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 8,87,329 (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	Relative Frequency in %
ANY	3	0
MORTGAGE	443557	49.985
NONE	50	0.006
OTHER	182	0.021
OWN	87470	9.857
RENT	356117	40.131

Term

Row ID	count	Relative Frequency in %
36 months	621125	69.995
60 months	266254	30.005

application_type

Row ID	count	Relative Frequency in %
INDIVIDUAL	886868	99.942
JOINT	511	0.058

Purpose

Row ID	count	Relative Frequency in %
car	8863	0.999
credit_card	206182	23.235
debt_consolidation	524215	59.075
educational	423	0.048
home_improvement	51829	5.841
house	3707	0.418
major_purchase	17277	1.947
medical	8540	0.962
moving	5414	0.61
other	42894	4.834
renewable_energy	575	0.065
small_business	10377	1.169
vacation	4736	0.534
wedding	2347	0.264

loan_condition

Row ID	count	Relative Frequency in %
Bad Loan	67429	7.599
Good Loan	819950	92.401

income_category

Row ID	count	Relative Frequency in %
High	16786	1.892
Low	729616	82.221
Medium	140977	15.887

interest_payments

Row ID	count	Relative Frequency in %
Low	465316	52.437
High	422063	47.563

Grade

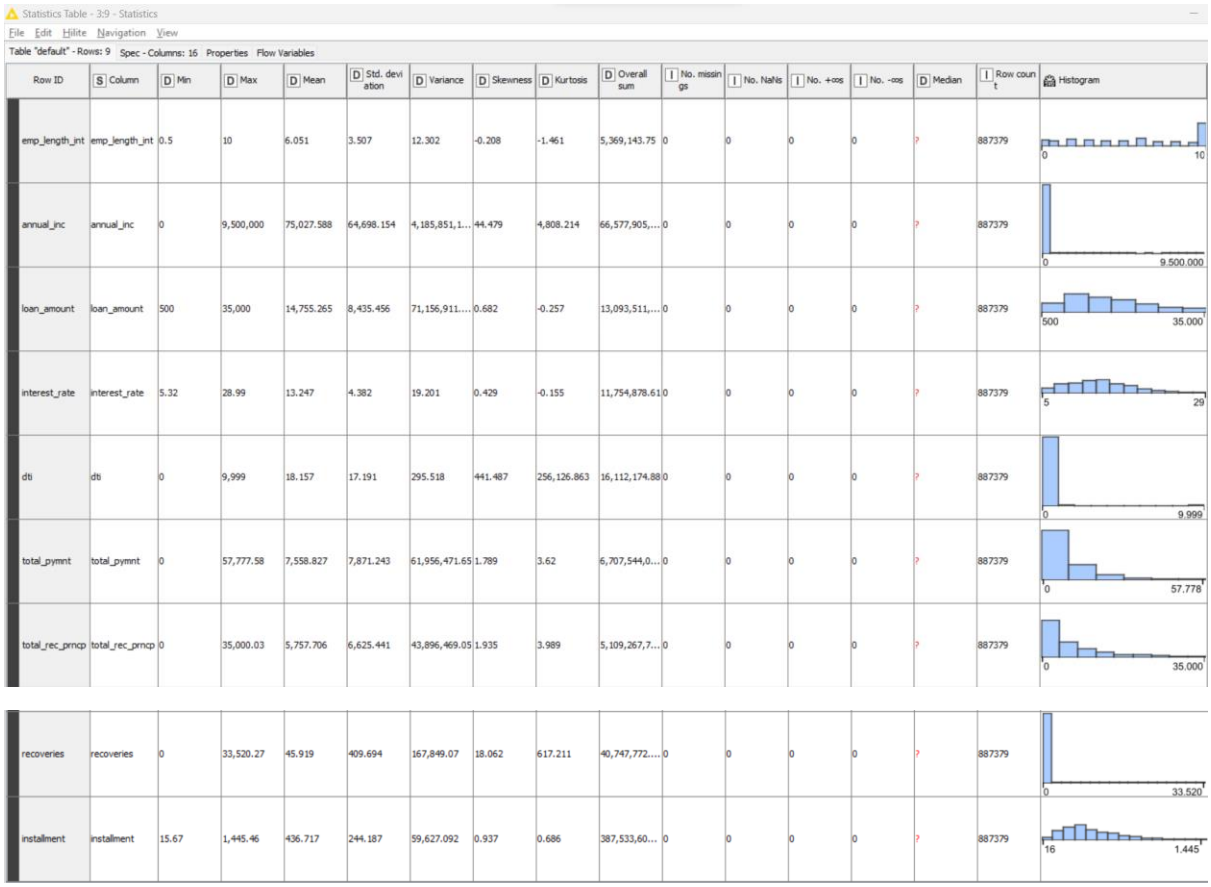
Row ID	count	Relative Frequency in %
A	148202	16.701
B	254535	28.684
C	245860	27.706
D	139542	15.725
E	70705	7.968
F	23046	2.597
G	5489	0.619

2.3.2 Descriptive Statistics: Non-Categorical Variables

2.3.2.1 Measures of Central Tendency and Dispersion

Variable Name	Type	Missing Values	Unique Values	Mean	Mean Absolute Deviation
emp_length_int	Number (double)	0	12	6.05056436	3.114290891
annual_inc	Number (integer)	0	45784	75027.5879	31746.89615
loan_amount	Number (integer)	0	1372	14755.2646	6870.451906
interest_rate	Number (double)	0	542	13.2467397	3.511116328
dti	Number (double)	0	4086	18.1570387	6.840026344
total_pymnt	Number (double)	0	505628	7558.82668	5883.399367
total_rec_prncp	Number (double)	0	260227	5757.70642	4878.070446
recoveries	Number (double)	0	23055	45.9192434	89.395893
installment	Number (double)	0	68711	436.717127	193.431199

Variable Name	Minimum	Maximum	25% quartile	50% quartile	75% quartile	Mean	Standard Deviation	Variance	Skewness
emp_length_int	0.5	10	3	6.05	10	6.05056436	3.507404696	12.3018877	-0.2079422
annual_inc	0	9500000	45000	65000	90000	75027.5879	64698.15428	4.19E+09	44.47843683
loan_amount	500	35000	8000	13000	20000	14755.2646	8435.455601	7.12E+07	0.681680 84
interest_rate	5.32	28.99	9.99	12.99	16.2	13.2467397	4.381867415	19.200762	0.429479176
dti	0	9999	11.91	17.65	23.95	18.1570387	17.19062569	295.517612	441.4852825
total_pymnt	0	57777.5799	1914.59	4894.99912	10616.8185	7558.82668	7871.243336	6.20E+07	1.788887581
total_rec_prncp	0	35000.03	1200.57	3215.32	8000	5757.70642	6625.441046	4.39E+07	1.935029 29
recoveries	0	33520.27	0	0	0	45.9192434	409.6938736	167849.07	18.06177786
installment	15.67	1445.46	260.7	382.55	572.6	436.717127	244.1865935	59627.0924	0.936949179



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
Any - 6, 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

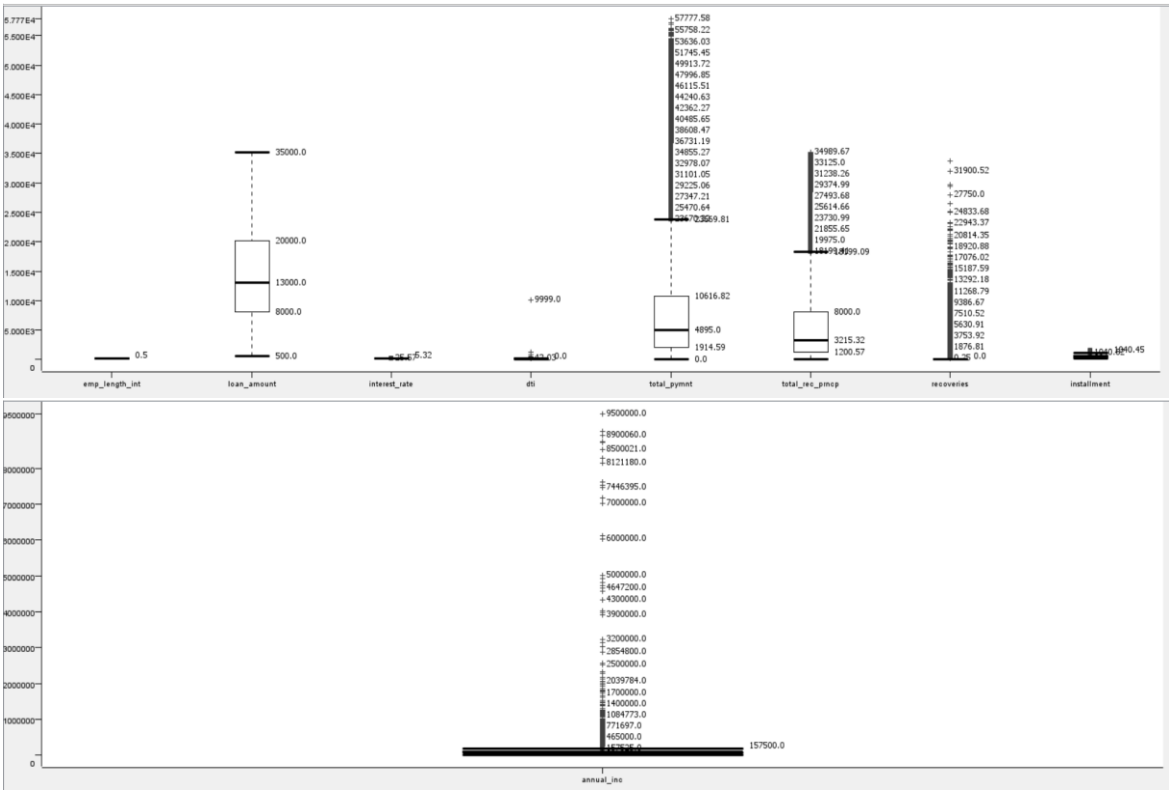
3.1.3 Outlier Statistics and Treatment

3.1.3.1 Outlier Statistics: Non-Categorical Variables

Row ID	Outlier column	Member count	Outlier count	Lower bound	Upper bound
Row0	emp_length_int	887379	0	-7.5	20.5
Row1	annual_inc	887379	39719	-22,500	157,500
Row2	loan_amount	887379	0	-10,000	38,000
Row3	interest_rate	887379	6308	0.675	25.515
Row4	dti	887379	81	-6.15	42.01
Row5	total_pymnt	887379	46324	-11,138.743	23,670.145
Row6	total_rec_prncp	887379	56944	-8,998.575	18,199.145
Row7	recoveries	887379	24677	0	0
Row8	installment	887379	23147	-207.15	1,040.45

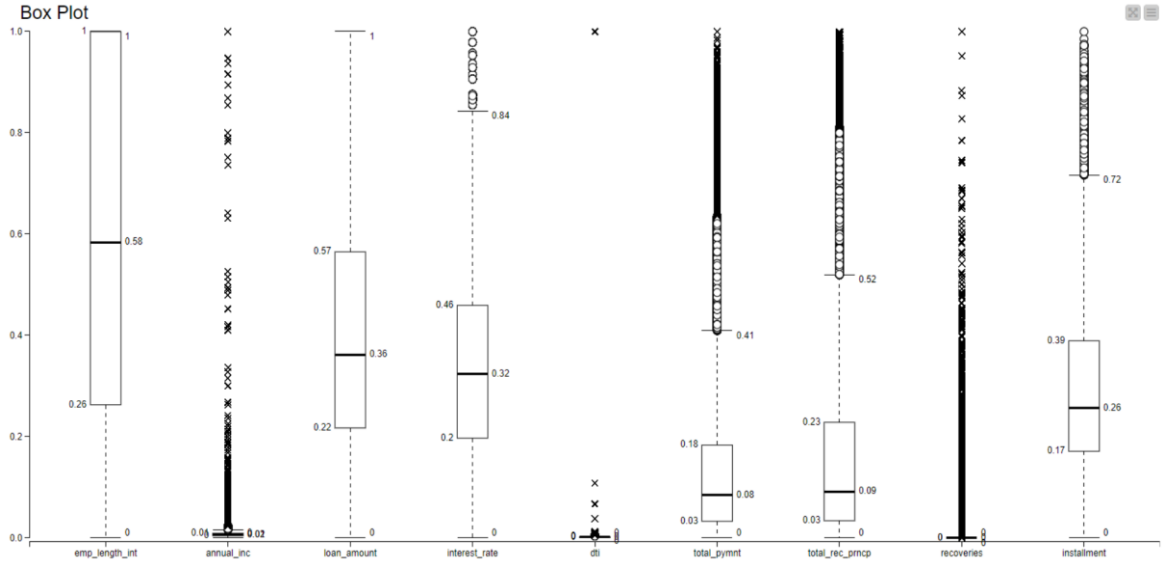
3.1.3.2 Normalization using Min-Max Scaler

Before Normalization



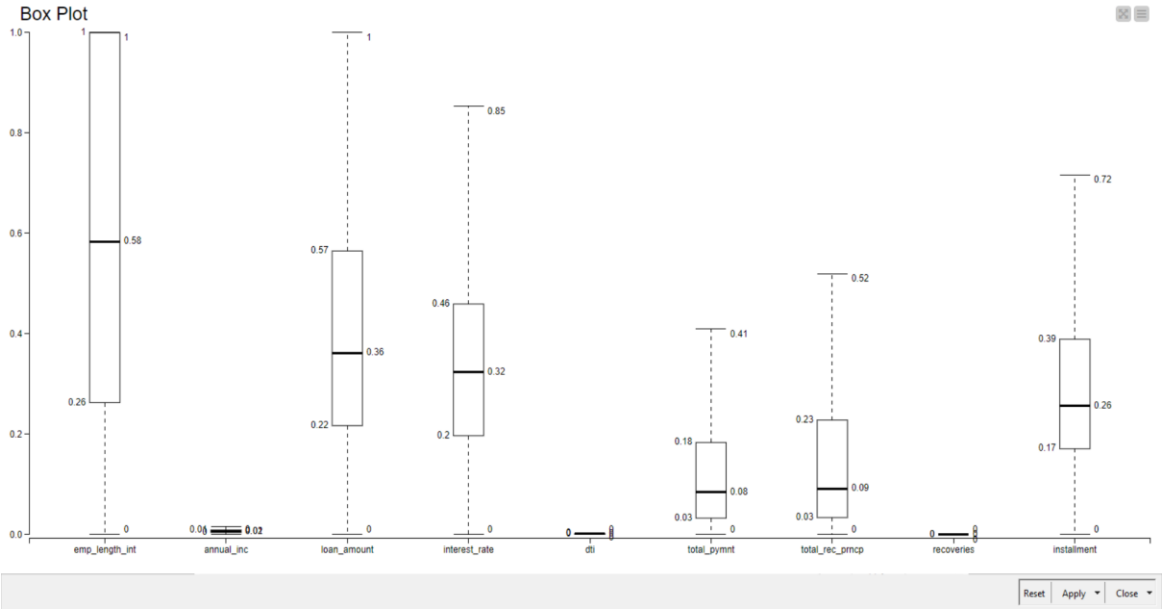
After Normalization

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.

3.1.4 Data Bifurcation

The training and testing data is bifurcated into 80% and 20%.

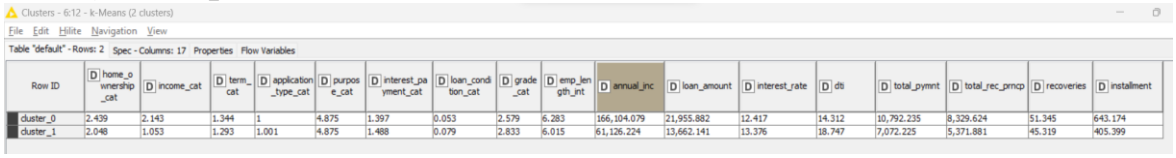
3.2 Data Analysis

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)

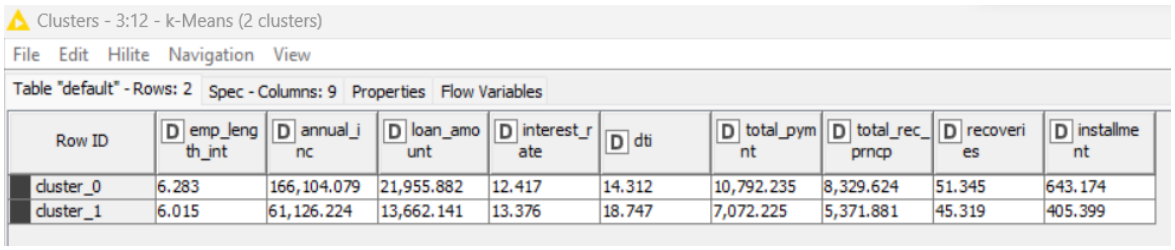


Clusters - 6:12 - k-Means (2 clusters)

File Edit Hilite Navigation View

Table "default" - Rows: 2 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	2.439	2.143	1.344	1	4.875	1.397	0.053	2.579	6.283	166,104.079	21,955.882	12.417	14.312	10,792.235	8,329.624	51.345	643.174
cluster_1	2.048	1.053	1.293	1.001	4.875	1.488	0.079	2.833	6.015	61,126.224	13,662.141	13.376	18.747	7,072.225	5,371.881	45.319	405.399



Clusters - 3:12 - k-Means (2 clusters)

File Edit Hilite Navigation View

Table "default" - Rows: 2 Spec - Columns: 9 Properties Flow Variables

Row ID	emp_length_int	annual_inc	loan_amount	interest_rate	dti	total_pymnt	total_rec_prncp	recoveries	installment
cluster_0	6.283	166,104.079	21,955.882	12.417	14.312	10,792.235	8,329.624	51.345	643.174
cluster_1	6.015	61,126.224	13,662.141	13.376	18.747	7,072.225	5,371.881	45.319	405.399

Cluster 1

- 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: 6.28 years
- annual_inc: $166,104.08
- loan_amount: $21,955.88
- interest_rate: 12.42%
- dti: 14.31%
- total_pymnt: $10,792.24
- total_rec_prncp: $8,329.62

- recoveries: $51.35
- installment: $643.17

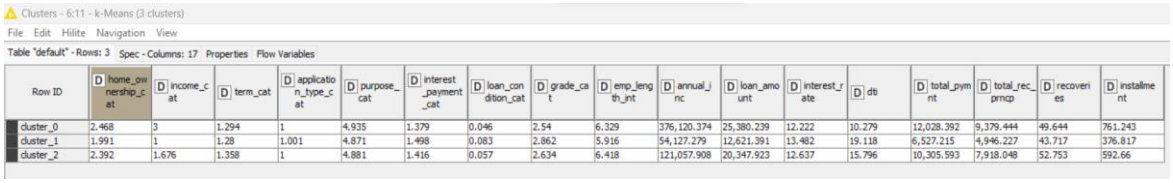
Cluster 2

- 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: 6.01 years
- annual_inc: $61,126.22
- loan_amount: $13,662.14
- interest_rate: 13.38%
- dti: 18.75%
- total_pymnt: $7,072.22
- total_rec_prncp: $5,371.88
- recoveries: $45.32
- installment: $405.40

Analysis of Cluster 1 and 2 (K=2)

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

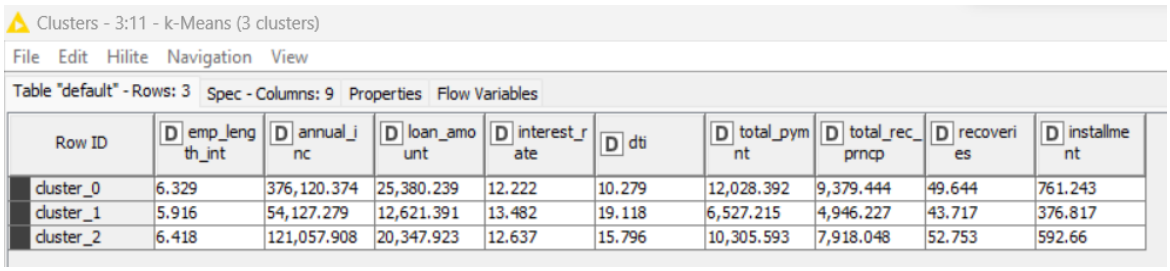


Clusters - 6:11 - k-Means (3 clusters)

File Edit Hilite Navigation View

Table "default" - Rows: 3 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	2.468	3	1.294	1	4.935	1.379	0.046	2.54	6.329	376,120.374	25,380.239	12.222	10.279	12,028.392	9,379.444	49.644	761.243
cluster_1	1.991	1	1.28	1.001	4.871	1.498	0.083	2.862	5.916	54,127.279	12,621.391	13.482	19.118	6,527.215	4,946.227	43.717	376.817
cluster_2	2.392	1.676	1.358	1	4.881	1.416	0.057	2.634	6.418	121,057.908	20,347.923	12.637	15.796	10,305.593	7,918.048	52.753	592.66



Clusters - 3:11 - k-Means (3 clusters)

File Edit Hilite Navigation View

Table "default" - Rows: 3 Spec - Columns: 9 Properties Flow Variables

Row ID	emp_length_int	annual_inc	loan_amount	interest_rate	dti	total_pymnt	total_rec_prncp	recoveries	installment
cluster_0	6.329	376,120.374	25,380.239	12.222	10.279	12,028.392	9,379.444	49.644	761.243
cluster_1	5.916	54,127.279	12,621.391	13.482	19.118	6,527.215	4,946.227	43.717	376.817
cluster_2	6.418	121,057.908	20,347.923	12.637	15.796	10,305.593	7,918.048	52.753	592.66

Cluster 1

- home_ownership: own
- income_category: High
- Term: 36 months
- application_type: Individual
- Purpose: Major purchase
- interest_payments: Low
- loan_condition: Good Loan
- Grade: B
- emp_length_int: 6.33 years
- annual_inc: $376,120.37
- loan_amount: $25,380.24
- interest_rate: 12.22%
- dti: 10.28%
- total_pymnt: $12,028.39
- total_rec_prncp: $9,379.44
- recoveries: $49.64
- installment: $761.24

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: 5.92 years
- annual_inc: $54,127.28

- loan_amount: $12,621.39
- interest_rate: 13.48%
- dti: 19.12%
- total_pymnt: $6,527.21
- total_rec_prncp: $4,946.23
- recoveries: $43.72
- installment: $376.82

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: 6.42 years
- annual_inc: $121,057.91
- loan_amount: $20,347.92
- interest_rate: 12.64%
- dti: 15.80%
- total_pymnt: $10,305.59
- total_rec_prncp: $7,918.05
- recoveries: $52.75
- installment: $592.66

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

Cluster 1: 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 2: Low-Income, Higher-Risk Borrowers

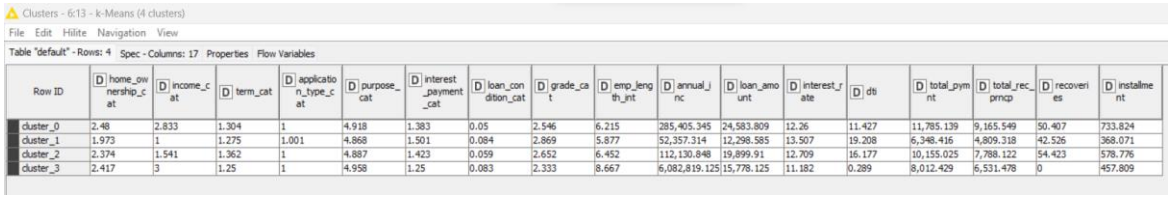
- This cluster comprises individuals with lower incomes, mostly renting their homes and seeking loans for debt consolidation.
- They have higher-interest payments and loans in this cluster are classified as bad loan indicating higher risk.
- The loans in this cluster have higher interest rates and higher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

Clusters - 3:13 - k-Means (4 clusters)

File Edit Hilite Navigation View

Table "default" - Rows: 4 Spec - Columns: 9 Properties Flow Variables

Row ID	emp_length_int	annual_inc	loan_amount	interest_rate	dti	total_pymnt	total_rec_prncp	recoveries	installment
cluster_0	6.215	285,405.345	24,583.809	12.26	11.427	11,785.139	9,165.549	50.407	733.824
cluster_1	5.877	52,357.314	12,298.585	13.507	19.208	6,348.416	4,809.318	42.526	368.071
cluster_2	6.452	112,130.848	19,899.91	12.709	16.177	10,155.025	7,788.122	54.423	578.776
cluster_3	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.

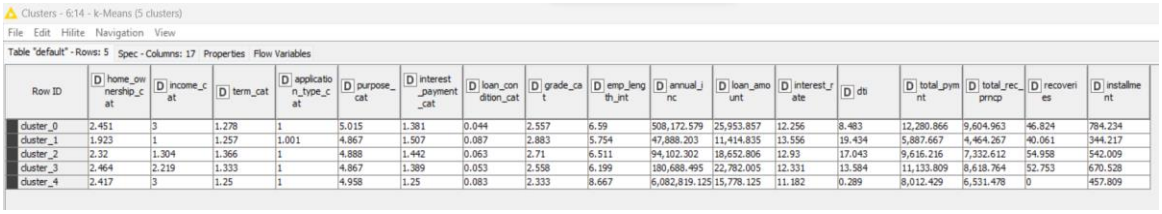
Cluster 3: Medium-Income, Moderate-Risk Borrowers

- 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

Clusters - 3:14 - k-Means (5 clusters)

File Edit Hilite Navigation View

Table "default" - Rows: 5 Spec - Columns: 9 Properties Flow Variables

Row ID	emp_length_int	annual_inc	loan_amount	interest_rate	dti	total_pymnt	total_rec_prncp	recoveries	installment
cluster_0	6.59	508,172.579	25,953.857	12.256	8.483	12,280.866	9,604.963	46.824	784.234
cluster_1	5.754	47,888.203	11,414.835	13.556	19.434	5,887.667	4,464.267	40.061	344.217
cluster_2	6.511	94,102.302	18,652.806	12.93	17.043	9,616.216	7,332.612	54.958	542.009
cluster_3	6.199	180,688.495	22,782.005	12.331	13.584	11,133.809	8,618.764	52.753	670.528
cluster_4	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 Si...
cluster_1	0.675
cluster_0	0.329
Overall	0.628

- Silhouette score of 0.675 indicates that cluster 0 has strong cohesion and is well-separated from other clusters.
- Silhouette score of 0.329 suggests that cluster 1 has moderate cohesion and separation from other clusters.

K=3

Row ID	D Mean Si...
cluster_1	0.362
cluster_0	0.608
cluster_2	0.333
Overall	0.54

- Silhouette score of 0.362 indicates that cluster 0 has moderate to strong cohesion and separation.
- Silhouette score of 0.608 indicates that cluster 1 has strong cohesion and separation making it a well-defined cluster.
- Silhouette score of 0.333 indicates that cluster 2 also shows moderate cohesion and separation

K=4

Row ID	Mean Si...
cluster_1	0.356
cluster_0	0.561
cluster_2	0.357
cluster_3	0.312
Overall	0.478

- Silhouette score of 0.356 indicates that cluster 0 exhibits moderate to strong cohesion and separation.
- Silhouette score of 0.561 indicates that cluster 1 shows strong cohesion and separation, indicating a well-defined cluster.
- Silhouette score of 0.357 indicates that cluster 2 has moderate to strong cohesion and separation.
- Silhouette score of 0.312 suggests moderate cohesion and separation for cluster 3.

K=5

Row ID	Mean Si...
cluster_1	0.334
cluster_0	0.524
cluster_2	0.365
cluster_3	0.324
cluster_4	0.328
Overall	0.429

- Silhouette score of 0.334 indicates that cluster 0 demonstrates moderate cohesion and separation.
- Silhouette score of 0.524 suggests strong cohesion and separation for Cluster 1.
- Silhouette score of 0.365 indicates that cluster 2 shows moderate cohesion and separation.
- Silhouette score of 0.324 indicates that cluster 3 exhibits moderate cohesion and separation.
- Silhouette score of 0.328 indicates moderate cohesion and separation for cluster 4.

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

Table "default" - Rows: 1 | Spec - Columns: 6 | Properties | Flow Variables

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	13,999.036	0.0	421,374.729	532,198.5	344,859.564	319,691.5

Term

Table "default" - Rows: 1 | Spec - Columns: 6 | Properties | Flow Variables

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	995.029	0.0	370,576.982	248,413.5	352,577.901	248,413.5

application_type

Table "default" - Rows: 1 | Spec - Columns: 6 | Properties | Flow Variables

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	18.98	1.3212494161973787...	354,840.687	354,753.5	354,968.913	354,753.5

Purpose

Table "default" - Rows: 1 | Spec - Columns: 6 | Properties | Flow Variables

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	128.233	0.0	361,195.565	426,276.5	354,003.337	426,276.5

loan_condition

Table "default" - Rows: 1 | Spec - Columns: 6 | Properties | Flow Variables

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	798.6	0.0	346,859.278	327,981.5	356,181.629	327,981.5

income_category

Table "default" - Rows: 1 | Spec - Columns: 6 | Properties | Flow Variables

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	504,938.448	0.0	649,124.948	640,083	310,254.625	291,839

interest_payments

Table "default" - Rows: 1 | Spec - Columns: 6 | Properties | Flow Variables

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	2,708.674	0.0	326,860.863	186,066	359,220.238	186,066

Grade

Table "default" - Rows: 1 | Spec - Columns: 6 | Properties | Flow Variables

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	3,816.899	0.0	317,456.061	220,217.5	360,649.227	420,397.5

The p-value associated with the Kruskal-Wallis test is 0.0, 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 0.0 which is 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

Table "default" - Rows: 1 | Spec - Columns: 8 | Properties | Flow Variables

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	25,908.064	0.0	426,634.769	532,198.5	333,602.004	319,691.5	412,370.232	532,198.5

Term

Table "default" - Rows: 1 | Spec - Columns: 8 | Properties | Flow Variables

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	3,912.067	0.0	352,626.039	248,413.5	347,766.074	248,413.5	375,419.387	248,413.5

application_type

Table "default" - Rows: 1 | Spec - Columns: 8 | Properties | Flow Variables

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	34.78	2.802554843750471...	354,798.419	354,753.5	354,987.437	354,753.5	354,858.191	354,753.5

p-value: 2.802 * 10^-5

Purpose

Table "default" - Rows: 1 | Spec - Columns: 8 | Properties | Flow Variables

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	234.633	0.0	370,725.31	426,276.5	353,081.059	426,276.5	359,574.009	426,276.5

loan_condition

Table "default" - Rows: 1 | Spec - Columns: 8 | Properties | Flow Variables

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	1,397.64	0.0	344,152.411	327,981.5	357,470.5	327,981.5	348,279.953	327,981.5

income_category

Table "default" - Rows: 1 | Spec - Columns: 8 | Properties | Flow Variables

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	429,541.296	0.0	703,196	703,196	291,839	291,839	518,802.241	640,083

interest_payments

Table "default" - Rows: 1 | Spec - Columns: 8 | Properties | Flow Variables

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	3,951.861	0.0	320,688.835	186,066	362,933.652	186,066	333,808.491	186,066

Grade

Table "default" - Rows: 1 | Spec - Columns: 8 | Properties | Flow Variables

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	5,635.023	0.0	310,871.774	220,217.5	365,681.533	420,397.5	326,443.906	220,217.5

The p-value associated with the Kruskal-Wallis test is 0.0, 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 0.0 which is 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

Table "default" - Rows: 1 Spec - Columns: 10 Properties Flow Variables

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	28,824.432	0.0	429,174.104	532,198.5	330,164.536	319,691.5	408,794.195	532,198.5	418,492.542	532,198.5

Term

Table "default" - Rows: 1 Spec - Columns: 10 Properties Flow Variables

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	5,181.064	0.0	356,228.969	248,413.5	345,866.285	248,413.5	376,739.101	248,413.5	337,151.375	248,413.5

application_type

Table "default" - Rows: 1 Spec - Columns: 10 Properties Flow Variables

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	39.811	1.1686883127914882...	354,819.78	354,753.5	354,993.861	354,753.5	354,861.624	354,753.5	354,753.5	354,753.5

Purpose

Table "default" - Rows: 1 Spec - Columns: 10 Properties Flow Variables

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	289.933	0.0	368,233.995	426,276.5	352,656.868	426,276.5	359,429.848	426,276.5	374,348.396	426,276.5

loan_condition

Table "default" - Rows: 1 Spec - Columns: 10 Properties Flow Variables

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	1,460.424	0.0	345,766.634	327,981.5	357,791.379	327,981.5	348,837.2	327,981.5	357,560.792	327,981.5

income_category

Table "default" - Rows: 1 Spec - Columns: 10 Properties Flow Variables

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	377,867.49	0.0	692,687.642	703,196	291,839	291,839	480,279.323	640,083	703,196	703,196

interest_payments

Table "default" - Rows: 1 Spec - Columns: 10 Properties Flow Variables

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	4,110.03	0.0	322,116.749	186,066	363,874.781	541,017.5	336,059.974	186,066	274,803.875	186,066

Grade

Table "default" - Rows: 1 | Spec - Columns: 10 | Properties | Flow Variables

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	5,884.219	0.0	311,718.387	220,217.5	366,974.875	420,397.5	329,412.483	220,217.5	281,028.75	220,217.5

The p-value associated with the Kruskal-Wallis test is 0.0, 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 0.0 which is 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

Table "default" - Rows: 1 | Spec - Columns: 12 | Properties | Flow Variables

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	35,785.751	0.0	423,189.935	532,198.5	320,293.902	319,691.5	398,290.254	532,198.5	426,125.662	532,198.5	418,492.542	532,198.5

Term

Table "default" - Rows: 1 | Spec - Columns: 12 | Properties | Flow Variables

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	8,859.376	0.0	347,233.028	248,413.5	339,789.342	248,413.5	378,201.688	248,413.5	366,538.297	248,413.5	337,151.375	248,413.5

application_type

Table "default" - Rows: 1 | Spec - Columns: 12 | Properties | Flow Variables

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	40.661	3.1584963644526454E-8	354,753.5	354,753.5	355,003.793	354,753.5	354,892.099	354,753.5	354,826.552	354,753.5	354,753.5	354,753.5

Purpose

Table "default" - Rows: 1 | Spec - Columns: 12 | Properties | Flow Variables

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	294.958	0.0	378,352.372	426,276.5	352,091.473	426,276.5	358,132.581	426,276.5	361,817.483	426,276.5	374,348.396	426,276.5

loan_condition

Table "default" - Rows: 1 | Spec - Columns: 12 | Properties | Flow Variables

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	1,689.428	0.0	343,442.182	327,981.5	358,770.971	327,981.5	350,262.79	327,981.5	346,894.617	327,981.5	357,560.792	327,981.5

income_category

Table "default" - Rows: 1 | Spec - Columns: 12 | Properties | Flow Variables

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	366,549.828	0.0	703,196	703,196	291,839	291,839	397,872.432	291,839	653,898.264	640,083	703,196	703,196

interest_payments

Table "default" - Rows: 1 | Spec - Columns: 12 | Properties | Flow Variables

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	4,274.819	0.0	321,346.971	186,066	365,935.445	541,017.5	342,986.455	186,066	324,155.86	186,066	274,803.875	186,066

Grade

Table "default" - Rows: 1 | Spec - Columns: 12 | Properties | Flow Variables

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	6,107.452	0.0	312,982.922	220,217.5	369,773.062	420,397.5	338,658.287	420,397.5	313,933.128	220,217.5	281,028.75	220,217.5

The p-value associated with the Kruskal-Wallis test is 0.0, 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 0.0 which is less than 0.05 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	5,858.6281	1	5,858.6281	476.8951	0.0
emp_length_int	Within Groups	8,721,091.28	709901	12.2849		
emp_length_int	Total	8,726,949.9081	709902			
annual_inc	Between Groups	8.96E14	1	8.96E14	336,176.8466	0.0
annual_inc	Within Groups	1.89E15	709901	2.66E9		
annual_inc	Total	2.79E15	709902			
loan_amount	Between Groups	5.59E12	1	5.59E12	88,348.6615	0.0
loan_amount	Within Groups	4.49E13	709901	63,287,453.1294		
loan_amount	Total	5.05E13	709902			
interest_rate	Between Groups	74,878.1439	1	74,878.1439	3,919.3699	0.0
interest_rate	Within Groups	13,562,401.7027	709901	19.1046		
interest_rate	Total	13,637,279.8466	709902			
dti	Between Groups	1,598,860.355	1	1,598,860.355	4,577.329	0.0
dti	Within Groups	2.48E8	709901	349.2999		
dti	Total	2.50E8	709902			
total_pymnt	Between Groups	1.12E12	1	1.12E12	18,633.4442	0.0
total_pymnt	Within Groups	4.29E13	709901	60,368,711.3594		
total_pymnt	Total	4.40E13	709902			
total_rec_prncp	Between Groups	7.11E11	1	7.11E11	16,563.3871	0.0
total_rec_prncp	Within Groups	3.05E13	709901	42,932,703.1582		
total_rec_prncp	Total	3.12E13	709902			
recoveries	Between Groups	2,952,298.5974	1	2,952,298.5974	17.3234	3.15E-5
recoveries	Within Groups	1.21E11	709901	170,422.173		
recoveries	Total	1.21E11	709902			
installment	Between Groups	4.60E9	1	4.60E9	86,406.2924	0.0
installment	Within Groups	3.78E10	709901	53,186.8132		
installment	Total	4.24E10	709902			

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 (Lower Bound)	Confidence Interval of the Difference (Upper Bound)	Minimum	Maximum
Row0	emp_length_int	cluster_1	616266	0	0	6.015	3.491	0.004	0.95	6.006	6.023	0.5	10
Row1	emp_length_int	cluster_0	93637	0	0	6.283	3.597	0.012	0.95	6.26	6.306	0.5	10
Row2	emp_length_int	Total	709903	0	0	6.05	3.506	0.004	0.95	6.042	6.058	0.5	10
Row3	annual_inc	cluster_1	616266	0	0	61,126.224	22,927.127	29.206	0.95	61,068.983	61,183.466	0	115,100
Row4	annual_inc	cluster_0	93637	0	0	166,104.079	129,393.357	422.852	0.95	165,275.294	166,932.864	110,000	9,000,000
Row5	annual_inc	Total	709903	0	0	74,972.921	62,662.099	74.371	0.95	74,827.156	75,118.686	0	9,000,000
Row6	loan_amount	cluster_1	616266	0	0	13,662.141	7,734.436	9.852	0.95	13,642.831	13,681.452	500	35,000
Row7	loan_amount	cluster_0	93637	0	0	21,955.882	9,278.932	30.323	0.95	21,896.449	22,015.315	925	35,000
Row8	loan_amount	Total	709903	0	0	14,756.095	8,435.853	10.012	0.95	14,736.471	14,775.718	500	35,000
Row9	interest_rate	cluster_1	616266	0	0	13.376	4.339	0.006	0.95	13.366	13.387	5.32	28.99
Row10	interest_rate	cluster_0	93637	0	0	12.417	4.574	0.015	0.95	12.387	12.446	5.32	28.99
Row11	interest_rate	Total	709903	0	0	13.25	4.383	0.005	0.95	13.24	13.26	5.32	28.99
Row12	dti	cluster_1	616266	0	0	18.747	19.86	0.025	0.95	18.698	18.797	0	9,999
Row13	dti	cluster_0	93637	0	0	14.312	7.227	0.024	0.95	14.266	14.359	0	39.99
Row14	dti	Total	709903	0	0	18.162	18.75	0.022	0.95	18.119	18.206	0	9,999
Row15	total_pymnt	cluster_1	616266	0	0	7,072.225	7,262.922	9.252	0.95	7,054.091	7,090.358	0	55,906.95
Row16	total_pymnt	cluster_0	93637	0	0	10,792.235	10,512.419	34.354	0.95	10,724.902	10,859.569	0	57,777.58
Row17	total_pymnt	Total	709903	0	0	7,562.898	7,871.034	9.342	0.95	7,544.588	7,581.208	0	57,777.58
Row18	total_rec_prncp	cluster_1	616266	0	0	5,371.881	6,098.569	7.769	0.95	5,356.655	5,387.107	0	35,000.02
Row19	total_rec_prncp	cluster_0	93637	0	0	8,329.624	8,983.961	29.359	0.95	8,272.08	8,387.167	0	35,000.03
Row20	total_rec_prncp	Total	709903	0	0	5,762.011	6,628.299	7.867	0.95	5,746.592	5,777.43	0	35,000.03
Row21	recoveries	cluster_1	616266	0	0	45.319	392.831	0.5	0.95	44.338	46.3	0	29,623.35
Row22	recoveries	cluster_0	93637	0	0	51.345	525.758	1.718	0.95	47.978	54.713	0	33,520.27
Row23	recoveries	Total	709903	0	0	46.114	412.827	0.49	0.95	45.153	47.074	0	33,520.27
Row24	installment	cluster_1	616266	0	0	405.399	220.966	0.281	0.95	404.848	405.951	15.69	1,445.46
Row25	installment	cluster_0	93637	0	0	643.174	286.162	0.935	0.95	641.342	645.007	25.28	1,424.57
Row26	installment	Total	709903	0	0	436.762	244.255	0.29	0.95	436.194	437.33	15.69	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=2 are statistically different.

K=3

	Source	Sum of Squares	df	Mean Square	F	p-value
emp_length_int	Between Groups	34,687.3453	2	17,343.6726	1,416.4636	0.0
emp_length_int	Within Groups	8,692,262.5628	709900	12.2443		
emp_length_int	Total	8,726,949.9081	709902			
annual_inc	Between Groups	1.33E15	2	6.65E14	324,357.0776	0.0
annual_inc	Within Groups	1.46E15	709900	2.05E9		
annual_inc	Total	2.79E15	709902			
loan_amount	Between Groups	8.98E12	2	4.49E12	76,748.1105	0.0
loan_amount	Within Groups	4.15E13	709900	58,512,171.8228		
loan_amount	Total	5.05E13	709902			
interest_rate	Between Groups	105,159.7306	2	52,579.8653	2,758.3591	0.0
interest_rate	Within Groups	13,532,120.1161	709900	19.062		
interest_rate	Total	13,637,279.8466	709902			
dti	Between Groups	1,990,132.1116	2	995,066.0558	2,853.2427	0.0
dti	Within Groups	2.48E8	709900	348.7492		
dti	Total	2.50E8	709902			
total_pymnt	Between Groups	2.09E12	2	1.05E12	17,721.671	0.0
total_pymnt	Within Groups	4.19E13	709900	59,007,283.0844		
total_pymnt	Total	4.40E13	709902			
total_rec_prncp	Between Groups	1.30E12	2	6.50E11	15,435.9406	0.0
total_rec_prncp	Within Groups	2.99E13	709900	42,103,487.8286		
total_rec_prncp	Total	3.12E13	709902			
recoveries	Between Groups	11,148,832.7137	2	5,574,416.3569	32.7116	6.22E-15
recoveries	Within Groups	1.21E11	709900	170,410.867		
recoveries	Total	1.21E11	709902			
installment	Between Groups	7.15E9	2	3.57E9	72,049.0329	0.0
installment	Within Groups	3.52E10	709900	49,593.8444		
installment	Total	4.24E10	709902			

Descriptive Statistics: K=3

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_1	518916	0	0	5.916	3.477	0.005	0.95	5.907	5.925	0.5	10
Row1	emp_length_int	cluster_2	183085	0	0	6.418	3.561	0.008	0.95	6.401	6.434	0.5	10
Row2	emp_length_int	cluster_0	7902	0	0	6.329	3.523	0.04	0.95	6.251	6.406	0.5	10
Row3	emp_length_int	Total	709903	0	0	6.05	3.506	0.004	0.95	6.042	6.058	0.5	10
Row4	annual_inc	cluster_1	518916	0	0	54,127.279	17,441.526	24.212	0.95	54,079.824	54,174.734	0	90,000
Row5	annual_inc	cluster_2	183085	0	0	121,057.908	32,142.488	75.12	0.95	120,910.676	121,205.141	82,000	248,400
Row6	annual_inc	cluster_0	7902	0	0	376,120.374	374,731.44	4,215.525	0.95	367,856.832	384,383.917	248,400	9,000,000
Row7	annual_inc	Total	709903	0	0	74,972.921	62,662.099	74.371	0.95	74,827.156	75,118.686	0	9,000,000
Row8	loan_amount	cluster_1	518916	0	0	12,621.391	6,985.575	9.697	0.95	12,602.385	12,640.398	500	35,000
Row9	loan_amount	cluster_2	183085	0	0	20,347.923	9,226.471	21.563	0.95	20,305.66	20,390.186	925	35,000
Row10	loan_amount	cluster_0	7902	0	0	25,380.239	8,930.306	100.461	0.95	25,183.309	25,577.169	1,000	35,000
Row11	loan_amount	Total	709903	0	0	14,756.095	8,435.853	10.012	0.95	14,736.471	14,775.718	500	35,000
Row12	interest_rate	cluster_1	518916	0	0	13.482	4.291	0.006	0.95	13.47	13.494	5.32	28.99
Row13	interest_rate	cluster_2	183085	0	0	12.637	4.563	0.011	0.95	12.616	12.657	5.32	28.99
Row14	interest_rate	cluster_0	7902	0	0	12.222	4.583	0.052	0.95	12.12	12.323	5.32	28.99
Row15	interest_rate	Total	709903	0	0	13.25	4.383	0.005	0.95	13.24	13.26	5.32	28.99
Row16	dti	cluster_1	518916	0	0	19.118	21.367	0.03	0.95	19.059	19.176	0	9,999
Row17	dti	cluster_2	183085	0	0	15.796	7.523	0.018	0.95	15.761	15.83	0	42.64
Row18	dti	cluster_0	7902	0	0	10.279	6.139	0.069	0.95	10.143	10.414	0	38.18
Row19	dti	Total	709903	0	0	18.162	18.75	0.022	0.95	18.119	18.206	0	9,999
Row20	total_pymnt	cluster_1	518916	0	0	6,527.215	6,554.317	9.099	0.95	6,509.382	6,545.048	0	55,906.95
Row21	total_pymnt	cluster_2	183085	0	0	10,305.593	10,064.863	23.522	0.95	10,259.49	10,351.697	0	57,777.58
Row22	total_pymnt	cluster_0	7902	0	0	12,028.392	11,530.653	129.714	0.95	11,774.119	12,282.664	0	54,091.939
Row23	total_pymnt	Total	709903	0	0	7,562.898	7,871.034	9.342	0.95	7,544.588	7,581.208	0	57,777.58
Row24	total_rec_prncp	cluster_1	518916	0	0	4,946.227	5,498.962	7.634	0.95	4,931.266	4,961.189	0	35,000.01
Row25	total_rec_prncp	cluster_2	183085	0	0	7,918.048	8,561.016	20.008	0.95	7,878.833	7,957.263	0	35,000.03
Row26	total_rec_prncp	cluster_0	7902	0	0	9,379.444	9,933.337	111.745	0.95	9,160.395	9,598.493	0	35,000.01
Row27	total_rec_prncp	Total	709903	0	0	5,762.011	6,628.299	7.867	0.95	5,746.592	5,777.43	0	35,000.03
Row28	recoveries	cluster_1	518916	0	0	43.717	366.7	0.509	0.95	42.72	44.715	0	25,000.29
Row29	recoveries	cluster_2	183085	0	0	52.753	513.099	1.199	0.95	50.403	55.103	0	33,520.27
Row30	recoveries	cluster_0	7902	0	0	49.644	615.796	6.927	0.95	36.065	63.224	0	27,750
Row31	recoveries	Total	709903	0	0	46.114	412.827	0.49	0.95	45.153	47.074	0	33,520.27
Row32	installment	cluster_1	518916	0	0	376.817	198.288	0.275	0.95	376.277	377.356	15.69	1,445.46
Row33	installment	cluster_2	183085	0	0	592.66	277.727	0.649	0.95	591.388	593.932	25.28	1,445.46
Row34	installment	cluster_0	7902	0	0	761.243	293.83	3.305	0.95	754.763	767.722	30.65	1,409.99
Row35	installment	Total	709903	0	0	436.762	244.255	0.29	0.95	436.194	437.33	15.69	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	48,129.9859	3	16,043.3286	1,312.2917	0.0
emp_length_int	Within Groups	8,678,819.9222	709899	12.2254		
emp_length_int	Total	8,726,949.9081	709902			
annual_inc	Between Groups	2.11E15	3	7.03E14	736,261.4902	0.0
annual_inc	Within Groups	6.78E14	709899	9.55E8		
annual_inc	Total	2.79E15	709902			
loan_amount	Between Groups	9.90E12	3	3.30E12	57,659.4403	0.0
loan_amount	Within Groups	4.06E13	709899	57,221,073.4734		
loan_amount	Total	5.05E13	709902			
interest_rate	Between Groups	107,647.8367	3	35,882.6122	1,882.7586	0.0
interest_rate	Within Groups	13,529,632.0099	709899	19.0585		
interest_rate	Total	13,637,279.8466	709902			
dti	Between Groups	2,074,914.8576	3	691,638.2859	1,983.8734	0.0
dti	Within Groups	2.47E8	709899	348.6302		
dti	Total	2.50E8	709902			
total_pymnt	Between Groups	2.38E12	3	7.92E11	13,520.7074	0.0
total_pymnt	Within Groups	4.16E13	709899	58,604,882.1001		
total_pymnt	Total	4.40E13	709902			
total_rec_prncp	Between Groups	1.47E12	3	4.89E11	11,676.6158	0.0
total_rec_prncp	Within Groups	2.97E13	709899	41,868,535.4906		
total_rec_prncp	Total	3.12E13	709902			
recoveries	Between Groups	20,709,728.5978	3	6,903,242.8659	40.5125	0.0
recoveries	Within Groups	1.21E11	709899	170,397.6391		
recoveries	Total	1.21E11	709902			
installment	Between Groups	7.84E9	3	2.61E9	53,719.9213	0.0
installment	Within Groups	3.45E10	709899	48,622.4913		
installment	Total	4.24E10	709902			

Descriptive Statistics: K=4

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_1	490278	0	0	5.877	3.472	0.005	0.95	5.868	5.887	0.5	10
Row1	emp_length_int	cluster_2	203535	0	0	6.452	3.55	0.008	0.95	6.437	6.468	0.5	10
Row2	emp_length_int	cluster_0	16066	0	0	6.215	3.56	0.028	0.95	6.16	6.27	0.5	10
Row3	emp_length_int	cluster_3	24	0	0	8.667	2.808	0.573	0.95	7.481	9.852	1	10
Row4	emp_length_int	Total	709903	0	0	6.05	3.506	0.004	0.95	6.042	6.058	0.5	10
Row5	annual_inc	cluster_1	490278	0	0	52,357.314	16,277.832	23.247	0.95	52,311.75	52,402.879	0	85,000
Row6	annual_inc	cluster_2	203535	0	0	112,130.848	26,013.249	57.66	0.95	112,017.836	112,243.86	76,000	199,000
Row7	annual_inc	cluster_0	16066	0	0	285,405.345	145,572.91	1,148.489	0.95	283,154.179	287,656.511	198,000	3,120,000
Row8	annual_inc	cluster_3	24	0	0	6,082,819.125	1,743,355.326	355,860.916	0.95	5,346,664.733	6,818,973.517	3,900,000	9,000,000
Row9	annual_inc	Total	709903	0	0	74,972.921	62,662.099	74.371	0.95	74,827.156	75,118.686	0	9,000,000
Row10	loan_amount	cluster_1	490278	0	0	12,298.585	6,735.712	9.62	0.95	12,279.73	12,317.439	500	35,000
Row11	loan_amount	cluster_2	203535	0	0	19,899.91	9,158.157	20.3	0.95	19,860.123	19,939.697	1,000	35,000
Row12	loan_amount	cluster_0	16066	0	0	24,583.809	9,011.333	71.094	0.95	24,444.456	24,723.162	925	35,000
Row13	loan_amount	cluster_3	24	0	0	15,778.125	9,527.279	1,944.748	0.95	11,755.108	19,801.142	1,225	35,000
Row14	loan_amount	Total	709903	0	0	14,756.095	8,435.853	10.012	0.95	14,736.471	14,775.718	500	35,000
Row15	interest_rate	cluster_1	490278	0	0	13.507	4.276	0.006	0.95	13.495	13.519	5.32	28.99
Row16	interest_rate	cluster_2	203535	0	0	12.709	4.557	0.01	0.95	12.69	12.729	5.32	28.99
Row17	interest_rate	cluster_0	16066	0	0	12.26	4.603	0.036	0.95	12.189	12.331	5.32	28.99
Row18	interest_rate	cluster_3	24	0	0	11.182	3.979	0.812	0.95	9.502	12.862	6.24	19.99
Row19	interest_rate	Total	709903	0	0	13.25	4.383	0.005	0.95	13.24	13.26	5.32	28.99
Row20	dti	cluster_1	490278	0	0	19.208	21.898	0.031	0.95	19.147	19.27	0	9,999
Row21	dti	cluster_2	203535	0	0	16.177	7.596	0.017	0.95	16.144	16.21	0	42.64
Row22	dti	cluster_0	16066	0	0	11.427	6.383	0.05	0.95	11.328	11.526	0	39.84
Row23	dti	cluster_3	24	0	0	0.289	0.457	0.093	0.95	0.096	0.482	0.03	2.36
Row24	dti	Total	709903	0	0	18.162	18.75	0.022	0.95	18.119	18.206	0	9,999
Row25	total_pymnt	cluster_1	490278	0	0	6,348.416	6,329.042	9.039	0.95	6,330.7	6,366.132	0	55,362.539
Row26	total_pymnt	cluster_2	203535	0	0	10,155.025	9,897.538	21.939	0.95	10,112.026	10,198.024	0	57,777.58
Row27	total_pymnt	cluster_0	16066	0	0	11,785.139	11,222.157	88.537	0.95	11,611.598	11,958.681	0	54,329.49
Row28	total_pymnt	cluster_3	24	0	0	8,012.429	11,434.459	2,334.049	0.95	3,184.08	12,840.777	0	43,114.63
Row29	total_pymnt	Total	709903	0	0	7,562.898	7,871.034	9.342	0.95	7,544.588	7,581.208	0	57,777.58
Row30	total_rec_prncp	cluster_1	490278	0	0	4,809.318	5,311.086	7.585	0.95	4,794.452	4,824.185	0	35,000.01
Row31	total_rec_prncp	cluster_2	203535	0	0	7,788.122	8,409.777	18.641	0.95	7,751.586	7,824.657	0	35,000.03
Row32	total_rec_prncp	cluster_0	16066	0	0	9,165.549	9,649.629	76.13	0.95	9,016.325	9,314.773	0	35,000.01
Row33	total_rec_prncp	cluster_3	24	0	0	6,531.478	9,703.719	1,980.763	0.95	2,433.957	10,628.999	0	35,000
Row34	total_rec_prncp	Total	709903	0	0	5,762.011	6,628.299	7.867	0.95	5,746.592	5,777.43	0	35,000.03
Row35	recoveries	cluster_1	490278	0	0	42.526	352.092	0.503	0.95	41.54	43.511	0	25,000.29
Row36	recoveries	cluster_2	203535	0	0	54.423	521.435	1.156	0.95	52.157	56.688	0	33,520.27
Row37	recoveries	cluster_0	16066	0	0	50.407	549.235	4.333	0.95	41.913	58.9	0	27,750
Row38	recoveries	cluster_3	24	0	0	0	0	0	0.95	0	0	0	0
Row39	recoveries	Total	709903	0	0	46.114	412.827	0.49	0.95	45.153	47.074	0	33,520.27
Row40	installment	cluster_1	490278	0	0	368.071	190.961	0.273	0.95	367.536	368.606	15.69	1,445.46
Row41	installment	cluster_2	203535	0	0	578.776	273.901	0.607	0.95	577.586	579.966	25.28	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	90,542.1393	4	22,635.5348	1,860.6024	0.0
emp_length_int	Within Groups	8,636,407.7688	709898	12.1657		
emp_length_int	Total	8,726,949.9081	709902			
annual_inc	Between Groups	2.32E15	4	5.80E14	878,529.6267	0.0
annual_inc	Within Groups	4.68E14	709898	6.60E8		
annual_inc	Total	2.79E15	709902			
loan_amount	Between Groups	1.18E13	4	2.95E12	54,180.5714	0.0
loan_amount	Within Groups	3.87E13	709898	54,519,838.2027		
loan_amount	Total	5.05E13	709902			
interest_rate	Between Groups	107,612.2048	4	26,903.0512	1,411.5958	0.0
interest_rate	Within Groups	13,529,667.6418	709898	19.0586		
interest_rate	Total	13,637,279.8466	709902			
dti	Between Groups	2,259,023.8605	4	564,755.9651	1,621.1319	0.0
dti	Within Groups	2.47E8	709898	348.3714		
dti	Total	2.50E8	709902			
total_pymnt	Between Groups	2.88E12	4	7.19E11	12,415.6521	0.0
total_pymnt	Within Groups	4.11E13	709898	57,902,799.462		
total_pymnt	Total	4.40E13	709902			
total_rec_prncp	Between Groups	1.74E12	4	4.35E11	10,477.9174	0.0
total_rec_prncp	Within Groups	2.95E13	709898	41,485,340.166		
total_rec_prncp	Total	3.12E13	709902			
recoveries	Between Groups	36,541,651.0454	4	9,135,412.7614	53.6193	0.0
recoveries	Within Groups	1.21E11	709898	170,375.5774		
recoveries	Total	1.21E11	709902			
installment	Between Groups	9.25E9	4	2.31E9	49,561.723	0.0
installment	Within Groups	3.31E10	709898	46,636.8831		
installment	Total	4.24E10	709902			

Descriptive Statistics: K=5

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_1	412680	0	0	5.754	3.457	0.005	0.95	5.743	5.764	0.5	10
Row1	emp_length_int	cluster_2	245855	0	0	6.511	3.517	0.007	0.95	6.497	6.525	0.5	10
Row2	emp_length_int	cluster_3	48589	0	0	6.199	3.604	0.016	0.95	6.166	6.231	0.5	10
Row3	emp_length_int	cluster_0	2755	0	0	6.59	3.442	0.066	0.95	6.461	6.719	0.5	10
Row4	emp_length_int	cluster_4	24	0	0	8.667	2.808	0.573	0.95	7.481	9.852	1	10
Row5	emp_length_int	Total	709903	0	0	6.05	3.506	0.004	0.95	6.042	6.058	0.5	10
Row6	annual_inc	cluster_1	412680	0	0	47,888.203	13,642.558	21.237	0.95	47,846.579	47,929.826	0	74,000
Row7	annual_inc	cluster_2	245855	0	0	94,102.302	17,418.082	35.129	0.95	94,033.451	94,171.154	64,480	138,280
Row8	annual_inc	cluster_3	48589	0	0	180,688.495	42,973.968	194.956	0.95	180,306.379	181,070.611	136,000	344,000
Row9	annual_inc	cluster_0	2755	0	0	508,172.579	239,094.3	4,555.21	0.95	499,240.605	517,104.552	345,000	3,120,000
Row10	annual_inc	cluster_4	24	0	0	6,082,819.125	1,743,355.326	355,860.916	0.95	5,346,664.733	6,818,973.517	3,900,000	9,000,000
Row11	annual_inc	Total	709903	0	0	74,972.921	62,662.099	74.371	0.95	74,827.156	75,118.686	0	9,000,000
Row12	loan_amount	cluster_1	412680	0	0	11,414.835	6,061.244	9.435	0.95	11,396.342	11,433.327	500	35,000
Row13	loan_amount	cluster_2	245855	0	0	18,652.806	8,831.577	17.811	0.95	18,617.896	18,687.716	1,000	35,000
Row14	loan_amount	cluster_3	48589	0	0	22,782.005	9,231.36	41.879	0.95	22,699.921	22,864.088	925	35,000
Row15	loan_amount	cluster_0	2755	0	0	25,953.857	9,013.98	171.734	0.95	25,617.117	26,290.597	1,500	35,000
Row16	loan_amount	cluster_4	24	0	0	15,778.125	9,527.279	1,944.748	0.95	11,755.108	19,801.142	1,225	35,000
Row17	loan_amount	Total	709903	0	0	14,756.095	8,435.853	10.012	0.95	14,736.471	14,775.718	500	35,000
Row18	interest_rate	cluster_1	412680	0	0	13.556	4.226	0.007	0.95	13.543	13.568	5.32	28.99
Row19	interest_rate	cluster_2	245855	0	0	12.93	4.547	0.009	0.95	12.912	12.948	5.32	28.99
Row20	interest_rate	cluster_3	48589	0	0	12.331	4.58	0.021	0.95	12.291	12.372	5.32	28.99
Row21	interest_rate	cluster_0	2755	0	0	12.256	4.574	0.087	0.95	12.085	12.427	5.32	28.99
Row22	interest_rate	cluster_4	24	0	0	11.182	3.979	0.812	0.95	9.502	12.862	6.24	19.99
Row23	interest_rate	Total	709903	0	0	13.25	4.383	0.005	0.95	13.24	13.26	5.32	28.99
Row24	dti	cluster_1	412680	0	0	19.434	23.613	0.037	0.95	19.362	19.506	0	9,999
Row25	dti	cluster_2	245855	0	0	17.043	7.763	0.016	0.95	17.013	17.074	0	42.64
Row26	dti	cluster_3	48589	0	0	13.584	6.898	0.031	0.95	13.523	13.646	0	39.97
Row27	dti	cluster_0	2755	0	0	8.483	5.653	0.108	0.95	8.272	8.694	0	37.51
Row28	dti	cluster_4	24	0	0	0.289	0.457	0.093	0.95	0.096	0.482	0.03	2.36
Row29	dti	Total	709903	0	0	18.162	18.75	0.022	0.95	18.119	18.206	0	9,999
Row30	total_pymnt	cluster_1	412680	0	0	5,887.667	5,770.329	8.982	0.95	5,870.062	5,905.273	0	47,589.603
Row31	total_pymnt	cluster_2	245855	0	0	9,616.216	9,318.769	18.794	0.95	9,579.38	9,653.052	0	56,809.052
Row32	total_pymnt	cluster_3	48589	0	0	11,133.809	10,755.381	48.793	0.95	11,038.174	11,229.444	0	57,777.58
Row33	total_pymnt	cluster_0	2755	0	0	12,280.866	11,912.677	226.96	0.95	11,835.838	12,725.895	0	53,320.407
Row34	total_pymnt	cluster_4	24	0	0	8,012.429	11,434.459	2,334.049	0.95	3,184.08	12,840.777	0	43,114.63
Row35	total_pymnt	Total	709903	0	0	7,562.898	7,871.034	9.342	0.95	7,544.588	7,581.208	0	57,777.58
Row36	total_rec_prncp	cluster_1	412680	0	0	4,464.267	4,857.526	7.562	0.95	4,449.447	4,479.088	0	35,000
Row37	total_rec_prncp	cluster_2	245855	0	0	7,332.612	7,889.842	15.912	0.95	7,301.425	7,363.8	0	35,000.03
Row38	total_rec_prncp	cluster_3	48589	0	0	8,618.764	9,201.951	41.746	0.95	8,536.942	8,700.586	0	35,000.02
Row39	total_rec_prncp	cluster_0	2755	0	0	9,604.963	10,301.433	196.262	0.95	9,220.127	9,989.8	0	35,000.01
Row40	total_rec_prncp	cluster_4	24	0	0	6,531.478	9,703.719	1,980.763	0.95	2,433.957	10,628.999	0	35,000
Row41	total_rec_prncp	Total	709903	0	0	5,762.011	6,628.299	7.867	0.95	5,746.592	5,777.43	0	35,000.03
Row42	recoveries	cluster_1	412680	0	0	40.061	326.941	0.509	0.95	39.064	41.059	0	24,862.1
Row43	recoveries	cluster_2	245855	0	0	54.958	497.114	1.003	0.95	52.993	56.923	0	29,623.35
Row44	recoveries	cluster_3	48589	0	0	52.753	547.142	2.482	0.95	47.888	57.618	0	33,520.27
Row45	recoveries	cluster_0	2755	0	0	46.824	746.926	14.23	0.95	18.921	74.727	0	27,750
Row46	recoveries	cluster_4	24	0	0	0	0	0	0.95	0	0	0	0
Row47	recoveries	Total	709903	0	0	46.114	412.827	0.49	0.95	45.153	47.074	0	33,520.27
Row48	installment	cluster_1	412680	0	0	344.217	171.85	0.268	0.95	343.692	344.741	15.69	1,302.02
Row49	installment	cluster_2	245855	0	0	542.009	260.047	0.524	0.95	540.981	543.037	21.99	1,445.46
Row50	installment	cluster_3	48589	0	0	670.528	288.559	1.309	0.95	667.963	673.094	25.28	1,409.99
Row51	installment	cluster_0	2755	0	0	784.234	299.188	5.7	0.95	773.057	795.411	45.75	1,406.45
Row52	installment	cluster_4	24	0	0	457.809	294.129	60.039	0.95	333.609	582.009	40.69	1,218.61
Row53	installment	Total	709903	0	0	436.762	244.255	0.29	0.95	436.194	437.33	15.69	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	0.675	
2	Cluster 1	0.329	0.628
	Cluster 0	0.362	
	Cluster 1	0.608	
3	Cluster 2	0.333	0.54
	Cluster 0	0.356	
	Cluster 1	0.561	
	Cluster 2	0.357	
4	Cluster 3	0.312	0.478
	Cluster 0	0.334	
	Cluster 1	0.524	
	Cluster 2	0.365	
	Cluster 3	0.324	
5	Cluster 4	0.328	0.429

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

Identity of cluster 2: Debt-Ridden customers who are struggling for solvency and even to those that want to start a small business

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.