Saimaa University of Applied Sciences
Faculty of Tourism and Hospitality, Imatra
Degree programme in Hotel, Restaurant and Tourism Management

Aleksandra Bogdanova

Customer Database Creation and Analysis for Small-Sized Tourism Agency in South Karelia

Thesis 2018

Abstract

Aleksandra Bogdanova

Customer Database Creation and Analysis for Small-Sized Tourism Agency in South Karelia, 53 pages, 1 appendix
Saimaa University of Applied Sciences
Faculty of Tourism and Hospitality, Imatra
Degree Programme in Hotel, Restaurant and Tourism Management
Thesis 2018
Instructors: Ms Jaana Tanhuanpää, Lecturer, Saimaa University of Applied Sciences

The objective of this constructive research was to create a low-budget customer database for small-sized tourism agency to optimize company's data management, working routine, customer service, and to gain knowledge from customer data to create a competitive advantage, and to meet the provisions of General Data Protection Regulations of the European Union.

The theoretical part of the thesis was supported by various sources like books, articles, regulations from the law, the Internet and the author's personal experience. The raw data for the empirical part of the thesis was provided by the partner company. The empirical part was naturally divided into data processing, the creation of a database, and following data analysis.

As a result of this thesis, a database with records of almost 6,000 customers and nearly 10,000 services provided was created. Analysis of customer data produced on its basis was visualized in several graphs. The thesis formed an understanding of the importance of careful and structured storing and processing of customer data, as well as its marketing benefit. The database with the results of the analysis and the basic rules of data storing were handed out to the tourism agency.

Keywords: Customer Relationship Management, customer database, customer analytics

Table of contents

Appendices

1 Introduction

This thesis performs the creation of a database as a budgetary and accessible tool for independent optimizing of the working process of a small tourism agency in South Karelia. In particular, helping management to optimize customer data handling of the company. Researcher will use visa application forms of the agency's clients for 6 years of work as a basis for the database.

In conditions of tough competition, small-sized companies suffer the most. Shortcomings in routine organization of company could lead to various risks in future, reduce competitiveness of company constantly and demolish company’s visibility in the market. The purpose of the thesis is to show how efficient data usage could help small business like tourism agency in its daily routine and marketing, thus creating a competitive advantage.

This thesis is particularly relevant considering General Data Protection Regulation that came into force in European Union on 25th of May 2018. Creating a database is crucial for the tourism agency to be able to comply with the provisions of this law.

1.1 Main objectives of the thesis

The main objective of this work is to create a data management tool for the tourism agency, based on the customers’ information that company owns in connection with the peculiarities of the workflow. During placement in case company, the author identified imperfections in everyday activities of the enterprise such as customer data handling in particular.

The author believes there is a common solution for data management problems of tourism agency, which will optimize the work of the enterprise and consequently allow conducting analysis of customers’ data in the future. According to the researcher's observations, there are imperfections and difficulties in the work of the travel agency that management started to take for granted and perceive as a norm. Following features of the working process can be considered as imperfections: increased size of paper documentation, insufficiently secured storage of data, unjustifiably increased amount of storage space, no possibility to address to client’s data if they had already used company’s services. Among the goals of creating an automated solution to the aforementioned problems, consideration should be given to reducing the processing time and obtaining efficient data to make management decisions, increasing the reliability of information processing, and increasing the amount of analytical data. Among the directions of the organization where new technologies are important, the following could be distinguished: decision-making, strategic planning, marketing and communications, customer service and interaction with consumer, financial management, research and development. Moreover, using database itself does not require any marketing education.

The secondary aim is to display through data analysis how much knowledge customer information could contain and why collecting customer data should be an essential part of everyday routine of the staff. In the case of particular tourism agency, this thesis work will show how small businesses can benefit from information they possess without having recourse to third-party organizations.

Taking into account that the company management does not have special knowledge in the field, as well as the fact that the researcher does not have enough experience in working with this kind of projects, the result of this work will be aimed mainly at creating a working foundation for further development of data management in the company.

1.2 Presentation of a partner

The concept of this research study matured during 3 months of deep involvement into the working process of an organization. This company is a small-sized tourism agency with 16 years of operation. It is based in South Karelia region next to the Finnish-Russian boarder. South Karelia region is considered to be the second tourism region in Finland largest by popularity. (YLE Uutiset 2018.) Office is situated in Lappeenranta in a periphery area of the city.

According to the YTJ (2018) website, the company was registered in trade register in 1999 as a branch of the Non-State Educational Institution based in Russia. In 2001, the company was incorporated in tax register and in 2002 started its activities. At the present, staff of the agency consist of one permanent worker (manager), one temporary worker (call-up) and from 0 to 2 trainees. It is a usual practice for this company to provide placement and internship for students and workers from both Imatra and Lappeenranta. Accounting services are outsourced. The annual turnover of the company, according to YTJ (2018) is from 200,000 to 400,000 euro.

There are standard services as booking cottages and hotels, creation of group and individual programs, booking tickets for international trains and ferries. However, agency is mainly aimed at document preparations for obtaining visas to Russia, as the most demanded touristic service in the region. Practice of traveling to Russia for gasoline, alcohol, cigarettes or food is common for residents of boarder region. It is also fairly common to travel to Russia for leisure, for instance, to St. Petersburg for music performances or hockey games. The experience of the internship indicated as well that there is a category of clients traveling on business or employment issues. For all cases, a visa is required. This service is advertised on the company's website and there is an online store where one can pre-order the visa. Visas are divided according to the duration and terms of readiness. Company offers visas from 1 to 3 years of duration, single-entry visas and group visas. The main clients of the company are Finnish citizens, but it also serves citizens of other countries like Russia, Latvia, Estonia, France, etc.

1.3 Research method

1.3.1 Constructive research

The research method in this thesis is a constructive research. This is a research where construction or design is the cornerstone of work. This could be, for instance, a product, software, concept or some solution by means of which the main knowledge is indicated (Koskinen, Zimmerman, Binder, Redström & Wensveen 2011, p.5). Constructive research is driven by researcher’s own experience of the company in case and by theoretical issues, while addressing both of them. The main aim of research is to create a solution to the case problem from scratch through practice with supplementing theory. This kind of research typically consists of theoretical part which indicates the problem, constructive part which provides solutions and new knowledge and part which validates those solutions. (Oyegoke 2011, p.576.) In this thesis, theoretical part of a research provides a background for work, constructive part is presented by data processing and database creation, while data analysis would represent a validation constituent.

Due to the fact that constructive research may contain elements of other types of research, the theoretical part will include an interview with the Software Engineer and analyst from EPAM Systems, who will advise the author on data analytics methods, as well as support the analysis in the R programming language, which the author of the research began studying specifically for thesis.

1.3.2 Research questions

The identified research problems are used to propose research questions that address the problem (Oyegoke 2011, p.576). To meet the objectives of this thesis, author should give answers to the following questions:

- How could a small tourism agency with limited budget optimize the data management to get a competitive advantage?
- What problems with routine data handling of tourism agency a database could solve?
- How could a small tourism agency use the data it possesses to make a basis for marketing actions?
- What knowledge could tourism agency gain from the existing customer data?

Answers to these questions will form the core of this thesis, as well as shed light on the phenomenon of using customer data for the benefit of small companies.

1.4 In brief

In the first part of the thesis author will describe the basic concepts involved in the creation of a database. The researcher will talk about the origin of customers' data in the tourism agency, range of problems concerning its storage and use in the company. Furthermore, theoretical part will describe the new regulations on the protection of personal data and its business impact. In addition to the technical element, the author will describe database in terms of business and marketing.

Practical part will be devoted directly to the creation of a database, the processing of clients' information, the identification of features important for a particular tourism agency database. Beyond database creation, researcher will analyze the data to demonstrate how the proposed solution will work and how useful structuring of information is.

The last part of the thesis will conclude the study. The author will mark the conclusions that follow from the database creation and data analysis, thereby confirming or disproving the benefits of maintaining a database in a small travel agency.

2 Data

2.1 Data

Data is what underlies any database. Depending on the field, interpretation of data may have slight differences. In common, data is anything a person can turn into information and make a use of for its purposes. For example, in tourism or customer service it could be demographic data (age, gender, place of birth), contact information (phone number, address, email address), etc. Davidson (1996, pp.4-5) defines data as organized, informative, gathered and concerned with the topic of interest. Katz & Katz (1997) mention that data in use must be justified and relevant to company's strategy. Whether data is informative or not depends on the field and purpose of research. It is worth mentioning that different sources interpret the concepts of "data" and "information" in different ways, in many cases attributing clearly opposite definitions to them. In this thesis, author addresses to the concepts of “data” and “information” interchangeably.

2.1.1 Customer data of the tourism agency

It is important to determine the most useful source of information in the tourism agency to create a database that will make benefit for the company. The customer data of the agency is represented by two main sources, not to mention e-mails, photos and copies of the documents. In this company, applications for Russian visas are the most informative source of information, as well as the second largest by number of items after bills and invoices issued to customers. Both of these sources can serve as a good basis for analyzing a business, but only the first can qualitatively affect the agency’s working cycle, as it is involved in and affects a much larger segment of the workflow than bills.

At the same time, accounting documents are much more reliable sources of information, since the tax purity of a company directly depends on them. In contrast, attitude for processing applications may be less responsible, since employees could make changes even at the very last stages, directly in the visa application center. In this regard, visa application forms can contain various typos and reflect reality badly. In addition, trainee errors often lead to incorrect data storing, which means that the list of visa orders in the company may be either incomplete or redundant.

Paper copies of applications and related documents are stored for a long time at the agency as an extra source to determine the cost of the service, date of issue, contact data, and so on. There are no special rules on the disposal of these papers – management destroys them when their quantity becomes too excessive. Information about the invoices is stored separately on paper, customer contact information is stored separately from the invoice data, the invoices themselves are stored separately from the first and latter. Constant addressing between stacks of different documents, which are frequently situated in different parts of the office, may seem accustomed to management, but absolutely unjustified and confusing for outsiders, including new employees or interns.

The existence of a large amount of paper documentation significantly complicates the process of training and the introduction of new employees or trainees, thus leading to disruptions in service. In turn, this often leads to increased workload of the manager, who is forced to recheck all the documents that pass through subordinates. Moreover, it is worth remembering that any paper documentation is at high risk of corruption, whether it is spilled coffee, broken water-supply pipes or a fire.

2.1.2 Visa applications

In general, application is a one to two page document containing questions of a demographic, occupation, contact and touristic nature, the answers to which are necessary for the Consul to make a decision. Personnel of agency (or customer) fills application in directly on the website of the consulate or, as it was before, in paper form. Over time, due to changes in the procedure for issuing visas, the content and appearance of the applications change. Employees of the tourism agency save them in pdf-format and store in the appropriate directory on the computers and cloud service. This is necessary in order to further print the applications and submit them to the visa center or Consulate along with the originals of the client's documents.

2.1.3 Origin of data in the applications

Official requirements of the Consulate of the Russian Federation condition the contents of information in the visa application forms. The Consulate regulates the rules for obtaining visas, as well as sets the time for examination of documents and the size of consular fees. There are several consular departments for processing Russian visas in Finland. Appointments for consular issues in Helsinki and Turku can be made online or by phone. Making an appointment to the consular department of Lappeenranta is impossible nowadays, which puts the tourism agencies of the region in a favorable position. (Russian Visa Application Center in Finland 2018.)

Tourism agencies offer Russian visa services, making it easier for customers to collect the necessary documents. Russian consulate issues tourist and business visas, group visas, student visas, single entry and multiple-entry visas for citizens of different countries. Regardless of the visa category, applicant must submit the following documents to the Russian consulate (or travel agency dealing with visa issues):

- Passport - a valid passport with at least two empty pages. The validity of the passport should not expire earlier than six months after the expiration date of the visa.
- Application - filled in by a foreign citizen or a travel agency on the official website of the consulate with personal signature of the applicant.
- Photography - meeting certain rules and not older than 6 months. Case company provides this service for free.
- Invitation - a document from the receiving party (the official company's letter of invitation or the Russian Federal Migration Service invitation). The need for this document is another advantage of travel agencies that take responsibility for receiving invitations.
- Insurance - which covers medical and repatriation costs throughout the validity period of the visa. (Russian Visa Application Centre in Finland 2018; Consular Department of Ministry of Foreign Affairs of the Russian Federation 2018)

Consequently, at the disposal of the travel agency there are such personal data of the clients as: the passport number, validity of the passport, citizen's ID, previous visas, insurance company, insurance number, home address, telephone number, past citizenship of Russia (if any), purpose of the trips, visa category, dates of planned and dates of previous trips, destinations, family ties in Russia, name, surname, and past names, gender, date of birth (age), place of birth, employment, place of work or study, position. This is not just a very sensitive data, but also marketing-relevant information.

2.2 General Data Protection Regulations

As was mentioned in the previous subchapter name, date of birth, passport details, home address and telephone, purchase and travel history are the examples of sensitive information. Gaps in handling customer data can lead to unpleasant consequences. Leaks, for instance, can not only lead to undesirable and intrusive marketing from companies that one has never even dealt with, but lead to crimes of varying severity - from house thefts, customer manipulations to serious frauds. In this regard, on May 25, 2018 in the territory of the European Union and the European Economic Area, a new law on the protection of personal data came into force.

New regulations concerning data collection, its utilization and distribution were set in 2016 in the EU. Earlier, companies' preparations for the implementation of new law were not visible for an average resident. The scandal about Cambridge Analytica and use of personal data in social media to influence the presidential elections in the United States of America naturally launched a series of events of general indignation and anxiety about safety of personal data. Talks about the protection of personal data have gone beyond the political halls and head offices of large corporations. The battle for the safety of personal data became more than just one of many formalities, it became a personal concern of everyone who ever left a digital footprint - first in America, and then around the world. In this regard, the topic of personal data protection appeared in the center of worldwide discussion. Starting from 25th of May 2018 every company, that implies managing a personal data of EU citizens – online-shops, banks, social media, tourism agencies, airlines, internet providers and many others – have new list of responsibilities. At the same time, this directive gave specific rights to everyone who have ever left a digital footprint. Main positions of new law are, in brief:

- Duty to request an agreement of the customer for collection and utilization of personal data.
- Duty to provide an opportunity to reject this agreement at any time.
- Duty to make policy of confidentiality simple and clear for reader without any specific juridical knowledge.
- Duty to request a separate agreement to use data for marketing actions.
- Fines for violation of the GDPR – up to 20 million dollars or 4% of profit depending on what is bigger.
- Duty to inform customers about data leaks in 72 hours after it became known.
- Right of the user to know the duration of personal data storage.

- Right to request company to share one’s personal data with competitor one is going to leave to
- Right to request and get all the copies of personal data stored and get for free.
- Right of the customer to request deletion of his personal data without court decision (‘right to be forgotten’). (Kalmikov 2018; Regulation (EU) 2016/679.)

While the company does not have any understanding of what information they store, where they store it, about whom and for how long, without being able to access this data on the first request, the tourism agency cannot fulfill the requirements of the new directive. The creation of a database increases the safety of data already available to the company, as well as make it possible to embody the rights of the client to receive a copy, the deletion or sharing of the stored personal data. Current data management of tourism agency does not provide this opportunity. Although, taking into account the aforementioned rules it is now relatively difficult to collect data to use it for marketing purposes as separate permissions of a client are now required, there are marketing solutions. With the introduction of the new directive, many companies, for instance Google, have created their user agreements in such way that in case user wants to use the services of the company, he or she has to accept the agreement. If the user is against it, he or she will not be able to use the services.

3 Database

There are two ways to define databases in this research case. The general definition of database shows the technical side of phenomena, while marketing and customer relationship management reflects the practical benefits of using databases in business.

Stephens (2009, p.5.) defines database as any tool that meets CRUD principles to some extent. These principles describe whether database can Create data, how well it could be Read or Updated, and whether data can be properly Deleted. This abstract definition may refer to such kinds of databases as, for instance, paper notes, libraries, archives, timetables and, among other things, the human brain. Those are called “physical databases” and have the list of flaws that particular computer database can solve.

Most of those databases mentioned are adjustable and suitable for the time being. However, over time the content of databases grows so much that the search for necessary information becomes almost impossible. It becomes significantly time consuming and structuring of this information is only a temporary measure. In addition, the human brain is prone to forget and distort data stored in memory. (Stephens 2009, pp.6-8.)

Database management system or simply DBMS is software or tool to manage databases (Stephens 2009, p.490). It could also refer to the interface that helps users operate within databases. In this thesis “database” (if not specified) means a computer database of relational model (RDM) or relational database management system (RDBMS).

3.1 Database design

The design of databases directly appeals to the goals of creating a database in each individual case. Any system is a set of interconnected segments. Disadvantages in one or more of the components put the entire system at great risk. In large projects, minor deficiencies can be invisible or even ignored at the beginning of the system operation. However, over time one error leads to a whole cascade of new problems. Ultimately, this makes the system absolutely inadequate and unhelpful. Neglected deficiencies are very difficult to fix, and often it is much easier to create a new system instead of renovation. This in turn leads to significant budgetary and time waste, and equally importantly the loss of numerous important data, that accumulated during the operation of the imperfect system. In order to avoid this there is a design. Database design not only helps to create a database that will bring maximum benefit to the enterprise, but also gives some guarantees for the future. (Stephens 2009, p.4).

3.2 Principles and advantages of database

In addition to the marketing benefits of creating a database, which will be described in the next chapter, there are a number of technical advantages and common database principles. As was mentioned earlier, any database must comply with the CRUD. CRUD in its terms partially coincides with the main features (at the same time advantages) of a good database. As many authors of database literature there are, as many minds of what the main features of a good database design are. Simon Allardice (2013) in his video courses for LinkedIn Learning Solutions describes 6 main problems that database exist to solve: *size, ease of updating, accuracy, security, redundancy and importance*.

3.2.1 Size

Size is the very first and simple feature of a database and at the same time the problem to solve. There is no serious necessity in possessing a database if the amount of data in the company is insignificant (Allardice 2013). However, it is hard to think of such an enterprise, especially in customer service and tourism with tons of loyalty programs, online shopping, thousands and thousands of different interactions with clients. Throughout the years of operations in connection with the peculiarities of the workflow, companies accumulate data of different types. At some point, storage solutions that work for small amounts of data reach the limits of efficiency. After hitting those limits, additional subproblems as, for instance, speed of data retrieval emerge. Moreover, it is not only hard to find information from long or divided tables, but technical specifications of the computer do not always meet the requirements for speedy use of voluminous spreadsheets. (Allardice 2013.) In the case company there are so many visa application forms, that it takes up to 5-7 minutes for the computer to display them in the folder.

3.2.2 Ease of updating

In case there is a spreadsheet or a physical database like notebook, it is impossible for several users simultaneously make changes into the file. Even with the use of cloud-services, employees will overwrite the changes of each other instead of updating it. (Allardice 2013.) Stephens (2009) uses term “easy error correction” which means that database should (and could) allow users not only to make changes synchronously, but allow to make massive changes such as fixing systematical typos in one or two steps. This is especially useful when company hires a large number of trainees. In case, for instance, when trainee during the whole summer placement types in “Laperanta” instead of “Lappeenranta” in the row “City”, database allows to change all incorrect entries at once.

3.2.3 Accuracy

Allardice (2013) notices that simple spreadsheets do not have any prevention features for users to type incorrect or incomplete data. This leads to the files full of unreliable data in the future. Stephens (2009) describes a similar concept under the definition of Validity. Computerized databases have protection mechanisms or validation features.

For instance, database has data type verification. Database will refuse user to type literal data into numerical row, or will reject to enter incomplete date of birth. Database is also able to verify entered data for presence of similar data in its different parts. For instance, the trainee of travel agency is trying to enter “Moscow” as a destination city of a customer, but accidently misspell and write “Maskow”. If necessary database will check the list for data resemblance and inform if there are no similarities found. User can do even more and limit the list of possible destinations for data entries. In case there are only Moscow, Petrozavodsk, Saint-Petersburg and Vyborg on the list, database will reject any different or incorrect data. Moreover, it is possible to forbid leaving blank spaces in the database. (Allardice 2013.)

3.2.4 Security

Companies store consciously or not a lot of information, which differs in the degree of informativeness, importance, sensitivity and in nature. The loss of some information may not bring the company any difficulties, while other information entails not only the loss of important data for the work and analysis, but entail much more serious problems with law too. Business-to-Consumers companies often store very sensitive data about their customers, such as passports, insurances, addresses and so forth.

Moreover, some data, not being sensitive by itself, in conjunction with other data carry a great potential for frauds from the outside. In the case of a travel agency, the client's vacation trip dates do not carry any specific information, but in conjunction with the home addresses of customers can serve as a basis for crimes such as home thefts. Physical databases such as notebooks, despite their high portability, have an increased risk of being stolen or corrupted. Spreadsheets are much more complicated to steal, but still not impossible. Database software allow user to control and restrict access to database parts. This software not only tracks who made changes and when, but also prohibit the ability to change or delete the data for individual users. In case when a criminal or a malicious employee wants, for example, to delete data they will not be able to do that if there is no access to this function. Unfortunately, budget database software often lacks this feature.

Stephens (2009) partially extends “security” phenomenon by describing the “Sharing” feature of the database. One more superiority of computer database over physical database is the ability of multiple users to work with database at the same time, which was mentioned also in “Ease of updating” subchapter. Separating access to the database into parts, which only particular users can work with, does not only offload the system, but also helps with security issues.

3.2.5 Redundancy

Allardice (2013) defines redundancy in simple terms as a repetition of the same data. The presence of duplicates in the data not only increases the storage space, but can also lead to various kinds of conflicts and inaccuracies in the use of this data. For instance, among the files of the tourism agency there are two identical records about the same person. Later, each employee makes changes to client’s data in different records instead of one. As a result, both entries become incomplete, inaccurate, unreliable and contradictory to each other. As an example, Allardice (2013) cites a situation in which a file ends with two records about the same product containing conflicting pricing information. For example, in the files of the tourism agency there are two entries about the same person. One of those says that invoice was set for 220 euros for a multiple-entry visa, and another that the invoice was set at 200 euros. Is the invoice actually for 220 or 200 euros? Or have there been 2 invoices for one person? Or were those invoices set for different customers? Such example is very closely related to the concept of consistency, which Stephens (2009) introduces. According to this concept, the database must carry consistent data and always perform the consistency of the results when searching. Besides the negative sides of redundancy, it is worth mentioning, that this phenomena is not negative itself and sometimes it is beneficial, for instance when it comes to database backups.

3.2.6 Importance

As was repeatedly indicated earlier, the data can be of varying degrees of sensitivity and importance. According to Allardice (2013), the importance of information is one of the six fundamental reasons for creating a database. In the researcher's own experience, failures in working with ordinary spreadsheets (in the same way as physical data carriers) happen occasionally. In some cases, especially when there were no backups for a long time, such failures can lead to a complete loss of critically important business information. Such incidents do not just lead to extra work and stress, but they are simply dangerous, since lost information can be invoices, contact details of customers who expect their visas tomorrow because of group trip to Saint-Petersburg. There are many more examples of information, the loss of which can cause material damage to business and sometimes irreparable harm to the reputation of the company.

3.3 Database model

There are several types or models of databases, each of which meets specific goals. The database model is responsible for the appearance of the database, methods of entry and retrieval of data, adaptability to changes and special features.

3.3.1 Flat files database

Common and comprehensible flat files database could be understood as a plain text file, lists, and spreadsheets with one record per line. Moving away from the concept of computer databases, the shopping list or phone numbers in the notebook are flat file databases.

These databases could be an intermediate stage in creating a database management system or tool for moving the data between different formats, for instance, a CSV (comma-separate values) files. As the name implies, CSV files contain records with commas as delimiters. When converting data table into the

CSV file, data will no longer be separated with columns, but with commas. (Tuffil n.d; Rouse 2018; Puckering 2018.) Researcher uses this file in the empirical part to transfer data to R-Studio for analysis.

One of the main disadvantages of flat files is an inability to make changes at the same time by several users. Taking into account that such database implies only one line per record, information about the ordered visa, its terms, type and cost, as well as information about the customer itself will be on the same line. This means that staff will have to input data about each new service provided to the same client on a separate line. This, in turn, creates duplicates, adds work to staff, increases the size of the storage space and makes the database search inconsistent. (Tuffil n.d; Rouse 2018; Puckering 2018.)

3.3.2 Relational database

The most commonly used and popular model is the relational data model, which is often used in the service sector (Stephens 2009, p.29). The author uses this model in this research work. The relational data model or RDM is a set of logically separated tables (relations) with data, allowing user to perform various data manipulations, make different specific queries without having to duplicate or reorganize the database, thereby simplifying the output of data from the database and opening up new opportunities for cross-comparison.

Database schema described in 3.4 subchapter determines the relations (tables) required for a particular database. Tourism agency could have, for instance, tables concerning customers, their orders, issued invoices and so on. In this case, database automatically assigns personal ID to each customer, called the primary key, which will serve as one of the main links between the relations (tables). The “contact data”, “orders” and “invoices” tables are linked to a table containing customer names through this key. This allows, for example, reporting all orders or invoices ever issued to a specific customer by his or her ID in the database in the future. Data separating into relations (tables) not only adds functionality to the base, but also optimizes its work.

3.4 Database schema

Working on any type of project implies a plan as a first step. Responding to the goal of creating a successful product, the plan helps to allocate resources and forces, sets the direction of work and establishes the necessary milestones. When the project is to create a high-quality database, such a plan is called a scheme. It defines the algorithm or rules of the database, thereby controlling the principles of input, storage and retrieval of information, as well as determining the basic interactions of the data. Schema is a database “portrait” of any creator, involved in the process. For instance, manager of the tourism agency and IT person have different “portraits” of what database consist of, while the perfect product is an interaction of those schemas. While the database design defines the basic features of the database, the schema defines the database functionality for a particular user, for example, establishes what relations are necessary in a particular database. (Hennig, Bradly, Linson, Purvis and Spaulding 2010, pp.6, 25-26.) The establishment of proper relations in the database allows increasing its performance, efficiency, the quality of reports and even the speed of task execution.

3.5 Database tools

The modern application market offers a great variety of programs for creating and manipulating databases. Most of them are solutions for large enterprises, while the cost of buying a license can reach thousands of euros. Such applications operate with millions of data records and require a separate staff for support and/or special knowledge and skills. (Hennig et al. 2010, p.115.)

3.5.1 Microsoft Access

Access is a DBMS and a part of Microsoft Office Software package. As with all types of databases, Access has its own advantages and disadvantages. The main disadvantages is that the access is not a free software. In addition, the Access is not properly adapted to the access distribution between different users of database. On the other hand, although Access is subjected to a fee, it comes in the same package with other Microsoft programs, which means that since the tourism agency uses such programs as Excel, Word and Power Point, they also have Access regardless whether they used it or not by default. This makes Access an available software. (Stephens 2009, pp.287-288). This plays a decisive role for research purposes of this thesis and author consider this option as an adequate choice to start the database management for tourism agency.

3.5.2 Other

Microsoft SQL Server and MySQL are widely used alternatives to each other. Although Microsoft SQL is great at linking with other products of Microsoft, as well as integrating with the cloud system, it only operate within Windows operational system and have both free and subscription versions (Lungariello 2017.) MySQL is an open-source DBMS and like other budgetary solutions lacks some of common DBMS features (Mullins 2015).

Very expensive and complicated for a tourism agency in question, but the most frequently recommended software in the world of databases are Oracle and IMB DB2. According to Lungariello (2017), being a system of high-performance with a large number of data manipulating techniques, Oracle requires considerable training because of complexity. IMB DB2 works well with many operating systems. It is considered one of the biggest competitors of the Oracle, and requires highly trained personnel too. (Mullins 2015.)

4 Data analysis

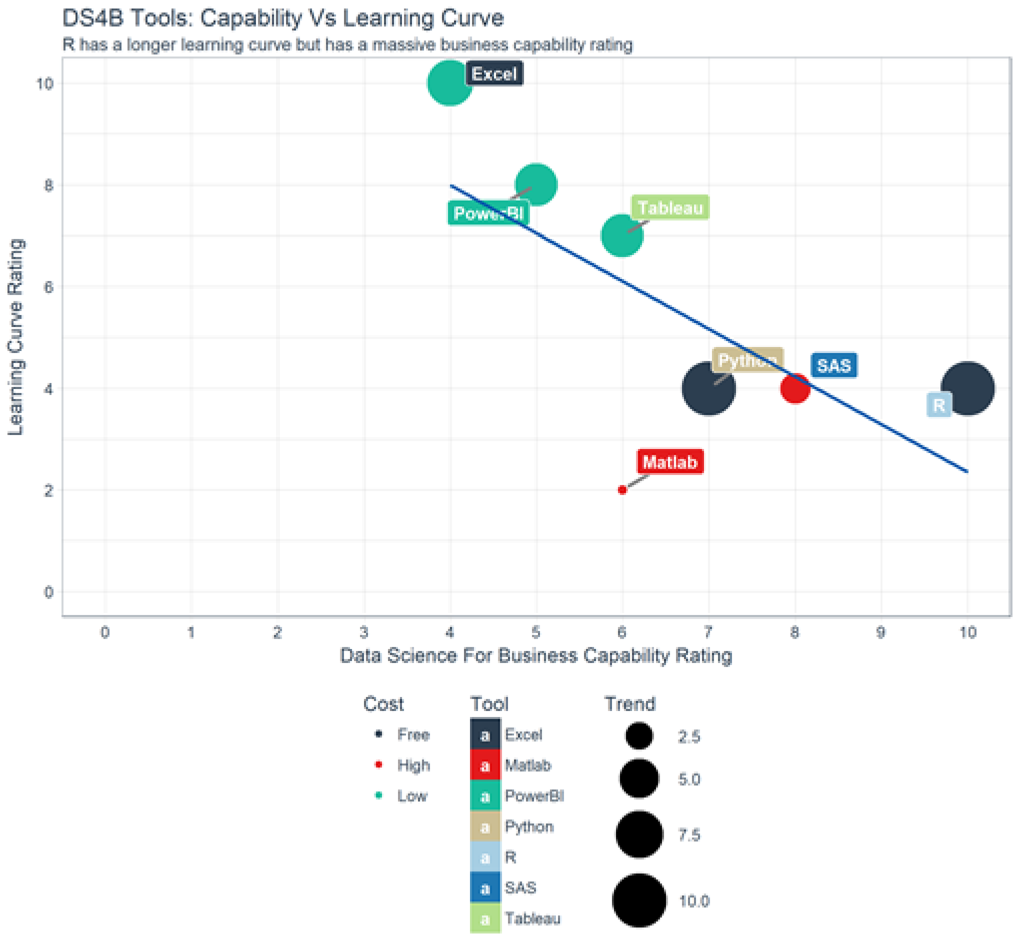
With the ability of more and better sources of data, decision makers must begin systematically incorporate such information into the decision process (Blattberg & Hoch 1990, p.887). There is almost no point in owning a database, if one does not get any knowledge from it.

4.1 Analysis tools

In order to demonstrate the exponent of the basic client information analysis is needed. Nowadays, there are many products to perform the analysis in software market. For some of them no special knowledge is required.

One of the most famous and popular is Microsoft Excel. Excel supports a number of software add-ons designed specifically for data analysis. In addition, Microsoft offers customer support and tutorials to help the user. Since Access is used to create the database in this research, it is doubly convenient to use Excel as a tool for analysis or to use the Microsoft Access's already built-in calculation feature. However, the author does not want to tie the analysis to the Office Suite of Microsoft as it is quite ordinary and familiar, while there are many other tools for analysis with their specific advantages.

Picture below (Graph 4.1) describes the relation between learning curve and business capability of popular tools for analysis, which include Excel, Matlab, PowerBI, Python, R, SAS and Tableau.



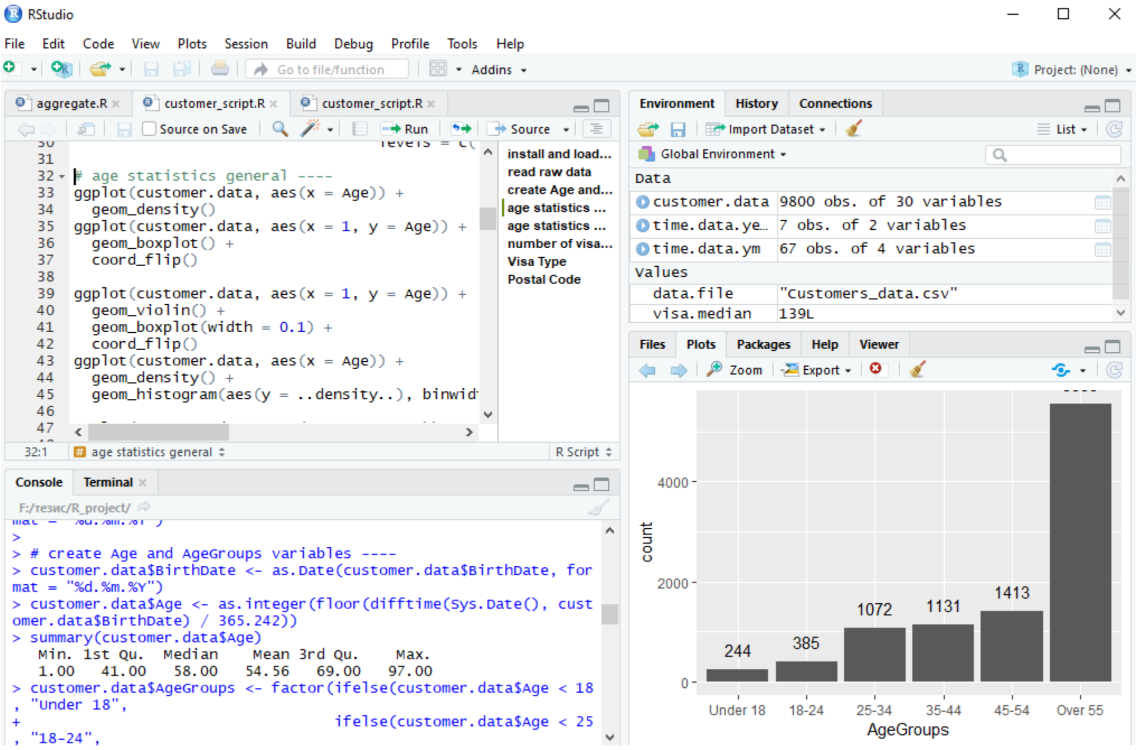
Graph 4.1. DS4B Tools: Capability vs Learning Curve (Dancho 2017).

Learning curve reflects the dynamics of success in learning while business capability reflects the tools' effectiveness for the business to perform its core functions. On the learning curve scale from 1 to 10 (where 0 is "hard to learn" and 10 is "easy to learn") Excel transcends all other tools, but at the same time is the lowest on scale of business capability (where 0 is “not effective” and 10 is “very effective”). Together with PowerBI and Tableau, also showing low business capability, they form top three most easily studied and at the same time low in price programs. All the programs remaining on the list are rather complicated in learning, and although they demonstrate a greater business capability, each of them was created to meet rather specific goals of the analytics. In addition, Matlab and SAS have a high cost, while Python and R are completely free and more likely than others to be in the trend. (Dancho 2017.) When comparing Python and R language Matt Dancho (2017) mentions that Python is created by software engineers and because of the field of usage lacks metrics, reports and graphic techniques necessary for business whereas R leads the market in its interactivity and topic-oriented packages which is the basis of its business capability. One of the main advantages of RStudio over Excel is that RStudio can fast process significantly large data sets.

4.2 R-programming language and R-studio

In order to draw a line under the choice of an adequate tool to analyze customer data, the researcher conducted an interview with the Software Engineer and at the same time analyst from EPAM Systems with more than 6 years of working experience in IT. Taking thesis goals and the author’s insufficient experience in analytics into consideration, the expert advised the R programming language as the most illustrative, relatively easy to study and very popular analytics tool. (Bogdanova 2018.)

R is a high-performance open source language to meet the needs of statistical analytics. Specialists and users often view R not just as a programming language or an analytics tool, but a whole environment — coherent, flexible, sufficient, and large. RStudio is the developing environment for R language that offers interface for working with R as well as powerful well-designed metrics visualizations - histograms, charts, clusters, modeling, and many more. (R-project 2018.) Figure 4.1 demonstrates the user interface of RStudio.



```
# age statistics general ----
ggplot(customer.data, aes(x = Age)) +
  geom_density()
ggplot(customer.data, aes(x = 1, y = Age)) +
  geom_boxplot() +
  coord_flip()

ggplot(customer.data, aes(x = 1, y = Age)) +
  geom_violin() +
  geom_boxplot(width = 0.1) +
  coord_flip()
ggplot(customer.data, aes(x = Age)) +
  geom_density() +
  geom_histogram(aes(y = ..density..), binwid
```

```
>
> # create Age and AgeGroups variables ----
> customer.data$BirthDate <- as.Date(customer.data$BirthDate, for
mat = "%d.%m.%Y")
> customer.data$Age <- as.integer(floor(difftime(Sys.Date(), cust
omer.data$BirthDate) / 365.242))
> summary(customer.data$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00   41.00   58.00   54.56   69.00   97.00
> customer.data$AgeGroups <- factor(ifelse(customer.data$Age < 18
, "Under 18",
+                                          ifelse(customer.data$Age < 25
, "18-24",
```

Data	
customer.data	9800 obs. of 30 variables
time.data.ye…	7 obs. of 2 variables
time.data.ym	67 obs. of 4 variables
Values	
data.file	"customers_data.csv"
visa.median	139L

Figure 4.1. RStudio interface with running script

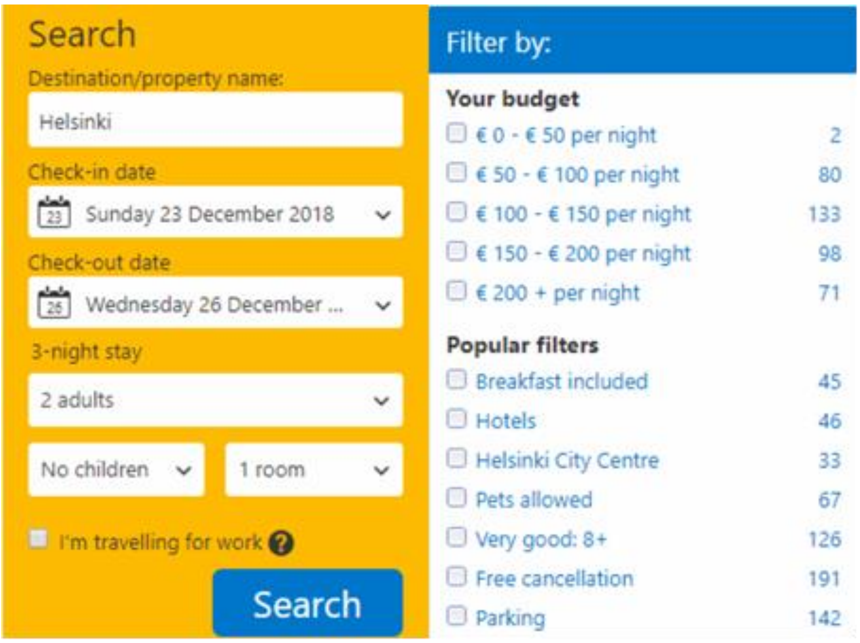
Due to the fact that R is an open source, free software language, it is supported by a vast community where one can contact other specialists and users on issues of interest, find ready-made and specific solutions for one’s challenges and tasks. Furthermore, this language is very extensive due to variety of packages that are created by both developers and common users. In addition, user in need could link R with other programming languages adding functionality and opening up new opportunities for work. (R-project 2018.)

5 Database marketing

Database marketing is the use of customer databases to enhance marketing productivity through more effective acquisition, retention, and development of customers (Blattberg, Kim & Neslin 2008, p.4).

Technologies are well-established in the lives of people, whether it is a business, private life or, for instance, medicine. It is difficult to underestimate the benefits of automation and digitalization in the service sector. The development of the Internet has given the service sector a new tool for interacting with customers, allowing the opening of digital trading platforms - online stores, booking services, and so on.

In online tourism services, one of the best-known examples of the database is Booking.com containing information on 28,989,662 accommodation units in 229 countries in 142,131 destinations worldwide (Booking.com 2018). Through the user interface of Booking.com (Figure 5.1), user can sort the data by categories upon one's individual preferences. In the direction to Helsinki, for 2 people during the Christmas holidays of 2018, Booking.com offers a list of 122 options to filter accommodation.



Search

Destination/property name: Helsinki

Check-in date: 23 Sunday 23 December 2018

Check-out date: 26 Wednesday 26 December ...

3-night stay

2 adults

No children

1 room

□ I'm travelling for work

Filter by:

Your budget	
□ € 0 - € 50 per night	2
□ € 50 - € 100 per night	80
□ € 100 - € 150 per night	133
□ € 150 - € 200 per night	98
□ € 200 + per night	71
Popular filters	
□ Breakfast included	45
□ Hotels	46
□ Helsinki City Centre	33
□ Pets allowed	67
□ Very good: 8+	126
□ Free cancellation	191
□ Parking	142

Figure 5.1. Interface of Booking.com

Among the options presented there are such filters as, for instance, budget, rating by stars, availability, special offers and discounts, things to do, meals, review score, etc. This helps business to better meet customer demand, apply individual approach and bring customer satisfaction.

Another illustrative example of using databases is the online shopping cart, which is a result of relational database operations. The website stores customers' id and products' id on the server. This serves to make up a shopping cart, to save it when switching between online store sections, and in some cases using additional tools such as "cookies" to save content of the cart throughout a particular time even after customer leaves the website. To check out from the online shop, contact details are necessary. Once the purchase is done, the order confirmation is sent to the customer. This usually contains order number, which is a primary reference for the store database in different situations.

In hospitality industry, hotels use specific databases in order to manage hotel bookings, sales and housekeeping. Opera PMS (property management systems) supported by Oracle is one of the examples.

5.1 Segmentation

One of the most logical examples of using databases in marketing is segmentation. Segmentation is the division of customers into groups with similar characteristics, buying habits and demands. Good segmentation allows the company to allocate those groups of clients that will create the greatest revenue, as well as more carefully and accurately meet their needs. Segmentation is needed primarily because it is impossible to satisfy the needs of all customers, being everything to everyone at once. Concentrating resources on those groups of people whose needs company’s resources can most effectively satisfy or reach, the tourism agency is able to become more competitive. (Morritt 2007, p.5-6.)

According to Morritt (2007) there are 10 segmentation types used in hospitality industry: geographical, demographic, purpose of trip, product, usage, brand loyalty, benefit, channel segmentation, lifestyle segmentation and organizational segments.

Geographical segmentation is a relatively cheap type and thus popular for hospitality. In this type, company uses locations to specify clients. A tourism agency can serve customers within a radius of 2 kilometers, some specific city districts, a city, region (for instance, South Karelia) or the whole country. Case company sometimes serves customers from abroad, which could also be a segment. Demographic segmentation separates customers according to their age, gender, nationality, marital status, income. It is common to use those two segmentation bases together to better serve the needs of enterprise and customers. (Morritt 2007, p.18-19.)

Purpose of trip segmentation usually refers to either business travelers, that are usually sponsored by their companies or leisure travelers who pay for themselves and prefer a more relaxed pace. Nowadays, it could also refer to specific groups of travelers such as, for instance, health and welfare travelers. Product segmentation usually implies dividing customers by price-product correlations. Examples include budget vacation packages, all-inclusive hotel offers or indecently expensive villas on Maldives. (Morritt 2007, p.20-21.) Neither of those segmentations is applicable for visa customers of travel agency. However, they still could efficiently fit to the rest of agency's services as, for instance, creating customized trip packages.

Segmentation by Usage specifies customers according to the knowledge that twenty per cent of customers produce eighty per cent of revenue, thus aiming at those who use the specific service more frequently. A similar type of segmentation is the division according to the degree of loyalty. (Morritt 2007, p.22-23.)

Benefit segmentation separates customers by the type of features they find important. For instance, those who would prefer hotels on the first-line of the beach or those who are looking for nightclubs. Channel segmentation divide customers by the service distributer, such as travel agencies. Lifestyle segmentation separates customers by their psychological profiles. And organizational segmentation implicates organizations, other companies, or corporate customers. This segmentation is applicable for group visas, when the whole company department goes on a two-days business ferry trip to Saint-Petersburg. (Morritt 2007, p.24-26.)

Morritt (2007) considers the existing customer database as one of the prime source to meet segmentation purposes of the company, revealing retention ratio, frequency of use of the services and many more features of customers, which could help to set the right segment and indicate matching customer profiles. The results of the analysis, which the researcher describes in the practical part of the thesis, could suggest the main segments of the travel agency.

5.2 Customer Relationship Management

Customer Relationship Management is closely related to the concept of database marketing, and even generally having a common goal. Nevertheless, these concepts still differ as they focus on different aspects of customer acknowledgment, while database marketing is one of the key parts of CRM of the company. The creation of a database for the tourism agency will set the foundation for the further installation of the CRM system. This system is not just a set of interrelated data, but also a tool to track all interactions with customers, often integrating with other applications, departments, social media, online stores and even a telephone. This system has advanced features and allows not only to read data, but also to manage it in a completely special manner. For instance, it allows management to set various reminders and announcements from "Make a data back-up!" to "This client has a birthday today". Some CRM Systems may also track all the incoming and outcoming phone calls not to lose a single client due to a missed call. One of the distinguishing features of CRM system is the ability to see and make records about interactions with a particular client. When the customer calls the company, CRM system is able to determine the number from the database and provide the employee with information about the customer and all his previous calls, queries, problems and offered solutions. Very high cost is the main disadvantage of CRM systems. (Salesforce 2018.)

5.3 Case examples

The simplest example of what database marketing may stand for in tourism agency is as follows:

- As part of the monthly work routine, manager sorts all customers in the database according to the visa expiry dates. To anyone whose visas are expiring in the near future (for instance, the current month + 1), manager sends the same offer to the e-mail or telephone. This could prevent customers from leaving for another company for the same service and consequently will increase customer retention.

- Sorting customers by both visa expiry date and the date of birth provides possibility for direct marketing – special offers or discounts for visa services during the whole birthday week.
- The manager can analyze and control seasonality of demand for his services, which makes it possible to balance profits by means of sales promotion, cross selling, upselling, and media advertising strategies.

The same examples from the CRM System’s point of view could be the following:

- As part of CRM System’s inner operations, manager gets the alert about everyone whose visas are expiring in the near month. Program helps to send customized attractive offer to those customers by the e-mail or telephone.
- As part of CRM System inner operations, manager gets a reminder to get in touch with everyone who has birthday in the near future to offer some extra benefits.
- As part of CRM System inner operations, manager gets alert if any changes happen in the buying behavior of customer, for instance, “Low sales. Forty-five customers fewer than in the same month of year 2017”.

6 Data converting and processing for the research

Despite the advantages of invoices as a data source, the author’s choice still fell on visa application forms, as they provide more opportunities for analysis and are more closely related to the routine work of staff. The partner company has provided a folder with 16,000 files for this thesis. Visa application forms in the company are presented by 1-2 page files of PDF format, which are stored in a shared folder. This data format does not imply either analysis or practical use. The search of the required application in the folder is complicated by the large number of files, which not only slows down the process, but often gives incomplete search results. About 14,000 of these files are the visa applications, and the remaining 2000 are various files stored in the folder by mistake, which also proves the need to create certain rules and the structure of data storage.

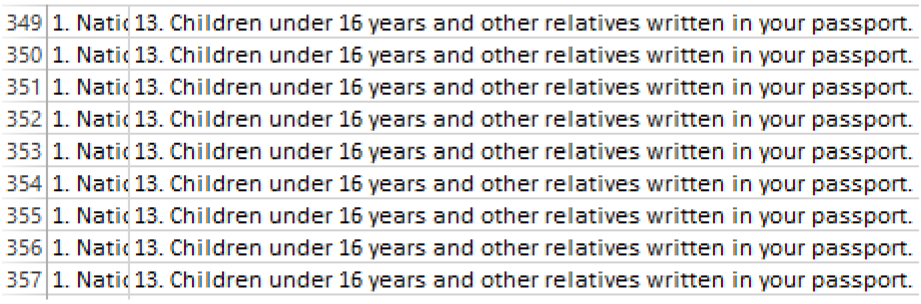
6.1 Data converting

Almost 10,000 applications out of 14,000 were taken for research work for several reasons. In order to be able to work with and manipulate data from the applications, researcher manually transferred data from PDF applications into Excel spreadsheets. The work was complicated by the fact that the appearance of applications changed during the storage of these data, and accordingly the data set in applications changed due to changes of Consulate policies. In addition, the way the computer copies data has changed. Concerning these changes, tourism agency possessed applications of 3 different kinds: from 2010 to 2012, from 2013 to 2015 and from 2016 to 2018. Author decided, that converting of all 3 types is unjustifiably time and power consuming, while it would not bring any specific benefits. Therefore, the data from the old type of applications (2010-2012) was not used for neither database nor analysis.

In the case of 3000 applications from 2016 to 2018, two-page files, the researcher was able to copy data sequentially to a line separated by the numbers of the sub-paragraphs from application as shown on the Figure 6.1. Later, these numbers and punctuation marks were used to separate data to the columns using “replace” and “text to columns” feature. These features were very widely used through the database creation process.

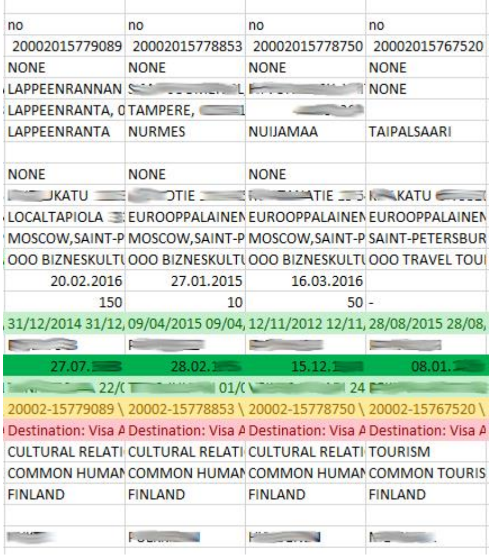
In the case of more than 6000 earlier applications (2013 – 2015), which were 1-page files, difficulties arose already at the stage of copying. Computer copied data to columns instead of lines as it was before, and the information lost sequence, which added work for manual processing. Moreover, the number of lines in the columns per each record differed due to differences in the amount of information provided for applications. For instance, due to lack of data management instructions towards applications, part of rows named “Previous names” contained “none” records, while some cells were left blank. As a result, columns containing customer data without previous names were shortened by one row, which led to inconsistency of data. In order to most effectively find the columns with extra or missing lines and to manually fix the discrepancies of rows that appeared in connection with this, some lines that were the same or similar for all applications were highlighted in bright color (Figure 6.2).

Researcher manually adjusted more than 6,000 columns into one format and then transposed the columns into lines. Author did this several times, since during the transposition missed errors and omissions became noticeable.



349	1. Natio	13. Children under 16 years and other relatives written in your passport.
350	1. Natio	13. Children under 16 years and other relatives written in your passport.
351	1. Natio	13. Children under 16 years and other relatives written in your passport.
352	1. Natio	13. Children under 16 years and other relatives written in your passport.
353	1. Natio	13. Children under 16 years and other relatives written in your passport.
354	1. Natio	13. Children under 16 years and other relatives written in your passport.
355	1. Natio	13. Children under 16 years and other relatives written in your passport.
356	1. Natio	13. Children under 16 years and other relatives written in your passport.
357	1. Natio	13. Children under 16 years and other relatives written in your passport.

Figure 6.1. Data transferring of applications (2016-2018) from pdf to Excel



no	no	no	no
20002015779089	20002015778853	20002015778750	20002015767520
NONE	NONE	NONE	NONE
LAPPEENRANNAN	[redacted]	[redacted]	NONE
LAPPEENRANTA, 0	TAMPERE, [redacted]	[redacted]	
LAPPEENRANTA	NURMES	NUIJAMAA	TAIPALSAARI
NONE	NONE	NONE	
[redacted]KATU	[redacted]OTIE	[redacted]ATIE	[redacted]KATU
LOCALTAPIOLA	EUROOPPALAINEN	EUROOPPALAINEN	EUROOPPALAINEN
MOSCOW,SAINT-P	MOSCOW,SAINT-P	MOSCOW,SAINT-P	SAINT-PETERSBUR
OOO BIZNESKULTU	OOO BIZNESKULTU	OOO BIZNESKULTU	OOO TRAVEL TOU
20.02.2016	27.01.2015	16.03.2016	
150	10	50	-
31/12/2014 31/12,	09/04/2015 09/04,	12/11/2012 12/11,	28/08/2015 28/08,
[redacted]	[redacted]	[redacted]	[redacted]
27.07.[redacted]	28.02.[redacted]	15.12.[redacted]	08.01.[redacted]
[redacted] 22/(	[redacted] 01/(	[redacted] 24	[redacted]
20002-15779089 \	20002-15778853 \	20002-15778750 \	20002-15767520 \
Destination: Visa A	Destination: Visa A	Destination: Visa A	Destination: Visa A
CULTURAL RELATI	CULTURAL RELATI	CULTURAL RELATI	TOURISM
COMMON HUMAN	COMMON HUMAN	COMMON HUMAN	COMMON TOURIS
FINLAND	FINLAND	FINLAND	FINLAND
[redacted]	[redacted]	[redacted]	[redacted]

Figure 6.2. Data transferring of applications (2013-2015) from pdf to Excel

In the next stage of work, researcher brought together the results of the conversion of two types of applications. This took some time, as the order of the data was still different. In addition, the applications from years 2016-2018 contained more columns due to a larger variety of data. In general, manual correction of almost 10,000 applications took 2 months, not including various revisions and rechecks. It is worth drawing attention that imperfections of raw data appeared until the end of work on the database.

It is worth noting that need of this complicated and time-consuming data transferring or converting happened because of poor data management in the tourism agency. In the case of an already existing database, employees of this agency would not have to undertake those actions. In such situation, the filling of the fields in the database would be primary while filling of the applications for the pdf format would occur using data from the database in the secondary stage.

6.2 Pre-processing of data

The converting of data from one format to another was only the initial stage. After the researcher transferred the data to a format that allows manipulations, the stage of preparing the data for transfer to the database began. At this stage, the author made decisions about what kind of information was important for the database and the appearance of this data.

6.2.1 Organization of columns

In order to ease further work with information, it was necessary to organize data logically. The latest visa applications served as a basis to organize the order of the columns. The data in the applications were sequentially joined into groups. There was demographic data of the applicant in the beginning. This data consisted of such sections as nationality, name, date of birth, passport data and place of birth. Next group contained information about the visa for which a customer submitted the application, for instance, visa type and duration. Following group described details about previous visas - the number of entries to Russia and the dates of last entry. In conclusion, there was contact information such as home address, telephone number, place of work and contact information of the employer, as well as data about relatives in Russia.

In the next step, the author divided specific columns into two or more. This was necessary in order to allow certain types of data sorting. For example, the columns responsible for the validity of the passport, the visa validity and the duration of the last trip to Russia contained both the start and the end dates. Using the previously used “text to columns” feature, the researcher reorganized columns so that the starting and expiry dates would stand separately. In addition, the author divided the column "Name" to make sorting and searching by last name possible, regardless of the first name. Moreover, the author separated the dates of changing the citizenship and the reason for these changes, which were previously in the same column.

Furthermore, the process of naming the columns is worth noting. There are specific rules on naming the data for further computerized manipulations. In this regard, the author designed column headings so that they did not contain spaces or any specific symbols, but were recognizable. The list of headings was as following: Nationality, DateCitizenshipChange, ReasonCitizenshipChange, Surname, Name, BirthDate, Gender, PassportNumber, PassportStartDate, PassportExpiryDate, VisaCategory, VisaPurpose, VisaType, VisaStartDate, VisaExpiryDate, Invitation, Destination, Children, HomeAddress, TelephoneNumber, PostalCode, Workplace, NumberEntries, LastEntryDateStart, LastEntryDateEnd, Insurance, PreviousNames, RussianRelatives.

6.2.2 Duplicates

A spreadsheet that resulted from data converting was a flat file database. As was mentioned previously, this means that there was only one record per line resulting in several records about the same persons that used visa services more than once. In addition to these records, there were absolutely identical records or records containing errors too. The author tried to get rid of some obvious duplicates already at the converting stage, but still detected more while sorting. In order to avoid accidentally deleting records of the same people who applied for the visa at travel agency several times, author had to determine criteria for record comparison. Both Excel and Access have features for revealing duplicates. However, Access could only find duplicates and report them while Excel could also delete them within one operation. “Remove Duplicates” feature of Excel suggested choosing which columns should be checked for matches. The researcher chose “Surname”, “Name”, “BirthDate” columns to determine a person and “VisaStartDate” column to determine the service those customers got at the travel agency. At the beginning, identification of clients by passport seemed suitable too, however, the passport number is not necessarily related to the client due to necessity to change passport every 5 years or earlier if there are no blank pages left. Overall, the author deleted 175 duplicates at the this stage. However, when database creation process started, more duplicates were revealed due to significant amount of typos and extra spaces in the surnames and names of the customers. The researcher managed to determine 5,660 customer profiles with 9,560 visa applications in total after cleaning all the duplicates out.

6.2.3 Other changes

In addition to the shortcomings in the organization and purity of the data, there were inaccuracies that appeared insignificant and small at first glance. One of the most global issues in this section was presence of blank cells. For instance, it was not possible to copy customers' gender from the applications of years 2013-2015. In this regard, about 6,000 cells in the corresponding column turned out to be blank, the same for VisaType column. In addition, as researcher described earlier, each employee marked absence of previous names differently. Some employees left the fields empty instead of typing "none". The same situation occurred with the columns about the change and the reason for the change of citizenship. The corresponding fields appeared to be blank for everyone who has never had Russian citizenship before.

From the database point of view as well as data analysis, empty fields are unacceptable. It is not clear, whether fields are left blank, because someone forgot to fill them in or accidentally erased the data, or because this data does not apply to the customer. In the case of data analysis, empty fields make the results unreliable. Although there are ways to make software recognize these fields as zero or "not applicable", it is considered time consuming and is only possible for skillful users. To solve this problem, the author replaced some empty fields in the spreadsheet with the expression "N/A" which means "not applicable" or "not available" depending on the situation.

Furthermore, some columns contained information of different types both within the columns and relative to each other. For instance, several columns with dates contained slash signs as delimiters, others had dots and some columns had both of them. While different delimiters could slightly interfere the analysis, the order of the data, for instance, having both 08.10.2018 and 2018.10.08 as dates, can significantly slow down the process and lead to false analysis results. For accuracy of data, the author replaced all delimiters with one kind of signs.

The need for analysis had even more impact on data processing. The researcher considered it crucial to find out where the customers come from. This is important, for instance, in order to understand how a particular marketing campaign works, to see its scope. In addition, it may reveal new potentially interesting areas for advertising. Unfortunately, the address was one of the most difficult parts (with the most diversified design) of the data processing.

Firstly, the applications did not support such common symbols of Finnish language as “Ä” and “Ö”. Instead of them, it is customary to use letter combinations “ae” and “oe”. However, for some reason employees did not follow this rule every time. In this regard, the name of the same town could have appeared in many different ways. It was not possible to replace all “ae” and “oe” with the necessary letters, since in some names such combinations are standing naturally.

Secondly, names of the towns contained the biggest amount of typos in comparison with other parts of visa applications, possibly because Finnish language was not native to some trainees. For instance, town of Lappeenranta appeared in 46 different ways. In order to determine the locations, author decided to separate postal codes from the HomeAddress column. This could have helped not only to determine the customers’ locations, but also to combine very small towns into larger regions.

However, due to the lack of data input masks or any rules that would regulate the order of address input, employees entered addresses differently. In this regard, the postal codes were before the town name or after, the address were either separated by commas or not. To get the postal codes out of the column, author manually harmonized the data to one type. By re-sorting and filtering, as well as transposing the column, the author revealed the rows that required changes. After the work was completed, the researcher separated the postal codes with “RIGHT()” function of Excel which returns the specified number of characters starting on the right side of the line.

7 Database creation process

In accordance to necessity of an automated solution, which will optimize the workflow and allow conducting an analysis of customer data in extremely limited budget and personnel, researcher has chosen database as a tool. The author defines database as an information structured upon particular algorithm or concept while providing opportunities for further database management system creation.

Despite the standing acquisition of applications for several years, the tourism agency did not or could not derive any benefit from data existence. Database must not only meet the requirements of the enterprise, but also demonstrate managers the benefits from gathering, structuring and analysis of customer data.

Due to the fact that the manager of tourism agency does not have any special knowledge neither in IT, nor in marketing, the database was created in a simplified format from a limited amount of data with the use of budget software. Although, this approach may bring a temporary solution on the scale of bigger company, it will prove that even a small company with a minimal budget and limited knowledge can independently optimize the working routine of the enterprise and create opportunities for future marketing research and data management development.

7.1 Database schema

After the author prepared the data, it was possible to proceed with the creation of the database. In order to create a useful relational database, it was necessary to determine the schema - which relations (tables) are applicable for the tourism agency, and what links between them needed to be established. Work with visa applications includes 3 main components: clients, visas and payments. The database may include many more units, for example, a separate table for attachments like insurances or photographs to unload the main table. However, the author set a limit of 3 basic components in order to make the schema more simple and understandable for the managers of the tourism agency, as well as to provide them with the opportunity to correct or improve the database, according to experience in using it.

In addition, the author distinguishes between a visa and orders. Visas, just like a product, contain additional information such as dates, type, enclosed documents, for instance, invitations, while the order is a kind of customer request, which could be satisfied by Consulate or not. Moreover, management could use “Order” table for other services of the company. In this regard, author decided to set 4 main relations (tables) for the database: customers, orders, visas and payments.

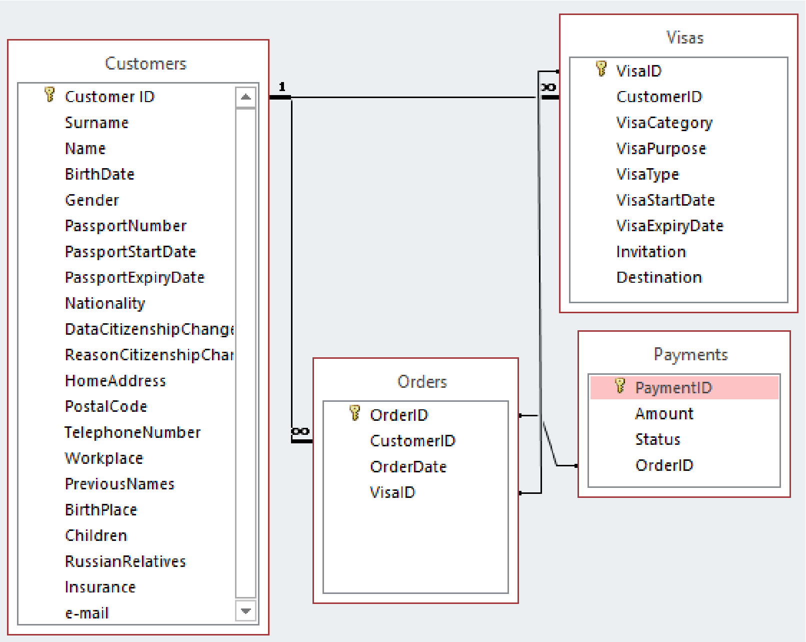


Figure 7.1. Database tables, fields and relations

Figure 7.1 describes dependencies between tables and their contents. Table “Customers” contains customer details. This table has relation “one-to-many” with “Visas” meaning that one customer could have many visas. The author set the same type of dependency between “Customers” and “Orders” that means that customer could have many orders through time. At the same time, one order could only apply to one visa and, consequently, one payment – establishing a “one-to-one” relation. Both “Visas” and “Orders” have “CustomerID” column with the primary key in order to create a relation. Database software creates this unique key automatically to every new record, which means that there could not be two similar customer ID’s. Access creates relations between other tables on the same principle.

7.2 Field settings

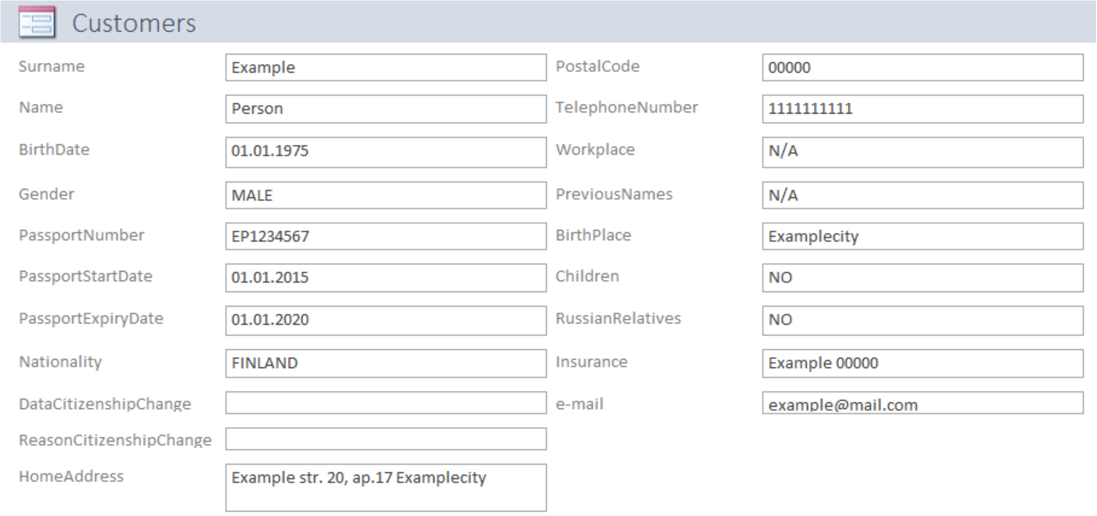
The author forbid leaving blank spaces in most of the fields, which means that the database will not allow user to continue, save or close the table until the database user enters the data. This measure is necessary, since it is compulsory to fill in all the fields in the visa applications, and empty cells can lead to an unreliable database. In case customer does not have information for some fields, employees should write “N/A” meaning “not applicable” or “not available”.

In addition, for data accuracy, the author set the data type and the length of the data in each field to reduce the possibility of errors to a minimum. Thus, all fields containing, for instance, the duration of a visa were designated as "date and time" fields. The date entered to this field may contain different delimiters as, for example, commas and slashes, which the database automatically translates into points. For postal codes and telephone numbers, author specified a number of digits so that all fields contain the same data format (either +35841234567 or 041234567). The researcher formed fields containing a limited set of data options with drop-down lists from which user can choose the appropriate option. For example, in the “Payment” table, the Status is determined by only two options “pending” and “paid”, since the travel agency has not yet provided any system for charging penalties for overdue bills. The researcher applied the same feature to the fields containing visa categories, types and visa purposes, since there is a limited number of data options.

7.3 Forms

There are several ways to enter or change data in the Access. One of them is the Forms that allow user to make data entries in a more compact way and in some cases eliminate the need to switch between different tables too. Managers of tourism agency can choose a convenient way to enter the data or independently adjust the Forms to suit their habits, but the author still created several forms for familiarization.

Firstly, the author created a Form for the Customers table. The main benefit of this form is that user can see all the fields for data entry at once. In the table view user has to scroll the table to the right or left to see the fields. Figure 7.2 shows the Form for table Customers.

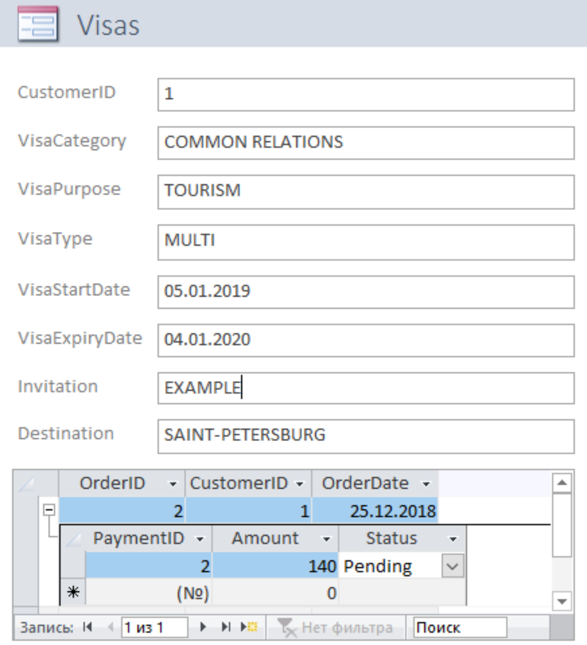


Customers

Surname	Example	PostalCode	00000
Name	Person	TelephoneNumber	1111111111
BirthDate	01.01.1975	Workplace	N/A
Gender	MALE	PreviousNames	N/A
PassportNumber	EP1234567	BirthPlace	Examplecity
PassportStartDate	01.01.2015	Children	NO
PassportExpiryDate	01.01.2020	RussianRelatives	NO
Nationality	FINLAND	Insurance	Example 00000
DataCitizenshipChange		e-mail	example@mail.com
ReasonCitizenshipChange			
HomeAddress	Example str. 20, ap.17 Examplecity		

Figure 7.2. Form for Customers table

Secondly, the author created a Form for Visas. This Form (Figure 7.3) allows user to automatically create an order and payment for a visa without swithiching between different tables or Forms.



Visas

CustomerID	1
VisaCategory	COMMON RELATIONS
VisaPurpose	TOURISM
VisaType	MULTI
VisaStartDate	05.01.2019
VisaExpiryDate	04.01.2020
Invitation	EXAMPLE
Destination	SAINT-PETERSBURG

OrderID	CustomerID	OrderDate
2	1	25.12.2018

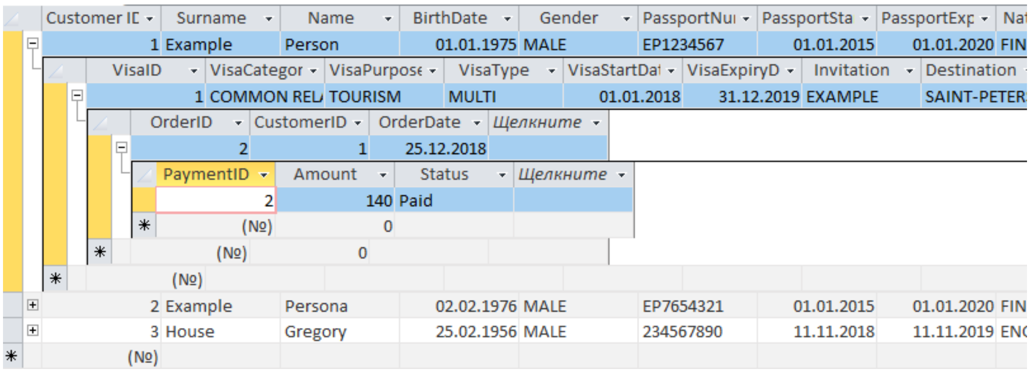
	PaymentID	Amount	Status
	2	140	Pending
✱	(№)	0	

Figure 7.3 Forms for Visas table including Order and Payment fields

There is still a possibility to fill in all the data throuhg the main table, however it is much easier and secure to make data entries with the help of Forms.

7.4 Queries

Straight from the Customers table user is able to see all the visas of the clients as well as corresponding orders and payments by clicking the “+” sign to the left from the particular customer record as shown in the Figure 7.4.



	Customer IĽ	Surname	Name	BirthDate	Gender	PassportNu	PassportSta	PassportExp	Na
⊟	1	Example	Person	01.01.1975	MALE	EP1234567	01.01.2015	01.01.2020	FIN

	VisaID	VisaCategor	VisaPurpose	VisaType	VisaStartDa	VisaExpiryD	Invitation	Destination
⊟	1	COMMON REL	TOURISM	MULTI	01.01.2018	31.12.2019	EXAMPLE	SAINT-PETER
✳	(№)							

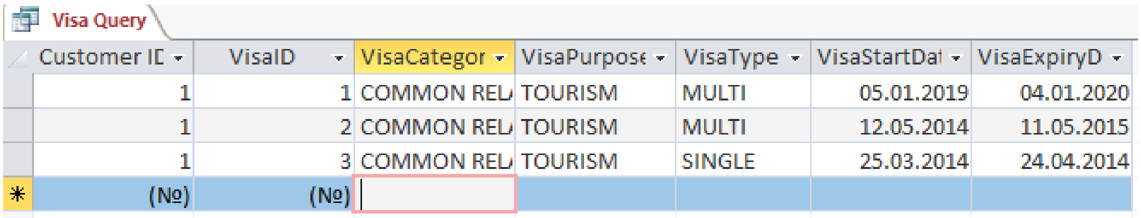
	OrderID	CustomerID	OrderDate	Щелкните
⊟	2	1	25.12.2018	
✳	(№)	0		

	PaymentID	Amount	Status	Щелкните
	2	140	Paid	
✳	(№)	0		

	Customer IĽ	Surname	Name	BirthDate	Gender	PassportNu	PassportSta	PassportExp	Na
⊞	2	Example	Persona	02.02.1976	MALE	EP7654321	01.01.2015	01.01.2020	FIN
⊞	3	House	Gregory	25.02.1956	MALE	234567890	11.11.2018	11.11.2019	ENG
✳	(№)								

Figure 7.4. View of customer’s visas and corresponding orders and payments

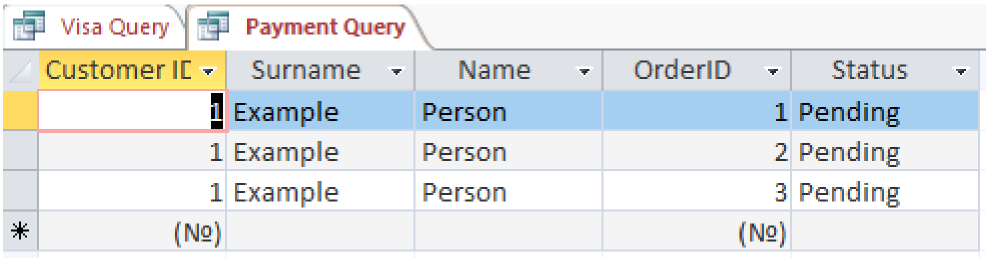
Employees can get the same result with the help of customized Queries. Figure 7.5 and Figure 7.6 show the basic and simple queries, which show all the visas of the customer and all the payments of the customer respectively.



Visa Query

	Customer IĽ	VisaID	VisaCategor	VisaPurpose	VisaType	VisaStartDa	VisaExpiryD
	1	1	COMMON REL	TOURISM	MULTI	05.01.2019	04.01.2020
	1	2	COMMON REL	TOURISM	MULTI	12.05.2014	11.05.2015
	1	3	COMMON REL	TOURISM	SINGLE	25.03.2014	24.04.2014
✳	(№)	(№)					

Figure 7.5. Visa query



Visa Query | Payment Query

	Customer IĽ	Surname	Name	OrderID	Status
	1	Example	Person	1	Pending
	1	Example	Person	2	Pending
	1	Example	Person	3	Pending
✳	(№)			(№)	

Figure 7.6. Payment Query

The user can see all the visas and payments of all customers in specific order or to see the data concerning only one customer by setting the CustomerID into the Query.

8 Data analysis

In order to demonstrate what kind of business knowledge the basic customer information carries, the researcher conducted an analysis of the data available at the tourism agency. The author used comprehensive and popular software called “RStudio” as a tool for this analysis. To make this possible, the researcher took online educational course at Stepik.org, which is a cloud-based educational platform. The data analysis process consisted of performing calculations using the R programming language and further visualizing the results using special packages for the RStudio.

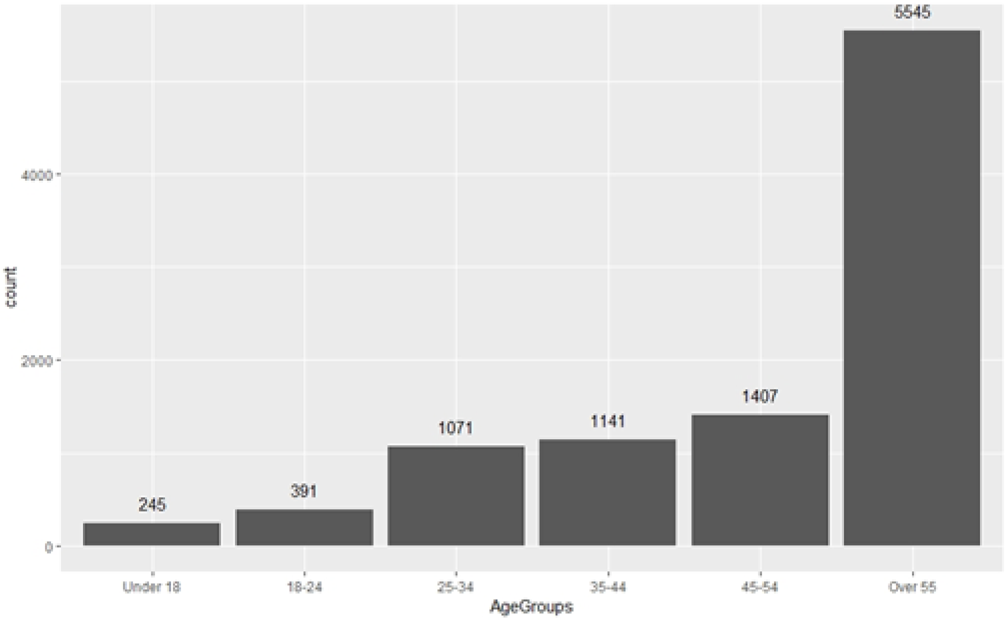
8.1 Process

The initial stage of the work included the data preparation for the analysis described in the chapter “Data processing and converting”. As soon as the spreadsheet was ready, the author saved it as a CVS file, which frequently used to transfer data from one source to another. In order to simplify the process, the researcher uploaded RStudio packages designed, for instance, to help to turn variables into plots and graphs or to deal with date and time operations. Appendix 1 contains the whole script, which illustrates the work done step by step. Due to lack of experience in creating codes and working with programming languages, difficulties arose several times throughout the the analysis process. The author got consultation of the specialist Valeria Bogdanova (2018) from EPAM Systems mentioned before, as well as turned to the R community on the Internet for advice and assistance. In order to get the most accurate results, the researcher only did the simplest calculations and visualizations.

8.2 Results

This subchapter demonstrates the visualized results of customer data calculations. For companies in the service sector, it is crucial to have a good and complete understanding of who their client is.

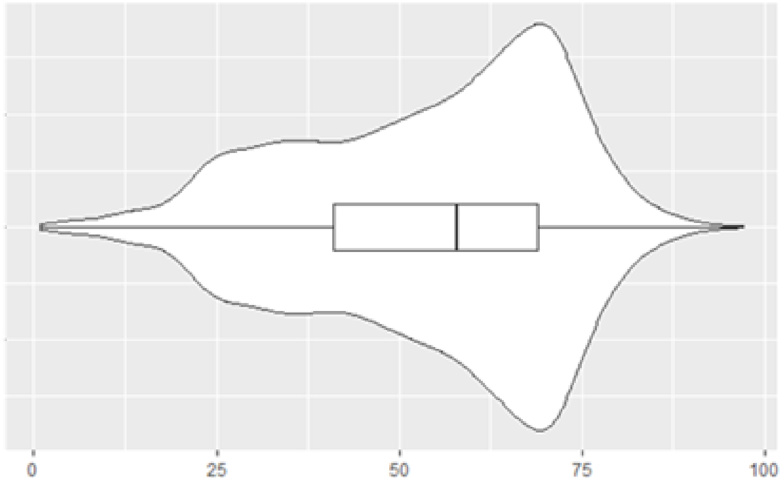
One of the main feature of customers that tourism agency could extract from visa applications is age. Graph 8.1 is a bar chart that presents age groups of customers by their number. They were determined as following: "Under 18", "18-24", "25-34", "35-44", "45-54", "Over 55". The calculations found that the most numerous age group among the clients of a travel agency are people over 55 years old.



Graph 8.1. Customers’ age groups by number bar chart

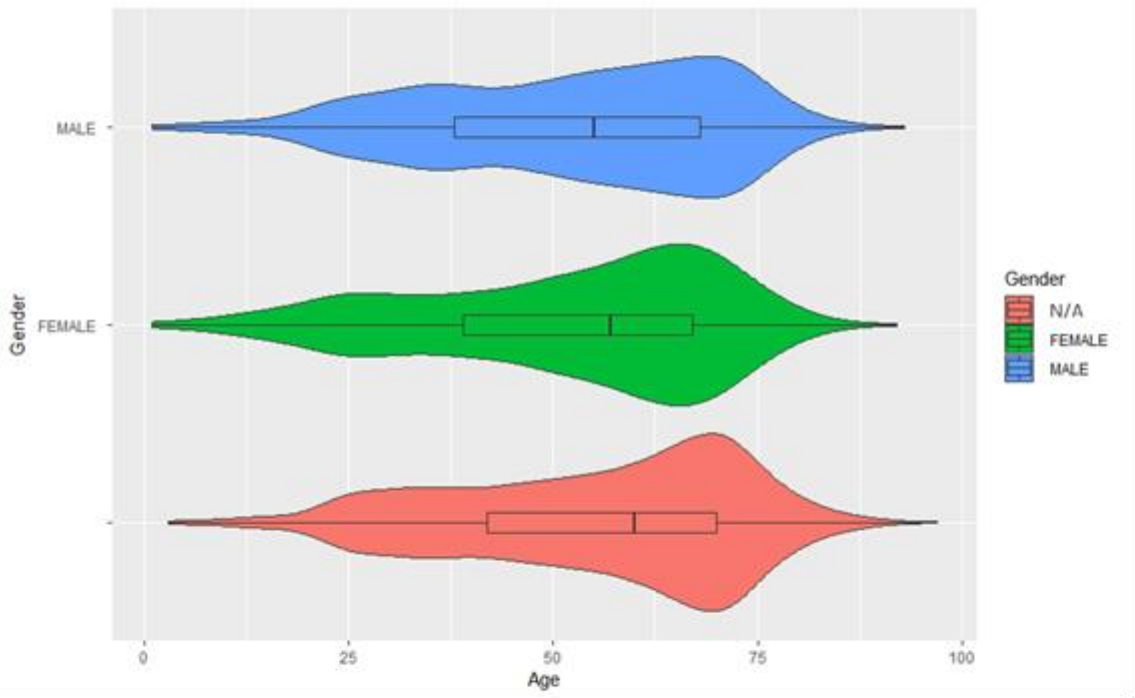
Age groups from “25 to 34”, “35-44” and “45-54” are the next with a slight difference in numbers between each other. Together, these groups make up a little more than half of the biggest age group of customers. Youth and children customers have the smallest share.

Graph 8.2 describes age of customers as well. It is a combination of violin graph and a box plot. Violin graph demonstrates a proportional increase in a number of clients relative to their age.



Graph 8.2. Customers’ age violin graph with a box plot

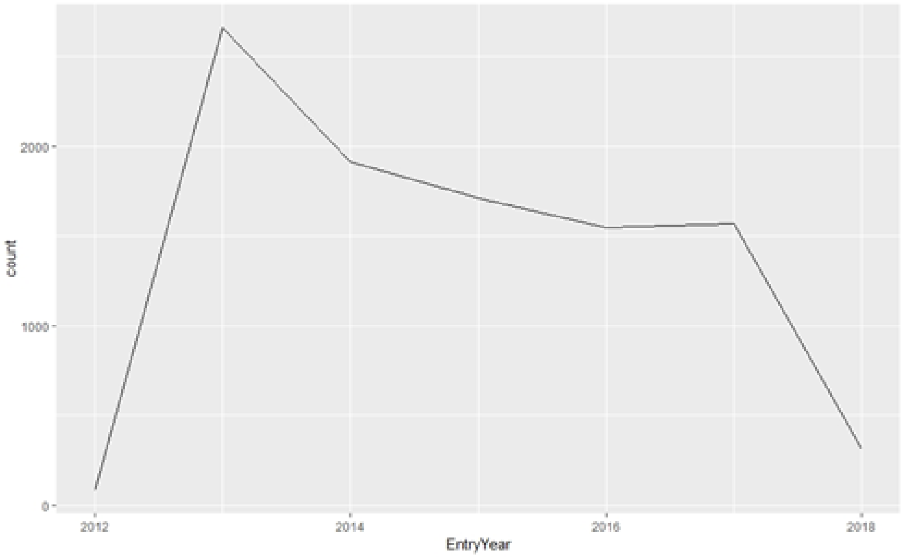
This graph shows a steady growth of Millennial customers with insignificant decline of clients around age of 40 and consequent substantial increase up to age of 70 with a sharp decline afterwards. Box plot describes 1st and 3rd quartiles meaning 25% observations are below the age of 41 and 75% of observations below the age of 69. The line inside the box plot shows a 2nd quartile or median of observations at the age of 58.



Graph 8.3. Customers' gender in relation to the age

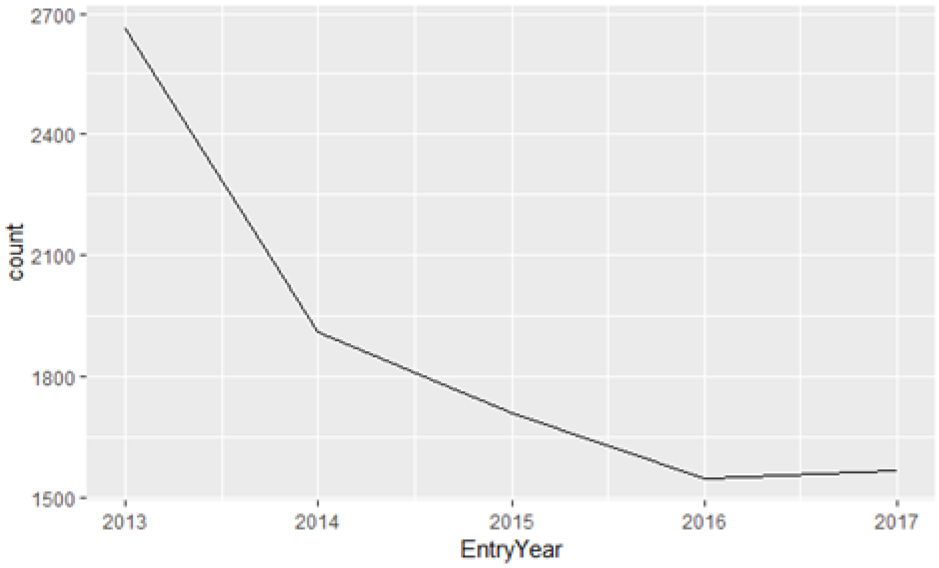
Graph 8.3 describes customers' gender proportions with the relation to the age. Despite the fact that author was not able to extract gender data from a significant number of applications, the results of analysis of existing data suggests that gender proportions of customers in tourism agency could be almost 50 to 50 percent. Moreover, this graph also shows that both males and females have rather similar age groups.

Graph 8.5 describes the amount of visas issued in the tourism agency per year. Firstly, author used all the applications for these calculations. However, due to the fact that some years were presented with one or two months instead of twelve, the researcher considered results of calculations and visualization non-indicative as shown on the Graph 8.4. To avoid inconsistency, author only took applications from January 2013 to December 2017 for the metrics.



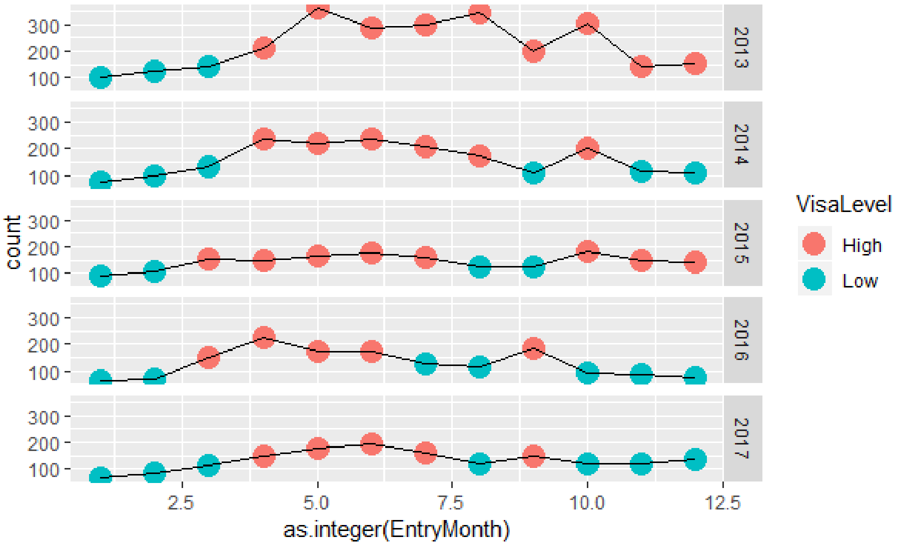
Graph 8.4. Amount of visas issued per year from 2012 to 2018. Non-indicative

Graph 8.5 demonstrates the steady decline in issued visas from year 2013 to 2016 with a slight increase in year 2017.



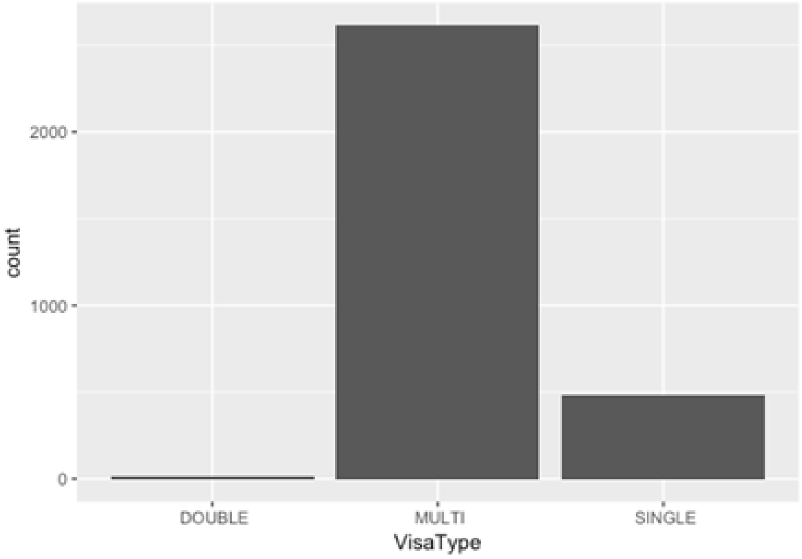
Graph 8.5. Amount of visas issued per year from 2013 to 2017

In order to take a closer look at the issue of the applied visas, the author made a visualization for each year separately. In the process of calculating, the author learned that the largest number of issued visas per month was 359 items. In order to determine high and low levels of visa applications, the author took median of 139 visas per month as the borderline.



Graph 8.6. Amount of visas issued per month of 2013-2017

Graph 8.6 confirms that year 2013 had a greater amount of issued visas, maintaining the high level from April to December. The graph reveals the high seasons from late spring to summer as well as a peak in autumn throughout the years, while winter could be considered as a low season.



Graph 8.7. Visa type (2016-2018)

By analogy with the gender, the researcher was not able to transfer the data about visa types from all the applications. Nevertheless, the author made calculations on the data that existed. Graph 8.7 shows that the multiple-entry visa prevail among other types over the period from 2016 to 2018.

Finally yet importantly, the author made calculations to determine customer locations by postal codes of their home addresses. In order not to overload the results author divided all locations according to postal code areas. Figure 8.1 shows these areas.

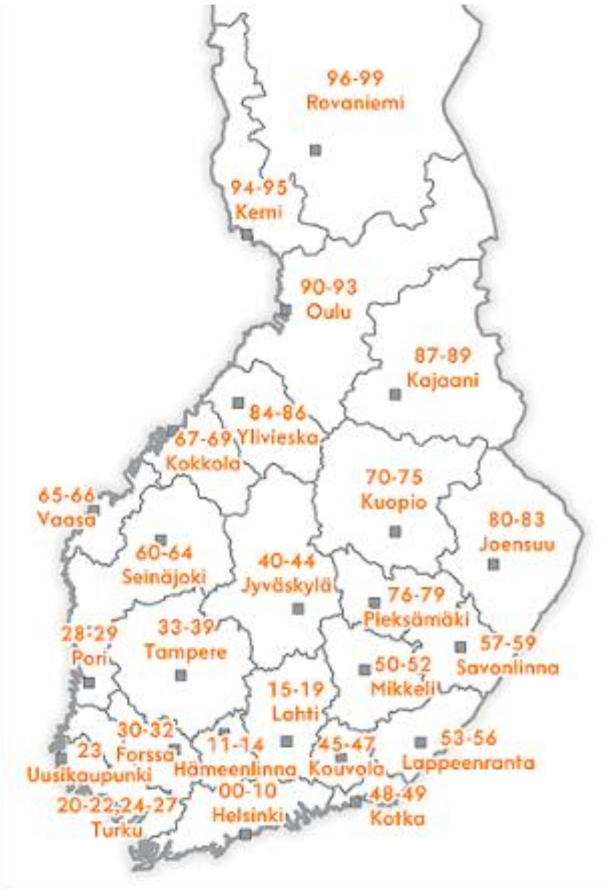
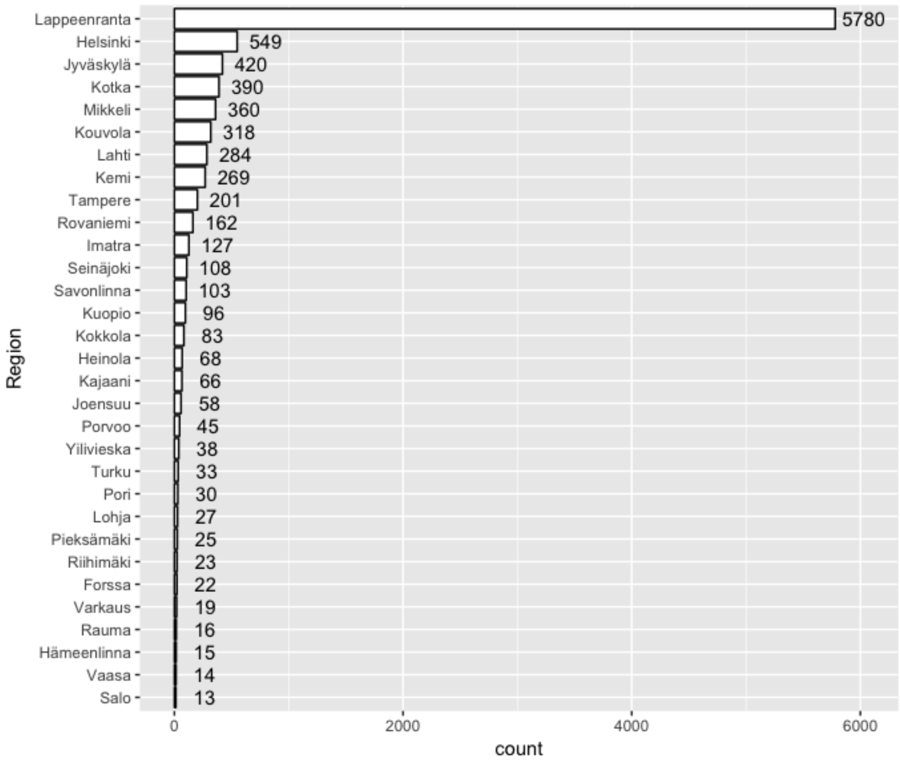


Figure 8.1. Postal areas of Finland (Posti 2018).

The Graph 8.8 clearly shows that the main segment of customers comes from Lappeenranta, which makes up 59.21% of the total amount. It is worth mentioning, that in order to proceed the visa application, customer should have an address in Finland. That is why there were no foreign postal codes, despite the fact that there were several customers with foreign citizenship.



Graph 8.8. Visa applications by customers' postal areas

For more accurate customer segmentation, this analysis can be carried out in conjunction with the Google Business Analytics tools available to the tourism agency.

9 Conclusions

The main goal of this thesis was to create a tool that will help a small-sized tourism agency in South Karelia to improve data management in order to optimize the working routine and gain knowledge from the existing customer data of the enterprise. The author had to create this tool using budget solutions that do not entail additional expenses for the company, as well as not implying the need for special long-term staff training. This was necessary in order to demonstrate that small businesses such as a tourism agency can independently create a tool for data management and analysis, the results of which can be incorporated into decision-making process to create a competitive advantage.

One of the main difficulties occurred during data processing due to the very low quality of data provided by tourism agency. Data processing was the most exhausting, the most time and power consuming part of the thesis, which took about 4 month to get satisfactory results. However, even after numerous manipulations with data, it was impossible to unambiguously judge its completeness or accuracy, since the company have never had any rules and procedures for storing data. In this regard, without having experience with working on a large amount of data previously, the author was unable to correctly estimate the volume of the work to be done at the preliminary stage.

Based on the experience with data handling obtained during the writing of the thesis, the author concludes that managers should create and discuss data management techniques as well as rules for working with data at the business planning stage regardless of the size of the enterprise. The later data management techniques will be introduced into the business, the greater would be the costs both in money and in time involved. Moreover, the later the company’s management understands the need to take measures to manage the data, the less stable and reliable the result of those measures would be. Furthermore, the stage at which managers decide whether or not disciplined data management is necessary directly affects whether the company can organize data management on its own or have to resort to professional help from outside.

Based on the results of the data analysis, it is possible to specify the main segment of the company's visa clients, which, if detected earlier, could be used to correctly select the advertising and marketing channels. It is widely known that despite the growing popularity of social networks among users of all ages, newspapers and television are still the main sources of information for older people who are the main customers of the company. However, the company advertises itself exclusively through Facebook, with a minimal amount of detail, which usually push away the older customers. In addition, the results obtained from data analysis indicate a long-term and constant decrease in client interest in the company. This information could have helped the company identify errors or problems at an early stage to respond appropriately and prevent further decline. The analysis helped to identify the seasonality of interest in visas too. This knowledge could have helped the company to plan marketing campaigns and discounts specifically for low seasons in order to balance the revenue.

The large amount of private data has complicated reporting on the thesis, depriving the author of the opportunity to fully demonstrate the results of the work done. Not being a professional in working on visa issues, the author was able to create a simple basic tool, which opens up opportunities for the experienced managers of tourism agency to develop and improve their data management according to their specific needs and habits. This tool improves the quality of customer data in the company, improves customer service, and reduces waste of resources such as paper. In addition, having an up-to-date database is the only way to meet the requirements of the law on the protection of personal data.

Author handed out the database and analysis results to the management of the tourism agency. Together with the manager author had a trial use of the database on the working day of the company. The manager noted the ease of use, accessibility of information in the database, as well as deeply appreciated the availability of the "Payment" table. The author included this table on her own initiative, which turned out to be significantly demanded from the point of view of the travel agency manager.

Furthermore, through the analysis the author demonstrated the knowledge content of the data, as well as showed the possibilities of using this knowledge for the benefit of the company. However, just having the database is not a solution to the problems. To make data serve the needs of the enterprise, managers and employees of the tourism agency must stick to the common rules of data handling as well as to create the list of their own specific rules. Despite all the difficulties that have arisen during this research, the author was able to execute the plan.

List of Figures

List of Graphs

References

Allardice, S. 2013. Database programming tutorial: What are databases? LinkedIn Learning. https://www.youtube.com/watch?v=Ls_LzOZ7x0c. Accessed on 5 June 2018.

Blattberg, R. & Hoch, S. 2010. Database Models and Managerial Intuition 50% Model + 50% Manager. **In** Allenby, G. (Eds.) 1990. Perspectives on Promotion and Database Marketing: The Collected Works of Robert C Blattberg. Singapore: World Scientific Printers Co.Pte.Ltd.

Blattberg, R., Kim, B. & Neslin, S. 2008. Database Marketing: Analyzing and Managing Customers. New York: Springer Science+ Business Media, LCC.

Bogdanova, V. 2018. Software Engineer. EPAM Systems. Saint-Petersburg. Interview on 15 May 2018.

Bogdanova, V. 2018. Software Engineer. EPAM Systems. Lappeenranta. Interview on 25 August 2018.

Booking.com 2018. https://www.booking.com/content/about.en-gb.html?aid=376363;label=booking-name-L%2AXf2U1sq4%2AGEkIwcLOALQS267777916057%3Apl%3Ata%3Ap1%3Ap22%2C119%2C000%3Aac%3Aap1t1%3Aneg%3Afi%3Atikwd-65526620%3Alp1005600%3Ali%3Adec%3Adm;sid=f1bf3c783c7e8c6e63793858b45a6897. Accessed on 23 September 2018.

Consular department: Ministry of Foreign Affairs of the Russian Federation, 2018. https://helsinki.mid.ru/obsaa-informacia-o-rossijskoj-vize. Accessed on 13 July 2018.

Dancho, M. 2017. 6 reasons to learn R for business. Business Science. https://www.business-science.io/business/2017/12/27/six-reasons-to-use-R-for-business.html. Accessed on 15 May 2018.

Davidson F. 1996. Principles of statistical data handling. SAGE Publications, Inc.

Hennig, T., Bradly, T., Linson, L., Purvis, L. & Spaulding, B. 2010. Microsoft Access Small Business Solutions: State-of-the-Art Database Models for Sales, Marketing, Customer Management, and More Key Business Activities. Indianapolis: Wiley Publishing, Inc. International Journal of Managing Projects in Business, Vol. 4 Iss 4 pp. 573 – 595.

Kalmikov, A. 2018. Evrosouz voorujil 500 mln partisan k bitve za lichnie Dannie v internete. BBC Russian Service. https://www.bbc.com/russian/features-44256440. Accessed on 25 May 2018.

Katz M. & Katz B. 1997. Marketing on a Restricted Budget. Chalford: Management Books 2000 Ltd.

Koskinen, I., Zimmerman, J., Biner, T., Redström, J. & Wensveen, S. 2011. Design Research through practice. From the Lab, Field, and Showroom. Waltham: Elsevier Inc.

Lungariello, R. 2017. The 10 Best Database Software Systems For Business Professionals. https://mytechdecisions.com/it-infrastructure/10-best-database-software-systems-business-professionals/. Accessed on 24 September 2018.

Morritt R. 2007. Segmentation Strategies for Hospitality Managers: target marketing for competitive advantage. Binghamton: The Haworth Press, Inc.

Mullins S., C. 2015. Which relational DBMS is best for your company? https://searchdatamanagement.techtarget.com/feature/Which-relational-DBMS-is-best-for-your-company. Accessed on 24 September 2018.

Oyegoke, A. 2011. The constructive research approach in project management research.

Puckering, G. 2018. What is the difference between a flat file and a database? Quora. https://www.quora.com/What-is-the-difference-between-a-flat-file-and-a-database. Accessed on 23 September 2018.

Regulation (EU) 2016/679

Rouse, M. 2018. Relational database. Techtarget SearhDataManagement. https://searchdatamanagement.techtarget.com/definition/relational-database. Accessed on 23 September 2018.

R-project, 2018. What is R? https://www.r-project.org/about.html. Accessed on 20 September 2018.

Russian Visa Application Center in Finland, 2018. http://russiavisacentre.com/. Accessed on 13 July 2018.

Salesforce, 2018. What is CRM? https://www.salesforce.com/eu/learning-centre/crm/what-is-crm/. Accessed on 23 September 2018.

Stephens, R. 2009. Beginning Database Design Solutions. Indianapolis: Wiley Publishing, Inc.

Tuffill, S. n.d. Advantages & Disadvantages of Flat File Databases. Techwalla. https://www.techwalla.com/articles/advantages-disadvantages-of-flat-file-databases. Accessed on 23 September 2018.

YLE Uutiset 2018. Inostrannie turisti v proshlom godu ostavili v Yujnoi Karelii rekordnie 360 mln evro. https://yle.fi/uutiset/osasto/novosti/inostrannye_turisty_v_proshlom_godu_ostavili_v_yuzhnoi_karelii_rekordnye_360_mln_yevro/10267047. Accessed on 15 July 2018.

YTJ 2018. Negosudarstvennoje Obrazovatelnoje Utsrezdenie Litzei Stolitshnyj, sivuliike Suomessa https://tietopalvelu.ytj.fi/yritystiedot.aspx?yavain=1679807&tarkiste=D4B70B4F023FADECB7F64B66E6C32DB8417F142E. Accessed on 28 May 2018.

Appendicies

Appendix 1 RStudio script for data analysis

```
# Created on 2018.08.31
# Author: Aleksandra Bogdanova

# install and load packages ----

# install.packages(c("dplyr", "ggplot2", "lubridate", "tidyr"))
library(dplyr)
library(ggplot2)
library(lubridate)
library(tidyr)


# read raw data ----
data.file <- "Customers_data.csv"
customer.data <-  read.csv2(data.file)
customer.data <-  read.csv2(data.file, stringsAsFactors = F)[, 1:30] ## for new
file

str(customer.data)
customer.data$EntryDate <- as.Date(customer.data$EntryDate, format =
"%d.%m.%Y")

# create Age and AgeGroups variables ----
customer.data$BirthDate <- as.Date(customer.data$BirthDate, format =
"%d.%m.%Y")
customer.data$Age <- as.integer(floor(difftime(Sys.Date(),
customer.data$BirthDate) / 365.242))
summary(customer.data$Age)
customer.data$AgeGroups <- factor(ifelse(customer.data$Age < 18, "Under
18",
                                  ifelse(customer.data$Age < 25, "18-24",
                                         ifelse(customer.data$Age < 35, "25-34",
                                                ifelse(customer.data$Age < 45, "35-44",
                                                       ifelse(customer.data$Age < 55, "45-54",
"Over 55"))))),
                                  levels = c("Under 18", "18-24", "25-34", "35-44", "45-54",
"Over 55"))

# age statistics general ----
ggplot(customer.data, aes(x = Age)) +
  geom_density()
ggplot(customer.data, aes(x = 1, y = Age)) +
  geom_boxplot() +
  coord_flip()

ggplot(customer.data, aes(x = 1, y = Age)) +
  geom_violin() +
  geom_boxplot(width = 0.1) +
  coord_flip()
ggplot(customer.data, aes(x = Age)) +
  geom_density() +
  geom_histogram(aes(y = ..density..), binwidth = 10)

ggplot(customer.data, aes(x = AgeGroups)) +
  geom_bar() +
  geom_text(stat='count', aes(label=..count..), vjust=-1)

# age statistics by sex ----
ggplot(customer.data, aes(x = Age, fill = Gender)) +
  geom_density(alpha = .3)
ggplot(customer.data, aes(x = Gender, y = Age, fill = Gender)) +
  geom_boxplot() +
  coord_flip()

ggplot(customer.data, aes(x = Gender, y = Age, fill = Gender)) +
  geom_violin() +
  geom_boxplot(width = 0.1) +
  coord_flip()

# number of visas by year ----
customer.data <- customer.data %>%
  mutate(EntryYear = as.integer(year(EntryDate)),
         EntryMonth = as.integer(month(EntryDate)))

time.data.years <- customer.data %>%
  group_by(EntryYear) %>%
  summarise(count = n()) %>%
  ungroup()

ggplot(time.data.years, aes(x = EntryYear, y = count)) +
  geom_line()
ggplot(filter(time.data.years, EntryYear > 2012 & EntryYear < 2018), aes(x =
EntryYear, y = count)) +
  geom_line()

# stats by year-month number of visas
time.data.ym <- customer.data %>%
  group_by(EntryYear, EntryMonth) %>%
  summarise(count = n()) %>%
  ungroup() %>%
  mutate(VisaLevel = ifelse(count > median(count), "High", "Low"))
summary(time.data.ym$count)
visa.median <- median(time.data.ym$count)

ggplot(filter(time.data.ym, EntryYear > 2012 & EntryYear < 2018),
       aes(x = as.integer(EntryMonth), y = count, color = VisaLevel, group = 1)) +
  geom_point(size = 5) +
  geom_line(color = "black") +
  facet_grid(EntryYear ~ .)

# Visa Type ----
table(customer.data$VisaType)

ggplot(customer.data, aes(x = VisaType)) +
  geom_bar()

ggplot(filter(customer.data, VisaType != ''), aes(x = VisaType)) +
  geom_bar()

ggplot(filter(customer.data, VisaType != '   DOUBLE  ' & VisaType != ''), aes(x =
VisaType)) +
  geom_bar()

# Postal Code ----

customer.data$region_number <- as.integer(substr(customer.data$PostalCode,
1, 2))
customer.data$region <- ''
customer.data$region[customer.data$region_number <= 99] <- 'Rovaniemi'
customer.data$region[customer.data$region_number <= 95] <- 'Kemi'
customer.data$region[customer.data$region_number <= 89] <- 'Kajaani'
customer.data$region[customer.data$region_number <= 86] <- 'Yilivieska'
customer.data$region[customer.data$region_number <= 83] <- 'Joensuu'
customer.data$region[customer.data$region_number <= 79] <- 'Varkaus'
customer.data$region[customer.data$region_number <= 77] <- 'PieksΓ¤mΓ¤ki'
customer.data$region[customer.data$region_number <= 75] <- 'Kuopio'
customer.data$region[customer.data$region_number <= 69] <- 'Kokkola'
customer.data$region[customer.data$region_number <= 66] <- 'Vaasa'
customer.data$region[customer.data$region_number <= 64] <- 'SeinΓ¤joki'
customer.data$region[customer.data$region_number <= 59] <- 'Savonlinna'
customer.data$region[customer.data$region_number <= 56] <- 'Imatra'
customer.data$region[customer.data$region_number <= 54] <- 'Lappeenranta'
customer.data$region[customer.data$region_number <= 52] <- 'Mikkeli'
customer.data$region[customer.data$region_number <= 49] <- 'Kotka'
customer.data$region[customer.data$region_number <= 47] <- 'Kouvola'
customer.data$region[customer.data$region_number <= 44] <- 'JyvΓ¤skylΓ¤'
customer.data$region[customer.data$region_number <= 39] <- 'Tampere'
customer.data$region[customer.data$region_number <= 32] <- 'Forssa'
customer.data$region[customer.data$region_number <= 29] <- 'Pori'
customer.data$region[customer.data$region_number <= 27] <- 'Rauma'
customer.data$region[customer.data$region_number <= 25] <- 'Salo'
customer.data$region[customer.data$region_number <= 22] <- 'Mariehamn'
customer.data$region[customer.data$region_number == 23] <- 'Turku'
customer.data$region[customer.data$region_number <= 21] <- 'Turku'
customer.data$region[customer.data$region_number <= 19] <- 'Heinola'
customer.data$region[customer.data$region_number <= 17] <- 'Lahti'
customer.data$region[customer.data$region_number <= 14] <- 'HГ¤meenlinna'
customer.data$region[customer.data$region_number <= 12] <- 'RiihimГ¤ki'
customer.data$region[customer.data$region_number <= 10] <- 'Lohja'
customer.data$region[customer.data$region_number <= 7] <- 'Porvoo'
customer.data$region[customer.data$region_number <= 5] <- 'Helsinki'


regions <- customer.data %>%
  select(region_number, region) %>%
  filter(region != '') %>%
  group_by(region) %>%
  summarise(count = n())

ggplot(regions, aes(x = reorder(region, count), y = count)) +
  geom_bar(stat = 'identity', color= 'black', fill = 'white') +
  coord_flip() +
  xlab('Region') +
  geom_text(aes(label = count), nudge_y = 250)
```