2018

Case Study 2

MINING FROM CLOTHING, SUPERMARKET, NEWS STORIES AND WEB LOG DATA

GROUP 1

9449949 - RICK MCCAKSER

9819363 - BRENDAN DONCASTER

Contents

Part 1 Clustering

Task 1: Data Preparation for Clustering

1. Can you identify data quality issues in this dataset such as unusual data types, missing values and others?

Annual Sales, Sales and TotalInvestment all suffered from NaN values, however a relation was able to be made between these and the stores with a floor size of 0. These were likely errors and given that they would damage clustering outcomes, each of the stores matching this criterion were removed.

2. What variables did you include in the analysis and what were their roles and measurement level set? Justify your choice.

They only field that was dropped from the data set was the store code as it is unique identifier and is therefore not able to provide meaningful patterns and relationships.
SFloorSize is a continuous value that was set to integer. Given that it is an uncapped integer following no interval between values, we felt setting it as continuous was justifiable.
AnnualSales, Sales and TotalInvestment were all extremely similar types of fields and as such they were all set to floats and were recognized as continuous fields. Given that an amount of money doesn't have any sort of interval but is a measure of wealth, setting as continuous was justifiable.

3. Identify a store that is underperforming in sales. Based on your reporting, the company does not want to focus their efforts on this store. Now onwards, the selected store should not be part of analysis.

Store with StoreCode 287 has a Sales value of 300 which is 2.8 times worse than the second lowest performing store.

Task 2: The First Clustering Model

1. Build a default clustering model with K= 3 and answer the followings:

a. How many records are assigned into each cluster?

```
Number of items per cluster:
0    212
2    140
1     32
```

b. Plot the cluster distribution using pair plot. Explain key characteristics of each cluster/ segment.

Refer to appendix 1 for the pair plot image.

Blue (cluster 0) represents stores with a floor size smaller than 500 and annual sales less than 1000000.

Orange (cluster 1) represents stores with annual sales roughly about >= 200000

Green (cluster 2) represents stores with average annual sales

2. What is the effect of using the standardization method on the model above? Does the variable normalization process enable a better clustering solution?

Refer to appendix 2 for the scaled pair plot image.

Standardizing the datasets allows for a better clustering solution. This is demonstrated by the pair plot show the more complex boundaries for each cluster, rather than a straight line as was used in the default approach.

Blue (cluster 0) now represents stores that an underperforming in sales (given the floor size they have), but still have decent annual sales, making them potential targets for an efficiency improvement. Outliers in this cluster have also had more total investment than other stores.

Orange (cluster 1) shows stores that are over performing, with higher sales, lower floor size and higher annual sales on average than other stores.

Green (cluster 2) shows stores that are small, have low sales and have a lower than average annual sales.

3. Interpret the (2.2) cluster analysis outcome. In other words, characterize the nature of each cluster by giving it a descriptive label by using distplot.

Refer to appendix 3 for a distribution graph for more information on these clusters.

Cluster 0 | Average Performance/Low Performance:

Average/low performing stores that don't have any particularly outstanding achievements.

Cluster 1 | Holiday Season/Rush period prone:

High sales with slightly above average annual sales. Likely subject to high load during busy periods (such as holidays) with average normal periods.

Cluster 2 | High Load/Consistently busy:

Stores that see a lot more income but have the floor size to accommodate, reflected by the above average annual sales and floor size but average sales.

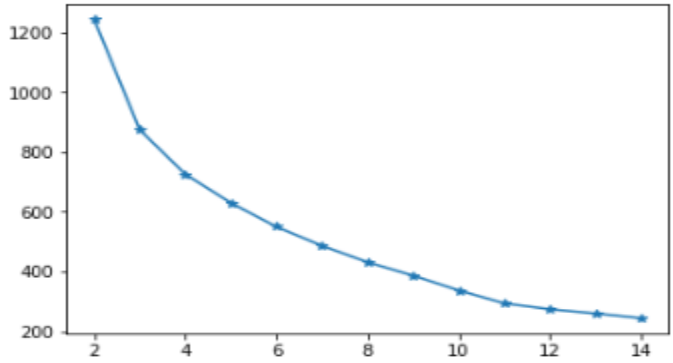
Cluster 3 | Requires reconsideration of assets:

A store that is performing around average but has larger floor size and a large amount of investment put into it, presents a possible experiment of money leak.

Task 3: Refining the clustering model

1. Using elbow method and silhouette, find the optimal K. What is the best K? Explain your reasoning. Evaluate the result.

Given that the standardised dataset provided better results in the clustering model, the elbow and silhouette strategies will be performed on it.



The elbow method shows that the optimal k is likely to be either 3 or 4 as they appear to have the highest bends or 'elbows' on the graph. However, since the result is not obvious the silhouette strategy will need to be used to refine it.

```
KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
    n_clusters=3, n_init=10, n_jobs=10, precompute_distances='auto',
    random_state=10, tol=0.0001, verbose=0)
Silhouette score for k=3 0.4991260137091022

KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
    n_clusters=4, n_init=10, n_jobs=10, precompute_distances='auto',
    random_state=10, tol=0.0001, verbose=0)
Silhouette score for k=4 0.5050002970308124
```

The silhouette strategy shows that k=3 and k=4 have very similar scores, however, k=4 is the more optimal value.

2. What is the best number of clusters that can describe the dataset effectively? Was this obtained with the default setting (i.e. the automated process) or manually specifying a clustering number?

As mentioned the ideal n-clusters is most likely 4. Given that the default was done with 3, the original default solution was close to providing a good model, however it had the capability to be further refined. Overall, manually finding and specifying the default number of clusters provided a better result. Refer to appendix 4 for a pair plot image of this refined cluster.

3. How the outcome of this study can be used by decision makers?

See which stores are underperforming
See which stores are utilising or under utilising their store size
See which stores require more attention regarding investment, either positively (needs more investment) or negatively (needs less investing)
See which stores might require more stock during busy periods (such as cluster 1)
See which stores are being poorly managed.

Part 2 Association Mining

Task 4: Association Mining

1. Can you identify data quality issues in this dataset for performing association analysis?

Quantity provides arbitrary data, which is one in all vars. This is completely useless and provides nothing to the problem and should be dropped as a result.
Transaction date could be useful if the desire was to find which days attract which kind of customers, however the focus of this model should be towards which items are bought together.

2. What variables did you include in the analysis and what were their roles and measurement level set? Justify your choice.

Transaction_ID - Interval input variable, used to group the products purchased in the same transaction to find patterns.
Product_Name - Nominal input variable, describes the product that is purchased and is the main focus for association.

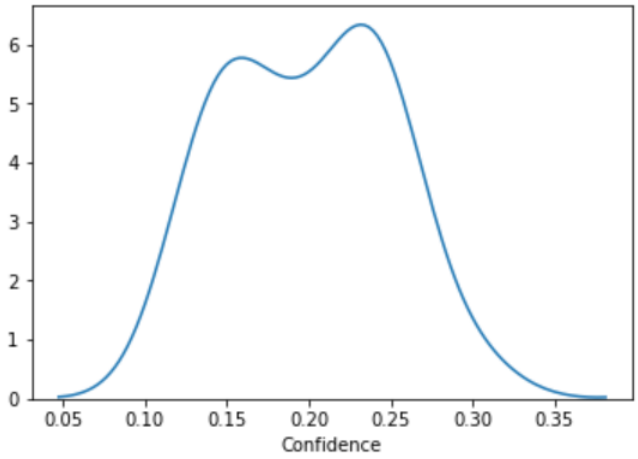
3. Conduct association mining and answer the following:

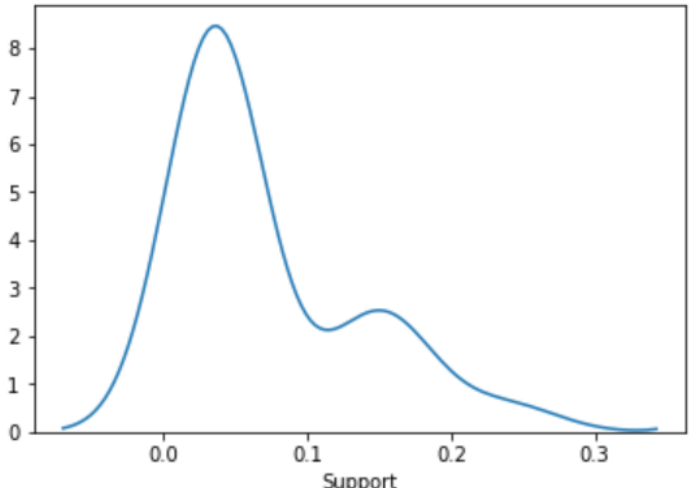
a) What is the highest lift value for the resulting rules? Which rule has this value?
The rule DVD media -> Exercise book has the highest lift of 1.738191

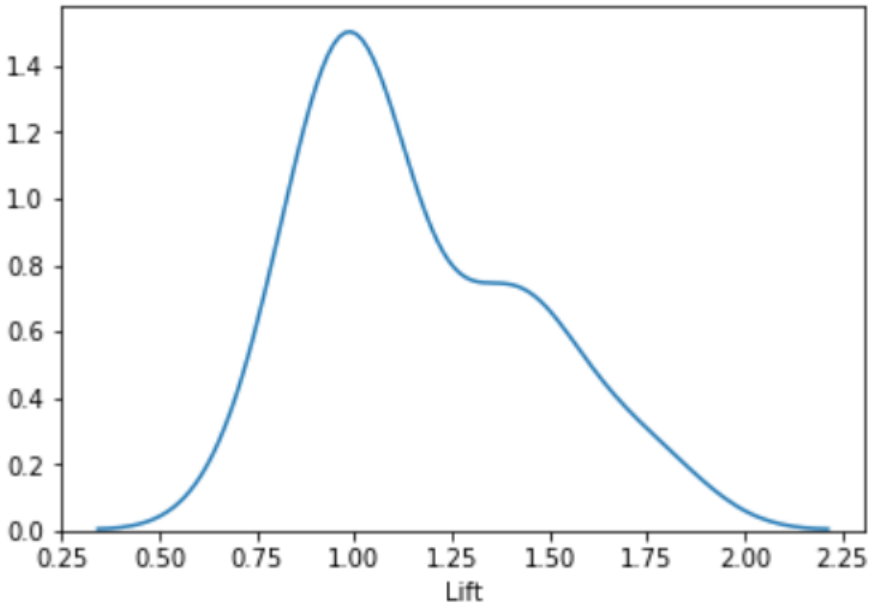
b) What is the highest confidence value for the resulting rules? Which rule has this value?
The rule DVD media -> Exercise book has the highest confidence of 0.297239

c) Plot the confidence, lift and support of the resulting rules. Interpret them to discuss the rule-set obtained.

As seen from the graphs and the tables, some of the rules do not supply information on useful relations, these mainly being the cases where items are bought on their own. Naturally these have high support and confidence but have a lift of 1, reducing their value to someone looking for relations. The rules with higher lift demonstrate more popular combinations which should be targeted and the rules with lower lift show item combinations that are still targetable but should be of a lower priority. Refer to appendix 5 for a list of the rules found.







4. The store is particularly interested in products that individuals purchase when they buy "Exercise book".

a) How many rules are in the subset?

4

	Left_side	Right_side	Support	Confidence	Lift
7	Exercise book	DVD media	0.043660	0.255314	1.738191
12	Exercise book	Flash Card	0.039780	0.232625	1.450053
14	Exercise book	Lanyards	0.033015	0.193065	1.430903
16	Exercise book	Sketching Markers	0.040535	0.237040	0.982325

b) Based on the rules, what are the other products these individuals are most likely to purchase?

DVD media, Flash Card, Lanyards and Sketching Markers

5. How the outcome of this study can be used by decision makers?

It can be used to choose how to more optimally position product locations within a shop, as placing them closer to products often bought together, they are more likely to be seen and then the customer may then realise they want to purchase that product as well.

Part 3: Text Mining (Clustering) the News Stories

Task 5: Text Mining

1. What variables did you include in the analysis and what were their roles and measurement level set? Justify your choice.
The only variable we used from this data set is the text associated with each document. On its own this text is not able to represent a measurement level but when it is processed for the model use, the string is bound to an interval representing how many times it occurs.

2. Can you identify data quality issues in order to perform text mining?
No data quality issues were present in the data

3. Based on the ZIPF plot, list the top 10 terms that will be least useful for clustering purpose.
Refer to Appendix 6 for the full ZIPF plot.

High Document Frequency/Low Term Frequency:
Say
Year
High Term Frequency/Low Document Frequency:
Trust india
Real venture
Winner believe
Lose people
Rainer schuettler
Pollard say
Get Three
Early Save

4. Did you disregard any frequent terms? Justify your selection.
The items given early for being less useful overall were disregarded despite being frequently

5. Justify the term weighting option selected.
Minimum document frequency of 2 was selected so that all terms which only existed solely in a single document would be removed, this is because a term only existing in one document cannot help to group together different documents
Maximum document frequency of 70% was selected to shrink the bounds of outlier items that appeared over so many documents that they could be considered stop words in this domain such as ‘say’.

6. What is the number of input features available to execute clustering? (FYI: Note how the original text data is converted into a feature set that can be mined for knowledge discovery.)

Initially there was a total of 36385 features available to use for clustering however using zipf to identify outliers and Tfidf vectorization to remove less useful features this input feature set was reduced to 6922.

7. State how many clusters are generated? Name each cluster meaningfully according to the terms that appear in the clusters?

7 clusters were generated:
Andy Roddick Australian open final tennis match
Kenteris and Thanou IAAF drug violations
South Africa cricket test match
Chelsea vs Liverpool soccer match
Kelly Holmes, Britain Olympic competitor
Paula Radcliffe, World cross country silver medalist
Players for the England world cup

8. Identify the first fifteen high frequent terms (that are not stop words or noise) in the start list?

Year
Play
Win
England
Game
First
World
6
Take
Cup
Get
Would
Match
Player
Final

9. Describe how these clusters can be useful in the online personalised news story service planned

They can be used to recommend/show users other news stories that are based on the same topic as the one they are reading/have been reading.

Part 4: Web Mining the Log Data for a Website

Task 6: Web Mining

a. Rationale behind selecting the data mining method.

The data mining method we decided to use was association rule mining.
Initially, clustering was considered as it is capable of combining a set of users and creating specialized groups that could then be targeted for particular deals or specials. Combined with a classification model this could be flexible enough to accept incoming changes. However, given the data set and a lack of understanding of the domain itself, this could potentially generate bad rules and incorrect recommendations as we had no idea how to identify the clusters themselves. Association mining however is able to provide an easily understandable set of rules without separation into clusters and gives information on which sections of the site should be targeted primarily. Therefore, in a given case where we had a domain expert and perhaps location data, a clustering classification model would be far superior however, in this case an association model is a safer option.

b. What variables did you include in the analysis and what were their roles and measurement level set? Justify your choice.

Variables:
user_id - Interval input variable, used to group the web pages visited by the same user to find frequent sequential patterns.

Request - Nominal input variable, describes the web pages that the user visited and is the main focus for association.

c. Can you identify data quality issues in order to perform web mining?

Web log data with requests favico.ico and robots.txt were removed.
Robots.txt was removed as this file responds to web scraping robots that are an automated service. Given the nature of this it could potentially skew data with results that aren't relevant to the target desired.
Favico.ico is a favourites image icon and is displayed in almost all cases where a user accesses the site, making it redundant data that should not be of interest to the client.

d. Discuss the results obtained. Discuss also the applicability of findings of the method. You should include only a high-level managerial kind of discussion

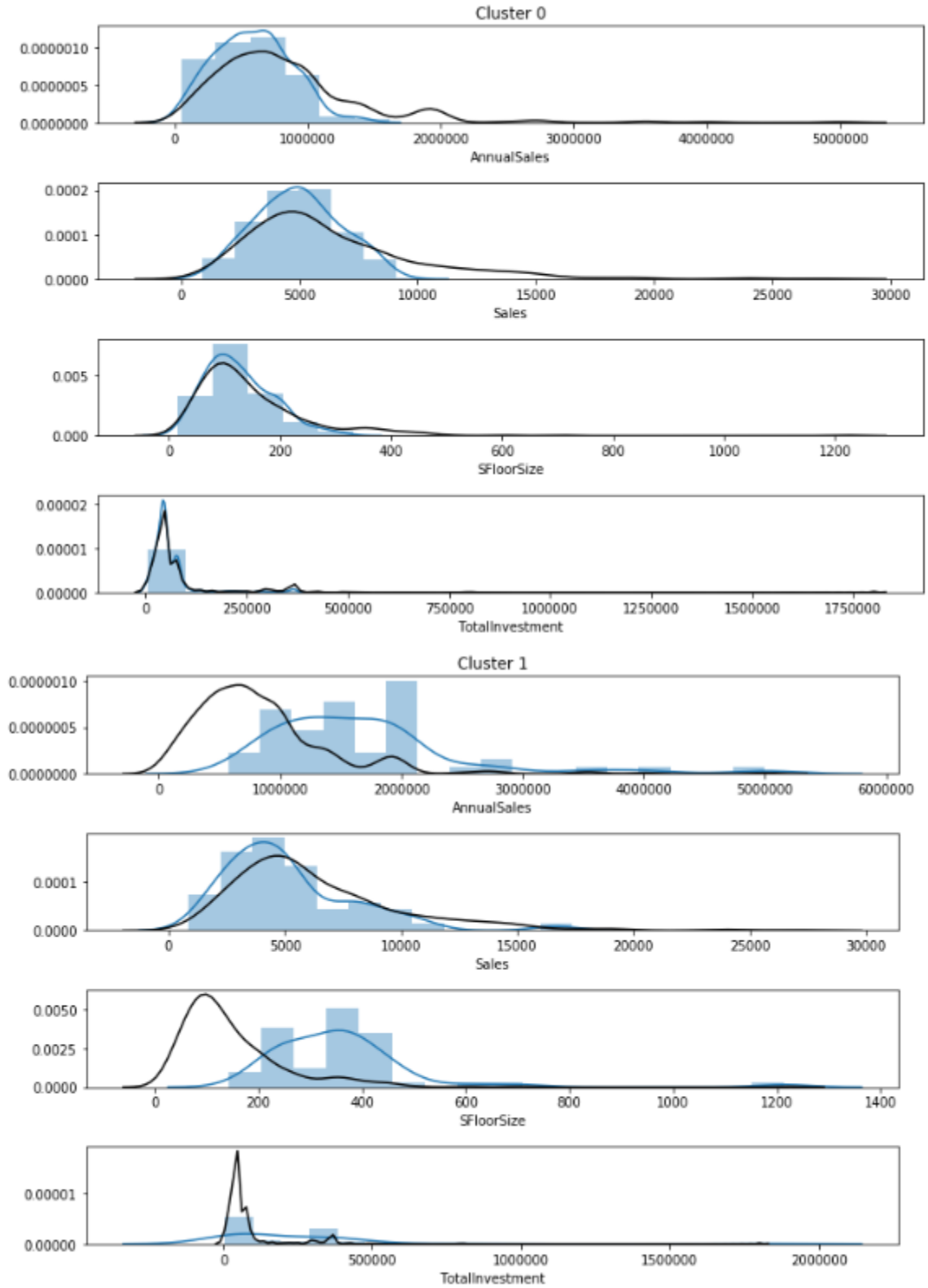
Users who access the richlands site have shown that they rarely navigate elsewhere besides the menu.js function. Assuming this function is just a menu bar or list of some sort, users aren't actually accessing the price list so either the data displayed on the homepage is already ideal or the site isn't attracting users to access the price list.
The new farm price list is being accessed in large amounts and from many different pages, therefore it may be worth improving accessibility of the pricelist or displaying extra important information on the pricelist page.
It should also be noted that a large number of users accessing the specials section of the new farm site are accessing it directly. It is probably safe to assume that some external resource is pointing users to it and given its success it should be implemented for other region pages if possible.
The users who interact with new farm and richlands have overwhelming traffic compared to the other regions and as a result sufficient rules could not be generated for these regions. It is likely that these regions are not being advertised properly or presence in these regions is weak. It should be considered to either abandon these regions or put more money into advertisement.

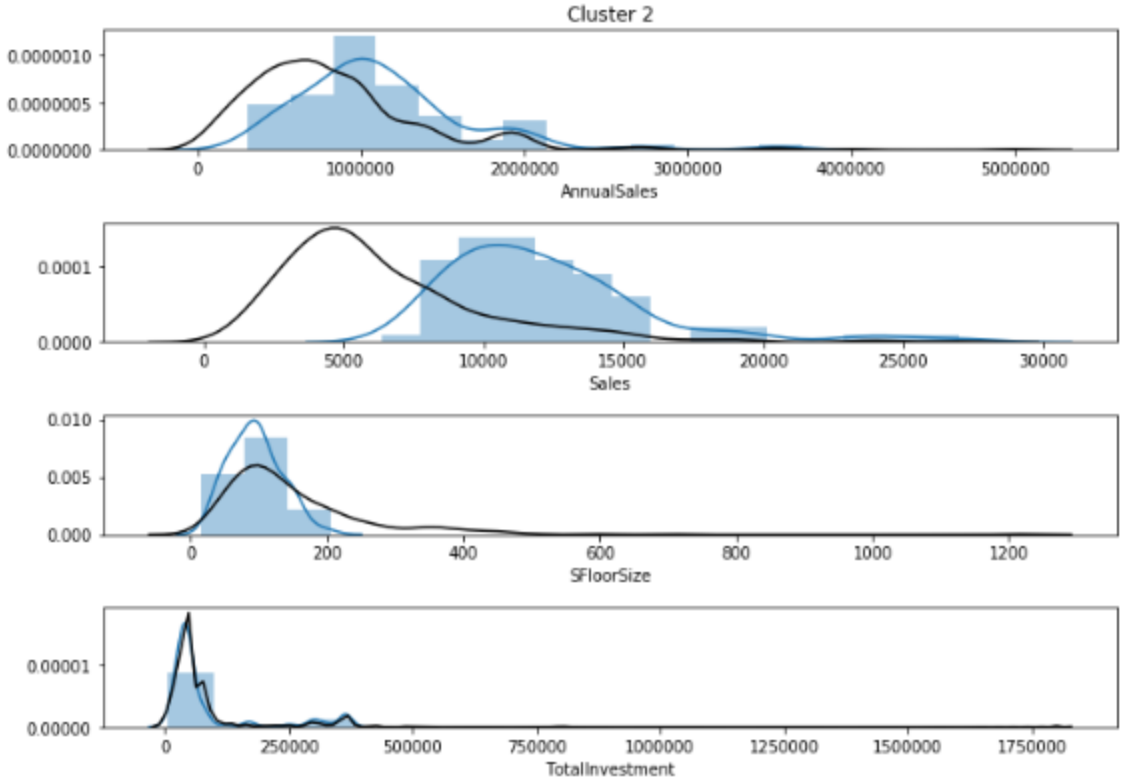
Appendix

Appendix 1: Pair Plot on a default clustering model (k=3)

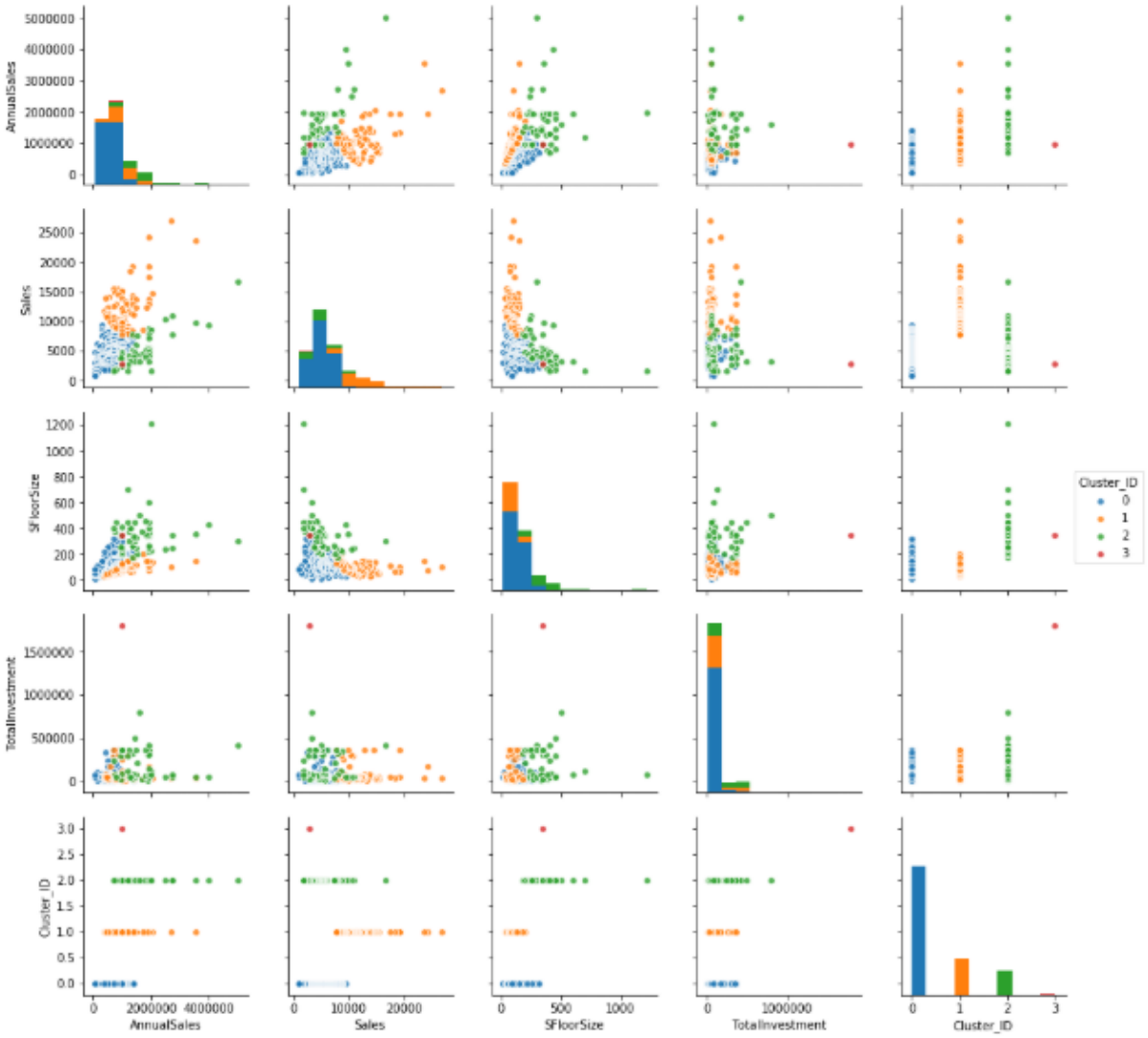
Appendix 2: Pair Plot on a standardised clustering model (k=3)

Appendix 3: Distribution model for the scaled clustering model




Cluster 2
0.0000010
0.0000005
0.0000000
0
1000000
2000000
3000000
4000000
5000000
AnnualSales
0.0001
0.0000
0
5000
10000
15000
20000
25000
30000
Sales
0.010
0.005
0.000
0
200
400
600
800
1000
1200
SFloorSize
0.00001
0.00000
0
250000
500000
750000
1000000
1250000
1500000
1750000
TotalInvestment

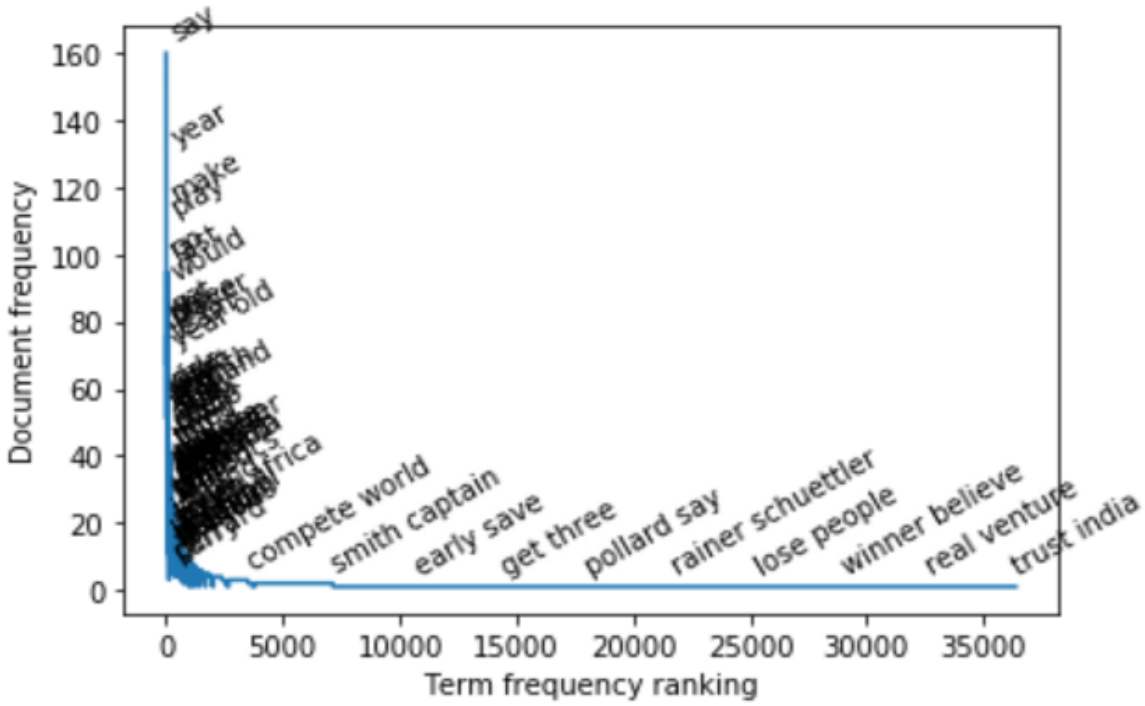
Appendix 4: Pair plot image of the K=3 cluster



Appendix 5: List of rules generated from association mining ordered by lift.

	Left_side	Right_side	Support	Confidence	Lift
6	DVD media	Exercise book	0.043660	0.297239	1.738191
7	Exercise book	DVD media	0.043660	0.255314	1.738191
12	Exercise book	Flash Card	0.039780	0.232625	1.450053
13	Flash Card	Exercise book	0.039780	0.247966	1.450053
14	Exercise book	Lanyards	0.033015	0.193065	1.430903
15	Lanyards	Exercise book	0.033015	0.244691	1.430903
9	Flash Card	DVD media	0.032080	0.199969	1.361397
8	DVD media	Flash Card	0.032080	0.218402	1.361397
10	DVD media	Sketching Markers	0.036335	0.247370	1.025136
11	Sketching Markers	DVD media	0.036335	0.150577	1.025136
3		Lanyards	0.134925	0.134925	1.000000
4		Sketching Markers	0.241305	0.241305	1.000000
1		Exercise book	0.171005	0.171005	1.000000
2		Flash Card	0.160425	0.160425	1.000000
0		DVD media	0.146885	0.146885	1.000000
5		Wristbands	0.143575	0.143575	1.000000
16	Exercise book	Sketching Markers	0.040535	0.237040	0.982325
17	Sketching Markers	Exercise book	0.040535	0.167982	0.982325
20	Lanyards	Sketching Markers	0.031630	0.234427	0.971495
21	Sketching Markers	Lanyards	0.031630	0.131079	0.971495

Appendix 6: Zipf plot



Appendix 7: List of clusters generated for text mining.

```
Top terms for cluster 0: 6, roddick, 7, open, final,
Top terms for cluster 1: kenteris, greek, thanou, iaaf, test,
Top terms for cluster 2: wicket, south, south africa, africa, test,
Top terms for cluster 3: liverpool, club, gerrard, league, chelsea,
Top terms for cluster 4: holmes, athletics, olympic, britain, conte,
Top terms for cluster 5: radcliffe, marathon, chepkemei, race, cross country,
Top terms for cluster 6: england, cup, play, game, player,
```

Appendix 8: List of rules generated using association for the web mining task.

	Left_side	Right_side	Support	Confidence	Lift
57	/richlands/	/richlands	0.043083	0.910256	18.988640
56	/richlands	/richlands/	0.043083	0.898734	18.988640
109	/newfarm/javascript/menu.js,/newfarm/specials/	/newfarm/specials	0.035194	0.753247	18.808343
59	/richlands/javascript/menu.js	/richlands/	0.036408	0.869565	18.372352
58	/richlands/	/richlands/javascript/menu.js	0.036408	0.769231	18.372352
64	/newfarm/,/	/newfarm	0.035194	0.865672	16.983653
108	/newfarm/javascript/menu.js,/newfarm/specials	/newfarm/specials/	0.035194	0.983051	16.701730
54	/newfarm/specials	/newfarm/specials/	0.038228	0.954545	16.217432
55	/newfarm/specials/	/newfarm/specials	0.038228	0.649485	16.217432
31	/newfarm/contact/	/newfarm/contact	0.043083	0.797753	14.939734
30	/newfarm/contact	/newfarm/contact/	0.043083	0.806818	14.939734
125	/newfarm/pricelist,/newfarm/javascript/menu.js...	/newfarm/pricelist/	0.086772	1.000000	9.526012
129	/newfarm/pdf/Web_Price_List.pdf,/newfarm/price...	/newfarm/pricelist/	0.076456	1.000000	9.526012
90	/newfarm/pricelist,/newfarm/	/newfarm/pricelist/	0.087379	1.000000	9.526012
121	/newfarm/pdf/Web_Price_List.pdf,/newfarm/javas...	/newfarm/pricelist/	0.075850	1.000000	9.526012
138	/newfarm/javascript/menu.js,/newfarm/pricelist...	/newfarm/pricelist/	0.037621	1.000000	9.526012
141	/newfarm/pdf/Web_Price_List.pdf,/newfarm/price...	/newfarm/pricelist/	0.075850	1.000000	9.526012
87	/newfarm/pdf/Web_Price_List.pdf,/newfarm/	/newfarm/pricelist/	0.076456	1.000000	9.526012
96	/newfarm/pdf/Web_Price_List.pdf,/newfarm/javas...	/newfarm/pricelist/	0.086165	0.986111	9.393706
133	/newfarm/pdf/Web_Price_List.pdf,/newfarm/javas...	/newfarm/pricelist/	0.085558	0.986014	9.392781
99	/newfarm/javascript/menu.js,/newfarm/pricelist	/newfarm/pricelist/	0.097087	0.975610	9.293670
111	/newfarm/pdf/Web_Price_List.pdf,/newfarm/price...	/newfarm/pricelist/	0.088592	0.973333	9.271985
117	/newfarm/pdf/Web_Price_List.pdf,/newfarm/javas...	/newfarm/pricelist	0.075850	1.000000	9.258427
126	/newfarm/pricelist/,/newfarm/javascript/menu.j...	/newfarm/pricelist	0.086772	1.000000	9.258427