# 5810 Final Project Report

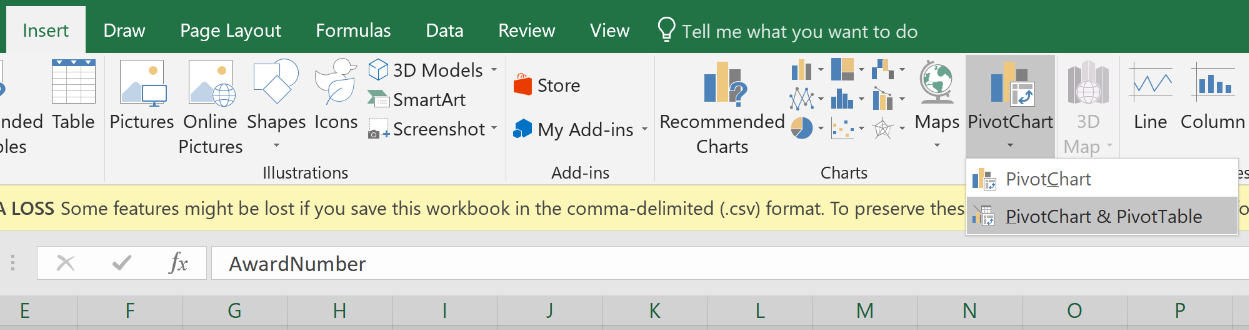
The Final Project involves the application of some data analysis techniques such as text mining and clustering analysis etc., to retrieve some useful information and these analyses are done on the provided dataset which was initially obtained from the National Science Foundation (NSF). It has the data about the various awards of NSF on Smart and Connected Health. The columns of this dataset have information about Award Number, Title of the Award, NSF Organization, Program, Start Date, Last Amendment Date, Principal Investigator, Co-Principal Investigator, Organization, Award Instrument, awarded amount, State, PI Email Address, Address fields of the Organization, contact details, End Date, Abstract etc., summing up to 25 columns and a total of 295 records.

## Section I

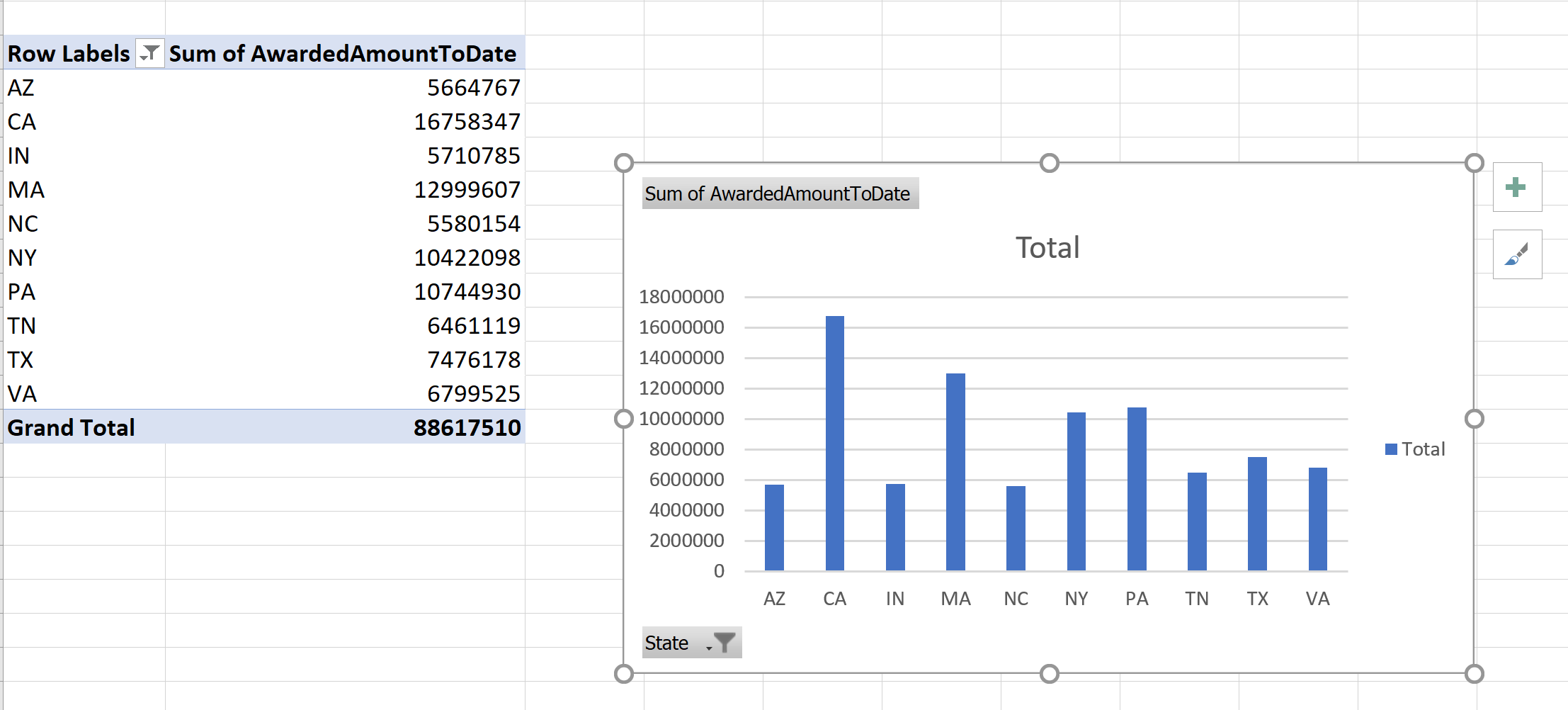
### Top 10 NSF funds on Smart and Connected Health among the states

In order to find out the states with highest funds we follow the below stated steps:

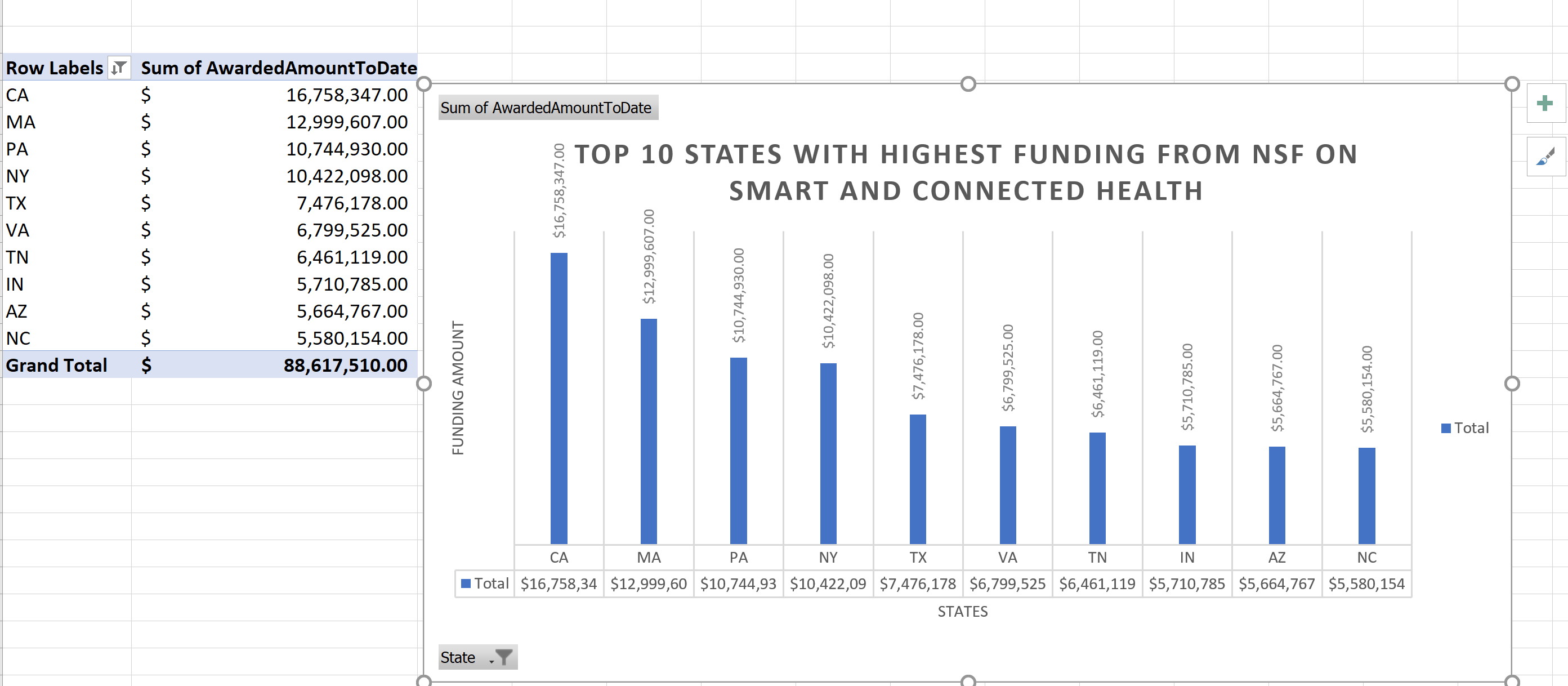
1. We select the entire data set and plot the Pivot Chart and Table by clicking the Pivot Chart under the Insert Menu



1. Then we can plot the states against the amount till date by selecting them from the Pivot Chart fields. Then we use the Top 10 value filter which filters the top 10 items in a field. So here the field we use is the “Sum of Award amount to Date”. So, by applying this filter it displays the top 10 funded states.



1. Then we can add descending sorting order based on the amount and the axis titles, title and the data labels for better understanding. So, the final chart is as shown below.

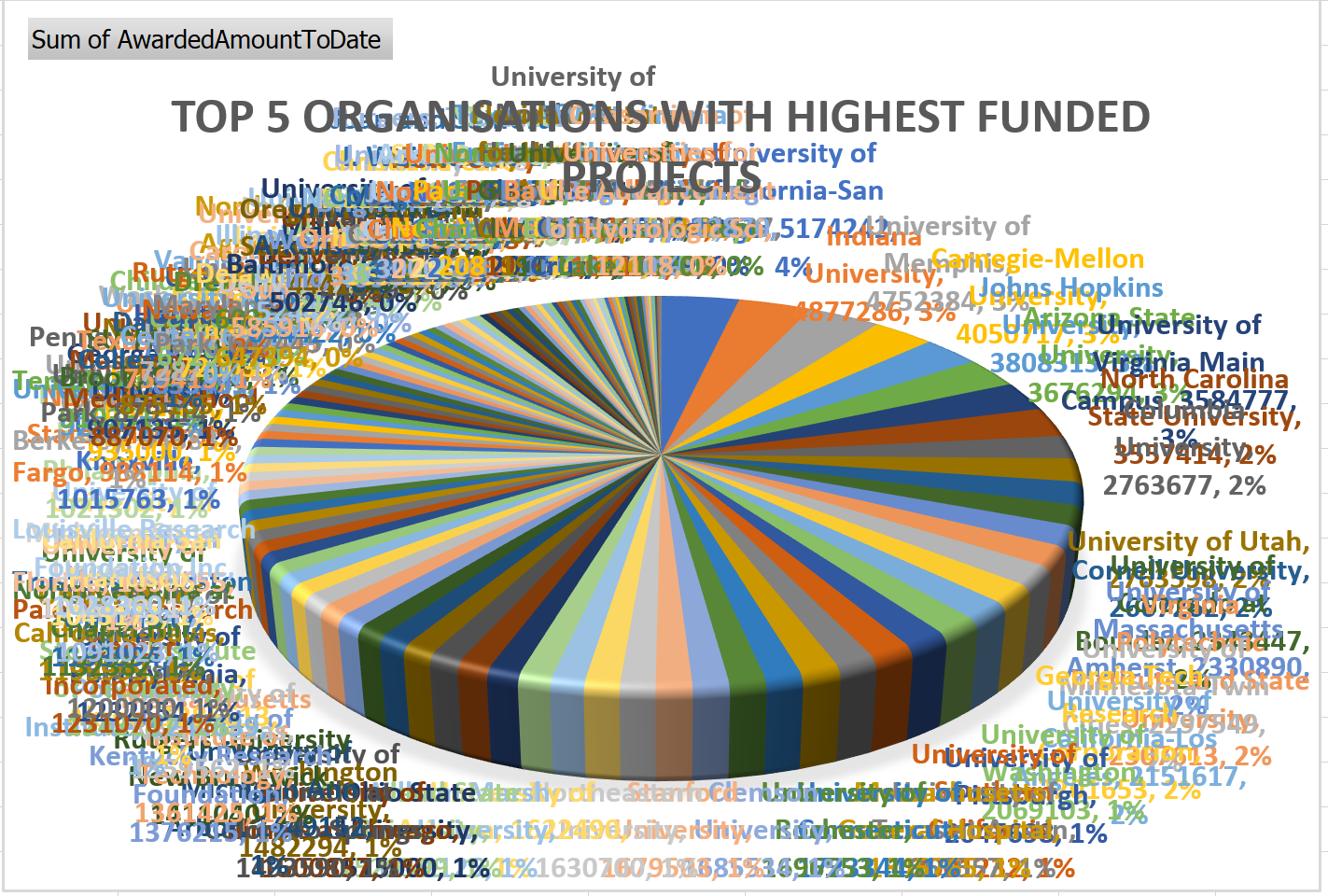


**Conclusion:**

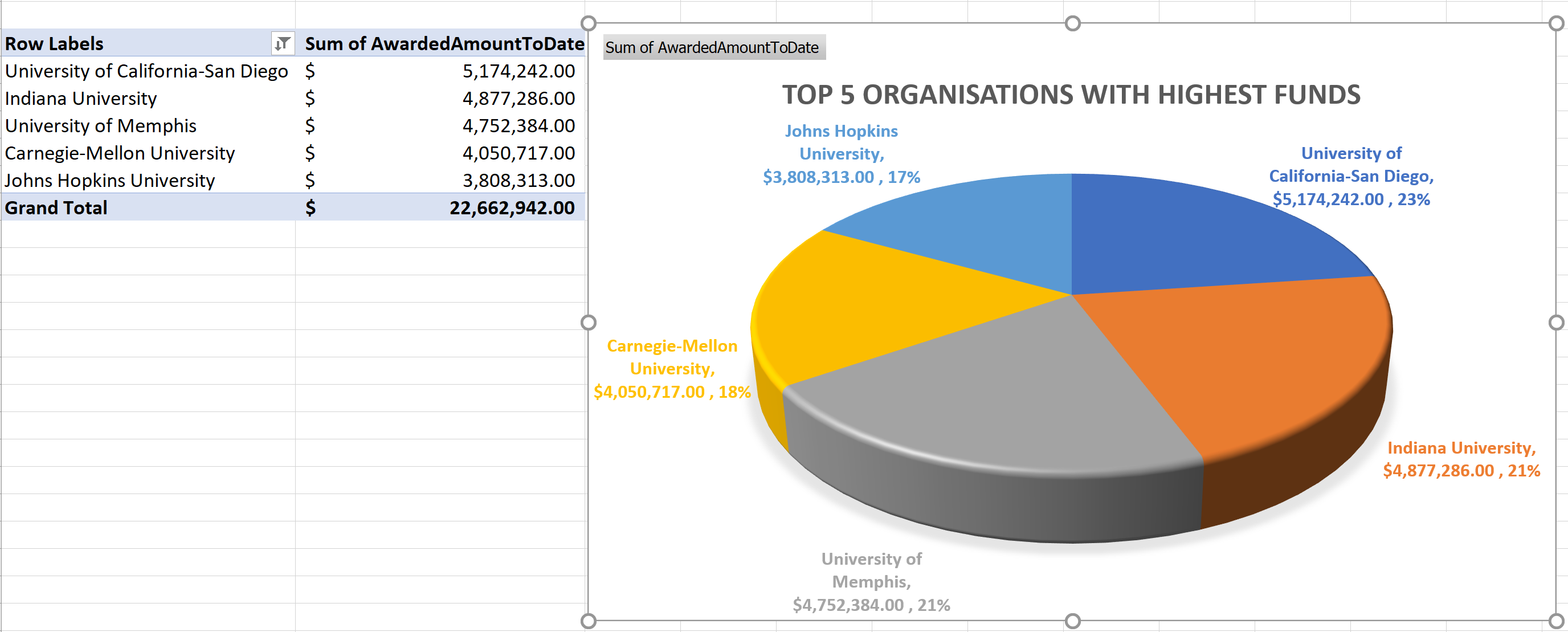
The graphs show that the California state ranks first with a total funding of $ 16,758,347 followed by Massachusetts with $ 12,999,607. Pennsylvania stands third with a fund of $ 10,744,930 followed by New York with a fund of $ 10,422,098. Texas, Virginia, Tennessee, Indiana, Arizona, North Carolina take the 5, 6, 7, 8,9 and 10 positions with funds as shown in the table above.

## Section II

1. **The top 5 Organizations with the highest fundings.**
2. The initial step is the data selection and the plotting of Pivot Chart and Table as shown in the above question. Then we graph the Organizations against the Amount Awarded to Date from the Pivot Chart Fields. Next, we change the chart type to Pie Chart which generates the below graph.



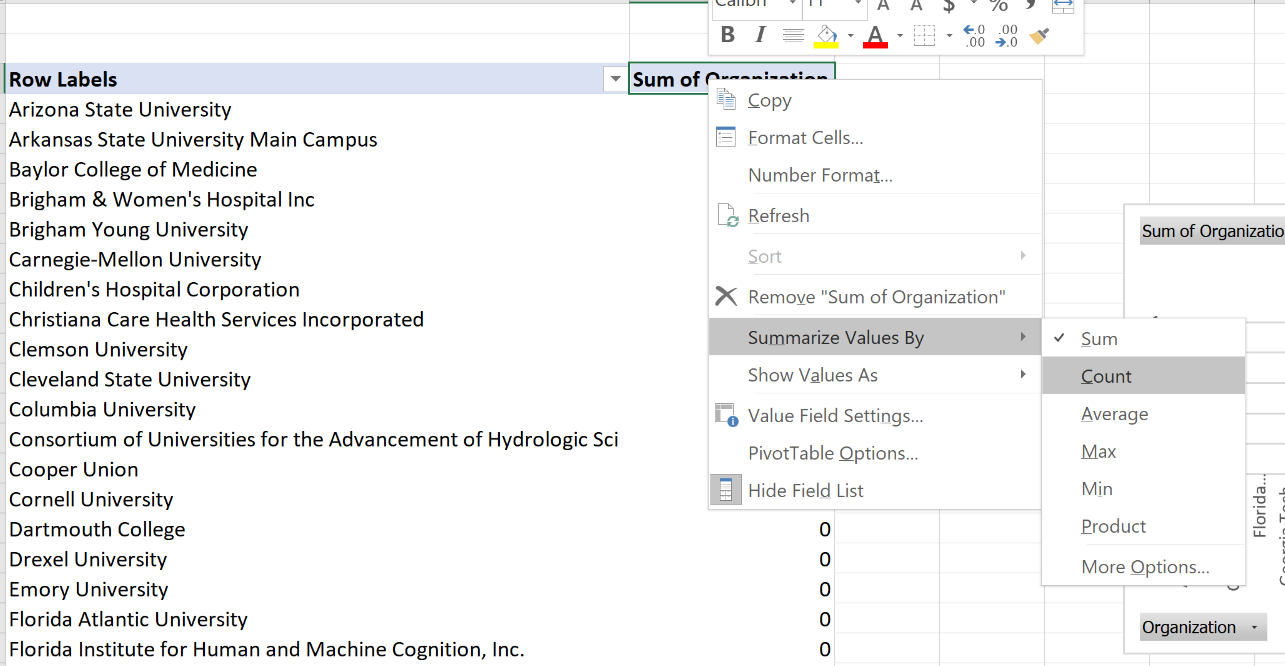
1. As we need just the top 5 organizations, they can be plotted by filtering using the Top10 variable filter in which we can manually enter the number of items to be plotted based on the highest funding amount. So, by applying that filter we get



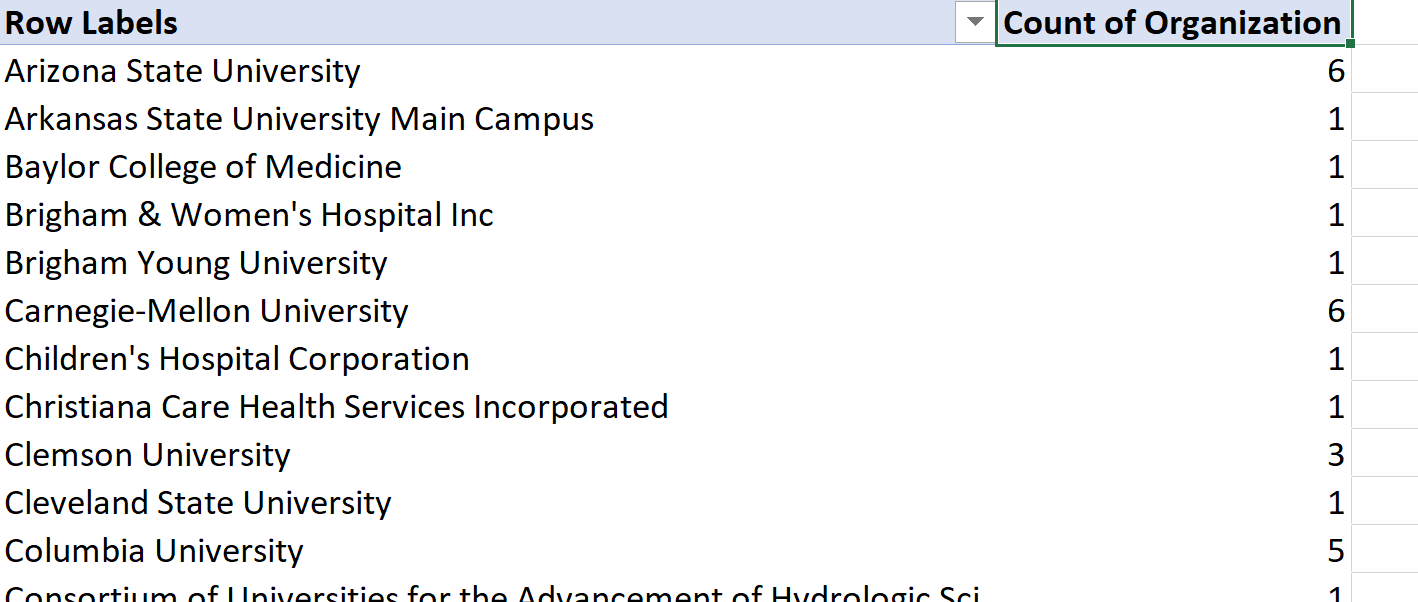
**Conclusion:**

The pie charts shows that the University of California-San Diego stands first with a total funding of $ 5,174,242 followed by Indiana University with a funding of $ 4,877,286, University of Memphis takes the third position with funding of $ 4,752,384. Carnegie-Mellon University and John Hopkins University stand fourth and fifth with fundings of $ 4,050,717 and $3,808,313 respectively. The Pie Charts also shows that they are differences in their fundings is not too high as the sector sizes are very close to each other, but we can notice the changes based on the percentage values.

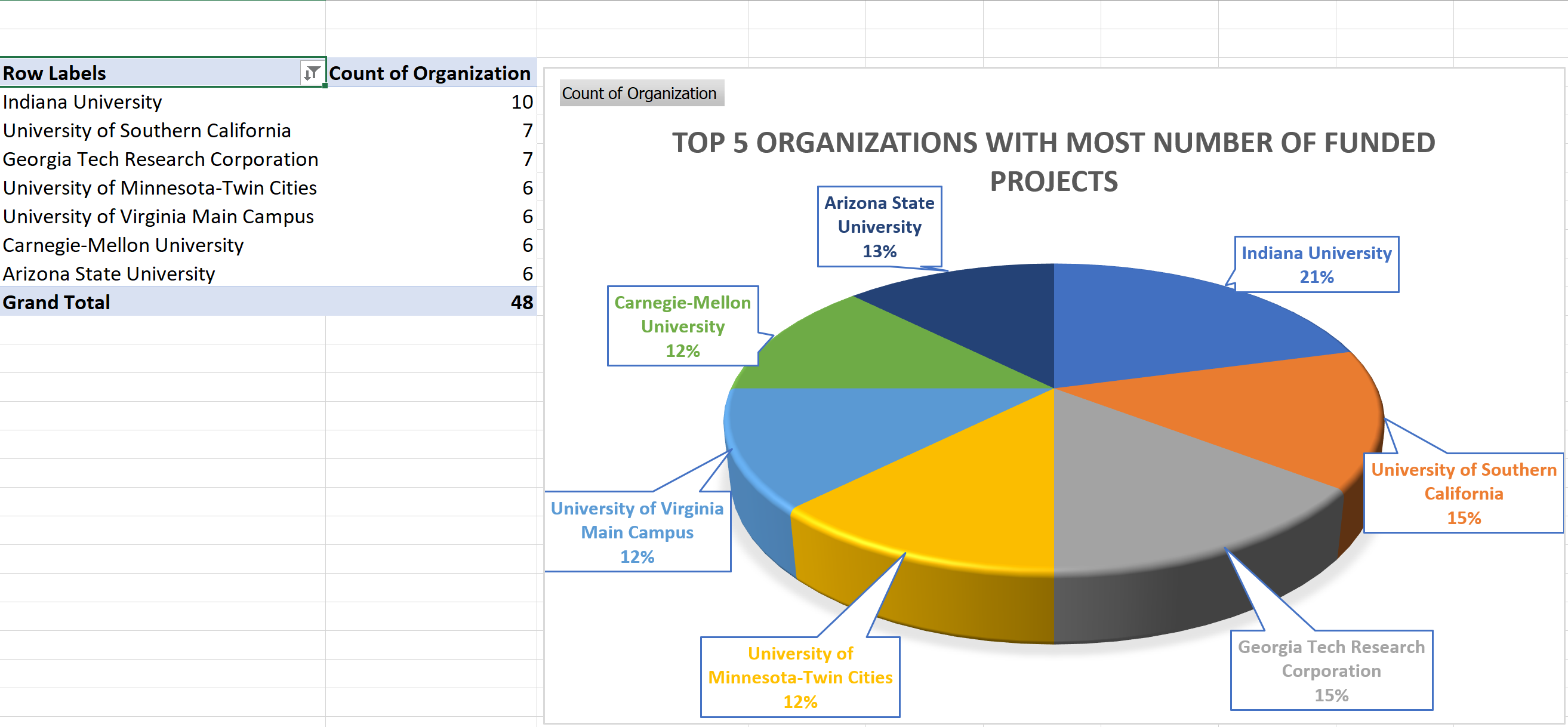
1. **The top 5 Organizations with the most number of funded projects**
2. The initial step is the data selection and the plotting of Pivot Chart and Table as we did it for the earlier questions or we can change the fields accordingly. But I prefer doing it in another sheet, so I inserted the Pivot Table and the chart. So, by selecting the shown in the above question. Then we graph the Organizations against the Amount Awarded to Date from the Pivot Chart Fields. Next, we change the chart type to Pie Chart which generates the below graph. By dragging the Organization into rows and Values field in the Pivot Table Fields, it lists all the organizations and their sum in two columns by default. So, we can display the counts of these organization by changing it from Sum to Count in the “Summarize Values by” option which appears when we right click on the Sum of Organization column.



1. When we change the option as said in the above point the table looks as shown below.



1. As we need just the top 5 organizations, they can be plotted by filtering using the Top10 variable filter in which we can manually enter the number of items to be plotted based on the highest funding amount. So, by applying that filter and changing the chart type to Pie Charts we get



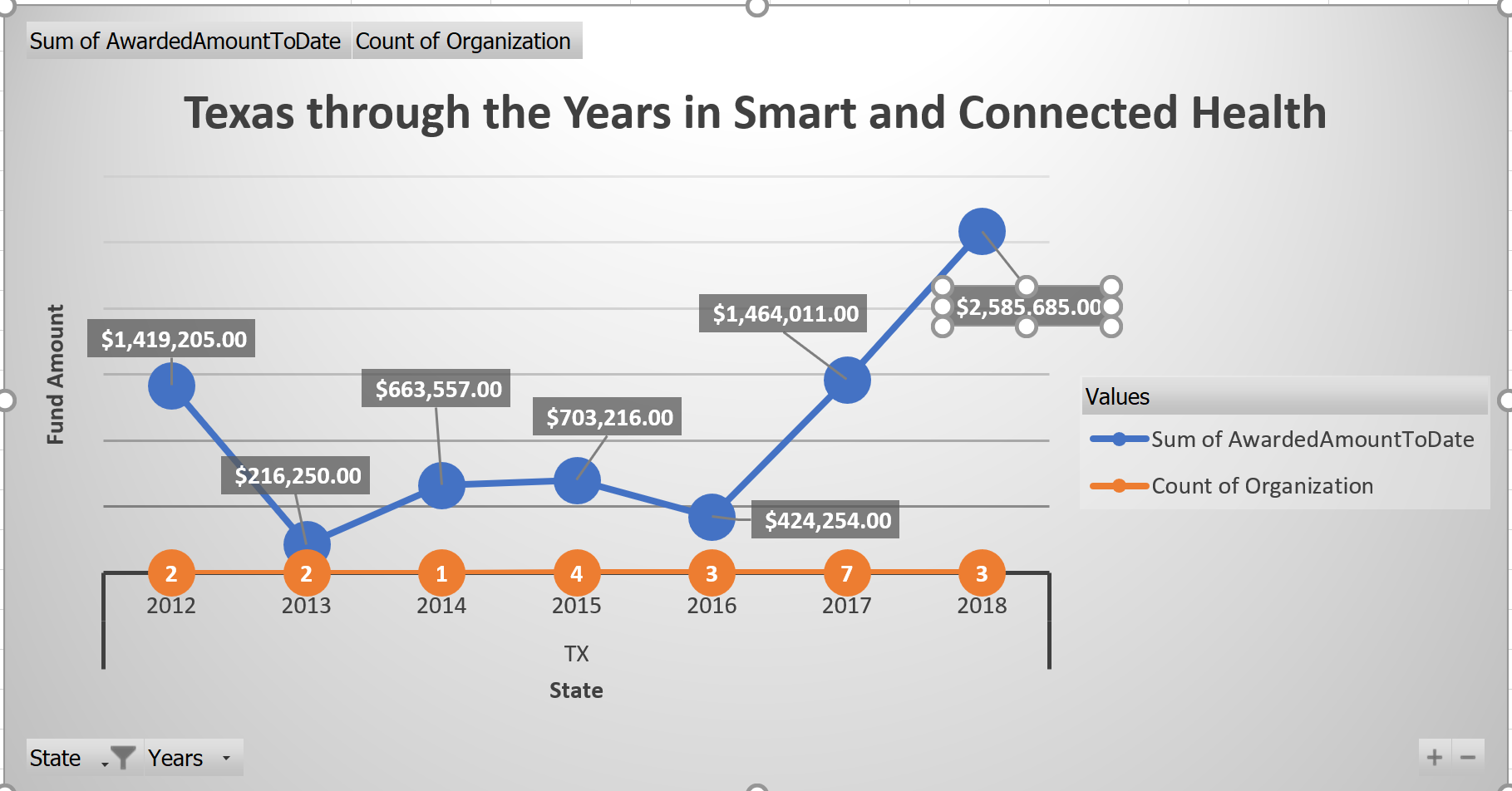
**Conclusion:**

The pie charts shows that the Indiana University has the most number of projects with a total of 10 projects. University of South California and Georgia Tech Research Corporation stand next with a total number of 7 projects. Whereas the University of Minnesota- Twin Cities, University of Virginia Main Campus, Carnegie Mellon University and the Arizona State University have a total of 6 projects each.

## Section III

**How is Texas doing in terms of Smart and Connected Health based on this dataset?**

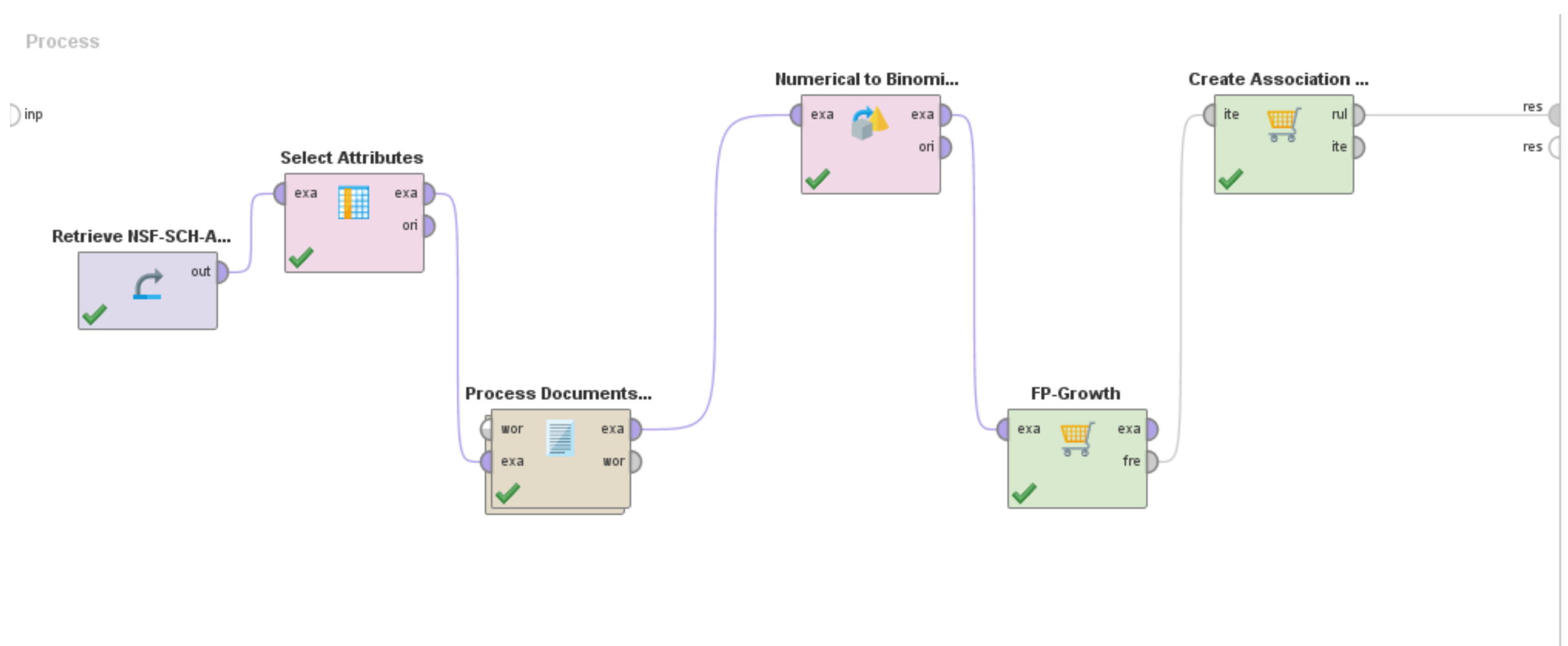
Texas when compared to other states in terms of funds, it stands fifth with a total funding of $ 7,476,178. It had its maximum funds in the year 2018 with a sum of $ 2,585,685 followed by the year 2017 with a sum of $ 1,464,011. Texas suffered lowest funds in the year 2013 with the sum of $ 216,250. There was a terrible decrease in the funds for the 2012- 2013 and 2015 – 2016. When the number of projects is considered, the Texas dealt with seven projects in the year 2017 followed by 4 projects in the year 2015. And the other years projects ranged between 1-3. Through the years there is a considerable increase in the number of organizations from Texas.



## Section IV

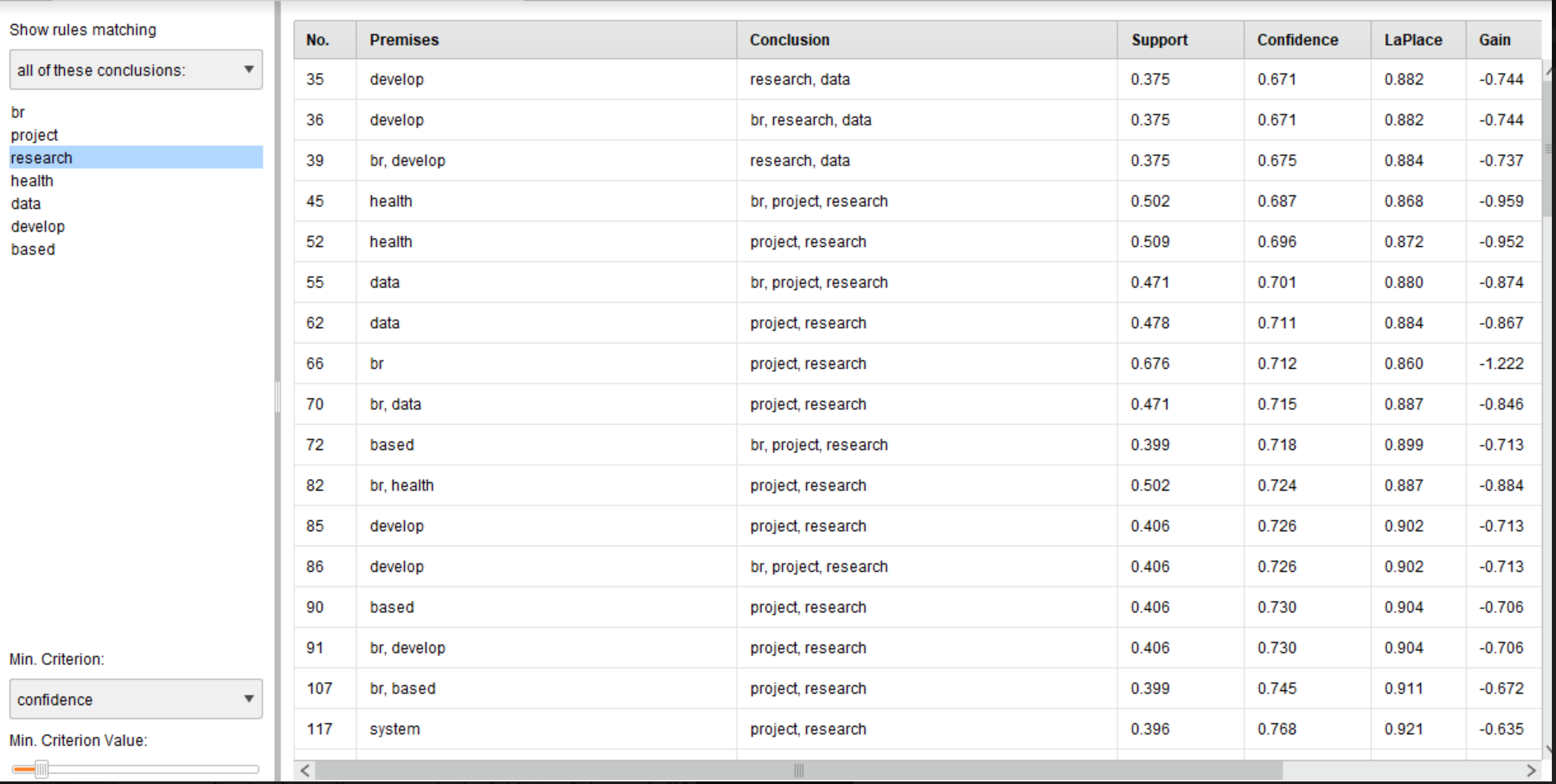
## Association Rules Analysis

We perform the Association Rule analysis using the software called Rapid Miner. Our goal is to analyze the relationships between words i.e., the co-occurrence of words. We initially import the dataset into the Rapid Miner by creating a new blank assignment. We edit the type of column that we would like to examine to abstract. Then we start adding blocks that would aid our purpose. We add several blocks like Select Attributes, Process Documents from Data, Numerical to Binomial, FP-Growth, Create Association Rules. The Process from Documents block has some blocks inside which are responsible for certain transformation on the data such as case transformation, tokenization of non-letters, removing the stop words etc., The model after adding all the blocks looks like as shown below.

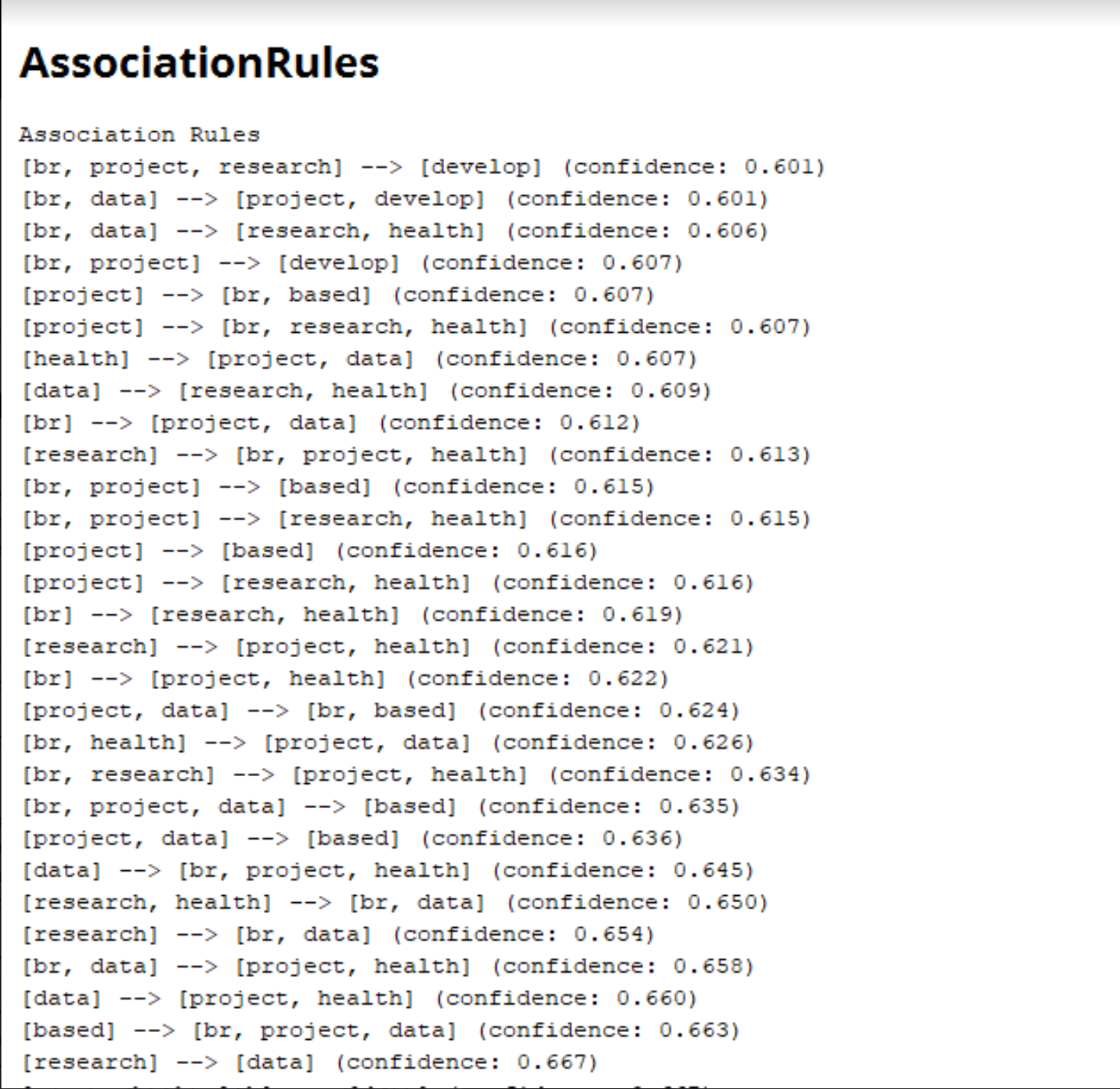


After running the model with min support as 0.5 and min confidence as 0.6, the results show that there are about 338 rules that have been created. The most frequently used words in this process are project, br, data, research, health develop and based. The graphical view of the association rules shows the words that are closely related to these above mentioned frequent words. When the word “research” is considered for high confidence such as 0.9 which means the words that are associated with the word research in more than 90% of the files, it shows that it is very closely related to students and br. When we try finding out the words at a lower confidence level we found out that words like project, health, develop, healthcare, information, data, support are closely related for confidence level of >= 0.8. Similarly, for the word “health” with confidence 0.88 we found the co-occurrences to be students and br again and for confidence level >=0.8 it is related to words like support, healthcare, students and br. In the same way the word “data” is most closely related to develop and br and when lower confidences are also considered its relation is found to be with words like research, project, time, health etc., The results for the obtained association rules are shown below.

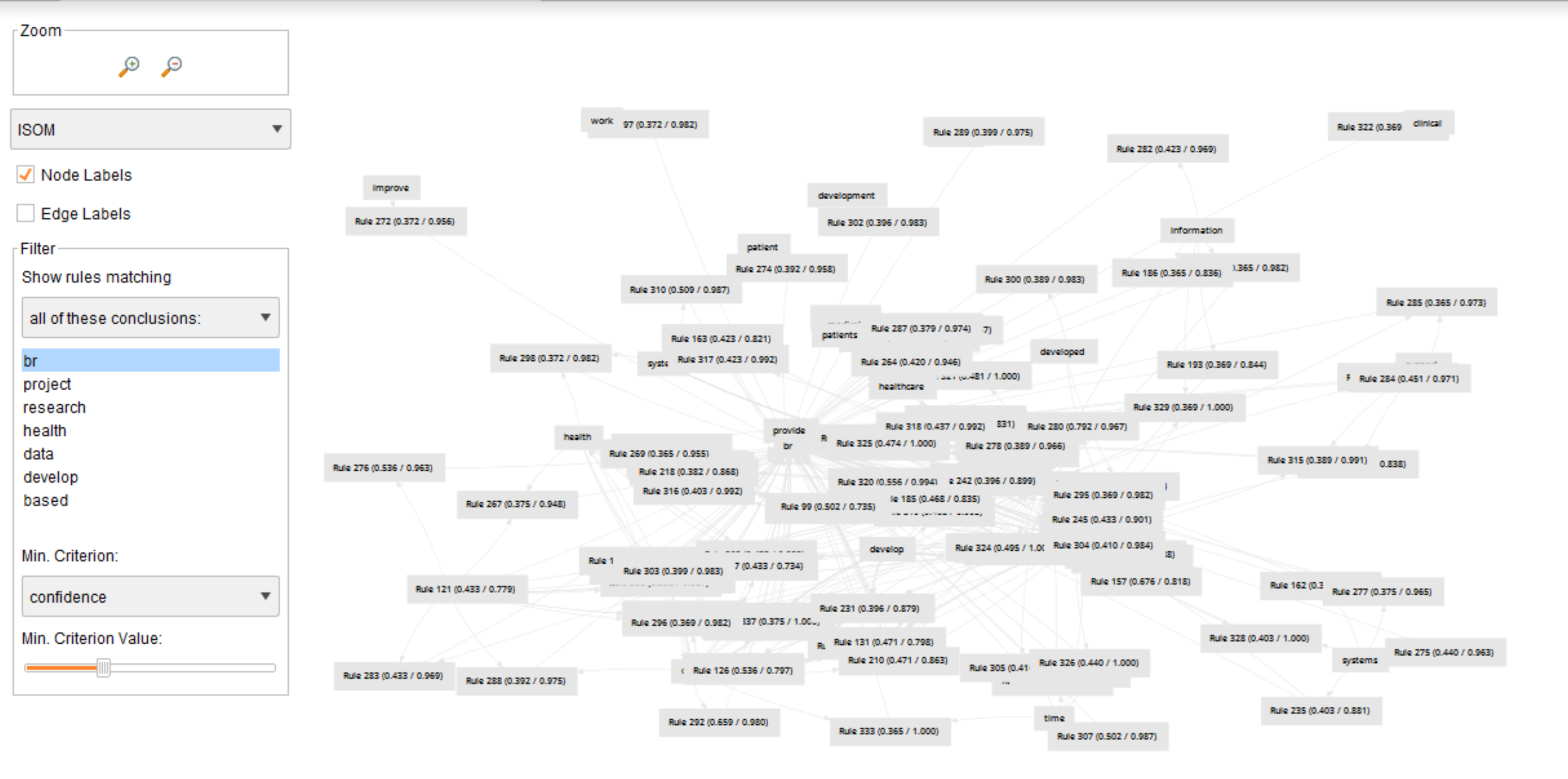
Tabular View:



Descriptive view:

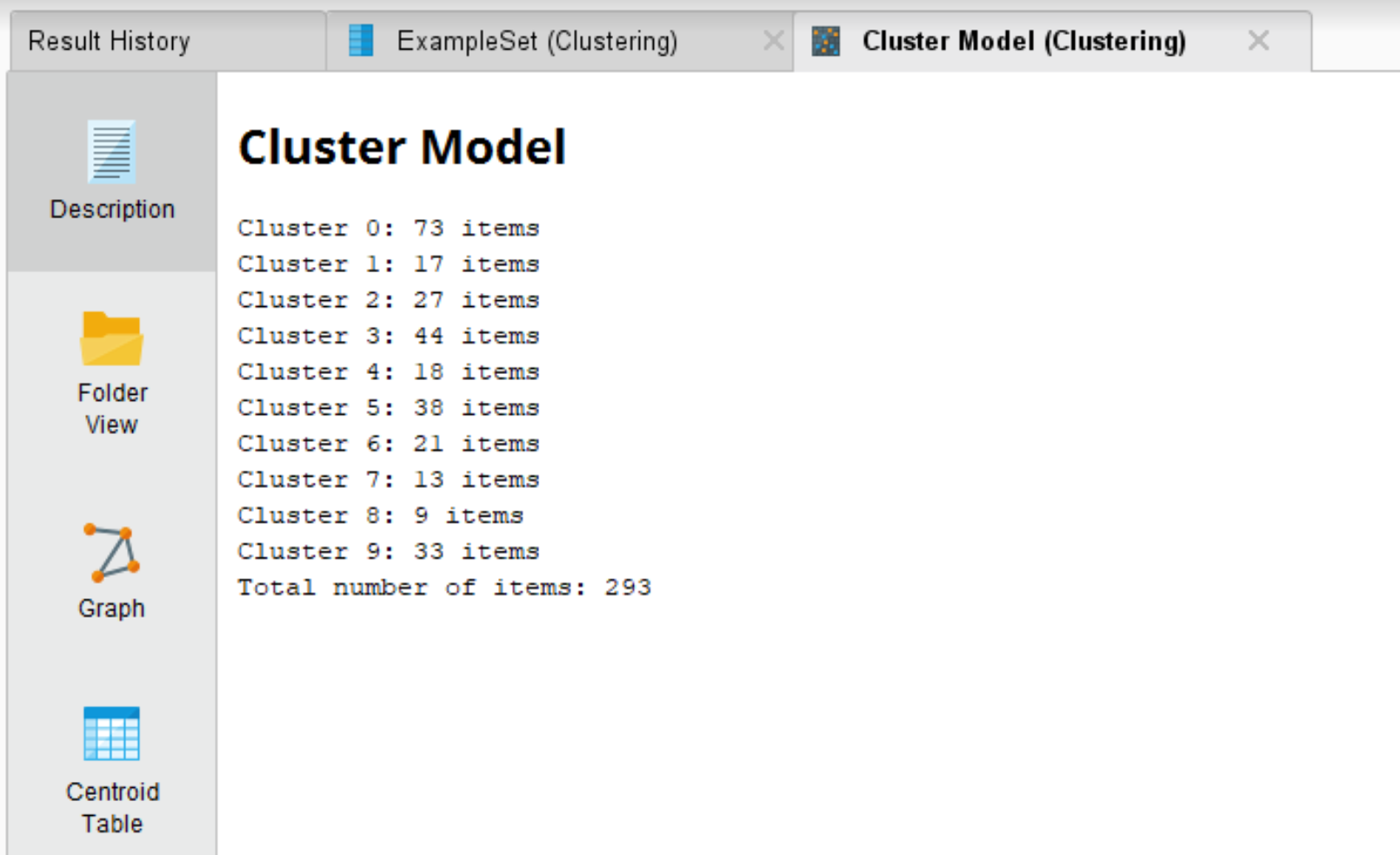


Graphical view:



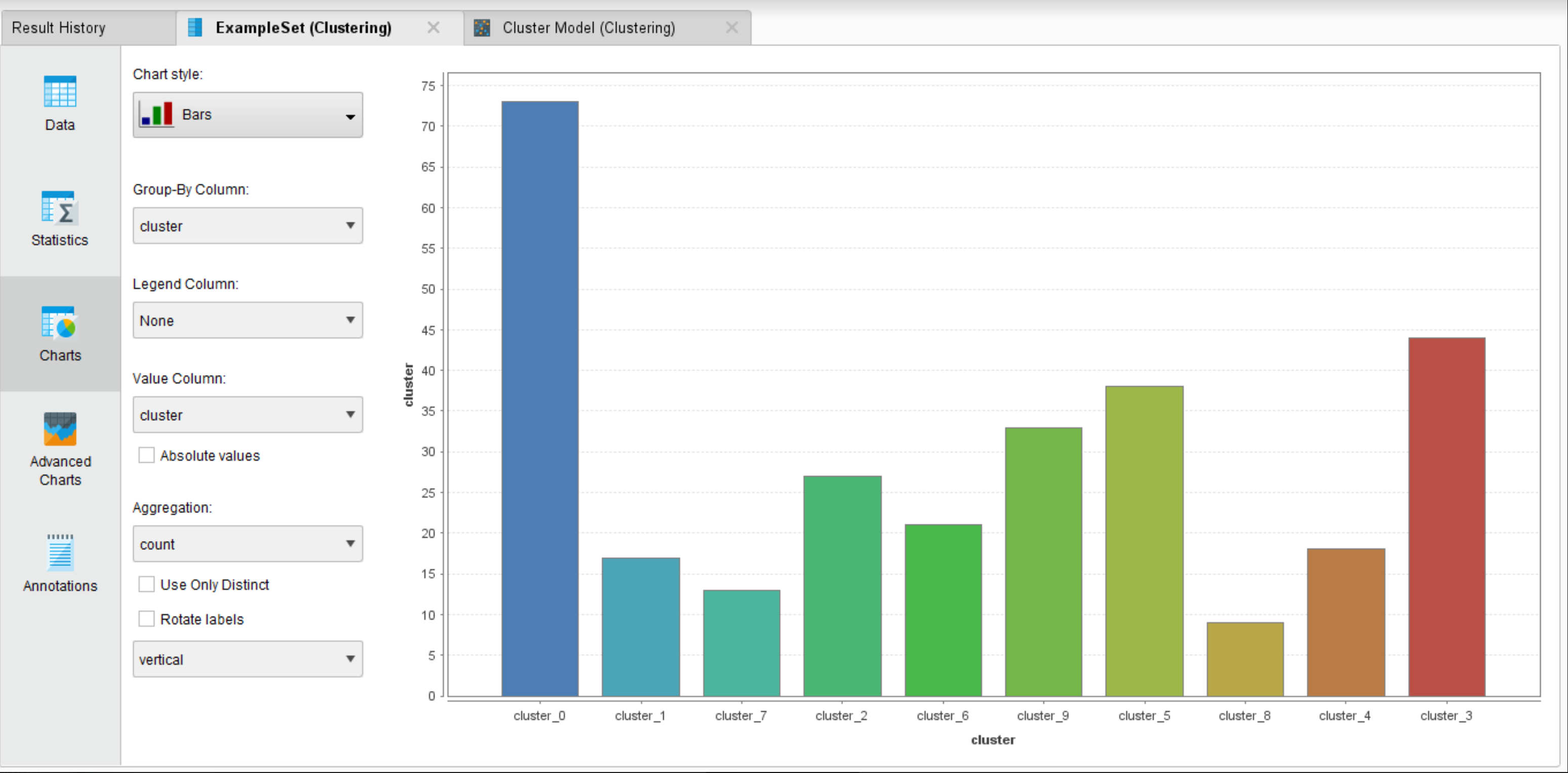
**Section V**

**Clustering Analysis of the Organizations** The Clustering Analysis is done by having the same initial steps as that of the Association Rules. Even this process has blocks like Select Attribute, Process Documents from Data. The next block in this is the Clustering block. The Clustering procedure which we do is the k-means clustering for k value of 5. This resulted in the formation of 10 clusters. The first cluster has the most number of items i.e., 73 and the ninth cluster has the lowest with just 9 items. The frequencies of clusters are as shown below.



The first 20 items in the formed clusters are given below:

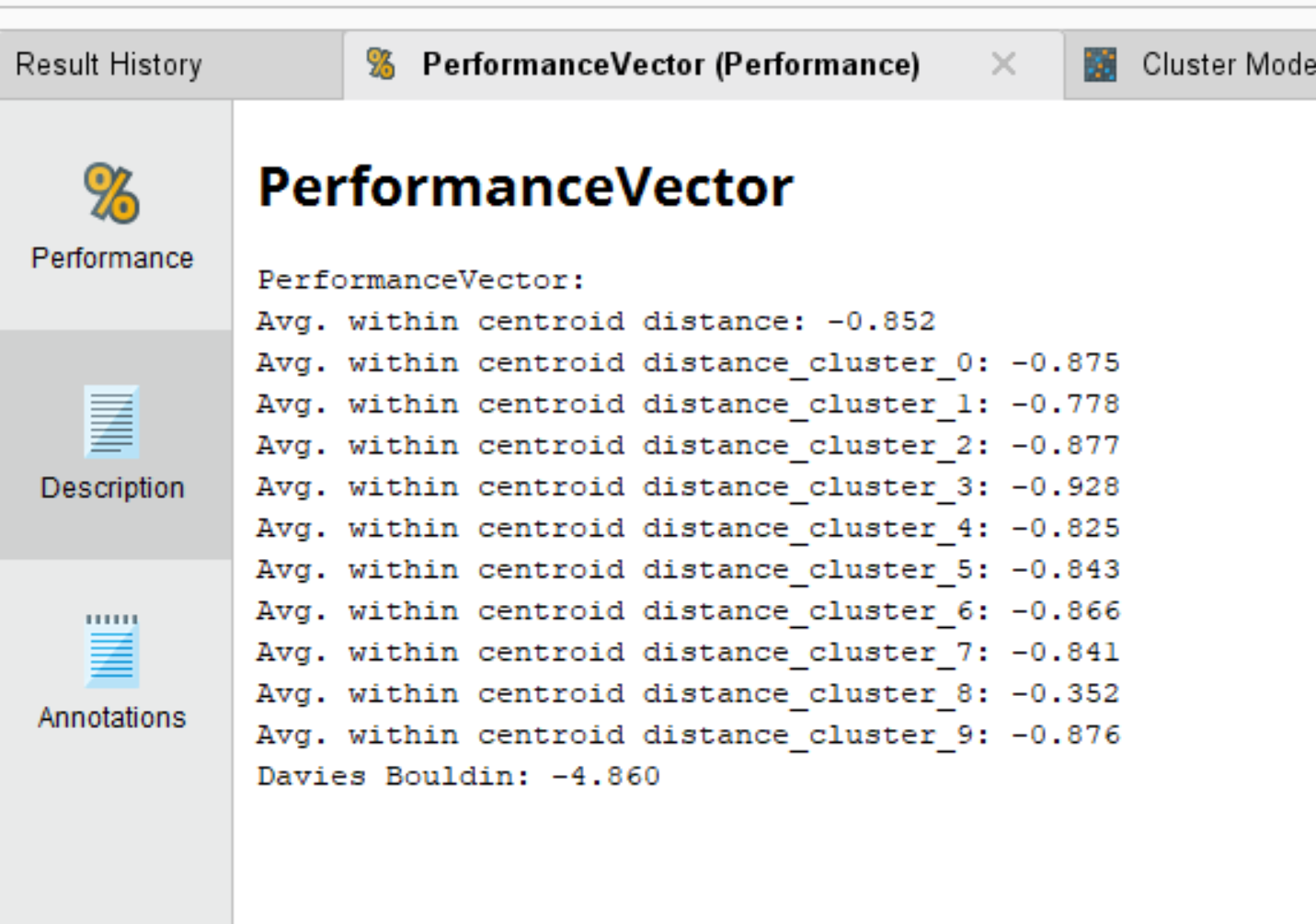
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Cluster 0** | **Cluster 1** | **Cluster 2** | **Cluster 3** | **Cluster 4** | **Cluster 5** | **Cluster 6** | **Cluster 7** | **Cluster 8** | **Cluster 9** |
| Sepsis | behavior | brain | therapy | heart | workshop | mobility | privacy | water | privacy |
| Clinical | driving | pain | individuals | smoking | conference | children | air | drinking | sleep |
| Patient | adhd | physiological | older | cardiac | student | impairments | coaching | hurricane | wifi |
| Telemedicine | dyadic | easysense | cell | failure | doctoral | limb | coach | disaster | grained |
| Phenotypes | behavioral | trauma | rehabilitation | acute | connected | fall | pulmonary | maria | monitoring |
| patients | fed | treatment | dialog | cessation | participants | orthosis | environmental | puerto | preserving |
| decision | change | drug | cognitive | patients | forum | soft | schemes | rico | infants |
| data | psychotherapy | safety | robot | wearable | area | body | data | natural | family |
| disease | behaviors | sensing | ha | fluid | science | garments | urban | disasters | signals |
| ehr | inspire | facial | feedback | decompensated | career | movement | goals | infrastructure | volume |
| models | symptoms | web | patient | atrial | international | ambulatory | manner | recovery | fine |
| risk | caregiver | signals | user | fibrillation | computing | sense | fatigue | environmental | breathing |
| drug | dynamics | compass | elderly | postoperative | students | visual | sms | hydroshare | wellbeing |
| care | dementia | injury | interaction | bioimpedance | informatics | physical | diseases | ontology | environments |
| hcv | indicators | vision | physical | hospitalization | health | upper | gas | prototype | ehrs |
| medical | obesity | dynamics | furnishings | simcardio | travel | activity | transportation | cyberinfrastructure | change |
| modeling | adolescent | image | activity | simvascular | consortium | face | sensitivity | impacted | waveform |
| electronic | ai | ddi | patients | device | scientific | motion | vehicle | response | techniques |
| image | preventive | iis | adults | monitoring | mhealth | shape | older | test | threatening |
| agitation | dynamical | self | self | continuously | meeting | mri | preserving | utilities | behavior |

The different result representations are given as  


The centroid table obtained is as shown below:



The cluster performance by distance operator gives us the following results. It has a Davies-Bouldin index of -11.622 which shows that the clusters formed are optimal.



**Section VI**

**Conclusion**

This final project involves the application of all the knowledge that we’ve learned from all the assignments that we’ve done till date. The questions have been answered using both Excel and Rapid Miner and the results have been visualized using graphs as above. I would like to thank Dr. Chen for giving us an opportunity to work on this.