Jalendar Reddy Maligireddy
11511290
Summer-2022

Sentiment Analysis on Zomato IPO Tweets

About data and source:-

Zomato had opened IPO in April, saying it plans to utilize the continues to finance development, which may incorporate mergers or takeovers. Zomato is advertising 1.23 billion offers, esteeming the IPO at 93.75 billion rupees. That incorporates issuing new offers worth up to 90 billion rupees as well as up to 3.75 billion rupees worth of stock sold by existing shareholders. Reuters detailed that final week Zomato's IPO drew $46.3 billion in offers and was more than 38 times

oversubscribed, with huge regulation speculators setting major bets. Zomato, at the side equal start-up Swiggy, rules India's $4.2 billion nourishment conveyance advertise, which is profoundly competitive but moreover exceptionally divided.

Separated from nourishment conveyance, Zomato too lets clients book tables and totals surveys for eateries. Tech mammoth Uber sold its India nourishment conveyance trade to Zomato final year in an all-stock exchange that gave the U.S. company a stake within the start-up. Zomato's other conspicuous sponsor incorporate Indian web company Data Edge, Alibaba-affiliate Subterranean insect Bunch and Singapore state speculator Temasek.

The data set is taken from Kaggle. It's a large number of dataset, extracted from Twitter.

Importing the required packages

```
In [1]:  import pandas as pd
```

```
In [2]:  df=pd.read_csv('zomato-ipo.csv')
```

Loading the data set into data frame using panda's package.

```
In [7]:  df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11494 entries, 0 to 11493
Data columns (total 36 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   id               11494 non-null  int64
 1   conversation_id  11494 non-null  int64
 2   created_at       11494 non-null  object
 3   date             11494 non-null  object
 4   time             11494 non-null  object
 5   timezone         11494 non-null  int64
 6   user_id          11494 non-null  int64
 7   username         11494 non-null  object
 8   name             11493 non-null  object
 9   place            9 non-null      object
 10  tweet            11494 non-null  object
 11  language         11494 non-null  object
 12  mentions         11494 non-null  object
 13  urls             11494 non-null  object
```

The data set contains 36 columns and 11494 rows, it's a raw data.

Jalendar Reddy Maligireddy
11511290
Summer-2022

```
In [7]:  df.columns

Out[7]: Index(['id', 'conversation_id', 'created_at', 'date', 'time', 'timezone',
       'user_id', 'username', 'name', 'place', 'tweet', 'language', 'mentions',
       'urls', 'photos', 'replies_count', 'retweets_count', 'likes_count',
       'hashtags', 'cashtags', 'link', 'retweet', 'quote_url', 'video',
       'thumbnail', 'near', 'geo', 'source', 'user_rt_id', 'user_rt',
       'retweet_id', 'reply_to', 'retweet_date', 'translate', 'trans_src',
       'trans_dest'],
      dtype='object')
```

It's a raw data it contains, the following columns

The Data dictionary of the data

The data type is integer and object (it may be string, date, time, URL)

Column Name	**Description**	**Data Type**
id	User id	Integer
conversation_id	The conversation id	Integer
created_at	the time and date of the conversation_id created	object
date	The tweet created date	object
time	The tweet created time	object
timezone	The tweet created timezone	object
user_id	The tweet owner user ID	object
username	The tweet person username	object
name	Name of the tweet person	object
place	The place where tweet or location	object
tweet	What's tweet	object
language	The language of the tweet	object
mentions	Any user in the tweet	object
urls	Any url tagged or linked	object
photos	Attached picture format	object
replies_count	The number of replies	integer
retweets_count	The number retweets	integer
likes_count	The number of likes	integer
hashtags	Any user mentioned	object
cashtags	cashtag	object
link	connections	object
retweet	Reply to the tweet	object
quote_url	The url of the quote	object
video	Any video in the tweet	object
thumbnail	thumbnail	object
near	Near by	object
geo	geography	object
source	the source taken	object

user_rt_id	the user ID of retweeted person	object
user_rt	The user retweet	object
retweet_id	The retweeted ID	object
reply_to	What's the reply	object
retweet_date	The retweeted date	object
translate	Any translation	object
trans_src	The translation source	object
trans_dest	The destination of the translation	object

Data Cleaning:-

For the data cleaning process, initially checking any null values.

Checking any null values in the data.

```
In [6]: df.isnull().sum()

Out[6]: id                    0
        conversation_id       0
        created_at            0
        date                  0
        time                  0
        timezone              0
        user_id               0
        username              0
        name                  1
        place             11485
        tweet                 0
        language              0
        mentions              0
        urls                  0
        photos                0
        replies_count         0
        retweets_count        0
        likes_count           0
        hashtags              0
        cashtags              0
        link                  0
        retweet               0
        quote_url         10604
        video                 0
        thumbnail          8846
        near              11494
        geo               11494
        source            11494
        user_rt_id        11494
        user_rt           11494
        retweet_id        11494
        reply_to              0
        retweet_date      11494
        translate         11494
        trans_src         11494
        trans_dest        11494
        dtype: int64
```

Found few columns had null values.

Removing all the null values and selecting required columns for analysis.

Jalendar Reddy Maligireddy
11511290
Summer-2022

Selecting required columns for the analysis which will be effective. By taking needed columns created a new data frame.

```
df1= df[['created_at',  'name','tweet','replies_count', 'retweets_count','likes_count', 'hashtags','retweet']]
```

```
df1
```

2]:

	created_at	name	tweet	replies_count	retweets_count	likes_count	hashtags	retweet
0	2021-07-15 04:04:31 UTC	Aman Singh	@ipo_mantra I haven't applied for Zomato IPO -...	1	0	2	[]	False
1	2021-07-15 04:03:51 UTC	TRIPURATEER	#PaytmIPO #PaytmIPOnews #zomatoipo #Zomatoshar...	0	0	0	['paytmipo'- 'paytmiponews'- 'zomatoipo'- 'zom...	False
2	2021-07-15 04:03:48 UTC	Ankit Kakkad	#wipro CMP 570😍 Trgt achieved 🚀 😊 (Safe players...	0	0	0	['wipro'- 'wipro'- 'nifty50'- 'hdfcbank'- 'ban...	False
3	2021-07-15 04:03:05 UTC	Saurabh Chandra	So much love poured by the startup community o...	0	0	0	['zomatoipo']	False
4	2021-07-15 04:02:51 UTC	Dr. Silvia Elaluf-Calderwood	Zomato IPO: India food delivery 'unicorn' open...	0	0	0	['india'- 'food'- 'unicorn']	False
...	...	...	...	...	...	...	...	...
11489	2020-09-10 12:05:45 UTC	Nikita Vashisht	Is this why Info Edge's shares zoomed suddenly...	0	0	0	['trading'- 'zomato'- 'ipo'- 'markets'- 'food'...	False
11490	2020-08-22 10:57:16 UTC	Digital Agents	Business news #84 | Ambience Mall illegal-Pata...	0	0	0	[]	False
11491	2020-05-22 14:07:31 UTC	Kushal Bhagia	Wow looks like Mumbai is allowing online order...	1	0	32	[]	False
11492	2015-06-23 10:44:25 UTC	zooey m	#TheIndianCapitalist: e-IPO for #Start-ups ht...	0	0	0	['theindiancapitalist'- 'start'- 'blog'- 'sebi...	False
11493	2013-03-24 13:52:00 UTC	IB Singapore	Indian restaurant guide Zomato Media planning ...	0	0	0	['zomato'- 'ipo']	False

11494 rows × 8 columns

Checking any null values in the required data.

```
df1.isnull().sum()
```

3]:

```
created_at        0
name              1
tweet             0
replies_count     0
retweets_count    0
likes_count       0
hashtags          0
retweet           0
dtype: int64
```

Dropping null value rows and cleaning the data, final dataset contains 9997 rows and 8 columns.

```
df2.shape
```

```
(9997, 8)
```

Exporting to excel file, the final dataset.

```
df2.to_csv("zomatoIpo.xlsx")
```

Jalendar Reddy Maligireddy
11511290
Summer-2022

The goal of the analysis (what did you want to find out?):-

- To find the most influential tweets about Zomato IPO.
- From the above data set we need to figure out which are dependent variables and what's the connection among them.
- Our task is to find out this Zomato IPO is good for investment or not, whether it leads to loss or gain.

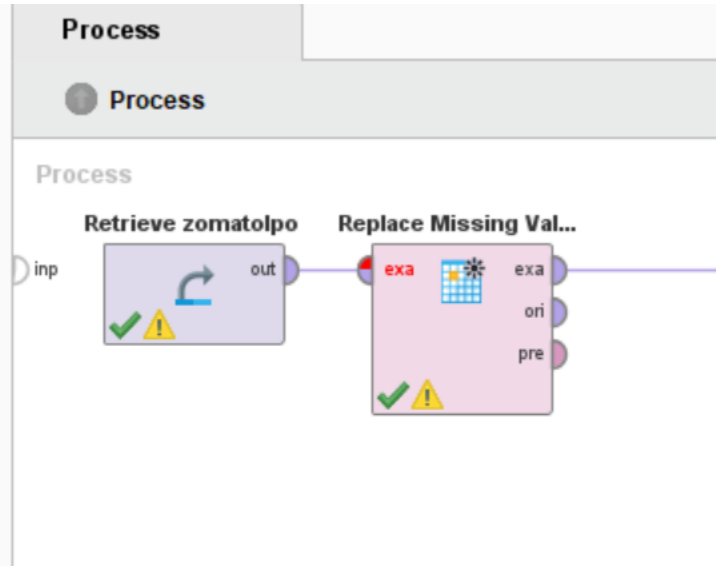
Data mining technique used and final result diagram:-

The data mining technique used are:

- Correlation analysis
- Association analysis
- k-means Cluster analysis

The process for data mining: -

First loading the data into rapid miner tool, replacing any missing values any present.



Checking any null present in the data.

The data contains 9997 rows and 8 columns

Jalendar Reddy Maligireddy
11511290
Summer-2022

Result History | ExampleSet (Replace Missing Values)

Data | Statistics | Visualizations | Annotations

Filter (8 / 8 attributes): Search for Attributes

Name	Type	Missing	Statistics		
created_at	Polynominal	0	Least: 2021-07- [...] 1 UTC (1)	Most: 2021-07- [...] 0 UTC (4)	Values: 2021-07-13 13:30:00 UTC (4), 2021-07-14 03
username	Text	0	Least: zindademocracy (1)	Most: shivanshnangia (143)	Values: shivanshnangia (143), etnowlive (91), ...[6525
tweet	Text	0	Least: ðŸ"»âœ–ï [...] ocare (1)	Most: @zomato IPO (60)	Values: @zomato IPO (60), #ZomatoIPO (48), ...[9754
replies_count	Integer	0	Min: 0	Max: 329	Average: 0.828
retweets_count	Integer	0	Min: 0	Max: 204	Average: 1.032
likes_count	Integer	0	Min: 0	Max: 4211	Average: 10.596
hashtags	Text	0	Least: ['ã¤ãƒªãƒ‰ãƒ¡ãƒ¢ã'] (1)	Most: [] (3146)	Values: [] (3146), ['zomatoipo'] (2508), ...[2562 more]

Showing attributes 1 - 8 | Examples: 9,997 Special Attributes: 0 Regular Attributes: 8

The sample data.

Result History | ExampleSet (Replace Missing Values)

Data | Statistics | Visualizations

Open in: Turbo Prep | Auto Model

Filter (9,997 / 9,997 examples): all

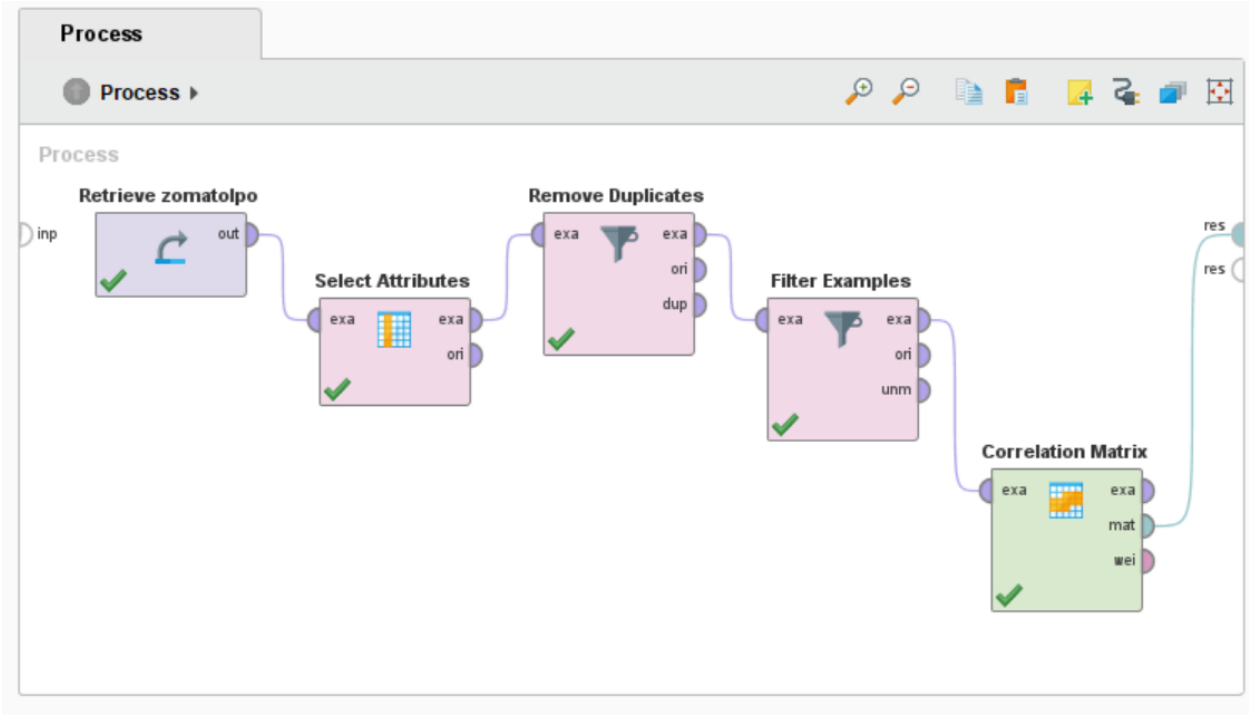
Row No.	created...	userna...	tweet	replies_...	retweet...	likes_c...	hashtags	retweet
1	2021-07-...	thakura4...	@ipo_m...	1	0	2	[]	false
2	2021-07-...	tripurateer	#Paytml...	0	0	0	['paytmip...	false
3	2021-07-...	ankitkak...	#wipro C...	0	0	0	['wipro', '...	false
4	2021-07-...	schandr...	So much...	0	0	0	['zomatoi...	false
5	2021-07-...	sileIf	Zomato I...	0	0	0	['india', 'f...	false
6	2021-07-...	k_ranjan	#zomatoi...	0	0	0	['zomatoi...	false
7	2021-07-...	akash_s...	GMP on t...	0	0	0	['ipo', 'zo...	false

Correlation analysis:-

This Administrator decides relationship between all Properties, and it can deliver a weights vector based on these relationships. Relationship could be a factual method that can appear whether and how unequivocally sets of Traits are related.

A relationship could be a number between -1 and +1 that measures the degree of affiliation between two Properties (call them X and Y). A positive esteem for the relationship infers a positive affiliation. In this case huge values of X tend to be related with huge values of Y and little values of X tend to be related with little values of Y. A negative esteem for the relationship suggests a negative or reverse affiliation. In this case huge values of X tend to be related with little values of Y and bad habit versa.

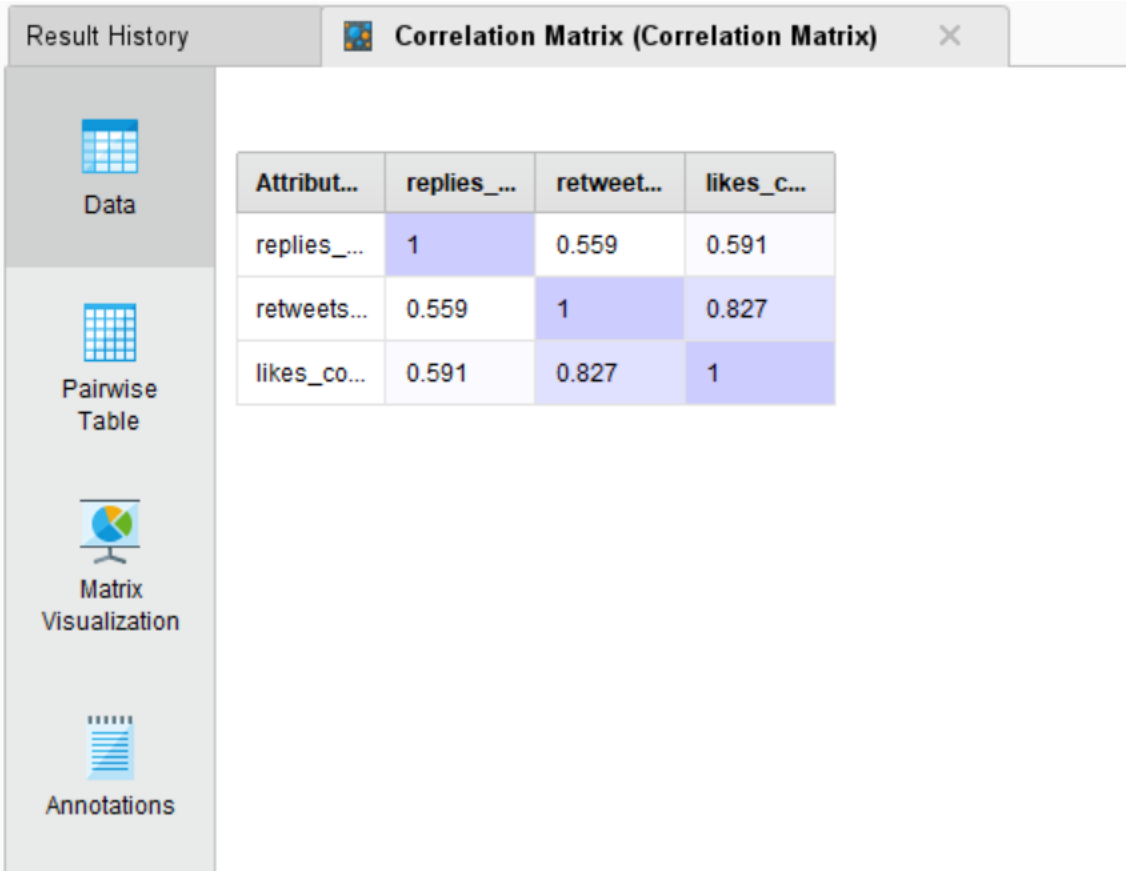
Jalendar Reddy Maligireddy
11511290
Summer-2022



After performing above operation, found the results are.

Retweets, likes, replies are important in analysis.

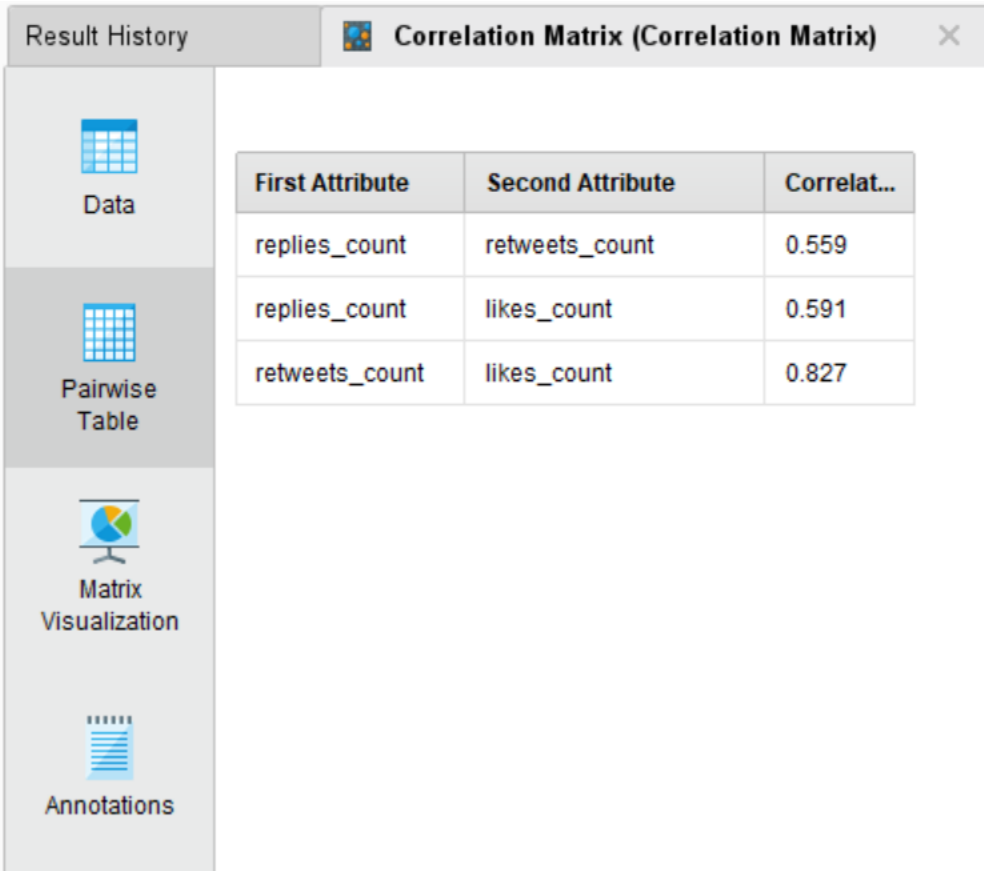
Jalendar Reddy Maligireddy
11511290
Summer-2022



Attribut...	replies_...	retweet...	likes_c...
replies_...	1	0.559	0.591
retweets...	0.559	1	0.827
likes_co...	0.591	0.827	1

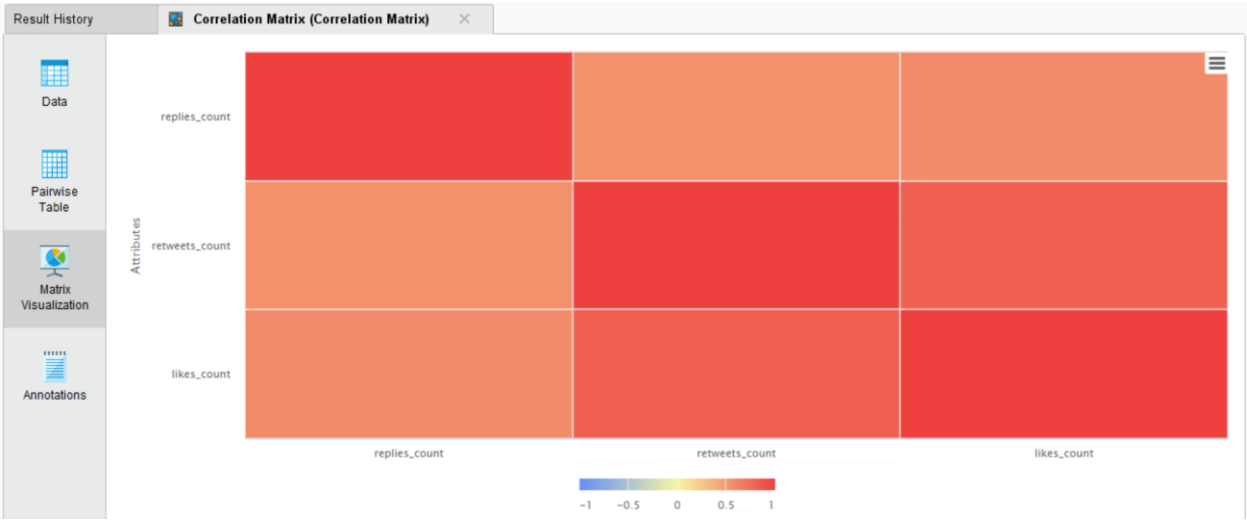
The pair wise table

Jalendar Reddy Maligireddy
11511290
Summer-2022



First Attribute	Second Attribute	Correlat...
replies_count	retweets_count	0.559
replies_count	likes_count	0.591
retweets_count	likes_count	0.827

The correlation matrix on the replies_count,retweet_count,likes_count.



Association analysis:-

This administrator creates a set of affiliation rules from the given set of visit item sets.

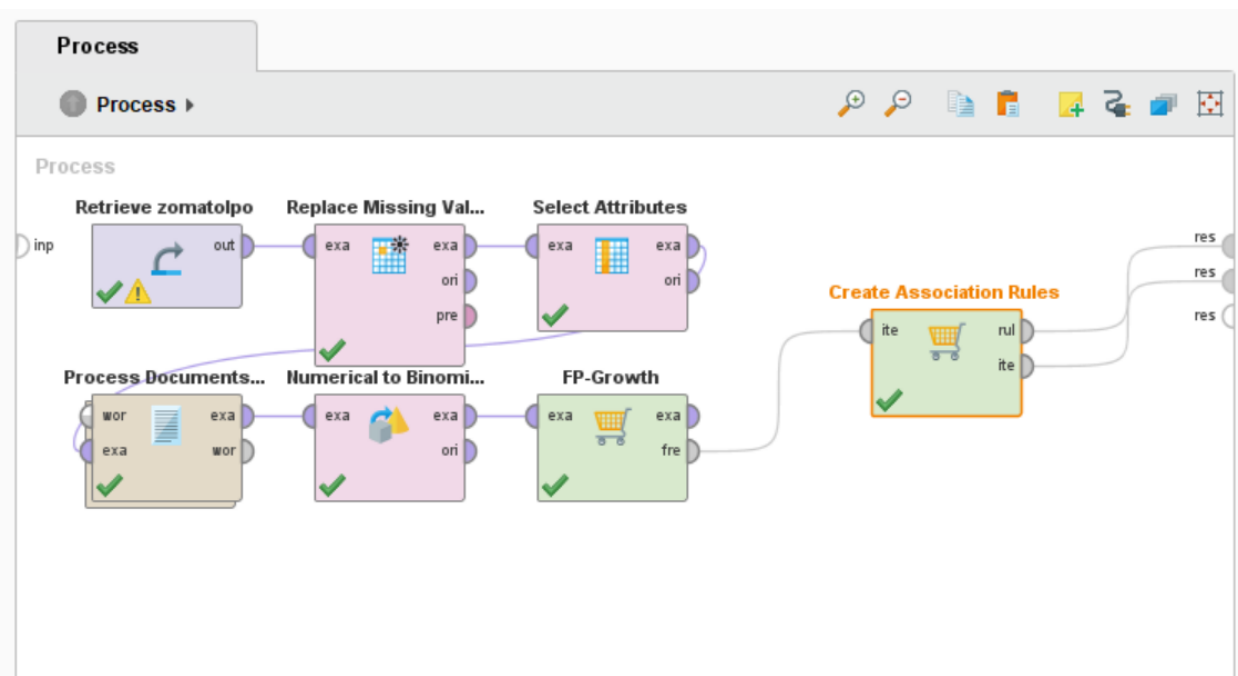
Jalendar Reddy Maligireddy
11511290
Summer-2022

Affiliation rules are if/then articulations that offer assistance reveal connections between apparently irrelevant information. An illustration of an affiliation run the show would be "In case a client buys eggs, he is 80% likely to moreover buy drain." An affiliation run the show has two parts, and forerunner (in case) and a resulting (at that point).

A predecessor is an thing (or itemset) found within the information. A resulting is an item (or itemset) that's found in combination with the antecedent. Association rules are made by analyzing information for visit if/then designs and utilizing the criteria back and certainty to recognize the foremost critical relationships.

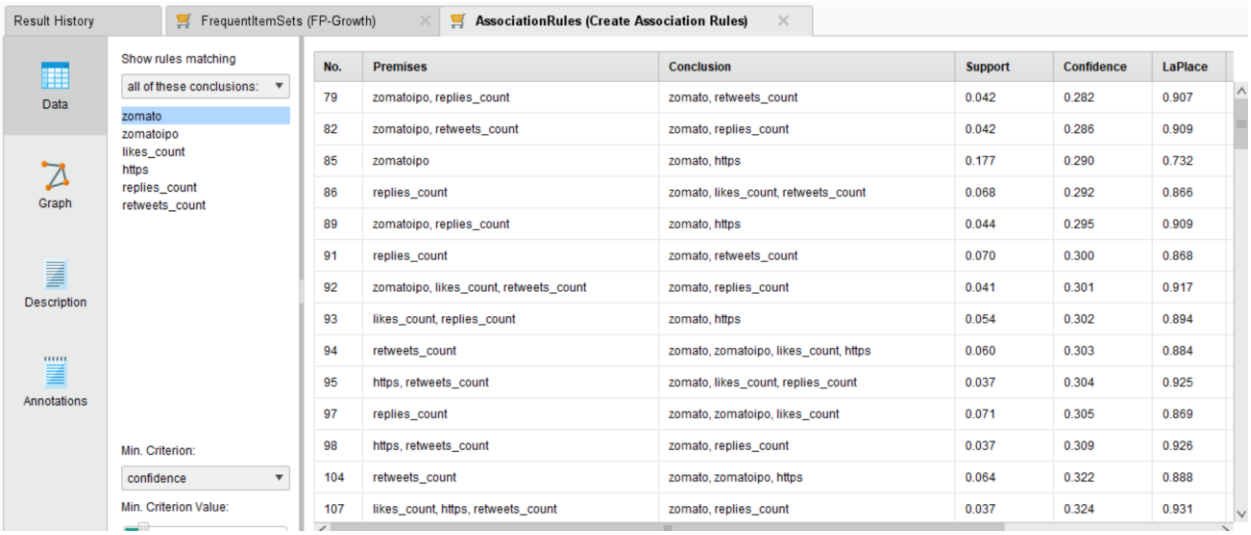
Support is an sign of how habitually the things show up within the database. Certainty demonstrates the number of times the if/then explanations have been found to be genuine. The visit if/then designs are mined utilizing the administrators just like the FP-Growth administrator. The Make Affiliation Rules administrator takes these visit item sets and creates affiliation rules.

The process of Association rule:



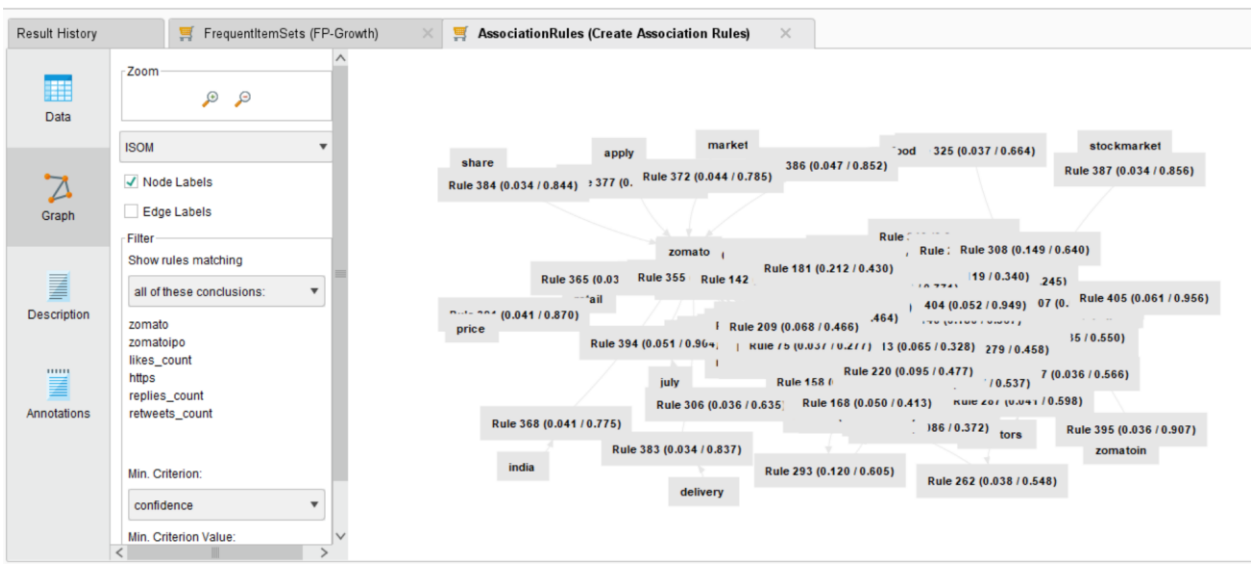
The Association rules are

Jalendar Reddy Maligireddy
11511290
Summer-2022



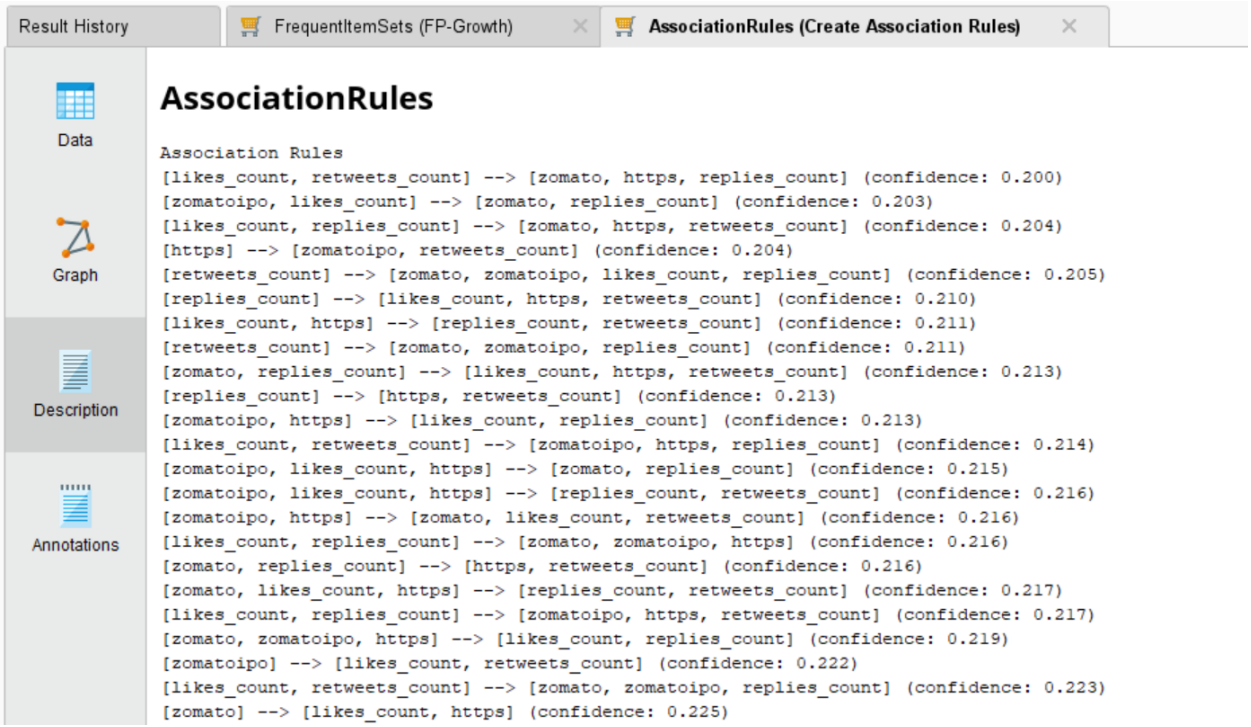
No.	Premises	Conclusion	Support	Confidence	LaPlace
79	zomatoipo, replies_count	zomato, retweets_count	0.042	0.282	0.907
82	zomatoipo, retweets_count	zomato, replies_count	0.042	0.286	0.909
85	zomatoipo	zomato, https	0.177	0.290	0.732
86	replies_count	zomato, likes_count, retweets_count	0.068	0.292	0.866
89	zomatoipo, replies_count	zomato, https	0.044	0.295	0.909
91	replies_count	zomato, retweets_count	0.070	0.300	0.868
92	zomatoipo, likes_count, retweets_count	zomato, replies_count	0.041	0.301	0.917
93	likes_count, replies_count	zomato, https	0.054	0.302	0.894
94	retweets_count	zomato, zomatoipo, likes_count, https	0.060	0.303	0.884
95	https, retweets_count	zomato, likes_count, replies_count	0.037	0.304	0.925
97	replies_count	zomato, zomatoipo, likes_count	0.071	0.305	0.869
98	https, retweets_count	zomato, replies_count	0.037	0.309	0.926
104	retweets_count	zomato, zomatoipo, https	0.064	0.322	0.888
107	likes_count, https, retweets_count	zomato, replies_count	0.037	0.324	0.931

The Association rule graph, how the word is dependent on other words.



The association rule description, which results out how much words are dependent on with a confidence level.

Jalendar Reddy Maligireddy
11511290
Summer-2022



The frequent item set:

It shows how many items the word appears; how much does it effect on different words.

Jalendar Reddy Maligireddy
11511290
Summer-2022

Result History | FrequentItemSets (FP-Growth) | AssociationRules (Create Associat

Data

Annotations

No. of Sets: 111

Total Max. Size: 5

Min. Size: 1

Max. Size: 5

Contains Item:

Update View

Size	Support	Item 1
1	0.748	zomato
1	0.608	zomatoipo
1	0.492	likes_count
1	0.465	https
1	0.233	replies_count
1	0.199	retweets_count
1	0.068	investors
1	0.064	zomatoindia
1	0.057	july
1	0.055	food
1	0.055	market
1	0.053	india
1	0.047	price
1	0.047	apply

By the above results we can conclude that most often used words on the tweet are ZOMATO, ZOMATOIPO, ZOMATOINDIA, JULY, etc..

k-means Cluster analysis:-

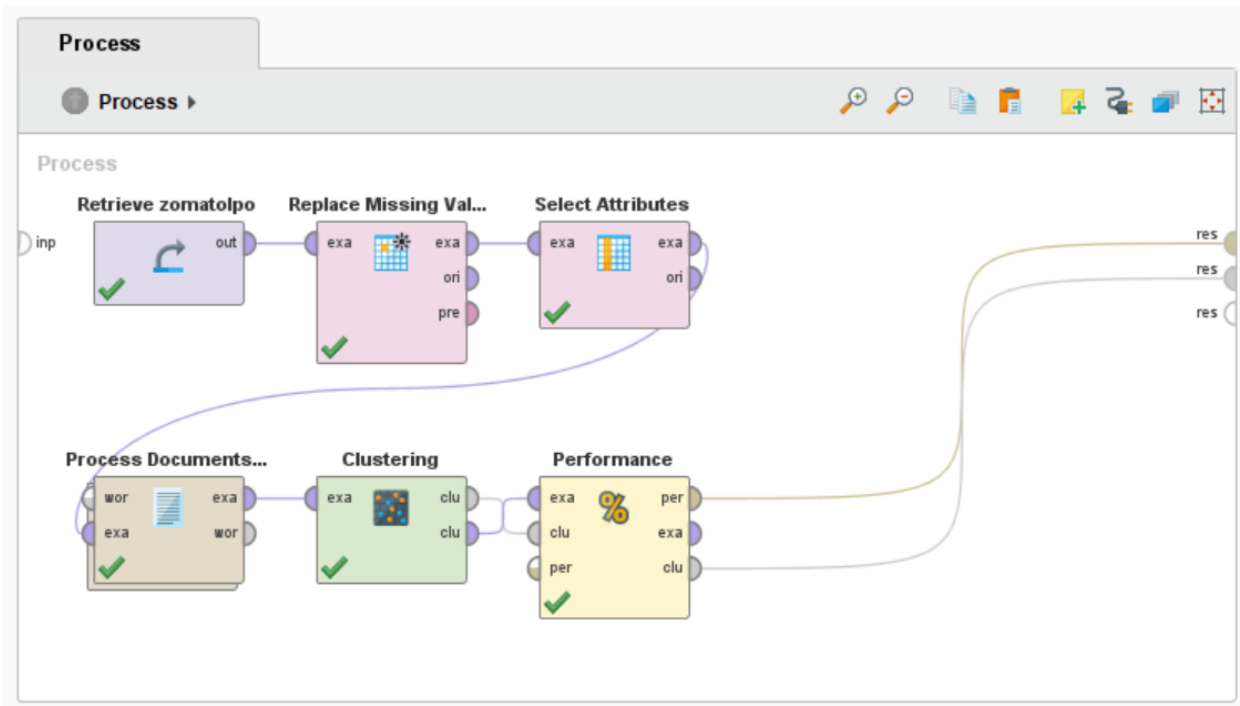
This administrator performs clustering utilizing the bit k-means calculation. Clustering is concerned with gathering objects together that are comparable to each other and disparate to the objects having a place to other clusters. Bit k-means employments bits to assess the separate between objects and clusters. K-means is an select clustering calculation.

This administrator performs clustering utilizing the part k-means calculation. The k-means is an select clustering calculation i.e. each protest is doled out to accurately one of a set of clusters. Objects in one cluster are comparable to each other. The similitude between objects is based on a degree of the remove between them. Part k-means employments parts to assess the remove between objects and clusters. Since of the nature of bits it is vital to entirety over all components of a cluster to calculate one remove. So this calculation is quadratic in number of cases and does not return a Centroid Cluster Show opposite to the K-Means administrator. This administrator

makes a cluster quality within the resultant Example Set in the event that the include cluster property parameter is set to genuine.

The process of clustering:

When k=4



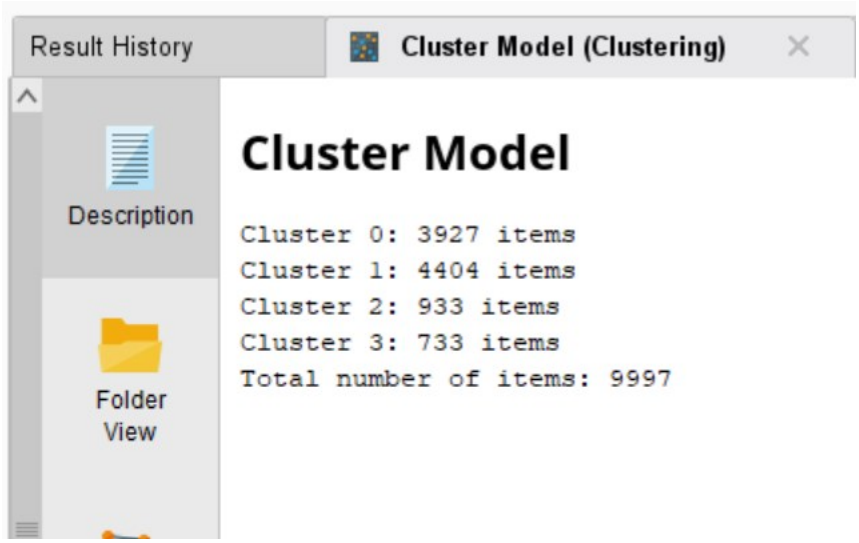
The total number of items are 9997 which are total number of rows in the data set.

Each cluster is divided the data in random way.

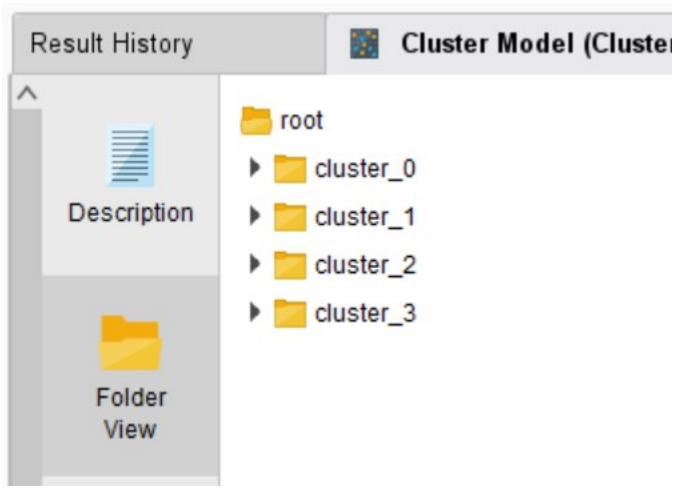
For example:

Cluster number	Number of data taken
0	3927
1	4404
2	933
3	733

Jalendar Reddy Maligireddy
11511290
Summer-2022

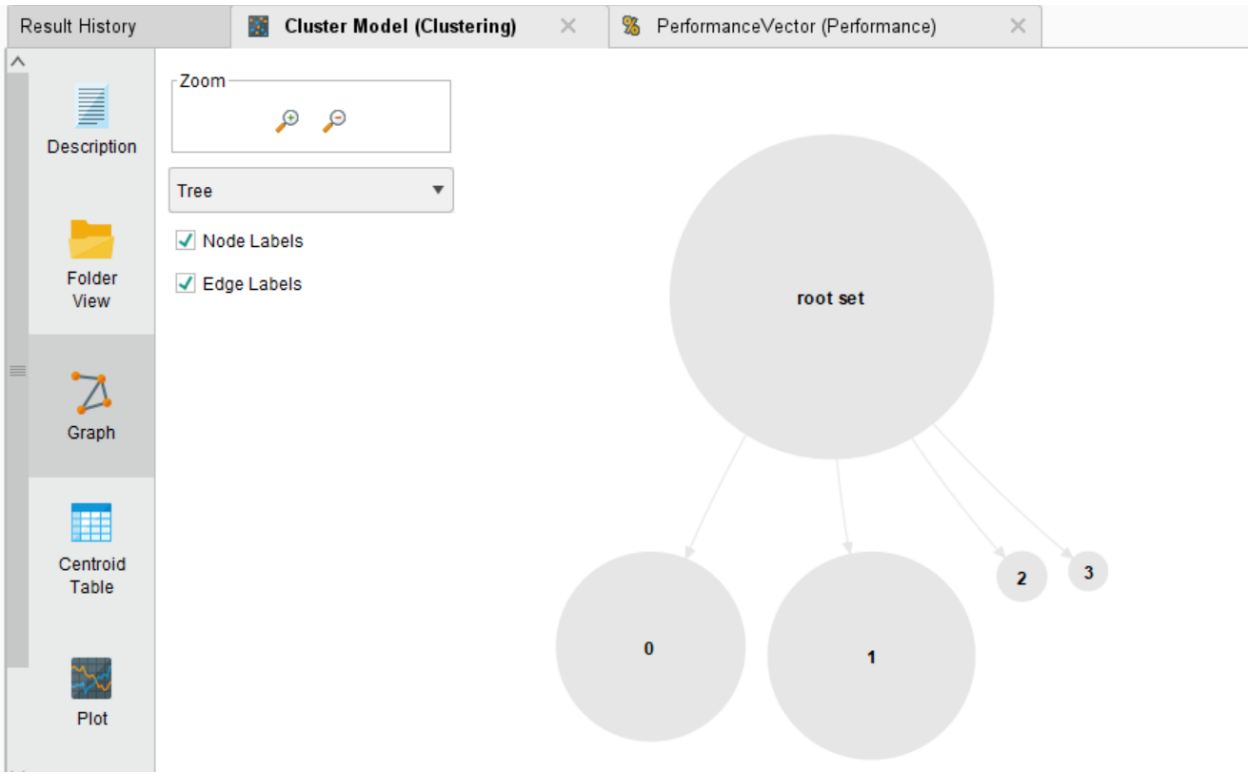


The folder view of cluster



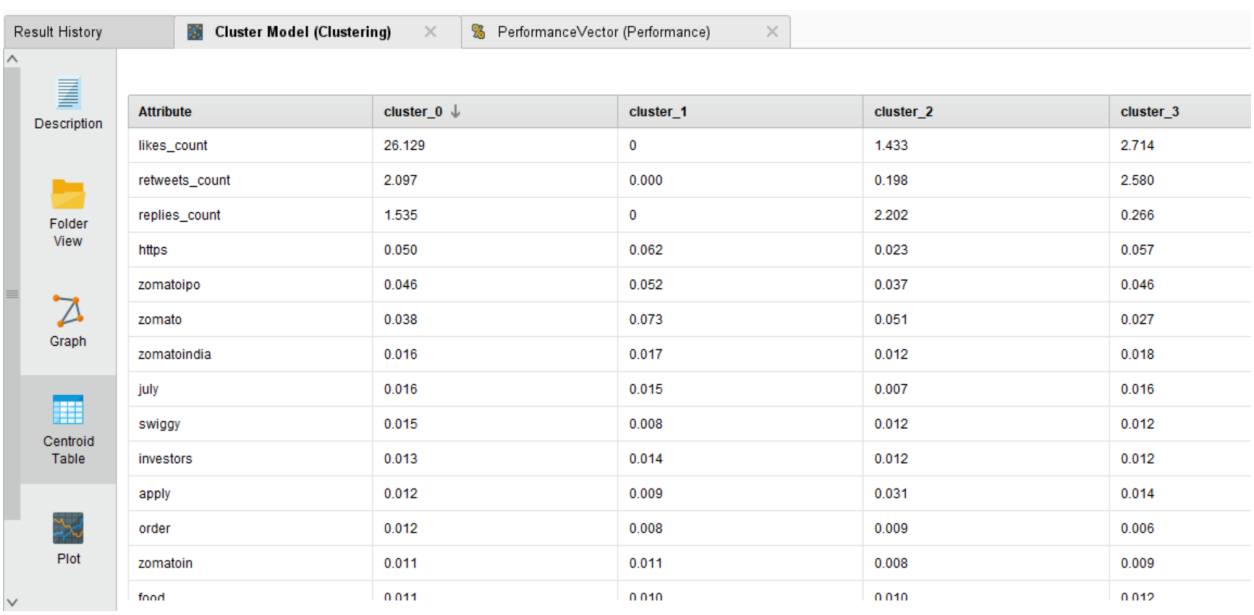
The graph of the cluster, when k=4 cluster 1 is larger than any other clusters, cluster 0 is almost near to cluster 1.

Jalendar Reddy Maligireddy
11511290
Summer-2022



The centroid table of the cluster.

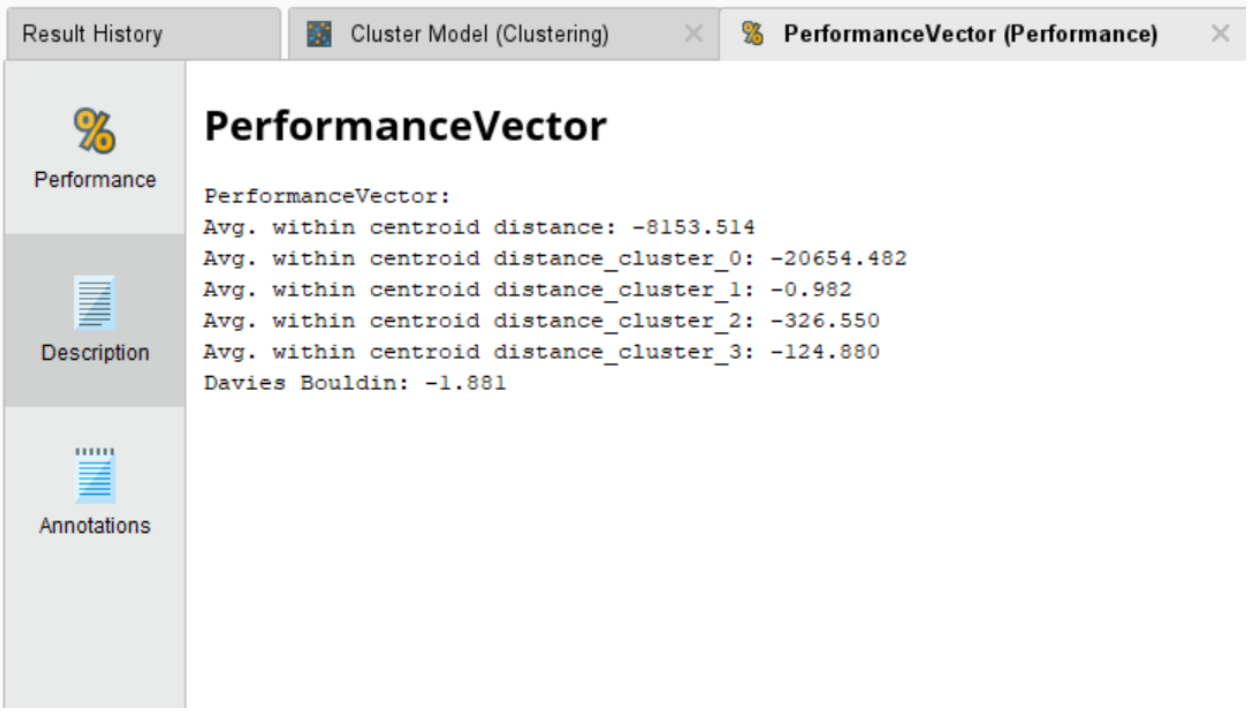
The attribute shows the words, clusters show how they appeared.



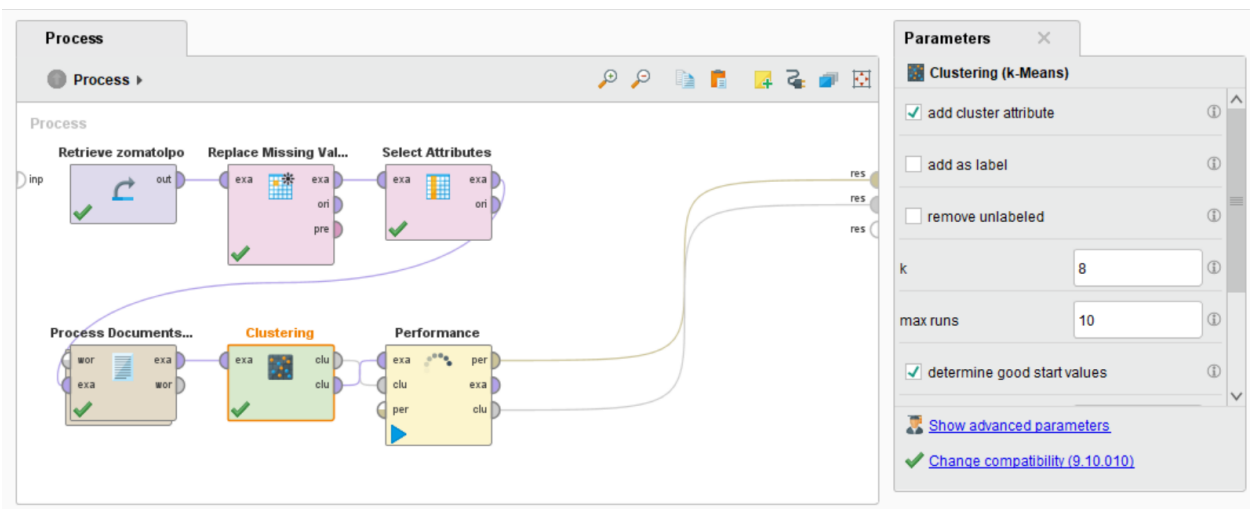
Attribute	cluster_0	cluster_1	cluster_2	cluster_3
likes_count	26.129	0	1.433	2.714
retweets_count	2.097	0.000	0.198	2.580
replies_count	1.535	0	2.202	0.266
https	0.050	0.062	0.023	0.057
zomatoipo	0.046	0.052	0.037	0.046
zomato	0.038	0.073	0.051	0.027
zomatoindia	0.016	0.017	0.012	0.018
july	0.016	0.015	0.007	0.016
swiggy	0.015	0.008	0.012	0.012
investors	0.013	0.014	0.012	0.012
apply	0.012	0.009	0.031	0.014
order	0.012	0.008	0.009	0.006
zomatoin	0.011	0.011	0.008	0.009
food	0.011	0.010	0.010	0.012

The performance vectors

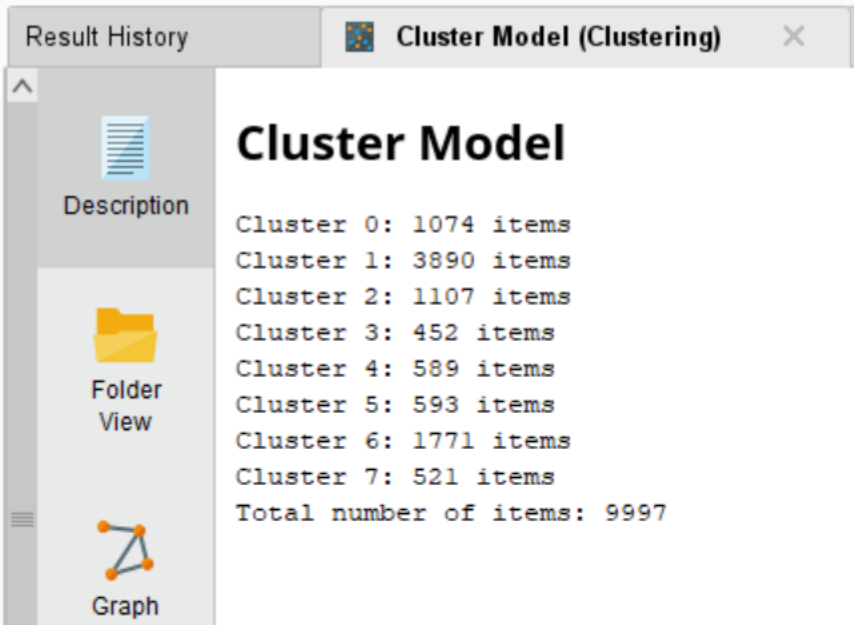
Jalendar Reddy Maligireddy
11511290
Summer-2022



Similarly for the cluster K=8



Jalendar Reddy Maligireddy
11511290
Summer-2022



Jalendar Reddy Maligireddy
11511290
Summer-2022

Result History | Cluster Model (Clustering) | PerformanceVector (Performance)

Performance
Description
Annotations

PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: -8154.866
Avg. within centroid distance_cluster_0: -0.922
Avg. within centroid distance_cluster_1: -20793.562
Avg. within centroid distance_cluster_2: -0.972
Avg. within centroid distance_cluster_3: -0.921
Avg. within centroid distance_cluster_4: -2.569
Avg. within centroid distance_cluster_5: -916.626
Avg. within centroid distance_cluster_6: -0.982
Avg. within centroid distance_cluster_7: -168.799
Davies Bouldin: -5.605
```

Result History | Cluster Model (Clustering) | PerformanceVector (Performance)

Description
Folder View
Graph
Centroid Table
Plot

Attribute	cluster_0	cluster_1 ↓	cluster_2	cluster_3	cluster_4	cluster_5	cluster_6	cluster_7
zomatoindia	0.000	0.017	0.057	0.004	0.008	0.014	0.004	0.016
july	0	0.017	0.000	0.010	0.006	0.011	0.035	0.015
swiggy	0.001	0.015	0.031	0	0.014	0.012	0.001	0.012
investors	0.003	0.014	0.002	0.094	0.010	0.013	0.007	0.011
apply	0.004	0.012	0.013	0.004	0.037	0.018	0.012	0.015
order	0.001	0.012	0.008	0.001	0.009	0.013	0.015	0.005
food	0.000	0.012	0.004	0.002	0.010	0.010	0.022	0.011
zomatoin	0	0.011	0.038	0.003	0.009	0.010	0.002	0.008
deepigoyal	0.001	0.011	0.020	0.001	0.004	0.007	0.002	0.008
market	0	0.011	0.001	0.004	0.009	0.010	0.023	0.009
indian	0.000	0.011	0.002	0.004	0.004	0.006	0.013	0.008
price	0.002	0.010	0.001	0.004	0.008	0.007	0.027	0.008
know	0.000	0.010	0.003	0.002	0.005	0.006	0.023	0.009

Jalendar Reddy Maligireddy
11511290
Summer-2022

Conclusion of the analysis (what did you find out?):-

a. I had done 3 analysis on the dataset
    - Correlation analysis
    - Association analysis
    - k-means Cluster analysis
b. I had selected k values for two different sizes.
    - K=4
    - K=8
c. When k=4, the cluster is divided into 9 parts, where average centroid distance cluster is -8154.866. where average centroid distance for k=4 is -8153.514.
d. We can observe from above clusters when k=4 cluster 1 is larger and almost cluster 0 is also same, similarly when k=8 cluster 1 is larger.
e. From the observation we got know that these cluster people had tweets a lot.
f. From the above two different k values, each had different significance:
    - When k=4, the cluster 1 had a highest of 0.073 & cluster 0 had 26.129.
    - When k=8, the cluster 1 had a highest of 26.125.
g. By all the observation made, finally we can conclude that, people are going to invest on ZOMATO IPO might give profit in future. It's a good for investors.

Jalendar Reddy Maligireddy
11511290
Summer-2022

Reference:

https://twitter.com/Gautam__Baid/status/1415338575269228549

https://twitter.com/zomato