

# GRAMENER CASE STUDY

## EXPLORATORY DATA ANALYSIS

### For a Bank (Loan criteria)

**Name:**

**Deepak Pandey**

**Vivek P**

**Kedarnath G**

**Ravindra S (Facilitator)**

## Overview

- This analysis is performed for Gramener, a **consumer finance company** which specializes in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Two **types of risks** are associated with the bank's 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. they are likely to default, then approving the loan may lead to a **financial loss** for the company

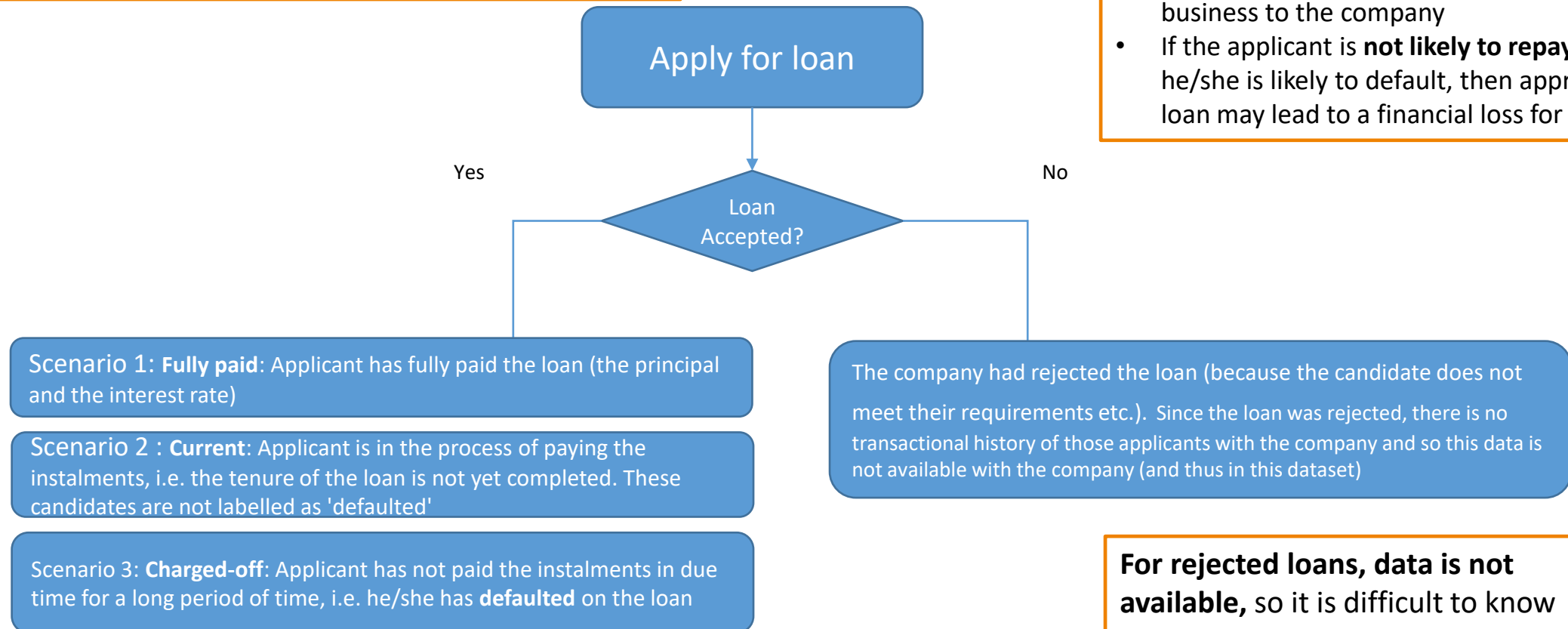
## Objective

- The dataset provided contains the information about past loan applicants and whether they **'defaulted' or not**.
- The aim is 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.
- Determine how **consumer attributes** and **loan attributes** influence the tendency of default and present a **'risk profile'** of individual seeking loan.

**Objective:** 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 utilize this knowledge for its portfolio and risk assessment.

**Risks associated with Bank's decision of giving loan:**

- 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



**For rejected loans, data is not available,** so it is difficult to know what requirements were not met by the customer.

## 1 Understand Business Objectives

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 utilize this knowledge for its portfolio and risk assessment.

2

## Determine Goals for Exploratory Data Analysis

Data Source: loan.csv, Data\_Dictionary.xls

1. **Univariate Analysis:** Determine columns to be analyzed and find details using attributes such as mean, median, SD, percentile, etc.
2. **Segmented Univariate Analysis:** Segment relevant data to understand impact on other variables
3. **Correlations:** Perform bivariate analysis to find correlations amongst numerical data attributes as well as categorical data

3

## Plan and establish milestones

Define checkpoints / milestones for analysis and execution

4

EXECUTE

### 1. Data Cleaning & processing

Using Python load data in Dataframe (DF).

Process and clean the dataset

Processed & clean master data

### 2. Univariate Analysis

Determine the relevant variables required for the analysis

Document the metadata of the dataset and each relevant column – such as unique values, mean, median, count, percentile

Data dictionary / Key variables identified

### 3. Segmented UV Analysis

Determine the target variable. Find the average.

Group by the impact variables and analyze how average compares

Observations from segmented UV analysis

### 4. Bivariate Analysis

Analyze the correlations among the numerical variables

Correlations analyzed between variables

### 5. Summarize Analysis

Consolidate the understanding from various analysis and conclude.

Risk profile for loan

5

## Consolidate & present recommendations

- Consolidate analysis and document
- Present the risk profile based on customer and loan attributes and give recommendations.


EDA

1. Business Understanding
2. Data requirements
3. Cleaning and preparation
4. Univariate Analysis
5. Segmented Univariate analysis
6. Summary of analysis

- The interest rate on a loan is a quantification of the risk associated with the loan, and as such was not treated as a cause of default. If a high interest loan gets defaulted on – it is because the loan was a high risk loan to begin with.
- The bottom and top 2.5% of annual income can be removed as statistical noise.
- States are vast and diverse, and cannot be used as a reliable metric. Zip code offer a more specific and actionable variable.
- Loans that are Current may be ignored for most analysis.



**GOAL : Prepare the given dataset for analysis by removing unwanted data and attributes.**

1. Remove columns – Of **111 columns** in original dataset, **52 remained** after columns with ‘Null’ values were removed.
2. Check percentage of empty values in remaining columns - Remove **two** columns where **values > 90%** are not available - ‘**mths\_since\_last\_record**’ and ‘**next\_pymnt\_d**’
3. Analyze other columns and eliminate the ones not relevant for analysis. Following are identified as not relevant. 
4. Other changes –
  - Convert the ‘**revol\_until**’ data type from object to float
  - Change ‘**emp\_length**’ from object to float type - assumed ‘0’ for ‘n/a’; assumed ‘0’ for ‘<1’ and ‘10’ for ‘10+’
  - Remove rows corresponding to **outliers** for column ‘annual\_inc’ from the dataset (eliminate individuals outside the [2.5-97.5] percentile range)
  - Column ‘**int\_rate**’ – data type converted to float by removing ‘%’
  - Column ‘**term**’ - data type converted to float

Column name	Reason for elimination
url	Not relevant for analysis
'tax_liens	All values are zero
'application_type	All values are 'INDIVIDUAL
'acc_now_delinq	All values are zero
'chargeoff_within_12_mths	All values are zero
'delinq_amnt	All values are zero
'collections_12_mths_ex_med	All values zero or Null
'policy_code	All values are '1 - publicly-available policy code'
'member_id	Use 'id' as unique key instead



# Results from Univariate and Segmented Univariate Analysis

**GOAL : Identify some key Loan and Borrower attributes that influence the chances of the loan being charged off**

## Key Insights

1. The median annual income of our borrowers is USD 59,000 which is significantly higher than the median US Individual income.
2. The Median Debt To Income ratio stood at 13.4%
3. The Charge Off rate for loans was 14.1%
4. About 30% of the principal was recovered from Charged Off Loans.



**While there are more insights present in the data, this presentation will focus on results of Segmented Univariate and Bivariate Analysis.**



**GOAL : Identify some key Loan and Borrower attributes that influence the chances of the loan being charged off**

Based on the Segmented Univariate Analysis, we have determine that the following variables can are indicators of an increased probability of loan default :

1. Term of the Loan
2. Purpose of the Loan
3. Existence of Public Records (Derived Variable)
4. Existence of Delinquent Accounts in the last two years (Derived Variable)
5. The Classification of the Loan
6. Revolving Credit Utilization
7. Variable Derived form the Loan Amount and the Borrowers Annual Income
  - Current Debt To Income Ratio
  - Increase in Debt to Income Ratio a new loan will cause
  - Final Debt to Income Ratio
8. High Risk Location (Zip Codes)





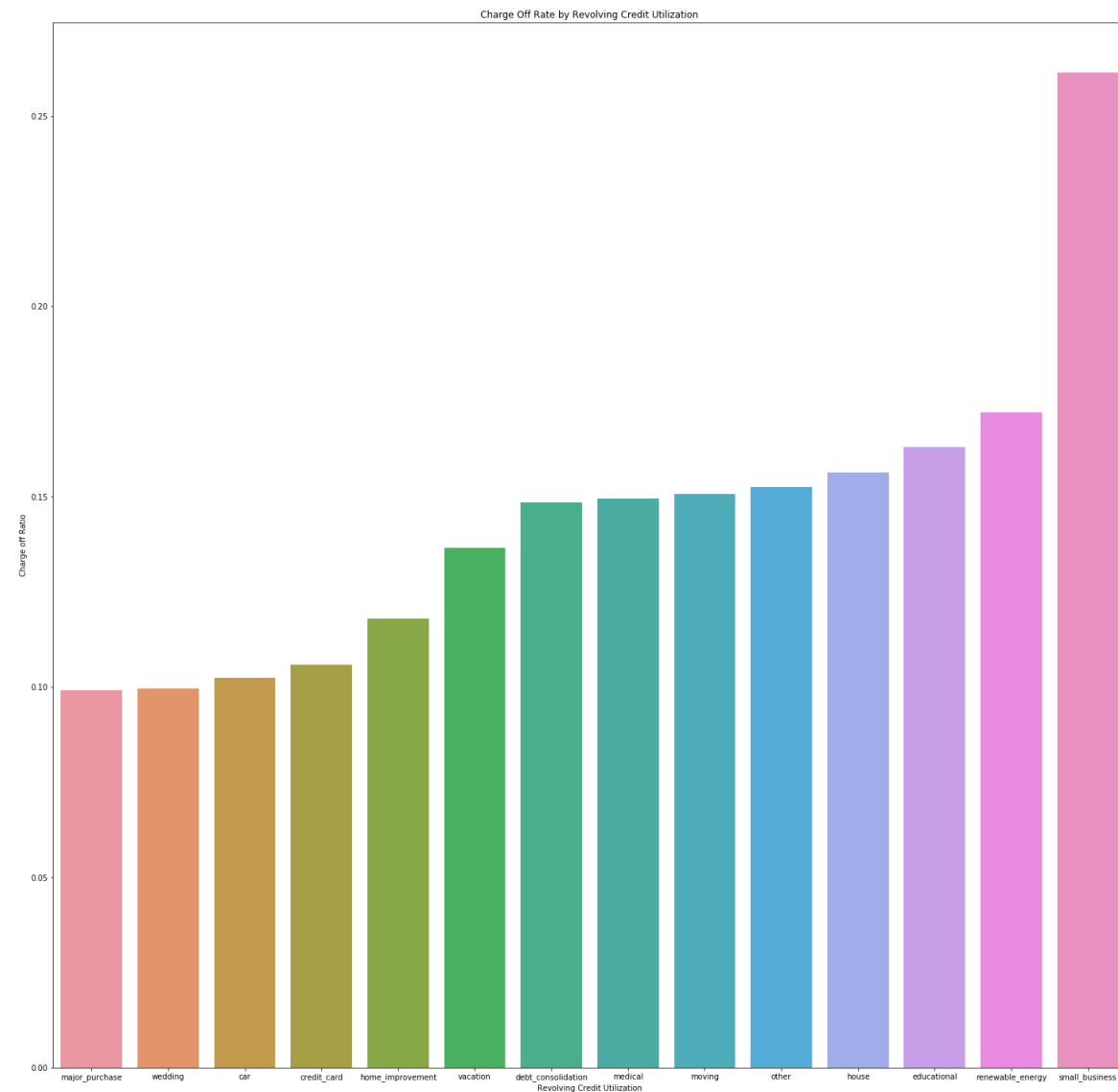
It has been observed that loans for a 60 month term have a much higher default rate than 36 month loans.

	Charged Off	Current	Fully Paid	Charge Off Rate
Term				
60	2311.0	1100.0	6782.0	0.226724
36	3008.0	0.0	24612.0	0.108907

Loans for some specific purposes posed a significantly higher risk than others

Some of the highest risk loans were

- Small Business – 26.1% Charge Off
- Renewable Energy – 17.2% Charge Off
- Education – 16.2% Charge Off





## Variable : Public Records and Delinquent Accounts

We determined the presence of public records and delinquent accounts increased the probability of a loan getting charged off.

	Charged Off	Current	Fully Paid	Charge Off %
Public Records				
Yes	449	46	1552	21.9%
No	4870	1054	29842	13.6%

	Charged Off	Current	Fully Paid	Charge Off%
Delinquent				
Yes	649	115	3319	15.9%
No	4670	985	28075	13.8%

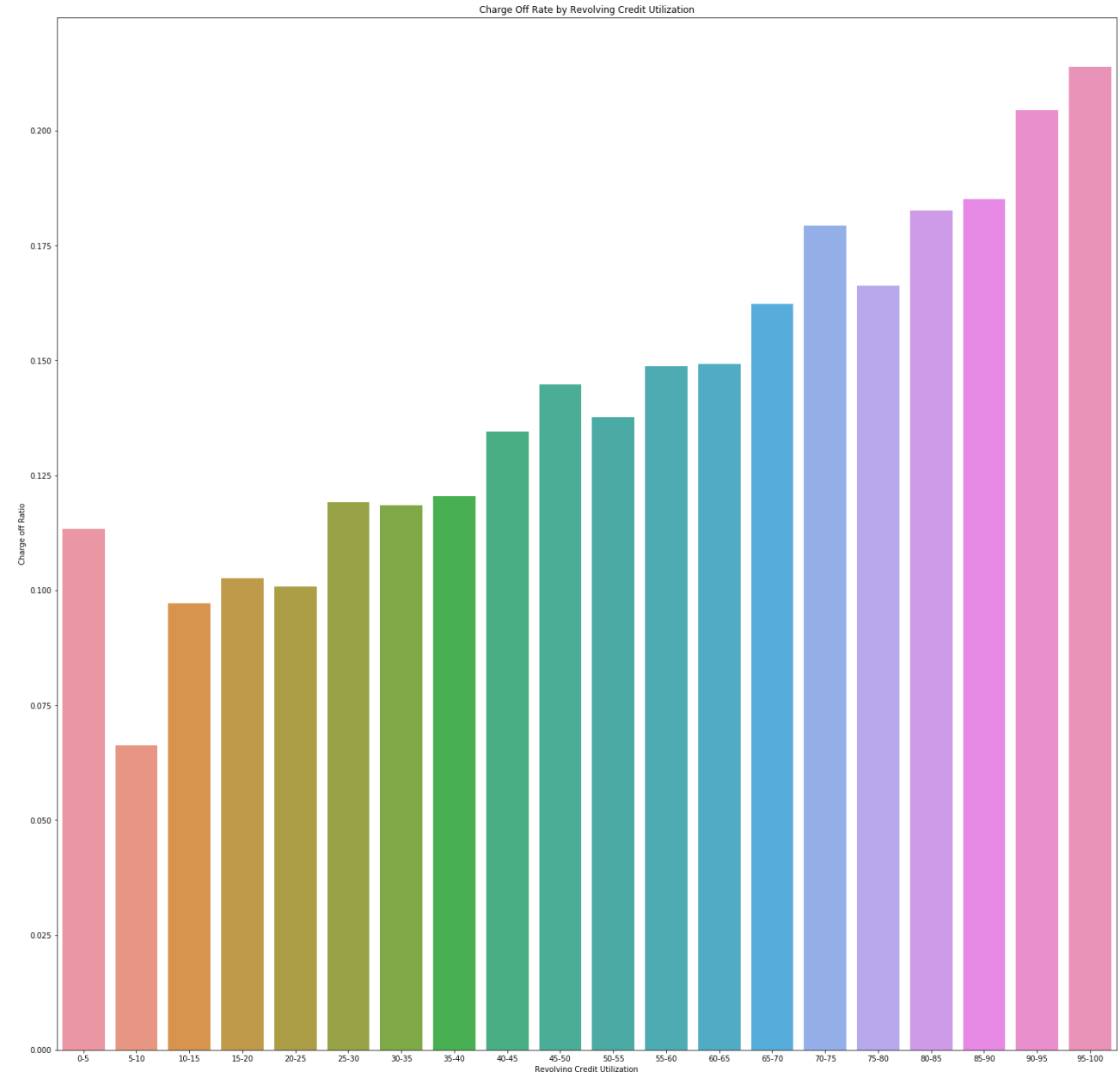
The existing Loan Classification System is a good predictor of the possibility a loan gets charged off.

- A and B rated loans had a much smaller Charge Off Rate than loans rated C or below.

Grade	Charged Off	Current	Fully Paid	Charge Off%
G	95	16	184	32.2%
F	301	66	613	30.7%
E	675	169	1843	25.1%
D	1063	214	3761	21.1%
C	1277	257	6160	16.6%
B	1343	339	9774	11.7%
A	565	39	9059	5.8%

A high revolving credit utilization is a strong indicator of a higher risk of loan default

- Beyond the 40-45% Revolving Credit Utilization range, the chances of charge off increase to over 14%.





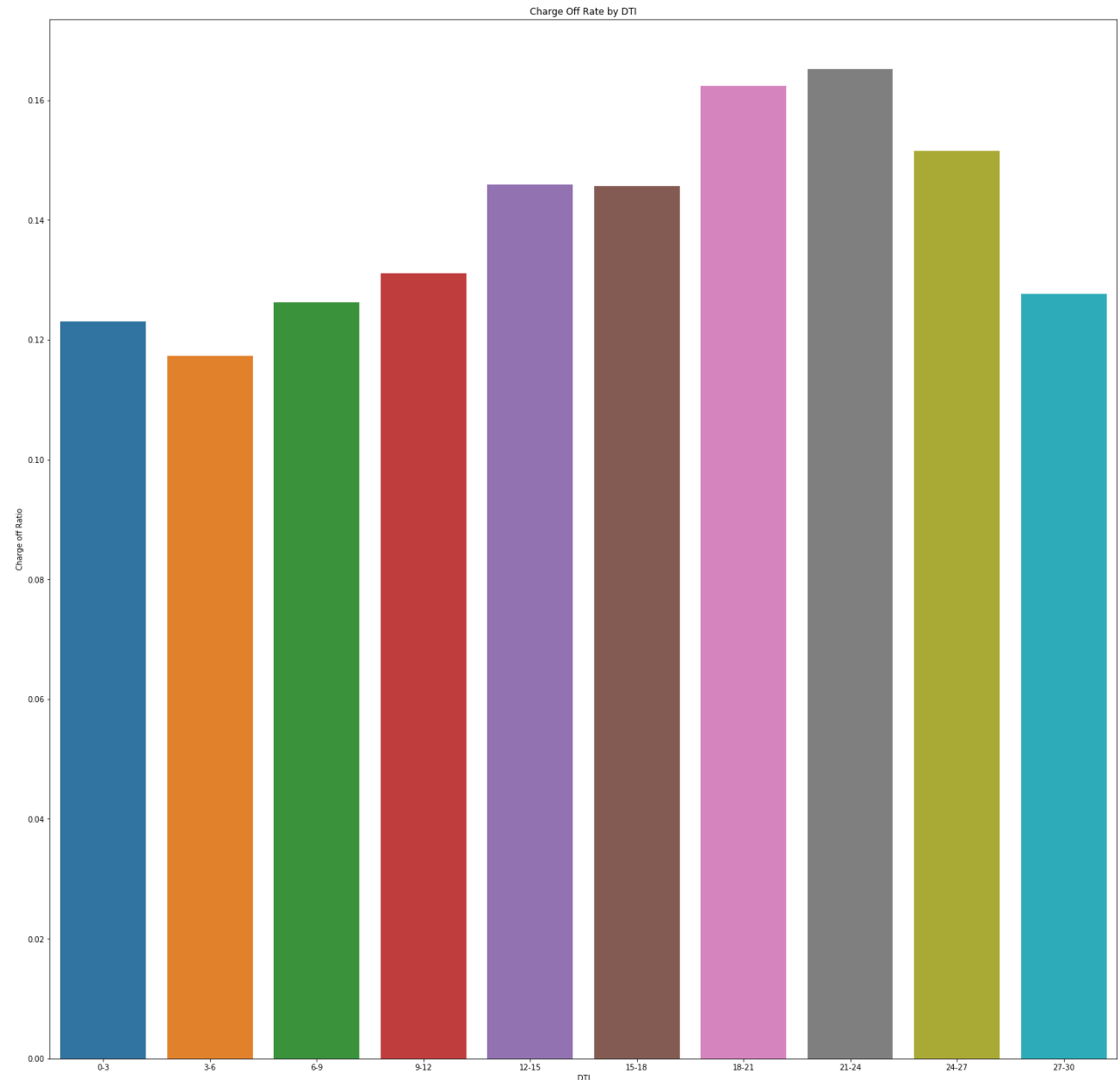
# Variables Derived from Annual Income, Amount of the Loan, and existing DTI



## Existing DTI

As existing DTI increases, the chances of a loan being charged off increase.

- At about 18% DTI, the probability of Charge Off increases beyond our sample's 14% probability.



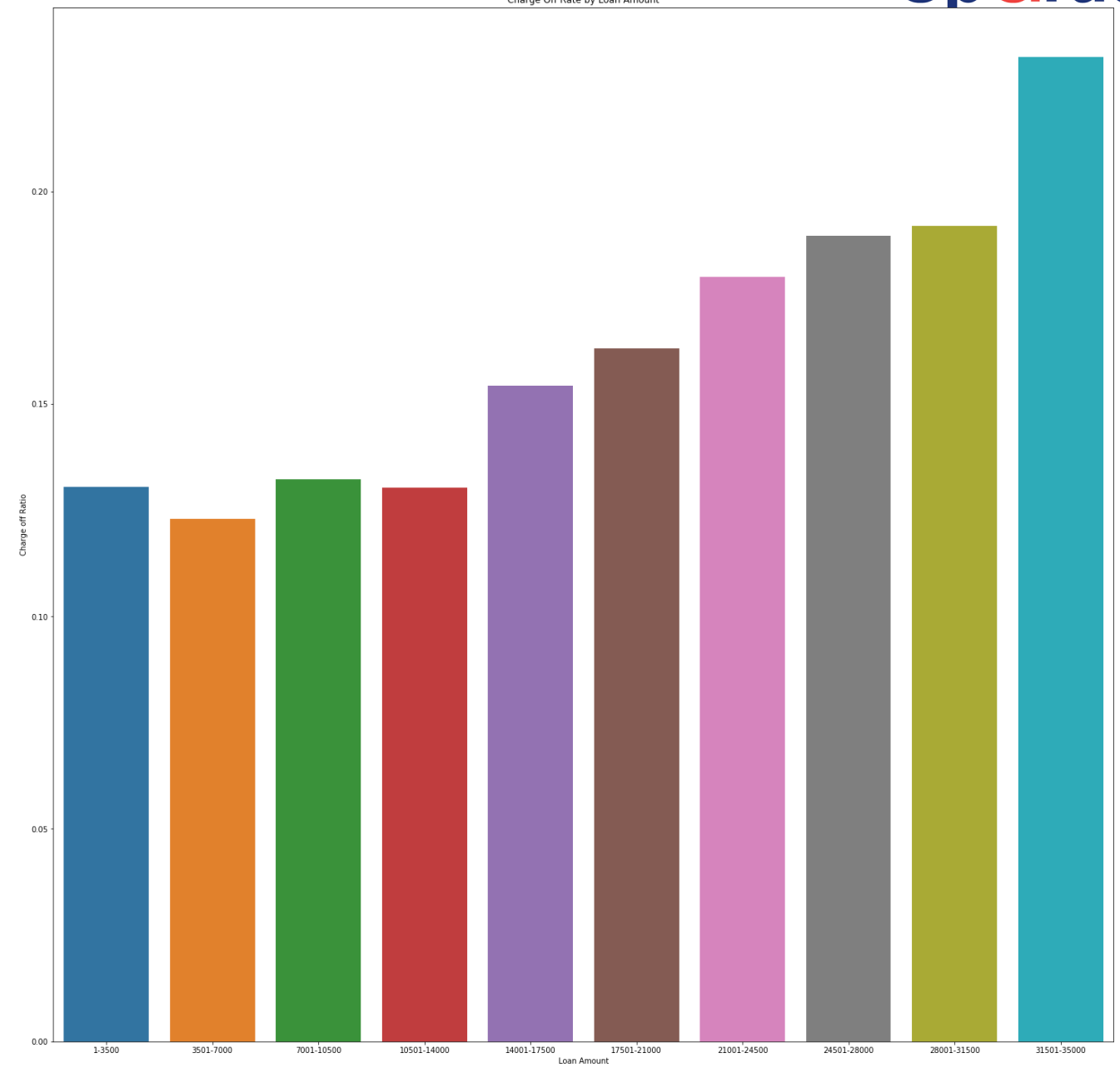


# Variables Derived from Annual Income, Amount of the Loan, and existing DTI



## Loan Amount

As the loan amount increase, the ratio of loans getting charged off increases.





# Variables Derived from Annual Income, Amount of the Loan, and existing DTI

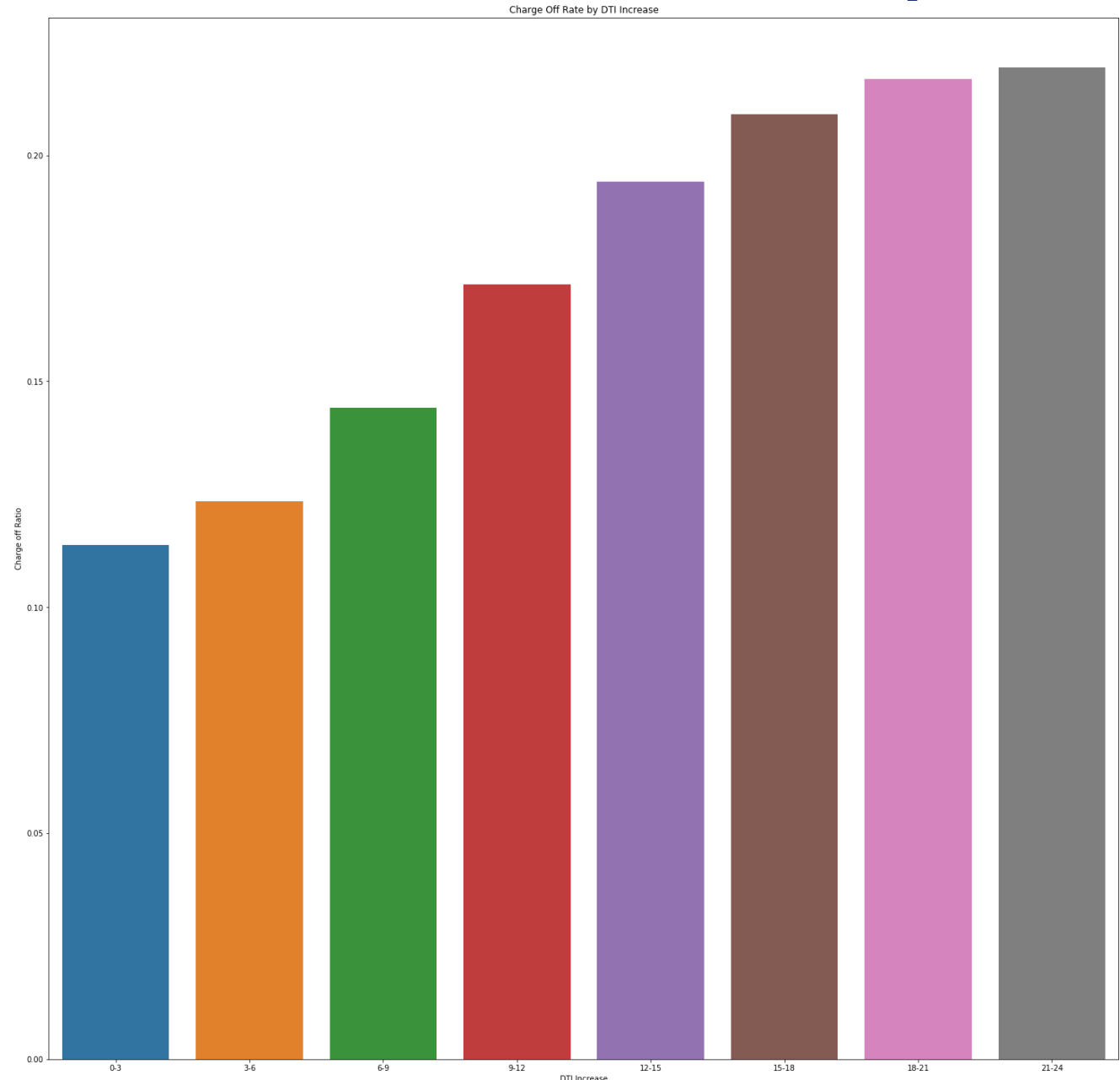


## Increase in DTI

We created a new variable – Increase in DTI. This variable represents the ratio of the loan’s installment to the borrower’s income.

If this ratio is high, the chances the loan gets charged off is also high.

If this ratio is above 9% there is a higher risk of default.







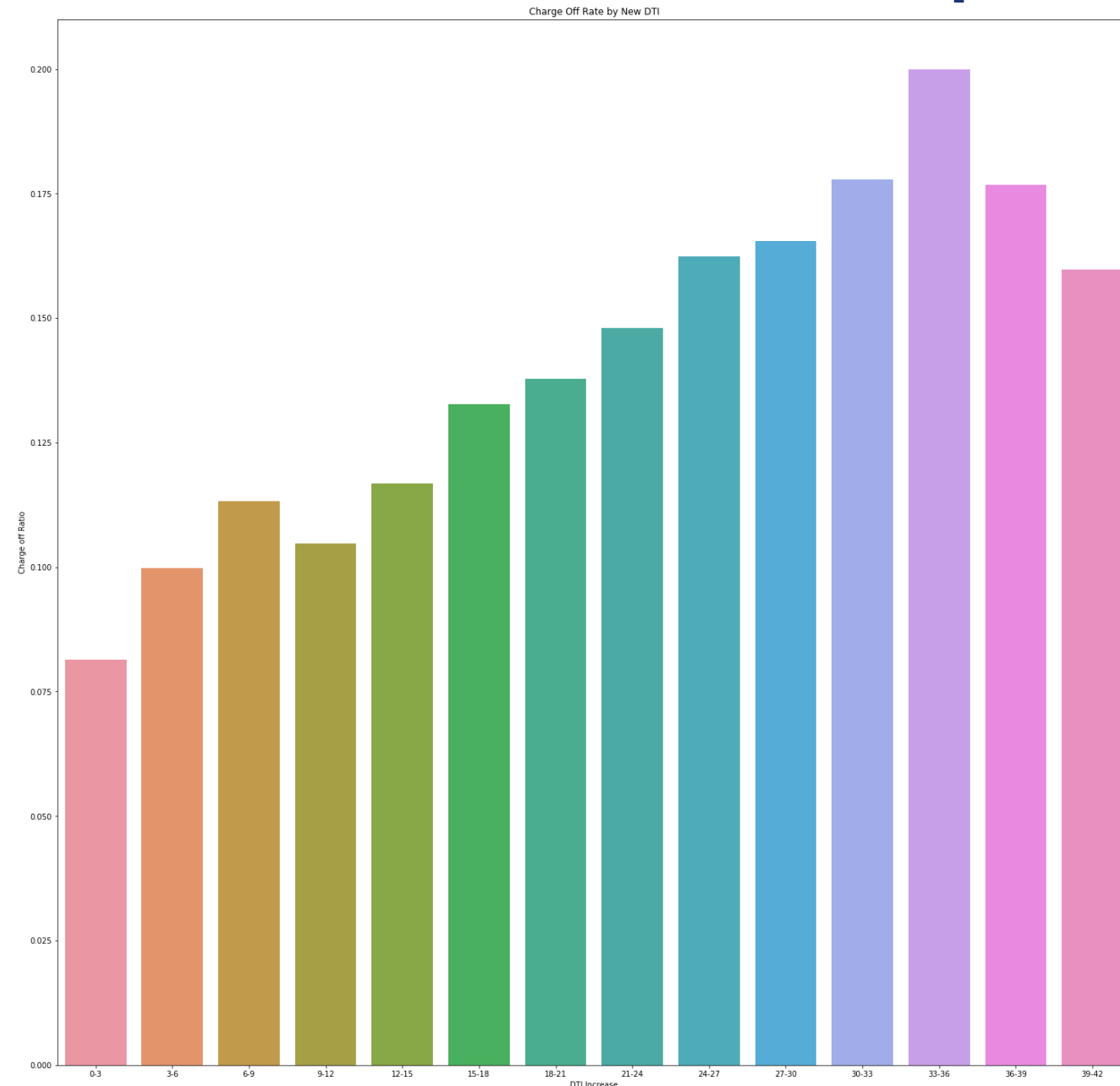
# Variables Derived from Annual Income, Amount of the Loan, and existing DTI



## New DTI

We created a new variable – New DTI. This variable represents the borrowers Debt to Income Ratio the requested loan is granted.

If this ratio is above 24% there is a higher risk of default.



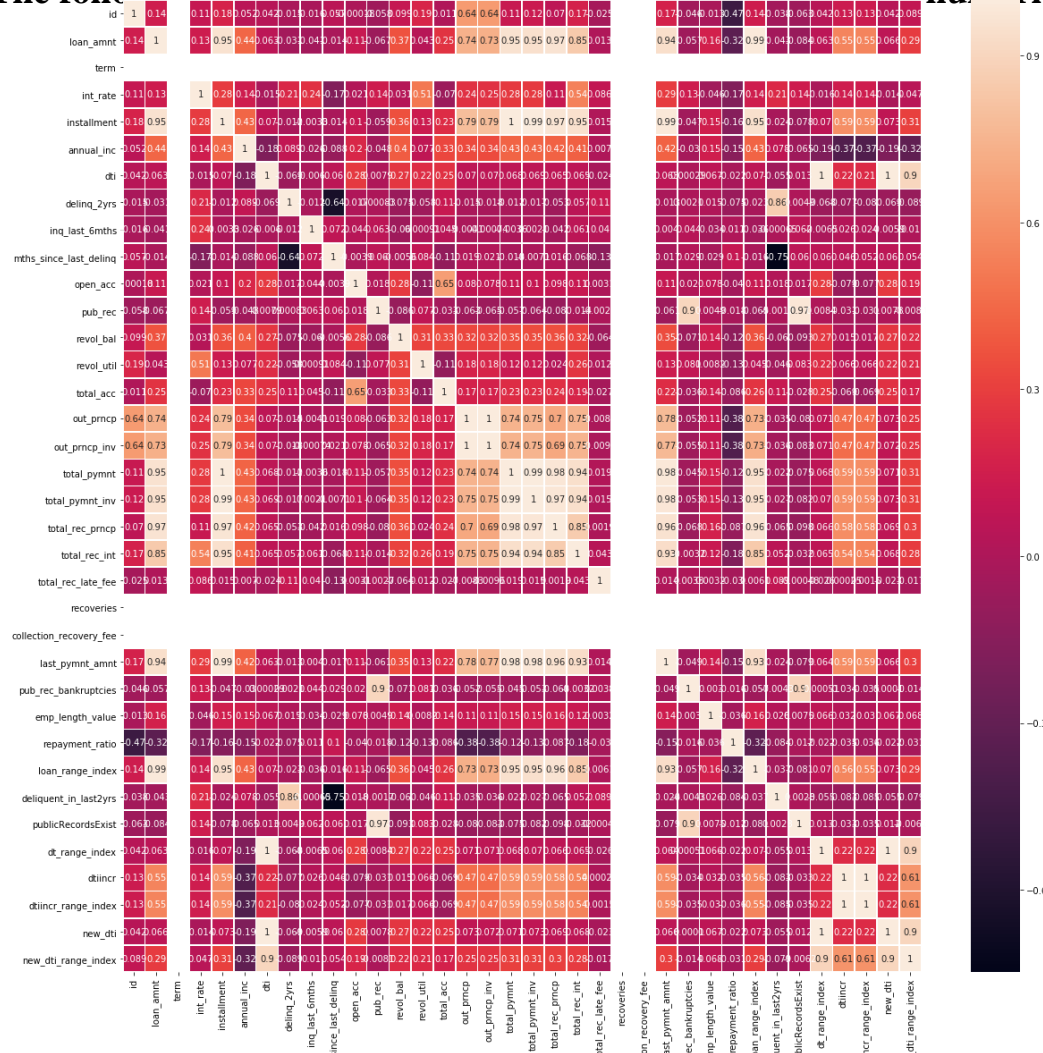
We have identified a list of high risk zip codes. If there are other risk factors associated with the loan, and if the borrower is resident of these zip codes, a closer look is warranted.

A sampling of this data is shown below

ZIP	Charged Off	Current	Fully Paid	totalCount	Charged Off%
912	15.0	1.0	30.0	46.0	32.6
935	31.0	1.0	67.0	99.0	31.3
986	15.0	1.0	39.0	55.0	27.8
206	16.0	1.0	42.0	59.0	27.1
321	16.0	1.0	43.0	60.0	26.7
082	13.0	2.0	34.0	49.0	25.5
106	8.0	0.0	24.0	32.0	25.0
072	9.0	1.0	26.0	36.0	25.0
484	8.0	0.0	24.0	32.0	25.0

## GOAL: Analyze the correlations between relevant numerical variables

The following man shows the correlations between various numerical variables.



## Key observations from the correlations:

- Funded amount has:**
  - negative correlation with ‘debt to loan’ ratio – indicating that higher the existing debt, the funded amount goes down*
  - A slight negative correlation with the ‘public record’ and ‘no. of bankruptcies’*
- Interest rate has positive correlation with all most all the other variable.**

Key ones are:

  - Higher the funded amount, higher the interest rate*
  - More credit use (revol\_util), higher the interest rate.*
- Annual\_inc:**
  - The Annual income is higher where the funded amount is high*
  - ‘Debt to income’ ratio has negative correlations with the annual income indicating individuals with higher debt tend to have lower income*
- Total payment towards loan:**
  - Is negatively correlated to ‘debt to loan’ ratio, indicating that higher the ration, less chances of payments towards loan*
  - Similarly, with higher public records, and bankruptcies, less chances of payments towards loan*
- Other correlations such as**
  - a positive one between public records and bankruptcies are important factor to look at when funding loan*
  - Number of open\_acc is another key factor to consider as it has positive correlations with factors that enable return of loan*

**Goal: Summarize the analysis and present a ‘Risk Profile’ based on the customer and loan attributes.**

**Risk Profile:** Following personal and loan attributes can be considered while making decision on granting loan to the individuals

**Personal Attributes:**

1. Annual Income
2. Presence of public records and bankruptcies – Higher the numbers, less chances of returning the loan
3. Current DTI
4. Utilization of revolving credit
5. Zip Code (as tie breaker)

**Loan Attributes:**

1. Term of the Loan
2. Purpose of the Loan
3. Amount of the Loan, Installment