Reading dataset

Loading loan.csv in a dataframe

Data Overview

```
Shape of the original dataset:  (39717, 111)

 A few records from top and bottom of the dataframe:
```

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_ra
0	1077501	1296599	5000	5000	4975.0	36 months	10.65
1	1077430	1314167	2500	2500	2500.0	60 months	15.27
2	1077175	1313524	2400	2400	2400.0	36 months	15.96
3	1076863	1277178	10000	10000	10000.0	36 months	13.49
4	1075358	1311748	3000	3000	3000.0	60 months	12.69
39712	92187	92174	2500	2500	1075.0	36 months	8.07
39713	90665	90607	8500	8500	875.0	36 months	10.28
39714	90395	90390	5000	5000	1325.0	36 months	8.07
39715	90376	89243	5000	5000	650.0	36 months	7.43
39716	87023	86999	7500	7500	800.0	36 months	13.75

10 rows × 111 columns

This is indeed an extensive dataset with **111 columns** and **39717 records**.

```
Datatypes of columns and their respective count:

float64    74
object     24
int64      13
dtype: int64
```

Detailed information about individual column datatype and count of non-null values present in them:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 39717 entries, 0 to 39716
Data columns (total 111 columns):
 #    Column                          Non-Null Count  Dtype  
---   ------                          --------------  -----  
 0    id                              39717 non-null  int64  
 1    member_id                       39717 non-null  int64  
 2    loan_amnt                       39717 non-null  int64  
 3    funded_amnt                     39717 non-null  int64  
 4    funded_amnt_inv                 39717 non-null  float64
 5    term                            39717 non-null  object 
 6    int_rate                        39717 non-null  object 
 7    installment                     39717 non-null  float64
 8    grade                           39717 non-null  object 
 9    sub_grade                       39717 non-null  object 
 10   emp_title                       37258 non-null  object 
 11   emp_length                      38642 non-null  object 
 12   home_ownership                  39717 non-null  object 
 13   annual_inc                      39717 non-null  float64
 14   verification_status             39717 non-null  object 
 15   issue_d                         39717 non-null  object 
 16   loan_status                     39717 non-null  object 
 17   pymnt_plan                      39717 non-null  object 
 18   url                             39717 non-null  object 
 19   desc                            26777 non-null  object 
 20   purpose                         39717 non-null  object 
 21   title                           39706 non-null  object 
 22   zip_code                        39717 non-null  object 
 23   addr_state                      39717 non-null  object 
 24   dti                             39717 non-null  float64
 25   delinq_2yrs                     39717 non-null  int64  
 26   earliest_cr_line                39717 non-null  object 
 27   inq_last_6mths                  39717 non-null  int64  
 28   mths_since_last_delinq          14035 non-null  float64
 29   mths_since_last_record          2786 non-null   float64
 30   open_acc                        39717 non-null  int64  
 31   pub_rec                         39717 non-null  int64  
 32   revol_bal                       39717 non-null  int64  
 33   revol_util                      39667 non-null  object 
 34   total_acc                       39717 non-null  int64  
 35   initial_list_status             39717 non-null  object 
 36   out_prncp                       39717 non-null  float64
 37   out_prncp_inv                   39717 non-null  float64
 38   total_pymnt                     39717 non-null  float64
 39   total_pymnt_inv                 39717 non-null  float64
 40   total_rec_prncp                 39717 non-null  float64
 41   total_rec_int                   39717 non-null  float64
 42   total_rec_late_fee              39717 non-null  float64
 43   recoveries                      39717 non-null  float64
 44   collection_recovery_fee         39717 non-null  float64
 45   last_pymnt_d                    39646 non-null  object 
 46   last_pymnt_amnt                 39717 non-null  float64
 47   next_pymnt_d                    1140 non-null   object 
 48   last_credit_pull_d              39715 non-null  object 
 49   collections_12_mths_ex_med      39661 non-null  float64
 50   mths_since_last_major_derog     0 non-null      float64
 51   policy_code                     39717 non-null  int64  
 52   application_type                39717 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                  39717 non-null  int64  
 57   tot_coll_amt                    0 non-null      float64
 58   tot_cur_bal                     0 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                0 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
 74   acc_open_past_24mths            0 non-null      float64
 75   avg_cur_bal                     0 non-null      float64
 76   bc_open_to_buy                  0 non-null      float64
 77   bc_util                         0 non-null      float64
 78   chargeoff_within_12_mths        39661 non-null  float64
 79   delinq_amnt                     39717 non-null  int64
 80   mo_sin_old_il_acct              0 non-null      float64
 81   mo_sin_old_rev_tl_op            0 non-null      float64
 82   mo_sin_rcnt_rev_tl_op           0 non-null      float64
 83   mo_sin_rcnt_tl                  0 non-null      float64
 84   mort_acc                        0 non-null      float64
 85   mths_since_recent_bc            0 non-null      float64
 86   mths_since_recent_bc_dlq        0 non-null      float64
 87   mths_since_recent_inq           0 non-null      float64
 88   mths_since_recent_revol_delinq  0 non-null      float64
 89   num_accts_ever_120_pd           0 non-null      float64
 90   num_actv_bc_tl                  0 non-null      float64
 91   num_actv_rev_tl                 0 non-null      float64
 92   num_bc_sats                     0 non-null      float64
 93   num_bc_tl                       0 non-null      float64
 94   num_il_tl                       0 non-null      float64
 95   num_op_rev_tl                   0 non-null      float64
 96   num_rev_accts                   0 non-null      float64
 97   num_rev_tl_bal_gt_0             0 non-null      float64
 98   num_sats                        0 non-null      float64
 99   num_tl_120dpd_2m                0 non-null      float64
 100  num_tl_30dpd                    0 non-null      float64
 101  num_tl_90g_dpd_24m              0 non-null      float64
 102  num_tl_op_past_12m              0 non-null      float64
 103  pct_tl_nvr_dlq                  0 non-null      float64
 104  percent_bc_gt_75                0 non-null      float64
 105  pub_rec_bankruptcies            39020 non-null  float64
 106  tax_liens                       39678 non-null  float64
 107  tot_hi_cred_lim                 0 non-null      float64
 108  total_bal_ex_mort               0 non-null      float64
 109  total_bc_limit                  0 non-null      float64
 110  total_il_high_credit_limit      0 non-null      float64
dtypes: float64(74), int64(13), object(24)
memory usage: 33.6+ MB
```

We have got some basic information about our dataset -

- Dimension
- Column names
- Datatypes
- Non-null count for every column

We observe that this dataset has many column attributes, and therefore, we will filter out some of the columns which are not important for our current objective of determining driving factors (or driver variables) behind loan default. This will enable us to focus on the column attributes that are strong determiners.

Data cleaning

Verifying if id and member_id columns are identifiers

```
id column an identifier: True
member_id column an identifier: True
```

```
Dropping 28 columns from dataframe...
Shape of the updated dataframe:  (39717, 83)
```

Inspecting columns with all null values

The following are the columns with all null values:

```
['mths_since_last_major_derog',
 'annual_inc_joint',
 'dti_joint',
 'verification_status_joint',
 'tot_coll_amt',
 'tot_cur_bal',
 'open_acc_6m',
 'open_il_6m',
 'open_il_12m',
 'open_il_24m',
 'mths_since_rcnt_il',
 'total_bal_il',
 'il_util',
 'open_rv_12m',
 'open_rv_24m',
 'max_bal_bc',
 'all_util',
 'total_rev_hi_lim',
 'inq_fi',
 'total_cu_tl',
 'inq_last_12m',
 'acc_open_past_24mths',
 'avg_cur_bal',
 'bc_open_to_buy',
 'bc_util',
 'mo_sin_old_il_acct',
 'mo_sin_old_rev_tl_op',
 'mo_sin_rcnt_rev_tl_op',
 'mo_sin_rcnt_tl',
 'mort_acc',
 'mths_since_recent_bc',
 'mths_since_recent_bc_dlq',
 'mths_since_recent_inq',
 'mths_since_recent_revol_delinq',
 'num_accts_ever_120_pd',
 'num_actv_bc_tl',
 'num_actv_rev_tl',
 'num_bc_sats',
 'num_bc_tl',
 'num_il_tl',
 'num_op_rev_tl',
 'num_rev_accts',
 'num_rev_tl_bal_gt_0',
 'num_sats',
 'num_tl_120dpd_2m',
 'num_tl_30dpd',
 'num_tl_90g_dpd_24m',
 'num_tl_op_past_12m',
 'pct_tl_nvr_dlq',
 'percent_bc_gt_75',
 'tot_hi_cred_lim',
 'total_bal_ex_mort',
 'total_bc_limit',
 'total_il_high_credit_limit']
```

```
Dropping 54 columns having zero non-null values...
Shape of the updated dataset:  (39717, 29)
```

Analysing missing values

```
Missing value percentage in each column:

loan_amnt                      0.0
funded_amnt                    0.0
funded_amnt_inv                0.0
term                           0.0
int_rate                       0.0
installment                    0.0
grade                          0.0
sub_grade                      0.0
emp_length                     2.7
home_ownership                 0.0
annual_inc                     0.0
verification_status            0.0
issue_d                        0.0
loan_status                    0.0
pymnt_plan                     0.0
purpose                        0.0
addr_state                     0.0
dti                            0.0
mths_since_last_delinq        64.7
mths_since_last_record        93.0
initial_list_status            0.0
next_pymnt_d                  97.1
collections_12_mths_ex_med     0.1
policy_code                    0.0
acc_now_delinq                 0.0
chargeoff_within_12_mths       0.1
delinq_amnt                    0.0
pub_rec_bankruptcies           1.8
tax_liens                      0.1
dtype: float64
```

Missing values percentage is varying from 0 to approximately 97% across the columns. Next, we drop any column with more than 60% missing values

```
No. of columns with more than 60% missing values: 3
['mths_since_last_delinq', 'mths_since_last_record', 'next_pymnt_d']
```

Dropping columns with more that 60% missing values

```
Shape of the updated dataframe:  (39717, 26)
```

Analysing columns with all values as zero

Analysing column tax_liens

```
Values in tax_liens -

0.0    39678
Name: tax_liens, dtype: int64
```

Analysing column delinq_amnt

```
Values in delinq_amnt -

0    39717
Name: delinq_amnt, dtype: int64
```

Dropping tax_liens and delinq_amnt columns, as they contain only zeros

```
Shape of the updated dataframe:  (39717, 24)
```

Analysing collections_12_mths_ex_med column

```
0.0    39661
Name: collections_12_mths_ex_med, dtype: int64
```

```
Shape of the updated dataframe:  (39717, 23)
```

Analysing acc_now_delinq column

```
0    39717
Name: acc_now_delinq, dtype: int64
```

```
Shape of the updated dataframe:  (39717, 22)
```

Analysing chargeoff_within_12_mths column

```
0.0    39661
Name: chargeoff_within_12_mths, dtype: int64
```

```
Shape of the updated dataframe:  (39717, 21)
```

Analysing and dropping the columns with only one value across all rows, since such columns do not contribute value to our current objective

Analysing initial_list_status

```
f    39717
Name: initial_list_status, dtype: int64
```

```
Shape of the updated dataframe:  (39717, 20)
```

Analysing policy_code column

```
1    39717
Name: policy_code, dtype: int64
```

```
Shape of the updated dataframe:  (39717, 19)
```

Analysing pymnt_plan column

```
n    39717
Name: pymnt_plan, dtype: int64
```

```
Shape of the updated dataframe:  (39717, 18)
```

Analysing loan_status column

```
Fully Paid     32950
Charged Off     5627
Current         1140
Name: loan_status, dtype: int64
```

We can drop the rows corresponding to the loan_status as "Current", since we can not perform prescriptive analysis on the ongoing loans

```
Shape of the updated dataframe:  (38577, 18)
```

Analysing pub_rec_bankruptcies column

```
0.0    36238
1.0     1637
2.0        5
Name: pub_rec_bankruptcies, dtype: int64
```

We note that almost 93.94% of the records have the number of public record bankruptcies as zero. Thus, we choose to drop this column as well.

```
Shape of the updated dataframe:  (38577, 17)
```

Analysing the grade and sub_grade columns

	grade	sub_grade
0	B	B2
1	C	C4
2	C	C5
3	C	C1
5	A	A4
6	C	C5
7	E	E1
8	F	F2
9	B	B5
10	C	C3

Apparantly, both grade and sub_grade are the LC-assigned loan grades, wherein grade is a higher-level classification, with sub_grade not adding any adidtional information as such to our current analysis. Therefore, we have decided to drop column sub_grade.

```
Shape of the updated dataframe:  (38577, 16)
```

Duplicate data inspection

```
False
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 38577 entries, 0 to 39716
Data columns (total 16 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   loan_amnt            38577 non-null  int64  
 1   funded_amnt          38577 non-null  int64  
 2   funded_amnt_inv      38577 non-null  float64
 3   term                 38577 non-null  object 
 4   int_rate             38577 non-null  object 
 5   installment          38577 non-null  float64
 6   grade                38577 non-null  object 
 7   emp_length           37544 non-null  object 
 8   home_ownership       38577 non-null  object 
 9   annual_inc           38577 non-null  float64
 10  verification_status  38577 non-null  object 
 11  issue_d              38577 non-null  object 
 12  loan_status          38577 non-null  object 
 13  purpose              38577 non-null  object 
 14  addr_state           38577 non-null  object 
 15  dti                  38577 non-null  float64
dtypes: float64(4), int64(2), object(10)
memory usage: 5.0+ MB
```

We have narrowed down our dataframe to 16 columns and 38577 records.

Fixing Data

Since int_rate is a continous variable, we need to fix datatype of this column from object to float

For convenience and to facilitate analysis, we extract the month and year from the column issue_d, and create two new columns issue_year and issue_month. Subsequently, since the data of issue_d is now available to us in the two new columns, we drop issue_d column

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 38577 entries, 0 to 39716
Data columns (total 17 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   loan_amnt            38577 non-null  int64  
 1   funded_amnt          38577 non-null  int64  
 2   funded_amnt_inv      38577 non-null  float64
 3   term                 38577 non-null  object 
 4   int_rate             38577 non-null  float64
 5   installment          38577 non-null  float64
 6   grade                38577 non-null  object 
 7   emp_length           37544 non-null  object 
 8   home_ownership       38577 non-null  object 
 9   annual_inc           38577 non-null  float64
 10  verification_status  38577 non-null  object 
 11  loan_status          38577 non-null  object 
 12  purpose              38577 non-null  object 
 13  addr_state           38577 non-null  object 
 14  dti                  38577 non-null  float64
 15  issue_year           38577 non-null  object 
 16  issue_month          38577 non-null  object 
dtypes: float64(5), int64(2), object(10)
memory usage: 5.3+ MB
```

Variable types

At this point of time we can broadly divide all variables into two types -

- Categorical variables
- Numeric variables

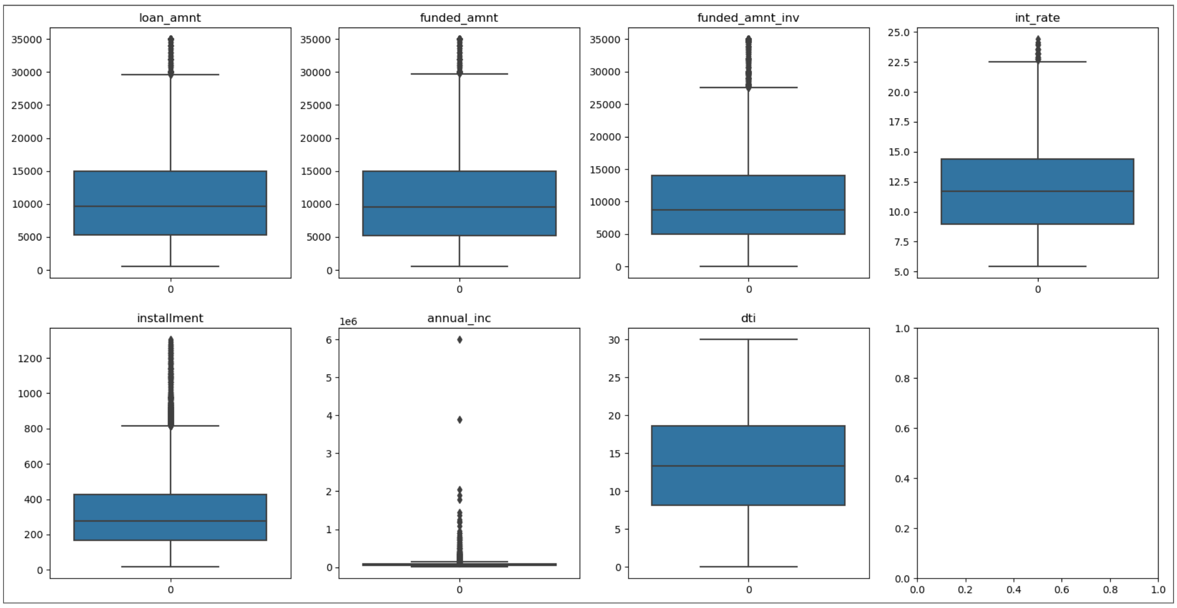
```
Shape of the updated dataframe after cleaning the data and creating new
columns:  (38577, 17)
Number of categorical variables 10
Number of continuous variables 7
```

Outlier detection

	count	mean	std	min	25%
loan_amnt	38577.0	11047.025430	7348.441646	500.00	5300.00
funded_amnt	38577.0	10784.058506	7090.306027	500.00	5200.00
funded_amnt_inv	38577.0	10222.481123	7022.720644	0.00	5000.00
int_rate	38577.0	11.932219	3.691327	5.42	8.94
installment	38577.0	322.466318	208.639215	15.69	165.74
annual_inc	38577.0	68777.973681	64218.681802	4000.00	40000.00
dti	38577.0	13.272727	6.673044	0.00	8.13

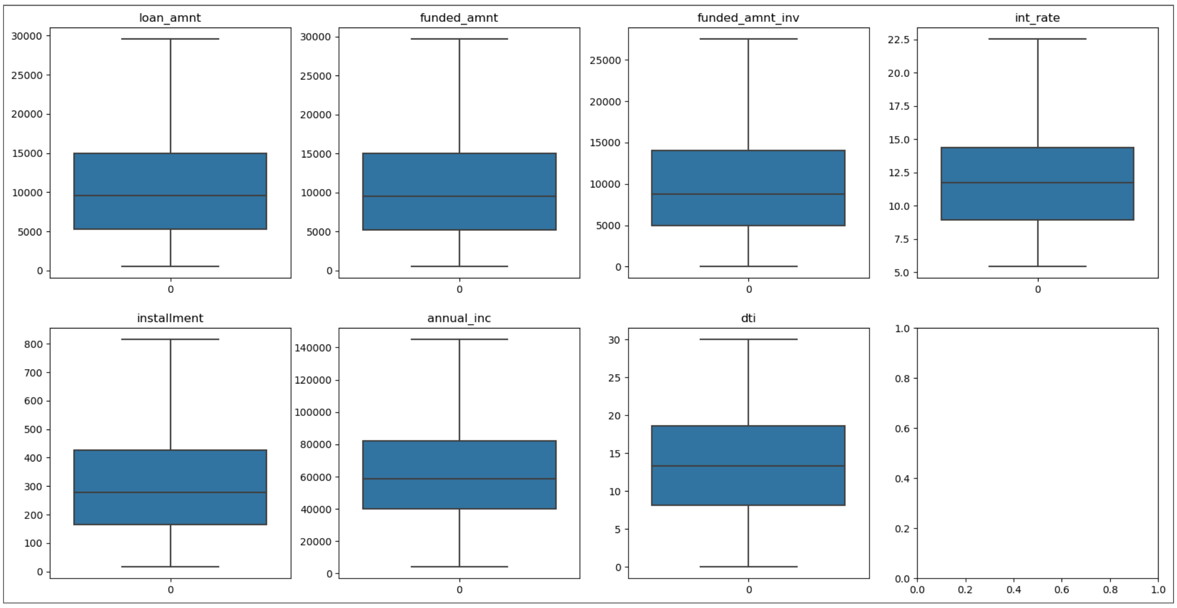
```
Columns with outliers: ['loan_amnt', 'funded_amnt', 'funded_amnt_inv',
'int_rate', 'installment', 'annual_inc']
```

Visualizing outliers through boxplots


loan_amnt
funded_amnt
funded_amnt_inv
int_rate
installment
annual_inc
1e6
dti

No outliers noted for dti. We proceed with treating the outliers for loan_amnt, funded_amnt, funded_amnt_inv, int_rate, installment, annual_inc by replacing the values less than the lower threshold by the lower threshold and the values more than the upper threshold by the upper threshold.

Final boxplot visualization post removal of outliers


loan_amnt
funded_amnt
funded_amnt_inv
int_rate
installment
annual_inc
dti

Checking dataframe shape after removal of outliers

```
Dataframe shape after the treatment of outliers remains the same as that
before the treatment of outliers: (38577, 17)
```

Univariate, Bivariate, and Multivariate Analyses

To explore individual variables and visualize all the variables for univariate analysis simulatenously, we use subplots, utilizing histograms for numeric variables and bar plots for categorical variables.

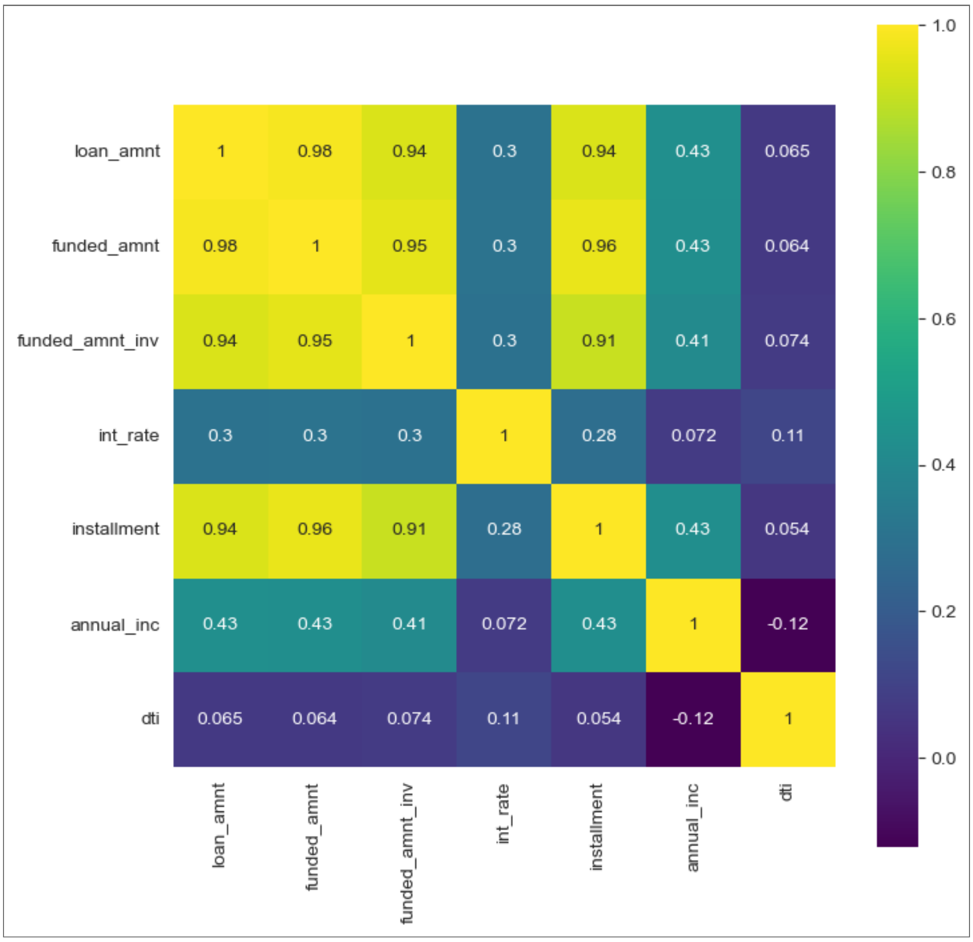

term
grade
emp_length
home_ownership
verification_status
loan_status
purpose
addr_state
issue_month
issue_year
loan_amnt
funded_amnt
funded_amnt_inv
int_rate
installment
annual_inc
dti

While the above plots provide an understanding of the spread, range, and distribution of the individual column variables across the records, it is more meaningful if we visualize the same plots specifically comparing between the defaulters (Charged Off) and non-defaulters (Fully Paid)


term
grade
emp_length
home_ownership
verification_status
loan_status
purpose
addr_state
issue_month
issue_year
loan_amnt
funded_amnt
funded_amnt_inv
int_rate
installment
annual_inc
dti
loan_status
Fully Paid
Charged Off

THE ABOVE ANALYSIS WITH RESPECT TO THE FULLY PAID AND CHARGED OFF LOANS FOR EACH VARIABLE REVEALS THE FOLLOWING:

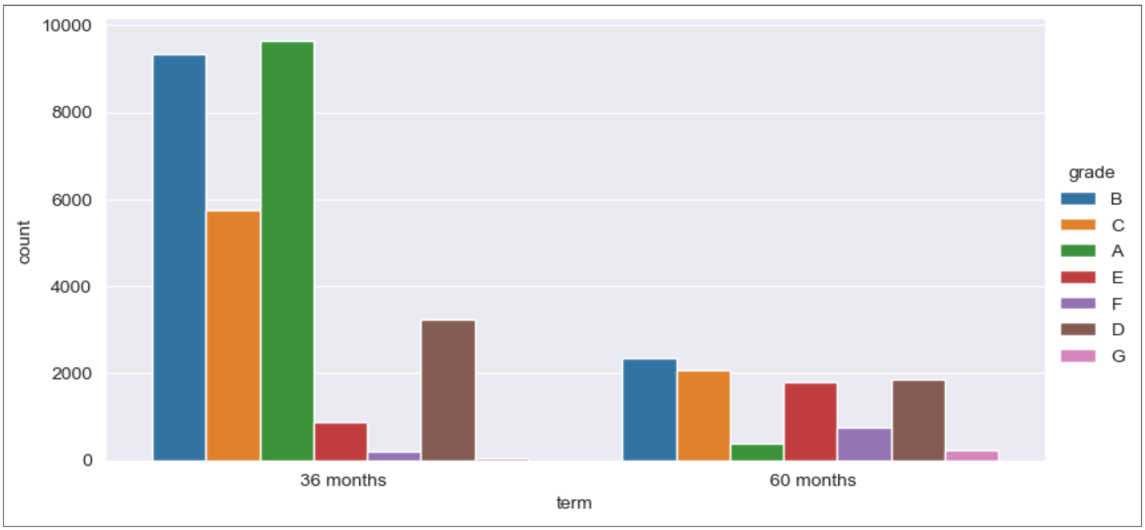
The number of defaulters appears to be more
• for shorter-term loans with a term of 36 months
• for grades B and C
• for 10 or more years of employment, which is slightly counter-intuitive
• for the home ownership status of 'RENT'
• when the income is not verified by LC
• when the purpose is debt consolidation
• for applicants filing with address as State CA, followed by NY and TX
• towards the last 4 months of the year, peaking in December
• in the loan issue year of 2011
• when the listed amount of the loan applied for by the borrower and the total funded amount committed to that loan at that point in time are around 5000, 10000, 12000, 15000
• when the total amount committed by investors for that loan at that point in time is in the range of 2500-12500 approximately
• when the interest rate is approximately 10-17%
• when the monthly payment or installment is approximately 50-400
• when the annual income is approximately 30000-70000
• dti is below 25 (dti is the ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income)

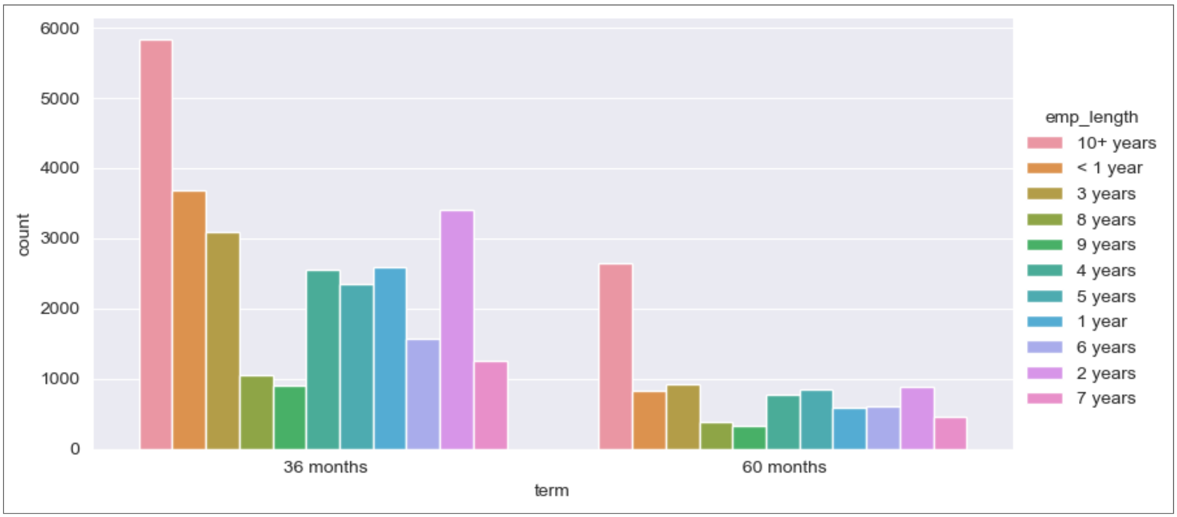


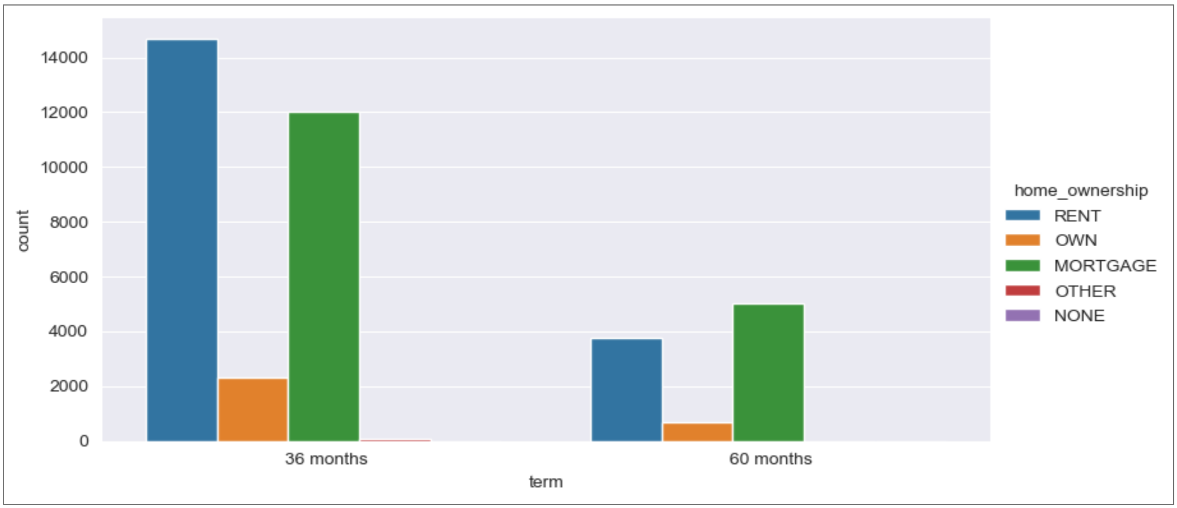

```
The top 10 pairs of variables with the largest absolute correlations:
 funded_amnt      loan_amnt          0.982832
                 installment        0.961555
                 funded_amnt_inv    0.953377
installment      loan_amnt          0.937448
funded_amnt_inv  loan_amnt          0.936486
                 installment        0.905209
annual_inc       loan_amnt          0.432883
                 installment        0.428981
                 funded_amnt        0.428814
                 funded_amnt_inv    0.408463
dtype: float64
```

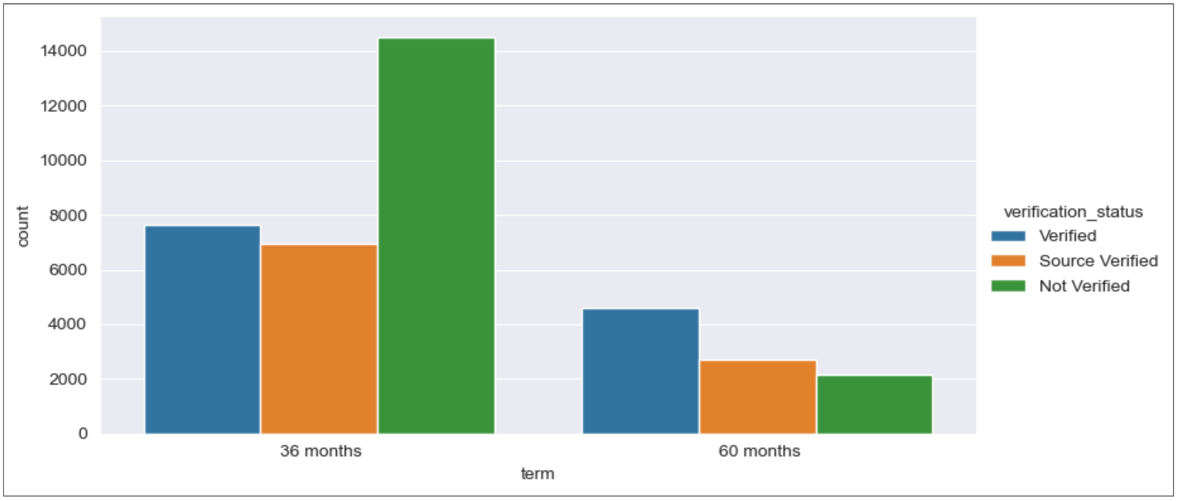
Based on the correlation coefficients, it seems that the three variables "funded_amnt", "loan_amnt", and "funded_amnt_inv" are highly correlated with each other. The next highly correlated pair is "installment" and "loan_amnt", followed by "funded_amnt_inv" and "installment".

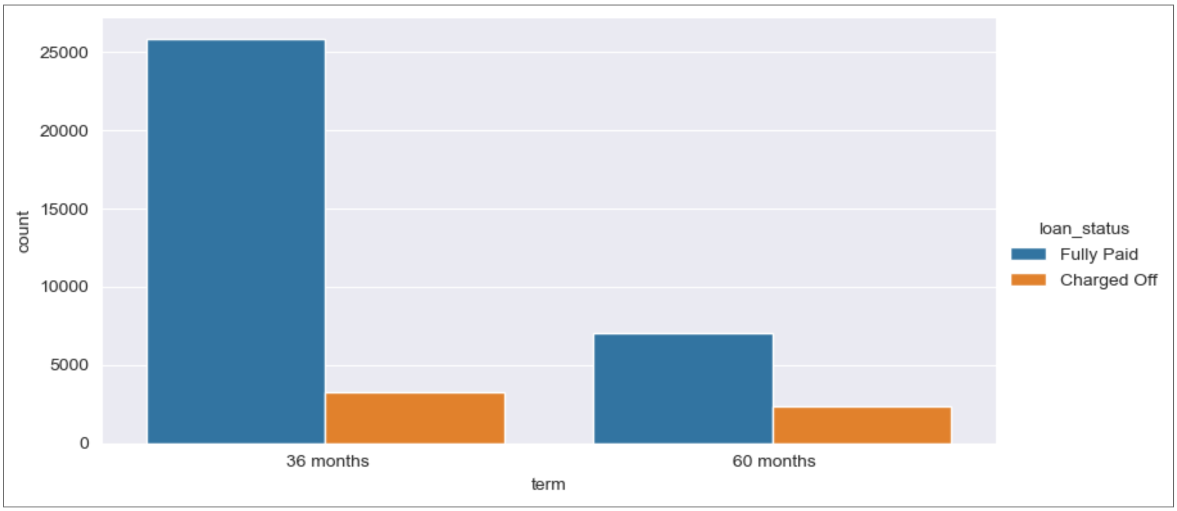
To understand the inter-relations between the catagorical variables, bar charts are used as shown below.

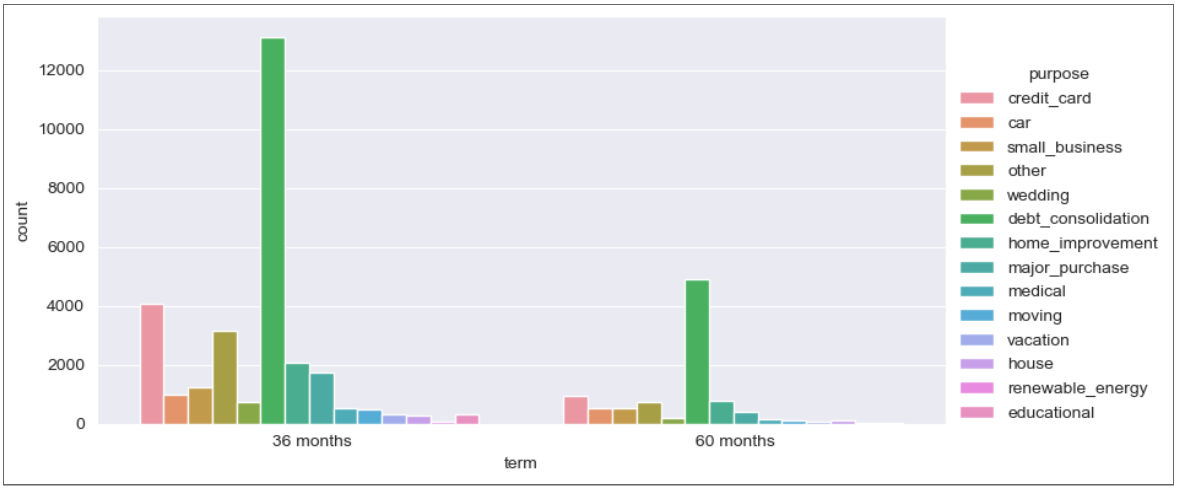

10000
8000
6000
4000
2000
0
count
36 months
60 months
term
grade
B
C
A
E
F
D
G

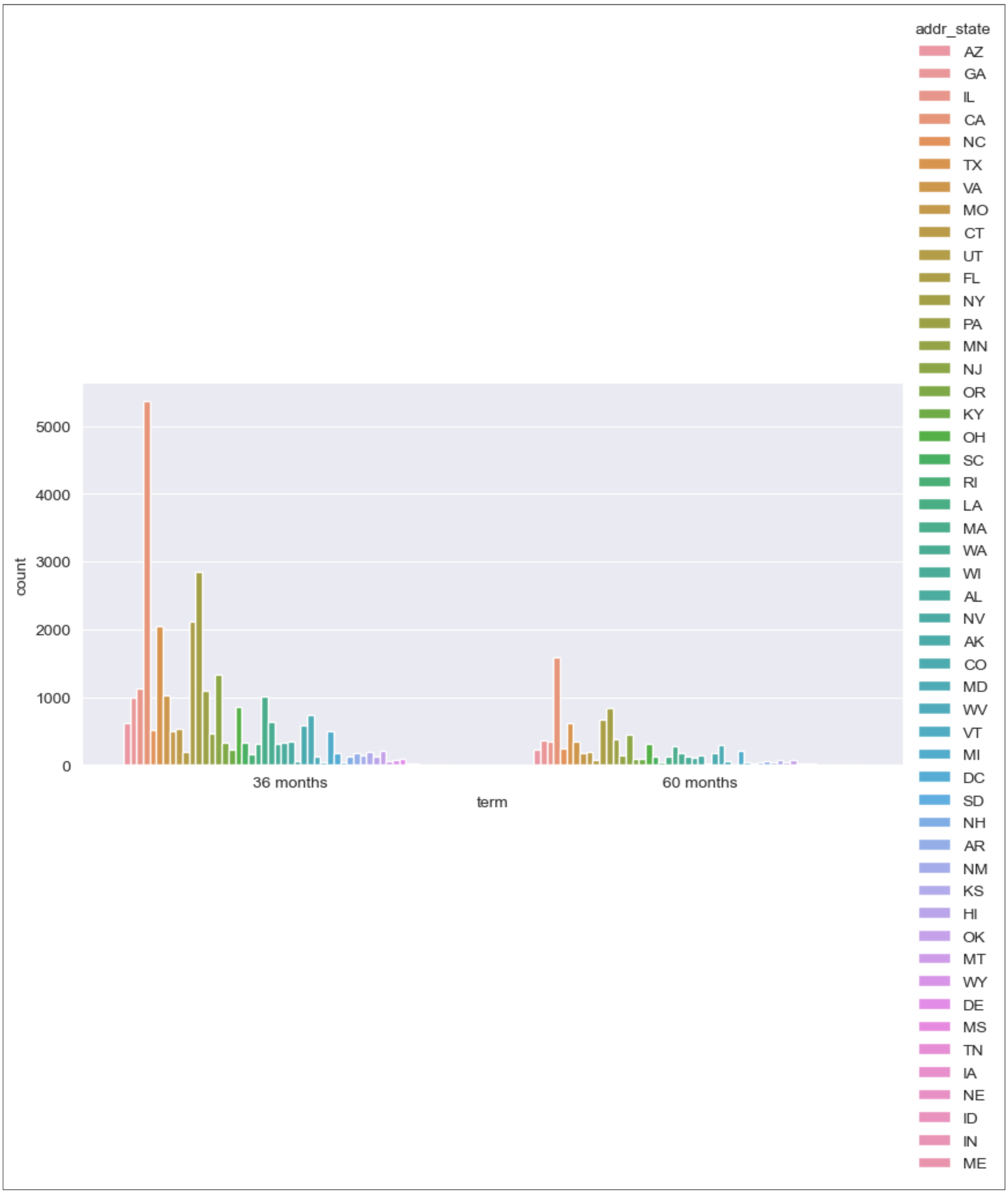

6000
5000
4000
3000
2000
1000
0
count
36 months
60 months
term
emp_length
10+ years
< 1 year
3 years
8 years
9 years
4 years
5 years
1 year
6 years
2 years
7 years

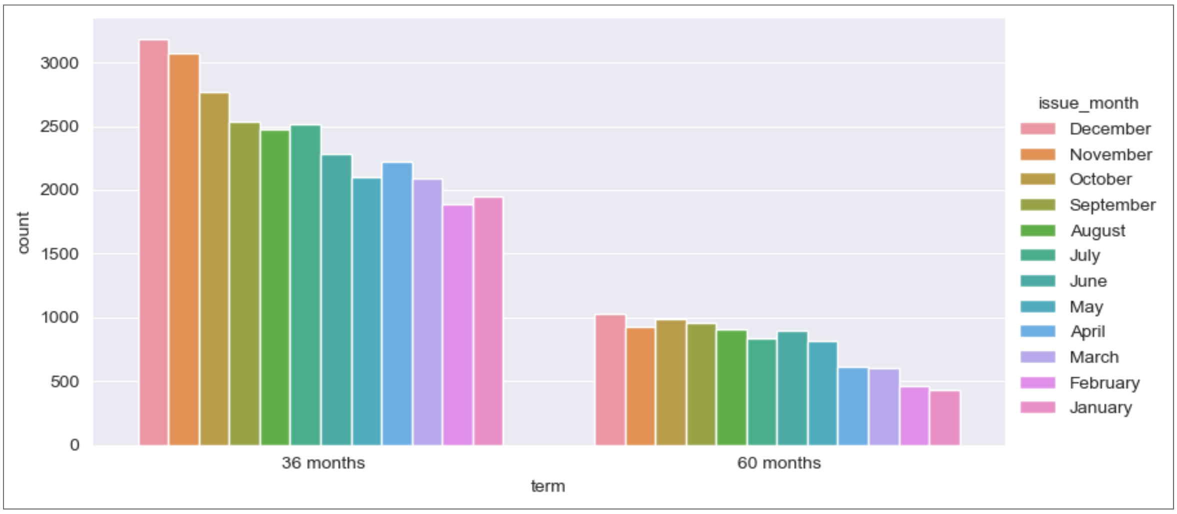

14000
12000
10000
8000
6000
4000
2000
0
count
36 months
60 months
term
home_ownership
RENT
OWN
MORTGAGE
OTHER
NONE

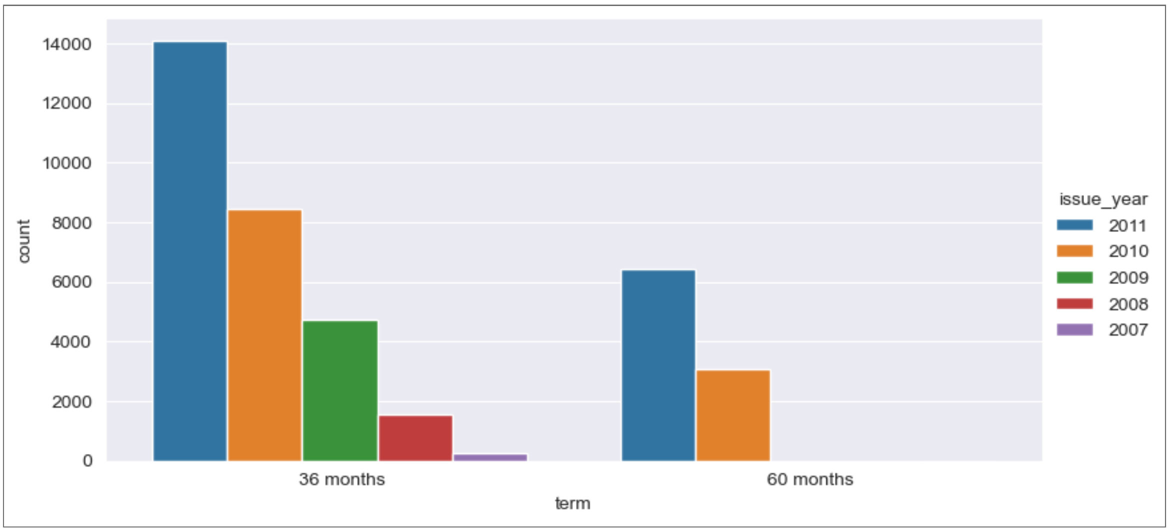

14000
12000
10000
8000
6000
4000
2000
0
count
36 months
60 months
term
verification_status
Verified
Source Verified
Not Verified

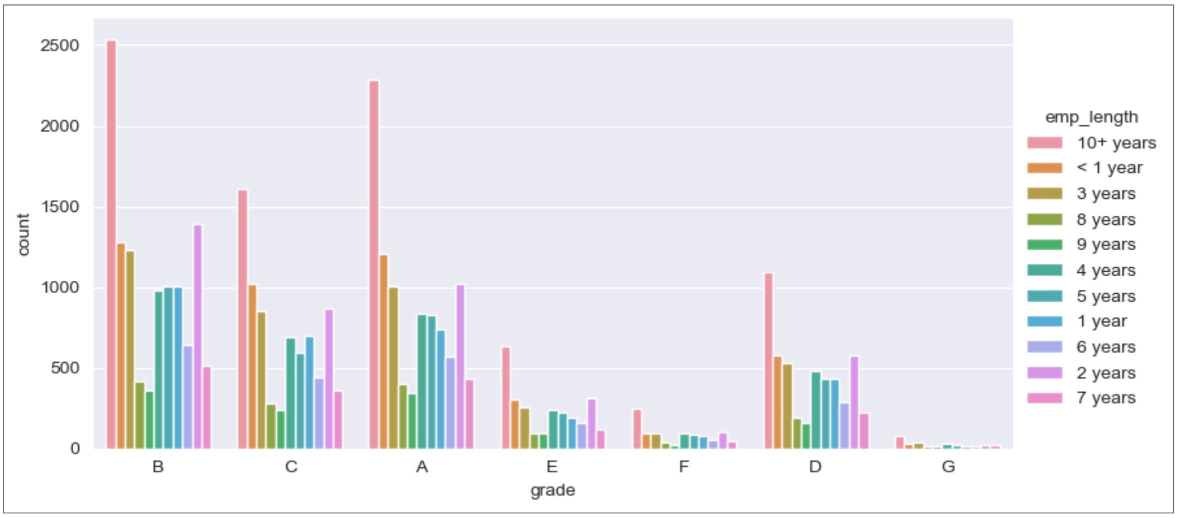

25000
20000
15000
10000
5000
0
count
36 months
60 months
term
loan_status
Fully Paid
Charged Off

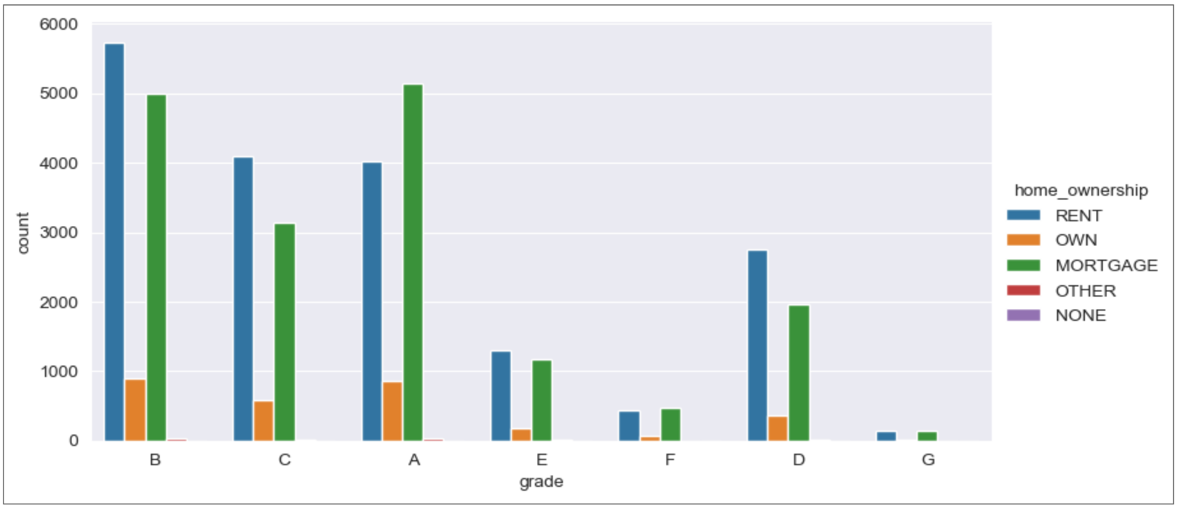

12000
10000
8000
6000
4000
2000
0
count
36 months
60 months
term
purpose
credit_card
car
small_business
other
wedding
debt_consolidation
home_improvement
major_purchase
medical
moving
vacation
house
renewable_energy
educational

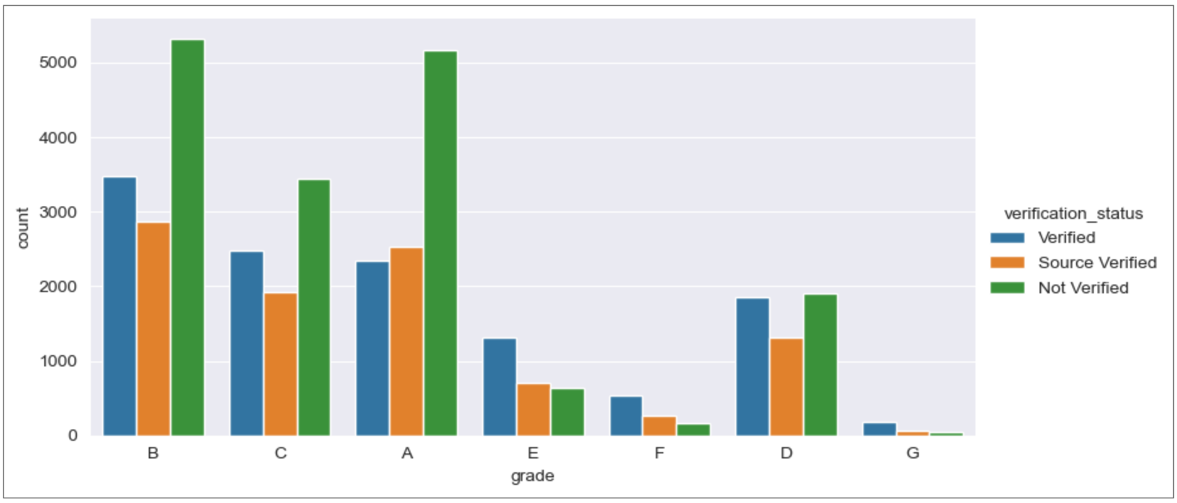

addr_state
AZ
GA
IL
CA
NC
TX
VA
MO
CT
UT
FL
NY
PA
MN
NJ
OR
KY
OH
SC
RI
LA
MA
WA
WI
AL
NV
AK
CO
MD
WV
VT
MI
DC
SD
NH
AR
NM
KS
HI
OK
MT
WY
DE
MS
TN
IA
NE
ID
IN
ME
5000
4000
3000
2000
1000
0
count
36 months
60 months
term

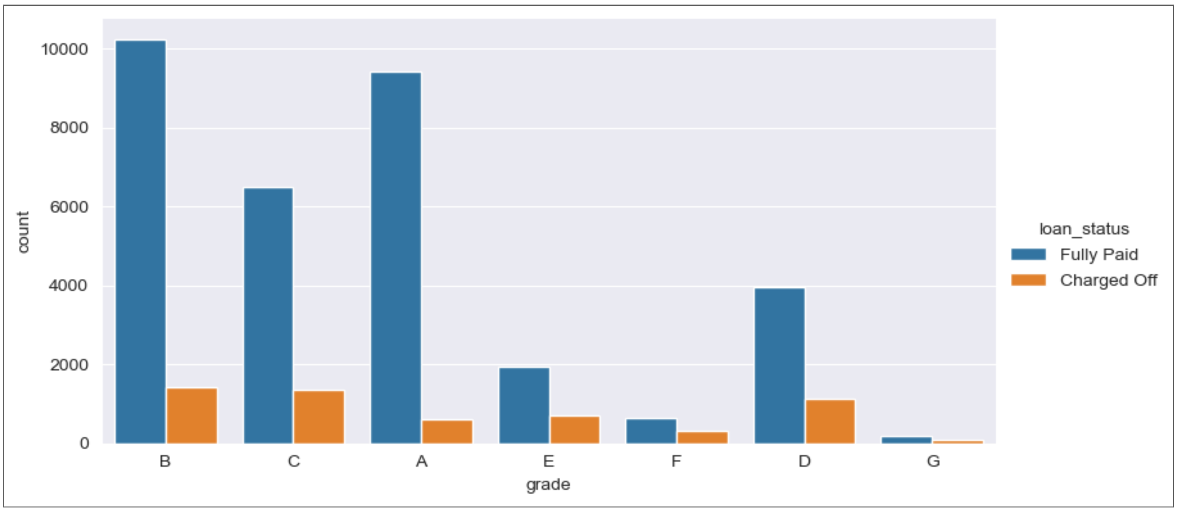

3000
2500
2000
1500
1000
500
0
count
36 months
60 months
term
issue_month
December
November
October
September
August
July
June
May
April
March
February
January

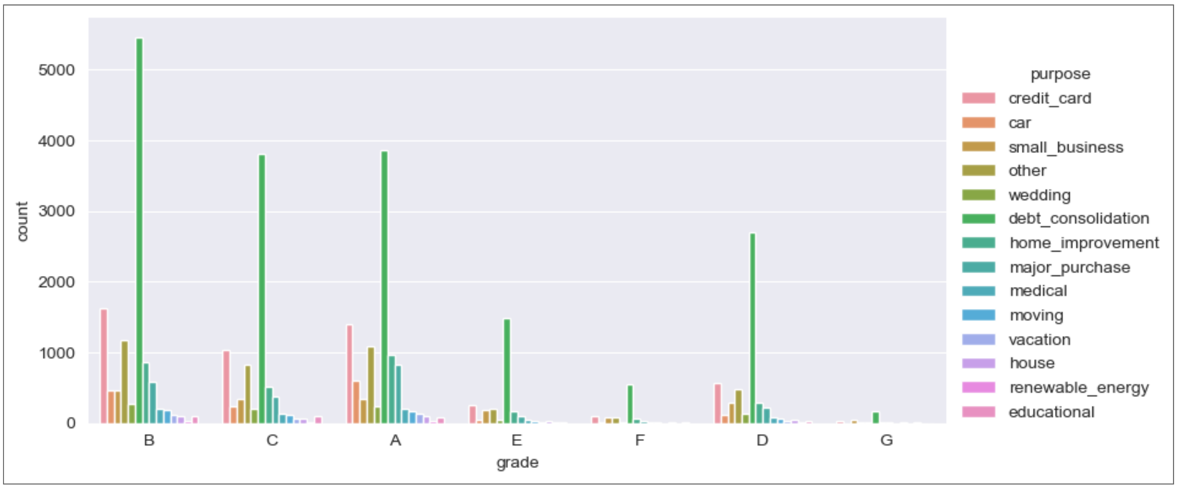

14000
12000
10000
8000
6000
4000
2000
0
count
36 months
60 months
term
issue_year
2011
2010
2009
2008
2007

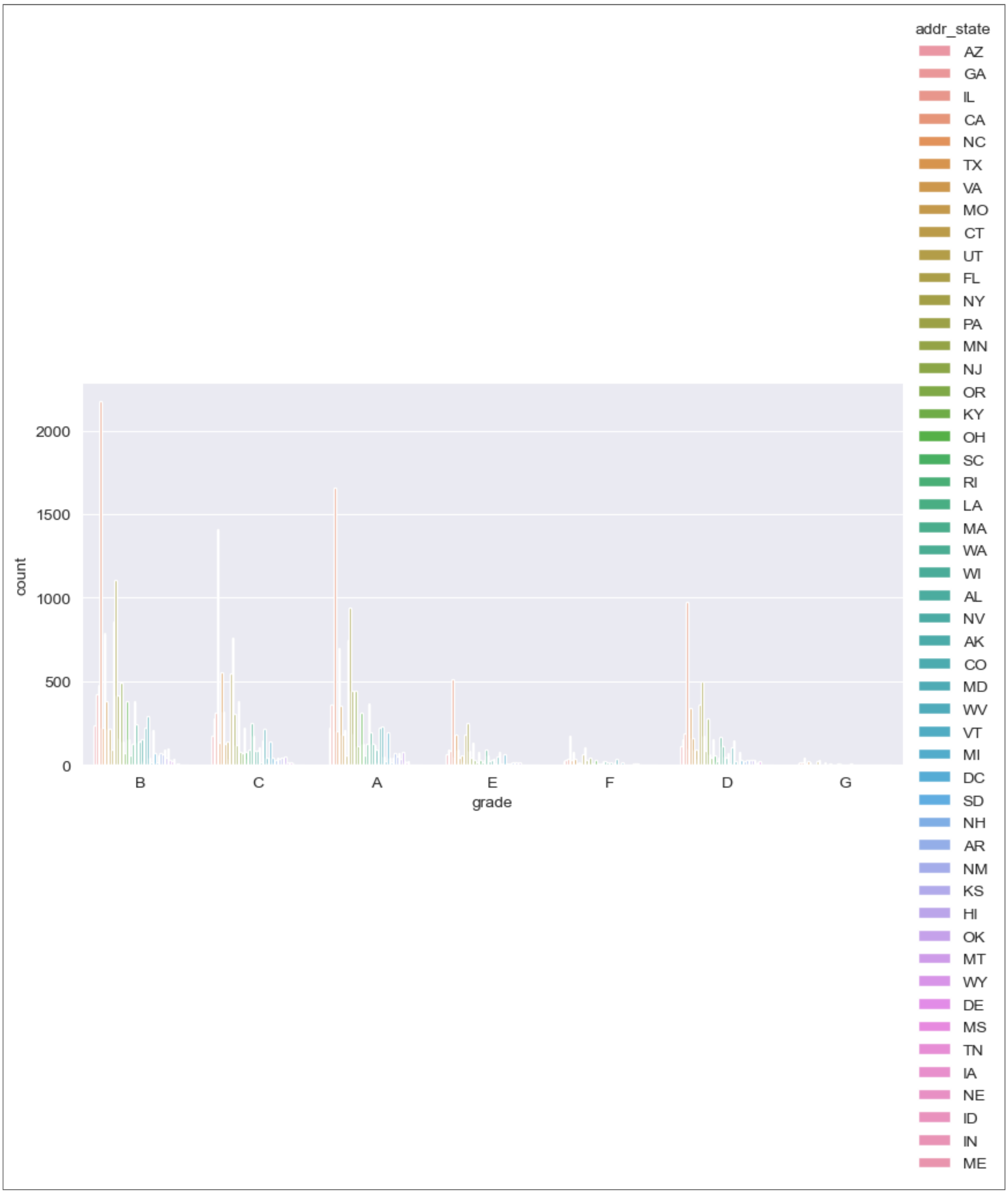

2500
2000
1500
1000
500
0
count
B
C
A
E
F
D
G
grade
emp_length
10+ years
< 1 year
3 years
8 years
9 years
4 years
5 years
1 year
6 years
2 years
7 years

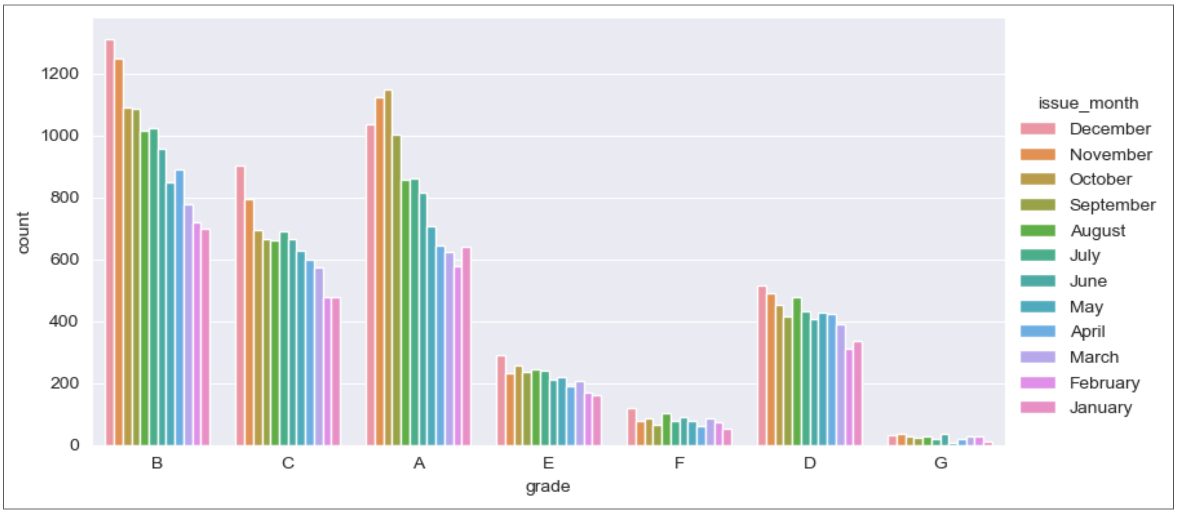

6000
5000
4000
3000
2000
1000
0
count
B
C
A
E
F
D
G
grade
home_ownership
RENT
OWN
MORTGAGE
OTHER
NONE

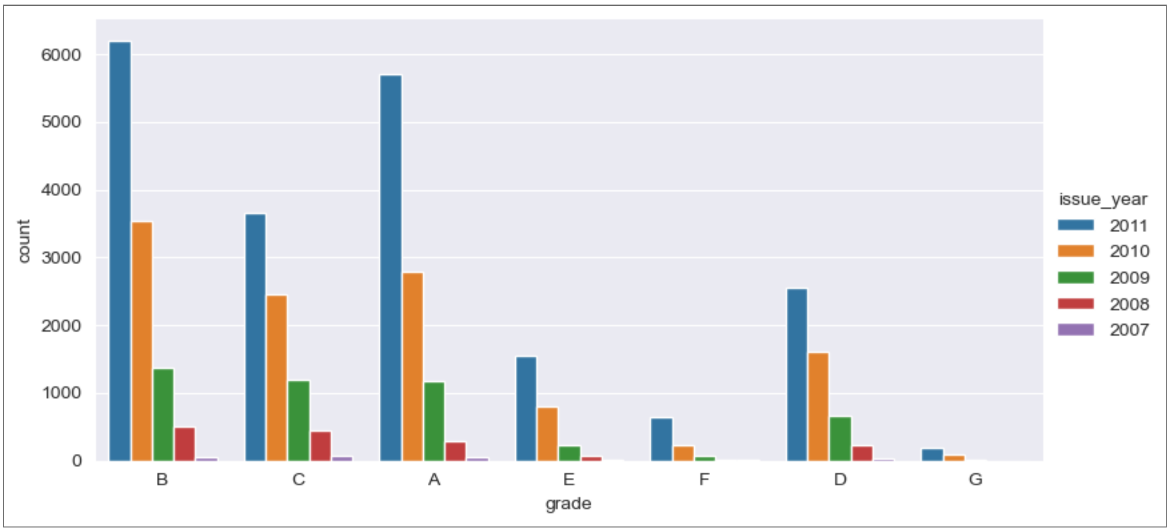

5000
4000
3000
2000
1000
0
count
B
C
A
E
F
D
G
grade
verification_status
Verified
Source Verified
Not Verified

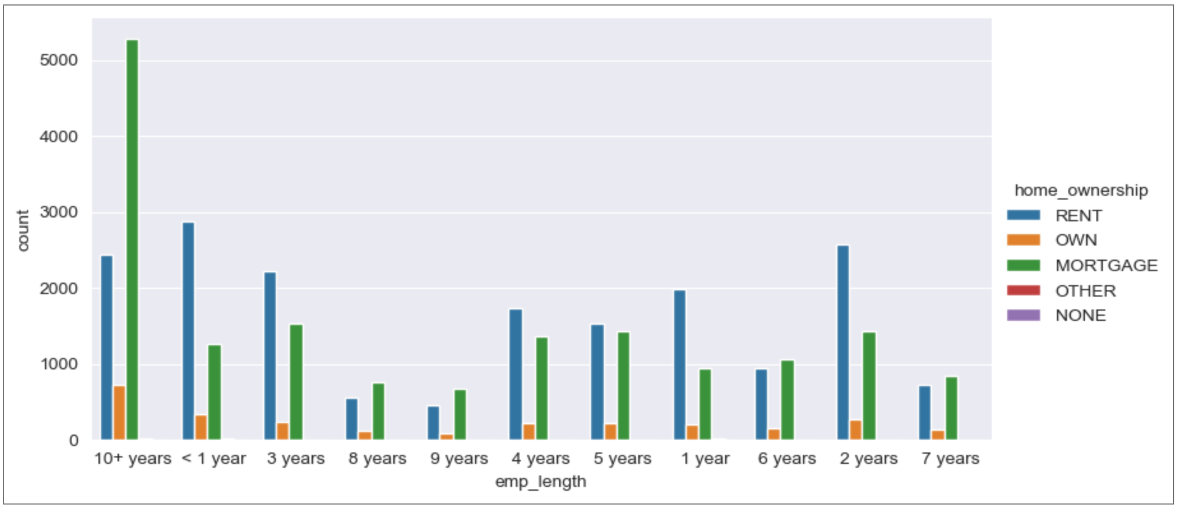

count
10000
8000
6000
4000
2000
0
B
C
A
E
F
D
G
grade
loan_status
Fully Paid
Charged Off

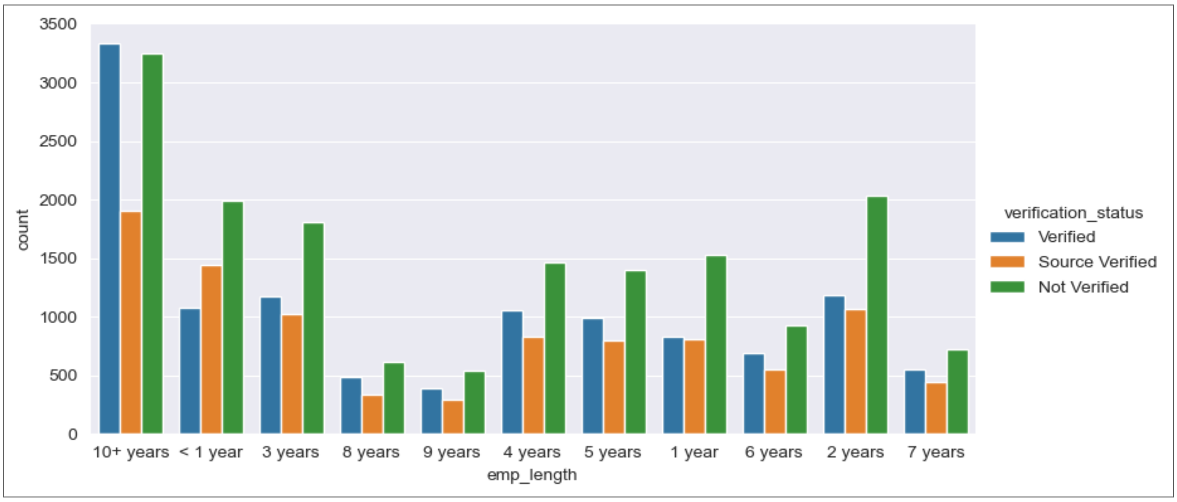

count
5000
4000
3000
2000
1000
0
B
C
A
E
F
D
G
grade
purpose
credit_card
car
small_business
other
wedding
debt_consolidation
home_improvement
major_purchase
medical
moving
vacation
house
renewable_energy
educational

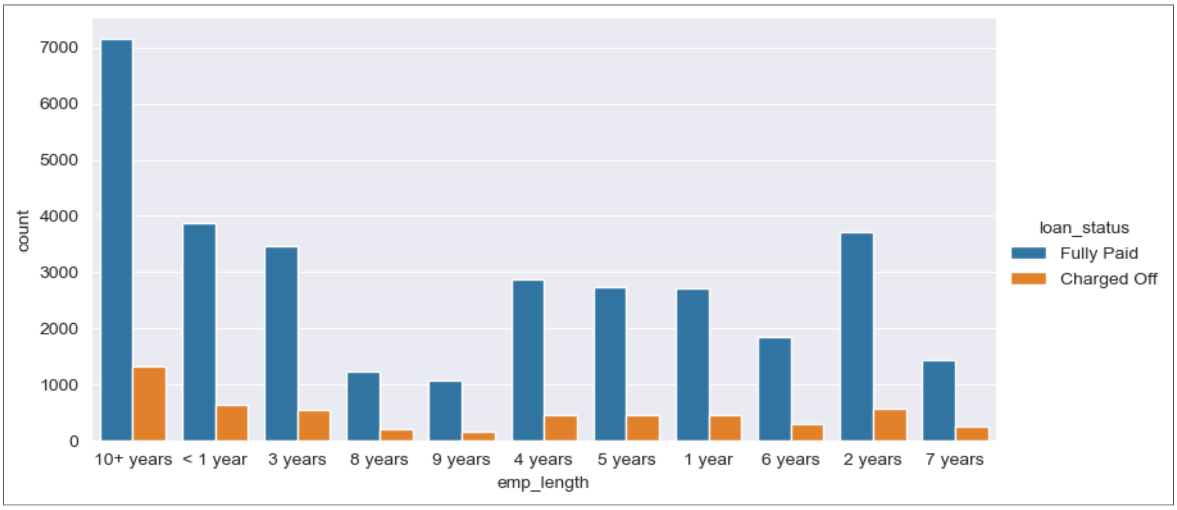

addr_state
AZ
GA
IL
CA
NC
TX
VA
MO
CT
UT
FL
NY
PA
MN
NJ
OR
KY
OH
SC
RI
LA
MA
WA
WI
AL
NV
AK
CO
MD
WV
VT
MI
DC
SD
NH
AR
NM
KS
HI
OK
MT
WY
DE
MS
TN
IA
NE
ID
IN
ME
count
2000
1500
1000
500
0
B
C
A
E
F
D
G
grade

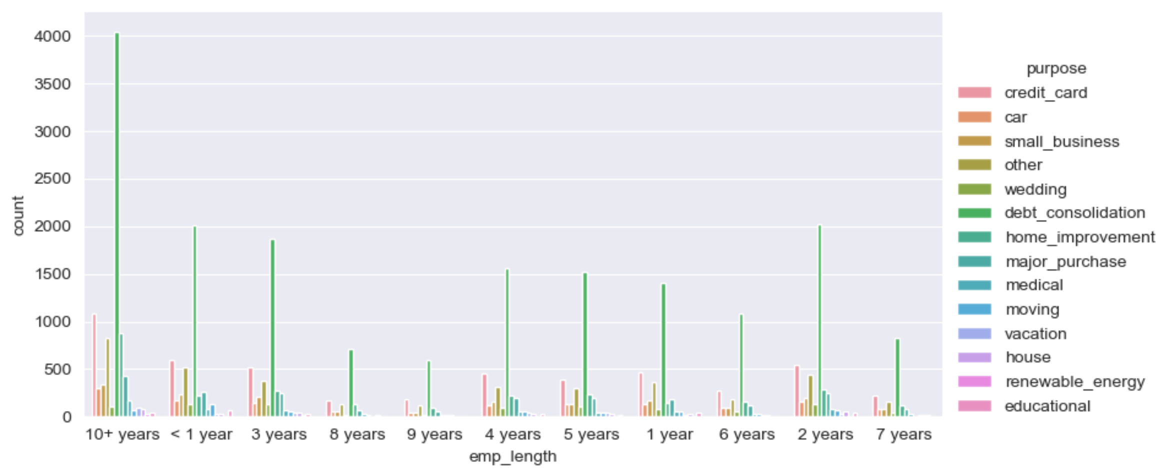

issue_month
December
November
October
September
August
July
June
May
April
March
February
January
count
1200
1000
800
600
400
200
0
B
C
A
E
F
D
G
grade

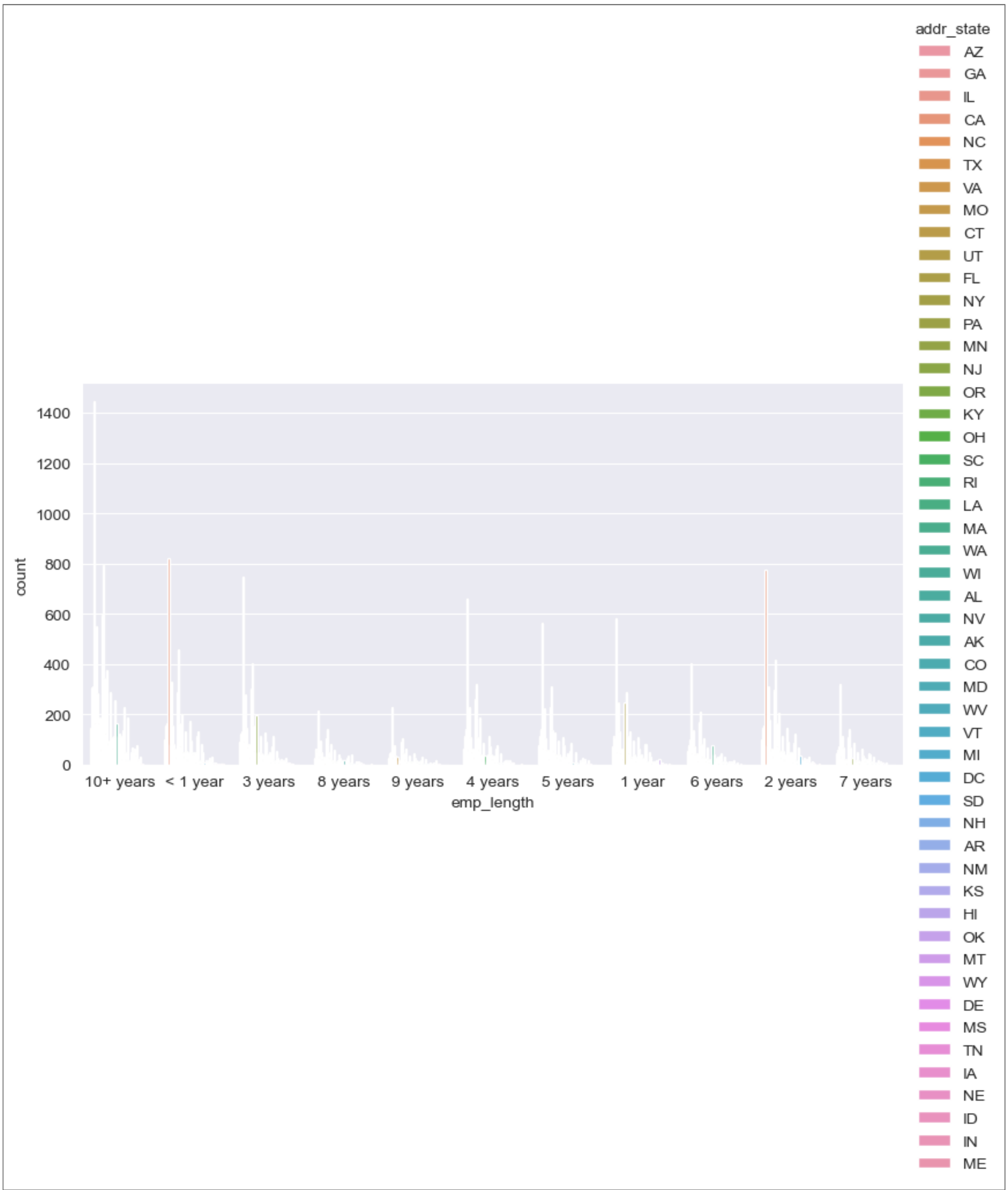

count
0
1000
2000
3000
4000
5000
6000
B
C
A
E
F
D
G
grade
issue_year
2011
2010
2009
2008
2007

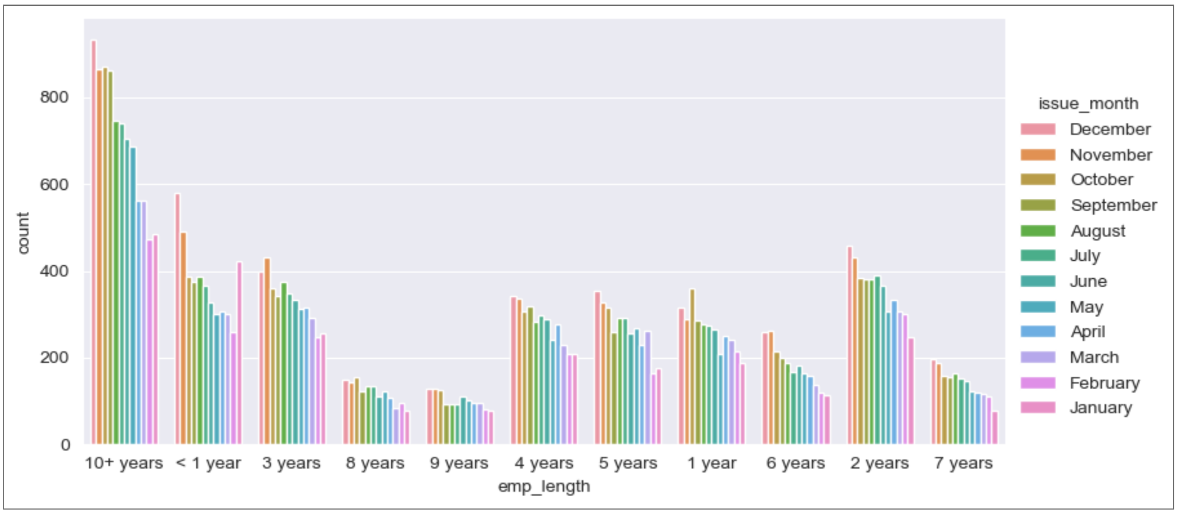

count
0
1000
2000
3000
4000
5000
10+ years
< 1 year
3 years
8 years
9 years
4 years
5 years
1 year
6 years
2 years
7 years
emp_length
home_ownership
RENT
OWN
MORTGAGE
OTHER
NONE

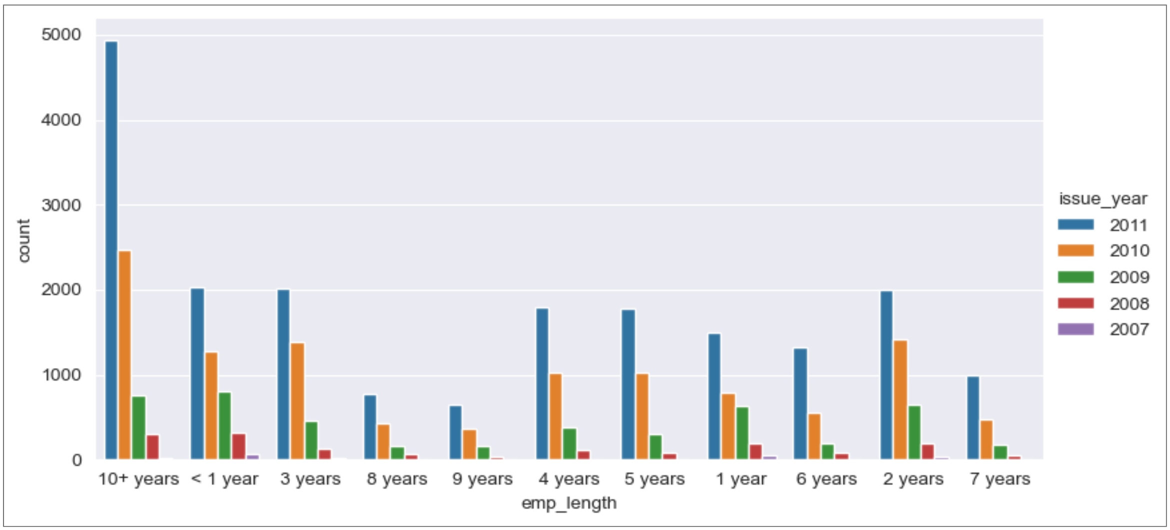

count
0
500
1000
1500
2000
2500
3000
3500
10+ years
< 1 year
3 years
8 years
9 years
4 years
5 years
1 year
6 years
2 years
7 years
emp_length
verification_status
Verified
Source Verified
Not Verified

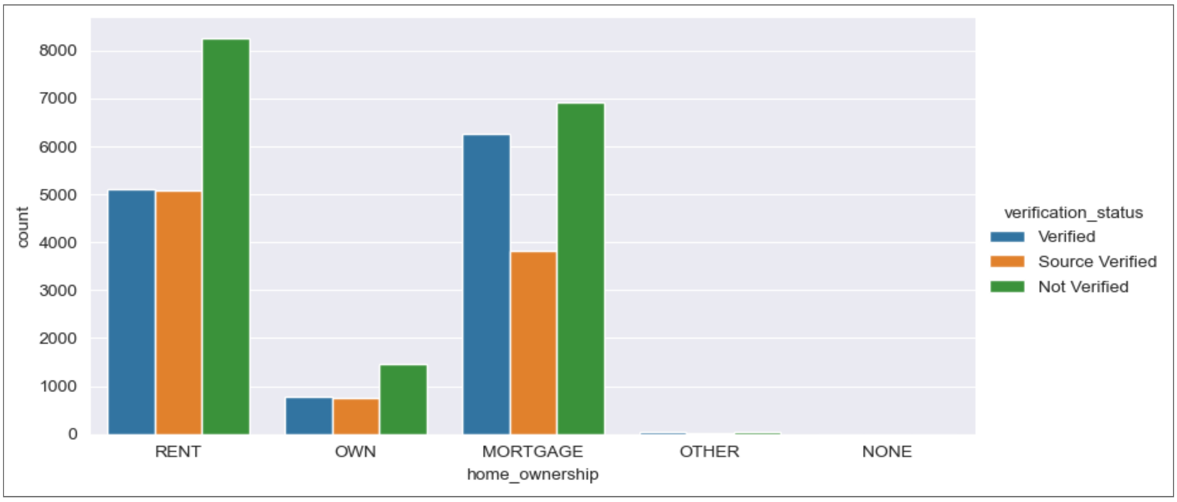

count
0
1000
2000
3000
4000
5000
6000
7000
10+ years
< 1 year
3 years
8 years
9 years
4 years
5 years
1 year
6 years
2 years
7 years
emp_length
loan_status
Fully Paid
Charged Off

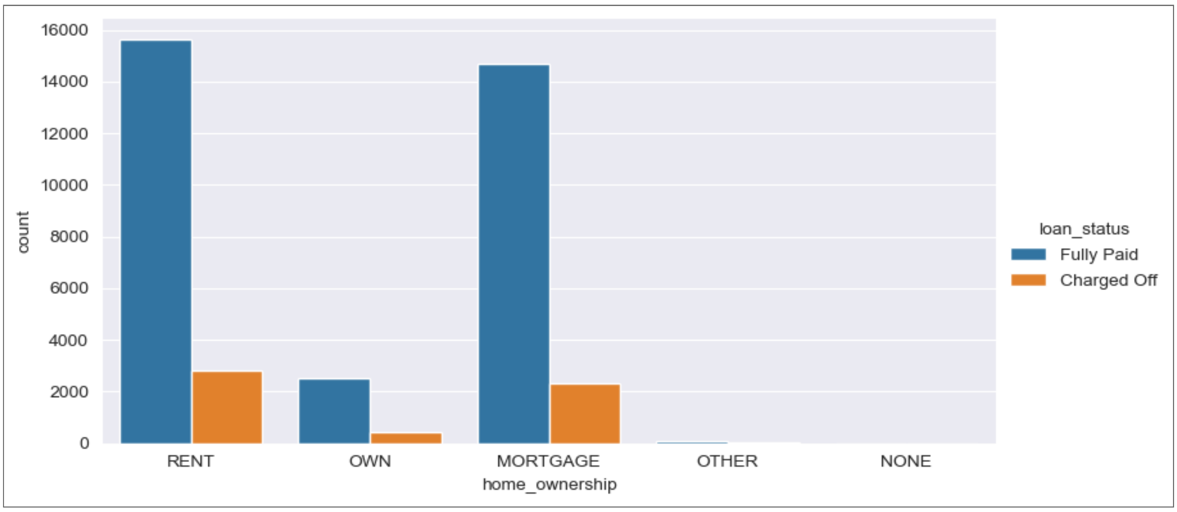

purpose
credit_card
car
small_business
other
wedding
debt_consolidation
home_improvement
major_purchase
medical
moving
vacation
house
renewable_energy
educational
count
4000
3500
3000
2500
2000
1500
1000
500
0
10+ years
< 1 year
3 years
8 years
9 years
4 years
5 years
1 year
6 years
2 years
7 years
emp_length

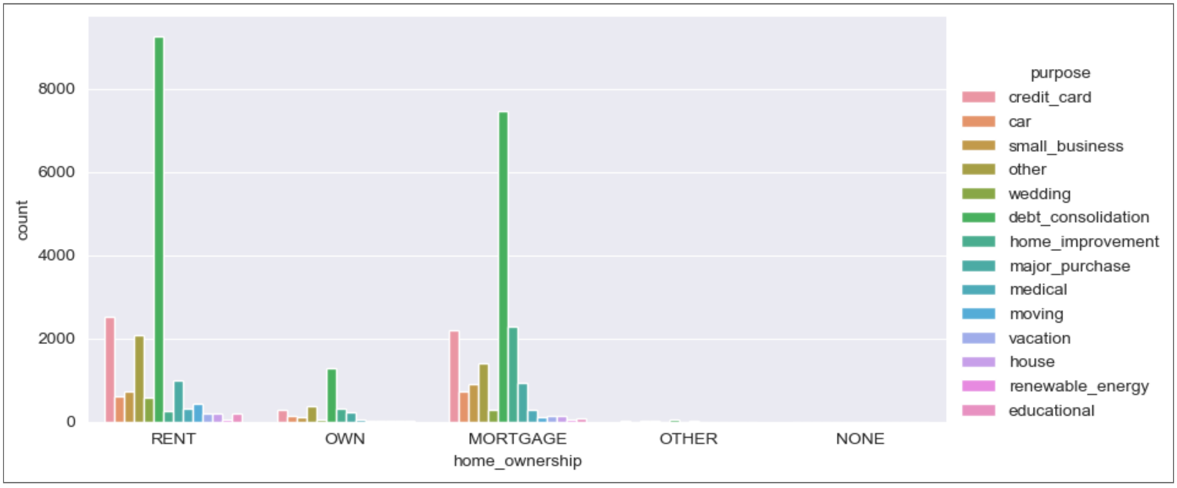

addr_state
AZ
GA
IL
CA
NC
TX
VA
MO
CT
UT
FL
NY
PA
MN
NJ
OR
KY
OH
SC
RI
LA
MA
WA
WI
AL
NV
AK
CO
MD
WV
VT
MI
DC
SD
NH
AR
NM
KS
HI
OK
MT
WY
DE
MS
TN
IA
NE
ID
IN
ME
count
1400
1200
1000
800
600
400
200
0
10+ years
< 1 year
3 years
8 years
9 years
4 years
5 years
1 year
6 years
2 years
7 years
emp_length

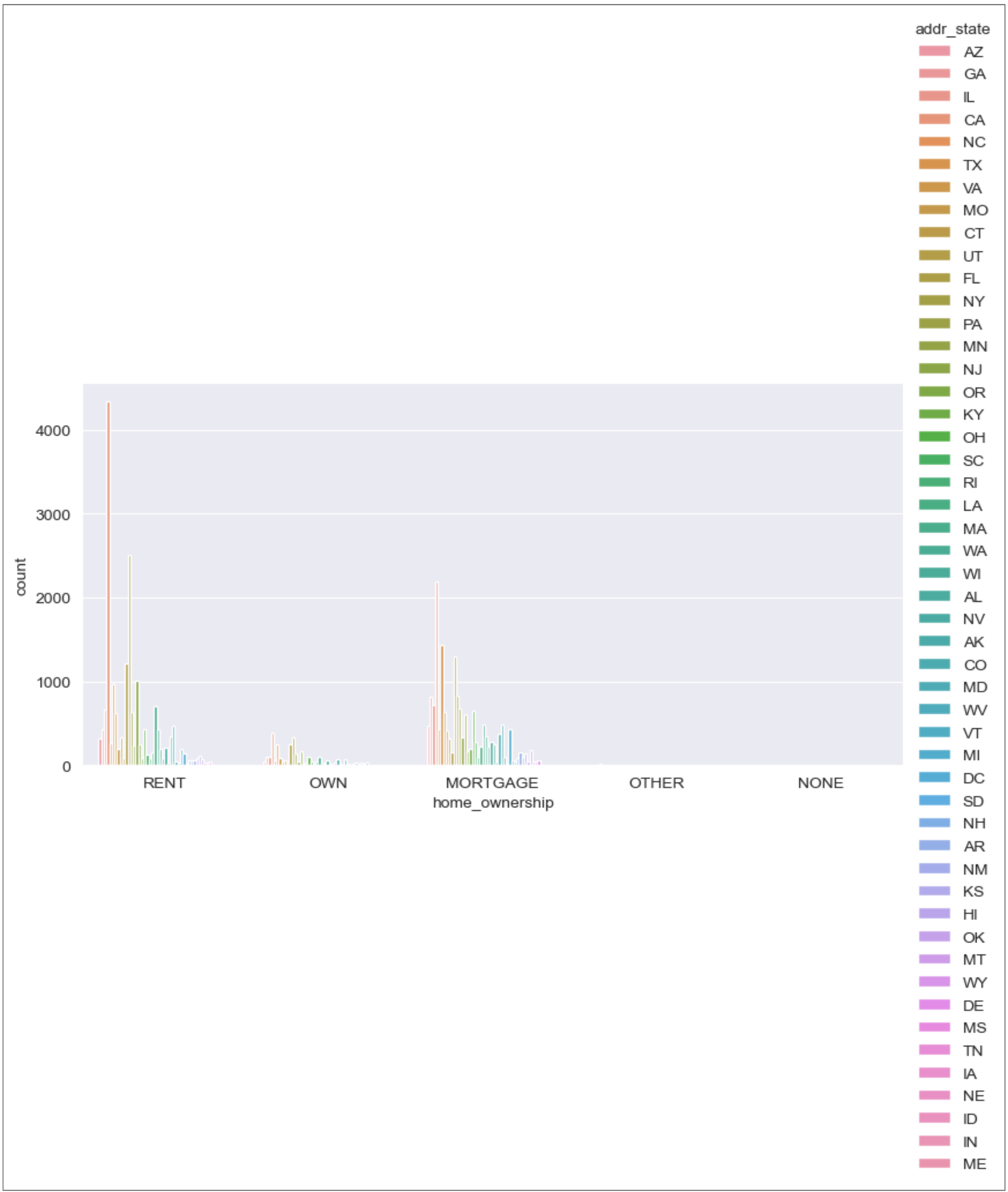

count
emp_length
10+ years
< 1 year
3 years
8 years
9 years
4 years
5 years
1 year
6 years
2 years
7 years
issue_month
December
November
October
September
August
July
June
May
April
March
February
January

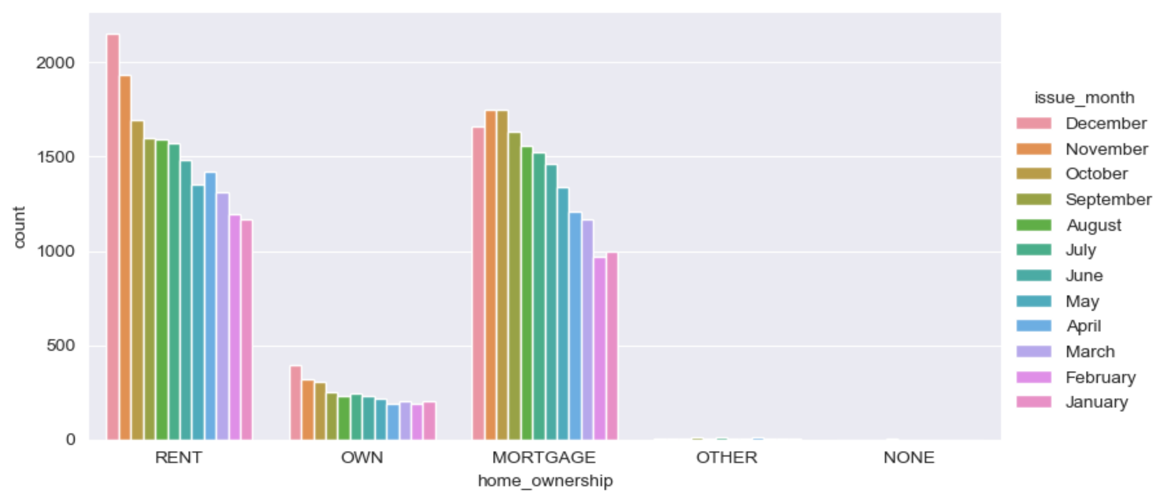

count
emp_length
issue_year
2011
2010
2009
2008
2007

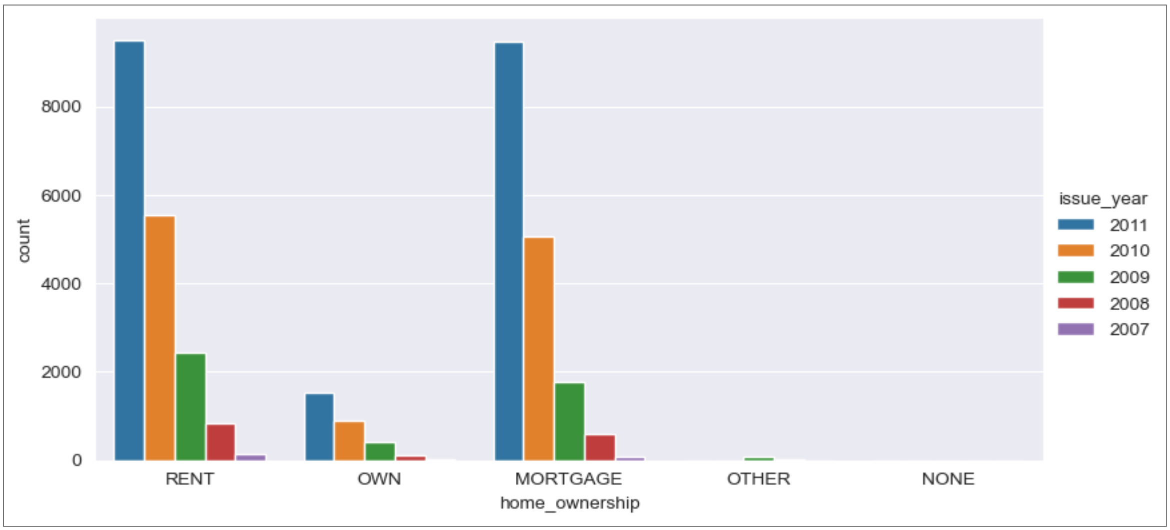

count
home_ownership
RENT
OWN
MORTGAGE
OTHER
NONE
verification_status
Verified
Source Verified
Not Verified

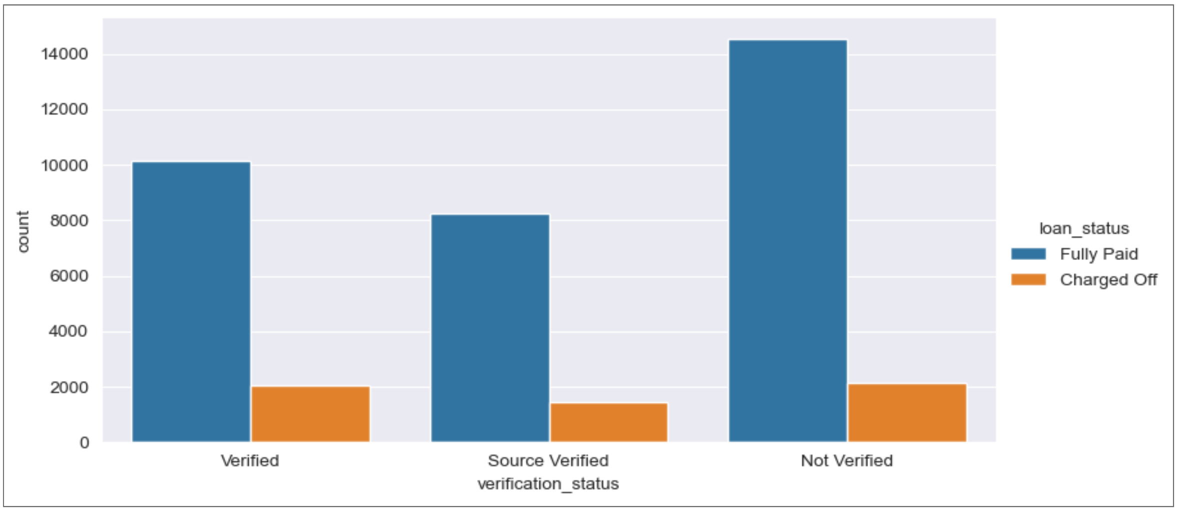

count
home_ownership
RENT
OWN
MORTGAGE
OTHER
NONE
loan_status
Fully Paid
Charged Off

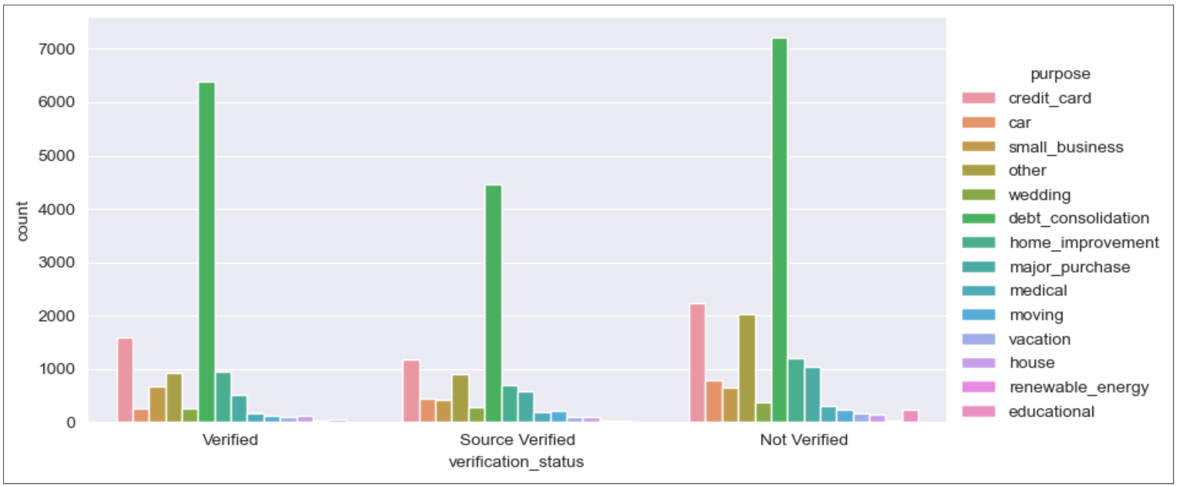

purpose
credit_card
car
small_business
other
wedding
debt_consolidation
home_improvement
major_purchase
medical
moving
vacation
house
renewable_energy
educational
count
8000
6000
4000
2000
0
RENT
OWN
MORTGAGE
OTHER
NONE
home_ownership

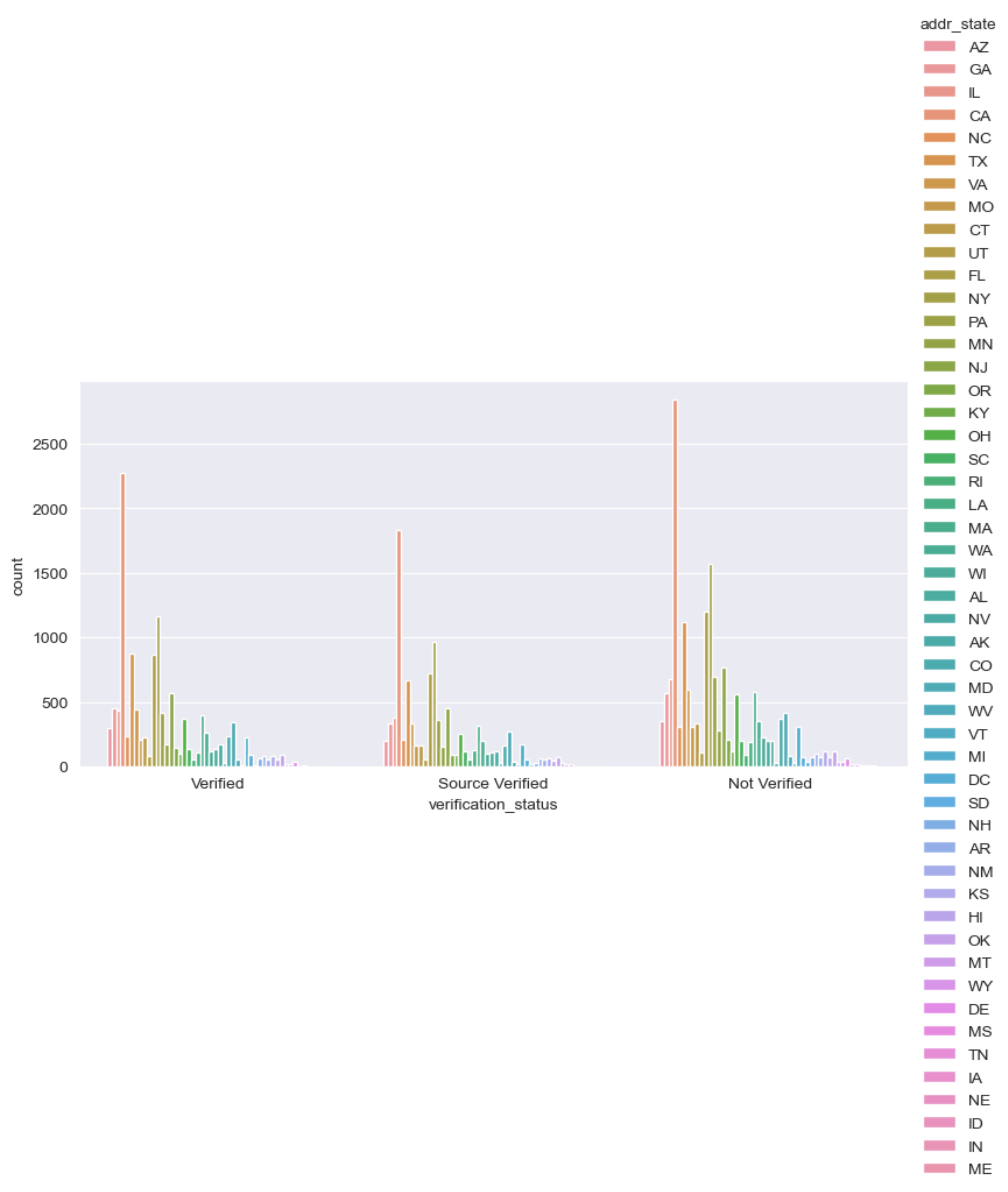

addr_state
AZ
GA
IL
CA
NC
TX
VA
MO
CT
UT
FL
NY
PA
MN
NJ
OR
KY
OH
SC
RI
LA
MA
WA
WI
AL
NV
AK
CO
MD
WV
VT
MI
DC
SD
NH
AR
NM
KS
HI
OK
MT
WY
DE
MS
TN
IA
NE
ID
IN
ME
count
4000
3000
2000
1000
0
RENT
OWN
MORTGAGE
OTHER
NONE
home_ownership

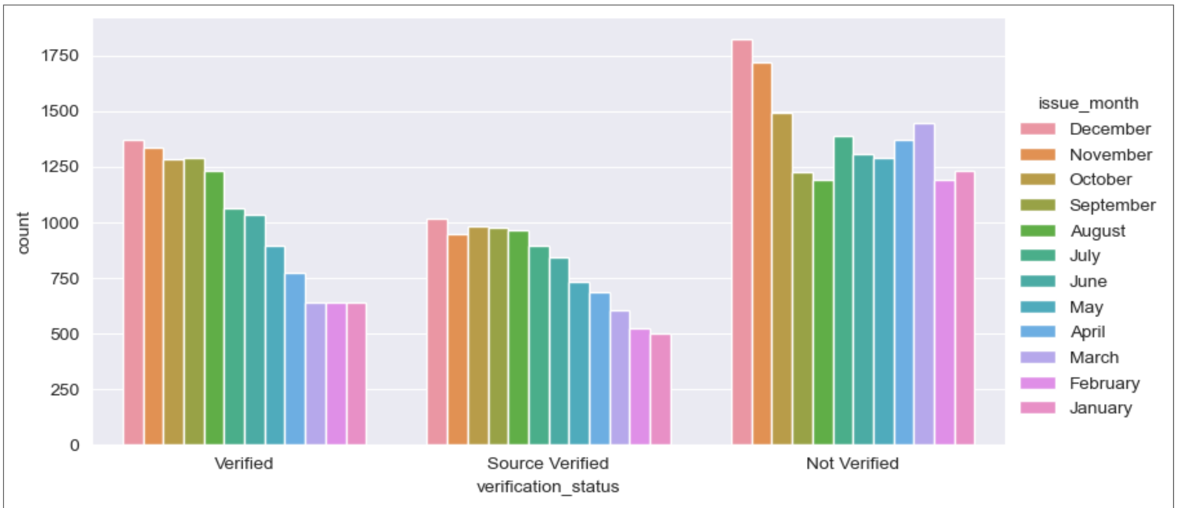

2000
1500
1000
500
0
count
RENT
OWN
MORTGAGE
OTHER
NONE
home_ownership
issue_month
December
November
October
September
August
July
June
May
April
March
February
January

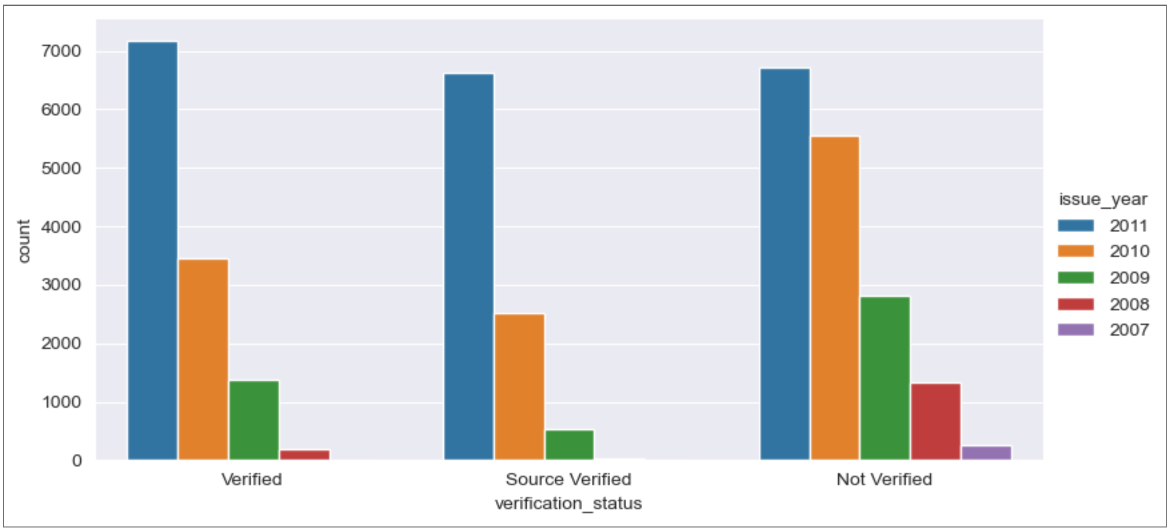

8000
6000
4000
2000
0
count
RENT
OWN
MORTGAGE
OTHER
NONE
home_ownership
issue_year
2011
2010
2009
2008
2007

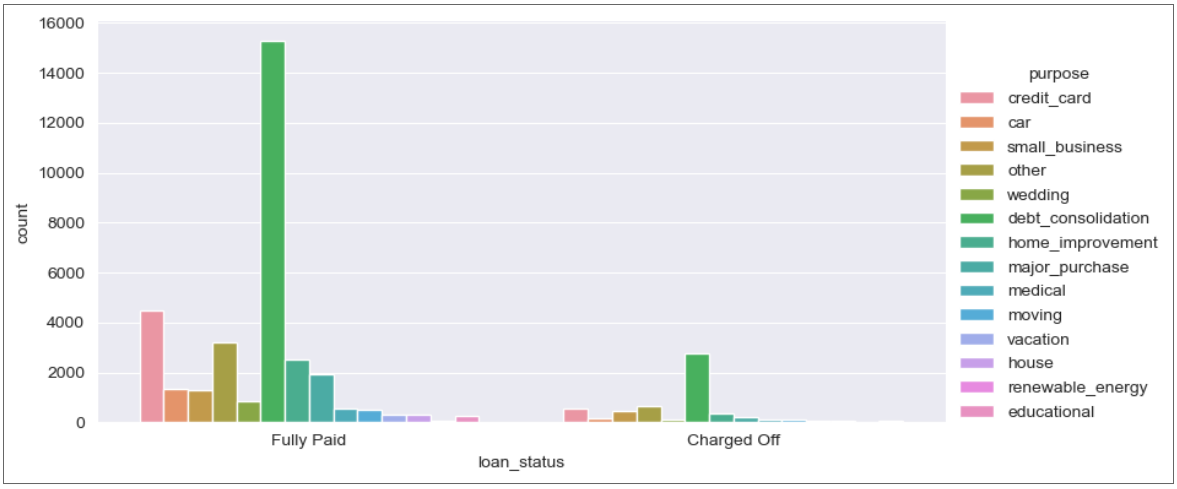

14000
12000
10000
8000
6000
4000
2000
0
count
Verified
Source Verified
Not Verified
verification_status
loan_status
Fully Paid
Charged Off

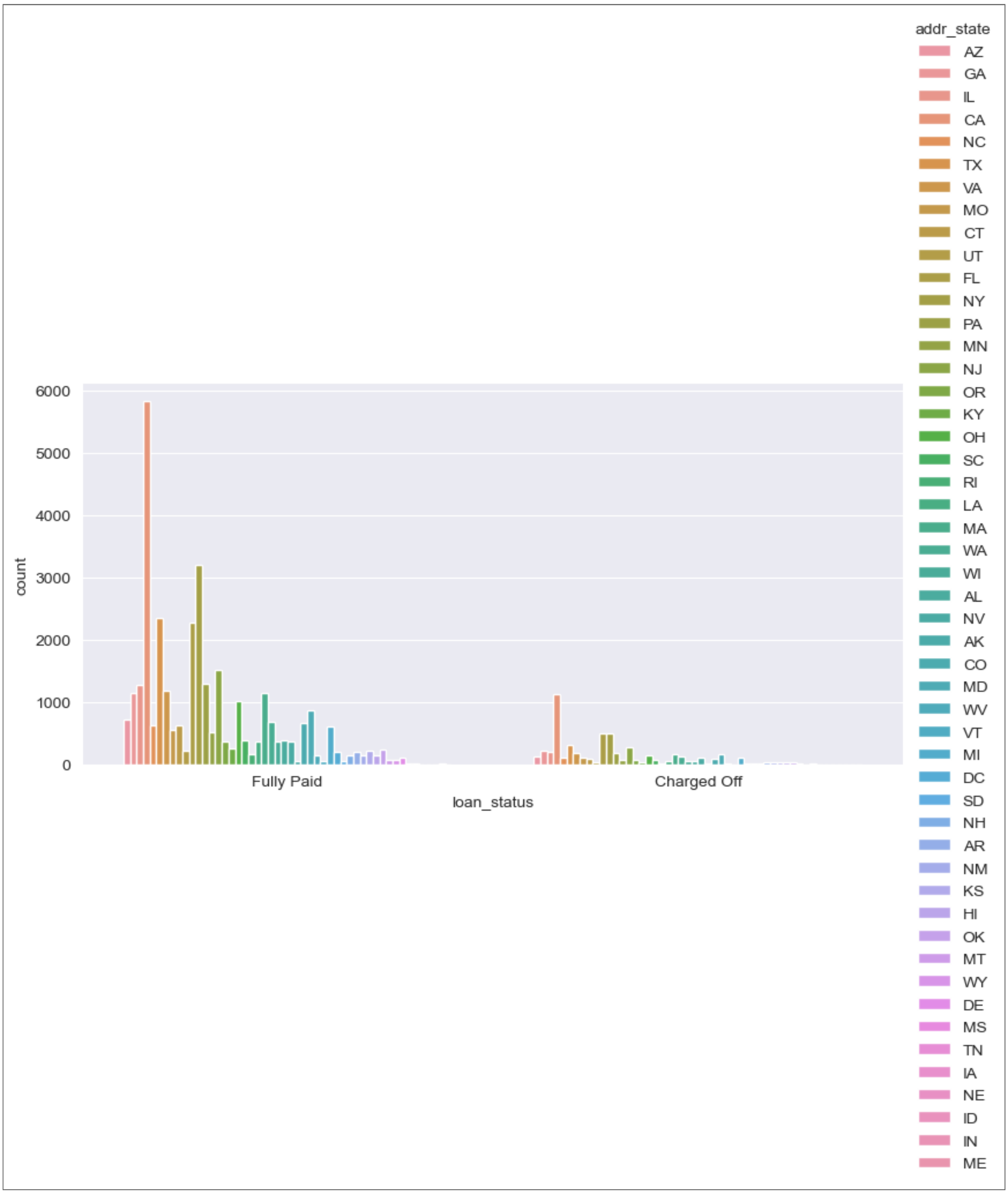

7000
6000
5000
4000
3000
2000
1000
0
count
Verified
Source Verified
Not Verified
verification_status
purpose
credit_card
car
small_business
other
wedding
debt_consolidation
home_improvement
major_purchase
medical
moving
vacation
house
renewable_energy
educational

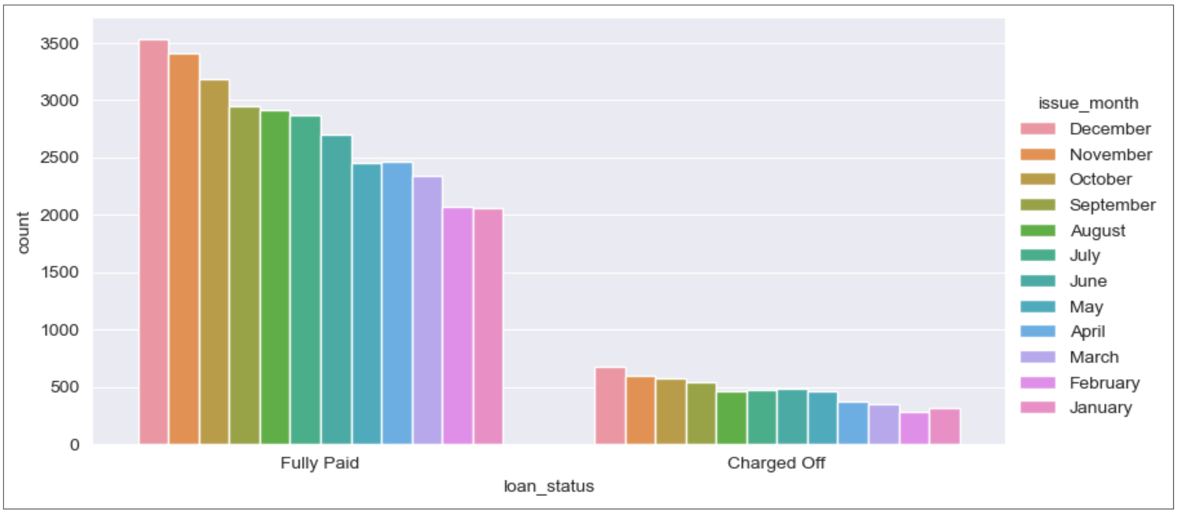

addr_state
AZ
GA
IL
CA
NC
TX
VA
MO
CT
UT
FL
NY
PA
MN
NJ
OR
KY
OH
SC
RI
LA
MA
WA
WI
AL
NV
AK
CO
MD
WV
VT
MI
DC
SD
NH
AR
NM
KS
HI
OK
MT
WY
DE
MS
TN
IA
NE
ID
IN
ME
count
2500
2000
1500
1000
500
0
Verified
Source Verified
Not Verified
verification_status

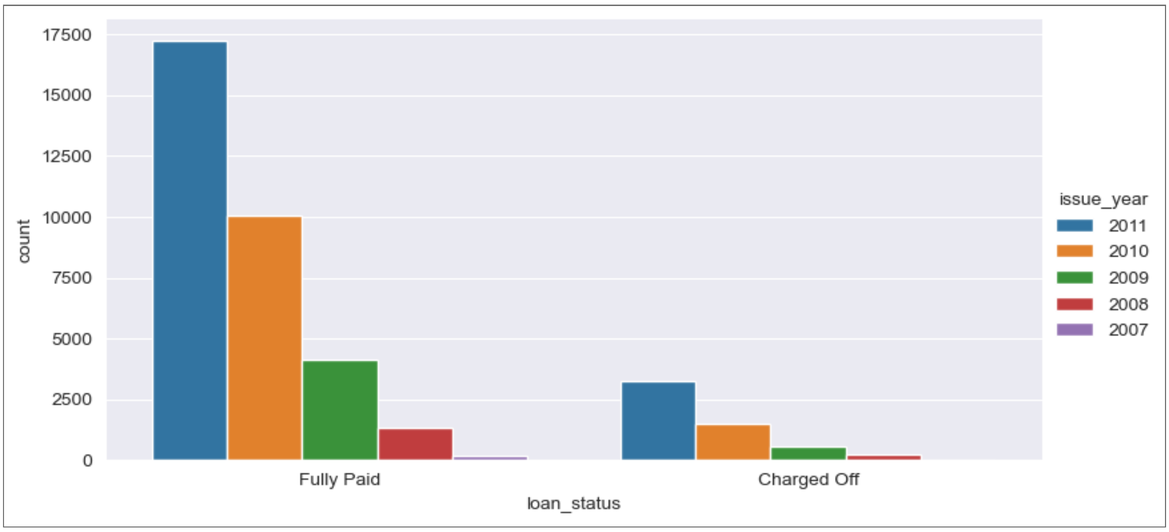

count
1750
1500
1250
1000
750
500
250
0
Verified
Source Verified
Not Verified
verification_status
issue_month
December
November
October
September
August
July
June
May
April
March
February
January


7000
6000
5000
4000
3000
2000
1000
0
count
Verified
Source Verified
Not Verified
verification_status
issue_year
2011
2010
2009
2008
2007

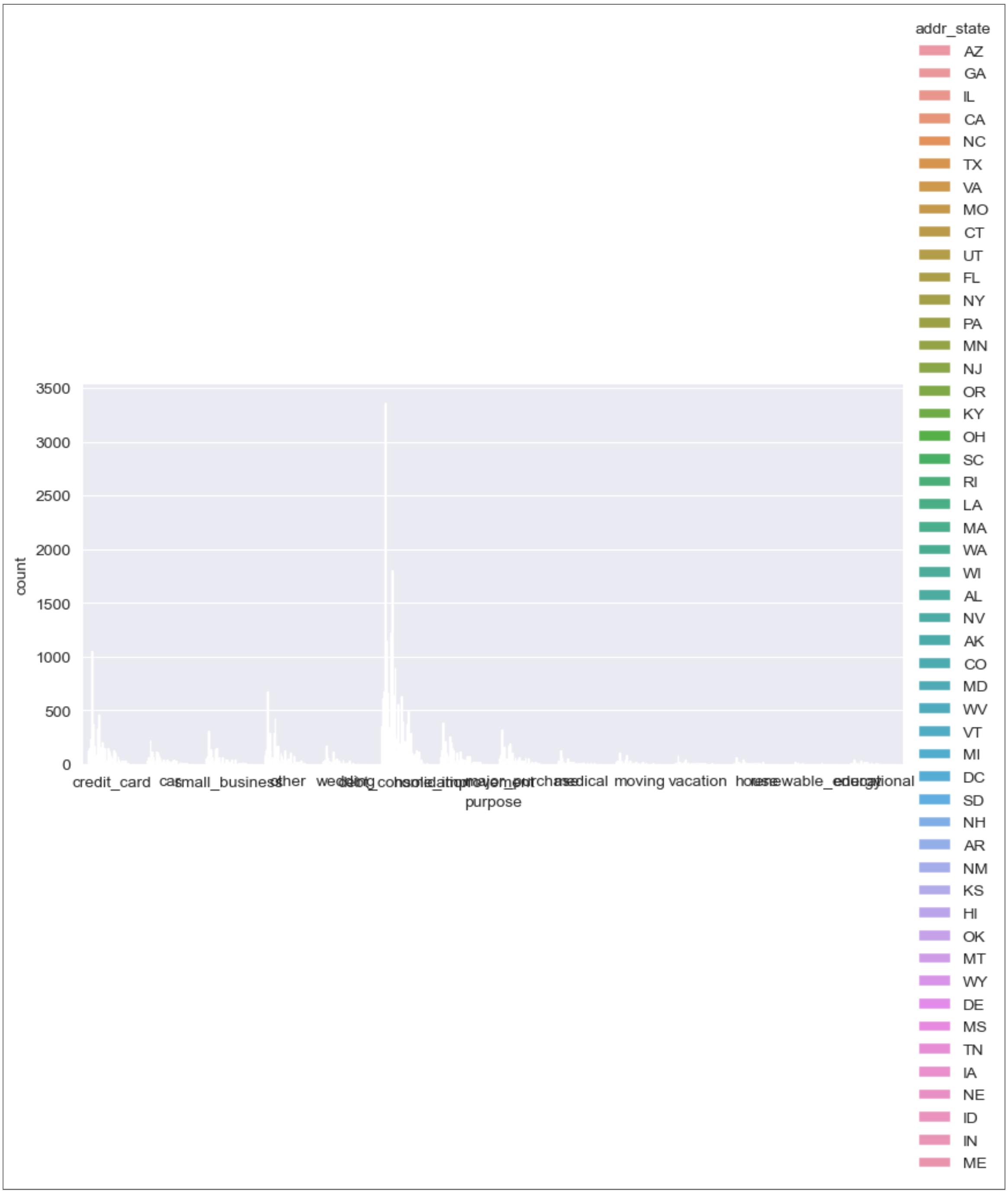

16000
14000
12000
10000
8000
6000
4000
2000
0
count
Fully Paid
Charged Off
loan_status
purpose
credit_card
car
small_business
other
wedding
debt_consolidation
home_improvement
major_purchase
medical
moving
vacation
house
renewable_energy
educational


addr_state
AZ
GA
IL
CA
NC
TX
VA
MO
CT
UT
FL
NY
PA
MN
NJ
OR
KY
OH
SC
RI
LA
MA
WA
WI
AL
NV
AK
CO
MD
WV
VT
MI
DC
SD
NH
AR
NM
KS
HI
OK
MT
WY
DE
MS
TN
IA
NE
ID
IN
ME
count
0
1000
2000
3000
4000
5000
6000
Fully Paid
Charged Off
loan_status

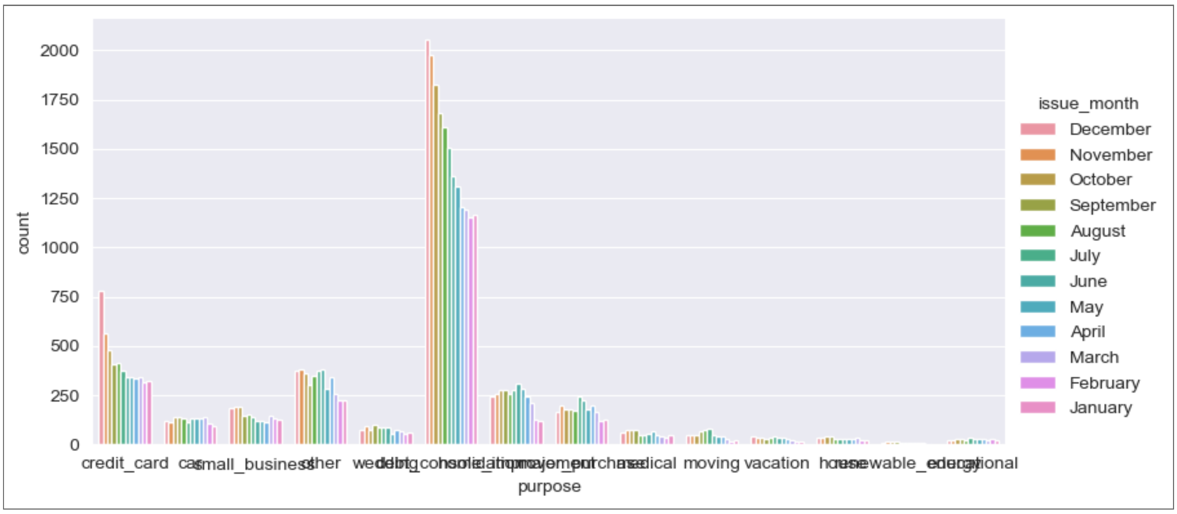

count
0
500
1000
1500
2000
2500
3000
3500
Fully Paid
Charged Off
loan_status
issue_month
December
November
October
September
August
July
June
May
April
March
February
January

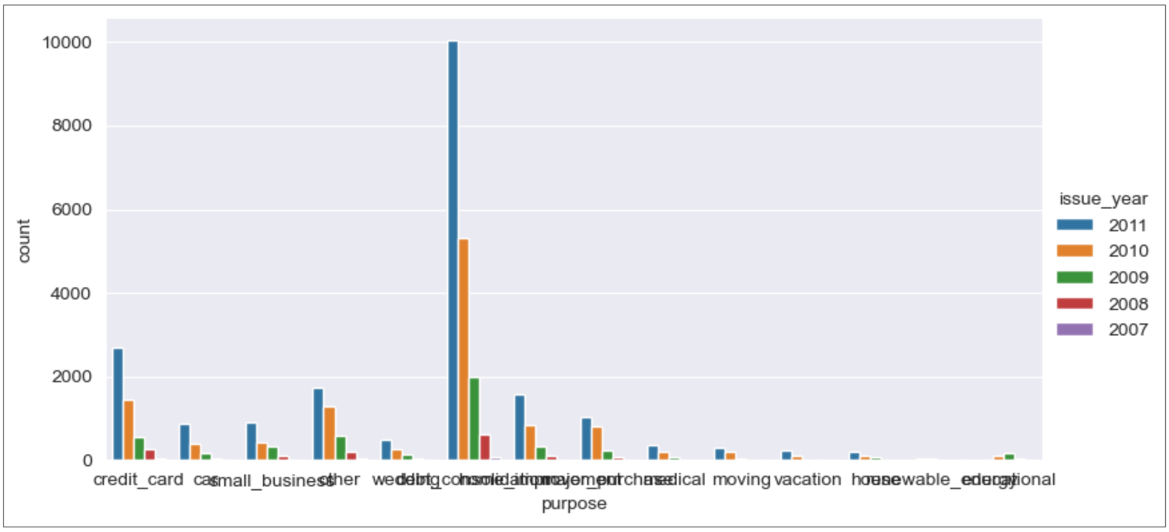


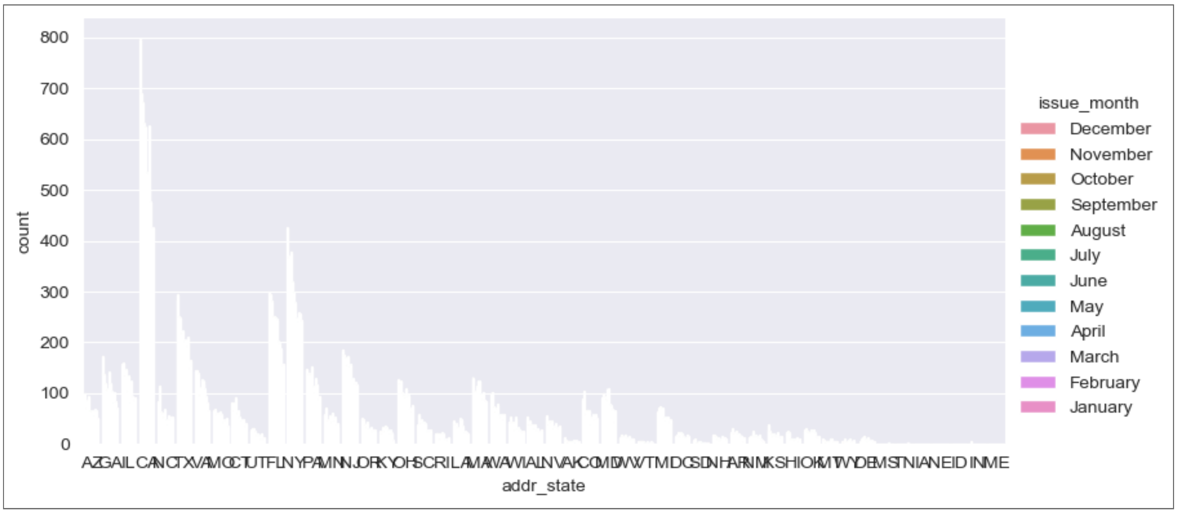
17500
15000
12500
10000
7500
5000
2500
0
count
Fully Paid
Charged Off
loan_status
issue_year
2011
2010
2009
2008
2007

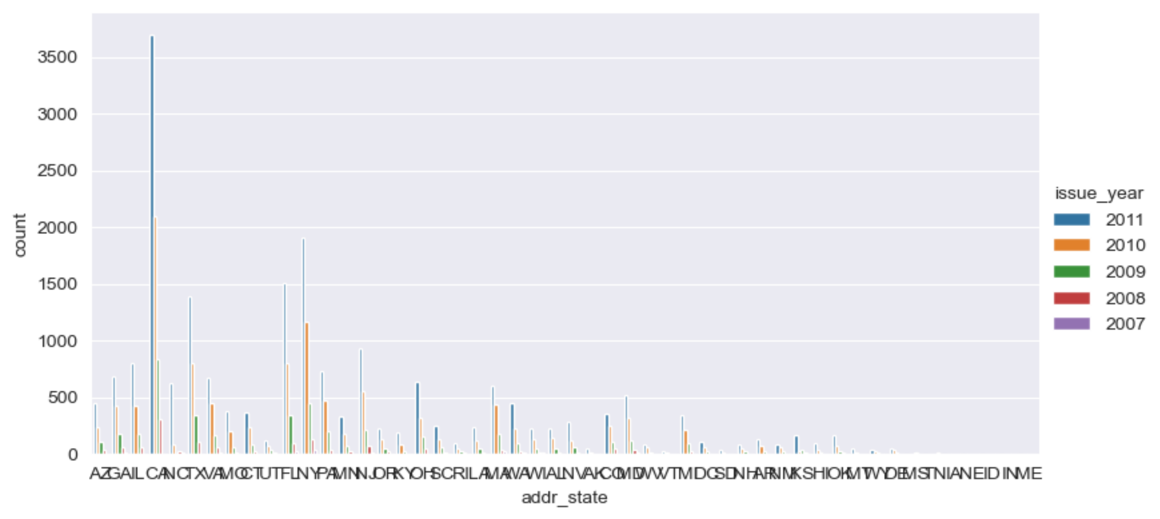


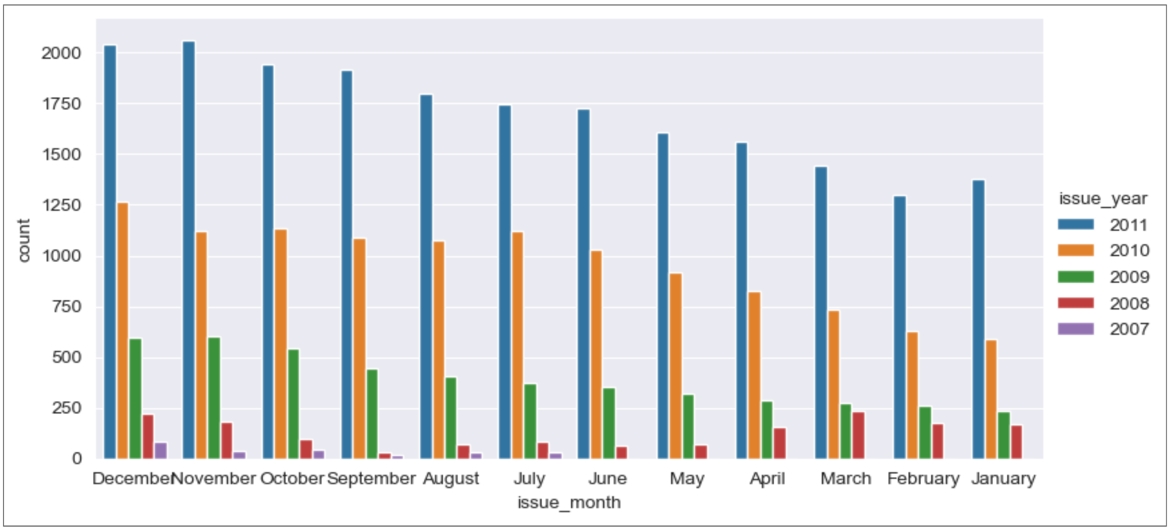
addr_state
AZ
GA
IL
CA
NC
TX
VA
MO
CT
UT
FL
NY
PA
MN
NJ
OR
KY
OH
SC
RI
LA
MA
WA
WI
AL
NV
AK
CO
MD
WV
VT
MI
DC
SD
NH
AR
NM
KS
HI
OK
MT
WY
DE
MS
TN
IA
NE
ID
IN
ME
3500
3000
2500
2000
1500
1000
500
0
count
credit_card
medical
moving
vacation
purpose


count
2000
1750
1500
1250
1000
750
500
250
0
credit_card
medical
moving
vacation
purpose
issue_month
December
November
October
September
August
July
June
May
April
March
February
January


count
10000
8000
6000
4000
2000
0
credit_card
medical
moving
vacation
purpose
issue_year
2011
2010
2009
2008
2007

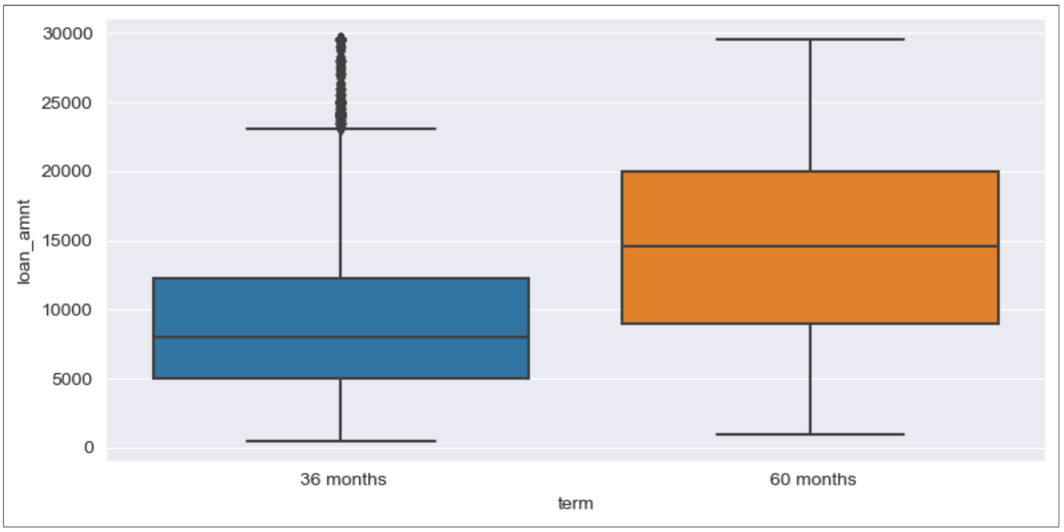

count
800
700
600
500
400
300
200
100
0
addr_state
issue_month
December
November
October
September
August
July
June
May
April
March
February
January

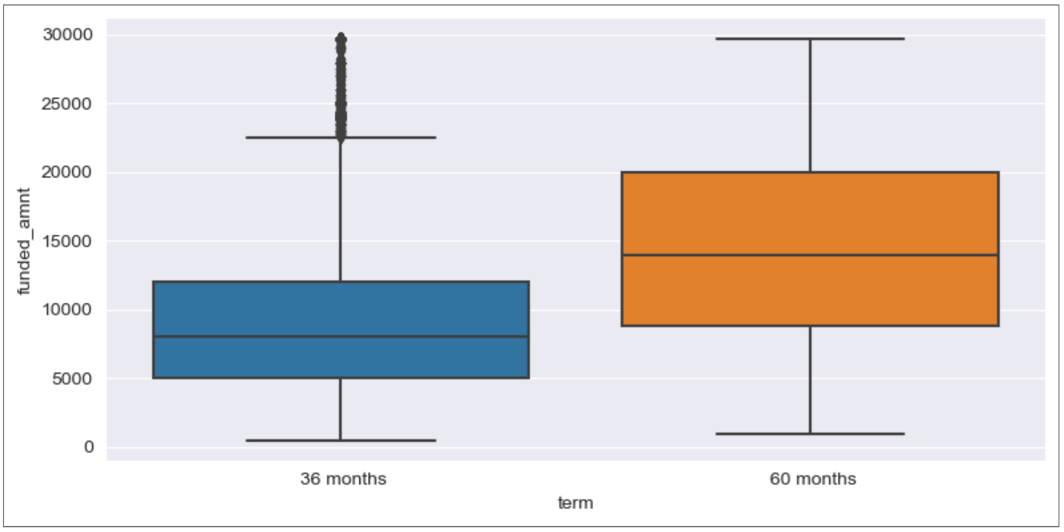

count
3500
3000
2500
2000
1500
1000
500
0
addr_state
issue_year
2011
2010
2009
2008
2007

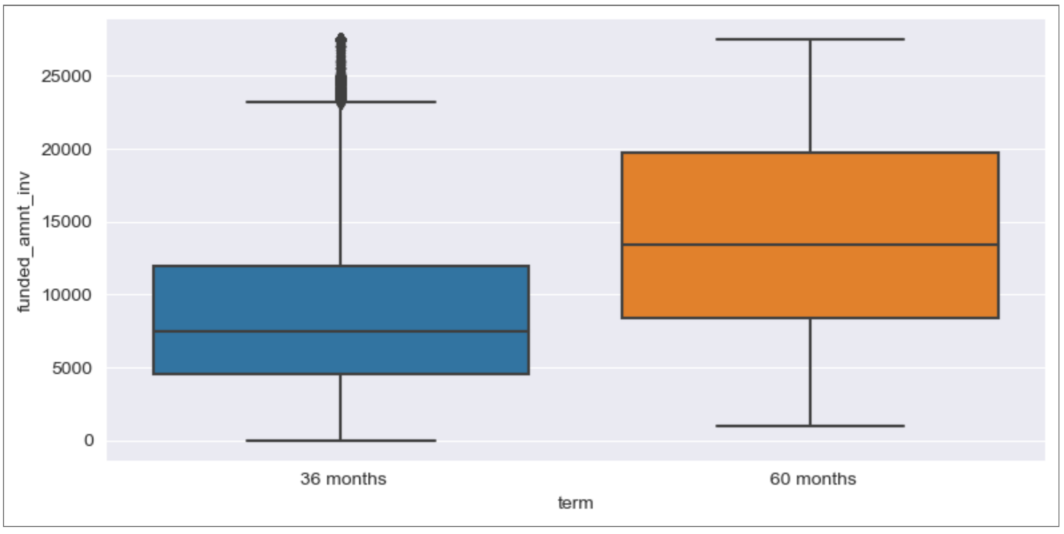

count
2000
1750
1500
1250
1000
750
500
250
0
Decembe
November
October
September
August
July
June
May
April
March
February
January
issue_month
issue_year
2011
2010
2009
2008
2007

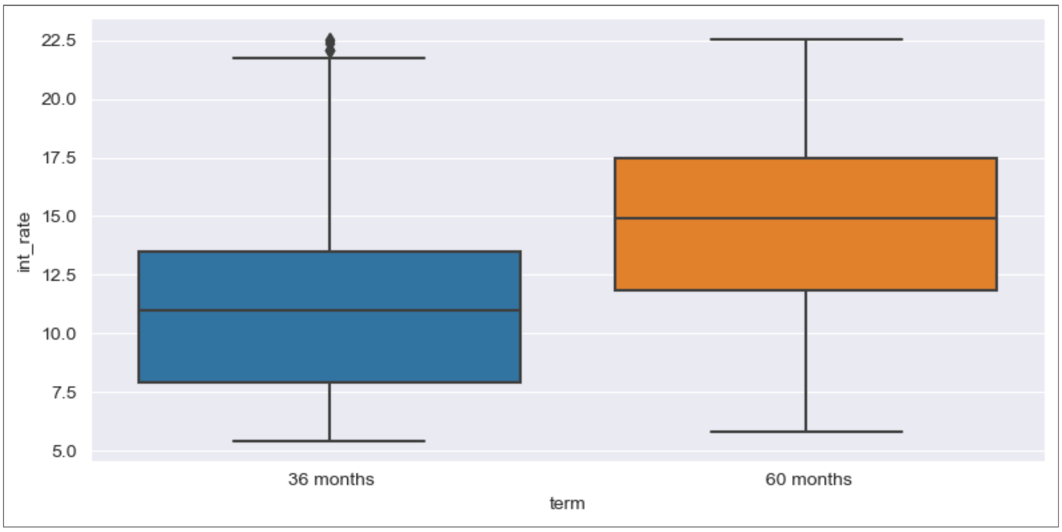
From the bar charts, the following inferences are drawn:
Most loans are for a term of 36 months.
Loans are mostly given to borrowers with grade A, B, or C.
The largest number of borrowers have been employed for 10+ years.
Majority of the borrowers are living in rented places or in their own dwellings.
There is substantial difference in the status of verification for loans drawn over a term of 36 months, compared with those drawn over 60 months.
Most loans are fully paid.
Debt consolidation is the most common purpose for taking a loan.
CA, NY, and TX are the states with the highest number of loan applicants.
The highest number of loans are issued in the last three months of the year.
Most loans were issued in the years 2011 and 2010 in the current dataset used in this study.
Overall, these inferences only reinforce the findings already drawn above (in particular, the inferences from univariate analysis).

Box plots are used to further visualize the relationships between pairs of categorical and numeric variables

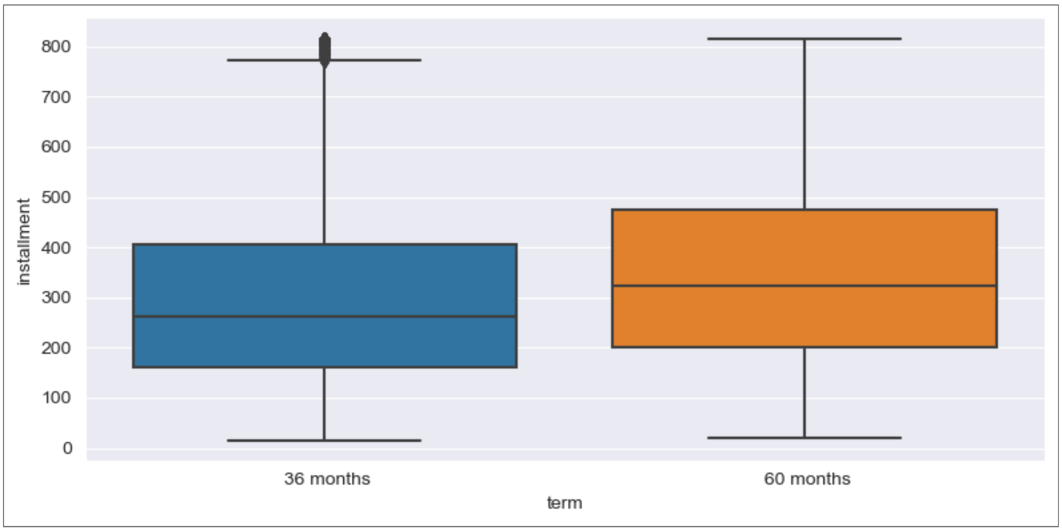

30000
25000
20000
15000
10000
5000
0
loan_amnt
36 months
60 months
term


30000
25000
20000
15000
10000
5000
0
funded_amnt
36 months
60 months
term

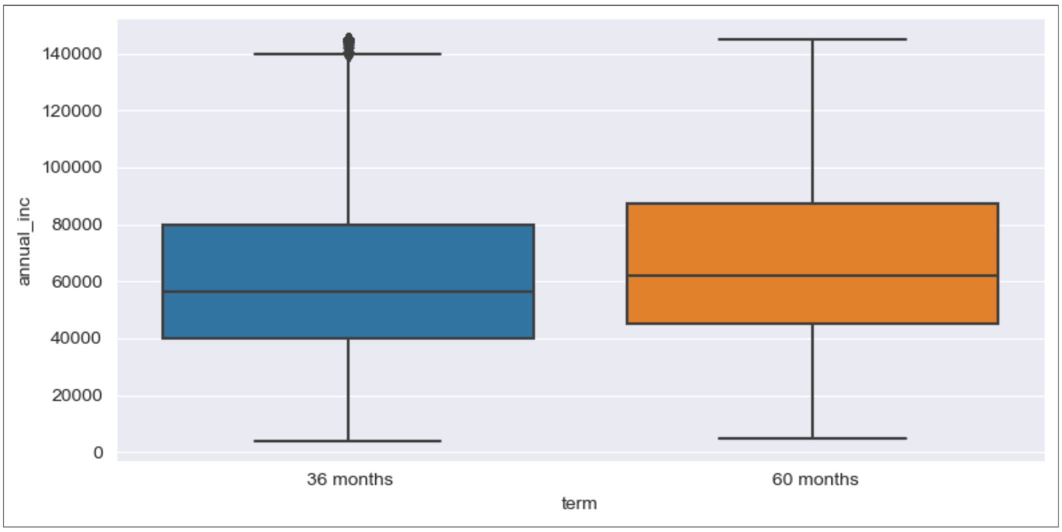

25000
20000
15000
10000
5000
0
funded_amnt_inv
36 months
60 months
term



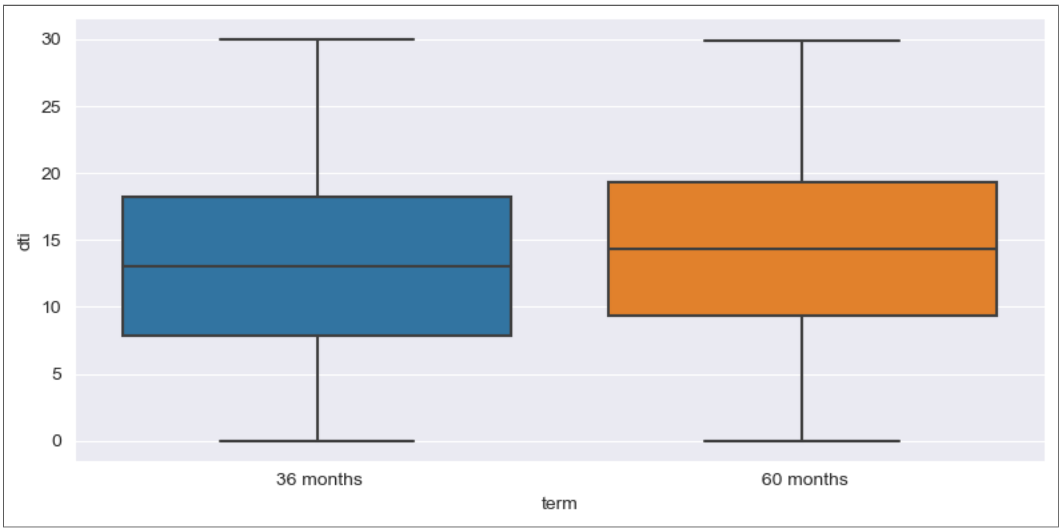
int_rate
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
36 months
60 months
term

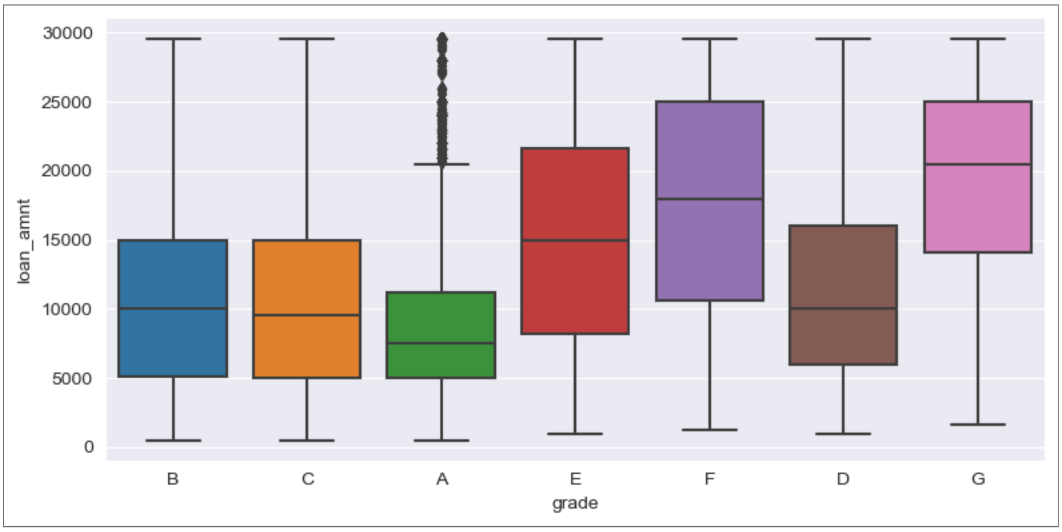


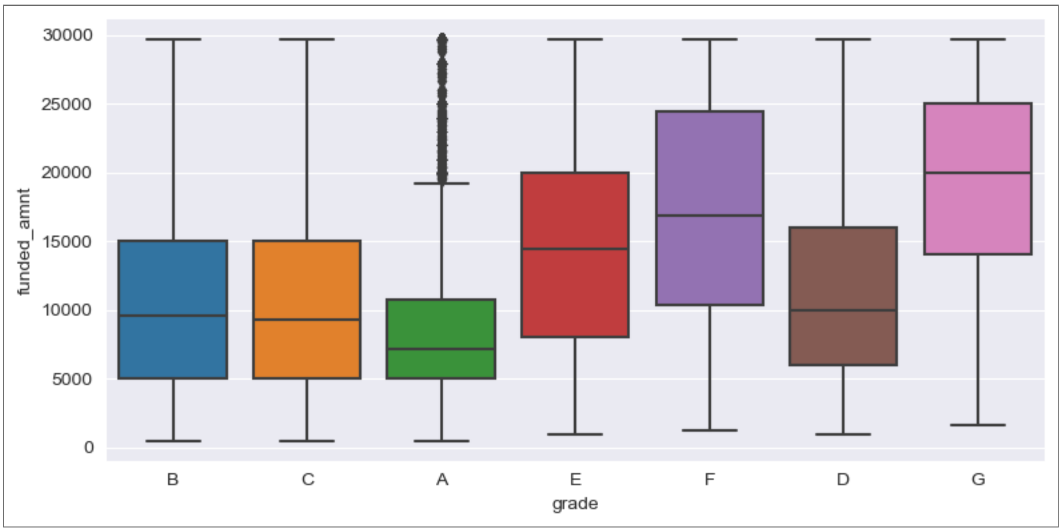
installment
800
700
600
500
400
300
200
100
0
36 months
60 months
term

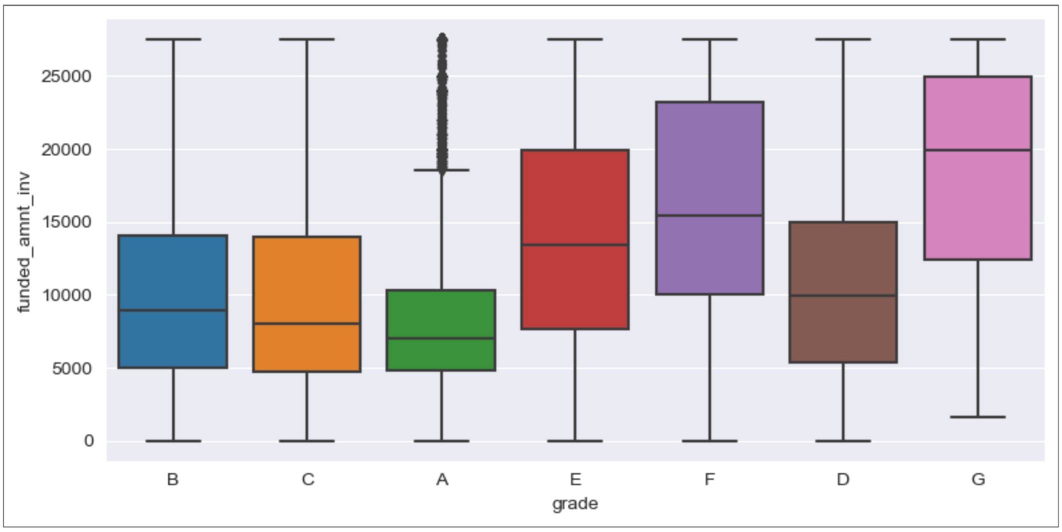


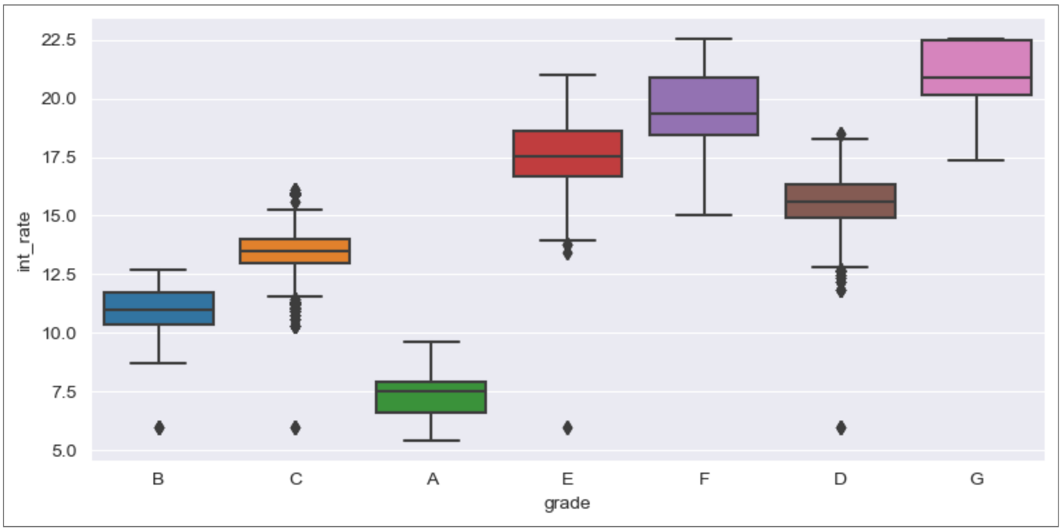
annual_inc
140000
120000
100000
80000
60000
40000
20000
0
36 months
60 months
term

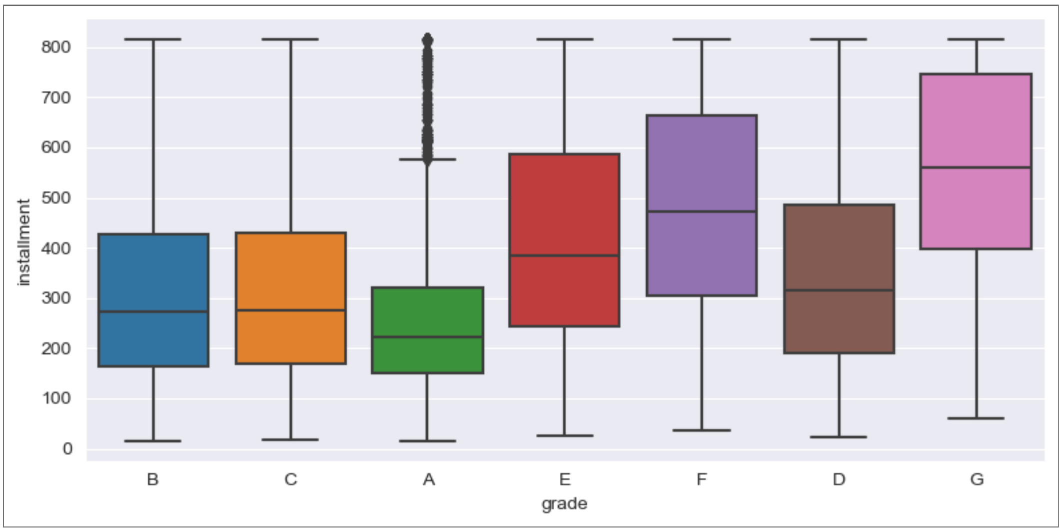

30
25
20
15
10
5
0
dti
36 months
60 months
term

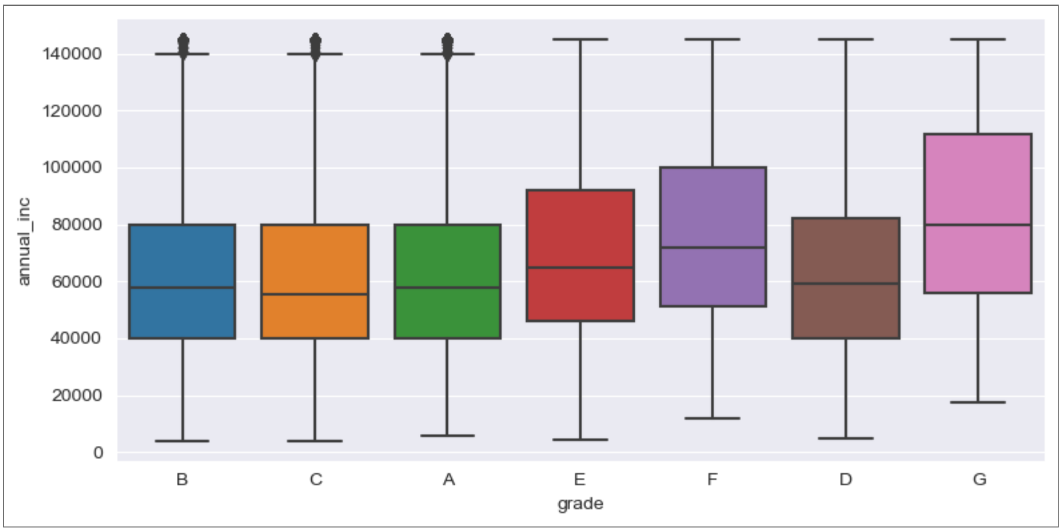

30000
25000
20000
15000
10000
5000
0
loan_amnt
B
C
A
E
F
D
G
grade

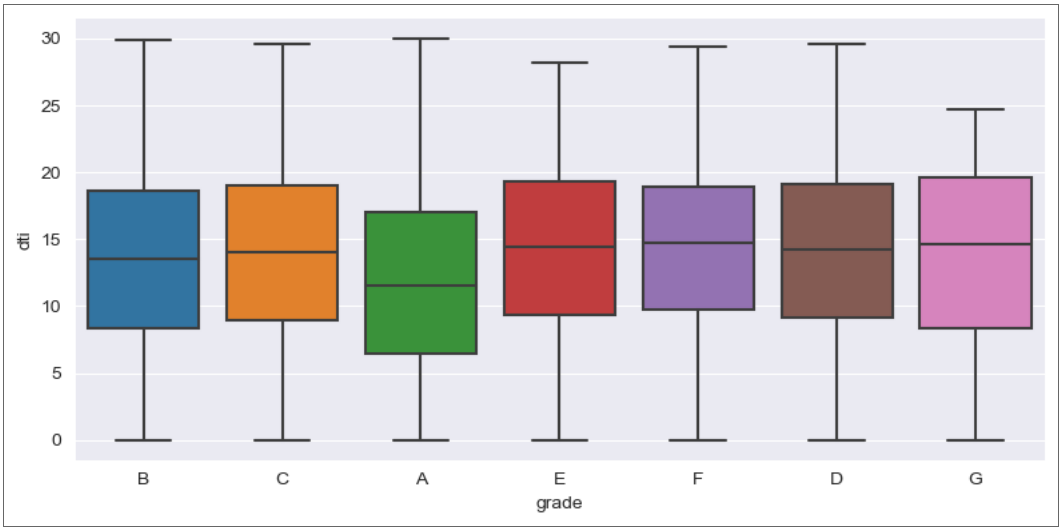

30000
25000
20000
15000
10000
5000
0
funded_amnt
B
C
A
E
F
D
G
grade

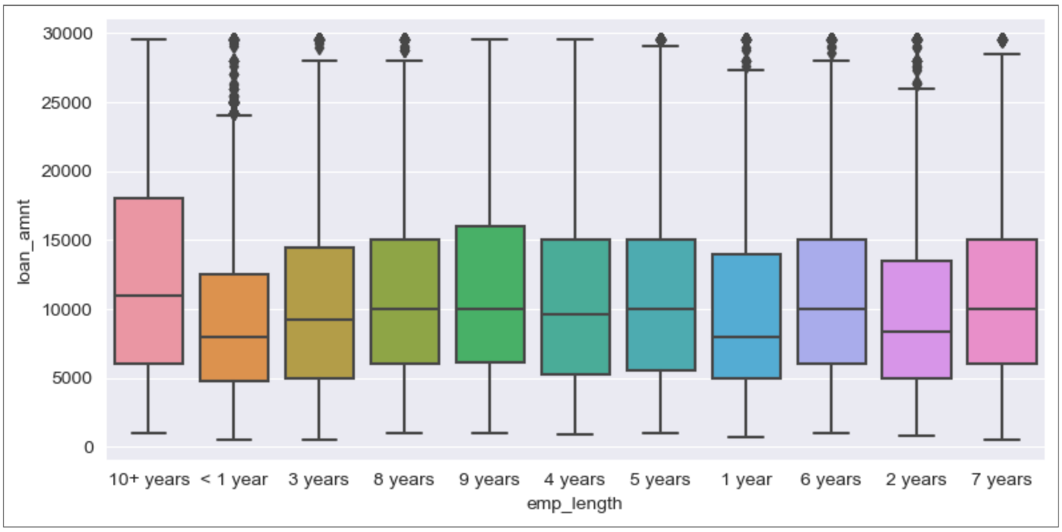

funded_amnt_inv
0
5000
10000
15000
20000
25000
B
C
A
E
F
D
G
grade

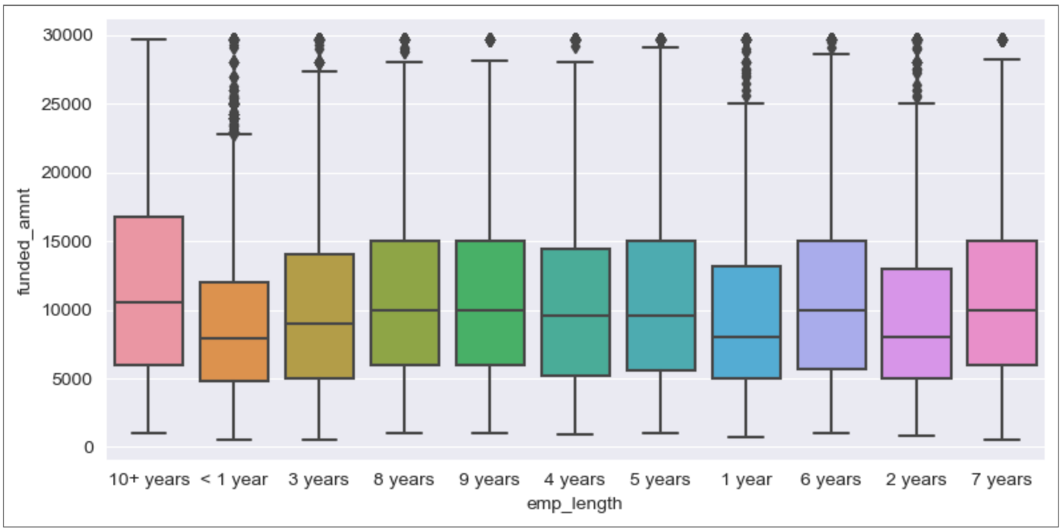

int_rate
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5
B
C
A
E
F
D
G
grade

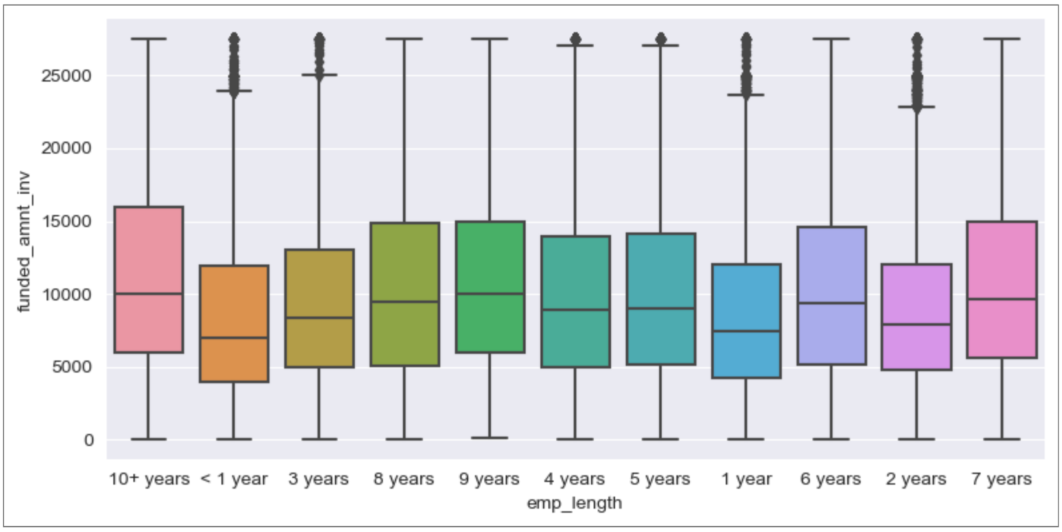

installment
0
100
200
300
400
500
600
700
800
B
C
A
E
F
D
G
grade

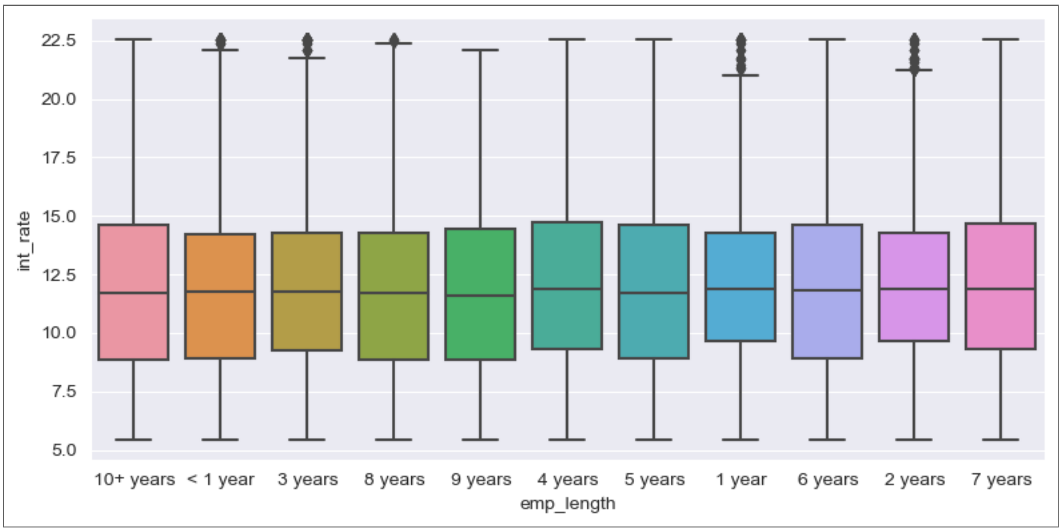

140000
120000
100000
80000
60000
40000
20000
0
annual_inc
B
C
A
E
F
D
G
grade

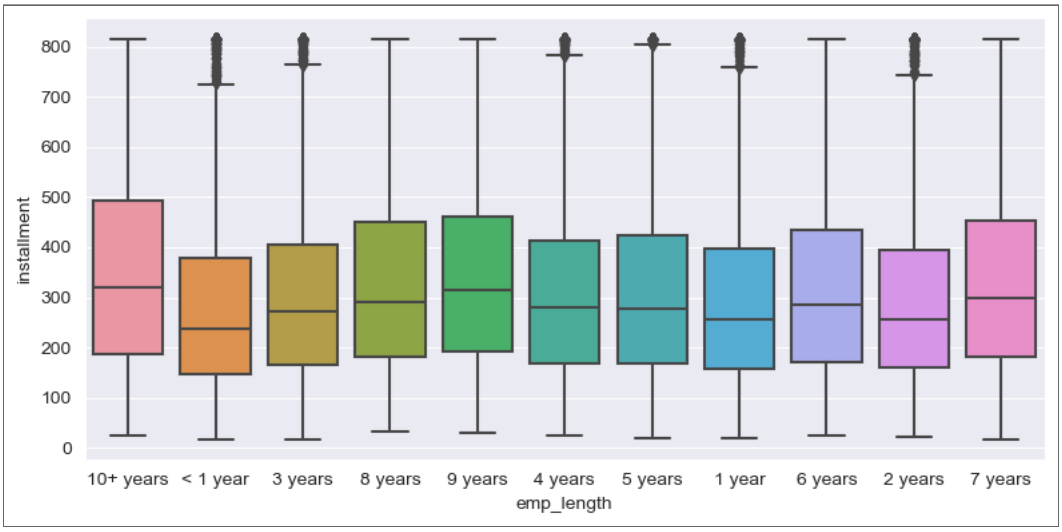

30
25
20
15
10
5
0
dti
B
C
A
E
F
D
G
grade


30000
25000
20000
15000
10000
5000
0
loan_amnt
10+ years
< 1 year
3 years
8 years
9 years
4 years
5 years
1 year
6 years
2 years
7 years
emp_length

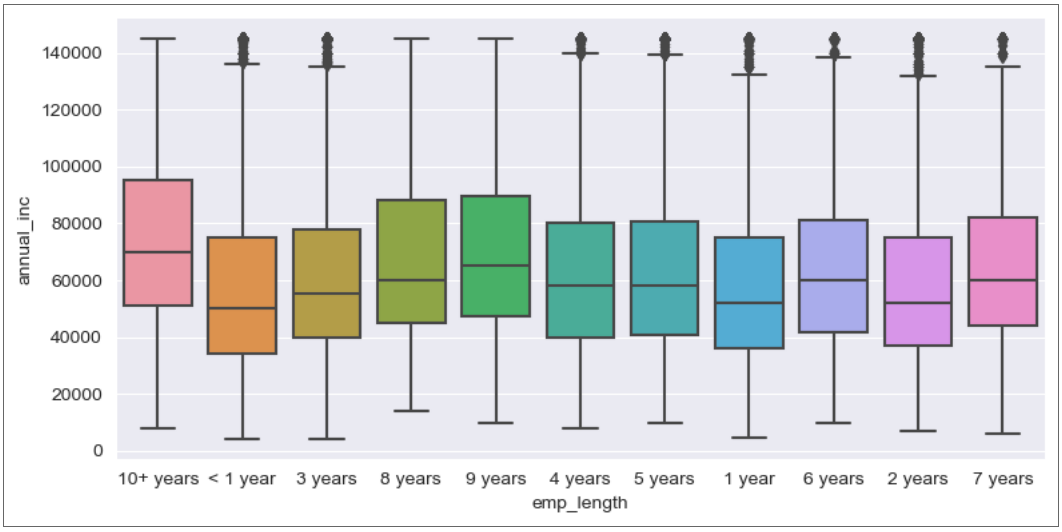

funded_amnt
30000
25000
20000
15000
10000
5000
0
10+ years
< 1 year
3 years
8 years
9 years
4 years
5 years
1 year
6 years
2 years
7 years
emp_length


funded_amnt_inv
25000
20000
15000
10000
5000
0
10+ years
< 1 year
3 years
8 years
9 years
4 years
5 years
1 year
6 years
2 years
7 years
emp_length

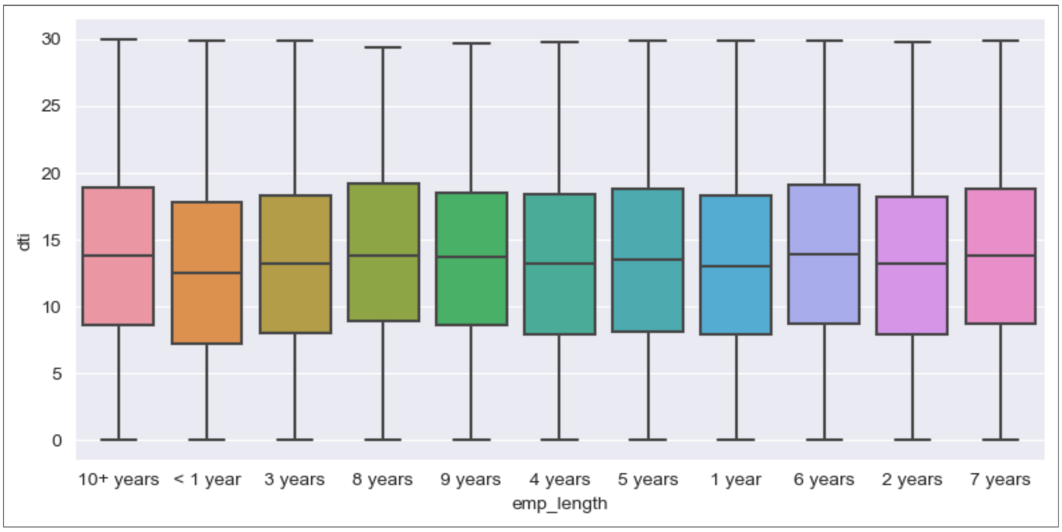

int_rate
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
10+ years
< 1 year
3 years
8 years
9 years
4 years
5 years
1 year
6 years
2 years
7 years
emp_length



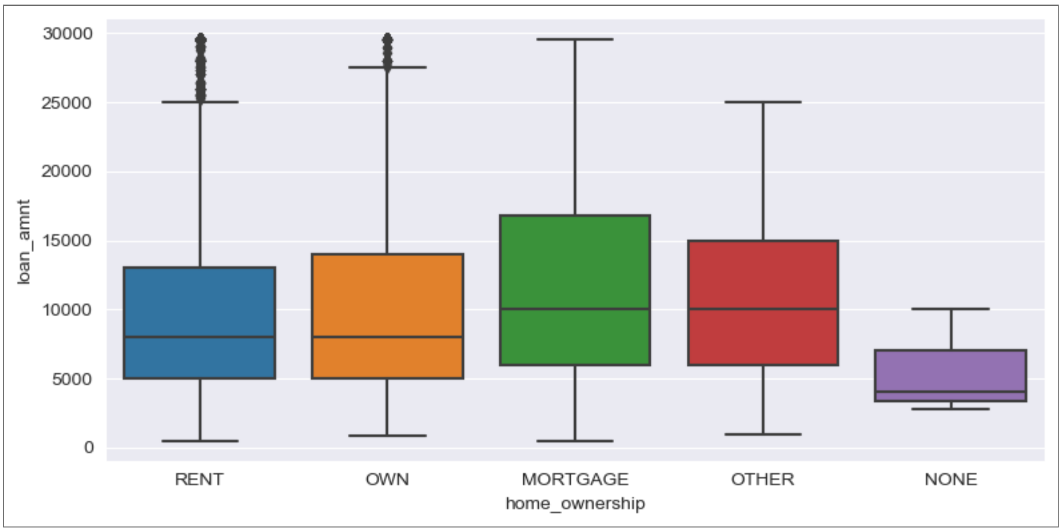
800
700
600
500
400
300
200
100
0
installment
10+ years
< 1 year
3 years
8 years
9 years
4 years
5 years
1 year
6 years
2 years
7 years
emp_length

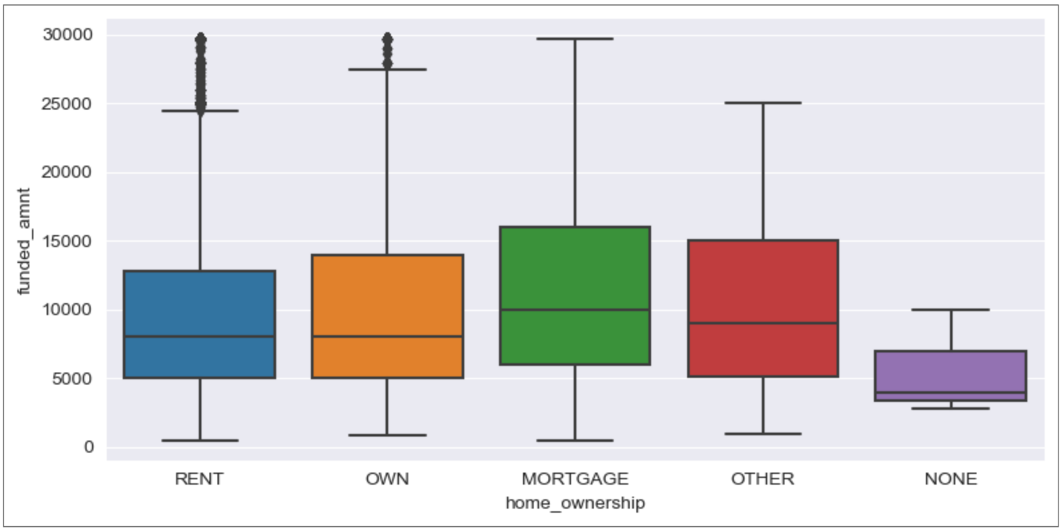


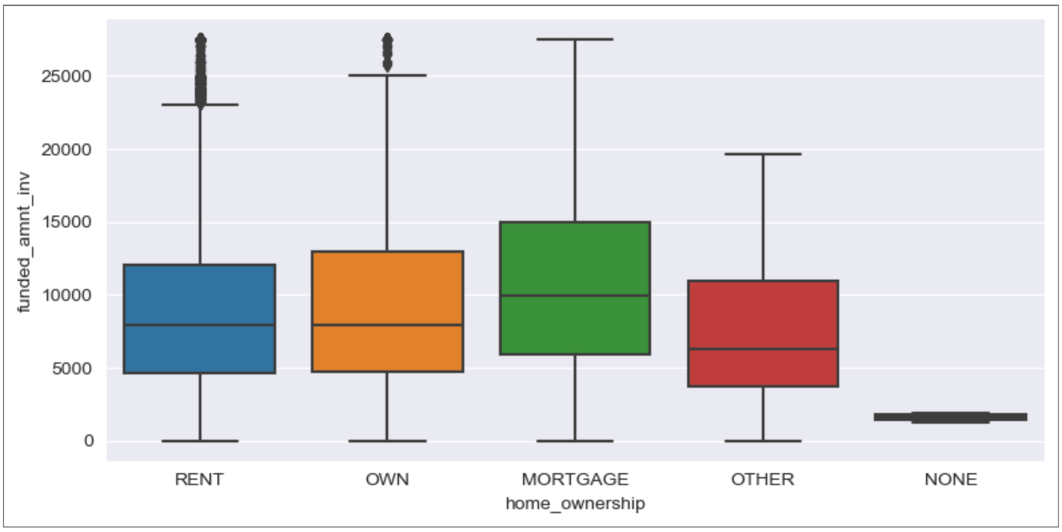
140000
120000
100000
80000
60000
40000
20000
0
annual_inc
10+ years
< 1 year
3 years
8 years
9 years
4 years
5 years
1 year
6 years
2 years
7 years
emp_length

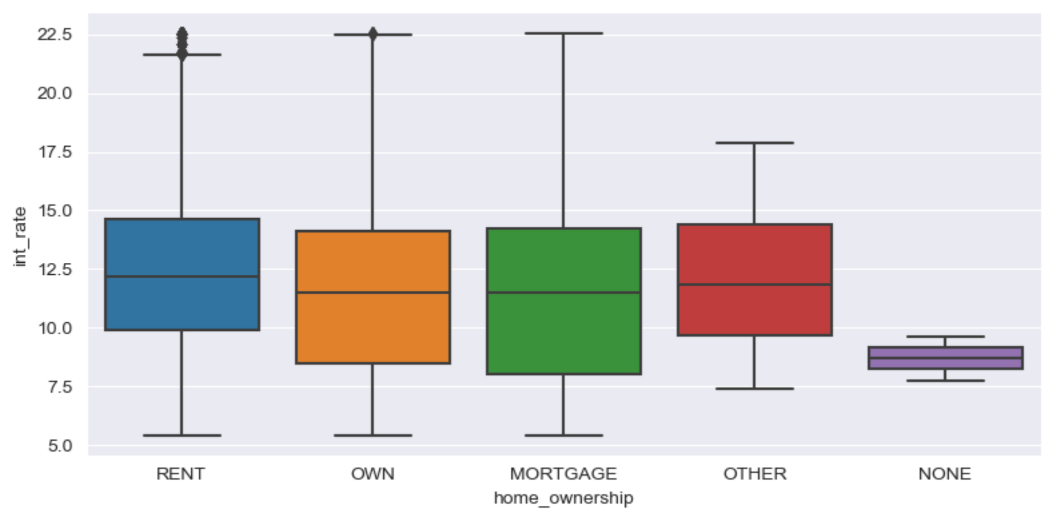


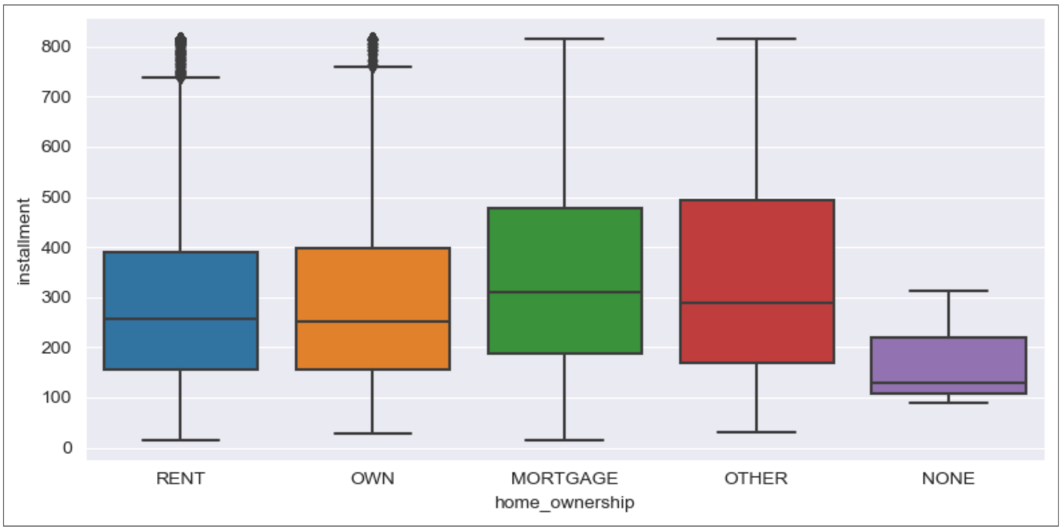
30
25
20
15
10
5
0
dti
10+ years
< 1 year
3 years
8 years
9 years
4 years
5 years
1 year
6 years
2 years
7 years
emp_length

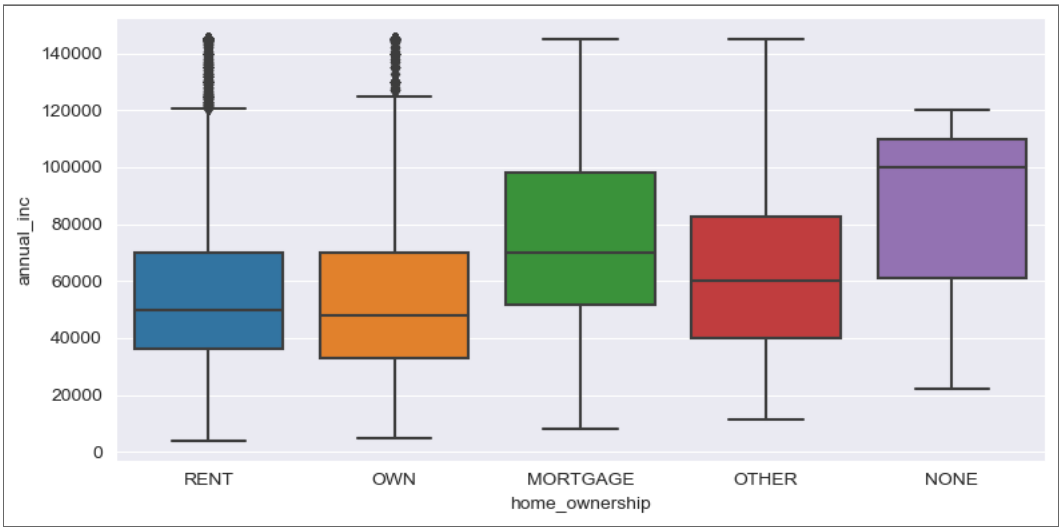

30000
25000
20000
15000
10000
5000
0
loan_amnt
RENT
OWN
MORTGAGE
OTHER
NONE
home_ownership


30000
25000
20000
15000
10000
5000
0
funded_amnt
RENT
OWN
MORTGAGE
OTHER
NONE
home_ownership

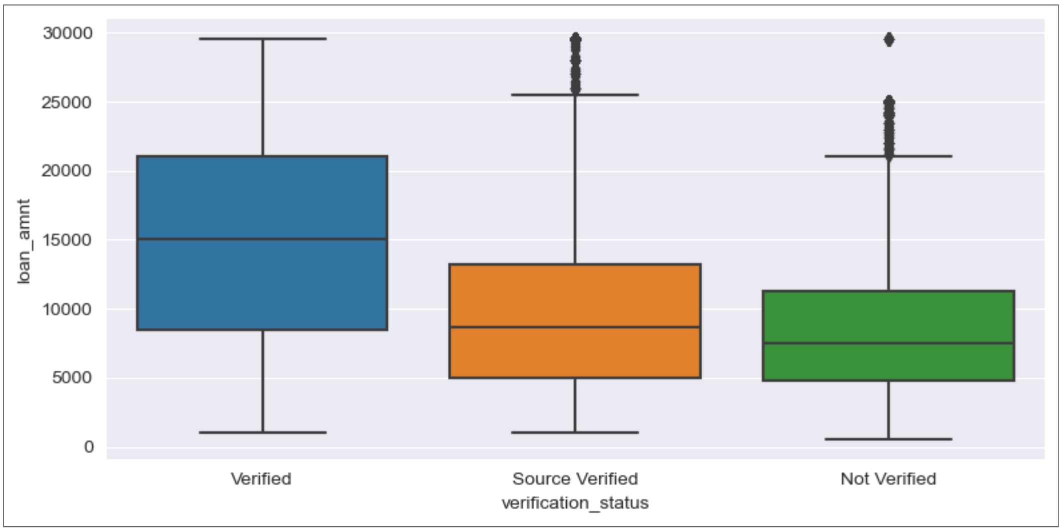

25000
20000
15000
10000
5000
0
funded_amnt_inv
RENT
OWN
MORTGAGE
OTHER
NONE
home_ownership

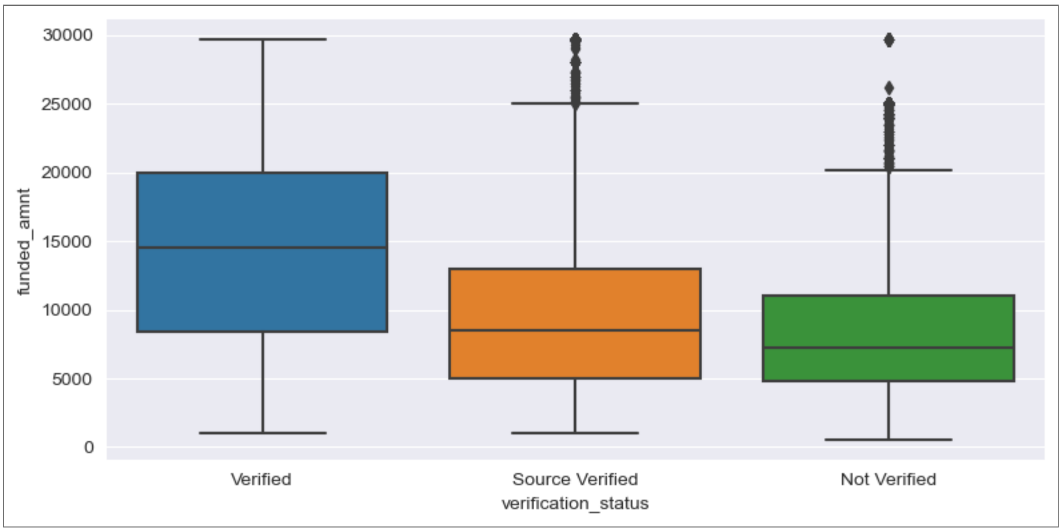

22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
int_rate
RENT
OWN
MORTGAGE
OTHER
NONE
home_ownership

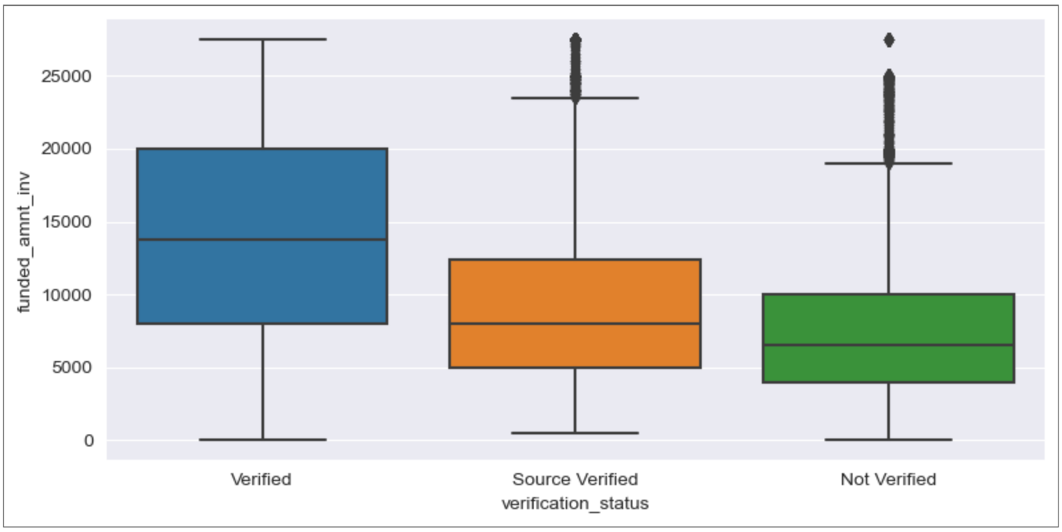

800
700
600
500
400
300
200
100
0
installment
RENT
OWN
MORTGAGE
OTHER
NONE
home_ownership

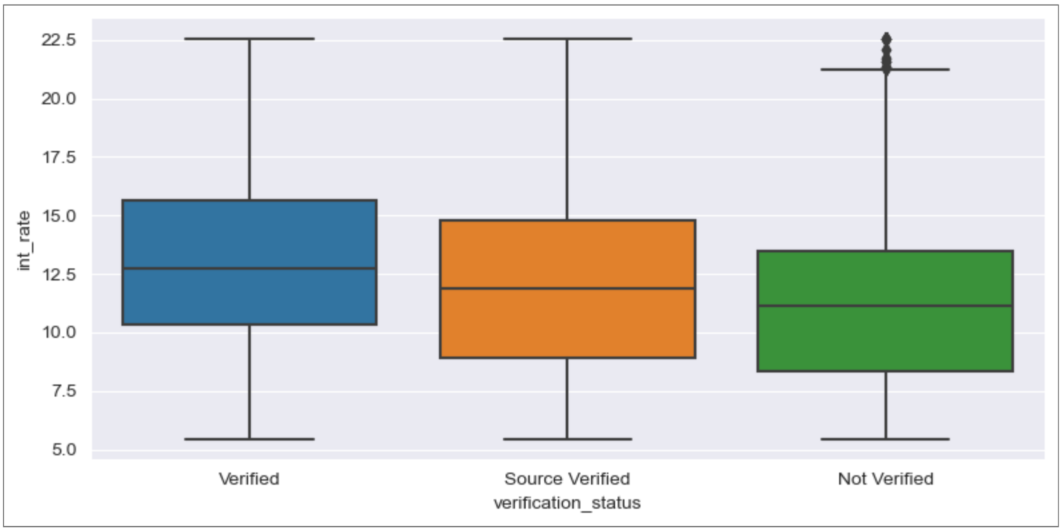

140000
120000
100000
80000
60000
40000
20000
0
annual_inc
RENT
OWN
MORTGAGE
OTHER
NONE
home_ownership

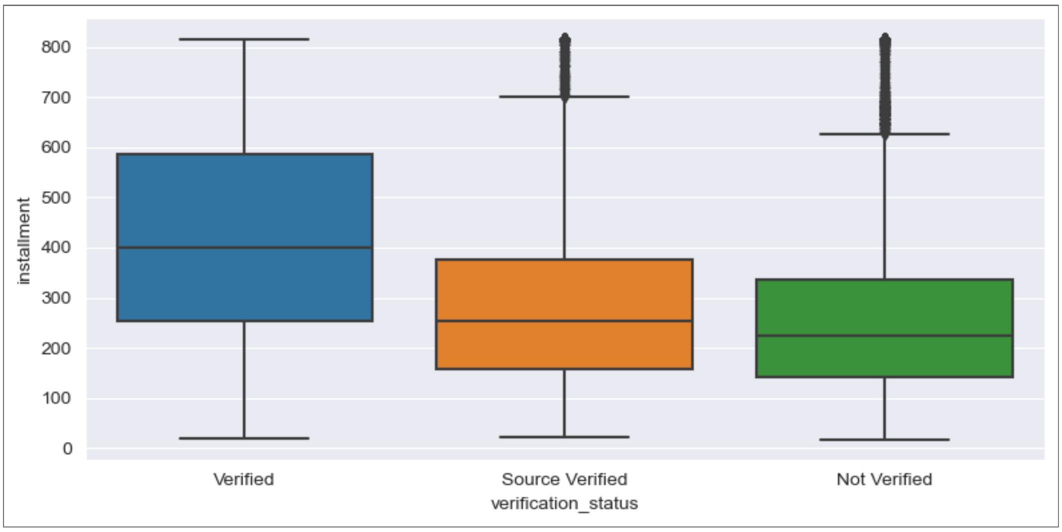

30
25
20
15
10
5
0
dti
RENT
OWN
MORTGAGE
OTHER
NONE
home_ownership

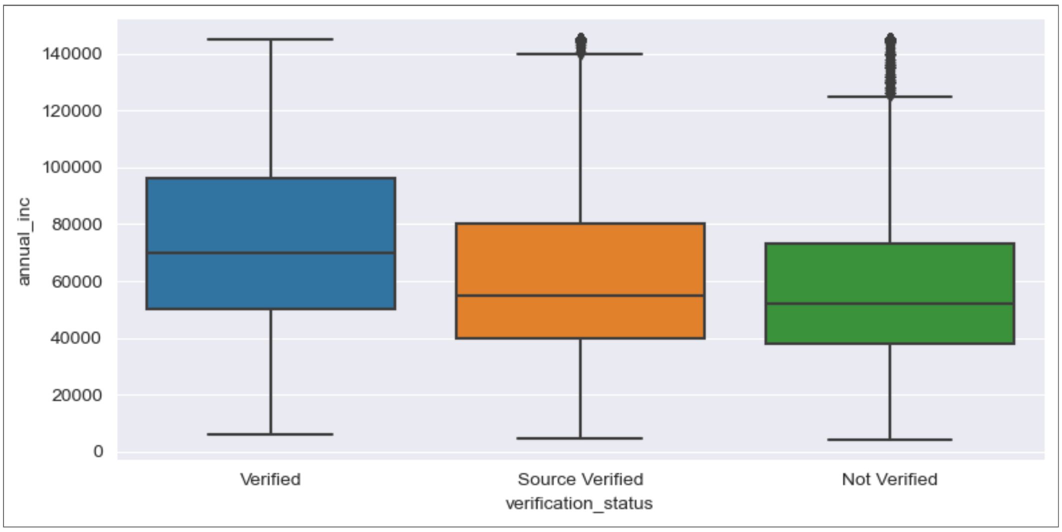

30000
25000
20000
15000
10000
5000
0
loan_amnt
Verified
Source Verified
Not Verified
verification_status

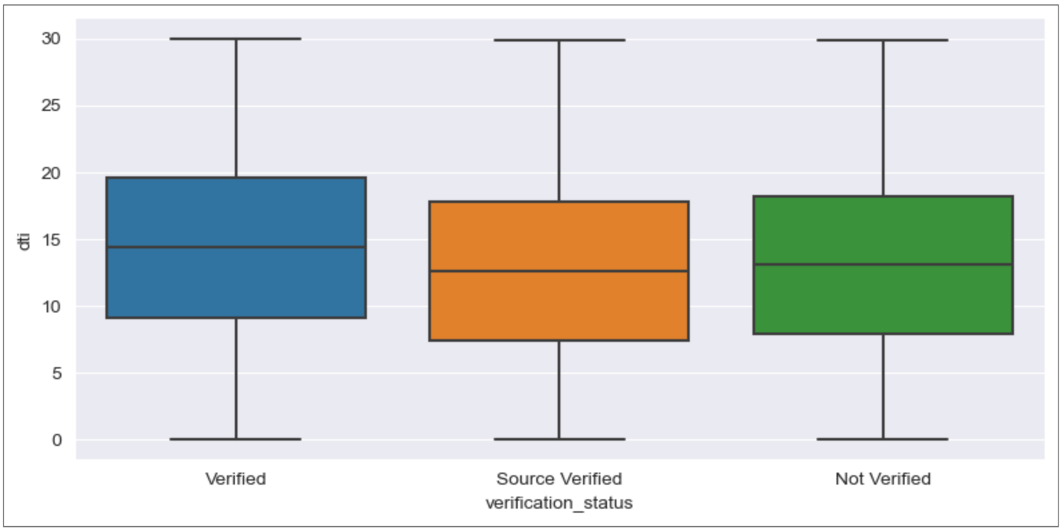

30000
25000
20000
15000
10000
5000
0
funded_amnt
Verified
Source Verified
Not Verified
verification_status

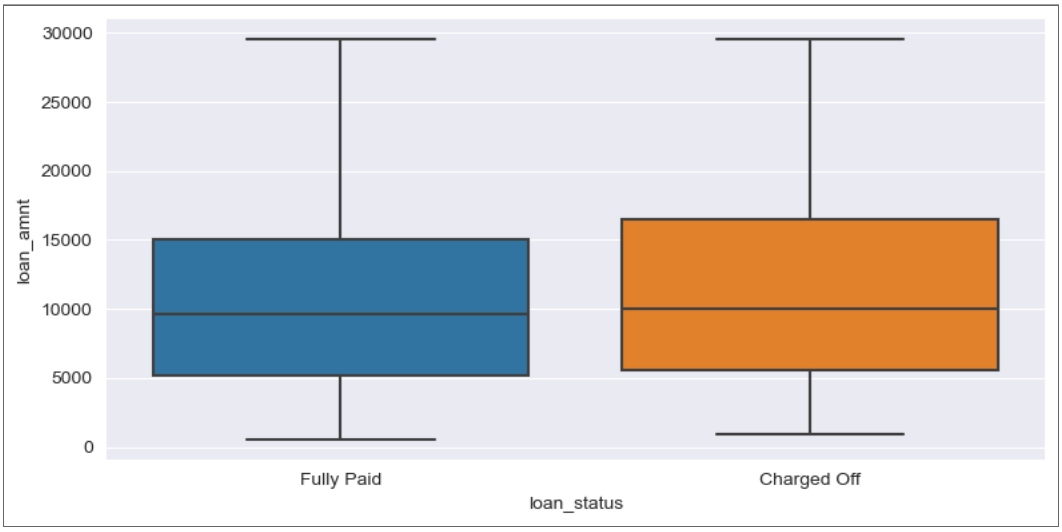

funded_amnt_inv
0
5000
10000
15000
20000
25000
Verified
Source Verified
Not Verified
verification_status

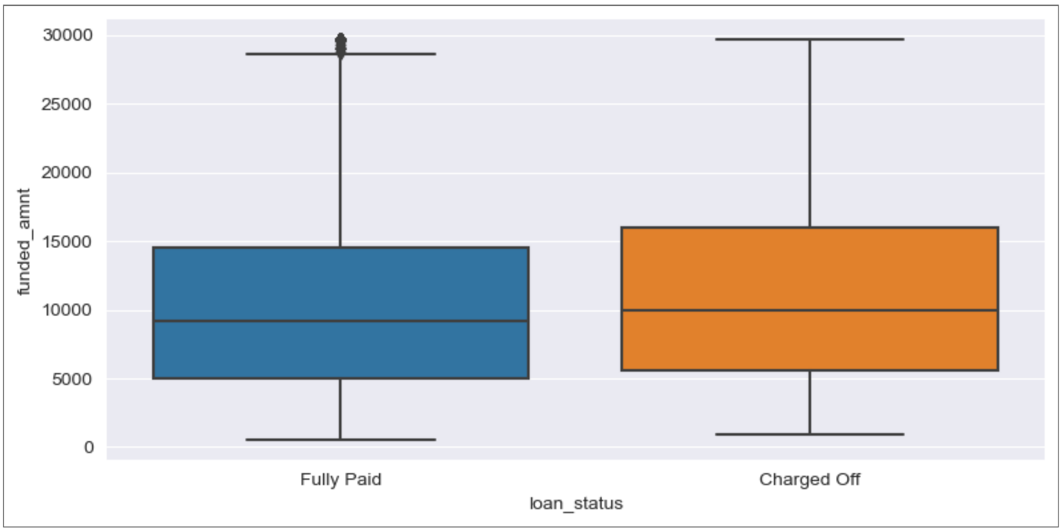

int_rate
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5
Verified
Source Verified
Not Verified
verification_status


installment
0
100
200
300
400
500
600
700
800
Verified
Source Verified
Not Verified
verification_status

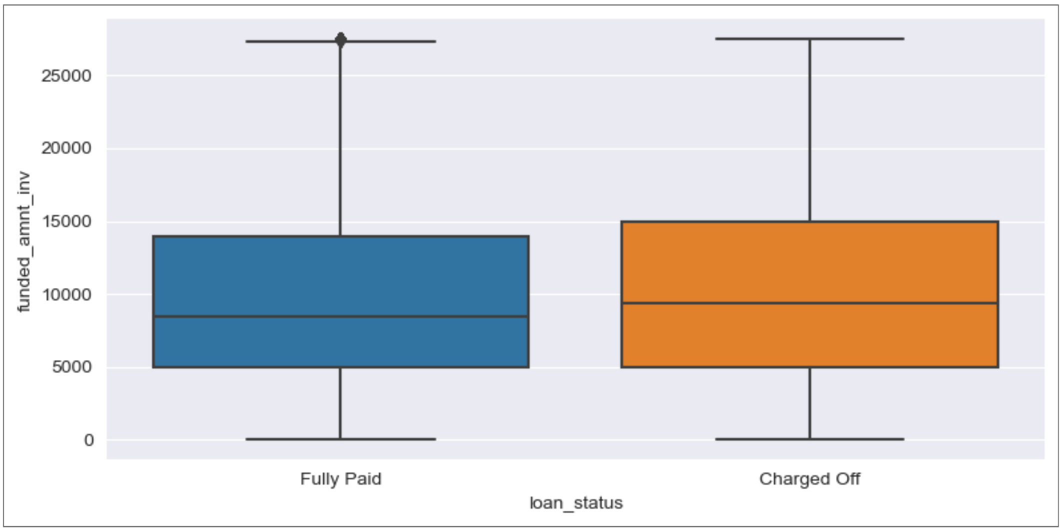

annual_inc
140000
120000
100000
80000
60000
40000
20000
0
Verified
Source Verified
Not Verified
verification_status


dti
30
25
20
15
10
5
0
Verified
Source Verified
Not Verified
verification_status

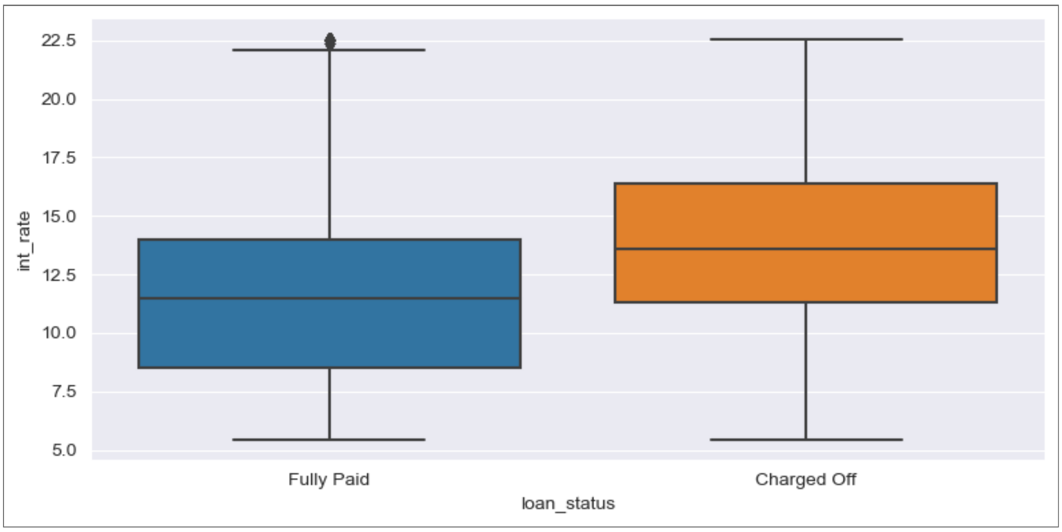

loan_amnt
30000
25000
20000
15000
10000
5000
0
Fully Paid
Charged Off
loan_status



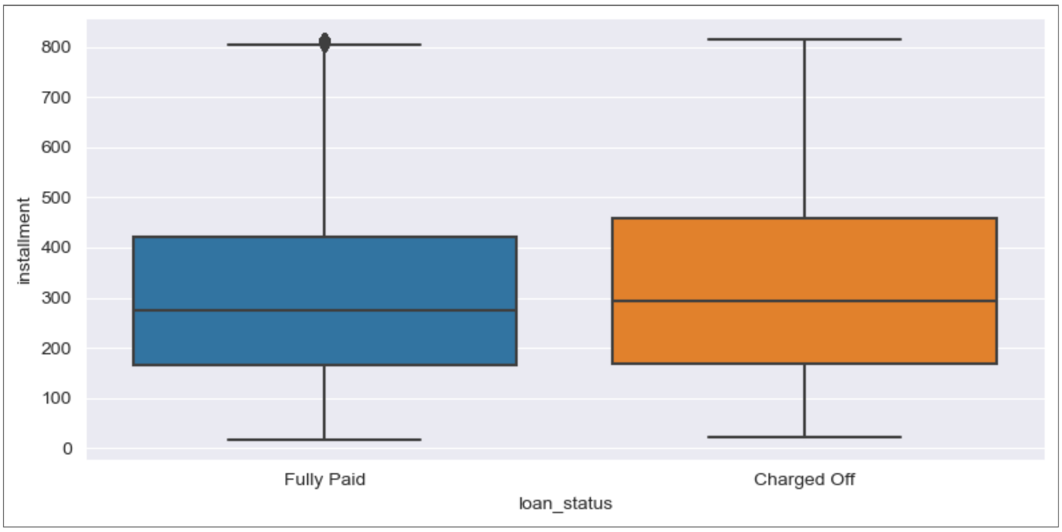
30000
25000
20000
15000
10000
5000
0
funded_amnt
Fully Paid
Charged Off
loan_status

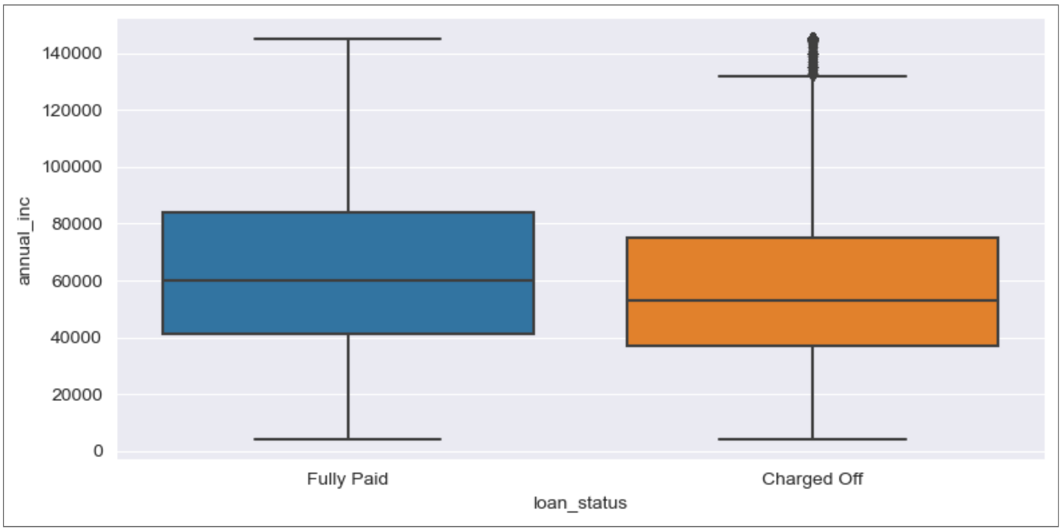


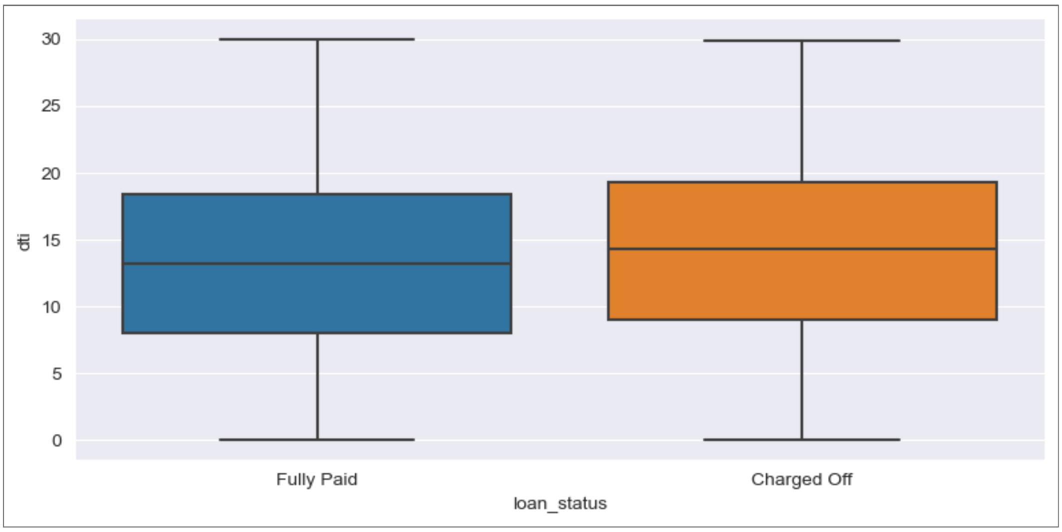
25000
20000
15000
10000
5000
0
funded_amnt_inv
Fully Paid
Charged Off
loan_status

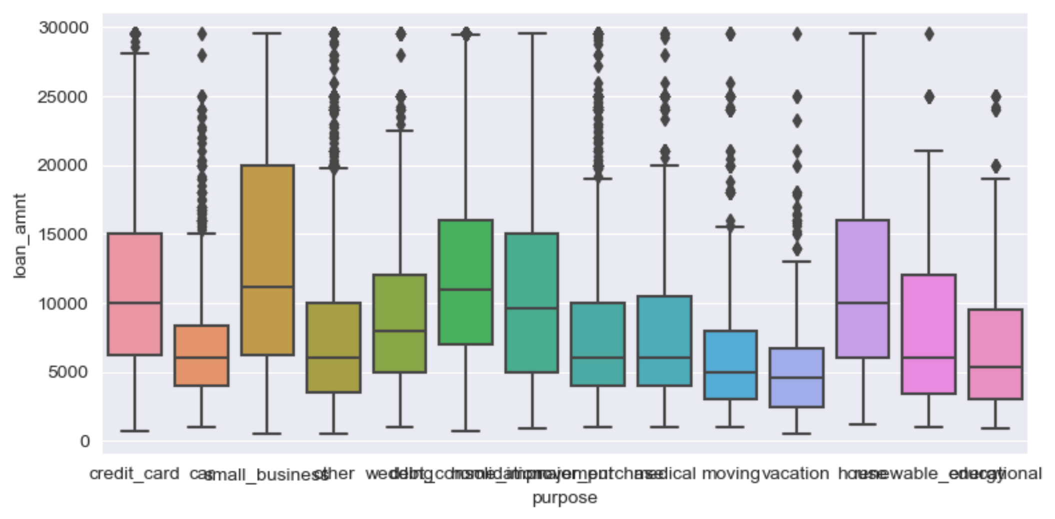


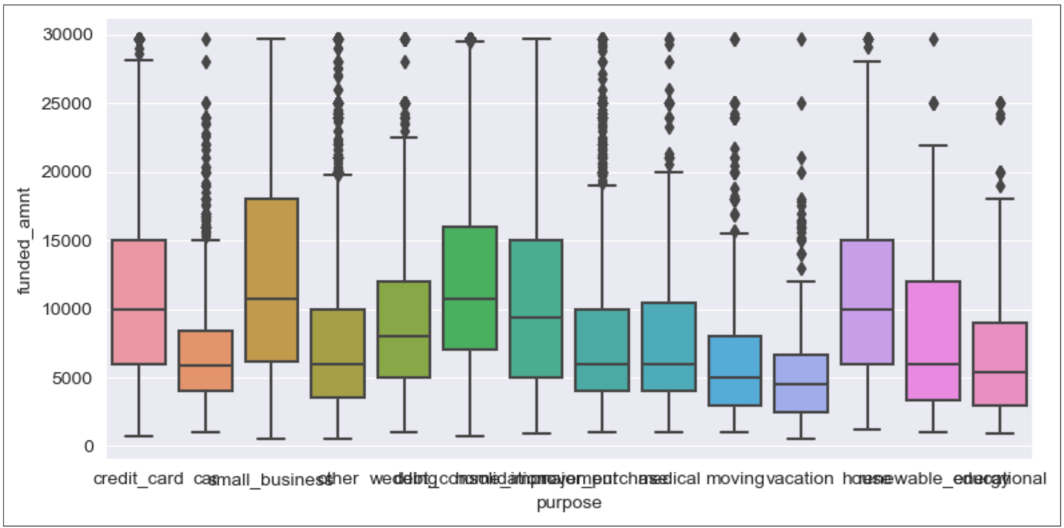
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
int_rate
Fully Paid
Charged Off
loan_status

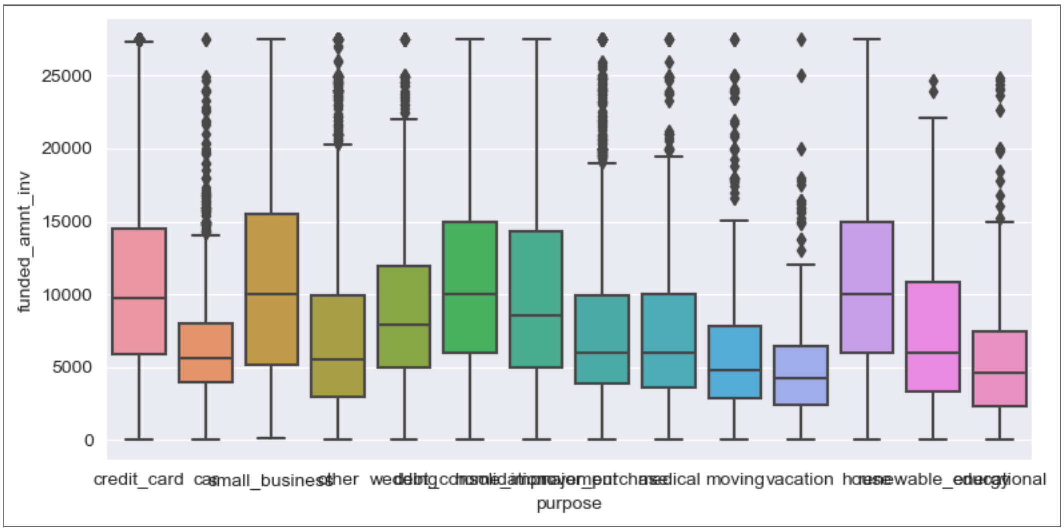

800
700
600
500
400
300
200
100
0
installment
Fully Paid
Charged Off
loan_status

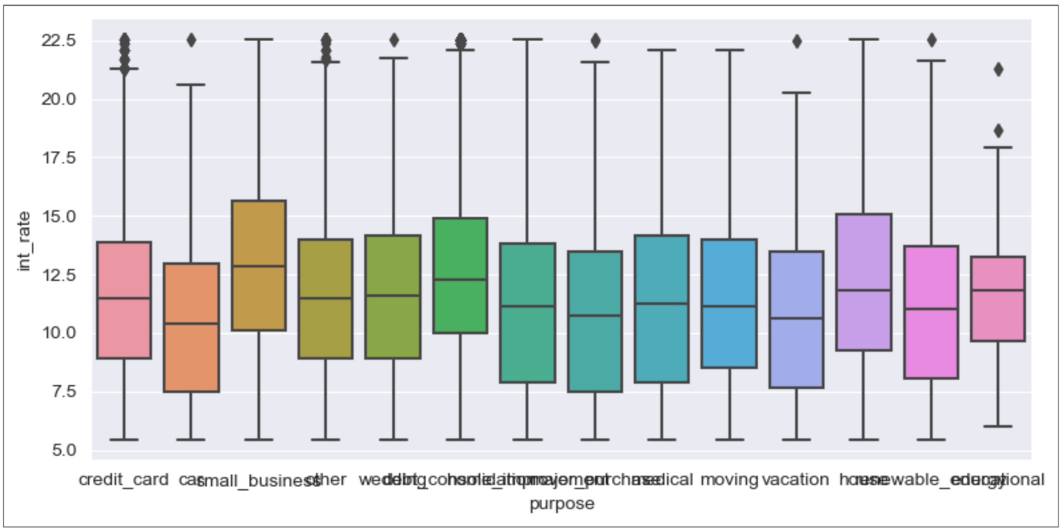

140000
120000
100000
80000
60000
40000
20000
0
annual_inc
Fully Paid
Charged Off
loan_status

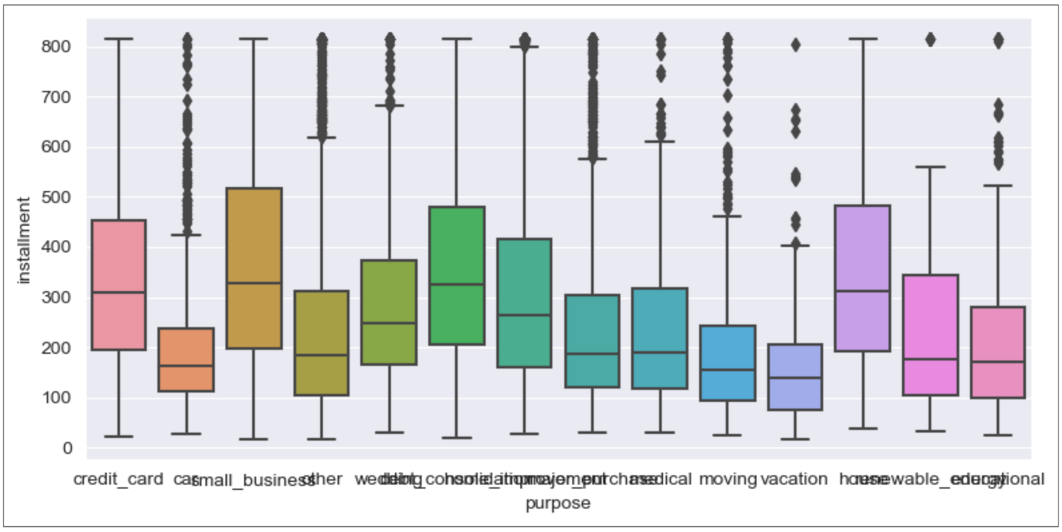

30
25
20
15
10
5
0
dti
Fully Paid
Charged Off
loan_status

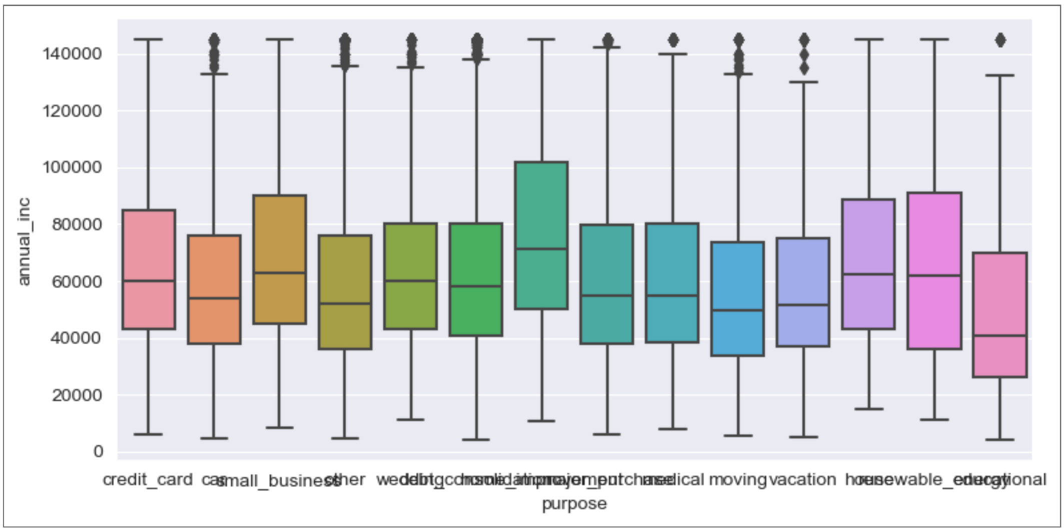

loan_amnt
purpose
credit_card
car
small_business
other
wedding
debt_consolidation
home_improvement
major_purchase
medical
moving
vacation
house
renewable_energy
educational

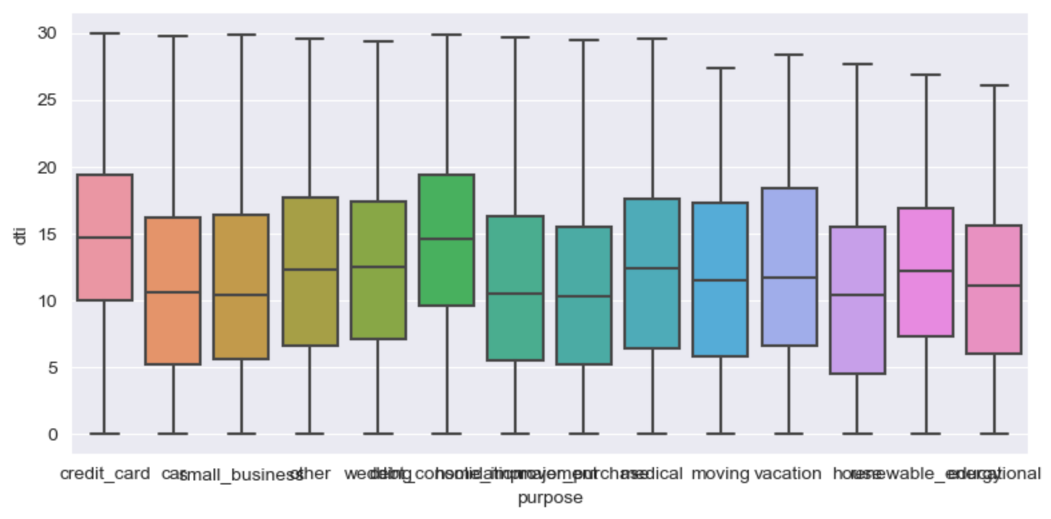

funded_amnt
purpose
credit_card
car
small_business
other
wedding
debt_consolidation
home_improvement
major_purchase
medical
moving
vacation
house
renewable_energy
educational

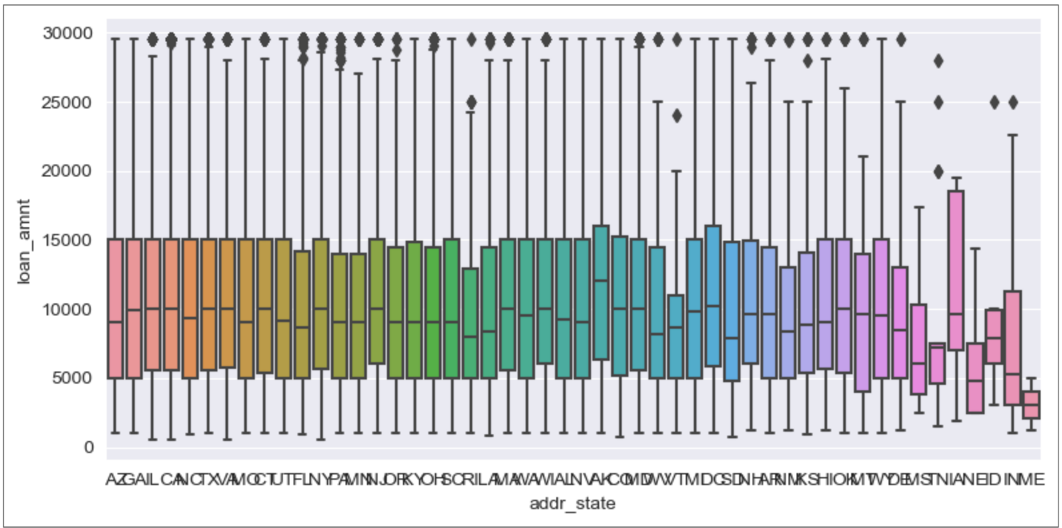

funded_amnt_inv
purpose
credit_card
car
small_business
other
wedding
debt_consolidation
home_improvement
major_purchase
medical
moving
vacation
house
renewable_energy
educational

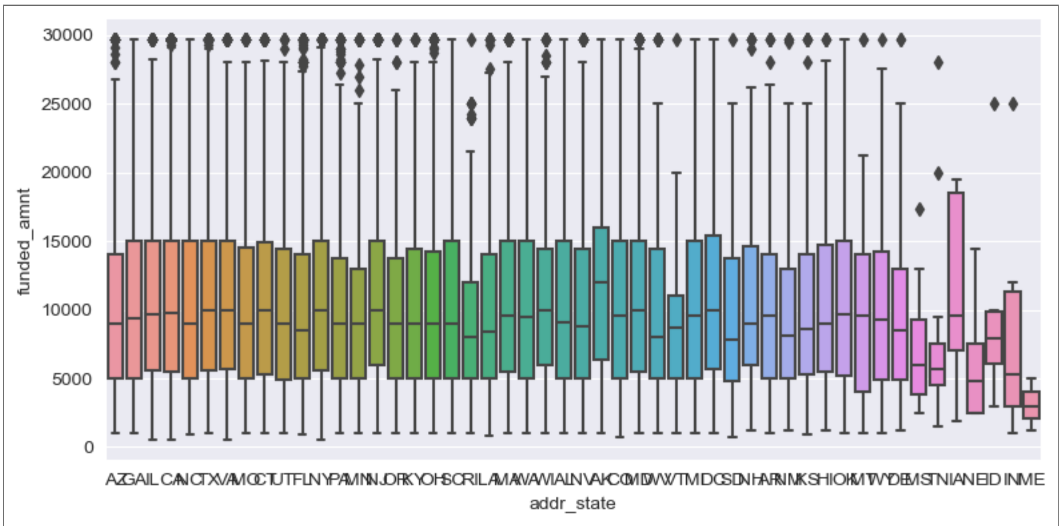

int_rate
purpose
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0

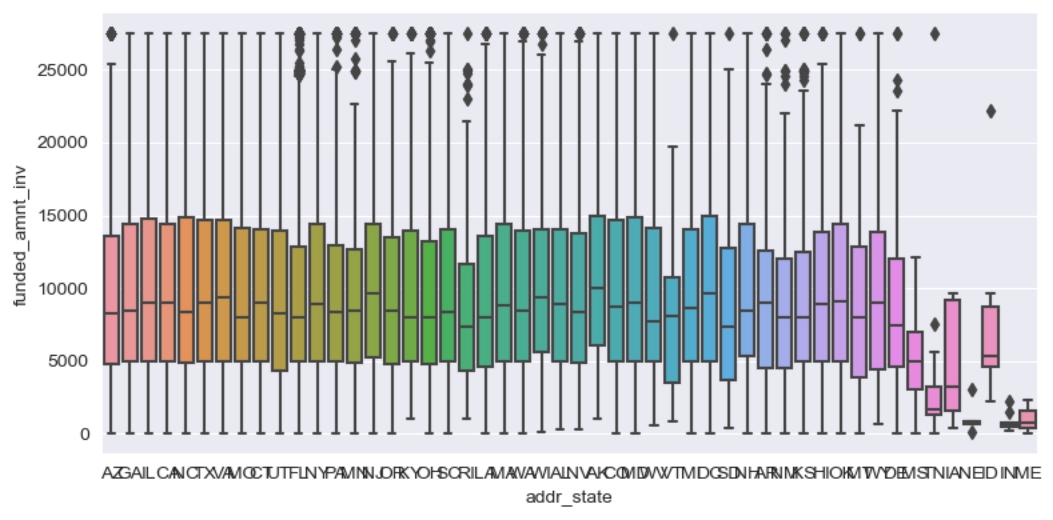

installment
purpose
800
700
600
500
400
300
200
100
0

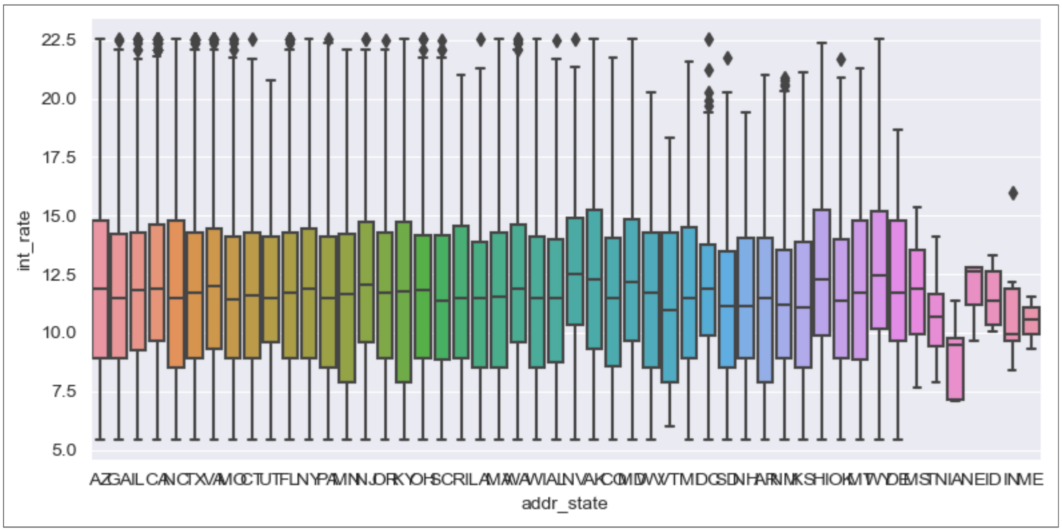

annual_inc
purpose
140000
120000
100000
80000
60000
40000
20000
0
credit_card
moving
vacation

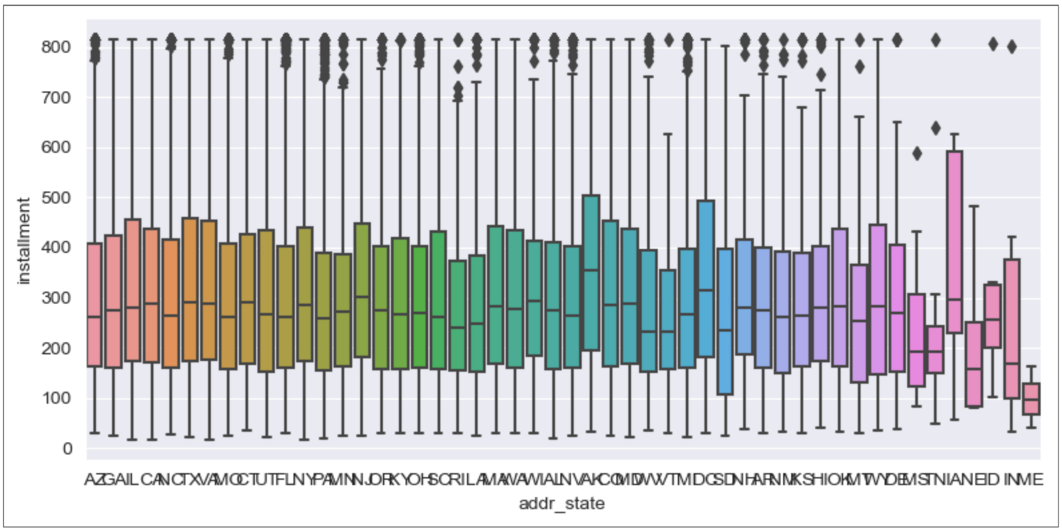

dti
0
5
10
15
20
25
30
credit_card
medical
moving
vacation
purpose

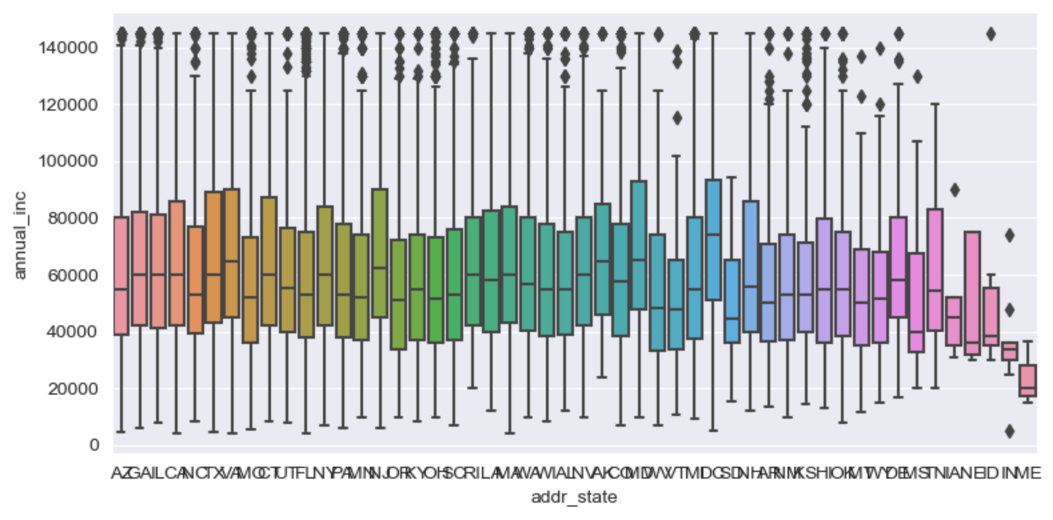

loan_amnt
0
5000
10000
15000
20000
25000
30000
addr_state


funded_amnt
0
5000
10000
15000
20000
25000
30000
addr_state

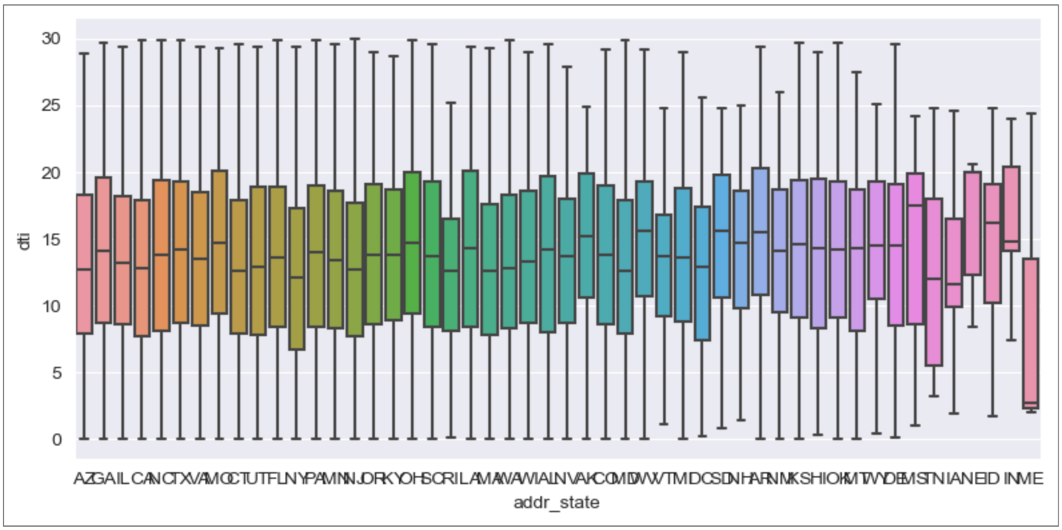

funded_amnt_inv
0
5000
10000
15000
20000
25000
addr_state


int_rate
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5
addr_state

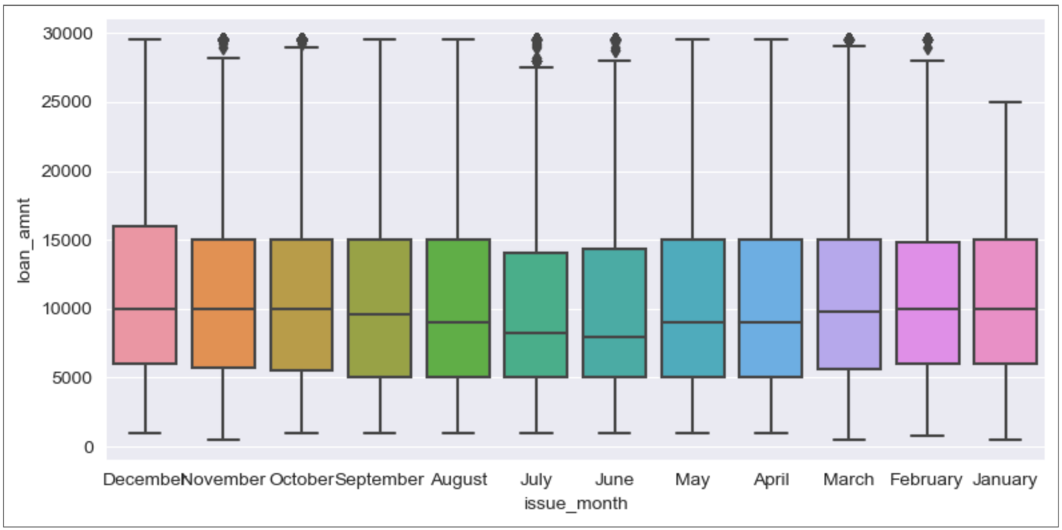

installment
0
100
200
300
400
500
600
700
800
addr_state



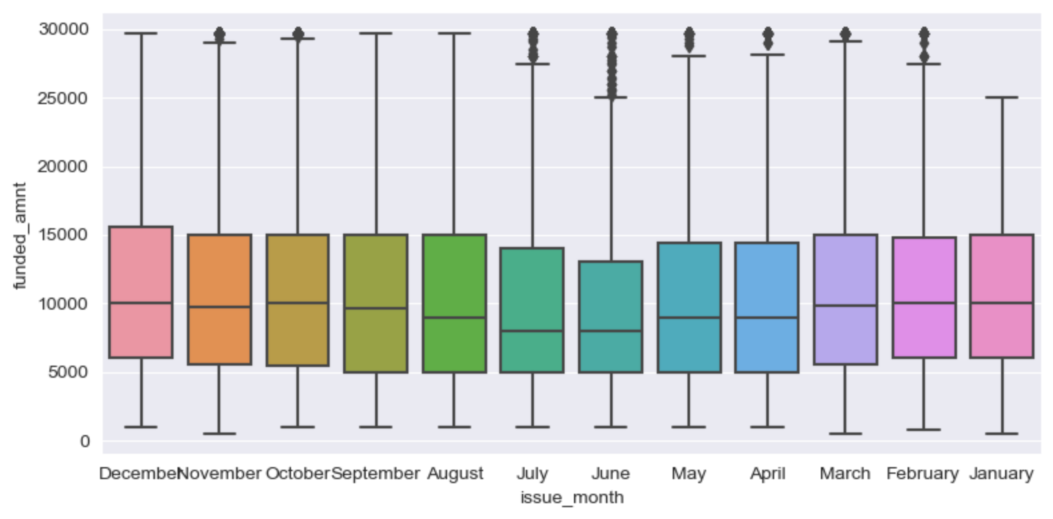
annual_inc
addr_state

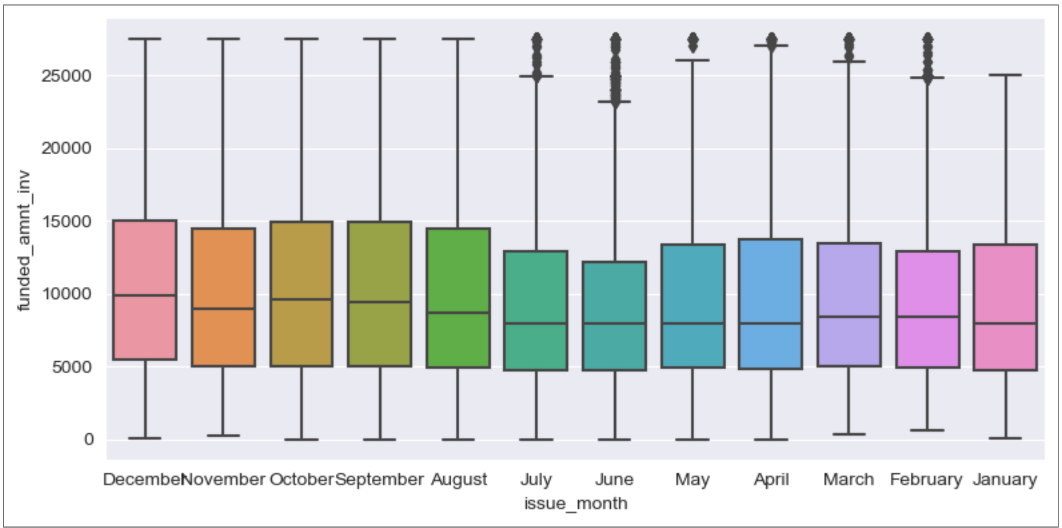


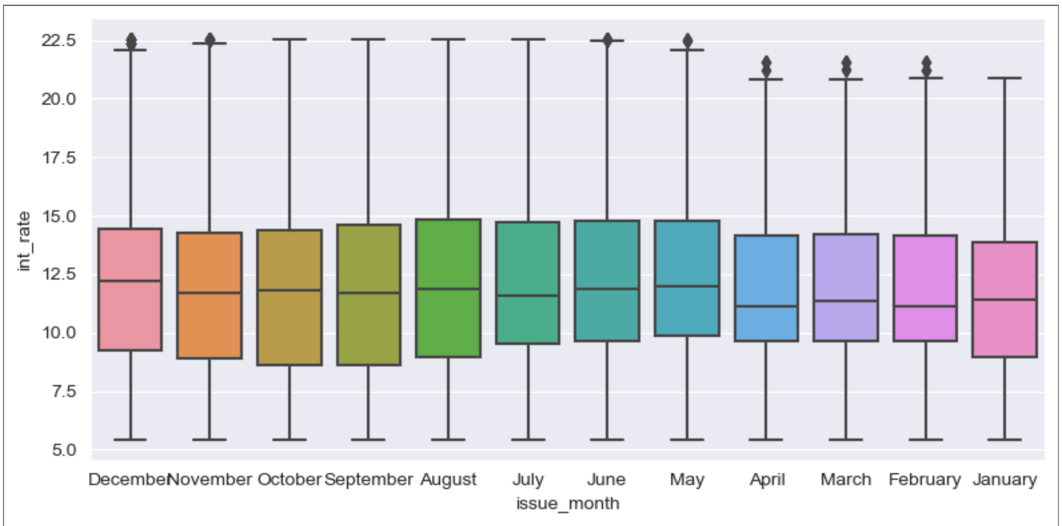
dti
addr_state

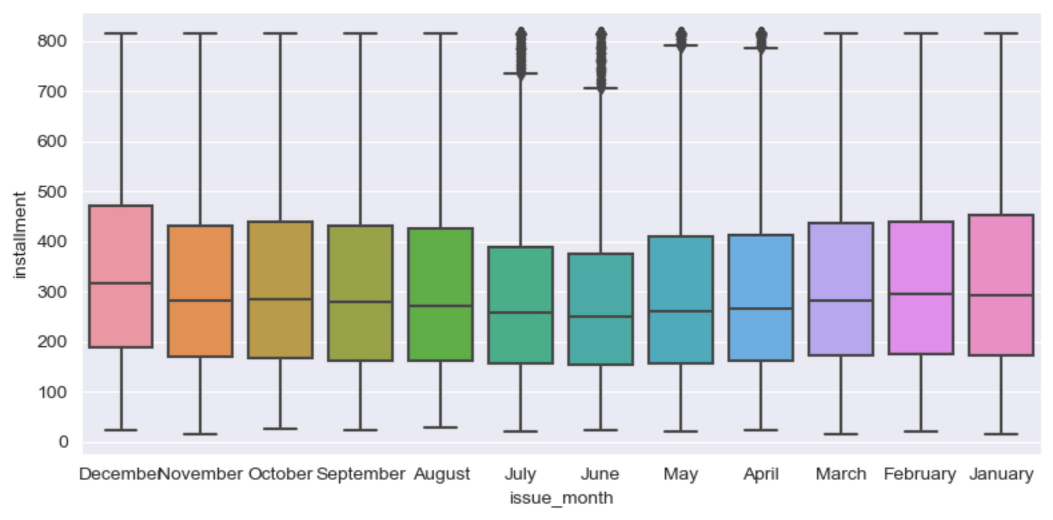


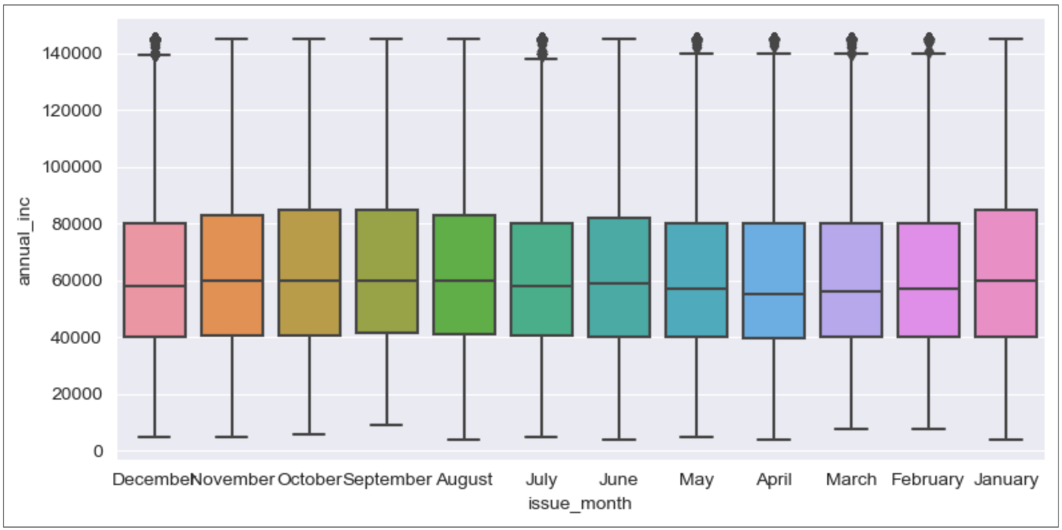
loan_amnt
DecembeNovember October September August July June May April March February January
issue_month

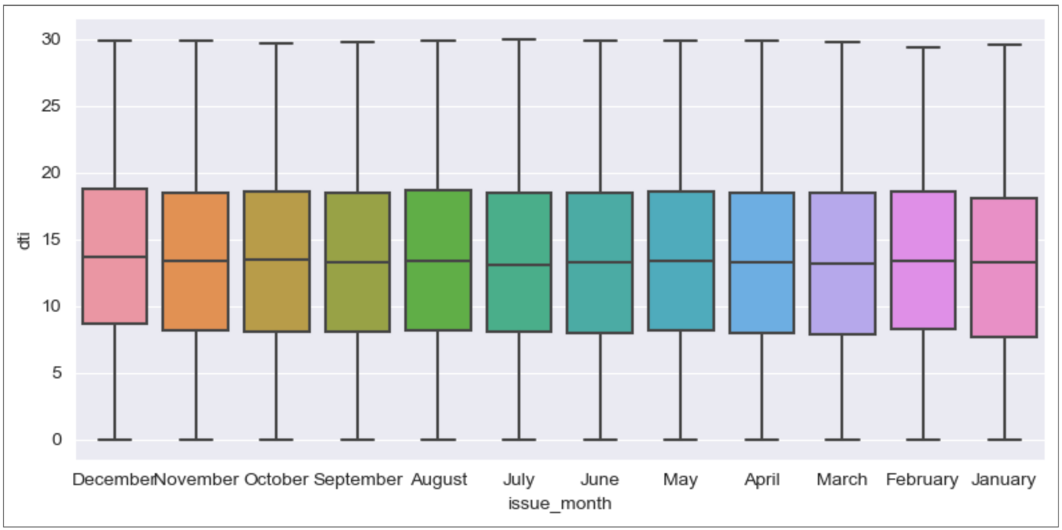

funded_amnt
0
5000
10000
15000
20000
25000
30000
December
November
October
September
August
July
June
May
April
March
February
January
issue_month

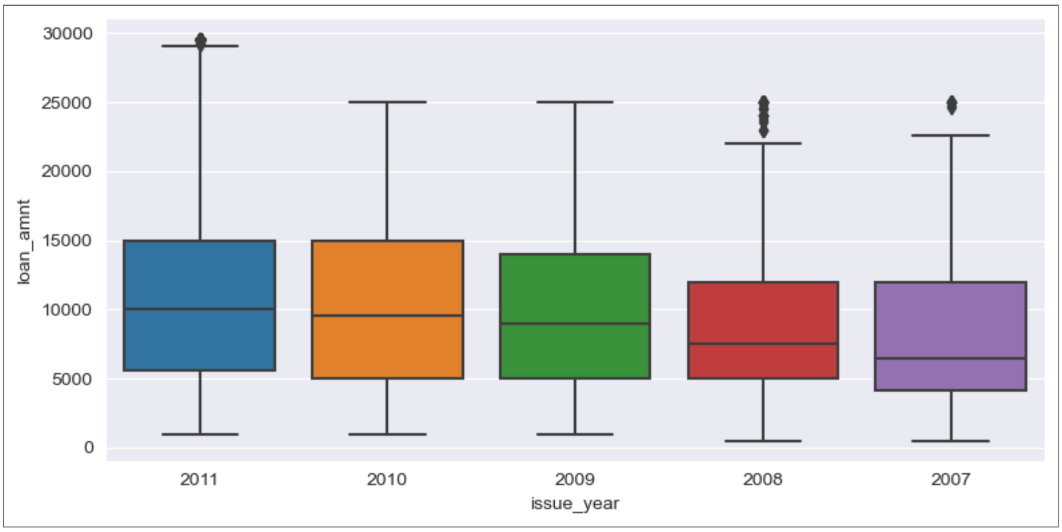

funded_amnt_inv
0
5000
10000
15000
20000
25000
December
November
October
September
August
July
June
May
April
March
February
January
issue_month

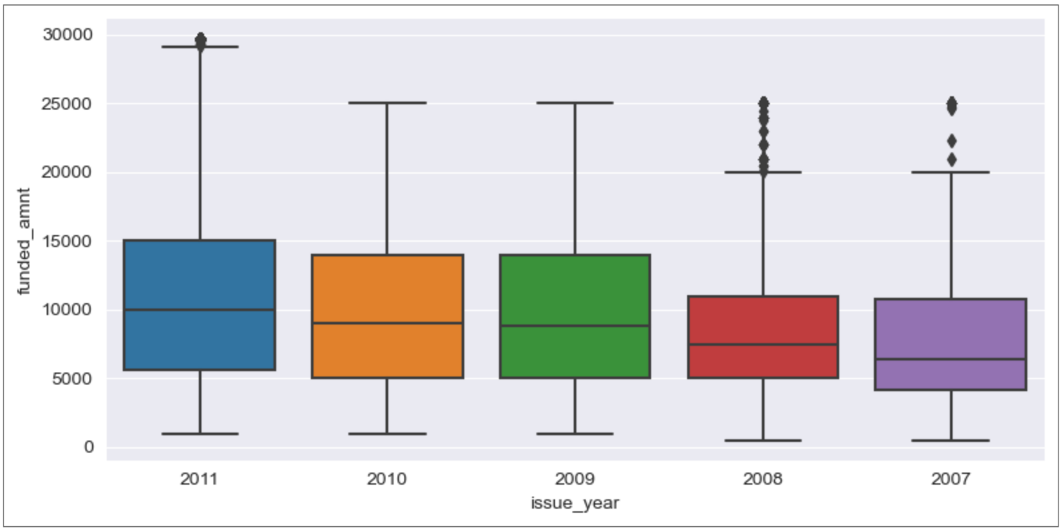

int_rate
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5
December
November
October
September
August
July
June
May
April
March
February
January
issue_month

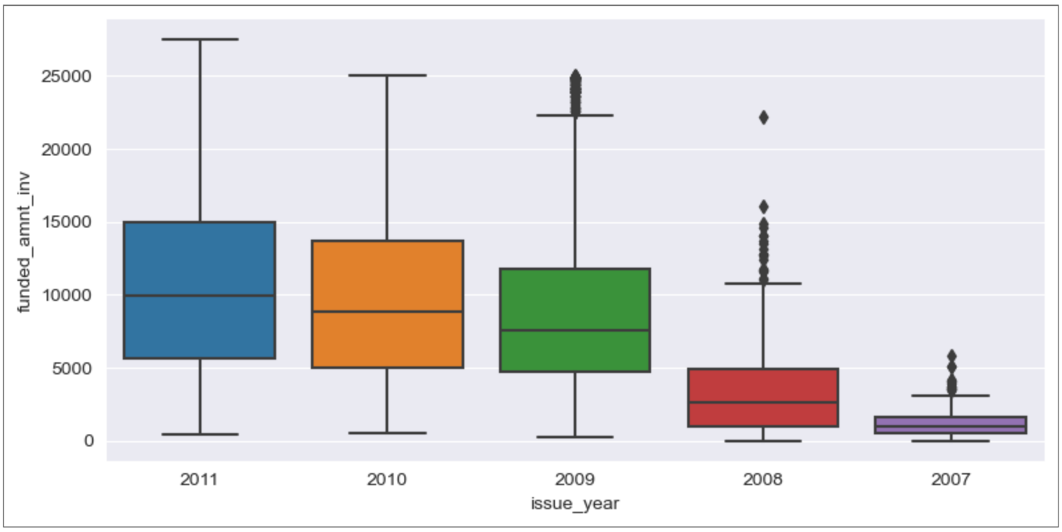

installment
800
700
600
500
400
300
200
100
0
DecembeNovember October September August July June May April March February January
issue_month

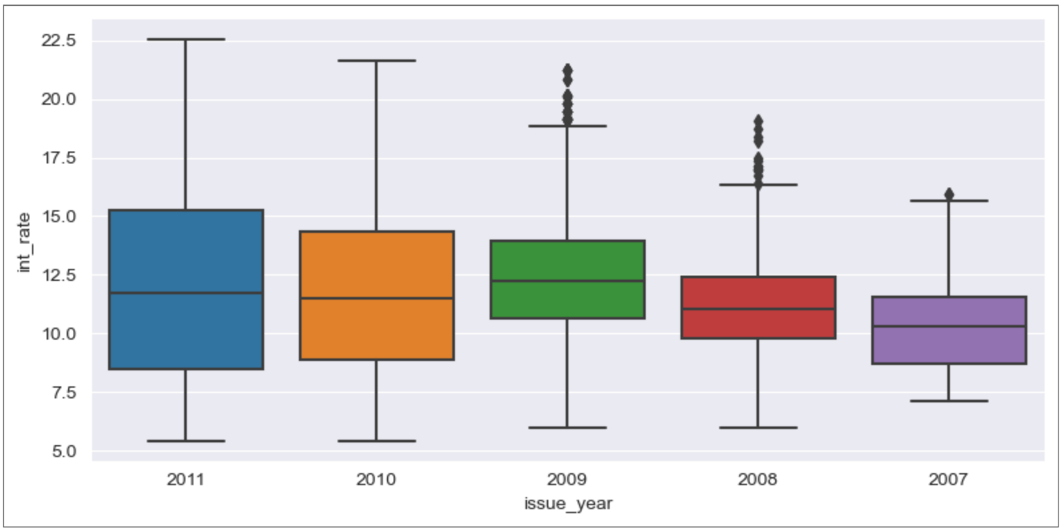

annual_inc
140000
120000
100000
80000
60000
40000
20000
0
DecembeNovember OctoberSeptember August July June May April March February January
issue_month

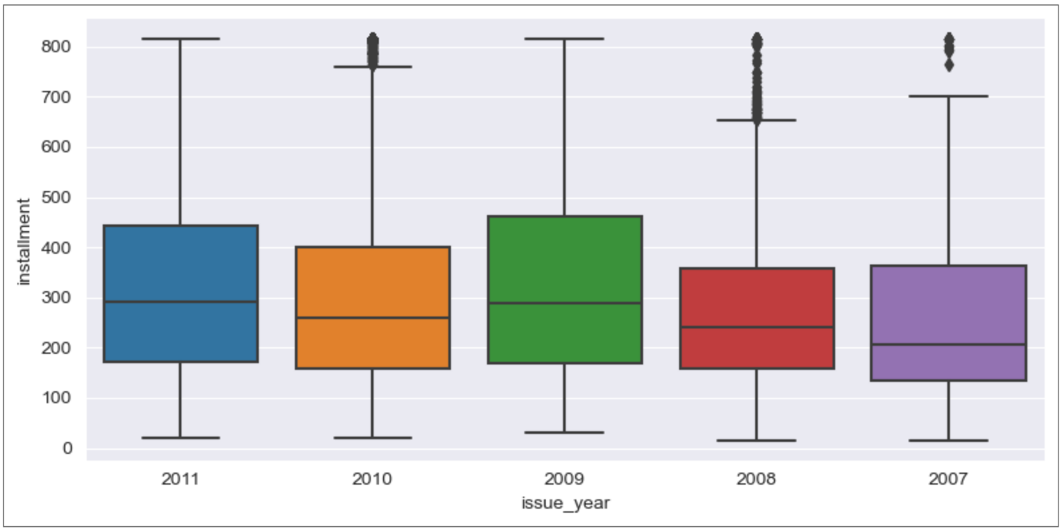

dti
30
25
20
15
10
5
0
DecemberNovember October September August July June May April March February January
issue_month

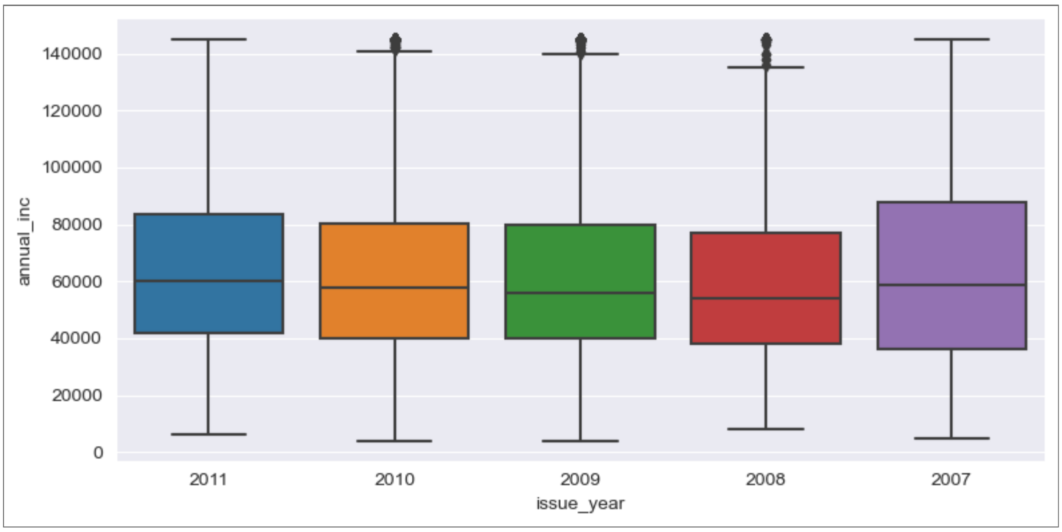

30000
25000
20000
15000
10000
5000
0
loan_amnt
2011
2010
2009
2008
2007
issue_year

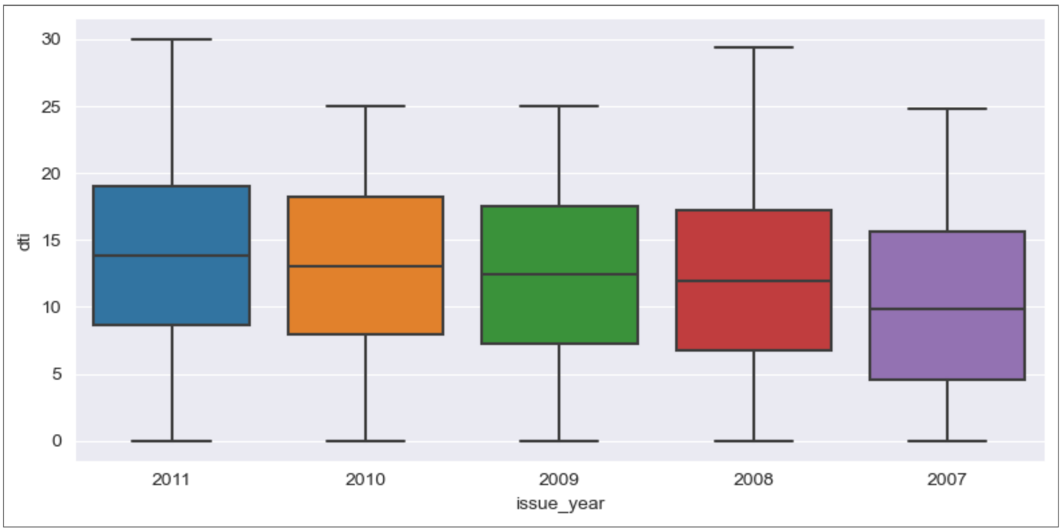

30000
25000
20000
15000
10000
5000
0
funded_amnt
2011
2010
2009
2008
2007
issue_year


25000
20000
15000
10000
5000
0
funded_amnt_inv
2011
2010
2009
2008
2007
issue_year


22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
int_rate
2011
2010
2009
2008
2007
issue_year


800
700
600
500
400
300
200
100
0
installment
2011
2010
2009
2008
2007
issue_year

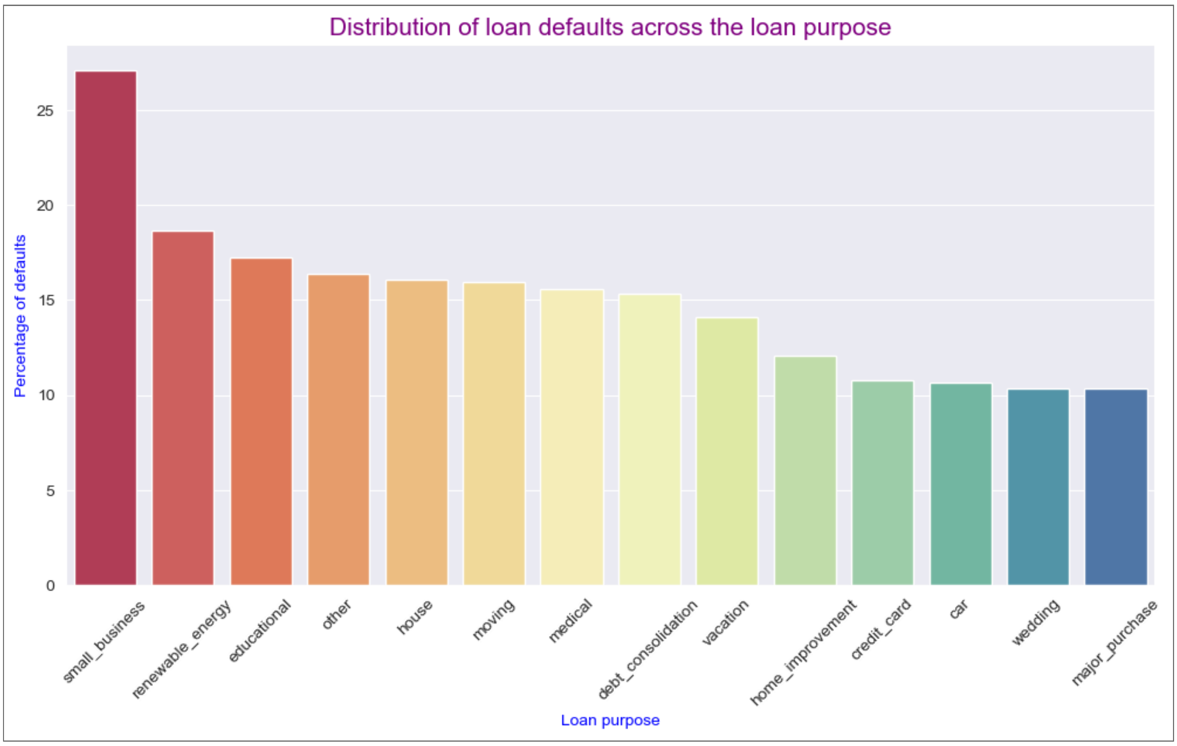

140000
120000
100000
80000
60000
40000
20000
0
annual_inc
2011
2010
2009
2008
2007
issue_year


30
25
20
15
10
5
0
dti
2011
2010
2009
2008
2007
issue_year

From the box plots, the following inferences are drawn:
The median dti in case of charged-off loan_status is higher than in case of fully paid loan_status
The median annual_inc is greater in case of charged-off loan_status than in case of fully paid loan_status
The median loan_amnt, installment, funded_amnt, and funded_amnt_inv are comparable for both cases of fully paid and charged-off loan_status
The median int_rate is higher in case of charged-off loan_status than in case of fully paid loan_status
The median dti, annual_inc,installment, int_rate, funded_amnt_inv, funded_amnt, and loan_amnt is highest in case of LC-verified profiles compared with source-verified or unverified status.
The median annual_inc varies the most across home_ownership type.
The median annual_inc and funded_amnt_inv vary considerably across the emp_length.
The int_rate is found to vary drastically across the grade.
The loan_amnt and funded_amnt vary slightly across the grade.
The median loan_amnt, funded_amnt, funded_amnt_inv is substantially less for the term of 36 months than 60 months. Nontheless, in case of annual_inc, dti, and installment, this difference is not significant.
The median of loan_amnt, funded_amnt, funded_amnt_inv, and installment vary significantly across different loan purposes.
The median of annual_inc varies significantly across addr_state.
While the median loan_amnt does not vary subtantially across the issue_year, the funded_amnt_inv and int_int show considerable variation.

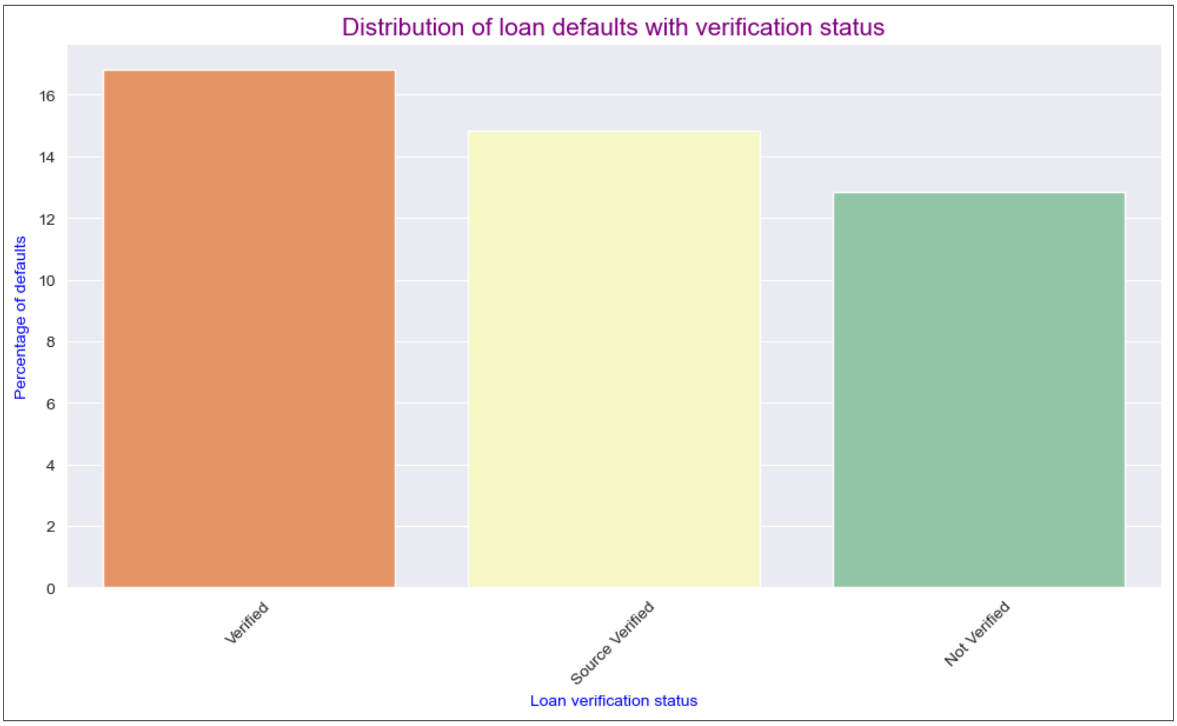
Our target variable for the analysis is loan_status which has two possible values, "Charged Off" and "Fully Paid". Next, we will concentrate our analysis to find out patterns where borrowers default on their loan.

Influence of the purpose of loan application on loan default


Distribution of loan defaults across the loan purpose
Percentage of defaults
0
5
10
15
20
25
small_business
renewable_energy
educational
other
house
moving
medical
debt_consolidation
vacation
home_improvement
credit_card
car
wedding
major_purchase
Loan purpose

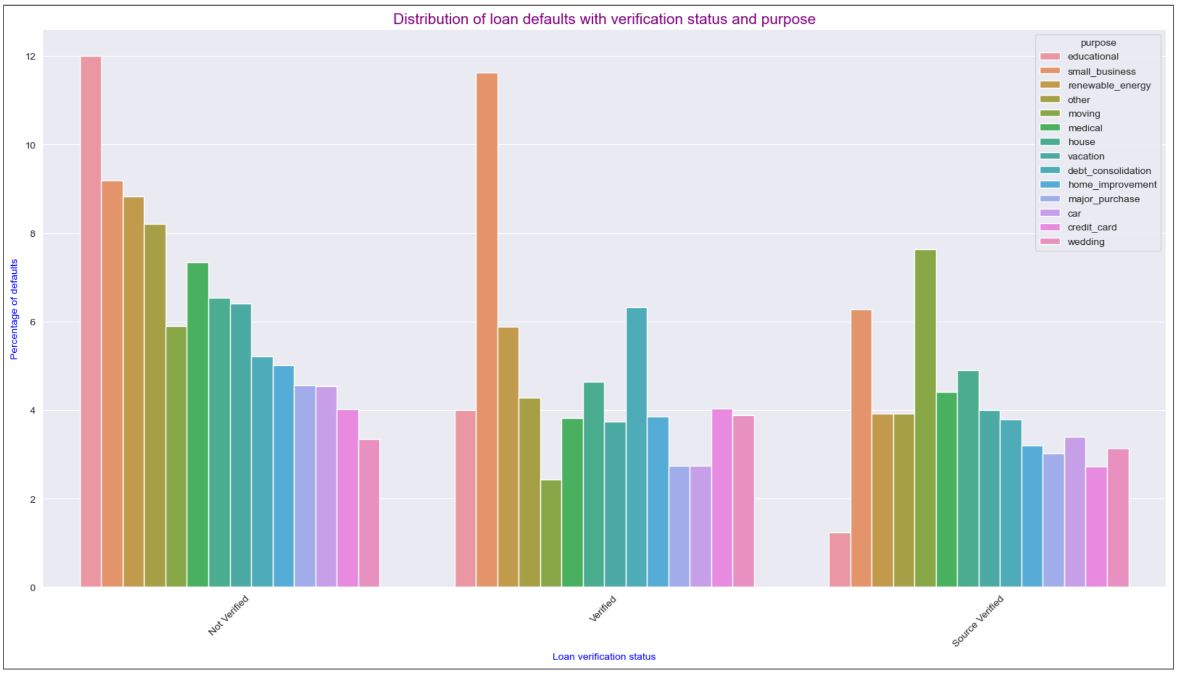
Above bar plot shows that loans taken for the purpose of Small business are extremely risky; ~27% of such loans have defaulted.

Influence of loan verification on loan default


Distribution of loan defaults with verification status
Percentage of defaults
0
2
4
6
8
10
12
14
16
Verified
Source Verified
Not Verified
Loan verification status

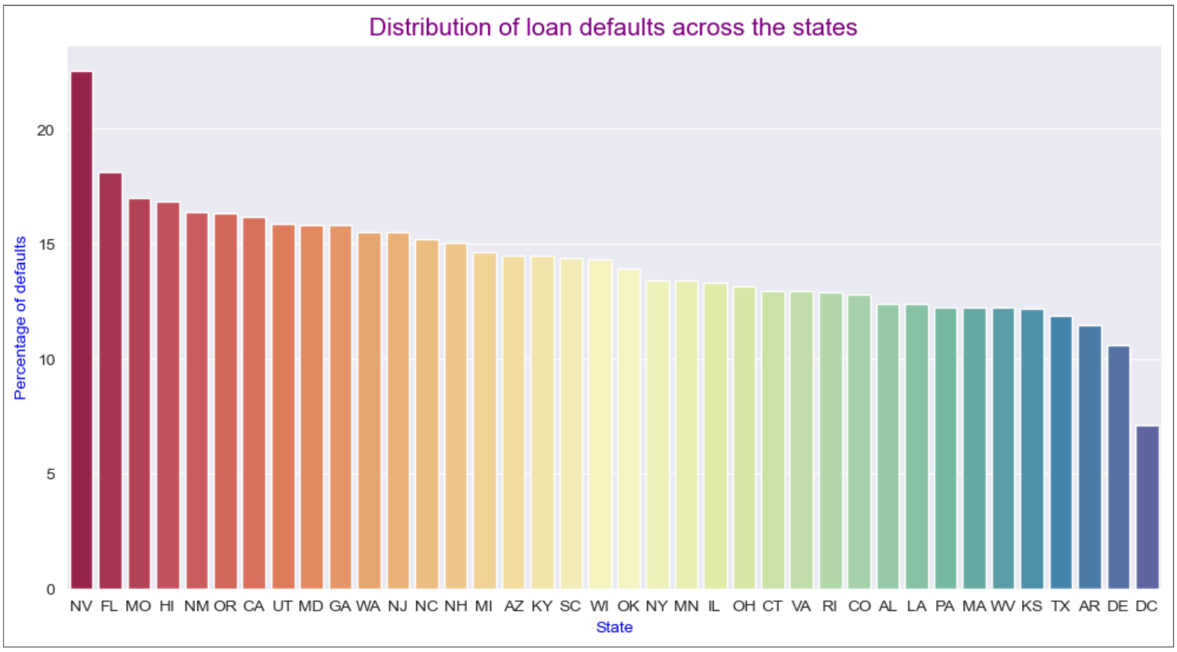
This is an interesting observation that verified applicants are defaulting more which suggests that LC's verification process is not effective and maybe there are a few loopholes within the process that need to be addressed.

Now, lets find out if this trend is same across all types of loan purpose.


Distribution of loan defaults with verification status and purpose
purpose
educational
small_business
renewable_energy
other
moving
medical
house
vacation
debt_consolidation
home_improvement
major_purchase
car
credit_card
wedding
Percentage of defaults
0
2
4
6
8
10
12
Not Verified
Verified
Source Verified
Loan verification status

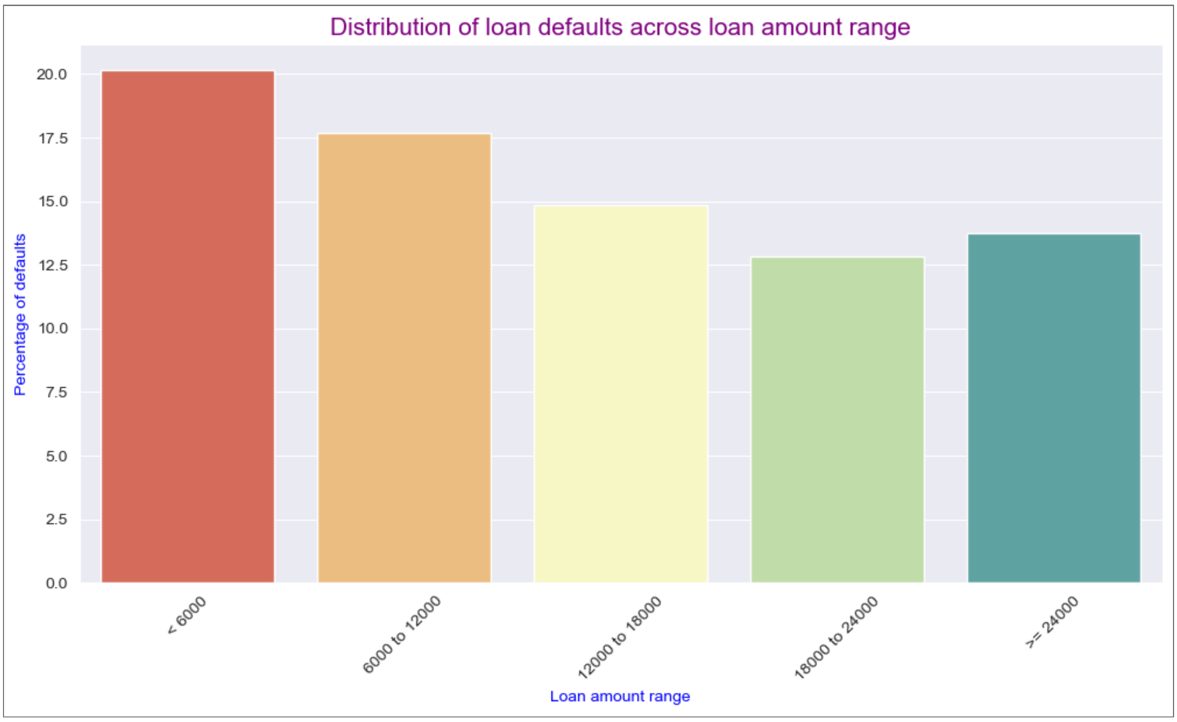
The following are the main observations from this plot:
Loans for educational purpose: These are highly likely to default if loan application profile is not verified.
Loans for Small business: These are most likely to default when income source is verified, compared with when they are not verified or are source-verified.
Loan for moving: These are most likely to default when source-verified.

Analyzing the default percentage across different states


Distribution of loan defaults across the states
Percentage of defaults
0
5
10
15
20
NV FL MO HI NM OR CA UT MD GA WA NJ NC NH MI AZ KY SC WI OK NY MN IL OH CT VA RI CO AL LA PA MA WV KS TX AR DE DC
State

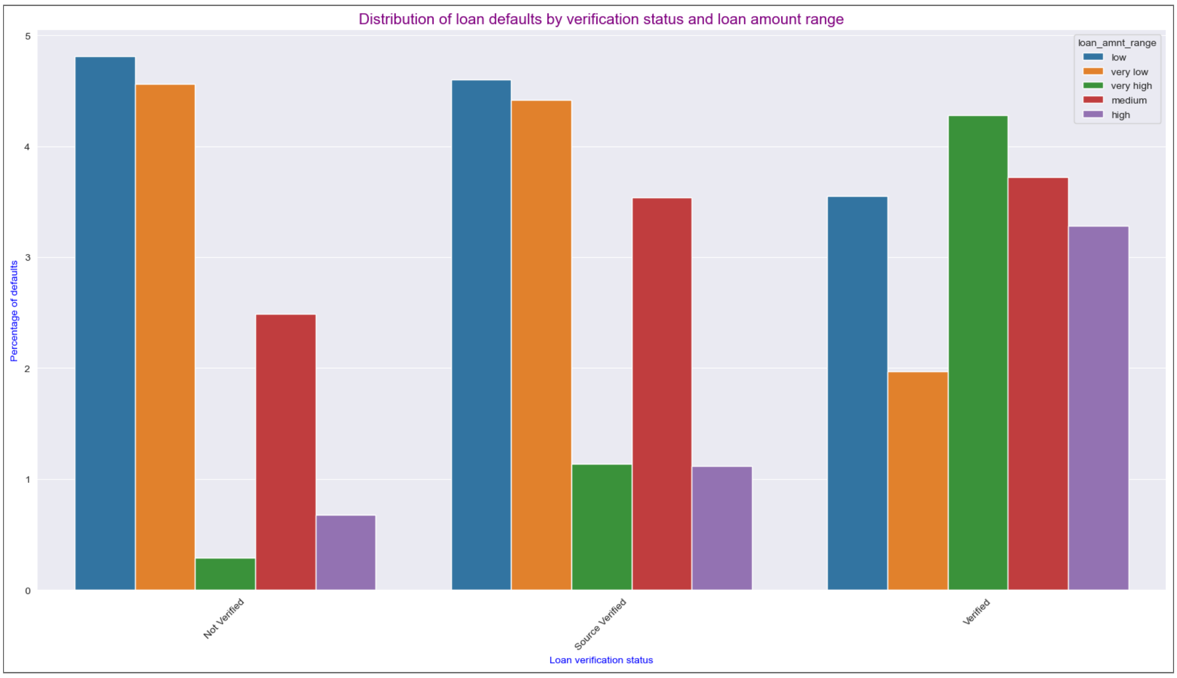
We observe that NV has the highest default rate of ~27%

Influence of loan amount on loan default


Distribution of loan defaults across loan amount range
Percentage of defaults
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
< 6000
6000 to 12000
12000 to 18000
18000 to 24000
>= 24000
Loan amount range

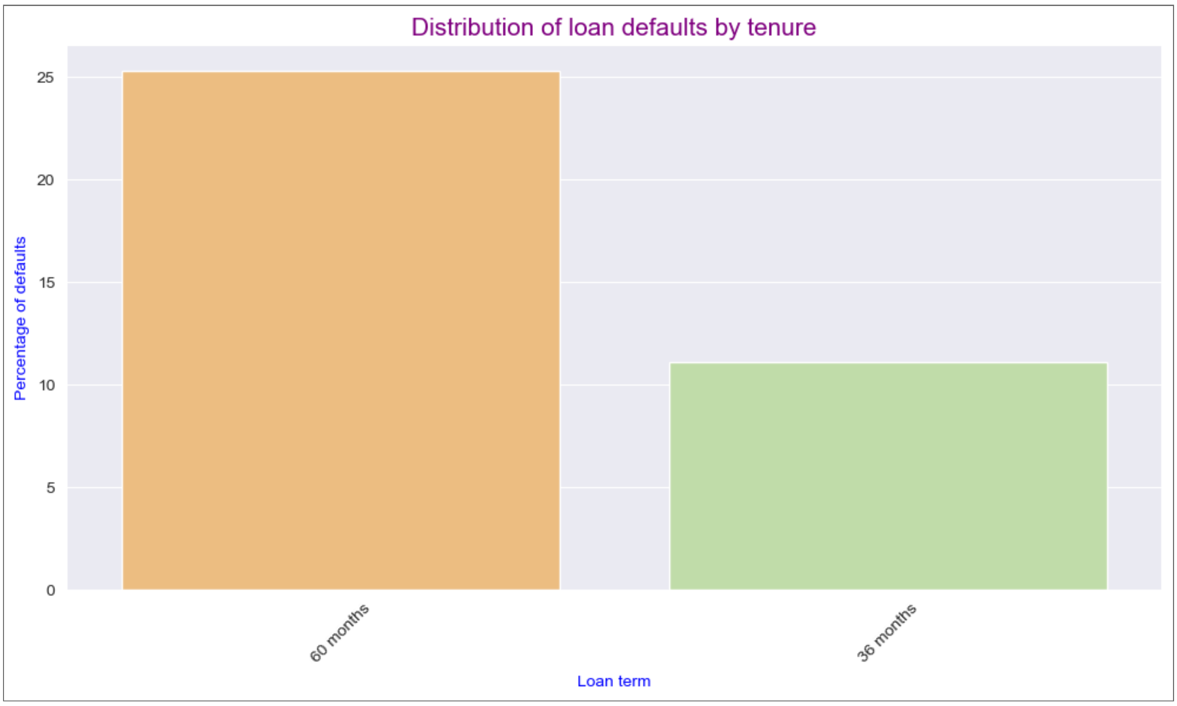
We observe that the higher-amount loans are more likely to default.

Now let's visulalize this trend against verification_status


Distribution of loan defaults by verification status and loan amount range
loan_amnt_range
low
very low
very high
medium
high
Percentage of defaults
0
1
2
3
4
5
Not Verified
Source Verified
Verified
Loan verification status

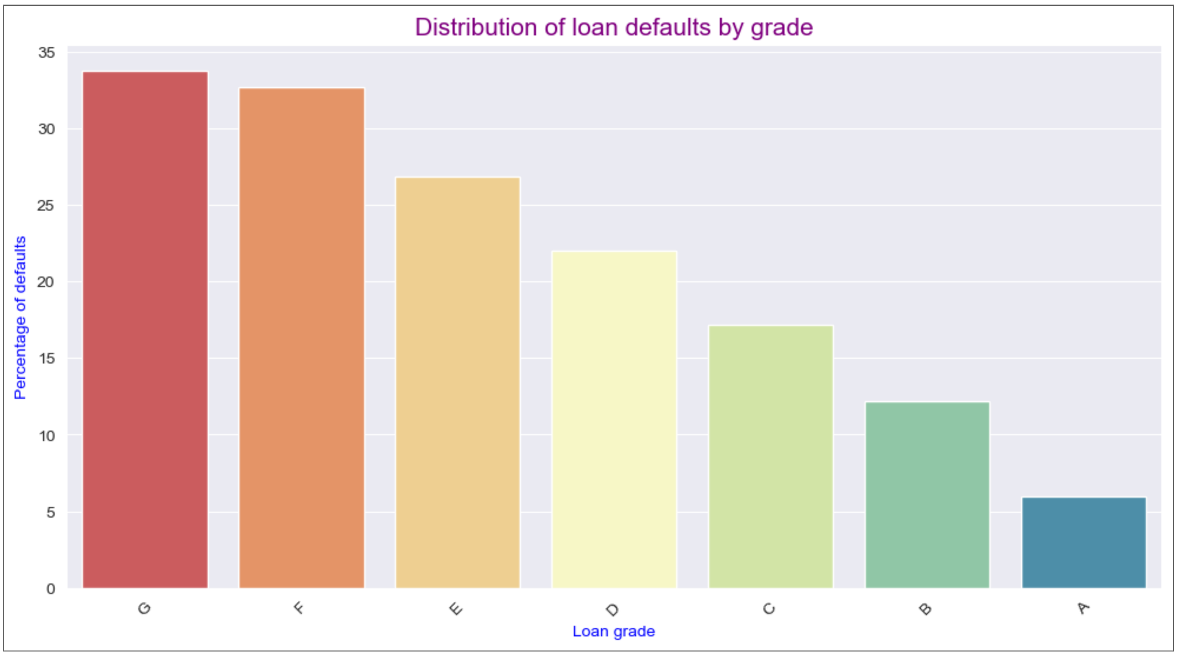
The verification process is apparently effective for loans with low and very low amount but not as effective for higher-amount loans. This necessitates revisiting the verification process at a policy level and making the necessary changes to the existing process, such that the loan applicant profiles for a wide-ranging loan amount are thoroughly verified.

Influence of loan term on loan default


Distribution of loan defaults by tenure
Percentage of defaults
0
5
10
15
20
25
60 months
36 months
Loan term

As observed here, loans with long tenure are more likely to default. The default rate is double over the 60-months tenure compared with that over the 36-months tenure.

Influence of loan grade on the default rate


Distribution of loan defaults by grade
Percentage of defaults
35
30
25
20
15
10
5
0
G
F
E
D
C
B
A
Loan grade

The default percent is observed to increase with the progression from grade A to G. The current dataset does not specify the parameters underlying this grade allocation. Perhaps, insight into those underlying parameters will be more revealing pertaining to this trend noted here.

Overall, based on all the preceding analyses and inferences, we summarize this study as follows:

• The median loan amount, installment, funded amount, and the total amount committed by investors for that loan at that point in time are comparable for both cases of fully paid and charged-off (defaulted) loan status.

• The median loan amount, funded amount, the total amount committed by investors for that loan at that point in time, and the installment vary substantially across the purpose underlying the loan.

• The loans drawn for the purpose of small business are extremely risky, and ~27% of such loans have defaulted.

• Loans for educational purpose are highly likely to default if loan application profile is not verified.

• The state NV has the highest default rate of ~27%.

• The higher-amount loans are more likely to default.

• The verification process is apparently effective for loans with low and very low amount but not as effective for higher-amount loans.

• Loans with long tenure are more likely to default.

• The default percent is observed to increase with the progression from grade A to G.

In particular, it is counter-intuitive that (1) the median annual income is greater in case of defaulted loan status than in case of fully paid loan status, (2) that the number of defaulters is maximum in case of applicants with 10 or more years of experience, and (3) that the default rate is the highest in case of verified profiles.

Overall, based on these observations, we make the following recommendations:

(1) The verification process needs to be revisited at a policy level. Based on the data, verification seems to be apparently effective for loans with low and very low amount but not as effective for higher-amount loans. Necessary changes are required in the existing verification process, such that the loan applicant profiles encompassing diverse ranges of loan amounts are thoroughly verified.

(2) Since the default default percent is observed to increase with the progression from grade A to G, a closer look into the parameters underlying this grade allocation could give us more insight pertaining to this trend noted here. The parameters considered during gradation are not provided in the original data file; thus, we could not proceed with further analysis of such parameters. We recommend invetigation of those parameters underlying gradation.

(3) Since loans characterized by longer terms and higher amounts have shown higher default rates, we recommend deciding on a upper limit on the amount of loan that an applicant can apply for and is sanctioned. Furthermore, shorter terms are recommended.

(4) We recommend treating the loans filed for from NV with additional verification and caution.

(5) The loans applied for the purpose of business should also be treated with additional verification and caution. For these types of loans, specifically focusing on income verification of the application is recommended, as it has been observed that such loans are less likely to default upon income verification.

(6) The loans applied for the purpose of education should be thoroughly verified.

Although by leveraging EDA we have arrived at this stage with the aforementioned inferences and recommendations, to further determine the order of significance and effectiveness of these parameters in predicting loan default, a statistical analysis such as logistic regression or decision tree analysis can be performed. These methods can help identify the most significant predictors of default by assessing the strength of the relationship between each predictor and the outcome variable (default or non-default). After performing the statistical analysis, the parameters can be ranked in order of significance in predicting loan default. This ranking can then be used to develop a predictive model for identifying risky borrowers and mitigating default risk.