

# Big Data Maestros

Ketua : Ade Khamelia Putra

Anggota:

1. Cahya Chrishariyani
2. Faridatul Husna
3. Fiqrah Maulani
4. Indra Bayu Permana
5. Muhammad Haidar Alessandro Abror

## Final Project - Stage 1



# Stage 1

## 1. Deskriptif Statistik

Berdasarkan hasil observasi yang telah tim kami lakukan terhadap dataset "Loan Predicton based on customer behavior" didapatkan bahwa:

- Tipe data untuk setiap fitur sudah sesuai

```
[ ] df.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
```

# Stage 1

## 1. Deskriptif Statistik

Berdasarkan hasil observasi yang telah tim kami lakukan terhadap dataset "Loan Predicton based on customer behavior" didapatkan bahwa:

- Dataset memiliki jumlah baris sebanyak 252000 dan jumlah fitur ada sebanyak 13.
- Tidak terdapat nilai null dalam dataset
- Tidak terdapat data yang duplikat

```
[ ] # Mengetahui jumlah kolom dan baris
    jumlah_baris, jumlah_kolom = df.shape

# Menampilkan hasil
print("Jumlah Baris:", jumlah_baris)
print("Jumlah Kolom:", jumlah_kolom)

Jumlah Baris: 252000
Jumlah Kolom: 13
```

```
[ ] df.isnull().sum()
```

Id	0
Income	0
Age	0
Experience	0
Married/Single	0
House_Ownership	0
Car_Ownership	0
Profession	0
CITY	0
STATE	0
CURRENT_JOB_YRS	0
CURRENT_HOUSE_YRS	0
Risk_Flag	0
dtype: int64	

```
[ ] df.duplicated().sum()
```

0



# Stage 1

- Nama kolom karena tidak konsisten terkait penulisan, maka kami sesuaikan dengan menggunakan huruf kecil untuk semua nama kolom serta untuk nama kolom married/single kami ganti menjadi marital\_status untuk memudahkan dalam processing data.
- Merubah fitur ID menjadi index

```
# Mengganti nama kolom
df.rename(columns={'Income': 'income', 'Age': 'age', 'Experience': 'experience', 'Married/Single': 'marital_status', 'House_Ownership': 'house_ownership', 'Car_Ownership': 'car_ownership'})

# Menampilkan DataFrame setelah perubahan
print("\nDataFrame Setelah Perubahan Nama Kolom:")
df.info()
```

```
DataFrame Setelah Perubahan Nama Kolom:
<class 'pandas.core.frame.DataFrame'>
Int64Index: 252000 entries, 1 to 252000
Data columns (total 12 columns):
#   Column                Non-Null Count  Dtype
---  -
0   income                252000 non-null  int64
1   age                  252000 non-null  int64
2   experience            252000 non-null  int64
3   marital_status        252000 non-null  object
4   house_ownership       252000 non-null  object
5   car_ownership         252000 non-null  object
6   profession            252000 non-null  object
7   city                  252000 non-null  object
8   state                 252000 non-null  object
9   current_job_yrs       252000 non-null  int64
10  current_house_yrs     252000 non-null  int64
11  risk_flag             252000 non-null  int64
dtypes: int64(6), object(6)
memory usage: 25.0+ MB
```

# Stage 1

- Pada "city" dan "state" kami melakukan cleaning terkait penulisan nama kota yang sebelumnya menggunakan simbol dan angka

```
# Menghilangkan angka dan tanda kurung siku dari data di kolom "city"
df.city = df['city'].str.replace(r'\\[\\d+\\]', '')
df.city = df.city.str.replace('_', ' ')

# Menampilkan DataFrame setelah perubahan
print(df.city)
```

```
Id
1      Rewa
2      Parbhani
3      Alappuzha
4      Bhubaneswar
5      Tiruchirappalli
...
251996      Kolkata
251997      Rewa
251998      Kalyan-Dombivli
251999      Pondicherry
252000      Avadi
Name: city, Length: 252000, dtype: object
```

```
[ ] # Menghilangkan angka dan tanda kurung siku dari data di kolom "state"
df.state = df['state'].str.replace(r'\\[\\d+\\]', '')
df.state = df.state.str.replace('_', ' ')

# Menampilkan DataFrame setelah perubahan
print(df.state)
```

```
Id
1      Madhya Pradesh
2      Maharashtra
3      Kerala
4      Odisha
5      Tamil Nadu
...
251996      West Bengal
251997      Madhya Pradesh
251998      Maharashtra
251999      Puducherry
252000      Tamil Nadu
Name: state, Length: 252000, dtype: object
```

# Stage 1

- Tidak ada nilai summary yang aneh pada dataset. Dimana mean dan median tidak terdapat perbedaan hal tersebut dapat mengindikasikan bahwa distribusi data cenderung simetris. Tetapi perlu dilakukan analisis lebih lanjut terkait distribusi data.

```
[ ] # melihat data summary  
df[kontinu].describe()
```

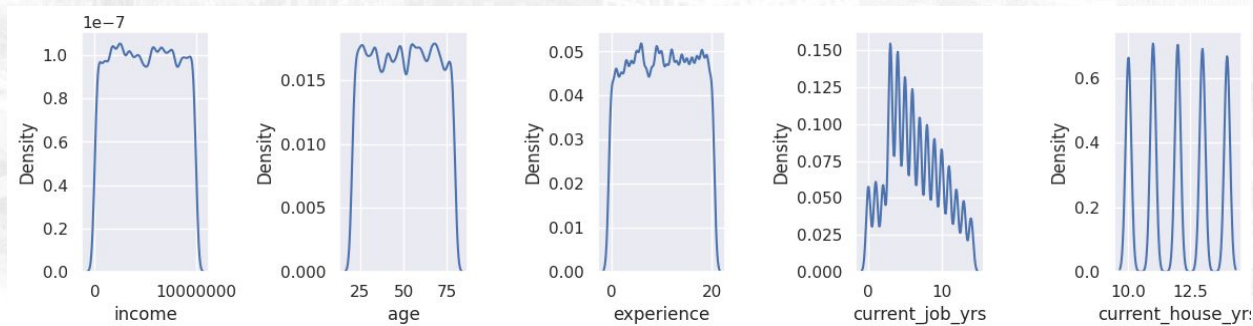
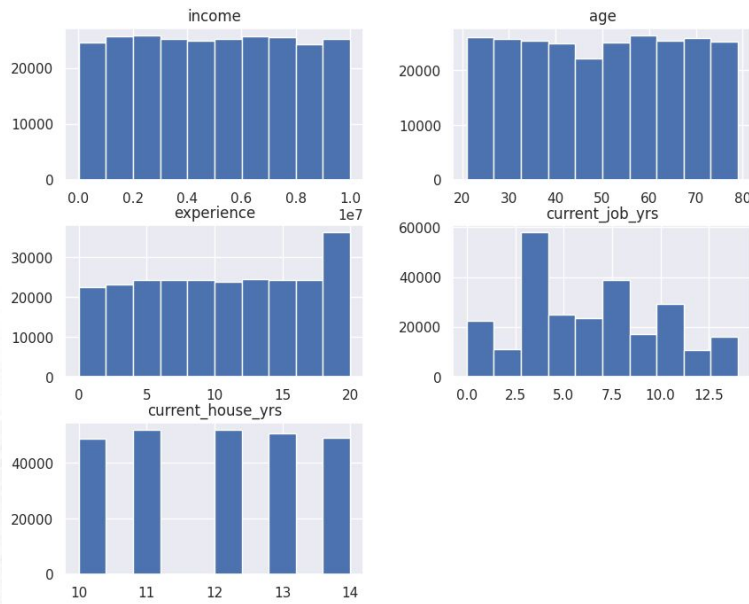
	income	age	experience	current_job_yrs	current_house_yrs
count	2.520000e+05	252000.000000	252000.000000	252000.000000	252000.000000
mean	4.997117e+06	49.954071	10.084437	6.333877	11.997794
std	2.878311e+06	17.063855	6.002590	3.647053	1.399037
min	1.031000e+04	21.000000	0.000000	0.000000	10.000000
25%	2.503015e+06	35.000000	5.000000	3.000000	11.000000
50%	5.000694e+06	50.000000	10.000000	6.000000	12.000000
75%	7.477502e+06	65.000000	15.000000	9.000000	13.000000
max	9.999938e+06	79.000000	20.000000	14.000000	14.000000

```
[ ] # melihat data summary  
df[diskrit].describe()
```

	marital_status	house_ownership	car_ownership	profession	city	state	risk_flag
count	252000	252000	252000	252000	252000	252000	252000
unique	2	3	2	51	316	28	2
top	single	rented	no	Physician	Aurangabad	Uttar Pradesh	0
freq	226272	231898	176000	5957	1543	29143	221004

# Stage 1

## 2. Univariate Analysis





# Stage 1

## 2. Univariate Analysis

### A. Distribusi Data

Berdasarkan histogram dan kdeplot terlihat bahwa:

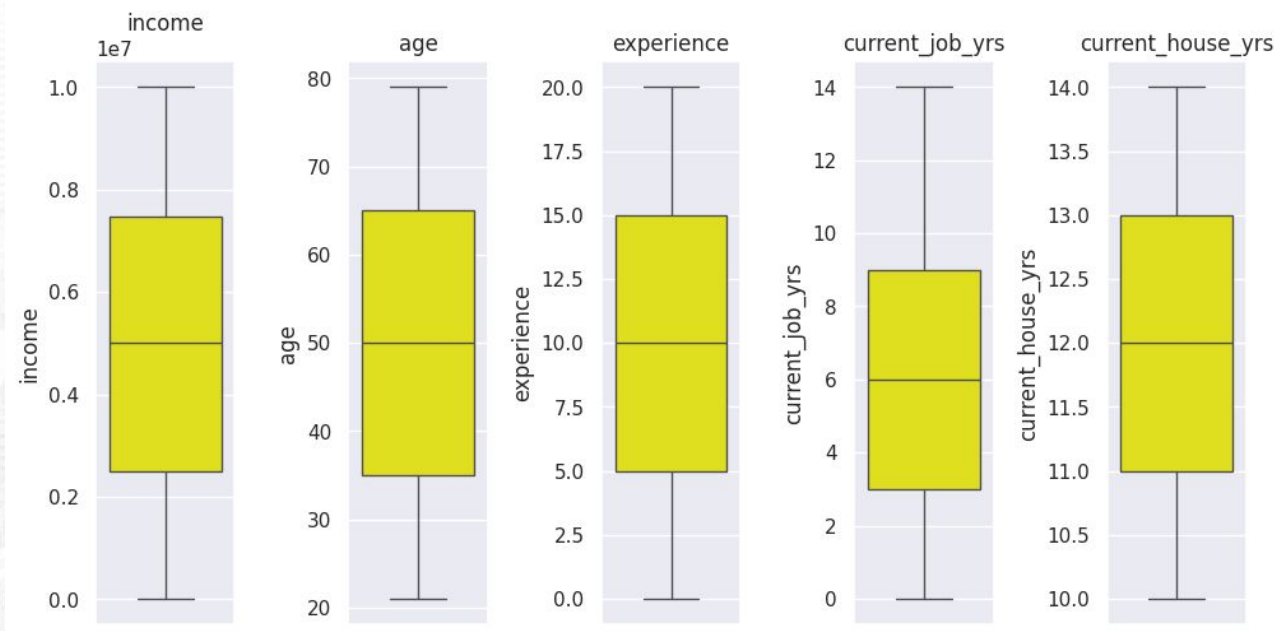
- Penyebaran distribusi untuk masing masing fitur cenderung rata kecuali pada fitur "current\_job\_yrs".
- Skewness multimodal : income, age, experience, current\_house\_yrs
- Skew positif (cenderung ke kanan) : current\_job\_yrs



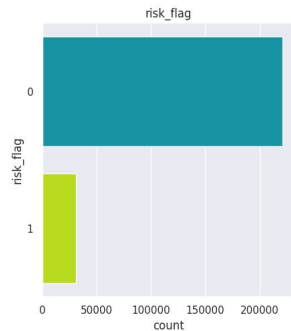
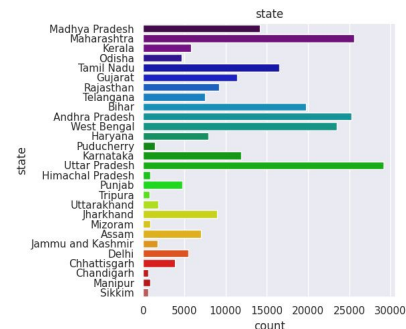
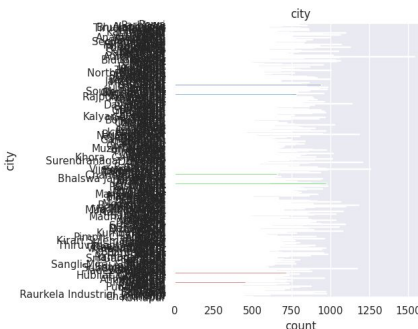
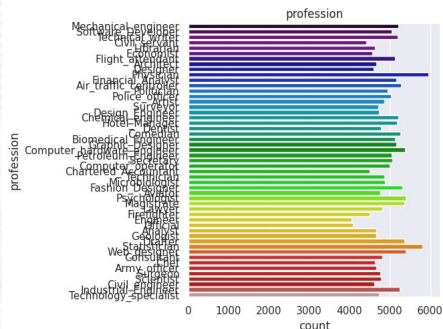
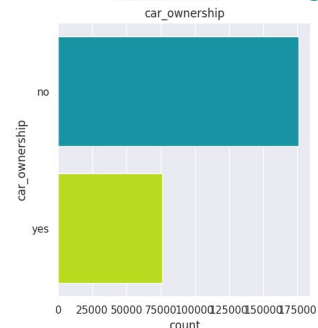
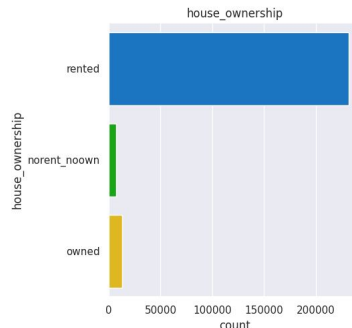
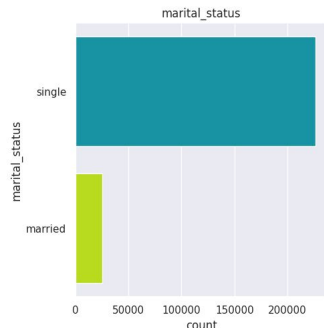
# Stage 1

## B. Outlier

Terlihat dari boxplot dibawah, tidak terdapat outlier pada dataset numerical dan dapat disimpulkan juga bahwa data cenderung berpusat pada nilai tengah.



# Distribusi Data Diskrit

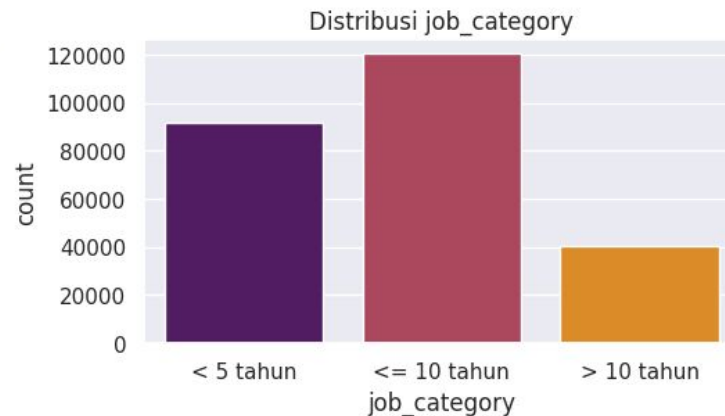
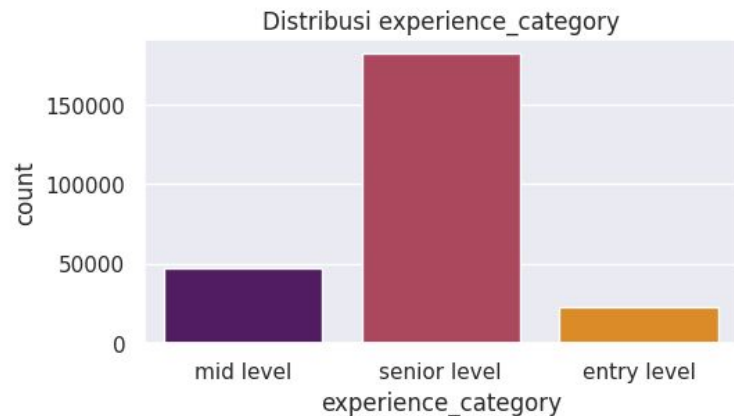
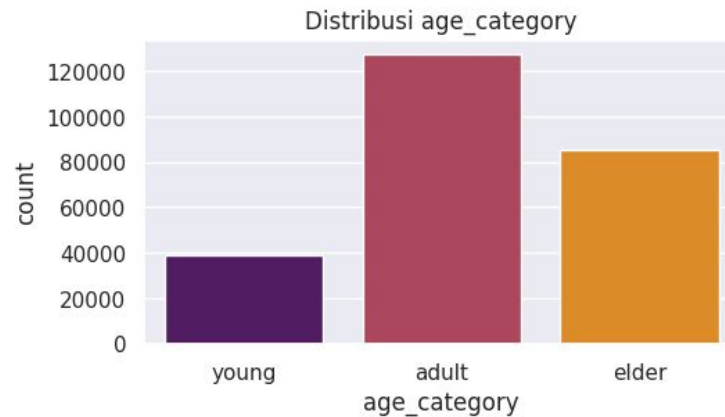
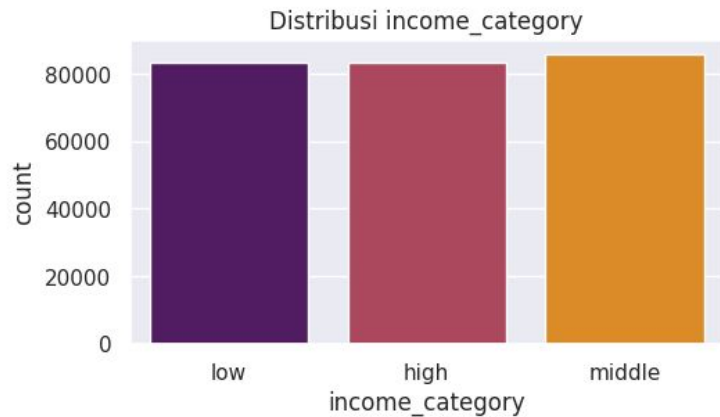


## Distribusi Data Diskrit

- Kolom-kolom seperti city, state, dan profession memiliki data yang sangat banyak, sehingga diperlukan pengelompokan data untuk menggabungkan kategori serupa menjadi satu kategori baru. Tujuannya adalah untuk menyederhanakan data.
- Ditemukan ketidakseimbangan dominasi kategori pada kolom marital\_status, house\_ownership, dan risk\_flag yang dapat memengaruhi hasil analisis. Oleh karena itu, perlu dipertimbangkan apakah tindakan seperti oversampling atau undersampling diperlukan untuk menangani ketidakseimbangan tersebut.
- Kami juga melakukan pengelompokan data untuk fitur income, age, experience, dan current job years untuk memudahkan proses analisis data.

# Stage 1

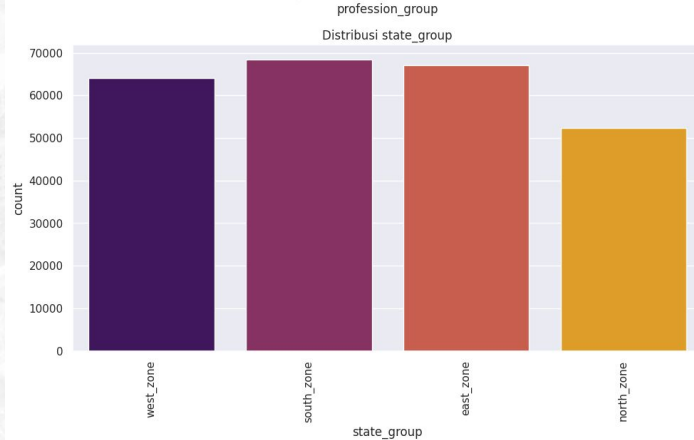
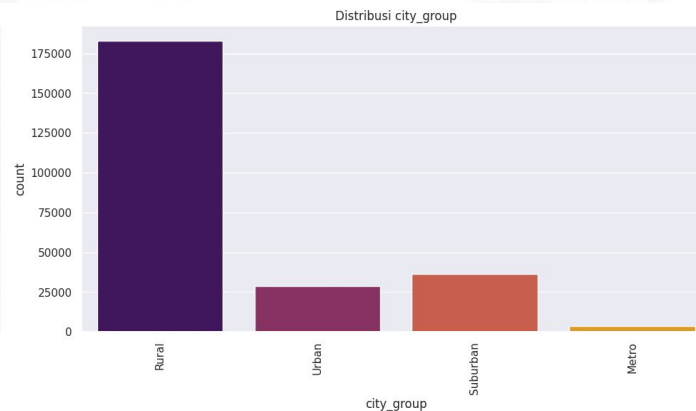
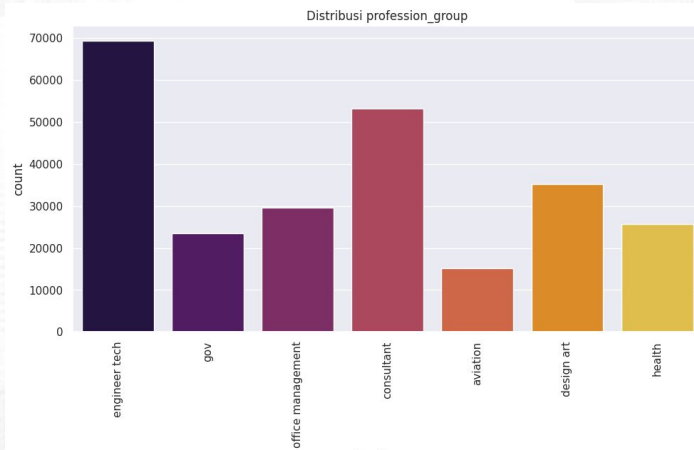
## Distribusi Data yang telah di kelompokkan





# Stage 1

## Distribusi Data yang telah di kelompokkan



# Stage 1

## Distribusi Data yang telah di kelompokkan

- Dari hasil distribusi di atas, dapat diketahui bahwa pada kolom `income_category` nilai yang paling tinggi adalah middle, namun secara garis besar distribusi ketiga kategori income data merata.
- Pada kolom `age_category` nilai yang mendominasi adalah adult.
- Pada kolom `experience_category` nilai yang mendominasi adalah senior level.
- Pada kolom `job_category` nilai yang mendominasi adalah  $\leq 10$  tahun.
- Pada kolom `profession_group` nilai yang mendominasi adalah engineer tech.
- Pada kolom `city_group` nilai yang mendominasi adalah Rural.
- Pada kolom `state_group` nilai yang paling tinggi ada pada south\_zone.

# Stage 1

## 3. Multivariate Analysis

### a. Heatmap fitur Kontinue

Berdasarkan korelasi heatmap dibawah terlihat bahwa `current_job_years` dan `experience` mempunyai korelasi yang besar sehingga berpotensi terjadi multikolinearitas.



## 3. Multivariate Analysis

### *Hasil Observasi Terkait Korelasi antar Fitur dan Label*

- Marital Status: Nilai p-value yang sangat rendah ( $3.77e-26$ ) menunjukkan bahwa terdapat hubungan yang signifikan antara status pernikahan dan risiko yang diidentifikasi (risk\_flag).
- House Ownership: Nilai p-value yang sangat rendah ( $1.84e-40$ ) menunjukkan bahwa terdapat hubungan yang signifikan antara kepemilikan rumah dan risiko yang diidentifikasi (risk\_flag).
- Car Ownership: Nilai p-value yang sangat rendah ( $1.74e-33$ ) menunjukkan bahwa terdapat hubungan yang signifikan antara kepemilikan mobil dan risiko yang diidentifikasi (risk\_flag).
- Profession: Nilai p-value yang sangat rendah ( $5.11e-98$ ) menunjukkan bahwa terdapat hubungan yang signifikan antara profesi dan risiko yang diidentifikasi (risk\_flag).
- City: Nilai p-value yang sangat rendah (0.0) menunjukkan bahwa terdapat hubungan yang signifikan antara kota dan risiko yang diidentifikasi (risk\_flag).
- State: Nilai p-value yang sangat rendah ( $6.50e-137$ ) menunjukkan bahwa terdapat hubungan yang signifikan antara negara bagian dan risiko yang diidentifikasi (risk\_flag).

Dari hasil ini, kita dapat menyimpulkan bahwa semua fitur kategorikal memiliki hubungan yang signifikan dengan label risk\_flag, seperti yang diindikasikan oleh nilai p-value yang sangat rendah pada uji chi-square. Oleh karena itu, fitur-fitur ini mungkin memiliki kontribusi yang signifikan dalam memprediksi risiko.

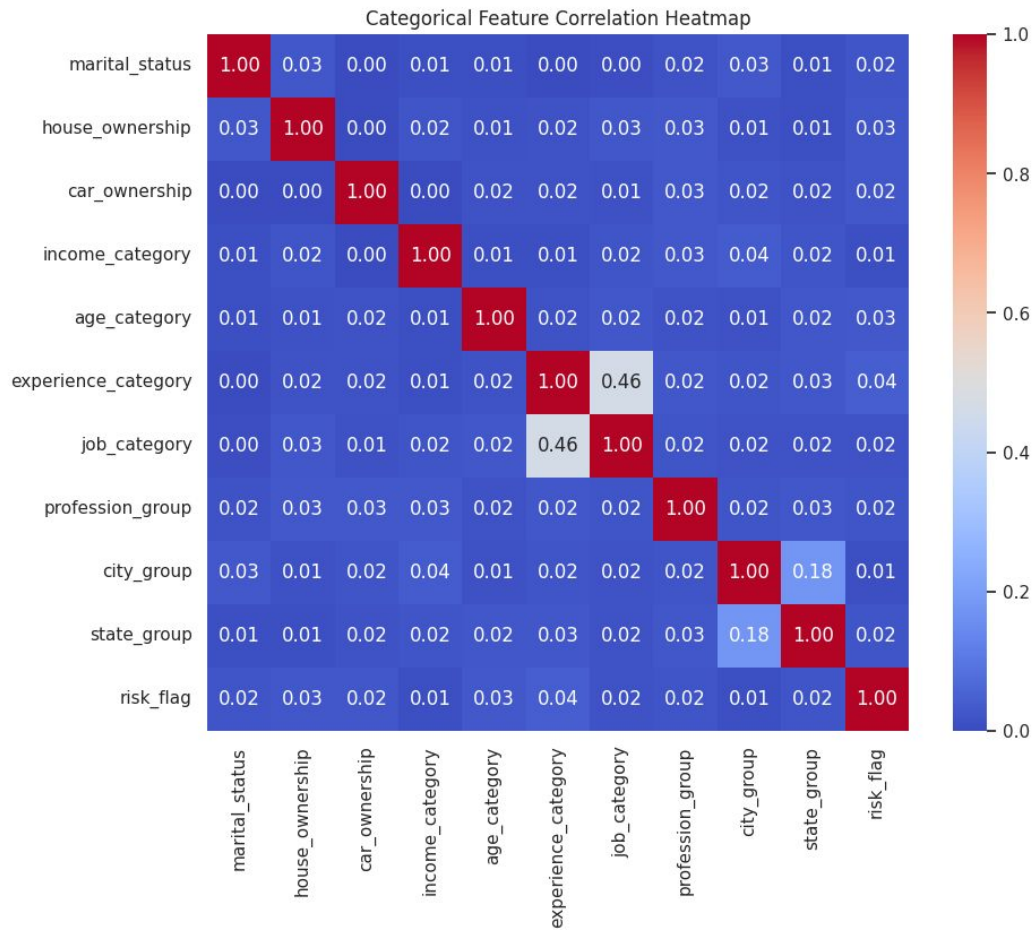


# Stage 1

## 3. Multivariate Analysis

### a. Heatmap fitur Diskrit

Terlihat dari korelasi heatmap disamping tidak ada data yang mempunyai korelasi yang signifikan. Hal tersebut mengindikasikan bahwa fitur-fitur tersebut bersifat independen satu sama lain

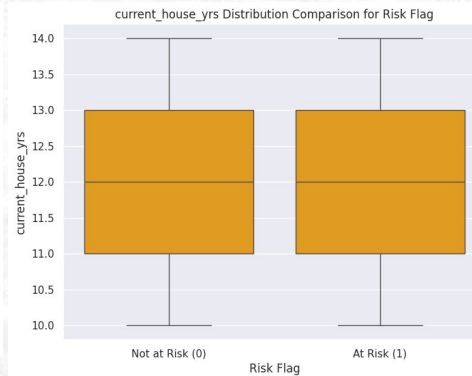
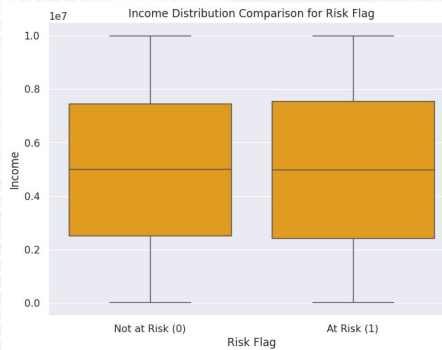
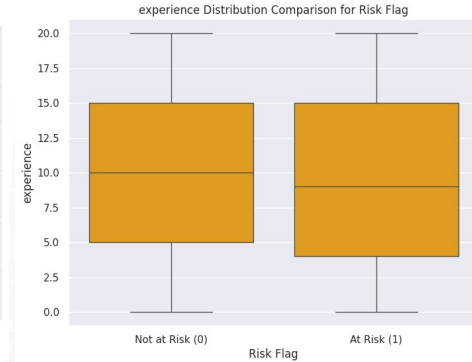
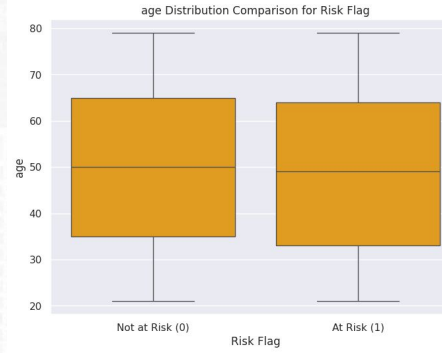


# Stage 1

## 3. Multivariate Analysis

### C. Distribusi fitur kontinu berdasarkan risk flag

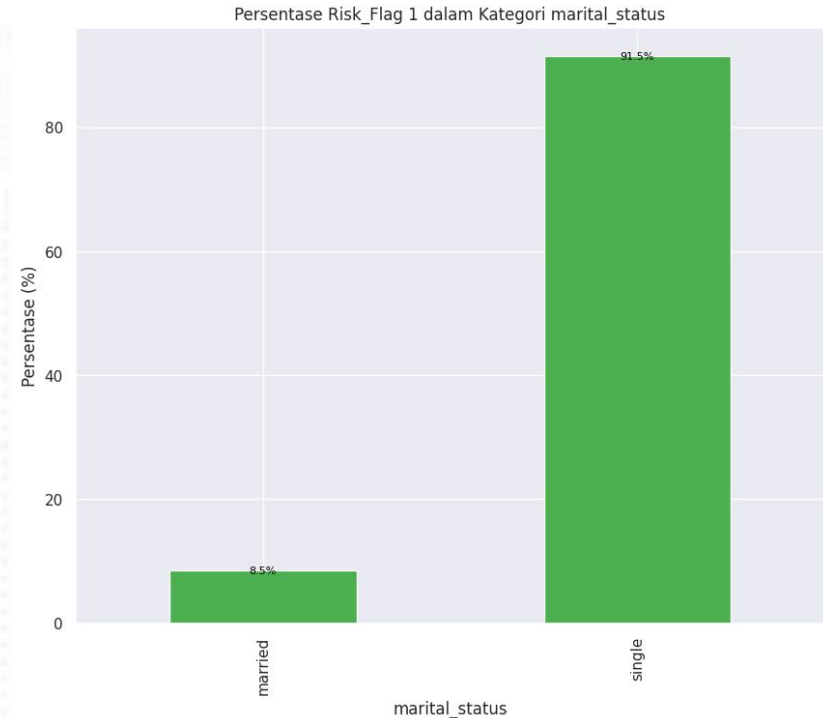
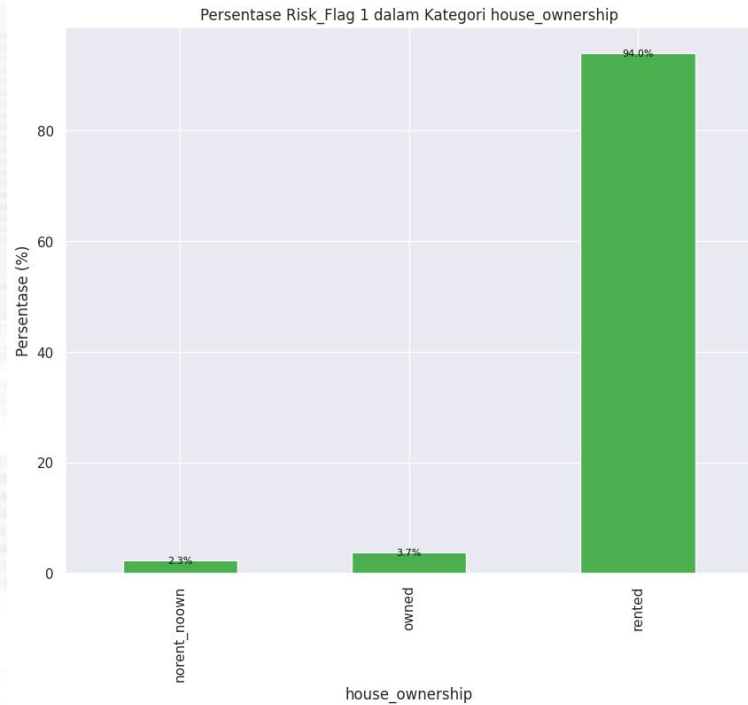
Dari distribusi data kontinu disamping berdasarkan risk flag-nya terlihat bahwa tidak terdapat perbedaan distribusi data dari masing-masing fitur terhadap risk flagnya



### 3. Multivariate Analysis

#### D. Perbandingan tingkat persentase risk flag di fitur Diskrit

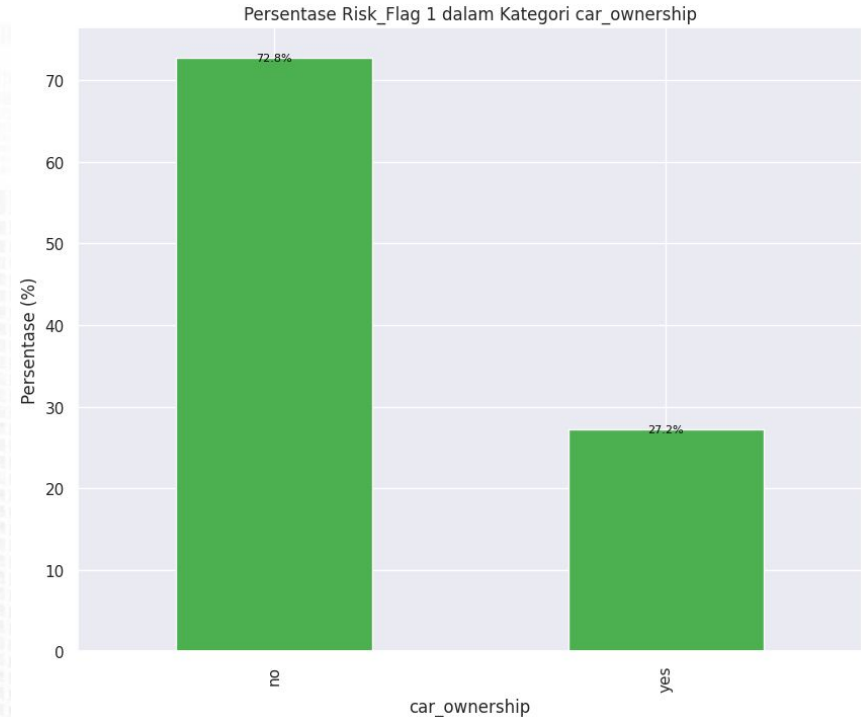
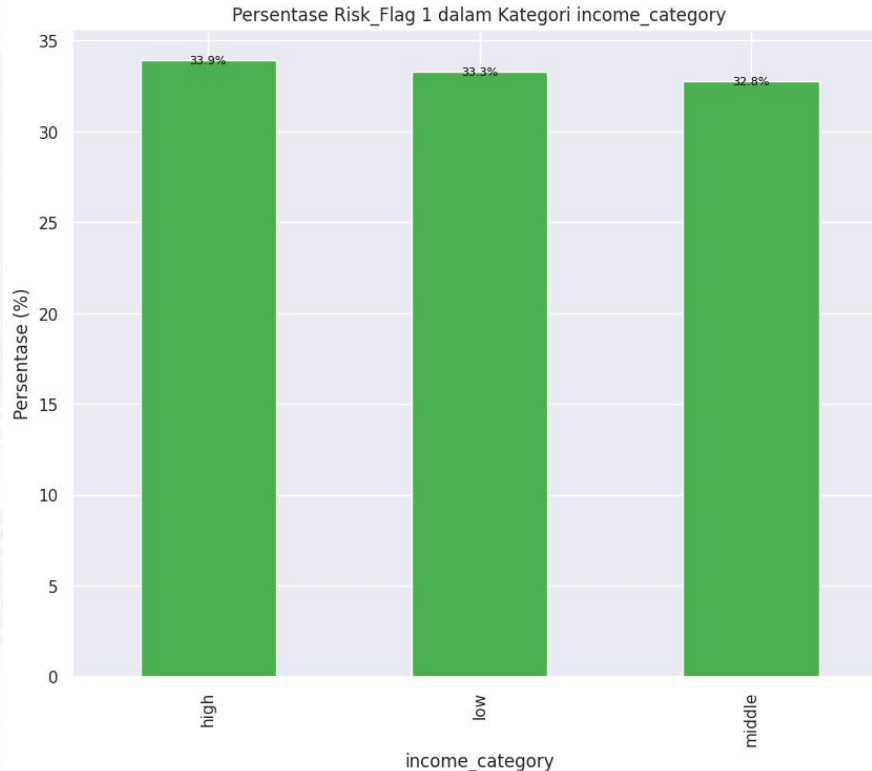
Dari perbandingan persentasi risk flag dari data house\_ownership dibawah dapat diketahui bahwa tingkat resiko dari customer yang menyewa rumah (rented) mempunyai resiko paling tinggi sebesar 94.0%



Dari perbandingan persentasi risk flag dari data marital\_status diatas dapat diketahui bahwa tingkat resiko dari customer yang single sebesar 91,5% mendominasi dibandingkan dengan yang sudah menikah

#### D. Perbandingan tingkat persentase risk flag...(lanjutan)

Dari perbandingan persentasi risk flag dari fitur income\_kategori dibawah dapat diketahui bahwa tingkat resiko dari customer dari setiap tingkat kategori income memiliki resiko yang seimbang diangka 32.8%-33.9%

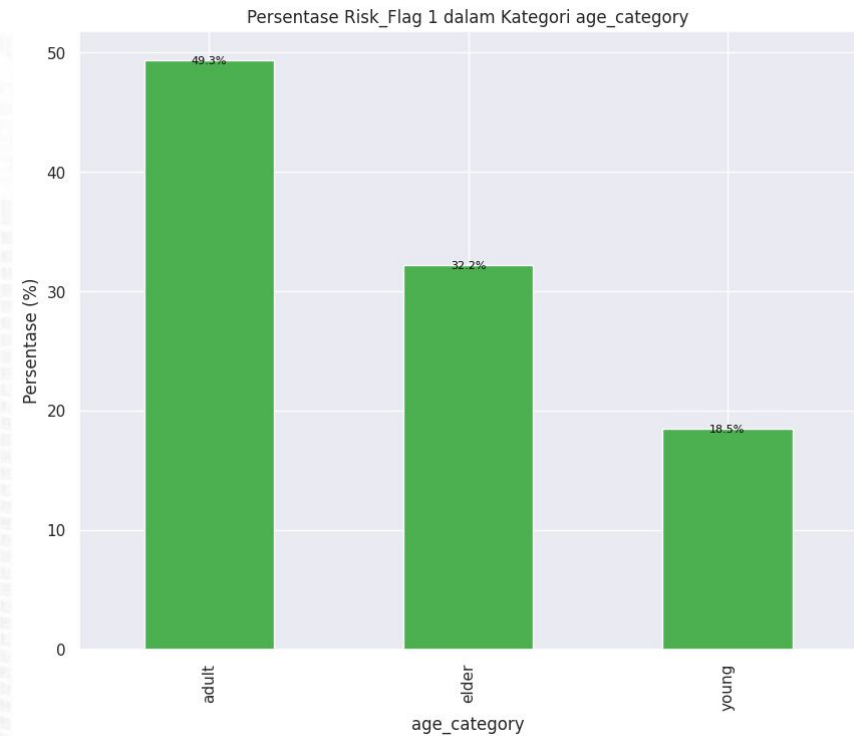
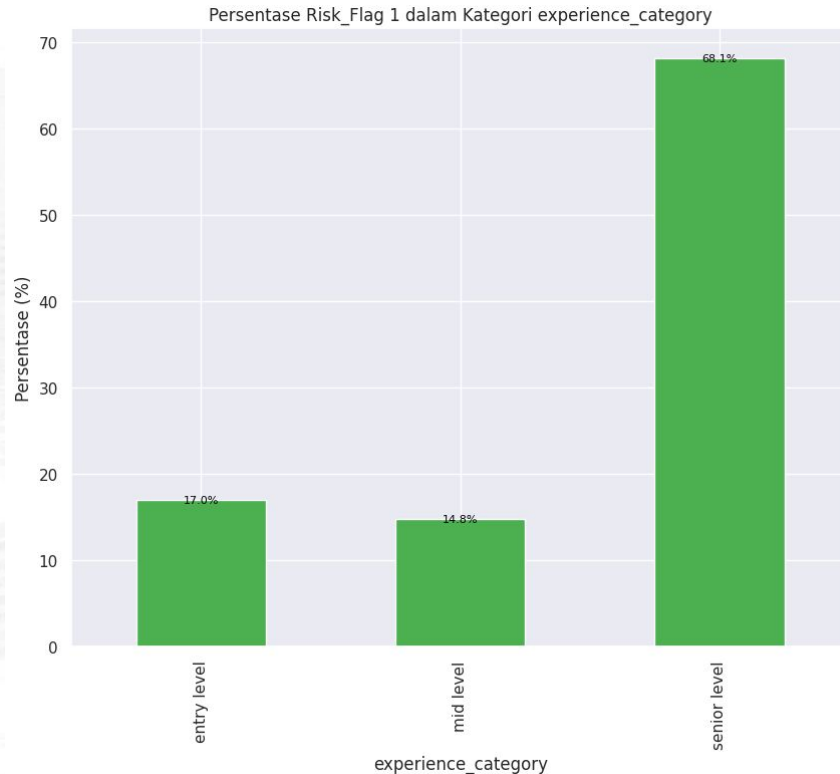


Dari perbandingan persentasi risk flag dari data car\_ownership diatas dapat diketahui bahwa tingkat resiko dari customer yang tidak memiliki mobil mempunyai tingkat resiko 72.8% dibandingkan yang mempunyai mobil dengan tingkat resiko 27.2%



#### D. Perbandingan tingkat persentase risk flag...(lanjutan)

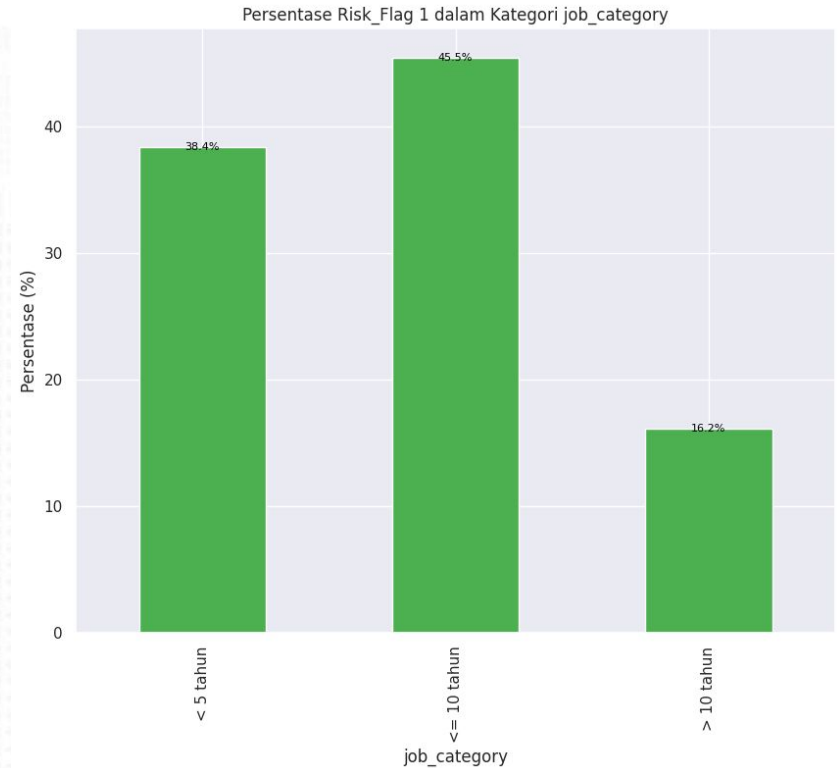
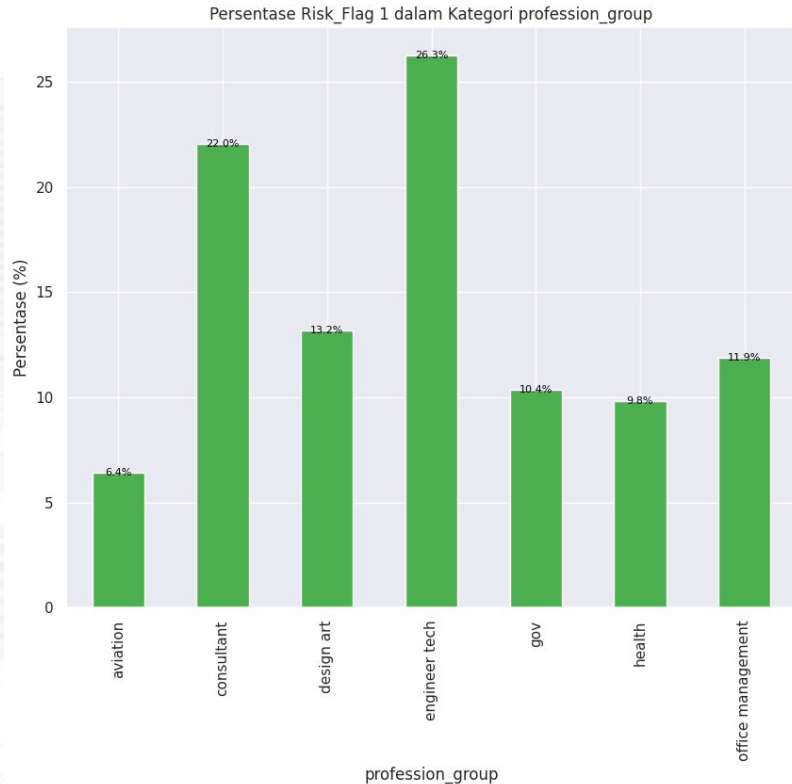
Dari perbandingan persentasi risk flag dari fitur experience\_kategori dibawah dapat diketahui bahwa tingkat resiko dari customer dari tingkat senior\_level mempunyai risk flag yang tinggi yaitu sebesar 68.1%



Dari perbandingan persentasi risk flag dari data age\_category diatas dapat diketahui bahwa tingkat resiko dari customer dengan kategori umur adult mempunyai tingkat resiko paling tinggi yaitu 49.3%

#### D. Perbandingan tingkat persentase risk flag...(lanjutan)

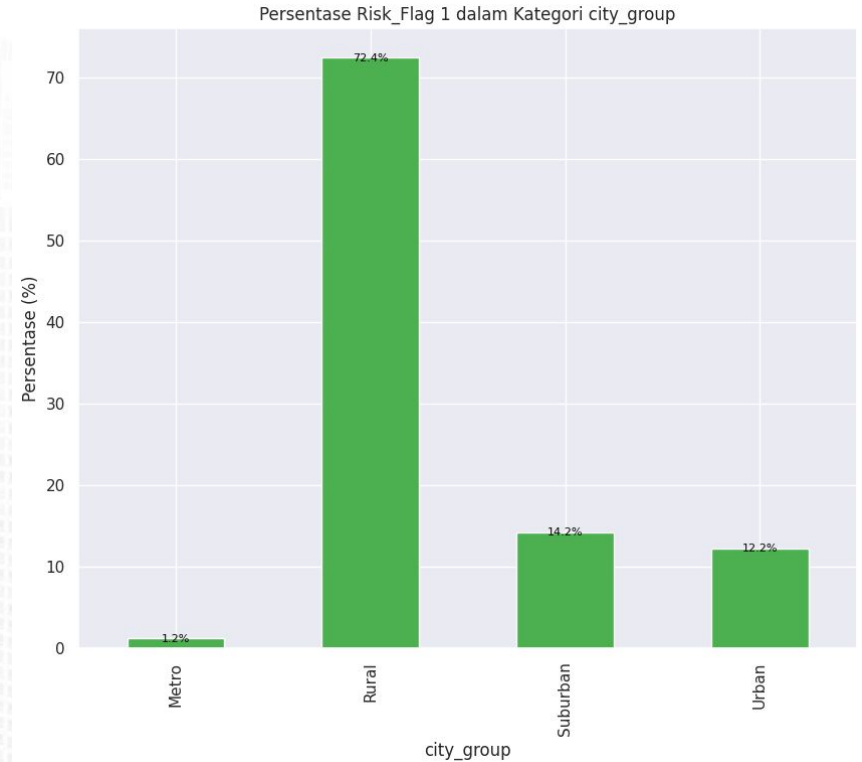
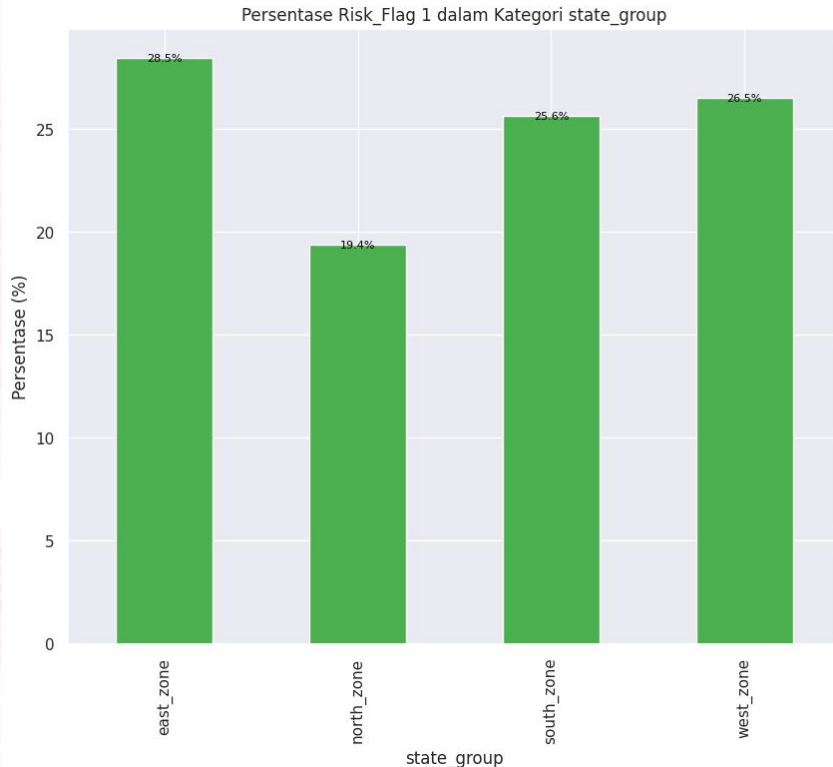
Dari perbandingan persentasi risk flag dari fitur profession\_group dibawah dapat diketahui bahwa tingkat resiko dari customer dari group engineer\_tech mempunyai risk flag yang paling tinggi yaitu sebesar 26.3%



Dari perbandingan persentasi risk flag dari data job\_category diatas dapat diketahui bahwa tingkat resiko dari customer dengan kategori lamanya bekerja di perusahaan sekarang 5-10 tahun mempunyai tingkat resiko paling tinggi yaitu 45.5%

#### D. Perbandingan tingkat persentase risk flag...(lanjutan)

Dari perbandingan persentasi risk flag dari fitur state\_group dibawah dapat diketahui bahwa tingkat resiko dari customer yang tinggal di east zone state mempunyai risk flag yang paling tinggi yaitu sebesar 28.5%



Dari perbandingan persentasi risk flag dari data city\_group diatas dapat diketahui bahwa tingkat resiko dari customer yang tinggal di rural Area mempunyai tingkat resiko paling tinggi yaitu 72.4%

## 3. Follow up data pre processing

- **Skewness pada Kolom Numerik**

Jika terdapat skewness, pertimbangkan untuk melakukan transformasi seperti log-transform.

- **Dominasi Nilai pada Kolom Kategorikal**

Jika terdapat kategori yang mendominasi, evaluasi apakah perlu menggabungkan beberapa kategori atau menerapkan strategi oversampling/undersampling.

- **Outlier pada Kolom Numerik**

Pertimbangkan untuk menangani outlier, misalnya dengan menghapusnya atau melakukan transformasi khusus.

- **Multikolinearitas**

Jika terdapat multicollinearity antar-feature, pertimbangkan untuk menggabungkan, menghapus, atau memilih subset feature agar model lebih stabil dan interpretatif.



# Stage 1

## 4. Business Insight

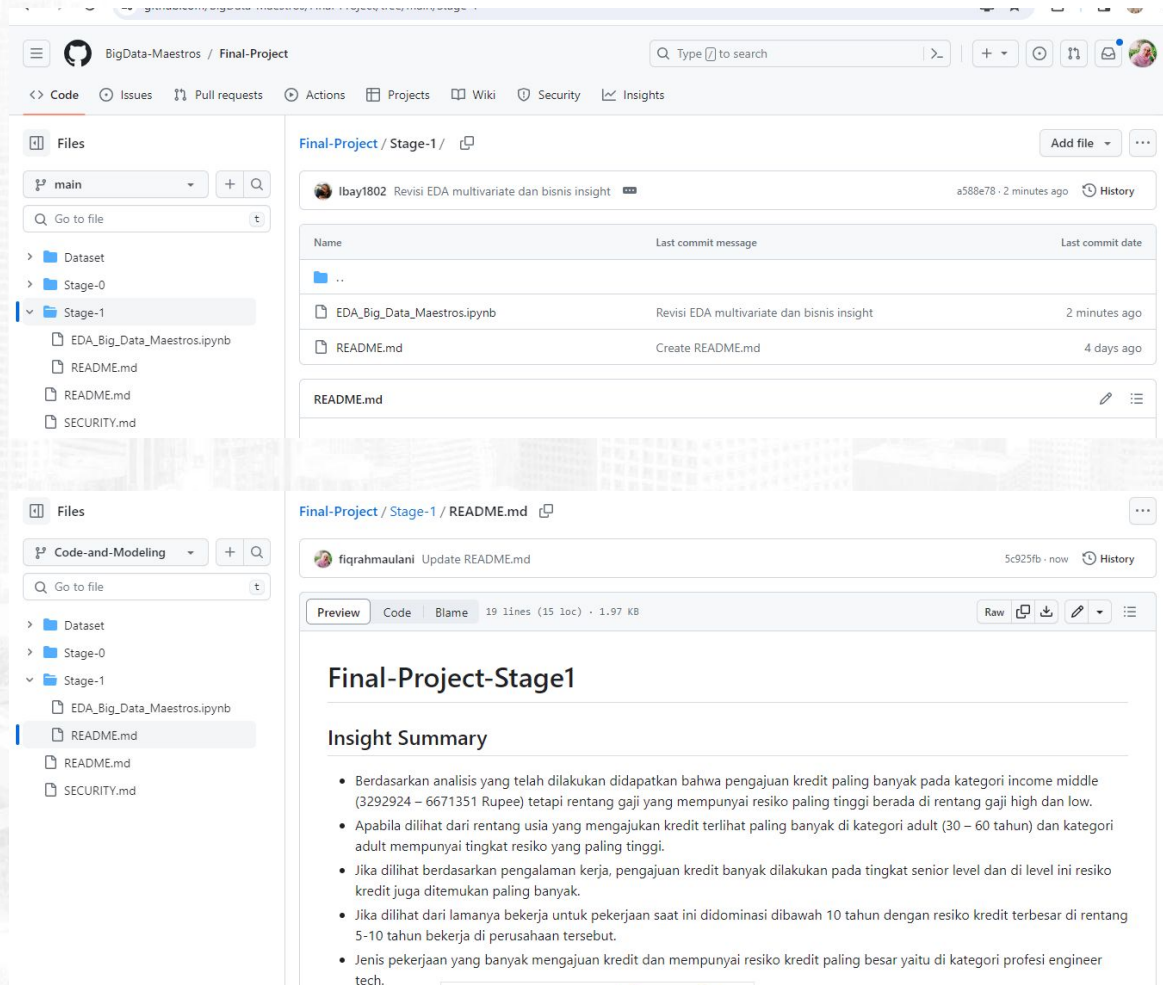
- Berdasarkan analisis yang telah dilakukan didapatkan bahwa pengajuan kredit paling banyak pada kategori income middle (3292924 – 6671351 Rupee) tetapi rentang gaji yang mempunyai resiko paling tinggi berada di rentang gaji high dan low.
- Apabila dilihat dari rentang usia yang mengajukan kredit terlihat paling banyak di kategori adult (30 – 60 tahun) dan kategori adult mempunyai tingkat resiko yang paling tinggi.
- Jika dilihat berdasarkan pengalaman kerja, pengajuan kredit banyak dilakukan pada tingkat senior level dan di level ini resiko kredit juga ditemukan paling banyak.
- Jika dilihat dari lamanya bekerja untuk pekerjaan saat ini didominasi dibawah 10 tahun dengan resiko kredit terbesar di rentang 5-10 tahun bekerja di perusahaan tersebut.
- Jenis pekerjaan yang banyak mengajukan kredit dan mempunyai resiko kredit paling besar yaitu di kategori profesi engineer tech.
- Apabila dilihat dari jenis kategori wilayah, pengajuan kredit banyak dilakukan pada wilayah Rural Area begitupun resiko kreditnya mempunyai resiko yang tinggi di wilayah ini.
- Apabila dilihat berdasarkan kategori zona state yang banyak melakukan pengajuan berada pada zona south\_zone tetapi tingkat resiko paling tinggi di temukan di wilayah east\_zone.

### 4. Business Recommendation

- Bank dapat melakukan peningkatan pemasaran pada customer dengan level income middle hingga high untuk semua rentang usia tetapi perlu di lihat juga profil dari nasabah tersebut sehingga dapat meminimalkan resiko.
- Dilihat berdasarkan pengalaman kerja customer, kategori yang memiliki kapabilabilitas kredit yang baik yaitu dengan pengalaman kerja diatas 5 tahun dengan tingkat lama bekerja untuk pekerjaan saat ini yaitu diatas 5 tahun tetapi perlu di lihat juga profil dari nasabah tersebut sehingga dapat meminimalkan resiko.
- Meningkatkan pemasaran pada kategori profesi yang masing rendah.
- Meningkatkan pemasaran untuk area urban hingga metropolitan dan untuk semua kategori zona state.

# Stage 1

## 5. Github



The screenshot displays the GitHub interface for the repository 'BigData-Maestros / Final-Project'. The left sidebar shows the file structure with folders 'Dataset', 'Stage-0', and 'Stage-1'. The 'Stage-1' folder is expanded, showing files 'EDA\_Big\_Data\_Maestros.ipynb', 'README.md', and 'SECURITY.md'. The main content area shows the 'Final-Project / Stage-1 / README.md' file. The commit history for this file is shown, with the latest commit by 'fiqrahmaulani' titled 'Update README.md' at '5c925fb · now'. The file content is displayed in a preview mode, showing the title 'Final-Project-Stage1' and a section 'Insight Summary'.

**Final-Project-Stage1**

### Insight Summary

- Berdasarkan analisis yang telah dilakukan didapatkan bahwa pengajuan kredit paling banyak pada kategori income middle (3292924 – 6671351 Rupee) tetapi rentang gaji yang mempunyai resiko paling tinggi berada di rentang gaji high dan low.
- Apabila dilihat dari rentang usia yang mengajukan kredit terlihat paling banyak di kategori adult (30 – 60 tahun) dan kategori adult mempunyai tingkat resiko yang paling tinggi.
- Jika dilihat berdasarkan pengalaman kerja, pengajuan kredit banyak dilakukan pada tingkat senior level dan di level ini resiko kredit juga ditemukan paling banyak.
- Jika dilihat dari lamanya bekerja untuk pekerjaan saat ini didominasi dibawah 10 tahun dengan resiko kredit terbesar di rentang 5-10 tahun bekerja di perusahaan tersebut.
- Jenis pekerjaan yang banyak mengajukan kredit dan mempunyai resiko kredit paling besar yaitu di kategori profesi engineer tech.