LENDING CLUB CASE STUDY

- SIDDHARTH JEETENDRA CHOUDHARI

Studying the dataset and drawing initial observations

- Initial Observations
  - Total Rows = 39716
  - Total Columns = 111
  - There are many columns which contains
    - All Null Values
    - Partial Null Values
    - No Null Values
  - There are many columns wherein the data type is not matching the Column header as per the header definition.

Data Cleaning based on initial observations

- Dropping all the columns which are completely null.

Data Cleaning based on initial observations

- Taking action on columns with partial Null values

Column Name	Comments	Action Taken
emp_length	Based on the values available it would be good decision to replace the null values of this coloumn with string “ < 1 year” so that these entries will be clubbed with the existing entries where data is already populated as “< 1 year”	NaN values are replaced with string “< 1 year”
emp_title	This data maybe used to analyse what category of emp titles are likely to default the loans. However there is a huge variety of different data / different strings in this coloum. So replacing the NaN values with common identified would be preferred	NaN values are replaced with string “Data Not Available”
desc	A lot of entries in this coloum are null and this coloum will have less impact on our analysis. However at higher level some NLP could be performed on this coloum to check for any pattern in description provided by the applicant. However at this stage this is not being considered. So removing this coloum from our dataset	Column Dropped

Data Cleaning based on initial observations

- Taking action on columns with partial Null values

Column Name	Comments	Action Taken
title	“title” coloumn has different values based on the user inputs. More appropriate column for analysis would be the “purpose” column which has more standardised values, probably due to dropdown options in the application form. So this coloum can be removed	Column Dropped
months_since_last_delinqent	Has around 25000 NaN entires	Column Dropped
mths_since_last_record	Has around 36000 NaN entires	Column Dropped
revol_util	This coloum has around 50 NaN entries, but this coloum is important for our analysis to study the spending on credit habits of the applicant. Since a very small portion of the values are NaN, we can replace these with average of revol_util of remaining entries.	1. Derieved metric coloumn is created to include only the numeric part of this column. 2. Mean and median is checked for Non-NaN values and those are found to be close-by indicating there could be no outliers. 3. Mean is computed of the Non-NaN values. 4. NaN values are replaced with the mean value.

Data Cleaning based on initial observations

- Taking action on columns with partial Null values

Column Name	Comments	Action Taken
last_payment_d	This date column will have no impact on our analysis	Column Dropped
next_payment_d	This date column will have no impact on our analysis	Column Dropped
collections_12_mths_ex_med	Has only 2 entries a. 0 b. NaN	Column Dropped
chargeoff_within_12_mths	Has only 2 entries a. 0 b. NaN	Column Dropped
tax_liens	Has only 2 entries a. 0 b. NaN	Column Dropped
pub_rec_bankruptcies	This coloum seems important for our analysis.	NaN replaced with string “Data Not Available”

Data Cleaning based on initial observations

- Checking if any columns are having common values throughout. If there is no variation in the data of any columns , we cannot use those columns to find any pattern since there is no change in data.
- After analysis it is observed that following columns contain common data throughout.
  - pymnt_plan
  - initial_list_status
  - policy_code
  - application_type
  - acc_now_delinq
  - delinq_amnt

  DROPPING THESE COLUMNS

Data Cleaning based on initial observations

- Checking if there any columns which will have not impact on our analysis, based on the column header.
    - URL

  DROPPING THIS COLUMN

loan_status

- Based on the problem statement we need to study patterns for applicants who have earlier fully paid the loan and applicants who have earlier defaulted the loans.
- If we scan the loan_status column we find 3 entries.
    - Fully Paid
    - Charged Off
    - Current
- Since we need to analyse only the past data, the ongoing data “Category Current” is of no use to us in this analysis. So removing the rows where the value under “loan_status” is “Current”.

Problem Statement

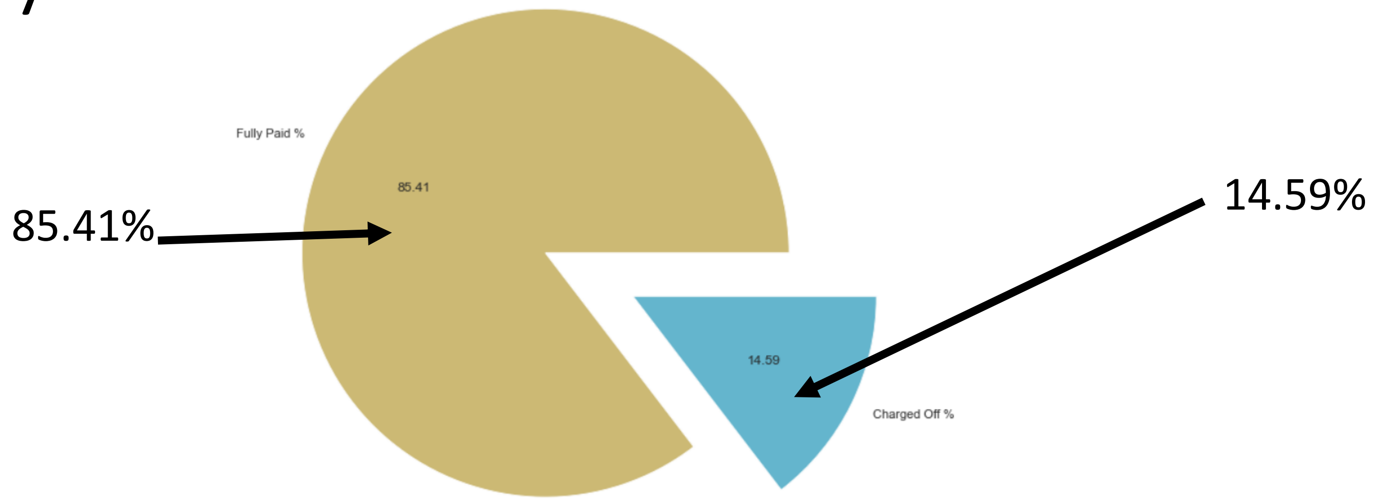
- To identify and present minimum 5 drivers which are an indicator for higher probability of loan being defaulted.
- These drivers can be used to analyse an incoming loan application and to deduce if the parameters indicate high risk to grant loan to the particular application.

Analysis

- Following columns are mainly analysed and respective deductions are presented in further slides.

Column Name	Type of Data	Type of Analysis
term	Ordered Categorical	Univariate
pub_rec_bankruptcies	Ordered Categorical	Univariate
purpose	Unordered Categorical	Univariate
state	Unordered Categorical	Univariate
revol_util	Ordered Categorical data converted to Numerical	Univariate
annual_inc + instalments	Numerical	Bivariate

Basis of Analysis



Above pie chart shows “% Fully paid” entries and “% of Charged off” entries from our cleaned data frame.

- We will assume that for different categories when we plot such pie chart the distribution of full paid and charged off loans shall be approximately equal to the figures as per this pie chart.
- So for any category 85% people are assumed to have fully paid the loan and 15% people’s loans have been charged off.
- Our rubric for further analysis against any category:-
  - If the fully paid % is more than 85% i.e. charged off % is less than 15% the scenario can be described as **LESS RISKY**
  - If the fully paid % is less than 85% i.e. charged off % is more than 15% then the scenario can be described as **MORE RISKY**

Analysis :- “term”

- 2 unique values under this column
  - 36 months
  - 60 months.
- Fig 1 shows that out of total entries
  - 24.5% people have taken loan for 60 months
  - 75.5% people have taken loan for 36 months.

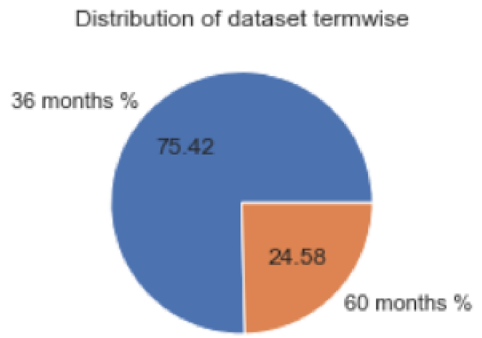


Fig - 1

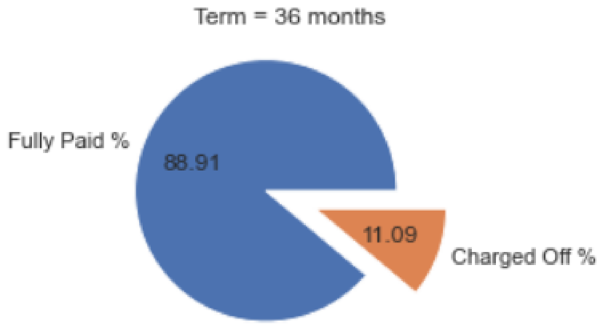


Fig - 2

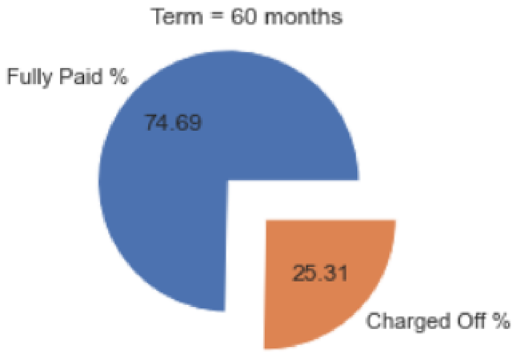


Fig - 3

Analysis :- “term”

- Fig 2 shows that out of the people who have taken loan for 36 months
  - 89% people have fully paid the loan
  - 11% people have defaulted the loan

- Fig 3 shows that out of the people who have taken loan for 60 months
  - 75% people have fully paid the loan
  - 25% people have defaulted the loan

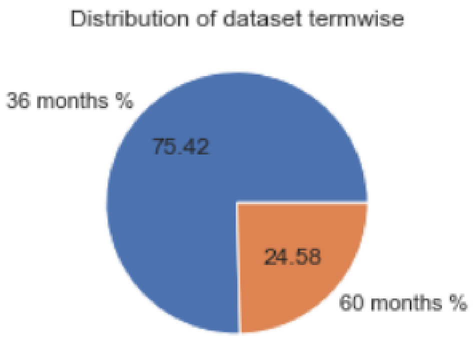


Fig - 1

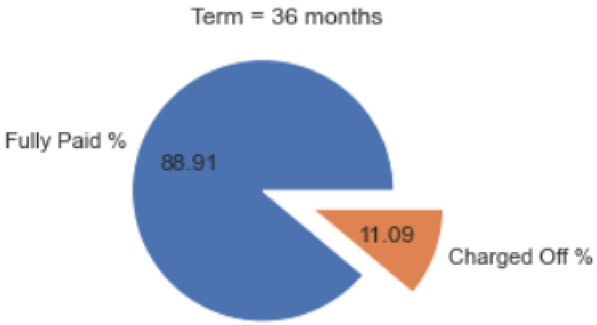


Fig - 2

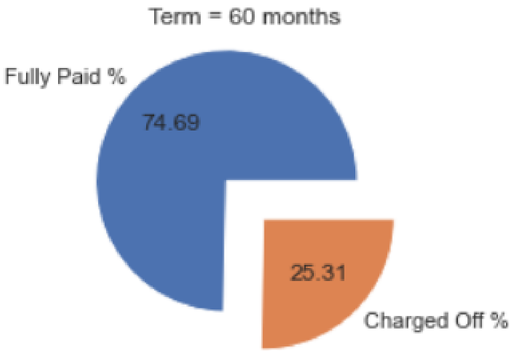


Fig - 3

Deductions from Analysis :- “term”

- Charts show that the **probability of a loan and hence the risk getting defaulted increases with increase in tenure**.
- An increase in loan tenure could mean more profit for the lender, however increase in tenure also increases the risk of loan getting defaulted, because the individual to whom loan is granted ,has higher probability to be affected by the economic dynamics (for e.g. loss of job due to recession). So, the earlier the individual repays the loan , the less is the risk of loan getting defaulted.

Analysis :- “pub_rec_bankruptcies”

- 3 unique values under this column
  - 0.0
  - 1.0
  - 2.0
- Fig 1 shows that out of total entries
  - 95.6% people have 0 public bankruptcies records
  - 4.32 % people have 1 public bankruptcies records
  - 0.01% people have 2 public bankruptcies records

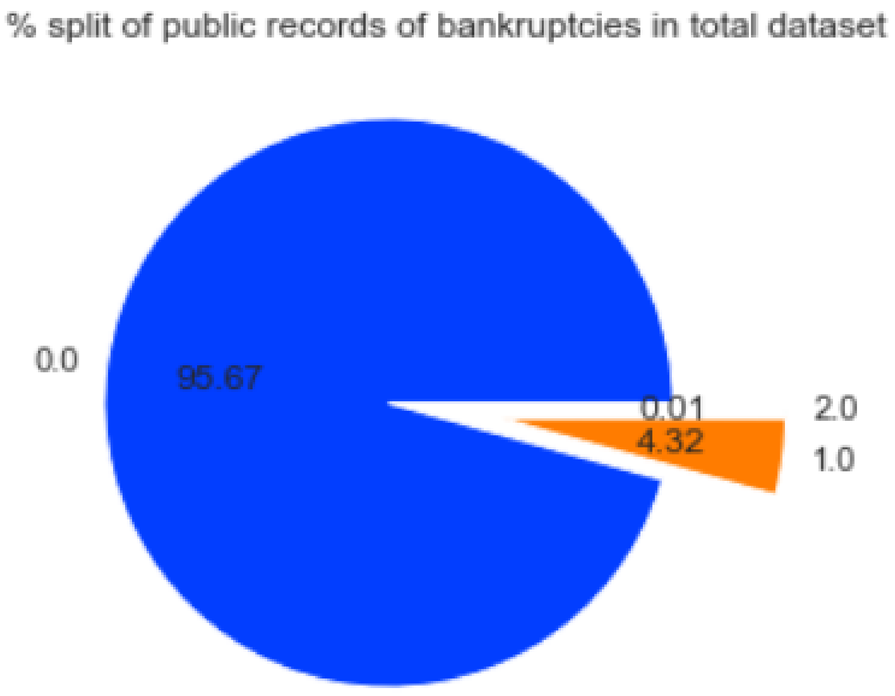


Fig - 1

Analysis :- “pub_rec_bankruptcies”

- Fig 2 shows that out of the people having 0 records for public bankruptcies
  - 85.8% people have fully paid the loan
  - 14.1% people have defaulted

- Fig 3 shows that out of the people having 1 record for public bankruptcies
  - 77.6% people have fully paid the loan
  - 22.3% people have defaulted

- Fig 4 shows that out of the people having 2 records for public bankruptcies
  - 60% people have fully paid the loan
  - 40% people have charged off

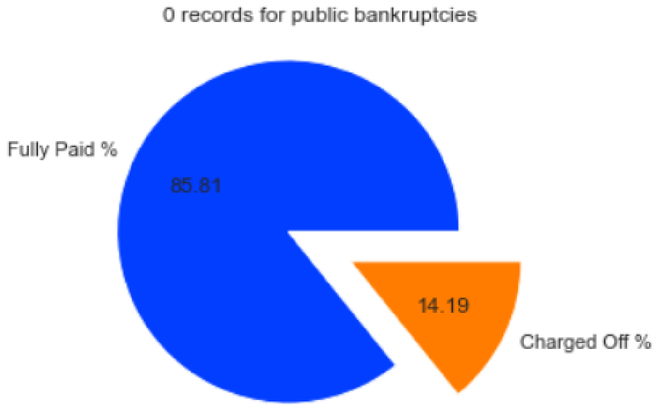


Fig - 2

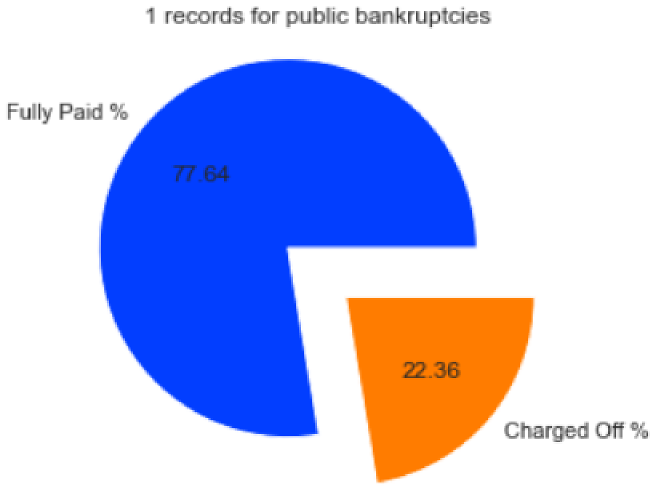


Fig - 3

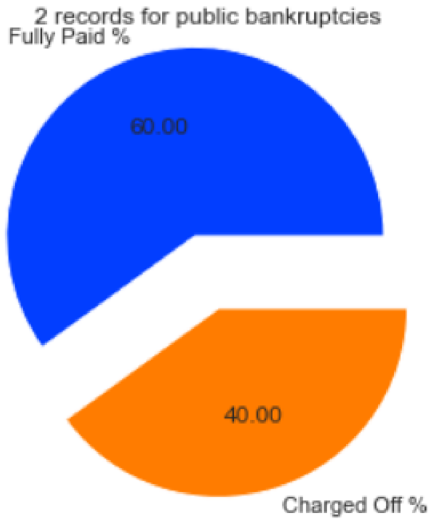


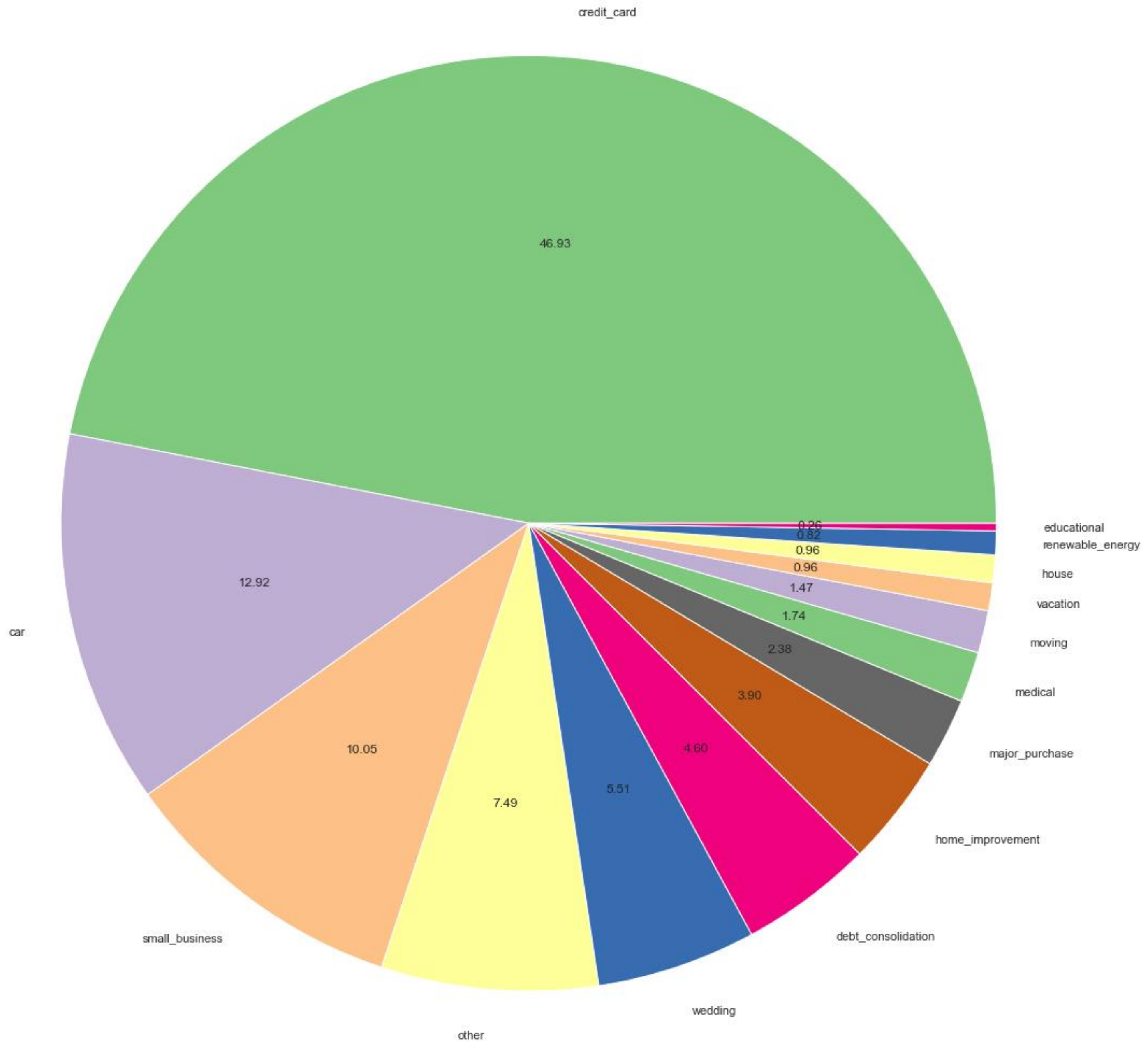
Fig - 4

Deductions from Analysis :- “pub_rec_bankruptcies”

- In the main dataset we can see than there are very few entries of details of loans with 1 or 2 public records of bankruptcies. This may be due to the fact that loans for applicants having 1 or 2 public bankruptcies records could have been rejected and that is the reason they are not appearing in our database.
- However out of the entries that we have, we can see that people having 1 or 2 public records of bankruptcies are more likely to default on their loans and so we can deduce that **the risk factor increases even if the applicant has 1 public record of bankruptcy**.

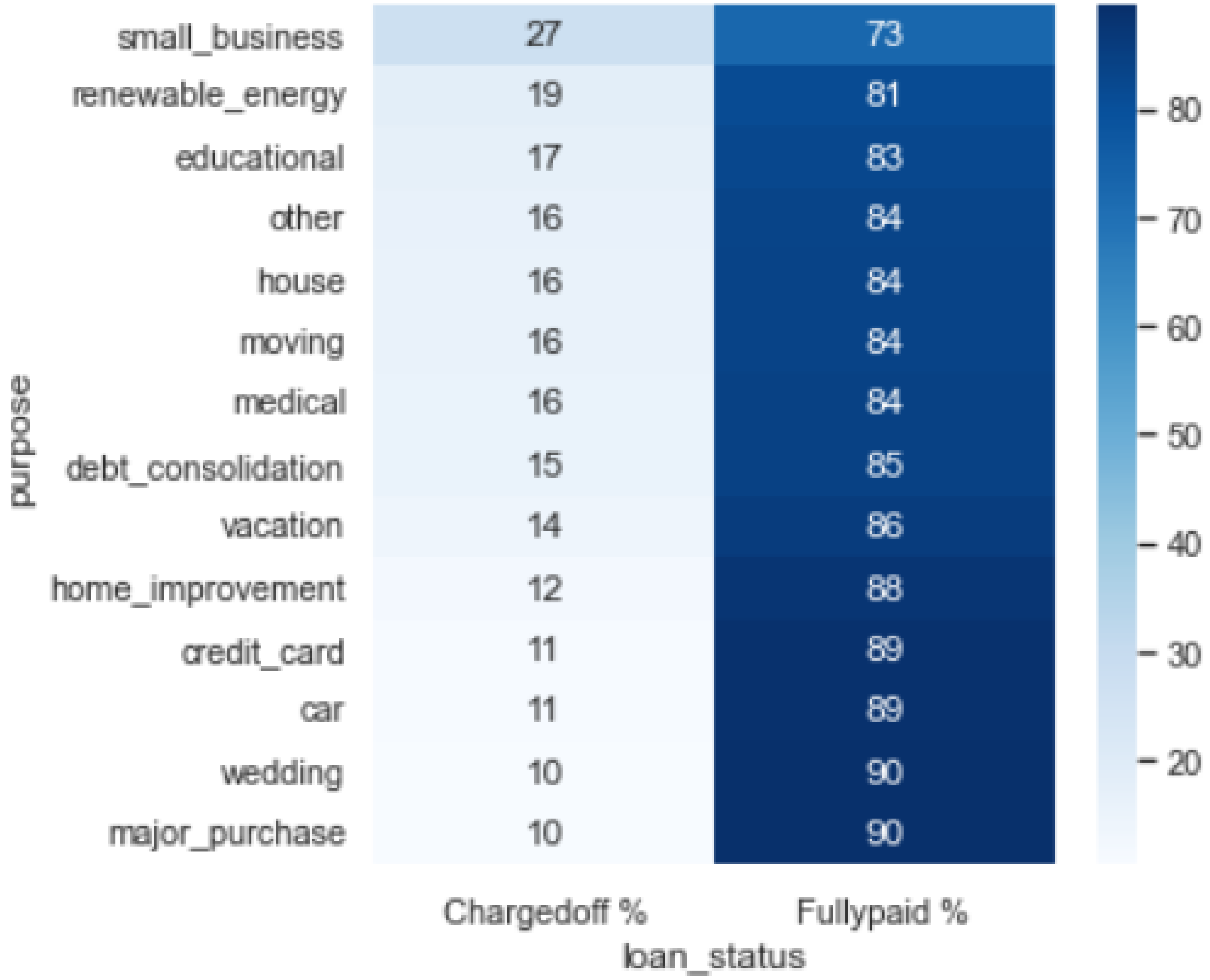
Analysis :- “purpose”

- Besides pie chart which shows ‘%’ wise distribution of loans granted w.r.t the purpose for which the loan is taken.



Analysis :- “purpose”

- Against each purpose we have computed what is the percentage of people who have defaulted the loans and what is the percentage of people who have fully paid the loans.
- This is shown in the besides heat map plot.



Deductions from Analysis :- “purpose”

- As we have already defined our basis
  - If the fully paid % is more than 85% i.e. charged off % is less than 15% the scenario can be described as **LESS RISKY**
  - If the fully paid % is less than 85% i.e. charged off % is more than 15% then the scenario can be described as **MORE RISKY**

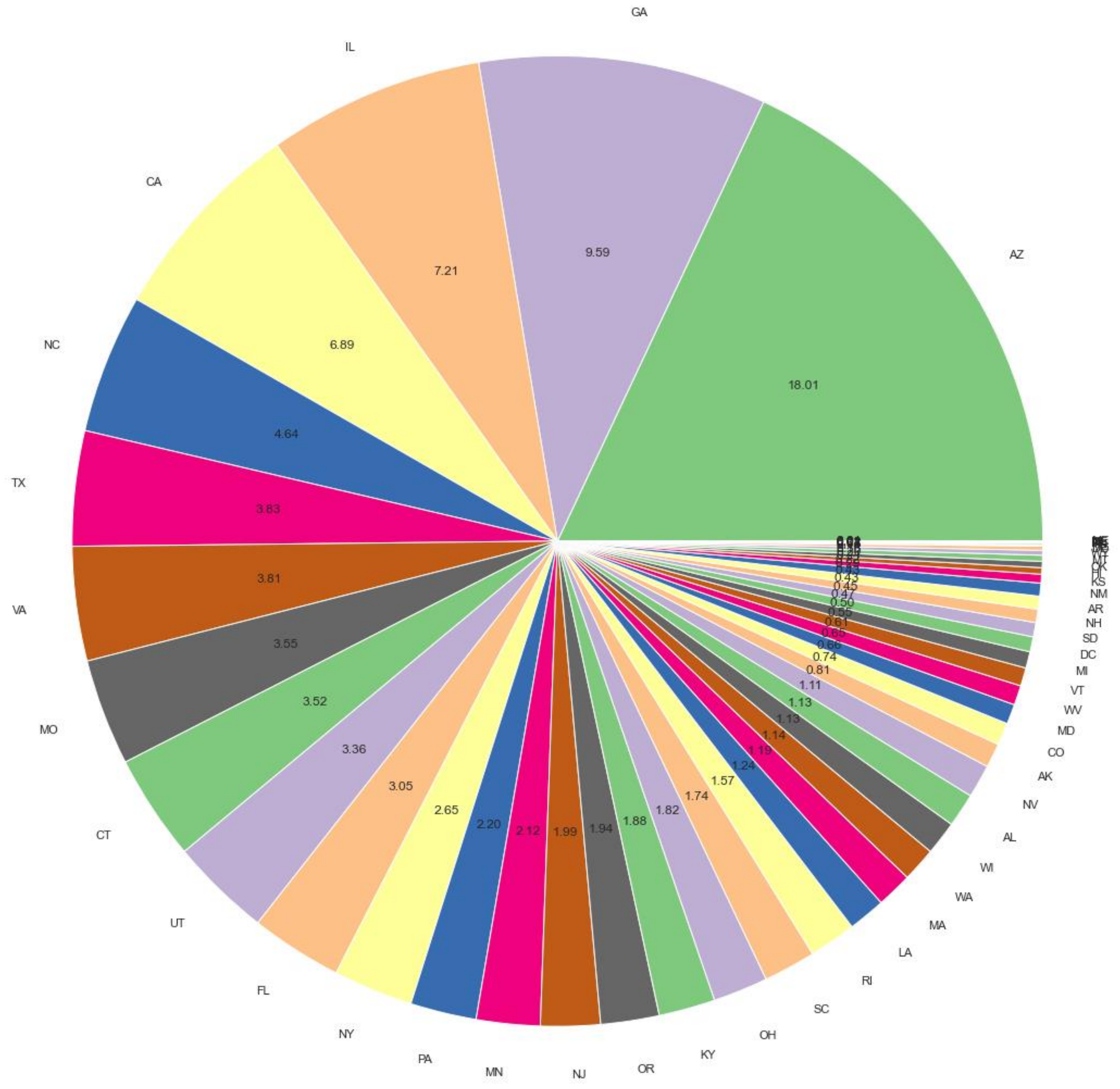
- Considering only 2 categories as above we can state that if any application claims for a **loan for below purposes** :-
  - Medical
  - Moving
  - House
  - Other
  - Educational
  - Renewable energy
  - Small business

  The application could be categorised to have a High Risk of defaulting the loan.

- **Additionally we can see that category small business carries the maximum risk out of all categories**. So lending club can formulate special rules while granting loans which is intended for forming / running a small business.

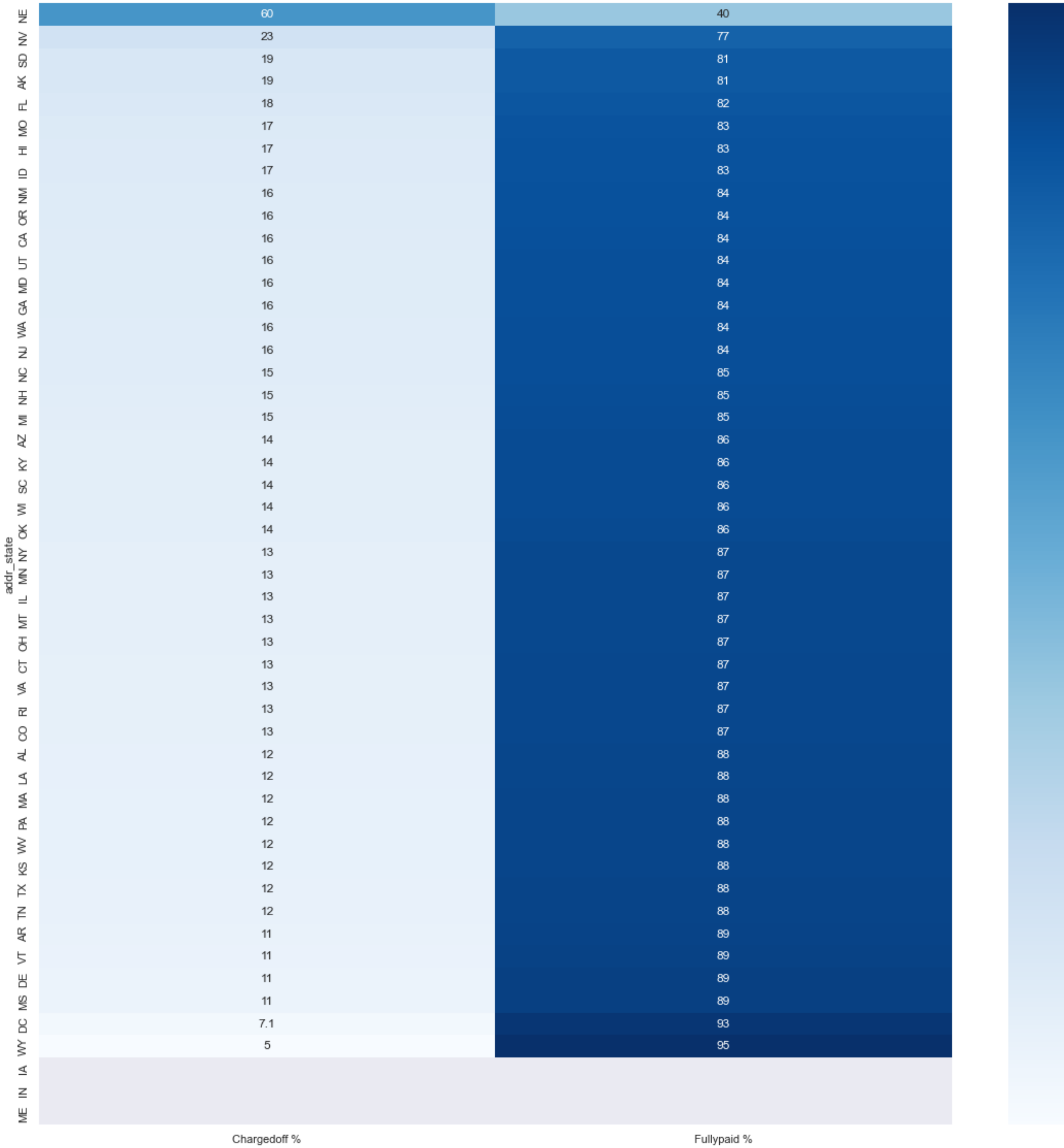
Analysis :- “addr_state”

- Besides pie chart which shows ‘%’ wise distribution of loans granted w.r.t the address state of the person who has applied for the loan.



Analysis :- “addr_state”

- Against each state we have computed what is the percentage of people who have defaulted the loans and what is the percentage of people who have fully paid the loans.
- This is shown in the besides heat map plot.



Deductions from Analysis :- “addr_state”

- As we have already defined our basis
  - If the fully paid % is more than 85% i.e. charged off % is less than 15% the scenario can be described as **LESS RISKY**
  - If the fully paid % is less than 85% i.e. charged off % is more than 15% then the scenario can be described as **MORE RISKY**

- Considering only 2 categories as above we can deduce that **loan applications from residents of below states can be categorised as Risky.**
  - NE
  - NV
  - SD
  - AK
  - FL
  - MO
  - HI
  - ID
  - NM
  - OR
  - CA
  - UT
  - MD
  - GA
  - WA
  - NJ

- This could be due to demographic , economic conditions of the particular state which can lead to conditions wherin people cannot repay the loans.
- For e.g. if people from any state are mostly dependant on farming as main source of income, but over the period due to changing climatic conditions there has been irregular rainfalls resulting in loss of crops. So it is more likely that any loan application from this particular state would be defaulted because of these conditions.

Analysis :- “revol_util_percent”

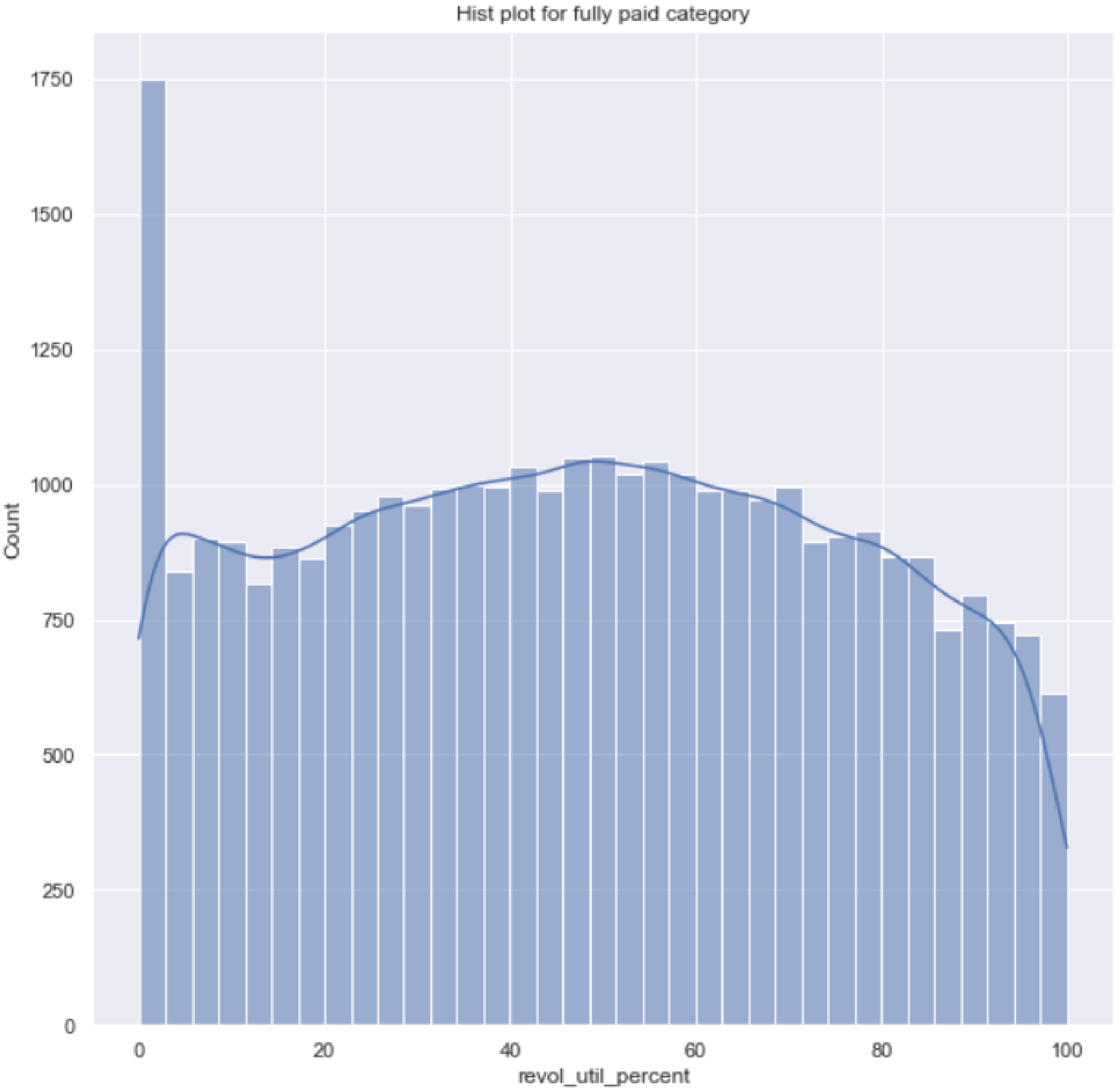


Fig - 1

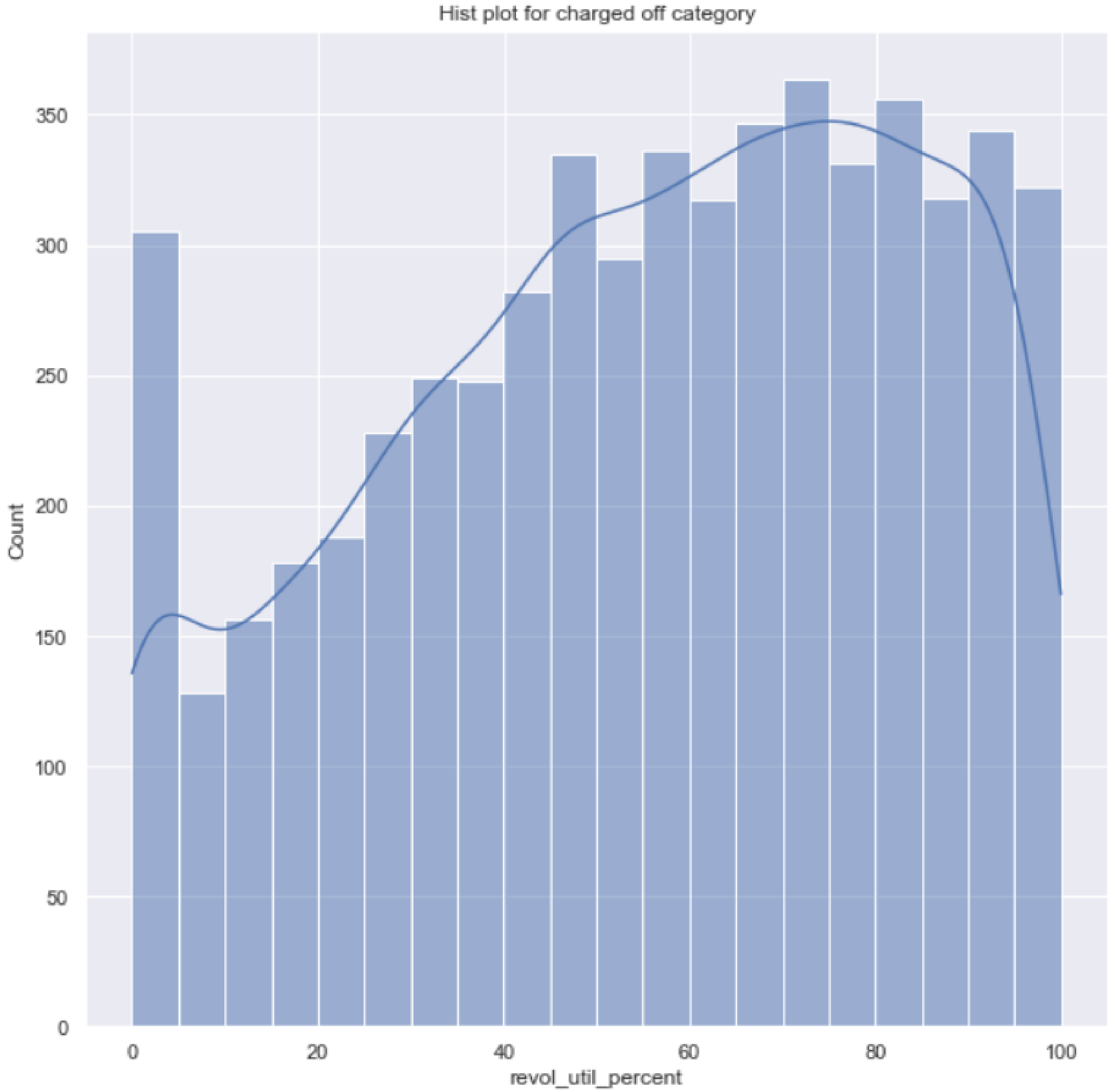


Fig - 2

Analysis :- “revol_util_percent”

- Our data set is distributed in 2 separate datasets
  - 1st contains data of only “Fully Paid” Loan Category
  - 2nd contains data of only “Charged Off” Loan Category
- A histplot is plotted for both categories
  - Fig 1 shows histplot of “revol_util_percent” for Fully Paid Category
  - Fig 2 shows histplot of “revol_util_percent” for Charged Off Category.

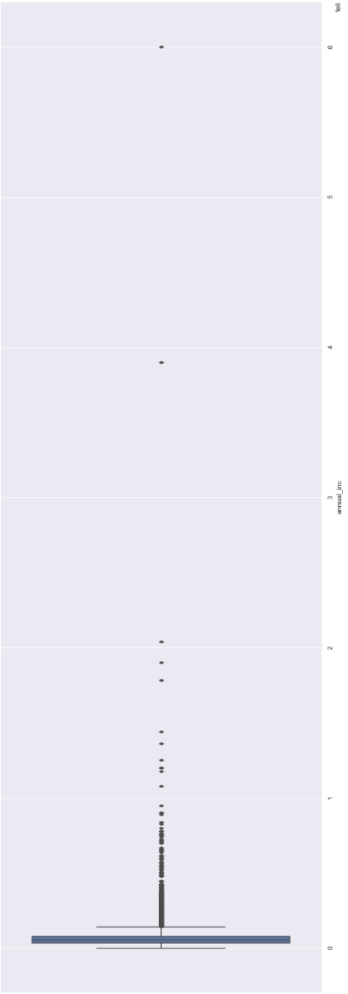
Deductions from Analysis :- “revol_util_percent”

- On analysing the histplots we observe that
  - Histplot of “Fully Charged” category shows a downward trend with increase in revol_util_percent
  - Histplot of “Charged Off” category shows a upward trend with increase in revol_util_percent
- Revol_util_percent indicates the spending habits of the individual wherein high percentage indicates that the person relies on credit for his / her expenses and is more prone to be in a situation wherein his / her credit expense is more than his / her’s returning capability.
- This could be the reason which shows an upward trend in histogram of Fig 2
- Since both graphs show an upward trend till 50% mean and median of revol_util_percent for whole dataset is also approximately near to 50%.
- **We can deduce that if the revol_util for any applicant is more than 50%, that application should be categorises as Risky.**

Analysis – “Annual_inc” and “Monthly Installment”

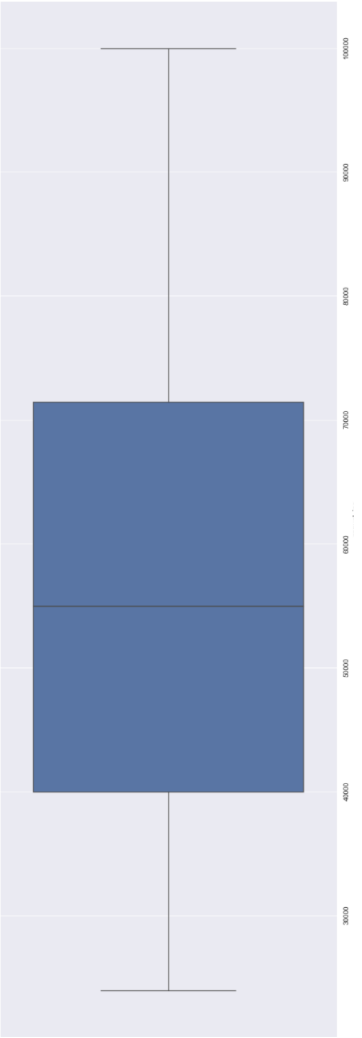
- To check if the ratio of monthly instalment to monthly salary has any pattern w.r.t defaulting of the loans.
- On scanning the annual_inc coloum we find that there are many outliers which will hamper our analysis and these need to be removed.
- On studying the plots it is decided to keep only the rows where the annual income is between 5 percentile and 85 percentile.

Analysis – “Annual_inc” and “Monthly Installment”



Mean =68777
Median = 58868

BEFORE REMOVING OUTLIERS
FROM ANNUAL_INC

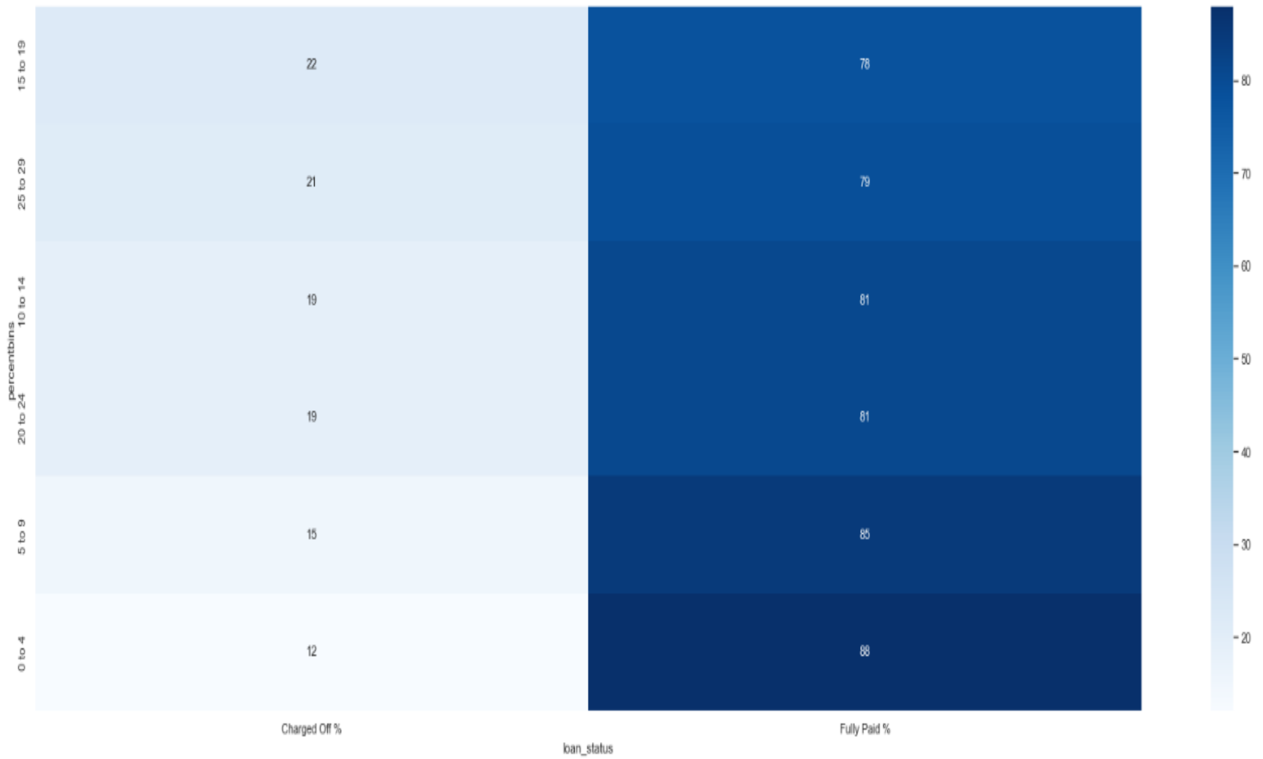


Mean =56822
Median = 55996

AFTER REMOVING OUTLIERS
FROM ANNUAL_INC

Analysis – “Annual_inc” and “Monthly Installment”

- The annual income is converted to monthly income
- Another column is derieved which shows how much percentage is the instalment w.r.t the monthly income.
- This coloum is coverted to bins of 5 units (0 to4 , 5 to 9 etc.)
- For each bin we have compute how many people have fully paid the loan and how many have defaulted.
- The besides heatmap is the representative of the same.



Deductions from Analysis – “Annual_inc” and “Monthly Installment”

- Based on the heatmap it is observed that **as the ratio of instalment to monthly income increases beyond 10% the risk of loan default increases.** Lending company can do this check and can take decisions accordingly like increasing the tenure to lower the instalment. However as seen in our earlier observation, higher tenure also has increased risk. So lending club has to take a call w.r.t which risk to adapt and which to mitigate.
- Another observation is that within a group of 0% to 9% there are more instances of default within the bin 5% to 9% than the bin 0% to 4%. Same observation is seen consistently for the group of 10 to 19% where instances of default within bin 15% to 19% is more than the bin 10% to 14%. Similar observation is for the bin 20% to 29% where instances of default within bin 25% to 29% is more than the bin 20% to 24%.
- Reason for this observation is not known, however if we trust the data then lending club can use this observation to plan the instalment – tenure – interest rate accordingly so that they can keep the ratio of instalment to monthly income within the first half of any bin and relatively reduce the risk.

THANK YOU

- SIDDHARTH JEETENDRA CHOUDHARI