LENDING CLUB

USE CASE ANALYSIS

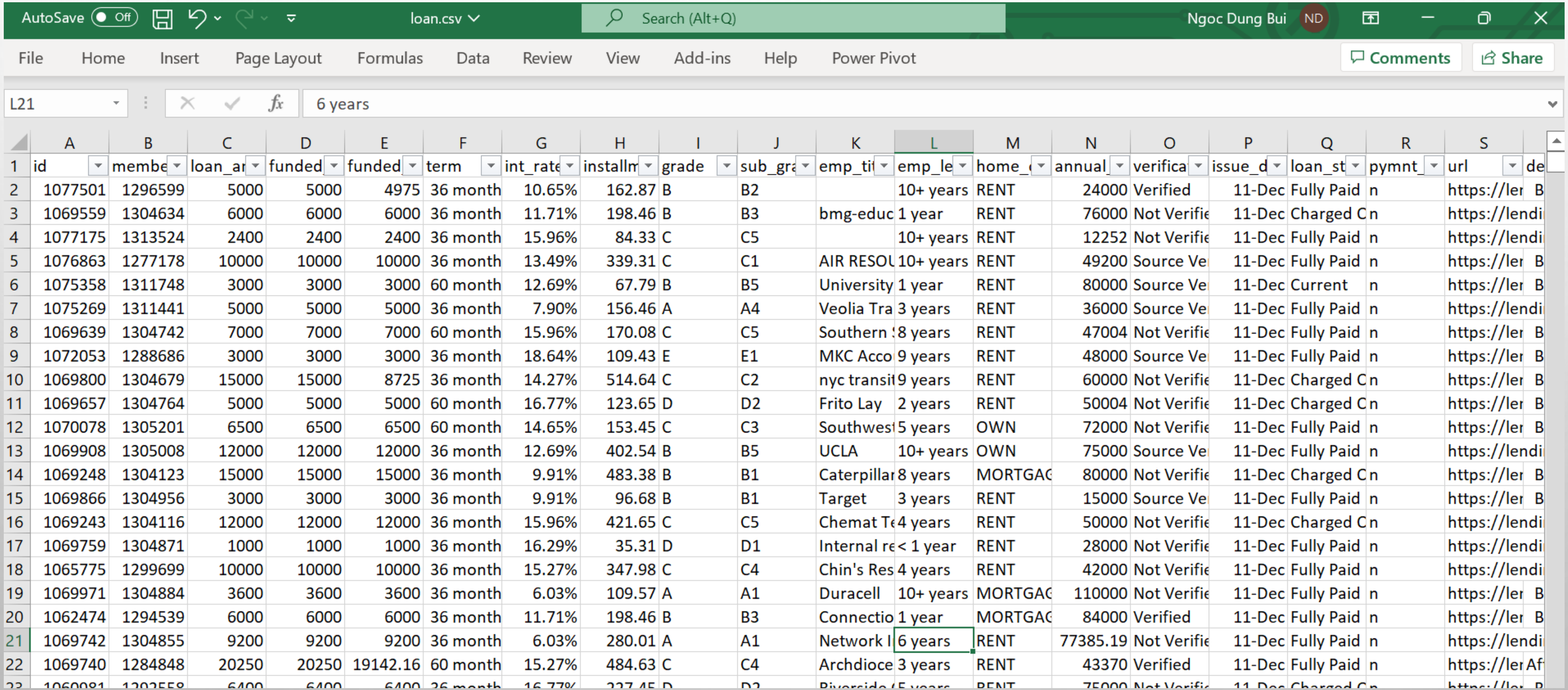
NGOC DUNG BUI

BUSINESS REQUIREMENTS

- CARRY OUT AN EDA (EXPLORATORY DATA ANALYSIS) ON A DATASET OF LOAN DATA TO IDENTIFY RISKY FACTORS THAT WOULD LEAD TO A FINANCIAL LOSS FOR THE LENDERS. BASED ON THAT, FINANCIAL INSTITUTIONS/LENDERS COULD TAKE ACTIONS SUCH AS DENYING THE LOAN, REDUCING THE AMOUNT OF LOAN, LENDING (TO RISKY APPLICANTS) AT A HIGHER INTEREST RATE, ETC

DATASET OVERVIEW

The dataset is stored in the format of a CSV file.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
1	id	membe	loan_ar	funded	funded	term	int_rate	installm	grade	sub_gra	emp_tit	emp_le	home_	annual_	verifica	issue_d	loan_st	pymnt_	url	de
2	1077501	1296599	5000	5000	4975	36 month	10.65%	162.87	B	B2		10+ years	RENT	24000	Verified	11-Dec	Fully Paid	n	https://ler	B
3	1069559	1304634	6000	6000	6000	36 month	11.71%	198.46	B	B3	bmg-educ	1 year	RENT	76000	Not Verifie	11-Dec	Charged C	n	https://lendi	
4	1077175	1313524	2400	2400	2400	36 month	15.96%	84.33	C	C5		10+ years	RENT	12252	Not Verifie	11-Dec	Fully Paid	n	https://lendi	
5	1076863	1277178	10000	10000	10000	36 month	13.49%	339.31	C	C1	AIR RESOU	10+ years	RENT	49200	Source Ve	11-Dec	Fully Paid	n	https://ler	B
6	1075358	1311748	3000	3000	3000	60 month	12.69%	67.79	B	B5	University	1 year	RENT	80000	Source Ve	11-Dec	Current	n	https://ler	B
7	1075269	1311441	5000	5000	5000	36 month	7.90%	156.46	A	A4	Veolia Tra	3 years	RENT	36000	Source Ve	11-Dec	Fully Paid	n	https://lendi	
8	1069639	1304742	7000	7000	7000	60 month	15.96%	170.08	C	C5	Southern !	8 years	RENT	47004	Not Verifie	11-Dec	Fully Paid	n	https://ler	B
9	1072053	1288686	3000	3000	3000	36 month	18.64%	109.43	E	E1	MKC Acco	9 years	RENT	48000	Source Ve	11-Dec	Fully Paid	n	https://ler	B
10	1069800	1304679	15000	15000	8725	36 month	14.27%	514.64	C	C2	nyc transit	9 years	RENT	60000	Not Verifie	11-Dec	Charged C	n	https://ler	B
11	1069657	1304764	5000	5000	5000	60 month	16.77%	123.65	D	D2	Frito Lay	2 years	RENT	50004	Not Verifie	11-Dec	Charged C	n	https://ler	B
12	1070078	1305201	6500	6500	6500	60 month	14.65%	153.45	C	C3	Southwest	5 years	OWN	72000	Not Verifie	11-Dec	Fully Paid	n	https://ler	B
13	1069908	1305008	12000	12000	12000	36 month	12.69%	402.54	B	B5	UCLA	10+ years	OWN	75000	Source Ve	11-Dec	Fully Paid	n	https://lendi	
14	1069248	1304123	15000	15000	15000	36 month	9.91%	483.38	B	B1	Caterpillar	8 years	MORTGAG	80000	Not Verifie	11-Dec	Charged C	n	https://ler	B
15	1069866	1304956	3000	3000	3000	36 month	9.91%	96.68	B	B1	Target	3 years	RENT	15000	Source Ve	11-Dec	Fully Paid	n	https://ler	B
16	1069243	1304116	12000	12000	12000	36 month	15.96%	421.65	C	C5	Chemat Te	4 years	RENT	50000	Not Verifie	11-Dec	Charged C	n	https://lendi	
17	1069759	1304871	1000	1000	1000	36 month	16.29%	35.31	D	D1	Internal re	< 1 year	RENT	28000	Not Verifie	11-Dec	Fully Paid	n	https://lendi	
18	1065775	1299699	10000	10000	10000	36 month	15.27%	347.98	C	C4	Chin's Res	4 years	RENT	42000	Not Verifie	11-Dec	Fully Paid	n	https://lendi	
19	1069971	1304884	3600	3600	3600	36 month	6.03%	109.57	A	A1	Duracell	10+ years	MORTGAG	110000	Not Verifie	11-Dec	Fully Paid	n	https://ler	B
20	1062474	1294539	6000	6000	6000	36 month	11.71%	198.46	B	B3	Connectio	1 year	MORTGAG	84000	Verified	11-Dec	Fully Paid	n	https://ler	B
21	1069742	1304855	9200	9200	9200	36 month	6.03%	280.01	A	A1	Network I	6 years	RENT	77385.19	Not Verifie	11-Dec	Fully Paid	n	https://lendi	
22	1069740	1284848	20250	20250	19142.16	60 month	15.27%	484.63	C	C4	Archdioce	3 years	RENT	43370	Verified	11-Dec	Fully Paid	n	https://ler	Af

THE DATASET CONTAINS 111 COLUMNS AND 39713 ROWS

5 rows × 111 columns

```
df_loan.describe()
```

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	installment	annual_inc	dti	delinq_2yrs	inq_last_6mths	...
count	3.971700e+04	3.971700e+04	39717.000000	39717.000000	39717.000000	39717.000000	3.971700e+04	39717.000000	39717.000000	39717.000000	...
mean	6.831319e+05	8.504636e+05	11219.443815	10947.713196	10397.448868	324.561922	6.896893e+04	13.315130	0.146512	0.869200	...
std	2.106941e+05	2.656783e+05	7456.670694	7187.238670	7128.450439	208.874874	6.379377e+04	6.678594	0.491812	1.070219	...
min	5.473400e+04	7.069900e+04	500.000000	500.000000	0.000000	15.690000	4.000000e+03	0.000000	0.000000	0.000000	...
25%	5.162210e+05	6.667800e+05	5500.000000	5400.000000	5000.000000	167.020000	4.040400e+04	8.170000	0.000000	0.000000	...
50%	6.656650e+05	8.508120e+05	10000.000000	9600.000000	8975.000000	280.220000	5.900000e+04	13.400000	0.000000	1.000000	...
75%	8.377550e+05	1.047339e+06	15000.000000	15000.000000	14400.000000	430.780000	8.230000e+04	18.600000	0.000000	1.000000	...
max	1.077501e+06	1.314167e+06	35000.000000	35000.000000	35000.000000	1305.190000	6.000000e+06	29.990000	11.000000	8.000000	...

8 rows × 87 columns

```
len(df_loan)
```

39717

STEP 1: CLEANSING DATA

Extract Loan dataframe to a sub-dataframe that contains columns in which there's at least 1 null value each

```
nulls_df = df_loan.loc[:, df_loan.isna().any()]
```

```
nulls_df.isna().sum()
```

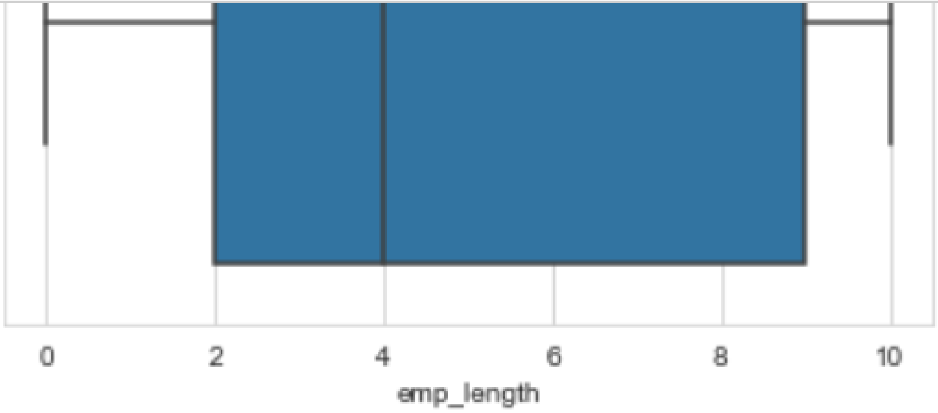
```
emp_title                      2459
emp_length                     1075
desc                          12940
title                            11
mths_since_last_delinq        25682
                              ...
tax_liens                        39
tot_hi_cred_lim               39717
total_bal_ex_mort             39717
total_bc_limit                39717
total_il_high_credit_limit    39717
Length: 68, dtype: int64
```

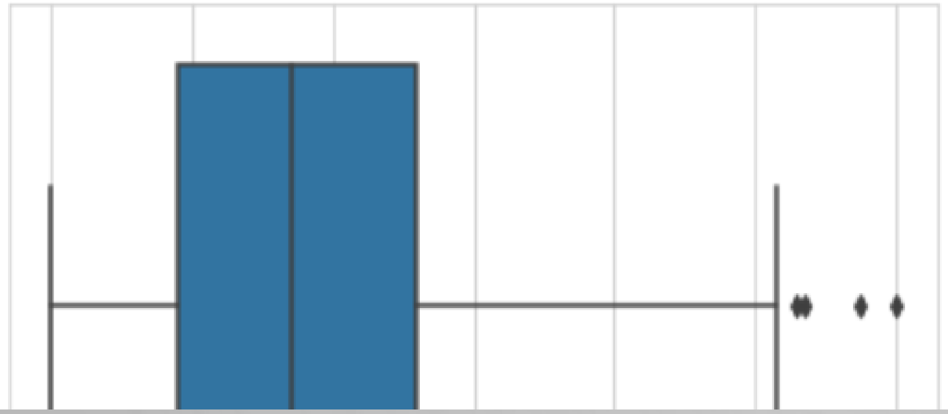
Columns in blue background should be converted to Date type and columns in orange should be converted to Numeric type

```
nulls_df.select_dtypes(include=["object"])
```

	emp_title	emp_length	desc	title	revol_util	last_pymnt_d	next_pymnt_d	last_credit_pull_d
0	NaN	10+ years	Borrower added on 12/22/11 > I need to upgra...	Computer	83.70%	Jan-15	NaN	May-16
1	Ryder	< 1 year	Borrower added on 12/22/11 > I plan to use t...	bike	9.40%	Apr-13	NaN	Sep-13
2	NaN	10+ years	NaN	real estate business	98.50%	Jun-14	NaN	May-16
3	AIR RESOURCES BOARD	10+ years	Borrower added on 12/21/11 > to pay for prop...	personel	21%	Jan-15	NaN	Apr-16
4	University Medical Group	1 year	Borrower added on 12/21/11 > I plan on combi...	Personal	53.90%	May-16	Jun-16	May-16
...	...	...	...	...	...	...	...	...
39712	FiSite Research	4 years	Our current gutter system on our home is old a...	Home Improvement	13.10%	Jul-10	NaN	Jun-10
39713	Squarewave Solutions, Ltd.	3 years	The rate of interest and fees incurred by carr...	Retiring credit card debt	26.90%	Jul-10	NaN	Jul-10

```python
#loop thru null df to visualize outliers of continuous variables
# remove some columns which contains only Null values
for column in nulls_df.columns:
    if df_loan[column].isna().sum() == len(df_loan.index):
        df_loan.drop(column, axis=1, inplace=True)
        continue
    if df_loan[column].dtype == 'float64':
        sns.boxplot(df_loan[pd.notna(df_loan[column])][column])
        plt.show()
```





Replace missing values in numeric columns with Median

```
for col in float_cols:
    if col in df_loan.columns:
        median = df_loan[col].median
        df_loan[col].fillna(median, inplace=True)
```

```
df_loan.isna().sum()
```

```
id                        0
member_id                 0
loan_amnt                 0
funded_amnt               0
funded_amnt_inv           0
term                      0
int_rate                  0
installment               0
grade                     0
sub_grade                 0
emp_title              2459
emp_length                0
home_ownership            0
annual_inc                0
verification_status       0
issue_d                   0
loan_status               0
pymnt_plan                0
```

Null values in some columns like emp_title, title, desc should be filled with a Description text like "Missing"

```
df_loan.emp_title.value_counts()
```

```
41]: US Army                                  134
     Bank of America                          109
     IBM                                       66
     AT&T                                      59
     Kaiser Permanente                         56
                                             ...
     Community College of Philadelphia          1
     AMEC                                       1
     lee county sheriff                         1
     Bacon County Board of Education            1
     Evergreen Center                           1
     Name: emp_title, Length: 28820, dtype: int64
```

```
42]: #replace Null values in emp_title with "missing"
     df_loan.emp_title.fillna("Missing", inplace=True)
```

```
43]: df_loan.emp_title.isna().sum()
```

```
43]: 0
```

Some constant features which contain only constants should be removed

Remove constant features, which contain a single value and do not bring a meaningful interpretation

```
print("Contant fields to be removed:")
for column in list(df_loan.columns):
    if df_loan[column].unique().size < 2:
        print(column)
        df_loan.drop(column, axis=1, inplace=True)
```

```
Contant fields to be removed:
pymnt_plan
initial_list_status
policy_code
application_type
```

Next step: Remove outliers which could be defined as a value that < (Q1 - 1.5 IQR) or > (Q3 + 1.5 IQR)

As the rule of thumb for outliers detection, any values that are not in the range (Q1 - 1.5 IQR) and (Q3 + 1.5 IQR) are considered as outliers and should be removed

```
Q1 = df_loan.quantile(0.25)
Q3 = df_loan.quantile(0.75)
IQR = Q3 - Q1
print(IQR)
```

```
id                          321534.00000
member_id                   380559.00000
loan_amnt                     9500.00000
funded_amnt                   9600.00000
funded_amnt_inv               9400.00000
int_rate                         5.34000
installment                    263.76000
emp_length                       7.00000
annual_inc                   41896.00000
dti                             10.43000
delinq_2yrs                      0.00000
inq_last_6mths                   1.00000
open_acc                         6.00000
pub_rec                          0.00000
revol_bal                    13355.00000
```

```
df_loan = df_loan[~((df_loan < (Q1 - 1.5 * IQR)) |(df_loan > (Q3 + 1.5 * IQR))).any(axis=1)]
```

Extract Charged off data into a separate DF for further analysis

Because the company wants to understand the driving factors (or driver variables) behind loan default, extract Charged off data into a separate DF for further analysis

```
grouped_df = df_loan.groupby('loan_status')
charged_off_df = grouped_df.get_group('Charged Off')
```

Extract verified/verified source data only

```
veri_grp_df = charged_off_df.groupby('verification_status')
source_verified = veri_grp_df.get_group('Source Verified')
verified = veri_grp_df.get_group('Verified')
```

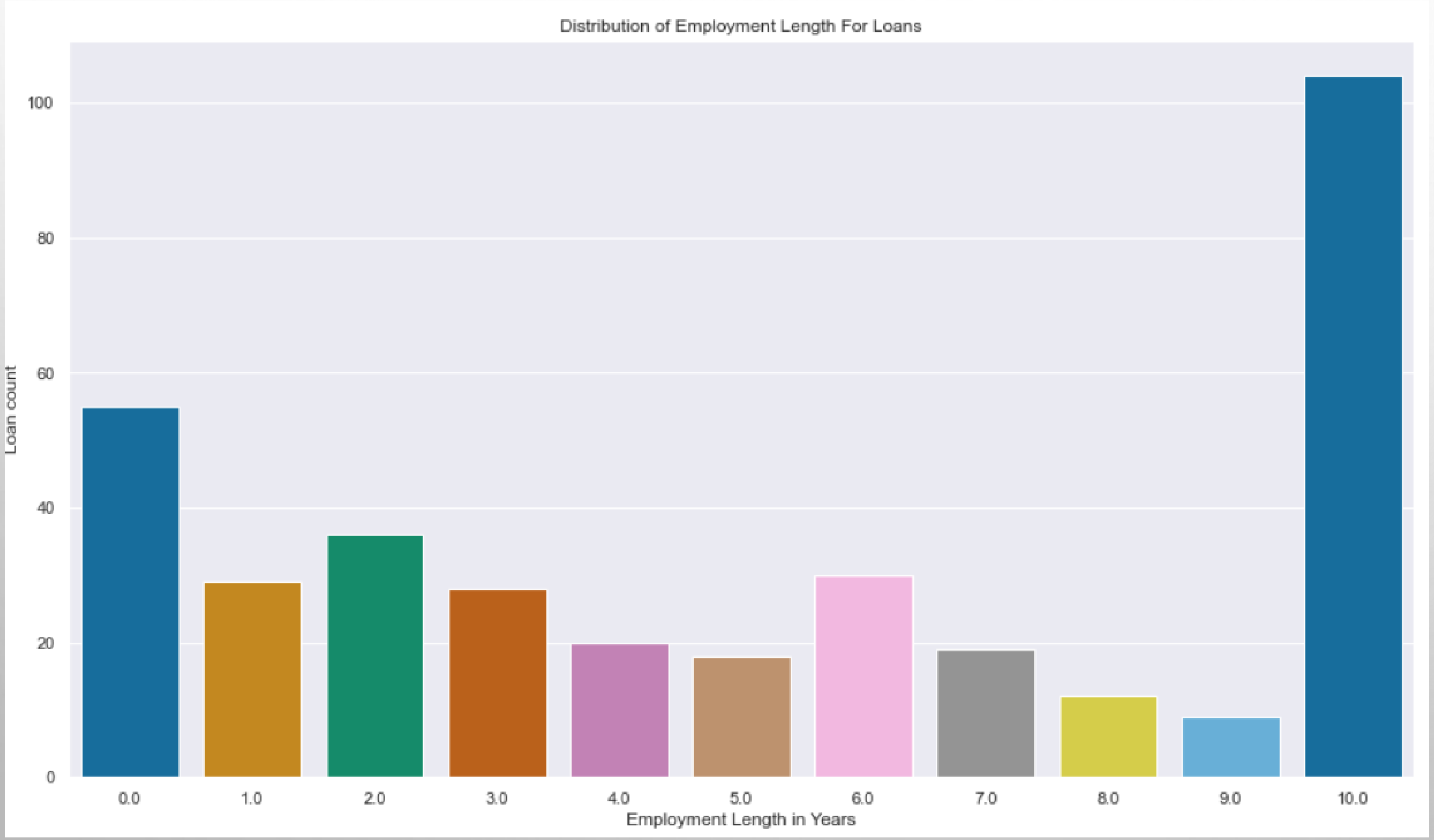
```
default_df = pd.concat([source_verified,verified])
```

```
default_df
```

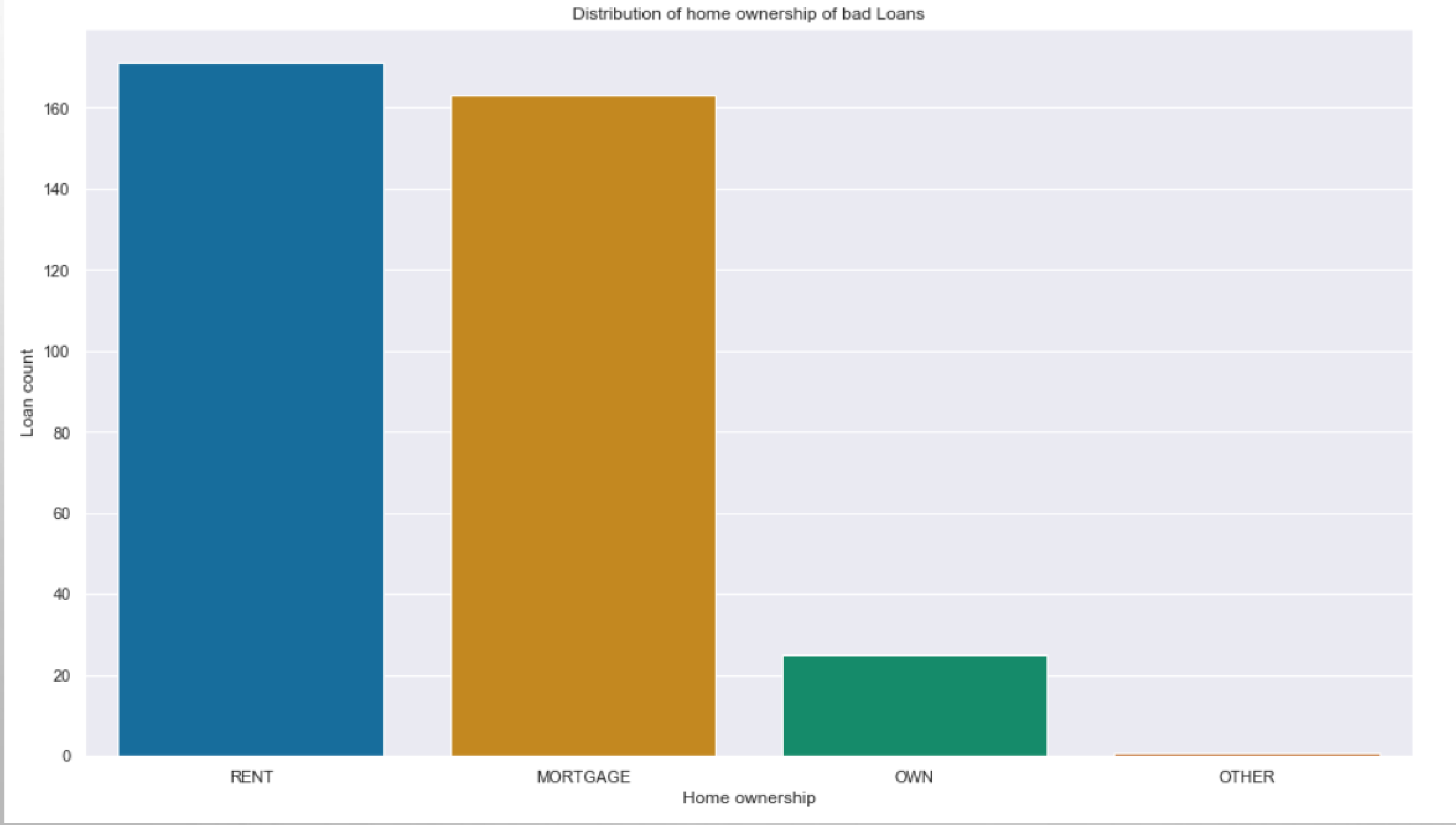
	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	...	total_acc	out_prncp	out_prncp_inv
204	1066835	1301027	10500	10500	10500.000000	36 months	16.29	370.66	D	D1	...	8	0.0	0.0
220	1066798	1300982	9500	9500	9500.000000	36 months	12.69	318.68	B	B5	...	42	0.0	0.0
239	1066344	1300716	15600	15600	15600.000000	60 months	12.69	352.48	B	B5	...	11	0.0	0.0
324	1065348	1299443	5000	5000	5000.000000	36 months	12.42	167.08	B	B4	...	40	0.0	0.0

STEP 2: EDA

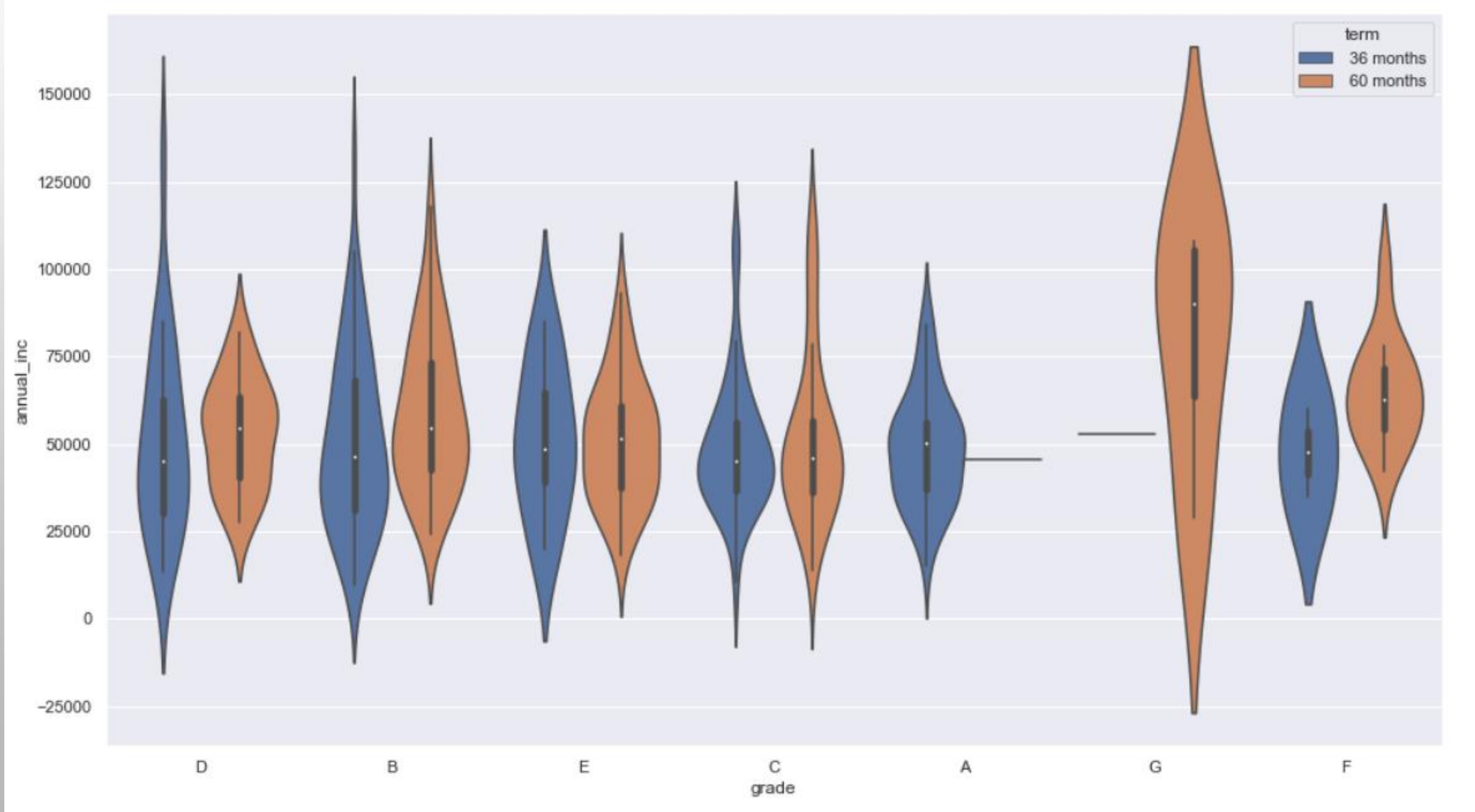
The current dataset showing that borrowers with 10+ years employments are the most who have a bad loan.



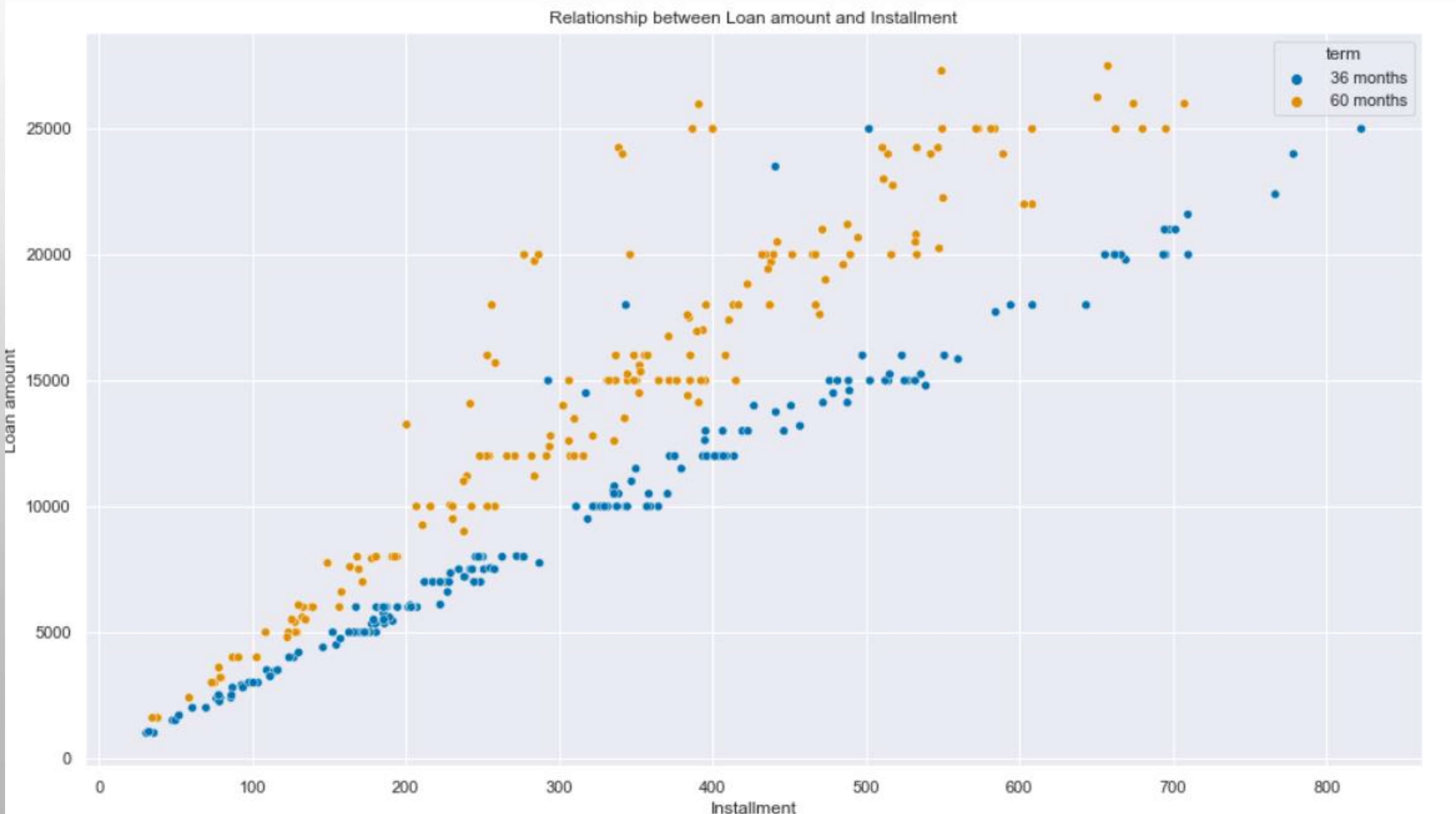
It's more likely to have a bad Loan if a borrower is a renter or having a mortgage instead of being the owner a home



The median of annual income is almost the same among different loan grades at about 50k, except that the median of grade G is greater than the rest, at about 85k, and the term of loan is 60 months



There's a strong relationship between Loan amount and monthly payment (installment) for bad Loans



The relationship between Annual income and Loan amount is weak. However, the density of the Loan amount is higher, if the term is shorter (36 months) and the loan amount is smaller (less than 15k)

The overall correlation among the dataset attributes showing that there're high correlations among some attributes of "loan amount", "funded amount", "installment"