

## CAB330: Case Study 2

**Due date: 22<sup>nd</sup> Oct, 2017**  
**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. 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.

### Task 1: Descriptive Data Mining - Clustering

The MODEL-CAR-SALES data set gives the number of four different car models sold at stores of a particular car agency. Each row represents an individual store. There are eight columns in the data set.

- Location Number and DEALER CODE contain the store identification number and store label code respectively.
- REPORT DATE is the date that the report was generated.
- The next four columns contain the number of each type of model sold. The sales numbers are over a specified time period.
- K\_SALES\_TOT is a derived variable that shows the total sales for each store.

Name	Description
Location Number	Numerical code for the store (unique, occurs only once in the table)
DEALER CODE	Text identifier for the store (also unique, only occurs once in the table)
REPORT DATE	Date of the data extraction
HATCH	Number of hatch back model cars sold by the store
SEDAN	Number of sedan model cars sold by the store
WAGON	Number of station wagon model cars sold by the store
UTE	Number of utility/tray back model cars sold by the store
K_SALES_TOT	The total sales for the store (thousands of dollars)

The company has noticed that stores seem to have an overall preference for certain combinations of model types, with some stores referencing a predominance of sales of two model types; e.g. *Hatch and Sedan* or *Hatch and Wagon*, thus creating segments in their market. They want to find the optimal number of product sale segments, to allow development of advertising to match the sales in stores of each segment.

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

1. Can you identify data quality issues in this dataset such as unusual data types, missing values and others?
2. Identify a car model that is underperforming in sales. Based on your reporting, the company does not want to focus their efforts on this car model anymore and has decided to drop it from manufacturing. Now onwards, the selected car product should not be part of analysis.
3. What variables did you include in the analysis and what were their roles and measurement level set? Justify your choice.
4. Build a clustering model with K=3 and answer the following:
  - a. How many records are assigned into each cluster?
  - b. Plot the cluster distribution using pairplot. Explain key characteristics of each cluster/segment
5. What is the effect of using the standardization method on the model above? Does the variable normalization process enable a better clustering solution?
6. Using elbow method and silhouette, find the optimal K. What is the best K? Explain your reasoning.
7. Using both pairplots and cluster variable distplots, interpret the cluster analysis outcome (for the model with best K). In other words, characterize the nature of each cluster by giving it a descriptive label.

## Task 2: Descriptive Data Mining - Association

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

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)

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

[Bar soap, Bows, Candy bars, Deodorant, Greeting cards, Magazines, Markers, Pain relievers, Pencils, Pens, Perfume, Photo processing, Prescription medications, Shampoo, Toothbrushes, Toothpaste, Wrapping paper]

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

1. Can you identify data quality issues in this dataset for performing association analysis?
2. What variables did you include in the analysis and what were their roles and measurement level set? Justify your choice.
3. Conduct association mining and answer the following:
  - a. What is the highest lift value for the resulting rules? Which rule has this value?
  - b. What is the highest confidence value for the resulting rules? Which rule has this value?
  - c. Plot the confidence, lift and support of the resulting rules using pairplot with number of items in a rule as hue. Interpret them to discuss the rule-set obtained.
4. The store is particularly interested in products that individuals purchase when they buy "Candy Bar".
  - a. How many rules are in the subset?
  - b. Based on the rules, what are the other products these individuals are most likely to purchase?

### Task 3: Text Mining

A leading news corporation is planning to start an online personalised news story service. They have a collection of individual stories. Perform text mining on the BBC dataset to determine clusters of stories based on similar topics. Answer the followings 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.
2. Based on the ZIPF plot, list 10 terms that will be not useful for clustering purpose.
3. Did you disregard any frequent terms? Justify their selection.
4. Justify the term weighting option selected.
5. Perform term/feature selection based on ZIPF plot. 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.)
6. Perform your clustering and find the optimal K using both elbow and silhouette method. State how many clusters are generated? Name each cluster meaningfully according to the terms that appear in the clusters?
7. Apply LSA/SVD on the dataset before performing document clustering. Observe and describe the changes in the clustering process and results.
8. Describe how these clusters can be useful in the online personalised news story service planned.

## Task 4: 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 user browsing patterns of the website and analyse those patterns to provide recommendations to improve the website.

You have been provided with a log file in CSV format, WEB\_LOG\_DATA. 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 organise 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. Rationale behind selecting the data mining method.
2. What variables did you include in the analysis and what were their roles and measurement level set? Justify your choice.
3. 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.

## Distribution of Marks (Total: 25 marks)

We would mark your data mining projects in the Week 13 practical class. You should be prepared to show your final diagrams 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.

**Clustering analysis and pre-processing (7 marks)**  
**Association Mining and pre-processing (5 marks)**  
**Text Mining and pre-processing (7 marks)**  
**Web Mining (6 marks)**

## Instructions

1. The assignment is due on 22<sup>nd</sup> Oct. 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) will be **marked in the practical class in Week 13**. We will check the code, plots and results, along with the assignment report, to assign you 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-pv.zip**. It includes four datasets:
  - a. MODEL-CAR-SALES to perform clustering
  - b. POS\_TRANSACTIONS to perform association mining
  - c. BBC to perform text mining
  - d. WEB\_LOG\_DATA to perform web mining
5. Name the case-study report as **casestudy2.doc**. The word 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 your **source code** and name the compressed file as **casestudy2.zip**. Submit the zip file on **Blackboard (under assessment panel Assignment 2)**.
6. The **project report** should be divided into four parts according to each task, each part starting from a new page. 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. Answer the questions in the case study for each model appropriately and succinctly. If a case-study step is about conducting a process, you do not have to provide an explanation or a screen shot. Include the final screen shot when you added all kinds of nodes in a particular analysis. However if a question such as “Examine the results of clustering/association mining” is asked, you then need to explain what, why or why not? 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.
8. 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.
9. 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 submission and demonstrates the ability to run the program and add some components.	1-5
Has demonstrated a task with a working model having a data source and process with substantial but incorrect implementation of at least one of the seven components. Questions were poorly answered.	6-11
Has implemented all tasks with at least two being substantially correct. Shows some understanding of concepts with some success in applying knowledge. Only basic questions were answered.	12
Has implemented all four tasks: One mining task is fundamentally correct, with substantially correct work flow process which may contain minor errors. Response to questions shows fundamental understanding of terms and concepts.	13-15
Has fundamentally correct implementation of all tasks i.e. 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.	16-18
Has implemented all of the requirements above with very few errors. A strong focus on application of tools, and evaluation and interpretation of results is evident.	19-21
All of the criteria above are met, extensive model generation and analyses have been conducted to produce exceptional outcomes. Have applied principles learnt in lectures to enhance the results.	22-25