Kelompok 5 – Finpro Rakamin

Loan Prediction

STAGE 1

EDA

1. Descriptive Statistics

Setelah melakukan function **info** terhadap dataset kita bisa melihat bahwa terdapat 13 variabel berbeda dimana ada 6 variabel dengan tipe data **int64** yaitu ['Income', 'Age', 'Experience', 'CURRENT_JOB_YRS', 'CURRENT_HOUSE_YRS'].

Untuk variable tersebut bisa dikategorikan sebagai tipe data numerical. Selain itu terdapat 6 variabel dengan tipe data **object** yaitu ['Married/Single', 'Car_Ownership', 'House_Ownership', 'Profession', 'CITY', 'STATE']

Setelah melakukan function **describe** kita bisa melihat bahwa untuk nilai summary di setiap kolom masih tidak ada masalah. Bisa dilihat bahwa untuk setiap kolom perbedaan antara mean dan median tidak begitu jauh.

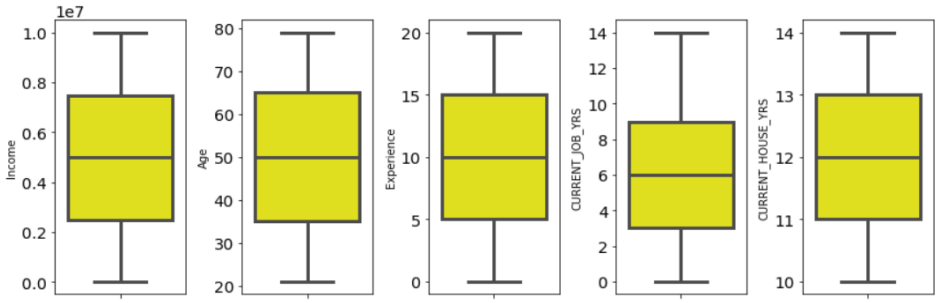
	Id	Income	Age	Experience	CURRENT_JOB_YRS	CURRENT_HOUSE_YRS	Risk_Flag
count	252000.000000	2.520000e+05	252000.000000	252000.000000	252000.000000	252000.000000	252000.000000
mean	126000.500000	4.997117e+06	49.954071	10.084437	6.333877	11.997794	0.123000
std	72746.278255	2.878311e+06	17.063855	6.002590	3.647053	1.399037	0.328438
min	1.000000	1.031000e+04	21.000000	0.000000	0.000000	10.000000	0.000000
25%	63000.750000	2.503015e+06	35.000000	5.000000	3.000000	11.000000	0.000000
50%	126000.500000	5.000694e+06	50.000000	10.000000	6.000000	12.000000	0.000000
75%	189000.250000	7.477502e+06	65.000000	15.000000	9.000000	13.000000	0.000000
max	252000.000000	9.999938e+06	79.000000	20.000000	14.000000	14.000000	1.000000

	House_Ownership	Profession	CITY	STATE
count	252000	252000	252000	252000
unique	3	51	317	29
top	rented	Physician	Vijayanagaram	Uttar_Pradesh
freq	231898	5957	1259	28400

Di dataset ini tidak terdapat nilai kosong di 13 variabel tersebut.

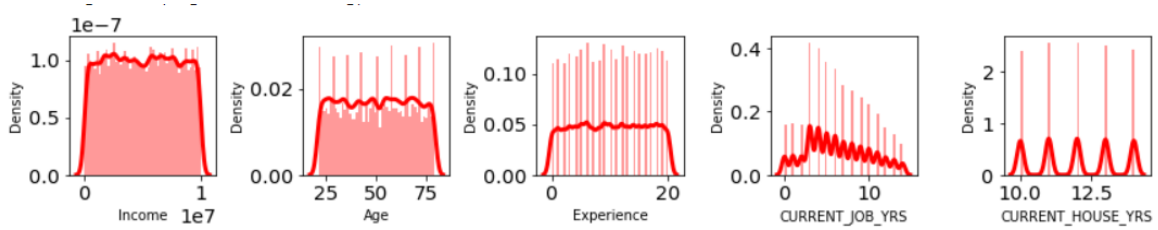
2. Univariate Analysis

Untuk tahap awal univariate analysis kita akan memvisualisasikan tiap distribusi dari data numerik menggunakan boxplot

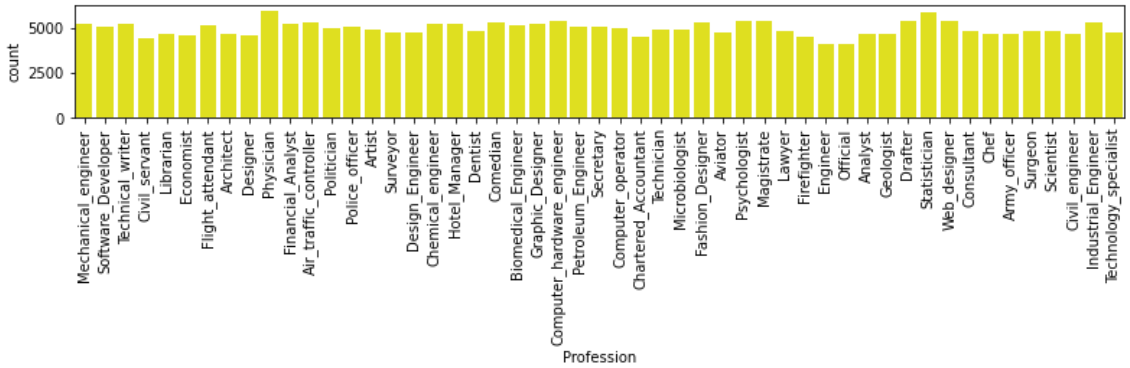


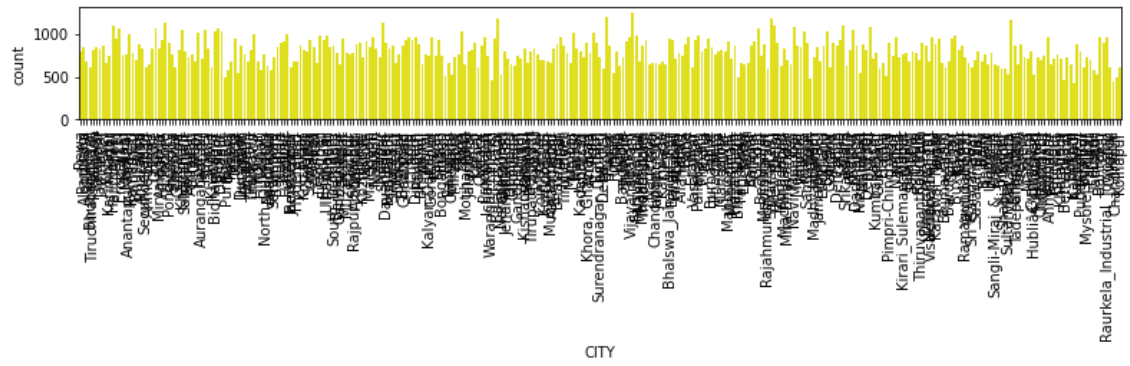
Dari distribusi boxplot diatas untuk tipe data yang numerik bisa dilihat bahwa tidak terdapat outlier sama sekali di masing-masing kolom.

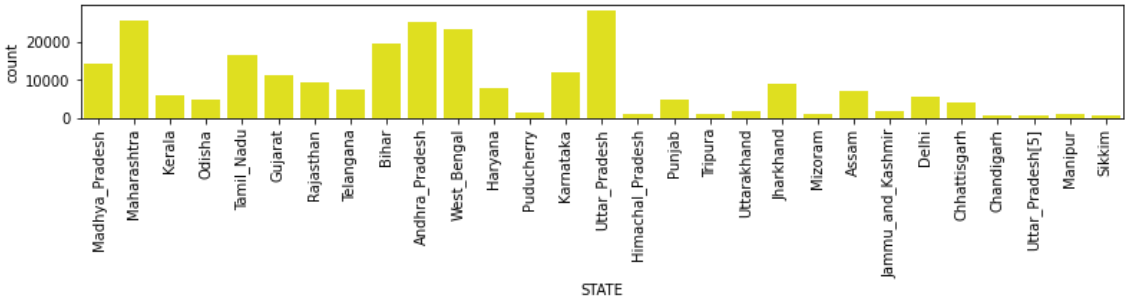
Setelah itu kita bisa melihat distribusinya menggunakan displot atau kdeplot seperti plot dibawah



Dari distribution plot diatas bisa dilihat bahwa untuk variabel 'Income', 'Age' dan 'Experience' distribusinya mendekati normal. Sedangkan untuk 'CURRENT_JOB_YRS' distribusinya skew ke kanan, dimana nilai mean lebih besar daripada nilai median. Terakhir untuk distribusi 'CURRENT_HOUSE_YRS' distribusinya merupakan multimodal dimana memiliki beberapa puncak. Di tahap data pre-processing sepertinya perlu di normalisasi agar distribusi yang skew menjadi normal.





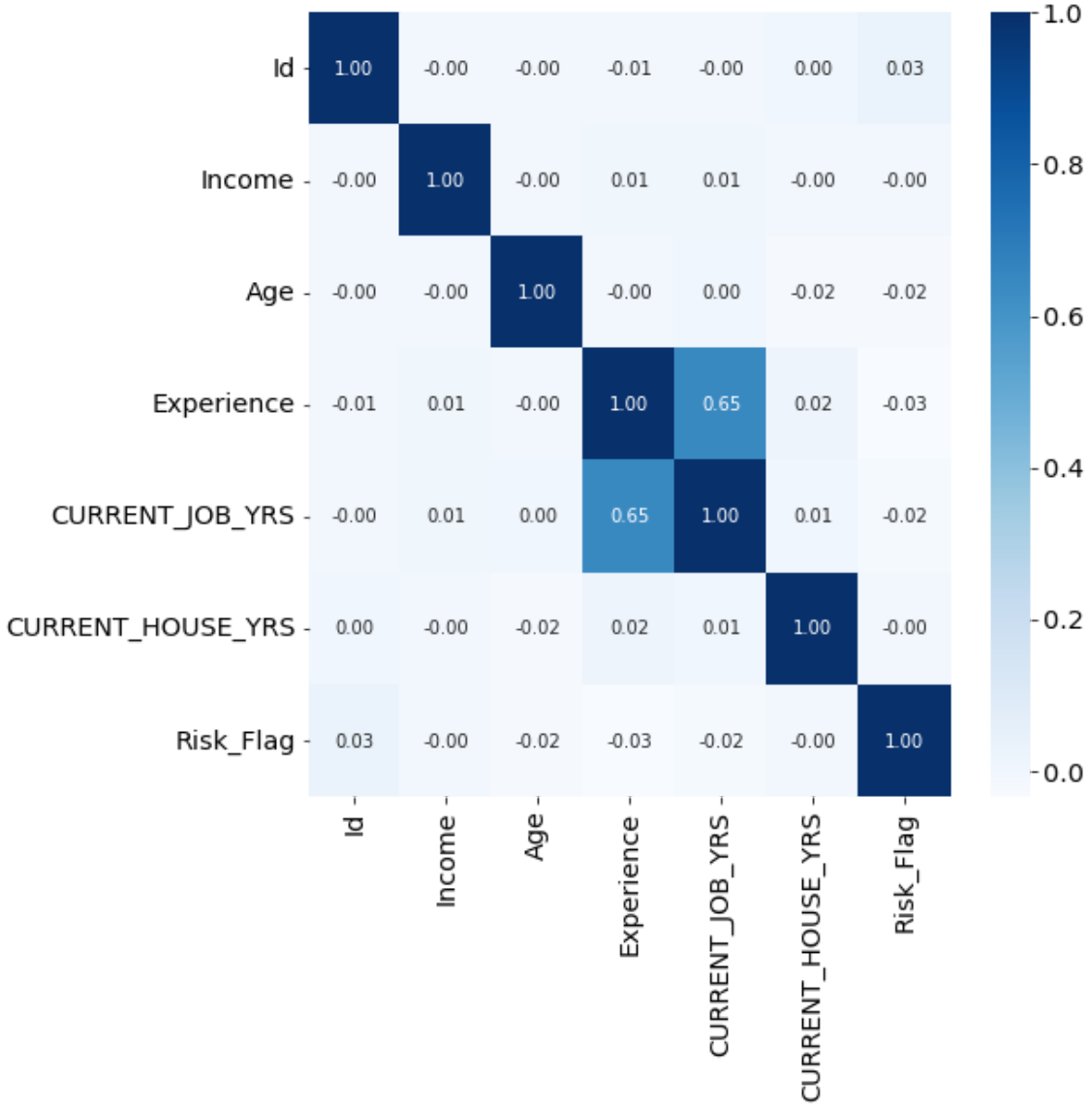


Untuk data kategorik, dapat dilihat untuk kolom 'Married/Single' didominasi oleh value 'singl e', kolom 'Car_Ownership' didominasi oleh value 'no', dan kolom 'House_Ownership' didomi nasi oleh value 'rented'. Ketiga kolom tersebut nantinya akan dilakukan encoding untuk meng ubah valuenya menjadi numerik.

Untuk kolom 'Profession','CITY', dan 'STATE', memiliki banyak unique value, sehinggauntu k data tersebut akan dilihat apakah kolom tersebut berpengaruh terhadap target atau tidak. Jik a tidak akan didrop, jika iya maka akan dilakukan encoding atau dilakukan pengerucutan dul u lalu encoding.

3. Multivariate Analysis

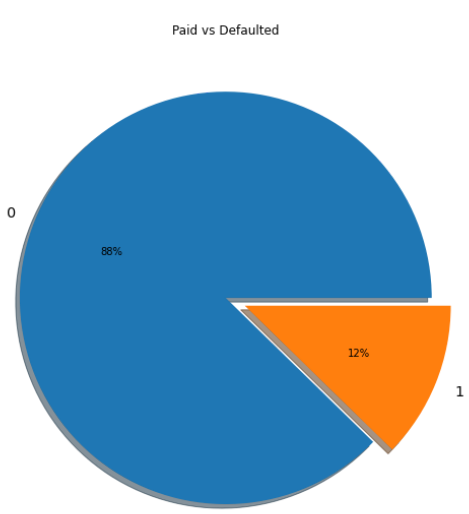
Untuk multivariate analysis bisa kita visualisasikan menggunakan heatmap

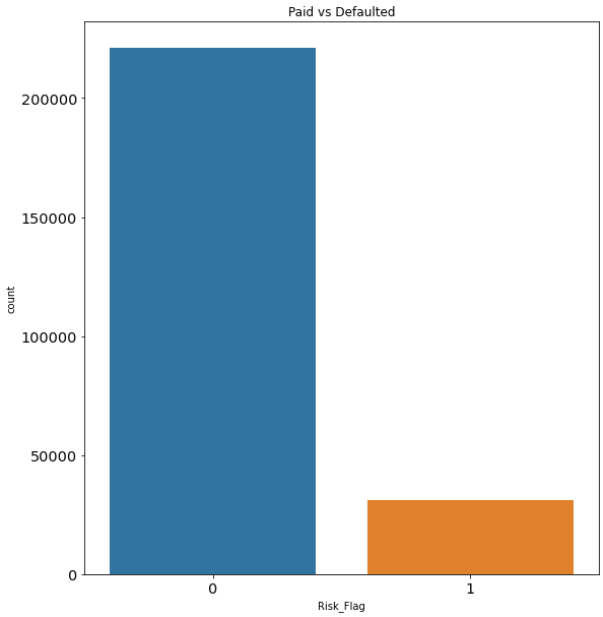


Pada heatmap dapat dilihat bahwa ada 3 fitur yang berkorelasi dengan target yaitu, CURRENT_JOB_YRS (negative correlation), Experience (negative correlation), dan Age (negative correlation). Lalu ada 2 fitur yang saling berkorelasi dengan nilai yang cukup besar yaitu CURRENT_JOB_YRS dan Experience dengan nilai koefisien korelasi sebesar 0.65.

4. Business Insight

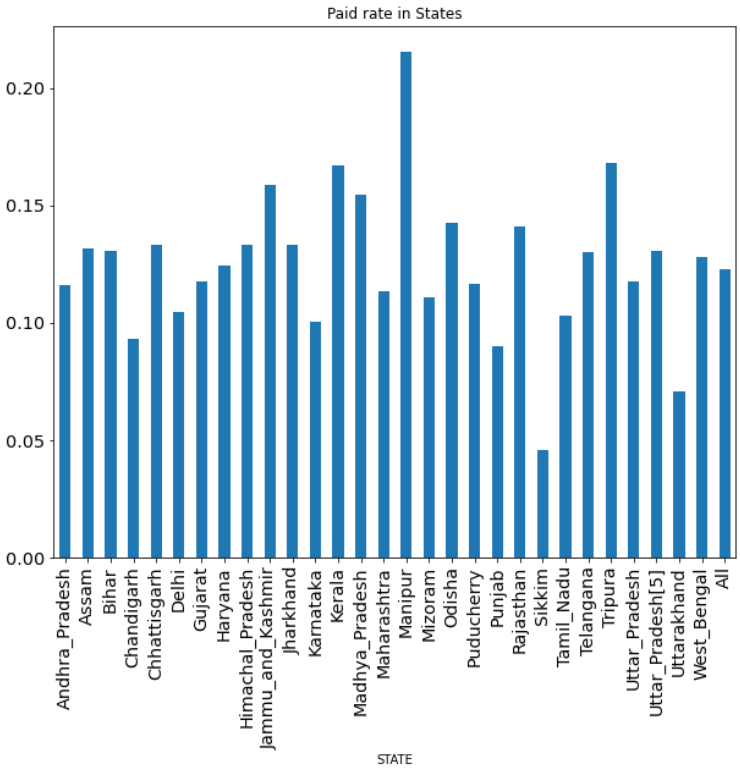
- Total User (Paid Vs Defaulted)





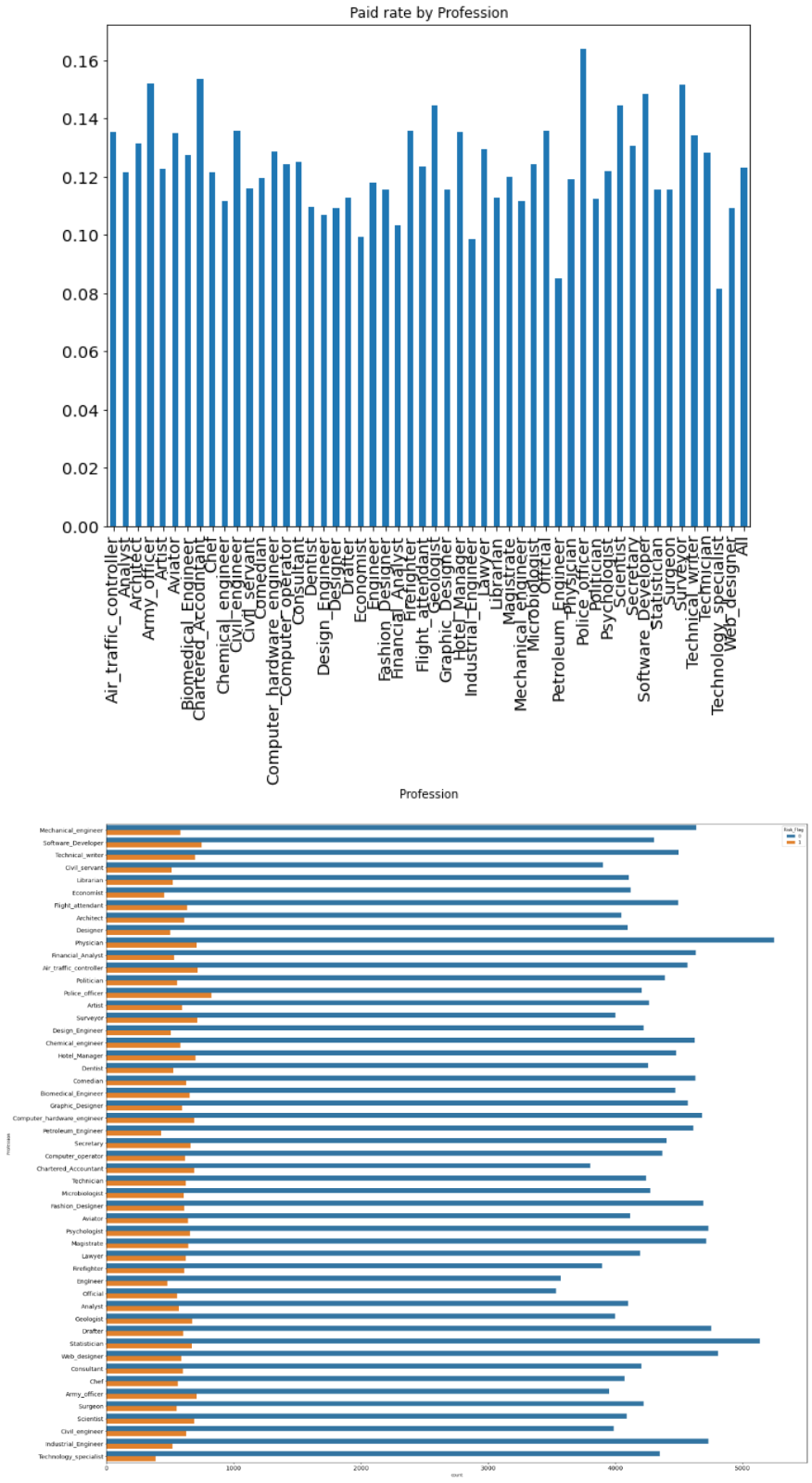
Presentase user yang default sebesar 12% dengan banyak nya user yang default sebesar 30996 user.

- Presentase Pinjaman yang berhasil bayar tiap state



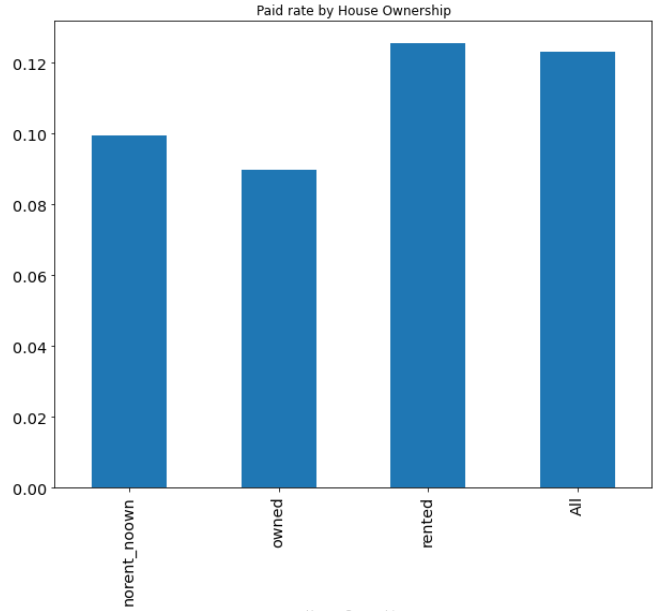
Dari hasil visualisasi diatas dapat kita lihat bahwa Manipur memiliki presentase yang default paling tinggi yaitu sebesar 21% dan yang berhasil bayar yaitu Sikkim.

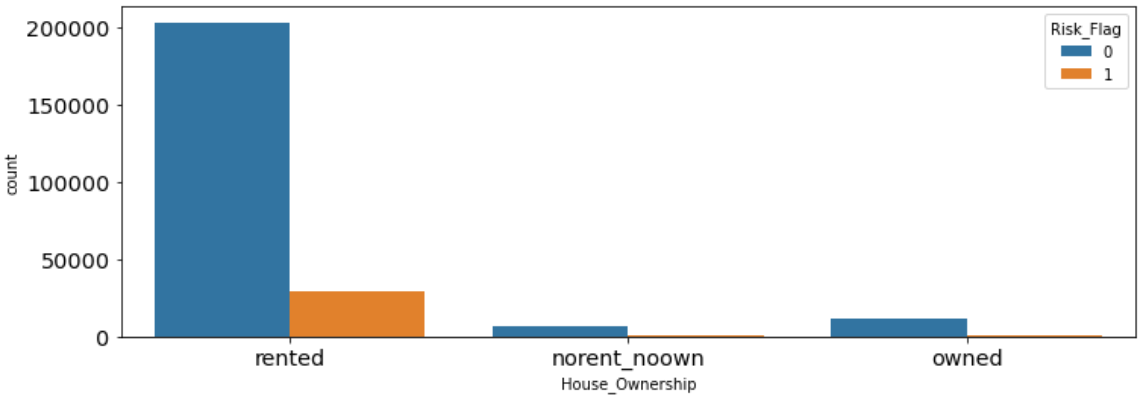
- Presentase pinjaman yang berhasil bayar tiap profesi



Profesi yang paling banyak default yaitu Police_officer.

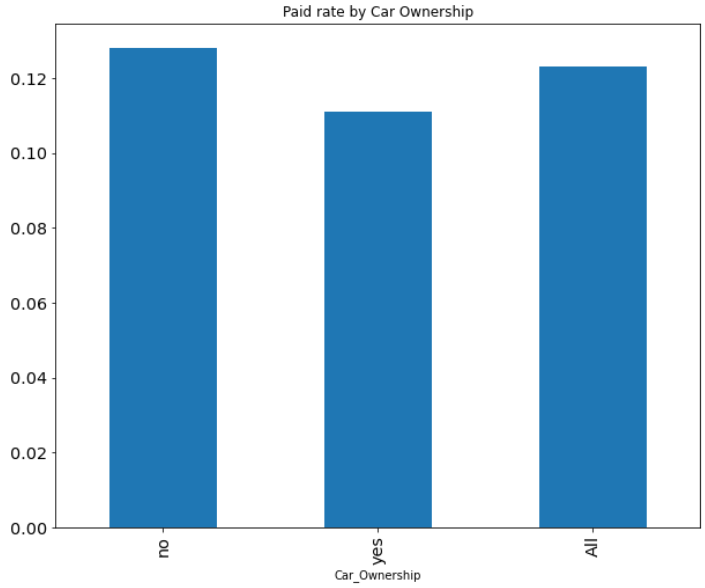
- House Ownership

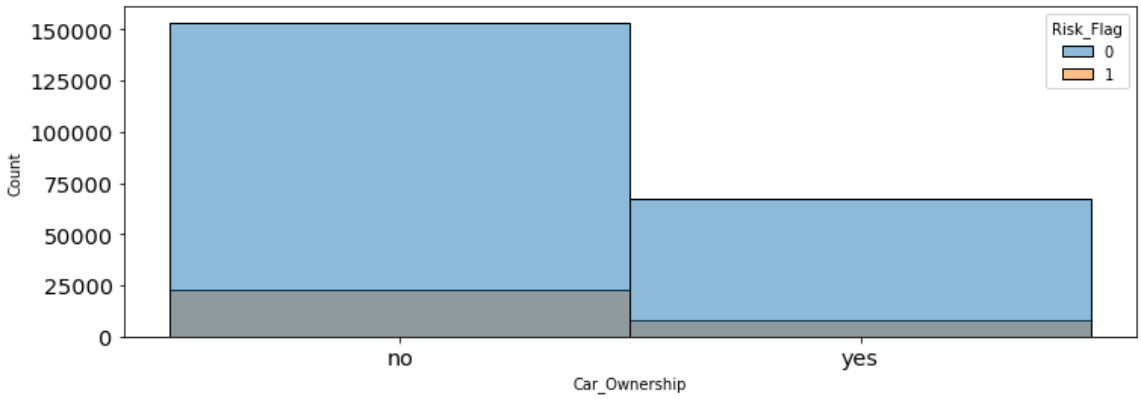




Dari hasil visualisasi didaptkan bahwa user yang rented cenderung yang paling banyak melakukan peminjaman dan risk flag yang paling tinggi.

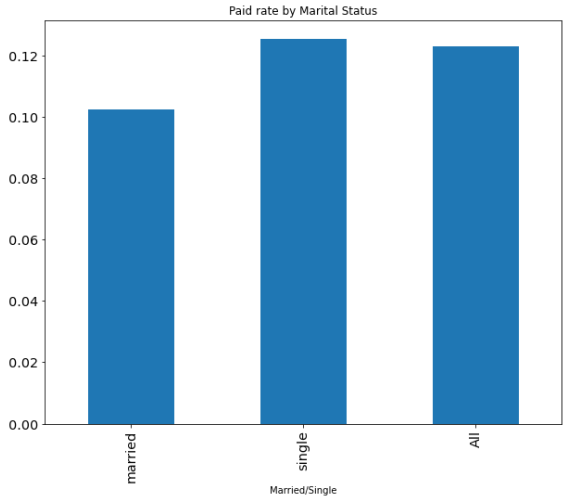
- Presentase pinjaman yang berhasil bayar berdasarkan car ownership

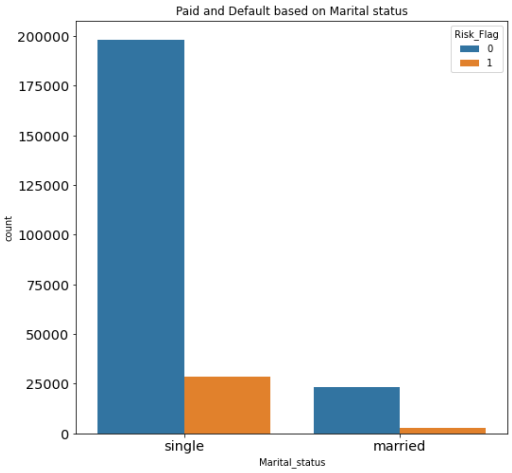




User yang paling banyak melakukan pinjaman dan default lebih banyak pada user yang tidak memiliki mobil dibandingkan dengan yang punya mobil.

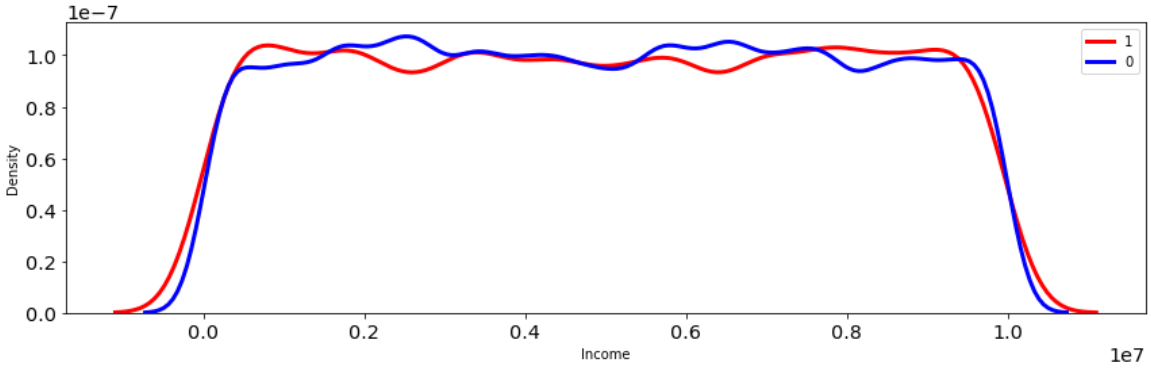
- Martial Status





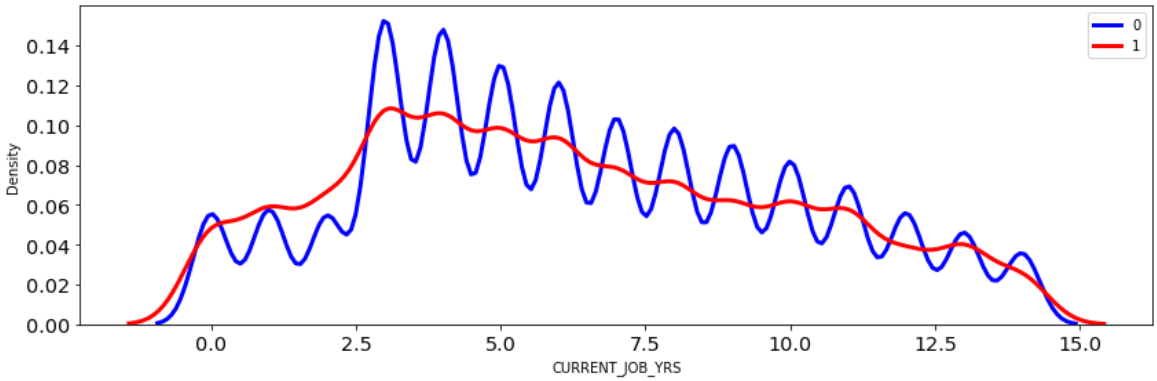
Default user didominasi oleh user dengan marital status single.

- Income dan Risk Flag



Dari hasil visualisai terlihat bahwa income dan risk flag terdistribusi secara uniform.

- Current Job Years dan Risk Flag



Dari grafik diatas didapat insight sebagai berikut:

- 88% data merupakan user dengan risk flag 0
- Data didominasi oleh user dengan martial status single, no car ownership, rented house ownership, dan user dengan current job years 3-5 tahun
- Presentase user yang berhasil membayar loan relative kecil untuk semua data kategorik (rata-rata kurang dari 20%)
- Distribusi income, age, dan experience uniform
- Distribusi data numerik untuk data dengan risk flag 1 hampir sama
- Perbandingan data kategorik dengan risk flag 0 dan dengan risk flag 1 hampir sama

Dari data yang diperoleh cukup sulit untuk menentukan alasan user gagal melakukan pembayaran, karena sebagian besar distribusi datanya uniform. sehingga kami berhipotesa bahwa alasan user gagal melakukan pembayaran yaitu total loan nya. Untuk mengatasi permasalahan tersebut dan mengurangi default rate dan revenue loss, kami merekomendasikan untuk melakukan segmentasi customer, dimana customer akan dibagi ke dalam beberapa tier berdasarkan data customer, dan jumlah pinjamannya berdasarkan tier customernya.

5. GIT

https://github.com/drestantav/raka_project_repo.git