# Project Assignment for Data Science on Exploratory Data Analysis

By

Darakhsha Anwar

&

Mohd Reeyaz

# LIST OF CONTENTS

1. Prerequisite
2. Import Library and set required parameters
3. Data sourcing
4. Data Cleaning and Manipulation
5. Derived Metrics
6. Univariate Analysis
7. Bivariate Analysis
8. Bivariate Analysis and Multivariate Analysis with Probability function
9. Conclusion

# CREDIT EDA CASE STUDY

## 1. Prerequisite:

1.1 Place load ‘Dataset.csv’ input file at directory before running this code.

1.2 Please make sure that you have following python libraries like Numpy,  Pandas, Scipy,  Seaborn,at your system.

## 2. Import Libraries and set required parameters:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline from scipy
import stats
```

# CREDIT EDA CASE STUDY

```
# seaborn : Advanced visualization
    import seaborn as sns
#Set it to None to remove
    pd. options. mode. chained_ assignment = None
    pd. options. display. float_ format = '{:.4f}'.format
 #Set it to convert scientific notations' such as 8.775e+11 to
    422510842796.00.
#Display required Columns
    pd.set.option('display.max_columns', 100)
```

# CREDIT EDA CASE STUDY

## 3.Data Sourcing:

3.1 df = pd. read _ csv('../input/.csv')

3.2 print(df.shape)

3.3 (307511, 122)

3.4 df.info()

3.5 df.describe

## 4.List of Columns & Missing Values are more than 50 Percent:

1. As we can see there are **307511** rows & **122** columns in the dataset.
2. It will be very difficult to look at each column one by one & find the NA or missing values.
3. So let's find out all columns where missing values are more than percentage(**50%**).
4. We will remove those columns as it is not feasible to impute missing values for those columns.

# Imputing Missing Values

1 For ‘AMT_GOODS_PRICE’, we see that the mean is 538396.21 and median is 450000. So, it would be better to impute the missing values with the median (i.e:450000) as the mean values is quite higher.

2. For ‘NAME_TYPE_SUITE’ column, we see that it’s a categorical data. So, for a categorical data it would be better to replace the missing values with the mode, which is NAME_TYPE_SUITE could be fill in by ‘unaccompanied’

3. For ‘EXT_SOURCE_2’ column the mean is 0.514 and median is 0.565 which is quiet close to each other. However, it is always a good pratice to impute the missing values with the median, as the mean might get affected. If a new value is added in the column.

4. For ‘OBS_30_CNT_SOCIAL_CIRCLE’ column, we that the mean is 1,422 and median as well as the mode turns out to be 0.0. So, how again it’s a better option to impute the missing values with median.

5. For ‘DEF_30_CNT_SOCIALL_CIRCLE’ column, we observed that mean is 0.143 and median and mode is 0.0. So, we can conclude that the missing values in the column can be imputed by the median

# CREDIT EDA CASE STUDY

## 5. Data Cleaning and Manipulation:

5.1 RemoveNulls(dataframe, axis, percent) will drop the columns/rows from the dataset based on the parameter values

5.2 Remove Null: Remove Null function will remove the rows and columns based on parameters provided.

5.3 dataframe : Name of the dataframe

5.4 Axis : axis = 0 defines drop rows, axis =1(default) defines drop columns

5.5 Percentang : percentage of data where column/rows values are null, default is 0.5(50%)

5.6 Remove columns and rows where NA values are more than or equal to 50%

5.7 Remove columns where number of unique value is only 1.

columns based on the count . Now let's look at each column from business perspective if that is required or not for our analysis such as Unique ID's.

5.8 Remove irrelevant columns: Till now we have removed the

5.9 Cast all continuos variables to numeric so that we can find a correlation matrix between them>

# CREDIT EDA CASE STUDY

## 6.Loan :

We will analysis only those categories which contain more than 80% of records.

## 7.Derived Metrics:

7.1.We will now derive some new columns based on our business understanding that will be helpful in our Risk analysis.

7.2. Loan amount to Annual Income ratio

7.3. Extract Year & Month for previous_data and application_data

7.4. Arranged order of months from Jan to Dec as per the alphabets order.

7.5. Create Bins for range of Loan Amount

7.6. For Example: {0 to 5000, and 5000 to 20000} as per the requirement.

7.7 Create Bins for range of Annual Income & Interest rates

# CREDIT EDA CASE STUDY

## 8.Univariate Analysis

### 8.1 Continuous Variables:

In case of continuous variables, we need to understand the central tendency and spread of the variable.

These are measured using various statistical metrics visualization methods such as Boxplot, Histogram etc.

### 8.2 Categorical Variables

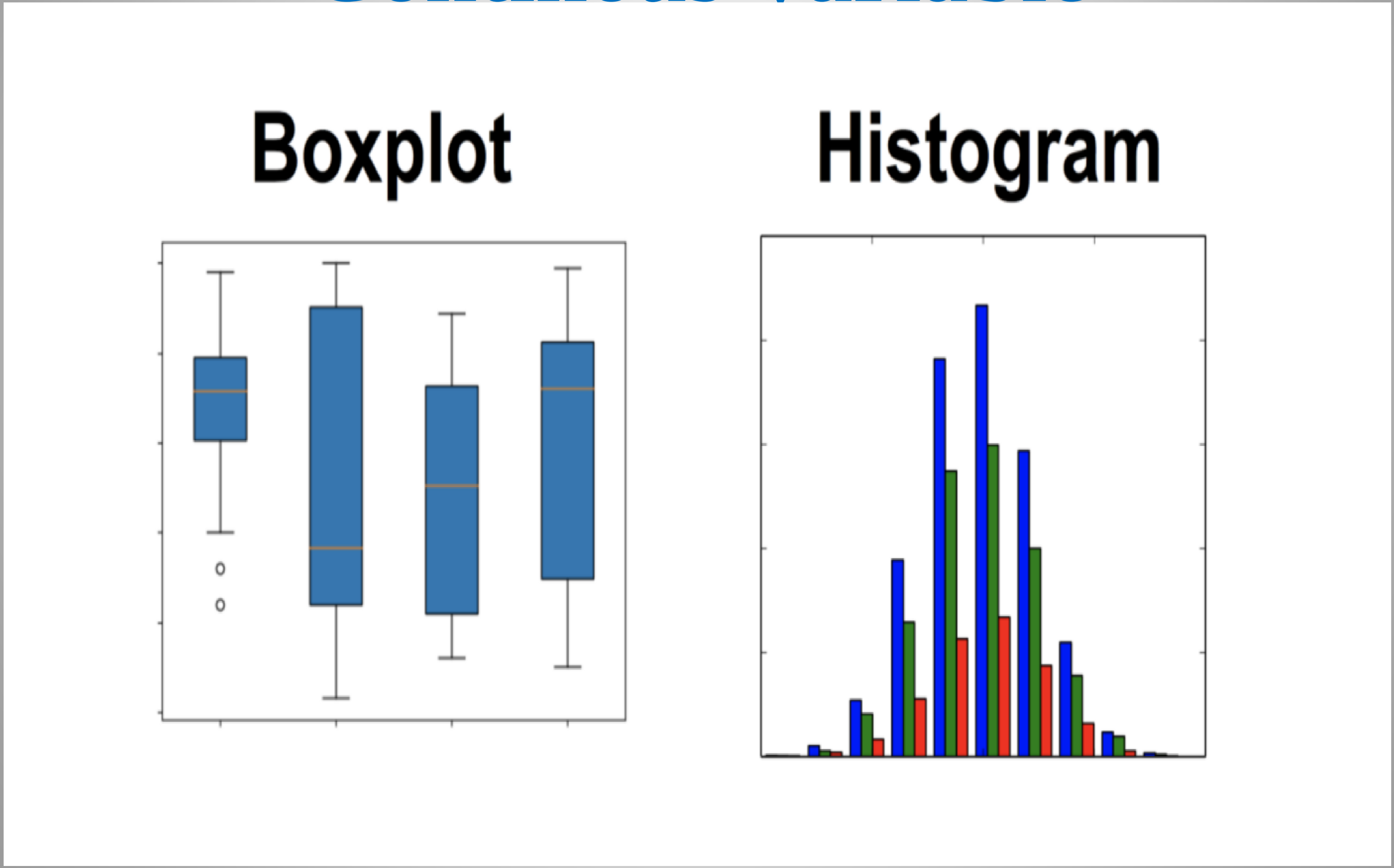
For categorical variables, will be use frequency  to understand distribution of each category.

It can be measured using two metrics, Count and Count% against each category. Bar chart can be used as visualization.

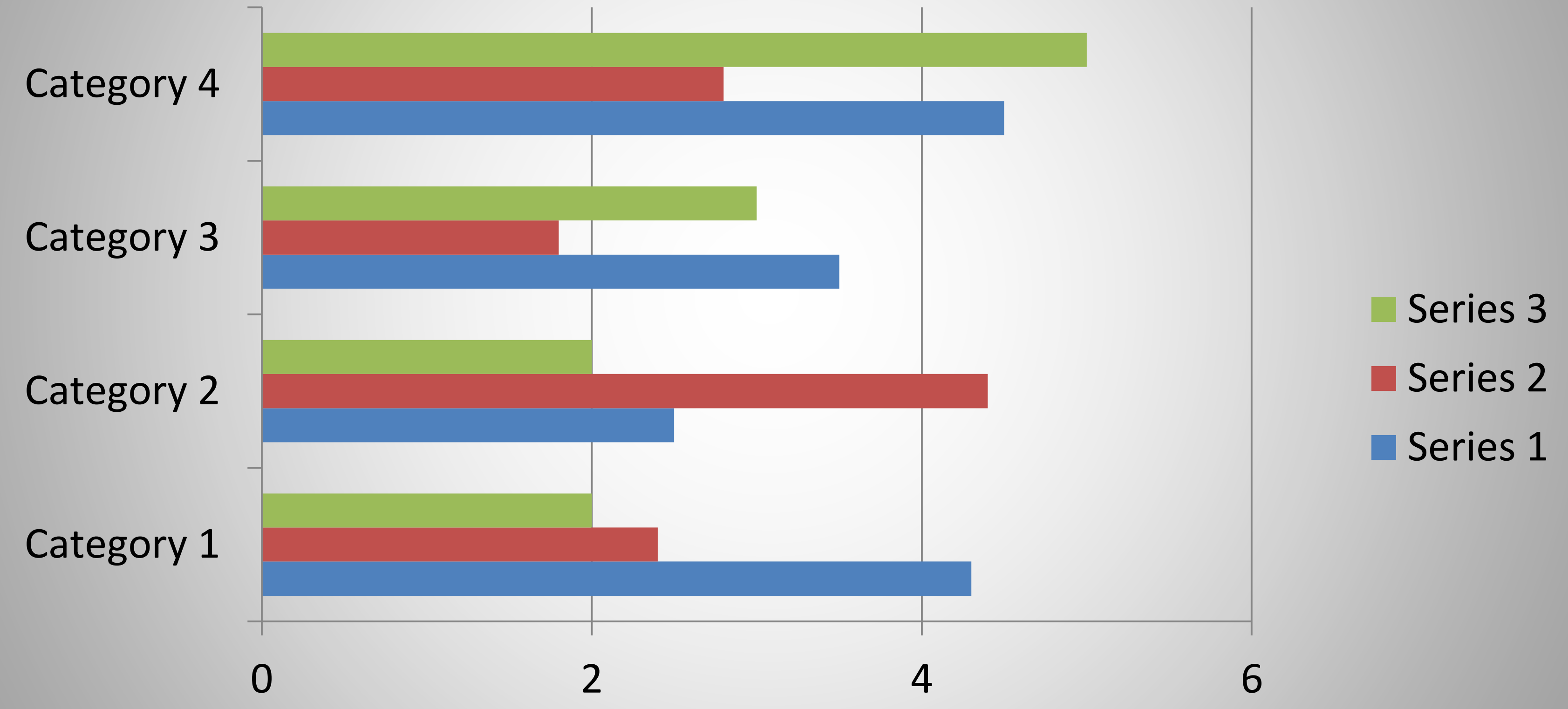
# Univariate Analysis For Categorical Data

1. From the first plot, ‘NAME_CONTRACT_TYPE’ with respect to TARGET variable, we observe that in comparison to who opted for Revolving loans, the clients who failed to pay their loans opted for cash loans.
2. Form the second plot ‘CODE_GENDER’ with respect to TARGET, we observe that there is slight high number of female clients who turned out to outliers in comparison to Male clients.
3. From the third plot, ‘Flag_OWN_REALITY’ with respect to TARGET Variable, we observe that the clients who own a flat or house turns out to be a defaulter in comparison to the clients who doesn’t.
4. From the third plot, ‘NAME_FAMILY_STATUS’ with respect to TARGET Variable. We observe that the married clients are the column ones who have paid their loans on time. However, for the clients who has difficulty in paying their loans are also married.

# Box Plot & Histogram for Continous Variable



# Bar Charts For Categorical Variables



# CREDIT EDA CASE STUDY

## **9.Bivariate/Multivariate Analysis:**

9.1 Bivariate/Multivariate Analysis finds out the relationship between two or more variables.

9.2 We can perform Bivariate/Multivariate analysis for any combination of categorical and continuous variables. The combination can be: Categorical & Categorical, Categorical & Continuous and Continuous & Continuous.

## **10.Correlation :**

10.1 Correlation is a statistical measure.

10.2 Correlation explains how one or more variables are related to each other.

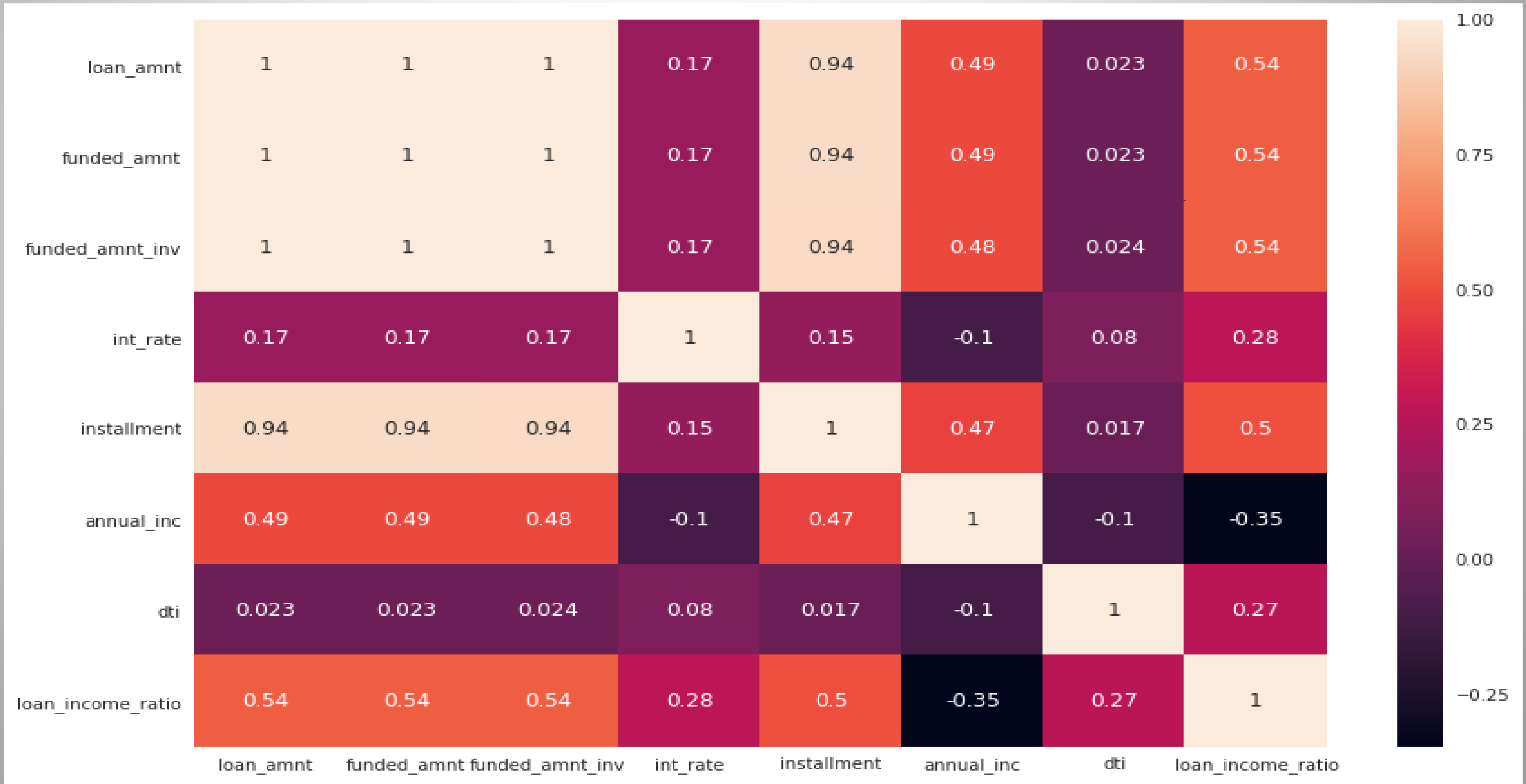
10.3 These variables can be input data features which have been used to forecast our target variable.

# CORRELATION DATA FOR TARGET 0 & 1

1. The maximum correlated variable for the appl_d_zero (i.e for the clients who are defaulters) are the OBS_60_CNT_SOCIAL_CIRCLE and OBS_30_CNT_SOCIAL_CIRCLE with the correlation coefficient of 1.00
2. We can say that both the variable are linearly related to each other. When one variable moves higher or lower, the other variable moves in the same direction with the same magnitude.
3. Similary, the maximum correlated variable for the appl_d_one(i.e for the clients who are defaulters is also OBS_60_CNT_SOCIAL_CIRCLE and OBS_30_CNT_SOCIAL_CIRCLE with the correlation coefficient of 1.00
4. Further, on observation we see that the correlation coefficient of AMT_GOODS_PRICE with AMT_CREDIT, CNT_FAM_MEMBERS and CNT_CHILDREN, DEF_60_CNT_SOCIAL_CIRCLE and DEF_30_CNT_SOCIAL_CIRCLE and DEF_30_CNT_SOCIAL_CIRCLE are almost similar for both the TARGETS 0 and 1.
5. For the defaulters clients the correlation coefficient of AMT_GOODS_PRICE with AMT_CREDIT and CNT_FAM_MEMBERS and CNT_CHILDREN are quiet high. So, while taking decisions we may look into these factors.

# CREDIT EDA CASE STUDY

**11.0 Heat Map**: are used for how 'loan’, and other kind of thing which are closely interrelated. So we can take any one column out of them for our analysis.



| | loan_amnt | funded_amnt | funded_amnt_inv | int_rate | installment | annual_inc | dti | loan_income_ratio |
|---|---|---|---|---|---|---|---|---|
| loan_amnt | 1 | 1 | 1 | 0.17 | 0.94 | 0.49 | 0.023 | 0.54 |
| funded_amnt | 1 | 1 | 1 | 0.17 | 0.94 | 0.49 | 0.023 | 0.54 |
| funded_amnt_inv | 1 | 1 | 1 | 0.17 | 0.94 | 0.48 | 0.024 | 0.54 |
| int_rate | 0.17 | 0.17 | 0.17 | 1 | 0.15 | -0.1 | 0.08 | 0.28 |
| installment | 0.94 | 0.94 | 0.94 | 0.15 | 1 | 0.47 | 0.017 | 0.5 |
| annual_inc | 0.49 | 0.49 | 0.48 | -0.1 | 0.47 | 1 | -0.1 | -0.35 |
| dti | 0.023 | 0.023 | 0.024 | 0.08 | 0.017 | -0.1 | 1 | 0.27 |
| loan_income_ratio | 0.54 | 0.54 | 0.54 | 0.28 | 0.5 | -0.35 | 0.27 | 1 |

# CREDIT EDA CASE STUDY

## 12.0 Bivariate/Multivariate Analysis with Probability of Charge off:

12.1 categorical variables vs probability of charged off

12.2 The main motive of this eda case to find what parameters are impacting the most on loan  that is if a applicant will successfully complete the loan term or will charge off.

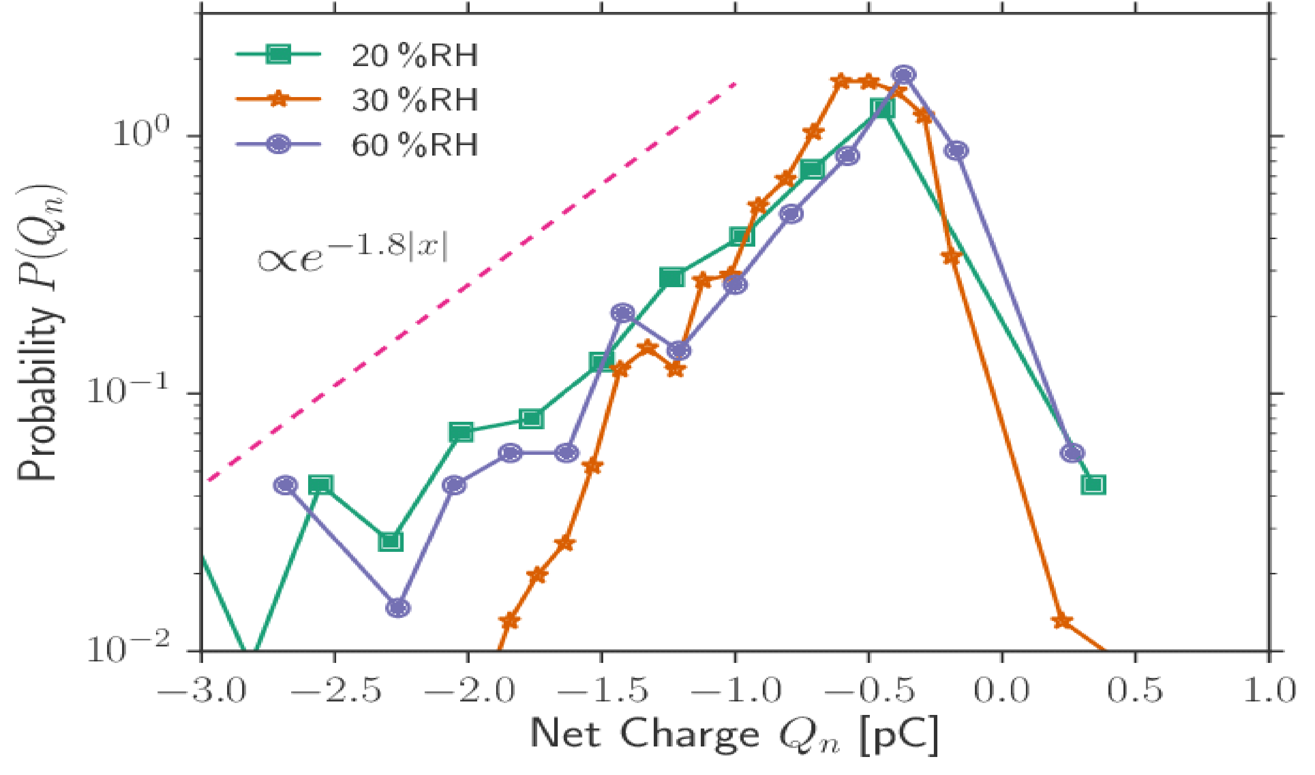
12.3 so we will be using a new term now probability of charged off that will be equal to :

12.4 probability of chargedoff =numberofapplicantswhochargedoff/totalno.ofapplicants

12.5 we will calculate this probability w.r.t each column in bivariate analysis & will see how the probability of charged off changes with these columns.

# CREDIT EDA CASE STUDY

- Probability Charges vs Loan and Grade and Annual income
- Probability charges vs Annual Income



Thank You