CREDIT RISK LOAN

TABLE OF CONTENT

DATA UNDERSTANDING

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 466285 entries, 0 to 466284
Data columns (total 74 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   id                          466285 non-null  int64
 1   member_id                   466285 non-null  int64
 2   loan_amnt                   466285 non-null  int64
 3   funded_amnt                 466285 non-null  int64
 4   funded_amnt_inv             466285 non-null  float64
 5   term                        466285 non-null  object
 6   int_rate                    466285 non-null  float64
 7   installment                 466285 non-null  float64
 8   grade                       466285 non-null  object
 9   sub_grade                   466285 non-null  object
 10  emp_title                   438422 non-null  object
 11  emp_length                  445277 non-null  object
 12  home_ownership              466285 non-null  object
 13  annual_inc                  466281 non-null  float64
 14  verification_status         466285 non-null  object
 15  issue_d                     466285 non-null  object
 16  loan_status                 466285 non-null  object
 17  pymnt_plan                  466285 non-null  object
 18  url                         466285 non-null  object
 19  desc                        125983 non-null  object
 20  purpose                     466285 non-null  object
 21  title                       466265 non-null  object
 22  zip_code                    466285 non-null  object
 23  addr_state                  466285 non-null  object
 24  dti                         466285 non-null  float64
 25  delinq_2yrs                 466256 non-null  float64
 26  earliest_cr_line            466256 non-null  object
 27  inq_last_6mths              466256 non-null  float64
 28  mths_since_last_delinq      215934 non-null  float64
 29  mths_since_last_record      62638 non-null   float64
 30  open_acc                    466256 non-null  float64
 31  pub_rec                     466256 non-null  float64
 32  revol_bal                   466285 non-null  int64
 33  revol_util                  465945 non-null  float64
 34  total_acc                   466256 non-null  float64
 35  initial_list_status         466285 non-null  object
 36  out_prncp                   466285 non-null  float64
 37  out_prncp_inv               466285 non-null  float64
 38  total_pymnt                 466285 non-null  float64
 39  total_pymnt_inv             466285 non-null  float64
 40  total_rec_prncp             466285 non-null  float64
 41  total_rec_int               466285 non-null  float64
 42  total_rec_late_fee          466285 non-null  float64
 43  recoveries                  466285 non-null  float64
 44  collection_recovery_fee     466285 non-null  float64
 45  last_pymnt_d                465909 non-null  object
 46  last_pymnt_amnt             466285 non-null  float64
 47  next_pymnt_d                239071 non-null  object
 48  last_credit_pull_d          466243 non-null  object
 49  collections_12_mths_ex_med  466140 non-null  float64
 50  mths_since_last_major_derog 98974 non-null   float64
 51  policy_code                 466285 non-null  int64
 52  application_type            466285 non-null  object
 53  annual_inc_joint            0 non-null       float64
 54  dti_joint                   0 non-null       float64
 55  verification_status_joint   0 non-null       float64
 56  acc_now_delinq              466256 non-null  float64
 57  tot_coll_amt                396009 non-null  float64
 58  tot_cur_bal                 396009 non-null  float64
 59  open_acc_6m                 0 non-null       float64
 60  open_il_6m                  0 non-null       float64
 61  open_il_12m                 0 non-null       float64
 62  open_il_24m                 0 non-null       float64
 63  mths_since_rcnt_il          0 non-null       float64
 64  total_bal_il                0 non-null       float64
 65  il_util                     0 non-null       float64
 66  open_rv_12m                 0 non-null       float64
 67  open_rv_24m                 0 non-null       float64
 68  max_bal_bc                  0 non-null       float64
 69  all_util                    0 non-null       float64
 70  total_rev_hi_lim            396009 non-null  float64
 71  inq_fi                      0 non-null       float64
 72  total_cu_tl                 0 non-null       float64
 73  inq_last_12m                0 non-null       float64
dtypes: float64(46), int64(6), object(22)
memory usage: 266.8+ MB
```

- Dalam dataset ini memiliki 74 Kolom dengan 466.275 entri
- Memiliki 22 Fitur Kategorikal dan 52 Fitur Numerikal
- Memiliki berbagai tipe data seperti int64, float64, dan object

FITUR NUMERIKAL

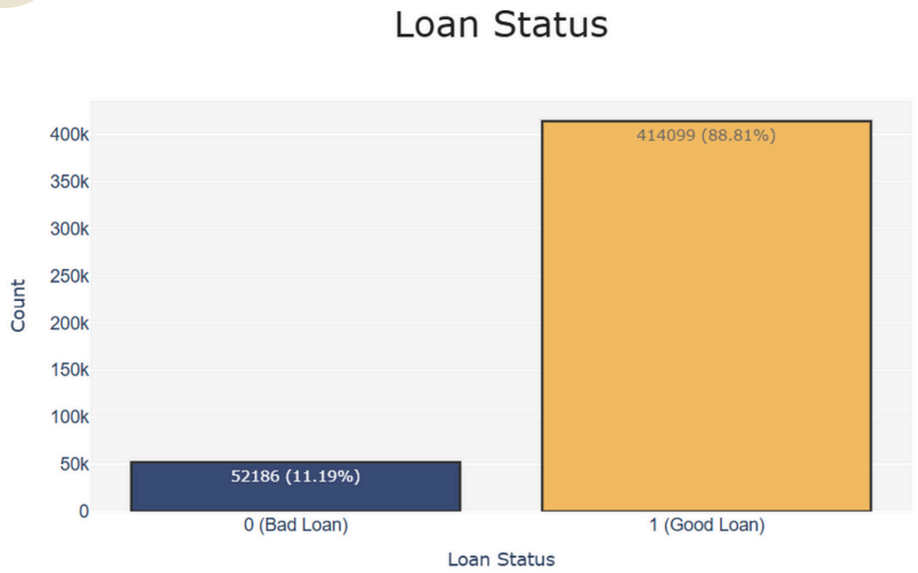
	count	mean	std	min	25%	50%	75%	max
id	466285.0	1.307973e+07	1.089371e+07	54734.00	3.639987e+06	1.010790e+07	2.073121e+07	3.809811e+07
member_id	466285.0	1.459766e+07	1.168237e+07	70473.00	4.379705e+06	1.194108e+07	2.300154e+07	4.086083e+07
loan_amnt	466285.0	1.431728e+04	8.286509e+03	500.00	8.000000e+03	1.200000e+04	2.000000e+04	3.500000e+04
funded_amnt	466285.0	1.429180e+04	8.274371e+03	500.00	8.000000e+03	1.200000e+04	2.000000e+04	3.500000e+04
funded_amnt_inv	466285.0	1.422233e+04	8.297638e+03	0.00	8.000000e+03	1.200000e+04	1.995000e+04	3.500000e+04
int_rate	466285.0	1.382924e+01	4.357587e+00	5.42	1.099000e+01	1.366000e+01	1.649000e+01	2.606000e+01
installment	466285.0	4.320612e+02	2.434855e+02	15.67	2.566900e+02	3.798900e+02	5.665800e+02	1.409990e+03
annual_inc	466281.0	7.327738e+04	5.496357e+04	1896.00	4.500000e+04	6.300000e+04	8.896000e+04	7.500000e+06
dti	466285.0	1.721876e+01	7.851121e+00	0.00	1.136000e+01	1.687000e+01	2.278000e+01	3.999000e+01
delinq_2yrs	466256.0	2.846784e-01	7.973651e-01	0.00	0.000000e+00	0.000000e+00	0.000000e+00	2.900000e+01
inq_last_6mths	466256.0	8.047446e-01	1.091598e+00	0.00	0.000000e+00	0.000000e+00	1.000000e+00	3.300000e+01
mths_since_last_delinq	215934.0	3.410443e+01	2.177849e+01	0.00	1.600000e+01	3.100000e+01	4.900000e+01	1.880000e+02
mths_since_last_record	62638.0	7.430601e+01	3.035765e+01	0.00	5.300000e+01	7.600000e+01	1.020000e+02	1.290000e+02
open_acc	466256.0	1.118707e+01	4.987526e+00	0.00	8.000000e+00	1.000000e+01	1.400000e+01	8.400000e+01
pub_rec	466256.0	1.605642e-01	5.108626e-01	0.00	0.000000e+00	0.000000e+00	0.000000e+00	6.300000e+01
revol_bal	466285.0	1.623020e+04	2.067625e+04	0.00	6.413000e+03	1.176400e+04	2.033300e+04	2.568995e+06
revol_util	465945.0	5.617695e+01	2.373263e+01	0.00	3.920000e+01	5.760000e+01	7.470000e+01	8.923000e+02
total_acc	466256.0	2.506443e+01	1.160014e+01	1.00	1.700000e+01	2.300000e+01	3.200000e+01	1.560000e+02
out_prncp	466285.0	4.410062e+03	6.355079e+03	0.00	0.000000e+00	4.414700e+02	7.341650e+03	3.216038e+04
out_prncp_inv	466285.0	4.408452e+03	6.353198e+03	0.00	0.000000e+00	4.413800e+02	7.338390e+03	3.216038e+04
total_pymnt	466285.0	1.154069e+04	8.265627e+03	0.00	5.552125e+03	9.419251e+03	1.530816e+04	5.777758e+04
total_pymnt_inv	466285.0	1.146989e+04	8.254158e+03	0.00	5.499250e+03	9.355430e+03	1.523131e+04	5.777758e+04
total_rec_prncp	466285.0	8.866015e+03	7.031688e+03	0.00	3.708560e+03	6.817760e+03	1.200000e+04	3.500003e+04
total_rec_int	466285.0	2.588677e+03	2.483810e+03	0.00	9.572800e+02	1.818880e+03	3.304530e+03	2.420562e+04
total_rec_late_fee	466285.0	6.501292e-01	5.265730e+00	0.00	0.000000e+00	0.000000e+00	0.000000e+00	3.586800e+02
recoveries	466285.0	8.534421e+01	5.522161e+02	0.00	0.000000e+00	0.000000e+00	0.000000e+00	3.352027e+04
collection_recovery_fee	466285.0	8.961534e+00	8.549144e+01	0.00	0.000000e+00	0.000000e+00	0.000000e+00	7.002190e+03
last_pymnt_amnt	466285.0	3.123914e+03	5.554737e+03	0.00	3.126200e+02	5.459600e+02	3.187510e+03	3.623444e+04
collections_12_mths_ex_med	466140.0	9.085253e-03	1.086484e-01	0.00	0.000000e+00	0.000000e+00	0.000000e+00	2.000000e+01
mths_since_last_major_derog	98974.0	4.285255e+01	2.166259e+01	0.00	2.600000e+01	4.200000e+01	5.900000e+01	1.880000e+02
policy_code	466285.0	1.000000e+00	0.000000e+00	1.00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
annual_inc_joint	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
dti_joint	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
verification_status_joint	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
acc_now_delinq	466256.0	4.002093e-03	6.863680e-02	0.00	0.000000e+00	0.000000e+00	0.000000e+00	5.000000e+00
tot_coll_amt	396009.0	1.919135e+02	1.463021e+04	0.00	0.000000e+00	0.000000e+00	0.000000e+00	9.152545e+06
tot_cur_bal	396009.0	1.388017e+05	1.521147e+05	0.00	2.861800e+04	8.153900e+04	2.089530e+05	8.000078e+06
open_acc_6m	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
open_il_6m	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
open_il_12m	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
open_il_24m	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
mths_since_rcnt_il	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
total_bal_il	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
il_util	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
open_rv_12m	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
open_rv_24m	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
max_bal_bc	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
all_util	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
total_rev_hi_lim	396009.0	3.037909e+04	3.724713e+04	0.00	1.350000e+04	2.280000e+04	3.790000e+04	9.999999e+06
inq_fi	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
total_cu_tl	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
inq_last_12m	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN

FITUR KATEGORIKAL

	count	unique	top	freq
term	466285	2	36 months	337953
grade	466285	7	B	136929
sub_grade	466285	35	B3	31686
emp_title	438697	205475	Teacher	5399
emp_length	445277	11	10+ years	150049
home_ownership	466285	6	MORTGAGE	235875
verification_status	466285	3	Verified	168055
issue_d	466285	91	Oct-14	38782
loan_status	466285	9	Current	224226
pymnt_plan	466285	2	n	466276
url	466285	466285	https://www.lendingclub.com/browse/loanDetail....	1
desc	125981	124435		234
purpose	466285	14	debt_consolidation	274195
title	466264	63098	Debt consolidation	164075
zip_code	466285	888	945xx	5304
addr_state	466285	50	CA	71450
earliest_cr_line	466256	664	Oct-00	3674
initial_list_status	466285	2	f	303005
last_pymnt_d	465909	98	Jan-16	179620
next_pymnt_d	239071	100	Feb-16	208393
last_credit_pull_d	466243	103	Jan-16	327699
application_type	466285	1	INDIVIDUAL	466285

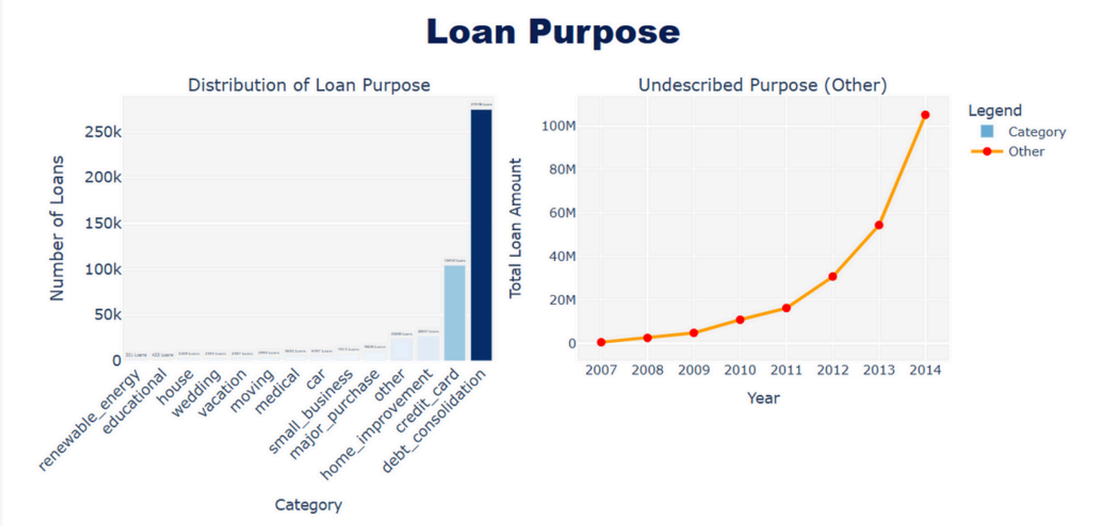
EXPLORATORY DATA ANALYSIS (EDA)

01 Status Pinjaman



Dari kategori loan_status, dipisahkan ke dalam berbagai kategori, lalu dibagi menjadi 2 kategori, yaitu good_loan dan bad_loan, yang akan digunakan sebagai label data.

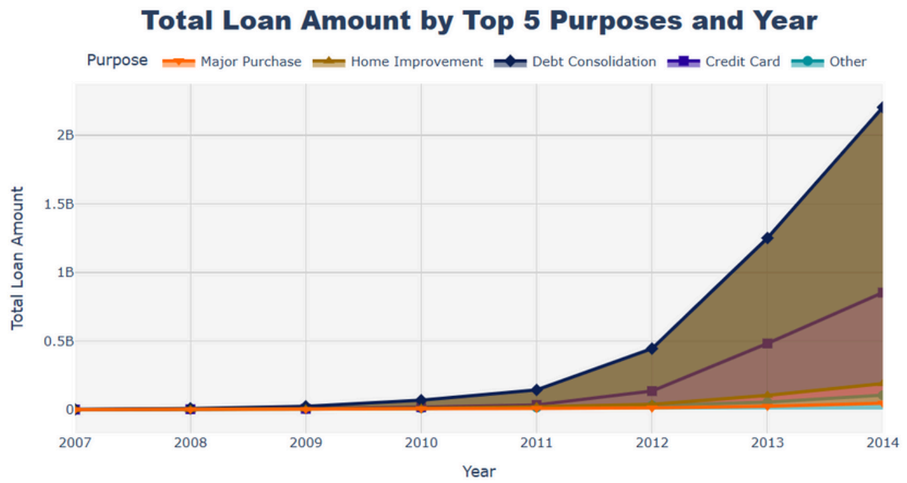
02 Tujuan Pinjaman



Ditampilkan juga tujuan uang dipinjam untuk mengetahui untuk apa pinjaman digunakan, peminjaman paling tinggi terdapat pada kategori debt_consolidation

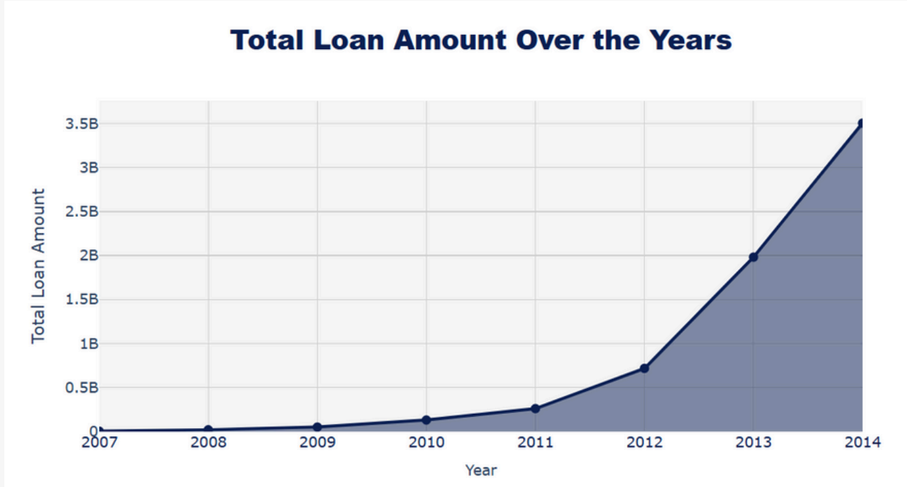
EXPLORATORY DATA ANALYSIS (EDA)

03 Total Besaran Pinjaman Berdasarkan Tujuan dan Tahun nya



Ditampilkan total besaran pinjaman berdasar tujuan dan tahun nya, dapat dilihat debt_consolidation merupakan total pinjaman tertinggi

04 Total Pinjaman



Grafik diatas menampilkan total pinjaman tiap tahun serta grafik kenaikan total pinjaman, dapat dilihat tahun 2012-2014 terjadi kenaikan pinjaman yg cukup besar

DATA PREPARATION

01 Duplikat Data

```
[101]: # For Categorical Data
print('shape before drop = ',df.shape)

df_clean = df.drop(columns=['member_id','id','emp_title','url','desc','title','zip_code','policy_code','application_type'], axis=1)
df_clean.drop_duplicates(inplace=True)

print('shape after drop = ',df_clean.shape)
```

```
shape before drop =  (466285, 77)
shape after drop =  (466285, 68)
```

```
[102]: # For Numerical Data
print('shape before drop = ',df_clean.shape)

df_clean = df_clean.drop(columns=['annual_inc_joint','dti_joint','verification_status_joint','open_acc_6m','open_il_6m','open_il_12m','open_il_24m','mths
'mths_since_rcnt_il','total_bal_il','il_util','open_rv_12m','open_rv_24m','max_bal_bc','all_util','inq_fi','inq_last_12m','total_cu_tl'], axis=1)

print('shape after drop = ',df_clean.shape)
```

```
shape before drop =  (466285, 68)
shape after drop =  (466285, 51)
```

Dataset Tidak Rapi

- Hapus data yang tidak diperlukan dengan banyak duplikat: Data yang berulang atau redundan harus dihilangkan untuk memastikan kualitas dataset.
- Hapus data yang memiliki nilai NaN di setiap baris: Baris yang seluruhnya berisi nilai NaN (tidak terdefinisi) harus dihapus karena tidak memberikan informasi yang berguna.
- Setelah proses pembersihan data, 51 dari 74 kolom tersisa

02 Konvert string ke int

```
[104]: # Drop kata months di kolom term
df_pre['term'] = df_pre['term'].str.replace(' months', '')
df_pre['term'] = df_pre['term'].astype(float)
```

```
[105]: # Drop data_years di kolom emp_length
df_pre['emp_length'] = df_pre['emp_length'].str.replace('\+ years', '')
df_pre['emp_length'] = df_pre['emp_length'].str.replace('< ', '')
df_pre['emp_length'] = df_pre['emp_length'].str.replace('< 1 year', str(0))
df_pre['emp_length'] = df_pre['emp_length'].str.replace(' years', '')
df_pre['emp_length'] = df_pre['emp_length'].str.replace(' year', '')
df_pre['emp_length'] = df_pre['emp_length'].astype(float)
```

Lakukan drop kata month pada kolom term serta years pada kolom emp_length agar dapat diubah menjadi data int

MODELLING

Modelling Dengan Data Tak Seimbang

	Latihan Akurasi	Uji Akurasi
Decision tree	1.0000	0.9844
Random Forest	1.0000	0.9907
Logistic Regression	0.9466	0.9468

Modelling Dengan Undersampling

	Latihan Akurasi	Uji Akurasi
Decision tree	1.0000	0.9970
Random Forest	1.0000	0.9983
Logistic Regression	0.9473	0.9476

Terlihat dari data diatas, Modelling dengan Undersampling menggunakan Algoritma Random Forest merupakan cara yg paling optimal.