**Demographic data and it’s use in a succession planning process - the role of data analytics.**

Sinéad M Duffy

A Thesis Submitted in Partial Fulfilment

of the requirements for the

Degree of

Master of Science in Data Analytics

Diagram

Description automatically generated with medium confidence

September 2023

Supervisor: Marina Soledad Iantorno

# CCT College Dublin

Assessment Cover Page

To be provided separately as a word doc for students to include with every submission.

|  |  |
| --- | --- |
| **Module Title:** | Capstone Project |
| **Assessment Title:** | Demographic data and it’s use in a succession planning process - the role of data analytics. |
| **Supervisor Name:** | Marina Soledad Iantorno |
| **Student Full Name:** | Sinéad Duffy |
| **Student Number:** | SBA22229 |
| **Assessment Due Date:** | Friday, 22 September 2023 |
| **Date of Submission:** | TBC |

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

**Acknowledgement and Dedication**

# Table of Contents

# Table of Figures

# Table of Tables

# Abstract

# Chapter 1: Introduction

# Chapter 2: Research Design

## Primary Data

## Problem Identification and Clarification

**Thesis Title** - Employee demographic data and it’s use in a succession planning process - the role of data analytics.

## Research Objectives

**Objective 1** - Does employee tenure have an impact on a succession planning role

**Objective 2** - Can previous level of education have an impact on succession

**Objective 3** - Are there other demographic features which have an impact such as age or gender?

## Validity Type

## Ethical Considerations

# Chapter 3: Literature Review

**Introduction**

People analytics as defined by Ferrar et al (2021) is ‘*The analysis of employee and workforce data to reveal insights and provide recommendations to improve business outcomes*’. Numerous authors outline the importance of using data analytics to empower business decisions within the Human Resources Function (Ferrar et al. 2021, Mattox et all 2020, Rasmussen and Ulrich 2015). Rasmussen and Ulrich (2015) however point out the need to ask the ‘*right question*’ when reviewing data generated by HR and propose that this question should be incorporated into the end-to-end analytics process to identify and confirm the impact of people on decisions.

Learning analytics on the other hand, focuses on the effectiveness of a learner’s experience and is routed in basic training evaluation models such as the Four Levels of Evaluation model developed by Don Kirkpatrick (Mattox et al 2020). Specifically in this research paper, the author will focus on training provided solely within a corporate structure. Using a Learning Management Systems (LMS) has provided an effective way of gathering, analysing and reporting on learning related data (Sin and Muthu, 2015, Arka et al 2022, Mattox et al 2020). LMS’s such as Moodle have long been used in academic circles and have provided rich data sources in understanding how students learn and interact with systems (Sin and Muthu, 2015, Arka et al 2022).

This paper is an attempt to identify if a link or relationship can be found between training undertaken by employees and area’s such as succession planning within a manufacturing environment. Initial investigations into academic literature on the use of HR and learning data uncovered different themes which will be outlined in the following sections.

**Themes of the Literature Review.**

The Literature Review of academic and related papers helped to uncover several themes with the opportunity for further analysis of data held within the HR Department focusing specifically on data relating to learners. How the analysis should be conducted was, as expected, discussed at length with different approaches being taken. Four main themes that were identified by the author and have been outlined in more detail in the following sections.

**Opportunities for use of Human Resource Data**

HR data provides a lot of opportunity for analysis within companies (Rasmussen and Ulrich, 2015; Mattox II, Parskey and Hall, 2020; Ferrar and Green, 2021). Mattox et al (2020) in their book ‘*Learning Analytics*’ outline the pressure from business leaders to provide better and more insightful information in a timely manner. The demand for information is coming not just from Senior Managers, but also from stakeholder who want to know more about the people function and how effective it is (Topno, 2012; Mattox II, Parskey and Hall, 2020; Ferrar and Green, 2021). David Ulrich outlines how people analytics can add value to companies by allowing teams to make informed decision led with data in support of the business (Ferrar and Green, 2021).

To balance out this desire, HR data is uniquely different from other types of data used for analysis (Tambe, Cappelli and Yakubovich, 2019; Bankins, 2021). By its nature, data gathered by HR is formed of generally small datasets where events that companies want to model or predict are infrequent and nonstandard (for example dismissal of employees) or the data is subject to interpretation such as performance management where employees with different roles and responsibilities cannot easily be compared (Chadwick and Dabu, 2009; Tambe, Cappelli and Yakubovich, 2019; Bankins, 2021). Another issue with HR data relates to external requirements on the company which are not evident in other functions. For example, the recruitment process is influenced by internal factors such as the company’s own recruitment goals, as well as external ones such as the statutory landscape (Tambe, Cappelli and Yakubovich, 2019). This fact forces companies to limit the use of historical data such as recruitment data as it’s use could make incorrect predictions based on outdated information, or based on practices that are no longer the same (Rasmussen and Ulrich, 2015; Tambe, Cappelli and Yakubovich, 2019). As Bhardwaj et al (2019) stated ‘*Human resource analytics is an area of study that uses the mix of art and science on human capital in order to get measurable return on investment*’, (Bhardwaj and Patnaik, 2019).

That being said, HR needs to prove its importance to the business, especially in terms of how impactful it’s action are on the overall financial health of the company (Dong, 2022; Losey, Meisinger and Ulrich, 2005), (pp 121). In monetary terms analysis has shown that small changes to processes can make cost savings for the business such as implementing training reminders to cut down on the amount of time to complete induction, or to uncover a link between engagement data and business performance (Ferrar and Green, 2021) (pp. 4). There is an opportunity for data analytics within HR, whilst also recognising the need for help from HR experts to interpret the results of any analysis (Edwards and Edwards, 2019) - (pp. 5). In truth HR need to refocus their role to become a ‘strategic partner’ of the business helping it to achieve its strategic goals (Bhardwaj and Patnaik, 2019; Dahlbom *et al.*, 2020; Losey, Meisinger and Ulrich, 2005) (pp 150). Academics are aligned on the need for HR to upskill and become ‘ambassadors’ for data analytics as a means of driving data driven decision making (Martin, 2019).

One could argue that the future for HR data is to become integrated into the wider information stream of the company as a method to identifying how individual’s performance affects the wider company performance (Rasmussen and Ulrich, 2015; Tambe, Cappelli and Yakubovich, 2019). Rasmussen et al (2015) outline that impactful HR analytics are about linking to strategic business operations rather than trying to identify patterns in big data (Rasmussen and Ulrich, 2015). Some academic’s espouse the opinion that to be used successfully, HR data must be taken away from the HR department for analysis (Rasmussen and Ulrich, 2015; Ferrar and Green, 2021). Experience in one case-study outlined by Ferrar et al (pp 20 - 26), recofirms that HR data is different to other types of data and to successfully analyse it HR must be included in system development (Ferrar and Green, 2021).

Use of machine learning or artificial intelligence to supplement the HR decision making process is another growing theme within the data analysis raises ethical issues and questions which should be considered as part of this research. Employee’s perception on the use of artificial intelligence

Bankins et al (2021) has proposed a framework to help with the ethical implementation of artificial intelligence within an organisation (Bankins, 2021).

Focusing on data gathered as part of the learning process and how such analysis might be completed is discussed in the following sections.

**Learning Analytics**

A range of methods is used to create a dataset, and this is particularly true for online or distance learning (Sin and Muthu, 2015, Arka et al 2022). Systems such as Moodle allow analysts to follow a student’s learning path through a module or full course (Sin and Muthu, 2015, Arka et al 2022). Shen and Chi (2016) analysed how different levels of learners reacted to different methods of learning using such online interactions. In practice companies use systems such as LMS’ to collate learning data from employee interactions. An LMS (a Learning Management System) is a system that allows companies to manage training within the company, which then allows companies to run reports, track training requirements, assign learnings etc (*The LMS Guidebook : Learning Management Systems Demystified*, 2018)(Chapter 1). The advantages of using such a system is advanced features such as dashboards and reports created displaying high-level overviews of the data contained within the LMS as well as the ability to interlink with existing systems within the HR department (*The LMS Guidebook : Learning Management Systems Demystified*, 2018). The amount of data incorporated into an LMS means that large datasets can potentially be extracted, and it may be necessary to use data mining techniques to focus such big data sources (Sin and Muthu, 2015, Arka et al 2022, Mattox et al 2021).

The advent of LMS systems has led to a culture of self-directed learning by employees within companies (*The LMS Guidebook : Learning Management Systems Demystified*, 2018). Self-directed learning is where the employee is in charge of their own learning journey, a method of learning that is gaining traction in recent times (Araka *et al.*, 2022; Mustafa Yağcı, 2022). The drive to this new method of learning is coming from both companies as they roll out new technologies and employees themselves as they become more data savvy (Mattox II, Parskey and Hall, 2020; Araka *et al.*, 2022). The drive towards digitisation has only increased since the onset of Covid-19 and the need for companies and employees to adapt to increasing digital offerings (Almeida, Duarte Santos and Augusto Monteiro, 2020). Kokoc et al (2021) present the theory that by giving learners (employees) access to a dashboard to support their individual learning journey they will have more motivation to develop based on consistent feedback on their progress (Kokoç and Altun, 2021).

The development of this new area of learning has given more scope to allow machine learning to analyse the resulting data to help predict different outcomes - especially within educational settings (Araka *et al.*, 2022; Mustafa Yağcı, 2022). Analysis completed by academics chart learner performance against system access, and compare the results to final exam results (Araka *et al.*, 2022; Mustafa Yağcı, 2022). In companies, a different but similar approach is needed to gauge employee progress. For clarity, learning analytics has many definitions, but the one used in this paper is that learning analytics is the method of collecting, analysing, interpreting and reporting data to inform and understand learning methods and environments with a view to making improvements (Mattox II, Parskey and Hall, 2020; Kokoç and Altun, 2021; Araka *et al.*, 2022; Mustafa Yağcı, 2022). Educational data mining has emerged as a new field in which to access learning data stored in data warehouses or data lakes and seeks to work to open learning data to new analysis methods (*Learning Analytics – A Growing Field and Community Engagement*, 2015; Araka *et al.*, 2022; Mustafa Yağcı, 2022).

Deloitte in their 2017 Global Human Capital Trends outline that HR leaders, and specifically Learning & Development (L&D) leaders should reassess how they think about employees learning journey and ‘inspire’ employees to develop deeper skills with a view to enabling employees to change positions within their respective companies (‘2017 Deloitte Global Human Capital Trends’, 2017) (pp 36). The Deloitte report goes on to outline a case study about AT&T where they focus on career development for their employees and encourage them to change roles every four years as part of employees ongoing development (pp 36). Numerous sources outline reasons that employee should ideally be seeking new experiences every three to five years such as keeping in touch with outside trends, that employees become comfortable with change as some of the key items (Ryan, 2016; Christian, 2022).

As outlined in the research objectives above, the succession planning process is critical to the business’ ability to develop its employees. Huselid et al (2005) agree and outline that it is better to identify roles that are critical for the business and then spend time investing in the development of employees going into those roles to ensure that the right people are in place to drive the business forward (Huselid, Beatty and Becker, 2005).

**Causal Discovery**

Causal Discovery is an area of analysis that has been growing steadily in the last number of years and numerous authors have made a study of using causal algorithms to help identify and infer relationships within data (Eberhardt, 2017; Malinsky and Danks, 2018). Eberhardt (2017) outlines in his article ‘Introduction to the foundations of causal discovery’ that these so called ‘causal relations’ are thought-provoking because of how they can be used to illustrate how a system or process will react if an intervention is put in place (Eberhardt, 2017). Eberhardt (2017) goes on to outline that when searching for a definition of causal discovery it is made up of three distinct elements - statistical inference (inference from data to the distribution ), causal discovery (inference of finding about the possible causal structure, given statistical quantities) and finally causal inference (is the deciding on the causal effects given the causal structure and associated quantities), (Spirtes and Zhang, 2016; Eberhardt, 2017).

Causal relationships and structures can be displayed with the use of graphs such as the Directed Acyclic Graphs (DAG) (Vowels, Cihan Camgoz and Bowden, 2023). An example of a DAG graph used by Vowels et al (2023) is displayed in Figure 1 below. The figure on the left-hand side indicates that B has an impact on both A and C, and that A also has an impact on A. Applying the CMC theory to the figure on the right-hand side, the external values of Ua, Ub and Uc are found to all have a causal relationship.

A picture containing sketch, diagram, drawing, circle

Description automatically generated

Figure 1- Transitioning from a typical DAG representation (left) to a structural equation model (right). Grey vertices are unobserved/latent random variables. Source - Vowels et al, 2023

As can be observed, causal discovery is routed in statistics and relies heavily upon the Causal Markov Condition with refinement such as adapting to include noise (Eberhardt, 2017), the concept of faithfulness (Malinsky and Danks, 2018) and fairness in the data (Loftus *et al.*, 2018) to name but a few. Malinsky et al (2017) define the Causal Markov Condition (CMC) as being ‘every variable X in V (the set of variables in the causal graph) is independent of its non-effects conditional on its direct causes. Malinsky et al (2017) goes on to explain Faithfulness as being ‘the only independencies among the variables in V are those entailed by the CMC’ (Malinsky and Danks, 2018).

A point raised in articles reviewed for this case study is that ‘causation is not correlation’, which is very true (Xiao *et al.*, 2022). However researchers have put forward the idea that there are links between different variables (dependent and independent), and the study of causation allows for the identification of such relationships as well as the opportunity to estimate the size and / or magnitude of the relationship (Vowels, Cihan Camgoz and Bowden, 2023).

Tambe et al (2019) put forward the argument that the HR Department generally contains small datasets which may not be suitable to use to clearly identify relationships with the dataset, but applying causal discovery to the data allows the data analyst to infer relationships (Tambe, Cappelli and Yakubovich, 2019). This is further complicated by any decision to use historical data for analysis, the danger with is that historical HR data my unwittingly contain a bias towards non-traditional employees within a workplace such as a bias towards men against women where historical data is largely collected on male employees as women were underrepresented at the time period (Loftus *et al.*, 2018; Tambe, Cappelli and Yakubovich, 2019; Vowels, Cihan Camgoz and Bowden, 2023). Therefore to limit potential bias when analysing HR data, whilst working with small datasets it is necessary to use other sources such as theory and prior research as a guide to identifying causal relations (Tambe, Cappelli and Yakubovich, 2019).

Tambe et al (2019) outline the benefits of using causal discovering and reasoning as a method for working with HR datasets, such allowing analysts to focus on the characteristics and behaviours of the variables in the dataset, decreasing the cost of data management as well as allowing users to articulate and display the relations between variables also (Tambe, Cappelli and Yakubovich, 2019). A natural disadvantage to the use of causal discover and reasoning is that the results are open to interpretation depending on how the data is read, they also are not strong on being able to predict outcomes to queries (Tambe, Cappelli and Yakubovich, 2019).

It is important to note that in some fields it is not ethical to seek causal relationships between variables, especially in scenarios where there may be ethical considerations (Eberhardt, 2017; Malinsky and Danks, 2018; Vowels, Cihan Camgoz and Bowden, 2023).

To facilitate causal discovery, several causal discovery algorithms have been developed to make the search for such casual relationships easier. This is the next theme that the author will review, including what kind of analysis is typically used in conjunction with causal discovery is also outlined.

Casual Search Algorithms

Causal search algorithms, as defined by Malinsky and Danks (2017) are used to investigate hypothesis-based relationships between variables for example A and B based on the context of event C. They go on to explain that causal search algorithms are exactly the same as other better known analysis methods such as regression, the main difference being that assumption that results can be expressed as a causal graph such as a DAG (Malinsky and Danks, 2018). The benefit of using causal search algorithms is they help explain questions like ‘what makes a person intelligent’ by displaying all possible connections between subjects such as math, logic and writing test scores, leaving the analyst to select the best model based on the data and their own experience (Malinsky and Danks, 2018).

As expected, when applying causal search algorithms, there are a myriad of implementations possible, and some of these are detailed in the following sections. However, it would be helpful to understand the data preparation process that is recommended by Malinsky et al to help specifically with causal search algorithms. The first step is to assume that variables are ‘semantically independent’, i.e that they are capable of being manipulated independently. Therefore, it is important to remove any unnecessary or redundant variables before beginning analysis. The second step / assumption is that variable is continuous or categorical in nature, therefore any mixed datasets should be ‘cleaned’ and with the recommendation that values be placed within a scale to minimise any potential bias that could occur in the results. The third step is ensuring that any proxy or estimated values are as accurate as possible and of a single unobserved causal factor. The fourth step is to consider the timeframe that the data collected represents - is it month by month / day by day etc as the analyst must be able to outline if the measurement are for the same individual or for different ones over time. Finally, the fifth step is for the analyst to consider their own knowledge of potential causal relationships (Malinsky and Danks, 2018).

Malinsky and Danks, (2018) also outline the different types of search algorithms that may be applied to data including:

* Constraint based algorithms which display connections between the causal graphs and independencies that are found in the data.
* Score base algorithms which compares the models based on some measure of model fit such as the Bayesian Information Criterion.
* Causal search algorithms with semi-parametric assumptions which allows the model to use the assumptions to display the relationships in more detail.
* Clustering causal analogues which cluster variables as opposed to individuals. This helps to show or imply patterns of ‘observed correlation.

Example so applying such causal search algorithms are outlined in the beginning with Mӓkelӓ et al who applied a constraint based algorithms to their paper on Earth system sciences (Mäkelä *et al.*, 2022). The authors outline that knowledge experts can help beyond initial inputs and quality check the results by helping to interpret the results based on their own prior experience.

Qui et al (2019) used fuzzy cognitive maps to display causal relationships in emergency cases as an effective way of modelling knowledge representation (Qiu, Gu and Wang, 2019).

Dong (2022) proposes an integrated approach to analysis of HR data including use of neural networks, artificial intelligence etc in a proposal for an integrated system called the Human Resource Intelligence System (IHRMS) (Dong, 2022)

Assaad and Devijer (2022) applied causal discovery to time series data sharing a number of different approaches to this type of analysis. They outlined a number of issues applying theory to practice and warn that causal discovery and time series is still an active area for research (Assaad, Devijver and Gaussier, 2022).

Peters et al (2014) suggest using structural equation models rather than graphs as being helpful in using noise in data to highlight causal relationships between variables (Peters *et al.*, no date)

Finally, Kalainatha et al (2020) propose a method for using causal discovery within Python called the Casual Discovery toolbox (CDT) (Kalainathan, Goudet and Dutta, no date). CDT is an open-source library that allows users to first graph dependencies and then apply causal discover algorithms.

In respect of training data, different techniques which were used by authors to complete the analysis are outlined in the articles reviewed. Some of these techniques are outlined below:

* clustering / classification techniques (Araka *et al.*, 2022; Mustafa Yağcı, 2022)
* data mining based on time data (Araka *et al.*, 2022)
* Random forests / Support Vectors / Naïve Bayes techniques (Mustafa Yağcı, 2022)

As is expected, several alternative means of analysis is possible, and the author will need to experiment to identify the method and / or algorithm that will give the best results for the individual dataset.

Conclusion

Reviewing work by other authors is a key step in completing any analysis. In this instance, the review was able to highlight the need for a more analytical mindset when considering HR data or learning data and outlined the benefits of same in terms of closer strategic connections to the business. Opportunities exist for such data to be used to allow better decision making by business leaders, but the HR department must enable this by being ‘ambassadors’ for their data. Systems such as an LMS should be used to leverage learning analysis within companies, as well as directing more effective and strategic learning outcomes-based on sound data analysis.

Analysis of LMS data can be completed using several causal machine learning algorithms to produce causal models that best suit the data at hand.

# Chapter 4: Methodology

As a rule, research methodologies form two distinct groups namely primary and secondary data (Saunders, Lewis and Thornhill, 2012). Primary data is collected specifically in order to answer the research questions posed at the start of this paper, secondary data is collected for some other use and is then applied to this data. As can be expected, within both areas there are different types of collection methods that can be applied. The purpose of this section is to outline the collection strategy or strategies that work best in the author’s opinion to answer the research questions posed in Section 1 above.

**Primary Research Data Collection**

The author identified that a quantitative approach to collecting primary research data would be the most appropriate method of gaining first person information from people well versed in the succession planning process. Such data collection would be completed using in person interviews. Interviews were to be completed online using the Microsoft Teams platform which enabled transcription to be completed automatically. Using this method of data collection allowed for in depth discussion on the area of succession planning. As well allowing a level of observation to be used by the Researcher to gauge reactions to questions posed, as well as allow the flow of the interview to be altered depending on the interviewee’s reaction to the question (Saunders et al, 2012). To overcome any potential bias that occurs on the interviewee’s behalf, the author has sought input from different individuals who have taken part in the succession planning process. The interviewees will all provide a different point of view. One interviewee manages the succession planning process for the manufacturing division of the company, another manages the process for office-based employees located around the wider European Economic Area. Yet another interviewee has experience in both areas of the business outlined previously and is well placed to share similarities and differences in the two processes. Carrying out a diverse range of interviews allows the author to construct a holistic view of the succession planning process currently used.

The author has opted to using unstructured / in-depth interviews which are informal in nature with no prepared questions rather there is a general area for discussion (Interview Techniques for UX Practitioners: A User-Centered Design Method, 2013), Saunders et al 2012). This allows the author and interviewee to have a conversation that moves organically through topics giving an opportunity to probe for further understanding where necessary.

Issues considered by the author when selecting this method of primary research including:

* Research Objectives - the objectives stated previously lean on utilising HR Data, specifically training data to enhance decision-making. Unstructured interviews with experts within the HR Community will enable allow for open discussion on the themes of this research as outlined previously in the literature review.
* Expertise of the Selected Participants - expertise of the HR Community is important to this research. As such, it is important to canvas a range of opinions to gain a holistic understanding of how training data can be utilised for data driven decision making in the succession planning process.
* Timeframe - it is important to recognise that the timeframe with which to complete the research project is limited to a maximum of 10 weeks. As such, a limited number of interviews could be carried out during the timeframe.
* Ethical Considerations - in the world of the General Data Protection Regulations (GDPR) gaining consent of data subjects is very important. This is no different when considering ethical considerations for a research project. As standard, all interviewees were informed of the confidentiality of the process, as well as highlighting their option to withdraw their consent to have their interview notes included as part of the finished research paper. Should a participant decide to withdraw their consent, the author will put measures in place to delete all contributions from the project.
* Bias - the author herself is an experienced HR practitioner and as such already has a view regarding the use of HR data to drive decision making. Therefore, the author is aware of the need to be impartial in asking questions of participants as well as noting their responses.

The availability of experts was an important consideration when deciding on a primary research method. Working with HR data can be a nuanced process and the author understand the importance of garnering a range of views so a holistic view can be considered of the succession planning process. Using the framework outlined above, the following points were considered:

• Research Objectives - would the subject matter of the research objectives lend themselves to facilitating participant interviews of 30 minutes or longer? Were themes outlined in the Literature Review of relevance for the HR Community as they stood?

• Expertise of the Selected Participants - again, a key point in considering who would be a suitable participant for the interview process. In the end, the selected participants were employees with the company who play a key role in integrating data from different systems to support the succession planning process. Other employees had integral knowledge of the company’s LMS - where if successful will be the source of future evolutions of any successful algorithm.

• Timeframe -having consideration to the timeframe of the data, the author chose to largely nominate experts within the company in which she works. This will encourage a greater response rate as the ‘collegiality’ of the request.

• Ethical Considerations - as outlined in the previous section, all participants will be asked for their consent prior to the interviews taking place and will be reminded of their ability to withdraw their consent at any time up to publication of this research paper.

• Bias - using inputs from one company will naturally lead to a degree of bias in the process. However, the author has selected those experts who are involved in either improving or implementing analytical systems, or who work in diverse geographies within the company. The author is confident that this will give the interview responses a rounded view, keeping bias (unconscious or otherwise) to a minimum.

The next step in the methodology is in selecting the experts to take part in the interview process. To ensure that the right mix of experts are chosen, Saunders et al (2012) outline that identifying the characteristics of the experts prior to selection will create a more rounded group of experts. To that end, the author has identified two characteristics that would be helpful in answering the research objectives posed at the start of the document. The three characteristics are:

• Have some involvement in implementing / improving processes with HR and the wider company.

• Be ‘outward looking’ in that they are knowledgeable of company strategy as well as best practices within the market.

A panel of experts has been selected and based on the criteria and framework outlined previously. Transcripts from the interviews can be seen in appendix XXX.

**Secondary Research Data Collection**

Initially, the author hoped to use data from the company’s own LMS system. However, the author was unable to gain sufficient data to work with. As an alternative, the Open University Learning Analytics Dataset (OULAD) was selected as the dataset upon which analysis could be completed (Kuzilek, Hlosta and Zdrahal, 2017). In order to closely mimic data typically found in a commercial operation, the employee added a column to the ‘student\_info.csv’ file called ‘Tenure’ to represent the number of years the employee / student is working within the company. The values of the ‘Tenure’ column were randomly generated using ‘***randint***’ with values between 0 and 20. These values were also subsequently grouped together into ‘bands’ to identify if tenure had an impact when analysing categorical data.

**Research Methodology and Validity**

Considering the research methodology outlined above, it is possible to say that the most relevant components of validity relevant to this research are accuracy, currency, and bias. It is however also possible to say all components of validity apply to the proposed research, some component’s more than others. The concepts of accuracy and currency are explored below.

Accuracy in this instance relates to comprehensive the data statistically is. In terms of primary data, accuracy does not apply as the data is not statistically based. The data captured from interviews will need to be transcribed and included in the appendices of this report. Furthermore, the main points and sentiments expressed will be used to verify if the proposed model will be useful or not.

Currency in this instance is a potential barrier to the methodology of this research. The author has chosen to use simulated data extracted from an educational institute learning management system. The data was originally released in 2017 and contains data from 2013. It is true to say that the data is not current, however, it closely mimics the data is contained within the company’s own LMS. That being the case, the author has chosen to accept the risk to the validity of the results of this research paper.

Bias has already been identified as a possible threat to validity when conducting in-depth interviews for primary research. The author will attempt to limit bias by ensuring that there is a clear purpose of the interview which is communicated in advance. By working with known participants there is already a degree of trust established between the parties to facilitate a frank discussion. Finally, the author will create several prompts based on key research themes that will help guide the interview process and stay within the research area.

Although three components have been listed, it is not unreasonable to assert that other components may also become more apparent as this research progresses.

Research Methodology and Ethics

As with all research, there are ethical considerations that will need to be planned for, some of which have been outlined above.

5.5.2. In respect of primary data collection, participants will be asked to participate, and will have the option to withdraw their consent to participate or have their data included at any stage of the process up to the submission data. Interviews will be conducted.in a professional manner with responses being confidentially and anonymously recorded for the purpose of the research. The author will have a master file noting the responses, but this will not be share as standard. This Master file will be stored in a secure location and will be password protected as added security. Finally, if participants have questions at the end of the interview process, or before, the author will undertake to resolve these queries as quickly and sensitively as possible.

5.5.3. In respect of secondary data collection taken as a download from the company’s LMS, the author will put in place the relevant requests necessary within the company to obtain access to the data. Again, any data received will be anonymised immediately so no data can be related back to any individual. Furthermore, the data will be stored in a safe location with the relevant passwords in place.

5.5.4. The author is also conscious of the General Data Protection Regulation’s (GDPR) and will put the necessary steps in place to ensure compliance at all stages of the research.

6. Sampling Strategy

6.1. A key part of any research project is deciding if a sampling strategy is necessary, and if it is, what is the appropriate strategy upon which to complete the analysis on. Looking at the two data inputs, the author determined that a different strategy will be needed for primary and secondary data collection. For primary data, namely in-depth unstructured interviews as outlined above, a non-probability sampling method will need to be applied. For secondary data, a probability-based sampling method is most appropriate. Both approaches are outlined in the following sections.

6.2. Primary Data

As outlined in Section 5, the chosen primary research methodology selected is that of in-depth unstructured interviews. This approach allows for participants to be selected based on their expertise with a given area, a form of sampling known as purposive or judgement sampling (Saunders et al, 2012). Purposive sampling does not have any statistically relevance as the chosen sample bears no resemblance to the overall population of the study. Saunders et al (2012) outlines that Heterogeneous sampling is that where ‘chosen participants with sufficiently diverse characteristics provide the maximum variation possible in the data collected’, (pp. 287). Saunders et al (2012) also advise that a minimum sample size of 5 in-depth unstructured interviews be carried out.

6.3. Secondary Data

When reviewing the secondary data requirements, it was clear that a different strategy was needed. Specifically relating to this project, the total population is approximately 3,000 employees worldwide. Narrowing the scope of the study to just employees in Ireland reduces the scope to approximately 550 employees. Focusing on just employees employed in the operations function again the population narrows to approximately 300 employees. As outlined above training records are collated from when a new employee starts work. Training records have been collated since the LMS was introduced in 2020. In addition, the timeframe by which to complete the analysis is limited to several weeks. Working with full populations of more than 50 is not recommended (Saunders, Lewis and Thornhill, 2012). Based on these reasons, it will be necessary to use a sample strategy to complete the research.

6.4. The author reviewed characteristics of the two sampling techniques of probability and non-probability closely. Each case has an equal chance of being selected for inclusion in the sample. Therefore, the author has identified that using a probability-based sampling strategy would best fit the research task, objectives, and population within scope of the project.

6.5. To clearly articulate the reasons behind this choice of sampling strategy, the author has utilised work completed by Saunders et al (2012) as a guide. In their book ‘Research Methods for Business Students’ they outlined a number of steps to help provide clarification when choosing a sampling strategy, namely identification of sampling frame, sample size, sampling technique and a check to ensure that the sample is representative of the overall population (Saunders, Lewis and Thornhill, 2012). Using this framework, the author will discuss each element comprehensively in the following sections.

6.6. Sampling Frame

6.6.1. Saunders et al (2012) outlines that a sampling frame is ‘a complete list of all the cases in the population from which your sample will be drawn’, (pp 293)

6.6.2. As outlined in previous sections, the population for this research is contained within an LMS system, which contains data for all employees working for the company. Employee records are created as part of the new hire process, and a learning profile is created. As employees complete training tasks and are signed off, their learning record is updated. For employees not in a manufacturing role, assigned learnings are marked ‘complete’ once the learning is completed. Other training sites within the company are linked to the LMS so employee records are automatically updated with no manual interaction.

6.6.3. A further consideration would be that the data is stored within the LMS is for all employees working for the company both in Ireland and in other sites around the world. By filtering to the Irish site, it’s possible to say that employees contained there have an equal chance of being selected to take part in the study.

6.6.4. In addition to filtering the data to only Irish based employees, it will be necessary to consider the different working patterns of employees. Some employees work full time (100%), whist other employees work less than this on a range of flexible work patterns such as weekend or part time.

6.7. Sample Size

6.7.1. As sampling size will have an impact on any results, the research found considered what the appropriate sample size should be. The population of employees is approximately 300. Saunders et al (2012) outline that a sample size of 30 is a ‘rule of thumb’ in order to carry out statistical analysis on the group. This will allow a degree of confidence in the analysis and is based on best practice.

Using a sample size of 30 for this population would give a 10% sample size. The author is suggesting a sample of 60, which would give a sample size of 20%.

6.8. Sampling Techniques

6.8.1. The author reviewed the five sampling techniques common to probability sampling and identified that systematic sampling would be the most suitable. The technique is discussed in more detail below.

6.8.2. Saunders et al (2012) outline that the systematic sampling technique involves dividing the population based on sample based on a regular interval. The interval is calculated using a sampling fraction. The inputs to the sampling fraction are the size of the same and the total population. The benefit of using a sampling fraction is that is it relatively easy to implement and explain. However, it’s important to ensure that the population is not pre-sorted in any way prior to the sampling method being applied. This will ensure that the applied sampling method is truly random.

6.8.3. For this piece of research, the sampling faction calculation is worked out below.

Sampling Fraction = Actual sample size 60 = 1

Total population 300 5

6.8.4. Based on the workings above, every 5th record will be taken to form part of the sample for this research.

6.9. Check for Representation

6.9.1. The final check of the sample is to assess the possibility of bias. This can be done in multiple ways, with one method being to draw another sample from the same data set, using a different sampling fraction to compare the results.

6.10. Applying the Sampling Strategy

6.10.1. The sampling strategy outlined above is normally applied to primary data collection. For the purposes of this research, a sampling strategy will be applied to both primary and secondary data collection.

6.10.2. Secondary data collection for the purpose of this research will be taken from the LMS system in the form of a download. The data downloaded will conform to the sampling strategy outlined above.

6.11. Now that a sampling strategy and methodology have been developed, it is necessary to consider the ethical implications associated with the research, and how the author proposes to work with them.

# Chapter 5: Implementation

# Chapter 6: Results

# Chapter 7: Discussion

# Chapter 8: Conclusion

# Appendix A: Workflow

# Appendix B: Interview Transcripts

# Appendix C: Data Permissions

# Appendix D: Consent Forms

# Reference List