RISK ANALYTICS : LOAN DEFAULT ANALYSIS

PRESENTER : DR. SAMUEL ISRAEL

PRODUCT MANAGER | BUSINESS ANALYST | DATA ANALYST | CSPO

DATE 18 th FEBRUARY, 20025

In [2]:

```python
# importing python libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error,r2_score
```

Importing & reading datasets

In [4]:

```python
# importing & reading the  application_data

df_app = pd.read_csv(r'C:\Users\dell\Downloads\application_data.csv')
df_app        # df_app = application_data
```

Out[4]:

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AM
0	100002	1	Cash loans	M	N	Y	0	202500.0	406597.5	
1	100003	0	Cash loans	F	N	N	0	270000.0	1293502.5	
2	100004	0	Revolving loans	M	Y	Y	0	67500.0	135000.0	
3	100006	0	Cash loans	F	N	Y	0	135000.0	312682.5	
4	100007	0	Cash loans	M	N	Y	0	121500.0	513000.0	
...	...	...	...	...	...	...	...	...	...	
307506	456251	0	Cash loans	M	N	N	0	157500.0	254700.0	
307507	456252	0	Cash loans	F	N	Y	0	72000.0	269550.0	
307508	456253	0	Cash loans	F	N	Y	0	153000.0	677664.0	
307509	456254	1	Cash loans	F	N	Y	0	171000.0	370107.0	
307510	456255	0	Cash loans	F	N	N	0	157500.0	675000.0	

307511 rows × 122 columns

In [5]:

```python
# importing & reading previous_application
df_prev = pd.read_csv(r'C:\Users\dell\Downloads\previous_application.csv')
df_prev     # df_prev = previous_application
```

Out[5]:

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	WEEKDAY_A
0	2030495	271877	Consumer loans	1730.430	17145.0	17145.0	0.0	17145.0	
1	2802425	108129	Cash loans	25188.615	607500.0	679671.0	NaN	607500.0	
2	2523466	122040	Cash loans	15060.735	112500.0	136444.5	NaN	112500.0	
3	2819243	176158	Cash loans	47041.335	450000.0	470790.0	NaN	450000.0	
4	1784265	202054	Cash loans	31924.395	337500.0	404055.0	NaN	337500.0	
...	...	...	...	...	...	...	...	...	
1670209	2300464	352015	Consumer loans	14704.290	267295.5	311400.0	0.0	267295.5	
1670210	2357031	334635	Consumer loans	6622.020	87750.0	64291.5	29250.0	87750.0	
1670211	2659632	249544	Consumer loans	11520.855	105237.0	102523.5	10525.5	105237.0	
1670212	2785582	400317	Cash loans	18821.520	180000.0	191880.0	NaN	180000.0	
1670213	2418762	261212	Cash loans	16431.300	360000.0	360000.0	NaN	360000.0	

1670214 rows × 37 columns

Data explorations & preprocessing

In [7]:

```python
# checking the basic informations of the datasets
df_app.head()
```

Out[7]:

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANN
0	100002	1	Cash loans	M	N	Y	0	202500.0	406597.5	24
1	100003	0	Cash loans	F	N	N	0	270000.0	1293502.5	35
2	100004	0	Revolving loans	M	Y	Y	0	67500.0	135000.0	6
3	100006	0	Cash loans	F	N	Y	0	135000.0	312682.5	29
4	100007	0	Cash loans	M	N	Y	0	121500.0	513000.0	21

5 rows × 122 columns

```
In [8]: df_app.tail()
```

Out[8]:

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AM
307506	456251	0	Cash loans	M	N	N	0	157500.0	254700.0	
307507	456252	0	Cash loans	F	N	Y	0	72000.0	269550.0	
307508	456253	0	Cash loans	F	N	Y	0	153000.0	677664.0	
307509	456254	1	Cash loans	F	N	Y	0	171000.0	370107.0	
307510	456255	0	Cash loans	F	N	N	0	157500.0	675000.0	

5 rows × 122 columns

```
In [9]: df_app.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 307511 entries, 0 to 307510
Columns: 122 entries, SK_ID_CURR to AMT_REQ_CREDIT_BUREAU_YEAR
dtypes: float64(65), int64(41), object(16)
memory usage: 286.2+ MB
```

```
In [10]: df_app.dtypes
```

```
Out[10]: SK_ID_CURR                      int64
         TARGET                          int64
         NAME_CONTRACT_TYPE             object
         CODE_GENDER                    object
         FLAG_OWN_CAR                   object
                                         ...
         AMT_REQ_CREDIT_BUREAU_DAY     float64
         AMT_REQ_CREDIT_BUREAU_WEEK    float64
         AMT_REQ_CREDIT_BUREAU_MON     float64
         AMT_REQ_CREDIT_BUREAU_QRT     float64
         AMT_REQ_CREDIT_BUREAU_YEAR    float64
         Length: 122, dtype: object
```

```
In [11]: df_app.describe()
```

Out[11]:

	SK_ID_CURR	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAY
count	307511.000000	307511.000000	307511.000000	3.075110e+05	3.075110e+05	307499.000000	3.072330e+05	307511.000000	30751
mean	278180.518577	0.080729	0.417052	1.687979e+05	5.990260e+05	27108.573909	5.383962e+05	0.020868	-1603
std	102790.175348	0.272419	0.722121	2.371231e+05	4.024908e+05	14493.737315	3.694465e+05	0.013831	436
min	100002.000000	0.000000	0.000000	2.565000e+04	4.500000e+04	1615.500000	4.050000e+04	0.000290	-2522
25%	189145.500000	0.000000	0.000000	1.125000e+05	2.700000e+05	16524.000000	2.385000e+05	0.010006	-1968
50%	278202.000000	0.000000	0.000000	1.471500e+05	5.135310e+05	24903.000000	4.500000e+05	0.018850	-1575
75%	367142.500000	0.000000	1.000000	2.025000e+05	8.086500e+05	34596.000000	6.795000e+05	0.028663	-1241
max	456255.000000	1.000000	19.000000	1.170000e+08	4.050000e+06	258025.500000	4.050000e+06	0.072508	-748

8 rows × 106 columns

```
In [12]: df_app.shape
```

```
Out[12]: (307511, 122)
```

```
In [13]: df_app.columns
```

```
Out[13]: Index(['SK_ID_CURR', 'TARGET', 'NAME_CONTRACT_TYPE', 'CODE_GENDER',
                'FLAG_OWN_CAR', 'FLAG_OWN_REALTY', 'CNT_CHILDREN', 'AMT_INCOME_TOTAL',
                'AMT_CREDIT', 'AMT_ANNUITY',
                ...
                'FLAG_DOCUMENT_18', 'FLAG_DOCUMENT_19', 'FLAG_DOCUMENT_20',
                'FLAG_DOCUMENT_21', 'AMT_REQ_CREDIT_BUREAU_HOUR',
                'AMT_REQ_CREDIT_BUREAU_DAY', 'AMT_REQ_CREDIT_BUREAU_WEEK',
                'AMT_REQ_CREDIT_BUREAU_MON', 'AMT_REQ_CREDIT_BUREAU_QRT',
                'AMT_REQ_CREDIT_BUREAU_YEAR'],
               dtype='object', length=122)
```

```
In [14]: df_prev.head()
```

Out[14]:

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	WEEKDAY_APPR_PI
0	2030495	271877	Consumer loans	1730.430	17145.0	17145.0	0.0	17145.0	
1	2802425	108129	Cash loans	25188.615	607500.0	679671.0	NaN	607500.0	
2	2523466	122040	Cash loans	15060.735	112500.0	136444.5	NaN	112500.0	
3	2819243	176158	Cash loans	47041.335	450000.0	470790.0	NaN	450000.0	
4	1784265	202054	Cash loans	31924.395	337500.0	404055.0	NaN	337500.0	

5 rows × 37 columns

```
In [15]: df_prev.tail()
```

Loading [MathJax]/extensions/Safe.js

Out[15]:

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	WEEKDAY_A
1670209	2300464	352015	Consumer loans	14704.290	267295.5	311400.0	0.0	267295.5	
1670210	2357031	334635	Consumer loans	6622.020	87750.0	64291.5	29250.0	87750.0	
1670211	2659632	249544	Consumer loans	11520.855	105237.0	102523.5	10525.5	105237.0	
1670212	2785582	400317	Cash loans	18821.520	180000.0	191880.0	NaN	180000.0	
1670213	2418762	261212	Cash loans	16431.300	360000.0	360000.0	NaN	360000.0	

5 rows × 37 columns

In [16]:

```
df_prev.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1670214 entries, 0 to 1670213
Data columns (total 37 columns):
 #   Column                       Non-Null Count    Dtype  
---  ------                       --------------    -----  
 0   SK_ID_PREV                   1670214 non-null  int64  
 1   SK_ID_CURR                   1670214 non-null  int64  
 2   NAME_CONTRACT_TYPE           1670214 non-null  object 
 3   AMT_ANNUITY                  1297979 non-null  float64
 4   AMT_APPLICATION              1670214 non-null  float64
 5   AMT_CREDIT                   1670213 non-null  float64
 6   AMT_DOWN_PAYMENT             774370 non-null   float64
 7   AMT_GOODS_PRICE              1284699 non-null  float64
 8   WEEKDAY_APPR_PROCESS_START   1670214 non-null  object 
 9   HOUR_APPR_PROCESS_START      1670214 non-null  int64  
 10  FLAG_LAST_APPL_PER_CONTRACT  1670214 non-null  object 
 11  NFLAG_LAST_APPL_IN_DAY       1670214 non-null  int64  
 12  RATE_DOWN_PAYMENT            774370 non-null   float64
 13  RATE_INTEREST_PRIMARY        5951 non-null     float64
 14  RATE_INTEREST_PRIVILEGED     5951 non-null     float64
 15  NAME_CASH_LOAN_PURPOSE       1670214 non-null  object 
 16  NAME_CONTRACT_STATUS         1670214 non-null  object 
 17  DAYS_DECISION                1670214 non-null  int64  
 18  NAME_PAYMENT_TYPE            1670214 non-null  object 
 19  CODE_REJECT_REASON           1670214 non-null  object 
 20  NAME_TYPE_SUITE              849809 non-null   object 
 21  NAME_CLIENT_TYPE             1670214 non-null  object 
 22  NAME_GOODS_CATEGORY          1670214 non-null  object 
 23  NAME_PORTFOLIO               1670214 non-null  object 
 24  NAME_PRODUCT_TYPE            1670214 non-null  object 
 25  CHANNEL_TYPE                 1670214 non-null  object 
 26  SELLERPLACE_AREA             1670214 non-null  int64  
 27  NAME_SELLER_INDUSTRY         1670214 non-null  object 
 28  CNT_PAYMENT                  1297984 non-null  float64
 29  NAME_YIELD_GROUP             1670214 non-null  object 
 30  PRODUCT_COMBINATION          1669868 non-null  object 
 31  DAYS_FIRST_DRAWING           997149 non-null   float64
 32  DAYS_FIRST_DUE               997149 non-null   float64
 33  DAYS_LAST_DUE_1ST_VERSION    997149 non-null   float64
 34  DAYS_LAST_DUE                997149 non-null   float64
 35  DAYS_TERMINATION             997149 non-null   float64
 36  NFLAG_INSURED_ON_APPROVAL    997149 non-null   float64
dtypes: float64(15), int64(6), object(16)
memory usage: 471.5+ MB
```

In [17]:

```
df_prev.dtypes
```

Out[17]:

```
SK_ID_PREV                       int64
SK_ID_CURR                       int64
NAME_CONTRACT_TYPE              object
AMT_ANNUITY                    float64
AMT_APPLICATION                float64
AMT_CREDIT                     float64
AMT_DOWN_PAYMENT               float64
AMT_GOODS_PRICE                float64
WEEKDAY_APPR_PROCESS_START      object
HOUR_APPR_PROCESS_START          int64
FLAG_LAST_APPL_PER_CONTRACT     object
NFLAG_LAST_APPL_IN_DAY           int64
RATE_DOWN_PAYMENT              float64
RATE_INTEREST_PRIMARY          float64
RATE_INTEREST_PRIVILEGED       float64
NAME_CASH_LOAN_PURPOSE          object
NAME_CONTRACT_STATUS            object
DAYS_DECISION                    int64
NAME_PAYMENT_TYPE               object
CODE_REJECT_REASON              object
NAME_TYPE_SUITE                 object
NAME_CLIENT_TYPE                object
NAME_GOODS_CATEGORY             object
NAME_PORTFOLIO                  object
NAME_PRODUCT_TYPE               object
CHANNEL_TYPE                    object
SELLERPLACE_AREA                 int64
NAME_SELLER_INDUSTRY            object
CNT_PAYMENT                    float64
NAME_YIELD_GROUP                object
PRODUCT_COMBINATION             object
DAYS_FIRST_DRAWING             float64
DAYS_FIRST_DUE                 float64
DAYS_LAST_DUE_1ST_VERSION      float64
DAYS_LAST_DUE                  float64
DAYS_TERMINATION               float64
NFLAG_INSURED_ON_APPROVAL      float64
dtype: object
```

In [18]:

```
df_prev.describe()
```

Out[18]:

	SK_ID_PREV	SK_ID_CURR	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	HOUR_APPR_PROCESS_START	NFLA

count	1.670214e+06	1.670214e+06	1.297979e+06	1.670214e+06	1.670213e+06	7.743700e+05	1.284699e+06	1.670214e+06
mean	1.923089e+06	2.783572e+05	1.595512e+04	1.752339e+05	1.961140e+05	6.697402e+03	2.278473e+05	1.248418e+01
std	5.325980e+05	1.028148e+05	1.478214e+04	2.927798e+05	3.185746e+05	2.092150e+04	3.153966e+05	3.334028e+00
min	1.000001e+06	1.000010e+05	0.000000e+00	0.000000e+00	0.000000e+00	-9.000000e-01	0.000000e+00	0.000000e+00
25%	1.461857e+06	1.893290e+05	6.321780e+03	1.872000e+04	2.416050e+04	0.000000e+00	5.084100e+04	1.000000e+01
50%	1.923110e+06	2.787145e+05	1.125000e+04	7.104600e+04	8.054100e+04	1.638000e+03	1.123200e+05	1.200000e+01
75%	2.384280e+06	3.675140e+05	2.065842e+04	1.803600e+05	2.164185e+05	7.740000e+03	2.340000e+05	1.500000e+01
max	2.845382e+06	4.562550e+05	4.180581e+05	6.905160e+06	6.905160e+06	3.060045e+06	6.905160e+06	2.300000e+01

8 rows × 21 columns

```
In [19]: df_prev.shape
```

```
Out[19]: (1670214, 37)
```

```
In [20]: df_prev.columns
```

```
Out[20]: Index(['SK_ID_PREV', 'SK_ID_CURR', 'NAME_CONTRACT_TYPE', 'AMT_ANNUITY',
                'AMT_APPLICATION', 'AMT_CREDIT', 'AMT_DOWN_PAYMENT', 'AMT_GOODS_PRICE',
                'WEEKDAY_APPR_PROCESS_START', 'HOUR_APPR_PROCESS_START',
                'FLAG_LAST_APPL_PER_CONTRACT', 'NFLAG_LAST_APPL_IN_DAY',
                'RATE_DOWN_PAYMENT', 'RATE_INTEREST_PRIMARY',
                'RATE_INTEREST_PRIVILEGED', 'NAME_CASH_LOAN_PURPOSE',
                'NAME_CONTRACT_STATUS', 'DAYS_DECISION', 'NAME_PAYMENT_TYPE',
                'CODE_REJECT_REASON', 'NAME_TYPE_SUITE', 'NAME_CLIENT_TYPE',
                'NAME_GOODS_CATEGORY', 'NAME_PORTFOLIO', 'NAME_PRODUCT_TYPE',
                'CHANNEL_TYPE', 'SELLERPLACE_AREA', 'NAME_SELLER_INDUSTRY',
                'CNT_PAYMENT', 'NAME_YIELD_GROUP', 'PRODUCT_COMBINATION',
                'DAYS_FIRST_DRAWING', 'DAYS_FIRST_DUE', 'DAYS_LAST_DUE_1ST_VERSION',
                'DAYS_LAST_DUE', 'DAYS_TERMINATION', 'NFLAG_INSURED_ON_APPROVAL'],
               dtype='object')
```

```
In [21]: # Checking & handling missing values
         df_app.isnull()
```

Out[21]:

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AM
0	False	False	False	False	False	False	False	False	False	
1	False	False	False	False	False	False	False	False	False	
2	False	False	False	False	False	False	False	False	False	
3	False	False	False	False	False	False	False	False	False	
4	False	False	False	False	False	False	False	False	False	
...	...	...	...	...	...	...	...	...	...	
307506	False	False	False	False	False	False	False	False	False	
307507	False	False	False	False	False	False	False	False	False	
307508	False	False	False	False	False	False	False	False	False	
307509	False	False	False	False	False	False	False	False	False	
307510	False	False	False	False	False	False	False	False	False	

307511 rows × 122 columns

```
In [22]: df_app.isnull().sum()
```

```
Out[22]: SK_ID_CURR                        0
         TARGET                            0
         NAME_CONTRACT_TYPE                0
         CODE_GENDER                       0
         FLAG_OWN_CAR                      0
                                       ...  
         AMT_REQ_CREDIT_BUREAU_DAY     41519
         AMT_REQ_CREDIT_BUREAU_WEEK    41519
         AMT_REQ_CREDIT_BUREAU_MON     41519
         AMT_REQ_CREDIT_BUREAU_QRT     41519
         AMT_REQ_CREDIT_BUREAU_YEAR    41519
         Length: 122, dtype: int64
```

```
In [23]: df_app.isnull().sum
```

```
Out[23]: <bound method DataFrame.sum of         SK_ID_CURR  TARGET  NAME_CONTRACT_TYPE  CODE_GENDER  FLAG_OWN_CAR  \
         0            False   False               False        False         False   
         1            False   False               False        False         False   
         2            False   False               False        False         False   
         3            False   False               False        False         False   
         4            False   False               False        False         False   
         ...            ...     ...                 ...          ...           ...   
         307506       False   False               False        False         False   
         307507       False   False               False        False         False   
         307508       False   False               False        False         False   
         307509       False   False               False        False         False   
         307510       False   False               False        False         False   

                 FLAG_OWN_REALTY  CNT_CHILDREN  AMT_INCOME_TOTAL  AMT_CREDIT  \
         0                 False         False             False       False   
         1                 False         False             False       False   
         2                 False         False             False       False   
         3                 False         False             False       False   
         4                 False         False             False       False   
         ...                 ...           ...               ...         ...   
         307506            False         False             False       False   
         307507            False         False             False       False   
         307508            False         False             False       False   
         307509            False         False             False       False   
                           False         False             False       False   
```

```
        AMT_ANNUITY  ...  FLAG_DOCUMENT_18  FLAG_DOCUMENT_19  \
0             False  ...             False             False   
1             False  ...             False             False   
2             False  ...             False             False   
3             False  ...             False             False   
4             False  ...             False             False   
...             ...  ...               ...               ...   
307506        False  ...             False             False   
307507        False  ...             False             False   
307508        False  ...             False             False   
307509        False  ...             False             False   
307510        False  ...             False             False   

        FLAG_DOCUMENT_20  FLAG_DOCUMENT_21  AMT_REQ_CREDIT_BUREAU_HOUR  \
0                  False             False                       False   
1                  False             False                       False   
2                  False             False                       False   
3                  False             False                        True   
4                  False             False                       False   
...                  ...               ...                         ...   
307506             False             False                        True   
307507             False             False                        True   
307508             False             False                       False   
307509             False             False                       False   
307510             False             False                       False   

        AMT_REQ_CREDIT_BUREAU_DAY  AMT_REQ_CREDIT_BUREAU_WEEK  \
0                           False                       False   
1                           False                       False   
2                           False                       False   
3                            True                        True   
4                           False                       False   
...                           ...                         ...   
307506                       True                        True   
307507                       True                        True   
307508                      False                       False   
307509                      False                       False   
307510                      False                       False   

        AMT_REQ_CREDIT_BUREAU_MON  AMT_REQ_CREDIT_BUREAU_QRT  \
0                           False                      False   
1                           False                      False   
2                           False                      False   
3                            True                       True   
4                           False                      False   
...                           ...                        ...   
307506                       True                       True   
307507                       True                       True   
307508                      False                      False   
307509                      False                      False   
307510                      False                      False   

        AMT_REQ_CREDIT_BUREAU_YEAR  
0                            False  
1                            False  
2                            False  
3                             True  
4                            False  
...                            ...  
307506                        True  
307507                        True  
307508                       False  
307509                       False  
307510                       False  

[307511 rows x 122 columns]>
```

In [24]:
```
missing_percentage = df_app.isnull().mean() * 100
missing_percentage
```

Out[24]:
```
SK_ID_CURR                     0.000000
TARGET                         0.000000
NAME_CONTRACT_TYPE             0.000000
CODE_GENDER                    0.000000
FLAG_OWN_CAR                   0.000000
                                 ...    
AMT_REQ_CREDIT_BUREAU_DAY     13.501631
AMT_REQ_CREDIT_BUREAU_WEEK    13.501631
AMT_REQ_CREDIT_BUREAU_MON     13.501631
AMT_REQ_CREDIT_BUREAU_QRT     13.501631
AMT_REQ_CREDIT_BUREAU_YEAR    13.501631
Length: 122, dtype: float64
```

In [25]:
```
df_app_cleaned = df_app.dropna()
df_app_cleaned
```

Out[25]:

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AM
71	100083	0	Cash loans	M	Y	Y	0	103500.0	573628.5	
124	100145	0	Cash loans	F	Y	Y	1	202500.0	260725.5	
152	100179	0	Cash loans	F	Y	N	0	202500.0	675000.0	
161	100190	0	Cash loans	M	Y	N	0	162000.0	263686.5	
255	100295	1	Cash loans	M	Y	N	1	225000.0	1019205.0	
...	...	...	...	...	...	...	...	...	...	
307358	456083	0	Cash loans	F	Y	Y	2	112500.0	361462.5	
307359	456084	0	Cash loans	F	Y	Y	1	99000.0	675000.0	
307407	456140	1	Cash loans	F	Y	Y	1	261000.0	711454.5	
307456	456195	0	Cash loans	F	Y	Y	0	94500.0	270000.0	
	456226	0	Cash loans	F	Y	Y	0	225000.0	500566.5	

8602 rows × 122 columns

In [26]:
```
missing_percentage = df_app_cleaned.isnull().mean() * 100
missing_percentage
```

Out[26]:
```
SK_ID_CURR                    0.0
TARGET                        0.0
NAME_CONTRACT_TYPE            0.0
CODE_GENDER                   0.0
FLAG_OWN_CAR                  0.0
                             ...
AMT_REQ_CREDIT_BUREAU_DAY     0.0
AMT_REQ_CREDIT_BUREAU_WEEK    0.0
AMT_REQ_CREDIT_BUREAU_MON     0.0
AMT_REQ_CREDIT_BUREAU_QRT     0.0
AMT_REQ_CREDIT_BUREAU_YEAR    0.0
Length: 122, dtype: float64
```

In [27]:
```
duplicates = df_app[df_app.duplicated()]
duplicates
```

Out[27]:

SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNU

0 rows × 122 columns

In [28]:
```
app_data = df_app_cleaned
app_data
```

Out[28]:

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AM
71	100083	0	Cash loans	M	Y	Y	0	103500.0	573628.5	
124	100145	0	Cash loans	F	Y	Y	1	202500.0	260725.5	
152	100179	0	Cash loans	F	Y	N	0	202500.0	675000.0	
161	100190	0	Cash loans	M	Y	N	0	162000.0	263686.5	
255	100295	1	Cash loans	M	Y	N	1	225000.0	1019205.0	
...	...	...	...	...	...	...	...	...	...	
307358	456083	0	Cash loans	F	Y	Y	2	112500.0	361462.5	
307359	456084	0	Cash loans	F	Y	Y	1	99000.0	675000.0	
307407	456140	1	Cash loans	F	Y	Y	1	261000.0	711454.5	
307456	456195	0	Cash loans	F	Y	Y	0	94500.0	270000.0	
307482	456226	0	Cash loans	F	Y	Y	0	225000.0	500566.5	

8602 rows × 122 columns

In [29]:
```
app_data.head()
```

Out[29]:

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_A
71	100083	0	Cash loans	M	Y	Y	0	103500.0	573628.5	
124	100145	0	Cash loans	F	Y	Y	1	202500.0	260725.5	
152	100179	0	Cash loans	F	Y	N	0	202500.0	675000.0	
161	100190	0	Cash loans	M	Y	N	0	162000.0	263686.5	
255	100295	1	Cash loans	M	Y	N	1	225000.0	1019205.0	

5 rows × 122 columns

In [30]:
```
app_data.tail()
```

Out[30]:

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AM
307358	456083	0	Cash loans	F	Y	Y	2	112500.0	361462.5	
307359	456084	0	Cash loans	F	Y	Y	1	99000.0	675000.0	
307407	456140	1	Cash loans	F	Y	Y	1	261000.0	711454.5	
307456	456195	0	Cash loans	F	Y	Y	0	94500.0	270000.0	
307482	456226	0	Cash loans	F	Y	Y	0	225000.0	500566.5	

5 rows × 122 columns

In [31]:
```
app_data.shape
```

Out[31]:
```
(8602, 122)
```

In [32]:
```
df_prev.isnull()
```

Out[32]:

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	WEEKDAY_A
0	False	False	False	False	False	False	False	False	
1	False	False	False	False	False	False	True	False	
2	False	False	False	False	False	False	True	False	
3	False	False	False	False	False	False	True	False	
4	False	False	False	False	False	False	True	False	
...	...	...	...	...	...	...	...	...	
1670209	False	False	False	False	False	False	False	False	
1670210	False	False	False	False	False	False	False	False	

1670211	False	False	False	False	False	False	False	False
1670212	False	False	False	False	False	False	True	False
1670213	False	False	False	False	False	False	True	False

1670214 rows × 37 columns

In [33]:
```
df_prev.isnull().sum()
```

Out[33]:
```
SK_ID_PREV                           0
SK_ID_CURR                           0
NAME_CONTRACT_TYPE                   0
AMT_ANNUITY                     372235
AMT_APPLICATION                      0
AMT_CREDIT                           1
AMT_DOWN_PAYMENT                895844
AMT_GOODS_PRICE                 385515
WEEKDAY_APPR_PROCESS_START           0
HOUR_APPR_PROCESS_START              0
FLAG_LAST_APPL_PER_CONTRACT          0
NFLAG_LAST_APPL_IN_DAY               0
RATE_DOWN_PAYMENT               895844
RATE_INTEREST_PRIMARY          1664263
RATE_INTEREST_PRIVILEGED       1664263
NAME_CASH_LOAN_PURPOSE               0
NAME_CONTRACT_STATUS                 0
DAYS_DECISION                        0
NAME_PAYMENT_TYPE                    0
CODE_REJECT_REASON                   0
NAME_TYPE_SUITE                 820405
NAME_CLIENT_TYPE                     0
NAME_GOODS_CATEGORY                  0
NAME_PORTFOLIO                       0
NAME_PRODUCT_TYPE                    0
CHANNEL_TYPE                         0
SELLERPLACE_AREA                     0
NAME_SELLER_INDUSTRY                 0
CNT_PAYMENT                     372230
NAME_YIELD_GROUP                     0
PRODUCT_COMBINATION                346
DAYS_FIRST_DRAWING              673065
DAYS_FIRST_DUE                  673065
DAYS_LAST_DUE_1ST_VERSION       673065
DAYS_LAST_DUE                   673065
DAYS_TERMINATION                673065
NFLAG_INSURED_ON_APPROVAL       673065
dtype: int64
```

In [34]:
```
df_prev.isnull().sum
```

Out[34]:
```
<bound method DataFrame.sum of          SK_ID_PREV  SK_ID_CURR  NAME_CONTRACT_TYPE  AMT_ANNUITY  \
0             False       False               False        False   
1             False       False               False        False   
2             False       False               False        False   
3             False       False               False        False   
4             False       False               False        False   
...             ...         ...                 ...          ...   
1670209       False       False               False        False   
1670210       False       False               False        False   
1670211       False       False               False        False   
1670212       False       False               False        False   
1670213       False       False               False        False   

         AMT_APPLICATION  AMT_CREDIT  AMT_DOWN_PAYMENT  AMT_GOODS_PRICE  \
0                  False       False             False            False   
1                  False       False              True            False   
2                  False       False              True            False   
3                  False       False              True            False   
4                  False       False              True            False   
...                  ...         ...               ...              ...   
1670209            False       False             False            False   
1670210            False       False             False            False   
1670211            False       False             False            False   
1670212            False       False              True            False   
1670213            False       False              True            False   

         WEEKDAY_APPR_PROCESS_START  HOUR_APPR_PROCESS_START  ...  \
0                             False                    False  ...   
1                             False                    False  ...   
2                             False                    False  ...   
3                             False                    False  ...   
4                             False                    False  ...   
...                             ...                      ...  ...   
1670209                       False                    False  ...   
1670210                       False                    False  ...   
1670211                       False                    False  ...   
1670212                       False                    False  ...   
1670213                       False                    False  ...   

         NAME_SELLER_INDUSTRY  CNT_PAYMENT  NAME_YIELD_GROUP  \
0                       False        False             False   
1                       False        False             False   
2                       False        False             False   
3                       False        False             False   
4                       False        False             False   
...                       ...          ...               ...   
1670209                 False        False             False   
1670210                 False        False             False   
1670211                 False        False             False   
1670212                 False        False             False   
1670213                 False        False             False   

         PRODUCT_COMBINATION  DAYS_FIRST_DRAWING  DAYS_FIRST_DUE  \
0                      False               False           False   
```

```
1                      False               False           False
2                      False               False           False
3                      False               False           False
4                      False                True            True
...                      ...                 ...             ...
1670209                False               False           False
1670210                False               False           False
1670211                False               False           False
1670212                False               False           False
1670213                False               False           False

         DAYS_LAST_DUE_1ST_VERSION  DAYS_LAST_DUE  DAYS_TERMINATION  \
0                            False          False             False
1                            False          False             False
2                            False          False             False
3                            False          False             False
4                             True           True              True
...                            ...            ...               ...
1670209                      False          False             False
1670210                      False          False             False
1670211                      False          False             False
1670212                      False          False             False
1670213                      False          False             False

         NFLAG_INSURED_ON_APPROVAL
0                            False
1                            False
2                            False
3                            False
4                             True
...                            ...
1670209                      False
1670210                      False
1670211                      False
1670212                      False
1670213                      False

[1670214 rows x 37 columns]>
```

In [35]:
```
missing_percentage = df_prev.isnull().mean() * 100
missing_percentage
```

Out[35]:
```
SK_ID_PREV                      0.000000
SK_ID_CURR                      0.000000
NAME_CONTRACT_TYPE              0.000000
AMT_ANNUITY                    22.286665
AMT_APPLICATION                 0.000000
AMT_CREDIT                      0.000060
AMT_DOWN_PAYMENT               53.636480
AMT_GOODS_PRICE                23.081773
WEEKDAY_APPR_PROCESS_START      0.000000
HOUR_APPR_PROCESS_START         0.000000
FLAG_LAST_APPL_PER_CONTRACT     0.000000
NFLAG_LAST_APPL_IN_DAY          0.000000
RATE_DOWN_PAYMENT              53.636480
RATE_INTEREST_PRIMARY          99.643698
RATE_INTEREST_PRIVILEGED       99.643698
NAME_CASH_LOAN_PURPOSE          0.000000
NAME_CONTRACT_STATUS            0.000000
DAYS_DECISION                   0.000000
NAME_PAYMENT_TYPE               0.000000
CODE_REJECT_REASON              0.000000
NAME_TYPE_SUITE                49.119754
NAME_CLIENT_TYPE                0.000000
NAME_GOODS_CATEGORY             0.000000
NAME_PORTFOLIO                  0.000000
NAME_PRODUCT_TYPE               0.000000
CHANNEL_TYPE                    0.000000
SELLERPLACE_AREA                0.000000
NAME_SELLER_INDUSTRY            0.000000
CNT_PAYMENT                    22.286366
NAME_YIELD_GROUP                0.000000
PRODUCT_COMBINATION             0.020716
DAYS_FIRST_DRAWING             40.298129
DAYS_FIRST_DUE                 40.298129
DAYS_LAST_DUE_1ST_VERSION      40.298129
DAYS_LAST_DUE                  40.298129
DAYS_TERMINATION               40.298129
NFLAG_INSURED_ON_APPROVAL      40.298129
dtype: float64
```

In [36]:
```
df_prev_cleaned = df_prev.dropna()
df_prev_cleaned
```

Out[36]:

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	WEEKDAY_A
598	2388655	414811	Consumer loans	14152.545	153387.0	138046.5	15340.5	153387.0	
21366	1184010	252161	Consumer loans	3136.275	29781.0	29781.0	0.0	29781.0	
24027	2144692	423348	Consumer loans	2640.195	26145.0	26014.5	2614.5	26145.0	
43927	2697394	178347	Consumer loans	10324.665	101002.5	101002.5	0.0	101002.5	
115115	2403906	268507	Consumer loans	13452.660	145800.0	131220.0	14580.0	145800.0	
...	...	...	...	...	...	...	...	...	
1603346	1928485	386819	Consumer loans	45418.500	562500.0	450000.0	112500.0	562500.0	
1619458	1347931	336203	Consumer loans	9207.180	113400.0	90720.0	22680.0	113400.0	
1644524	2002593	168701	Consumer loans	3518.460	38524.5	34668.0	3856.5	38524.5	
1645311	2396619	341729	Consumer loans	17179.380	171477.0	167571.0	17149.5	171477.0	
1663414	1328802	105065	Consumer loans	6357.375	68553.0	61695.0	6858.0	68553.0	

71 rows × 37 columns

In [37]:
```
missing_percentage = df_prev_cleaned.isnull().mean() * 100
missing_percentage
```

Out[37]:
```
SK_ID_PREV                     0.0
SK_ID_CURR                     0.0
NAME_CONTRACT_TYPE             0.0
AMT_ANNUITY                    0.0
AMT_APPLICATION                0.0
AMT_CREDIT                     0.0
AMT_DOWN_PAYMENT               0.0
AMT_GOODS_PRICE                0.0
WEEKDAY_APPR_PROCESS_START     0.0
HOUR_APPR_PROCESS_START        0.0
FLAG_LAST_APPL_PER_CONTRACT    0.0
NFLAG_LAST_APPL_IN_DAY         0.0
RATE_DOWN_PAYMENT              0.0
RATE_INTEREST_PRIMARY          0.0
RATE_INTEREST_PRIVILEGED       0.0
NAME_CASH_LOAN_PURPOSE         0.0
NAME_CONTRACT_STATUS           0.0
DAYS_DECISION                  0.0
NAME_PAYMENT_TYPE              0.0
CODE_REJECT_REASON             0.0
NAME_TYPE_SUITE                0.0
NAME_CLIENT_TYPE               0.0
NAME_GOODS_CATEGORY            0.0
NAME_PORTFOLIO                 0.0
NAME_PRODUCT_TYPE              0.0
CHANNEL_TYPE                   0.0
SELLERPLACE_AREA               0.0
NAME_SELLER_INDUSTRY           0.0
CNT_PAYMENT                    0.0
NAME_YIELD_GROUP               0.0
PRODUCT_COMBINATION            0.0
DAYS_FIRST_DRAWING             0.0
DAYS_FIRST_DUE                 0.0
DAYS_LAST_DUE_1ST_VERSION      0.0
DAYS_LAST_DUE                  0.0
DAYS_TERMINATION               0.0
NFLAG_INSURED_ON_APPROVAL      0.0
dtype: float64
```

In [38]:
```
duplicates = df_prev[df_prev.duplicated()]
duplicates
```

Out[38]:

SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	WEEKDAY_APPR_PR

0 rows × 37 columns

In [39]:
```
prev_data = df_prev_cleaned
prev_data
```

Out[39]:

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	WEEKDAY_A
598	2388655	414811	Consumer loans	14152.545	153387.0	138046.5	15340.5	153387.0	
21366	1184010	252161	Consumer loans	3136.275	29781.0	29781.0	0.0	29781.0	
24027	2144692	423348	Consumer loans	2640.195	26145.0	26014.5	2614.5	26145.0	
43927	2697394	178347	Consumer loans	10324.665	101002.5	101002.5	0.0	101002.5	
115115	2403906	268507	Consumer loans	13452.660	145800.0	131220.0	14580.0	145800.0	
...	...	...	...	...	...	...	...	...	
1603346	1928485	386819	Consumer loans	45418.500	562500.0	450000.0	112500.0	562500.0	
1619458	1347931	336203	Consumer loans	9207.180	113400.0	90720.0	22680.0	113400.0	
1644524	2002593	168701	Consumer loans	3518.460	38524.5	34668.0	3856.5	38524.5	
1645311	2396619	341729	Consumer loans	17179.380	171477.0	167571.0	17149.5	171477.0	
1663414	1328802	105065	Consumer loans	6357.375	68553.0	61695.0	6858.0	68553.0	

71 rows × 37 columns

In [40]:
```
prev_data.head()
```

Out[40]:

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	WEEKDAY_AI
598	2388655	414811	Consumer loans	14152.545	153387.0	138046.5	15340.5	153387.0	

21366	1184010	252161	Consumer loans	3136.275	29781.0	29781.0	0.0	29781.0
24027	2144692	423348	Consumer loans	2640.195	26145.0	26014.5	2614.5	26145.0
43927	2697394	178347	Consumer loans	10324.665	101002.5	101002.5	0.0	101002.5
115115	2403906	268507	Consumer loans	13452.660	145800.0	131220.0	14580.0	145800.0

5 rows × 37 columns

In [41]:
```
prev_data.tail()
```

Out[41]:

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	WEEKDAY_A
1603346	1928485	386819	Consumer loans	45418.500	562500.0	450000.0	112500.0	562500.0	
1619458	1347931	336203	Consumer loans	9207.180	113400.0	90720.0	22680.0	113400.0	
1644524	2002593	168701	Consumer loans	3518.460	38524.5	34668.0	3856.5	38524.5	
1645311	2396619	341729	Consumer loans	17179.380	171477.0	167571.0	17149.5	171477.0	
1663414	1328802	105065	Consumer loans	6357.375	68553.0	61695.0	6858.0	68553.0	

5 rows × 37 columns

In [42]:
```
prev_data.shape
```

Out[42]:
```
(71, 37)
```

Data analysis & visualization

In [44]:
```
# unique values in gender & repayment status
print(app_data['CODE_GENDER'].unique())
```

```
['M' 'F']
```

In [45]:
```
print(app_data['TARGET'].unique())
```

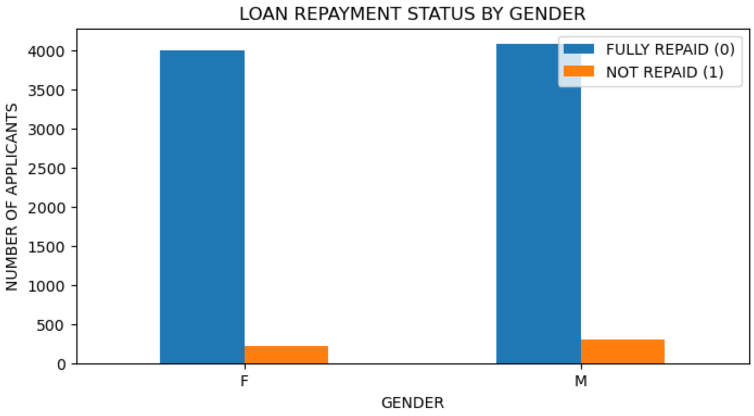
```
[0 1]
```

In [46]:
```
# count of loan repayment status by gender

loan_status_by_gender = app_data.groupby(['CODE_GENDER','TARGET']).size().unstack()
loan_status_by_gender
```

Out[46]:

TARGET	0	1
CODE_GENDER		
F	3997	224
M	4079	302

In [47]:
```
loan_status_by_gender.plot(kind = 'bar', figsize = (8,4), stacked = False)

plt.title('LOAN REPAYMENT STATUS BY GENDER')
plt.xlabel('GENDER')
plt.ylabel('NUMBER OF APPLICANTS')
plt.legend(['FULLY REPAID (0)', 'NOT REPAID (1)'])
plt.xticks(rotation = 0)

plt.show()
```



In [48]:
```
loan_status_by_gender_percentage = loan_status_by_gender.div(loan_status_by_gender.sum(axis=1),axis=0) * 100
loan_status_by_gender_percentage
```

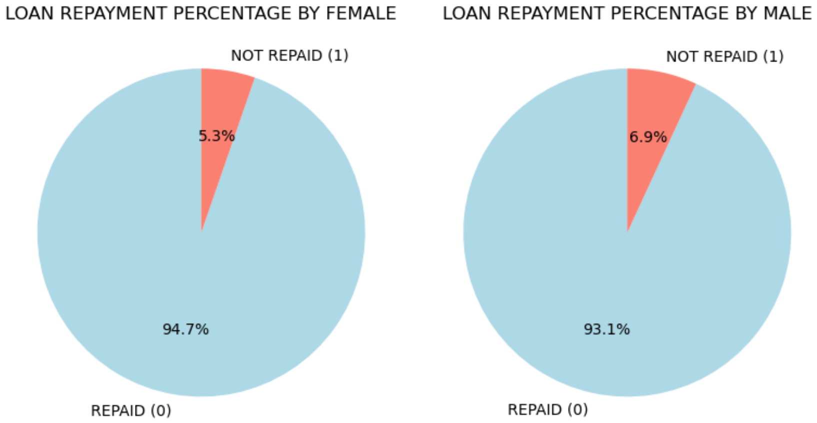
Out[48]:

TARGET	0	1
CODE_GENDER		
F	94.693201	5.306799
M	93.106597	6.893403

In [49]:
```
fig, axes = plt.subplots(1,2 ,figsize = (8,6))
colors = ['lightblue', 'salmon']

 # for Female applicants
axes[0].pie(loan_status_by_gender_percentage.loc['F'],labels = ['REPAID (0)', 'NOT REPAID (1)'], autopct = '%1.1f%%',startangle = 90, colors = colors)
```

Loading [MathJax]/extensions/Safe.js

```
axes[0].set_title('LOAN REPAYMENT PERCENTAGE BY FEMALE')

# for Male applicants
axes[1].pie(loan_status_by_gender_percentage.loc['M'],labels = ['REPAID (0)', 'NOT REPAID (1)'], autopct = '%1.1f%%',startangle = 90, colors = colors)
axes[1].set_title('LOAN REPAYMENT PERCENTAGE BY MALE')

plt.tight_layout()
plt.show()
```



In [50]:
```
# Loan repayment statue by age category

print(app_data[['DAYS_BIRTH','TARGET']].describe())
```

```
         DAYS_BIRTH       TARGET
count   8602.000000  8602.000000
mean  -14189.009416     0.061149
std     3259.202657     0.239617
min   -24835.000000     0.000000
25%   -16299.750000     0.000000
50%   -13883.500000     0.000000
75%   -11664.500000     0.000000
max    -7715.000000     1.000000
```

In [51]:
```
app_data['AGE_YEARS'] = ( -app_data['DAYS_BIRTH']) //365
app_data['AGE_YEARS']
```

```
C:\Users\dell\AppData\Local\Temp\ipykernel_5616\2804735630.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  app_data['AGE_YEARS'] = ( -app_data['DAYS_BIRTH']) //365
```

Out[51]:
```
71        42
124       44
152       31
161       38
255       31
          ..
307358    41
307359    38
307407    31
307456    55
307482    38
Name: AGE_YEARS, Length: 8602, dtype: int64
```

In [52]:
```
# define age categories

bins = [20,30,40,50,60,70]
labels = ['20-30','31-40','41-50','51-60','61-70']

app_data['AGE_GROUP'] = pd.cut(app_data['AGE_YEARS'],bins = bins, labels = labels,right = False)
app_data['AGE_GROUP']

app_data[['AGE_YEARS','AGE_GROUP','TARGET']].head()
```

```
C:\Users\dell\AppData\Local\Temp\ipykernel_5616\2981949013.py:6: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  app_data['AGE_GROUP'] = pd.cut(app_data['AGE_YEARS'],bins = bins, labels = labels,right = False)
```

Out[52]:

	AGE_YEARS	AGE_GROUP	TARGET
71	42	41-50	0
124	44	41-50	0
152	31	31-40	0
161	38	31-40	0
255	31	31-40	1

In [53]:
```
loan_status_by_age = app_data.groupby(['AGE_GROUP','TARGET']).size().unstack()
loan_status_by_age
```

```
C:\Users\dell\AppData\Local\Temp\ipykernel_5616\964034161.py:1: FutureWarning: The default of observed=False is deprecated and will be changed to True in a fut
ure version of pandas. Pass observed=False to retain current behavior or observed=True to adopt the future default and silence this warning.
  loan_status_by_age = app_data.groupby(['AGE_GROUP','TARGET']).size().unstack()
```

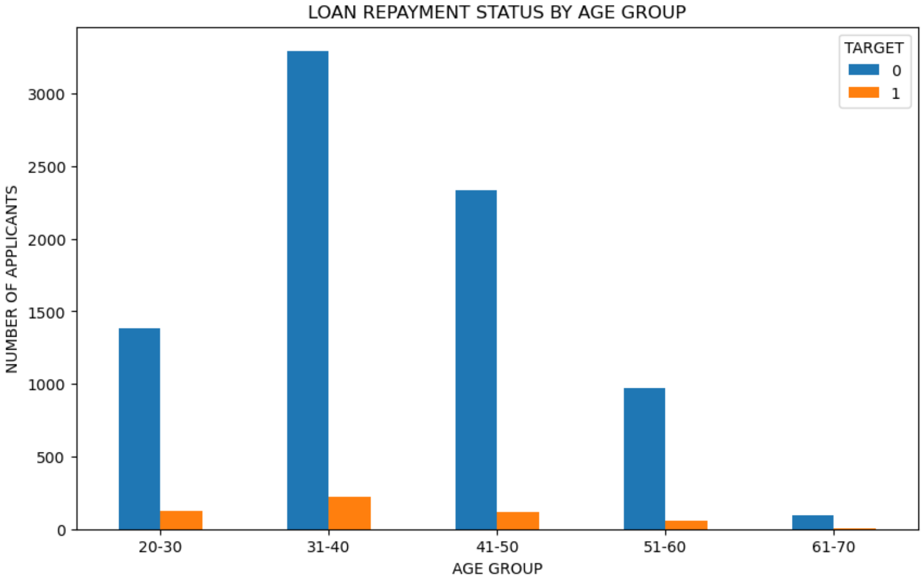
Loading [MathJax]/extensions/Safe.js

Out[53]:

TARGET AGE_GROUP	0	1
20-30	1382	123
31-40	3295	224
41-50	2336	119
51-60	969	55
61-70	94	5

In [54]:

```
loan_status_by_age.plot(kind = 'bar', stacked = False, figsize = (10,6))
plt.title('LOAN REPAYMENT STATUS BY AGE GROUP')
plt.xlabel('AGE GROUP')
plt.ylabel('NUMBER OF APPLICANTS')
plt.xticks(rotation = 0)

plt.show()
```



In [55]:

```
# Loan repayment by educational level

# unique education level & repayment status
print(app_data['NAME_EDUCATION_TYPE'].unique())
print(app_data['TARGET'].unique())
```

```
['Secondary / secondary special' 'Higher education' 'Incomplete higher'
 'Lower secondary' 'Academic degree']
[0 1]
```

In [56]:

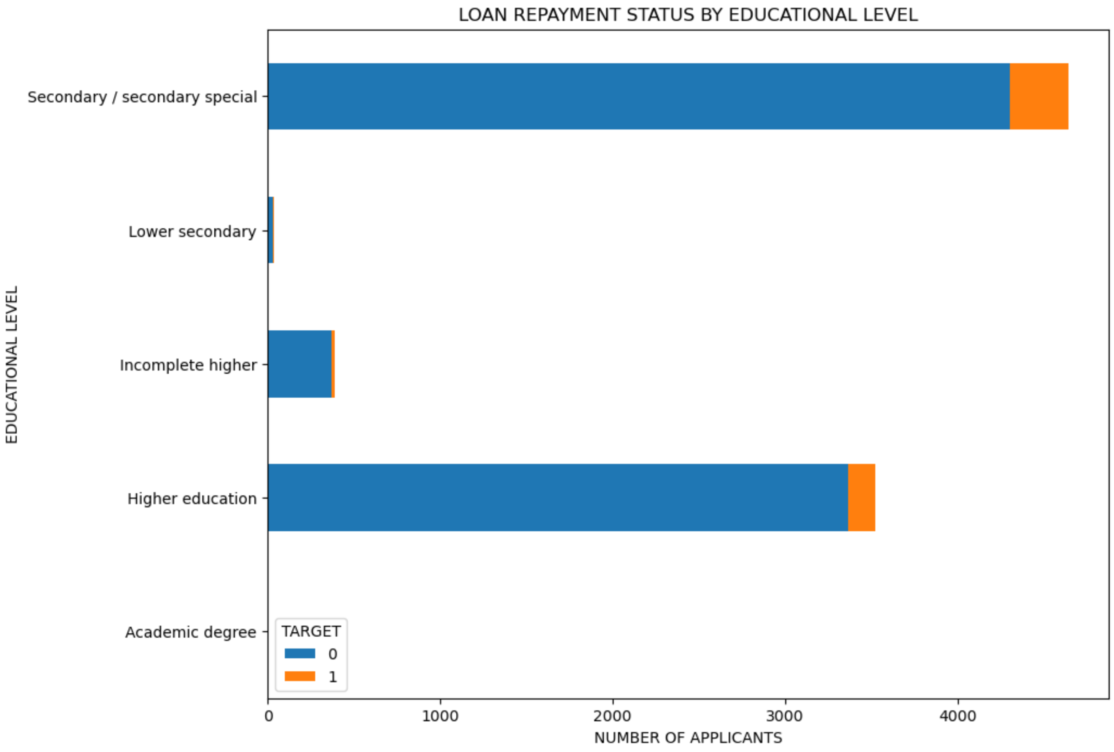
```
loan_status_by_edu = app_data.groupby(['NAME_EDUCATION_TYPE', 'TARGET']).size().unstack()
loan_status_by_edu
```

Out[56]:

TARGET NAME_EDUCATION_TYPE	0	1
Academic degree	6.0	NaN
Higher education	3364.0	159.0
Incomplete higher	370.0	22.0
Lower secondary	32.0	3.0
Secondary / secondary special	4304.0	342.0

In [57]:

```
loan_status_by_edu.plot(kind = 'barh', figsize =(10,8),stacked = True)

plt.title('LOAN REPAYMENT STATUS BY EDUCATIONAL LEVEL')
plt.ylabel('EDUCATIONAL LEVEL')
plt.xlabel('NUMBER OF APPLICANTS')
plt.xticks(rotation = 0)
plt.show()
```



In [58]:

```
# Loan repayment status by housing type

print(app_data['NAME_HOUSING_TYPE'].unique())
print(app_data['TARGET'].unique())
```

```
['House / apartment' 'With parents' 'Municipal apartment'
 'Office apartment' 'Co-op apartment' 'Rented apartment']
[0 1]
```

In [59]:

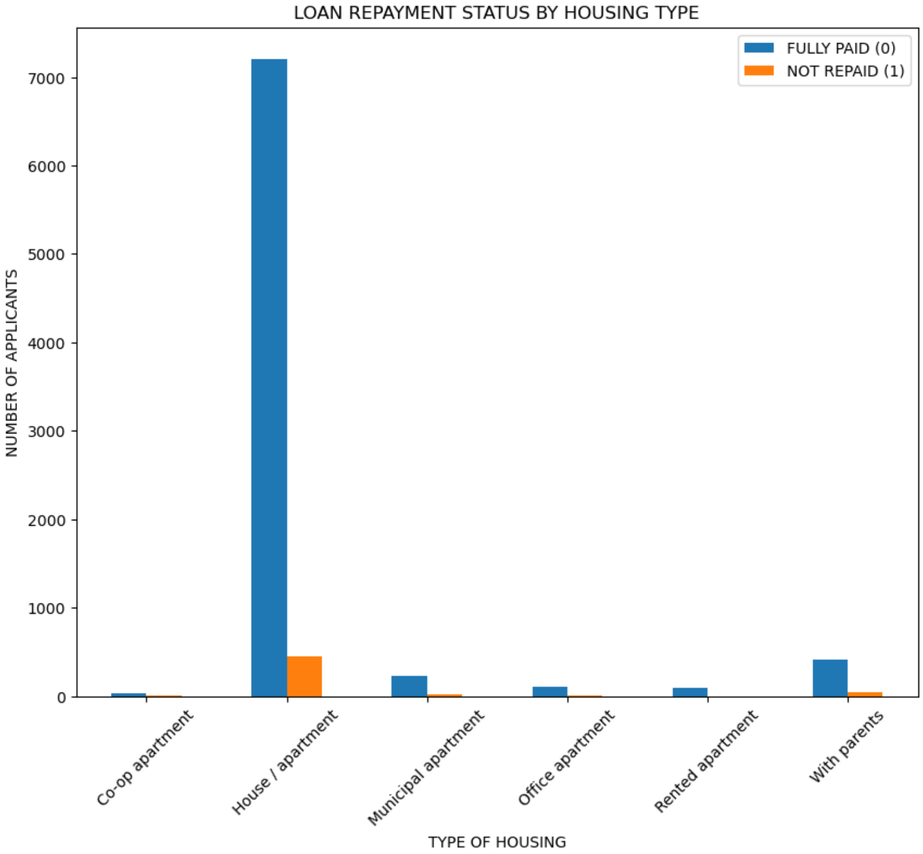
```
loan_status_by_housing = app_data.groupby(['NAME_HOUSING_TYPE','TARGET']).size().unstack()
loan_status_by_housing
```

Out[59]:

TARGET	0	1
NAME_HOUSING_TYPE		
Co-op apartment	31	4
House / apartment	7201	447
Municipal apartment	233	23
Office apartment	105	5
Rented apartment	95	3
With parents	411	44

In [60]:

```
loan_status_by_housing.plot(kind = 'bar', figsize = (10,8))
plt.title('LOAN REPAYMENT STATUS BY HOUSING TYPE')
plt.xlabel('TYPE OF HOUSING')
plt.ylabel('NUMBER OF APPLICANTS')
plt.legend(['FULLY PAID (0)', 'NOT REPAID (1)'])
plt.xticks(rotation = 45)
plt.show()
```



In [61]:

```
# Loan repayment type by income level

print(app_data['NAME_INCOME_TYPE'].unique())
print(app_data['TARGET'].unique())
```

```
['Working' 'Commercial associate' 'State servant']
[0 1]
```

In [62]:

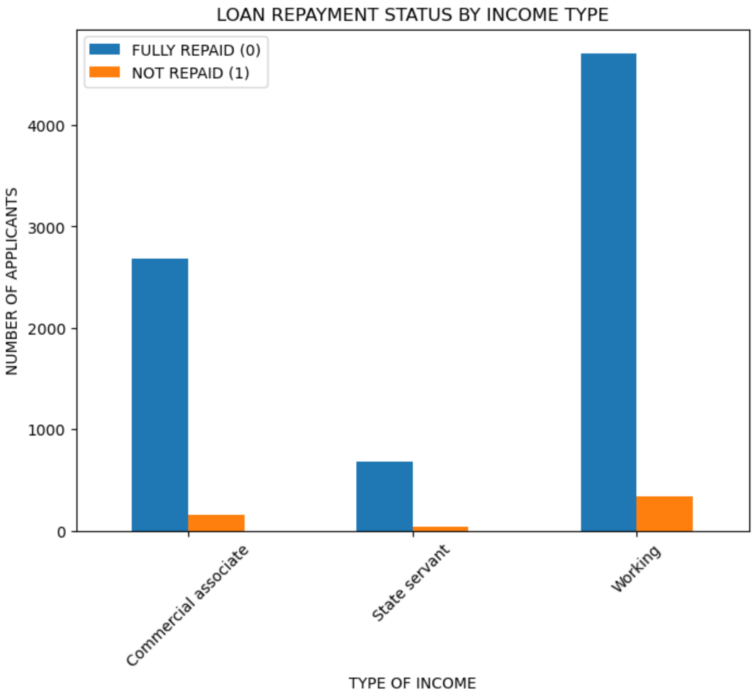
```
loan_status_by_income = app_data.groupby(['NAME_INCOME_TYPE','TARGET']).size().unstack()
loan_status_by_income
```

Out[62]:

TARGET	**0**	**1**
NAME_INCOME_TYPE		
Commercial associate	2683	153
State servant	680	36
Working	4713	337

In [63]:

```
loan_status_by_income.plot(kind = 'bar',figsize = (8,6))
plt.title('LOAN REPAYMENT STATUS BY INCOME TYPE')
plt.xlabel('TYPE OF INCOME')
plt.ylabel('NUMBER OF APPLICANTS')
plt.xticks(rotation = 45)
plt.legend(['FULLY REPAID (0)', 'NOT REPAID (1)'])

plt.show()
```



In [64]:

```
# Loan repayment status by family status

print(app_data['NAME_FAMILY_STATUS'].unique())
print(app_data['TARGET'].unique())
```

```
['Married' 'Separated' 'Single / not married' 'Widow' 'Civil marriage']
[0 1]
```

In [65]:

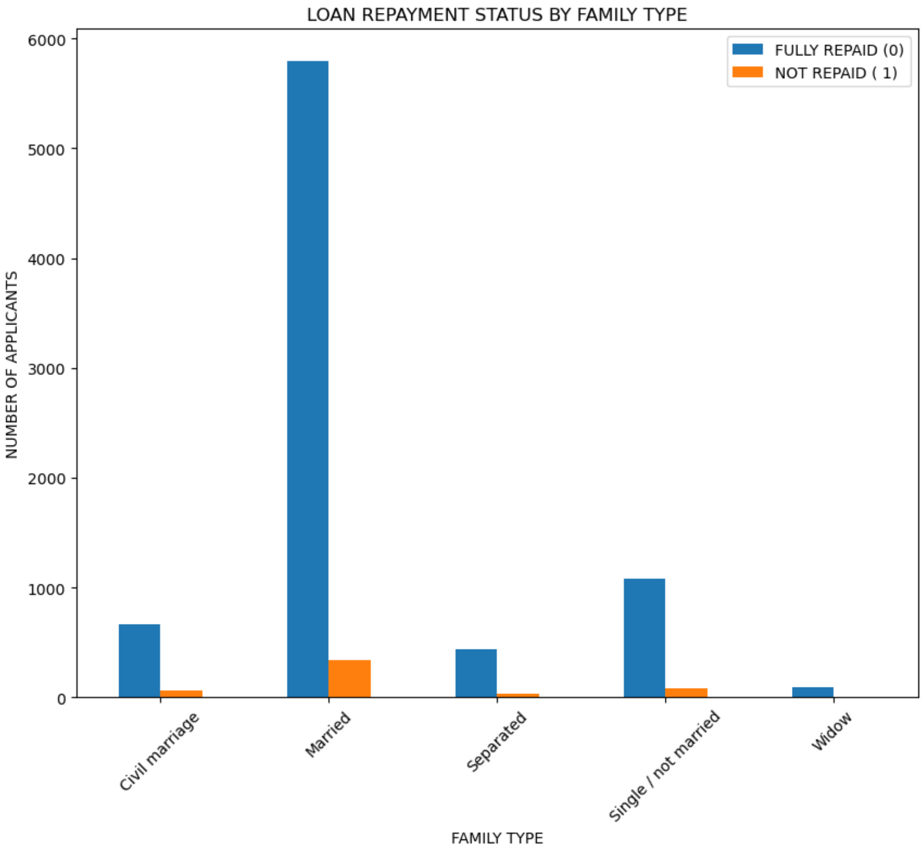
```
loan_status_by_family = app_data.groupby(['NAME_FAMILY_STATUS','TARGET']).size().unstack()
loan_status_by_family
```

Out[65]:

TARGET	0	1
NAME_FAMILY_STATUS		
Civil marriage	669	64
Married	5796	343
Separated	437	33
Single / not married	1077	82
Widow	97	4

In [66]:

```
loan_status_by_family.plot(kind = 'bar', figsize = (10,8))
plt.title('LOAN REPAYMENT STATUS BY FAMILY TYPE')
plt.xlabel('FAMILY TYPE')
plt.ylabel('NUMBER OF APPLICANTS')
plt.legend(['FULLY REPAID (0)','NOT REPAID ( 1)'])
plt.xticks(rotation = 45)

plt.show()
```



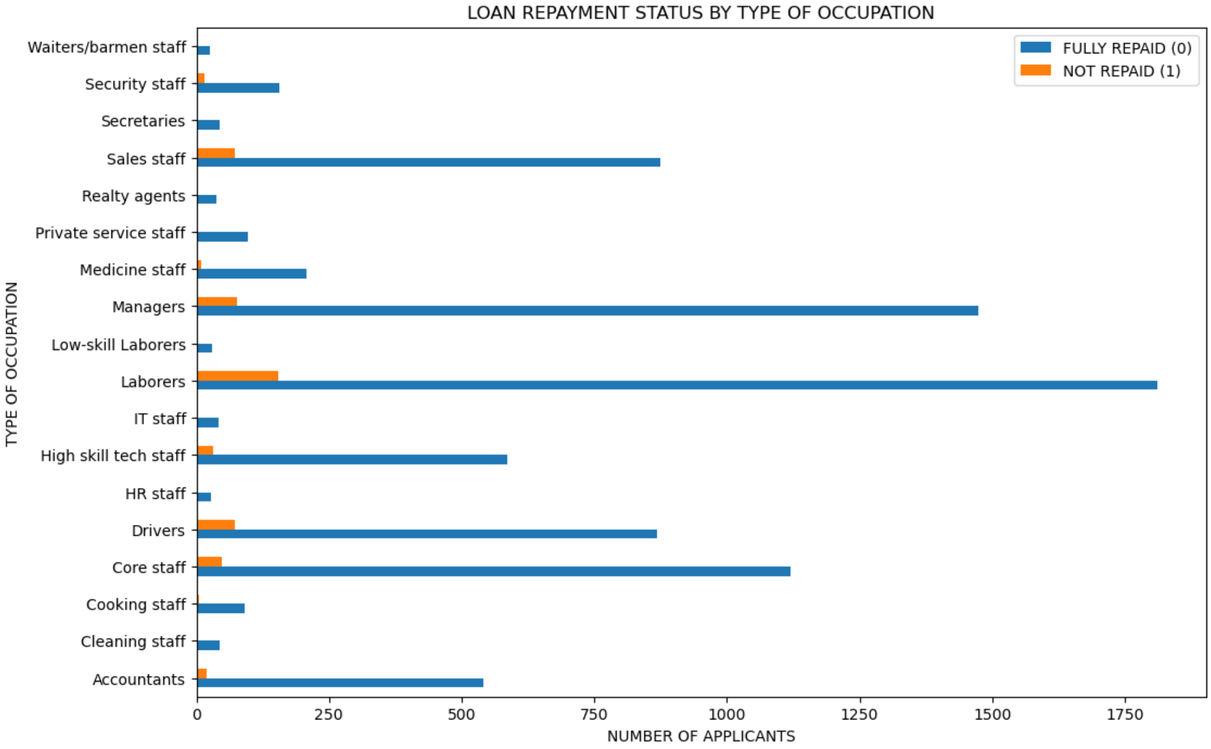
In [67]:

```
# Loan repayment status by occupation type

print(app_data['OCCUPATION_TYPE'].unique())
print(app_data['TARGET'].unique())
```

```
['Laborers' 'Managers' 'Drivers' 'Core staff' 'Sales staff'
 'High skill tech staff' 'Medicine staff' 'Accountants'
 'Private service staff' 'Cooking staff' 'HR staff' 'Cleaning staff'
 'Security staff' 'Secretaries' 'IT staff' 'Realty agents'
 'Waiters/barmen staff' 'Low-skill Laborers']
[0 1]
```

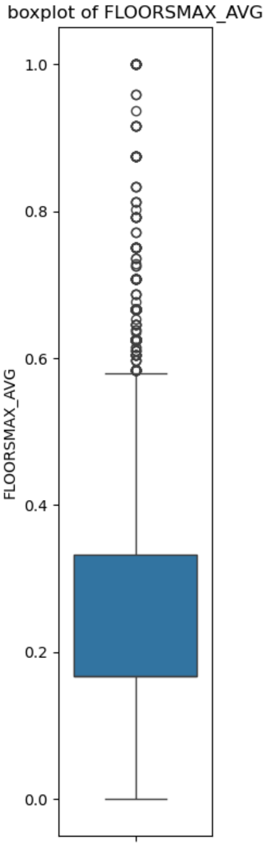
In [68]:

```
loan_status_by_occupation = app_data.groupby(['OCCUPATION_TYPE','TARGET']).size().unstack()

loan_status_by_occupation.plot(kind = 'barh', figsize = (12,8))
plt.title('LOAN REPAYMENT STATUS BY TYPE OF OCCUPATION')
plt.ylabel('TYPE OF OCCUPATION')
plt.xlabel('NUMBER OF APPLICANTS')
plt.legend(['FULLY REPAID (0)','NOT REPAID (1)'])
#plt.xticks(rotation = 45)

plt.show()
```

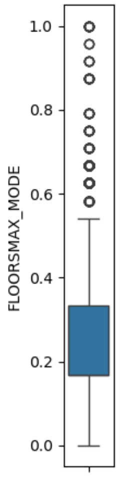


In [73]:

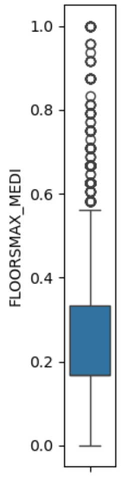
```
# Presence of outliers in the application_data
# Boxplot
boxplot_features = ['FLOORSMAX_AVG', 'FLOORSMAX_MODE', 'FLOORSMAX_MEDI','REGION_RATING_CLIENT','EXT_SOURCE_3','FLAG_DOCUMENT_3']

plt.figure(figsize = (10,8))
for i, feature in enumerate(boxplot_features,1):
    plt.subplot(1,6,i)
    sns.boxplot(y = app_data[feature])
    plt.title(f'boxplot of {feature}')
    plt.tight_layout()

    plt.show()
```



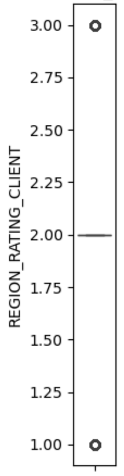
boxplot of FLOORSMAX_MODE

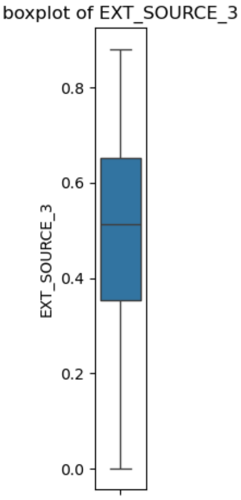


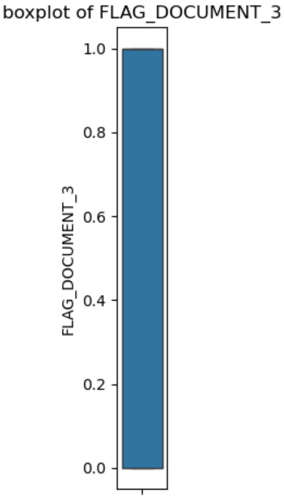
boxplot of FLOORSMAX_MEDI



boxplot of REGION_RATING_CLIENT





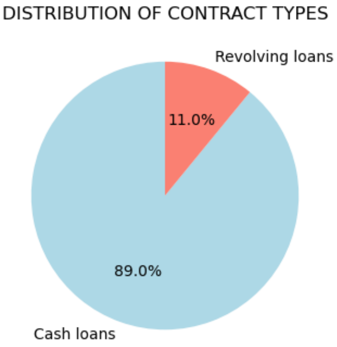


In [74]:
```
# Contract type distribution for current application
print(app_data['NAME_CONTRACT_TYPE'].unique())
contract_counts = app_data['NAME_CONTRACT_TYPE'].value_counts()
contract_counts
```

```
['Cash loans' 'Revolving loans']
```

Out[74]:
```
NAME_CONTRACT_TYPE
Cash loans         7660
Revolving loans     942
Name: count, dtype: int64
```

In [75]:
```
plt.figure(figsize = (6,4))
plt.pie(contract_counts, labels= contract_counts.index, autopct = '%1.1f%%', startangle = 90, colors = ['lightblue','salmon'])
plt.title('DISTRIBUTION OF CONTRACT TYPES')

plt.show()
```



In [148…
```
# Handling outliers in the application dataset
num_cols = app_data.select_dtypes(include = ['number'])
Q1 = num_cols.quantile(0.25)
Q3 = num_cols.quantile(0.75)
IQR = Q3 - Q1

lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

outliers = (num_cols<lower_bound) | (num_cols>upper_bound)
         outliers = app_data[outliers.any(axis = 1)]
```

```
app_data_outliers
```

Out[148…

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AM
71	100083	0	Cash loans	M	Y	Y	0	103500.0	573628.5	
124	100145	0	Cash loans	F	Y	Y	1	202500.0	260725.5	
152	100179	0	Cash loans	F	Y	N	0	202500.0	675000.0	
161	100190	0	Cash loans	M	Y	N	0	162000.0	263686.5	
255	100295	1	Cash loans	M	Y	N	1	225000.0	1019205.0	
...	...	...	...	...	...	...	...	...	...	
307358	456083	0	Cash loans	F	Y	Y	2	112500.0	361462.5	
307359	456084	0	Cash loans	F	Y	Y	1	99000.0	675000.0	
307407	456140	1	Cash loans	F	Y	Y	1	261000.0	711454.5	
307456	456195	0	Cash loans	F	Y	Y	0	94500.0	270000.0	
307482	456226	0	Cash loans	F	Y	Y	0	225000.0	500566.5	

7969 rows × 124 columns

In [150…

```
app_data1 = app_data_outliers
app_data1
```

Out[150…

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AM
71	100083	0	Cash loans	M	Y	Y	0	103500.0	573628.5	
124	100145	0	Cash loans	F	Y	Y	1	202500.0	260725.5	
152	100179	0	Cash loans	F	Y	N	0	202500.0	675000.0	
161	100190	0	Cash loans	M	Y	N	0	162000.0	263686.5	
255	100295	1	Cash loans	M	Y	N	1	225000.0	1019205.0	
...	...	...	...	...	...	...	...	...	...	
307358	456083	0	Cash loans	F	Y	Y	2	112500.0	361462.5	
307359	456084	0	Cash loans	F	Y	Y	1	99000.0	675000.0	
307407	456140	1	Cash loans	F	Y	Y	1	261000.0	711454.5	
307456	456195	0	Cash loans	F	Y	Y	0	94500.0	270000.0	
307482	456226	0	Cash loans	F	Y	Y	0	225000.0	500566.5	

7969 rows × 124 columns

In [154…

```
# Correlation between the variables
numeric_data = app_data1.select_dtypes(include = ['number'])
print(numeric_data.columns)
```

```
Index(['SK_ID_CURR', 'TARGET', 'CNT_CHILDREN', 'AMT_INCOME_TOTAL',
       'AMT_CREDIT', 'AMT_ANNUITY', 'AMT_GOODS_PRICE',
       'REGION_POPULATION_RELATIVE', 'DAYS_BIRTH', 'DAYS_EMPLOYED',
       ...
       'FLAG_DOCUMENT_19', 'FLAG_DOCUMENT_20', 'FLAG_DOCUMENT_21',
       'AMT_REQ_CREDIT_BUREAU_HOUR', 'AMT_REQ_CREDIT_BUREAU_DAY',
       'AMT_REQ_CREDIT_BUREAU_WEEK', 'AMT_REQ_CREDIT_BUREAU_MON',
       'AMT_REQ_CREDIT_BUREAU_QRT', 'AMT_REQ_CREDIT_BUREAU_YEAR', 'AGE_YEARS'],
      dtype='object', length=107)
```

In [208…

```
# correlation between selected variables - new features
new_features = ['AMT_CREDIT',
                'AMT_ANNUITY',
                'DAYS_EMPLOYED',
                'AMT_INCOME_TOTAL',
                'AGE_YEARS',
                'EXT_SOURCE_1',
                'CNT_CHILDREN'
                ]
new_features
```

Out[208…

```
['AMT_CREDIT',
 'AMT_ANNUITY',
 'DAYS_EMPLOYED',
 'AMT_INCOME_TOTAL',
 'AGE_YEARS',
 'EXT_SOURCE_1',
 'CNT_CHILDREN']
```

In [210…

```
corr_matrix = app_data1[new_features].corr()
corr_matrix
```

Out[210…

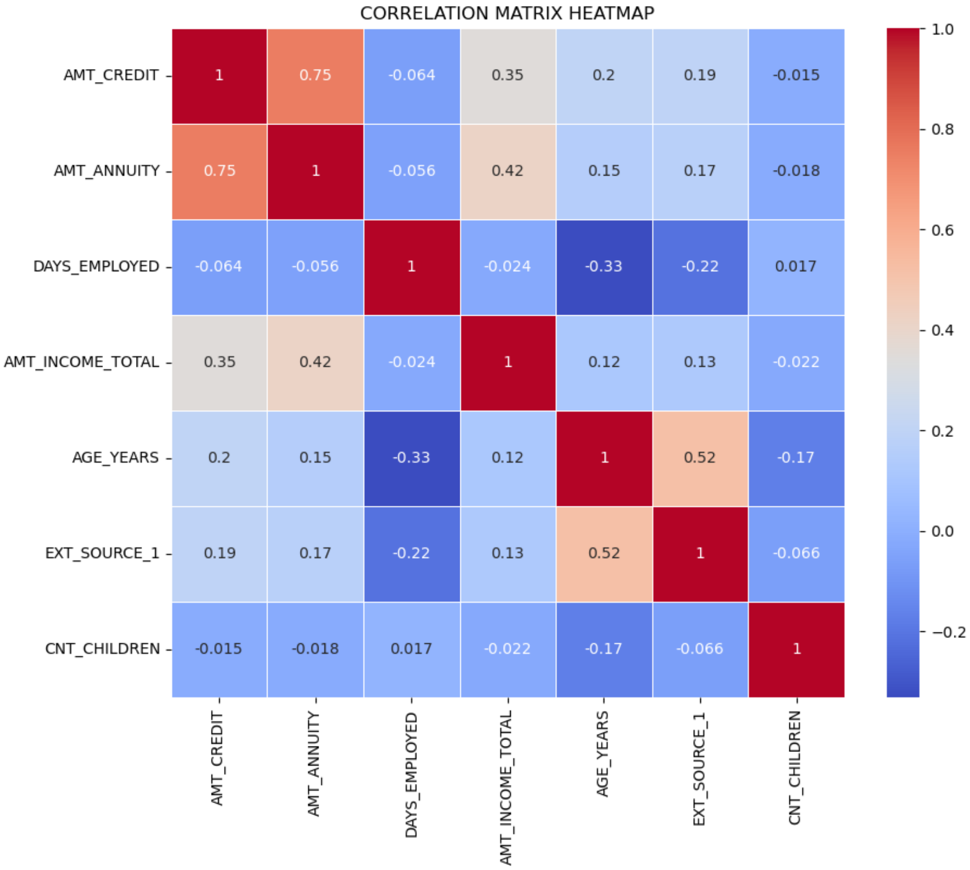
	AMT_CREDIT	AMT_ANNUITY	DAYS_EMPLOYED	AMT_INCOME_TOTAL	AGE_YEARS	EXT_SOURCE_1	CNT_CHILDREN
AMT_CREDIT	1.000000	0.750385	-0.064243	0.347891	0.200651	0.193545	-0.014793
AMT_ANNUITY	0.750385	1.000000	-0.055789	0.416533	0.153747	0.169014	-0.018006
DAYS_EMPLOYED	-0.064243	-0.055789	1.000000	-0.023574	-0.332295	-0.215603	0.016572
AMT_INCOME_TOTAL	0.347891	0.416533	-0.023574	1.000000	0.119941	0.130522	-0.022210
AGE_YEARS	0.200651	0.153747	-0.332295	0.119941	1.000000	0.516535	-0.173183
EXT_SOURCE_1	0.193545	0.169014	-0.215603	0.130522	0.516535	1.000000	-0.065797
CNT_CHILDREN	-0.014793	-0.018006	0.016572	-0.022210	-0.173183	-0.065797	1.000000

In [ ]:

Loading [MathJax]/extensions/Safe.js

In [212…

```
# visualizing the correlation matrix

plt.figure(figsize = (10,8))
sns.heatmap(corr_matrix, annot = True, cmap = 'coolwarm', linewidths = 0.5)
plt.title('CORRELATION MATRIX HEATMAP')

plt.show()
```



In [216…

```
# Regression analysis of the application data

features = ['AMT_ANNUITY', 'AMT_INCOME_TOTAL','EXT_SOURCE_1']
target = 'AMT_CREDIT'

app_data1_filtered = app_data1[features + [target]].dropna()
app_data1_filtered
```

Out[216…

	AMT_ANNUITY	AMT_INCOME_TOTAL	EXT_SOURCE_1	AMT_CREDIT
71	24435.0	103500.0	0.270766	573628.5
124	16789.5	202500.0	0.647045	260725.5
152	53329.5	202500.0	0.674832	675000.0
161	24781.5	162000.0	0.534999	263686.5
255	31032.0	225000.0	0.262005	1019205.0
...	...	...	...	...
307358	16051.5	112500.0	0.653115	361462.5
307359	21906.0	99000.0	0.383096	675000.0
307407	47673.0	261000.0	0.766549	711454.5
307456	15075.0	94500.0	0.823222	270000.0
307482	34969.5	225000.0	0.470808	500566.5

7969 rows × 4 columns

In [218…

```
x = app_data1_filtered[features]
y = app_data1_filtered[target]

print(x.head(), y.head())
```

```
     AMT_ANNUITY  AMT_INCOME_TOTAL  EXT_SOURCE_1
71       24435.0          103500.0      0.270766
124      16789.5          202500.0      0.647045
152      53329.5          202500.0      0.674832
161      24781.5          162000.0      0.534999
255      31032.0          225000.0      0.262005 71      573628.5
124     260725.5
152     675000.0
161     263686.5
255    1019205.0
Name: AMT_CREDIT, dtype: float64
```

In [220…

```python
# splitting data into traing & testing

x_train, x_test, y_train, y_test = train_test_split(x,y, test_size = 0.2, random_state = 42)

print(f'Training Set Size: {x_train.shape} , Testing Set Size: {x_test.shape}')
```

```
Training Set Size: (6375, 3) , Testing Set Size: (1594, 3)
```

In [226…

```python
regressor = LinearRegression()
regressor.fit(x_train, y_train)

# get model coefficients

print('Intercept:', regressor.intercept_)
print('Coefficients:', dict(zip(features, regressor.coef_)))
```

```
Intercept: -38881.41624700499
Coefficients: {'AMT_ANNUITY': 20.222233324442733, 'AMT_INCOME_TOTAL': 0.10444815223918028, 'EXT_SOURCE_1': 158357.8829596226}
```

In [228…

```python
# making predictions

y_pred = regressor.predict(x_test)

comparison = pd.DataFrame({'Actual': y_test, 'Predicted': y_pred})
print(comparison.head())
```

```
           Actual     Predicted
105177   450000.0  5.306111e+05
205113  1354500.0  8.317769e+05
272820   625536.0  7.916978e+05
252771  1113840.0  1.173895e+06
251202   225000.0  2.302246e+05
```

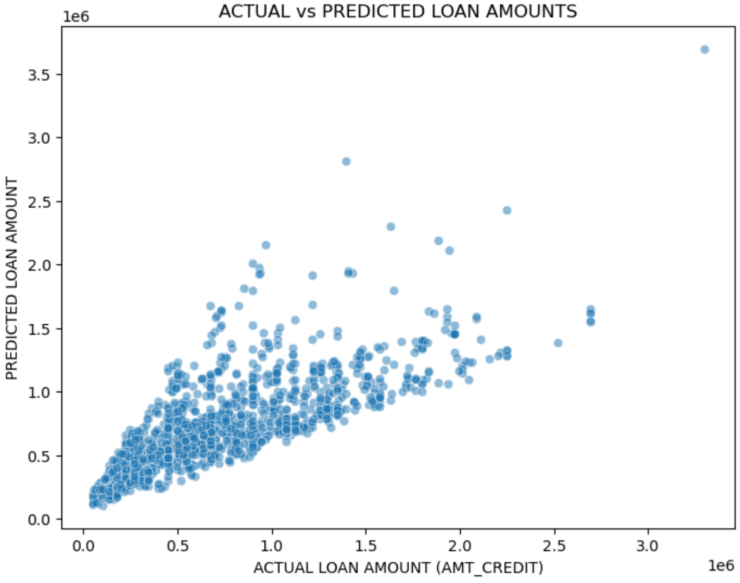
In [232…

```python
# calculate MSE & R2

mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test,y_pred)

print(f'Mean Squared Error: {mse:.2f}')
print(f'R-Squared: {r2:.4f}')
```

```
Mean Squared Error: 95726793472.44
R-Squared: 0.5776
```

In [238…

```python
# scatter plot for Actual vs Predicted values

plt.figure(figsize = (8,6))
sns.scatterplot(x=y_test, y=y_pred, alpha = 0.5)
plt.xlabel('ACTUAL LOAN AMOUNT (AMT_CREDIT)')
plt.ylabel('PREDICTED LOAN AMOUNT')
plt.title('ACTUAL vs PREDICTED LOAN AMOUNTS')

plt.show()
```



In [ ]: