**Data analysis**

To make the best decisions regarding the data given, it is important to look at and understanding the data. All work completed in this chapter will be documented in 1 - data analysis notebook. The data given contains merger trees in matrix format of three variables. The matrices have 29 rows and 10 columns, corresponding to 29 timesteps and 10 possible branches (one main branch and nine possible subbranches). But not all merger trees have 10 branches, in fact, the data contains merger trees with a total of five to ten branches. Here are a table of the number of branches:

|  |  |
| --- | --- |
| **Total number of branches** | **Count** |
| 5 | 7679 |
| 6 | 8161 |
| 7 | 7311 |
| 8 | 6187 |
| 9 | 4916 |
| 10 | 4094 |
| total | 38348 |

[3] states that around 1000 merger trees per given number of branches are required for the GAN model in [3] to converge and obtain decent generated merger trees, which implies that it should be possible to reconstruct the results with all merger trees with five to ten branches.

Another interesting aspect of the merger trees with different number of branches is whether they “behave” similarly or not in terms of length of branches.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Total number of branches** | **Average main branch length** | **Subbranch minimum length** | **Subbranch maximum length** | **Subbranch average length** | **Subbranch standard deviation** |
| 5 | 24.23 | 1 | 26 | 6.64 | 5.22 |
| 6 | 24.13 | 1 | 26 | 6.65 | 5.25 |
| 7 | 24.13 | 1 | 26 | 6.57 | 5.23 |
| 8 | 24.20 | 1 | 26 | 6.51 | 5.25 |
| 9 | 24.39 | 1 | 26 | 6.54 | 5.31 |
| 10 | 24.55 | 1 | 26 | 6.61 | 5.37 |
| **Total** |  |  |  |  |  |
|  | 24.24 | 1 | 26 | 6.58 | 5.27 |

All merger trees with different number of branches have the same behavior in terms of length of branches, they are almost inseparable, the only difference is the total number of branches. There is no real difference in the distribution of branch length within each merger tree category. Here is the distribution length of branches of all merger trees, and by looking at the same plots for each number of branch category, they have an almost identical distribution (see 1 - data analysis notebook). Therefore, we can conclude that there are no difference in branch length behavior depending on number of branches in a merger tree, the only difference is the total number of branches in the merger tree.

A graph of a bar

Description automatically generated

**Variables**

* **distance**

**A graph of a graph

Description automatically generated**To handle variables in the best way possible, it is important to understand them as well. The first variable is distance. The distance values in all trees ranges from 0.0 to 1.775 with the following distribution. The absolute majority of the distance values are centered at the lower end, close to 0.0. The mean of the nonzero distance values is 0.0657, more than 95% of all nonzero distance values is between 0.0 and 0.145 and 99.17% of all nonzero distance values are in the range 0.0 – 0.2. It is extremely interesting, that only 301 out of 1 518 348 nonzero distance values are above 0.5. This finding introduces the possibility of different normalization methods for the distance variable, since one of the main generative features we are interested in is the structure of the merger trees. And to capture the structure, the models must be able to distinguish zero entries (no halo) with nonzero entries (halo present). Having a variable that is centered so close to zero might make it hard for the model to predict the structure correctly, or at least the structure within the distance variable.

* **Mass**

The second variable, mass, might be the most important variable in merger trees, since the mass assembly history of a galaxy formation merger tree is extremely important [2]. The mass variable has a minimum nonzero value of 7.6377 and a maximum nonzero value of 10.9763 and a mean nonzero value of 9.1345. The distribution of the values have two peaks at around 8.6 and 10.0 and according to standard quartile outlier theorem, there are no outliers.

**A graph of a graph

Description automatically generated**

* **Subhalo**

A blue rectangular graph with white text

Description automatically generatedThe last variable is the subhalo variable, which is a categorical variable. There is not much to analyze for this variable except checking the ratio between main haloes (1.0) and subhaloes (0.5). 77% of all progenitors are main haloes, which is as expected.

**Variables in diff number of branches merger trees:**

In terms of the differences of minimum, maximum and mean value of all three variables within merger trees with different number of branches, there are no peculiar patterns to mention.