APANPS5310_D01

Loan RDMS for Lending Club

Project Report

Group 3:

Jennifer Wang

Kathy Yuehan Wang

Li Su

Yimei Yang

Table of Contents

Chapter 1 Problem Description

1.1 Client Introduction

As the trailblazer in peer-to-peer lending, Lending Club has evolved into America's largest online marketplace that allows borrowers to apply for personal loans, auto refinancing, business loans, and elective medical procedures.

Mission: Company mission is to transform the banking system to make credit more affordable and investing more rewarding.

Operation: By building "bridge" between investors and borrowers, Lending Club help borrowers to acquire fundings they need and help investors to earn a profit depending on the risk they take.

1.2 Problem Statement

Management Dilemma

Since 2007, Lending Club has been bringing borrowers and investors together, transforming the way people access credit. Over the last 10 years, we've helped millions of people take control of their debt, grow their small businesses, and invest for the future. However, with more and more customers coming in, Lending Club is facing a lot of archived, unstructured data that need to be organized for easy access and analysis. Meanwhile, Lending Club is also trying use these data to get a full understanding of current business circumstance, identify more valuable customers and explore new market.

Management Problem

How to structure archived raw data to build data analysis in order to get business insight for both data analysts and manager/C-executives?

Unstructured Dataset

The raw loan data we used is a reduced sample set that contains 5,000 records of loans issued through the 2008-2018, and 97 attributes covering different information such as current loan status, loan owner demographic information, customer financial credit history, payments, collections, secondary applicants, etc. This dataset contains redundant information in many places. The original dataset can be found at Kaggle website:https://www.kaggle.com/wendykan/lending-club-loan-data

After carefully analyzing raw dataset, we noticed that there are many co-relationships among these 97 attributes. For example, a member's *employee title*, *employee length*, *home ownership*, *income* are only dependent on member id, thus should be reorganized in a separate table. And the *zipcode* (first three numbers) and *address_state* columns are only dependent on on each other. We also identified that the interest rate issued to each member is only determined by the member's credit *grade* and *sub_grade*. Therefore this raw dataset is perfect to be restructured as a relational database to reduce redundancy and improve data integrity. Based on the data insight, we established and populated the raw data into 17 tables. Here we provide a screenshot and the link to the complete dataset:

https://docs.google.com/spreadsheets/d/1W7Z0xgIgOECHlfH1VGxTsCD9twrWq0eetEgDW0N3ve4/edit#gid=1302175229

memb	loan_amr	funded_ar	installment	issue_d	dti	total_rec_pr	total_rec_i	out_prncp	last_credit_pull_d	disbursemen	debt_	term	loan_status	purpose	title	emp_title	emp_length	home_owners
NA	5200	5200	185.17	Dec-2018	21.27	5200	135.31	0	Jan-2019	Cash	N	36 months	Fully Paid	debt_consolidatic	Debt consolidation	Athletic Director	5 years	OWN
NA	2500	2500	76.58	Dec-2018	10.96	2500	12.56	0	Jan-2019	DirectPay	N	36 months	Fully Paid	debt_consolidatic	Debt consolidation		n/a	OWN
NA	39500	39500	1209.92	Dec-2018	11.86	1998.68	336.11	37501.32	Feb-2019	DirectPay	N	36 months	Current	debt_consolidatic	Debt consolidation		< 1 year	MORTGAGE
NA	36000	36000	1131.27	Dec-2018	13.37	1777.19	380.7	34222.81	Feb-2019	Cash	N	36 months	Current	credit_card	Credit card refinancing	Professor of Eng	10+ years	MORTGAGE
NA	16000	16000	502.79	Dec-2018	17.69	1547.24	192.93	14452.76	Feb-2019	Cash	N	36 months	Current	debt_consolidatic	Debt consolidation	Goods flow Lead	5 years	RENT
NA	30000	30000	951.35	Dec-2018	18.98	1467.57	398.42	28532.43	Feb-2019	Cash	N	36 months	Current	credit_card	Credit card refinancing	Division Controlle	10+ years	MORTGAGE
NA	30000	30000	978.19	Dec-2018	28.73	1426.72	511.79	28573.28	Feb-2019	Cash	N	36 months	Current	home_improvem	Home improvement	Table Games De	7 years	OWN
NA	25000	25000	815.16	Dec-2018	31.83	1188.94	426.49	23811.06	Feb-2019	DirectPay	N	36 months	Current	home_improvem	Home improvement	Director	10+ years	MORTGAGE
NA	20000	20000	628.49	Dec-2018	23.91	1187.84	296.44	18812.16	Feb-2019	DirectPay	N	36 months	Current	credit_card	Credit card refinancing	Registered Nurse	9 years	MORTGAGE
NA	25000	25000	880.66	Dec-2018	25.58	1096.14	642.76	23903.86	Feb-2019	Cash	N	36 months	Current	debt_consolidatic	Debt consolidation	gentlemen sheet	10+ years	RENT
NA	40000	40000	856.4	Dec-2018	6.85	1028.55	672.34	38971.45	Feb-2019	Cash	N	60 months	Current	debt_consolidatic	Debt consolidation	Technology Cons	1 year	MORTGAGE
NA	20000	20000	612.62	Dec-2018	10.45	1012.62	198.26	18987.38	Feb-2019	Cash	N	36 months	Current	credit_card	Credit card refinancing	Accountant	2 years	RENT
NA	20000	20000	617.73	Dec-2018	1.44	1004.39	223.27	18995.61	Feb-2019	Cash	N	36 months	Current	credit_card	Credit card refinancing		< 1 year	RENT
NA	20000	20000	622.68	Dec-2018	9.93	996.49	240.47	19003.51	Feb-2019	Cash	N	36 months	Current	debt_consolidatic	Debt consolidation	Subject Matter E	2 years	MORTGAGE
NA	20000	20000	628.49	Dec-2018	26.77	987.34	251.44	19012.66	Feb-2019	Cash	N	36 months	Current	credit_card	Credit card refinancing	director	4 years	MORTGAGE
NA	39075	39075	865.26	Dec-2018	16.58	966.78	776.55	38108.22	Feb-2019	Cash	N	60 months	Current	debt_consolidatic	Debt consolidation	Accountant	10+ years	MORTGAGE
NA	40000	40000	909.72	Dec-2018	34.02	959.27	816.9	39040.73	Feb-2019	Cash	N	60 months	Current	debt_consolidatic	Debt consolidation	Systems Enginee	8 years	MORTGAGE
NA	40000	40000	921.64	Dec-2018	16.27	944.59	868.56	39055.41	Feb-2019	Cash	N	60 months	Current	credit_card	Credit card refinancing	Foreman	10+ years	RENT
NA	20000	20000	679.29	Dec-2018	13.71	911.7	431.81	19088.3	Feb-2019	Cash	N	36 months	Current	debt_consolidatic	Debt consolidation	Project Manager	10+ years	MORTGAGE
NA	17500	17500	540.51	Dec-2018	18.74	878.83	185.13	16621.17	Feb-2019	Cash	N	36 months	Current	credit_card	Credit card refinancing	Probation & Paro	10+ years	MORTGAGE
NA	20000	20000	712.16	Dec-2018	32.85	866.72	520.02	19133.28	Feb-2019	Cash	N	36 months	Current	debt_consolidatic	Debt consolidation	Door installer	10+ years	OWN
NA	35000	35000	775.03	Dec-2018	12.29	865.96	661.16	34134.04	Feb-2019	Cash	N	60 months	Current	credit_card	Credit card refinancing	Vice President	4 years	MORTGAGE
NA	16000	16000	490.1	Dec-2018	15.44	810.11	158.61	15189.89	Feb-2019	DirectPay	N	36 months	Current	credit_card	Credit card refinancing	Mechanic	10+ years	RENT
NA	30000	30000	619.99	Dec-2018	16.04	802.41	422.89	29197.59	Feb-2019	Cash	N	60 months	Current	debt_consolidatic	Debt consolidation	Lineman	5 years	MORTGAGE
NA	35000	35000	833.02	Dec-2018	34.73	794.83	1061.05	34205.17	Feb-2019	DirectPay	N	60 months	Current	credit_card	Credit card refinancing	Account Manage	10+ years	MORTGAGE
NA	18000	18000	650.48	Dec-2018	23.64	767.56	497.46	17232.44	Feb-2019	Cash	N	36 months	Current	debt_consolidatic	Debt consolidation	Procurement Spe	10+ years	RENT
NA	15000	15000	471.37	Dec-2018	23.84	740.51	185.17	14259.49	Feb-2019	Cash	N	36 months	Current	debt_consolidatic	Debt consolidation		n/a	RENT
NA	16000	16000	543.43	Dec-2018	31.53	729.36	381.61	15270.64	Feb-2019	DirectPay	N	36 months	Current	debt_consolidatic	Debt consolidation	Realtor	2 years	RENT
NA	35275	35275	933.01	Dec-2018	12.02	700.66	1126.32	34574.34	Feb-2019	Cash	N	60 months	Current	other	Other	Nurse	10+ years	OWN
NA	28000	28000	620.02	Dec-2018	20.56	692.76	519.75	27307.24	Feb-2019	Cash	N	60 months	Current	debt_consolidatic	Debt consolidation	Asset Manager	1 year	MORTGAGE
NA	4000	4000	131.55	Dec-2018	10.64	689.84	66.98	3310.16	Feb-2019	Cash	N	36 months	Current	major_purchase	Major purchase	General Manage	3 years	MORTGAGE

Fig 1.1 Raw Dataset

1.3 Client Needs

My client requires two levels of access to the data:

- Analysts: write and execute SQL code as well as access the database through Python/R
- Managers/C- executives: high-level overview through visualizations and interactive dashboards that automatically update when new data is stored in the database

Chapter 2 Project Proposal

2.1 Our Solution

After conducting researches to understand Lending Club's data availability, business scope and analytics demands, our team proposed to build a Relational Database Management System(RDMS) on PostgreSQL for Lending Club, together with an interactive dashboard connected to the database system to display visualized data insights.

Our design of the loan database aims to address three main data perspectives for Lending Club: Current Business Analysis, Customer Analysis and Risk Analysis.

2.2 Proposal Justification

RDMS is the primary data model in use today, and are the most popular way to interact with data. It organizes data into tables of columns and rows, with a unique key identifying each row. Data stored in a structured form can can significantly reduce iteration time. These features allow us to create systems that are consistent and accurate that support simple operations using SQL - the primary interface for querying and maintaining the relational database.

Using RDBMS can bring a systematic view to raw data of Lending Club. It is easy to understand and execute and hence enables better decision making. Meanwhile, RDMS allows triggers to enforce business rules that automatically invoke a procedure whenever a special event in the database occurs, such as a trigger to ensure each loan when input as new entry must be assigned with a credit grade among levels A, B,C,D,E,F,G.

Our choice of PostgreSQL is based on the fact that PostgreSQL is a powerful, open source relational database system that is known for strong reputation for reliability, feature robustness, and performance.

Chapter 3 Team Structure and Timeline

Task & Responsibility	Details	Team Member	Timeline
Explanation of Dataset	Why choose this dataset: detail the reasoning behind choices, what are the motivation. describe the source, type, and extent (how much) of the data.	Yuhan Wang, Li Su	Apr 5
Scenario Justification	Research performed in making the decision and initial plan of action, how work will improve the decision-making for the company that hired us, and what other benefits will be?	Yimei Yang, Yuehan Wang	Apr 5
Normalization Plan	Provide reasons of breaking the dataset and all steps in details	Yimei Yang, Yuehan Wang	Apr 7
Database Schema Design	Include all **integrity constraints**, **trigger**, and so on with SQL code for normalized tables	Yuhan Wang, Li Su	**Apr 7**
ER Diagram	Relationships between each table in PDF version (**Lucidchart**)	Yuhan Wang, Li Su	Apr 8
Populate dataset	Transforming and entering the data to your database system **Explain plan and reasoning**	Yimei Yang, Yueha Wang	Apr 8
ETL process	Explain all work for the process in detail with code listed(GitHub repo/Gist)	Yimei Yang, Yueha Wang, Yuhan Wan Li Su	Apr 13
Analytical Procedures	**At least 10 analytical procedures**, such as "How are the salespeople performing?" with SQL or R code	Yimei Yang, Yueha Wang, Yuhan Wan Li Su	Apr 13

Customer Interaction Plan	**What implement for analysts?**(How will analysts run the code for the analytical procedures you designed?) **C level officers?**(How will executives review the results?) **What tools?(**What tools/programming languages did you implement?) What were the **benefits** of performing databa actions with a programming language? Plan for **redundancy and performance**.	Yimei Yang, Yueha Wang	**Apr 21**
Conclusion	**Showcase the dashboards** you produced in Metabase. Paste screenshots and explain wha presented. What are the **benefits of these dashboards.** **Summarize your goals and how these were achieved**. Make sure to emphasize all **benefits of RDMS, ETL and analysis.** What **insights** were made possible due to you work, how did your client benefit from your RDMS implementation?	Yuhan Wang, Li Su	Apr 21
Preparation for Presentati	Transforming our project into PPT and present for the class.	Yimei Yang, Yueha Wang, Yuhan Wan Li Su	

Chapter 4 Database Design

4.1 Normalization Plan

The concepts for 1NF include each table cell should contain a single value and each record needs to be unique. We divided the loan database into customer-related table and account-related table, connect these tables with member_id, account_id, credit_id, status_id.
Then based on 1NF, we singled column primary key to build the 2NF, so no non-prime attribute is dependent on the proper subset of any candidate key of table.
We segment the customer-related tables in 1NF into personal information and credit-related information, then divided account-related tables into loan-related and payment-related tables, connected with member_id, account_id, credit_id, status_id, address_id, pymts_info_id.
The standard of 3NF is to has no transitive functional dependencies. Based on the outcome of 2NF, we move forward to be more specific. We divided the database into 16 tables: terms_info, statuses_info, loans_purpose, collections, applications, payments_info, pymts_schedule, secondary_applicants, interest_rates, addresses, inquires, customer_credit_histories, accounts_info, customers, bankcards, loans_info.

4.2 ETL Explanation

Our goal for design the ETL process is to help the company analyzing their business data much conveniently for taking critical business decisions. In addition, whenever new information is inserted,updated or deleted, the well designed ETL system will help the database to be automatically updated. It also allows sample data comparison between the source and the target system.
In the extraction step, data is extracted from the original dataset we have into the staging area. When we extract the data, validations are done in this stage included reconcile records with the source data, make sure that no spam/unwanted data loaded, check data type in each columns.
In the transaction step, data extracted from source server is raw and not usable in its original form. Therefore we cleaned, mapped and transformed the data. We have performed some customized operations on data such as create id columns in several tables. We have also checked several criterias of the raw data such as different account numbers are generated by various applications for the same customer. We also did filtering and set data threshold validation check such as select only certain columns to load and invalid credit grade can not be inserted.
In the load step, loading data into the database is the last step of the ETL process. We populated all the data into the accorded tables and formed the data warehouse.

4.3 Information Architecture: ERD (see attached PDF)

ERD in Lucidchart:
https://www.lucidchart.com/invitations/accept/146f7ec2-69c9-4e2b-b639-142631b7065d

4.4 R Codes Listings for Implementation

https://github.com/sqlfinal/loan/blob/master/sql%20final.md

4.5 Triggers Created in Loan Database

In our Loan database two triggers are created, and will be provoked whenever an event associated with the table occurs in order to enforce business rules of Lending Club:

Trigger Name	Description	SQL Implementation

customers_audits	This trigger ensures that any insert, update or delete of a row in the customers table is recorded in the customers_audits table, so a user accidently insert or update a wrong information into the customers table, we could check it in the customers_audits table and revise it accordingly.	CREATE TABLE customers_audits(
		change_id int primary key,
		member_id int,
		emp_title varchar(100),
		emp_length varchar(100),
		home_ownership varchar(100),
		annual_inc float,
		verification_status varchar(100),
		debt_settlement_flag varchar(10),
		grade varchar(10),
		sub_grade varchar(10),
		address_id int,
		loan_id int,
		bankcard_id int,
		account_info_id int,
		credit_id int,
		updated_at date NOT NULL,
		operation CHAR(3) NOT NULL,
		CHECK(operation = 'INS' or operation='DEL' or operation='UPD')
		);
		
		CREATE OR REPLACE FUNCTION process_customers_audit() RETURNS TRIGGER AS $customers_audits$
		BEGIN
		IF (TG_OP = 'DELETE') THEN
		INSERT INTO customers_audits SELECT 'DEL', now(), user, OLD.*;
		RETURN OLD;
		ELSIF (TG_OP = 'UPDATE') THEN
		INSERT INTO customers_audits SELECT 'UPD', now(), user, NEW.*;
		RETURN NEW;
		ELSIF (TG_OP = 'INSERT') THEN
		INSERT INTO customers_audits SELECT 'INS', now(), user, NEW.*;
		RETURN NEW;
		END IF;
		RETURN NULL;

		END; $customers_audits$ LANGUAGE plpgsql; CREATE TRIGGER customers_audits AFTER INSERT OR UPDATE OR DELETE ON customers FOR EACH ROW EXECUTE PROCEDURE process_customers_audit();

add_customers	This trigger ensure that any new customer we insert has a credit grade that fall into A,B,C,D,E,F,G which are the only classes that the bank has.	CREATE OR REPLACE FUNCTION add_customers()
		RETURNS TRIGGER AS
		$body$
		BEGIN
		IF (SELECT grade
		FROM customers
		WHERE customers.grade=NEW.grade)
		+NEW.grade NOT IN ('A','B','C','D','E','F','G')
		THEN RAISE EXCEPTION 'the customer grade has error';
		END IF;
		RETURN NEW;
		END;
		$body$
		LANGUAGE plpgsql;
		CREATE TRIGGER add_customers
		BEFORE INSERT OR UPDATE ON customers
		FOR EACH ROW EXECUTE PROCEDURE add_customers()

Chapter 5 User Interaction

5.1 Users

There are multiple users for this system defined by their role:

- For the data analyst, who will also play the role of a database administrator that manages, backs up and ensures the availability of the data produced and consumed by the organization by installing and upgrading the database server, and insert and update records in the database, will write SQL queries to retrieve meaningful information from the database system and generate data analysis reports.
- For the Managers/C- executives who are non-technical personnel responsible for decision making, data reports provided by the data analyst will be used to facilitate decision making process in terms of optimizing business strategy, identify new business opportunities and potential risks, etc.

5.2 Tools

Loan Database uses a Postgresql database back-end to store the data, and is connected with Metabase to show dashboards of insights, which are achieved by SQL queries.

5.2.1 Tools for Data Analyst

Loan RMDB built on PostgreSQL can be interacted by directly executing SQL queries in Pgadmin, the most popular and feature rich Open Source administration and development platform for PostgreSQL. Alternatively, the database can be accessed by programming tools such as Python or R. Data analysts of Lending Club, who are more familiar with programming languages, can use APIs to connect with the Loan Database server, send SQL commands to the database server, and fetch results into program variables. Programming languages are very efficient and powerful tools to handle data retrieval and entry, and support rich analysis packages. Our team use R to create and populate the Loan database.
Besides, data analysts will also use Metabase, an open source database visualization platform that is easily connected to Loan Database, to modify the interactive dashboards supported by SQL queries, and to generate analytics reports.

5.2.2 Tools for Managers/C- executives

Managers/C- executives will access the pre-modified Metabase to get direct insights that are displayed in user-friendly interactive charts and graphs without going through complex queries and programming steps. All interactions will be done in a few clicks.

5.3 Plan for Database Redundancy and Performance

There might be some redundancy and performance requirements for our database system, which is why we may need to consider a proper enterprise resource planning (ERP) system. There are two ways to consider this, one is on-premise, the other is in the cloud. Although cloud-based ERP systems are more common recently, there are still several reasons why a small or midsize business might choose a traditional on-premise system.

	Cloud ERP	On-Premise ERP
Cost	Predictable costs over time Cheaper upfront investment No additional hardware investments (e.g., server infrastructure) May end up spending more money over the course of the system's life cycle	Reduce initial price of system Upfront investment can be seen as riskier Have to pay associated hardware and IT costs
Security	Data security is in the hands of the vendor While vendors pledge strict data security standards, some organizations might not have total peace of mind with this arrangement.	Data security is in the hands of the organization Some organizations might not be as adept at practicing proper data security protocols
Customization	Offer greater stability and continuous updates from vendor as a result of less customization Organizations can work with vendors to see what changes can be made Less customizable in general	Greater ability to customize Customizations can delay implementation time Customizations can result in headaches when vendor updates software
Implementation	Typically take less time to implement Shorter implementation times are largely a result of less customization	Organization has more control over the implementation process Implementation process can take significantly longer.

For cost consideration: Since our company needs to minimize expense and predict cost over time, we need to get cheaper upfront investment, thus Cloud ERP is more suitable.

For security consideration: On the one hand, it may sound like Cloud ERP is not as safe as On-Premise ERP. However, the data security is in the hands of the vendor. We could choose a better vendor with more strict rules to make sure data security. On the other hand, on-premise ERP may sound safer, but it has risk of data consistency. If the local system breaks down, it would be hard to abstract the data.

For customization consideration: Although in general. Cloud ERP is less customizable compared to on-premise, our database is used for analysts and C-level officers to get data insight, which means the customization is not highly needed.
For implementation consideration: Since we have a pretty large dataset, a cloud ERP can significantly cut off the running time and improve implementation efficiency.
Based on all above, we will choose to store the database into cloud system instead of local storage to maximize the database performance.

Chapter 6 Analytical Procedures

As mentioned before, the Loan Database System can provide various data insights, both for analysts and for C-level review.

In this chapter, we will offer three specific examples, as in current business analysis, customer analysis and risk analysis, which cover the most important business perspectives for Lending Club. Inside the questions under each perspective, we provide data analyst version (query) or Managers/C-executives (reports) based on the content of each question.

SQL code listings used to generate C-executive dashboards can be found at:
https://docs.google.com/document/d/1Gz7rGjQBAZvMW5jrRQQf_njdpQhlZvopt6AjeKkYwis/edit

6.1 Current Business Analysis

Current business analysis refers to operational procedures implemented on a day-to-day by data analysts of Lending Club. They will run queries to retrieve information to answer basic operation questions such as:

1. Is there a pattern for customers' credit history and funded loan amount? - For Data Analyst

In order to analyze answer this question, we provide three perspectives for analyst:

1) Is there a difference between fund amount and customers with different count of public record of bankruptcy?

Query:

```
select avg(loan_amnt) as avg_loan_amnt, pub_rec_bankruptcies
from loans_info as L, customer_credit_histories as A, customers as C
where L.member_id = C.member_id AND C.credit_id = A.credit_id
group by A.pub_rec_bankruptcies
order by A.pub_rec_bankruptcies
```

Output:

	avg_fund_amnt numeric	pub_rec_bankruptcies integer
1	15493.863019891501	0
2	13032.398897058824	1
3	11788.392857142857	2
4	10666.6666666666666667	3
5	10800.0000000000000000	4

Fig 6.1 Average Fund Amount in Each Public Record of Bankruptcy

2) Is there a difference among fund amount and customers who have higher late fee and who have lower?

Query:

```
select avg(funded_amnt) as avg_fund_amnt, total_rec_late_fee
from loans_info
group by total_rec_late_fee
order by total_rec_late_fee
```

Output:

	avg_fund_amnt numeric	total_rec_late_fee double precision
1	15194.402985074627	0
2	28000.000000000000	6.6e-09
3	7000.0000000000000000	2.3
4	9000.0000000000000000	7
5	1200.0000000000000000	14.9132948
6	2200.0000000000000000	14.93382961
7	1500.0000000000000000	14.97454607
8	10275.0000000000000000	14.99999994
9	7104.4444444444444444	15
10	2400.0000000000000000	15.00000001
11	4500.0000000000000000	15.00000004

Fig 6.2 Average Fund Amount with Late Fee

3) Is there a difference among fund amount and whether or not a customer has acc_now_delinq?

Query3:

```
select avg(funded_amnt) as avg_fund_amnt, acc_now_delinq
from loans_info as L, accounts_info as A, customers as C
where L.member_id = C.member_id AND C.account_info_id = A.account_info_id
group by acc_now_delinq
order by acc_now_delinq
```

Output 3:

	avg_fund_amnt numeric	acc_now_delinq integer
1	15204.512635379061	0
2	13680.769230769231	1
3	19800.0000000000000000	2

Fig 6.3 Average Fund Amount in Delinq Account

Summary:

- It is not likely that people with less acc_now_delinq to receive more average fund amount.
- It is not likely that people with less late fee to receive more average fund amount.
- People with more public record bankruptcies tend to receive less average fund amount.

2. Which current loans have high risk to go bad? - For Data Analyst

We have set three conditions as a boundary to consider a loan as high risk to become bad:

a.status: 'Current'

b.pub_rec_bankruptcies>0

c.delinq_2yrs >0

Query:

```
select l.loan_id,c.member_id,issue_d,last_credit_pull_d,emp_title,
emp_length,home_ownership,annual_inc,verification_status,sub_grade,delinq_2yrs,pub_rec_bankruptcies,acc_now_delinq,ad
dr_state,zip_code
from loans_info l
join customers c on l.member_id=c.member_id
join customer_credit_histories cr on c.credit_id=cr.credit_id
join accounts_info a on c.account_info_id=a.account_info_id
join addresses ad on c.address_id=ad.address_id
join statuses_info s on l.status_id=s.status_id
where s.loan_status = 'Current'
and pub_rec_bankruptcies>0
and delinq_2yrs >0
order by sub_grade;
```

Output:

	loan_id integer	member_id integer	issue_d	last_credit_ character v	emp_title character varying	emp_length character var	home_owner character var	annual_inc double preci	verification_status character varying (	sub_grad characte	delinq_2yrs integer	pub_rec_bankru integer	acc_now_deli integer	addr_state character va	zip_code character varying (2
1	2066	2066	Dec-2...	Feb-2019	Registered Nur...	9 years	MORTGAGE	80000	Verified	A4	1	1	0	LA	704xx
2	2330	2330	Oct-20...	Feb-2019	MACHINE OPE...	10+ years	OWN	35360	Verified	A5	1	1	0	FL	322xx
3	2140	2140	Jun-20...	Feb-2019	Title clerk	6 years	MORTGAGE	70000	Source Verified	A5	1	1	0	MD	208xx
4	2517	2517	Aug-2...	Feb-2019	leasing agent	7 years	OWN	160000	Verified	B2	1	1	0	FL	331xx
5	1946	1946	Sep-2...	Feb-2019	Athletic Trainer	10+ years	RENT	67000	Source Verified	B3	1	1	0	CA	922xx
6	1965	1965	Jun-20...	Feb-2019	Technician	1 year	MORTGAGE	78000	Source Verified	B3	9	1	0	GA	316xx
7	1840	1840	Nov-2...	Feb-2019		n/a	OWN	50000	Source Verified	B4	1	1	0	MO	633xx
8	834	834	Aug-2...	Feb-2019		n/a	MORTGAGE	54700	Verified	B5	1	1	0	SC	295xx
9	1415	1415	Sep-2...	Feb-2019	Machine adjust...	10+ years	MORTGAGE	46800	Source Verified	B5	1	1	0	GA	301xx
10	707	707	May-2...	Feb-2019	LPN	1 year	RENT	35000	Verified	C1	1	1	0	PA	161xx
11	892	892	Jul-20...	Feb-2019	C	< 1 year	MORTGAGE	45248	Source Verified	C1	1	1	0	OH	442xx
12	1676	1676	Dec-2...	Feb-2019	Dean of Studen...	2 years	MORTGAGE	95000	Source Verified	C1	1	2	0	CA	933xx
13	1767	1767	May-2...	Feb-2019	Dental Hygienist	7 years	MORTGAGE	45495	Not Verified	C1	1	1	0	WI	532xx
14	1808	1808	Nov-2...	Feb-2019	sales	2 years	RENT	66000	Not Verified	C1	3	1	0	IL	606xx
15	2152	2152	Nov-2...	Feb-2019	Locomotive En...	10+ years	OWN	80000	Not Verified	C4	1	1	0	VA	242xx
16	1862	1862	Jun-20...	Feb-2019	tandem extrusi...	10+ years	MORTGAGE	92000	Source Verified	C4	3	1	0	WI	539xx
17	2326	2326	Dec-2...	Feb-2019	Court Assistant	2 years	RENT	55596	Source Verified	C4	1	1	0	NY	130xx
18	2087	2087	Oct-20...	Feb-2019	Mammographer	10+ years	RENT	55000	Source Verified	D2	1	1	0	MD	211xx
19	2482	2482	Oct-20...	Feb-2019	Assistant Prop...	1 year	MORTGAGE	46000	Source Verified	D2	1	1	0	NV	895xx
20	2055	2055	Jun-20...	Feb-2019	O.R Sterile Tech.	10+ years	OWN	65000	Not Verified	D3	1	3	0	IL	600xx
21	1817	1817	Sep-2...	Feb-2019	Owner	10+ years	MORTGAGE	82700	Verified	D3	1	1	0	TN	377xx
22	2056	2056	Sep-2...	Feb-2019	Dental hygienist	3 years	MORTGAGE	90000	Source Verified	D5	1	1	0	TX	751xx
23	1632	1632	May-2...	Feb-2019	Service Advisor	< 1 year	MORTGAGE	70000	Source Verified	E3	1	1	0	MD	217xx
24	1521	1521	Dec-2...	Feb-2019	Supervisor Co...	6 years	MORTGAGE	70000	Source Verified	E3	1	1	0	ID	838xx
25	1506	1506	Nov-2...	Feb-2019	Rn	4 years	RENT	70000	Verified	E4	1	1	0	TX	781xx

Fig 6.4 Account with High Potential Risks

Summary:

The output shows there are 25 current loans are at high risks of becoming bad. In addition, in this segment, more people tend to have a credit subgrade with C or less and only few people with a credit subgrade of A. Several people with more than 1 time public record bankruptcies and over 3 to 9 times records on delinq_2yrs. So we suggest that our client to put more attentions on people have similar characteristics with this segment since they are at high risks of becoming bad loan.

3. **What is the average loan amount for people with different credit grade? - For Data Analyst**

Query:

select avg(loan_amnt) as avg_loan_amnt, avg(funded_amnt) as avg_fund_amnt, grade, sub_grade
from loans_info as L, customers as C
where L.member_id = C.member_id
group by C.grade, C.sub_grade
order by sub_grade;

Output:

	avg_loan_amnt numeric	avg_fund_amnt numeric	grade character varying (10)	sub_grade character varying (10)
1	4885.121951219512	4885.121951219512	A	A1
2	3943.750000000000	3930.681818181818	A	A2
3	4417.187500000000	4386.093750000000	A	A3
4	5459.137055837563	5459.137055837563	A	A4
5	4759.770114942529	4759.770114942529	A	A5
6	3835.887096774194	3823.185483870968	B	B1
7	5537.068965517241	5537.068965517241	B	B2
8	4597.373188405797	4597.373188405797	B	B3
9	4732.866043613707	4732.866043613707	B	B4
10	3886.419753086420	3886.419753086420	B	B5
11	4973.824451410658	4973.824451410658	C	C1

Fig 6.5 Average Fund Amount in Credit Grade

6.2 Customer Analysis

1. **Could we consider people with better credit as our main customers? - For Managers/C-executives**

Metabase Dashboard:

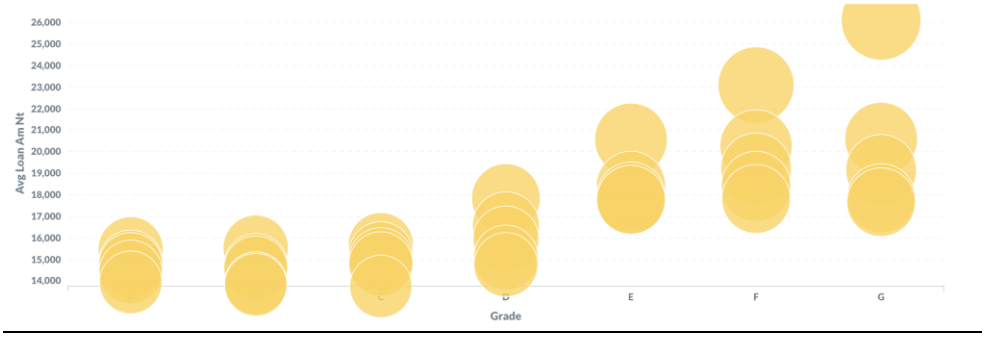


Fig 6.6 Average Fund Amount in Credit Grade

Summary:

From Fig 6.6 we can see that people with better credit grade do not tend to receive more average fund amount, which means they are not our main customers. The reason could be that:

- People with lower credit may need more money, which means they have higher needs.
- People with better credit intend to maintain their credit history, which makes them not to borrow too much money to avoid the pressure.

2. When should we suggest customers to apply for individual application or for joint application? - For Managers/C- executives

Considering that personal status may vary because of the region distinct, we should provide the recommendation based on the geographic information.

- First, make sure that there are difference among individual application and joint application, then we could get the benchmark for each region.
- Second, when we are trying to find a benchmark for recommendation of application type, we may need to consider the geographic condition. Since people in difference place may have difference benchmark of income and lifestyle, which may also lead to varieties of satisfactory account number.

Metabase Dashboard:

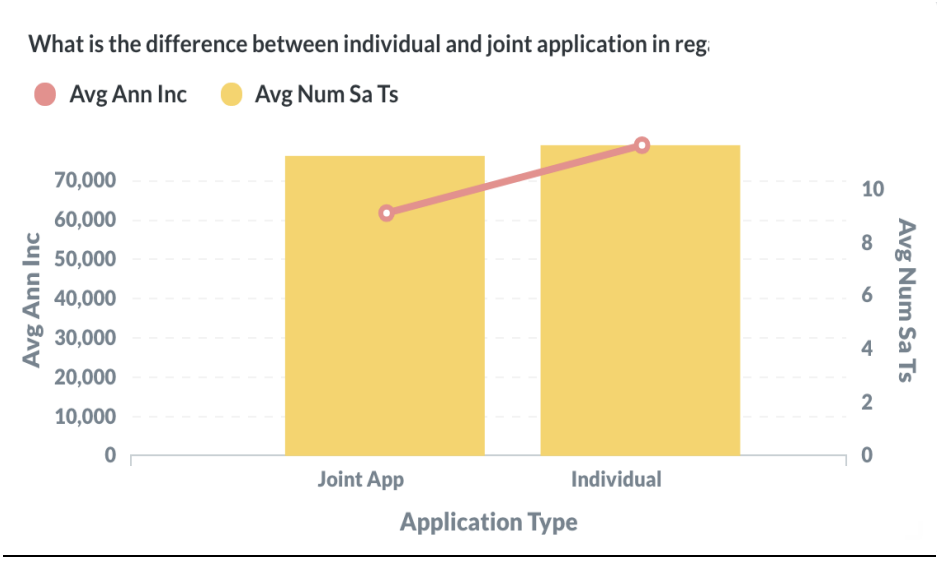


Fig 6.7 Application Type for Average Loan Amount

Summary:

Generally speaking, people with lower average annual income and lower number of satisfactory accounts should consider joint application instead of individual application. This aligns with reality since people with lower income may need more evidence to support their loans. Thus, the joint application is a more suitable application type.

Here we can see the average annual income and average number of satisfactory accounts for customers we have. Based on this we can get our benchmark for recommendation of application type for each state. Also here we can check our insight we get before: the lower average annual income and average satisfactory accounts, the more suitable for joint application type.

3. Where does Lending Club's best market locate? - For Managers/C- executives

Metabase Dashboard:

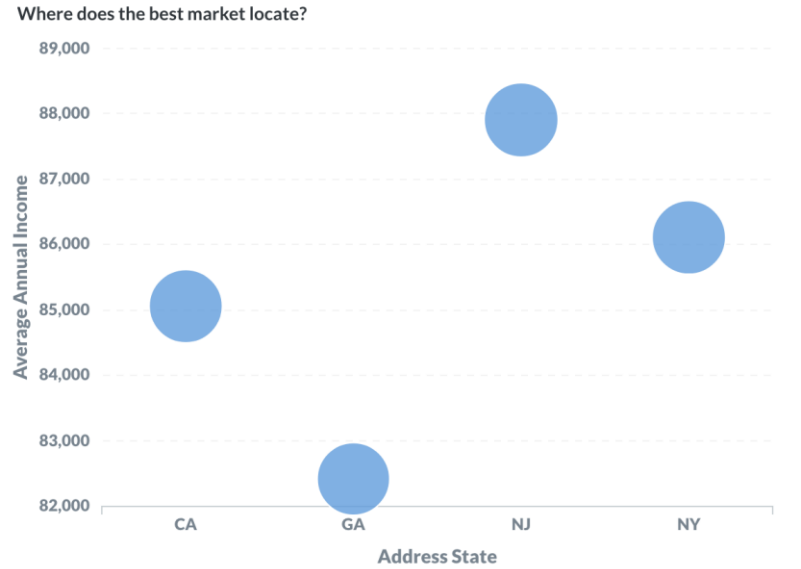


Fig 6.8 Best Market Location

Summary:

When considering the best market, we choose the top 10 places of amount of customers, top 10 places of average annual income and top 10 places where people have 'OWNER' type of home ownership. We put them together and find out that these four states show up in each terms. Meanwhile, after we finish the risk analysis, we can see that these four states are not in the top 10 'highest bad/good loan ratio' list. Thus, we consider that these four states have more potential customers with good credit and personal financial statement. In the future, when we launch new products, we may consider these four market places.

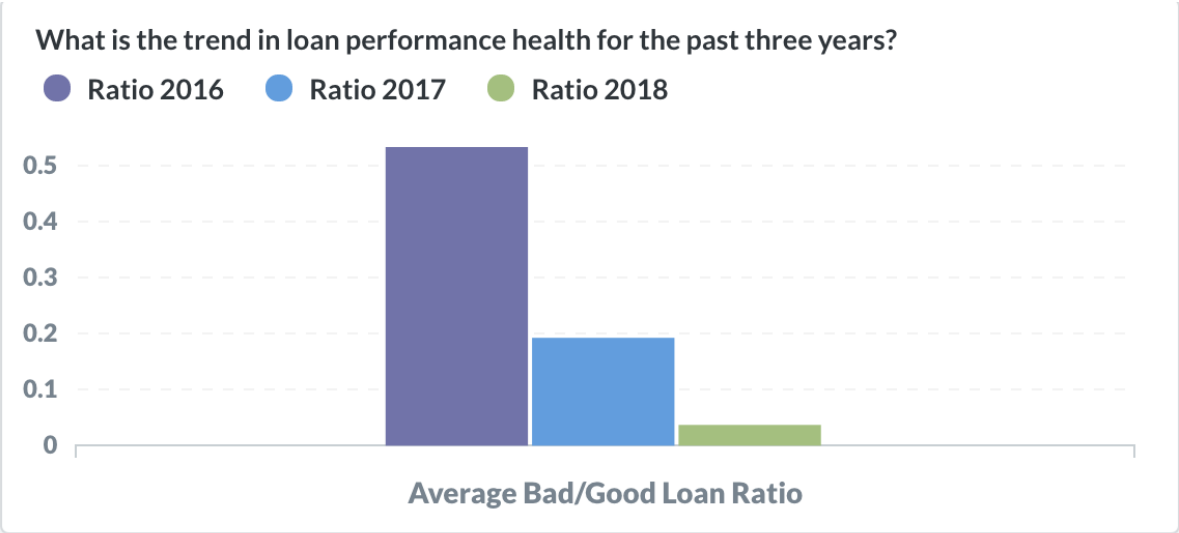
6.3 Risk Analysis

1. What is the trend in loan performance health for the past three years? - For Managers/C-executives

What determines a bad loan? We defined a criteria of delinquent or bad loans - loans in status among "Late Loans: Loans that are not paid within the expected time to the financial institution", "Charged Off Loans: Are loans that are unlikely to be collected by the financial institution or the credit company", "Default Loans: Are loans that are not going to be paid back", and "Grace Period: Payments to the financial institution that will be paid later on time but with no penalty to the client". Meanwhile the criteria of a good loan is that loans in status of "Current".

What is the metric for loan performance health? To facilitate risk assessment, a bad/good loan ratio will be used as the major metric to quantify loan performance health.

Metabase Dashboard:



What is the trend in loan performance health for the past three years?

Avg Ratio	Ratio 2016	Ratio 2017	Ratio 2018
0.17	0.53	0.19	0.036

Fig 6.9 Trend in Loan Performance Health in Past 3 Years

Summary:

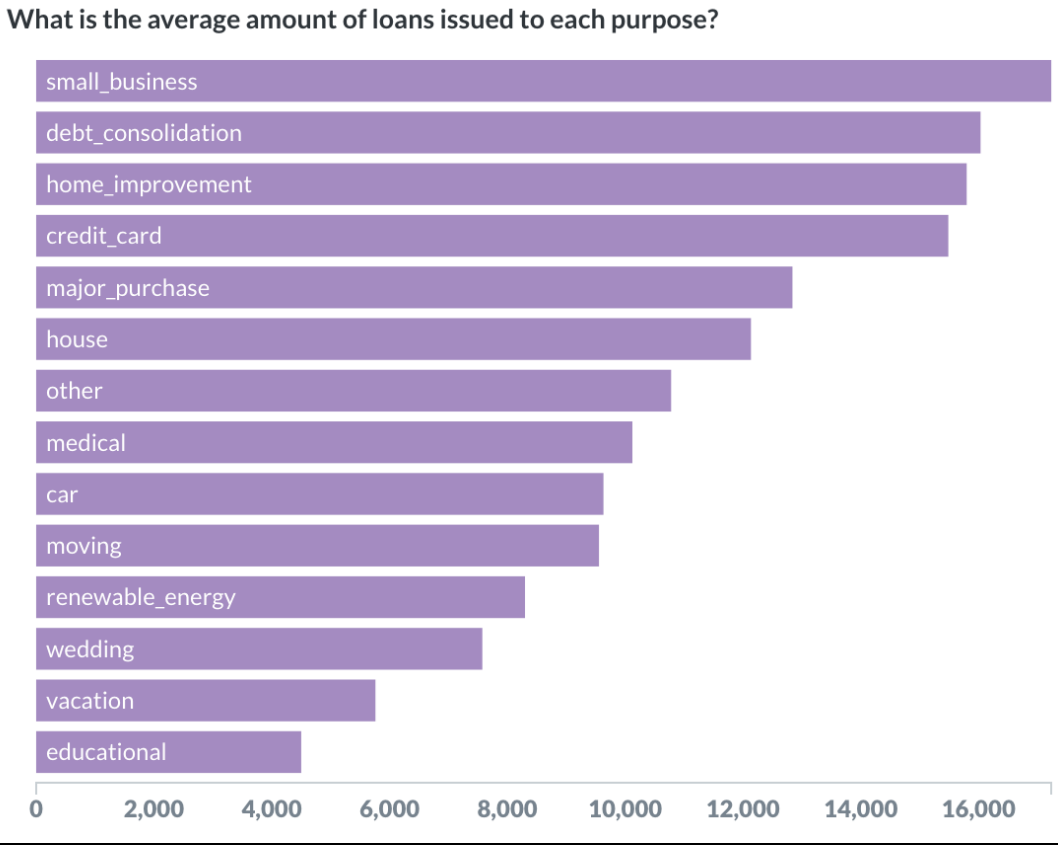
The bad/good loan ratio has declined from 0.5330 among loans issued in 2016 to 0.1914 among loans issued in 2017, then to 0.0357 among loans issued in 2018. This trend shows a positive tendency of loan performance health of Lending Club.

2. What loan purposes may result in poor loan performance health? - For Managers/C- executive

Our main aim is to see if there are purposes that contribute to a "higher" risk whether the loan will be repaid or not. In order to explore the relationship between loan purposes and loan performance health, we consider this into two perspectives:

- The average amount of loans issued to each purpose
- The bad/good loan ratio for each loan purpose

Metabase Dashboard:



ratio_small_bus numeric	ratio_debt_consl numeric	ratio_house numeric	ratio_home_imp numeric	ratio_credit_card numeric	ratio_med numeric	ratio_other numeric	ratio_car numeric	ratio_moving numeric	ratio_maj_purchase numeric
0.8779	0.3552	0.5072	0.3766	0.2219	0.1942	0.2872	0.6359	0.2983	0.3492

Fig 6.10 Average Amount of Loans in Each Purpose

Summary:

- From above result, the top three purposes that are applied with relatively higher loan amount are credit_card, major_purchase, and house while the purposes with least average loan amount are education and vacation.
- The 5 highest average loan amount are issued to purposes: “small_business”, ”debt_consolidation”,”home_improvement”,”credit_card” and “major_purchase”.
- For C-level insight: Loans with purpose for“small_business” has the biggest poor/good loan ratio, which is 0.88, follows by purpose “car” with a ratio of 0.64, and “house” 0.51. When approve loan applications, we need to pay special attention to these categories, and be more discreet in assessing the risks.

3. Which region has the highest bad/good loan ratio?- For Managers/C- executive

Metabase Dashboard:

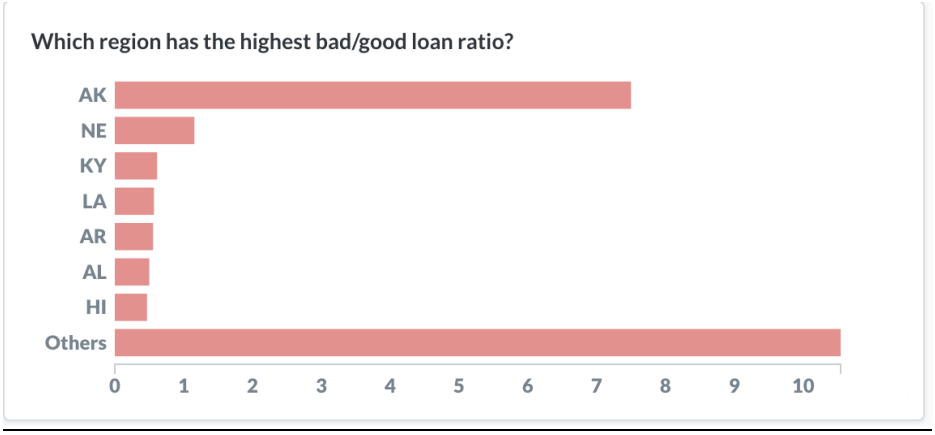


Fig 6.11 Loan Performance in Each Region

Summary:

AK, NE, KY, LA, AR, AL, HI, NC, OH, TX are the top 10 states with highest bad/good loan ratio. When issue loans to members from these states, company should be more discreet with members' risk evaluation.

Chapter 7 Conclusion

7.1 Goals Achieved

Our final goal is to establish a structured relational database that provides data insight to help make better business decisions. In order to achieve our final goal, we used several steps:

1. By building Loan Relational Database, we managed to reduce Lending Club's reduncent data by reorganizing data into dependent tables with a normalization plan.
2. We created a consistent and accurate system that supports very easy and simple SQL queries.
3. Triggers were stored in Loan Database to enforce Lending Club's business rules.
4. The Loan RDMS is connected with Metabase to generate interactive dashboards for managers/ C-executives.

7.2 Benefits of RDMS

Simplicity

The original dataset contains 97 attributes and will lead to inefficiency in terms of data retrieval and analysis. By transforming data into a RDMS, the simple structured tables were created that are intuitively familiar to most users, will benefits both analysts and management.

Ease of Data Retrieval

Loan Database users can retrieve data through simple structured language and can join independent tables by easy syntax. More importantly, analysts can create multiple views to improve retrieval efficiency and grant different access to different database users.

Data Integrity

The predefined data types and triggers ensure new data inserted fall within acceptable ranges for Loan RDMS. Referential integrity prevents records from becoming incomplete or orphaned. This feature helps to ensure accuracy and consistency of the data.

Flexibility

The Loan RDMS is scalable and extensible, providing a flexible structure to meet changing requirements and increasing amounts of user data of Lending Club.

Normalization

The normalization process we took provided a set of rules, qualities and objectives for the design and review of a database structure. Specifically, we designed a 3NF database and separated the original dataset into 17 structured tables that ensures the system to be robust and dependable.

7.3 Benefit of ETL

As introduced in chapter 4.2, ETL processes the heterogeneous data and makes it homogeneous, and in turn makes it seamless for analysts to run analysis and derive business intelligence. ETL, compared with traditional methods which always involve program writing, is much easier and faster to apply.
For Loan Database, the original data is highly unstructured and inefficient to analyze. ETL helps to retrieve data from wide source, transform to organized types and load into target database.

7.4 Benefit of Analysis: Business Insights for Clients

From the analytics procedures we reach several conclusions that will benefit Lending Club:
1. Late fee is not correlated with average fund amount the customers apply, however, people with more public record bankruptcies tend to receive less average fund amount.
2. When compare current loan to determine which may have higher risk to bad, we found it efficient to compare it by credit grade/subgrade. When people are assigned to a lower credit subgrade, their loan have a relatively higher chance to become bad than others. In addition, most members with bad loans have record of bankruptcies and delinquent of 2 years. So we suggest Lending Club to put more attention on customers in these categories.
3. We suggest that people with lower average annual income and lower number of satisfactory accounts consider a joint application rather than individual application.
4. The top 3 purposes for highest average amount of loan are small business, debt consolidation and home improvement. However, loans with purpose of "small_business" have the biggest poor/good loan ratio, which is 0.88, follows by purpose "car" of 0.64, and "house" 0.51. When approve loan applications, Lending Club need to pay special attention to these categories, and be more discreet in assessing the risks.
5. Overall, four states consist top four markets of Lending Club: CA, GA, NJ and NY.
AK, NE, KY, LA, AR, AL, HI, NC, OH, TX are the top 10 states with highest bad/good ratio. When issue loans to members from these states, Lending Club should be more discreet.
6. Finally, there is a positive tendency of loan performance health of Lending Club which shows that bad/good loan ratio has declined from 2016 to 2018.

7.5 Benefits for Users of RDMS:

As mentioned in 1.3, Loan RDMS and Metabase serve for two groups of users:
Analysts: The Loan RDMS allows analysts to write and execute SQL code to run day-to-day operational analysis and generate report for management.
Managers/C- executives: Metabase, connected with Loan RDMS, will provide visualized insights and interactive dashboards that automatically update when new data is inserted to database, making it straightforward for non-tech executives.
Overall, Loan RDMS helps Lending Club achieve a more smooth operation process and facilitate business decision-making.