Artificial Intelligence DT2016

Task 3 - K-means clustering

By *Özgun Mirtchev*

Professor *Franziska Klügl*

Örebro University

April 17, 2017

1 Introduction

Machine learning is a topic that has been mentioned a lot more in recent times. Almost all of the technical devices that are used today have some sort of machine learning built-in. In this lab K-means clustering is used to group up similar objects from a base of available data. While clustering is not exactly machine learning, since it gives a result based on data, it still counts as some sort of learning.

1.1 Objective

This lab requires a working implementation of k-means clustering algorithm. The provided unclustered data are values of customers spending in different attributes: Fresh, Milk, Grocery, Frozen, Detergents Paper and Delicatessen. The customers has their own spending channel as well which includes hotels, restaurants, cafes and retail. Besides channels the customers have a region attribute as well, which speaks about where the customers lives. The file data has the data separated by ";" needs to be read into the program by an appropriate procedure into appropriate data structures. From that k-means clustering will be done and an analysis of the result will be done. A method for deciding the optimal k-value for a cluster will also be implemented.

1.2 K-means clustering

K-means clustering is a very commonly used clustering algorithm. It uses a distance measure between unclustered data objects to cluster them up together depending on randomly picked centroids. The centroids are recalculated depending on all the objects from the database, and the way the objects are placed in the cluster is dependent on the distance between the object and the nearest centroid. The process is repeated until the centroid reaches a state where its position stops changing. Depending on which k value was given, the produced result will be k amount of clusters. An example may be seen in figure 1where the k value is 3, clustering the data into three groups.

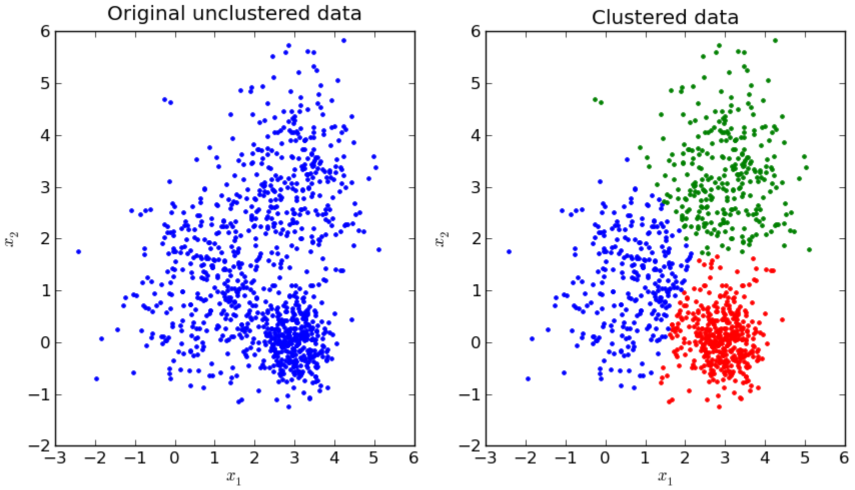


Figure 1: K-means clustering example

1.3 Elbow-curve

Selecting a good k value is also important. There are a few alternatives to choose from when wanting to know the optimal k value. One approach is to do multiple scenario runs with different k values and choose the best one out of these according to a criterion. In this lab the Elbow-method was applied to figure the best k value of an unclustered data.

This makes it easier to see which k value would be the most optimal for the current cluster group, instead of doing it manually for each attribute. It's important to not increase the k value too much to avoid overfitting, eventhough it may give a more 'precise' result.

2 Results

2.1 Scenario Preparation

First task was to implement a way to store all of the data that was going to be read into an apporpriate data structure, to be able to handle the data easily.

Since the data is about the annual spending of costumers, a class named Client was created to store the information for each customer. In this class, the channel and the region of the client are stored as member variables, while the spending of the client was stored into an array, to make it easier to calculate the new position of the centroid, depending on which attribute was selected to do clustering on.

A cluster class was also created to store the randomly picked centroid, together with a list of the clients that are going to be assigned to the cluster. The initial cluster has only one client in it, which is the randomly picked centroid.

```
class Client:
  init(costumer_data):
    client_data = costumer_data
    channel = 0
    region = 0
```

```
class Cluster:
  init(clients):
    clients = clients
    centroid = clients[0]
```

Figure 2: Class pseudocode

2.2 Program procedure

Upon starting the program in the console, a menu is displayed listing all the available attributes to choose:

```
Attributes:
 1. Fresh
 2. Milk
 3. Grocery
 4. Frozen
 5. Detergents paper
 6. Delicatessen
```

The user enters the number of the chosen attribute, which in turn loads the corresponding data into the class instances. The data is stored in a file named “costumer.data”, which is filled with 440 values for each attribute. If the loading of data is successful, a message will be displayed, notifying the user with how many values were loaded.

```
Loading client data... (../files/costumer.data)
Data loading completed: 440 values
```

Immediately after the loading has finished the user is prompted with a choice of which method of K-means should be run. The available methods are manual and elbow-method. With manual the user is able to manually choose a desired k value.

Here the user needs to choose which method they would like to run.

```
K-means
 1. Manual K-value
 2. Elbow-method
```

By choosing manual-method, the program prompts the user again by asking which value the K-means clustering should be run with.

By choosing elbow-method, k-means clustering is currently run on the dataset with the k value in a range of 1-9 . The elbow-method calculates the variance between each cluster and detects the elbow-curve by the biggest change in the value of sum of squared errors (SSE) from the specific cluster. Below one may see an example of the elbow-calculations from 1-9 clusters. Disregarding the variance of the first cluster, the biggest difference of variance is between 2 and 3. Due to how the elbow-method was implemented the optimal value of k is chosen as 3.

```
          Cluster           Percentage of variance
          1                 0.0
          2                 60.42500844611611
          3                 79.3975499703917
          4                 85.91375760172765
          5                 90.98892241443801
          6                 92.41320818922254
          7                 95.38843859520377
          8                 95.78655678993113
          9                 95.98528234810792
Elbow curve detected. Optimal K-value: 3
```

After the clustering process, an analysis report is printed to give information about what is inside each cluster. Most important values of a cluster is to see what the size of it is and how the different data is distributed. For this reason, there are four versions of the analysis print-out:

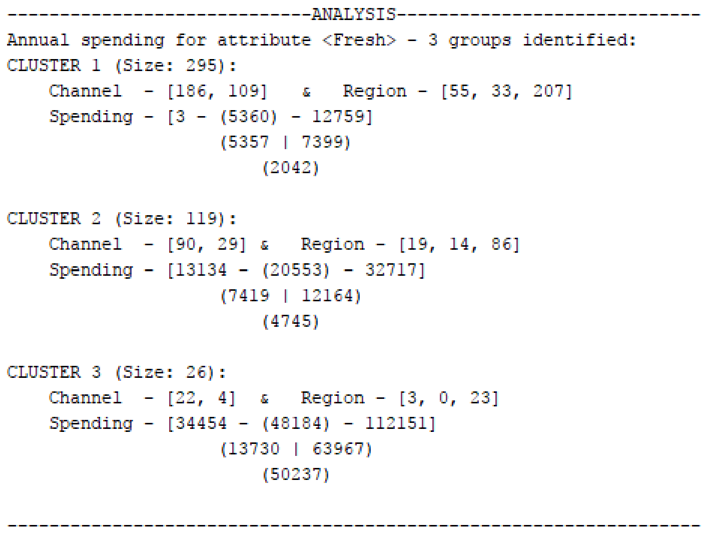
1. A **compact** version (see Figure 3) which gives a more detailed output but easy to glance over, with differences between the highest and lowest value in the cluster.

2. A **compacttext** version (see Figure 4) which prints out the values in a compact text form for better overview of the shopping and habitat of the group.

3. A **full** version (see Figure 5) which prints out relevant information for a user not interested in differences between clusters.

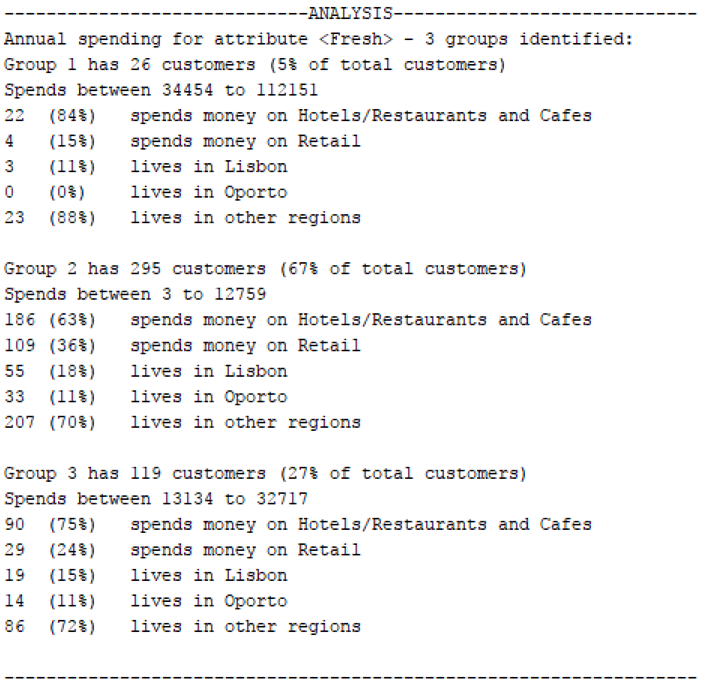
4. A **fulltext** version (see Figure 6) prints out all of the values and tells it in a natural spoken language. More useful if one wants to present it in an

article or some kind.



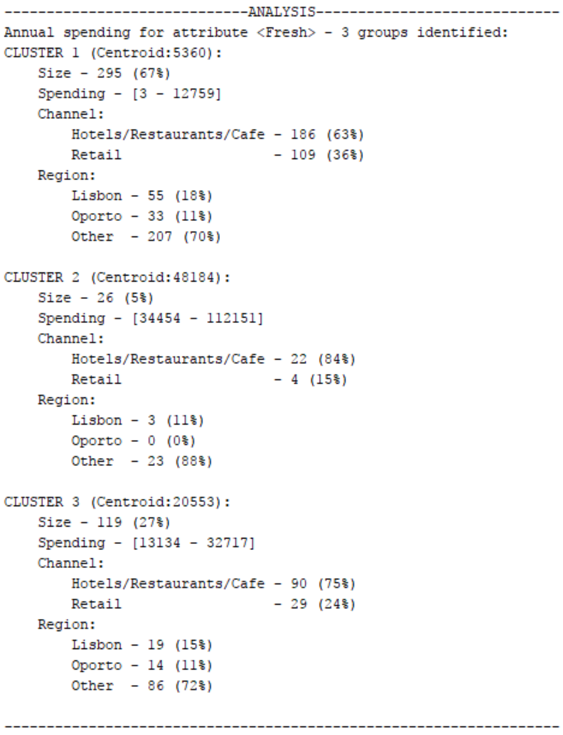
```
-----------------------------ANALYSIS-----------------------------
Annual spending for attribute <Fresh> - 3 groups identified:
CLUSTER 1 (Size: 295):
    Channel  - [186, 109]   &   Region - [55, 33, 207]
    Spending - [3 - (5360) - 12759]
                    (5357 | 7399)
                        (2042)

CLUSTER 2 (Size: 119):
    Channel  - [90, 29] &   Region - [19, 14, 86]
    Spending - [13134 - (20553) - 32717]
                    (7419 | 12164)
                        (4745)

CLUSTER 3 (Size: 26):
    Channel  - [22, 4]  &   Region - [3, 0, 23]
    Spending - [34454 - (48184) - 112151]
                    (13730 | 63967)
                        (50237)

------------------------------------------------------------------
```

Figure 3: Compact analysis output



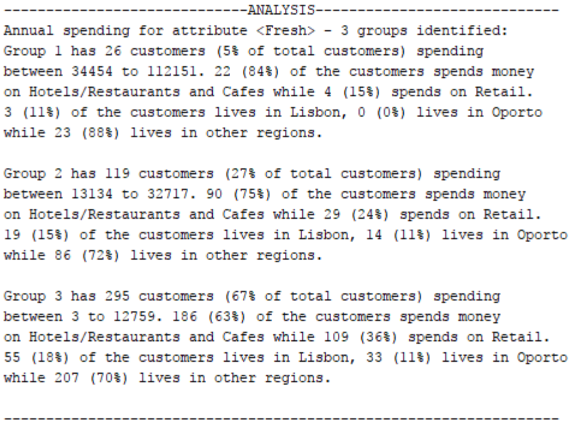
```
-----------------------------ANALYSIS-----------------------------
Annual spending for attribute <Fresh> - 3 groups identified:
Group 1 has 26 customers (5% of total customers)
Spends between 34454 to 112151
22  (84%)   spends money on Hotels/Restaurants and Cafes
4   (15%)   spends money on Retail
3   (11%)   lives in Lisbon
0   (0%)    lives in Oporto
23  (88%)   lives in other regions

Group 2 has 295 customers (67% of total customers)
Spends between 3 to 12759
186 (63%)   spends money on Hotels/Restaurants and Cafes
109 (36%)   spends money on Retail
55  (18%)   lives in Lisbon
33  (11%)   lives in Oporto
207 (70%)   lives in other regions

Group 3 has 119 customers (27% of total customers)
Spends between 13134 to 32717
90  (75%)   spends money on Hotels/Restaurants and Cafes
29  (24%)   spends money on Retail
19  (15%)   lives in Lisbon
14  (11%)   lives in Oporto
86  (72%)   lives in other regions

------------------------------------------------------------------
```

Figure 4: Compact text analysis output



```
-----------------------------ANALYSIS-----------------------------
Annual spending for attribute <Fresh> - 3 groups identified:
CLUSTER 1 (Centroid:5360):
    Size - 295 (67%)
    Spending - [3 - 12759]
    Channel:
        Hotels/Restaurants/Cafe - 186 (63%)
        Retail                  - 109 (36%)
    Region:
        Lisbon - 55 (18%)
        Oporto - 33 (11%)
        Other  - 207 (70%)

CLUSTER 2 (Centroid:48184):
    Size - 26 (5%)
    Spending - [34454 - 112151]
    Channel:
        Hotels/Restaurants/Cafe - 22 (84%)
        Retail                  - 4 (15%)
    Region:
        Lisbon - 3 (11%)
        Oporto - 0 (0%)
        Other  - 23 (88%)

CLUSTER 3 (Centroid:20553):
    Size - 119 (27%)
    Spending - [13134 - 32717]
    Channel:
        Hotels/Restaurants/Cafe - 90 (75%)
        Retail                  - 29 (24%)
    Region:
        Lisbon - 19 (15%)
        Oporto - 14 (11%)
        Other  - 86 (72%)

------------------------------------------------------------------
```

Figure 5: Full analysis output

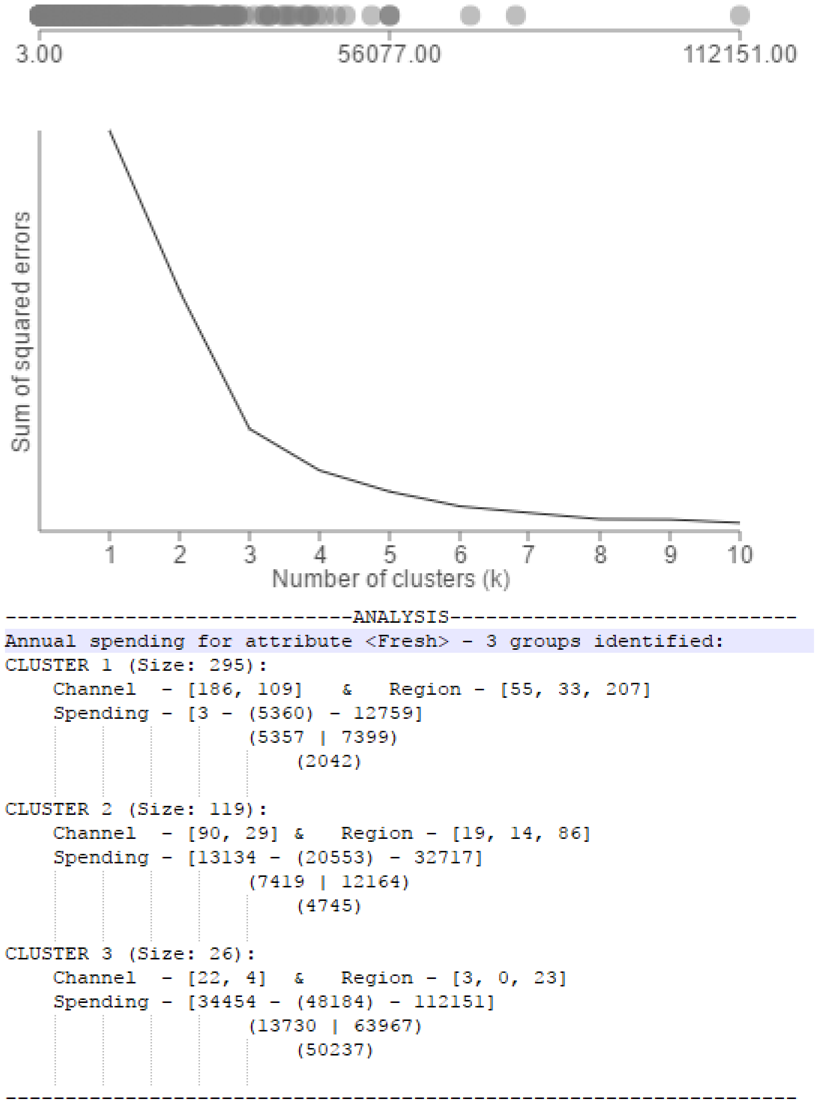


```
-----------------------------ANALYSIS-----------------------------
Annual spending for attribute <Fresh> - 3 groups identified:
Group 1 has 26 customers (5% of total customers) spending
between 34454 to 112151. 22 (84%) of the customers spends money
on Hotels/Restaurants and Cafes while 4 (15%) spends on Retail.
3 (11%) of the customers lives in Lisbon, 0 (0%) lives in Oporto
while 23 (88%) lives in other regions.

Group 2 has 119 customers (27% of total customers) spending
between 13134 to 32717. 90 (75%) of the customers spends money
on Hotels/Restaurants and Cafes while 29 (24%) spends on Retail.
19 (15%) of the customers lives in Lisbon, 14 (11%) lives in Oporto
while 86 (72%) lives in other regions.

Group 3 has 295 customers (67% of total customers) spending
between 3 to 12759. 186 (63%) of the customers spends money
on Hotels/Restaurants and Cafes while 109 (36%) spends on Retail.
55 (18%) of the customers lives in Lisbon, 33 (11%) lives in Oporto
while 207 (70%) lives in other regions.

------------------------------------------------------------------
```

Figure 6: Full text analysis output

2.3 Tests

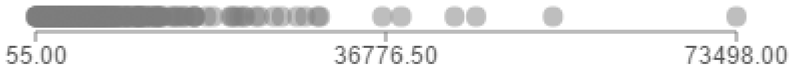
In this section some tests will show how the elbow-method in the program chooses the k-value for each attribute, together with diagram of the elbow-curve and data points complemented with an analysis output.

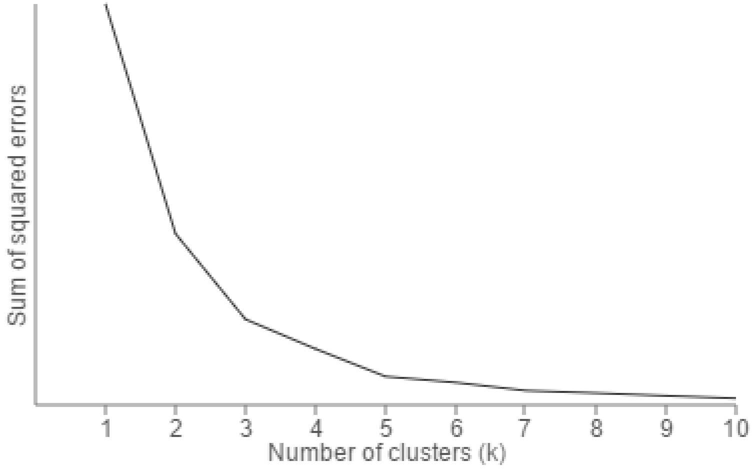


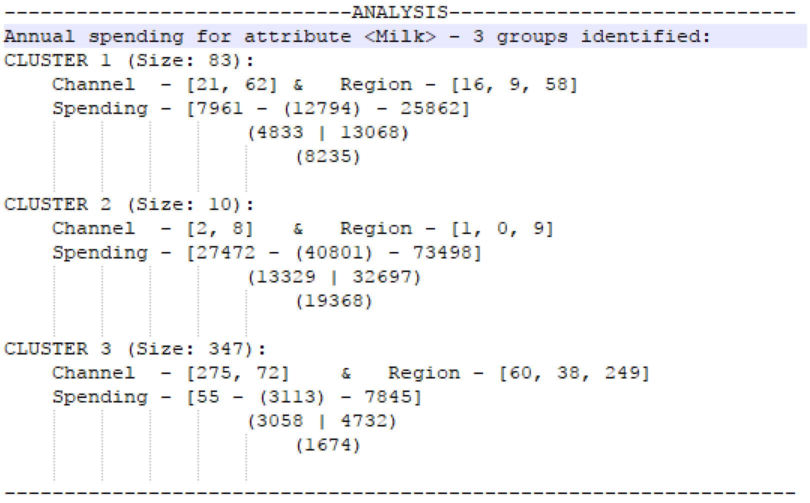
```
-----------------------------ANALYSIS-----------------------------
Annual spending for attribute <Fresh> - 3 groups identified:
CLUSTER 1 (Size: 295):
    Channel  - [186, 109]   &   Region - [55, 33, 207]
    Spending - [3 - (5360) - 12759]
                (5357 | 7399)
                    (2042)

CLUSTER 2 (Size: 119):
    Channel  - [90, 29] &   Region - [19, 14, 86]
    Spending - [13134 - (20553) - 32717]
                (7419 | 12164)
                    (4745)

CLUSTER 3 (Size: 26):
    Channel  - [22, 4]  &   Region - [3, 0, 23]
    Spending - [34454 - (48184) - 112151]
                (13730 | 63967)
                    (50237)

------------------------------------------------------------------
```

Table 1: Attribute: Fresh

The first attribute, Fresh, received the k value of 3 by the elbow-method. Looking at the analysis data it seems like a fair number to put it, but perhaps even making 4 clusters would make it group up the three highest points together, instead of grouping them up with the 23 others, as can be seen in cluster 3. Doing this would perhaps make the clustering a bit fair and not make the higher numbers span over to the lower side too much, however it would most likely be overfitted, so keeping it at 3 is a better number.



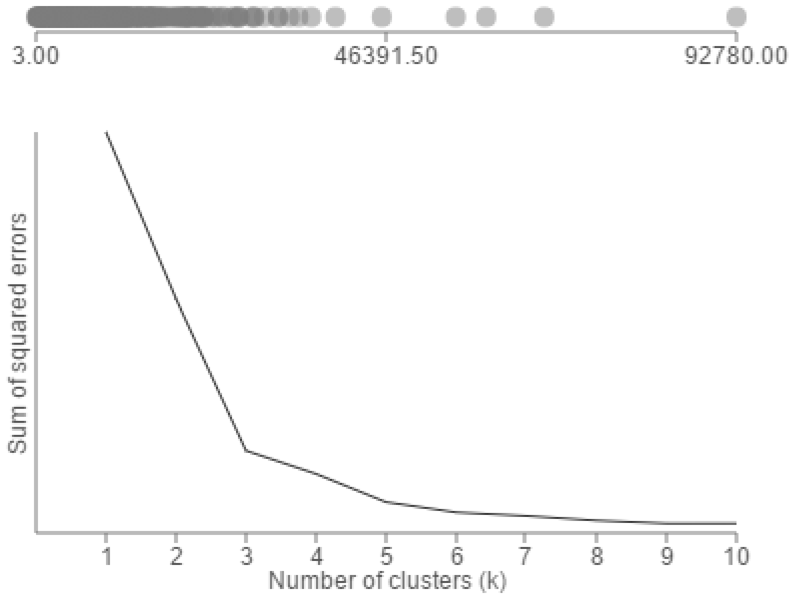


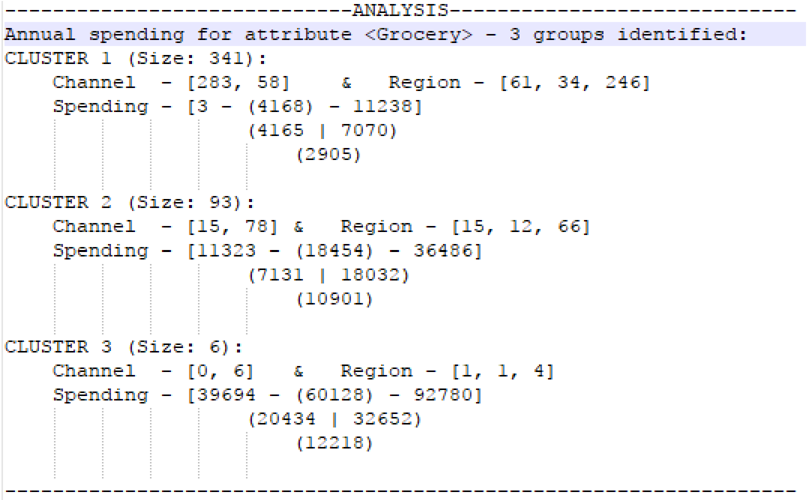


```
-----------------------------ANALYSIS-----------------------------
Annual spending for attribute <Milk> - 3 groups identified:
CLUSTER 1 (Size: 83):
    Channel  - [21, 62] &   Region - [16, 9, 58]
    Spending - [7961 - (12794) - 25862]
                    (4833 | 13068)
                        (8235)

CLUSTER 2 (Size: 10):
    Channel  - [2, 8]   &   Region - [1, 0, 9]
    Spending - [27472 - (40801) - 73498]
                    (13329 | 32697)
                        (19368)

CLUSTER 3 (Size: 347):
    Channel  - [275, 72]    &   Region - [60, 38, 249]
    Spending - [55 - (3113) - 7845]
                    (3058 | 4732)
                        (1674)

------------------------------------------------------------------
```

Table 2: Attribute: Milk

As the previous attribute the elbow-algorithm detects the k value to be at 3, looking at the diagram it makes sense.

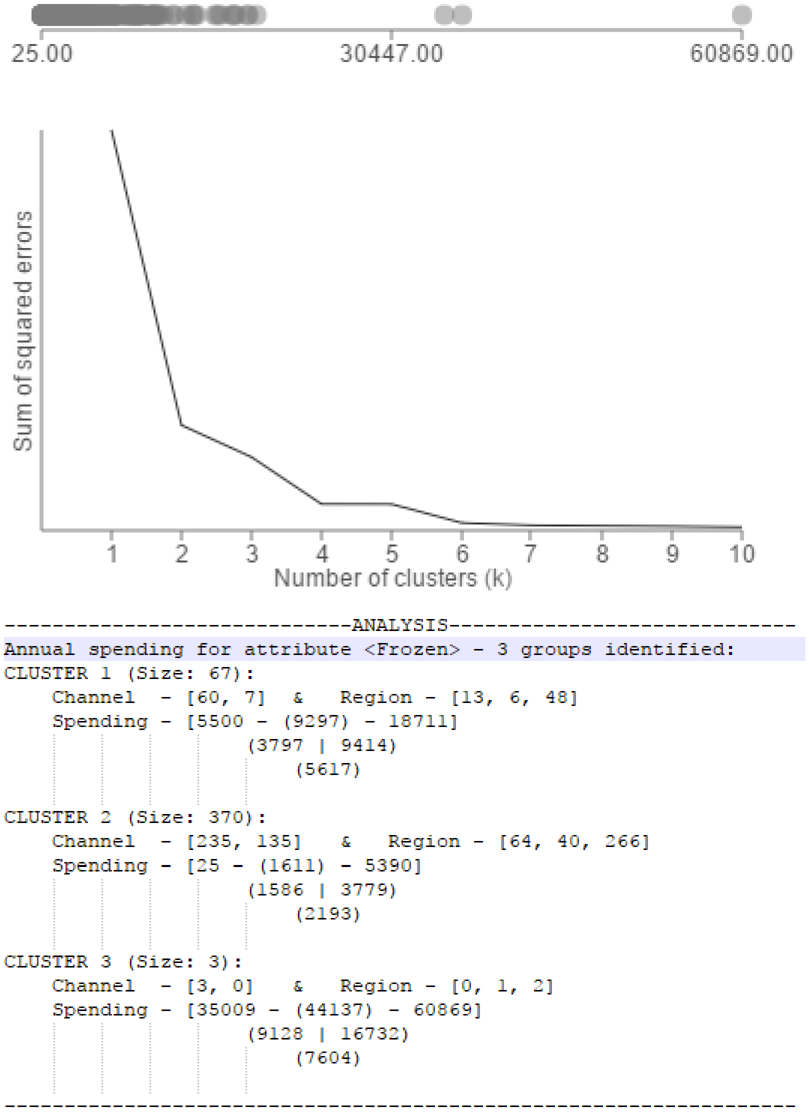




```
-----------------------------ANALYSIS-----------------------------
Annual spending for attribute <Grocery> - 3 groups identified:
CLUSTER 1 (Size: 341):
    Channel  - [283, 58]    &   Region - [61, 34, 246]
    Spending - [3 - (4168) - 11238]
                   (4165 | 7070)
                       (2905)

CLUSTER 2 (Size: 93):
    Channel  - [15, 78] &   Region - [15, 12, 66]
    Spending - [11323 - (18454) - 36486]
                   (7131 | 18032)
                       (10901)

CLUSTER 3 (Size: 6):
    Channel  - [0, 6]   &   Region - [1, 1, 4]
    Spending - [39694 - (60128) - 92780]
                   (20434 | 32652)
                       (12218)

------------------------------------------------------------------
```

Table 3: Attribute: Grocery

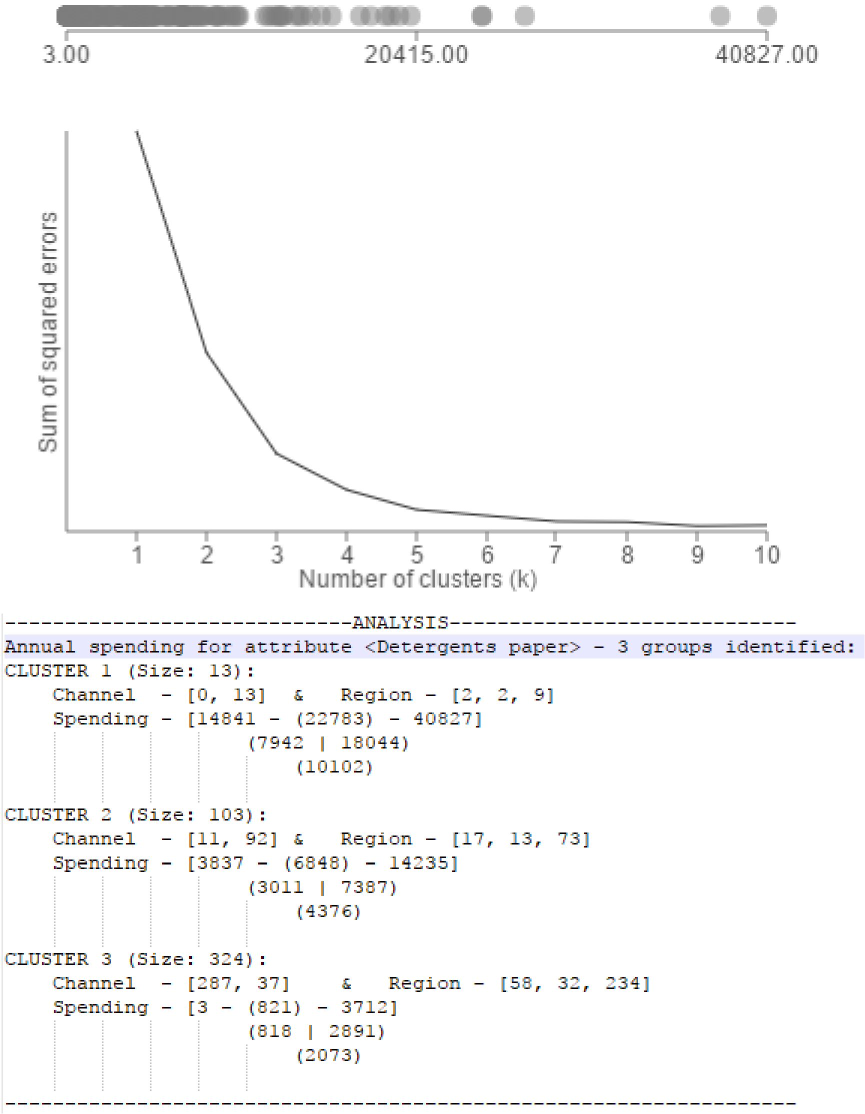
Once again there is clear elbow detected at k = 3. Looking at the 6 data points in the diagram, and the size of cluster 3 in the analysis report, which is also 6, it seems to make sense putting it as k = 3.



```
----------------------------ANALYSIS----------------------------
Annual spending for attribute <Frozen> - 3 groups identified:
CLUSTER 1 (Size: 67):
    Channel  - [60, 7]  &   Region - [13, 6, 48]
    Spending - [5500 - (9297) - 18711]
                    (3797 | 9414)
                        (5617)

CLUSTER 2 (Size: 370):
    Channel  - [235, 135]   &   Region - [64, 40, 266]
    Spending - [25 - (1611) - 5390]
                    (1586 | 3779)
                        (2193)

CLUSTER 3 (Size: 3):
    Channel  - [3, 0]   &   Region - [0, 1, 2]
    Spending - [35009 - (44137) - 60869]
                    (9128 | 16732)
                        (7604)

----------------------------------------------------------------
```

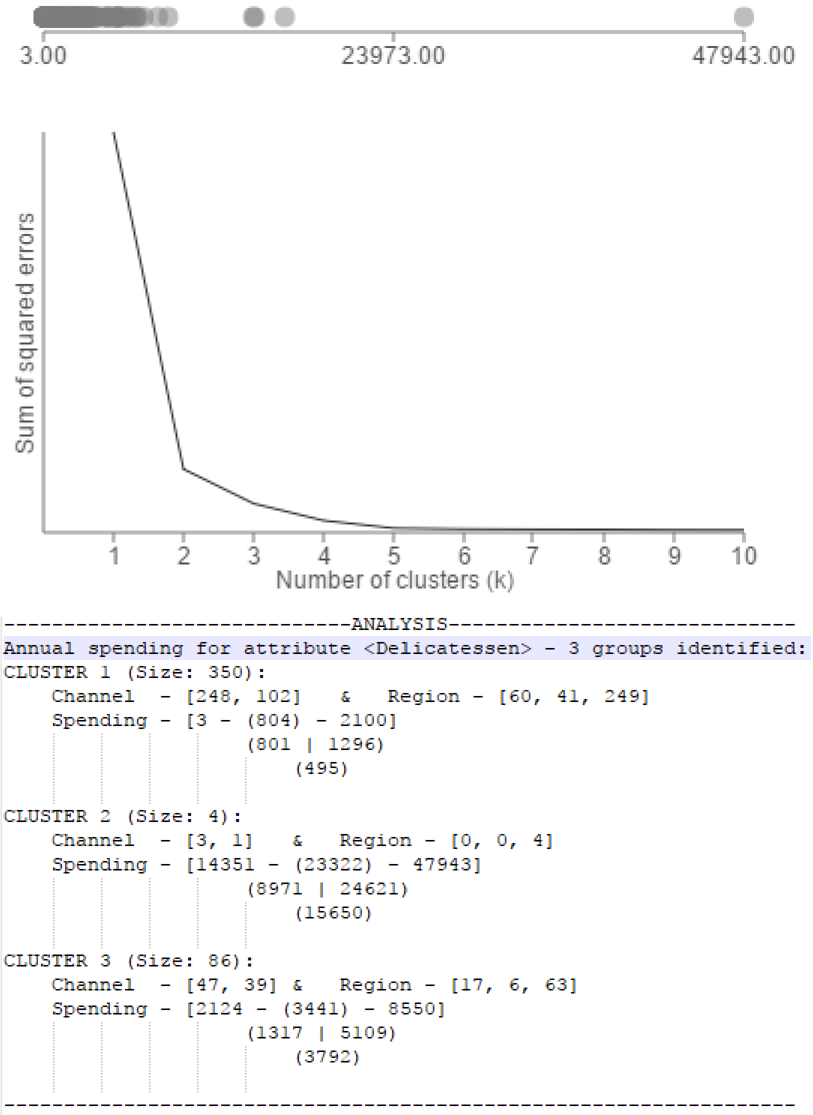
Table 4: Attribute: Frozen

The same situation as the previous one, cluster 3 has the size of 3, corresponding to the amount of points in the diagram. There are a lot of values grouped up at the start making a big cluster at the size 370 and a second cluster a bit smaller at 67 where the points are a bit more spread.



```
-----------------------------ANALYSIS-----------------------------
Annual spending for attribute <Detergents paper> - 3 groups identified:
CLUSTER 1 (Size: 13):
    Channel  - [0, 13]  &   Region - [2, 2, 9]
    Spending - [14841 - (22783) - 40827]
                    (7942 | 18044)
                        (10102)

CLUSTER 2 (Size: 103):
    Channel  - [11, 92] &   Region - [17, 13, 73]
    Spending - [3837 - (6848) - 14235]
                    (3011 | 7387)
                        (4376)

CLUSTER 3 (Size: 324):
    Channel  - [287, 37]    &   Region - [58, 32, 234]
    Spending - [3 - (821) - 3712]
                    (818 | 2891)
                        (2073)

------------------------------------------------------------------
```

Table 5: Attribute: Detergents paper



```
-----------------------------ANALYSIS-----------------------------
Annual spending for attribute <Delicatessen> - 3 groups identified:
CLUSTER 1 (Size: 350):
    Channel  - [248, 102]   &   Region - [60, 41, 249]
    Spending - [3 - (804) - 2100]
                (801 | 1296)
                    (495)

CLUSTER 2 (Size: 4):
    Channel  - [3, 1]   &   Region - [0, 0, 4]
    Spending - [14351 - (23322) - 47943]
                (8971 | 24621)
                    (15650)

CLUSTER 3 (Size: 86):
    Channel  - [47, 39] &   Region - [17, 6, 63]
    Spending - [2124 - (3441) - 8550]
                (1317 | 5109)
                    (3792)

------------------------------------------------------------------
```

Table 6: Attribute: Delicatessen

3 Conclusion

The reason why all the attributes received k-value 3 might be because of the small amount of but very high values in the cluster data. This was countered by using the Minkowski distance during the k-means clustering, which gives a middle value of two inputs, no matter how small or big they are. However the elbow-method doesn't consider the unbalanced data and just picks out the value corresponding to the curve. Perhaps balancing the data before doing the clustering might've produced fluctuating values for k, however this is not an easy thing to accomplish without changing the real values and affecting the end result. Had there been more time to spend, a good alternative would've been implemented to alleviate this problem.

Improvements that could've been made to this program, are better code structure. I noticed that I added more things as I was doing the programming, which I didn't consider beforehand. Coding everything in classes probably would've helped a bit with the readability of the code as well, since now it's mostly a big mess of big functions everywhere.

Overall I've learned a lot about how unsupervised machine learning and clustering works. It was a bit difficult to understand how I would setup everything at the start and a lot of information finding was required for me to get going. Either way, I feel like I managed to complete it in a good way and it didn't require an immense amount of time when compared to the previous task.