MURDOCH UNIVERSITY

ICT616 Data Resources Management

Semester 2, 2020

ASSIGNMENT 2

Question 1.B

For this project the US states covid-19 data set is considered. In order to analyse the dataset, at first the data set is checked for any kind of inconsistencies such as missing values. When checked using python code, it is found that there are no missing values.

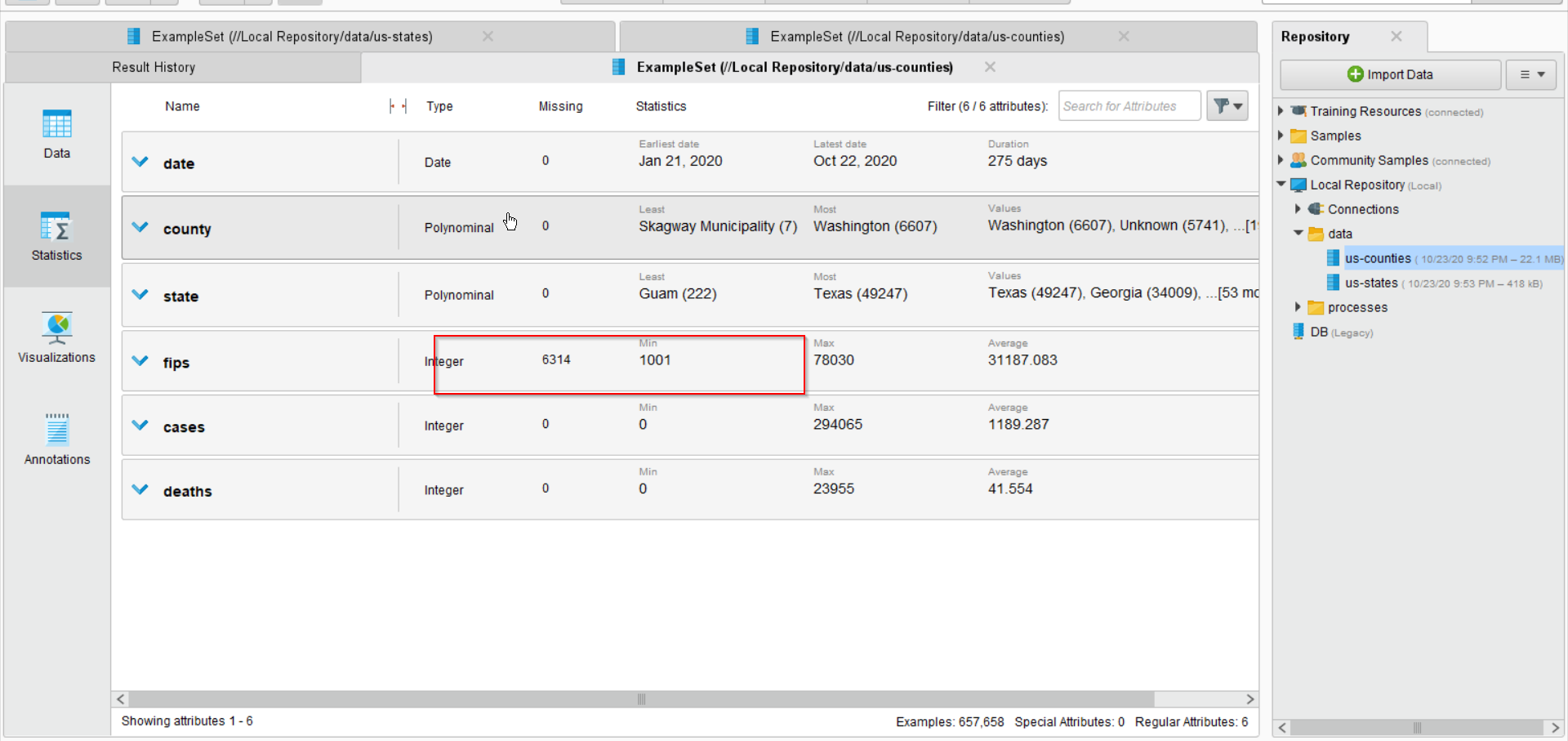
For preparing the data different tools are used in created workspace inside the RapidMiner workspace .

Following table shows the descriptive statistics for the dataset,

|  |  |  |  |
| --- | --- | --- | --- |
|  | fips | cases | deaths |
| count | 12279 | 12279 | 12279 |
| mean | 31.85577 | 56404.21 | 2033.969 |
| std | 18.61078 | 110606 | 4500.918 |
| min | 1 | 1 | 0 |
| 25% | 17 | 1738.5 | 42 |
| 50% | 31 | 14755 | 418 |
| 75% | 46 | 62707 | 1904.5 |
| max | 78 | 857076 | 32879 |

For the us-counties dataset, following is the descriptive statistics;

|  |  |  |  |
| --- | --- | --- | --- |
|  | fips | cases | deaths |
| count | 615984 | 621964 | 621964 |
| mean | 31170.61 | 1113.573 | 40.07909 |
| std | 16266.08 | 6650.846 | 415.1296 |
| min | 1001 | 0 | 0 |
| 25% | 18169 | 18 | 0 |
| 50% | 29195 | 97 | 2 |
| 75% | 46091 | 462 | 11 |
| max | 78030 | 282135 | 23883 |



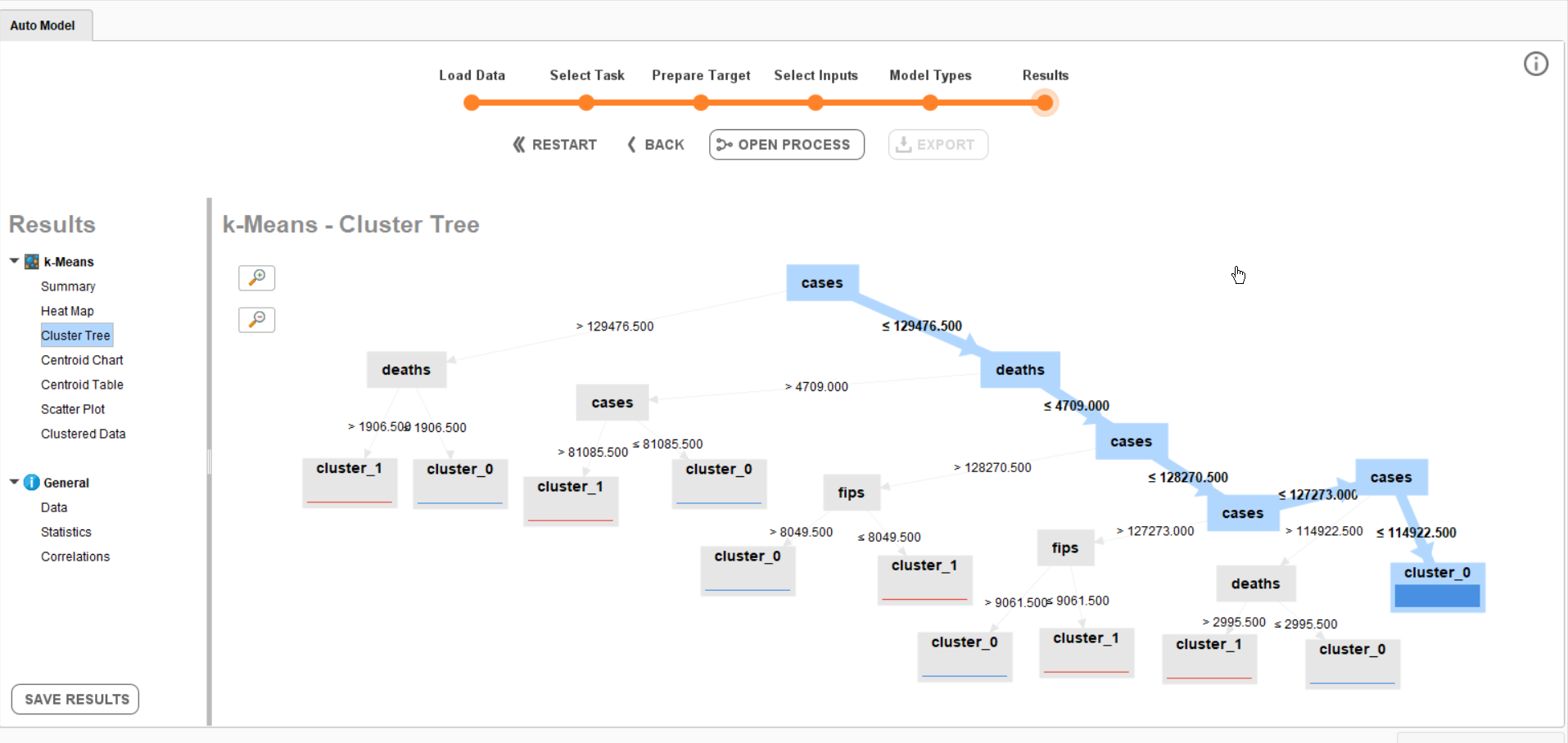
When explored in the data statistics tab it is found that there are no mission values in us-states dataset but in case of the us-counties data set, it is found that there are 6314 missing values for the column of the fips.

In order to cleanse and prepare the dataset, the missing values are replaced with the average value of that column using the “replace missing value operator.” The operator can also replace the missing values by minimum, maximum, zeroes or with the user specified values.

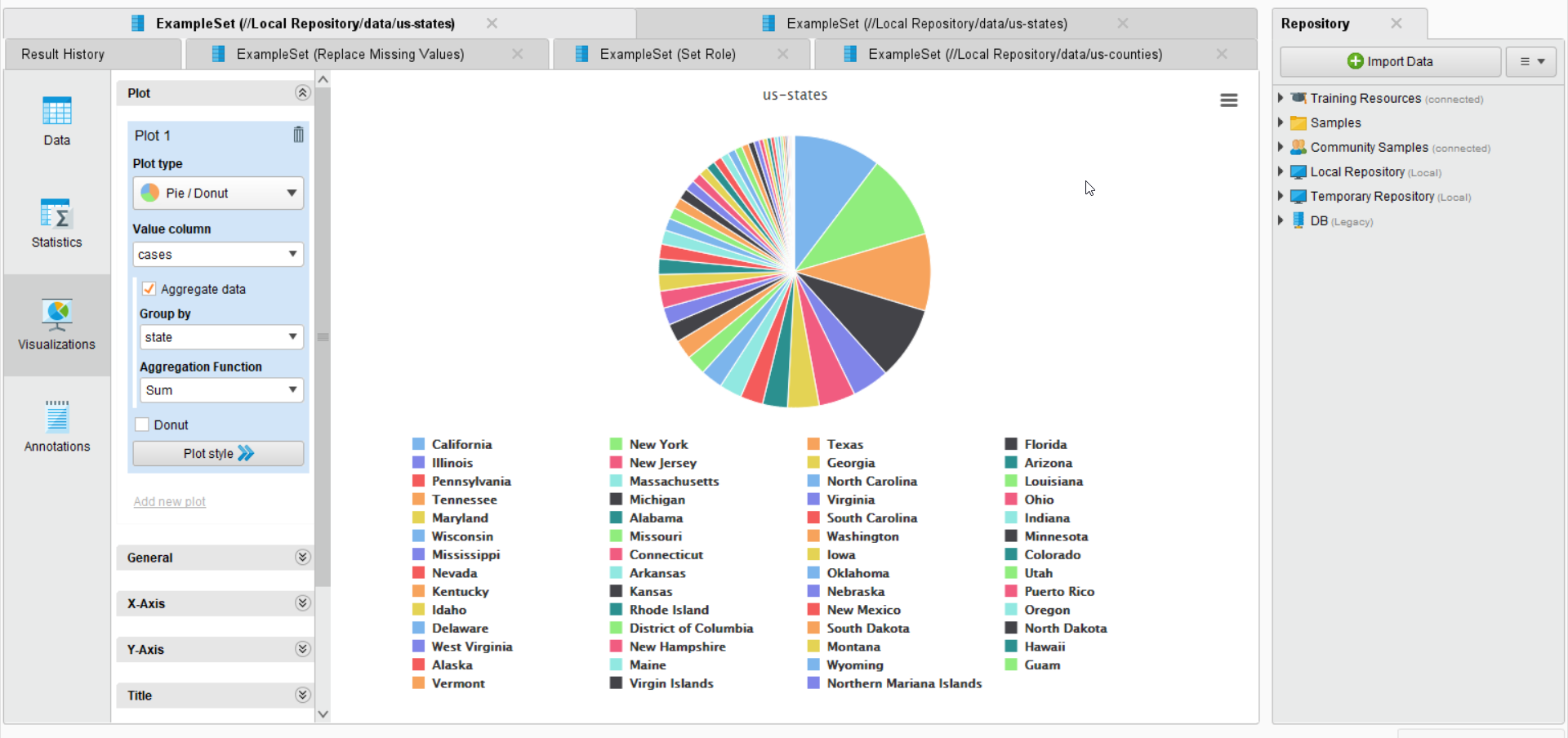
For the recorded cases in the different states, the following plot is created.

Here, as expected New York had the highest number of cases. But surprisingly the cases in California, Texas and Florida is comparatively high too.

Following cluster tree is generates using the dataset;



Highest number of cases are recorded for the California city which can be seen in the following visualization;



Highest number of deaths are recorded for New York and lowest is for Northern Mariana Islands.

