GRAMENER CASE STUDY

SUBMISSION

Group Name: United By Chance

1. Ashish Gaurav
2. Ayushman Priye
3. Chandan Agrawal
4. Rahul SP

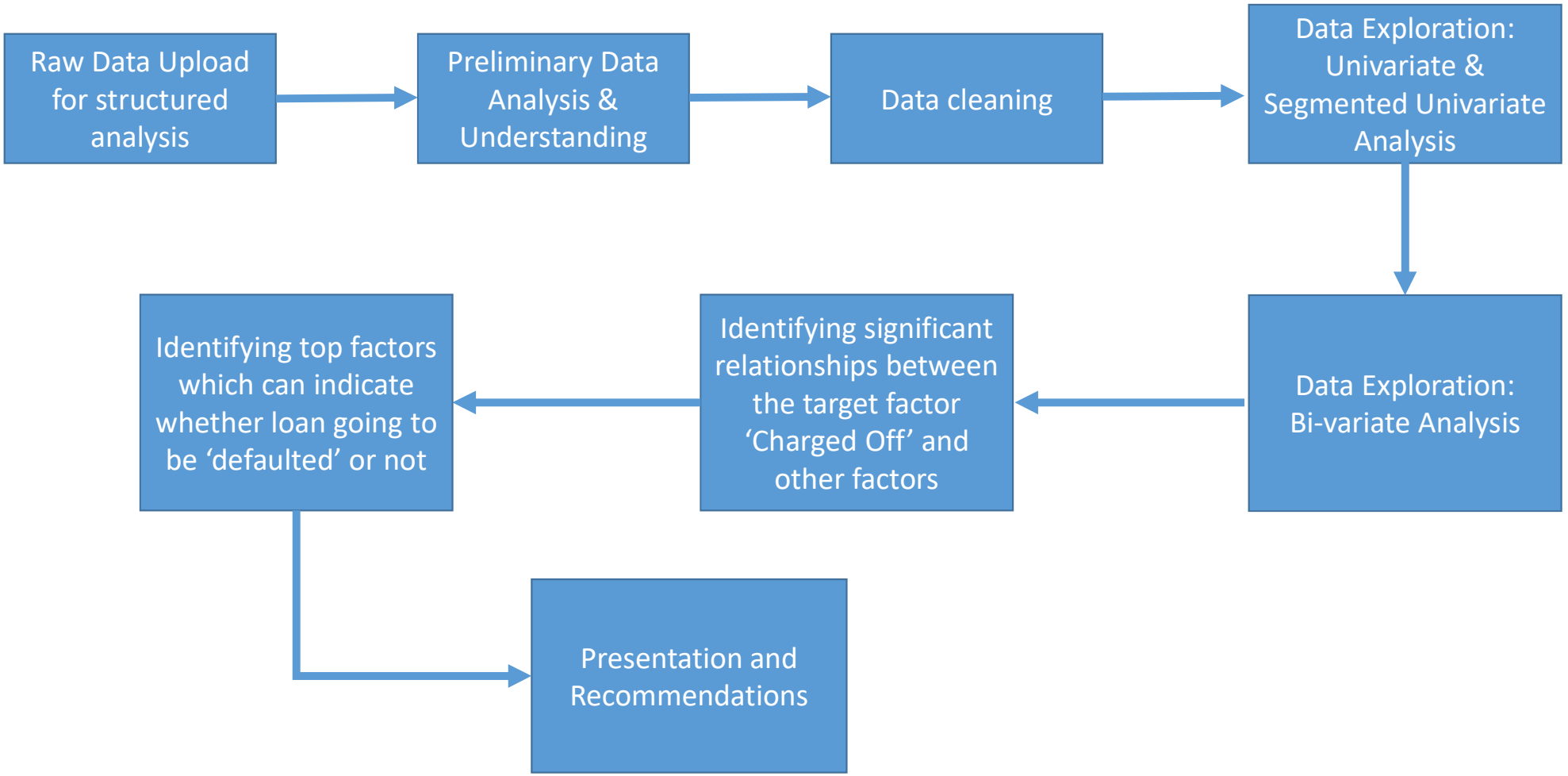
<Abstract - Objective and Problem Statement>

- The company specialises in lending various types of loans to urban customers. When the company receives a loan application, it has to make a decision for loan approval based on the applicant's profile. Two types of risks are associated with the decision:
  - If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
  - 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
- By analysing the past data about loan applicants and whether they 'defaulted' or not, company aims to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.
- If we are able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.
- In other words, 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 can utilise this knowledge for its portfolio and risk assessment.

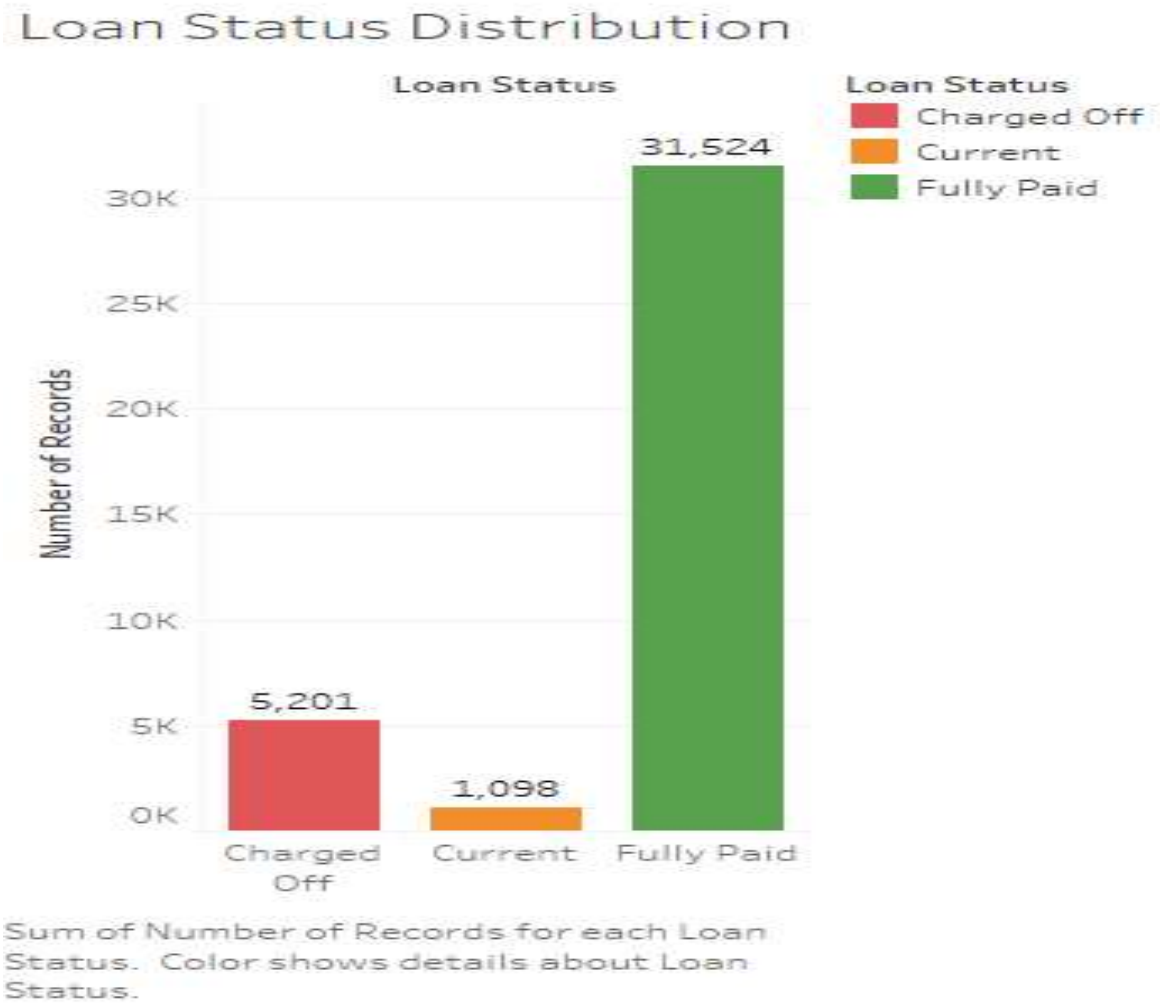
<Some Preliminary Observations>

- The dataset contains a large number of columns with detailed information about the loans applications.
- The column ‘loan_status’ is the target variable for further analysis. Here the value ‘Charged Off’ means those loans got Defaulted. We need to identify other attributes which can indicate whether loan is going to be defaulted or not. Thus, primarily we need to identify the relationship between ‘loan_status’ and other columns.
- A large number of columns are either having no values (NA) or the value is same in all records in such values. These columns won’t be contributing in any way in the analysis and thus can be safely dropped.
- Also after going through the available literature, it can be found that some columns (like funded_amnt_inv , out_prncp, out_prncp_inv etc.) do not seem to contribute the loan Defaulting behaviour. Thus these loans can also be dropped before doing further analysis.

<Approach followed>

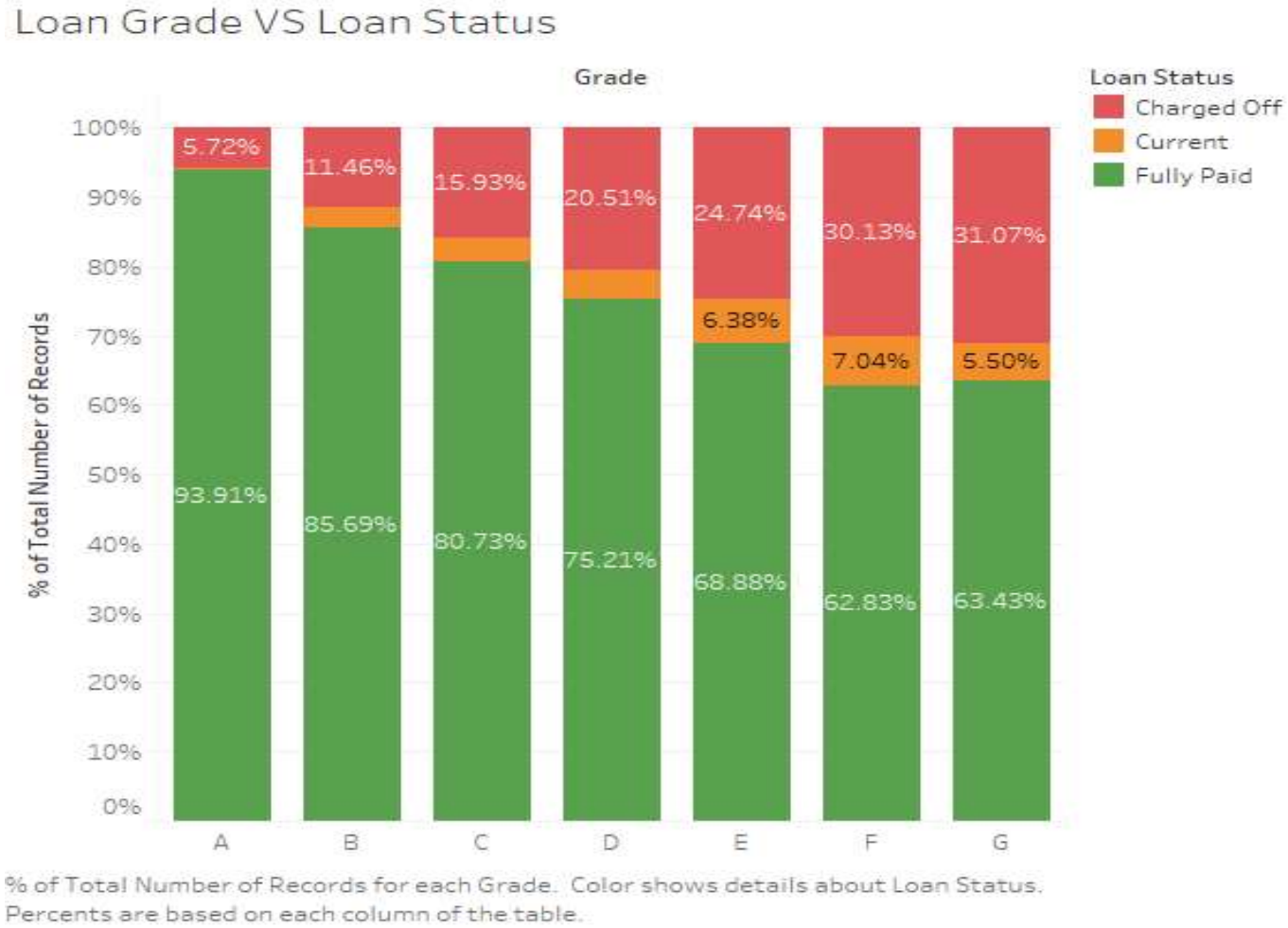


<Some Preliminary Analysis>



- This graph shows the distribution of loans across different status.
- Out of the total records available, 5201 loans got defaulted.
- It means about 13.75% of the loans considered for the analysis got defaulted because of different reasons.

<Some Important Drivers - 1>



- The Grade of Loan turns out be a good indicator of the 'Default'.
- More than 20% of the loans in Grades **D, E, F and G** are getting Defaulted.
- Loans in Grades **F and G** are more problematic with more than 30% cases of Default.

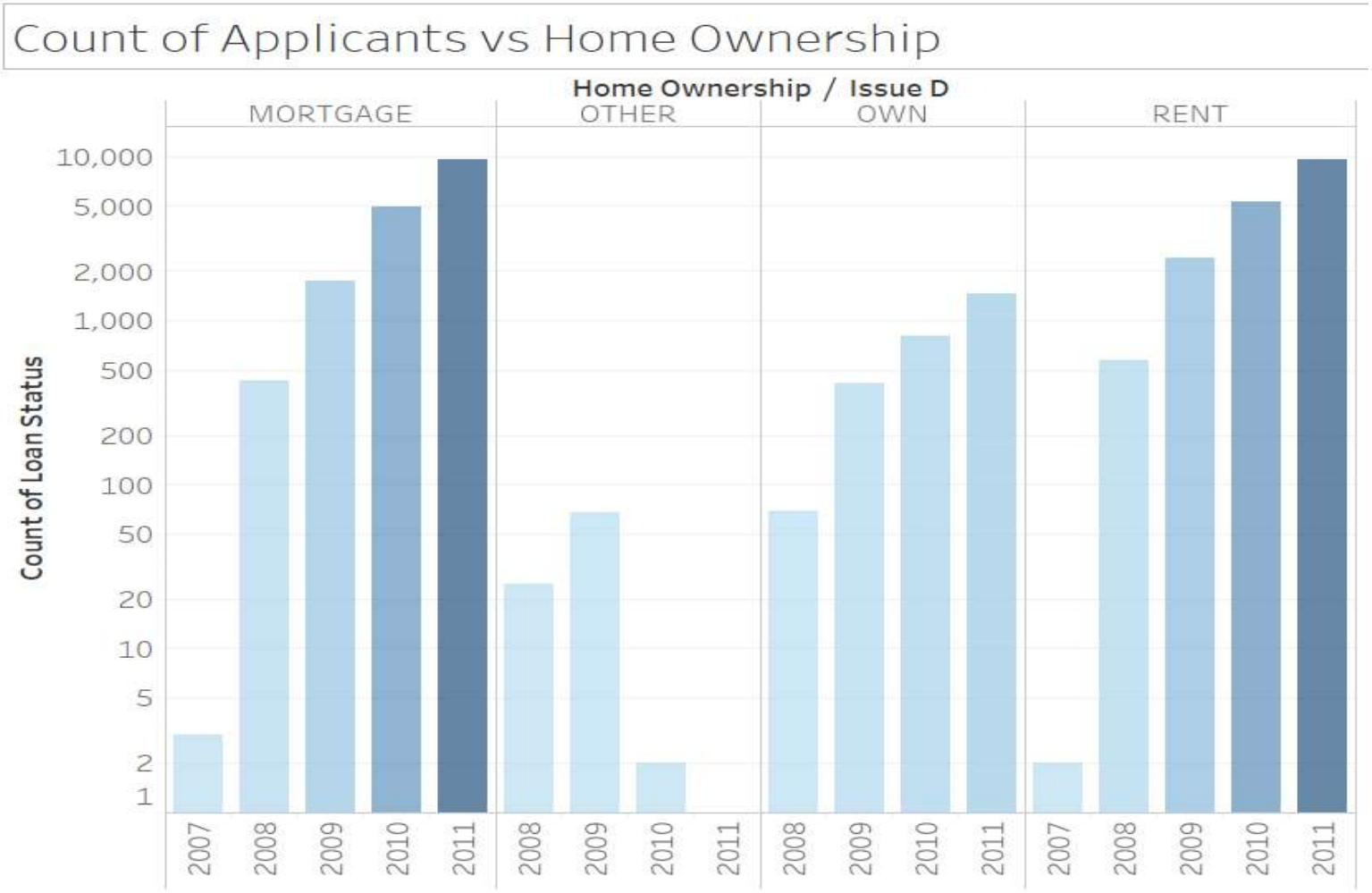
<Some Important Drivers – 2.a.>

Home Ownership Status VS Loan Status

% of Total Number of Records for each Loan Status broken down by Home Ownership. Color shows details about Loan Status. Percents are based on each pane of the table.

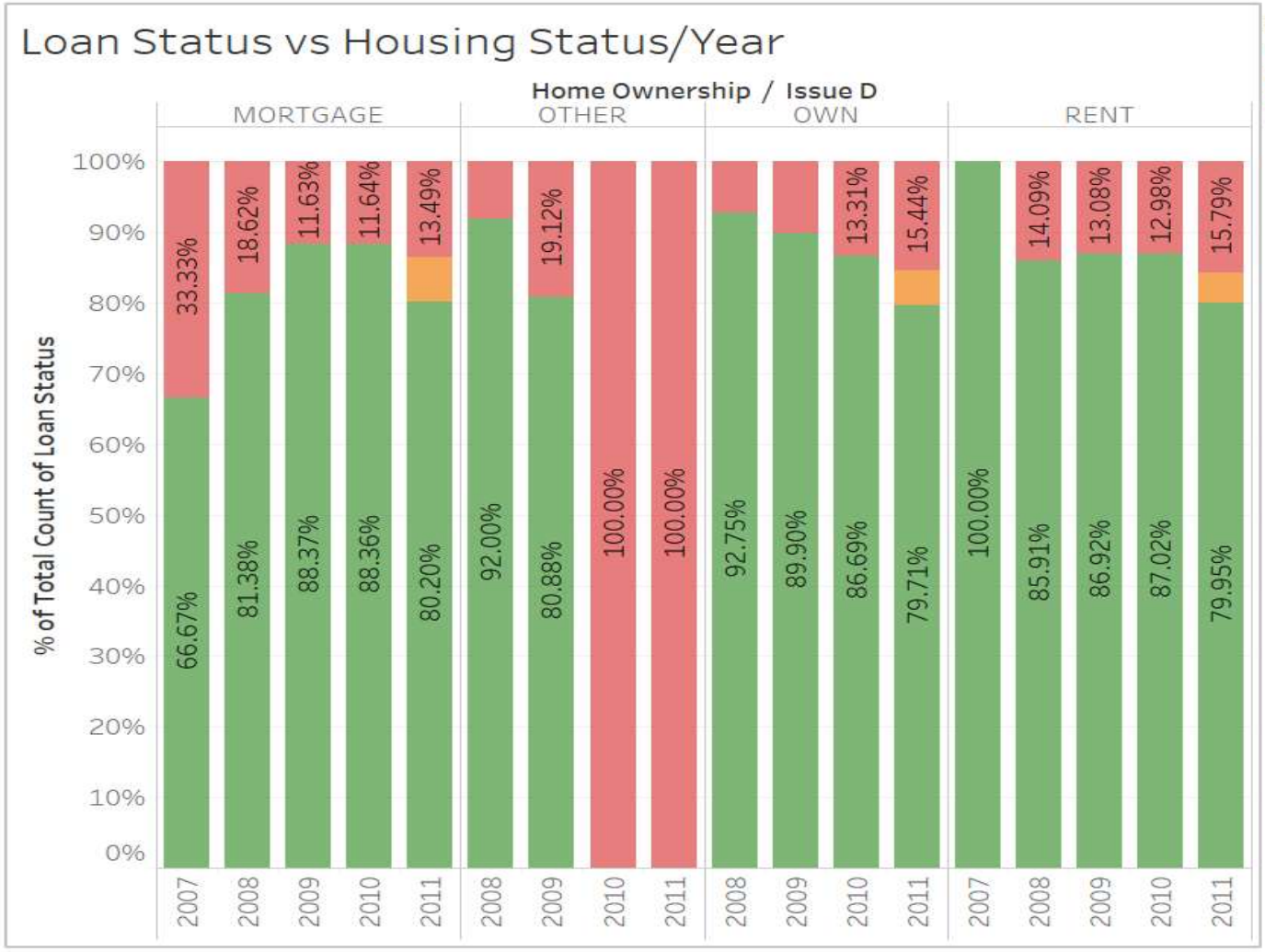
- More than 18% of the loans where Home Ownership Status is **'OTHER'** has got defaulted.
- This is much higher than for the other values of Home Ownership Status.

<Some Important Drivers – 2.b.>



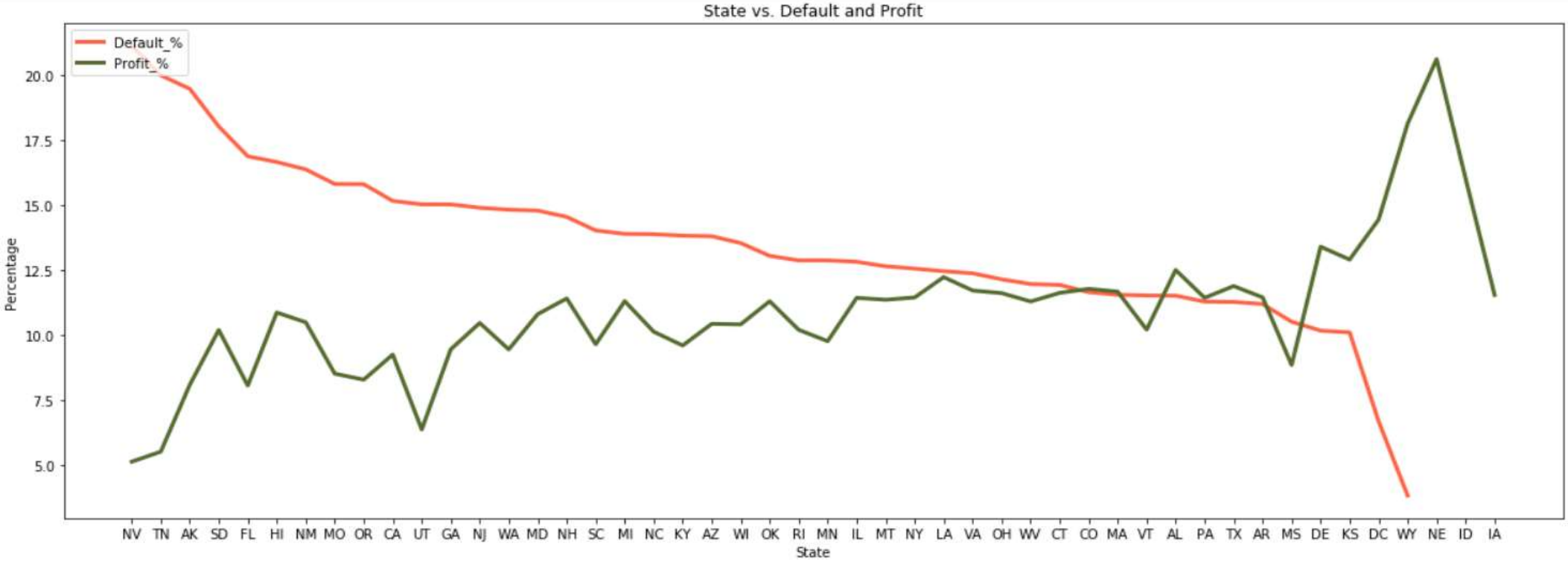
- But we also have to note that the number of applicants with 'Other' status as far less than other categories.
- We see an interesting pattern. See graph on next page.

<Some Important Drivers – 2.c.>



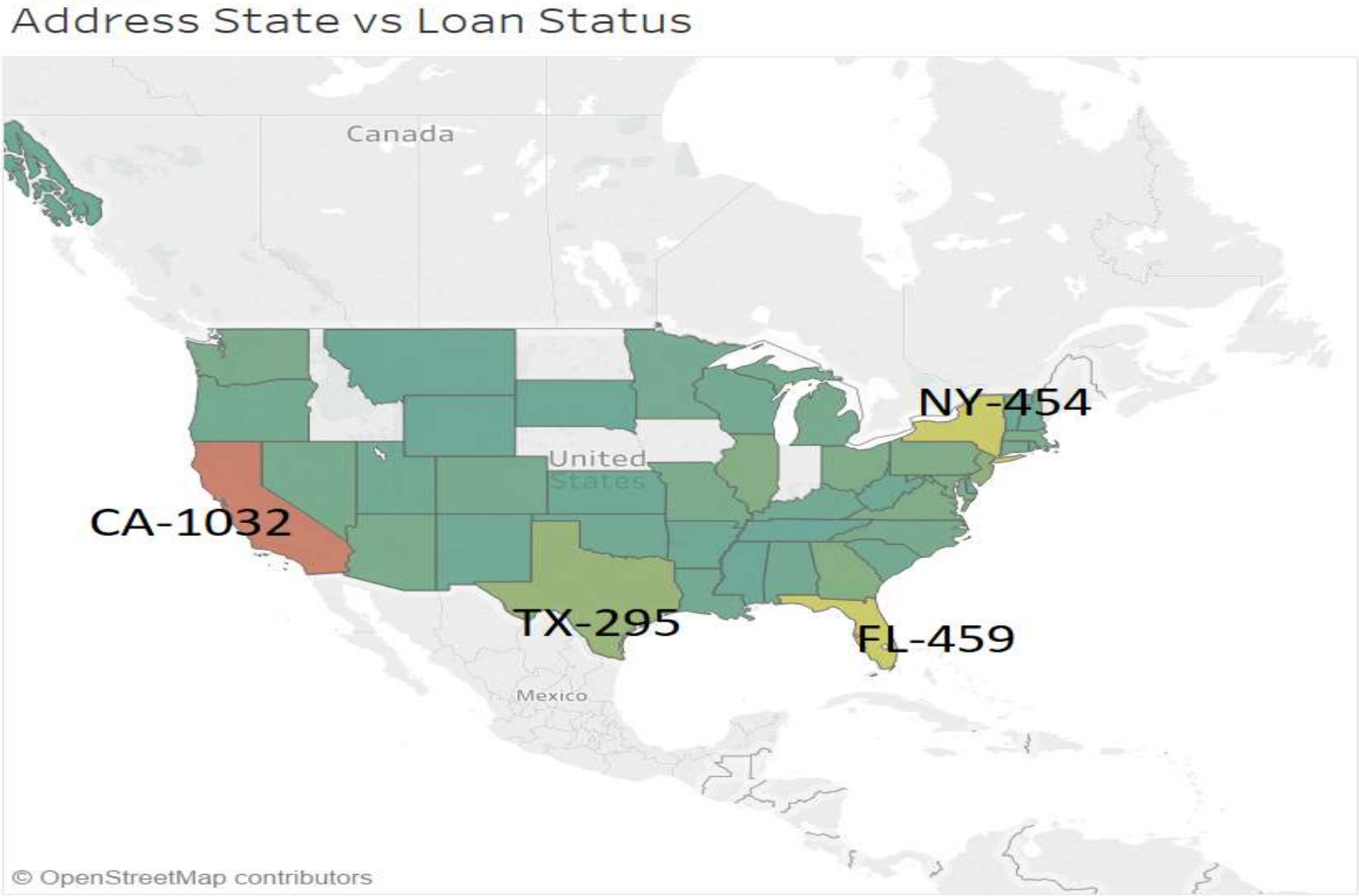
- Look at 'Charged Off' status in 'Mortgage segment. Year 2007 has 33.33% defaulters, then it suddenly reduces to 18.62% the next year and decreases more in further years.
- And applicants with 'Rent', 'Own' status seem more likely to be defaulters when compared with 'Mortgage' status for further years like 2010, 2011. *(Note: We are not comparing with the category 'Other' because of it's sparsity. Look previous graph 2.b.).*
- This might be because after the housing bubble burst in 2008, mortgage was given to people only after stringent scrutiny. Thus the people who were qualified for mortgage, were less likely to default on another loan as well, as they had already been thoroughly verified in their mortgage process and considered fit for giving loans.

<Some Important Drivers – 3.a.>



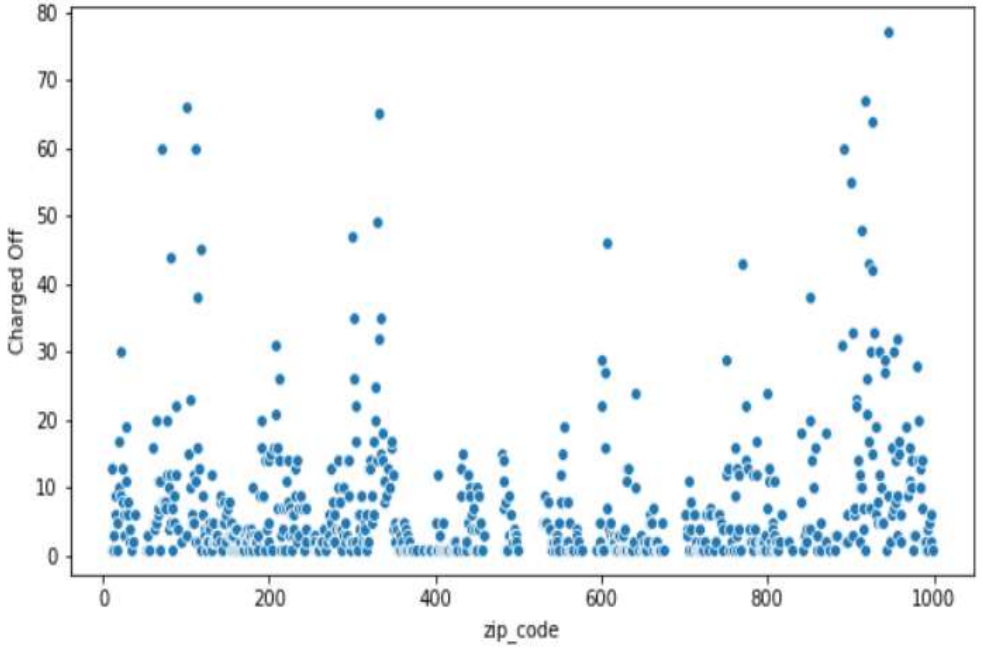
- Address State **'NV' (Nevada)** is seeing the highest rate of Default in loans.
- Both **Nevada and Tennessee** are the Address States having more than 20% as Default cases.
- States like **KANSAS, WYOMING, DELAWARE, WASHINGTON D.C**., etc. show higher profits and lower defaulting.

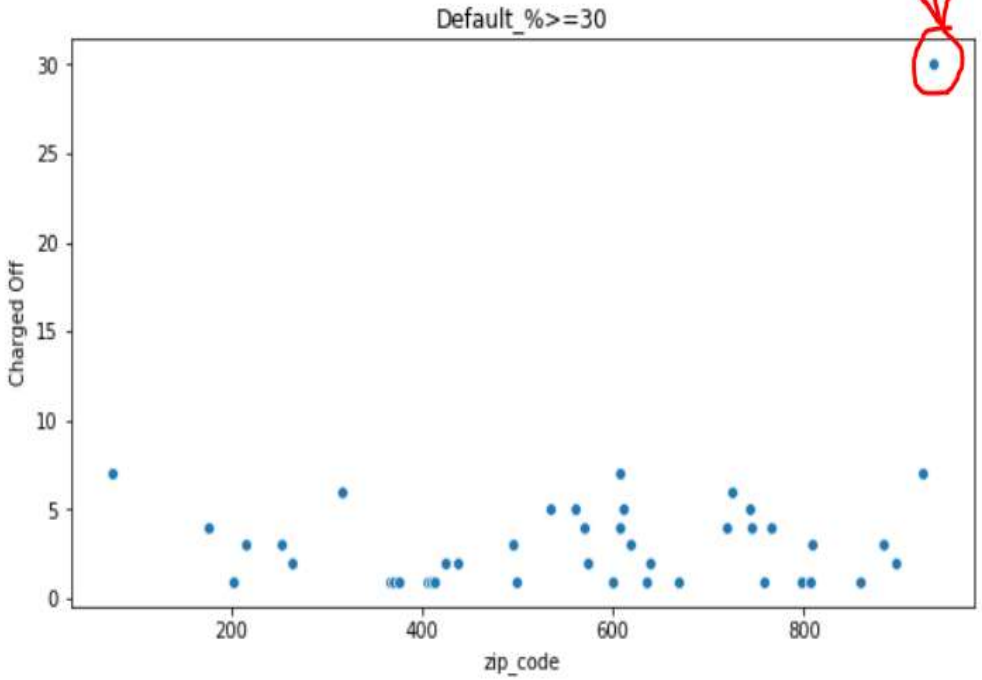
<Some Important Drivers – 3.b.>



- The top number of defaulters are:
- 1. California(CA),
- 2. Florida(FL),
- 3. New York(NY),
- 4. Texas(TX).

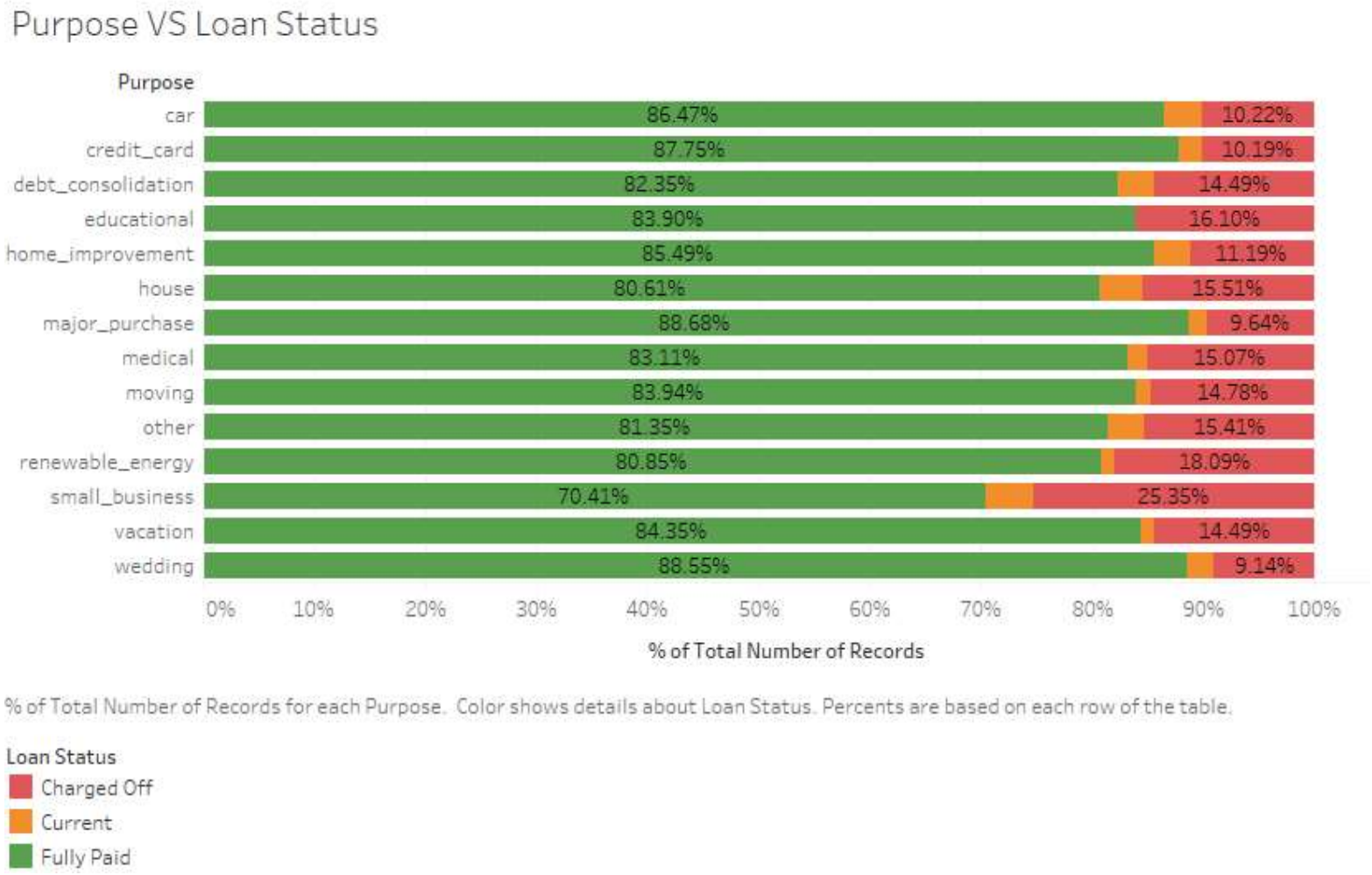
<Some Important Drivers – 3.c.>





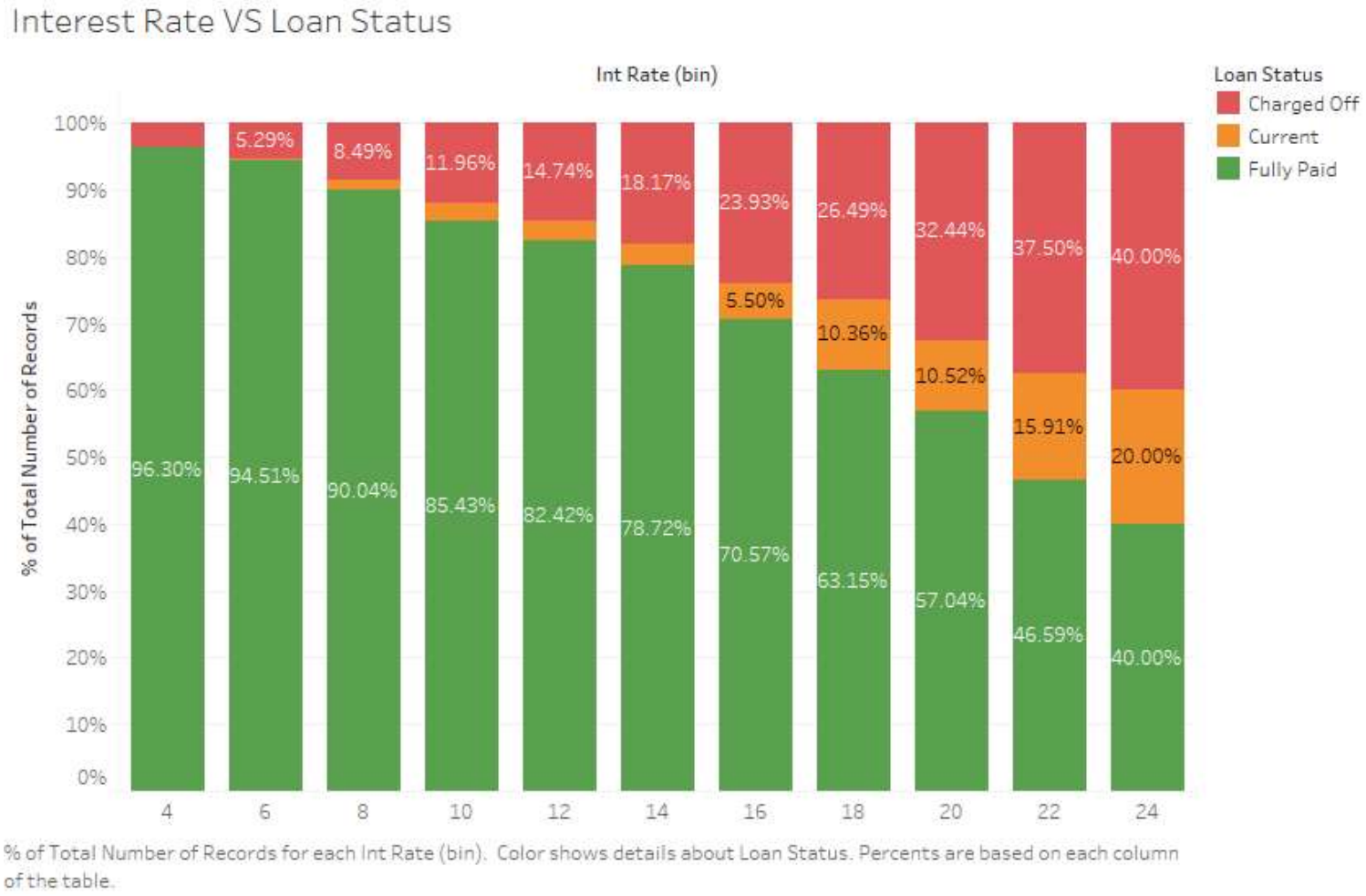
- X-axis shows number of Charged-Off applicants. Y-axis shows the zip-code. Eg. 800 is shown for 800xx.
- It can be seen that zip codes ranging around 100, 300 and from 900 to 1000 have high number of 'Charged Off' status. *(See left graph).* We can see on the right graph (points are only displayed if Default% is >=30%.), that zip-code 935xx has maximum 'Charged Off' applicants. This zip-code belongs to **California**.

<Some Important Drivers - 4>



- Loans for purposes like 'small_business' and 'renewable_energy' are seeing a higher percentage of default.
- On the other hand, 'wedding', 'major_purchase', 'credit_card', 'car', etc. seem to be a good choice for giving loans.

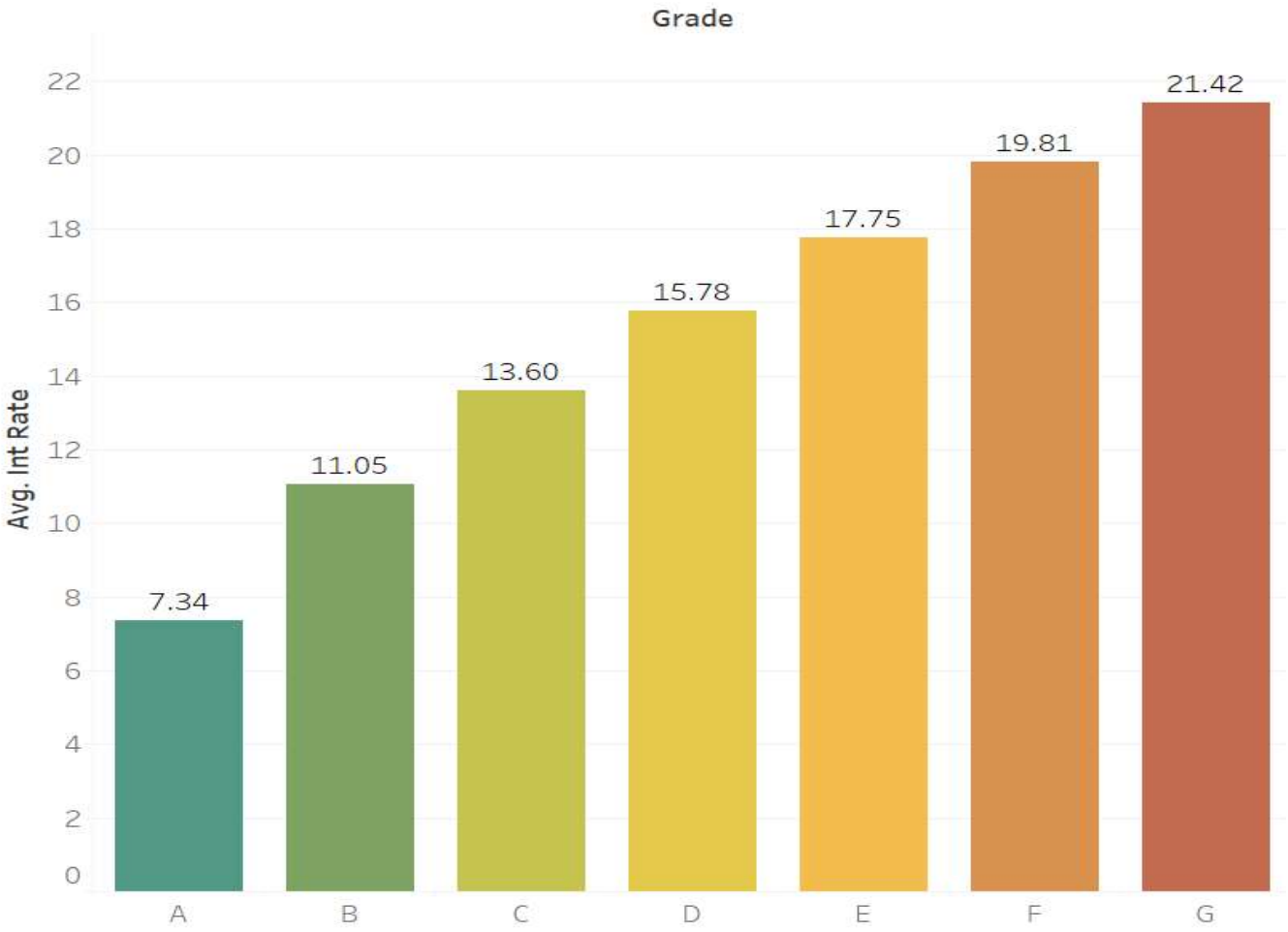
<Some Important Drivers – 5.a.>



- Interest Rates are segregated in different bins here with a bin-size of 2.
- As was expected, as the Interest Rate increases, the percentage of Default cases is also increasing.
- Loans with interest rate more than 24% are having Default cases of 40%.

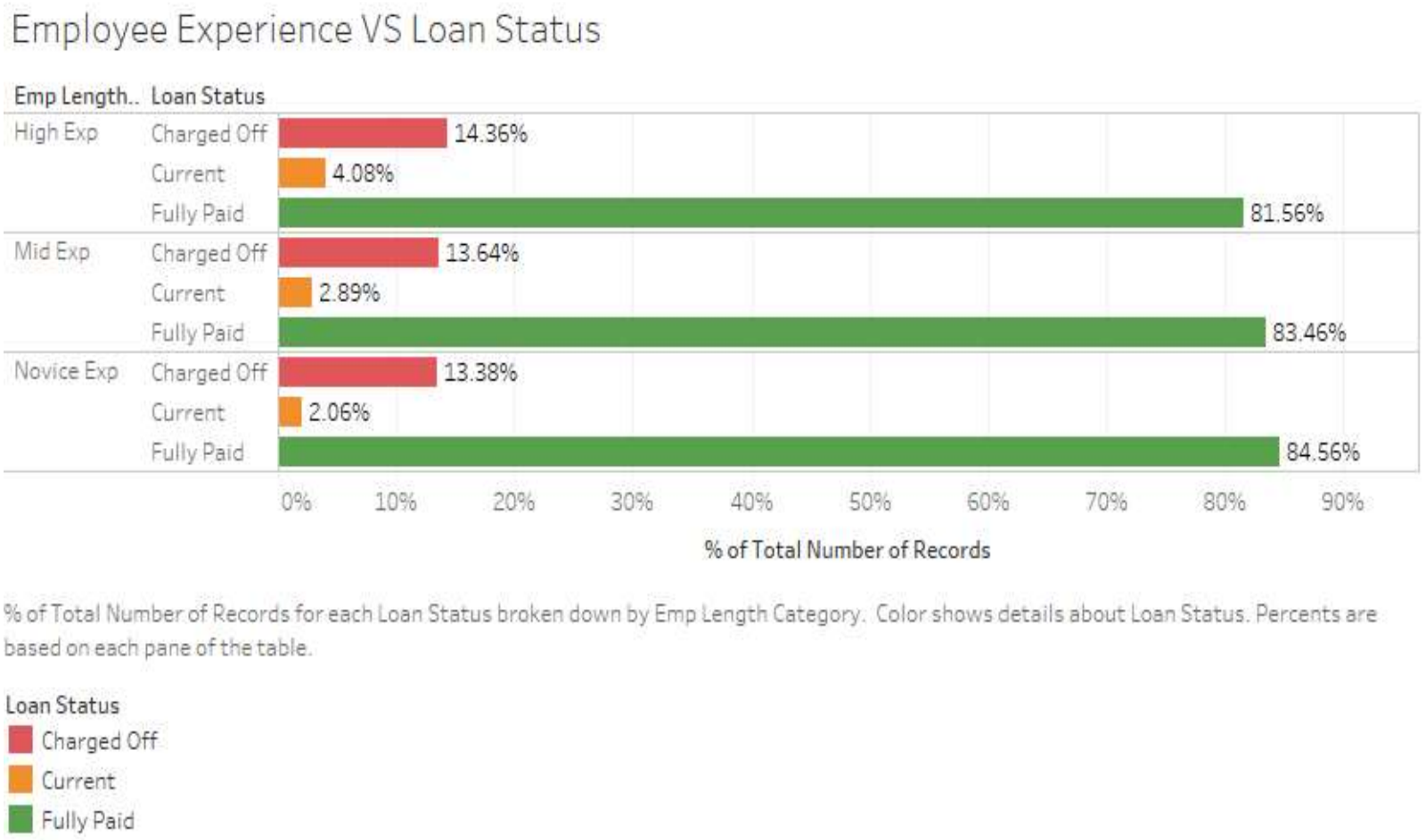
<Some Important Drivers – 5.b.>

Grade vs Interest



- Interest rate is also associated with grade, which is in turn related with defaulting behaviour.
- As grade changes from A to G, interest rate increases as well.

<Some Important Drivers – 6.a.>



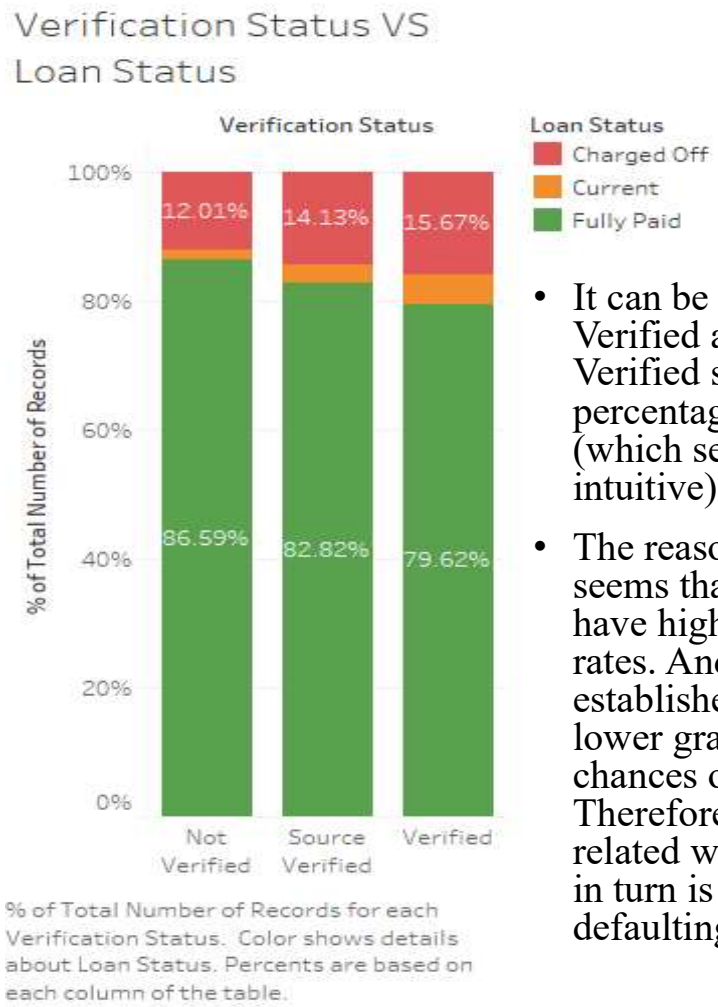
- It appears that people who have just begun their carrier show lesser chance of defaulting.
- See the explanation for this in next graph 6.b.

<Some Important Drivers – 6.b.>

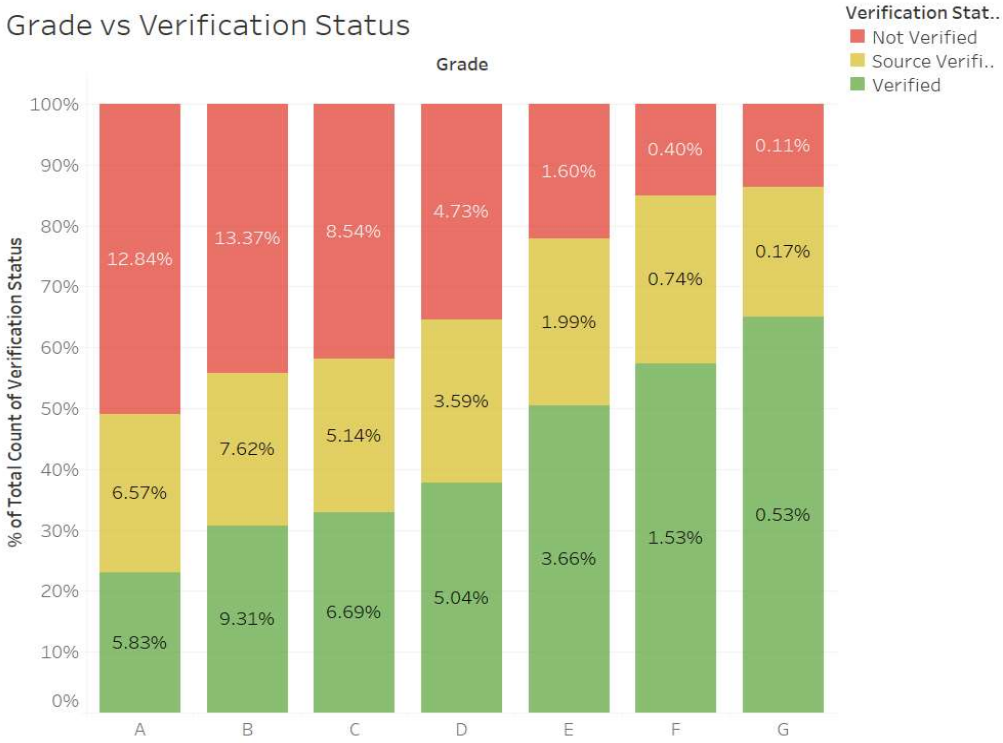
emp_length_category grade	High Exp	Mid Exp	Novice Exp	Novice_Grade_%
A	3005	2641	3900	69.08
B	3428	3180	4852	73.43
C	2184	2138	3385	78.32
D	1491	1437	2123	72.51
E	884	771	1086	65.62
F	337	292	380	60.41
G	106	87	116	60.10

- The reason for inference in previous slide 6.a. seems that better grade loans like A,B and C (which have lower interest rates) are more likely to be given to novices and lower grade loans (with higher interest rates) like D,E,F and G are less likely to be given to novices.

<Some Important Drivers - 7>



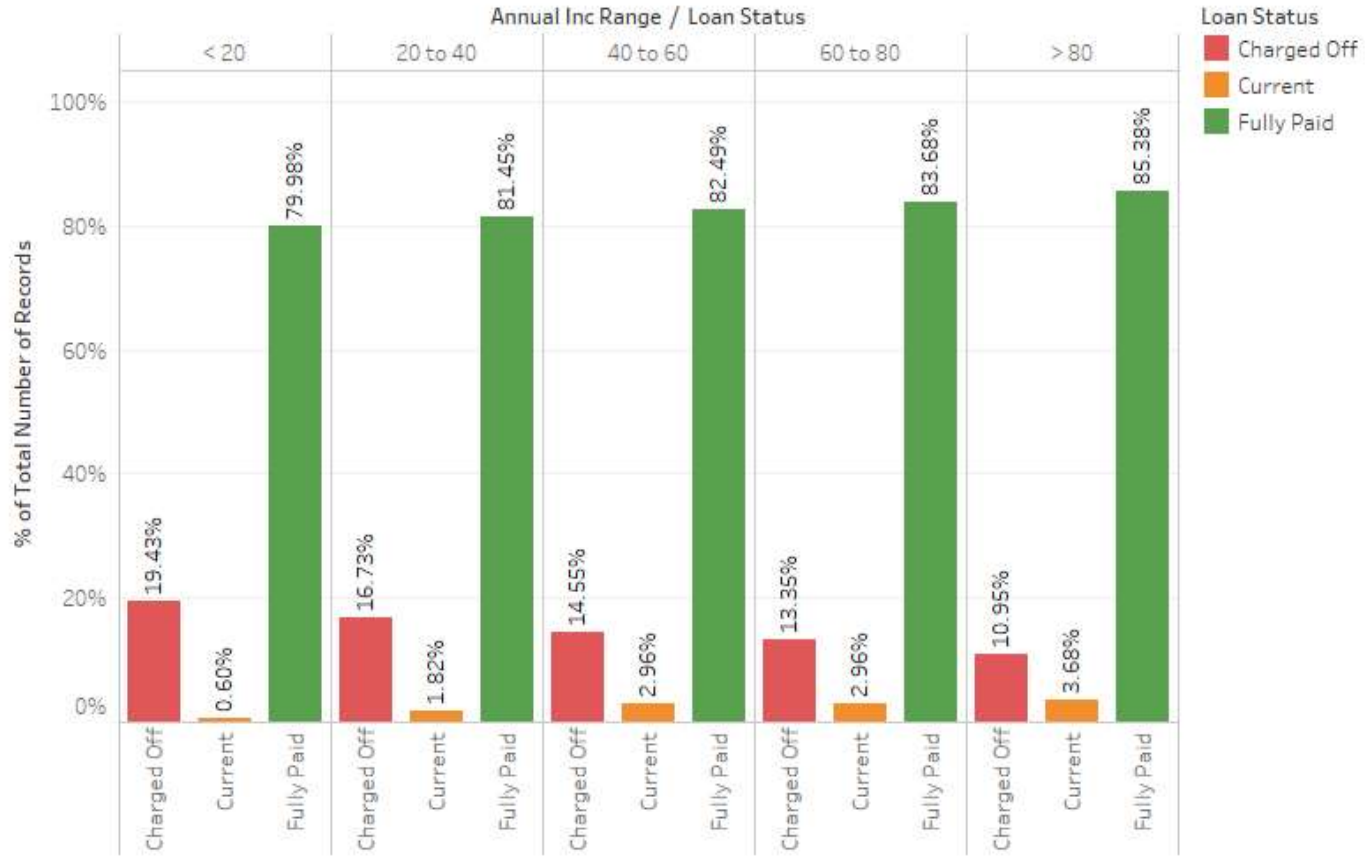
- It can be seen that, Verified and Source Verified show a higher percentage of default (which seems counter-intuitive).
- The reason behind this seems that lower grades have higher verification rates. And we had established earlier that lower grades have more chances of defaulting. Therefore, verification is related with grade, which in turn is related with defaulting.



- It is also to be noticed that, higher grades are not verified much, because it is known that they have lower chances of defaulting. So, essentially verification is only done more strictly where it seems necessary. Thus, grade seems to be a better predictor of defaulting behavior than verification status.

<Some Important Drivers - 8>

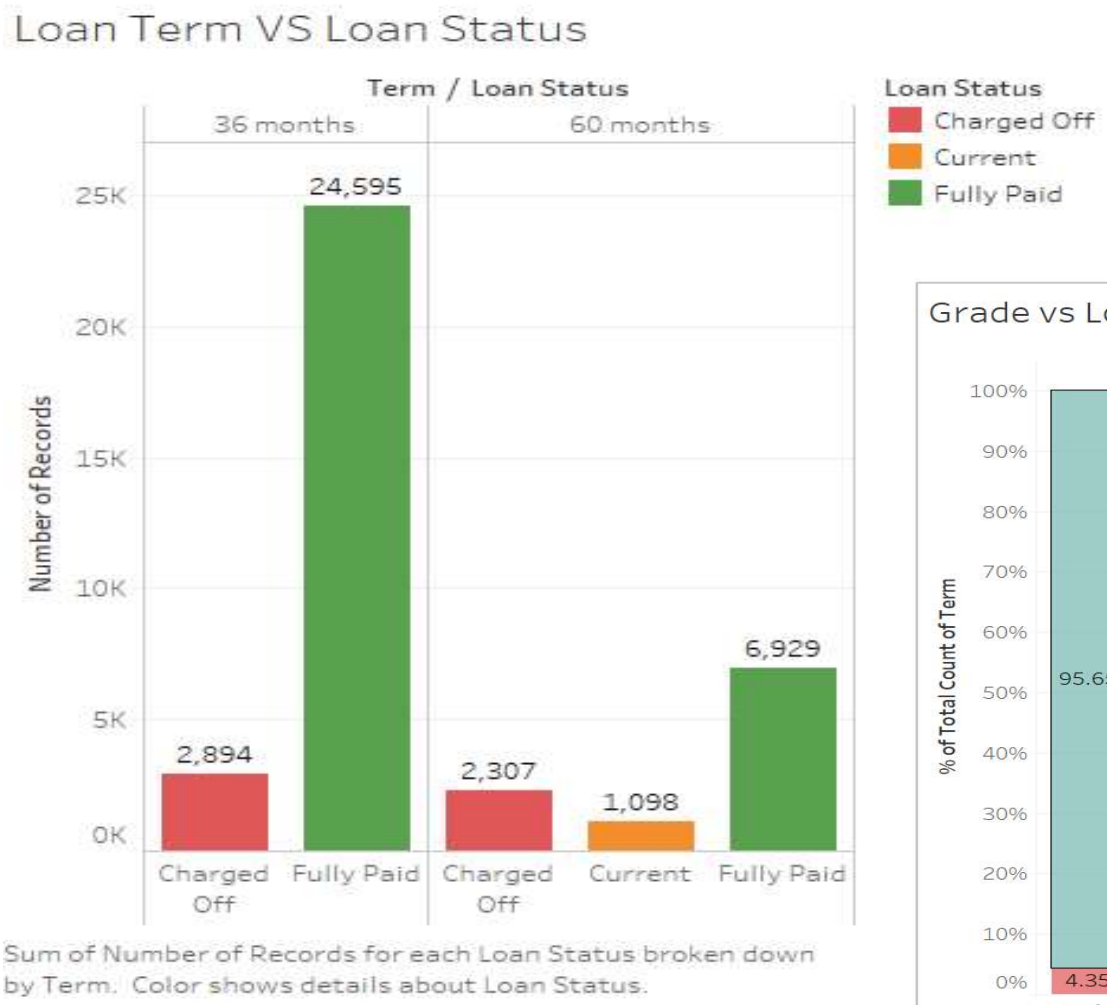
Annual Inc Range VS Loan Status



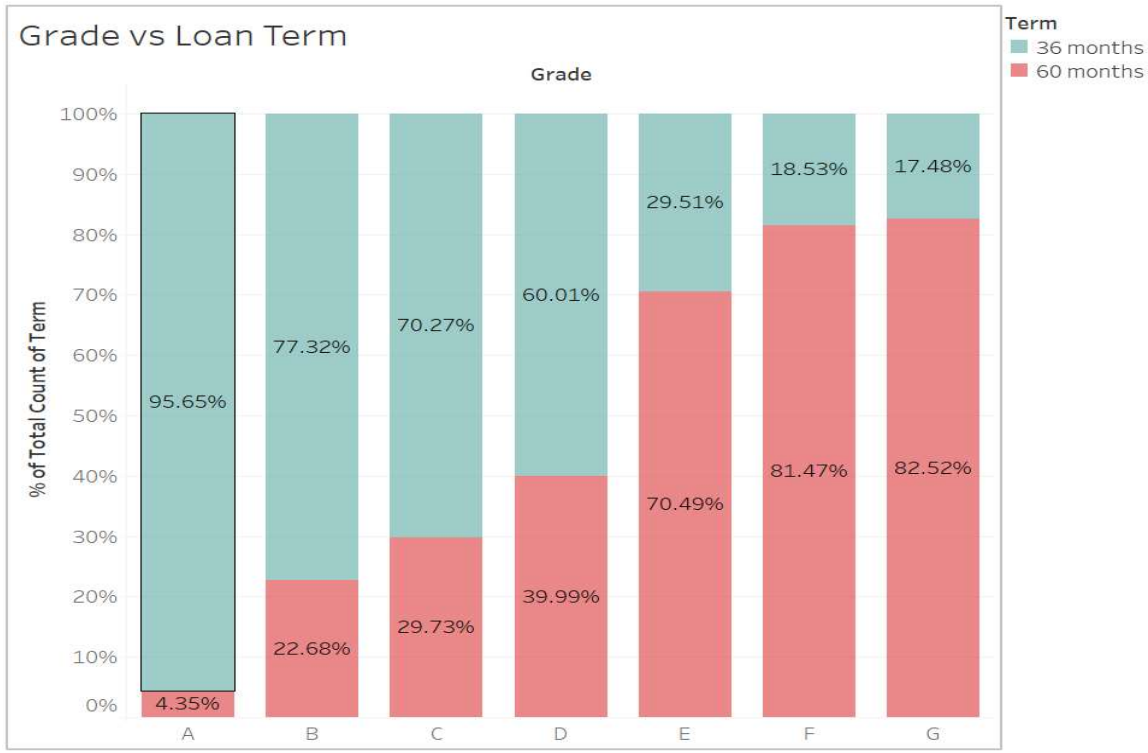
% of Total Number of Records for each Loan Status broken down by Annual Inc Range. Color shows details about Loan Status. Percents are based on each pane of the table.

- It can be seen that as the salary range increases, less percentage of people default.
- *Note: The salaries are in thousands. Eg. 40 to 60 is 40,000 to 60,000 dollars.*

<Some Important Drivers - 9>



- The graph shows that 60 months term has higher chances of defaulting.
- The reason seems that, 60 months term is essentially related with grade. Higher share of poor grades like D,E,F,G are associated with 60 month term.



<Some Recommendations>

- Verification process needs to be reviewed and if necessary changed to make it more robust.
- Company should be more cautious while giving loans to applications with grades D, E, F, or G.
- As about 25% of applications with ‘Small Business’ as purpose got defaulted, it might be a good idea to stop extending loans for this purpose.
- Interest Rate of the loan is calculated keeping other risks in consideration. However, it is easier to be tempted to extend a very risky loan with very high interest rate to reap profits. But as analysed a large chunk of such loans are getting defaulted. Utmost caution is needed here.
- Due diligence must be done before extending loans to applicants from states like Nevada or Tennessee and also if Home Ownership Status of the applicant is not ‘Mortgage’.
- It is good to be more careful when providing loan to people with salary less than 20,000 dollars.