Introduction

Lending Club Case Study

Name : Pankaj S Patil

ML Batch:- Dec 2021

Batch No :- C36

Data Sourcing

Importing Libraries

- Libraries which are required for importing the data, doing analysis, making presentation. The following libraries needs to be included in this case study

  - **Seaborn** :- Used for making visuals.
  - **Pandas** :- Used for loading dataset and perform operation once data loaded into dataframe.
  - **Matplotlib** :- Used for making /displaying visuals. Can be used to set the properties of visuals.
  - **Numpy** ;- As this is one of the important library, but in this case study it can be optional.

  There is one additional libraries which we need to add and that is

  - **Warning** :- This library used for filtering/not display the warning, while creating some visual it the code throws some warning which doesn't look much of the clean notebook so to avoid showing warning we filter those out using this library.

Data Sourcing

Check Data

- Once the data is loaded into data-frame using **Pandas** library check the data, looked for the number of columns and rows, datatypes of columns, uniqueness of data etc.
- Once went through the data, by now we realized there are several columns which has almost all “NULL” or “NaN” values present. Checked for uniqueness of rows also (duplicates) by using one or two fields (which ideally should be unique), found there are no duplicate data present in the dataset.

Lets proceed with **Data Cleaning** activity ⟶

Data Cleaning

Cleaning Data

- Once we are done with loading the data and identified those columns which can be cleaned then we can proceed with some data cleaning steps which has been taught to us in DA.

- As while understanding the data, we found there are 54 columns which has all the values “NULL” or “NaN”, as a starting point of cleaning activity we dropped those columns as column with no values won’t be helpful for analysing the motive.

- After dropping all the NULL values column, we end up having 57 columns of data with no duplicates rows. So after taking the look at the data as per the understanding giving by coach there are several rows which are specific to Customer Behavioural like deling_2_year, revol_bal, last_pymnt_d etc so we can drop those columns also since we are analysis a data for Diagnostic purpose and not performing any Prediction out of it so these variables came into picture if we are building Predictive model. As our main moto stays, with the existing data need to show who all are could be defaulter or not defaulters.

- After removing all the columns, we have 24 columns which are mainly focused on the borrowers characteristics like the loan amount, emp_length, emp_title, Grade, int_rate etc.

Data Analysis

Rectify Data & Datapoints

- After cleaning up all the unnecessary and invalid data we have the relevant data at our hand which is still raw and needs to be rectified for better analysis.

- Even though we cleaned the data there is still a possibility of having NULL values in the relevant columns which going forward will cause us trouble while making the data presentable. So checked for the NULL values in the rectified data where found there are some values in emp_title, emp_length, title and pub_rec_bankruptcies.

- Now we have identified columns having NULL values in the rectified data lets try to handle these missing values by either replacing them with some aggregated value of the respective field or by dropping those.

- emp_title :- Since there are few records having NULL values in this column, instead of dropping those replaced it with "**Not Available**". If we would've dropped those column there would be loss of data in terms of loan_amnt, int_rate etc which would've directly impacted in analysing the data.

Data Analysis

Rectify Data & Datapoints

- emp_length :- In this field also there are few NaN values, so we cannot replace them with some hardcoded one, since these values can be used to identify defaulters so instead of some hardcoded value will fill the missing value with the average of emp_length.

- Along with filling the missing value, we considered the "< 1 year" category to be "0 year" as per stated in the Data Definition file (**Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years**.)

- Then we separated the numeric value from the string value using regex, and also changed the datatype of the same field from "object" to "int" so that the column will have numeric value all the way to end and to check the values statistically, we can get idea about the range of the data and some info on the same.

- After performing all the above steps, we could see there is in increase in 5 years category since we have replaced NaN with the mean value and the count of 0 years is same as < 1 year category as we replaced the value with 0 year.

Data Analysis

Rectify Data & Datapoints

- int_rate :- Going forward, when we are done with emp_length and emp_title column, tried to get the details of int_rate column using describe, the values of column containing special character, the describe function won't work properly.

- To do the smooth analysis like to check how the interest rates is varying we need to have some numeric values. To do that so me just took out the last character from the string and made the column type from object to float.

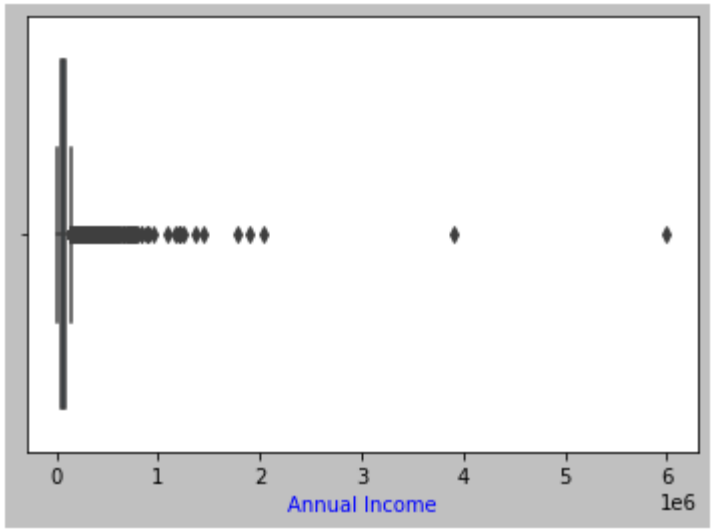
- After handling int_rate, went further and checked the issue_d column, which don't have any NULL values, but since the values represented as Month-Year type, more like a date field, we can derive Month and Year from it. For now only derived the Month from it and named the column **issue_m**, since I think it will helps to show the Interest Rate distribution month wise.

- After checking and deriving some metric we further proceeded with checking outlier, after going through int_rate, dti variables we found annual income is one of the important field since its dealing with annual income of the borrower. While checking, found the there are outlier present in that field and to visualize the same, thought of displaying it using boxplot, since its very useful to display outliers.

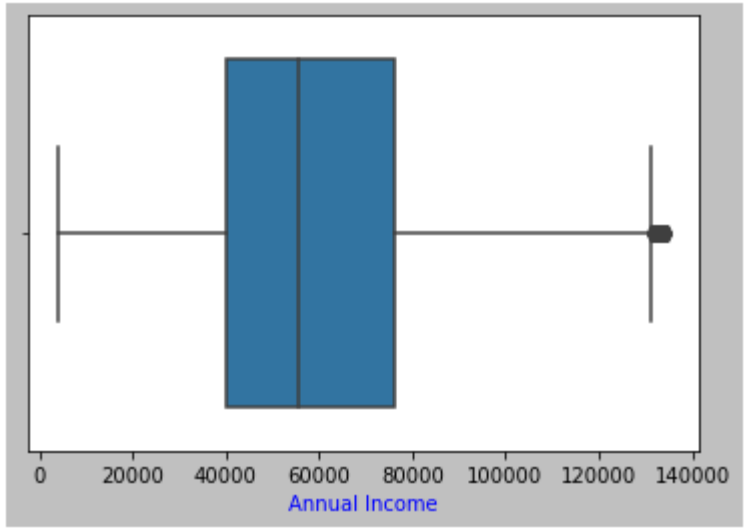
Data Analysis

Rectify Data & Datapoints

- To handle the outlier, took the 97 percentile of that field and updated the data-frame to have a data which is outlier free. Below are the visuals.



Before Outlier Removal



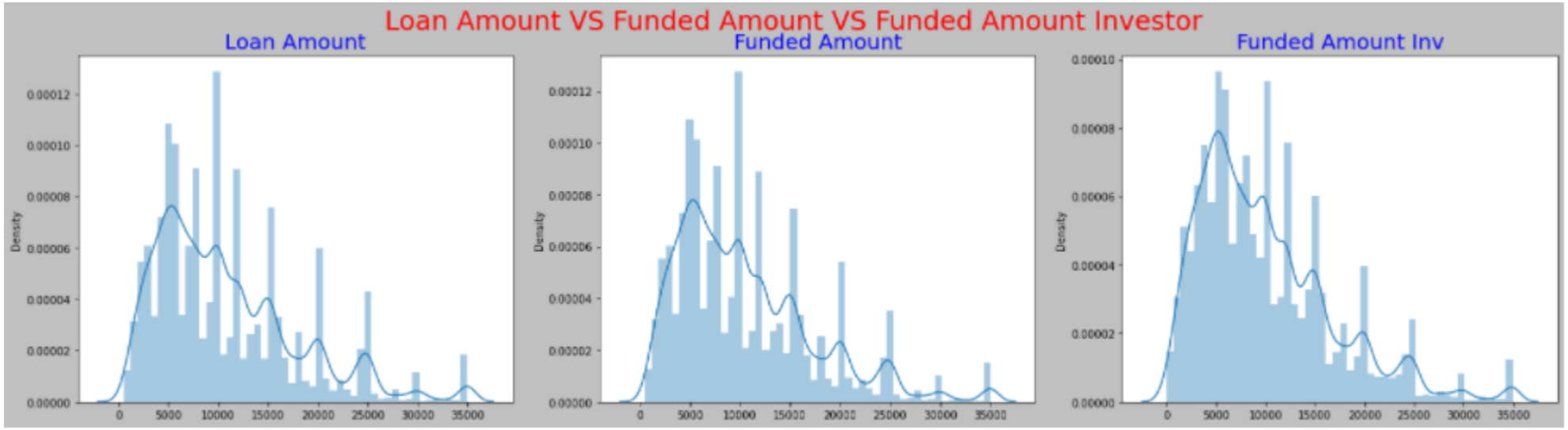
After Outlier Removal

Univariate Analysis

Describe Data

- Since Univariate deals with single variable, we an describe the information for the particular column about the data range it has & kind of data it contains.

- For proceeding, looked at the data again, found we have 4 important value we should be looking for those are loan_amnt, funded_amnt, funded_amnt_inv and annual_inc. As we already know about annual_inc come while removing the outliers, then we can compare all the remaining three variables to check the comparison between them. Below is the comparison visual.



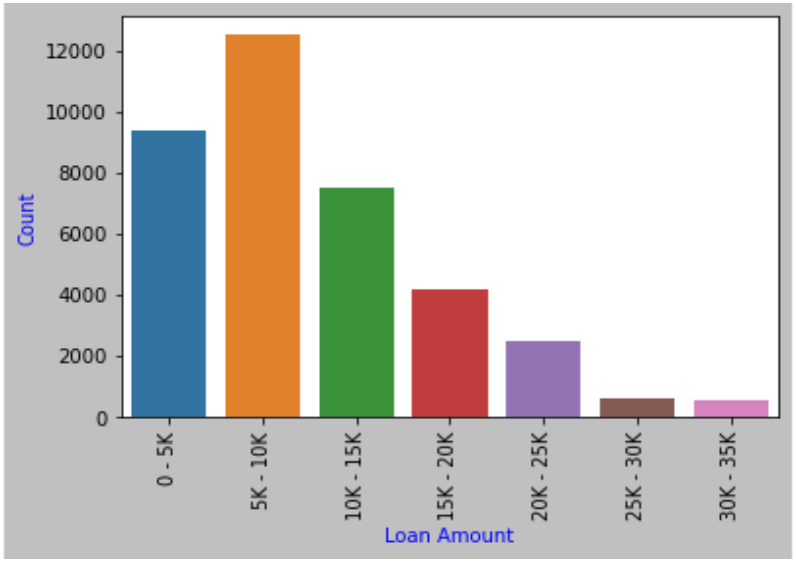
Univariate Analysis

Describe Data

- We are doing the comparison of three amount so that we can find how these variable are varying from each other. From the previous graph it looks like the 3 variables almost flows the same pattern where we can say the most of the amount falls from 5K to 15K in all the three visuals, so if we focus on one variable that will be also give the close analysis for other variables too.

- Now to check our assumptions is correct or not, thought of creating bins for the loan_amnt field to see how the loan_amnt is varying. Below is the snap of bins we created for the loan amount.

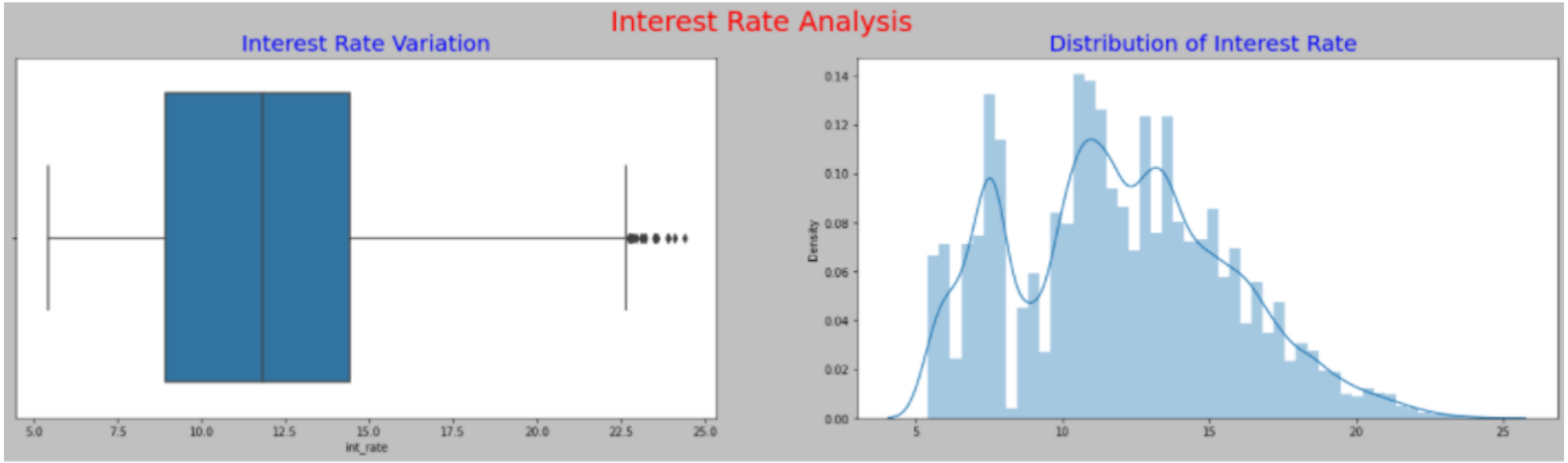
- The images show our assumption we pretty unclear as the most of the loan amount falls into “5K – 10K” category followed by “0 - 5K” and “10K – 15K”.



Univariate Analysis

Describe Data

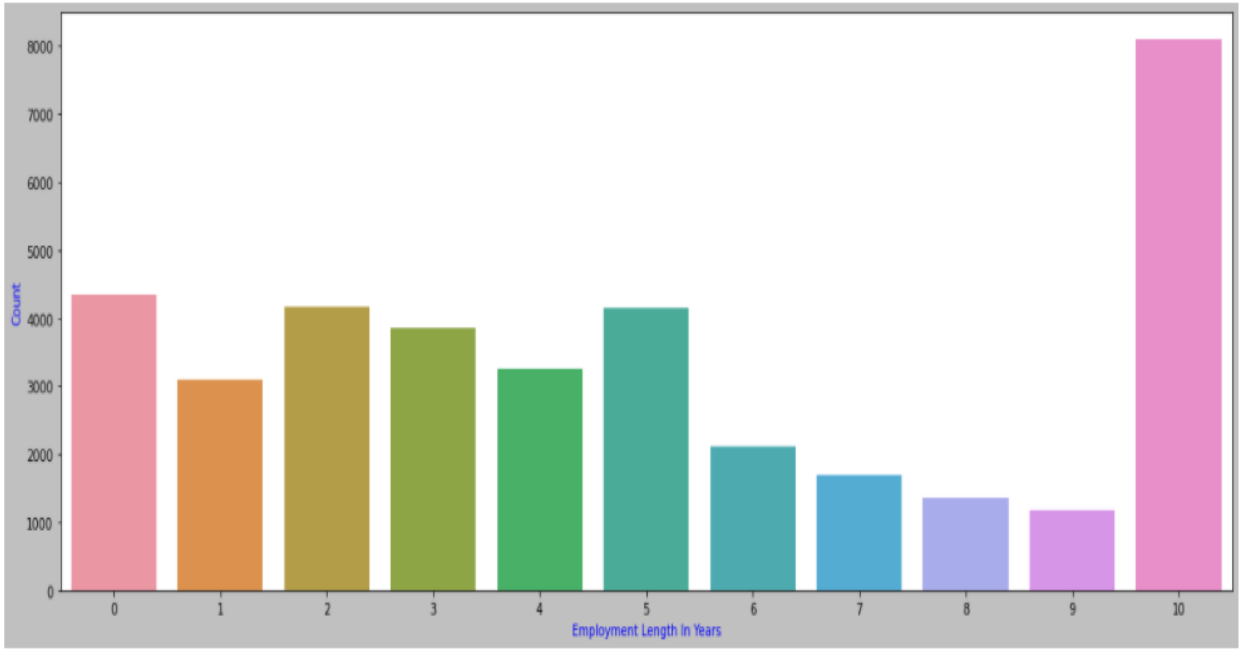
- As we are in Univariate Analysis phase, we already took a look at one of the Quantative variable that is loan amount, lets take a look at the values of interest rate, after plotting the below box plot and the histogram we could say the most interest rate varies from approx. 9 to 15 and post that, it is declining as per the **kde** and the **graph frequencies**.



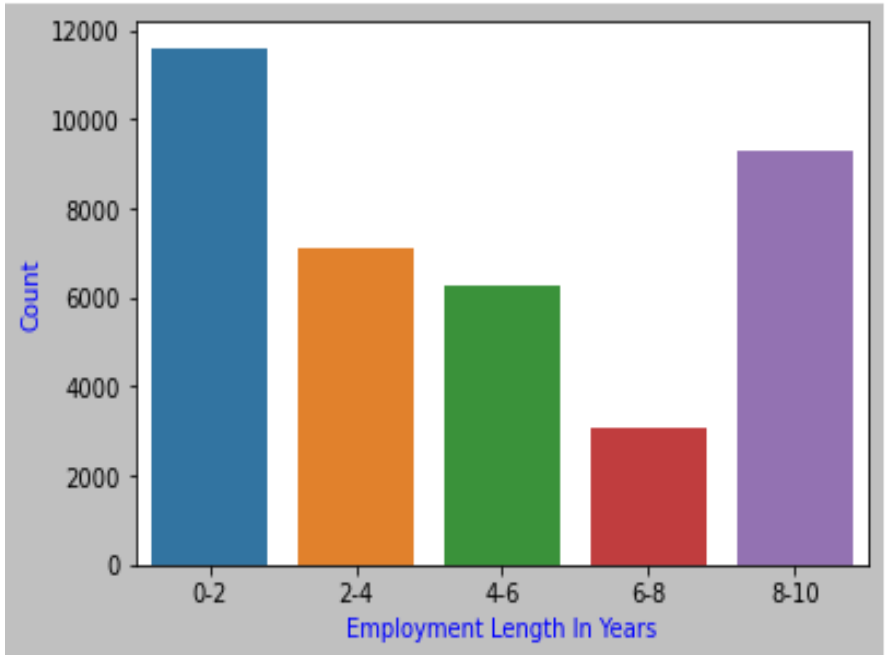
Univariate Analysis

Describe Data

- The distribution of the employment length variable, form which we can see the **10+ years** employment length are more as compared to other variables and the employment length **5 years** is the average years of the overall employment length.
- This is also a quantitative variable, we won't be having much of the provision to display the info on visual as most of the visual don't work with both of the values as numeric, so creating a bins for this seems better idea to me for categorising the data, doing so we'll have data in short and simple way.



Non - Grouped

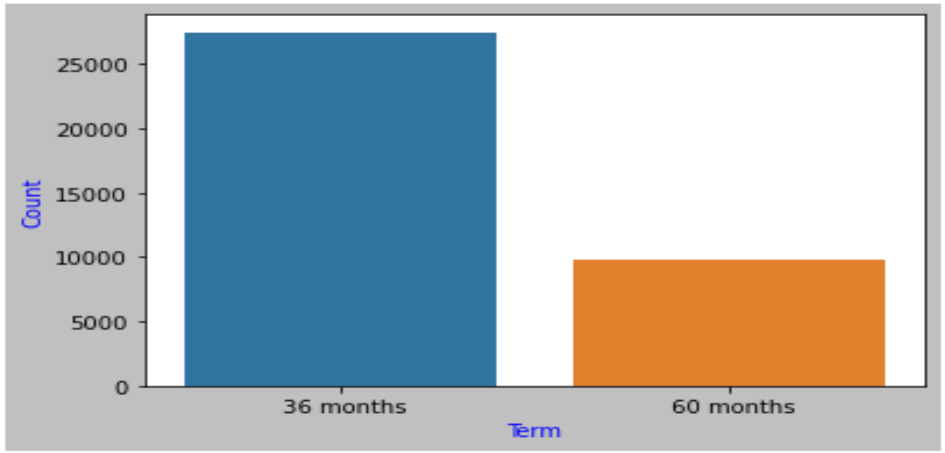


Grouped (Binned)

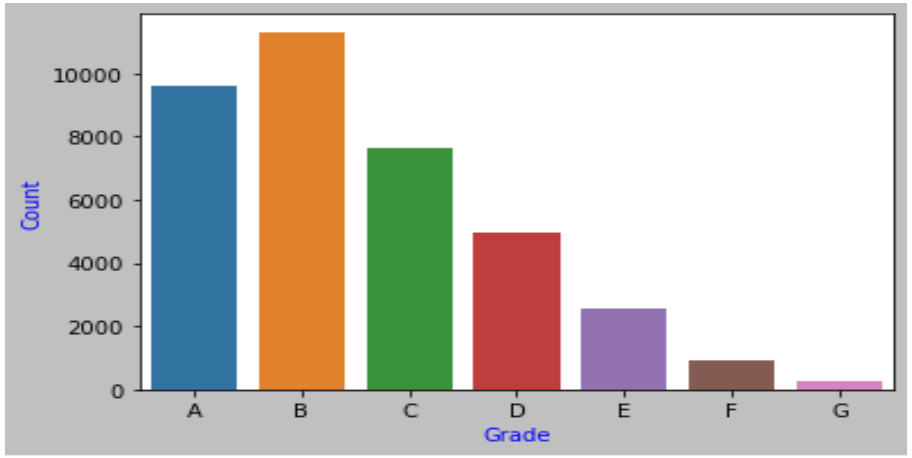
Univariate Analysis

Describe Data

- Going forward, tried to get the insights on some **Ordered categorical** data which seems to be important.



Term Distribution



Grades Distribution

- From the above visual, we can identify more number of loan application are applied for 36 months duration where only 30% approx. are applied for 60 months long term duration.

- And from the second, it shows more number are applicants are from B grade followed by A & C. Since higher the grade lesser the interest rate will be so from the graph we can assume most of the borrower has lower interest rate.

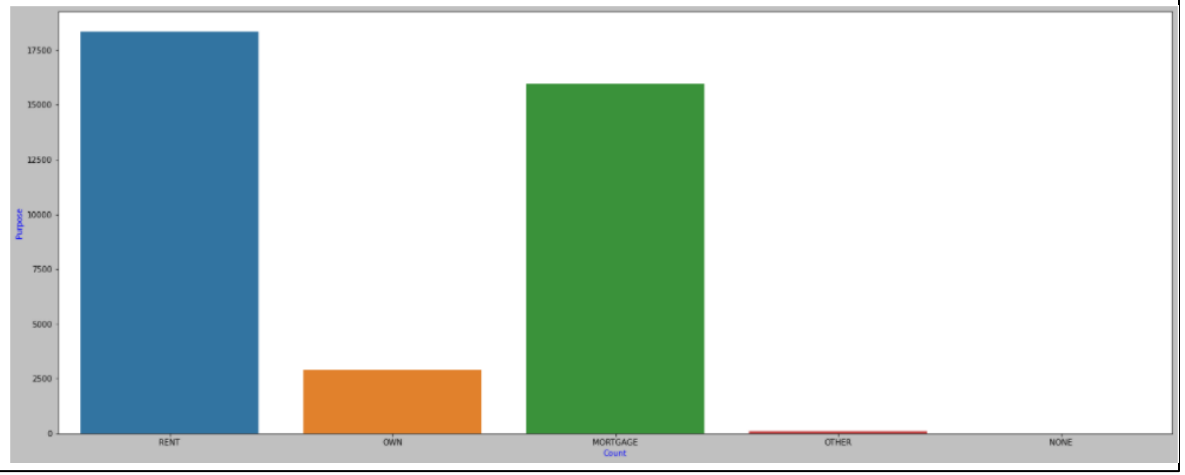
Univariate Analysis

Describe Data

- While studying the data, we found there were some columns which are focused on borrower and comes under **Unordered categorical** variables.

- Distribution of purpose with the spread across the loan amount
- Debt consolidation, credit card are likely purpose borrower mentions.

- Distribution of home ownership states the reliability of borrower's as we can see the Rented borrower has high count of loan could be for paying rent along with monthly expenses followed by Mortgage which does make sense for paying off some.



Bivariate Analysis

Categorisation

- After describing or analysing single fields values, lets deep dive into Bivariate analysis which will give us meaningful insights as we'll be dealing with two variables in this category.

- For proceeding to this, listed all the available fields and tried several combination to visualize the variables and show some insights but was struggling in some points like trying to show the comparison of two numeric value in bar chart such as DTI vs Annaul Income, Interest Rate vs DTI on the bar chart but was unable to do it as it require one of the axes in numeric format. We could have gone to other visuals too but pretty comfortable with bar charts.

- Workaround for such issue, created a bin for annual income and DTI, doing so will have the same data in the categorical format and can be with the multiple variables.

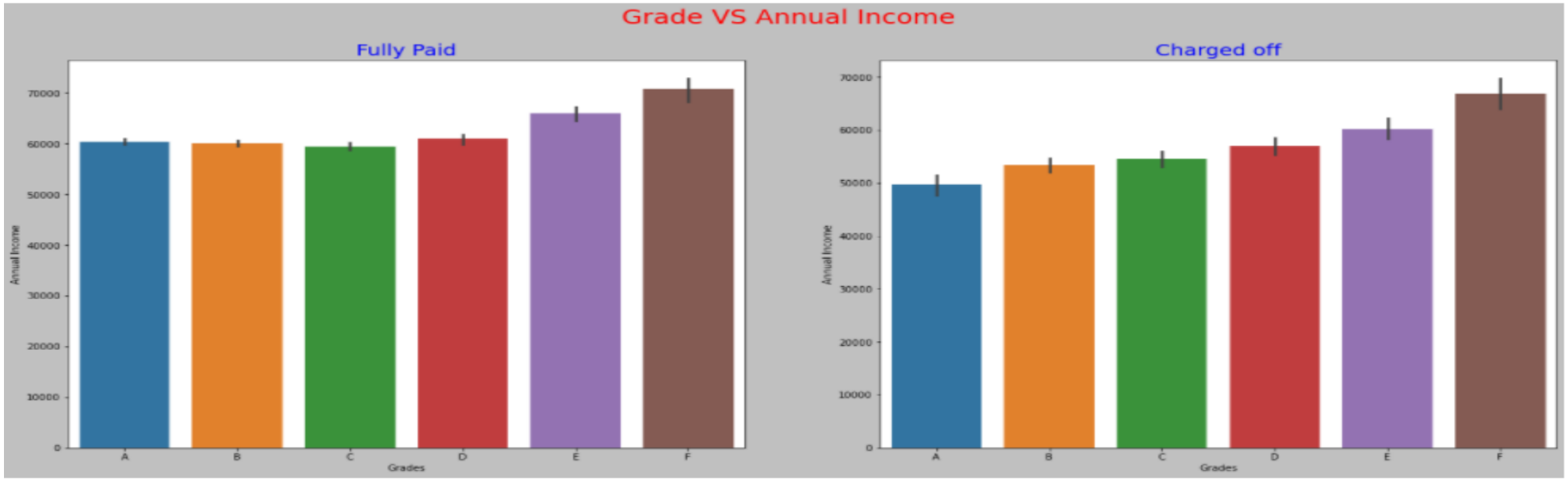
- Along with that, I thought of dividing the dataset into categories based on the status which are
  - paid_df :- which consist of all the column with 'Fully Paid' loans only.
  - chargedoff_df :- which consist of all the column with "Charged Off" loans only

Bivariate Analysis

Grade VS Annual Income

- The motive for creating two different dataset was to have a data which is more focus to one particular category and showing the comparison for these two different dataset will also give meaningful points to notice with respect to loan amount, interest rate, dti etc.

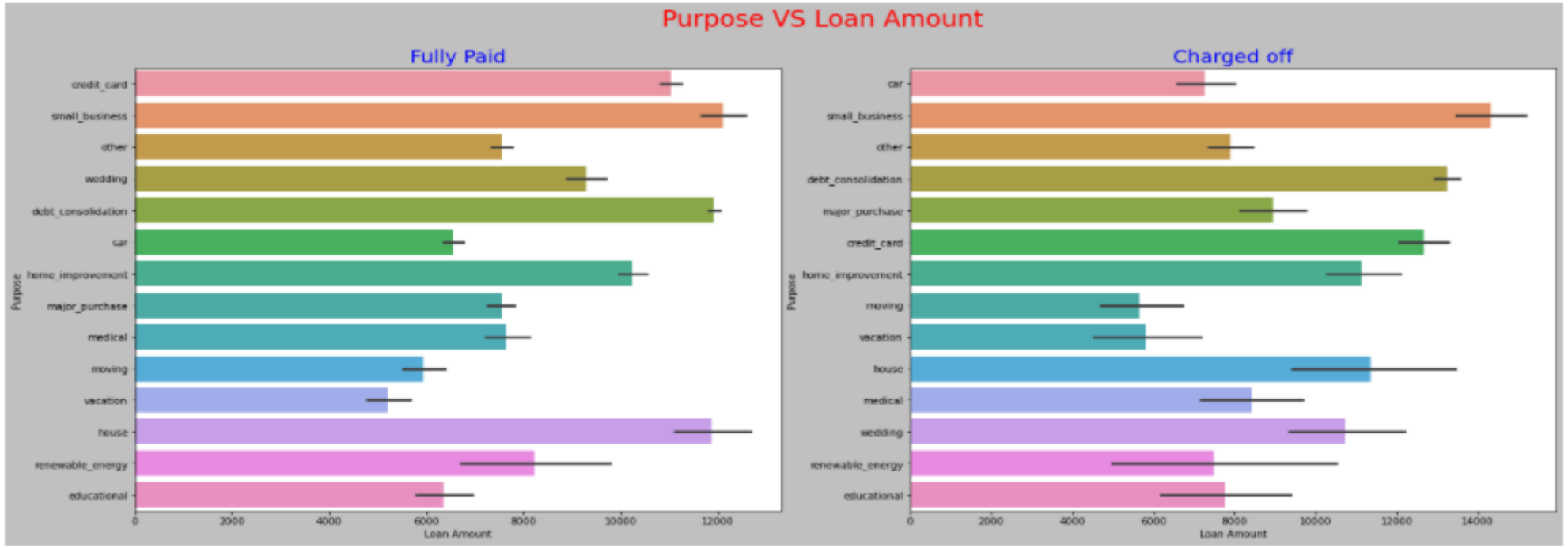
- Having said that lets try to compare some variable. Below is the snap of Grade vs Annual Income which tells us in Fully Paid visual Annual income are higher as compared to Charged Off in terms of Grades which shows the borrowers who has lesser income are likely to get charged off.



Bivariate Analysis

Purpose vs Loan Amount

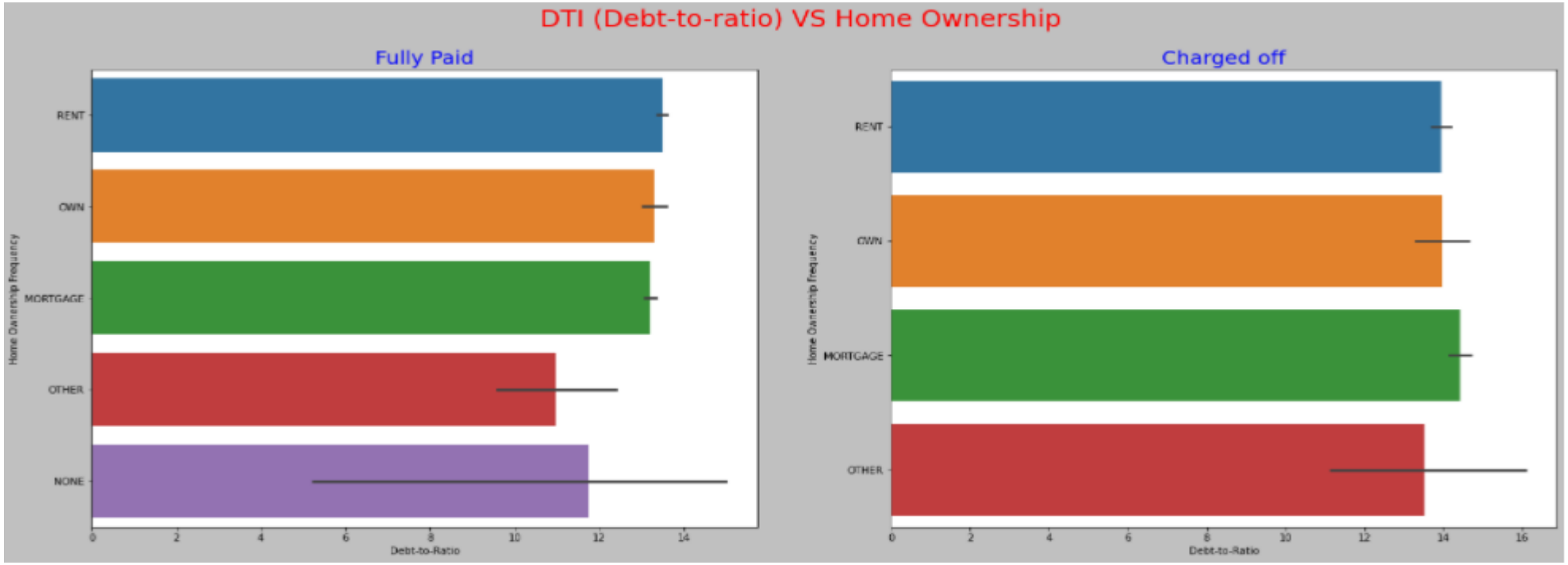
- From the below graphs it has been seen the loan amount sanctioned for small business, debt consolidation, house, credit card bills are higher in charged off category so we have hardcore data in our hand which is stating such purposes of borrower are likely to get charged off.
- Recommendation would, if the purpose of the loan is any one of the above category, then the loan should be sanctioned to only those borrower who has annual income higher, lesser or average DTI percentage, should have OWN or at least RENT home ownership.



Bivariate Analysis

DTI vs Home Ownership

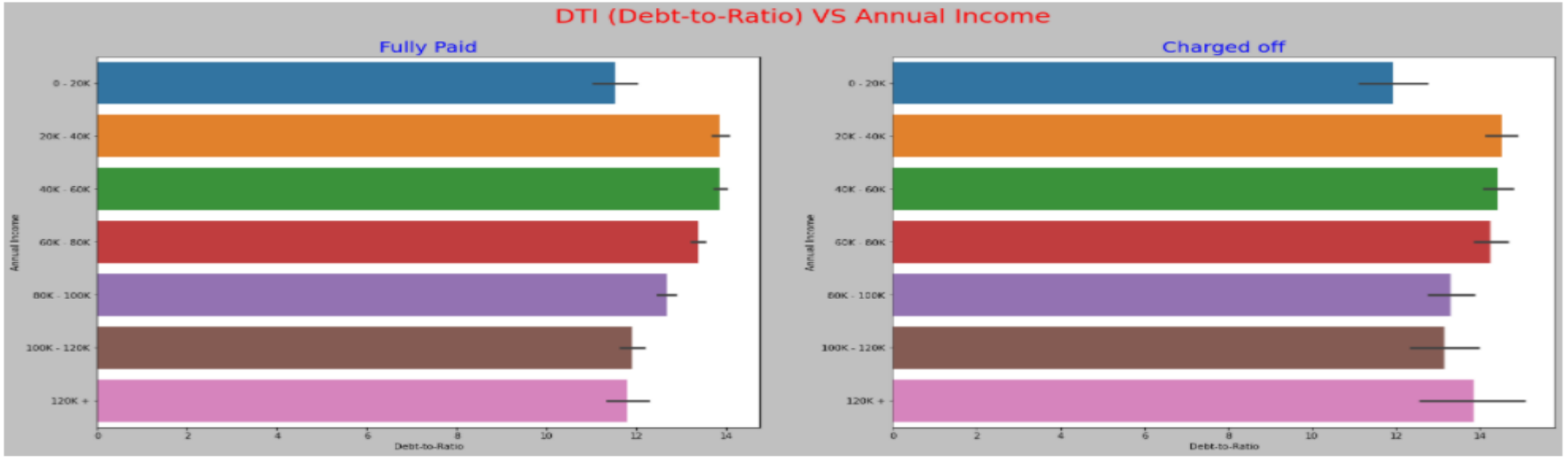
- The comparison of DTI and Home Ownership gives as the data that on the charged off category we have increase in DTI rate in each category which states the higher the DTI higher the chances of getting defaulter. As DTI is the component which is based on the monthly income and the monthly debt it has to be as much as low. The lesser the DTI, more the trust for lending loan.



Bivariate Analysis

DTI vs Annual Income

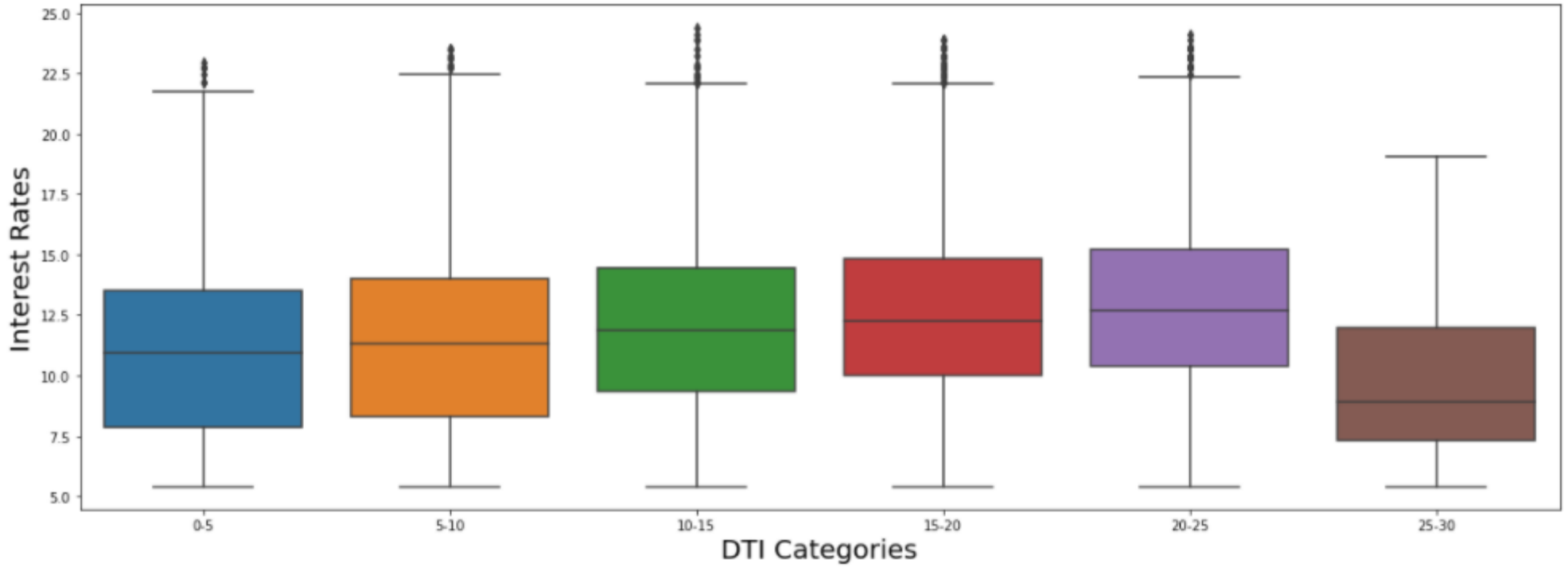
- When DTI compared with the annual income it clearly show for the borrower who has low annual income so ideally they should having higher DTI, considering with the little amount of income many thing to manage but with the higher income like 80K + borrowers category also has a higher DTI in charged off which doesn't seems much convincing, so such folks which has a higher income plus higher DTI can be avoided for lending loan.
- For such borrower who has higher income plus DTI, with the detail statement of monthly expenses, understanding the situation of the borrowers that why since working so many years he or she has higher number of monthly debts.



Bivariate Analysis

DTI vs Interest Rate

- Below graphs shows as the interest rate is getting increase the DTI is also getting increase but for the last category for DTI which 25-30 has been dropped, as there are less number of borrowers falls into that category. Apart from that we could state the borrower whose DTI is higher will have higher interest rate and if one has lower DTI rate will have lower interest rate.

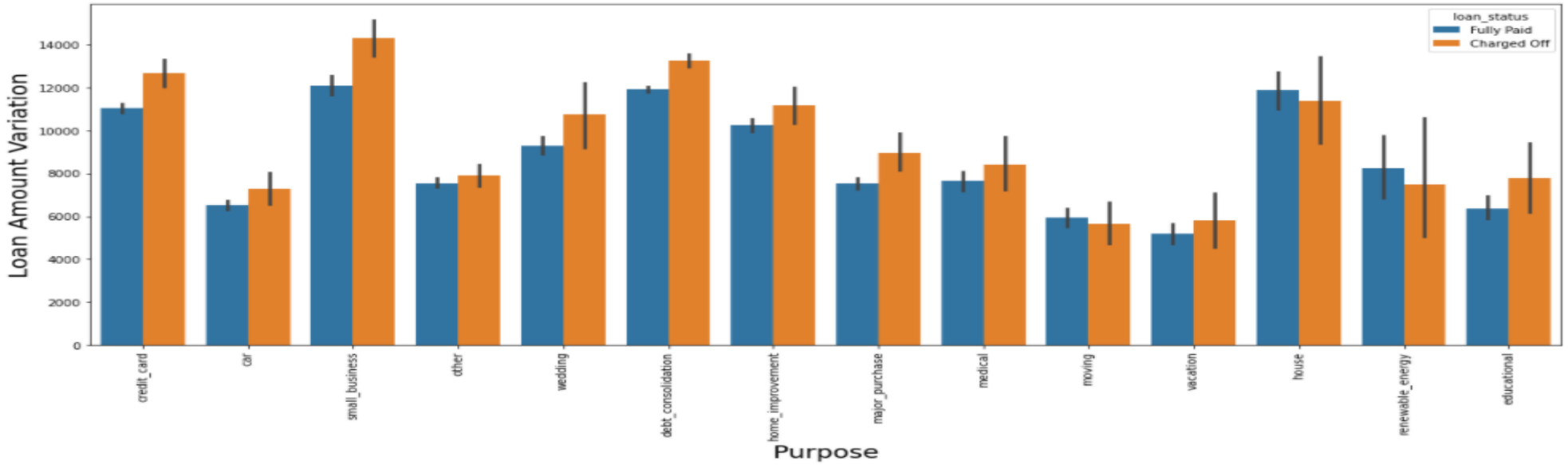


Multivariate Analysis

Purpose vs Loan Amount vs Loan Status

- Below is the comparison which has been made to see how the borrower is requesting for the loan, after plotting the visuals we could see the charged off category has some purpose which are higher than fully paid like the small business, credit card, major_purchases, debt_consolidation etc. so for those there are high chances for making it to defaulters.

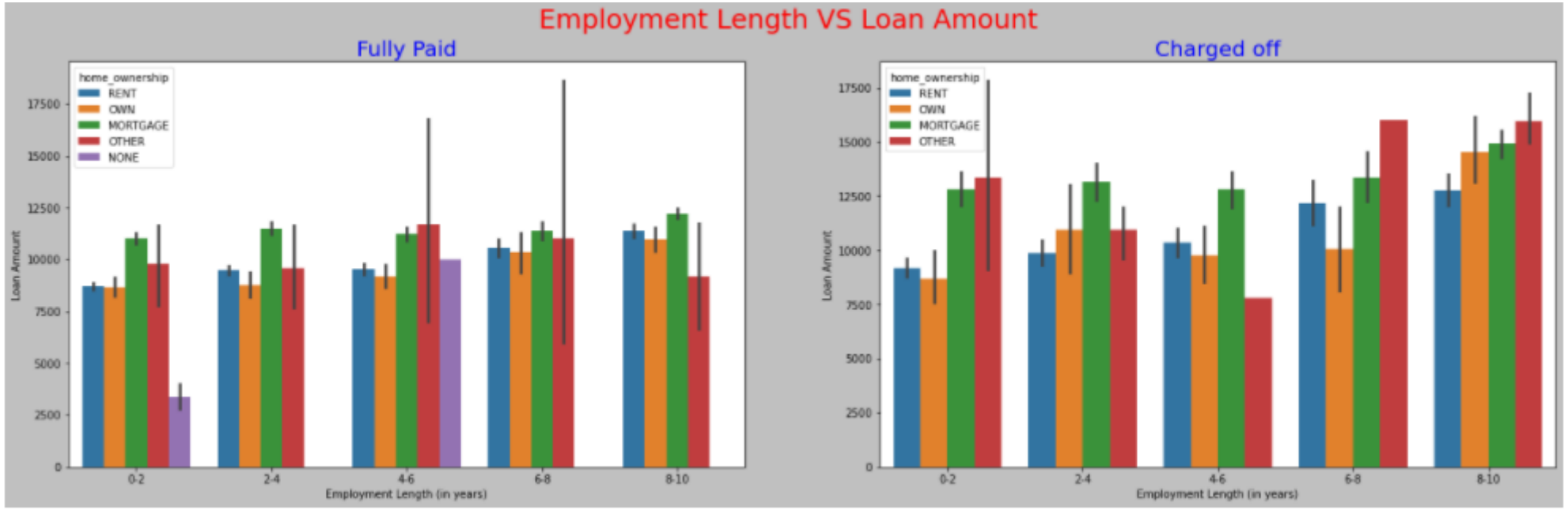
- As we saw in the previous visual the same factors were showing high risk of getting charged off so we can say by looking at both the visuals for such purposes some hard background check is needed and no compromise can be made



Multivariate Analysis

Employment length vs Loan Amount vs Loan Status

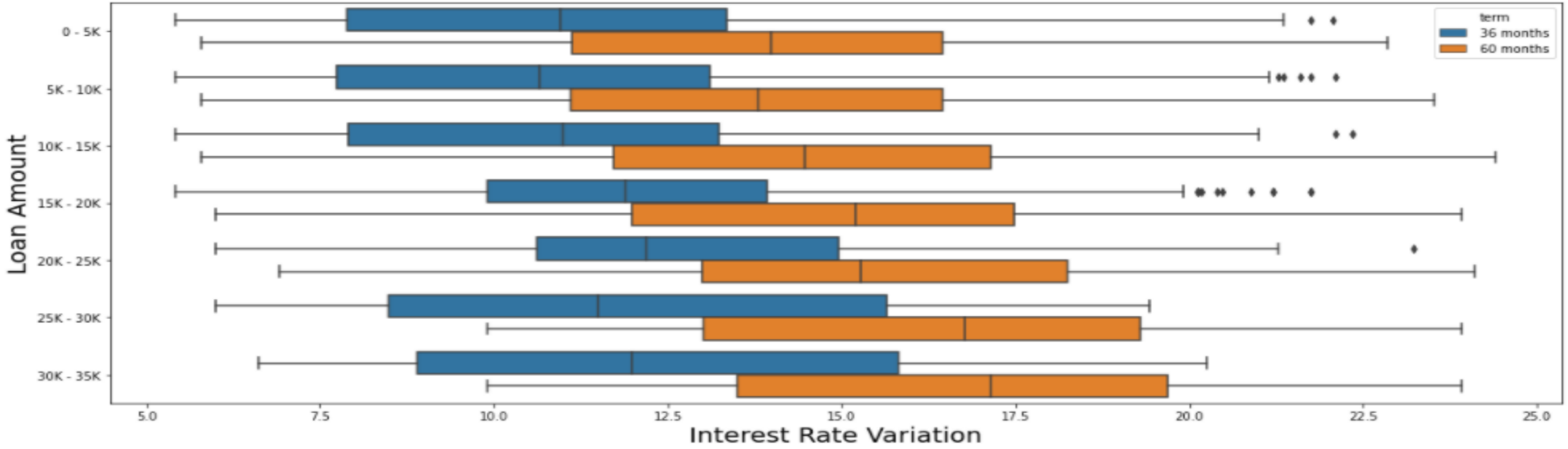
- Below visual shows the comparison of Fully Paid and Charged Off category where it looks the numbers are high in ChargedOff category.
- Mortgage rate and Other is high in Charged Of category for every emp_length, so there is high chance of these categories borrower making to defaulter or Charged off category.
- Since the borrower is already having the mortgage, Rent or could be several others personal instalment over his head, so there would have been chances to miss out the instalments if the borrower's still struggling on financial situation.



Multivariate Analysis

Loan Amount vs Interest Rate vs Term

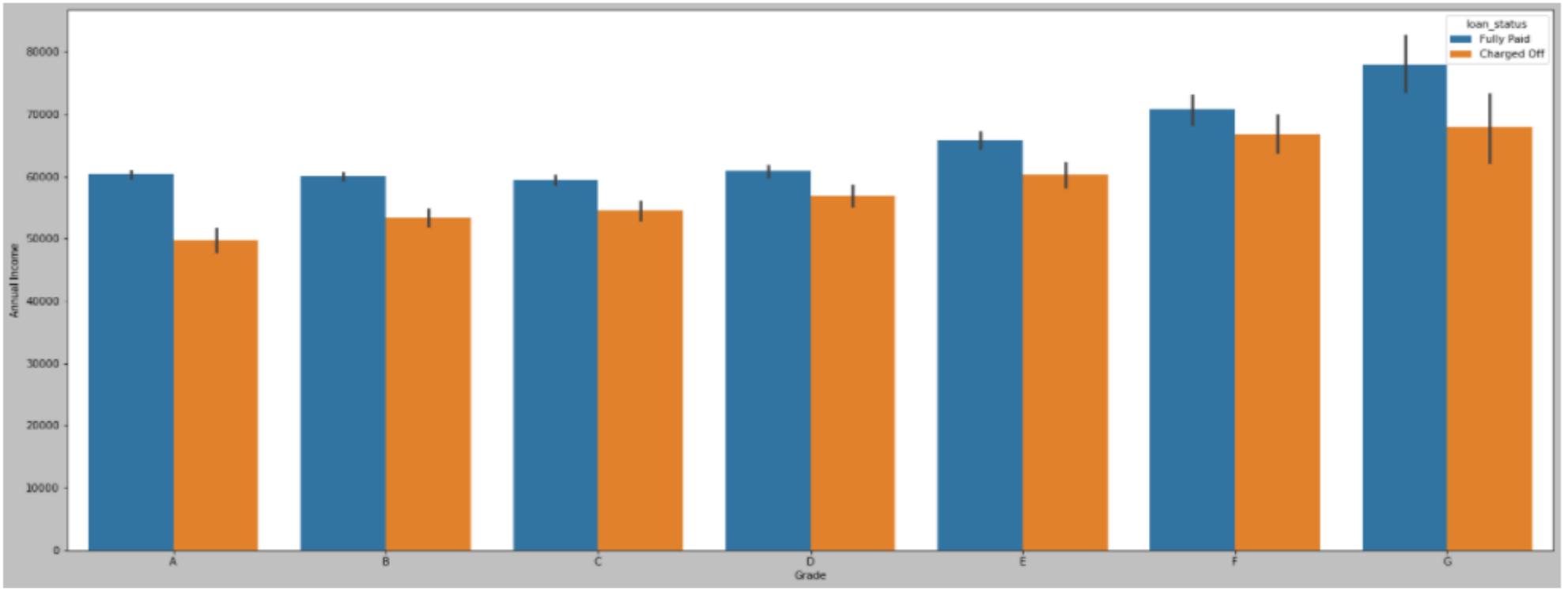
- The below graph represents the comparison of interest rate and loan amount category, as from the below graph we can see the interest rate is getting increase as the loan amount getting increased, so there is a strong positive correlation between these 2 variables. (fig below)
- Recommendation would be, if the person is going to apply for the loan for longer period of time it is suggested that borrower should have higher income rate and lesser DTI since the interest rate would be higher for such loan as a lender we should be making sure that borrower is capable of returning the loan.
- The loan amount and interest rate has a strong positive correlation since the loan amount is getting increased the interest rate is also increasing so now the below chart shows if we have the loan for a longer period of time it could lead us to have a higher interest rates too



Multivariate Analysis

Grades vs Annual Income vs Loan Status

- The comparison of annual income and grade with loan status clearly shows the charged off category has less annual income which make them more likely to get defaulters.
- Since the interest rate gets higher as the Grades proceeds, if the borrower has lower income and falls into end grades like F or G, he might gets the higher interest rate on his/her loan amount which in turn causes the monthly instalment to get higher.

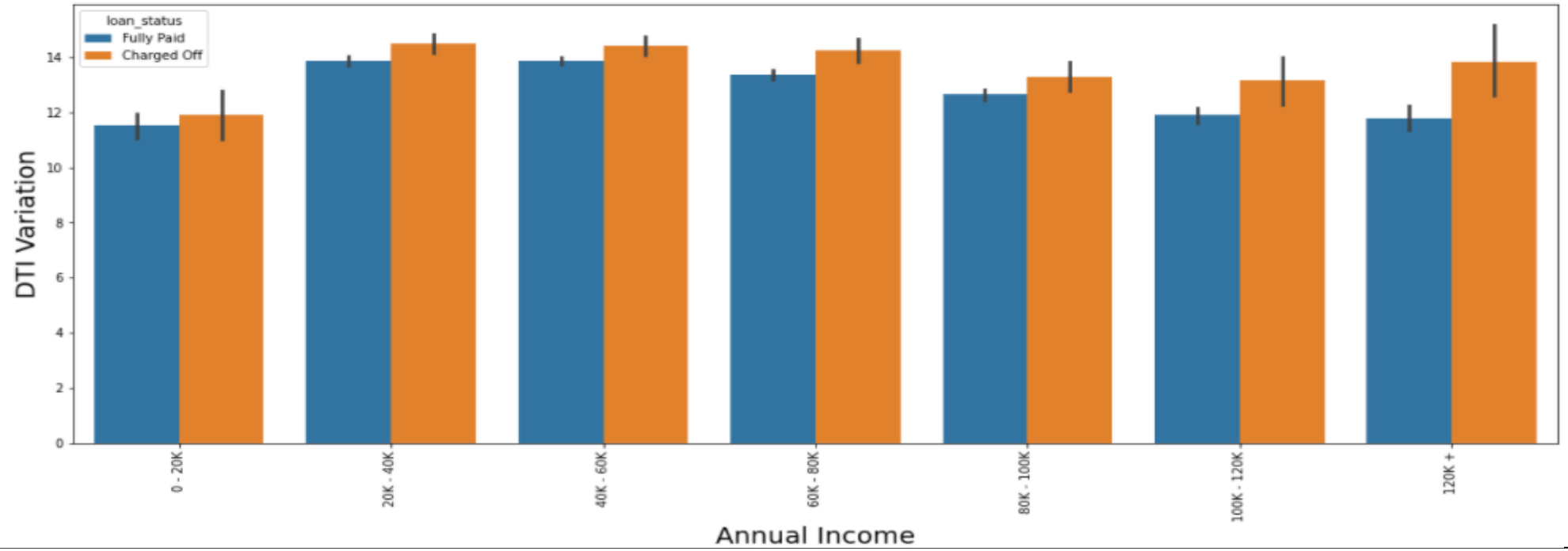


Multivariate Analysis

Annual Income vs Debt-to-Ratio vs Loan Status

- The below graphs show, the DTI percentage is higher than annual income and most of them falls into Charged Off category which have been seen on the earlier visual too.

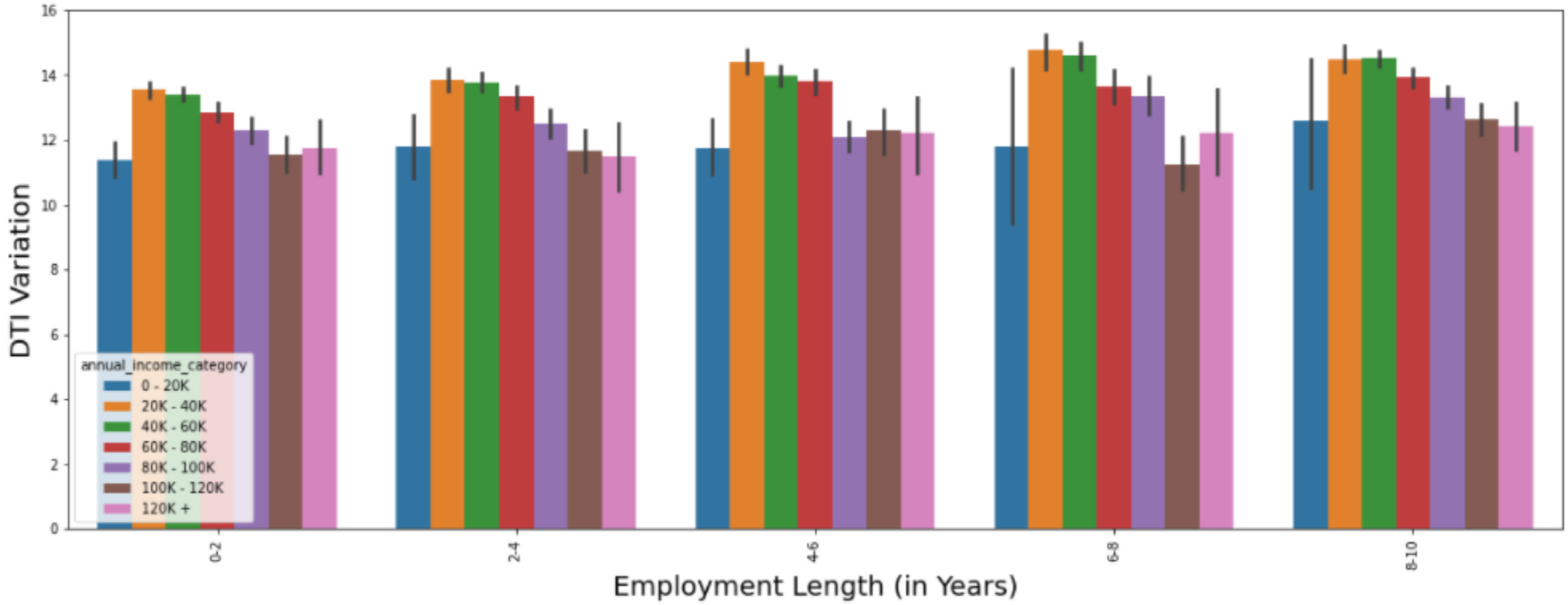
- Those who has lower income, the graph does make sense since lower income mean lesser amount to manage monthly expenses, but it has been seen those who falls above 1 mils still have higher DTI which is less convincing so might need to check thoroughly such borrower would be my recommendation.



Multivariate Analysis

Employment Length vs Debt-to-Ratio vs Annual Income

- The below visual shows that even with the employment length is higher on each category of the annual income there has been an increased, so we could say that those who has higher income will have higher DTI rate even though they have good amount of employment experience.



Recommendations

Points to look out for

- After going through all the metrics and analysing every fields singly and with combination of two or more fields together some recommendation can be followed. They are listed below :-
  - Borrower who has **lower income**, can be checked in which grades they fall into if the **grades are lower** and **higher DTI** is there, such application can be avoided considering going further such borrowers might find struggle to repay the credit.
  - If the purpose of the loan falls into **Small Business, Debt Consolidation, Credit Card Bill Payment** category, then the loan should be sanctioned to only those borrower who has annual income higher, lesser or average DTI percentage, should have OWN or at least RENT home ownership.
  - If the person is going to apply for the loan for longer period of time i.e. **60 months**, it is suggested that borrower should have **higher income rate and lesser DTI** since the interest rate would be higher for such longer loan.
  - Borrowers which falls into **Mortgage, Others** category should have **lesser DTI** as we have seen there has been in increase of DTI in every category of home ownership when we checked it for Charged Off. Those who have higher DTI and falls into above two category, if the detail background check professionally and the capability of borrower if found convincing loan could sanctioned the for such borrowers.

Recommendations

Points to look out for

- Annual income plays vital role as based on that DTI varies, and from the analysis it has been seen those who has higher annual income also has the higher DTI percentage which even with the employment length of 7 to 10+ years when compared to less employment length of 4 to 6 years. So for such borrowers, details monthly expenses distribution can be looked to check how they have distributed the their income.

Conclusion

- By studying the data, we could there will be borrowers who will have higher income with higher DTI, could have purpose of Small Business, Debt Consolidation with lower annual income, could have much of the employment experience but still have higher debts to pay, could have mortgage home ownership with higher DTI etc such borrower cant be neglected completely, since that'll be loss for the lender's credit. Lender can sanctioned loan with the details background check of the borrowers in terms of professionally, personally and also check the back up of the borrower and post doing all if lender gets enough confidence about repaying the loan amount then loan could be sanctioned.

End

Thankyou