|  |  |
| --- | --- |
| Library Management using Analytics and Big Data Technologies  MS5103: Business Analytics Project | **Name: Student ID:**    Doireann Casey 16435532  Issac Antu Joseph  19230120  Kapil Arora  19230069  Varun Dua  19230587    **Group Number:** 22  **Supervisor:** Dr Anatoli Nachev |



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  | **Declaration of Originality** | |
| **Project Details** |  |  |  |  |  |
| Module Code: | **MS5103** | | | Assignment Title: **Group 22** **M.Sc. Business Analytics - Major Project** | |

Group Members:

|  |  |  |
| --- | --- | --- |
| ***Student ID*** | ***Student Name*** | ***Contact Details (Email, Telephone)*** |
| 16435532 | Doireann Casey | D.Casey11@nuigalway.ie  +353 852022170 |
| 19230120 | Issac Antu Joseph | I.Joseph1@nuigalway.ie  +353 894858113 |
| 19230069 | Kapil Arora | K.Arora2@nuigalway.ie  +91 9999374373 |
| 19230587 | Varun Dua | V.Dua2@nuigalway.ie  +353 894876463 |

I/We hereby declare that this project is my/our own original work. I/We have read the University *Code of* *Practice for Dealing with Plagiarism\** and am/are aware that the possible penalties for plagiarism includeexpulsion from the University*.*

|  |  |
| --- | --- |
| ***Signature*** | ***Date*** |
| Doireann Casey | 29th June 2020 |
| Issac Antu Joseph | 29th June 2020 |
| Kapil Arora | 29th June 2020 |
| Varun Dua | 29th June 2020 |

\* http://www.nuigalway.ie/plagiarism

# Executive Summary

The following report explores the developments transpiring in library institutions to exploit the capabilities of analytics to help drive more informed decisions to better improve inventory management. Over the last number of years, analytics has become increasingly prevalent in diverse industries to uncover new insights and exhibit new values and opportunities. The defects of shelving inefficiencies identified in the James Hardiman Library provoked this study. Recognising the volume of qualitative data in the James Hardiman Library and the insights that could be exploited allowed this analysis to be materialised using the analytical platform that has been devised.

This report delves into the innovative capabilities of analytics in library services through enabling analysis to categorise demand into three distinct categories and evolving loan periods to correspond to the designated demand levels. The discovery of these idle resources has unlocked new and improved prospects for the James Hardiman Library and the analytical platform is crucial for this evolution.

Contents

[1 Introduction 4](#_Toc44313838)

[1.1 Objective 4](#_Toc44313839)

[1.2 Stakeholders 4](#_Toc44313840)

[2 Background 5](#_Toc44313841)

[3 Analytical Methodology 7](#_Toc44313842)

[3.1 Inspired Framework 7](#_Toc44313843)

[3.2 Overview of Data 7](#_Toc44313844)

[3.3 Tools Used 9](#_Toc44313845)

[3.4 Data Preparation 9](#_Toc44313846)

[3.5 Data Analysis Techniques 10](#_Toc44313847)

[4 Findings of Analyses 11](#_Toc44313848)

[4.1 Exploratory Analyses 11](#_Toc44313849)

[4.1.1 Exploratory Analyses on Stage 1 Data (Shannon Library) 11](#_Toc44313850)

[4.1.2 Exploratory Analyses on Stage 2 Data (Hardiman Library) 14](#_Toc44313851)

[4.2 Clustering 16](#_Toc44313852)

[4.2.1 Exploratory Clustering on Stage 1 Data (Shannon Library) 16](#_Toc44313853)

[4.2.2 Principal Component Analysis (PCA) on Stage 1 Data (Shannon Library) 19](#_Toc44313854)

[4.2.3 Finalised Clustering on Stage 2 Data (Hardiman Library) 20](#_Toc44313855)

[4.3 Classification (Hardiman Library) 22](#_Toc44313856)

[5 Discussion 27](#_Toc44313857)

[5.1 Clustering 27](#_Toc44313858)

[5.2 Classification 28](#_Toc44313859)

[6 Analytics Platform (R Shiny Application) 29](#_Toc44313860)

[6.1 Model Creation and Training 29](#_Toc44313861)

[6.2 Demand Categorisation and Dynamic Loan Periods 32](#_Toc44313862)

[7 Conclusion 33](#_Toc44313863)

[8 Appendix 34](#_Toc44313864)

[9 References 34](#_Toc44313865)

# Introduction

Showers, B. (2015). *Library analytics and metrics*. London: Facet Publishing.

“*Libraries, along with archives, museums, and galleries, find themselves ideally placed to exploit the full potential of analytics*.”

“*The variety and scope of the data collected and generated by libraries and organizations such as museums and archives are significant: transactional data on catalogue searches, item check-outs, log-ins to online resources and services, swipes through the entrance gates; manually collected statistics on space usage, student satisfaction, external visitors to the library. The applications of the data are equally varied and overlapping, including management functions (collections development and management, usage statistics), impact (demonstrating value, benchmarking, improving learner outcomes) and improving services and meeting user requirements (recommendation services, collections management/development).*”

James Hardiman Library at the National University of Ireland Galway has a functional BI environment, which collects both catalogue and transactional data. However, with limited analytical tools and reports, the data is under-utilized in producing insights on the routine functioning of the library and its management.

The following points were identified through our discussions with the stakeholders at the library:

* Books tend to run its course as newer editions are released to the market or newer books come in. Some books see an increase in demand during certain semesters. Shelf utilisation can be improved and made more efficient based on the seasonal demands of various books
* There are situations where high demand for a book can lead to other books in the same domain to be ignored. This can be countered through clustering different *tags* associated with a book to recommend a potential alternative to the user. The clustering of book *tags* may be done based on choices generated from a programme’s cohort
* Books are currently loaned for either 1,3 or 81 days. A more dynamic system can be used to gauge future demand for a book and adjust loan days accordingly. This can incur in a book being available for a longer time during lean periods

## Objective

Analyse catalogue and transaction data of a library to gather insights and create recommendations for:

* Categorise books in three different ways-
  + Rarely used
  + Requested a few times last year
  + In-demand books
* Shelving books efficiently based on demand (trends and seasonality)
* Predicting future requirements for books, journals, and articles across different categories
* Prescribing dynamic period for book loans

## Stakeholders

Primary stakeholder: Library staff

# Background

Over the past number of years, analytics has become an increasingly popular concept in the foreground of technology. The term Big Data was first coined in 2005 and its importance has continued to intensify its purpose over the years. The proliferation of analytics has become very applicable in today’s age establishing advantageous forecasts for enterprises.

 The demand for analytics has become an increasingly important aspect of library management in recent years. The incorporation of analytics has highlighted the huge opportunities that big data has on the integrated library system. Scholars in the library and information science claim an increase in the three V’s: volume, velocity, and variety provide an insight into the daily interactions of users in the library which offers more innovative developments to the current library management approach. According to Jim Tallman, an expert in technological growth within industries, “Libraries are 8-10 years behind other industries concerning analytics.” (Tay, 2016). Henceforth, as the technological world enhances its growth, libraries need to enhance their technological literacy.

Previous work demonstrated by Gerrard et al. (2018) recommended a contemporary approach to the preservation of big data in the long run. He proposed to improve the usefulness and capabilities to process digital resources at the volumes necessary for big data. Gerrard’s technique to manage big data in library management is due to the vast amount of catalogue and transactional data that is generated daily. Library big data is categorised into these two groups. Catalogue data consists of the inherent data and information of the library files while transactional data is the data generated via the users and the library itself.

The challenges associated with implementing data science within the confines of a diverse library environment are explored by two perspectives, the ‘skills gap’ and the ‘management gap.’ The skills gap is associated with teaching librarians’ computational skills and growing their knowledge in the effort to work efficiently with data. The ability of library managers to administer managerial and organisational support through understanding the value of data science is called the management gap.

Frameworks are implemented to leverage data science capabilities in decision-making and for the operational management of library services. The findings linked with academic libraries are displaced using a multi-faceted framework. This consists of two sections: Structures and skills, and services and stakeholders. Without the use of a multi-faceted framework, there is a wide variety of concerns that have an impact on the progress of integrating data science into library environments (Burton et al, 2018).

A close up of a sign

Description automatically generated

Figure 1: Multi-faceted framework for academic libraries

(Shifting to Data Savvy: The Future of Data Science In Libraries, 2018)

Burton et al (2018) mention that the structure and skills connect to form an axis that links the expertise of data-savvy librarians with organisational structures that can influence data science. Numerous drivers are recognised which include transforming research needs to improve productivity and generate new insights. This involves a joint initiative at both the local and wider community level. (Burton et al, 2018).

The other two components that form that axis are services and stakeholders. Drivers associated with services and stakeholders include physical space which is a place for innovations, creativity, and strategic planning to flourish. This will help with problem-solving which sheds light on hard-to-see problems. (Burton et al, 2018).

The development of this multi-faceted framework allows librarians to make the necessary changes to their library environment so that they can provide the highest possible service to their customers. Library professionals have recognised this strategy and it has been employed by numerous experts in the field. Fang et al. (2018) focused on bibliographical data and used a probabilistic generative topic model to examine hot and cold topics from electronic library references. The topics examined match the experts' results which provides value for libraries to generate more advanced services. Also, Yi et al. (2018) employed book-borrowing record data and used an association rule-mining algorithm to provide personalized recommendation service for the readers.

Having experienced the NUI Galway library first-hand, we recognised a gap in the library's services that we felt analytics could fulfil. The variety of data generated by libraries is significant which facilitates the incorporation of analytics to improve the day-to-day processes in the library environment. From personal experience, we identified major concerns associated with inventory management with reoccurring issues of certain books being unavailable due to insignificant and limited stock levels. This was raised to our attention as we feel adequate inventory management is required to maximise the potential of scholars. Recognising the ability of analytics in the library allowed us to aid in the improvement of book-keeping via enhanced shelving efficiency. Like big data, literature is continually evolving as new editions and new information become available and older versions become outdated. Enabling analytical tools to extract high or low demand books is one of our goals to extend prior work in the industry. Addressing the high level of idle and unused books and maximising the library's floor space is an effective practice of analytical forecasting. Using our interactive dashboard, we will be able to take an inside and informative look at current stock levels in the library and make real-time informed decisions as to what resources should be archived, increased, or updated. Employing our dashboard for analysing seasonality and trends is imperative for NUIG library staff to maximise inventory management. Evolving shelving efficiency in this manner builds on the existing knowledge of librarians and revolutionizes libraries into the 21st century. If successful, our project aims to bridge the ‘skills gap’ and ‘management gap’ for the library to operate at its true potential. By incorporating our easy to use interactive dashboard, ‘The James Hardiman Library’ in NUIG will be to the forefront of its field and in turn, these advancements will be funnelled to students leading to a more well-rounded experience. Our project should help the NUIG library increase and free up vital fixed space. This should in turn maximise student and staff college experience and provide required resources for a cohesive and successful degree. We also hope to increase physical space to allow students to flourish in a streamlined service.

# Analytical Methodology

A structured methodology can help design and understand a data mining project. A popular option for this is CRISP-DM, which stands for the Cross-Industry Standard Process for Data Mining (Piatetsky, 2014). CRISP-DM provides details of the typical phases involved in a data mining project, including the tasks to be involved in each phase as well as an explanation of how the tasks are related to each other (IBM Knowledge Center, 2012). IBM Knowledge Center states that the sequences of phases in the CRISP-DM model are not fixed and can move back and forth. This makes it flexible and customisable as per need.

## Inspired Framework

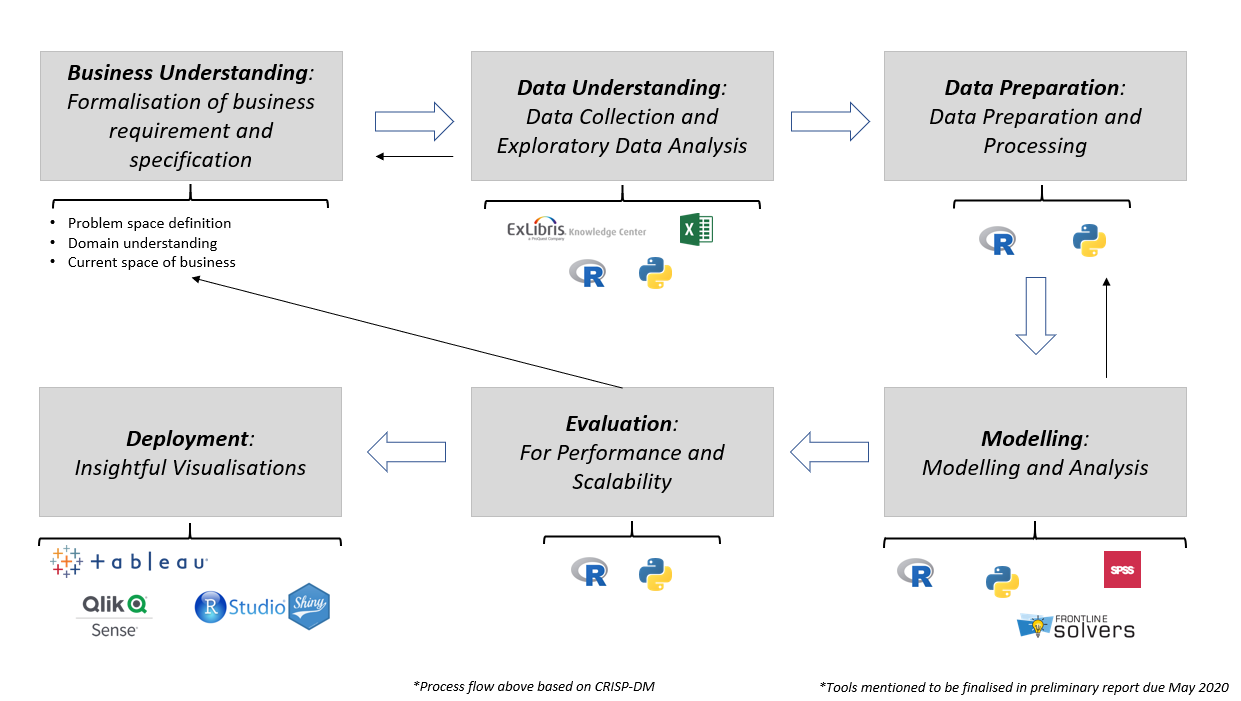


Figure 2: CRISP-DM Inspired Framework

\*Note: Tools depicted are identified in section 4.3

## Overview of Data

The data has been sourced from NUI Galway’s library network. Data has been sourced in two stages:

* **Stage 1**: 2019 data from Shannon College Library
* **Stage 2**: 2019 data from James Hardiman Main Library

Data dictionary of fields used-

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset** | **Field** | **Description** | **Field Values** |
| ***Library Loan Data -  Details about all assets that have been loaned from the library***  ***3088 records  [Primary Dataset]*** | Title | Title of the book | String |
| Edition | Edition of the book | String |
| Author | Author of the book | String |
| ISBN | 10- and 12-digit ISBN associated with the book | String |
| Publisher | Publisher of the book | String |
| Publication Date | The publication date of the book | Year |
| Barcode | Barcode assigned to book by the library | String |
| Patron Group | Patron group that has loaned out the book (Undergraduate, Staff, Postgraduate (Taught), External User (Borrower), DSS Undergraduate) | Categorical |
| Loan Date | Loan date of the book | Date |
| Due Date | Due date of the book | Date |
| Return Date | The actual return date of the loaned book | Date |
| Renewals | Number of renewals that have been triggered for the loaned book | Numerical |
| Recalls | Number of times the loaned book has been recalled in the loaned period | Numerical |
| Auto Renewals | Number of times the loaned book has been auto-renewed by the library management system | Numerical |
| Policy Name | 1 Day/ 14 Days/ 3 Hours/ 7 Days/ 84 Days Loan | Categorical |
| Item Policy | Standard loan, No loan, O/Access/Loan, Desk 3HR, Desk 24HR, No Item Policy, Library Use Only, None | Categorical |
| ***Reading Lists -  Contains reading lists that have been provided by different lecturers***  ***4790 records***  ***[Primary Dataset]*** | Title | Title of the book | String |
| Author(s) | Author of the book | String |
| List Appearance | Associated module from programme | Categorical |
| Importance | Whether the book is core/recommended/suggested read | Categorical |
| ISBN10 | ISBN formats associated with the books | Numerical |
| ISBN13 | Numerical |
| Volume | The volume of the book | String |
| Edition | Edition of the book | String |
| Publisher | Publisher of the book | String |
| Format | Electronic resource/hardcover/text/recording | Categorical |
| Period | The associated academic year for which the book has been prescribed | Categorical |
| Date Added | The date on which the book was added to the reading list | Date |
| ***Library Asset Requests -  Requests that have been made to loaned books***  ***213 records***  ***[Primary Dataset]*** | Title | Title of the book | String |
| Edition | Edition of the book | String |
| Author | Author of the book | String |
| ISBN | 10 and 12 digits ISBN associated with the book | String |
| Publisher | Publisher of the book | String |
| Publication Date | The publication date of the book | Year |
| # of requests | Number of requests that have been made for the book while it was loaned | Numerical |
| Request Date | Date of the book request | Date |
| ***Library Stocktake -  Details of all books available at the library***  ***4743 records***  ***[Auxiliary Dataset]*** | Title | Title of the book | String |
| Edition | Edition of the book | String |
| Author | Author of the book | String |
| ISBN | 10- and 12-digit ISBN associated with the book | String |
| Publisher | Publisher of the book | String |
| Publication Date | The publication date of the book | Year |
| Barcode | Barcode, assigned by the library, for the book | String |

Figure 3: Data Dictionary

## Tools Used

1. **ExLibris Alma**: Ex Libris Alma is a secure, scalable end-to-end library software system for managing the acquisition, sharing, cataloguing, and use of all kinds of resources, including physical and electronic books, physical and electronic periodicals, and digital resources (such as audio, image, and video files; ExLibris 2020)
2. **Microsoft Excel**: Microsoft Excel is a spreadsheet program that is used to record and analyse numerical data (Guru99 2020)
3. **R and RStudio**:
   1. R- R is a language and environment for statistical computing and graphics (R 2020)
   2. RStudio- RStudio is a development environment for R
4. **IBM SPSS**: A statistical analysis tool used for identifying the correlation between variables in this case

## Data Preparation

For preliminary data exploration, MS Excel was used, and this enabled data preparation for *clustering*. In terms of production usage, these steps will be reflected in R, using RStudio and will include sub-steps such as Data Reduction, Dimension Reduction, Outlier Management, and Missing Value Treatment.

Data preparation and transformation steps included:

* Unmerge and fill empty cells with required content across various datasheets:

Cleaning the datasheets ensures consistent and usable data by eliminating the risk of missing data which are commonly not accepted by algorithms and reducing the inefficiency of structural errors by unmerging the data.

* In *Library Loan Data,* 
  + Standardise *Auto-Renewal* entries

Ensuring the data is in the same format is crucial to establish accurate results for analysis.

* + Aggregate *Auto-Renewal*, *Recalls,* and *Renewals* at an asset level

Seeks identifiable patterns and trends within these columns for enhanced decision-making.

* In *Library Asset Requests Data,* aggregate *Requests* at an asset level

This step allows for meaningful data to be extracted for better decision-making and analysis.

* Merge *Library Loan Data* and *Library Asset Requests Data* to form a unified database at the library asset level and a monthly period
* The variables in the unified database include all variables noted above in Fig.3. for both Library Loan Data and Library Requests Data.

## Data Analysis Techniques

1. Exploratory Analysis – Using SPSS

Descriptive Statistics:

* + 1. Histograms of the variables *Renewals, Requests, Recall* and *Auto-Renewals*
    2. Scatter plots
    3. Spearman Correlation: Correlation is an analysis between two variables that measure the strength of association. Spearman Correlation is non-parametric and does not make assumptions with regards to data distribution (StatisticsSolutions n.d)

1. Clustering – Using R

Two major clustering algorithms deployed include Hierarchical Clustering and K-Means clustering. A stakeholder requirement is to label the historical data into three specific categories based on book demand. Due to the fixed number of clusters for this specific task, the K-means clustering technique has been used to create the clusters. For fulfilment of further requirements, in particular the identification of dynamic loan periods, other clustering methods will be explored, evaluated, and compared as per need.

A technique that can be used to manage and identify the optimum number of clusters is through a WSS Elbow Chart (Shapiro, 2018).

* Some features of K-Means Clustering:
  + 1. Clustering promotes a better understanding of data by grouping them into distinct groups. This is done based on various patterns in the data (Sharma, P 2019)
    2. All data points in a single cluster are like one another (Sharma, P 2019)
    3. Data points from different clusters should be as different as possible (Sharma, P 2019)
    4. K-Means clustering is relatively inexpensive to run, and the number of clusters can be specified in advance. It is ideal to be used as a precursor to other machine learning activities (StackOverflow, 2013)
* ggplot library in R: *ggplot* packages enable the creation of graphs to represent both univariate and multivariate numerical and categorical data utilizing elegant and complex plots. Grouping can be represented by colour, symbol, size, and transparency (Kabacoff, R 2017)

1. Principal Component Analysis (PCA) – Using R

Principal Component Analysis (PCA) is a dimension-reduction tool that can be used to reduce a large set of variables to a small set that still contains most of the information in the large set (Thorne 2019).

To select a subset of variables from a larger set, based on which original variables have the highest correlations with the principal component (Thorne 2019).

Principal Component Analysis (PCA) has been performed as an exploratory venture. A key aspect of the fields involved in the analyses is that they are dependent on the period as well as the size of the library under consideration. For example, with Shannon College library, the data fields responsible for book renewals and recalls are very sparsely populated. In such a situation, data interpretation from the Shannon College dataset may vary significantly when compared to similar interpretations made from the James Hardiman library data. The data present variability as per the period utilised for aggregation (e.g. month vs year).

Different fields have been used individually and in groups, as output fields for clustering activities. Usage of PCA ensures all possible variations in clustering have been attempted, before further discussion and interpretation with stakeholders.

1. Decision Tree Classification – Using R

Decision Trees are supervised learning algorithms used to continuously split data based on a certain parameter. These trees are built because of recursive partitioning, where data is split into smaller subsets, which are then split into further smaller subsets. This process continues until the algorithm can determine that the different subsets are sufficiently homogenous (Chakure 2019)

Chakure (2019) mentions that for a Decision Tree classifier, the algorithm starts at the root of the tree, and the data is split on a variable that returns the most homogenous branches. This process is iterative and continues until all samples at each leaf node belong to the same class.

Decision trees are relatively inexpensive to construct, fast, easy to interpret for smaller trees, and avoids unimportant variables/features (Chakure 2019).

# Findings of Analyses

Exploratory analyses, exploratory clustering, and principal component analysis were performed on both Stage 1 data (i.e. Shannon College Library data) and Stage 2 data (i.e. James Hardiman Library data)

## Exploratory Analyses

### Exploratory Analyses on Stage 1 Data (Shannon Library)

* Histogram distributions:

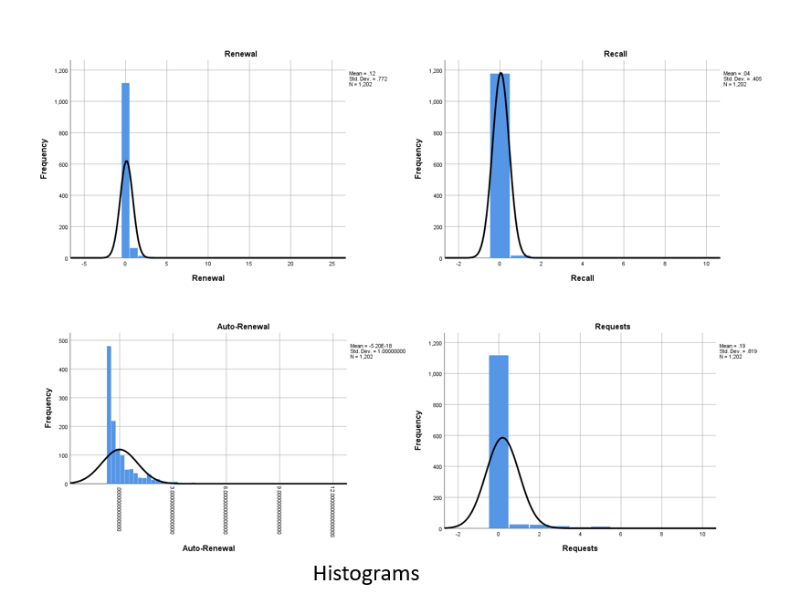


Figure 4: Histogram distributions on Stage 1 data

The histograms for Requests, Renewal, Recall, and Auto-Renewal are positively (right) skewed. However, most of the data is centred around values 0 to 10, and the long tail towards the right is sparsely distributed. Evidence of normal distributions enables the use of specific algorithms attuned to normally distributed data.

*The Shannon Library dataset is small. The range of variations in the fields Recalls and Renewals is very small. Most of the insights around demand would thus need to be inferred from either Auto-Renewal or Requests, as far as this dataset is concerned.*

* Scatter plots:

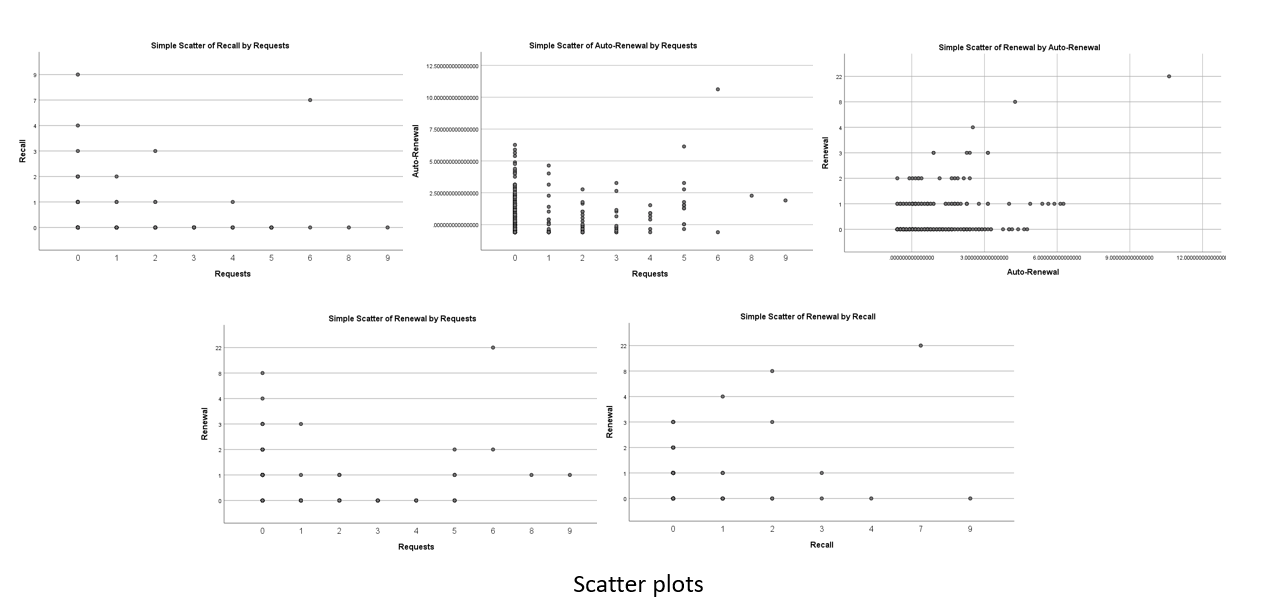


Figure 5: Scatterplots on Stage 1 data

The plots do not satisfy the pre-test condition for Pearson correlation test – Linearity and Homoscedasticity.

* Spearman Correlation Tests:
* All pairs of the 4 variables show a weak positive correlation.
* Renewal and Auto-Renewal show a stronger positive correlation, compared to other pairs (shown in the table below).
* All correlations are significant at 0.01 level of α.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Correlations** | | | | |
|  | | | Auto-Renewal | Renewal |
| Spearman's rho | Auto-Renewal | Correlation Coefficient | 1.000 | .313\*\* |
| Sig. (2-tailed) | . | .000 |
| N | 1202 | 1202 |
| Renewal | Correlation Coefficient | .313\*\* | 1.000 |
| Sig. (2-tailed) | .000 | . |
| N | 1202 | 1202 |

The correlation tests reveal that the 4 variables under observation are all positively correlated – indicating the possibility for dimension reduction or feature selection at later stages. This is especially true for the variables Renewal and Auto-Renewal which exhibit stronger positive correlation as discussed above.

### Exploratory Analyses on Stage 2 Data (Hardiman Library)

* Histogram distributions:

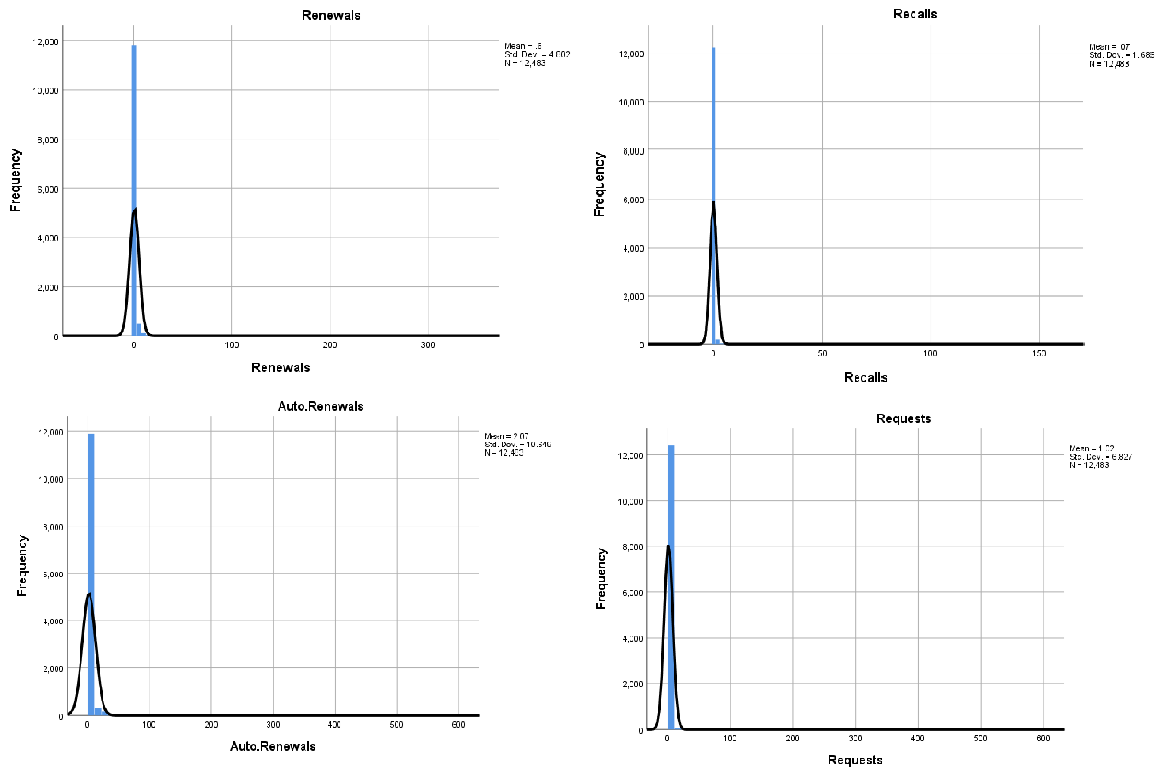


Figure 6: Histogram distributions on Stage 2 data

The histograms for Requests, Renewal, Recall, and Auto-Renewal are positively (right) skewed. However, most of the data is centred around values 0 to 10, and the long tail towards the right is sparsely distributed. Evidence of normal distributions enables the use of specific algorithms attuned to normally distributed data.

While variables such as Requests and Recalls show heterogeneous demand (i.e. demand from various students), Auto-Renewals and Renewals show demand from the current holder of the book. Additionally, Auto-Renewal, Renewal, and Recalls are slow-moving variables. Irrespective of the number of requests (2 or 20 or 100), a book can be recalled only once before it goes to the student next in the queue. The fields related to Renewals are updated in subsequent months. Thus, we will need to aggregate data over a larger period, if we want to have adequate data to classify demand.

* Scatter plots:

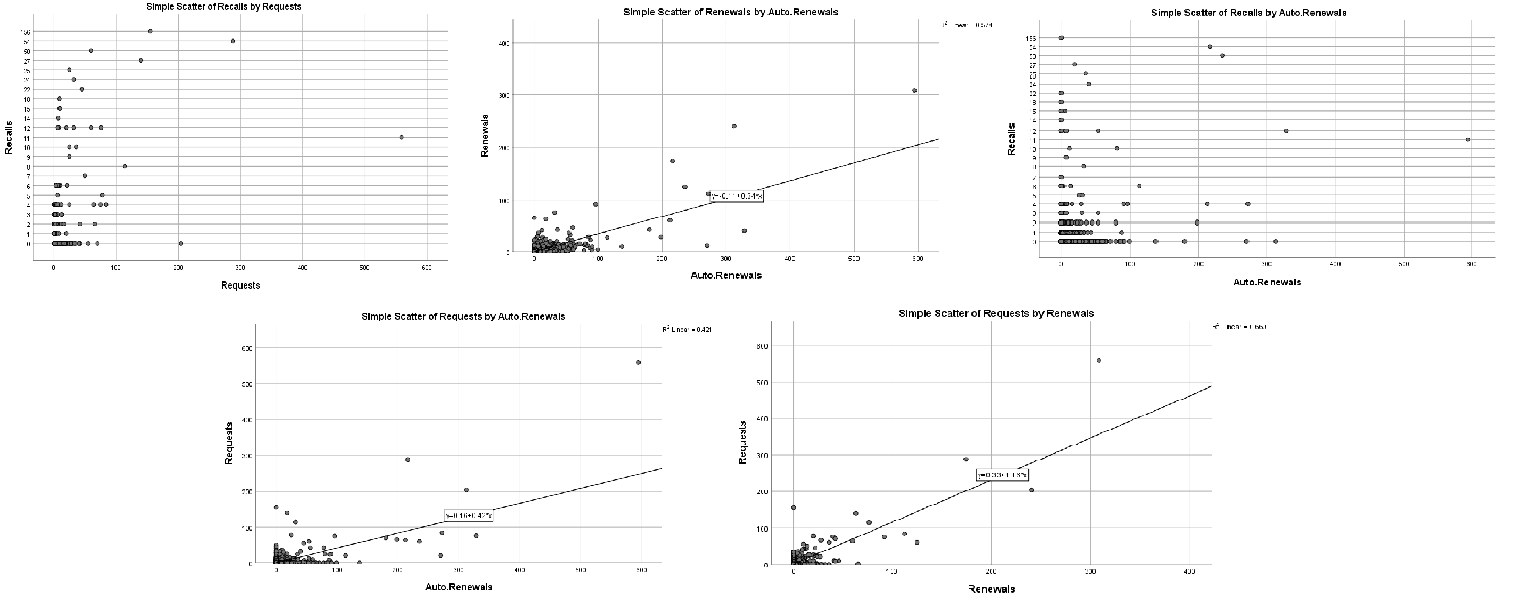


Figure 7: Scatterplots on Stage 2 data

The plots which map Recall against other variables do not satisfy the pre-test condition for Pearson correlation test – Linearity and Homoscedasticity.

Plots that contain any combination of the other 3 variables (Requests, Renewals, and Auto-Renewals) fairly satisfy the conditions of Linearity and Homoscedasticity. We run Pearson Correlation tests for these 3 variables and the Spearman correlation test for Recalls.

* Pearson Correlation Tests:
* All pairs of the 3 variables (Requests, Renewals, and Auto-Renewals) show a strong positive correlation.
* Renewal and Requests show a very strong positive correlation (shown in the table below).
* All correlations are significant at 0.01 level of α.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | Renewals | Auto.Renewals | Requests |
| Renewals | Pearson Correlation | 1 | .758\*\* | .813\*\* |
| Sig. (2-tailed) |  | .000 | .000 |
| N | 12483 | 12483 | 12483 |
| Auto.Renewals | Pearson Correlation | .758\*\* | 1 | .649\*\* |
| Sig. (2-tailed) | .000 |  | .000 |
| N | 12483 | 12483 | 12483 |
| Requests | Pearson Correlation | .813\*\* | .649\*\* | 1 |
| Sig. (2-tailed) | .000 | .000 |  |
| N | 12483 | 12483 | 12483 |

* Spearman Correlation Tests:
* All pairs against Recalls show a weak positive correlation.
* All correlations are significant at 0.01 level of α.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | | Renewals | Auto.Renewals | Requests |
|  | Recalls | Correlation Coefficient | .164\*\* | .140\*\* | .041\*\* |
| Sig. (2-tailed) | .000 | .000 | .000 |
| N | 12483 | 12483 | 12483 |

*Observations:*

1. *Recalls do not change like Renewals, Requests, and Auto. Renewals. This could be due to the way this field is updated (as discussed before, irrespective of the demand, recall field can increase only 1 point at a time)*
2. *While we can look at some form of dimension reduction if we want to cluster data based on the fields Renewals, Requests, and Auto.Renewals (as they are strongly correlated), we believe that (as discussed in the previous sections) these fields indicate different rates at which they portray demand. Requests are a more immediate form of update on the demand, while Renewals and Auto-Renewals take a longer time to reflect. Clubbing these two forms of fields would not be the best way to cluster our data.*
3. *We will cluster our data using two different sets of variables – {Requests, Recalls} and {Renewals, Auto-Renewals}. Performing our clustering in this way we can ensure that we can use these clusters for any future requirements in analysing short term demand (using only set 1 above) and long term demand (using a combination of both the sets above). Additionally, as we now know that Requests, Renewals, and Auto-Renewals are strongly positively correlated, we can safely say that the clusters formed from both the sets should mostly overlap.*

## Clustering

Clustering the data is important for achieving the following objectives:

1. Classification of book demand into 3 categories: Low, Moderate and High
2. Classification of book demand into different levels for dynamic loan periods

The dataset we have obtained from the library is not labelled. We need to use clustering to label our data into these categories. Once we have clustered and labelled our data, we can build models for classification of future data using the labelled data as our training and testing platform.

### Exploratory Clustering on Stage 1 Data (Shannon Library)

Concerning the data obtained from Shannon Library, we observe the following pattern for the Sum of Squared Errors at different values of cluster size. This chart can be used to pick the optimum number of cluster size. In this case, the chart depicts an optimum value of either 3 or 4. Since we are currently working on the Stage 1 data, our primary goal is exploration. We thus choose to go ahead with 3 clusters.

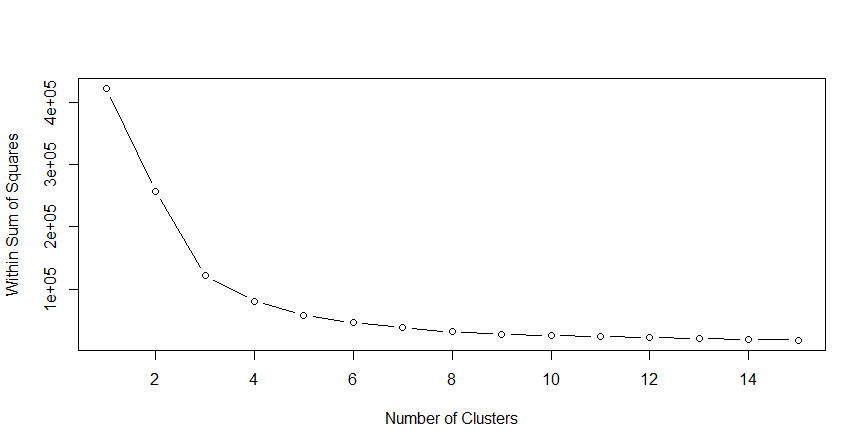


Figure 8: Elbow-chart from Stage 1 data

As mentioned above, our primary goal with stage 1 dataset is to explore the variables and their interactions. We have thus performed clustering in the following ways:

1. Based on individual variables (Requests, Renewals, Recalls, Auto-Renewals)
2. Based on all 4 variables together
3. Based on 2 variables at a time

* Clustering based on individual variables:

Clustering data based on Recalls and Renewals (independently) is resulting in the majority of the data being classified in the Low Demand category. This seems to be due to low variability in the data under these variables (the majority of the records are 0). This can be observed in the bar charts below:

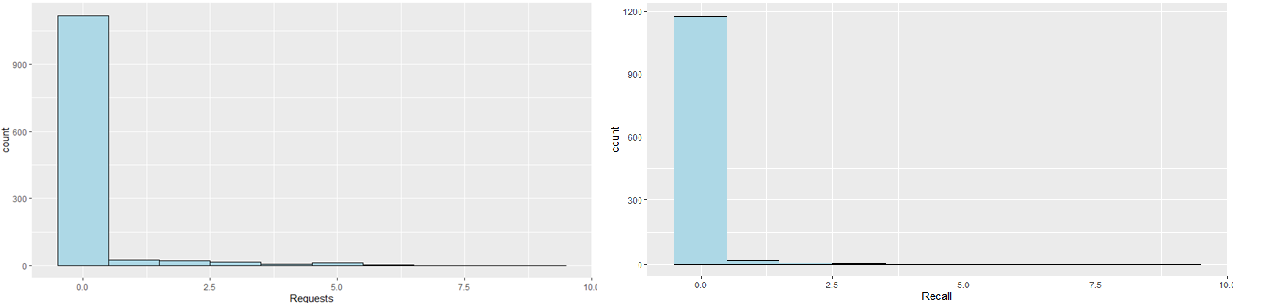


Figure 9: Cluster distributions across Request and Recall variables

Clustering based on Auto-Renewals and Requests (independently) depicts more balanced clusters, thus these two variables seem to be projecting more variances in demand. The variability in their data can be seen below. The values in the field Auto-Renewals is standardised (thus some values are below zero).

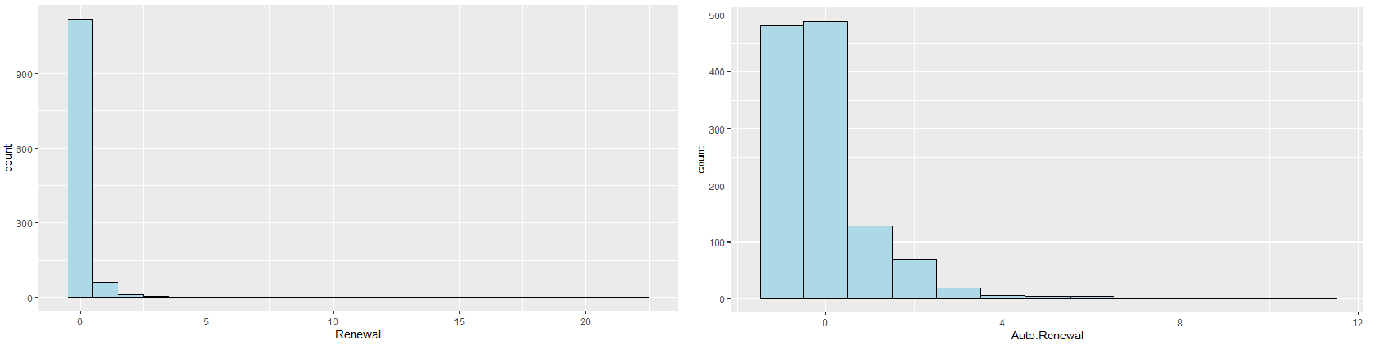


Figure 10: Cluster distributions across Renewal and Auto-Renewal variables

* Clustering based on all variables together:

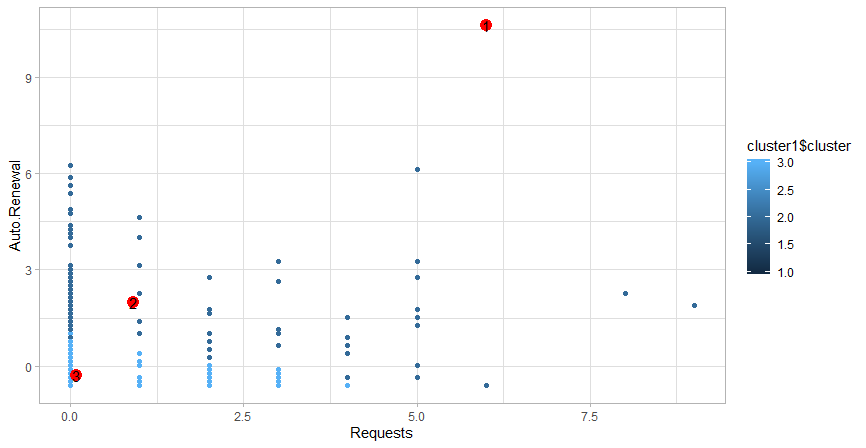


Figure 11: Clustering based on Request and Auto-Renewal variables

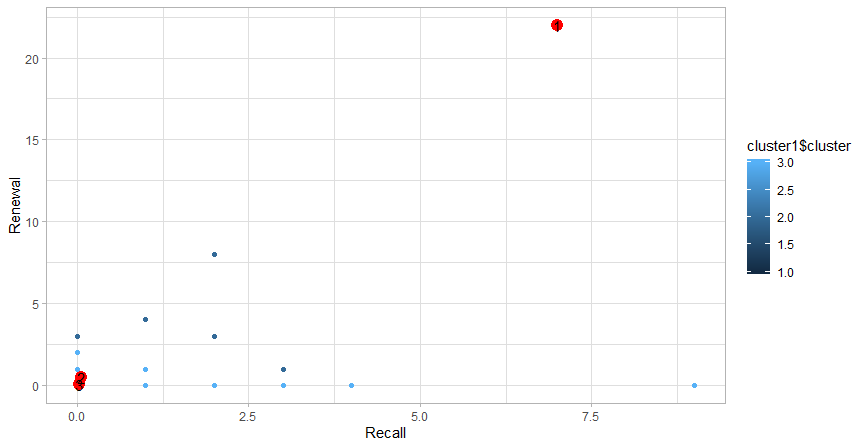


Figure 12: Clustering based on Renewal and Recall variables

As expected, Recalls and Renewals drive the data points into Low Demand clusters, due to the low variability of these features. Clustering data based on all 4 variables is not the best plausible solution in the case of the Shannon Library dataset.

* Clustering based on *Requests* and *Auto-Renewals*:

Requests and Auto Renewals have been considered for this round of clustering due to the lack of variation for other two data fields Higher Requests can be taken as a sign of greater demand, while higher Auto-Renewals, although not a real sign of demand, still keeps books out of circulation and can be interpreted as “Medium” level of demand

Based on the above considerations, clustering based on Requests and Auto-Renewals should lead to better classification of the data (for Shannon Library)

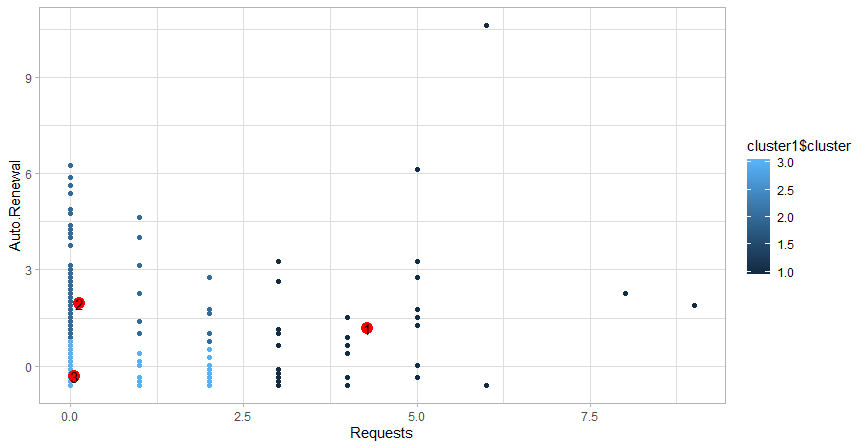


Figure 13: Clustering based on Request and Auto-Renewal variables

### Principal Component Analysis (PCA) on Stage 1 Data (Shannon Library)

The cluster creation process results in the data that can be plotted in different dimensions. PCA attempts to identify the direction of the movement of the data (David & Jacobs, 2014). For example, the dataset in hand contains four primary input variables. The count of variables has been reduced to two variables and this reduction will plot the data almost the same as the initial four variables from the dataset. The latter two variables can be considered as the principal components and have been used in place of the former four variables for further analyses.

As we have seen in the previous section, Recalls and Renewals do not have much variance in their values. While removing them from our cluster analysis could be one of the solutions to obtain a better output (as seen in section 5.2.1), another solution is PCA. As discussed in section 4.5, PCA will help us identify alternate variables that capture the intricacies of the data, in a lesser number of dimensions. In other words, PCA uses old characteristics to create new components that can summarise our data better.

* Summary of PCA and Rotation Matrix:

## Importance of components:  
## PC1 PC2 PC3 PC4  
## Standard deviation 1.3778 0.9602 0.8629 0.6597  
## Proportion of Variance 0.4746 0.2305 0.1861 0.1088  
## Cumulative Proportion 0.4746 0.7051 0.8912 1.0000

## Rotation (n x k) = (4 x 4):  
## PC1 PC2 PC3 PC4  
## Renewal 0.6074635 -0.2145690 -0.08979826 -0.759529089  
## Recall 0.4900378 -0.5450755 0.47171661 0.490141961  
## Auto Renewal 0.5120104 0.2322040 -0.70787720 0.427593891  
## Requests 0.3587428 0.7764883 0.51800539 0.006315521

The Rotation matrix above determines the direction of the new feature space (this feature space is plotted below).

From the above set of components obtained via PCA, we observe that:

* Approx. 90% of the variances in the data is captured by the 3 components PC1, PC2, PC3.
  + Approx. 70% of the variances in the data is captured by the 2 components PC1, PC2.

Plotting the projection of PC1, PC2, and PC3 on the data:

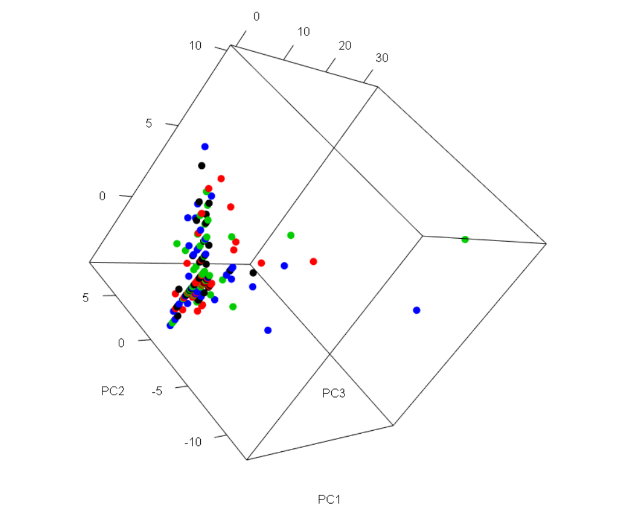


Figure 13: Clustering based on Request and Auto-Renewal variables

* Clustering the data based on PC1, PC2, and PC3:

As the principal components capture variability in all the data variables, the above set of clusters also accounts for the low variability of Renewals and Recalls. The clusters obtained using the principal components are very similar to the ones obtained with {*Requests, Auto-Renewals*} in section 5.2.1.

*Findings and insights from the above exploratory analyses were discussed with stakeholders. The use of data from James Hardiman Library (i.e. Stage 2 data) can ensure a higher degree of variability. Therefore, final clustering and classification activities were performed on Stage 2 data.*

### Finalised Clustering on Stage 2 Data (Hardiman Library)

Based on observation from exploratory clustering, the clustering activity on Hardiman Library data has been performed on two different sets of variables. This is due to the inherently different nature of the viable variables involved. For example, *Requests* and *Recalls* are variables that are focussed on immediate changes in demand while *Renewals* and *Auto-renewals* are dependent on updates that happen in subsequent months based on initial demand.

Additionally, as we are using 2 groups of variables (each set with 2 variables), we will not be requiring any form of Principal Component Analysis to reduce the dimensionality of the data.

**Group 1 clusters based on *Requests* and *Recalls*:**

Based on the WSS chart (elbow chart), below, a minimum value of 10 clusters has been identified to be the optimum number of clusters for this data (and variables). However, we have used only 3 variables to denote the demand across these 10 clusters (*Low, Moderate, High*). The definition of the range of these labels is subjective and can be altered based on credibility and feasibility from the stakeholders.

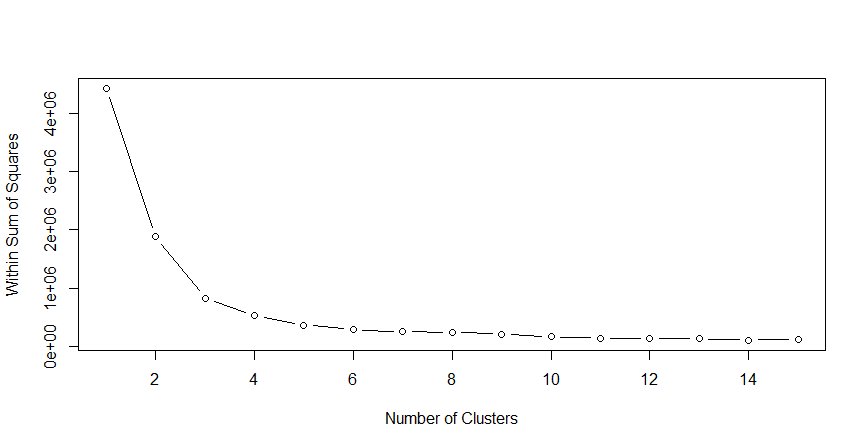


Figure 14: Elbow chart from Group 1 clustering

**Group 2 clusters based on *Renewals* and *Auto-renewals*:**

Based on the WSS chart (elbow chart), below, a minimum value of 10 clusters has been identified to be the optimum number of clusters for this data (and variables). However, we have used only 3 variables to denote the demand across these 10 clusters (*Low, Moderate, High*). The definition of the range of these labels is subjective and can be altered based on credibility and feasibility from the stakeholders.

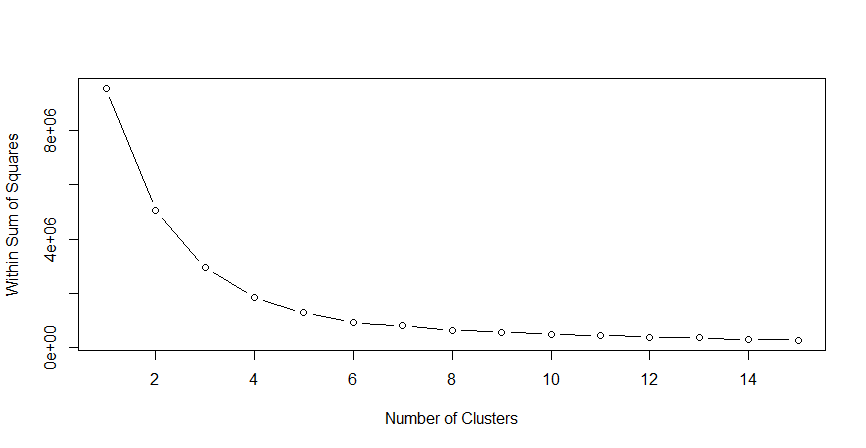


Figure 15: Elbow chart from Group 2 clustering

## Classification (Hardiman Library)

Based on the output clusters we obtained from the previous discussion, we labelled our data into Low, Moderate, and High categories. This accomplishes our objective of categorizing existing data into 3 different categories based on demand.

Closely tied to this objective, we next want to use the labelled data as a source of information for a model to classify demand based on new data from the Library. Thus, our labelled data now becomes our training and test data for a Decision Tree Classification model. The dependent variables, in this case, are the labels we have generated: Demand and Loan Period, while the independent variables are the number of Requests, Recalls, Renewals, and Auto-Renewals.

Two Decision Tree models were utilised for classification purposes. The difference between the two models lies in the splitting criterion used; while one uses the Gini Index, the other uses Information Gain. Both the models displayed the same results and precision rates, and thus we have decided to go ahead with the Information Gain DT model.

Initially, we split the data into the ratio 80:20 for training and testing. Based on this test strategy, we achieved good test results. We plan to use other test strategies in the future to ensure our models give comparable results across all test strategies. The confusion matrices from the classification activities are provided below.

* Demand-based Classifications

The Charts provided below depict how our models classify demand based on the two sets of variables. The first model is based on Requests and Recalls, while the second one is based on Renewals and Auto-Renewals. For example, in the first case below, if the number of Requests in a new data row is less than 4, then according to the decision tree below, this data row is classified as the Low Demand category.

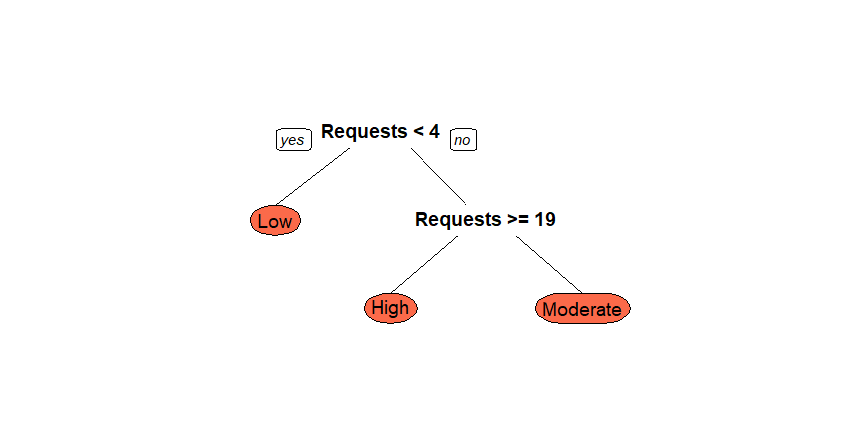
The statistics below the charts depict the accuracy of the prediction of the respective model. In the first case below, the model predicts the class of the demand with 99.98 % accuracy. This high level of accuracy should be expected because we have manually cluster and labelled our data, before building our classification models on it.

Figure 15: Decision Tree based on {Requests, Recalls}

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction High Low Moderate  
## High 30 0 0  
## Low 0 5154 1  
## Moderate 0 0 81  
##   
## Overall Statistics  
##   
## Accuracy : 0.9998   
## 95% CI : (0.9989, 1)  
## No Information Rate : 0.9787   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.9954   
##   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: High Class: Low Class: Moderate  
## Sensitivity 1.000000 1.0000 0.98780  
## Specificity 1.000000 0.9911 1.00000  
## Pos Pred Value 1.000000 0.9998 1.00000  
## Neg Pred Value 1.000000 1.0000 0.99981  
## Prevalence 0.005697 0.9787 0.01557  
## Detection Rate 0.005697 0.9787 0.01538  
## Detection Prevalence 0.005697 0.9789 0.01538  
## Balanced Accuracy 1.000000 0.9955 0.99390

The model below depicts the Decision Tree for Demand classification based on {Renewals, Auto-Renewals}

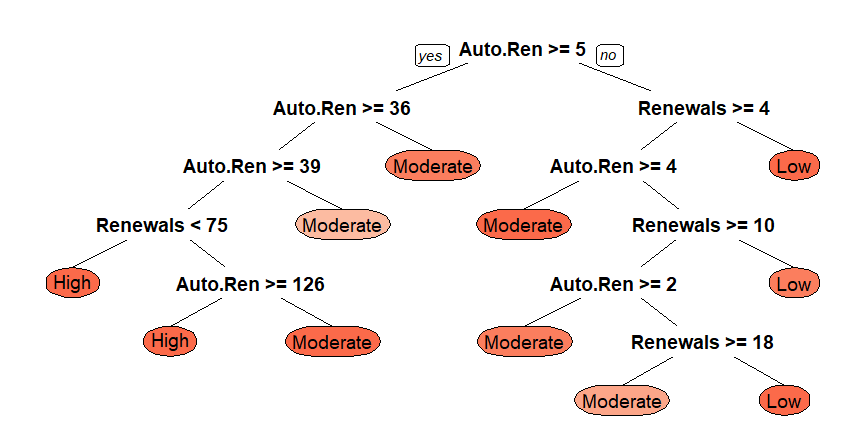


Figure 16: Decision Tree based on {Renewals, Auto-Renewals}

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction High Moderate Low  
## High 56 1 0  
## Moderate 0 653 2  
## Low 0 0 4554  
##   
## Overall Statistics  
##   
## Accuracy : 0.9994   
## 95% CI : (0.9983, 0.9999)  
## No Information Rate : 0.8652   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.9976   
##   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: High Class: Moderate Class: Low  
## Sensitivity 1.00000 0.9985 0.9996  
## Specificity 0.99981 0.9996 1.0000  
## Pos Pred Value 0.98246 0.9969 1.0000  
## Neg Pred Value 1.00000 0.9998 0.9972  
## Prevalence 0.01063 0.1242 0.8652  
## Detection Rate 0.01063 0.1240 0.8648  
## Detection Prevalence 0.01082 0.1244 0.8648  
## Balanced Accuracy 0.99990 0.9990 0.9998

* Dynamic Loan Period Classification

The Charts provided below depict how our models classify loan periods based on the two sets of variables. The first model is based on Requests and Recalls, while the second one is based on Renewals and Auto-Renewals. For example, in the first case below, if the number of Requests in a new data row is less than 4 and the number of Recalls is less than 6, then according to the decision tree below, this data row is classified into a 10-day loan category. (Note: the number of days at present is linearly divided based on the demand levels. The stakeholders at the library will decide the number of days in the final version of the product).

The statistics below the charts depict the accuracy of the prediction of the respective model. In the first case below, the model predicts the class of the demand with 99.94 % accuracy. This high level of accuracy should be expected because we have manually clustered and labelled our data, before building our classification models on it.

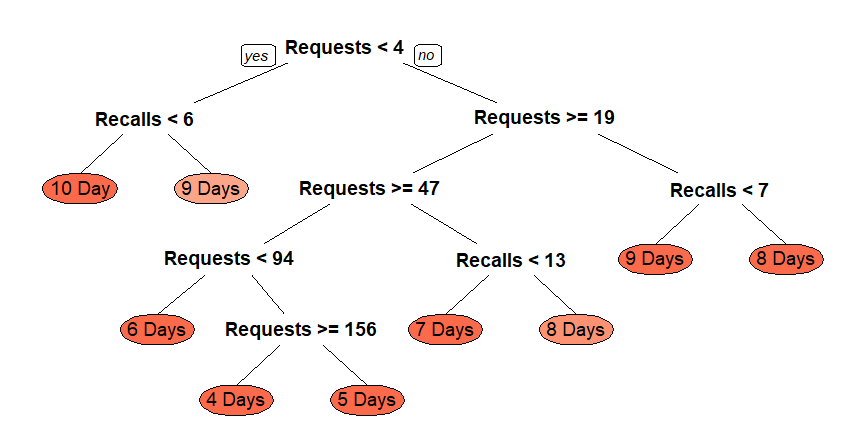


Figure 17: Decision Tree based on {Requests, Recalls}

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 4 Days 3 Days 6 Days 1 Day 5 Days 10 Day 9 Days 7 Days 2 Days 8 Days  
## 4 Days 2 0 0 0 0 0 0 0 0 0  
## 3 Days 0 0 0 0 0 0 0 0 0 0  
## 6 Days 0 0 7 0 0 0 0 0 0 0  
## 1 Day 0 0 0 0 0 0 0 0 0 0  
## 5 Days 1 0 0 0 3 0 0 0 0 0  
## 10 Day 0 0 0 0 0 5154 0 0 0 0  
## 9 Days 0 0 0 0 0 0 76 0 0 0  
## 7 Days 0 0 0 0 0 0 0 14 0 1  
## 2 Days 0 0 0 0 0 0 0 0 0 0  
## 8 Days 0 0 0 0 0 0 0 1 0 4  
##   
## Overall Statistics  
##   
## Accuracy : 0.9994   
## 95% CI : (0.9983, 0.9999)  
## No Information Rate : 0.9793   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.986   
##   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: 4 Days Class: 3 Days Class: 6 Days Class: 1 Day  
## Sensitivity 0.66667 NA 1.00000 NA  
## Specificity 1.00000 1 1.00000 1  
## Pos Pred Value 1.00000 NA 1.00000 NA  
## Neg Pred Value 0.99981 NA 1.00000 NA  
## Prevalence 0.00057 0 0.00133 0  
## Detection Rate 0.00038 0 0.00133 0  
## Detection Prevalence 0.00038 0 0.00133 0  
## Balanced Accuracy 0.83333 NA 1.00000 NA  
## Class: 5 Days Class: 10 Day Class: 9 Days Class: 7 Days  
## Sensitivity 1.00000 1.0000 1.00000 0.93333  
## Specificity 0.99981 1.0000 1.00000 0.99981  
## Pos Pred Value 0.75000 1.0000 1.00000 0.93333  
## Neg Pred Value 1.00000 1.0000 1.00000 0.99981  
## Prevalence 0.00057 0.9793 0.01444 0.00285  
## Detection Rate 0.00057 0.9793 0.01444 0.00266  
## Detection Prevalence 0.00076 0.9793 0.01444 0.00285  
## Balanced Accuracy 0.99990 1.0000 1.00000 0.96657  
## Class: 2 Days Class: 8 Days  
## Sensitivity NA 0.80000  
## Specificity 1 0.99981  
## Pos Pred Value NA 0.80000  
## Neg Pred Value NA 0.99981  
## Prevalence 0 0.00095  
## Detection Rate 0 0.00076  
## Detection Prevalence 0 0.00095  
## Balanced Accuracy NA 0.89990

The model below depicts the Decision Tree for Loan Period classification based on {Renewals, Auto-Renewals}

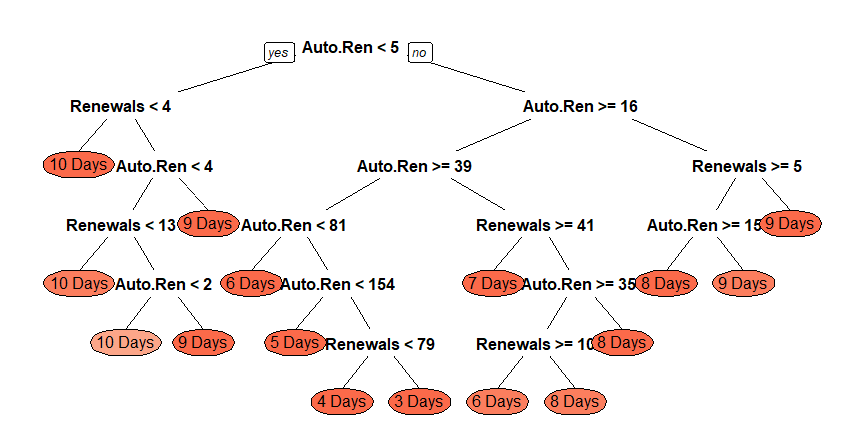


Figure 18: Decision Tree based on {Renewals, Auto-Renewals}

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 6 Days 5 Days 7 Days 4 Days 10 Days 8 Days 3 Days 1 Day 9 Days 2 Day  
## 6 Days 40 0 1 0 0 0 0 0 0 0  
## 5 Days 0 8 0 0 0 0 0 0 0 0  
## 7 Days 0 0 1 0 0 0 0 0 0 0  
## 4 Days 0 0 0 3 0 0 0 0 0 0  
## 10 Days 0 0 1 0 4556 0 0 0 2 0  
## 8 Days 1 0 0 0 0 229 0 0 0 0  
## 3 Days 0 0 0 1 0 0 3 0 0 1  
## 1 Day 0 0 0 0 0 0 0 0 0 0  
## 9 Days 0 0 2 0 0 0 0 0 419 0  
## 2 Day 0 0 0 0 0 0 0 0 0 0  
##   
## Overall Statistics  
##   
## Accuracy : 0.9983   
## 95% CI : (0.9968, 0.9992)  
## No Information Rate : 0.8648   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.993   
##   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: 6 Days Class: 5 Days Class: 7 Days Class: 4 Days  
## Sensitivity 0.975610 1.000000 0.2000000 0.7500000  
## Specificity 0.999809 1.000000 1.0000000 1.0000000  
## Pos Pred Value 0.975610 1.000000 1.0000000 1.0000000  
## Neg Pred Value 0.999809 1.000000 0.9992406 0.9998101  
## Prevalence 0.007783 0.001519 0.0009491 0.0007593  
## Detection Rate 0.007593 0.001519 0.0001898 0.0005695  
## Detection Prevalence 0.007783 0.001519 0.0001898 0.0005695  
## Balanced Accuracy 0.987709 1.000000 0.6000000 0.8750000  
## Class: 10 Days Class: 8 Days Class: 3 Days Class: 1 Day  
## Sensitivity 1.0000 1.00000 1.0000000 NA  
## Specificity 0.9958 0.99980 0.9996201 1  
## Pos Pred Value 0.9993 0.99565 0.6000000 NA  
## Neg Pred Value 1.0000 1.00000 1.0000000 NA  
## Prevalence 0.8648 0.04347 0.0005695 0  
## Detection Rate 0.8648 0.04347 0.0005695 0  
## Detection Prevalence 0.8654 0.04366 0.0009491 0  
## Balanced Accuracy 0.9979 0.99990 0.9998101 NA  
## Class: 9 Days Class: 2 Day  
## Sensitivity 0.99525 0.0000000  
## Specificity 0.99959 1.0000000  
## Pos Pred Value 0.99525 NaN  
## Neg Pred Value 0.99959 0.9998102  
## Prevalence 0.07992 0.0001898  
## Detection Rate 0.07954 0.0000000  
## Detection Prevalence 0.07992 0.0000000  
## Balanced Accuracy 0.99742 0.5000000

# Discussion

## Clustering

Discussion based on the analysis and results in section 5.2.3:

The data have been clustered on two sets of variables:

* Group 1 (using *Requests* and *Recalls*): This indicates the current heterogeneous demand for books. These variables are immediately affected by changes in demand. For example, any request against a book reflects immediately in this month’s data. Similarly, a recall is reflected immediately in the data
* Group 2 (using *Renewals* and *Auto-renewals*): These variables tend to change slowly based on demand. For example, if a book is loaned in a month, the renewals, and auto-renewals against the book, if any, will be updated in subsequent months. Additionally, these fields do not portray demand from individuals other than the borrower, limiting their predictive value.

Such differentiation is important for two reasons:  
1. Group 1 cluster can be used to build models that classify data based on short-term past data (for example previous month’s data).  
2. Group 2 on the other hand, will be effective with long term data (aggregated monthly), for example, 1 years’ data aggregated per month

Together, the two groups of clusters can be used to gain better insights into the trends in demand.

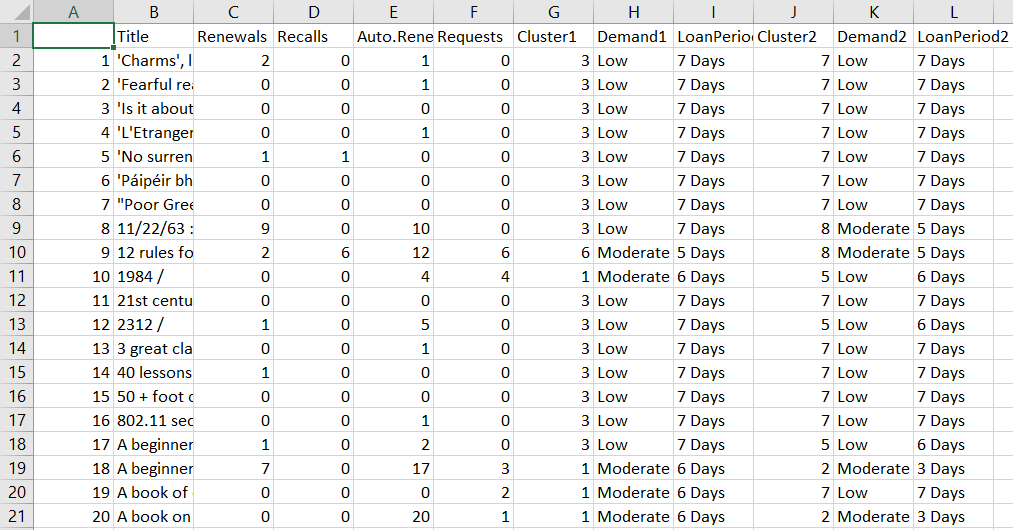


Figure 19: Screenshot of file with categorisation output

From the above snippet, *Demand1* refers to clusters formed based on *Request* and *Recall* variables and *Demand2* refers to clusters formed based on *Renewal* and *Auto-renewal* variables. The clusters formed through the separate clustering activities have been encapsulated into the keywords: Low, Moderate, and High. This has been done to improve ease in understanding and consuming the demand clusters generated by the two clustering activities. The *LoanPeriod* variables have been included, on a trial basis, due to the requirement to have dynamic loan period for library assets and would serve as a micro-level classification for the same (current methodology relies on fixed loan periods that have been set for different assets).

The clustering activities have helped in assigning demand tags to each library asset. This will enable higher efficiency and accuracy with regards to classification

## Classification

Discussion based on the result and analysis from section 5.3:

A high level of accuracy and precision can be seen with different decision tree activities performed as part of classifying books. This can be attributed to clustering and tagging the library assets (from previous steps; refer above). The models used are built on *Information Gain*

The models' aid in two primary objectives:

* Demand-based Classification: Classifying library assets based on “High”, “Moderate” and “Low” demand, leveraging the two clustering groups and subsequent decision tree classification
* Dynamic Loan Period: Assigning dynamic loan periods for library assets, in place of fixed loan periods that are currently used. Dynamic loan periods can ensure assets keep up with demand surrounding them

# Analytics Platform (R Shiny Application)

As a final deliverable to the client (Hardiman library), we have developed an Analytics Platform for library management. It covers the following functionalities:

1. Model Creation and Training
2. Demand Categorisation
3. Dynamic Loan Periods

## Model Creation and Training

Under this tab, the user is prompted to upload data, run clustering and classification, evaluate charts and models, and select labels for the clusters. A basic summary and chronology of features is as follows:

1. Upload Loan and Request Data Files - User is prompted to upload multiple data files associated with Loans and Requests.

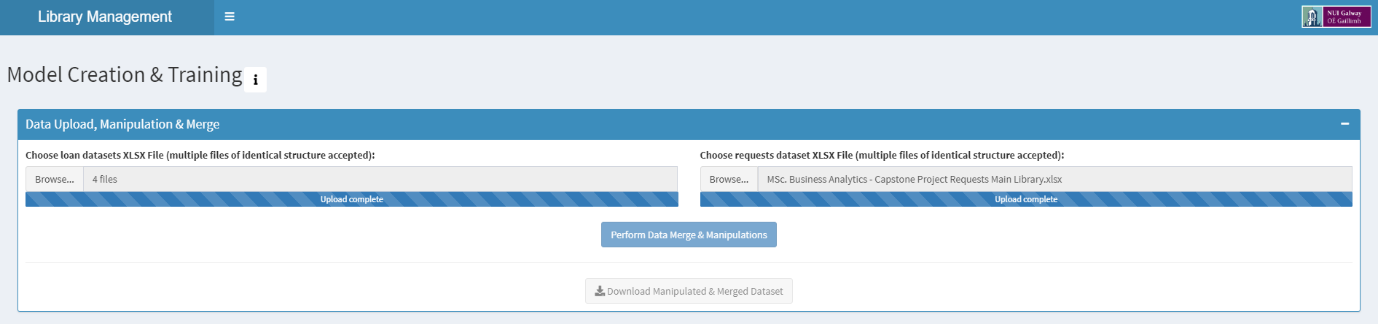


Figure 20: Screenshot of Tool: Data Upload

1. Run clustering algorithm and produce WSS Charts - Clustering is done to label the raw data based on their features. Optimum number of clusters can be determined with the WSS Chart. Associated User Manual, available on the tool, lists stepwise procedure for audience with no technical experience.

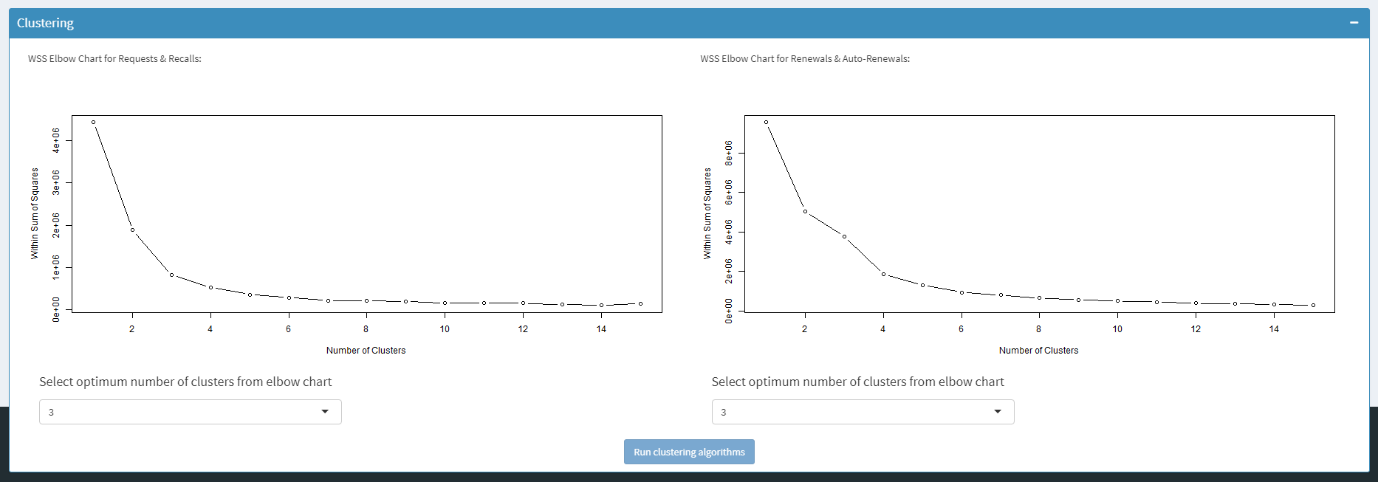


Figure 21: Screenshot of Tool: WSS Charts

1. Obtain user input on cluster labels

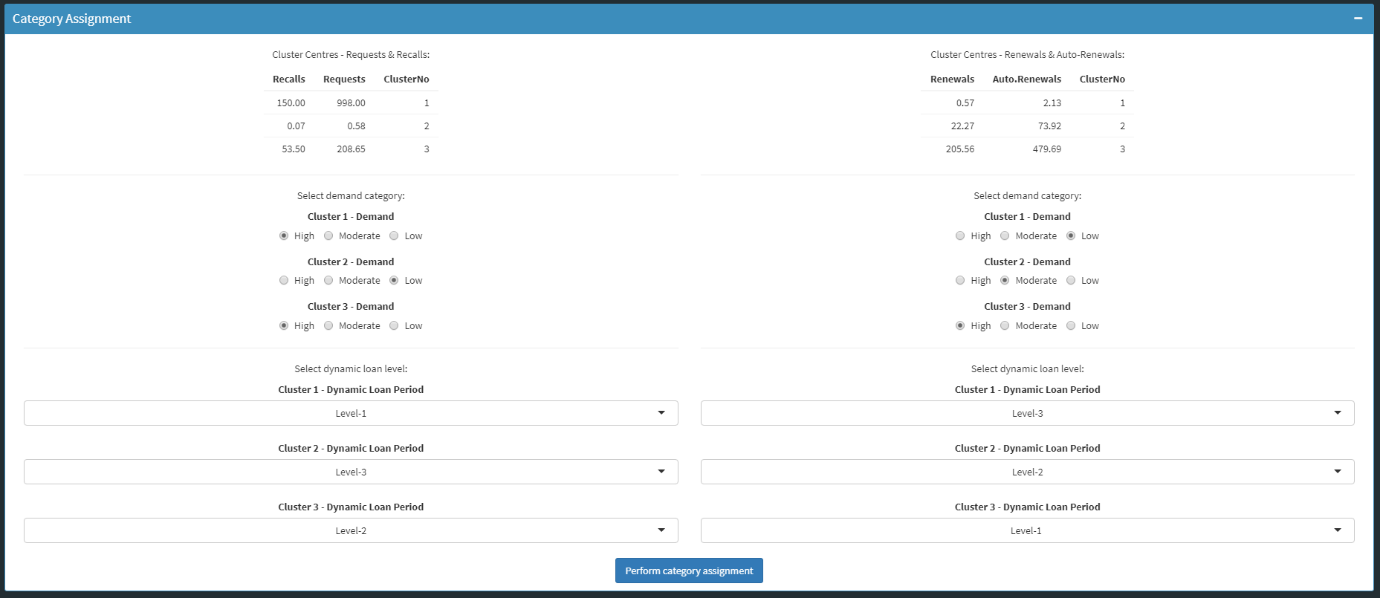


Figure 22: Screenshot of Tool: Cluster Labels

1. Build Classification Model on labelled data

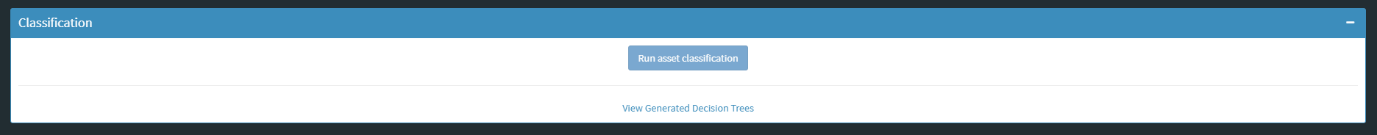


Figure 23: Screenshot of Tool: Classification Button

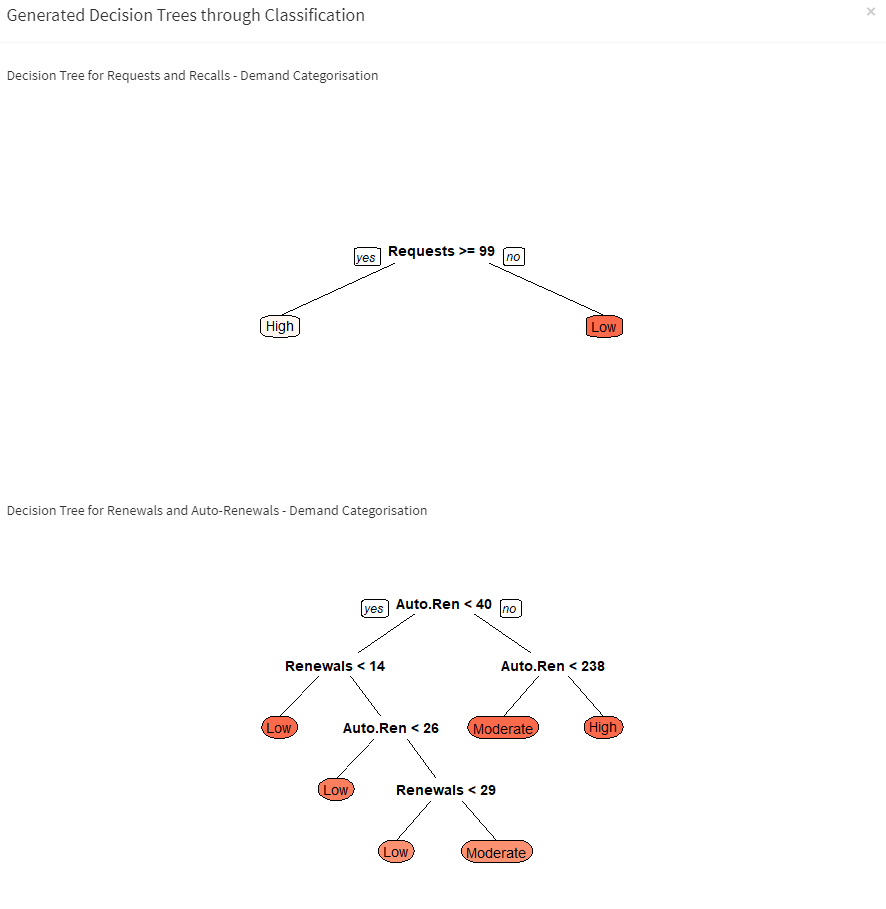


Figure 24: Screenshot of Tool: Decision Trees

1. Upload New Data for Classification

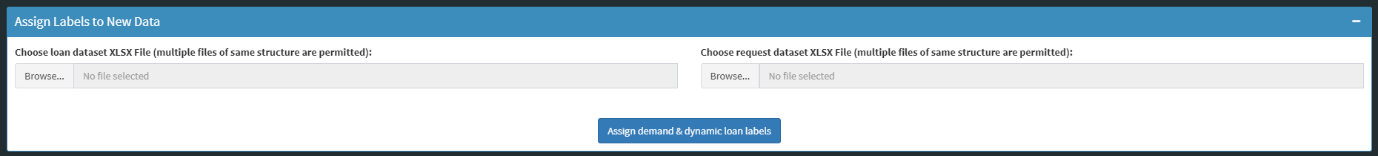


Figure 25: Screenshot of Tool: New Data for Classification

There are two paths that a user can follow in this tab. One path allows the user to run all the steps above, chronologically. The other path allows the user to utilise an older data file containing clustered data, to run only classification steps, i.e., steps 4 and 5.

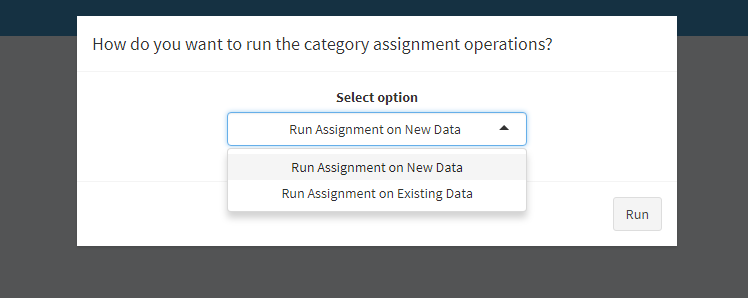


Figure 26: Screenshot of Tool: Two Paths

This part of the tool expects users to have some basic understanding of reading outputs from clustering algorithms and WSS charts. To help the user navigate through these tasks, some resources have been included along with the tool - these include pop-up documents for help, and a user manual.

This part of the application covers the objectives related to clustering data and classifying it into demand levels and dynamic loan periods. However, prediction and classification tasks are not one-off events. The models and outputs are bound to change with changes in data - factors such as size, velocity, etc, affect the models and their classifications. To provide a platform that can handle such changes in the future, this tool was built to dynamically work with new data, build fresh models and update outputs based on the new models.

## Demand Categorisation and Dynamic Loan Periods

The other two tabs for Demand Categorisation and Dynamic Loan Periods are descriptive dashboards for providing tabular and visual information on the classifications. In its current form, the application has no database or data models to store information. Thus, every new instance of the application will not be able to recall models created in the previous session. In every new session, the user is expected to create and run the classification models first, and then move on to the dashboards for visual representations.

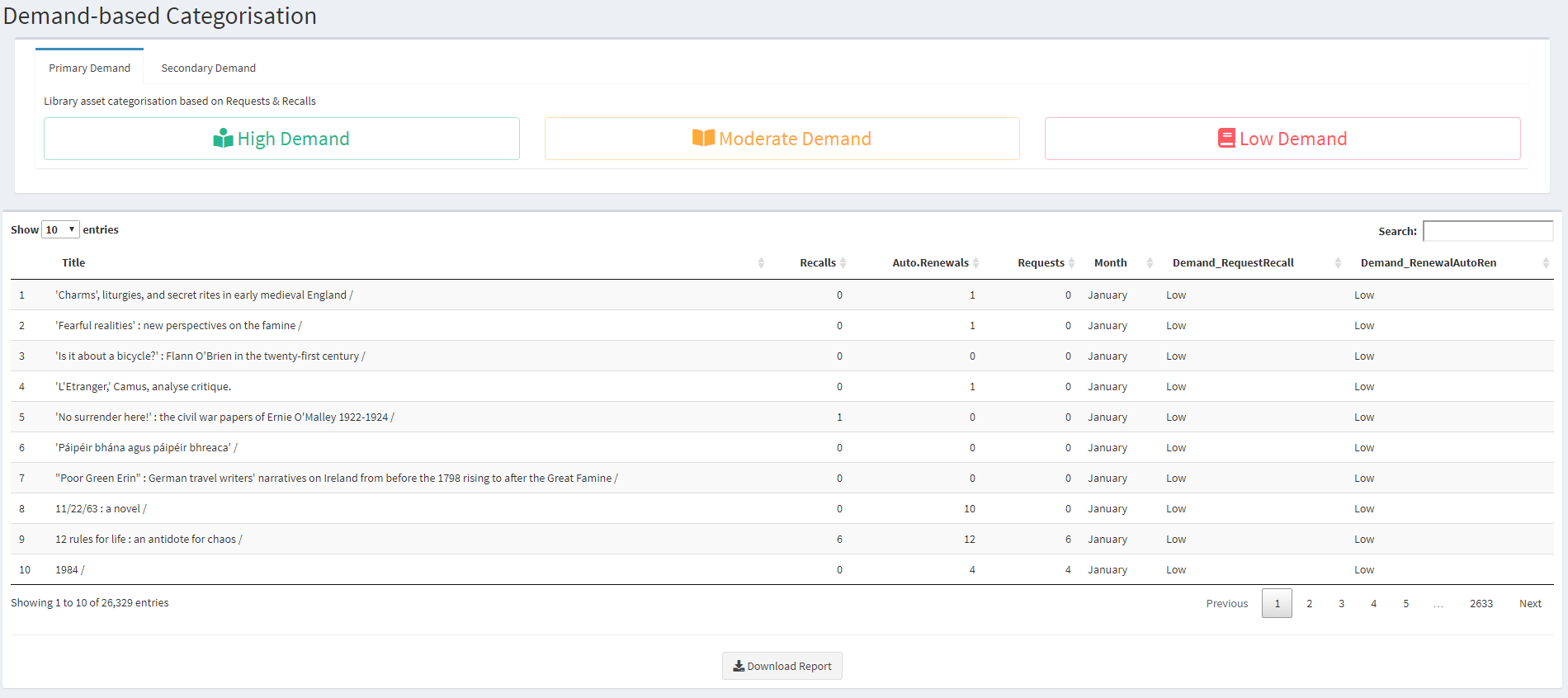


Figure 27: Screenshot of Tool: Demand Categorisation

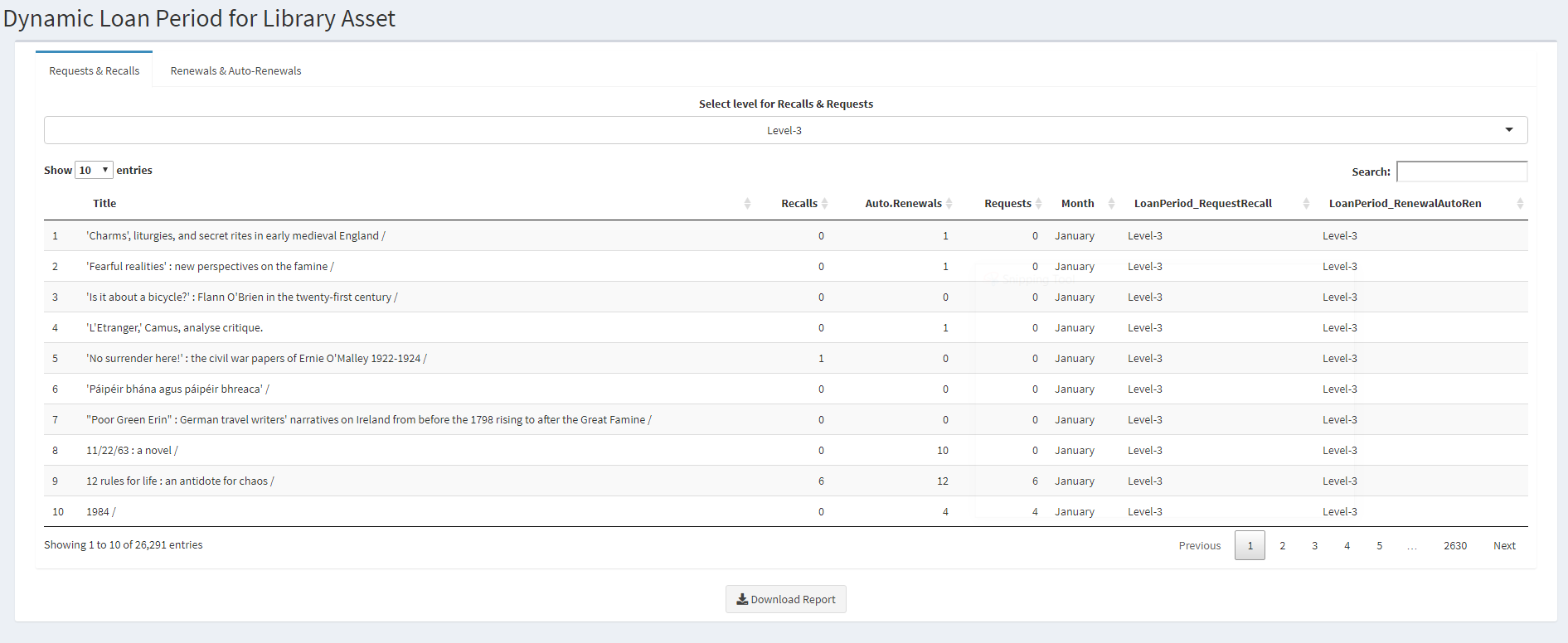


Figure 28: Screenshot of Tool: Dynamic Loan Periods

# Conclusion

This research project integrates a comprehensive view of the world of business with a detailed analytical perspective. Recognising the gap in the “James Hardiman Library” to facilitate analytics lead to the formulation of an analytical platform devised to reduce the latencies of workflow for NUIG library stakeholders. Working on the data provided by the library department posed a challenge from data manipulations to the implementation of clustering and classification algorithms. The basic idea behind the approaches that were intended during this project was to reduce the complexities and provide an accessible and systematic platform to the employees who oversee various activities of the library. Numerous technologies were employed to initiate the analysis process i.e. R programming, SPSS, R shiny, and MS excel. The classification of books was executed based on primary and secondary demands categories which further classified as “High”, “Moderate” and “Low”. Demand is categorized based on the feature selection of (“Recalls”, “Requests”, “Renewals”, “Auto-Renewals”). Another hallmark that this analysis provides is the remodelling of the library's dynamic loan periods. This characteristic is associated with the demand categorisation and administers a dynamic approach to the user in terms of days the books can be loaned.

Overall, the analytical platform will provide a dynamic and modern approach to the stakeholders in the management of laborious tasks of the library. Library management will shift towards a swift and robust approach by implementing this analytical platform. Incorporating this platform meets the specifications of our clients due to elements that enhance shelving efficiency and overall library performance. Stakeholders can now improve categorisation of the library's assets based on demand levels and can coordinate the availability of books as per the suggestions of the dynamic loan period. The objective of this operation allows ease of entanglement in the current system. The platform is convenient and user-friendly and does not require a technical background to implement. Furthermore, the user-guide will be available to aid in the understanding of the operatives of the application to ensure best practice and avoid confusion. Ultimately, the stakeholders can avail of this platform whenever they find it relevant to do so as it will be installed and monitored on their systems. To conclude, this research project can gather insights and create recommendations for the James Hardiman library fulfilling the requirements of our stakeholders.

# Appendix



# References

* Burton, M., Lyon, L., Erdmann, C. and Tijerina, B. (2020). *Shifting to Data Savvy: The Future of Data Science In Libraries*. [ebook] Available at: http://d-scholarship.pitt.edu/33891/1/Shifting%20to%20Data%20Savvy.pdf [Accessed 20 Feb. 2020].
* A concept data science framework for libraries. (2018). [ebook] Available at: https://www.ifla.org/files/assets/big-data/publications/a\_concept\_data\_science\_framework\_for\_libraries.pdf [Accessed 20 Feb. 2020].
* Liu, S. and Shen, X. (2018). Library management and innovation in the Big Data Era. *Library Hi Tech*, 36(3), pp.374-377.
* Mukhopadhyay, B. and Mukhopadhyay, S. (n.d.). *Application of Data Mining Techniques for Library Management Information System*. [ebook] Available at: https://ir.inflibnet.ac.in/bitstream/1944/226/1/cali\_57.pdf [Accessed 23 Feb. 2020].
* Sharma, P. (2019). The Most Comprehensive Guide to K-Means Clustering You’ll Ever Need. Available: https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/. Last accessed 23rd Feb 2020.
* Kabacoff, R. (2017). Graphics with ggplot2. Available: https://www.statmethods.net/advgraphs/ggplot2.html. Last accessed 23rd Feb 2020.
* Thorne. (2019). Introduction to Principal Components and FactorAnalysis. Available: ftp://statgen.ncsu.edu/pub/thorne/molevoclass/AtchleyOct19.pdf. Last accessed 23rd Feb 2020.
* Ex Libris. (2020). Alma Overview. Available: https://knowledge.exlibrisgroup.com/Alma/Product\_Documentation/010Alma\_Online\_Help\_(English)/010Getting\_Started/010Alma\_Introduction/010Alma\_Overview. Last accessed 23rd Feb 2020.
* Guru99. (2020). Introduction to Microsoft Excel 101. Available: https://www.guru99.com/introduction-to-microsoft-excel.html. Last accessed 23rd Feb 2020.
* R. (2020). R: What is R?. Available: https://www.r-project.org/about.html. Last accessed 23rd Feb 2020.
* Piatetsky, G., 2014. CRISP-DM, Still The Top Methodology For Analytics, Data Mining, Or Data Science Projects - Kdnuggets. [online] KDnuggets. Available at: <https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html> [Accessed 14 April 2020].
* Knowledge Center, IBM, 2012. CRISP-DM Help Overview. [online] Ibm.com. Available at: <https://www.ibm.com/support/knowledgecenter/SS3RA7\_15.0.0/com.ibm.spss.crispdm.help/crisp\_overview.htm> [Accessed 14 April 2020].
* StackOverflow. 2013. Why Do We Use K-Means Instead Of Other Algorithms?. [online] Available at: <https://stats.stackexchange.com/questions/58855/why-do-we-use-k-means-instead-of-other-algorithms> [Accessed 14 April 2020].
* van Rijmenam, D., 2013. A Short History of Big Data. [Blog] Datafloq, Available at: <https://datafloq.com/read/big-data-history/239> [Accessed 14 April 2020].
* Tay, A., 2016. 5 reasons why library analytics is on the rise. [Blog] Musings about librarianship, Available at: <http://musingsaboutlibrarianship.blogspot.com/2016/11/5-reasons-why-library-analytics-is-on.html> [Accessed 14 April 2020].
* DATAVERSITY. 2020. A Brief History Of Analytics - DATAVERSITY. [online] Available at: <https://www.dataversity.net/brief-history-analytics/> [Accessed 12 April 2020].
* Smart Vision - Europe. 2020. What Is The CRISP-DM Methodology?. [online] Available at: <https://www.sv-europe.com/crisp-dm-methodology/> [Accessed 12 April 2020].
* Bhatia, R., 2020. Is K-Means Clustering Really The Best Unsupervised Learning Algorithm?. [online] Analytics India Magazine. Available at: <https://analyticsindiamag.com/is-k-means-clustering-really-the-best-unsupervised-learning-algorithm/> [Accessed 12 April 2020].
* Tutorialspoint.com. 2020. Data Mining - Cluster Analysis - Tutorialspoint. [online] Available at: <https://www.tutorialspoint.com/data\_mining/dm\_cluster\_analysis.htm> [Accessed 12 April 2020
* Paulson, V., 2018. K – Means Clustering Algorithm. [Blog] Step Up Analytics, Available at: <https://stepupanalytics.com/k-means-clustering-algorithm/> [Accessed 14 April 2020].
* Shapiro, D. (2018). Elbow Clustering for Artificial Intelligence. Available: https://towardsdatascience.com/elbow-clustering-for-artificial-intelligence-be9c641d9cf8. Last accessed 17th Apr 2020.
* David, C & Jacobs, D. (2014). Principal Component Analysis: A Method for Determining the Essential Dynamics of Proteins. Methods in Molecular Biology. 1084 (n.k), p193-226.
* StatisticsSolutions. (n.d). Correlation (Pearson, Kendall, Spearman). Available: https://www.statisticssolutions.com/correlation-pearson-kendall-spearman/. Last accessed 22nd May 2020.
* Chakure, Afroz. (2019). Decision Tree Classification. Available: https://towardsdatascience.com/decision-tree-classification-de64fc4d5aac. Last accessed 22nd May 2020.