Term Project:

Analyzing and Minimizing Loan Investment Risks Using Random Forest Classifier

GROUP-3:
Aradhya Alva Rathnakar
Bhavan Kumar Basavaraju
Mahamaya Panda
Reddysaketh Reddy Chappidi
Shashi Kumar Kadari Mallikarjuna

TABLE OF CONTENTS:

1. Introduction 2
2. Problem Statement 2
3. Motivation 2
4. Objectives 3
5. Literature Review 3
6. Crisp-DM Methodology 5
   - 6.1. Business Understanding 5
   - 6.2. Data Understanding 5
   - 6.3. Data Preparation 6
   - 6.4. Modeling 6
   - 6.5. Data Evaluation 6
   - 6.6. Deployment 6
7. Data Flow Diagram 7
8. Data Collection 8
9. Data Pre-Processing 8
   - 9.1. Exploratory Data Analysis 8
   - 9.2. Data Cleaning 20
   - 9.3. Data Transformation 22
   - 9.4. Data Reduction 23
   - 9.5. Removing Outliers 23
10. Modeling 24
    - 10.1. Train and Test Data Split 24
    - 10.2. Data Normalization 25
    - 10.3. Model Building 26
11. Data Visualization 30
12. Deployment 34
13. Discussion 34
14. Impact 35
15. Conclusion 35
16. Future Scope 36
17. References 37

ABSTRACT

The lending industry is constantly faced with the challenge of identifying potential loan defaulters to minimize financial risks. The dataset from LendingClub loan defaulters prediction provides valuable information for analyzing and predicting loan default risks. In this report, we explore the potential of using this dataset to analyze and develop a predictive model that can help lenders make informed decisions and minimize the risks associated with loan investments. By leveraging the power of the Random Forest Classifier, we aim to identify key factors and patterns that contribute to loan defaults and build a robust model that can accurately classify potential defaulters. Through analysis, we aim to provide insights and recommendations to improve risk assessment in loan investments and enhance the overall lending process.

1. INTRODUCTION :

Banking and finance are the backbones of the world economy. The size of a particular country's banking sector shows how big that country's GDP might be.
Most legal transactions happen through the banking system, and there is a need to ensure the security of funds in financial institutions to build trust.
Banks make the majority of their money by lending out the investors' money to the people who need it for buying a house, car, personal reasons, etc., for a specific interest rate. These banks will face a loss if the borrowers don't return the borrowed money.
In this project, we look at different factors that could be considered before giving out a loan to ensure that the loan is only sanctioned to people who can repay it.

2. PROBLEM STATEMENT :

Online financial services that emerged with the growth of blockchain technology and the crypto market have disrupted traditional banking systems by offering lower interest rates due to reduced infrastructure costs. However, to maintain viability, such services require a thorough vetting process to reduce the number of defaulters. To address this problem, predictive analytics, specifically the use of a Random Forest Classifier, can be employed to analyze various factors such as credit history, income, employment status, debt-to-income ratio, etc. of borrowers to detect potential defaulters. This approach will enable well-informed decision-making and reduce the risk of defaults.

3. MOTIVATION :

Lending money is a crucial aspect of the financial industry. Banks and other lending institutions have been using traditional credit scoring methods and models to assess the risk associated with lending money. Traditional risk assessment methods have relied on credit scores and financial history, but these measures don't always provide a complete picture of the borrower's ability to repay the loan. In the last 4 months, many banks including First Republic Bank went under and were acquired by JPMorgan Chase Bank. It has become all the more important in the recent times to keep the operating costs and losses low for any financial institution to survive in the market.

The "Lending Club dataset" provides a rich source of information that can be used to visualize the performance of the financial institution and also to build predictive models that can help lenders better assess risk and make more informed lending decisions. By leveraging machine

learning techniques like Random Forest Classifier, lenders can more accurately predict the likelihood of defaulters and take steps to minimize their risks, leading to more sustainable lending practices and better outcomes for both lenders and borrowers.

4. OBJECTIVES :

For our project, we aim to analyze the performance of LendingClub company by using various data exploration and cleansing techniques to preprocess the data, use Random Forest Classifier to predict possible loan defaulters, and visualize the company data using visualization tools like PowerBI. Below are the key objectives:

- **Python and PowerBI:**

  Visualize historical data to identify patterns and trends in the loan data.

- **PowerBI:**

  Provide insights into borrower creditworthiness and identify areas for improvement to reduce risk and improve profitability for investors.

- **Google Forms and PowerBI:**

  Visualize the interest rates that the current SJSU students are paying using the live survey to analyze the market competition that LendingClub has in terms of interest rates offered for different loan categories

- **Python libraries such as pandas, and sklearn:**

  Cleanse the data and develop a machine learning model to predict loan defaulters and improve loan acceptance/interest rates.

5. LITERATURE REVIEW :

[1] "Credit risk assessment: a challenge for financial institutions" by Evangelos Kalapodas; Mary E. Thomsom :

The focus of this research is to examine the assessment of financial credit risk, which is a significant concern due to the absence of a standardized approach employed by financial institutions. Through a thorough evaluation of the commonly used credit risk assessment methods, credit scoring, and portfolio models, certain limitations are identified when these methods are used individually. Interviews with industry experts support the notion that combining multiple credit risk assessment methods is crucial for achieving effective results. As a result, this study proposes a framework that enhances credit risk assessment by leveraging the strengths of these methods while effectively addressing their limitations.

[2] "Data Visualization and its Key Fundamentals: A Comprehensive Survey" by Muscan, Gurpreet Singh, Jaspreet Singh, and Chander Prabha :

The author emphasizes the critical role of data visualization in today's data-driven corporate environment. It mentions that data visualization is widely used to aid decision-making processes and is closely tied to the key revenues of many industrial businesses. It acknowledges the increasing need for effective and accurate data processing, particularly due to the volume, velocity, and validity of data in

contemporary datasets. The study aims to address this need by examining different phases of data visualization, its various types, and its applications in areas such as scientific research and prediction. Hence, it suggests that data visualization plays a crucial role in understanding and extracting insights from large volumes of data, contributing to scientific research and facilitating prediction in various domains.

[3] “The Application Study of Credit Risk Model In Financial Institution via Machine-learning Algorithms" by Yuanzhang Wang; Jiongcheng Lu; Jiehan Qin; Chenyi Zhang; Yiyang Chen :

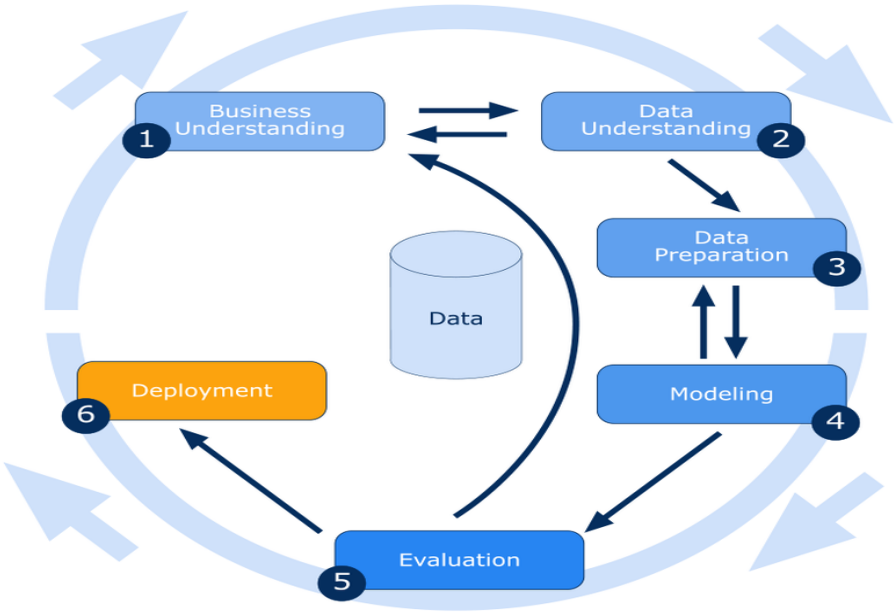
Credit scoring technology, a type of statistical model, is widely employed in assessing the risk associated with loan applicants. It utilizes borrower-provided information, historical data, and banking system data to predict credit risk. This research focuses on credit risk analysis and evaluation using machine learning algorithms, including Random Forest. The study conducts data preprocessing and creates a credit scorecard. The base score of the credit scorecard is 750 points, and for every decrease of 60 points, the probability of default doubles. Experimental results demonstrate the viability of machine learning algorithms in credit evaluation. The evaluation parameters such as precision, and F1 score are examined within the study's training and testing samples, indicating that Random Forest can be applied to financial risk analysis.

[4] “Comparative Breast Cancer Detection with Artificial Neural Networks and Machine Learning Methods” by Muhammed Coşkun Irmak; Mehmet Bilge Han Taş; Sedat Turan; Abdulsamet Haşiloğlu :

The prevalence of cancer worldwide is on the rise due to factors such as environmental pollution, excessive technological and biological waste, climate change, and increased consumption of processed foods. The objective of this study is to achieve the fastest and most accurate cancer detection. The researchers compare artificial neural networks and traditional machine learning methods to obtain high-performance results. The study utilizes various machine learning algorithms, including k-Nearest Neighbors (kNN), Random Forest (RF), Xgboost (XGB), and Artificial Neural Network (ANN). Hence, the study aims to contribute to cancer detection by exploring the effectiveness of different machine-learning techniques, highlighting the potential of artificial neural networks for achieving accurate and efficient detection rates.

6. CRISP-DM METHODOLOGY :

For the development of this project, the CRISP-DM methodology was used. This helped in planning the different phases of the project effectively.



6.1. BUSINESS UNDERSTANDING :

In this phase, the problem statement was discussed to get a deeper understanding of the issue. The research was done by the team to get an idea of how financial institutions work and the problems related to it in recent years with the advancement in technology. Different online resources and research papers were referred to to come up with a solution to the problem which involved analysis and prediction to reduce the risks involved in the banking sector focusing on loss due to loan defaulters.

6.2. DATA UNDERSTANDING :

In this phase, the different data sources were explored and we chose the LendingClub dataset from Kaggle which had almost 8 years of data about the loans provided by the company. A Live Google survey was also done to get information from SJSU students to understand the interest rates they are paying on the different loans they might have to explore the competition that LendingClub has. Exploratory data analysis was performed on the LendingClub data as well as on the live Google survey data to understand the data we are dealing with.

6.3. DATA PREPARATION:

In this phase, the source data had to be cleaned. The live survey had a few questions for which options and checkboxes were provided but there were a few other columns like Loan type, annual income, email, etc which were short answers. These short answer column values had to be cleansed where there were some null values. There was also an email column for which the email format was checked and cleansed to meet the email standards. The data for LendingClub from Kaggle also had a lot of null values which had to be handled. There was an address column that had to be split to get the zip code and state to perform analysis. The data columns also had to be cleaned to have them in the right format. Once this pre-processing was done, the data had to be split into train and test to train the Random Forest Classifier and then find the accuracy of that model using the test dataset.

6.4. MODELING :

To solve our problem to predict the possible loan defaulters, Random Forest Classifier was used. The model had to be trained using the training dataset. Along with modeling the Random Forest Classifier, visualizations were also created using PowerBI to see how different aspects of the business have been over some time. The live survey data was also visualized to compare how LendingClub is performing when compared to its competitors in the market in terms of interest rates.

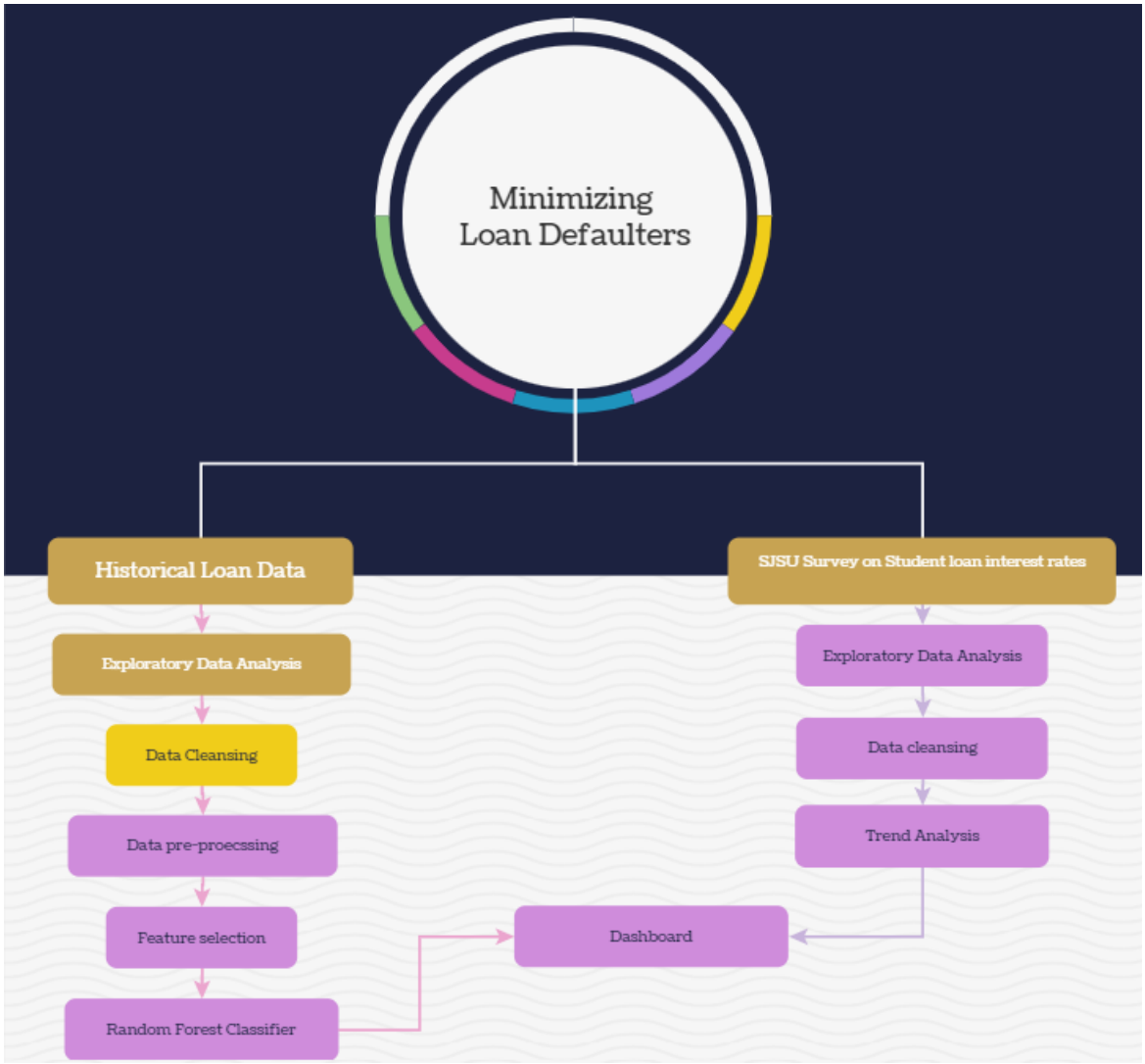
6.5. DATA EVALUATION :

The Random Forest Classifier is run against the test dataset to check the accuracy of the model. The prediction accuracy has to be within the acceptable range where if it is too good then the model is an over-fit and if the accuracy is too less, then it is an under-fit. The model accuracy is visualized in Python. The rest of the data is also visualized in Power BI to see the LendingClub's performance as compared to its competitors and analyze the sustainability of its business model.

6.6. DEPLOYMENT :

The visualizations are created on Power BI Desktop and then deployed onto the powerBI workspace. The Power BI deployment pipeline was used to perform the initial development and then was pushed to the production workspace once the development was done and tested. The dashboard could be shared with multiple people in our university by providing them access to the report.

7. DATA FLOW DIAGRAM :



The primary objective of the project was to analyze the performance of LendingClub company and minimize the loan defaulters to reduce the loss that could be incurred by the company that could bring down the collective interest rates of everyone who took a loan from the company. SJSU's live Google survey was performed to get the interest rates to understand the competition that LendingClub is facing. Exploratory data analysis is performed on this data and data cleansing is performed. This cleansed data is visualized using powerBI to perform trend analysis the visualizations are added to the dashboard.
The historical loan data for LendingClub was taken from Kaggle and exploratory data analysis was performed to understand the data we are dealing with. Data cleansing and pre-processing were done to prep the data to train and test the Random Forest Classifier and also analyze the performance of the company using PowerBI.

8. DATA COLLECTION :

Name	Data Collection Method	Description
Minimizing risks for Loan investment	Kaggle	Analyze the loan data of LendingClub company, and predict the possible defaulters using Random Forest Classifier helping to identify high-risk and low-risk loan groups
SJSU live survey dataset	Google form	Live survey filled out by SJSU students used for trend analysis

9. DATA PRE-PROCESSING :

Data pre-processing is an important step in understanding the kind of data we are dealing with. The data has to be cleaned and the right features need to be selected to train the machine learning model to make the right predictions. Exploratory data analysis was performed on the data which led to data cleansing to get the data ready to be used for training the machine learning model. Two data sources were considered: the Lending Club dataset: https://www.kaggle.com/code/faressayah/lending-club-loan-defaulters-prediction/input and SJSU Google Survey dataset

9.1 . EXPLORATORY DATA ANALYSIS :

LENDING CLUB DATA

The exploration of data is the key to solving any complex problem. Initial data exploration was done using various visuals to get a good understanding of the data. Python libraries like pandas, matplotlib, seaborn, etc were used for exploratory data analysis. Once the data from Kaggle was loaded into a data frame using the Pandas library, we went on to look at the kind of data we were dealing with by displaying the top 5 rows in the dataset.

```
#to display all columns
pd.set_option('display.max_columns', None)

#explore the kind of data present in the dataframe
lendingdf.head()
```

	loan_amnt	term	int_rate	installment	grade	sub_grade	emp_title	emp_length	home_ownership	annual_inc	verification_status	issue_d	loan_status	purpose	titl
0	10000.0	36 months	11.44	329.48	B	B4	Marketing	10+ years	RENT	117000.0	Not Verified	Jan-2015	Fully Paid	vacation	Vacatio
1	8000.0	36 months	11.99	265.68	B	B5	Credit analyst	4 years	MORTGAGE	65000.0	Not Verified	Jan-2015	Fully Paid	debt_consolidation	Det consolidatio
2	15600.0	36 months	10.49	506.97	B	B3	Statistician	< 1 year	RENT	43057.0	Source Verified	Jan-2015	Fully Paid	credit_card	Credit car refinancin
3	7200.0	36 months	6.49	220.65	A	A2	Client Advocate	6 years	RENT	54000.0	Not Verified	Nov-2014	Fully Paid	credit_card	Credit car refinancin
4	24375.0	60 months	17.27	609.33	C	C5	Destiny Management Inc.	9 years	MORTGAGE	55000.0	Verified	Apr-2013	Charged Off	credit_card	Credit Car Refinanc

The dataset contained different kinds of data types at first glance, and it consists of loan-related data like loans taken to date, interest rate, loan amount, customer details, etc.
We further went on to look at the different numbers of non-null records in each column and the data types of each column.

```
#explore the dataframe to see how many not null values are there in each column
lendingdf.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 396030 entries, 0 to 396029
Data columns (total 27 columns):
 #   Column                Non-Null Count   Dtype  
---  ------                --------------   -----  
 0   loan_amnt             396030 non-null  float64
 1   term                  396030 non-null  object 
 2   int_rate              396030 non-null  float64
 3   installment           396030 non-null  float64
 4   grade                 396030 non-null  object 
 5   sub_grade             396030 non-null  object 
 6   emp_title             373103 non-null  object 
 7   emp_length            377729 non-null  object 
 8   home_ownership        396030 non-null  object 
 9   annual_inc            396030 non-null  float64
 10  verification_status   396030 non-null  object 
 11  issue_d               396030 non-null  object 
 12  loan_status           396030 non-null  object 
 13  purpose               396030 non-null  object 
 14  title                 394275 non-null  object 
 15  dti                   396030 non-null  float64
 16  earliest_cr_line      396030 non-null  object 
 17  open_acc              396030 non-null  float64
 18  pub_rec               396030 non-null  float64
 19  revol_bal             396030 non-null  float64
 20  revol_util            395754 non-null  float64
 21  total_acc             396030 non-null  float64
 22  initial_list_status   396030 non-null  object 
 23  application_type      396030 non-null  object 
 24  mort_acc              358235 non-null  float64
 25  pub_rec_bankruptcies  395495 non-null  float64
```

There were a few null values in a few columns which had to be handled in the data cleansing process.

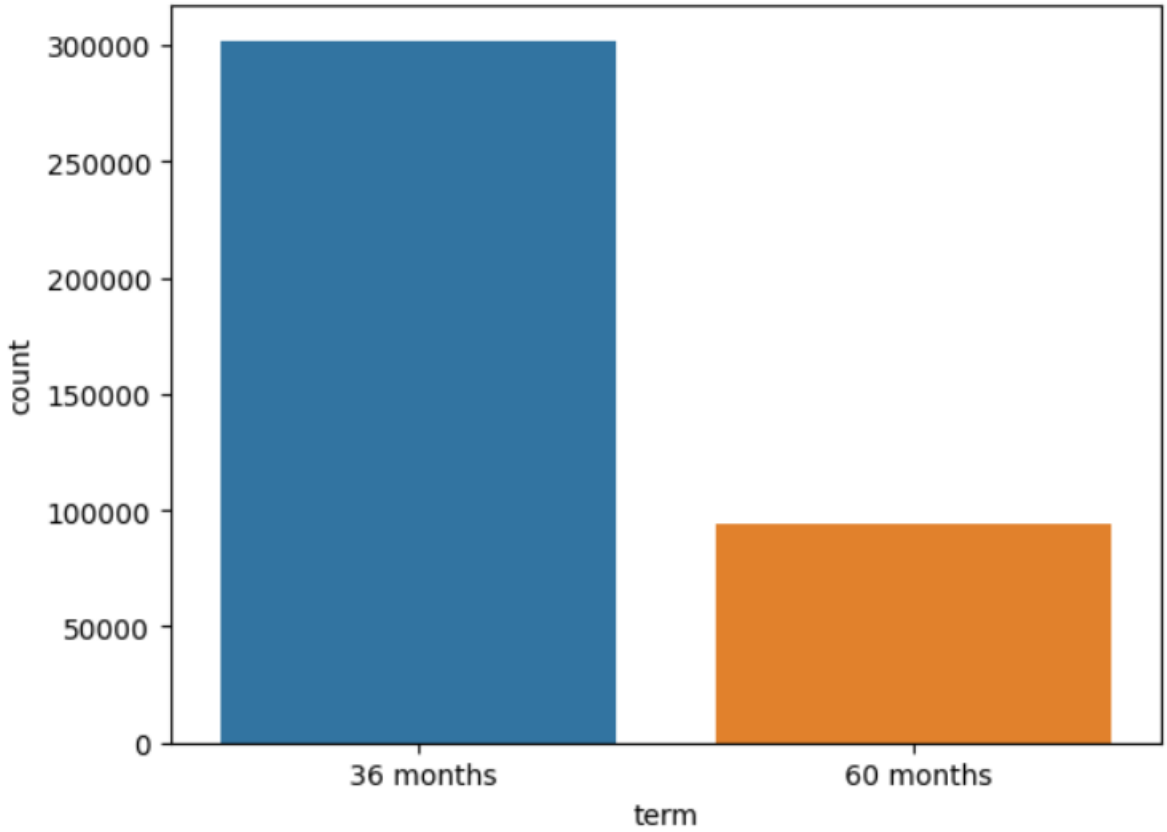
To get a better understanding of the numerical data, we used the ‘describe’ function.

```
[ ] #check the distribution of the numerical values in the dataframe
    lendingdf.describe()
```

	loan_amnt	int_rate	installment	annual_inc	dti	open_acc	pub_rec	revol_bal	revol_util	total_acc	mort_acc	pub_rec_bankruptcie
count	396030.000000	396030.000000	396030.000000	3.960300e+05	396030.000000	396030.000000	396030.000000	3.960300e+05	395754.000000	396030.000000	358235.000000	395495.0000
mean	14113.888089	13.639400	431.849698	7.420318e+04	17.379514	11.311153	0.178191	1.584454e+04	53.791749	25.414744	1.813991	0.1216
std	8357.441341	4.472157	250.727790	6.163762e+04	18.019092	5.137649	0.530671	2.059184e+04	24.452193	11.886991	2.147930	0.3561
min	500.000000	5.320000	16.080000	0.000000e+00	0.000000	0.000000	0.000000	0.000000e+00	0.000000	2.000000	0.000000	0.0000
25%	8000.000000	10.490000	250.330000	4.500000e+04	11.280000	8.000000	0.000000	6.025000e+03	35.800000	17.000000	0.000000	0.0000
50%	12000.000000	13.330000	375.430000	6.400000e+04	16.910000	10.000000	0.000000	1.118100e+04	54.800000	24.000000	1.000000	0.0000
75%	20000.000000	16.490000	567.300000	9.000000e+04	22.980000	14.000000	0.000000	1.962000e+04	72.900000	32.000000	3.000000	0.0000
max	40000.000000	30.990000	1533.810000	8.706582e+06	9999.000000	90.000000	86.000000	1.743266e+06	892.300000	151.000000	34.000000	8.0000

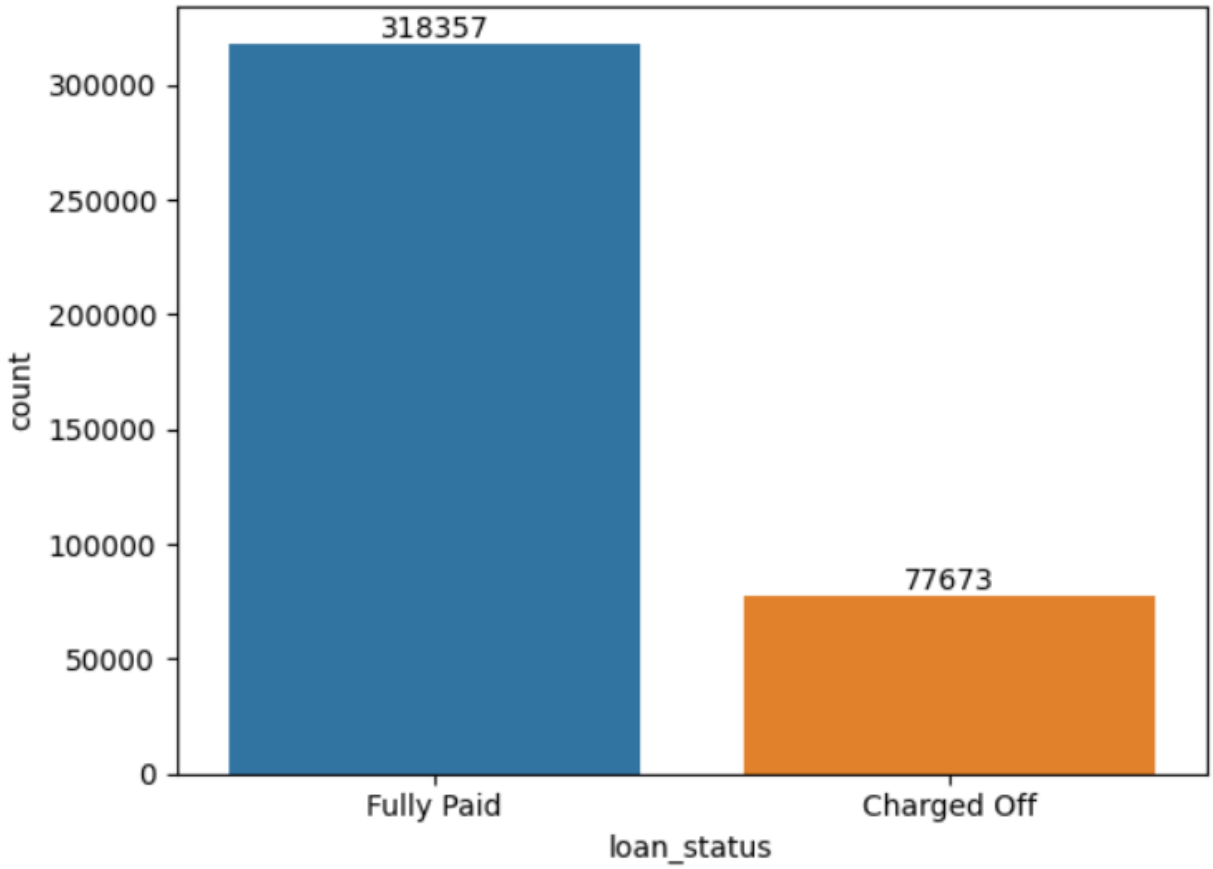
We better understood the data by performing this initial source data exploration.

To understand the number of loans that were approved for each loan repayment term to see how long the people usually prefer to repay the loan, we used a bar plot.

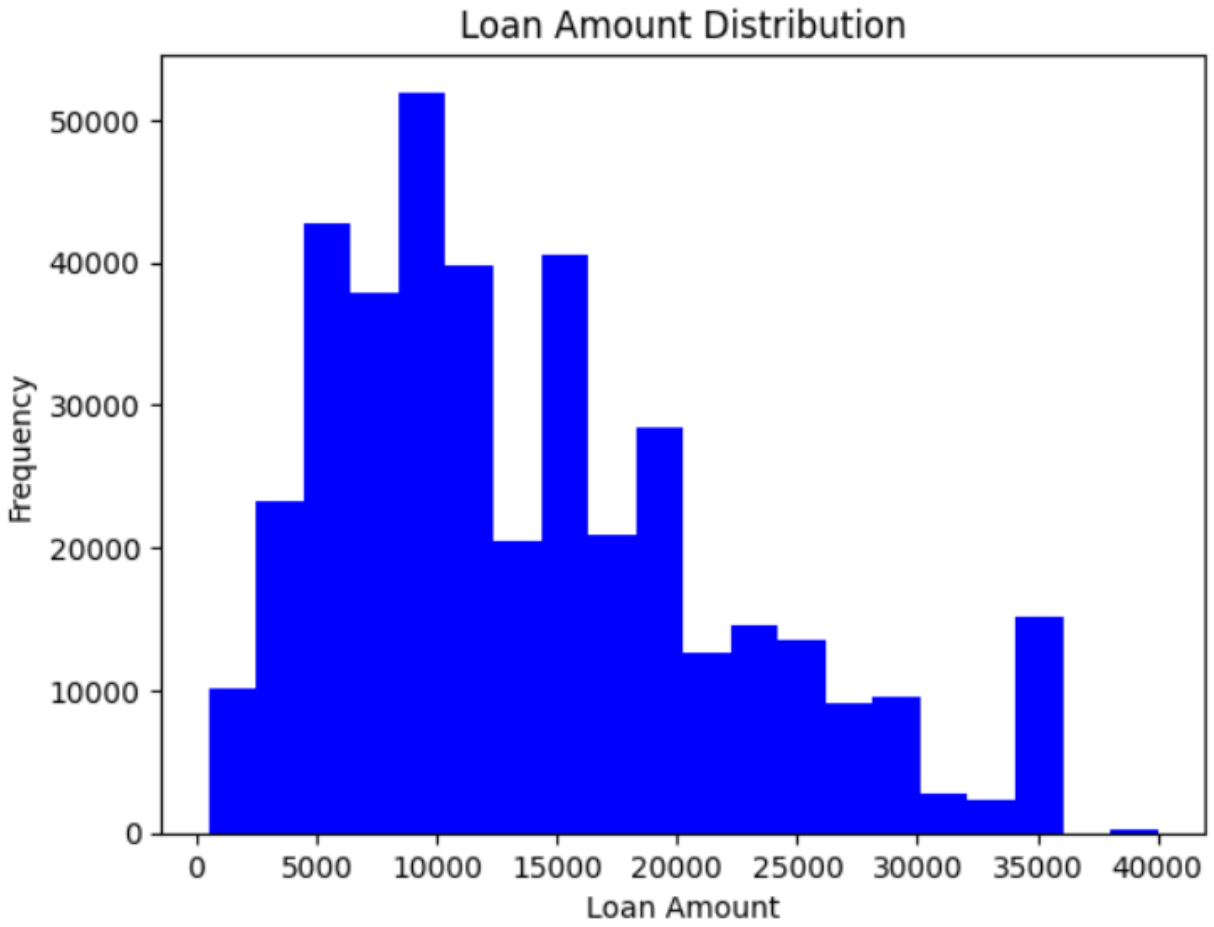


From the above plot, we can see that people prefer 36 months loan repayment terms with almost 300,000 people taking a loan for a repayment period of 36 months as compared to a little less than 100,000 people who took a loan for a loan repayment period of 60 months.

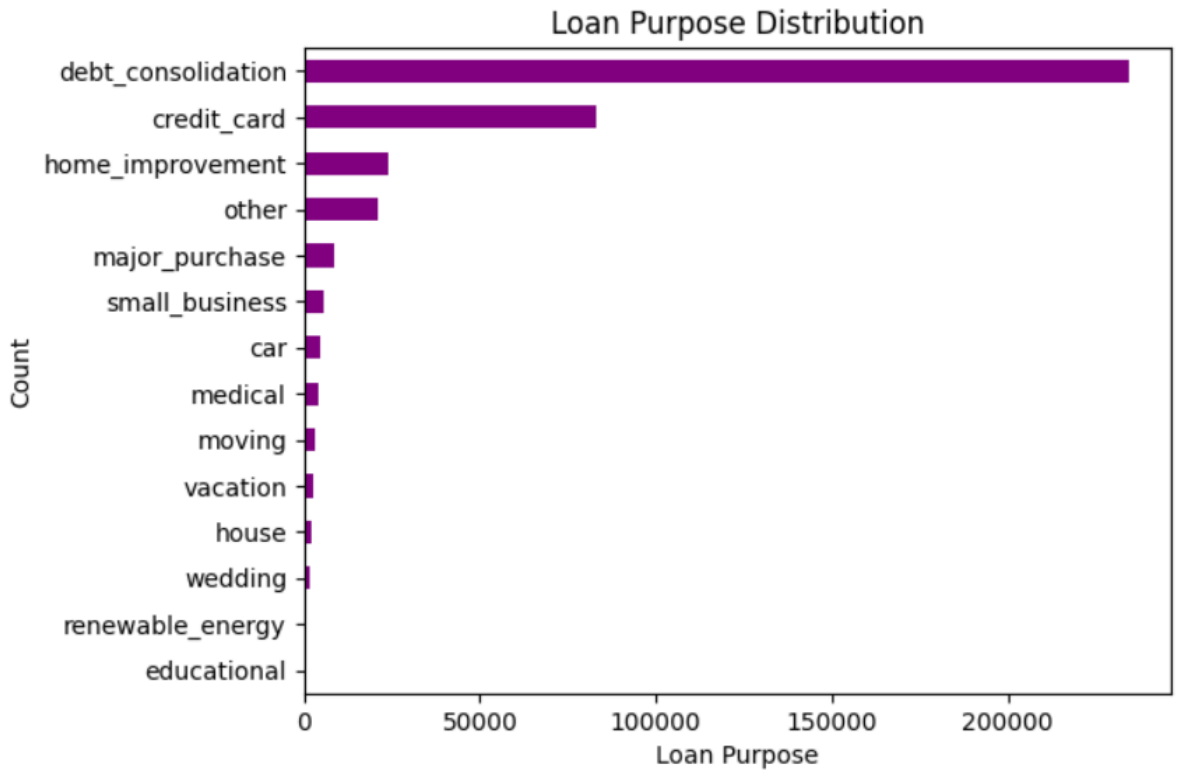
We also wanted to know the number of people who repaid the loan on time and the number of people who charged off/ did not pay the loan to understand the impact it might have on a financial institution.



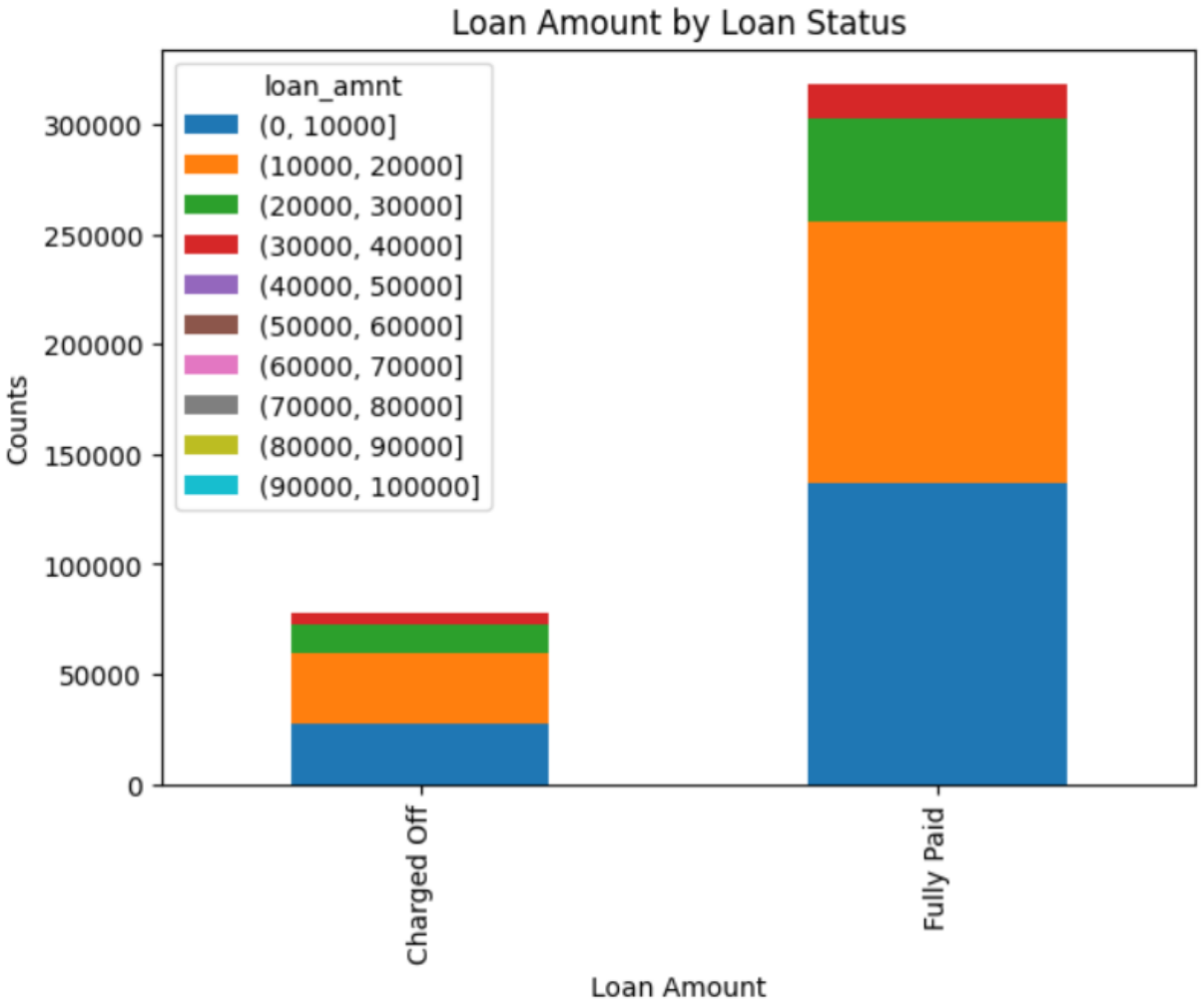
From the bar plot, we can see that almost 20 % of the people who have taken a loan did not repay the loan and this has a great impact on the interest rates the financial institution is charging.



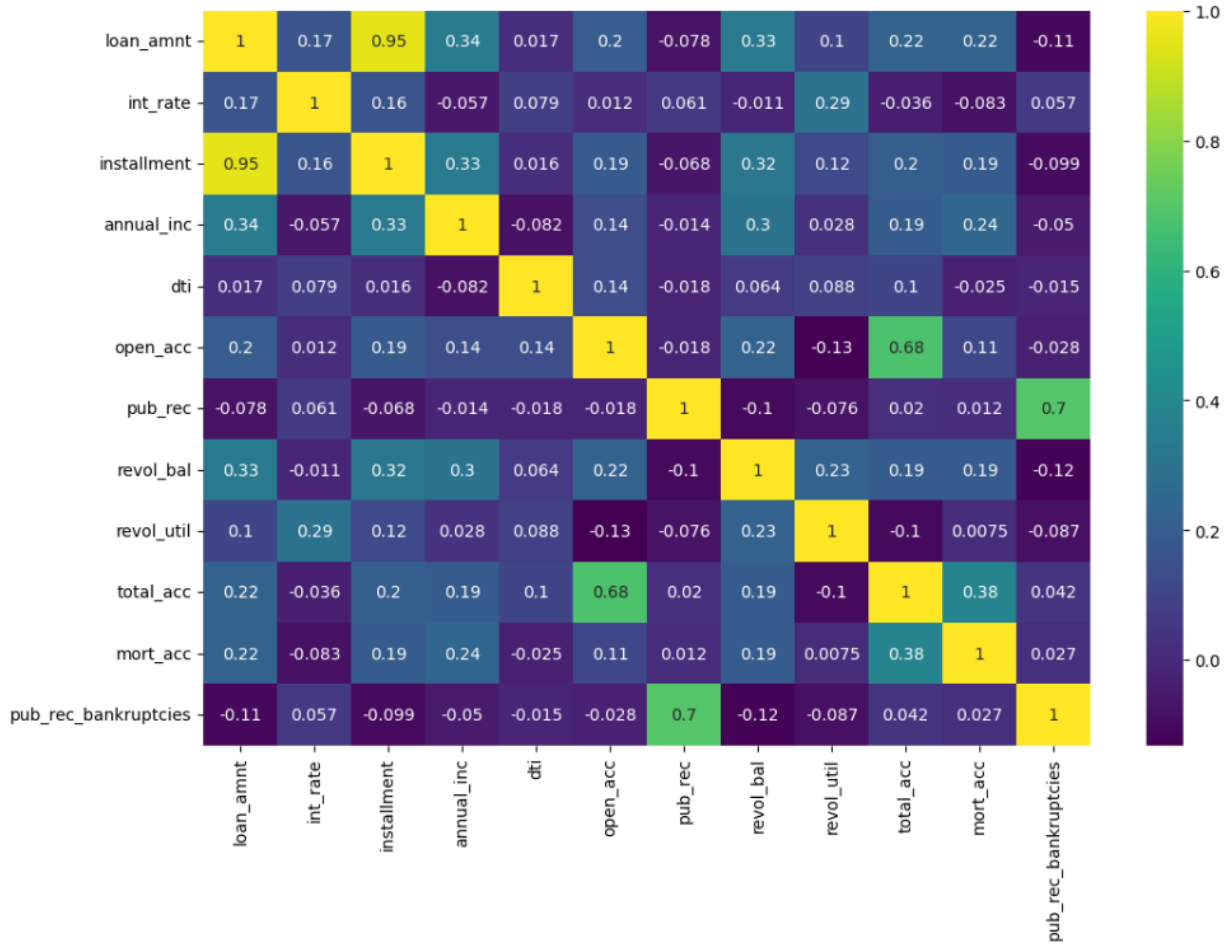
From the above histogram, we can see that the majority of loans are between 0 and 20,000 dollars, with the highest frequency occurring at around 10,000 dollars. The distribution is right-skewed, indicating that there are a few loans with very high amounts that are skewing the distribution toward the right.



From this horizontal bar chart, we can say that debt consolidation is the most common loan purpose, followed by credit card and home improvement. The range of counts varies widely between loan purposes, with the highest count being for debt consolidation (over ~200,000), and the lowest count being for educational purposes (less than ~1,500).

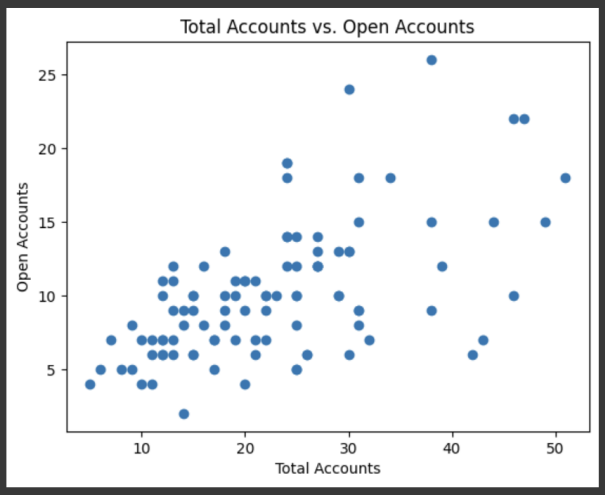


The above plot is grouping the loans by their status and bins them based on their loan amount into different categories, ranging from 0-10,000 to 90,000-100,000. The resulting plot is a stacked bar chart that shows the count of loans in each loan amount category, separated by their loan status. We can see that the majority of loans fall in the 0-20,000 category, and most of those loans are fully paid. There are fewer loans in the higher loan amount categories and the proportion of charged-off loans increases as the loan amount increases.



Based on the heatmap, we can see that there is a strong positive correlation between the loan amount and installment, indicating that borrowers who take out larger loans also have higher monthly installment payments. There is a moderate positive correlation between the loan amount and annual income, indicating that borrowers with higher incomes tend to take out larger loans.

```
total_accounts = lendingdf['total_acc']
open_accounts = lendingdf['open_acc']

sampled_df = lendingdf.sample(n=100)

sampled_total_accounts = sampled_df['total_acc']
sampled_open_accounts = sampled_df['open_acc']

plt.scatter(sampled_total_accounts, sampled_open_accounts)

plt.xlabel('Total Accounts')
plt.ylabel('Open Accounts')
plt.title('Total Accounts vs. Open Accounts')

# Display the plot
plt.show()
```

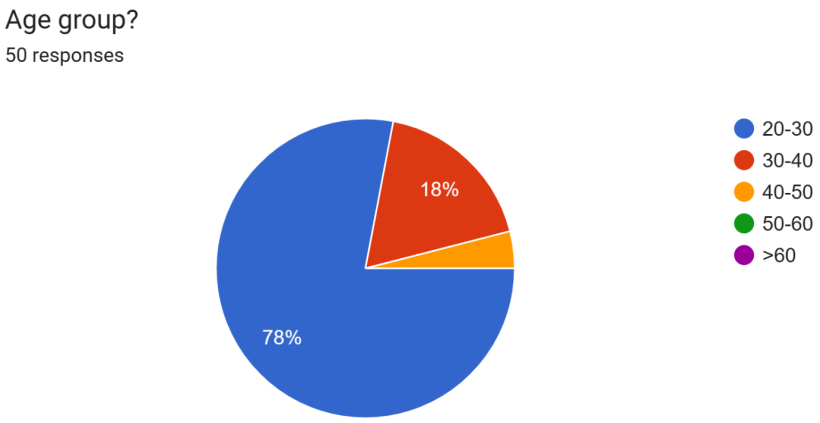


The above code allows us to visually analyze the relationship between the total number of accounts and the number of open accounts using a scatterplot. The code also incorporates random sampling to select a smaller subset of data points, which can be useful for better visual clarity or when working with large datasets.

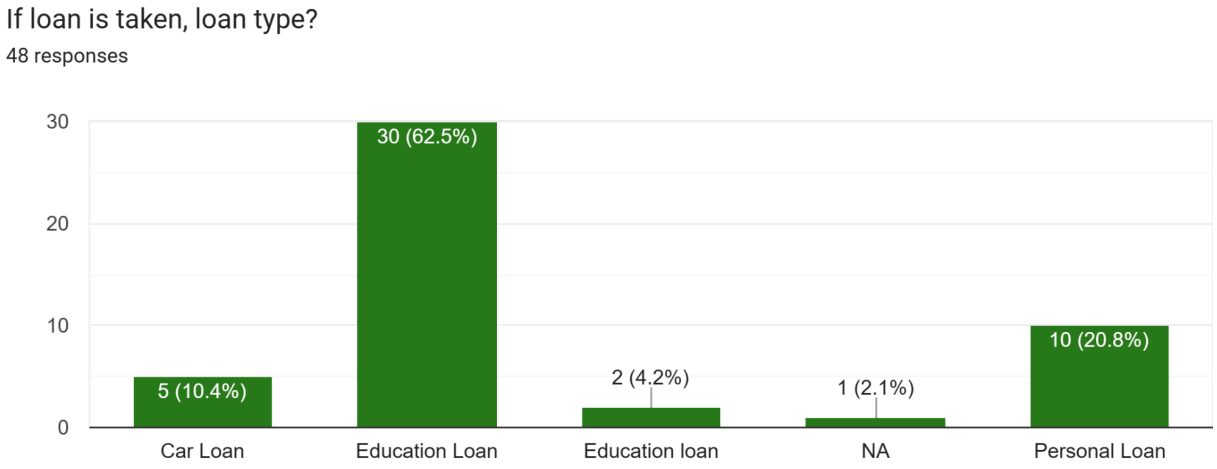
Hence, it provides a simple but useful way to analyze and visualize the relationship between "total_acc" and "open_acc" variables. It can help in gaining insights, identifying patterns, and making decisions.

SJSU GOOGLE SURVEY :

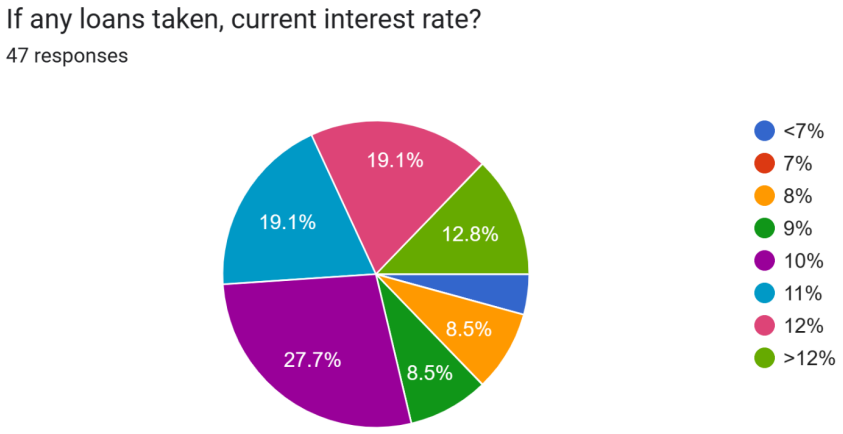
Based on the Live survey, Google Forms provides summary responses to different questions which helps in understanding the data.



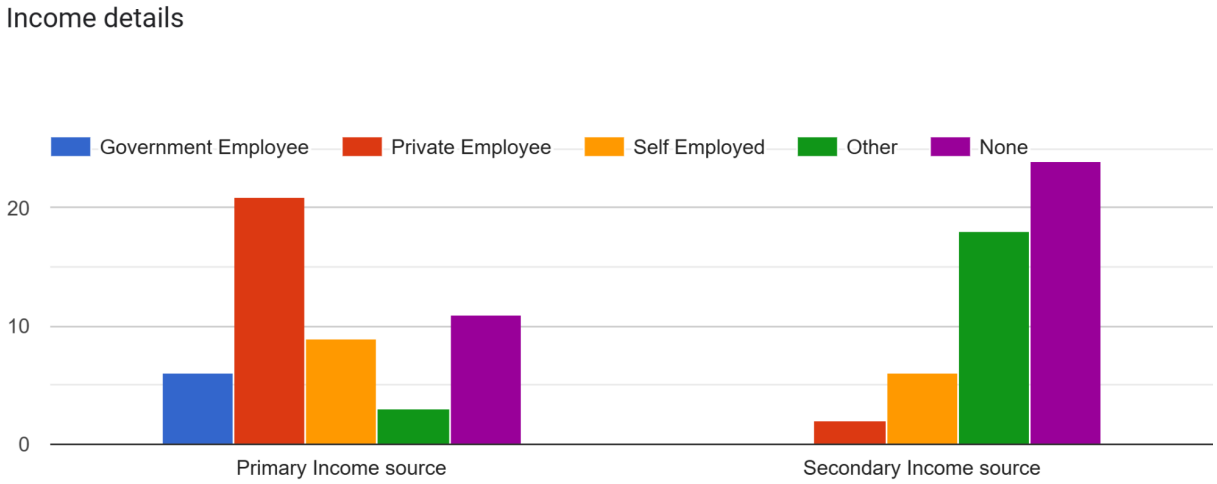
From this pie chart, we can see that the majority of the respondents belong to the 20-30 years age group at 78% followed by the 30-40 age group at 18%. The other 4% belong to other age groups.



From the above bar chart, we can see that the majority of the respondents took loans for educational purposes with almost 67% followed by loans for personal reasons with around 20%.



The above pie chart shows the different interest rates that the respondents are paying currently. From the chart, we can see that almost 28% of the respondents are paying 10% APR. An equal number of people are paying 11% APR and 12% APR with around 19% of respondents respectively. Only 4% of the respondents are paying less than 7% APR



Exploring the income of the respondents who have taken loans, we can see that the primary income source for the majority of the respondents is by working at a private organization and the majority of the respondents do not have any secondary source of income.

9.2. DATA CLEANING :

LENDING CLUB DATA

This step involves checking the missing values or inconsistent values and removing them if the entire row is missing. It then replaces the null values in the numerical column with the average value of the entire column.

```
[16] # Count of Missing values
     for column in lendingdf.columns:
         if lendingdf[column].isna().sum() != 0:
             missing = lendingdf[column].isna().sum()
             portion = (missing / lendingdf.shape[0]) * 100
             print(f"'{column}': number of missing values '{missing}' ==> '{portion:.3f}%'")

     'emp_title': number of missing values '22927' ==> '5.789%'
     'emp_length': number of missing values '18301' ==> '4.621%'
     'title': number of missing values '1755' ==> '0.443%'
     'revol_util': number of missing values '276' ==> '0.070%'
     'mort_acc': number of missing values '37795' ==> '9.543%'
     'pub_rec_bankruptcies': number of missing values '535' ==> '0.135%'

[17] # Drop row only if the entire row has no value
     lendingdf.dropna(how='all', inplace=True)
```

```
for col in lendingdf.columns:
    if lendingdf[col].dtype != 'object':
        lendingdf[col].fillna(lendingdf[col].mean(), inplace=True)

# Print the number of null values after replacement
lendingdf_after_replace = lendingdf.isnull().sum()
print("\nNull values after replacement:\n", lendingdf_after_replace)

Null values after replacement:
 loan_amnt                   0
term                         0
int_rate                     0
installment                  0
grade                        0
sub_grade                    0
emp_title                22927
emp_length               18301
home_ownership               0
annual_inc                   0
verification_status          0
issue_d                      0
loan_status                  0
purpose                      0
title                     1755
dti                          0
earliest_cr_line             0
open_acc                     0
pub_rec                      0
revol_bal                    0
revol_util                   0
total_acc                    0
initial_list_status          0
application_type             0
mort_acc                     0
pub_rec_bankruptcies         0
address                      0
dtype: int64
```

SJSU GOOGLE SURVEY :

```
[48] # Count of Missing values
     for column in surveydf.columns:
         if surveydf[column].isna().sum() != 0:
             miss = surveydf[column].isna().sum()
             port = (miss / surveydf.shape[0]) * 100
             print(f"'{column}': number of missing values '{miss}' ==> '{port:.3f}%'")
```

```
'Gender': number of missing values '3' ==> '6.000%'
'If loan is taken, loan type?': number of missing values '3' ==> '6.000%'
'If any loans taken, current interest rate?': number of missing values '3' ==> '6.000%'
'Income details (per annum)': number of missing values '1' ==> '2.000%'
```

```
[49] # Drop row only if the entire row has no value
     lendingdf.dropna(how='all', inplace=True)
```

```
[50] import re

     # regular expression pattern to match email format
     email_pattern = r'\b[A-Za-z0-9._%+-]+@[A-Za-z0-9.-]+\.[A-Z|a-z]{2,}\b'

     # replace invalid formats with null in the Email column of the DataFrame
     surveydf['Email'] = surveydf['Email'].apply(lambda x: x.strip() if isinstance(x, str) else None)
     surveydf['Email'] = surveydf['Email'].apply(lambda x: x if re.match(email_pattern, str(x)) else None)

     # print the updated DataFrame
     surveydf
```

	Timestamp	Name	Email	Age group?	Gender	Place of stay?	Any current loans?	If loan is taken, loan type?	If any loans taken, current interest rate?	Loan source?	Income details [Primary Income source]	Income details [Secondary Income source]	Income details (per annum)
0	08/05/2023 23:16:04	Mahamaya Panda	mahamaya.panda@sjsu.edu	20-30	NaN	San Jose	No	NaN	NaN	Commericial	None	Other	NaN
1	08/05/2023 23:17:43	Aradhya Alva Rathnakar	None	20-30	NaN	1234	Yes	Education loan	10%	Credit Union	Government Employee	None	2 billion >?
2	08/05/2023 23:26:05	Saketh	reddysakethreddy.chappidi@sjsu.edu	20-30	NaN	Sanjose	Yes	Education loan	10%	Credit Union	Government Employee	Self Employed	50000
3	09/05/2023 22:52:41	Shashi Kumar	shashikumar.kadarimallikarjuna@sjsu.edu	20-30	Male	United States of America	Yes	Car Loan	<7%	Credit Union	Private Employee	Other	50000
4	09/05/2023 22:53:55	Milan	milan123@gmail.com	20-30	Male	United States of America	Yes	Education Loan	10%	Commericial	Other	None	20000
5	09/05/2023 22:54:57	Jack	jackrocket@gmail.com	30-40	Male	United States of America	Yes	Education Loan	12%	Commericial	Private Employee	None	32000
6	09/05/2023 22:55:58	Daniel Lee	Danny.Lee@sjsu.edu	20-30	Male	United States of America	Yes	Education Loan	8%	Credit Union	Private Employee	None	10000
7	09/05/2023 22:56:54	Dhiren L	dlal1222@gmail.com	20-30	Male	United States of America	Yes	Personal Loan	>12%	Commericial	Self Employed	None	60000
8	09/05/2023 22:57:54	Jessica Holmes	Jessica.holmes@sjsu.edu	20-30	Female	United States of America	Yes	Education Loan	11%	Commericial	Private Employee	None	20000
9	09/05/2023 22:58:41	Christopher Tan	chris.tan@gmail.com	20-30	Male	United States of America	Yes	Education Loan	11%	Commericial	Private Employee	None	15000
10	09/05/2023 22:59:25	Krina Marinez	Krina.Marinez@sjsu.edu	20-30	Female	United States of America	Yes	Education Loan	9%	Commericial	Self Employed	Other	8000
11	09/05/2023 23:00:10	Jack Tran	Jack.Tran@sjsu.edu	30-40	Male	United States of America	Yes	Education Loan	12%	Commericial	Private Employee	None	27000
12	09/05/2023 23:01:57	Leslie Kao	Leslie.kao@sjsu.edu	20-30	Female	United States of America	Yes	Education Loan	8%	Credit Union	Private Employee	None	40000
13	09/05/2023 23:02:54	Judah marinez	judah.mar@gmail.com	20-30	Male	United States of America	Yes	Personal Loan	>12%	Commericial	Self Employed	Other	23000
14	09/05/2023 23:03:39	Wafia Mani	wafy.mani@gmail.com	20-30	Female	United States of America	Yes	Education Loan	11%	Commericial	Private Employee	Other	35000
15	09/05/2023 23:04:18	Emauel	manny.gari@sjsu.edu	20-30	Male	United States of America	Yes	Education Loan	10%	Commericial	Private Employee	None	45000
16	09/05/2023 23:05:06	Kishan	k.manny@gmail.com	20-30	Male	United States of America	Yes	Personal Loan	12%	Commericial	Self Employed	None	32000
17	09/05/2023 23:05:55	Piyanshika	p.anshika@gmail.com	20-30	Female	United States of America	Yes	Car Loan	<7%	Credit Union	Private Employee	None	60000
18	09/05/2023 23:06:37	Anne Lee	Anne.Lee@sjsu.edu	20-30	Female	United States of America	Yes	Education Loan	11%	Commericial	Private Employee	None	22000

This step involves the cleansing of survey data in terms of missing values and improper data. Data values that are null in all the columns have been dropped. The email format of email column data is been filtered down according to the right format and the improper data is being replaced with null values.

9.3. DATA TRANSFORMATION :

This step involves the transformation of data that are of the type object to float if the values inside the column are numerical but are input in string format. This step also converts two columns to date-time format which makes the operations and analysis much easier and more efficient.

```
[ ] # list of columns that are currently of type object
    object_cols = lendingdf.select_dtypes(include='object').columns

    # Convert all numeric columns that are currently of type object to float
    for col in object_cols:
        if lendingdf[col].str.isnumeric().all():
            lendingdf[col] = pd.to_numeric(lendingdf[col], downcast='float')
```

```
lendingdf.dtypes
```

```
loan_amnt              float64
term                     int64
int_rate               float64
installment            float64
grade                   object
sub_grade               object
emp_title               object
emp_length              object
home_ownership          object
annual_inc             float64
verification_status     object
issue_d                 object
loan_status             object
purpose                 object
title                   object
dti                    float64
earliest_cr_line        object
open_acc               float64
pub_rec                float64
revol_bal              float64
revol_util             float64
total_acc              float64
initial_list_status     object
application_type        object
mort_acc               float64
pub_rec_bankruptcies   float64
street                  object
state                   object
zip_code                object
dtype: object
```

```
# Convert 'issue_d' and 'earliest_cr_line' columns to datetime
lendingdf['issue_d'] = pd.to_datetime(lendingdf['issue_d'])
lendingdf['earliest_cr_line'] = pd.to_datetime(lendingdf['earliest_cr_line'])

# Print the resulting dataframe
print(lendingdf[['issue_d', 'earliest_cr_line']].dtypes)
print(lendingdf[['issue_d', 'earliest_cr_line']].head())
```

```
issue_d             datetime64[ns]
earliest_cr_line    datetime64[ns]
dtype: object
     issue_d earliest_cr_line
0 2015-01-01       1990-06-01
1 2015-01-01       2004-07-01
2 2015-01-01       2007-08-01
3 2014-11-01       2006-09-01
4 2013-04-01       1999-03-01
```

9.4. DATA REDUCTION :

This step breaks the data into simpler proportions to enable easier processing and analysis of data. The term column is shortened down to a numerical value to make it easier for conditions and comparisons. The address column is broken down into simpler aspects such as street, state, and pin code to get a deeper and independent focus for analysis.

```
[ ] term_values = {' 36 months': 36, ' 60 months': 60}
    lendingdf['term'] = lendingdf.term.map(term_values)
```

```
lendingdf.head()
```

	loan_amnt	term	int_rate	installment	grade	sub_grade	emp_title	emp_length	home_ownership	annual_inc	verification_status	issue_d	loan_status	purpose	titl
0	10000.0	36	11.44	329.48	B	B4	Marketing	10+ years	RENT	117000.0	Not Verified	Jan-2015	Fully Paid	vacation	Vacatio
1	8000.0	36	11.99	265.68	B	B5	Credit analyst	4 years	MORTGAGE	65000.0	Not Verified	Jan-2015	Fully Paid	debt_consolidation	Del consolidatio
2	15600.0	36	10.49	506.97	B	B3	Statistician	< 1 year	RENT	43057.0	Source Verified	Jan-2015	Fully Paid	credit_card	Credit car refinancin
3	7200.0	36	6.49	220.65	A	A2	Client Advocate	6 years	RENT	54000.0	Not Verified	Nov-2014	Fully Paid	credit_card	Credit car refinancin
4	24375.0	60	17.27	609.33	C	C5	Destiny Management Inc.	9 years	MORTGAGE	55000.0	Verified	Apr-2013	Charged Off	credit_card	Credit Car Refinanc

```
# Split the address column into street, state, and zip code
split_address = lendingdf['address'].str.split(',', expand=True)
split_address2 = split_address[1].str.split(' ', n=1, expand=True)
lendingdf['street'] = split_address[0]
lendingdf[''] = split_address2[1]
split_state_zip = lendingdf[''].str.split(' ', n=1, expand=True)
lendingdf['state'] = split_state_zip[0]
lendingdf['zip_code'] = split_state_zip[1]

# Remove unnecessary columns
lendingdf.drop(['address', ''], axis=1, inplace=True)

# Print the resulting dataframe
print(lendingdf[['street', 'state', 'zip_code']].head())
```

```
                                     street state zip_code
0      0174 Michelle Gateway\r\nMendozaberg    OK    22690
1  1076 Carney Fort Apt. 347\r\nLoganmouth    SD    05113
2  87025 Mark Dale Apt. 269\r\nNew Sabrina    WV    05113
3             823 Reid Ford\r\nDelacruzside    MA    00813
4              679 Luna Roads\r\nGreggshire    VA    11650
```

9.5. REMOVING OUTLIERS :

This step involves removing the unnecessary outliers from the dataset such that data values are more intrinsic and consistent. Elimination of outliers is accomplished by calculating the interquartile range of each numerical column and then removing the outliers which are 1.5 times less than the first quartile or 1.5 times more than the third quartile and then being replaced by null values. Later the null values are replaced by the average value of that respective column such that the entire column values are now consistent and mannerly.

```
[33] # Remove everything except the number from emp_length
     lendingdf['emp_years'] = lendingdf['emp_length'].str.extract('(\d+)')
     lendingdf = lendingdf.drop('emp_length', axis=1)
```

```
[34] # Print the resulting dataframe
     print(lendingdf[['emp_years']].head())
```

```
  emp_years
0        10
1         4
2         1
3         6
4         9
```

```
# Select numerical columns
num_cols = ['loan_amnt', 'term', 'int_rate', 'installment', 'annual_inc', 'revol_bal']

# Calculate IQR for each column
Q1 = lendingdf[num_cols].quantile(0.25)
Q3 = lendingdf[num_cols].quantile(0.75)
IQR = Q3 - Q1

# Remove outliers and replace with column average
lendingdf[num_cols] = np.where((lendingdf[num_cols] < (Q1 - 1.5 * IQR)) | (lendingdf[num_cols] > (Q3 + 1.5 * IQR)), np.nan, lendingdf[num_cols])
lendingdf[num_cols] = lendingdf[num_cols].fillna(lendingdf[num_cols].mean())
lendingdf[num_cols].head()
```

	loan_amnt	term	int_rate	installment	annual_inc	revol_bal
0	10000.0	36.0	11.44	329.48	117000.0	36369.0
1	8000.0	36.0	11.99	265.68	65000.0	20131.0
2	15600.0	36.0	10.49	506.97	43057.0	11987.0
3	7200.0	36.0	6.49	220.65	54000.0	5472.0
4	24375.0	36.0	17.27	609.33	55000.0	24584.0

10. MODELING :

10.1. TRAIN AND TEST DATA SPLIT :

LendingClubDefaultersPrediction.ipynb

File Edit View Insert Runtime Tools Help Last edited on May 10

Comment Share

+ Code + Text

RAM Disk

Train and Test data Split

```
from sklearn.model_selection import train_test_split
```

```
w_p = cleanseddata.loan_status.value_counts()[0] / cleanseddata.shape[0]
w_n = cleanseddata.loan_status.value_counts()[1] / cleanseddata.shape[0]

print(f"Weight of positive values {w_p}")
print(f"Weight of negative values {w_n}")
```

```
Weight of positive values 0.8038709188697826
Weight of negative values 0.1961290811302174
```

```
train, test = train_test_split(cleanseddata, test_size=0.30, random_state=40)

print(train.shape)
print(test.shape)
```

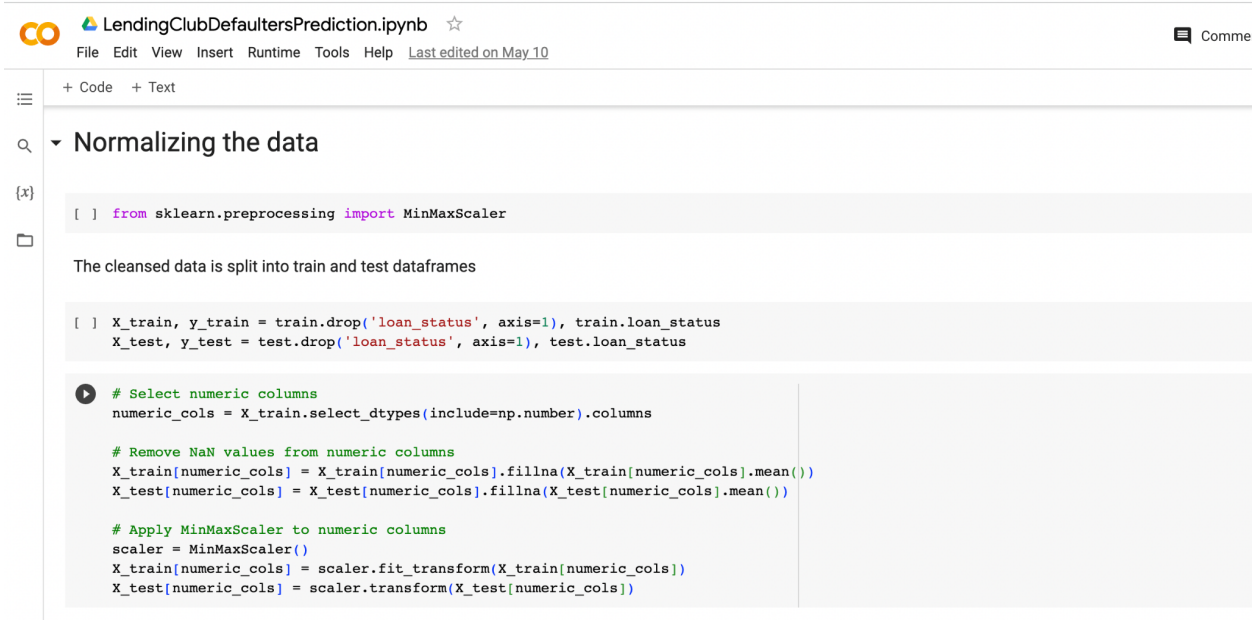
```
(265340, 29)
(130690, 29)
```

In the above "Train and Test Data Split" step, we are dividing the dataset into two separate sets: a training set and a testing set. It is done to evaluate the performance of our predictive model on unseen data.

I. We randomly select a portion of the dataset to be used as the training set. The subset of data is used to train the machine learning model, allowing it to learn patterns and relationships between the input features and the target variable.
II. The remaining portion of the dataset is set aside as the testing set. This set is not used during the training phase but is used to evaluate the model's performance after training.

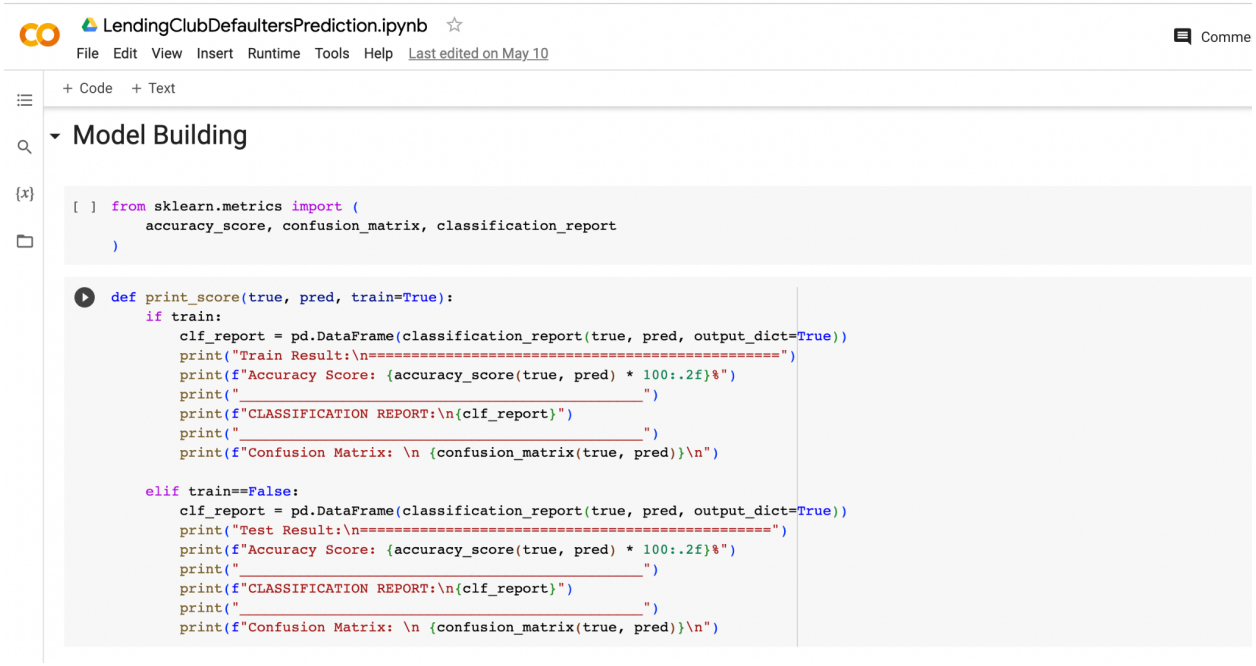
By splitting the data into these two sets, we can assess the model's accuracy, evaluate its performance metrics and make any necessary adjustments or improvements before deploying the model for real-world predictions.

10.2. DATA NORMALIZATION:



In the above step, we first remove any NaN (missing) values from the numeric columns. It is done to ensure that the dataset does not contain any missing or invalid values, which could negatively impact the performance of the scaler. Once the NaN values are removed, we are scaling the numerical features of the dataset to a standardized range, typically between 0 and 1, to ensure fairness in their contribution to the model and prevent features with larger magnitudes from dominating the learning process.

10.3. MODEL BUIDLING :

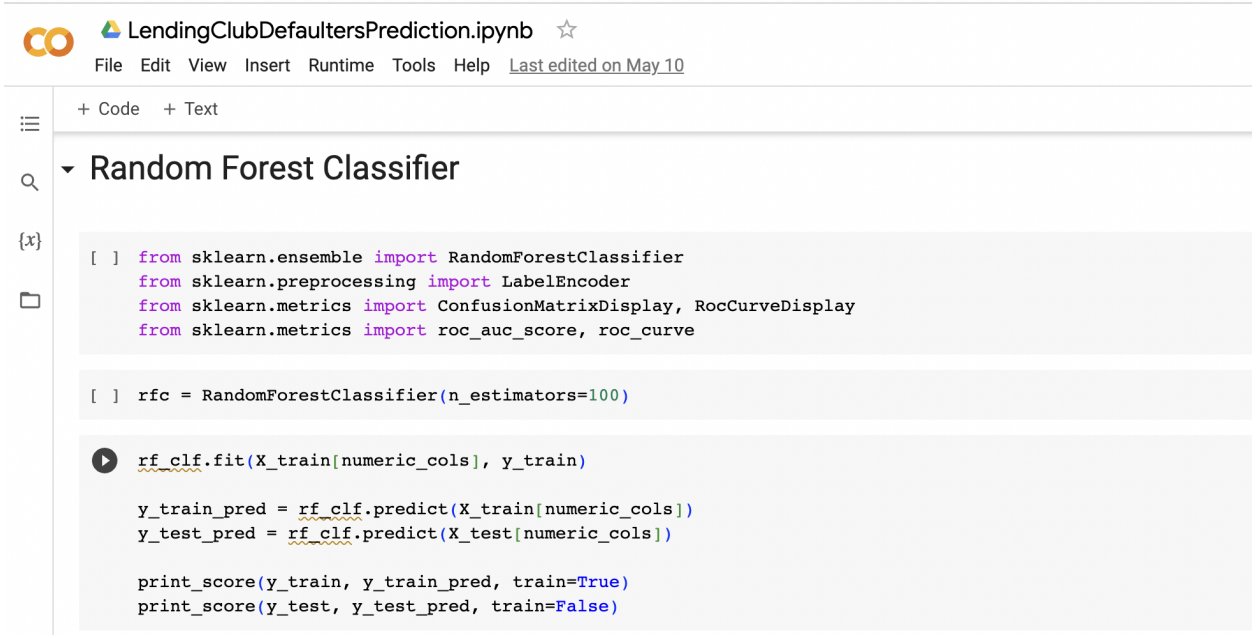


The above code defines a function called "**print_score**" that is used to print the evaluation metrics and results of a classification model.

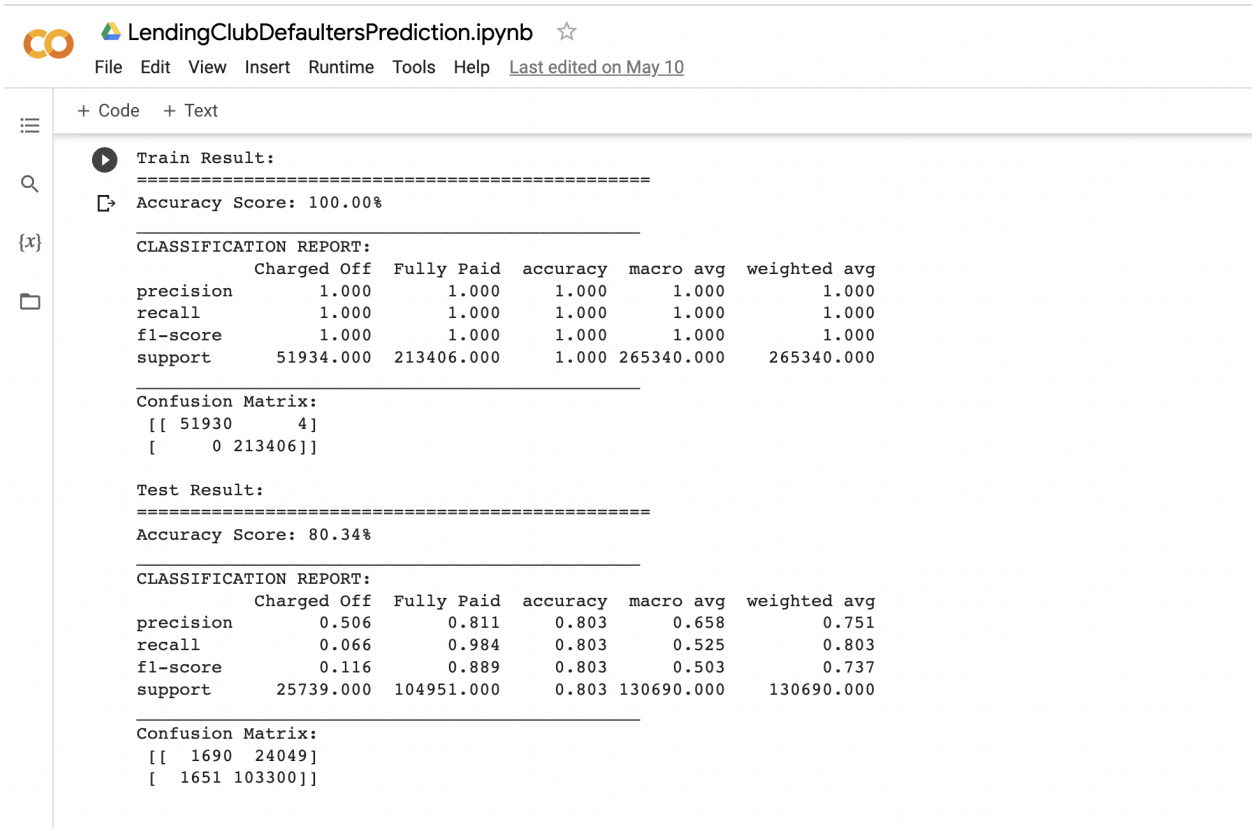
If the train parameter is set to **True**, it means the function is being called to print the evaluation metrics for the training set. It calculates and displays the accuracy score, classification report (which includes precision, recall, F1-score, and support), and the confusion matrix.
If the train parameter is set to **False**, it means the function is being called to print the evaluation metrics for the test set. Similar to the previous case, it calculates and displays the accuracy score, classification report, and confusion matrix for the test set.

Hence, this function helps in assessing the performance of a classification model on either the training or test data by providing relevant evaluation metrics and results.

RANDOM FOREST CLASSIFIER :

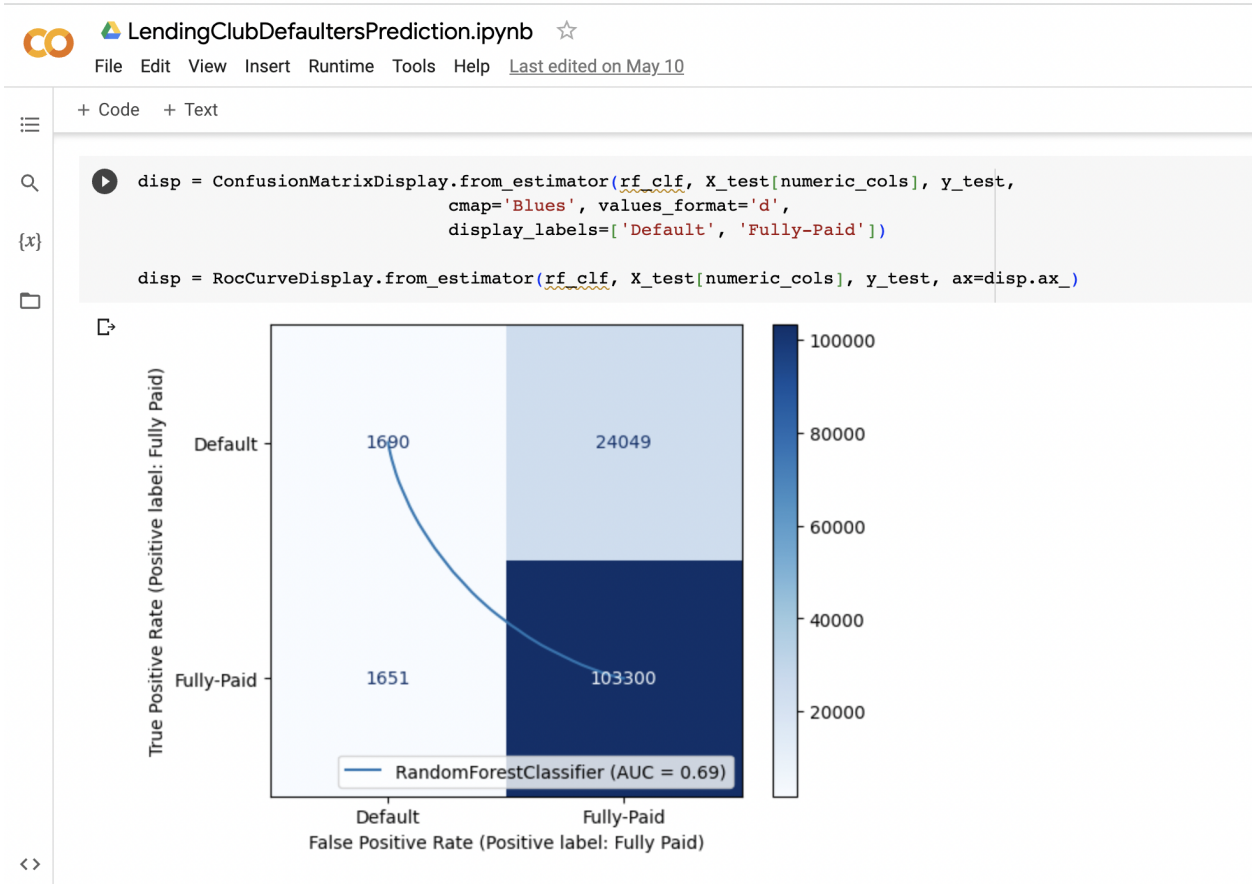


In the above, the code is first creating an instance of the Random Forest Classifier using the RandomForestClassifier() function. It then fits the classifier to the training data using the fit() method, where X_train[numeric_cols] represents the input features and y_train represents the target variable.
After training the classifier, it makes predictions on both the training data (X_train[numeric_cols]) and the test data (X_test[numeric_cols]) using the predict() method. The predicted values are stored in the variables y_train_pred and y_test_pred, respectively.
Finally, the print_score() function is called to print the performance metrics of the classifier for both the training and test sets, providing insights into the accuracy, classification report, and confusion matrix of the model's predictions.



Hence, in the training set, the Random Forest Classifier achieved perfect accuracy of 100%. It correctly classified all instances of both the "Charged Off" and "Fully Paid" categories. The confusion matrix shows only four misclassifications.
In the test set, the classifier achieved an accuracy of 80.34%. The precision for "Charged Off" is relatively low at 50.6%, indicating some false positives, while the precision for "Fully Paid" is higher at 81.1%. The recall (true positive rate) for "Charged Off" is low at 6.6%, suggesting missed predictions, while the recall for "Fully Paid" is high at 98.4%. Overall, the model performs better in predicting "Fully Paid" instances than "Charged Off" instances. The confusion matrix shows a larger number of misclassifications compared to the training set.

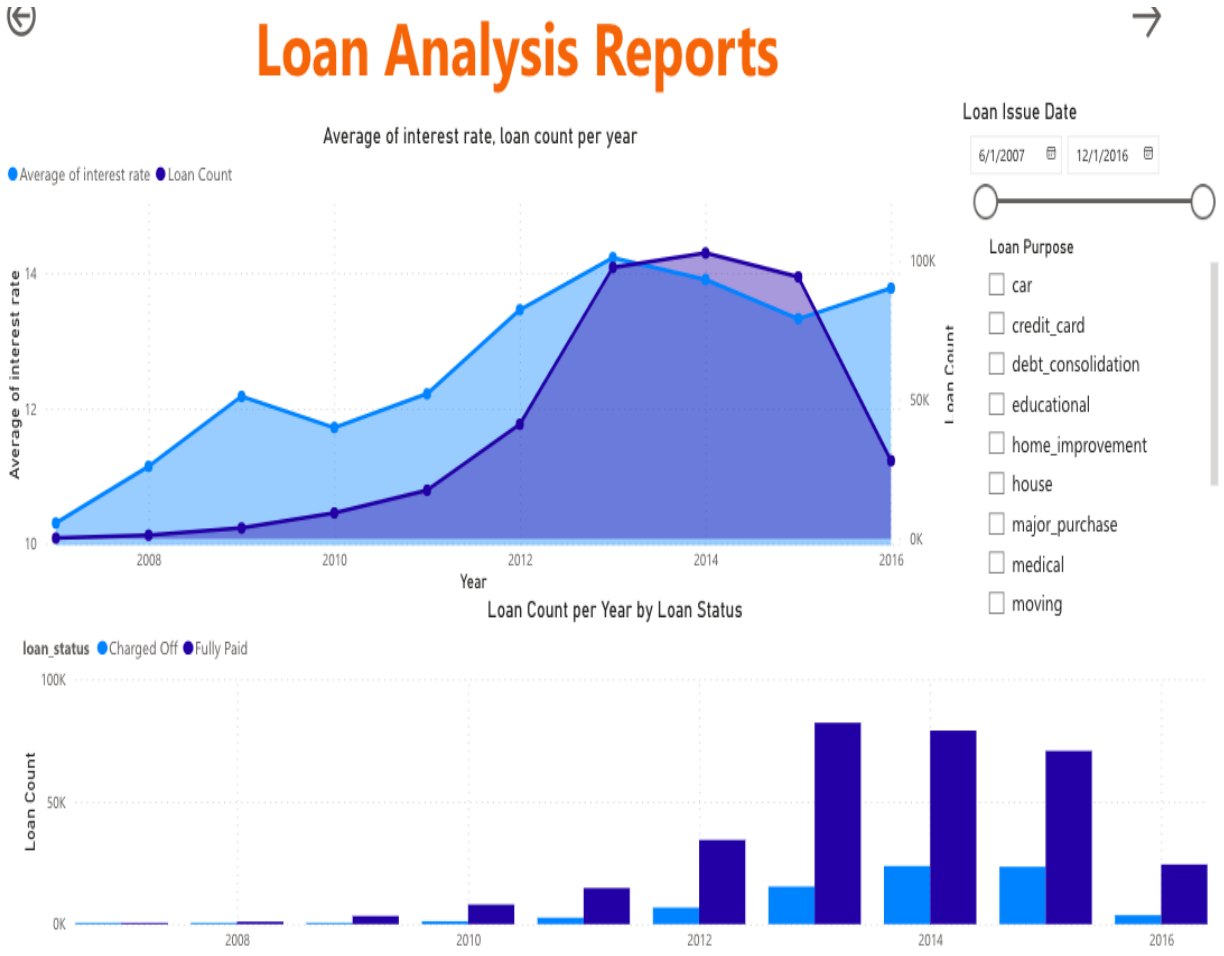
CONFUSION MATRIX :



In the above, the code first creates a ConfusionMatrixDisplay using the trained Random Forest Classifier (rf_clf) and the test data (X_test[numeric_cols] and y_test).
It visualizes the confusion matrix, which displays the true positive, false positive, true negative, and false negative counts, using a blue color map (cmap='Blues'). The values_format='d' parameter formats the values in the confusion matrix as integers. The display_labels parameter specifies the labels to be shown on the confusion matrix, which are 'Default' and 'Fully-Paid'.

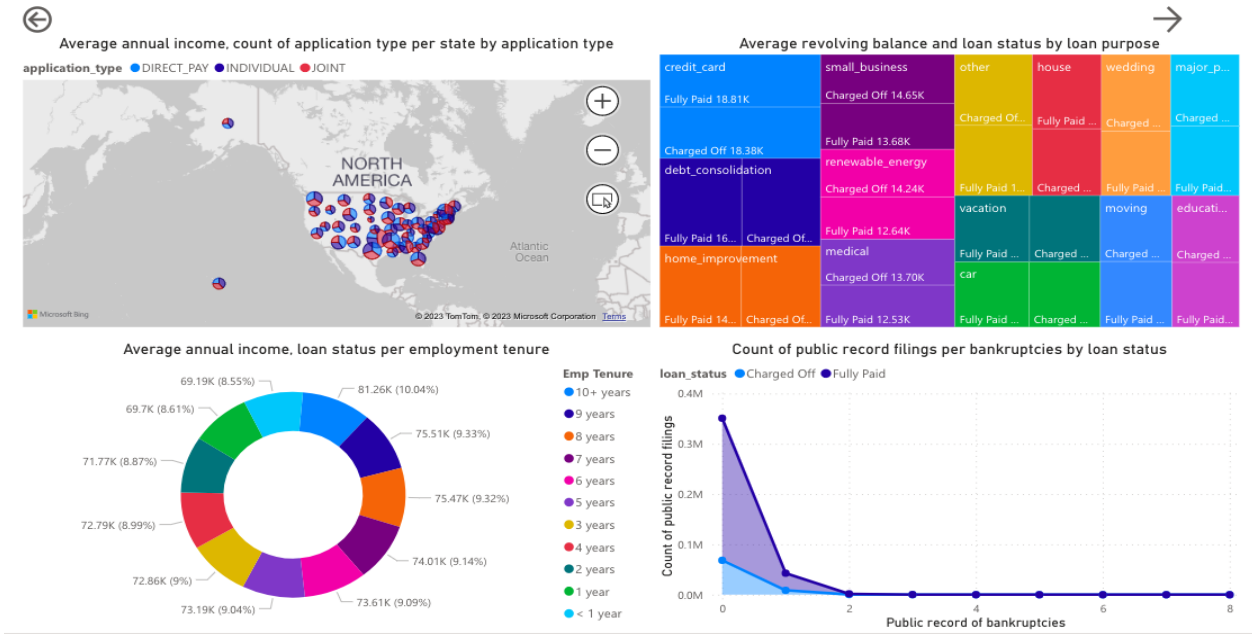
Secondly, it creates a RocCurveDisplay using the same classifier and test data, but this time it is overlaid on the ConfusionMatrixDisplay (ax=disp.ax_). The RocCurveDisplay visualizes the receiver operating characteristic (ROC) curve, which shows the trade-off between the true positive rate and false positive rate for different classification thresholds.

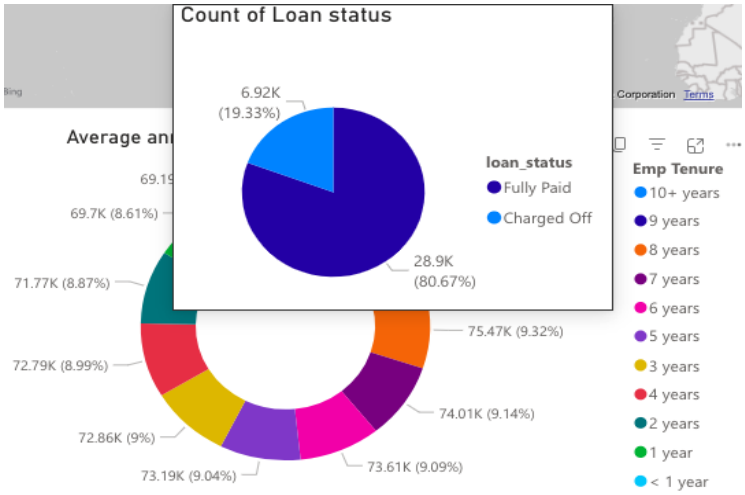
11. DATA VISUALIZATIONS :



The above report provides insights into how the average interest rates and the number of loans issued have changed over time. The user can filter the data by issue date and purpose, which can help identify trends and potential areas for improvement in the lending process. Thus, the visual allows the user to understand the impact of verification status on interest rates and loan volume over time.

The side-by-side bar chart is filtered to show only issue dates on or after Friday, June 01, 2007. It can help to identify trends in the frequency of loans being issued over time, as well as potential changes in lending practices or regulations that may have affected the loan issuance volume. Hence, it may help to identify any seasonality in the loan issuance patterns, such as increased activity during certain times of the year.

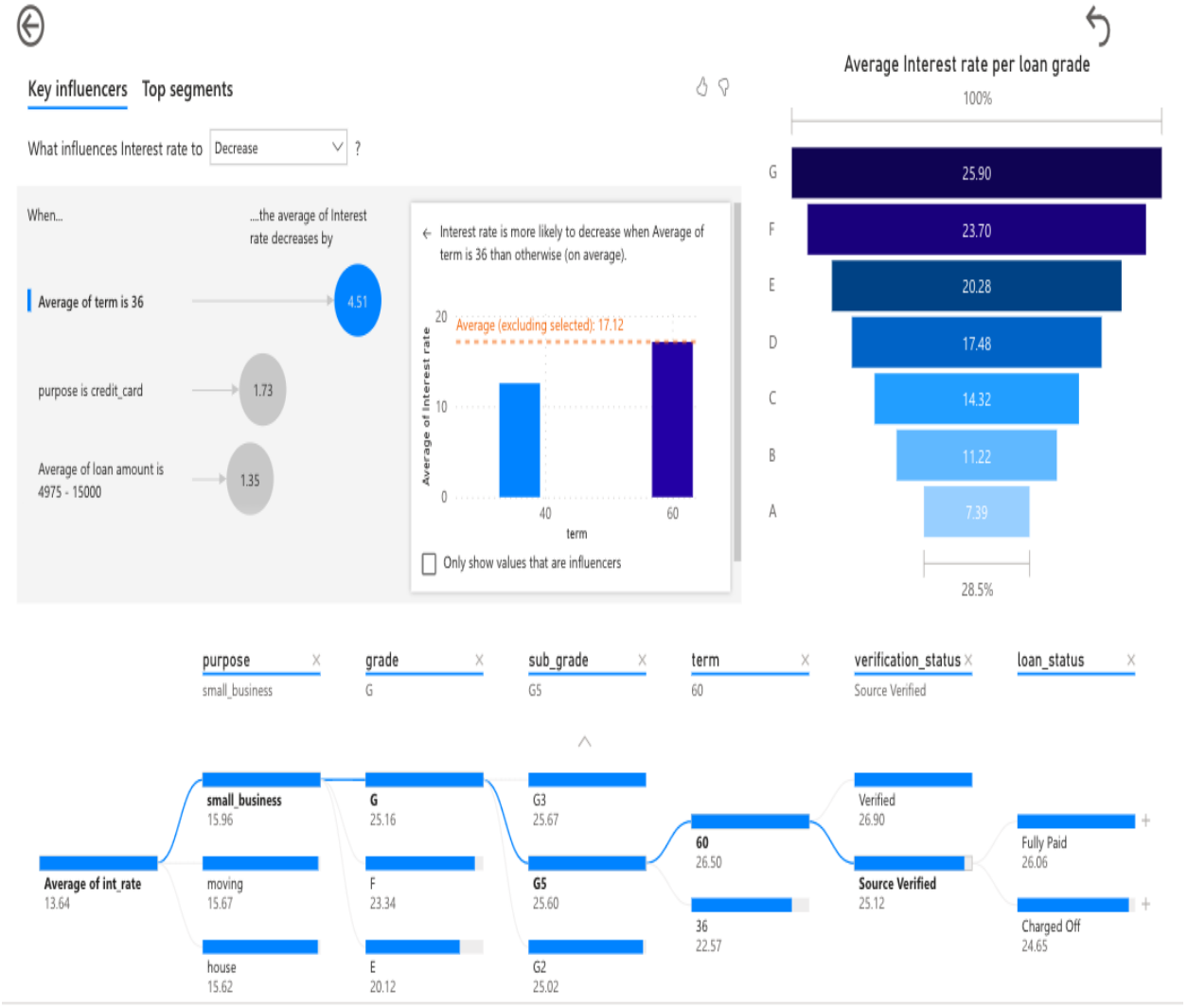




The above report displays a wide variety of valuable insights considering a range of influential factors. The top left map shows us the average annual income in each state through the size of the pie chart. The pie chart is characterized by different application types and their count per each state. Coming to the tree map, the average revolving balance is displayed according to its loan status which is characterized by the purpose of loan differentiated by colors.

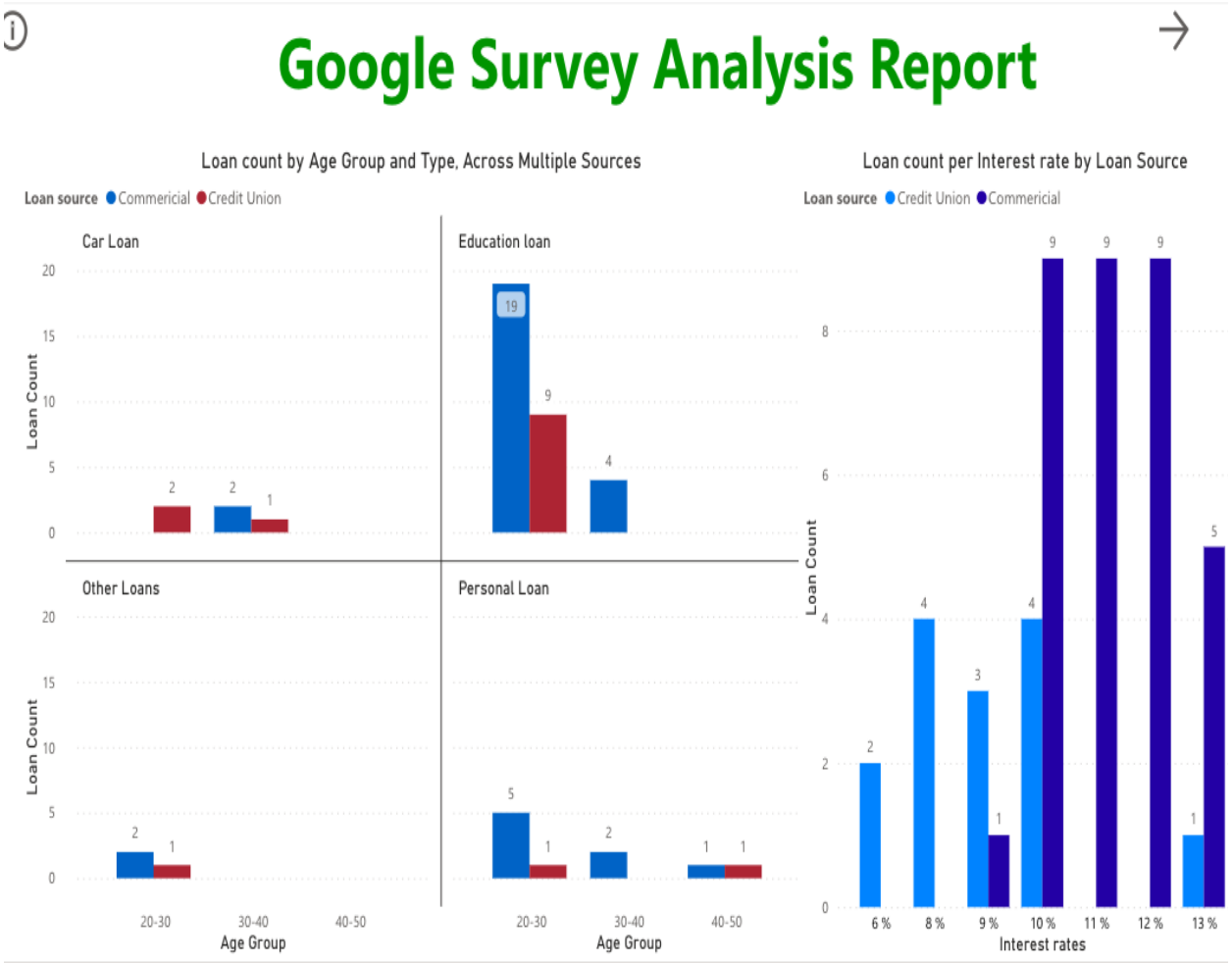
The donut chart displays the count of loan status according to their annual income and tenure of the loan amount being taken. It is observed above that when we hover around the donut chart a visual representation of the pie chart is displayed showing the details of the count of loan status according to each category in the donut chart. It is being divided according to the loan status and represented accordingly. The different colors represent the category of tenure period they belong to.

The line graph is plotted with the total number of public filings against the total number of bankruptcies to grasp an overview of data distribution under the category of the different loan statuses. It is observed from the graph that the majority of the applicants have fully paid the loan amount before recording a bankruptcy while there still exists a considerable amount of the population who have gone bankrupt before completely paying back their loan amount.



The above report is created to display insights about what factors influence the interest rate to increase or decrease. When the choice is given as decreased, it is shown that the interest rate decreases when the average term revolves around 36 months. The second factor is the purpose of the loan amount which is influential when it is for credit cards. The third factor is when the average loan amount is around $4975 to $15000. The degree of influence is displayed by the top left plot in the report. The plot in the top right corner differentiates each loan grade by its average interest rate.

The third plot displays a network around the attributes of the data connecting each influential factor to the next metric. It is shown that when the average interest rate is 13.64, the purpose is a small business, the grade of the loan is G, the subgrade is G5, the term is 60, and the verification status is ‘Source Verified’, then the stats of fully paid is 26.06 while charged off is 24.65



The above report is created to analyze and visualize SJSU Google survey data. From the data, we focus on the analysis of different types of loans that the respondents have taken and the current interest rates they are paying, in general, to compare it with the Lending Club company to compare the competition. Clustered bar charts were used for comparison between the different loan sources. On the left side, we have the analysis of the number of loans taken by each age group from different types of banks. We can see that a total of 3 people have not mentioned the loan type but out of them 2 people who belong to the 20-30 years age group have taken the loan from the commercial bank but the other person who belongs to the same age group has taken the loan from a credit union. Out of the different kinds of loans taken, a lot of our respondents have education loans. In the age group 20-30, the majority of them have education loans with around 28 people. Out of them, 19 people have taken the loan from commercial banks and 9 people have taken the loan from credit unions.

From our survey data, we can see that it is a common trend to take loans from a commercial bank. We further analyzed the interest rates that the people are paying currently for the loans they have taken. The clustered bar chart on the right shows the number of people paying a certain interest rate for each type of bank. We can see that commercial banks usually charge a higher interest rate as compared to credit unions. The majority of the people from the survey have interest rates greater than 9% and the majority of them have loans from commercial banks. Commercial banks have no loans with less than 7%.

12. DEPLOYMENT :

The Power BI report is deployed from a local environment to Power BI Online. It enables increased user collaboration and accessibility.

It is crucial to consider scalability and the dynamic nature of the users' environment. Scalability in this context involves ensuring that the Power BI reports and dashboards can handle increased user demands and growing data volumes. It can be achieved by optimizing data models, leveraging data refresh schedules, and utilizing the cloud-based infrastructure provided by Power BI Online. It ensures that the system can efficiently handle larger user loads without compromising performance.

The deployment strategy should prioritize flexibility and modularity to facilitate seamless integration with Power BI Online. It involves considering factors such as data connectivity options, data refresh frequency, and security settings. It is also important to actively seek user feedback and engagement to ensure that the deployed reports and dashboards align with user expectations.

As a result, publishing the Power BI report to Power BI Online increases accessibility, encourages collaboration, makes it possible to update data, makes sharing and distribution easier, and ensures security and governance. It enables users to fully utilize the report's potential and make data-driven decisions in a secure and collaborative setting.

13. DISCUSSION :

- The analysis of survey data is more tilted towards the different types of loans the applicants have taken and the amount of interest they are paying. The majority of the loans are Education loans which are commercial-oriented compared to credit unions. But it is evident that commercial loans are of higher interest than credit union loans because there is no loan given by a commercial bank that is less than 7%.

- It is observed from the Loan Analysis Report that the number of loans taken dramatically increased during the period of 2012 and then came to a stable state in 2016. This might indicate a seasonal trend in the loan system which can be caused by many potential factors such as an improved education system, profitable real estate, high trading outcomes, low capital for business start-ups, etc for which many applicants are interested in taking loans.

- It is observed from the analysis that the higher the tenure of the loan amount, the more likely it is to be fully paid compared to being charged off. Joint accounts are slowly rising above the norms and seem to be a feasible choice for individual applicants. Among the major reasons for loan, ‘Credit card’, ’Debt consolidation’, and ‘Home Improvement’ tops the list. There still exists a considerable amount of the population who go bankrupt before they are able to pay back the loan and the bank has to bear the losses.

- It can be incited out of the reports that the interest rate is highly dependent on the term period, second on the purpose of the loan, and third, the amount of loan taken. It is also seen that the higher the loan grade the higher is the interest rate which can be a useful piece of information for loan applicants.

14. IMPACT :

The analysis of the "LendingClub" dataset has had a significant impact on the domain of credit risk assessment and loan investments.

- **Better Risk Assessment:** The lending sector has been significantly impacted by the development and enhancement of a prediction model for identifying potential loan defaulters using the dataset. Lenders are able to make more accurate risk assessments and well-informed decisions about loan approvals by utilizing the insights obtained from the dataset. As a result, there are now fewer prospective defaulters, which reduces financial losses for lenders and boosts overall loan investment profitability. As a result, the impact of this dataset and the resulting predictive model are seen in the improved efficiency and effectiveness of loan operations, enabling lenders to optimize their lending practices and achieve better financial outcomes.

- **Improved Decision-Making:** The dataset has offered insightful information about the elements that affect loan default. Lenders can gain a more thorough understanding of the risk variables related to loan applicants by examining the correlations and linkages within the data using our Power BI reports. It enables them to distribute resources effectively, analyze creditworthiness more effectively, and make data-driven decisions.

- **Industry Impact:** The dataset analysis revealed information that had an impact on the lending sector as a whole. The findings can be used by lending platforms and financial institutions to improve their loan approval workflows, risk management procedures, and portfolio management. It may result in greater profitability, lower default rates, and more responsible lending practices.

- **Risk reduction:** Using the Random Forest Classifier and the analysis, lenders may identify high-risk loan applicants and put the right risk reduction measures in place. It can entail changing interest rates, tightening eligibility requirements, or offering more assistance to debtors who are thought to be at risk of default. The dataset has thereby helped to reduce financial losses and enhance the performance of loan portfolios.

15. CONCLUSION :

This analysis focuses on using the Random Forest Classifier to minimize risks in loan investments. By analyzing the Lending Club loan defaulters prediction dataset, valuable insights have been gained using different visualization techniques with the help of Python and PowerBI. The study successfully developed a predictive model that can identify potential loan defaulters based on key factors in the dataset. The application of the machine learning technique, ie; the Random Forest algorithm, enables lenders to make informed decisions and improve the accuracy of loan investment assessments. It is important to note that the model should be used as a supportive tool rather than the sole determinant in decision-making. Further research and refinement of the model are necessary to ensure its effectiveness in real-world lending scenarios.
Hence, our findings highlight the significance of leveraging data-driven approaches to mitigate risks and optimize loan investment strategies.

16. FUTURE SCOPE :

- **Model Optimization:** Investigating various techniques to optimize the Random Forest Classifier model specifically for loan default prediction. It can include tuning hyperparameters, exploring different ensemble methods, or experimenting with feature selection techniques to improve the model's accuracy and performance.

- **Feature Engineering:** Exploring additional feature engineering techniques to enhance the predictive power of the model. This may involve creating new features based on domain knowledge, deriving ratios, or incorporating external data sources that provide valuable insights into borrower behavior.

- **Ensemble Methods:** Exploring the combination of Random Forest Classifier with other machine learning algorithms or ensemble methods to create hybrid models. It can potentially improve the robustness and stability of the predictions by leveraging the strengths of different algorithms.

- **Imbalanced Data Handling:** Addressing the issue of imbalanced data, where the number of non-defaulting loans significantly outweighs the number of defaulting loans.

- **Interpretability and Explainability:** Investigating methods to enhance the interpretability and explainability of the Random Forest Classifier model. It can involve techniques such as feature importance analysis etc, to provide insights into the factors driving loan default predictions.

- **Real-time Prediction and Monitoring:** Extending the model to enable real-time loan default predictions, monitoring, and setting up automated monitoring systems to identify and flag potential default risks in real-time.

- **Integration with Portfolio Management Systems:** Integrating the loan default prediction model with portfolio management systems used by lenders and investors. It would allow for seamless integration of risk assessment and decision-making processes, enabling lenders to make more informed loan investment decisions and manage their portfolios more effectively.

17. REFERENCES :

“Credit risk assessment: a challenge for financial institutions" by Evangelos Kalapodas; Mary E. Thomsom: *https://ieeexplore.ieee.org/document/8132501/authors#authors*

"Data Visualization and its Key Fundamentals: A Comprehensive Survey" by Muscan, Gurpreet Singh, Jaspreet Singh, and Chander Prabha **:** *https://ieeexplore.ieee.org/document/9835803*

“The Application Study of Credit Risk Model In Financial Institution via Machine-learning Algorithms" by Yuanzhang Wang; Jiongcheng Lu; Jiehan Qin; Chenyi Zhang; Yiyang Chen: *https://ieeexplore.ieee.org/document/9532215*

“Comparative Breast Cancer Detection with Artificial Neural Networks and Machine Learning Methods” by Muhammed Coşkun Irmak; Mehmet Bilge Han Taş; Sedat Turan; Abdulsamet Haşiloğlu: *https://ieeexplore.ieee.org/document/9477991*

SAYAH, F. (2023, February). Lending Club Dataset, Version 1. Retrieved March 20, 2023 from https://www.kaggle.com/code/faressayah/lending-club-loan-defaulters-prediction/input.