Lending Club Case Study

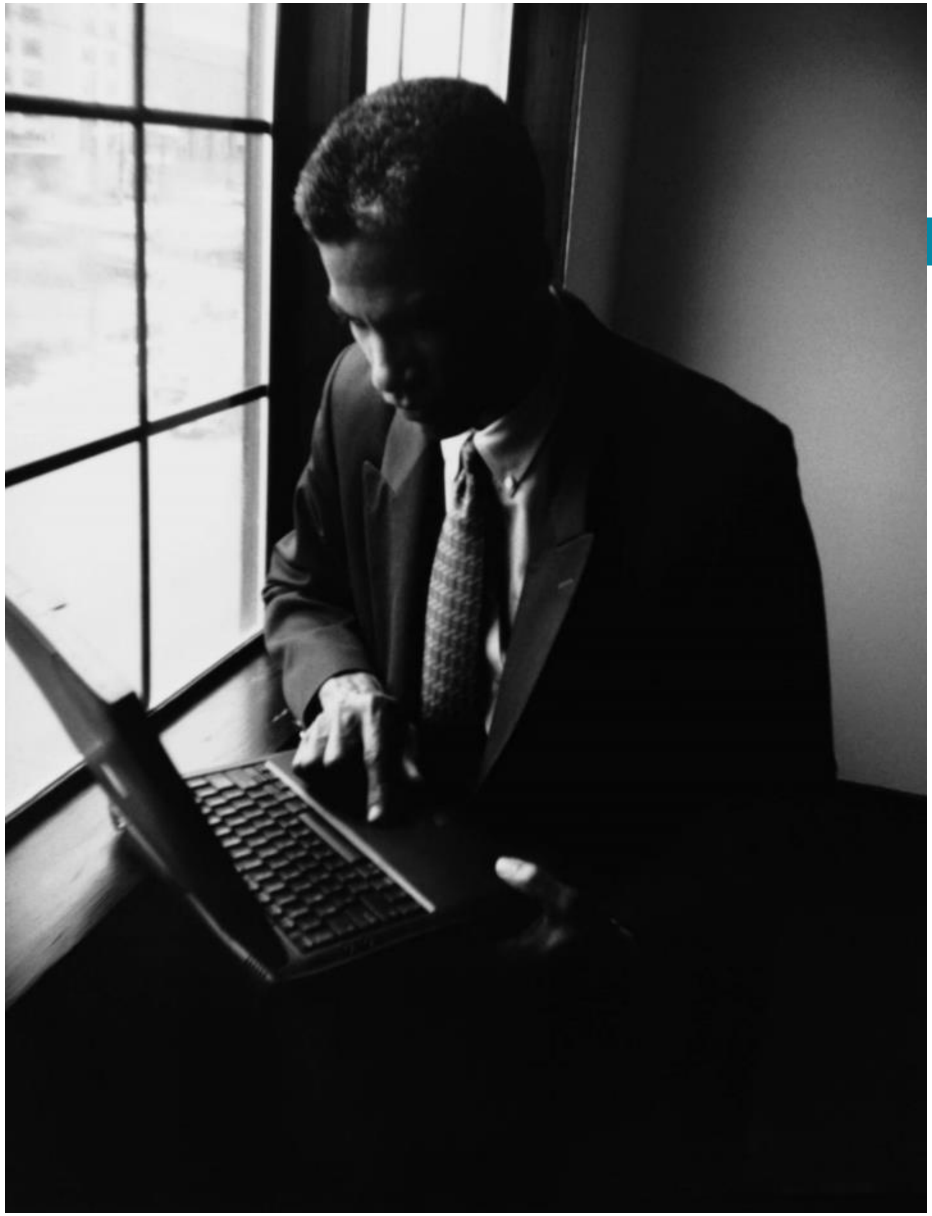
Mukund S – MLC50

OVERVIEW

Over the past few years, we have attempted to collate data over different customer behaviors, patterns, requirements and reason for caution in extreme cases.

We hope this data helps you as our client make an informed decision when investing in a specific customer needs.

OVERVIEW



SOLUTION

DATA UNDERSTANDING

We attempt to decipher the data collected, various variables from the set and their corresponding significance

DATA CLEANING

No Analysis can be completed without cleaning out/replacing missing data, removing unnecessary and redundant columns that otherwise contribute to noise

DATA ANALYSIS

We attempt various methods to analyze the data to arrive at a various pattern's observable from the set

RECOMMENDATIONS

Based on these we provide a comprehensive set of recommendations to help you invest with greater returns

Data Understanding

The Dataset has the following characteristics:

- **111 Columns:**
  - Float Variables 74
  - Int Variables 13
  - String/Object Variables 24
- **54** of these columns contain null values and can be omitted in the next step:
  - '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', etc.
- **39717** rows or entries
- **Variables Types(nonEmpty Columns):**
  - **Ordered Categorical:** grade, sub_grade, emp_length, issue_d, zip_code, earliest_cr_line, etc.
  - **Unordered Categorical:** id, member_id, emp_title, home_ownership, verification_status, loan_status, pymnt_plan, url, etc.
  - **Numerical Variables (nonEmpty Columns):** loan_amnt, funded_amnt, funded_amnt_inv, term, int_rate, Installment, annual_inc, dti, delinq_2yrs, inq_last_6mths, mths_since_last_delinq, mths_since_last_record, etc.

Data Cleaning

We start with identifying and removing columns with only null values

```
RangeIndex: 39717 entries, 0 to 39716
Data columns (total 57 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
```

Anything with a high Percentage of missing data is targeted next

```
next_pymnt_d               97.129693
mths_since_last_record     92.985372
mths_since_last_delinq     64.662487
```

Data Cleaning

Working through the Columns by order of missing values

- **desc** The empty values here can be replaced with 'Other' as this is a description field and can't be easily substituted
- **emp_title** As again here the employer is unknown its better to leave empty columns as 'Other' value
- **emp_length** Based on a similar age group a median value can be substituted
- **title** As this is a categorical field it is better to replace empty values here with 'Other' too
- **revol_util** Drop as this would be a Customer Behavior Variable
- **last_pymnt_d** Drop as this would be a Customer Behavior Variable
- **last_credit_pull_d** Drop as this would be a Customer Behavior Variable
- **collections_12_mths_ex_med** Taking mean value 0
- **chargeoff_within_12_mths** Taking mean value 0
- **pub_rec_bankruptcies** Taking mean value 0
- **tax_liens** Taking mean value 0

```
next_pymnt_d                  97.129693
mths_since_last_record        92.985372
mths_since_last_delinq        64.662487
desc                          32.580507
emp_title                      6.191303
emp_length                     2.706650
pub_rec_bankruptcies           1.754916
last_pymnt_d                   0.178765
collections_12_mths_ex_med     0.140998
chargeoff_within_12_mths       0.140998
revol_util                     0.125891
tax_liens                      0.098195
title                          0.027696
last_credit_pull_d             0.005036
```

Data Cleaning

We finally drop the columns that are Customer behavior variables

Customer behaviour variables
delinq_2yrs
earliest_cr_line
inq_last_6mths
open_acc
pub_rec
revol_bal
revol_util
total_acc
out_prncp
out_prncp_inv
total_pymnt
total_pymnt_inv
total_rec_prncp
total_rec_int
total_rec_late_fee
recoveries
collection_recovery_fee
last_pymnt_d
last_pymnt_amnt
last_credit_pull_d
application_type

Data Cleaning

We then drop Columns with equal data for all rows

	pymnt_plan	initial_list_status	collections_12_mths_ex_med	policy_code	acc_now_delinq	chargeoff_within_12_mths	delinq_amnt	tax_liens
0	n	f	0.0	1	0	0.0	0	0.0
1	n	f	0.0	1	0	0.0	0	0.0
2	n	f	0.0	1	0	0.0	0	0.0
3	n	f	0.0	1	0	0.0	0	0.0
4	n	f	0.0	1	0	0.0	0	0.0
...	...	...	...	...	...	...	...	...
39712	n	f	0.0	1	0	0.0	0	0.0
39713	n	f	0.0	1	0	0.0	0	0.0
39714	n	f	0.0	1	0	0.0	0	0.0
39715	n	f	0.0	1	0	0.0	0	0.0
39716	n	f	0.0	1	0	0.0	0	0.0

39717 rows × 8 columns

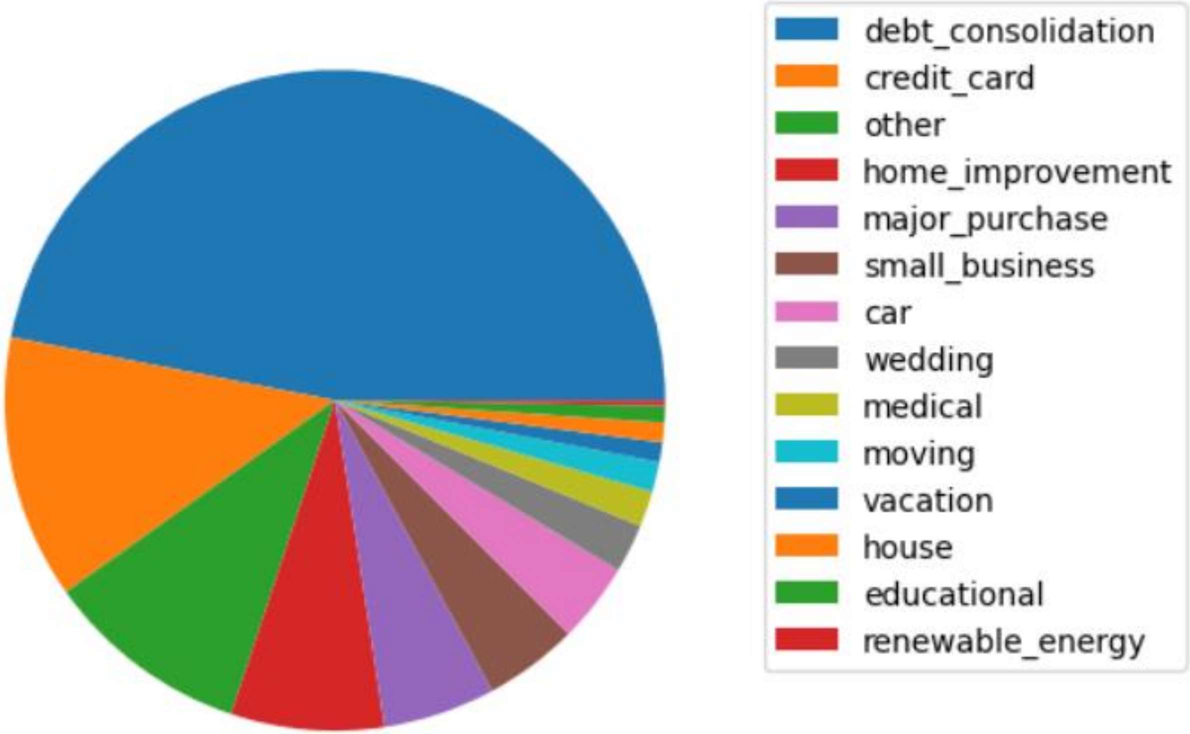
Data Cleaning

After some final set of cleaning and data type conversions here's what our initial 111 column dataset looks like

```
Data columns (total 24 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  int64
 6   int_rate              39717 non-null  float64
 7   installment           39717 non-null  float64
 8   grade                 39717 non-null  object
 9   sub_grade             39717 non-null  object
 10  emp_title             39717 non-null  object
 11  emp_length            39717 non-null  int64
 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  datetime64[ns]
 16  loan_status           39717 non-null  object
 17  desc                  39717 non-null  object
 18  purpose               39717 non-null  object
 19  title                 39717 non-null  object
 20  zip_code              39717 non-null  object
 21  addr_state            39717 non-null  object
 22  dti                   39717 non-null  float64
 23  pub_rec_bankruptcies  39717 non-null  float64
```

Data Analysis - Univariate

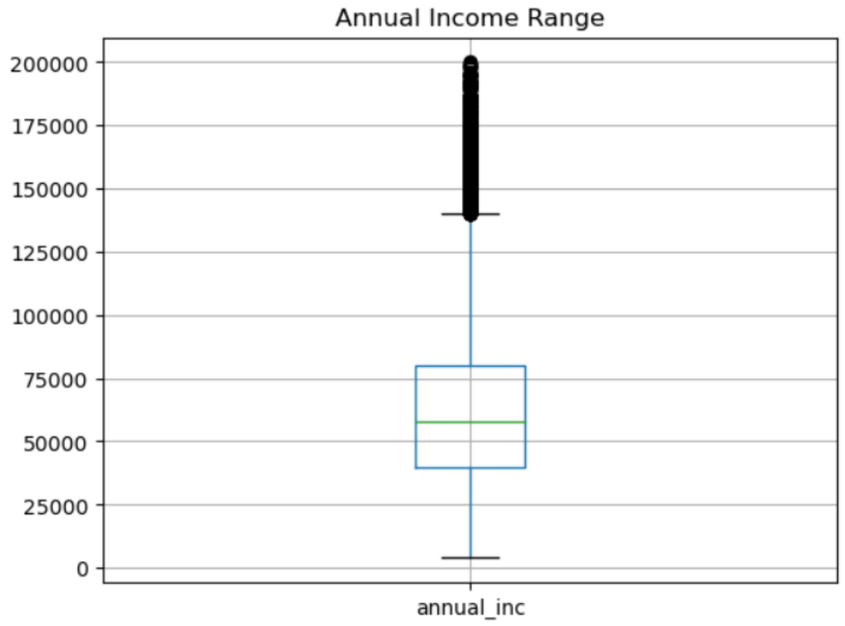
We took a look into the different loan purposes and a large subset appears to be set towards debt consolidation and credit card payments

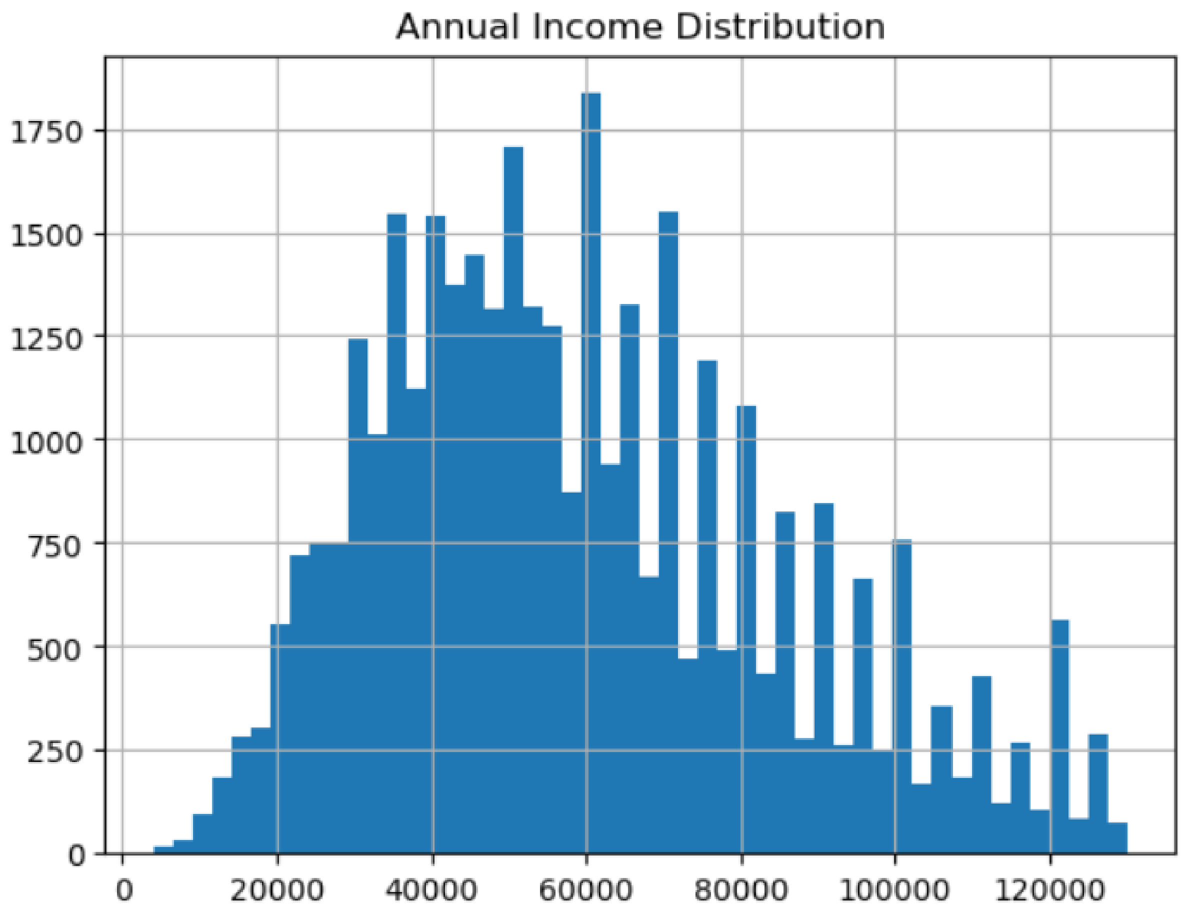


Data Analysis - Univariate

A quick look into the income range our customers earn on an average with some earning over the norm

Most earn below $130K PA and are roughly around the 40K to 80K range

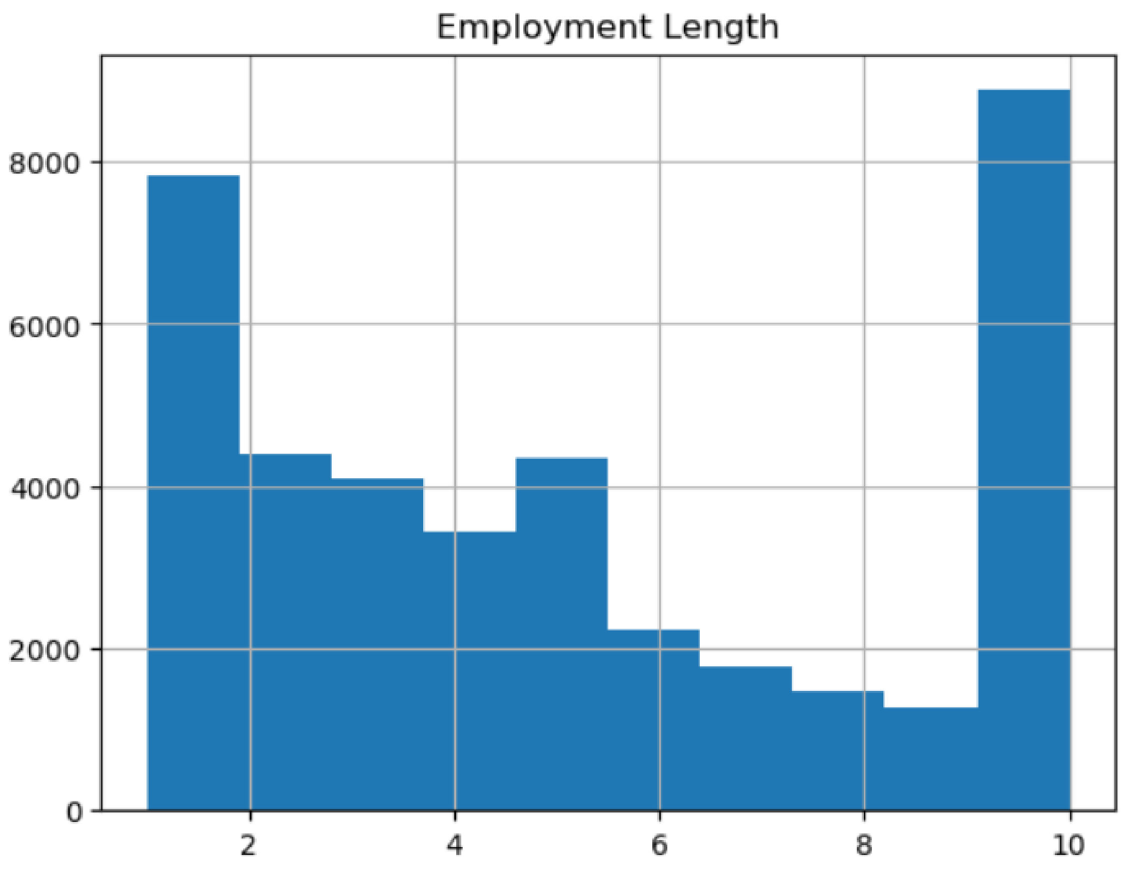




Data Analysis - Univariate

Average length of Employment

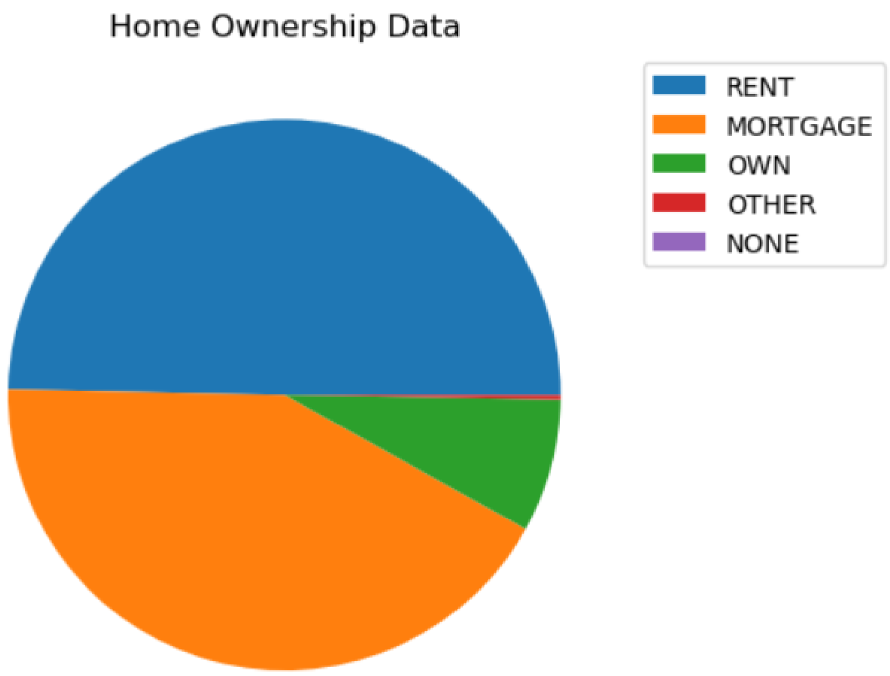
There's an interesting distribution here of folks both in the 1-2 year career mark and those well over 10.



Data Analysis - Univariate

Home Ownership

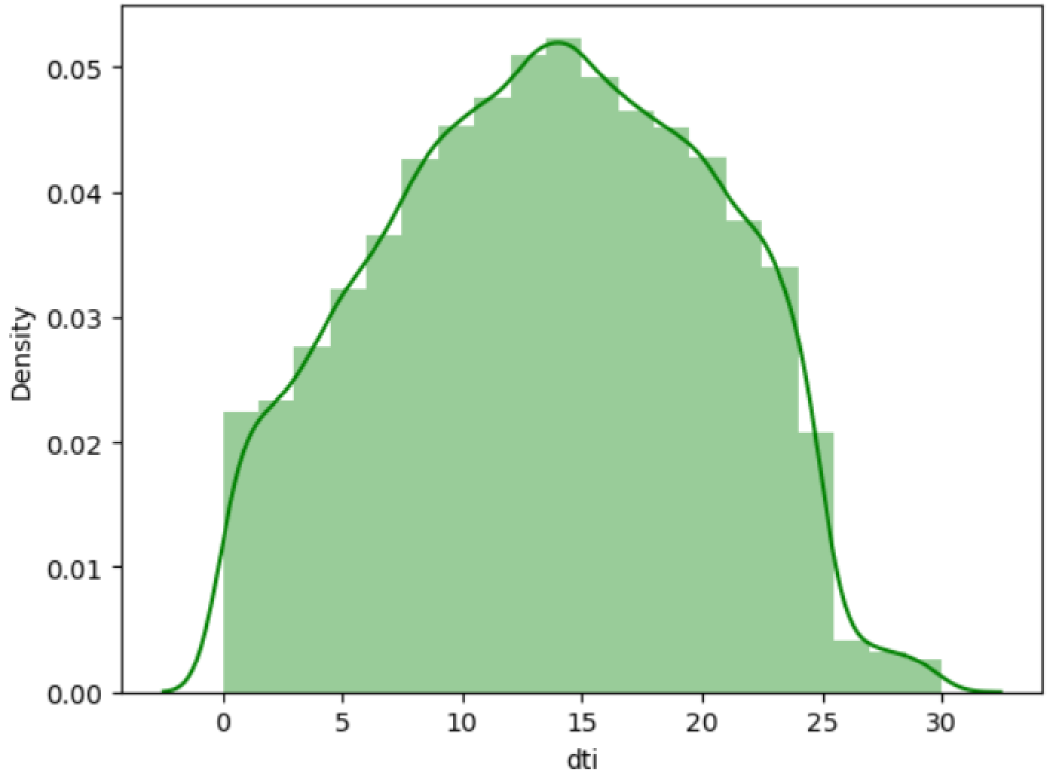
Another interesting insight here is that 50% stay at rented premises. Though it may appear that the other 50 own, a large portion of these have those homes under mortgage which might also imply a large part of the income going into those payments



Data Analysis - Univariate

Debt to Income Ratio

A large subset of our customers have a health 15% DTI denoting a good capacity to repay.



Data Analysis – Segmented Univariate

State Distribution

A larger portion of our customers are from the state of California which might help us if we're looking to scale up the branch or employees in the location to cater to the demand

	count	mean	std	min	25%	50%	75%	max
addr_state								
CA	6397	61538	26700	4080	40800	59000	79000	129996
NY	3404	60046	25133	7200	41000	55016	75000	129996
FL	2622	55651	25177	4000	37000	51000	70003	129996
TX	2424	61531	26791	4800	41912	58000	80000	129996
NJ	1627	62571	26070	6000	42702	60000	80000	129996
PA	1400	56246	25435	6000	37000	50815	73614	129996
IL	1377	59797	25789	8000	40000	56004	75288	129996
VA	1276	64880	27050	4200	43150	61950	84000	129700
GA	1257	59365	24919	6000	40000	56000	75000	128000
MA	1180	60673	25527	4200	41930	58000	76000	129600

Data Analysis – Segmented Univariate

Loan Purpose

Here's a quick look into the loan requirement reasons and their contributing factors. A good portion of our large loans go into financing small businesses and at the same time we cater to the smallest of loan requirements such as vacations.

purpose	count	mean	std	min	25%	50%	75%	max
small_business	1584	12229	7802	500	6000	10000	16000	35000
debt_consolidation	16998	11796	6825	700	6500	10000	15000	35000
house	332	11396	6922	1200	6000	10000	15000	35000
credit_card	4678	10914	6367	725	6000	10000	14500	35000
home_improvement	2458	9976	6868	900	5000	8000	13000	35000
wedding	870	9241	5660	1000	5000	8000	12000	35000
renewable_energy	93	7778	6351	1000	3200	5600	10875	35000
medical	629	7646	5471	1000	4000	6000	10000	35000
major_purchase	2011	7471	5327	1000	4000	6000	10000	35000
other	3660	7426	5703	500	3369	6000	10000	35000
car	1431	6545	3868	1000	4000	5600	8000	30000
educational	311	6495	4856	900	3000	5000	8400	25000
moving	531	5780	4552	1000	3000	4800	7000	35000
vacation	360	5177	3946	500	2475	4200	6400	29700

Data Analysis – Segmented Univariate

Employment Duration and corresponding loan amounts

emp_length	count	mean	std	min	25%	50%	75%	max
10	7652	11925	7517	1000	6000	10000	16000	35000
9	1132	11302	6619	1000	6000	10000	15000	35000
7	1616	10916	6718	500	6000	10000	15000	35000
8	1315	10899	6908	1000	6000	9600	15000	35000
6	2033	10669	6695	1000	5500	9600	14400	35000
4	3152	10254	6528	900	5000	9000	14000	35000
3	3748	9927	6279	500	5000	8800	13000	35000
5	4008	9882	6618	1000	5000	8000	13406	35000
2	4040	9444	6112	800	5000	8000	12000	35000
1	7250	9150	6118	500	4800	7750	12000	35000

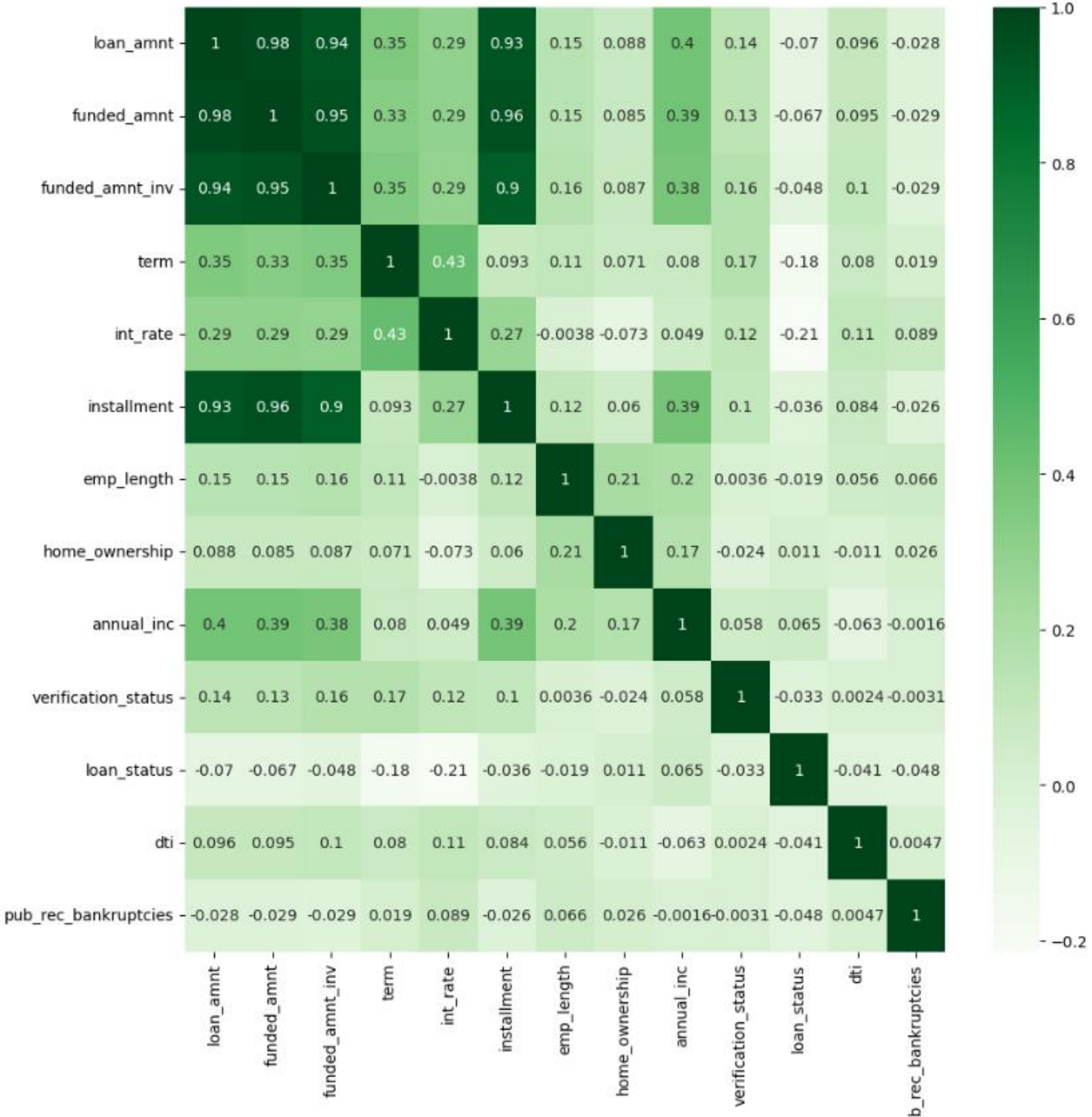
As we noticed the pattern earlier our customer demographic are from both the experienced and fresher category. However, we might need to look at the max loan amounts dispensed here for lower employment duration as this ceiling appears to be set uniformly for all customers.

Data Analysis – Bivariate

Heatmap

A few co-related patterns observed here:

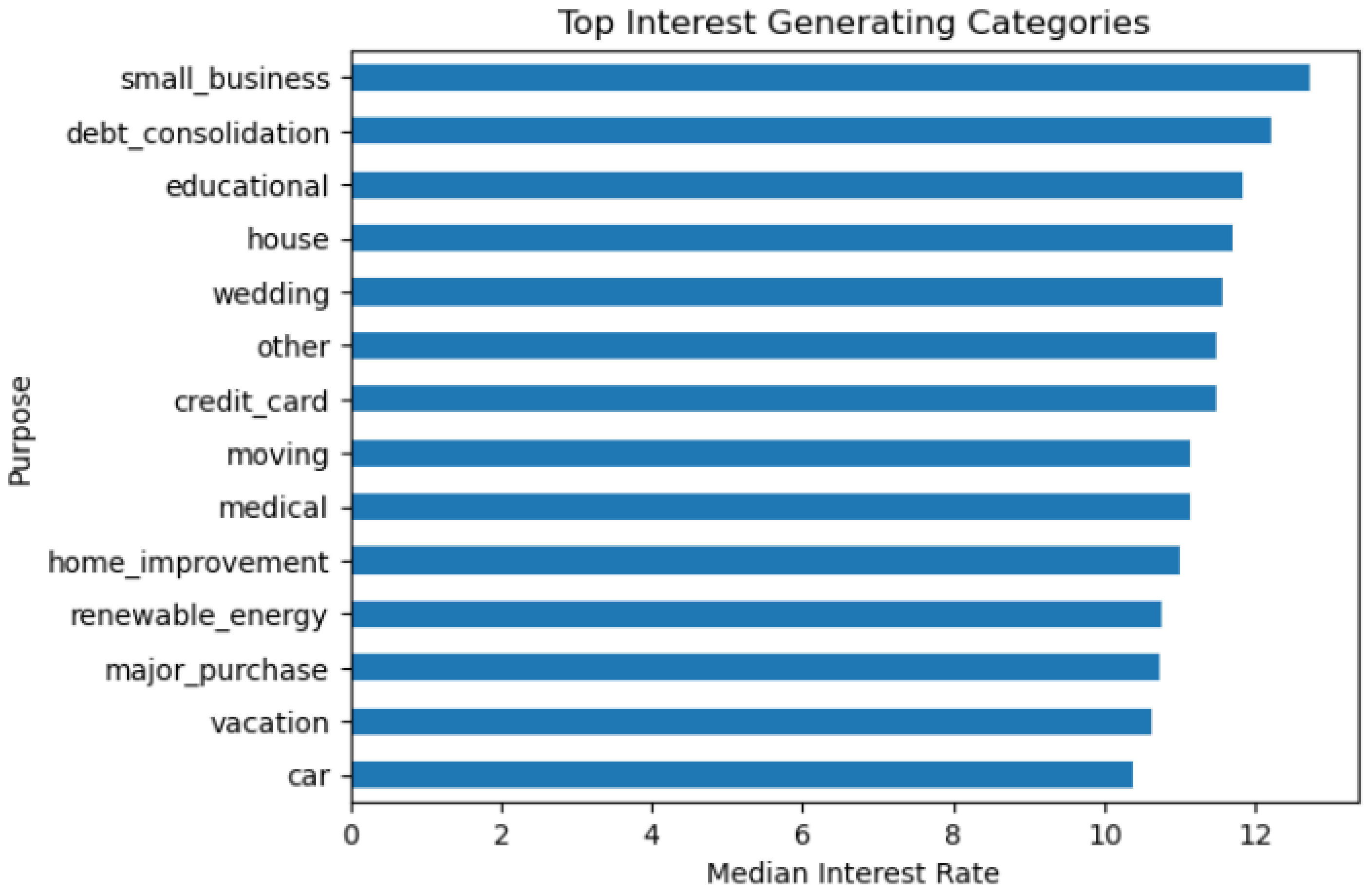
- The final funded amount appears to be around 95% correlated to the actual request amount
- Factors that co-related to the annual income even if partially:
    - Loan amount
    - Installments
    - Small value but interesting – Employment length & Home Ownership



Data Analysis – Bivariate

Top Interest generating purposes

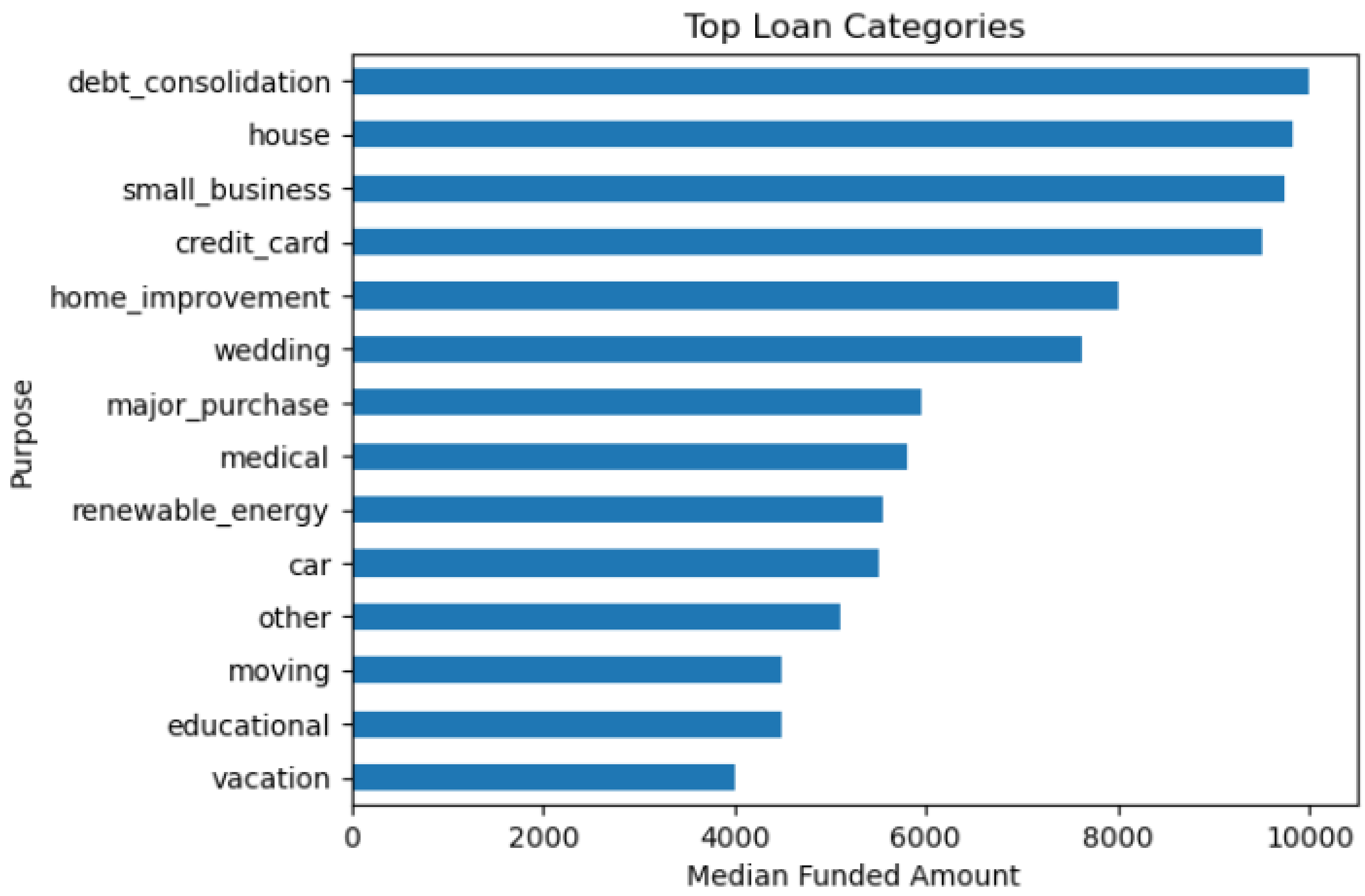
Here again, our best earnings come from small businesses



Data Analysis – Bivariate

Top loan requirement purposes

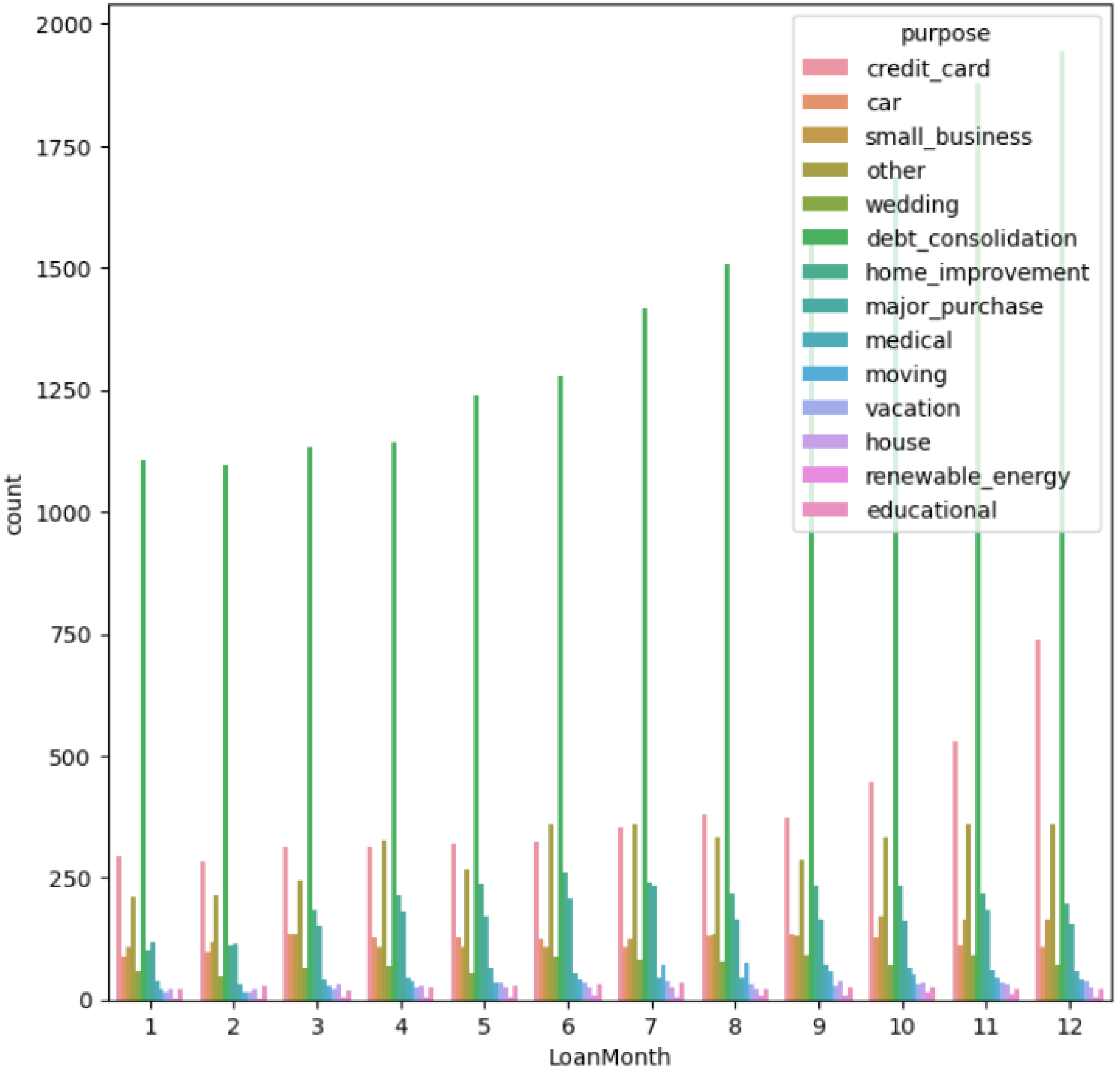
Here again, our best earnings come from small businesses



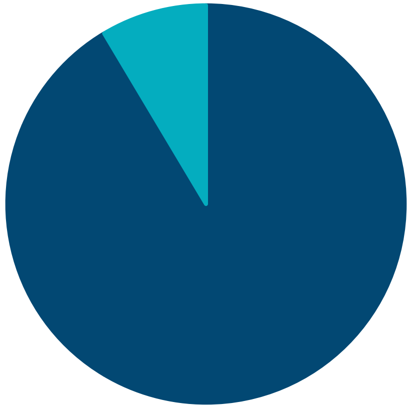
Data Analysis – Derived Columns

We attempted to analyze the months we noticed a spike in requirements and this appears to follow a progressive increase into the end of the year.

A large spike is noticed in debt consolidation and education.

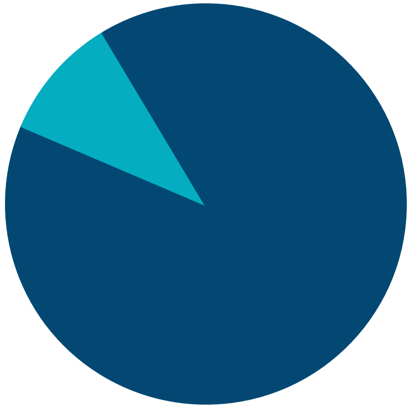


RECOMMENDATIONS



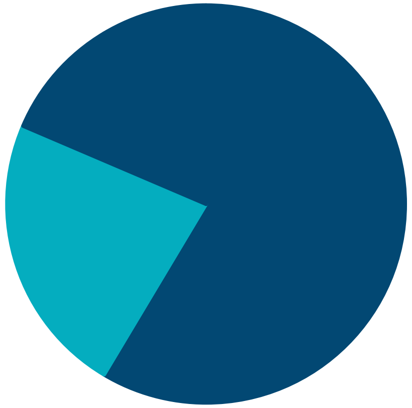
DEBT CONSOLIDATION

A Large part of our market appears to be from debt consolidation and credit card. We might want to gather additional insights on past loans



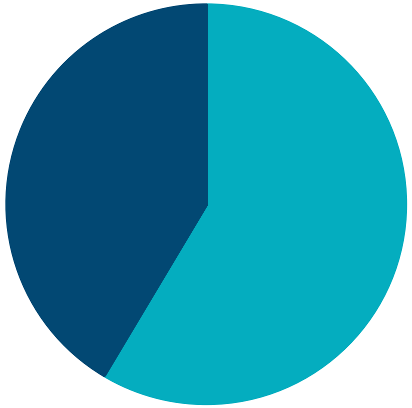
MAX LOAN Ceiling

This currently appears to be 35k for both 1 year and 10 year experienced customers. We might want to either lower the ceiling for freshers or increase for experienced



OFFICE EXPANSION IN CA

As majority of our customers are from CA we might want to expand our business there and also increase advertising in the other regions



SMALL BUSINESS SEGMENT

We can attempt to diversify more into funding small businesses as these appear to generate more investment return

THANK YOU

MUKUND S

MLC50