**CAB330 Data & Web – Analytics**

**Case Study 2**

Donghyeon Kim n9967273

Jack Teys n9996532

Vladislav Kireyev n9897810

**CAB330: Case Study 2**

**Mining from Clothing, Supermarket,**

**News stories and Web Log Data**

**Due date: 21st Oct 2018**

**Weighting: 25%**

**Introduction**

The purpose of this assignment is to give you an understanding that data mining methods can be applied to various types of data sets such as record data, transactional data, text data and web logs, and show you the benefits of applying mining techniques to data domains of any kind. This assignment is divided into four parts: Clustering, Association mining, Text Mining and Web Mining. You will use Python with all of the libraries you have learned to use so far

**Distribution of Marks (Total: 25 marks)**

We would mark your data mining projects in the Week 13 practical class. Every member in the team should complete each of the tasks and should be prepared to show the final models and results panels to your marker. The marker will ask each individual student questions and will assign individual mark (~15%).

In data mining, there is hardly ever a single solution. Also, many times, there is no correct or wrong solution. You may find that your project partner may have different solution as yours. Your group should decide on a single project that you would like to be marked. Submit the report discussing the final project components.

The marks are distributed as follows:

**Part 1: Clustering pre-processing and K-means analysis (7 marks)**

**Part 2: Association Mining and its pre-processing (5 marks)**

**Part 3: Text Mining (Clustering) the News Stories (7 marks)**

**Part 4: Web Mining the Log Data for a Website (6 marks)**

**Part 1: Clustering pre-processing and K-means analysis**

|  |  |
| --- | --- |
| **Name** | **Description** |
|  |  |
| StoreCode | Numerical code for the store (occurs only once in the table) |
|  |  |
| AnnualSales | Annual sales in Dutch guilders |
|  |  |
| Sales | Sales per square meter |
|  |  |
| SFloorSize | Sales floor space of the store (in m2). |
|  |  |
| TotalInvestment | Investment in shop-premises and automation. |
|  |  |

MENS\_CLOTHING\_SALES\_2018.csv is a data set contains the StoreCode, annual sales, Sales per square meter and other characteristics of 400 Dutch men’s fashion stores. Each row represents an individual store. There are five columns in the data set.

The purpose of this task is to identify different clusters of men's fashion stores based on the investment and sales. This analysis helps for any new investors to decide investment and location that can return good sales. The analysis can also identify the fashion interests of people from various locations.

Your task is to conduct k-means clustering on this dataset, and find and describe the **minimum number** **of effective clusters**. Answer the followings in relation to this data and analysis.

**Task 1. Data Preparation for Clustering.**

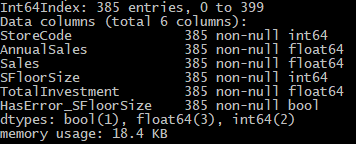
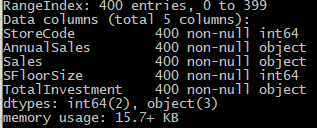
1. Can you identify data quality issues in this dataset such as unusual data types, missing values and others?

There were several data quality issues in this dataset. Firstly, unusual data-types. When the dataset was imported into a data frame using the pandas library, using the function ‘dataset.info()’ has displayed ‘Annual Sales’, ‘Sales’, ‘TotalInvestment’ columns as object data-type.

This was because row values in AnnualSales, Sales and TotalInvestment contains empty input (Second data quality issue) when SFloorSize is equal to 0m^2, hence why SFloorSize is set as ‘int64’ as every row has a numeric value but Rows in AnnualSales, Sales and TotalInvestment has no value when SFloorSize is 0 and the python has interpreted the column as object.

Therefore, empty values has been replaced as NaN (not a number) and the column has been converted as ‘float’ data-type using “.replace(‘’, np.nan)” and “.astype(float)”

After converting each column to the correct data-type, rows which had an inputs as 0 (erroneous inputs, also NaN values) had been dropped using ‘df[df[‘SFloorSize’] >1]



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

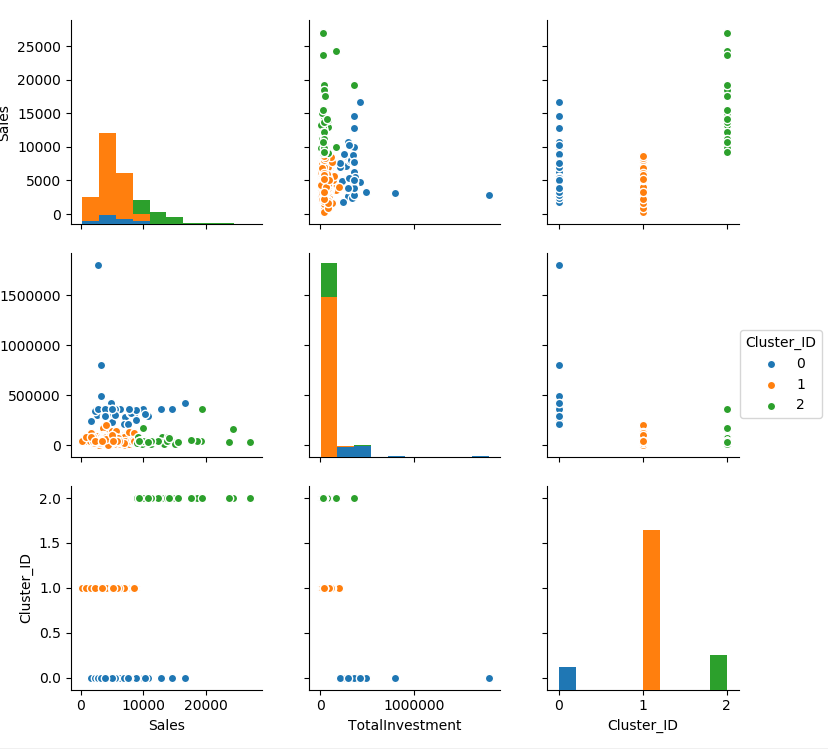
**Included variables**

|  |  |  |
| --- | --- | --- |
| **Variables Included** | **Roles and Measurement level** | **Reason for choosing** |
| Annual Sales | **Role:** Independent variable  **Measurement level:** Quantitative – Interval Continous | This variable is important, as it is the main scope of this case study; which is to define, which characteristics that stores have (such as investment values and store sizes), which outcomes to a **High Sales Profit.** |
| Store Floor Size | **Role:** Independent variable  **Measurement level:** Quantitative – Interval Continous | This variable is used to determine how **store floor sizes** affect on **annual sales** outcome. |
| Total Investment | **Role:** Independent variable  **Measurement level:** Quantitative – Interval Continous | This variable is used to determine the relationship between **investment amount** and **sales**. |

**Removed variables**

|  |  |  |
| --- | --- | --- |
| **Variables Not Included** | **Roles and Measurement level** | **Reason for choosing** |
| Sales | **Role:** Independent variable  **Measurement level:** Quantitative – Interval Continous | We believed Sales was a redundant variable and only decreased the processing speed of our data mining project. Sales can be obtained from dividing **Annual Sales value** by **Store Floor Size.**  Also, more depth analysis is shown below with each picture and its brief explanation |

**Dropping Annual Sales and Investment**

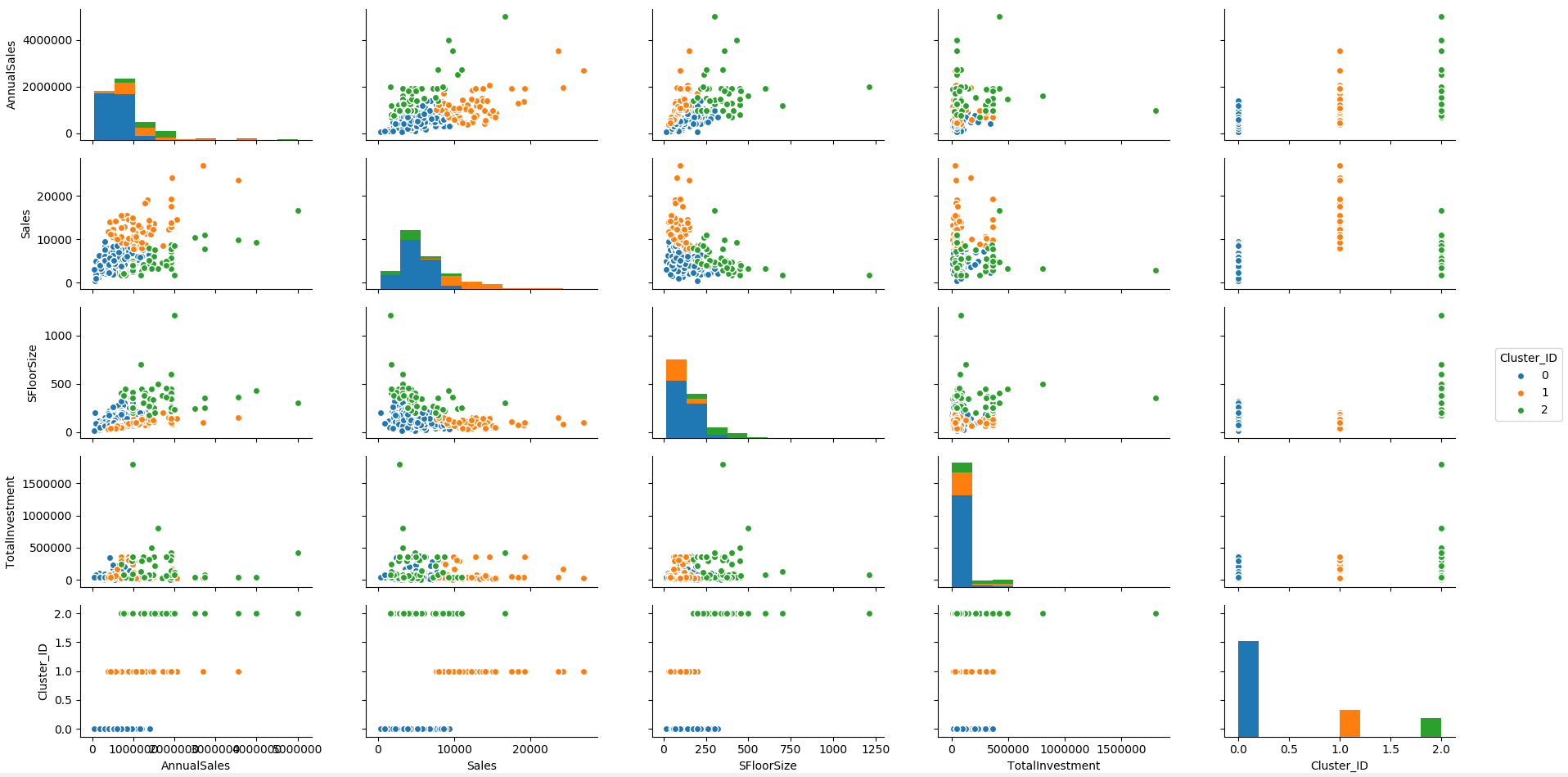


**Explanation why this wasn’t chosen:**

In this model with 2 columns (Sales and Total Investment), we can clearly visualize each cluster. They are well distinguishable, however, there are not many characteristics we can analyse for our clients.

Only visible characteristics I can find is, lower investments tend to do better than higher investments in sale. Where the below 2 cluster models could explain the relationship between the sales and its floor size .

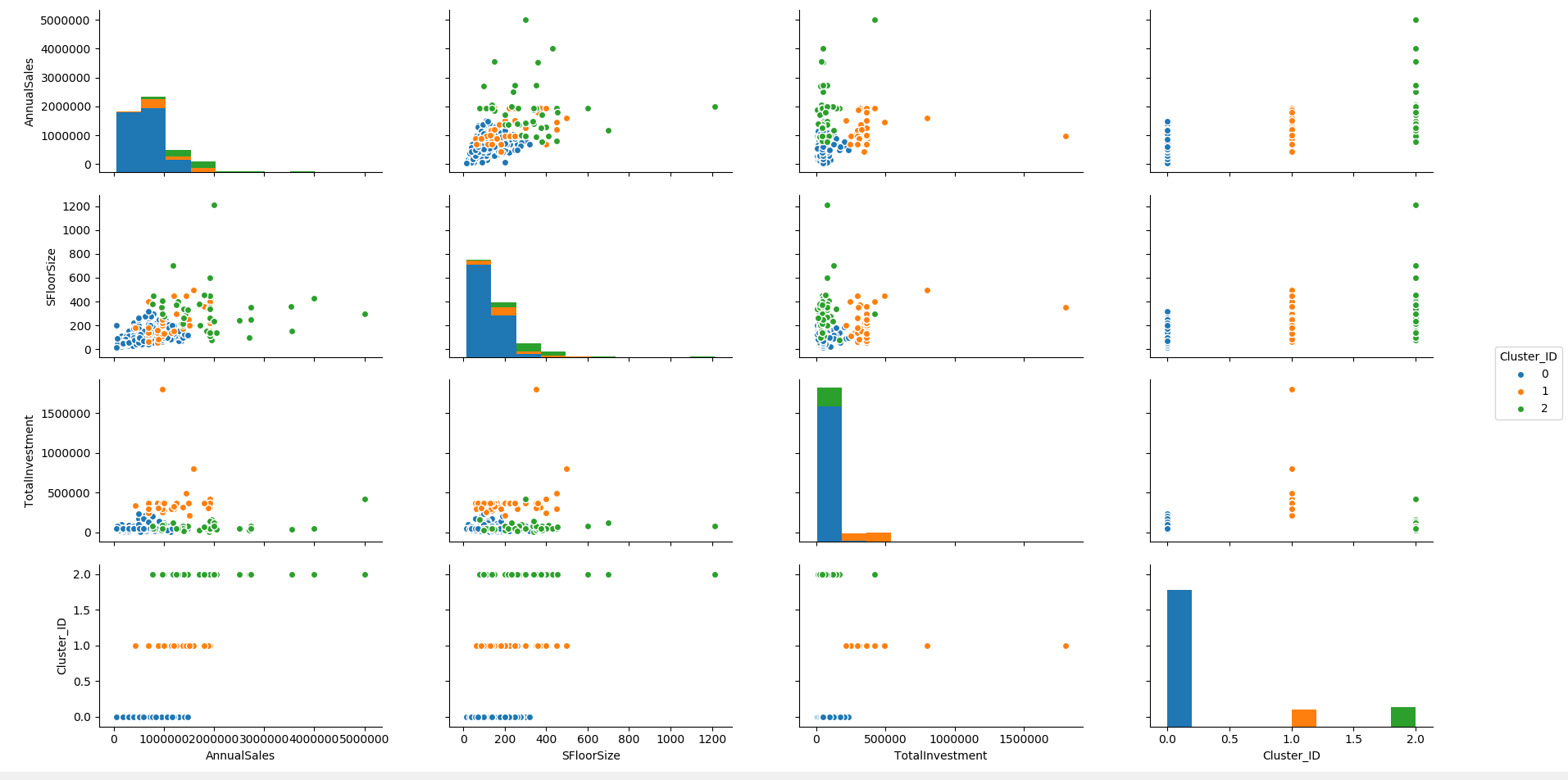
**Dropping no variables**



**Explanation why this wasn’t chosen:**

This cluster model is not too bad. For the first 3 rows of the cluster (up to column 2 3rd row), each cluster is well clustered and clearly visible. However, after that last 4 scatter plots are plotted on top of each other, making it harder to identify each cluster.

**Dropping Sales**



**Explanation of why this was used for data mining task:**

Each cluster is clearly distinguishable, also, there are enough variety of variables to identify different aspects of characteristics from this cluster model.

1. 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.***

Using pandas sort function, **‘dataframe.sort\_values(‘Sales’, ascending=True)** was applied to the pre-processed dataset containing only store and sales (only the necessary columns to identify the underperforming store),to sort data frame in an ascending order based on the **‘sales’** column. With this, it has been identified, store (store code) **number** **287** is the underperforming store out of all with the sales (Annual sale / Store floor size) of **$300 per m^2**

**Therefore, store code number 287 was dropped (using df = df[df.Sales != 300]) from the data frame and removed from the analysis.**

Resulting with **total of 384 rows of the data- frame.**



**Task 2. The first clustering model**

1. Build a default clustering model with K= 3 and answer the followings:
   1. How many records are assigned into each cluster?

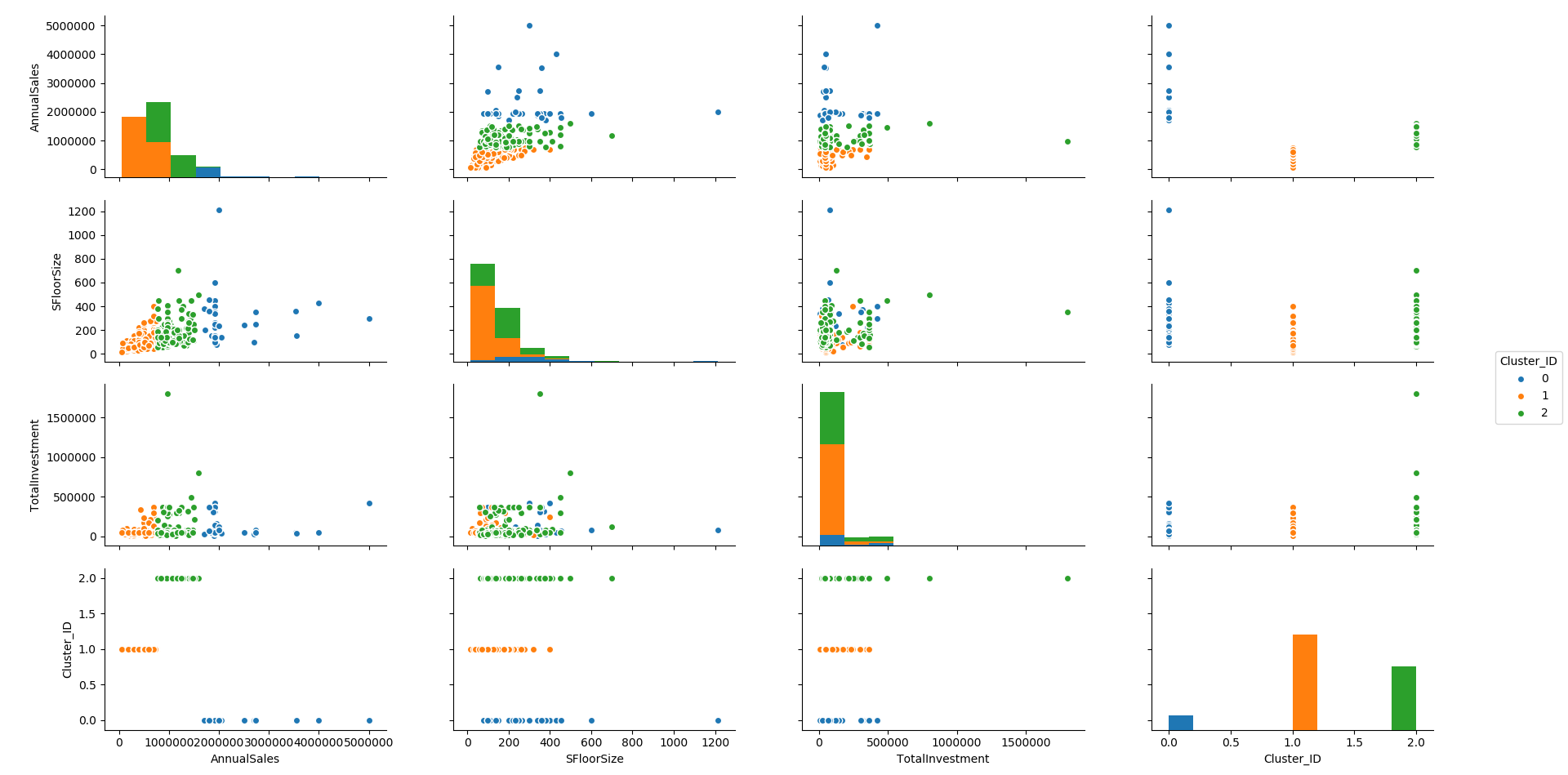
**Non-standardised k = 3 model**

|  |  |
| --- | --- |
| **Cluster ID** | **Assigned Record (Rows)** |
| Cluster 0 | 32 records |
| Cluster 1 | 212 records |
| Cluster 2 | 140 records |
|  | |

* 1. Plot the cluster distribution using pairplot. Explain key characteristics of each cluster/ segment.

|  |  |
| --- | --- |
| **Cluster ID** | **Key Characteristic** |
| Cluster 0 | High Sales (Average of 2.5 million dutch gliders annual sales), variety of floor size (mostly in 100m^2 to 400m^2 range) and around 80% of the investments in the mean range.  Also, small dataset of 32 members (8% of the dataset)  **High Sales, Average Floor Size and Low Investment and small data range** |
| Cluster 1 | Low Sales (Average of 400,000 dutch gliders annual sales), low to median range floor size (14m^2 to 220m^2) and mean range investments (180, 000)  Also, biggest dataset membership of 212 members (55% of the dataset)  **Low Sales, Slightly below Average Floor Size, Medium Investment and Majority (55%) of the data range.** |
| Cluster 2 | Mean range sales (Average of 1 milion dutch gliders), Mean range floor size and low to high investments.  **Medium Sales, Medium Floor Size and evenly distributed investment in low and high range** |

**Non-standardised, k = 3 cluster.**



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

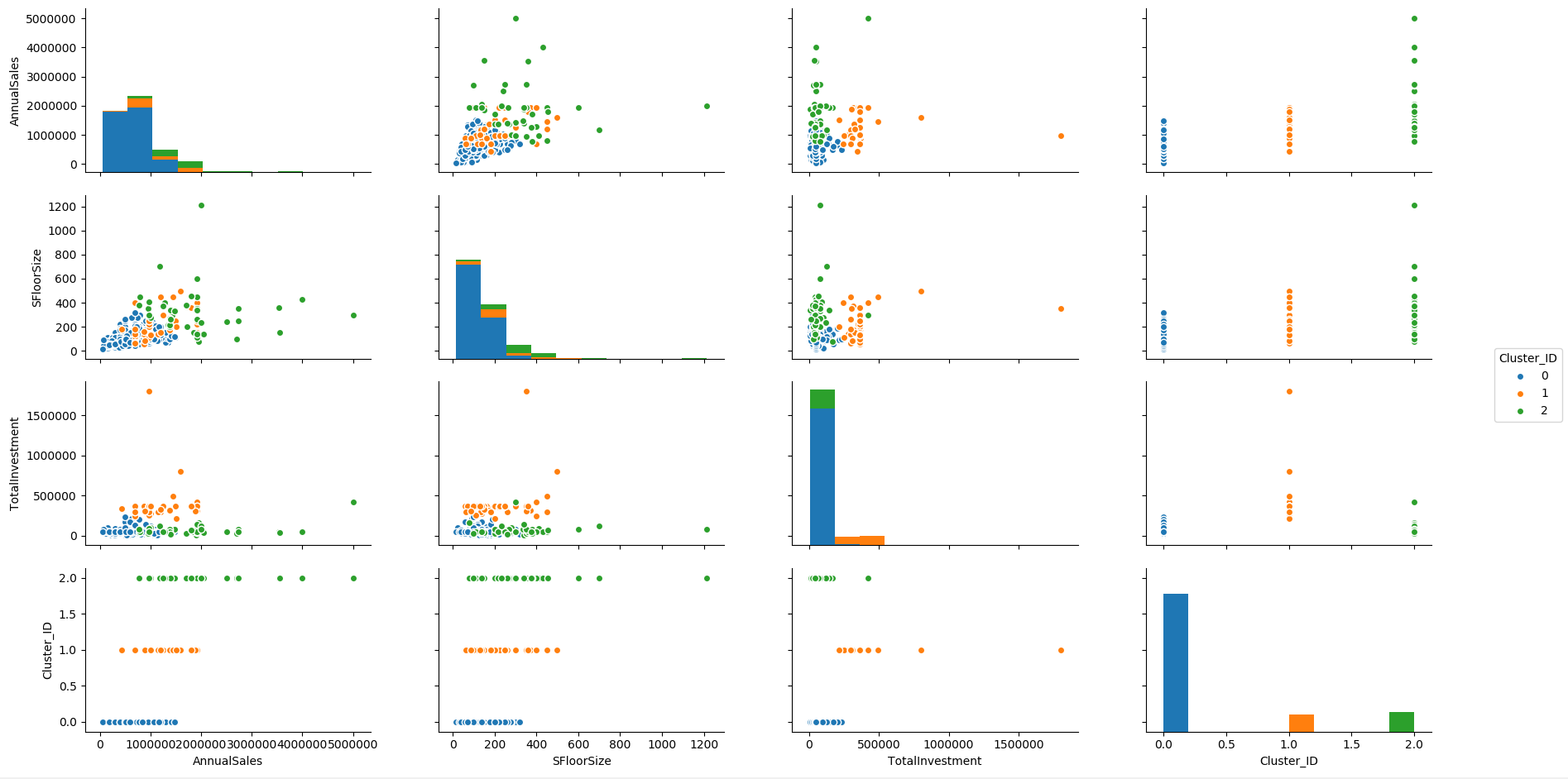
After variable normalization process by using standard scaler function on the data-frames, it has distinguished each cluster more unique. As seen in the first diagram below, each clusters are more defined into their own cluster (high intra-class similarity) and less overlapping with other clusters (inter cluster similarity), unlike the unnormalized version performed above.

This is because variable normalization process evenly transforms every data row into similar scale within the interval [0, 1], resulting in more “rounded dataset” or evenly distributed dataset, meaning every data in the dataset will not be over-ruled by outliers, resulting in abnormal clusters.

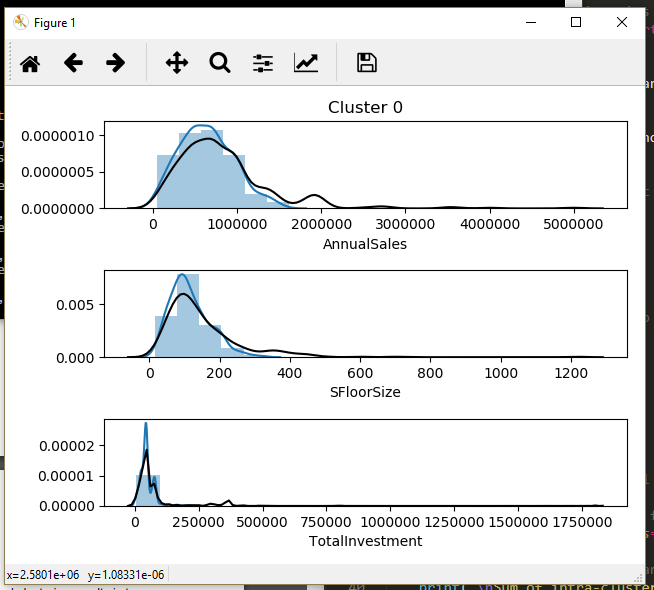
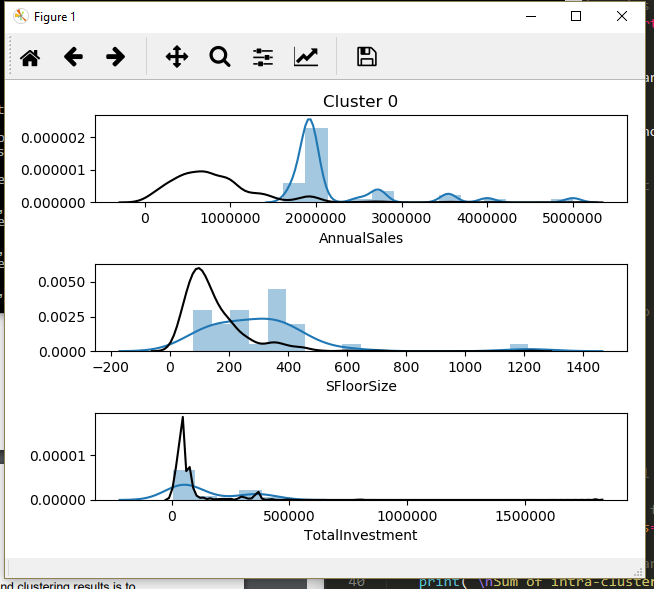
This condition is evident in the second and third diagram below, where second photo is the each variable distribution in cluster 1 where variables were not normalized.

The black lines show the distribution of all record in the clusters where blue line shows the distribution indicates the specific cluster (cluster 0). It can be clearly seen that, un-normalized cluster modelling have abnormal cluster characteristics (blue line) compared to the rest of the model. Which would’ve also affected the black line.

However, after normalization, variables are more clustered together, as it was scaled between 0 and 1, and we can see the characteristic of each cluster more clearly and is logical compared to the un-normalized model.



**Non-Standardised variable distribution Standardised variable distribution**

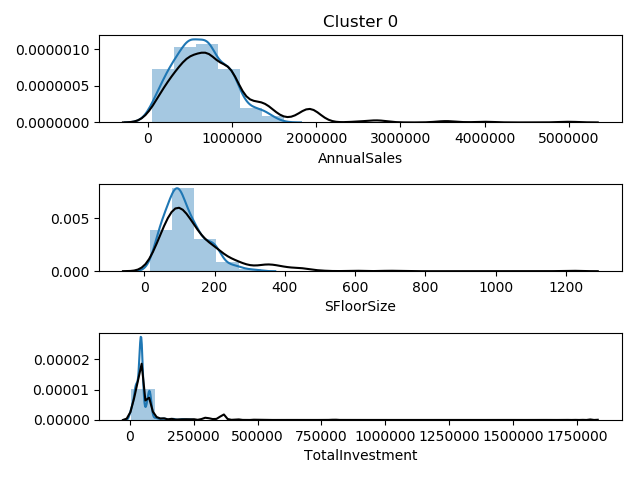


1. Interpret the (best out of 2.1 and 2.2) cluster analysis outcome. In other words, characterize the nature of each cluster by giving it a descriptive label by using distplot.

Using the 2.2 model (k value of 3 and normalized variables);

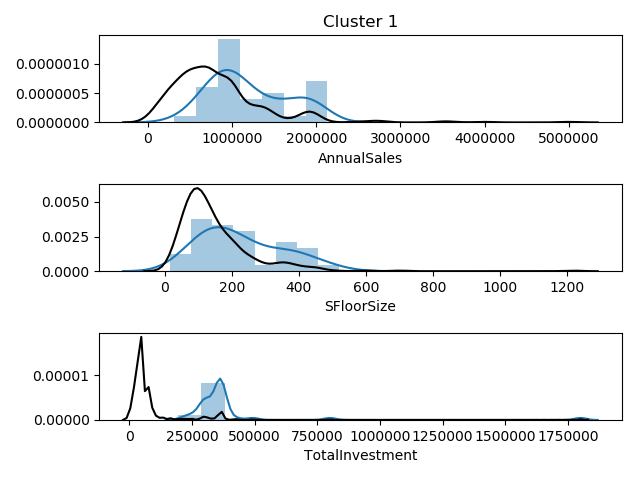
Cluster 0: Slightly left leaning ‘Annual Sales’, slightly left leaning ‘Store Floor Size’ and medium ‘Total Investment’.

Stores in cluster 0 has average store investment rate with with slightly smaller sized store floor size and below average annual sales.



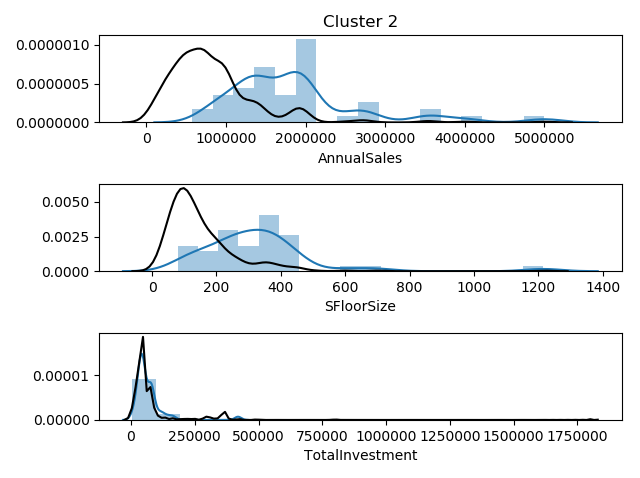
Cluster 1: Right leaning ‘Annual Sales’, Right leaning ‘SFloorSize’ and heavily negatively skewed “Total Investment”.

Stores in cluster 1 have been heavily invested with above average store floor sizes and above average annual sales.



Cluster 2: Heavily negatively skewed “Annual Sales”, right leaning “Store Floor Size” and average “Total Investment”

Stores in cluster 2 has average median total investment rate with large store floor sizes and extremely higher income.



**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.

**Method used**

In the elbow method, two variables are plotted against each other, x being the number of clusters (or K value) and y being the clustering error (known as inertia in sklearn). As K values are inversely correlated with the clustering error. As the task requires, we need to define the optimal K with low clustering error, also, with the minimal clusters. Which can be defined in the K point in the elbow graph where it stops after decreasing rapidly, known as the local minima.

**Brief Explanation**

In the diagram shown below, it rapidly decreases until k-mean value of 3, which then, we can conclude the optimal k value can be defined at either value of **3, 4 or 5**.

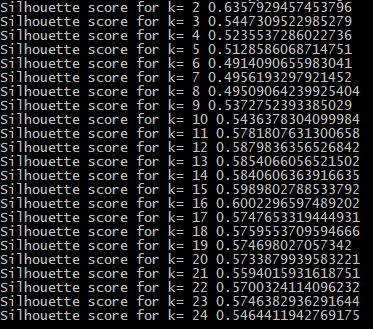
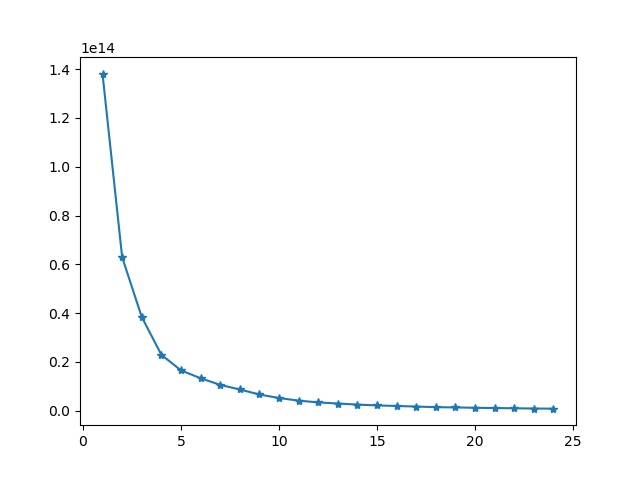
To define the optimal k value between **3, 4 or 5**, we use silhouette score to determine the quality of the clusters (higher the better), which measures how similar an object is to its own cluster compared to other clusters.

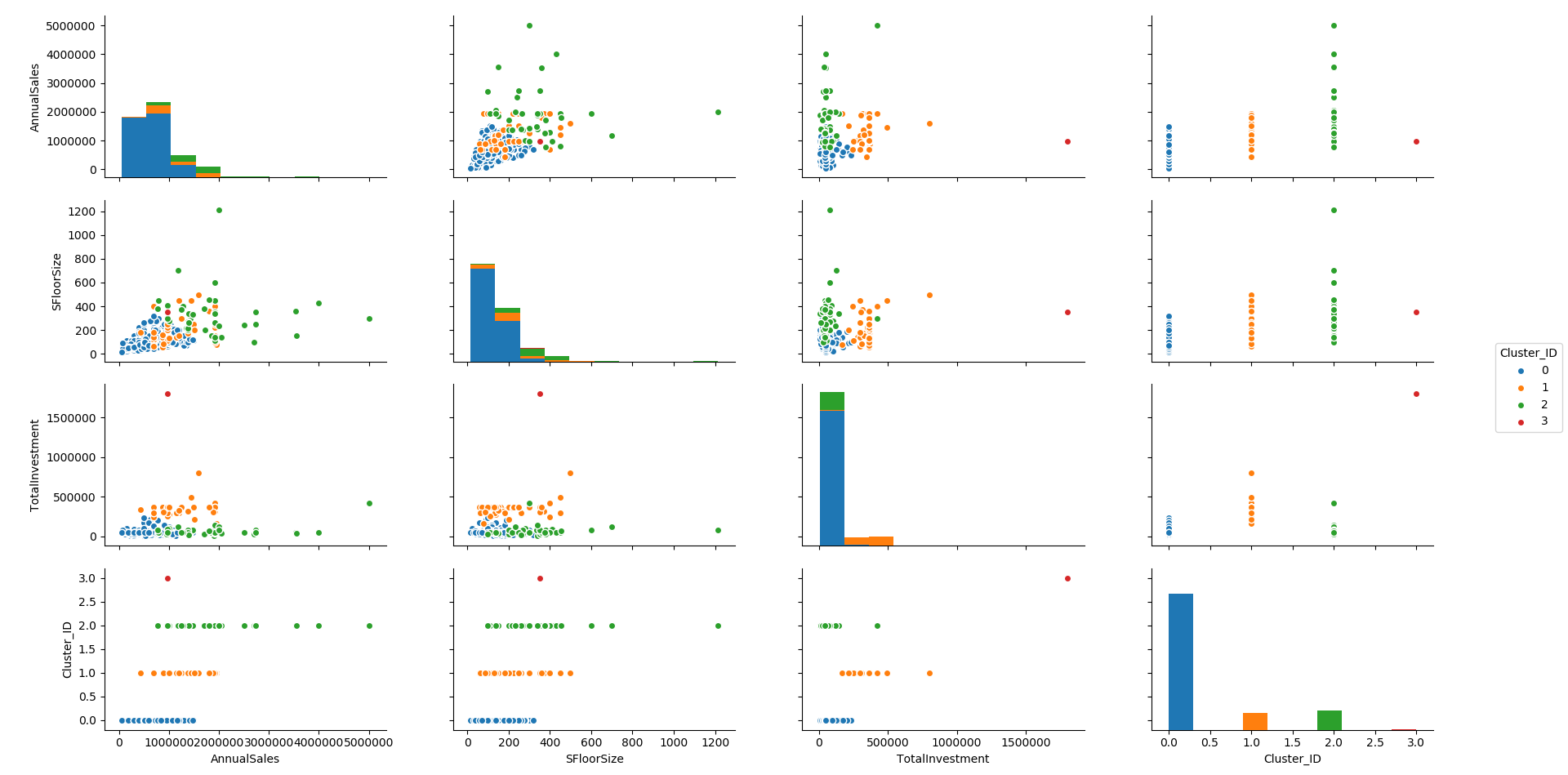
From the second diagram, it is visible that, k value of 3 has a higher silhouette score compared to k value of both 4 and 5. Therefore, **we can conclude that, the optimal k value for our clustering model is 3.**

**Reasoning**

As explained previously, k values are inversely correlated with the clustering error, also, we can see some increased silhouette score after the k value of 5, however, if we choose a k value, well over the elbow point, it is most likely being overfitted by the model, which can be seen when the y value increases again after decreasing. Therefore, k value of 3 was chosen over k values such as 16 which has much higher silhouette score.

It is also evident from the third diagram, using the k value of 4, it has failed to define valuable 4th cluster (red), as there is only 1 data contained in the cluster and it is impossible to gather any useful information.





1. 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?

The best number of clusters that can describe the dataset effectively is at **3 clusters with scaled variables using standard scaler.**

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

From this clustering model study outcome, we can conclude that, stores should be invested with average median rate with adequately larger than median store floor sizes to return good sales from the store. As seen from characteristics in Cluster 2.

**Part 2: Association Mining and its pre-processing**

A supermarket store is interested in determining the associations between items purchased from a stationary department and electronics department. The store has chosen to conduct a market basket analysis of specific items purchased from these two departments.

The POS\_TRANSACTIONS\_2018 data set includes over 400,000 transactions made over the past three months. The following products are represented in the data set:

[A4 copy paper, Drink bottle, Exercise book, USB Flash Drive, DVD media, Sketching Markers, Watercolor Set, Mini Stationery Set, Lanyards, Wristbands, Laminator, Power Bank, Photo Frame, Certificate Frame, Digital Clock, Flash Card, Puzzle]

|  |  |
| --- | --- |
| Name | Description |
| LOCATION | Point of sale device identification number |
|  | (e.g. for Register 3) |
| TRANSACTION\_ID | Unique transaction identification number for a |
|  | given sale. A sale may include several products and thus |
|  | the same transaction id may occur over several rows. |
|  |  |
| TRANSACTION\_DATE | Date of transaction |
| PRODUCT\_NAME | Product Purchased |
| QUANTITY | Quantity of this product purchased (always |
|  | set to 1 by a point of sale device) |

Your task is to conduct association analysis on this data set. Answer the followings in relation to this data and analysis.

**Task 4. Association Mining**

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

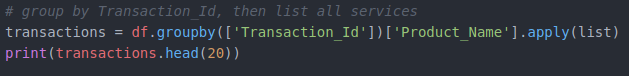
No quality issues were found, due to all variables being of the correct type and no empty values detected.

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

“Transaction\_Id” and “Product\_Name” variables were used in the analysis.

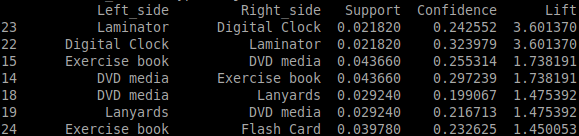
“Transaction\_Id” corresponded to the unique transaction occurring at the given Location and Transaction\_Date, while “Product\_Name” corresponded to one item purchased.

To produce a list of the products purchased in one transaction, “Transaction\_Id” values were grouped and the followed by generating a list of the “Product\_Name” values, belonging to the transaction.



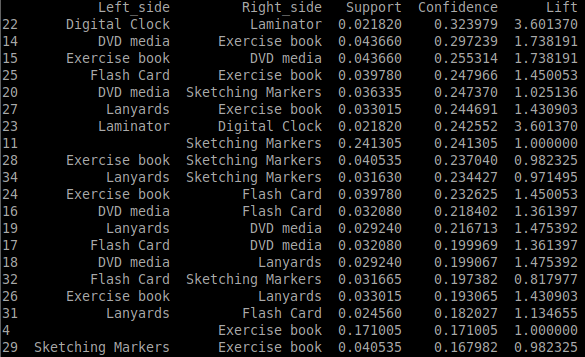
1. Conduct association mining and answer the following:
   * 1. What is the highest lift value for the resulting rules? Which rule has this value?

After the association mining was performed, the produced rules were sorted by lift in the descending order, to determine the rule with the highest lift value. The rules with “Laminator” and “Digital Clock” bought together had the highest Lift value of 3.601370.



* + 1. What is the highest confidence value for the resulting rules? Which rule has this value?

To determine the rule with the highers t confidence, the rules were sorted by confidence in the descending order. The rule with “Digital Clock” and “Laminator” bought together had highest confidence of 0.323979.



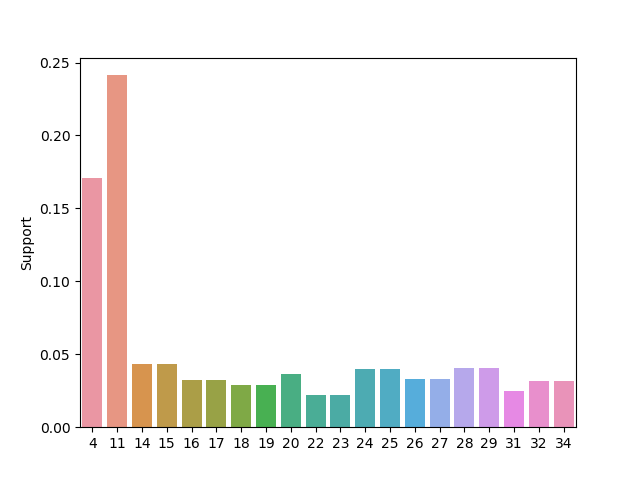
* + 1. Plot the confidence, lift and support of the resulting rules. Interpret them to discuss the rule-set obtained.

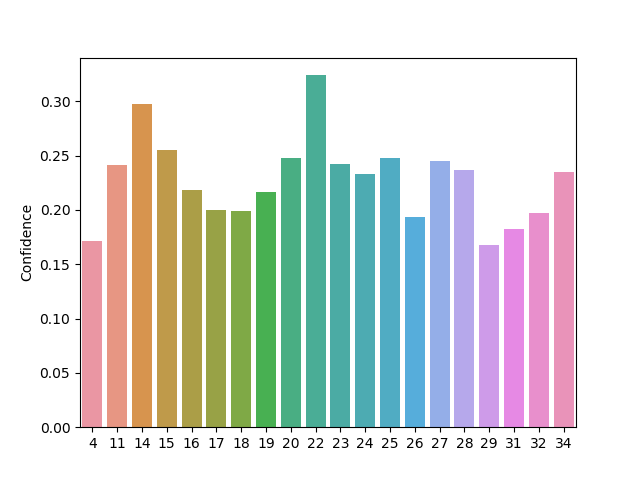
From the produced graphs it can be seen, that the rule with an index of 11 has the highest confidence. However, it does not have any importance, as it does not only have a lift of 1, but also is a rule with only one purchased item. Same follows for the rule with an index 11, which has second highest confidence of the visualised rules.

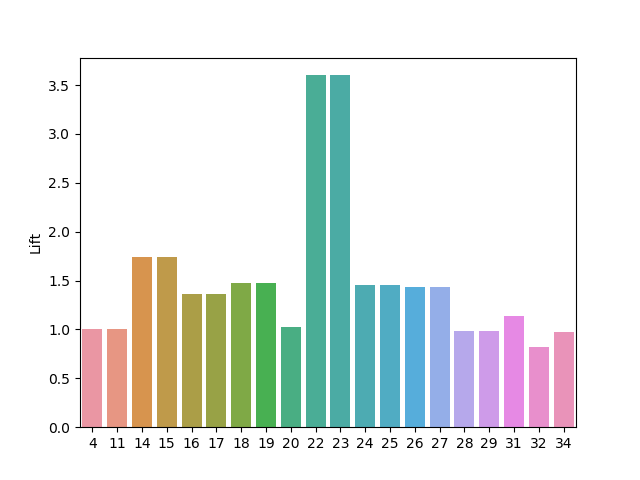
The rule with an index of 22, displaying “Digital Clock” and “Laminator” items, performed with the highest confidence of 0.323979. Also, it is one of two rules with the highest Lift of 3.601370. It has support of 0.2182, which means that this rule is rarely occurring. However, based on it’s Confidence and Lift the rule is strong and can be important for the decision making.

The rule with an index of 14, displaying “DVD media” and “Exercise book”, performed with second highest confidence of 0.297239. However it does not have a very high Lift, which is 1.738191. It has a low support of 0.04366.

The rule with an index of 23, displaying “Laminator” and “Digital Clock” items, had a lift of 3.601370, which is the highest lift value of the produced rules. The items of this rule are the same as in the rule with an index of 22, however it has a lower confidence of 0.242552.







1. The store is particularly interested in products that individuals purchase when they buy “Exercise book”.
   1. How many rules are in the subset?

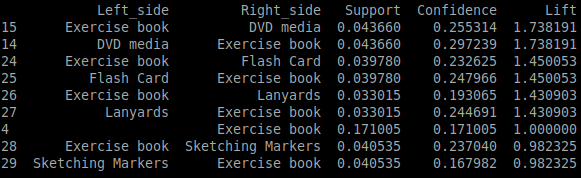
There are 9 rules in the subset, where “Exercise book” is purchased.



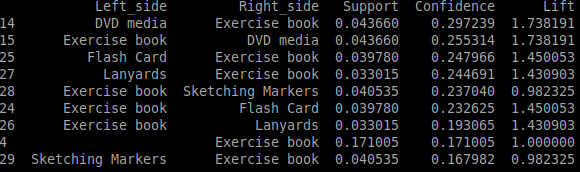
* 1. Based on the rules, what are the other products these individuals are most likely to purchase?

Based on the confidence and lift values, it was determined that the likeliest product to be purchased with “Exercise book” is “DVD media”. Second most likely item to be purchased is “Flash Card”.

Subset rules sorted by lift:



Subset rules sorted by confidence:



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

“Digital Clock” and “Laminator” are most likely items to be bought together, because this rule has both highest confidence and lift.

The store was particularly interested in the association rules including “Exercise book”. From the analysis of this subset, it was determined that “DVD media” and “Flash Card” were two items, most likely to be bought with “Exercise book”. One of the examples, how this could be used, is that exercise books and DVDs with flash cards could be placed at the different parts of the store, so the customer would see variety of other products.

**Part 3: Text Mining (Clustering) the News Stories**

**Task 5. Text Mining**

A leading news corporation, BBC, is planning to start an online personalised news story service. They have a collection of individual stories in the form of a compressed single file. Perform text mining on the BBC(bbc.json) dataset to determine clusters of stories based on similar topics. Answer the following in relation to this data and analysis.

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

**Included variables**

|  |  |  |
| --- | --- | --- |
| **Variables Included** | **Roles and Measurement level** | **Reason for choosing** |
| **text** | **Role:** Independent variable  **Measurement level:** Qualitative – Categorical Nominal | Each ‘text’ variable holds individual story from the BBC news.  Each story is pre-processed for text mining. |

**Removed variables**

|  |  |  |
| --- | --- | --- |
| **Variables Not Included** | **Roles and Measurement level** | **Reason for choosing** |
| **id** | **Role:** Independent variable  **Measurement level:** Qualitative – Categorical Nominal | As the name suggest, this variable is a unique identifier starting from 0 and incrementing by 1, and provides no real value to our analysis, therefore, removed from analysis. |

1. Can you identify data quality issues in order to perform text mining?

As text mining is unstructured data-mining task and is heavily dependent on text pre-processing, many data pre-processing steps were required before performing text mining, and this task was no different as this data-set (json file) was delivered in data rows of paragraph text.

Firstly, Lemmatization, which is cleaning up redundant words of different forms into lemma form, such as Ponies, Pony’s Ponys’ into one single lemma form of Pony.

Secondly, removal of stopwords or common words that do not add value into our data mining task and also, excluding punctuations.

Then pre-processed words were contained into a tf/idf version of matrix.

1. Based on the ZIPF plot, list the top 10 terms that will be least useful for clustering purpose.

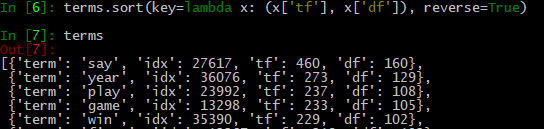
The top 10 terms that will be the least useful for clustering purpose would be the top 5 most common term and last 5 most common term.

Which were;

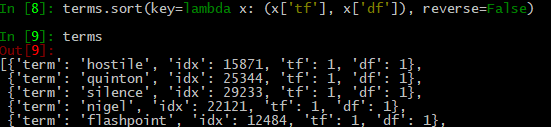
**Top 5 common:** say, year, play, game, win.

**Last 5 common**: hostile, quinton, silence, nigel, flashpoint.

**Top 5 most frequent terms**



**Last 5 most frequent terms**



1. Did you disregard any frequent terms? Justify your selection.

Yes, only the terms that have occurred in minimum of once in 2 documents and maximum of 80% of all documents have been kept and rest terms that did not comply with these two rules have been disregarded.

This can be evaluated through previously plotted ZIPF plot, as ZIPF’s law is the frequency of any word that is inversely proportional to its rank in the frequency table, meaning as terms that appear on the left side are most frequent words and terms that appears as it goes right, are rare words.

As seen in the plot previously, top-left of the, and bottom-right, do not provide valuable information and only takes up memory and decrease the efficiency when processing. Therefore, terms not following the two rules were all disregarded.

1. Justify the term weighting option selected.

Term weights are determined on how close the individual term is to its own cluster’s centroid, which would be have higher term value or its importance/relevance to the cluster.

1. What is the number of input features available to execute clustering?

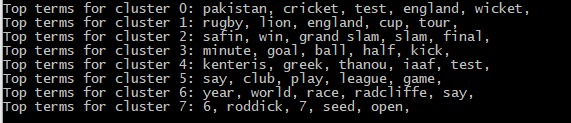
(FYI: Note how the original text data is converted into a feature set that can be mined for knowledge discovery.)

After disregarding invaluable terms, input features have been decreased from 36,385 to 6,923.

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

8 clusters were generated in this text-mining model.

|  |  |  |
| --- | --- | --- |
| **Cluster-ID** | **Top-Terms** | **Generalised name** |
| Cluster 0 | Pakistan, cricket, test, England, wicket | **Cricket** |
| Cluster 1 | Rugby, lion, England, cup, tour | **Rugby** |
| Cluster 2 | safin, win, grand slam, slam, final | **Tennis (Australian Grand Slam)** |
| Cluster 3 | minute, goal, ball, half, kick | **Soccer** |
| Cluster 4 | kenteris, Greek, thanou, iaaf, test | **Athletics** |
| Cluster 5 | say, club, play, league, game | **Sport clubs** |
| Cluster 6 | year, world, race, Radcliffe, say | **Marathon** |
| Cluster 7 | 6, Roddick, 7, seed, open | **Tennis** |



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

Premier, television, graham, mile, wasps, subuse, benitez, cole, recall, Russian, wkt, el, robben, morgan, Edward.

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

From these clusters, we can group each individual stories into this cluster model, then, we identify which kind of stories that users are interested in and show them the relevant stories.

Or we can gather individual news stories then automatically group them into this cluster model, and have different sports sections such as; cricket, rugby, tennis, soccer, etc.

**Part 4: Web Mining the Log Data for a Website**

**Task 6. Web Mining**

For an e-commerce business, the website structure and site plan were established with the efficiency and usability in mind, but its effectiveness was not verified. Only basic statistics have been produced through simple report and query techniques, but they provide no means for sophisticated web site analysis and predictions. Your task is to determine the patterns of user browsing the website and analyse those patterns to provide the results and recommendations to the website owner.

You have been provided with a log file in CSV format, WEB\_LOG\_DATA.csv. This was originally a text file and was processed with the steps required for web usage mining as explained in the lecture. The processing steps were: (1) removing unproductive items from the log file such as graphics, sound etc; and (2) identifying users and sessions based on IP address, date and time. The goal of user session identification is to divide the page access of each user into individual sessions.

The dataset consists of 6 columns namely IP address, timestamp, request, step, session id and user id.

Your task is to **apply a data mining operation**, such as classification or clustering or association mining, to the pre-processed data set. Answer the followings in relation to this data and the analyses that you have chosen.

1. For each data mining operation:
   1. Rationale behind selecting the data mining method.

The association operation was used for the task to indicate user activity/behaviour across the site.

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

To obtain the best indication of user behaviour the variables “Session” along with “Request” were grouped and used in the association mining. This will allow for the generation of rules that will show popular user navigation.

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

A major data quality issue found was that request URLs had instances of the same URL with minor variations (i.e. “/newfarm/pricelist” and “/newfarm/pricelist/”, note the extra forward slash). As such, rows where the Request column contained a forward slash at the end was replaced with and empty string.

* 1. Discuss the results obtained. Discuss also the applicability of findings of the method. You should include only a high-level managerial kind of discussion on the findings. It should not just be interpretation of results as shown in results panel.

The rules generated from the association mining show that users who visit the “newfarm” page also view the price subset pages such as “/newfarm/pricelist” and “/newfarm/pdf/Web\_Price\_List.pdf”.

Users are also rather interested in information from pages such as the “guarantee” and “more” as these link to each other

**Instructions**

1. The assignment is due on 21st Oct at 11:59pm. It is a firm deadline.
2. You should submit the assignment report via Blackboard Assignment.
3. The assignment (record, transaction, text and web mining project DMProj2) will be **marked in the** **practical class in Week 13**. We will check the outputs along with the assignment report, to assignyou marks. The entire team should be present to show the project result and answer the questions raised by marker. We will ask questions to each student, and will assign about 15% of total marks as per individual performance.
4. The datasets required for this assignment can be found on BlackBoard with the file named as **casestudy2-data.zip**. It includes four datasets:
   1. MENS\_CLOTHING\_SALES\_2018.csv to perform clustering
   2. POS\_TRANSACTIONS\_2018.csv to perform association mining
   3. bbc.json to perform text mining
   4. WEB\_LOG\_DATA.csv to perform web mining
5. Name the case-study report as **casestudy2.docx or casestudy2.pdf.** The submission file should include a cover page with Student ID number and full name (as in QUT-Virtual) for all students, along with the group name. Combine this file with your **team contract**, and name the compressed file as **casestudy2.zip.** Submit the zip file on **Blackboard (under assessment panel**

**Assignment 2.**

6 . A report should be submitted via online submission answering each question of the case study. There is no need of including introduction, summary, conclusion or references in the report. The report should just include responses to the questions set in the case-study. Some answers may require screen shots. Use them as needed. You can even include your own table detailing those results based on the outcomes. While you may like to go into extreme detail about, you will not have the space to do so. Rather, write down the important points and attach the important screen dumps, to show that you have thought the matter through.

7 . This is a group assignment. The team size is three. You can continue the same group as in case study 1. If you have formed a new group after assignment 1, please notify the lecturing staff. They will remove you from the existing group. In this case, you need to register your new team at Blackboard.

1. The group is to be ARRANGED and MANAGED by you. As in real life, the performance of the individuals in the team shall be judged by the performance of the team together, so choose your partners carefully.
2. Of course, the work your group hand in must be your own; no collaboration or borrowing from others groups is permitted. Read the Assessment Policies on Blackboard or QUT Website.

**Assignment Criteria Sheet**

|  |  |
| --- | --- |
| **Criteria** | **Comments and scoring** |
| Non Submission of all components/ evidence of plagiarism | 0 |
| Has demonstrated a task with a working model with /without | 1-5 |
| submission and demonstrates the ability to run the program and |  |
| add some components. |  |
| Has demonstrated a task with a working model having a data | 6-11 |
| source, and models with substantial but incorrect |  |
| implementation of at least one of the four parts. |  |
| Questions were poorly answered. |  |
| Has implemented all tasks with at least two being substantially | 12 |
| correct. Shows some understanding of concepts with some |  |
| success in applying knowledge. Only basic questions were |  |
| answered. |  |
|  |  |
| Has implemented all the tasks: | 13-15 |
| One mining task is fundamentally correct, with substantially |  |
| correct results which may contain minor errors. Response to |  |
| questions shows fundamental understanding of terms and |  |
| concepts. |  |
| Has fundamentally correct implementation of all tasks i.e. | 16-18 |
| selection of correct variables in data, correct allocations, |  |
| understanding, and explanation of clusters, findings association |  |
| rules, finding clusters in text data with good term features, and |  |
| application of an appropriate data mining operation to the web |  |
| log data. Shows competency in applying text mining. Many |  |
| questions have been reasonably answered. Demonstrate a good |  |
| understanding of the methods and terms used in clustering, |  |
| association mining, text mining and web mining, during written |  |
| and verbal analyses. Some minor errors are allowed. Written |  |
| application is required to be of reasonable standard. |  |
| Has implemented all of the requirements above with very few | 19-21 |
| errors. A strong focus on application of tools, and evaluation and |  |
| interpretation of results is evident. |  |
| All of the criteria above are met, extensive model generation and | 22-25 |
| analyses have been conducted to produce exceptional outcomes. |  |
| Have applied principles learnt in lectures to enhance the results. |  |
|  |  |