Describing financial product holding status through visualizing big financial data

Jiangsu Du

August 17, 2017

MSc in High Performance Computing with Data Science

The University of Edinburgh

2017

Abstract

With the development of digital technology within the financial services industry, companies are increasing their focus on being able to understand the data that is being collected in order to better understanding their customer bases. This UK based financial company seeks to gain a fuller understanding of its customers from this data, therefore this project will aim to utilize data analysis techniques to better understand the personal financial product holding conditions of its existing customers. The results will be concluded in a dynamic report, and the report will be uploaded on to an online server to assist a wider range of staff in the company to answer simple specific questions in this area. Also all analytical code developed as part of this project will be packaged for later use.

Contents

1 Introduction **1**

1.1 Background 1

1.2 Project goal 1

1.2.1 Danger Zone 1

1.2.2 Goal description 2

1.3 Research methodology 3

1.4 Report structure 4

2 Context review **6**

2.1 Teradata[1] 6

2.2 SQL[2] 6

2.3 Data visualization platform 7

2.4 Google Maps JavaScript API 9

2.5 Product overview 10

2.5.1 Current Account product 11

2.5.2 Saving Account product 11

2.5.3 Credit Card product 12

2.5.4 Loan product 12

2.5.5 Mortgage product 12

2.6 ETL vs. ELT 13

2.7 CRISP-DM 13

3 Business & Data Understanding **16**

3.1 Data Description 16

3.1.1 Data Source 16

3.1.2 Data volume 18

3.1.3 Data quality 18

3.2 Code review . . . 18
3.3 Business analysis and data availability . . . 19
3.3.1 Basic information . . . 19
3.3.2 Personal features . . . 19
3.3.3 Geographical distribution . . . 20

4 Data preparation & processing **23**
4.1 Data preparation . . . 23
4.2 Data cleaning . . . 23
4.3 Data filtering . . . 24
4.4 Data aggregation . . . 25
4.4.1 Time period distribution . . . 26
4.4.2 Customer holding relationship . . . 28
4.4.3 Geographical distribution . . . 30
4.5 Performance . . . 30

5 Data visualization & Analysis **31**
5.1 Basic Aggregation charts . . . 31
5.2 Holding Frequency charts . . . 33
5.3 Geographical Distribution map . . . 36

6 Evaluation **40**
6.1 Evaluating Environments and Work Practice . . . 40
6.2 Evaluating Visual Data Analysis and Reasoning . . . 41
6.3 Evaluation User Experience . . . 42

7 Conclusions **43**
7.1 Review . . . 43
7.2 Goal accomplishment . . . 43
7.3 Future work . . . 44

A SQL Syntax example **45**

B SQL Code example **48**
B.1 SQL code file 1 . . . 48
B.2 SQL code file 2 . . . 50
B.3 SQL code file 3 . . . 52

List of Tables

4.1 Fields list 25
4.2 The structure of analysis 1 26
4.3 The data structure of analysis 2 (sum) 28
4.4 The data structure of analysis 2(average) 28
4.5 The structure of analysis 3 30

5.1 The data structure of geographical analysis 37

List of Figures

1.1 Project workflow 5

2.1 Multiple platform report[3] 8
2.2 Data exploration 8
2.3 Pure heatmap display[5] 9
2.4 Fusion table with dummy data 10
2.5 Pure fusion table layer display[6] 10
2.6 CRISP-DM Process Diagram [7] 15

3.1 UK map 21

4.1 Data cleaning process[10] 24
4.2 The flow chart of merging the table for the first analysis 27
4.3 The flow chart of merging the table for the second analysis 29

5.1 Screen shot of the basic aggregation charts 32
5.2 Pie chart 33
5.3 Holding Frequency example 1 34
5.4 Holding Frequency example 2 34
5.5 Holding Frequency example 3 35
5.6 Holding Frequency example 4 35
5.7 Holding Frequency example 5 36
5.8 Holding Frequency example 6 36
5.9 Overall Screenshot of the interface 38
5.10 The comparison between different radius and another colour schema 39
5.11 The dialog which renders corresponding information 39

Acknowledgements

I would like to thank my supervisors, Terry Sloan and Greg Milne, for their immense help and useful advice during the entirety of this project. As a non-native English speaker, it is truly a tough process for me to learn the project and get used to the work environment. However, they are extremely patient and considerate to help me finally finish the project. Moreover, Ally Hume provided me some valuable suggestions in preparation step.

To the company, thank you for providing the data and work environment necessary for this project. It has been a great pleasure to work with the team in the company and be able to utilise their expertise and advice.

To EPCC, I love the experience of studying here for a whole year. The staff here are all very friendly and help.

Finally, I would also like to thank my family for supporting me throughout not only this project, but the entire time of my studies.

Chapter 1

Introduction

1.1 Background

With the development of digital technology, financial services are becoming increasingly convenient. More and more financial companies have moved their services to the Internet, thereby creating a large amount of financial data. In addition to convenience, digital technology can also make a wider range of customized financial services possible. Digitalized information can also be stored for the long-term and easily reused. Thus, it is meaningful to utilize this data by using data analysis technology.

This project is an industry attached project between the Edinburgh Parallel Computing Center (EPCC), part of The University of Edinburgh, and a major UK based financial services company. The company pays great attention to digital technology and applies digital technology to almost all the products it provides. With many years' accumulation, it has established a rich and comprehensive data warehouse which contains systematic information about its products. To utilize this data better, one of the tasks is to get an understanding about the personal financial products that their customers hold.

1.2 Project goal

1.2.1 Danger Zone

In this project, data analysis technology is applied to a real problem. The design process between domain experts and data analysts to understand each other is one of the most dangerous and time-consuming steps. Generally, data analysts do not have knowledge in a specific area and domain experts don't know how the data analysis technology can benefit their work. On the one hand, it can be hard for domain experts to express their views accurately within limited times of communication. On the other hand, because of the lack of the relevant knowledge, data analysts usually understand domain experts wrongly and get the incorrect analysis as a result. Even if data analysts and domain experts can understand each other correctly, many policy limitations in big organizations especially financial organizations can block the process and delay the project.

The entire process of understanding the project goal is extremely tough. In the beginning, the title of this project was "Visualizing data for business drivers" from which little can be understood, because of confidential reasons, after first communication with the person in this company, the main direction of this project was still unclear. Until finishing the project preparation course, only the developing environment was understood correctly and the title changed to "Optimizing and customizing data visualization for non-analytical stakeholders". In other words, this author misunderstood the topic of this project after three months' effort. Then the employment background screening was finished and the company representative described the whole project, but the practical meaning of the project was still quite vague since a significant level of domain specific knowledge is needed to help understand its purpose. Finally, after another three weeks' financial knowledge study, a clear idea of this project was slowly formed.

Overall, this process reflected the importance and difficulty of communication between domain expertise and data analyst in data science.

1.2.2 Goal description

There are many different products provided by this major UK based financial services company. For personal products, there are saving, credit cards, loans, mortgages, insurance, investments, and current accounts. For high net worth customers, it also provides a private service. Moreover, every product claimed above can be divided into different sub-types further. Also, there are business products for organizations and companies.

This project will focus on personal product holding conditions of the customers within this company. In this company, a customer can hold more than one product. With all the customers' information, we can get the holding condition of these products which can help internal staff make decisions. Personal products include saving products, mortgage products, credit card products, loan products, and current products. Products differ in necessity and some products have a closer relationships. Here is a list of some examples to help understand the importance of the project:

- On average, every savings account customer will hold a savings account product and a current account product together. The transactions between these two products are normally very frequent, so most people will choose to open these two products in the same organization to improve the convenience. If the holding numbers of these differ a lot, this indicates that there maybe some pain points. The staff can investigate these and provide solutions to improve its products.
- Whilst every mortgage account customer will hold a mortgage account, they might not necessarily hold a current account as these two products are rarely considered together. If many customers only hold a mortgage product, it may reflect that a competitor has a more attractive current account products. However, if the rate of holding these two products is higher than the estimated rate, this might suggest that the current account is a market leading product.

Thus, understanding the financial product holding condition of customers can help staff improve their products. However, processing financial big data is very complicated. Lots of issues need to be overcome. In this project, issues include:

- Financial data is very important for a financial company as it can be utilized by criminals or business competitors. Disclosure of confidential data is very likely to cause serious damage to this company. Thus, there are strict policies to protect data safety, thereby even causing delay in doing the project.
- Big data technology is expected to be combined with business logistic and professional knowledge. In fact, only in this way can a meaningful report be produced. The first step of this project is to study the relevant financial knowledge and fully understand the financial products of the company.
- Relational database is responsible for the data storage and management in the company. Before collecting the data from raw dataset, the usage of the database should be learned.
- SQL is used to process data in the relational database. It is necessary to master it and use it to prepare data for the visualization.
- The records of different products are stored in many tables and the structure of each table differs a lot. It is worthwhile to carefully consider the method to merge, filter, and simplify these tables into a clear and analyze-prepared table.
- Products such as current accounts have many different types with different benefits associated with them. The personal product holding condition can be grouped by the product type (Current account, Saving account, Loan, Mortgage, Credit Card) at a high level, however it can also be divided into more categories at a lower product level. For instance, credit cards can differ in whether they can help customers earn cash back or not. Moreover, there are many fields that can be displayed. It is difficult to describe the whole dataset, and only the important features should be chosen and reflected.
- The way to visualize the informative data and make it easier for people to understand the holding condition clearly needs careful consideration.

So, it is almost impossible to use conventional methods to analyze the raw dataset. Scientific and modern methodology must be used to support the project. From another direction, the goal of the project can be thought as an implementation of data analysis technology and the problem is how to apply the technology to financial domain.

To sum up, the goal of the project is to use data analysis techniques to describe and display the financial product holding conditions of customers who bank with this UK based financial company from several valuable directions in a comprehensive way.

1.3 Research methodology

To make the financial product holding condition visualized, a Data Science approach to problem solving will be applied. This project will follow the typical Data Science project life cycle - CRISP-DM as the main part of the research methodology. The project will involve business & data understanding, data preparation & process, data visualization & Analysis and Evaluation. The general process of the project will be:

- Understanding the business and data held within the financial company and choosing how to describe the product holding condition.
- Defining required data.
- Filtering, simplifying, and merging relevant data from raw tables to new tables.
- To create the visualization of the selected data.

Evaluating and repeating the above four steps to achieve a good and accurate product holding report. Details of the entire process of the project is introduced in Figure 1.1.

1.4 Report structure

This chapter attempts to introduce the reader to the general background of introducing the problem along with detailed description of the goal. It also covered the high-level details of the research methodology to be applied throughout this project. Chapter 2 is about a review of the general background in which this project was conducted. It will describe the data warehouse and visualization tools used in the project as well as the products waiting to be analyzed. It will also briefly include the Data Science process used. Chapter 3 provides the necessary information about the business and the data. Also, it will decide what information should be included in the report. Then in chapter 4, it introduces how this data is processed. Chapter 5 displays details of visualization. Chapter 6 evaluates the report. Finally, chapter 7 concludes, discusses the results and suggests future work.

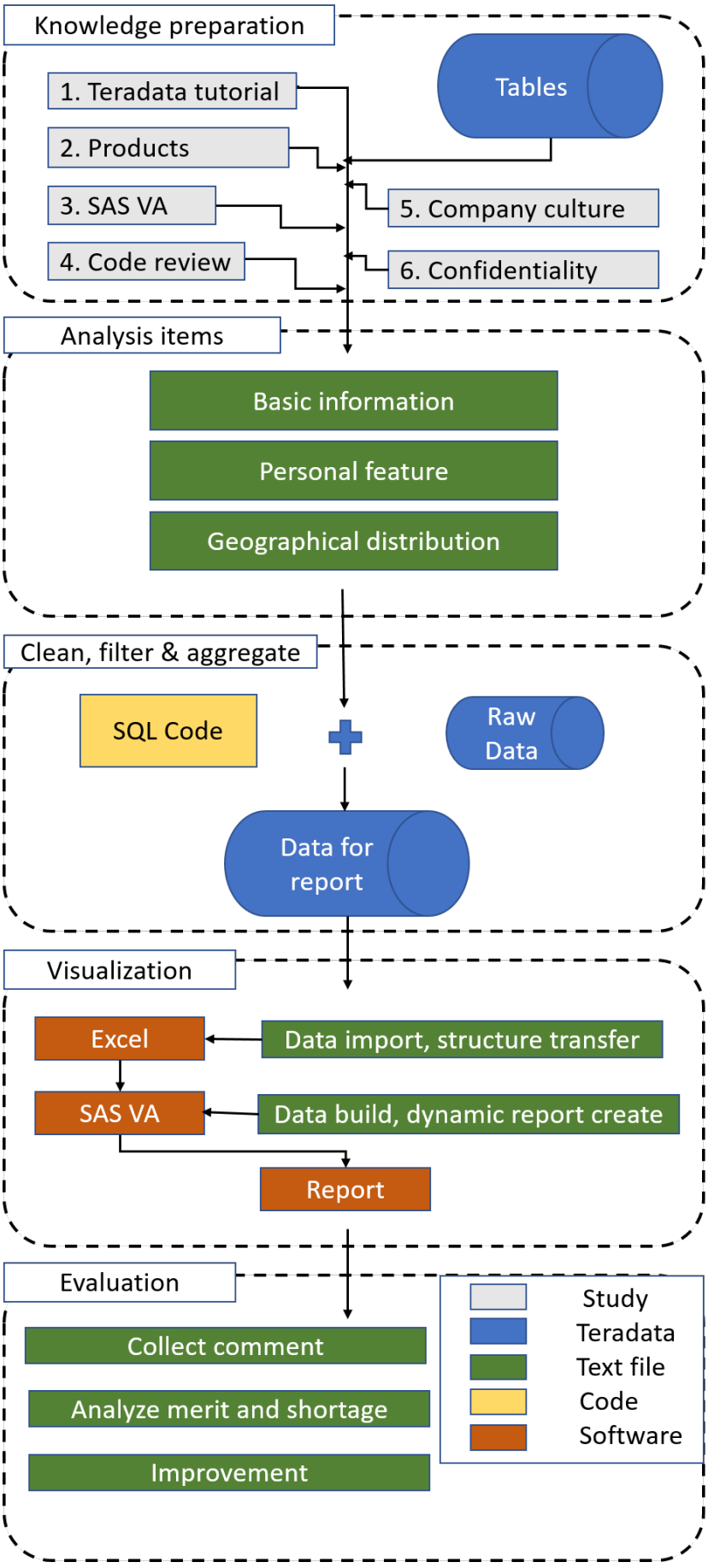

Figure 1.1: Project workflow

Chapter 2

Context review

2.1 Teradata[1]

With the development of digital technology, the speed of creating data becomes increasingly fast. To store and manage big data, the concept of data warehousing is introduced. A data warehouse is a federated repository which can store and manage all the data created by an enterprise's various businesses. A repository can be physical or logical. So, the data warehouse is an important concept of managing data.

The Teradata corporation provides database-related products and service. It developed a popular relational database management system which is also called Teradata. Teradata is one of the world's largest commercial databases. Teradata has a long history of providing solution to enterprise data warehouse. Its relational database management system is also referred to as a " Data Warehouse system". With great parallelism and scalability, Teradata allows you to start small with a single node and grow large with many nodes through linear expandability. It can divide a huge repository of data into smaller tasks and allocate them to many individual processors to be executed concurrently.

In order to handle billions of rows of data and up to and beyond a terabyte of data, a solution called YNET technology is created. It is the interconnect that allows hundreds of individual processors to share the same bandwidth. It can be understood as a sophisticated communication bus.

Teradata provides solutions to manage big financial data in the company. In this project, all the data sources will be collected from a system supported by Teradata corporation.

2.2 SQL[2]

In this project, SQL is going to be utilized as a programming language to select and manipulate relevant data in relational database to complete the process of preparing data.

SQL stands for Structured Query Language. It is specially designed for communicating with relational database and users can manage data stored in databases by using it. SQL has been widely used in popular database management systems. Comparing with the old read/write

APIs, SQL has many obvious advantages. Firstly, it is designed following the concept of accessing multiple records with one single command. Secondly, programmers don't need to provide the detailed information of reaching a record.

The SQL has been form into a standard in the American National StAndards Institute (ANSI) in 1986 and the International Organization for Standardization (ISO) in 1987, but implementations of SQL in different database management systems have slight difference. So, it is essential to get familiar with the SQL in Teradata. The standard SQL commands include "Select", "Insert", "Update", "Delete", "Create", and "Drop", which can be used to accomplish almost everything that one needs to do with a database. Apart from these fundamental commands, advanced commands such as "JOIN", "WHERE", "ORDER BY", "COUNT", "SUM", "GROUP BY", and "IN" are also used in the code for this project. The detail of these commands can be found in Appendix A.

2.3 Data visualization platform

The choice of visualization tool is mainly based on the computer system which is provided by the company. Because of confidentiality, all the intermediate data can only be accessed within the company's internal network. Upon inspection, it became apparent that the company could not provide the author with access to a typical analytics development environment such as R or Python. So, integrated visualization tools will be used in the visualization step. Excel and SAS VA are chosen as potential choices. Actually, these integrated tools usually perform more robust and speedy than writing new code.

Microsoft Excel is a spreadsheet software which is a part of Microsoft office family. It includes storage, calculation, graph tools, pivot tables, and it can also be extended by using a specific programming language. Meanwhile, it has its own file formats and the newer versions support many external file formats including CSV, XML, and DBF. Excel can also support high performance computing when combined with other Microsoft software. It is a powerful software tool and its power is only partly utilized in this project.

In this company, there is a distributed implementation of SAS Visual Analytics. SAS Visual Analytics is developed by SAS Institute and it aims at helping staff in different domains explore valuable trend and information from big data. The company cooperates with SAS Institute and they place SAS VA into a distributed cluster. The platform is an easy-to-use, web-based product that you can use to explore and view data, interact with and create reports, and display reports on a mobile device or on the web (Figure 2.1). You can explore your data by using interactive visualization such as charts, histograms, and tables. As displayed in SAS VA support page[4], it empowers organizations to explore huge volumes of data very quickly to identify patterns and trends and to identify opportunities for further analysis.

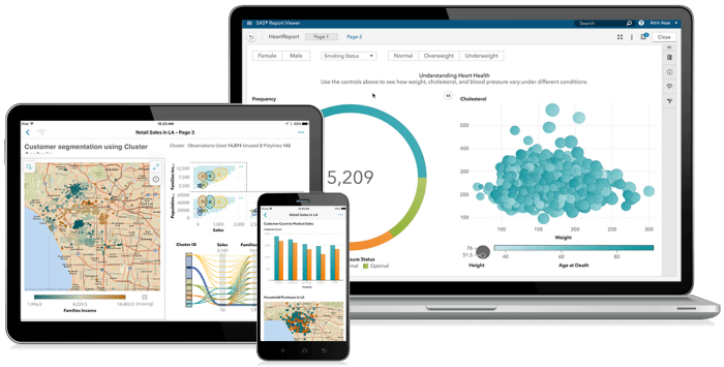

Figure 2.1: Multiple platform report[3]

It mainly provides three functions: data build, data exploration, and report build. In terms of data build, users can connect SAS VA to multiple data sources from local files to enterprise database. After importing data, users can view, edit, and join the data using a graphical interface. Data exploration can help user mine potential information from a dataset. It provides both automatic and manual methods for users to inspect the structure of the imported data. As shown in Figure (2.2), users can drag several columns of data which might have a potential link into the panel, then the software will automatically choose suitable charts to visualize the data from which users can have an impression about the data. If the chart chosen automatically cannot get useful or right information, users can choose other charts to identify patterns and trends of the data further. As for report build, this function is used for creating final report when users confirm what display they want.

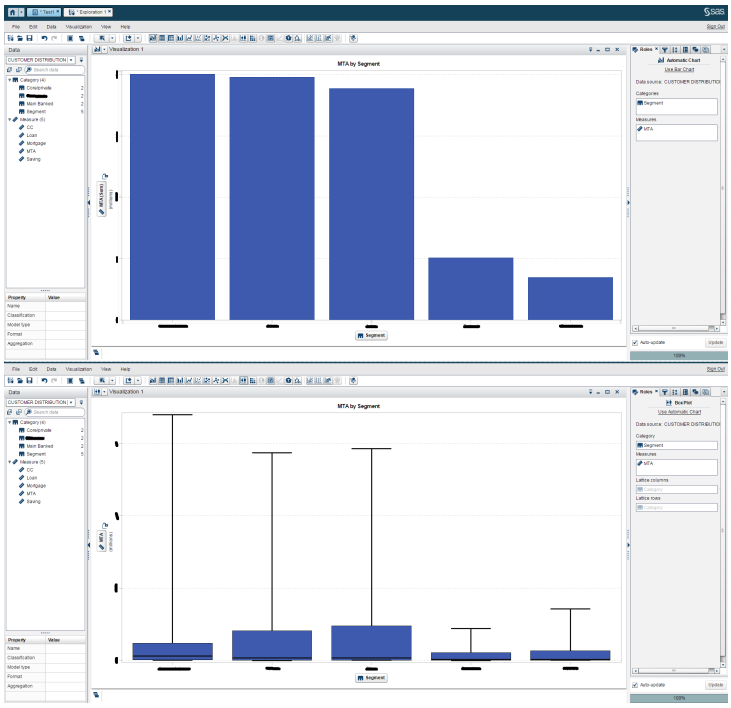

Figure 2.2: Data exploration

In this project, both Excel and SAS VA are going to be used together to create the final report.

2.4 Google Maps JavaScript API

Apart from Excel and SAS VA, Google Maps JavaScript API can be used as the supplement for geographical analysis. The map can scale up and down to view different range of the map. When choosing higher scale, the map tends to depict more detailed information and vice versa. Moreover, there are two types of map. The default type is road map and users can choose to view Google Earth satellite images. In this project, Heatmap Layer and Fusion Tables Layer are going to overlay on the top of the map and visualize geographical distribution of products.

A heatmap is a visualization which can be used to describe the intensity of data at geographical points. A colored layer will cover on the top of the map. By default, the higher intensity areas are, the deeper a red will be used to color it, and the area of lower intensity is going to be represented by cold colors like blue and green. The intensity can be used to reflect the quantity of accounts in a branch. It specifies a point by coordinate value and weight value (intensity):

```
location: new google.maps.LatLng(57.172, -2.771), weight: 180
```

The Figure 2.3 displays the effect of overlaying a Heatmap Layer. To customize how the heatmap rendered, five options can be installed.

1. Dissipating : Specifies whether heatmaps dissipate on zoom. If setting false, this layer will only appear on a particular zoom layer.
2. Gradient: Choose color schemes.
3. Opacity: The opacity of the heatmap, between 0 and 1.
4. MaxIntensity: The maximum intensity of the heatmap. When scaling the map, the size of heatmap points will change dynamically to make colored areas more obvious. This property allows you to specify a fixed maximum and avoids unexpectedly high intensity.
5. Radius: The radius of effect for each point, in pixels.

Figure 2.3: Pure heatmap display[5]

Fusion Tables Layer allows users to render data stored in Google Fusion Table on the Map. Google Fusion Table is a relational database table. It allows users to store relevant features about locations and provides interface to display them on the map. Figure 2.4 displays how fusion table stores data. For depicting geographical information, a fusion table must have a column to store coordinate.

Location	Current	Saving	Loan
51.348005, -0.007628	245	225	50
51.450702, -0.452072	526	480	120
51.696236, 0.370972	276	124	45
51.532694, 0.091137	573	352	100
51.594092, 0.436816	413	312	95

Figure 2.4: Fusion table with dummy data

Figure 2.5 displays the effect of adding Fusion table layer. Each row in fusion table is going to become a point on the map and it pops up a dialog to display the corresponding information of the point when clicking.

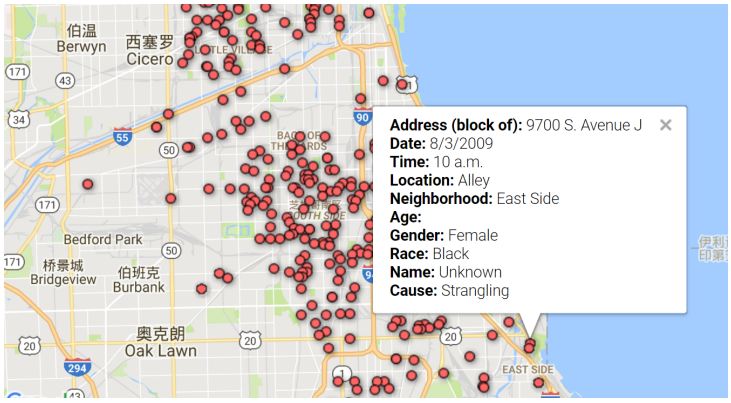

Figure 2.5: Pure fusion table layer display[6]

2.5 Product overview

The products in this company can be divided into personal products, business products, private products, corporate products, and international products according to the customers who use them. In this project, only personal products will be taken into consideration.

In terms of personal products, there are saving, current, mortgage, loan, and credit card products separately. Moreover, each product can be divided further. There are many choices of each product. For example, a current account can be packaged or non-packaged. A packaged current product is combined with some other services like insurance and travel cards, and a non-packaged current product only has the basic function. A particular product is called an

account type, so a product can be divided into several account types. Also, because of the difference between different accounts of the same products, many customers have the option to open several accounts of the same general financial product in this company to satisfy their needs.

2.5.1 Current Account product

The current account can be also called a bank account. It is usually the most fundamental product in a financial institute. It provides cash withdrawal, money transfer, use of a debit card as well as new technologies such as Apple Pay and cheque use service to customers. In general, customers may be charged a service fee for using it. Some financial institutes provide free service if customers choose add-on services such as an overdraft.

In this company, the current product can differ in:

- **Cash back**. A cash back account will return money to customers on a percentage basis when they perform certain transactions.
- **Packaged**. A packaged account has other non-banking benefits such as insurances, travel cards, discount tickets and restaurant vouchers.
- **Life cycle**. Like Student account, Youth account, and Graduate accounts. These accounts have benefits that are linked to the particular type of person/ age group that these accounts are aimed at.
- **Overdraft**. A customer may choose to have unsecured borrowing with the account.

Therefore, current products cover a wide range of customer types, and contains a variety of account types.

2.5.2 Saving Account product

The saving account is a money deposit account held at a retail bank. Customers store an amount of money in the bank and bank will return interest to the customers as a reward. There are four main types of saving product in this company:

- **Easy Access account**. There are no withdrawal penalties, so customers can withdrawal money out whenever they want with the interest being credited on the first business day of a month. This account is only for customers aged 16 or over. Because of the flexibility, the rate of this type of account is usually very low. Moreover, customers need to pay tax for the interest they earn.
- **Fixed term**. This account is usually for a large amount of money (£5,000 to £500,000). Customers store this money in the bank for a fixed period of time. The bank will pay in a higher rate that agreed in advance. In case of the unexpected, customers can take the money out with penalties.

- **ISA**. ISA stands for Individual Saving Account. It is a tax-free account which is for qualified UK residents who can enjoy tax-free privilege. So, he is also not necessary to pay for the tax of the interest he earns. The individual account can also be divided into instant access and fixed term.
- **Under 16**. The children's account is aimed at people aged between 7 and 16. Parents can use child's name to open this account, and the account is still managed by the parents.

2.5.3 Credit Card product

Credit card is a payment card that cardholders can pay a merchant for goods and services based on the cardholders' promise. Normally, the card issuers who are members of international payment organizations create a revolving account and give a line of credit to the card holders. The difference between a credit card account and a current account is that credit card has a delay of payment, which offers useful protection against the purchase of faulty goods. In the company, there are 4 main types of credit card:

- **Cash back credit card**. This account returns awards for customers' everyday spend. It requires the cardholders a mainland UK resident, over 18, and earning more than a particular amount of money per annum. Moreover, annual fee applies.
- **Advanced credit card**. This account provide a simple low rate on purchases and balance transfers. It requires the cardholders a UK resident, over 18, and earning more than a particular amount of money per annum, annual fee applies.
- **Cash back& travel credit card**. The advantage of this account is substantial spending power with some special travel benefits. It requires the cardholders a mainland UK resident, over 18, and earning more than a particular amount of money per annum. Moreover, annual fee applies.
- **Student credit card**. This account is designed for students. It doesn't have an annual fee. It requires the card holder a UK resident, over 18 years old and have a Student Current Account in this company.

2.5.4 Loan product

In general, a Loan allows a customer to borrow a fixed amount of money, then they pay back the principle as well as interest in instalments over an agreed period of time. For this product, the amount of money that can be lent is determined by the credit worthiness of the applier, the ability to pay the loan, and the purpose of money use.

2.5.5 Mortgage product

When a person wants to purchase a real estate asset but requires funding, mortgage can be chosen to assist in the purchase. This loan is secured against the asset, therefore it can be

considered as a secured form of lending. There the bank can lend much more money to the customer than that of common loan due to having this security over the asset. In general, a customer will borrow part of the amount needed for the property, and put forward their own capital to make up the difference. As well as the basic type of mortgage, many banks provide more flexible mortgage products so customers can apply for additional borrowing for example for home improvements and/or extensions.

2.6 ETL vs. ELT

In traditional database, the process of storing and using data is called **Extract, Transform and Load** (ETL). Data will be processed before storing. In this way, the database system can save the storage and speed up the loading time. Doing some basic transforming work can reduce the amount of data dramatically but it will hide much useful information simultaneously.

ELT is an alternative to ETL and it stands for **Extract, Load and Transform**. Firstly, it stores data produced by business directly without any processing. Secondly, it loads required data to processing units when analyzing. Generally, ELT has a great demand on storage and processing speed but it can record information without loss. This concept is always applied in data warehouse. The data warehouse in this project employs ELT and it is very common that tables contain all kinds of data of a business from the beginning of the company without any classification. For example, the monthly snapshots of credit card portfolio table contains approximately 40 columns and more than half billion rows, which lead to a long time in preparing data.

2.7 CRISP-DM

In this project, the Cross Industry Standard Process for Data Mining model is followed. It is a data processing model which gives the common processing steps to data mining experts to tackle problems. It can be seen from Figure (2.6), the process of data-mining can be roughly divided into 6 steps:

- **Business understanding**: Analysts should determine several items of a project in this step. Firstly, the problem waiting to be solved and possible goal (it can be modified then) of the project should be defined. To understand a business problem, analysts must study relevant context first. In this way, what data is required and what problems should be focused on can be known, and it is possible to get a correct result. Generally, domain experts will work with analysts together in this step. After defining the goal, the relevant requirement of achieving the goal should also be determined. To sum up, the first step is to have a first impression and come up a preliminary plan for a project.
- **Data understanding**: After understanding business step, analysts start to touch business data in this step. Because all kinds of data are stored in the data warehouse in the company without cleaning, data requires being selected and processed before it is used for reporting. When accessing the data warehouse, analysts should first get familiar with the

data and identify the data quality. Based on business logistic, analysts need to determine the data which may include relevant information and make sure its availability. Additionally, this step can turn back to previous step when having new understanding about business.

- **Data preparation**: This step transforms raw data from different data sources to report-ready data. Generally, this step consumes the largest amount of calculation. In general, a table in data warehouse will contain all kinds of attributes relate to a particular stuff. It only considers how to store these data properly and efficiently. Thus, analysts need to check the data quality and do necessary cleaning work. After making the data correct in both technology and logic , selected attributes should be extracted from tables and reconstructed.

- **Modeling**: An appropriate model or visualization program should be determined in this step. The goal of any data analysis is abstracting useful information or summarizing a common trend. Typically, there has been some mature solutions such as linear regression for specific problem type. When changing the model type, the form of data also changes. Therefore, stepping back to the previous phase is often needed.

- **Evaluation**: Until now, there has had a result for this project. However, whether the prediction or visualization can effectively solve the real problem or not are still unknown. Thus, the effect of the solution should be evaluated. If the result cannot satisfy the requirement, all the previous steps need assessing again.

- **Deployment**: If the solution actually take effect, it can be deployed to improve the real business.

Figure 2.6: CRISP-DM Process Diagram [7]

Chapter 3

Business & Data Understanding

The first as well as the most important step in analyzing financial data is to have a comprehensive and complete understanding of the business and the data. To learn the business in the company, the first two months are occupied by internal courses and reading relevant documents.

The aim of the project is to create a report which can reflect the personal product holding status of customers. However, because the complexity in business logic, the holding condition of the products can be described from many directions. Thus, the question faced currently should be what kind of information are the most meaningful and useful for the company. On the contrary, data limitation also needs considering when exploring what kind of valuable information should be mined since the accessible data only contains limited information. Basically, the process of mining valuable information relevant to product holding status is a trade-off between business logic and data limitation.

This section will combine data understanding and business understanding to explain the reasons for choosing the content to be analyzed.

3.1 Data Description

The company is experienced in providing digitalized financial services and the data collected is pretty rich and wide. The data warehouse is supported by Teradata corporation, thereby data is stored in the form of relational tables.

3.1.1 Data Source

Because of confidential reason and policy, the technical group doesn't give full right of accessing the data warehouse. In fact, only 11 tables which are relevant to personal products were accessible in this project. Moreover, only query authority and create temporary table authority were available. Here is the type of information that is contained in these tables. Data volume and data quality will also be introduced as below.

- **Table 1:** Customer to account relationships. This table stores all the saving, mortgage, loan, and current accounts which belong to a primary customer. In this company, ac-

counts that can be held by more than one person and the concept of primary customer represents the customer who utilizes the account the most. For each row, this table stores the type code of the account, the id of the customer who holds the account (customer id), the number of the branch that the account is opened in (branch number), the number of the account (account number), the relationship type between accounts and customers, the end date of the customer to account relationship, and the start date of the customer to account relationship. The primary key is comprised by customer id, branch number, and account number. It is worth mentioning that the branch numbers of saving, loan, and current accounts are associated the branch that they are opened at. However, mortgages and credit cards tend to have a centralized branch numbers associated with them. In fact, branch numbers historically were used to understand account volumes of our customer base as it helped to provide an easy to remember unique number and it can also provide a geographical split. But for mortgage and credit card accounts, the centralized branch don't provide this level of detail.

- **Table 2:** Account type. This table stores all the account types within this company. Furthermore, it includes some basic attributes such as whether the account belongs to personal product and a description that relates to the account type. In this project, it can be used to link account code to detailed account type information.

- **Table 3:** Account switch record. In this company, some products support account switch, this is particularly relevant to current accounts, which means customers are capable of switching the account type that they hold to a different type without changing account specific information. To record these actions, this table is created to store relevant information like account number, branch number and switch date.

- **Table 4:** Core or Private customer. The company differentiates customers into core or private. Most of customers are core customers. These standard retail customers are the foundation of the company. As for private customers, generally, they are customers with much more capital. Customers will be grouped to benefit providing more precise service. This table is a list that stores the mapping between customers and their 'segment'. Because a private customer can change to a core customer and vice versa, this table is the month snapshot collections.

- **Table 5:** Current account interaction times. A snapshot will be created each month to record how many times customers interact with each channel (for example branch, telephone, online). Because a customer may open more than one current account in different financial companies, the interaction frequency can reflect whether an account is the customer's main account to a great extent.

- **Table 6:** Customer segment. In order to better understanding the types of customers that the bank has, a model has been introduced to split them into certain groups depending on certain behaviors. The table stores the what segment a customer is by segment code. Examples of these segments are: 'Established Households', 'Everyday Banking', 'University / Graduates', 'Retired', and others. Because the segment might change, this table is the month snapshot collections.

- **Table 7:** Segment mapping. To save storage, whole expression of each segment is stored in this table.

- **Table 8:** Out of scope customers. This company excludes customers who meet certain criteria, as they are no longer considered relevant. Because the scope might change, this table is the month snapshot collections.
- **Table 9:**The Monthly snapshot of account balance. Each month, the company updates this table and add each account's balance of last month. Due to system reason, this table only contains information of saving, loan, mortgage, and current account.
- **Table 10:** Monthly snapshot of credit card portfolio. Each month, the company updates this table and add snapshot of credit card of last month.
- **Table 11:** Monthly snapshot of mortgage portfolio. Because mortgages are more complex products, this table contains more bespoke mortgage data.

3.1.2 Data volume

The company has existed in financial domain for many years and the earliest customer can even date back to the beginning of the last century. Thus, except for segment mapping and account type tables, the record number of the other tables are in billions. Even though the distributed server of the data warehouse has more than 200 processors, the iterations of cleaning so large dataset will consume a large amount of time and calculation resource.

3.1.3 Data quality

Financial field stresses strictly on rigorous style of work, so almost no missing data or non-logical data exists in these large tables. These tables are all with high quality and almost no need to do any repair work. Further work will be discussed in section 4.2.

3.2 Code review

A previous SQL code was reviewed, this code described the number of new personal products held by customers who had a particular type of Current account. It also viewed the customer holding trend of personal products over a 24 month period. The detailed process of organizing data in the SQL code is concluded into a flowchart. However, because of confidential reasons, it cannot be included within this dissertation.

The code also uses tables listed in section 3.1.1. Lots of business logics are learned from the code. Here list relevant logics learned from the previous code:

- If an account is closed, the end date of the account will record the day it closed. Otherwise, all the active accounts' end dates are populated with a date far into the future.
- Customer Id,Branch Number, and Account Number constitute the primary key for Saving, Current, Loan, and Mortgage accounts.
- A field called Personal records whether an account belongs to personal products.

- Customer segments are banded together to provide a high level overview.
- In Customer to account relationship table, a field is used for record the relation between an account and a customer. In this analysis, this field should be filtered to only show customers who own the account and exclude any related only customers.
- By default, only the data of particular brands in the company should be taken into account.
- The analysis range of credit card should be within retailed brand.
- There are many other useful details that benefit this project, but it cannot be claimed for confidential reasons.

3.3 Business analysis and data availability

After introducing the data source that can be accessed, this section is going to combine the business logic and the data together to explain the reasons for choosing the resolution. In data science, a method called Exploration Data Analysis (EDA) is going to be used to summarize the main characteristics of a dataset. Normally, it uses visual methods to explore what potential information might the data tell us. However, It is worth mentioning that some of the accessible tables contain more than 40 fields, so it is impractical to explore all the possible relationships in the data. In fact, most of the potential relationships among these fields are meaningless. Besides, there are many fixed analysis strategy of financial big data. That is also the reason why exploration data analysis (EDA) has not been used in this project.

3.3.1 Basic information

To give an as comprehensive as possible description of the product holding status, some basic information must be included. Therefore, the holding number of each product is supposed to be contained in the report. With these data, persons who viewed this report can basically have a first impression of personal products. Moreover, the holding condition is changing constantly with time and the changing trend of the holding number can reflect how competitive a product in the market to a great extent. So, it is of great importance to include the holding number of each product. Also, it can display the sale trend if adding time dimension.

Moving to **data availability**, the Customer to account relationships table (**Table 1**) has records of the start date of saving , mortgage, loan, and current accounts, from which we can obtain the holding number of the four products in any specific time. As for credit card product, there are monthly snapshots of credit card in monthly snapshot of credit card portfolio table (**Table 10**). Obviously, the holding numbers of each product in each month can be acquired.

3.3.2 Personal features

Customers have the different demand for their banking services, so they have different potential interactions with the company for different products. Generally, it is vital to attach tags to a

customer in precision marketing. In this way, a company can market products more accurately to to better meet their needs whilst increasing profitably. Thus, the holding quantity of different products for each particular customers is very valuable.

Firstly, in this company, customers are differentiated by core and private. Core customers are the most common customers in the company and they are the most fundamental customer group. The company also provides more considerate services to private customers in order to attract them. Private customers sometimes are not defined only by the amount of money they keep, some particular guests with great fame can also be grouped into the private group. Overall, the company can generally earn more benefits from a private customer than a core customer, but core customers greatly outnumber private customers. As for **data availability**, whether a customer is core or private can be known from **Table 4**, so the distribution of core and private customers can be demonstrated separately.

Secondly, whether the customer is in scope, as if they meet criteria they should be excluded. As for **data availability**, **table 8** is the special table for recording whether a customer belongs to this group, so the corresponding information is accessible.

Thirdly, whether a customer uses this company as the main financial service provider should be taken into consideration. In general, a financial company can not cover all kinds of services in all places. Even if a giant company can provide all services in all places, other companies may have more attractive services with same functionalities. So, particularly for current accounts, customers tend to either open several accounts in one company or across several companies. In order to facilitate the management of funds, most of customers mainly use large, responsible companies. In turn, for the financial company, customers who use them as their main bank tend to be more profitability. If a customer mainly use the services provided by this company, they will be called a main banked customer. The company judges a customer as a main banked customer by the number of debit transaction on their current account. That is to say, if the number of transaction of a customer's current account is larger than a set amount, they will be labeled as a main banked customer. As for **data availability**, current account interaction number table (**Table 5**) contains the interaction number of every current account in each month. Through simple judgment, whether a customer is main banked can be determined.

Lastly, it is learned from the customer segment table that customers are divided into several groups in the company. According to the knowledge gained from the code review, they are 'accumulators', 'active', 'premium', 'retired', and 'other' respectively. Combining **Table 6** and **Table 7**, it is easy to attach the 'segment' to each customer.

Overall, personal product customers in this company are going to labeled by four conditions: core or private, whether in scope, main banked, and segment. The report is expected to show the distribution of different groups.

3.3.3 Geographical distribution

The range of targeted holding condition report covers the business over the United Kingdom including Scotland, England, Northern Ireland, and Wales. The company's sales differs in different areas. On the one hand, there is no doubt that the population and the degree of affluence can be some of the main reasons. On the other hand, the product applicability and product competitiveness are also of great importance. The object of precise marketing can focus not only

on individuals but also on areas. If the orientation of products doesn't fit the local situation, the company is going to be failed in the competition with other financial companies. For instance, according to a news of The Guardian[8] in April 2009, the city dwellers of St Albans earned an average of £43,500 in 2008, and each citizens paid £10,500 in income tax. Comparing with St Albans, the people living in Hull earned an average of only £17,300 and paid £2,360 in income tax in the same year. Thus, if the company markets the products in St Albans directly to Hull without any change, they are very likely to lack attraction and the possible solution can be shifting the focus of the business to low-end products, while the reverse transplantation can face the same situation.

So, if we can describe the customer distribution in UK map (Figure (3.1)) [9], it can help non-analytical stakeholders observe the geographical distribution of the sales. Moreover, combining the map with the local population and estimated market share, staffs can find out the places where existing potential problems. After targeting places, more detailed analysis can be applied to find the pain point and improve services. Except for the function claimed above, geographical analysis can be used to assist more decisions.

Generally, branches and the Internet are the most common business processing point. The branches of the company spread over the UK. When a customer wants to open a new account, he can go to a branch of the company. Normally, customers prefer choosing branches in the neighborhood. So, the number of accounts that a branch helps customers open can reflect the sales in the corresponding region. The report tends to contain the sales of each branch.

Figure 3.1: UK map

Moving to the **data availability** (confidential reason is not considered here), an internal logic should be introduced. To begin with, branches of the company act as proxies. Customers can open new accounts in every branch of the company. In Customer to account relationships table, branch number field records which branch the account is opened in. However, all the branch numbers of mortgage accounts are replaced by several centralized branches' number (there are several sub-brands of the company). So, the information of mortgage accounts' opening branch is not accessible. Having discussed four types of products, the information of

credit card accounts' opening branch faces the same situation as mortgage product. In Table 10, the branch number of credit card are also replaced by several centralized branches' number. Besides, a mapping between branch numbers and their true locations is essential and there must have this kind of mapping in the company. Overall, saving, current, and loan accounts can carry out geographical distribution more easily.

Chapter 4

Data preparation & processing

Having determined what information should be included in the report and verified the data availability, the next step should be finding out the method of merging the required data together to a usable format. Because all the data is stored in relational tables, it can be more convenient to organize the final information in relational tables.

Any Data Science project needs to get the data in the format that is required before modeling or visualization. The required fields are expected to be grouped into several tables for easy visualization. In this section, the information is going to be formed into a expected structure.

4.1 Data preparation

Because of the privacy concerns of the customers' personal data, the access of the data warehouse is managed strictly. After joining the company, there is a 3-week delay to get the access authority since the complicated approval processes. And the permissions are limited to data queries and creating temporary tables over 11 tables claimed above.

4.2 Data cleaning

In most commercial applications of data science, the prepared raw datasets often go through a pre-processing stage which involves cleaning. The transformations applied at this stage can be done to reduce the in-class variability, normalize the magnitudes of features to a similar scale, embed domain information into the data, or remove useless information. According to the data analytics course note[10], the basic process of cleaning data is shown in Figure (4.1).

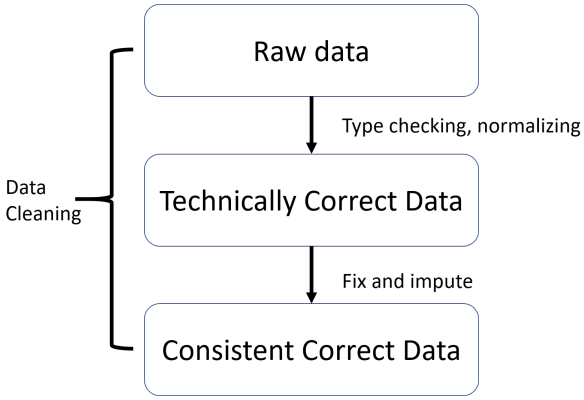

Figure 4.1: Data cleaning process[10]

A data set is a collection of data that describes attribute values of a lot of real-world objects. The technically correct dataset should satisfy that all the data should be stored in formats fit the real-world logic and all these in a format should keep consistent in the entire dataset in each value[11]. From Technically Correct Data to Consistent Data, the missing values, special values, obvious errors and outliers are either removed, corrected or imputed from Technically Correct Data.

Detecting the data quality in the company, the items are listed below:

- All the fields are stored in corresponding carefully selected data formats, character encoding, and string normalization,.
- In terms of missing data, some fields which represent whether a customer has corresponding attributes only have one-side value. For example, if a customer meets certain criteria and is out of scope, he will be recorded in the table, however, the content of the in scope customers missed.
- As for obvious errors, all the data is strictly collected from the real business and customers also attach great importance on their properties, so no obvious errors are found.
- As for outliers, all the data is collected from the real business strictly. Even if there is outliers data, it is the important part of the report.

Overall, the data is with great quality and some missing data can be corrected easily by the SQL code shown below:

```
CASE WHEN customer_id IS NOT NULL
THEN 1 ELSE 0
END AS scope
```

4.3 Data filtering

These candidate tables contain lots of extra information. To make it simple, Table (4.1) is drawn to list the fields required and the meaning of these fields.

Table	Name	Description	Example
Table 1	customer_id	The id of the account's customer	99999999
	branch_no	The number of the branch relate	888888
	account_no	The number of the account	77777777
	type_code	The codes which represent the account's category	AA
	open_date	The opening date of the account	2001-03-20
	relation	the relation between the account and the customer	OJ
	end_date	the date that the account is canceled	2050-06-30
	brand	the sub-brand of the account	FAKEBANK
Table 2	type_code	The codes which represent the account's category	AA
	category	The product name of the account	Saving
	personal	whether the type is personal product	personal
	brand	the sub-brand of the account	FAKEBANK
Table4	customer_id	The id of the account's customer	99999999
	Core_Private	Core or private	Core
	period	The month that creates the snapshot	2001-01-01
Table5	customer_id	The id of the account's customer	99999999
	trans_quan	The quantity of the account's transactions	13
	period	The month that creates the snapshot	2001-01-01
Table6	customer_id	The id of the account's customer	99999999
	segment _code	The code that represent the accounts' segment	YP
	period	The month that creates the snapshot	2011-01-01
Table7	segment_code	The code that represent the accounts' segment	YP
	segment_name	The name of the segment	Young Potentials
	period	The month that creates the snapshot	2011-01-01
Table8	customer_id	The id of the account's customer	99999999
	scope	whether customer is in scope	out of scope
	period	The month that creates the snapshot	2011-01-01
Table10	customer_id	The id of the account's customer	99999999
	period	The month that creates the snapshot	2011-01-01
	account_id	Credit card's number	77777777
	brand	The sub-brand of the company	FAKEBANK
	retail_banked	Whether retailed product	0/1
	status	The status of the account	0/1/NULL

Table 4.1: Fields list

4.4 Data aggregation

All the engineered features on the raw data has been targeted. The aggregation step needs to merge them into usable structure. After combining data with the business, three final tables are planned to be created:

4.4.1 Time period distribution

This table is going to include personal feature information, time period and the quantity of personal products together. In other words, the structure of the data aggregation should benefit the displaying of all these information. Table (4.2) shows the structure of the final table.

Main banked	Scope	Core/Private	Segment	Product	Number	snaptime

Table 4.2: The structure of analysis 1

This table's information is collected from Table 1, Table 2, Table 4, Table 5, Table 6, Table 7, Table 8, and Table 9. It can be thought that Current, Saving, Loan and Mortgage products' information is mainly collected from Table 1 and the Credit card product's information is mainly collected from Table 9. For credit card data, records which satisfies certain criteria are selected which can be simple in logic. By contrast, it is much more difficult to get required records from Customer to Account Relationship table since the table which stores credit card data have the snapshot of each month while this table doesn't. To achieve the goal, a temporary table which contains the first date of every targeted month is created and the data is extracted from Table 9 rather than inserting directly since the insert authority is not accessible. All the account open dates are formatted to the first date of the month. Then 'Left Join' the temporary table with the table on period to create these snapshots of required months. After adjusting two sets of snapshots from 01/04/2016 to 01/04/2017 to a same structure, use "UNION" command to combine them together. Then merge different account records with the same customer to one record and calculate how many accounts of each product customers hold. Until now, records of how many accounts of each product customers hold in each month has been prepared.

Final step is to add labels to customers and group customers with same labels. For main banked label, if the transaction quantity of a customer's current account is larger than a threshold, the customer is going to be labeled as main banked. Current Account Interaction time table has records of every month and the latest month is selected to reflect whether a customer is main banked customer. For customer segment, initially, codes stored in Customer Segment table should be replaced by corresponding names and these segments are going to be classified further into 5 types. Also, the latest month will be choose to reflect this label. For core/private and scope labels, the records in the latest snapshot will be selected. To have a more detailed view, the flow chart is shown in Figure 4.2 below. The Merge Table 4 in the figure is the final table ready for visualization. Because of the confidential reason, only part of the SQL code is shown in Appendix B.

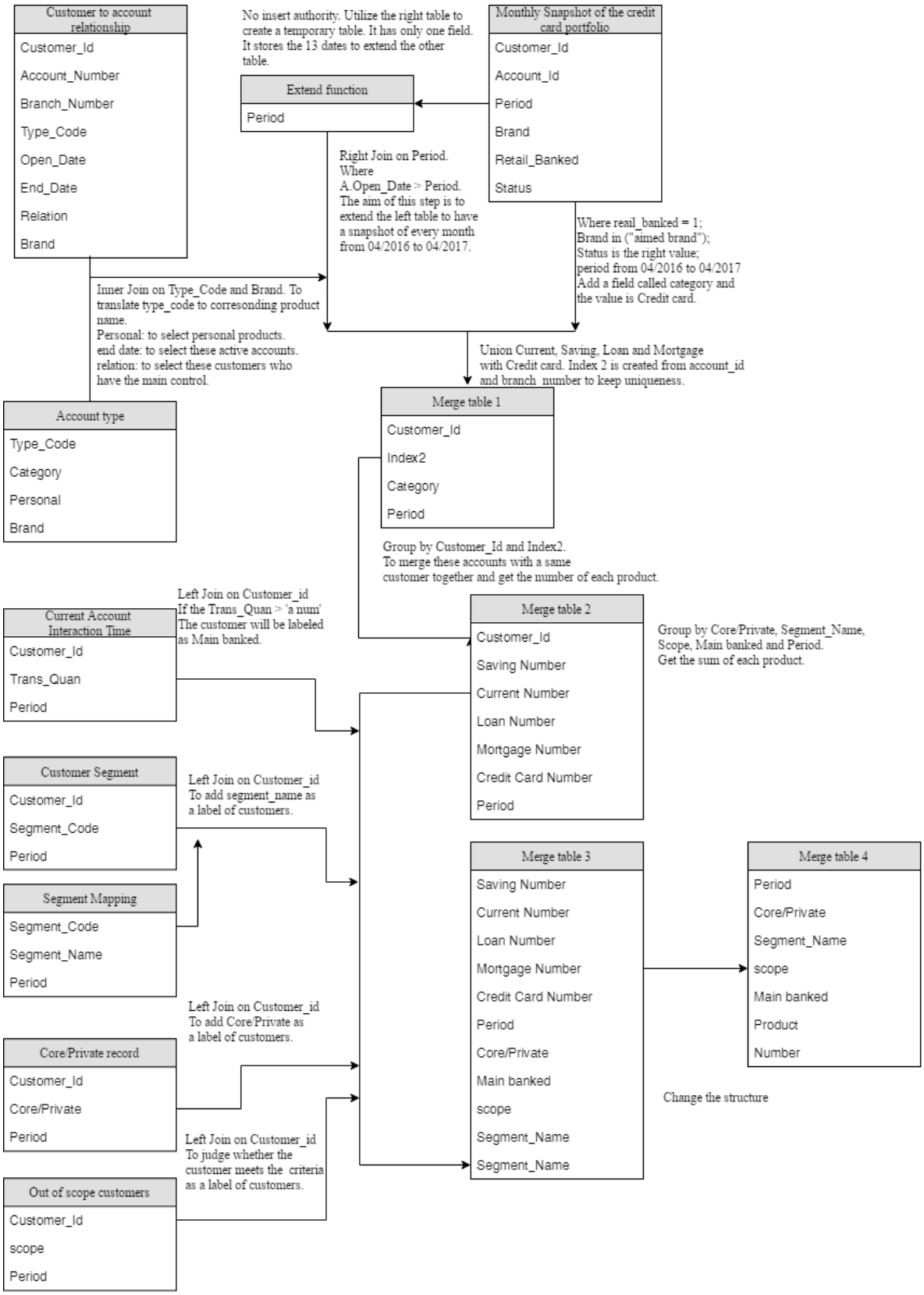

Figure 4.2: The flow chart of merging the table for the first analysis

4.4.2 Customer holding relationship

Customers can hold more than one account in each product. It is of great importance to know the relationships among their quantity. The structure of this data aggregation should be suitable for displaying the relationships of these products' quantity. Table (4.3) shows the structure of the final table.

Product	Number	Saving sum	Loan sum	Current sum	Mortgage sum	Credit card sum

Table 4.3: The data structure of analysis 2 (sum)

This analysis uses less tables to collect information. It uses Table 1, Table 2, and Table 9. Firstly, select Saving, Current, Loan and Mortgage accounts which meets the requirements from the joining result of Table 1 and Table 2. The first requirement is that an account must be personal account and the customer is the main owner of this account in the record. Also, the account should have not been ended. The result of the first step only keeps customer id, account number, branch number and account type. Then merge these accounts with the same customer together and calculate how many accounts of each product customers hold, after which the data of Credit Card (pre-select these meet requirements) should be added from Table9. Until now, the table records customers and how many accounts of each of five products they hold. Then group this table respectively by Saving Number, Current Number, Loan Number, Mortgage Number and Credit Card Number to acquire 5 new tables. Finally, use the "UNION" operation to merge these five tables to one table. Figure 4.3 below is the aggregation flow chart and the temporary table 10 is the final table ready for visualization. Because of the confidential reason, only part of the SQL code is shown in Appendix B.

Apart from the sum, there is another version which uses the same logic but calculate the average number when grouping by each of products' number. Table (4.4) shows the structure of the final table.

Product	Number	Saving avg	Loan avg	Current avg	Mortgage avg	Credit card avg

Table 4.4: The data structure of analysis 2(average)

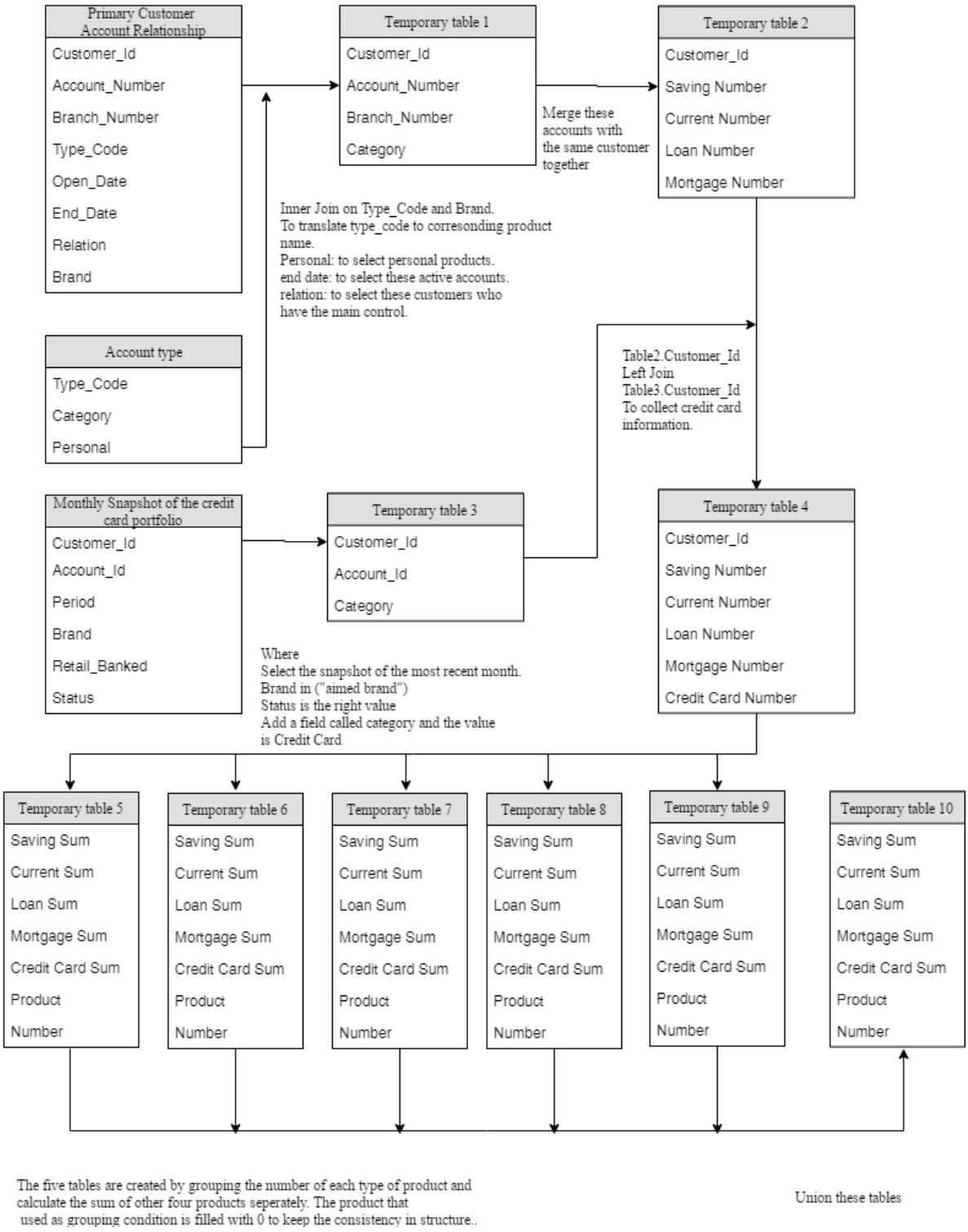

Figure 4.3: The flow chart of merging the table for the second analysis

4.4.3 Geographical distribution

Table (4.5) shows its structure.

Branch num	Saving num	Current num	Loan num

Table 4.5: The structure of analysis 3

Because of confidential reason, the data used for the demo of this analysis is not collected from the database but by fabricating. The produce of dummy data is going to be demonstrated together with the visualization step.

4.5 Performance

The data warehouse is shared over the entire company. Thus, almost every time the running time may be different. Moreover, some business tables will be updated automatically at the beginning of every month and the code is going to run extremely slow. So, it can be difficult to have a precise evaluation of the performance. According to the past running time, the shortest time can be less than 15 minutes while the longest can be more than 1 hour to get two final tables.

The code has been slightly optimized when programming. For this project, the code is only used for once, so, it might be meaningless to make further optimization work. However, if the code is added into the regular analysis, more optimization work should be applied.

Chapter 5

Data visualization & Analysis

Data visualization is a general concept that makes an effort to assist people to easily understand the features and importance of data by putting it in a visual context. Rodriguez and Kaczmarek(2006) [13] claimed in their book: "You can rely on data visualizations to see outliers, trends, correlations, and patterns". Making data visualized has become one of the basic processes in data science and it is a science as well as an art. Both implementation mode and user-friendly expression should be taken into account.

The visualization tool SAS VA can import data by local files, web source, and database. In the project, importing data from database directly is the most ideal way. However, the company prevents creating permanent table for this project, therefore the final output was transferred to Excel then imported by local file. Because the copy operation loses the meta information of each field, extra completion work is carried out to add this data back.

5.1 Basic Aggregation charts

Aggregation plots is the visualization for 'Time period distribution'. It aims at reflecting the sum of five products in 13 months, how the percentages change for each product in a month, and the changing trend of the sum of five products over the year. Moreover, there are five drop-down lists locate on the upper right corner of the report. Through operating these lists, the bar chart and the pie chart would display the corresponding group's information.

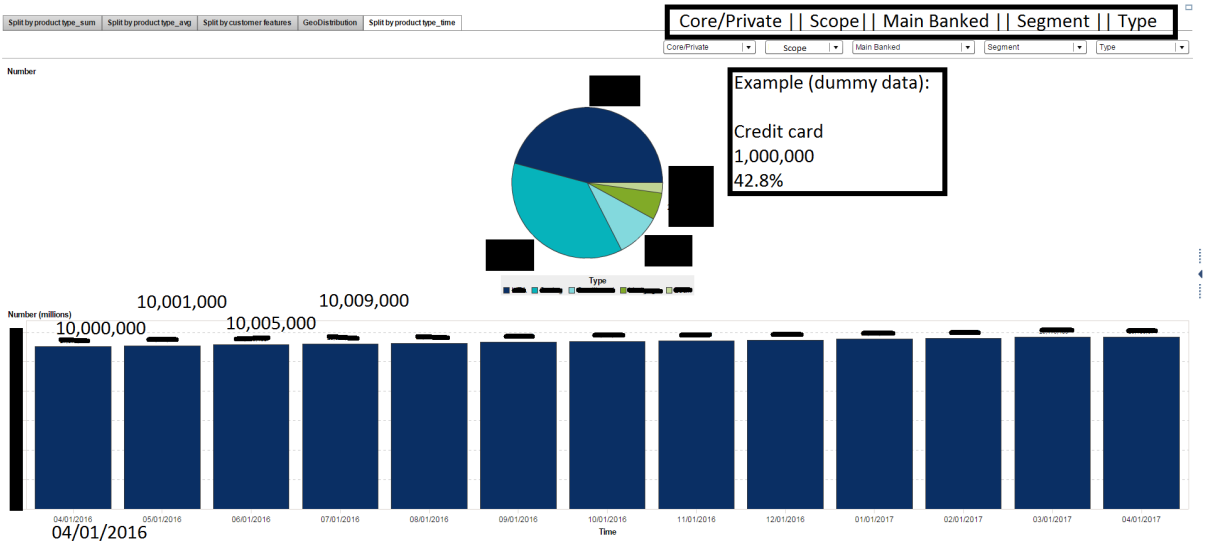

Figure 5.1: Screen shot of the basic aggregation charts

When opening the dynamic report, the first appearance of this report is an initialization interface (Figure 5.1) which displays the sum of five products at the beginning of every month from 04/2016 to 04/2017. It is quite clear when using the visualization on a screen, however, the information displayed in the picture is too small to have a clear look. Some annotations are added in the picture to assist understanding.

The set of charts is composed of 5 drop-down lists, 1 pie chart, and 1 bar chart. Firstly, the five drop-down lists can be operated to choose a particular group of customers. The first drop-down list is responsible for choosing the customers are core customer, private customer, or all. The second drop-down list can choose whether the group of customers are out of scope. Whether customers have the main banked label is decided by the third drop-down list. The forth drop-down list has six items: 'Active', 'Accumulators', 'Premium', 'Retired', 'Other', and 'All' to choose corresponding group of customers. Lastly, the fifth drop-down list can choose the type of product. The five drop-down lists can be used together. The bar chart displays the sum of five products when choosing nothing in the last list, and it will switch to display the corresponding quantity of the product chosen.

The bar chart can display the quantity trend of selected customers' accounts. The vertical axis is the number of selected accounts and the scale can change dynamically to fit the quantity of the accounts. The horizontal axis has 13 points to represent months. For example, if a user has a demand of knowing the credit card holding quantity of retired customers. He should choose 'retired' in the forth drop-down list and choose 'credit card' in the last drop-down list. Then the bar chart will display of quantity of retired customers' credit card of each month.

The pie chart has tight link with the bar chart. It is mainly responsible for displaying how many percentage each product account for, which means it tends to become a solid circle when selecting only one product in the fifth drop-down list. The pie chart displays the percentage composition of five products in one of these months. When the bar chart displays the sum of five products, users can designate the pie chart to display a particular month's distribution by clicking the corresponding bar in the bar chart.

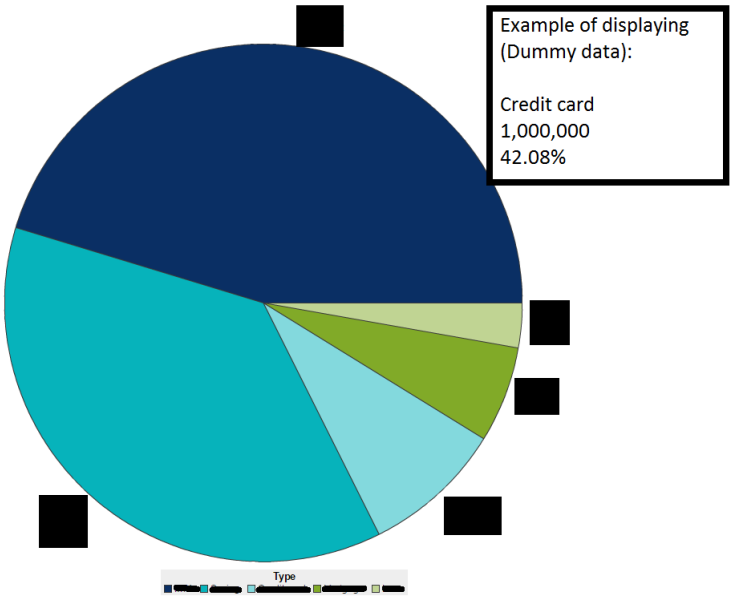

Figure 5.2: Pie chart

5.2 Holding Frequency charts

The Holding Frequency chart is the visualization for 'customer holding relationship table'. There are two visualization versions for average and sum respectively. Each one is composed of 1 drop-down list and 1 compound bar chart. Because they have the same structures, here we mainly discuss the sum one.

The drop-down list has 6 items: 'Current', 'Saving', 'Loan', 'Mortgage', 'Credit Card', and 'All'. After choosing one type of these products, the horizontal axis will represent the number of the product that a customer holds and the vertical axis will show the number of other products that customers hold with this condition.

For example, as shown in Figure 5.3, after choosing Credit Card in the drop-down list, the horizontal axis displays from 0 to 7, which means customers may not hold a credit card in the company and the largest holding number of credit card is 7. As for the vertical axis, it represents the number of the other four products associated with the number of credit cards each customer holds. Figure 5.4 displays the visualization of average when choosing credit card. It is worth mentioning that the visualization can display concrete quantity when putting the mouse over bars.

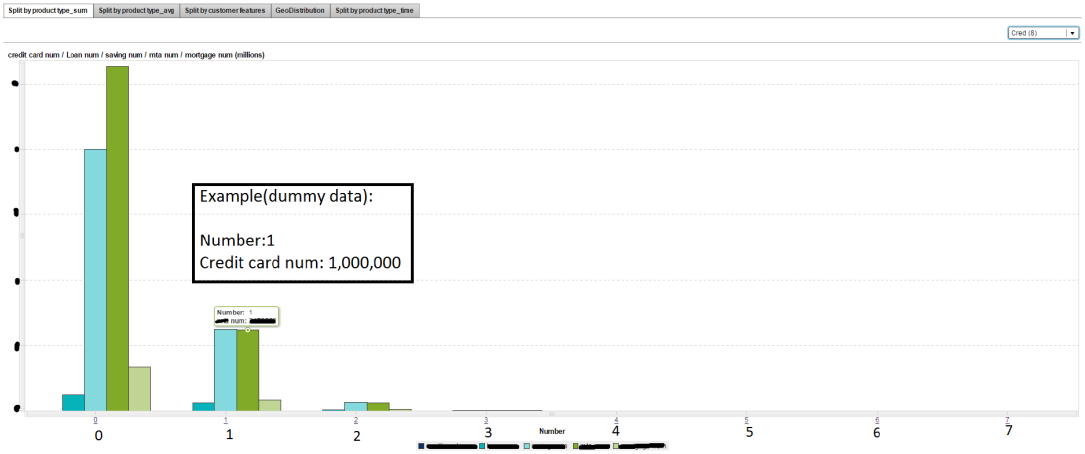

Figure 5.3: Holding Frequency example 1

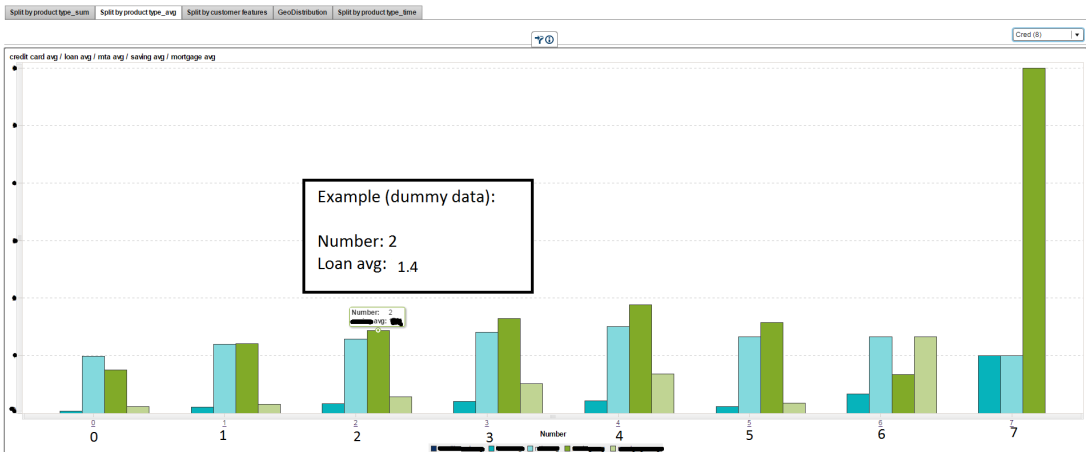

Figure 5.4: Holding Frequency example 2

Because a customer can hold more than one account of any product, this visualization aims to display the potential relationship among these products. For example, when clicking 0 in the horizontal axis, the chart is going to become Figure 5.5. The five bar sets display the sum of other four products when customers don't hold a particular product. Obviously, we can find out that Current account is the most popular account type in the company. Also, customers always tend to open Saving Accounts and Current Accounts simultaneously. In fact, because it is more convenient for customers to transfer their money between Saving accounts and Current accounts, these two products are very likely to be used together. In turn, if the relationship between Saving accounts and Current Accounts is not so tight, further investigation might be required.

Figure 5.5: Holding Frequency example 3

Also, the display content can be narrowed by choosing both the type of product and the number of accounts held by a customer. Figure 5.6 displays the total amount of each of the other four products held by customers who don't have Loan account.

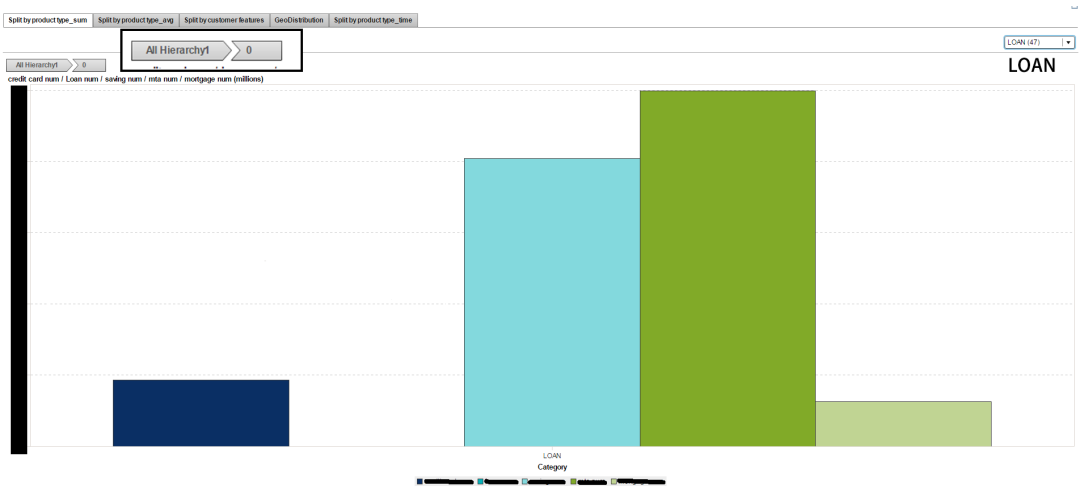

Figure 5.6: Holding Frequency example 4

Besides sum, average is also very necessary since sum value can easily have a comparison within the same bar set but the average values can be compared over different bar sets. The Figure 5.7 displays the average values when clicking 0 in horizontal axis. There are five bar sets and each of them has four bars. To choose credit card further in the drop-down list, the display tends to become Figure 5.8. It is a more detailed demonstration of the selected situation.

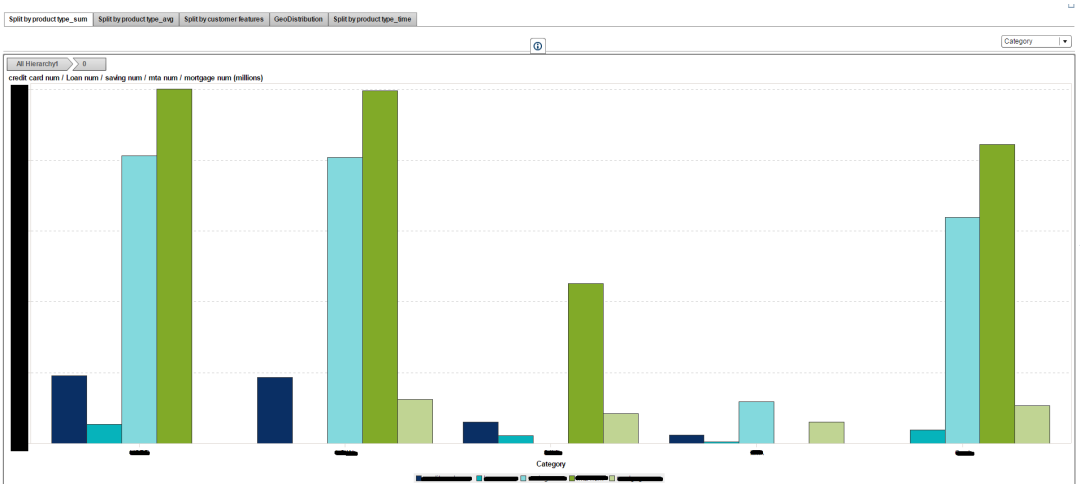

Figure 5.7: Holding Frequency example 5

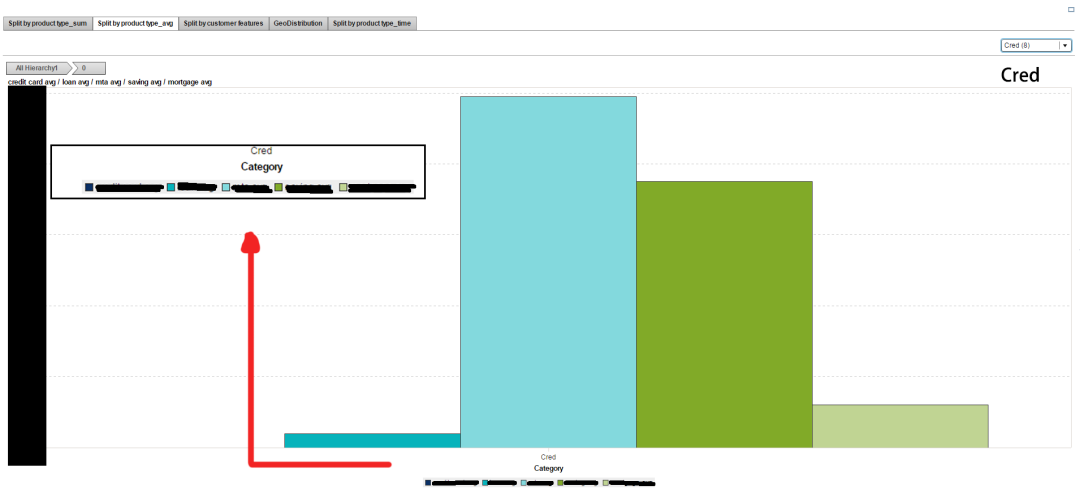

Figure 5.8: Holding Frequency example 6

5.3 Geographical Distribution map

After browsing the softwares installed on the internal computer, only SAS VA provides a possible geographic analysis function. Heat map can be selected to display how prosperous the sales over different places. However, according to the document of SAS VA [12], it can only locate a country or region by SAS Map ID Value, which means it can only draw heat map in coarse grain and it is impossible to customize a coordinate on the internal analysis software. After validating the unavailability from these professional and Integrated software tools, geographic information system was considered but installing new software personally is prohibited on the internal computers. Finally, this visualization is prepared to be implemented by using external tools. Because of the confidential problem that the company prohibits data migration from in-

ternal network to any networks that is likely to leak their data to public, dummy data was used to demonstrate the effectiveness of this visualization.

Branches normally locate around cities, so **Dummy data** is created manually rather than random generation automatically:

1. Open Google Map.
2. Select several big cities around the UK and the cities are London, Cambridge, Oxford, Bristol, Cardiff, Southampton, Plymouth, Brighton, Manchester, Liverpool, Edinburgh, Glasgow, Dundee, Aberdeen, Inverness, and Northern Ireland.
3. For each city, randomly and evenly select 4 to 6 points around the city since the the real site selection needs to consider how to cover as wide as possible by limited branches. Google map will return the latitude and longitude of clicked points.
4. For each point, dummy quantity of Current accounts, Saving accounts, and Loan accounts needs to be fabricated. It is worth paying attention that the the quantity of Current accounts slightly outnumbers Saving accounts and it is several times larger than the quantity of Loan accounts.

The number of dummy branch points is much less than the real and they sometimes locate in a rural area which will never happen. So, the dummy data is only fabricated for displaying the effect of this visualization. The structure of the dummy table is shown in Table 5.1:

Longitude	**Latitude**	**Saving num**	**Current num**	**Loan num**

Table 5.1: The data structure of geographical analysis

Many digital map API providers were contrasted for the project and Google Maps JavaScript API was finally decided. Abandoned map APIs exist one or more of shortages listed below:

- The API doesn't have multiple language versions.
- It is difficult to add a dynamic information layer on the map. Some providers provide interface of adding new information layer, but it doesn't have a ideal method of managing information.
- The API documents are not clear. For example, the pictures and demos in the document of the Bing Maps API provided by Microsoft can always fail to access.
- Some API don't have detailed geographic information about the UK.

The Heatmap Layer needs to switch among the three products to display them respectively. In this way, three buttons are added. The heatmap will change to display corresponding data when clicking a button. In order to demonstrate the quantity of accounts in a branch, the quantity is going to be scaled to weight. Consequently, the more accounts a branch deals with, a warmer color (deeper red) will be painted.

As for Fusion Table Layer, it loads an interact-able point in the same location as the center of each heat circle. When clicking a point, it is going to load the corresponding information of the branch. Here only the quantity of three accounts are displayed and more information can be added in real use.

Figure 5.9 is the overall screen shot of the interface:

1. Upper left corner can choose the type of displaying the map, road map type or satellite type.
2. Bottom right corner can change the zoom of the map or the same functionality can be achieved by using mouse wheel. In this way, users can view the overall geographical distribution of products. Also, they can use large scale to view a small range of area in detail.
3. The first button in the center console controls whether display the heatmap. Sometimes the heatmap layer might influence the Fusion Table layer's view, so users can cancel the display of the heatmap Layer.
4. The second button can change the color schema of the heatmap, which can be compared in Figure 5.9 and Figure 5.10. More color schema can fit more users' habits.
5. The third button can decrease or rise the number of pixel a point covers to help users view more clearly. Because colored area influenced by a branch changes with the scaling, different radius can help users compare the difference among different branches more clearly.
6. The forth button can change the opacity of the heatmap.
7. The last three buttons of the center console is responsible for switching the demonstration among different products.

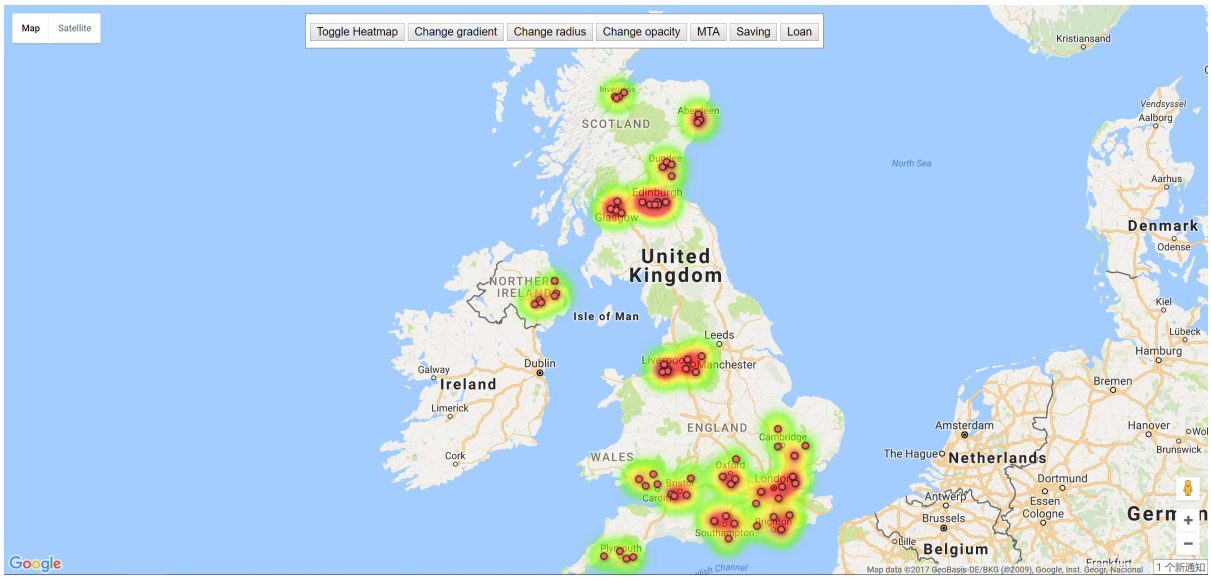

Figure 5.9: Overall Screenshot of the interface

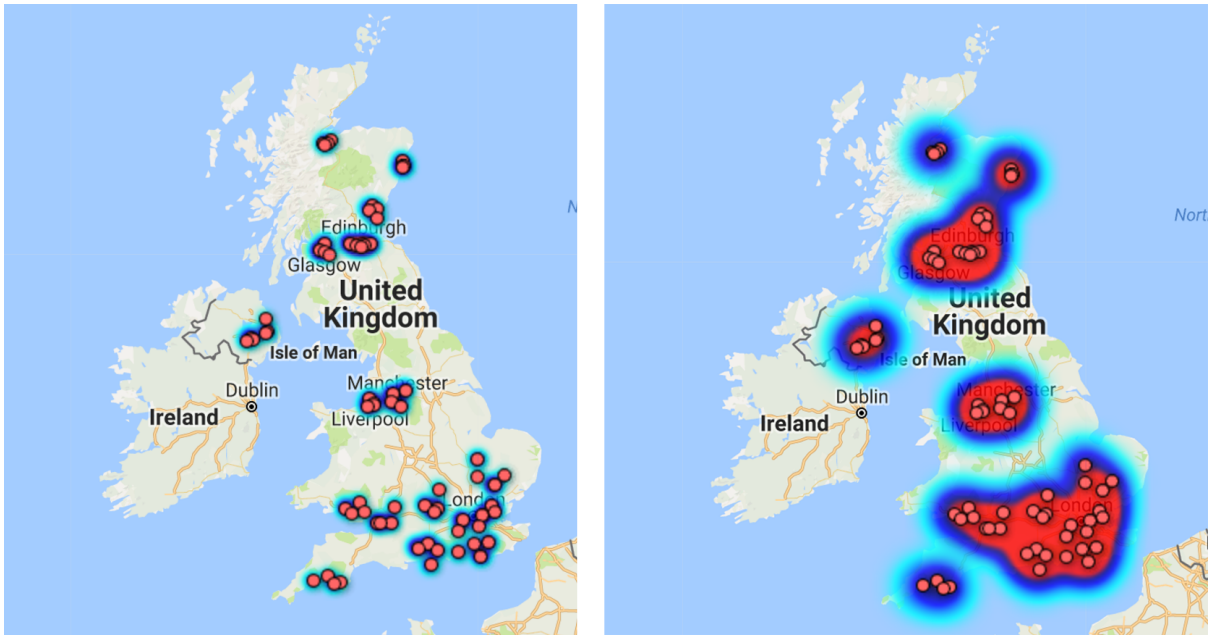

Figure 5.10: The comparison between different radius and another colour schema

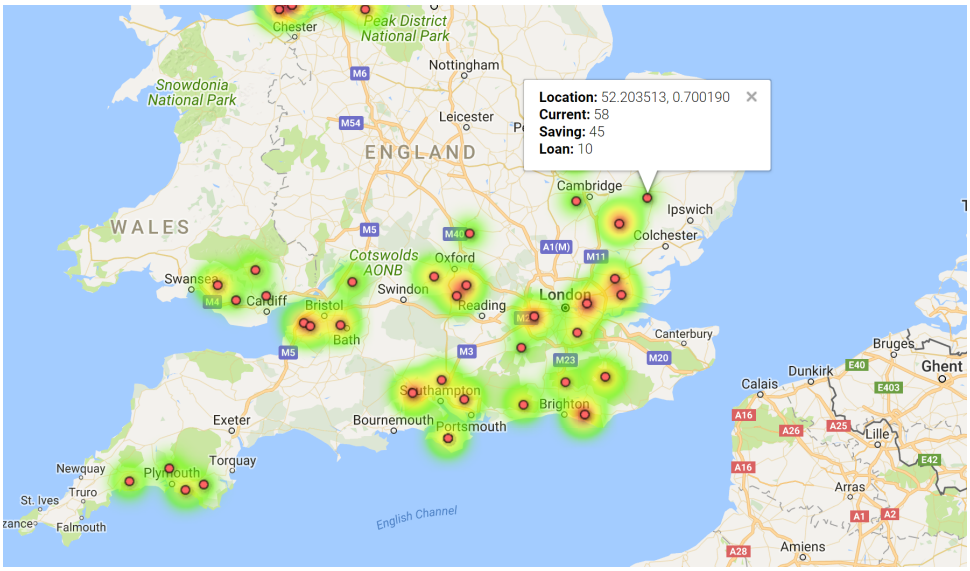

Figure 5.11: The dialog which renders corresponding information

With this visualization, it is easy to learn the holding condition of each product in geographic terms.

Chapter 6

Evaluation

In fact, whether a single report can reflect a subject perfectly is very debatable, so it is difficult to have the evaluation step. In this project, the evaluation step is going to be carried in a mixed way. On the one hand, the evaluation work is going to be based on some standard questions. On the other hand, a comment will be given by the company manager of this project. Then, the evaluation will start with standard questions and conclude with the comment.

Lam, Heidi, et al (2011) systematically concluded the seven most commonly encountered evaluation situations after reviewing over 800 papers. Some scenarios claimed in the paper will relate to how to evaluate a visualization tool and other aspects. As for this project, several relevant scenarios will be selected to evaluate this visualization.

6.1 Evaluating Environments and Work Practice

Within Financial Services, there is increasing focus on being able to understand the customer base in order to better meet customer needs. To achieve this goal, this company is becoming more customer centric. The report uses SQL to interrogate the Teradata data warehouse, Excel to import data, SAS VA and Google Map JavaScript API to create the visualization.

- For **user group**, the report will be used by non-analytical stakeholders in the company and help provide further insight that will in turn drive better informed commercial decisions.
- For **data use**, the data used in this project is only a small part of data produced by personal products. There is no doubt that much valuable information has not been mined. The project was set as part of getting full understanding of customers and much more information can be mined further.
- For **chart use**, bar chart, pie chart, drop-down list, chart link and map are used in the project. Except for the geographical distribution, the chart use is very basic but enough. In terms of the map, it is an innovation in understanding their customers. However, whether and how it can be used in the company is an issue for further investigation.

Overall, the analysis produced by this project can give stakeholders the ability to answer specific simple question by themselves and they can access it anywhere with Internet connection.

6.2 Evaluating Visual Data Analysis and Reasoning

The goal of the project is to organize data and make it easier for staff to recognize valuable information. This section will list the possible information the analysis can recognize from. For the **first set of charts**, the visualization can help non-analysis stakeholders recognize:

- In the case of different groupings, the quantity of the five products from April/2016 to April/2017.
- In the case of different groupings, the holding trend of the five products from April/2016 to April/2017.
- In the case of different groupings, the sum of the five products from April/2016 to April/2017.
- In the case of different groupings, the trend of the sum of the five products from April/2016 to April/2017.
- In the case of different groupings, how many percentages each of five products account for respectively from April/2016 to April/2017.

For example, with the data, staff can deduce the investment potential of the different group of people and pay more attention to attract particular groups.

For the **second set of charts**, the visualization can help staff recognize the corresponding sum(average) of other four products when customers have a particular number of a product. Further, for example, staff can deduce the potential relational among these products. Outliers data can be detected easily such as a customer hold more than 25 saving accounts and 15 current accounts together. So, for instance, staff can have some products with tight relationship tied together to increase sales.

For the **map visualization**, it can help staff recognize:

- The sales condition in different areas. Staff can compare the sales condition in different scales. For example, if they choose a small scale, the sales condition in Scotland and England can be compared; if they choose a large scale, the sales of two branches can be compared.
- The sales condition of different products. Staff can ask the report to display the sales condition of Saving, Current and Loan products.
- The detailed sales information of a branch.

With the information, staff can draw up more particular sales strategy to fit local condition.

6.3 Evaluation User Experience

Evaluation of user experience seeks to know how people react to the report. The goal of this evaluation is to understand to what extent can the report help the staff.

- In terms of the **ease of use**, the report created by SAS VA can be operated by clicking or touching and it doesn't need any knowledge on programming. Because staff in the company is familiar with the interface of SAS VA and the report will be only shown to them, there is almost no learning cost. The report created by Google Map JavaScript API can also be operated by mouse or finger. Because Google Map is pretty popular in the world, it might not cost much time on getting familiar with it.

- As for **accessibility**, both tools can publish report as a web service. SAS VA is very professional and it can fit both PC and mobile end varying in screen size. Google Map JavaScript API can also display the visualization on both PC and Mobile end, but the display effect is not stable varying in screen size. Normally, it displays clearly on PC while it is hard to operate on mobile end. The largest problem of Google Map JavaScript API is how to protect the data safety and it can possibly be solved by implementing a copy in their internal server or use other software tools such as ArcGIS.

- Moving to **user-friendly display**, if lots of information is in a screen, users can ask SAS VA to dynamically display data in a different level, which means any information can be viewed clearly. Moreover, linking pie chart and bar chart in the 'Basic Aggregation charts' can display the multi-dimension information in a understandable way. Google Map JavaScript API can scale to different levels to have a good view. If users want to know the detailed data, they can directly click the corresponding point.

- For **response time**, the report created by SAS VA responses user operation slowly while that of Google Map JavaScript API can react quickly.

The visualization is based on two mature tools. Moreover, SAS VA is an internal tool and the staff in the company are very familiar with it.

Chapter 7

Conclusions

This chapter will provide a conclusion to the work down throughout this project and give a focus on some future work that can be done in this area. It will start with a complete review of the project, covering the motivation, methodology and results. Then the chapter examine how well the goal is achieved. Finally, the chapter will conclude some recommendations for future work in this area.

7.1 Review

In order to better meet customer needs, the company focuses on gaining a fuller understanding of customers, with metrics such as age, income, assets, product holdings, channel usage and spending habits now all being considered on a more regular basis. Within this mind, this project tries to analyze which of personal financial products existing customers hold.

The methodology of this project was broken down into business & data understanding, data preparation & processing, data visualization & analysis and evaluation. Chapter 3 describes the business logic and the structure of the accessible data for this project in detail. Based on the preparation work, which metrics should be taken into account is determined. Then in chapter 4, the required data is cleaned, filtered, researched how to organize it into suitable structure and finally aggregated into three tables. After getting the data, chapter 5 describes the process of visualization. The first visualization is visualized by 5 drop-down lists, 1 bar chart and 1 pie chart. The second visualization is visualized by a drop-down list and a compound bar chart. The third table uses external tool and the effect has to be displayed by a demo for the confidential reason. Finally, chapter 6 evaluates the report relying on a standard process and comment. It evaluates the report from environment & work practice, visual data analysis & reasoning and user experience.

7.2 Goal accomplishment

As part of the main goal, the final output created by SAS VA in this project has been uploaded and an online dashboard has been developed to allow users to easily view and interrogate the

data. From the final report created with SAS VA, staff can easily gain an impression on which of personal financial products existing customers hold at a high level and they can use the report to explore more potential information to help them make decisions from a lower level. As for the demo created with Google Map JavaScript API, it provides a new thinking that focuses on the geographical difference to the company.

7.3 Future work

This section describes some potential improvements that can be implemented to improve the report. These include enhancements that may be made possible by an extended data, as well as utilizing new visualization tools.

1. Holding Frequency charts calculate the sum of each products. In future work, it can calculate the sum of customers who hold a particular number of a product.

2. Holding Frequency charts calculate the sum and average of all products held by existing customers. In future work, it can be extended to calculate the sum and average of customers in different groups. To be more exact, to add customer features such as scope, segment, main banked and core/private to the analysis.

3. To divide the customers by more features. In the project, customers are divided by four features and it can be divided by more features such as customer age.

4. To divide the products by more features. Due to the complexity and time limit, many features in the data haven't been mined. For example, in this project, we only know a product contains a number of types but different types of a product are not taken into consideration. Many valuable classifications are still waiting to be mined.

5. Apply the geographical distribution to real use.

6. To evaluate the code performance and optimize the code. The code runs on Teradata cluster, so it is difficult to evaluate the performance. However, as the business rising, it is going to be more and more important to consider the performance since more and more regular analysis tends to run on the cluster.

Appendix A

SQL Syntax example

- **Select:** The SQL Select statement is used to select data from a database. It will return a result table called the result-set which contains the data selected.

```
SELECT column1, column2, ...
FROM table_name;
```

- **Update:** The SQL UPDATE statement is used to alert the existing records in a table.

```
UPDATE table_name
SET column1 = value1, column2 = value2, ...
WHERE condition;
```

- **INSERT INTO:** The INSERT INTO statement is used to insert new records in a table.

```
INSERT INTO table_name (column1, column2, column3, ...)
VALUES (value1, value2, value3, ...);
```

- **DELETE:** The DELETE statement is used to delete existing records in a table.

```
DELETE FROM table_name
WHERE condition;
```

- **CREATE:** The CREATE TABLE statement is used to create a new table in a database.

```
CREATE TABLE table_name (
    column1 datatype,
    column2 datatype,
    column3 datatype,
   ....
);
```

- **DROP:** The DROP TABLE statement is used to drop an existing table in a database.

```
DROP TABLE table_name;
```

- **JOIN:** A JOIN is used to merge rows from two or more tables, based on a related column between them. Besides, there are "INNER JOIN", "LEFT JOIN", "RIGHT JOIN", and "FULL JOIN". The LEFT JOIN keyword returns all records from the left table, and the matched records from the right table. If there is no match field, the right side is going to be occupied by NULL. The RIGHT JOIN can get the reverse result of the LEFT JOIN. As for FULL JOIN, it returns all records when there is a match in either left or right table records. INNER JOIN syntax is shown below and the code for other JOIN is only different in key words.

```
SELECT Orders.OrderID, Customers.CustomerName, Orders.
    OrderDate
FROM Orders
INNER JOIN Customers ON Orders.CustomerID=Customers.
    CustomerID;
```

- **WHERE:** The WHERE clause is used to filter records.

```
SELECT column1, column2, ...
FROM table_name
WHERE condition;
```

- **IN:** The IN operator is used to specify multiple values in a WHERE clause.

```
SELECT column_name(s)
FROM table_name
WHERE column_name IN (value1, value2, ...);
```

- **ORDER BY:** The ORDER BY keyword allows to sort the returned table in ascending or descending order.

```
DROP TABLE table_name;
```

- **GROUP BY:** The GROUP BY statement is often used with aggregate functions such as COUNT, MAX, MIN, SUM, and AVG to group the returned table by one or more columns and and get the math result.

```
SELECT column_name(s)
FROM table_name
WHERE condition
GROUP BY column_name(s)
ORDER BY column_name(s);
```

- **COUNT, SUM, MAX, MIN, AVG:** These aggregate functions can calculate the corresponding result of multiple rows. They are often use with GROUP BY and WHERE statement.

  The COUNT() function returns the number of rows that matches a specified condition.

  The SUM() function returns the total sum of the content of rows that matches a specified condition.

The Max/Min() function returns the max/min value of the content of rows that matches a specified condition.

The AVG() function returns the average of the content of rows that matches a specified condition.

```
SELECT COUNT(column_name)
FROM table_name
WHERE condition;
```

- **UNION:**The UNION operator is used to combine two of more tables with the same structure.

```
SELECT column_name(s) FROM table1
UNION
SELECT column_name(s) FROM table2;
```

Appendix B

SQL Code example

In this appendix, part of SQL files is listed. The SQL code is divided into three files and header of these files are listed first.

B.1 SQL code file 1

This file is used for aggregating the data for 'Time period distribution' (section 4.4.1).

```
/*
    Name:           Personal customer distribution _ product
       distribution of a year
    Author:         Jiangsu Du
    Written:        01/07/17
    Purpose:      The code collects the data of personal
       products(Saving, Current, Loan, Mortgage, and Credit
       Card)from 04/2016 to 04/2017. Firstly, it collects the
       information of credit card. Secondly, it collects the
       information of other four products. Then use 'UNION'
       operation to merge them together. Group it by customer
       to know the holding condition of each customer. Add
       features to customers, group them by their features, get
        the sum of each product.

    Edit History:
*/
```

Part of the code is listed below:

```
--Create a function table for later selection
--Because I dont have the right of writing fields into
   database, I use 'per_kpi_acct_t' to create a new volatile
   table. And use it to attach time to rows, duplicate a record
    into 13 records.
CREATE VOLATILE TABLE func_12
```

```
AS
(
SELECT
vperiod
FROM BAC_EDW_DM_TAC.per_kpi_acct_t
WHERE vperiod>='2016-04-01' AND vperiod <= '2017-04-01' AND
   ACCT_ID='00000000' --a specific row which satisfy
   requirements
)
WITH DATA AND STATISTICS PRIMARY INDEX (vperiod)
ON COMMIT PRESERVE ROWS;

--Select all savings, mta, loan, and mortgage accounts.
   Mortgage also exists in this table. The mortgage snapshot is
    not used.
--drop table account_4time_overview;
--select custid, account_no from account_4time_overview group
   by custid, account_no having count(*)>1;

CREATE VOLATILE TABLE account_4time_overview
AS
(
  SELECT
  A.CUSTOMER_ID AS custid,
  A.branch_no,
  A.account_no,
  B.categrya,

  CASE WHEN A.account_open_date gt '1900-12-01' THEN A.
     account_open_date -    EXTRACT(DAY FROM A.
     account_open_date) + 1
  END  AS start_date
  FROM BAC_EDW_DM_TAC.TMIPRAC_PRMYCST_AC A
  JOIN
  (
    SELECT
    brand,
    categrya,
    acctype
        FROM bac_edw_dm_tac.LK_ACCOUNTS_V
    WHERE  personal IN ('Personal')) B
  ON A.account_type = B.acctype AND A.brand = B.brand
  WHERE PRMY_END_DATE='2099-06-30' AND RELATION_TYPE IN ('O','
     OJ') AND account_open_date <=  DATE'2017-04-01'
)
WITH DATA AND STATISTICS PRIMARY INDEX (custid,branch_no,
```

```
    account_no)
ON COMMIT PRESERVE ROWS;

CREATE VOLATILE TABLE account_5time_sum
AS
(
SELECT
A.custid,
A.snaptime,
SUM(A.loan_count) AS loan_count,
SUM(A.mortgage_count) AS mortgage_count,
SUM(A.saving_count) AS saving_count,
SUM(A.mta_count) AS mta_count,
SUM(A.cc_count) AS cc_count
FROM
(
SELECT
custid,
categrya,
snaptime,
CASE WHEN categrya = 'Loan' THEN 1 ELSE 0
END AS LOAN_COUNT,
CASE WHEN categrya = 'Mortgage' THEN 1 ELSE 0
END AS MORTGAGE_COUNT,
CASE WHEN categrya = 'Savings' THEN 1 ELSE 0
END AS SAVING_COUNT,
CASE WHEN categrya = 'Money Transmission' THEN 1 ELSE 0
END AS MTA_COUNT,
CASE WHEN categrya = 'Credit card' THEN 1 ELSE 0
END AS CC_COUNT
FROM account_5_addtime
) A
GROUP BY 1,2
)
WITH DATA AND STATISTICS PRIMARY INDEX (custid,snaptime)
ON COMMIT PRESERVE ROWS;
```

B.2 SQL code file 2

The result of 'SQL code file 1' can have the distribution of 13 months while this file can produce the distribution of only 1 month. At the beginning, this file is the test version of file 1, then it is combined with SQL code file 3 to get the table for 'Customer holding relationship' (section 4.4.2).

```
/*
```

```
    Name:            Personal customer distribution _
    Author:          Jiangsu Du
    Written:         01/07/17
    Purpose:       This file has to rely on 'Customer
       distribution_a particular month.sql' since it needs the
       table account_5_sum. It groups records by the number of
       each product.  Then Union these tables into one table.

    Edit History:
*/
```

Part of the code is listed below:

```
--select savings, loan, mortgage, and mta accounts which are
   running by a particular month. and merge accounts owned by a
    customer together.
CREATE VOLATILE TABLE account_4_sum
AS
(
SELECT
custid,
SUM(loan_count) AS loan_count,
SUM(mortgage_count) AS mortgage_count,
SUM(saving_count) AS saving_count,
SUM(mta_count) AS mta_count
FROM
(
  SELECT
  A.CUSTOMER_ID AS custid,
  A.branch_no,
  A.account_no,


  CASE WHEN B.categrya = 'Loan' THEN 1 ELSE 0
  END AS LOAN_COUNT,
  CASE WHEN B.categrya = 'Mortgage' THEN 1 ELSE 0
  END AS MORTGAGE_COUNT,
  CASE WHEN B.categrya = 'Savings' THEN 1 ELSE 0
  END AS SAVING_COUNT,
  CASE WHEN B.categrya = 'Money Transmission' THEN 1 ELSE 0
  END AS MTA_COUNT
  FROM BAC_EDW_DM_TAC.TMIPRAC_PRMYCST_AC A
  JOIN
  (
    SELECT
    brand,
    categrya,
```

```
    acctype
        FROM bac_edw_dm_tac.LK_ACCOUNTS_V
    WHERE  personal IN ('Personal')
        ) B
  ON A.account_type = B.acctype AND A.brand = B.brand
  WHERE PRMY_END_DATE='2099-06-30' AND PRMY_START_DATE <=
    DATE'2017-04-01' AND RELATION_TYPE IN ('O','OJ')
  ) C
  GROUP BY 1
)
WITH DATA AND STATISTICS PRIMARY INDEX (custid)
ON COMMIT PRESERVE ROWS;
```

B.3 SQL code file 3

This file is used for aggregating the table for 'Customer holding relationship' (section 4.4.2). And it has to rely on SQL code file 2.

```
/*
    Name:           Personal customer distribution _ product
       distribution of a particular month
    Author:         Jiangsu Du
    Written:        01/07/17
    Purpose:      This file only create the product
       distribution of a particular month. Firstly, it collects
        credit card data. Second, it collected the data of
       other four accounts. Thirdly, Merge the two parts
       together and group by customer. Then, add features to
       customers. Group records by features and get the sum of
       each product,

    Edit History:
```

Part of the code is listed below:

```
--This file should be run after 'account_5_sum table'  of
   Customer distribution_a particular month.sql 'having been
   created.
--This analysis is to group customers with the same quantity
   of a particular account. Eache of the first 5 tables selects
    customers with corresponding kind of account and the same
   quantity together.
--The last table is to combine five tables together.
CREATE VOLATILE TABLE  savings_distribution
AS
(
```

```
SELECT
saving_count AS num,
'SAVINGS' AS category,

SUM(loan_count) AS loan_count,
SUM(mortgage_count) AS mortgage_count,
SUM(mta_count) AS mta_count,
SUM(cc_num) AS cc_num,
0 AS saving_count,

SUM(CAST(loan_count AS FLOAT))/COUNT(CAST(loan_count AS FLOAT)
   ) AS loan_average,
SUM(CAST(mortgage_count AS FLOAT))/COUNT(CAST(mortgage_count
   AS FLOAT)) AS mortgage_average,
SUM(CAST(mta_count AS FLOAT))/COUNT(CAST(mta_count AS FLOAT))
   AS mta_average,
SUM(CAST(cc_num AS FLOAT))/COUNT(CAST(cc_num AS FLOAT)) AS
   cc_average,
0 AS saving_average

FROM account_5_sum
GROUP BY num
)
WITH DATA AND STATISTICS PRIMARY INDEX (num)
ON COMMIT PRESERVE ROWS;

--drop table savings_distribution; drop table cc_distribution;
    drop table mortgage_distribution; drop table
   loan_distribution; drop table mta_distribution;
--drop table Direction_5_merge
CREATE VOLATILE TABLE Direction_5_merge
AS
(
SELECT * FROM cc_distribution
UNION ALL
SELECT * FROM mortgage_distribution
UNION ALL
SELECT * FROM  savings_distribution
UNION ALL
SELECT * FROM mta_distribution
UNION ALL
SELECT * FROM loan_distribution
)
WITH DATA AND STATISTICS PRIMARY INDEX (category,num)
ON COMMIT PRESERVE ROWS;
```

Bibliography

[1] Teradata.com. (2017). *Business Analytics, Hybrid Cloud & Consulting — Teradata*. [online] Available at: http://www.teradata.com/ [Accessed 13 Aug. 2017].

[2] En.wikipedia.org. (2017). *SQL*. [online] Available at: https://en.wikipedia.org/wiki/SQL [Accessed 13 Aug. 2017].

[3] SAS VA(2017). *Analytics, Business Intelligence and Data Management* [online] Available from: https://www.sas.com [Accessed 13 Aug. 2017].

[4] SAS VA support(2017). *SAS VA SUPPORT* [online] Available from: https://www.support.sas.com [Accessed 14 Aug. 2017].

[5] Google Map Heatmap Layer. *Google Map JavaScript API* [online] Available from: https://developers.google.com [Accessed 11 Aug. 2017].

[6] Google Map Fusion Table Layer. Google Map JavaScript API [online] Available from: https://developers.google.com [Accessed 10 Aug. 2017].

[7] CRISP-DM. *CRISP-DM— Wikipedia*. [online] Available from:https://en.wikipedia.org [Accessed 12 Aug. 2017].

[8] Collinson, P. and Finch, J. (2017). *St Albans and Hull: tale of two cities is uncovered in tax returns.* [online] the Guardian. Available at: https://www.theguardian.com [Accessed 6 Aug. 2017].

[9] Google Map Screen Shot. *Google Map* [online] Available from: http://google.maps.com/ [Accessed 7 Aug. 2017].

[10] Sloan, Terry, Data Analytics with HPC - Cleaning Techniques. ËPCC, 2017.

[11] Edwin de Jonge and Mark van der loo, 2013. An introduction to data cleaning with R. S̈tatistics Netherlands Discussion Paper.

[12] SAS VA Document(2017). *SAS VA document* [online] Available from: http://support.sas.com/ [Accessed 3 Aug. 2017].

[13] Rodriguez, J. and Kaczmarek, P., 2016. *Visualizing Financial Data.* John Wiley & Sons.

[14] Lam, Heidi and Bertini, Enrico and Isenberg, Petra and Plaisant, Catherine and Carpendale, Sheelagh, 2011. Seven guiding scenarios for information visualization evaluation. l̈am2011seven.