PROJECT PORTFOLIO

COMP 602 – Enterprise Information Management

Liliana Quyen Tang

Program: Master of Science in Information Systems

Instructor: Jon Dron

December 9, 2018

Table of Contents

A. Analyse business data needs and requirements for data-driven systems

1. My Introduction section in the essay
2. Statement of the research problem section in my essay

B. Argue the strengths and weaknesses of different approaches to data management

1. Literature Review section in my essay
2. Argue strengths and weaknesses of R versus SQL versus Python and under which circumstances, one should stand out

C. Solve problems in technology, technique and process relating to database management and design

1. Essay

2. Presentation

D. Independently and reflectively research issues, technologies, processes and tools in information management

1. Essay

2. Presentation

3. Reflection

    a. Week 1

    b. Week 2

    c. Week 3

    d. Week 4

E. Critically evaluate information and data technologies in the context of organisational needs

1. Evaluate when to use Python versus R versus SQL in the presentation and when answering questions during the presentation
2. Questions asked during the presentation

F. Be a reflective practitioner in the information management field

1. Project Week 1
2. Project Week 2
3. Project Week 3
4. Project Week 4

A. Analyse business data needs and requirements for data-driven systems

1. My Introduction section in the essay:

Over the past decades, with the increasing demand in artificial intelligence, the tremendous need to save time from shoppers, and the paramount call for attention and improvement in technology to attract more customers from retail industry, mining customer behaviours has been proving its crucial importance. Supermarkets start saving costs for hiring cashiers and associates by moving to ecommerce. Two largest retailers, Costco and Walmart, both have websites for online shopping across North America countries, including the United States and Canada while still running physical stores. Customers move toward hybrid shopping style – purchasing products both online and in-store - to save time and money while still being able to main a healthy lifestyle for meals made from fresh and carefully chosen ingredients in-store. However, there has been a gap in demand and supply in the market. Customers have become more well-informed and knowledgeable shoppers. Entrepreneurs have developed more interest in retail, so the question of attracting those knowledgeable shoppers comes into focus. To seek patterns in customer behaviours, the most effective way is through data mining, like David Ciancio, Senior Customer Strategist at Dunnhumby, once states, "Grocery companies need to think of themselves more like tech companies" (Haddon 2017).

Data mining is a growing yet challenging field due to limited access to real-life data for experimental purposes. Big companies, such as Amazon, tend to refuse to share their sales and marketing strategies (Haddon 2017). Therefore, through academic research, this paper aims to propose an approach in data mining to implement a system that acts as a placeholder for three well-known and existing rule-mining based, event-prediction, and web-personalization systems. Beginning with a concept walk-through of three data mining algorithms employed, the paper then uses a diagram to describe business processes of the proposed system. Later, it makes clear the concept of data mining in CCSS and how data mining stands out from other statistical techniques, such as hypothesis testing. Lastly, the author ends the discussion by briefly discussing strengths and weaknesses of the proposed system.

2. Statement of the research problem in my essay:

The author's initial scenario is to design and develop a database for a Point-of-Sale (POS) system in grocery stores, so this research further takes the full advantage of those information stored in the database by applying appropriate data mining techniques. In addition to gain customer satisfaction, the CCSS system plays a crucial role in boosting profit for retailers by incorporating three common and existing systems, including rule-mining based systems acting as a recommendation tool for customers; event-prediction systems contributing as a supply-demand forecast for store managers to order an appropriate amount of items to minimize waste from expired products; and finally web personalization systems that greatly benefits e-grocery shopping model by coming up with a wide-ranging recommendation as advertisement while customer are using Internet for other websites. The author of this paper often promises to herself that she would stop spending on clothes, and advertisement about fashion deals shows up in one of the websites she was using for school and work, she usually ended up re-visiting the fashion website again and breaking promises to herself. She soon figured out to download the content of the app as a pdf file to her local storage, so she could avoid seeing those advertisement.

Therefore, the author thinks web personalization is a powerful innovation that can dramatically foster sales.

B. <u>Argue the strengths and weaknesses of different approaches to data management</u>

1. Literature Review section in my essay:

Griva, A., Bardaki, C., Pramatari, K. & Papakiriakopoulos, D. (2018). Retail Business Analytics: Customer Visit Segmentation Using Market Basket Data. Expert Systems with Applications, 100 (2018), 1-16

Griva, Bardaki, Pramatari, and Papakiriakopoulos (2018) examine an effective business analytics approach in retail through a variety of studies that have a strong emphasis in the application of data mining techniques on customer behaviours. Griva et al. (2018) further categorize those approaches into two streams, corresponding to the main focus of each study, including customer segmentation and market basket. While customer segmentation mainly focuses on the whole picture of the total purchases a shopper makes, regardless of the number of their visits, market basket analysis tries to seek out how each item during a single visit relates to one another in one single purchase. They then point out the research gap with respect to the two above approaches by analyzing each individual's scope of the analysis and product taxonomy. Basket analysis is not a practical approach since it does not cover full scope of area of interest and retailers need to understand a bigger picture of each visit. Customer segmentation, on the other hand, depicts a much broader extent, though is still not sufficient to provide a detailed set of information that fulfills business needs. Since each approach has their own pros and cons, the authors later propose another approach that employs clustering technique to achieve the most out of strengths from above mentioned systems while eliminating their weaknesses. The paper wraps up with a step-by-step set of processes of the proposed framework, covering business and data understanding; modelling, such as product taxonomy adjustment, clustering sampling, and input data adjustment and clustering; evaluation which takes into consideration both technical and business perspective. This resource greatly enriches the author's knowledge in clustering technique which is one of the three main data mining algorithms she employs in her proposed system. She completely agrees with the arguments regarding the strengths and weaknesses of two common approaches in data mining, and how clustering sample stands out from the crowd by its ability to accomplish both "segmentation" and "characterization" in studying customer behaviour through their visits.

Min, H. (2006). Developing the Profiles of Supermarket Customers through Data Mining. The Service Industries Journal, 26)7), 747-763

Min (2006) conduct an experiment to study customers behaviour by selectively distribute questionnaire to regular grocery shoppers. The author introduces and employs the concept of decision tree through "IF-THEN" statements to predict customer behaviours, so management team can have a clear target marketing or promotional actions. Furthermore, Min describe three

major steps in constructing decision trees, including data collection, data formatting, and rule induction. The outcome points out some crucial implication. First, the volume of grocery purchases negatively depends on the number of stores customers have to visit frequently and their shopping frequency. For example, if a customer often goes to store A to buy vegetables, store B to buy meat, and store C to buy canned food, meaning they have to visit three stores, so in comparison with a customer who is able to shop everything at one store, the customer visiting more stores will have lower volume of purchases. Similarly, if a customer goes shopping every week, their order will be much smaller than a customer going shopping monthly. Management team can use this finding to comfortably accommodate their customers. In particular, if the majority of customers shop with low volume of purchases, the store should have more self-checkout than cash registers run by cashiers, and vice versa. This accommodation in turn is a remarkable win in customer satisfaction. Secondly, the author figures that marital status also affects either a customer prefers self-checkout. In fact, married couples often go shopping for grocery together while single customers go on their own and often prefer fast checkout from self-checkout. Therefore, depending on if the store has more unmarried shoppers or married customers, management team should develop an innovative approach accordingly to increase customer satisfaction. The study also shows that very young (under 20) of older (over 50) often more appreciate employee courtesy which is in turn the first reason for them to become a loyal customer since they often need more assistance from employees due to their lack of experience or physical limitation. Fourth, the outcome of the classification technique from the survey points out that senior citizens who rely on their limited retirement savings are often more sensitive to price, so supermarkets should have more discount promotion for these customers. This study strengthens my previous knowledge in the application of classification techniques through a real-life example that perfectly fits my scenario. Although the article does not provide much advanced technical details, in fact, data formatting phase was conducted in Excel rather than other common programming languages in data science, such as R, Python, or SQL, it provides us with a set of specific recommendations for each outcome of data mining process.

Witten, I. & Frank, E. (2005). Data mining: practical machine learning tools and techniques. 2nd ed. San Francisco, CA: Elsevler

The book comprehensively covers machine learning – a broader term that include data mining - tools and techniques. Due to limited time and background information in data mining, the author of this paper was not able to fully understand each chapter of the book. However, the book is ranked among one of the most helpful books for a beginner and recommended by instructor of COMP 682 – Data Mining course at Athabasca University, so the author refer and circle back to this book whenever coming across a difficult term during her literature review. The book clearly explains the concept in common language with real-life examples, as such, it is very straightforward to understand. In addition, the content is clearly divided into many independent sections, so the author can skip chapters that are not particularly relevant to her research study.

Yang, Y., Liu, H. & Cai, Y. (2013). Discovery of Online Shopping Patterns Across Websites. INFORMS Journal on Computing, 25(1), 161 – 176

Yang, Liu, and Cai (2013) employ market basket analysis which is an application of association rule technique in data mining to define online shopping patterns. The authors employ OSP-Tree and OSP-Level methods to mine the complete set of frequent online shopping patterns. OSP-Tree is the common approach due to it efficiency. However, when memory is limited, OSP-

Level comes into play. This paper provides with a significant insight as for how data mining is used in practice and with the real-life dataset from comScore Inc. The author of this paper particularly enjoys how the article first introduces the theory, then jumps to step-by-step processes including input, output, major steps/ algorithm for each process, and lastly displays the outcome graphically. Not only this article adds to the author's knowledge of how association rule is used to discover online shopping patterns, the author can greatly benefit from the way this article is divided into sections logically.

The literature review processes include other articles as well, as per references page; however, the author decides not to analyze them in depth since she does not completely agree with their techniques and opinions, but still aims to cite them to demonstrate a few of her arguments.

2. **Argue strengths and weaknesses of R versus SQL versus Python and under which circumstances, one should stand out**

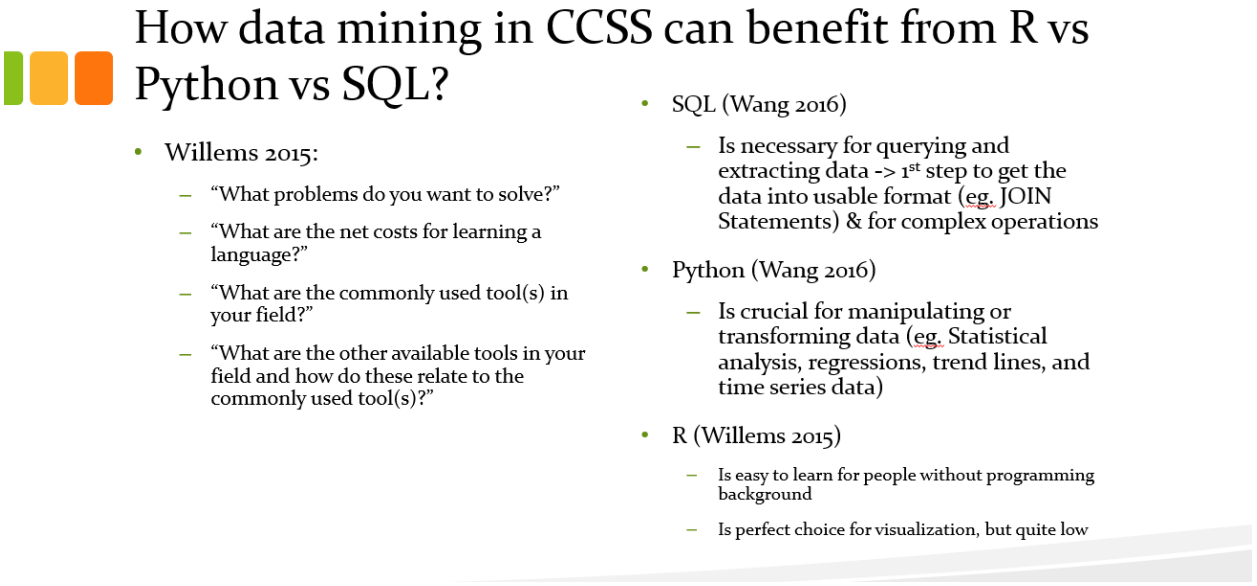

Figure 1. Recommendation as for when to use SQL versus Python versus R in data mining slide

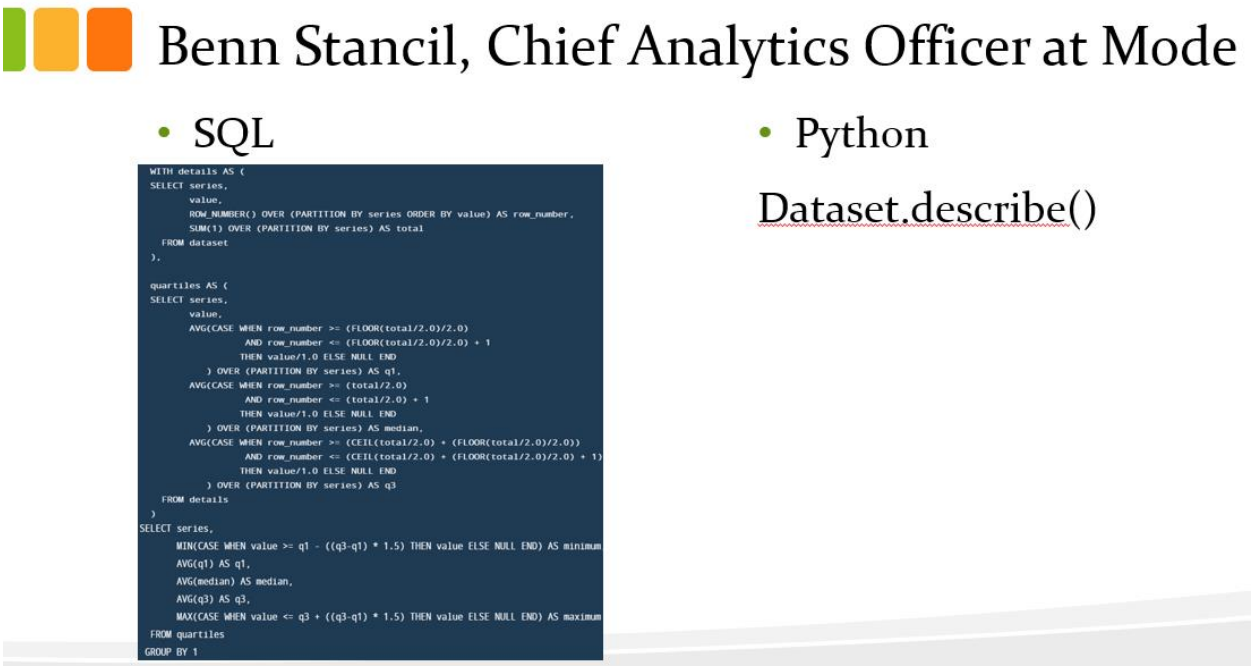

Figure 2. Code snippets to present the quantities of a dataset performed in SQL versus Python slide

C. Solve problems in technology, technique and process relating to database management and design

My whole essay and presentation is to propose a system that can take into account advantages of existing applications of data mining in grocery shopping while attempting to eliminate as many weaknesses as possible.

1. **Essay:** https://landing.athabascau.ca/file/download/3763863

2. **Presentation:** https://landing.athabascau.ca/file/download/3761252

D. Independently and reflectively research issues, technologies, processes and tools in information management

My end products are essay and presentation that are all individual work and results of literature review process. In addition, my weekly reflections also demonstrate challenges I have gone through and lessons I have taken away from those experience.

1. **Essay:** https://landing.athabascau.ca/file/download/3763863

2. **Presentation:** https://landing.athabascau.ca/file/download/3761252

3. **Reflections:**
    a. *Project Week 1*: https://landing.athabascau.ca/pages/view/3713453/week-10-project-week-1

b. *Project Week 2*: https://landing.athabascau.ca/pages/view/3733731/week-11-project-week-2
c. *Project Week 3:* https://landing.athabascau.ca/pages/view/3761643/week-11-project-week-3
d. *Project Week 4:* https://landing.athabascau.ca/pages/view/3763125/week-12-project-week-4

E. Critically evaluate information and data technologies in the context of organisational needs

1. Evaluate when to use Python versus R versus SQL in the presentation and when answering questions during the presentation

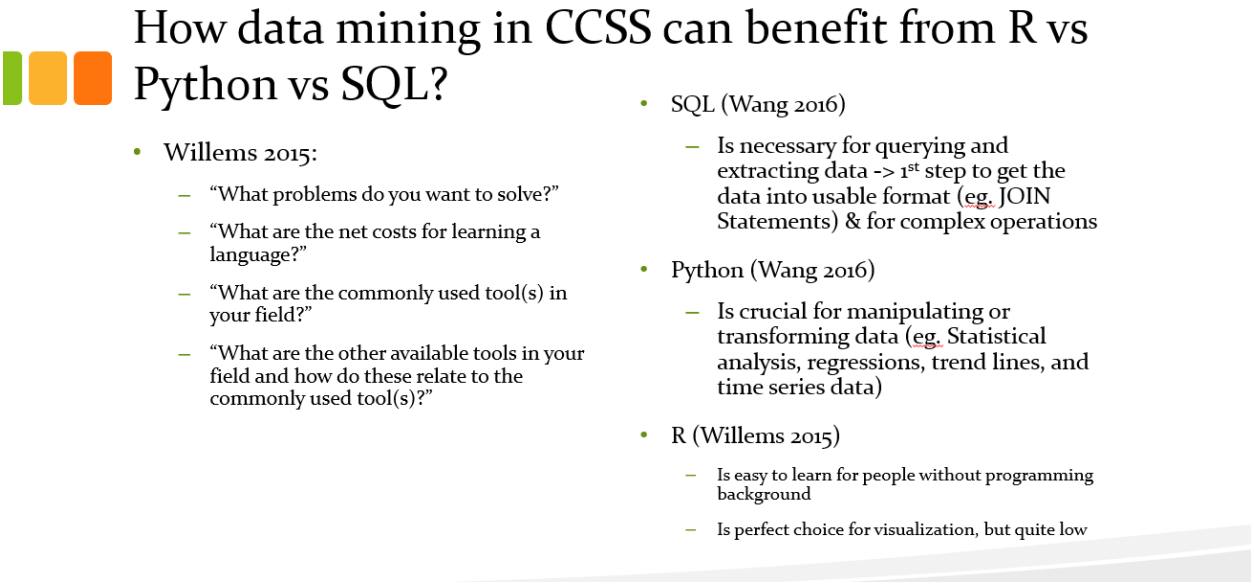

Figure 1. Recommendation as for when to use SQL versus Python versus R in data mining slide

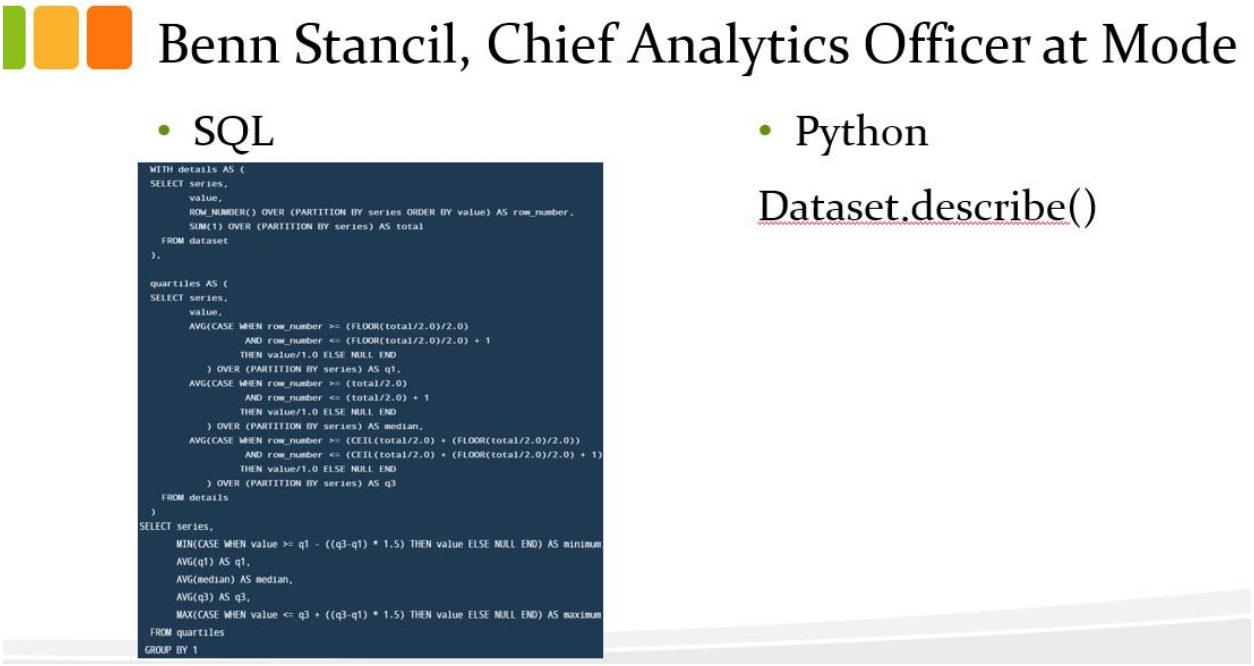

Figure 2. Code snippets to present the quantities of a dataset performed in SQL versus Python slide

2. Questions asked during the presentation

- I asked a classmate about why he used NATURAL JOIN in his SELECT statement. Through my research and experience, NATURAL JOIN is quite restricted in query development because it does not specify which columns should be joined together. Unfortunately, time was out and I couldn't elaborate on my question, but I have been researching about under which scenarios, NATURAL JOIN is encouraged to use.
- I was also curious to know which encryption techniques, such as hash algorithms Cloud Computing from distributors, such as Microsoft and Amazon, provides to ensure data privacy – that was another question asked during another classmate's presentation.

F. Be a reflective practitioner in the information management field

My weekly reflections during research period:

a. Project Week 1: https://landing.athabascau.ca/pages/view/3713453/week-10-project-week-1
b. Project Week 2: https://landing.athabascau.ca/pages/view/3733731/week-11-project-week-2
c. Project Week 3: https://landing.athabascau.ca/pages/view/3761643/week-11-project-week-3
d. Project Week 4: https://landing.athabascau.ca/pages/view/3763125/week-12-project-week-4