PROJECT REPORT - TEAM 6
LoanStats – Data Preprocessing

<u>Team Members</u>

Aravind Senthil Kumar

Mohit Kumar Dhiman

Anjani Korkonda Bhattar

Poojith Routhu

Sneha Jyotindrakumar

Note: **"The work contained and presented here is our team's work and our team's work alone."**

Contents

1.0 Executive Summary

We started our data pre-processing by obtaining the random sample data set “loan_sample_final” of size 30000 from the main data table “LoanStats modified fall 2018 Group Project 1” where we first checked the data type of all variables in JMP and made necessary corrections.

We then used **Summary table** to find missing data across all fields and found that it constituted less than 5% of the total sample. We also used various combinations of **Missing Data Pattern** and found out that 1089 rows were missing data in 8 fields in the dataset. We further analyzed other missing data in the fields. We then retained, modified or deleted the rows as applicable and saved the new version of the sample as “**loan_sample_final_v2**”.

We tried different methods for modifying the data like contingency tables and distributions for each variable, recoding the necessary variables and formatting them to be used for analysis, and applying formulas for creating new columns.

We tried to take actions on the observed inconsistencies in the data. We also tried to check for various outlier methods like univariate and multivariate outlier analysis to see for the potential outliers. We transformed some variables to best fit the data using Shash transform, Johnson Si transform, etc.

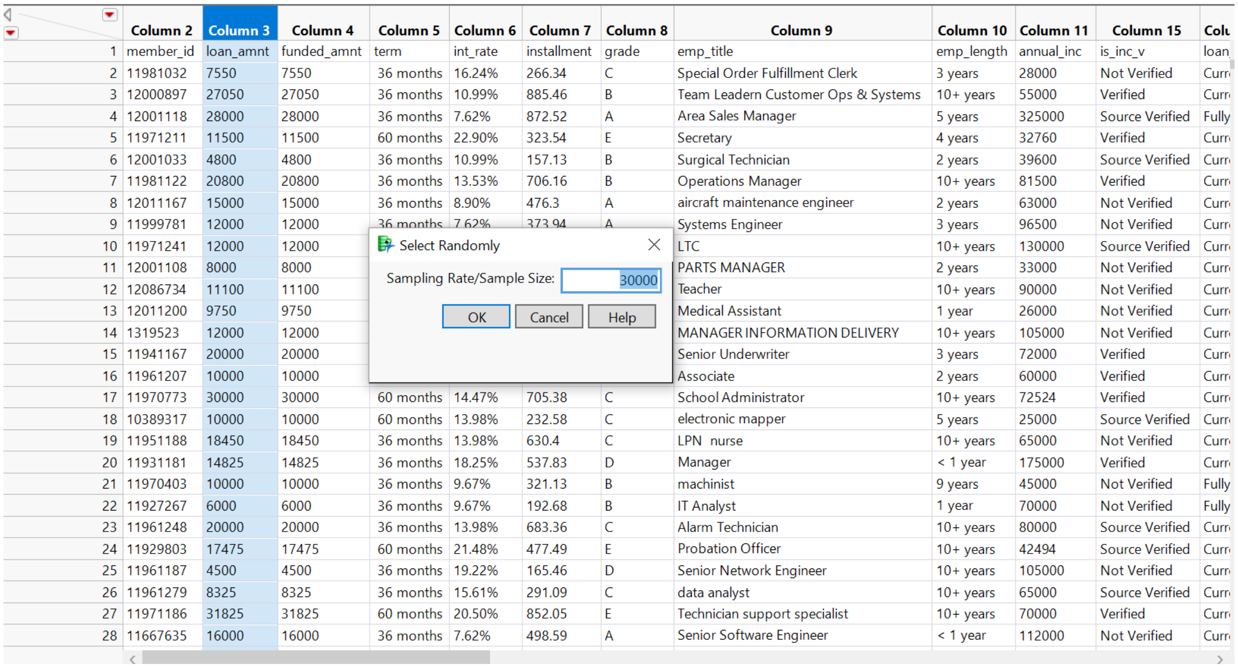
For data reduction we performed bivariate and multivariate correlation analysis between the fields and took necessary actions based on the insights from the analysis for the data preparation.

We also performed the Principal Component Analysis for finding the cumulative percentages of information captured by the principal components. From the observations of PCA we decided to retain 18 Principal Components for modeling purposes for 98% information retain.

Finally, we have two data sets, one with the Principal Components and one with the original continuous fields. We have decided to keep both data sets and try modeling on both data sets to see which provides more accurate results.

2.0 Sampling

1. Open the data table LoanStats modified fall 2018 Group Project 1
2. Go to **Row → Row selection → Select Randomly**
3. Enter the sample size value as 30000



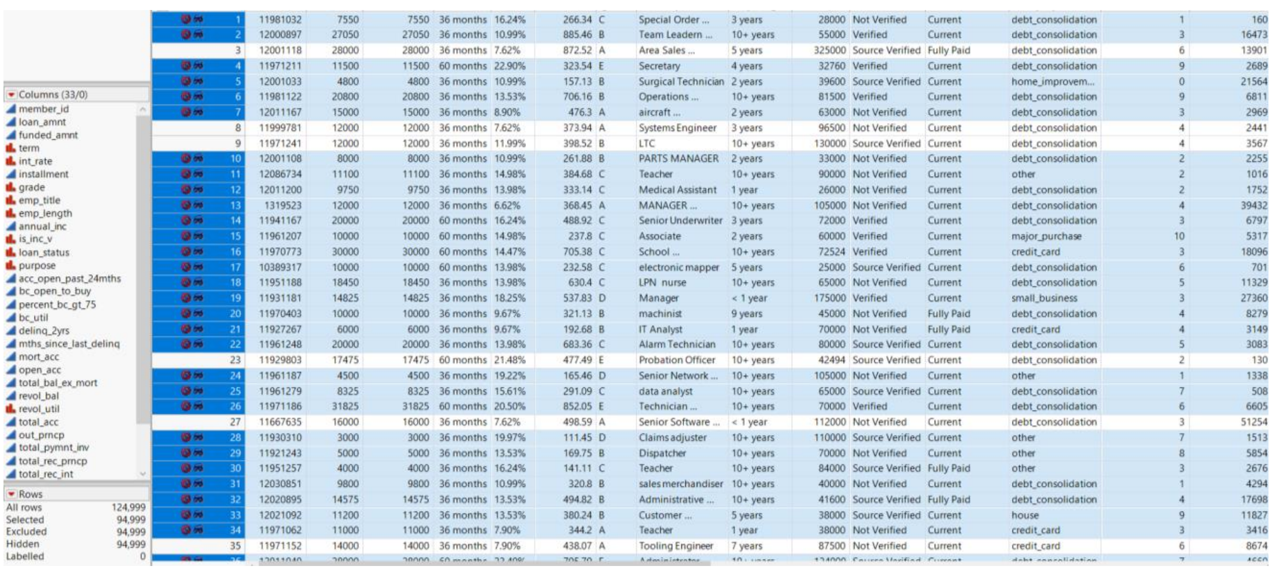
	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Column 10	Column 11	Column 15	Colu
1	member_id	loan_amnt	funded_amnt	term	int_rate	installment	grade	emp_title	emp_length	annual_inc	is_inc_v	loan
2	11981032	7550	7550	36 months	16.24%	266.34	C	Special Order Fulfillment Clerk	3 years	28000	Not Verified	Curr
3	12000897	27050	27050	36 months	10.99%	885.46	B	Team Leadern Customer Ops & Systems	10+ years	55000	Verified	Curr
4	12001118	28000	28000	36 months	7.62%	872.52	A	Area Sales Manager	5 years	325000	Source Verified	Fully
5	11971211	11500	11500	60 months	22.90%	323.54	E	Secretary	4 years	32760	Verified	Curr
6	12001033	4800	4800	36 months	10.99%	157.13	B	Surgical Technician	2 years	39600	Source Verified	Curr
7	11981122	20800	20800	36 months	13.53%	706.16	B	Operations Manager	10+ years	81500	Verified	Curr
8	12011167	15000	15000	36 months	8.90%	476.3	A	aircraft maintenance engineer	2 years	63000	Not Verified	Curr
9	11999781	12000	12000	36 months	7.62%	373.94	A	Systems Engineer	3 years	96500	Not Verified	Curr
10	11971241	12000	12000					LTC	10+ years	130000	Source Verified	Curr
11	12001108	8000	8000					PARTS MANAGER	2 years	33000	Not Verified	Curr
12	12086734	11100	11100					Teacher	10+ years	90000	Not Verified	Curr
13	12011200	9750	9750					Medical Assistant	1 year	26000	Not Verified	Curr
14	1319523	12000	12000					MANAGER INFORMATION DELIVERY	10+ years	105000	Not Verified	Curr
15	11941167	20000	20000					Senior Underwriter	3 years	72000	Verified	Curr
16	11961207	10000	10000					Associate	2 years	60000	Verified	Curr
17	11970773	30000	30000	60 months	14.47%	705.38	C	School Administrator	10+ years	72524	Verified	Curr
18	10389317	10000	10000	60 months	13.98%	232.58	C	electronic mapper	5 years	25000	Source Verified	Curr
19	11951188	18450	18450	36 months	13.98%	630.4	C	LPN nurse	10+ years	65000	Not Verified	Curr
20	11931181	14825	14825	36 months	18.25%	537.83	D	Manager	< 1 year	175000	Verified	Curr
21	11970403	10000	10000	36 months	9.67%	321.13	B	machinist	9 years	45000	Not Verified	Fully
22	11927267	6000	6000	36 months	9.67%	192.68	B	IT Analyst	1 year	70000	Not Verified	Fully
23	11961248	20000	20000	36 months	13.98%	683.36	C	Alarm Technician	10+ years	80000	Source Verified	Curr
24	11929803	17475	17475	60 months	21.48%	477.49	E	Probation Officer	10+ years	42494	Source Verified	Curr
25	11961187	4500	4500	36 months	19.22%	165.46	D	Senior Network Engineer	10+ years	105000	Not Verified	Curr
26	11961279	8325	8325	36 months	15.61%	291.09	C	data analyst	10+ years	65000	Source Verified	Curr
27	11971186	31825	31825	60 months	20.50%	852.05	E	Technician support specialist	10+ years	70000	Verified	Curr
28	11667635	16000	16000	36 months	7.62%	498.59	A	Senior Software Engineer	< 1 year	112000	Not Verified	Curr

Select Randomly

Sampling Rate/Sample Size: 30000

OK Cancel Help

4. When the sample rows are highlighted, go to **Rows → Row selection → Invert Row Selection**
5. Go to **Rows → Hide and Exclude**



Columns (33/0)
member_id
loan_amnt
funded_amnt
term
int_rate
installment
grade
emp_title
emp_length
annual_inc
is_inc_v
loan_status
purpose
acc_open_past_24mths
bc_open_to_buy
percent_bc_gt_75
bc_util
delinq_2yrs
mths_since_last_delinq
mort_acc
open_acc
total_bal_ex_mort
revol_bal
revol_util
total_acc
out_prncp
total_pymnt_inv
total_rec_prncp
total_rec_int

Rows	
All rows	124,999
Selected	94,999
Excluded	94,999
Hidden	94,999
Labelled	0

1	11981032	7550	7550	36 months	16.24%	266.34	C	Special Order ...	3 years	28000	Not Verified	Current	debt_consolidation	1	160
2	12000897	27050	27050	36 months	10.99%	885.46	B	Team Leadern ...	10+ years	55000	Verified	Current	debt_consolidation	3	16473
3	12001118	28000	28000	36 months	7.62%	872.52	A	Area Sales ...	5 years	325000	Source Verified	Fully Paid	debt_consolidation	6	13901
4	11971211	11500	11500	60 months	22.90%	323.54	E	Secretary	4 years	32760	Verified	Current	debt_consolidation	9	2689
5	12001033	4800	4800	36 months	10.99%	157.13	B	Surgical Technician	2 years	39600	Source Verified	Current	home_improvem...	0	21564
6	11981122	20800	20800	36 months	13.53%	706.16	B	Operations ...	10+ years	81500	Verified	Current	debt_consolidation	9	6811
7	12011167	15000	15000	36 months	8.90%	476.3	A	aircraft ...	2 years	63000	Not Verified	Current	debt_consolidation	3	2969
8	11999781	12000	12000	36 months	7.62%	373.94	A	Systems Engineer	3 years	96500	Not Verified	Current	debt_consolidation	4	2441
9	11971241	12000	12000	36 months	11.99%	398.52	B	LTC	10+ years	130000	Source Verified	Current	debt_consolidation	4	3567
10	12001108	8000	8000	36 months	10.99%	261.88	B	PARTS MANAGER	2 years	33000	Not Verified	Current	debt_consolidation	2	2255
11	12086734	11100	11100	36 months	14.98%	384.68	C	Teacher	10+ years	90000	Not Verified	Current	other	2	1016
12	12011200	9750	9750	36 months	13.98%	333.14	C	Medical Assistant	1 year	26000	Not Verified	Current	debt_consolidation	2	1752
13	1319523	12000	12000	36 months	6.62%	368.45	A	MANAGER ...	10+ years	105000	Not Verified	Current	debt_consolidation	4	39432
14	11941167	20000	20000	60 months	16.24%	488.92	C	Senior Underwriter	3 years	72000	Verified	Current	debt_consolidation	3	6797
15	11961207	10000	10000	60 months	14.98%	237.8	C	Associate	2 years	60000	Verified	Current	major_purchase	10	5317
16	11970773	30000	30000	60 months	14.47%	705.38	C	School ...	10+ years	72524	Verified	Current	credit_card	3	18096
17	10389317	10000	10000	60 months	13.98%	232.58	C	electronic mapper	5 years	25000	Source Verified	Current	debt_consolidation	6	701
18	11951188	18450	18450	36 months	13.98%	630.4	C	LPN nurse	10+ years	65000	Not Verified	Current	debt_consolidation	5	11329
19	11931181	14825	14825	36 months	18.25%	537.83	D	Manager	< 1 year	175000	Verified	Current	small_business	3	27360
20	11970403	10000	10000	36 months	9.67%	321.13	B	machinist	9 years	45000	Not Verified	Fully Paid	debt_consolidation	4	8279
21	11927267	6000	6000	36 months	9.67%	192.68	B	IT Analyst	1 year	70000	Not Verified	Fully Paid	credit_card	4	3149
22	11961248	20000	20000	36 months	13.98%	683.36	C	Alarm Technician	10+ years	80000	Source Verified	Current	debt_consolidation	5	3083
23	11929803	17475	17475	60 months	21.48%	477.49	E	Probation Officer	10+ years	42494	Source Verified	Current	debt_consolidation	2	130
24	11961187	4500	4500	36 months	19.22%	165.46	D	Senior Network ...	10+ years	105000	Not Verified	Current	other	1	1338
25	11961279	8325	8325	36 months	15.61%	291.09	C	data analyst	10+ years	65000	Source Verified	Current	debt_consolidation	7	508
26	11971186	31825	31825	60 months	20.50%	852.05	E	Technician ...	10+ years	70000	Verified	Current	debt_consolidation	6	6605
27	11667635	16000	16000	36 months	7.62%	498.59	A	Senior Software ...	< 1 year	112000	Not Verified	Current	debt_consolidation	3	51254
28	11930310	3000	3000	36 months	19.97%	111.45	D	Claims adjuster	10+ years	110000	Source Verified	Current	other	7	1513
29	11921243	5000	5000	36 months	13.53%	169.75	B	Dispatcher	10+ years	70000	Not Verified	Current	other	8	5854
30	11951257	4000	4000	36 months	16.24%	141.11	C	Teacher	10+ years	84000	Source Verified	Fully Paid	other	3	2676
31	12030851	9800	9800	36 months	10.99%	320.8	B	sales merchandiser	10+ years	40000	Not Verified	Current	debt_consolidation	1	4294
32	12020895	14575	14575	36 months	13.53%	494.82	B	Administrative ...	10+ years	41600	Source Verified	Fully Paid	debt_consolidation	4	17698
33	12021092	11200	11200	36 months	13.53%	380.24	B	Customer ...	5 years	38000	Source Verified	Current	house	9	11827
34	11971062	11000	11000	36 months	7.90%	344.2	A	Teacher	1 year	38000	Not Verified	Current	credit_card	3	3416
35	11971152	14000	14000	36 months	7.90%	438.07	A	Tooling Engineer	7 years	87500	Not Verified	Current	credit_card	6	8674

6. Export this data to excel to remove the headers 'Column1', 'Column2' etc.
7. Import it back to JMP and save the sample as loan_sample_final

3.0 Key Observations

Below are the highlights of our initial observations

1. *loan_amnt* and *funded_amnt* have a correlation of 0.9997
2. There are 1447 rows with no data in any column – 4.84% of the total data
3. The variables *int_rate*, *emp_length*, *revol_util* are displayed as nominal variables
4. *mnths_since_last_delinq* has 17,932 missing values
5. *revol_balance* has an extreme outlier value '1746716' in Row 2699
6. The following fields have the same set of 1089 values missing
    a. *acc_open_past_24mnths*
    b. *bc_open_to_buy*
    c. *percent_bc_gt_75*
    d. *bc_util*
    e. *mort_acc*
    f. *total_bal_ex_mort*
    g. *num_rev_accts*
    h. *total_cur_bal*

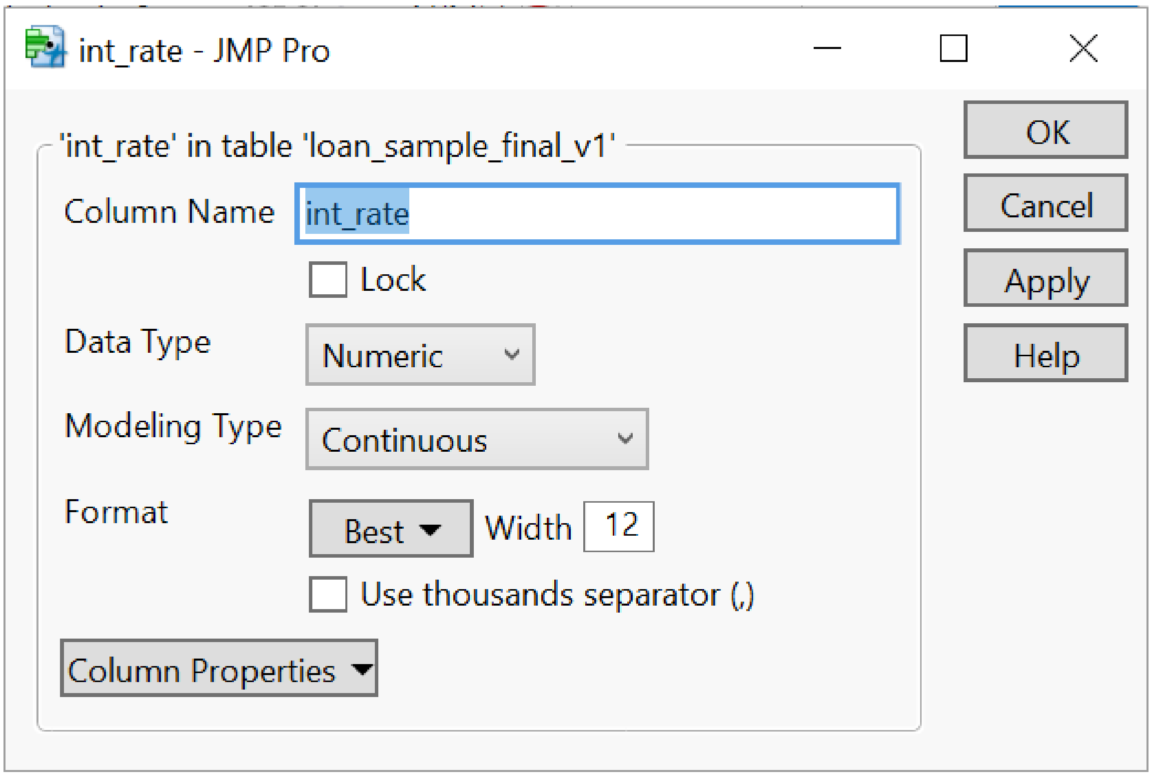
Note: The above data are just the highlights of our initial observations. For detailed analysis and modifications information, please refer to the Data Preprocessing section

4.0 Data Preprocessing

4.1 Changing variable types

The variables *int_rate*, *emp_length*, *Revol_util* are displayed as nominal variables and need to be changed to continuous variables for further analysis.

1. *emp_length*
   a. To change this variable to continuous, the term 'years' has to be removed from the values.
   b. The word 'years' was eliminated using **Recode** and the resultant data was saved in a new column 'emp_length2'
   c. By this, the variable emp_length was changed to continuous
2. *int_rate*
   a. Each value in this variable contains the percentage sign (%)
   b. The symbol was removed by accessing **Columns → Column Info**, changing the Data Type to 'Numeric' and Modeling Type to 'Continuous'



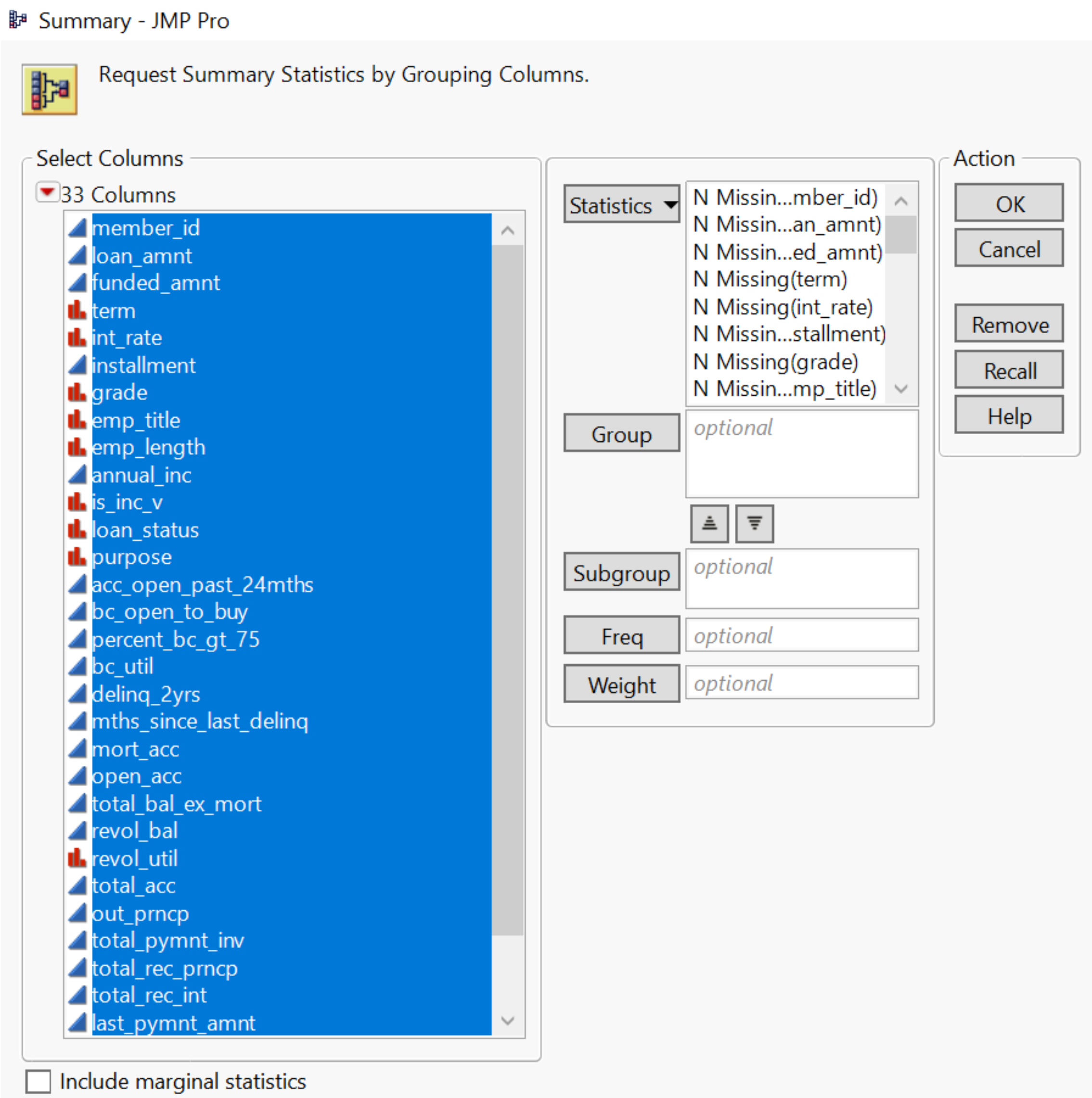
3. *revol_util*
   a. Every value in this variable also contains the percentage sign (%)
   b. Similarly, the symbol was removed by accessing Columns → Column Info, changing the Data Type to 'Numeric' and Modeling Type to 'Continuous'

4.2 Data Cleaning

4.2.1 Missing Values

4.2.1.1 Observations

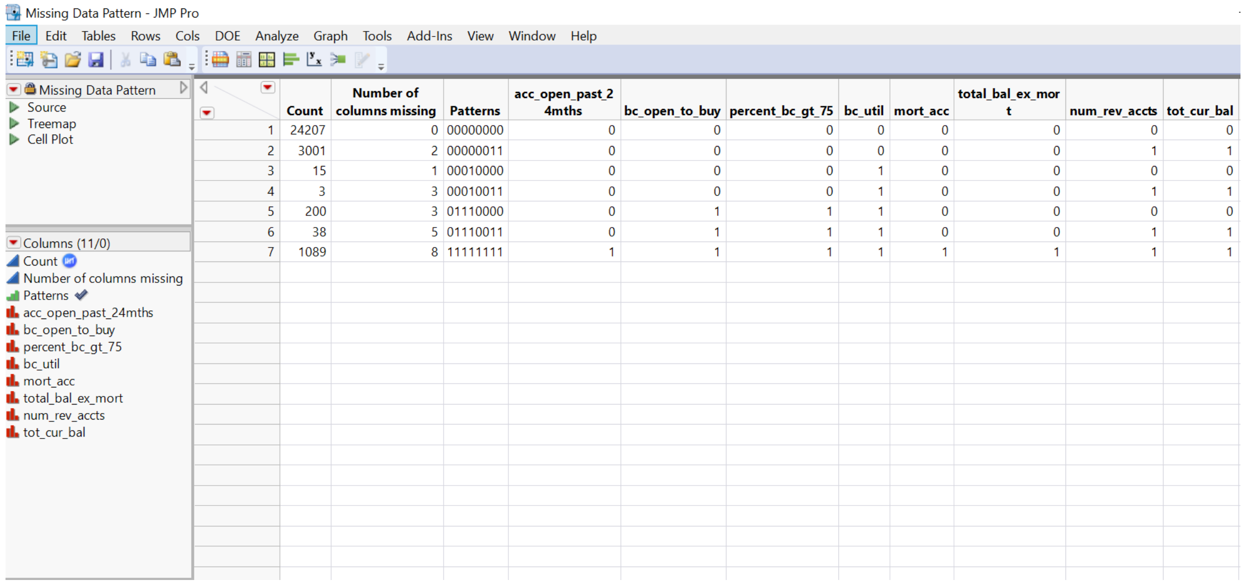
1. Missing data across all fields was identified using **Tables → Summary**
2. All the fields in the left pane were selected under the Statistics ‘N missing’



3. From the obtained Summary table, it was observed that 1447 rows were completely blank and contained no value for any variable.

32/0 Cols	member_id	loan_amnt	term	int_rate	installment	grade	emp_title	emp_length	annua
29974	•	•	.	•	•				
29975	•	•	.	•	•				
29976	•	•	.	•	•				
29977	•	•	.	•	•				
29978	•	•	.	•	•				
29979	•	•	.	•	•				
29980	•	•	.	•	•				
29981	•	•	.	•	•				
29982	•	•	.	•	•				
29983	•	•	.	•	•				
29984	•	•	.	•	•				
29985	•	•	.	•	•				
29986	•	•	.	•	•				
29987	•	•	.	•	•				
29988	•	•	.	•	•				
29989	•	•	.	•	•				
29990	•	•	.	•	•				
29991	•	•	.	•	•				
29992	•	•	.	•	•				
29993	•	•	.	•	•				
29994	•	•	.	•	•				
29995	•	•	.	•	•				
29996	•	•	.	•	•				
29997	•	•	.	•	•				
29998	•	•	.	•	•				
29999	•	•	.	•	•				
30000	•	•	.	•	•				

4. On further analysis using various combinations of missing data patterns (**Tables → Missing Data Pattern**), it was observed that all the below variables were missing data in 1089 common rows:
    a. *acc_open_past_24mnths*
    b. *bc_open_to_buy*
    c. *percent_bc_gt_75*
    d. *bc_util*
    e. *mort_acc*
    f. *total_bal_ex_mort*
    g. *num_rev_accts*
    h. *total_cur_bal*



	Count	Number of columns missing	Patterns	acc_open_past_24mths	bc_open_to_buy	percent_bc_gt_75	bc_util	mort_acc	total_bal_ex_mort	num_rev_accts	tot_cur_bal
1	24207	0	00000000	0	0	0	0	0	0	0	0
2	3001	2	00000011	0	0	0	0	0	0	1	1
3	15	1	00010000	0	0	0	1	0	0	0	0
4	3	3	00010011	0	0	0	1	0	0	1	1
5	200	3	01110000	0	1	1	1	0	0	0	0
6	38	5	01110011	0	1	1	1	0	0	1	1
7	1089	8	11111111	1	1	1	1	1	1	1	1

4.2.1.2 Deletion

1. Based on the above observations, the 1447 rows that were missing data for all the variables were deleted as they provide no information and constitute less than 5% of the total sample. The updated version of the sample was saved as loan_sample_final_v1
2. The 1089 rows that were commonly missing the values for the below 8 variables were also deleted. The updated version of the sample was saved as loan_sample_final_v2
3. The below table clearly depicts the count of missing values in the
   a. initial sample
   b. sample after deleting the 1447 blank rows across all variables and
   c. sample after deleting the 1089 rows that were commonly missing in 8 variables

Variable	**Missing Values**		
	Initial Sample	*Sample after deleting the 1447 rows with no data*	*Sample after deleting the 1089 rows commonly missing in the 8 variables listed above in Page 9*
member_id	1447	0	0
loan_amnt	1447	0	0
funded_amnt	1447	0	0
term	1447	0	0
int_rate	1447	0	0
installment	1447	0	0
grade	1447	0	0
emp_title	3231	1784	1784
emp_length	1447	0	0
annual_inc	1447	0	0
is_inc_v	1455	8	8
loan_status	1447	0	0

purpose	1447	0	0
acc_open_past_24mths	2536	1089	0
bc_open_to_buy	2774	1327	238
percent_bc_gt_75	2774	1327	238
bc_util	2792	1345	256
delinq_2yrs	1447	0	0
mths_since_last_delinq	17,932	16,485	16,485
mort_acc	2536	1089	0
open_acc	1447	0	0
total_bal_ex_mort	2536	1089	0
revol_bal	1447	0	0
revol_util	1474	27	27
total_acc	1447	0	0
out_prncp	1447	0	0
total_pymnt_inv	1447	0	0
total_rec_prncp	1447	0	0
total_rec_int	1447	0	0
last_pymnt_amnt	1447	0	0
num_rev_accts	5578	4131	3042
tot_cur_bal	5578	4131	3042
policy_code	1447	0	0

4.2.1.3 Modification

For the remaining missing data, the respective steps taken are provided in the below table.

Field Name	Missing Values	Steps Taken
emp_title	1784	• These are nominal data with no values in common • Some values are invalid employee titles • Since there is no additional data available from the source, this variable may not be of any use for data analysis or modeling • Hence, no modifications were made to this field
is_inc_v	8	• We tried to predict the missing values in this field using the contingency tables by checking with the variables *loan_status* and *purpose* • No relation could be found between these variables and we replaced them with the nominal value "Unknown" so that we can perform analysis in the future using this nominal variable
bc_open_to_buy	238	• Values missing in same rows as *percent_bc_gt_75* • We did not impute this field since mean and median were too far and there was no significant correlation with any other field

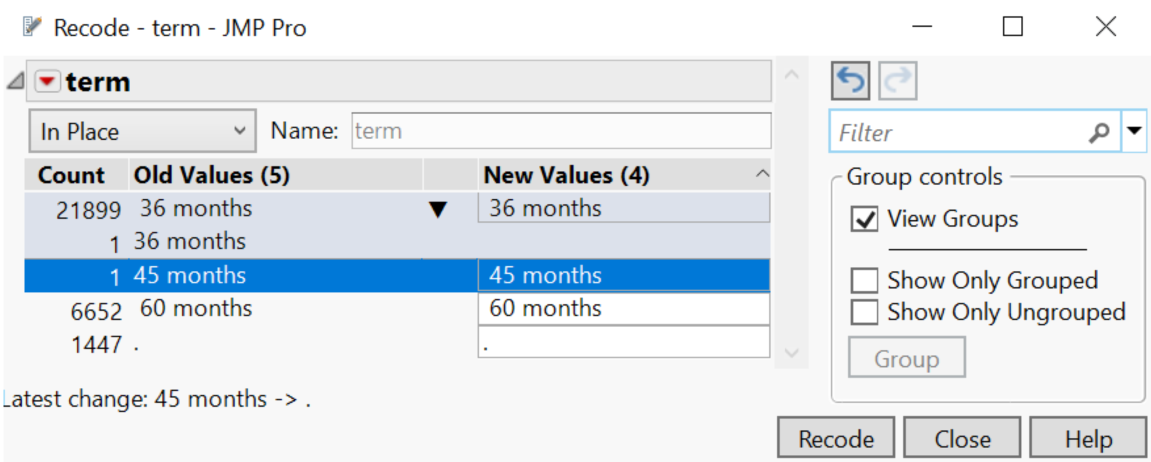
percent_bc_gt_75	238	• A new column 'percent_bc_gt_75 2' was added and the median(value=50) was used to fill the missing values using **Recode** • The median is used since the distribution is slightly right skewed • Values were missing in same rows as *bc_open_to_buy* • The screenshots and steps of this process are provided below
bc_util	256	• A new column *'bc_util2'* was added and the median (value=72.2) was used to fill the missing values using **Recode** • The median is used since distribution is right skewed • The screenshots and steps of this process are provided below
mths_since_last_delinq	16,485	• This variable is missing more than 55% of data from the sample • No additional data being available from the source, no modifications were made to this variable
revol_util	27	• A new column *'revol_util2'* was added and the median (0.604) was used to fill the missing values using **Recode** • The median is used since the distribution is skewed • The screenshots and steps of this process are provided below
num_rev_accts	3042	• A new column *'num_rev_accts 2'* was added and the median (64) was used to fill the missing values using **Recode** • The median is used since the distribution is right skewed • The screenshots and steps of this process are provided below

tot_cur_bal	3042	• Upon analysis, it was observed that if the number of mortgage accounts is 0, total balance except mortgage will be the same as the total current balance
		• Hence, the below formula was applied to get the missing values for total current balance field in the additional column created as *'tot_cur_bal_calc'*
		    ○ If mort_acc ==0 => total_bal_ex_mort
		    ○ Else => tot_cur_bal

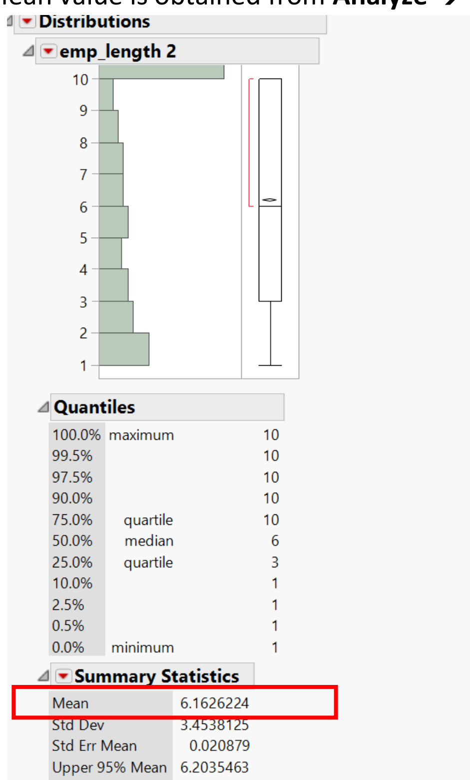
4.2.2 Resolve inconsistencies

The below inconsistencies were observed in the data set.

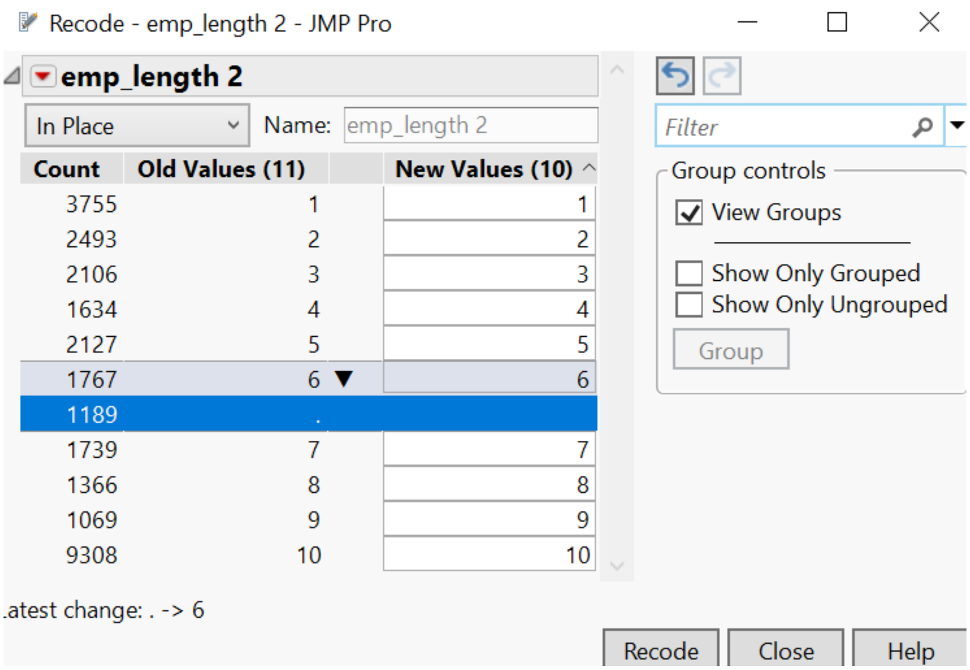
1. In the variable *'term'*,
   a. 1 record marked as 'NA' was replaced with the most occurring value '36 months' using **Recode**
   b. The value '45' was changed to '45 months' for data consistency using **Recode**



2. In the variable '*policy*', there was a record with value '22'. As per data dictionary, the valid value for policy is either 1 or 2. Assuming '22' to be a typo, we replaced it with '2'.
3. In the variable *'emp_length 2'*, the '.' Values were replaced with the mean
   a. Mean value is obtained from **Analyze → Distribution**



   b. Replace '.' value by Mean value using **Cols→Recode**



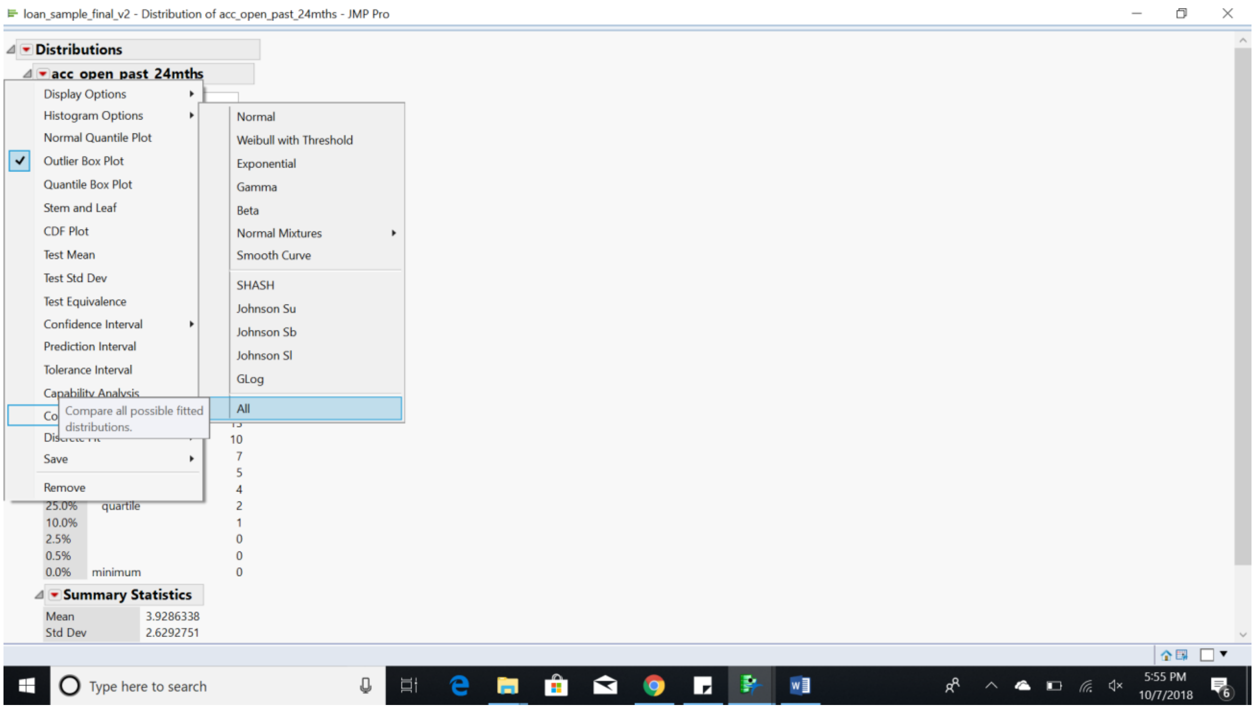
4.2.3 Outlier Detection & Analysis

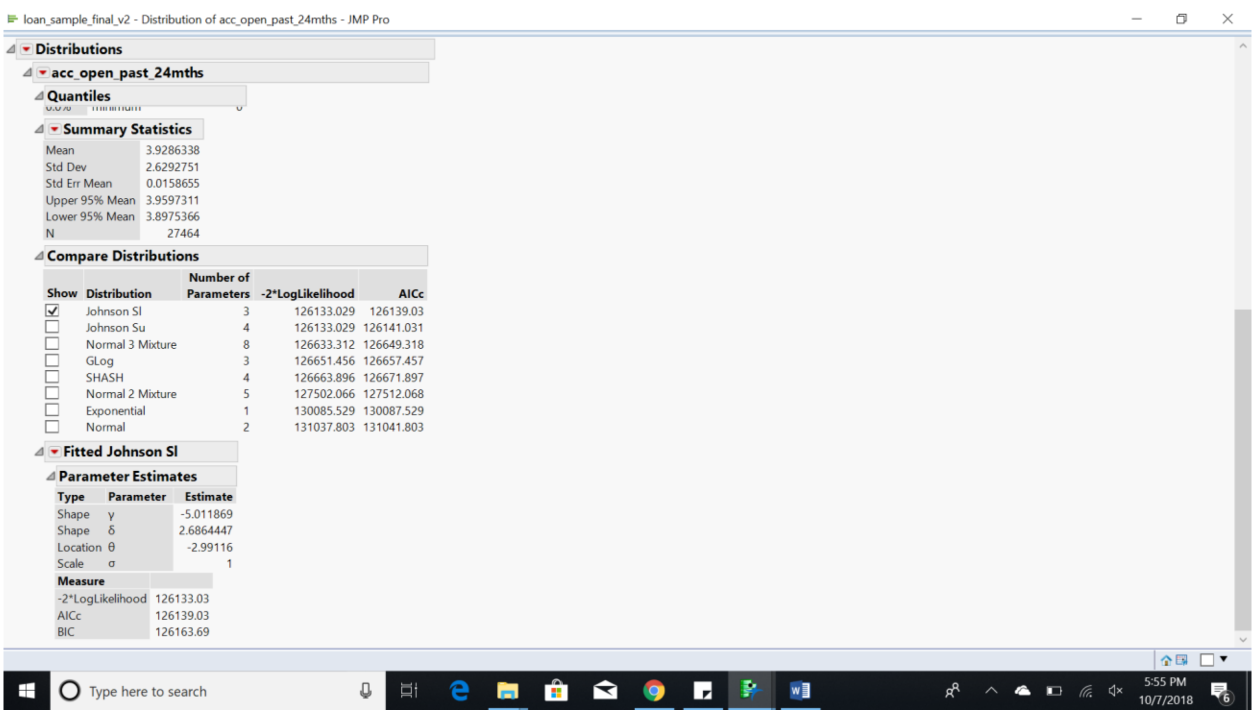
We analyzed the distribution on each continuous field to determine the number of outliers. Below are a few fields that we modified because of the presence of too many outliers.

Field Name	Outlier Analysis & Transformation
installment	• Applied continuous fit and Saved **Standardized - Gamma Distribution** as new column *'Standardized installments'* • The screenshots and steps of transformation are provided below
bc_open_to_buy	• Applied continuous fit and Saved **SHASH transform** as new column *'SHASH Transform bc_open_to_buy'* • The screenshots and steps of transformation are provided below
total_acc	• Applied continuous fit and Saved **Johnson SI transform** as new column *'Johnson SI Transform total_acc'* • The screenshots and steps of transformation are provided below
total_pymnt_inv	• Applied continuous fit and Saved **Johnson SI transform** as new column *'Johnson SI Transform total_pymnt_inv'* • The screenshots and steps of transformation are provided below
total_rec_prncp	• Applied continuous fit and Saved **Generalized Logarithm transform** as new column *'Generalized Logarithm Transform total_rec_prncp'* • The screenshots and steps of transformation are provided below
total_rec_int	• Applied continuous fit and Saved **Johnson SI transform** as new column *'Johnson SI Transform total_rec_int'* • The screenshots and steps of transformation are provided below
last_pymnt_amnt	• Applied continuous fit and Saved **Johnson SU transform** as new column *'Johnson SU Transform last_pymnt_amnt'* • The screenshots and steps of transformation are provided below

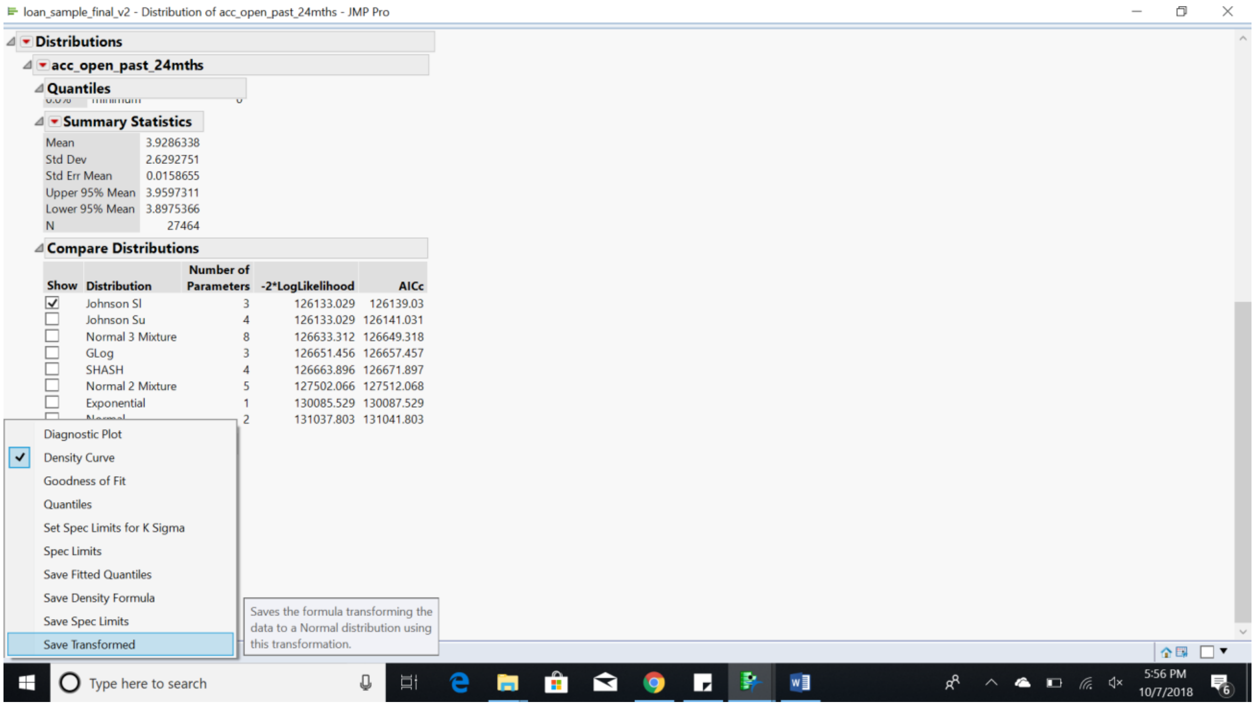
Steps for analyzing outlier and transforming the data to the distribution that fits the best

1. *Select a column and click on* ***Analyze → Distribution***
2. *Click the Red arrow and select* ***Continuous Fit → All***





3. *Save the selected fit*

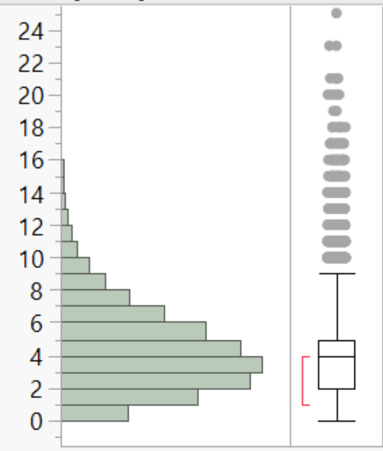

loan_sample_final_v2 - Distribution of acc_open_past_24mths - JMP Pro
Distributions
acc_open_past_24mths
Quantiles
Summary Statistics
Mean 3.9286338
Std Dev 2.6292751
Std Err Mean 0.0158655
Upper 95% Mean 3.9597311
Lower 95% Mean 3.8975366
N 27464
Compare Distributions
Show Distribution Number of Parameters -2*LogLikelihood AICc
Johnson Sl 3 126133.029 126139.03
Johnson Su 4 126133.029 126141.031
Normal 3 Mixture 8 126633.312 126649.318
GLog 3 126651.456 126657.457
SHASH 4 126663.896 126671.897
Normal 2 Mixture 5 127502.066 127512.068
Exponential 1 130085.529 130087.529
2 131037.803 131041.803
Diagnostic Plot
Density Curve
Goodness of Fit
Quantiles
Set Spec Limits for K Sigma
Spec Limits
Save Fitted Quantiles
Save Density Formula
Save Spec Limits
Save Transformed
Saves the formula transforming the data to a Normal distribution using this transformation.
Type here to search
5:56 PM
10/7/2018

Screenshots for outlier analysis and transformations

1. *Acc_open_past_24months*: Box-whisker plot

Distributions

acc_open_past_24mths



Quantiles

100.0%	maximum	31
99.5%		13
97.5%		10
90.0%		7
75.0%	quartile	5
50.0%	median	4
25.0%	quartile	2
10.0%		1
2.5%		0
0.5%		0
0.0%	minimum	0

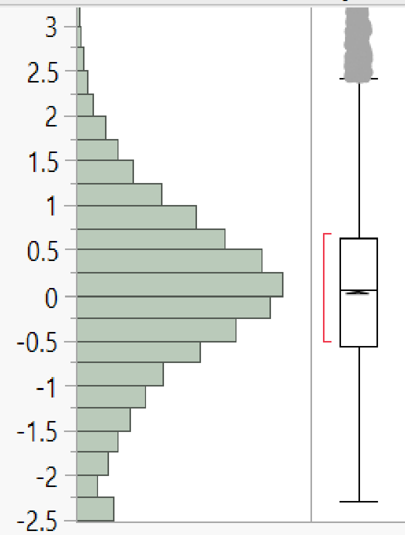
Summary Statistics

Mean	3.9286338
Std Dev	2.6292751
Std Err Mean	0.0158655
Upper 95% Mean	3.9597311
Lower 95% Mean	3.8975366
N	27464

2. *bc_open_to_buy*: SHASH Transform Box-whisker plot

Distributions

SHASH Transform bc_open_to_buy



Quantiles

100.0%	maximum	4.5526408726
99.5%		2.7142263153
97.5%		1.9659139705
90.0%		1.2251172151
75.0%	quartile	0.6374716507
50.0%	median	0.0576410924
25.0%	quartile	-0.555325929
10.0%		-1.300554924
2.5%		-2.166999723
0.5%		-2.28427418
0.0%	minimum	-2.28427418

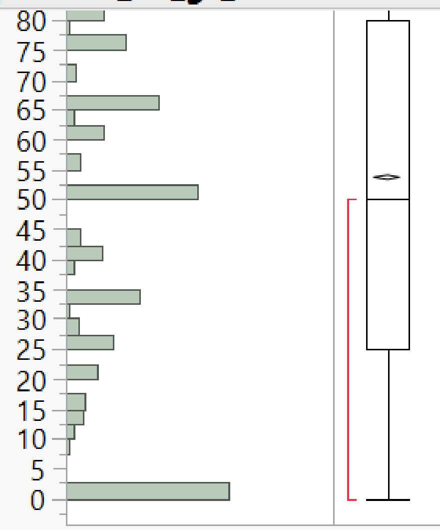
Summary Statistics

Mean	0.0230426
Std Dev	0.9850684
Std Err Mean	0.00597
Upper 95% Mean	0.0347441
Lower 95% Mean	0.0113411
N	27226

3. *percent_bt_gt_75*: Box-whisker plot

Distributions

percent_bc_gt_75



Quantiles

100.0%	maximum	100
99.5%		100
97.5%		100
90.0%		100
75.0%	quartile	80
50.0%	median	50
25.0%	quartile	25
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

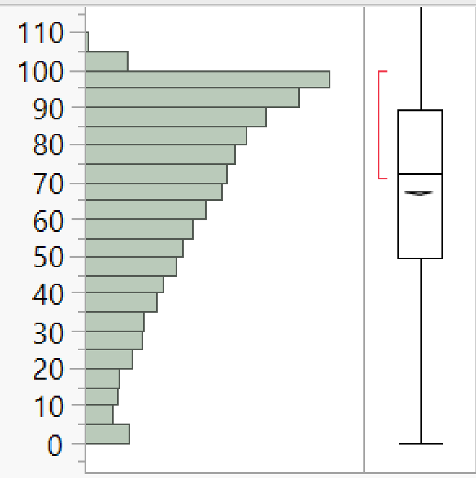
Summary Statistics

Mean	53.762001
Std Dev	34.197715
Std Err Mean	0.2072551
Upper 95% Mean	54.168231
Lower 95% Mean	53.35577
N	27226

4. *bc_util*: Box-whisker plot

Distributions

bc_util



Quantiles

100.0%	maximum	165.7
99.5%		102.2
97.5%		99.7
90.0%		96.5
75.0%	quartile	89.1
50.0%	median	72.2
25.0%	quartile	49.5
10.0%		27.4
2.5%		6.9
0.5%		0
0.0%	minimum	0

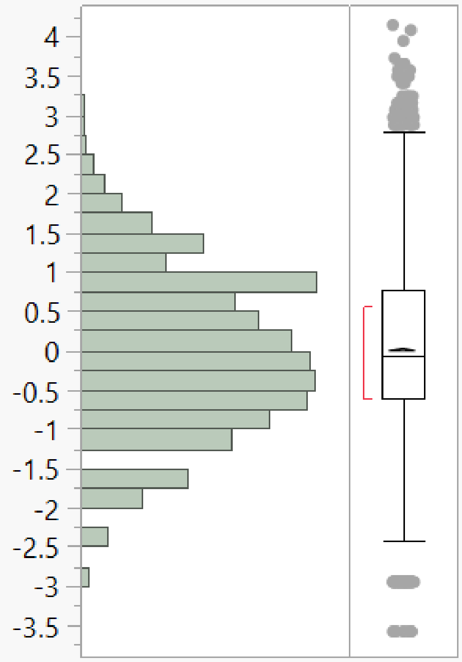
Summary Statistics

Mean	66.941951
Std Dev	26.098616
Std Err Mean	0.1582228
Upper 95% Mean	67.252076
Lower 95% Mean	66.631826
N	27208

5. *open_acc*: Johnson Sl Transform Box-whisker plot

Distributions

Johnson Sl Transform open_acc



Quantiles

100.0%	maximum	4.1447116175
99.5%		2.4574879284
97.5%		1.9835064025
90.0%		1.2746597155
75.0%	quartile	0.7617597362
50.0%	median	-0.077273518
25.0%	quartile	-0.594034421
10.0%		-1.207779127
2.5%		-1.963519775
0.5%		-2.418678673
0.0%	minimum	-3.578056139

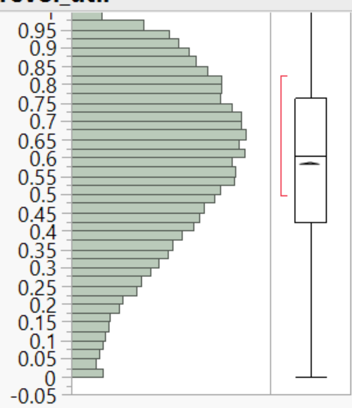
Summary Statistics

Mean	4.069e-15
Std Dev	1.0000175

6. *revol_util*: Box-whisker plot

Distributions

revol_util



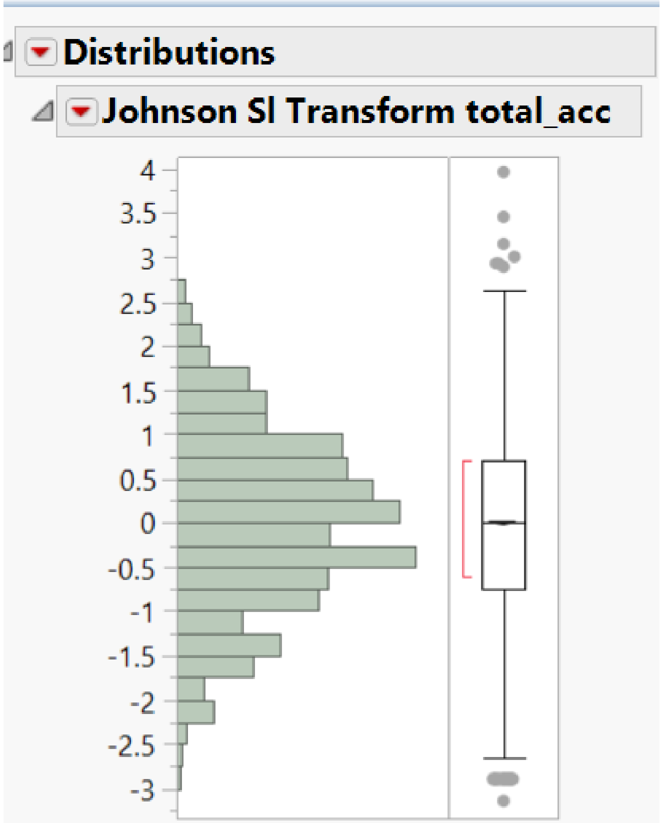
Quantiles

100.0%	maximum	1.225
99.5%		0.981365
97.5%		0.95
90.0%		0.8753
75.0%	quartile	0.762
50.0%	median	0.604
25.0%	quartile	0.426
10.0%		0.264
2.5%		0.092
0.5%		0.012
0.0%	minimum	0

Summary Statistics

Mean	0.5834921
Std Dev	0.2284021
Std Err Mean	0.0013523
Upper 95% Mean	0.5861428
Lower 95% Mean	0.5808415
N	28526

7. *total_acct*: Johnson Sl Transform Box-whisker plot



Quantiles

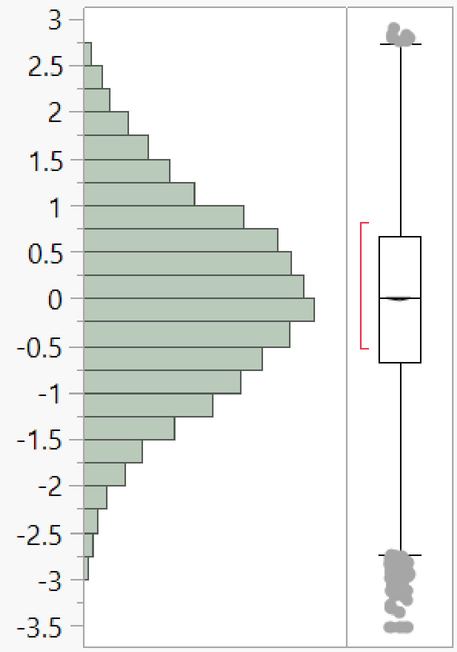
100.0%	maximum	3.9622081425
99.5%		2.4349910184
97.5%		1.9549459912
90.0%		1.3267985082
75.0%	quartile	0.700814379
50.0%	median	0.0173763093
25.0%	quartile	-0.731393023
10.0%		-1.254620035
2.5%		-1.881237593
0.5%		-2.448412634
0.0%	minimum	-3.139205739

Summary Statistics

8. total_pymnt_inv: Johnson Sl Transform Box-whisker plot

Distributions

Johnson Sl Transform total_pymnt_inv



Quantiles

100.0%	maximum	2.894813127
99.5%		2.4539217323
97.5%		1.9758368284
90.0%		1.2960864768
75.0%	quartile	0.68502328
50.0%	median	-0.000312165
25.0%	quartile	-0.684993767
10.0%		-1.280789084
2.5%		-1.957464634
0.5%		-2.60724114
0.0%	minimum	-3.514853933

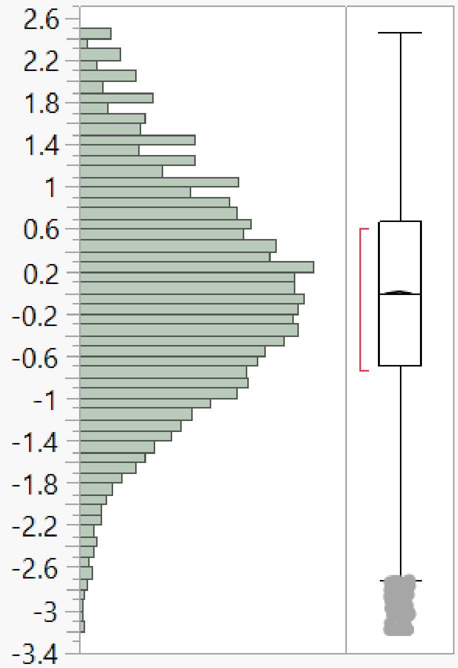
Summary Statistics

Mean	7.539e-15
Std Dev	1.0000175

9. *Total_rec_prncp*: Generalized Logarithm Transform Box-whisker plot

Distributions

Generalized Logarithm Transform total_rec_prncp



Quantiles

100.0%	maximum	2.4717841582
99.5%		2.4380514063
97.5%		2.0327558259
90.0%		1.3195629945
75.0%	quartile	0.676318123
50.0%	median	-0.013936406
25.0%	quartile	-0.682995546
10.0%		-1.263080022
2.5%		-1.957229254
0.5%		-2.648026599
0.0%	minimum	-3.185934994

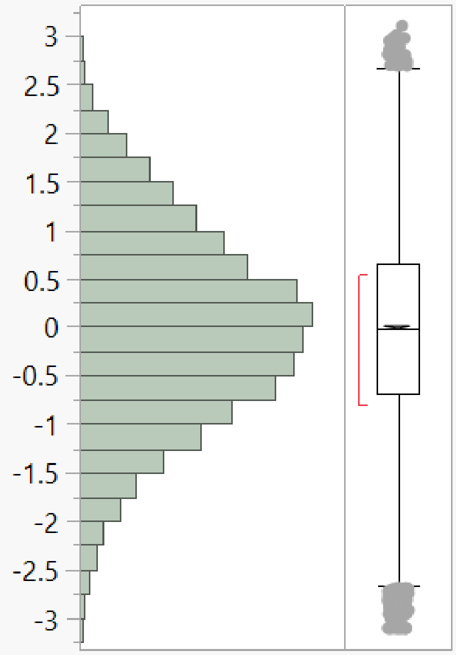
Summary Statistics

Mean	1.914e-15

10. *total_rec_int*: Johnson SI Transform Box-whisker plot

Distributions

Johnson Sl Transform total_rec_int



Quantiles

100.0%	maximum	3.1026533481
99.5%		2.4105732281
97.5%		1.9377306311
90.0%		1.3331654893
75.0%	quartile	0.6675239854
50.0%	median	-0.002401037
25.0%	quartile	-0.673209195
10.0%		-1.267132158
2.5%		-1.997058995
0.5%		-2.593001928
0.0%	minimum	-3.10040521

Summary Statistics

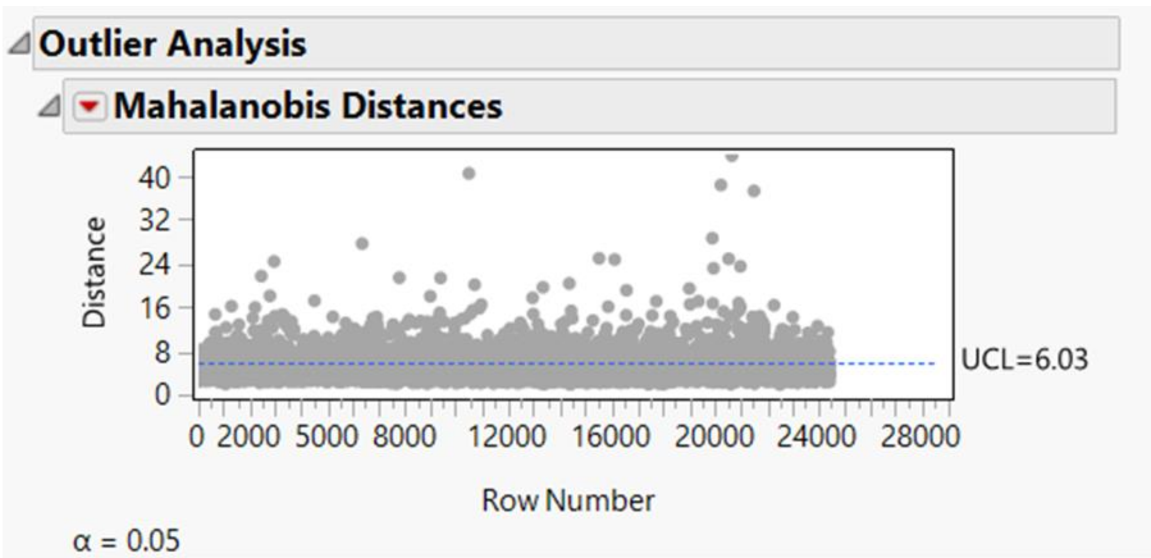
Mean	4.504e-14
Std Dev	1.0000175

Multivariate Outlier Analysis

To consider the potential outliers with respect to other variables we have done multivariate outlier analysis. We used Mahalanobis Distances to check for the potential outliers.

We included all continuous variables except the variables member_id and policy_code for the analysis.

We can see that the points above the UCL can be considered as potential outliers. But we have to analyze various other parameters to actually eliminate them which depends on other factors and depends on the impact of the variable that we need to predict. So, we haven't excluded any of these from our dataset.



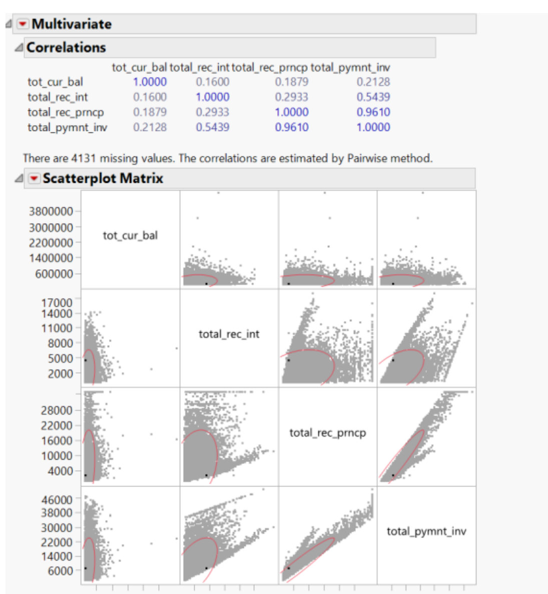
4.3 Data Reduction

4.3.1 Correlation

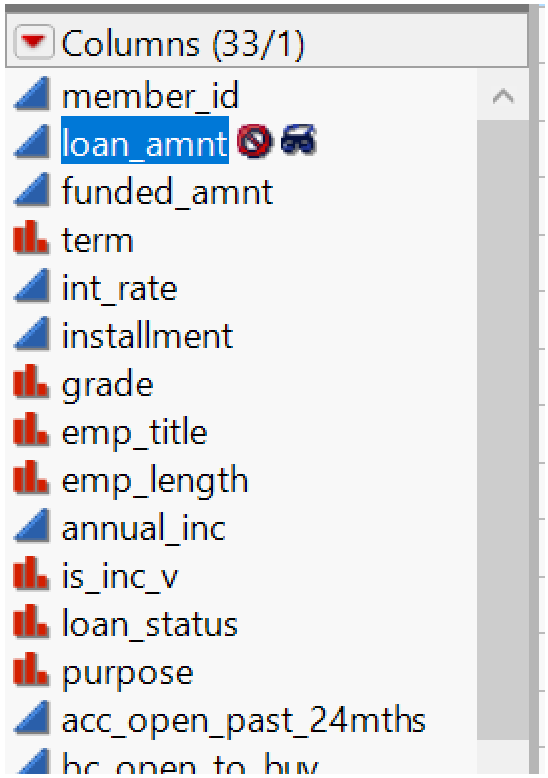
- The Multivariate Correlation was analyzed for all the continuous variables using **Analyze → Multivariate Methods → Multivariate**.
- The correlation matrix obtained is displayed below

Row	member_i	loan_amn	funded_ai	installmer	annual_in	acc_open_	bc_open_t	percent_b	bc_util	delinq_2yı	mths_sinc	mort_acc	open_acc	total_bal_	revol_bal	total_acc	out_prncp	total_pym	total_rec_	total_rec_	last_pymr	num_rev_
member_i	1	0.035204	0.035666	0.019122	0.01645	0.005085	-0.00314	-0.01415	-0.00378	0.051286	-0.05715	0.036668	0.023712	0.020711	-0.00465	0.030779	0.346569	-0.35566	-0.32998	-0.22063	-0.10434	-0.01705
loan_amn	0.035204	1	0.999697	0.954852	0.419536	-0.00716	0.189127	0.001631	0.030875	0.016215	-0.04097	0.238367	0.200949	0.280554	0.317395	0.241547	0.680699	0.635963	0.48026	0.738688	0.230492	0.192785
funded_ai	0.035666	0.999697	1	0.955214	0.419468	-0.00724	0.189038	0.001715	0.030894	0.016198	-0.04072	0.238219	0.200874	0.280402	0.317475	0.241324	0.681302	0.635737	0.480022	0.738494	0.230026	0.192696
installmer	0.019122	0.954852	0.955214	1	0.416077	0.003343	0.150531	0.032969	0.067775	0.026882	-0.03916	0.19745	0.196317	0.272237	0.313482	0.222291	0.59643	0.661363	0.527144	0.682449	0.231008	0.186368
annual_in	0.01645	0.419536	0.419468	0.416077	1	0.046757	0.200223	-0.04379	-0.02629	0.087035	-0.07823	0.323142	0.182724	0.425191	0.35223	0.266972	0.262423	0.29809	0.253199	0.262878	0.125409	0.141436
acc_open_	0.005085	-0.00716	-0.00724	0.003343	0.046757	1	0.056149	-0.11044	-0.11694	-0.0621	0.121088	0.066934	0.437191	0.163022	-0.03387	0.379094	-0.03812	0.016121	0.009864	0.02183	0.056364	0.290013
bc_open_t	-0.00314	0.189127	0.189038	0.150531	0.200223	0.056149	1	-0.48356	-0.59754	-0.03633	-0.02655	0.151311	0.255989	0.11335	0.204709	0.211052	0.090792	0.138385	0.163771	-0.01981	0.080441	0.288223
percent_b	-0.01415	0.001631	0.001715	0.032969	-0.04379	-0.11044	-0.48356	1	0.832625	-0.02563	0.041574	-0.04729	-0.09716	0.04306	0.078444	-0.08406	0.030255	-0.00592	-0.05486	0.145626	-0.03246	-0.14328
bc_util	-0.00378	0.030875	0.030894	0.067775	-0.02629	-0.11694	-0.59754	0.832625	1	-0.0162	0.042921	-0.04463	-0.09123	0.067251	0.111896	-0.08086	0.059361	0.006238	-0.04863	0.169076	-0.03699	-0.14658
delinq_2yı	0.051286	0.016215	0.016198	0.026882	0.087035	-0.0621	-0.03633	-0.02563	-0.0162	1	-0.60712	0.106049	0.055412	0.047602	-0.02184	0.133455	0.038744	-0.0136	-0.02757	0.035546	-0.00607	0.090904
mths_sinc	-0.05715	-0.04097	-0.04072	-0.03916	-0.07823	0.121088	-0.02655	0.041574	0.042921	-0.60712	1	-0.09504	-0.03007	-0.0407	-0.01863	-0.06459	-0.05949	0.006241	0.020487	-0.03776	0.004291	-0.04876
mort_acc	0.036668	0.238367	0.238219	0.19745	0.323142	0.066934	0.151311	-0.04729	-0.04463	0.106049	-0.09504	1	0.132286	0.160393	0.1922	0.426387	0.170097	0.141934	0.122179	0.118946	0.077764	0.220646
open_acc	0.023712	0.200949	0.200874	0.196317	0.182724	0.437191	0.255989	-0.09716	-0.09123	0.055412	-0.03007	0.132286	1	0.397503	0.220627	0.664781	0.146607	0.117687	0.089167	0.135776	0.050587	0.614058
total_bal_	0.020711	0.280554	0.280402	0.272237	0.425191	0.163022	0.11335	0.04306	0.067251	0.047602	-0.0407	0.160393	0.397503	1	0.55327	0.422153	0.190671	0.177517	0.139285	0.187966	0.073865	0.115714
revol_bal	-0.00465	0.317395	0.317475	0.313482	0.35223	-0.03387	0.204709	0.078444	0.111896	-0.02184	-0.01863	0.1922	0.220627	0.55327	1	0.197997	0.211652	0.198939	0.158187	0.205301	0.053702	0.20313
total_acc	0.030779	0.241547	0.241324	0.222291	0.266972	0.379094	0.211052	-0.08406	-0.08086	0.133455	-0.06459	0.426387	0.664781	0.422153	0.197997	1	0.152886	0.164103	0.139925	0.143068	0.099237	0.767786
out_prncp	0.346569	0.680699	0.681302	0.59643	0.262423	-0.03812	0.090792	0.030255	0.059361	0.038744	-0.05949	0.170097	0.146607	0.190671	0.211652	0.152886	1	-0.01769	-0.20502	0.581379	-0.34792	0.116432
total_pym	-0.35566	0.635963	0.635737	0.661363	0.29809	0.016121	0.138385	-0.00592	0.006238	-0.0136	0.006241	0.141934	0.117687	0.177517	0.198939	0.164103	-0.01769	1	0.961014	0.543853	0.713802	0.143373
total_rec_	-0.32998	0.48026	0.480022	0.527144	0.253199	0.009864	0.163771	-0.05486	-0.04863	-0.02757	0.020487	0.122179	0.089167	0.139285	0.158187	0.139925	-0.20502	0.961014	1	0.293311	0.803302	0.1257
total_rec_	-0.22063	0.738688	0.738494	0.682449	0.262878	0.02183	-0.01981	0.145626	0.169076	0.035546	-0.03776	0.118946	0.135776	0.187966	0.205301	0.143068	0.581379	0.543853	0.293311	1	0.029855	0.110374
last_pymr	-0.10434	0.230492	0.230026	0.231008	0.125409	0.056364	0.080441	-0.03246	-0.03699	-0.00607	0.004291	0.077764	0.050587	0.073865	0.053702	0.099237	-0.34792	0.713802	0.803302	0.029855	1	0.064538
num_rev_	-0.01705	0.192785	0.192696	0.186368	0.141436	0.290013	0.288223	-0.14328	-0.14658	0.090904	-0.04876	0.220646	0.614058	0.115714	0.20313	0.767786	0.116432	0.143373	0.1257	0.110374	0.064538	1
tot_cur_b	-0.00115	0.315994	0.315949	0.28307	0.550163	0.091973	0.170841	-0.01381	0.000938	0.084359	-0.08672	0.528805	0.242224	0.498818	0.419478	0.321912	0.207573	0.212802	0.187878	0.160027	0.09581	0.137084
policy_coc	0.01075	0.011946	0.011953	0.009002	-0.0003	0.004754	0.00427	-0.00731	-0.00447	-0.00207	0	-0.00223	0.009042	0.004565	0.000609	0.006155	0.016749	-0.00214	-0.00388	0.004511	-0.00166	0.007858

- From the above matrix, it can be observed that *loan_amnt* and *funded_amnt* have a correlation of 0.9997
- *bc_util* and *percent_bc_gt_75* have a correlation of 0.8326
- *total_rec_prncp* and *total_pymnt_inv* have a correlation of 0.9610



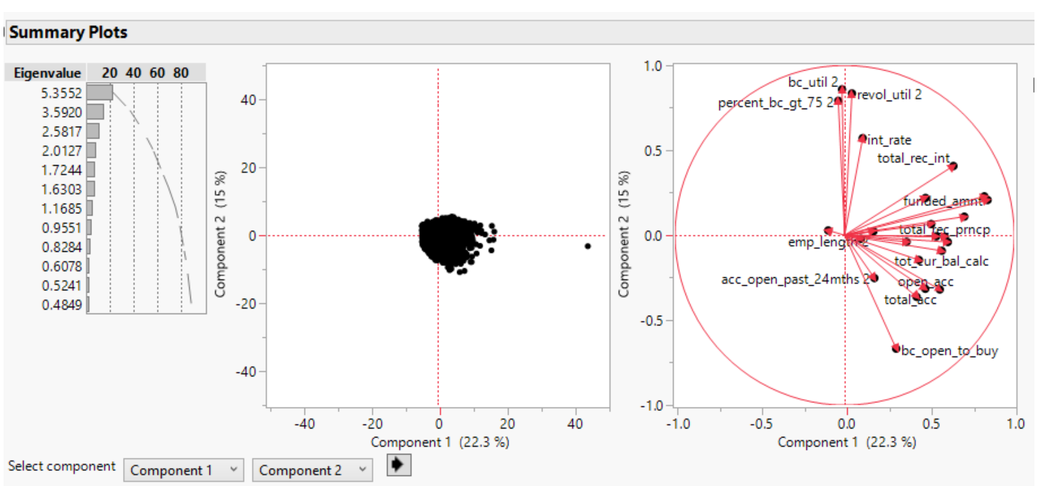
1. Between *loan_amnt* and *funded_amnt*, we decided to hide and exclude the *loan_amnt* column and retain only the *funded_amnt* column



2. Though *percent_bc_gt_75* and *bc_util* have a correlation of 0.8326, we decided to retain both the fields and reduce dimensionality using Principal Component Analysis

4.3.2 Principal Component Analysis

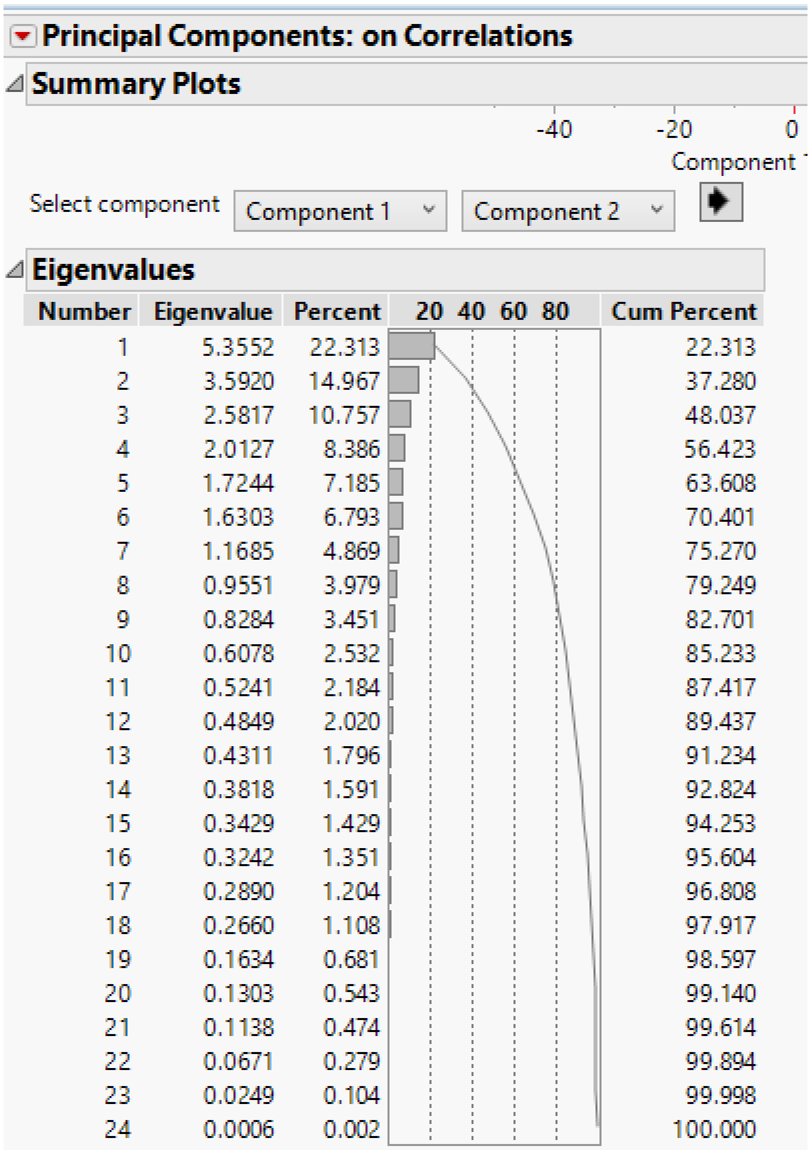
- We performed principle component analysis for the cleaned dataset. The screenshots of the Summary Plots and Eigen Values are provided below.



We excluded the following fields from the analysis

- member_id: This variable is excluded as it doesn't provide any information for building the model since this is unique at row level. (all unique records)

- policy_code: This variable had only one record with value 2 and all other rows are all value 1. So this is considered very insignificant for PCA.
- mths_since_last_delinq: This variable has almost 60% of data missing. So this is excluded from the PCA analysis since this would impact the analysis and cause missing values.



Principal Components: on Correlations

Summary Plots

-40 -20 0

Component

Select component Component 1 Component 2

Eigenvalues

Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	5.3552	22.313		22.313
2	3.5920	14.967		37.280
3	2.5817	10.757		48.037
4	2.0127	8.386		56.423
5	1.7244	7.185		63.608
6	1.6303	6.793		70.401
7	1.1685	4.869		75.270
8	0.9551	3.979		79.249
9	0.8284	3.451		82.701
10	0.6078	2.532		85.233
11	0.5241	2.184		87.417
12	0.4849	2.020		89.437
13	0.4311	1.796		91.234
14	0.3818	1.591		92.824
15	0.3429	1.429		94.253
16	0.3242	1.351		95.604
17	0.2890	1.204		96.808
18	0.2660	1.108		97.917
19	0.1634	0.681		98.597
20	0.1303	0.543		99.140
21	0.1138	0.474		99.614
22	0.0671	0.279		99.894
23	0.0249	0.104		99.998
24	0.0006	0.002		100.000

The above analysis shows the cumulative percentages of information captured by the principal components. We can actually go ahead and use these principal components to build our model. Though the question of how many components are to be included is always there and completely depends on the expected accuracy and business needs, we would recommend to use 18 variables as they constitute almost 98% of the available information from all the original continuous variables.

5.0 Updated Data dictionary

In the updated loan_data_dictionary we have updated the column "BrowseNotesFile". In this we have stated whether the variable is included in the dataset or not.

We have also added some extra variables in the dataset and their respective descriptions.

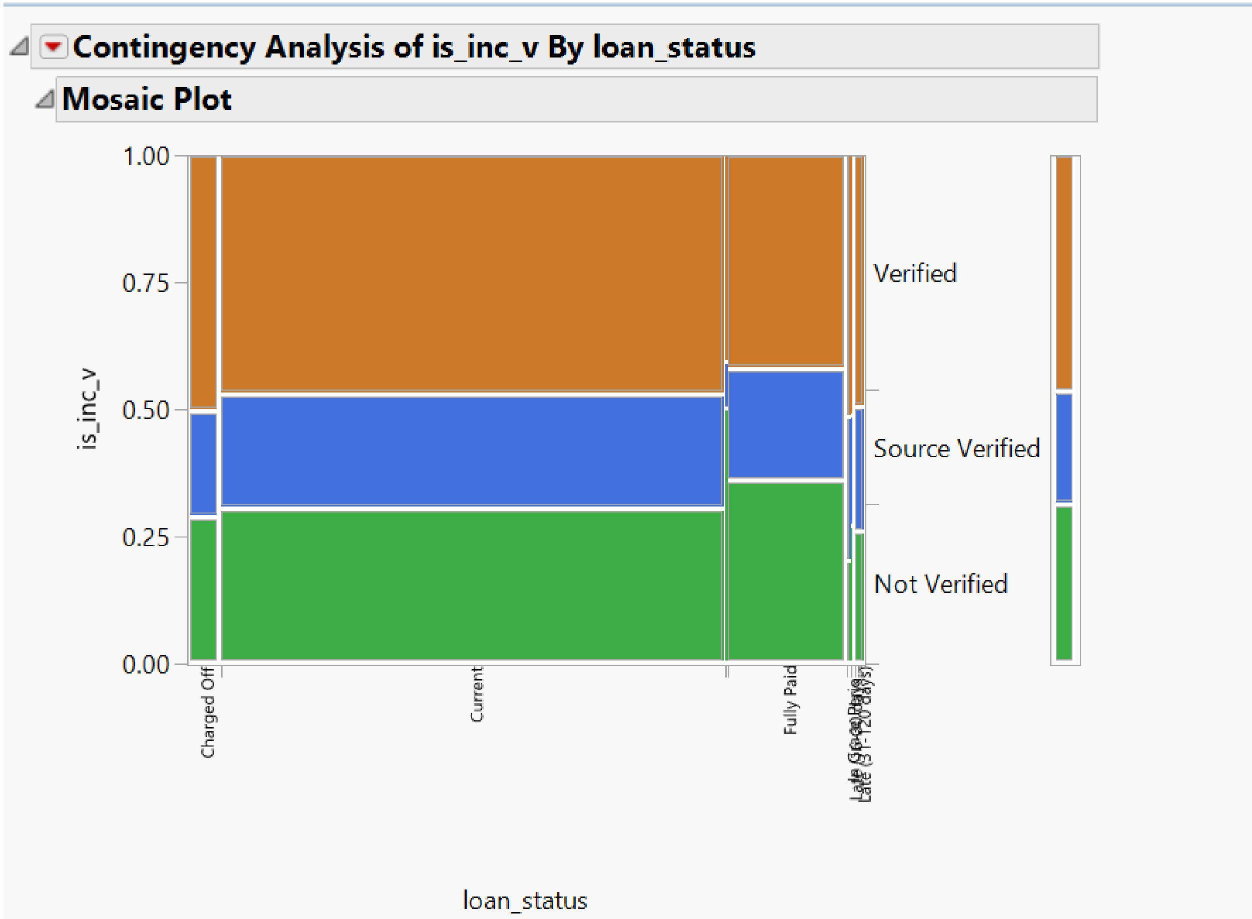
Added variables in the data set:

Field added	**Field Description**
acc_open_past_24mths 2	Imputed column for acc_open_past_24mths
bc_util 2	Included column for bc_util
emp_length 2	Imputed column for emp_length
num_rev_accts 2	Imputed column for num_rev_accts
percent_bc_gt_75 2	Imputed column for percent_bc_gt_75
revol_util 2	Imputed column for revol_util
tot_cur_bal_calc	Imputed column for tot_cur_bal
SHASH Transform bc_open_to_buy	Transformed variable for bc_open_to_buy
Johnson Sl Transform open_acc	Transformed variable for open_acc
Johnson Sl Transform total_acc	Transformed varaible for total_acc
Generalized Logarithm Transform total_rec_prncp	Transformed variable for total_rec_prncp
Johnson Sl Transform total_rec_int	Transformed variable for total_rec_int
Johnson Sl Transform total_pymnt_inv	Transformed variable for total_pymnt_inv
Johnson Su Transform last_pymnt_amnt	Transformed variable for last_pymnt_amnt
Std installment	Standerdized column for installment

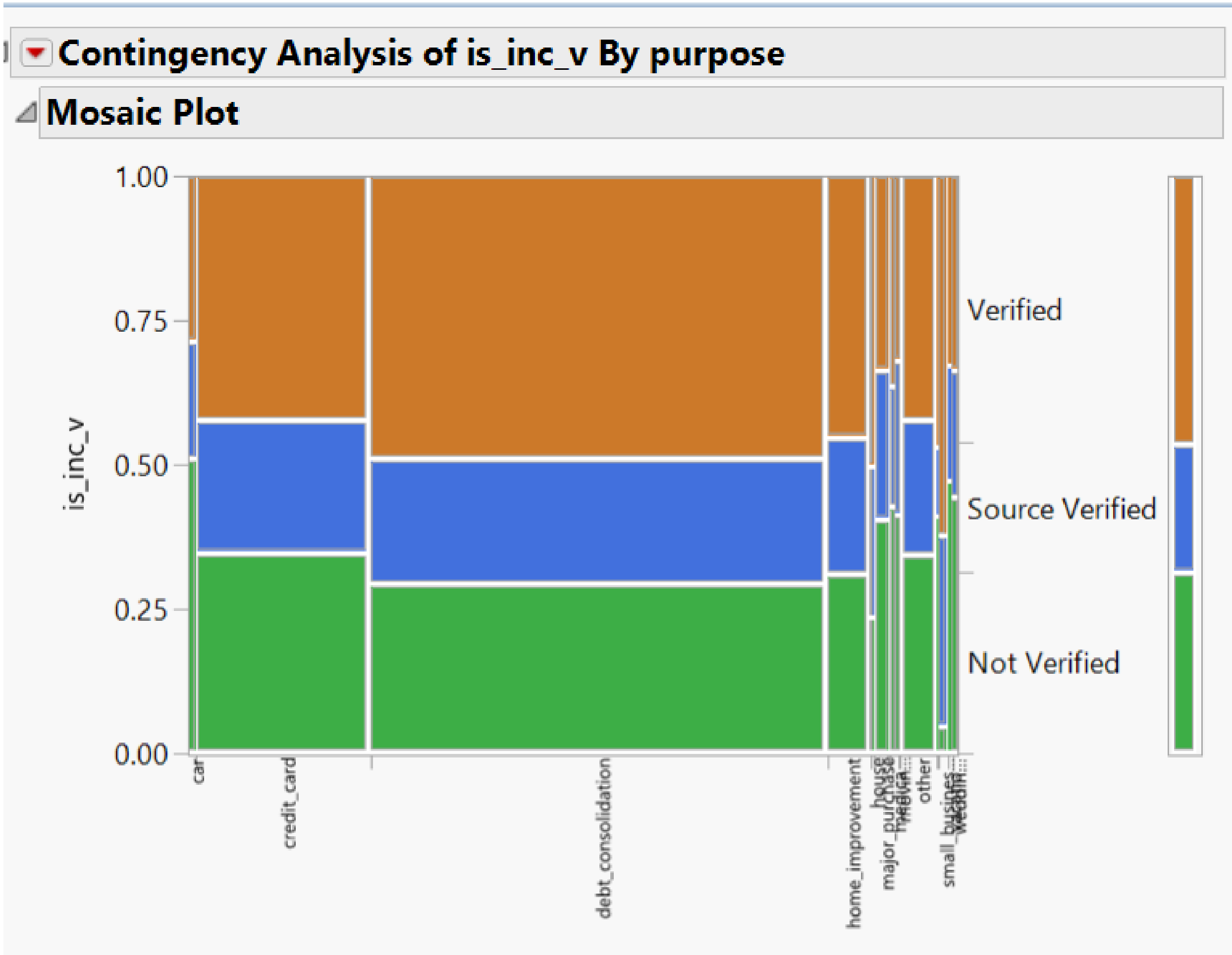
6.0 Appendix

This section includes trials that did not prove fruitful in our analysis

1. We tried to predict the missing values in this field using the contingencies tables by checking with the variables loan_status and the variable purpose
    a. Is_inc_v & loan status – Mosaic plot – no relation



    b. Is_inc_v & purpose – Mosaic plot – no actual relation


Contingency Analysis of is_inc_v By purpose
Mosaic Plot
1.00
0.75
0.50
0.25
0.00
is_inc_v
car
credit_card
debt_consolidation
home_improvement
house
major_purchase
other
Verified
Source Verified
Not Verified