

**FOUNDATION FOR ORGANIZATION RESEARCH AND EDUCATION (FORE) SCHOOL OF MANAGEMENT**

**NEW DELHI**

**Academic Year 2023 – 25**

**MACHINE LEARNING FOR MANAGERS**

**PROGECT REPORT**

**Submitted to: Prof. AMARNATH MITRA**

**Submitted by:**

**SUBHAJIT PAUL**

**ROLL: 321173**

**SEC: C**

**Project Contents**

1. **Project Objectives | Problem Statements**
2. **Description of Data**
3. **Analysis of Data**
4. **Results | Observations**
5. **Managerial Insights**

**Project 1: Report**

**Project Title: Segmentation of Consumer Data**

**1. Project Objectives | Problem Statements**

1.1. PO1 | PS1: Segmentation of Consumer Data using Unsupervised Machine Learning Clustering Algorithms

1.2. PO2 | PS2: Identification of Appropriate Number of Segments or Clusters

1.3. PO3 | PS3: Determination of Segment or Cluster Characteristics

**2. Description of Data**

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

2.1.1. Data Source (Website Link)

-- <https://www.kaggle.com/datasets/parisrohan/credit-score-classification>

2.1.2. Data Size (in KB | MB | GB …)

-- 10 MB

2.1.3. Data Shape | Dimension: Number of Variables | Number of Records

-- Number of Variables: 27, Number of Records: 60000

**2.2. Description of Variables**

2.2.1. Index Variable(s): ID

2.2.2. Variables or Features having Categories | Categorical Variables or Features (CV)

Categorical variables are:

* + - ID
    - Customer\_ID
    - Name
    - Occupation
    - Credit\_Mix
    - Payment\_of\_Min\_Amount
    - Credit\_Score

2.2.2.1. Variables or Features having Nominal Categories | Categorical Variables or Features - Nominal Type:

* Name
* Occupation

2.2.2.2. Variables or Features having Ordinal Categories | Categorical Variables or Features - Ordinal Type:

* Credit\_Mix
* Payment\_of\_Min\_Amount
* Credit\_Score

2.2.3. Non-Categorical Variables or Features:

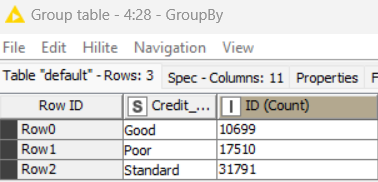
* ID
* Customer\_ID
* Age
* Annual\_Income
* Num\_Bank\_Accounts
* Num\_Credit\_cards
* Interest\_Rate
* Num\_of\_Loan
* Delay\_from\_due\_date
* Num\_of\_Delayed\_Payment
* Changed\_Credit\_Limit
* Num\_Credit\_Inquiries
* Outstanding\_Debt
* Credit\_Utilization\_Ratio
* Credit\_History\_Age\_days
* Total\_EMI\_per\_month
* Monthly\_Balance

**2.3. Descriptive Statistics**

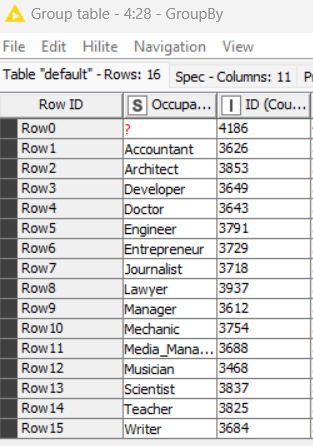
2.3.1. Descriptive Statistics: Categorical Variables or Features

2.3.1.1. Count | Frequency Statistics

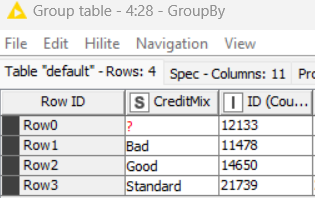
Credit Score:



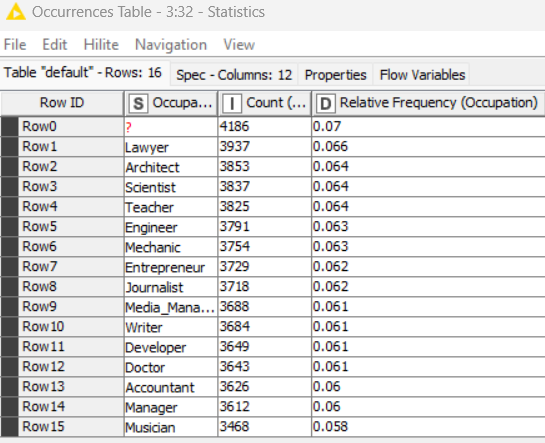
Occupation:

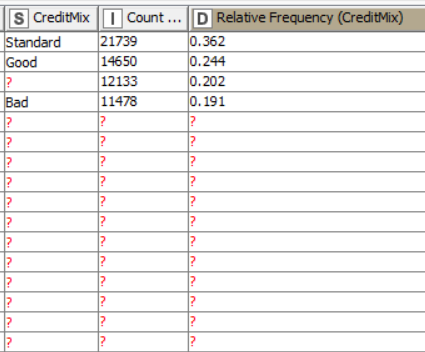


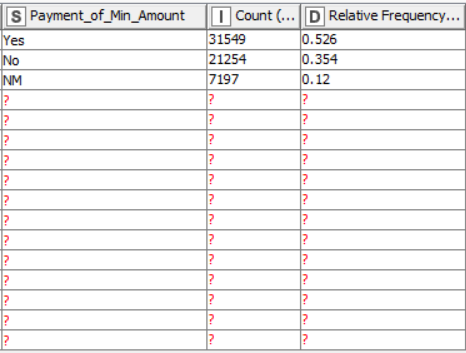
Credit Mix:

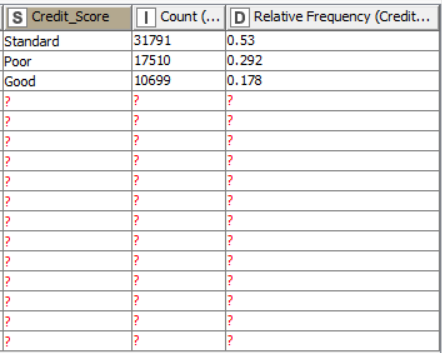


2.3.1.2. Proportion (Relative Frequency) Statistics





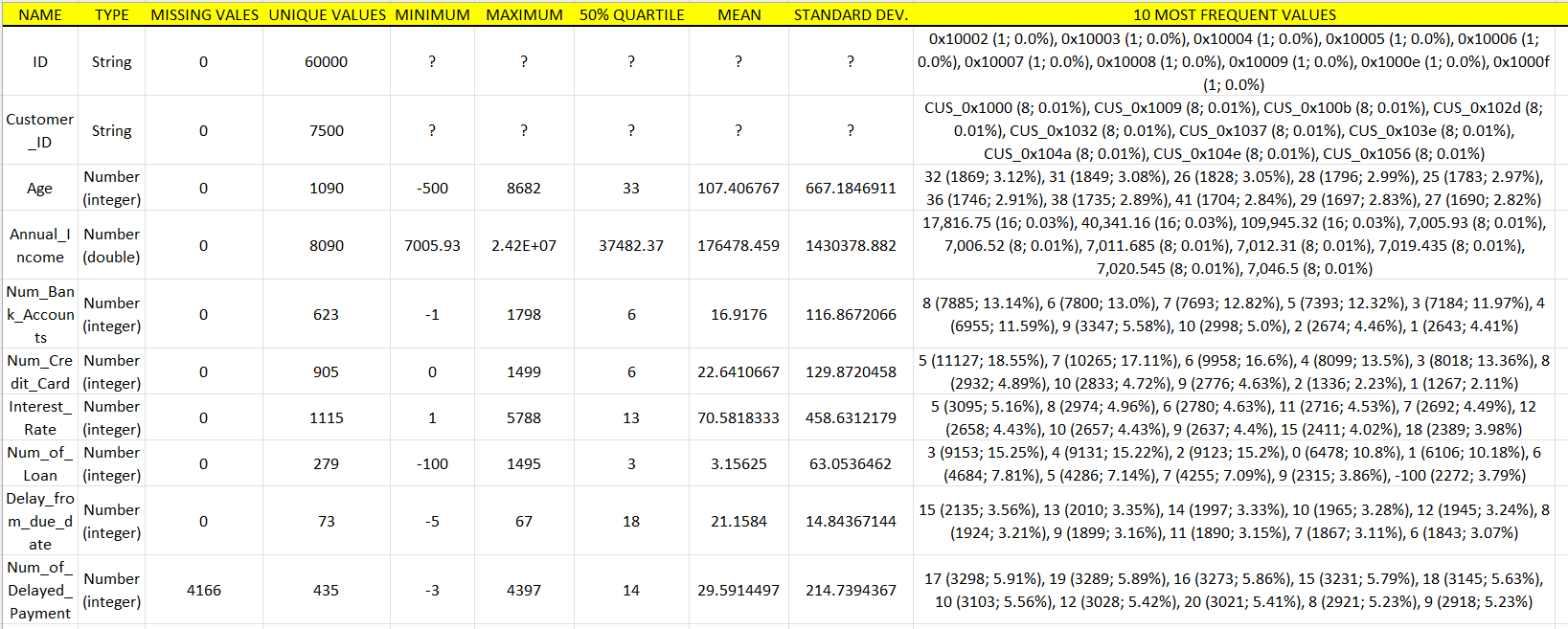


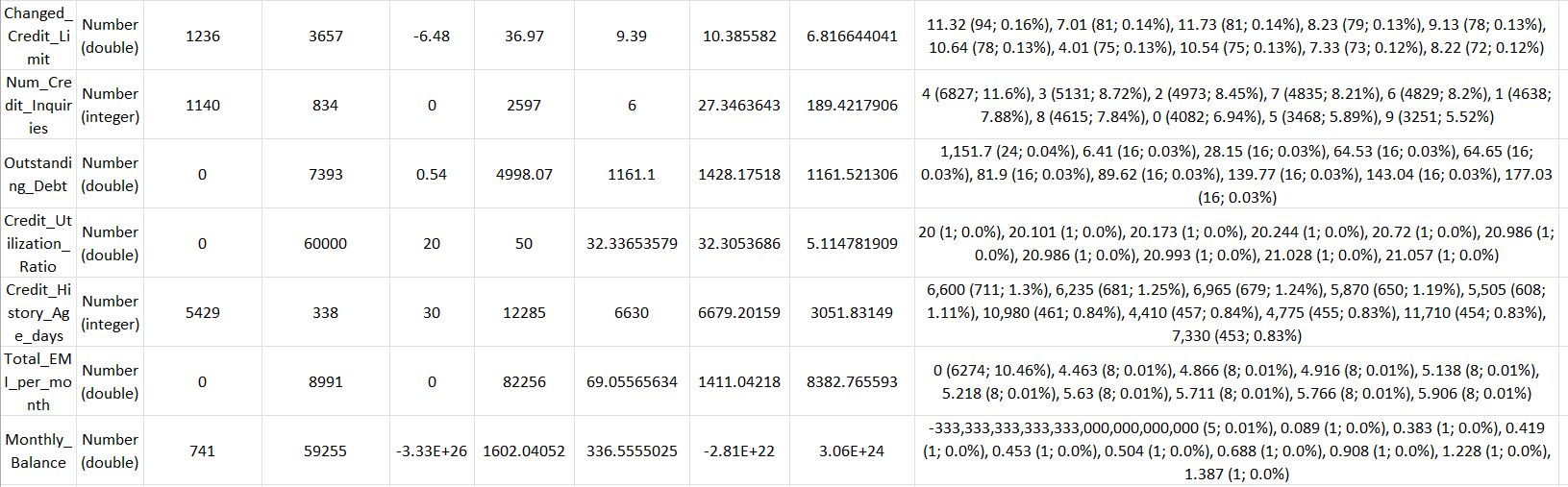


2.3.2. Descriptive Statistics: Non-Categorical Variables or Features

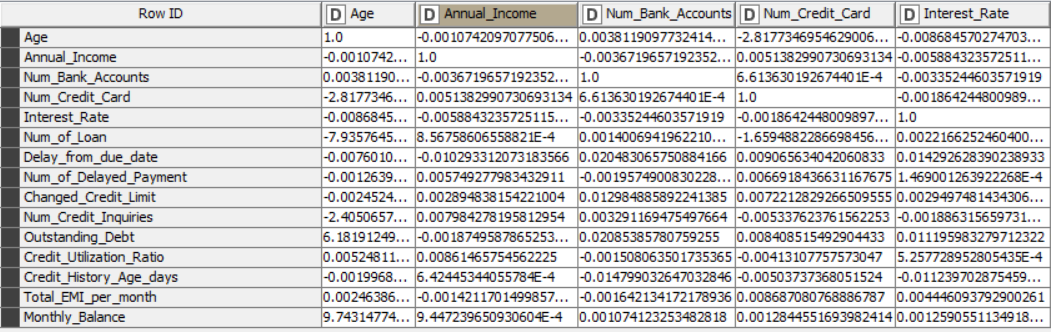
2.3.2.1. Measures of Central Tendency

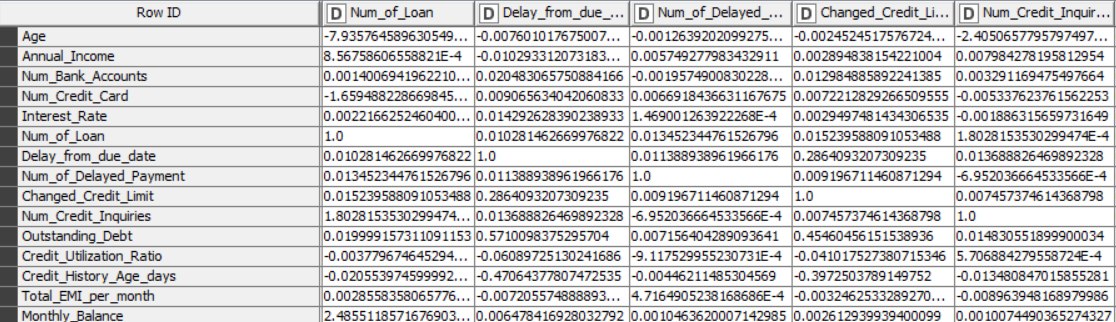
2.3.2.2. Measures of Dispersion

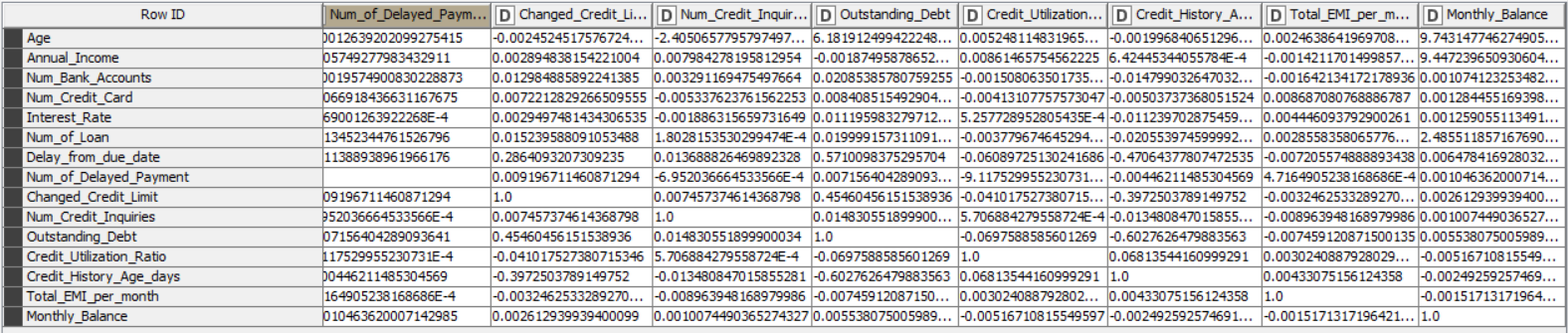




2.3.2.3. Correlation Statistics (with Test of Correlation)







Where the p value is less than 0.05 in the matrix, there the two variables are correlated. For example,

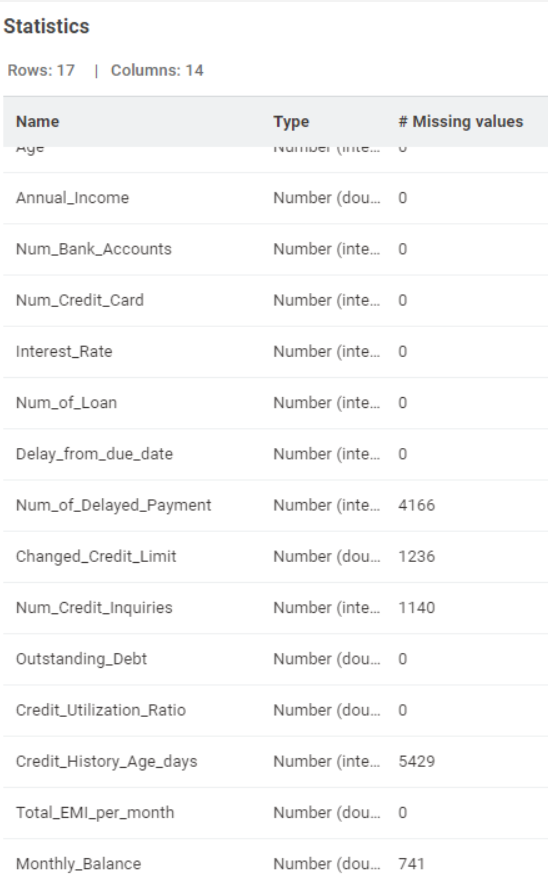
Annual income and Changed Credit Limit are correlated.

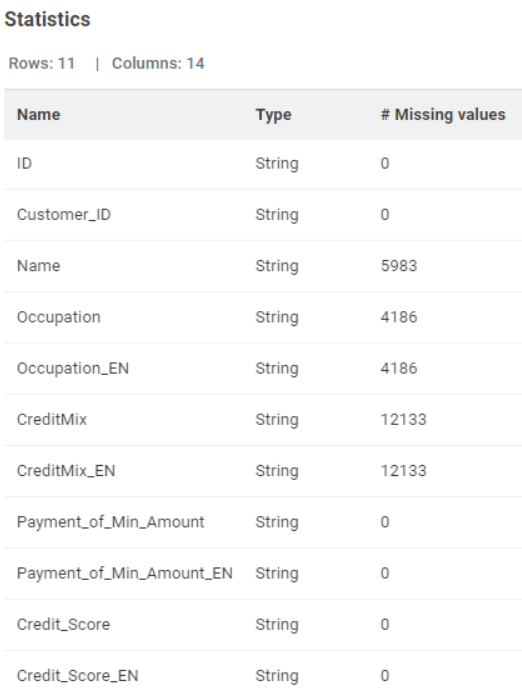
**3. Analysis of Data**

**3.1. Data Pre-Processing**

**3.1.1. Missing Data Statistics and Treatment**

3.1.1.1.1. Missing Data Statistics: Records

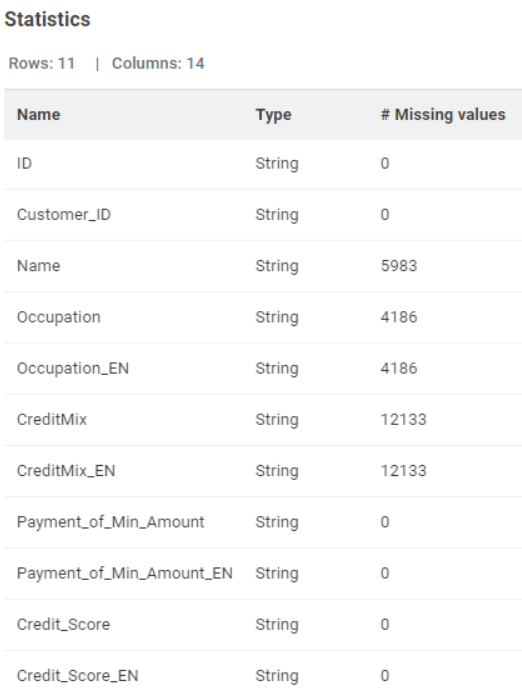




3.1.1.1.2. Missing Data Treatment: Records

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

3.1.1.2.1. Missing Data Statistics: Categorical Variables or Features

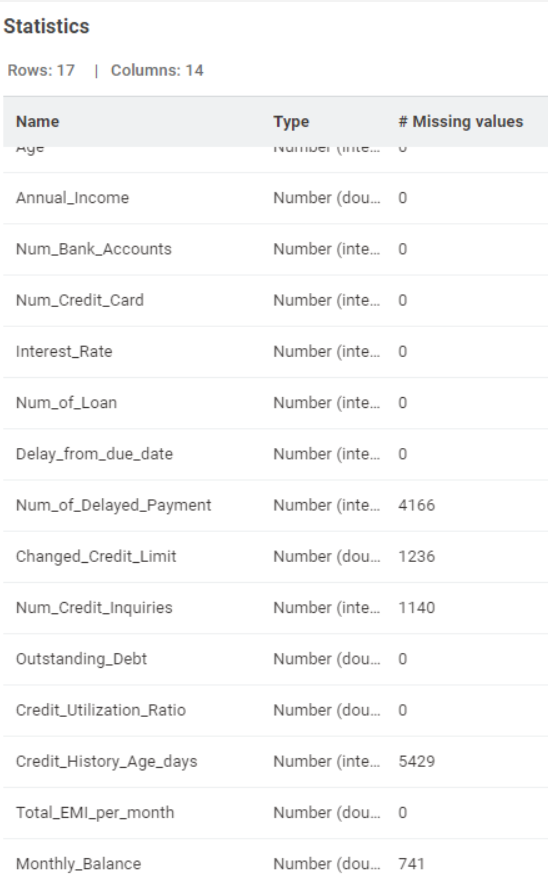


3.1.1.2.2. Missing Data Treatment: Categorical Variables or Features

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

3.1.1.2.2.2. Imputation of Missing Data using Descriptive Statistics: Mode – Most Frequent Value

3.1.1.3.1. Missing Data Statistics: Non-Categorical Variables or Features

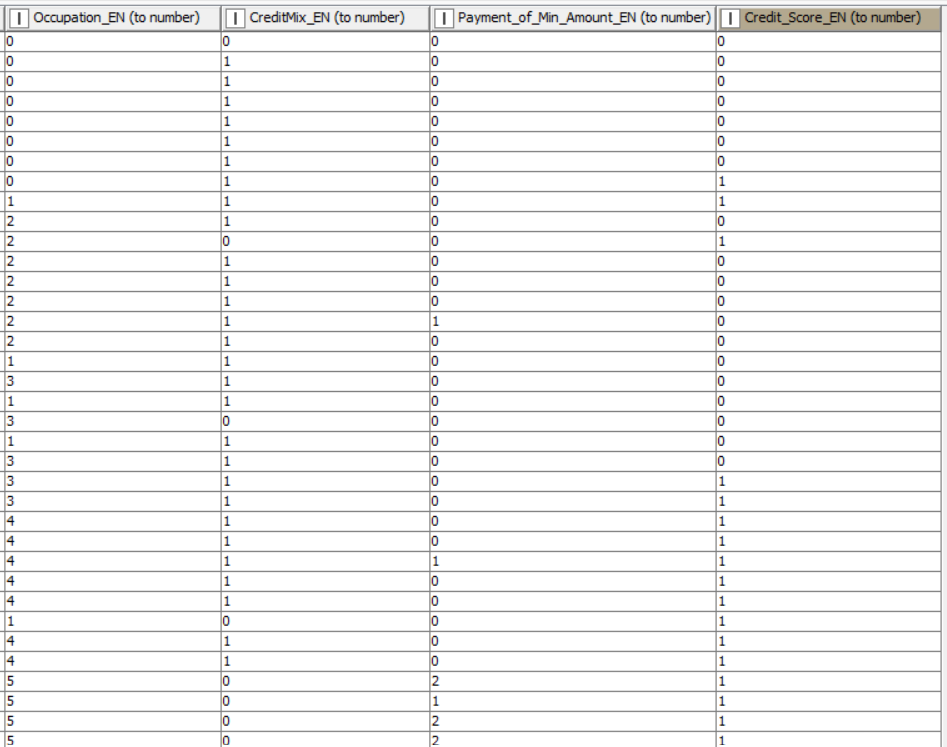


3.1.1.3.2. Missing Data Treatment: Non-Categorical Variables or Features

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

3.1.1.3.2.2. Imputation of Missing Data using Descriptive Statistics: Mean

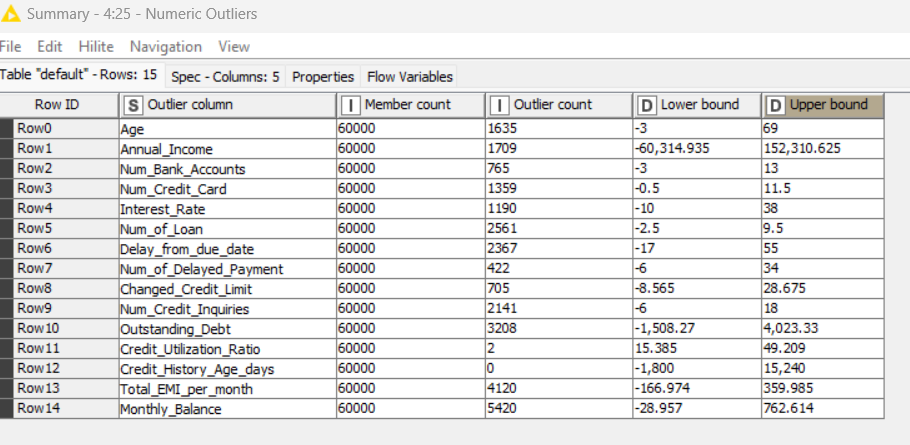
**3.1.2. Numerical Encoding of Categorical Variables or Features** (Encoding Schema - Alphanumeric Order)



Occupation, Credit Mix, Payment of minimum amount, Credit Score has been encoded to numeric characters.

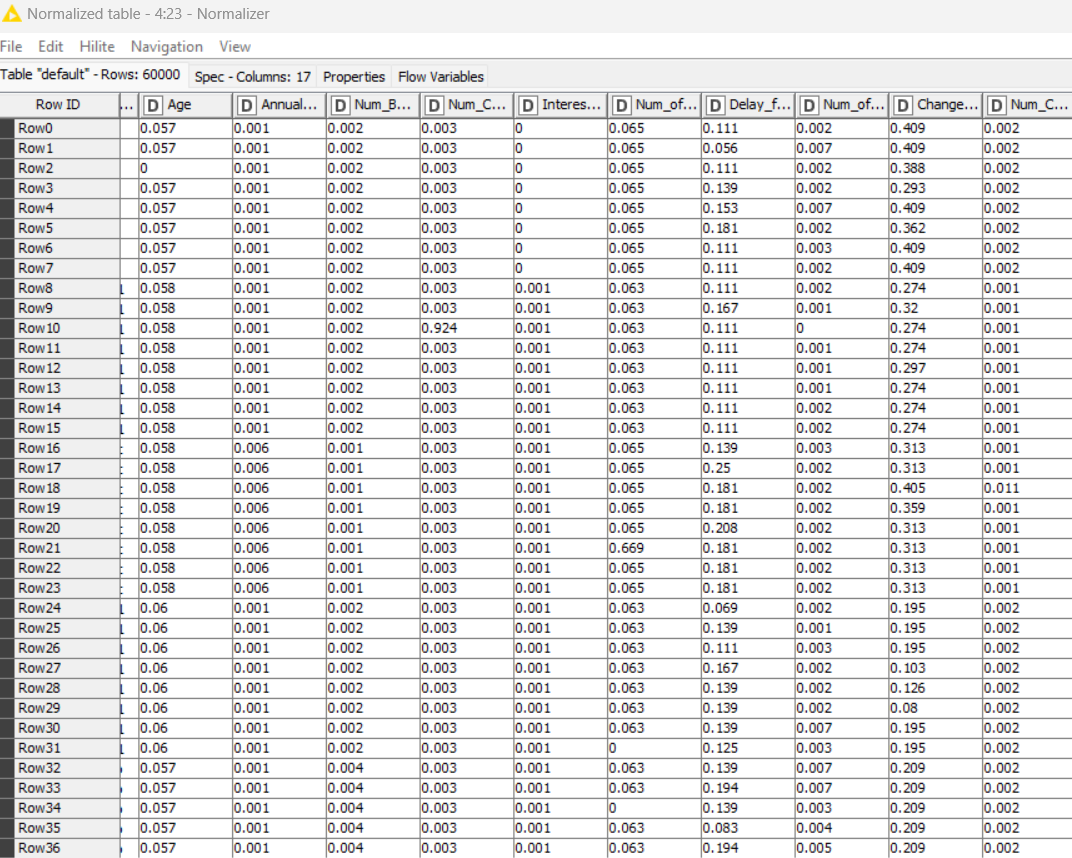
**3.1.3. Outlier Statistics and Treatment** (Scaling | Transformation)

3.1.3.1.1. Outlier Statistics: Non-Categorical Variables or Features



3.1.3.1.2. Outlier Treatment: Non-Categorical Variables or Features

3.1.3.1.2.2. Normalization using Min-Max Scaler:



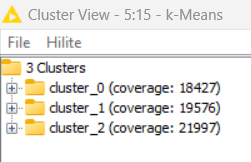
We have done normalization using MIN-MAX scaler where minimum value is 0 and maximum value is 1.

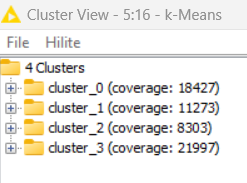
**3.1.4. Data Bifurcation: Training & Testing Sets**

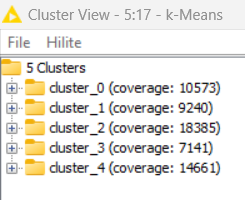
Not required

**3.2. Data Analysis**

3.2.1.1. PO1 | PS1:: Unsupervised Machine Learning Clustering Algorithm: K-Means (Base Model) | Metrics Used - Euclidean Distance

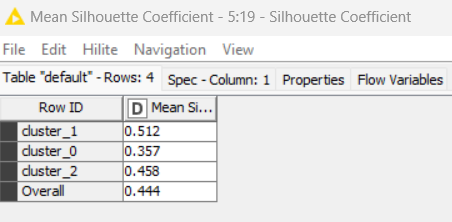


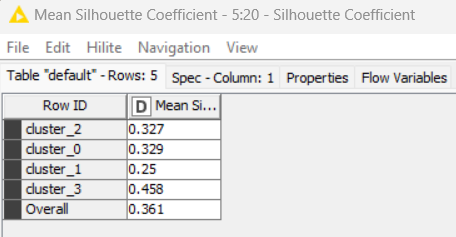


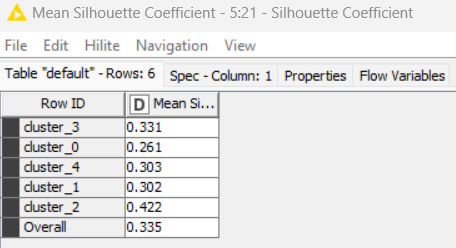


Here are the snippets of k-means for 3,4 & 5 clusters

3.2.2.1.1. PO2 | PS2: Clustering Model Performance Evaluation: Silhouette Score:





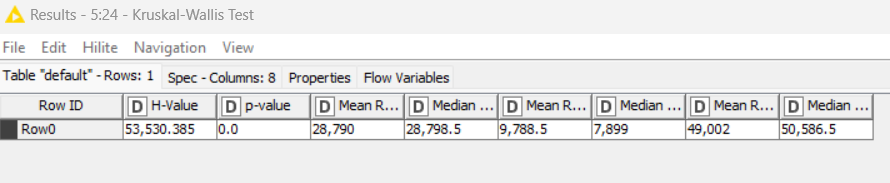


The overall mean Silhouette Coefficient is maximum for 3 clusters. It means 3 clusters can segment the data into better format.

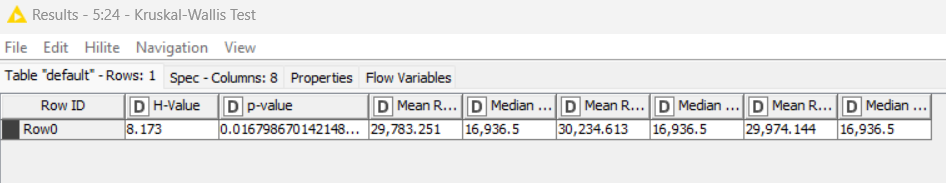
3.2.3.1. PO3 | PS3:: Cluster Analysis: Base Model (K-Means)

3.2.3.1.1. Cluster Analysis with Categorical Variables or Features: Chi-Square Test of Independence

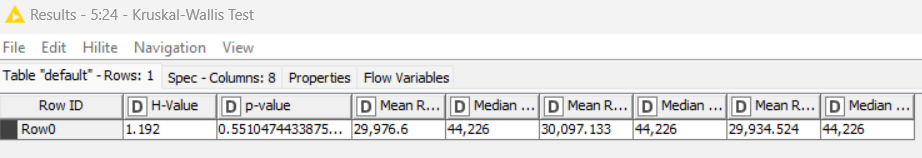
With Occupation:



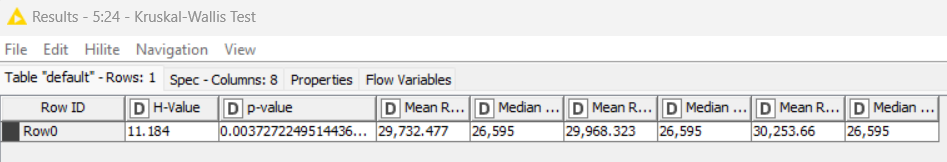
With Credit Mix:



With Payment of minimum amount:

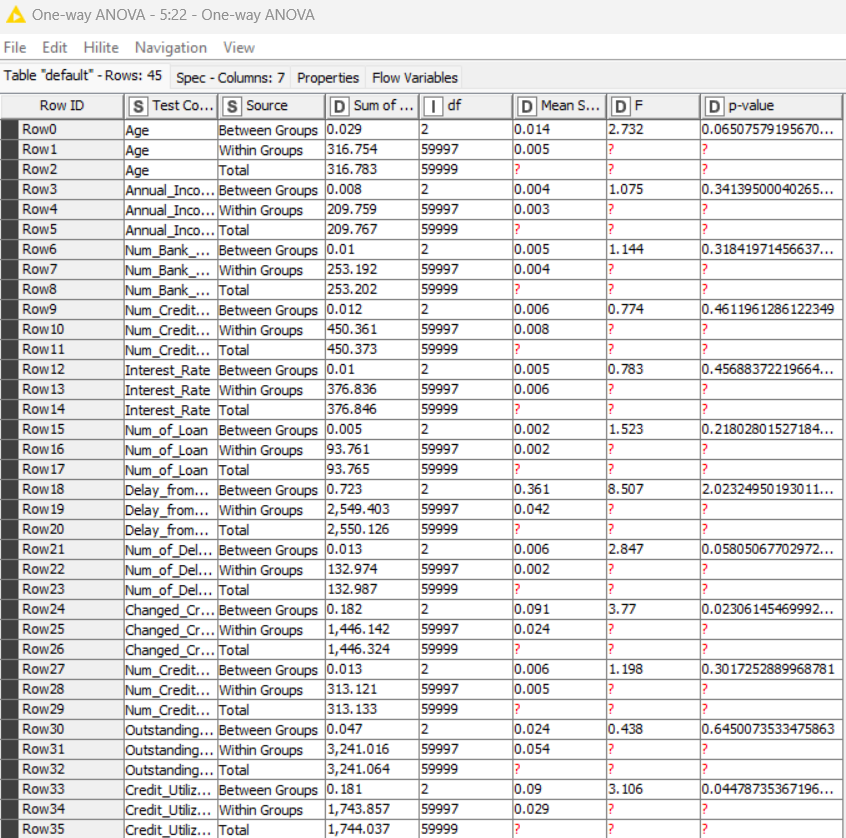


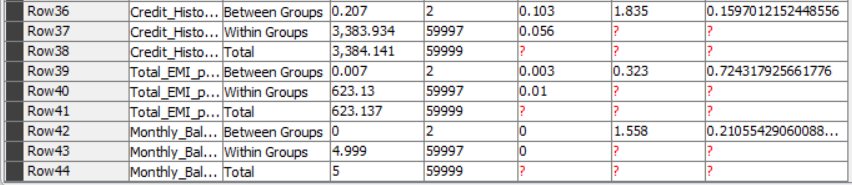
With Credit Score:



In case of Occupation, Credit Mix and Credit Score, the P value is less than 0.05. So, these variables can define the cluster significantly.

3.2.3.1.2. Cluster Analysis with Non-Categorical Variables or Features: Analysis of Variance (ANOVA)





Changed\_Credit\_Limit and Credit\_Utilization\_Ratio has p value less than 0.05. Therefore, these two variables can distinguish the clusters significantly.

**4. Results | Observations**

4.1. Appropriate Number of Segments | Clusters: Base Model (K-Means): 3

4.3. Cluster Analysis: Base Model (K-Means)

4.3.1. Categorical Variables or Features: Contributing or Significant | Non-Contributing or Non-Significant

Contributing variables are:

* Occupation
* Credit\_Mix
* Credit\_Score

Other variables are non-contributing.

4.3.2. Non-Categorical Variables or Features: Contributing or Significant | Non-Contributing or Non-Significant

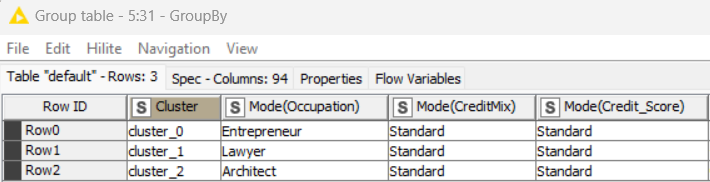
* Changed\_Credit\_Limit
* Credit\_Utilization\_Ratio

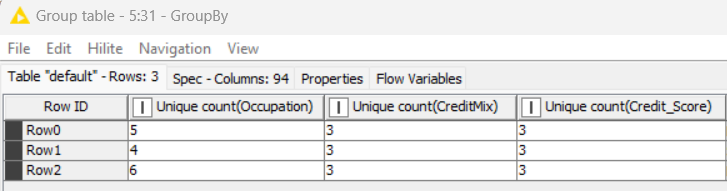
**5. Managerial Insights**

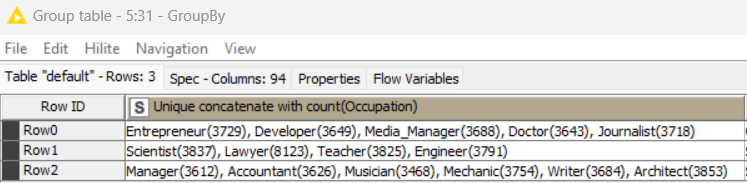
5.1. Appropriate Number of Segments | Clusters (Given the Appropriate Model):

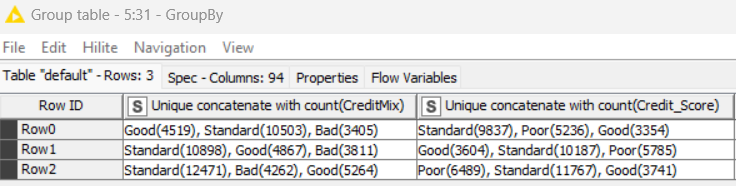
Appropriate Number of segments are 3

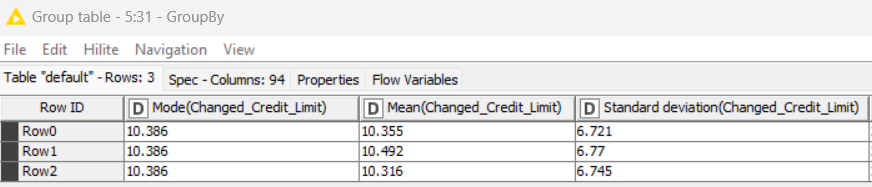
5.3. Segment | Cluster - (Heterogeneous) Identity | Nomenclature

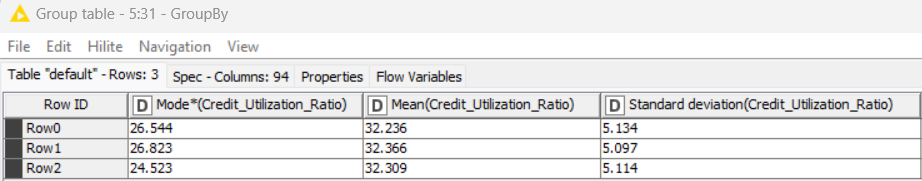


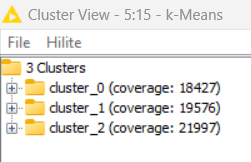












CLUSTER 0: Entrepreneur – 20.23%

Standard Credit Mix – 57.01%

Standard Credit Score – 56.99%

Mean Changed\_Credit\_Limit is 10.355% with s.d. 6.721%

Mean Credit\_Utilization\_Ratio is 32.236% with s.d. 5.134%

CLUSTER 1: Lawyer – 41.49%

Standard Credit Mix – 55.67%

Standard Credit Score – 52.07%

Mean Changed\_Credit\_Limit is 10.492% with s.d. 6.77%

Mean Credit\_Utilization\_Ratio is 32.366% with s.d. 5.097%

CLUSTER 2: Architect – 17.51%

Standard Credit Mix – 56.69%

Standard Credit Score – 53.49%

Mean Changed\_Credit\_Limit is 10.316% with s.d. 6.745%

Mean Credit\_Utilization\_Ratio is 32.309% with s.d. 5.114%

Change in credit limit is highest for lawyers, and lowest for Architects. Credit Utilization ratio is highest for Lawyers, and lowest for Entrepreneurs.

Therefore, most valuable customer base for credit card is lawyers.