­­­

|  |
| --- |
| Arden University |
| Msc Data Analytics and Finance |
| DAT7003-Data Design |
| Data Design Report |
| STU74936 |
| Eugene Kozlovski |
|  |

1. **Business Understanding**
   1. **Background to the Case**

The aim of this research paper is to measure the service quality that we provide to customers that have a credit facility at the researcher’s place of work, and to provide measures to improve this service. The organization chosen is a limited liability bank and is based within the Caribbean region. The services offered are online banking, credit facilities, savings and demand deposit accounts and other common banking services.

* 1. **Key Organizational Objectives and Business Question**

The banks main source of income is through the interest generated in its credit facilities. Customer satisfaction plays a major role in attracting and retaining new credit customers. The bank also faces a major challenge with the emergence of new competitors within the market that are aggressively targeting it customers.

Therefore, at the end of this study the researcher aims to reach the following objectives:

1. To asses wether our customers are satisfied with the level of service that we provide.
2. To analyze the relationship between service quality and customer retention
3. To suggest new creative measures to improve on our service.

**The case study question posed “Are the number of loans taken by customers affected by the customer service?**

* 1. **Data Sources**

The researcher is using the Inferential method to collect data. This data will be gathered from questionnaire’s provided to a selected sample of customers and a general conclusion will be derived from the results. The goal is to establish whether our service is meeting customer standards, what aspects of our service that customers are impressed with and improvements that can be made to widen our market share.

A sample of fifty customers were targeted, with a mixture of both male and female, ranging from an age group of 25-55yrs. The sample selected were of customers that resided in the country that the bank was established and is currently operating from. Data was gathered during the months of February- March 2020.

The risks involved in using this method of data collection are the sample used to collect the data may not depict a true picture of all our customers, customers may also provide inaccurate information, the relevant stakeholders may retain certain information that they feel is highly confident. The competitors may become more aggressive when they realize that the researcher is conducting market research.

For this project declarative knowledge will be sourced from the loans department database, from the Senior Executive Team, and from the internal and external customers. The business rule applied is “whether the customer received or was declined a loan from the institution within the last five years?

1. **DATA COLLECTION TOOL**
   1. **General Evaluation of Data Collection Instruments**

**Table 2.1 Comparison of Data Collection Instruments**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **No.** | **Type of Data Collection Instrument** | **Strengths** | **Weaknesses** | **Applicability to the Case Study** | **References** |
|  | Interviews | In deph Information  High Response Rate  Easier to reach specific persons  More personalised approach | Requires good interviewing skills  Expensive  Difficult to analyze and interpret results  Time-Consuming | It is good for assessing attitudes and opinions and easier to ask open ended questions. Researcher would be able to get a high response rate. | *Data Collection Methods Pros and Cons,*  Available from: <https:///www.michigan.gov/documents/metormichican/Data-Collection-Metjhods-Pros-and-Cons>  (Accessed June 7, 2020)  *Data Collection Methods,* Available from: <https://www.formpl.us/blog/data-collection-method> ( Accessed June 7, 2020)  Greeson, M., 2012 *Evaluating the work of sexual assault nurse examiner (SANE) programs in the criminal justice system (online) Pg 21* |
|  | Questionnaires | Easy to administer  Inexpensive  Easy to visualize and analyze  Respondents identity are protected | Subject to misinterpretation  More prone to error  Answers maybe biased  Possibility of getting a low response rate | It is a quick and inexpensive way to gather data from a large group of people.  Easy to analyze and interpret the data received. | *Data Collection Methods Pros and Cons,*  Available from: <https:///www.michigan.gov/documents/metormichican/Data-Collection-Methods-Pros-and-Cons>  (Accessed June 7, 2020)  *Data Collection Methods,* Available from: <https://www.formpl.us/blog/data-collection-method> ( Accessed June 7, 2020)  Greeson, M., 2012 *Evaluating the work of sexual assault nurse examiner (SANE) programs in the criminal justice system (online) Pg 21* |
|  | Focus Groups | Allows for clarification  Relatively low-cost  Efficient in gathering information  A diverse group can generate insights and fresh ideas | Maybe challenging to interpret group results  Lack of confidentiality  Input may be inbalanced as some group members maybe more outspoken | It is good for assessing opinions and information can be quickly gathered. | *Data Collection Methods Pros and Cons,*  Available from: <https:///www.michigan.gov/documents/metormichican/Data-Collection-Metjhods-Pros-and-Cons>  (Accessed June 7, 2020)  Greeson, M., 2012 *Evaluating the work of sexual assault nurse examiner (SANE) programs in the criminal justice system (online) Pg 21* |
|  | Observations | Generates relevant quantifiable data  Setting is Flexible and unstructured  Records actual behavious vs. self-reports  Useful for studying a small unit | Requires good observation skills  Observations cannot be generalized to the entire population  Requires consistency across observations | This tool is more relavant to study behaviors and may not provide much insight for the case study at hand. | *Data Collection Methods Pros and Cons,*  Available from: <https:///www.michigan.gov/documents/metormichican/Data-Collection-Metjhods-Pros-and-Cons>  (Accessed June 7, 2020)  Greeson, M., 2012 *Evaluating the work of sexual assault nurse examiner (SANE) programs in the criminal justice system (online) Pg 21* |

* 1. **Selection of a Data Collection Instrument for the chosen case study**

In this report, quantitative research will be selected. The data collection tool used to collect the data was a questionnaire. This was sent via customer’s emails with a link attached to complete the survey via survey monkey. The researcher will also do face to face interviews with internal customers because of time limitations.

The questionnaire only comprised of five questions, as the researcher was aware that because of the Corona Crisis respondents would not put aside much time to aid in this report. The first question was an introductory one, inquiring when last the customer had a loan with the bank. The other questions were aimed at gathering research information. They were geared at getting customer feedback on how satisfied they were with the service and ways they think it can be improved.

* 1. **Analysis of the chosen instruments strength and weaknesses**

The potential strengths of a questionnaire are that is it less costly to administer, questionnaires will be sent via email so that eliminates travel cost and a free basic plan will be used on Survey Monkey to gather the data collected. It is also quicker to administer as a large number of emails can be sent within a short space of time. It also allows customers to feel free to give their opinions as this is done online rather than a face to face interview. Self-Completion questionnaires are also more convenient for customers as they can complete it wherever they choose to within the speed that they are comfortable with.

Although this data collection tool comes with several advantages there are also disadvantages when using this method. It will be difficult to ask a lot of open ended questions as respondents may not want to type a lot. There is a chance that some respondents may not interpret the questions clearly and the researcher will not be present to guide these individuals. It will be difficult to ask many questions as respondents can get bored or tired of completing the questionnaire. There is also the possibility of getting a very low response from customers since they are not obligated to respond to the emails sent.

**Table 2.2 Evaluation of the chosen data instrument**

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | **Advantages** | **Limitations** | **References** |
| 1 | Cost-Effective | Difficult to ask open ended questions |  |
| 2 | Quicker to administer | Misinterpretation of Questions |  |
| 3 | More Convenient | Possibility of getting a low response rate |  |
| 4. | Easy to visualise and analyse | Respondents maybe biased. |  |

1. **Data Collection**

The company’s main data software was used to collect key information required for this report. This data software is highly protected therefore the researcher had to make a request to her manager to gain the necessary permissions. The necessary access was temporarily provided with from the Management Information System department to generate reports and to view same. The researcher user details had to be temporarily modified to give her these permissions in order to carry out this report.

The data was gathered automatically by running various reports that were specific to the business analysis question. The reports generated are in the form of excel worksheets and are in tabular formats.

Two reports were generated to collect the necessary data. One report showed shows all the loan customers for the period 2018 to 2019 and contains the following fields customer name, loan account number, customer sex, the type of loan, the date taken and the loan amount. The loan number is this table key as every loan has a unique number to identify the loan. The other report will provide details of customers with multiple loans at the institution within a 10 year period, and contains the fields of customer names, customer information file number, loan numbers, number of loans taken, dates these loans were taken and the branch that the facility was completed. The table key in this table was the customer information file number.

The data collected from this source allowed the researcher to identify a sample of customers to target when sending out the questionnaires. A total of forty customers where selected that has taken a loan with the company within the last two years. A sample of customers will be selected on the following basis. The type of credit facility that the customer has with the institution, the sex of the customer, the date the loan was taken, and the branch that the loan was processed. A total of ten internal customers was also interviewed as the loan process for internal customers differs from external customers.

The data collected from the institutions data software will also provide the researcher with the email addresses of customers in a quick manner so that a mail merge can be created to send out one mass mail to customers. The data gathered can also be used as a measure of customer satisfaction by using it to calculate an average of the customers who uses our credit facilities repeatedly over the last ten years of operation.

The limitations of this data source is that is does not provide any information on the customer’s loan experience. The value and amount of loans that a customer has taken with us is not a clear indication of their satisfaction, persons may choose this institution only because its local, it gives back a lot to the community and for customer loyalty. We also had quite a few promotions which brought in a lot of customers therefore numbers are not a good measure of customer service.

***500 words/Missing 500***

1. **Data Integration**
   1. **Data Sets adopted for this study**
   2. **Data Analytics Model**
   3. **Data Cleaning & Preparation Procedures**

IBM defines **Data Integration** as “the combination of technical and business processes used to combine data from disparate sources into meaningful and valuable information.” In Lenzirini research paper *Data Integration: A Theoretical Perspective* he stated that one of the main task in the design of a data integration system is to establish the mapping between the sources and the global schema, and such a mapping should be properly taken into account when creating a data integration system.

Extract, transform, and load (ETL) technology supports data consolidation.  
ETL pulls data from sources, transforms it into an understandable format, and then transfers it to another database or data warehouse. The ETL process cleans, filters, and transforms data, and then applies business rules before data populates the new source.

The report ran to gather data from the internal source was accessed from the master server which came from different database sources. The data integrated was taken from the loans software and the banks main software. Information from these two sources were then pulled together to formulate a report. The report generated gathered the data in a tabular form and is exported into an excel spreadsheet.

The data gathered in this report was crucial in order to conduct the study as it was the quickest way possible to gather the research sample. It provided important information on the sample selected to respond to the questionares as it relates to demographics and the amount of business they have with the institution. It also provided data on customer trends, for example if our customer has repeat credit facilities at our institutions and how often customers use our loan services.

Data cleaning is the process of detecting and removing errors and inconsistencies from data as a means of improving the quality. The information gathered during this research are keen to having errors seeing that the data collected was manually entered and there is a room for human error. Other customers were added to the sample to gain feedback, as a result of time limitations it was impossible to contact each customer and verify their email address. For example, some of the email addresses uploaded from our database were incorrect or has since been changed. The researcher observed that after the initial contact with the customer no further update was done on the customer database hence the possibility of having inaccurate information. A recommendation will be made to update customer accounts when they visit the bank to carry out business in order that we have the most correct information.

An update anomaly that was identified in the report was redundancy of data. The report generated from our internal database some of the information was repeated. For example, the customer’s name was included in the table twice

* 1. **Exploratory Analysis**

Exploratory Data Analysis refers to the process of performing initial investigations on the data collected, as to identify trends, to discover anomalies, and to test hypothesis and assumptions with the use of summary statistics and graphical representations.

**498 words- Misisng 752 words**

1. **Data Analysis, Inferential Statistics or Modeling**
   1. **Justification of the Statistical Model**

Statistical modeling is the process of applying statistical analysis to a dataset. A statistical model is a mathematical representation (or mathematical model) of observed data. (Storbierski). A statistical model should summarise the data as closely as possible (be 'a good fit'). We cannot measure a population, so the best we can do is make generalizations from a sample to a population using a representative summary, i.e. a statistical model.

Hypothesis Testing

The hypothesis in this study is whether customer service affects persons decision in taking a loan at our financial institution?

The independent variable is wether or not the customer has taken a credit facility with us within the last five years. This is easily measured by using the report generated form our internal data source which informs the date the loan was taken, the customer name and the number of credit facilities that a particular customer has with the bank.

**Null Hypothesis and Alternative Hypothesis**

The alternative hypothesis of this report was developed as follows:

H1: There is a significant positive relationship between customer satisfaction and continued use of our credit facilities. Its clear that customer satisfaction plays a mjor role in retaining customers and widening the banks market share. Study shows that dissatisfied customers is more than likely to share their bad experience with atleast ten individuals.

**The SPSS Sharpio-Wilk Test**

The Shapiro-Wilk test is based on the correlation between the data and the corresponding normal scores. It is a test to determine if a random sample comes from a normal distribution.

The Shapiro-Wilk test is one of the most popular tests for normality assumption diagnostics which has good properties of power and it based on correlation within given observations and associated normal scores. The Shapiro-Wilk test statistic is derived by Shapiro and Wilk (1965)

The Shapiro-Wilk Test is more suitable for small sample sizes (< 50 samples).

**Pearson’s Correlation**

The Karl Pearson’s product-moment correlation coefficient (or simply, the Pearson’s correlation coefficient) is a measure of the strength of a linear association between two variables and is denoted by r or rxy(x and y being the two variables involved).

This method of correlation attempts to draw a line of best fit through the data of two variables, and the value of the Pearson correlation coefficient, r, indicates how far away all these data points are to this line of best fit.

* 1. **Analysis and Discussion**

**Figure 1.1 Bar Graph Showing the rating of Customer Service**

**Figure 2. 2 Line Graph Showing how likely a customer would refer a customer**

* 1. **Assumptions and Limitations**

The following limitations were observed:

1. The accuracy of the data collected could be improved if a larger sample size was selected to get feedback from a wider range of customers.
2. The researcher received less responses than anticipated as the research was conducted during the Corona epidemic and customers focus was on survival.
3. Data was not available from the other banks sharing that same market space. Data should have been collected from the competitor’s market share in order to gain insight on what the competitors are doing differently to attract new customers and to gain customer loyalty.

***530 words, 990 words missing***

1. **Deployment, Ethics and Conclusion**
   1. **Analysis of Risks and Challenges Associated with this deployment**

Some of the challenges associated with the deployment of this project is that management may not be susceptible to change. The organsation also face human resource limitations, as presently there are no persons trained in Data Science within the workforce . Data Analytics also requires computing resources and analytical tools which is a cost that management may not be willing to bear as we are expecting a plunge in the economy due to the Corona epidemic.

* 1. **Ethical Aspects of Deployment**

The researcher made the respondents aware of the content and purpose of the survey when they were given the questionnaires. Therefore, they are fully aware what the information provided will be used for.

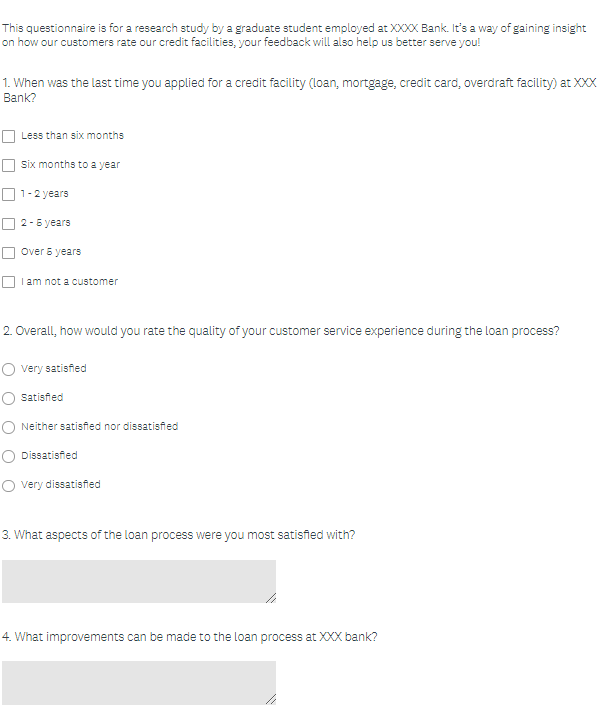
The report will be used internally hence the data collected will remain within the institution which is a highly trusted operative. Management will follow the same code of ethics when dealing with this information hence respondents is guaranteed confidentiality. The researcher has aslo kept the respondents of the survey details anonymous.

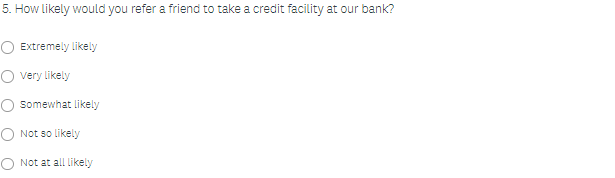
* 1. **Conclusion and Recommendations**

The following recommendations were made based on the findings of the study:

Appendix:

***Questionnaire:***





References:

Bryman, A., 2015. Social research methods. Oxford University Press. Available from:<https://www.academia.edu/30520568/Social_Research_Methods_4th_Edition_by_Alan_Bryman.pdf>  [Accessed March 2020].

Data Integration: A Theoretical Perspective, December 2001, Conference: Proceedings of the Twenty-first ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 3-5, Madison, Wisconsin, USA, Maurizio Lenzerini

*(PDF) Data Integration: A Theoretical Perspective*. Available from: <https://www.researchgate.net/publication/220266329_Data_Integration_A_Theoretical_Perspective> [accessed Apr 07 2020].

<https://www.globalscape.com/blog/5-types-data-integration>

<https://www.spss-tutorials.com/spss-variable-types-and-formats/>

<https://www.talend.com/resources/what-is-data-integration/>

<https://www.boostlabs.com/explanatory-exploratory-data-analytics-what-you-need-to-know/>

<https://www.northeastern.edu/graduate/blog/statistical-modeling-for-data-analysis/> By Tim Stobierski  |  October 29, 2019

<https://researchskills.epigeum.com/courses/researchskills/413/course_files/html/stm_1_10.html>

Peat J, Barton B. Medical Statistics: A guide to data analysis and critical appraisal. Blackwell Publishing; 2005. [[Google Scholar](https://scholar.google.com/scholar_lookup?title=Medical+Statistics:+A+guide+to+data+analysis+and+critical+appraisal.&author=J+Peat&author=B+Barton&publication_year=2005&)] [[Ref list](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3693611/#A3505R10)]

<https://www.toppr.com/guides/business-mathematics-and-statistics/correlation-and-regression/karl-pearsons-coefficient-correlation/>