Stage 2
DataRider

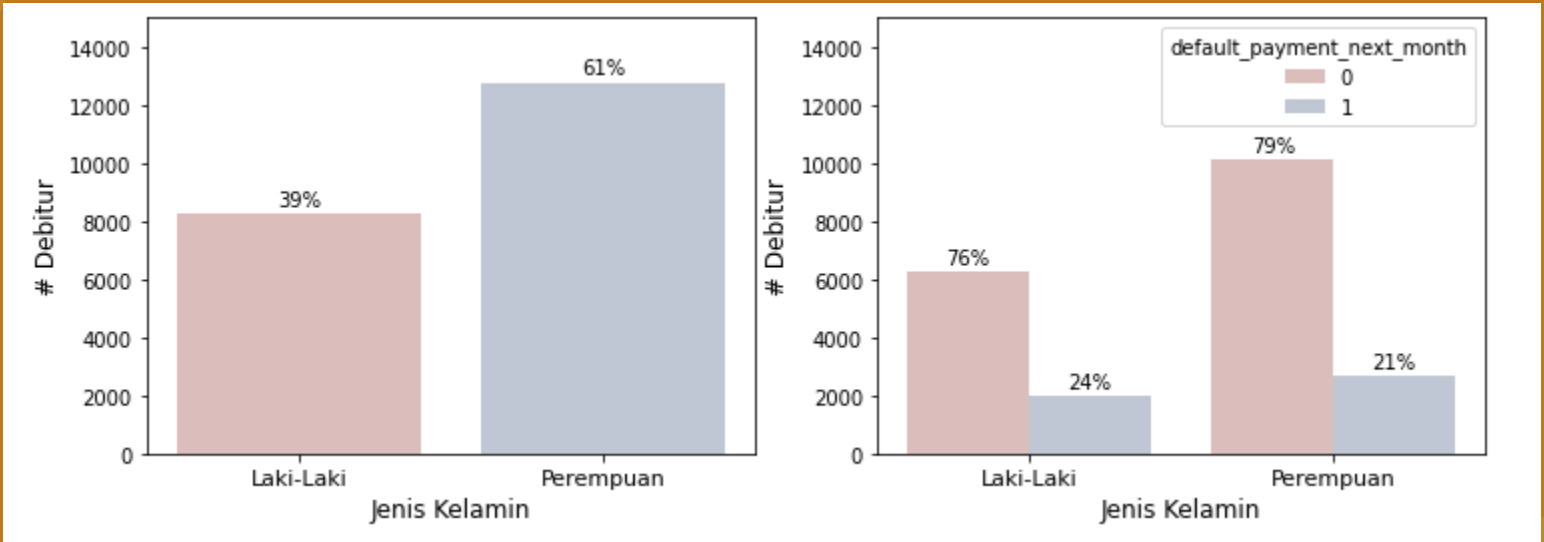
Cristanto
Steven Benny
Tri Setiawan
Ulva Dewiyanti

Eksplorasi Berbagai Attributes

1. Categoricals
2. Numericals

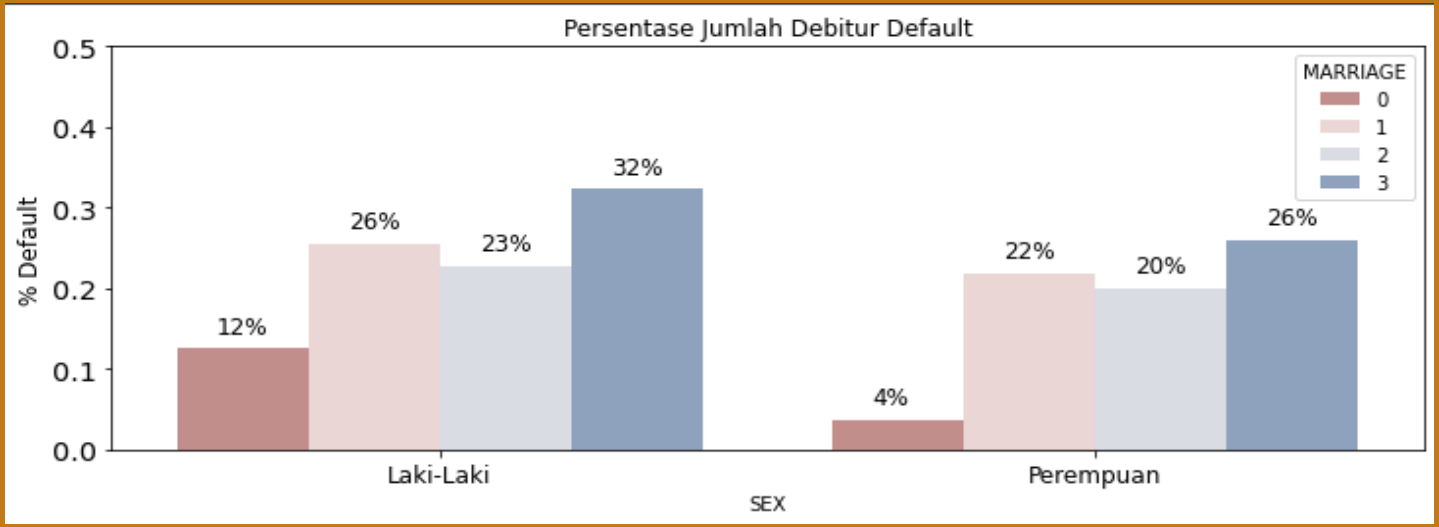
Categoricals

SEX



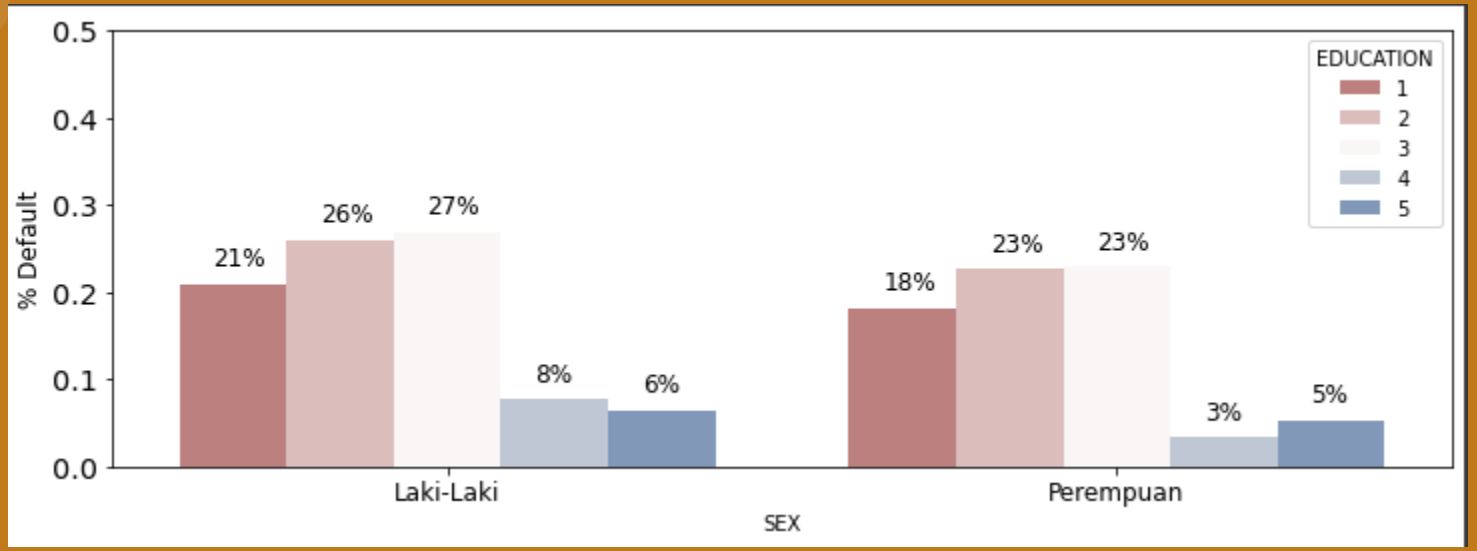
Melalui visualisasi kolom SEX di atas dapat disimpulkan bahwa sekitar 61% debitur adalah perempuan, namun tampak laki-laki memiliki peluang gagal bayar yang sedikit lebih tinggi yaitu 24% dibandingkan perempuan yang hanya 21%

SEX & MARRIAGE



1. Berdasarkan grafik diatas, di setiap kategori MARRIAGE ternyata dapat disimpulkan bahwa laki-laki dengan status Bercerai lebih memungkinkan gagal bayar, diikuti oleh laki-laki yang Menikah kemudian Lajang
2. Jika melihat dari Perempuan yang memilki kesamaan grafik dengan laki-laki, hanya berbeda di nilai besarannya

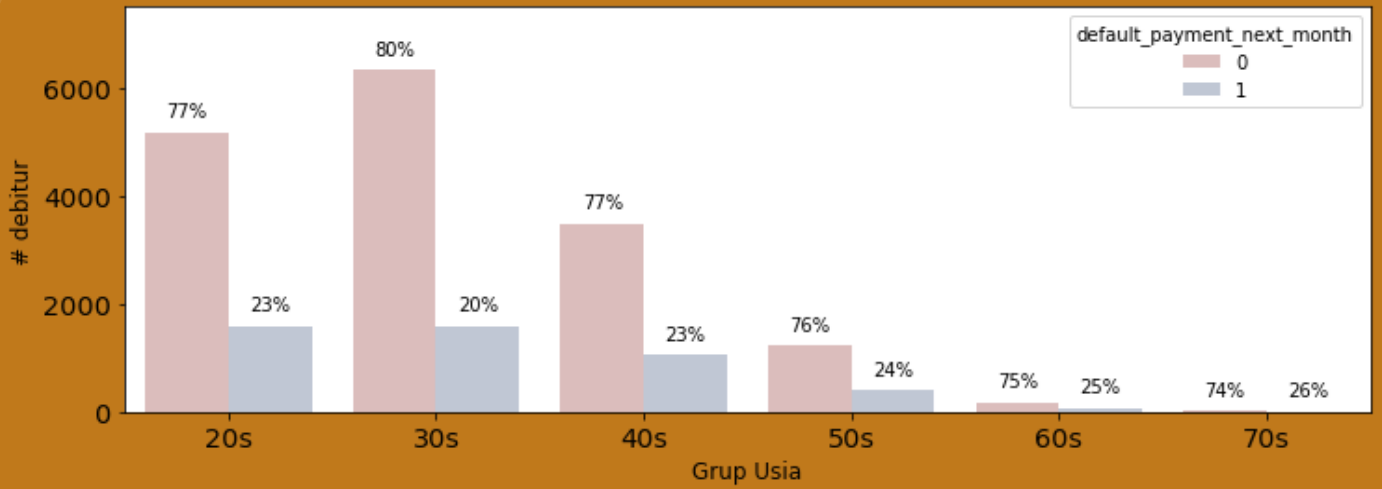
SEX & EDUCATION



1. Berdasarkan grafik diatas, di setiap kategori EDUCATION ternyata dapat disimpulkan bahwa memang laki-laki SMA lebih cenderung gagal bayar diikuti oleh laki laki yang berkuliah di universitas.
2. Nilai perempuan yang SMA dan Berkuliah di universitas memiliki kesamaan nilai gagal bayar

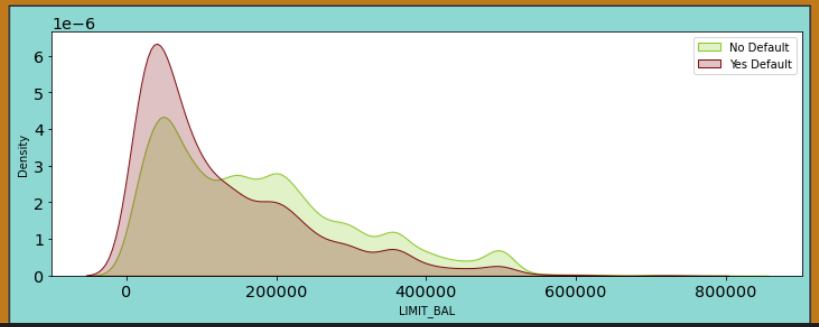
Numericals

AGE



1. Debitur didominasi oleh usia antara 25 sampai 40 tahun, berdasarkan grafik diatas juga terlihat bahwa dalam range usia 25 - 40 tahun tersebut memiliki peluang default yang lebih rendah.
2. Peluang default paling rendah yaitu debitur dengan usia 30an (30-39 tahun), sementara default yang tinggi berada pada usia-usia lanjut yaitu default tertinggi pada range usia 70-79 tahun, disusul oleh 60-69 tahun, kemudian 50-59 tahun.

LIMIT_BAL

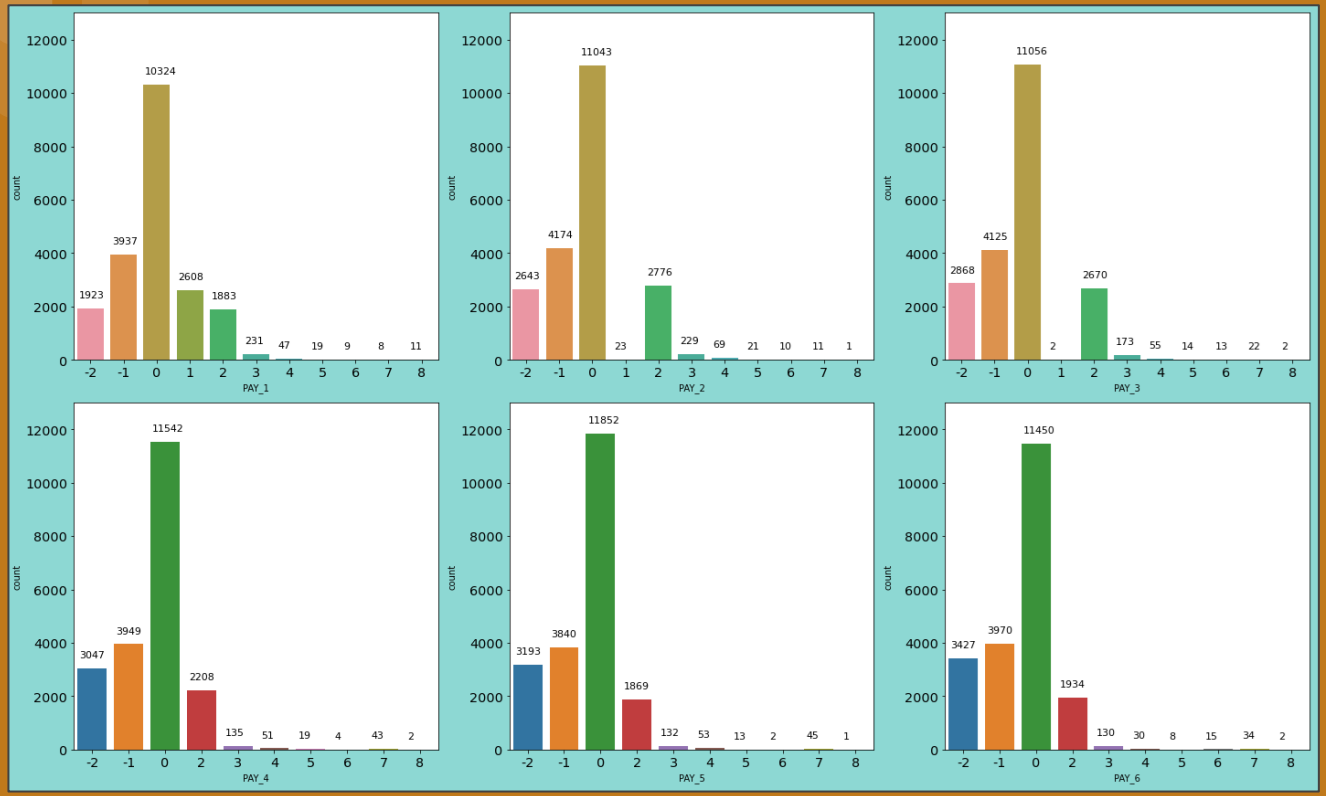



```
1 df_train_20D8GL3[df_train_20D8GL3['LIMIT_BAL'] > 600000].groupby(['default_payment_next_month']).count()
```

default_payment_next_month	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_1	PAY_2	PAY_3	PAY_4	PAY_5	P
0	43	43	43	43	43	43	43	43	43	43	43	
1	5	5	5	5	5	5	5	5	5	5	5	

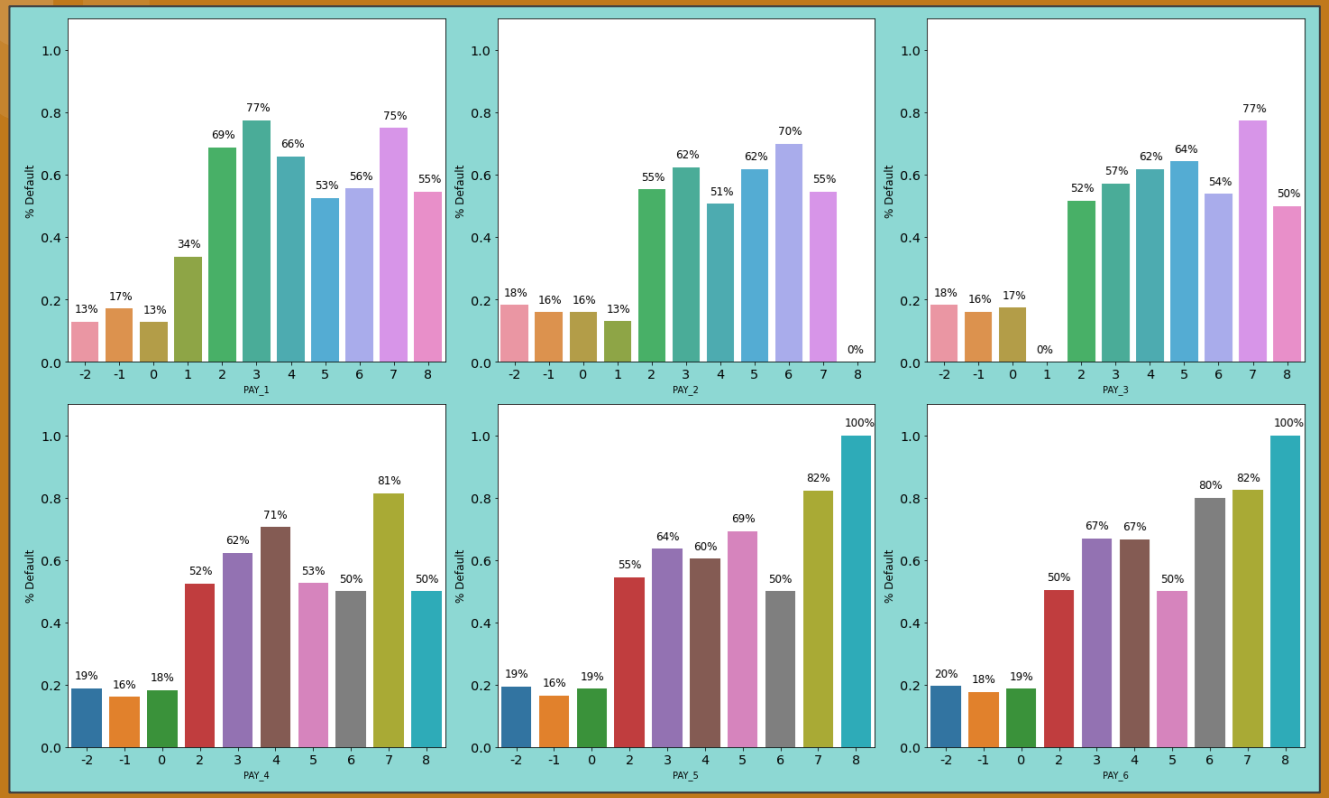
1. Dapat disimpulkan bahwa data pada kolom LIMIT_BAL memiliki distribusi yang right-skewed dan berdasarkan grafik tersebut terdapat keberadaan beberapa outlier dengan nilai ekstrim positif.
2. Terlihat bahwa terdapat sejumlah kecil debitur dengan LIMIT_BAL yang lebih dari 600000, yaitu sebanyak 48 orang, dimana dari 48 orang tersebut terdapat 5 orang yang default.
3. Sebagian besar debitur memiliki limit kredit sebesar 200000 atau kurang, dan tampak dalam range tersebut terdapat jumlah debitur default yang tinggi dibandingkan limit kredit lainnya.
4. Selain itu dapat pula disimpulkan bahwa, semakin besar nilai LIMIT_BAL maka kecenderungan untuk default juga akan semakin menurun.

PAY_X



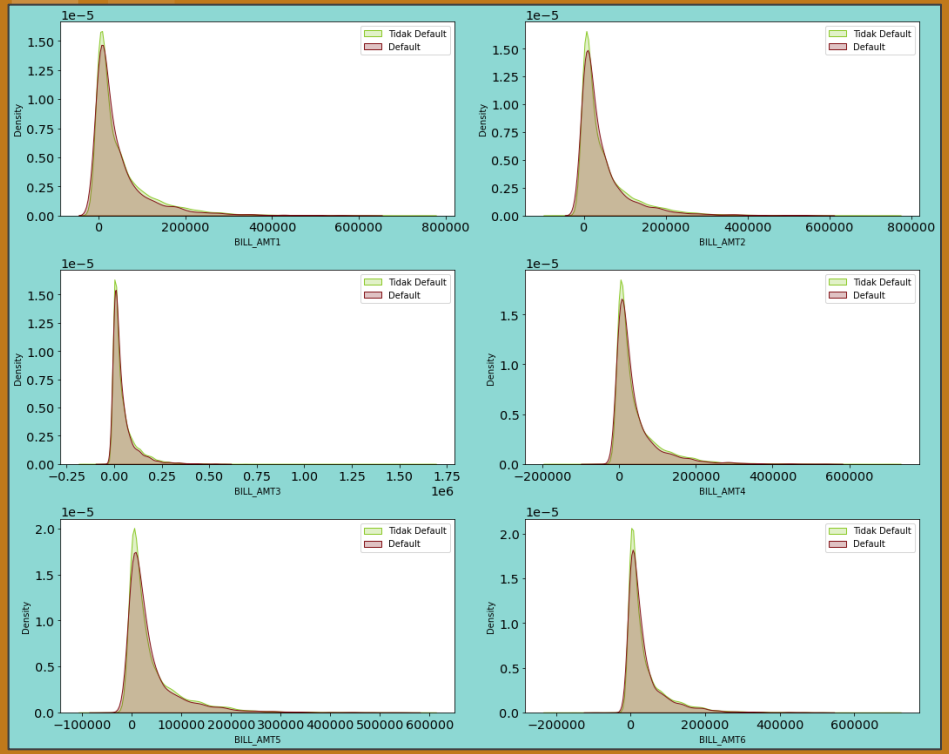
Dari semua repayment status terlihat bahwa jumlah debitur didominasi oleh kategori 0 dan disusul oleh kategori -1

PAY_X

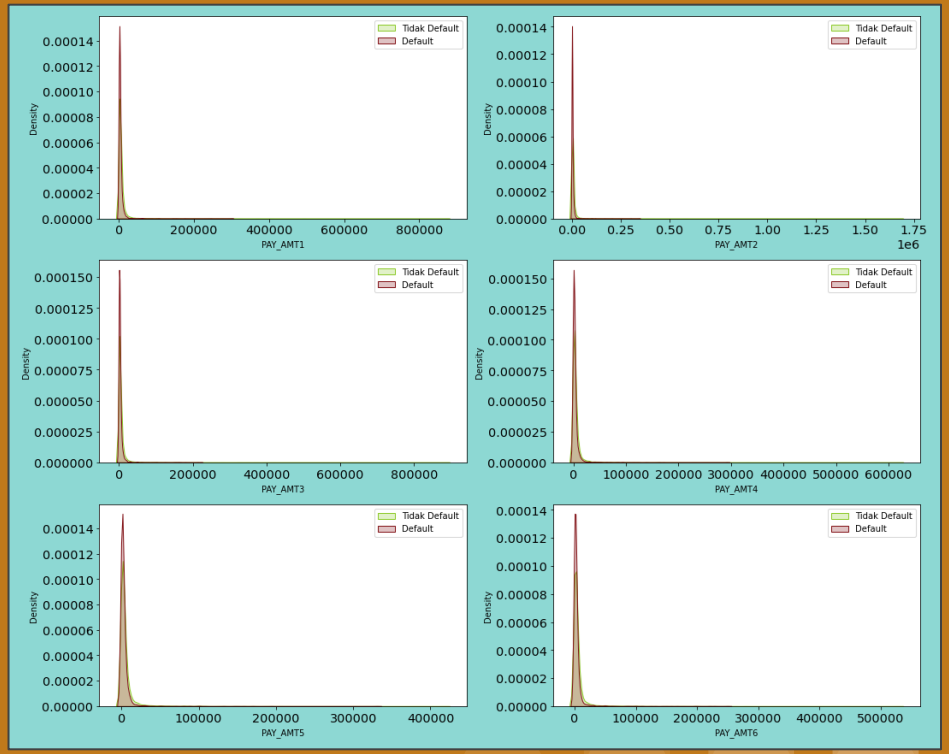


Namun terlihat jelas dari grafik kedua yang diagungkan dengan default payment, bahwa kemungkinan untuk default pada status pembayaran 0 dan -1 tersebut jauh lebih rendah dari status pembayaran lainnya.

BILL_AMT & PAY_AMT right skew

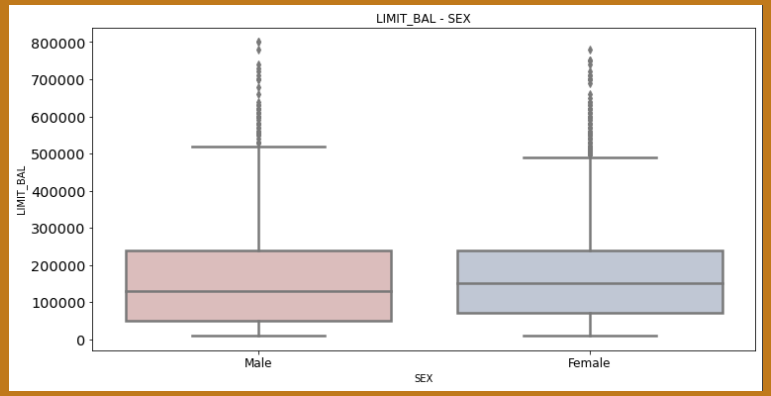


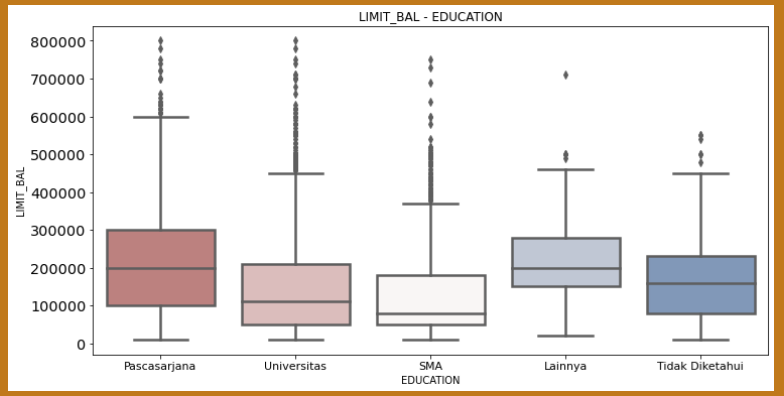
BILL_AMT

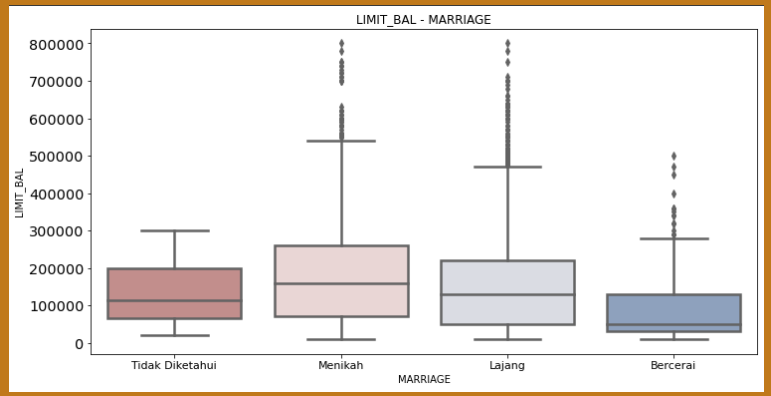


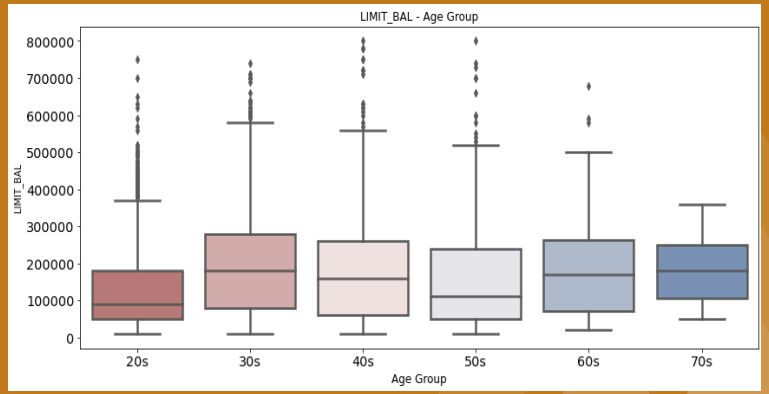
PAY_AMT

LIMIT_BAL vs Demografi







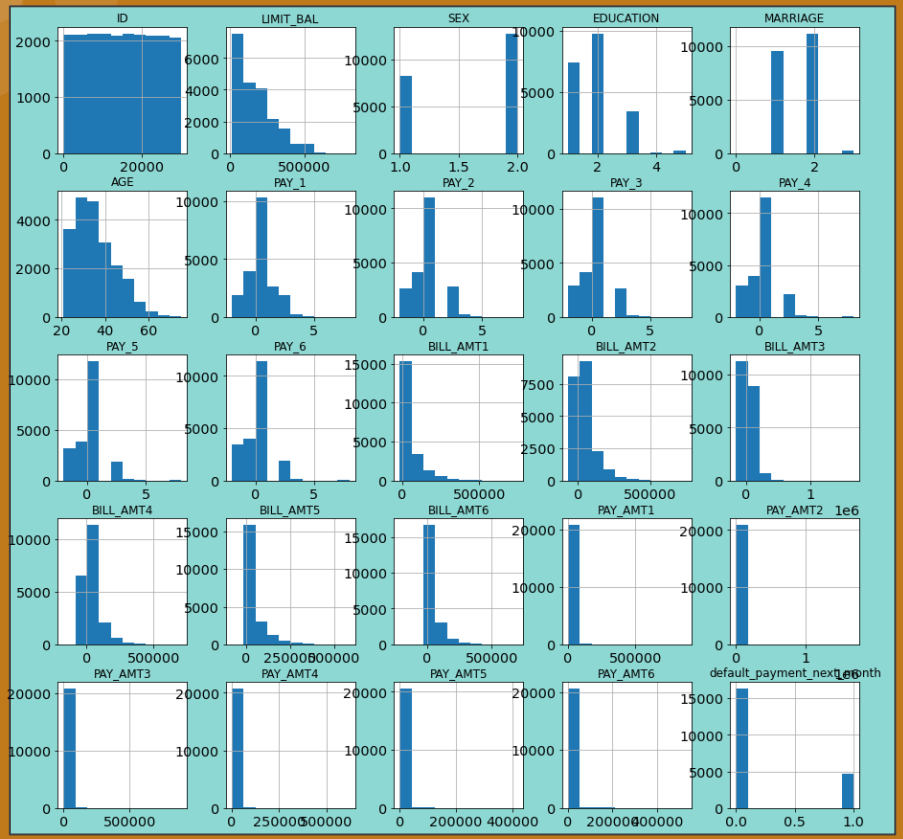


Eksplorasi Berbagai Attributes (Numerical Data)

1. Secara keseluruhan dari grafik-grafik antara Limit credit dengan data demografis, terlihat bahwa sebagian besar pemberian limit kredit berbanding lurus dengan probabilitas default pada kategori tersebut, maksudnya adalah terlihat bahwa semakin kecil peluang default yang ditunjukkan pada bagian univariate analysis untuk categorical data, maka limit kredit yang diberikan juga akan semakin besar
2. Namun terdapat pengecualian yaitu pada kategori usia 60 tahun keatas dimana walaupun memiliki peluang gagal bayar yang tinggi dibandingkan kategori usia lainnya, namun kategori ini tetap mendapatkan limit kredit yang tinggi

Target Output Olahan Data

Target Output Olahan Data



1. Terdapat 4645 (22%) dari 21000 entries orang yang gagal bayar bulan depan. Sisanya tidak gagal bayar berjumlah 16355 (78%) dari 21000
2. Jumlah kredit yang diberikan (LIMIT_BAL) memiliki rata-rata sekitar 167.214
3. Sebagian besar pelanggan adalah wanita
4. Dari pendidikan, yang tertinggi adalah dari universitas, dan yang kedua adalah dari sekolah pascasarjana, dan yang ketiga adalah sekolah menengah
5. Mayoritas sudah menikah, dan tertinggi kedua adalah single
6. usia rata-rata adalah sekitar 35 tahun

Pengecekan apakah ada data bermasalah

Categorical

Pengamatan:
1. Pada kolom `MARRIAGE` nilai 0 ada 36. Nilai 0 belum terdefinisi
2. Pada kolom `EDUCATION` nilai 0 ada 10. Nilai 0 belum terdefinisi
3. Pada kolom `PAY_1` nilai 0 ada 10324, dan nilai -2 ada 1923. nilai 0 dan -2 belum terdefinisi
4. Pada kolom `PAY_2` nilai 0 ada 11043, dan nilai -2 ada 2643. nilai 0 dan -2 belum terdefinisi
5. Pada kolom `PAY_3` nilai 0 ada 11056, dan nilai -2 ada 2868.
6. Pada kolom `PAY_4` nilai 0 ada 11542, dan nilai -2 ada 3047.
7. Pada kolom `PAY_5` nilai 0 ada 11852, dan nilai -2 ada 3193.
8. Pada kolom `PAY_6` nilai 0 ada 11450, dan nilai -2 ada 3427.

Numericals

Pengamatan:
1. Nilai negatif pada BILL_AMT1 ada 366 dari 19027 (0.019%)
2. Nilai negatif pada BILL_AMT2 ada 422 (0.022%)
3. Nilai negatif pada BILL_AMT3 ada 415
4. Nilai negatif pada BILL_AMT4 ada 428
5. Nilai negatif pada BILL_AMT5 ada 427
6. Nilai negatif pada BILL_AMT6 ada 455

Kesimpulan

1. Melihat dari urutnya deskripsi kolom, membuat kami `merubah nama kolom PAY_0 menjadi PAY_1

2. setelah melihat dari beberapa sample data, kami menympulkan untuk tiap-tiap data yang belum terdefinisi akan kami definisikan sebagai:

- Nilai 0 pada kolom MARRIAGE akan didefinisikan sebagai unknown karena sedikitnya entries data, hanya sebanyak 36 dari 21000.
- Nilai 0 pada kolom EDUCATION akan didefinisikan sebagai `unknown` juga karena sedikitnya entries data, hanya sebanyak 10 dari 21000
- Nilai 0 dan -2 pada kolom PAY_AMT1 - PAY_AMT6 akan didefinisikan sebagai, nilai 0 = OnTime` & nilai -2 = Telah membayar 2 bulan sebelumnya
- Nilai minus pada nilai min untuk kolom BILL_AMT1 - BILL_AMT6 akan didefinisikan sebagai kelebihan bayar

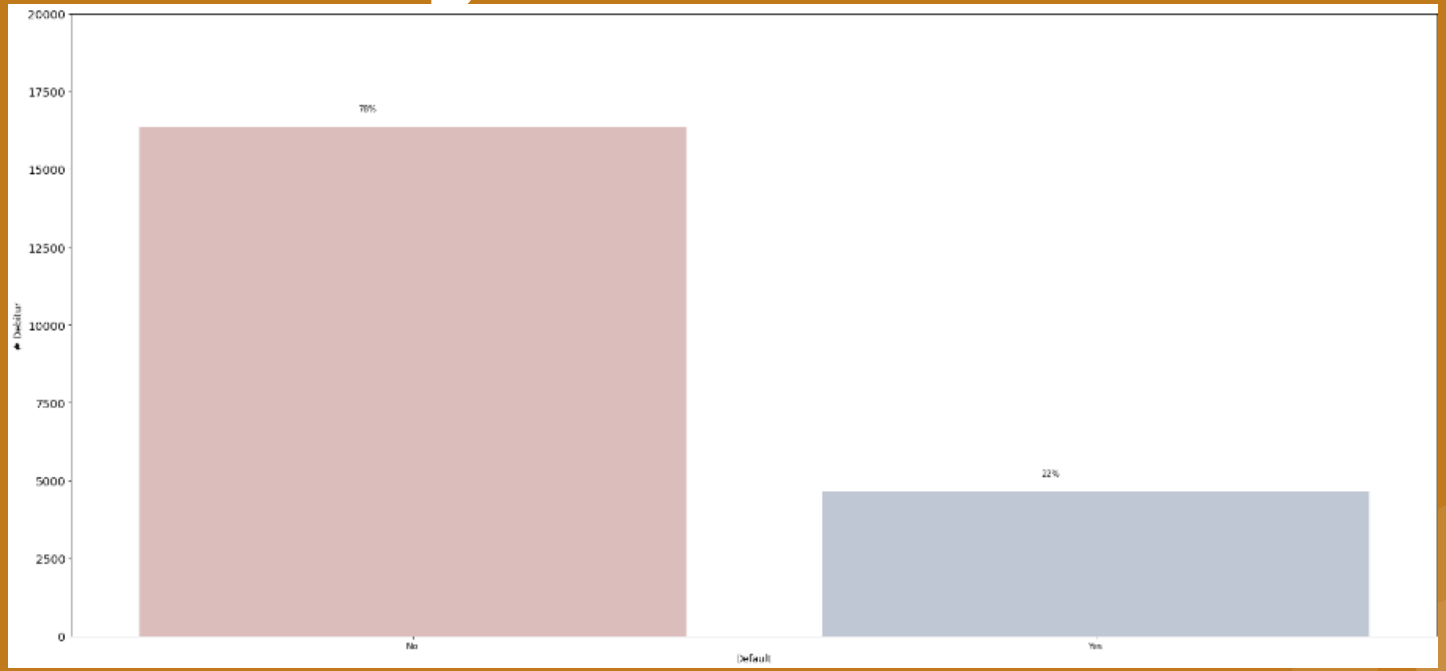
Handling Outlier Data

```
print(f'Jumlah baris sebelum memfilter outlier: {len(df_train_20D8GL3)}')

filtered_entries = np.array([True] * len(df_train_20D8GL3))

for col in numericals:
    zscore = abs(stats.zscore(df_train_20D8GL3[col])) # hitung absolute z-scorenya
    filtered_entries = (zscore < 3) & filtered_entries # keep yang kurang dari 3 absolute z-scorenya

df_train_20D8GL3 = df_train_20D8GL3[filtered_entries] # filter, cuma ambil yang z-scorenya dibawah 3

print(f'Jumlah baris setelah memfilter outlier: {len(df_train_20D8GL3)}')
```

```
Jumlah baris sebelum memfilter outlier: 21000
Jumlah baris setelah memfilter outlier: 19027
```

Nilai entries data menjadi 19027, setelah membuang 0,094% dari entries data aslinya

Handling imbalanced data



Dilihat dari visualisasi diatas, nilainya adalah 7:3 yang berarti termasuk imbalanced

Handling Data Bermasalah Lainnya

```
 0   ID                          19027 non-null  int64
 1   LIMIT_BAL                   19027 non-null  float64
 2   SEX                         19027 non-null  int64
 3   EDUCATION                   19027 non-null  int64
 4   MARRIAGE                    19027 non-null  int64
 5   AGE                         19027 non-null  int64
 6   PAY_1                       19027 non-null  int64
 7   PAY_2                       19027 non-null  int64
 8   PAY_3                       19027 non-null  int64
 9   PAY_4                       19027 non-null  int64
 10  PAY_5                       19027 non-null  int64
 11  PAY_6                       19027 non-null  int64
 12  BILL_AMT1                   19027 non-null  float64
 13  BILL_AMT2                   19027 non-null  float64
 14  BILL_AMT3                   19027 non-null  float64
 15  BILL_AMT4                   19027 non-null  float64
 16  BILL_AMT5                   19027 non-null  float64
 17  BILL_AMT6                   19027 non-null  float64
 18  PAY_AMT1                    19027 non-null  float64
 19  PAY_AMT2                    19027 non-null  float64
 20  PAY_AMT3                    19027 non-null  float64
 21  PAY_AMT4                    19027 non-null  float64
 22  PAY_AMT5                    19027 non-null  float64
 23  PAY_AMT6                    19027 non-null  float64
 24  default_payment_next_month  19027 non-null  int64
 25  AGE_BIN                     19027 non-null  category
dtypes: category(1), float64(13), int64(12)
memory usage: 4.3 MB
```

Merubah range nilai ftrs, agar persebaran data mendekati distribusi normal

One Hot Encoding

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 19027 entries, 0 to 20999
Columns: 104 entries, ID to PAY_6_8
dtypes: category(1), float64(13), int64(12), uint8(78)
memory usage: 5.7 MB
```

Kolomnya bertambah menjadi 104 kolom, agar memudahkan algoritma machine learning

Drop Outdate Columns

```
[82] df_train_20D8GL3 = df_train_20D8GL3.drop(columns=['ID', 'AGE', 'AGE_BIN', 'EDUCATION', 'MARRIAGE','PAY_1','PAY_2','PAY_3','PAY_4'
```

```
[83] df_train_20D8GL3.head()
```

	LIMIT_BAL	SEX	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_A
0	1.761155	2	-0.397004	-0.540791	-0.712392	-0.690787	-0.534897	-0.678256	1.513495	-0.047927	0.026852	1.089062	-0.176
1	0.743843	2	-0.793698	-0.794430	-0.785910	-0.717856	-0.705773	-0.672282	-0.645628	-0.578376	-0.183319	-0.548482	-0.128
2	1.422051	2	0.162232	0.289651	0.032249	0.033473	0.087047	0.142101	0.211848	-0.273432	-0.225144	0.005366	0.014
3	-0.866903	2	-0.750750	-0.794430	-0.785910	-0.768595	-0.745729	-0.727111	-0.645628	-0.578376	-0.573685	-0.548482	-0.569
4	-0.612574	1	-0.108444	-0.421304	-0.376270	-0.302857	-0.240373	-0.208157	-0.388385	-0.349668	-0.312279	-0.358697	-0.398

5 rows × 93 columns

Setelah melakukan drop pada beberapa kolom diatas, jumlah kolom menjadi 93 dan akan dipakai untuk membuat model Machine Learning

Penjelasan proses handling untuk setiap data yang bermasalah

1. Merubah nama kolom `PAY_0` menjadi `PAY_1`, karena melihat dari deskripsi kolom yang sudah berurutan
2. Mendefinisikan nilai-nilai yang belum terdefinisi
3. Menghandling outlier dengan membuang 9,4% (1973) entries data
4. Kami telah menyimpulkan datanya tidak terlalu imbalance sehingga tidak perlu melakukan handling
5. Melakukan Standarization agar persebaran data mendekati distribusi normal
6. Melakukan OneHot Encoding agar memudahkan jalannya algoritma Machine Learning kedepannya
7. Melakukan Drop Outdated Columns

Terima Kasih