LENDING CLUB – CASE STUDY

SIDDHARTH SUJIR

ABHISHEK

BATCH: EPGML C55 JULY 2023

CASE STUDY – PROBLEM STATEMENT

- The purpose of this case study about a consumer finance company which specializes in lending loans to urban customers.
- The company wants to understand the risk associated to approving loan to customers.
- Not approving loan to customers who are likely to repay will result in loss to the business.
- Approving loan to customers who are not likely to pay will result in financial loss to the company.
- Objective is to identify the driving factors behind loan default using the given dataset.

UNDERSTANDING OF THE DATA

- The given dataset contains data from period of 2007 to 2011 with records of approved customers and their data of the loan status along with other variables.
- The **target variable** is the **loan_status** using which we can identify if a customer has repaid the loan or has defaulted.
- **Categorical Variables**
  - **Ordered Categorical Variables:** *emp_length, income_bucket, interest_bucket, issued_quarter, term*
  - **Unordered Categorical Variables:** *addr_state, purpose, grade, verification_status, home_ownership*
- **Continuous Variables**
  - *int_rate, annual_income, dti, open_acc, revol_bal, revol_util, total_acc, total_payment, total_due, loan_amount*

DATA CLEANING

- Based on the initial understanding of the data, we could find out that the shape of the data frame created is (39717, 111) which means it has 39717 records and 111 columns.
- As a next step, the columns with null values were identified and dropped from the data frame as they won't add any significance in our analysis.
- Also, the following columns having same value for all records were removed as they won't have much significance in our analysis.
  'pymnt_plan','url','initial_list_status','collections_12_mths_ex_med','acc_now_delinq','chargeoff_within_12_mths','delinq_amnt','tax_liens', 'application_type','policy_code','out_prncp','out_prncp_inv','pub_rec'

DATA CLEANING..

- The following columns had most of their values as 0. The 25th percentile and 75th percentile were 0 and hence it won't add much value for our analysis. These were also removed from the data frame.
  'total_rec_late_fee','recoveries','collection_recovery_fee','pub_rec_bankruptcies'
- After these column removal, the shape of the data frame is (39717, 39).
- Removed data with **loan_status** as 'Current' as they won't provide insights on the defaulters.
- The following columns were transformed to remove some characters from the string and converted to float and int to perform analysis on them as continuous numeric variables.

  'revol_util',' int_rate','term'.

DERIVED COLUMNS

- The following columns were derived from available columns to derive additional insights from the data.
  - **total_due** – By multiplying **installment** with **term**.
  - **updated_issue_dt** – derived by converting **issue_dt** from MMM-YY format to yyyy-mm-dd format.
  - **upd_earliest_cr_line -** derived by converting **earliest_cr_line** from MMM-YY format to yyyy-mm-dd format.
  - **income_bucket –** derived by placing **annual_inc** into buckets of ['VL','L','M','H','VH'].
  - **interest_bucket -** derived by placing **interest_rate** into buckets of ['VL','L','M','H','VH'].
  - **monthly_in –** derived by dividing **annual_inc** by 12.

DERIVED COLUMNS../ DATA FRAMES

- **no_years_of_credit** – Number of years of credit from the loan issue date to earliest credit line. Derived by subtracting **updated_issue_dt** from **upd_earliest_cr_line.** There were few numbers in negative. All the negative years were converted as 0 years of credit.
- **issued_month –** Extracted the month from **updated_issue_dt.**
- **issued_quarter –** Extracted quarter of the year from **updated_issue_dt**.

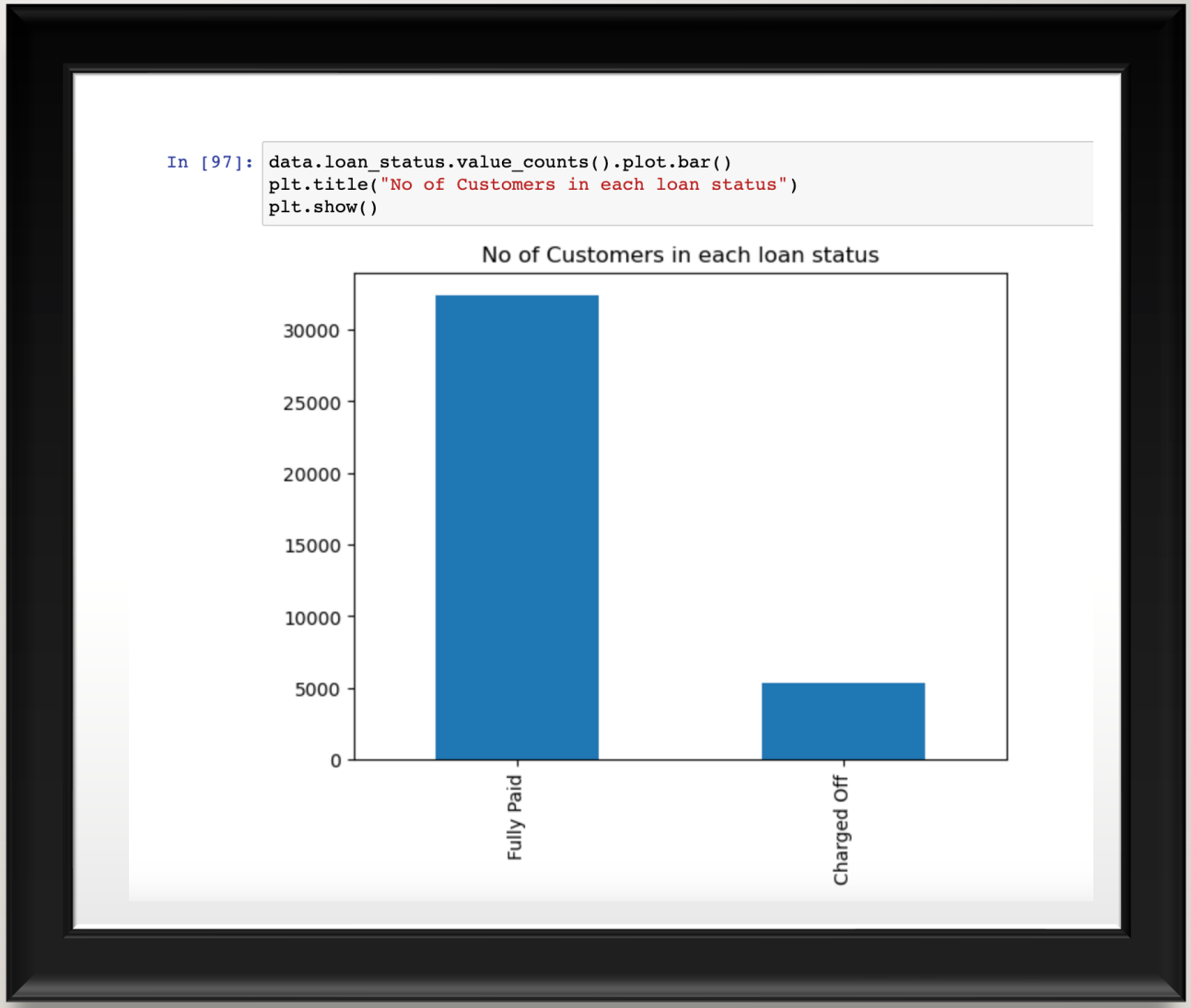
- **data_defaulters –** Dataframe containing defaulters.
- **data_fullypaid –** Dataframe containing customers who have fully paid their loans.

Shape of data frame after the derived columns - (37708, 48)

ANALYSIS

- The dataset consists of a greater number of **Fully Paid** customers compared to those who have **Charged Off**

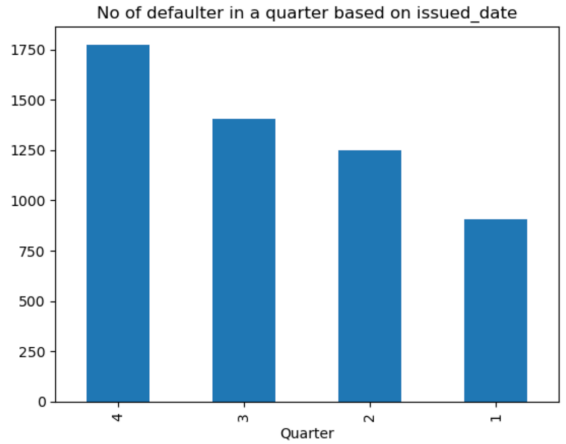
```
In [97]: data.loan_status.value_counts().plot.bar()
         plt.title("No of Customers in each loan status")
         plt.show()
```



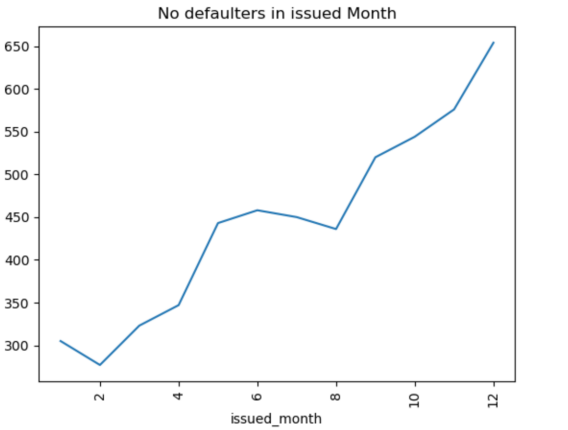
UNIVARIATE ANALYSIS

- Analysis on **issued_month** indicate that there are more number of defaulters whose loan was issue in the month of Oct, Nov, Dec or the 4th quarter of the year.

```
In [1090]: data_defaulters.issued_quarter.value_counts().plot.bar()
           plt.title('No of defaulter in a quarter based on issued_date')
           plt.xlabel('Quarter')
           plt.show()
```



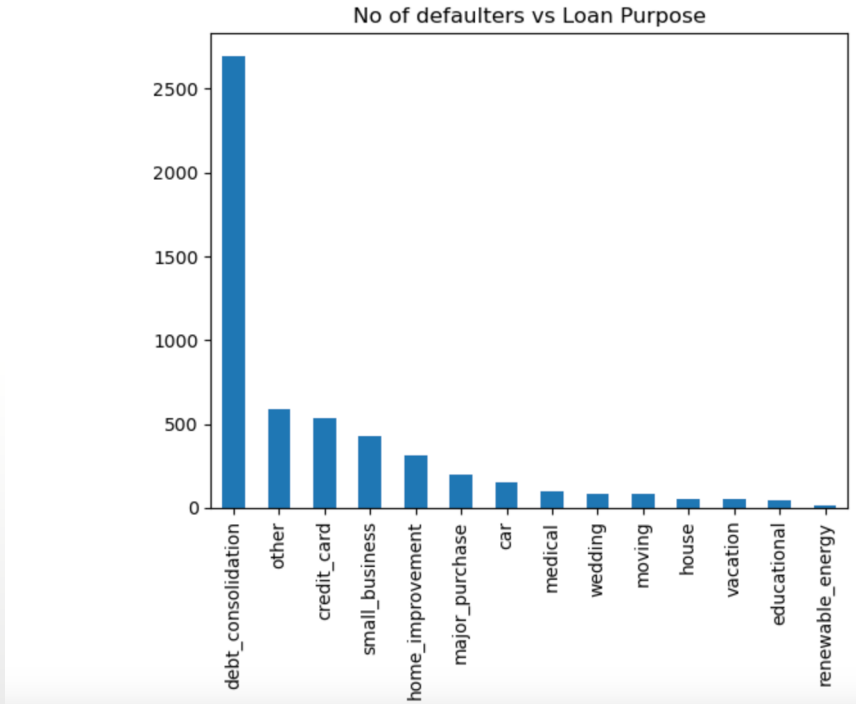

```
In [1087]: # Line chart also depicts the same.
           data_defaulters.groupby(by=['issued_month']).member_id.count().plot()
           plt.xticks(rotation=90)
           plt.title('No defaulters in issued Month')
           plt.show()
```



UNIVARIATE ANALYSIS..

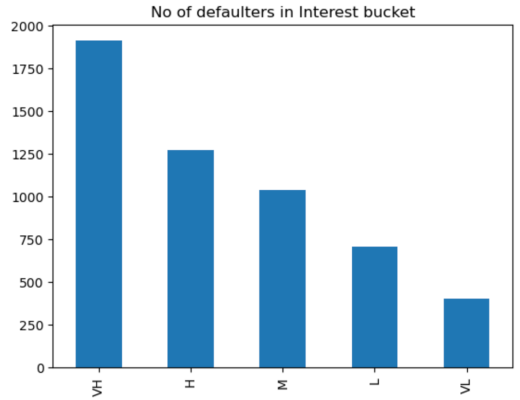
- Analysis on the variable '**purpose**' against the no of defaulters indicate people who purchase loan for **debt_consolidation** are likely to default more.

In [96]:
```
# no of defaulters against purpose of loan
data_defaulters.purpose.value_counts().plot.bar()
plt.title("No of defaulters vs Loan Purpose")
plt.show()
```



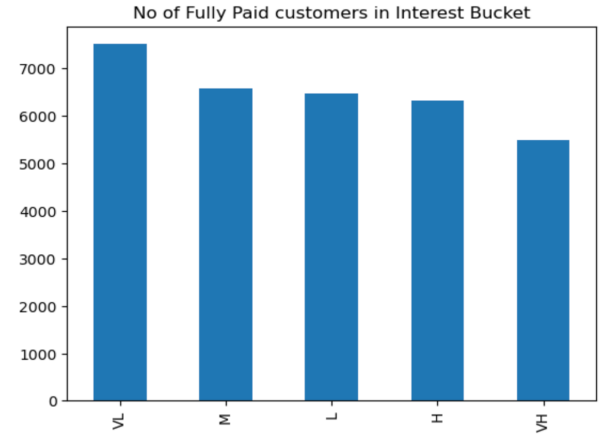
```
In [68]: # Plotting a bar chart of interest rate bucket on the defaulters data
         # People in VH interest rate bucket, default more compared to people in lower interest bucket.
         data_defaulters.interest_bucket.value_counts().plot.bar()
         plt.title("No of defaulters in Interest bucket")
```

```
Out[68]: Text(0.5, 1.0, 'No of defaulters in Interest bucket')
```




```
In [88]: # Plotting a bar char on interest bucket of fully paid customers
         # Indicates high no of people in VL intrest rate bucket pay off the loans
         data_fullypaid.interest_bucket.value_counts().plot.bar()
         plt.title("No of Fully Paid customers in Interest Bucket")
```

```
Out[88]: Text(0.5, 1.0, 'No of Fully Paid customers in Interest Bucket')
```



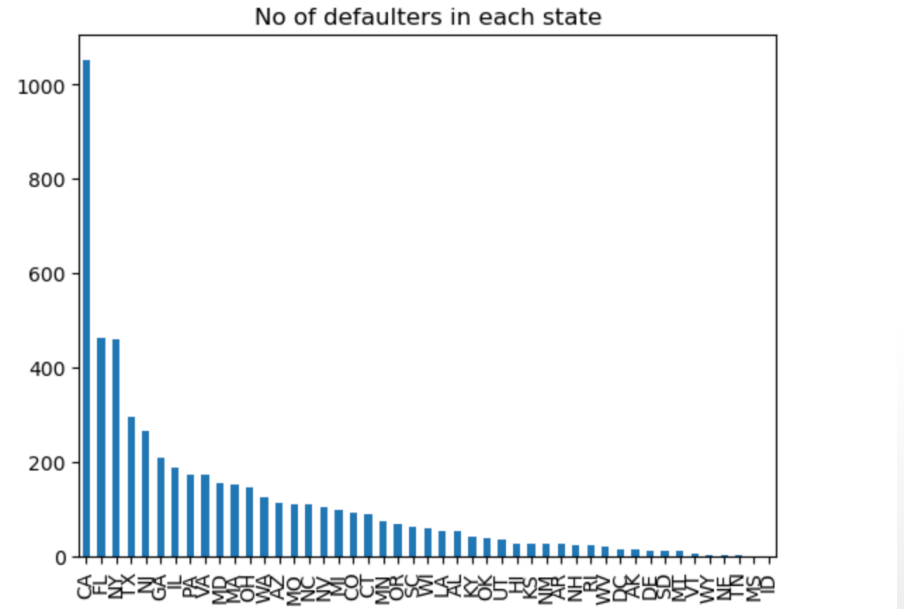
UNIVARIATE ANALYSIS

- Analysis on **interest_bucket** variable indicates that people in VH bucket tend to default more. Likewise, more number of people in VL bucket fully pay their loans.

UNIVARIATE ANALYSIS..

- Analysis on **addr_state** over defaulters' dataset shows that there are more number of defaulters from the state of **California** as compared to the other states.

```
In [95]: # Plotting a bar chart on addr_state, the chart shows that people in Califor
         data_defaulters.addr_state.value_counts().plot.bar()
         plt.title("No of defaulters in each state")
         plt.show()
```



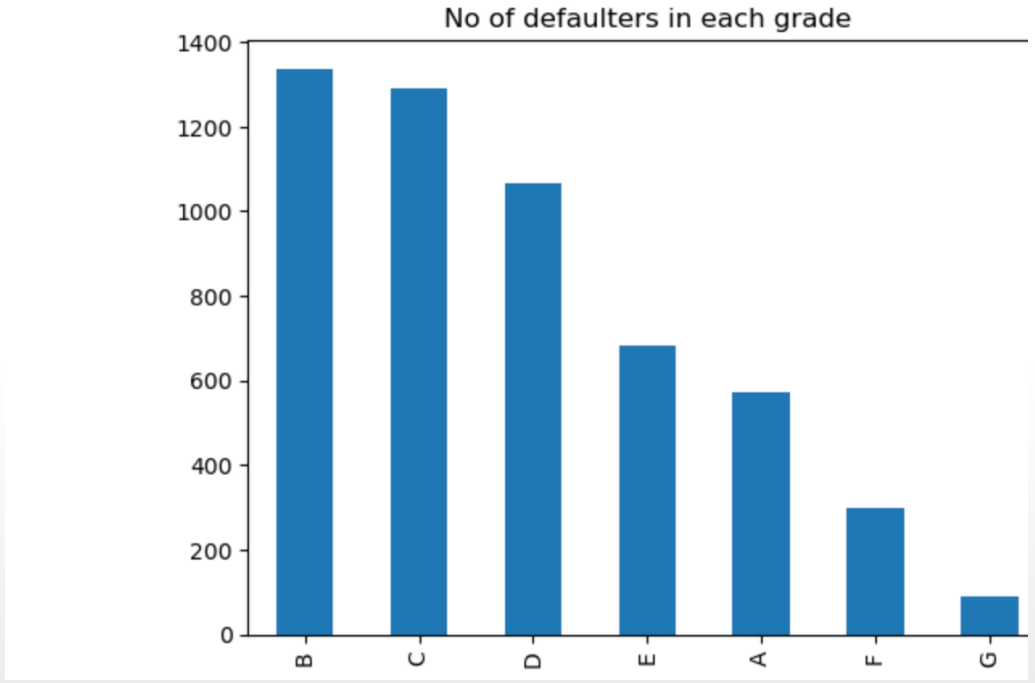
SEGMENTED UNIVARIATE ANALYSIS..

- Analysis on **home_ownership** variable over **charged_off** customers reveal that the people with homes on RENT and MORTAGE are more likely to default as compared to people with OWN and OTHER types.

UNIVARIATE ANALYSIS..

- Analysis on **grade** variable on defaulters' data indicate that there are more number of defaulters in B, C, D grade.

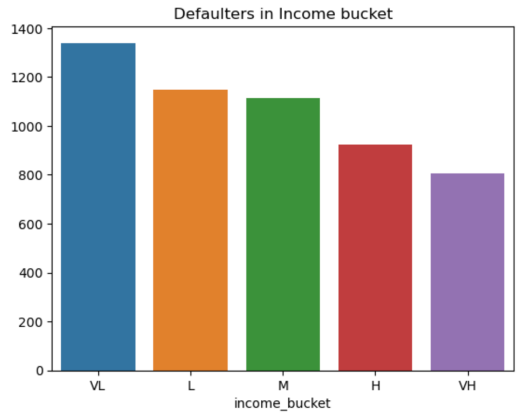
```
In [99]: data_defaulters.grade.value_counts().plot.bar()
         plt.title('No of defaulters in each grade')
         plt.show()
```



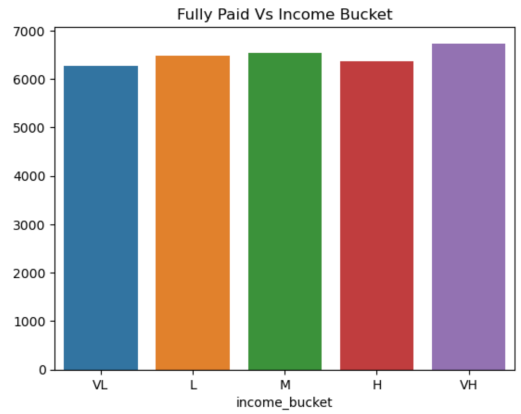
SEGMENTED UNIVARIATE ANALYSIS..

- Analysis on the **income_bucket** on defaulters' bucket indicate that more number of people in **VL, L** bucket default.
- More number of people in **VH** bucket pay of their loan.

```
In [1047]: ivot on the annual salary bucket shows that people in VL and L bucket have high no of defaulters
           ult=pd.pivot_table(data=data_defaulters,index='loan_status',columns='income_bucket',values='member_id',aggfunc='count')

In [1083]: sns.barplot(result)
           plt.title('Defaulters in Income bucket', fontsize=12)

Out[1083]: Text(0.5, 1.0, 'Defaulters in Income bucket')
```




```
In [1084]: # Pivot on the annual salary bucket shows that people in VH and H bucket are more likely to pay off the loan
           result2=pd.pivot_table(data=data_fullypaid,index='loan_status',columns='income_bucket',values='member_id',aggfunc='cou
           sns.barplot(result2)
           plt.title('Fully Paid Vs Income Bucket')

Out[1084]: Text(0.5, 1.0, 'Fully Paid Vs Income Bucket')
```



SEGMENTED UNIVARIATE ANALYSIS..

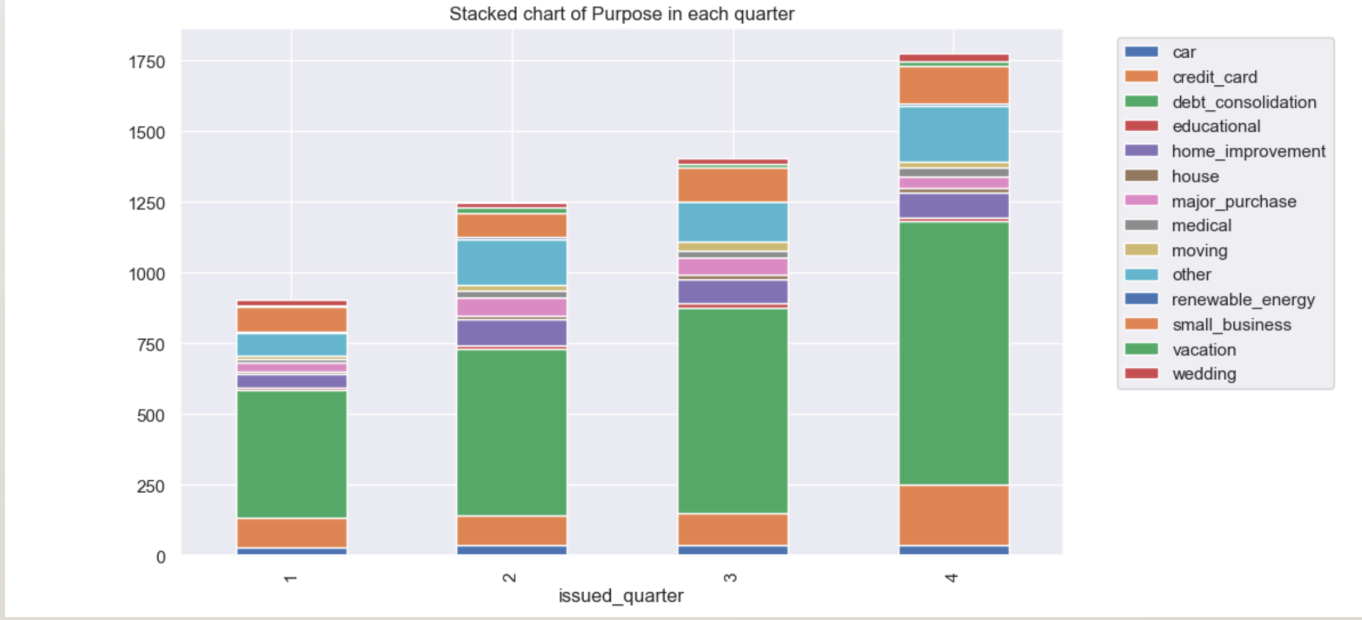
- Analysis on the **emp_length** variable indicate that there are more number of defaulters with 10+ years of experience.

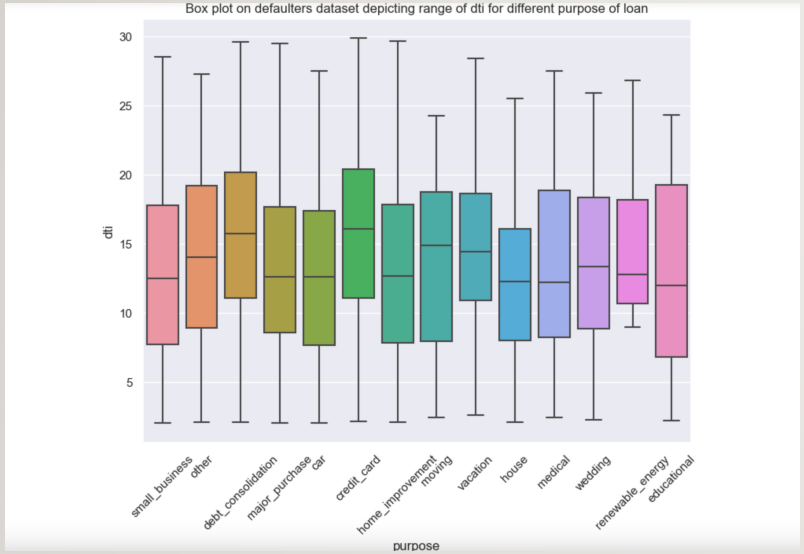
```
In [1085]: # Plotting a bar chart to get count of defaulters based on no of years
           # 10 + years of experience default more.
           data_defaulters.emp_length.value_counts().plot.bar()
           plt.xticks(rotation=90)
           plt.title('No of Defaulters in emp_length')

Out[1085]: Text(0.5, 1.0, 'No of Defaulters in emp_length')
```

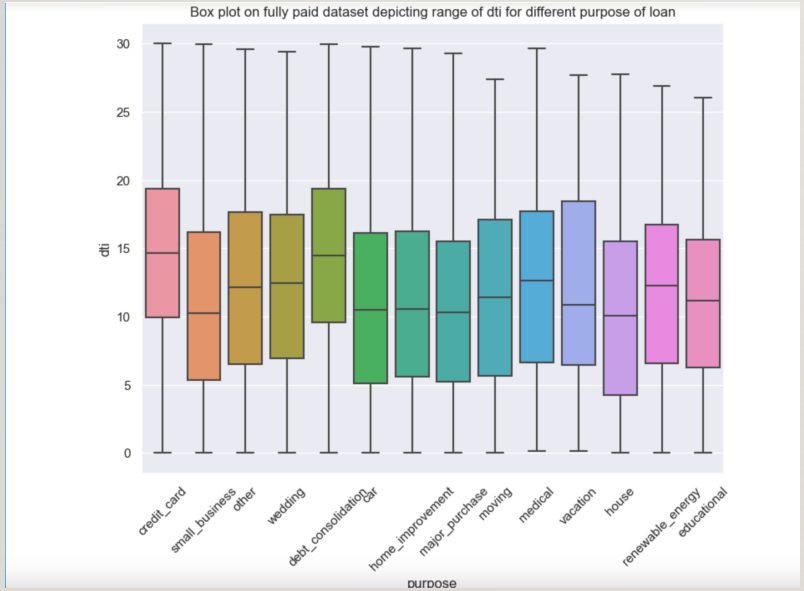
BIVARIATE ANALYSIS

- Analysis of **issued_quarter** vs **purpose** over defaulters' dataset indicate that there are more defaulters in the 4th quarter of the years whose purpose of loan is *debt_consolidation*.

```
In [101]: # Stacked bar plot on issued quarter vs purpose shows people who get loan for debt_consolidation in 4th quarter are lik
monthly_pivot.plot(kind='bar', stacked='True', figsize=[10,6])
plt.title('Stacked chart of Purpose in each quarter')
plt.legend(bbox_to_anchor=(1.05, 1.0), loc='upper left')
plt.show()
```



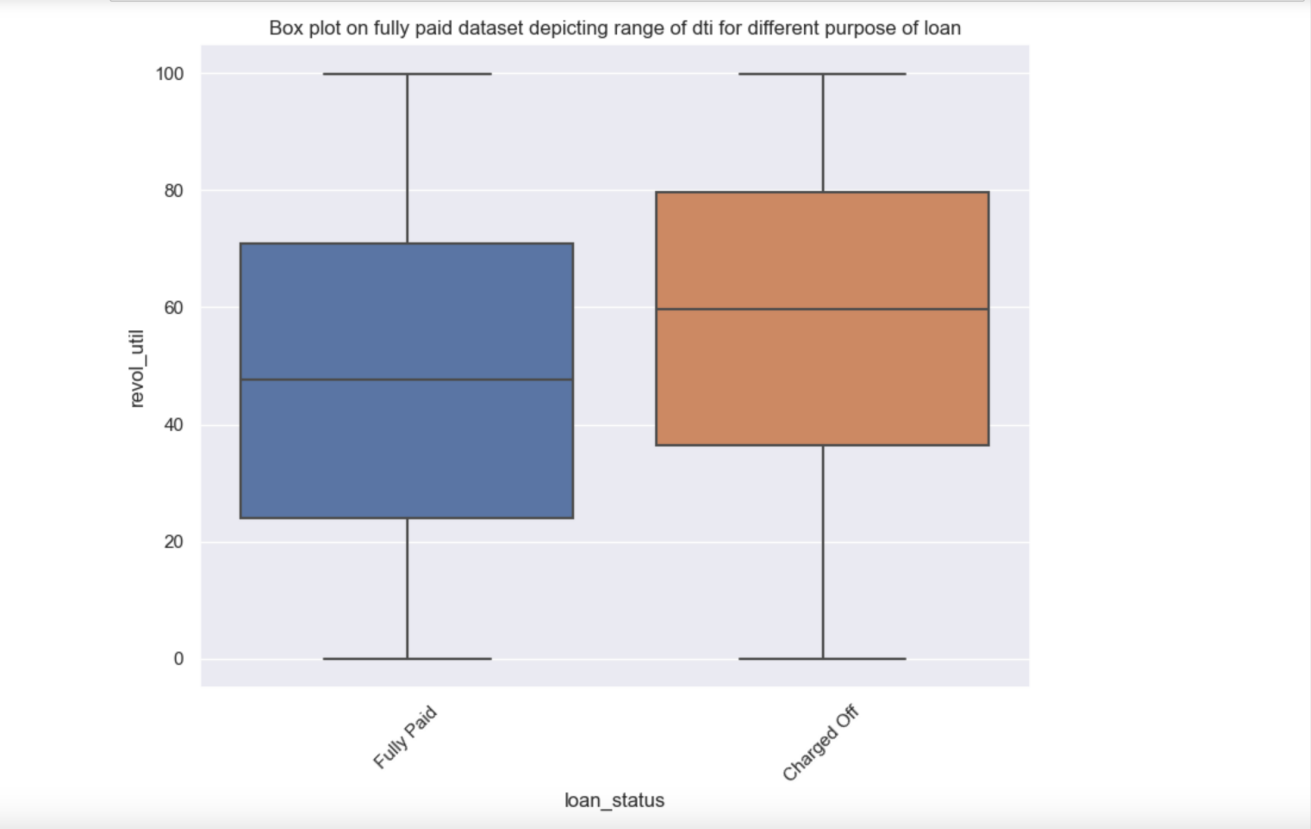




BIVARIATE ANALYSIS..

- Analysis of **dti** and **purpose** over defaulter's dataset show that the people who applied for loan for *debt_consolidation* have higher median dti compared to those who have fully paid

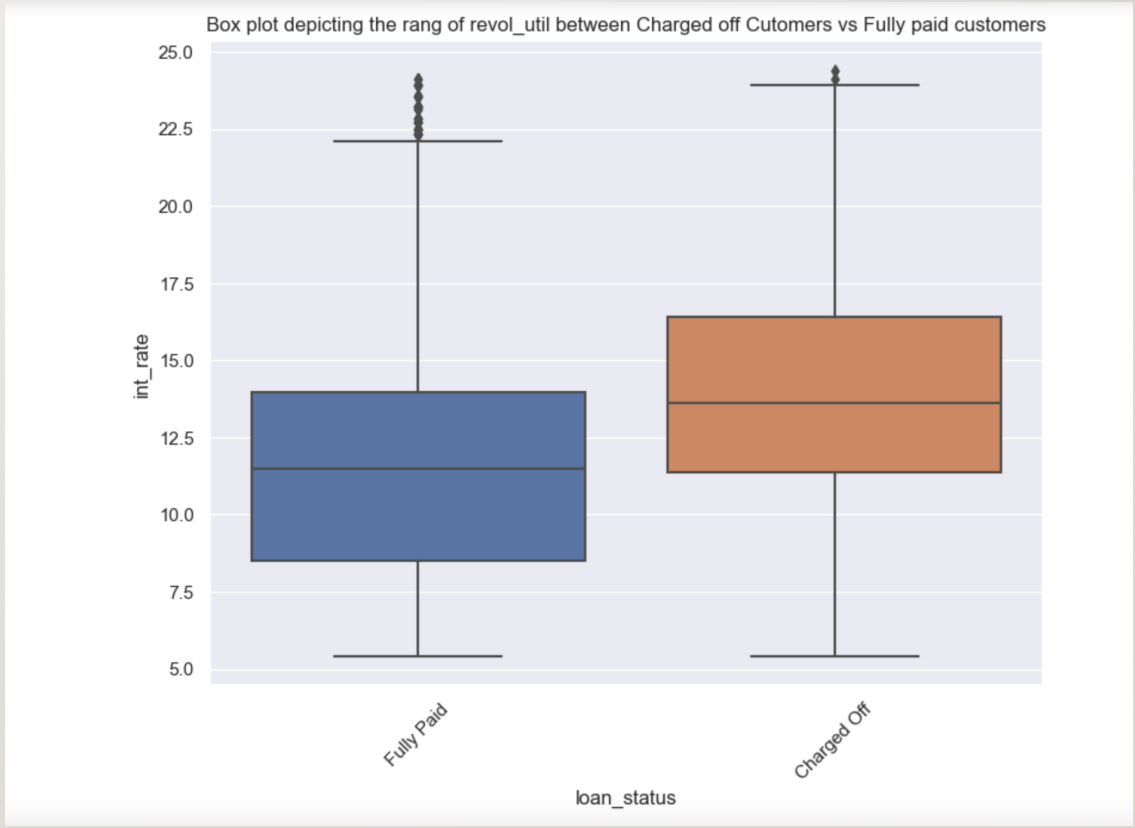
BIVARIATE ANALYSIS..



- Analysis on **revol_util** against **loan_status** indicate that the charged off customers have higher median as compared to fully paid customers. The value in 25th and 75th percentile are also considerably higher.

BIVARIATE ANALYSIS..

- Analysis of **int_rate** against **loan_status** indicate people with higher interest rate default more compared to the ones with lower interest rate. The median of int_rate for people who have defaulter is higher compared to the fully paid customers.



BIVARIATE ANALYSIS..

- Analysis on **inq_last_6mths,** defaulters have higher median and 75th percentile compared to those who have fully paid off.

CONCLUSION FROM UNIVARIATE AND SEGMENTED ANALYSIS

- From the analysis the following can be concluded:
  - The given dataset contains a larger number of *Fully Paid* customers compared to *Charged Off* customers *[Slide 8].*
  - More number of loan applicants are from California state. More number of defaulters are from California state *[Slide 12].*
  - People who have been issued loan in the 4th quarter of the year have defaulted more *[Slide 9].*
  - People who have taken loan for **debt consolidation** have defaulted more compared to other category of loan application purpose *[Slide 10].*
  - People who have **VH** interest tend to default more. Contrary to it, people who have VL interest fully pay off their loan *[Slide 11].*
  - People in **Rented home** and **home on Mortgage,** have a larger number of defaulters *[Slide 13].*
  - People in VL incomer bucket tend to have more defaulters. People in VH income bucket have more number of fully paid *[Slide 15].*
  - People with 10+ years of **employment length** default more compared to other levels of experience *[Slide 16].*

CONCLUSION FROM BIVARIATE ANALYSIS

- Customers with loan purpose as ***debt_consolidation,*** and 4th quarter of the years have defaulted more *[Slide 17]*.
- Customers who apply for loan for ***debt_consolidation*** defaulted have higher median *debt to income ratio (dti)* *[slide 18]*.
- Customers who have defaulted have higher **revol_util** median compared to fully paid customers. A higher **revol_util** means they already have a higher credit balance to pay off *[Slide 19]*.
- Customers who have defaulted on loan have higher median **int_rate** *slide[20]*.
- Customers who have defaulted have higher number of inquires in the last 6 months. The median and 75th percentile value of **inq_last_6mths** variable is higher for those who have defaulted on loans *[Slide 21]*.