# Lending Club Case Study


G. Kranthi Kiran
Nitin Kumar

# Agenda

1. Problem Statement Understanding
2. Solution Flow Diagram
3. Solution
   a. Data Preprocessing
   b. Data Cleaning
   c. EDA
4. Business Drivers and Recommendations
5. Appendix

# 1. Problem Statement Understanding

**Dataset Given :** Historical Loan Data

**GOAL** :

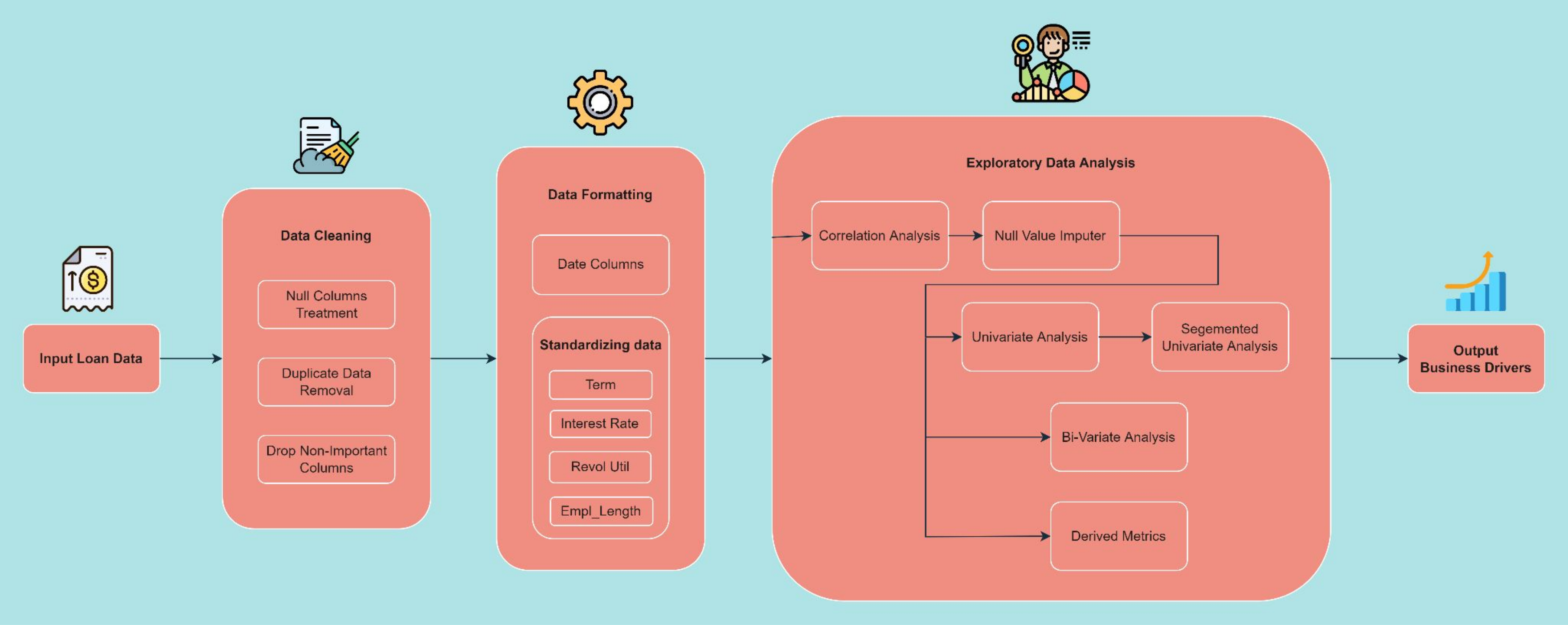
- Identify profit-makers and loss-making sectors
- Analyse the data and make business decisions which drives the business to make more profits and reduce losses

**Assumption Made :**

- We are neglecting current loans and considering only Fully Paid and Charge Off loans

upGrad

# 2. Solution Flow Diagram



# 3. Solution

1. Data Preprocessing
2. Data Cleaning
3. EDA

# 3.1 Data Preprocessing

1. **Null Column Treatment**
    a. Removing columns with **100% null values**
2. Removing **duplicate rows**
3. Removing columns which **only have 1 unique value** i.e has **no information**
4. **Dropping Columns** :
    a. **ID columns** : fully unique columns
    b. **ZipCode** : ZipCode contain partial information; lets drop that cause we can use addr_state instead of that
    c. **Last payment information** : As we're analysing the behaviour of completed loans and charged off loans; there's no point of having last payment information
    d. **Post Loan Approval Features** : We can't have this information before approving a loan which is our actual goal
    e. **Heavily Null Features** : Have > 60% null values so isn't contributing a lot of information for our case study

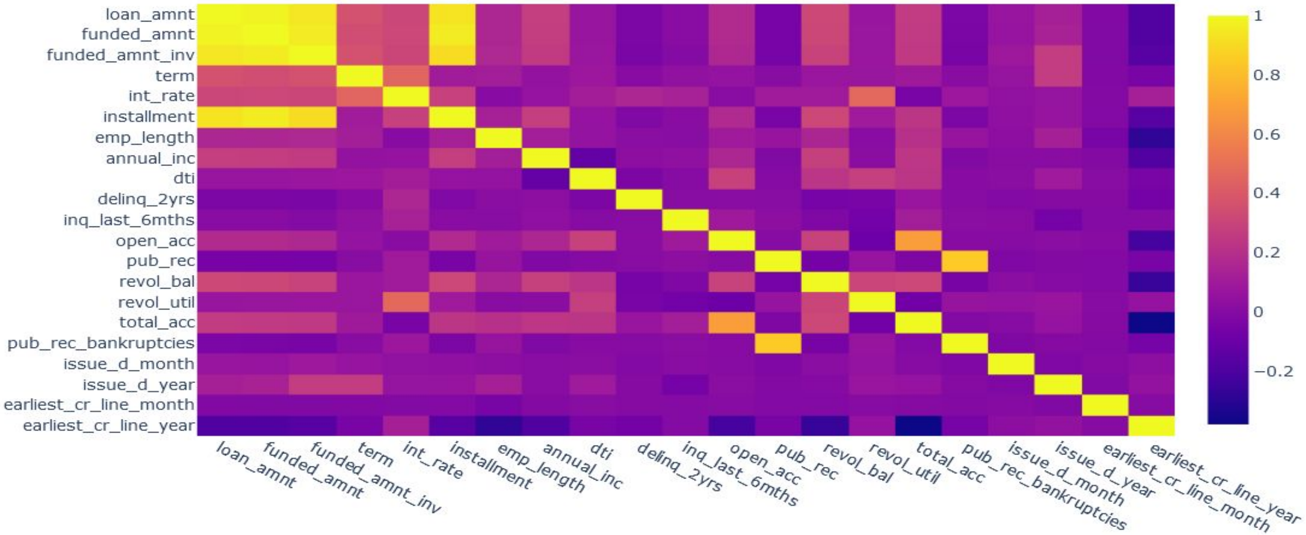
# 3.2 Data Cleaning

1. **Date Formatting** : The columns which contained the date values are **issue_d, earliest_cr_line** :
    a. **Converted** the **dates into datetime from string data-type** so that it helps to visualize data better
    b. **Derived features** :
        i. issue_d_month
        ii. issue_d_year
        iii. earliest_cr_line_month
        iv. Earliest_cr_line_year
2. **Standardizing columns** :
    a. **term** : removed the keyword “months” and converted into integer data-type
    b. **int_rate** : removed the “%” character and converted into float data-type
    c. **revol_util** : removed the “%” character and converted into float data-type
    d. **emp_length** : removed the keyword “years” and converted into integer data-type
        i. < 1 year is converted to 0
        ii. +10 years is converted to 10

# 3.3 EDA

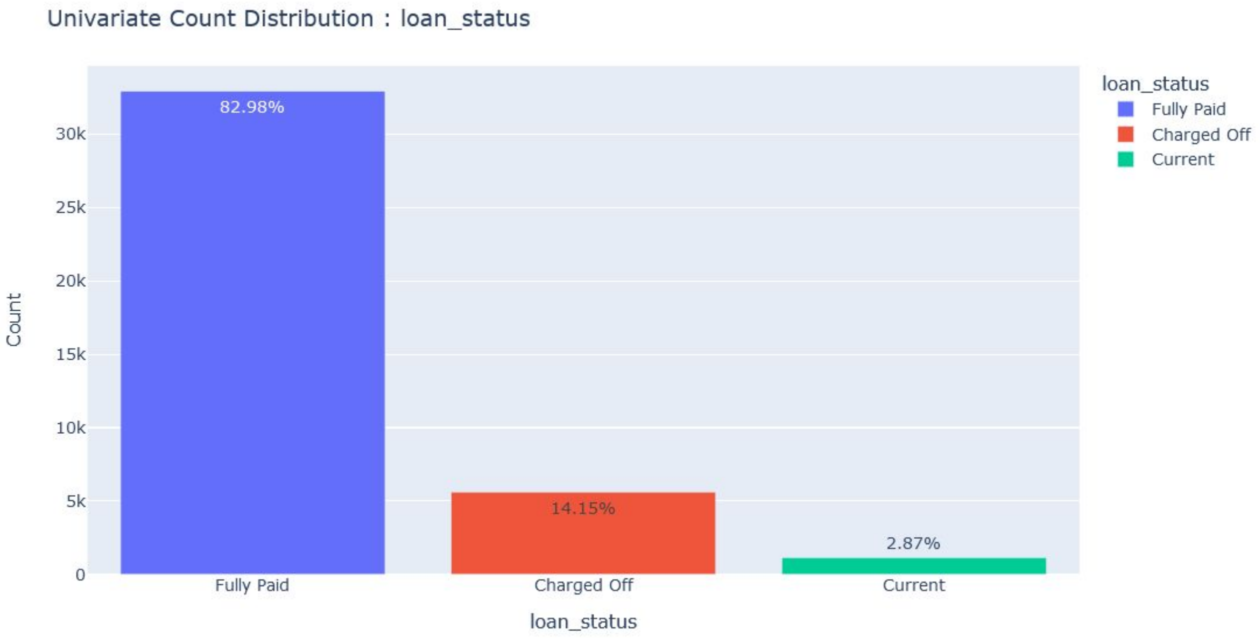
1. Correlation Analysis
2. Target Distribution
3. Null Value Imputing
4. Feature Distinction
    a. Categorical Feature Analysis
    b. Numerical Feature Analysis
    c. Date Feature Analysis
5. Bi-Variate Analysis

# Correlation Analysis




- **'loan_amnt', 'funded_amnt', 'funded_amnt_inv' and 'installment'** have **huge correlation(>0.9)** within each other
- **public records related fields** i.e 'pub_rec' and 'pub_rec_bankrupcies' have **correlation(0.84)**
- **number of accounts fields** i.e 'open_acc' and 'total_acc' have **correlation(0.68)**

# Target Distribution

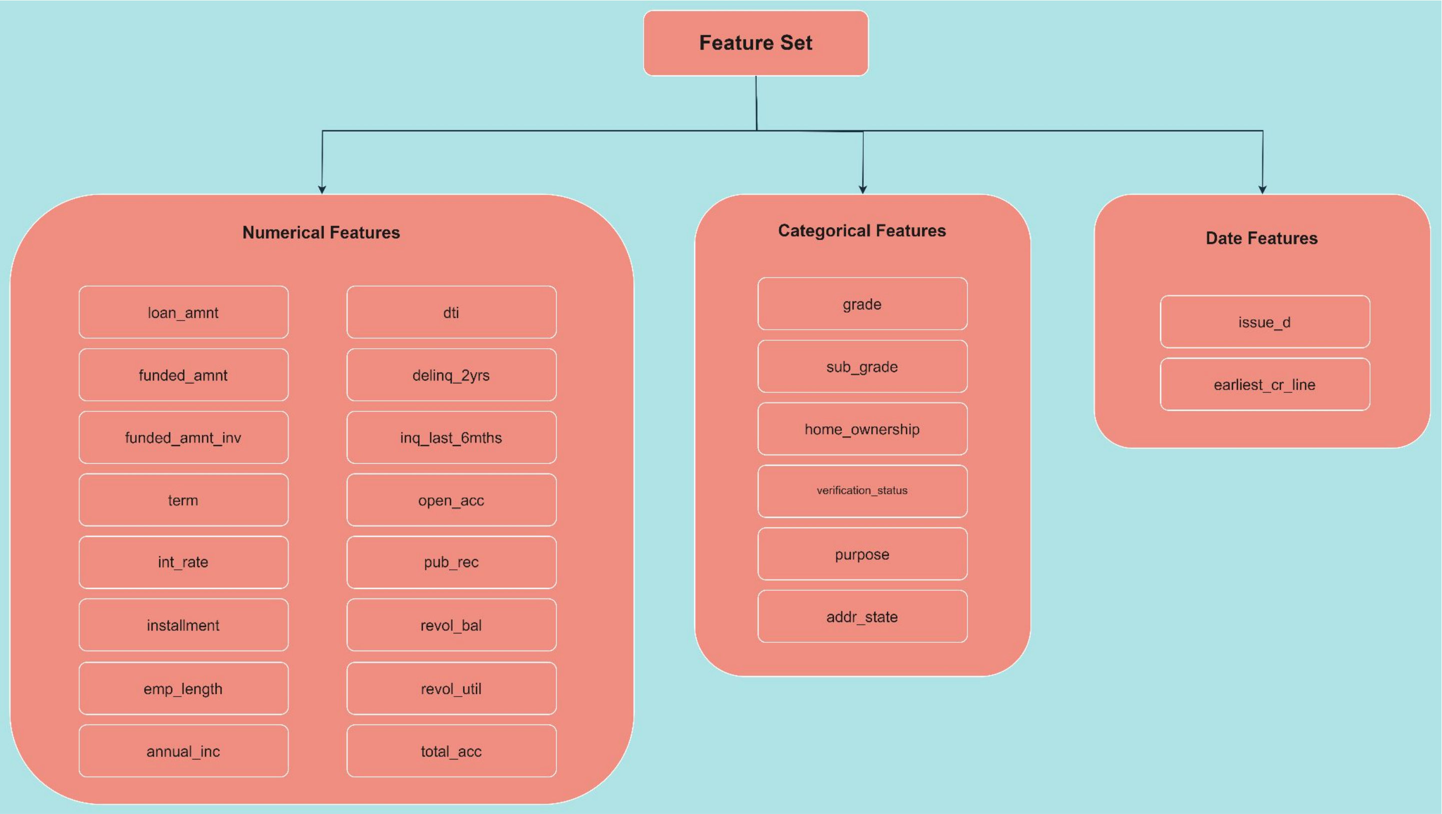


# Null Value Imputing

The following features have null values :

1. **emp_title** : filled with “**NA**”
2. **title** : filled with “**NA**”
3. **pub_rec_bankruptcies** : **imputed** with values of “**pub_rec**” values as both of them are **heavily correlated**
4. **emp_length** : filled with **mode** of emp_length i.e 10 years
5. **revol_util** : **dropped** the rows containing revol_util as null as they’re **quite insignificant** in number (**~0.1%**)

# Features Distinction



# EDA

# Categorical Columns Analysis

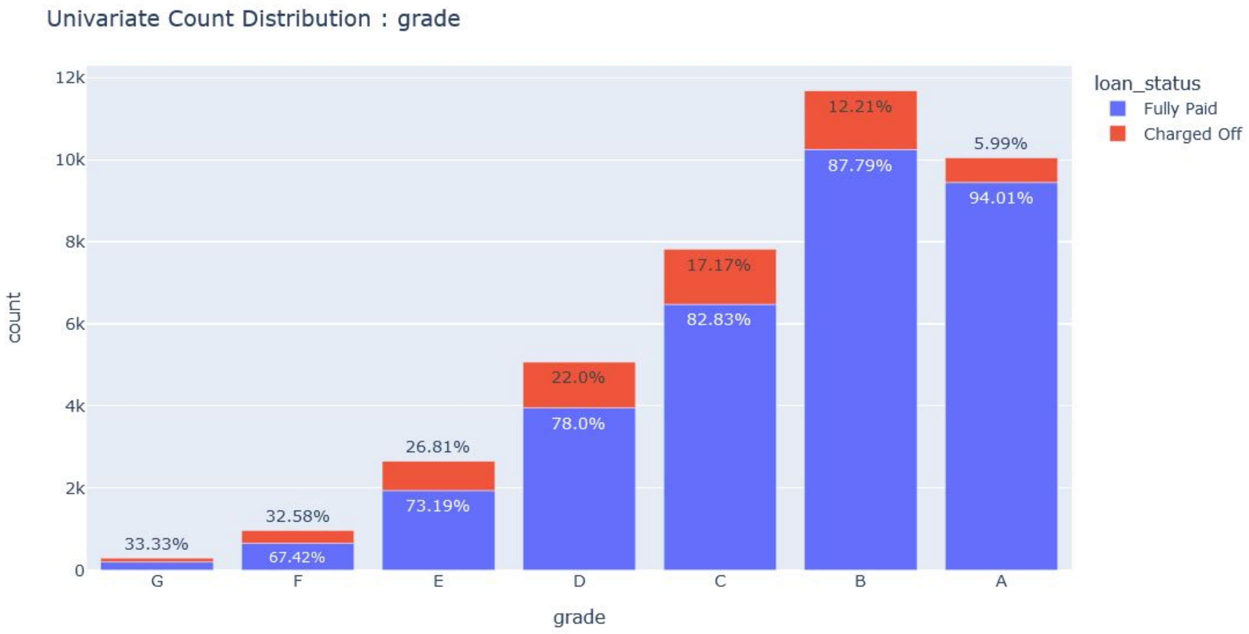
The **X Axis** in the plots is **sorted** as :

- **higher "charged off percentage"** to the **left** and **lower** percentage to the **right**

# Grade

- lower grades i.e G, F, E, D have so much higher defaulting percentages

- the lowest "Charged Off" percentage comes from A Grade of only 6% defaulters



# SubGrade

- The Grade "F5" is very very risky as it almost has ~50% of loan defaulting percentage

- The Grade "A1" is more than safe and has ~2% of loan defaulting percentage

- The right-most i.e safest bets for loan lending are very clear i.e A1, A2, .., B4, B5, etc so this means the internal grading algorithm of the lending club is very robust and reliable

# Verification Status

- Not Verified Customers have the least loan defaulting percentage i.e ~13% where as Verified Customers have 15-17%; which doesn't make sense

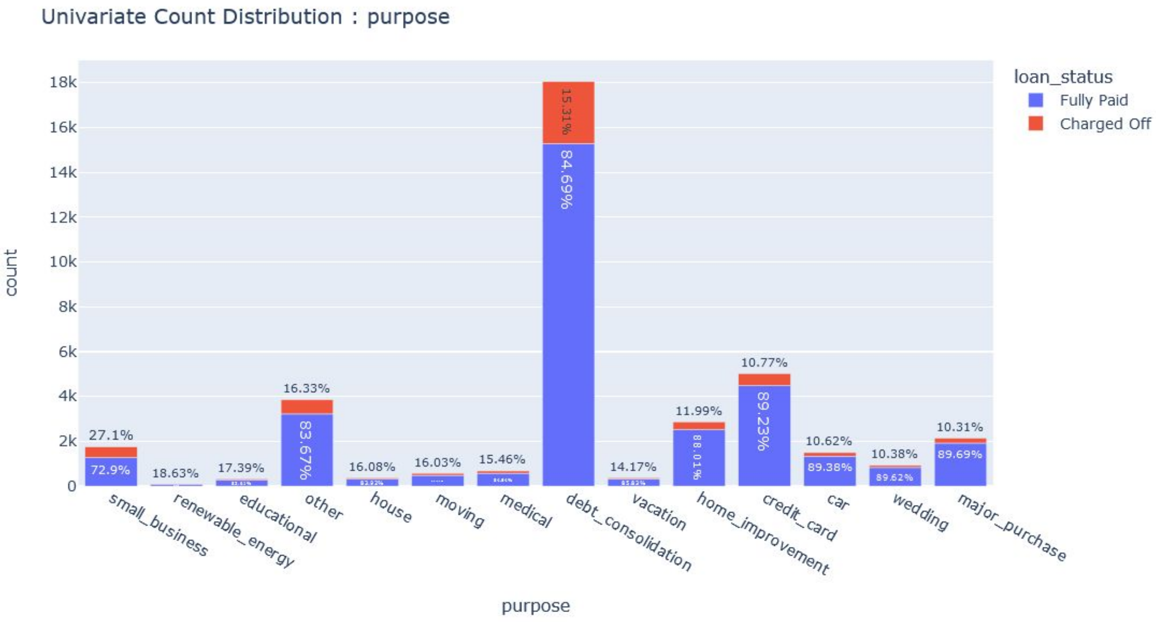
upGrad

# Purpose

- **"small business"** have the **highest** tendency of loan defaulting i.e **~27%**

- **Most of the loans** have a purpose of "**Debt Consolidation**" which has a okay-ish default percentage i.e **~15%** which the LC can live through

- "**Major Purchase", "Wedding", "Car' "Credit Card**" are the most **safe bets** as they have the **least loan default percentage i.e 10-11%**

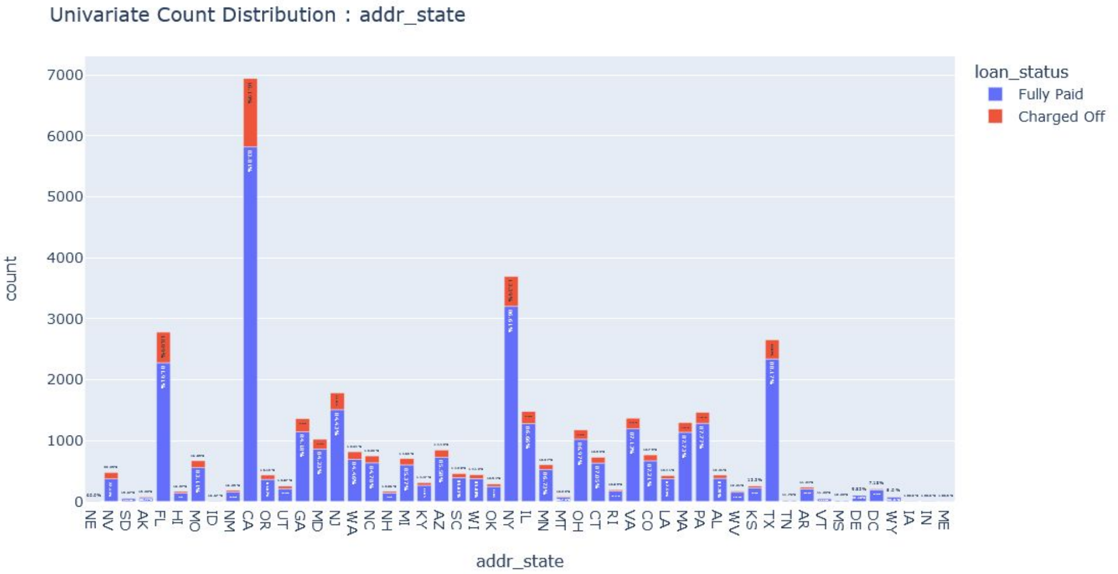


# Address State

- **Most loans are from CA** i.e Canada which also has a **higher default percentage i.e ~16.2% than normal**

- **FL** i.e Florida is also a state where the **loans are heavily taken and also has a higher default percentage i.e ~18%**

- **TX** - Texas, **PA** - Pennsylvania are **some good business making states** i.e have **go amount of loans taken** and also repaid too i.e have **lower default percentage <12%**



# EDA

# Numerical Columns Analysis

1. **Box Plot** : shows a **distribution of the loan status with respect to numerical column** we're analysing
2. **Distribution Plot** : shows a distribution of the **segmented numerical column**
3. **Yellow Trend Line** (top-right of each slide): shows the **loan default percentage trend line** (the **percentages are calculated for segmented numerical column**)

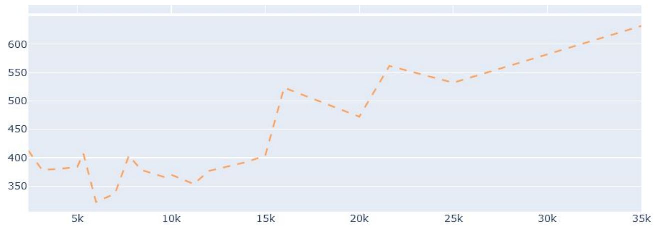
- **NOTE** : The **percentage in trend-line is scaled** up to show in the plot; you can **see the actual percentage by hovering on the line**

# Loan Amount

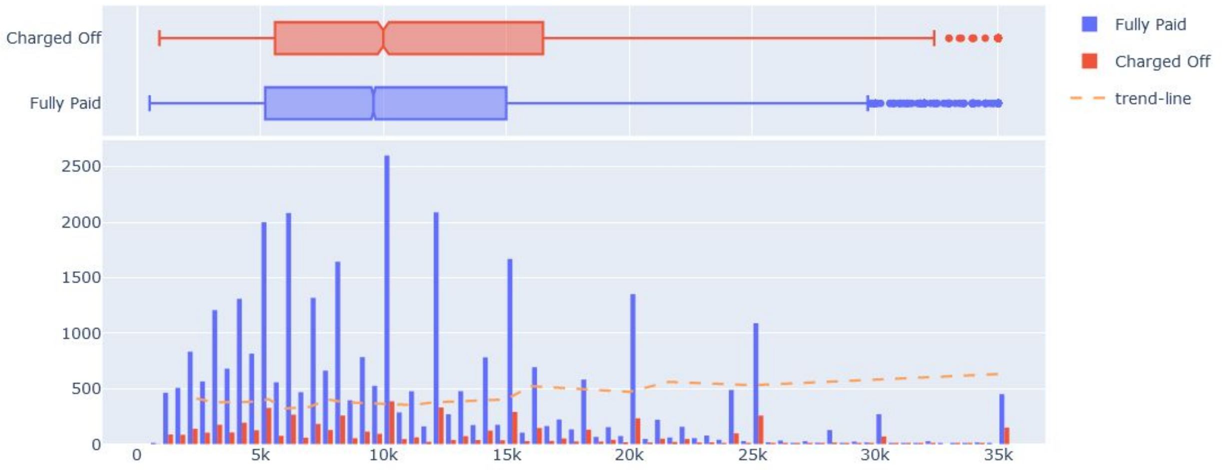
- The **loan defaulting percentages increases with Loan Amount taken** i.e steadily upward sloped trend line

- The **higher amount loans(> 20K) are having higher default rates (>17%)**

- As the **Funded Amount and Funded Amount Investor columns are heavily correlated** with this column; **they both also have similar trend lines**



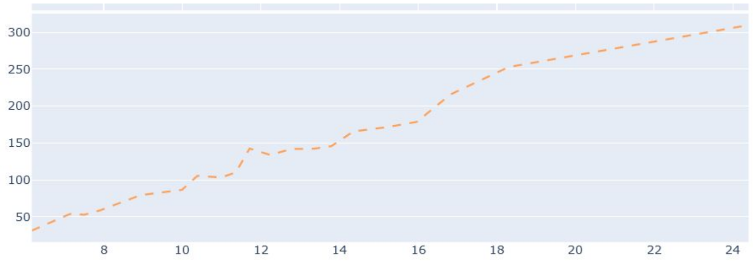

Univariate Numerical Distribution : loan_amnt



upGrad

# Interest Rates

- The **loan defaulting percentages increases with the Interest Rates** i.e steadily **upward sloped trend line**



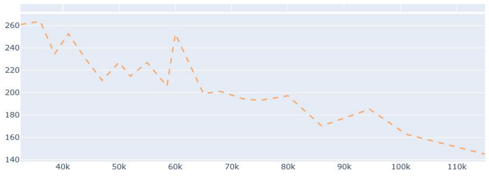

# Annual Income

- Having a **higher annual income the lower the default percentage** i.e **heavily downward sloped trend line**

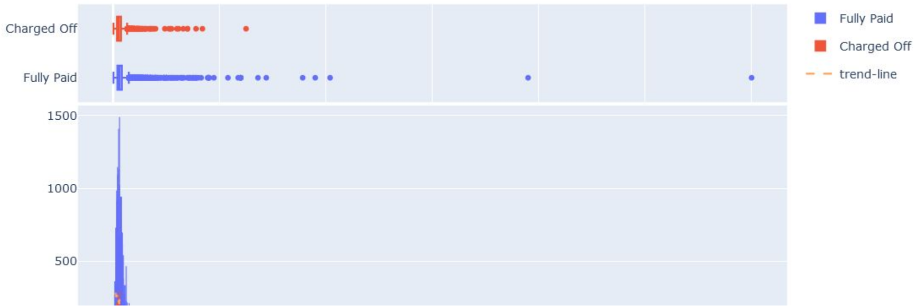
- Outliers on the right

- **Outliers are removed** by only taking data till the **95th percentile**

- As the annual income increases the loan defaulting percentage drastically comes down



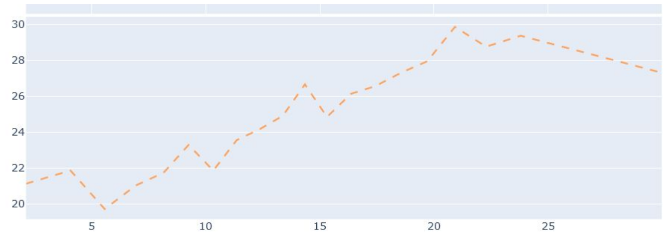

Before Outlier Removal : annual_inc



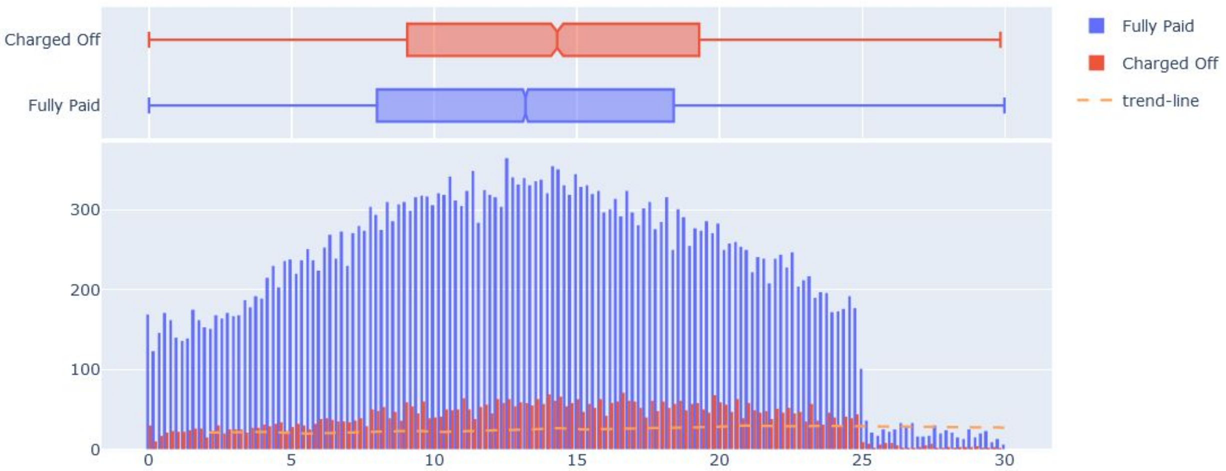

After Outlier Removal : annual_inc

# DTI

- The loan **defaulting percentages increases with the DTI** i.e steadily upward sloped trend line

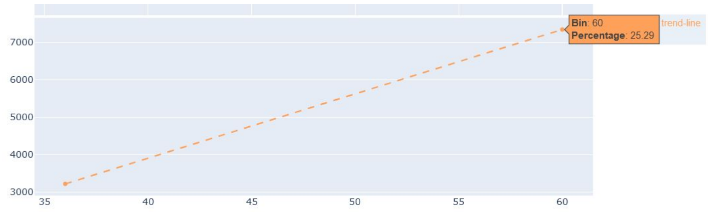



Univariate Numerical Distribution : dti

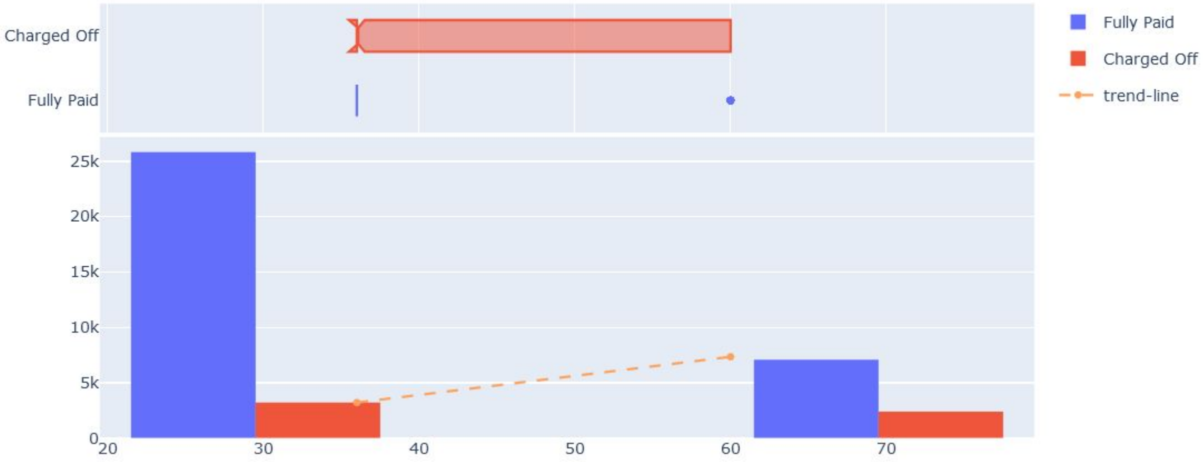


# Term

- Longer loans(60 months) has a very high default percentage (25%) compared to 36 Months loan (11%)



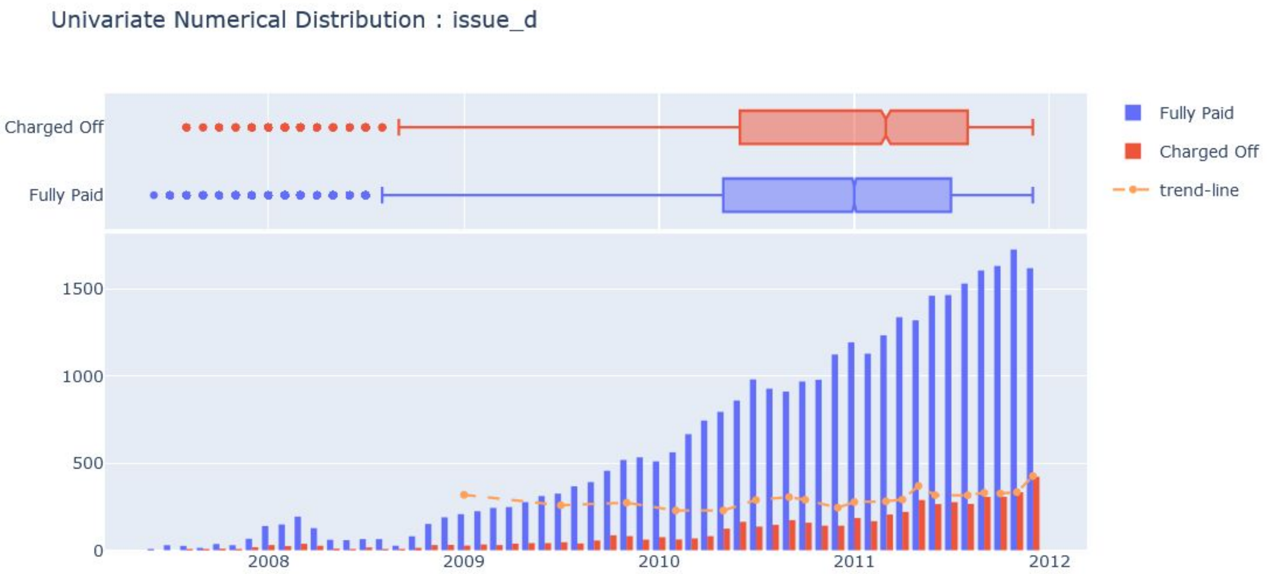

Univariate Numerical Distribution : term



# Issue Date

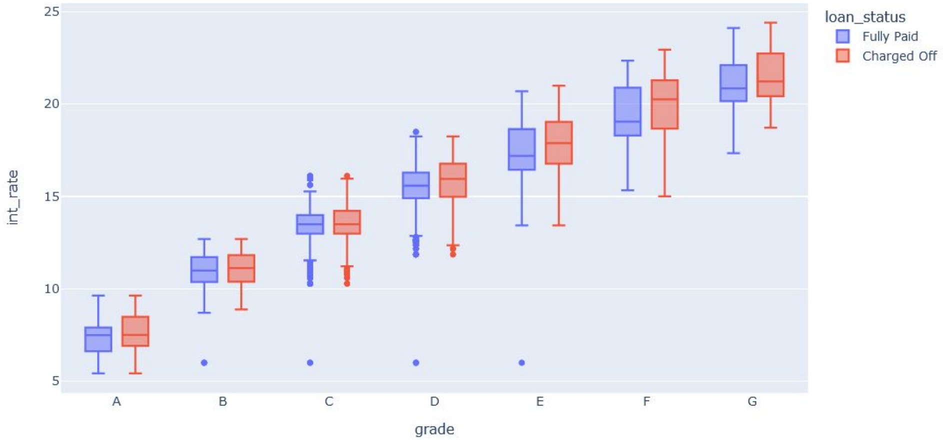
There's an **increasing loan defaulting trend** seen for the **latest approved loans** i.e **starting from Jan 2011**

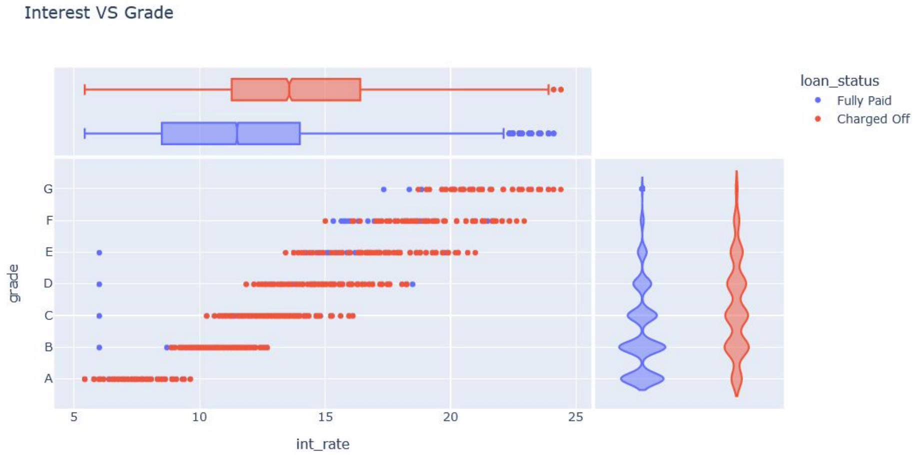
- Can be an effect of **Financial Crisis in 2011-2012 period** in US, Canada

# Bi-Variate Analysis

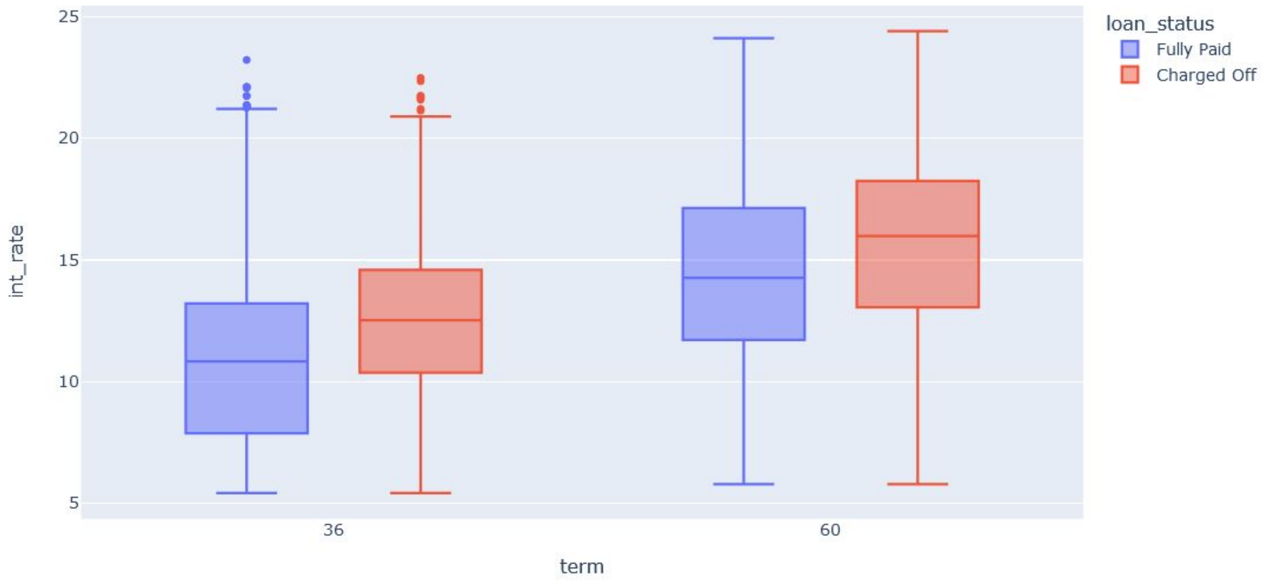
# Interest vs Grade



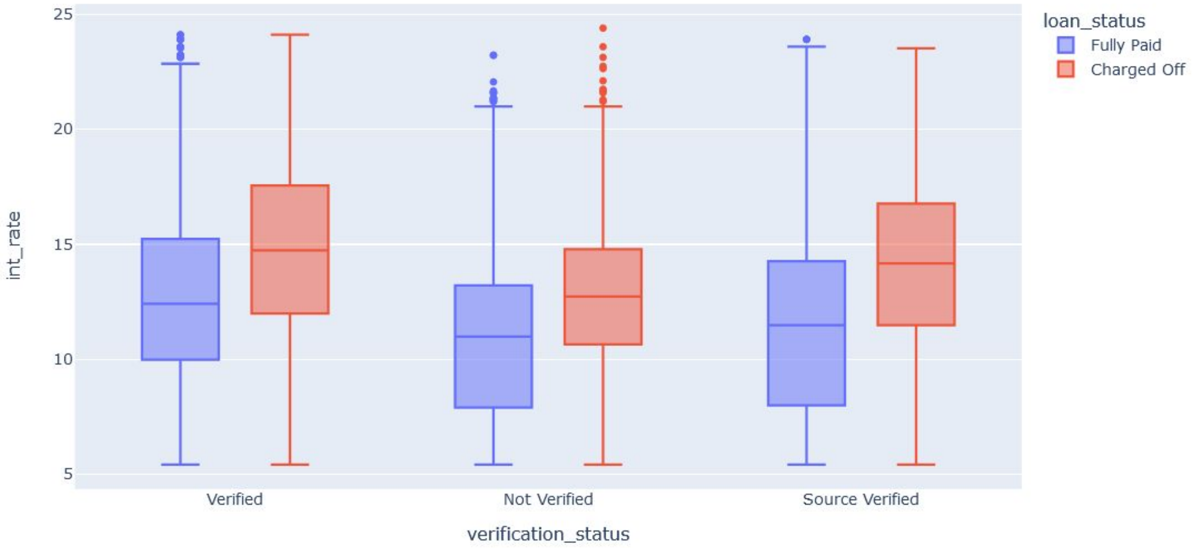

- It is clearly visible that as the **grade increases the interest rates linearly increase** too

# Term vs Interest Rate



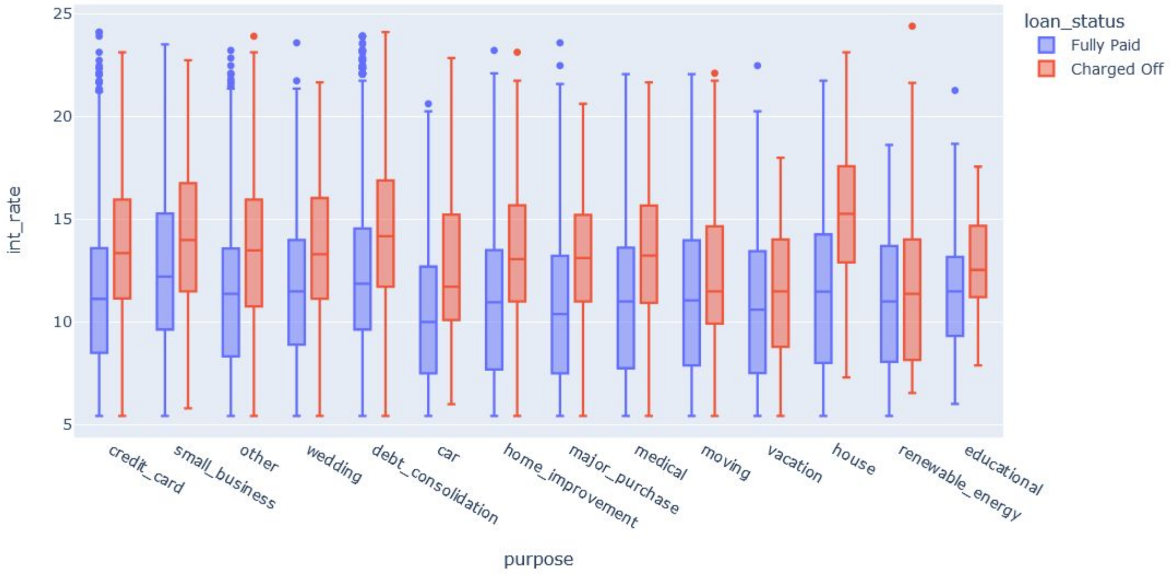

- Longer Termed Loans have higher interest rate margins

# Verification Status vs Interest Rate




- Why does "Verified" Sources of income have higher interest rate margins than "Not Verified" ones?

# Purpose vs Interest Rate



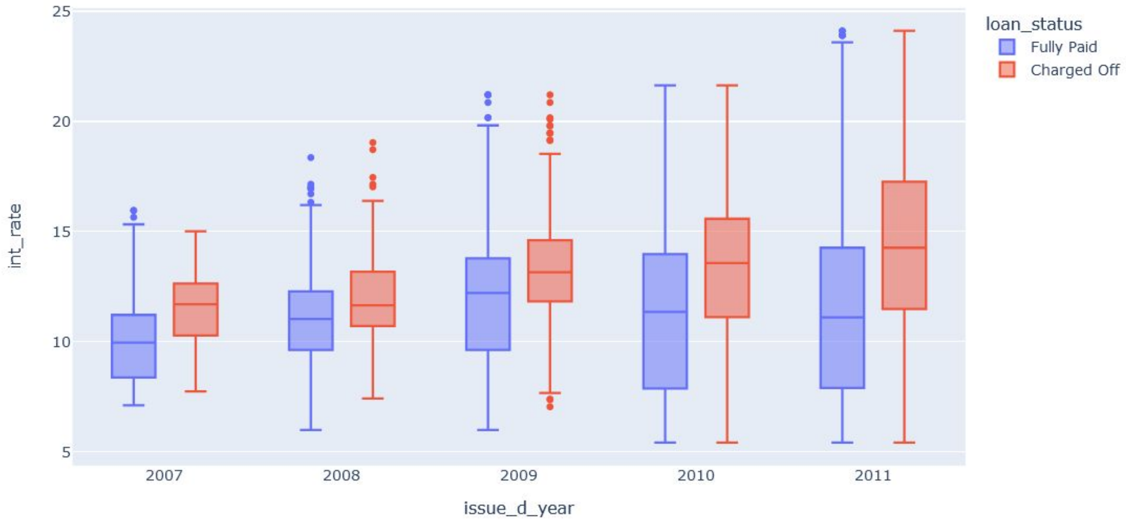

- Purpose = "House" has unusually high interest rates(mean: 15.5) for "Charged Off" applications
- Purpose = "Small Business" have normal interest rate margins but we've already seen that it is the most riskiest sector of loan lending i.e these have the highest loan defaulting rates (27%)

# Issue Year vs Interest Rate

- **2008, 2009** have smaller boxes(quantiles) showing that the **LC usually used to play very safe** by giving loans on usual industry set interest rate (9-14%)

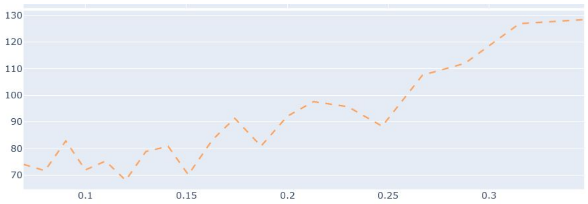
- But **starting from 2010**, the **LC has lowered the profit margin** (**decrease in interest rate) with giving out loans at a lower interest rates (7-14%)**

- The **clear increase in interest rates for "Charged Off" loan applications** is very clearly visible, example : in **2010**: the **25% percentile** of interest rate for **"Fully Paid"** applications was **7.88**, whereas was **11.12** for "Charged Off" applications which is a very **huge increase in interest rate**

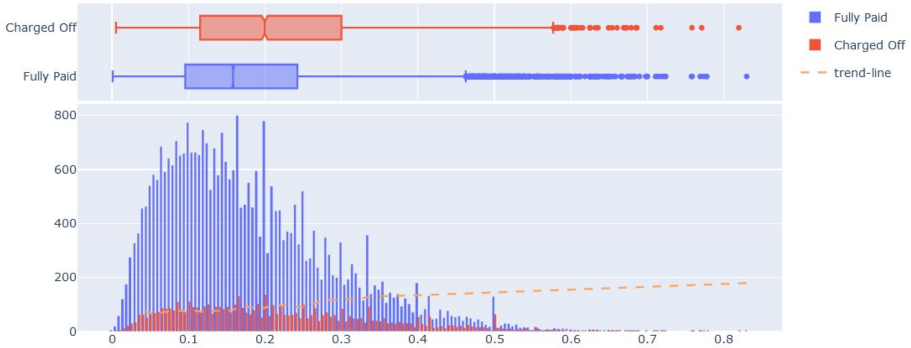


upGrad

# Loan Amount to Annual Income Ratio




Before Outlier Removal : loan_to_annual_income_ratio




After Outlier Removal : loan_to_annual_income_ratio


- We can clearly see in the plots that **higher the ratio, higher the defaulting tendency**

# 4. Business Drivers & Recommendation

| | | |
|---|---|---|
| 01 | Loan Amount to Annual Income Ratio | • OBSERVED : This derived ratio is directly proportional to the loan defaulting percentage<br>• RECOMMENDATION : Interest Rates should also increase proportionally with this metric |
| 02 | Purpose | • OBSERVED : "Small Business" have normal interest rate margins but we've already seen that it is the most riskiest sector of loan lending i.e these have the highest loan defaulting rates (27%)<br>• RECOMMENDATION : Historically riskier sections should have higher interest rate margins and vice-versa |
| 03 | Issue Date | • OBSERVED :<br>○ There's an increasing loan defaulting trend seen for the latest approved loans i.e starting from Jan 2011<br>■ Can be an effect of Financial Crisis in 2011-2012 period in US, Canada<br>• RECOMMENDATION : Observing the current loan defaulting trend and global economic situation; the loans should be given at a higher rate than normal as we've got the highest loan-defaulting rates currently |

# 4. Business Drivers & Recommendation

| | | |
|---|---|---|
| 04 | Loan Term | • OBSERVED : Longer loans(60 months) has a very high default percentage (25%) compared to 36 Months loan (11%)<br>• RECOMMENDATION : Longer Loans should have higher interest rates compared to shorter-term loans |
| 05 | Interest Rates | • OBSERVED : The loan defaulting percentages increases with the Interest Rates |
| 06 | Public Image | • OBSERVED : As soon as there is a public derogatory record or a publicly recorded bankruptcy the loan default rates shoots up and also the interest rates shoot up<br>• RECOMMENDATION : Should have higher interest rates if the public image is not good because the loan-defaulting can go upto 33%(some cases) hence for such high risk -> set higher Interest Rate |
| 07 | Verification | • OBSERVED : “Not Verified” customers have the least loan defaulting percentage i.e ~13% where as Verified Customers have 15-17%<br>• RECOMMENDATION : Verification process should be looked into and corrected as we should be having lower risk for Verified Profiles and higher for Not Verified |

# END!

# 5. Appendix

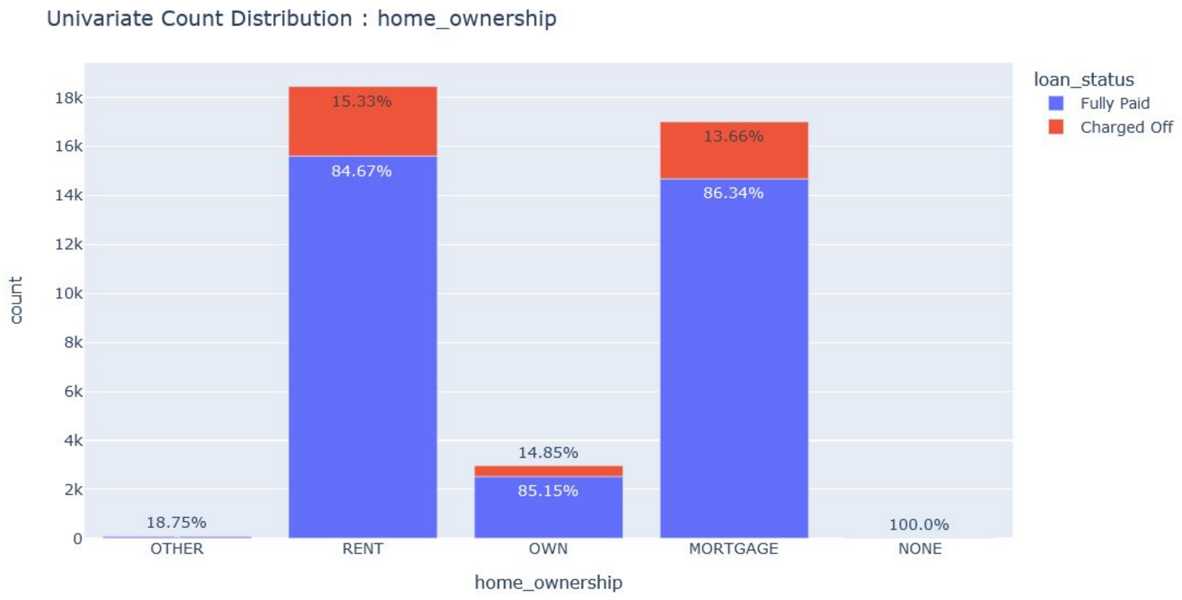
Attaching all the other feature plots which we deemed weren't as important.

upGrad

# Home Ownership

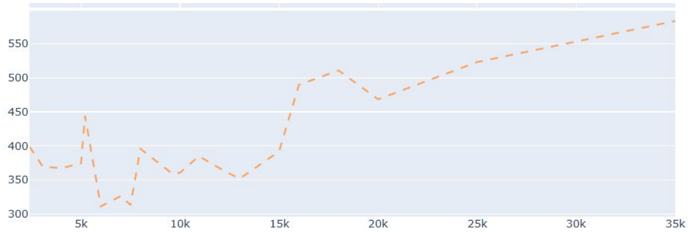
- People who have own homes tend not to take loans as compared to people who are rented and on mortgage

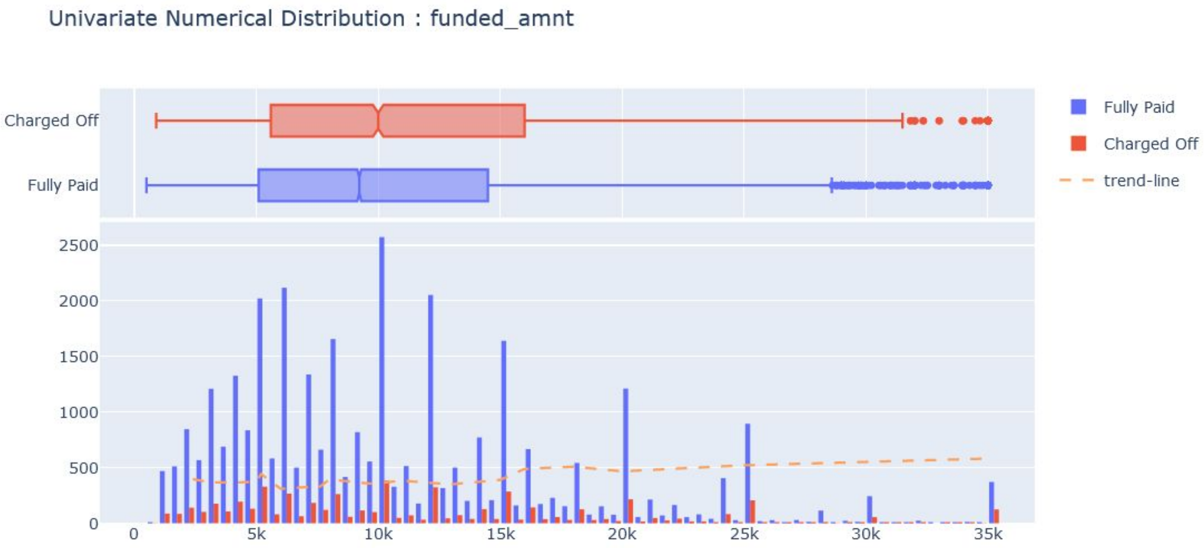
- It doesn't matter who the loan is given to i.e Rented, Mortgaged, or Own Home; the loan defaulting percentage is almost ~15%



# Funded Amount

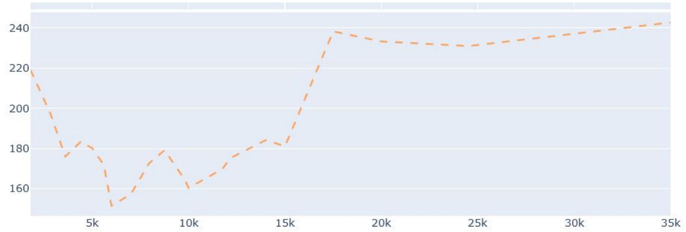
- Similar trend line as Loan Amount i.e higher amount loans(> 20K) are having higher default rates (>17%)

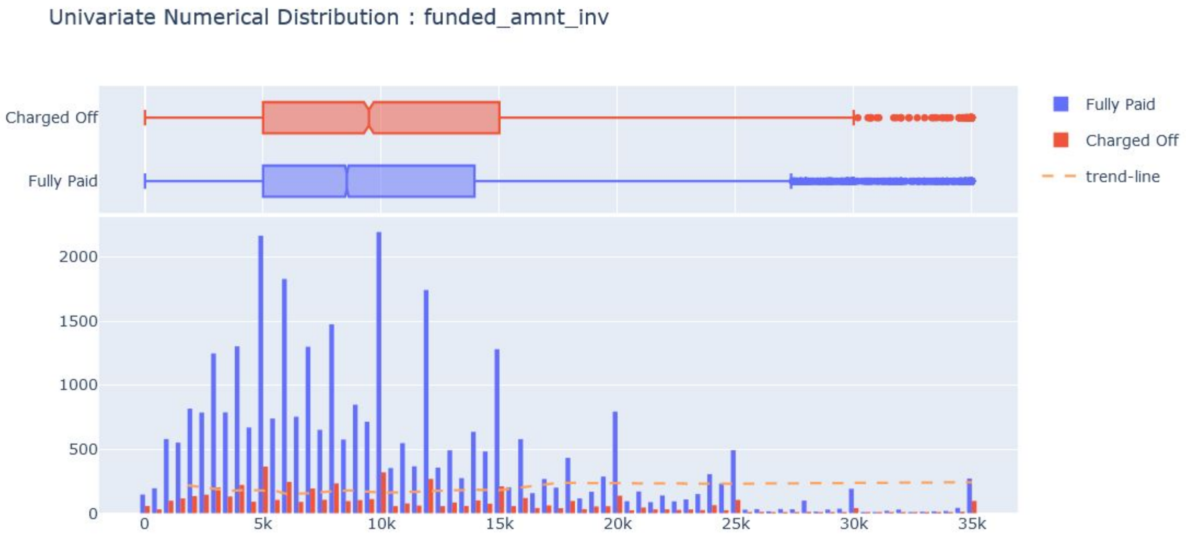


# Funded Amount Investor

- Similar trend line as Loan Amount i.e higher amount loans(> 20K) are having higher default rates (>17%)

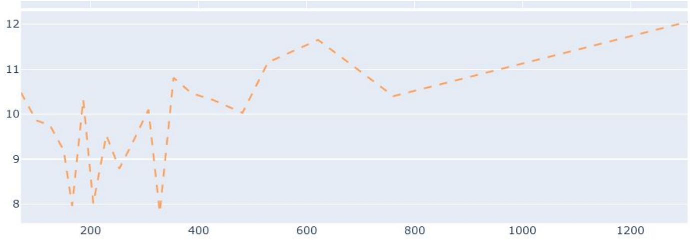


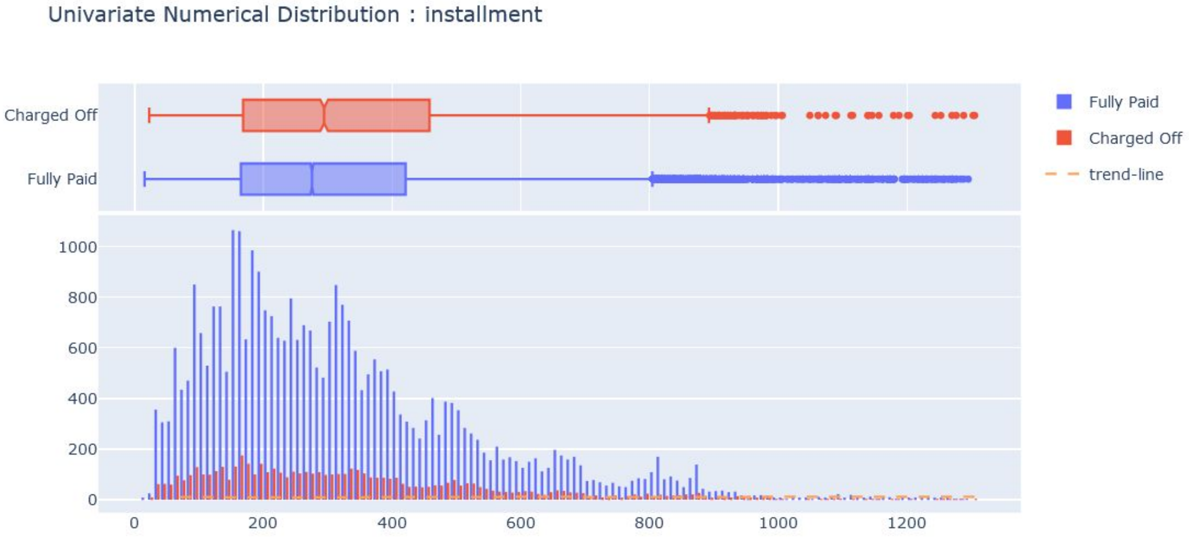



# Installment Amount

- outliers on the right i.e having many total accounts

- Actions : Perform an outlier removal



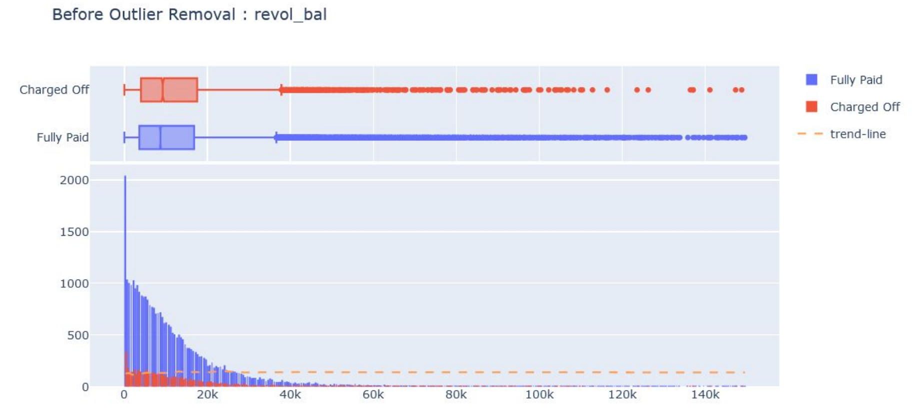



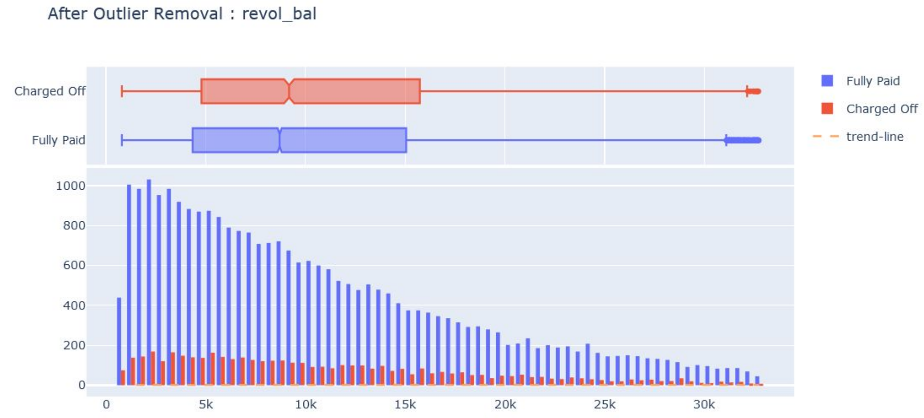
# Revolving Balance

- outliers on the right i.e having many total accounts

- Actions : Perform an outlier removal

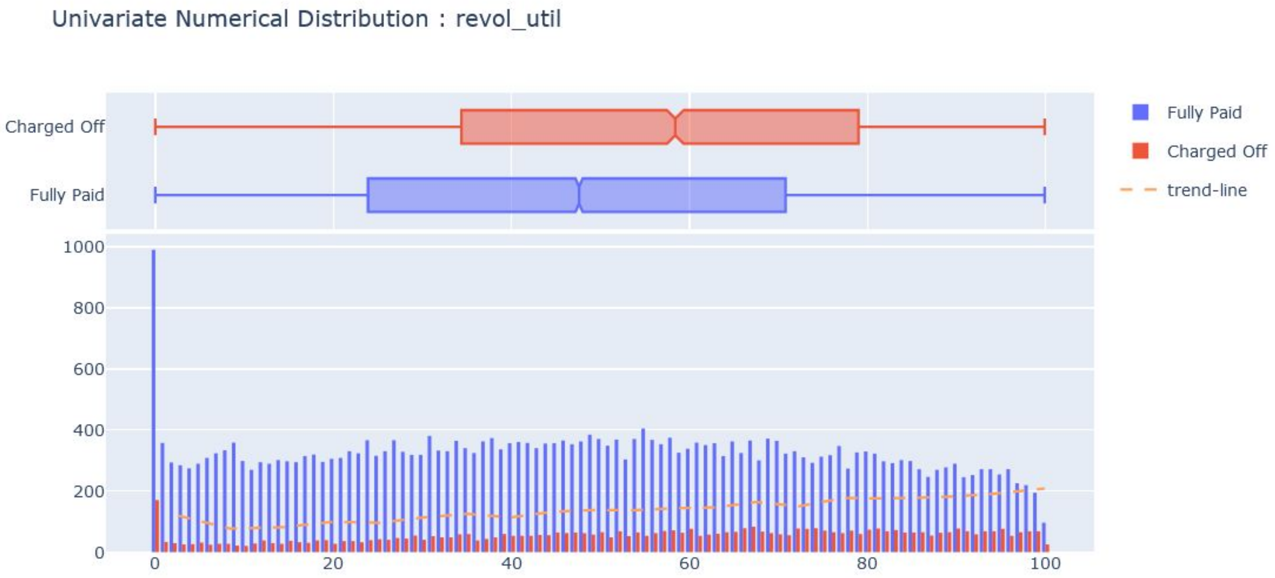
- As the revolving balance increases the loan defaulting percentage goes up too



# Revolving Utilization Rate

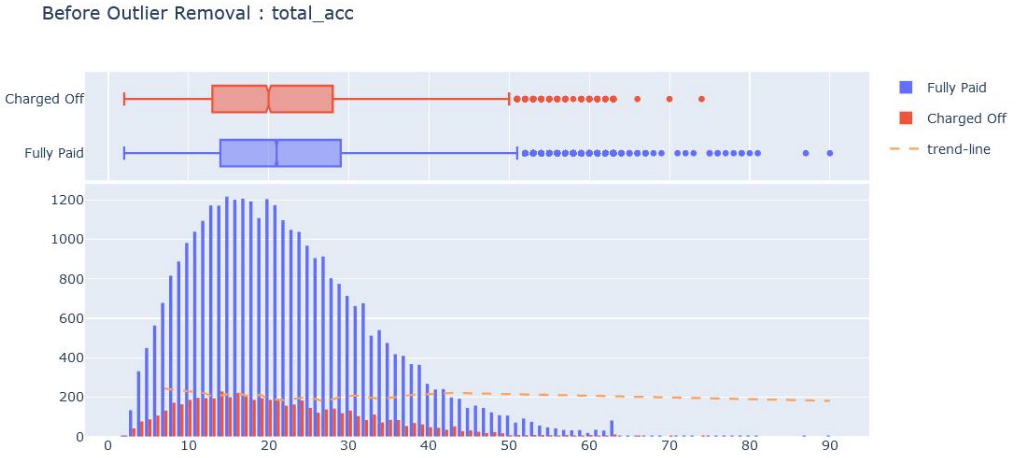
- The loan defaulting percentages increases with the Revolving Utilization i.e heavily sloped upward trend line

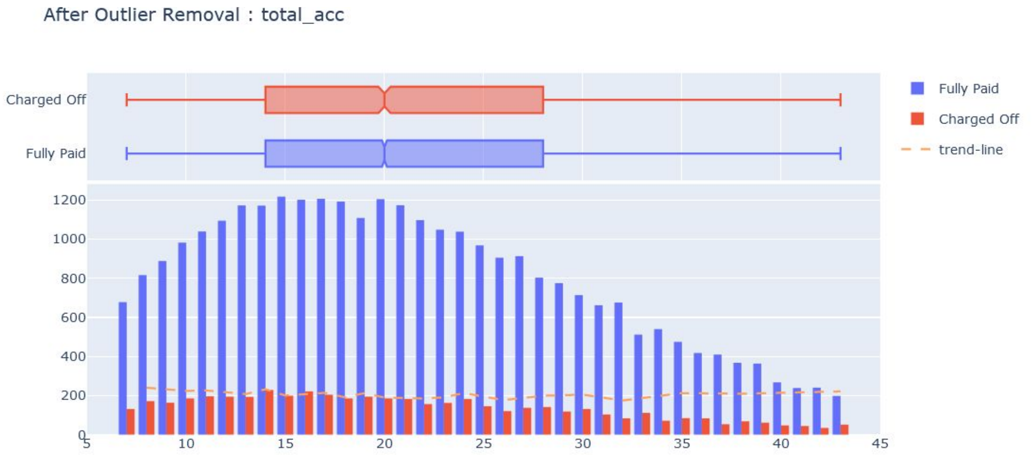


# Total Accounts

- outliers on the right i.e having many total accounts

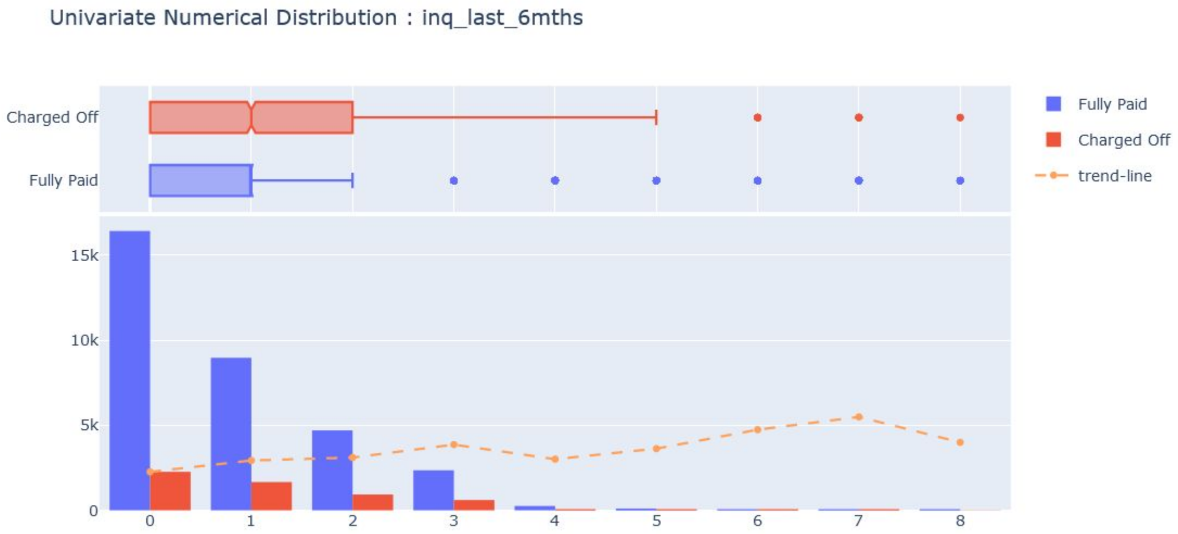
- Actions : Perform an outlier removal



# Number of Inquiries in past 6 months

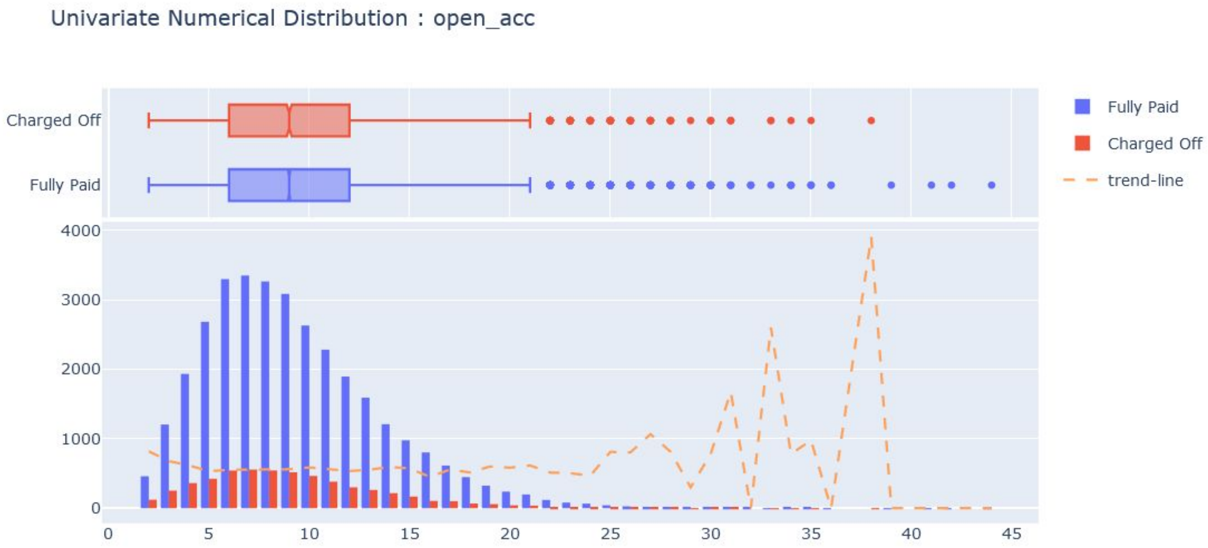
- As the number of inquiries increase the loan default rates also increase than normal i.e has a upward trend line



# Open Accounts

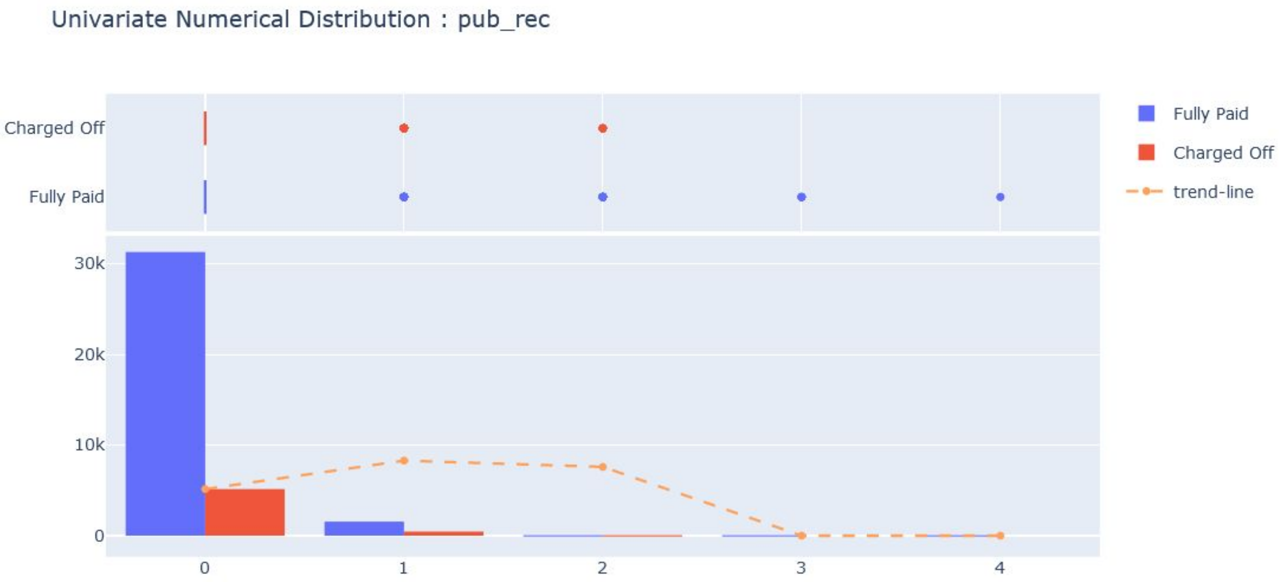
- Due to outliers on the right i.e having many open accounts the the trend line is not very stable

- Actions : Perform an outlier removal

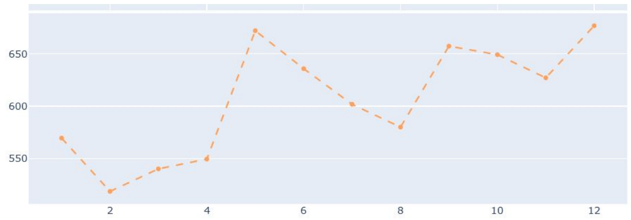


# Derogatory Public Records

- The Loan defaulting percentages increases as soon as you have a derogatory public record (> 23%)

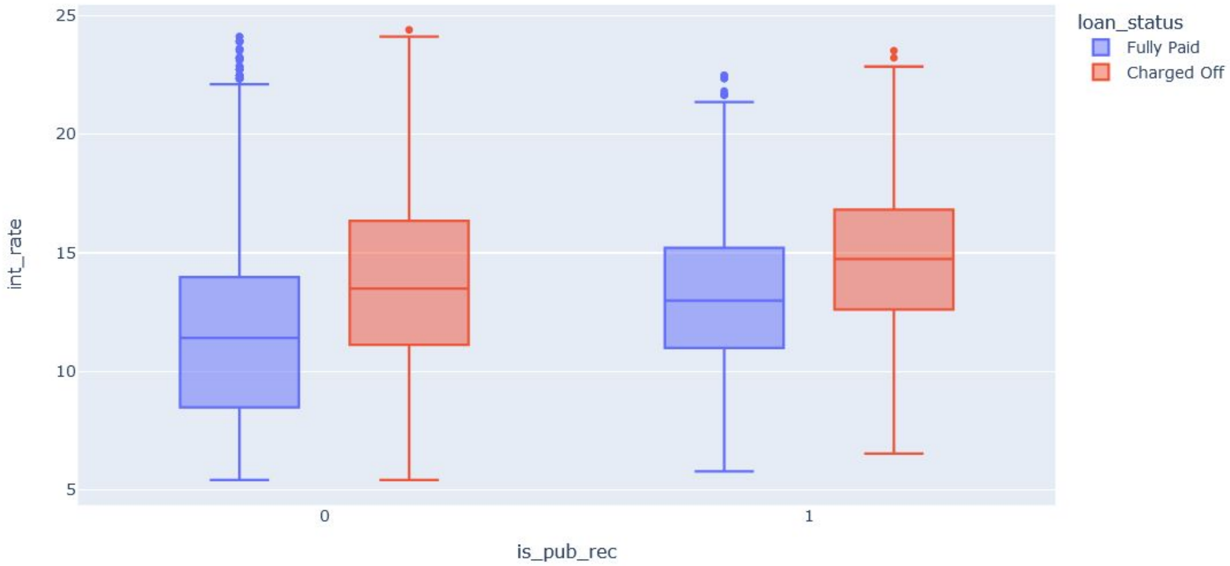


# Issue Date - Month




- Loans issued in Feb, March, April have lower loan default rates (12-13%) than normal (15-16%)

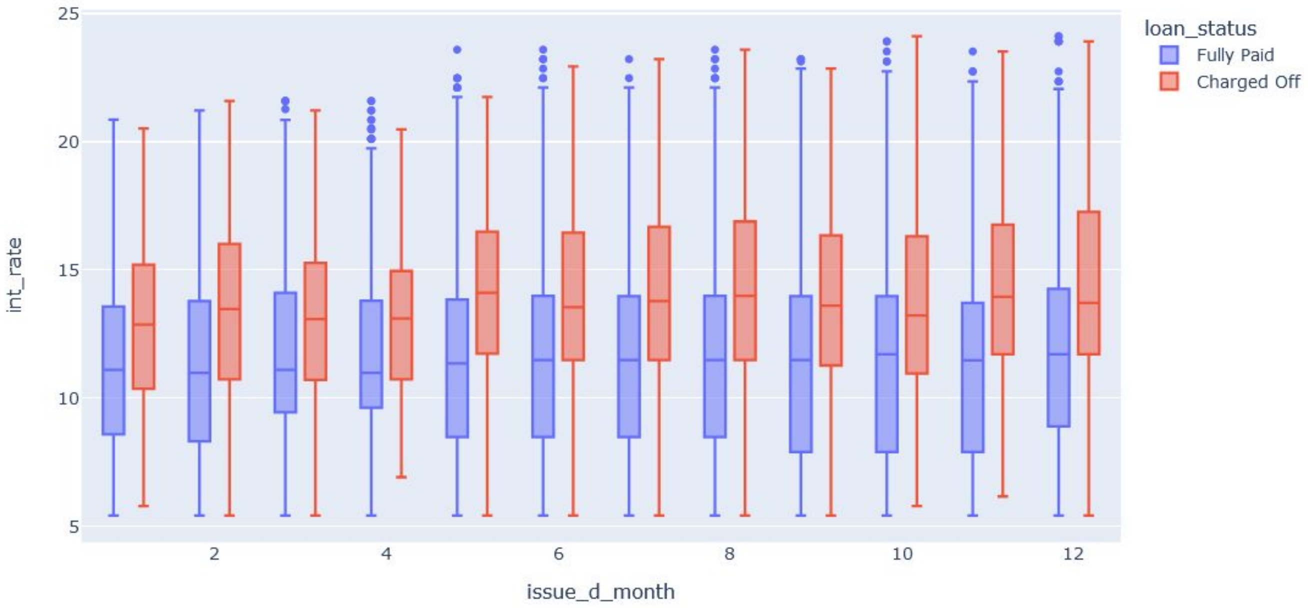
# Public Records vs Interest Rate




Created a derived feature i.e is_pub_rec : Binary Feature whether there is a public derogatory record of the applicant or not

- As soon as there is a public derogatory record the interest rates shoot up

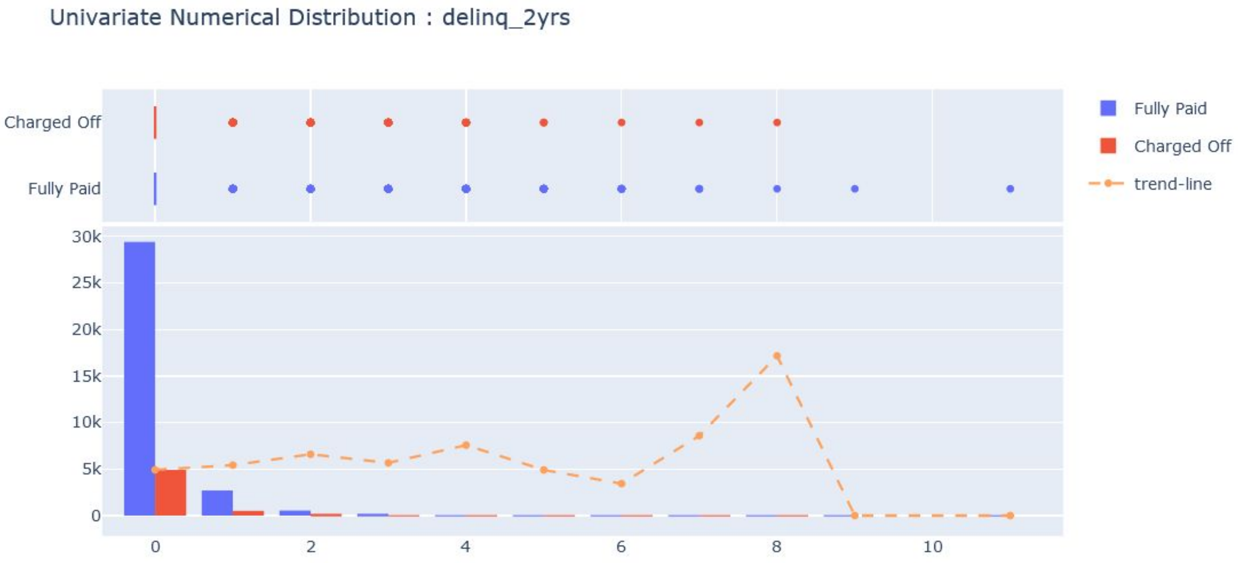
# Issue Month vs Interest Rate




- Nothing interesting information gain except that lower limits (25 percentile limit) for loans in months March and April are higher than normal i.e ~9.5 wheras the usual interest rate is around 8.5

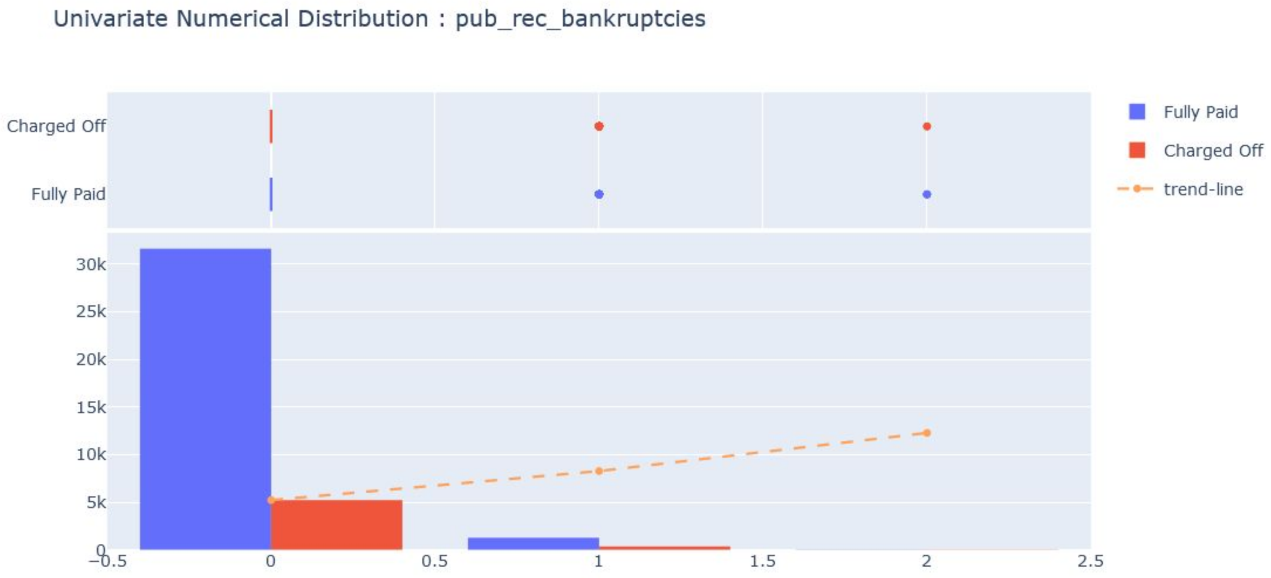
# Number of Delinquency in past two years

The Loan defaulting percentages increases as soon as you have a delinquency i.e has a upward trend line



# Public Record Bankruptcies

- The Loan defaulting percentages increases as soon as you have a publically recorded bankruptcy (22% and 33% for 1 and 2 bankruptcies reported respectively)

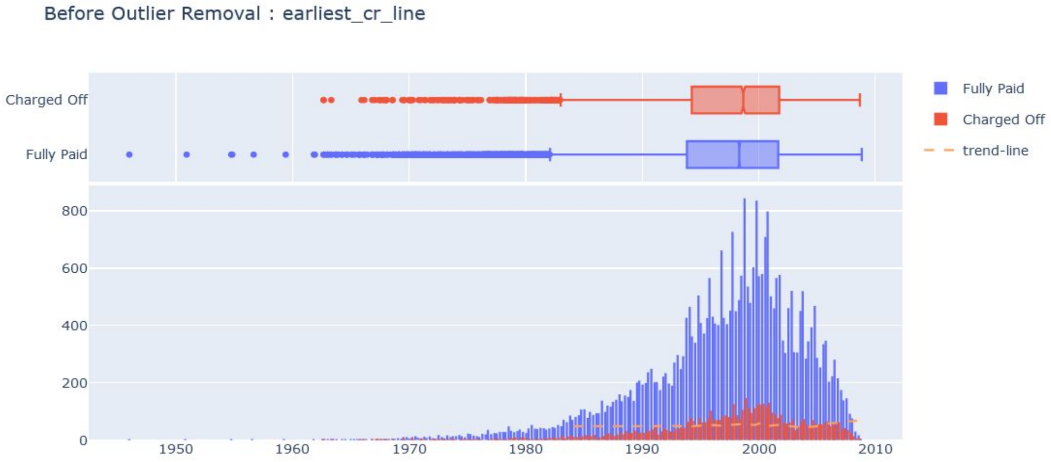


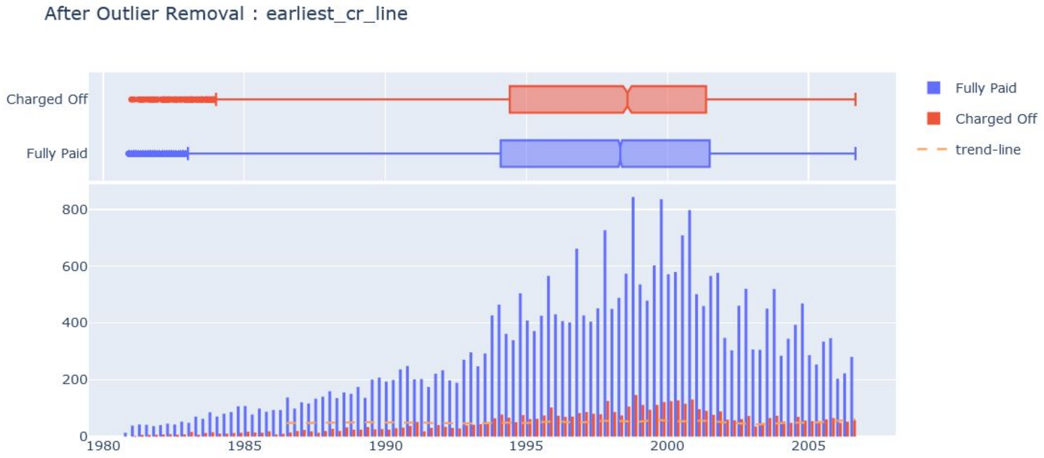
upGrad

# Earliest Credit Line

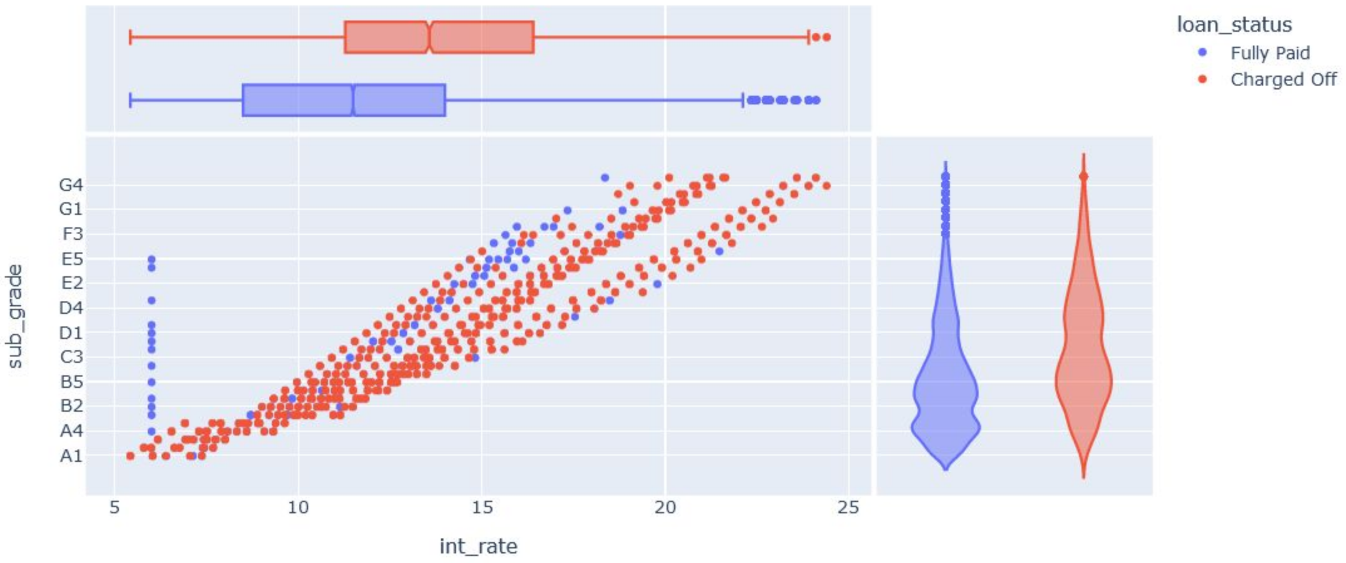
- Heavily Left Skewed i.e too many outliers in the left (outlier-ish loan applications which have very old applicants)

- Actions : Perform Outlier Removal



# Interest vs SubGrade




- It is clearly visible that as the sub-grade increases(lower grades) the interest rates linearly increase too