

**Academic Year 2023-2025**

**Customer Segmentation (Loan Data)**

**on the basis of Risk**

**Machine Learning for Managers**

**Submitted to:**

**Prof. Amarnath Mitra**

**Submitted by:**

**321066 - Akriti Agarwal**

***Contents of the Report***

|  |  |
| --- | --- |
|  | Objective of the Project |
| 2. | Description of Data |
| 3. | Analysis of Data |
| 4. | Observations |
| 5. | Managerial Insights |

1. **Objectives of the Project**

1.1. PO1 | Segmentation of Consumer Data using Unsupervised Machine Learning Clustering Algorithms

1.2. PO2 | Identification of Appropriate Number of Segments or Clusters

1.3. PO3 | Determination of Segment or Cluster Characteristics

1. **Description of Data**
   1. **Dimension of data**

2.1.1 Number of variables-The number of variables in the csv is 41.

* + 1. Number of observations- There are 212999 observations.
  1. **Description of Variables**

2.2.1. Index Variable(s): 2- Id, Member Id

2.2.2. Categorical Variables -30

2.2.2.1. Nominal Variables

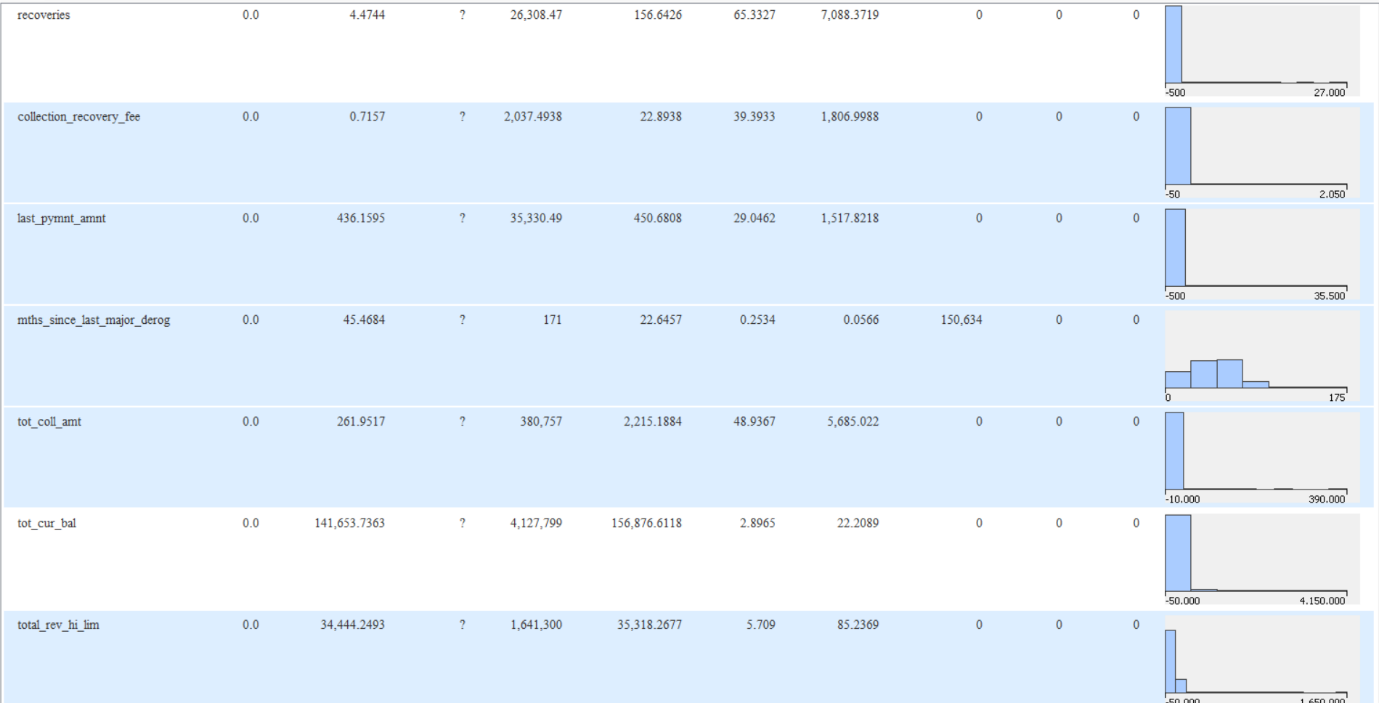
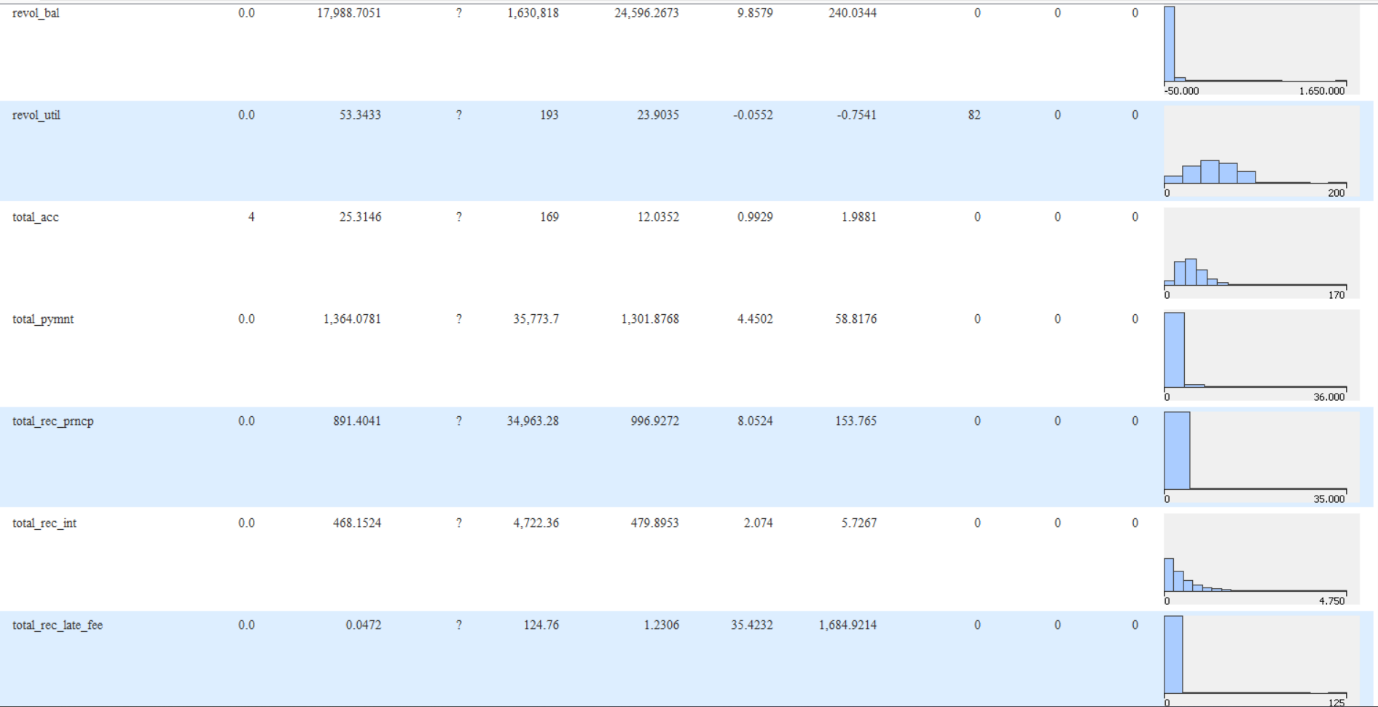
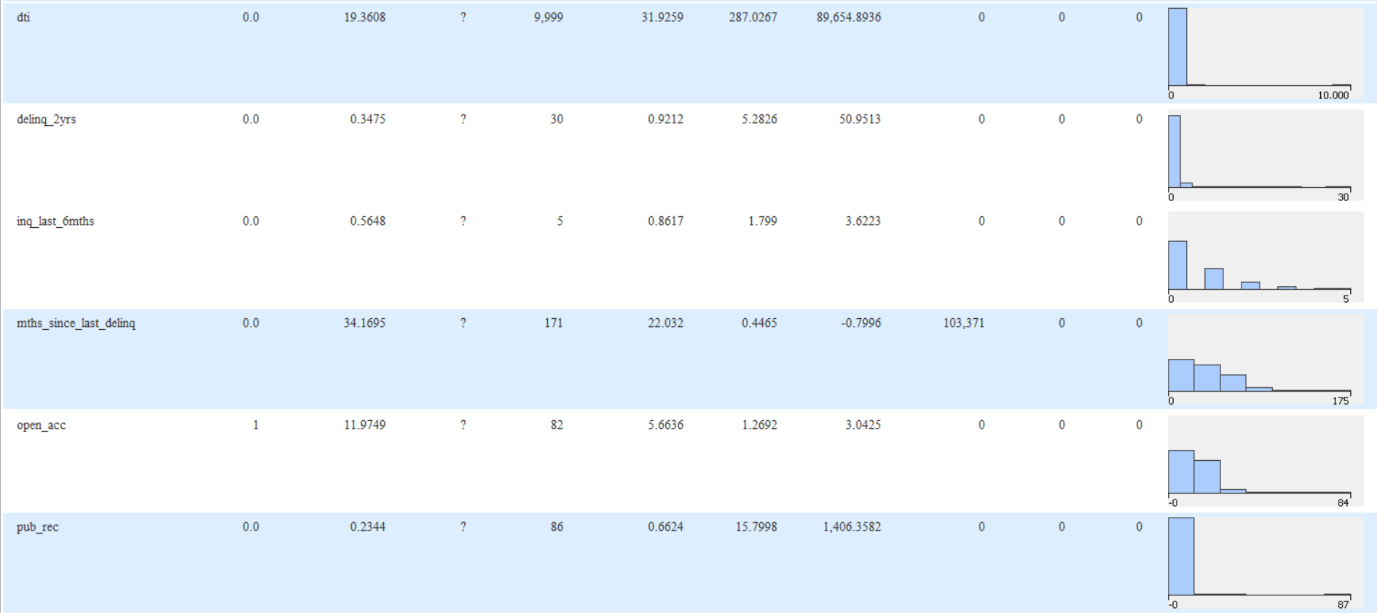
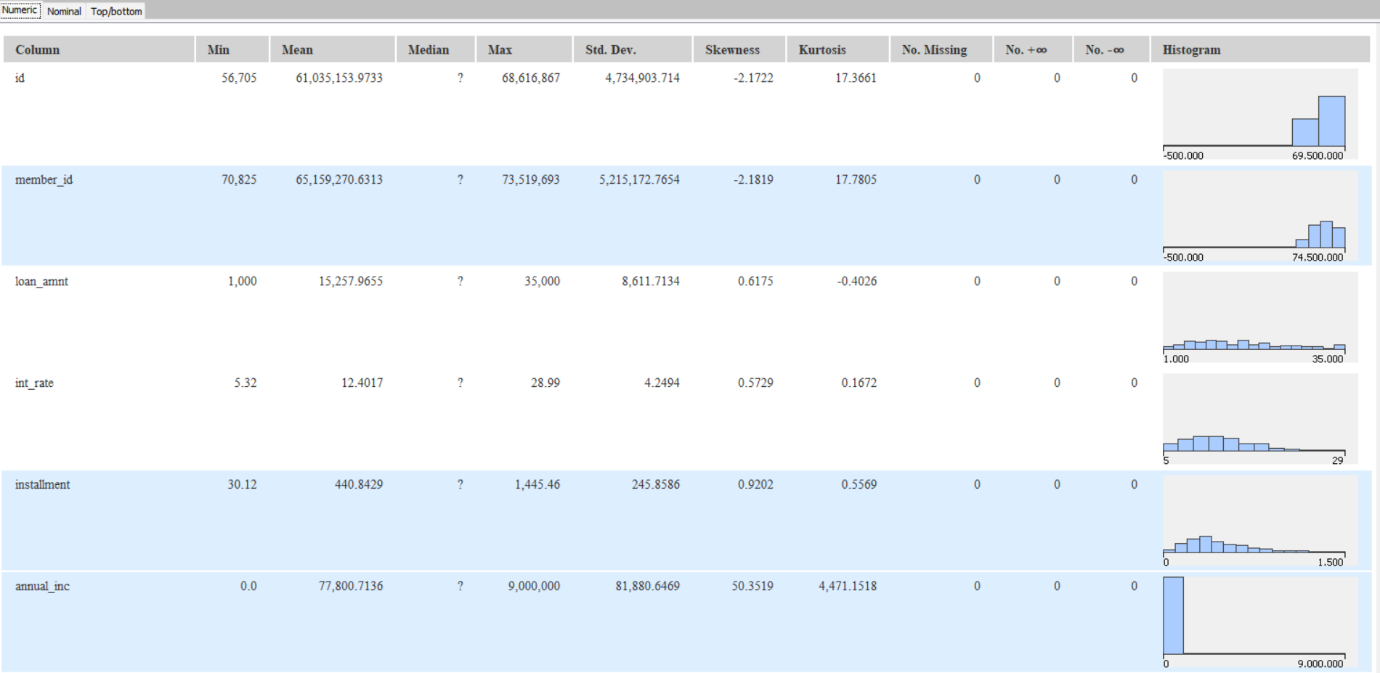
* Loan amount
* term
* interest rate
* instalment
* employment length
* annual income
* issue date
* dti
* earliest cr line
* months since last delinq
* open account
* revol bal
* revol util
* total acc
* total payment
* total recovery principal
* total recovery interest
* late fees
* recoveries
* collection
* last payment date
* last payment amount
* last credit pull date
* months since last major derog
* total collected amt
* total current balance
* total rev hi lim.

2.2.2.2. Ordinal Variables

* Delinq 2yrs
* inq last 6 months
* public recovery

2.2.3. Non-Categorical Variables-10

* Employee title
* Grade
* sub grade
* home ownership
* verification
* loan status
* purpose
* title
* initial list status
* application type.
  1. **Descriptive Statistics**

****

* 1. **Source of Data**- [https://www.kaggle.com/datasets/rounak02/financial-data](https://www.kaggle.com/datasets/rounak02/financial-data%20)

1. **Analysis**
   1. **Data Preprocessing**
      1. Missing Data Treatment
         1. *Stats: Categorical missing data variables*-

employment length, months since last delinq, mths since last major derog, last payment date

*Non-categorical missing data variables*-

employee title

* + - 1. *Proposed Treatment: Categorical Data*- Mode

*Non-Categorical*- Mean

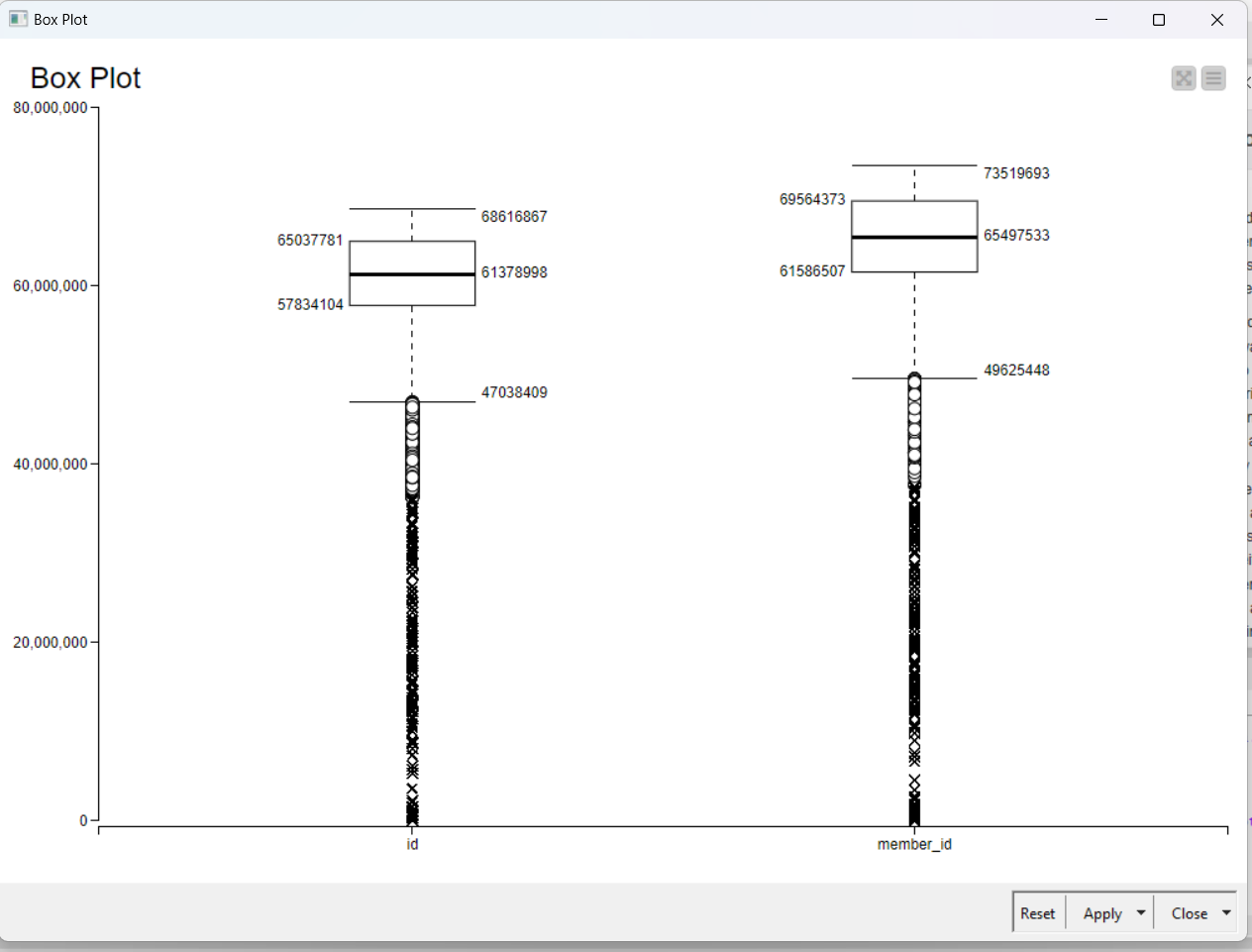
3.1.1.3 *Variables excluded due to more than 50% missing data*- None

* + 1. Categorical Data- Numerical Encoding

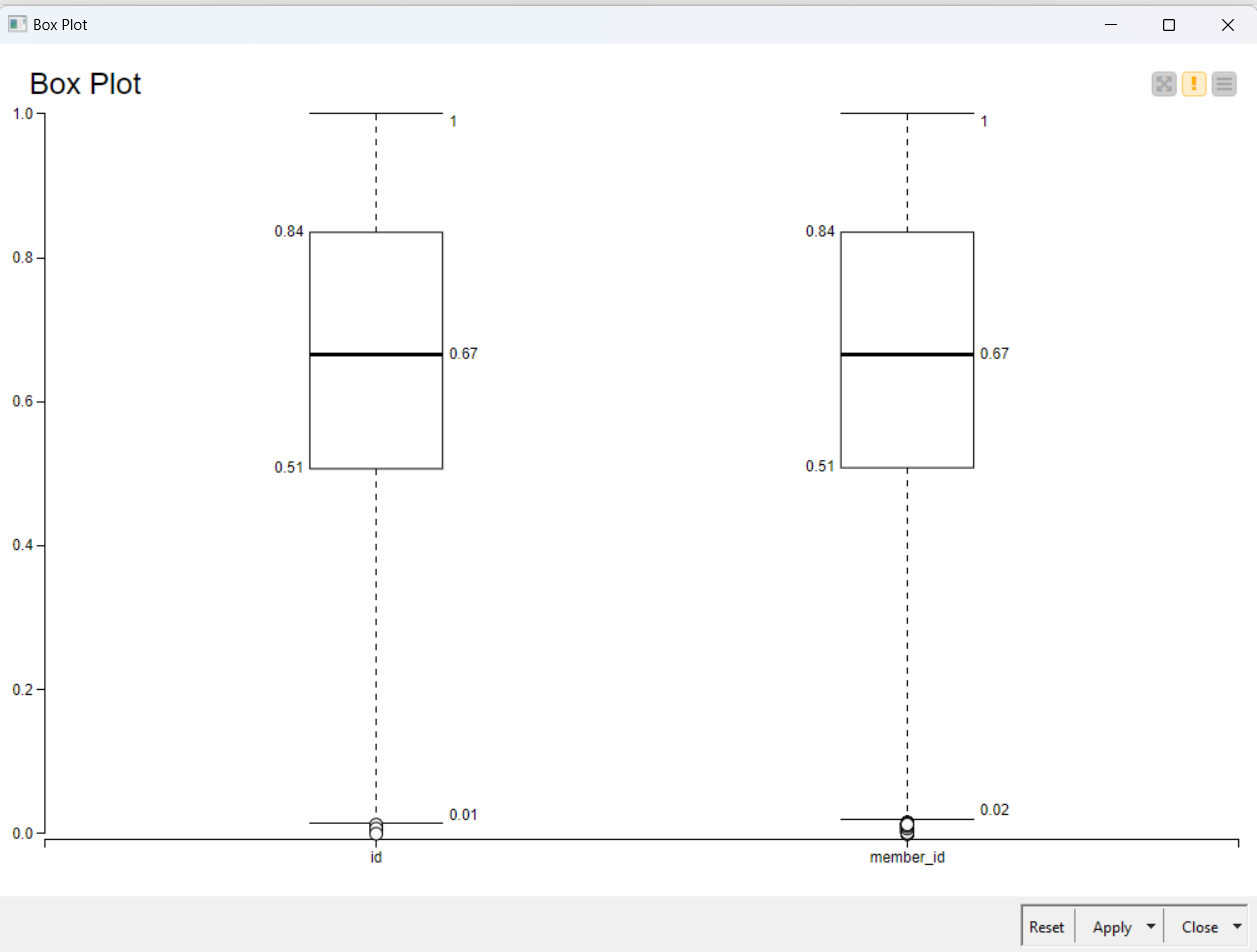
Encoding Schema- Alphanumeric

* + 1. Non-Categorical Data- Outlier Treatment

Using Box Plot- *Variables that have outliers*: id, member id



Normalisation for outliers using min/max scaler



* 1. **Data Analysis**

3.2.1. PO1 | Unsupervised Machine Learning Clustering Algorithm: K-Means

3.2.2. PO2 | Determining the appropriate number of segments/cluster

Silhoutte Score for K=3,4,5

K=3 SS= 0.581

K=4 SS= 0.541

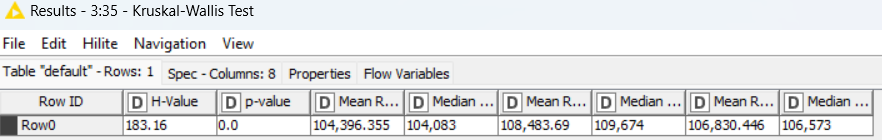
K=5 SS= 0.581

3.2.3. PO3 | Cluster Analysis: Base Model (K-Means)

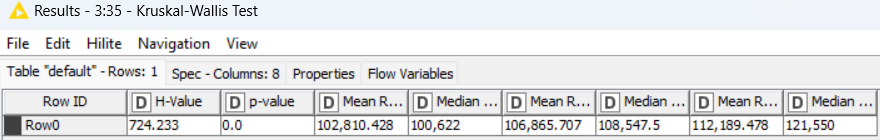
3.2.3.1 Cluster Analysis with Categorical Variables:

Kruskal wallis test for K= 3

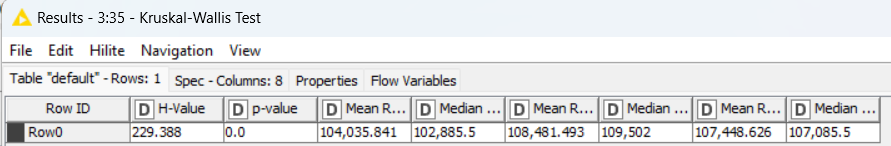
Loan Amnt



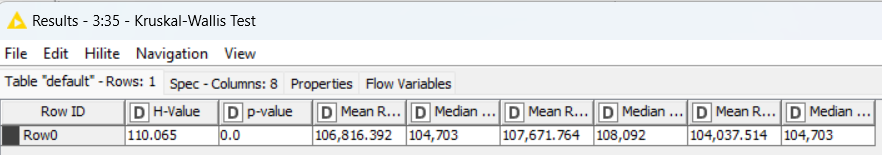
Interest Rate



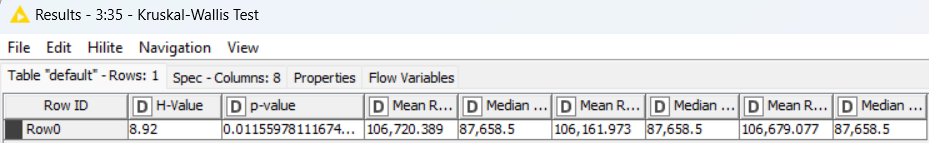
Installment



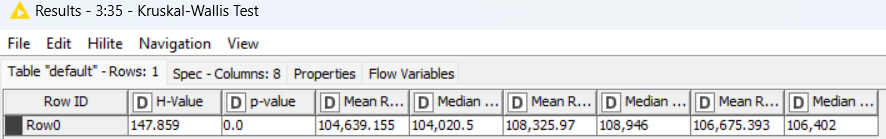
Annual Income



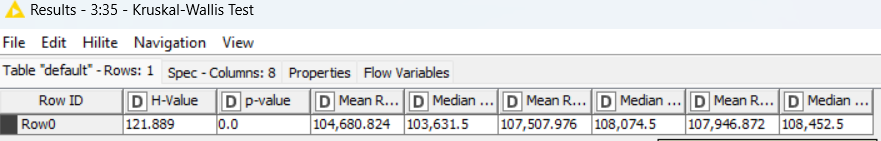
Pub\_rec



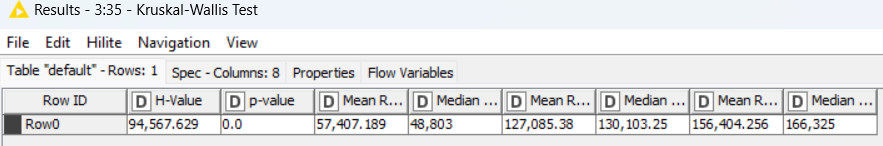
Revol\_bal



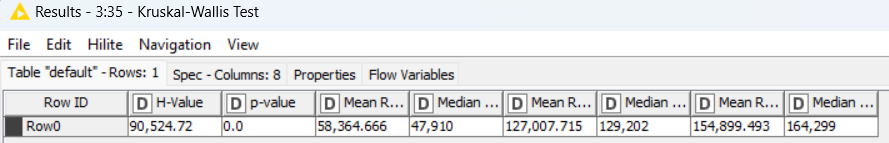
Revol\_util



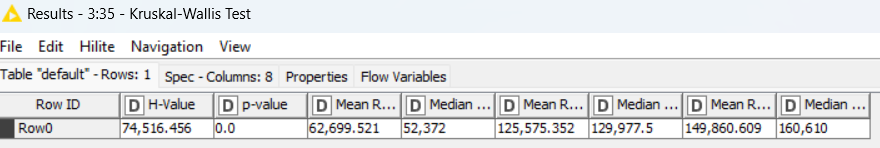
Total-pymnt



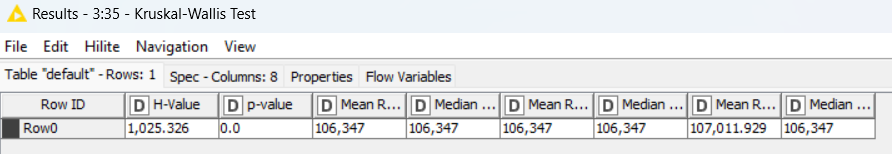
Total\_rec\_prncpl



Total\_rec\_int



Recoveries



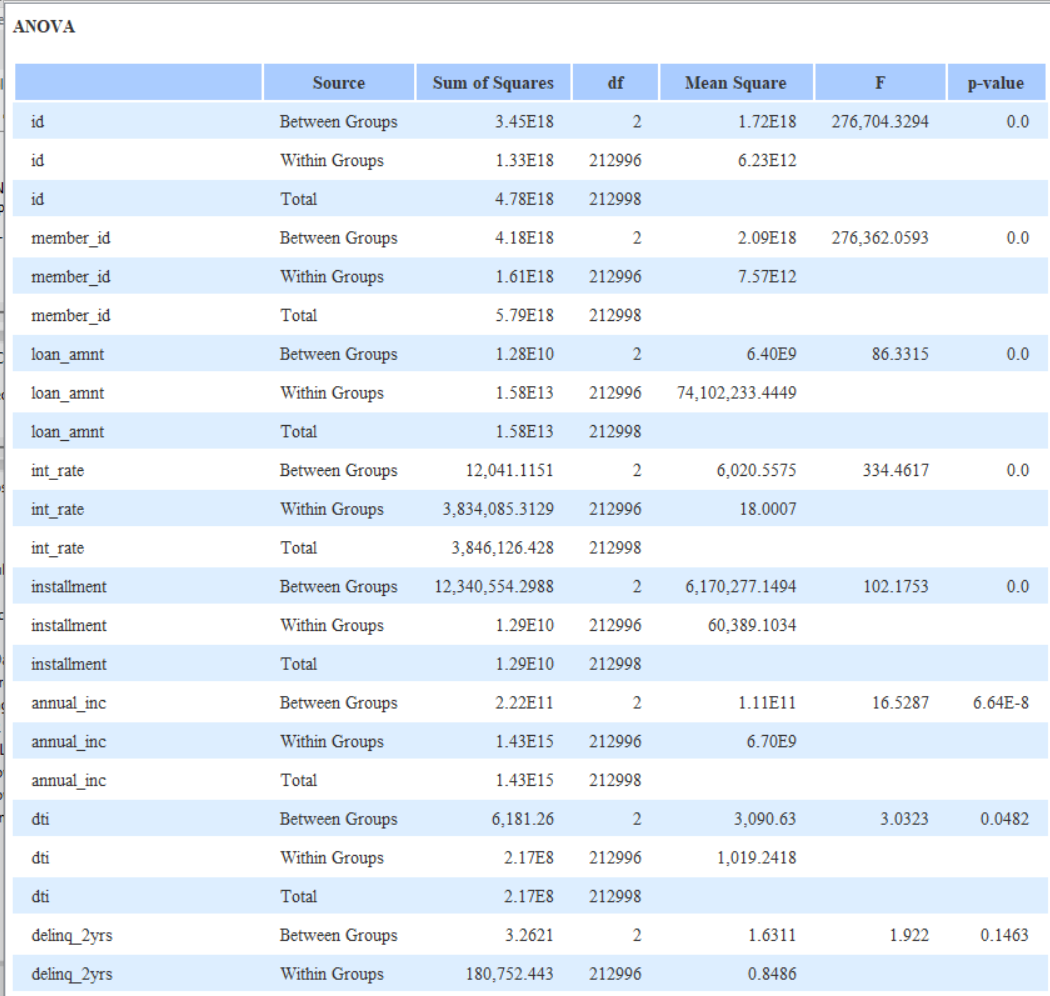
The p-value associated with the Kruskal-Wallis test is 0.0, which is less than the significance level of 0.05. Therefore, we reject the null hypothesis and conclude that there are statistically significant differences between the medians of cluster 0, cluster 1 and cluster 2.

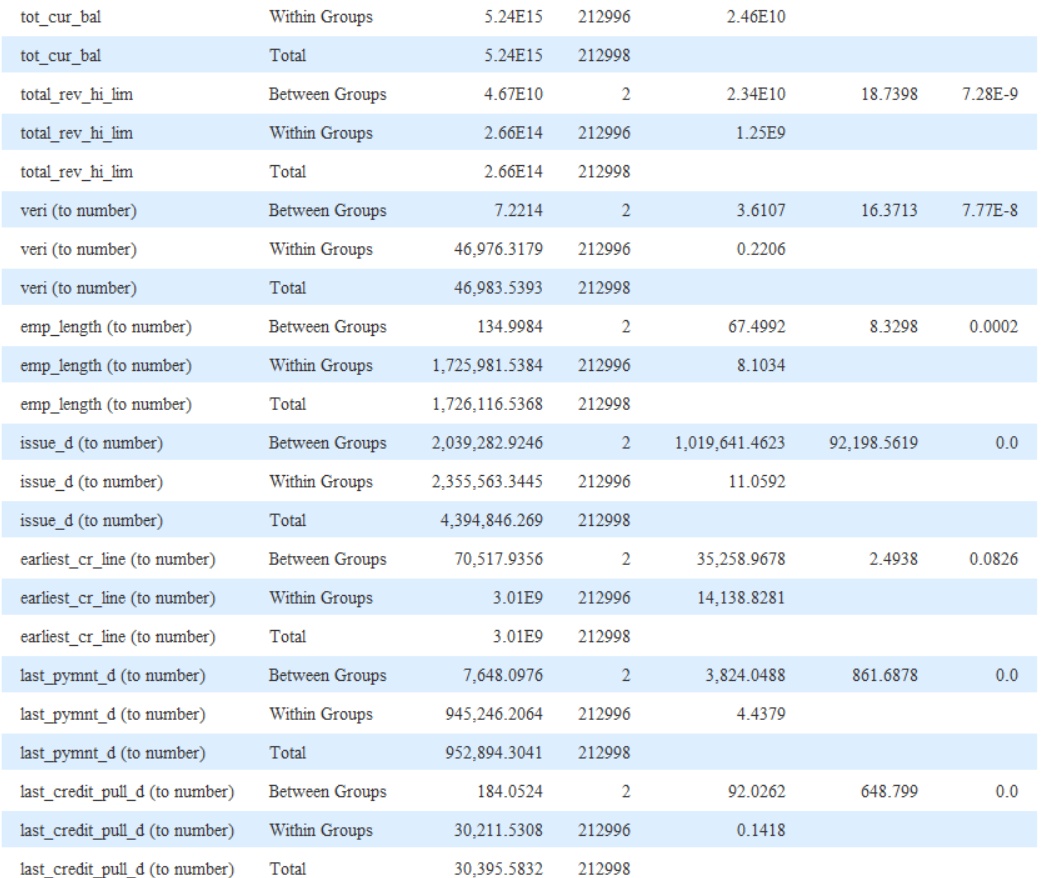
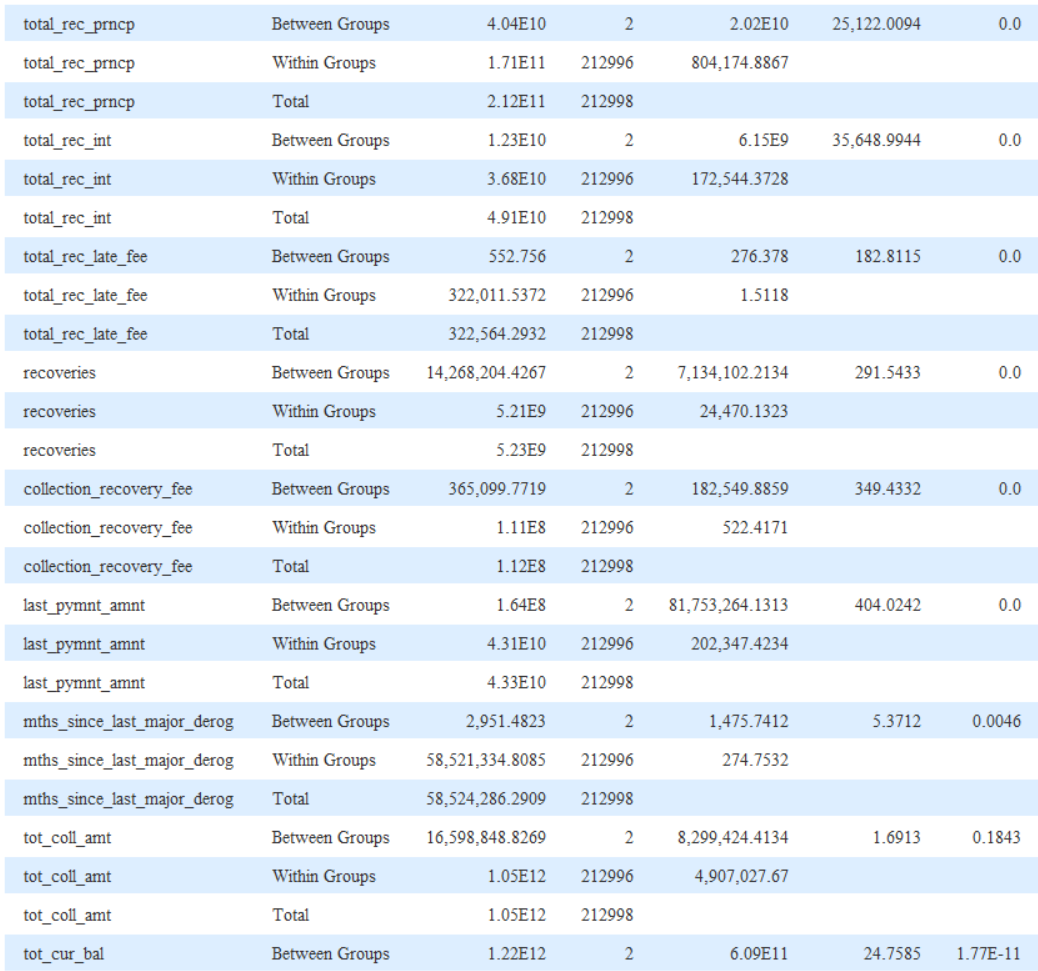
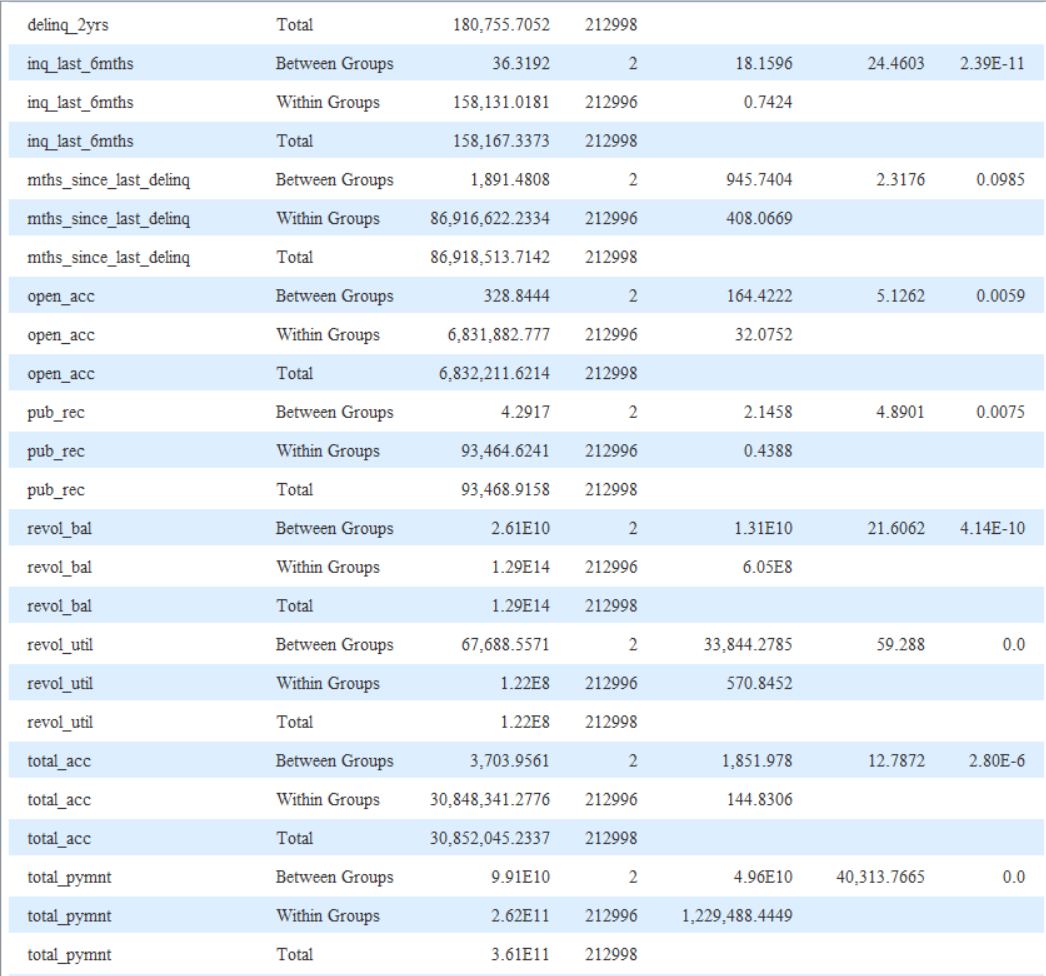
The mean and median ranks of each cluster indicate the average and middle positions of the observations within each group. The differences in these values between the two clusters suggest variations in the distribution of data points, contributing to the rejection of the null hypothesis.

We see that 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 0,1 and 2 as indicated by the Kruskal-Wallis test results.

3.2.3.2. Cluster Analysis with Non-Categorical Variables:

Analysis of Variance (ANOVA) for K= 3





The null hypothesis is rejected in all of the variables since the p-value for all the variables is less than 0.05, indicating that there are significant differences in the variables between the groups.

1. **Observations** 
   1. **Companies have been segmented.**
   2. **Appropriate number of segments**

|  |  |
| --- | --- |
| Cluster number | Silhouette Score |
| 3 | 0.581 |
| 4 | 0.541 |
| 5 | 0.581 |

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.

Thus, we will take clustering with 3 as the appropriate number of clusters due to highest 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 and the variety of services will attract the customers more.

* 1. **Cluster Analysis**

4.3.1. 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.

4.3.2. 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.

1. **Managerial Insights**

**1. Segmentation of Consumer Data**

*Targeted Marketing*: By grouping consumers with similar characteristics, you can tailor marketing campaigns, product offerings, and promotions to resonate better with each segment.

*Improved Customer Experience*: Understanding customer needs and preferences within each segment allows for personalized recommendations and improved customer service strategies.

*Resource Allocation*: Resources can be efficiently allocated by focusing efforts on high-value segments with targeted campaigns and offerings.

**2. Identifying the Right Number of Segments**

*Actionable Segments*: Having too many segments can lead to scattered efforts, while too few might miss crucial consumer distinctions. Finding the optimal number ensures actionable insights.

*Business Needs*: The ideal number of segments should be aligned with your business goals and resources.

**3. Understanding Segment Characteristics**

*Customer Profiling*: Each segment can be profiled based on demographics, purchase behavior, and other relevant factors, providing a clear picture of your customer base.

*Targeted Communication*: Insights into segment characteristics allow for crafting targeted communication strategies for each group, leading to increased engagement and conversion rates.

*Product Development*: Understanding segment needs can inform product development efforts, leading to offerings that cater to specific consumer preferences.

Managerial Actions based on KNIME Analysis:

*Decision Making*: Use the segmentation results to make data-driven decisions about marketing campaigns, product development, and resource allocation.

*Personalization*: Personalize the customer journey based on segment profiles, fostering stronger relationships and brand loyalty.

*Performance Monitoring*: Continuously monitor the performance of your segmentation strategy and refine it as needed based on new data and market trends.

**Appropriate Number of Segments**

Here are some managerial insights that can be gained from performing K-means clustering and selecting 3 clusters:

Risk Assessment: The 3 clusters represent different customer segments with varying risk profiles.

Low-risk borrowers: This cluster represent customers with a high credit score, stable income, and low delinquency rates. They should be offered lower interest rates or streamlined loan application processes.

Moderate-risk borrowers: This segment consists of customers with average creditworthiness and repayment history. They could qualify for standard loan options with moderate interest rates.

High-risk borrowers: This cluster represents individuals with a lower credit score, higher debt-to-income ratio, or history of missed payments. They can be considered stricter loan approval criteria or higher interest rates.

**Cluster - (Heterogeneous) 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.