menunjukkan bahwa kontrak aktif cenderung memiliki jumlah cicilan yang lebih tinggi dan variasi yang lebih besar.
- Status "Approved": Menunjukkan distribusi yang lebih menyebar dengan outlier hingga lebih dari 60 cicilan, tetapi rentang interquartile-nya cukup lebar, menunjukkan variabilitas cicilan yang tinggi.
- Status "Completed", "Signed", "Returned to the store", dan "Demand": Cenderung memiliki median yang lebih rendah dan distribusi yang lebih sempit, dengan cicilan mayoritas di bawah 20. Ini menunjukkan bahwa kontrak yang sudah selesai atau ditandatangani biasanya melibatkan cicilan yang lebih sedikit.
- Status "Canceled" dan "Amortized debt": Memiliki distribusi cicilan yang cukup rendah dengan sedikit atau tanpa outlier, menunjukkan cicilan yang stabil dan lebih rendah.

2. Distribusi CNT_INSTALMENT Berdasarkan Target (Grafik Kanan):

- Target 0 (Tidak Gagal Bayar): Rentang cicilan lebih lebar dengan outlier yang lebih banyak hingga lebih dari 80 cicilan. Ini menunjukkan bahwa nasabah yang tidak gagal bayar cenderung memiliki cicilan yang lebih bervariasi, termasuk beberapa kasus dengan cicilan sangat tinggi.
- Target 1 (Gagal Bayar): Distribusinya relatif serupa, tetapi sedikit lebih sempit dengan lebih sedikit outlier dibandingkan target 0. Ini mungkin menunjukkan bahwa meskipun ada nasabah yang gagal bayar dengan jumlah cicilan tinggi, kebanyakan dari mereka berada di kisaran cicilan yang lebih rendah hingga menengah.



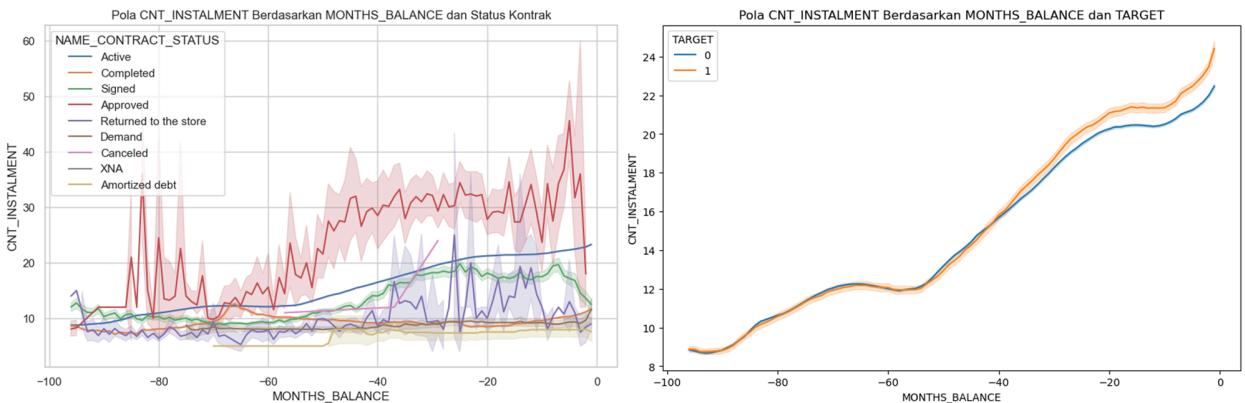
1. Rata-rata CNT_INSTALMENT Berdasarkan Status Kontrak (Grafik Kiri):

- Status "Approved": Memiliki rata-rata jumlah cicilan yang paling tinggi dibandingkan status kontrak lainnya, dengan rata-rata lebih dari 25 cicilan.
- Status "Active": Memiliki rata-rata yang cukup tinggi, sekitar 15 cicilan. Ini menunjukkan bahwa kontrak aktif cenderung memiliki jumlah cicilan yang signifikan.

- Status "Canceled": Rata-rata cicilan cukup tinggi, meskipun tidak sebanyak "Approved", dan juga memiliki rentang variabilitas yang besar (ditunjukkan dengan bar error yang lebar).
- Status "Completed", "Signed", "Returned to the store", dan "Demand": Memiliki rata-rata cicilan yang lebih rendah, sekitar 10 cicilan atau kurang, menunjukkan bahwa kontrak yang sudah selesai atau ditandatangani melibatkan lebih sedikit cicilan.
- Status "Amortized debt": Memiliki rata-rata cicilan terendah, dengan sekitar 5 cicilan, mengindikasikan bahwa kontrak dengan status ini biasanya melibatkan cicilan yang lebih sedikit.

2. Rata-rata CNT_INSTALMENT Berdasarkan Target (Grafik Kanan):

Target 0 (Tidak Gagal Bayar) dan Target 1 (Gagal Bayar): Kedua kelompok memiliki rata-rata yang hampir sama, dengan rata-rata cicilan sekitar 17 cicilan. Ini menunjukkan bahwa jumlah cicilan rata-rata tidak terlalu memengaruhi apakah seorang nasabah gagal membayar atau tidak, sehingga variabel ini mungkin bukan indikator utama untuk memprediksi default.



1. Plot Kiri - Pola CNT_INSTALMENT Berdasarkan Status Kontrak:

- Status kontrak yang berbeda mempengaruhi jumlah cicilan (CNT_INSTALMENT) seiring berjalannya waktu (MONTHS_BALANCE).
- Status "Completed" (merah) menunjukkan lonjakan yang sangat signifikan dalam jumlah cicilan sebelum akhir periode cicilan (MONTHS_BALANCE mendekati 0).
- Status "Active" (biru tua) memiliki pola yang lebih stabil dan naik secara perlahan seiring waktu.
- "Approved", "Signed", dan "Returned to the store" memiliki pola yang lebih rendah dan stabil dibandingkan "Completed".
- Beberapa status kontrak, seperti "XNA" dan "Amortized debt", memiliki jumlah cicilan yang rendah dan tidak menunjukkan perubahan signifikan sepanjang waktu.

2. Plot Kanan - Pola CNT_INSTALMENT Berdasarkan TARGET:

- Perbedaan antara TARGET 0 dan TARGET 1 relatif kecil, tetapi TARGET 1 (warna oranye) menunjukkan sedikit jumlah cicilan yang lebih tinggi dibandingkan TARGET 0 (biru), terutama mendekati akhir periode cicilan (MONTHS_BALANCE mendekati 0).
- Pola ini mengindikasikan bahwa klien dengan TARGET 1 (kemungkinan lebih berisiko atau gagal bayar) memiliki cicilan yang sedikit lebih tinggi, terutama di periode akhir kontrak.

BUSINESS RECOMMENDATIONS:

1. Manajemen Risiko dan Intervensi Dini:

- Pantau Kontrak Berisiko Tinggi: Fokus pada kontrak dengan cicilan masa depan yang tinggi (CNT_INSTALMENT_FUTURE), terutama yang ditandai TARGET = 1. Menawarkan rencana pembayaran yang disesuaikan atau konseling keuangan dapat mencegah terjadinya gagal bayar.
- Selidiki Pembatalan Kontrak: Pahami alasan di balik pembatalan kontrak dan tawarkan program intervensi dini, seperti restrukturisasi pinjaman, untuk pelanggan yang menunjukkan tanda-tanda kesulitan keuangan sebelum pembatalan.

2. Segmentasi Pelanggan:

- Segmentasi Berdasarkan Status Kontrak: Gunakan status kontrak untuk membedakan strategi pelanggan. Tawarkan program loyalitas untuk kontrak Completed atau insentif pembayaran untuk kontrak Active.
- Targetkan Retensi untuk Pelanggan Jangka Panjang: Berikan penghargaan loyalitas bagi pelanggan jangka panjang dengan nilai CNT_INSTALMENT_FUTURE yang tinggi, terutama yang ditandai TARGET = 1.

3. Ketentuan Kontrak yang Fleksibel dan Produk Baru:

- Sesuaikan Rencana Pembayaran: Tawarkan rencana pembayaran yang fleksibel bagi pelanggan dengan jumlah cicilan masa depan yang tinggi untuk mencegah gagal bayar, terutama untuk kontrak Approved.
- Perluas Pembiayaan Jangka Pendek: Sediakan opsi pembiayaan jangka pendek bagi pelanggan dengan total cicilan yang lebih rendah yang mungkin ingin mengonsolidasikan utang mereka.

4. Mencegah Gagal Bayar dan Re-engagement Pelanggan:

- Sistem Peringatan Dini: Implementasikan sistem peringatan dini bagi pelanggan dengan saldo tertunggak yang meningkat. Tawarkan intervensi seperti restrukturisasi pinjaman atau cuti pembayaran sementara.
- Kelola Pembatalan Kontrak: Proaktif tawarkan bantuan keuangan atau negosiasi ulang kepada pelanggan yang mendekati pembatalan.

- Kampanye Retargeting: Libatkan kembali pelanggan yang telah menyelesaikan kontrak Completed atau Signed melalui kampanye yang menawarkan produk baru berdasarkan riwayat pembayaran mereka yang positif.

5. Peluang Loyalitas dan Upsell:

- Insentif untuk Kontrak yang Selesai: Berikan penghargaan kepada pelanggan yang menyelesaikan kontrak mereka dengan sukses dengan menawarkan produk keuangan baru dengan syarat yang menarik.
- Peluang Cross-Sell untuk Kontrak Aktif: Identifikasi pelanggan dengan jumlah cicilan yang tinggi untuk peluang upsell atau penawaran yang dipersonalisasi.

6. Produk Keuangan yang Disesuaikan:

- Sesuaikan Ketentuan untuk Kontrak yang Disetujui: Karena kontrak Approved cenderung memiliki jumlah cicilan yang lebih tinggi, pertimbangkan untuk menawarkan opsi pembayaran yang fleksibel atau restrukturisasi untuk meringankan proses pembayaran.
- Produk Jangka Pendek Baru: Perkenalkan produk keuangan khusus bagi pelanggan yang lebih memilih rencana cicilan jangka pendek, terutama dalam status Demand atau Returned.

7. Penargetan Berbasis Data:

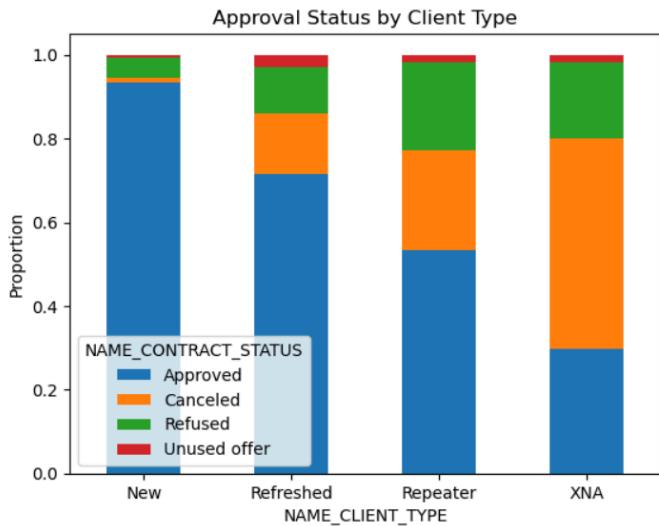
Manfaatkan hubungan antara total dan cicilan masa depan untuk membuat penawaran yang dipersonalisasi atau segmentasi pelanggan untuk program loyalitas, terutama untuk pelanggan jangka panjang.

2.3.6. HC_previous_application

1. Tingkat Persetujuan Berdasarkan Tipe Nasabah

Insight: Berdasarkan visualisasi data status kontrak berdasarkan tipe nasabah, nasabah baru memiliki tingkat persetujuan yang lebih tinggi dibandingkan nasabah lama.

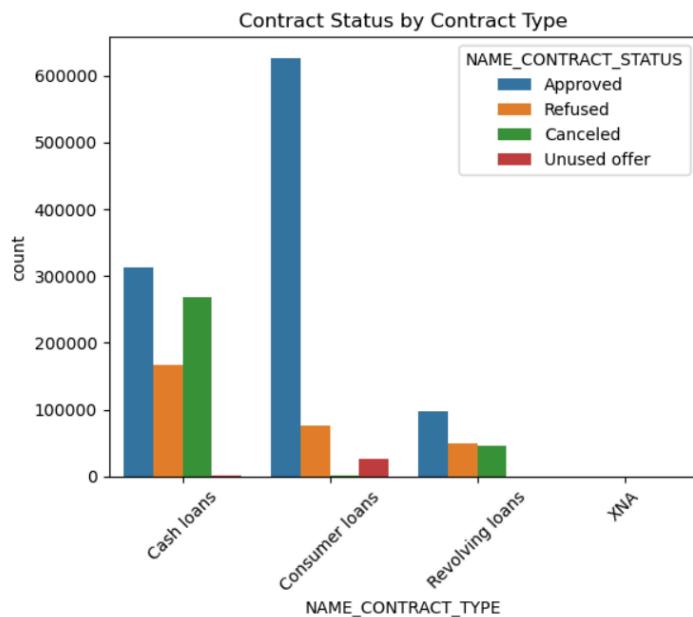
Rekomendasi: Bisnis dapat memberikan program loyalitas untuk nasabah berulang, seperti penawaran suku bunga lebih rendah, untuk mempertahankan mereka sebagai pelanggan jangka panjang.



2. Status Aplikasi Kredit Berdasarkan Tipe Kontrak

Insight: Tipe kontrak **Consumer loans** lebih sering disetujui dibandingkan tipe kontrak lain seperti **Cash loans**. Hal ini menunjukkan bahwa peminjaman berupa barang akan mudah disetujui karena terdapat barang berupa fisik.

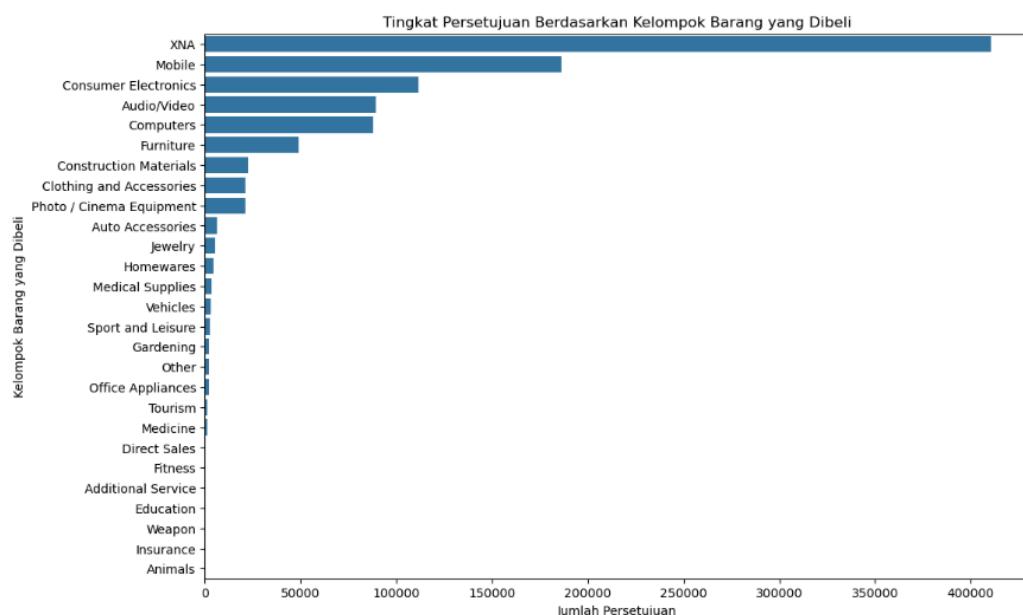
Rekomendasi: Perusahaan dapat mempertimbangkan untuk memprioritaskan kerjasama dengan penyedia barang-barang yang seringkali dibutuhkan konsumen, yang bisa dilihat juga pada data kategori produk yang paling populer. Evaluasi kembali risiko dan penilaian kredit pada tipe kredit barang (Cash loans) untuk melihat apakah ada cara untuk meningkatkan persetujuan.



3. Kategori Produk Paling Populer

Insight: Berdasarkan visualisasi data pada kategori produk kelompok barang yang dibeli, barang elektronik dan furniture adalah kategori yang paling sering diajukan dan disetujui. Hal tersebut menunjukkan bahwa nasabah lebih cenderung membeli barang-barang rumah tangga dan teknologi melalui kredit, serta seringkali disetujui kreditnya.

Rekomendasi: Bisnis dapat memprioritaskan kerjasama dengan penyedia barang-barang elektronik dan furniture untuk memberikan penawaran spesial atau diskon, karena ini adalah produk yang paling diminati oleh nasabah.



2.4. Pre-processing Recommendations

2.4.1. HC_application_train | test

1. Handling missing values

Ada 3 kolom yang memiliki data yang kosong: AMT_ANNUITY, AMT_GOODS_PRICE, NAME_TYPE_SUITE.

Untuk fitur AMT_ANNUITY dan AMT_GOODS nilai yang hilang dapat diisi dengan rata-rata (mean) atau median dan untuk NAME_TYPE_SUITE nilai yang hilang dapat diisi dengan modus.

2. Handling Outliers

Setelah mengidentifikasi outliers, untuk menangani outliers dapat menggunakan metode IQR(Inter Quartile Range).

3. Feature Engineering

- Berdasarkan multivariate analysis heatmap, ada 3 fitur yang berkorelasi kuat. Maka dari itu fitur yang dipilih adalah fitur AMT_CREDIT, karena dapat mengetahui besaran nilai kredit yang dikeluarkan.
- Untuk fitur AMT_INCOME_TOTAL dan AMT_CREDIT perlu dilakukan transformasi fitur dikarenakan besaran nilainya memiliki rentang yang besar.

2.4.2. HC_bureau

Recommendations for Pre-Processing Action

Berdasarkan hasil analisa masing-masing feature pada Bureau Dataset, terdapat beberapa insight untuk proses cleansing data atau pre-processing di tahapan berikutnya.

1. Handling missing values

Missing values atau nilai yang kosong pada Bureau Dataset sangat besar, berkisar 0,0007% - 71,47% data. Dengan hampir tiga perempat dari dataset hilang, hasil analisis dapat menjadi sangat bias atau tidak representatif. Ini dapat mengarah pada kesimpulan yang salah atau interpretasi yang keliru. Besarnya nilai missing values ini juga bisa menjadi potensi untuk dilakukan penghapusan kolom, agar menghindari kesalahan representasi ketika nilai kosong dimasukkan sebuah nilai.

2. Handling Outliers

Banyak dari fitur ini menunjukkan distribusi yang sangat terpusat dengan beberapa outliers. Data terpusat pada nilai negatif (-) dan nol (0), sedangkan sebaran data tidak merata karena nilai maksimum terlalu jauh daripada nilai terpusat yang lainnya. Hal ini perlu dilakukan analisa secara tepat, karena nilai outliers tersebut harus benar diketahui apakah memang merepresentasikan data atau kondisi yang sesungguhnya, atau justru sebaran data yang ada tidak merepresentasikan data yang benar karena didominasi oleh missing values.

3. Feature Engineering

Feature engineering terutama features selection menjadi tahapan yang krusial pada Bureau Dataset, karena pemahaman terhadap feature dataset akan memberikan dampak yang besar terutama untuk permodelan sistem. Bureau dataset memiliki banyak nilai kosong, distribusi sebaran data yang terpusat dan tidak merata, serta membutuhkan pemahaman data outliers harus mendalam. Sehingga, perlu membuat keputusan yang tepat untuk memilih feature yang relevant sesuai tujuan bisnis. Features yang dianggap tidak relevant dan beresiko terlalu bias (seperti banyak missing values) menyebabkan model yang dibangun dengan data ini tidak akurat atau tidak dapat diandalkan.

2.4.3. HC_credit_card_balance

1. Perlu diteliti lebih lanjut apakah nilai 0 di Q1, Q2, Q3 di beberapa kolom dipengaruhi karena ada data yang null, NaN atau emang bernilai 0.
2. Perlu juga diteliti lebih lanjut mengenai nilai max yang ada apakah masih dikatakan wajar atau tidak, mengingat bahwa tidak semua orang memiliki karakteristik atau perilaku yang sama.
3. Terdapat beberapa kolom yang memiliki outliers, maka perlu ditindaklanjutin dengan mengganti nilai outliers dengan median
4. Menghapus feature yang kurang relevan (tidak berkorelasi antar feature atau memiliki pengertian yang sama dengan feature lainnya. Berikut kolom-kolom tersebut CNT_INSTALMENT_MATURE_CUM, AMT_RECEIVABLE_PRINCIPAL, AMT_RECVABLE, AMT_PAYMENT_CURRENT

2.4.4. HC_installments_payments

1. Melakukan Transformasi Fitur (menggunakan standarisasi/log transformation):
 - NUM_INSTALMENT_VERSION
 - NUM_INSTALMENT_NUMBER
 - DAYS_INSTALMENT
 - DAYS_ENTRY_PAYMENT
 - AMT_INSTALMENT
 - AMT_PAYMENT
2. Melakukan Handling Outlier (menggunakan metode IQR) :
 - NUM_INSTALMENT_VERSION
 - NUM_INSTALMENT_NUMBER
 - DAYS_ENTRY_PAYMENT
 - AMT_INSTALMENT
 - AMT_PAYMENT
3. Pengisian Missing Value (menggunakan median):
 - DAYS_ENTRY_PAYMENT
 - AMT_PAYMENT
 - TARGET
4. Melakukan Class Imbalance Handling (menggunakan undersampling/SMOTE):
 - TARGET

2.4.5. HC_POS_CASH_balance

1. Transformasi dan Skala Fitur:
 - Standarisasi fitur numerik seperti CNT_INSTALMENT, CNT_INSTALMENT_FUTURE, dan MONTHS_BALANCE menggunakan Min-Max atau Standard Scaling.

- Normalisasi fitur yang skewed menggunakan log transformations.
2. Penanganan Outlier:
Tangani outlier pada CNT_INSTALMENT dan CNT_INSTALMENT_FUTURE menggunakan log transformation atau menghapus kasus ekstrim, terutama untuk kontrak dengan TARGET = 1.
 3. Encoding Kategorikal:
Terapkan One-Hot Encoding atau Target Encoding pada NAME_CONTRACT_STATUS, tergantung pada model yang digunakan.
 4. Feature Engineering:
 - Buat fitur baru seperti remaining_installments_percentage: $= \frac{\text{remaining_instalments_pct}}{\text{CNT_INSTALMENT_FUTURE/CNT_INSTALMENT}}$
 - Perkenalkan fitur berbasis waktu seperti average installment count over different periods, terutama pada bulan-bulan terakhir.
 - Konversi MONTHS_BALANCE menjadi fitur time-lagged atau rolling windows untuk menangkap tren temporal dalam perilaku pelanggan.
 5. Penanganan Ketidakseimbangan Kelas:
Atasi ketidakseimbangan pada TARGET = 1 menggunakan metode SMOTE atau undersampling.
 6. Dimensionality Reduction dan Penanganan Missing Data:
 - Hapus identifier (SK_ID_PREV, SK_ID_CURR) yang tidak berkontribusi pada pemodelan prediktif.
 - Tangani kategori langka pada NAME_CONTRACT_STATUS dengan cara menggabungkannya dengan kategori lain atau menghapusnya.

2.4.6. HC_previous_application

1. Mengatasi outlier pada feature AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_DOWN_PAYMENT, AMT_GOODS_PRICE, RATE_DOWN_PAYMENT dengan di-transformasi, capping, atau diganti dengan nilai median.
2. Drop feature SELLERPLACE_AREA, DAYS_FIRST_DRAWING, DAYS_FIRST_DUE, DAYS_LAST_DUE_1ST_VERSION, DAYS_LAST_DUE, DAYS_TERMINATION karena memiliki banyak nilai kosong dan tidak relevan untuk analisis terhadap model.

III. LAMPIRAN (LINK)

Repository : <https://github.com/Bramasta66/Home-Credit-Default-Risk>