# Lending Club Case Study

Prepared by:
Ejaz Sayyed
Manikandan

# Agenda

- Problem Statement
- Technologies Used
- Data Understanding
- Analysis
- Conclusion(s)

# Problem Statement

- Understand the **driving factors (or driver variables)** behind loan default, i.e., the variables which are strong indicators of default

# Technologies Used

We have used below technologies for analysis

- **Python** – Open-source language best suited for handling data. Python provides rich libraries like matplotlib & seaborn for visualization and many other libraries for statistical analysis and data manipulation
- **Jupyter Notebooks** – An open-source web application to create and share documents that contain live code (e.g. Python), equations, visualizations, and narrative text.

# Data Understanding – As-IS

After analyzing the raw data, below are the initial observations:

- 39717 records with 111 columns
- Enough columns to derive the useful metrics and perform loan default analysis
- But also, more than 50 columns which had no data
- Many attributes which are not required for this analysis
- Some missing values in the data

# Data Cleansing Activities

We used below techniques to cleanse the data and make it suitable for analysis -

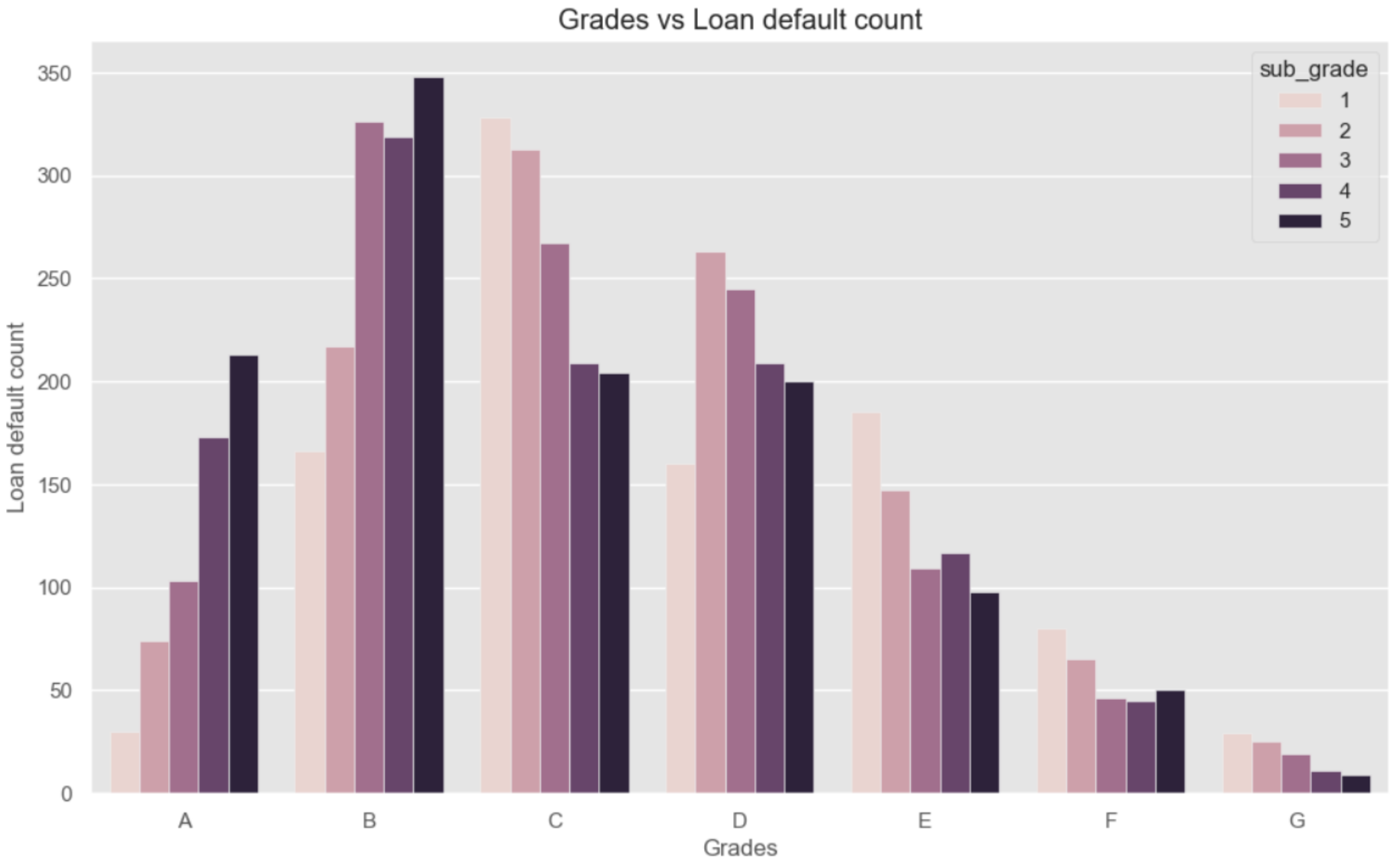
- Fill in some columns with missing values i.e. imputation (using mean, median, mode techniques)
- Converted columns from String to Day/Month/Year
- Removing columns where ‘all’ values are null
- Dropping columns which are not relevant for this analysis
- Converting few column values from String to Integer/Float for analysis

# Analysis on Data - I

- Univariate Analysis (i.e. using one variable/field/column/attribute at a time)

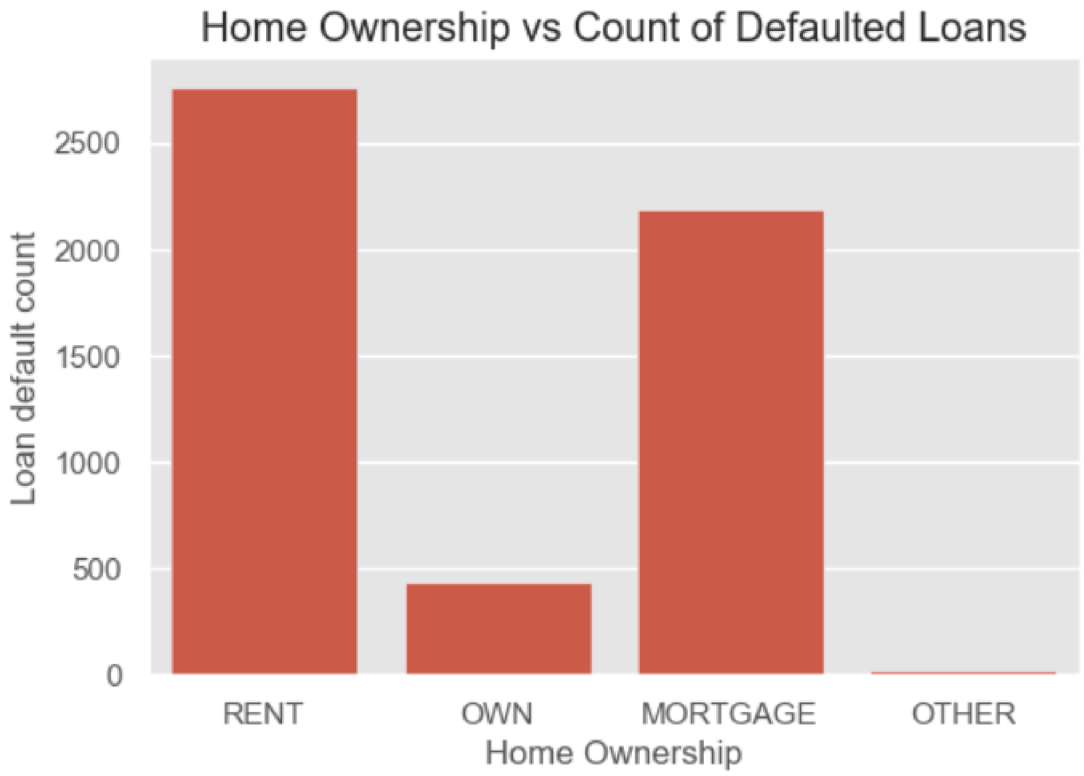
# Observation 1

Many defaulters are from B and C grades and specifically, borrowers from 5th subgrade of B are higher in %.



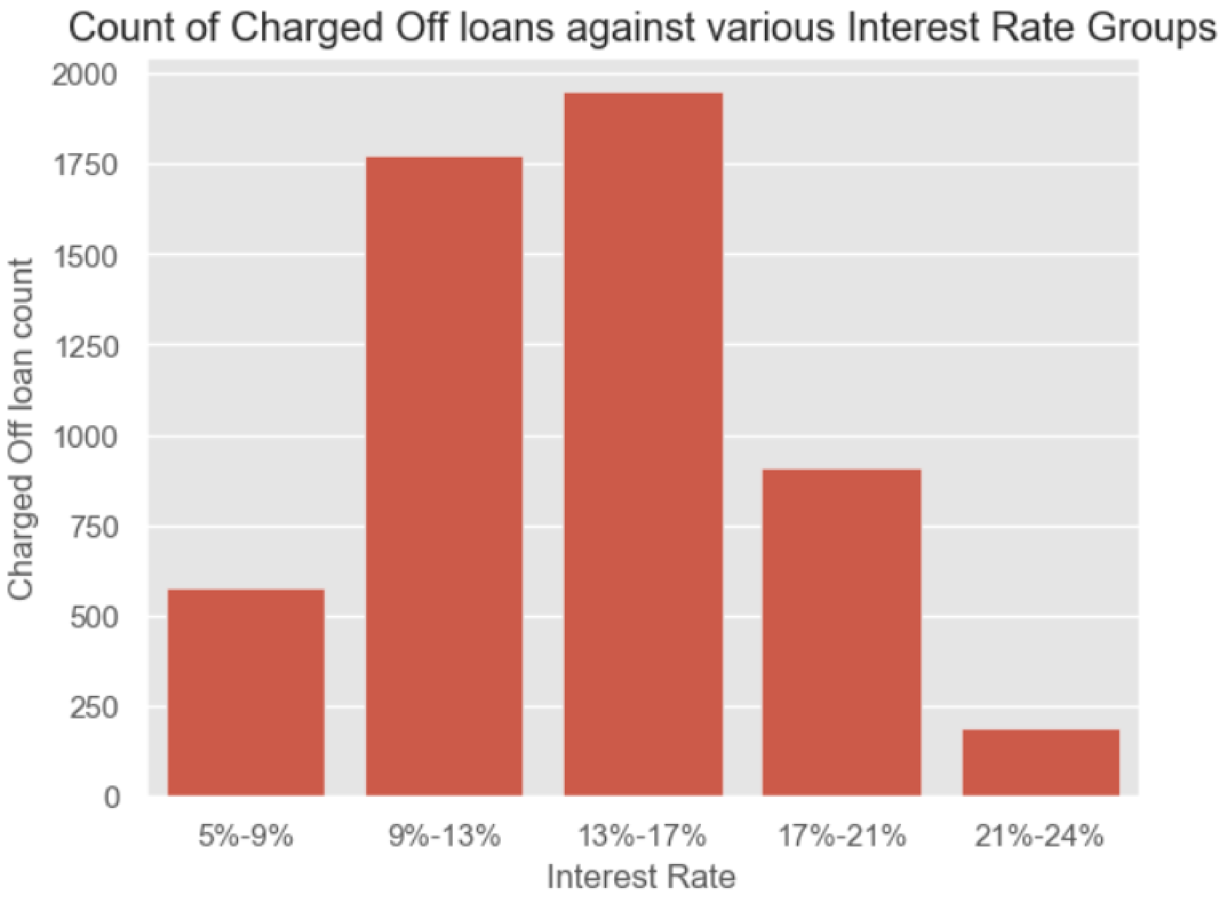
# Observation 2

Loan Defaulters have ‘Home Ownership’ mostly as ‘**Rent**’ or ‘**Mortgage**’



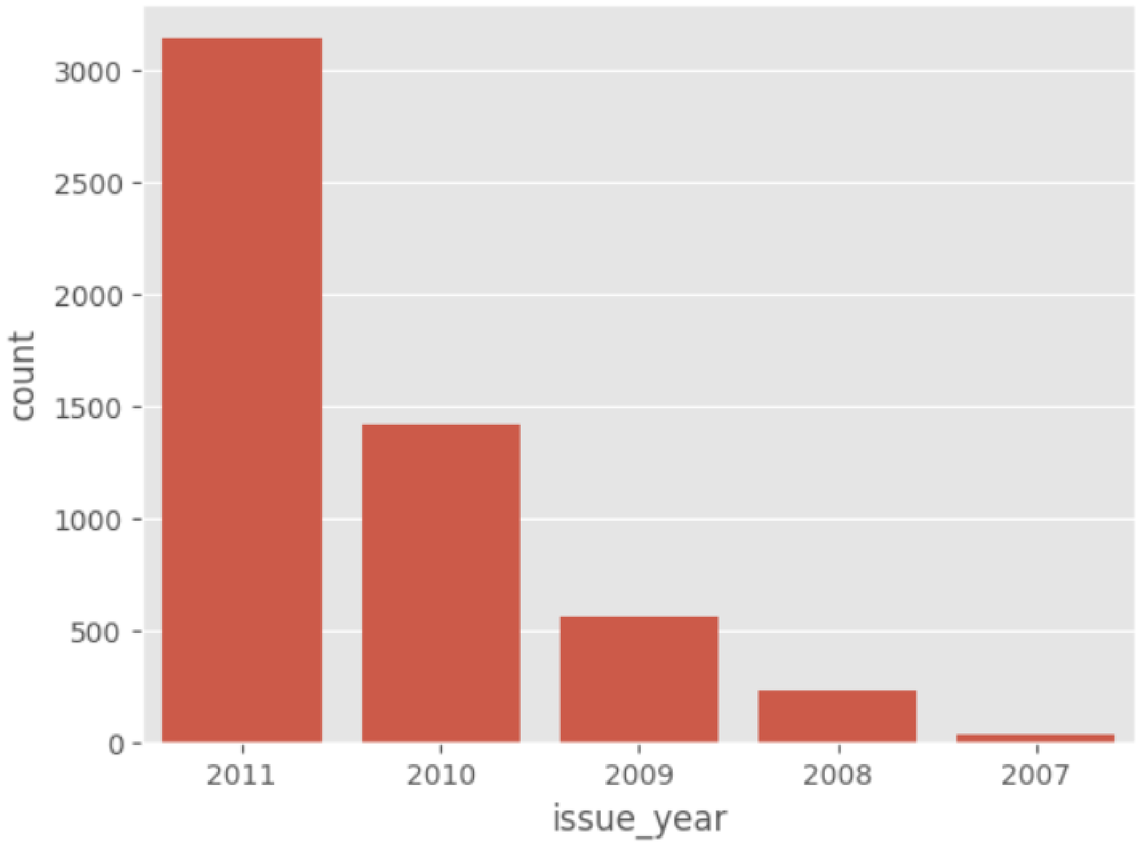
# Observation 3

Loans are ‘Charged Off’ or ‘Defaulted’ when ‘Interest Rate’ is high.



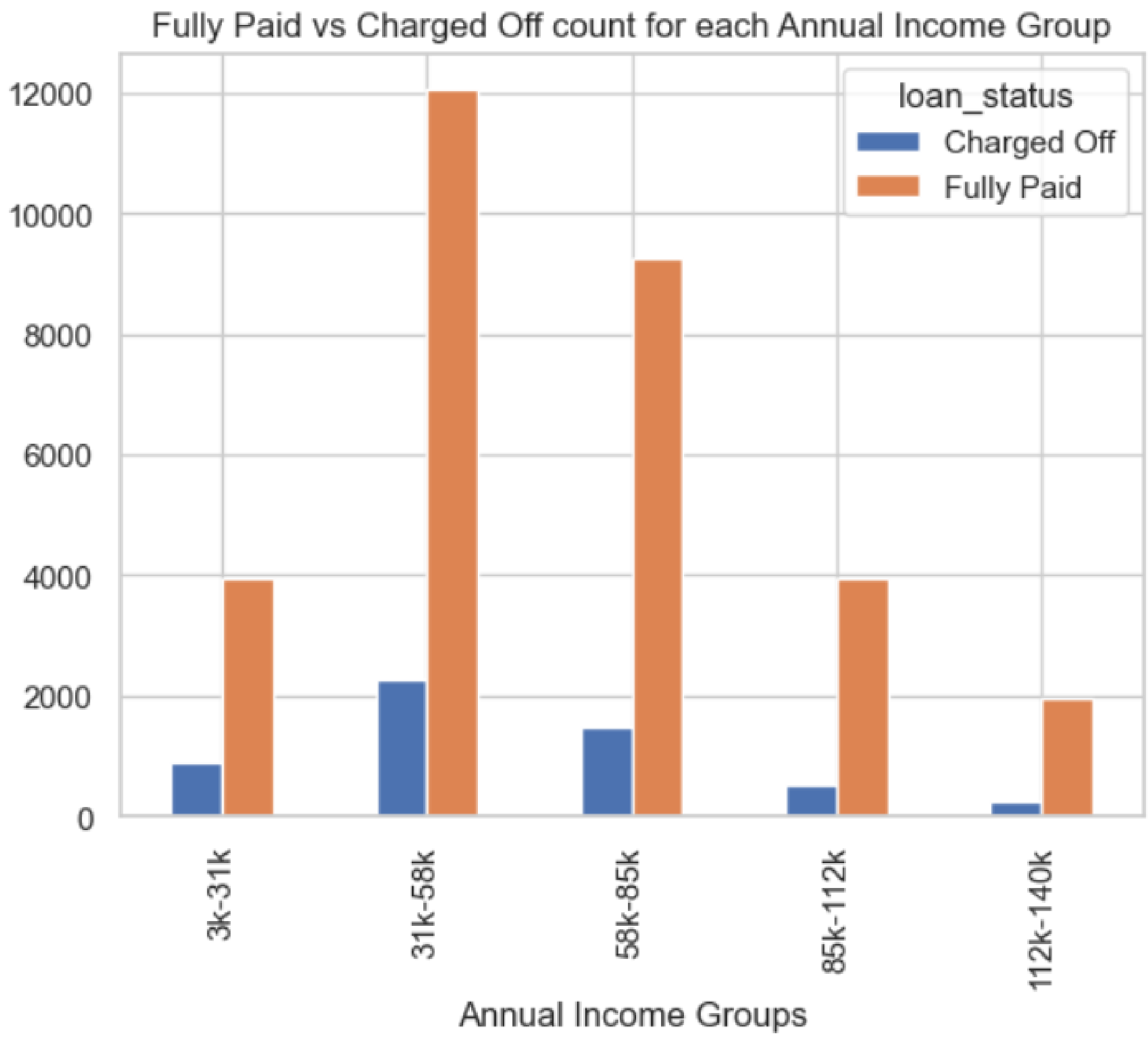
# Observation 4

Most Loan defaults have occurred in year 2011.



# Observation 5

Charged Off (i.e. Defaulters) count is high in the annual income group of 31k-58k.



# Analysis on Data - II

- Bivariate or Multivariate Analysis (i.e. using **two or more** variables/fields/columns/attributes at a time)

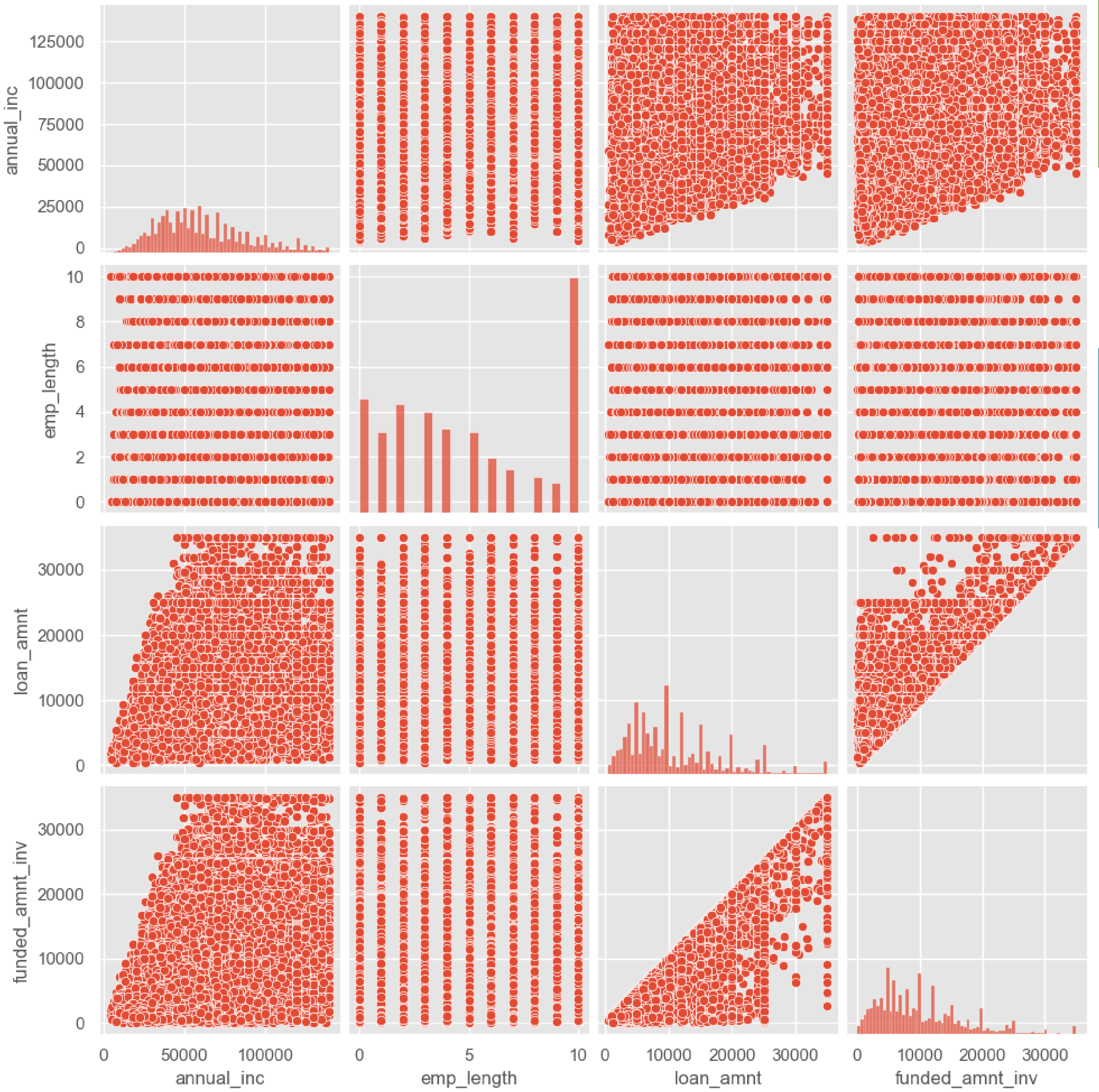
# Observation 6

This is called as Pairplot and is used to show relationship between multiple variables/attributes in the dataset.

From the graph, it can be observed that –

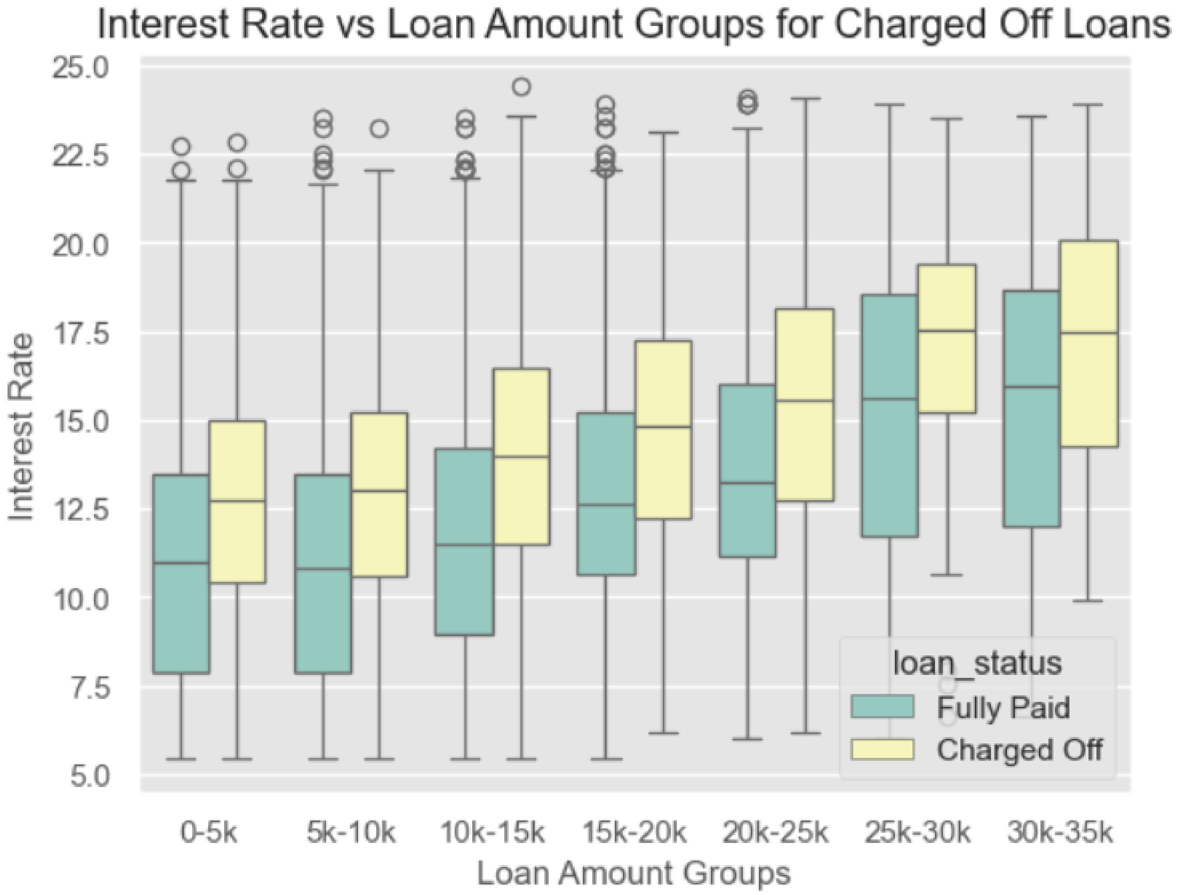
- Loan Amount and Funded Amount are related

- No relationship between Employment Length/Duration and Loan Amount!



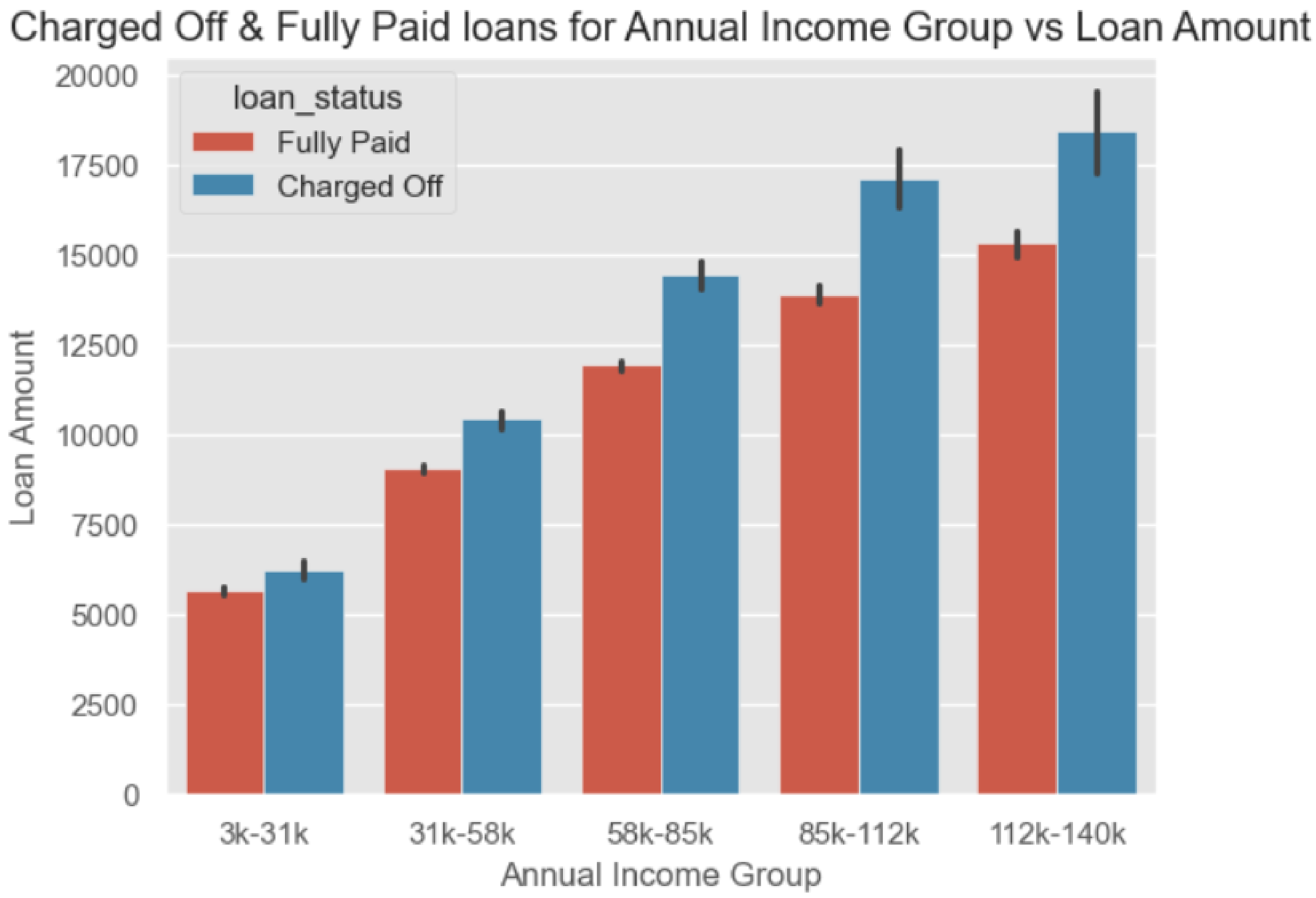
# Observation 7

For high loan amounts, more interest is being charged for the loans which are defaulted.



# Observation 8

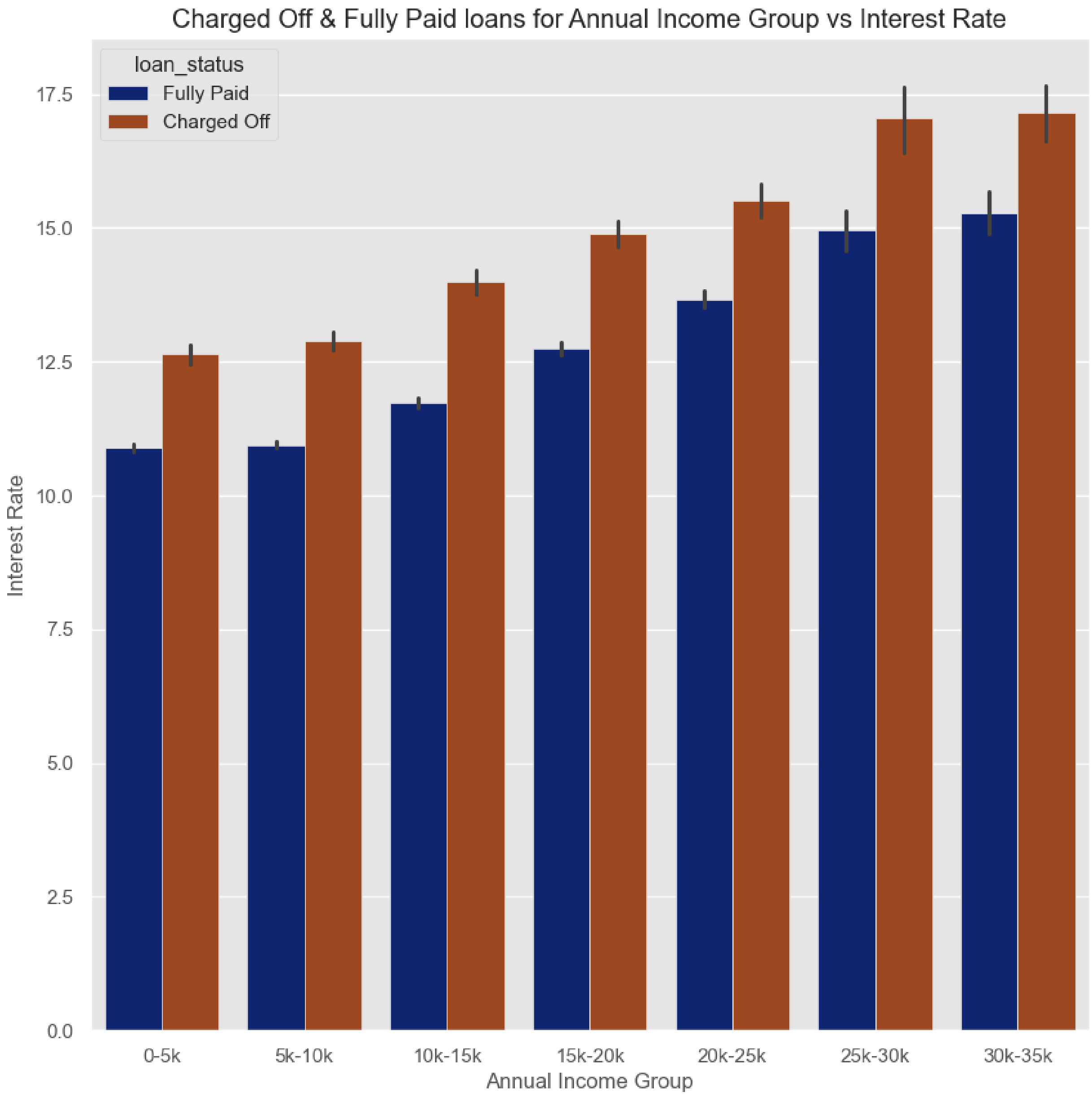
As the Loan Amount gets higher, % of loan defaults is increased. These high amount loans are taken by people with high annual income.



# Observation 9

High Interest Rate Loans are taken by people with high Annual Income.

‘Fully Paid’ Loans are always taken at Low Interest Rates than ‘Charged Off’ loans.



# Observation 10

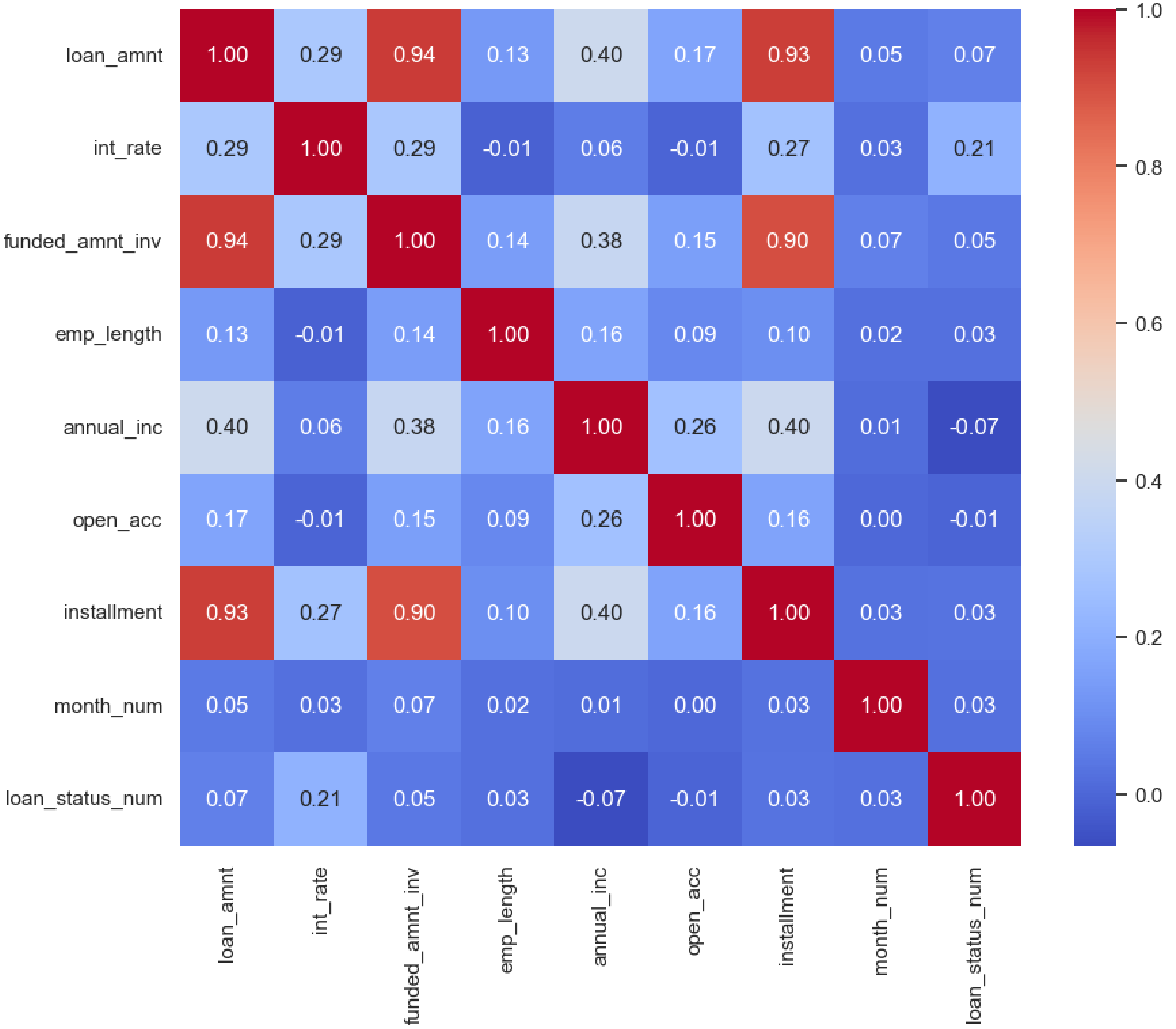
Heatmap is shown. It shows relationship of various attributes with every other attribute.

There is high correlation between –

- Funded Amount Investment & Installment
- Loan Amount & Installment
- Loan Status & Interest Rate

There is low correlation between –

- Loan Status & Annual Income of a person



| | loan_amnt | int_rate | funded_amnt_inv | emp_length | annual_inc | open_acc | installment | month_num | loan_status_num |
|---|---|---|---|---|---|---|---|---|---|
| loan_amnt | 1.00 | 0.29 | 0.94 | 0.13 | 0.40 | 0.17 | 0.93 | 0.05 | 0.07 |
| int_rate | 0.29 | 1.00 | 0.29 | -0.01 | 0.06 | -0.01 | 0.27 | 0.03 | 0.21 |
| funded_amnt_inv | 0.94 | 0.29 | 1.00 | 0.14 | 0.38 | 0.15 | 0.90 | 0.07 | 0.05 |
| emp_length | 0.13 | -0.01 | 0.14 | 1.00 | 0.16 | 0.09 | 0.10 | 0.02 | 0.03 |
| annual_inc | 0.40 | 0.06 | 0.38 | 0.16 | 1.00 | 0.26 | 0.40 | 0.01 | -0.07 |
| open_acc | 0.17 | -0.01 | 0.15 | 0.09 | 0.26 | 1.00 | 0.16 | 0.00 | -0.01 |
| installment | 0.93 | 0.27 | 0.90 | 0.10 | 0.40 | 0.16 | 1.00 | 0.03 | 0.03 |
| month_num | 0.05 | 0.03 | 0.07 | 0.02 | 0.01 | 0.00 | 0.03 | 1.00 | 0.03 |
| loan_status_num | 0.07 | 0.21 | 0.05 | 0.03 | -0.07 | -0.01 | 0.03 | 0.03 | 1.00 |

# Conclusions

Below are the main factors having influence on loan defaults –

- Interest Rate (and thus Installment)
- Grade
- Year (2011) – Can be due to market conditions in the year!
- Home Ownership

# Thank you