CAPP: Customer Analytic and Personalization Platform
8

Phanomphorn	Kanyawongha	Ploy	ID 55070503430	k.phanomphorn@gmail.com
Wipada	Waisaya	Wi	ID 55070503444	wipada.wai@gmail.com

Advisor: Assoc. Prof. Dr. Tiranee Achalakul
Co-advisor: Dr. Warasinee Chaisangmongkon

4 May 2016

_____

I have read this report and approve its content.

Abstract (English)

Retailers' ultimate goal is to increase their profit which can be done in many different ways, such as; increasing customers' loyalty or increasing the sales. Many retailers have their own online website as another channel for commercializing and selling their products. The data gathered from customers' interaction with the website can be kept as a log. So, these logs can be used to analyze for many useful information. For example, customer purchasing pattern can be analyze for recommend another item, which can increase the chance for additional purchase from customers. Without an appropriate knowledge and tools for working with these big amount of data, the retailer will miss the opportunity to utilize those data and gain benefit from them. So, this project focuses on increasing Thai retailers' profit by making use of the data that are available to them, for example; customers' personal information, transaction details, and products data.

Although the knowledge of data science and data analytics are crucial for analyzing the data for further use, not many retailers have been able to fully utilize their data yet. One of the main reason causing this is because data science and analytics are quite new in Thailand. Another key reason is even though there are data analytics solutions available, they are usually in English and in western context. Therefore, those solutions are not very suitable to be applied to the data in Thailand which usually contain Thai and involve Thai cultural context.

This project purpose is to help retailers to utilize their data by using data analytics and customer journey techniques. A Software as a Service (SaaS) analytic platform will be developed for retailers to analyze their data via website. It will analyze the data received from retailers and provide a result of the analysis back via the web interface which will be a benefit to their marketing strategy decision. It has the potential to be commercialized upon completion. Since it can help retailers to increase their profit, we are quite certain that it will become a great product in the near future.

Abstract (Thai)

โครงงานนี้เป็นการพัฒนาเว็บแอปพลิเคชั่นและแพลตฟอร์มที่ใช้การวิเคราะห์ข้อมูลขนาดใหญ่เพื่อช่วยให้ผู้ประกอบการธุรกิจค้าปลีกได้รับผลกำไรสูงสุด โดยการนำข้อมูลต่าง ๆ เช่น ข้อมูลส่วนตัวของลูกค้า ประวัติการซื้อและพฤติกรรมการเลือกซื้อสินค้า เป็นต้น มาทำการวิเคราะห์ด้วยเทคนิคทางวิทยาศาสตร์ในการจัดกลุ่มลูกค้า แบ่งประเภทสินค้า ศึกษาหารูปแบบที่เกิดขึ้นซ้ำ ๆ จากประวัติการซื้อ นำมาประมวลผลข้อมูลบนกลุ่มก้อนเมฆ เพื่อให้ได้ความรู้ในเชิงลึกอันเป็นประโยชน์ เช่น การรับรู้และเข้าใจความต้องการที่แท้จริงของลูกค้า แนวโน้มของสินค้าในตลาด ปัญหาและอุปสรรคที่ทำให้ไม่สามารถไปถึงเป้าหมายทางการค้า เป็นต้น ซึ่งทำให้ง่ายต่อการจัดการ วางแผนกลยุทธ์ทางการตลาด โฆษณา และทำกิจกรรมส่งเสริมการขาย นำไปสู่การลดต้นทุนและเพิ่มรายได้ เพิ่มโอกาสในการขายให้มากยิ่งขึ้น

ในการพัฒนาโครงงานนี้ได้มีการวางแผนจัดการระบบฐานข้อมูลและออกแบบโมเดลต่างๆ เพื่อสามารถนำไปประยุกต์เข้ากับธุรกิจค้าปลีกประเภทอื่น ๆ ในประเทศไทย ให้สามารถตอบโจทย์ได้อย่างตรงจุด ใช้งานได้อย่างเหมาะสม และมีประสิทธิภาพ เพื่อเพิ่มผลกำไร และยังเป็นการเพิ่มอัตราการเติบโตทางเศรษฐกิจโดยรวมของประเทศอีกด้วย

Acknowledgements

This project would not be possible without the guidance of our advisor, co-advisor, committee members, lecturers, friends, and family. We would like to express our gratitude to everyone who has helped and guided us in this project.

Firstly, we would like to sincerely express our gratitude to our advisor and co-advisor, Assoc. Prof. Dr. Tiranee Achalakul and Dr. Warasinee Chaisangmongkon for their continuous support on this project, for their patience, motivation, new ideas, knowledge, and guidance which have led us since the very beginning until the end of the project. Without them, this project would not be able to come this far.

Beside our advisors, we would like to thank our committee, Dr. Priyakorn Pusawiro and Dr. Sally E. Goldin who guided and commented on our project so that it can be improved. We can further improved our presentation by Dr. Priyakorn Pusawiro guidance and Dr. Sally E. Goldin gave us a detailed comment on the documentation so that we can correct our mistakes and organize the content appropriately.

We are also thankful to Mr. Rajchawit Sarochawikasit and Asst. Prof. Dr. Santitham Prom-on who taught us in Big Data Experience class. We got techniques and knowledge that can be adopted in this project from them and they also gave us advice on big data analytics.

Our gratitude also goes to our seniors and friends who have helped us in this project. In particular, we are grateful to Mr. Nopparoot Kitcharoen who introduced us to tools and techniques to be used in this project and Mr. Koetkao Sriratanaban, Mr. Thanik Sitthichoksakulchai, and Mr. Thanasit Yiamwinya who provide us technical support, trouble shooting, and environment setup.

Finally, we would like to thank our families who understand and support us throughout this project and our lives.

Table of Contents

List of Figures

Chapter 1
Introduction

1.1 Problem Statement and Approach

Retailers' ultimate goal is to increase their profit which can be done in many different ways, such as; increasing customers' loyalty or increasing the sales. Many retailers have their own online website as another channel for commercializing and selling their products. The data gathered from customers' interaction with the website can be kept as a log. So, these logs can be used to analyze for many useful information. For example, customer purchasing pattern can be analyze for recommend another item, which can increase the chance for additional purchase from customers. Without an appropriate knowledge and tools for working with these big amount of data, the retailer will miss the opportunity to utilize those data and gain benefit from them. So, this project focuses on increasing Thai retailers' profit by making use of the data that are available to them, for example; customers' personal information, transaction details, and products data.

Although the knowledge of data science and data analytics are crucial for analyzing the data for further use, not many retailers have been able to fully utilize their data yet. One of the main reason causing this is because data science and analytics are quite new in Thailand. Another key reason is even though there are data analytics solutions available, they are usually in English and in western context. Therefore, those solutions are not very suitable to be applied to the data in Thailand which usually contain Thai and involve Thai cultural context.

This project purpose is to help retailers to utilize their data by using data analytics and customer journey techniques. A Software as a Service (SaaS) analytic platform will be developed for retailers to analyze their data via website. It will analyze the data received from retailers and provide a result of the analysis back via the web interface which will be a benefit to their marketing strategy decision. It has the potential to be commercialized upon completion. Since it can help retailers to increase their profit, we are quite certain that it will become a great product in the near future.

1.2 Objectives

The objective of this project is to utilizing potential data of retailers to ultimately increase profit by

- Anticipating demand
- Providing availability at right place and time (inventory)
- Relevant promotions and timely offers offering
- Accelerating customer acquisition
- Improving customer loyalty
- Developing more innovative product

1.3 Scope

This project aims to develop a Software as a Service (SaaS) analytic platform for retailers on the cloud. The services will be available as a web-based platform. Each service will provide an analytic result corresponding with the model. The platform consists of 2 main components, which are front-end web-interface and analytic backend. However, the full scope of the project is too big to be done within one academic year, so some features will be limited. The feature included in this project are as follow:

Web-interface

Front end web-interface will be developed by using PHP, HTML5, and CSS. Retailers can submit the sample data to be analyzed and view the analysis results via dashboard. There are several functions for the web-interface which are as follow:

- Registration and login system

  Retailers are required to create their own username and password to login in order to use the services provided by the analytic platform.

- Data retrieval

  In the real scope of the project, the data will be retrieved via various online channels such as e-commerce platform or social media. However, in this project, the data that used for analyze will be pre-retrieved via offline source instead because the main focus for this project is the development of the analytics backend.

- Analysis Dashboard

  After the data has been analyzed, which take some times depend on the size of the data, retailers will be able to view the analysis results from the analytics engine on the cloud via graphical-interface dashboard. The categories of the analysis are:

  - Customer segmentation

    The result of the analysis will show the segments of each customer depends on the similarity between them or various factors. Similar customers or customers who belong to the same criteria will be grouped together as one segment.

  - Product classification

    The analysis will classify each product into several categories. Retailers can view the number of products and products detail in each category.

- Sales and revenue

  The report for sales and revenue will mainly focused on income from each product category. It will also include the ranking of popular products.

- Social media

  The analysis will take the data of content posted, number of views and likes to compute the potential to reach the customers of each social media. The report will also show popular posts in various period of time to see the content and feedback of those posts.

- Generated recommender

  Recommender will generated top 5 items for each customer separately and send the recommendation via emails to encourage their purchasing. The recommended items will come from the recommendation analytics backend on the cloud.

Analytic backend

The analytics model will be hosted on the cloud for high computation power and storage. The data will be retrieved from several retailers channels and analyze using each analysis category model. The analysis model will be constructed beforehand and tune-up periodically to enhance the analytics performance. After the data has been analyzed, it will send the result back to the web engine to visualize the result to retailers.

For this project, there will be some limitation and restriction on the platform due to the limited development time and the full scope of this project. The fully-developed system will take more than one academic year to implement. Also, all of the knowledge and technology required for the project is quite complex and required considerable time to learn. So, at the end of this academic year, a proof of concept prototype will be submitted instead of the fully-developed platform.

1.4 Tasks and Schedule

Semester 1/2015: August – December 2015

Task name	August			September				October				November				December				
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Pre-Development Phase																				
Meet with advisor																				
Background research																				
- Determine scope of problem																				
- Research on cloud computing																				
- Research on machine learning																				
Documentation																				
- Draft proposal																				
- Proposal report due																				
- Create proposal presentation																				
Proposal presentation																				
Data science and configuration process																				
- Gather and preprocess data																				
- Explore data for insights derivation																				
- Setup cloud environment for development																				
Documentation																				
- Draft report 1																				
- Report draft 1 due																				
- Create final presentation																				
Final presentation																				

Semester 2/2015: January – May 2015

Task name	January				February				March				April				May			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Development Phase																				
Model research																				
- Determine scope of each model																				
- Research on model construction																				
Documentation																				
- Draft report 2																				
- Report draft 2 due																				
Model construction																				
- Construct a model to solve pre-defined problem																				
- Implement in cloud environment																				
Documentation																				
- Draft report 3																				
- Report draft 3 due																				
- Draft final report																				
- Final report due																				
- Create final presentation 3																				
- Final presentation due																				
Demo and open house																				

Chapter 2
Background, Theory and Related Research

2.1 Machine learning

2.1.1 Overview concept

Machine learning is a data analytic method that allows computer to learn and analyze the data without being explicitly programmed [1]. The analytic process relies on mathematical or logical formulas, which are called a model, to identify relationships for the variables in the data. These relationships can grow and change when applied to new dataset.

2.1.2 Training data and testing data

In order to apply machine learning concept, that machine needed to be taught and trained before it can automates the analytic process by using analytic model. So, there are 2 types of the data that are used for constructing the analytic model.

- Training set
  The training set is the set of data that used to train the analytic model in order to find or learn an analytic pattern or relationships among variables of the data.
- Testing set
  The testing set is the data that used to estimate error rate and evaluate performance of the final model constructed by training set. Testing set and training set should not be the same set of data because it will cause a situation where the model become too specific with the training data and become unsuitable to use with an unseen data.

2.1.3 Learning method for machine learning

The machine can be trained in many different ways in order to be able to figure out the analytic model or hidden pattern in the data. There are 2 main types of learning method for machine learning.

- Supervised learning
  The supervised learning is the machine was trained by the set of example data. The training data is labelled with outputs to indicate which inputs should produce which outputs.
- Unsupervised learning
  In the unsupervised learning method, the training data do not have outputs labelled with inputs. The goal for this method can be to find hidden patterns in the data.

2.2 Data analytic model

2.2.1 Overview concept

A Data analytic model is a mathematical or logical formula that describes relationships among variables in a dataset [2]. The mathematical formula of the model can be different even for the same problem with the same dataset. For the best solution, the model need to minimize the prediction error when applied to a new or unseen dataset.

2.2.2 Types of analytic model

Data analytic model can be different depends on the expected output for the problem. The expected output also depends on the question that needs to be answered. There are mainly 2 types of analytic model that is used for analyze the data in order to answer those questions.

- Classification

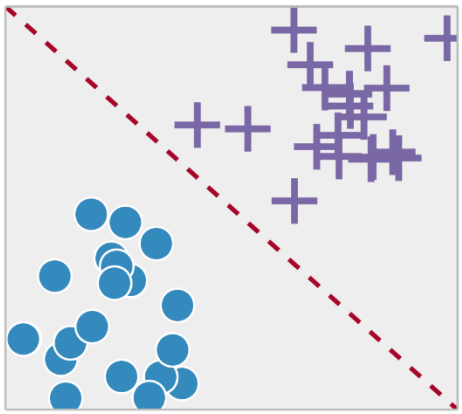


Figure 1: Illustration of classification model.
Taken from http://ipython-books.github.io/featured-04/

A Classification model is a model that categorizes each data record into class variables with different characteristic. The above figure illustrated how the data can be categorized into different classes. The circle dot and plus sign are the data point in the dataset. The dash line is decision line between circle dot data point and plus sign data point, which is created by the model. Data that are similar with each other can be define in the same class. The class that model needed to predict is called ‘target variable’. Each attribute in data record that is used for categorized is called a ‘parameter’. Not all of parameters are necessary for predicting target variable. The model can be optimized by adjusting parameters that are used in prediction process.

case ID | predictors | target

CUST_ID	CUST_GENDER	EDUCATION	OCCUPATION	AGE	AFFINITY_CARD
101501	F	Masters	Prof.	41	0
101502	M	Bach.	Sales	27	0
101503	F	HS-grad	Cleric.	20	0
101504	M	Bach.	Exec.	45	1
101505	M	Masters	Sales	34	1
101506	M	HS-grad	Other	38	0
101507	M	< Bach.	Sales	28	0
101508	M	HS-grad	Sales	19	0
101509	M	Bach.	Other	52	0
101510	M	Bach.	Sales	27	1

Figure 2: An example demographic data of customer with target variable for classification problem. Taken from http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm

The above figure shows an example demographic data table. A target value of 1 indicated that this customers increased spending with an affinity card, while a value of 0 indicated that this customers did not increase spending with an affinity card. So, one possible problem that can be solved with this data is to indicate the characteristics of the customer who will increase spending with an affinity card. If the customer is in the class that will spend more with affinity card, the company can offer them a card to gain more profit for the company.

- Regression

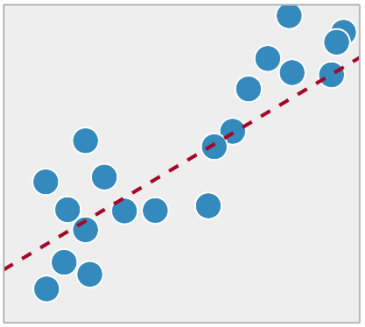


Figure 3: Illustration of regression model. Taken from http://ipython-books.github.io/featured-04/

Regression model is a model that estimates data values in numerical format. The above figure illustrated how the model estimates data value. The dash line is the estimation line, which is represents the model. It goes along the trends of the data to estimates the next data value. The construction of regression model also taken from the historical data value. The target variable is a dependent variable that is influenced by one or more independent variable in the record. The relationship of target variable and other independent variable can be derived in mathematical formula for predicting unseen data value.

Advertising	Sales
$120	$1,503
$160	$1,755
$205	$2,971
$210	$1,682
$225	$3,497
$230	$1,998
$290	$4,528
$315	$2,937
$375	$3,622
$390	$4,402

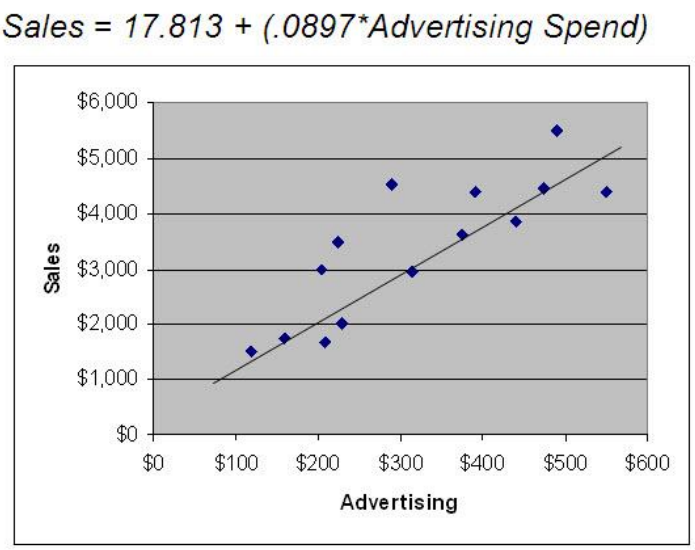


Figure 4: An example data about advertising cost and sales. Taken from http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/regress.htm

From the above example data, there is a mathematical formula which illustrates the relation between advertising cost and sales. Advertising cost is an independent variable and can be used to estimate sales value. Model's formula can be drawn as a line passing through the data to estimate sales amount depends on advertising cost. So, the trends of this example data is, if advertising cost is increase, sales value is also increase. The model also indicates how much each additional dollar in advertising will increase sales.

The regression equation fitting the data might not be linear. If the data has an increasing and decreasing in different period, the regression equation might become parabola, or any other shape.

Both classification and regression are types of model that can be used to analyze and predicted data value. However, there are many other types of analytics model which are not included in this report because there are too many and some of them are very advanced and complex. Classification and regression models were mentioned because they are the fundamental models and related with the analytics engine of the project.

2.2.3 Model validation

After the model has been constructed from example data, it needs to be validated to determine whether it is a good fit to the data or not. The quality of the model depends on many factors such as the data that were used to build the model, or even the relation of model's parameters. A model with low quality or badly constructed can affect further prediction of the unseen data. Therefore, in order to be able to predict an unseen data value properly, the model need to be well-adjusted and validated before deploying in the real system environment.

- Error estimation and accuracy

One of the measurement for model quality is the error rate of the model. Error estimation is the approach to investigate the correctness of the model prediction. Each type of model can be evaluated differently. The accuracy of the classification model can be measured by

$$accuracy = \frac{Number\ of\ correct\ decision\ made}{Total\ number\ of\ decision\ made}$$

$$accuracy = 1 - error\ rate$$

Figure 5: accuracy measurement formula

This error estimation approach assumes that all errors have the same importance, but that is rarely true in the real world. Different errors can have an impact on different cost and benefit. So, the error should be investigated separately. If one model has lower accuracy but the error prediction has a small impact on the benefit, it might be better than a model with high accuracy but the error prediction has a large impact on the benefit. These all depend on the problem that needs to be solved.

- Model fitting

As from the model error rate evaluation, the quality of the model can be measure. However, the model with very low error rate does not necessarily mean that the model is a good fit to the data. There are 2 type of model that is not good enough for applying to the real data.

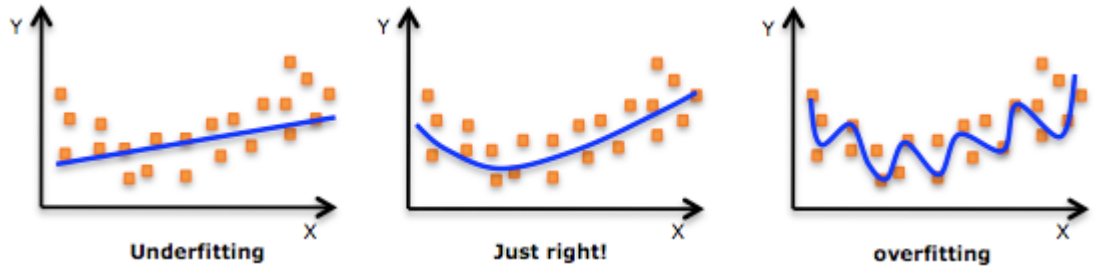


Figure 6: The illustration of underfitting, properly fit, and overfitting model. Taken from http://pingax.com/regularization-implementation-r/

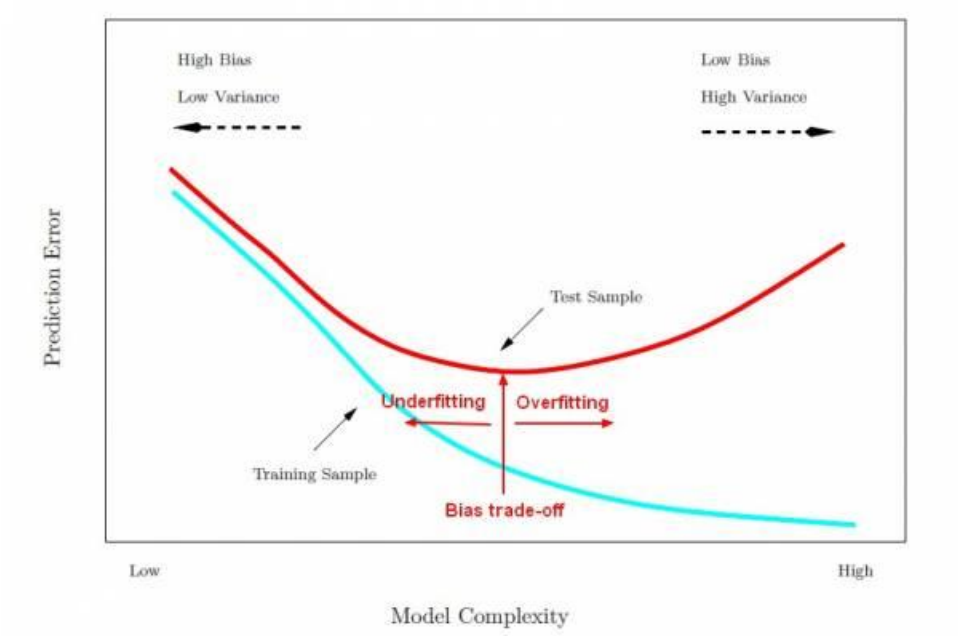


Figure 7: The error graph of training error and testing error : http://gerardnico.com/wiki/data_mining/overfitting

- Overfitting

  The overfitting situation happens when the model was trained with training set for too long or with too many training attempts. The model can perfectly fit with all the data in training set and become too specific with the training set, so the error rate from the training set is very low. There might be some misunderstanding error rate is low and accuracy is high, so the model is good and can be used for real problem. However over-fitted model means the model can only work well with the specific data and it might have lower performance when applying to the unseen dataset.

  The error rate graph in the figure above also shows the problem with the overfitting model. The lower curve represents an error rate from training set. The upper curve represents error from the testing or unseen dataset. The goal of the model construction is to predict an unseen dataset with a minimum error, which is a bias trade-off point.

  In the situation where the model over-fitted, the training error becomes lower as the model become more complex, that is, too specified with the correct answer in the training set. However, when apply to an unseen dataset, an error become higher than a bias trade-off point because the model is constructed to perfectly fit with the training data, but not the unseen dataset.

- Underfitting

  The under-fitting situation is opposite with overfitting situation. An underfitting model means that the model has low performance on training set. Thus, the model has failed to capture the relationship between variables in dataset. The error rate for unseen data will still high and model will show low performance on unseen data. This can be occurred from model with bad structure or too few training data items.

  Figure 7 shows that underfitting model has high error rate on both training set and testing set. The model poorly performs prediction because it still cannot discover the relation of the variable in the dataset. The error rate can be reduced until certain point before the model will become overfitting.

- Cross validation

  Cross validation is one of the model validation methods that can be used to prevent overfitting model. The idea of the method is to use different portion of data as a training set. There are many methods to perform cross validation.

  - Holdout method

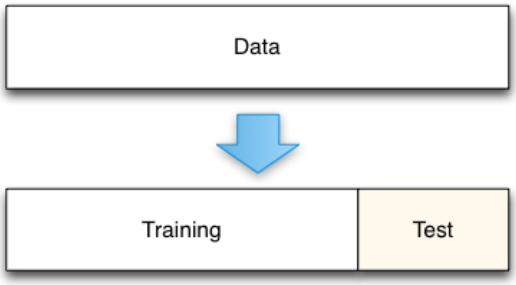


Figure 8: Holdout validation data split.
Taken from, http://scott.fortmann-roe.com/docs/MeasuringError.html

This holdout method is the simplest way to perform cross validation. It also called 2-fold cross validation. From the figure above, it illustrated that the data is separated into 2 sets as a training set and testing set. The testing set is 'holding-out' from the data at the beginning as 'new' data to estimate the prediction error.

  - K-fold cross validation

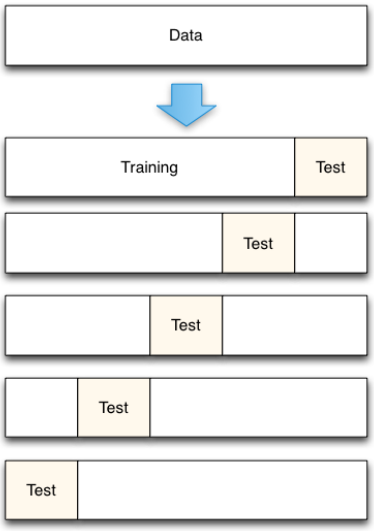


Figure 9: 5-fold cross validation data split.
Taken from http://scott.fortmann-roe.com/docs/MeasuringError.html

K in K-fold cross validation is stand for number of times that the data is divided. For example, the above example is a 5-fold cross validation as the data is divided into 5 portions. The model is evaluated 5 times, each time with each different set of training and testing data. However, the downside of this is method is the evaluation needs to be applied for more than once.

- Leave-one-out cross validation

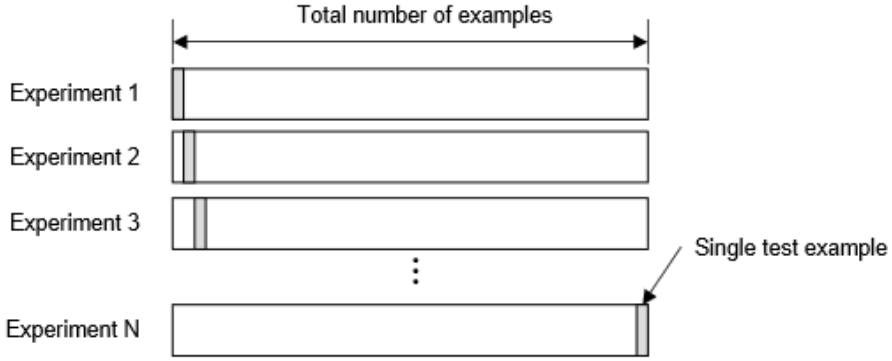


Figure 10: Leave-one-out cross validation data split

The test data for leave-one-out cross validation is a single data record, instead of a set of data, that is left out from the model construction. As from above figure, single test example is used as a testing data, so the experiment needs to be conducted N times where N is the number of examples in the dataset.

- Mean Absolute Error (MAE)

MAE is a statistical measure for measuring how close a prediction is to the actual data [43]. It is given by the equation

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|\,.$$

Figure 11: MAE equation

Which means, it is the average of the absolute error between the pair of forecast (f_i) and the true value (y_i). MAE is a common standard measure for data that occur over a sequence of time or order (this is called time series in statistic term).

2.3 Model induction algorithms

There are many algorithm that can be used to construct the model for prediction and classification. Each algorithm is suitable for different problems depending on the question that needs to be answered and the type of data available. Also, even for the same question, the algorithm for model induction can be different, and provide different quality of answers. The data scientist needs to evaluate the answer quality and choose the algorithm that provides the best answers for the problem.

2.3.1 Decision tree

Decision tree is one of the algorithms for inducing classification model. It can be applied to the data via classifier to categorized data into each class or category. The decision tree can be constructed by the labeled data as a supervised learning. An internal node in the decision tree is called 'decision node', where it can divides the data into sub-categories. As the data goes along the path from one classify node to the next, it will eventually reach the 'leaf node' where it represent the class that the data belongs to.

RID	*age*	*income*	*student*	*credit_rating*	Class: *buys_computer*
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Figure 12: Example table of customer demographic with the class whether customer buy the computer or not. Taken from http://scriptslines.com/blog/example-generate-decision-tree-by-id3/

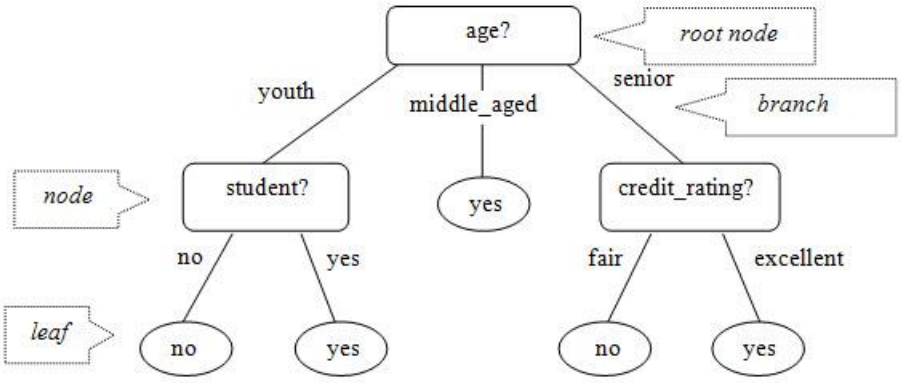


Figure 13: Decision tree model for data in figure 11.
Taken from http://scriptslines.com/blog/example-generate-decision-tree-by-id3/

Figure 12 shows the customer information with class of buying a computer or not. This information can be used to construct the decision tree model in figure 12. This model can be applied to the new data to classify it and predict whether a new customer will buy a computer or not. At the first node, the data will be classified by age. If the age of customer is a youth, it will be applied to the next node. Then, it will be classify whether customer is a student or not. If customer is a student, this customer will be classify into class of customer who buy a computer. The order of decision node is order by the attribute that gain the most information about the data, which means it can deduct the most uncertainty of data.

The above decision tree model is only an example for the example data, different data will result in different decision tree model. Also, not every branch will use all the attribute, each branch can use some of attributes as a child node to separate data into different class.

2.3.2 K-means Clustering

Clustering is one of the classification models that is constructed from unsupervised training. It can be applied to find a pattern or structure with in the unlabeled data and classify them into the same class, which is called cluster. K in K-means clustering stands for the number of clusters as user-defined constant.

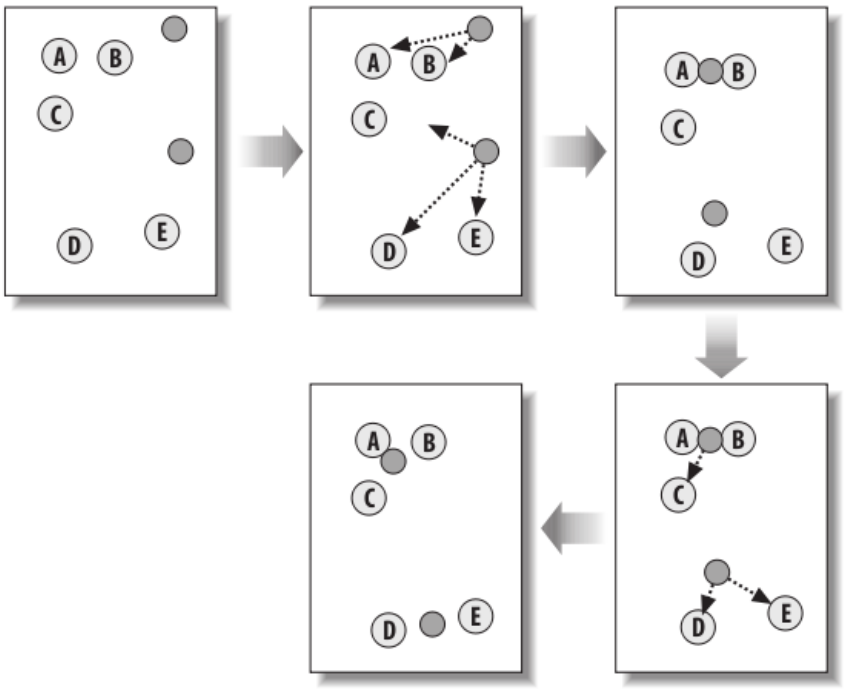


Figure 14: Example of K-means clustering. Taken from http://mines.humanoriented.com/classes/2010/fall/csci568/portfolio_exports/sphilip/kmeans.html

Figure 13 illustrates how K-means algorithm is applied to classify data into 2 clusters. The goal of this algorithm is to minimize the distance between each data point and its cluster to classify them into 2 different groups.

The mechanism for this is to pre-define 2 center points in the data, which is called centroids, and assign them randomly in dataset. Each data point will be assign to their nearest cluster. Then, those centroids are moved to the average location of all the data points that are assigned to it. The process is repeated multiple times until the solution converges to clusters that represent the true structure of the data.

2.3.3 Collaborative filtering

Collaborative filtering is one of the popular recommendation algorithms for the model-based recommender system. The general idea of collaborative filtering is the process of filtering large amount of data for information or pattern about customer interests or preference by using similarity value.

Similarity measures

Similarity measures are a fundamental concept for collaborative filtering. There are several similarity measures for determining the similarity between two entities, but this project mainly use the following measurements:

- Euclidean distance

$$d(x,y)=\sqrt{\sum_{k=1}^{n}(x_k-y_k)^2}$$

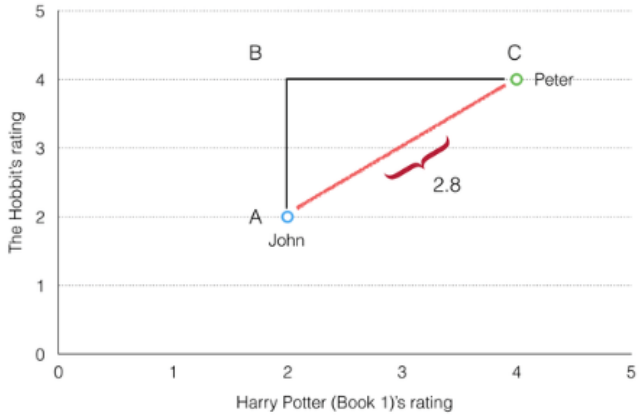


Figure 15: Ilustration of euclidean distance

Euclidian distance is a distant between vectors in linear perspective. It can represents similarity between each data with their distance. The less distance between two items, the more similarity they are.

Figure 14 shows the example of Euclidean distance between John and Peter, and also shows the formula for calculating Euclidean distance, where X and Y are the data vectors, and k is the dimension of the vector. The distance between John and Peter is calculated from

$$\sqrt{(2-4)^2+(2-4)^2}=2.8$$

- Cosine distance

$$cos(\theta) = \frac{x \cdot y}{||x|| * ||y||}$$

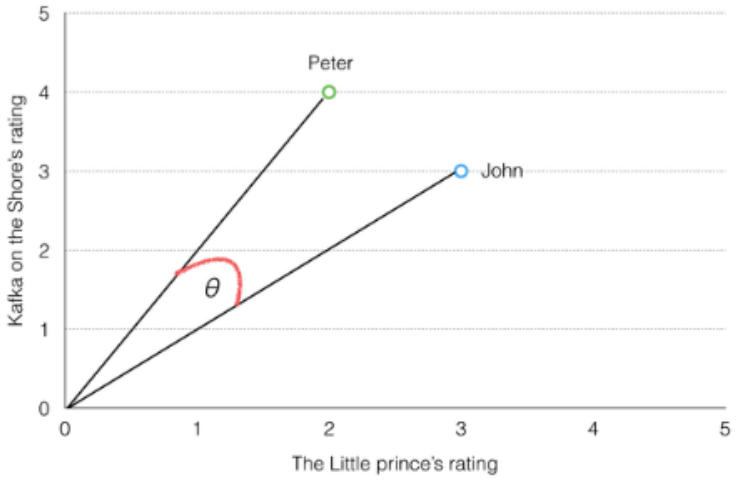


Figure 16: Ilustration of cosine distance

Cosine distance is the similarity measures that based on the angle between data vectors. If the angle between data vector is high, cosine value will become less, which indicate lower the similarity between them.

Figure 15 shows the example of cosine distance along with cosine distance formula, where X and Y is the data vectors. As for example, the cosine distance between John and Peter is

$$\frac{2 * 3 + 4 * 3}{\sqrt{2^2 + 4^2}\sqrt{3^2 + 3^2}} = 0.95$$

- Pearson correlation score

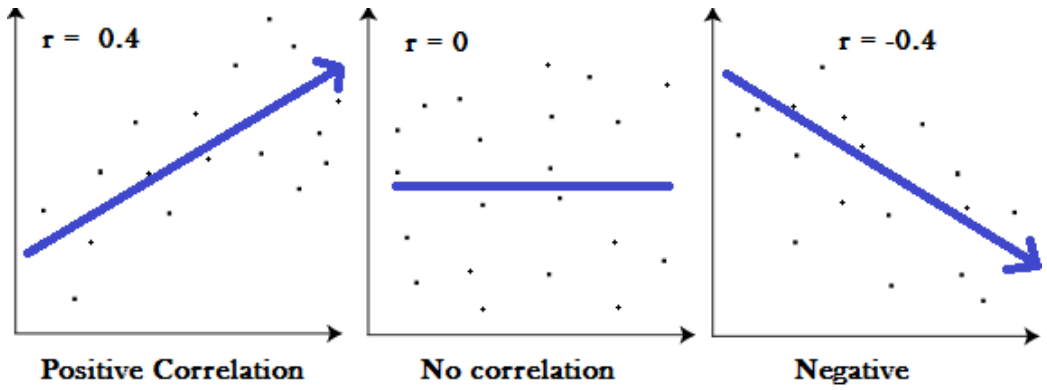


Figure 17: Ilustration of pearson correlation score

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

Figure 18: Pearson correlation score formula. Taken from http://www.statisticshowto.com/what-is-the-pearson-correlation-coefficient/

Pearson correlation score is the score that measures of how well the data are related as a linear relationship. The possible result for Pearson correlation will be between -1 and 1. If the score is positive, it means that those data are related in positive trend and vice versa. If the score is zero, it means that those data have no relation between each other. The formula of Pearson correlation score is showed in Figure 17, where N is a sample size, X and Y are data value of each set of data to be compare.

Types of collaborative filtering

There are many type of collaborative filtering, but this project mainly involved with 2 types of collaborative filtering as follow:

- Item-based collaborative filtering

  Item-based collaborative filtering is a form of collaborative filtering which focused on the similarity between items based on customer preference or purchase. So, even a new customer who only visit the sites for the first time, the recommendation with item-based collaborative filtering can still generated relevant recommended items.

Customer	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
A	X	X	X			
B	X	X				
C			X		X	
D			X	X	X	
E		X	X			
F	X	X		X	X	
G	X		X			
H	X					
I						X

Figure 19: An example customer purchasing matrix, taken from http://www.dummies.com/how-to/content/how-to-use-itembased-collaborative-filters-in-pred.html

Figure 18 shows an example customer purchasing for each items. The "X" sign means that a customer has once purchased this item. The each pair of items needed to be calculated their similarity with each other by using one of the similarity measurement method. In this example, the cosine distance will be used to determine similarity between each item. For example, the similarity between item 1 and item 2 is

$$\frac{1*1+1*1+0*0+0*0+0*1+1*1+1*0+1*0+0*0}{\sqrt{1^2+1^2+0^2+0^2+0^2+1^2+1^2+1^2+0^2}\,\sqrt{1^2+1^2+0^2+0^2+1^2+1^2+0^2+0^2+0^2}} = 0.67$$

Item 6	0	0	0	0	0	
Item 5	0.26	0.29	0.52	0.82		0
Item 4	0.32	0.35	0.32		0.82	0
Item 3	0.40	0.45		0.32	0.52	0
Item 2	0.67		0.45	0.35	0.29	0
Item 1		0.67	0.40	0.32	0.26	0
	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6

Figure 20: Similarity table between each pair of item. Taken from http://www.dummies.com/how-to/content/how-to-use-itembased-collaborative-filters-in-pred.html

After similarity table has been computed, the recommender will select the certain number of the most similar item with the item that customer currently view or purchase. For example, recommender will select item 3 to customer A because customer bought item 1 and item 2, which both have highest similarity value with item 3.

A noticeable example in real world is the recommendation in most e-commerce websites which states that "Customer who bought/view this item also buy…" The algorithm automatically select the certain number of the items that have the highest similarity value with the item that customer is viewing.

- User-based collaborative filtering

Unlike an item-based collaborative filtering, user-based collaborative filtering focused on the similarity between users instead of items. The general idea is to select recommended item from other user who are similar with current user. This algorithm requires several user profiles to calculate the similarity and to be used as references.

2.4 Development platform and environment

2.4.1 Hadoop ecosystem

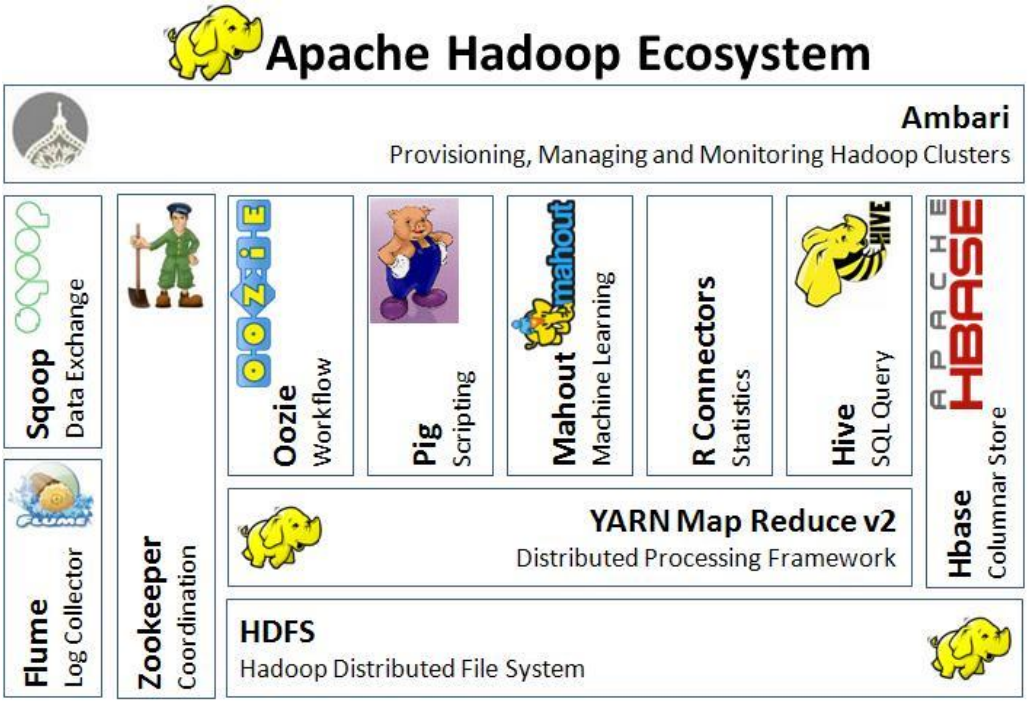


Figure 21: Hadoop ecosystem. Taken from http://thebigdatablog.weebly.com/blog/the-hadoop-ecosystem-overview

Apache Hadoop is an open-source framework for distributed storage and distributed processing [21]. It is the framework for commodity computer cluster to be able to work with very large data. There are many additional modules or libraries that work on top of Hadoop platform. Each module is suitable for different tasks. However, there are 4 core modules included in the foundation of the framework

- Hadoop Common
  The libraries and utilities essential for Hadoop services.

- Hadoop Distributed File System (HDFS)
  The primary distributed file system used by Hadoop applications. The data stored in HDFS will be distributed across multiple machines in cluster to improve I/O performance and redundancy.

- Yet Another Resource Management (YARN)
  Resource management for the processes running on Hadoop.

- MapReduce
  A distributed data processing model and execution environment

In addition, there are many software components that can run on top of Hadoop. Those software components relies on Hadoop to be able to run efficiently. Hadoop also relies on those software component to enhance its capability for big data work, which is why they were call “Hadoop ecosystem”. Software components that are related to this project are as follow:

- Hbase

  Hbase is a NoSQL database built on top of HDFS in Hadoop ecosystem. NoSQL is different from traditional relational SQL database, which is in tabular form. It is suitable for managing and analyzing massive amounts of unstructured data.

  There are different types of NoSQL database and HBase is a column-oriented database. It stored data by a column rather by a row. There are some situation where column-oriented database is more efficient that row-oriented database. For example, if we need all customers name in the customer table, in column-oriented database, it can easily access customer name field to get the data without accessing unrelated fields. However, in row-oriented database, it need to access every record in the database and all fields in the record.

- Mahout

  Mahout is a library for machine learning algorithm on top of the Apache Hadoop. It provides many type of machine learning algorithm such as collaborative filtering, clustering and classification.

2.4.2 Cloudera

Figure 22: Logo of Cloudera

Cloudera is a platform that provides Apache Hadoop distribution for enterprise developed by Cloudera Inc. [30]. It provide several services and software for an easier use of Hadoop ecosystem. The services related with the project are:

Figure 23: Logo of Cloudera Impala

Cloudera Impala

Cloudera impala is a SQL query engine runs on Apache Hadoop. It provides a high-performance and low-latency queries in order to efficiently explore a massive amounts of data.

Figure 24: Logo of Hue

Hue

Hue is an open-source web-interface for visualizing data and managing configuration of Apache Hadoop and its ecosystem. It provide easy-to-use interface for navigating and configuring Apache Hadoop system.

2.4.3 Microsoft Azure

Figure 25: Logo of Microsoft Azure

Microsoft Azure is a cloud computing platform developed by Microsoft [22]. Microsoft and Microsoft partner have their own datacenter to host them as a cloud. It also provide many services along with the cloud service such as SQL database, machine learning services is also one of the services available.

Microsoft Azure is a closed-source cloud platform and need to be purchase in order to use their services. The pricing is pay-as-you-go model, which customers can only pay when they needs to use cloud storage or services from Microsoft Azure.

2.4.4 SendGrid

Figure 26: Logo of SendGrid

SendGrid is an email delivery and management service founded by Isaac Saldana, Jose Lopez, and Tim Jenkins in 2009. It also provides SMTP relay service for web-app to be able to deliver emails with credentials. SendGrid email service is an important integration to the system because it provide credentials for emails in order to prevent our emails to be recognized as spam mails. This can help improve deliverability and ensure that recipient will definitely get an email.

Another reason for choosing SendGrid service is that SendGrid is an add-on service available on Microsoft Azure, which is the cloud-based development platform of this project. So, the integration is easier and it is compatible with this project.

2.5 Big data analytic tools

2.5.1 Tableau

Figure 27: Example visualiztation from Tableau. Taken from http://www.lavastorm.com/tableau/

Tableau is one of the big data visualization tools developed by Tableau Software company[22]. It can be applied to an enormous amount of data from the organization. Its main purpose is to visualize big amount of data into graphical form to illustrated the properties or components of the data, as manual exploration of the data is nearly impossible for human to analyze. It is a useful tool for data exploration to gain an insight of the data and determine the question that the data can answer.

There are many visualzation methods for many types of data, for example, bar graph, pie chart, scatter plot, etc. Visualization from Tableau can illustrated their organization data to a bussiness owner or staff in the organization for an insight data. Tableau supports many types of data format, such as Excel, Access, Firebird 2.0, IBM DB2, MS SQL Server, and Microsoft Power pivot. Also, Tableau has a graphical interface and visualization for easy-access with VizQL technology, which simplified SQL command into interactive actions. For example, user can drag and drop the field name into the graph axis and the data will be quered automatically without having the user input SQL commands directly.

For this project, Tableu was used only to gain an insight of an existing data as the first step to analyze a potential question or information that would be useful for retailers and develop analytics engines. It is proprietary software and require a use to purchase license in order to fully utilize the software. However, the university licenses were given to us by our advisor for one academic year to analyze and explore the data for this project.

2.5.2 RapidMiner

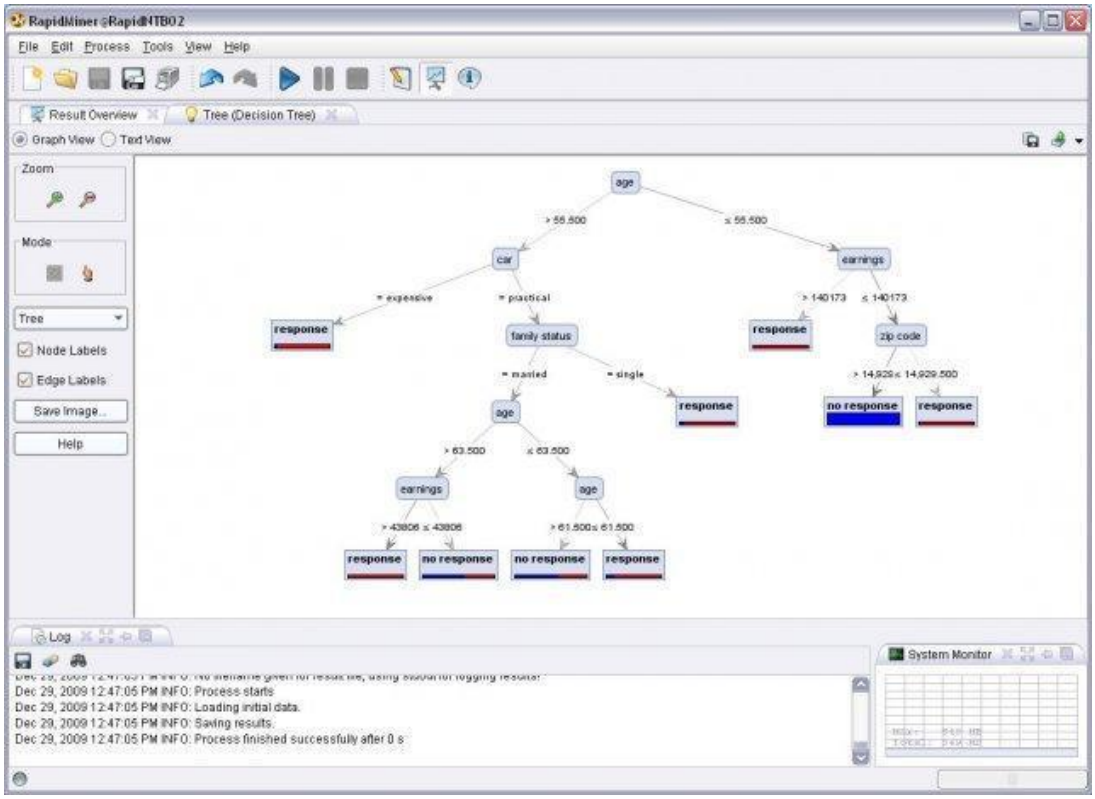


Figure 28: RapidMiner program screenshot. Taken from http://www.predictiveanalyticstoday.com/top-free-software-for-text-analysis-text-mining-text-analytics/

RapidMiner is an integrated environment for machine learning, data mining, text mining and business analytics developed by RapidMiner company [27]. It is an open source program that can be used for data mining analytics with graphical support for easier use. There are many well-known algorithms for machine-learning and data mining integrated in the program. The result of data analysis from RapidMiner can be illustrated in graphical for better understanding.

2.6 Data science process

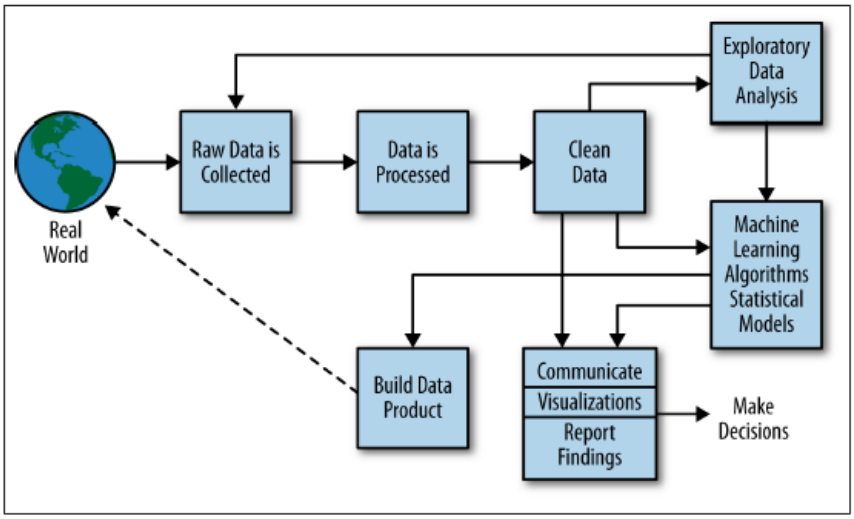


Figure 29: Data science process illustration

The data analytic main idea is to perform analytic method on data and get the result. However, in the real world, the process is not as simple as the ideal step. There are many steps for data science process in order to get the practical results for the real world problems.

First, the data in the real world are usually not clean enough, meaning that the data is not ready for analysis yet. Some of the records in the dataset might be incomplete or some of the attributes in the dataset might be irrelevant and need to be filtered out before analysis. So, the step of data processing is essential in order to get a cleaned data that can be used for analytic.

After the data is cleaned and ready for analytics, the step of exploratory data analysis can be applied to gain the insight of the collected data. Exploratory data analysis means to look and explore the data to gain an insight into it. However, because the data is very large, viewing each data explicitly would take too much time and effort. So, using visualization tool to visualize data composition or data value is the better way to explore the data. The goal of the data exploration is to know the data and to determine what questions these data can answer.

If the collected data do not have enough information, it might be better to collect the data again instead of continuing the process. Data that have too little relevant information can also affect the quality of the analytic process. Thus, the result might not be good enough to use to solve the problem.

Next, machine learning and model induction process can be applied to the cleaned data. There are many types of machine learning algorithm and model induction depending on the problem and desired answer. For example, if it is a problem that the data need to be classified into several classes, it can be used to create a classification model.

Then, the result can be evaluated for its performance. It can be deployed into a real world environment to see the actual outcome of the solution. If the model cannot perform well enough for the real world problems, data science process can be repeated again to find another solution. Or, the model from the analytic process can be evaluated as a report or visualization in order to help business owner or stakeholder decide on the best solution to the problem.

2.7 Customer journey

Figure 30: Customer journey illustration. Taken from http://www.exacttarget.com/blog/uk/7-technology-trends-transforming-consumer-communication/

The "customer journey" is a term used to describe the common purchasing behavior flow of a customer from the start until the end of the purchasing. The study of customer behavior can be used to define the point of interest in order to influence them to increase purchasing chance of the customer. There are many steps in the customer journey

First, awareness stage is the stage that customers have a need of a particular item. This is the step that initiate the possibility of purchasing from customer. Also, retailer or business can influence in the stage of customer awareness to increase the chance of purchasing from the customer. For example, retailer can advertise the product across media to reach the customer. If the customers see the advertisement, they may become interested in the product and may increase the chance of purchasing.

Next is the stage of evaluation. After customers have a need to buy something, they may try to find information related to the product in order to consider whether they will buy the product or not. Detailed information of the product may increase more chance of purchasing.

Next is the stage of purchasing. This is the step where customers interact directly with the retailer or provider to purchase the item. The quality of retailer service can affect customer experience and improve customer satisfaction.

Then, it comes to usage stage, where customers consume or utilize the item that they purchased. If they are satisfied with the product, it may lead to the next stage, which is repurchase. Also, it may lead to the advocacy step where current customer suggests the product to another customer and create another customer journey cycle.

2.8 Competitive Analysis

Since CAPP aims to provide an affordable analytics solution for SME, so the main focus of a comparison will be on pricing. Also, available features of each platform will be compared to investigate each platform capability and differentiate CAPP from other cloud-based marketing analytics platform.

2.8.1 Richrelevance

Figure 31: Richrelevant Logo

RichRelevance is one of the omnichannel personalization solution platform developed by RichRelevance company in USA[28]. It provides personalized and innovative customer experience across web, mobile, and in store. It also offers a platform called the Relevance Cloud™ which provides personalization solutions as a cloud-based platform.

- Features
  - Search optimization

    RichRelevance arranges the search result best on each customer past behaviors and preferences. It sort the result which has the most relevant to customer interest as the first result to attract their attention and create more conversion rate.

    It also provides product suggestions within the search bar to gain an impact on customers and improve customer satisfaction.

  - Content personalization

    The content of each pages that customer visits will be different based on their interests and behaviors. If the customer usually views some certain product, the content of the page will recommended that product to engage customers across browsing page

  - Product recommendation

    Identifies and evaluates recommendation performance to ultimately recommend an effective product to customers. It also gathers customer data from multiple touchpoint such as mobile, web-browser, any cloud or device.

- API services

  RichRelevance provides a ready-to-build API-based personalization tools to manage customer data and integrate personalization into any form of application or services.

- Consultant services

  Retailer can contact RichRelevance to consult on any issues regarding personalization matters. It also provides an A/B testing to validate their solutions from RichRelevance.

- Pricing

  RichRelevance has a performance-based business model. The service price will depends on the revenue and profit gain of the client after the services has been integrated.

2.8.3 Custora

Figure 32: Logo of Custora

Custora is a predictive marketing platform which focuses on customer lifetime value. It provides analysis about customer behavior and their value to the business, which can help increase their income by turning members into customers, or turning one-time buyer into repeated customers.

It also provides several additional features other than predictive analytics, such as customer segmentation to categorized customer into different group, or automated marketing for each customer group.

- Features
  - Predictive Customer Lifetime Value Analysis

    Customer Lifetime Value (CLV) is means the amount of revenue generated from certain customers over their lifetime. It can be useful for optimizing advertising campaign from CLV analysis.

  - Persona Analysis

    Each customer has a unique persona, some of them might like fancy cloth while the others like simple cloth. By using their historical transaction data, Custora can identify existing customers' persona and predict new customers' persona in order to presents them with a relevant recommended items and contents.

- Churn Detection

  Detect and identify if the customer is fading away from the business by analyzed their shopping behavior. Each customer might have different fading period depends on their shopping pattern, such as customer who orders every week will be considered as fading if he/she is idle for 30 days while customer who orders every month will be considered as fading if he/she is idle for 90 days.

  Churn detection can alerts retailer to approach their fading customers at the right time to turns them back to be a regular customer.

- Customer Segmentation

  Segment customers into several groups based on different criteria, for example, demographic, preference, or shopping style. This can be used to help marketer to target an appropriate customer segment with a certain advertising campaign.

- Cohort Analysis

  Analyze different cohorts, or groups, of customer behave over time. This can be used to compares an income from different groups of customer.

- Trend Analysis

  Visualized the number of customers acquired, profit, revenue, or customer lifetime value over time to see the summary of the business.

- Email Integration

  Custora enables user to easily integrate other email services provider for any marketing email campaign.

- Automated Marketing

  Custora let users to design a specific email content for each specific customer.

- Pricing

  Custora has a subscription-based business model. The service price will be different for different tier of user. The information about Custora's pricing model are in following figure:

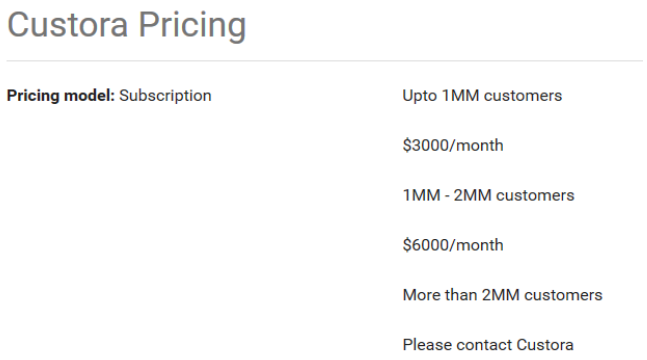


Figure 33:Custora pricing model. Taken from https://www.getapp.com/ business-intelligence-analytics-software/a/custora/pricing/

2.8.4 Futurelytics

Figure 34: Logo of Futurelytics

Futurelytics is marketing analytics and automation platform for e-commerce websites. There are many features integrated with Futurelytics, such as customer segmentation and product recommendation. Futurelytics also provides marketing automation to optimize marketing campaigns and emails marketing.

- Features
  - Overview visualization

    Futurelytics provides an analytics dashboard to visualize the overall income and customers for retailer to see their business status.

  - Behavioral Segmentation

    Segment each customer into several groups based on their purchasing behavior.

- Product Recommendation

  Predicts up to four recommended product for each customer to be embedded with the marketing email campaign.

- Marketing Automation

  Integrated an email service provider and provide a personalize email campaign to each customers.

- Pricing

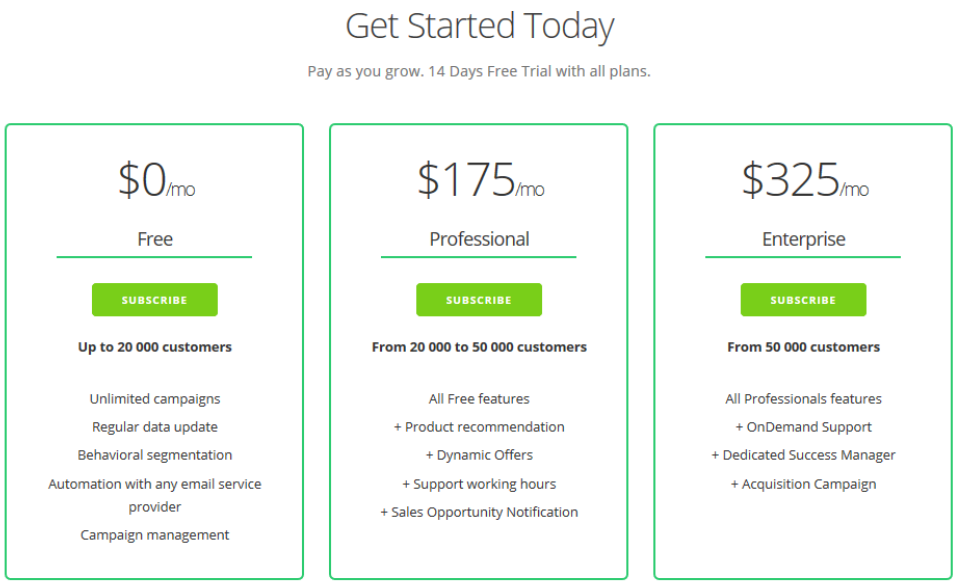


Figure 35: Futurelytic's pricing model

Futurelytic has a subscription-based pricing model and provides 14 days free trial for first-timer. Each different tier will have a different access to Futurelytics' services and different limited number of customer.

2.8.5 CAPP Advantages

Our cloud-based big data analytics solution offers affordable service for SMEs to sustain and further develop their businesses. SMEs will be able to improve their customers' loyalty through personalized customer experience and increase their profit. They will be able to utilize our analytic results to compete with large-sized enterprises and pave their way into AEC and other international markets.

There are many other big data analytic solutions offered by various IT and consulting companies nowadays. However, the price of those solutions is very high and a private server is required to integrate the solution. Those solutions are not suitable for SMEs because of limited funding. On the other hand, CAPP will be the first cloud-based big data analytics solution that are readily integrated into popular data sources and the service will be offered at affordable price for SMEs. Our SaaS pricing model is low-commitment and ideal for SMEs who want to figure out whether data analytics can help their businesses. Moreover, our technology is highly scalable and can be applied to different businesses with different data sizes; therefore, even though our main targets are SMEs, we can also customize solutions for larger enterprise, which provides funding for sustainable development of the technology.

Furthermore, as mentioned in the problem section, there is no existing tools that offer single customer view. CAPP will become the first service that answer this need by gathering the data from the available channels and interpret them together in our cloud. This provides precise insights that will help enterprises connect each customer with the right offer through the right contact channel.

2.9 Web Application Development

In this section, the technologies for web application development will be discussed.

- Hyper Text Markup Language (HTML)

Figure 36: a version of HTML logo

HTML is a markup language suitable for creating web page [31]. The web browser can read HTML and shows the content of the website. It is written in the form of tag, such as fonts, image, colors, or tables, to achieve the desire format. It is often used with Java script and CSS to create a website. The current version is HTML 5.

- PHP

Figure 37: a version of PHP logo

PHP is a scripting language that is suitable for web development. It is a server-side scripting language [32] meaning that it will be executed at the server side not the client side. It can also be embedded in HTML [33]. Moreover, it can also be used for general purpose programming. It is an open-source language with a strong community support.

- JavaScript

Figure 38: a version of JavaScript logo

JavaScript is a programming language that is used for making web interactive [34]. JavaScript may sound similar to Java but they are 2 different programming language. Unlike PHP, JavaScript is a client-side scripting language which means it will be executed on the clients' (users') computer.

- Cascading Style Sheets (CSS)

Figure 39: a version of CSS logo

CSS is a styling language used for customizing the appearance of a markup language document such as HTML [35]. It is used for web page styling including font size, font style, background, transparency, section size, etc. CSS can transform a website from outdated into a very nice looking one.

- MySQL

Figure 40: Logo of MySQL

MySQL is a widely used relational database management system [36]. It is also the most popular database to use with php on the web [37]. It is open source and free. It runs on the server and uses standard SQL. MySQL and php are also able to work cross-platform, for example, developing on Linux and then run on Windows server and vice versa.

- Bootstrap

Figure 41: a version of Bootstrap logo

According to W3school, “Bootstrap is the most popular HTML, CSS, and JavaScript framework for developing responsive, mobile-first web sites.” [38] This project will use a Bootstrap Theme so that the website will be neat and beautiful. The theme that was used is called Gentelella which is a Bootstrap Admin Template by Colorlib [39].

- Chart JS

Figure 42: a version of Chart JS logo

Chart JS is an open source JavaScript library that uses HTML 5 canvas element to display different types of charts interactively [40]. According to Chart JS official website, “Chart.js is an easy way to include animated, interactive graphs on your website for free. Simple yet flexible JavaScript charting for designers and developers” [41]. Our main focus on it is using mix chart type in the same chart and chart animation.

- ECharts

Figure 43: a version of ECharts logo

ECharts is used for plotting graphs similar to Chart JS. ECharts is developed by Baidu. It is a free and open source data visualization library for browser [42]. We use ECharts for the graphs that we think are more beautiful than Chart JS or not available in Chart JS such as pyramid graphs.

Chapter 3
Design and Methodology

3.1 Dataset Introduction

The dataset for customers and transactions that was used for the first half of our project is "Dunnhumby – The Complete Journey" [30]. It contains retail store data over 2 years from 2500 households. For customer segmentation, we first need to explore the customer demographic. Each record will illustrates characteristics of each customer in various aspect. The customer demographic and transactions table structure are as follow:

hh_demographic

This table contains demographic information for a portion of households. Due to nature of the data, the demographic information is not available for all households.

Variable	Description
HOUSEHOLD_KEY	Uniquely identifies each household
AGE_DESC	Estimated age range
MARITAL_STATUS_CODE	Marital Status (A - Married, B- Single, U - Unknown)
INCOME_DESC	Household income
HOMEOWNER_DESC	Homeowner, renter, etc.
HH_COMP_DESC	Household composition
HOUSEHOLD_SIZE_DESC	Size of household up to 5+
KID_CATEGORY_DESC	Number of children present up to 3+

AGE_DESC	MARITAL_STATUS_	INCOME_DESC	HOMEOWNER_DESC	HH_COMP_DESC	HOUSEHOLD_SIZE_	KID_CATEGORY_DESC	household_key
65+	A	35-49K	Homeowner	2 Adults No Kids	2	None/Unknown	1
45-54	A	50-74K	Homeowner	2 Adults No Kids	2	None/Unknown	7
25-34	U	25-34K	Unknown	2 Adults Kids	3	1	8
25-34	U	75-99K	Homeowner	2 Adults Kids	4	2	13
45-54	B	50-74K	Homeowner	Single Female	1	None/Unknown	16
65+	B	Under 15K	Homeowner	2 Adults No Kids	2	None/Unknown	17

Figure 44: Customer demographic table description and example data from Dunnhumby dataset user guide

transaction_data

This table contains all products purchased by households within this study. Each line found in this table is essentially the same line that would be found on a store receipt.

Variable	Description
HOUSEHOLD_KEY	Uniquely identifies each household
BASKET_ID	Uniquely identifies a purchase occasion
DAY	Day when transaction occurred
PRODUCT_ID	Uniquely identifies each product
QUANTITY	Number of the products purchased during the trip
SALES_VALUE	Amount of dollars retailer receives from sale
STORE_ID	Identifies unique stores
COUPON_MATCH_DISC	Discount applied due to retailer's match of manufacturer coupon
COUPON_DISC	Discount applied due to manufacturer coupon
RETAIL_DISC	Discount applied due to retailer's loyalty card program
TRANS_TIME	Time of day when the transaction occurred
WEEK_NO	Week of the transaction. Ranges 1 - 102

household_key	BASKET_ID	DAY	PRODUCT_ID	QUANTITY	SALES_VALUE	STORE_ID	RETAIL_DISC	TRANS_TIME	WEEK_NO	COUPON_DISC	COUPON_MATCH_DISC
2375	26984851472	1	1004906	1	1.39	364	-0.6	1631	1	0	0
2375	26984851472	1	1033142	1	0.82	364	0	1631	1	0	0
2375	26984851472	1	1036325	1	0.99	364	-0.3	1631	1	0	0
2375	26984851472	1	1082185	1	1.21	364	0	1631	1	0	0
2375	26984851472	1	8160430	1	1.5	364	-0.39	1631	1	0	0

Figure 45: Transaction table description and example data from Dunnhumby dataset user guide

The customer demographic and transactions analysis will be the main focus for the first phase of this project, which focuses on exploring the data and gain some valuable insight. The insight from those data can be used as a fundamental knowledge for the second phase of this project, which will focus on items analysis and recommendation. More tables and their relationships in the dataset can be found in the appendices.

This Dunnhumby data set is simply a testing set to be implemented in the prototype since it does not model the behavior of Thai people. It is solely for the purpose of learning the mechanism and pattern of big data analytic solution for this project.

3.2 Data Exploratory

For the first phase of this project, the main purpose is to explore the data and gain an insight. To be able to explore the data efficiently, suitable processing power will be needed. The size of the whole dataset is 2.15 GB of comma separated values (.csv) files and some files are over 600 MB. It needs more processing power than our PC in order to run all processing tasks smoothly, so we moved dataset to a Hadoop server to be processed later.

3.2.1 Storing data in Cloudera Hadoop

Step 1: Connect to the server with winSCP program to transfer the dataset to server.

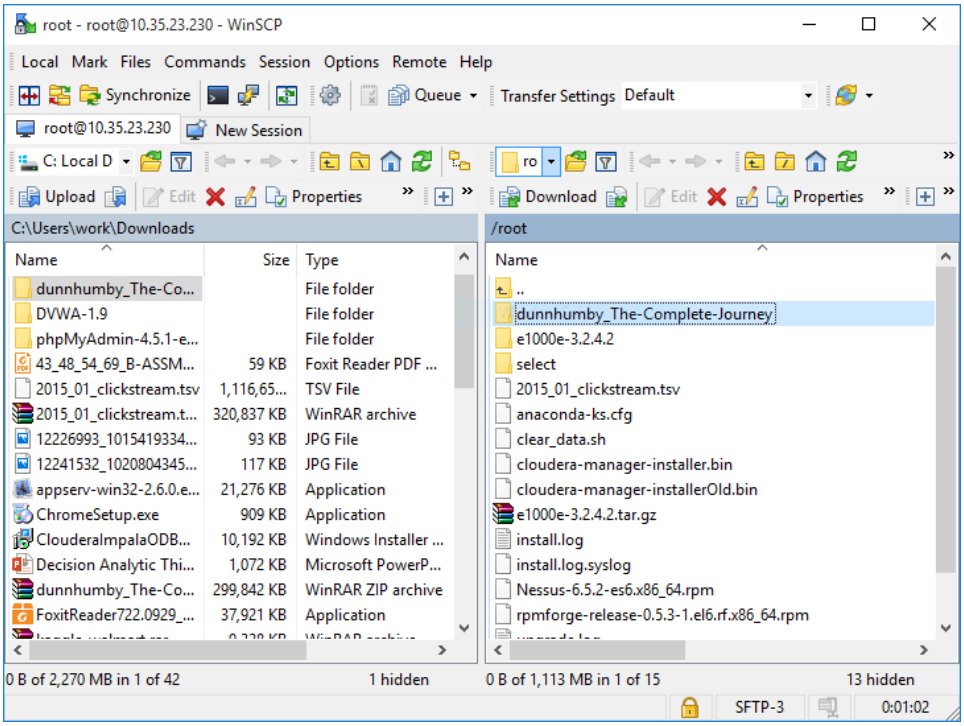


Figure 46: WinSCP GUI showing the process of moving the dataset to Hadoop server

Step 2: Check the uploaded files with Hue. Go to Metastore Manager, create a new database and import the files as new table.

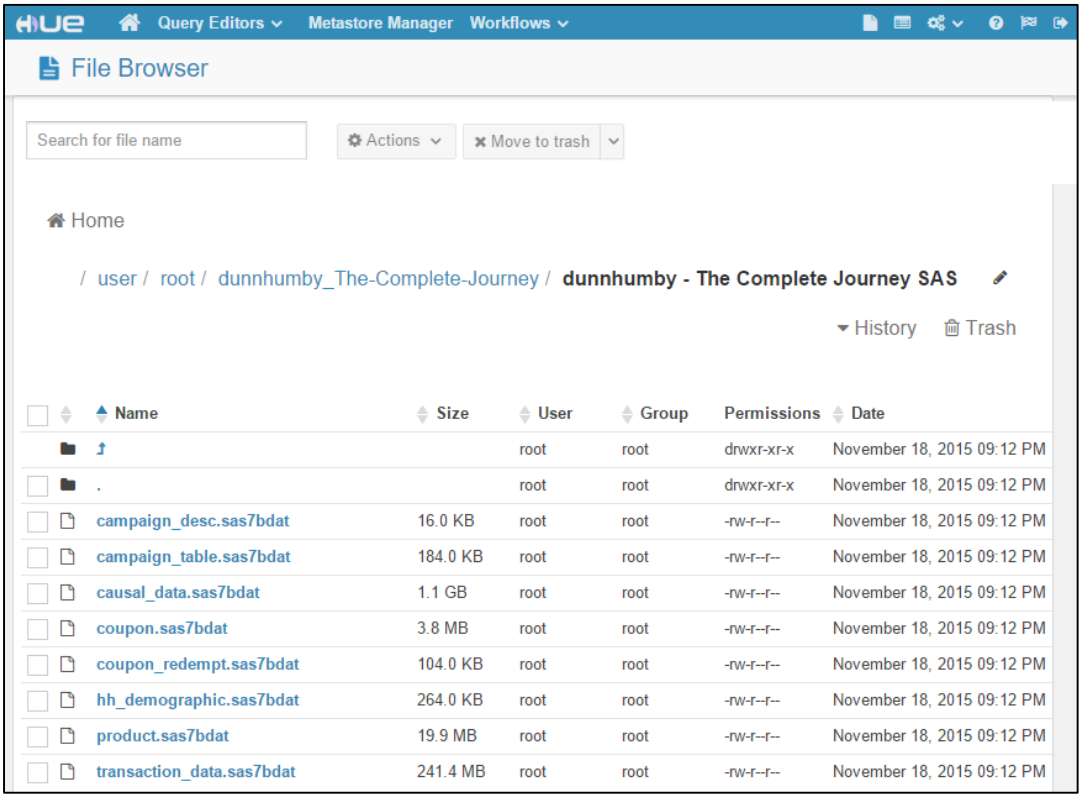


Figure 47: Hadoop UI showing the dataset in the server

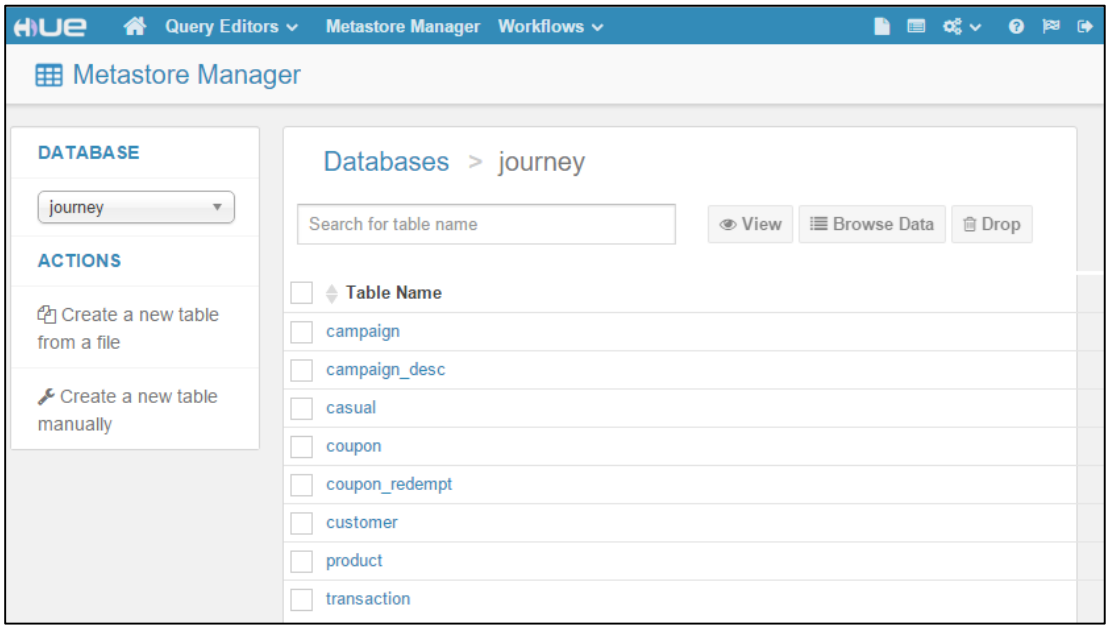


Figure 48: The result of creating database as importing files as tables

3.2.2 Data visualization using Tableau connected to Cloudera Hadoop using Impala

At first, Tableau in our PC cannot work smoothly on the dataset. However, once we connect Tableau to Impala, the visualization can be done more smoothly, the load time decreases and our PC won't freeze because it run out of memory anymore.

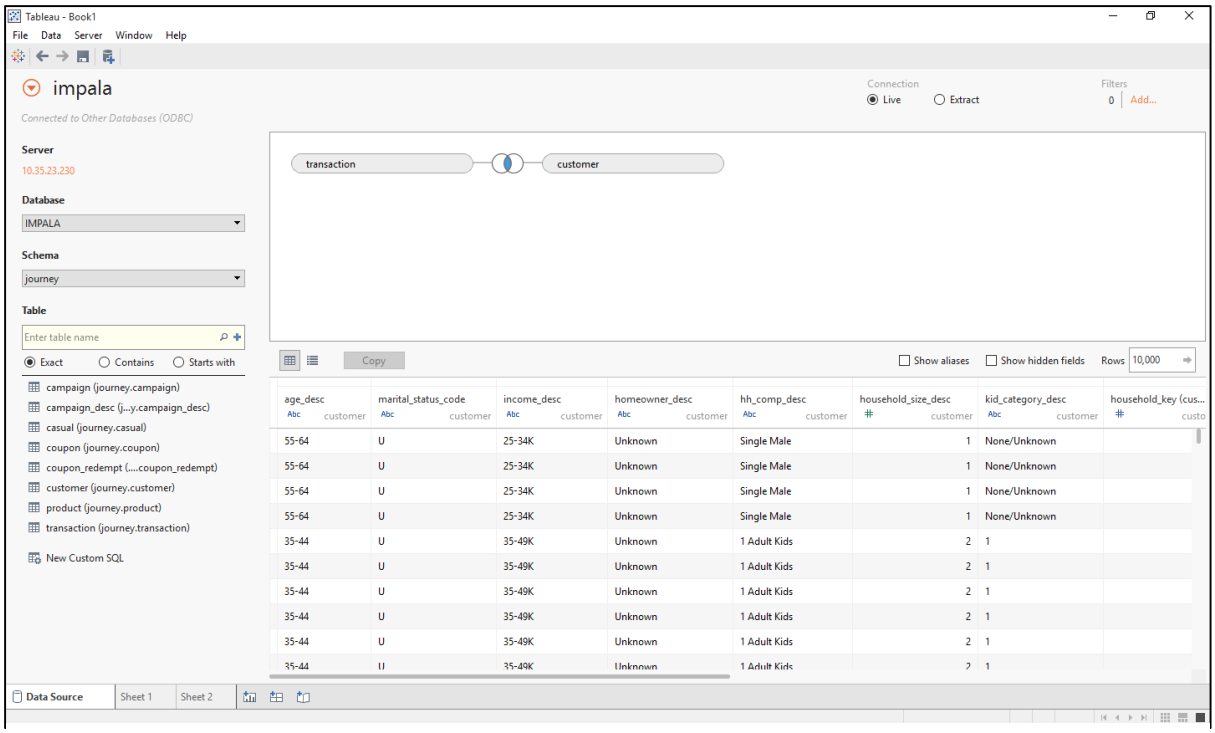
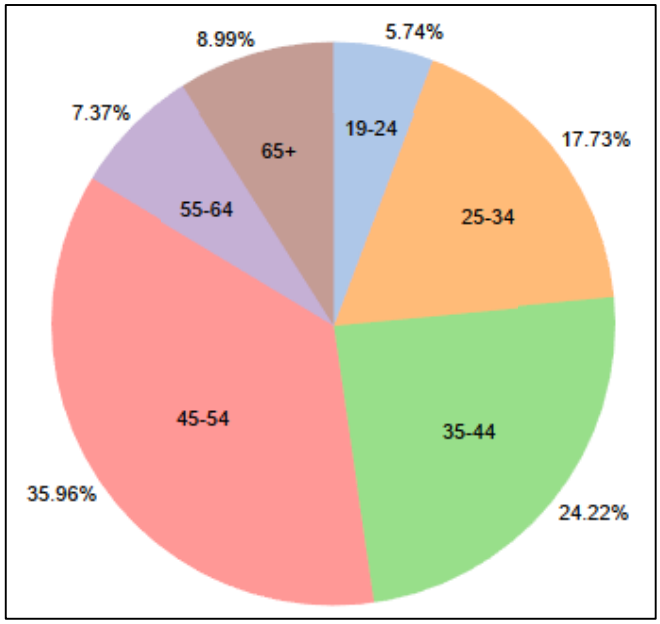


Figure 49: Tableau GUI showing dataset in Impala

We started by visualizing a few attributes first, then we gradually go into more dept.



First, we explore the customer composition by age group because it is easy to understand and the age group may give us further information to be utilized.

Figure 50: Pie chart showing customer age group composition

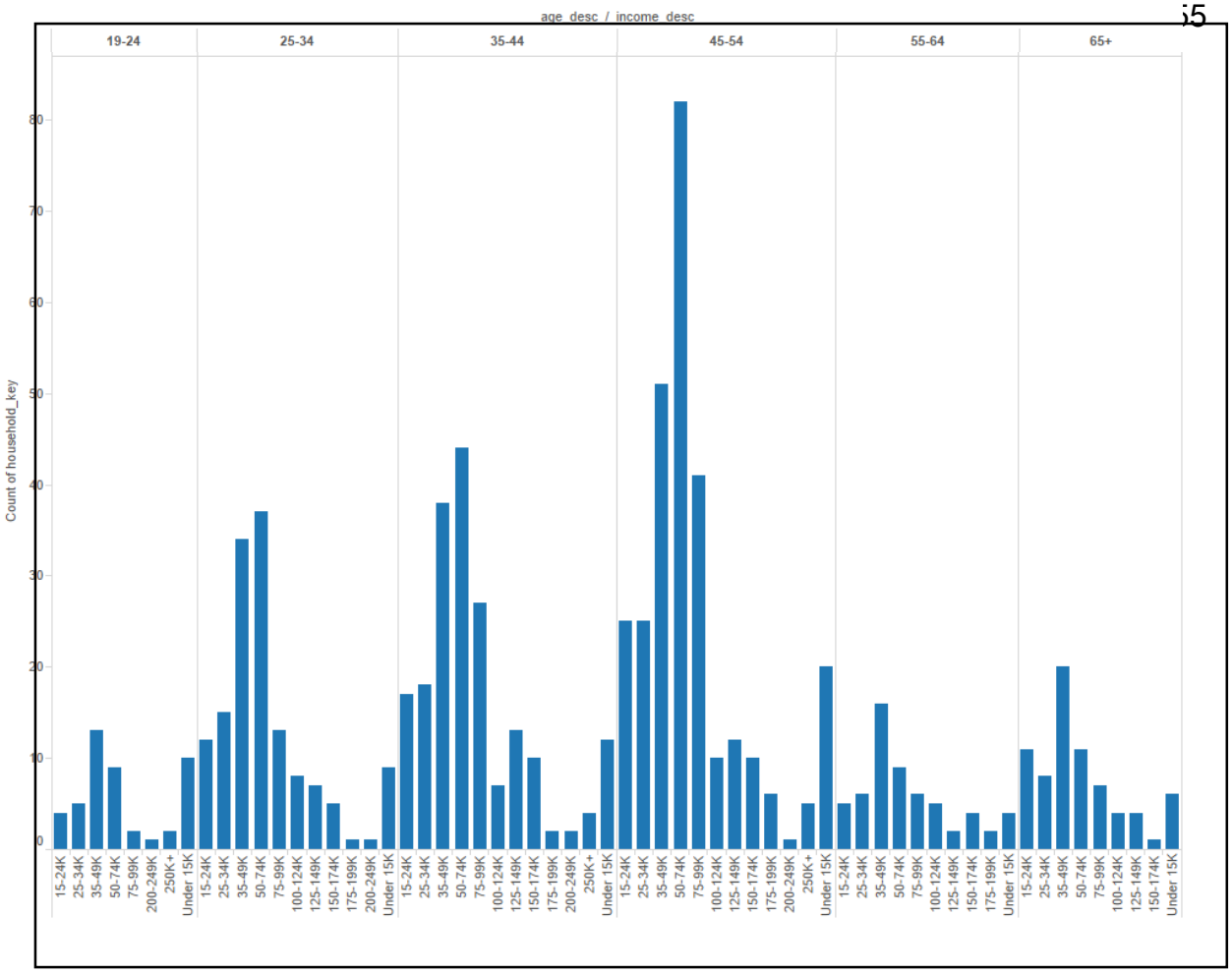


Figure 51: Histogram showing number of customer in age group and their income

From the age group composition in Figure 26 and income of each age group in Figure 27, we can see that customer aged 45-54 is the biggest group and they have high purchasing power. Therefore, this may imply that this age group should be our high priority target.

By examining customers, shopping behavior may be derivable. For example, we can combine the quantity that bought the product and product ID to see whether the customer may be buying a lot of goods to be sold in their own store or the customer is buying groceries for their own household.

3.3 Trial phase modeling with RapidMiner

In the initial state, we try modeling data in RapidMiner since it is easy to implement and get the result. However, we found out that RapidMiner in our PC will freeze because of the size of our input and it cannot handle many operations, such as joining tables, without crashing. Therefore, we can only try on a single table at a time. We will move the data modeling and machine learning to Cloudera Hadoop in a later phase of this project.

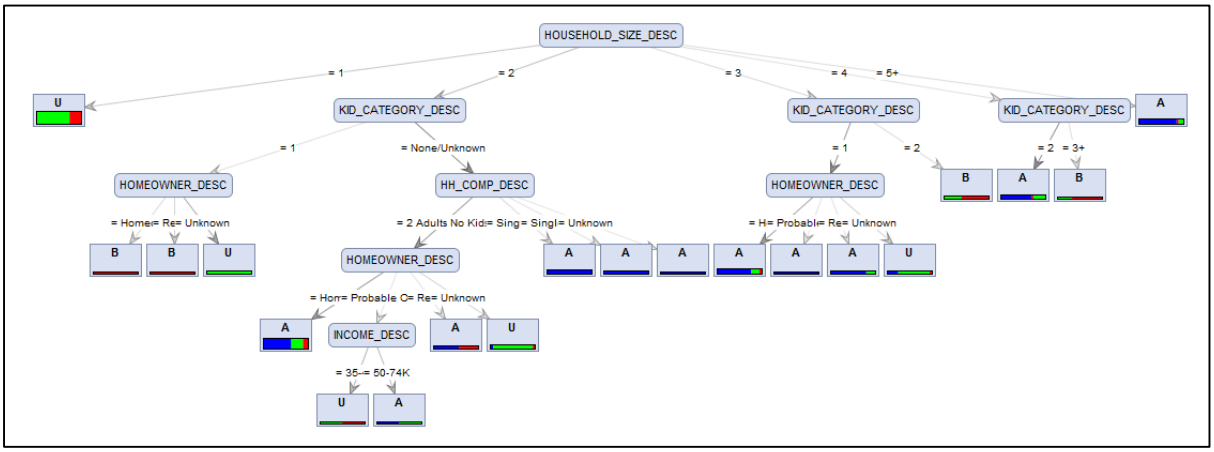


Figure 52: Sample decision tree classifying customers' marital status in RapidMiner

3.4 Project Plan and Design

After the data exploration was done, we start to design and plan the functionality and structure of the whole system. Our application can be separated into 3 modules:

Module 1: Data gathering. We will create enterprise data pool from 4 main sources.

- Transaction/customer/product data from enterprise's e-commerce platform
- Social media data from Facebook, Twitter, and others
- Business emails
- Clickstreams

Module 2: Analytic engine, report interface, and marketing tools. We will develop the following analytic capabilities.

- Customer segmentation
  - Segments from level of engagement and RFM (recency, frequency, monetary value)
  - Segments from product choices (use machine learning to discover customer segments for each business)
  - Segments from contact preference (social media, email, shop messaging)

- Product trends
    - Trending products (both views and purchases)
    - Market basket discovery
    - Generate social media campaign for hot products and product baskets
- Personalized recommendation
    - Generate personalized up-selling and cross-selling emails for specific customer segment.
- Targeted promotional campaign
    - Generate targeted discount offer for particular customer segments.
- Social media analytics
    - Rank social media channels for impact (likes and shares), traffic generation, and sale generation. Identify hot marketing content across channels at a glance.
    - Hashtag/competitor discovery
- Clickstream
    - Browsing depth (For example, businesses can segment customer based on how much time they spent browsing the site -> ‘not interested’, ‘interested’, ‘very interested’. Can contact the ‘very interested’ group to give lucrative offer).
    - To discover what effects conversion rate, where do customers get stuck, visualize customer journeys.

Module 3: APIs. Allow other popular digital marketing tools to be integrated to our solution.

However, not all of the above functionality and modules will be implemented in this senior project prototype. The features of our proof-of-concept prototype are as followed:

3.4.1 Prototype Features

Web Application:

- Login
- Registration
- View analytic results
    - Show graphs and/or dashboards appropriated for each result
- Create campaign by sending E-mails to customers
    - The targeted customers comes from the analytic result**.**
    - The recommended items comes from the analytic result**.**
    - The E-mail content is auto generated.
    - User selects an E-mail template
    - User reviews and can edit the content before sending to customers.

Data source:

We use offline transaction, customers, and products data from Dunnhumby.com. Other sources of data will be included in the future work but not in this senior project.

Analytic Engine:

- Customer segmentation
  Segment customers into different groups and generate reports and dashboards
  - Segments from level of engagement and RFM (recency, frequency, monetary value)
  - Segments from product choices (use machine learning to discover customer segments for each business)

- Product trend and classification
  - Trending products (both views and purchases)
  - Market basket discovery
  - Item sets

- Personalized recommendation
  - Generate personalized up-selling and cross-selling emails for specific customer segment.

Marketing Tool:

Generate E-mail to targeted group of customers with recommended products.

3.5 Architectural Overview

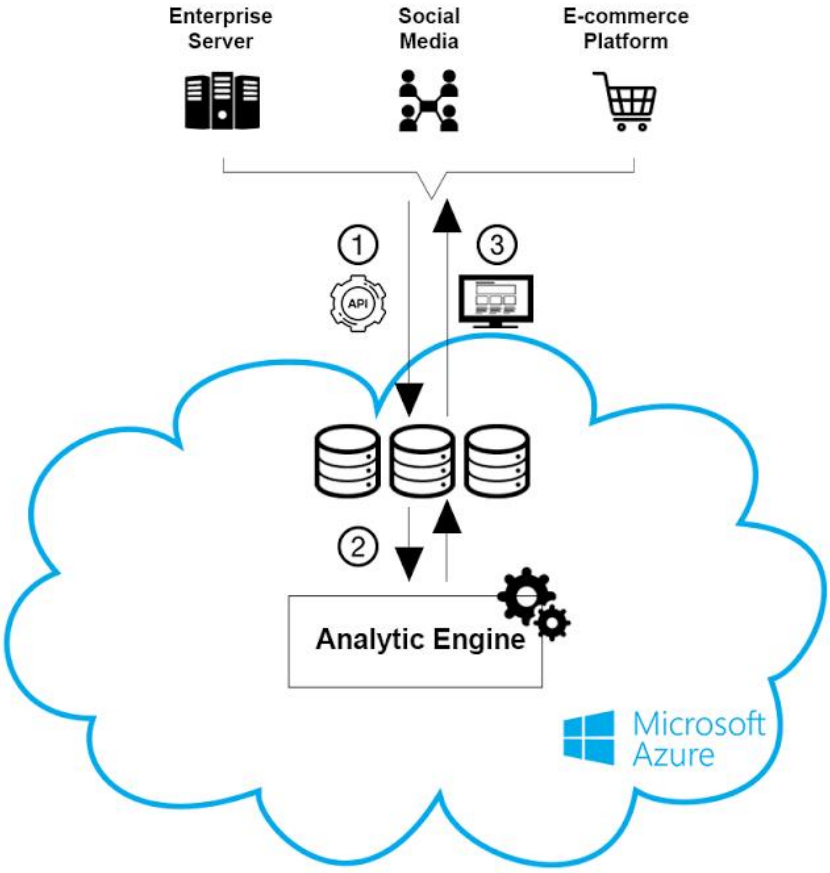


Figure 53: An architectural Overview of CAPP

1. APIs get data from SMEs' server and other sources, namely social media and e-commerce platform. The data is stored in CAPP cloud data storage.
   Remark: The prototype will use offline data already stored in the cloud.
2. The data are fed to the analytic engine for analysis and machine learning model training
3. The results are sent back to the dashboard as easy-to-interpret reports. Then SMEs can utilize the analytic results to offer personalized recommendation to customers.

3.6 Semester 1/2015 Progress

For the first semester, our main focus is on doing researches, acquiring data, data preprocessing, data exploratory, and simple model construction. This process was done by both our group and another group who also work on this project because before we can do any analytics, we need to have basic understanding of data science process and the data itself.

Analytic Trials

- **Classification**

We want to know what kind of customer buy the product. This is quite useful in a lot of situation such as when we want to recommend the product in the future. We will know which group of customer we want to target. We filter which customer purchase the target product and merge it together with customer demographic table. The first example we select the "garden" in the product category and classify the customer group while assuming people who is a home owner and married is likely to purchase the product in this category.

```
> class <-dtnov[, list(buy=("GARDEN" %in% DEPARTMENT)), by=list(household_
key)]
> cust <-merge(cust, class, by="household_key")
```

Filter

	household_key	buy	AGE_DESC	MARITAL_STATUS_CODE	INCOME_DESC	HOMEOWNER_DESC	HH_COMP_DESC	HOUSEHOLD_SIZE_DESC	KID_CATEGORY_DESC
1	1	FALSE	65+	A	35-49K	Homeowner	2 Adults No Kids	2	None/Unknown
2	7	TRUE	45-54	A	50-74K	Homeowner	2 Adults No Kids	2	None/Unknown
3	8	TRUE	25-34	U	25-34K	Unknown	2 Adults Kids	3	1
4	13	TRUE	25-34	U	75-99K	Homeowner	2 Adults Kids	4	2
5	16	FALSE	45-54	B	50-74K	Homeowner	Single Female	1	None/Unknown
6	17	FALSE	65+	B	Under 15K	Homeowner	2 Adults No Kids	2	None/Unknown
7	18	FALSE	45-54	A	100-124K	Homeowner	2 Adults No Kids	2	None/Unknown
8	19	FALSE	35-44	B	15-24K	Unknown	Single Female	1	None/Unknown
9	20	TRUE	25-34	A	75-99K	Renter	2 Adults No Kids	2	None/Unknown
10	22	FALSE	45-54	A	75-99K	Homeowner	2 Adults No Kids	2	None/Unknown
11	25	FALSE	35-44	U	50-74K	Unknown	Unknown	1	None/Unknown
12	27	FALSE	45-54	U	25-34K	Probable Renter	Single Female	1	None/Unknown
13	31	TRUE	35-44	B	35-49K	Homeowner	2 Adults No Kids	2	None/Unknown
14	39	TRUE	35-44	A	125-149K	Homeowner	2 Adults No Kids	2	None/Unknown
15	40	TRUE	45-54	U	Under 15K	Homeowner	2 Adults No Kids	2	None/Unknown
16	42	FALSE	65+	U	50-74K	Unknown	Single Male	1	None/Unknown
17	43	TRUE	35-44	B	15-24K	Homeowner	2 Adults Kids	5+	3+
18	46	TRUE	45-54	A	150-174K	Homeowner	2 Adults Kids	5+	3+

Showing 1 to 19 of 801 entries

Figure 54: garden purchased customer table

Attribute	Class FALSE (0.79)	TRUE (0.21)
AGE_DESC		
65+	57.0	17.0
45-54	225.0	65.0
25-34	116.0	28.0
35-44	147.0	49.0
19-24	44.0	4.0
55-64	47.0	14.0
[total]	636.0	177.0
MARITAL_STATUS_CODE		
A	259.0	83.0
U	278.0	68.0
B	96.0	23.0
[total]	633.0	174.0
HOMEOWNER_DESC		
Homeowner	386.0	120.0
Unknown	188.0	47.0
Renter	38.0	6.0
Probable Renter	12.0	1.0
Probable Owner	11.0	2.0
[total]	635.0	176.0

Figure 55: probability of purchase garden product

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         611               76.2797 %
Incorrectly Classified Instances       190               23.7203 %
Kappa statistic                          0.0723
Mean absolute error                      0.3222
Root mean squared error                  0.4168
Relative absolute error                 95.807  %
Root relative squared error            101.7207 %
Total Number of Instances              801

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.938     0.883     0.796      0.938     0.862      0.086     0.586     0.841     FALSE
               0.117     0.062     0.339      0.117     0.174      0.086     0.586     0.293     TRUE
Weighted Avg.  0.763     0.708     0.699      0.763     0.715      0.086     0.586     0.724

=== Confusion Matrix ===

   a   b   <-- classified as
 591  39 |   a = FALSE
 151  20 |   b = TRUE
```

Figure 56: validation

Next we select “Diapers and Disposable” product to classify the group of customer while assuming people who has a kid and has a household size more than 2 will likely to purchase the product.

```
> class <-dtnov[, list(buy=("DIAPERS & DISPOSABLES" %in% COMMODITY_DESC))
, by=list(household_key)]
> cust <-merge(cust, class, by="household_key")
```

Filter

	household_key	buy	AGE_DESC	MARITAL_STATUS_CODE	INCOME_DESC	HOMEOWNER_DESC	HH_COMP_DESC	HOUSEHOLD_SIZE_DESC	KID_CATEGORY_DESC
1	1	FALSE	65+	A	35-49K	Homeowner	2 Adults No Kids	2	None/Unknown
2	7	FALSE	45-54	A	50-74K	Homeowner	2 Adults No Kids	2	None/Unknown
3	8	TRUE	25-34	U	25-34K	Unknown	2 Adults Kids	3	1
4	13	FALSE	25-34	U	75-99K	Homeowner	2 Adults Kids	4	2
5	16	FALSE	45-54	B	50-74K	Homeowner	Single Female	1	None/Unknown
6	17	FALSE	65+	B	Under 15K	Homeowner	2 Adults No Kids	2	None/Unknown
7	18	FALSE	45-54	A	100-124K	Homeowner	2 Adults No Kids	2	None/Unknown
8	19	FALSE	35-44	B	15-24K	Unknown	Single Female	1	None/Unknown
9	20	FALSE	25-34	A	75-99K	Renter	2 Adults No Kids	2	None/Unknown
10	22	TRUE	45-54	A	75-99K	Homeowner	2 Adults No Kids	2	None/Unknown
11	25	FALSE	35-44	U	50-74K	Unknown	Unknown	1	None/Unknown
12	27	FALSE	45-54	U	25-34K	Probable Renter	Single Female	1	None/Unknown
13	31	FALSE	35-44	B	35-49K	Homeowner	2 Adults No Kids	2	None/Unknown
14	39	FALSE	35-44	A	125-149K	Homeowner	2 Adults No Kids	2	None/Unknown
15	40	FALSE	45-54	U	Under 15K	Homeowner	2 Adults No Kids	2	None/Unknown
16	42	FALSE	65+	U	50-74K	Unknown	Single Male	1	None/Unknown

Showing 1 to 17 of 801 entries

Figure 57: diaper purchased customer table

Then we use Naïve Bay algorithm in Weka to classify the group of customer. Our assumption is correct, people who has a kid and has a household size larger than 2 tend to purchase the diaper.

```
                          Class
Attribute                 FALSE   TRUE
                         (0.68) (0.32)
=====================================

HH_COMP_DESC
  2 Adults No Kids       190.0   67.0
  2 Adults Kids          100.0   89.0
  Single Female          116.0   30.0
  Unknown                 52.0   23.0
  Single Male             72.0   25.0
  1 Adult Kids            25.0   24.0
  [total]                555.0  258.0

HOUSEHOLD_SIZE_DESC
  2                      231.0   89.0
  3                       64.0   47.0
  4                       25.0   30.0
  1                      196.0   61.0
  5+                      38.0   30.0
  [total]                554.0  257.0
```

Figure 58: probability of purchase diaper

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         536               66.9164 %
Incorrectly Classified Instances       265               33.0836 %
Kappa statistic                          0.2152
Mean absolute error                      0.3795
Root mean squared error                  0.4865
Relative absolute error                 87.9668 %
Root relative squared error            104.7674 %
Total Number of Instances              801

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.778    0.567    0.749      0.778    0.763      0.216    0.629     0.771     FALSE
                 0.433    0.222    0.472      0.433    0.451      0.216    0.629     0.439     TRUE
Weighted Avg.    0.669    0.459    0.662      0.669    0.665      0.216    0.629     0.666

=== Confusion Matrix ===

   a   b   <-- classified as
 427 123 |   a = FALSE
```

Figure 59: validation

This classification only use customer demographic as a training model. To improve the implementation of the model, the further work is to add the product purchase pattern of each customer into a model.

- **Association Rules**

1. Loading necessary libraries.

```
> library(arules)
> library(arulesViz)
```

2. Load the dataset and plot the top 20 items to see what product is the most popular.

```
> transactionID <- read.transactions("C:/Users/Ploy/Desktop/trans.csv", sep=",")
> itemFrequencyPlot(transactionID,topN=20,type="absolute")
```

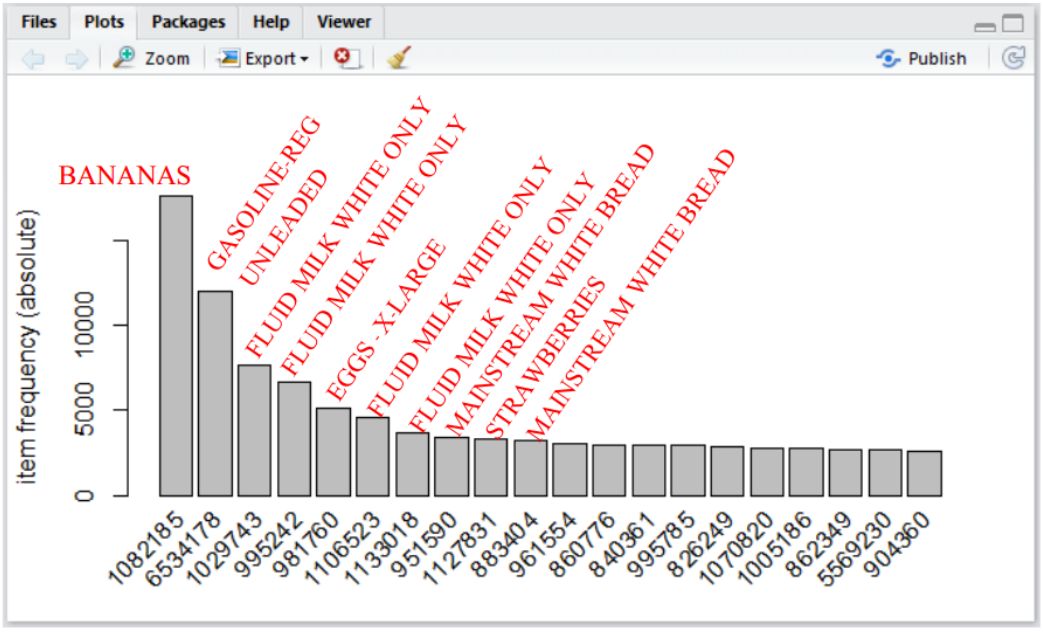


Figure 60: top selling product

From the above graph, *BANANAS* is the best seller product, However, there are more *FLUID MILK WHITE ONLY* when combining different product ID. This is because there are different sizes and brands for each product and each product has their own product ID. We will discuss about the different results from association rules with product ID and product name later.

3. Get the association rules

We set the min support to 0.001 and min confidence to 0.5 because there are many transactions, so we expect the support, which is the proportion of transactions which contains the item-set, to be low and the confidence is set to 0.5 because we think the middle
is appropriate and we can adjust it later. With this setup, we got 32 rules.

```
> rules <- apriori(transactionID, parameter = list(supp = 0.001, conf = 0.5))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport support minlen maxlen
        0.5    0.1    1 none FALSE            TRUE   0.001      1     10
 target   ext
  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 139

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[207777 item(s), 139873 transaction(s)] done [0.41s].
sorting and recoding items ... [1653 item(s)] done [0.04s].
creating transaction tree ... done [0.06s].
checking subsets of size 1 2 3 done [0.02s].
writing ... [32 rule(s)] done [0.02s].
creating S4 object  ... done [0.03s].
>
```

Figure 62: Apriori association by product ID

4. Inspect the rules after sorting the rules by confidence

```
Console ~/
> rulesConf<-sort(rules, by="confidence", decreasing=TRUE)
> inspect(rulesConf)
    lhs                  rhs       support     confidence lift
28 {1070820,1127831} => {1082185} 0.001215388 0.6882591  5.453091
27 {1126899,1127831} => {1082185} 0.001165343 0.6367188  5.044736
25 {1127831,866211}  => {1082185} 0.001422719 0.6317460  5.005337
23 {860776,878996}   => {1082185} 0.001000908 0.6306306  4.996499
24 {1029743,878996}  => {1082185} 0.001093849 0.6120000  4.848888
26 {1029743,866211}  => {1082185} 0.001294031 0.6114865  4.844820
3  {5586942}         => {1082185} 0.001222538 0.6021127  4.770551
30 {1127831,961554}  => {1082185} 0.001036655 0.5967078  4.727728
32 {1029743,1127831} => {1082185} 0.001744440 0.5865385  4.647156
31 {1127831,981760}  => {1082185} 0.001251135 0.5737705  4.545995
17 {968215}          => {1082185} 0.001994667 0.5728953  4.539061
14 {901666}          => {1082185} 0.001730141 0.5667447  4.490330
21 {1024306,1127831} => {1082185} 0.001072401 0.5639098  4.467869
29 {1127831,860776}  => {1082185} 0.001122447 0.5627240  4.458474
16 {1098248}         => {1082185} 0.002001816 0.5577689  4.419214
1  {1029112}         => {1082185} 0.001029505 0.5538462  4.388134
11 {5585635}         => {1082185} 0.001637199 0.5531401  4.382540
2  {1070538}         => {1082185} 0.001000908 0.5490196  4.349893
20 {933835,951590}   => {1082185} 0.001115297 0.5473684  4.336811
15 {880427}          => {1082185} 0.001580005 0.5338164  4.229438
12 {879528}          => {1082185} 0.001751589 0.5314534  4.210716
7  {998444}          => {1082185} 0.001508511 0.5288221  4.189868
18 {1062002}         => {1082185} 0.002516569 0.5277361  4.181264
19 {901062}          => {1082185} 0.004904449 0.5260736  4.168092
4  {1055853}         => {1082185} 0.001015207 0.5220588  4.136283
8  {7024990}         => {1082185} 0.001401271 0.5171504  4.097393
10 {1127179}         => {1082185} 0.001715842 0.5150215  4.080525
6  {865174}          => {1082185} 0.001193940 0.5091463  4.033977
9  {1098844}         => {1082185} 0.001415570 0.5076923  4.022456
22 {862349,981760}   => {1082185} 0.001151044 0.5015576  3.973851
```

Figure 63: Sorting the rule by confidence

5. Examining the rules. From step 4, we can see that the rhs of the rules are all 1082185 which are BANANAS. This is not very surprising since it is the best seller product. We mapped the product ID with their respective product name to get a clear picture of the rules.

1. FRZN BAGGED VEGETABLES - PLAIN
2. APPLES BRAEBURN (BULK&BAG)
3. YOGURT NOT MULTI-PACKS
4. YOGURT NOT MULTI-PACKS
5. YOGURT NOT MULTI-PACKS
6. APPLES GRANNY SMITH (BULK&BAG)
7. APPLES RED DELICIOUS (BULK&BAG
8. BLUEBERRIES
9. CLEMENTINES
10. NECTARINES YELLOW FLESH
11. YOGURT NOT MULTI-PACKS
12. INSTORE CUT FRUIT
13. MELON HALVES/QUARTERS
14. PEARS ANJOU
15. PREMIUM BREAD
16. PEARS BARTLETT
17. KIWI FRUIT
18. APPLES GOLD DELICIOUS (BULK&BA
19. APPLES GALA (BULK&BAG)
20. MEAT: SAUS DRY BULK, MAINSTREAM WHITE BREAD
21. CANTALOUPE WHOLE, STRAWBERRIES
22. FLUID MILK WHITE ONLY, EGGS - X-LARGE
23. CUCUMBERS, GRAPES RED
24. FLUID MILK WHITE ONLY, GRAPES RED
25. STRAWBERRIES, GRAPES WHITE
26. FLUID MILK WHITE ONLY, GRAPES WHITE
27. FLUID MILK WHITE ONLY, STRAWBERRIES
28. FLUID MILK WHITE ONLY, STRAWBERRIES
29. STRAWBERRIES, CUCUMBERS
30. STRAWBERRIES, CARROTS MINI PEELED
31. STRAWBERRIES, EGGS - X-LARGE
32. FLUID MILK WHITE ONLY, STRAWBERRIES

We visualized the rules and the result is shown in Figure 15 . The green circles with numbers are the product IDs, the arrows points from lhs to rhs, and the small yellow, orange, and red circles symbolized the support.

We found that product 901062 which is Apple has the most support which is shown by the biggest red circle in the graph.

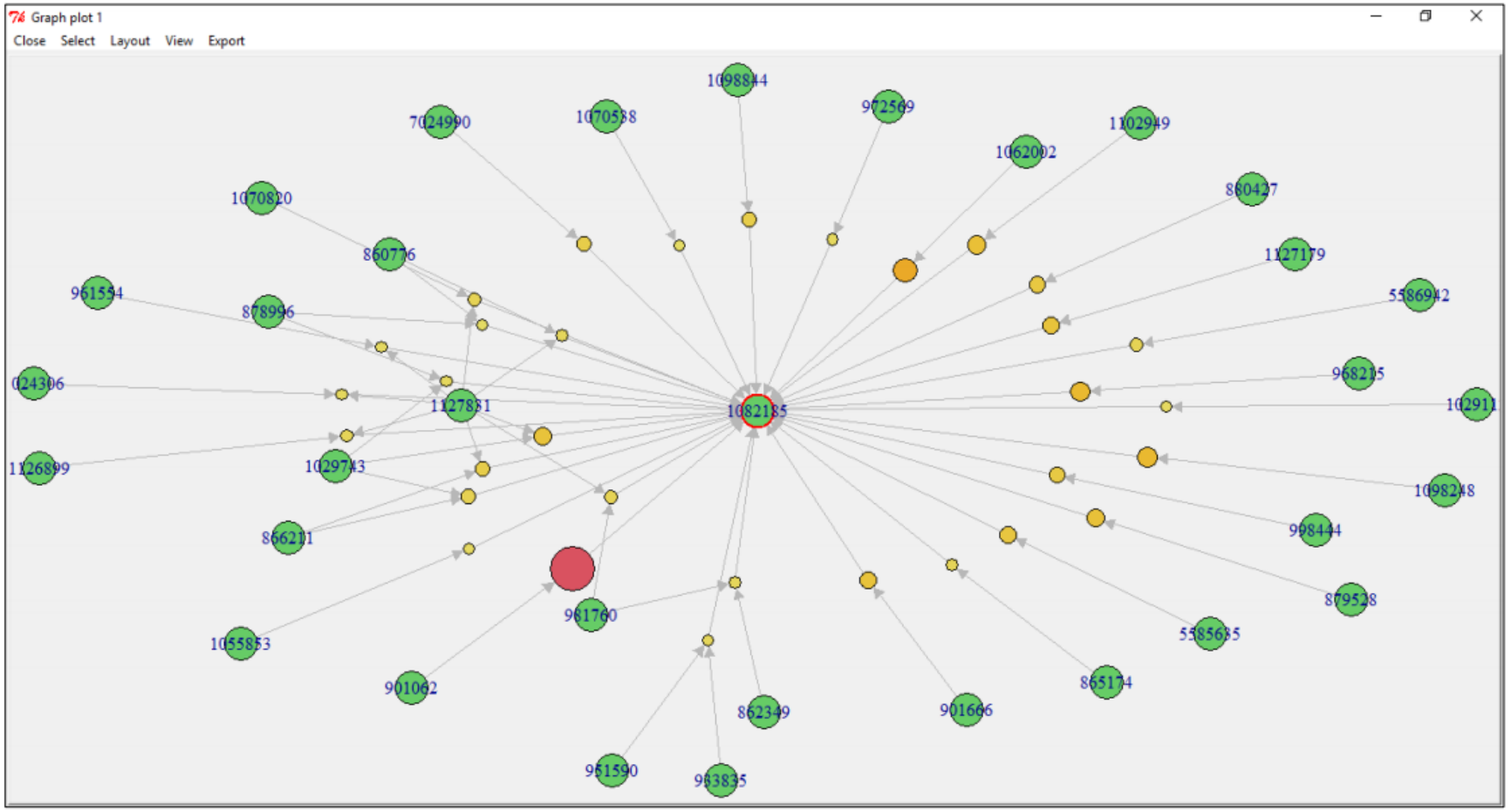
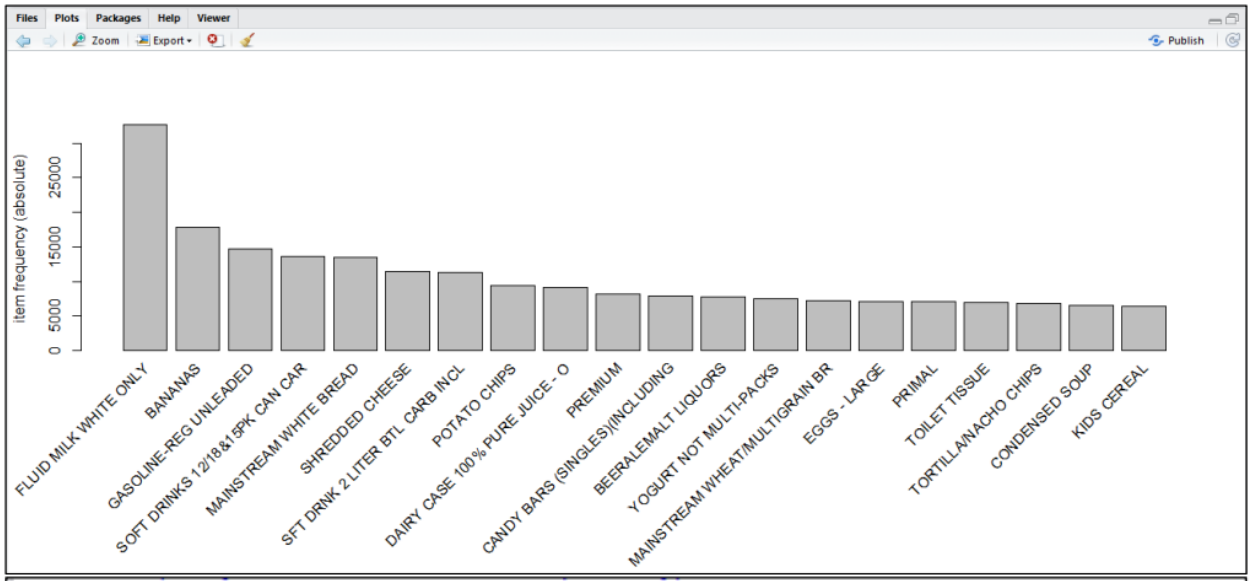


Figure 65 Association rules visualization

6. Load another dataset which has been preprocessed to contain product name instead of product ID in each transaction. Plot the top 20 items again to see the differences.

```
> transactionName <- read.transactions("C:/Users/Ploy/Desktop/product_name.csv", format="basket", sep=',', rm.duplicates=TRUE)
> itemFrequencyPlot(transactionName,topN=20,type="absolute")
```

Note: We load the data by putting rm.duplicates=TRUE so that the same product with different ID will only appear once in a transaction.

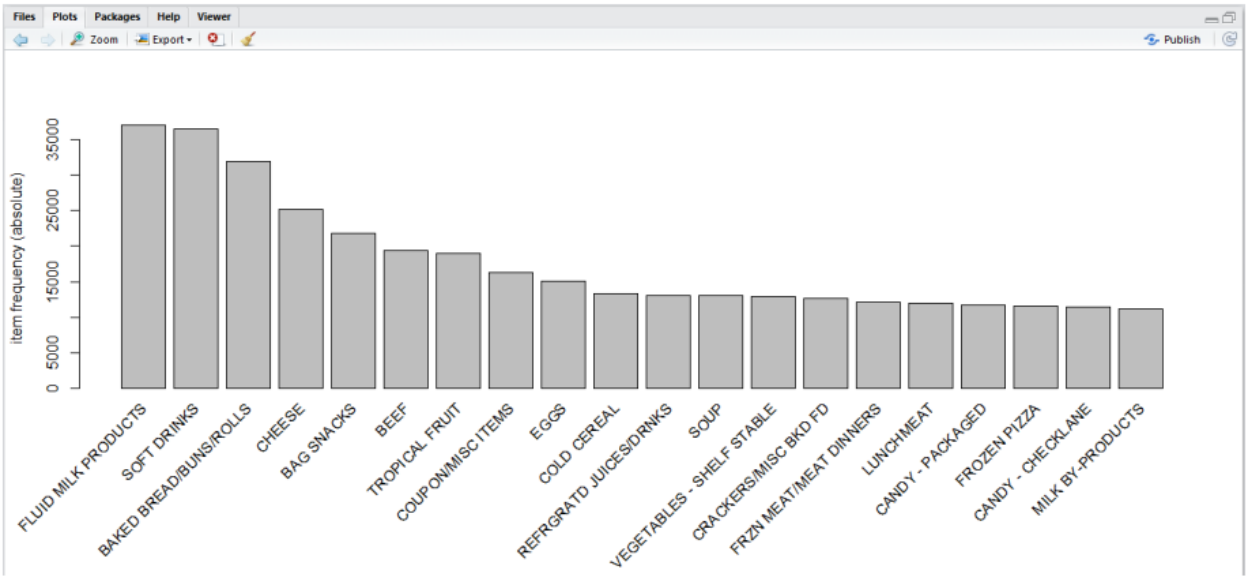


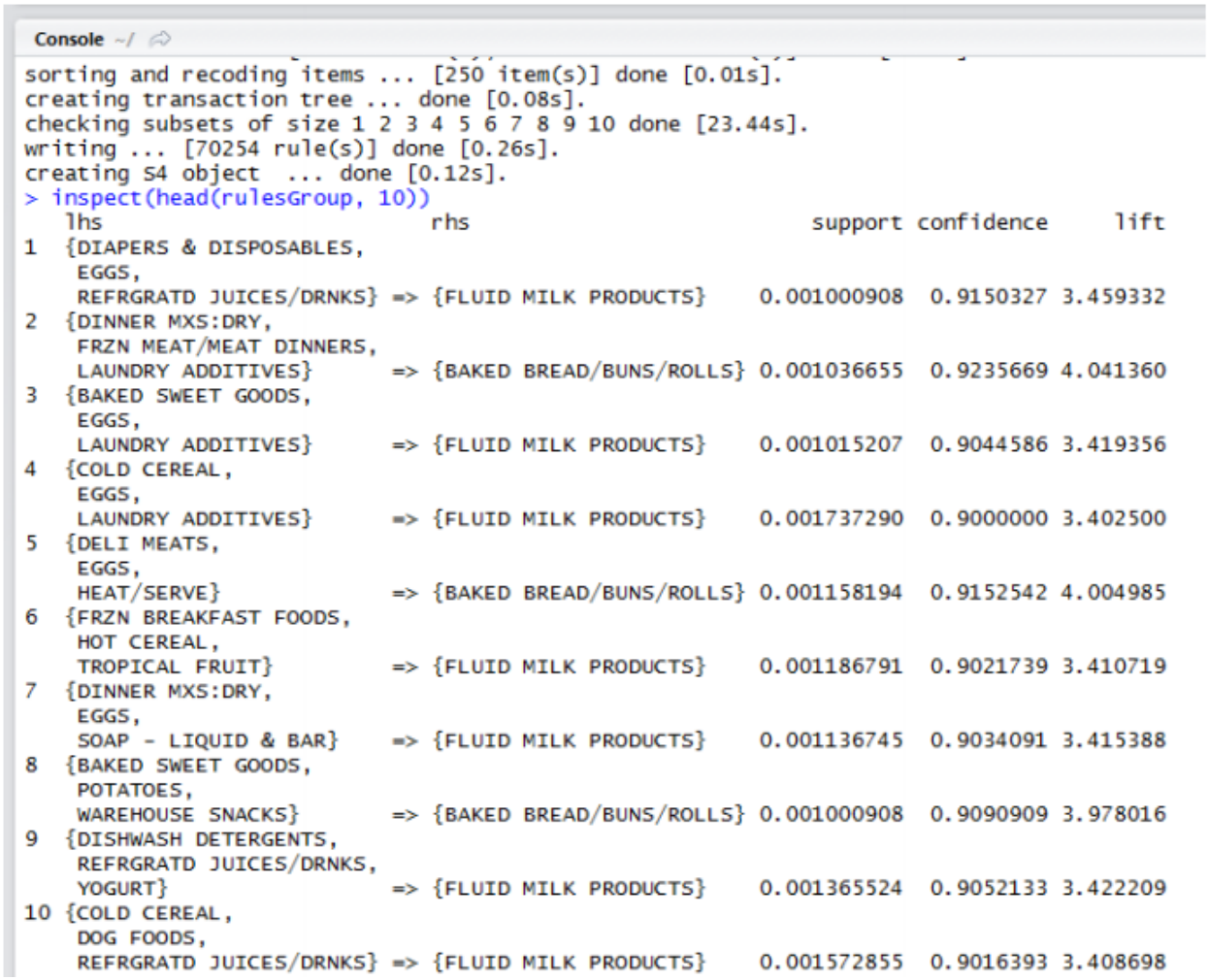

```
> rulesName <- apriori(transactionName, parameter = list(supp = 0.001, conf = 0.5))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport support minlen maxlen target   ext
        0.5    0.1    1 none FALSE            TRUE   0.001      1     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 139

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[142148 item(s), 139873 transaction(s)] done [0.17s].
sorting and recoding items ... [970 item(s)] done [0.05s].
creating transaction tree ... done [0.06s].
checking subsets of size 1 2 3 4 5 done [0.44s].
writing ... [18033 rule(s)] done [0.01s].
creating S4 object  ... done [0.08s].
>
```

Figure 67 apriori association by product name

We got 18,033 rules which is a lot more than the previous 32 rules. So, we will only look at the top 10 rules by confidence.

```
Console ~/
> inspect(head(rulesNameConf, 10))
   lhs                                                rhs                          support confidence     lift
1  {BANANAS,
    CHEESE CRACKERS (CHEEZ-ITS/GOL,
    MEAT:HAM BULK}                                 => {FLUID MILK WHITE ONLY} 0.001093849  0.9107143 3.900675
2  {BANANAS,
    DAIRY CASE 100% PURE JUICE - O,
    SOFT DRINKS 12/18&15PK CAN CAR,
    TOILET TISSUE}                                 => {FLUID MILK WHITE ONLY} 0.001072401  0.8982036 3.847090
3  {DAIRY CASE 100% PURE JUICE - O,
    IWS SINGLE CHEESE,
    TUNA}                                          => {FLUID MILK WHITE ONLY} 0.001029505  0.8944099 3.830842
4  {ALL FAMILY CEREAL,
    BANANAS,
    MAINSTREAM WHEAT/MULTIGRAIN BR,
    SHREDDED CHEESE}                               => {FLUID MILK WHITE ONLY} 0.001029505  0.8944099 3.830842
5  {BANANAS,
    CHEESE: PROCESSED BULK,
    DAIRY CASE 100% PURE JUICE - O}                => {FLUID MILK WHITE ONLY} 0.001015207  0.8930818 3.825153
6  {ALL FAMILY CEREAL,
    BANANAS,
    EGGS - LARGE,
    SHREDDED CHEESE}                               => {FLUID MILK WHITE ONLY} 0.001322628  0.8894231 3.809483
7  {BANANAS,
    DAIRY CASE 100% PURE JUICE - O,
    MAINSTREAM WHITE BREAD,
    PREMIUM}                                       => {FLUID MILK WHITE ONLY} 0.001015207  0.8875000 3.801246
8  {DAIRY CASE 100% PURE JUICE - O,
    IWS SINGLE CHEESE,
    KIDS CEREAL}                                   => {FLUID MILK WHITE ONLY} 0.001379823  0.8853211 3.791913
9  {CHOCOLATE MILK,
    KIDS CEREAL,
    SOFT DRINKS 12/18&15PK CAN CAR}                => {FLUID MILK WHITE ONLY} 0.001136745  0.8833333 3.783400
10 {DAIRY CASE 100% PURE JUICE - O,
    MAINSTREAM WHITE BREAD,
    PRETZELS}                                      => {FLUID MILK WHITE ONLY} 0.001029505  0.8780488 3.760765
>
```





```
Console ~/
sorting and recoding items ... [250 item(s)] done [0.01s].
creating transaction tree ... done [0.08s].
checking subsets of size 1 2 3 4 5 6 7 8 9 10 done [23.44s].
writing ... [70254 rule(s)] done [0.26s].
creating S4 object  ... done [0.12s].
> inspect(head(rulesGroup, 10))
   lhs                                rhs                          support confidence     lift
1  {DIAPERS & DISPOSABLES,
    EGGS,
    REFRGRATD JUICES/DRNKS}        => {FLUID MILK PRODUCTS}     0.001000908  0.9150327 3.459332
2  {DINNER MXS:DRY,
    FRZN MEAT/MEAT DINNERS,
    LAUNDRY ADDITIVES}             => {BAKED BREAD/BUNS/ROLLS} 0.001036655  0.9235669 4.041360
3  {BAKED SWEET GOODS,
    EGGS,
    LAUNDRY ADDITIVES}             => {FLUID MILK PRODUCTS}     0.001015207  0.9044586 3.419356
4  {COLD CEREAL,
    EGGS,
    LAUNDRY ADDITIVES}             => {FLUID MILK PRODUCTS}     0.001737290  0.9000000 3.402500
5  {DELI MEATS,
    EGGS,
    HEAT/SERVE}                    => {BAKED BREAD/BUNS/ROLLS} 0.001158194  0.9152542 4.004985
6  {FRZN BREAKFAST FOODS,
    HOT CEREAL,
    TROPICAL FRUIT}                => {FLUID MILK PRODUCTS}     0.001186791  0.9021739 3.410719
7  {DINNER MXS:DRY,
    EGGS,
    SOAP - LIQUID & BAR}           => {FLUID MILK PRODUCTS}     0.001136745  0.9034091 3.415388
8  {BAKED SWEET GOODS,
    POTATOES,
    WAREHOUSE SNACKS}              => {BAKED BREAD/BUNS/ROLLS} 0.001000908  0.9090909 3.978016
9  {DISHWASH DETERGENTS,
    REFRGRATD JUICES/DRNKS,
    YOGURT}                        => {FLUID MILK PRODUCTS}     0.001365524  0.9052133 3.422209
10 {COLD CEREAL,
    DOG FOODS,
    REFRGRATD JUICES/DRNKS}        => {FLUID MILK PRODUCTS}     0.001572855  0.9016393 3.408698
```

Figure 69: top 20 frequency by product category

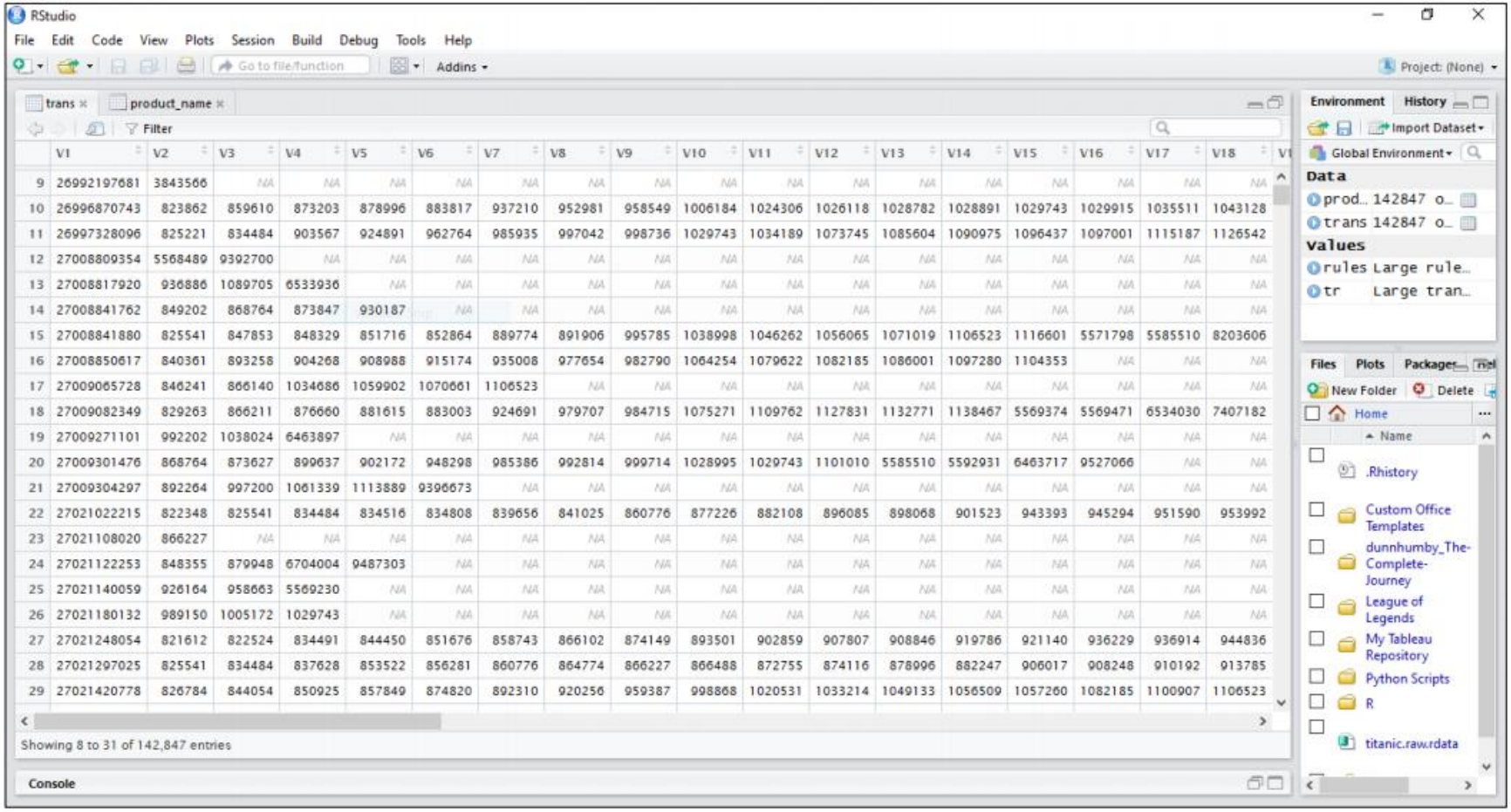
By using the product group for the association, we can see more relationships among various products.

From all the experiments, we can see that association rules can show different results based on different data given.

The product ID was very specific and do not provide the clear picture because the same products are divided into multiple IDs. Moreover, only a limited set of rules with low confidence can be mined.

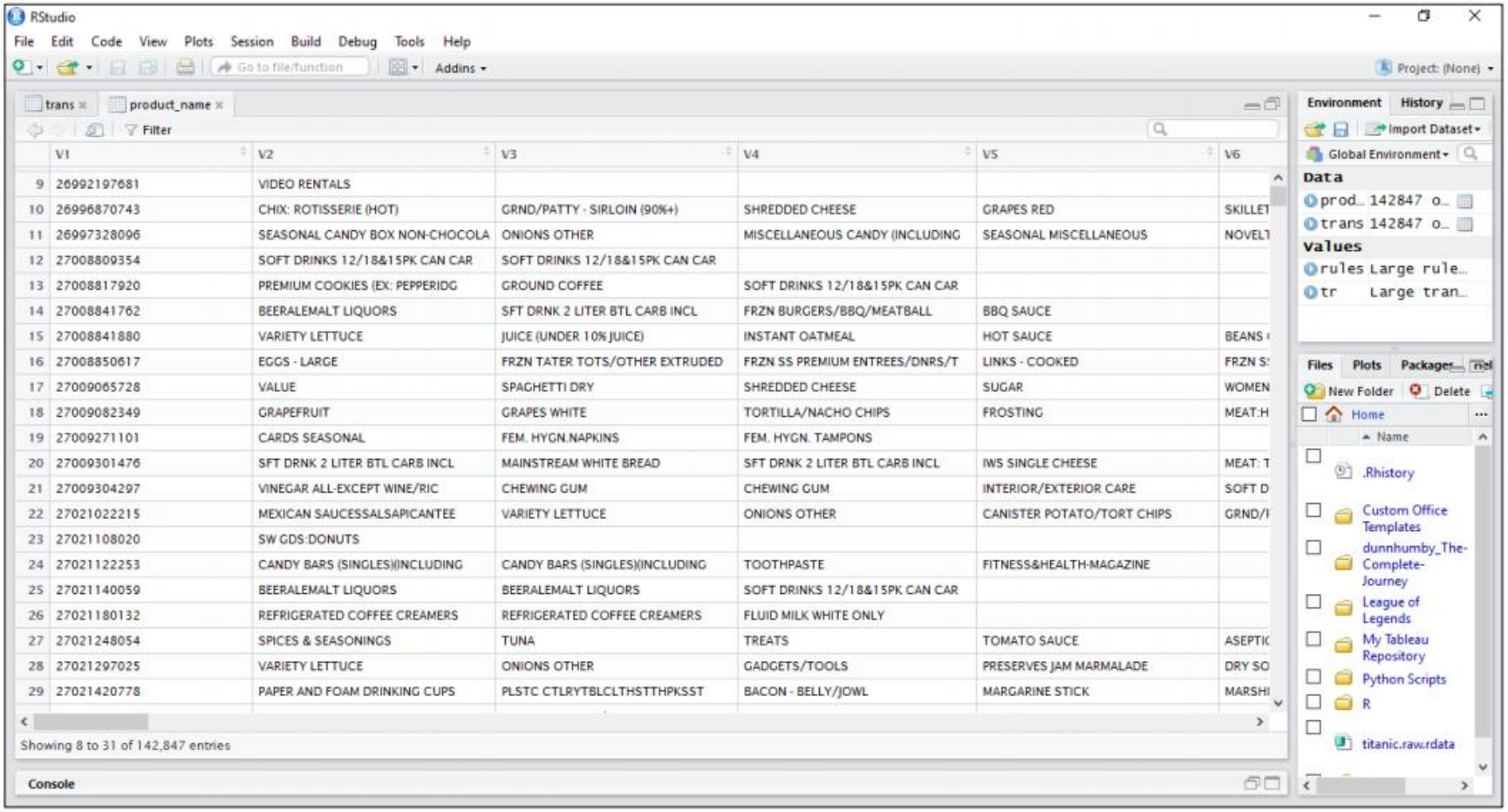
The product name group provided a clearer picture because the same products with different IDs are now groups together. We can derive a lot of rules with decent confidence from this product name association.

The product category group provided the big picture of what types of product customers are buying (which are mostly household items). We can get even more rules from this but since it is product category, we will not get the as detailed as the product name.



	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18
9	26992197681	3843566	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
10	26996870743	823862	859610	873203	878996	883817	937210	952981	958549	1006184	1024306	1026118	1028782	1028891	1029743	1029915	1035511	1043128
11	26997328096	825221	834484	903567	924891	962764	985935	997042	998736	1029743	1034189	1073745	1085604	1090975	1096437	1097001	1115187	1126542
12	27008809354	5568489	9392700	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
13	27008817920	936886	1089705	6533936	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
14	27008841762	849202	868764	873847	930187	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
15	27008841880	825541	847853	848329	851716	852864	889774	891906	995785	1038998	1046262	1056065	1071019	1106523	1116601	5571798	5585510	8203606
16	27008850617	840361	893258	904268	908988	915174	935008	977654	982790	1064254	1079622	1082185	1086001	1097280	1104353	NA	NA	NA
17	27009065728	846241	866140	1034686	1059902	1070661	1106523	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
18	27009082349	829263	866211	876660	881615	883003	924691	979707	984715	1075271	1109762	1127831	1132771	1138467	5569374	5569471	6534030	7407182
19	27009271101	992202	1038024	6463897	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
20	27009301476	868764	873627	899637	902172	948298	985386	992814	999714	1028995	1029743	1101010	5585510	5592931	6463717	9527066	NA	NA
21	27009304297	892264	997200	1061339	1113889	9396673	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
22	27021022215	822348	825541	834484	834516	834808	839656	841025	860776	877226	882108	896085	898068	901523	943393	945294	951590	953992
23	27021108020	866227	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
24	27021122253	848355	879948	6704004	9487303	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
25	27021140059	926164	958663	5569230	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
26	27021180132	989150	1005172	1029743	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
27	27021248054	821612	822524	834491	844450	851676	858743	866102	874149	893501	902859	907807	908846	919786	921140	936229	936914	944836
28	27021297025	825541	834484	837628	853522	856281	860776	864774	866227	866488	872755	874116	878996	882247	906017	908248	910192	913785
29	27021420778	826784	844054	850925	857849	874820	892310	920256	959387	998868	1020531	1033214	1049133	1056509	1057260	1082185	1100907	1106523

Figure 71: Transaction Table with transaction ID as the first column followed by series of items in the transaction as product ID.



	V1	V2	V3	V4	V5	V6
9	26992197681	VIDEO RENTALS				
10	26996870743	CHIX: ROTISSERIE (HOT)	GRND/PATTY - SIRLOIN (90%+)	SHREDDED CHEESE	GRAPES RED	SKILLET
11	26997328096	SEASONAL CANDY BOX NON-CHOCOLA	ONIONS OTHER	MISCELLANEOUS CANDY (INCLUDING	SEASONAL MISCELLANEOUS	NOVELT
12	27008809354	SOFT DRINKS 12/18&15PK CAN CAR	SOFT DRINKS 12/18&15PK CAN CAR			
13	27008817920	PREMIUM COOKIES (EX: PEPPERIDG	GROUND COFFEE	SOFT DRINKS 12/18&15PK CAN CAR		
14	27008841762	BEERALEMALT LIQUORS	SFT DRNK 2 LITER BTL CARB INCL	FRZN BURGERS/BBQ/MEATBALL	BBQ SAUCE	
15	27008841880	VARIETY LETTUCE	JUICE (UNDER 10% JUICE)	INSTANT OATMEAL	HOT SAUCE	BEANS
16	27008850617	EGGS - LARGE	FRZN TATER TOTS/OTHER EXTRUDED	FRZN SS PREMIUM ENTREES/DNRS/T	LINKS - COOKED	FRZN S
17	27009065728	VALUE	SPAGHETTI DRY	SHREDDED CHEESE	SUGAR	WOMEN
18	27009082349	GRAPEFRUIT	GRAPES WHITE	TORTILLA/NACHO CHIPS	FROSTING	MEAT:H
19	27009271101	CARDS SEASONAL	FEM. HYGN.NAPKINS	FEM. HYGN. TAMPONS		
20	27009301476	SFT DRNK 2 LITER BTL CARB INCL	MAINSTREAM WHITE BREAD	SFT DRNK 2 LITER BTL CARB INCL	IWS SINGLE CHEESE	MEAT: T
21	27009304297	VINEGAR ALL-EXCEPT WINE/RIC	CHEWING GUM	CHEWING GUM	INTERIOR/EXTERIOR CARE	SOFT D
22	27021022215	MEXICAN SAUCESSALSAPICANTEE	VARIETY LETTUCE	ONIONS OTHER	CANISTER POTATO/TORT CHIPS	GRND/I
23	27021108020	SW GDS:DONUTS				
24	27021122253	CANDY BARS (SINGLES)(INCLUDING	CANDY BARS (SINGLES)(INCLUDING	TOOTHPASTE	FITNESS&HEALTH-MAGAZINE	
25	27021140059	BEERALEMALT LIQUORS	BEERALEMALT LIQUORS	SOFT DRINKS 12/18&15PK CAN CAR		
26	27021180132	REFRIGERATED COFFEE CREAMERS	REFRIGERATED COFFEE CREAMERS	FLUID MILK WHITE ONLY		
27	27021248054	SPICES & SEASONINGS	TUNA	TREATS	TOMATO SAUCE	ASEPTIC
28	27021297025	VARIETY LETTUCE	ONIONS OTHER	GADGETS/TOOLS	PRESERVES JAM MARMALADE	DRY SO
29	27021420778	PAPER AND FOAM DRINKING CUPS	PLSTC CTLRYTBLCLTHSTTHPKSST	BACON - BELLY/JOWL	MARGARINE STICK	MARSHI

Figure 72: Transaction Table with transaction ID as the first column followed by series of items in the transaction as product name.

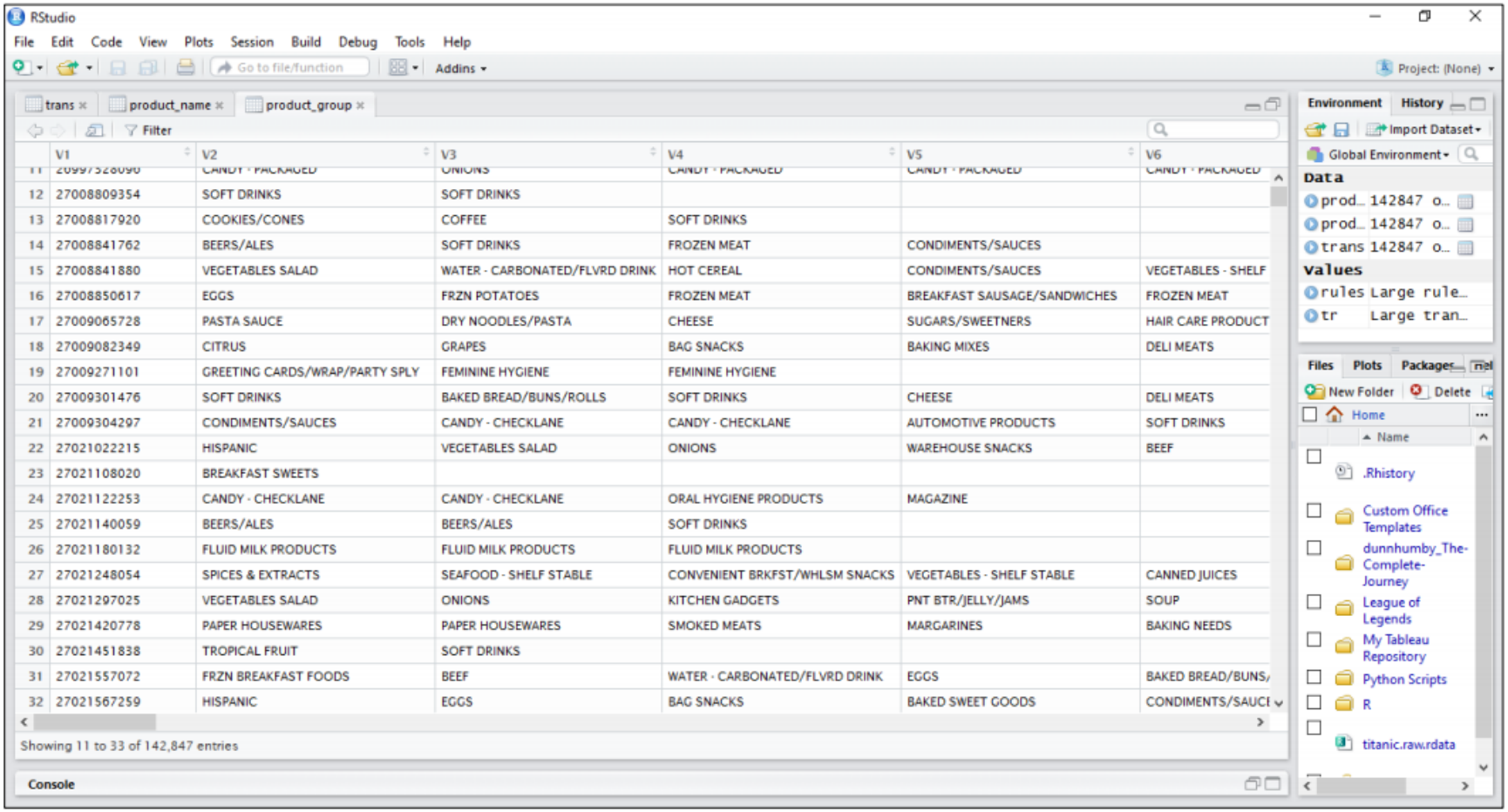


Figure 73: Transaction Table with transaction ID as the first column followed by series of items in the transaction as product category.

3.7 Semester 2/2015 Progress

As for the second semester, the project is divided into 2 parts; front-end and back-end. The front-end part mainly consists of user interface and the back-end part mainly is the analytics engine. Our group's responsibility is the front-end part while the other group was assigned the back-end part.

Designing the front-end web application

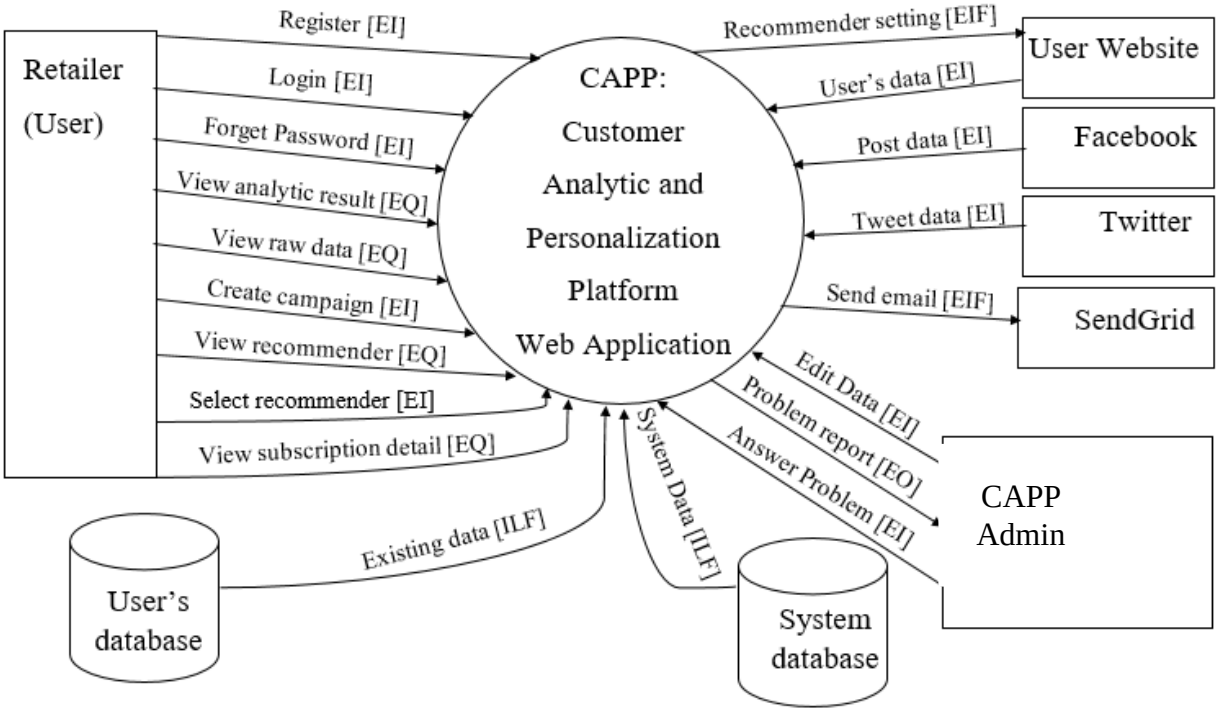


Figure 74: Architectural context diagram

Our customer analytic and personalization platform web application is design for 2 types of users; retailers and CAPP admin. For this prototype version only the retailer side is focused on.

The retailer can register to use CAPP and their member data will be kept in system database for authenticating access to the system. The user's existing data will be kept in user's database which is not the same as system database. For example, a company has products' data, customers' data, and transaction data. The data will be kept in the user's database while the company login details will be kept in system database.

Once the retailers are registered, they can log in to the system and view different kinds of data including raw data, analytic results, recommenders list, overview of their company performance, and their subscription details. The user can create e-mail campaign by choosing a recommender and entering the campaign information. The e-mail will be sent to their customer by SendGrid. They can also link the recommender to their e-commerce website to increase customer engagement.

Retailers can link their Facebook and twitter to CAPP to analyze the feedback and efficiency for their social media campaigns. The information from the analytic will later be used for customer profiling and also for campaign evaluation.

- User Registration

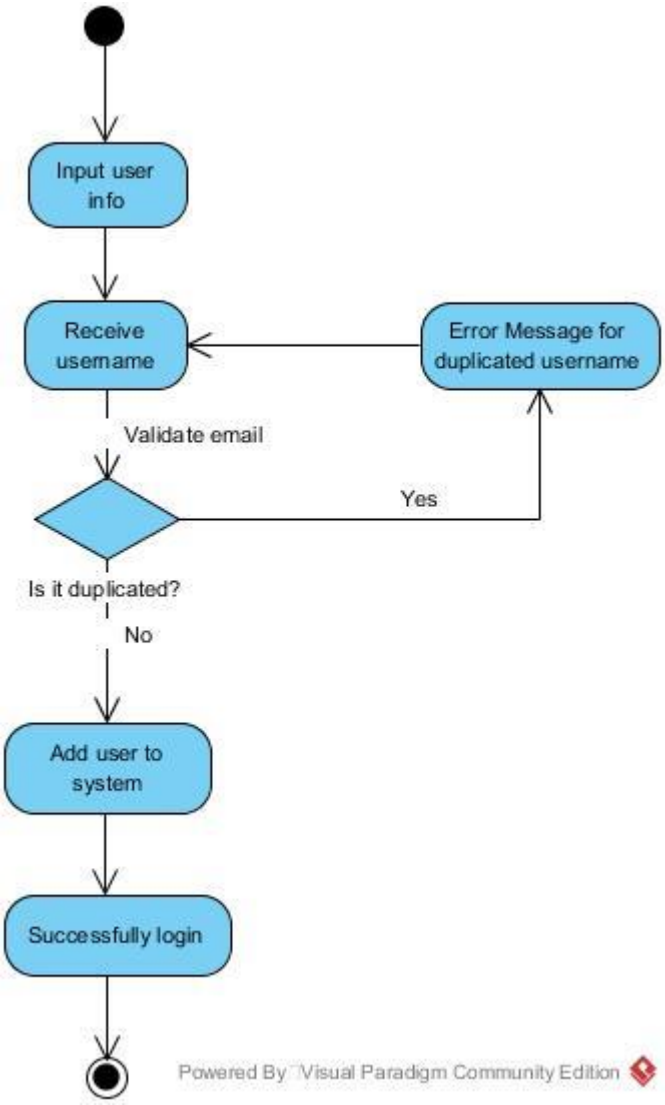


Figure 75: Registration activity diagram

Users need to be registered before using CAPP services. The registration only contains a few simple steps so that users can easily access the website. First, the users will enter their personal information along with the username and password for log in. Then, the username will be checked whether it already exists or not. If the username exists, an error message will prompt the user to try a different username. If the username does not already exist, the user will be added to the system and they can use their registered username and password to log in to the system.

- User Login

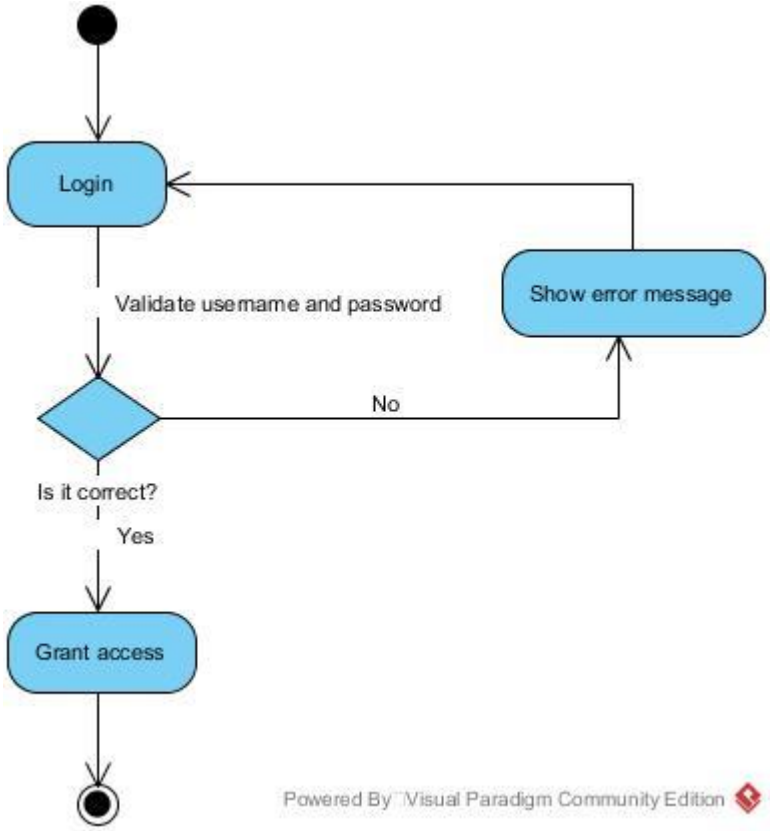


Figure 76: Login activity diagram

Once the users entered their username and password at the login screen, the information will be validated. If it matches with the information in the database, the user will be granted access to their company CAPP homepage. If the information does not match, users will be prompt to try again. This prototype version uses this simple mechanism for user login but for the actual deploy application, a more sophisticated and secure way to logged in will be implemented so that the security is more reliable.

- Forget password

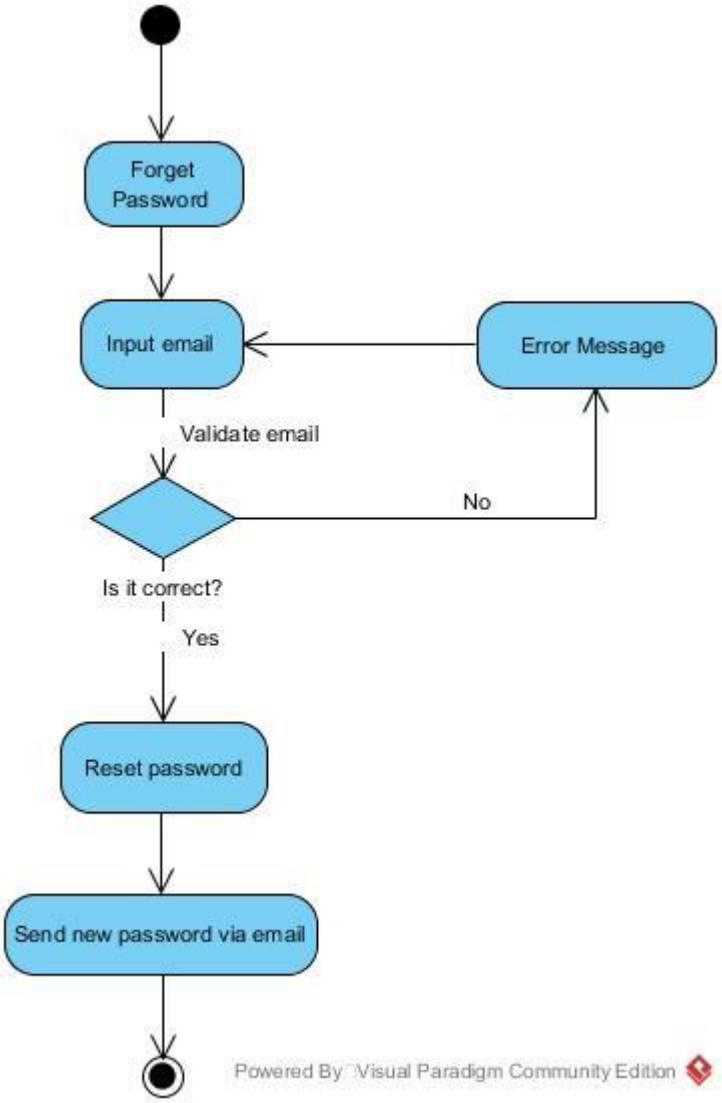


Figure 77: Forget password activity diagram

If an existing user forget his/her password, he/she can choose the forget password function. Once the user entered the e-mail address that was used for registration, the system will validate the email address. If the e-mail exists in the user database, a reset password e-mail will be sent. If not, the user will be prompted with error message to check their input e-mail address again.

- Create Campaign

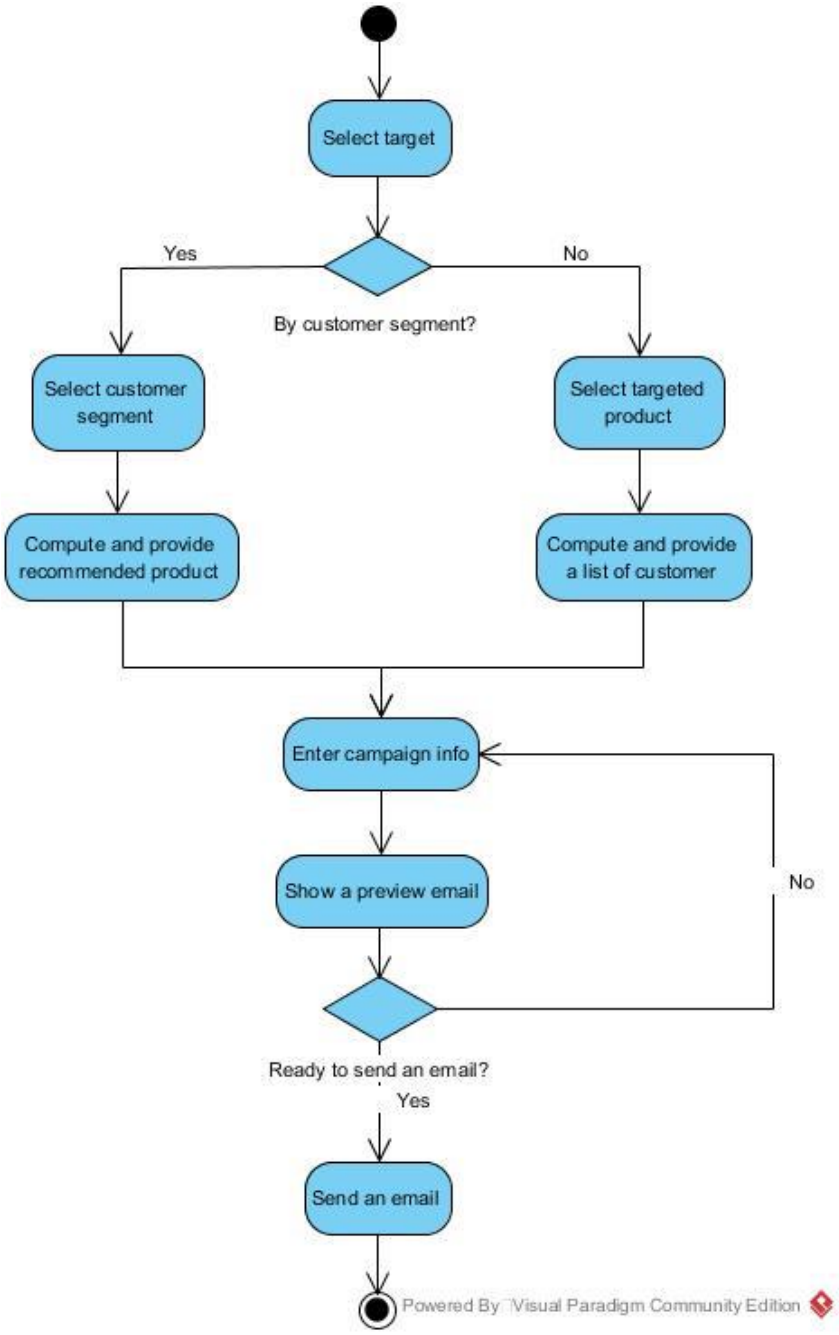


Figure 78: create campaign activity diagram

Once a user is logged in, user can choose to create an e-mail campaign to send to the customers. First, user can select target customer or target product. If the user selects customer segment, the system will provide the appropriate product. If the user selects a product, the system will provide appropriate customer. The user can then enter the campaign information and view the e-mail preview. If the user wants to edit the content, he/she can go back. If not, he/she can send the e-mail to the customers.

- Recommender Selection

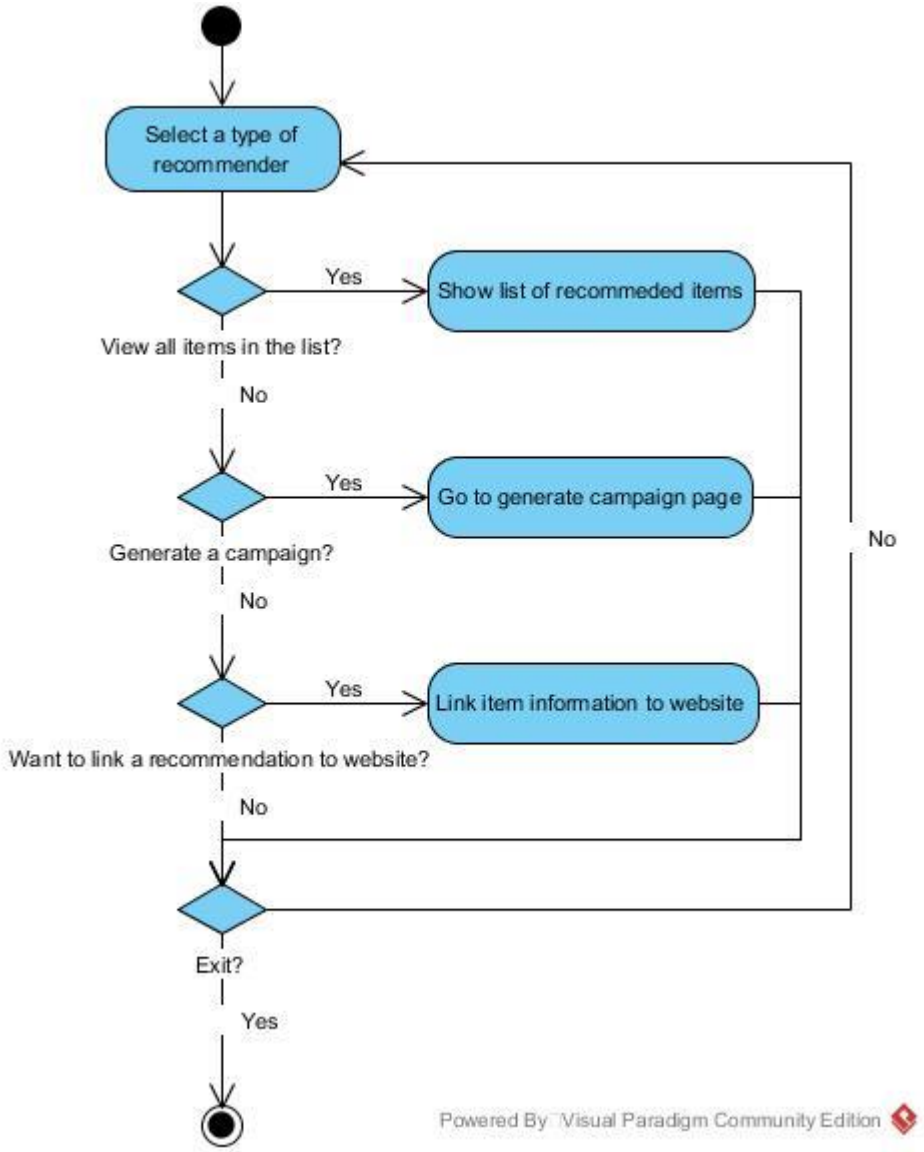


Figure 79: Recommender selection activity diagram

Logged in users can view the recommender that CAPP provided. On the recommender page, users will be able to view all the recommenders with the sample results and generate campaign based on the recommender that they like. They can also select a recommender to be integrated into their recommendation content in their e-commerce website.

Designing and implementing the web application user interface.

The development of CAPP web application is done mainly by php and html. The web content database is SQL database. The graphs and dashboards are design then presented with Chart JS and E-charts. The design of dashboards and actual graphical interface of the web application is discussed in Chapter 4. The web interface uses a bootstrap theme named Gentelella and modified with CSS to suit the design of the websites.

The data that is used in all of the analytics is the Dunnhumby dataset which is our temporary dataset. The dataset is kept in Microsoft Azure cloud sql database because it can handle this size of dataset easily. The data of users (log in information and company information) will also be kept in Microsoft Azure sql database for ease of access for the web application.

3.8 Migration to Microsoft Azure

3.8.1 Hosting a website on Microsoft Azure

Method 1: Creating an App Service

Step 1: Go to portal.azure.com and login with a Microsoft account. There is a free trial for free users but our project received a quota for developing on Azure for the imagine cup so we can use the cloud up to $150 per month. The homepage of Azure shows the dashboard, recent activities, and available resources.

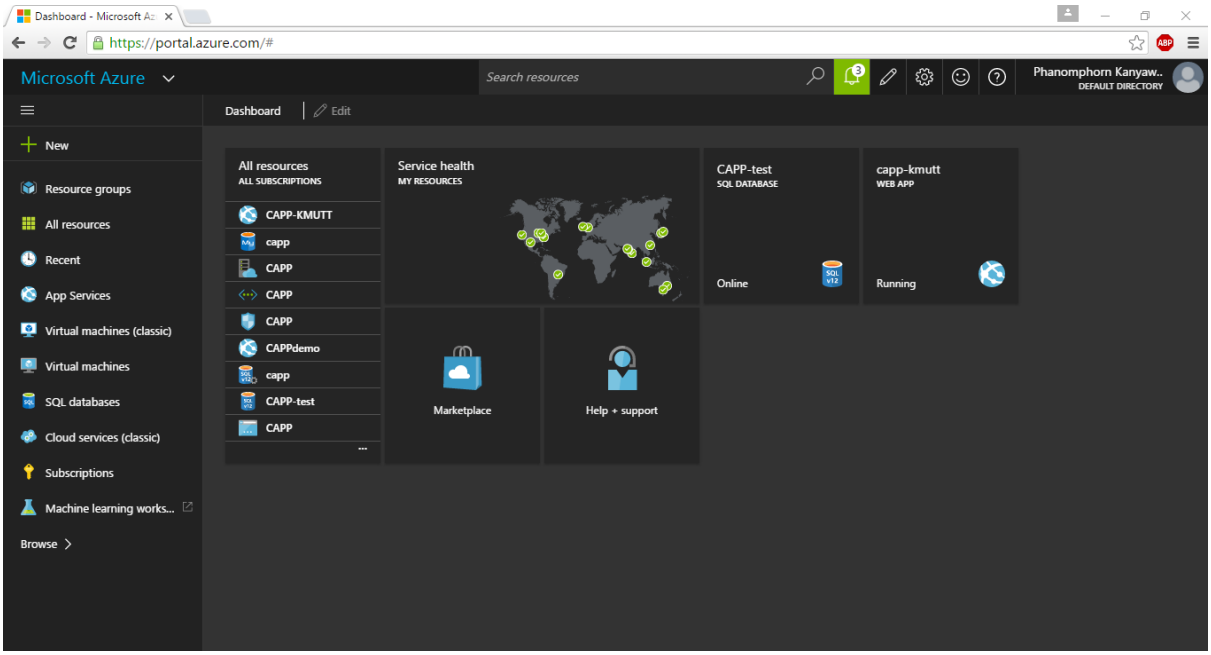


Figure 80: Home - Azure Dashboard

Step 2: Select App, add a new app service, enter information for the app service, and create a new application.
Step 3: Select the created application on the list of app services and select Settings to start setting.
Step 4: Select Deployment source on the Setting tab to choose the source code for deployment. Azure supports many existing shared storage such as GitHub, Dropbox, and OneDrive so we can easily selected the code in those storage and deploy it.
Step 5: After the deployment is successful, the website can be accessed.

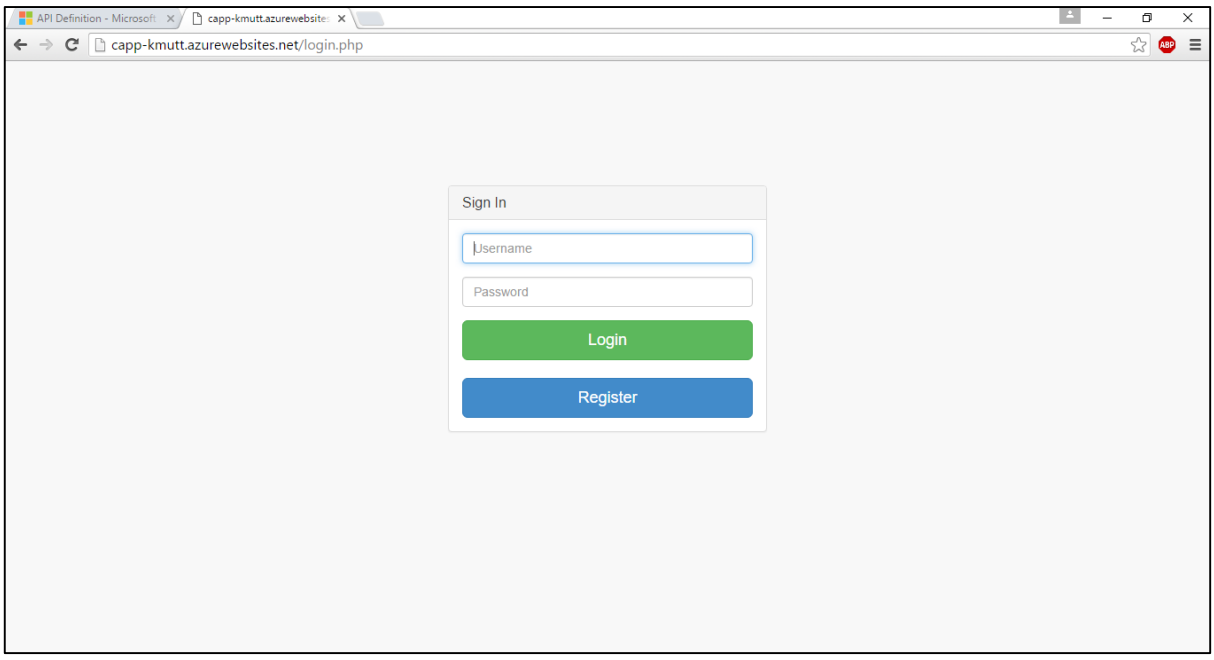


Figure 81: Successful deployment - Access website

Method 2: Creating a virtual machine with web server
Step 1: Go to Virtual machines and add a new virtual machine.
Step 2: Select a virtual machine. There are many OS to choose from, we will choose Ubuntu here for familiarity.
Step 3: After the creation, a setup will be shown. Enter the information of the virtual machine and wait for the deployment to finish.
Step 4: The virtual machine that we created does not have GUI so we use a secure shell program such as Putty to access it remotely and we can use it normally on the command line.

3.8.2 Creating a MySQL Database for web content.

Step 1: Search for MySQL providers in azure. We selected ClearDB.
Step 2: After the creation, enter the information and select a pricing plan. Then, wait for the deployment to be finished.
Step 3: Select a database management tool from ClearDB website, download and install it. We chose MySQL Workbench.

Step 4: Add the database to MySQL Workbench by specifying the Hostname, port, and username given in Azure.

Step 5: A database with the same name as the project will be available, We can now create databases and tables.

Step 6: We can now execute SQL query from the php code in our website. For example, when filling out the registration form, the user data can be saved into the user table in the database.

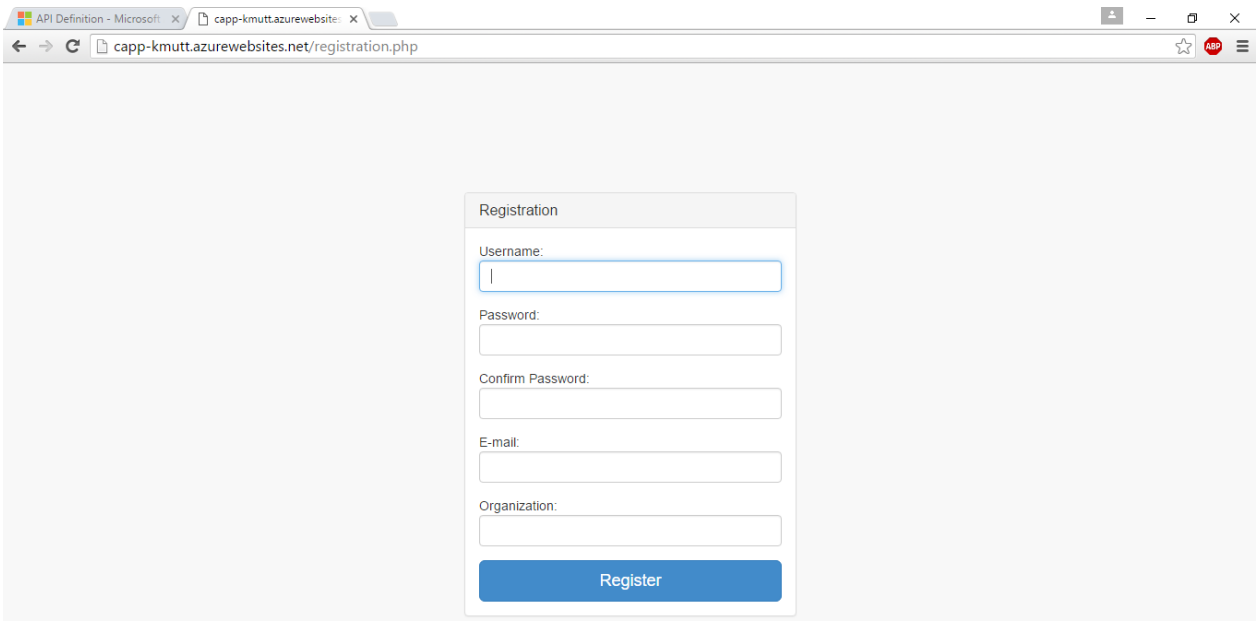


Figure 82: Sample registration form that require a database query

3.8.3 Email delivery configuration

CAPP will used SendGrid as an email delivery service for delivering marketing campaigns or promotions email to retailers' customer. In order to successfully send an email to customers, it need to be authenticated by SMTP server, otherwise an email will be clarified as a spam. The steps to acquire SendGrid SMTP server authentication are as follow:

Step 1: Go to 'Marketplace' in azure and search for 'SendGrid Email Delivery'
Step 2: Create a service

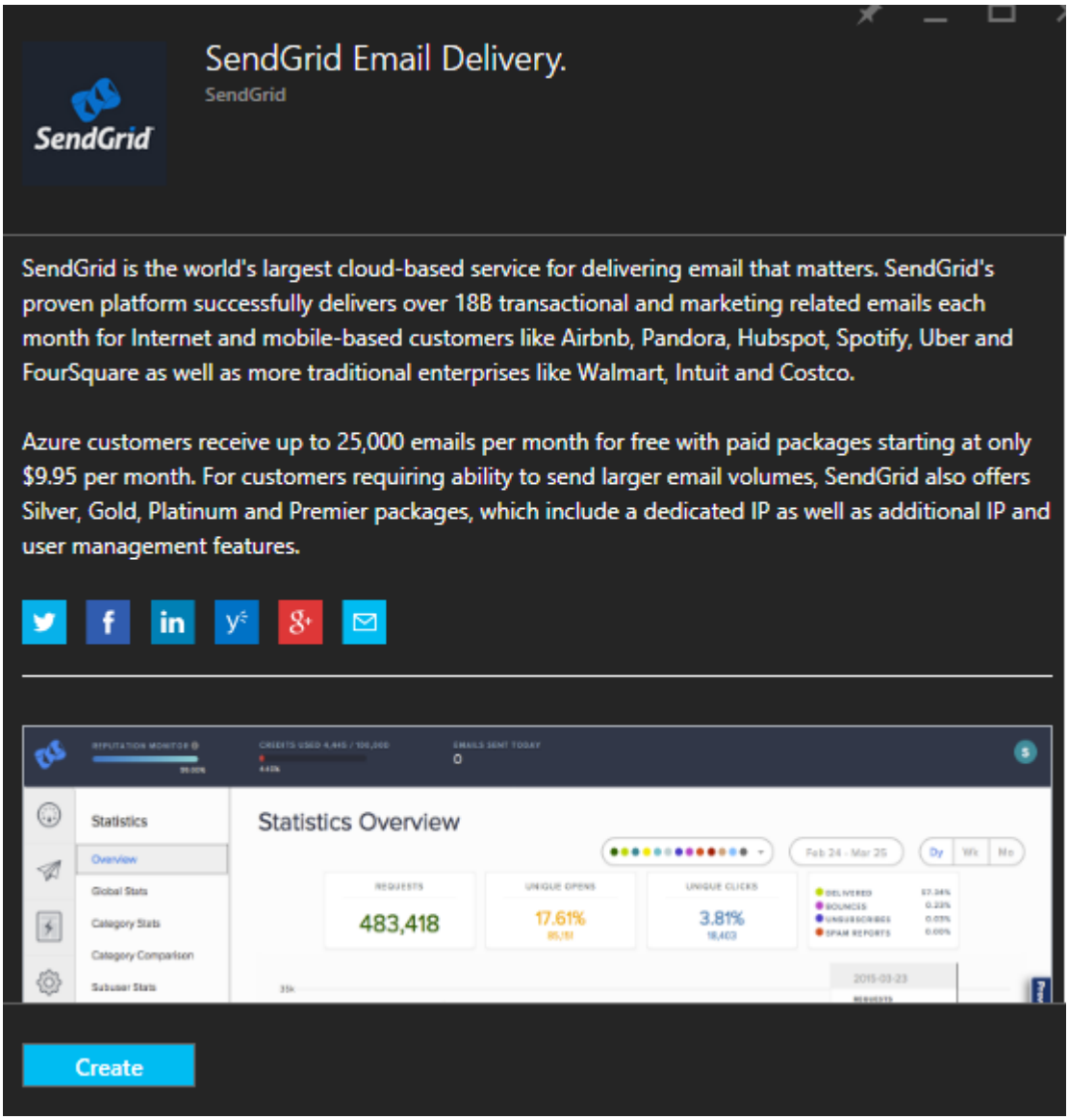


Figure 83: SendGrid information screen

Step 3: Go to SendGrid email service that was created
Step 4: From SendGrid setting panel, select 'Configuration' to view a username, password, and SMTP server

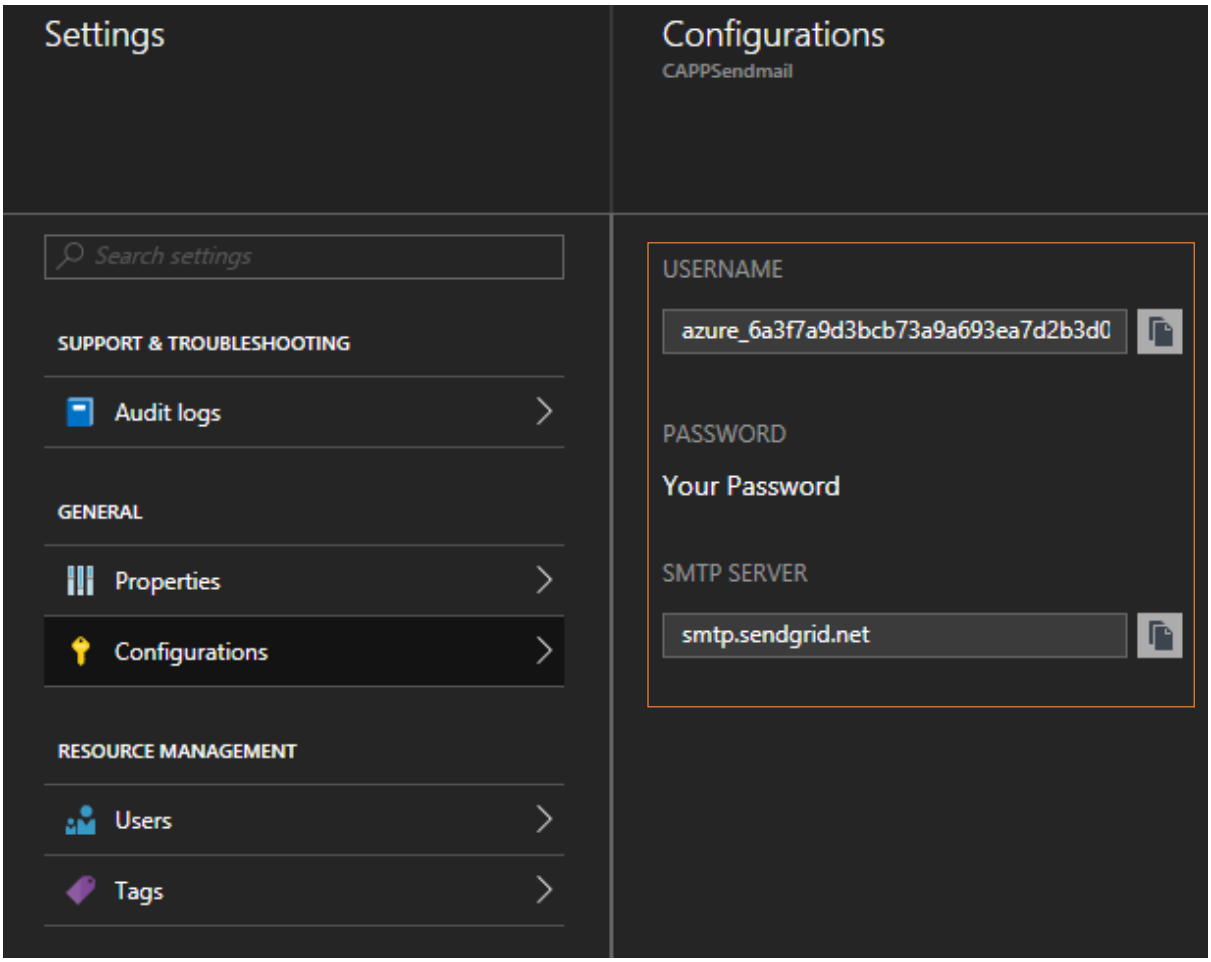


Figure 84: Sendgrid setting panel

Step 8: Use username and password acquire from Sendgrid setting panel to config in mail sending script.

```
$mail = new PHPMailer;

$mail->isSMTP();                                      // Set mailer to use SMTP
$mail->Host = 'smtp.sendgrid.net';                    // Specify main and backup SMTP servers
$mail->SMTPAuth = true;                               // Enable SMTP authentication
$mail->Username = 'azure_6a3f7a9d3bcb73a9a693ea7d2b3d00e8@azure.com';   // SMTP username
$mail->Password = '[redacted]';                       // SMTP password
$mail->SMTPSecure = 'tls';                            // Enable TLS encryption, `ssl` also accepted
$mail->Port = 587;                                    // TCP port to connect to
```

Figure 85: Mail sending script

Figure 86: Sample email sent from script

Step 9: Check Sendgrid dashboard to see the performance and number of sent emails.

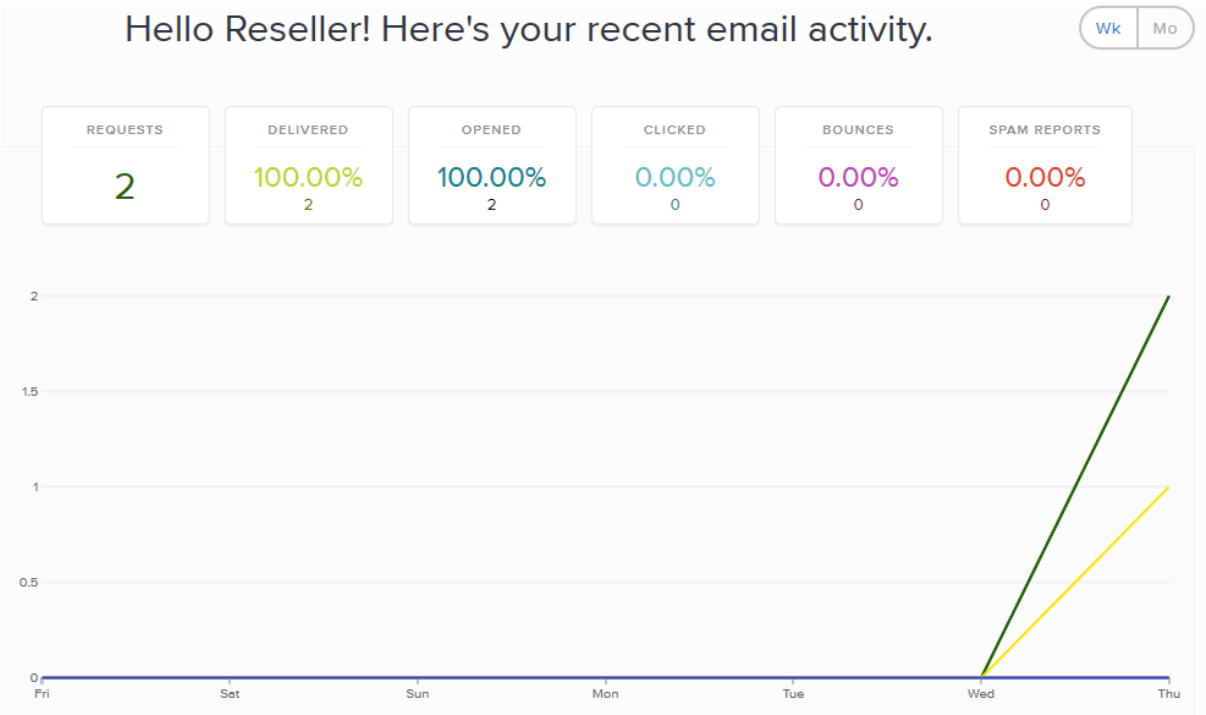


Figure 87: SendGrid analytics dashboard

Chapter 4
Results and Discussion

4.1 Recommender System

The recommender in CAPP can be divided into 2 main part of work. The first half in semester 1 was done together by the 2 teams to explore the use of machine learning models and libraries for data analytics. In the first semester, the main focus was mahout machine learning library and our team focused on user-based recommendation while the other team focused on item-based recommendation. The second half which is the second semester only the other team is working on the recommender system while our team work on GUI and web application development.

User-based recommendation is a type of collaborative filtering that focuses on the similarity among users (customers) while item-based recommendation focuses on the similarity of items purchased.

- Preprocessing

Preprocessing is the step to clean, filter the unnecessary data out and reformat them into an appropriated format. Before computation, the dataset was preprocessed into 3 columns; household key, product ID, and rating. The number of product purchase was used as rating.

Household key	Product ID	Number of purchase
1	820165	3
1	821815	1
1	821867	1
1	823721	1
1	823990	1

Figure 88: Sample preprocessed data

- User-based recommendation result

We used mahout in hadoop server at the laboratory to process the dataset. We randomly select 90 percent of the dataset as training set and another 10 percent was reserved for testing set.

From the transactional data, a list of recommended product will be generated for each household. The recommender can rate all the products and we can choose to view the top n recommendation of each household to use in the marketing module. The following is a sample of top 10 recommendation for household 1:

1. GROCERY: Instant breakfast
2. NUTRITION: Non-dairy beverages

3. NUTRITION: Dairy cheese
4. DRUG GM: Adult diaper
5. NUTRITION: Juice
6. DRUG GM: Infant formula
7. DRUG GM: Tobacco
8. DRUG GM: Adult diaper 2
9. DRUG GM: Cigarettes
10. NUTRITION: Soy beverage

Once the dataset is passed through the model, we will get the list of recommended items and the similarity matrix of the customers. The following figure shows the top 5 most similar customers of each customer and the recommended items list comes from their similar customers' transaction history.

User	Related User 1	Related User 2	Related User 3	Related User 4	Related User 5
3	159	1518	14	965	193
7	1349	1485	1774	1794	1015
9	2454	2257	1487	1566	2333
14	193	965	1061	2239	2327
16	1015	1534	1694	1774	2485
19	196	2142	804	1461	367
25	1860	388	1518	1498	356
26	2258	192	2468	2372	2356
31	314	475	1047	1076	2394
34	422	587	1323	1706	2132

Figure 89: sample top 5 user similarity result

- Evaluation and Model Comparison

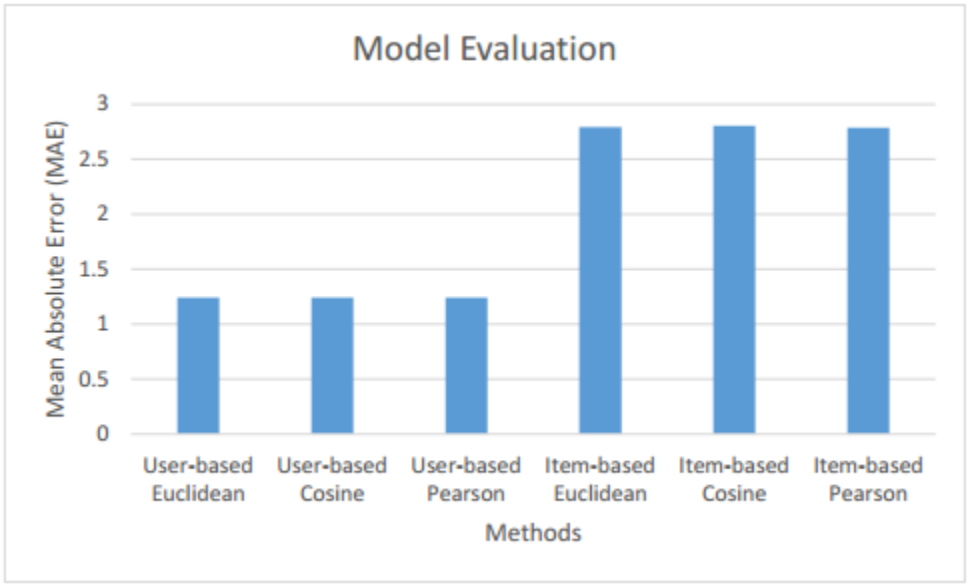


Figure 90: Graph showing the comparison of each model using Mean Absolute Error (MAE)

The user-based recommender yielded an overall better result than item-based recommender on the Mean Absolute Error (MAE) evaluation. It is done by measuring the MAE of each similarity algorithms as explain in Chapter 2 namely Euclidean distance, Cosine distance, and Pearson correlation.

However, the evaluation of a model cannot be justify with only statistical value. We need to implement it and collect data in order to really evaluate a recommender. Therefore, a majority voting algorithm will be used to decide the recommended item so that the best product will be chosen. The voting algorithms will be implement by the other team for semester 2.

4.2 Actual web prototype

4.2.1 Landing page

Figure 91: CAPP logo

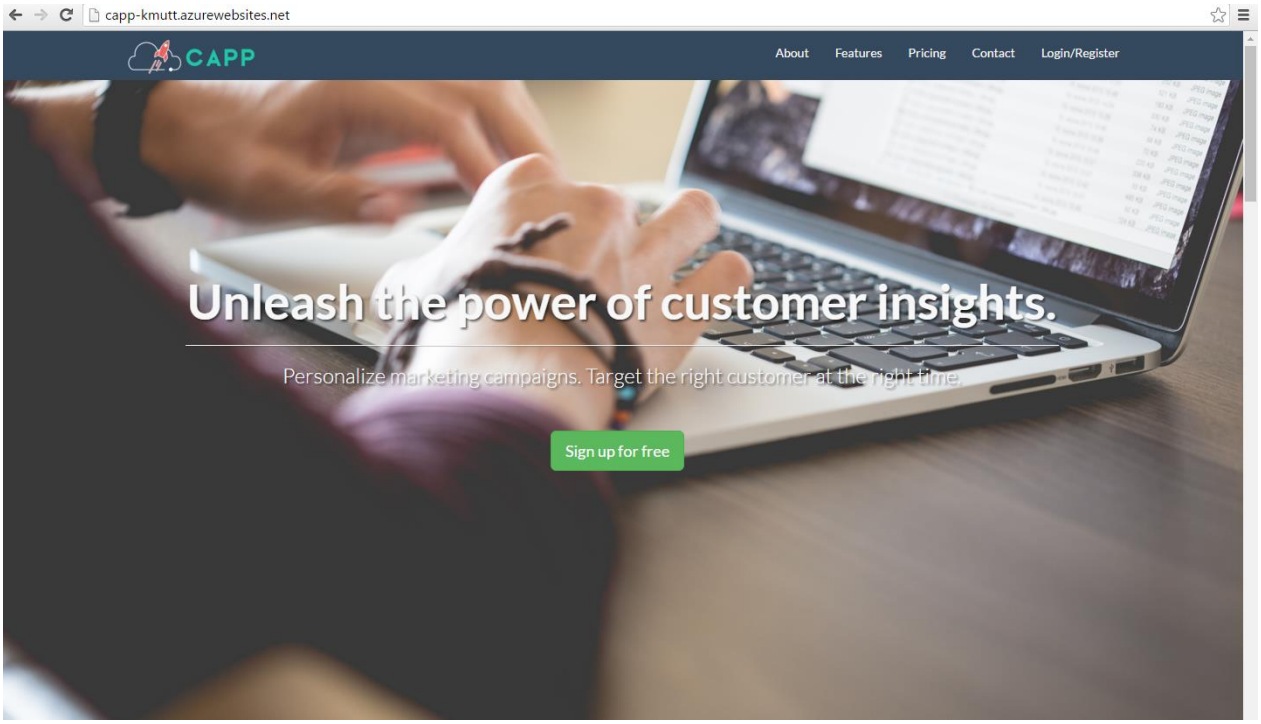


Figure 92: CAPP landing page

The above figure shows the landing page of our prototype, CAPP. The purpose of the landing page is to attract visitors' interest as much as possible in order to increase the chance of selling. So, we designed the first page with a beautiful background picture and catchy slogan. The navigation bar on the top of the page shows the project logo and headings for each content section which are:

- Features: Show the features of CAPP
- Pricing: Shows the pricing tier of CAPP
- Contact: Shows the section for contacting staffs and developers
- Login / Register: Show the login and register screen to access the dashboard

4.2.2 Analytics Dashboard

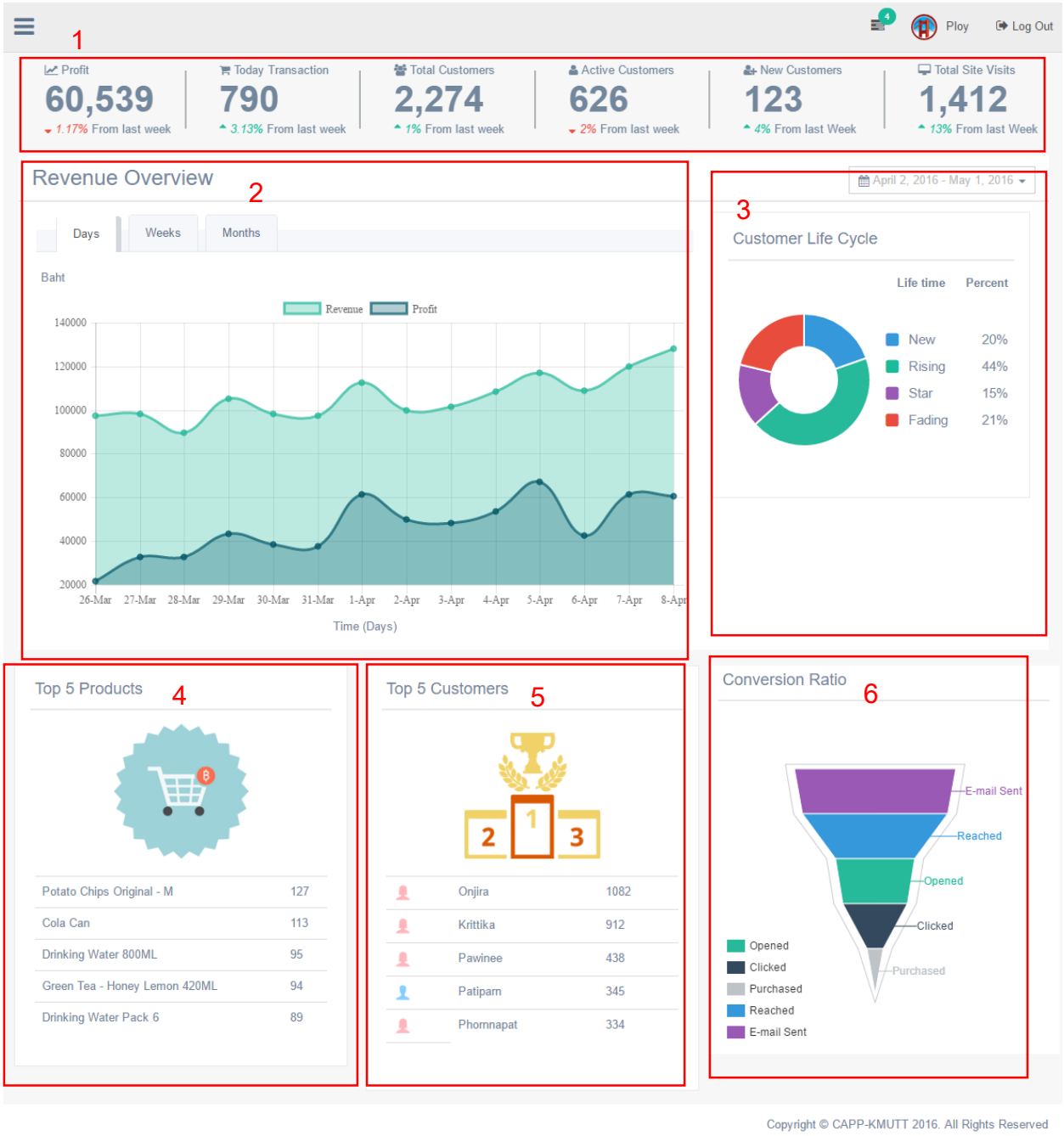


Figure 93: Analytics dashboard homepage content

1. Summary tab
    - Shows the summary data of the business. Its show the total profit, transaction, customer with comparison to the last week
2. Revenue Overview
    - Shows the business revenue and profits. Users can select the time period to show the data, or select to show only profit or revenue

3. Customer Lift Cycle
    - Shows the percentage of customer in different life cycle, which are new customer, raising customer, star customer, and fading customer.

4. Top 5 Product
    - Shows the 5 most popular items from customer transaction.

5. Top 5 Customer
    - Shows the 5 customer who spent the most for the business.

6. Conversion Rate
    - Shows the ratio rate from sent email to actual purchase.

2.4.3 Email Campaign

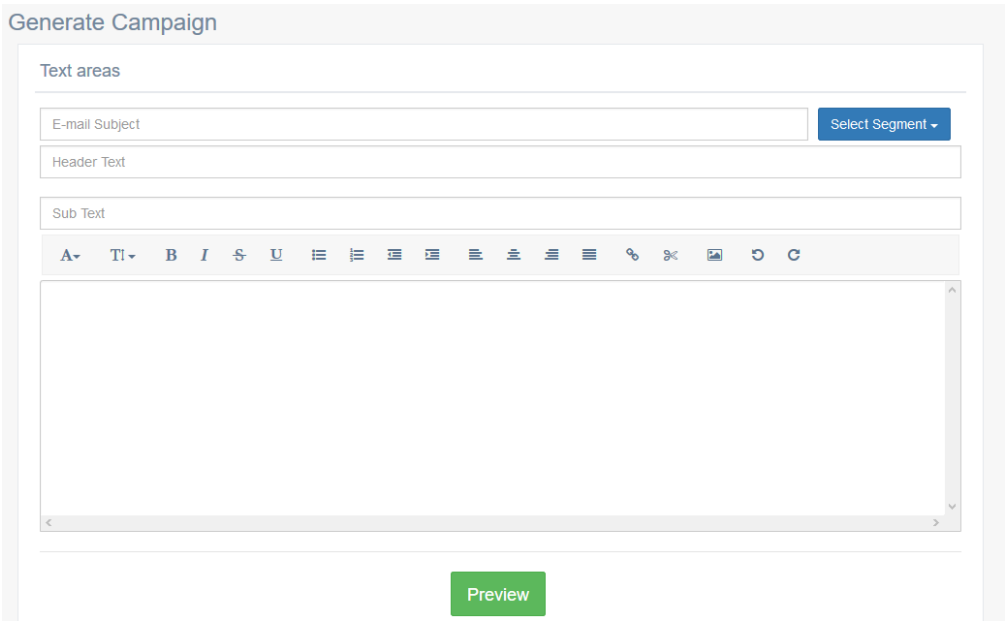


Figure 94: Email template

This page will let user compose their marketing email before sending to their customers. User can select customer in each specific segment as a recipient of the email.

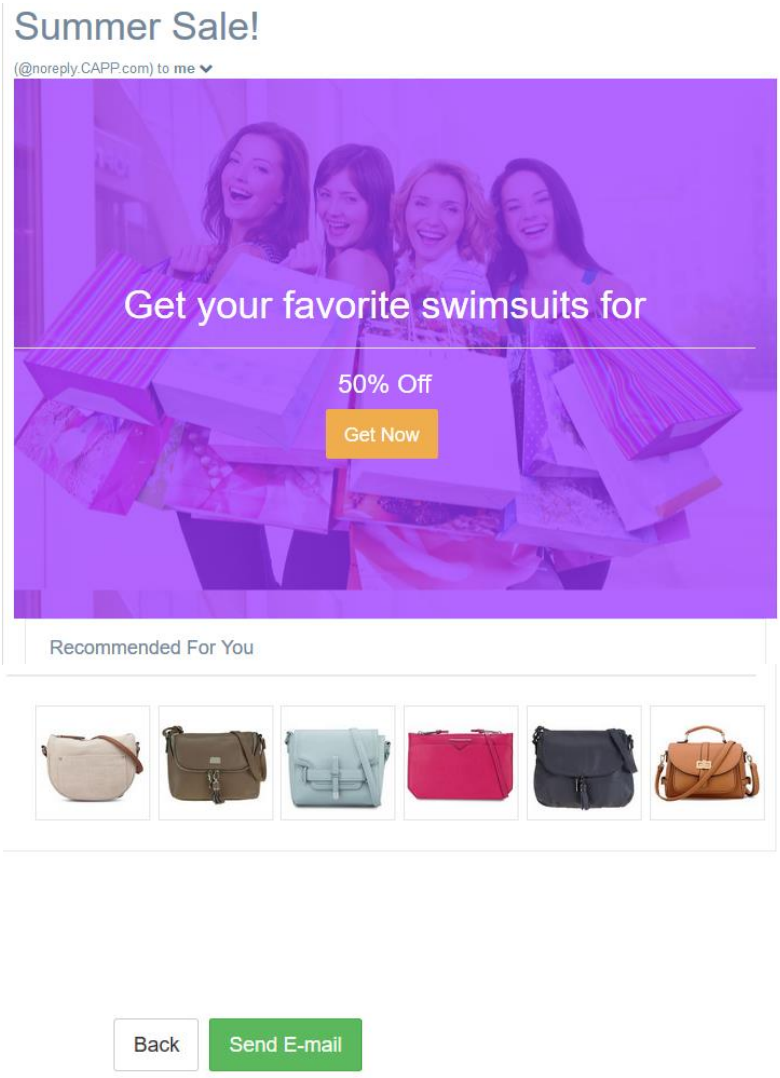


Figure 95: Sample mail content

2.4.4 Recommendation

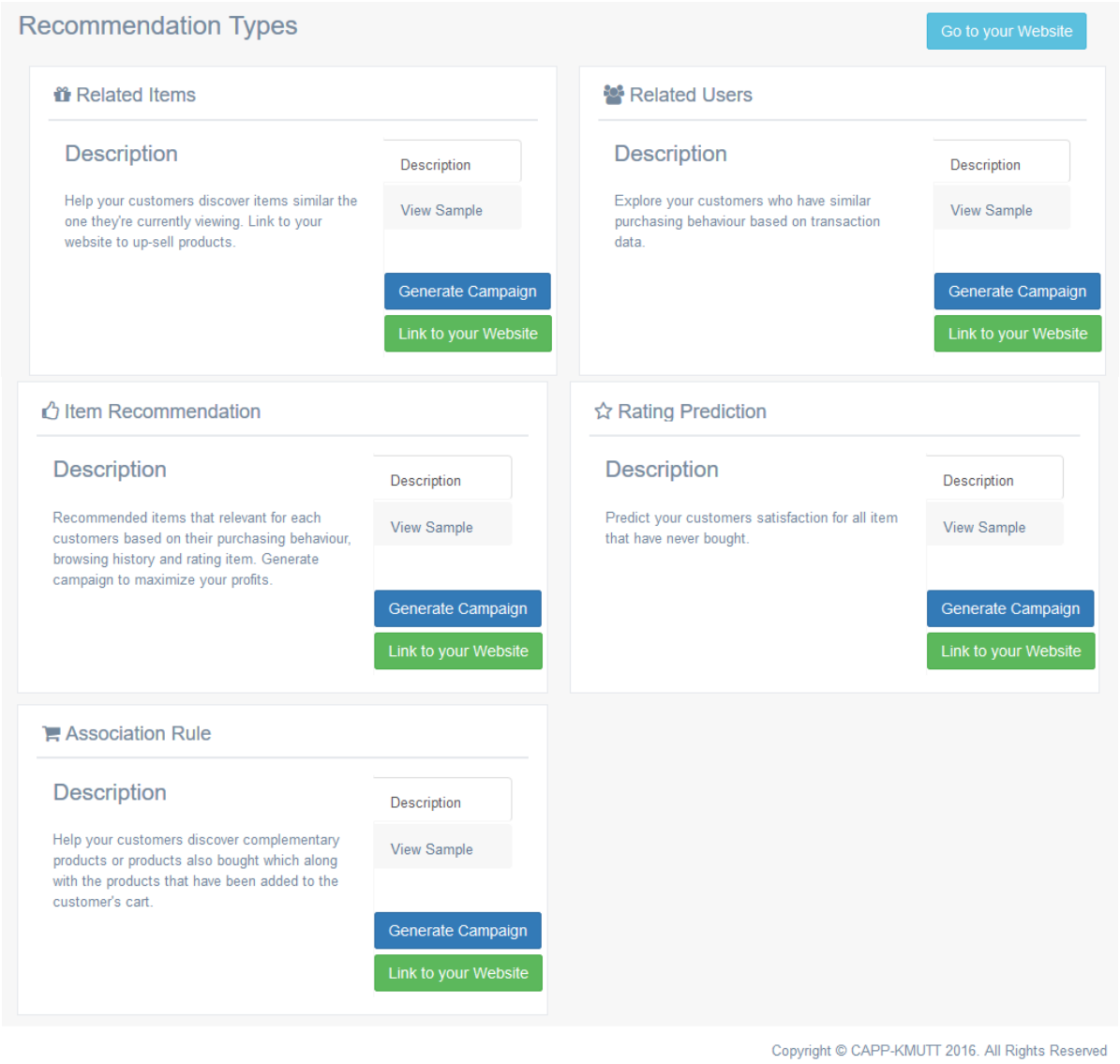


Figure 96: Recommendation page

Recommendation page shows the type of recommendation along with the list of recommended items for each type. The type of recommendation are as follow:

- Related items: Shows the items that related with each other
- Related users: Shows the user who has similar purchasing behavior
- Item recommendation: Shows the recommended item for each user.
- Rating Prediction: Predicts customer satisfaction for each item.
- Association Rule: Shows the items that usually bought together.

Items Recommendation Result Table

Show 10 entries Search:

User	Item 1	Item 2	Item 3	Item 4	Item 5
3	BANANAS	SOFT DRINKS 20PK&24PK CAN CARB	ADULT CEREAL	SHREDDED CHEESE	FLUID MILK WHITE ONLY
7	BATH OILS	ADULT CEREAL	BANANAS	BABY FOOD - BEGINNER	SHREDDED CHEESE
9	CIGARETTES	TEA SWEETENED	SNACK CAKE - MULTI PACK	COFFEE FILTERS	NON DAIRY CREAMER: DRY
14	SOFT DRINKS 20PK&24PK CAN CARB	ADULT CEREAL	FLUID MILK WHITE ONLY	SHREDDED CHEESE	SFT DRNK 2 LITER BTL CARB INCL
16	ADULT CEREAL	SALAD BAR OTHER	SHREDDED CHEESE	SOFT DRINKS 12/18&15PK CAN CAR	FLUID MILK WHITE ONLY
19	ADULT CEREAL	BANANAS	SOFT DRINKS 20PK&24PK CAN CARB	SHREDDED CHEESE	SFT DRNK 2 LITER BTL CARB INCL

Figure 97: Sample list of recommended item for each user

4.3 Migration to Microsoft Azure Cloud.

4.3.1 Comparison between using App service and virtual machine for website on Azure.

Both app service and virtual machine have their own advantages and disadvantages. By using App service for the website, we gain full access to the Azure APIs and other tools such as Azure machine learning. However, we are limited to what Azure provide. By using a virtual machine to run a web server, we can customize everything we want. We can also install many additional packages that Azure does not provide but we will not be able to connect to other services that Azure and its partner provide easily (we can only work in the virtual machine which is disconnected from the outside Azure environment.).

For now, we have chosen to work with the App service because we want to use Azure machine learning as our selling point in the Microsoft Imagine Cup. We want to utilize the Azure cloud's ability as much as possible because the competition is held by Microsoft and they require every team to use their product. If we had use the virtual machine and install other packages, it would defeat the purpose of utilizing the Azure cloud. However, we may consider the virtual machine at a later stage because it is very promising to be able to install the tools that we are familiar with and work with them instead.

Chapter 5
Conclusion

5.1 Project Accomplishment

The progress of the project is coming along really well. We already do most part of an analytics engine on our own clusters at first. However, we decided to migrate to Microsoft Azure cloud instead because of the performance and for competition purpose. So, the workload has been increase but it should be done by the end of the semester. The table below shows the accomplishment of the project of our project and remaining tasks:

Component	Completion
Analytic Engine	Completed
• User based recommender	Completed
• Item base recommender	Completed
• Ensemble	Completed
• Migration to MS Azure cloud	Completed
Designing Reports	Completed
• Report and dashboard types	Completed
• Report and dashboard interfaces	Completed
Create E-mail Recommender	Completed
• Design E-mailing interface	Completed
• Auto generate E-mail	Completed
Database	Completed
• SQL database for web	Completed
• MongoDB for storing data	Completed
• Migration to MS Azure cloud	Completed
Web Interface	Completed
• Login/Registration	Completed
• About	Completed
• Features	Completed
• Pricing	Completed
• CSS	Completed
• Dashboards	Completed
• Generate E-mail Recommender	Completed
• Hosted on MS Azure cloud	Completed
Migration to MS Azure cloud	Completed
• Analytic Engine	Completed
• Database	Completed
• Web	Completed

5.2 Discussions

5.2.1 Change of this project

The project scope has been modified to clearly explain our work. The details were also modified since our project has passed the first round of Microsoft Imagine Cup. We modified our project to better suit the retailers use. Our implementation were also modified significantly because we need to migrate to Microsoft Azure Cloud and use Azure Machine Learning Service. Therefore, there has been a lot of delay in our working process since we have to accommodate the competition.

5.2.2 Problems

The dataset that we were supposed to get from an actual anonymous retailer in Thailand did not arrived. So, we need to find available dataset online and we got the customer data from Dunnhumby.com and will be analyzing transactional data with customer and product data.

Secondly, everything crashes because the dataset was large. Even though it was just a temporary dataset, it was still big enough to crash Tableau, RapidMiner, and Microsoft Excel in our PC. Therefore, we need to move to Hadoop earlier than expected. At first, we planned to explore the data on our PC and continue the implementation on Hadoop in a later phase but our PC cannot even handle the visualization that has many joined tables. Fortunately, the Cloudera Hadoop Server configuration was finished early so we get to try using it for the first time.

Later, when we change to Hadoop and finish our analytic engine, another change must be made since we participated in a Microsoft competition, we need to use the Azure cloud. So, we have to migrate everything again. However, it will be easier than the first time that we did it since we now know how to handle large amount of data.

As we have spent a lot of time searching for dataset and trying out different dataset, there is not much time left to explore the dataset that we newly got in dept. Therefore, we need to speed up in the next phase and work during the vacation to compensate for the time in the first half, to catch up to our planned schedule.

During the vacation, we made changes to our project in order to compete in the Microsoft Imagine Cup competition. Now we are gradually applying the changes to our project and migrating to use Microsoft Azure cloud.

5.2.3 Knowledge gained and Future work

We learned that different techniques used for machine learning yield different results and by preparing data differently, we could gain more knowledge from the same data source.

Calibrating the parameters to provide good results is the real challenge for us. The parameters and input attributes are variables that we can change quite freely and they affect the efficiency of the model greatly.

By working on this project, we gained the skills necessary for data mining and data analytics application. In the near future we are going to apply the knowledge and parts of this project for real world problem. We are going to collaborate with big data experience center and work with a company to pilot a new project based on CAPP and get to address real needs of the clients

References

[1] Machine learning. (n.d.). In *Whatis.com*. Retrieved from http://whatis.techtarget.com/definition/machine-learning

[2] Wayne. (2011, November 29). *Part VI - Crafting Analytical Models*. Retrieved from http://www.b-eye-network.com/blogs/eckerson/archives/2011/11/part_vi_-_craft.php

[3] (n.d.). *Classification*. Retrieved from http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm

[4] (n.d.). *Machine Learning 101: General Concepts*. Retrieved from http://www.astroml.org/sklearn_tutorial/general_concepts.html

[5] Analytical Modeling. (n.d.). In *Open Energy Information*. Retrieved from http://en.openei.org/wiki/Definition:Analytical_Modeling

[6] (n.d.). *Classification modeling*. Retrieved from http://b-course.hiit.fi/obc/whatiscl.html

[7] (n.d.). *Regression*. Retrieved from http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/regress.htm

[8] Gondaliya, A. (2014, May 22). *Regularization implementation in R: Bias and Variance diagnosis.* Retrieved from http://pingax.com/regularization-implementation-r/

[9] Frost, J. (2015, September 3). *The Danger of Overfitting Regression Models.* Retrieved from http://blog.minitab.com/blog/adventures-in-statistics/the-danger-of-overfitting-regression-models

[10] Rossant, C. (n.d.). *Introduction to Machine Learning in Python with scikit-learn.* Retrieved from http://ipython-books.github.io/featured-04/

[11] (n.d.). *Line of Best Fit (Least Square Method).* Retrieved from http://hotmath.com/hotmath_help/topics/line-of-best-fit.html

[12] Schneider , J. (1997, February 7). *Cross Validation*. Retrieved from https://www.cs.cmu.edu/~schneide/tut5/node42.html

[13] Fortmann-roe, S. (May 2012). *Accurately Measuring Model Prediction Error*. Retrieved from http://scott.fortmann-roe.com/docs/MeasuringError.html.

[14] Kaewchinporn, C. (2013, August 22). *Example of decision tree induction by ID3*. Retrieved from http://scriptslines.com/blog/example-generate-decision-tree-by-id3/

[15] (n.d.). *Model-based recommendation systems*. Retrieved from http://www.cs.carleton.edu /cs_comps/0607/recommend/recommender/modelbased.html

[16] (n.d.). *Amazon.com Recommendation*. Retrieved from https://www.cs.umd.edu/ ~samir/498/Amazon-Recommendations.pdf

[17] Bari, A., Chaouchi, M., and Jung, T. (n.d.). *How to Use Item-Based Collaborative Filters in Predictive Analysis*. Retrieved from http://www.dummies.com/how-to/content/ how-to-use-itembased-collaborative-filters-in-pred.html

[18] Bari, A., Chaouchi, M., and Jung, T. (n.d.). *Basics of User-Based Collaborative Filters in Predictive Analysis*. Retrieved from http://www.dummies.com/how-to/content/basics-of-userbased-collaborative-filters-in-predi.html

[19] Shimodaira, H. (2015, January 20). *Similarity and recommender systems*. Retrieved from http://www.inf.ed.ac.uk/teaching/courses/inf2b/learnnotes/inf2b-learn-note02-2up.pdf

[20] S.Perone, C. (2013, September 12). *Machine Learning :: Cosine Similarity for Vector Space Models (Part III)*. Retrieved from http://blog.christianperone.com/2013/09/machine-learning-cosine-similarity-for-vector-space-models-part-iii/

[21] (n.d.). *What is Hadoop?*. Retrieved from http://www.sas.com/en_us/insights/big-data/ hadoop.html

[22] (n.d.). *Microsoft Azure Cloud Computing Platform and Services.* Retrieved from https://azure.microsoft.com/en-us/?b=16.01

[23] (n.d.). *Business Intelligence and Servicres | Tableu Software.* Retrieved from http://www.tableau.com/

[24] (n.d.). *RapidMiner - #1 Open Source Predictive Analytics Platforms.* Retrieved from https://rapidminer.com/

[25] (n.d.) *Doing Data Science*. Retrieved from http://semanticommunity.info/Data_Science /Doing_Data_Science#The_Data_Science_Process

[26] Beechler, D. (2014, September 30). 7 Technology Trends Transforming Consumer Communication. Retrieved from http://www.exacttarget.com/blog/uk/7-technology-trends-transforming-consumer-communication/

[27] (n.d.). *RapidMiner - #1 Open Source Predictive Analytics Platforms.* Retrieved from https://rapidminer.com/

[28] (n.d.). *RichRelevance - Omnichannel Ecommerce Recommendations and Content Optimization.* Retrieved from http://www.richrelevance.com

[29] (n.d.). *HubSpot - Inbound Marketing & Sales Software.* Retrieved from http://www.hubspot.com/

[29] (n.d.). *Cloudera.* Retrieved from http://www.cloudera.com/

[30] Dunnhumby. (2014). *THE COMPLETE JOURNEY USER GUIDE*. Retrieved from http://dunnhumby.skybzz.com/dunnhumby_The-Complete-Journey.zip

[31] HTML. (n.d.). Retrieved May 28, 2016 from https://en.wikipedia.org/wiki/HTML

[32] PHP. (n.d.). Retrieved May 28, 2016 from https://en.wikipedia.org/wiki/PHP

[33] What is PHP?. (n.d.). Retrieved May 28, 2016 from http://php.net/manual/en/intro-whatis.php

[34] Stephen, C. (n.d.). *What Is JavaScript?.* Retrieved from http://javascript.about.com/od/reference/p/javascript.htm

[35] riteshgaur, fscholz, et al. (Dec 12, 2015). *CSS developer guide.* Retrieved from https://developer.mozilla.org/en-US/docs/Web/Guide/CSS

[36] MySQL System Properties. (n.d.). Retrieved from http://db-engines.com/en/system/MySQL

[37] PHP MySQL Database. (n.d.). Retrieved from http://www.w3schools.com/php/php_mysql_intro.asp

[38] Bootstrap 3 Tutorial. (n.d.). Retrieved from http://www.w3schools.com/bootstrap/

[39] Gentelella - Bootstrap Admin Template by Colorlib. (n.d.). Retrieved from https://colorlib.com/polygon/gentelella/

[40] Aurelio, R. (January 07, 2015). *Creating Beautiful Charts with Chart.js*. Retrieved from https://www.sitepoint.com/creating-beautiful-charts-chart-js/

[41] Chart.js. (n.d.). Retrieved from http://www.chartjs.org/

[42] ECharts. (n.d.). Retrieved from https://ecomfe.github.io/echarts/index-en.html

[43] Mean absolute error. (n.d.). Retrieved May 28, 2016 from https://en.wikipedia.org/wiki/Mean_absolute_error

Appendices

Dunnhumby dataset tables and relationship:

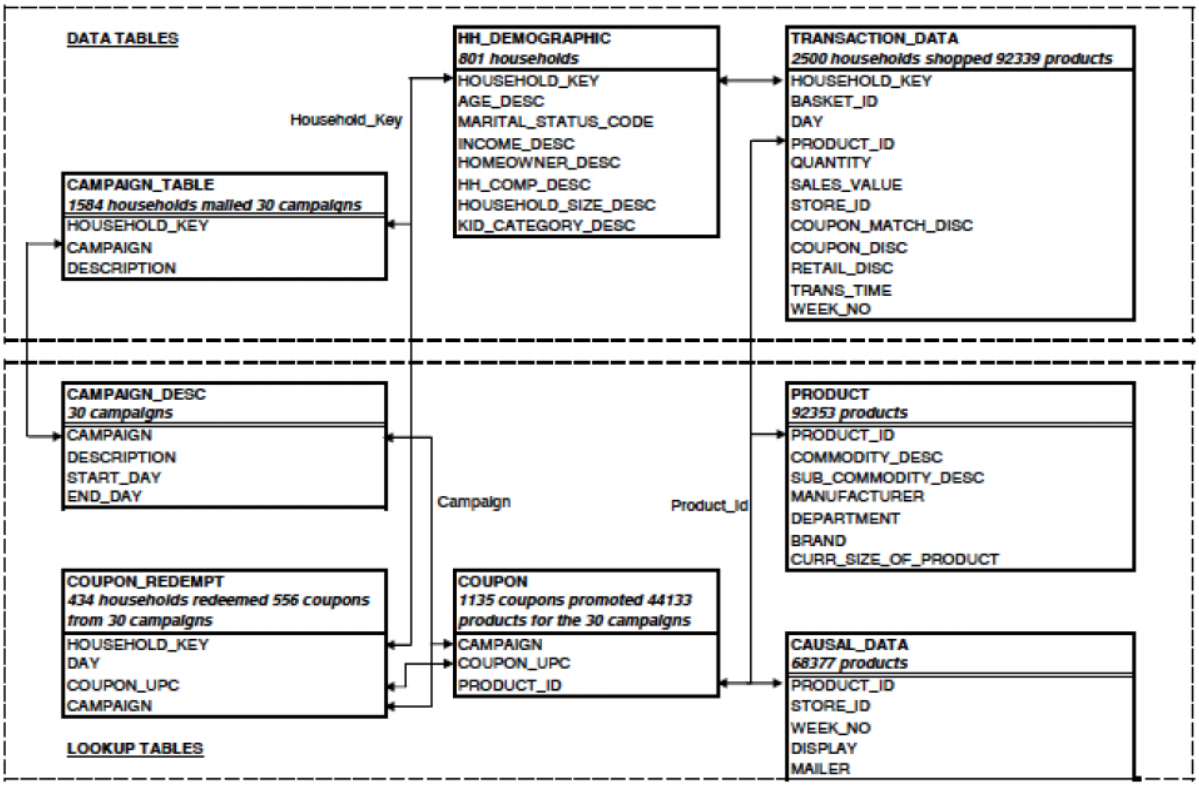


*Figure **98:** ER Diagram showing the relationship of tables extracted from Dunnhumby user guide*

hh_demographic

This table contains demographic information for a portion of households. Due to nature of the data, the demographic information is not available for all households.

Variable	Description
HOUSEHOLD_KEY	Uniquely identifies each household
AGE_DESC	Estimated age range
MARITAL_STATUS_CODE	Marital Status (A - Married, B- Single, U - Unknown)
INCOME_DESC	Household income
HOMEOWNER_DESC	Homeowner, renter, etc.
HH_COMP_DESC	Household composition
HOUSEHOLD_SIZE_DESC	Size of household up to 5+
KID_CATEGORY_DESC	Number of children present up to 3+

AGE_DESC	MARITAL_STATUS	INCOME_DESC	HOMEOWNER_DESC	HH_COMP_DESC	HOUSEHOLD_SIZE_	KID_CATEGORY_DESC	household_key
65+	A	35-49K	Homeowner	2 Adults No Kids	2	None/Unknown	1
45-54	A	50-74K	Homeowner	2 Adults No Kids	2	None/Unknown	7
25-34	U	25-34K	Unknown	2 Adults Kids	3	1	8
25-34	U	75-99K	Homeowner	2 Adults Kids	4	2	13
45-54	B	50-74K	Homeowner	Single Female	1	None/Unknown	16
65+	B	Under 15K	Homeowner	2 Adults No Kids	2	None/Unknown	17

Figure 99: Customer demographic table and description

transaction_data

This table contains all products purchased by households within this study. Each line found in this table is essentially the same line that would be found on a store receipt.

Variable	Description
HOUSEHOLD_KEY	Uniquely identifies each household
BASKET_ID	Uniquely identifies a purchase occasion
DAY	Day when transaction occurred
PRODUCT_ID	Uniquely identifies each product
QUANTITY	Number of the products purchased during the trip
SALES_VALUE	Amount of dollars retailer receives from sale
STORE_ID	Identifies unique stores
COUPON_MATCH_DISC	Discount applied due to retailer's match of manufacturer coupon
COUPON_DISC	Discount applied due to manufacturer coupon
RETAIL_DISC	Discount applied due to retailer's loyalty card program
TRANS_TIME	Time of day when the transaction occurred
WEEK_NO	Week of the transaction. Ranges 1 - 102

household_key	BASKET_ID	DAY	PRODUCT_ID	QUANTITY	SALES_VALUE	STORE_ID	RETAIL_DISC	TRANS_TIME	WEEK_NO	COUPON_DISC	COUPON_MATCH_DISC
2375	26984851472	1	1004906	1	1.39	364	-0.6	1631	1	0	0
2375	26984851472	1	1033142	1	0.82	364	0	1631	1	0	0
2375	26984851472	1	1036325	1	0.99	364	-0.3	1631	1	0	0
2375	26984851472	1	1082185	1	1.21	364	0	1631	1	0	0
2375	26984851472	1	8160430	1	1.5	364	-0.39	1631	1	0	0

Figure 100: Transaction table and description

campaign_table

This table lists the campaigns received by each household in the study. Each household received a different set of campaigns.

Variable	Description
HOUSEHOLD_KEY	Uniquely identifies each household
CAMPAIGN	Uniquely identifies each campaign. Ranges 1 - 30
DESCRIPTION	Type of campaign (TypeA, TypeB or TypeC)

DESCRIPTION	household_key	CAMPAIGN
TypeA	17	26
TypeA	27	26
TypeA	212	26
TypeA	208	26
TypeA	192	26

Figure 101: Campaign table and description

campaign_desc
This table gives the length of time for which a campaign runs. So, any coupons received as part of a campaign are valid within the dates contained in this table.

Variable	Description
CAMPAIGN	Uniquely identifies each campaign. Ranges 1 - 30
DESCRIPTION	Type of campaign (TypeA, TypeB or TypeC)
START_DAY	Start date of campaign
END_DAY	End date of campaign

DESCRIPTION	CAMPAIGN	START_DAY	END_DAY
TypeB	24	659	719
TypeC	15	547	708
TypeB	25	659	691
TypeC	20	615	685
TypeB	23	646	684

Figure 102: Campaign description table and description

product
This table contains information on each product sold such as type of product, national or private label and a brand identifier.

Variable	Description
PRODUCT_ID	Number that uniquely identifies each product
DEPARTMENT	Groups similar products together
COMMODITY_DESC	Groups similar products together at a lower level
SUB_COMMODITY_DESC	Groups similar products together at the lowest level
MANUFACTURER	Code that links products with same manufacturer together
BRAND	Indicates Private or National label brand
CURR_SIZE_OF_PRODUCT	Indicates package size (not available for all products)

PRODUCT_ID	MANUFACTURER	DEPARTMENT	BRAND	COMMODITY_DESC	SUB_COMMODITY_DESC	CURR_SIZE_OF_PRODUCT
25671	2	GROCERY	National	FRZN ICE	ICE - CRUSHED/CUBED	22 LB
26081	2	MISC. TRANS.	National	NO COMMODITY DESCRIPTION	NO SUBCOMMODITY DESCRIPTION	
26093	69	PASTRY	Private	BREAD	BREAD:ITALIAN/FRENCH	
26190	69	GROCERY	Private	FRUIT - SHELF STABLE	APPLE SAUCE	50 OZ
26355	69	GROCERY	Private	COOKIES/CONES	SPECIALTY COOKIES	14 OZ

Figure 103: Product table and description

coupon

This table lists all the coupons sent to customers as part of a campaign, as well as the products for which each coupon is redeemable. Some coupons are redeemable for multiple products. One example is a coupon for any private label frozen vegetable. There are a large number of products where this coupon could be redeemed.

For campaign TypeA, this table provides the pool of possible coupons. Each customer participating in a TypeA campaign received 16 coupons out of the pool. The 16 coupons were selected based on the customer's prior purchase behavior. Identifying the specific 16 coupons that each customer received is outside the scope of this database.

For campaign TypeB and TypeC, all customers participating in a campaign receives all coupons pertaining to that campaign.

Variable	Description
CAMPAIGN	Uniquely identifies each campaign. Ranges 1 - 30
COUPON_UPC	Uniquely identifies each coupon (unique to household and campaign)
PRODUCT_ID	Uniquely identifies each product

COUPON_UPC	PRODUCT_ID	CAMPAIGN
10000089061	27160	4
10000089064	27754	9
10000089073	28897	12
51800009050	28919	28
52100000076	28929	25

Figure 104: Coupon table and description

coupon_redempt

This table identifies the coupons that each household redeemed.

Variable	Description
HOUSEHOLD_KEY	Uniquely identifies each household
DAY	Day when transaction occurred
COUPON_UPC	Uniquely identifies each coupon (unique to household and campaign)
CAMPAIGN	Uniquely identifies each campaign

household_key	DAY	COUPON_UPC	CAMPAIGN
1	421	10000085364	8
1	421	51700010076	8
1	427	54200000033	8
1	597	10000085476	18
1	597	54200029176	18

Figure 105: Coupon redemption table and description

causal_data

This table signifies whether a given product was featured in the weekly mailer or was part of an in-store display (other than regular product placement).

Variable	Description
PRODUCT_ID	Uniquely identifies each product
STORE_ID	Identifies unique stores
WEEK_NO	Week of the transaction
DISPLAY	Display location (see below)
MAILER	Mailer location (see below)

Field	Contents
DISPLAY	0 - Not on Display 1 - Store Front 2 - Store Rear 3 - Front End Cap 4 - Mid-Aisle End Cap 5 - Rear End Cap 6 - Side-Aisle End Cap 7 - In-Aisle 9 - Secondary Location Display A - In-Shelf
MAILER	0 - Not on ad A - Interior page feature C - Interior page line item D - Front page feature F - Back page feature H - Wrap front feature J - Wrap interior coupon L - Wrap back feature P - Interior page coupon X - Free on interior page Z - Free on front page, back page or wrap

PRODUCT_ID	STORE_ID	WEEK_NO	display	mailer
26190	286	70	0	A
26190	288	70	0	A
26190	289	70	0	A
26190	292	70	0	A
26190	293	70	0	A

Figure 106: Causal data table and description