Data structure

* The data contains 27 columns (at **leaves level**)
* Around 6500 rows
* They could be grouped into 16 columns (at **level 2**)
* ~~And also could be grouped into 4 columns (at level 1)~~

Data Values –

**[ LEAVES LEVEL ]**

* The data has values between 0 and 1 with **a step of 0.1**
* However, the majority of the cells are of **value 0**

|  |  |  |  |
| --- | --- | --- | --- |
| **= 0** | = 0.1 | = 0.2 | >= 0.3 |
| **85%** | 7% | 3% | 5% |

* Column wise; only 7 columns contain **6%** to **25%** of the total value of the data (Rest are less than **3%**)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | precedence | reason | result | arg2-as-denier | conjunction | arg2-as-instance | arg2-as-detail |
| Value/total value | 7% | 6% | 22% | 6% | 23% | 6% | 16% |
| Count/total count countif > 0 | 1.1% | 1.1% | 2.5% | 1.0% | 2.8% | 1.1% | 2.2% |

**[ LEVEL TWO ]**

* A little better, butthe majority of the cells are of **value 0**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **= 0** | = 0.1 | = 0.2 | = 0.3 | >= 0,4 |
| **78%** | 9% | 5% | 3% | 5% |

* Column wise; only **6** columns contain **6%** to **28%** of the total value of the data (Rest are less than **3%**)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | synchronous | asynchronous | cause | condition | negative-condition | purpose | concession | contrast | similarity | conjunction | disjunction | instantiation | level-of-detail | equivalence | manner | exception | substitution |
| Value/total value | 1% | **8%** | **28%** | 0% | 0% | 0% | **8%** | 3% | 2% | **23%** | 0% | **6%** | **18%** | 0% | 0% | 0% | 1% |
| Count/total count **countif > 0** | 1% | **2%** | **5%** | 0% | 0% | 0% | 2% | 1% | 1% | **4%** | 0% | 2% | 4% | 0% | 0% | 0% | 0% |

1. Is the data **compositional data**?
   1. Originally each row sums up 1
   2. We removed the last two columns for a reason and not not all the rows sum up to 1
      1. Discogem:

|  |  |
| --- | --- |
| **# of entities** (out of 6505) | **Sum of the row** |
| 3,871 (60%) | = 1 |
| 1,547 (24%) | = 0.9 |
| 1,087 (16%) | <= 0.8 |

* + 1. Quadc

|  |  |
| --- | --- |
| **# of entities** (out of 900) | **Sum of the row** |
| 830 (92%) | = 1 |
| 26 (3%) | = 0.9 |
| 44 (5%) | <= 0.8 |

Note: the sum of all the columns in each row adds up to 1 (or 0.9, 0.8 ..), and we are trying to find the correlation between a pair of columns at a time

1. If transformation is needed, are any of the three below suitable for the transformation?
   1. Additive logratio transform
   2. Isometric logratio transform
   3. Center log ratio transform
2. After transformation, what is the best correlation coefficient to use

['pearson', 'spearman', 'kendall']

|  |  |
| --- | --- |
| 'pearson' | * Assumption: Assumes a linear relationship and that data is normally distributed. * Use Case: Suitable for continuous data when you want to measure linear associations. |
| 'spearman' | * Assumption: Non-parametric and does not assume a linear relationship but assumes a monotonic relationship. * Use Case: Appropriate for both continuous and ordinal data. Particularly useful when the relationship is expected to be monotonic but not necessarily linear. |
| 'kendall' | * Assumption: Non-parametric and makes no assumptions about the data distribution. * Use Case: Suitable for both continuous and ordinal (ranked) data. Useful when the data may not follow a linear relationship. |

[source](Assumption:%20Non-parametric%20and%20makes%20no%20assumptions%20about%20the%20data%20distribution.%20Use%20Case:%20Suitable%20for%20both%20continuous%20and%20ordinal%20(ranked)%20data.%20Useful%20when%20the%20data%20may%20not%20follow%20a%20linear%20relationship.)

1. It is safe to say that our data are not liner? Examples

Plot the data

A quick understanding of pearson and what not

Check if they are normal distributes

Then go to the bindar