FOUNDATION FOR ORGANISATIONAL RESEARCH AND EDUCATION

NEW DELHI

Academic Session 2023-2025

Customer Segmentation (Loan Data) on the basis of Risk

Machine Learning for Managers

FMG 32 Section A

Submitted to:

Prof. Amarnath Mitra

Submitted by:

321035 - Prityush Agarwal

Table of Contents

1. Project Objective

→ The first objective is to segment the consumer data of the bank using unsupervised learning -algorithms using K-means clustering.

→ The second objective is to identify the number of appropriate clusters using a performance matrix (Silhouette score).

→ The third objective is to determine the characteristics of each cluster (to sell the product/service).

2. Data Description

2.1 Dimension of Data

2.1.1 Number of Variables: The number of variables in the csv file is 30.

2.1.2 Number of records: The number of records in the csv file is 1,06,485 (excluding naming column).

2.2 Description of variables

2.2.1 Index variables: id – gives the loan a unique identification (year, issue_d and final_d won't be used for evaluation purpose another variable term is being used to gauge how much time it took to repay the loan).

2.2.2 Variables having categorical or non-categorical variables

2.2.2.1 Variables or Features having Nominal Categories:

→home_ownership - home ownership status provided by the borrower during registration

→term – Term of the loan

→application_type – Explains the status whether the account is individual or joint

→purpose – This variable tells the purpose why the loan was taken

→loan_condition – This variable tells the status of the loan whether the loan is good or bad

region – The region the loan was taken from

2.2.2.2 Variables or Features having Ordinal Categories:

→income_category – This variable tells the bracket under which the person earns

→interest_payments – This variable tells whether the interest payments on the loan is low or high

→grade – This variable tells the assigned grade of the loan

2.2.2.3 Non-Categorical Variables:

→emp_length_int – The number of years the person is employed

→annual_inc – This variable tells the annual income the person earns

→loan_amount – The variable tells the amount of loan that has been taken by the person

→interest_rate – The variable tells the interest rate at which the loan needs to be paid

→dti - A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested loan, divided by the borrower's self-reported monthly income.

→total_pymnt – This variable explains the total payment done against the loan

→total_rec_prncp – This variable explains the total received principal gotten from the loan

→recoveries – This variable explains the recoveries made from the bad loan

→installment – This variable explains the instalment made against the loan

2.3 Descriptive Statistics

2.3.1 Descriptive Statistics: Categorical Variables

2.3.1.1 home_ownership

Row ID	count	Relativ...
MORTGAGE	53079	49.846
NONE	5	0.005
OTHER	24	0.023
OWN	10502	9.862
RENT	42875	40.264

Term

Row ID	count	Relative Frequency in %
36 months	74473	69.938
60 months	32012	30.062

application_type

Row ID	count	Relative Frequency in %
INDIVIDUAL	106427	99.946
JOINT	58	0.054

Purpose

Row ID	count	Relative Frequency in %
car	1043	0.979
credit_card	24716	23.211
debt_consolidation	63092	59.25
educational	50	0.047
home_improvement	6193	5.816
house	433	0.407
major_purchase	2023	1.9
medical	1007	0.946
moving	661	0.621
other	5164	4.85
renewable_energy	67	0.063
small_business	1188	1.116
vacation	585	0.549
wedding	263	0.247

loan_condition

Row ID	count	Relative Frequency in %
Bad Loan	8084	7.592
Good Loan	98401	92.408

income_category

Row ID	count	Relative Frequency in %
High	2020	1.897
Low	87400	82.077
Medium	17065	16.026

interest_payments

Row ID	count	Relative Frequency in %
High	50610	47.528
Low	55875	52.472

Grade

Row ID	count	Relative Frequency in %
A	17753	16.672
B	30642	28.776
C	29537	27.738
D	16703	15.686
E	8487	7.97
F	2729	2.563
G	634	0.595

2.3.2 Descriptive Statistics: Non-Categorical Variables

2.3.2.1 Measures of Central Tendency and Dispersion

Name	Type	# Missing valu...	# Unique values	Minimum	Maximum	25% Quantile	50% Quantile (...	75% Quantile	Mean	Mean Absolut...	Standard Devi...	Skewness	Kurtosis
emp_length_int	Number (doubl...	0	12	0.5	10	3	6.05	10	6.054	3.115	3.508	-0.209	10.711
annual_inc	Number (doubl...	0	9543	3,000	7,500,000	45,000	65,000	90,000	75,216.112	31,824.916	62,577.359	34.794	-22,560.676
loan_amount	Number (doubl...	0	1316	500	35,000	8,000	13,000	20,000	14,786.903	6,867.874	8,427.684	0.677	1.924
interest_rate	Number (doubl...	0	480	5.32	28.99	9.99	12.99	16.2	13.235	3.512	4.381	0.428	1.186
dti	Number (doubl...	0	3996	0	104	11.9	17.71	23.98	18.159	6.823	8.323	0.238	2.998
total_pymnt	Number (doubl...	0	84359	0	55,145.01	1,918.7	4,868.8	10,577.075	7,558.313	5,891.31	7,894.445	1.8	-26.929
total_rec_prncp	Number (doubl...	0	52999	0	35,000.03	1,203.92	3,200	7,999.975	5,761.855	4,889.643	6,648.441	1.94	-29.24
recoveries	Number (doubl...	0	2944	0	31,900.52	0	0	0	45.972	89.481	415.898	19.633	-5,618.971
installment	Number (doubl...	0	28257	15.67	1,445.46	261.65	382.87	573.35	437.586	193.484	244.202	0.938	-5.098

Table "default" - Rows: 9 Spec - Columns: 16 Properties Flow Variables

Row ID	Column	Min	Max	Mean	Std. deviation	Variance	Skewness	Kurtosis	Overall sum	No. missings	No. NaNs	No. +oos	No. -oos	Median	Row count	Histogram
emp_length_int	emp_length_int	0.5	10	6.054	3.508	12.304	-0.209	-1.461	644,634.9	0	0	0	0	?	106485	0 10
annual_inc	annual_inc	3,000	7,500,000	75,216.112	62,577.359	3,915,925,8...	34.795	3,076.887	8,009,387,668	0	0	0	0	?	106485	3,000 7,500,000
loan_amount	loan_amount	500	35,000	14,786.903	8,427.684	71,025,850....	0.677	-0.262	1,574,583,350	0	0	0	0	?	106485	500 35,000
interest_rate	interest_rate	5.32	28.99	13.235	4.381	19.196	0.428	-0.162	1,409,325.24	0	0	0	0	?	106485	5 29
dti	dti	0	104	18.159	8.323	69.271	0.238	-0.409	1,933,637.52	0	0	0	0	?	106485	0 104
total_pymnt	total_pymnt	0	55,145.01	7,558.313	7,894.445	62,322,261....	1.8	3.673	804,846,93...	0	0	0	0	?	106485	0 55,145
total_rec_prncp	total_rec_prncp	0	35,000.03	5,761.855	6,648.441	44,201,766....	1.94	3.988	613,551,10...	0	0	0	0	?	106485	0 35,000
recoveries	recoveries	0	31,900.52	45.972	415.898	172,971.14	19.633	766.331	4,895,331.188	0	0	0	0	?	106485	0 31,901
installment	installment	15.67	1,445.46	437.586	244.202	59,634.835	0.938	0.695	46,596,375.72	0	0	0	0	?	106485	16 1,445

Source of Data

Link of the data: https://www.kaggle.com/datasets/mrferozi/loan-data-for-dummy-bank

3. Analysis

3.1 Data Pre-Processing

3.1.1 Missing Data Statistics and Treatment

3.1.1.1 Missing Data Statistics: 0

3.1.1.2 Missing Data Treatment: 0

3.1.1.2.1 Removal of Records with More Than 50% Missing Data: None

3.1.1.3 Missing Data Statistics of categorical Variables: 0

3.1.1.3.1 Missing Data Treatment: Categorical Variables or Features: 0

3.1.1.3.1.1 Removal of Variables or Features with More Than 50% Missing Data: None

3.1.1.4 Missing Data Statistics of non-categorical Variables: 0

3.1.1.4.1 Missing Data Treatment of non-categorical Variables: 0

3.1.1.4.1.1 Removal of Variables or Features with More Than 50% Missing Data: None

3.1.2 Numerical Encoding of Categorical Variables

In this case, category to number node will be used to encode the categorical variables.

home_ownership
mortagage - 3, none - 5, other - 4, own - 2, rent – 1

Term
36 months – 1, 60 months – 2

application_type
Individual - 1, Joint - 2

Purpose
Credit card – 1, car – 2, small business – 3, other – 4, wedding – 5, debt consolidation – 6, home improvement – 7, major purchase – 8, medical – 9, moving – 10, vacation – 11, house – 12, renewable energy – 13, educational - 14

loan_condition
Good Loan – 1, Bad Loan - 2

Region
Munster – 1, Leinster – 2, Cannught – 3, Ulster – 4, Northern-Irl - 5

income_category
Low – 1, Medium – 2, High – 3

interest_payments
Low – 1, High – 2

Grade
B – 1, C – 2, A – 3, E – 4, F – 5, D – 6, G – 7

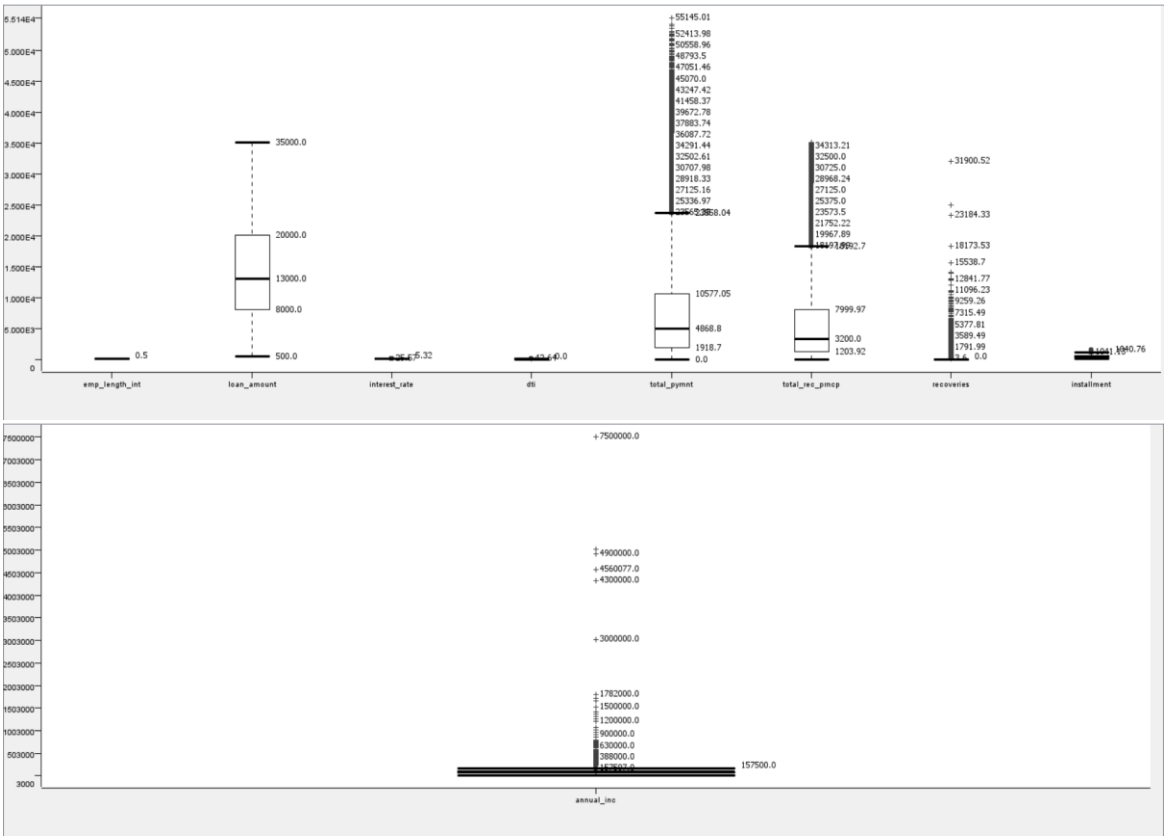
3.1.3 Outlier Statistics and Treatment

3.1.3.1 Outlier Statistics: Non-Categorical Variables

Row ID	S Outlier c...	I Membe...	I Outlier ...	D Lower ...	D Upper ...
Row0	emp_length_int	106485	0	-7.5	20.5
Row1	annual_inc	106485	4839	-22,500	157,500
Row2	loan_amount	106485	0	-10,000	38,000
Row3	interest_rate	106485	746	0.675	25.515
Row4	dti	106485	10	-6.22	42.1
Row5	total_pymnt	106485	5684	-11,068.806	23,564.544
Row6	total_rec_prncp	106485	6909	-8,990.151	18,194.039
Row7	recoveries	106485	2973	0	0
Row8	installment	106485	2800	-205.9	1,040.9

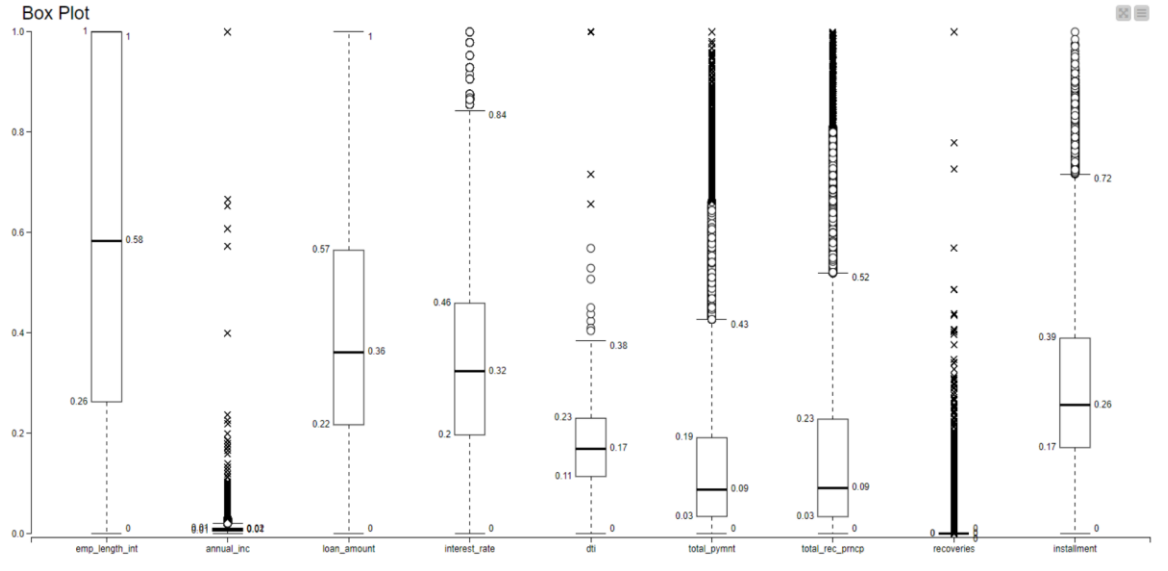
3.1.3.2 Normalization using Min-Max Scaler

Before Normalization



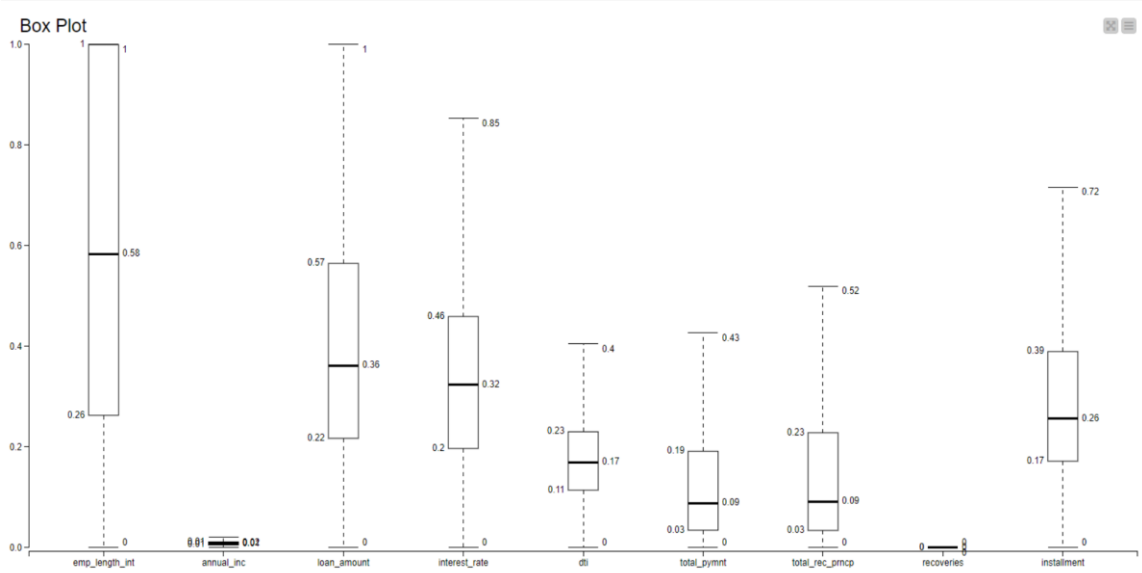
After Normalization

Min-Max Scaler Normalization (between 0 and 1) for variables: annual_inc, interest_rate, dti, total_pymnt, total_rec_prncp, recoveries, installment



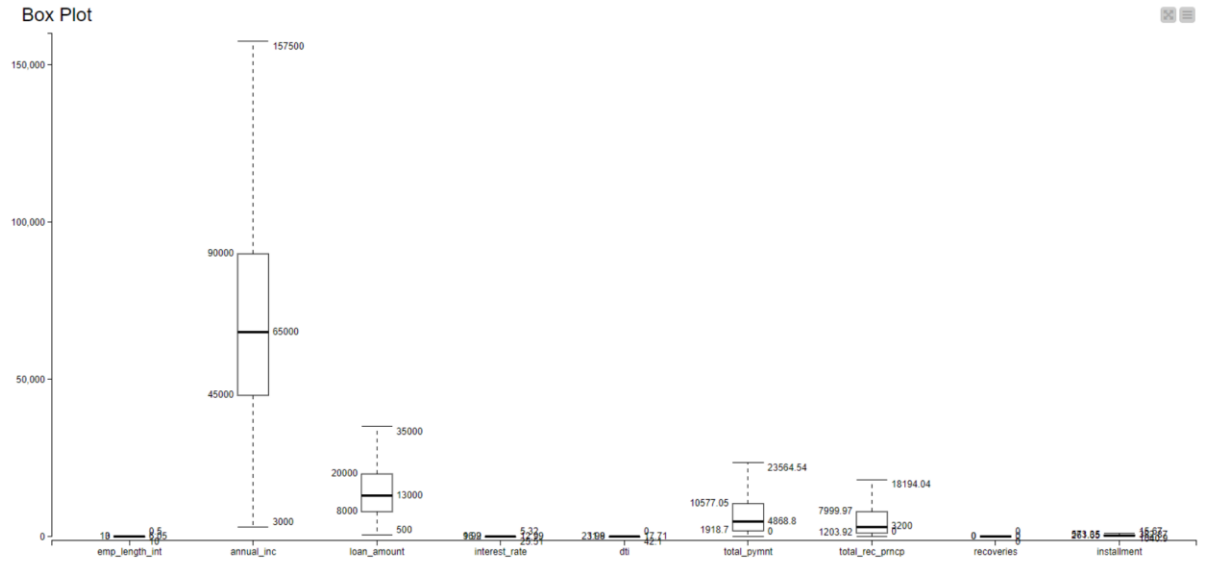
After treating Outliers

Outlier Treatment using replacement strategy where the values are replaced to the closest permitted value



As the number of outliers in annual income and recoveries are more thus it will be part of the data so that segment will not be ignored when unsupervised learning will be used.

After de-normalizing the data, we get the following box plot: -



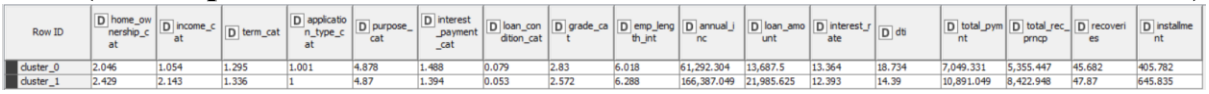
3.2 Data Analysis

3.2.1 Unsupervised Machine Learning Algorithm

K-means clustering is a popular unsupervised machine learning algorithm used for partitioning a dataset into a predefined number of non-overlapping clusters. The algorithm aims to group data points into clusters in such a way that the similarity (or distance) between data points within the same cluster is maximized, while the similarity between data points in different clusters is minimized.

In this project, K-means will be the clustering algorithm used for unsupervised learning. The metrics used in k-means is Euclidean distance.

K=2 (This represents the total number of clusters that will be formed are 2)



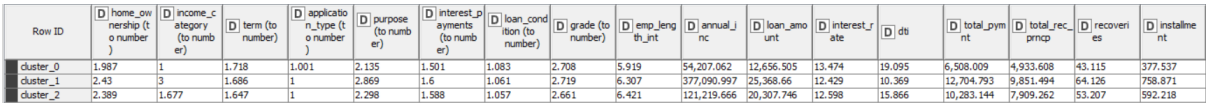
Row ID	home_ownership_cat	income_cat	term_cat	application_type_cat	purpose_cat	interest_payment_cat	loan_condition_cat	grade_cat	emp_length_int	annual_inc	loan_amount	interest_rate	dti	total_pymnt	total_rec_prncp	recoveries	installment
cluster_0	2.046	1.054	1.295	1.001	4.878	1.488	0.079	2.83	6.018	61,292.304	13,687.5	13.364	18.734	7,049.331	5,355.447	45.682	405.782
cluster_1	2.429	2.143	1.336	1	4.87	1.394	0.053	2.572	6.288	166,387.049	21,985.625	12.393	14.39	10,891.049	8,422.948	47.87	645.835

	Cluster 1	Cluster 2
home_ownership	rent	own
income_category	Low	Medium
Term	36 months	36 months
application_type	Individual	Individual
Purpose	Other	Other
interest_payments	High	Low
loan_condition	Bad Loan	Good Loan
Grade	C	B
emp_length_int	6.018	6.288
annual_inc	$ 61,292.30	$ 1,66,387.05
loan_amount	$ 13,687.50	$ 21,985.63
interest_rate	13.364%	12.393%
dti	18.734 %	14.390 %
total_pymnt	$ 7,049.33	$ 10,891.05
total_rec_prncp	$ 5,355.45	$ 8,422.95
recoveries	$ 45.68	$ 47.87
installment	$ 405.78	$ 645.84

Analysis of Cluster 1 and 2 (K=2)

Cluster 2 represents individuals with higher income and better creditworthiness, while cluster 1 represents individuals with lower income and potentially higher risk loans. Cluster 2 has a higher average loan amount, higher annual income, lower interest rate and a lower debt-to-income ratio compared to Cluster 1. Cluster 1 has a higher average interest rate and debt-to-income ratio indicating potentially riskier loans.

K=3 (This represents the total number of clusters that will be formed are 3)



Row ID	home_ownership (to number)	income_category (to number)	term (to number)	application_type (to number)	purpose (to number)	interest_payments (to number)	loan_condition (to number)	grade (to number)	emp_length_int	annual_inc	loan_amount	interest_rate	dti	total_pymnt	total_rec_prncp	recoveries	installment
cluster_0	1.987	1	1.718	1.001	2.135	1.501	1.083	2.708	5.919	54,207.062	12,656.505	13.474	19.095	6,508.009	4,933.608	43.115	377.537
cluster_1	2.43	3	1.686	1	2.869	1.6	1.061	2.719	6.307	377,090.997	25,368.66	12.429	10.369	12,704.793	9,851.494	64.126	758.871
cluster_2	2.389	1.677	1.647	1	2.298	1.588	1.057	2.661	6.421	121,219.666	20,307.746	12.598	15.866	10,283.144	7,909.262	53.207	592.218

	Cluster 1	Cluster 2	Cluster 3
home_ownership	rent	own	own
income_category	Low	High	Medium
Term	36 months	36 months	36 months
application_type	Individual	Individual	Individual
Purpose	Car (Debt consolidation)	Car or small business (Major purchase)	Car or small business (Major purchase)
interest_payments	High	Low	Low
loan_condition	Bad Loan	Good Loan	Good Loan
Grade	C	B	B
emp_length_int	5.919	6.307	6.421
annual_inc	$ 54,207.06	$ 3,77,091.00	$ 1,21,219.67
loan_amount	$ 12,656.50	$ 25,368.66	$ 20,307.75
interest_rate	13.47%	12.43%	12.60%
dti	19.10%	10.37%	15.87%
total_pymnt	$ 6,508.01	$ 12,704.79	$ 10,283.14
total_rec_prncp	$ 4,933.61	$ 9,851.49	$ 7,909.26
recoveries	$ 43.11	$ 64.13	$ 53.21
installment	$ 377.54	$ 758.87	$ 592.22

Analysis of Cluster 1,2 and 3 (K=3)

Cluster 1: Low-Income, Higher-Risk Borrowers

- This cluster comprises individuals with lower incomes, mostly renting their homes and seeking loans for debt consolidation.
- They have higher-interest payments and loans in this cluster are classified as bad loan indicating higher risk.
- The loans in this cluster have higher interest rates and higher debt-to-income ratios.

- Borrowers in this cluster tend to have shorter employment lengths and lower loan amounts.
- This cluster represents higher-risk borrowers who may face financial challenges.

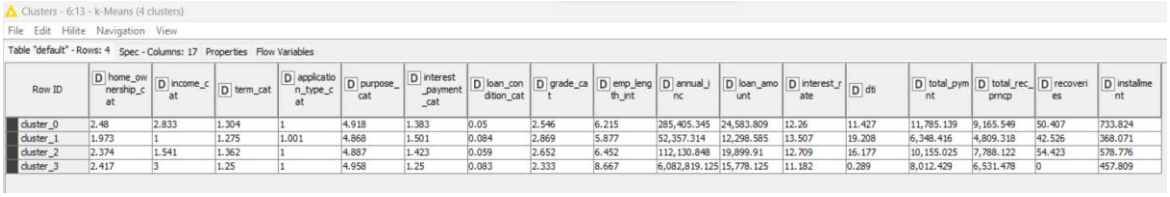
Cluster 2: High-Income, Low-Risk Borrowers

- This cluster consists of individuals with high incomes, owning their homes and seeking loans for major purchases.
- They have low-interest payments, good loan conditions and relatively lower debt-to-income ratios.
- The loans in this cluster are graded as 'B' indicating good creditworthiness.
- Borrowers in this cluster tend to have longer employment lengths and higher loan amounts.
- Overall, this cluster represents financially stable borrowers with low risk.

Cluster 3: Medium-Income, Moderate-Risk Borrowers

- This cluster consists of individuals with moderate incomes, owning their homes and seeking loans for major purchases.
- They have low-interest payments and good loan conditions similar to cluster 1.
- The loans in this cluster are also graded as 'B' indicating moderate creditworthiness.
- Borrowers in this cluster have moderate employment lengths and loan amounts.
- Overall, this cluster represents borrowers with moderate financial stability and moderate risk

K=4 (This represents the total number of clusters that will be formed are 4)


Clusters - 6:13 - k-Means (4 clusters)

File Edit Hilite Navigation View

Table "default" - Rows: 4 Spec - Columns: 17 Properties Flow Variables

Row ID	home_ownership_cat	income_cat	term_cat	application_type_cat	purpose_cat	interest_payment_cat	loan_condition_cat	grade_cat	emp_length_int	annual_inc	loan_amount	interest_rate	dti	total_pymnt	total_rec_prncp	recoveries	installment
cluster_0	2.48	2.833	1.304	1	4.918	1.383	0.05	2.546	6.215	285,405.345	24,583.809	12.26	11.427	11,785.139	9,165.549	50.407	733.824
cluster_1	1.973	1	1.275	1.001	4.868	1.501	0.084	2.869	5.877	52,357.314	12,298.585	13.507	19.208	6,348.416	4,809.318	42.526	368.071
cluster_2	2.374	1.541	1.362	1	4.887	1.423	0.059	2.652	6.452	112,130.848	19,899.91	12.709	16.177	10,155.025	7,788.122	54.423	578.776
cluster_3	2.417	3	1.25	1	4.958	1.25	0.083	2.333	8.667	6,082,819.125	15,778.125	11.182	0.289	8,012.429	6,531.478	0	457.809

Cluster 1

- home_ownership: own
- income_category: Medium-High
- Term: 36 months
- application_type: Individual
- Purpose: Major purchase
- interest_payments: Low
- loan_condition: Good Loan
- Grade: B

- emp_length_int: 6.21 years
- annual_inc: $285,405.34
- loan_amount: $24,583.81
- interest_rate: 12.26%
- dti: 11.43%
- total_pymnt: $11,785.14
- total_rec_prncp: $9,165.55
- recoveries: $50.41
- installment: $733.82

Cluster 2

- home_ownership: rent
- income_category: Low
- Term: 36 months
- application_type: Individual
- Purpose: Debt consolidation
- interest_payments: High
- loan_condition: Bad Loan
- Grade: C
- emp_length_int: 5.88 years
- annual_inc: $52,357.31
- loan_amount: $12,298.58
- interest_rate: 13.51%
- dti: 19.21%
- total_pymnt: $6,348.42
- total_rec_prncp: $4,809.32
- recoveries: $42.53
- installment: $368.07

Cluster 3

- home_ownership: own
- income_category: Medium
- Term: 36 months
- application_type: Individual
- Purpose: Major purchase
- interest_payments: Low
- loan_condition: Good Loan
- Grade: B
- emp_length_int: 6.45 years
- annual_inc: $112,130.85
- loan_amount: $19,899.91
- interest_rate: 12.71%
- dti: 16.18%
- total_pymnt: $10,155.02
- total_rec_prncp: $7,788.12
- recoveries: $54.42

- installment: $578.78

Cluster 4

- home_ownership: own
- income_category: High
- Term: 36 months
- application_type: Individual
- Purpose: Major purchase
- interest_payments: High
- loan_condition: Good Loan
- Grade: A
- emp_length_int: 8.67 years
- annual_inc: $6,082,819.13
- loan_amount: $15,778.13
- interest_rate: 11.18%
- dti: 0.29%
- total_pymnt: $8,012.43
- total_rec_prncp: $6,531.48
- recoveries: $0.00
- installment: $457.81

Analysis of Cluster 1,2,3 and 4 (K=4)

Cluster 1: High-Income, Low-Risk Borrowers

- This cluster consists of individuals with high to medium-high incomes, owning their homes and seeking loans for major purchases.
- They demonstrate characteristics of low-risk borrowers with good creditworthiness and low-interest payments.

Cluster 2: Low-Income, Higher-Risk Renters

- This cluster comprises individuals with lower incomes, mostly renting their homes and seeking loans for debt consolidation.
- They exhibit higher risk due to their lower incomes, higher debt-to-income ratios and higher likelihood of loans classified as bad loan.

Cluster 3: Medium-Income, Moderate-Risk Borrowers

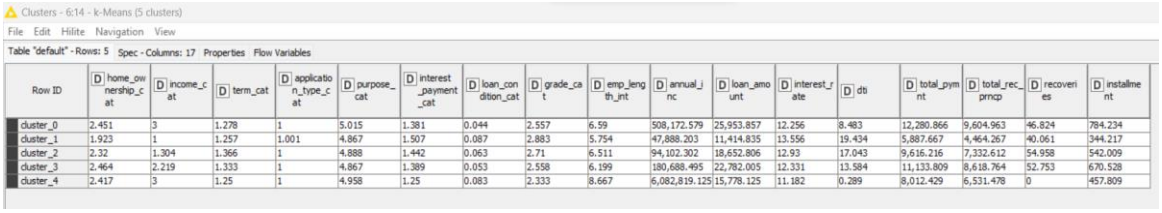
- This cluster consists of individuals with moderate incomes, owning their homes and seeking loans for major purchases.
- They represent borrowers with moderate financial stability and risk, similar to cluster 1 but with slightly lower incomes.

Cluster 4: High-Income Outlier

- This cluster represents an outlier group with extremely high incomes, owning their homes and seeking loans for major purchases.

- Despite the high incomes, they have relatively low loan amounts and debt-to-income ratios suggesting they may be managing their finances differently.

K=5 (This represents the total number of clusters that will be formed are 5)



Clusters - 6:14 - k-Means (5 clusters)

File Edit Hilite Navigation View

Table "default" - Rows: 5 Spec - Columns: 17 Properties Flow Variables

Row ID	home_ownership_cat	income_cat	term_cat	application_type_cat	purpose_cat	interest_payment_cat	loan_condition_cat	grade_cat	emp_length_int	annual_inc	loan_amount	interest_rate	dti	total_pymnt	total_rec_prncp	recoveries	installment
cluster_0	2.451	3	1.278	1	5.015	1.381	0.044	2.557	6.59	508,172.579	25,953.857	12.256	8.483	12,280.866	9,604.963	46.824	784.234
cluster_1	1.923	1	1.257	1.001	4.867	1.507	0.087	2.883	5.754	47,888.203	11,414.835	13.556	19.434	5,887.667	4,464.267	40.061	344.217
cluster_2	2.32	1.304	1.366	1	4.888	1.442	0.063	2.71	6.511	94,102.302	18,652.806	12.93	17.043	9,616.216	7,332.612	54.958	542.009
cluster_3	2.464	2.219	1.333	1	4.867	1.389	0.053	2.558	6.199	180,688.495	22,782.005	12.331	13.584	11,133.809	8,618.764	52.753	670.528
cluster_4	2.417	3	1.25	1	4.958	1.25	0.083	2.333	8.667	6,082,819.125	15,778.125	11.182	0.289	8,012.429	6,531.478	0	457.809

Cluster 1

- home_ownership: own
- income_category: High
- Term: 36 months
- application_type: Individual
- Purpose: Other
- interest_payments: Low
- loan_condition: Good Loan
- Grade: B
- emp_length_int: 6.59 years
- annual_inc: $508,172.58
- loan_amount: $25,953.86
- interest_rate: 12.26%
- dti: 8.48%
- total_pymnt: $12,280.87
- total_rec_prncp: $9,604.96
- recoveries: $46.82
- installment: $784.23

Cluster 2

- home_ownership: rent
- income_category: Low

- Term: 36 months
- application_type: Individual
- Purpose: Debt consolidation
- interest_payments: High
- loan_condition: Bad Loan
- Grade: C
- emp_length_int: 5.75 years
- annual_inc: $47,888.20
- loan_amount: $11,414.83
- interest_rate: 13.56%
- dti: 19.43%
- total_pymnt: $5,887.67
- total_rec_prncp: $4,464.27
- recoveries: $40.06
- installment: $344.22

Cluster 3

- home_ownership: own
- income_category: Medium
- Term: 36 months
- application_type: Individual
- Purpose: Other
- interest_payments: Low
- loan_condition: Good Loan
- Grade: B
- emp_length_int: 6.51 years
- annual_inc: $94,102.30
- loan_amount: $18,652.81
- interest_rate: 12.93%
- dti: 17.04%
- total_pymnt: $9,616.22
- total_rec_prncp: $7,332.61
- recoveries: $54.96

- installment: $542.01

Cluster 4

- home_ownership: own
- income_category: High
- Term: 36 months
- application_type: Individual
- Purpose: Other
- interest_payments: Low
- loan_condition: Good Loan
- Grade: B
- emp_length_int: 6.20 years
- annual_inc: $180,688.49
- loan_amount: $22,782.00
- interest_rate: 12.33%
- dti: 13.58%
- total_pymnt: $11,133.81
- total_rec_prncp: $8,618.76
- recoveries: $52.75
- installment: $670.53

Cluster 5

- home_ownership: own
- income_category: High
- Term: 36 months
- application_type: Individual
- Purpose: Other
- interest_payments: High
- loan_condition: Good Loan
- Grade: A
- emp_length_int: 8.67 years
- annual_inc: $6,082,819.13
- loan_amount: $15,778.13

- interest_rate: 11.18%
- dti: 0.29%
- total_pymnt: $8,012.43
- total_rec_prncp: $6,531.48
- recoveries: $0.00
- installment: $457.81

Analysis of cluster 1,2,3,4 and 5 (K=5)

Cluster 1: High-Income, Low-Risk Borrowers

- This cluster consists of individuals with high incomes, owning their homes and seeking loans for various purposes.
- They demonstrate characteristics of low-risk borrowers with good creditworthiness and low-interest payments.

Cluster 2: Low-Income Renters in Financial Distress

- This cluster comprises individuals with lower incomes, mostly renting their homes and seeking loans for debt consolidation.
- They exhibit higher risk due to their lower incomes, higher debt-to-income ratios and higher likelihood of loans classified as bad loan.

Cluster 3: Medium-Income Homeowners

- This cluster consists of individuals with moderate incomes, owning their homes and seeking loans for various purposes.
- They represent borrowers with moderate financial stability and risk.

Cluster 4: High-Income Homeowners

- This cluster represents individuals with high incomes, owning their homes and seeking loans for various purposes.
- Despite the high incomes, they have relatively moderate loan amounts and debt-to-income ratios.

Cluster 5: High-Income Outlier

- This cluster represents an outlier group with extremely high incomes, owning their homes and seeking loans for various purposes.
- They have exceptionally high incomes and very low debt-to-income ratios suggesting they may be managing their finances differently.

3.2.2 Clustering Model Performance Evaluation

The silhouette score is a metric used to evaluate the quality of clustering in unsupervised learning. It measures how similar an object is to its own cluster (cohesion) compared to other clusters (separation). A silhouette score ranges from -1 to 1, where a higher score indicates better clustering:

- Silhouette Score of 1 indicates that clusters are well-separated.
- Silhouette Score of 0 indicates overlapping clusters.
- Silhouette Score close to -1 indicates that samples have been assigned to the wrong clusters.

K=2

Row ID	D Mean Silhouette Coefficient
cluster_0	0.681
cluster_1	0.312
Overall	0.632

- Silhouette score of 0.681 indicates that cluster 0 has strong cohesion and is well-separated from other clusters.
- Silhouette score of 0.312 suggests that cluster 1 has moderate cohesion and separation from other clusters.

K=3

Row ID	D Mean Silhouette Coefficient
cluster_0	0.609
cluster_2	0.355
cluster_1	0.266
Overall	0.539

- Silhouette score of 0.355 indicates that cluster 0 has moderate to strong cohesion and separation.
- Silhouette score of 0.609 indicates that cluster 1 has strong cohesion and separation making it a well-defined cluster.
- Silhouette score of 0.266 indicates that cluster 2 also shows moderate cohesion and separation

K=4

Row ID	D Mean Silhouette Coefficient
cluster_0	0.59
cluster_2	0.367
cluster_1	0.32
cluster_3	0.615
Overall	0.519

- Silhouette score of 0.367 indicates that cluster 0 exhibits moderate to strong cohesion and separation.
- Silhouette score of 0.59 and 0.615 indicates that cluster 0 and cluster 1 shows strong cohesion and separation, indicating a well-defined cluster.
- Silhouette score of 0.32 suggests moderate cohesion and separation for cluster 1.

K=5

Row ID	D Mean Silhouette Coefficient
cluster_4	0.556
cluster_0	0.361
cluster_2	0.367
cluster_1	0.391
cluster_3	0.593
Overall	0.475

- Silhouette score of 0.556 and 0.593 suggests strong cohesion and separation for Cluster 4 and cluster 3.
- Silhouette score of 0.391 indicates that cluster 1 exhibits moderate cohesion and separation.
- Silhouette score of 0.361 and 0.367 indicates moderate cohesion and separation for cluster 0 and cluster 2.

To be able to identify, which cluster gives the better results or performance, silhouette score will be compared for each cluster. The closer the score to 1 better the performance of the cluster.

3.2.3 Cluster Analysis using Base Model as K-Means

3.2.3.1 Cluster Analysis with Categorical Variables

The Kruskal-Wallis test is a non-parametric statistical test used to determine whether there are statistically significant differences between the medians of two or more independent groups.

The test is appropriate when the data do not meet the assumptions required for parametric tests like ANOVA.
In KNIME, Kruskal-Wallis Test is used to analyze the categorical variable. The variables that have $p < 0.05$, those variables will be significant in the analysis of clusters.

K=2

home_ownership

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1
Row0	2,022.824	0.0	51,752.936	48,126.5	62,999.707	79,917

Term

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1
Row0	98.241	0.0	52,953.212	37,237	55,140.489	37,237

application_type

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1
Row0	4.85	0.027652175980004134	53,246.276	53,214	53,221.548	53,214

Purpose

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1
Row0	14.766	1.2173845392360327E-4	53,118.132	63,920.5	54,060.613	63,920.5

loan_condition

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1
Row0	126.092	0.0	53,432.642	49,201	52,001.251	49,201

income_category

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1
Row0	75,414.162	0.0	46,514.633	43,700.5	97,299.306	95,933

interest_payments

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1
Row0	433.455	0.0	53,905.945	27,938	48,902.14	27,938

Grade

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1
Row0	606.601	0.0	54,124.781	63,164	47,469.235	33,074.5

The p-value associated with the Kruskal-Wallis test for all variables is less than the significance level of 0.05. Therefore, we reject the null hypothesis and conclude that there are statistically significant differences between the medians of cluster 0 and cluster 1.

The mean and median ranks of each cluster indicate the average and middle positions of the observations within each group. The differences in these values between the two clusters suggest variations in the distribution of data points, contributing to the rejection of the null hypothesis.

We see that all the categorical variables have p-value less than 0.05 indicating that there are significant differences in the distributions of the data between cluster 0 and cluster 1 as indicated by the Kruskal-Wallis test results.

K=3

home_ownership

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1	Mean Rank of Group cluster_2	Median Rank of Group cluster_2
Row0	3,884.119	0.0	50,021.688	48,126.5	62,999.381	79,917	61,857.723	79,917

Term

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1	Mean Rank of Group cluster_2	Median Rank of Group cluster_2
Row0	501.579	0.0	52,229.931	37,237	53,934.352	37,237	56,053.812	37,237

application_type

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1	Mean Rank of Group cluster_2	Median Rank of Group cluster_2
Row0	8.351	0.015364712550923976	53,249.67	53,214	53,214	53,214	53,225.539	53,214

Purpose

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1	Mean Rank of Group cluster_2	Median Rank of Group cluster_2
Row0	23.475	7.98856315498142E-6	53,009.759	63,920.5	55,046.746	63,920.5	53,819.864	63,920.5

loan_condition

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1	Mean Rank of Group cluster_2	Median Rank of Group cluster_2
Row0	203.329	0.0	53,618.554	49,201	52,441.456	49,201	52,224.306	49,201

income_category

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1	Mean Rank of Group cluster_2	Median Rank of Group cluster_2
Row0	64,371.577	0.0	43,700.5	43,700.5	105,475.5	105,475.5	77,765.422	95,933

interest_payments

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1	Mean Rank of Group cluster_2	Median Rank of Group cluster_2
Row0	642.969	0.0	54,502.316	27,938	49,225.999	27,938	49,883.895	27,938

Grade

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1	Mean Rank of Group cluster_2	Median Rank of Group cluster_2
Row0	894.109	0.0	54,912.683	63,164	47,908.88	33,074.5	48,789.626	33,074.5

The p-value associated with the Kruskal-Wallis test which is less than the significance level of 0.05. Therefore, we reject the null hypothesis and conclude that there are statistically significant differences between the medians of cluster 0 and cluster 1.

The mean and median ranks of each cluster indicate the average and middle positions of the observations within each group. The differences in these values between the two clusters suggest variations in the distribution of data points, contributing to the rejection of the null hypothesis.

We see that all the categorical variables have p-value less than 0.05 indicating that there are significant differences in the distributions of the data between cluster 0, cluster 1 and cluster 2 as indicated by the Kruskal-Wallis test results.

K=4

home_ownership

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1	Mean Rank of Group cluster_2	Median Rank of Group cluster_2	Mean Rank of Group cluster_3	Median Rank of Group cluster_3
Row0	4,312.318	0.0	49,501.109	48,126.5	63,643.503	79,917	61,318.623	79,917	74,618.583	79,917

Term

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1	Mean Rank of Group cluster_2	Median Rank of Group cluster_2	Mean Rank of Group cluster_3	Median Rank of Group cluster_3
Row0	710.68	0.0	51,927.819	37,237	53,614.289	37,237	56,345.478	37,237	46,110.75	37,237

application_type

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1	Mean Rank of Group cluster_2	Median Rank of Group cluster_2	Mean Rank of Group cluster_3	Median Rank of Group cluster_3
Row0	8.333	0.03961085291652433	53,250.34	53,214	53,235.749	53,214	53,226.11	53,214	53,214	53,214

Purpose

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1	Mean Rank of Group cluster_2	Median Rank of Group cluster_2	Mean Rank of Group cluster_3	Median Rank of Group cluster_3
Row0	37.388	3.808571225061286E-8	52,962.112	63,920.5	55,581.922	63,920.5	53,722.729	63,920.5	67,811.5	63,920.5

loan_condition

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1	Mean Rank of Group cluster_2	Median Rank of Group cluster_2	Mean Rank of Group cluster_3	Median Rank of Group cluster_3
Row0	222.027	0.0	53,675.911	49,201	52,485.157	49,201	52,273.581	49,201	49,201	49,201

income_category

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1	Mean Rank of Group cluster_2	Median Rank of Group cluster_2	Mean Rank of Group cluster_3	Median Rank of Group cluster_3
Row0	56,584.351	0.0	43,700.5	43,700.5	103,783.733	105,475.5	71,927.265	95,933	105,475.5	105,475.5

interest_payments

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1	Mean Rank of Group cluster_2	Median Rank of Group cluster_2	Mean Rank of Group cluster_3	Median Rank of Group cluster_3
Row0	661.018	0.0	54,650.827	81,180.5	49,535.142	27,938	50,189.991	27,938	36,811.75	27,938

Grade

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1	Mean Rank of Group cluster_2	Median Rank of Group cluster_2	Mean Rank of Group cluster_3	Median Rank of Group cluster_3
Row0	896.733	0.0	55,085.623	63,164	47,955.272	33,074.5	49,279.869	33,074.5	40,991	33,074.5

The p-value associated with the Kruskal-Wallis test is less than the significance level of 0.05. Therefore, we reject the null hypothesis and conclude that there are statistically significant differences between the medians of cluster 0 and cluster 1.

The mean and median ranks of each cluster indicate the average and middle positions of the observations within each group. The differences in these values between the two clusters suggest variations in the distribution of data points, contributing to the rejection of the null hypothesis.
We see that all the categorical variables have p-value less than 0.05 indicating that there are significant differences in the distributions of the data between cluster 0, cluster 1, cluster 2 and cluster 3 as indicated by the Kruskal-Wallis test results.

K=5

home_ownership

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1	Mean Rank of Group cluster_2	Median Rank of Group cluster_2	Mean Rank of Group cluster_3	Median Rank of Group cluster_3	Mean Rank of Group cluster_4	Median Rank of Group cluster_4
Row0	5,299.825	0.0	59,737.421	79,917	63,586.18	79,917	63,608.296	79,917	74,618.583	79,917	48,057.36	48,126.5

Term

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1	Mean Rank of Group cluster_2	Median Rank of Group cluster_2	Mean Rank of Group cluster_3	Median Rank of Group cluster_3	Mean Rank of Group cluster_4	Median Rank of Group cluster_4
Row0	1,264.664	0.0	56,653.496	37,237	52,976.122	37,237	54,829.669	37,237	46,110.75	37,237	51,019.389	37,237

application_type

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1	Mean Rank of Group cluster_2	Median Rank of Group cluster_2	Mean Rank of Group cluster_3	Median Rank of Group cluster_3	Mean Rank of Group cluster_4	Median Rank of Group cluster_4
Row0	8.182	0.08511926781124224	53,232.747	53,214	53,214	53,214	53,221.258	53,214	53,214	53,214	53,251.913	53,214

Purpose

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1	Mean Rank of Group cluster_2	Median Rank of Group cluster_2	Mean Rank of Group cluster_3	Median Rank of Group cluster_3	Mean Rank of Group cluster_4	Median Rank of Group cluster_4
Row0	36.198	2.6352057502787574E-7	53,619.617	63,920.5	55,816.96	63,920.5	54,329.187	63,920.5	67,811.5	63,920.5	52,869.559	63,920.5

loan_condition

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1	Mean Rank of Group cluster_2	Median Rank of Group cluster_2	Mean Rank of Group cluster_3	Median Rank of Group cluster_3	Mean Rank of Group cluster_4	Median Rank of Group cluster_4
Row0	254.664	0.0	52,504.862	49,201	52,398.009	49,201	52,118.596	49,201	49,201	49,201	53,823.852	49,201

income_category

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1	Mean Rank of Group cluster_2	Median Rank of Group cluster_2	Mean Rank of Group cluster_3	Median Rank of Group cluster_3	Mean Rank of Group cluster_4	Median Rank of Group cluster_4
Row0	55,076.749	0.0	59,701.301	43,700.5	105,475.5	105,475.5	97,989.528	95,933	105,475.5	105,475.5	43,700.5	43,700.5

interest_payments

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1	Mean Rank of Group cluster_2	Median Rank of Group cluster_2	Mean Rank of Group cluster_3	Median Rank of Group cluster_3	Mean Rank of Group cluster_4	Median Rank of Group cluster_4
Row0	707.956	0.0	51,128.483	27,938	49,825.217	27,938	49,021.624	27,938	36,811.75	27,938	55,033.168	81,180.5

Grade

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1	Mean Rank of Group cluster_2	Median Rank of Group cluster_2	Mean Rank of Group cluster_3	Median Rank of Group cluster_3	Mean Rank of Group cluster_4	Median Rank of Group cluster_4
Row0	913.841	0.0	50,650.773	63,164	47,394.184	33,074.5	47,570.426	33,074.5	40,991	33,074.5	55,507.526	63,164

The p-value associated with the Kruskal-Wallis test which is less than the significance level of 0.05 except for application_type variable. Therefore, we reject the null hypothesis and conclude that there are statistically significant differences between the medians of cluster 0 and cluster 1.

The mean and median ranks of each cluster indicate the average and middle positions of the observations within each group. The differences in these values between the two clusters suggest variations in the distribution of data points, contributing to the rejection of the null hypothesis.
We see that all the categorical variables have p-value less than 0.05 except application_type indicating that there are significant differences in the distributions of the data between cluster 0, cluster 1, cluster 2, cluster 3 and cluster 4 as indicated by the Kruskal-Wallis test results.

3.2.3.2 Cluster analysis with Non-Categorical Variables

In KNIME, ANOVA is used to analyze the non-categorical variables. The variables that have $p < 0.05$, those variables are significant in the analysis of clusters.

K=2

	Source	Sum of Squares	df	Mean Square	F	p-value
emp_length_int	Between Groups	893.74	1	893.74	72.6893	0.0
emp_length_int	Within Groups	1,309,245.995	106483	12.2954		
emp_length_int	Total	1,310,139.7351	106484			
annual_inc	Between Groups	1.35E14	1	1.35E14	51,077.8088	0.0
annual_inc	Within Groups	2.82E14	106483	2.65E9		
annual_inc	Total	4.17E14	106484			
loan_amount	Between Groups	8.43E11	1	8.43E11	13,353.2951	0.0
loan_amount	Within Groups	6.72E12	106483	63,112,070.506		
loan_amount	Total	7.56E12	106484			
interest_rate	Between Groups	11,537.053	1	11,537.053	604.4209	0.0
interest_rate	Within Groups	2,032,524.0514	106483	19.0878		
interest_rate	Total	2,044,061.1044	106484			
dti	Between Groups	231,007.8241	1	231,007.8241	3,442.6246	0.0
dti	Within Groups	7,145,247.8234	106483	67.1022		
dti	Total	7,376,255.6475	106484			
total_pymnt	Between Groups	1.81E11	1	1.81E11	2,979.4038	0.0
total_pymnt	Within Groups	6.46E12	106483	60,626,512.031		
total_pymnt	Total	6.64E12	106484			
total_rec_prncp	Between Groups	1.15E11	1	1.15E11	2,670.6974	0.0
total_rec_prncp	Within Groups	4.59E12	106483	43,120,672.9072		
total_rec_prncp	Total	4.71E12	106484			
recoveries	Between Groups	58,567.8212	1	58,567.8212	0.3386	0.5606
recoveries	Within Groups	1.84E10	106483	172,972.2147		
recoveries	Total	1.84E10	106484			
installment	Between Groups	7.05E8	1	7.05E8	13,303.9303	0.0
installment	Within Groups	5.64E9	106483	53,012.0922		
installment	Total	6.35E9	106484			

Descriptive Statistics: K=2

Row ID	Test Column	Group	N	Missing Count	Missing Count (Group Column)	Mean	Standard Deviation	Standard Error Mean	Confidence Interval Probability	Confidence Interval of the Difference...	Confidence Interval of the Difference...	Minimum	Maximum
Row0	emp_length_int	cluster_0	92377	0	0	6.018	3.494	0.011	0.95	5.995	6.04	0.5	10
Row1	emp_length_int	cluster_1	14108	0	0	6.288	3.587	0.03	0.95	6.229	6.347	0.5	10
Row2	emp_length_int	Total	106485	0	0	6.054	3.508	0.011	0.95	6.033	6.075	0.5	10
Row3	annual_inc	cluster_0	92377	0	0	61,292.304	22,996.362	75.662	0.95	61,144.008	61,440.601	3,000	115,200
Row4	annual_inc	cluster_1	14108	0	0	166,387.049	128,504.598	1,081.897	0.95	164,266.388	168,507.711	110,589	7,500,000
Row5	annual_inc	Total	106485	0	0	75,216.112	62,577.359	191.767	0.95	74,840.252	75,591.972	3,000	7,500,000
Row6	loan_amount	cluster_0	92377	0	0	13,687.5	7,713.113	25.377	0.95	13,637.76	13,737.239	500	35,000
Row7	loan_amount	cluster_1	14108	0	0	21,985.625	9,317.536	78.446	0.95	21,831.862	22,139.389	1,000	35,000
Row8	loan_amount	Total	106485	0	0	14,786.903	8,427.684	25.826	0.95	14,736.283	14,837.522	500	35,000
Row9	interest_rate	cluster_0	92377	0	0	13.364	4.336	0.014	0.95	13.336	13.392	5.32	28.99
Row10	interest_rate	cluster_1	14108	0	0	12.393	4.579	0.039	0.95	12.317	12.468	5.32	28.99
Row11	interest_rate	Total	106485	0	0	13.235	4.381	0.013	0.95	13.209	13.261	5.32	28.99
Row12	dti	cluster_0	92377	0	0	18.734	8.321	0.027	0.95	18.681	18.788	0	104
Row13	dti	cluster_1	14108	0	0	14.39	7.29	0.061	0.95	14.27	14.51	0	39.94
Row14	dti	Total	106485	0	0	18.159	8.323	0.026	0.95	18.109	18.209	0	104
Row15	total_pymnt	cluster_0	92377	0	0	7,049.331	7,263.763	23.899	0.95	7,002.489	7,096.172	0	53,851.24
Row16	total_pymnt	cluster_1	14108	0	0	10,891.049	10,588.842	89.149	0.95	10,716.306	11,065.793	0	55,145.01
Row17	total_pymnt	Total	106485	0	0	7,558.313	7,894.445	24.192	0.95	7,510.896	7,605.729	0	55,145.01
Row18	total_rec_prncp	cluster_0	92377	0	0	5,355.447	6,102.476	20.078	0.95	5,316.094	5,394.8	0	35,000.01
Row19	total_rec_prncp	cluster_1	14108	0	0	8,422.948	9,034.786	76.065	0.95	8,273.85	8,572.046	0	35,000.03
Row20	total_rec_prncp	Total	106485	0	0	5,761.855	6,648.441	20.374	0.95	5,721.922	5,801.787	0	35,000.03
Row21	recoveries	cluster_0	92377	0	0	45.682	403.867	1.329	0.95	43.078	48.287	0	31,900.52
Row22	recoveries	cluster_1	14108	0	0	47.87	487.405	4.104	0.95	39.826	55.913	0	15,538.7
Row23	recoveries	Total	106485	0	0	45.972	415.898	1.275	0.95	43.474	48.47	0	31,900.52
Row24	installment	cluster_0	92377	0	0	405.782	220.075	0.724	0.95	404.363	407.201	15.67	1,445.46
Row25	installment	cluster_1	14108	0	0	645.835	288.095	2.426	0.95	641.081	650.589	25.28	1,409.99
Row26	installment	Total	106485	0	0	437.586	244.202	0.748	0.95	436.12	439.053	15.67	1,445.46

The null hypothesis is rejected in all of the variables except recoveries (p-value is greater than 0.05) since the p-value is less than 0.05, indicating that there are significant differences in employee length, annual income, income category, interest rate, debt-to-income ratio, total payment, total received principal and instalment (emp_length_int, annual_inc, income_cat, interest rate, debt-to-income ratio, total_pymnt, total_rec_prncp, installment) between the groups. We can further see in descriptive statistics that the means of the clusters i.e. for K=2 are statistically different.

K=3

	Source	Sum of Squares	df	Mean Square	F	p-value
emp_length_int	Between Groups	5,232.3483	2	2,616.1741	213.4829	0.0
emp_length_int	Within Groups	1,304,907.3868	106482	12.2547		
emp_length_int	Total	1,310,139.7351	106484			
annual_inc	Between Groups	2.01E14	2	1.00E14	49,382.342	0.0
annual_inc	Within Groups	2.16E14	106482	2.03E9		
annual_inc	Total	4.17E14	106484			
loan_amount	Between Groups	1.33E12	2	6.64E11	11,345.2646	0.0
loan_amount	Within Groups	6.23E12	106482	58,550,504.1613		
loan_amount	Total	7.56E12	106484			
interest_rate	Between Groups	16,433.7463	2	8,216.8731	431.5138	0.0
interest_rate	Within Groups	2,027,627.3581	106482	19.042		
interest_rate	Total	2,044,061.1044	106484			
dti	Between Groups	285,460.9657	2	142,730.4828	2,143.3743	0.0
dti	Within Groups	7,090,794.6818	106482	66.5915		
dti	Total	7,376,255.6475	106484			
total_pymnt	Between Groups	3.23E11	2	1.61E11	2,719.4842	0.0
total_pymnt	Within Groups	6.31E12	106482	59,294,731.1701		
total_pymnt	Total	6.64E12	106484			
total_rec_prncp	Between Groups	2.01E11	2	1.00E11	2,371.2464	0.0
total_rec_prncp	Within Groups	4.51E12	106482	42,317,845.3334		
total_rec_prncp	Total	4.71E12	106484			
recoveries	Between Groups	2,472,468.359	2	1,236,234.1795	7.1479	0.0008
recoveries	Within Groups	1.84E10	106482	172,951.1696		
recoveries	Total	1.84E10	106484			
installment	Between Groups	1.06E9	2	5.32E8	10,715.719	0.0
installment	Within Groups	5.29E9	106482	49,644.1647		
installment	Total	6.35E9	106484			

Descriptive Statistics: K=3

Row ID	Test Col...	Group	N	Missing ...	Missing ...	Mean	Standa...	Standa...	Confide...	Confide...	Confide...	Minimum	Maximum
Row0	emp_length_int	cluster_0	77618	0	0	5.919	3.482	0.012	0.95	5.894	5.943	0.5	10
Row1	emp_length_int	cluster_2	27684	0	0	6.421	3.552	0.021	0.95	6.38	6.463	0.5	10
Row2	emp_length_int	cluster_1	1183	0	0	6.307	3.505	0.102	0.95	6.107	6.506	0.5	10
Row3	emp_length_int	Total	106485	0	0	6.054	3.508	0.011	0.95	6.033	6.075	0.5	10
Row4	annual_inc	cluster_0	77618	0	0	54,207.062	17,468.179	62.7	0.95	54,084.171	54,329.954	3,000	90,000
Row5	annual_inc	cluster_2	27684	0	0	121,219.666	32,359.05	194.483	0.95	120,838.47	121,600.862	82,325	249,000
Row6	annual_inc	cluster_1	1183	0	0	377,090.997	372,102.591	10,818.582	0.95	355,865.233	398,316.762	249,000	7,500,000
Row7	annual_inc	Total	106485	0	0	75,216.112	62,577.359	191.767	0.95	74,840.252	75,591.972	3,000	7,500,000
Row8	loan_amount	cluster_0	77618	0	0	12,656.505	6,965.876	25.003	0.95	12,607.499	12,705.511	500	35,000
Row9	loan_amount	cluster_2	27684	0	0	20,307.746	9,258.578	55.645	0.95	20,198.678	20,416.814	1,000	35,000
Row10	loan_amount	cluster_1	1183	0	0	25,368.66	8,979.68	261.077	0.95	24,856.434	25,880.886	1,375	35,000
Row11	loan_amount	Total	106485	0	0	14,786.903	8,427.684	25.826	0.95	14,736.283	14,837.522	500	35,000
Row12	interest_rate	cluster_0	77618	0	0	13.474	4.286	0.015	0.95	13.444	13.504	5.32	28.99
Row13	interest_rate	cluster_2	27684	0	0	12.598	4.557	0.027	0.95	12.545	12.652	5.32	28.99
Row14	interest_rate	cluster_1	1183	0	0	12.429	4.752	0.138	0.95	12.158	12.7	5.32	26.06
Row15	interest_rate	Total	106485	0	0	13.235	4.381	0.013	0.95	13.209	13.261	5.32	28.99
Row16	dti	cluster_0	77618	0	0	19.095	8.389	0.03	0.95	19.036	19.154	0	104
Row17	dti	cluster_2	27684	0	0	15.866	7.567	0.045	0.95	15.776	15.955	0	42.64
Row18	dti	cluster_1	1183	0	0	10.369	6.081	0.177	0.95	10.022	10.716	0	36.81
Row19	dti	Total	106485	0	0	18.159	8.323	0.026	0.95	18.109	18.209	0	104
Row20	total_pymnt	cluster_0	77618	0	0	6,508.009	6,554.495	23.527	0.95	6,461.897	6,554.121	0	52,862.58
Row21	total_pymnt	cluster_2	27684	0	0	10,283.144	10,093.135	60.661	0.95	10,164.245	10,402.043	0	55,145.01
Row22	total_pymnt	cluster_1	1183	0	0	12,704.793	11,604.799	337.4	0.95	12,042.823	13,366.763	0	52,189.51
Row23	total_pymnt	Total	106485	0	0	7,558.313	7,894.445	24.192	0.95	7,510.896	7,605.729	0	55,145.01
Row24	total_rec_prncp	cluster_0	77618	0	0	4,933.608	5,503.527	19.754	0.95	4,894.89	4,972.326	0	35,000
Row25	total_rec_prncp	cluster_2	27684	0	0	7,909.262	8,576.694	51.547	0.95	7,808.227	8,010.297	0	35,000.03
Row26	total_rec_prncp	cluster_1	1183	0	0	9,851.494	10,025.772	291.491	0.95	9,279.596	10,423.392	0	35,000
Row27	total_rec_prncp	Total	106485	0	0	5,761.855	6,648.441	20.374	0.95	5,721.922	5,801.787	0	35,000.03
Row28	recoveries	cluster_0	77618	0	0	43.115	369.556	1.326	0.95	40.515	45.715	0	24,862.1
Row29	recoveries	cluster_2	27684	0	0	53.207	518.493	3.116	0.95	47.099	59.315	0	31,900.52
Row30	recoveries	cluster_1	1183	0	0	64.126	562.32	16.349	0.95	32.05	96.203	0	13,000
Row31	recoveries	Total	106485	0	0	45.972	415.898	1.275	0.95	43.474	48.47	0	31,900.52
Row32	installment	cluster_0	77618	0	0	377.537	197.326	0.708	0.95	376.149	378.925	15.67	1,445.46
Row33	installment	cluster_2	27684	0	0	592.218	279.375	1.679	0.95	588.927	595.509	25.28	1,424.57
Row34	installment	cluster_1	1183	0	0	758.871	295.646	8.596	0.95	742.006	775.735	47.43	1,343.92
Row35	installment	Total	106485	0	0	437.586	244.202	0.748	0.95	436.12	439.053	15.67	1,445.46

The null hypothesis is rejected in all of the variables since the p-value is less than 0.05, indicating that there are significant differences in employee length, annual income, income category, interest rate, debt-to-income ratio, total payment, total received principal, recoveries and instalment (emp_length_int, annual_inc, income_cat, interest rate, debt-to-income ratio, total_pymnt, total_rec_prncp, recoveries, installment) between the groups. We can further see in descriptive statistics that the means of the clusters i.e. for K=4 are statistically different.

K=4

	Source	Sum of Squares	df	Mean Square	F	p-value
emp_length_int	Between Groups	7,488.6818	3	2,496.2273	204.046	0.0
emp_length_int	Within Groups	1,302,651.0533	106481	12.2336		
emp_length_int	Total	1,310,139.7351	106484			
annual_inc	Between Groups	3.23E14	3	1.08E14	122,250.8975	0.0
annual_inc	Within Groups	9.38E13	106481	8.81E8		
annual_inc	Total	4.17E14	106484			
loan_amount	Between Groups	1.47E12	3	4.90E11	8,572.7845	0.0
loan_amount	Within Groups	6.09E12	106481	57,209,937.1904		
loan_amount	Total	7.56E12	106484			
interest_rate	Between Groups	16,229.5326	3	5,409.8442	284.0698	0.0
interest_rate	Within Groups	2,027,831.5718	106481	19.0441		
interest_rate	Total	2,044,061.1044	106484			
dti	Between Groups	298,909.0647	3	99,636.3549	1,499.0616	0.0
dti	Within Groups	7,077,346.5828	106481	66.4658		
dti	Total	7,376,255.6475	106484			
total_pymnt	Between Groups	3.67E11	3	1.22E11	2,075.2397	0.0
total_pymnt	Within Groups	6.27E12	106481	58,881,349.6729		
total_pymnt	Total	6.64E12	106484			
total_rec_prncp	Between Groups	2.26E11	3	7.53E10	1,788.5483	0.0
total_rec_prncp	Within Groups	4.48E12	106481	42,082,450.384		
total_rec_prncp	Total	4.71E12	106484			
recoveries	Between Groups	4,684,022.2877	3	1,561,340.7626	9.0286	5.66E-6
recoveries	Within Groups	1.84E10	106481	172,932.0244		
recoveries	Total	1.84E10	106484			
installment	Between Groups	1.17E9	3	3.91E8	8,032.1596	0.0
installment	Within Groups	5.18E9	106481	48,631.3249		
installment	Total	6.35E9	106484			

Descriptive Statistics: K=4

Row ID	Test Col...	Group	N	Missing ...	Missing ...	Mean	Standa...	Standa...	Confide...	Confide...	Confide...	Minimum	Maximum
Row0	emp_length_int	cluster_0	73256	0	0	5.878	3.477	0.013	0.95	5.853	5.904	0.5	10
Row1	emp_length_int	cluster_2	30775	0	0	6.464	3.542	0.02	0.95	6.424	6.504	0.5	10
Row2	emp_length_int	cluster_1	2448	0	0	6.137	3.546	0.072	0.95	5.997	6.278	0.5	10
Row3	emp_length_int	cluster_3	6	0	0	8.667	2.422	0.989	0.95	6.125	11.209	4	10
Row4	emp_length_int	Total	106485	0	0	6.054	3.508	0.011	0.95	6.033	6.075	0.5	10
Row5	annual_inc	cluster_0	73256	0	0	52,406.316	16,289.448	60.185	0.95	52,288.355	52,524.278	3,000	85,000
Row6	annual_inc	cluster_2	30775	0	0	112,116.054	25,985.587	148.127	0.95	111,825.719	112,406.388	77,000	197,000
Row7	annual_inc	cluster_1	2448	0	0	282,140.083	132,181.575	2,671.562	0.95	276,901.327	287,378.839	197,000	1,782,000
Row8	annual_inc	cluster_3	6	0	0	4,876,679.5	1,473,247.065	601,450.596	0.95	3,330,601.524	6,422,757.476	3,000,000	7,500,000
Row9	annual_inc	Total	106485	0	0	75,216.112	62,577.359	191.767	0.95	74,840.252	75,591.972	3,000	7,500,000
Row10	loan_amount	cluster_0	73256	0	0	12,321.378	6,700.511	24.756	0.95	12,272.856	12,369.9	500	35,000
Row11	loan_amount	cluster_2	30775	0	0	19,896.55	9,183.28	52.348	0.95	19,793.946	19,999.154	1,000	35,000
Row12	loan_amount	cluster_1	2448	0	0	24,317.218	9,199.44	185.933	0.95	23,952.616	24,681.82	1,375	35,000
Row13	loan_amount	cluster_3	6	0	0	20,600	10,111.38	4,127.953	0.95	9,988.758	31,211.242	8,000	35,000
Row14	loan_amount	Total	106485	0	0	14,786.903	8,427.684	25.826	0.95	14,736.283	14,837.522	500	35,000
Row15	interest_rate	cluster_0	73256	0	0	13.497	4.267	0.016	0.95	13.466	13.527	5.32	28.99
Row16	interest_rate	cluster_2	30775	0	0	12.677	4.561	0.026	0.95	12.626	12.728	5.32	28.99
Row17	interest_rate	cluster_1	2448	0	0	12.425	4.685	0.095	0.95	12.239	12.61	5.32	28.99
Row18	interest_rate	cluster_3	6	0	0	10.967	4.652	1.899	0.95	6.085	15.848	7.26	19.52
Row19	interest_rate	Total	106485	0	0	13.235	4.381	0.013	0.95	13.209	13.261	5.32	28.99
Row20	dti	cluster_0	73256	0	0	19.19	8.409	0.031	0.95	19.129	19.251	0	104
Row21	dti	cluster_2	30775	0	0	16.226	7.631	0.043	0.95	16.141	16.312	0	42.64
Row22	dti	cluster_1	2448	0	0	11.634	6.549	0.132	0.95	11.375	11.894	0	38.52
Row23	dti	cluster_3	6	0	0	0.508	0.752	0.307	0.95	-0.281	1.298	0.04	2.02
Row24	dti	Total	106485	0	0	18.159	8.323	0.026	0.95	18.109	18.209	0	104
Row25	total_pymnt	cluster_0	73256	0	0	6,320.224	6,312.973	23.324	0.95	6,274.508	6,365.94	0	51,003.725
Row26	total_pymnt	cluster_2	30775	0	0	10,162.094	9,936.054	56.639	0.95	10,051.079	10,273.108	0	55,145.01
Row27	total_pymnt	cluster_1	2448	0	0	11,876.162	11,290.677	228.199	0.95	11,428.678	12,323.645	0	52,189.51
Row28	total_pymnt	cluster_3	6	0	0	6,893.887	5,614.697	2,292.191	0.95	1,001.623	12,786.15	756.97	16,184.97
Row29	total_pymnt	Total	106485	0	0	7,558.313	7,894.445	24.192	0.95	7,510.896	7,605.729	0	55,145.01
Row30	total_rec_prncp	cluster_0	73256	0	0	4,791.017	5,305.681	19.603	0.95	4,752.595	4,829.438	0	35,000
Row31	total_rec_prncp	cluster_2	30775	0	0	7,798.816	8,430.795	48.058	0.95	7,704.62	7,893.013	0	35,000.03
Row32	total_rec_prncp	cluster_1	2448	0	0	9,207.328	9,722.919	196.513	0.95	8,821.979	9,592.676	0	35,000
Row33	total_rec_prncp	cluster_3	6	0	0	5,380.6	5,360.598	2,188.455	0.95	-245.002	11,006.202	586.18	15,600
Row34	total_rec_prncp	Total	106485	0	0	5,761.855	6,648.441	20.374	0.95	5,721.922	5,801.787	0	35,000.03
Row35	recoveries	cluster_0	73256	0	0	41.515	347.971	1.286	0.95	38.995	44.035	0	24,862.1
Row36	recoveries	cluster_2	30775	0	0	55.751	538.334	3.069	0.95	49.737	61.766	0	31,900.52
Row37	recoveries	cluster_1	2448	0	0	56.527	505.611	10.219	0.95	36.488	76.566	0	13,000
Row38	recoveries	cluster_3	6	0	0	0	0	0	0.95	0	0	0	0
Row39	recoveries	Total	106485	0	0	45.972	415.898	1.275	0.95	43.474	48.47	0	31,900.52
Row40	installment	cluster_0	73256	0	0	368.461	189.593	0.7	0.95	367.088	369.834	15.67	1,445.46
Row41	installment	cluster_2	30775	0	0	579.018	274.831	1.567	0.95	575.948	582.089	25.28	1,424.57
Row42	installment	cluster_1	2448	0	0	727.66	299.958	6.063	0.95	715.772	739.548	47.43	1,409.99
Row43	installment	cluster_3	6	0	0	632.68	323.562	132.094	0.95	293.123	972.237	255.04	1,115.77
Row44	installment	Total	106485	0	0	437.586	244.202	0.748	0.95	436.12	439.053	15.67	1,445.46

The null hypothesis is rejected in all of the variables since the p-value is less than 0.05, indicating that there are significant differences in employee length, annual income, income category, interest rate, debt-to-income ratio, total payment, total received principal, recoveries and instalment (emp_length_int, annual_inc, income_cat, interest rate, debt-to-income ratio, total_pymnt, total_rec_prncp, recoveries, installment) between the groups. We can further see in descriptive statistics that the means of the clusters i.e. for K=4 are statistically different.

K=5

	Source	Sum of Squares	df	Mean Square	F	p-value
emp_length_int	Between Groups	13,684.0465	4	3,421.0116	280.9732	0.0
emp_length_int	Within Groups	1,296,455.6886	106480	12.1756		
emp_length_int	Total	1,310,139.7351	106484			
annual_inc	Between Groups	3.53E14	4	8.82E13	145,861.1067	0.0
annual_inc	Within Groups	6.44E13	106480	6.04E8		
annual_inc	Total	4.17E14	106484			
loan_amount	Between Groups	1.75E12	4	4.37E11	8,006.5249	0.0
loan_amount	Within Groups	5.81E12	106480	54,604,936.0625		
loan_amount	Total	7.56E12	106484			
interest_rate	Between Groups	15,779.3895	4	3,944.8474	207.0952	0.0
interest_rate	Within Groups	2,028,281.7149	106480	19.0485		
interest_rate	Total	2,044,061.1044	106484			
dti	Between Groups	328,335.3731	4	82,083.8433	1,240.1229	0.0
dti	Within Groups	7,047,920.2744	106480	66.1901		
dti	Total	7,376,255.6475	106484			
total_pymnt	Between Groups	4.39E11	4	1.10E11	1,883.8191	0.0
total_pymnt	Within Groups	6.20E12	106480	58,205,566.3444		
total_pymnt	Total	6.64E12	106484			
total_rec_prncp	Between Groups	2.65E11	4	6.62E10	1,587.6856	0.0
total_rec_prncp	Within Groups	4.44E12	106480	41,715,411.8497		
total_rec_prncp	Total	4.71E12	106484			
recoveries	Between Groups	6,646,587.6091	4	1,661,646.9023	9.6096	9.13E-8
recoveries	Within Groups	1.84E10	106480	172,915.2172		
recoveries	Total	1.84E10	106484			
installment	Between Groups	1.38E9	4	3.44E8	7,364.7425	0.0
installment	Within Groups	4.97E9	106480	46,713.2845		
installment	Total	6.35E9	106484			

Descriptive Statistics: K=5

Row ID	Test Col...	Group	N	Missing ...	Missing ...	Mean	Standa...	Standa...	Confide...	Confide...	Confide...	Minimum	Maximum
Row0	emp_length_int	cluster_4	61790	0	0	5.759	3.461	0.014	0.95	5.732	5.786	0.5	10
Row1	emp_length_int	cluster_0	36920	0	0	6.522	3.518	0.018	0.95	6.487	6.558	0.5	10
Row2	emp_length_int	cluster_2	7336	0	0	6.15	3.584	0.042	0.95	6.068	6.232	0.5	10
Row3	emp_length_int	cluster_1	433	0	0	6.513	3.413	0.164	0.95	6.191	6.835	0.5	10
Row4	emp_length_int	cluster_3	6	0	0	8.667	2.422	0.989	0.95	6.125	11.209	4	10
Row5	emp_length_int	Total	106485	0	0	6.054	3.508	0.011	0.95	6.033	6.075	0.5	10
Row6	annual_inc	cluster_4	61790	0	0	47,975.267	13,665.686	54.976	0.95	47,867.514	48,083.02	3,000	74,000
Row7	annual_inc	cluster_0	36920	0	0	94,281.163	17,319.076	90.135	0.95	94,104.495	94,457.83	64,800	138,000
Row8	annual_inc	cluster_2	7336	0	0	180,408.77	41,985.867	490.2	0.95	179,447.837	181,369.704	136,115	334,000
Row9	annual_inc	cluster_1	433	0	0	488,213.848	203,896.693	9,798.65	0.95	468,954.891	507,472.805	335,000	1,782,000
Row10	annual_inc	cluster_3	6	0	0	4,876,679.5	1,473,247.065	601,450.596	0.95	3,330,601.524	6,422,757.476	3,000,000	7,500,000
Row11	annual_inc	Total	106485	0	0	75,216.112	62,577.359	191.767	0.95	74,840.252	75,591.972	3,000	7,500,000
Row12	loan_amount	cluster_4	61790	0	0	11,466.865	6,057.075	24.367	0.95	11,419.106	11,514.625	500	35,000
Row13	loan_amount	cluster_0	36920	0	0	18,617.541	8,817.183	45.888	0.95	18,527.599	18,707.482	1,000	35,000
Row14	loan_amount	cluster_2	7336	0	0	22,815.884	9,355.344	109.227	0.95	22,601.768	23,030	1,000	35,000
Row15	loan_amount	cluster_1	433	0	0	25,831.928	8,967.216	430.937	0.95	24,984.935	26,678.922	1,500	35,000
Row16	loan_amount	cluster_3	6	0	0	20,600	10,111.38	4,127.953	0.95	9,988.758	31,211.242	8,000	35,000
Row17	loan_amount	Total	106485	0	0	14,786.903	8,427.684	25.826	0.95	14,736.283	14,837.522	500	35,000
Row18	interest_rate	cluster_4	61790	0	0	13.546	4.217	0.017	0.95	13.513	13.579	5.32	28.99
Row19	interest_rate	cluster_0	36920	0	0	12.89	4.55	0.024	0.95	12.843	12.936	5.32	28.99
Row20	interest_rate	cluster_2	7336	0	0	12.403	4.599	0.054	0.95	12.298	12.509	5.32	28.99
Row21	interest_rate	cluster_1	433	0	0	12.409	4.83	0.232	0.95	11.953	12.865	5.32	25.89
Row22	interest_rate	cluster_3	6	0	0	10.967	4.652	1.899	0.95	6.085	15.848	7.26	19.52
Row23	interest_rate	Total	106485	0	0	13.235	4.381	0.013	0.95	13.209	13.261	5.32	28.99
Row24	dti	cluster_4	61790	0	0	19.415	8.477	0.034	0.95	19.348	19.482	0	104
Row25	dti	cluster_0	36920	0	0	17.054	7.782	0.041	0.95	16.975	17.134	0	42.64
Row26	dti	cluster_2	7336	0	0	13.71	6.977	0.081	0.95	13.55	13.869	0	39.94
Row27	dti	cluster_1	433	0	0	8.702	5.839	0.281	0.95	8.15	9.253	0	36.81
Row28	dti	cluster_3	6	0	0	0.508	0.752	0.307	0.95	-0.281	1.298	0.04	2.02
Row29	dti	Total	106485	0	0	18.159	8.323	0.026	0.95	18.109	18.209	0	104
Row30	total_pymnt	cluster_4	61790	0	0	5,878.515	5,770.114	23.213	0.95	5,833.018	5,924.012	0	42,695.93
Row31	total_pymnt	cluster_0	36920	0	0	9,566.659	9,323.044	48.521	0.95	9,471.558	9,661.761	0	55,145.01
Row32	total_pymnt	cluster_2	7336	0	0	11,246.693	10,909.957	127.378	0.95	10,996.996	11,496.389	0	52,770.344
Row33	total_pymnt	cluster_1	433	0	0	13,545.798	11,619.176	558.382	0.95	12,448.315	14,643.281	0	49,618.68
Row34	total_pymnt	cluster_3	6	0	0	6,893.887	5,614.697	2,292.191	0.95	1,001.623	12,786.15	756.97	16,184.97
Row35	total_pymnt	Total	106485	0	0	7,558.313	7,894.445	24.192	0.95	7,510.896	7,605.729	0	55,145.01
Row36	total_rec_prncp	cluster_4	61790	0	0	4,461.362	4,864.791	19.571	0.95	4,423.003	4,499.72	0	35,000
Row37	total_rec_prncp	cluster_0	36920	0	0	7,298.192	7,891.373	41.07	0.95	7,217.694	7,378.689	0	35,000.03
Row38	total_rec_prncp	cluster_2	7336	0	0	8,703.101	9,316.97	108.779	0.95	8,489.863	8,916.339	0	35,000.02
Row39	total_rec_prncp	cluster_1	433	0	0	10,522.168	10,045.775	482.769	0.95	9,573.299	11,471.036	0	35,000
Row40	total_rec_prncp	cluster_3	6	0	0	5,380.6	5,360.598	2,188.455	0.95	-245.002	11,006.202	586.18	15,600
Row41	total_rec_prncp	Total	106485	0	0	5,761.855	6,648.441	20.374	0.95	5,721.922	5,801.787	0	35,000.03
Row42	recoveries	cluster_4	61790	0	0	39.415	328.567	1.322	0.95	36.824	42.006	0	24,862.1
Row43	recoveries	cluster_0	36920	0	0	56.223	515.627	2.684	0.95	50.963	61.482	0	31,900.52
Row44	recoveries	cluster_2	7336	0	0	49.626	503.684	5.881	0.95	38.098	61.153	0	14,004.01
Row45	recoveries	cluster_1	433	0	0	46.388	387.655	18.63	0.95	9.773	83.004	0	4,273.82
Row46	recoveries	cluster_3	6	0	0	0	0	0	0.95	0	0	0	0
Row47	recoveries	Total	106485	0	0	45.972	415.898	1.275	0.95	43.474	48.47	0	31,900.52
Row48	installment	cluster_4	61790	0	0	345.425	171.251	0.689	0.95	344.075	346.775	15.67	1,223.45
Row49	installment	cluster_0	36920	0	0	541.106	259.848	1.352	0.95	538.455	543.756	30.54	1,445.46
Row50	installment	cluster_2	7336	0	0	672.589	292.997	3.421	0.95	665.884	679.295	25.28	1,409.99
Row51	installment	cluster_1	433	0	0	778.37	300.199	14.427	0.95	750.015	806.725	54.98	1,343.08
Row52	installment	cluster_3	6	0	0	632.68	323.562	132.094	0.95	293.123	972.237	255.04	1,115.77
Row53	installment	Total	106485	0	0	437.586	244.202	0.748	0.95	436.12	439.053	15.67	1,445.46

The null hypothesis is rejected in all of the variables since the p-value is less than 0.05, indicating that there are significant differences in employee length, annual income, income category, interest rate, debt-to-income ratio, total payment, total received principal, recoveries and instalment (emp_length_int, annual_inc, income_cat, interest rate, debt-to-income ratio, total_pymnt, total_rec_prncp, recoveries, installment) between the groups. We can further see in descriptive statistics that the means of the clusters i.e. for K=5 are statistically different.

4. Results and observation

4.1 Appropriate Number of Segments or Clusters

Cluster number	Cluster	Silhouette Score	Mean
	Cluster 0	0.681	
2	Cluster 1	0.312	0.632
	Cluster 0	0.609	
	Cluster 1	0.355	
3	Cluster 2	0.266	0.539
	Cluster 0	0.59	
	Cluster 1	0.367	
	Cluster 2	0.32	
4	Cluster 3	0.615	0.519
	Cluster 0	0.556	
	Cluster 1	0.361	
	Cluster 2	0.367	
	Cluster 3	0.391	
5	Cluster 4	0.593	0.475

The silhouette score for all the clusters is present. The analysis of the table will be done on 2 factors: -

1) Higher the silhouette score i.e. close to 1 more are the clusters separated and close to 0 indicates the clusters are overlapping
2) Sometimes having a smaller number of clusters can be very simplistic and the service provider may take simple decisions according it which will eventually hamper their market penetration and having simplified services/products may forgone the people who are the potential customers. Having more services will give the service provider a unique value proposition to attract customers.

Despite clustering with 2 has a higher silhouette score it will eventually affect the service provider in the long run as new service provider will try to copy these minimal number of services affecting their position in the market.

Thus, we will take clustering with 3 as the appropriate number of clusters despite having a lower silhouette score. The reason for taking appropriate segments and clusters as 3 because in finance (including insurance, mutual funds and banking services), the greater number of services the better the hold in the market and happier the customers as they have a service which is specifically made for them.

The resources used will be higher for clustering with 3 than 2, but the variety of services will attract the customers more.

4.2 Cluster analysis

4.2.1 Categorical Variables

It has been observed that all the variables are contributing to the cluster for making the service or product. This is because the p-value is less than 0.05 (confidence level at 95% for the model) which in turn tells that all the categorical variables are significant for the process of making the clusters.

4.2.2 Non-Categorical Variables

It has been observed that all the variables are contributing to the cluster for making the service or product. This is because the p-value is less than 0.05 (confidence level at 95% for the model) which in turn tells that all the non-categorical variables are significant for the process of making the clusters.

5. Managerial Insights

5.1 The managerial insights that can concluded by doing the k-means clustering as well as selecting the appropriate number of clusters as 3 are: -

→Insights for cluster 1 which represent low-Income and high-risk borrowers

- The customers have lower incomes and higher risk profiles, focus on implementing risk mitigation strategies such as stricter underwriting criteria and risk-based pricing for loans will help the bank to mitigate non-performing assets as well as higher rewards if the customer is successful in his/her venture.
- This cluster may require the institutional bank (can be a private or public bank or a NBFC) to offer financial literacy programs as well as budgeting tools to help customers manage their debt and improve their financial well-being. It is essential for those who want to venture into a start-up but don't have enough seed money.
- Developing alternative financial products such as micro-loans or secured credit cards tailored to the needs of this segment will help to provide access to credit while minimizing risk for the institutional bank.
- Providing credit counselling services to assist customers in improving their credit scores and financial stability.

→Insights for cluster 2 which represent the high-Income and low-risk borrowers

- The customers represent high incomes with good credit profiles. This segment represents financially stable individuals who are likely to qualify for premium financial products and services.
- Offering exclusive and tailored loan products with competitive interest rates and flexible terms for major purchases to attract and retain these high-value customers.
- Personalized Services can be personalized financial advice and wealth management services to help them maximize their wealth and achieve their financial goals.

- There are opportunities to cross-sell investment products, insurance and other high-value financial services to increase customer engagement and loyalty.

→Insights for cluster 3 which represent medium-income and moderate-risk borrowers

- In this cluster offering customized loan products and financial solutions tailored to the needs and preferences of this segment such as flexible repayment options and rewards programs will be beneficial to the institutional bank.
- The institutional bank can provide financial planning services to help customers achieve their short-term and long-term financial goals such as saving for retirement or purchasing a home.
- Implement customer retention strategies to maintain loyalty and prevent losing out on repeat customers such as offering incentives for referrals and regular financial health check-ups.
- There are opportunities to upsell additional financial products and services based on customer needs and life stages such as mortgages, insurance and investment products.

5.2 Cluster (Heterogenous) Identity

Identity of cluster 1: Debt-Ridden customers who are struggling for solvency and even to those that want to start a small business

Identity of cluster 2: Affluent Purchasers who are present in the upper strata in terms of income

Identity of cluster 3: Middle-Class Consumers represent the majority of customers who may contribute to maximum for the institutional bank in terms of profit. They represent customers who have a steady stability.