A&E Inpatient Clustering Using K-Prototypes for Length of Stay Analysis: A Comparative Study Before and After COVID-19

Lancaster University

Yi-Chun Huang
MSc Data Science

A dissertation submitted for the degree of
Master of Science in Data Science

Supervised by *Dr. Nicola Rennie*

School of Computing and Communications
Lancaster University

September, 2023

Declaration

I declare that the work presented in this dissertation is, to the best of my knowledge and belief, original and my own work. The material has not been submitted, either in whole or in part, for a degree at this, or any other university. Estimated word count is: **6715**

Name: **Yi-Chun Huang**
Date: **September, 2023**

A&E Inpatient Clustering Using K-Prototypes for Length of Stay Analysis: A Comparative Study Before and After COVID-19

Yi-Chun Huang, MSc Data Science.
School of Computing and Communications, Lancaster University
A dissertation submitted for the degree of *Master of Science* in Data Science.
September, 2023

Abstract

This comprehensive study explores evolving patterns in Accident and Emergency (A&E) inpatient demographics and healthcare utilisation within the Wrightington, Wigan and Leigh Teaching Hospitals NHS Foundation Trust (WWL). Utilising the K-Prototypes clustering algorithm to examine Length of Stay (LoS), significant trends affecting healthcare delivery were identified. Notably, an increase was observed in both average waiting times for hospital admission and LoS. Additionally, a stronger correlation between these variables was apparent post-pandemic. Intriguingly, patients tended to cluster around longer waiting times rather than high LoS, underscoring an emerging challenge in emergency department throughput. Post-pandemic demographic shifts indicated an increase in average age and comorbidity scores. Furthermore, a noticeable shift occurred in the dominant Healthcare Resource Group (HRG) Chapter from 'Respiratory' to 'Musculoskeletal' within the high-LoS cluster, suggesting potential long-term effects of COVID-19. Despite stable levels in the severity of initial patient assessments, a marked reallocation was observed towards 'Majors' within emergency departments, consequently reducing allocations to 'Minors'. These findings call into question the efficacy of existing acuity code systems and necessitate further investigations into healthcare resource management.

Acknowledgements

We extend our deepest gratitude to the healthcare professionals at the Wrightington, Wigan and Leigh Teaching Hospitals NHS Foundation Trust (WWL) for their unwavering commitment and exceptional service, especially during the formidable challenges posed by the COVID-19 pandemic. Their invaluable contributions have been pivotal to the success of this research.

Our profound thanks go to Thomas Ingram, the Principal Data Scientist at WWL, and Brian Wood, the Senior Data Scientist at WWL. Thomas has consistently guided us, helping us maintain focus on our research objectives and facilitating the development of our skills in data science. Brian has been crucial in equipping us with timely knowledge and essential tools to advance our research. Their collective expertise and mentorship have not only elevated the quality of this study but have also prepared us for forthcoming academic pursuits.

Special acknowledgement is given to Dr. Nicola Rennie for her invaluable academic guidance. Her expert counsel and constructive feedback have been instrumental in elevating this research to high academic standards. We remain deeply appreciative of her support and confidence in our abilities.

Finally, we offer our heartfelt gratitude to our friends and family for their steadfast support and love, which have consistently inspired us to push forward.

To all who have participated in this challenging yet rewarding journey, your contributions have been invaluable. For this, we are deeply grateful.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Company Background

Wrightington, Wigan and Leigh Teaching Hospitals NHS Foundation Trust (WWL) is a medium-sized emergency and community foundation trust located in the north-western region of England, specifically in Greater Manchester. Committed to serving over 300,000 local residents, WWL also extends its expertise to a broader regional, national, and international area. The Trust is comprised of multiple specialised medical centres, each with a focus on distinct healthcare domains. Notable facilities include the Royal Albert Edward Infirmary (Wigan Infirmary), offering A&E, General Surgery, General Medicine, and Maternity services; Leigh Infirmary and the Hanover Treatment Centre, specialising in Elderly Medicine and Outpatient Services; Wrightington Hospital, focusing on Orthopaedic Surgery and Rheumatology; the Thomas Linacre Centre, for comprehensive Outpatient Services; and the WWL Eye Unit, dedicated to Ophthalmology and Orthoptic Outpatient Services.

The foundation of WWL aligns with the broader context of the UK's National Health Service (NHS), acknowledged as the world's largest publicly funded healthcare system (Tennison et al., 2021). Operating within this framework, NHS Foundation Trusts exemplify a distinctive model of healthcare entities. These trusts maintain a partially autonomous status while remaining integral components of the NHS. This distinctive setup empowers them with a significant degree of financial and operational independence. Differing from conventional NHS hospitals, NHS Foundation Trusts possess their governance boards and exercise increased control over their financial planning (Klein, 2004). This autonomy equips them to make decisions tailored to specific circumstances effectively.

1.2 Research Background

Recent research from WWL reveals a significant shift in the demographic profiles of patients admitted to WWL-affiliated hospitals since 2016. These patients are generally older and have longer hospital stays compared to previous years. Additionally, a rise in Length of Stay (LoS) is observed across all age groups when compared to the pre-COVID-19 era. Extended hospital stays are correlated with worsening patient outcomes (Hassan et al., 2010) and also limit overall bed availability, thus exacerbating the demand for hospital beds.

As depicted in Figure 1.1, the shaded region highlights the timeframe where the effects of the COVID-19 pandemic were most acute. For the purposes of this study, we've marked the onset of the pandemic with the initiation of the UK's first national lockdown. Conversely, the end of the pandemic phase, for our analysis, is marked by the point at which mortality rates returned to their pre-pandemic norms (UK Government, 2023). The segments on either side of this shaded area give us insights into the pre-pandemic and post-pandemic durations. A striking observation from this visualization is the notable uptick in both average LoS and total LoS in the aftermath of the pandemic.

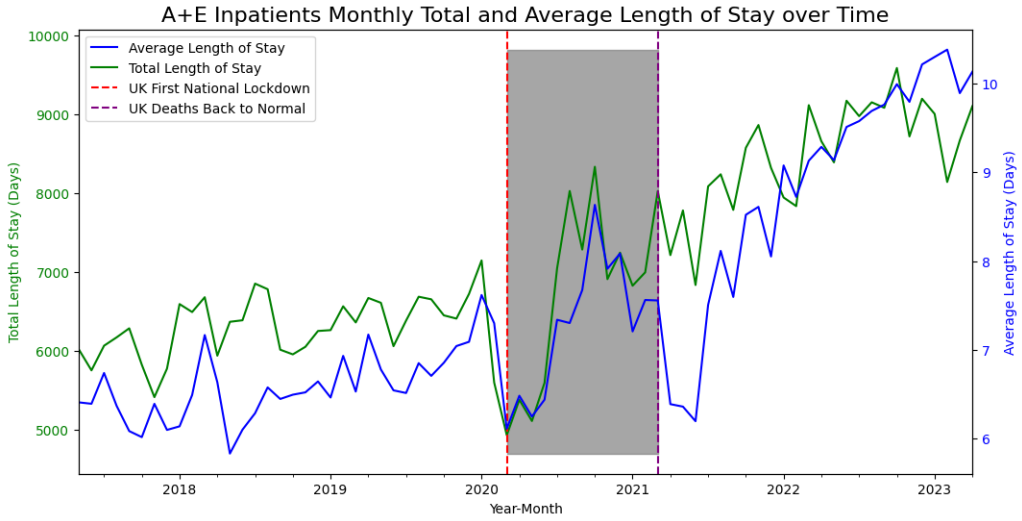

Figure 1.1: Monthly total and average Length of Stay for A&E inpatients over time

1.3 Aims and Objectives

The primary aim of this paper is to examine factors affecting changes in patient Length of Stay (LoS) at our hospital. This research specifically targets the periods before and after the COVID-19 pandemic and outlines the following objectives:

- **Analysis of Patient Characteristics and Demographics:** Initially, this study aims

to perform a thorough analysis of patient characteristics and demographics to identify trends and variations in length-of-stay patterns over time. This investigation enables us to understand the evolution of patient demographics and their potential association with changes in Length of Stay.

- **Evaluating the COVID-19 Pandemic's Impact:** The second objective is to assess the unique impact of the COVID-19 pandemic on patient Length of Stay. By comparing pre- and post-pandemic LoS data, we aim to identify changes brought about by the pandemic's specific conditions.
- **Advanced Clustering for Multifaceted Patient Categorisation:** To offer a comprehensive perspective on hospital admissions, we employ advanced clustering methods that incorporate multiple variables, including patient demographics and key performance indicators. This multi-dimensional approach aims to uncover nuanced patient profiles, thus providing a more refined understanding of factors affecting Length of Stay.
- **Optimising Operational Strategies:** Insights derived from this research aim to inform the development of strategic interventions aimed at optimising bed capacity and improving patient outcomes, particularly in a post-pandemic environment.

Through longitudinal studies spanning pre- and post-pandemic periods, this research aspires to furnish invaluable insights into the dynamics of patient Length of Stay, thereby informing and enhancing healthcare management strategies in an ever-evolving healthcare landscape.

Chapter 2

Related Work

2.1 Factors Affecting Length of Stay

In recent years, the impact of Length of Stay (LoS) on resource allocation, patient outcomes, and healthcare costs has garnered increased attention, leading to a heightened focus on determinants of LoS (McDermott and Stock, 2007). According to (Buttigieg, Abela, and Pace, 2018), critical factors affecting LoS include patient demographics, medical severity, and hospital-specific variables. Research by (Bahrmann et al., 2019) and (Daghistani et al., 2019) emphasizes the role of patient age and comorbidities, indicating that elderly patients with multiple chronic conditions are likely to experience extended hospital stays. Furthermore, (Higgins et al., 2003) highlights that the initial severity of illness upon admission also plays a crucial role, with patients admitted via the Accident & Emergency (A&E) department generally experiencing longer stays.

However, LoS is not solely influenced by patient-related factors; hospital characteristics are also impactful. Research by (Freitas et al., 2012) demonstrates a strong correlation between hospital type and extended LoS, suggesting that teaching hospitals with over 1,000 beds typically report longer stays. (Moisoglou et al., 2019) argues that greater nursing experience and a higher ratio of registered nurses can significantly reduce patient LoS. Additionally, (Salway et al., 2017) posits that overcrowding in A&E departments exacerbates issues detrimental to hospitals, such as longer waiting times, increased LoS, medical errors, and elevated patient mortality rates.

2.2 Medical Applications of Clustering

The application of clustering techniques in evaluating patient LoS has emerged as a groundbreaking approach in healthcare data analytics, providing valuable insights for hospital management and resource allocation (Vranas et al., 2017). Clustering allows healthcare

practitioners to differentiate between various patient groups, facilitating a more nuanced understanding of each category. This targeted approach aids in summarising specific traits and clinical test results for each patient cluster, thereby not only enhancing precise medical treatments but also reducing psychological and financial burdens on patients, shortening the duration of medical care, and ultimately improving the quality of life (Gong et al., 2022) (Chudasama, Khunti, and Davies, 2021).

Earlier studies primarily used conventional statistical methods for LoS analysis, including parametric and non-parametric tests (Qualls, Pallin, and Schuur, 2010), as well as multiple regression analyses (Liu, Phillips, and Codde, 2001). The advent of machine learning, however, has given rise to more advanced and dynamic models, particularly clustering algorithms like K-means and density-based clustering. For example, a study by (El-Darzi et al., 2009) employed K-means clustering to segment patients based on LoS intervals, identifying distinct patient groups with markedly different LoS metrics. Another study by (Panchami and Radhika, 2014) used the density-based clustering method DBSCAN to develop classification training sets. Subsequent research by (Ogbuabor and Ugwoke, 2018) showed that K-means generally outperforms DBSCAN in terms of clustering accuracy and computational efficiency. In the current landscape of smart healthcare, machine learning algorithms have become invaluable tools with significant predictive utility (Iwase et al., 2022).

Chapter 3

Data

3.1 Data Collection

WWL's data are securely stored in a Qlik Sense environment, utilizing a proprietary data file format denoted as .qvd. These data are accessible through SQL-like syntax for relevant information retrieval. Following rigorous filtering, four datasets were earmarked for further analysis:

1. *Inpatient Spells*: This dataset, consisting of 281 columns, incorporates a wide array of both medical and administrative data relevant to inpatient spells. It offers detailed insights into patient care, treatment outcomes, and other aspects.

2. *Pathway*: Introduced in 2021 and managed by the Integrated Discharge Team (IDT), this dataset aims to monitor patients' post-discharge status.

3. *Patient Info*: Comprising 108 columns, this dataset provides comprehensive demographic data, thereby offering a holistic understanding of the patient population within the healthcare system.

4. *A&E Flow*: This dataset contains 118 columns and records patient visits to and interactions within the Accident & Emergency (A&E) department, thus presenting a complete picture of patient flow.

3.2 Data Integration & Data Creation

The *Inpatient Spells* dataset acted as the primary table, while the remaining datasets were amalgamated using unique patient IDs and matching dates as primary keys. Subsequently, the following columns were appended:

1. 'COVID Period': As outlined in Figure 1.1, the time span from March 23, 2020, to March 23, 2021, was labelled as 'during_covid'. Periods before and after this interval were designated as 'pre_covid' and 'post_covid', respectively.

2. 'Arrival to Admission (Hrs)': Aligned with the findings of (Mentzoni, Bogstrand, and Faiz, 2019), which suggest a link between emergency department overcrowding and extended LoS, this column quantifies the time each patient spent in the emergency department, in hours.

3.3 Data Pre-processing

Data preparation commenced with a multi-step procedure aimed at ensuring both data quality and completeness. Initially, columns where a single value constituted more than 95% of entries were carefully removed, thereby eliminating columns with minimal variability and analytical utility.

Subsequently, columns with a high degree of unique values—comprising more than 1% of the dataset—were excluded to reduce complexity and noise in later analyses.

Moreover, columns containing more than 20% missing data were selectively omitted due to their limited informative value and the challenges tied to data imputation.

Finally, for columns with remaining missing entries—most of which were categorical—rows with these gaps were systematically eliminated to maintain dataset integrity, ensuring a consistent and reliable basis for subsequent analyses.

3.4 Redefining Data Type & Feature Selection

In light of the variety of data storage methods in the database, meticulous scrutiny was applied to data type validation to secure the integrity of future research endeavors. This stage required a thorough review of data types to confirm their appropriate interpretation by the model.

For this purpose, a nuanced feature selection strategy was deployed, incorporating two distinct analytical techniques: correlation analysis and Analysis of Variance (ANOVA). The former focused on retaining features with absolute correlation coefficients above the 0.1 threshold in relation to LoS, while the latter employed ANOVA to assess the impact of categorical features on LoS.

Table 3.1 summarizes the chosen features, encapsulating the collective outcomes of both the data type reassessment and feature selection processes.

Table 3.1: Description of Columns and their P-values or Correlations

Column	P-value or Correlation	Value
Source Admission Description	P-value	<0.001
Last Non-ADL Ward Code	P-value	<0.001
Specialty Description	P-value	<0.001
Spell HRG Chapter Code	P-value	<0.001
VTE Flag	P-value	<0.001
Medically Optimized	P-value	<0.001
Outlier	P-value	<0.001
Dementia Diagnosis Flag	P-value	<0.001
IDT_pathway	P-value	<0.001
Referral Source Description	P-value	<0.001
Arrival Mode Description	P-value	<0.001
Acuity Code	P-value	<0.001
A&E Investigation Description 1	P-value	<0.001
A&E Investigation Description 2	P-value	<0.001
Attendance Type	P-value	<0.001
COVID Period	P-value	<0.001
Patient Age	Correlation	0.1936
Comorbidity Score	Correlation	0.2425
Arrival to Admission (Hrs)	Correlation	0.1253

3.5 Exploratory Analysis

This section delves into both visual and analytical findings, targeting the revelation of insights into the correlations between various variables and the length of stay (LoS).

3.5.1 Pair Plot of Numeric Variables

Figure 3.1 showcases a pair plot of selected numerical variables present in the dataset. This visualization provides a comprehensive view of the interrelationships among these numerical variables. Interestingly, most of the variables display a right-skewed distribution, except for 'Patient Age'. In this context, 'Patient Age' indicates the age of the patient at the time of admission, whereas 'Comorbidity Score' serves as an index to quantify the presence and severity of multiple comorbid conditions, based on the algorithm suggested by (Quan et al., 2005). Additionally, 'Arrival to Admission (Hrs)' specifies the duration, in hours, between a patient's arrival at the A&E and their subsequent hospital admission. The variable 'Spell LoS (Days)', the primary focus of this study, documents the length of the patient's hospital stay in days.

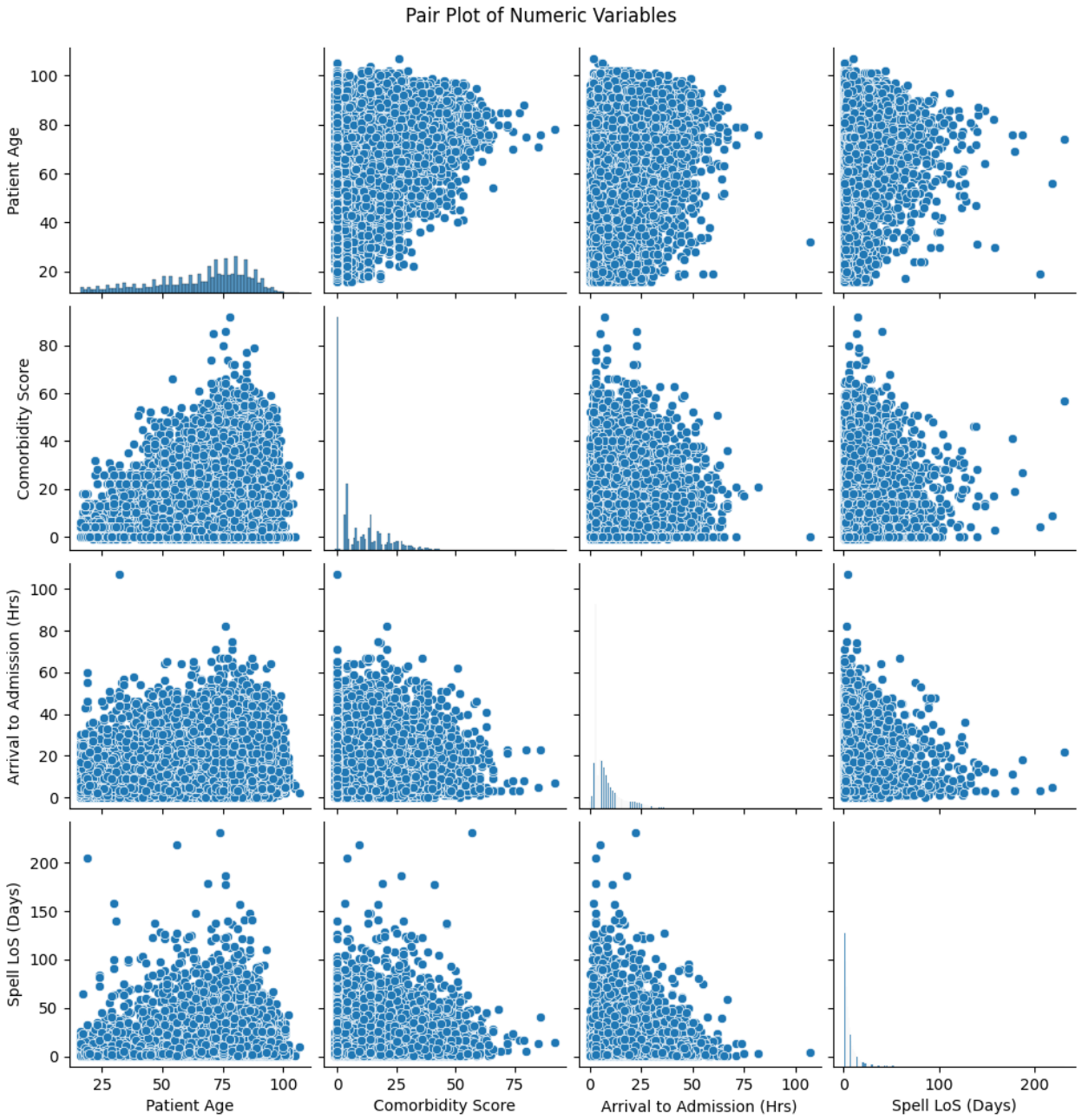

Figure 3.1: Pair Plot of Numeric Variables

Table 3.2 summarizes the selected features and their detailed associated statistics

Table 3.2: Summary Statistics of Numeric Variables

Variable	Count	Mean	Std	Min	Median	Max
Patient Age	68773.0	65.894	19.577	16.0	71.0	107.0
Comorbidity Score	68773.0	9.769	10.999	-1.0	5.0	92.0
Arrival to Admission (Hrs)	68773.0	8.124	7.813	0.1	5.0	107.0
Spell LoS (Days)	68773.0	6.987	9.535	1.0	4.0	231.0

3.5.2 Source Admission Description

This variable encompasses two categories: 'NON NHS NURSING HOME', which represents patients admitted from nursing homes, and 'USUAL ADDRESS', signifying patients admitted from other locations. As illustrated in Figure 3.2, the composite chart uncovers a notable trend: irrespective of the source of admission, patients tend to experience an extended LoS following a pandemic outbreak. Furthermore, the distribution of these categories appears to remain consistent pre- and post-pandemic.

Figure 3.2: Composite Chart of 'Source Admission Description'

3.5.3 Last Non-ADL Ward Code

This variable identifies the specific ward from which patients were discharged. Figure 3.3 presents a composite chart for 'Last Non-ADL Ward Code', exhibiting a noticeable shift in category distribution pre- and post-pandemic. This change could be attributed to modifications in the hospital's organizational structure that led to the creation of new wards.

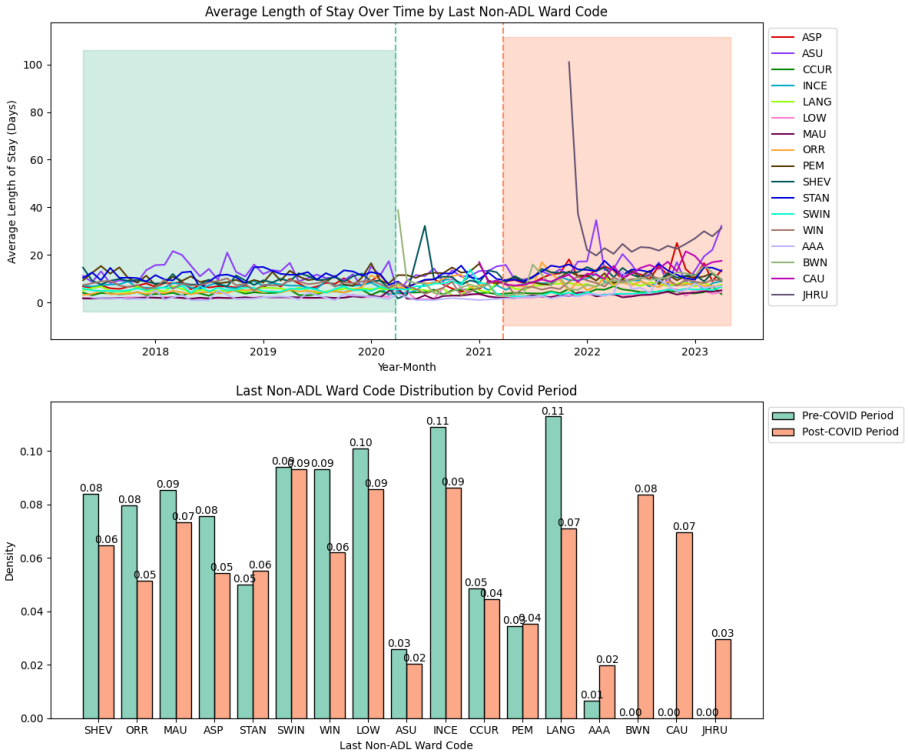

Figure 3.3: Composite Chart of 'Last Non-ADL Ward Code'

3.5.4 Specialty Description

Figures 3.4 illustrate the distribution of specialties to which patients were assigned. Specifically, 'general medicine' holds the majority, its share growing from 63% pre-pandemic to 69% post-pandemic. Following this, 'general surgery' remains the next most prevalent specialty, although its proportion declined from 16% to 13%. Figure 3.5 clarifies the average LoS across these specialties after the pandemic, indicating an overall upward trend. Notably, 'general medicine' saw its average LoS increase from 6.43 to 8.56 days, while 'general surgery' experienced a rise from 5.16 to 6.13 days. It's worth mentioning that 'elderly medicine' recorded the highest average LoS both before and after the pandemic, doubling its proportion and witnessing an increment from 10.83 to 14.13 days post-pandemic.

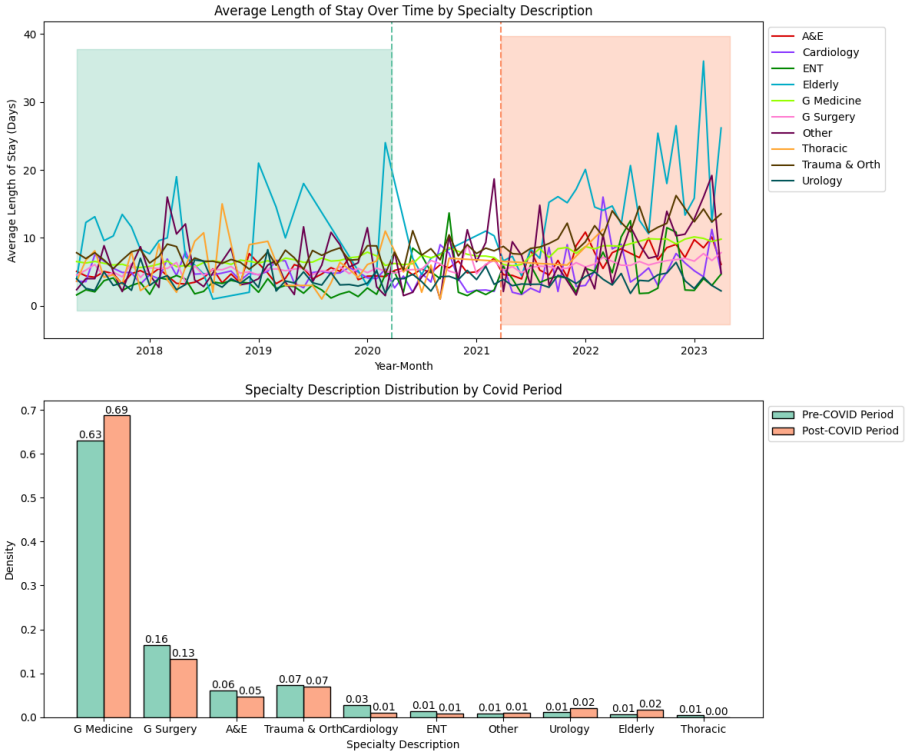

Figure 3.4: Composite Chart of 'Specialty Description'

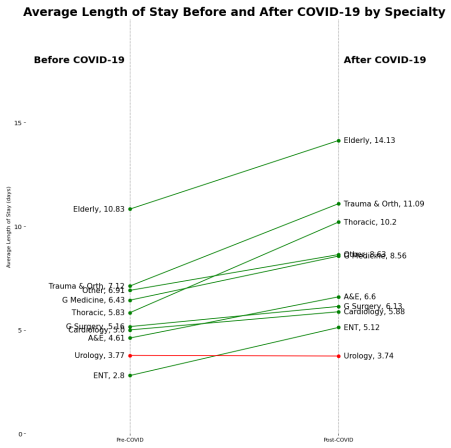

Figure 3.5: Average LoS Pre- and Post-COVID-19 by Specialty

3.5.5 Spell HRG Chapter Code

Within the NHS framework, Healthcare Resource Groups (HRGs) act as categories of patient cases anticipated to consume comparable levels of resources. Table 3.3 provides detailed descriptions of these codes.

Table 3.3: HRG Code Explanation

HRG Code	HRG Description
A	Nervous System
F	Digestive System
J	Skin, Breast and Burns
W	Infectious Diseases, Immune System Disorders and other Healthcare contacts
H	Musculoskeletal System
D	Respiratory System
E	Cardiac
S	Haematology, Chemotherapy, Radiotherapy and Specialist Palliative Care
C	Ear, Nose, Mouth, Throat, Neck and Dental
L	Urinary Tract and Male Reproductive System
M	Female Reproductive System and Assisted Reproduction
K	Endocrine and Metabolic System
V	Multiple Trauma, Emergency Medicine and Rehabilitation
G	Hepatobiliary and Pancreatic System
N	Obstetrics
P	Diseases of Childhood and Neonates
Y	Vascular Procedures and Disorders and Imaging Interventions
B	Eyes and Periorbita
U	Undefined Groups

As shown in Figure 3.6, patients with HRG code 'D' are the most prevalent, followed by those with code 'F'. The distribution exhibits minimal fluctuation before and after the pandemic. Concurrently, Figure 3.7 demonstrates that the average LoS for all HRG codes underwent an upward shift following the pandemic.

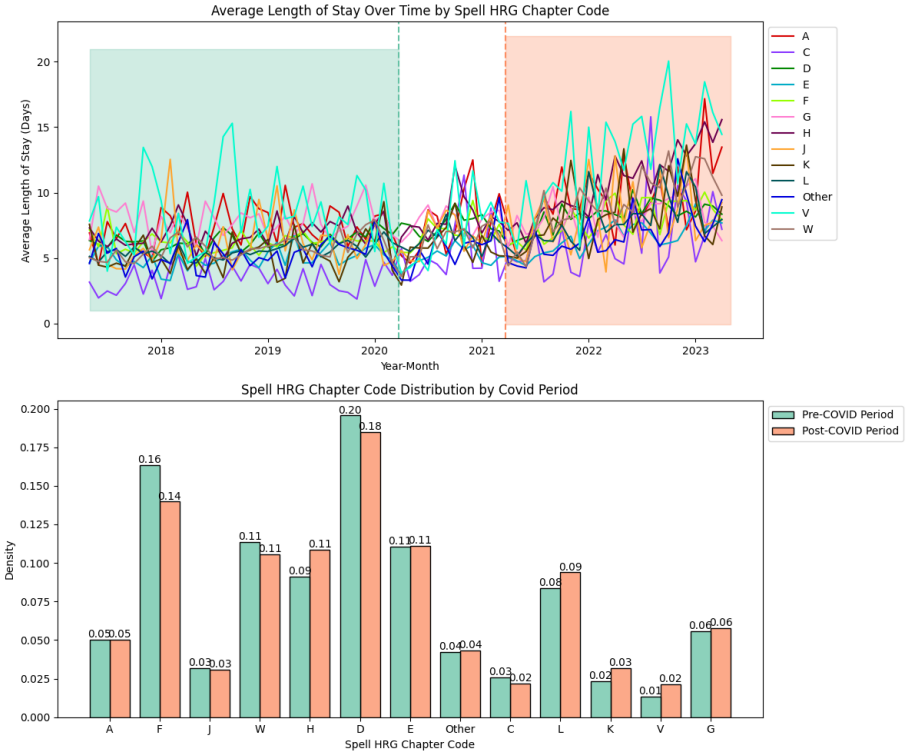

Figure 3.6: Composite Chart of 'Spell HRG Chapter Code'

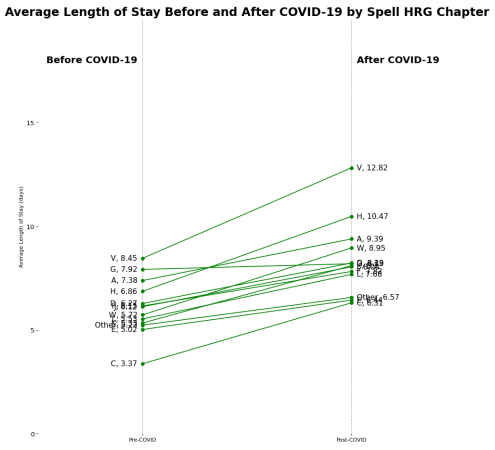

Figure 3.7: Average LoS Pre- and Post-COVID-19 by HRG Chapter

3.5.6 VTE Flag

This variable indicates whether patients are at risk of Venous Thromboembolism (VTE). A value of 1 signifies the presence of VTE risk, while a value of 0 indicates its absence. Figure 3.8 reveals that the ratio of patients identified as having VTE risk increased post-pandemic. Correspondingly, their average LoS also saw an upward trend. Conversely, patients without VTE risk displayed a declining pattern in average LoS subsequent to the pandemic.

Figure 3.8: Composite Chart of 'VTE Flag'

3.5.7 Dementia Diagnosis Flag

This indicator distinguishes whether patients have been diagnosed with dementia. A value of 1 signifies a dementia diagnosis, whereas a value of 0 denotes patients without such a diagnosis. Figure 3.9 reveals a 4-percentage-point decline in the ratio of patients diagnosed with dementia subsequent to the pandemic. Regardless of their dementia status, the average LoS for all patients increased following the pandemic.

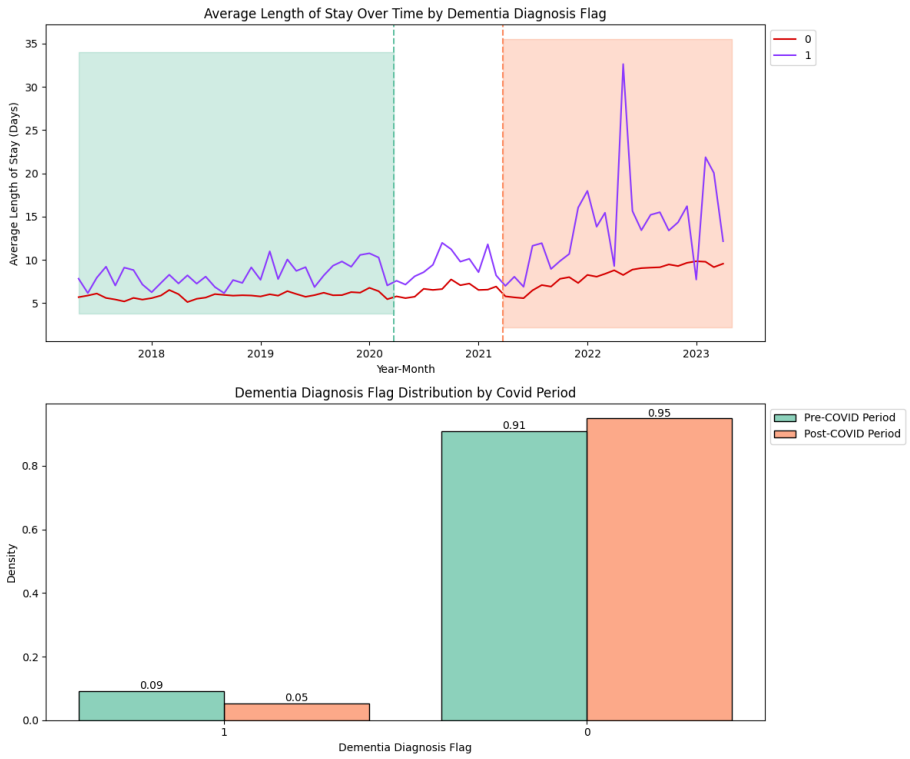

Figure 3.9: Composite Chart of 'Dementia Diagnosis Flag'

3.5.8 Referral Source Description

This metric enumerates the various sources from which patients are referred. Figure 3.10 indicates that the majority of patients were directly admitted to the Accident and Emergency (A&E) department, followed by those categorized as 'self-referral'. The proportion of the former increased from 52% pre-pandemic to 59% post-pandemic, while that of the latter decreased from 27% to 25%. Remarkably, there was a 50% reduction in the number of patients referred to A&E by General Practitioners (GPs) after the pandemic. Additionally, the average LoS generally rose across most referral sources post-pandemic.

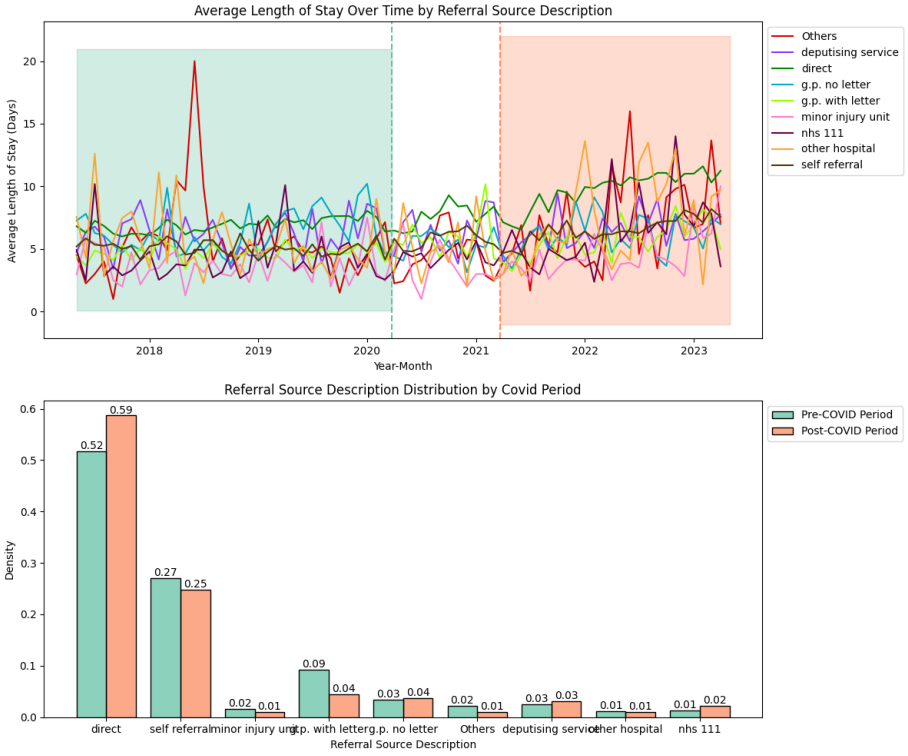

Figure 3.10: Composite Chart of 'Referral Source Description'

3.5.9 Arrival Mode Description

This variable delineates the method by which patients arrive at the Accident and Emergency (A&E) department. Figure 3.11 indicates that a substantial 65% of patients arrive via ambulance services, while the remaining 35% use alternative means. Interestingly, the distribution of this variable remains relatively stable both pre- and post-pandemic.

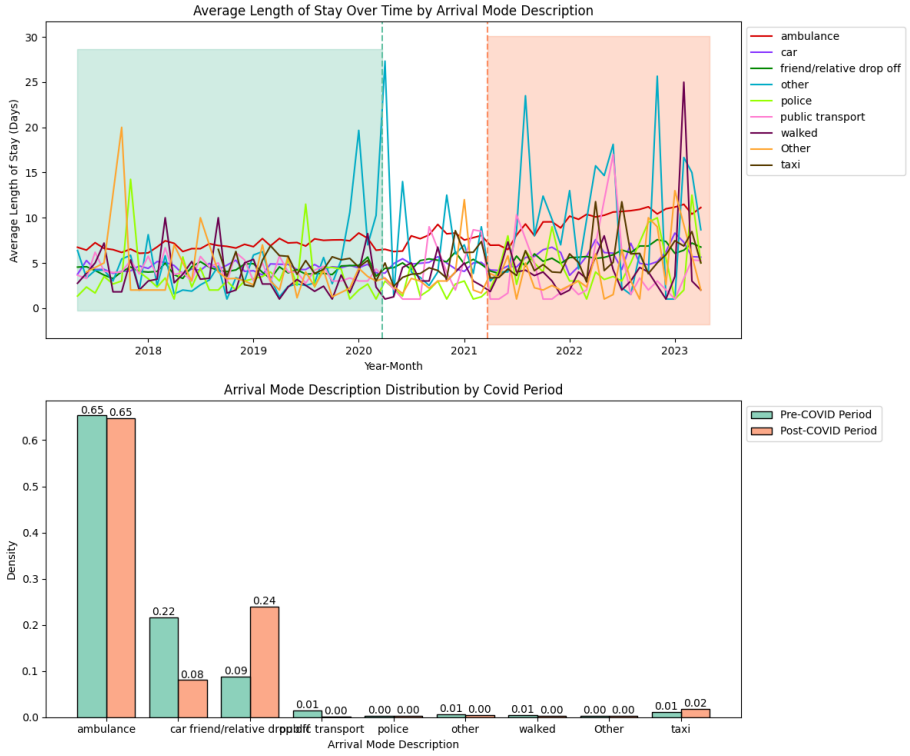

Figure 3.11: Composite Chart of 'Arrival Mode Description'

3.5.10 Acuity Code

This variable classifies patients based on their assigned acuity code, ranging from 1 to 5. A code of 1 indicates the most severe symptoms, while a code of 5 signifies the least severe. Figure 3.12 demonstrates that a substantial number of patients are classified under acuity codes 2, 3, and 4. Interestingly, post-pandemic data show a slight decrease in the prevalence of codes 2 and 3, alongside a minor increase in the proportion of code 4 patients.

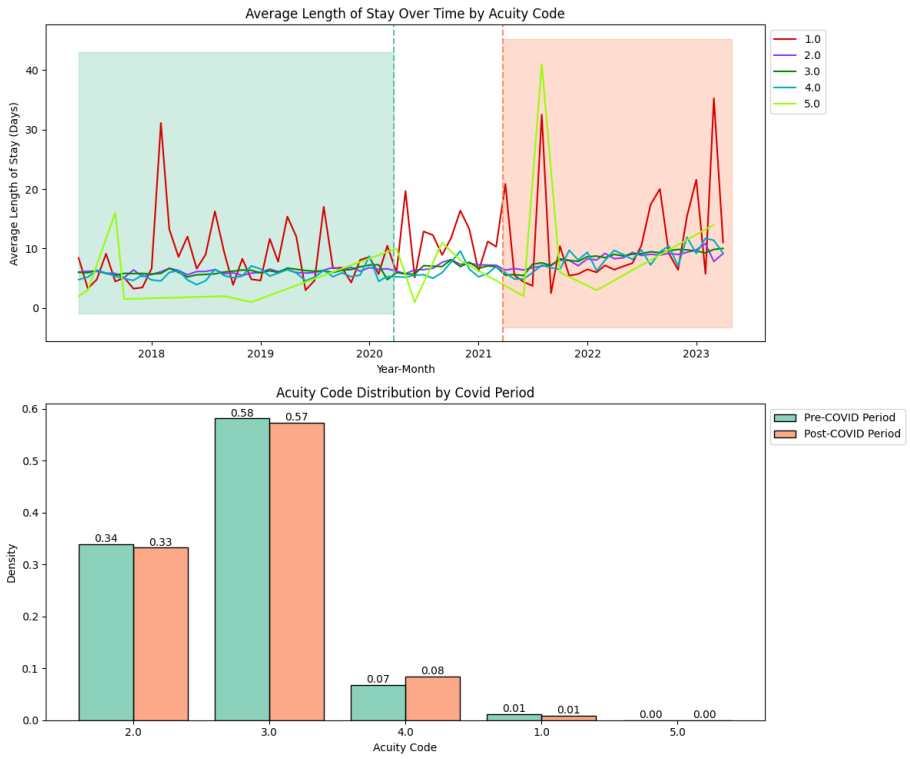

Figure 3.12: Composite Chart of 'Acuity Code'

3.5.11 Attendance Type

This metric categorizes patients based on their initial assessment and treatment locations, primarily divided into 'Resus', 'Majors', and 'Minors'. The 'Resus' category is for critically ill patients, 'Majors' for those with moderate severity, and 'Minors' for patients with less severe conditions. As depicted in Figure 3.13, the average LoS for the 'Majors' and 'Resus' categories has increased post-pandemic, while the LoS for 'Minors' remains relatively stable. The post-pandemic proportion of 'Majors' has risen by 12 percentage points, whereas that of 'Minors' has decreased by 11 percentage points.

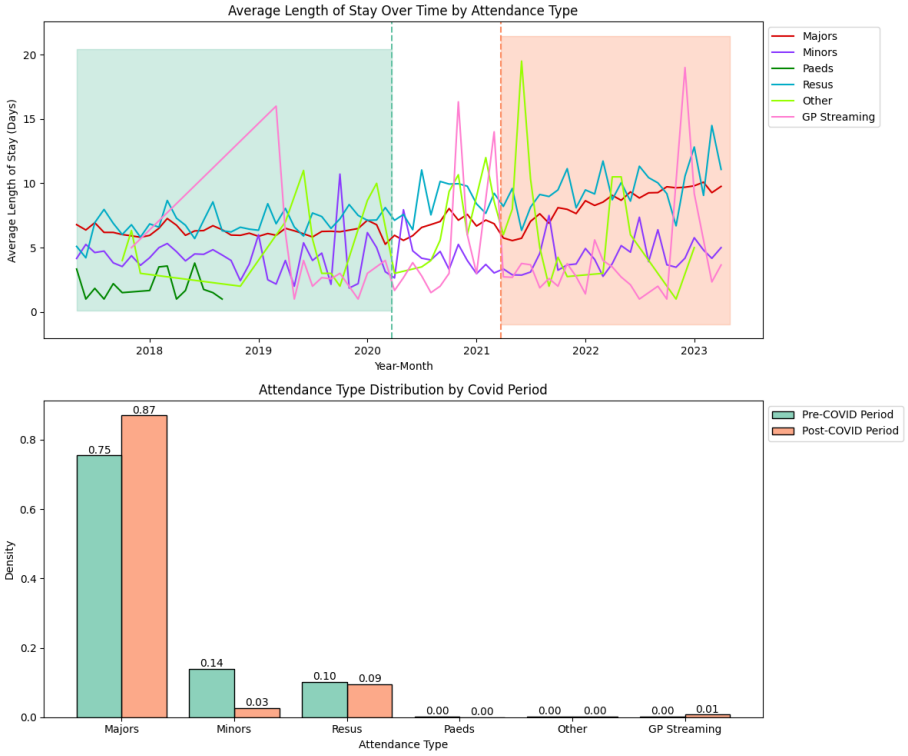

Figure 3.13: Composite Chart of 'Attendance Type'

3.5.12 Key Observations

- Across most categorical variables, an upward trend in the average LoS is evident, regardless of the category. Notable exceptions include 'Patients with Non-VTE Risk', patients routed for transfer to 'Minors', and those categorized under the 'Urology' specialty.
- Variables influenced by organizational restructuring pre- and post-pandemic are omitted from our analysis. These include terms like 'last non ADL ward code', 'medically_optimised', and 'A&E Investigation Description1'.
- The variables included in the model are listed in Table 3.4.

Table 3.4: Columns Used In The Model

Numerical Columns	**Categorical Columns**
Patient Age	Source Admission Description
Comorbidity Score	Specialty Description
Arrival to Admission (Hrs)	Spell HRG Chapter Code
Spell LoS (Days)	VTE Flag
	Dementia Diagnosis Flag
	Referral Source Description
	Arrival Mode Description
	Acuity Code
	Attendance Type

Chapter 4

Methodology

4.1 Statistical Analysis: Analysis of Variance (ANOVA)

For initial feature selection, and specifically to understand the impact of categorical variables on Length of Stay (LoS), we employed the Analysis of Variance (ANOVA) technique (St, Wold, et al., 1989). This statistical method is widely used to test the differences between two or more means. It is particularly effective for comparing mean differences among groups that are categorized based on multiple variables.

The null hypothesis (H_0) for ANOVA posits that the means of the different groups are equal. Conversely, the alternative hypothesis (H_a) suggests that at least one group mean differs from the others. In our study, the categorical variables functioned as the independent variables, and LoS served as the dependent variable. We performed the analysis at a 0.05 significance level using Python's `f_oneway` function from the SciPy library (Virtanen et al., 2020).

4.2 Correlation Analysis

Beyond ANOVA, we also utilized correlation analysis to investigate the relationships between continuous variables and Length of Stay (LoS). For this purpose, we used Python's `pearsonr` function from the SciPy library (Virtanen et al., 2020) to compute the Pearson correlation coefficient. This metric gauges the strength and direction of linear relationships between continuous variables and LoS (Cohen et al., 2009).

Our choice of the Pearson correlation coefficient was predicated on its appropriateness for quantifying linear associations. It yields a value between -1 and 1: -1 signifies a perfect negative linear correlation, 1 signifies a perfect positive linear correlation, and 0 implies no linear correlation (Benesty, Chen, and Y. Huang, 2008). This measure is notably relevant for scrutinizing how fluctuations in continuous variables might influence the LoS.

4.3 Clustering Technology

The essence of our methodology is founded on the application of advanced clustering techniques. These techniques serve to unearth unique patterns and categorize groupings within the data. This section elaborates on the chosen clustering technique and explicates its foundational principles.

4.3.1 K-Prototypes

Addressing the challenge of clustering extensive datasets comprising both numeric and categorical attributes, we employ the K-Prototypes algorithm, as available in the Python `kmodes` library (Vos, 2021). Initially proposed by Huang (Z. Huang, 1997), this algorithm amalgamates the benefits of the K-Means algorithm (Hartigan and Wong, 1979) for numerical attributes with those of the K-Modes algorithm (Chaturvedi, Green, and Caroll, 2001) for categorical attributes.

The K-Prototypes algorithm melds the principles of both K-Means and K-Modes. It employs the Euclidean distance measure for numerical attributes and a basic matching dissimilarity measure for categorical attributes. The cost function J is articulated as follows:

$$J = \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\left(\alpha\sum_{j=1}^{p}(x_{ij} - z_{kj})^2 + (1-\alpha)\delta\left(g_{ij}, h_{kj}\right)\right)$$

In this equation, w_{ik} denotes the weight for cluster k and object i, α is a balancing factor, p represents the number of numeric attributes, x_{ij} and g_{ij} are the numeric and categorical attributes, respectively, and z_{kj} and h_{kj} signify their corresponding cluster centroids.

In our Python implementation, we set the number of clusters (`n_clusters`) to 5, as guided by the elbow method. We also employ the 'Huang' initialization method (`init='Huang'`) to intelligently select the initial centroids. To bolster robustness, the algorithm runs with 10 different initializations (`n_init=10`). We fix the random state to 1 (`random_state=1`) to guarantee the reproducibility of our results.

4.3.2 Pre- and Post-Pandemic Data Clustering

In our study, we initially establish a baseline model by implementing clustering techniques on pre-pandemic data. This prototype model adopts a multi-dimensional strategy, integrating factors such as patient attributes, length of stay (LoS), and other key hospital performance metrics to create distinct clusters. The centroids of these pre-pandemic clusters function as an analytical foundation for deciphering standard patient behaviours and admission trends prior to the advent of COVID-19.

Subsequently, we apply these predefined centroids to post-pandemic patient data. Comparing the composition of clusters before and after the COVID-19 outbreak allows us to

discern notable alterations in patient characteristics, LoS, and other pivotal metrics. This method facilitates the identification of trends, comprehension of deviations, and the potential discovery of new patterns or subgroups that have come into being either directly or indirectly due to the pandemic.

4.3.3 Evaluation Metric: Sum of Squared Errors (SSE)

In cluster analysis, an effective evaluation metric is crucial for assessing the performance and efficacy of the clustering algorithm. A widely employed metric is the Sum of Squared Errors (SSE), which quantifies the compactness of clustering. SSE calculates the squared deviations between each data point and its cluster centroid and sums these across all clusters (Thinsungnoena et al., 2015):

$$SSE = \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - \mu_k\|^2 \tag{4.1}$$

Here, K denotes the number of clusters, C_k represents the k^{th} cluster, x_i is the i^{th} data point in C_k, and μ_k is the centroid of C_k. Lower SSE values suggest that data points are more proximal to the centroids of their respective clusters, implying a superior clustering model.

In our research, we calculated the SSE for various numbers of clusters to ascertain the optimum value of K. Utilizing the elbow method, we plotted SSE against the number of clusters (K) to pinpoint the 'elbow point' at which additional clusters would not yield a significant reduction in SSE. This 'elbow point' informed our final selection of the number of clusters.

Chapter 5

Results

5.1 Model Training

In this section, we outline the training procedure for our patient clustering model, which employs the K-Prototypes algorithm. The dataset is segmented into pre-pandemic and post-pandemic periods, and model training focuses solely on the pre-pandemic data. We use the Sum of Squared Errors (SSE) as the evaluative criterion to select the optimal number of clusters. By assessing the SSE across different numbers of clusters, our aim is to pinpoint the most suitable model configuration.

As shown in Figure 5.1, a clear 'elbow point' was evident when the number of clusters reached five. Accordingly, we chose this as the optimal number of clusters, allowing us to achieve a well-defined clustering structure for the dataset.

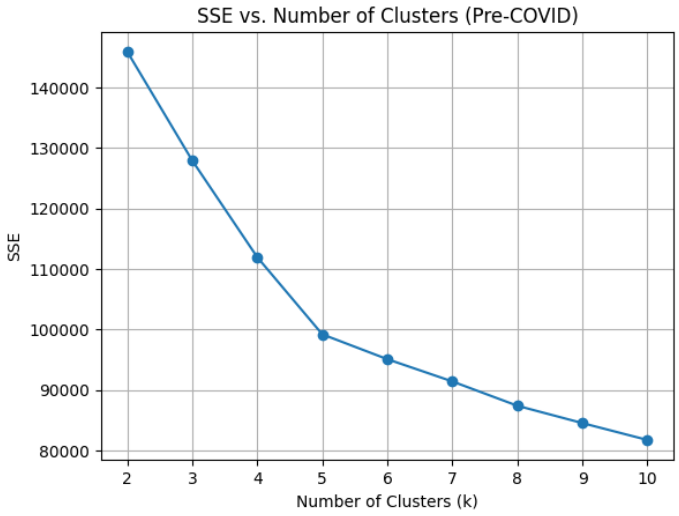

Figure 5.1: SSE vs. Number of Clusters

5.2 Model Outcomes

Table 5.1 delineates the centroids of the five identified clusters. Labels are assigned to each cluster based on its distinguishing characteristics. Specifically, Cluster 0 is primarily composed of elderly patients and is thus labeled 'Elderly'. Cluster 1 is characterized by extended stays in the emergency department, earning it the label 'Extended A&E Stay'. Cluster 2, representing the majority with normal attributes, is aptly termed 'Normal Majority'. Cluster 3, indicative of extended length of stay, is labeled 'High LoS'. Finally, Cluster 4, which predominantly includes a younger population, is labeled 'Young Adults'.

Table 5.1: Cluster Centroids

	Cluster 0	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
Patient Age	77.24	67.17	74.58	74.11	38.21
Comorbidity Score	23.86	7.50	5.04	15.45	1.38
Arrival To Admission (hrs)	4.79	12.33	3.89	5.65	4.16
LoS (days)	7.01	5.37	4.62	34.19	3.25
Referral Source	Direct	Direct	Direct	Direct	Self referral

5.3 Cluster Comparison Before and After COVID-19

5.3.1 Numeric Variables

Table 5.2 outlines the average shifts in numeric variables across the different clusters, both pre and post-pandemic. Analysis of the cluster counts reveals a notable decrease in the sizes of clusters 0, 2, and 4. Conversely, the size of cluster 4 remains constant, while cluster 1 experiences a significant increase. Our analysis indicates a post-pandemic inclination for patient characteristics to more closely align with cluster 1.

In terms of specific variables, all clusters show a general upward trend post-pandemic, affecting factors such as age, comorbidity scores, time from emergency department to hospitalization, Length of Stay (LoS), and emergency department waiting assessment time. Importantly, cluster 1 has almost tripled in size, with the average hospitalization time rising from 5.37 days to 7.1 days and the average waiting time in the emergency department extending from 12.33 hours to 20.47 hours.

Turning to cluster 3, the proportions remain largely unchanged before and after the pandemic. However, minor changes in patient characteristics are observed. For instance, LoS has lengthened from 34.19 days to 41.88 days, and the average emergency department waiting time has doubled in the post-pandemic period.

Cluster 4, when compared to the other clusters, demands fewer medical resources. Nevertheless, this cluster has seen a considerable reduction in its size following the pandemic.

In summary, as depicted in Figure 5.2, post-pandemic patients are more widely distributed along both the X-axis and Y-axis. Here, the X-axis corresponds to LoS, while the Y-axis reflects the waiting time for hospital admission. This suggests that post-pandemic patients not only have longer LoS but also endure more extended waiting times for admission.

Table 5.2: Mean Numerical Variables Before and After COVID-19 for Each Cluster

	Cluster 0	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
Pre-COVID					
Count	6474	4291	11985	1416	8986
Patient Age	77.24	67.16	74.58	74.11	38.21
Comorbidity Score	23.86	7.50	5.04	15.45	1.38
Arrival to Admission (hrs)	4.79	12.33	3.89	5.65	4.16
Length of Stay (days)	7.01	5.37	4.62	34.19	3.25
Post-COVID					
Count	3358	12474	4006	1385	3330
Patient Age	78.03	69.58	74.58	75.85	38.43
Comorbidity Score	25.79	11.15	5.19	18.21	1.58
Arrival to Admission (hrs)	6.28	20.47	4.52	11.73	4.91
Length of Stay (days)	7.44	7.10	5.02	41.88	3.42

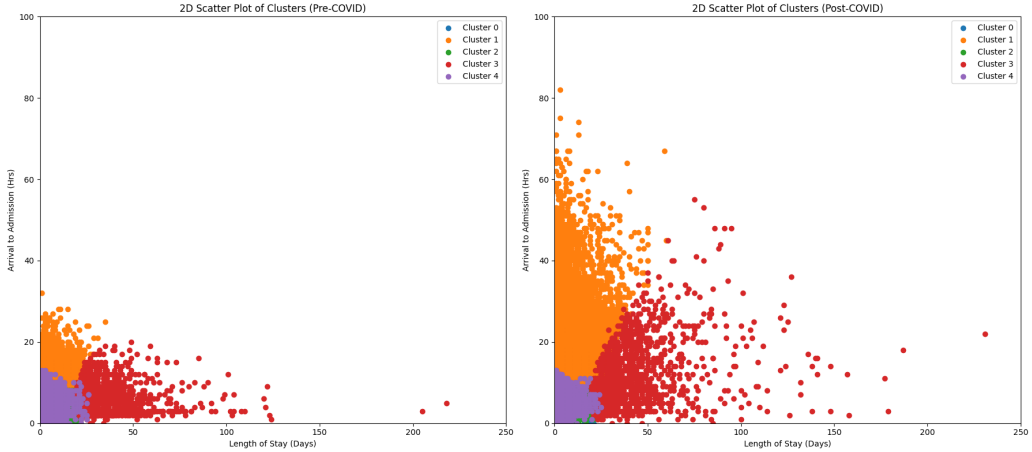

Figure 5.2: Clustering Visualization of Length of Stay and A&E Waiting Time for Admission

5.3.2 Category Variables

Based on Figure 5.3, one can observe that Cluster 4, both pre and post-pandemic, predominantly consists of HRG Chapter F patients. These patients mainly exhibit conditions

related to the Digestive System and consume the least amount of healthcare resources among all clusters. Conversely, the majority of the remaining clusters are primarily composed of HRG Chapter D, which represents patients with conditions pertaining to the Respiratory System. Notably, in Cluster 3—characterized by a higher Length of Stay (LoS)—HRG Chapter H, representing Musculoskeletal System conditions, has supplanted Chapter D as the majority post-pandemic.

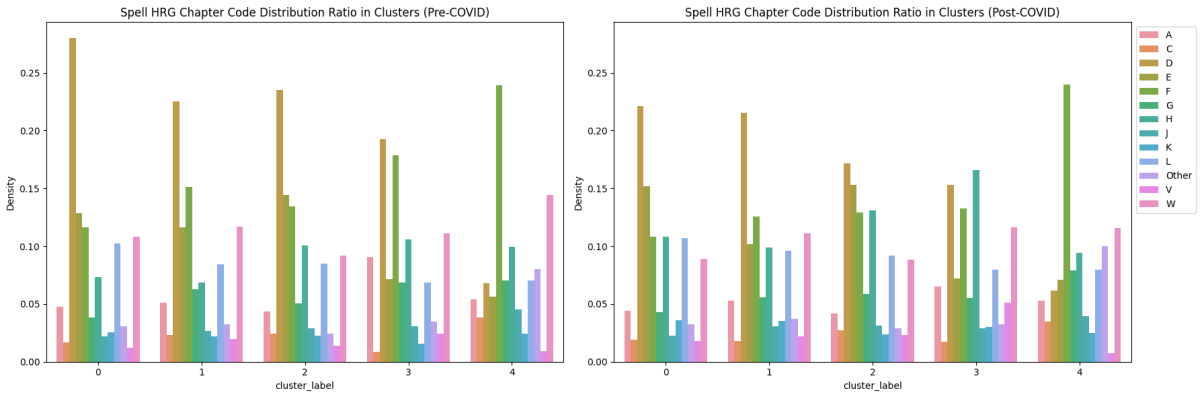

Figure 5.3: Comparison of HRG Chapter Distribution by Cluster

Figure 5.4 showcases the referral pathways through which patients arrived at the Accident and Emergency (A&E) department. Across all clusters, there is an apparent increase in the proportion of patients who directly present to the A&E. Notably, within Cluster 4, the number of patients referred from General Practitioners (GPs) to A&E has notably declined. This trend aligns with earlier exploratory data analysis (EDA) insights, suggesting that patients increasingly prefer to directly seek care at the A&E department post-pandemic, rather than initially visiting local clinics or general practitioners.

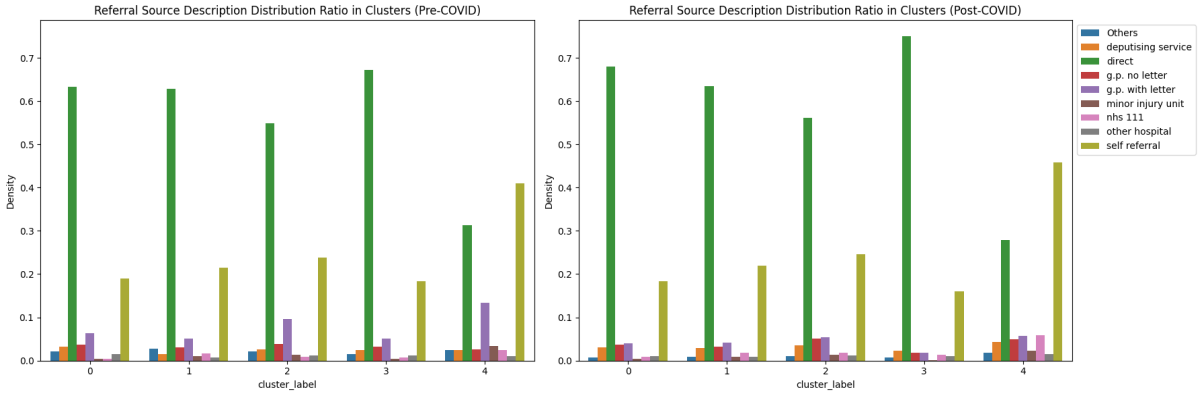

Figure 5.4: Comparison of Referral Source by Cluster

Continuing to examine the modes of arrival at A&E, Figure 5.5 indicates a decline in the percentage of patients arriving by ambulance in all clusters post-pandemic. Noteworthy is the significant uptick in the proportion of patients in Cluster 4 arriving at A&E independently.

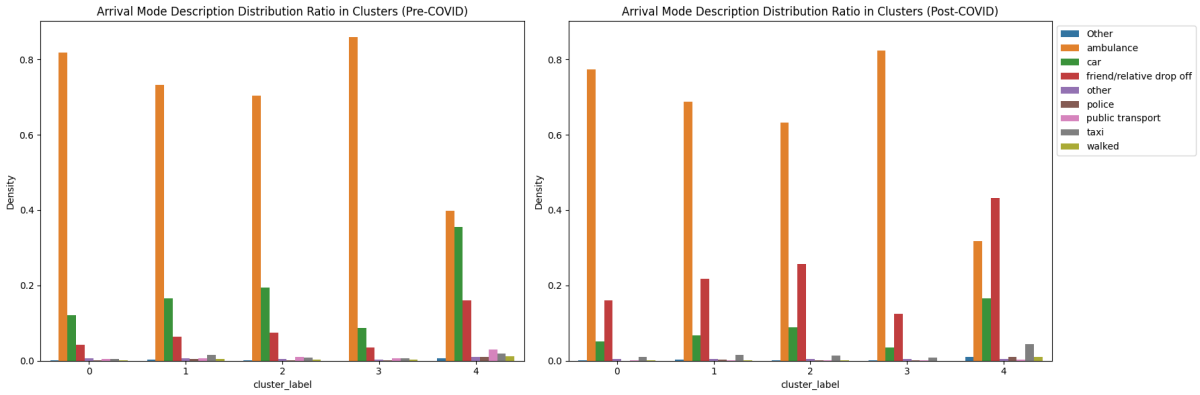

Figure 5.5: Comparison of Arrival Mode by Cluster

Figure 5.6 portrays the acuity codes for patients arriving at A&E before and after the pandemic. The data reveals that the proportion of patients with acuity codes 2 and 3 either declined or remained stable across all clusters. Conversely, patients classified under acuity code 4 exhibited an increasing or stable trend. Thus, based on the acuity codes, it can be inferred that the severity of conditions for patients arriving at A&E did not escalate post-pandemic but rather experienced a slight decline.

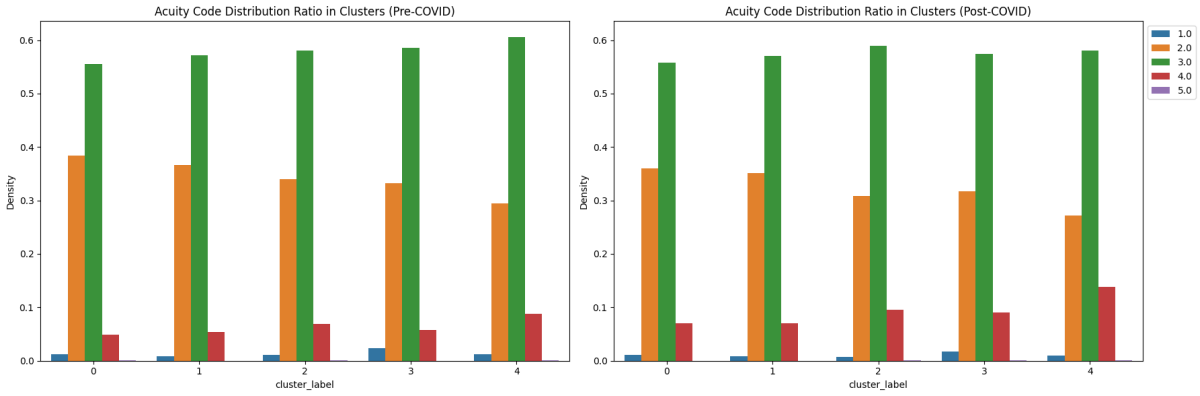

Figure 5.6: Comparison of Acuity Code by Cluster

In Figure 5.7, the distribution of patients among different acuity categories (Resus, Majors, Minors) pre and post-pandemic is presented. Notably, there is a marked increase

in the proportion of patients classified as 'Majors' across all clusters after the pandemic. Conversely, the 'Minors' category witnessed a significant decline in its utilization rate.

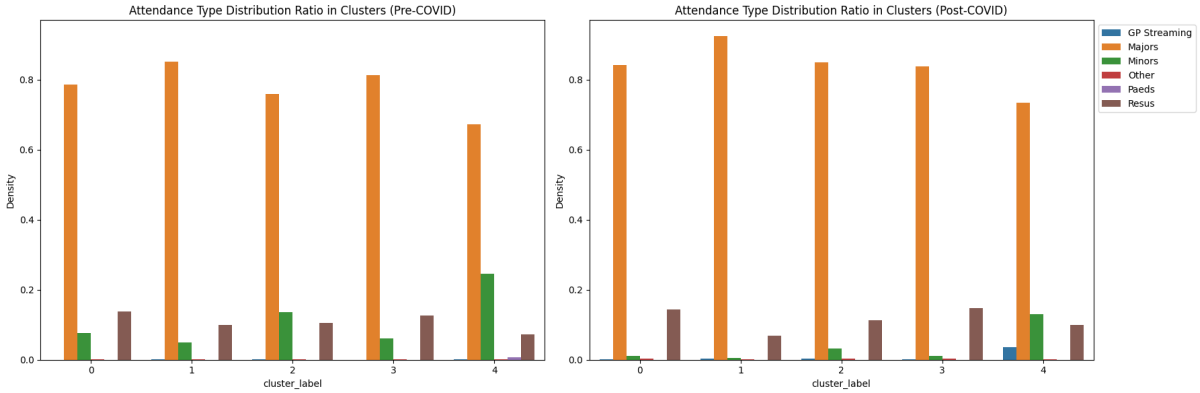

Figure 5.7: Comparison of Attendance Type by Cluster

To explore the marked discrepancies in the utilization rates of 'Majors' and 'Minors' areas pre- and post-pandemic—despite the consistent distribution of acuity codes across all clusters—Figures 5.8 and 5.9 offer insights into the area assignments given to patients based on their initial acuity code categorization. Importantly, a post-pandemic trend indicates that a higher proportion of patients with acuity codes 3 and 4 were subsequently allocated to the 'Major' area.

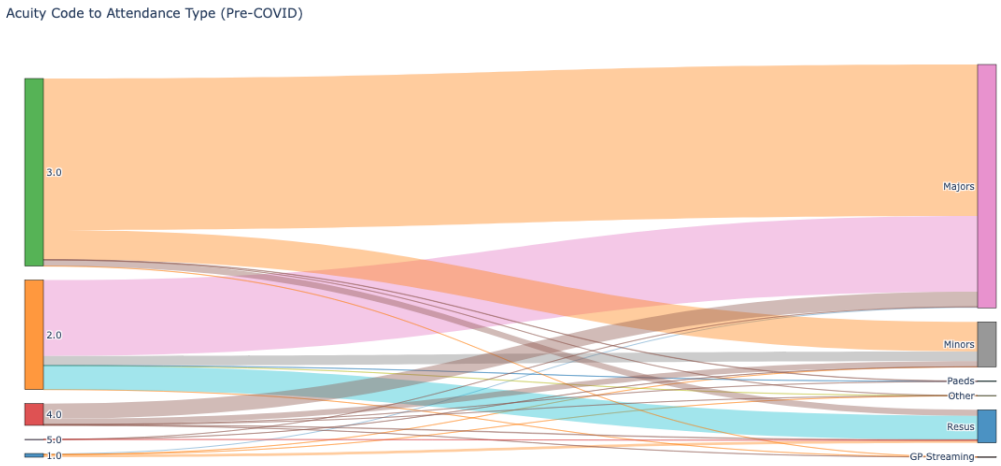

Figure 5.8: Patient Flow from Acuity Code to Attendance Type (Pre-COVID)

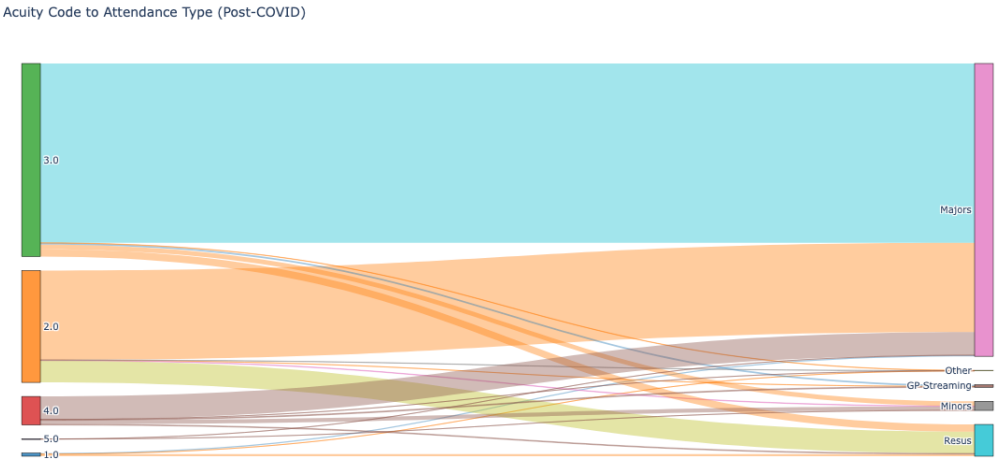

Figure 5.9: Patient Flow from Acuity Code to Attendance Type (Post-COVID)

5.4 Correlation Between Waiting Assessments and Acuity Code

To conclude, we turn our attention to the duration that patients waited for assessments upon their arrival at the Accident and Emergency (A&E) department. As delineated in Figure 5.10, it becomes evident that post-COVID, patients categorized under Acuity Codes 3 and 4 experienced a doubled waiting time, while those under Acuity Code 2 endured a waiting time one and a half times longer than in the pre-COVID period.

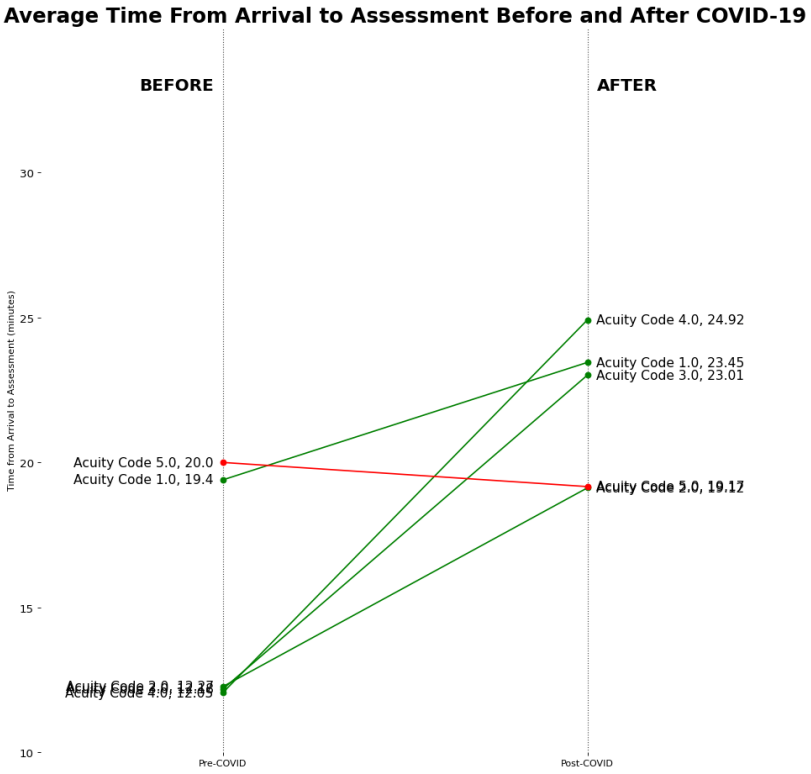

Figure 5.10: Slope Plot of Acuity Code and Waiting Assessment Before and After COVID

Chapter 6

Discussion

In this paper, we have conducted a comprehensive analysis of the attributes of inpatients admitted to the emergency department at the Wrightington, Wigan and Leigh Teaching Hospitals NHS Foundation Trust (WWL), both before and after the COVID-19 outbreak. By employing the K-Prototypes clustering algorithm, which focuses on Length of Stay (LoS), we scrutinized the features and behavioural shifts of patients across various clusters. Our analysis has uncovered several significant trends and associations that warrant further investigation.

Firstly, we noted a conspicuous increase in both the average LoS and the waiting time for admission following the onset of the COVID-19 pandemic. The relationship between these two variables has also strengthened (Lauks et al., 2016). Through cluster analysis, we observed a prevalent shift of patients towards clusters characterized by extended waiting times for admission, as opposed to increased LoS. This could indicate that the challenges confronting WWL in the post-outbreak environment are more related to throughput issues in the emergency department, rather than LoS. This finding lends support to existing arguments suggesting that increased LoS is a consequence of emergency department overcrowding (Mentzoni, Bogstrand, and Faiz, 2019).

Secondly, post-outbreak data reveal a rise in both the average age of patients, from 64 to 67, and comorbidity scores, from 8.4 to 11.2. Additionally, our clustering outcomes indicate that patients with musculoskeletal conditions now constitute the majority within high-LoS clusters, overtaking those with respiratory conditions. This shift could point to the long-term impact of COVID-19 on the musculoskeletal system (Ramani et al., 2021; Disser et al., 2020), and it implies an increased demand for healthcare resources to manage these cases.

On another note, according to acuity code standards, there has been no notable alteration in the initial severity of conditions that lead to admissions, either pre- or post-outbreak. Nevertheless, there has been a marked increase in the proportion of patients classified as 'Majors' across all clusters, accompanied by a corresponding decline in the 'Minors' category (Dickson, Mason, and Bailey, 2017). This raises questions about the efficacy of acuity codes in allocating healthcare resources optimally. In our study, we found that nearly 70% of

patients coded as 4 (Standard level emergency care) were assigned to 'Majors' pre-outbreak, a proportion that escalated to 81% post-outbreak.

Lastly, the waiting times for initial patient assessments have seen a considerable increase post-outbreak. For patients coded as 4 and 3, waiting times surged from 12 minutes to 24 minutes, whereas for those coded as 2, it escalated from 12 minutes to 19 minutes. These observations necessitate urgent attention from healthcare management to enhance patient experience and uphold the standard of care.

One limitation worth mentioning is our application of ANOVA to detect differences between clusters, despite the awareness that our data distributions are skewed. ANOVA traditionally assumes a normal distribution of data; therefore, skewed data may influence the precision of our conclusions. This limitation suggests that future research could benefit from utilizing non-parametric statistical methods to secure more reliable outcomes.

Another point of consideration lies in our choice to determine the number of clusters based on pre-pandemic data. While this method permits a direct comparison between the pre- and post-pandemic eras, it restricts our comprehension of whether the clustering structure has evolved as a result of the pandemic. For instance, it is conceivable that six or more clusters might now offer a more accurate representation of the post-pandemic patient population. Follow-up studies should contemplate re-assessing the optimal number of clusters in the post-pandemic context.

Furthermore, our analysis is confined to data collected from a single healthcare trust, which imposes a limitation on the generalizability of our findings. Conducting multi-center studies could yield more resilient and universally applicable insights, enriching our understanding of the broader impact of the COVID-19 pandemic on healthcare systems.

Chapter 7

Conclusion

In this study, we undertook a comprehensive analysis of A&E inpatient characteristics at Wrightington, Wigan and Leigh Teaching Hospitals NHS Foundation Trust (WWL) both before and after the COVID-19 outbreak. Employing a K-Prototypes clustering algorithm, we categorized patients based on their length of stay (LoS) and deeply investigated the traits and behavioural shifts within each cluster. Our primary observations are as follows: firstly, both the average LoS and waiting times experienced a marked rise post-outbreak, with the correlation between the two factors also intensifying. Furthermore, there is an increasing trend in the number of patients whose characteristics align more closely with clusters characterized by extended waiting times for admission than those with high LoS. This implies that the main challenges for WWL might increasingly pertain to A&E throughput rather than LoS.

Secondly, post-outbreak data showed an increase in the average age from 64 to 67 years and in comorbidity scores from 8.4 to 11.2. Notably, musculoskeletal disorders have become more prevalent among high-LoS patients, suggesting that COVID-19 could have long-term ramifications on this health domain.

Further, while no significant change in the severity of conditions was observed, a noteworthy increase in the assignment of patients to the 'Majors' area was recorded, alongside a decrease in 'Minors' usage. Lastly, the waiting time from patient arrival at A&E to the initial assessment also increased significantly, warranting immediate attention and remedial action from hospital management.

To sum up, the COVID-19 pandemic has had widespread impacts on all aspects of inpatient care at WWL hospitals. These emergent trends present new operational challenges and necessitate proactive adaptations in healthcare processes and resource allocation. Based on our findings, we propose several actionable recommendations for the NHS. Primarily, in light of the elongated initial evaluation waiting times in A&E, establishing a Rapid Assessment Team is advised to hasten the assessment process and potentially diminish overall waiting times. Secondly, given the re-categorisation trends towards 'Majors' and away from

'Minors', a reassessment of the existing acuity code criteria is recommended for more accurate and efficient resource allocation. Additionally, the observed uptick in older patients with higher comorbidity scores prompts a review of community support options for the elderly to potentially alleviate inpatient burdens. Lastly, considering the increased prevalence of musculoskeletal conditions in high-LoS clusters, focused investments in specialized and perhaps preventive care for such conditions are warranted.

Appendix A

Project Specification

Understanding Patient Length of Stay Dynamics

Yi-Chun Huang
Project hosted by WWL
21 July 2023

Project Background / Context / Motivation

The duration of patients' hospital stays, known as the length-of-stay (LoS), is a critical measure that indicates how long a patient remains in the hospital from admission to discharge. Several factors, including the seriousness of the illness, age, presence of other medical conditions, and shifts in population demographics, can affect the LoS. Our Trust's recent data has revealed significant changes in patient demographics, characterized by an upsurge in the average age and LoS since 2016. Additionally, the COVID-19 pandemic has had a further impact on LoS, resulting in extended stays for patients across all age groups. Prolonged LoS is linked to unfavorable patient outcomes and has implications for bed availability and demand. To tackle these challenges, our objective is to comprehend the contributing elements to the increased LoS and explore potential strategies to adapt to the post-pandemic population.

Project Aims

The main goal of this project is to explore the factors that have led to an increase in the length-of-stay of patients at our hospital, with a specific focus on both the periods before and after the pandemic. The project aims to achieve the following objectives:

1. Analyzing patient characteristics and demographics to detect trends and changes in length-of-stay patterns over time.

2. Investigating the impact of the COVID-19 pandemic on length-of-stay and assessing how it varies among different age groups.

3. Identifying specific patient subgroups with exceptionally long length-of-stay and comprehending the key factors contributing to their extended stays.

4. Developing predictive models using clustering or classification techniques to categorize patients based on their length-of-stay patterns.

5. Generating practical insights to guide operational strategies for managing bed capacity and enhancing patient outcomes in a post-pandemic population.

Project Data

For this project, we will utilize electronic health records and administrative data from our Trust's hospital system. The dataset will contain admission and discharge timestamps, patient demographics (age, gender, etc.), disease severity scores, comorbidity information, and relevant clinical variables.

Project Deliverables

The following outcomes are expected to be delivered from this project:

1. A comprehensive analysis report providing insights into the changes in patient demographics, length-of-stay trends, and the impact of the COVID-19 pandemic on length-of-stay.

2. Developed predictive models or clustering/classification techniques that will enable the identification of patient subgroups based on their length-of-stay patterns.

3. Identification of key factors influencing prolonged length-of-stay in specific patient groups.

4. Recommendations and actionable insights to optimize bed capacity and enhance patient outcomes in a post-pandemic population.

5. A final presentation summarizing the findings and proposing potential strategies to senior leadership and stakeholders.

Candidate Techniques

To address the challenge of understanding length-of-stay (LoS) patterns, we can utilize the following techniques:

1. Exploratory Data Analysis (EDA): By conducting EDA, we can gain valuable insights into the distribution of LoS, patient characteristics, and how these variables have changed over time.

2. Time-Series Analysis: Implementing time-series analysis will help us examine the variations in LoS patterns across different periods and understand any temporal trends.

3. Machine Learning: Leveraging clustering or classification algorithms will enable us to categorize patients into different groups based on their LoS behavior, allowing us to identify distinct patient subgroups.

4. Feature Importance Analysis: Conducting feature importance analysis will assist in identifying the key factors that significantly influence prolonged LoS in specific patient subgroups.

5. Data Visualization: Creating informative visualizations will aid in effectively presenting the findings and making complex patterns and trends more accessible to stakeholders.

By employing these techniques, we can gain a comprehensive understanding of the factors impacting LoS and develop actionable insights to optimize patient care and hospital resource management.

Draft Project Timeline / Plan

The project will be conducted over a duration of 14 weeks. The tentative timeline includes the following milestones:

- Weeks 1-3: Familiarise myself with the company environment and data and exploratory data analysis
- Weeks 4-6: Analyse LoS trends and patient characteristics over time
- Weeks 7-8: Develop predictive models or clustering/classification techniques
- Week 9: Provide interim reports to senior leadership and stakeholders, summarise progress reports
- Weeks 10-14: Writing and providing final reports to senior leaders and stakeholders

References

Bahrmann, Anke et al. (2019). "The Charlson Comorbidity and Barthel Index predict length of hospital stay, mortality, cardiovascular mortality and rehospitalization in unselected older patients admitted to the emergency department". In: *Aging Clinical and Experimental Research* 31, pp. 1233–1242.

Benesty, Jacob, Jingdong Chen, and Yiteng Huang (2008). "On the importance of the Pearson correlation coefficient in noise reduction". In: *IEEE Transactions on Audio, Speech, and Language Processing* 16.4, pp. 757–765.

Buttigieg, Sandra C, Lorraine Abela, and Adriana Pace (2018). "Variables affecting hospital length of stay: a scoping review". In: *Journal of health organization and management* 32.3, pp. 463–493.

Chaturvedi, Anil, Paul E Green, and J Douglas Caroll (2001). "K-modes clustering". In: *Journal of classification* 18, pp. 35–55.

Chudasama, Yogini V, Kamlesh Khunti, and Melanie J Davies (2021). "Clustering of comorbidities". In: *Future Healthcare Journal* 8.2, e224.

Cohen, Israel et al. (2009). "Pearson correlation coefficient". In: *Noise reduction in speech processing*, pp. 1–4.

Daghistani, Tahani A et al. (2019). "Predictors of in-hospital length of stay among cardiac patients: a machine learning approach". In: *International journal of cardiology* 288, pp. 140–147.

El-Darzi, Elia et al. (2009). "Length of stay-based clustering methods for patient grouping". In: *Intelligent patient management*, pp. 39–56.

Dickson, Jon M, Suzanne M Mason, and Andy Bailey (2017). "Emergency department diagnostic codes: useful data?" In: *Emergency Medicine Journal*.

Disser, Nathaniel P et al. (2020). "Musculoskeletal consequences of COVID-19". In: *The Journal of bone and joint surgery. American volume* 102.14, p. 1197.

Freitas, Alberto et al. (2012). "Factors influencing hospital high length of stay outliers". In: *BMC health services research* 12, pp. 1–10.

Gong, Xuran et al. (2022). "Managing hospital inpatient beds under clustered overflow configuration". In: *Computers & Operations Research* 148, p. 106021.

Hartigan, John A and Manchek A Wong (1979). "Algorithm AS 136: A k-means clustering algorithm". In: *Journal of the royal statistical society. series c (applied statistics)* 28.1, pp. 100–108.

Hassan, Mahmud et al. (2010). "Hospital length of stay and probability of acquiring infection". In: *International Journal of pharmaceutical and healthcare marketing* 4.4, pp. 324–338.

Higgins, Thomas L et al. (2003). "Early indicators of prolonged intensive care unit stay: Impact of illness severity, physician staffing, and pre–intensive care unit length of stay". In: *Critical care medicine* 31.1, pp. 45–51.

Huang, Z. (1997). "Clustering Large Data Sets with Mixed Numeric and Categorical Values". In: *Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining, (PAKDD)*, pp. 21–34.

Iwase, Shinya et al. (2022). "Prediction algorithm for ICU mortality and length of stay using machine learning". In: *Scientific reports* 12.1, p. 12912.

Klein, Rudolf (2004). *The first wave of NHS foundation trusts*.

Lauks, Juliane et al. (2016). "Medical team evaluation: effect on emergency department waiting time and length of stay". In: *PloS one* 11.4, e0154372.

Liu, Yingxin, Mike Phillips, and Jim Codde (2001). "Factors influencing patients' length of stay". In: *Australian Health Review* 24.2, pp. 63–70.

McDermott, Christopher and Gregory N Stock (2007). "Hospital operations and length of stay performance". In: *International Journal of Operations & Production Management* 27.9, pp. 1020–1042.

Mentzoni, Ida, Stig Tore Bogstrand, and Kashif Waqar Faiz (2019). "Emergency department crowding and length of stay before and after an increased catchment area". In: *BMC health services research* 19, pp. 1–11.

Moisoglou, Ioannis et al. (2019). "Nursing staff and patients' length of stay". In: *International Journal of Health Care Quality Assurance* 32.6, pp. 1004–1012.

Ogbuabor, Godwin and FN Ugwoke (2018). "Clustering algorithm for a healthcare dataset using silhouette score value". In: *Int. J. Comput. Sci. Inf. Technol* 10.2, pp. 27–37.

Panchami, VU and N Radhika (2014). "A novel approach for predicting the length of hospital stay with DBSCAN and supervised classification algorithms". In: *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*. IEEE, pp. 207–212.

Qualls, Munirih, Daniel J Pallin, and Jeremiah D Schuur (2010). "Parametric versus nonparametric statistical tests: the length of stay example". In: *Academic Emergency Medicine* 17.10, pp. 1113–1121.

Quan, Hude et al. (2005). "Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data". In: *Medical care*, pp. 1130–1139.

Ramani, Santhoshini Leela et al. (2021). "Musculoskeletal involvement of COVID-19: review of imaging". In: *Skeletal radiology* 50, pp. 1763–1773.

Salway, RJ et al. (2017). "Emergency department (ED) overcrowding: evidence-based answers to frequently asked questions". In: *Revista Medica Clinica Las Condes* 28.2, pp. 213–219.

St, Lars, Svante Wold, et al. (1989). "Analysis of variance (ANOVA)". In: *Chemometrics and intelligent laboratory systems* 6.4, pp. 259–272.

Tennison, Imogen et al. (2021). "Health care's response to climate change: a carbon footprint assessment of the NHS in England". In: *The Lancet Planetary Health* 5.2, e84–e92.

Thinsungnoena, Tippaya et al. (2015). "The clustering validity with silhouette and sum of squared errors". In: *learning* 3.7.

UK Government (2023). *COVID-19 Deaths in the UK*. Accessed on: September 1, 2023. URL: `https://coronavirus.data.gov.uk/details/deaths`.

Virtanen, Pauli et al. (2020). "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17, pp. 261–272. DOI: `10.1038/s41592-019-0686-2`.

Vos, Nelis J. de (2021). *kmodes categorical clustering library*. `https://github.com/nicodv/kmodes`.

Vranas, Kelly C et al. (2017). "Identifying distinct subgroups of intensive care unit patients: A machine learning approach". In: *Critical care medicine* 45.10, p. 1607.