LENDING CLUB CASE STUDY

Group Facilitator & Member Details :

Name: Soumayadeep Manna

Email: soumayadeepmanna@gmail.com

Contact: 7003706325

Table of Contents

01. Introduction

Brief introduction about the case study

02. Business Requirements

Brief idea about the business requirements and details

03. Data understanding

Understanding the data and performing data cleaning steps

04. Data analysis

Analyzing the data and deriving driving factors and impact parameters

05. Conclusion

Concluding analysis with validated outcome

Introduction

Lending Club is a consumer finance company which specializes in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

1. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company

2. If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company

Business Requirements

When a person applies for a loan, there are **two types of decisions** that could be taken by the company:

1. Loan accepted: If the company approves the loan, there are 3 possible scenarios described below:

1. ***Fully paid***: *Applicant has fully paid the loan (the principal and the interest rate)*
2. ***Current***: *Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.*
3. ***Charged-off***: *Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has* ***defaulted*** *on the loan*

2. Loan rejected: The company had rejected the loan (because the candidate does not meet their requirements etc.). Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available with the company

Business Requirements

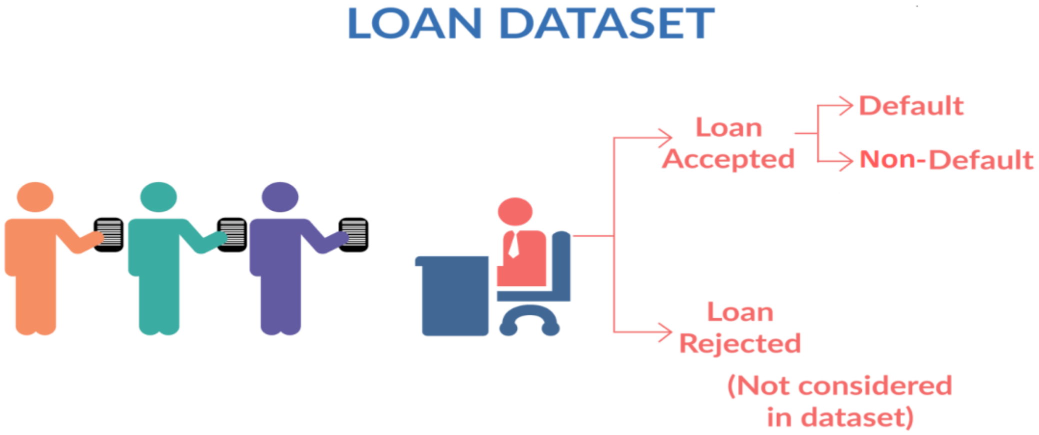
Objectives:

1. Prepare the data for the analysis

2. Finding the major risk factors for the loan lending or approvals

3. Identify the driving factors of the risk analysis

4. Reduce the risk factor for the lending of loan by identifying the risks.



Data Understanding

Lending Club dataset provided which contains the complete loan data for all loans issued through the time period 2007 to 2011.

1. **Sourcing the Data into Data frame for the analysis**
2. **Data Cleaning**

    1.Fix rows and columns

    2. Fix missing values

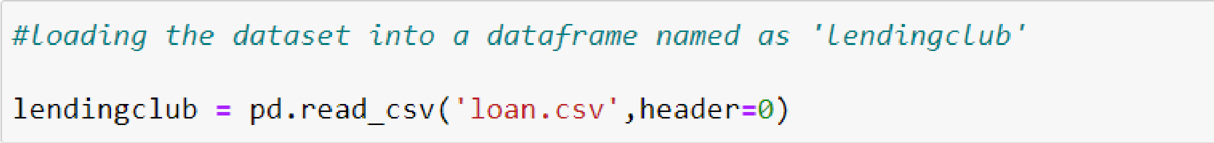
    3. Standardise values

    4. Fix invalid values

    5. Filter data

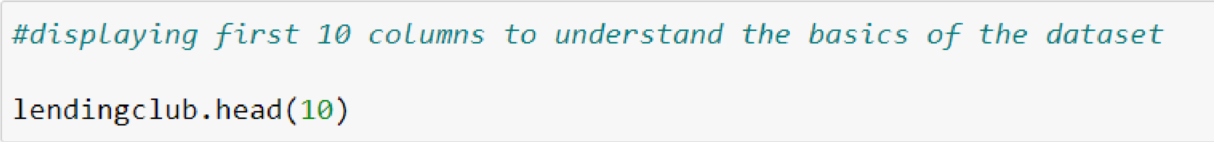
3. Preparing the data for the analysis

1. Data Sourcing and Understanding



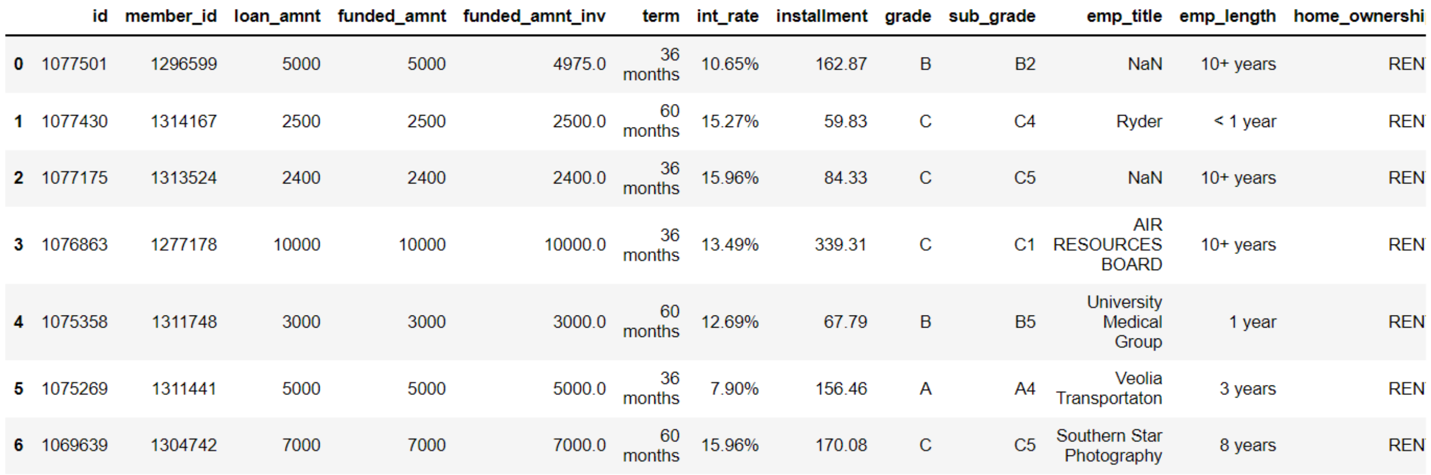
```
#loading the dataset into a dataframe named as 'lendingclub'

lendingclub = pd.read_csv('loan.csv',header=0)
```



```
#displaying first 10 columns to understand the basics of the dataset

lendingclub.head(10)
```

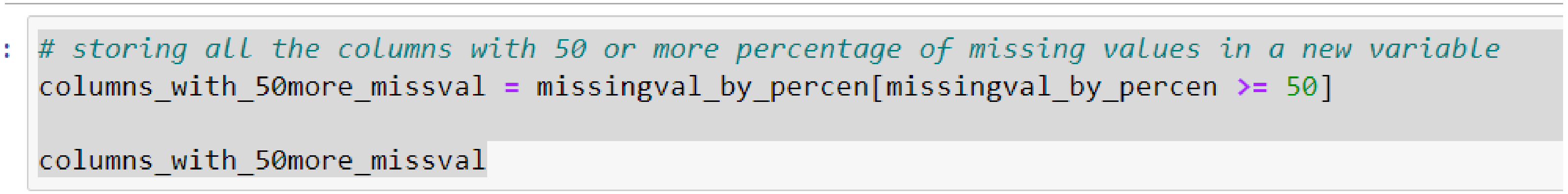


	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	emp_title	emp_length	home_ownershi
0	1077501	1296599	5000	5000	4975.0	36 months	10.65%	162.87	B	B2	NaN	10+ years	REN
1	1077430	1314167	2500	2500	2500.0	60 months	15.27%	59.83	C	C4	Ryder	< 1 year	REN
2	1077175	1313524	2400	2400	2400.0	36 months	15.96%	84.33	C	C5	NaN	10+ years	REN
3	1076863	1277178	10000	10000	10000.0	36 months	13.49%	339.31	C	C1	AIR RESOURCES BOARD	10+ years	REN
4	1075358	1311748	3000	3000	3000.0	60 months	12.69%	67.79	B	B5	University Medical Group	1 year	REN
5	1075269	1311441	5000	5000	5000.0	36 months	7.90%	156.46	A	A4	Veolia Transportaton	3 years	REN
6	1069639	1304742	7000	7000	7000.0	60 months	15.96%	170.08	C	C5	Southern Star Photography	8 years	REN

2. Data Cleaning

2.1 *Fixing Columns*

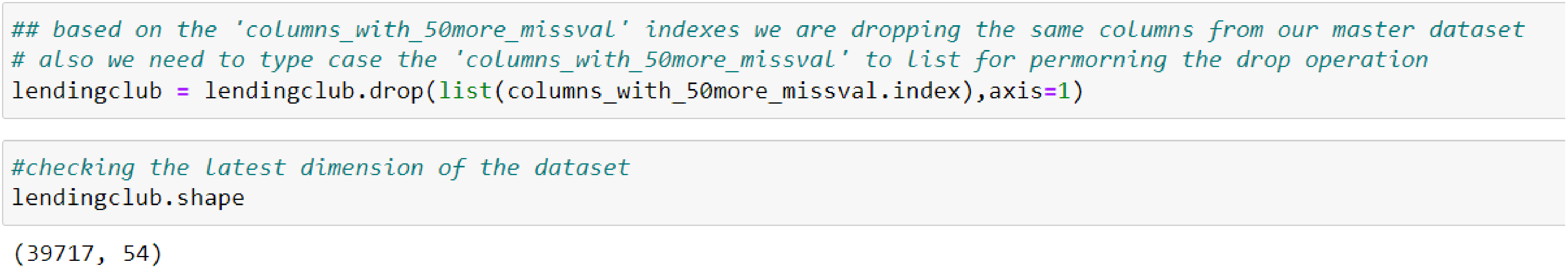
- Filtering columns where missing value percentage is greater than equals 50



```
# storing all the columns with 50 or more percentage of missing values in a new variable
columns_with_50more_missval = missingval_by_percen[missingval_by_percen >= 50]

columns_with_50more_missval
```

- Dropping those columns from the dataset



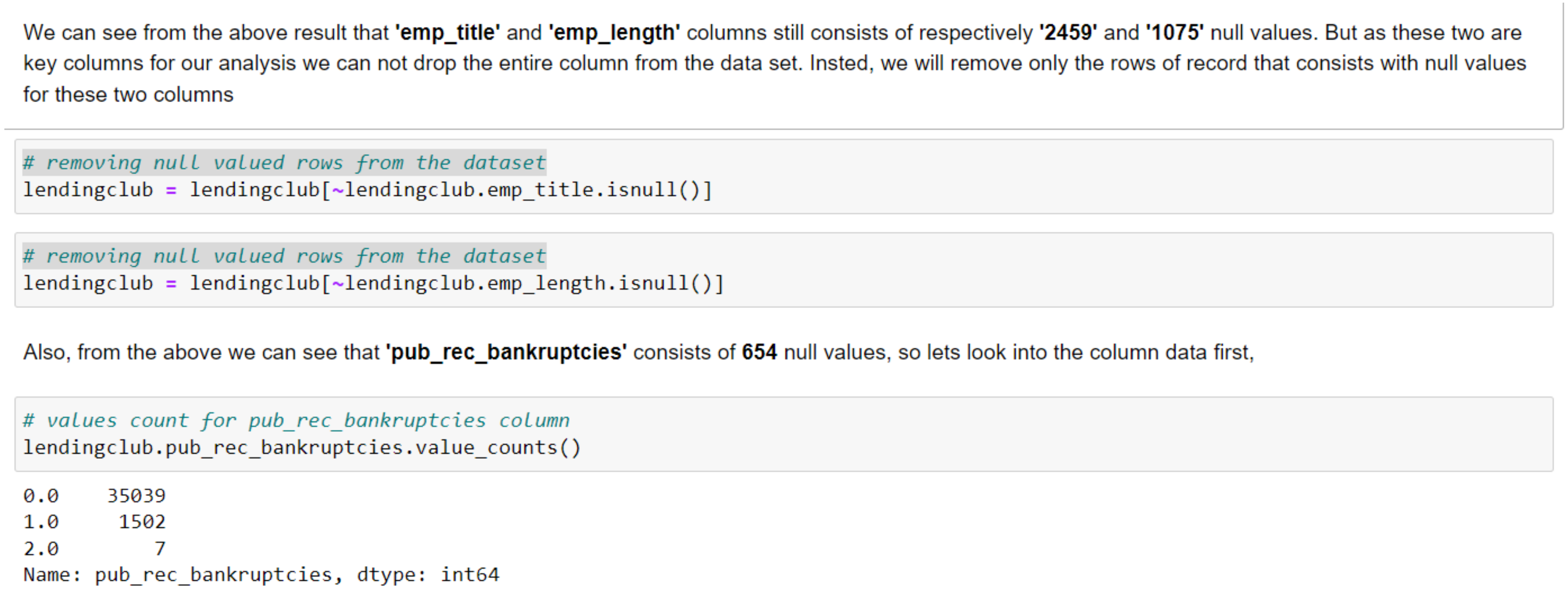
```
## based on the 'columns_with_50more_missval' indexes we are dropping the same columns from our master dataset
# also we need to type case the 'columns_with_50more_missval' to list for permorning the drop operation
lendingclub = lendingclub.drop(list(columns_with_50more_missval.index),axis=1)
```

```
#checking the latest dimension of the dataset
lendingclub.shape
```

```
(39717, 54)
```

2.2 Fixing Missing Values

- We are finding out the count of missing values per columns if the values are not high, we are row wise dropping those missing values *(also if possible we can replace those values by mean or average as well)*



We can see from the above result that **'emp_title'** and **'emp_length'** columns still consists of respectively **'2459'** and **'1075'** null values. But as these two are key columns for our analysis we can not drop the entire column from the data set. Insted, we will remove only the rows of record that consists with null values for these two columns

```
# removing null valued rows from the dataset
lendingclub = lendingclub[~lendingclub.emp_title.isnull()]
```

```
# removing null valued rows from the dataset
lendingclub = lendingclub[~lendingclub.emp_length.isnull()]
```

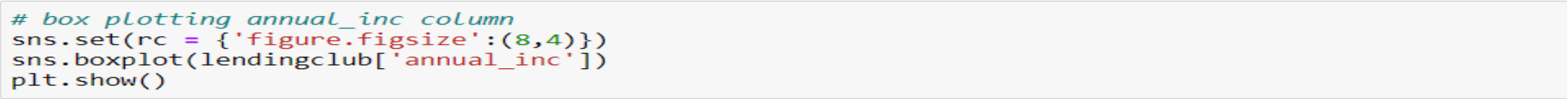
Also, from the above we can see that **'pub_rec_bankruptcies'** consists of **654** null values, so lets look into the column data first,

```
# values count for pub_rec_bankruptcies column
lendingclub.pub_rec_bankruptcies.value_counts()
```

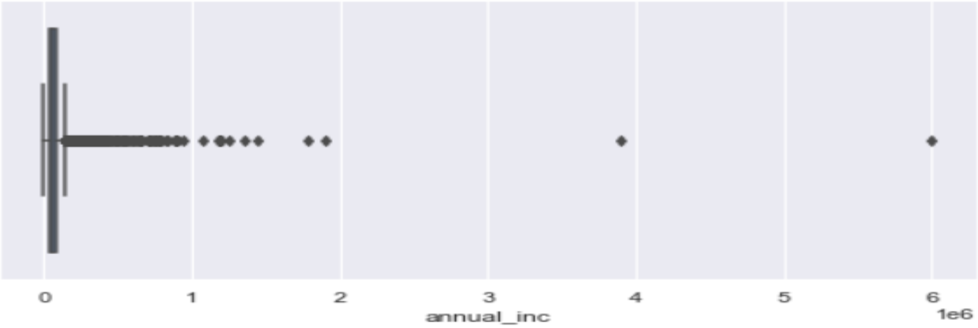
```
0.0    35039
1.0     1502
2.0        7
Name: pub_rec_bankruptcies, dtype: int64
```

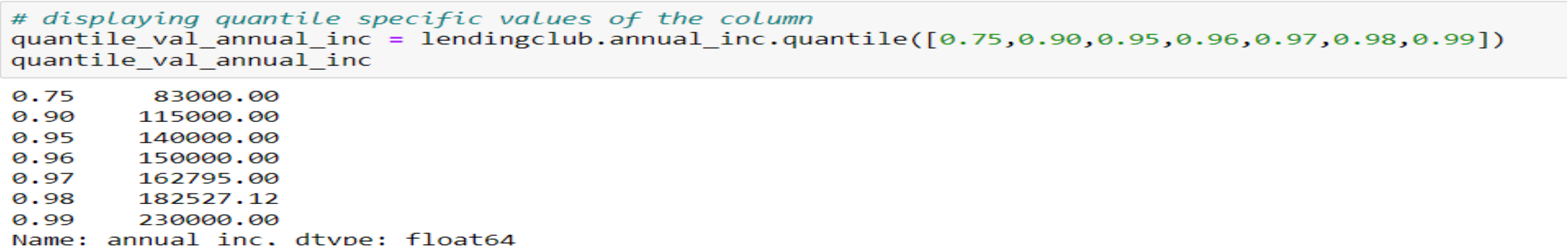
2.3 Standardising Values

- Outliers treatment is one of the methods to remove unnecessary high values from the data so that the result not get effected and the column values are standardised
- here we performed the same and removed outliers values from few columns



```
# box plotting annual_inc column
sns.set(rc = {'figure.figsize':(8,4)})
sns.boxplot(lendingclub['annual_inc'])
plt.show()
```



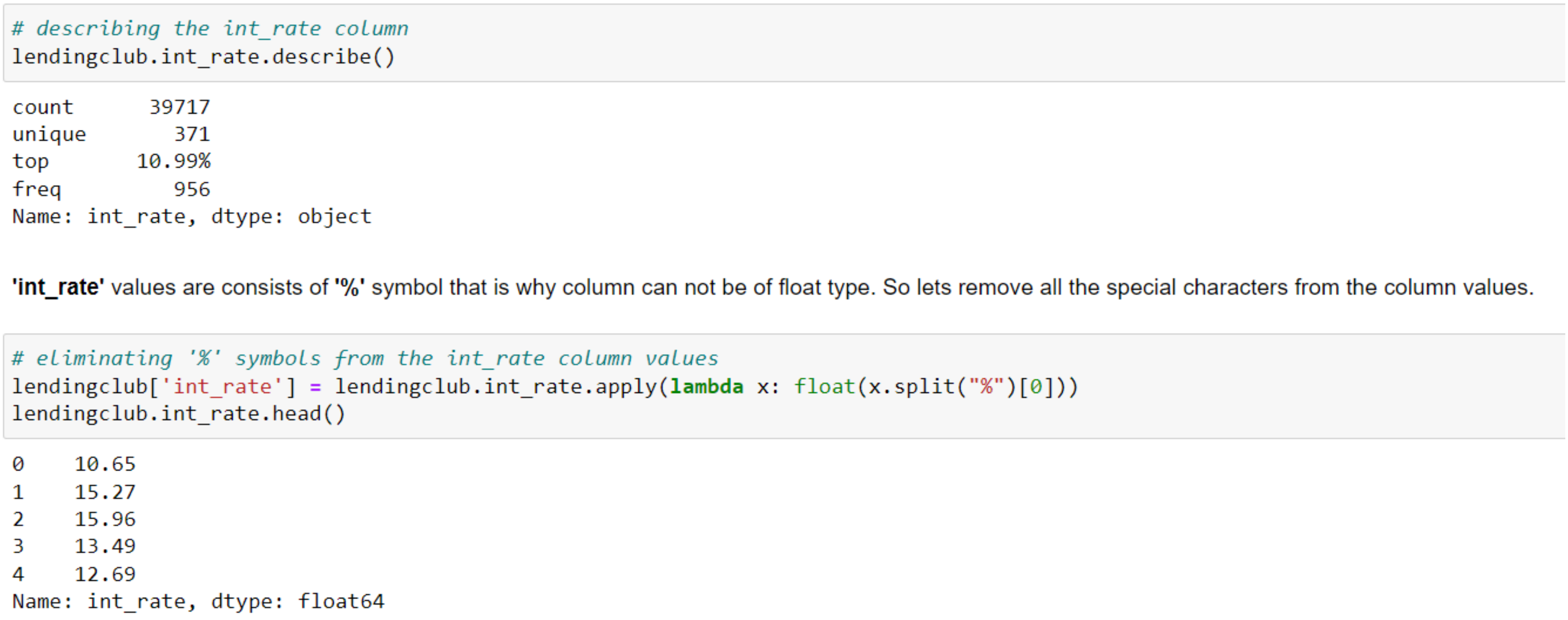


```
# displaying quantile specific values of the column
quantile_val_annual_inc = lendingclub.annual_inc.quantile([0.75,0.90,0.95,0.96,0.97,0.98,0.99])
quantile_val_annual_inc
```

```
0.75      83000.00
0.90     115000.00
0.95     140000.00
0.96     150000.00
0.97     162795.00
0.98     182527.12
0.99     230000.00
Name: annual_inc, dtype: float64
```

2.3 Fix invalid Values

- Sometimes few data presented in the dataset in invalid format as part of data cleaning those column values needs to be fixed and presented in valid format
- here we performed the same interest rate is a numeric value but due to % symbol it considered as object so we removed the symbol and converted to numeric



```
# describing the int_rate column
lendingclub.int_rate.describe()
```

```
count      39717
unique       371
top       10.99%
freq         956
Name: int_rate, dtype: object
```

'int_rate' values are consists of **'%'** symbol that is why column can not be of float type. So lets remove all the special characters from the column values.

```
# eliminating '%' symbols from the int_rate column values
lendingclub['int_rate'] = lendingclub.int_rate.apply(lambda x: float(x.split("%")[0]))
lendingclub.int_rate.head()
```

```
0    10.65
1    15.27
2    15.96
3    13.49
4    12.69
Name: int_rate, dtype: float64
```

“

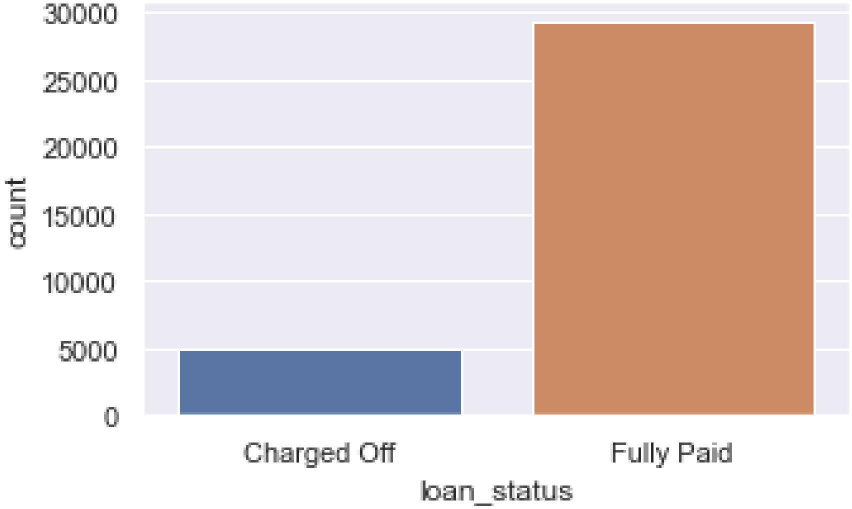
Data Analysis :

Once all the data cleaning is done, and we have a clean dataset. We can start with data analysis steps.

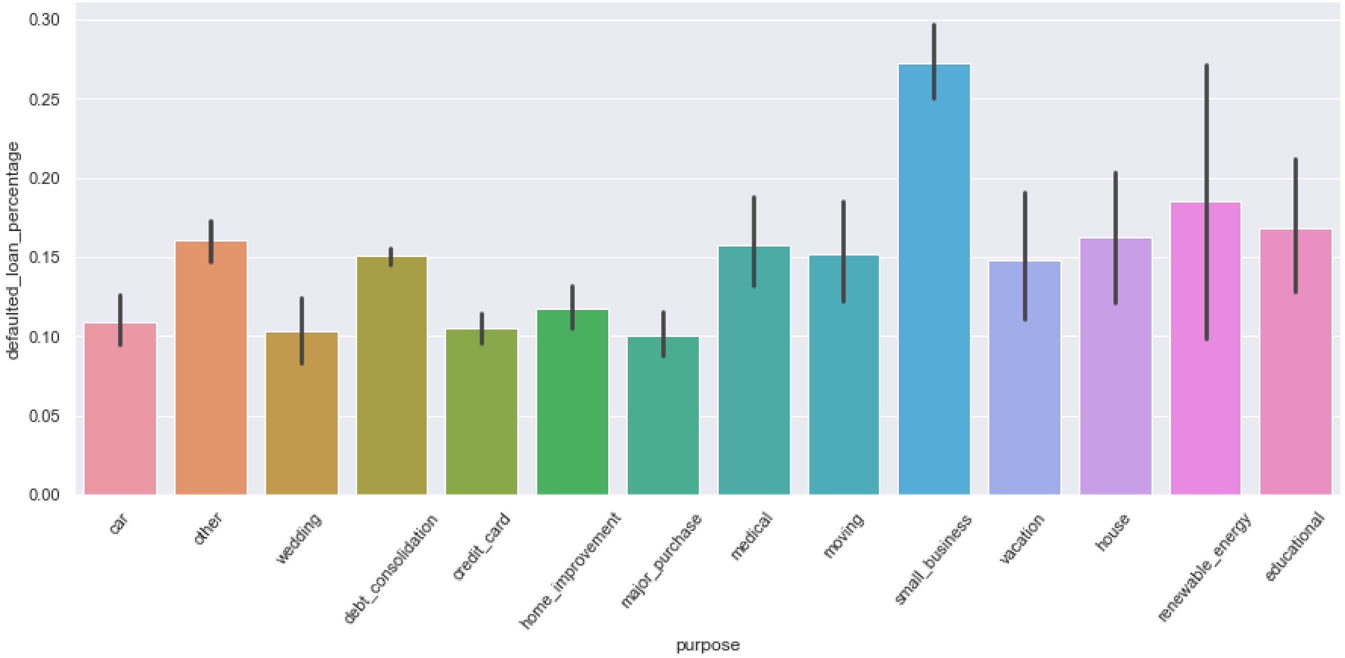
Univariate Analysis

As the term "univariate" suggests, this deals with analyzing variables one at a time. It is important to separately understand each variable before moving on to analyzing multiple variables together.

Univariate analysis helped to understand the singular behaviour the columns and their impact on the analysis in our case study. Lets look at the few of the univariate analysis :

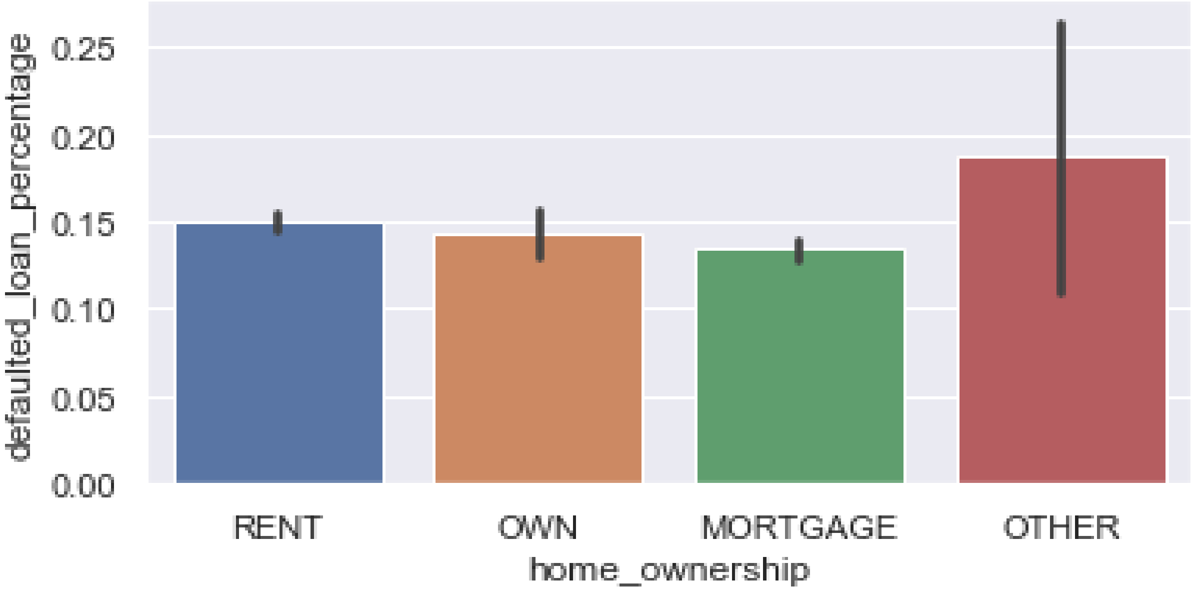


- Loan Status have two different categories where Fully paid lies around 25-30 k and Charged off lies between 0-5k

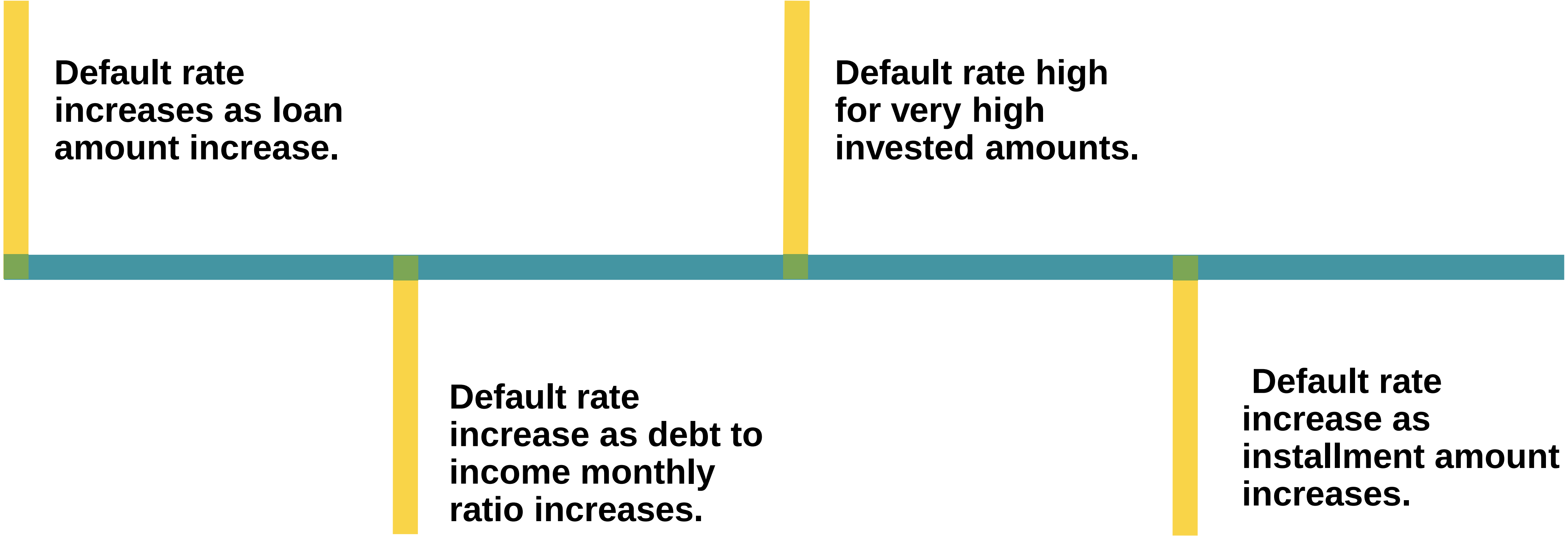


<<< Small business loans default the most, then renewable energy and education

Home ownership is not much effective as defaulting rate is almost avg for all the categories >>>



Few More Observations from the univariate analysis:



Default rate increases as loan amount increase.

Default rate increase as debt to income monthly ratio increases.

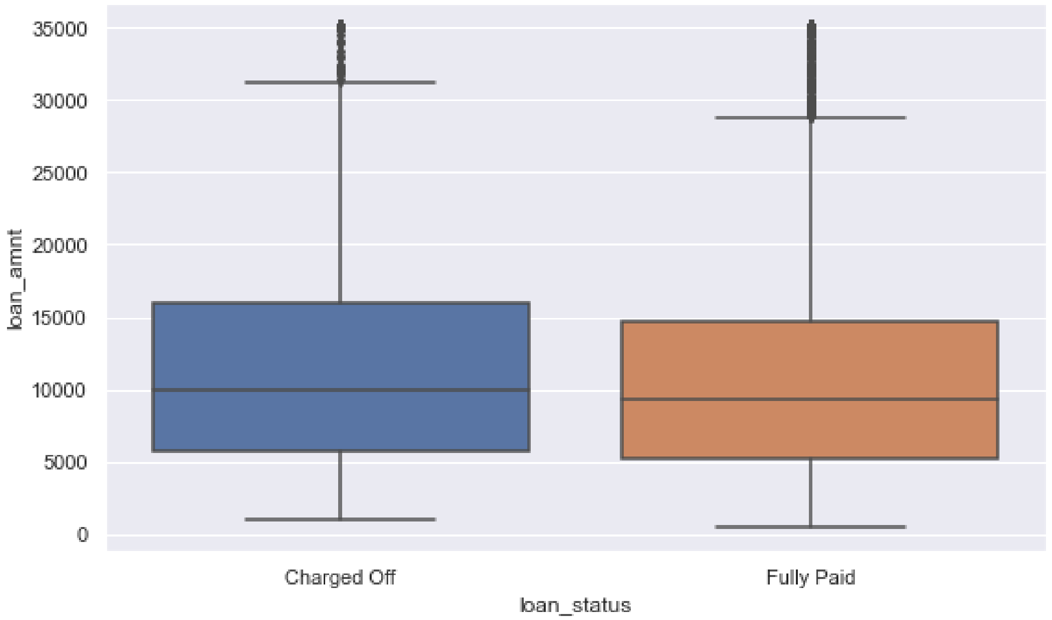
Default rate high for very high invested amounts.

Default rate increase as installment amount increases.

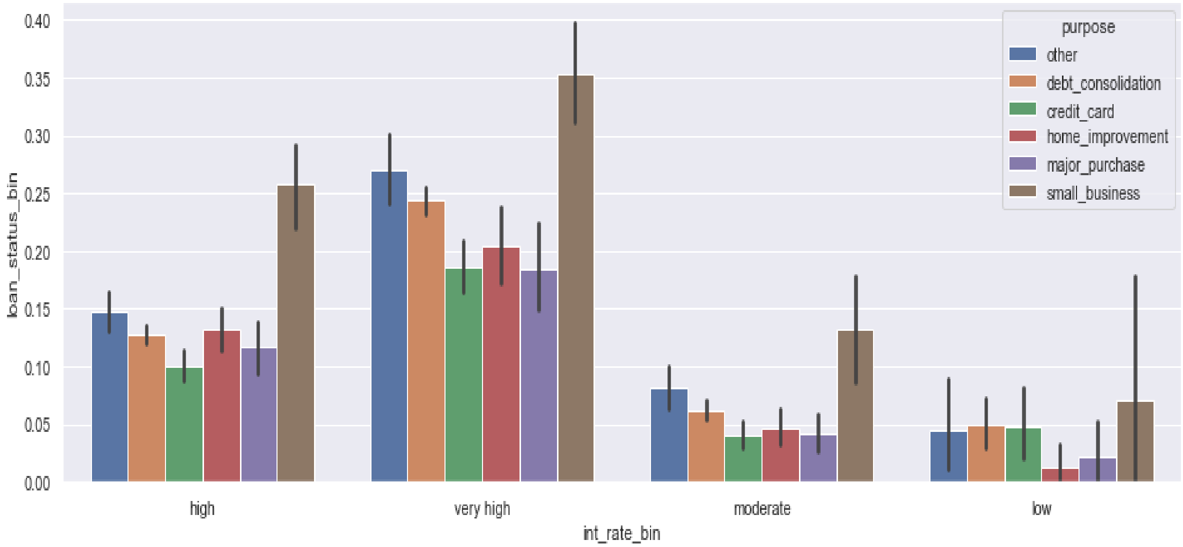
Segmented and Bivariate Analysis

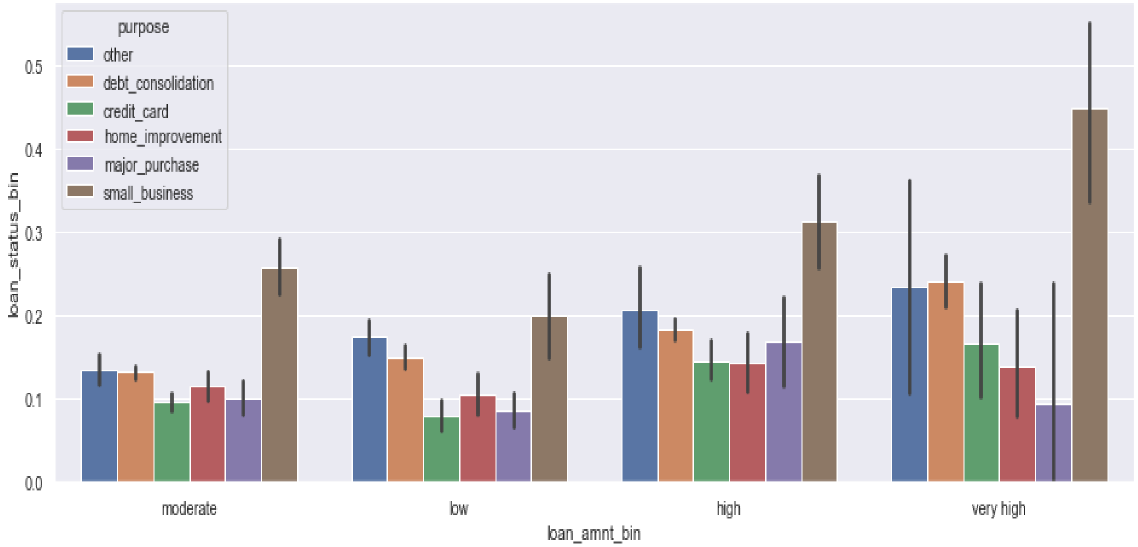
As part of the further analysis we will be performing segmented univariate and bivariate analysis on the columns to get more detailed insights on the records.

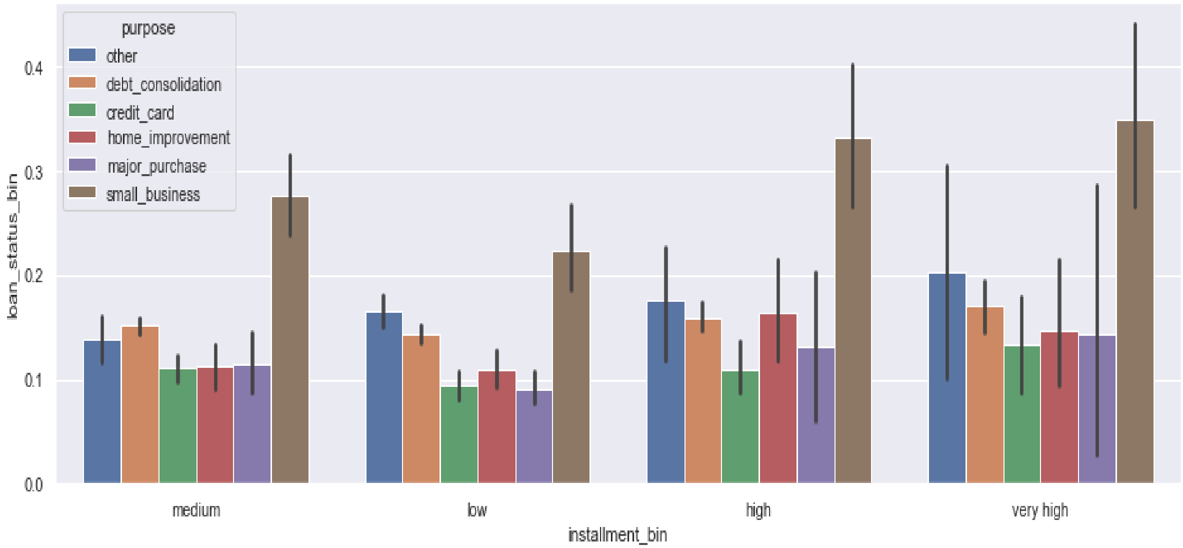
Based on those analysis we will be giving our observations more precisely.



- **More the loan amount increase there are high chances of getting charged off i.e. being a defaulter.**

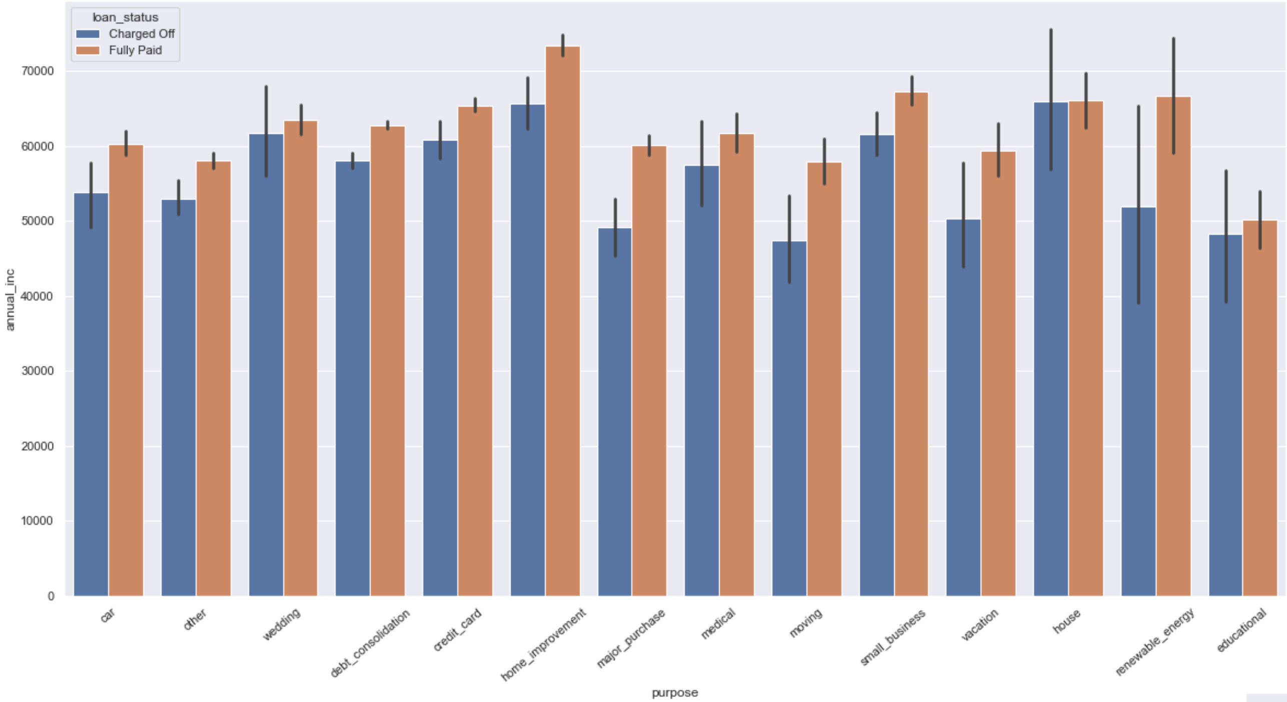






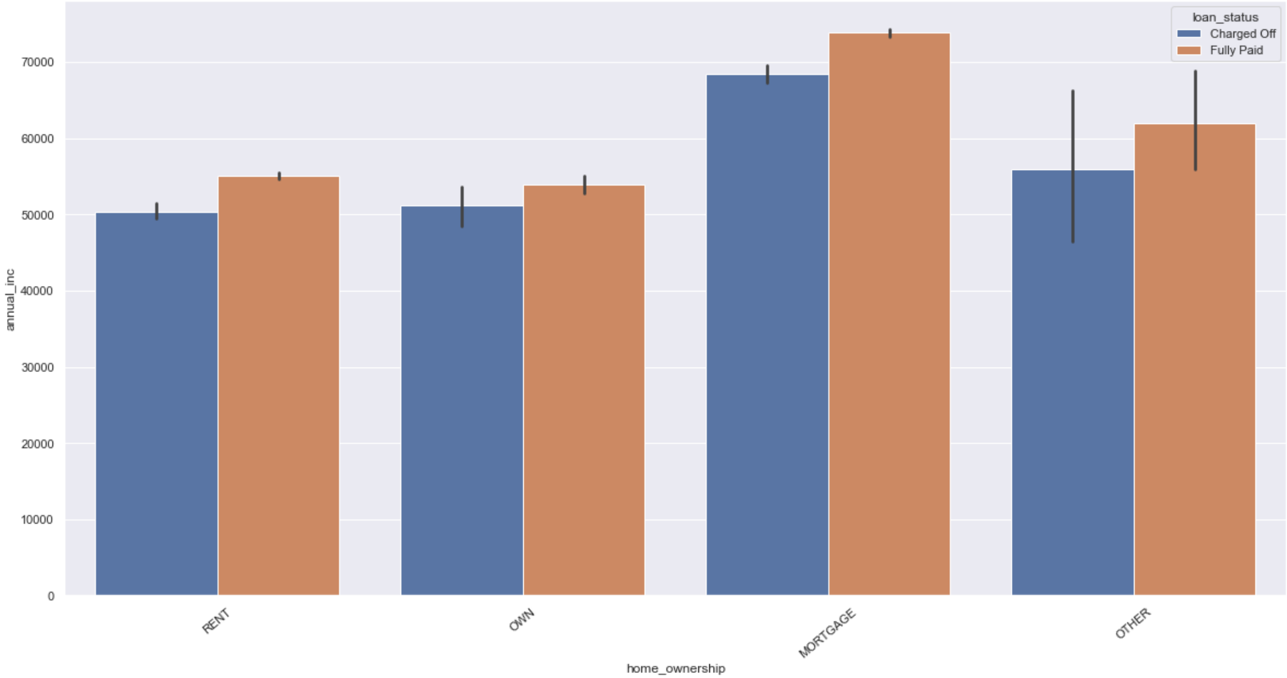
- Across all **'home ownership', 'loan amount', 'annual income' and 'installment'** categories default rate is very high when purpose of the loan is mainly **"Small Business"**

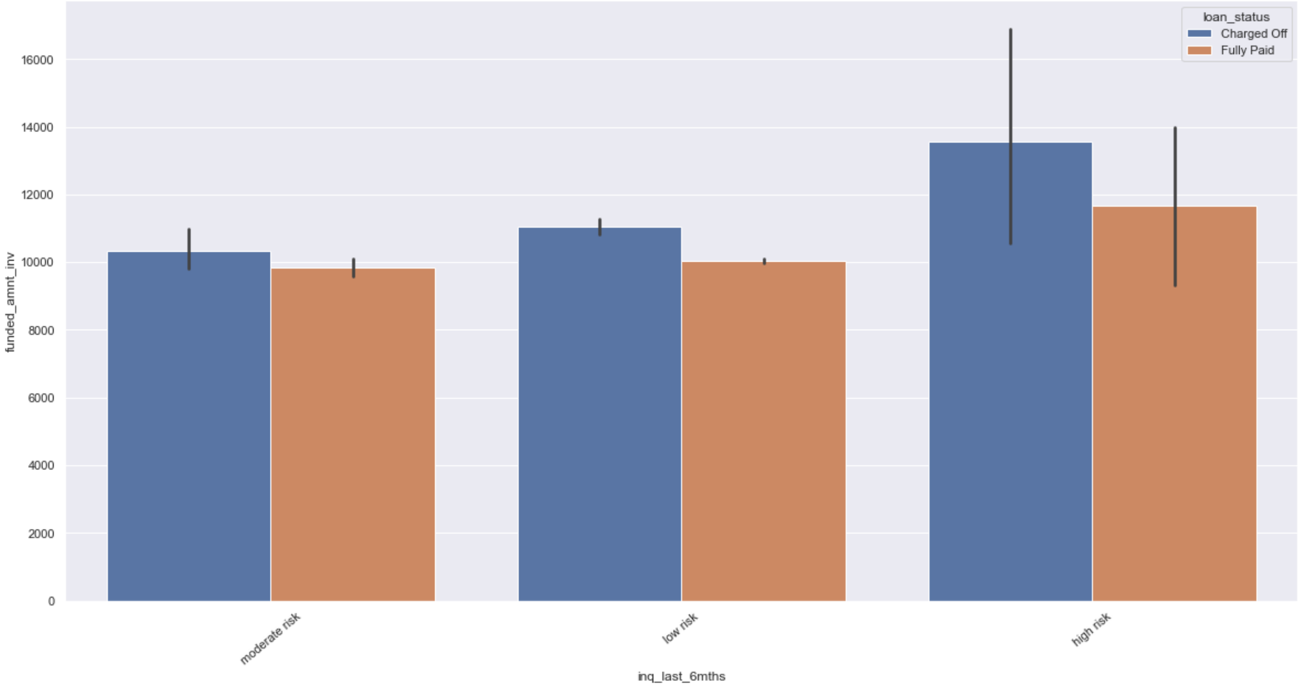
- Across **'interest rate'** categories, ***for very high interest rate default rate is higher for all the purposes and for low interest default rate is lower for all the purposes.***



- **Customers with high income are higher non-defaulter.**

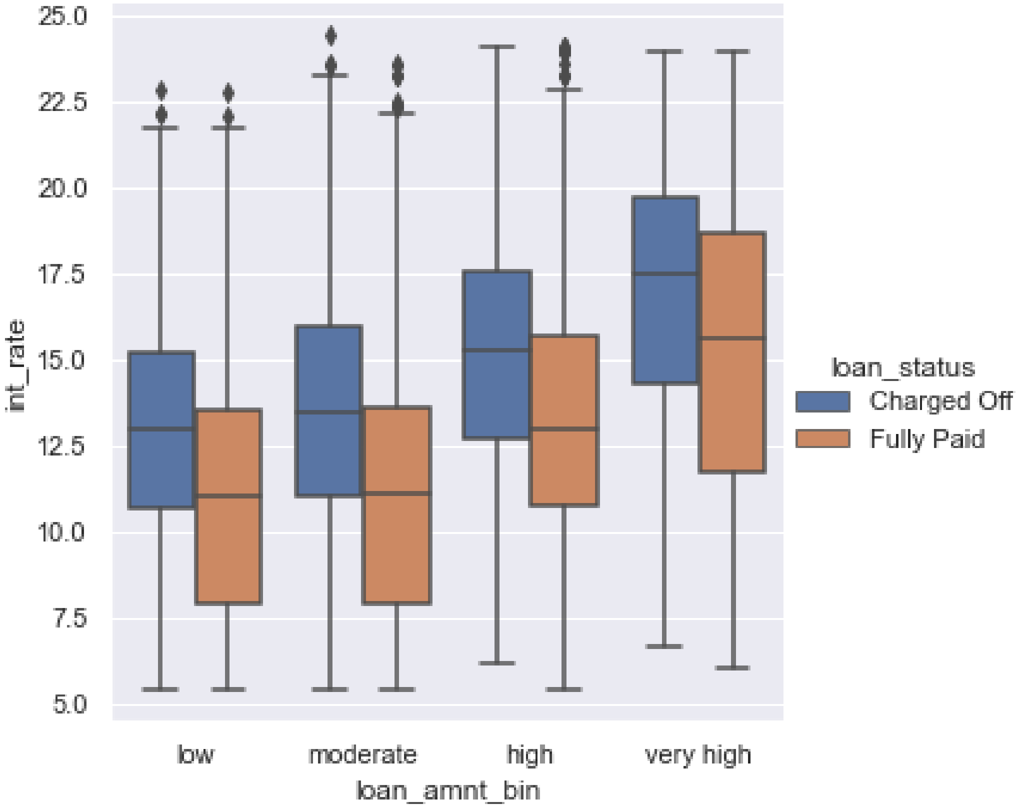
- **People with rented, mortgaged and other home ownership have slightly have less defaulting rate with the increase in annual income.**





- Customers with high risk(i.e. high number of inquiries in last 6 month) for high invested value have more default rate.

- Interest rate, Loan amount is directly proportional to the default rate. The more interest rate and loan amount the chances of charged off is very high.



Conclusions

The below conclusions are based on the above analysis based on Fully Paid and Charged Off loans and default rate. Conclusions as follows:

1.Chances of being a Defaulter or getting Charged off is very high when purpose of the loan is for small business.

2.Chances of being a Defaulter or getting Charged off increase with the interest rate of the loans. Higher the interest rate, high chances of being charged off and vice-versa.

3.Applicants who have taken a loan in the range 14k - 16k and taken loan for 60months term have high probability of getting defaulted.

4.Applicants who have taken a loan for small business and the loan amount is greater than 14k have high probability of getting defaulted.

5.Grade G loans have the highest interest rate above 20 %.

6.Applicant from the verified sourced and with the loan amount above 16k have high probability of getting defaulted.

7.Applicants with from than 8 years more employment experience and applied for loan amount above 14k have high probability of getting defaulted .

8.Applicant with home ownership as MORTGAGED and annual income between rage of 6-7k have high probability of getting defaulted.

Thank you