Case Study 2

CAB 330

Jack Farrel – N8640866

Anton Polison – N9907637

Dennis Tran – N9937072

2017

Table of Contents

[Task 1: Data Selection and Distribution. 2](#_Toc492835509)

[Task 2: Predictive Modelling Using Decision Trees 5](#_Toc492835510)

[Task 3: Predictive Modelling Using Regression 8](#_Toc492835511)

[Task 4: Predictive Modelling Using Neural Networks 11](#_Toc492835512)

[Task 5: Comparing Predictive Models 13](#_Toc492835513)

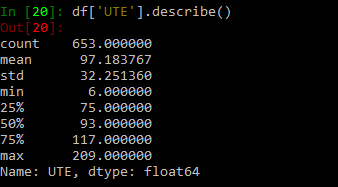
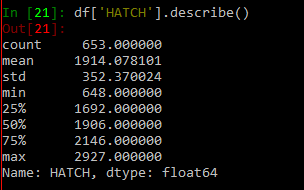
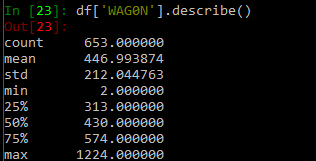
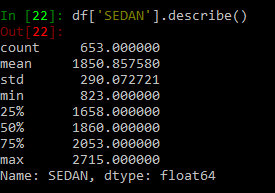
[Appendix A 16](#_Toc492835514)

# Task 1: Descriptive Data Mining - Clustering.

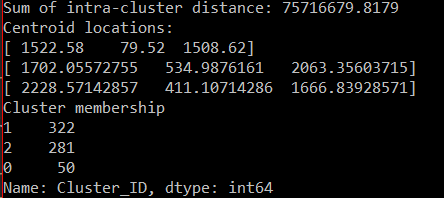
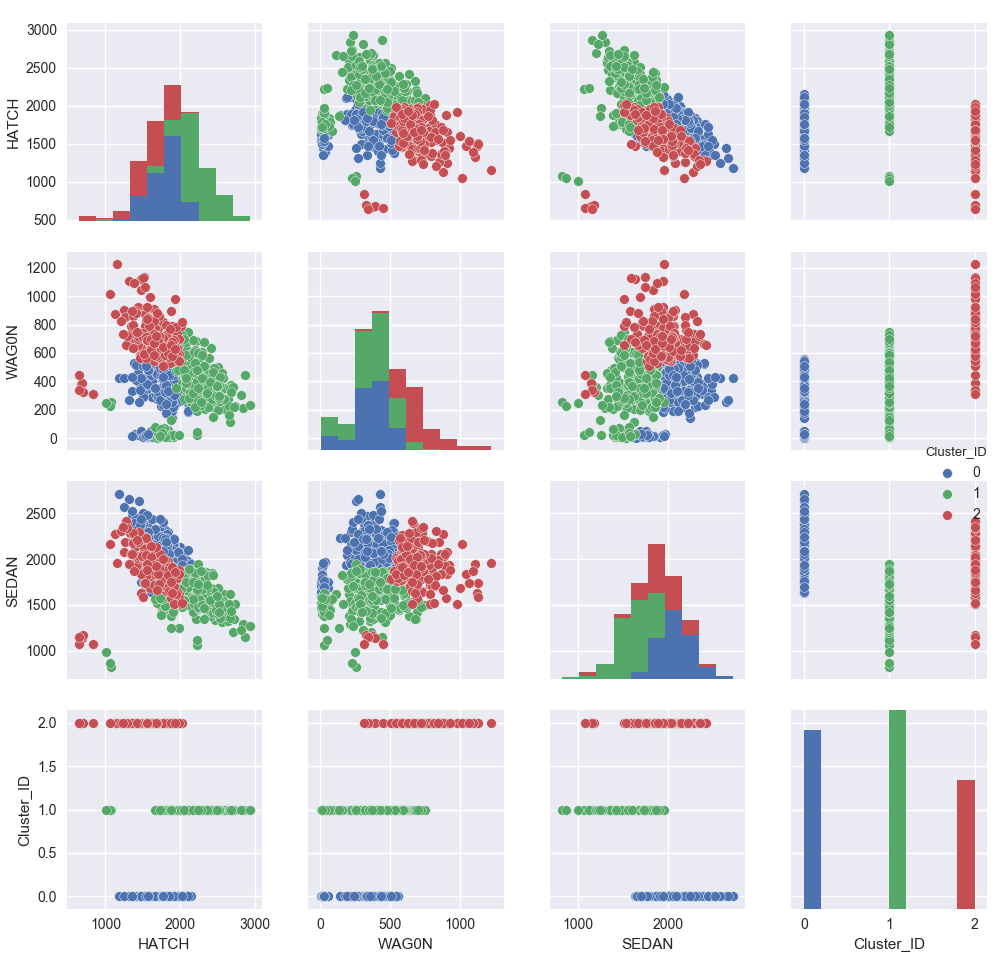
1. There were some changes made to the original dataset to ensure that the information from the analysis was accurate. There are several quality problems with “model\_car\_sales.csv” data set such as missing values and redundant or irrelevant variables
   1. ***Missing Values:*** The dataset contains 22 rows which have empty values. These values were deleted as they are irrelevant for our research.
   2. ***Dropped Variables:*** In total 4 variables were dropped due to various reasons. The below table shows the dropped variables and their corresponding reasons.

|  |  |
| --- | --- |
| *Variable* | *Reason For Drop* |
| *REPORT\_DATE* | Irrelevant. The variable has the date when the report was generated and doesn’t provide any useful information. |
| *DEALER\_CODE* | Redundancy. “LOCATION” variable provides more organised information which is at the same time easier to work with as it is numerical variable. |
| *K\_SALES\_TOT* | Irrelevant. The company wants to find the minimum number of product sale segments and it is not interested in identifying what is a number of total car sales. |

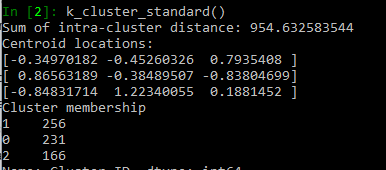
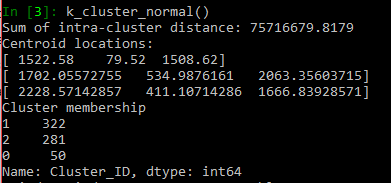
1. After preprocessing and analyzing the dataset it was identified that “UTE” is clearly underperforming in sales in comparison with “HATCH”, “SEDAN” and “WAGON”.

Form the tables above we can determine that “mean” of the variables “HATCH”, “WAGON”, “SEDAN” is significantly higher than of “UTE” variable.

1. “HATCH”, “SEDAN”, “WAGON” and “LOCATION” variables were included in the analyses. These were the variables for the three best performing cars and the locations of the dealerships. These variables were chosen as they best fit the task of determining the optimum number of sale segments. By looking at sales data per location, the analysis can determine the best results.  
     
     
     
   1.   
      From cluster membership table we can identify that 231 records were assigned to **cluster 0**, 256 records to **cluster 1** and 166 records to cluster 2.
   2. The pairplot shows the sales distribution on different variables.

* If we look at the first row first column we can notice that **cluster 0** has average to high sales. cluster 2 has low to average sales and finally, **cluster 1** has high sales.
* Wagon variable distribution is the following: **cluster 0** has low to average sales, cluster 2 has average to high sales and **cluster 1** has low to average sales with a slight win over **cluster 0.**
* SEDAN variable has the following distribution: **cluster 0** has high sales, **cluster 1** has low to average sales and cluster 2 average to high sales.

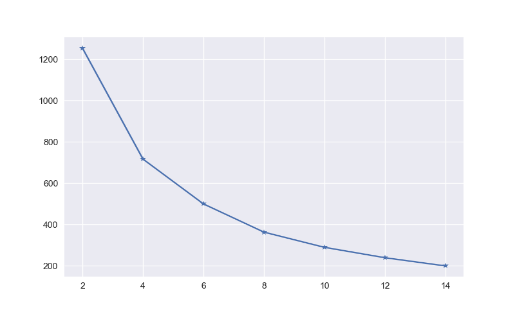
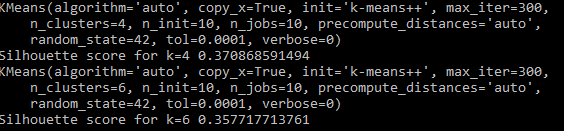
 

Normal distribution Standardization

The standardization made dramatic changes to our model. First and most obvious change is that intra-cluster distance was significantly reduced by standardization which is one of the main purpose of clustering. Furthermore, cluster membership changed as well “cluster 0” 50 =>231, “cluster 1” 322=>256, “cluster 2” 281 => 166.

If we look at the “pairplots” we can see 2 absolutely different results due to increased “cluster 0” membership and changed intra-cluster distance.

As we can see, all the above-mentioned facts proving the fact that standardization brings better clustering solution and better visualization outcome.

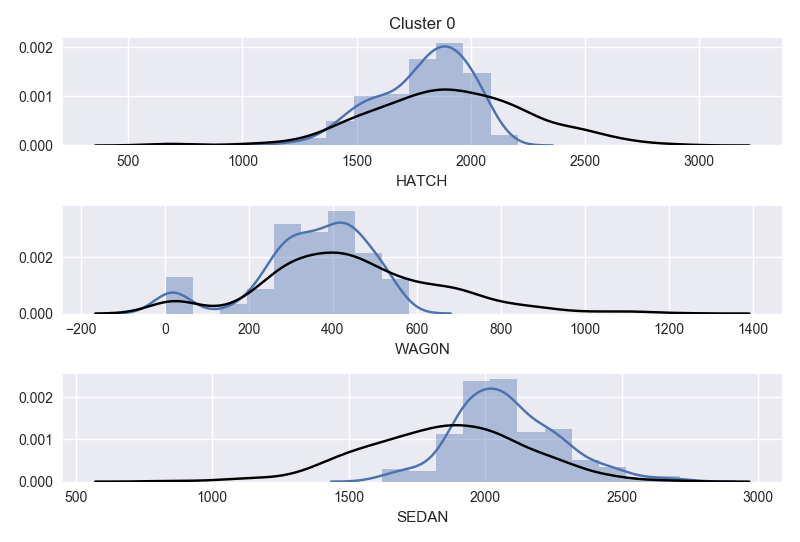
  “Elbow Method” “Silhouette Method”

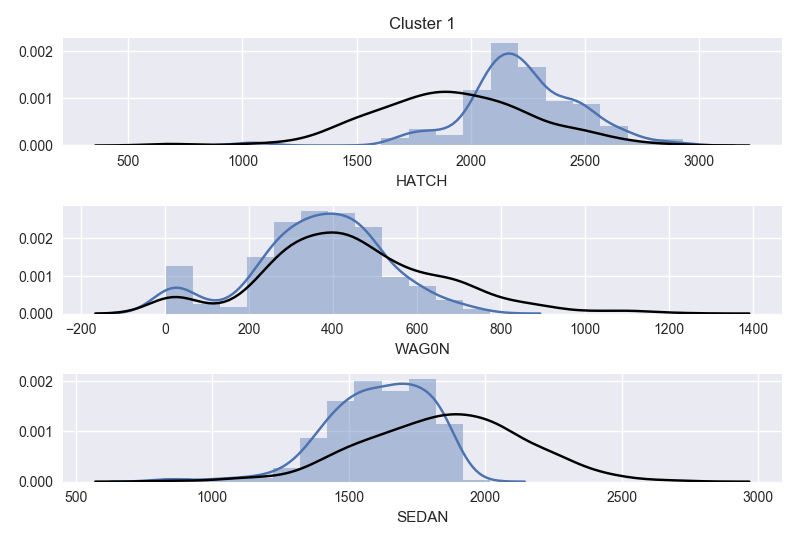
First, we look at the elbow method to determining better K. The graph shows that our best K is between 4 and 6. Though, this method doesn’t always provide with the best K.

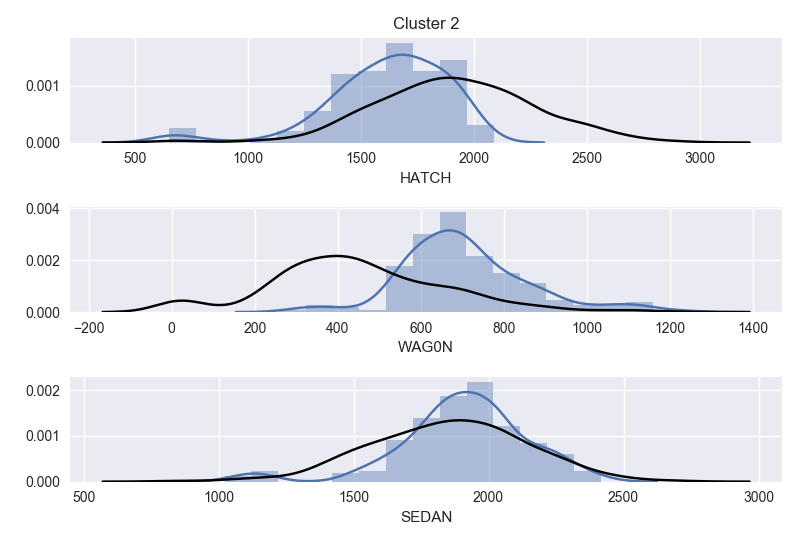
“Silhouette Method” is a more precise technique in identifying best K. It measures how close each point in a cluster to the point in the neighborhoodlike cluster. Silhouette value lies in the range [-1, 1]. One of the main goal of clustering is to minimise inter-cluster similarity. Therefore, higher the value better the cluster configuration. In our case, k=4 has a better inter-cluster similarity (~0.3709) in comparison with k=6 which has (~0.3577). Based on this method we can state that k=4 is the best model.

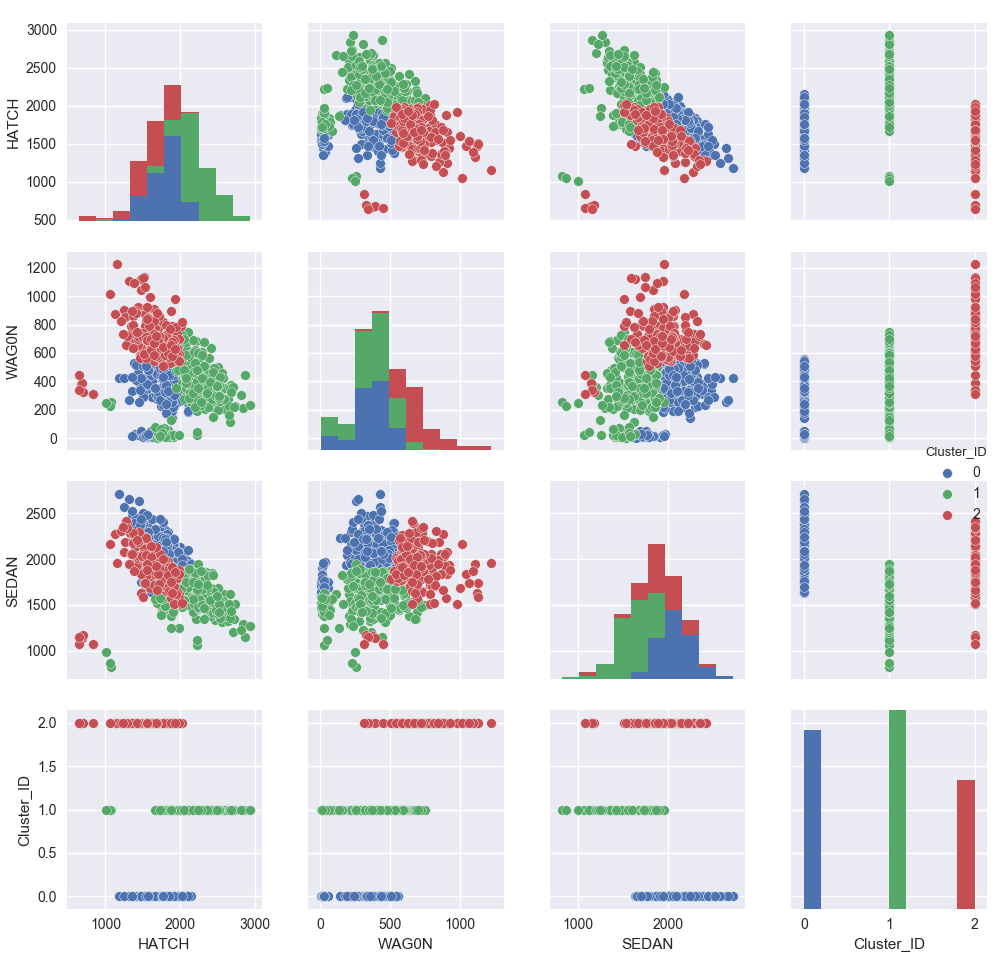
However, even though K=4 is the best model for this task we will use K=3 in our research as it’s easier to interpret and provide more distribution understandably.











After careful examination of all distplots and pairplot with K=3 we can sum up our findings.

**Cluster 0**: has mostly average sales for HATCH. At the same time WAGON has low to medium sales. On the other hand, SEDAN has average to high sales.

**Cluster 1**: showing that sales for HATCH is medium to high, WAGON has low to medium sales and SEDAN is sold on an average level.

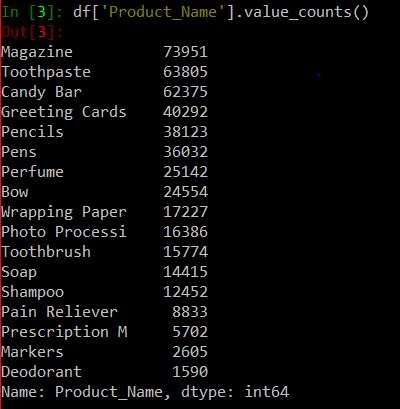
**Cluster 2**: sales on HATCH are from low to medium, WAGON has medium to high sales and SEDAN has medium to high sales.

# Task 2: Descriptive Mining – Association

|  |  |
| --- | --- |
| *Variable* | *Importance* |
| *Location* | Irrelevant. The variable has no use to our research |
| *Transactin\_Date* | Irrelevant. The transaction date’s values are similar and of no use. |
| *Quantity* | Irrelevant. Each row of Quantity variable has a value of “1”. |

1. “pos\_transaction.csv” doesn’t have obvious erroneous or missing values. Some of the variables were dropped due to irrelevancy in our research.

1. “Transaction\_Id” and “Product\_Name” were 2 most valuable variable which we were using to perform association mining. “Transaction\_Id” variable include the transaction ids which group the products under the particular transaction. At the same time “Product\_Name” attribute has the following values:

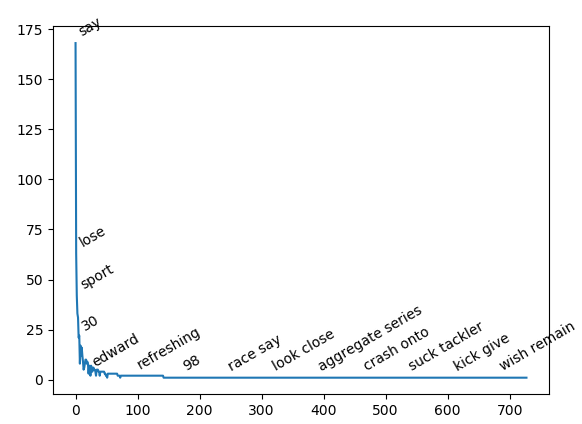


**3**.



* 1. The table is illustrating the association rules between different items. It’s noticed that 2 resulting rules have the highest lift of 3.601370. They are “**Perfume => Toothbrush**” and “**Toothbrush =>Bow**” with lift level of 3.081236.
  2. Furthermore, it is identified that the highest confidence value has “**Toothbrush=> Perfume**” rule with the value 0.323979.
  3. If we plot together with the confidence, lift and support we get the following scenario. The customers who buy **Perfume** are likely to purchase **Toothbrush** with it. The lift of this rule is quite high which denotes strong association rules between these two products. On the other hand, the rule “**Toothbrush => Perfume**” has the same lift level as “**Toothbrush => Perfume**” but higher confidence level implying that if there is **Toothbrush** in the transaction it’s likely that **Perfume** will be in the transaction as well. Further, we can observe interesting rules is **“Toothbrush => Bow”** with a quite a high lift level of 3.081236 but if we take a closer look at the confidence level it appears to be significantly lower than confidence level of other association rules with the same lift level. However, from the business prospective this rule is “interesting” as it has a positive correlation (positive lift value). Furthermore, there are a couple of other association rules which can attract business’s attention such as **“Candy Bar, Magazine => Greeting Cards”**, **“Pencils, Toothpaste => Candy Bar”**, **“Magazine, Toothpaste => Greeting Cards”**. Lift level is approximately the same around 2.7 for these rules. Also, it can’t go without saying that confidence level for these rules is almost twice higher than for the rules with higher lift.
  4. From the table above, we can notice 4 association rules for items which are bought together with “Candy Bar”. They are **“Pencils, Toothpaste => Candy Bar”**, **“Magazine**, **Greeting Cards => Candy Bar”**, **“Magazine, Toothpaste => Candy Bar”**, **“Toothpaste, Greeting Cards => Candy Bar”**.
  5. If we look closer to the above rule, we can state that the association rule **“Pencils, Toothpaste => Candy Bar”** has the highest Lift and confidence level. It means that the customers who bought pencils and toothpaste will likely to buy candy bar as well. It would be a wise business decision for the store to place these items close to each other to increase their sales. From the question 4.a. we notice that **Magazine** and **Greeting Cards** are also frequently bought with Candy Bars.

# Task 3: Text Mining

1. The **TEXT** variable was the only variable used for the text mining analysis. This variable was chosen as it was the only one relevant to the information trying to be gathered. The goal is to determine patterns in text to allow for groupings of similar articles. With that in mind the **TEXT** variable has all the required information, and thus it was chosen for the analysis
2. From the ZIPF plot, which can be seen below, ten terms can be seen which would not be useful for clustering  
     
   These ten terms are:  
   - wish remain  
   - kick give  
   - suck tackler  
   - crash onto  
   - aggregate series  
   - look close  
   - race say  
   - 98  
   - refreshing  
   - say  
   The first 9 terms listed were chosen due to low frequency occurrences (less than 2), making these not useful to the analysis. “say” was chosen due to its high occurrence. The phrase appears too many times to be useful in clustering
3. “Say” was disregarded as it occurred in nearly 90% of documents. This high frequency occurrence makes it non-beneficial to clustering as it could occur in all clusters. The next highest frequency occurred in less than half the documents, therefore it was decided to keep this term.
4. The feature weightings where as follows:  
   **min\_df** = 2,  
   **max\_df** = 0.8  
   **min\_df** was set as 2 as the lowest occurring terms appeared in less than 2 documents. This eliminates any phrases occurring once. While **max\_df** was set as 0.8 to ensure the highest occurring phrase was eliminated. This means that any occurrence of over 80% will be unused, as “say” was nearly 90% it will be eliminated.
5. After feature selection was performed based on the above information, the number of unique tokens available for clustering is 6917.
6. Based on elbow and silhouette methods, the optimum k value is 6-8. The clusters (9) generated are shown below with their top terms and named accordingly:

|  |  |
| --- | --- |
| Cluster 0 (Countries) | england, wale, ireland, win, nation, |
| Cluster 1 (Olympics) | world, race, year, olympic, indoor, |
| Cluster 2 (Football) | club, player, game, cup, season, |
| Cluster 3 (Tennis) | roddick, nadal, moya, hewitt, 6, |
| Cluster 4 (Cricket) | pakistan, ball, day, wicket, test, |
| Cluster 5 (Premier League) | chelsea, liverpool, club, want, arsenal, |
| Cluster 6 (Drugs in Sport) | drug, ban, greek, iaaf, test, |
| Cluster 7 (Tennis) | 6, open, 7, federer, henman, |
| Cluster 8 (Australian Open) | capriati, open, australian, australian open, clijsters, |

The name football was chosen for cluster 2 based on the phrases but also on the basis that the BBC are a British company. This makes the most likely sport to be football. What is noticed is that two clusters can be called tennis based on the player names within them, however a look at more terms could determine a point of difference.

1. By applying LSA/SVD the clustering is based on document concept similarity, rather than just word similarity. This can allow for a more accurate clustering, as the documents are similar. This is done by reducing the components significantly, in this case to 100. This helps to significantly reduce the time taken for clustering when compared to the previous method. Below is a table showing the new results of clusters and there possible names:

|  |  |
| --- | --- |
| Cluster 0 (Liverpool News) | club, liverpool, want, player, play, |
| Cluster 1 (Australian Sport) | open, year, australian, athletics, australian open, |
| Cluster 2 (Cricket) | cricket, pakistan, ball, test, england, |
| Cluster 3 (Rugby) | bath, tindall, sri, sri lanka, lanka, |
| Cluster 4 (Six Nation Cup) | england, ireland, wale, nation, six nation, |
| Cluster 5 (Olympics) | indoor, world, olympic, record, birmingham, |
| Cluster 6 (Rugby) | kafer, saracen, club, head coach, sinderberry, |
| Cluster 7 (EPL) | arsenal, chelsea, henry, wenger, mourinho**,** |
| Cluster 8 (Tennis) | 6, roddick, 7, moya, nadal, |

Cluster 3 was chosen as rugby as Mike Tindall is a rugby player who plays for Bath. However cluster 6 is a better representation of rugby as all terms directly relate to the topic. The clusters from SVD appear to give a better grouping. The terms within the clusters appear to have more relevance to each other, with the exception of cluster three. It can be seen that the clusters from this method do share some similarities, such as tennis and cricket remaining essentially the same. Other clusters such as the Six Nation Cup cluster have been given more relevant terms, making them more recognisable. Six Nation Cup was previously countries.

1. These clusters can be useful to the planned goal of an online personalised news service. With this information articles of a similar nature can be grouped together. For example, if a user was to read an article on a recent Premier League game, the service could then offer up other articles that have been classified in the same cluster. This would offer the user articles targeted at their particular interest.  
   The information could be used in a more direct way by asking the reader to select what types of news they would like to see. Once selected news from clusters relating to their choice can be delivered to them, meaning they obtain only the news they want.

# Task 4: Web Mining

1. Association Mining was chosen as the web mining technique to use for this task. Association mining is used for establishing links between variables, based on commonality, from a set of records within a data set. The goal of this task is to determine user browsing patterns to ensure the most efficient website design. With this in mind, establishing links between pages of the site visited based on sessions would be effective in determining browsing patterns.   
     
   We want to discover what pages are likely to be visited given another page is visited first. This type of analysis can be seen in the association rule (item set A) => (item set B). In this case, item set is web page. With this information the association technique was chosen.
2. For the task, the variables chosen were **user\_id** and **request**.   
   **user\_id:-** provides the information for the session. This variable is the same as the session variable, thus either could have been used. The reason this variable was chosen of IP is due to the fact an IP could be obtain via a proxy, making this information unreliable. As well as this, the user\_id only provides information on a session, where as the same IP could be used for multiple sessions. This could skew the data as a session on a different day could have no relevance to a previous session. Therefore it was decided that individual session data should be used to ensure browsing habits are determined more effectively.  
   **request:-** provides information on links clicked within the webpage. This variable was chosen as requests can be catalogued per session. This information can be used to determine the likelihood of certain links being requested if another link has been requested in a given session. This is the only variable that provides information on the pages visited on the website, making it important to determining browsing behaviour.  
   For the apriori function, the arguments ***transaction***and ***min\_support***were used.   
   ***transaction:-***this is a list of the transactions (in this case, requestions made). It is a list of lists of requests. The data used here was the request lists  
   ***min\_support:-*** This argument is to determine the minimum support of the relationship. This was set as 0.04, as it was found to give the best results.
3. PLACE HOLDER