# Stage 1 - Exploratory Data Analysis (EDA) & Business Insight

## 1. Descriptive Statistics

```
[ ] # melihat informasi awal dari dataset
    data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 252000 entries, 0 to 251999
Data columns (total 13 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   Id                 252000 non-null  int64
 1   Income             252000 non-null  int64
 2   Age                252000 non-null  int64
 3   Experience         252000 non-null  int64
 4   Married/Single     252000 non-null  object
 5   House_Ownership    252000 non-null  object
 6   Car_Ownership      252000 non-null  object
 7   Profession         252000 non-null  object
 8   CITY               252000 non-null  object
 9   STATE              252000 non-null  object
 10  CURRENT_JOB_YRS    252000 non-null  int64
 11  CURRENT_HOUSE_YRS  252000 non-null  int64
 12  Risk_Flag          252000 non-null  int64
dtypes: int64(7), object(6)
memory usage: 25.0+ MB
```

```
[ ] # Melihat deskripsi data kategori
    data[kategorikal].describe()
```

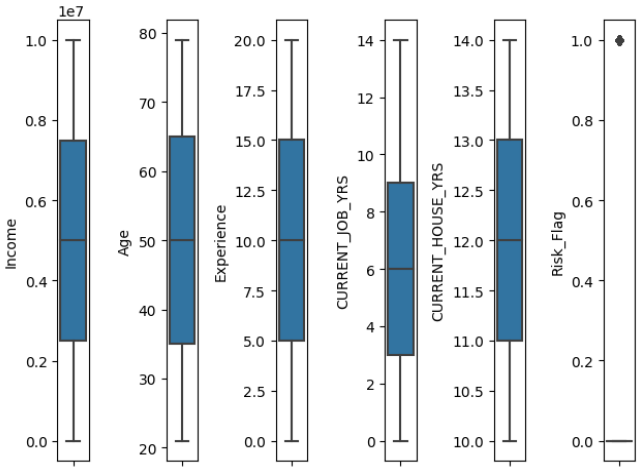
| | Married/Single | House_Ownership | Car_Ownership | Profession | CITY | STATE |
|---|---|---|---|---|---|---|
| count | 252000 | 252000 | 252000 | 252000 | 252000 | 252000 |
| unique | 2 | 3 | 2 | 51 | 317 | 29 |
| top | single | rented | no | Physician | Vijayanagaram | Uttar_Pradesh |
| freq | 226272 | 231898 | 176000 | 5957 | 1259 | 28400 |

```
[ ] # Melihat deskripsi data numerik
    data.describe()
```

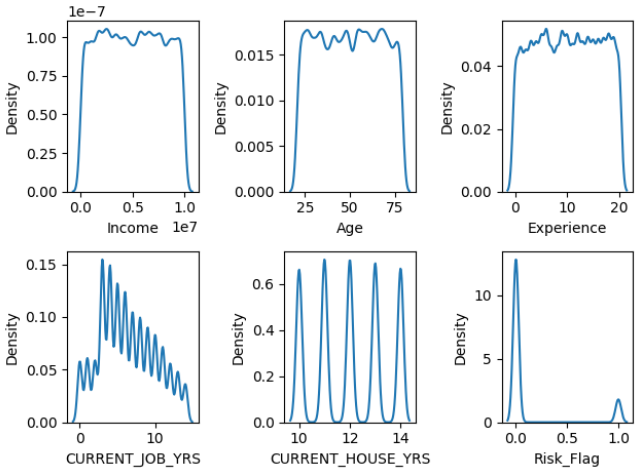
| | Id | Income | Age | Experience | CURRENT_JOB_YRS | CURRENT_HOUSE_YRS | Risk_Flag |
|---|---|---|---|---|---|---|---|
| count | 252000.000000 | 2.520000e+05 | 252000.000000 | 252000.000000 | 252000.000000 | 252000.000000 | 252000.000000 |
| mean | 126000.500000 | 4.997117e+06 | 49.954071 | 10.084437 | 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 |

1. Data terdiri dari 252000 baris and 13 kolom dan tidak ada nilai null, semua tipe data sudah sesuai.
2. 7 kolom dengan variabel kontinu: Id, Income, Age, Experience, CURRENT_JOB_YRS, CURRENT_HOUSE_YRS, Risk_Flag.
3. 6 kolom dengan variabel kategori : Married/Single, House_Ownership, Car_Ownership, Profession, CITY, STATE.
4. Masing-masing kolom kategorikal memiliki nilai unique sebagai berikut:
   a. Married/Single: married, single
   b. House_Ownership: rented, owned, norent_noown
   c. Car_Ownership: yes, no
   d. Profession: Physician, …, Engineer (ada 51 jenis profesi)
   e. City: Vijayanagaram, …, Karaikudi (ada 317 nama kota)
   f. State: Uttar Pradesh, …, Sikkim (ada 29 state)
5. Dari hasil pengecekan nilai unique tiap kolom sudah sesuai dan tidak redundan, untuk memperkecil scope distribusi, kolom-kolom yang memiliki banyak nilai unique akan dikelompokkan di proses pre-processing.

## 2. Univariate Analysis



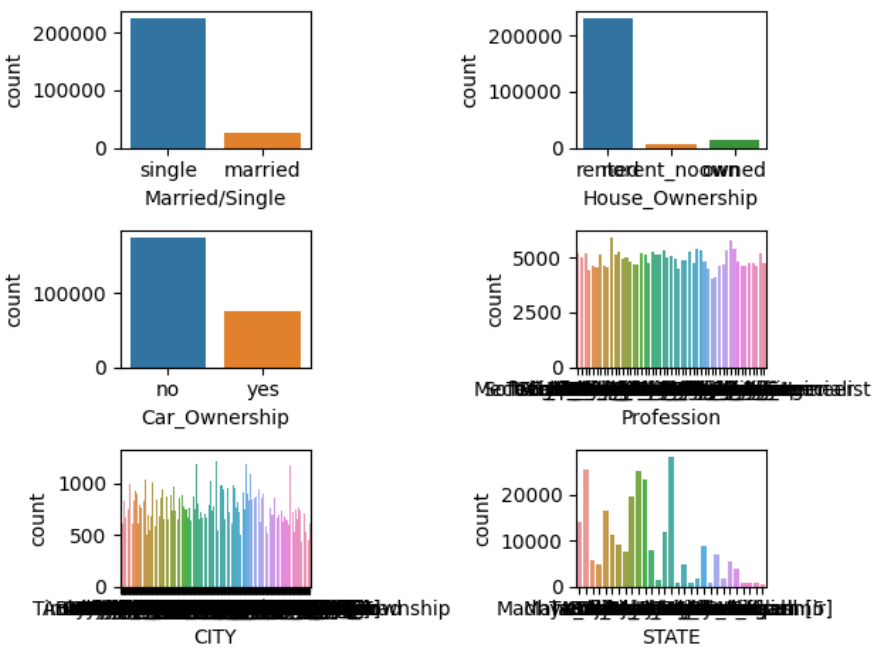

*Distribusi kolom numerik menggunakan boxplot*




*Distribusi kolom numerik menggunakan distplot*

Distribusi pada kolom numerik dibagi menjadi 2, yang pertama menggunakan boxplot dan yang kedua menggunakan distribution plot. Pada boxplot tidak ada outlier yang terlihat kecuali pada kolom Risk_Flag karena inputnya hanya angka 0 dan 1, tidak adanya outlier menandakan bahwa distribusi data normal. Lalu pada distplot, grafik pada setiap kolom memiliki distribusi yang berbeda:

- Income, age dan experience : uniform distribution, karena data memiliki nilai yang seragam dan tidak terjadi lonjakan sehingga data memiliki probabilitas yang sama.
- Current_job_yrs : Skewness positive, ekor distribusi berada di sebelah kanan dengan nilai terbanyak sehingga distribusi sebagian besar berada pada nilai rendah.
- current_house_yrs : multimodal distribution curve, karena data memiliki modus yang lebih dari satu.



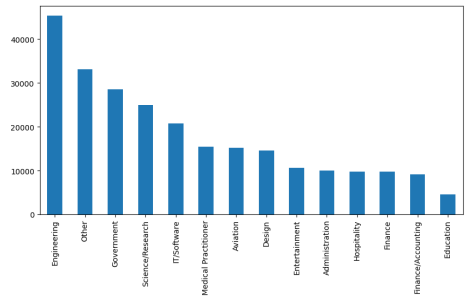

*Distribusi kolom kategorikal menggunakan countplot*

- Kolom kategorikal dengan banyak kategori terjadi pada kolom city, state dan profession sehingga data perlu dikelompokan menjadi beberapa kategori yang serupa menjadi satu kategori baru untuk menyederhanakan data.
- Kolom kategorikal dengan Dominasi Kategori Tertentu pada kolom married/single, dan house ownership yang mengindikasikan ketidakseimbangan dalam data sehingga perlu diperhatikan apakah akan mempengaruhi analisis seperti oversampling dan undersampling diperlukan

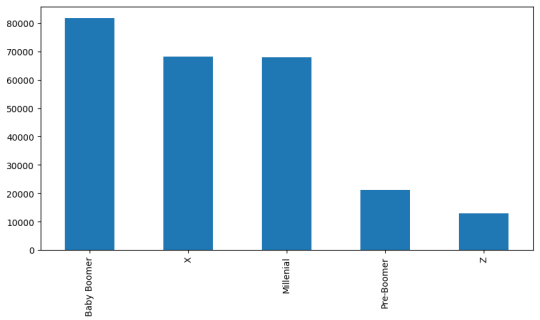
**Rekomendasi Data Pre Processing yang Dapat Dilakukan:**

1. **Transformasi Data** untuk data distribusi yang miring atau non-normal, pertimbangkan untuk menerapkan transformasi pada data numerik.
2. **Feature Engineering** untuk kolom yang terlalu banyak kategori, dapat dikelompokan menjadi beberapa kategori yang lebih umum atau menggunakan metode seperti one-hot encoding atau label encoding.
3. **Handling Imbalance** untuk data kategorikal yang tidak seimbang sehingga akan diputuskan apakah oversampling (menambahkan data minoritas) atau undersampling (mengurangi data mayoritas) diperlukan untuk menyeimbangkan dataset atau tidak
4. **Feature Selection** Berdasarkan hasil observasi distribusi,dipilih fitur-fitur yang memiliki dampak signifikan dalam analisis.

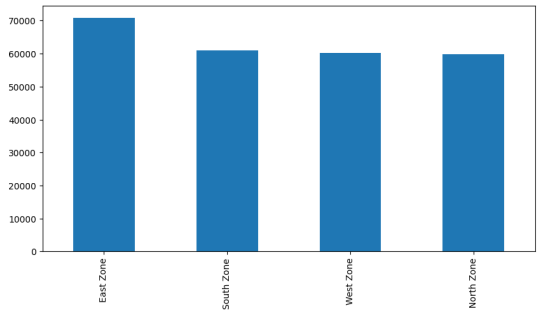
**Hasil grouping pada kolom yang memiliki terlalu banyak kategori:**



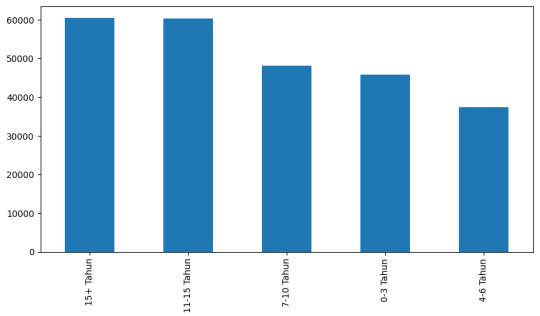

*Pengelompokan profesi berdasarkan field*



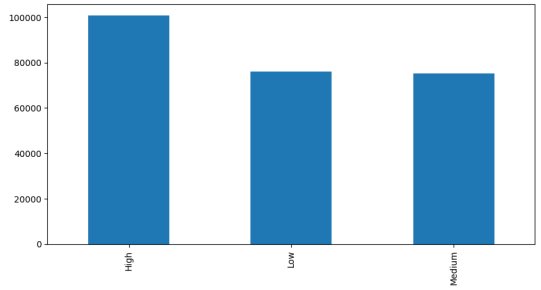

*Pengelompokan umur berdasarkan generasi*



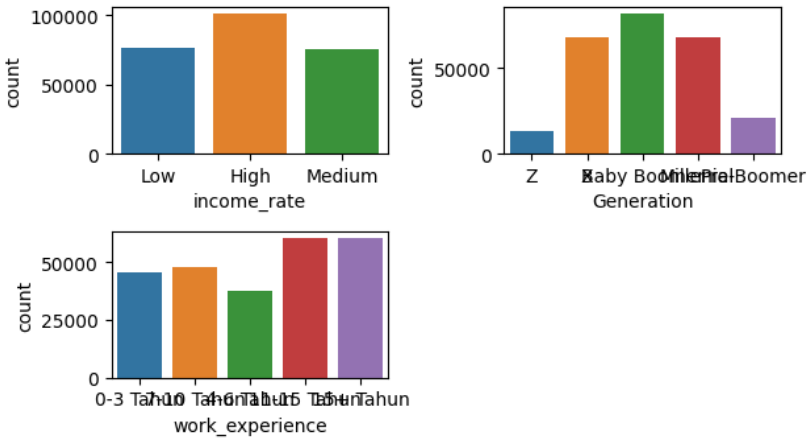

*Pengelompokan state berdasarkan zona*




*Pengelompokan experience berdasarkan tahun kerja*

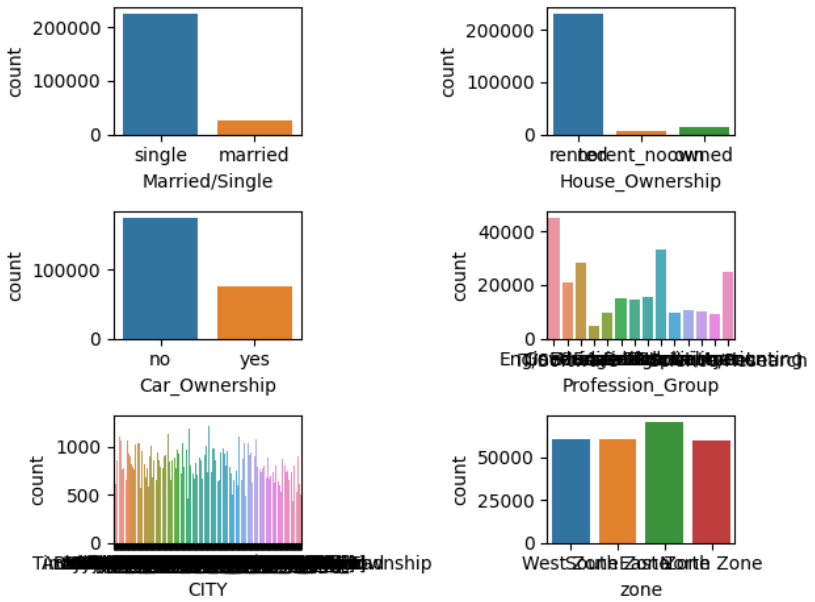



*Pengelompokan income berdasarkan level pendapatan*




*Distribusi kolom numerik menggunakan count plot setelah grouping*

Pada kolom Income dilakukan grouping berdasarkan income rate apakah low, medium, atau high. Lalu pada kolom Age dilakukan grouping berdasarkan generasi umur, dan pada kolom Experience dilakukan grouping berdasarkan lama bekerja dalam tahun




*Distribusi kolom kategorik menggunakan count plot setelah grouping*

Pada kolom Profession dilakukan grouping berdasarkan field/bidang pekerjaan. Lalu pada kolom STATE dilakukan grouping berdasarkan zone/daerah bagian tersebut apakah termasuk zona barat/timur/utara/selatan. Untuk kolom CITY tidak bisa diterapkan grouping karena nama kota tidak bisa mewakili satu sama lain, sehingga untuk kolom CITY akan diterapkan Label Encoding.

### 3. Multivariate Analysis

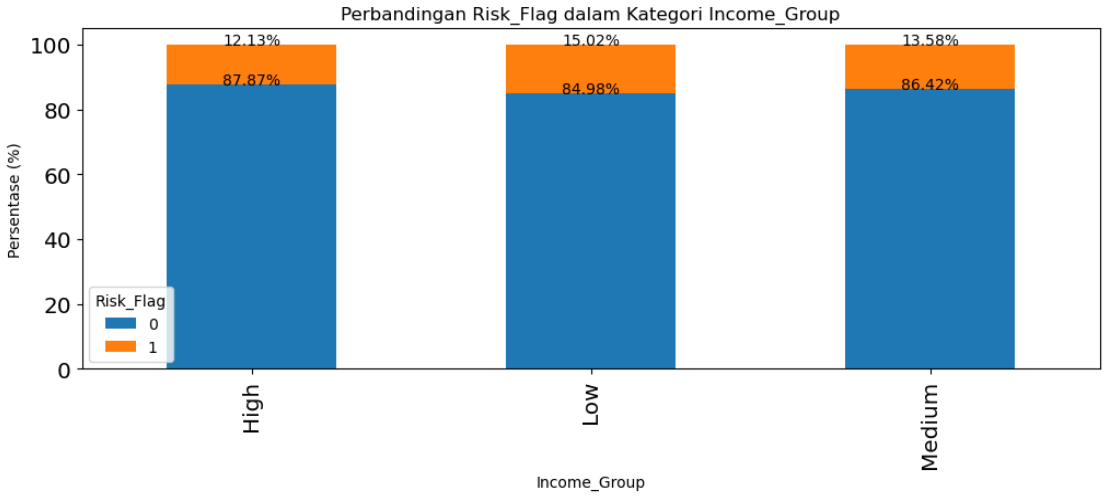

*Correlation heatmap antara feature dan label*

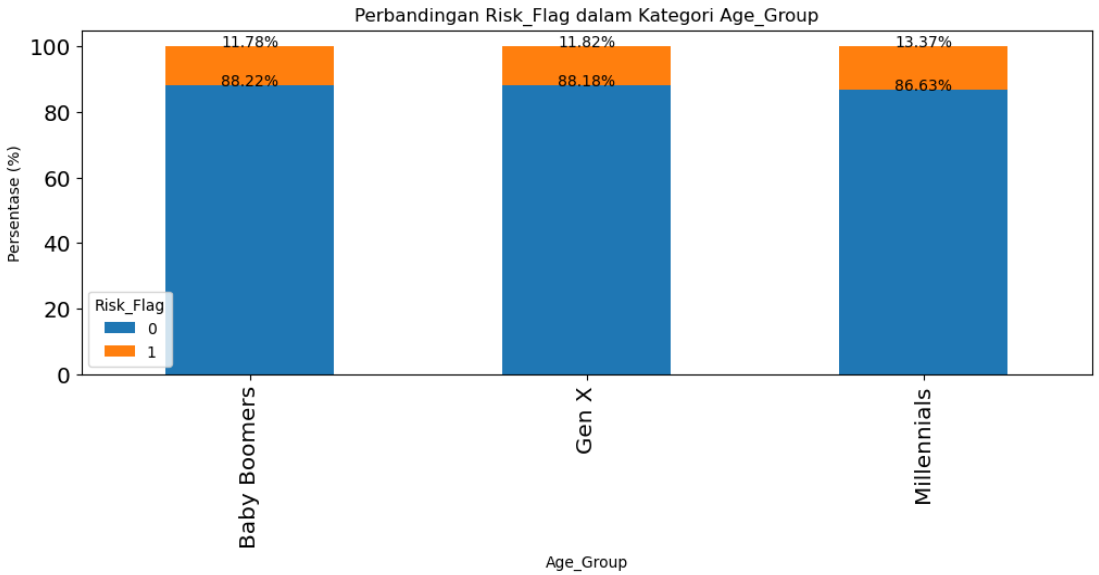
Feature yang relevan dan harus dipertahankan:

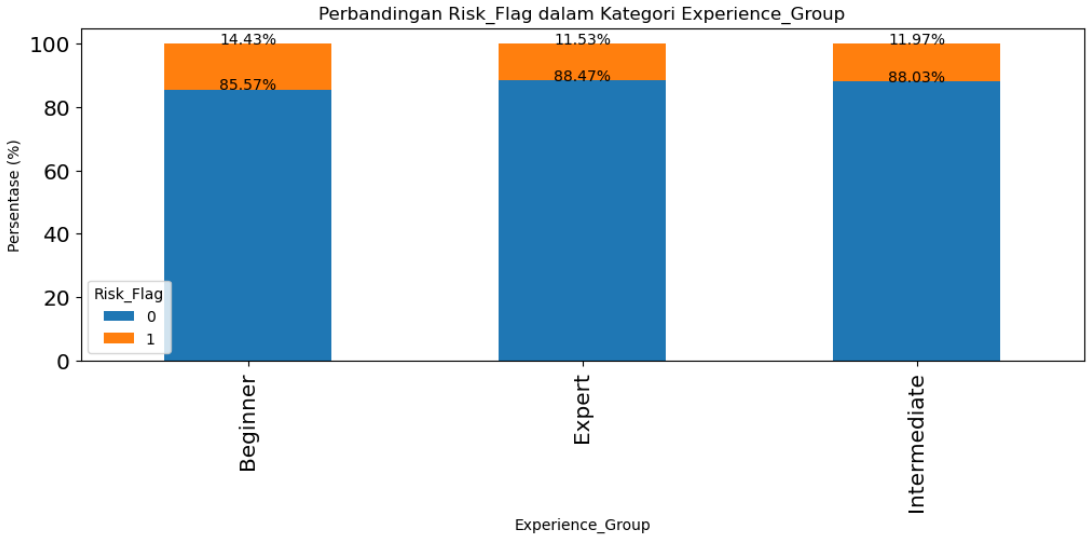
- **Experience**: Fitur yang relevan karena korelasi negatif yang lemah antara Experience dengan Risk_Flag **-0.03**. Artinya, semakin tinggi pengalaman seseorang, sedikit lebih rendah juga tanda risiko.
- **Age**: Fitur yang relevan karena terdapat perbedaan yang cukup signifikan dalam tingkat risiko antara perbedaan umur seseorang.
- **Income** mungkin adalah fitur yang relevan karena terdapat perbedaan yang signifikan dalam tingkat risiko antara perbedaan pendapatan.

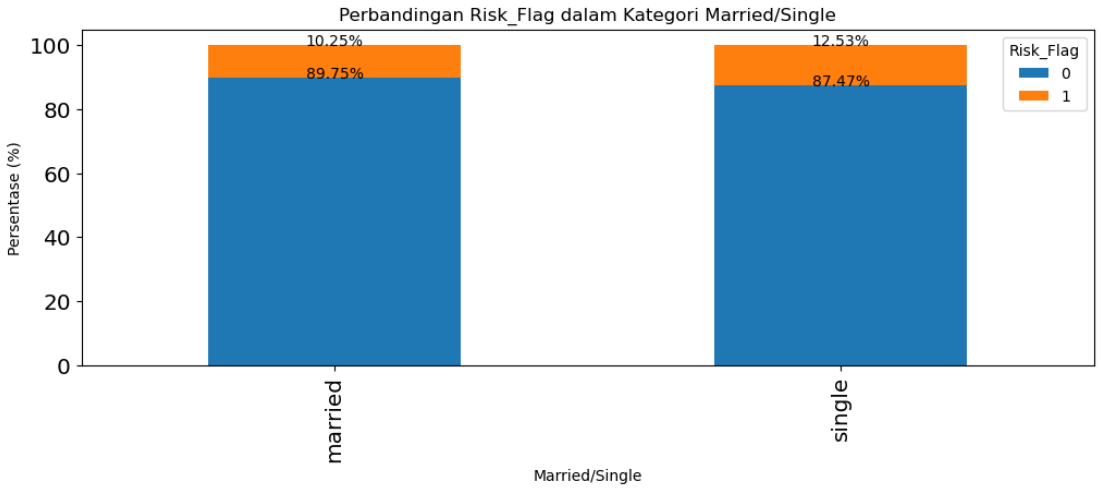
Pola menarik pada korelasi antar fitur ditemukan pada fitur **Experience** dan **CURRENT_JOB_YRS**, korelasi positif yang kuat antara dua fitur ini memiliki korelasi **0.65**. Artinya, semakin tinggi pengalaman seseorang, biasanya semakin tinggi juga tahun pekerjaan saat ini. Mengingat korelasi yang kuat antara dua fitur ini, mungkin lebih baik jika menggabungkan kedua fitur ini menjadi satu fitur baru untuk menghindari multikolinearitas dalam model.

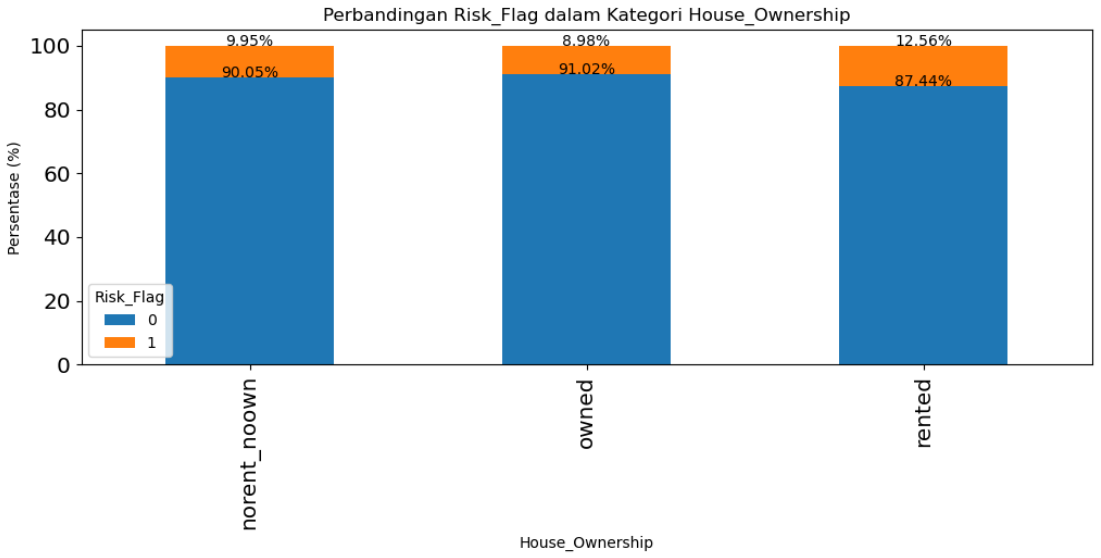
## Category plots antara fitur dan target:

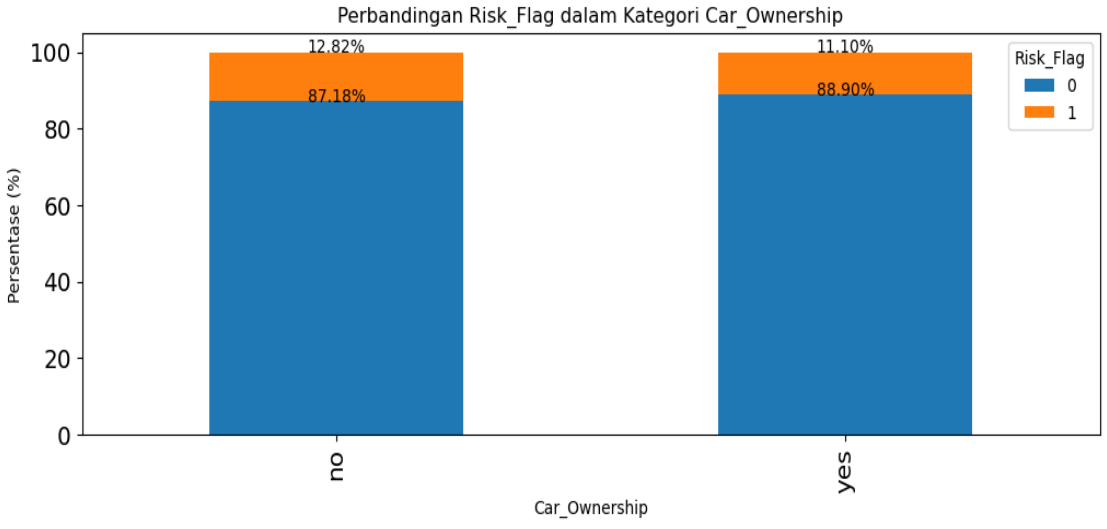


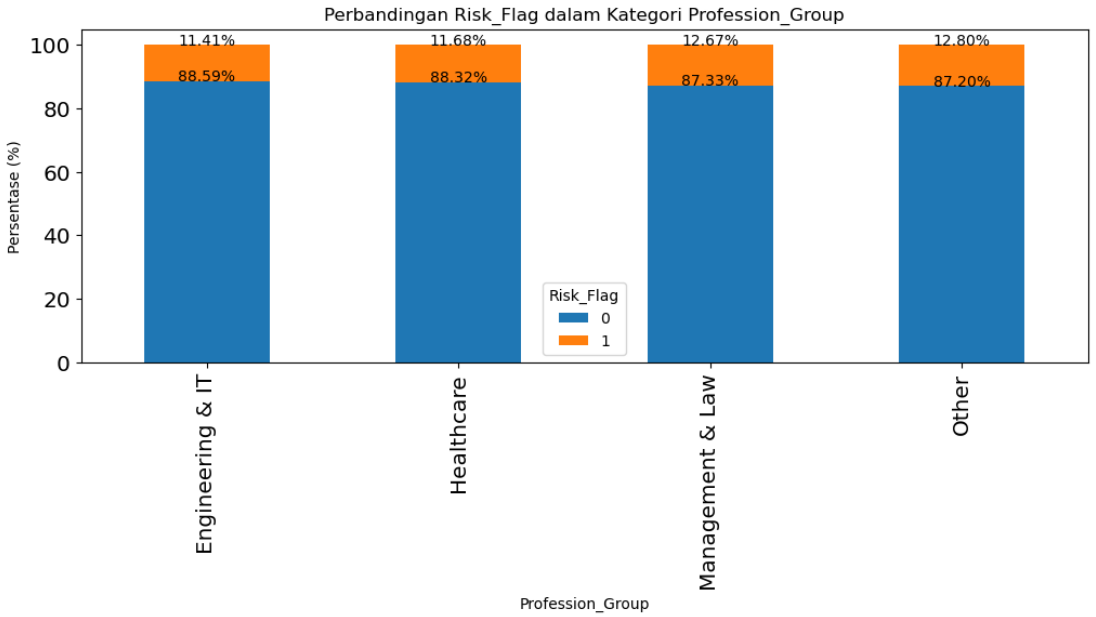



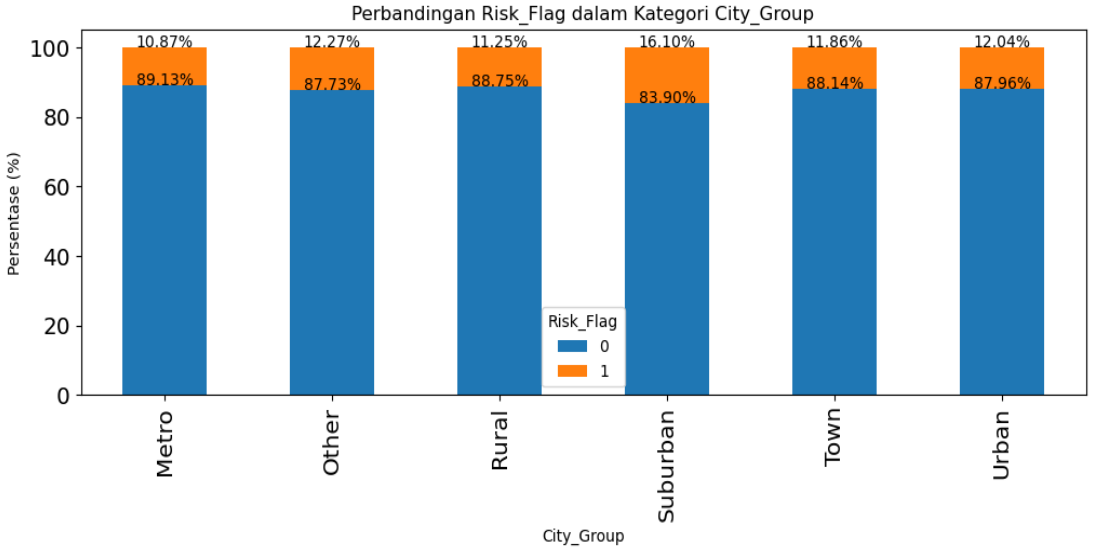



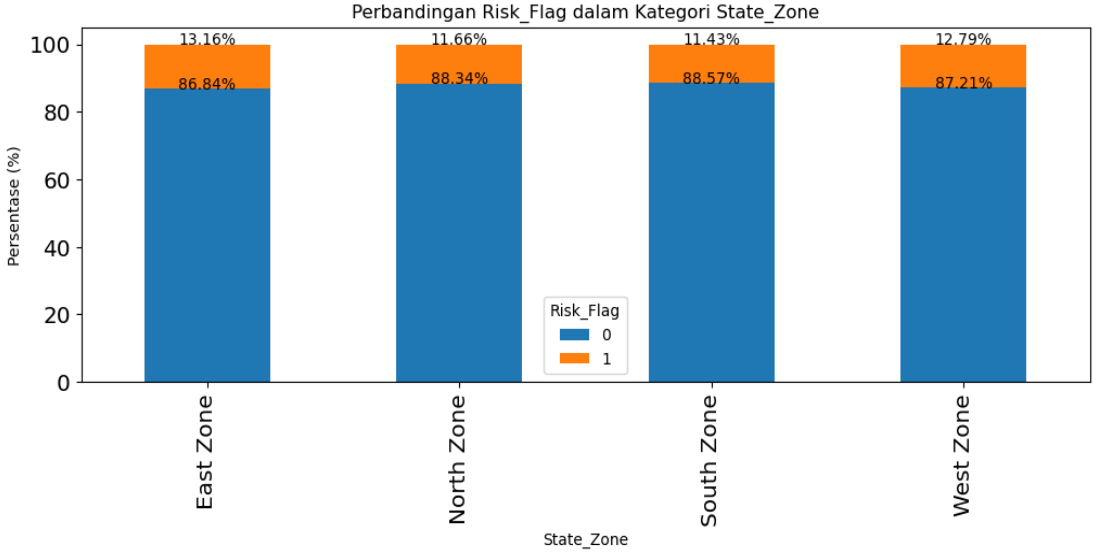

Perbandingan Risk_Flag dalam Kategori Married/Single
Persentase (%)
100
80
60
40
20
0
10.25%
89.75%
12.53%
87.47%
Risk_Flag
0
1
married
single
Married/Single

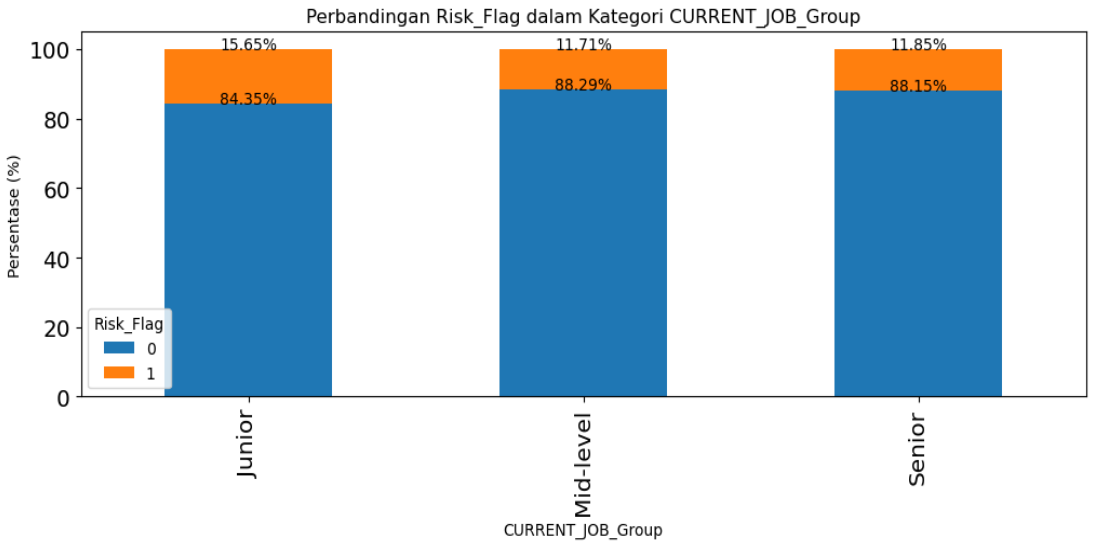

Perbandingan Risk_Flag dalam Kategori House_Ownership
Persentase (%)
100
80
60
40
20
0
9.95%
90.05%
8.98%
91.02%
12.56%
87.44%
Risk_Flag
0
1
norent_noown
owned
rented
House_Ownership

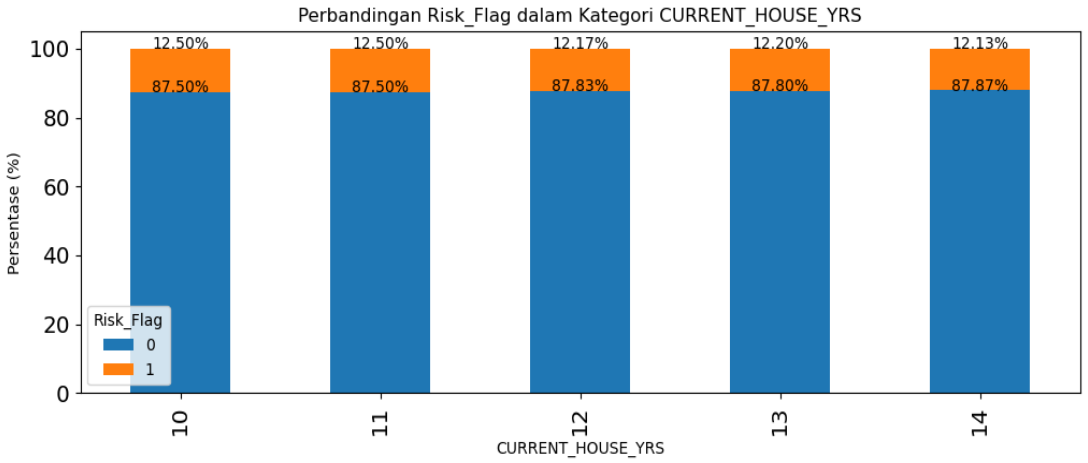

Perbandingan Risk_Flag dalam Kategori Car_Ownership
Persentase (%)
100
80
60
40
20
0
12.82%
87.18%
11.10%
88.90%
Risk_Flag
0
1
no
yes
Car_Ownership


Perbandingan Risk_Flag dalam Kategori Profession_Group
Persentase (%)
11.41%
88.59%
11.68%
88.32%
12.67%
87.33%
12.80%
87.20%
Risk_Flag
0
1
Engineering & IT
Healthcare
Management & Law
Other
Profession_Group


Perbandingan Risk_Flag dalam Kategori City_Group
Persentase (%)
10.87%
89.13%
12.27%
87.73%
11.25%
88.75%
16.10%
83.90%
11.86%
88.14%
12.04%
87.96%
Risk_Flag
0
1
Metro
Other
Rural
Suburban
Town
Urban
City_Group


Perbandingan Risk_Flag dalam Kategori State_Zone
Persentase (%)
13.16%
86.84%
11.66%
88.34%
11.43%
88.57%
12.79%
87.21%
Risk_Flag
0
1
East Zone
North Zone
South Zone
West Zone
State_Zone




## 4. Business Insight

Dalam penentuan pemberian kredit kepada nasabah, dapat menerapkan prinsip analisa 5C yang meliputi Character (Watak), Capacity (Kemampuan), Capital (Modal), Condition (Kondisi), dan Collateral (Jaminan).

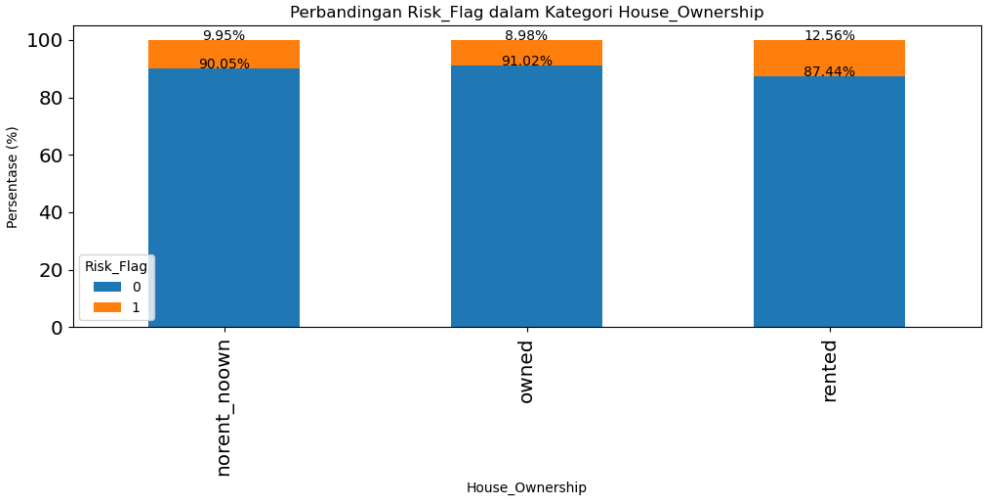
Dalam kemampuan nasabah, nasabah dengan income low lebih berpotensi gagal bayar pinjaman dengan persentase 15%, dibanding dengan nasabah yang memiliki income medium 14%, dan income high 12%. Selain itu, pada feature Age tidak ada perbedaan yang signifikan antara 3 generasi yaitu Gen X, Millennials, dan Baby Boomers dengan risiko gagal bayar pinjaman. Dan untuk feature Experience juga tidak ada perbedaan yang signifikan, namun semakin lama pengalaman nasabah maka semakin kecil nasabah berpotensi gagal bayar pinjaman. Begitu juga dengan feature Current_Job_Yrs, tidak ada perbedaan yang signifikan, namun nasabah dengan current job rendah atau junior maka potensi gagal bayar pinjaman lebih tinggi daripada mid-level dan senior.

1. Capacity

   Nasabah dengan penghasilan income Low cenderung lebih tinggi dalam tingkat risiko kredit dengan persentase 15.02% dari total nasabah dengan kategori penghasilan yang sama

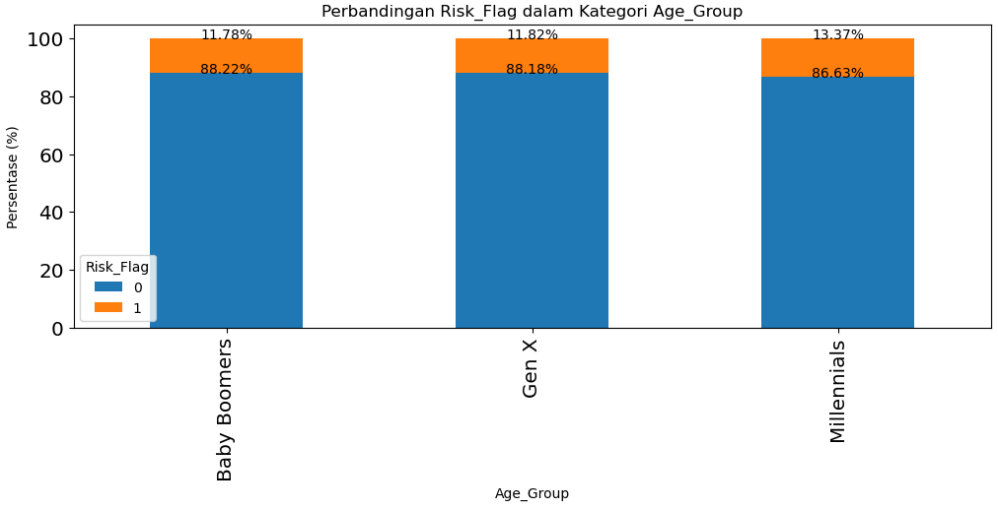
2. Capital

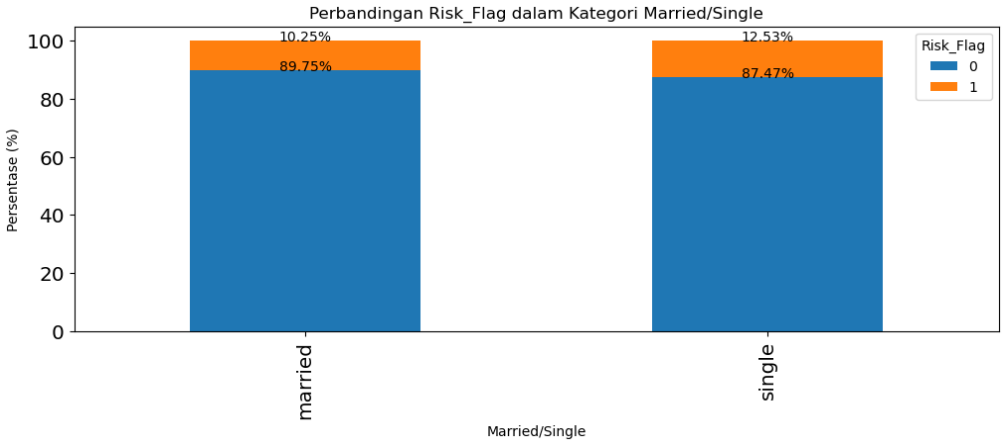
   Dari data didapatkan bahwa nasabah dengan modal atau kepemilikan harta tidak memiliki mobil dan menyewa rumah memiliki angka risiko yang lebih tinggi dengan persentase masing masing yaitu 11.10% dan 12.56% di setiap kategori kelasnya. Dari data data dapat dirumuskan hipotesis dengan pemberian kredit lebih baik diberikan kepada nasabah yang memiliki harta tetap dibandingkan dengan kepemilikan sementara.



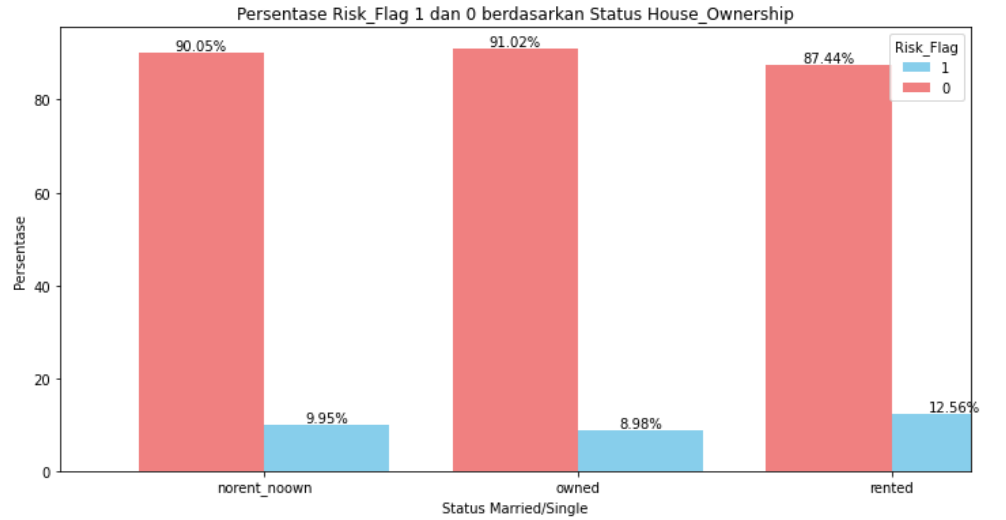

3. Condition

   Nasabah dengan status single dan rentang usia Millennials cenderung memiliki tingkat resiko cukup tinggi dalam gagal bayar pinjaman dengan persentase daripada nasabah yang sudah menikah dan/atau nasabah berusia selain millennials dengan masing-masing persentase 12.53% dan 13.37%  di setiap kategori kelasnya.






4. Collateral




Nasabah yang menyewa rumah cenderung lebih tinggi untuk tingkat risiko gagal bayar pinjaman dengan persentase 12.56% dari pada nasabah yang memiliki rumah pribadi 9% dan juga nasabah yang belum mempunyai rumah pribadi 9.95%. Sehingga dalam hal ini, lembaga pemberi pinjaman tidak disarankan untuk memberikan pinjaman terutama dalam jumlah banyak kepada nasabah dengan kepemilikan rumah menyewa dan tanpa kepemilikan rumah.

**5. Git (15 poin) - ALYA**

Link Repository:
https://github.com/AlyaniNS/Loan-Prediction-Based-on-Customer-Behavior