# Progress Report FinPro Stage 1


**Dino Kuning**

- Fany Okpiani
- Nadhilah Farhana
- Raditya Satria Gantara
- Rafindra Prihaztama

# Dino Kuning's Group Member

**Fany Okpiani**
*Business / Data Analyst*
Rakamin Academy

**Nadhilah Farhana**
*Data Scientist*
Rakamin Academy

**Rafindra Prihaztama**
*Data Engineer*
Rakamin Academy

**Raditya Satria G.**
*Project Manager*
Rakamin Academy

# Outline - Stage 1

# Data Train

Your selected dataframe has 122 columns.
There are 67 columns that have missing values.

| | Missing Values | Percent Missing (%) |
|---|---|---|
| COMMONAREA_MEDI | 214865 | 69.87 |
| COMMONAREA_MODE | 214865 | 69.87 |
| COMMONAREA_AVG | 214865 | 69.87 |
| NONLIVINGAPARTMENTS_MODE | 213514 | 69.43 |
| NONLIVINGAPARTMENTS_MEDI | 213514 | 69.43 |
| ... | ... | ... |
| EXT_SOURCE_2 | 660 | 0.21 |
| AMT_GOODS_PRICE | 278 | 0.09 |
| AMT_ANNUITY | 12 | 0.00 |
| CNT_FAM_MEMBERS | 2 | 0.00 |
| DAYS_LAST_PHONE_CHANGE | 1 | 0.00 |

67 rows × 2 columns

Terdapat 67 variable yang memiliki missing values, beberapa di antanya memiliki nilai persentase missing values >50% yang nantinya akan dihapus pada tahap data cleaning. Variable tersebut dihapus telah kehilangan banyak informasi.

# Data Quality Assesment

| | Jumlah Unique Kategori | Kategori |
|---|---|---|
| NAME_CONTRACT_TYPE | 2 | [Cash loans, Revolving loans] |
| CODE_GENDER | 3 | [M, F, XNA] |
| FLAG_OWN_CAR | 2 | [N, Y] |
| FLAG_OWN_REALTY | 2 | [Y, N] |
| NAME_TYPE_SUITE | 7 | [Unaccompanied, Family, Spouse, partner, Child... |
| NAME_INCOME_TYPE | 8 | [Working, State servant, Commercial associate,... |
| NAME_EDUCATION_TYPE | 5 | [Secondary / secondary special, Higher educati... |
| NAME_FAMILY_STATUS | 6 | [Single / not married, Married, Civil marriage... |
| NAME_HOUSING_TYPE | 6 | [House / apartment, Rented apartment, With par... |
| OCCUPATION_TYPE | 18 | [Laborers, Core staff, Accountants, Managers, ... |
| WEEKDAY_APPR_PROCESS_START | 7 | [WEDNESDAY, MONDAY, THURSDAY, SUNDAY, SATURDAY... |
| ORGANIZATION_TYPE | 58 | [Business Entity Type 3, School, Government, R... |
| FONDKAPREMONT_MODE | 4 | [reg oper account, nan, org spec account, reg ... |
| HOUSETYPE_MODE | 3 | [block of flats, nan, terraced house, specific... |
| WALLSMATERIAL_MODE | 7 | [Stone, brick, Block, nan, Panel, Mixed, Woode... |
| EMERGENCYSTATE_MODE | 2 | [No, nan, Yes] |

Selain pengecekan missing values data numerik, pada tahap ini juga dilakukan pengecekan data kategorik.

Terdapat penamaan yang tidak umum pada variable CODE_GENDER yang dapat diindikasi sebagai missing values.

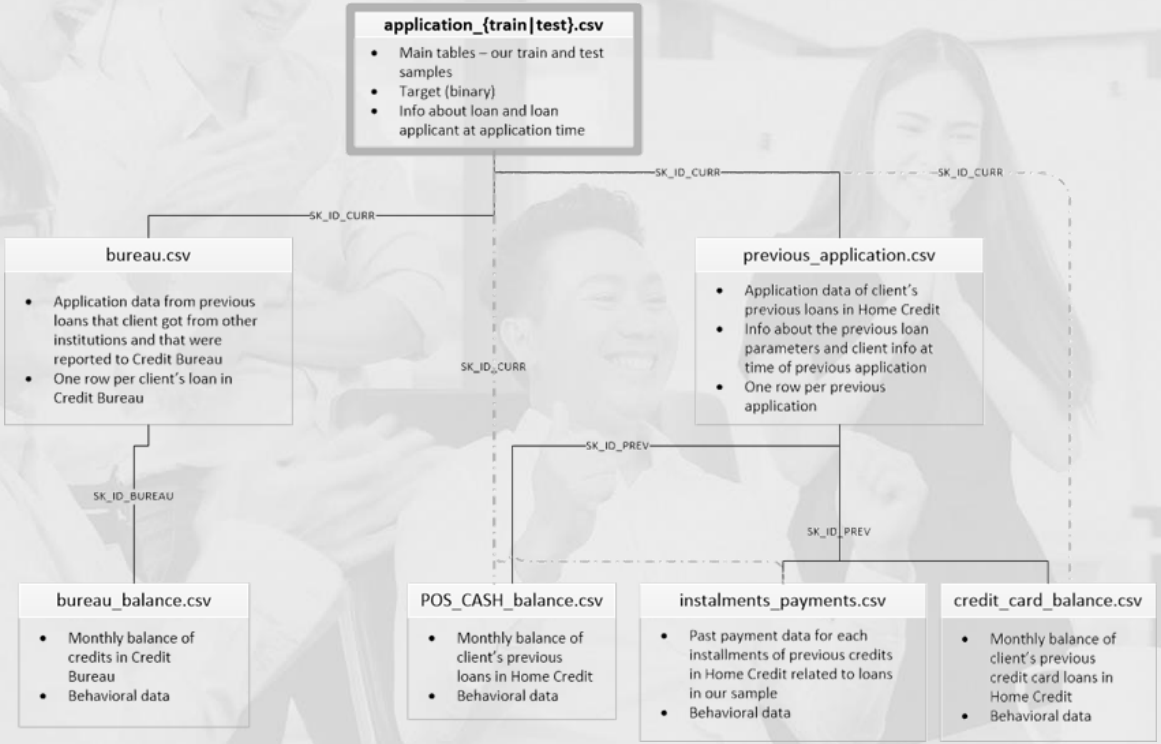
# Data Cleaning

| | Missing Values | Percent Missing (%) |
|---|---|---|
| FLOORSMAX_AVG | 153020 | 49.76 |
| FLOORSMAX_MODE | 153020 | 49.76 |
| FLOORSMAX_MEDI | 153020 | 49.76 |
| YEARS_BEGINEXPLUATATION_AVG | 150007 | 48.78 |
| YEARS_BEGINEXPLUATATION_MODE | 150007 | 48.78 |
| YEARS_BEGINEXPLUATATION_MEDI | 150007 | 48.78 |
| TOTALAREA_MODE | 148431 | 48.27 |
| EMERGENCYSTATE_MODE | 145755 | 47.40 |
| OCCUPATION_TYPE | 96391 | 31.35 |
| EXT_SOURCE_3 | 60965 | 19.83 |
| AMT_REQ_CREDIT_BUREAU_HOUR | 41519 | 13.50 |
| AMT_REQ_CREDIT_BUREAU_QRT | 41519 | 13.50 |
| AMT_REQ_CREDIT_BUREAU_MON | 41519 | 13.50 |
| AMT_REQ_CREDIT_BUREAU_WEEK | 41519 | 13.50 |
| AMT_REQ_CREDIT_BUREAU_DAY | 41519 | 13.50 |
| AMT_REQ_CREDIT_BUREAU_YEAR | 41519 | 13.50 |
| NAME_TYPE_SUITE | 1292 | 0.42 |
| DEF_30_CNT_SOCIAL_CIRCLE | 1021 | 0.33 |
| OBS_60_CNT_SOCIAL_CIRCLE | 1021 | 0.33 |
| DEF_60_CNT_SOCIAL_CIRCLE | 1021 | 0.33 |
| OBS_30_CNT_SOCIAL_CIRCLE | 1021 | 0.33 |
| EXT_SOURCE_2 | 660 | 0.21 |
| AMT_GOODS_PRICE | 278 | 0.09 |
| AMT_ANNUITY | 12 | 0.00 |
| CNT_FAM_MEMBERS | 2 | 0.00 |
| DAYS_LAST_PHONE_CHANGE | 1 | 0.00 |

Tahap Preprocessing:

- ❑ Variabel yang masih memiliki persentase missing values <50% akan ditangani pada tahap preprocessing dengan cara:
  - Numerik: Imputasi nilai median
  - Kategorik: Imputasi nilai modus
- ❑ Data kategorik yang memiliki nilai tidak umum (contoh: CODE_GENDER -> XNA) akan diubah menjadi NaN.

# Data Integration

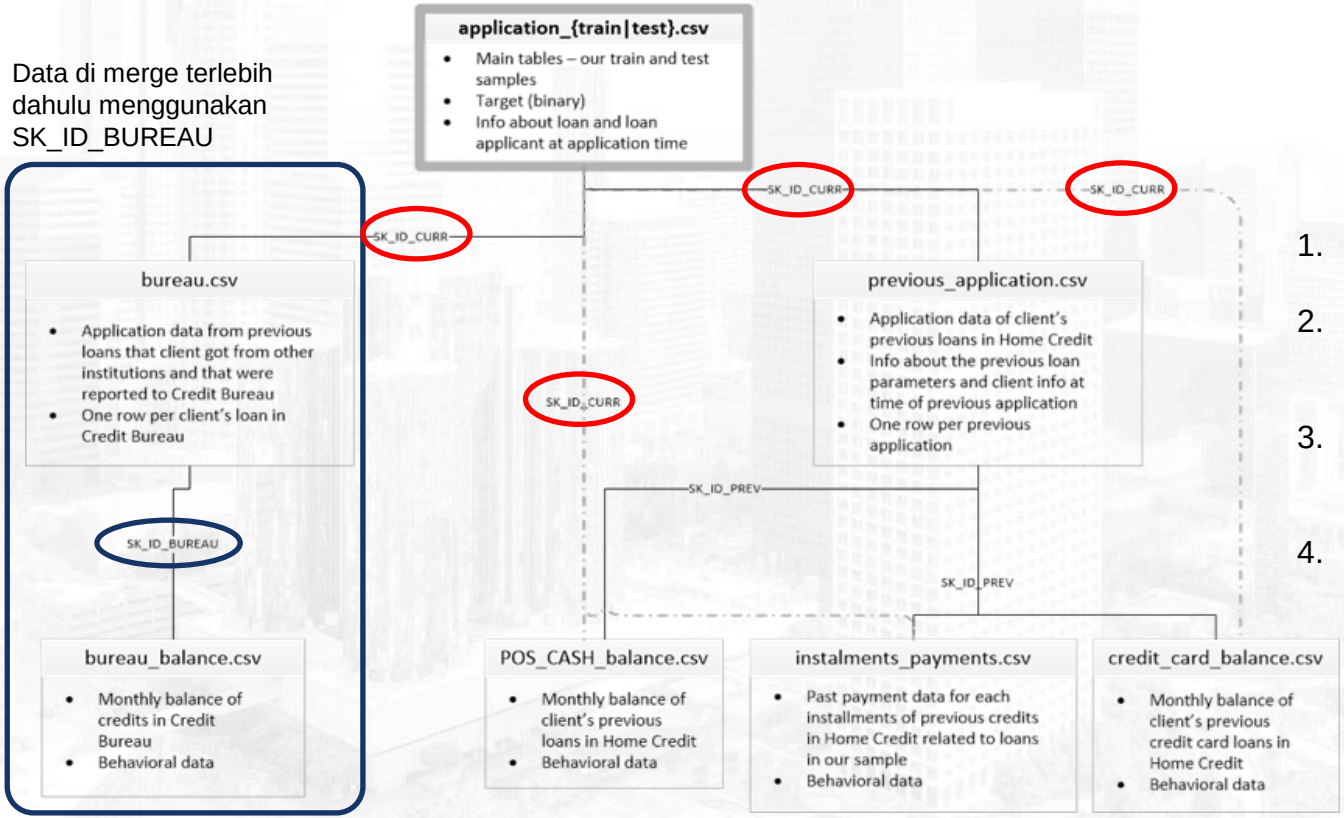



Pada penelitian ini akan dilakukan dua skema, yaitu:

1. Modelling dengan data app train
2. Modelling dengan merge seluruh dataset seperti gambar di samping

# Data Integration

Data di merge terlebih dahulu menggunakan SK_ID_BUREAU




1. Merge data app train dan app test (app train/test)
2. Aggregasi statistic di masing-masing dataset by SK_ID_CURR
3. Seluruh dataset akan di merge by SK_ID_CURR dengan app train/test secara bertahap
4. Preprocessing data yang telah di merge (all dataset)

## Variabel: DAYS_EMPLOYED → tambah kolom DAYS_EMPLOYED_ANOM

| | DAYS_EMPLOYED |
|---|---|
| count | 307511.000000 |
| mean | 63815.045904 |
| std | 141275.766519 |
| min | -17912.000000 |
| 25% | -2760.000000 |
| 50% | -1213.000000 |
| 75% | -289.000000 |
| max | 365243.000000 |

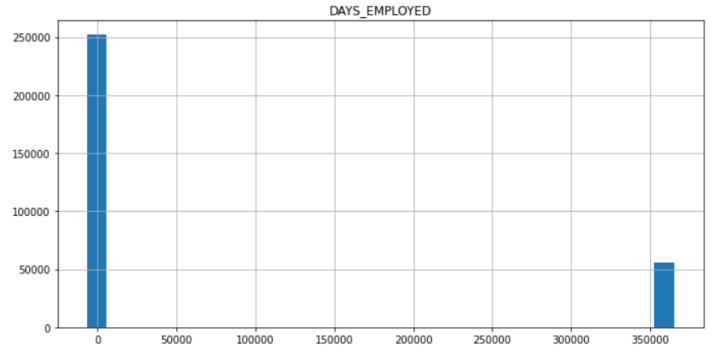
Melihat Anomali pada DAYS_EMPLOYED

```
anom = df[df['DAYS_EMPLOYED'] == 365243]
non_anom = df[df['DAYS_EMPLOYED'] != 365243]
print('Non-anomalies default adalah %0.2f%% dari data pinjaman' % (100 * non_anom['TARGET'].mean()))
print('Anomalies default adalah %0.2f%% dari data pinjaman' % (100 * anom['TARGET'].mean()))
print('Terdapat %d anomali DAYS_EMPLOYED' % len(anom))
```

```
Non-anomalies default adalah 8.66% dari data pinjaman
Anomalies default adalah 5.40% dari data pinjaman
Terdapat 55374 anomali DAYS_EMPLOYED
```

Jika dilihat nilai maksimum pada DAYS_EMPLOYED, terlihat bahwa terdapat kejanggalan pada nilai maksimumnya, yaitu nilai maksimumnya mencapai ribuan tahun.

# Feature Engineering

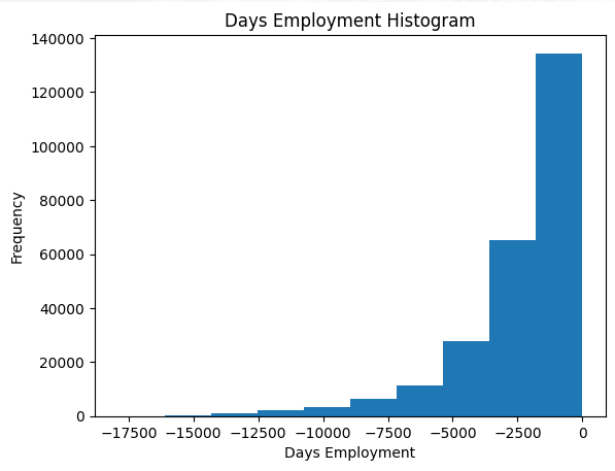



```
# Buat kolom anomali
df['DAYS_EMPLOYED_ANOM'] = df["DAYS_EMPLOYED"] == 365243

# Ganti anomali dengan NaN
df['DAYS_EMPLOYED'].replace({365243: np.nan}, inplace = True)

df['DAYS_EMPLOYED'].plot.hist(title = 'Days Employment Histogram');
plt.xlabel('Days Employment')
```

Dipisahkan antara yang anomali dan tidak, sehingga distribusinya dapat terlihat lebih jelas sebarannya



# Feature Engineering

## Variabel: DAYS_BIRTH → abs(DAYS_BIRTH)

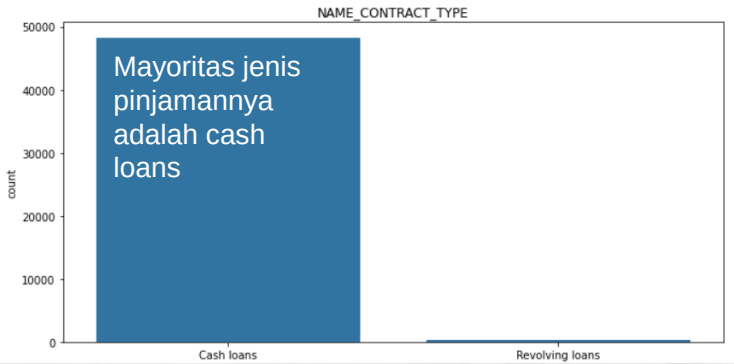
| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| DAYS_BIRTH | 307511.0 | -16038.090267 | 4241.743719 | -23204.0 | -19682.0 | -15750.0 | -12413.0 | -9407.0 |
| REGION_RATING_CLIENT_W_CITY | 307511.0 | 2.031521 | 0.502737 | 1.0 | 2.0 | 2.0 | 2.0 | 3.0 |
| REGION_RATING_CLIENT | 307511.0 | 2.052463 | 0.509034 | 1.0 | 2.0 | 2.0 | 2.0 | 3.0 |
| DAYS_LAST_PHONE_CHANGE | 307511.0 | -945.073230 | 785.624919 | -2522.0 | -1570.0 | -757.0 | -274.0 | 0.0 |
| DAYS_ID_PUBLISH | 307511.0 | -2989.319875 | 1469.619709 | -4944.0 | -4299.0 | -3254.0 | -1720.0 | -375.0 |
| REG_CITY_NOT_WORK_CITY | 307511.0 | 0.230454 | 0.421124 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| FLAG_EMP_PHONE | 307511.0 | 0.819889 | 0.384280 | 0.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| REG_CITY_NOT_LIVE_CITY | 307511.0 | 0.078173 | 0.268444 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| FLAG_DOCUMENT_3 | 307511.0 | 0.710023 | 0.453752 | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 |
| DAYS_REGISTRATION | 307511.0 | -4916.494451 | 3332.170721 | -11416.0 | -7479.5 | -4504.0 | -2010.0 | -330.0 |
| LIVE_CITY_NOT_WORK_CITY | 307511.0 | 0.179555 | 0.383817 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| DEF_30_CNT_SOCIAL_CIRCLE | 307511.0 | 0.114357 | 0.318245 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| DEF_60_CNT_SOCIAL_CIRCLE | 307511.0 | 0.083799 | 0.277086 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| FLAG_WORK_PHONE | 307511.0 | 0.199368 | 0.399526 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| CNT_CHILDREN | 307511.0 | 0.400509 | 0.664724 | 0.0 | 0.0 | 0.0 | 1.0 | 2.0 |

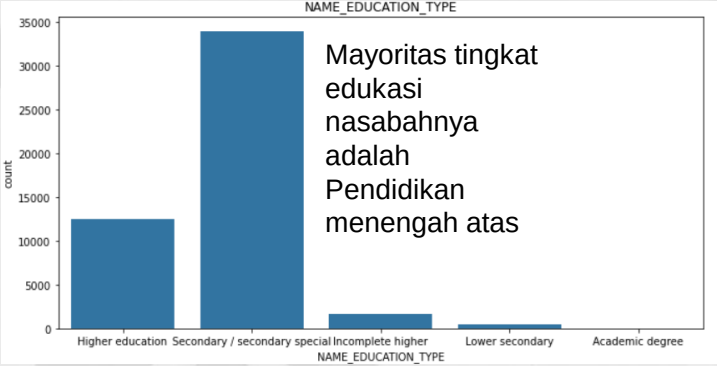
```
Most Positive Correlations:
DAYS_BIRTH                    0.077822
REGION_RATING_CLIENT_W_CITY   0.060893
REGION_RATING_CLIENT          0.058899
DAYS_LAST_PHONE_CHANGE        0.055368
DAYS_ID_PUBLISH               0.051594
REG_CITY_NOT_WORK_CITY        0.050994
FLAG_EMP_PHONE                0.045982
REG_CITY_NOT_LIVE_CITY        0.044395
FLAG_DOCUMENT_3               0.044346
DAYS_REGISTRATION             0.041567
LIVE_CITY_NOT_WORK_CITY       0.032518
DEF_30_CNT_SOCIAL_CIRCLE      0.031962
DEF_60_CNT_SOCIAL_CIRCLE      0.030790
FLAG_WORK_PHONE               0.028524
CNT_CHILDREN                  0.018593
Name: TARGET, dtype: float64
```

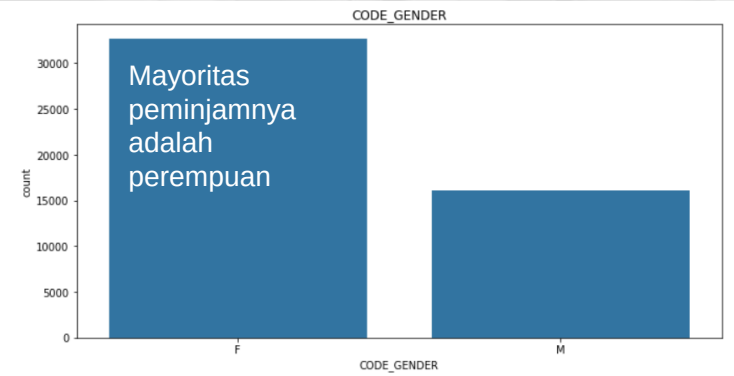
Angka-angka di kolom DAYS_BIRTH adalah negatif karena dicatat berdasarkan aplikasi pinjaman saat ini. Untuk melihat statistik DAYS_BIRTH dalam beberapa tahun, kita dapat mengalikannya dengan -1 dan membaginya dengan jumlah hari dalam setahun.

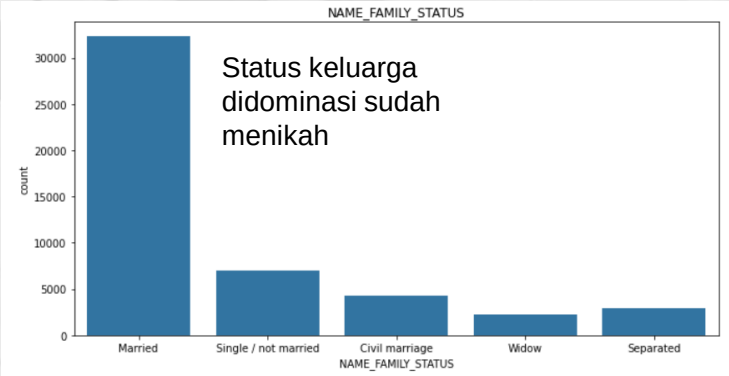
- Tidak terdapat kejanggalan dalam statistik umur. Nilai-nilainya masih dalam batas normal.
- Korelasinya positif, tetapi nilai fitur ini sebenarnya negatif, yang berarti bahwa seiring bertambahnya usia klien, mereka cenderung tidak membayar pinjamannya. Hal tersebut nampaknya kurang sesuai, sehingga akan diambil nilai absolut dari fitur tersebut dan korelasinya akan negatif.

# Exploratory Data Analysis



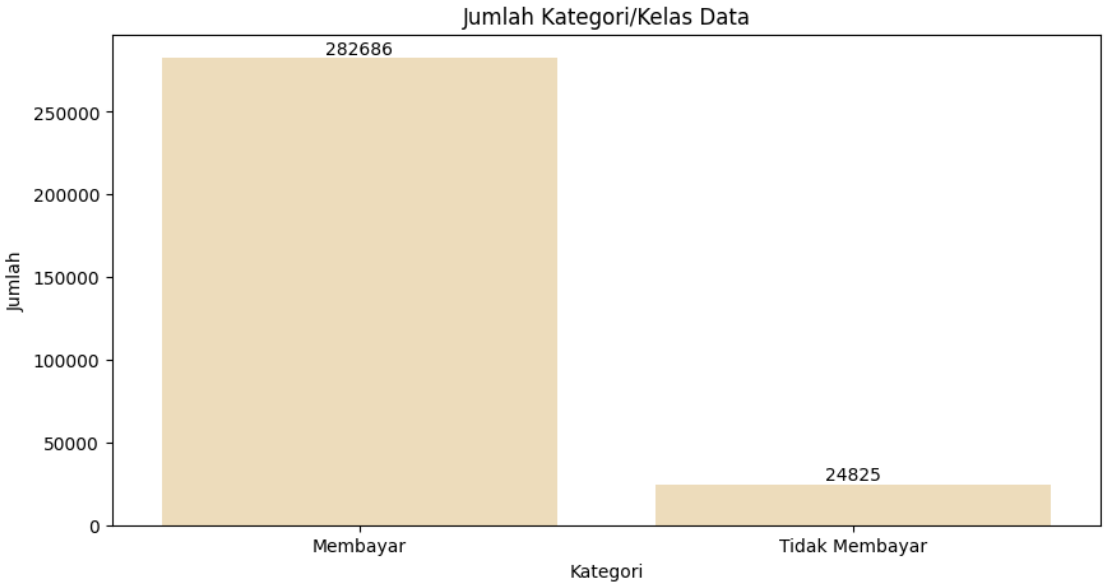




Dari visualisai di samping dapat terlihat beberapa gambaran kondisi data yang digunakan dalam penelitian ini

# Exploratory Data Analysis




Terdapat ketidakseimbangan kelas pada dataset. Untuk mengani permasalahan ini, perlu dilakukan resampling, yaitu salah satunya dengan SMOTE oversampling.


Thank you!
Raih Mimpi #TanpaBatas