AI based Tax Fraud Detection

T Edward Sam, Senior Technical Director
Arun Sankar, Technical Director
Arun K Varghese, Senior Systems Analyst
Ankith Kirti, Systems Analyst

Objectives

- Identify suspicious taxpayers likely to evade tax
- Estimate the amount of tax that could be recovered
- Provide the suspicious taxpayer list for scrutiny, assessment and issue of notices and recovery
- Compare the identified suspicious list with the actual list of confirmed tax evaders

Python and open source AI Libraries and Algorithms

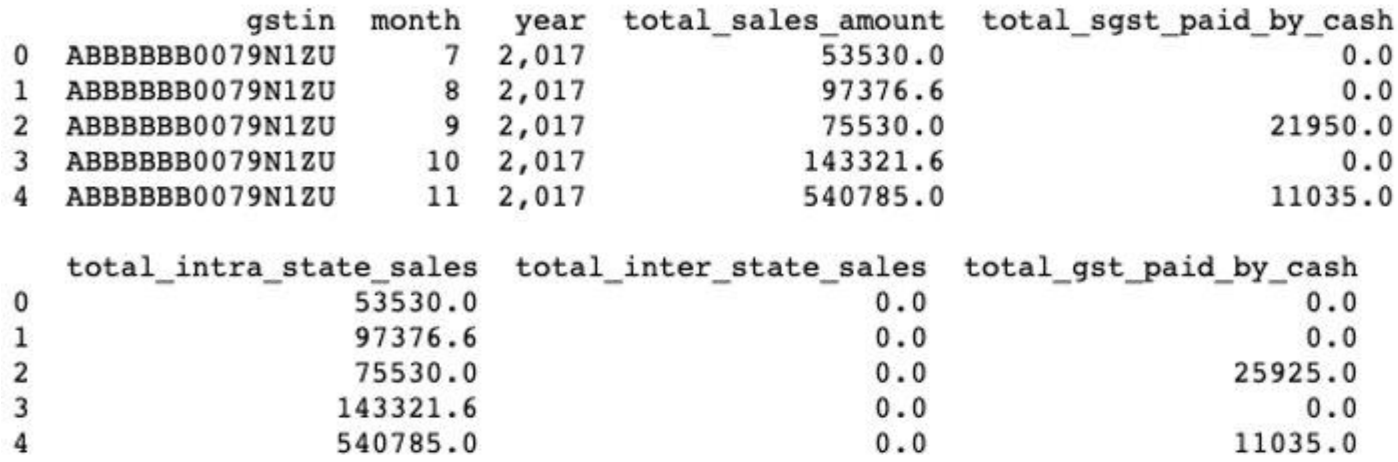
- AI Libraries used
  - Sklearn
  - Pandas
  - Plotly
- AI algorithms used of clustering:
  - DBSan
  - K-Mediods
  - K-Means
  - Silhouette scoring

- Fraud Detection
  - Graph based weighted MAD
  - Benford Analysis

Base Data for Clustering

- Identified 7 key base data fields from monthly GST return filing for correlation analysis
  - Total SGST Liability
  - Total CGST Liability
  - Total Liability
  - Total ITC available
  - Total IGST ITC
  - Total SGST paid by cash
  - Total Exempt Sales
- Identified 5 major business sectors for independent study
  - Gold
  - Timber
  - Steel
  - Vehicle
  - Textile

Base Data Sample

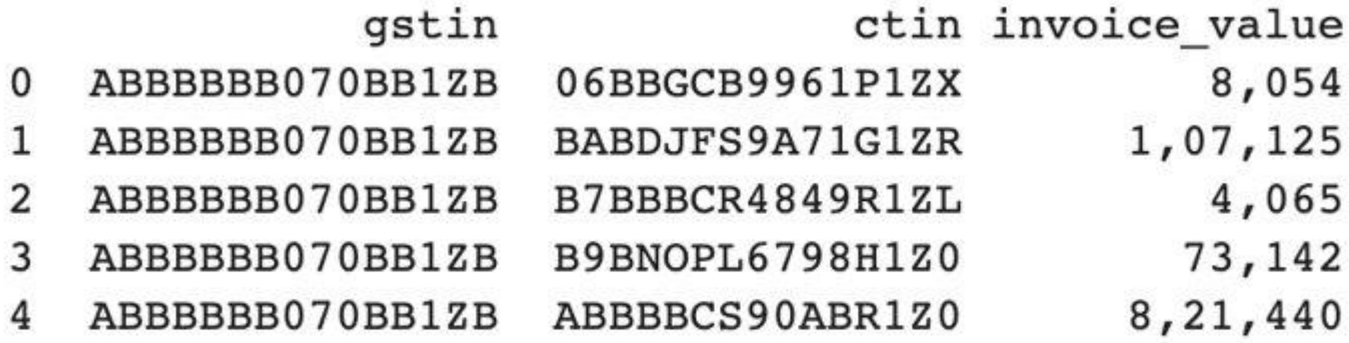


```
             gstin  month   year  total_sales_amount  total_sgst_paid_by_cash
0  ABBBBBB0079N1ZU      7  2,017             53530.0                      0.0
1  ABBBBBB0079N1ZU      8  2,017             97376.6                      0.0
2  ABBBBBB0079N1ZU      9  2,017             75530.0                  21950.0
3  ABBBBBB0079N1ZU     10  2,017            143321.6                      0.0
4  ABBBBBB0079N1ZU     11  2,017            540785.0                  11035.0

   total_intra_state_sales  total_inter_state_sales  total_gst_paid_by_cash
0                  53530.0                      0.0                     0.0
1                  97376.6                      0.0                     0.0
2                  75530.0                      0.0                 25925.0
3                 143321.6                      0.0                     0.0
4                 540785.0                      0.0                 11035.0
```

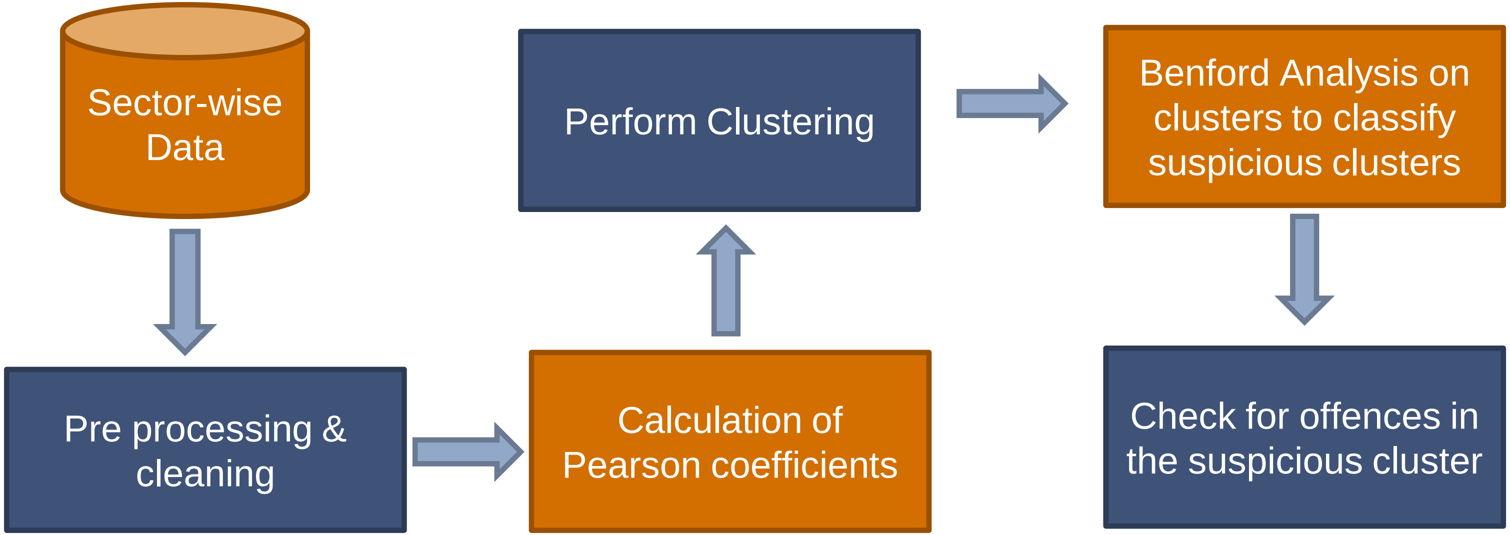
Base data for Benford Analysis

- Business to business invoice data from GSTR-1 return



```
            gstin            ctin invoice_value
0  ABBBBBB070BB1ZB  06BBGCB9961P1ZX         8,054
1  ABBBBBB070BB1ZB  BABDJFS9A71G1ZR      1,07,125
2  ABBBBBB070BB1ZB  B7BBBCR4849R1ZL         4,065
3  ABBBBBB070BB1ZB  B9BNOPL6798H1Z0        73,142
4  ABBBBBB070BB1ZB  ABBBBCS90ABR1Z0      8,21,440
```

Approach



Sector-wise Data

Pre processing & cleaning

Calculation of Pearson coefficients

Perform Clustering

Benford Analysis on clusters to classify suspicious clusters

Check for offences in the suspicious cluster

Activities Carried out

- Step 1: Identifying important sectors
  - Gold
  - Timber
  - Vehicle
  - Steel
  - Textiles

- Step 2: Identifying important parameter
  - Total Sales Amount
  - Total SGST liability
  - Total CGST liability
  - Total Liability
  - Total SGST paid in cash
  - Total Exempt sales
  - Total ITC
  - IGST ITC

Activities Carried out

- Step 3 : Extraction and cleaning of sector-wise data from July-2017 to Dec-2019 (30 months)
- Step 4 : Heat map analysis to identify parameter pairs with high overall correlation (for each sector)

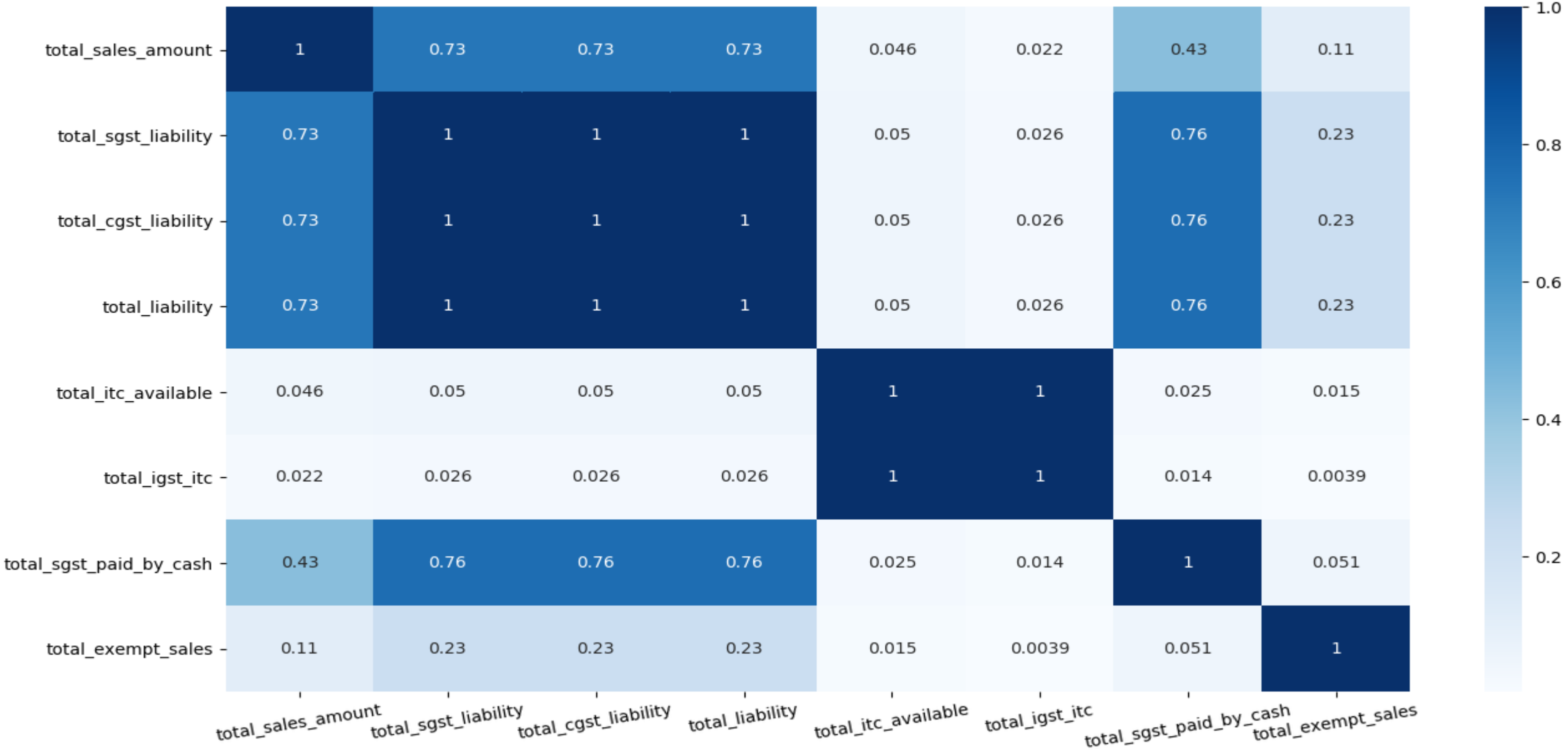
Activities Carried out

- Step 5 : Calculate Pearson's correlation coefficient between the pairs identified
  - Calculated for all taxpayers using period wise data for 30 months
  - The Pearson correlation coefficient measures the linear relationship between two datasets.
  - It varies between -1 and +1 with 0 implying no correlation. Correlations of -1 or +1 imply an exact linear relationship.
  - Positive correlations imply that as x increases, so does y. Negative correlations imply that as x increases, y decreases.

Activities Carried out

- Step 6 : Taxpayer clustering based on the correlation coefficients
  - DBSan
  - K-Mediods
  - K-Means

Feature Correlation Heatmap

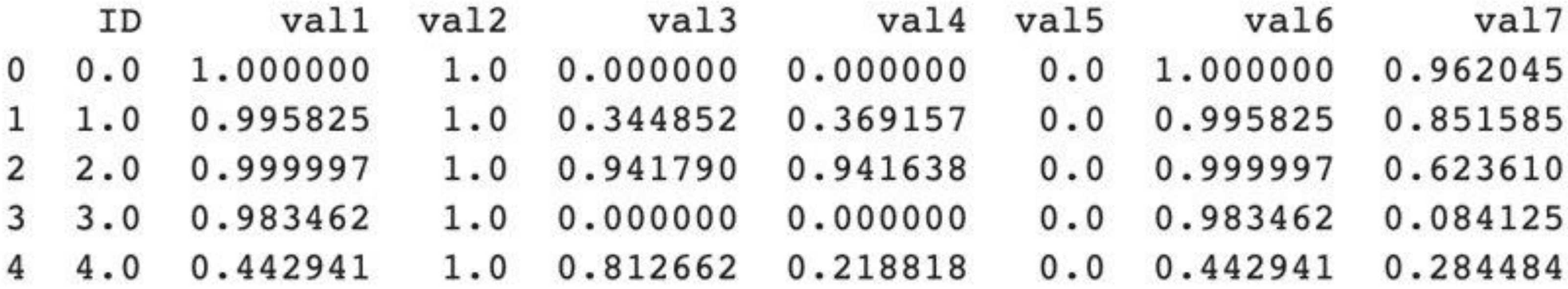


	total_sales_amount	total_sgst_liability	total_cgst_liability	total_liability	total_itc_available	total_igst_itc	total_sgst_paid_by_cash	total_exempt_sales
total_sales_amount	1	0.73	0.73	0.73	0.046	0.022	0.43	0.11
total_sgst_liability	0.73	1	1	1	0.05	0.026	0.76	0.23
total_cgst_liability	0.73	1	1	1	0.05	0.026	0.76	0.23
total_liability	0.73	1	1	1	0.05	0.026	0.76	0.23
total_itc_available	0.046	0.05	0.05	0.05	1	1	0.025	0.015
total_igst_itc	0.022	0.026	0.026	0.026	1	1	0.014	0.0039
total_sgst_paid_by_cash	0.43	0.76	0.76	0.76	0.025	0.014	1	0.051
total_exempt_sales	0.11	0.23	0.23	0.23	0.015	0.0039	0.051	1

Data used for clustering

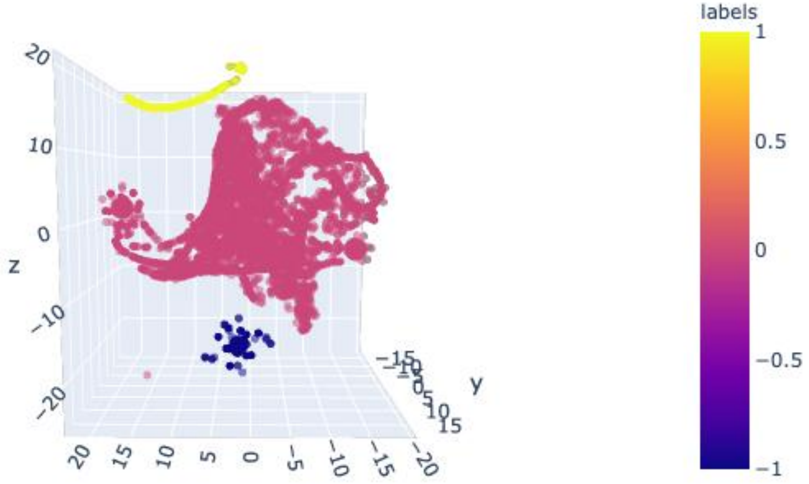
- Pearson coefficients were calculated form base data
  - Val1 = Total sales amount vs Total sales amount
  - Val2 = Total GST liability Vs. Total SGST liability
  - Val3 = Total SGST liability Vs. Total SGST paid in cash
  - Val4 = Total sales amount Vs. Total SGST paid in cash
  - Val5 = Total sales amount Vs. Total Exempt sales
  - Val6 = Total Liability Vs. Total ITC
  - Val7 = Total ITC Vs. IGST ITC

Sample Pearson Coefficient Data

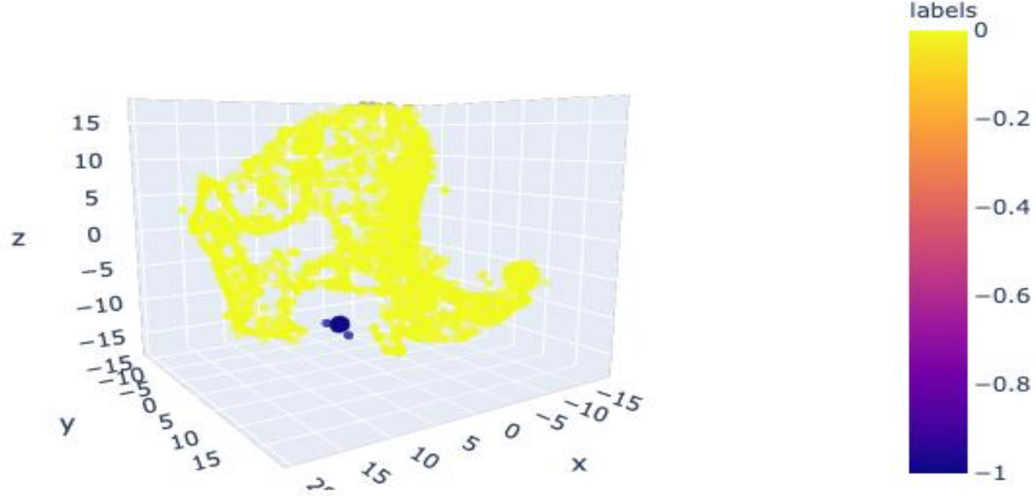


```
    ID      val1  val2      val3      val4  val5      val6      val7
0  0.0  1.000000   1.0  0.000000  0.000000   0.0  1.000000  0.962045
1  1.0  0.995825   1.0  0.344852  0.369157   0.0  0.995825  0.851585
2  2.0  0.999997   1.0  0.941790  0.941638   0.0  0.999997  0.623610
3  3.0  0.983462   1.0  0.000000  0.000000   0.0  0.983462  0.084125
4  4.0  0.442941   1.0  0.812662  0.218818   0.0  0.442941  0.284484
```

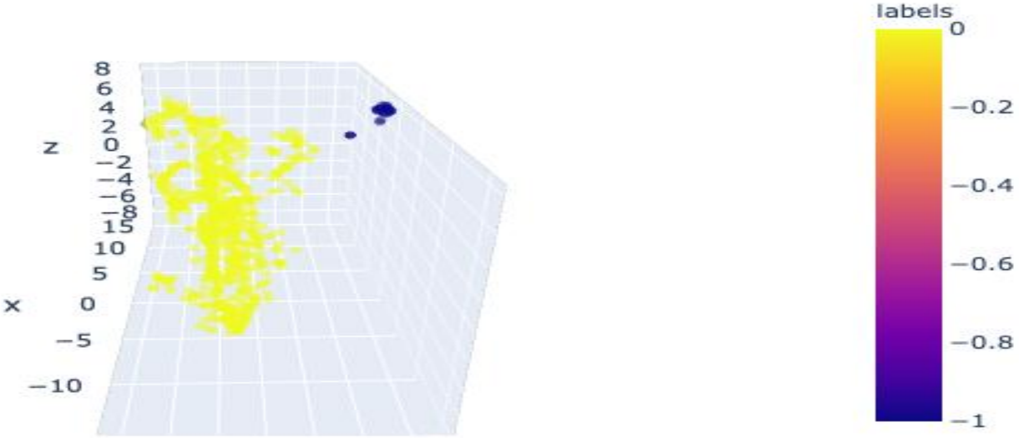
Cluster Analysis on Timber Sector data



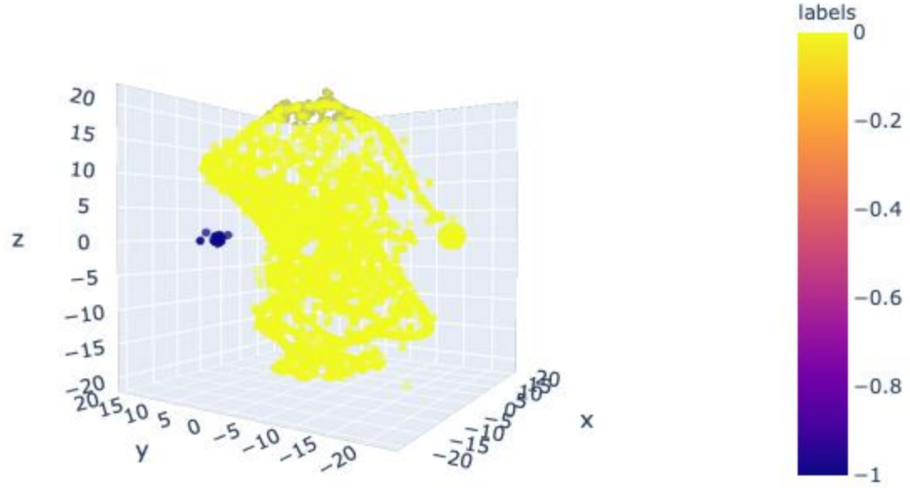
Cluster Analysis on Vehicle Sector data



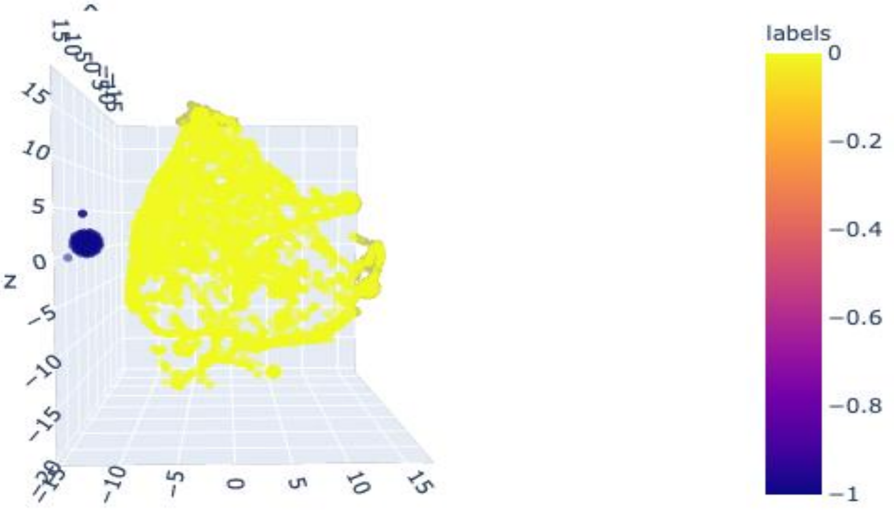
Cluster Analysis on Textile Sector data



Cluster Analysis on Steel Sector data



Cluster Analysis on Gold Sector data



Cluster Analysis Results

Gold Sector	Cluster 1	Cluster 2	Total
No. of observations	4677	78	4755

Steel Sector	Cluster 1	Cluster 2	Total
No. of observations	5434	51	5485

Vehicle Sector	Cluster 1	Cluster 2	Total
No. of observations	3920	50	3970

Textile Sector	Cluster 1	Cluster 2	Total
No. of observations	861	22	883

Timber Sector	Cluster 1	Cluster 2	Cluster 3	Total
No. of observations	5458	264	418	6140

Activities Carried out

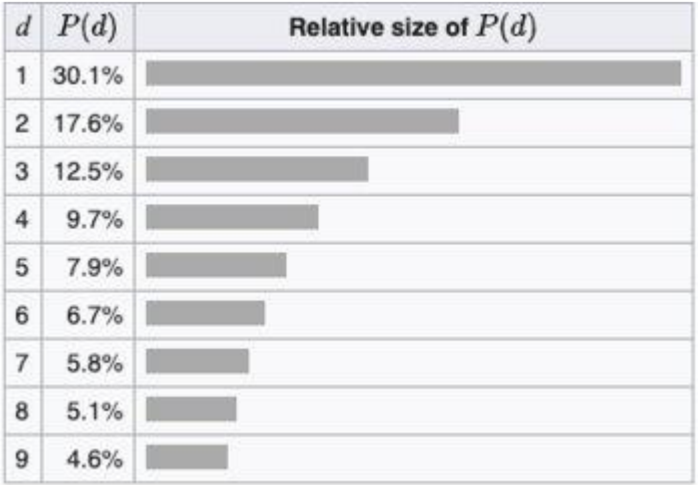
- Step 6 : Validation of clustering results
  - Silhoutte score was used as a metric for cluster validation
  - Estimated average distance between clusters
  - Silhouette Coefficient for a sample is (b - a) / max(a, b)

Clustering Algorithm		Timber	Gold	Steel	Vehicle	Textile
Silhouette score	DBScan	0.75	0.87	0.90	0.90	0.85
	K-Means	0.68	0.44	0.48	0.57	0.49
	K-Mediods	0.35	0.48	0.37	0.53	0.34

Activities Carried out

Step 6 : Benford Analysis

- Benford's law, also called the Newcomb–Benford law, is an observation about the frequency distribution of leading digits in many real-life sets of numerical data. The law states that in many naturally occurring collections of numbers, the leading significant digit is likely to be small.



d	$P(d)$	Relative size of $P(d)$
1	30.1%	
2	17.6%	
3	12.5%	
4	9.7%	
5	7.9%	
6	6.7%	
7	5.8%	
8	5.1%	
9	4.6%	

Benford Analysis

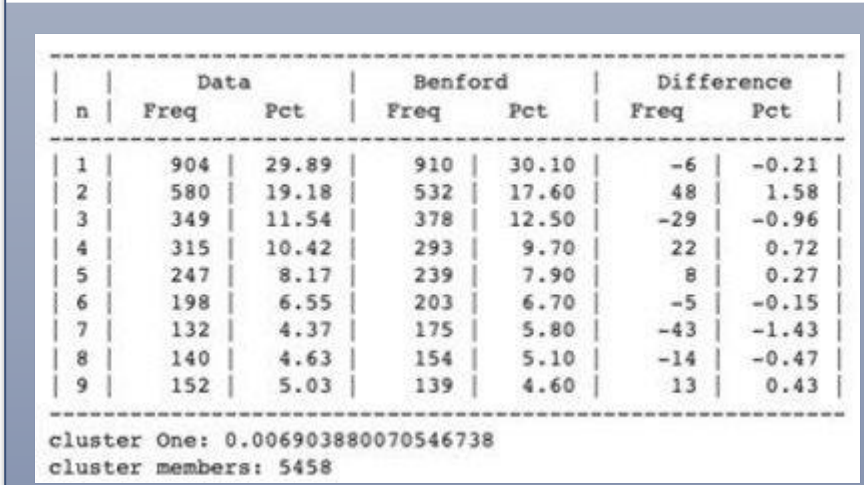
- Mean Absolute Deviation (MAD) is the measure we used to evalute the conformity to Benford’s Law.

$$\text{Mean Absolute Deviation} = \frac{\sum_{i=1}^{K} |AP - EP|}{K}$$

- According to research values greater than 0.012 are said to be non- conformal to Benfords law.

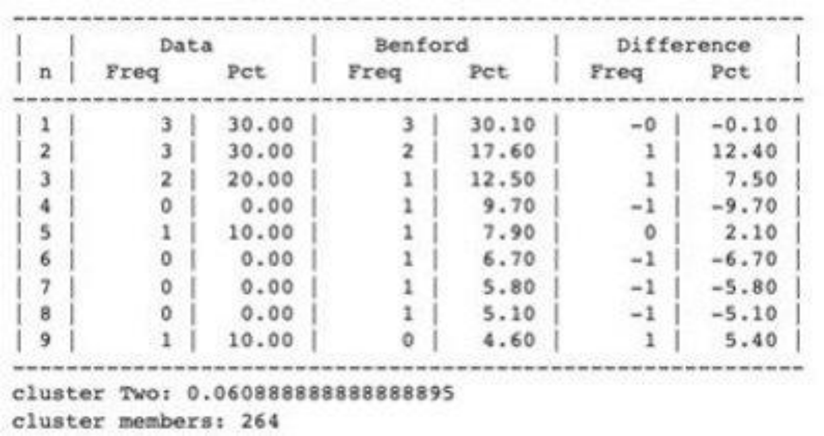
Benford Analysis – Timber Sector

Cluster-1: 5458 (Conformant to Benford)



```
--------------------------------------------------------------------
|   |        Data        |      Benford       |     Difference     |
| n |  Freq       Pct    |  Freq       Pct    |  Freq       Pct    |
--------------------------------------------------------------------
| 1 |     904 |   29.89 |     910 |   30.10 |        -6 |   -0.21 |
| 2 |     580 |   19.18 |     532 |   17.60 |        48 |    1.58 |
| 3 |     349 |   11.54 |     378 |   12.50 |       -29 |   -0.96 |
| 4 |     315 |   10.42 |     293 |    9.70 |        22 |    0.72 |
| 5 |     247 |    8.17 |     239 |    7.90 |         8 |    0.27 |
| 6 |     198 |    6.55 |     203 |    6.70 |        -5 |   -0.15 |
| 7 |     132 |    4.37 |     175 |    5.80 |       -43 |   -1.43 |
| 8 |     140 |    4.63 |     154 |    5.10 |       -14 |   -0.47 |
| 9 |     152 |    5.03 |     139 |    4.60 |        13 |    0.43 |
--------------------------------------------------------------------
cluster One: 0.006903880070546738
cluster members: 5458
```

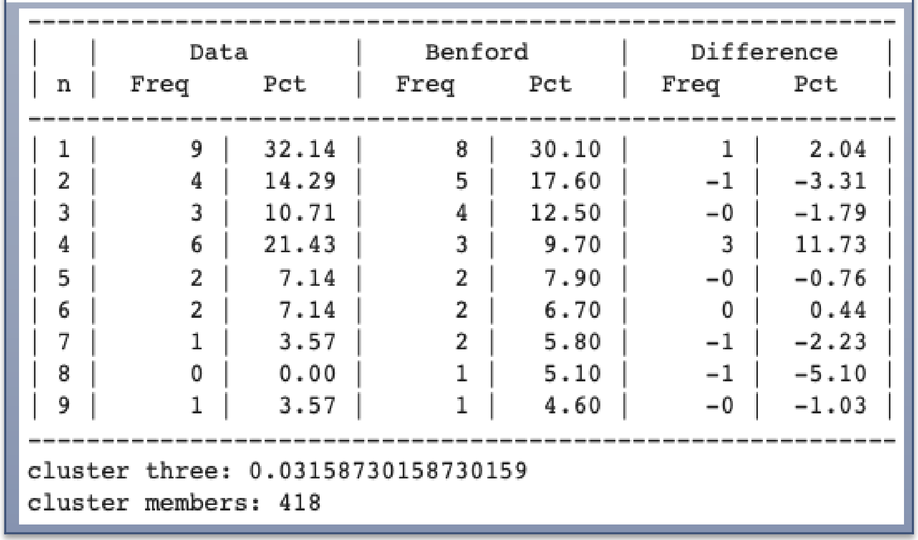
Cluster-2: 264 (Suspicious)



```
--------------------------------------------------------------------
|   |        Data        |      Benford       |     Difference     |
| n |  Freq       Pct    |  Freq       Pct    |  Freq       Pct    |
--------------------------------------------------------------------
| 1 |          3 |  30.00 |       3 |  30.10 |        -0 |   -0.10 |
| 2 |          3 |  30.00 |       2 |  17.60 |         1 |   12.40 |
| 3 |          2 |  20.00 |       1 |  12.50 |         1 |    7.50 |
| 4 |          0 |   0.00 |       1 |   9.70 |        -1 |   -9.70 |
| 5 |          1 |  10.00 |       1 |   7.90 |         0 |    2.10 |
| 6 |          0 |   0.00 |       1 |   6.70 |        -1 |   -6.70 |
| 7 |          0 |   0.00 |       1 |   5.80 |        -1 |   -5.80 |
| 8 |          0 |   0.00 |       1 |   5.10 |        -1 |   -5.10 |
| 9 |          1 |  10.00 |       0 |   4.60 |         1 |    5.40 |
--------------------------------------------------------------------
cluster Two: 0.06088888888888895
cluster members: 264
```

Benford Analysis – Timber Sector

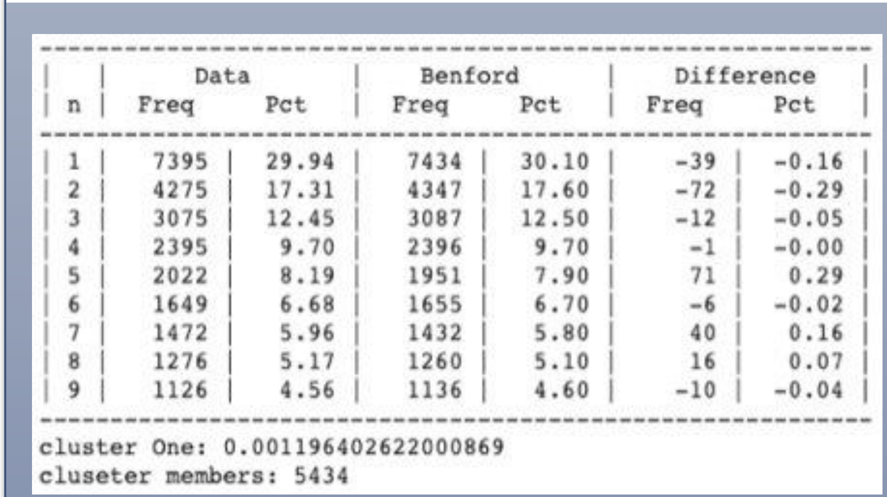
Cluster-2: 418(Suspicious)



```
-------------------------------------------------------------
|   |        Data        |       Benford      |     Difference     |
| n |  Freq       Pct    |  Freq       Pct    |  Freq       Pct    |
-------------------------------------------------------------
| 1 |        9 |  32.14 |        8 |  30.10 |        1 |   2.04 |
| 2 |        4 |  14.29 |        5 |  17.60 |       -1 |  -3.31 |
| 3 |        3 |  10.71 |        4 |  12.50 |       -0 |  -1.79 |
| 4 |        6 |  21.43 |        3 |   9.70 |        3 |  11.73 |
| 5 |        2 |   7.14 |        2 |   7.90 |       -0 |  -0.76 |
| 6 |        2 |   7.14 |        2 |   6.70 |        0 |   0.44 |
| 7 |        1 |   3.57 |        2 |   5.80 |       -1 |  -2.23 |
| 8 |        0 |   0.00 |        1 |   5.10 |       -1 |  -5.10 |
| 9 |        1 |   3.57 |        1 |   4.60 |       -0 |  -1.03 |
-------------------------------------------------------------
cluster three: 0.03158730158730159
cluster members: 418
```

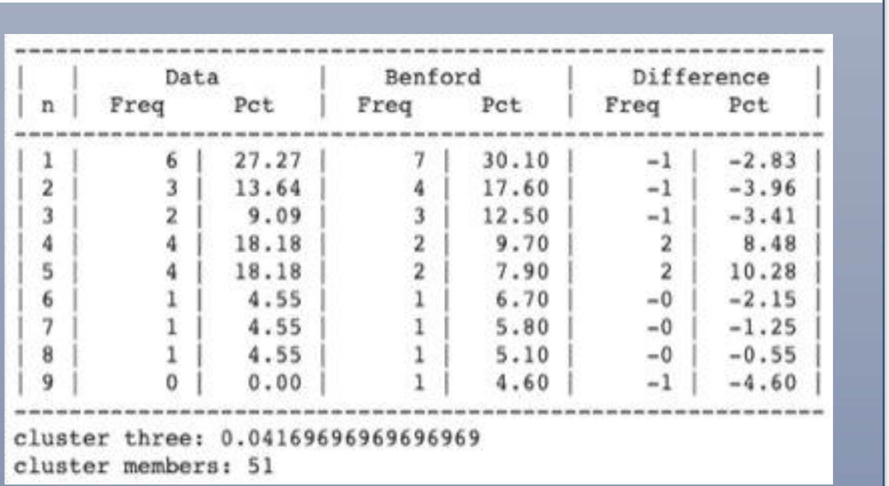
Benford Analysis – Steel Sector

Cluster-1: 5434 (Conformant to Benford)



```
--------------------------------------------------------------
|   |       Data         |      Benford       |    Difference      |
| n |  Freq        Pct   |  Freq        Pct   |  Freq        Pct   |
--------------------------------------------------------------
| 1 |    7395 |   29.94 |    7434 |   30.10 |      -39 |   -0.16 |
| 2 |    4275 |   17.31 |    4347 |   17.60 |      -72 |   -0.29 |
| 3 |    3075 |   12.45 |    3087 |   12.50 |      -12 |   -0.05 |
| 4 |    2395 |    9.70 |    2396 |    9.70 |       -1 |   -0.00 |
| 5 |    2022 |    8.19 |    1951 |    7.90 |       71 |    0.29 |
| 6 |    1649 |    6.68 |    1655 |    6.70 |       -6 |   -0.02 |
| 7 |    1472 |    5.96 |    1432 |    5.80 |       40 |    0.16 |
| 8 |    1276 |    5.17 |    1260 |    5.10 |       16 |    0.07 |
| 9 |    1126 |    4.56 |    1136 |    4.60 |      -10 |   -0.04 |
--------------------------------------------------------------
cluster One: 0.001196402622000869
cluseter members: 5434
```

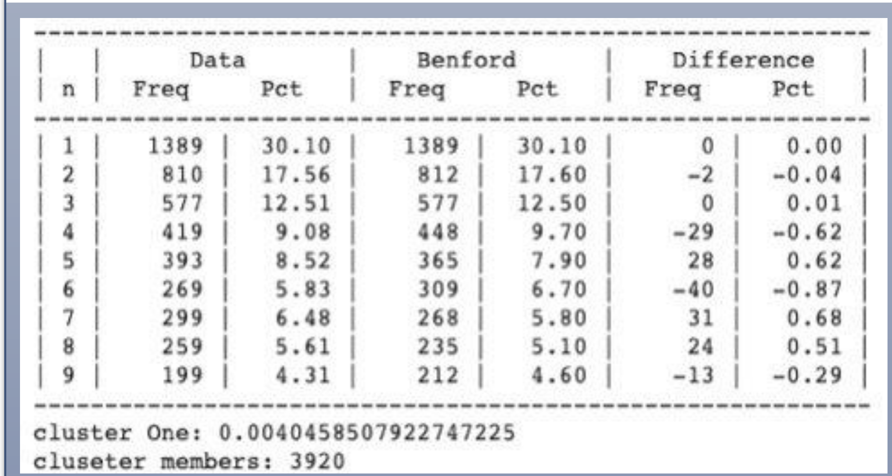
Cluster-2: 51 (Suspicious)



```
--------------------------------------------------------------
|   |       Data         |      Benford       |    Difference      |
| n |  Freq        Pct   |  Freq        Pct   |  Freq        Pct   |
--------------------------------------------------------------
| 1 |       6 |   27.27 |       7 |   30.10 |       -1 |   -2.83 |
| 2 |       3 |   13.64 |       4 |   17.60 |       -1 |   -3.96 |
| 3 |       2 |    9.09 |       3 |   12.50 |       -1 |   -3.41 |
| 4 |       4 |   18.18 |       2 |    9.70 |        2 |    8.48 |
| 5 |       4 |   18.18 |       2 |    7.90 |        2 |   10.28 |
| 6 |       1 |    4.55 |       1 |    6.70 |       -0 |   -2.15 |
| 7 |       1 |    4.55 |       1 |    5.80 |       -0 |   -1.25 |
| 8 |       1 |    4.55 |       1 |    5.10 |       -0 |   -0.55 |
| 9 |       0 |    0.00 |       1 |    4.60 |       -1 |   -4.60 |
--------------------------------------------------------------
cluster three: 0.04169696969696969
cluster members: 51
```

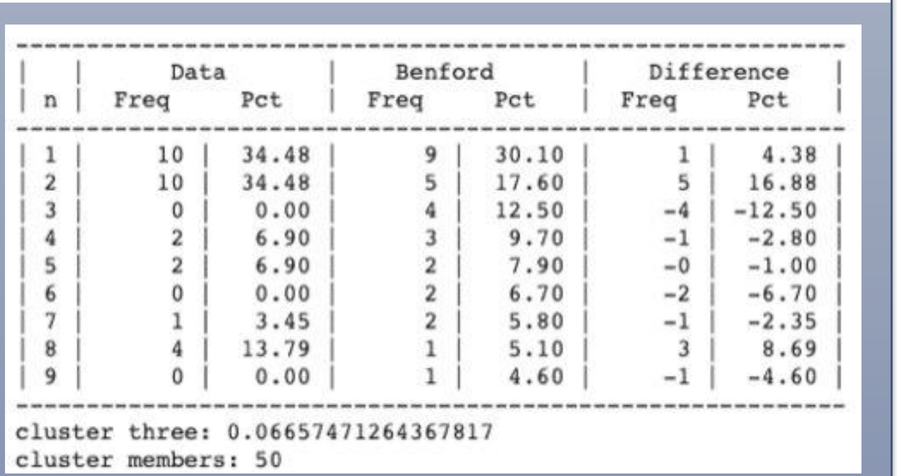
Benford Analysis – Vehicle Sector

Cluster-1: 3920 (Conformant to Benford)



```
---------------------------------------------------------------
|   |       Data        |      Benford      |    Difference    |
| n |  Freq        Pct  |  Freq        Pct  |  Freq       Pct  |
---------------------------------------------------------------
| 1 |   1389 |   30.10  |   1389 |   30.10  |     0 |    0.00  |
| 2 |    810 |   17.56  |    812 |   17.60  |    -2 |   -0.04  |
| 3 |    577 |   12.51  |    577 |   12.50  |     0 |    0.01  |
| 4 |    419 |    9.08  |    448 |    9.70  |   -29 |   -0.62  |
| 5 |    393 |    8.52  |    365 |    7.90  |    28 |    0.62  |
| 6 |    269 |    5.83  |    309 |    6.70  |   -40 |   -0.87  |
| 7 |    299 |    6.48  |    268 |    5.80  |    31 |    0.68  |
| 8 |    259 |    5.61  |    235 |    5.10  |    24 |    0.51  |
| 9 |    199 |    4.31  |    212 |    4.60  |   -13 |   -0.29  |
---------------------------------------------------------------
cluster One: 0.0040458507922747225
cluseter members: 3920
```

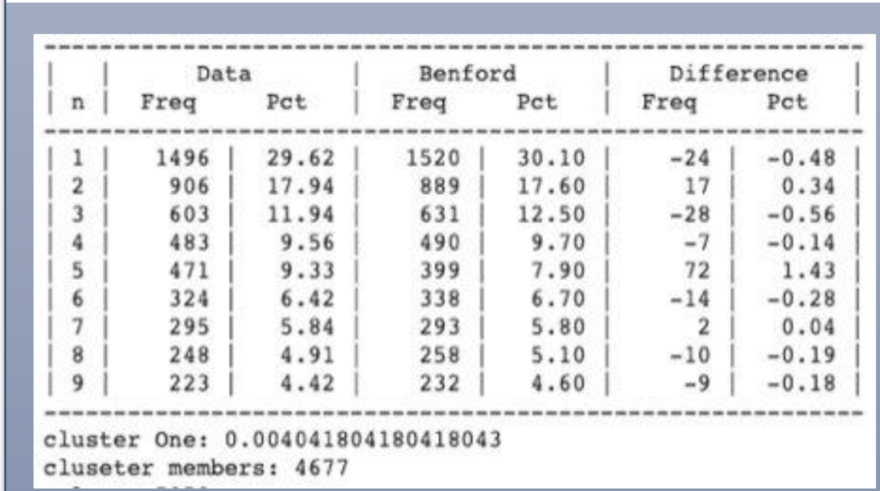
Cluster-2: 50 (Suspicious)



```
---------------------------------------------------------------
|   |       Data        |      Benford      |    Difference    |
| n |  Freq        Pct  |  Freq        Pct  |  Freq       Pct  |
---------------------------------------------------------------
| 1 |     10 |   34.48  |      9 |   30.10  |     1 |    4.38  |
| 2 |     10 |   34.48  |      5 |   17.60  |     5 |   16.88  |
| 3 |      0 |    0.00  |      4 |   12.50  |    -4 |  -12.50  |
| 4 |      2 |    6.90  |      3 |    9.70  |    -1 |   -2.80  |
| 5 |      2 |    6.90  |      2 |    7.90  |    -0 |   -1.00  |
| 6 |      0 |    0.00  |      2 |    6.70  |    -2 |   -6.70  |
| 7 |      1 |    3.45  |      2 |    5.80  |    -1 |   -2.35  |
| 8 |      4 |   13.79  |      1 |    5.10  |     3 |    8.69  |
| 9 |      0 |    0.00  |      1 |    4.60  |    -1 |   -4.60  |
---------------------------------------------------------------
cluster three: 0.06657471264367817
cluster members: 50
```

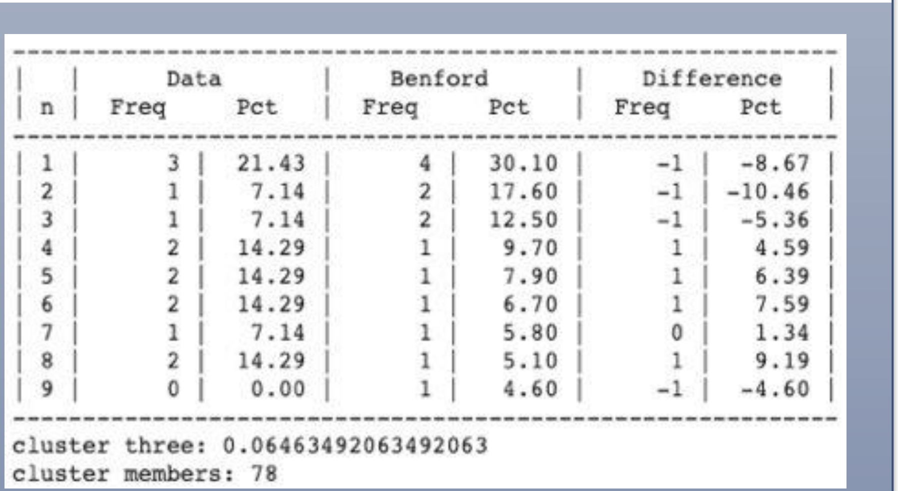
Benford Analysis – Gold Sector

Cluster-1: 4677 (Conformant to Benford)



```
----------------------------------------------------------------------
|   |        Data          |        Benford        |     Difference    |
| n |   Freq        Pct    |   Freq        Pct     |   Freq       Pct  |
----------------------------------------------------------------------
| 1 |     1496 |   29.62 |     1520 |   30.10 |      -24 |   -0.48 |
| 2 |      906 |   17.94 |      889 |   17.60 |       17 |    0.34 |
| 3 |      603 |   11.94 |      631 |   12.50 |      -28 |   -0.56 |
| 4 |      483 |    9.56 |      490 |    9.70 |       -7 |   -0.14 |
| 5 |      471 |    9.33 |      399 |    7.90 |       72 |    1.43 |
| 6 |      324 |    6.42 |      338 |    6.70 |      -14 |   -0.28 |
| 7 |      295 |    5.84 |      293 |    5.80 |        2 |    0.04 |
| 8 |      248 |    4.91 |      258 |    5.10 |      -10 |   -0.19 |
| 9 |      223 |    4.42 |      232 |    4.60 |       -9 |   -0.18 |
----------------------------------------------------------------------
cluster One: 0.004041804180418043
cluseter members: 4677
```

Cluster-2: 78 (Suspicious)

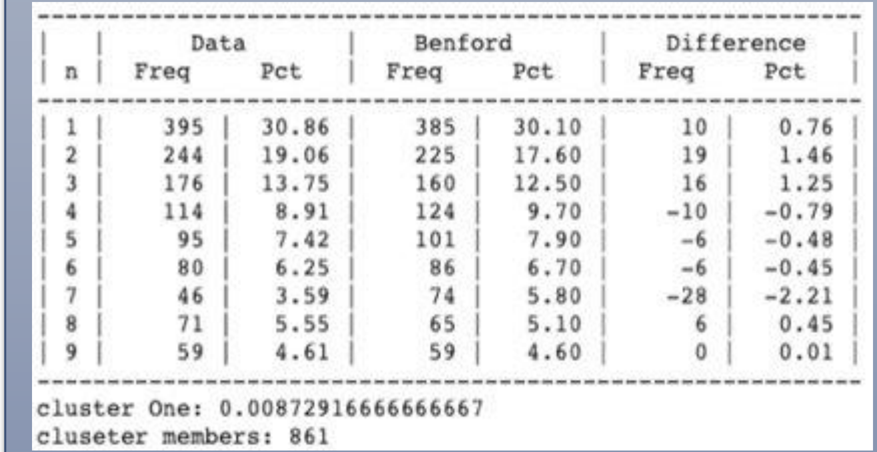


```
----------------------------------------------------------------------
|   |        Data          |        Benford        |     Difference    |
| n |   Freq        Pct    |   Freq        Pct     |   Freq       Pct  |
----------------------------------------------------------------------
| 1 |        3 |   21.43 |        4 |   30.10 |       -1 |   -8.67 |
| 2 |        1 |    7.14 |        2 |   17.60 |       -1 |  -10.46 |
| 3 |        1 |    7.14 |        2 |   12.50 |       -1 |   -5.36 |
| 4 |        2 |   14.29 |        1 |    9.70 |        1 |    4.59 |
| 5 |        2 |   14.29 |        1 |    7.90 |        1 |    6.39 |
| 6 |        2 |   14.29 |        1 |    6.70 |        1 |    7.59 |
| 7 |        1 |    7.14 |        1 |    5.80 |        0 |    1.34 |
| 8 |        2 |   14.29 |        1 |    5.10 |        1 |    9.19 |
| 9 |        0 |    0.00 |        1 |    4.60 |       -1 |   -4.60 |
----------------------------------------------------------------------
cluster three: 0.06463492063492063
cluster members: 78
```

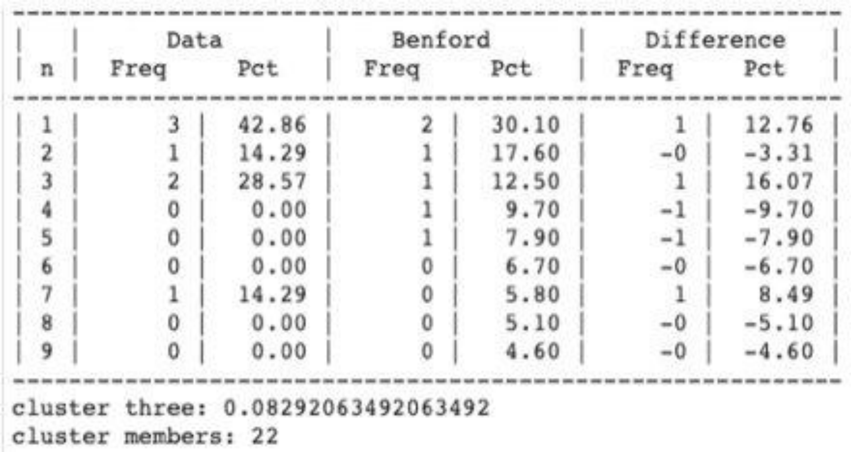
Benford Analysis – Textile Sector

Cluster-1: 861 (Conformant to Benford)

Cluster-2: 22 (Suspicious)



```
------------------------------------------------------------
|   |       Data       |      Benford      |    Difference    |
| n |  Freq      Pct   |  Freq      Pct    |  Freq      Pct   |
------------------------------------------------------------
| 1 |    395 |  30.86 |    385 |  30.10 |     10 |   0.76 |
| 2 |    244 |  19.06 |    225 |  17.60 |     19 |   1.46 |
| 3 |    176 |  13.75 |    160 |  12.50 |     16 |   1.25 |
| 4 |    114 |   8.91 |    124 |   9.70 |    -10 |  -0.79 |
| 5 |     95 |   7.42 |    101 |   7.90 |     -6 |  -0.48 |
| 6 |     80 |   6.25 |     86 |   6.70 |     -6 |  -0.45 |
| 7 |     46 |   3.59 |     74 |   5.80 |    -28 |  -2.21 |
| 8 |     71 |   5.55 |     65 |   5.10 |      6 |   0.45 |
| 9 |     59 |   4.61 |     59 |   4.60 |      0 |   0.01 |
------------------------------------------------------------
cluster One: 0.008729166666666667
cluseter members: 861
```



```
------------------------------------------------------------
|   |       Data       |      Benford      |    Difference    |
| n |  Freq      Pct   |  Freq      Pct    |  Freq      Pct   |
------------------------------------------------------------
| 1 |      3 |  42.86 |      2 |  30.10 |      1 |  12.76 |
| 2 |      1 |  14.29 |      1 |  17.60 |     -0 |  -3.31 |
| 3 |      2 |  28.57 |      1 |  12.50 |      1 |  16.07 |
| 4 |      0 |   0.00 |      1 |   9.70 |     -1 |  -9.70 |
| 5 |      0 |   0.00 |      1 |   7.90 |     -1 |  -7.90 |
| 6 |      0 |   0.00 |      0 |   6.70 |     -0 |  -6.70 |
| 7 |      1 |  14.29 |      0 |   5.80 |      1 |   8.49 |
| 8 |      0 |   0.00 |      0 |   5.10 |     -0 |  -5.10 |
| 9 |      0 |   0.00 |      0 |   4.60 |     -0 |  -4.60 |
------------------------------------------------------------
cluster three: 0.08292063492063492
cluster members: 22
```

Usability of Model

- Sector wise suspicious dealer list shall be provided to department
- Prioritization of assessment based on results
- Tuning the model to improve accuracy based on feedback
- Development and integration of the AI module in GST backend for future analysis

Thank You