**Method followed for EDA:**

1. Descriptive analysis of all the features to be done i.e both the numerical and categorical.

2. At first analysis of numerical category feature is done i.e between numerical vs numerical

3. Then analysis is done between the categorical feature:

In order to show association between two categorical variable we can use a contigency table. The contingency table is a display for two categorical variables. Its rows list the categories of one variable and its columns list the categories of the other variable. Each entry in the table is the number of observations in the sample at a particular combination of categories of the two categorical variables.

4. Then analysis is done between response variable(Target variable) and categorical variable

5. Finally between response variable(Target variable) and numerical variable is done

***Task-1:***

**On Numerical features alone:**

**1. Months\_loan\_taken\_for:**

*- majority of the loan taken durations are less than 25 and almost 90 percentage of values present between 4-36 and rest are in between 36-72*

*- it is positively skewed data .*

**2. Principal\_loan\_amount:**

*- max value is almost 1crore 84 lakhs but the 95 percentage of amount is less than 91 lakhs i.e less than half of max amount*

*- 32 lakh 71 thousand is the mean loan amount taken*

*- only roughly 5 % of amt taken are more than 91 lakhs which is low*

**3. EMI\_rate\_in\_percentage\_of\_disposable\_income:**

*- It has a very less spread of values*

*- almost half of the total applicant have 4 % of disposable income as EMI\_rate and 4 % is the max value so all the loan applicants have less than or equal to 4 % of disposable income as EMI\_rate*

**4. Has\_coapplicant:**

*- we can say than less than 95 % of people don't have coapplicants and roughly 3 percentage of the applicants have coapplicants*

**5. Has\_guarantor:**

*- only 5 % of people have guaranteer for the loan which they have applied*

**6. Number\_of\_existing\_loans\_at\_this\_bank:**

*- this value ranges from 1-4*

*- and 95 % of applicants have less than 2 loans at this bank and 99 % have less than 3 loans at this bank.*

**7. Primary\_applicant\_age\_in\_years:**

*- ranges from 19-75 years of age*

*- 50 % of applicants are less than 33 years of age*

*- 95 % of applicants are less than 60 years of age , implies 5% of applicants are from 60-75 years of age*

**8. Number\_of\_dependents:**

*- all applicants have atleast 1 and atmost 2 dependents*

*- only 10% of applicants have 2 dependents*

**9. Years\_at\_current\_residence:**

*- this value ranges from minimum of 1 to maximum of 4*

*- 50 % of the people have stayed in current residence atleast 3 years and 25 % of the people have stayed in 4 years.*

**10. Foreign\_worker:**

*- more than 95% of people(applicants) are foreign workers.*

**On categorical features alone:**

1. Purpose of loan for electronic equipment is 28.34% and new vehicle is 23.68% which constitutes more than 50%

2. No of people who are

- critical/pending loans at other banks are 29.3%

- delay in paying off loans in the past are 8.8%

3. 71.3% have their own house

4. 77.8% are skilled/employed/qualified/officials and rest are unemployed/unskilled

5. 73.8% have low Savings\_account\_balance

**On numerical vs target variavble:**

1. Months\_loan\_taken\_for , Principal\_loan\_amount:

- All the descriptive variables(mean,25,50,75th,std\_deviation) for applicants who defaulted is found to be higher than non defaulted ( things to be noted: since the size of non defaulted size is less difference might not be statistically significant)

2. Primary applicants who defaulted have mean age of 33.97 and who did not defaulted have 36.24. This doesn't mean people with less age have more probabality of getting defaulted .

3. Years\_at\_current\_residence did not have any difference on the target.

**categorical vs target variable:**

1. People who bought loan for career development have higher chance of default that is 88.9%

2. People who owns real estate have default rate of 78.72%

3. People who have critical/pending loans at other banks have default rate of 82.94%

4. skilled employee / official & unskilled - resident have default rate of 70.48% and 72% respectively

3. People with high saving bank account balance have higher chance of default i.e 82.54%--(since this feature is severely imbalanced less weightage has to be given to this observation )

4. People who are employed for atleast 4 years and 7 years have default rate of 77.6% and 74.7% which higher compared to 0 and 4 years.

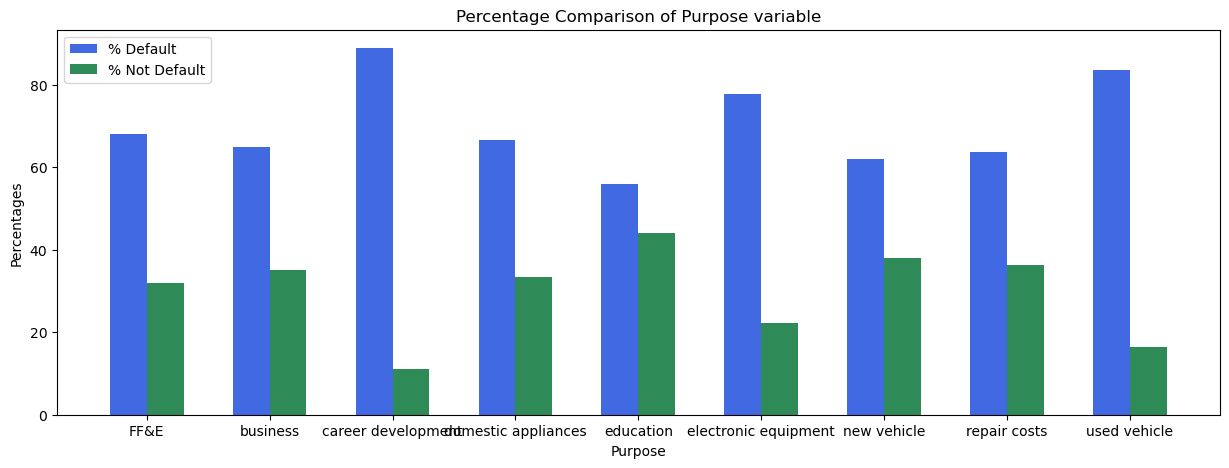
*—---------------------------------------------------------------------------------------------------------------------------*

**Q2. How would you segment customers based on their risk (of default).?**

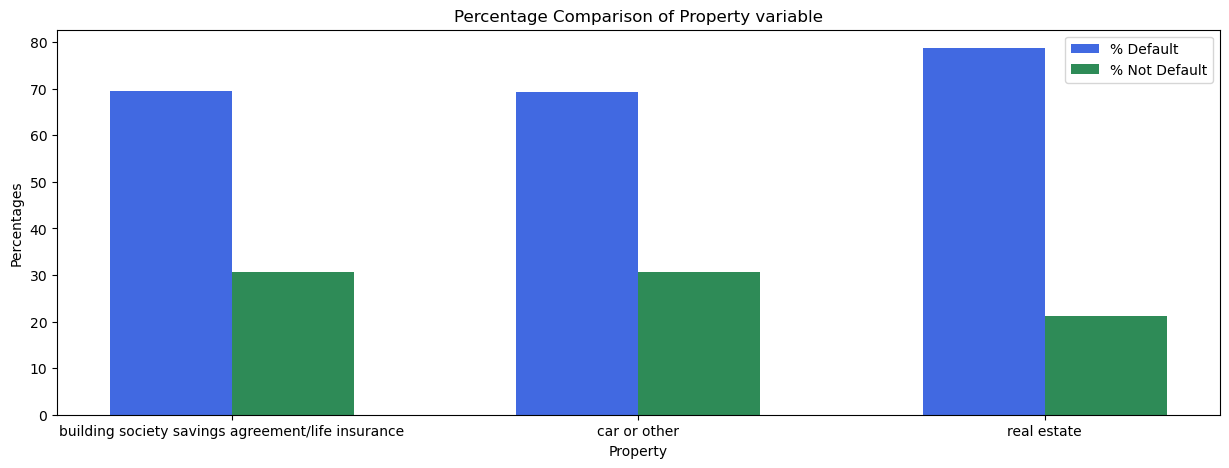
Ans- By data analysis important feature which determines default rates are

Customers having loan for :

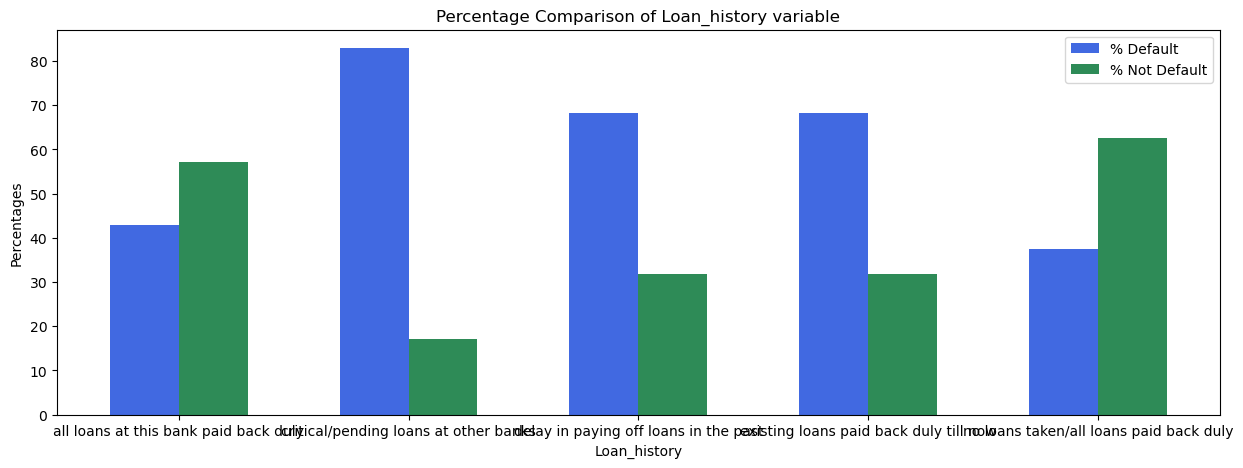
***a)career development (88.9% default)***



***b)real estate ( 78.72% default )***

******

***c)critical/pending loans at other banks (82.94% default )***

******

***d)people having less age have more probabality of getting defaulted .***

**Q3.Which of these segments / sub-segments would you propose be approved?**

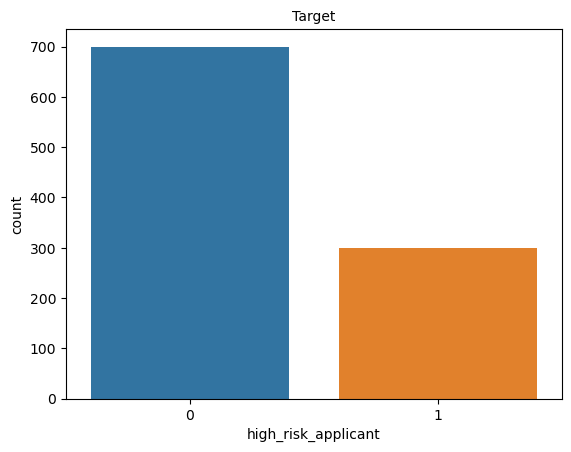
Ans-

* Numerical features are not statistically significant to propose any solutions.
* People who have paid back back the loans or no due amount of loans left and no loans taken have very low chance of getting default(37.5% which is minimum among all). — so these categories should be given a higher priority .

**Q4.Tell us what your observations were on the data itself (completeness, skews).**

Ans-

* As observed from the bar graph data have 30-70 split which is slightly imbalanced.
* It has positive skew towards non-default rates which is understandable since in real world ideally there are less default rates .



**Interesting observation:**

* Roughly 78% are skilled/employed/qualified/officials so data is imbalanced largely on that feature so we have gotten majority of defaults in employed/qualified professionals.

**Data in a story format:**

People who are relatively young and taken a career development loans and owns a real estate property have higher chance of getting defaulted.