

**FOUNDATION FOR ORGANISATIONAL**

**RESEARCH AND EDUCATION**

**NEW DELHI**

**Academic Session 2023-2025**

**Loan Applicant Profiling for Risk Assessment**

**Machine Learning for Managers**

**FMG 32 Section B**

**Submitted to Submitted by:**

**Prof. Amarnath Mitra 321105 – Samriddhi Pandey**

Contents

[1. Project Objectives 3](#_Toc161523832)

[2. Description of Data 4](#_Toc161523833)

[3. Analysis of Data 8](#_Toc161523834)

[4. Results | Observations 15](#_Toc161523835)

[5. Managerial Insights 16](#_Toc161523836)

# **1. Project Objectives**

**1.1. Segmentation of Loan Applicants:**

The objective is to segment loan applicants into distinct groups. This will aid in identifying patterns and similarities in loan approval processes.

**1.2. Identification of Appropriate Number of Segments or Clusters:**

This is to determine the optimal number of clusters for the given dataset and to evaluate the quality of the clusters generated. This involves utilizing appropriate metrics such as silhouette score to ensure the effectiveness and reliability of the clustering results.

**1.3. Determination of Segment or Cluster Characteristics:**

This is to ensure the interpretability of the generated clusters, meaning that each cluster should represent a meaningful and interpretable segment of loan applicants. This will facilitate the understanding of the characteristics and traits associated with each cluster, aiding in the development of actionable insights for loan approval processes.

# **2. Description of Data**

**2.1. Data Source, Size, Shape**

**2.1.1. Data Source:** *https://www.kaggle.com/datasets/yaminh/applicant-details-for-loan-approve*

**2.1.2. Data** Size (2 MB)

**2.1.3. Data Shape** (Dimension: 13 columns | 1,00,000 rows)

**2.2. Description of Variables**

**2.2.1. Index Variable:** Applicant ID

**2.2.2. Variables or Features having Categories**

**2.2.2.1. Categorical Variables or Features (Nominal Type):**

* **Occupation:** Profession or occupation of the loan applicant.
* **Residence City:** City where the loan applicant resides.
* **Residence State:** State where the loan applicant resides.

**2.2.2.2. Categorical Variables or Features (Ordinal Type):**

* **Marital Status:** Marital status of the loan applicant.
* **House Ownership:** Ownership status of the applicant's residence.
* **Vehicle Ownership(car):** Ownership status of the applicant's vehicle.
* **Loan Default Risk:** Indicator of loan default risk, with values indicating whether the loan applicant is at risk of defaulting on the loan

**2.2.3. Non-Categorical Variables or Features:**

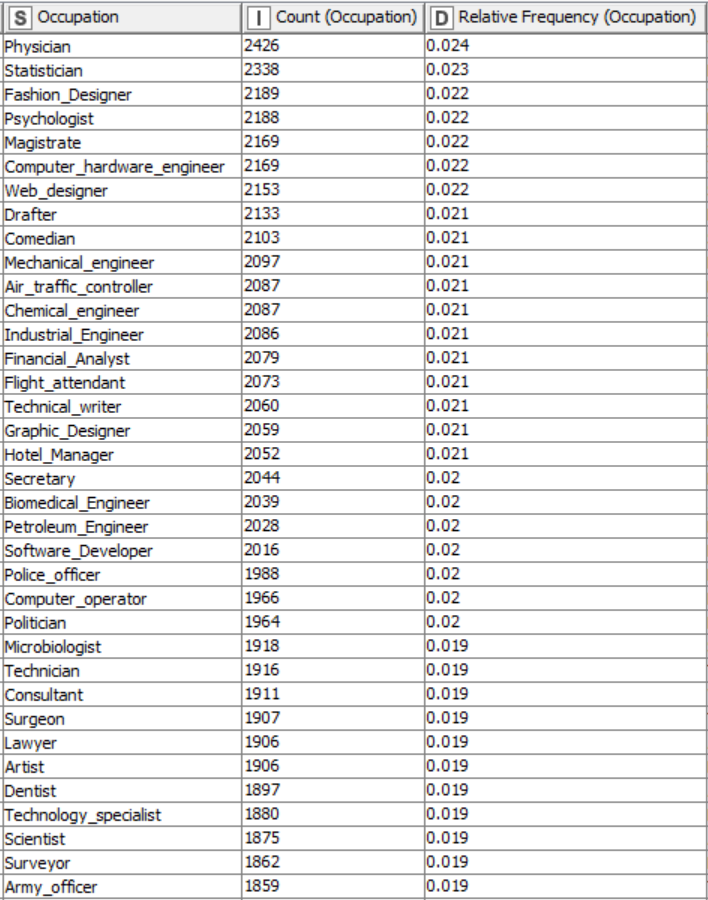
* **Annual Income:** Annual income of the loan applicant.
* **Applicant Age:** Age of the loan applicant.
* **Work Experience:** Number of years of work experience of the loan applicant.
* **Years in Current Employment:** Number of years the applicant has been in their current job.
* **Years in Current Residence:** Number of years the applicant has been residing in their current residence.

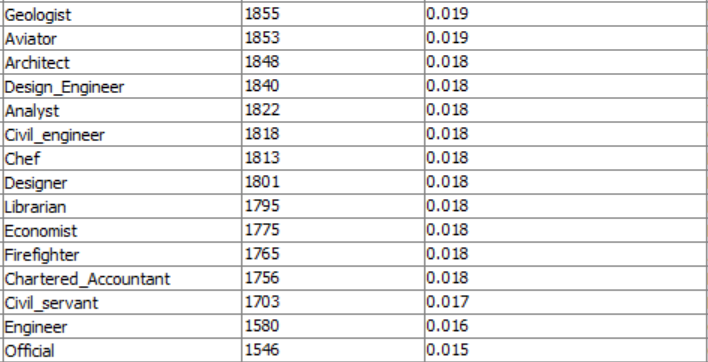
**2.3. Descriptive Statistics**

**2.3.1. Descriptive Statistics:** Categorical Variables or Features

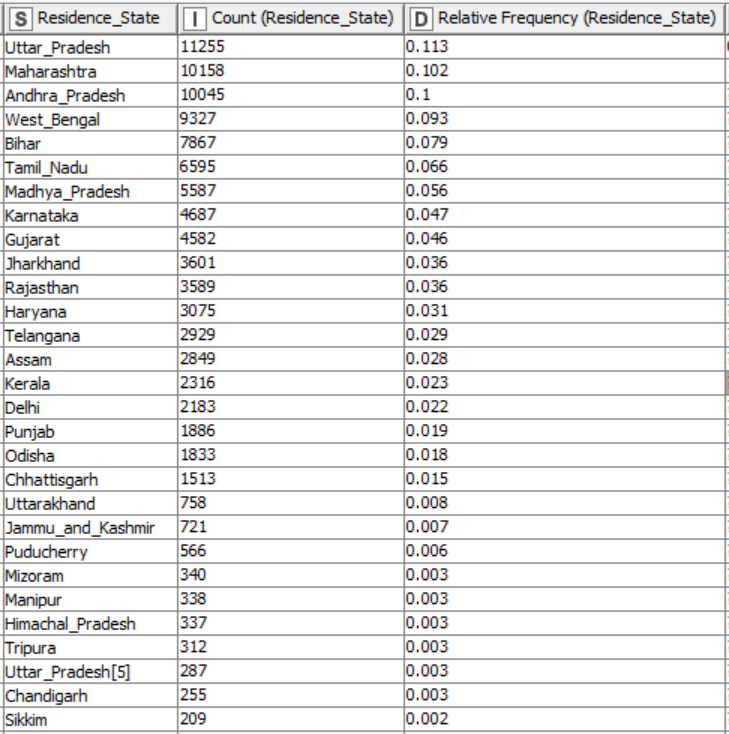
**2.3.1.1. Count | Frequency Statistics**

**Occupation**

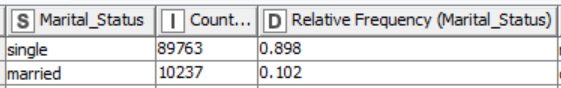




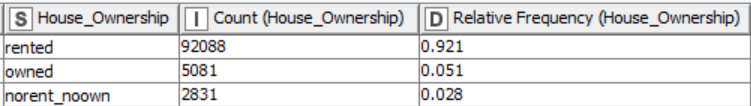
**Residence State**



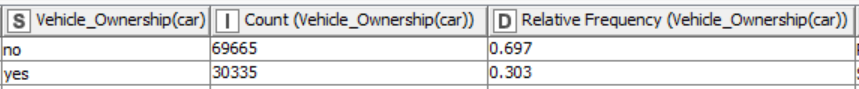
**Marital Status**



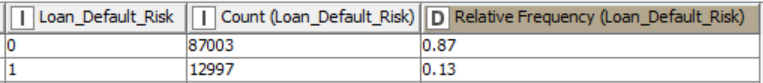
**House Ownership**



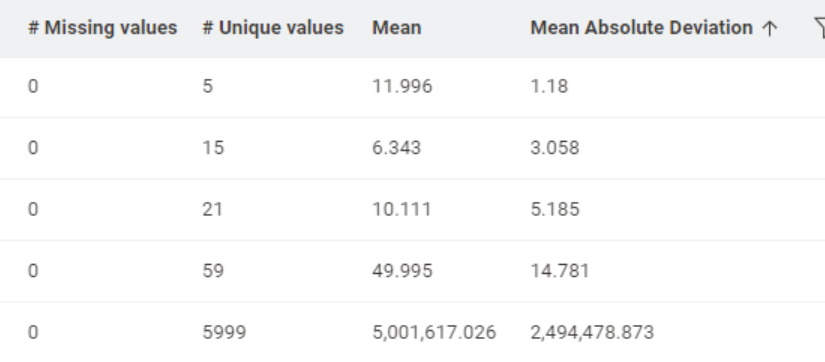
**Vehicle Ownership(car)**

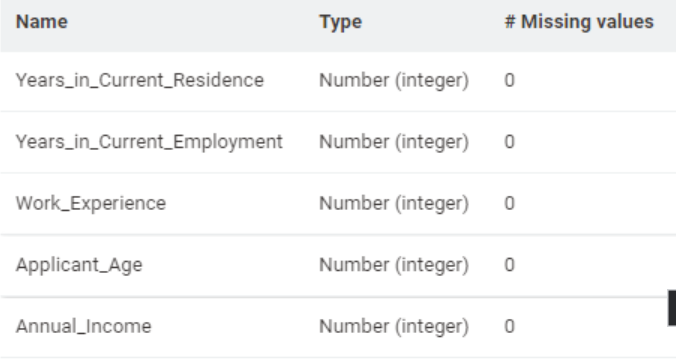


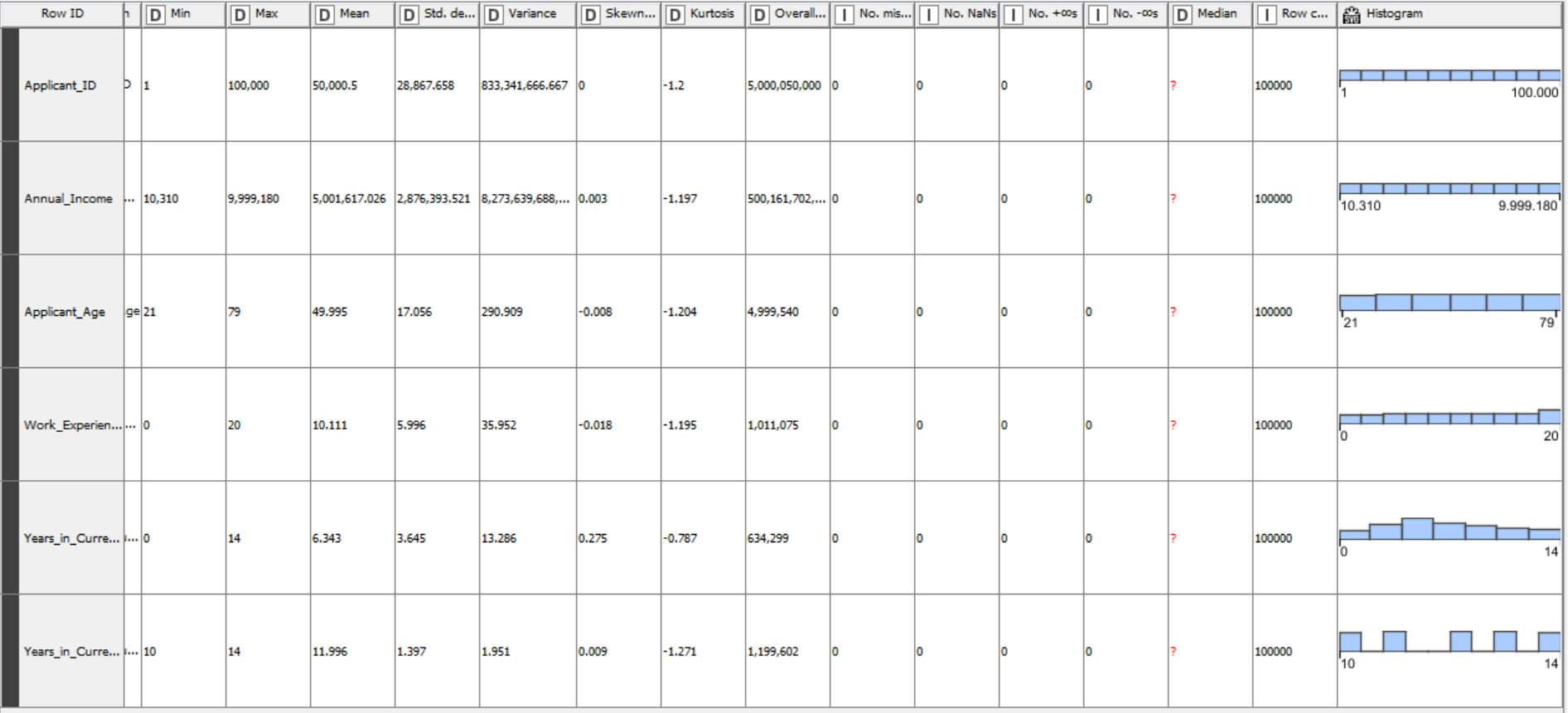
**Loan Default Risk**



**2.3.2. Descriptive Statistics: Non-Categorical Variables or Features**

** 2.3.2.1. Measures of Central Tendency**





# **3. Analysis of Data**

**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 or Features**

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

variables.

* **Marital Status**

Single-0, Married-1

* **House Ownership**

Norent\_noown- 2, Rented-1, Owned-0

* **Vehicle Ownership(car)**

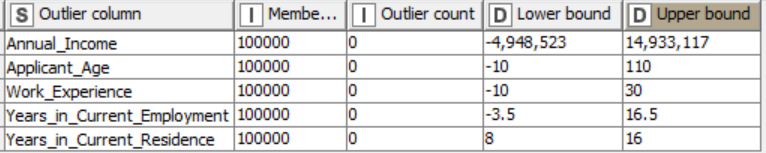
Yes-1, No-0

* **Residence State**

Punjab-0, West Bengal-1, Madhya Pradesh-2, Andhra Pradesh-3, Assam-4, Bihar-5, Rajasthan-6,Yttar Pradesh-7, Chandigarh-8, Karnataka-9, Delhi-10, Haryana-11, Gujarat-12, Maharashtra-13, Chhattisgarh-14, Kerala-15, Tripura-16, Tamil Nadu-17, Sikkim-18, Mizoram-19, Jharkhand-20, Uttar Pradesh-21, Odisha-22, Telangana-23, J&K-24, Himanchal Pradesh-25, Uttarakhand-26, Punducherry-27, Manipur-28

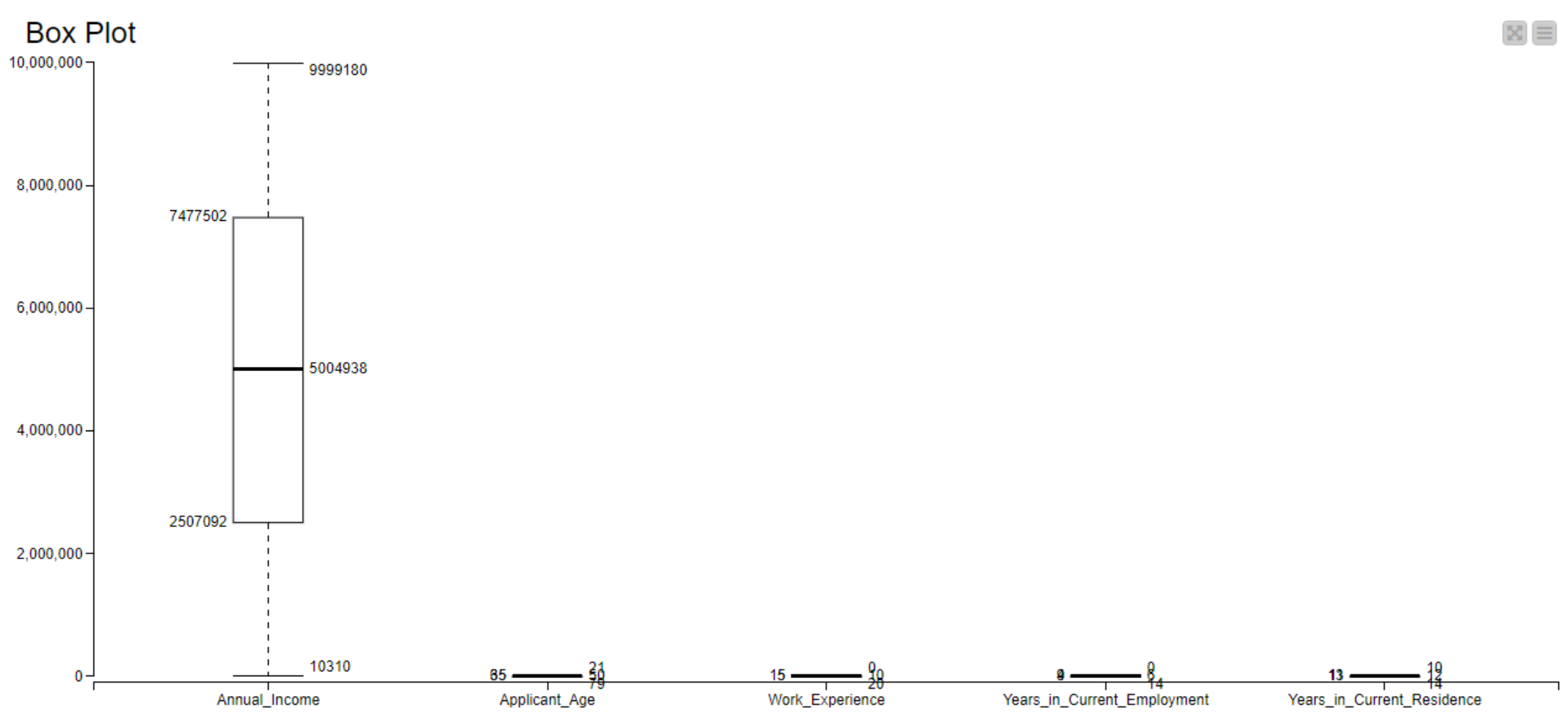
**3.1.3. Outlier Statistics and Treatment**

**3.1.3.1. Outlier Statistics: Non-Categorical Variables**

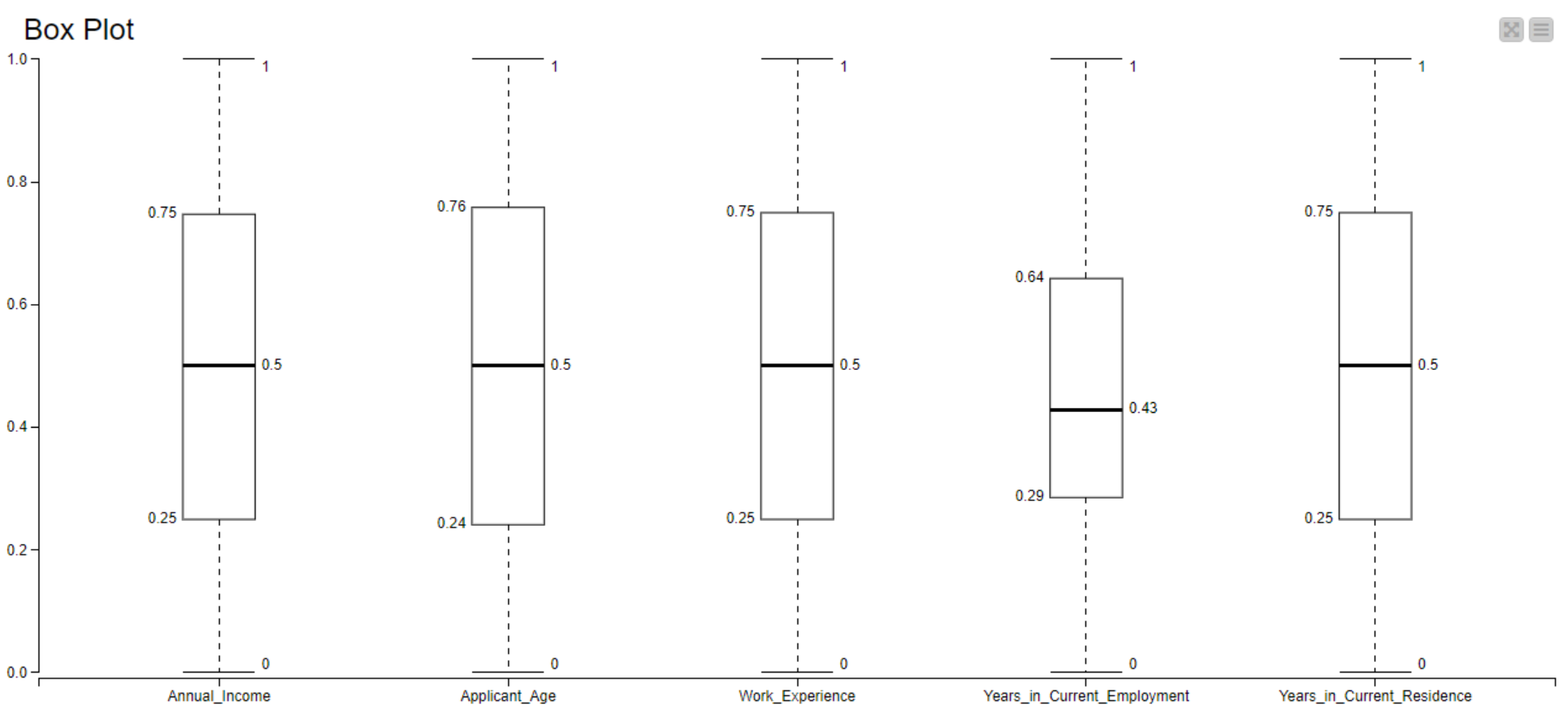


**3.1.3.1.2. Normalization using Min-Max Scaler**

**Before Normalization**



**After Normalization**



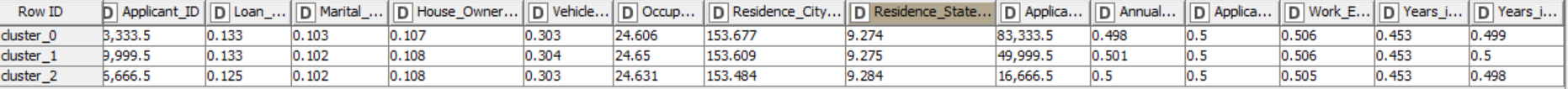
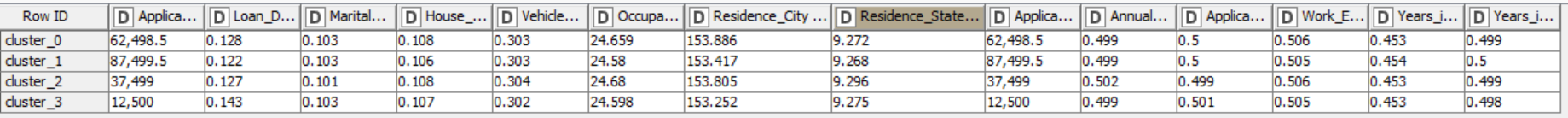
**3.1.4. Data Bifurcation: Training & Testing Sets** (Not Required)

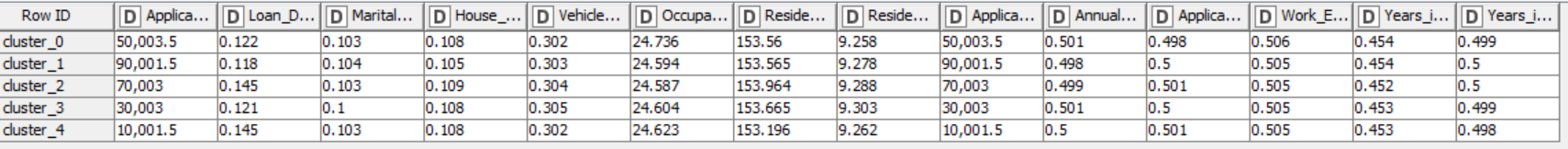
**3.2. Data Analysis**

**3.2.1. Unsupervised Machine Learning Clustering Algorithm: K-Means**

K-means clustering is an unsupervised machine learning algorithm used for partitioning data into K distinct clusters based on similarity. It iteratively assigns data points to the nearest cluster centroid and recalculates centroids until convergence. It relies on minimizing the within-cluster sum of squared distances to determine optimal cluster centroids.

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

**K=3 (This represents the total number of clusters that will be formed are 3)K=4 (This represents the total number of clusters that will be formed are 4)**

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

**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:

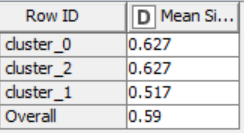
• Silhouette Score of 1 indicates that clusters are well-separated.

• Silhouette Score of 0 indicates overlapping clusters.

• Silhouette Score close to -1 indicates that samples have been assigned to the wrong

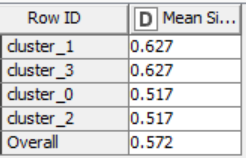
clusters.

**K=3**



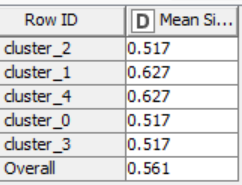
* **Cluster 0:** 0.627 (silhouettes imply good separation from other clusters)
* **Cluster 1:** 0.517 (silhouettes suggest some possible mixing with other clusters)
* **Cluster 2:** 0.627 (silhouettes imply good separation from other clusters)
* Overall silhouette score of 0.59 indicates fair cluster cohesion on average.

**K=4**



* **Cluster 0:** 0.517(silhouette score suggests some possible mixing with data points from other clusters)
* **Cluster 1:** 0.627 (silhouette score for this cluster implies good separation from other clusters)
* **Cluster 2:** 0.517 (silhouette score suggests some possible mixing with data points from other clusters)
* **Cluster 3:** 0.627 (silhouette score for this cluster implies good separation from other clusters)
* Overall: 0.572 (overall silhouette score of 0.572 indicates fair cluster cohesion on average)

**K=5**



* **Cluster\_0:** 0.517 (silhouette score suggests some possible mixing with data points from other clusters)
* **Cluster\_1:** 0.627 (score implies good separation from other clusters)
* **Cluster\_2:** 0.517 (silhouette score suggests some possible mixing with data points from other clusters)
* **Cluster\_3:** 0.517 (score suggests some possible mixing with data points from other clusters)
* **Cluster\_4:** 0.627 (score implies good separation from other clusters)
* Overall: 0.561 - The overall silhouette score of 0.561 indicates fair cluster cohesion on average.

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. Since the overall value of Mean Silhouette Coefficient of K=3 is maximum and closest to 1, so this will be choice for us.

**3.2.3 Cluster Analysis: Base Model (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.

**Loan\_Default\_Risk**

The p-value associated with the Kruskal-Wallis test is less than the significance level of 0.05.

Therefore, we reject the null hypothesis and conclude that there are statistically significant differences between the medians of cluster 0 and cluster 1.

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

hypothesis.

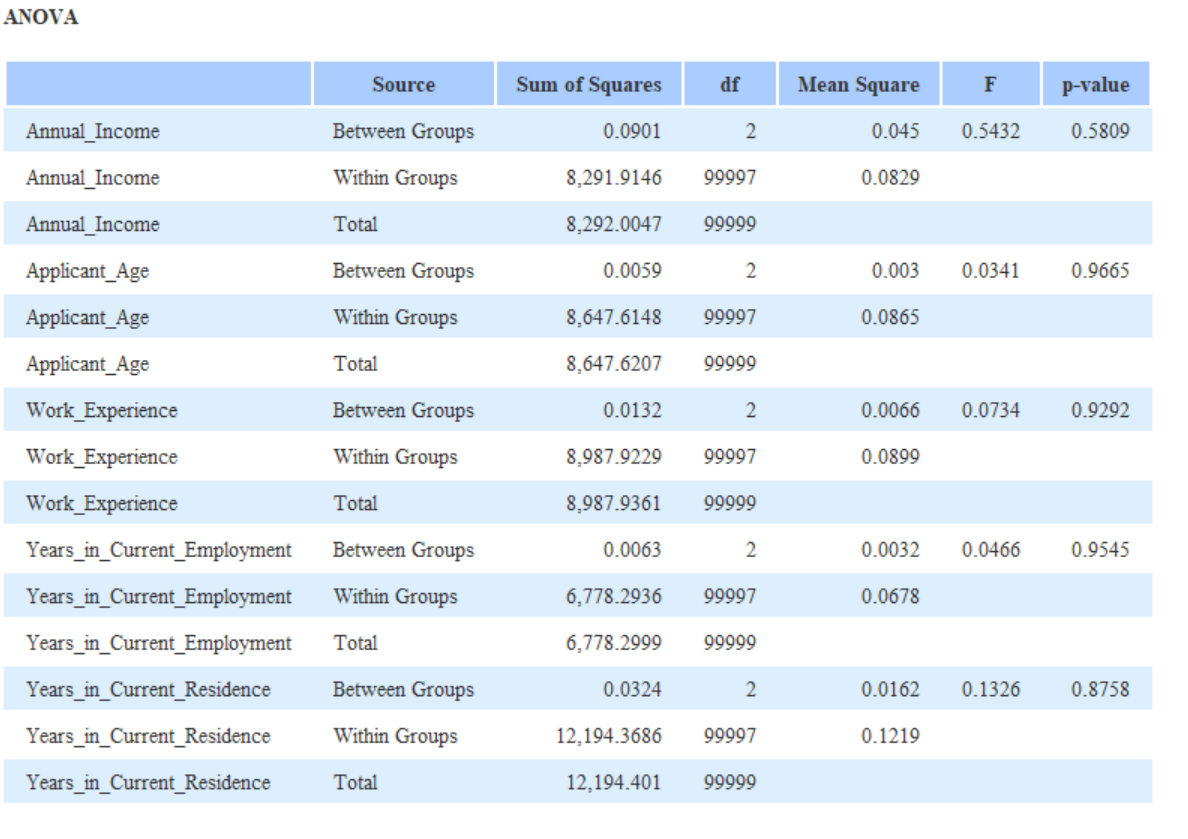
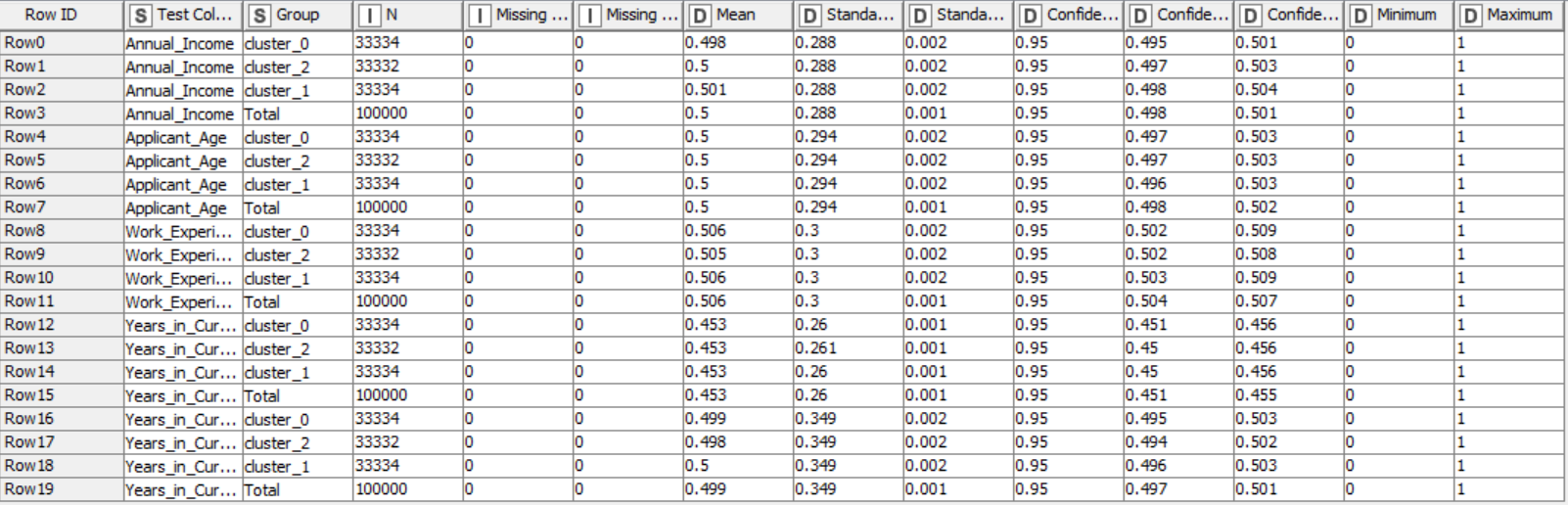
**Hence we see that loan default risk is significant for the cluster.**

**3.2.3.2. Cluster Analysis with Non-Categorical Variables (ANOVA)**

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.

**For K = 3**

**Descriptive Statistics:**

**As the p-value>0.05 hence we can say that Annual Income, Applicant age, work experience and others are not significant.**

# **4. Results | Observations**

**4.1. Appropriate Number of Segments**

|  |  |  |
| --- | --- | --- |
| **Cluster Number** | **Cluster** | **Silhouette Score** |
| 3 | Cluster 0 | 0.627 |
| Cluster 1 | 0.517 |
| Cluster 2 | 0.627 |
| 4 | Cluster 0 | 0.517 |
| Cluster 1 | 0.627 |
| Cluster 2 | 0.517 |
| Cluster 3 | 0.627 |
| 5 | Cluster 0 | 0.517 |
| Cluster 1 | 0.627 |
| Cluster 2 | 0.517 |
| Cluster 3 | 0.517 |
| Cluster 4 | 0.627 |

The silhouette score for all the clusters is present.

Higher the silhouette score i.e. close to 1 more are the clusters separated and close to 0 indicates the clusters are overlapping.

**4.2. Cluster Analysis**

**4.2.1 Categorical Variables**

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

**4.2.2 Non-Categorical Variables**

It has been observed that no variables are contributing to the cluster. This is because the p-value is greater than 0.05 (confidence level at 95% for the model)

which in turn tells that all the non-categorical variables are not 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 are:**

* **Decision Making:** Choosing the appropriate number of clusters is crucial for actionable insights. While clustering with 2 clusters provides a faster solution and relatively lower memory usage, it may oversimplify the dataset. Clustering with 3, 4, or 5 clusters offers a more nuanced understanding of the data's underlying patterns but requires more computational resources.
* **Resource Utilization:** The choice of the number of clusters impacts computational resources, such as time and memory. As the number of clusters increases, so does the processing time and memory usage. Therefore, it is essential to strike a balance between resource efficiency and the granularity of clustering.
* **Interpretability:** Fewer clusters are easier to interpret and may lead to more straightforward decision-making processes. However, too few clusters may overlook valuable insights hidden in the data. Conversely, too many clusters can complicate interpretation and hinder decision-making.

**Recommendation:** Considering the evaluation metrics and their implications, **clustering with 3 clusters** strikes a balance between resource utilization and granularity of clustering. It provides a moderately fast solution with reasonable memory usage while offering a more detailed understanding of the data compared to clustering with 2 clusters. Therefore, we recommend proceeding with clustering utilizing 3 clusters for further analysis and decision-making.