9. Bank Loan Defaulter Prediction

Writing a Data Description Report

Data Quantity

- What is the format of the data?

  The total 3 of the given files have a common file format of Comma-separated values(.CSV)

FILE NAME	FILE FORMAT
train	.csv(Comma-separated values)
test	.csv(Comma-separated values)
submission	.csv(Comma-separated values)

- Identify the method used to capture the data--for example, ODBC.

  We can't identify how the data is collected but on visualizing the data we can tell the data collection should have been taken by the banks itself from all the data they have gathered from giving out all the previous loans to the borrowers.

- How large is the database (in numbers of rows and columns)?

FILE NAME	NO. OF ROWS	NO. OF COLUMNS
train	67463	35
test	28913	34
submission	28913	1

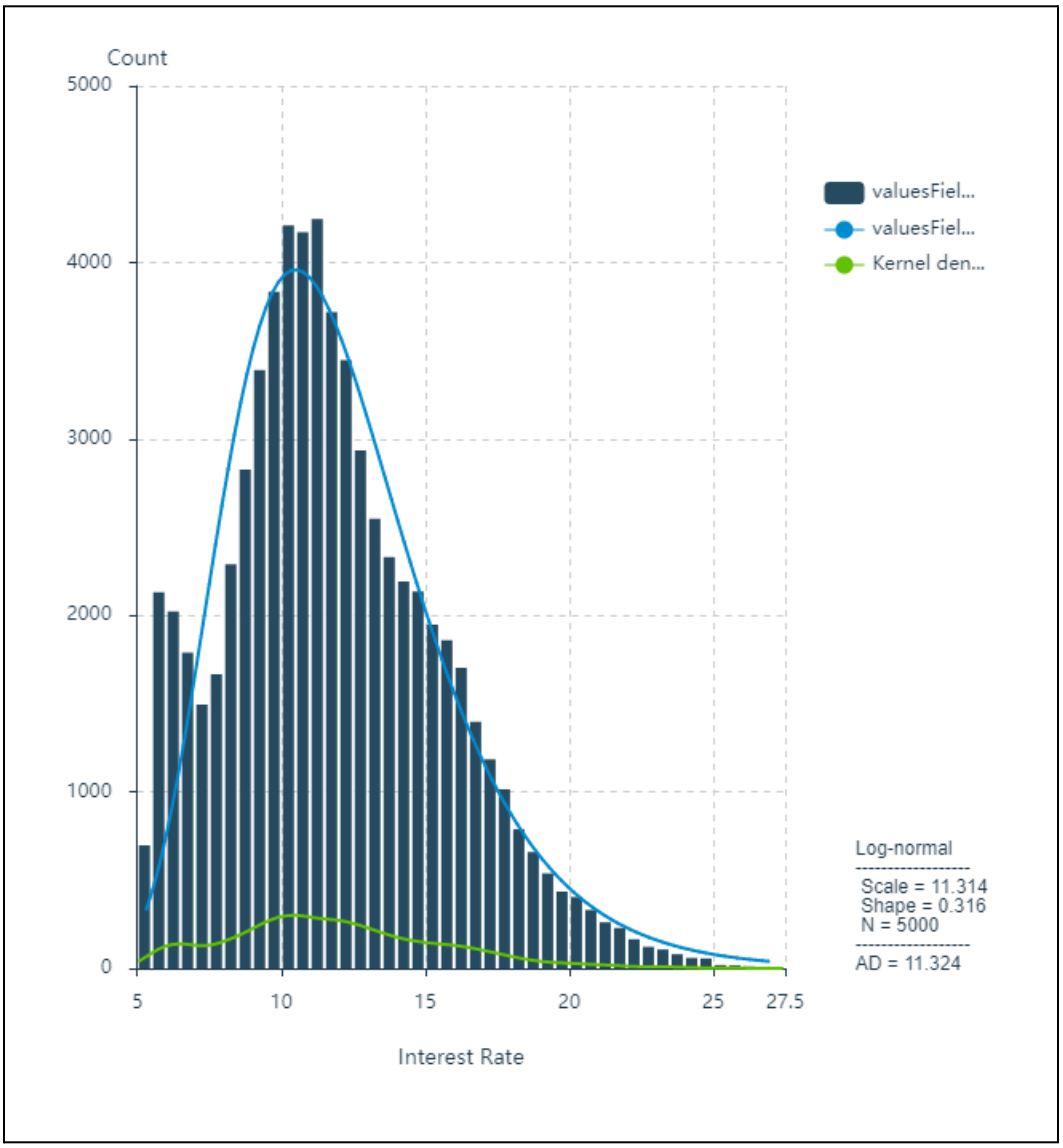
Data Quality

- Does the data include characteristics relevant to the business question?

  Accuracy

  We can check accuracy by checking given data with real life scenario:

  ➢ Interest Rates Range given in dataset:



➢ Current Interest Rate Range in Real Life by banks:

Bank	Interest Rate(p.a.)
HDFC Bank	10.5% p.a. - 21.00% p.a.
ICICI Bank	10.75% p.a. - 19.00% p.a.
Yes Bank	10% p.a. onwards - 24% p.a.
Citibank	10.50% p.a. - 16.49% p.a.
Kotak Mahindra Bank	10.99% and above
Axis Bank	12% p.a.- 21% p.a.
IndusInd Bank	10.49% p.a. - 26.5% p.a.
Home Credit Cash Loan	24% p.a. - 49.5% p.a.
Aditya Birla Capital	14% p.a. -26% p.a.
State Bank of India	10.65% p.a. - 13.65% p.a.
Federal Bank	10.49% p.a. - 17.99% p.a.
IIFL	11.75% p.a. - 34% p.a.

Completeness

- **id**: Unique ID of the loan application.
- **grade**: LC assigned loan grade.
- **annual_inc**: The self-reported annual income provided by the borrower during registration.
- **loan_amount**: Total loan amount given to the borrower.
- **interest_rate**: Interest rate at which a particular loan is given.
- **home_ownership**: Type of home ownership.
- **dti (Debt-To-Income Ratio)**: A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage

and the requested LC loan, divided by the borrower's self-reported monthly income.

- **loan_title**: A category provided by the borrower for the loan request.
- **term**: The number of payments on the loan. Values are in months and can be either 36 or 60.
- **delinquency**: 1 when the borrower had at least one event of delinquency in two years.
- **revolving_utilities**: Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
- **total_received_late_fee**: Late fees received to date.
- **total_current_balance**: Total balance in the bank of a particular borrower.
- **recoveries**: The recovery rate is the estimated percent of a loan or an obligation that will still be repaid to creditors in the event of a default or bankruptcy.

Reliability

In the realm of data quality characteristics, reliability means that a piece of information doesn't contradict another piece of information in a different source or system.

The data provided to us doesn't have contradicted pieces of information.

Relevance

ATTRIBUTES	DESCRIPTION
ID	Unique ID of the loan application.
Loan Amount	Total loan amount given to the borrower.

Funded Amount	The aggregate amount of Purchase Prices paid by the Banks
Funded Amount Investor	Aggregate amount
Term	Period of time for which loan is taken
Batch Enrolled	Loan enrolled in a batch
Interest Rate	The amount charged over and above the principal amount by the lender from the borrower
Grade	A particular level of rank, quality, proficiency, or value
Sub Grade	Subset for grades
Employment Duration	Home ownership type
Home Ownership	Amount for home the borrower is paying
Verification Status	Status for verification of a loan
Payment Plan	Plan for repayment of loan
Loan Title	A category provided by the borrower for the loan request.
Debt to Income	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income
Delinquency - two years	1 when the borrower had at least one event of delinquency in two years
Inquires - six months	If borrower has looked into getting any loans in last six months
Open Account	Total accounts opened for a particular borrower
Public Record	Documents or pieces of information that are not considered confidential
Revolving Balance	A line of credit they can keep using and repaying over and over
Revolving Utilities	How much of credit balance has been used
Total Accounts	Total no. of accounts
Initial List Status	A request by a Borrower to borrow money on the terms of a Loan
Total Received Interest	Principal loan amount x interest rate x loan term
Total Received Late Fee	A late fee is a charge imposed on a consumer who fails to make the payment on a debt or other financial obligation by the due date

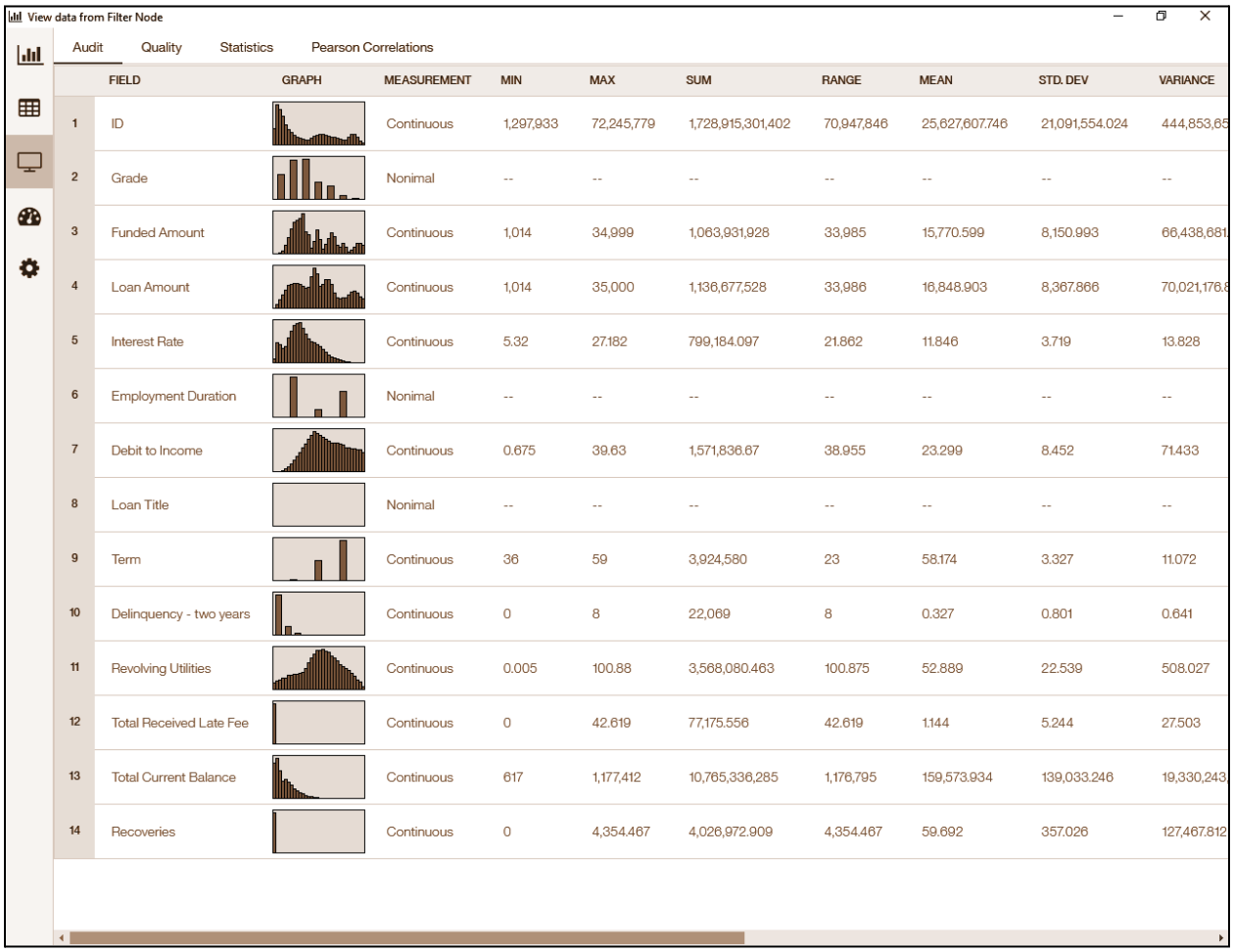
Recoveries	The recovery rate is the estimated percent of a loan or an obligation that will still be repaid to creditors in the event of a default or bankruptcy
Collection Recovery Fee	Covers costs incurred to collect tax and fee liabilities that are unpaid for more than 90 days
Collection 12 months Medical	Paying the debt in an agreed-on timeframe, no interest will be charged
Application Type	Whether a borrower is identified as individual or asking for loan in joint status
Last week Pay	Payment received in last week
Accounts Delinquent	A delinquency is when your account is late by 30 days or more.
Total Collection Amount	Total amount collected from a particular borrower.
Total Current Balance	Total current balance in a borrower's account.
Total Revolving Credit Limit	A line of credit that remains available over time, even if you pay the full balance.
Loan Status	Indicates where your loan is in the process.

- What data types are present (symbolic, numeric, etc.)?

ATTRIBUTES	DATA TYPES
ID	Integer
Loan Amount	Integer
Funded Amount	Integer
Funded Amount Investor	Real
Term	Integer
Batch Enrolled	String
Interest Rate	Real
Grade	String
Sub Grade	String
Employment Duration	String
Home Ownership	Real
Verification Status	String
Payment Plan	String

Loan Title	String
Debt to Income	Real
Delinquency - two years	Integer
Inquires - six months	Integer
Open Account	Integer
Public Record	Integer
Revolving Balance	Integer
Revolving Utilities	Real
Total Accounts	Integer
Initial List Status	String
Total Received Interest	Real
Total Received Late Fee	Real
Recoveries	Real
Collection Recovery Fee	Real
Collection 12 months Medical	Integer
Application Type	String
Last week Pay	Integer
Accounts Delinquent	Integer
Total Collection Amount	Integer
Total Current Balance	Integer
Total Revolving Credit Limit	Integer
Loan Status	Integer

- Did you compute basic statistics for the key attributes? What insight did this provide into the business question?



	FIELD	GRAPH	MEASUREMENT	MIN	MAX	SUM	RANGE	MEAN	STD. DEV	VARIANCE
1	ID		Continuous	1,297,933	72,245,779	1,728,915,301,402	70,947,846	25,627,607.746	21,091,554.024	444,853,65
2	Grade		Nonimal	--	--	--	--	--	--	--
3	Funded Amount		Continuous	1,014	34,999	1,063,931,928	33,985	15,770.599	8,150.993	66,438,681
4	Loan Amount		Continuous	1,014	35,000	1,136,677,528	33,986	16,848.903	8,367.866	70,021,176.8
5	Interest Rate		Continuous	5.32	27.182	799,184.097	21.862	11.846	3.719	13.828
6	Employment Duration		Nonimal	--	--	--	--	--	--	--
7	Debit to Income		Continuous	0.675	39.63	1,571,836.67	38.955	23.299	8.452	71.433
8	Loan Title		Nonimal	--	--	--	--	--	--	--
9	Term		Continuous	36	59	3,924,580	23	58.174	3.327	11.072
10	Delinquency - two years		Continuous	0	8	22,069	8	0.327	0.801	0.641
11	Revolving Utilities		Continuous	0.005	100.88	3,568,080.463	100.875	52.889	22.539	508.027
12	Total Received Late Fee		Continuous	0	42.619	77,175.556	42.619	1.144	5.244	27.503
13	Total Current Balance		Continuous	617	1,177,412	10,765,336,285	1,176,795	159,573.934	139,033.246	19,330,243
14	Recoveries		Continuous	0	4,354.467	4,026,972.909	4,354.467	59.692	357.026	127,467.812

	MEAN	STD. DEV	VARIANCE	MEAN STD. ERR.	SKEWNESS	SKEWNESS STD. ERR.	KURTOSIS	KURTOSIS STD. ERR.	UNIQUE	VALID
346	25,627,607.746	21,091,554.024	444,853,651,136,038.06	81,203.688	0.557	0.009	-1.09	0.019	--	67,463
	--	--	--	--	--	--	--	--	7	67,463
	15,770.599	8,150.993	66,438,681.38	31.382	0.673	0.009	-0.617	0.019	--	67,463
	16,848.903	8,367.866	70,021,176.815	32.217	0.288	0.009	-0.798	0.019	--	67,463
	11.846	3.719	13.828	0.014	0.563	0.009	0.149	0.019	--	67,463
	--	--	--	--	--	--	--	--	3	67,463
	23.299	8.452	71.433	0.033	0.081	0.009	-0.905	0.019	--	67,463
	--	--	--	--	--	--	--	--	109	67,463
	58.174	3.327	11.072	0.013	-6.381	0.009	39.593	0.019	--	67,463
	0.327	0.801	0.641	0.003	4.635	0.009	30.673	0.019	--	67,463
	52.889	22.539	508.027	0.087	-0.237	0.009	-0.545	0.019	--	67,463
	1.144	5.244	27.503	0.02	5.084	0.009	25.99	0.019	--	67,463
5	159,573.934	139,033.246	19,330,243,372.283	535.286	1.512	0.009	3.125	0.019	--	67,463
67	59.692	357.026	127,467.812	1.375	7.371	0.009	58.177	0.019	--	67,463

- Are you able to prioritize relevant attributes? If not, are business analysts available to provide further insight?

id: Unique ID of the loan application.
grade: LC assigned loan grade.
annual_inc: The self-reported annual income provided by the borrower during registration.
loan_amount: Total loan amount given to the borrower.

interest_rate: Interest rate at which a particular loan is given.
home_ownership: Type of home ownership.
dti (Debt-To-Income Ratio): A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
loan_title: A category provided by the borrower for the loan request.
term: The number of payments on the loan. Values are in months and can be either 36 or 60.
delinquency: 1 when the borrower had at least one event of delinquency in two years.
revolving_utilities: Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
total_received_late_fee: Late fees received to date.
total_current_balance: Total balance in the bank of a particular borrower.
recoveries: The recovery rate is the estimated percent of a loan or an obligation that will still be repaid to creditors in the event of a default or bankruptcy.

Data Exploration Report

- What sort of hypotheses have you formed about the data?

1. The people who are not verified are more likely to default on their loan amounts.
2. The people who have high debt income are more likely to default on their loan amounts.
3. Loan amount is a significant predictor of loan default. Customers with higher loan amounts are more likely to default on their loans compared to those with lower loan amounts.
4. Loan term is a significant predictor of loan default. Customers with longer loan terms are more likely to default on their loans compared to those with shorter loan terms.
5. Employment status is a significant predictor of loan default. Customers who are unemployed or self-employed are more likely to default on their loans compared to those who are employed or in other employment status categories.

- Which attributes seem promising for further analysis?

id: Unique ID of the loan application.
loan_amount: Total loan amount given to the borrower.
interest_rate: Interest rate at which a particular loan is given.
home_ownership: Type of home ownership.
dti (Debt-To-Income Ratio): A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.

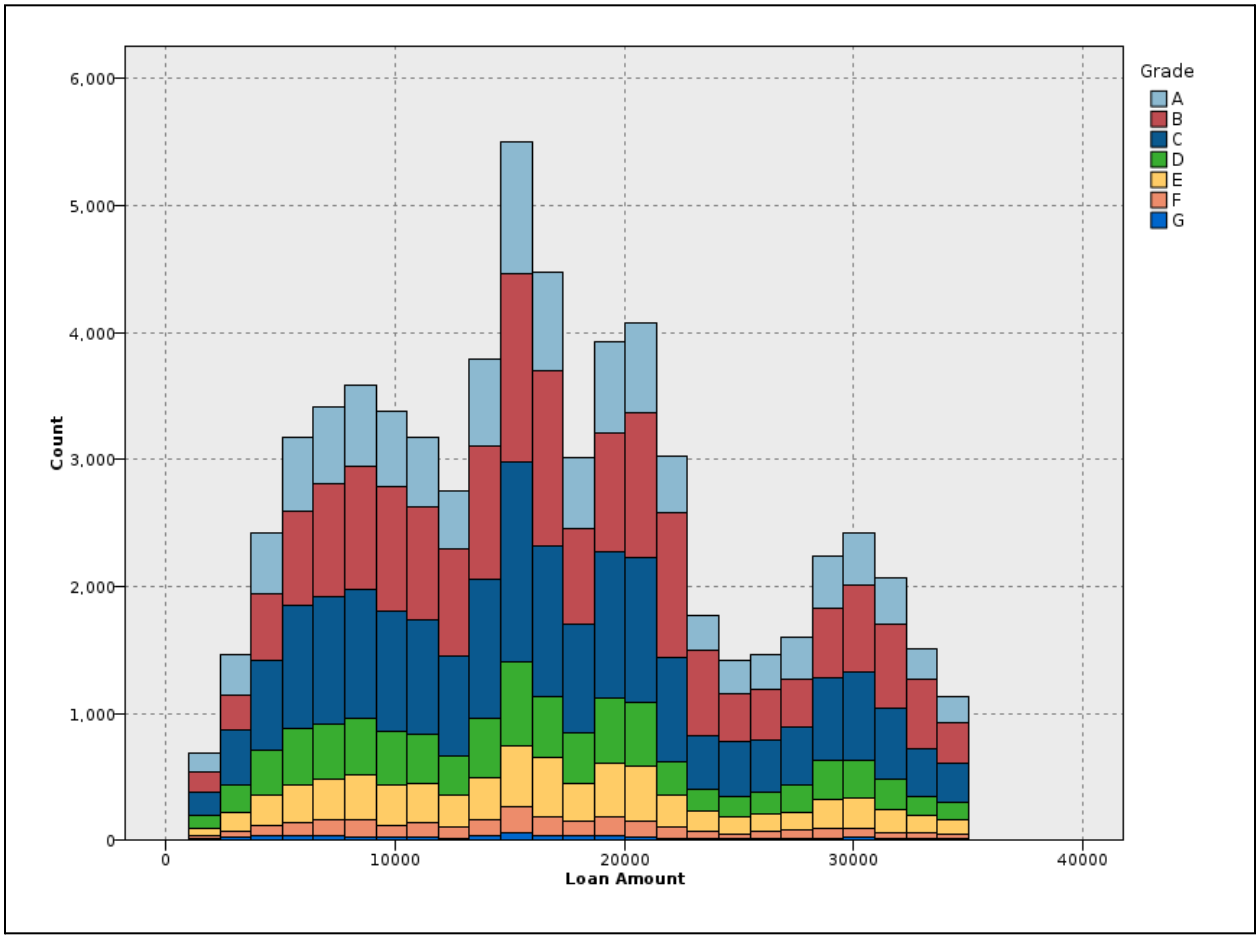
loan_title: A category provided by the borrower for the loan request.
verification_status: Status for verification of a loan.
term: The number of payments on the loan. Values are in months and can be either 36 or 60.
delinquency: 1 when the borrower had at least one event of delinquency in two years.
total_received_late_fee: Late fees received to date.
total_current_balance: Total balance in the bank of a particular borrower.
recoveries: The recovery rate is the estimated percent of a loan or an obligation that will still be repaid to creditors in the event of a default or bankruptcy.

- Have your explorations revealed new characteristics about the data?

Univariate Analysis:

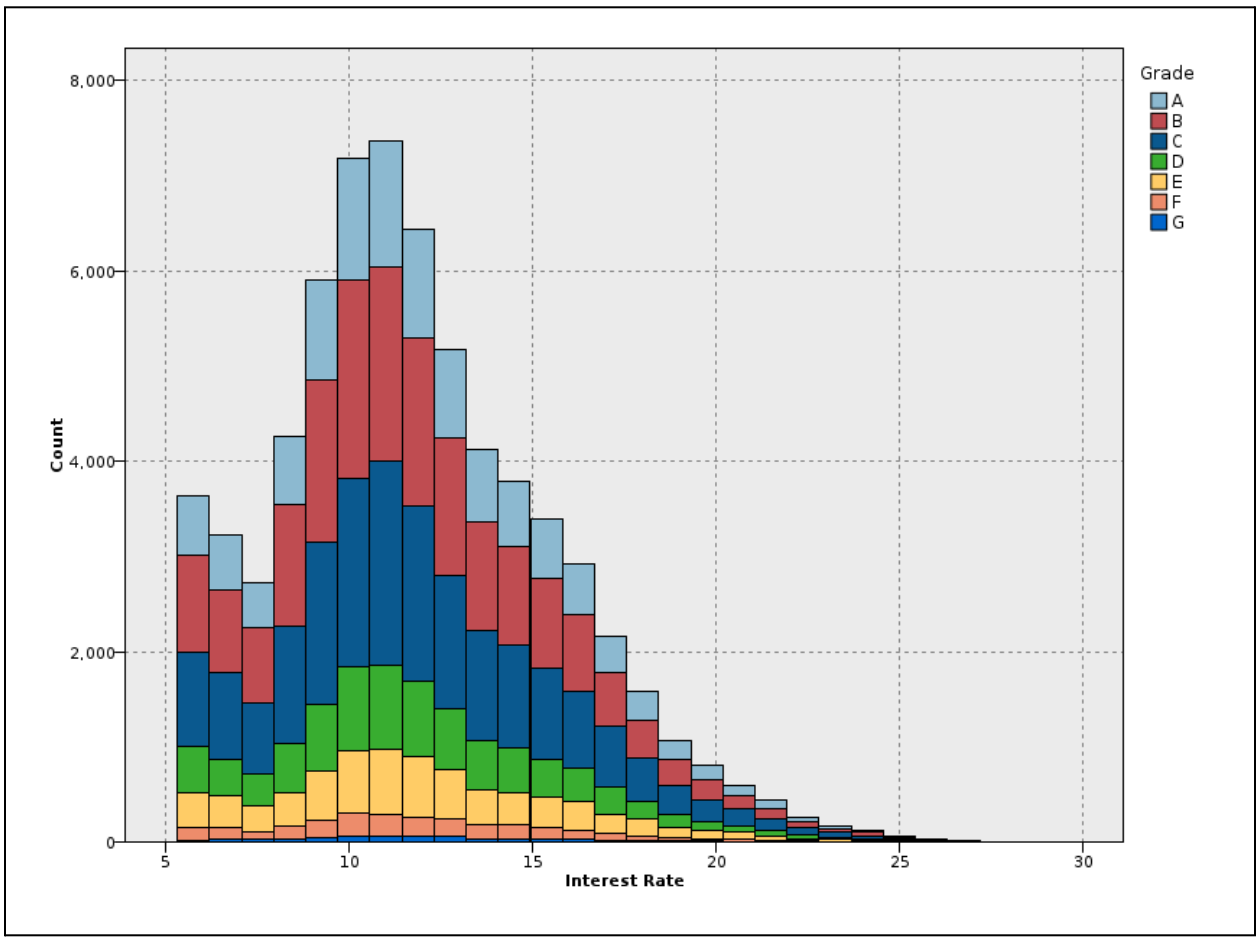
1. Loan Amount:

Insights: Most of the loan amounts are distributed between 10000 to 20000.



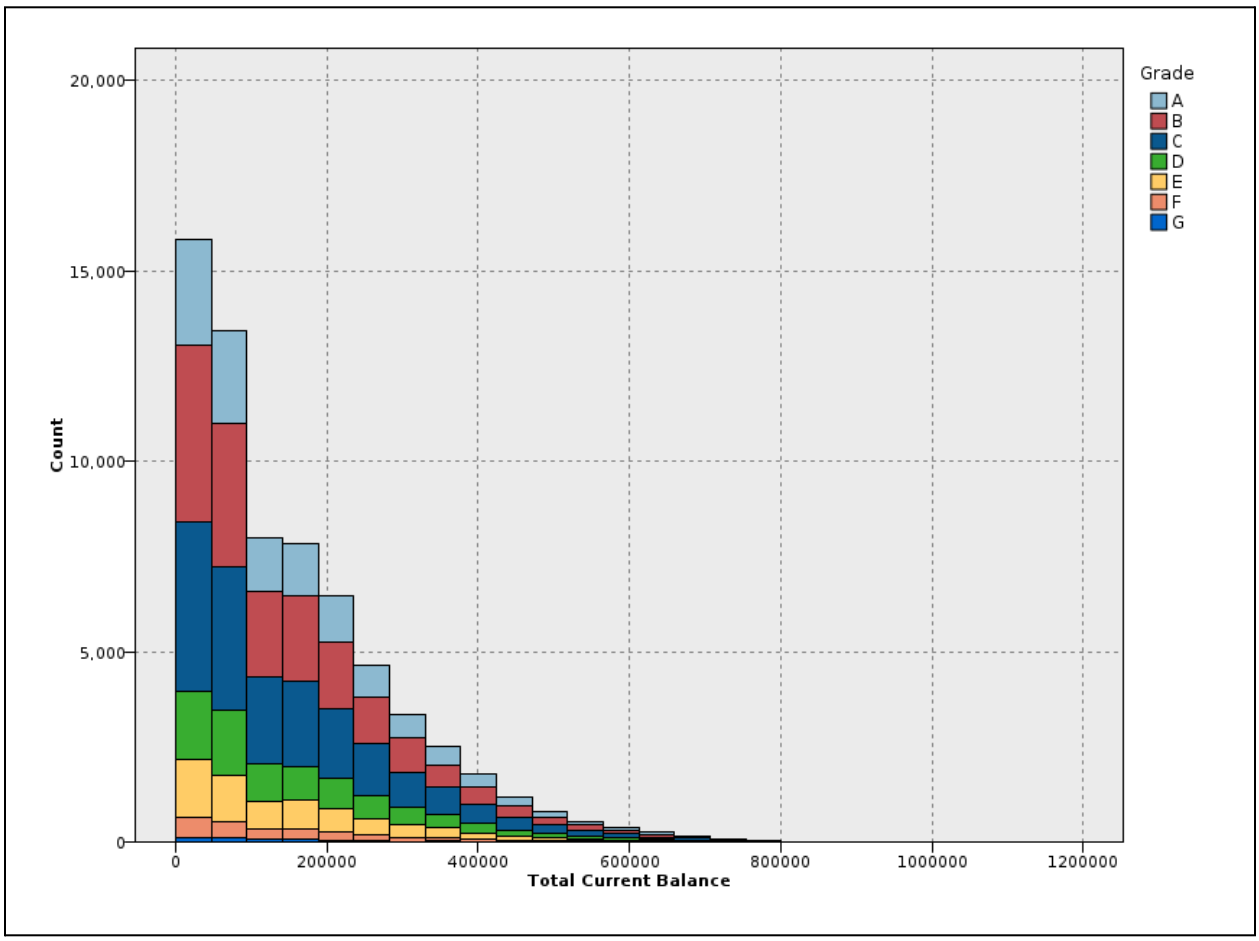
2. Interest Rate:

Insights: Most of the loans interest rates are distributed between 10% to 13%.



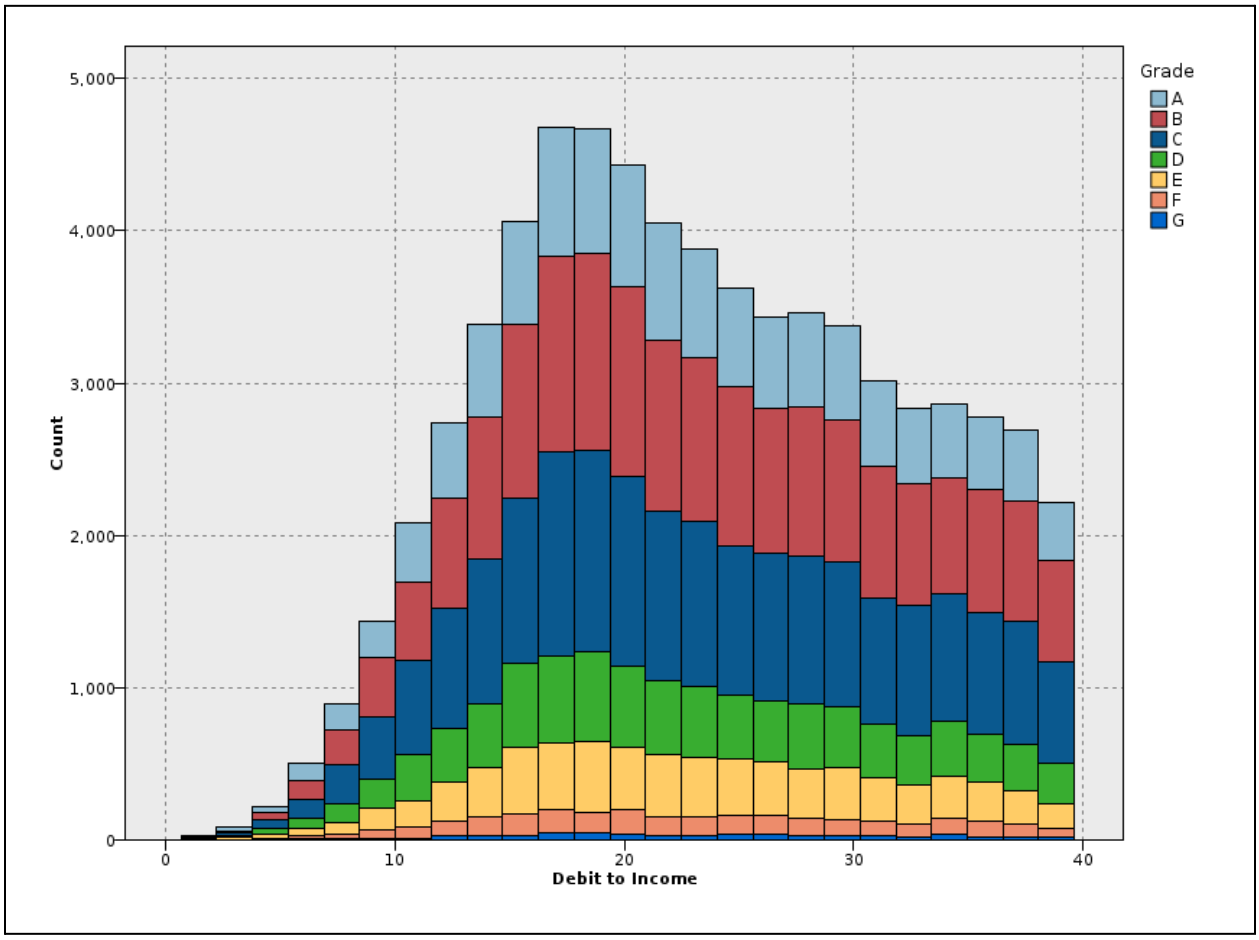
3. Total Current Balance:

Insights: Most of the applicants have total balance less than 200000.



4. dti (Debt to Income Ratio):

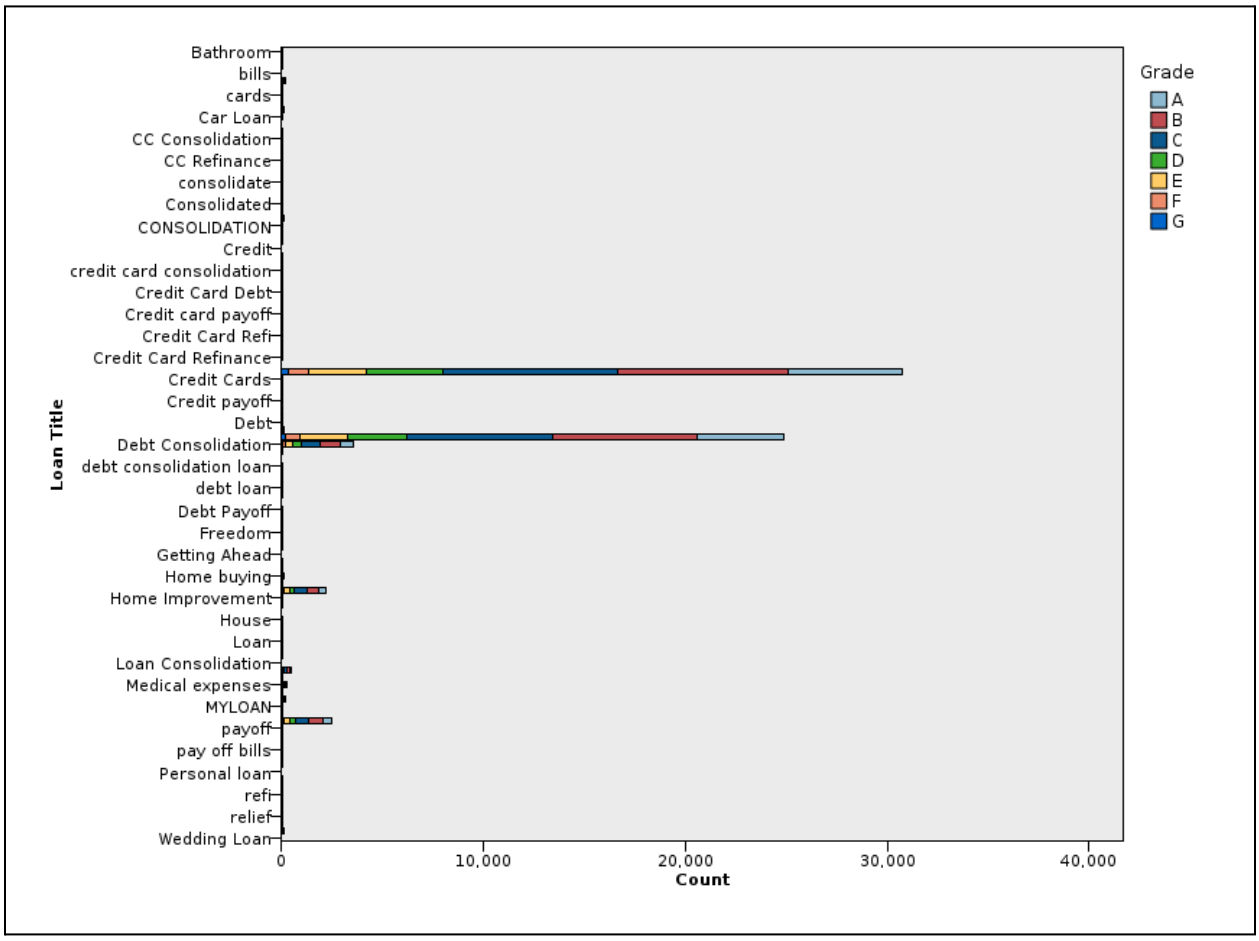
Insights: Most of the applicants have a debt-to-income (dti) ratio between 15 to 20.



Categorical Variables:

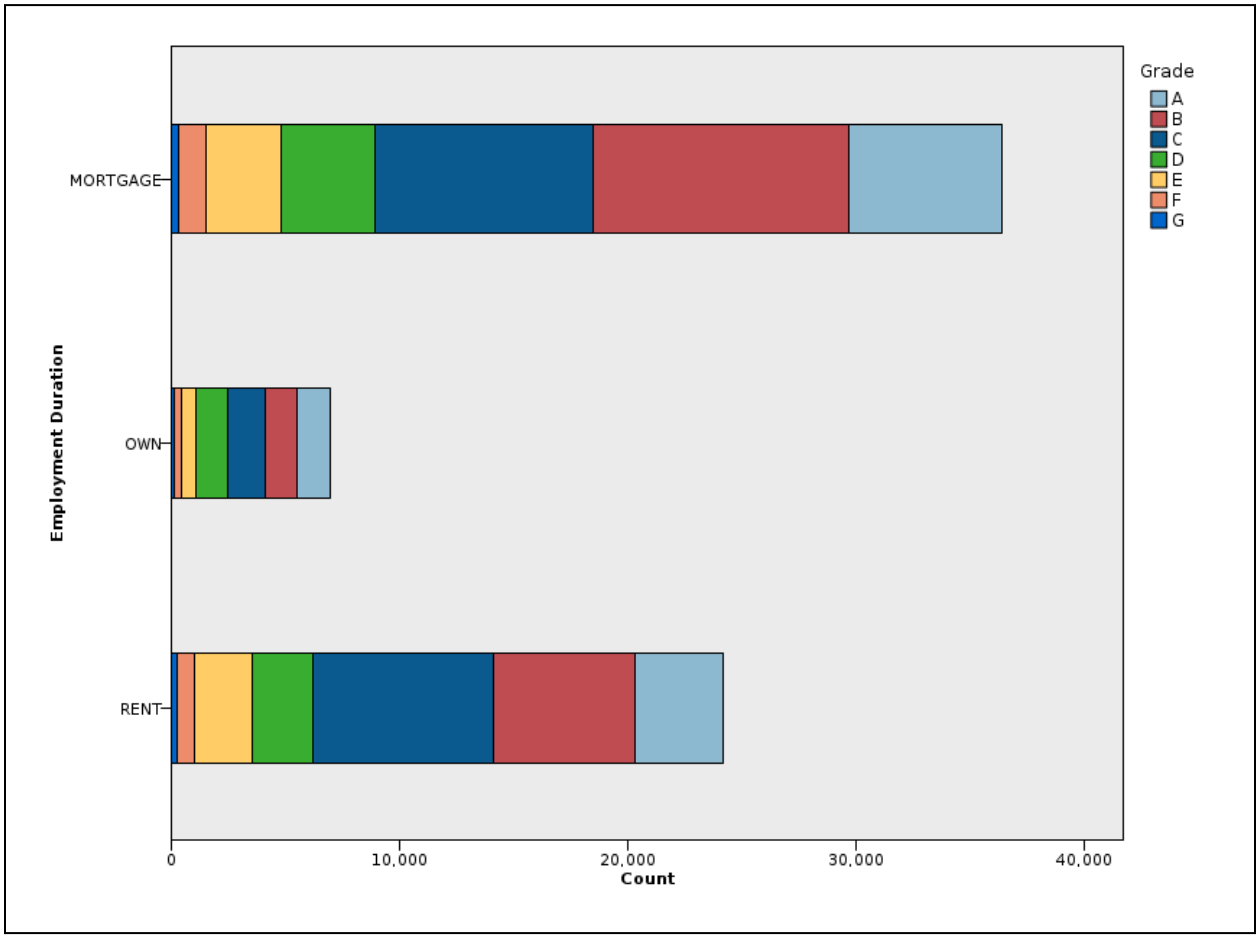
5. Purpose of Loan:

Insights: Insights: Approx. 60% of the applicants applied for a loan for paying their other loans and credit card bills (Debt Consolidation).



6. Home Ownership:

Insights: 30% of applicants are living in rented homes whereas 60% applicants were mortgaged their home.



- Can you identify particular subsets of data for later use?

  - ➢ Subsets of people with higher & lower income levels.
  - ➢ Subset of data in which a person has its own home ownership.
  - ➢ Subset of person with verification & non-verification status.
  - ➢ Subset of inquiries of the last six months.
  - ➢ Subsets of whether recoveries are high are not.

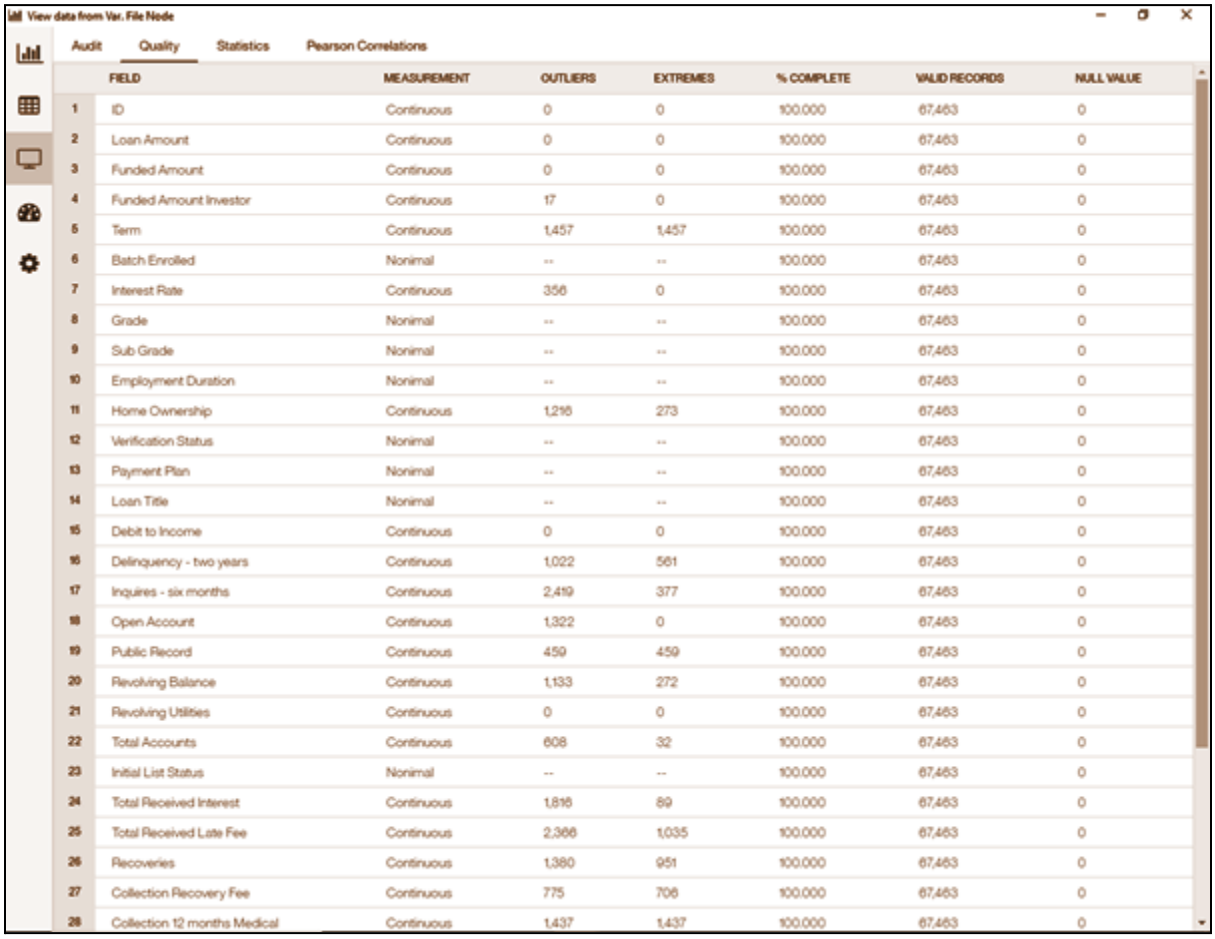
- Take another look at your data mining goals. Has this exploration altered the goals?

    ➢ No, because we have to predict that a person can pay the loan back or not & after visualizing the data we came to the conclusion that it doesn't alter the main goal.

Writing a Data Quality Report

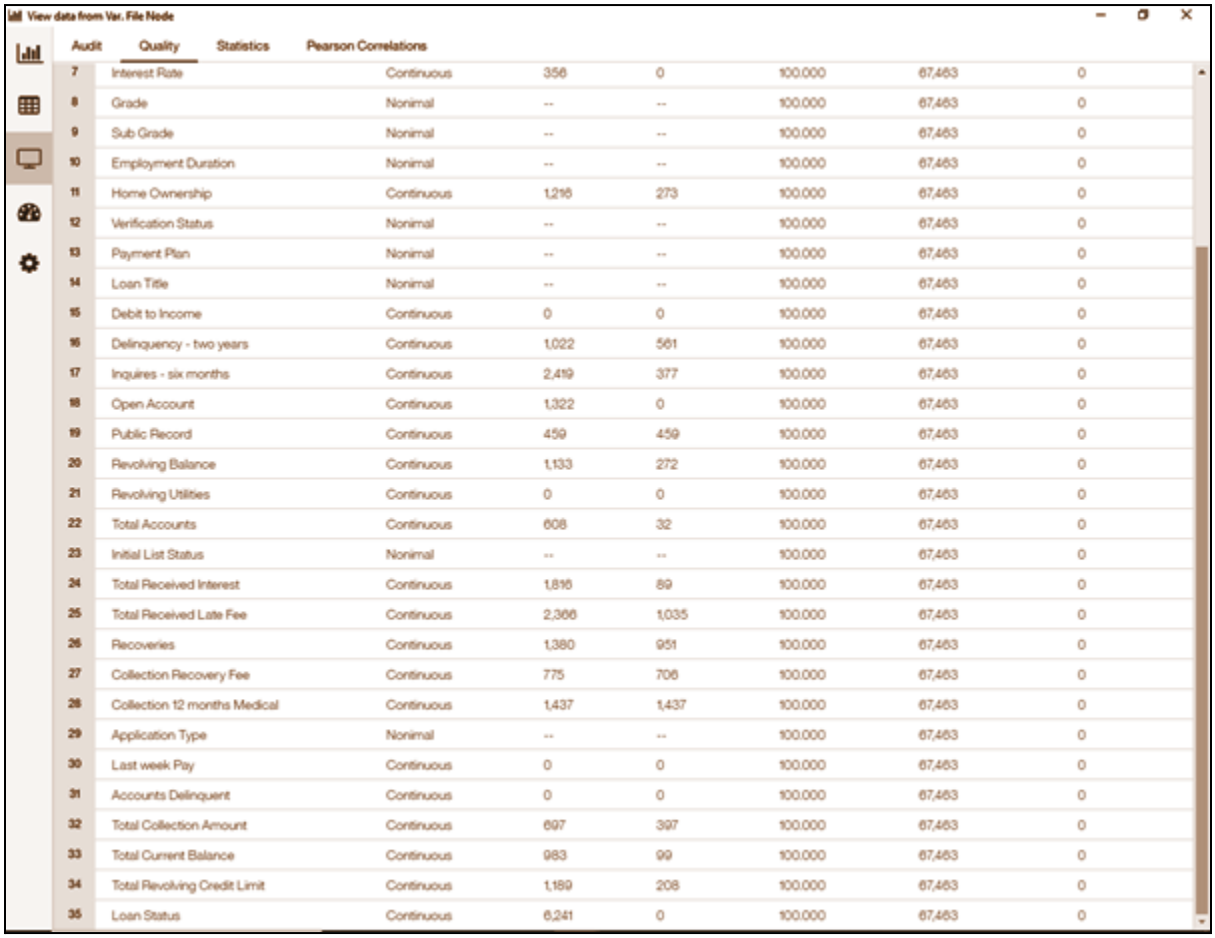
- Have you identified missing attributes and blank fields? If so, is there meaning behind such missing values?

No we haven't identified any missing values or blank fields in any attributes


View data from Var. File Node

Audit | Quality | Statistics | Pearson Correlations

	FIELD	MEASUREMENT	OUTLIERS	EXTREMES	% COMPLETE	VALID RECORDS	NULL VALUE
1	ID	Continuous	0	0	100.000	67,463	0
2	Loan Amount	Continuous	0	0	100.000	67,463	0
3	Funded Amount	Continuous	0	0	100.000	67,463	0
4	Funded Amount Investor	Continuous	17	0	100.000	67,463	0
5	Term	Continuous	1,457	1,457	100.000	67,463	0
6	Batch Enrolled	Nonimal	--	--	100.000	67,463	0
7	Interest Rate	Continuous	356	0	100.000	67,463	0
8	Grade	Nonimal	--	--	100.000	67,463	0
9	Sub Grade	Nonimal	--	--	100.000	67,463	0
10	Employment Duration	Nonimal	--	--	100.000	67,463	0
11	Home Ownership	Continuous	1,216	273	100.000	67,463	0
12	Verification Status	Nonimal	--	--	100.000	67,463	0
13	Payment Plan	Nonimal	--	--	100.000	67,463	0
14	Loan Title	Nonimal	--	--	100.000	67,463	0
15	Debit to Income	Continuous	0	0	100.000	67,463	0
16	Delinquency - two years	Continuous	1,022	561	100.000	67,463	0
17	Inquires - six months	Continuous	2,419	377	100.000	67,463	0
18	Open Account	Continuous	1,322	0	100.000	67,463	0
19	Public Record	Continuous	459	459	100.000	67,463	0
20	Revolving Balance	Continuous	1,133	272	100.000	67,463	0
21	Revolving Utilities	Continuous	0	0	100.000	67,463	0
22	Total Accounts	Continuous	608	32	100.000	67,463	0
23	Initial List Status	Nonimal	--	--	100.000	67,463	0
24	Total Received Interest	Continuous	1,816	89	100.000	67,463	0
25	Total Received Late Fee	Continuous	2,366	1,035	100.000	67,463	0
26	Recoveries	Continuous	1,380	951	100.000	67,463	0
27	Collection Recovery Fee	Continuous	775	706	100.000	67,463	0
28	Collection 12 months Medical	Continuous	1,437	1,437	100.000	67,463	0


View data from Var. File Node

Audit | Quality | Statistics | Pearson Correlations

7	Interest Rate	Continuous	356	0	100,000	67,463	0
8	Grade	Nominal	--	--	100,000	67,463	0
9	Sub Grade	Nominal	--	--	100,000	67,463	0
10	Employment Duration	Nominal	--	--	100,000	67,463	0
11	Home Ownership	Continuous	1,216	273	100,000	67,463	0
12	Verification Status	Nominal	--	--	100,000	67,463	0
13	Payment Plan	Nominal	--	--	100,000	67,463	0
14	Loan Title	Nominal	--	--	100,000	67,463	0
15	Debit to Income	Continuous	0	0	100,000	67,463	0
16	Delinquency - two years	Continuous	1,022	561	100,000	67,463	0
17	Inquires - six months	Continuous	2,419	377	100,000	67,463	0
18	Open Account	Continuous	1,322	0	100,000	67,463	0
19	Public Record	Continuous	459	459	100,000	67,463	0
20	Revolving Balance	Continuous	1,133	272	100,000	67,463	0
21	Revolving Utilities	Continuous	0	0	100,000	67,463	0
22	Total Accounts	Continuous	608	32	100,000	67,463	0
23	Initial List Status	Nominal	--	--	100,000	67,463	0
24	Total Received Interest	Continuous	1,816	89	100,000	67,463	0
25	Total Received Late Fee	Continuous	2,366	1,035	100,000	67,463	0
26	Recoveries	Continuous	1,380	951	100,000	67,463	0
27	Collection Recovery Fee	Continuous	775	706	100,000	67,463	0
28	Collection 12 months Medical	Continuous	1,437	1,437	100,000	67,463	0
29	Application Type	Nominal	--	--	100,000	67,463	0
30	Last week Pay	Continuous	0	0	100,000	67,463	0
31	Accounts Delinquent	Continuous	0	0	100,000	67,463	0
32	Total Collection Amount	Continuous	697	397	100,000	67,463	0
33	Total Current Balance	Continuous	983	99	100,000	67,463	0
34	Total Revolving Credit Limit	Continuous	1,189	208	100,000	67,463	0
35	Loan Status	Continuous	6,241	0	100,000	67,463	0

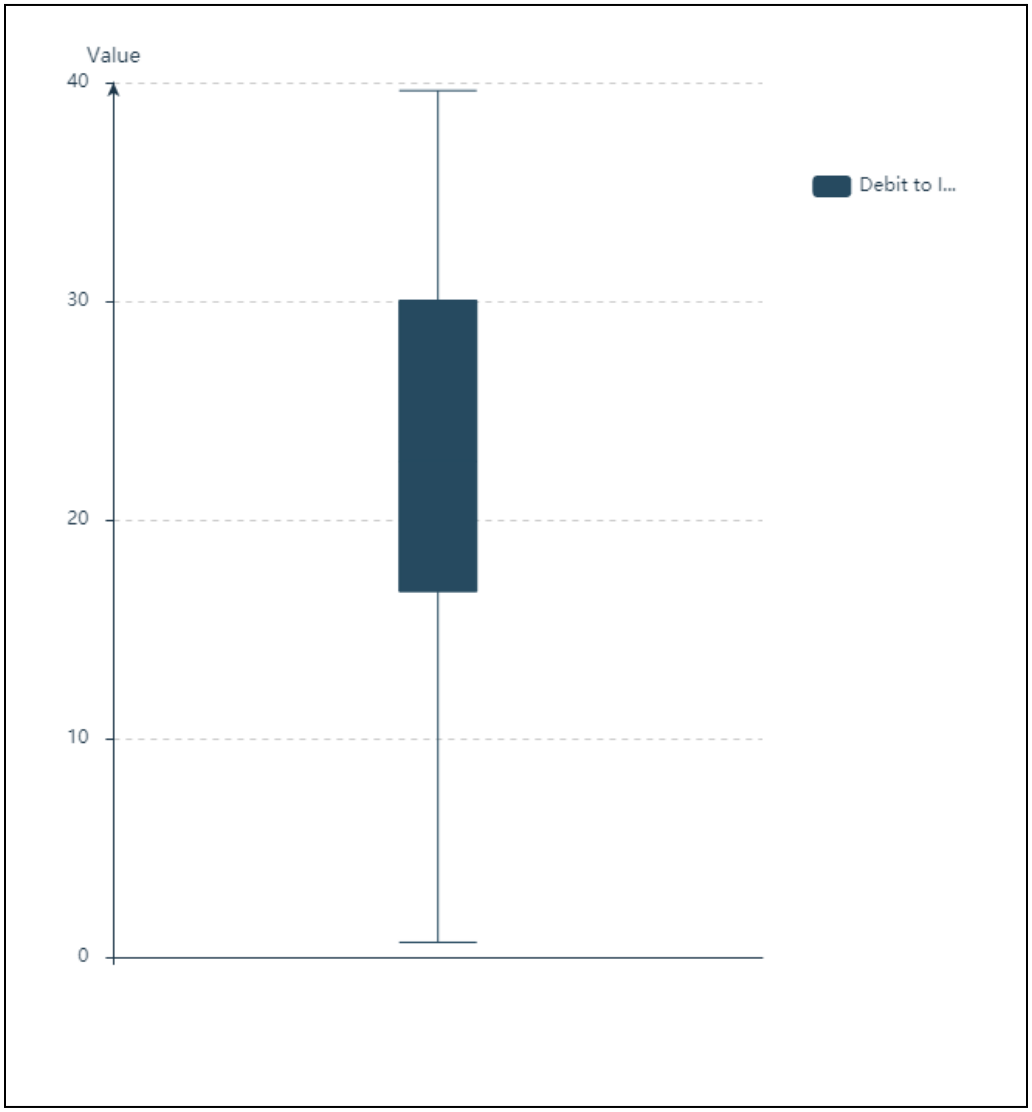
- Are there spelling inconsistencies that may cause problems in later merges or transformations?

  There are no spelling inconsistencies in the attributes so it will not cause any problems in later merger or transformation.

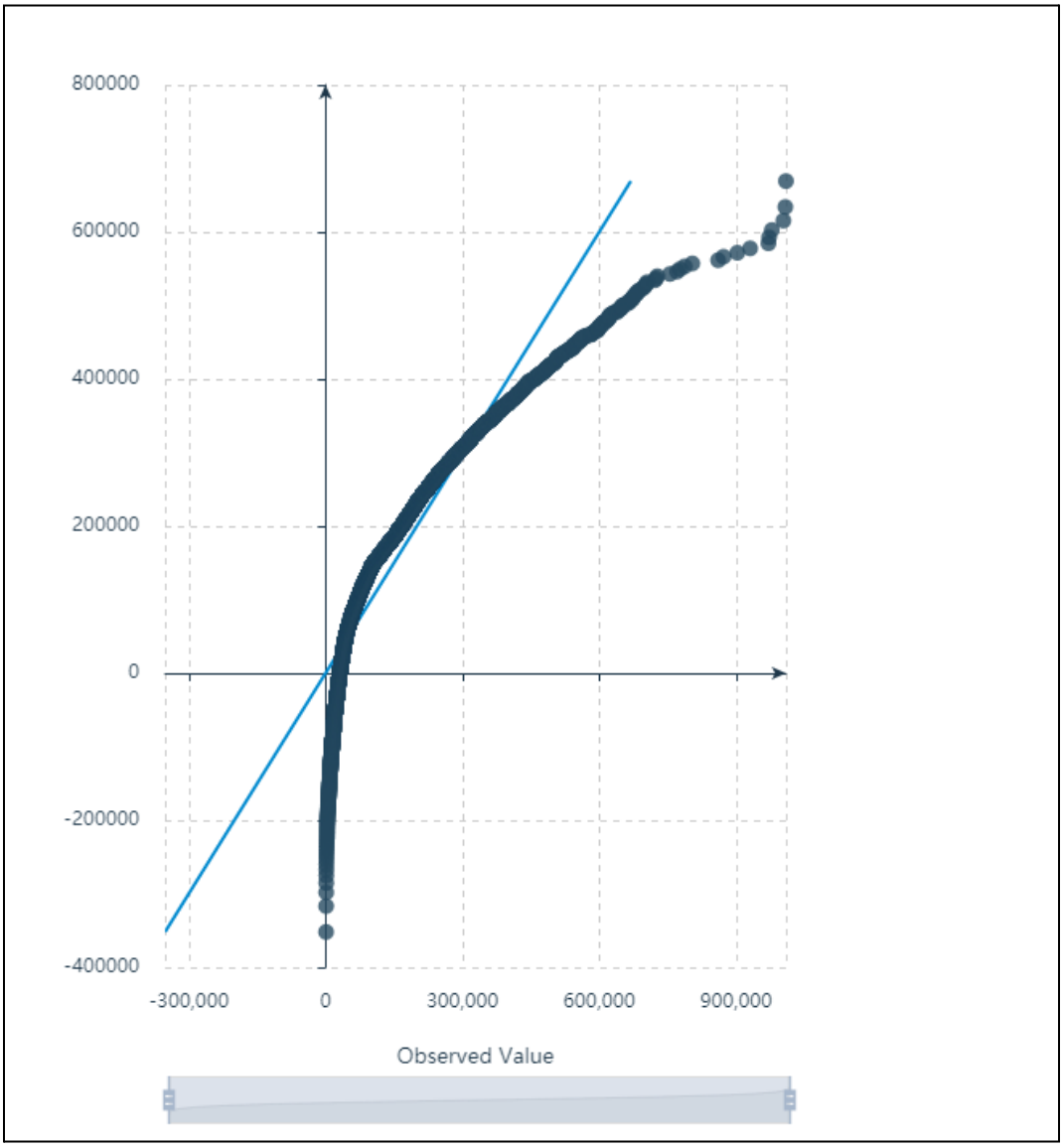
- Have you explored deviations to determine whether they are "noise" or phenomena worth analyzing further?

Yes, we explored datasets where there are outliers and extremes are present in some columns but they are valid outliers, all of these outliers and extremes didn’t affect the further analysis.

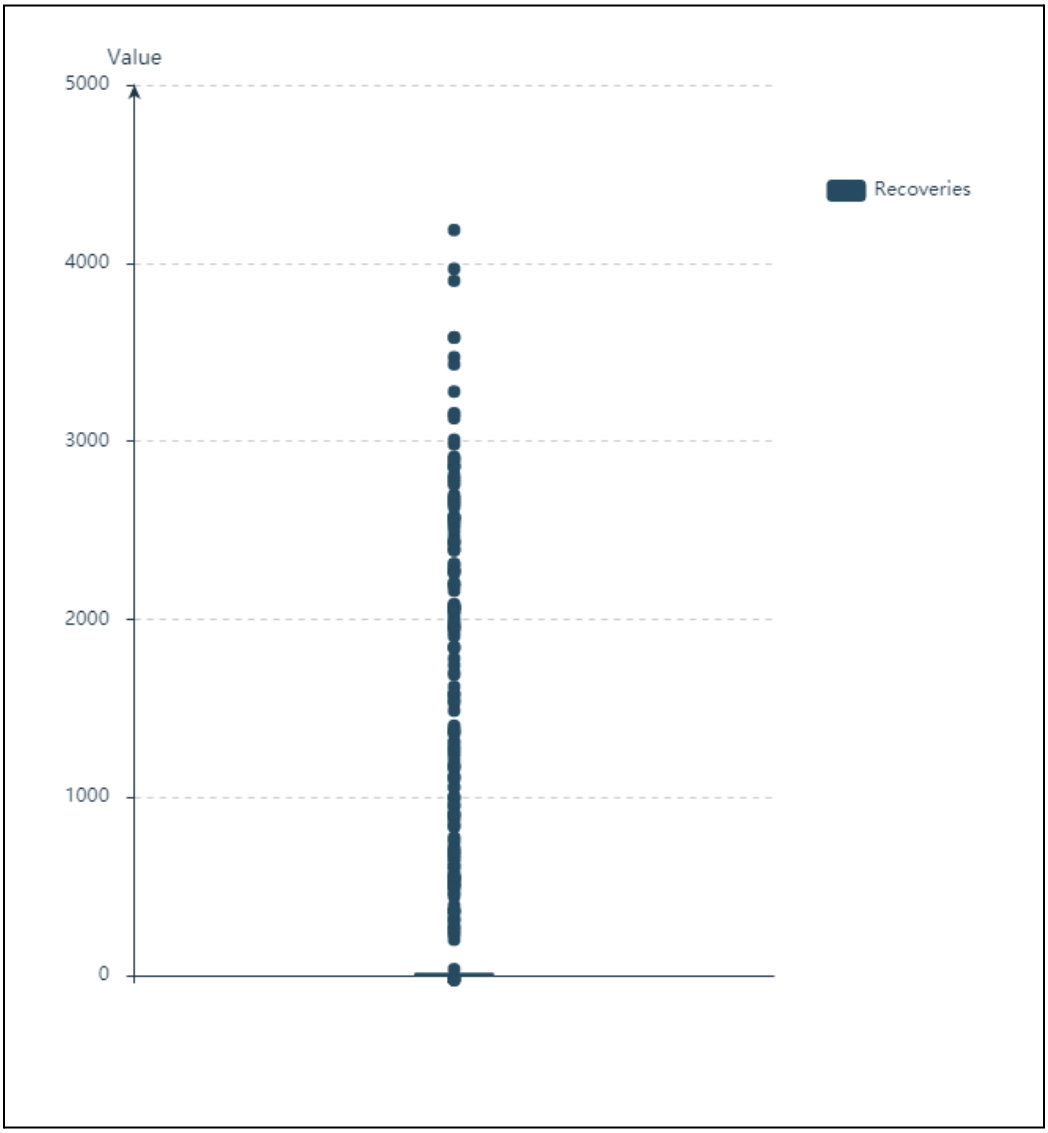
➢ Box plot for debt to income



➢ Q-Q plot of total_current_balance



➢ Box plot for recoveries attribute



- Have you conducted a plausibility check for values? Take notes on any apparent conflicts (such as teenagers with high income levels).

Yes, there is plausibility in our data.

- Have you considered excluding data that has no impact on your hypotheses?

  Yes , we have considered excluding the data that has no impact on our hypothesis.

- Is the data stored in flat files? If so, are the delimiters consistent among files? Does each record contain the same number of fields?

  Yes , data is stored in a flat file and comma delimiter is consistent in the file.

  Yes , each record contains the same number of fields.