



CASE STUDY: LENDING CLUB

Kompal Chaudhary



Table of content

Data Understanding

1

Data Analysis

3

Segmented Univariate
data analysis

5

Data Cleaning
(Missing value
treatment and
data-type validation)

2

Univariate data analysis

4

Conclusion

6

A decorative pattern of hexagons in various shades of blue and teal on the left side of the slide. Some hexagons contain icons: a lightbulb, a thumbs-up, a network node, a smartphone, a magnifying glass, a gear, and a speech bubble.


1

DATA UNDERSTANDING



Data Understanding : loan.csv


As per the given dataset:

- 
1. The shape of data is 39717, with 111 columns.
 2. Important columns in the dataset: loan_amount, term, interest rate, grade, sub grade, annual income, purpose of the loan etc.
 3. Here the target variable, is loan_status, i.e which is compared across all the independent variables.
 4. Analyse data by comparing the mean default_rate across various independent variables . Thus, depicting most affecting variables.

A decorative graphic on the left side of the slide. It features a large, light blue hexagon in the center, surrounded by several smaller hexagons in various shades of blue and teal. These smaller hexagons contain white icons: a lightbulb, a thumbs-up, a smartphone, a magnifying glass, a gear, and a speech bubble. There is also a small network diagram icon with a central node and five connecting lines.

2

DATA CLEANING



Data Validation and Treatment

As per the given dataset:

1. There are approximately more than 50% of the columns, which have more than 90% missing data
2. All such columns are removed from the data. Also, columns such as description and months since last delinquency are also removed, since while
3. There is data-type altered for columns such as interest rate and employment length

A decorative graphic on the left side of the slide. It features a large, light blue hexagon in the center containing the number '3'. Surrounding this central hexagon are several smaller hexagons of varying shades of blue and teal. Some of these smaller hexagons contain white icons: a lightbulb, a thumbs-up, a smartphone, a magnifying glass, and a gear. There is also a network-like icon with a central node and several connecting lines. The entire graphic is set against a dark blue background.

3

DATA ANALYSIS

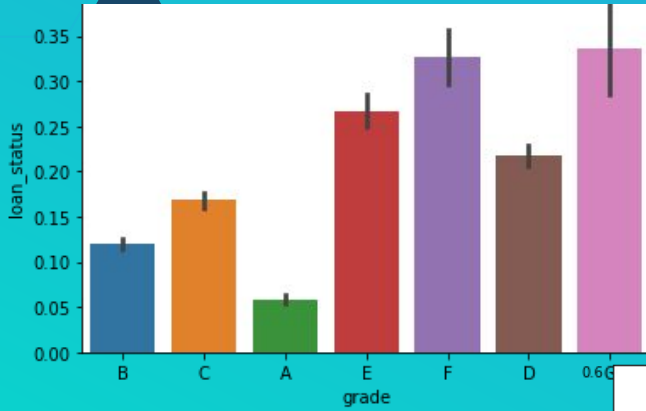


Feature Selection

As per the given dataset:

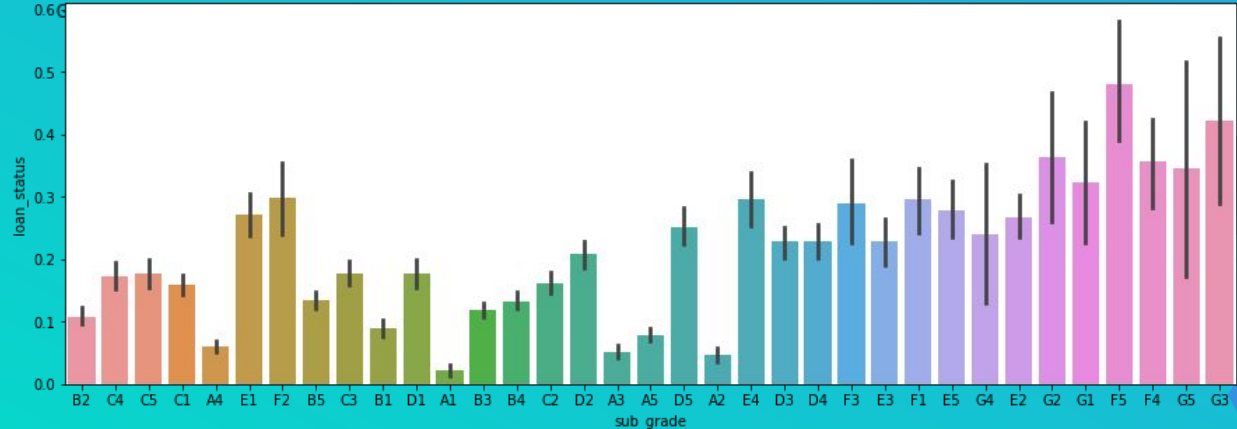
1. Here, we have broadly three types of variables -
 - a. Demographic variables such as age, occupation etc
 - b. Loan characteristics such as loan amount, interest rate, purpose of loan etc.
 - c. Customer behaviour variables such as delinquent 2 years, revolving balance, next payment date etc. As per nature of variables, these are generated after the loan is dispersed.
2. Here, since customer behaviour variables are not present at the loan application/ approval time, therefore these cannot be utilised as credit-approval predictors. Thus, utilising Demographic variables and Loan characteristics variables.

Univariate Analysis : Categorical variables

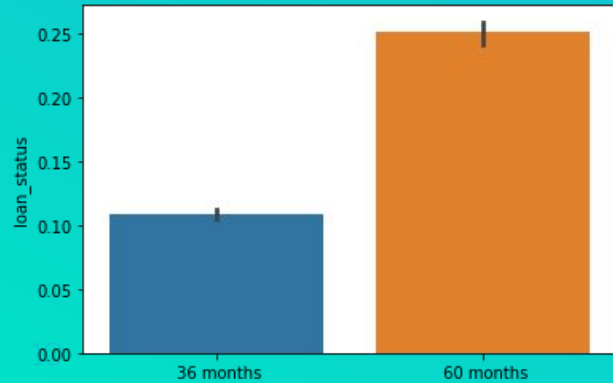
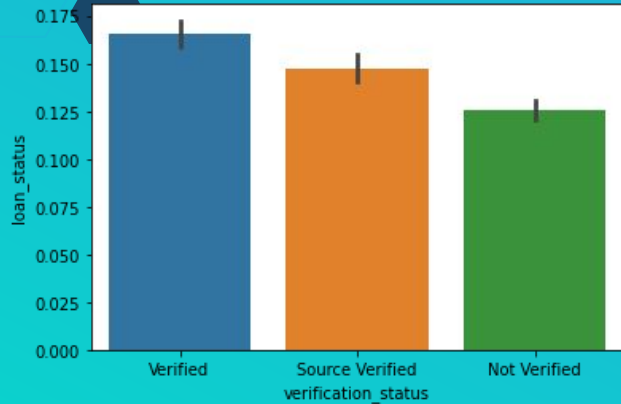


As per the figure:

- ❖ The Defaulter rate decreases as the grade and subgrade increases.
- ❖ This is due to the fact that grade depicts the likelihood of risk of not paying the loan by customer

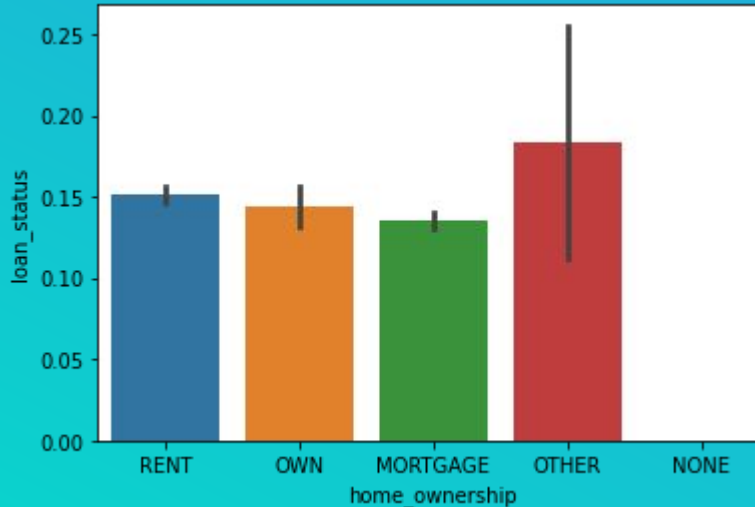


Univariate Analysis : Categorical variables

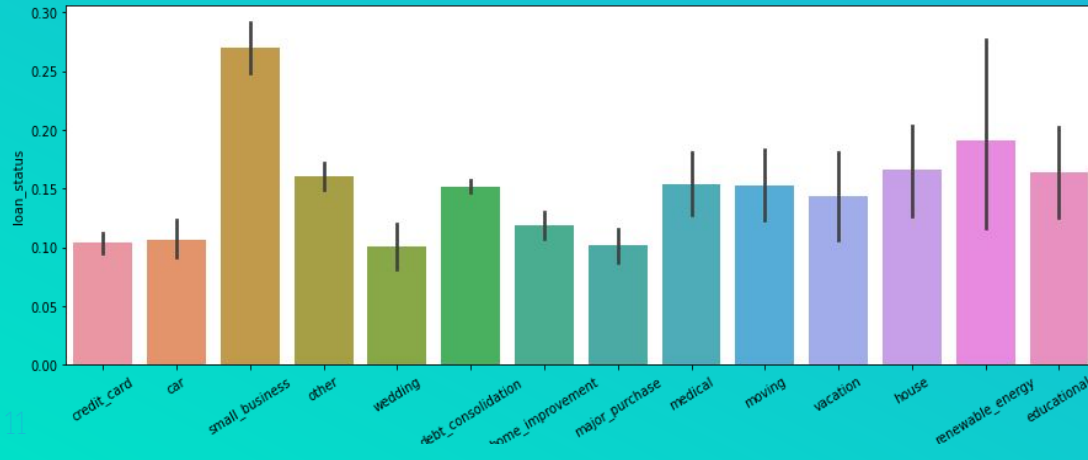
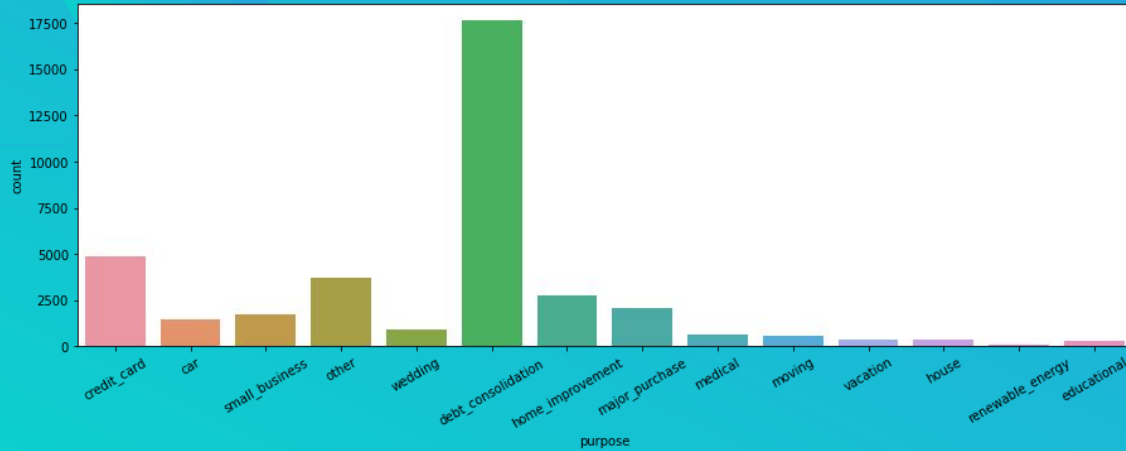


As per the figure:

- ❖ Surprisingly, the verified customers are more defaulters than not verified customers
- ❖ The default rate also increases if the loan is long term
- ❖ Also, the customers living on rent are more likely to default than customers owning house



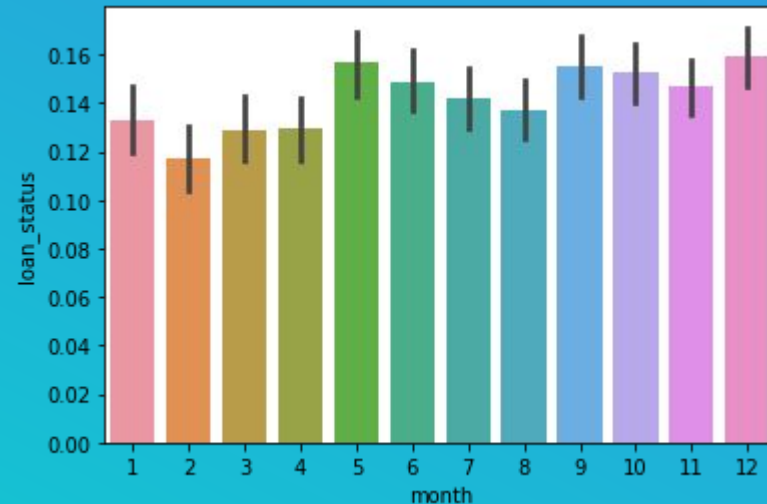
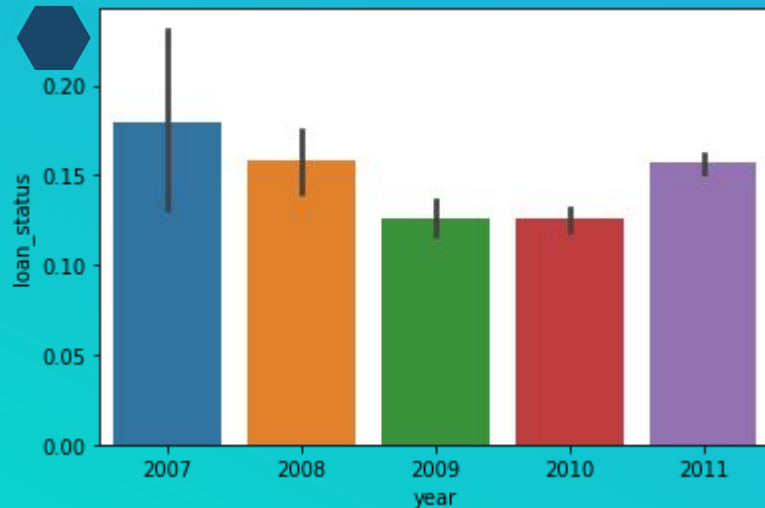
Univariate Analysis : Categorical variables



As per the figure:

- ❖ As far as the purpose of the loan is considered, people are taking more loans for debt consolidation than other categories combined together.
- ❖ There are least number of loans taken for the purpose of house, renewable energy or education.
- ❖ There are more defaulters who take loans for the purpose of small business. That also, depicts the risk factor in starting small businesses.
- ❖ The customers taking loans for renewable energy, education and house are next most defaulters.

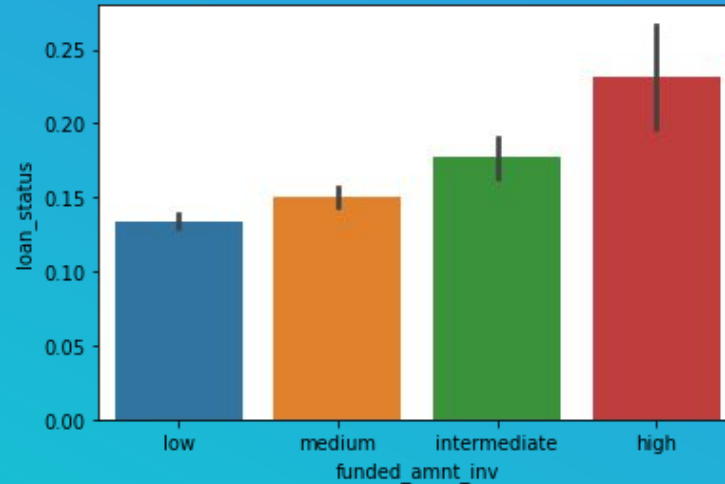
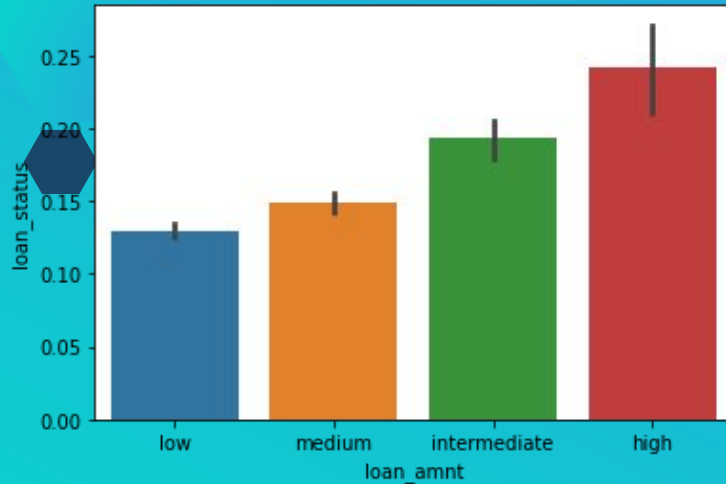
Univariate Analysis : Derived Metrics



As per the figure:

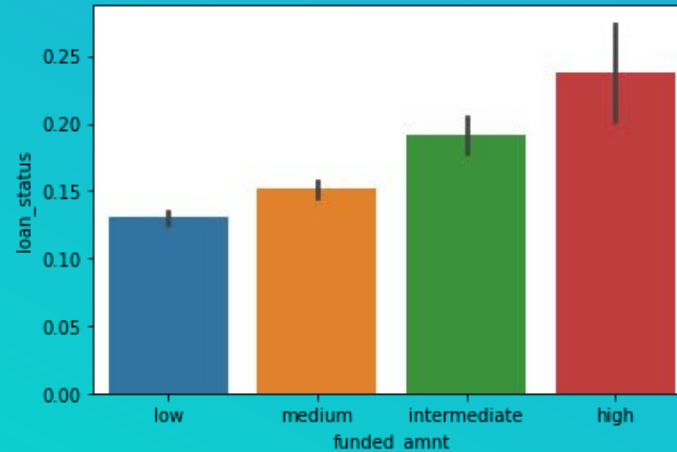
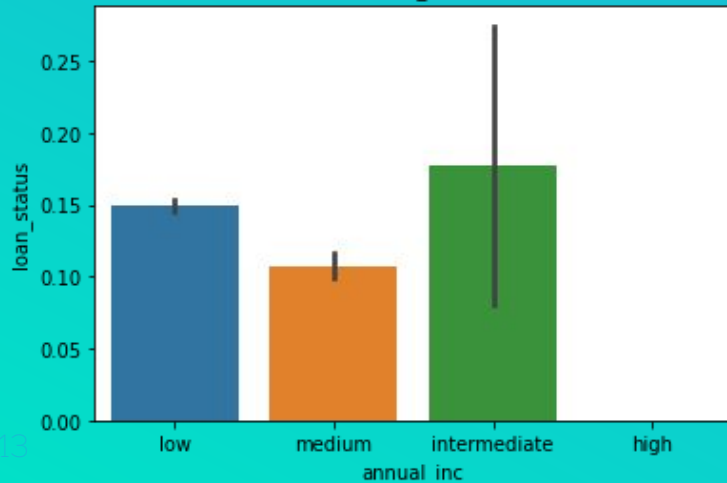
- ❖ There are some derived variables that can be picked from date column: such as month and year
- ❖ There are more default-rate for loans in 2007 as comparison to later years.
- ❖ Also, there is not much effect of the default-rate on the months of the year. There are very slight variations.

Univariate Analysis : Continuous variables

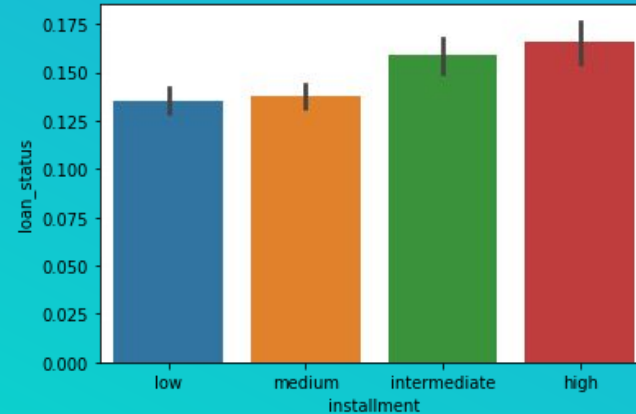
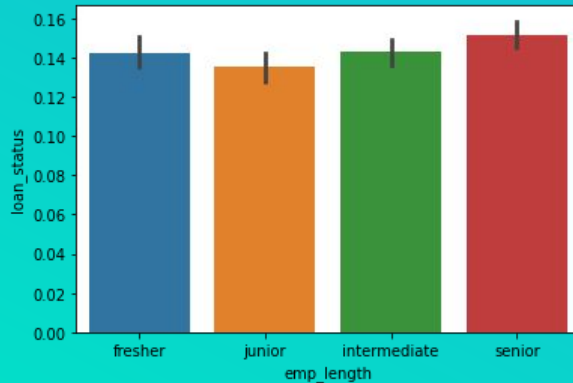
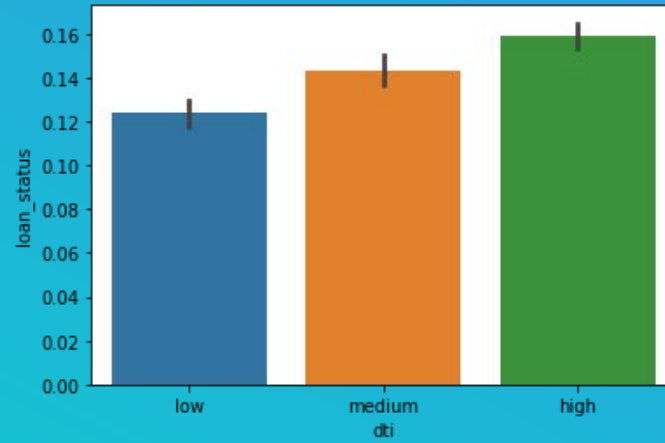
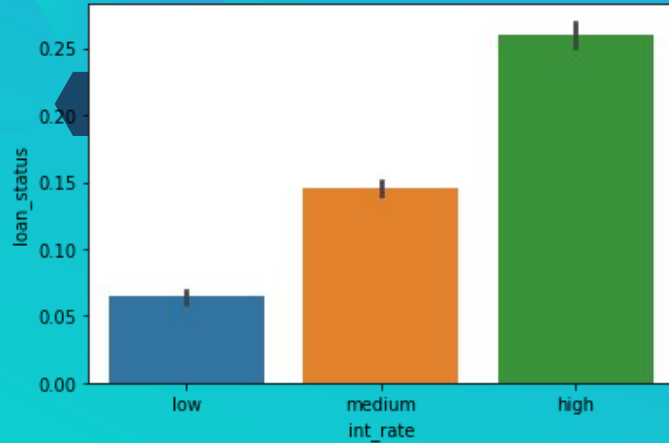


As per the figure:

- ❖ The default-rate is directly proportional to the loan amount, funded amount, funded amount investor and annual income
- ❖ Performed Binning on categorical variables



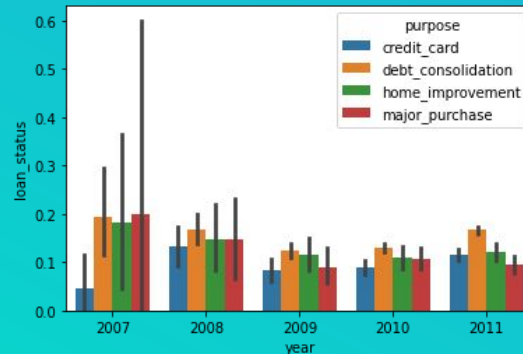
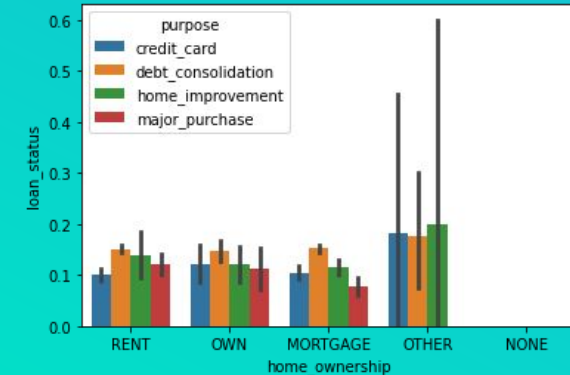
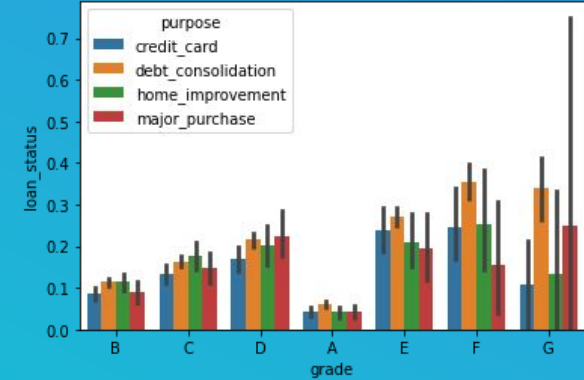
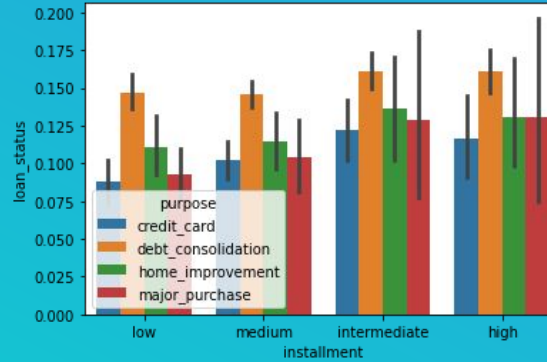
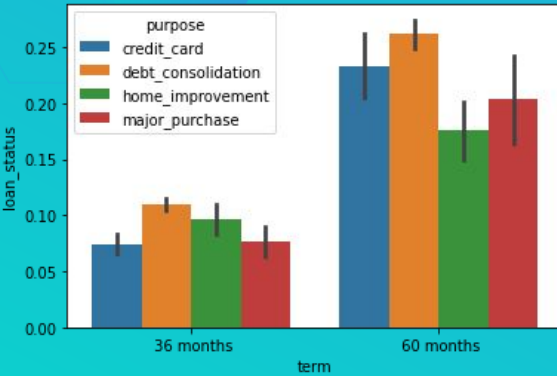
Univariate Analysis : Continuous variables



As per the figure:

- ❖ The default-rate is also directly proportional to the interest rate, dti, and installment
- ❖ Though, considering the length of the employment, freshers and senior are more likely to default a loan payment.

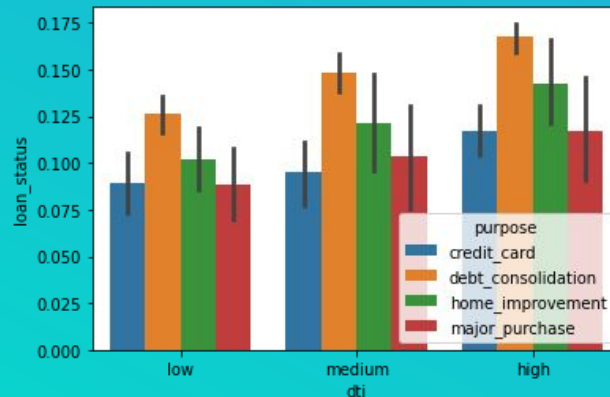
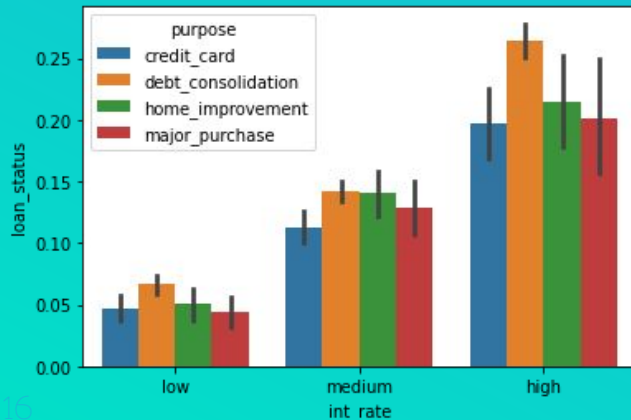
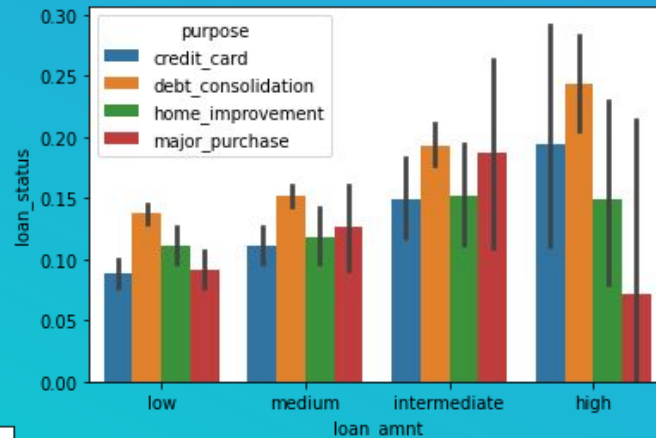
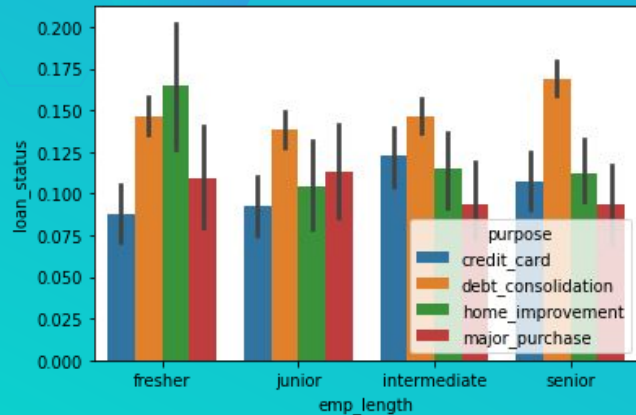
Segmented Univariate Analysis : wrt Purpose



As per the figure:

- ❖ For segmented univariate analysis, we are looking into 'purpose of the loan', since it effects- applicant-type, interest-rate, income, and thus the default-rate.
- ❖ Thus, all categorical and continuous variables are validated and analysed

Segmented Univariate Analysis : wrt Purpose



As per the figure:

- ❖ Generally, debt_consolidation loans does have very high default rates across all categories
- ❖ The credit card loans are less defaulter amongst almost all the four purpose of loan, when comparing between different categories

A decorative graphic on the left side of the slide. It features a large cyan hexagon in the center containing the number '4'. Surrounding this central hexagon are several smaller hexagons of varying shades of blue and cyan. Some of these smaller hexagons contain white icons: a lightbulb, a thumbs-up, a smartphone, a magnifying glass, and a gear. There is also a network-like icon with a central node and radiating lines, and a speech bubble icon. The entire graphic is set against a dark blue background.

4

CONCLUSION



1. There is a lot of missing data within the dataset, approximately 50% of the columns have more than 90% missing data.
2. Maximum number of loans are taken for the purpose of debt_consolidation. Also, this category contain the most defaulters.
3. There are more percentage of default-rate in long-term rate, that is, 36 months.
4. Small business loans default the most, then renewable energy and education.
5. It is important to analyse grade and subgrade, since the grades have positive effect on defaulters.





THANK YOU

