Group Project in Data Science (CSC8633): Group 10 Report

Group 10

2023-03-24

Abstract

This project is an investigation into the factors affecting the success of candidates at HC-One, a large-scale care home operator and recruiter for workers in care homes. It is important that HC-One discovers candidates that can deliver an appropriate quality of care. In addition, the application process contains a large amount of noise and a large volume of applications, so we aim to discover insights about the types of candidates that are more successful, to help HC-One prioritise candidates appropriately in future application cycles.

In order to understand which candidates are best for HC-One to approach first, we investigated the length of service of candidates, in terms of the amount of time spent at the company, and learned about the characteristics that correlate with this. Candidates that work at the company for a longer period are more efficient for the company to employ, as the company does not have to incur as much cost in the future to train and hire new employees. By learning about the factors which influence employees' retention at the company, HC-One can improve the efficiency of its hiring process in future.

We have conducted exploratory data analysis into a number of areas to investigate factors influencing candidates' retention rates. Employment history, academic background and the application channel for employees are factors that are commonly considered by employers, and we tested the significance of these aspects in practice. In addition, we looked at how the advertised salary for a position; the level of competition from other care homes in a candidates' catchment area; and the type of demographic, measured by the Acorn consumer classification, influenced the retention of candidates across the past six years of employment. Additionally, we considered how these factors differ in their impact on the three different jobs available in the data: nurses, carers and senior carers.

We discovered that educational background and employment history were both significant in influencing the performance of employees at HC-One. Also, the channel through which employees applied was important, with referrals having the longest tenures at the company. Advertised salary had a slight negative correlation with the sum of months of service, which could be caused by a number of extra factors. The Acorn classification, however, was not seen to have any significant association with retention rates. When considering the job types separately, advertised salary was seen to have a stronger negative correlation with carers, with nurses having almost no correlation. The level of competition in a candidate's local area showed us that the most-retained are those with medium levels of competition. Nurses were seen to be most dependent on having specific educational qualifications, typically a Diploma- or BSc-level qualification in a related subject.

Our analysis is unique to this field in considering the length of service as the key variable in determining successful candidates for the company. The data only contains applicants who were successful in the application process, so looking at the problem from a unique angle can provide new insights and enable HC-One to learn about areas that they may not have considered before. In addition, it enables us to order the candidates based on this variable, rather than simply splitting candidates into binary categories of success.

CYCLE 1

Business Understanding

Background

HC-One is the largest care home operator in Britain with over 275 homes across the UK with 19,000 employees. Due to the size of the organization, they receive about 2500-3000 applications per week. This poses a huge challenge for HC-One as the recruitment team has to go through thousands of applications and may miss out on hiring the best candidates for the job. Fortunately HC-One has collected applicants data over the last 6 years. The data consists of three different files, representing the candidates' schools, their employment history and candidates' application data & background information. The data of successful applicants for carers, nurses and senior carers positions will be used in this project as the ground truth for our analysis. HC-One would like a more efficient way of assessing candidates in an easier and more accurate way. This would mean that the process would be less time consuming for them and they would focus their hiring resources on candidates who will be successful in their role, whilst requiring fewer resources in future if candidates are retained successfully at a higher rate.

Business Objectives

To make the recruitment process for HC-One as efficient as possible, we aim to use the data provided to uncover insights that can be useful in improving the recruitment process.

To analyse how the different variables provided in the datasets have an impact on the success of a candidate, we came up with the following hypotheses. Does the channel of application affect or determine whether an applicant will be successful or not? How does a candidate's Acorn demographic Category affect the candidate's success in terms of retention in the company? Does the advertised salary have anything to do with whether a candidate is successful or not and does it affect their retention? Does the candidate's educational background determine whether their application will be successful or not? How does the number of competitors in the applicant's area of residence affect the time they serve at HC-One?

Situation Assessment

Requirements, assumptions, and constraints

This project requires that the development of the solution and appropriate reporting is done within two weeks. The data to be used for this project does not contain some personal information about the candidates and is to be used with discretion. As a result, the results of the analysis will not be available for public use, and will be heavily customised for the purposes of HC-One.

To get insights from the data we had to make assumptions on some of the variables. These assumptions include:

1) The sum of a candidates' months of service is equal to the total months of service at HC-One. Also we made the assumption that if a candidate left HC-One and came back later, this does not affect their total months of service.

2) Current employees who have less than 6 months of experience were removed from the dataset under the assumption that their data would not be informative as we don't know how long they will be at HC-One, and it would be difficult to predict this with a high accuracy. This removal is unlikely to bias the data heavily, as characteristics of successful employees are unlikely to have changed significantly over time.

3) Assuming that current employees who have worked for a period of 6 months will work for 12 months & 24 months as this matches closer to the observed retention rate for current employees (79% vs 70% for 12 months given 6 that they stayed for 6 months, and 57% vs 40% for 24 months given 6 months of work), so would be more accurate as a representation of the data than removing these observations.

We decided to work on this project using R's ProjectTemplate given the constraint on the privacy of the data involved. The data does not contain detailed information about the candidates which was a significant constraint in terms of drilling down to candidate specific characteristics that can be indicators into whether or not a candidate is likely to be successful. Something like age of the candidate would have been helpful in determining the validity of the years of experience provided or the education attained. Also, we were unable to use data on candidate's locations outside of the demographics of their local area. Another significant constraint was the two weeks time limit within which the project was to be delivered.

Determining data mining goals

The main goal is to clean the data to a state where we can analyse it to get insights into how we can improve the recruitment process for HC-One. The success will be measured by the number of actionable insights that we uncover that can help in making the recruitment process more efficient. Being able to discover that some characteristics do or do not have an association can assist HC-One in being able to refine their hiring process in future.

Project Plan

Initial Assessment of Tools and Techniques

Tools

For this project, the team agreed to use use R language for the analysis in the R studio. We will use libraries within R studio to aid in the data preprocessing and visualization of insights such as ggplot2, tidyverse, dplyr and many others. The project will be organized and managed using ProjectTemplate, which provides a clear framework for us to organise and collaborate on our work.

Techniques

CRISP-DM: The project life-cycle will be carried out using the CRISP-DM structure for data mining .This structure will help us in organizing tasks in phases that we can carry out sequentially to achieve our solution.

Data Understanding

The data for this analysis has been sourced from HC-One Ltd, containing data regarding successful applicants to the company over a period of six years. The data is stored in three separate spreadsheets. The first of these contains background data on each candidate, including information on the application, the candidate's job, their skills and qualifications, and background demographic information relating to the area that the applicant lives in. In addition, there are other spreadsheets containing more specific information on each candidate's educational background and employment history.

The three spreadsheets are easy to combine using each candidate's candidate ID number. This enables us to use the more detailed information on candidates' educational background and employment history in our analysis. However, these spreadsheets appear to have been filled from optional, free-entry fields in forms, meaning that there is a lot of missing data for a large portion of the dataset, and much of this data requires more thorough cleaning before being analysed. An issue encountered when loading the data for cleaning and analysis was password protection when accessing the spreadsheet. The data is fully anonymised, so individual candidates can only be identified by their candidate ID number. However, due to the nature of the information contained within this dataset, the data is password-protected in order to protect its confidentiality. To work around this issue, we decided to save a copy of the data locally and use this copy for our analysis, making sure that the data was not transferred to any location which would make it publicly available.

The spreadsheet containing candidate applications data contains 24483 observations on 129 variables. Candidates can be identified by the vacancy ID, vacancy application ID and candidate ID. We decided to include the vacancy ID as an index variable for different employees, treating the same employee in different roles as independent observations. From the employee's personal information, we decided to use the length of service for the company as our measure of "success" for the employee. As the primary objective of the analysis is to improve the hiring process for future applicants to the company, being able to predict how long an employee will stay with the company is very useful to determine whether the company will need to invest more resources in finding new employees. For example, an employee that stays with the company for six months takes half as much expense for the company to hire as two employees who are employed for three months each. We have decided to investigate the factors affecting an employee's probability to remain with the company for periods of six months, twelve months and twenty-four months. An important question when using this variable for analysis was how to model this for current employees, meaning those who are still active employees of HC-One. The difficulty arises as there is no way to tell for how long these employees will stay with the company. As a result, we decided to exclude employees from the dataset who were currently working with the company but had been doing so for a period of less than six months. This followed from our initial definition of six months as a successful retention period for an employee. For twelve- and twenty-four-month periods, however, we learnt that current employees are significantly more likely to stay for these lengths of time than past employees. A possible reason for this is that long-term employees are likely to be the ones still working with the company, as there is not a long period of time for employees to have served this length of time and left. As a result, we have decided to assume that current employees that have stayed for a period of at least six months will stay for twelve months and twenty-four months. Based on the current employees in the dataset, 79% that stayed for six months have also reached twenty-four months, and 57% of those that stayed for six months have reached twenty-four months. We believe that this assumption is likely to lead to a better projection of the true retention rates than dropping this data, which would introduce bias away from long-term employees. We generated additional binary variables to represent our definitions of "success", indicating whether an individual worked at the company for a period of at least six months, at least twelve months and at least twenty-four months. There are 13480 employees within our dataset who worked for HC-One for a period of at least six months, and 5811 who left the company before reaching six months. A total of 1800 employees from the dataset were still actively employed by HC-One, meaning that it is unlikely for there to be systematic bias introduced by dropping these observations. As there are numerous observations containing duplicate candidate IDs in the dataset, likely relating to individuals who have worked for multiple different periods at HC-One, we needed to create a new variable representing the total length of time served at the company by each individual.

We have also decided to include the date of application for each applicant, the number of hours they were contracted for (in order to differentiate between full-time and part-time workers), and their home code. An aspect of this data that we are interested in is exploring whether there is a difference in the types of candidates that are successful in the different jobs. For this dataset, we have three different job types: carer, senior carer and nurse. We have included the indicator variable for an individual's job to be able to partition this data for analysis and comparison.

In addition to these variables being required for all analysis, we selected several to test each of our hypotheses. To investigate how the channel that an individual applied through influences their longevity as an employee, we included both the type of channel that they applied through and the channel itself. Our hypothesis for this phase of the analysis is that applicants who are referred via employee referrals have a higher retention rate than applicants from other avenues. The Acorn demographic category is a system categorising the population of a catchment area based on socioeconomic demographic factors. It splits catchment areas into six primary groups, based on wealth-related factors, then into subtypes of the groups and sub-categories of these types. We want to examine whether the Acorn demographic profile is related to the retention of candidates. This data can help us to understand more about the living situation and background of successful applicants, and to produce insights for the company for the future based on this. Another aspect we want to investigate in this project is how the advertised salary for a job posting influences the longevity and success of the applicants who end up taking that job. This only requires the variable for advertised salary in addition to our choices for general characteristic variables. However, it is important that the advertised salary is taken in the context of the year in which the job application was advertised, such as being standardised according to the national minimum wage for that year. This variable required extra cleaning, as the different advertised salaries are phrased in many ways, and in some cases are imprecise, or simply give a range of salaries.

To look at an individual's educational background, we chose to analyse both vocational qualifications and academic qualifications, and to understand if there is a difference in retention between people with each of these types of qualifications. There are a total of 3486 individuals in this dataset with at least one academic qualification, and 5055 people with at least one vocational qualification. This means that a large portion of the dataset is not relevant when comparing which of these qualifications leads to higher longevity for employees at the company. Academic qualification information is missing for many individuals, so it is difficult to analyse these individuals for this hypothesis. There are additional issues when cleaning the data for the data stored in the spreadsheet relating to candidates' educational background, as there are many missing values, multiple entries for several candidates and alternative names for equivalent qualifications. As it appears to be a free-form entry for candidates to provide their education details, many also contain errors and inconsistencies which could make it difficult to process this data.

Our final part of exploratory analysis relates to how the number of competitors, meaning other care homes, are in the catchment area of candidates. This may have an impact on candidates' retention if it is easier or more attractive for them to move to an alternative care home and there is a higher quantity of alternative job opportunities for them. We can understand this by looking at the variables indicating the count of care homes within a 15-minute drive of the candidate's home and within a 500 metre walk, to understand how being able to commute by car and by foot respectively influence their decisions. In addition, we can look at the number of beds in care homes within these two distances as this may be a more accurate representation of the number of jobs available at nearby competitor care homes. Finally, the average age of the care homes within a 15-minute drive for the candidate can be a useful indicator to add to this analysis, as it represents the relative attractiveness of working at alternative care homes.

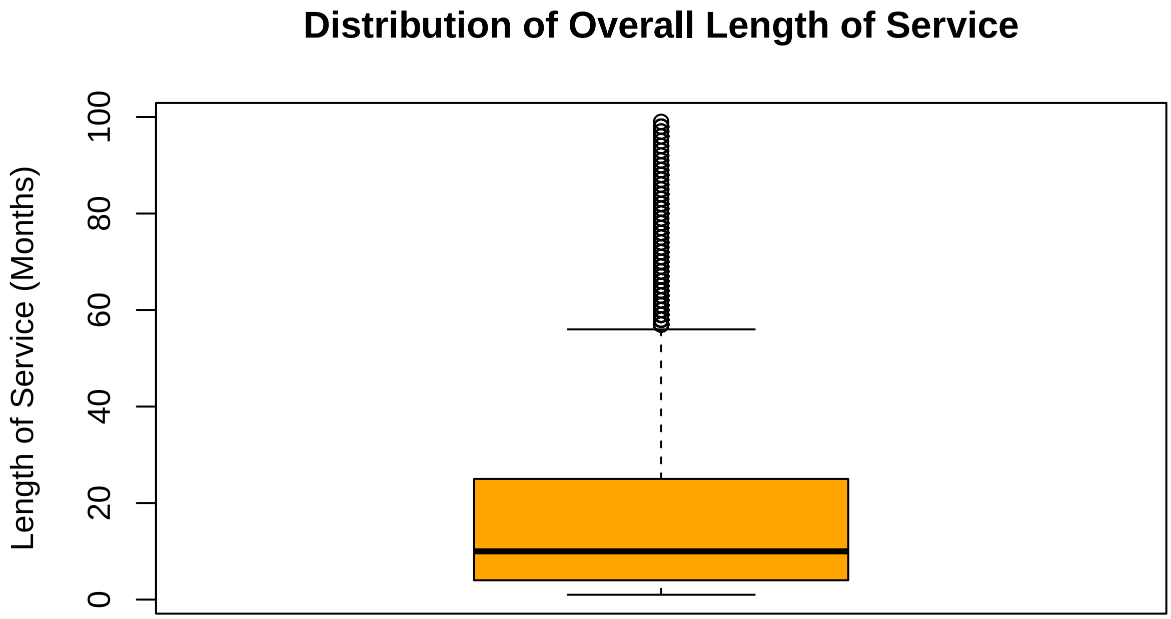


Figure 1: Boxplot showing the distribution of overall length of service of all employees in the dataset

Figure 1 shows the overall distribution of lengths of service for HC-One for all employees remaining in the dataset, with the exclusion of any outliers in the data (lengths of service < 72 months). These outliers have been removed in order to improve the interpretability of the plot. There are 473 employees within the dataset with a total service length of over 72 months. For the dataset including the outliers, there is a median of 11 months and a mean of 19.1 months. The first- and third-quartiles for this data are 4 months and 26 months respectively. This data is skewed right, indicating that a large portion of the employees tend to stay for a shorter time period than the mean value. With regards to our chosen thresholds for a candidate to be considered a successful retention, 70% of all employees in this dataset have worked at the company for at least six months, 49% have been there for twelve months and 28% have stayed for 24 months. We can use these statistics to compare the retention rate of applicants with different demographic and background profiles in order to assist the company to make the best hiring decisions possible in future, and to increase the longevity of their employees.

Data Preparation

This phase of our cycle is the cleaning and scrubbing of the data to prepare the final dataset for modeling. It is often referred to as data munging. This process is done in the munge directory of our ProjectTemplate. This data munging involves several steps including selecting, cleaning, constructing, integrating, and formatting the data.

Here we are determining based on our technical question which dataset we need to refine and drop based on our business requirements. For our initial cleaning, we select the Candidate Application data and remove the initial extra rows. Now we rename the relevant columns such as “Vocational” and “Academic” and pick the relevant columns from the total dataset like “Current Employee”, “Applied Through” and “Advertised

Salary” which would be necessary for our analysis.

We also convert some columns to match their types like matching all the vocational education together and binding the channels together. After having a well-defined dataset, we now replace all the Not found values with “NA” in the educational column and similarly, in the Channel column we replace Null with “NA”. This enables us to drop any observations which do not contain appropriate data for us to use. While doing this, we also need to convert the numeric columns to their numeric type and also handle the dates with a well-defined format. As we can see after all this cleaning we are left with many duplicate rows (approx. 2300 rows) in the Candidate Applications dataset having duplicate candidate IDs. To process, we add up the “length of service in months” for these and produce only one row for each candidate, to represent their total amount of time spent at the company.

Further, we create a length of service column for our analysis where we filter out all the candidates who are currently employed and have experience of fewer than 6 months in the current job position, which removes around 1800 rows. And also, we add three columns for the retention of the employee over 6, 12, and 24 months. These are binary variables, with candidates that reached that length of time regarded as a success, and candidates that didn’t as unsuccessful.

As our business understanding revolves around education data we create a subset from the Candidate Schools and merge, via a left-join, with the main data frame i.e. our Candidate Application dataset. Now here we handle all the variations of the different types of qualifications, for example all the different values of GCSsE, NVQs, bachelor’s degrees, diplomas and certificates. As we can see there are many candidates with specialisations like Nursing or Healthcare, so we make an education specialisation of each degree to dig more into the analysis of education, and how specialisation impacts candidates’ retention. As we need to make a left join with the Candidate Applications data, we need to rename the ‘candidate id’ column of both the datasets and convert them to numerical type.

To further expand on the analysis, we now clean the channels column by removing all the null values and then do an inner join with the Employee History dataset. Now for our analysis, we look at the ‘Application Date’ column and clean all the invalid date formats to a structured data format [DD-MM-YYYY] by using regular expressions. We also see some cases where only the year is mentioned, so we convert them to a valid date format. Now we remove all the null and NA values and do a left join to create a data frame for the analysis. We can also see that the competitor data show some relevance to the retention of employees. Here, the six columns of data are filtered. The idea is to combine the different candidates that retain in the job according to the length of the service broken down into different categories i.e. six months, twelve months, and twenty-four months. Therefore, when selecting the data, not only the six competition types were selected, but also the candidates’ retention data.

After observing the data, it was found that there were some NA values in the data, which would lead to subsequent inability to process the data for statistical purposes, so it was necessary to filter the data for null values. The filtered variables were renamed for ease of subsequent processing with different variable names, with the following effect.

‘Count of Care homes within 15 min drive’ -> CH15

‘Count of Care home Beds within 15 min drive’ -> CHB15

‘Average age of the care homes within 15 min drive’ -> AVG

‘Count of Care Homes within walking distance (<500 meters) from candidates address’ -> CHwalk500dis

‘Sum of care home beds within walking distance from candidates address’ -> CHBdis

‘Count of bus stops walking distance from candidates address’ -> BSdis.

We can also see that the advertised salary shows strong promise with the data. The Advertised Salary column has different kinds of values where rows contain ‘Competitive salary’, Range of salary (e.g., £7.50 - £9.55), £6.90 per hour, etc. This makes it very difficult to compare the data as so many values are stored in different formats. So, these values are replaced with numerical values. The rows which contain a range of possible salaries for a position are replaced with the average of those salaries. After replacing these values, all the ‘NA’ values are dropped from the data frame. After cleaning the records, the column ‘Advertised Salary’ is converted to numeric values, and a new column is added where the salaries are divided into various ranges, to help with visualisation of the data. Also, the length of service is divided into different ranges, to reflect whether employees reached a maximum of six months, twelve months, twenty-four months or stayed for longer than this. New subset data frames of the cleaned are then created, dividing the observations into nurses, carers, and senior carers. It is then observed that there are some outliers in the data frame, however the values where the advertised salaries are greater than 21 are kept in the data for analysis purposes, but ignored for visualisations for easier interpretation.

Modelling

In this phase of the project, we run some analysis to see if the objective of the project can be achieved using this dataset. The summaries below will help in answering the hypotheses questions.

How does channel type affect the months of experience?

The channel type variable represents the avenue through which an applicant applied to HC-One, for example job boards, referrals and HC-One’s website. To understand the effect that channel type has on an individual’s employment at the company, we explored a high level overview of the distribution of the candidates per channel type. The summary below indicates that the the job board channel has the highest number of applicants with 34% of the candidates applying through job boards. This led us to investigate what proportion of the applicants were successful when applying through various channel types.

channel_type	Channel
Advertisement	189
Golden Bees	1
HC-One Careers Website	916
Job Board	6644
Other	351
Referral	523
Social Media	180
Social Referral	24

We define success in this case to be in terms of retention. Therefore, a longer total of months of service would mean that the candidate was successful in their role, and that HC-One needs to incur less costs in recruitment for this role. For this phase of the analysis we made the assumption that candidates that stayed longer than 6 months are considered a success in this context. Figure 2 indicates the number of candidates that were successful for each type of channel.

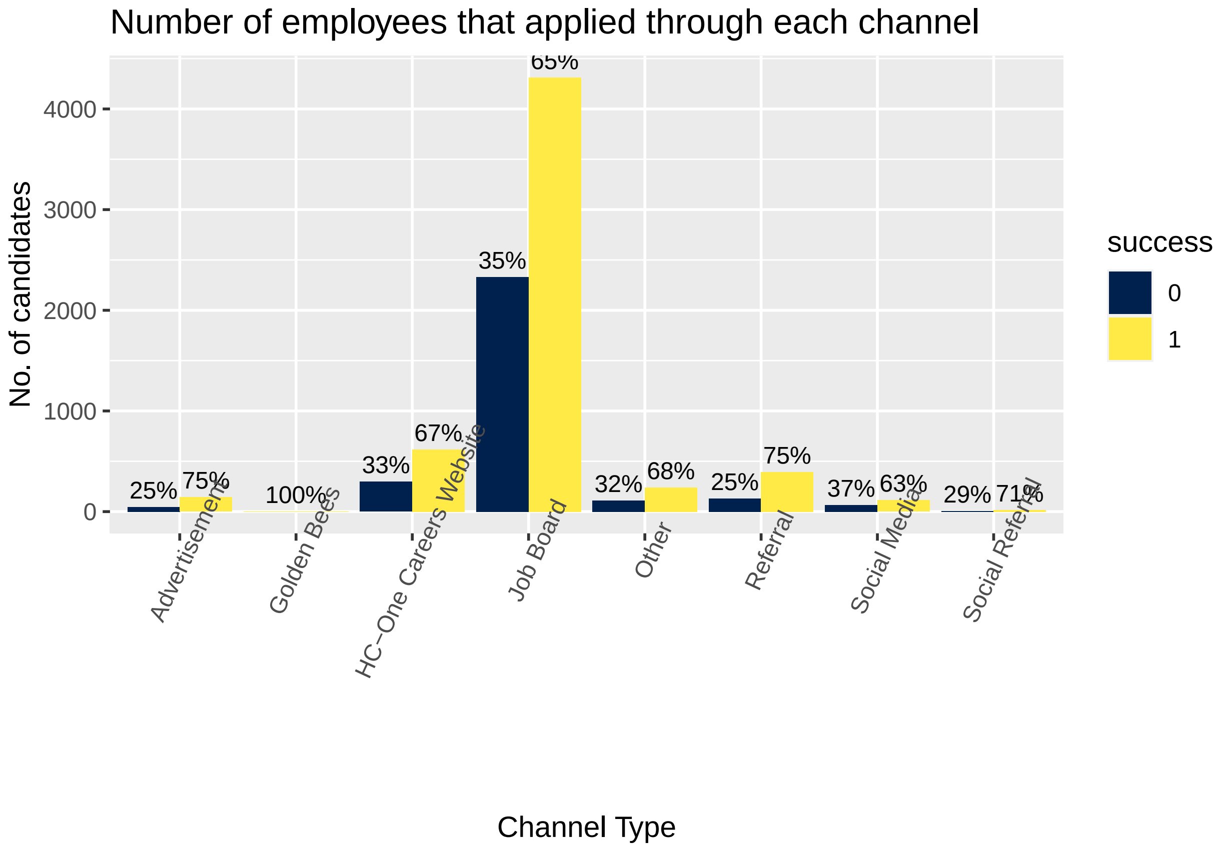


Figure 2: Bar chart showing the number of candidates by application channel, by success

From Figure 2, we can see that the applicants applying through all channels combined had over 50% of them were successful in their role. The plot also indicates that for the referral and advertisement channels we observed the highest percentages of success with 75% success. This however does not give us the entire picture into the average total of service months for candidates who applied through each channel.

Using a box plot, we can be able to check what the median months of service served for the candidates from each application channel was at HC-One over the observation period. From Figure 3, we can deduce that candidates who applied through the advertisement channel tended to have the longest stay at HC-One, with a median length of service of about 20 months. Candidates who applied through referrals also have a high retention with a median of about 18 months, with most of the candidates from this channel falling under the upper quartile.

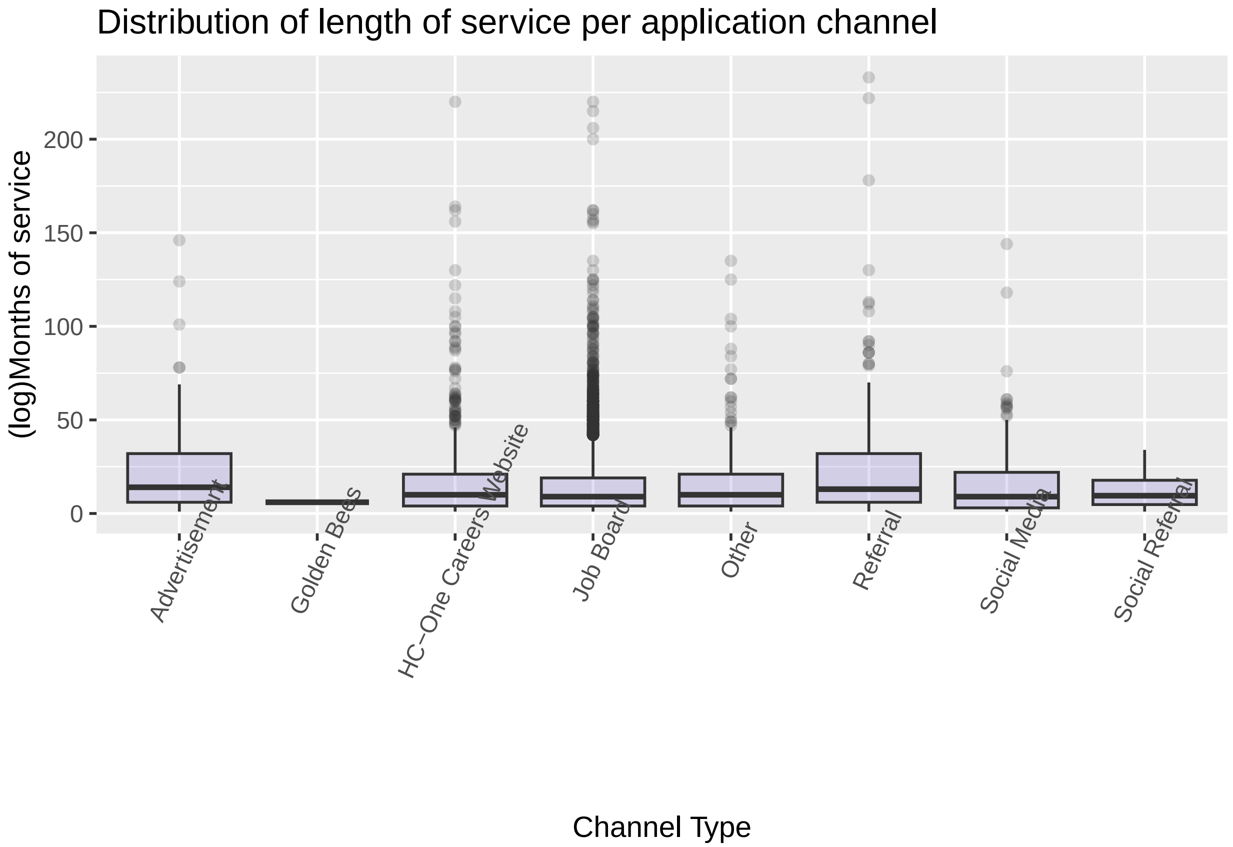


Figure 3: Boxplot showing the distribution of channels by service months

Next, we want to test the statistical significance of our observations on the relationship between channel type and length of service. Since the data is not normally distributed, we will use the Kruskal-Wallis rank sum test. This is a useful test when comparing multiple groups on an ordered variable, as in this case.

Our null hypothesis for this test is that there is no difference between the number of months served for employees who applied through different channels. The alternative hypothesis here is that there is a significant difference between the number of months served in the channels.

The test results in a p-value of 3.576e-15 which is less than 0.05, therefore we reject the null hypothesis and we can conclude that there is a difference between the number of months served for applicants from each channel, which we can observe from Figure 3.

From the box plots, we draw that referrals and advertisement channels have the highest median. We can use the Kruskal-Wallis test for these specific subsets to test the statistical significance. The p-value for referrals is p-value = 1.078e-13 and advertisement has a p-value = 1.045e-5 which is < 0.05 which validates the conclusions that referrals and advertisements have a higher median, and that they are significantly different from the other channels in terms of the number of months of service that candidates from the channel remain at HC-One.

Next, we would like to investigate whether the level of experience prior to employment at HC-One has an effect on the success of a candidate. This part of the analysis involved grouping the candidates into different buckets based on their months of service.

- Group 1: [0-50]
- Group 2: [50-100]
- Group 3: [100-150]
- Group 4: [150-200]

- Group 5: [200-250]
- Group 6: [250-300]

Figure 4 represents the median months of experience that the candidates served based on the group of months of service they belong to.

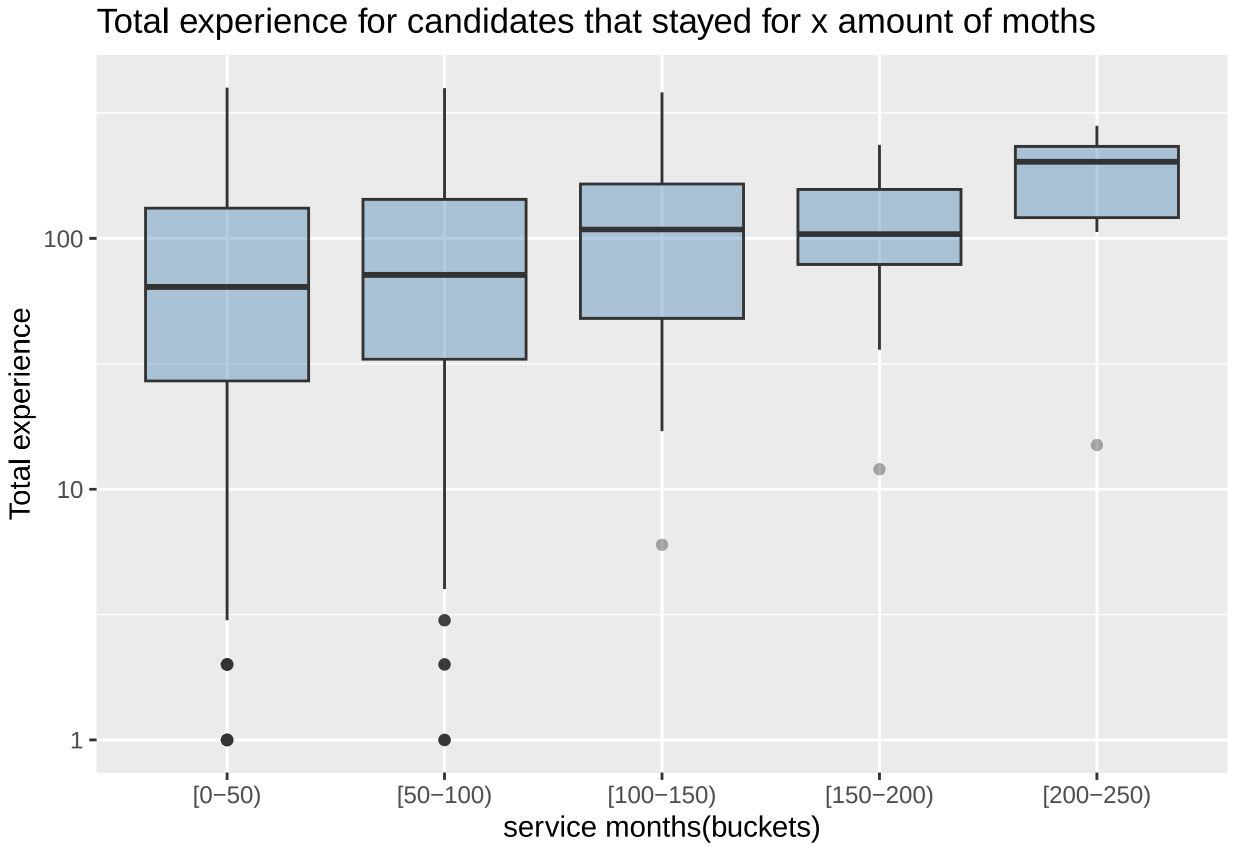


Figure 4: Box plot showing the service months in brackets versus candidates' total experience

Figure 4 indicates that applicants in the [200-250] group had the highest level of months of experience with over 150 months which is over 12.5 yrs, the median level of experience for the [100-150] is slightly higher than that of the group between [150-200] with the months of experience being 110 months. This shows a clear pattern that candidates with fewer months of service tended to have less experience.

To test the statistical significance of this results we will create a new hypothesis. Our null hypothesis is that there is no difference in the total experience for the different groups of candidates in terms of months of service. The alternative hypothesis is that there is a difference in the total experience for the different groups of months of service.

The test shows that p-value = 4.27e-5, which is less than 0.05. Therefore we reject the null hypothesis and conclude that there is a difference in the total experience for groups with different lengths of service at HC-One.

Does the candidate's Acorn Demographic Category have a relationship with their retention?

A candidate characteristic that may prove important to identifying candidate retention is the Acorn Demographic Category of where they live. The Acorn classification places across the United Kingdom into a hierarchy of categorisation according to an assessment of their inhabitants' sociological, economic and other lifestyle attributes. These categorisations are not specific to each candidate, merely providing a generalisation based on their catchment area around their registered address. An employee may happen to live in a catchment area with a high social status, but of poor means themselves. However, consideration of this variable might provide insight to the Business Objective. Without more specific data on the candidate's individual attributes in this field, using their catchment area is a good approximation of their likely attributes.

The highest level of Acorn classification [3] is divided into six categories:

1) Affluent Achievers

- Acorn describes this category as "some of the most financially successful people in the UK. They live in wealthy, high status rural, semi-rural and suburban areas of the country."

2) Rising Prosperity

- "These are generally younger, well educated, and mostly prosperous people living in our major towns and cities. Most are singles or couples, some yet to start a family, others with younger children. Often these are highly educated younger professionals moving up the career ladder."

3) Comfortable Communities

- "This category contains much of middle-of-the-road Britain, whether in the suburbs, smaller towns or the countryside. All lifestages are represented in this category."

4) Financially Stretched

- "This category contains a mix of traditional areas of Britain. Housing is often terraced or semi-detached, a mix of lower value owner occupied housing and homes rented from the council or housing associations, including social housing developments specifically for the elderly. This category also includes student term-time areas. There tends to be fewer traditional married couples than usual and more single parents, single, separated and divorced people than average."

5) Urban Adversity

- "This category contains the most deprived areas of large and small towns and cities across the UK. Household incomes are low, nearly always below the national average. The level of people having difficulties with debt or having been refused credit approaches double the national average. The numbers claiming Jobseeker's Allowance and other benefits is well above the national average. Levels of qualifications are low and those in work are likely to be employed in semi-skilled or unskilled occupations."

6) Not Private Households

- This category is made-up of areas where most residents do not live in private households. This includes military bases, holiday accommodation, children's homes, hospitals and prisons.

Of the 19291 employees recorded in the Candidate Applications dataset, 107 of these do not have Acorn categories recorded. The counts of each Acorn category for the remaining are shown in Table 1, and are plotted in Figure 5. We see an overwhelming majority of employees, 69.3%, are from category 4 ("Financially Stretched") and 5 ("Urban Adversity"), and therefore tend to be from areas with lower social status.

Category code	Count of Employees
1	1275
2	957
3	3544
4	6595
5	6692
6	121

Table 1: A table showing the number of employees in the CandidateApplications dataset for each of the Acorn Demographic groups

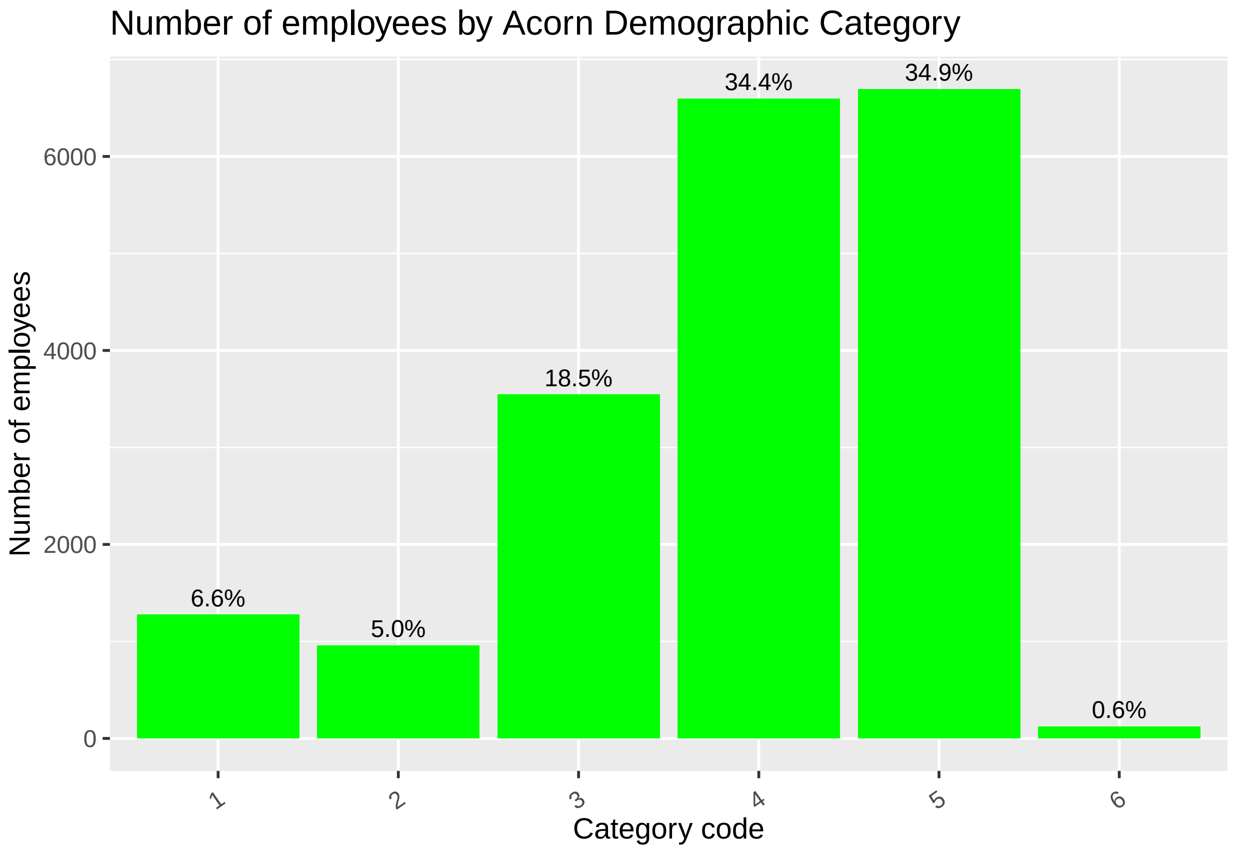


Figure 5: Bar chart showing number of employees in the CandidateApplications dataset for each of the Acorn Demographic groups, excluding candidates without a classification.

Figure 6 shows, for each Acorn Demographic Category, the percentage of employees retained for four retention brackets: <6 months, 6-12 months, 13-23 months and >=24 months. Each Category sees a very similar set of percentages of each bracket, with the majority of the differences being less than 3%. The largest percentage difference is 4.7% and is found for the >=24 month bracket between Category 1 (Affluent Achievers), with 33.1% retention and Category 4 (Financially Stretched), with a retention rate of 37.8%.

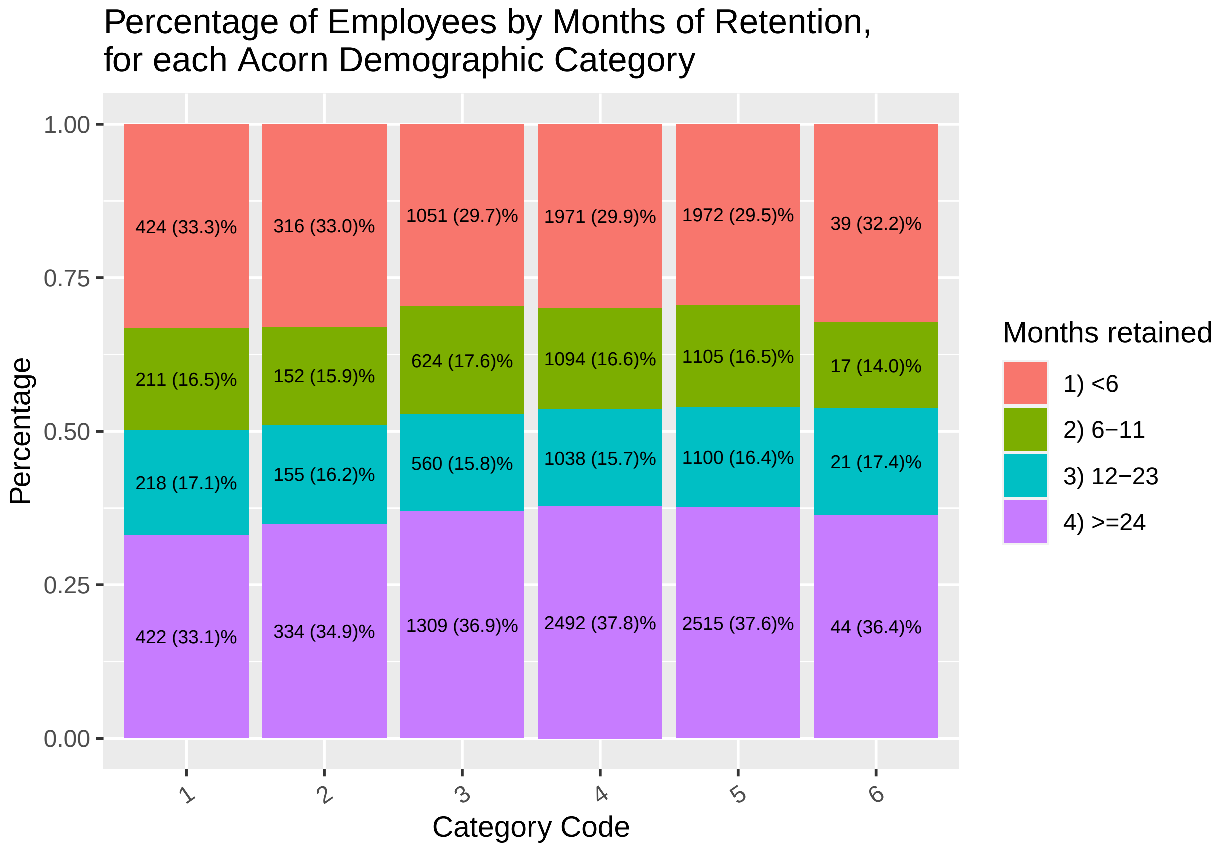


Figure 6: Stacked bar chart showing the percentage of employees by brackets of months of retention for each of the Acorn Demographic groups

To test for the statistical significance of the relationship between category code and months retained, a Kruskal-Wallis test is used. This is because the data is not normally distributed, therefore breaking an assumption of the standard ANOVA test. The Anderson-Darling Normality Test's p-value for candidates' months of service is 3.7×10^{-24}, therefore it is not normally distributed.

The Kruskal-Wallis rank sum test returns a p-value of 0.0663601, meaning it is found not to be statistically significant. We may therefore conclude that there is likely not a meaningful connection between an employee's Acorn Demographic Category and the level of retention observed by employees of that category at HC-One.

What is the relationship between the level of competition of care homes in a candidate's local area and the candidate's retention rate?

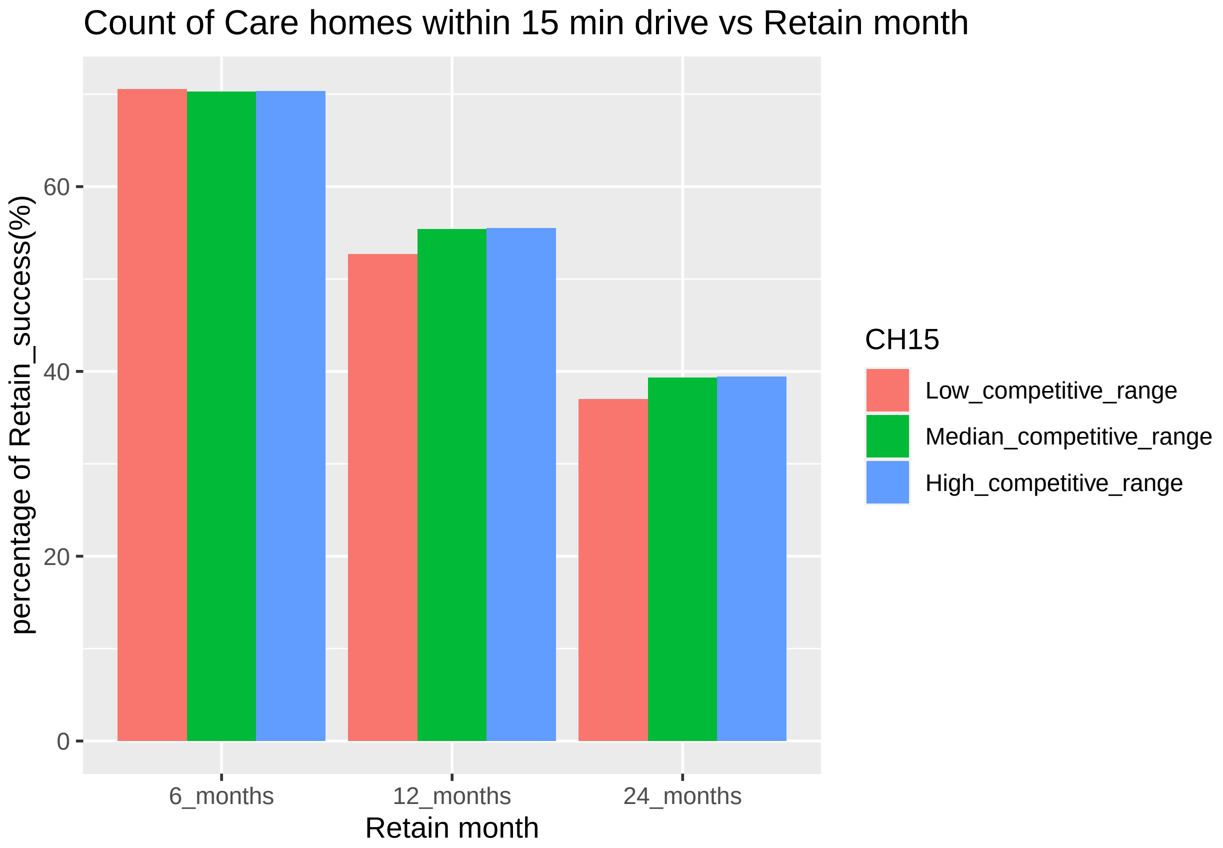
An exploratory analysis was carried out on the obtained variables in order to understand the relationship between the level of competition and the level of retention observed for candidates. Initially, a principal components analysis was carried out to discover the influence of different variables, extracting the different principal components of the variables and plotting the data to demonstrate these relationships. We divided the competition zones into six different types, based on the amount of care homes, the amount of care home beds, and the overall level of competition in the candidate's local area. When comparing the six different variables for competition, some linear relationships can be found between the six competing variables, but the correlations that are more evident only occur between similar types, for example between the number of care homes within a 15 minute drive and the count of care home beds within a 15 minute drive. However, in the results of the principal components analysis, it can be found that the level of retention does not have a significant relationship with the six different competition types, and does not show a clear relationship, but it is possible to analyse the proportion of candidates retained who live in the high and low competition zones under different thresholds for retention, which is convenient to use as one of the reference indicators.

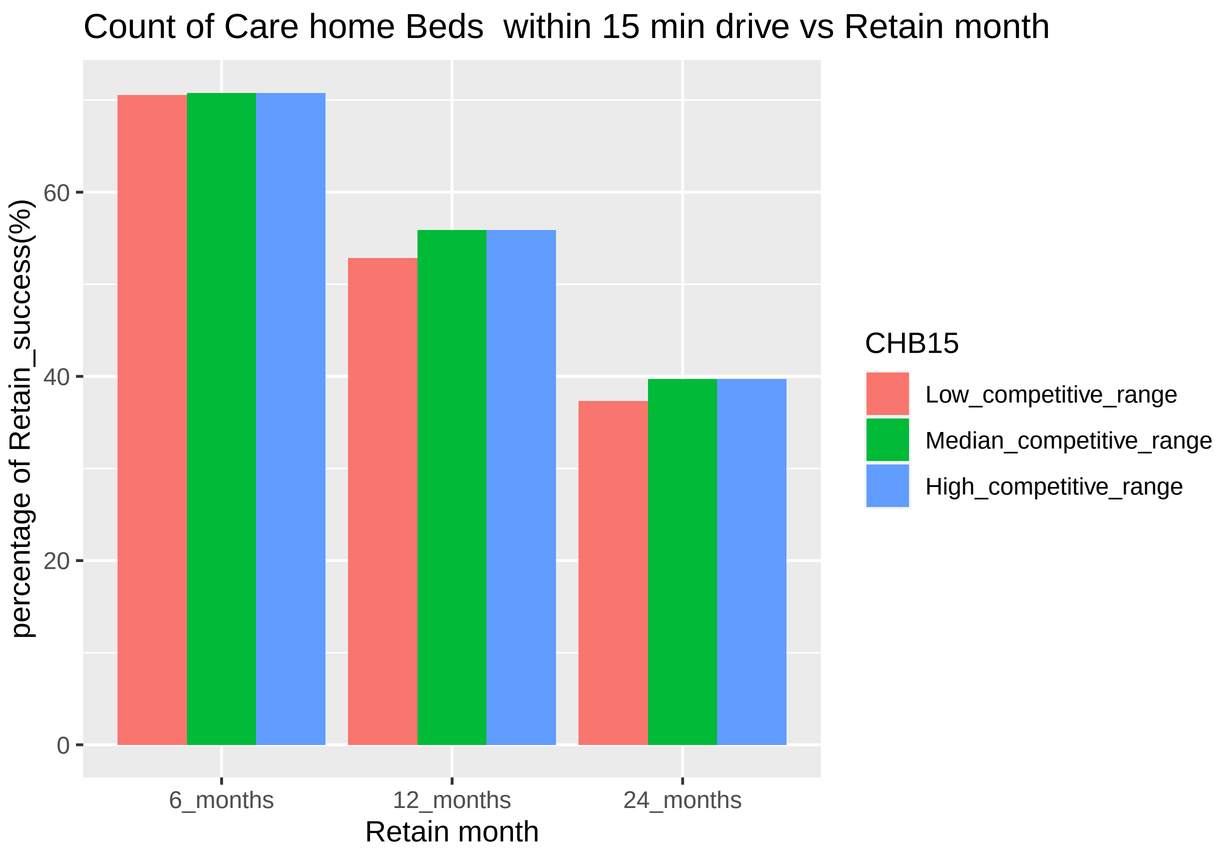
Distinguishing between the different competition zones

The data is partitioned using the method above and then distinguished by a function to obtain different competition brackets. The higher the number of competitors in the competition interval, the higher the competition level.

Linking the relationship between the two types of data

After generating the different competition intervals, they were analysed against the level of retention for candidates to obtain the probability of successful retention for the different brackets as a way to analyse the relationship between the competitor data and levels of retention. In Figures 7 and 8, we display the relationships that the number of care homes in a 15 minute drive of the candidate's home, and the number of care home beds in this area, have with the number of months of retention for candidates at HC-One.





Figures 7 & 8: (1) Bar chart showing the count of care homes within 15 minutes drives versus retention month brackets (2) Bar chart showing the count of care homes bed within 15 minutes drives versus retention month brackets

Comparing the respective relationships of 'Count of Care homes within 15 min drive' and ' Count of Care home Beds within 15 min drive' with length of retention shows a similar pattern, and as the retention time increases, it is difficult to see the differences due to the similar probabilities of the three intervals. For example, if the count of care homes within 15 min drive is held constant, the longer the time, the lower the retention rate. The shorter the time, the higher the retention rate. When looking at retention rates for 6 months, 12 months and 24 months, the highest retention rate was 70.56% (low), 70.28% (medium) and 70.34% (high) for 6 months, followed by 52.68% (low), 55.37 (medium) and 55.52% (high) for 12 months. The lowest is found at 24 months at 36.98% (low), 39.30% (medium), 39.42 (high). As a result, there is not a large difference observed in retention rates for candidates from areas with different competition levels across each of the time splits. The low competition retention rate decreases from 70.56% at six months to 52.68% at 12 months and 36.99% at 24 months. In contrast, the proportion of competitive intervals also gradually decreases and the low competitive interval appears significantly lower than the other two competitive intervals, which can be used as an aid in the initial screening of candidates.

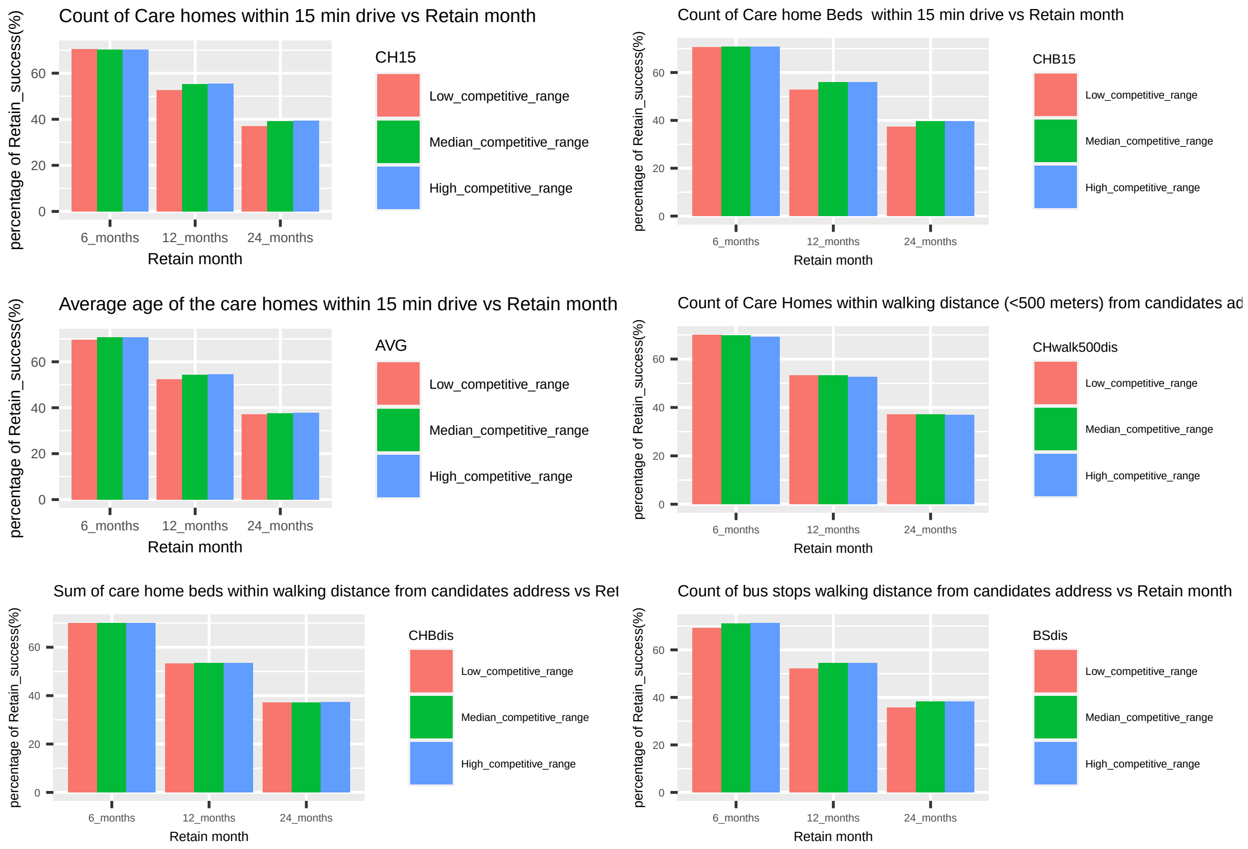


Figure 9: Bar chart combination of all competition variables versus retention brackets

The six different competition types show a similar effect, with competitiveness all decreasing with retention time. The small difference in the data for the three competitive types also indicates that there is not a strong relationship between retention time and competitiveness and can only be used as a reference criterion for the slightly lower retention rate prevalent for low competitiveness, which can assist in assessment of future candidates. However, two similar sets of competitive types represent more approximate effects, and have significant auto-correlation, hence one or the other should be chosen as one of the auxiliary criteria for candidate selection, such as either selecting the amount of care homes within a 15 minute drive, or alternatievly the count of care home beds within a 15 minute drive, for future selection.

Evaluation Cycle 1

In the exploratory data analysis of competitors in a candidate's local area during the first cycle, we can conclude that the data related to competition data and access to bus stops has a positive correlation for recruiters on retention time. For example, with respect to the count of care homes within a 15 minute drive, the longer the time, the lower the retention rate, for a constant level of competition. The shorter the time the higher the retention rate. When looking at retention at 6 months, 12 months and 24 months, the highest overall retention rate is about 70.56% at 6 months, followed by 52.68% at 12 months and the lowest overall retention rate is observed for 24 months at 36.99%. In order to get a better idea of what the percentage of the different intervals of the data is, we compared the retention time across different levels of competition. Analysis of the graphs shows that the six different competitive types perform similarly overall, with competitiveness all decreasing with retention time. The subtle difference is that this competitiveness decreases slightly faster in the low range compared to the middle and high ranges, while the data in the middle and high ranges is relatively stable for longer periods of service.

How does an employee's educational background influence their retention?

Educational background and qualification is an important factor in employee retention since it has a potential impact on an individual's competencies, expertise, and career trajectory, all of which can influence job satisfaction and commitment to an organisation. Moreover, highly-skilled employees with specialised education can contribute significantly to a company's productivity and success, increasing the likelihood of their retention. The distribution of qualifications in the Education Department is depicted in the pie chart in Figure 10.

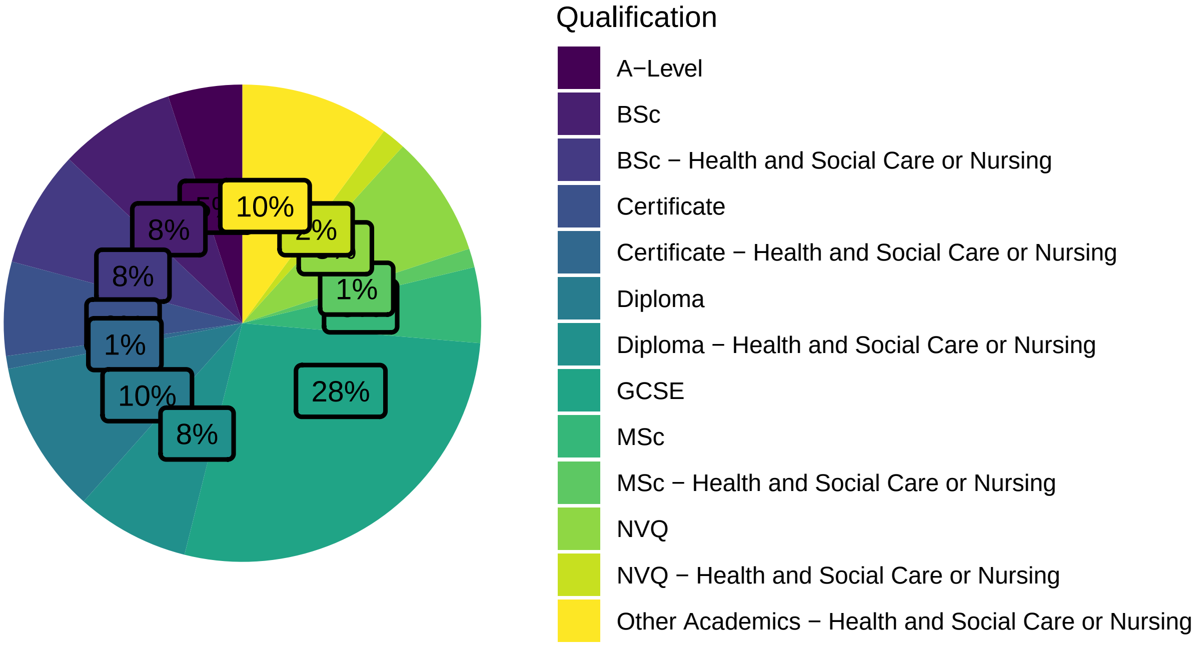


Figure 10: Count of Employees Distribution by Qualifications

Figure 10 illustrates the breakdown of employee education, with each segment representing a different type of education. We investigated the highest level of education observed for each employee for which data is available. Based on the pie chart, the most common maximum education level for employees is a GCSE-level education, accounting for 27.57% of the total distribution.

Diploma-level education, which accounts for 10.35% of the total, and A-Level education, which collectively accounts for 5.03%, are also important aspects of the distribution. The chart also involves BSc and MSc levels of education, but their percentages are comparatively small, with BSc reporting for 7.91% and 5.11% of the total, respectively.

The chart also includes information on education in courses specifically related to health, social care, or nursing, which is represented by different segments, at certificate-level education, diploma-level education, NVQ-level education, and other academic qualifications. In accordance with the data, diploma-level education in health and social care or nursing is more common than other types of education in this field, taking account for 7.68% of the total distribution.

Overall, Figure 10 illustrates the educational distribution of employees, with most of employees having GCSE-level education and a relatively small proportion having higher education qualifications, such as BSc or MSc. According to the data, there are a significant number of employees with qualifications related to health and social care or nursing.

To further analyse the data presented in the pie chart, it is useful to compare it with the pie chart distribution of qualifications based on Health and Social Care or Nursing. This chart provides insights into the distribution of qualifications specifically in this field.

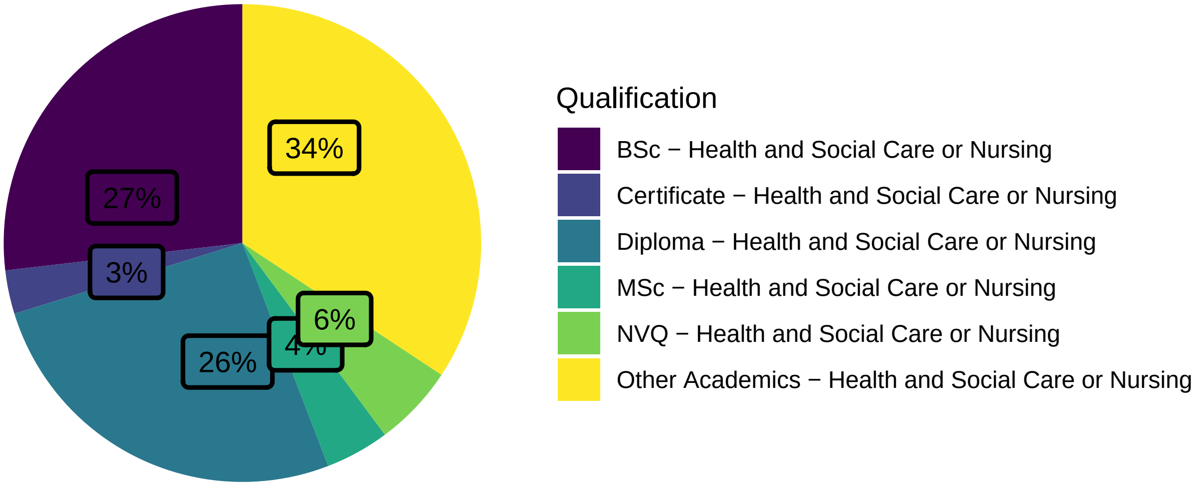


Figure 11: Count of Employees Distribution by Health and Social Care or Nursing Qualifications

Figure 11 represents the distribution of employee education in the Health and Social Care or Nursing field. The graph displays six different categories of educational qualifications, namely BSc, Certificate, Diploma, MSc, NVQ, and other qualifications in Health Care or Nursing. The highest percentage of employees (34.3%) fall under the category of "Other Academics - Health and Social Care or Nursing," indicating that a significant proportion of employees have pursued miscellaneous qualifications in Health Care and Nursing other than those listed. The second-highest proportion of employees (26.83%) have a BSc degree in Health and Social Care or Nursing, while 26.07% hold a Diploma in the same field. Additionally, 5.52% of employees have an NVQ qualification, 4.33% hold an MSc degree, and 2.95% have a Certificate in Health and Social Care or Nursing. These statistics represent the highest-level qualification obtained by each employee that has at least one health, social care or nursing qualification.

This distribution of employee education can provide insights into the skills and expertise present in the organization. For example, having a significant proportion of employees with BSc or Diploma degrees may indicate that the organisation prioritises hiring individuals with formal education and training in the field. Moreover, the proportion of employees with MSc degrees may indicate that the organization values higher education and encourages employees to pursue advanced degrees to enhance their skills and knowledge. Similarly, having a relatively low proportion of employees with Certificates may suggest that the organization prefers employees with more extensive training and education.

When compared to the pie chart distribution of qualification with respect to employee retention, it may be possible to identify whether there is a correlation between employee education and retention rates. For example, if the organisation has a higher retention rate for employees with BSc degrees, it may indicate that the organisation values this particular qualification and provides opportunities for employees with BSc degrees to grow within the organization, and help us to make projections for future candidates based on their educational background.

A comparative bar graph can be created using the given data to visually represent the percentage of retention and non-retention for each qualification type which can be used to make informed decisions on employee retention strategies for HC-One.

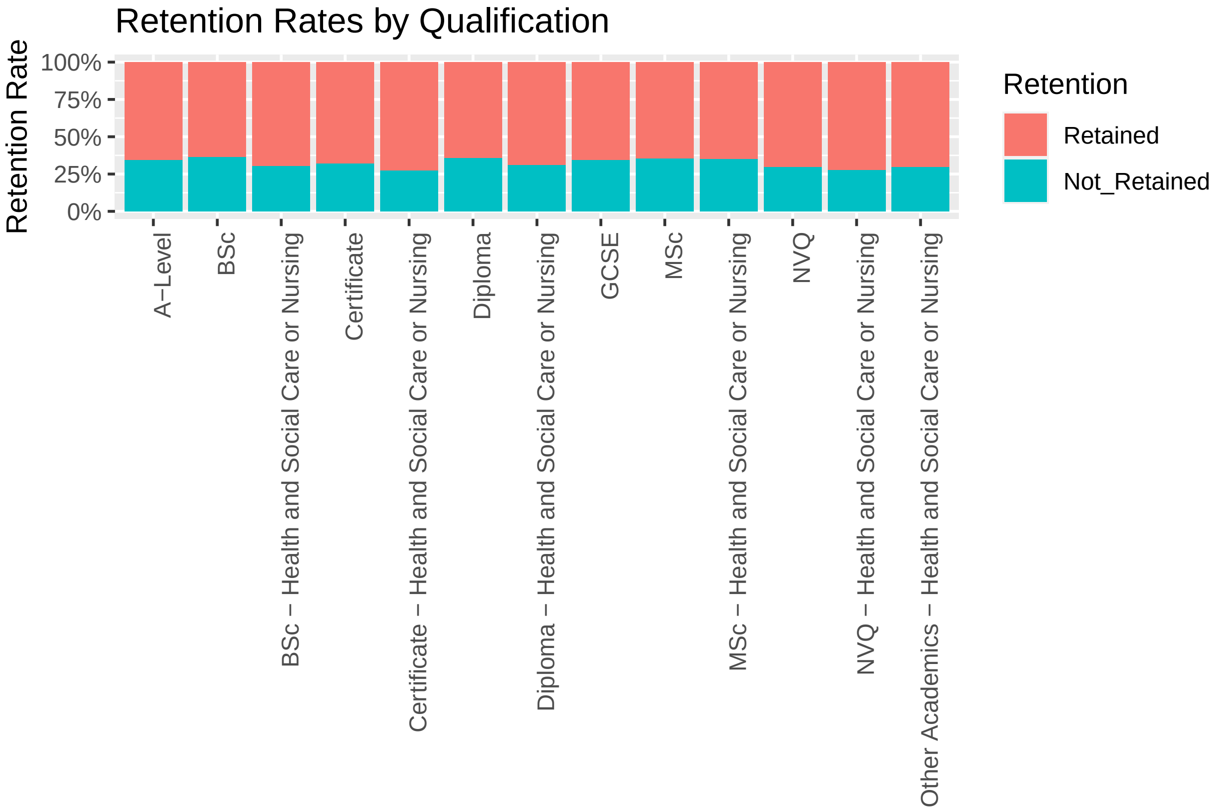


Figure 12: Percentage of Employees by Retention for each Qualification

From Figure 12, we can observe that academic qualifications related to health and social care or nursing have a higher retention rate than other academic qualifications. For instance, the qualifications with the highest retention rates are “Certificate - Health and Social Care or Nursing” at 72.6% and “NVQ - Health and Social Care or Nursing” at 72.4%. This suggests that people who have pursued qualifications in health and social care or nursing are more likely to remain in their jobs at the company longer than those with alternative qualifications. On the other hand, qualifications with the lowest retention rates are “BSc” at 63.7%, “MSc” at 64.7%. This implies that people who have pursued these academic qualifications are more likely to leave their jobs.

In conclusion, the graph indicates that academic qualifications related to health and social care or nursing have a higher retention rate than other academic qualifications. This could be attributed to the high demand for these skills in the job market, leading to more job security and career growth opportunities.

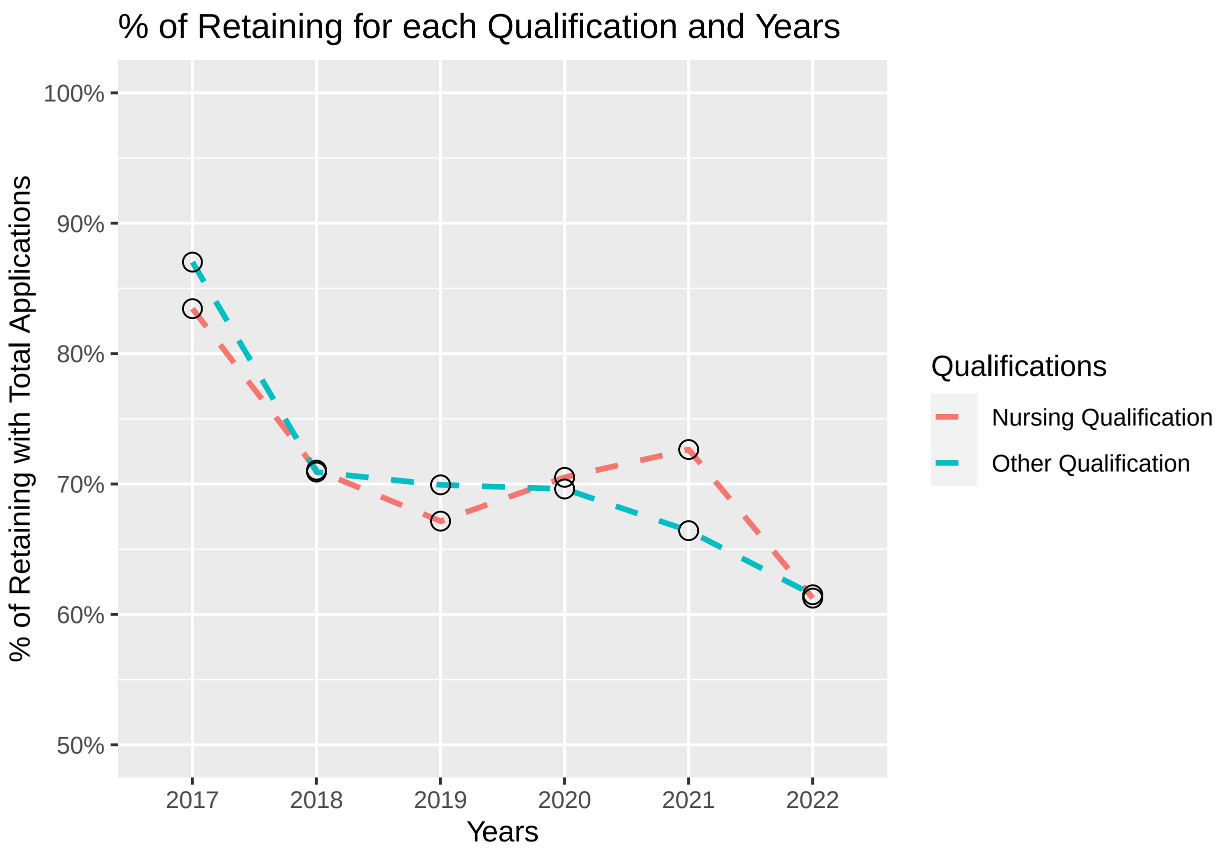


Figure 13: Trend Line for all qualification, with regards to retention.

Figure 13 depicts that the percentage of employee retention for nursing qualifications was highest in 2017, at 83.45%, and decreased to 61.25% in 2022. On the other hand, the percentage of employee retention for other qualifications was highest in 2017, at 87.02%, and decreased to 61.51% in 2022. It is observed that applicants having nursing qualifications are more likely to be retained when compared to applicants with other qualifications, in recent years.

How does the Advertised Salary for a position influence employee retention?

In order to investigate the impact that the advertised salary for a job opening has on the quality of an employment decision, we decided to explore the association between the overall length of time an employee has worked at the company for, and the salary that was advertised for the position when they applied for the job. We divided the lengths of service by employees into four different bands, indicating whether an individual left the company before six months, between six months and a year, between one year and two years, or stayed for over two years. Additionally, after pre-processing the data on advertised salaries, we decided to split these into intervals of £1 per hour, for all intervals from £8 to £21, due to a low amount of salaries either below £7 or above £21 per hour.

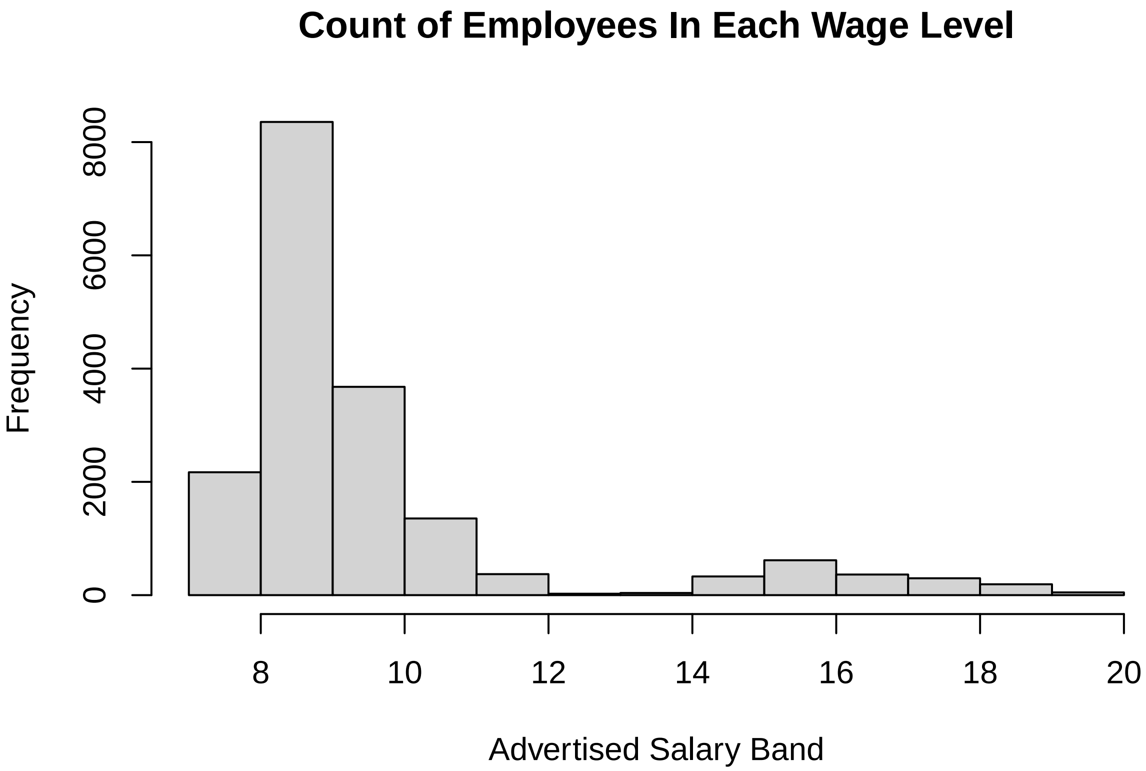
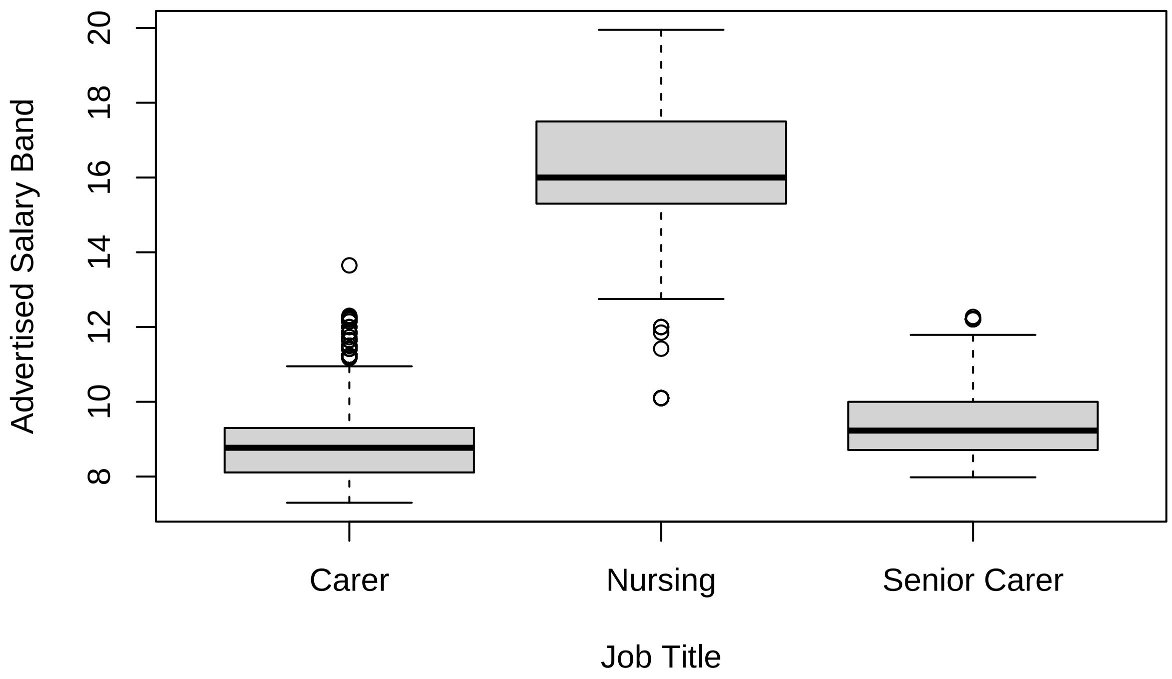


Figure 14: Histogram showing the number of employees by bracket of advertised salary

Figure 14 shows how many employees at HC-One across the period were advertised a salary within each band. The majority of observations are at the lower end of the distribution, with a very low amount in the mid-range of salaries, compared to those at the higher end. A possible reason for this is the different jobs that have been observed in the dataset. There are 13393 employees in the dataset who were employed as a carer, 1024 senior carers and 2003 nurses. Carers tend to earn less than the other two roles, with nurses earning significantly more with almost no overlap with the other two groups. Figure 15 below shows a summary of the distribution of salaries for each role.


Advertised Salary Band
8
10
12
14
16
18
20
Carer
Nursing
Senior Carer
Job Title

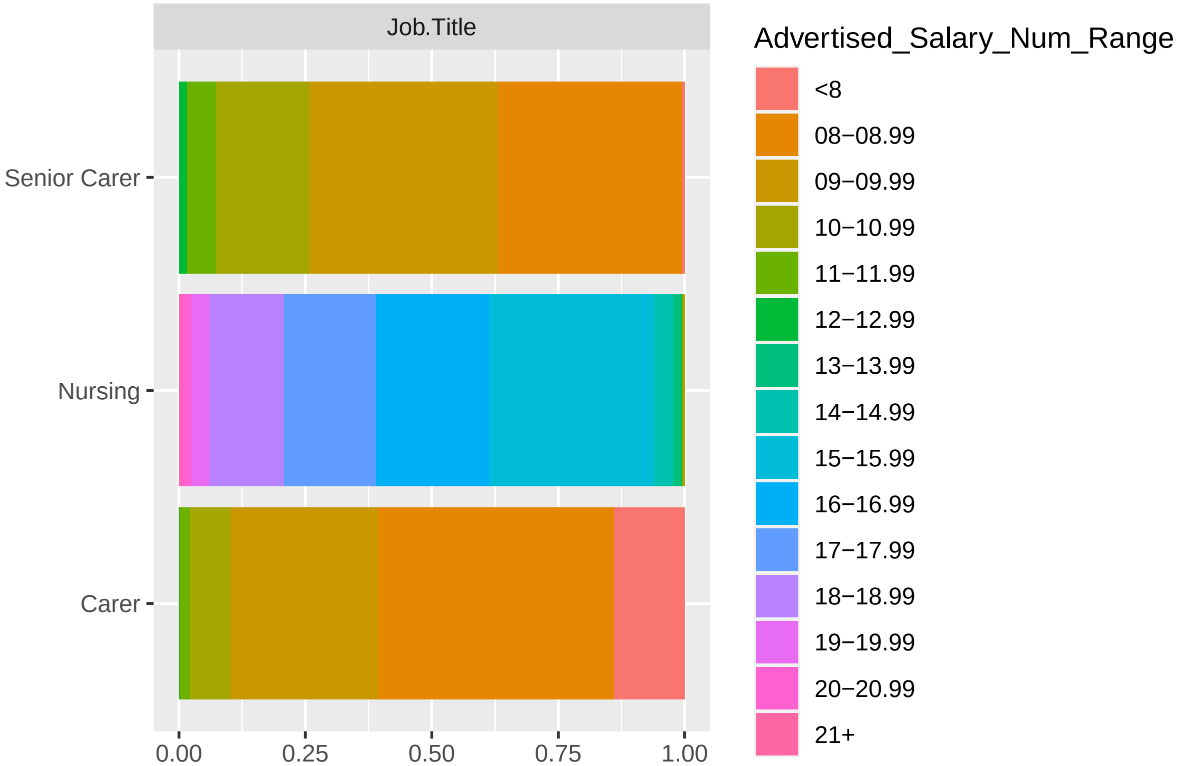


Figure 15: Boxplot showing the distribution of different salary brackets when split by job titles

In addition, Figure 15 shows the distribution across different salary bands when splitting candidates according to their job. Candidates working a nursing role almost all applied to jobs with a higher advertised salary than candidates working as a carer or senior carer. In addition, the lowest salaries (below £8 per hour), are almost all earned by carers. This means that conclusions drawn from this data overall about the correlation between advertised salary and retention rate may be biased if there is a significant difference in retention rate between different job roles. When considering the data overall, we discovered that there was a slight negative correlation between advertised salary and job longevity. This became more negative when removing outliers, of which some were very implausible values generated by issues in cleaning the data.

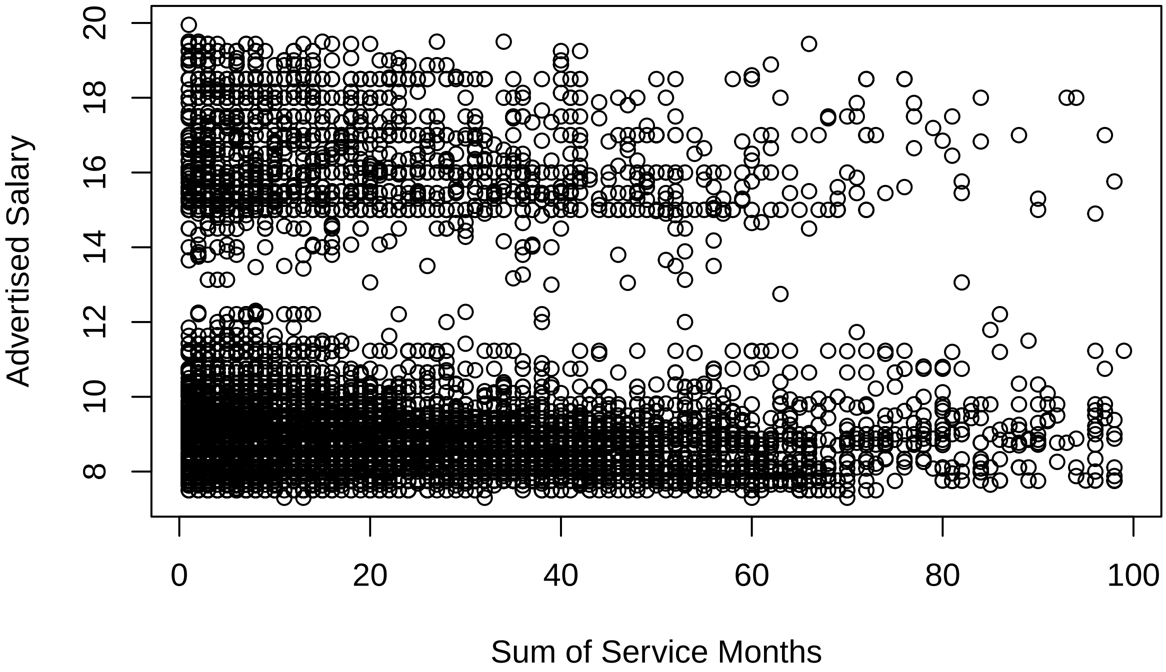


Figure 16: Bar chart showing the number of months of service by advertised salary

Figure 16 shows the overall association between the two variables in consideration, the number of months of service at the company and the advertised salary for the position of the given candidate. The issue of different job titles having different salaries again shows up here. It can loosely be observed that there are two distinct bands of salaries, meaning that an overall correlation coefficient may not be the best representation of the relationship between these variables. By controlling for job title, we can further analyse how this relationship exists without biases introduced by the different jobs, and we can consider whether the relationship is more significant for a particular job.

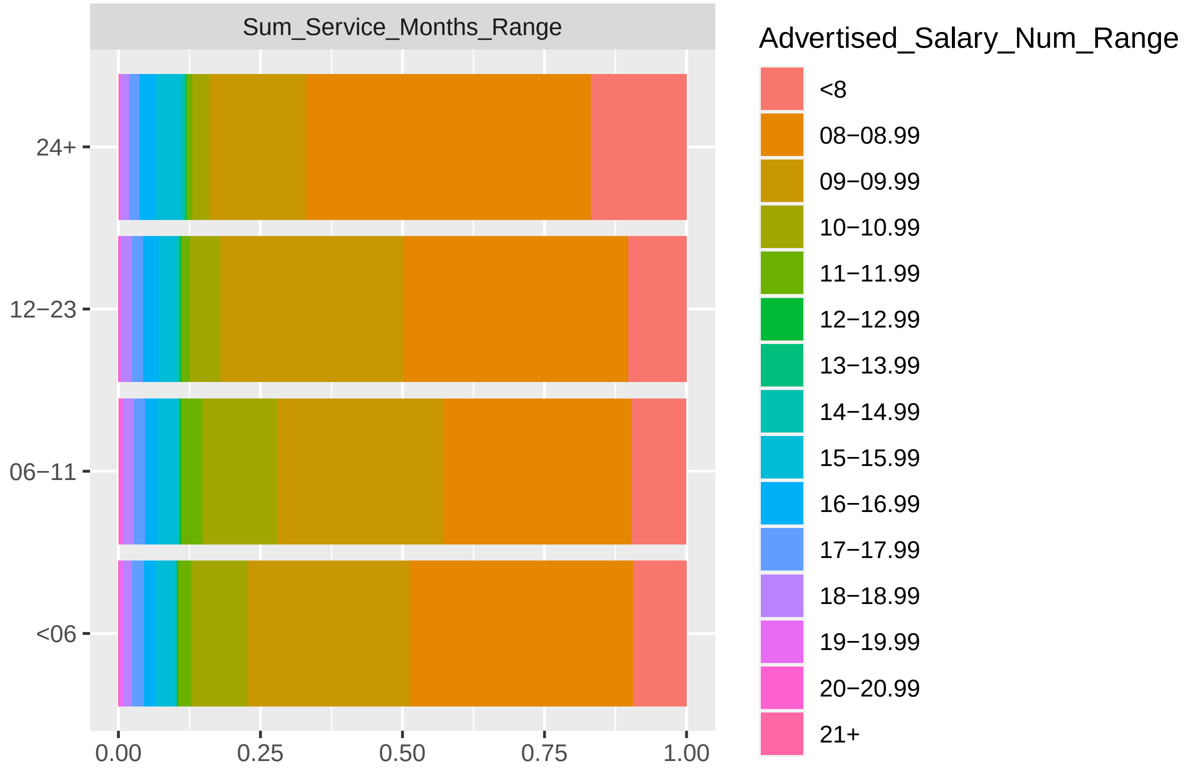


Figure 17: Bar chart showing the distribution of advertised salary bracket by length of service bracket

Figure 17 shows the distribution of advertised salaries for each band of lengths of service, for all jobs taken together. Salaries in this plot are ordered in descending order from left to right. An interesting aspect of this analysis is the proportion of those that have worked at the company for at least 24 months that are on lower salaries. The proportion for this category of employees earning less than £9 per hour is significantly higher than for the other three categories. An issue here, however, is that the vast majority of employees who were advertised these salaries are working as carers, so by considering the distribution for each role separately, we can learn further about possible conclusions from this data. The salary band that appears to have the lowest relative retention rate is people earning from £10-11, and over longer periods, people earning from £9-10. These are generally individuals earning on the higher end working as a carer or senior carer, so the company may wish to consider the impacts of this on its hiring process, and the longevity of prospective employees for these roles. By further considering the split across job roles, we can understand how these differences in retention rate are correlated across advertised salary bands in the absence of bias from individuals' chosen jobs.

Evaluation

During this project, we investigated five different areas in relation to their impact on successful retention of candidates over time at HC-One. Our selected variables were considered in terms of their impact to predict whether candidates would be retained across a period of six months, twelve months or twenty-four months. Through our analysis, the company can draw conclusions about the aspects of candidate's applications that make them more successful for the company.

Through exploration of candidates' Acorn demographic groups, we aimed to learn about how people from different backgrounds, based on estimated social status of their local neighbourhood, perform in terms of retention at the job. However, we discovered that there was no significant relationship between candidates from different Acorn groups in terms of job retention. This held when looking at candidates' retention rates for each of the periods that we analysed. From this, we believe that it would be unwise for the company to use this type of categorisation in order to prioritise future candidates for jobs, without further evidence to suggest there is a correlation.

An aspect which we found to have an association with higher retention rates was candidates' educational background. Through investigation of the relationship between the type of degree studied, for those candidates with degrees, we learnt that candidates with a background in a specialised field of study related to the job, such as nursing, were more likely to be retained in the job for a longer period, when compared to candidates with more generalised fields of study. In addition, from the trendline graph produced during the analysis, the percentage of employees that have been retained having taken nursing qualifications is higher than those with other qualifications, in recent years. This is an avenue where it would be useful to explore further in terms of the breakdown of retention rates for different qualifications in each of the different jobs, to see whether these specialised qualifications only hold for certain areas of work. However, this data already shows that targeting candidates with specific qualifications could be a good strategy for the company in future to ensure the highest retention rates.

Employment history is an aspect that is commonly associated with candidates with higher potential during hiring processes. During our investigation into the data on candidates' previous employment history, we found data that seemed to match this assumption. Specifically, we found that candidates with a higher length of service at HC-One, for example those with between 200 and 250 months of service, had a higher median total experience than those at the lower end of lengths of service, between 0 and 50 months. This means that when trying to optimise for longer-term employment choice, it may be best for HC-One to select candidates with more prior experience. Another factor that seemed to be significant in comparison of retention rates is the channel through which an employee applied for the job. Our conclusion for this analysis is that applicants through advertisements and referrals have a significantly higher median retention rate than other applicants in terms of total number of months of service. As a result, this would be another useful metric for the company to use to prioritise future applicants.

When looking at alternative care homes in the local area of a candidate, we split the dataset into six different competition types, based on the number of other care homes and number of beds in other care homes, to test the hypothesis that this affects a candidate's propensity to remain at HC-One. We discovered that the performance of the six different types was relatively similar. A small difference that we noticed was that the competitiveness decreased in the lower range compared to the middle and high ranges, while the data in the middle and high ranges is relatively stable. This may mean that candidates whose decisions are affected by competitiveness of other care homes are more likely to leave the job in the short-term, but this does not have a large long-term impact.

Finally, we looked at how the advertised salary for a position affects the retention rate for candidates. We noticed that there was a slight negative correlation between advertised salary and overall length of service. There could be several factors causing this. However, an issue that we encountered was that there was a large gap between salaries for different jobs. Specifically, nurses almost always earned more than senior carers and carers, with senior carers mostly earning more than carers, and many more carers earning minimum wage. This means that it is difficult to draw clear conclusions about the overall correlation between these variables as the distribution is so separated. For this reason, we decided to do another cycle of this analysis to see how advertised salary affects retention rate when controlling for each of the different job types.

The next step of this project is to separate the dataset into three different subsets, one for each job title. By considering this separation, we can learn about the differences in variables affecting the retention rate in each job, as there are significantly different characteristics for the individuals that are applying for each job, and we can learn different decisions for the company to make for new applicants based on the job that they are applying to.

CYCLE 2

Business Understanding

Business Objectives

The business objective for this cycle remains the same as for the first cycle of data mining, which is to build insights based on our selected avenues of investigation that will help HC-One in improving efficiency in the recruitment process. In this cycle, we will focus on the same objective but by breaking down all the hypotheses by job title.

Business Success Criteria

Success for this project would be to uncover the insights that can be useful in discovering ways to improve recruitment efficiency. Specifically in this cycle, our goal is to find out if there is any relationship between length of retention at the company, job title and the all the other variables from our analysis in relation to length of service. There are three different jobs which we will subset the data into: nurses, carers and senior carers. We hope to understand how relationships are similar and how they differ across the three roles.

Data Preparation

As we will be using the same data set as for the previous cycle, we don't have to clean or scrub the data anymore, as it is sufficiently pre-processed for this analysis. Our only preparation is creating the subsets necessary for splitting the data based on job title. We will be using the data frames to further expand our analysis in this cycle. In this phase, we can do a recheck of the data set to ensure the data quality is maintained and there is no loss of any data or introduction of new erroneous values after our previous cycle of CRISP-DM.

Modelling

After getting an understanding of the data, and constructing the data, the next step of this project was to feed the data into graphical summaries, to derive insights from our data set.

We will find out in this phase the relationship between all the hypotheses questions with the job titles.

Application Channel & Employment History

This analysis is a deep dive into how the channel of application affects success of a candidate grouped by the job title. This will allow us to see retention of candidates in different roles.

```
channels_plot2
```

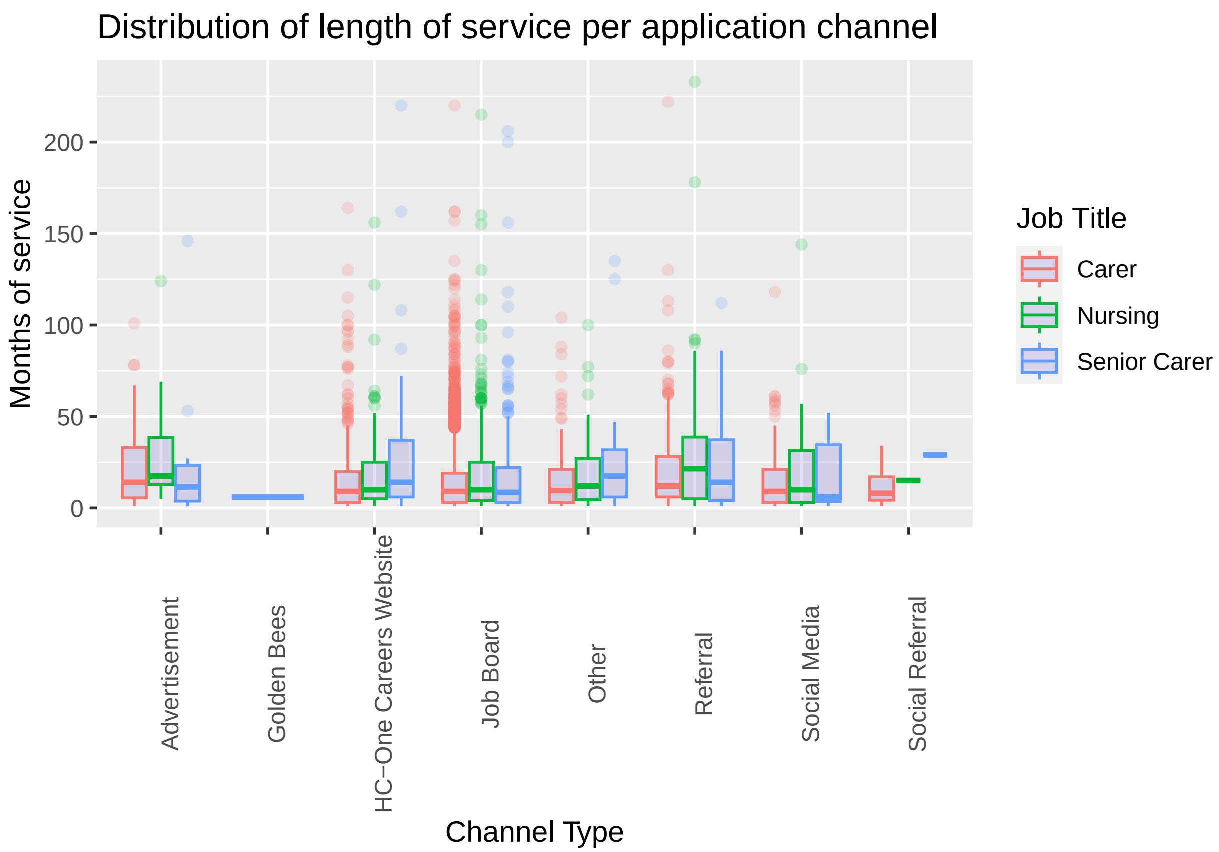


Figure 18: Boxplot showing channel type and months of service, grouped by job title

Our null hypothesis for the application channel's relationship with retention is that there is no significant difference across job titles for each channel type in terms of length of service. The alternative hypothesis is that the job titles in each channel type do have a difference in terms of length of service.

The statistical results show a p-value of 2.91e-07, which is less than 0.05. Therefore, we reject our null hypothesis and conclude that the different job titles have a difference in the months of service across the different channels.

From the plot containing information on the different channels, there were different job titles having the highest median in terms of total months of service. The overall highest median was slightly over 25 months in the social referral channel for senior carer candidates. However for this group there seem to have been very few candidates, so it is difficult to draw significant conclusions from this data. This is followed by a median of about 23 months in the referral channel by nursing candidates. This aligns with the observation in the first cycle where referrals had a high median.

```
job_box
```

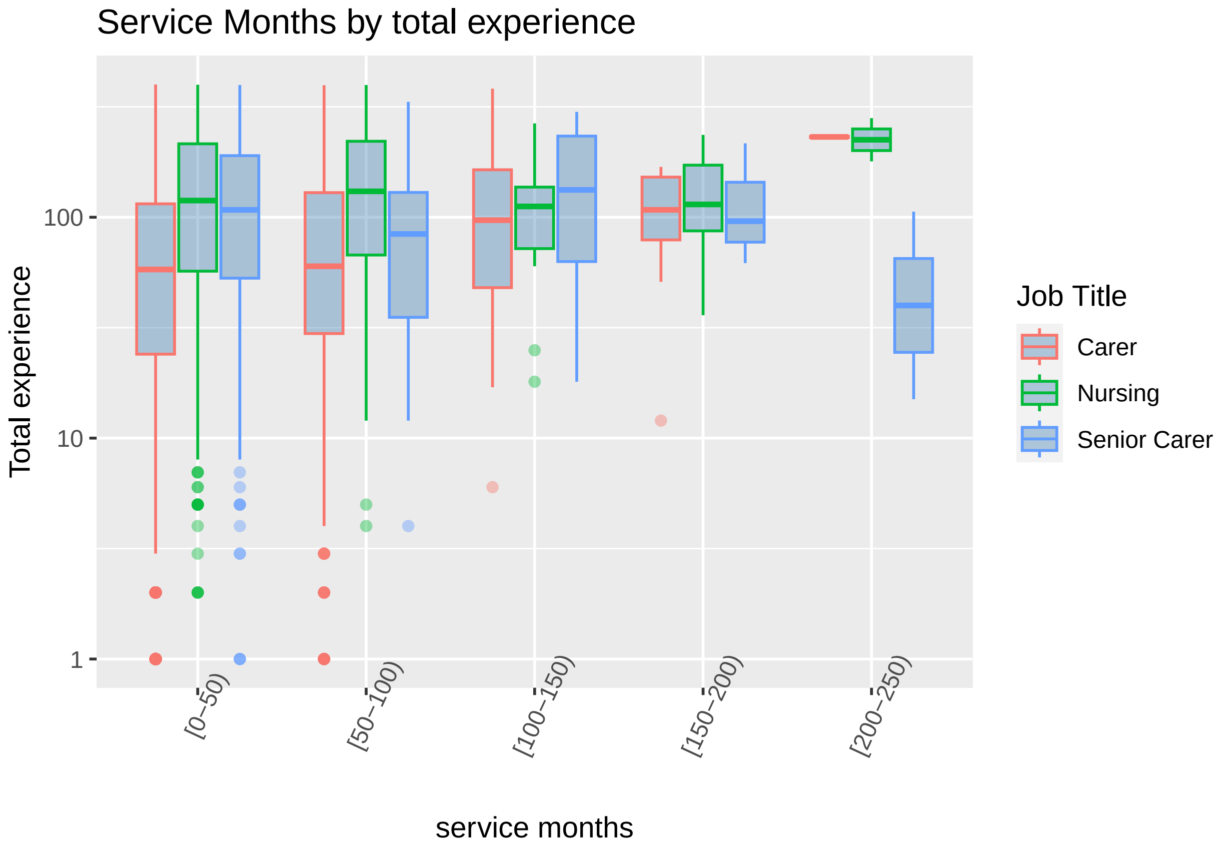


Figure 19: Boxplot showing the months of experience versus the months of service brackets, grouped by job title

With regards to employment history and experience, from Figure 19, employees in the nursing role have the overall highest median with over 20 years of experience and have served the longest since they fall under the [200-250] group of service months. Notably, employees in the nursing role have the highest median in terms of total experience in every group of months of service except for the [100-150] months of service where senior carers have the highest median.The statistical tests show that there is a significant difference in the total experience for the different job titles in the different groups, in relation to their months of service.

Acorn Demographics

Again, we would like to look at the retention of specific job roles by their Acorn Demographic Category, to see if we can uncover any insights about the relationship for a particular job.

Figure 20 shows the counts of employees in each demographic category broken down by job title. We see that for all three roles, Categories 4 (Financially Stretched) and 5 (Urban Adversity) make up the majority of employees. carers have 71.3% in this bracket, senior carers have 69.1% and nurses have 55.7%. This gap between nurses and the two carer roles continues in the upper-bands, with carers making up 10.8% in categories 1 (Affluent Achievers) and 2, senior carers making up 10%, and nurses making up 18.3%. This gap shows that nurses tend to live in locations with higher social status than both categories of carers, reflecting the higher status and pay that nurses achieve. The numbers for carers and senior carers show that a similar, but much less significant, gap exists between the two roles, likely a result of senior carers' greater experience and therefore higher pay than their carer colleagues.

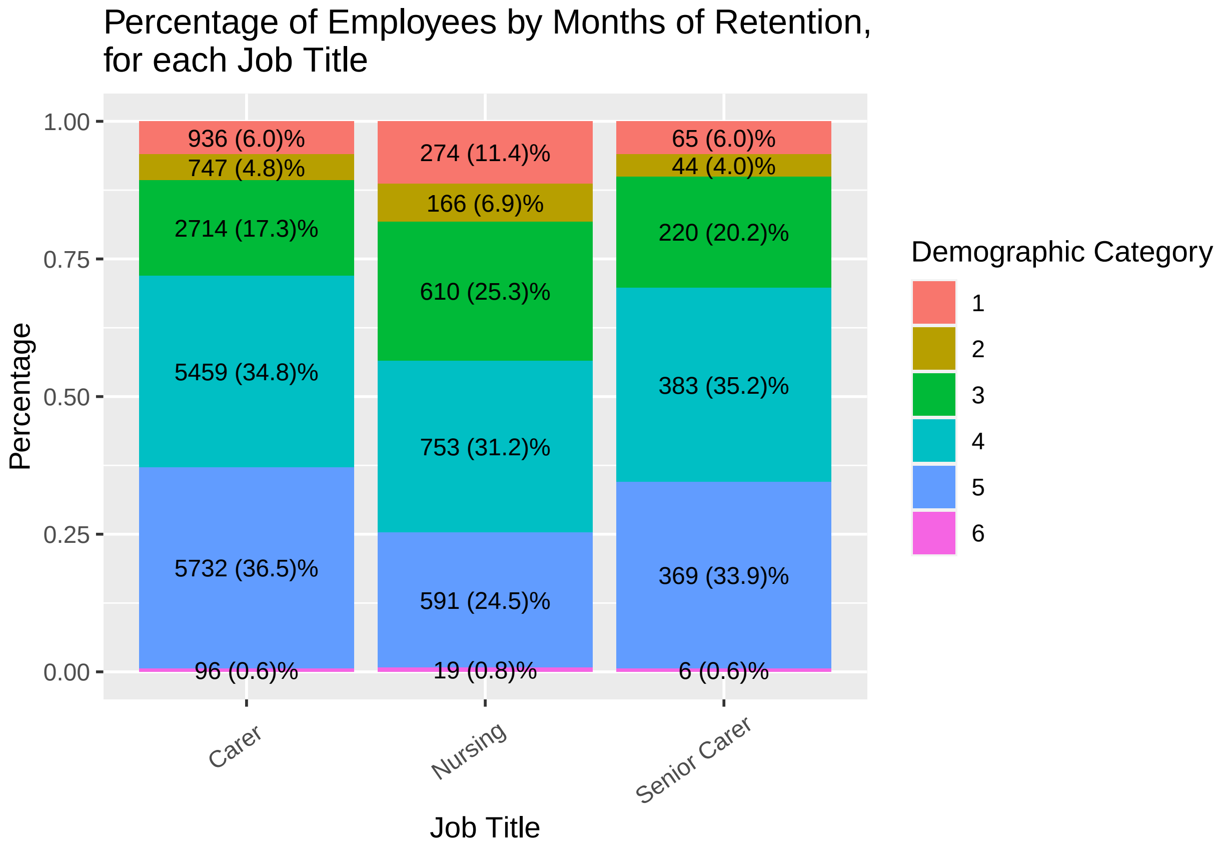


Figure 20: Stacked bar chart showing the percentage of employees by Acorn Demographic Category for each of the Job Titles

Figures 21, 22 and 23 show stacked bar charts for the brackets of employee retention, for each Acorn demographic category, for each of the three jobs. The only sizable visual difference between categories is found for category 6 (Not Private Households) for senior Carers and nurses likely due to their small number (6 senior carers and 19 nurses lived in households from this category). Excluding category 6, for carers, the largest difference in retention brackets is found for the >=24 months retained bracket, with 5.3% more carers from category 5 (Urban Adversity) than category 1 (Affluent Achievers). This is again matched for senior carers, with 11.5% more senior carers retained for >=24 months from Category 5 than Category 1. For nurses, the largest difference is found in the 6-12 month retention bracket with 7.1% more found in this bracket for category 3 (Comfortable Communities) than category 5 (Urban Adversity).

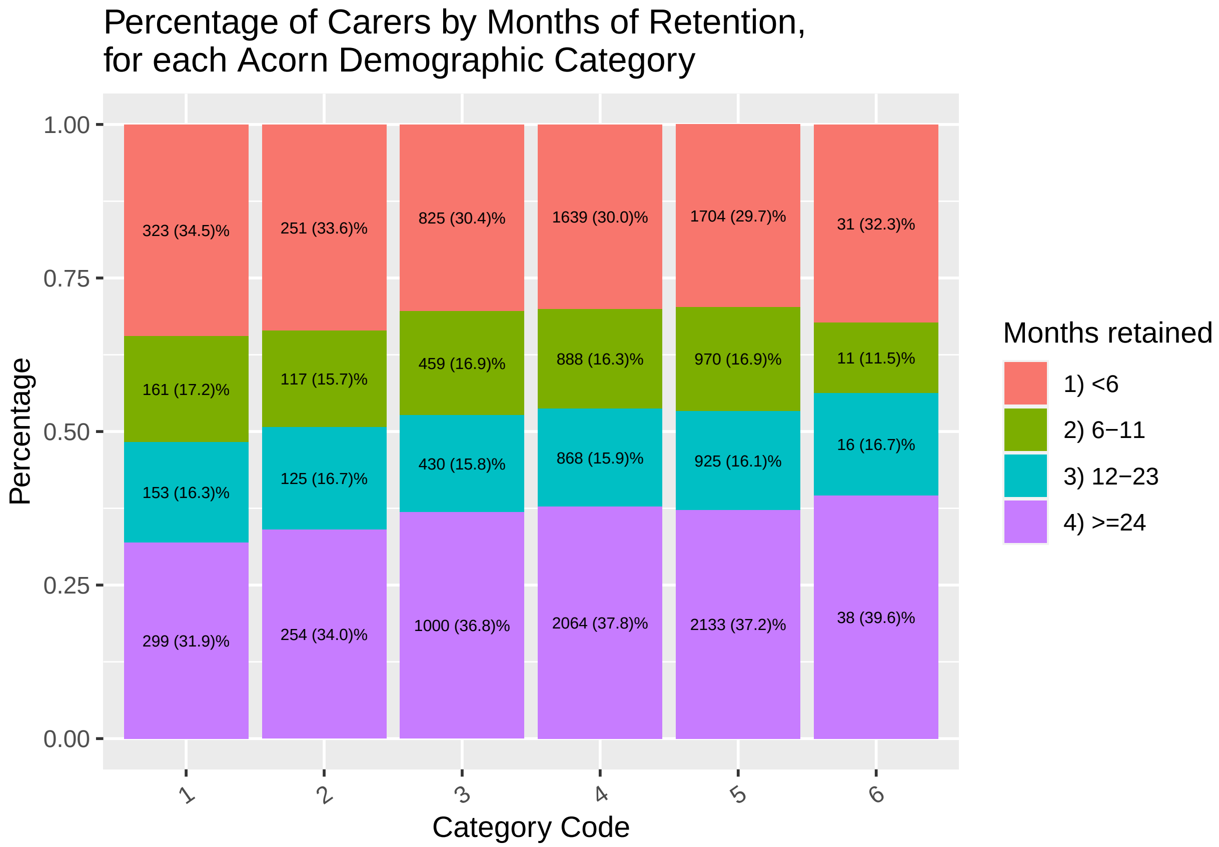


Figure 21: Stacked bar chart showing the percentage of carers by brackets of months of retention for each of the Acorn demographic groups

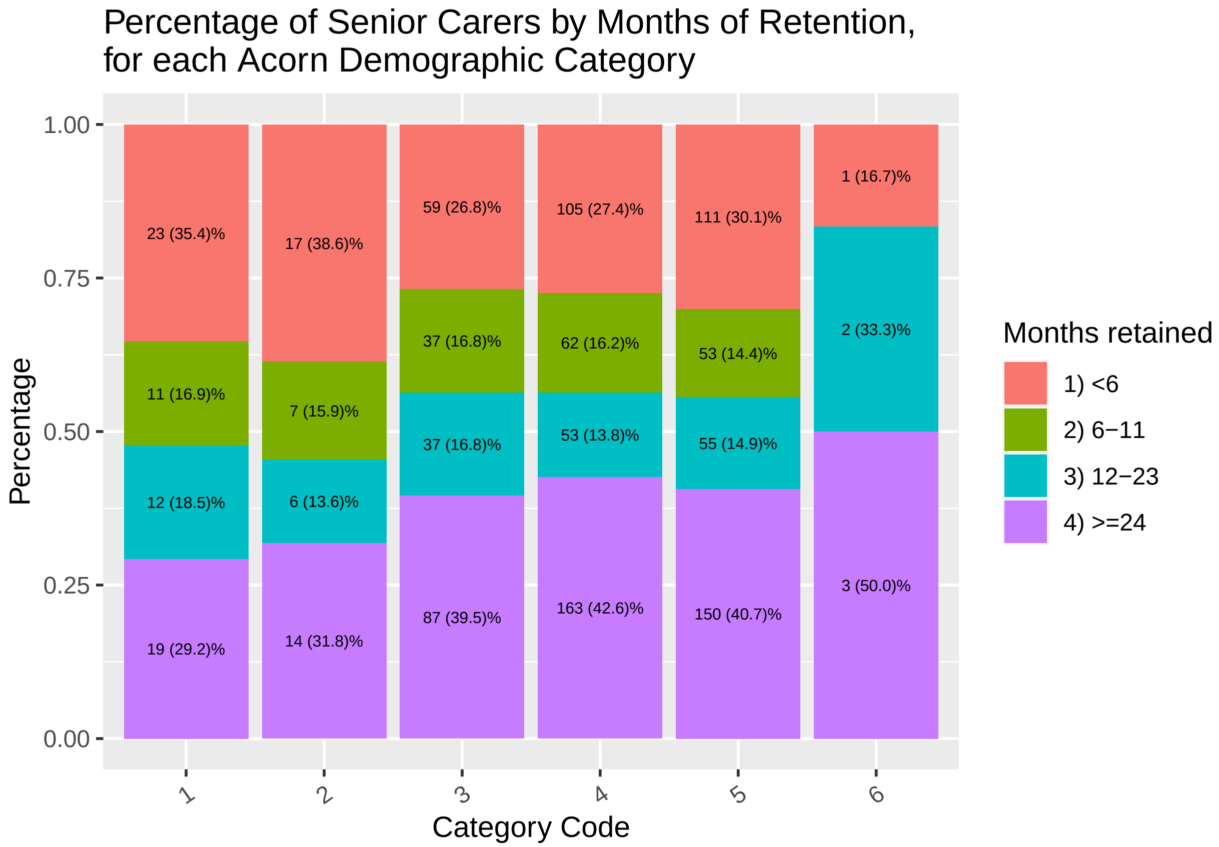


Figure 22: Stacked bar chart showing the percentage of senior carers by brackets of months of retention for each of the Acorn demographic groups

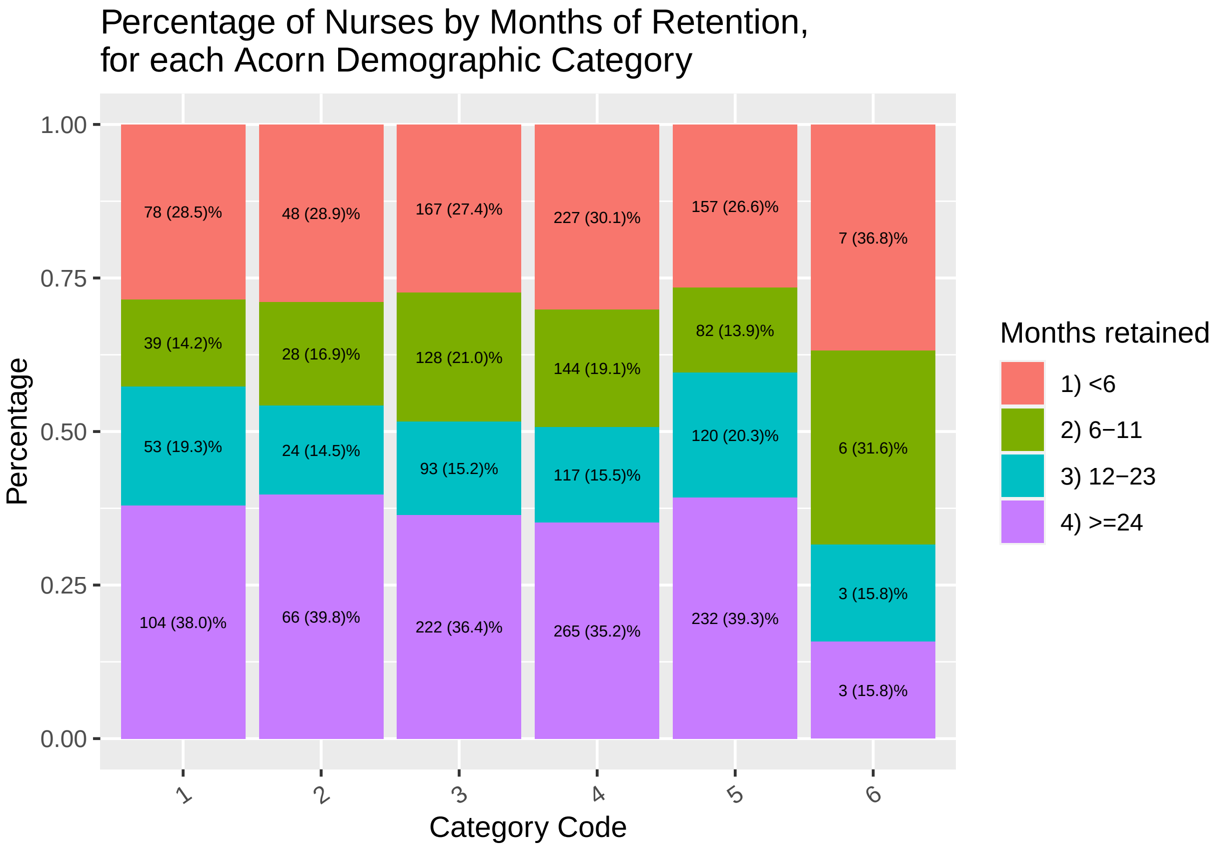
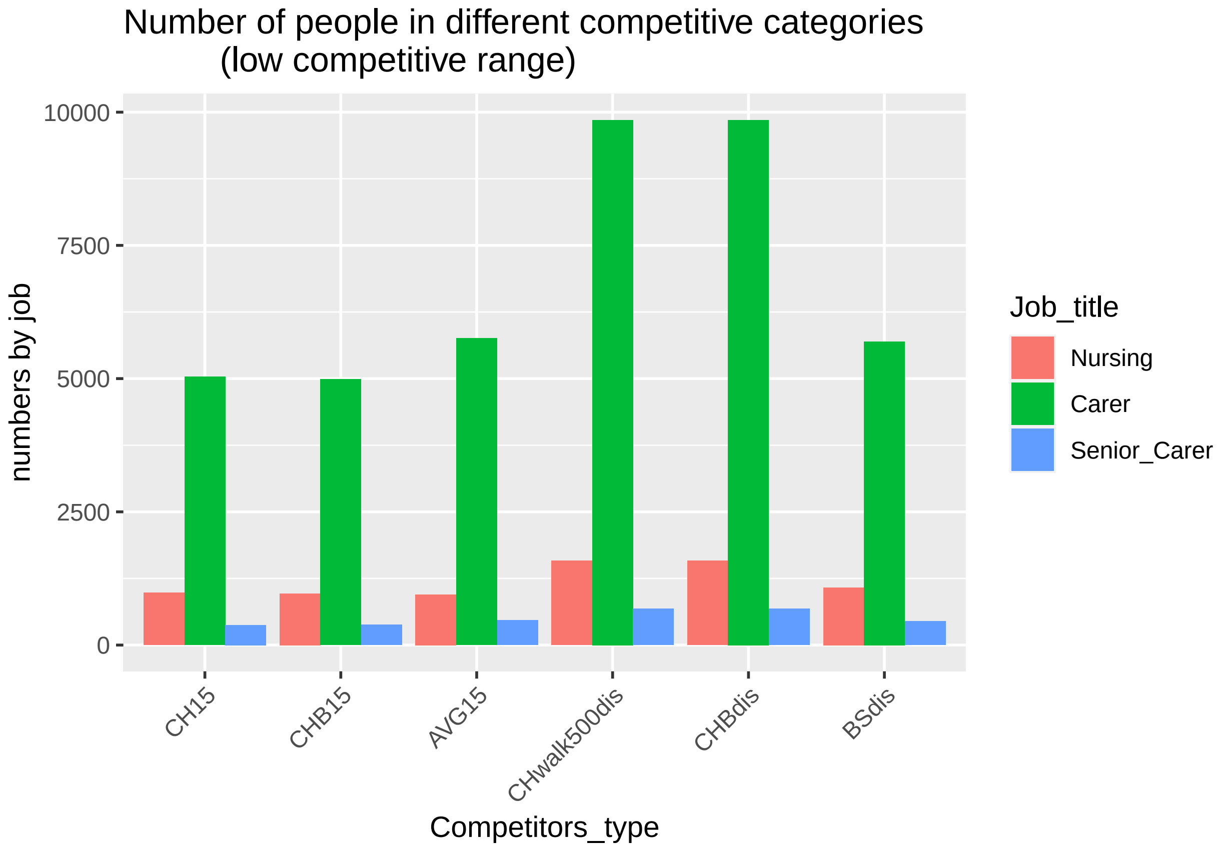


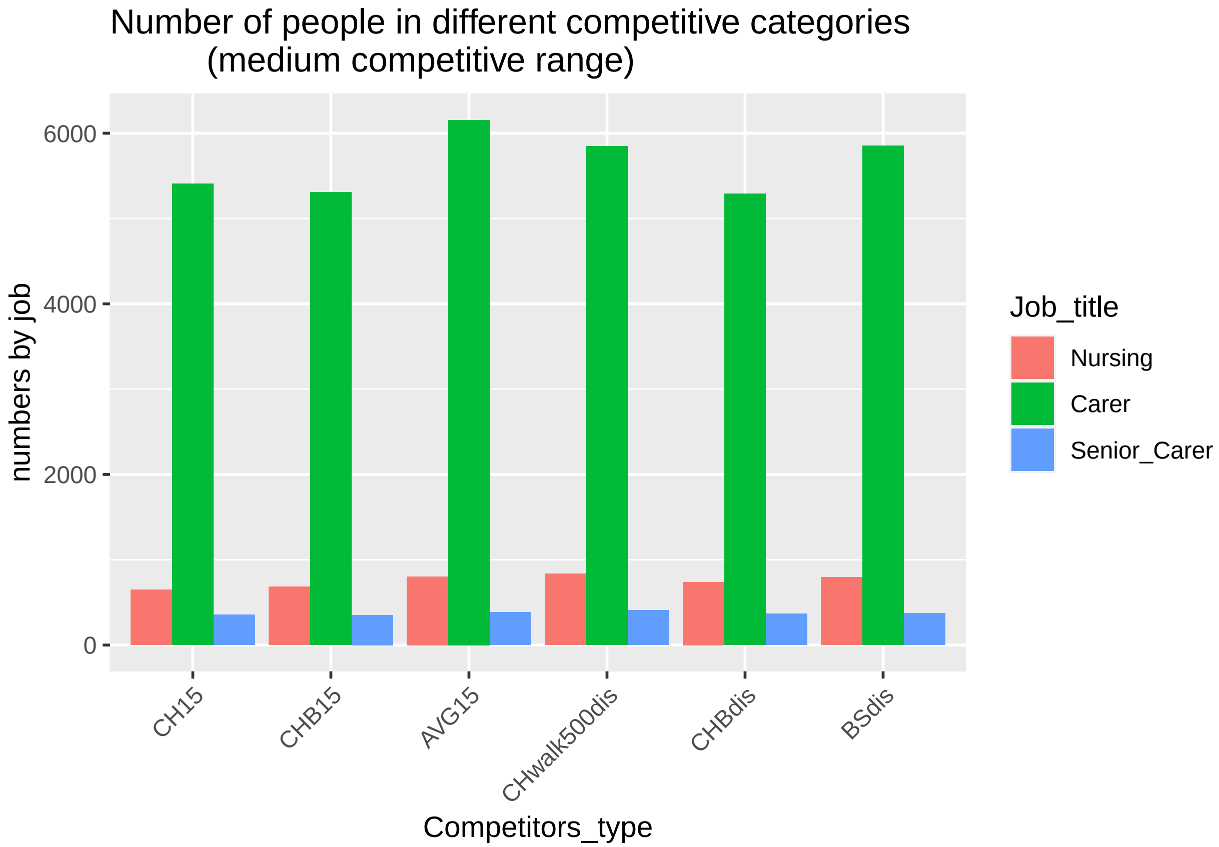
Figure 23: Stacked bar chart showing the percentage of nurses by brackets of months of retention for each of the Acorn demographic groups

To test for the statistical significance for between category code and months retained, we used a Kruskal-Wallis test. For carers, the Kruskal-Wallis rank sum test returns a p-value of 0.7366702. For senior carers, the p-value is 0.6306326. For nurses, the p-value is 0.3362162. All of these results are above 0.05, and thus mean none of the differences in retention between demographic categories are found to be statistically significant. We may therefore conclude that there is likely not a meaningful connection between an each employee's Acorn demographic category, broken down by the three job titles, and the different brackets of retention.

Competition in Candidates' Local Areas

The method used to analyse the effect of competition from other care homes around candidates' homes across the different jobs is to count the number of different occupations in different competition ranges in order to clarify the relationship between competition zones and occupations.


Number of people in different competitive categories
(low competitive range)
numbers by job
0
2500
5000
7500
10000
CH15
CHB15
AVG15
CHwalk500dis
CHBdis
BSdis
Competitors_type
Job_title
Nursing
Carer
Senior_Carer


Number of people in different competitive categories
(medium competitive range)
numbers by job
6000
4000
2000
0
CH15
CHB15
AVG15
CHwalk500dis
CHBdis
BSdis
Competitors_type
Job_title
Nursing
Carer
Senior_Carer

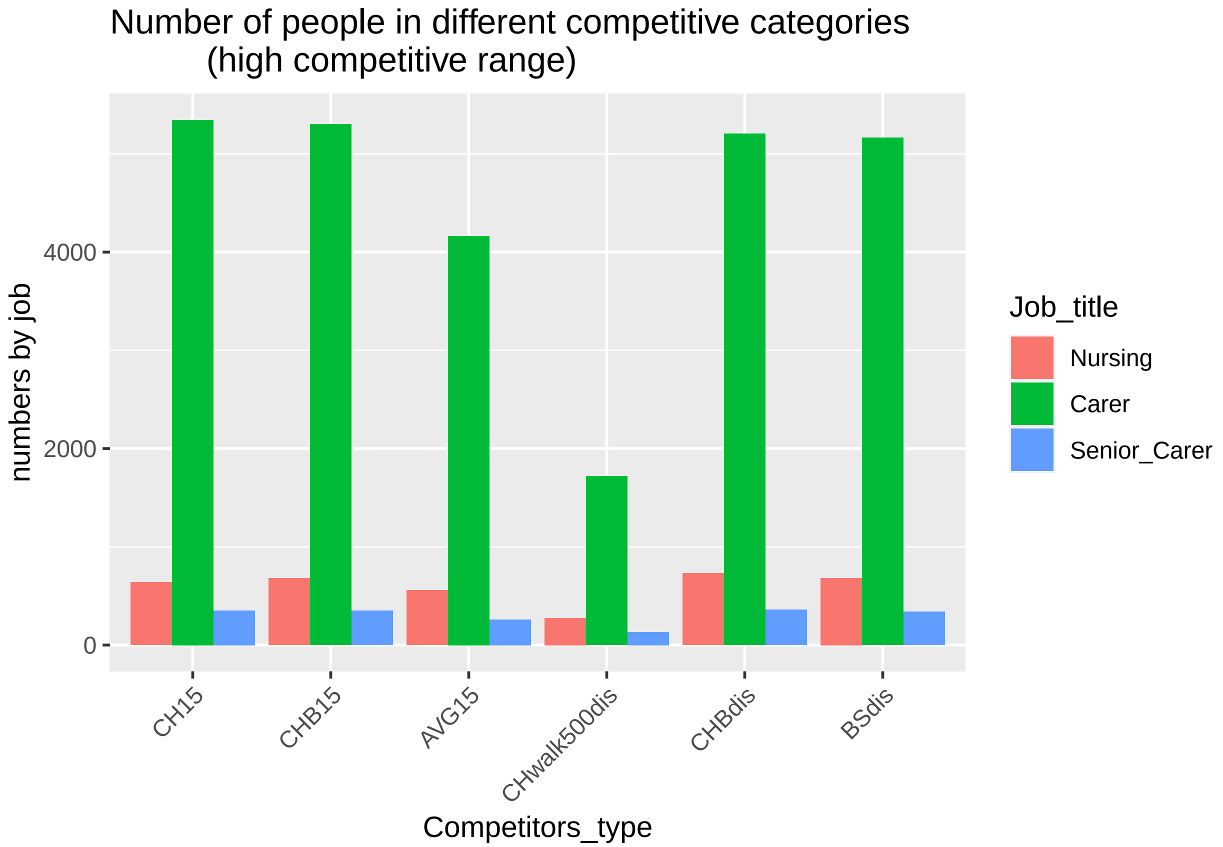
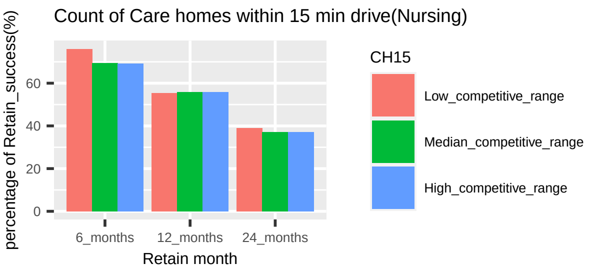
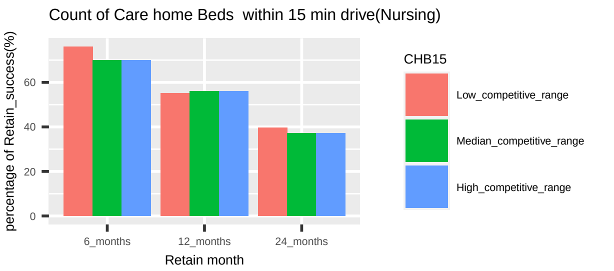
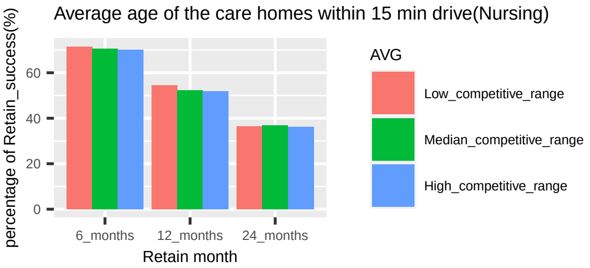


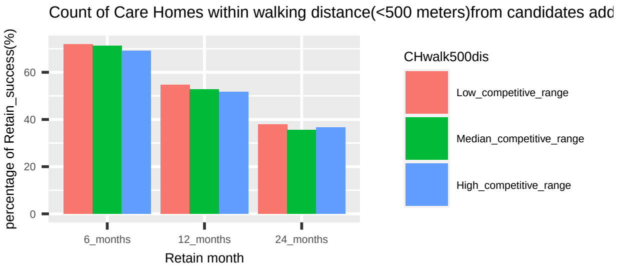
Figure 24: Bar charts showing the number of employees in each of the brackets of competition, grouped by job title

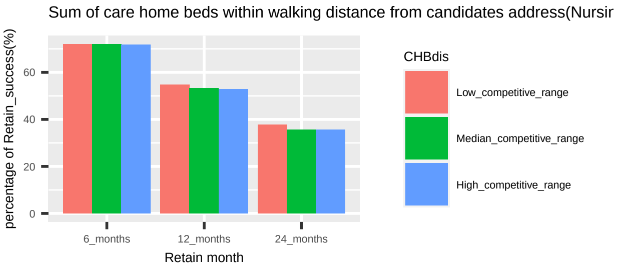
From Figure 24, it can be clearly seen that the job of carer has the most candidates overall, almost evenly distributed across the different competition ranges. For the low competition range, ‘chwalk500dis’ and ‘chbdis’ have the highest number of people, while in the high competition range this effect is reversed. What is also shown uniformly is that the demand in terms of competition for senior carer is the lowest Next, we will calculate the retention percentage for the three competitive ranges, high, medium and low, for the six different competition types by job title.

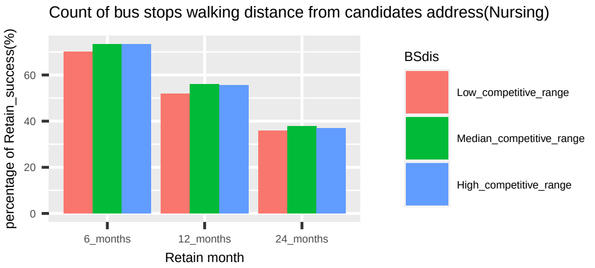

Count of Care homes within 15 min drive(Nursing)
percentage of Retain_success(%)
0
20
40
60
6_months
12_months
24_months
Retain month
CH15
Low_competitive_range
Median_competitive_range
High_competitive_range

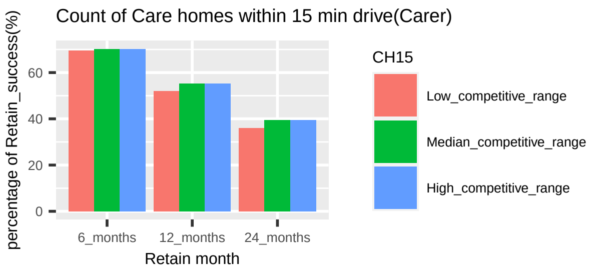

Count of Care home Beds within 15 min drive(Nursing)
percentage of Retain_success(%)
0
20
40
60
6_months
12_months
24_months
Retain month
CHB15
Low_competitive_range
Median_competitive_range
High_competitive_range

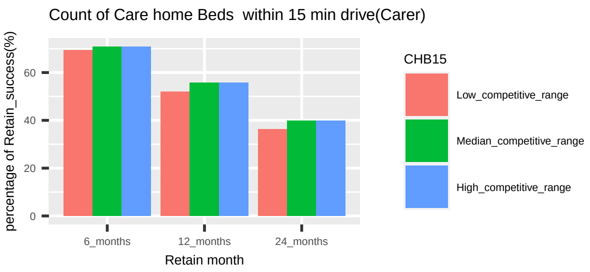

Average age of the care homes within 15 min drive(Nursing)
percentage of Retain_success(%)
0
20
40
60
6_months
12_months
24_months
Retain month
AVG
Low_competitive_range
Median_competitive_range
High_competitive_range

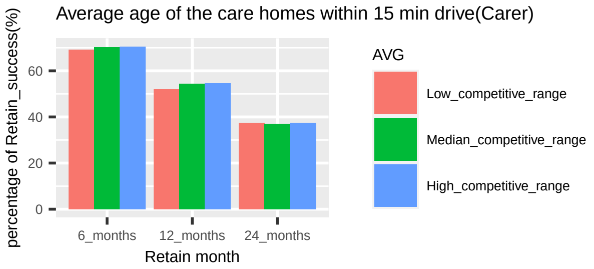

Count of Care Homes within walking distance(<500 meters)from candidates add
percentage of Retain_success(%)
0
20
40
60
6_months
12_months
24_months
Retain month
CHwalk500dis
Low_competitive_range
Median_competitive_range
High_competitive_range

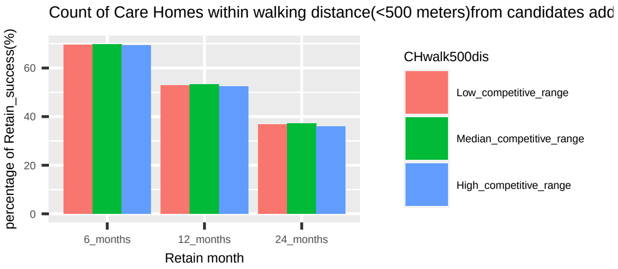

Sum of care home beds within walking distance from candidates address(Nursir
percentage of Retain_success(%)
0
20
40
60
6_months
12_months
24_months
Retain month
CHBdis
Low_competitive_range
Median_competitive_range
High_competitive_range

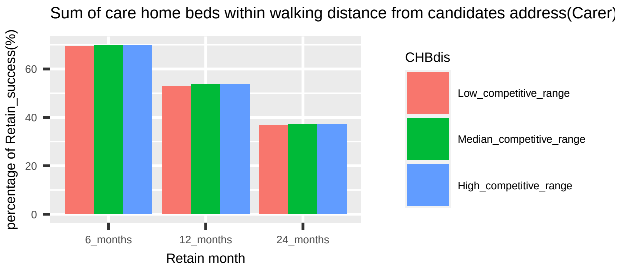

Count of bus stops walking distance from candidates address(Nursing)
percentage of Retain_success(%)
0
20
40
60
6_months
12_months
24_months
Retain month
BSdis
Low_competitive_range
Median_competitive_range
High_competitive_range

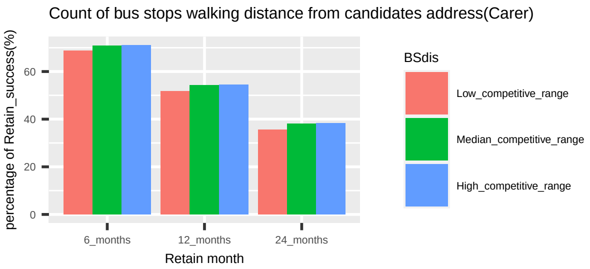

Count of Care homes within 15 min drive(Carer)


Count of Care home Beds within 15 min drive(Carer)


Average age of the care homes within 15 min drive(Carer)


Count of Care Homes within walking distance(<500 meters)from candidates add


Sum of care home beds within walking distance from candidates address(Carer)


Count of bus stops walking distance from candidates address(Carer)

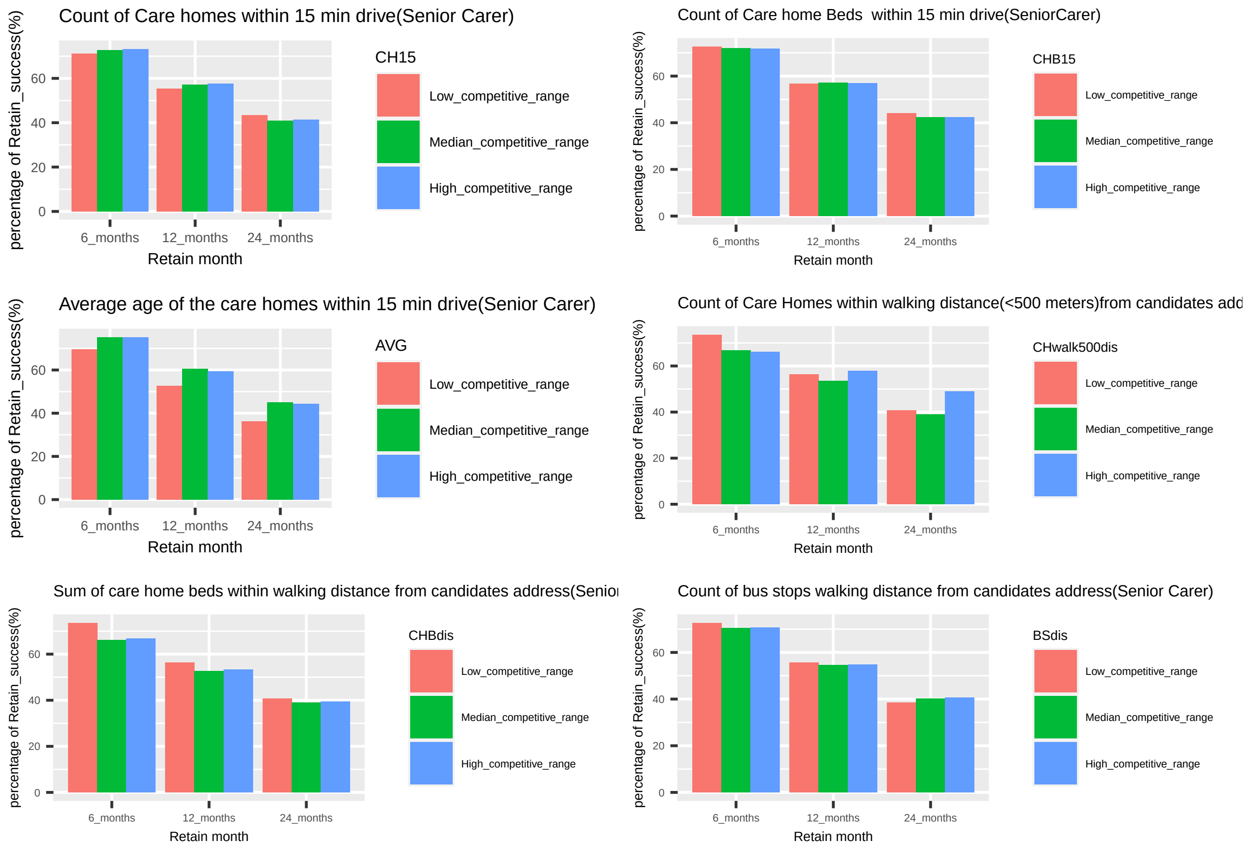


Figure 25: Bar charts showing the percentage of retention (for 6 months) in each of the brackets of competition versus retention brackets, grouped by job title

The three graphs in Figure 25 show the percentages of different competition ranges and retention times for each of the different job titles. In the plot for nurses, the proportions of the three competition ranges are similar, and it can be seen that the retention rate of the low competition range is relatively high in any retention time. Compared with carers, the number of candidates for nursing is very small. When screening candidates for this occupation, the six competition types show that the retention percentage decreases almost in proportion as their amount of time spent at the company increases. When referring to the admission of candidates, we can focus on the count of care homes within a 15 minute drive and the count of care home beds within a 15 minute drive, the retention ratio of this is the highest among the three competition intervals. And in respect to the count of bus stops within walking distance from a candidate's address, the retention rate of people in the low competition interval is low. We also observe a similar percentage for carers. In the three competition intervals, there is almost no obvious fluctuation in each data, and it still decreases with the increase of retain time, and the retention rate in the low competition interval is always less than or equal to the medium and high competition range. For the senior carer job, the sample size of the number of candidates is relatively low compared to the other two jobs, and the retention situation is similar to the first two careers.

Educational Background & Qualifications

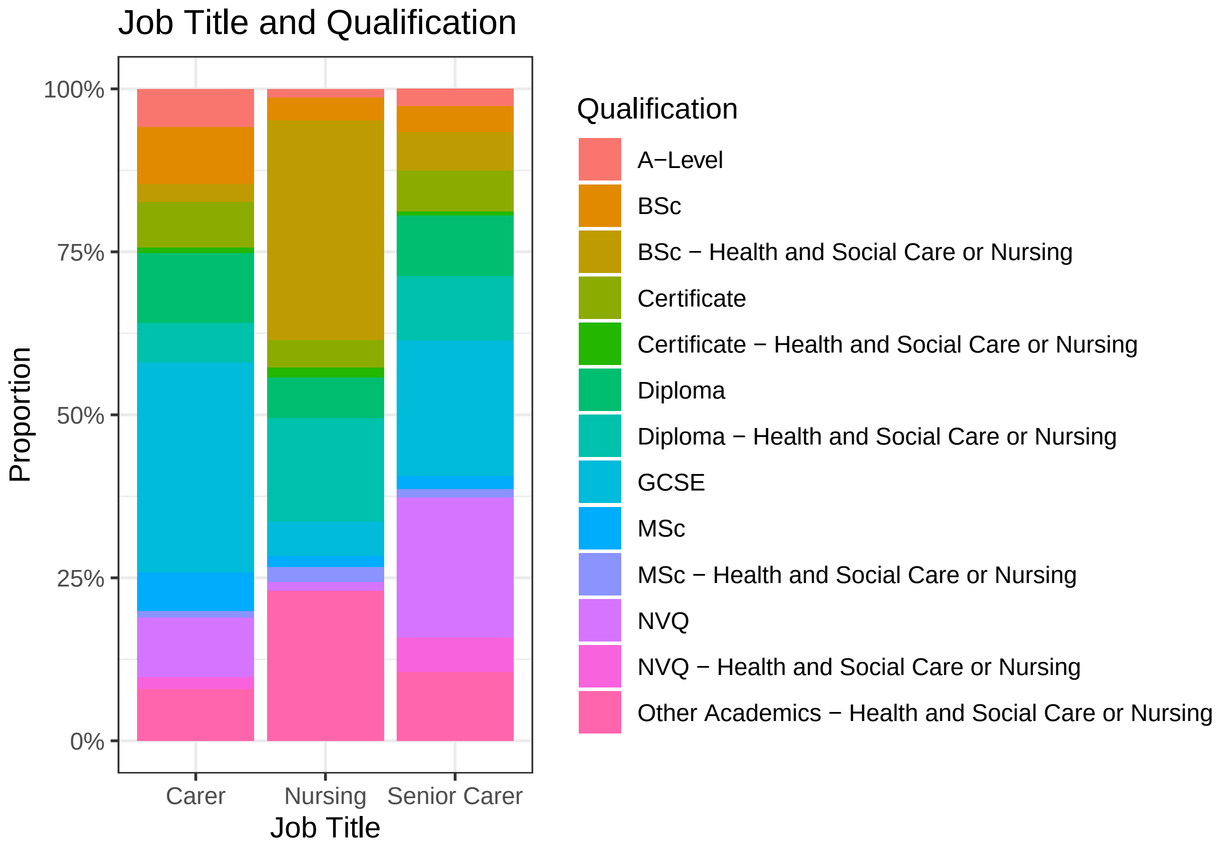


Figure 26: Bar chart showing the percentage distribution of employees versus qualifications, grouped by job title

The provided comparison bar graph in Figure 26 is an effective visualisation tool to represent the percentage distribution of employees retained for at least 6 months in HC-One based on their qualifications and job title. For instance, the graph shows that the highest percentage of retained employees are carers, followed by nurses and senior carers. The highest number of employees in this dataset work as carers, with the majority of these having GCSE qualifications. The minimum educational qualifications required and observed for carers and senior carers are similar, indicating that both roles require a similar level of fundamental knowledge and skills. However, it does not necessarily imply that the two roles have identical responsibilities or capabilities. The reason for the similarity in educational qualifications between carers and senior carers could be that senior carers have greater experience, making education less relevant compared to other factors in defining their abilities. However, there are also a significant number of carers who have a diploma in Health and Social Care or Nursing. Nurses, on the other hand, have higher qualifications overall, with the majority holding a BSc or a Diploma in Health and Social Care or Nursing. The nursing profession demands a higher level of education than the carer and senior carer roles. This suggests that the nursing profession places a greater emphasis on theoretical knowledge and technical skills than is required for carers and senior carers.

Advertised Salaries

After exploring the retention rate with the advertised salaries in cycle 1, we can further explore the retention rate based on the different job roles. There are three main job roles namely - ‘Carer’, ‘Senior Carer’ and ‘Nursing’. The histogram in Figure 27 shows the range of salaries that are advertised for nursing roles. The salary range lies between £10 to £21 with an average of £16.50 that is advertised for majority of the employees who have this job.

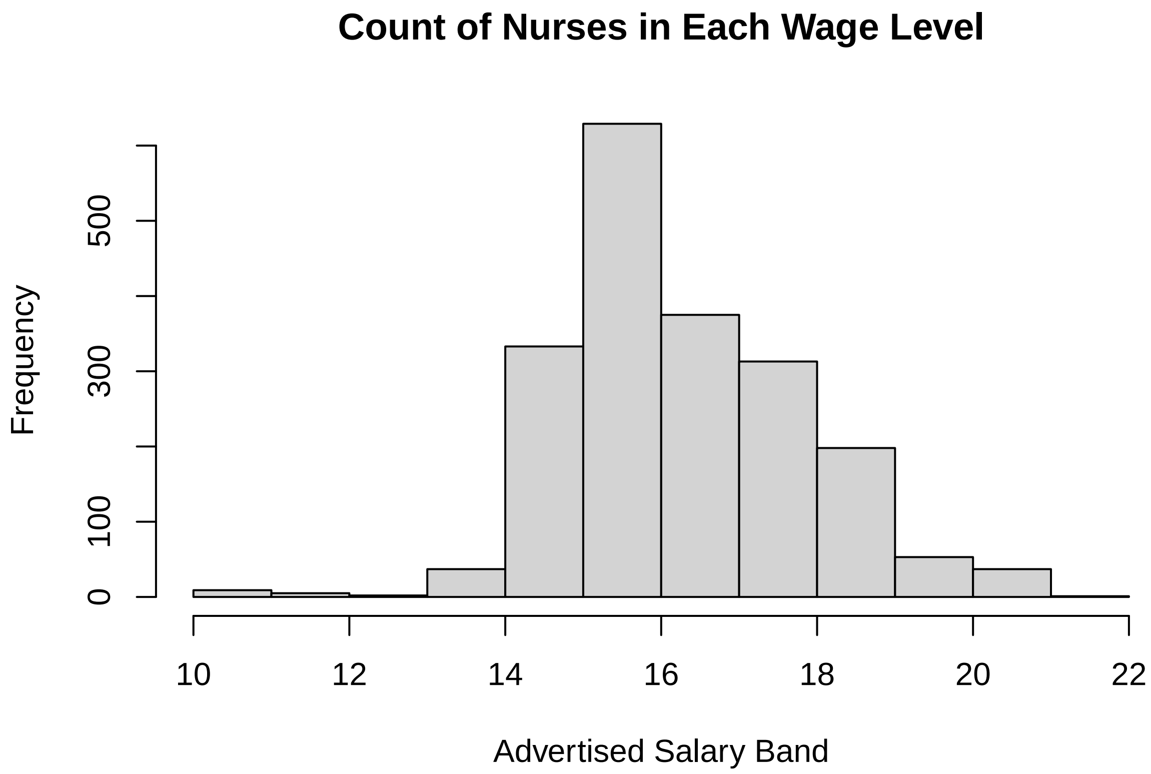


Figure 27: Histogram showing the range of salaries advertised for nursing roles by count of nurses

Figure 28 illustrates the retention period for the nurses based on the advertised salary in decreasing order. Here, it is evident that nurses with salaries in the range from £15 to £17 are retained more often for more than 24 months compared to those earning alternative salaries. The distribution is more uniform for the length of service for less than 6 months to 23 months. Overall, the correlation coefficient for this data did not indicate a strong association between these two variables for nurses.

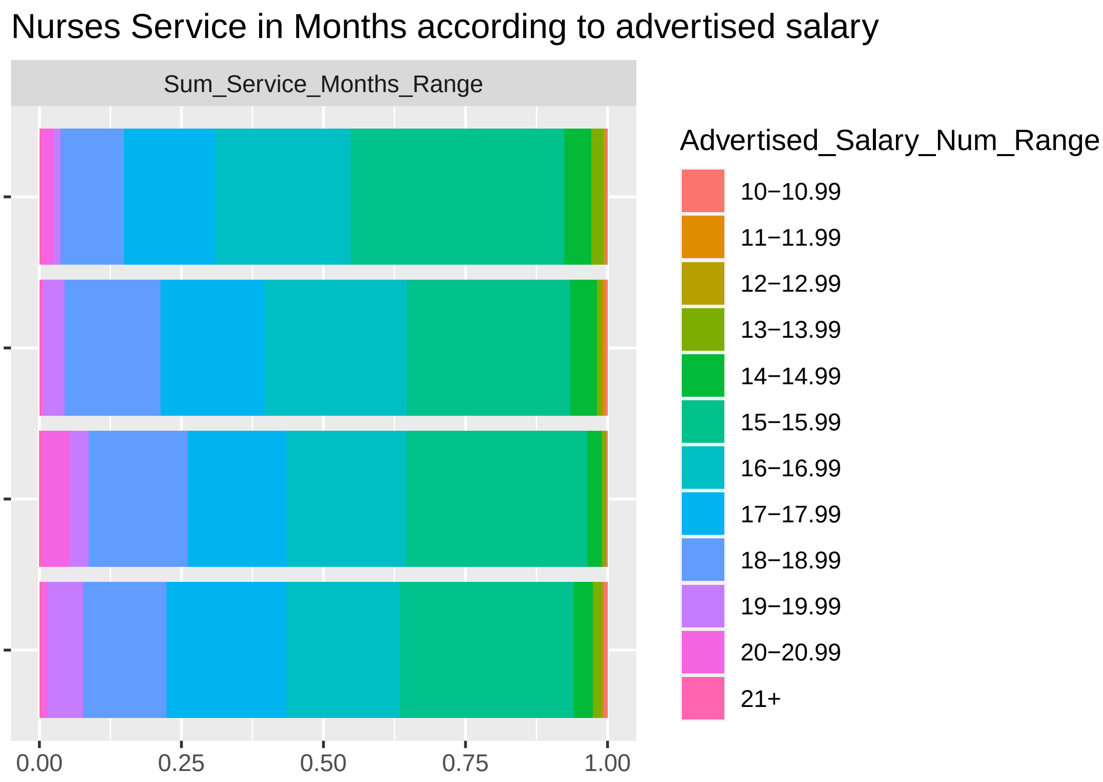


Figure 28: Bar chart showing the range of advertised salaries versus months of service for nurses

Figure 29 shows the range of advertised salaries for the carers job role. The salary ranges from £7 to £14 where the majority of the Carers have applied for £8 to £10. The average advertised salary for Carers is £8.8.

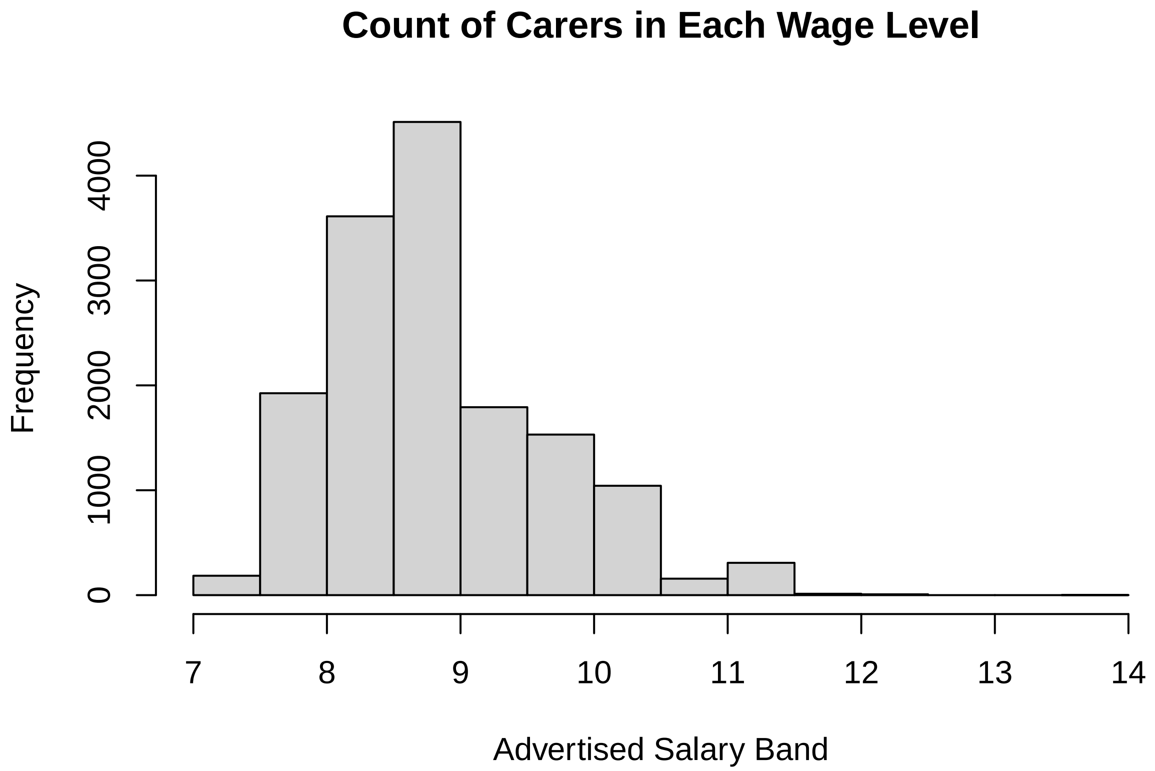


Figure 29: Histogram showing advertised salary and the count of carers

Figure 30 shows the retention rate of carers where it can be observed that the employees with lower salaries are retained for a longer period of time. The employees getting salaries in range £8 to £9 are more likely to retain till 24 months as compared to the people getting salaries in £9 to £10 band. The correlation for this job was observed to be stronger between these variables than for nurses or senior carers.

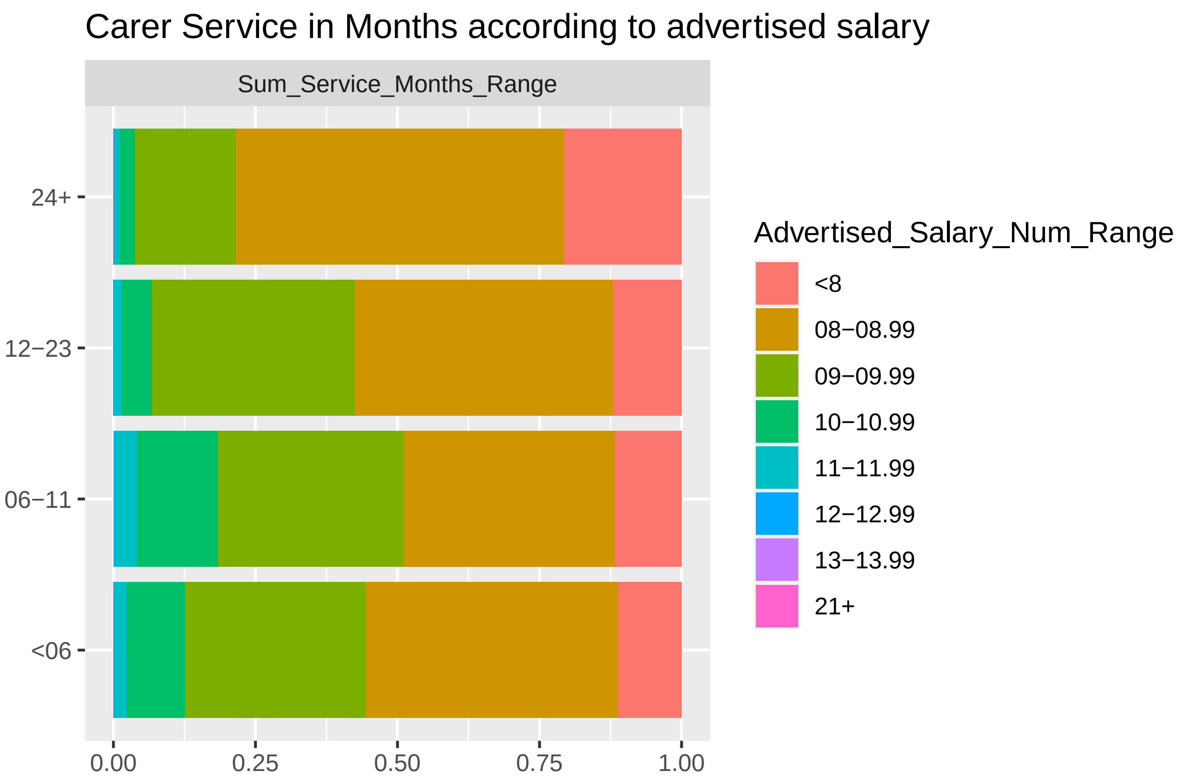


Figure 30: Bar chart showing advertised salary brackets versus the carer's month of service

Figure 31 illustrates the range of salaries advertised for the job 'Senior Carer' where advertised salaries range from £8 to £11 with a majority of employees applying for the range from £8.50 to £9.50. The average salary for the post of senior carer is £9.45.

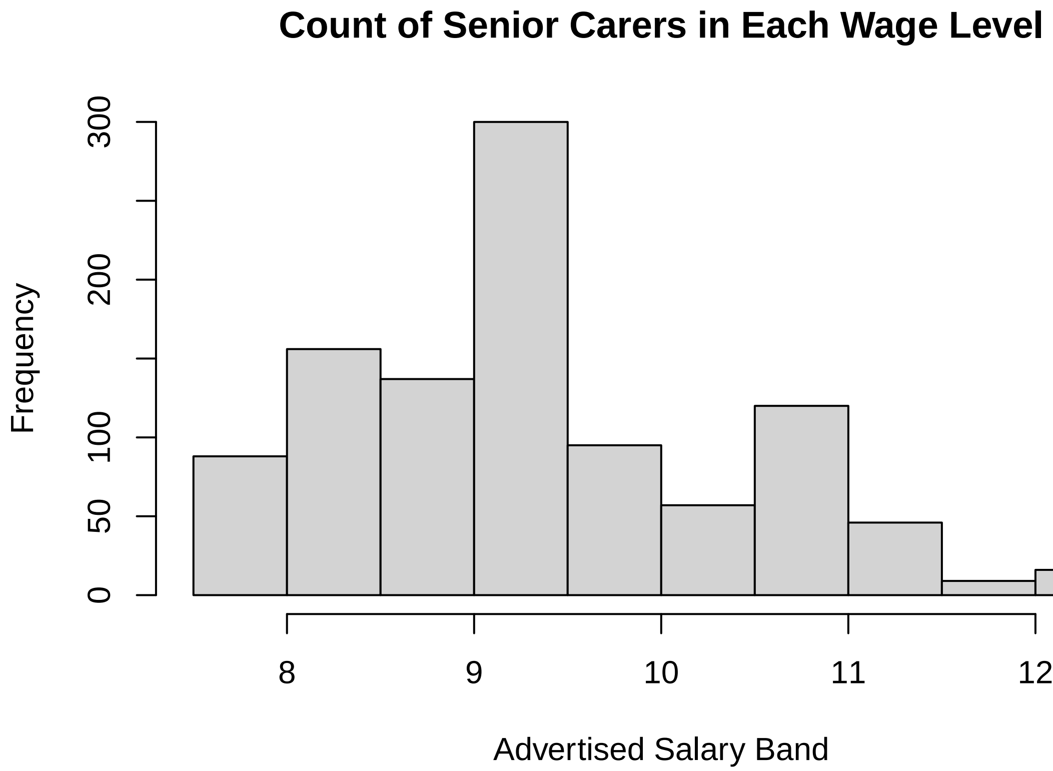


Figure 31: Histogram showing the advertised salary versus the number of Senior Carers

Figure 32 shows that the retention period of senior carers in the company is higher for the range £8 to £9 as more people tend to stay for more than 24 months.

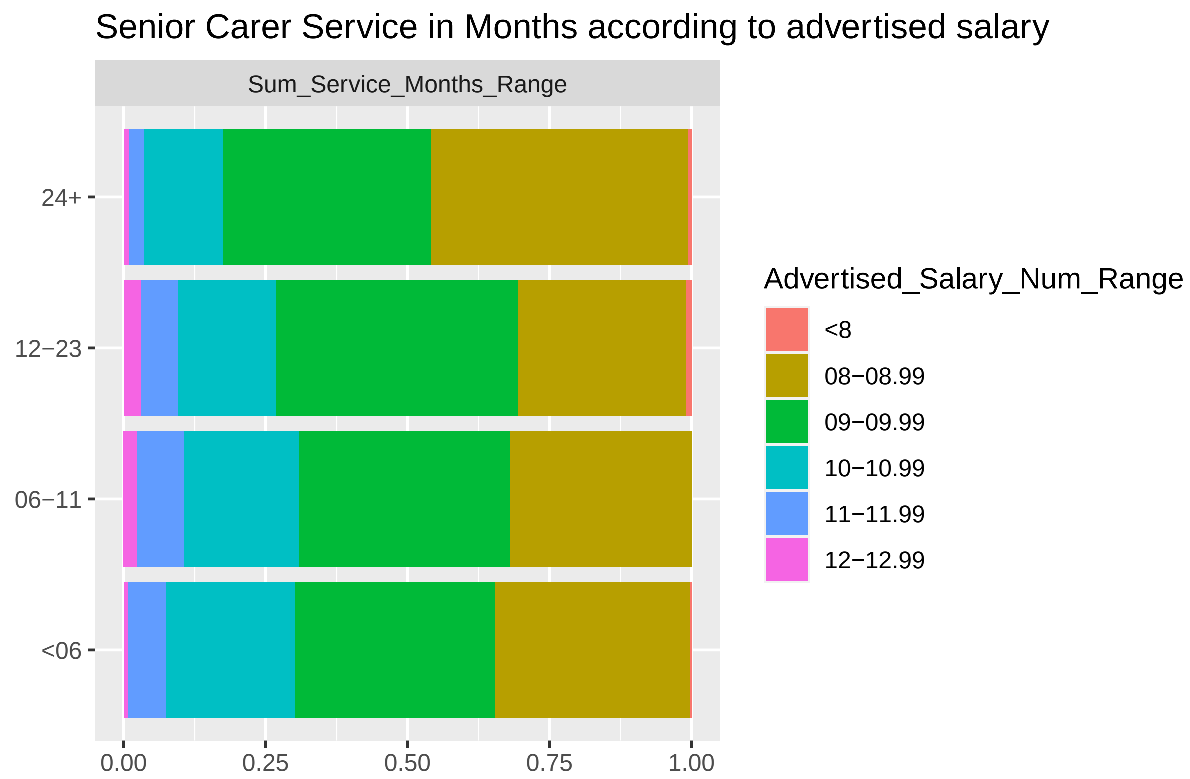


Figure 32: Bar chart showing the advertised salary brackets versus months of service, for Senior Carers

The range of salaries for carers and senior carers is almost the same whereas the nursing salaries are higher than the other two roles. The negative correlation between advertised salary and the length of service is strongest for carer and for nurses there is hardly any correlation, whereas the senior carer post lies somewhere in the middle.

Evaluation

When dividing the dataset into three subsets based on the job which each candidate was applying for, we were able to gather insights about the type of applicants that are more successful in each job, when controlling for factors that are exclusive to each given role. We observed that carers had the highest aggregate attention rate, followed by nurses and then senior carers. As applicants are only in direct competition with others applying for the same type of job, it is important to consider the factors that differentiate candidates for the same job from each other.

When considering the channel type through which an applicant applied, we discovered some separation in the channels that produced the most successful applicants for each job. Referrals was the most successful avenue for the dataset overall, but this mostly held for nurses, with a significantly higher median length of service than nurses who applied through other channels. This difference was not as large for the other two roles but was still a relatively successful portal for applicants. Job boards consistently seemed to have the lowest median length of service across the three roles, across a large sample. Social referrals performed very successfully in terms of job retention for senior carers, however this was across a very small sample. Overall, the statistical test for the effect of channel type showed that there is a significant difference in the length of service for the different job titles in each channel type.

For the Acorn demographic categories that we looked at, we were unable to discover any statistically significant difference in the lengths of service time for candidates from different types of catchment areas, for each of the jobs. This suggests that this would be unlikely to be a useful angle for the company to use in future to investigate candidates that they are considering for employment. The level of competition with regards to alternative care homes in the catchment areas of candidates showed that candidates in the middle zone of competition had the most stable retention rates, with retain percentage remaining constant with increasing service time. This zone saw slower drops in retain percentage than both the low zones and high zones of competition levels, suggesting that under consideration, the middle level of competition is the most stable for selection of candidates for long-term roles.

Nurses were seen to have higher levels of experience than carers and senior carers across almost all the different bands for months of service. The only bracket where this did not apply was for lengths of service between 100 and 150 months. In this case, senior carers with this length of service tended to have the most experience across the three jobs. For carers and senior carers, the requirements for education were similar, suggesting that similar types of specific experience and education was required for these roles than for nurses. A particular level of basic knowledge and skills is required for candidates to have a successful period working in these roles. With regards to nursing, most employees for this role had a BSc- or a Diploma-level qualification in health, social care or nursing. As a result, the nursing profession is implied to have a higher retention rate due to the higher level of education required for entry into this profession, compared to becoming a carer or senior carer.

The advertised salary for a position was originally observed for the dataset to have a slight negative correlation with the length of service for candidates. When we compared this phenomenon across the three jobs, we found that there was almost no correlation for nurses, who consistently earn higher than senior carers and carers. Carers saw the strongest association, with carers earning less tending to stay for a longer time at the company than their peers. The correlation for senior carers was close to the overall correlation for the dataset, suggesting a slight negative correlation. These statistics should be unaffected by the different sample sizes of the three roles as they were only compared with the same role.

Our next step for our analysis is to use the conclusions we have produced to better visualise the data. Currently, this report contains important insights for the company about the current situation of employees and factors influencing their retention, in addition to the possibility of examining future applicants based on this information. However, by creating a dashboard visualisation, we hope to be able to make our conclusions and the trends that we have discovered more easily interpretable for HC-One. The interactivity will add another useful aspect to our data by making it more accessible to the company for future use and enabling generation of custom insights.

PowerBI Dashboard

Power BI was utilised for creating a data visualisation of the HC-One job applicant data. The data included information on job titles, such as carer, nursing, and senior carer, as well as variables related to employee retention, including education level, advertised salary, length of employment, and competitors in the nearby locality. Additionally, we analysed the relationship between employee retention and the job portal from which the candidate applied, as well as Acorn demographics and candidate employment history.

The resulting dashboard can be used by the employer to gain more specific and customisable insights into factors that affect employee retention and to inform recruitment and retention strategies. By analysing the data, the employer can determine which job portals yield the highest retention rates and adjust recruitment efforts accordingly. Additionally, the employer can identify demographic trends among successful candidates and leverage this information to attract and retain similarly qualified individuals. Finally, the employer can use the dashboard to analyse the competitive landscape and make informed decisions about salary, benefits, and other incentives to maximise retention of high-performing employees.

Deployment

From the gathered insights during this project, we discovered some useful variables for HC-One's future hiring processes such as channel type, advertised salaries and educational background which with more data can be used to build a predictive model. This would help HC-One prioritise candidates early on in the application process depending on those variables that would likely make a good hire, including retention. This could be applied in HC-One by applying a weighting system for prioritising candidates, for example candidates that applied via a referral with the relevant educational background for application are processed first.

The interactive dashboard that we created can be integrated with a recruitment data pipeline that can be used by HC-One to gain further insights into their hiring process. Utilising an interactive dashboard, the employer can extract all dependent variables which affect employee retention in HC-One. This again can be used to prioritise candidates.

When using relevant results of data analysis to prioritise job candidates, HC-One must ensure that the process remains fair and ethical. It is important to consider factors beyond the identified variables, such as diversity, inclusion, and discrimination. The company must avoid making decisions based solely on a candidate's background, but instead, use data insights as a supplement to informed decision-making. Profiling candidates based on their background and information on their local area may be unfair to the candidate themselves. Additionally, it is crucial to maintain transparency throughout the recruitment process, informing candidates of the variables used to assess their suitability for the role. HC-One must strive to create a welcoming, inclusive workplace for all employees, regardless of their background or educational qualifications.

Bibliography

[1] R. Chakraborty, K. Mridha, R. N. Shaw and A. Ghosh, "Study and Prediction Analysis of the Employee Turnover using Machine Learning Approaches," 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON), Kuala Lumpur, Malaysia, 2021, pp. 1-6, doi: 10.1109/GUCON50781.2021.9573759.

[2] P. Gupta, S. F. Fernandes, and M. Jain, "Automation in Recruitment: A New Frontier," Journal of Information Technology Teaching Cases, vol. 8, no. 2. SAGE Publications, pp. 118-125, Nov. 2018. doi: 10.1057/s41266-018-0042-x.

[3] "Acorn - the smarter consumer classification | caci," CACI. [Online] [Accessed: 24-Mar-2023]