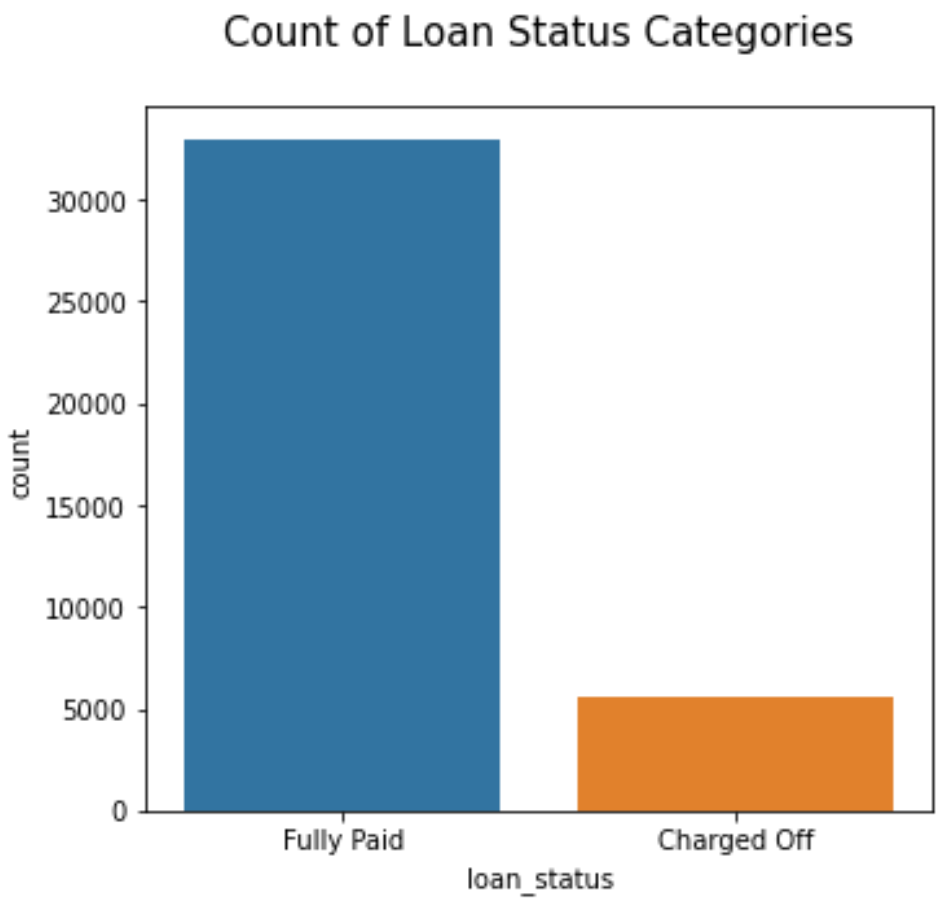
Use Case

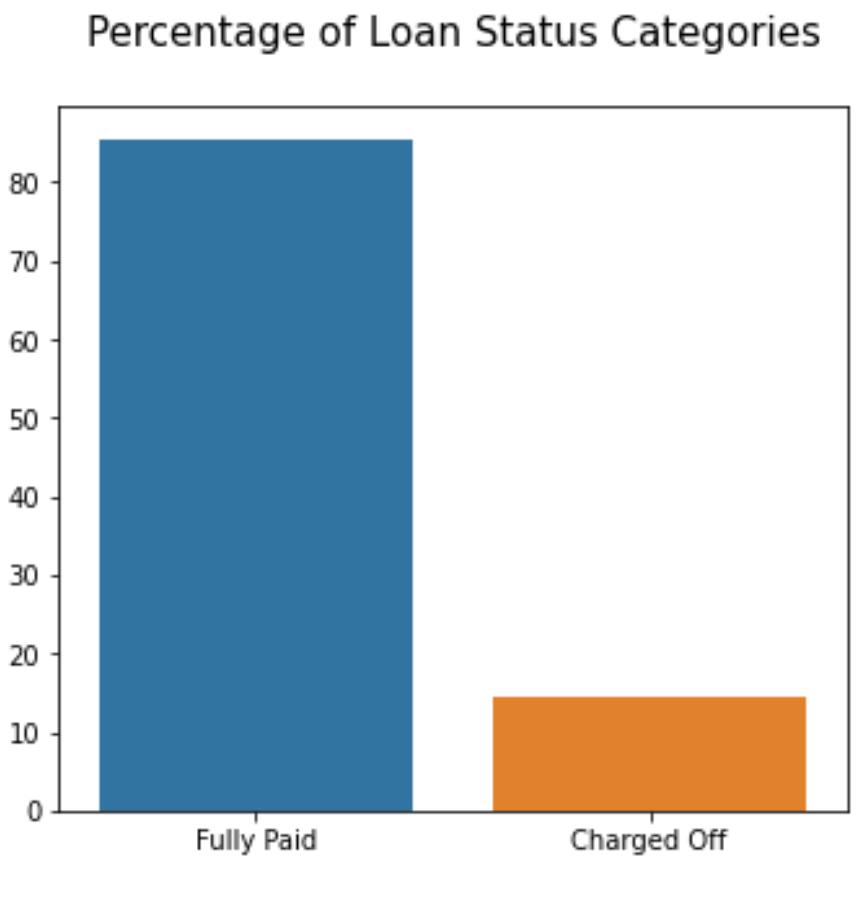
- Our client is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.
- The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.
- The company wants to utilize this knowledge for its portfolio and risk assessment

Data Understanding

1) The loan dataset had 39717 rows and 111 columns.
2) Their are 74 columns of float type, 13 of integer type and 24 of textual/object type.
3) Loan status is the depending or target column.
4) Loan status consists of 3 unique values, Fully paid, Charged off and Current respectively.
5) We will retain only those rows where loan status is not equal to Current as goal is to find out which customer is likely to default which can only be analyzed for either fully paid and charged off loans.
6) After removing rows when loan status equals ‘current’ we were left with 38577 rows

Loan Status





Inference :

1) The fully paid category has 32950(~85%) datapoints and Charged Off has 5627(~15%) datapoints respectively.
2) The overall default rate of whole dataset is 14.6%.

Removing Irrelevant Columns

- 55 columns had columns were 100% empty and 1 column was more than 90% empty. These were removed
- Removed 11 columns as they were having having 0 variance.
- Removed revol_bal as the revol_bal info is also contained in revol_util
- Removed title as the revol_bal is also contained in purpose
- Removed other irrelevant features like id, member_id, url, desc, earliest_cr_line, last_credit_pull_d, zip_code.
- Removed 16 post-loan features as we have to do EDA on features which are applicable before loan approval

Segregrating the remaining features into continuous and categorical/numeric discrete type

- continuous = ['loan_amnt’ ’int_rate’, 'annual_inc’, 'dti’, 'revol_util’, 'total_acc']

- Categorical = ['term’, 'grade’, 'sub_grade’, 'emp_title’, 'emp_length’, 'home_ownership’, 'verification_status’, 'purpose’, 'addr_state’, 'open_acc’, 'pub_rec’, 'pub_rec_bankruptcies’]

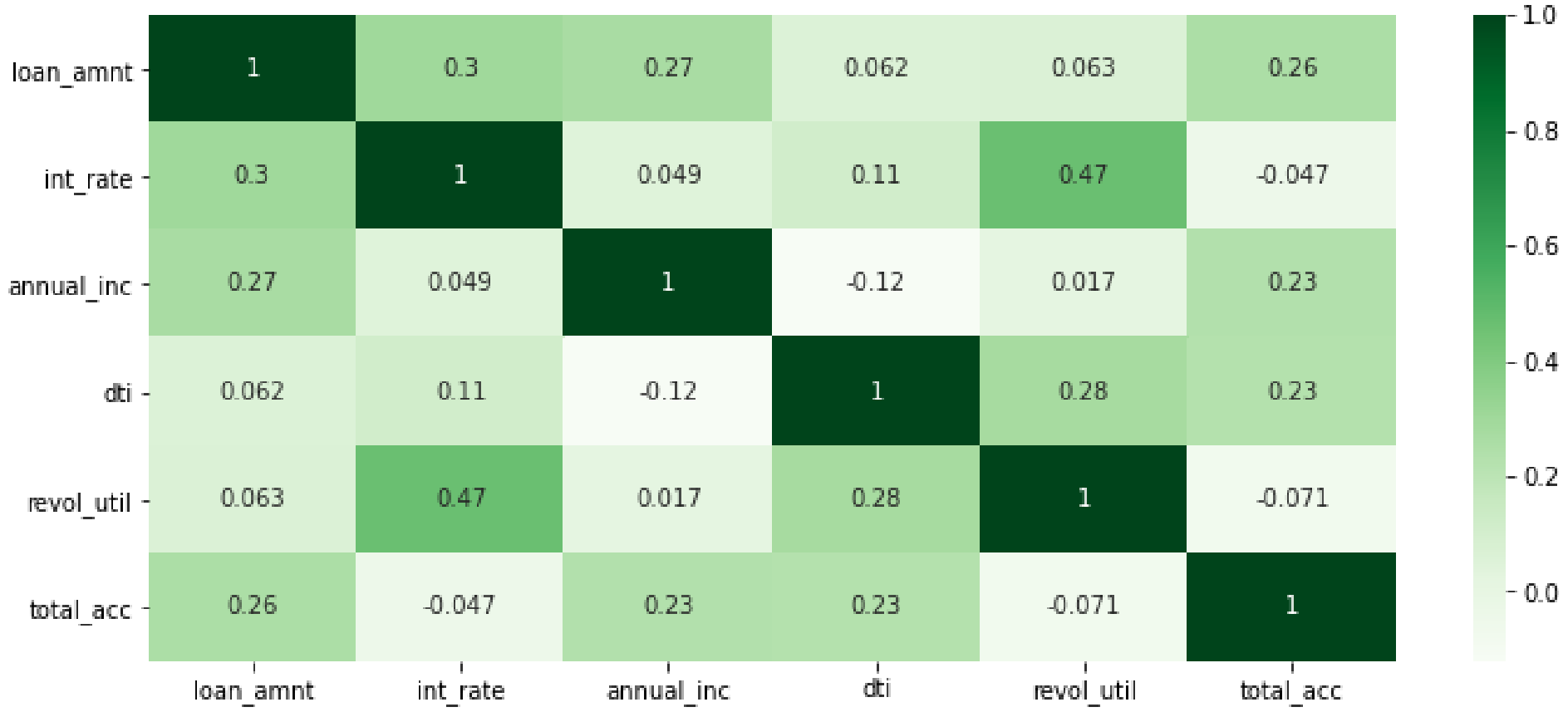
- target_column = 'loan_status’

- We have 6 continuous and 12 categorical/numerical discrete features respectively.

Data Cleaning and Missing Value Imputation

- % character was removed from int_rate and revol_util columns and there datatypes were changed to float
- ‘years’ and ‘months’ words were removed from emp_length and term columns respectively.
- The missing values of revol_util column were imputed with its median
- The missing values in each of emp_title, emp_length, pub_rec_bankruptcies columns were imputed with a new value 'unknown_col_name’ eg; missing values in 'emp_length' will be imputed with 'unknown_emp_length'

Correlation Analysis between continuous variables

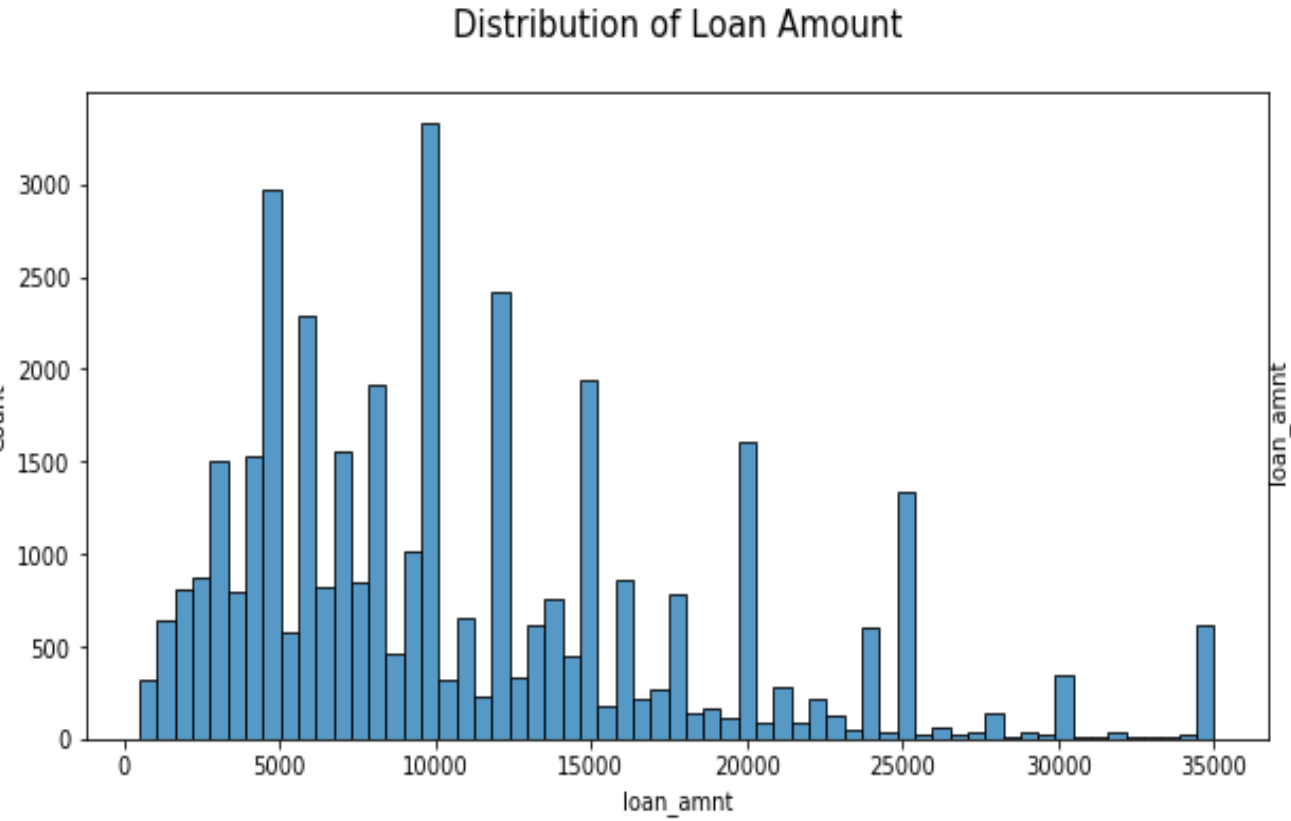


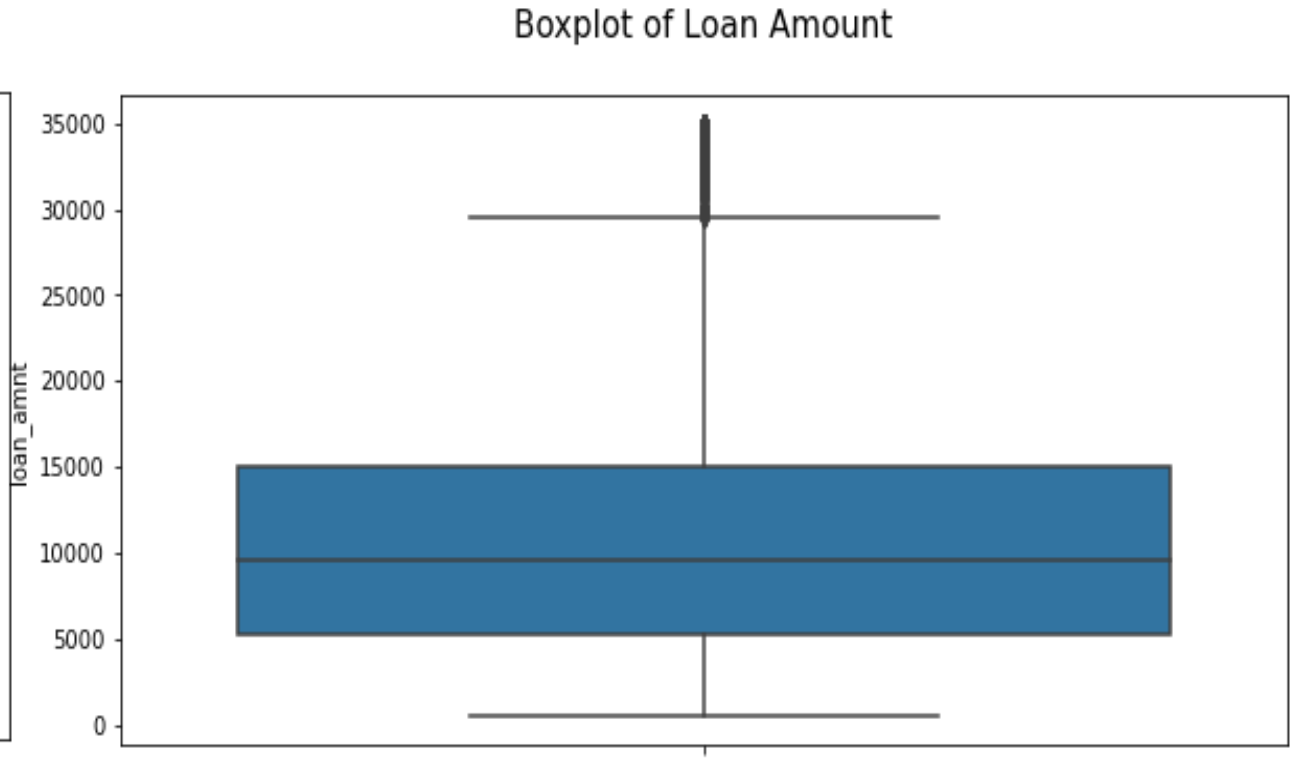
Inference:

- No strong linear correlation exists between any pair of continuous variables

Univariate Analysis(Continuous Features)

1) Loan Amount

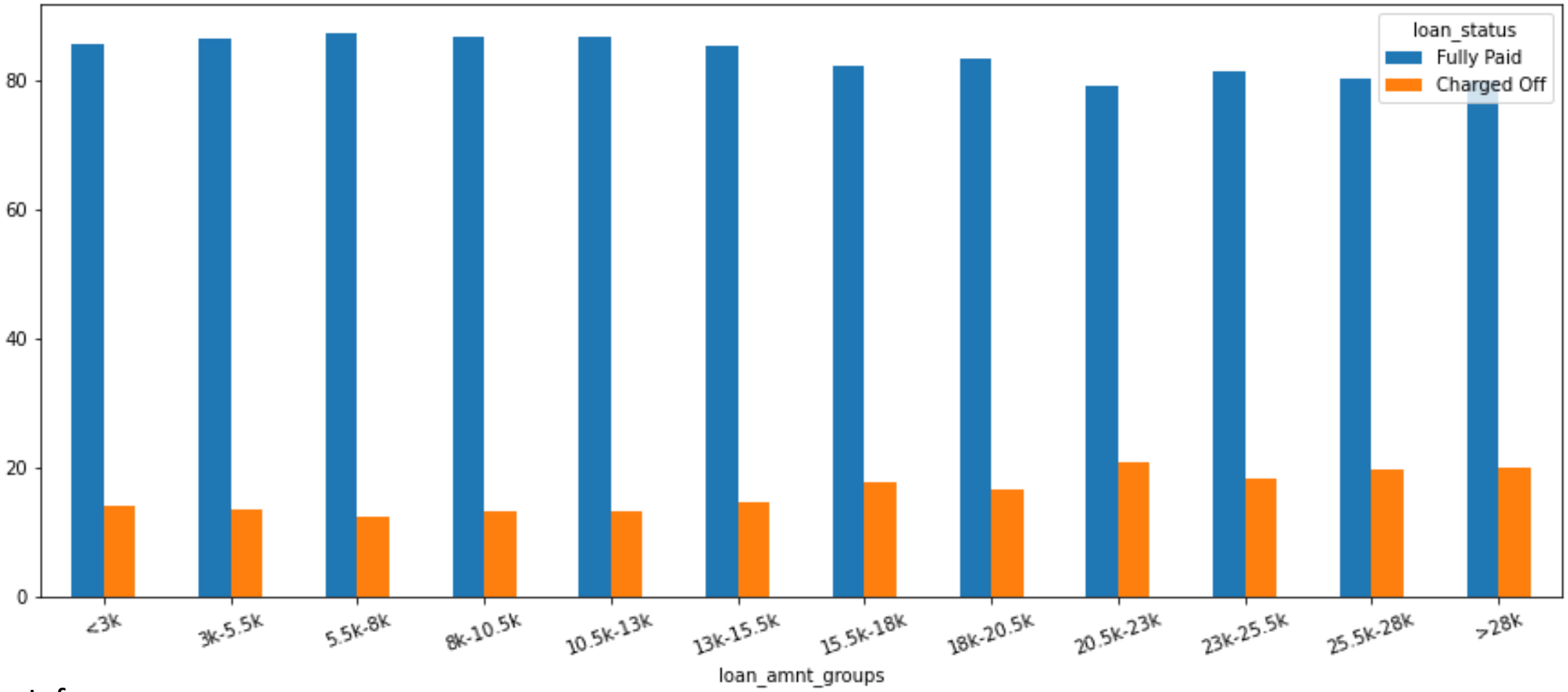




Inference:

1) Most of the loan amount(about 95%) are less than 25000
2) 50% of loan amount are less than 9600
3) 75% of loan amount are less than 15000
4) The max value of loan amount is 35000
5) Their are are outliers as shown in whiskers after loan amount of around 29000

Relation of loan amount with loan status

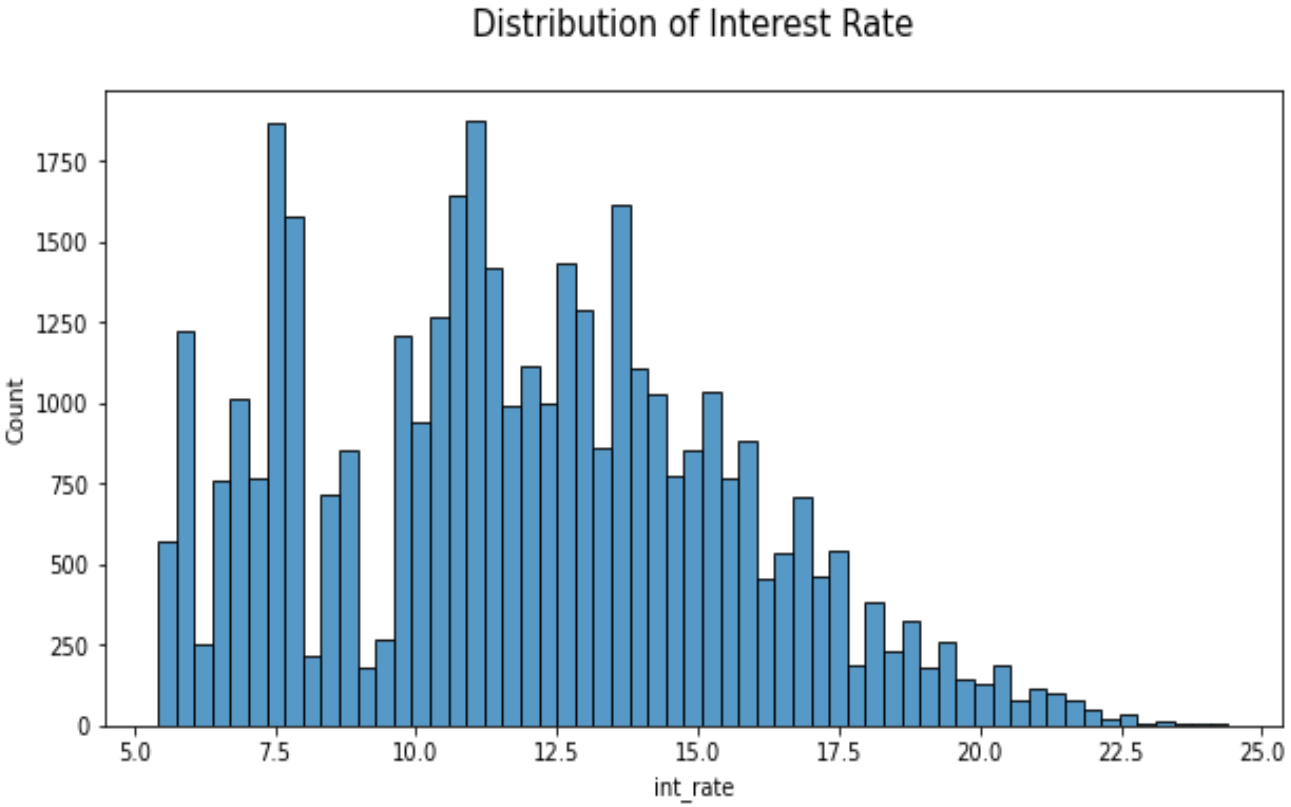


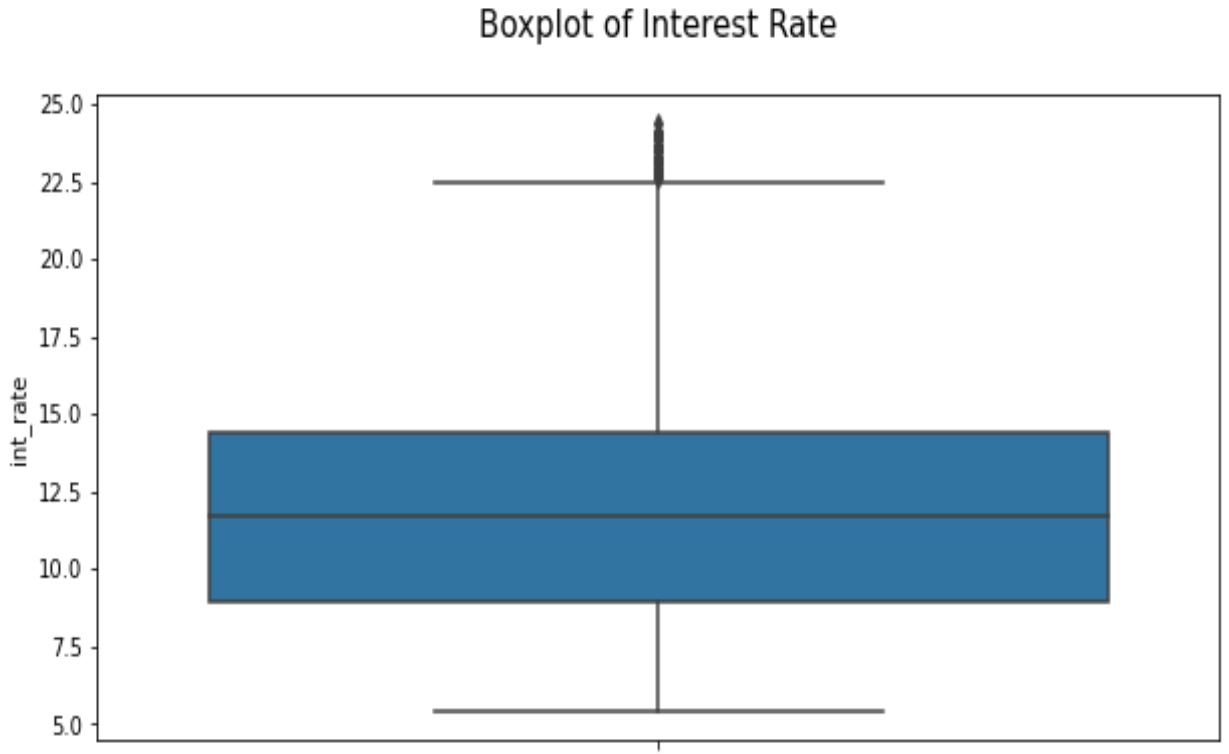
Inference:

1) In general we see that with increasing loan amount the probability of defaults increase

2) The highest default rate is when the loan amount is between 20.5k-23k(~21%), followed by when loan amount is greater than 28k(~20%)

2) Interest rate

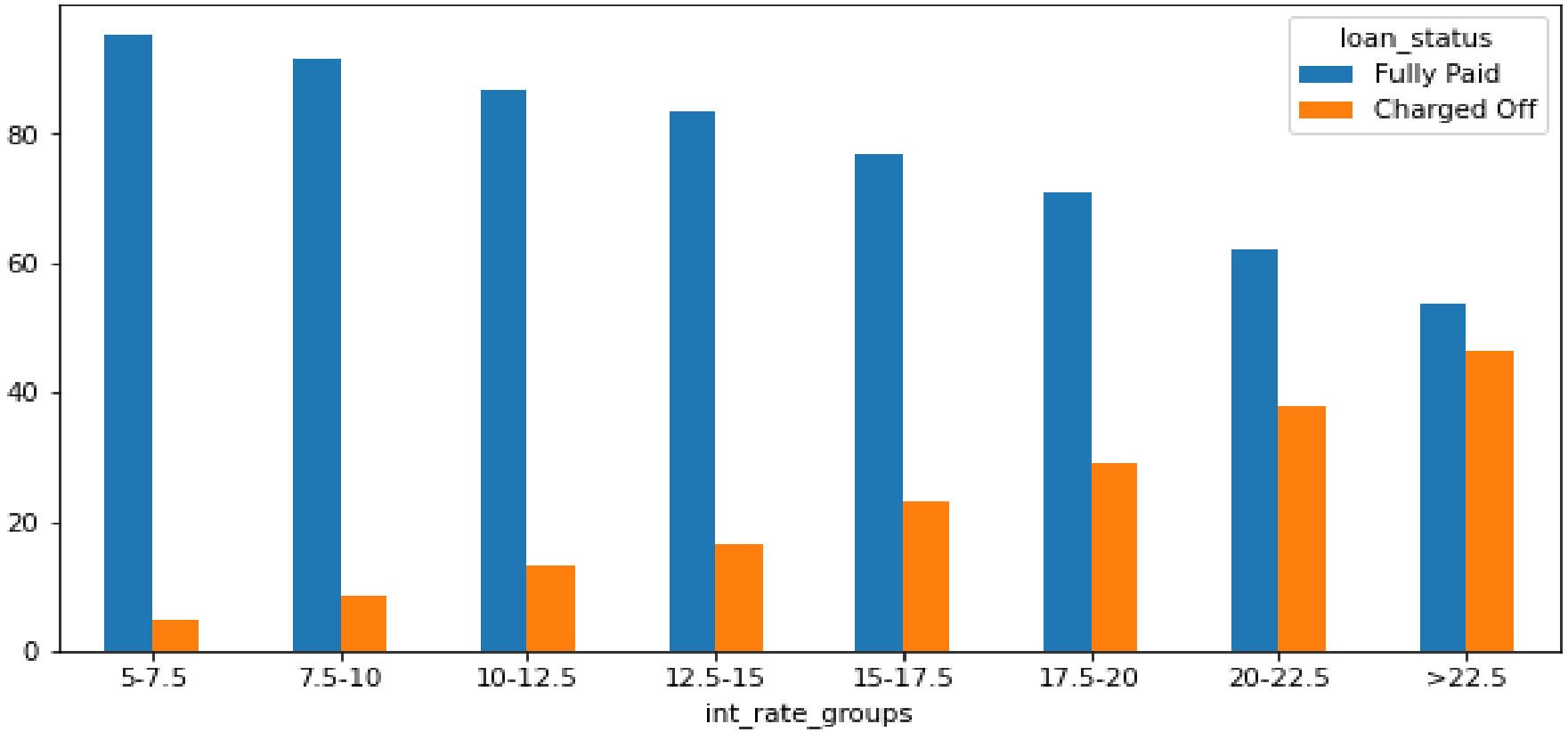




Inference:

1) Most of the interest rate(about 95%) is less than 18.3
2) 50% of interest rate is less than 11.7
3) 75% of interest rate is less than 14.38
4) The max value of interest rate is 24.4
5) There are outliers after interest rate of 22.5 as shown via boxplot and points greater than 22.5 are capped at 22.5

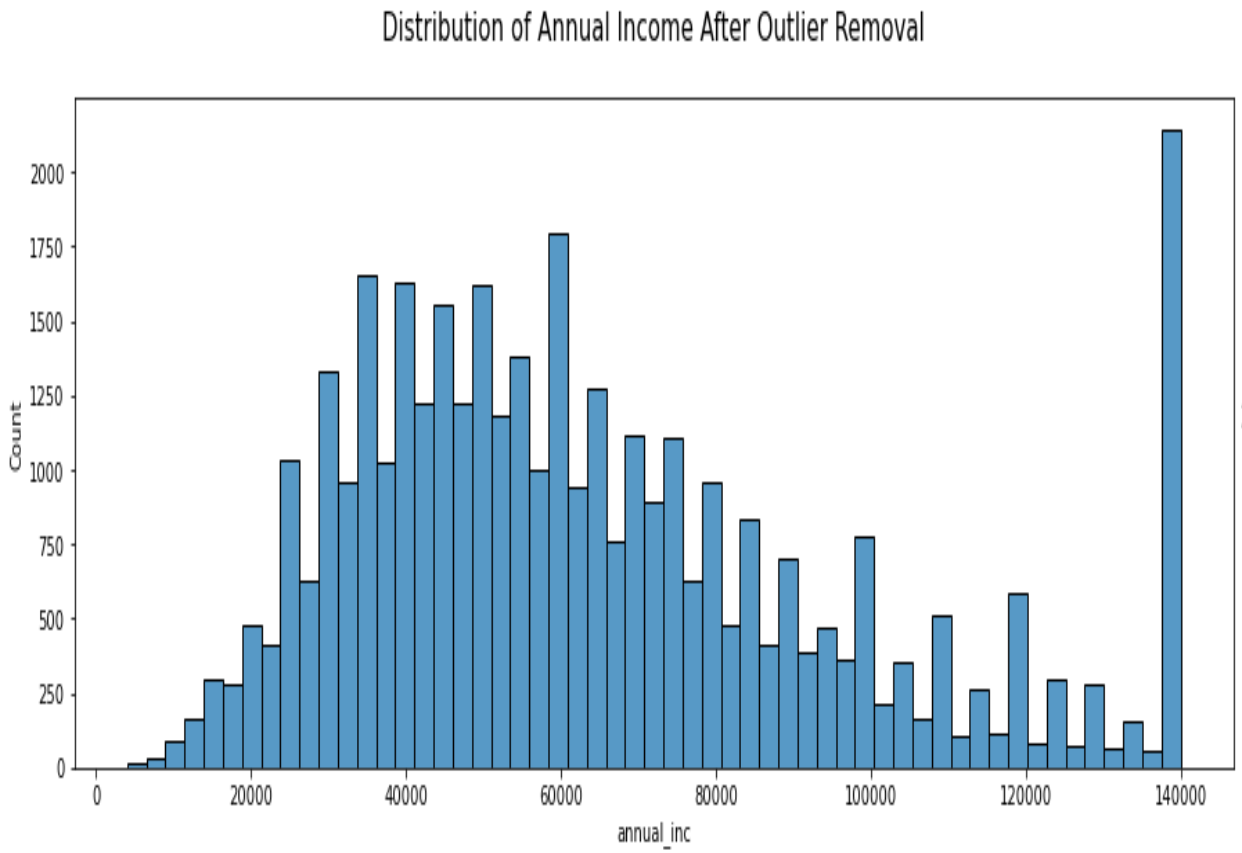
Relation of Interest Rate with loan status

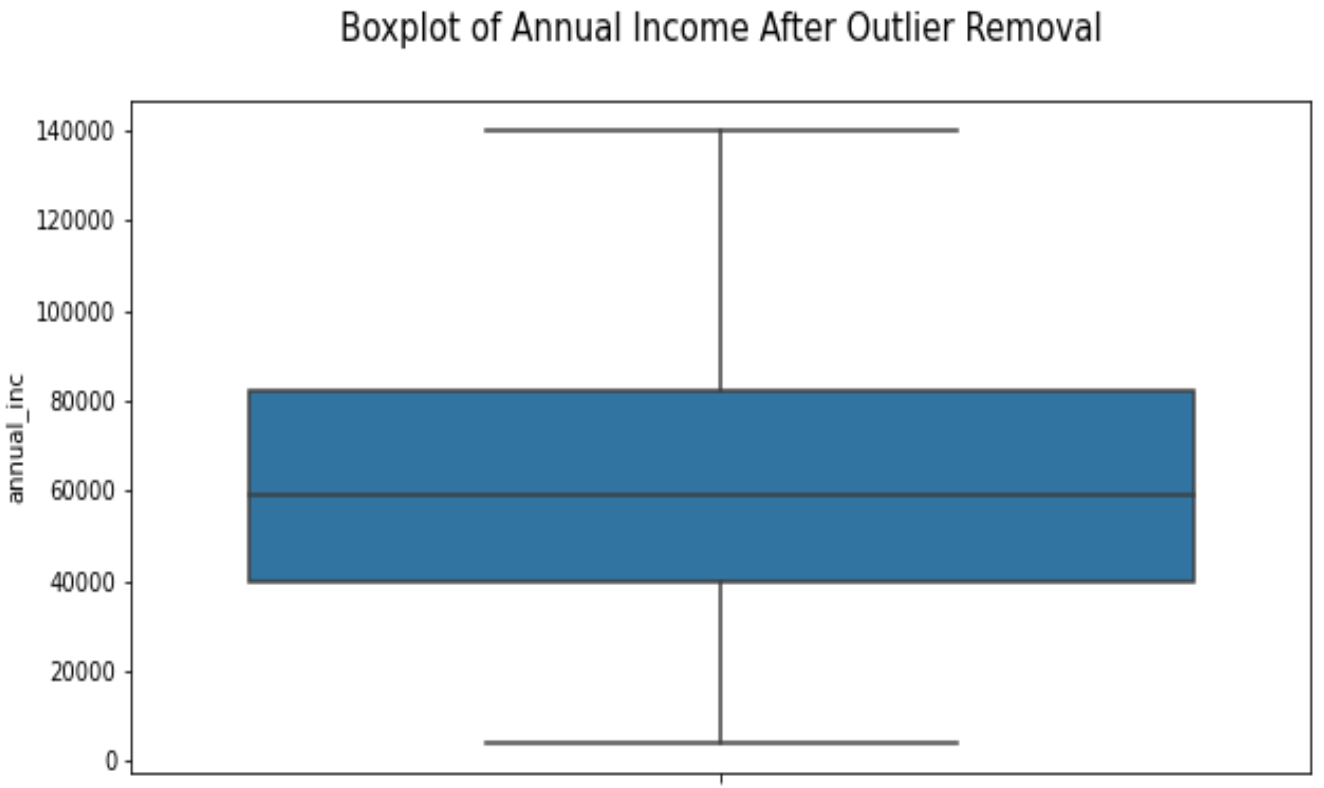


Inference:

1) In general we see that with increasing interest rate the probability of defaults increase
2) The highest default rate is when the interest rate is greater than 22.5(~46%), followed by when interest rate is between 20-22.5(~38%)
3) The lowest default rate is when the interest rate is between 5-7.5(~4.8%) followed by when interest rate is between 7.5-10(~8.3%)

3) Annual Income(after outlier removal)

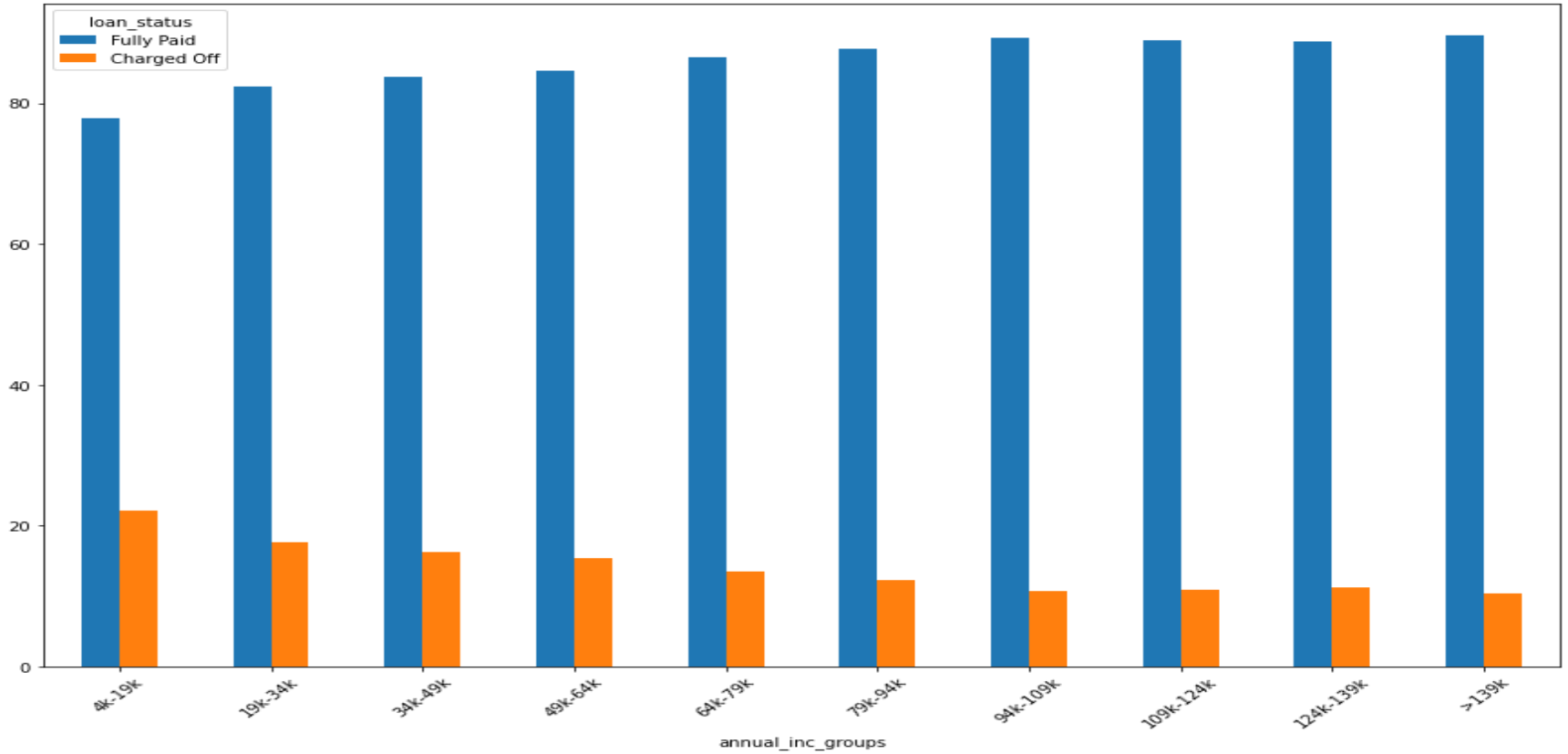




Inference:

1) Most of the Annual Income(about 95%) are less than 140000
2) 50% of loan amount are less than 58868
3) 75% of loan amount are less than 82000 and the max value of annual income is 6000000
5) Their are outliers after Annual Income of 140000 , so points greater than 140000 are capped at 140000

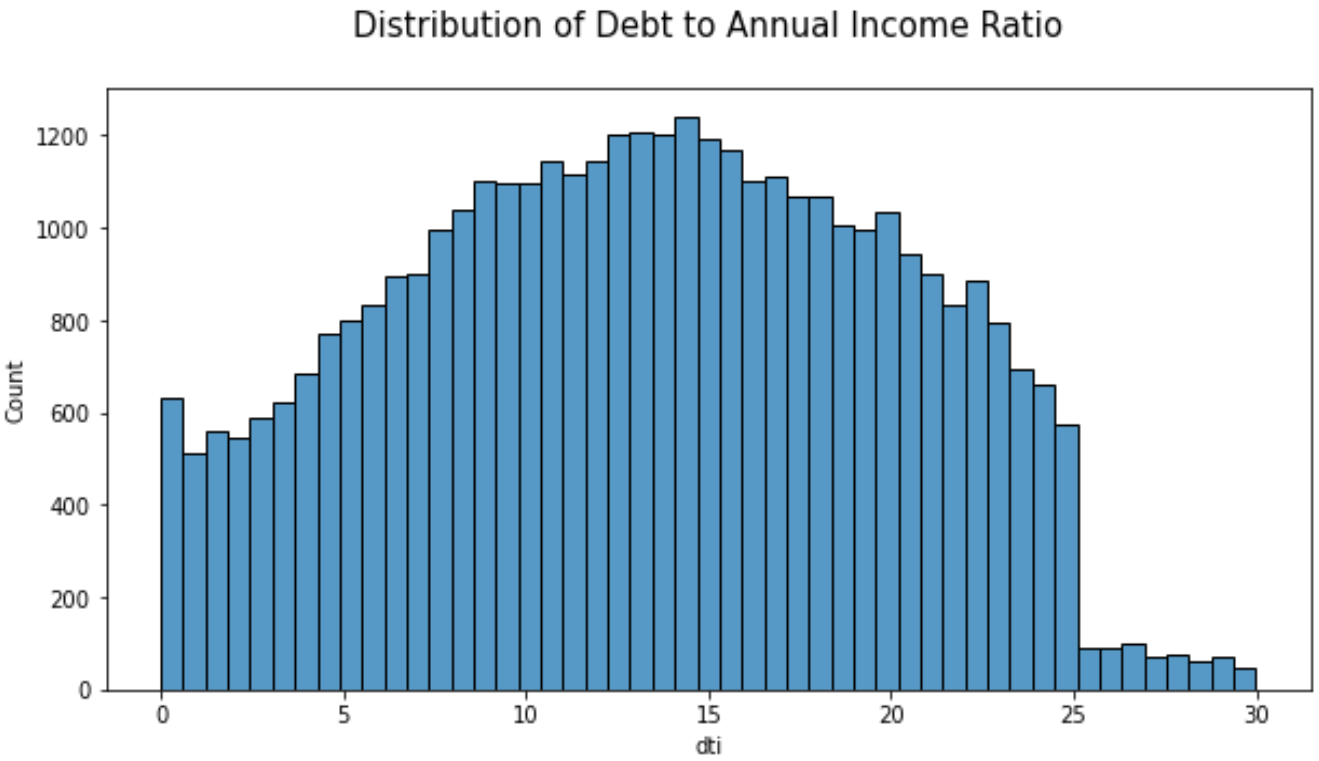
Relation of Annual Income with Loan Status

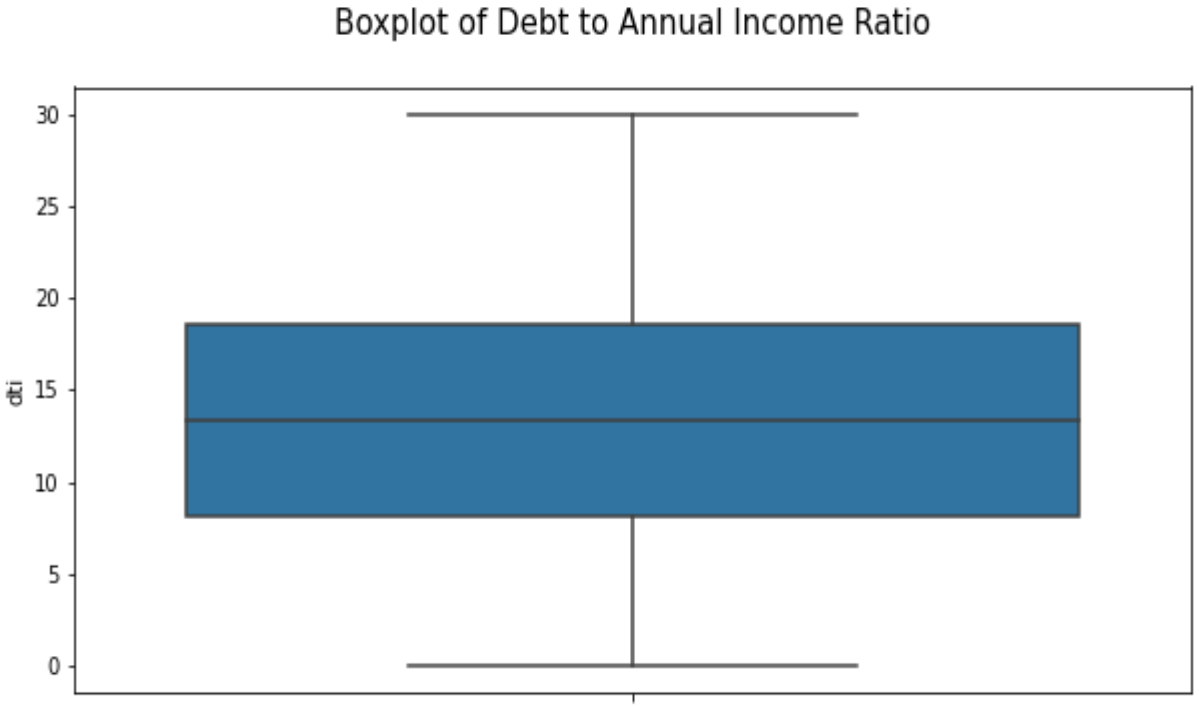


Inference:

1) In general we see that with increasing annual income the probability of defaults decrease
2) The highest default rate is when the annual income is between 4k-19k(~22%), followed by when annual incomes is between 19k-34k(~17.6%)
3) The lowest default rate is when the annual income is greater than 139k(~10.4%)

4) Debt to Annual Income Ratio(dti)





Inference:

1) Most of the dti points(about 95%) are less than 22.3

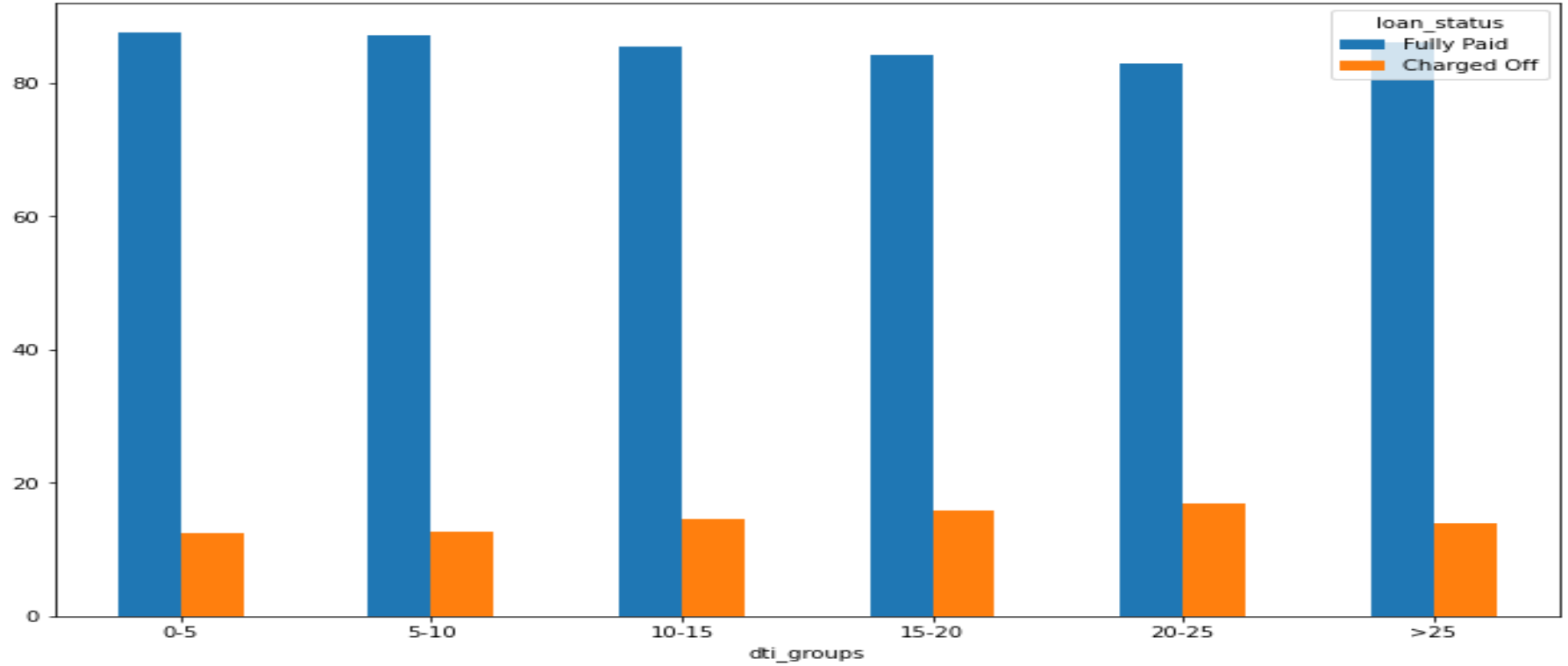
2) 50% of dti points are less than 13.37

3) 75% of dti points are less than 18.56

4) The max value of dti is 29.99

5) Their are no outliers as shown via boxplot

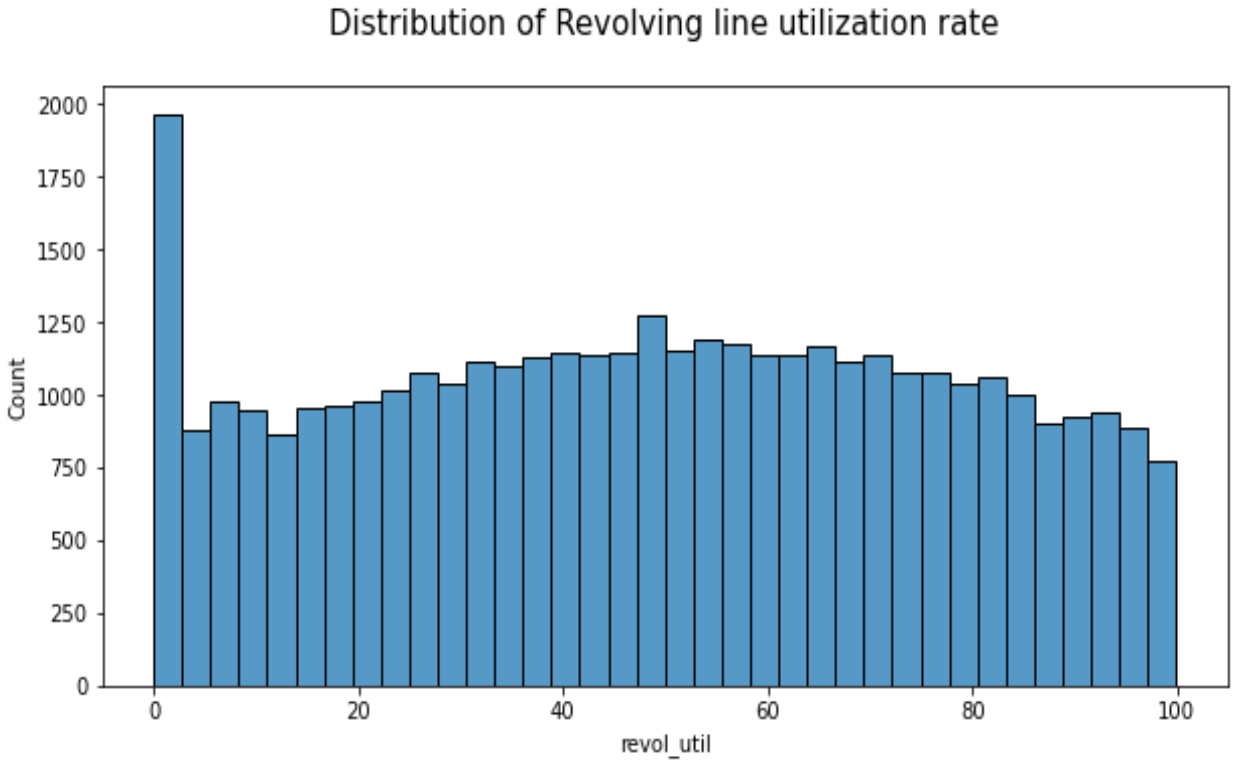
Relation of dti with Loan Status

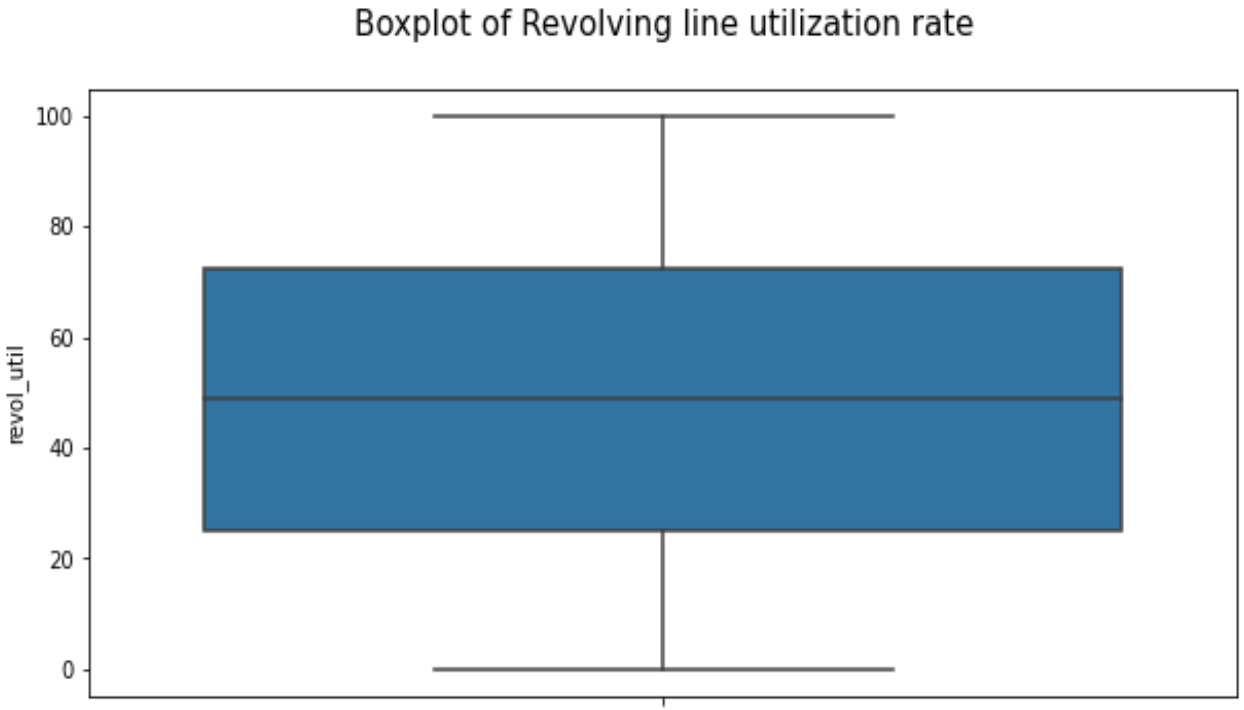


Inference:

1) In general we see that with increasing dti the probability of defaults decrease with exception of dti greater than 25 but this could a random variation.
2) The highest default rate is when the dti is between 20-25(~17%), followed by when dti is between 15-20(~16%)
3) The lowest default rate is when the dti is less than 5(~12.4%)

5) Revolving line utilization rate(revol_util)





Inference:

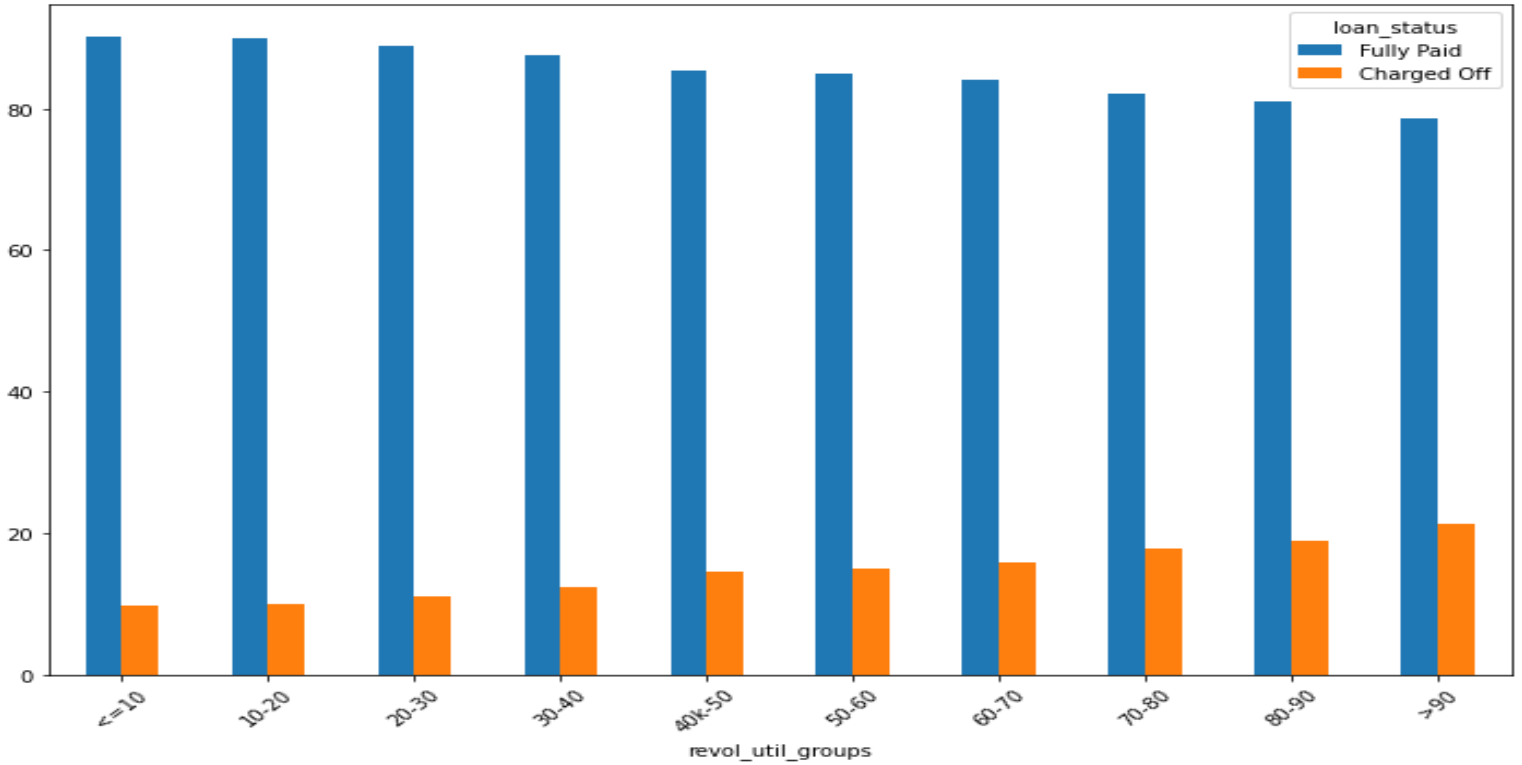
1) Most of the revol_util(about 95%) are less than 93.5

2) 50% of revol_util are less than 49.1

3) 75% of revol_util are less than 72.2 and the max value of revol_util is 99.9

4) Their are no outliers as shown via boxplot

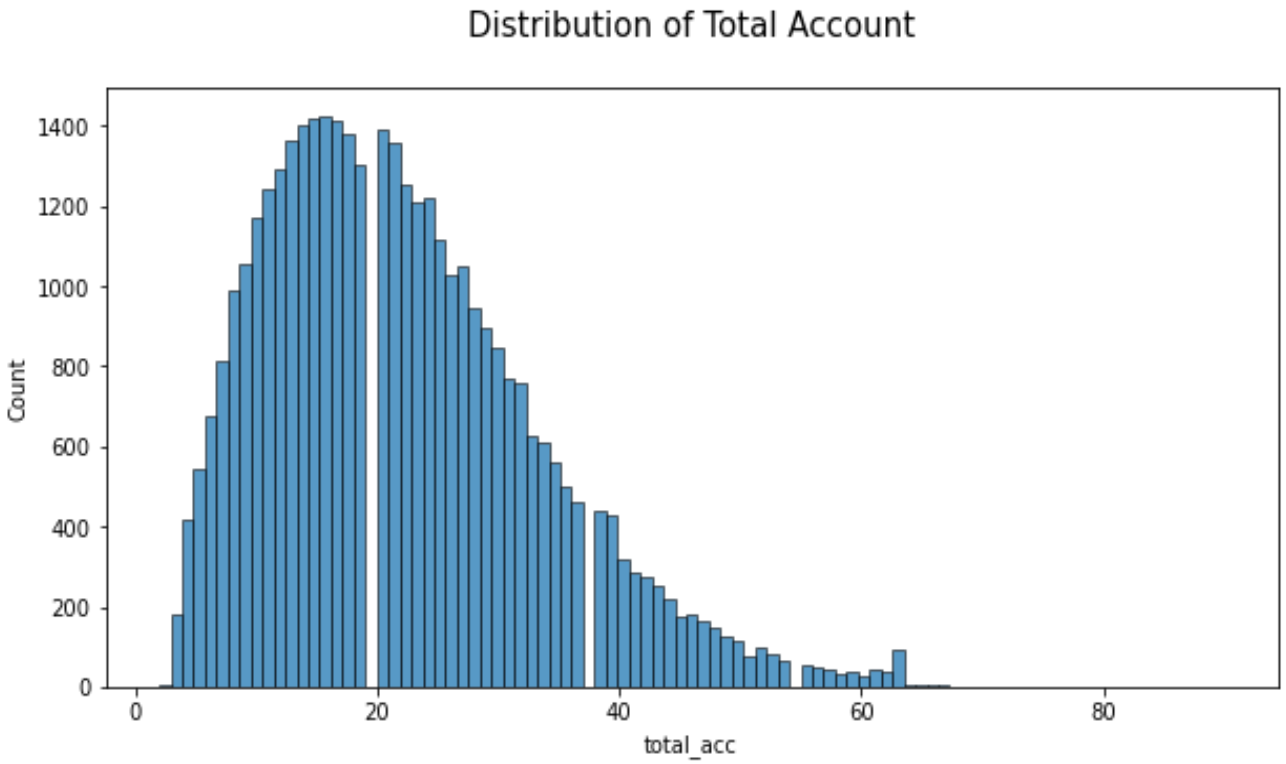
Relation of revol_util with Loan Status

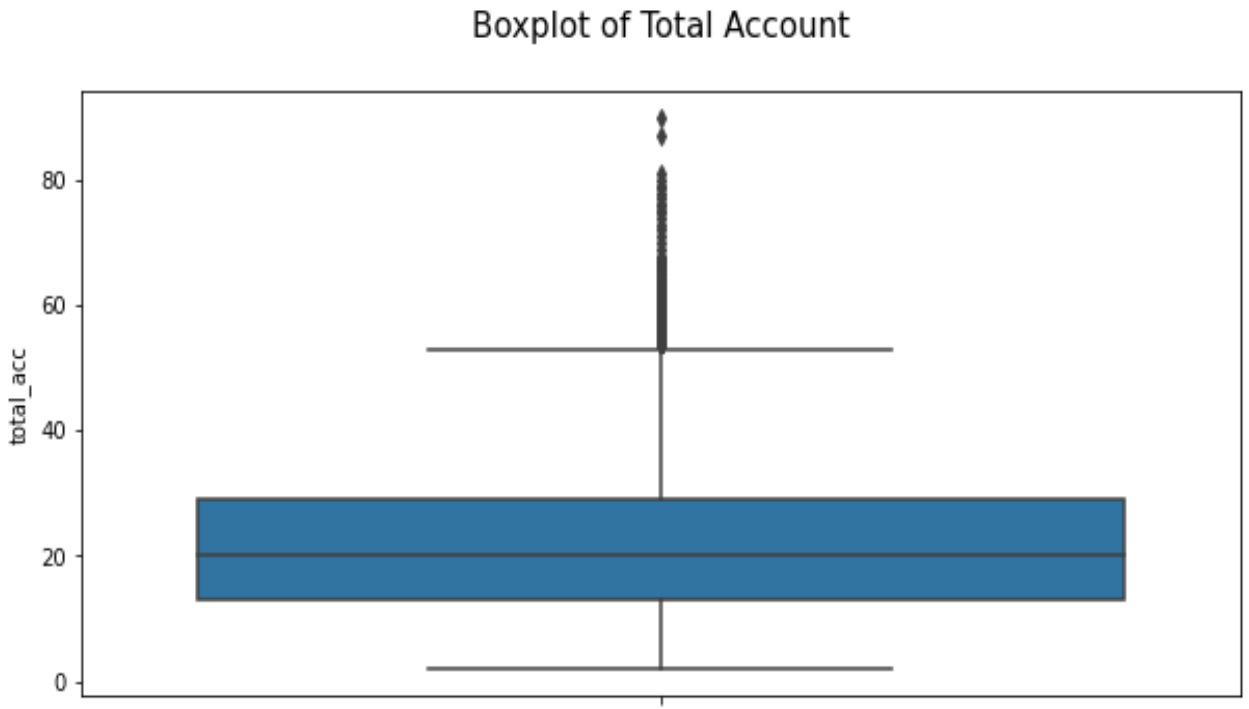


Inference:

1) In general we see that with increasing revol_util rate the probability of defaults increases.
2) The highest default rate is when revol_util is greater than >90(~21%), followed by when revol_util is between 80-90(~19%).
3) The lowest default rate is when revol_util is less than 10(~9.8%).

6) Total number of credit lines(total acc)





Inference:

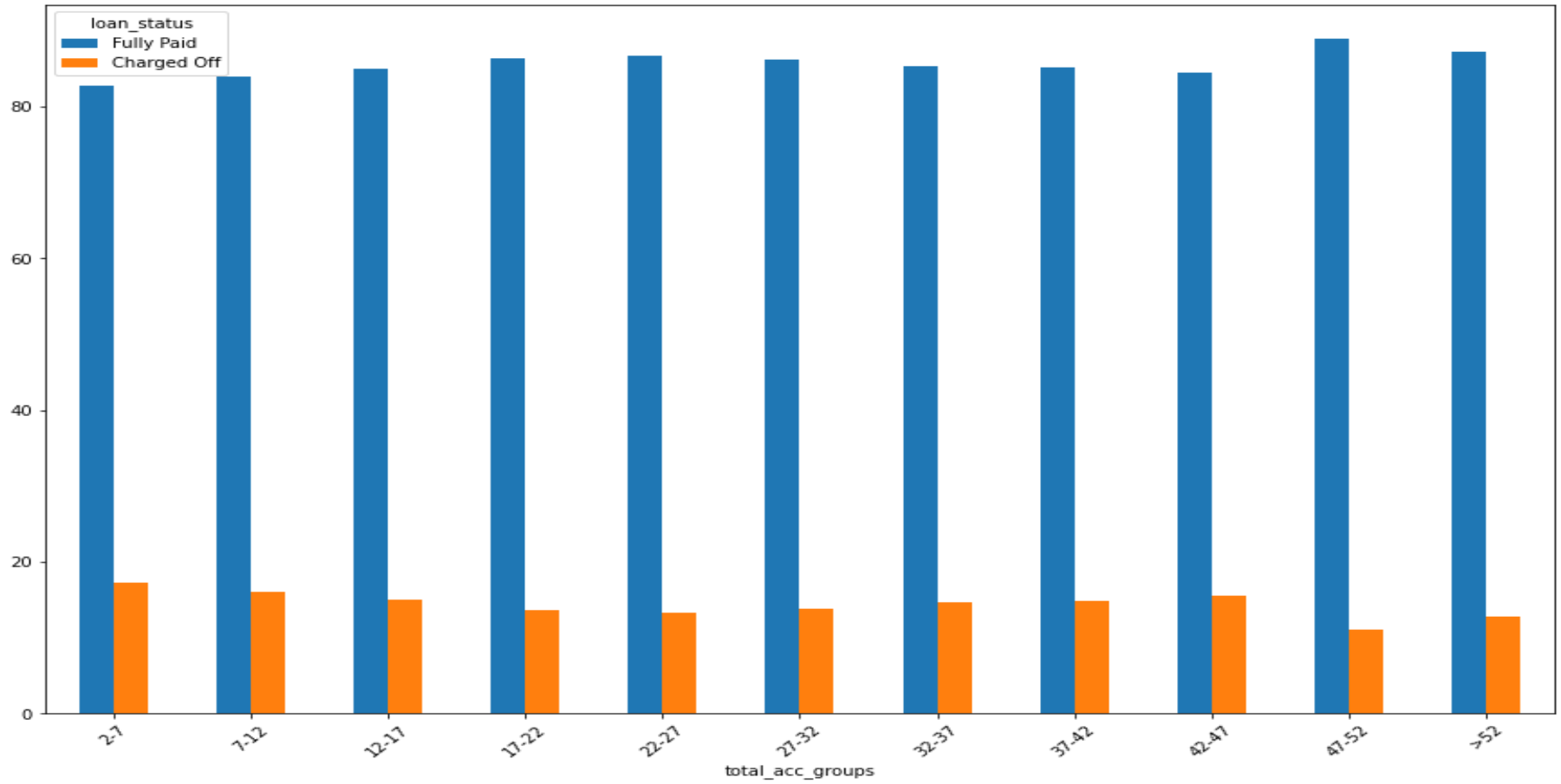
1) Most of the credit lines(about 95%) are less than 43

2) 50% of credit lines are less than 20

3) 75% of credit lines are less than 29 and the max value of credit lines is 90

4) There are outliers after credit lines value of 53 and values greater than 53 will be capped at 53

Relation of total credit lines with Loan Status

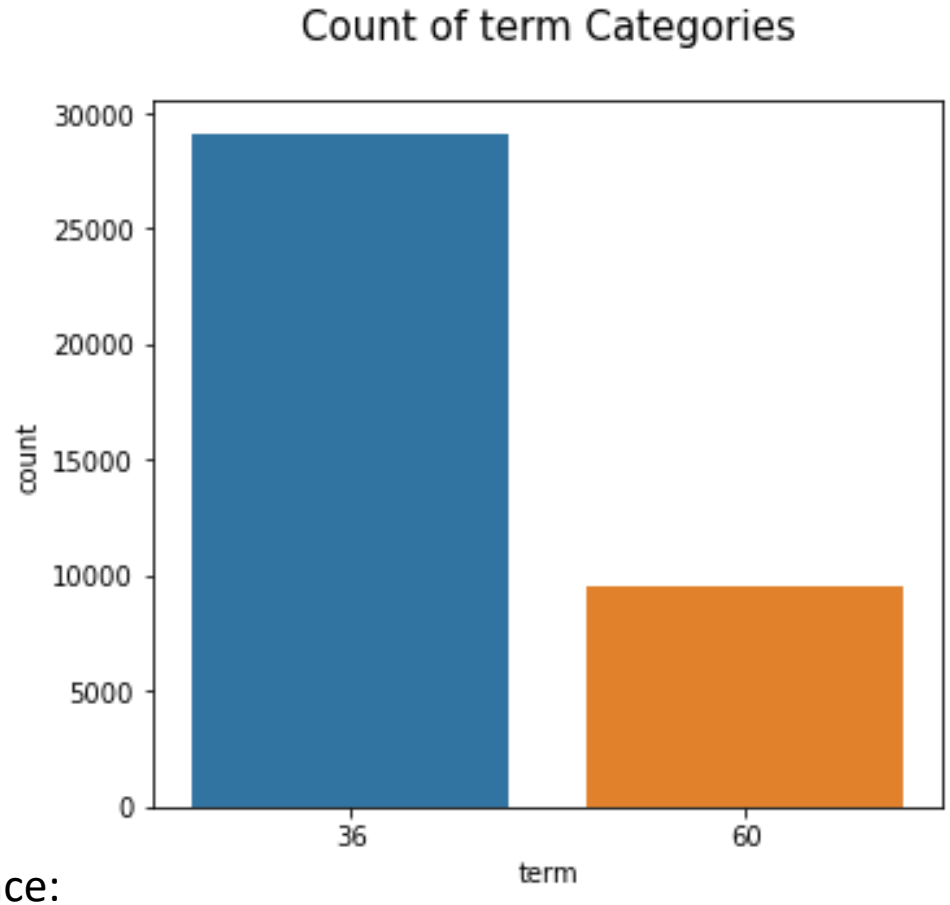


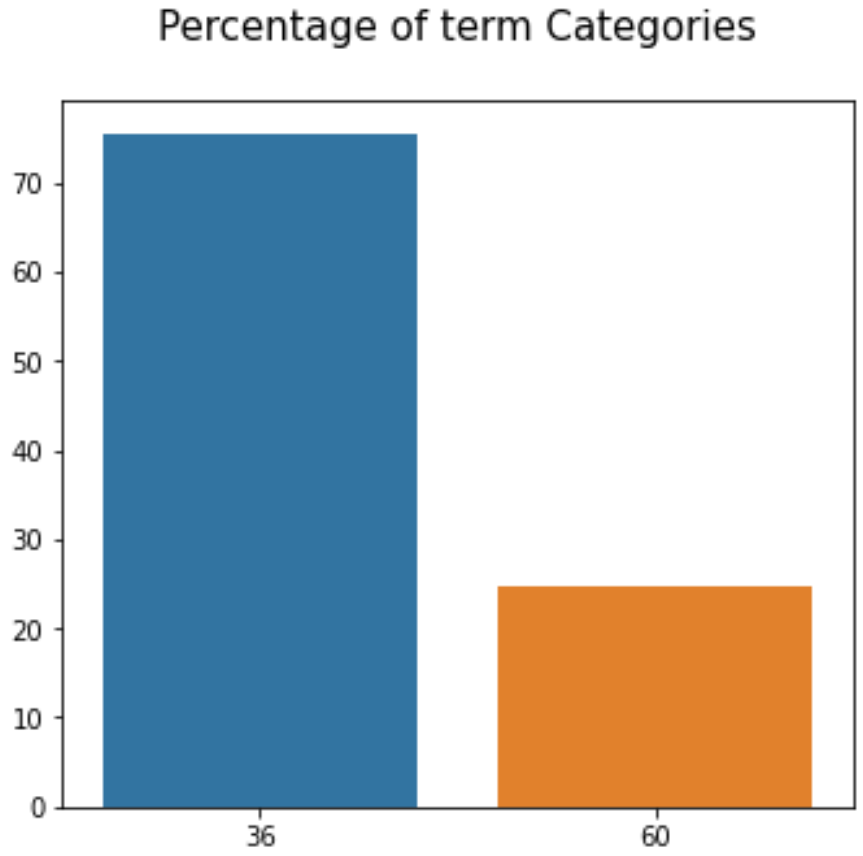
Inference:

1) Their seems to be no general relationship of total credit lines with loan status
2) With increasing total credit lines first decreases default rate then after a certain number of total account the default rate starts to increase
3) The highest default rate is when total credit lines is between 2-7(~17%)
4) The lowest default rate is when total credit lines is between 47-52(~11%)

Univariate Analysis(Categorical Features)

1) Term

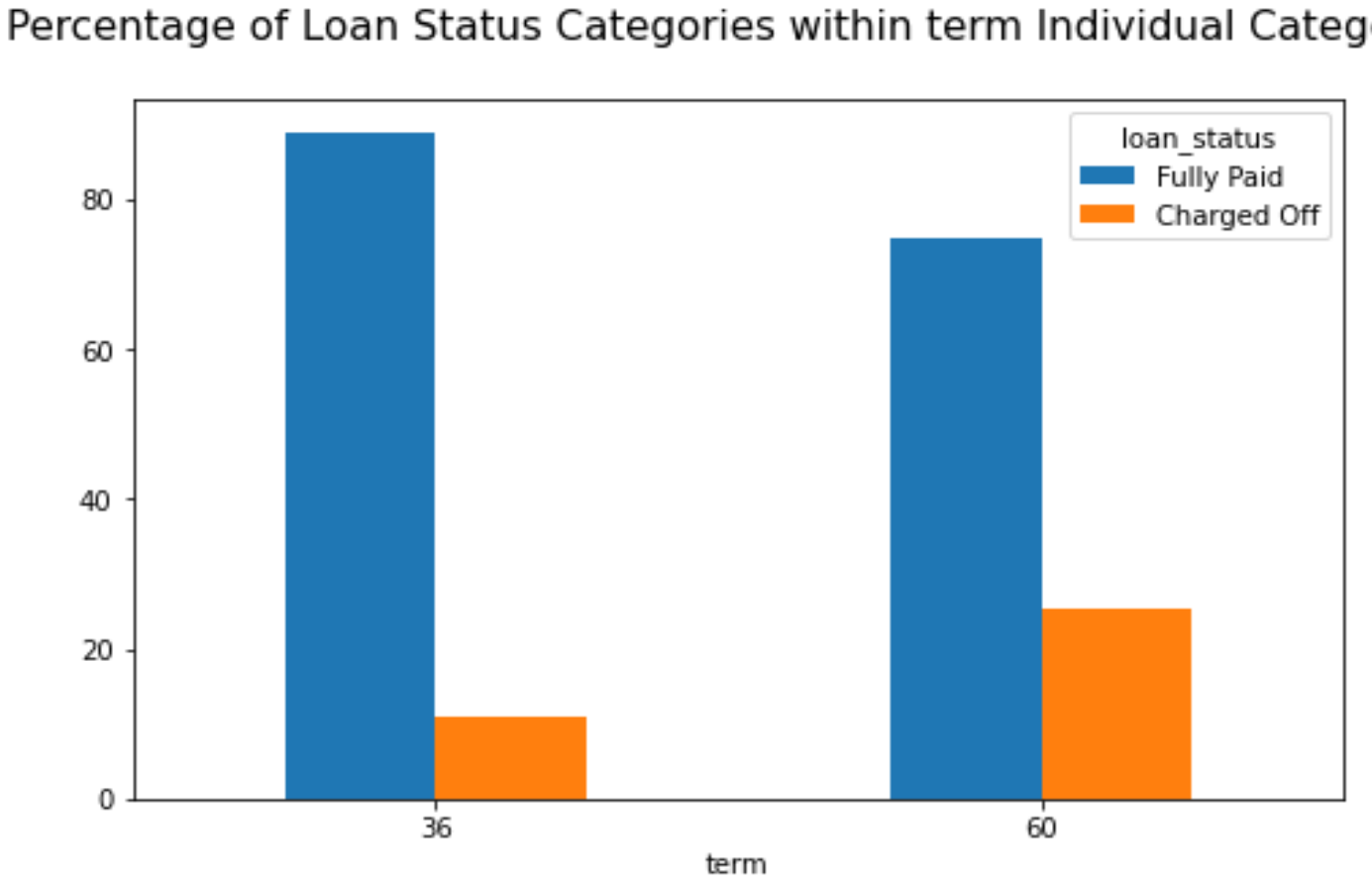




Inference:

1) Term column consists of 2 unique categories, 36 and 60 namely
2) The counts of 36 and 60 is 29096 and 9481 respectively.
3) 36 is 75% and 60 is 25% of datapoints

Relation of term with Loan Status

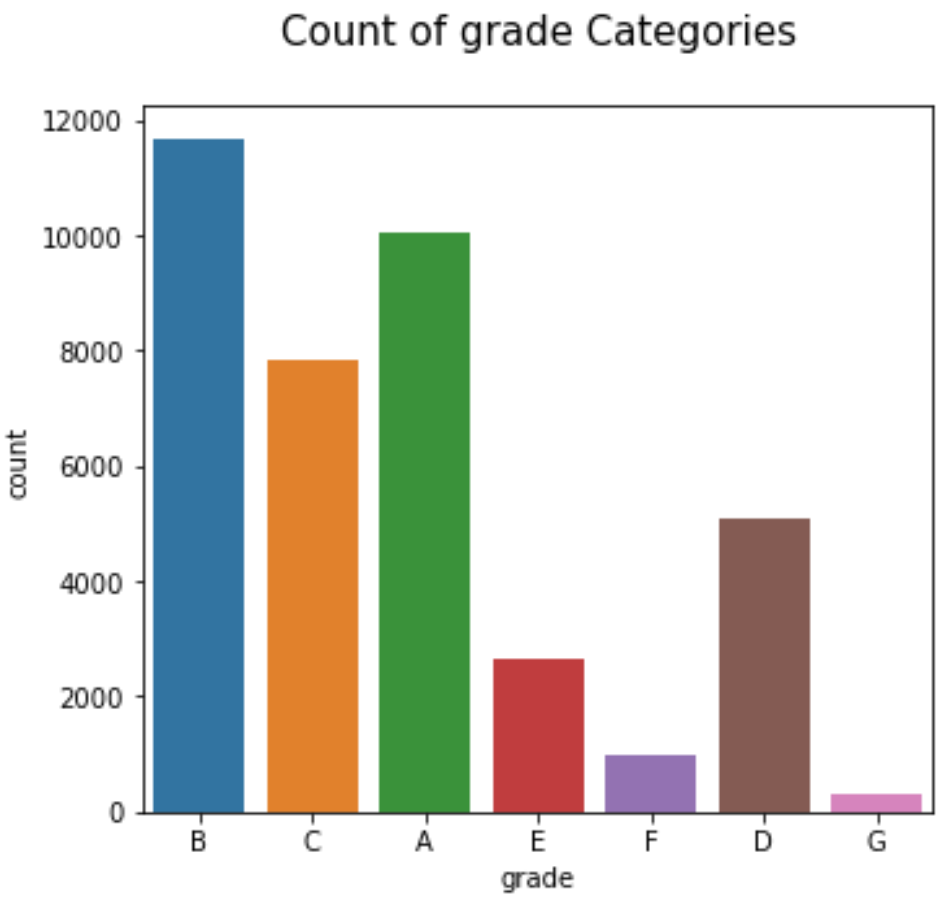


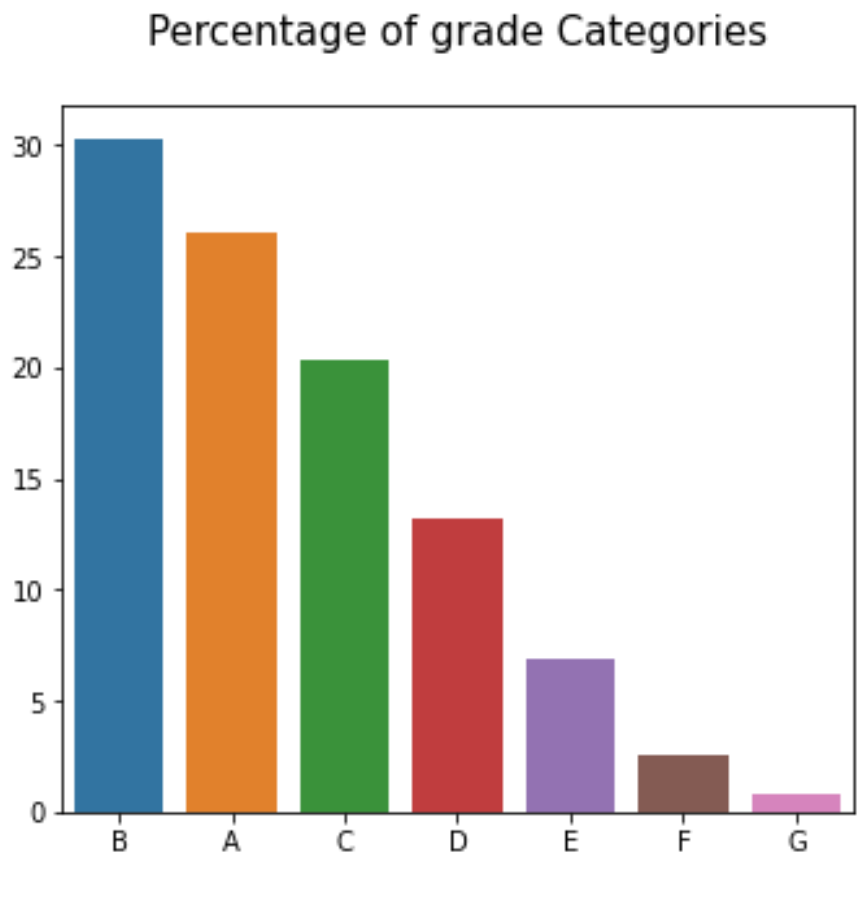
Inference:

1) The probability of default is more high for a loan given for 60 months than for 36 months

2) For 36 months the default rate is 11% and for 60 months its 25%

2) Grade





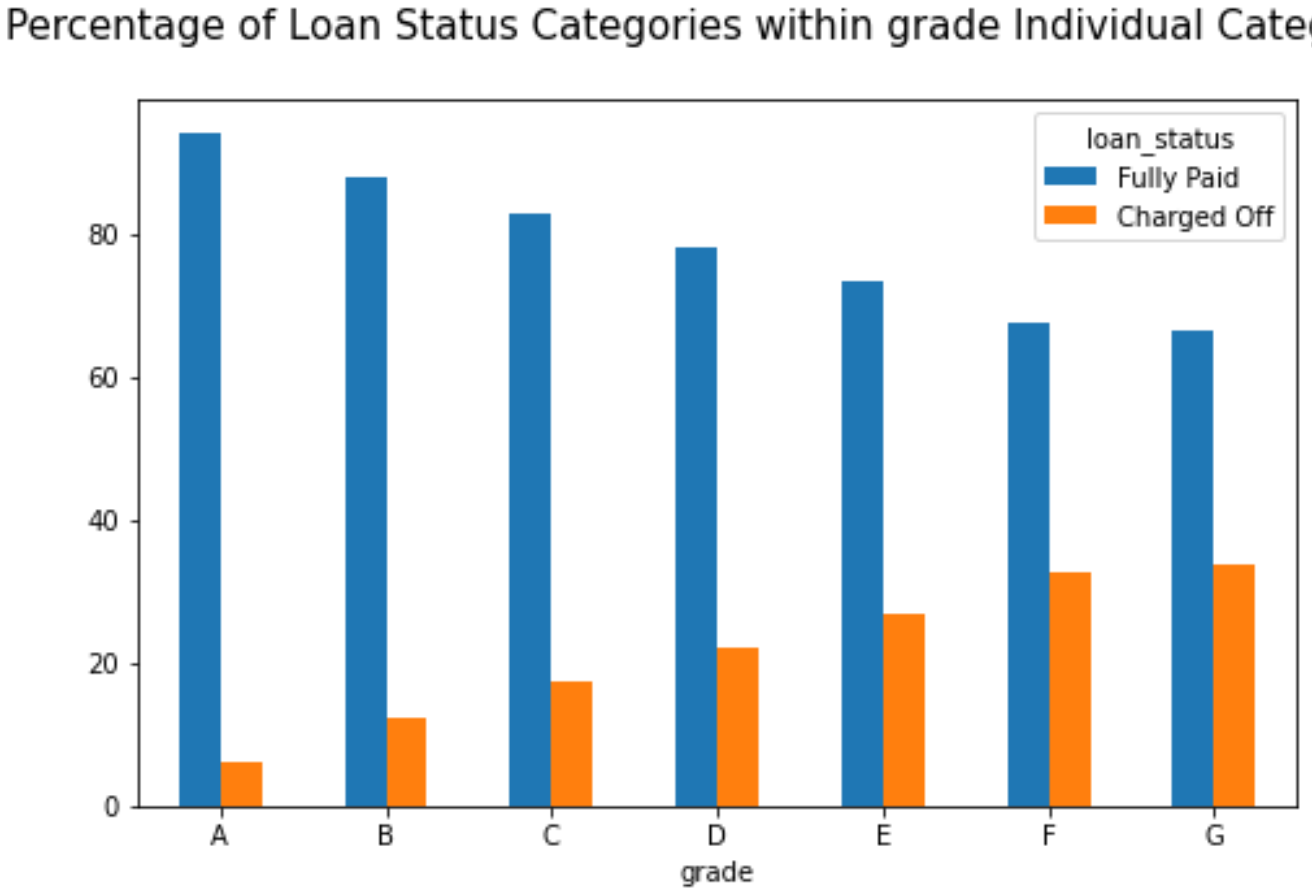
Inference:

1) Grade column consists of 7 unique categories, A,B,C,D,E,F,G namely

2) The count of B is highest with 11675(~30%) datapoints, followed by A with count a of 10045(~26%).

3) G is the lowest category in terms of count of datapoints; i.e 299(~0.8%)

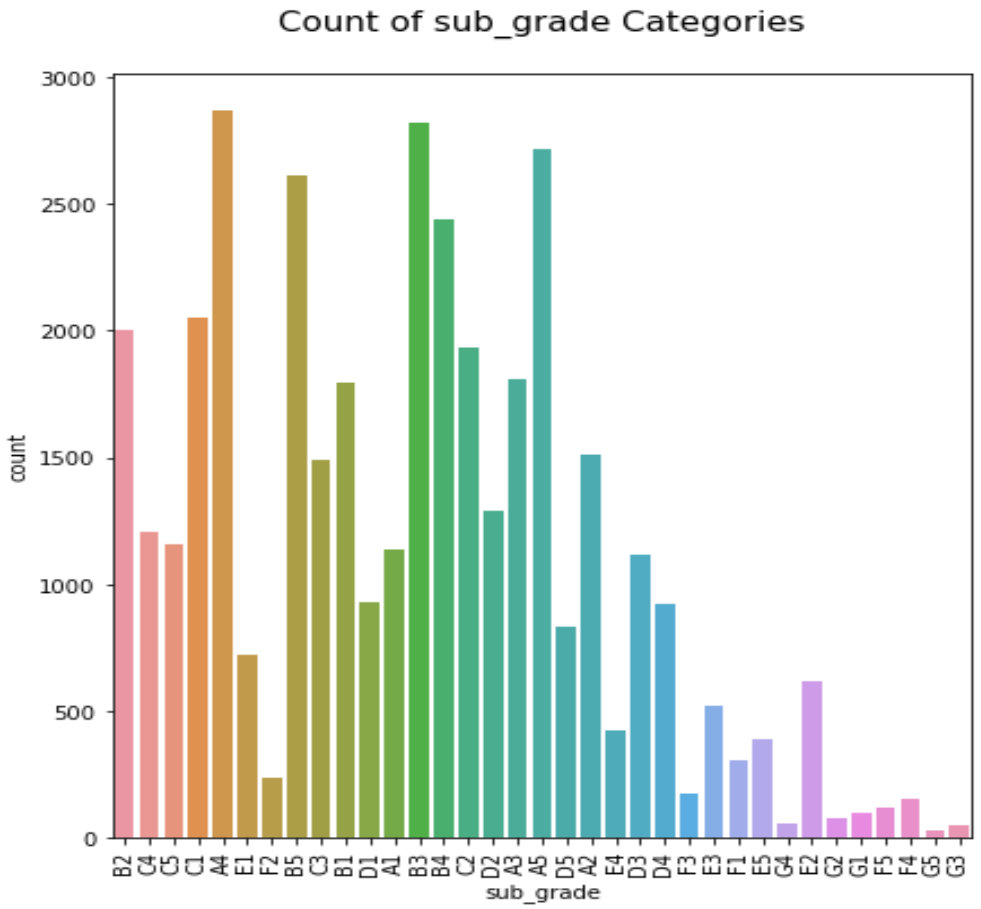
Relation of Grade with Loan Status

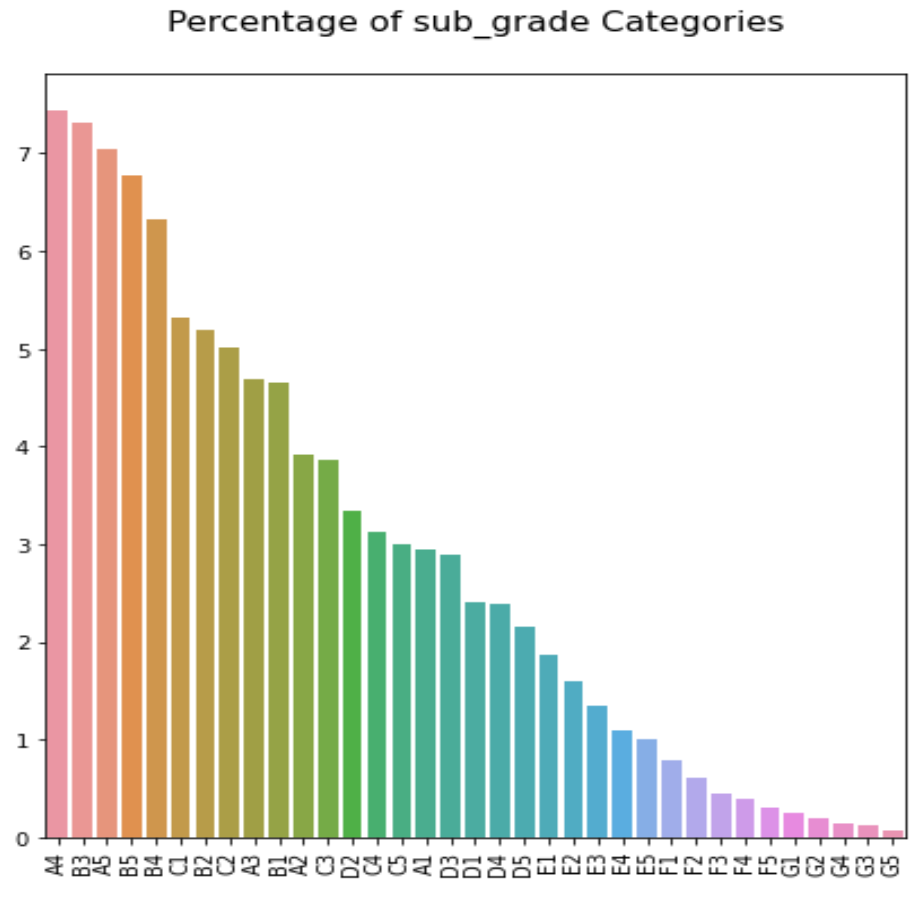


Inference:

1) The probability of default increases as we move in alphabetical order of grades; i.e the probability of default is more in G grade (~34%), less in F(~33%), more less in E(~27%) and same pattern follows till A(~6%)
2) Their is a order associated with grades

3) Sub-grade

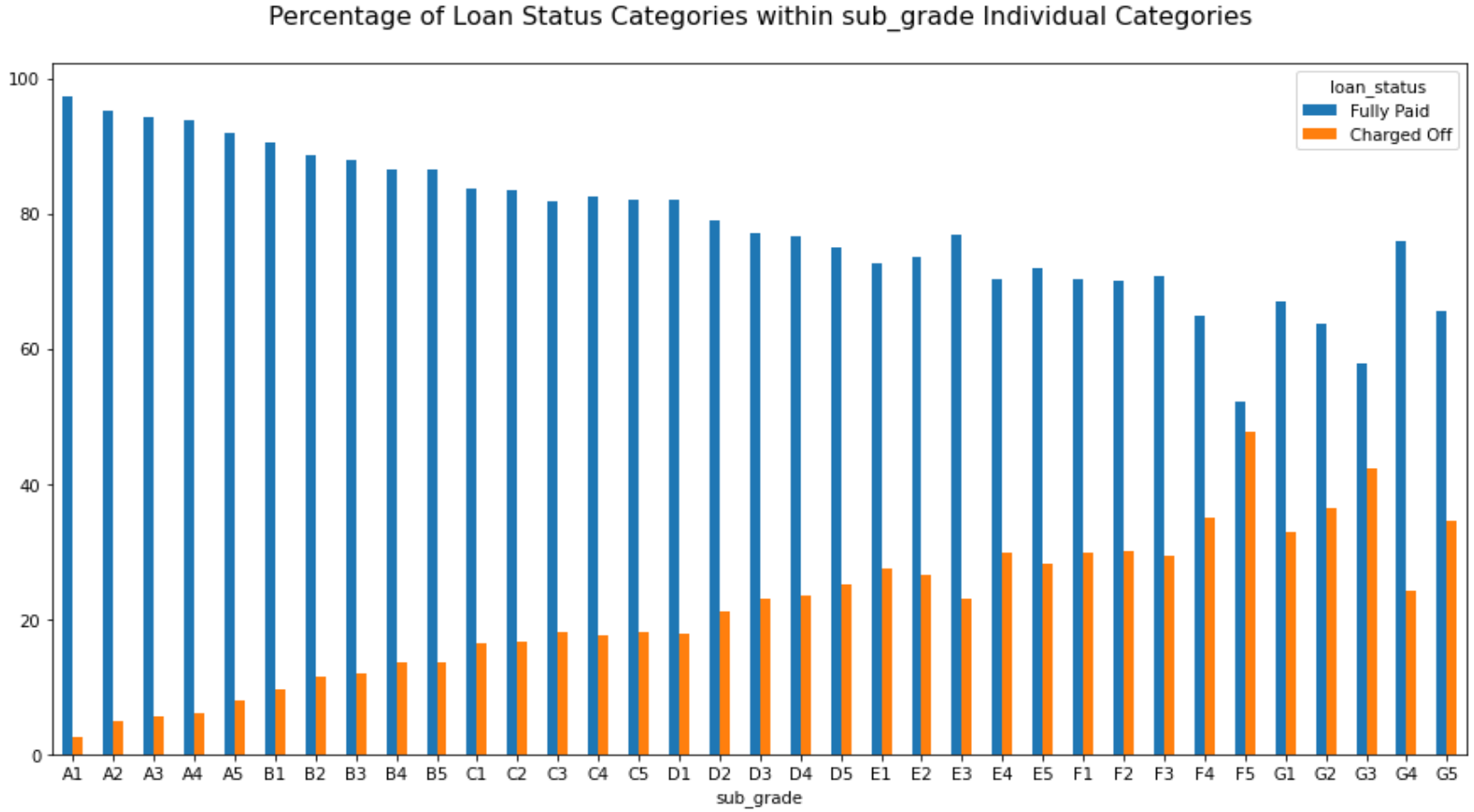




Inference:

1) Sub-Grade column consists of 35 unique categories, A grade has 5 unique sub grades A1, A2, A3, A4, A5 and same pattern follows for other 6 grades as well.
2) The count of A4 is highest with 2873(~7.4%) datapoints, followed by B3 with count a of 2825(~7.3%).
3) G5 is the lowest category in terms of count of datapoints; i.e 29(~0.08%)

Relation of Sub-grade with Loan Status

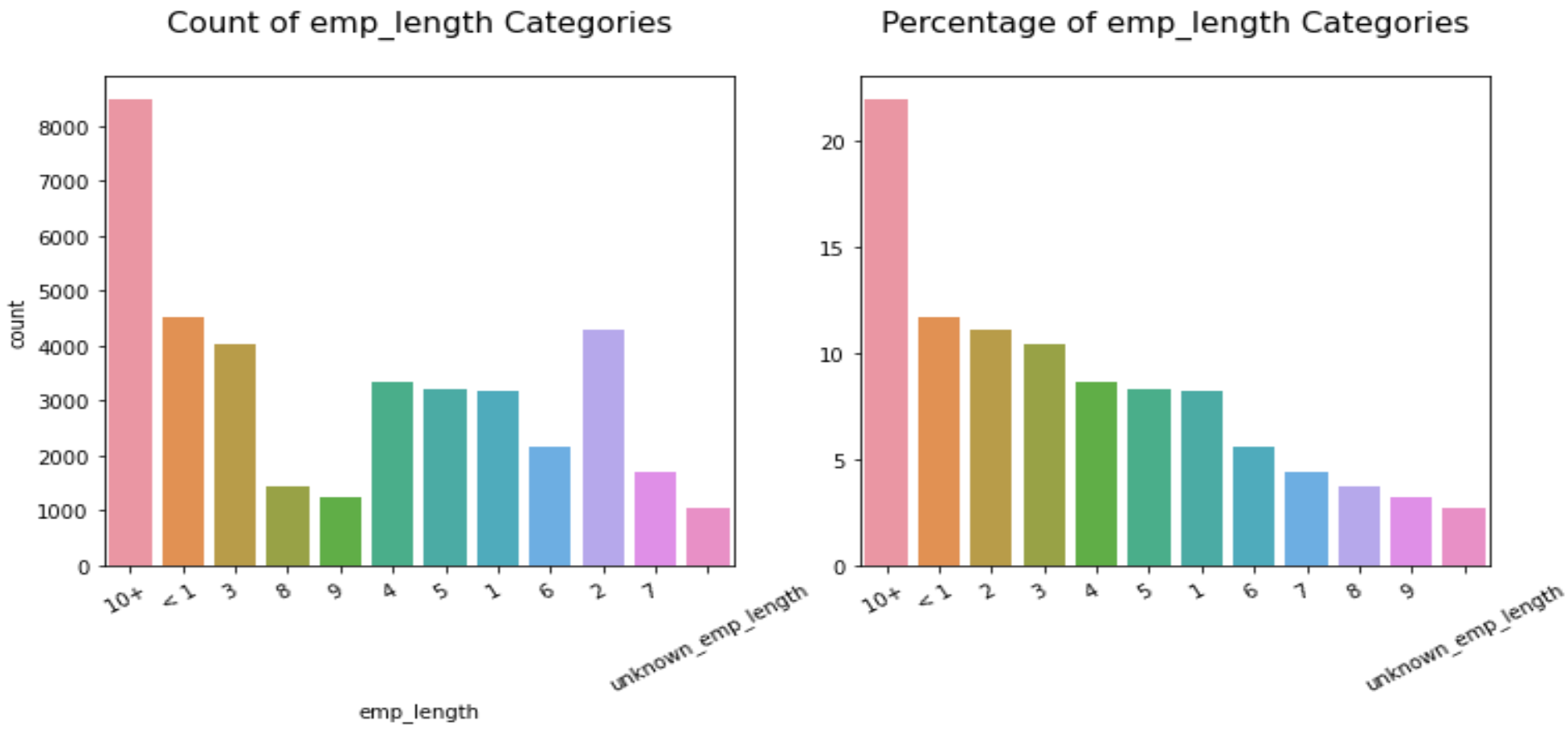


Inference:

1) Sub Grade column follows the same pattern as that of grade with loan status column leaving a few exceptions; i.e the A sub grades default rate is lower than of B, B sub grades have a default rate less than of C and same pattern follows till G sub grades

2) Within each grade the default rate increases with alphabetical order, eg; within A grade the default rate of A1 is less than of A2, A2 is less than of A3, A3 is less than of A4 and A4 is less than of A5.

4) Employment length in years(emp_length)



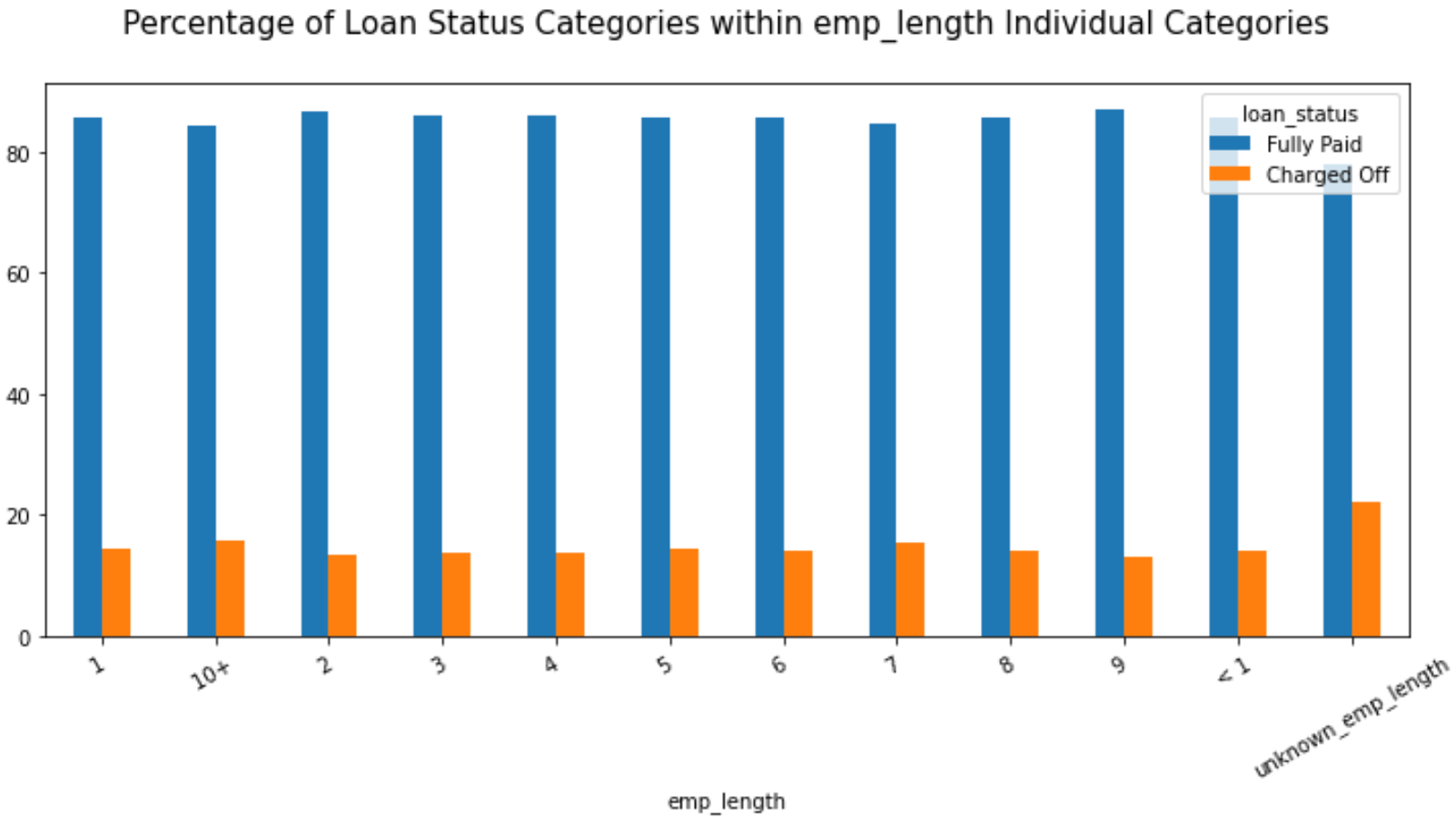
Inference:

1) Employee Length column consists of 11 unique categories, < 1, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10+ namely

2) The count of 10+ is highest with 8488(~22%) datapoints, followed by <1 with count a of 4508(~11.7%).

3) unknown_emp_length is the lowest category in terms of count of datapoints; i.e 1033(~2.7%)

Relation of Employee Length with Loan Status

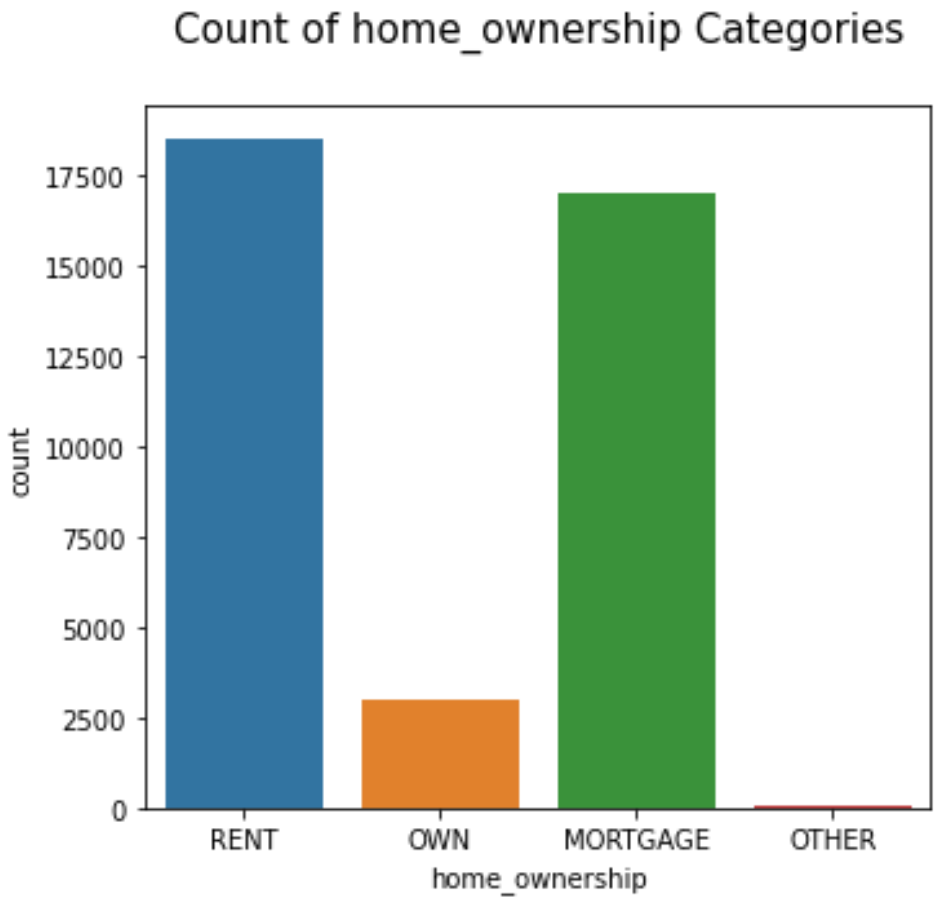


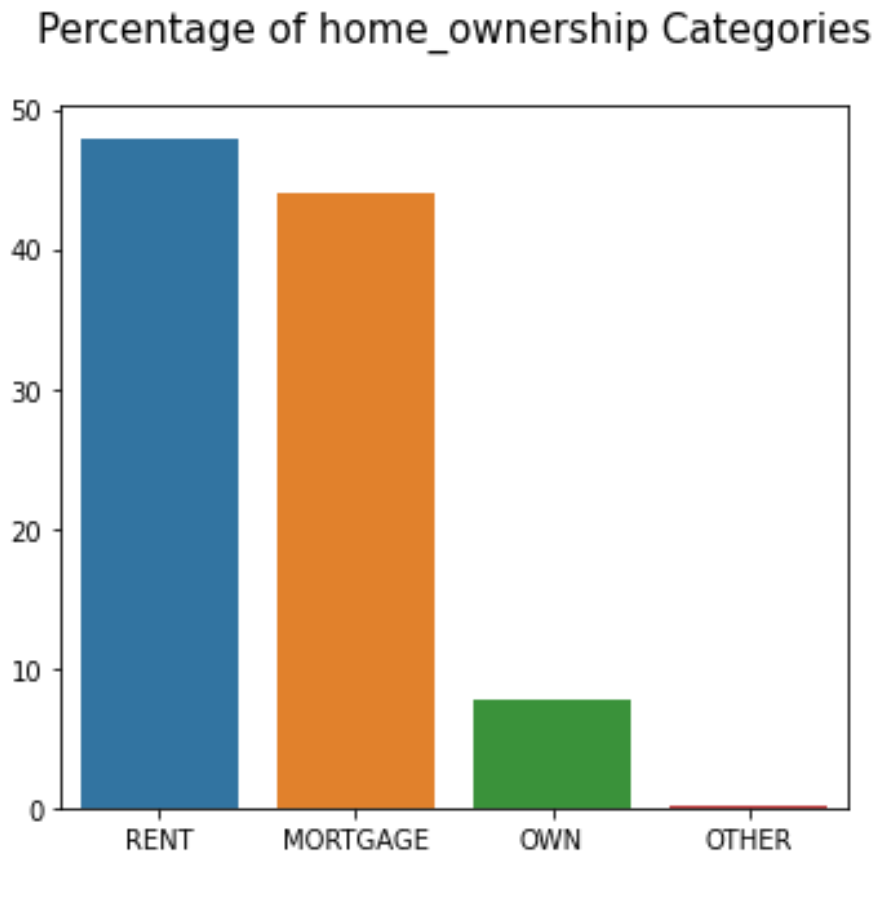
Inference:

1) Their is no general pattern of employee length with loan status.

1) Borrowers whose employee length is unknown are most risky as their default rate is highest(~22%)

5) Home Onwership

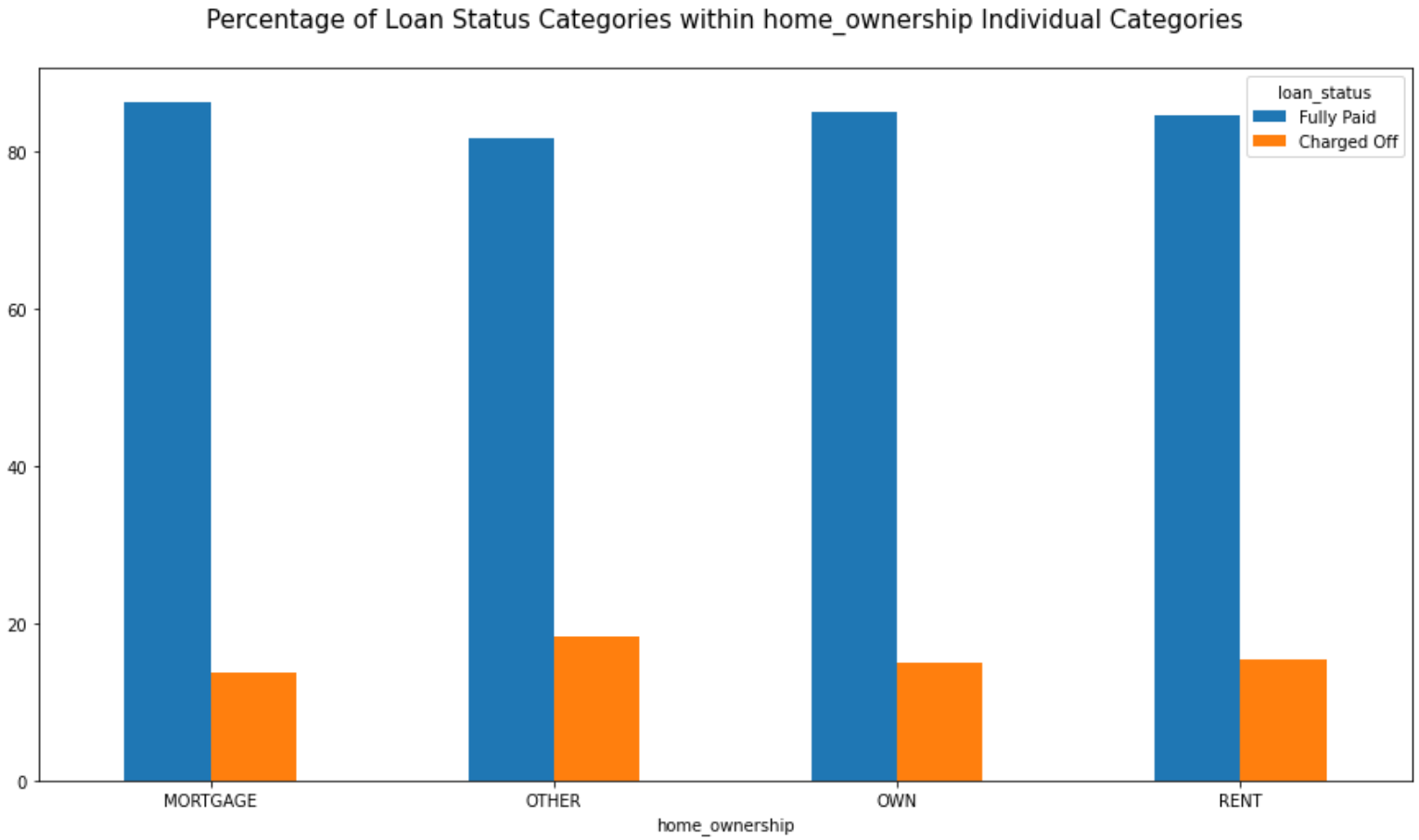




Inference:

1) 'Rent' type of home_ownership has maximum datapoints 18483(~48%),
   followed by MORTGAGE 17021 datapoints(~44%)
2) 'OTHER' type of home_ownership has least datapoints 98(~0.25%)

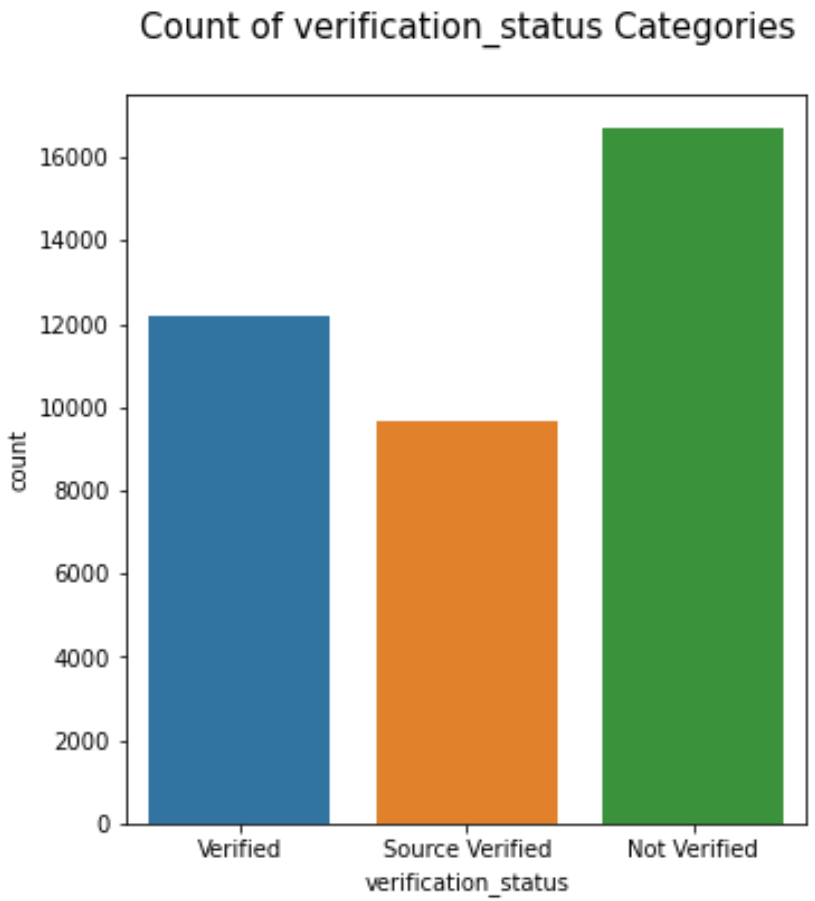
Relation of home ownership with loan status

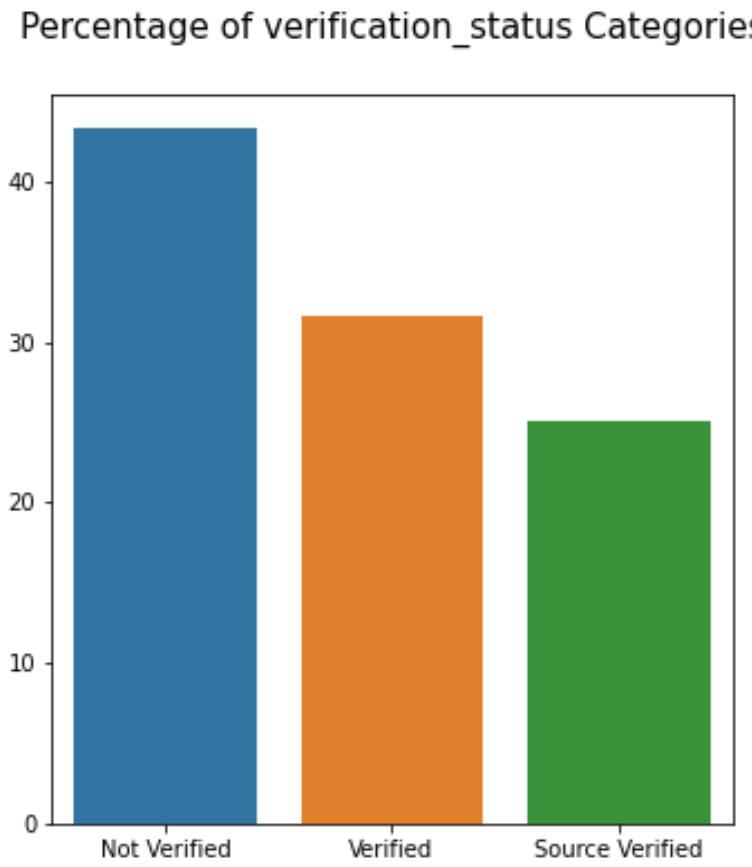


Inference:

1) OTHER category has highest default rate of ~18%

6) Verification Status

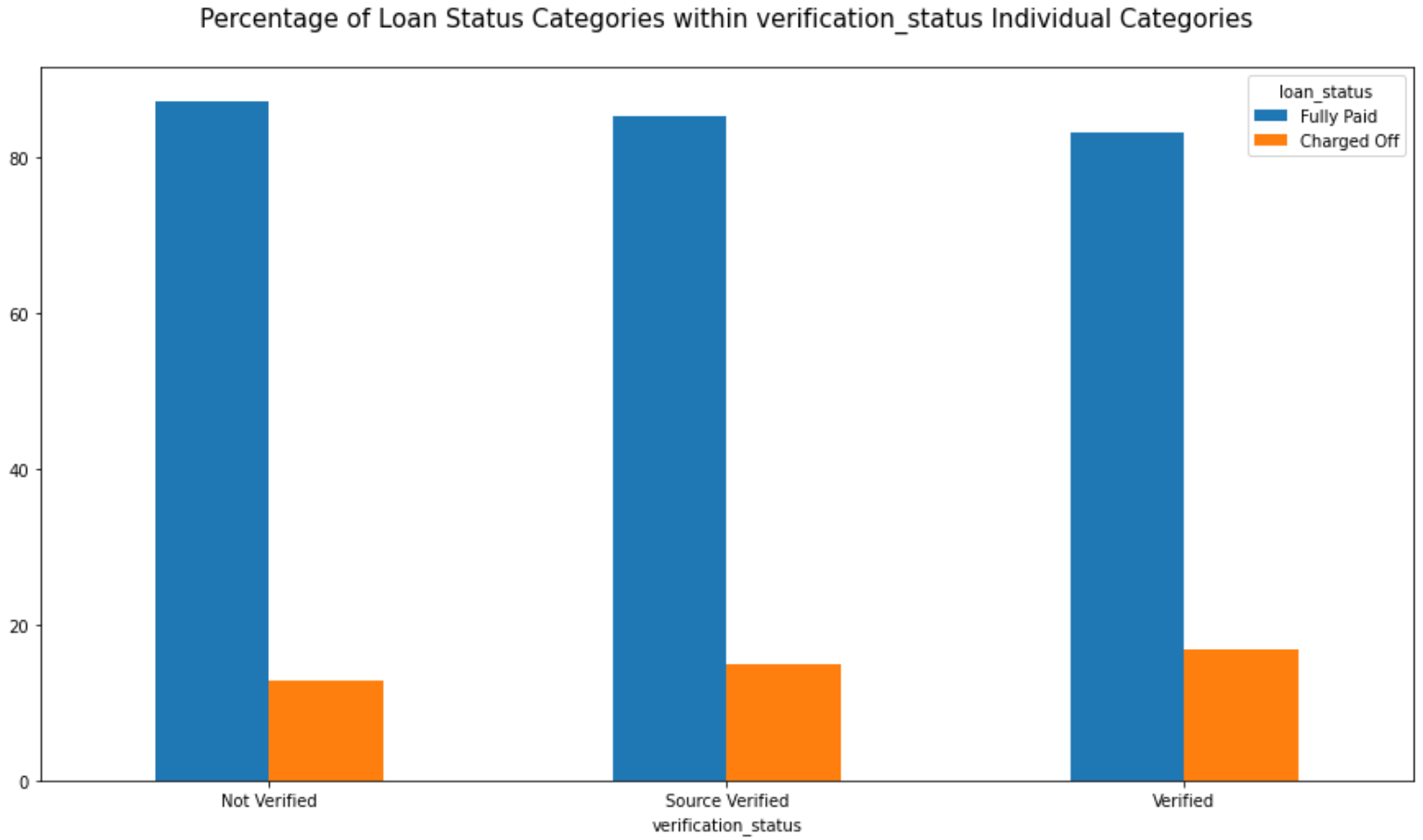




Inference:

1) Verification Status column consists of 3 unique categories, Verified, Source Verified, Not Verified namely
2) The count of Not Verified is highest with 16694(~43%) datapoints, followed by Verified with a count of 12206(~32%).
3) Source Verified is the lowest category in terms of count of datapoints; i.e 9677(~ 25%)

Relation of Verification Status with loan status



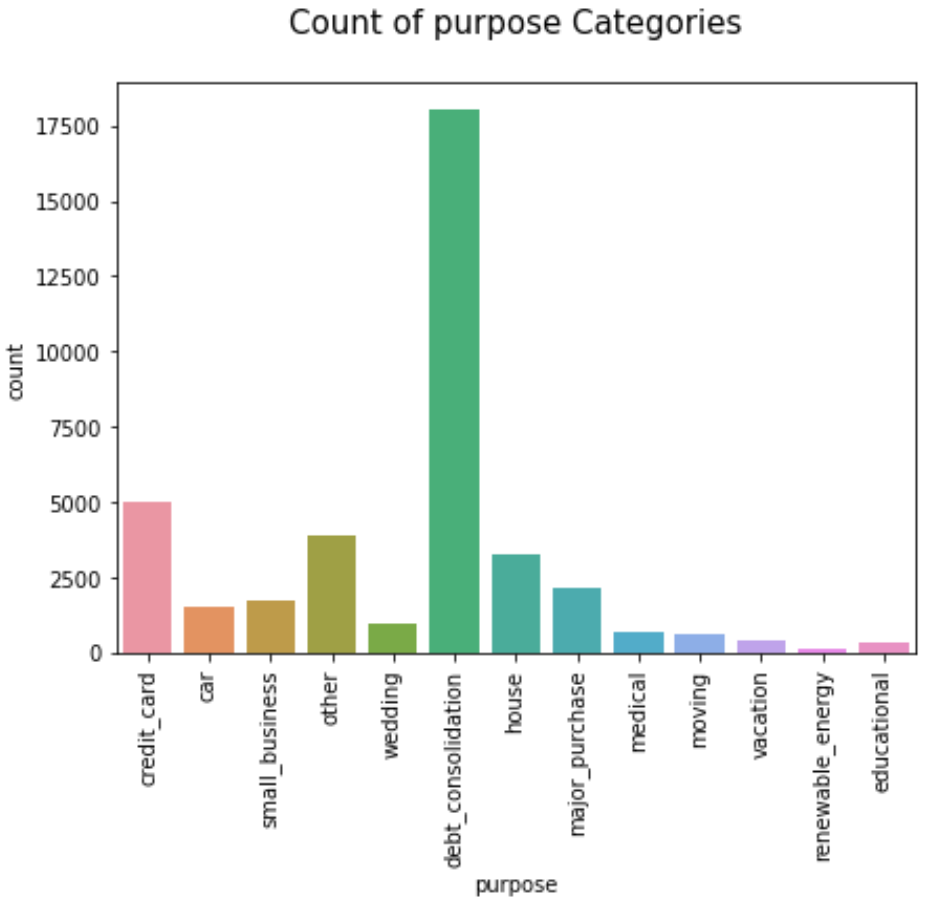
Inference:

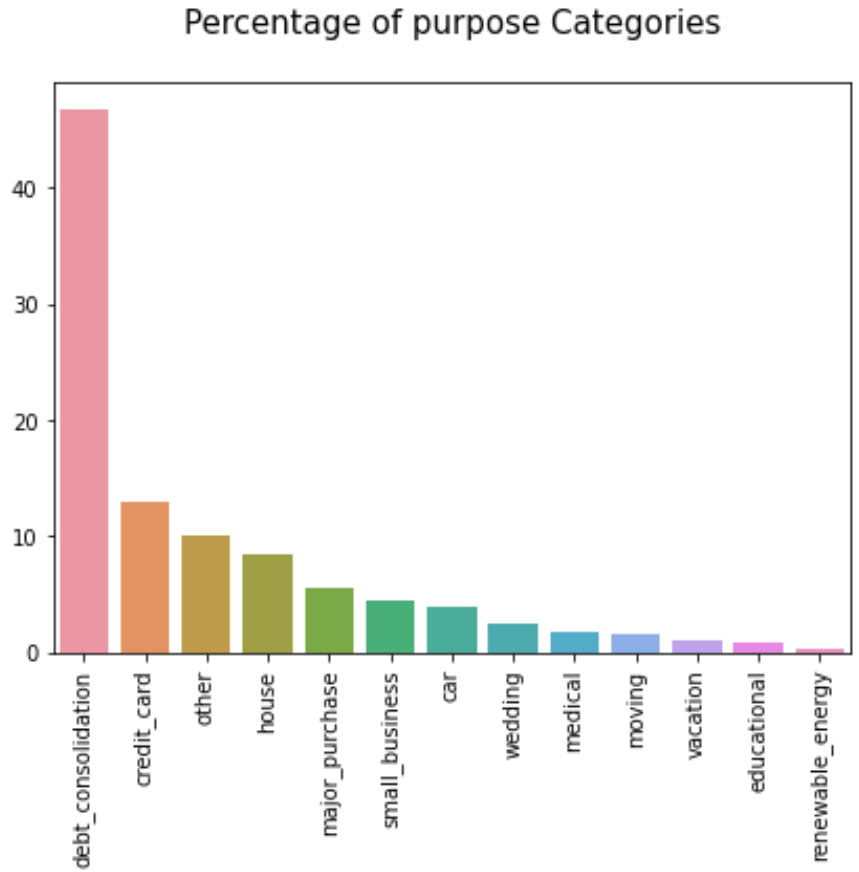
1) The default rate of Verified is highest(~17%)

2) The default rate of Not Verified is lowest(~13%)

3) The default rate of Source Verified is ~15%

7) Purpose





Inference:

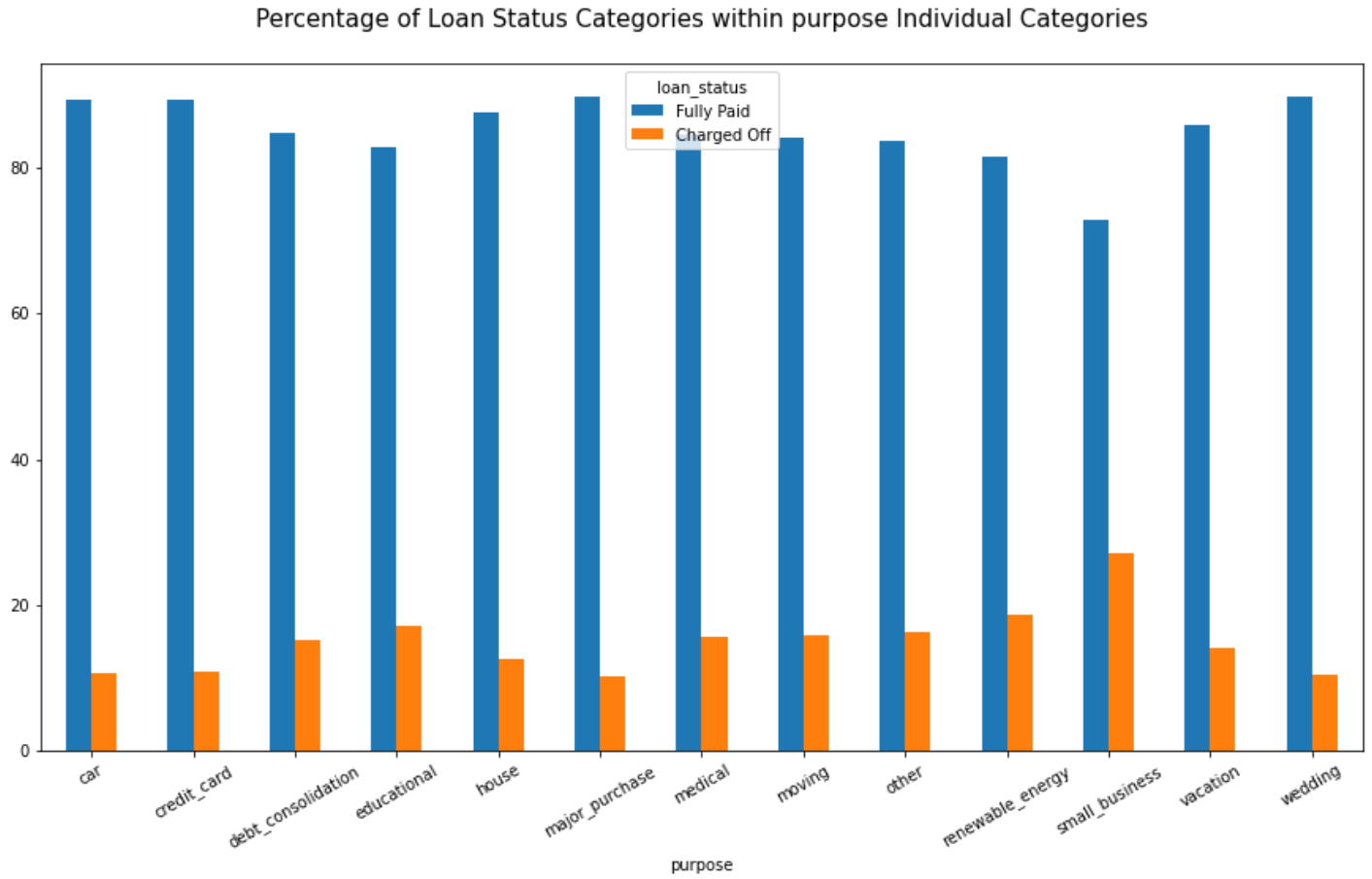
1) Purpose column consists of 13 unique categories

2) The count of debt_consolidation is highest with 18055(~47%) datapoints,

   followed by credit_card with a count of 5027(~13%).

3) Renewable_energy is the lowest category in terms of count of datapoints; i.e 102(~0.26%)

Relation of purpose with loan status



Inference:

1) The probability of default is highest for small_business(~27%),followed by renewable_energy(~18.6%)

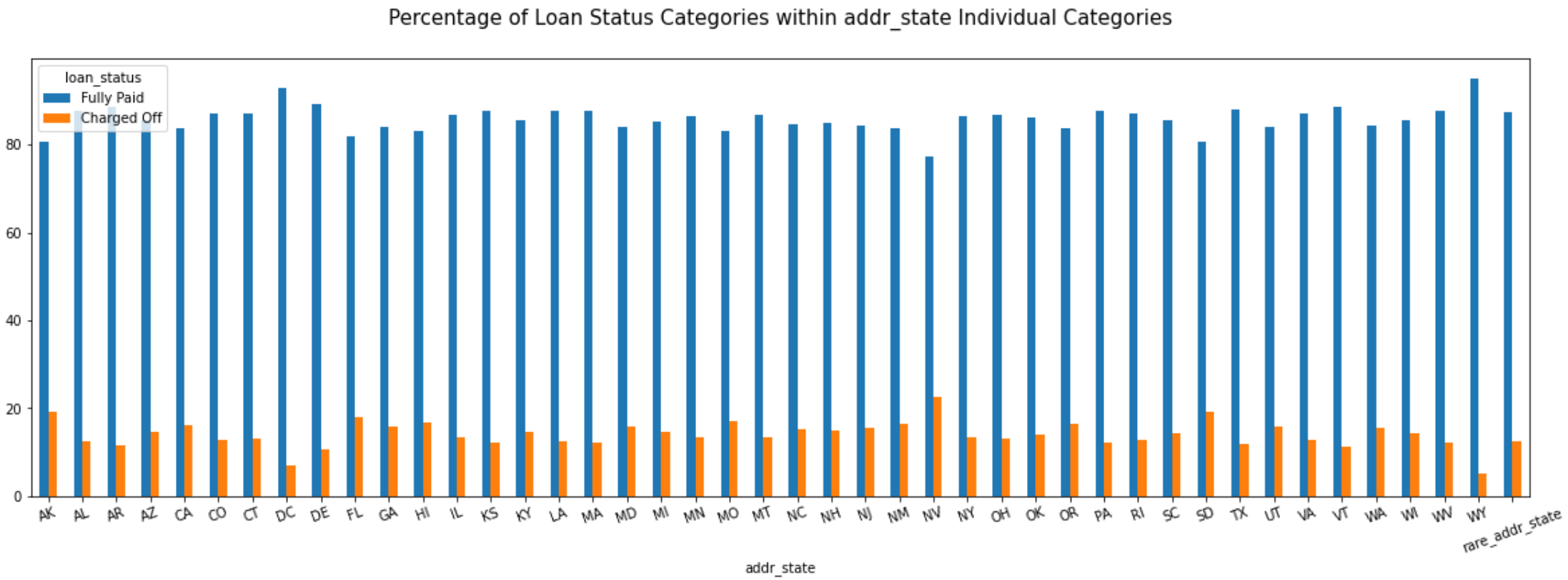
2) The probability of default is lowest for major_purchase(~10.3%), wedding(~10.4%)

8) Address State(addr_state)

- Replaces datapoints where addr_state is in list [MS, TN, IN, ID, NE, IA, ME] value with a new value of 'rare_addr_state' as the individual count of these values is very low.

- CA has highest datapoints(6949), followed by NY(3698), followed by FL(2781)

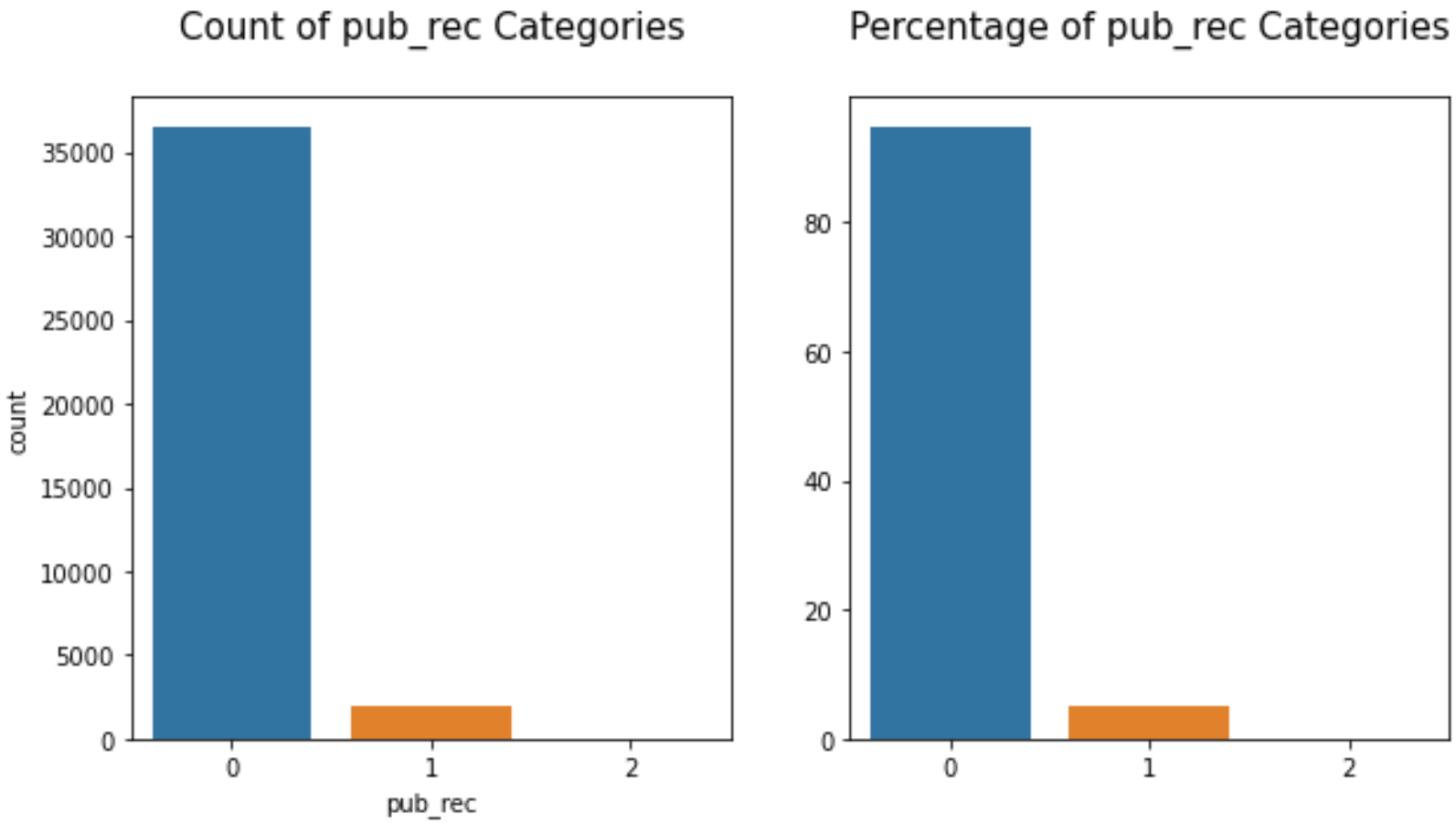
Relation of address state with loan status



Inference:

1) The address NV has highest default rate of ~22.5% followed by SD with ~19.35% followed by AK with a default rate of 19.2%
2) WY has the lowest default rate of 5%

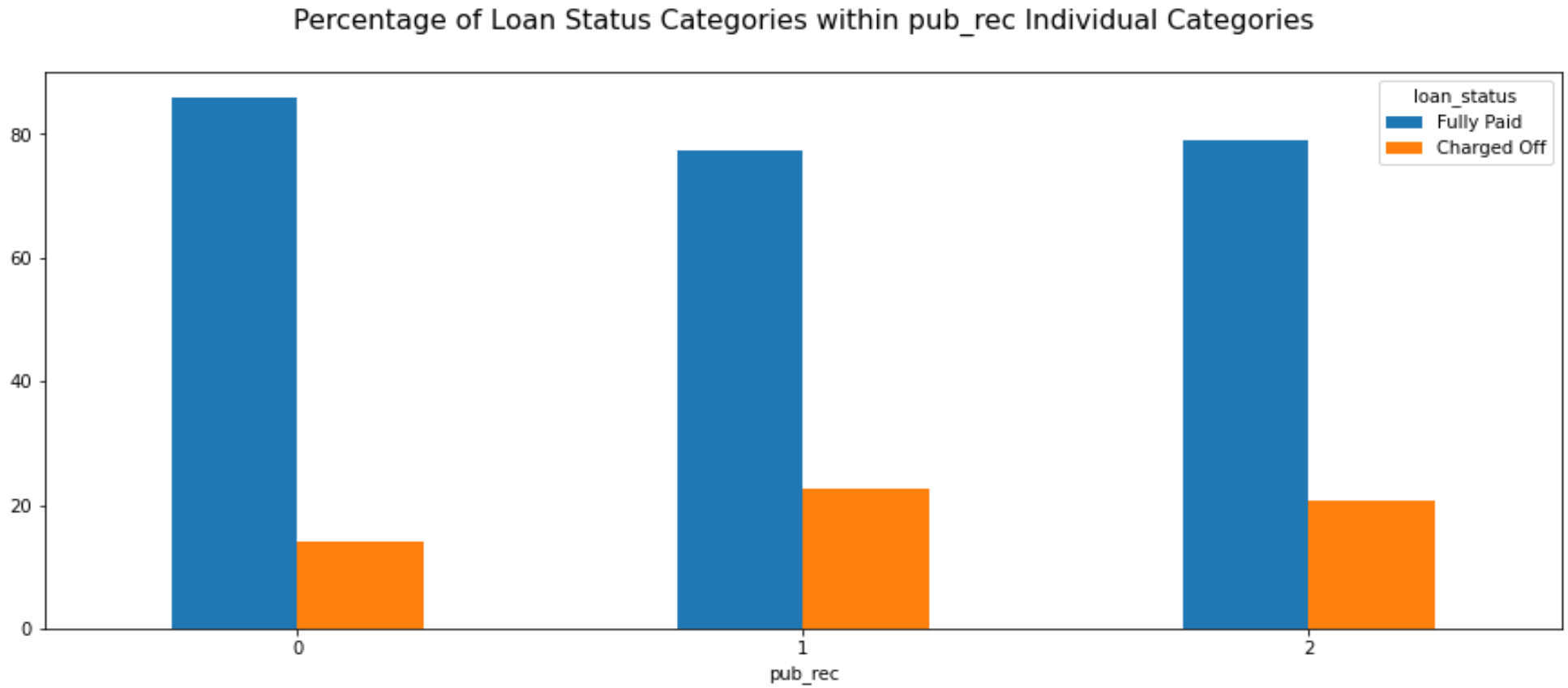
Number of derogatory public records(pub_rec)



Inference:

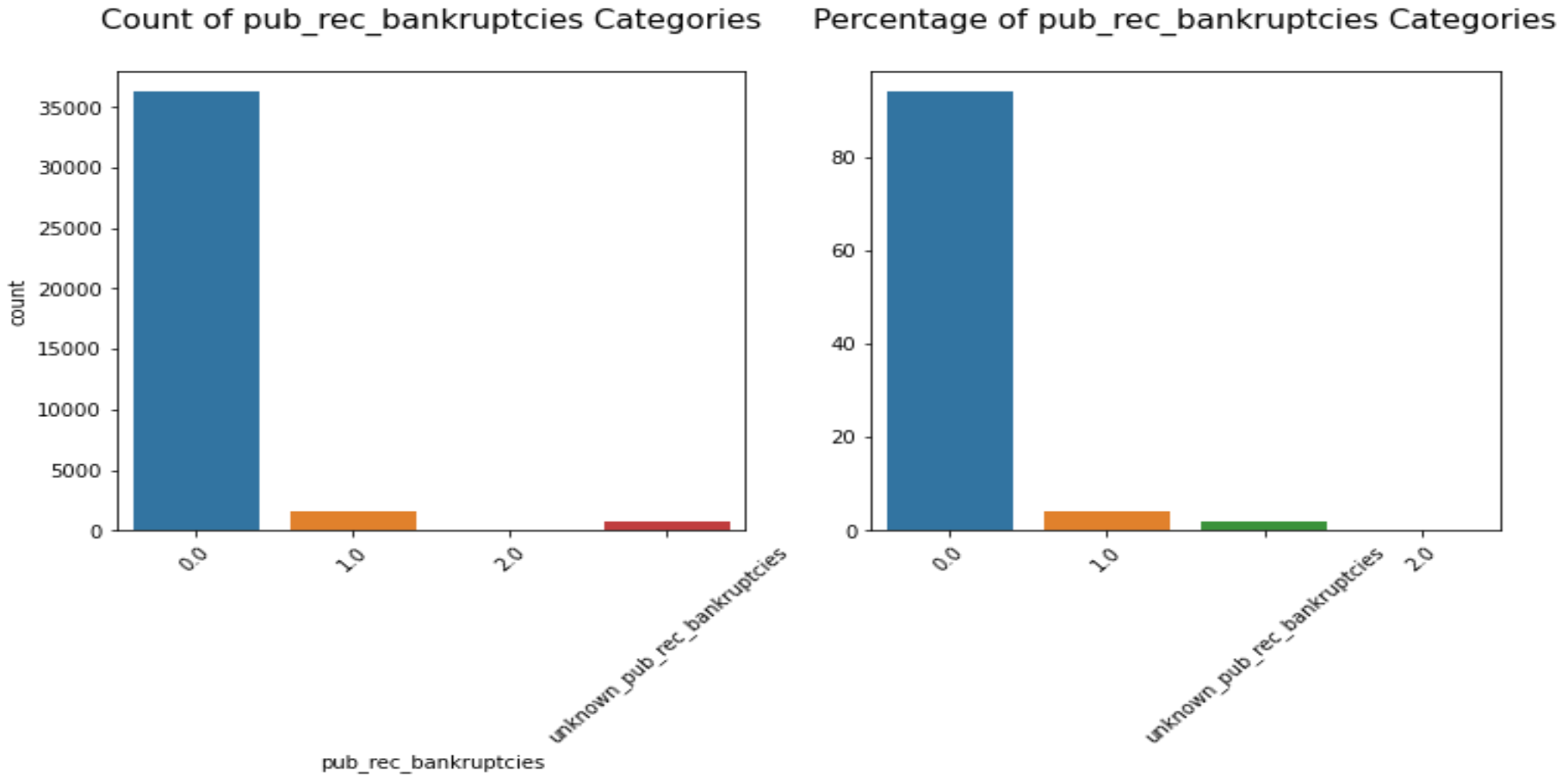
1) Customers who have 0 derogatory public records are maximum in number 36516(~95%)

Relation of derogatory public records with loan status



Inference: Customers with known derogatory public records more prone for default

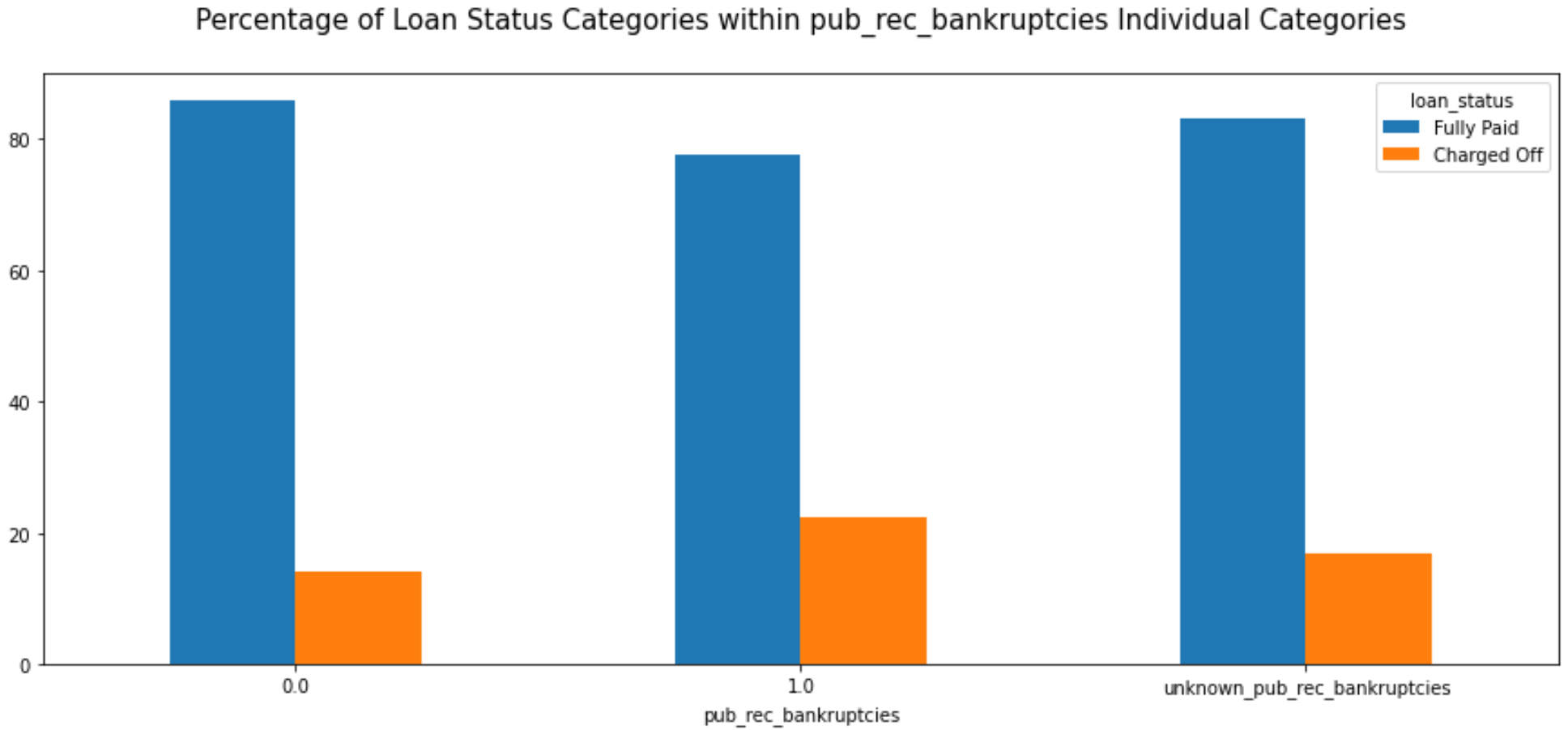
Number of public record bankruptcies (pub_rec_bankruptcies)



Inference:

1) Customers who have 0 public record of bankruptcies are maximum in number 36238(~94%)

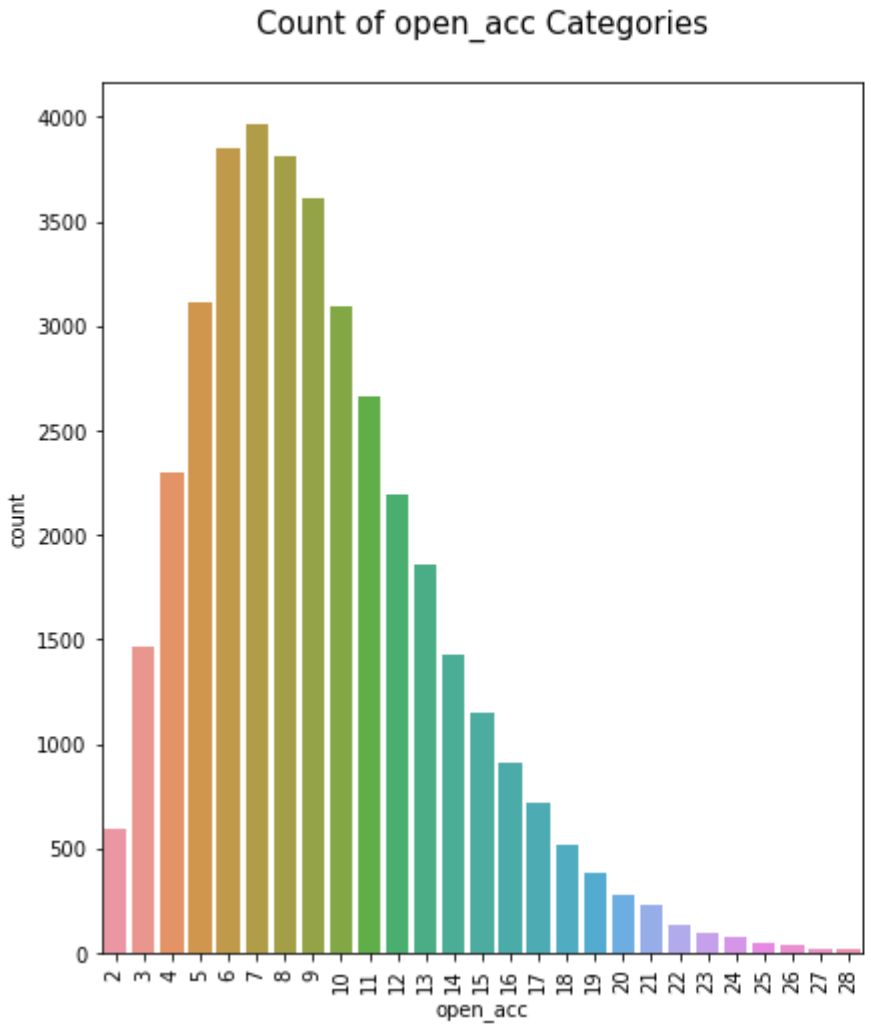
Relation of Number of public record bankruptcies with loan status

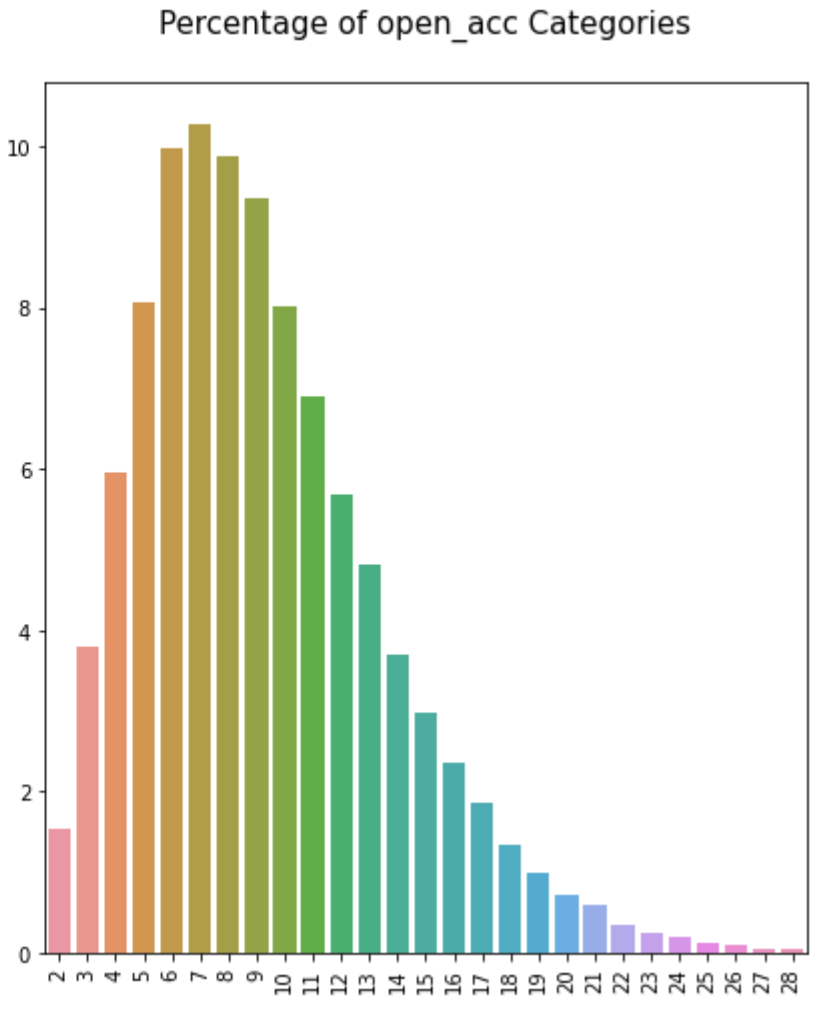


Inference: Customers with known public bankruptcies more prone for default

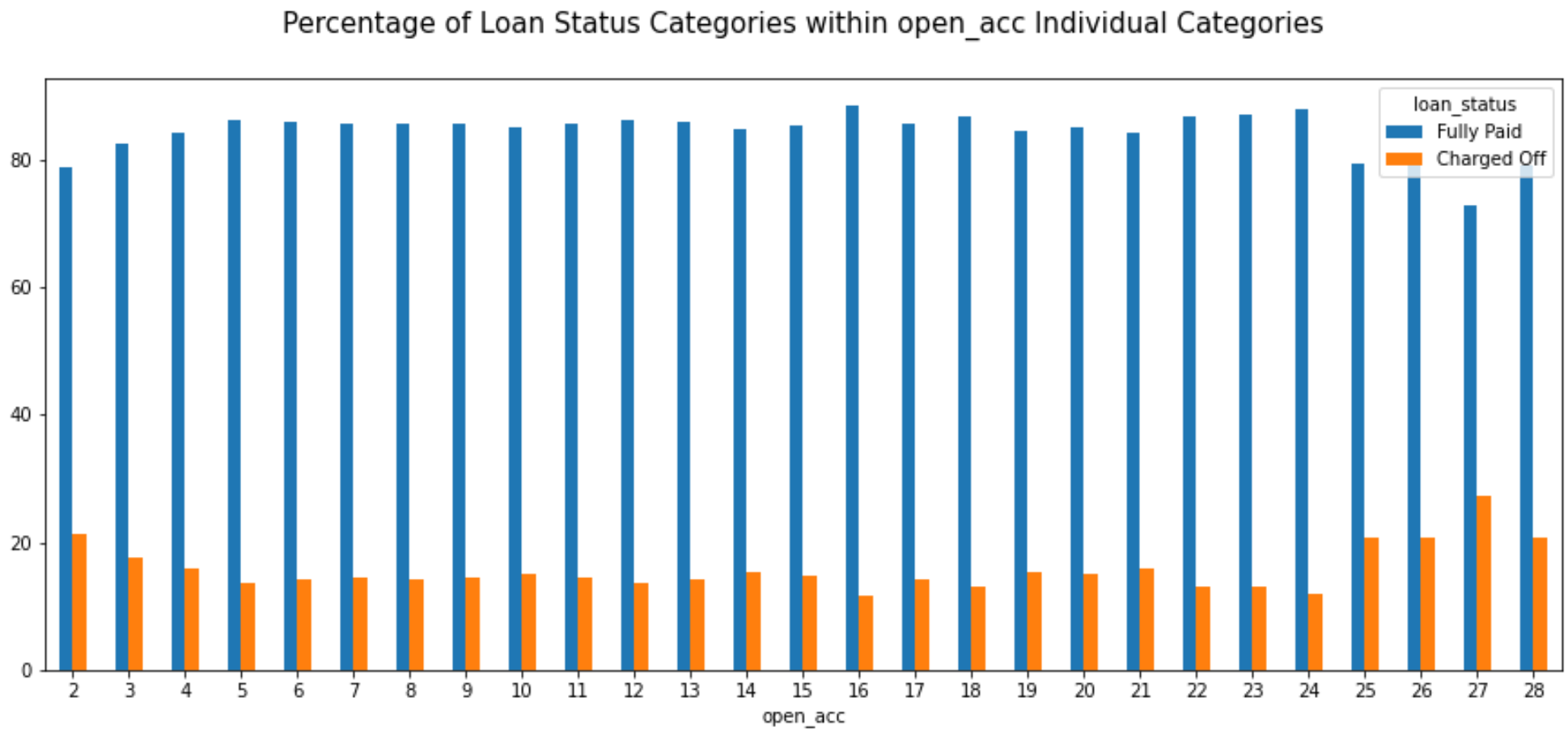
Number of open credit lines(open_acc)

- The open_acc column when it takes a value from [30,29,31,34,35,32,33,36,42,41,39,38,44] has a very low count , so replacing these values with the mode of the column





Relation of open credit lines with loan status



Inference:

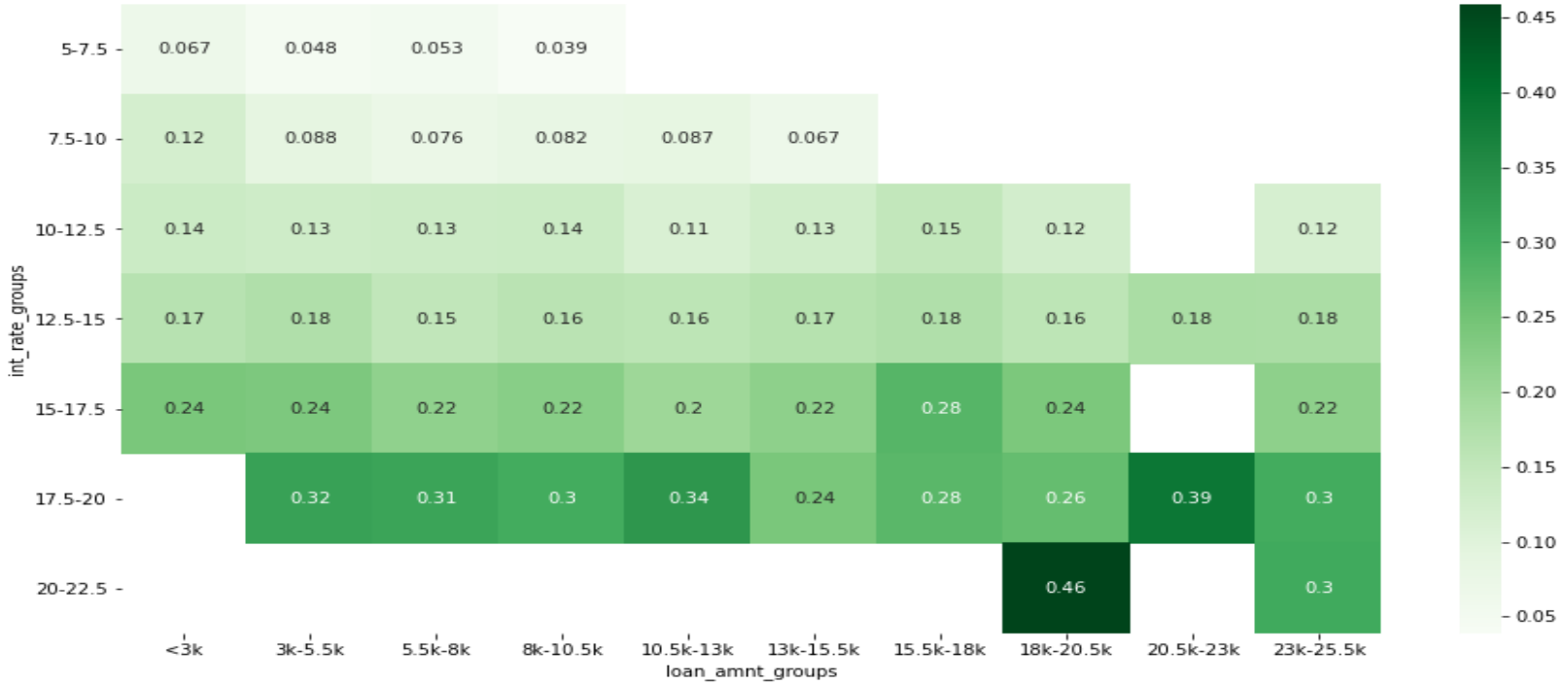
1) Their is no general pattern/ relationship of number of open accounts with loan status

2) The default rate is higher than 20% when number of open accounts exceed 20

Multivariate Analysis

Two Continuous variables with Loan Status

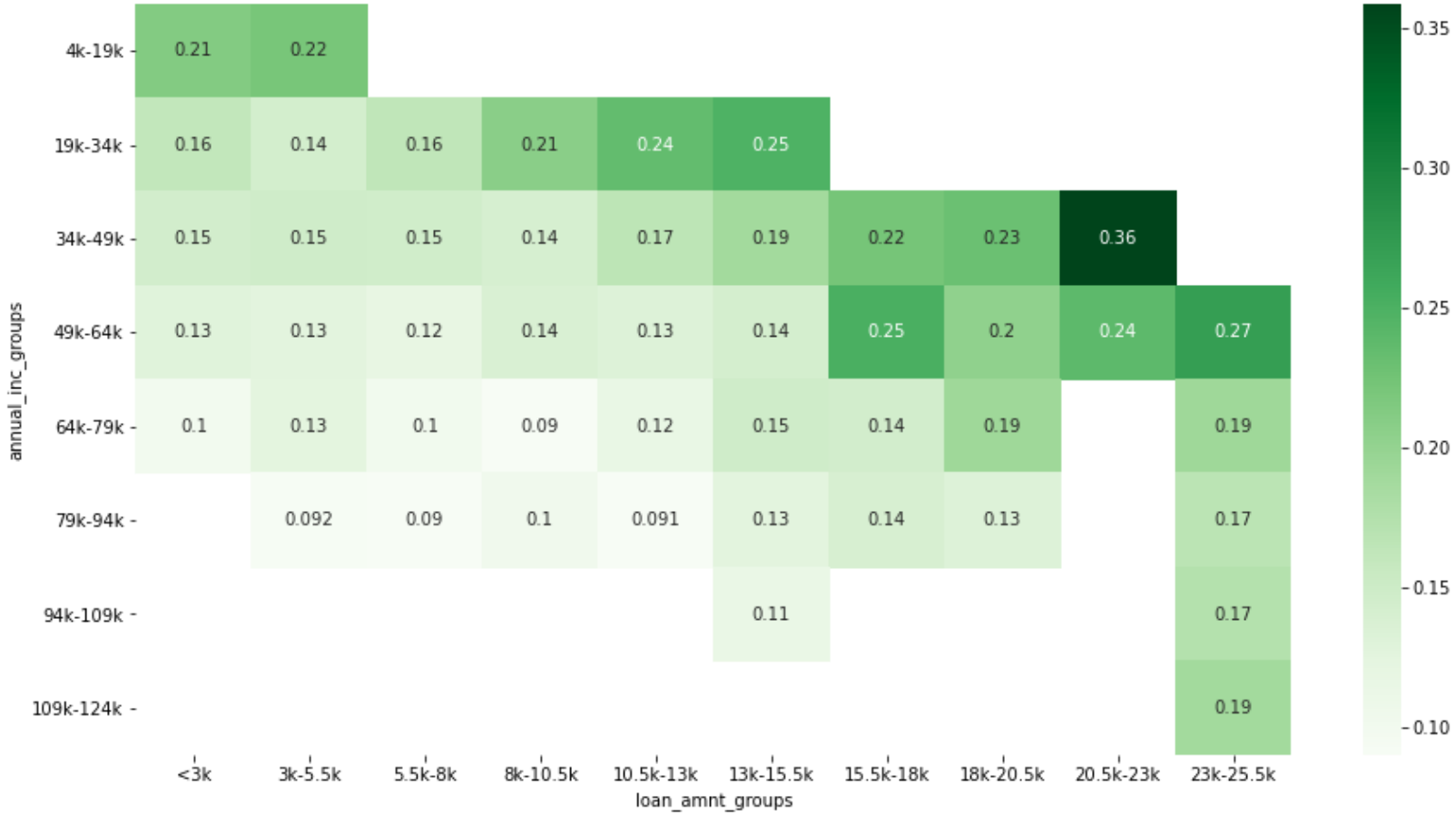
1) Loan Amount and Interest Rate with Loan Status



Inference:

1) The bivariate analysis of loan amount with loan status showed us that with increasing loan amount the default rate increases. Same analysis is shown in this graph also
2) The bivariate analysis of interest rate with loan status showed us that with interest rate the default rate increases. Same analysis is shown in this graph also.
3) Customers having any loan amount but with interest rate greater than 15% are much more prone to default

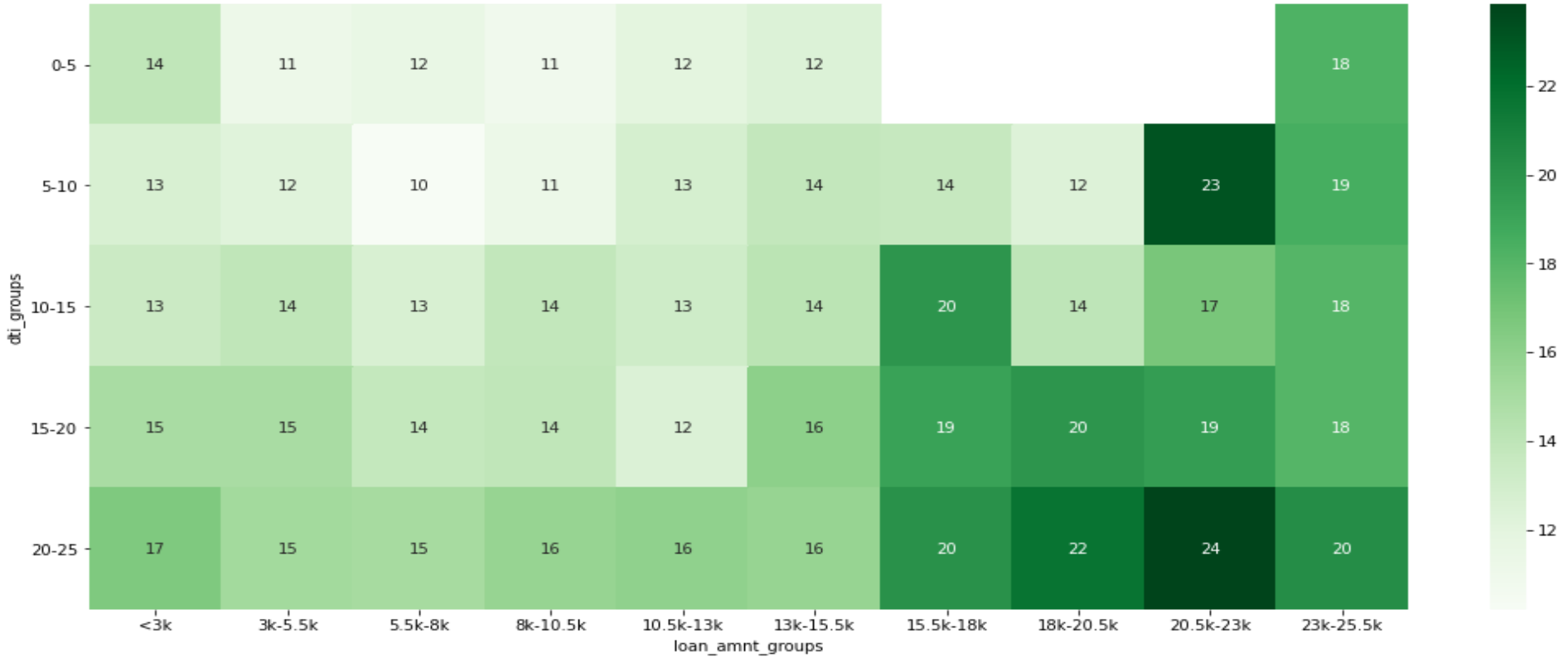
2) Loan Amount and Annual Income with Loan Status



Inference:

1) The bivariate analysis of loan amount with loan status showed us that with increasing loan amount the default rate increases. Same analysis is shown in this graph also.
2) The bivariate analysis of annual income with loan status showed us that with increasing annual income the default rate decreases. Same analysis is shown in this graph also.
3) Customers having loan amount greater than 10.5K and annual income less than 64k have more prone for default

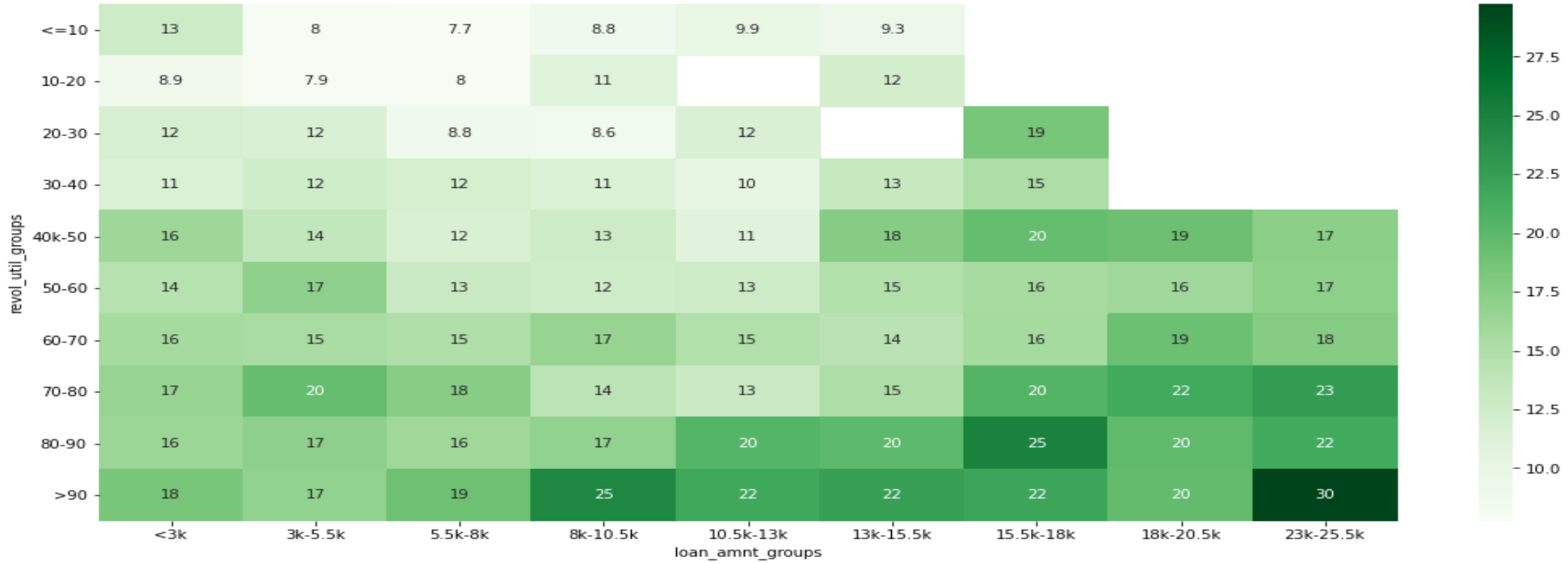
3) Loan Amount and Debt to Income Ratio with Loan Status



Inference:

1) The bivariate analysis of loan amount with loan status showed us that with increasing loan amount the default rate increases. Same analysis is shown in this graph also.
2) The bivariate analysis of dti with loan status showed us that with increasing dti the default rate increases. Same analysis is shown in this graph also.
3) Customers having loan amount greater than 15.5K and dti greater than 10 have more prone for default

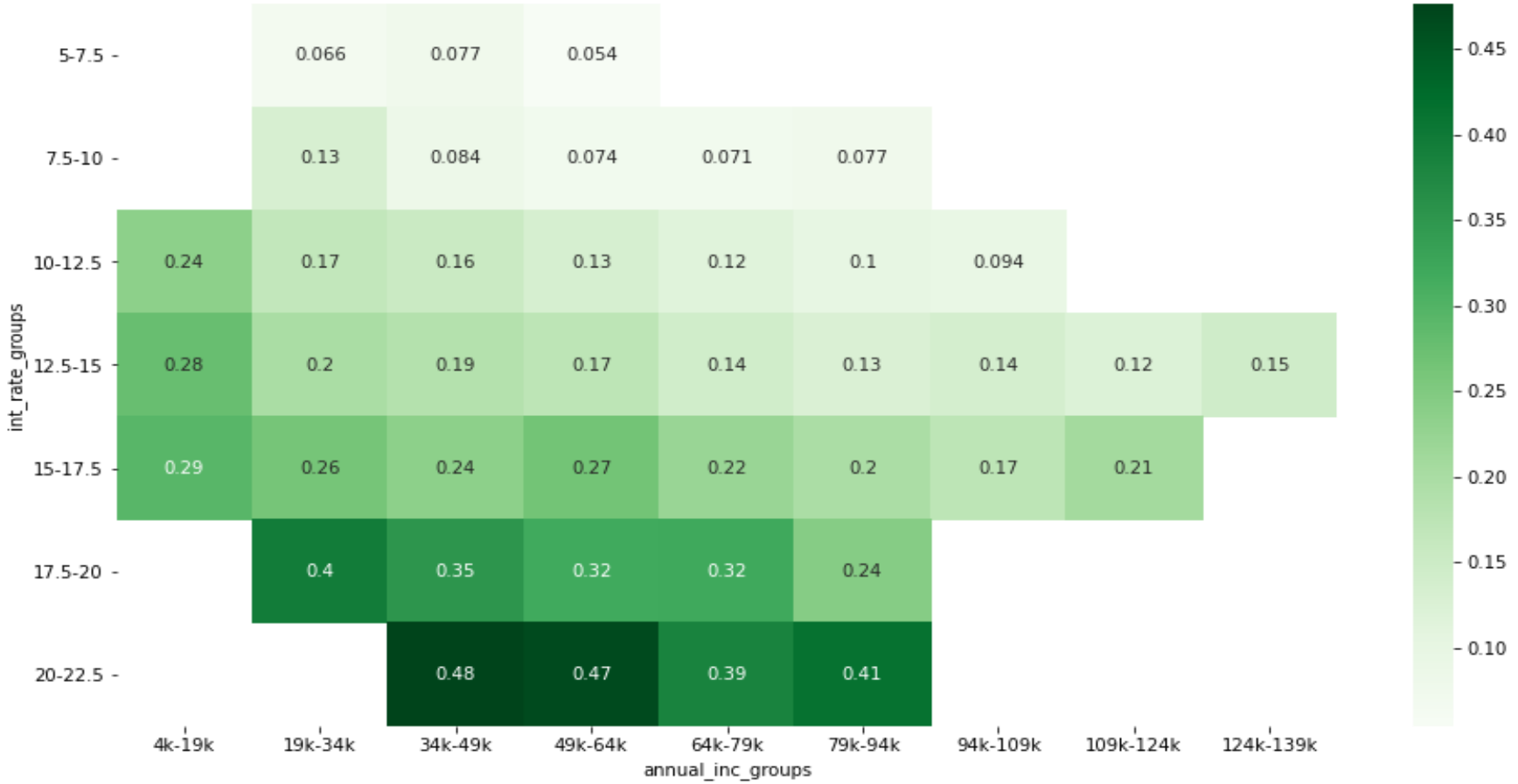
4) Loan Amount and revolving line utilization rate with Loan Status



Inference:

1) The bivariate analysis of loan amount with loan status showed us that with increasing loan amount the default rate increases. Same analysis is shown in this graph also.
2) The bivariate analysis of revolving rate income with loan status showed us that with increasing revolving rate the default rate increases. Same analysis is shown in this graph also.
3) Customers having loan amount greater than 10.5K and revolving rate greater than 70 have more prone for default

5) Annual Income and Interest Rate with Loan Status

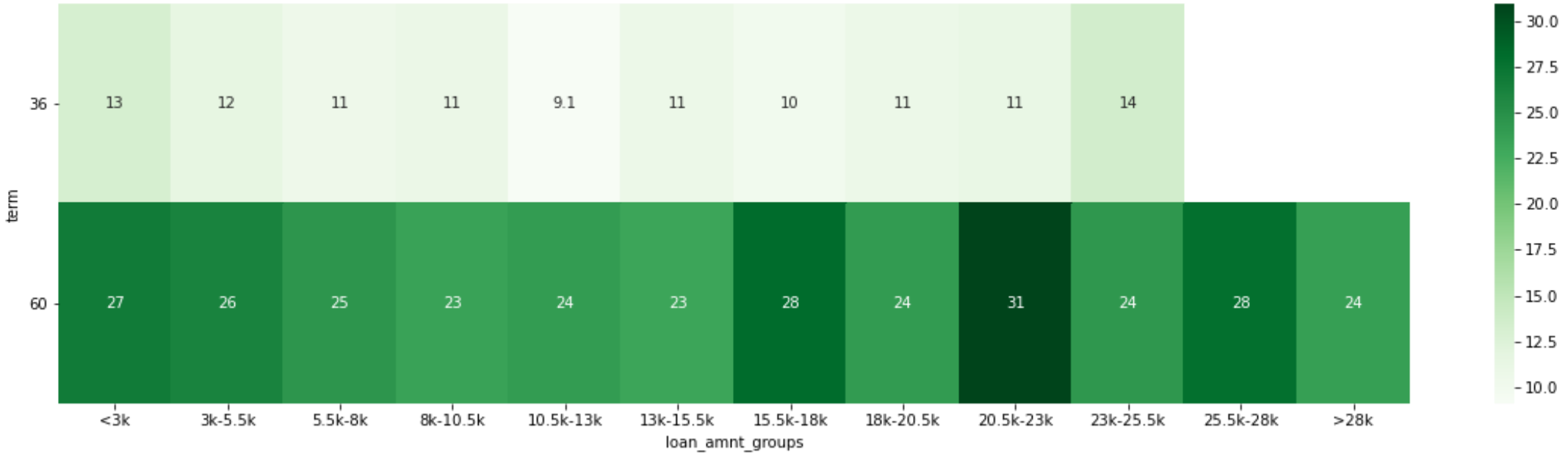


Inference:

1) The bivariate analysis of annual income with loan status showed us that with increasing annual income the default rate decreases. Same analysis is shown in this graph also.
2) The bivariate analysis of interest rate with loan status showed us that with interest rate the default rate increases. Same analysis is shown in this graph also.
3) Customers having annual income less than 94k and interest rate greater than 15 have more prone for default

1 Continuous + 1 Categorical with Loan Status

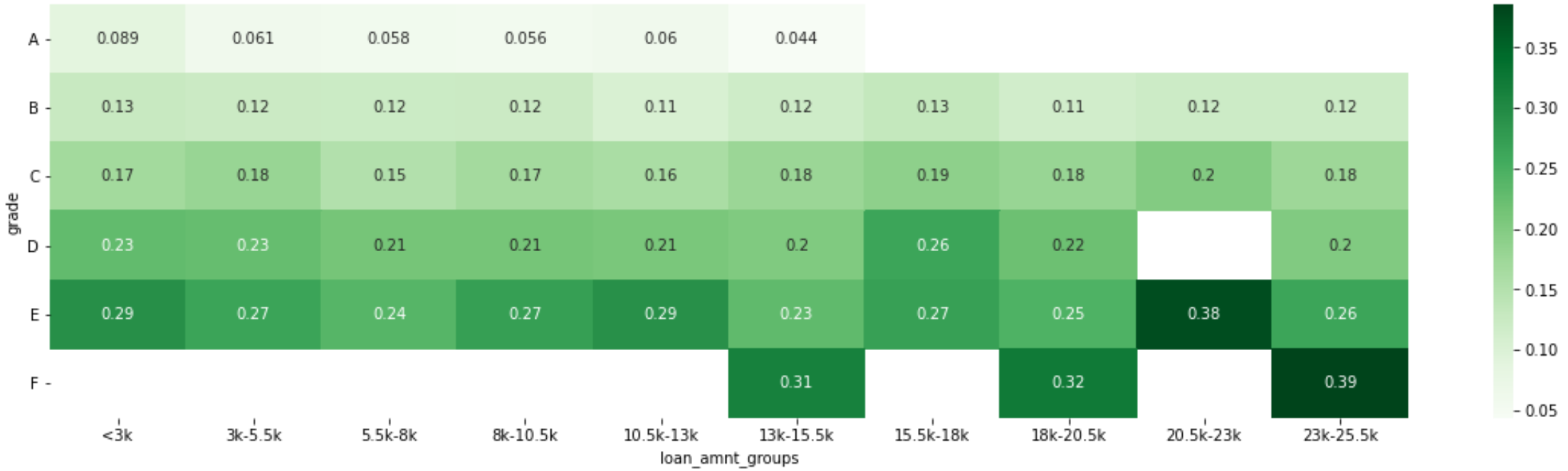
6) Loan Amount and term with loan status



Inference:

1) The bivariate analysis of loan amount with loan status showed us that with increasing loan amount the default rate increases. Same analysis is shown in this graph also.
2) The bivariate analysis of term with loan status showed us that customers who take loan for 60 months are more prone to deafult. Same analysis is shown in this graph also.
3) Customers taking any but for term of 60 months are more prone for default than customers taking the same amount for 30 months.

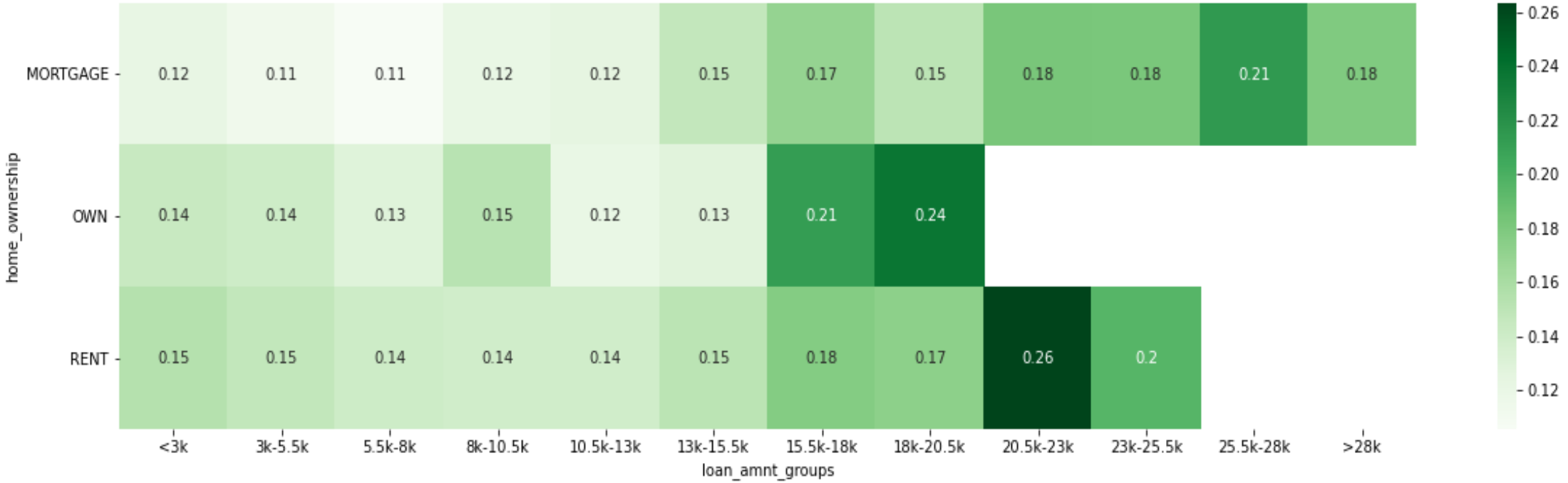
7) Loan Amount and grade with loan status



Inference:

1) The bivariate analysis of loan amount with loan status showed us that with increasing loan amount the default rate increases. Same analysis is shown in this graph also.
2) The bivariate analysis of grade with loan status showed us that with increasing alphabetical order of grades the default rate also increases. Same analysis is shown in this graph also.
3) Customers taking any but beloning to grade C, D, E or F are more prone for default.

8) Loan amount and home ownership with loan status



Inference:

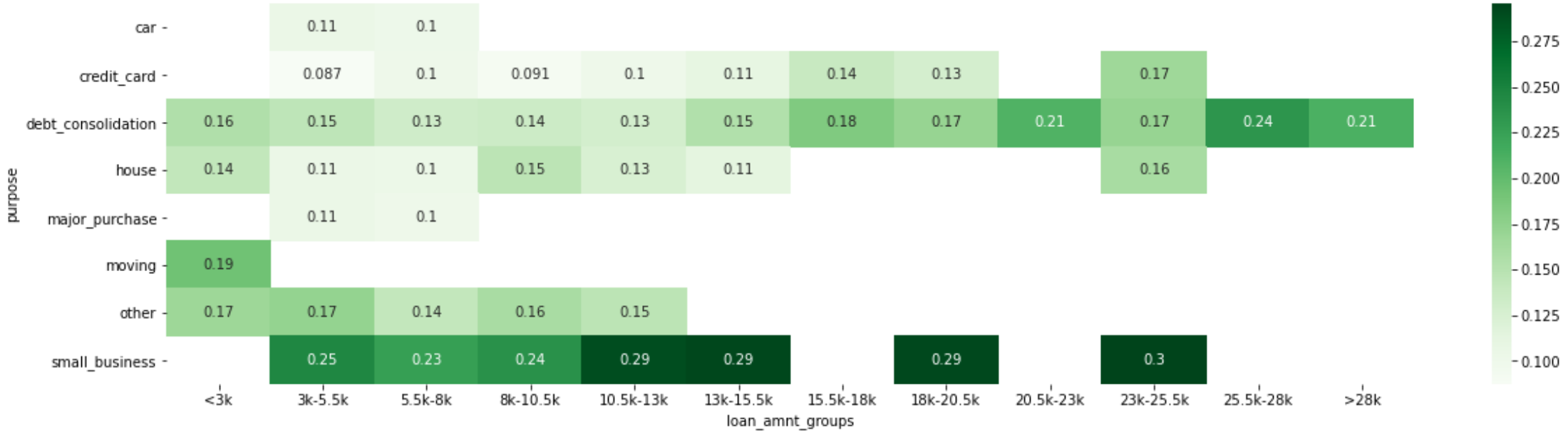
1) The bivariate analysis of loan amount with loan status showed us that with increasing loan amount the default rate increases. Same analysis is shown in this graph also.
2) The bivariate analysis of home onwership with loan status showed us no clear relationship. Same analysis is shown in this graph also.

9) Loan amount and verification status with loan status

Inference:

1) The bivariate analysis of loan amount with loan status showed us that with increasing loan amount the default rate increases. Same analysis is shown in this graph also
2) The bivariate analysis of verification status with loan status showed us no clear relationship. Same analysis is shown in this graph also.

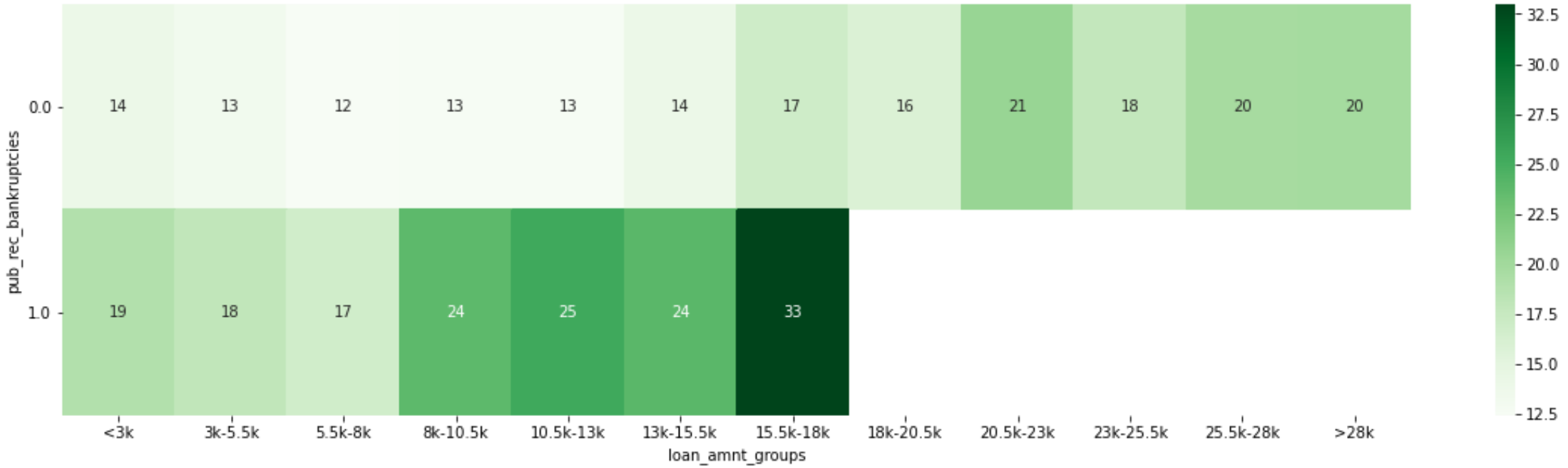
10) Loan amount and purpose with loan status



Inference:

1) The bivariate analysis of loan amount with loan status showed us that with increasing loan amount the default rate increases. Same analysis is shown in this graph also.
2) The bivariate analysis of purpose with loan status showed us that when purpose was small business the default risk was highest. Same analysis is shown in this graph also.
3) Customers taking any loan amount for small business purpose are most prone for default

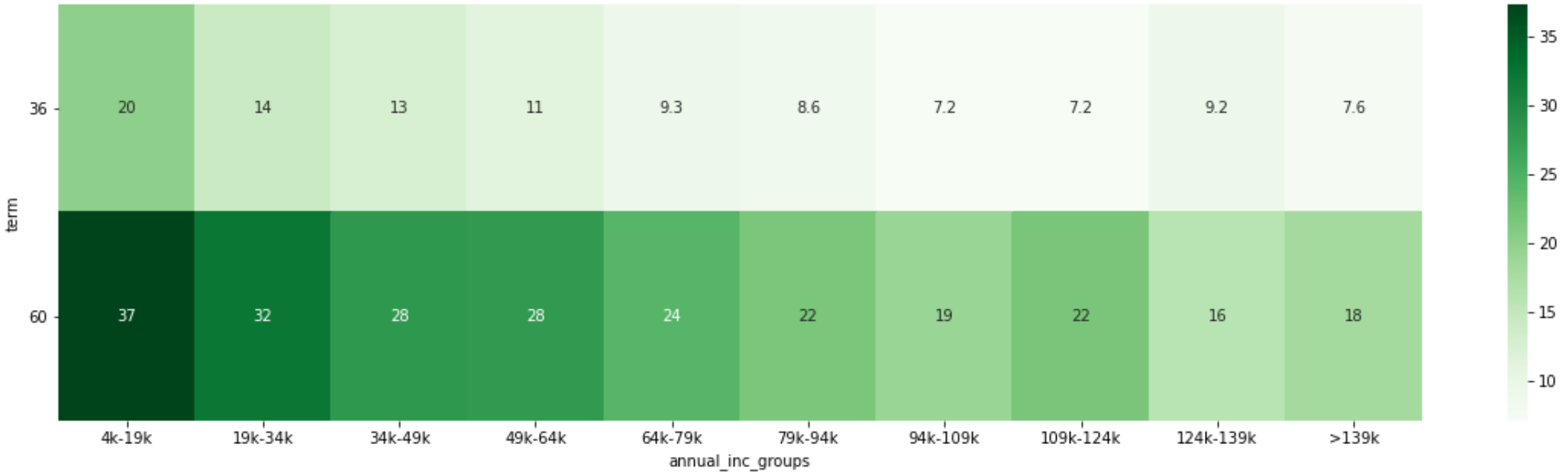
11) loan amount and number of public record bankruptcies with loan status



Inference:

1) The bivariate analysis of loan amount with loan status showed us that with increasing loan amount the default rate increases. Same analysis is shown in this graph also.
2) The bivariate analysis of pub_rec_bankruptcies with loan status showed us that customers having atleast one public record of bankruptcy are more for default.
   Same analysis is shown in this graph also.
3) Customers taking any loan amount but having a public record of bankruptcy are more prone for default.

12) annual income and term with loan status



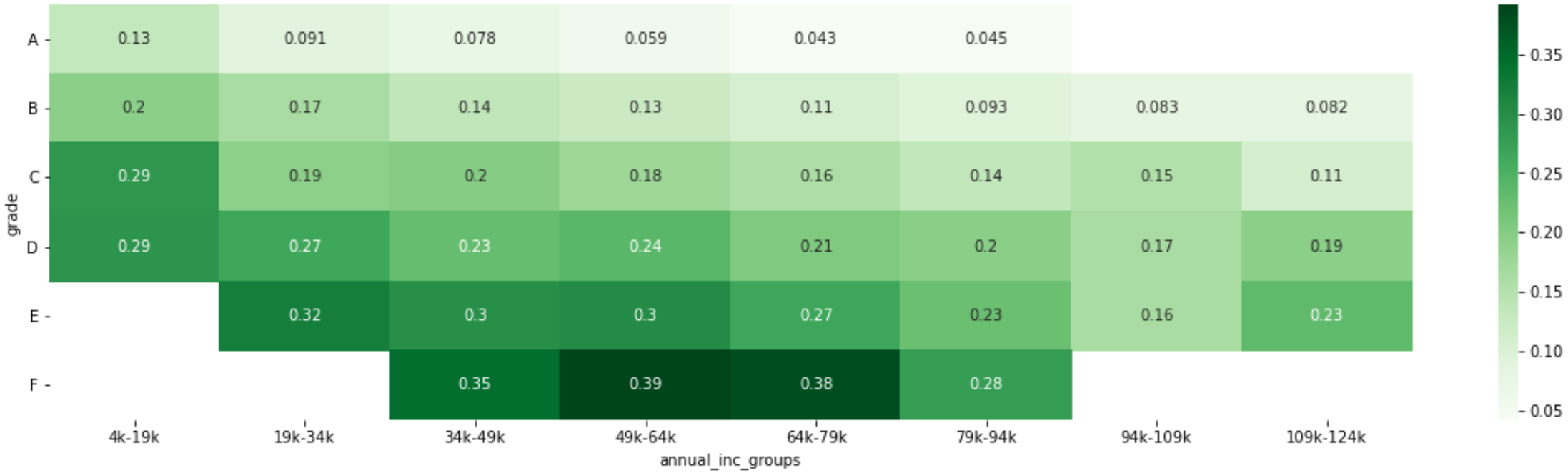
Inference:

1) The bivariate analysis of annual income with loan status showed us that with increasing annual income the default rate decreases. Same analysis is shown in this graph also.

2) The bivariate analysis of term with loan status showed us that customers who take loan for 60 months are more prone to deafult. Same analysis is shown in this graph also.

3) Customers having any annual income but taking loan for a term of 60 months are more prone for default.

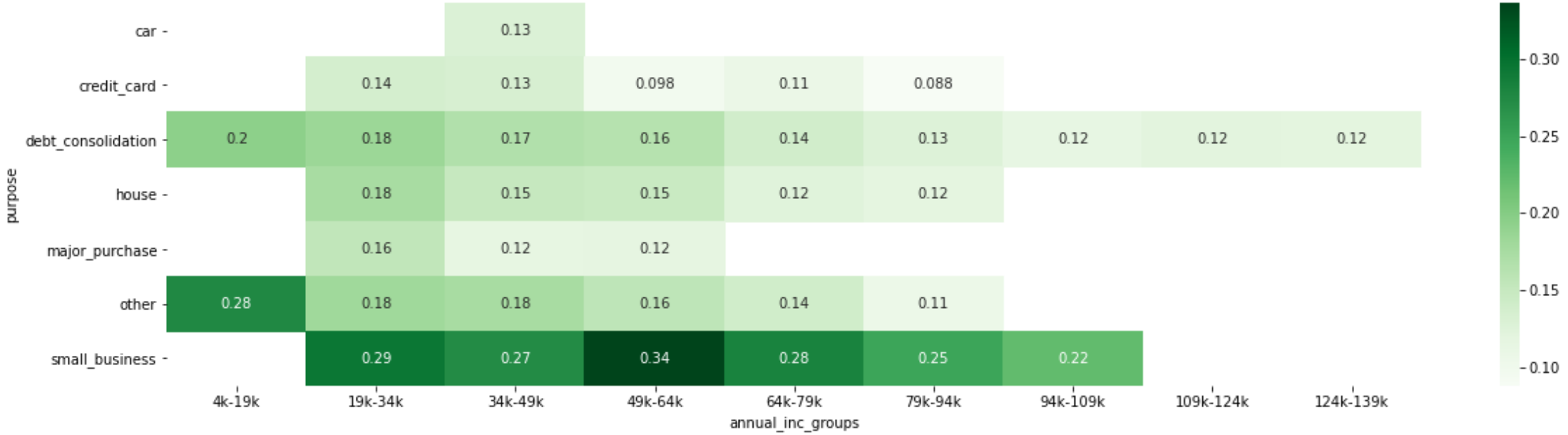
13) annual income and grade with loan status



Inference:

1) The bivariate analysis of annual income with loan status showed us that with increasing annual income the default rate decreases. Same analysis is shown in this graph also.
2) The bivariate analysis of grade with loan status showed us that with increasing alphabetical order of grades the default rate also increases. Same analysis is shown in this graph also.
3) Customers having annual income less than 79k and beloning to grade C, D, E or F are more prone for default

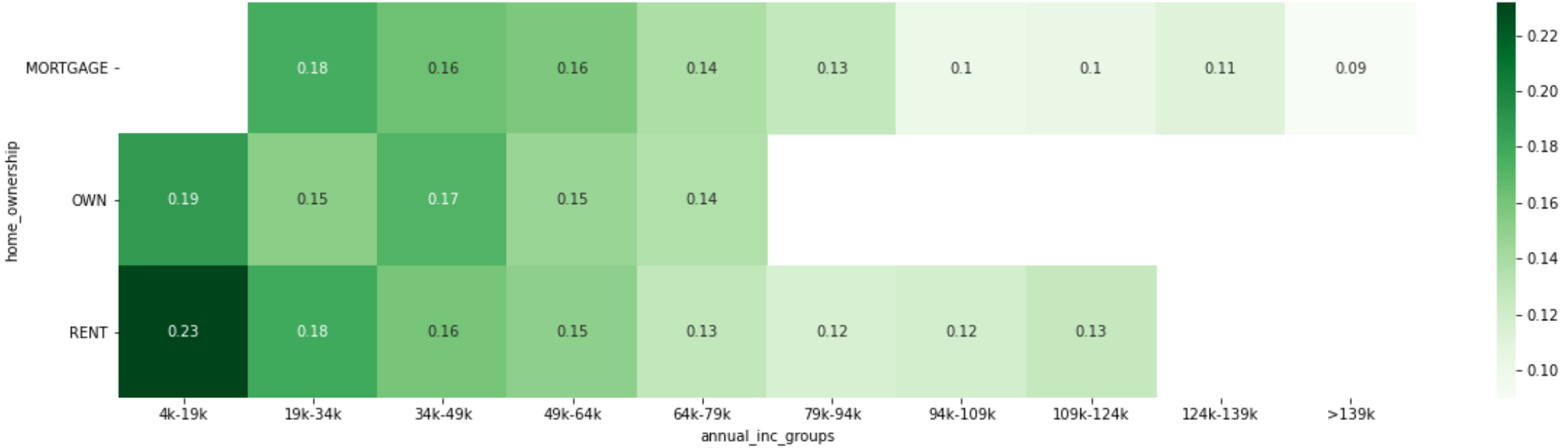
14) annual income and purpose with loan status



Inference:

1) The bivariate analysis of annual income with loan status showed us that with increasing annual income the default rate decreases. Same analysis is shown in this graph also.
2) The bivariate analysis of purpose with loan status showed us that when purpose was small business the default risk was highest. Same analysis is shown in this graph also.
3) Customers having any annual income and taking for small business purpose are most prone for default

15) annual income and home ownership with loan status



Inference:

1) The bivariate analysis of annual income with loan status showed us that with increasing annual income the default rate decreases. Same analysis is shown in this graph also.
2) The bivariate analysis of home onwership with loan status showed us no clear relationship. Same analysis is shown in this graph also.

Overall Analysis

Based on the EDA, we have identified the following features as most important in detecting default loans

a) loan amount

1) With increasing loan amount the probability of defaults also increases.

2) when the loan amount is greater than 13k, the default rate is always greater than the overall dataset default rate of 14.5%

b) Interest Rate

1) With increasing interest rate the probability of defaults also increases

2) when the interest rate is greater than 12.5%, the default rate is always greater than the overall dataset default rate of 14.5%

c) Annual Income

1) With increasing annual income the probability of defaults decreases

2) when the annual income is less than 64k, the default rate is always greater than the overall dataset default rate of 14.5%

d) Debt to Annual Income Ratio

1) With increasing debt to annual income ratio the probability of defaults also increases

2) when this ratio is greater than 10, the default rate is always greater than the overall dataset default rate of 14.5%

e) Revolving line utilization rate

1) With increasing revolving line utilization rate the probability of defaults also increases

2 when this rate is greater than 40, the default rate is always greater than the overall dataset default rate of 14.5%

f) term

1) The probability of default is more high for a loan given for 60 months than for 36 months

2) For 36 months the default rate is 11% and for 60 months its 25%

g) grade

1) The probability of default increases as we move in alphabetical order of grades; i.e the probability of default is more in G grade (~34%), less in F(~33%), more less in E(~27%) and same pattern follows till A(~6%).

2) Their is a order associated with grade and from grade C onward till G, the default rate is always greater than the overall dataset default rate of 14.5%

h) sub-grade

1) Sub Grade column follows the same pattern as that of grade with loan status column leaving a few exceptions; i.e the A sub grades default rate is lower than of B, B sub grades have a default rate less than of C and same pattern follows till G sub grades

2) Within each grade the the default rate increases with alphabetical order,eg; within A grade the default rate of A1 is less than of A2, A2 is less than of A3, A3 is less than of A4 and A4 is less than of A5.

3) The default rate of F5 is highest(~48%) and that of A1 is lowest(~2.6%) among all sub-grades

4) From all sub-grade C onward till all sub-grade of G, the default rate is always greater than the overall dataset default rate of 14.5%

i) purpose

1) The probability of default is highest for small_business(~27%),followed by renewable_energy(~18.6%)

2) When the loan is taken for debt_consolidation, educational, medical, moving, other, renewable_energy, small_business then the default rate is always greater than the overall dataset default rate of 14.5%

j) derogatory public records: Customers with known derogatory public records more prone for default with rate greater than the overall dataset default rate of 14.5%

k) known public bankruptcies: Customers with known public bankruptcies more prone for default with rate greater than the overall dataset default rate of 14.5%

l) Customers taking any loan amount but with interest rate greater than 15% are much more prone to default with default rate greater than 14.5%

m) Customers taking loan amount greater than 10.5K and annual income less than 64k have more prone for default with default rate greater than 14.5%

n) Customers taking loan amount greater than 15.5K and dti greater than 10 have more prone for default with default rate greater than 14.5%

o) Customers having loan amount greater than 10.5K and revolving rate greater than 70 have more prone for default with default rate greater than 14.5%

p) Customers having annual income less than 94k and interest rate greater than 15 have more prone for default with default rate greater than 14.5%

q) Customers taking any loan amount but for term of 60 months are more prone for default than customers taking the same amount for 30 months with default rate greater than 14.5%

r) Customers taking any loan amount but belonging to grade C, D, E or F are more prone for default with default rate greater than 14.5%

s) Customers taking any loan amount but having a public record of bankruptcy are more prone for default with default rate greater than 14.5%

t) Customers having any annual income but taking loan for a term of 60 months are more prone for default with default rate greater than 14.5%

u) Customers having annual income less than 79k and belonging to grade C, D, E or F are more prone for default with default rate greater than 14.5%

v) Customers having any annual income and taking loan for small business purpose are most prone for default with default rate greater than 14.5%

x) Customers having annual income less than 64k and taking loan for debt consolidation, housing or other purposes are most prone for default with default rate greater than 14.5%