Hospital Stay Analytics: Understanding Patient Data

By Elena Justo

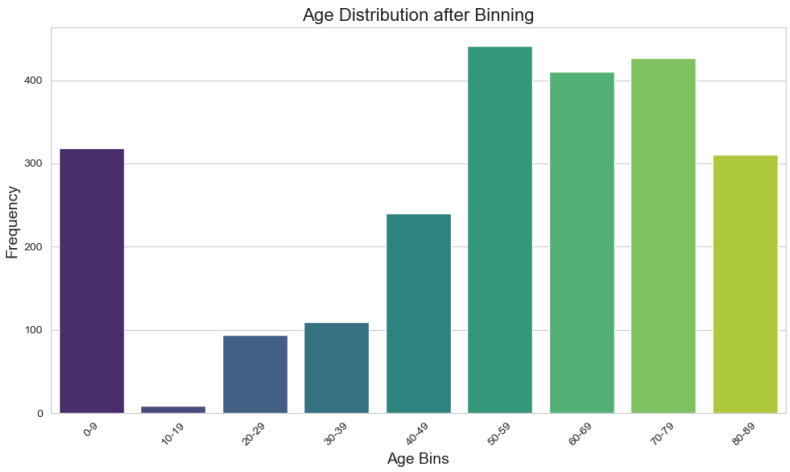
Student ID: 24429298

Table of Contents

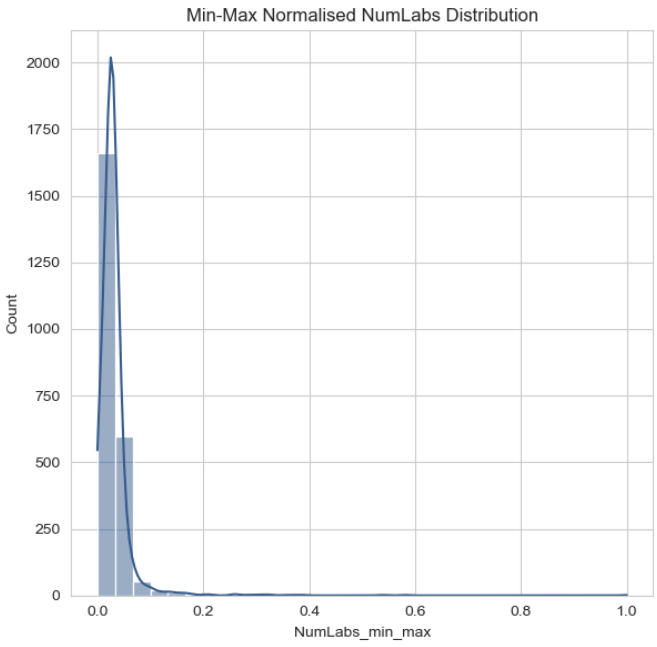
1C. Executive Summary

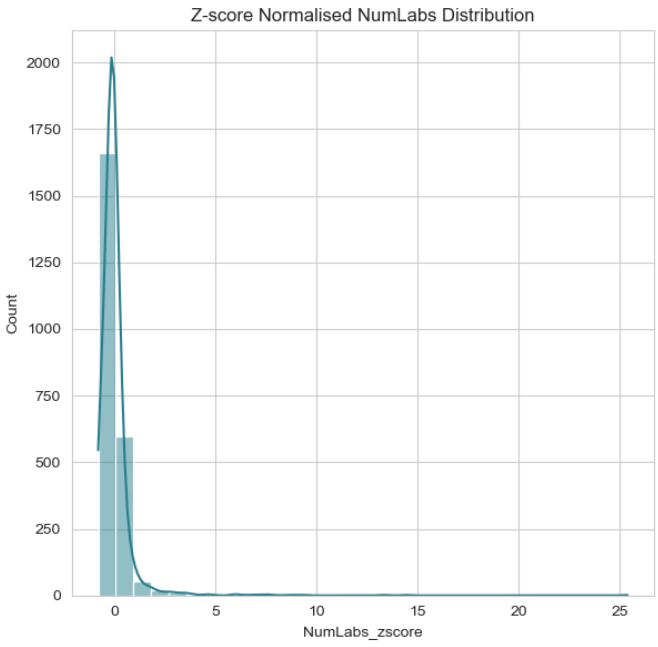
1C1. Key Findings

Demographic Distribution:
The majority of the patients in the dataset fall within the age brackets of 50s, 60s, and 70s. Newborns, especially those with premature births, also constitute a significant fraction of the dataset.

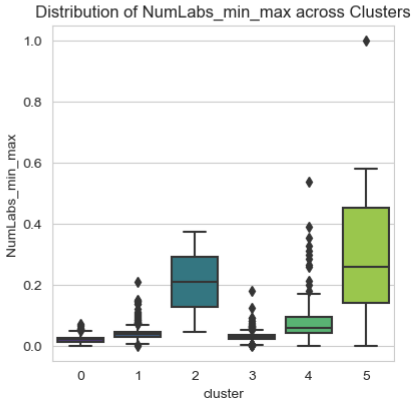


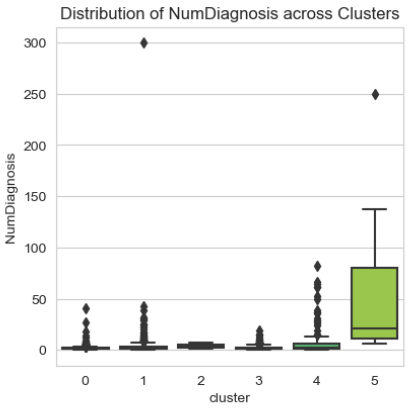
Attribute Distributions and Transformations:
The NumLabs attribute showed a massive positive skew due to outliers, which could be indicative of patients with complex medical needs or potential miscommunication in healthcare procedures.





Upon delving deeper into the key patient groups that might underlie the observed skews, a notable pattern emerges: the frequency of labs requested is paralleled by the number of diagnoses. This suggests that lab requests are judiciously made and often result in identifying additional conditions in the patients they're administered to. Clustering analysis has pinpointed two significant patient groups, namely Cluster 4 and Cluster 5, that appear to be intricately connected to this trend. The accompanying figures underscore their correlation with lab counts.



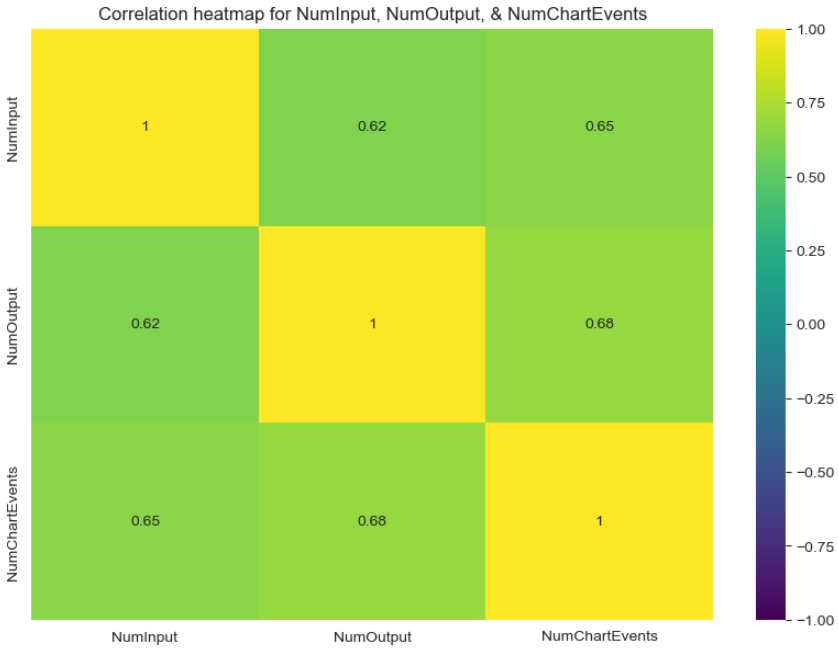


Further analysis should look at segmenting patient demographics and integrating additional features which could allow insights into health conditions these patients face, their potential risk factors, and importantly, provide actionable data for informed resource allocation decisions by the client.

Notable Associations:

A strong positive correlation exists between NumDiagnosis & NumProcs, indicating that as the number of diagnoses for a patient rises, the number of procedures they undergo also increases.

NumTransfers is positively correlated with both NumDiagnosis & NumProcs, suggesting patients requiring more diagnoses and procedures often need more transfers.



Several features, like NumInput, NumOutput, & NumChartEvents, are highly intercorrelated, suggesting an association between medications/fluids given and monitored outputs/events.

Data Correlation Insights:

Pre-processing didn't significantly alter the correlation strengths between attributes, suggesting the inherent relationships between the features remained consistent.

1C2. Interesting Patterns, Associations and Recommendations

Patient Demographics:

The dataset appears to lack representation from non-binary or other gender identities.

Recommendation: Engage with the client to ensure the inclusion of diverse gender identities to provide comprehensive care to all patients.

Admission Insights:

Patients were admitted from only 8 unique locations. This information could be utilised for understanding patient demographics and potential resource allocation.

The dataset has 1,083 unique diagnoses for a total of 2,359 patients. This could indicate a vast variety of conditions treated or potential limitations in diagnostic capabilities.

Recommendation: Dive deeper to understand if there's a need for specialisation or expansion in the diagnosis spectrum.

Resource Allocation:

The strong correlation between certain attributes like NumDiagnosis, NumProcs, and NumTransfers can aid in predicting resource requirements.

Recommendation: Consider dynamic resource allocation models that take into account these correlations to optimise hospital operations.

Introduction

Objective of this report

The Analytics Unit of our company has received a dataset from a potential client in the healthcare sector. The primary goal of this report is to explore and understand this dataset, derive insights from its attributes, and identify potential associations, outliers, and patterns that might hold value for the client.

Scope of this report

This report offers:

- An analysis of 2,359 patient records from the provided dataset.
- Data preprocessing and transformation techniques applied to specific attributes for improved clarity and pattern recognition.
- Visualisations to highlight key data distributions and relationships.
- Potential insights that could inform strategies for hospital operations, efficient resource allocation, and areas warranting more detailed data collection or further investigation.

Tools and techniques used

The following tools were used to conduct the analysis in this report:

- Python (Version 3.11.3): As the primary programming language for data analysis.
- Pandas: Employed for data manipulation and analysis.
- Matplotlib & Seaborn: For data visualisation to better understand distributions, relationships, and patterns.
- Numpy: To perform numerical operations on the dataset.
- Sklearn: To apply data preprocessing techniques, including normalisation, binning, and discretisation. As well as to perform scaling and clustering analysis.

The relevant Jupyter notebook, original dataset provided by the client as well as the output CSV file created from this analysis is in the same submission as this report.

Background of the dataset

The dataset provided by the client contains 2359 records of patients that have been admitted to a hospital. Each record contains the following information for each patient:

Feature	**Description of Feature**
gender	Gender of patient (e.g., male, female, non-binary, etc).
age	Age of patient.
LOSdays	Length of admission to the hospital in days.
admit_location	Location of admission (External transfer from a different facility, Admitted from the Emergency Department, Referred from a physician in a community setting, etc)./
AdmitDiagnosis	Diagnosis at admission.
Insurance	Type of insurance.
NumCallouts	**The client did not go into detail in what this attribute specifically referred to.** **More clarification is required.**
NumDiagnosis	Number of diagnoses during admission.
NumProcs	Number of procedures during admission.
AdmitProcedure	Procedure undergone at admission.
NumCPTevents	Number of Clinical Procedural Terminology (CPT) events during admission. CPT codes are added to a patient's electronic medical record (eMR) to detail the exact medical, surgical, and diagnostic services provided to a patient.
NumInput	Number of inputs administered to the patient during admission. These include the amount of fluids, medications and nutrients given to a patient.
NumLabs	Number of pathology tests conducted during admission.
NumMicroLabs	Number of microbiology pathology tests conducted during admission.

NumOutput	Number of outputs from the patient. Likely related to the amount of fluids excreted by a patient or could be related to the data outputs from laboratory tests or imagery results.
NumTransfers	Number of times the patient has been transferred. The client has not specified whether this attribute refers only to internal transfers between a single hospital's departments or whether it also includes transfers to external health facilities.
NumChartEvents	Number of chart events documented on the patient's eMR. Chart events can include administered treatments, clinical notes, administered tests, etc.
ExpiredHospital	Whether the patient has passed away whilst admitted to the hospital.
TotalNumInteract	Number of interactions the patient has had with healthcare professionals. This may include clinician to patient interactions involving consultations, therapies, diagnostic tests, etc.
Marital status	The marital status of the patient.

1A. Initial Data Exploration

1A.1 Attribute Type Identification

Feature	Attribute	Comment
gender	Nominal	Categorical data; e.g., male, female, non-binary
age	Ratio	Continuous data with a true zero point; can represent ages as exact values with meaningful differences between them
LOSdays	Ratio	LOS stands for "Length of Stay"; this is continuous data with a true zero point and can represent the length of stay in days with exact values
admit_location	Nominal	Categorical data; e.g., ER, ICU, general ward
AdmitDiagnosis	Nominal	Categorical data representing different diagnosis types
Insurance	Nominal	Categorical data; e.g., private, government, uninsured
NumCallouts	Ratio	Number of callouts can be measured on a ratio scale with a true zero point
NumDiagnosis	Ratio	Number of diagnoses can be measured on a ratio scale with a true zero point
NumProcs	Ratio	Number of procedures can be measured on a ratio scale with a true zero point
AdmitProcedure	Nominal	Categorical data representing different procedure types
NumCPTevents	Ratio	Number of CPT (Current Procedural Terminology) events can be measured on a ratio scale with a true zero point
NumInput	Ratio	Continuous data representing a number of inputs; it has a true zero point
NumLabs	Ratio	Continuous data representing the number of laboratory tests; it has a true zero point
NumMicroLabs	Ratio	Continuous data representing the number of microbiology laboratory tests; it has a true zero point
NumOutput	Ratio	Continuous data representing a number of outputs; it has a true zero point

NumTransfers	Ratio	Continuous data representing the number of transfers; it has a true zero point
NumChartEvents	Ratio	Continuous data representing the number of chart events; it has a true zero point
ExpiredHospital	Nominal	Categorical data representing whether a patient expired in the hospital; e.g., yes, no
TotalNumInteract	Ratio	Continuous data representing the total number of interactions; it has a true zero point
Marital status	Nominal	Categorical data representing marital status; e.g., married, single, divorced

1A.2 Summary Properties

Summary Statistics

The dataset offers a comprehensive overview of patient metrics across various attributes. This section delves into the specifics of these metrics, shedding light on central tendencies, distributions, and potential data quality concerns.

	age	LOSdays	NumCallouts	NumDiagnosis	NumProcs	NumCPTevents	NumInput	NumLabs	NumMicroLabs	NumOutput	NumTransfers	NumChartEvents	ExpiredHospital	TotalNumInteract
count	2359.0	2359.0	2359.0	2359.0	2359.0	2359.0	2359.0	2359.0	2359.0	2359.0	2359.0	2359.0	2359.0	2359.0
mean	53.11	10.08	0.1	2.74	0.78	1.08	30.28	45.44	1.3	7.07	1.12	534.11	0.11	640.19
std	25.98	12.9	0.16	9.71	2.8	1.5	50.76	55.86	6.47	7.38	3.51	549.47	0.31	658.01
min	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
25%	42.0	3.69	0.0	0.83	0.22	0.08	4.82	27.41	0.15	1.69	0.38	210.72	0.0	268.27
50%	59.0	6.25	0.0	1.43	0.43	1.01	13.94	38.88	0.51	5.17	0.67	426.5	0.0	507.69
75%	73.0	11.67	0.16	2.41	0.71	1.56	34.87	51.49	1.36	10.39	1.07	711.5	0.0	826.98
max	88.0	144.67	2.38	300.0	75.0	35.29	600.0	1462.5	292.31	144.83	100.0	10710.38	1.0	10850.84

Age

- Average: 53 years
- Range: 0 - 88 years
- Insight: The data suggests a higher concentration of elderly patients, with 50% of all patients falling between 42 and 73 years.

LOSdays

- Average length of admission: 10 days.
- Interquartile range for length of admission is between 3 to 11 days.
- Longest amount of days admitted was 144 days. (Big outlier)

NumDiagnosis

- Average number of diagnoses was 2.7.
- Interquartile range of number of diagnosis was 0.8 to 2.4 per patient.
- Largest amount was 300. This is an outlier, possibly representing a patient with a complex disorder?

Data Quality Concerns:

A potential data quality concern arises from the presence of floating-point values in attributes traditionally expected to be integers. It's vital to consult with the client to clarify the nature and intention behind these entries.

A list of attributes that demonstrate this issue are:

Attribute	**Number of Records with Float Values**
NumDiagnosis	2308
NumProcs	2075
NumCPTevents	1751
NumLabs	2264
NumMicroLabs	1920
NumOutput	2033
NumTransfers	2317
NumChartEvents	2214
TotalNumInteract	2309

Frequency

The table below represents the unique count of values for each feature in the dataset:

Feature	**Unique Values**
gender	2
age	73
LOSdays	626
admit_location	8
AdmitDiagnosis	1083
Insurance	5
NumCallouts	82
NumDiagnosis	568
NumProcs	252
AdmitProcedure	378
NumCPTevents	341
NumInput	1752
NumLabs	1941
NumMicroLabs	430
NumOutput	1221
NumTransfers	265
NumChartEvents	2265
ExpiredHospital	2
TotalNumInteract	2333
Marital status	6

Comments Regarding Frequency

gender

The dataset identifies only two genders, suggesting binary classifications of male or female. This raises a concern about the inclusivity of patients with non-binary or other gender identities. Hospitals should be supportive environments, ensuring patient comfort to facilitate recovery. Discussing the gender identification system with the client could lead to more inclusive patient data collection in the future.

Admit_location

Only 8 unique admit locations are identified over the dataset's duration. This information is crucial for understanding patient demographics and optimising resource allocation, potentially indicating which facilities have the highest demand or priority.

AdmitDiagnosis

The dataset highlights 1083 unique diagnoses from a total of 2359 patients. This high variability suggests a broad scope of conditions treated at the hospital. However, the diversity might also indicate possible inconsistencies in diagnostic criteria or documentation. A deep dive into these diagnoses can provide more accurate insights into the hospital's treatment capacities or potential areas of improvement.

Data Pre-Processing

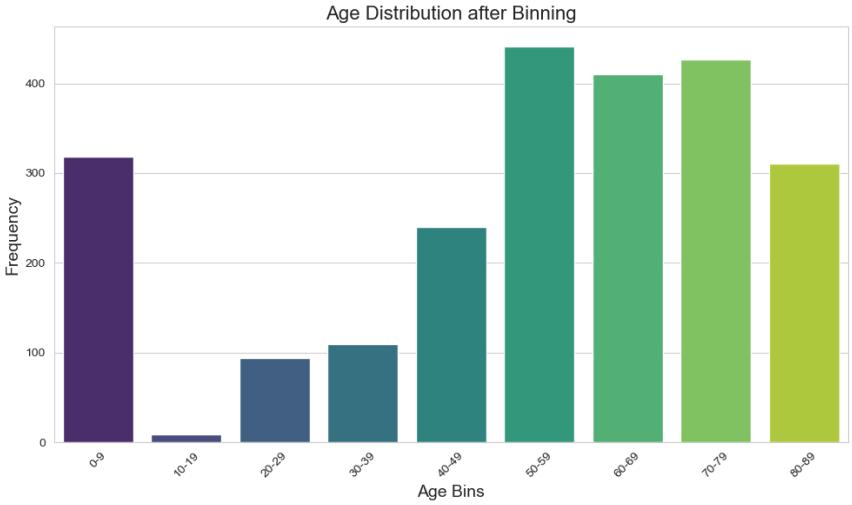
Some attributes in the dataset are continuous and contain a significant number of unique values. These attributes include:

- Age
- NumLabs
- LOSDays
- Marital Status

For more actionable insights and clearer visual representations, we applied specific pre-processing techniques to these attributes. The detailed methodologies and associated code are available in the report's Appendices.

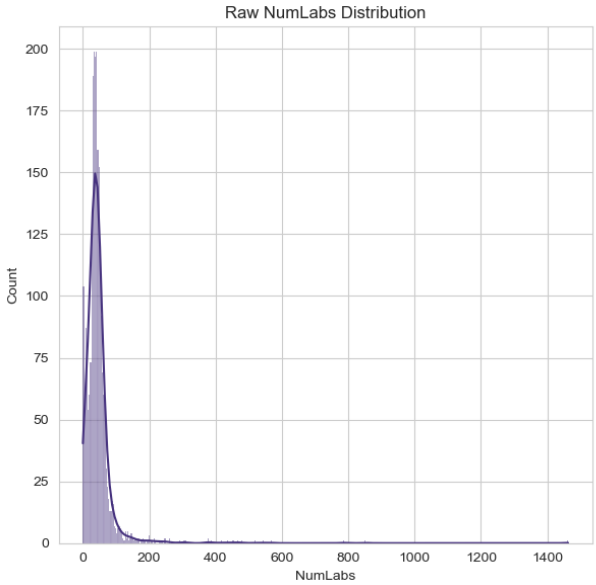
Binning the “Age” Attribute

Post-binning, the histogram clearly illustrates that; Individuals in their 50s (i.e., the 50-59 bin), 60s, and 70s are predominant in the dataset. Whilst newborns and people in their 80s follow closely in frequency.



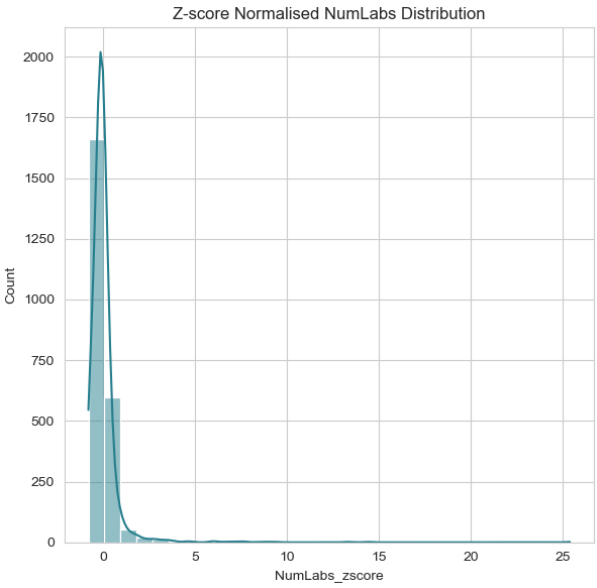
“NumLab” Attribute Normalisation

The NumLabs attribute was normalised, evident from the Min-Max and Z-score Normalised Distribution figures above. This decision stemmed from the pronounced positive skew caused by 139 outliers surpassing the upper IQR boundary, as depicted in the "Raw NumLabs Distribution" figure.



The outliers could represent patients with intricate medical needs, warranting multiple diagnostic tests. Conversely, these outliers might also highlight potential communication breakdowns among healthcare professionals, leading to excessive test orders. Future inquiries in this direction might encompass:

- Reviewing the admit_diagnosis.
- Analysing the length of stay.
- Investigating the number of diagnoses.



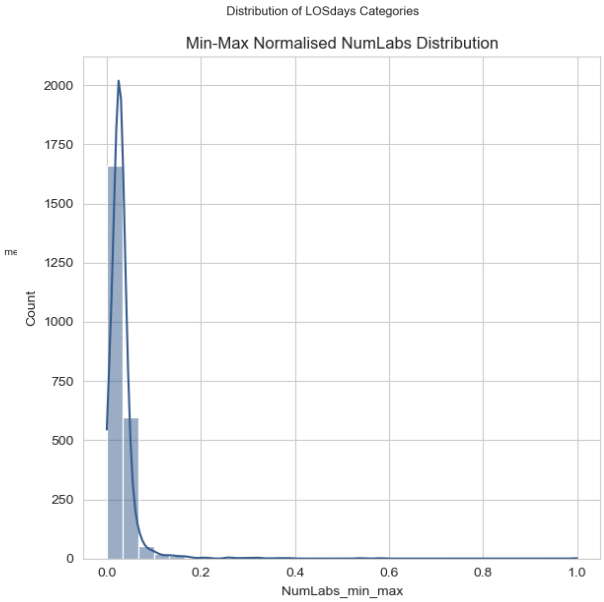
Regardless of the outlier interpretation, normalising the NumLabs values remains essential for subsequent clustering due to the skew magnitude introduced by these outliers.

“LOSdays” Attribute Discretisation

We discretised the LOSdays attribute using the following criteria:

- Shorter periods: 0 - 5 days stay
- Medium periods: 5 - 15 days stay
- Longer periods: 15 - 50 days stay
- Very long periods: 50+ days stay

Comment: A significant proportion of patients have shorter and medium stays. Specifically, shorter stays range from 1 to 5 days, while medium stays span 5 to 15 days.

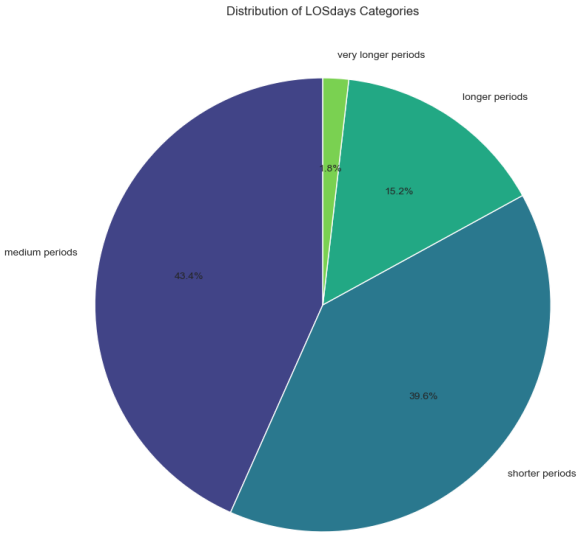


“LOSdays” Attribute Discretisation

We discretised the LOSdays attribute using the following criteria:

- Shorter periods: 0 - 5 days stay
- Medium periods: 5 - 15 days stay
- Longer periods: 15 - 50 days stay
- Very long periods: 50+ days stay

Comment: A significant proportion of patients have shorter and medium stays. Specifically, shorter stays range from 1 to 5 days, while medium stays span 5 to 15 days.

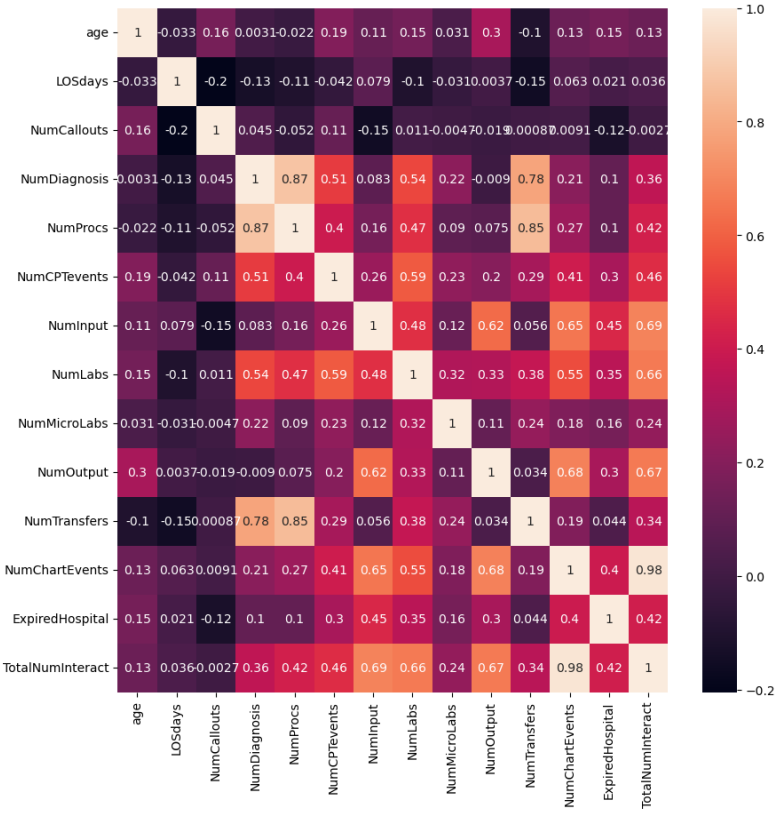


1A.3 Dataset Exploration

Raw Correlations between attributes before preprocessing

Legend:

- **Lighter Shades**: Indicate higher positive correlations.
- **Dark Purple**: Represents negligible or no correlation.
- **Shades of Dark to Black**: Indicate negative correlations.



Reference for Correlation Coefficients:

- **0.00 to 0.19**: Very weak correlation.
- **0.20 to 0.39**: Weak correlation.
- **0.40 to 0.59**: Moderate correlation.
- **0.60 to 0.79**: Strong correlation.

Attributes with Strong Correlation

NumDiagnosis & NumProcs: The relationship between these attributes is noteworthy, displaying a positive correlation (0.874762). This implies that as the number of diagnoses for a patient rises, so does the count of medical procedures they undergo.

NumTransfers, NumDiagnosis & NumProcs: A notable observation is that patients with a higher frequency of transfers tend to have a surge in both diagnoses and procedures. This may infer that intricate cases, which require a larger number of procedures and diagnoses, often necessitate multiple transfers.

```
Strongly Correlated Pairs:
NumDiagnosis        NumProcs            0.874762
                    NumTransfers        0.781260
NumProcs            NumDiagnosis        0.874762
                    NumTransfers        0.852218
NumInput            NumOutput           0.623768
                    NumChartEvents      0.651933
                    TotalNumInteract    0.686398
NumLabs             TotalNumInteract    0.662861
NumOutput           NumInput            0.623768
                    NumChartEvents      0.680269
                    TotalNumInteract    0.665836
NumTransfers        NumDiagnosis        0.781260
                    NumProcs            0.852218
NumChartEvents      NumInput            0.651933
                    NumOutput           0.680269
                    TotalNumInteract    0.977825
TotalNumInteract    NumInput            0.686398
                    NumLabs             0.662861
                    NumOutput           0.665836
                    NumChartEvents      0.977825
                    NumLabs_min_max     0.662861
                    NumLabs_zscore      0.662861
NumLabs_min_max     TotalNumInteract    0.662861
NumLabs_zscore      TotalNumInteract    0.662861
```

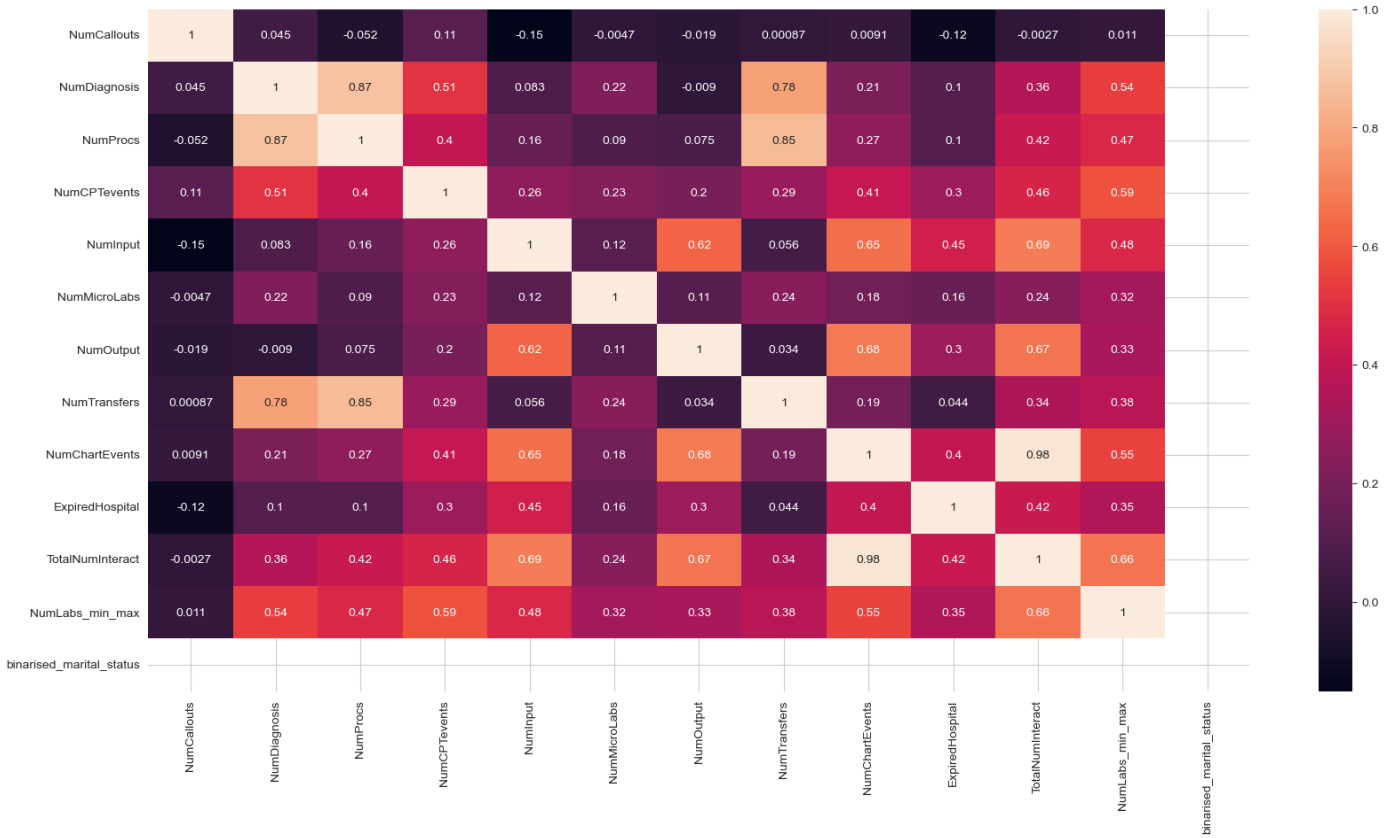
NumInput, NumOutput, & NumChartEvents: These attributes are interlinked. This suggests that an increment in inputs (like medications and fluids) often leads to a heightened monitoring of outputs (such as urine output and drainage). Simultaneously, there seems to be an increase in charted events.

TotalNumInteract: This attribute shares a very strong correlation with numerous attributes, with NumChartEvents being the most significant (correlation coefficient of 0.977825).

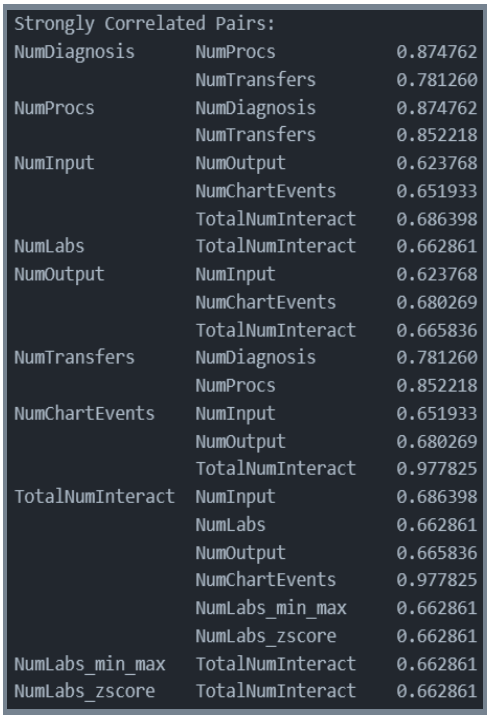
Correlations between attributes after preprocessing

The following correlation analysis uses columns from our refined dataset, after the preprocessing steps:

```
Index(['gender', 'admit_location', 'AdmitDiagnosis', 'insurance',
       'NumCallouts', 'NumDiagnosis', 'NumProcs', 'AdmitProcedure',
       'NumCPTevents', 'NumInput', 'NumMicroLabs', 'NumOutput', 'NumTransfers',
       'NumChartEvents', 'ExpiredHospital', 'TotalNumInteract',
       'age_equiwidth_bin', 'NumLabs_min_max', 'LOSdays_category',
       'binarised_marital_status'],
      dtype='object')
```

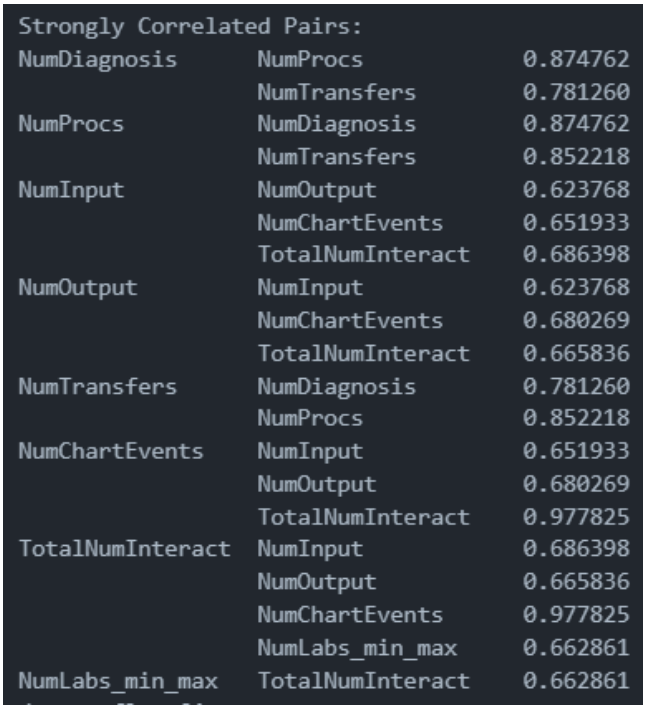


For clarity, we provide a comparative visualisation: on the left, we depict correlations from the raw dataset; on the right, the correlations post preprocessing.



Strongly Correlated Pairs:

NumDiagnosis	NumProcs	0.874762
	NumTransfers	0.781260
NumProcs	NumDiagnosis	0.874762
	NumTransfers	0.852218
NumInput	NumOutput	0.623768
	NumChartEvents	0.651933
	TotalNumInteract	0.686398
NumLabs	TotalNumInteract	0.662861
NumOutput	NumInput	0.623768
	NumChartEvents	0.680269
	TotalNumInteract	0.665836
NumTransfers	NumDiagnosis	0.781260
	NumProcs	0.852218
NumChartEvents	NumInput	0.651933
	NumOutput	0.680269
	TotalNumInteract	0.977825
TotalNumInteract	NumInput	0.686398
	NumLabs	0.662861
	NumOutput	0.665836
	NumChartEvents	0.977825
	NumLabs_min_max	0.662861
	NumLabs_zscore	0.662861
NumLabs_min_max	TotalNumInteract	0.662861
NumLabs_zscore	TotalNumInteract	0.662861



Strongly Correlated Pairs:

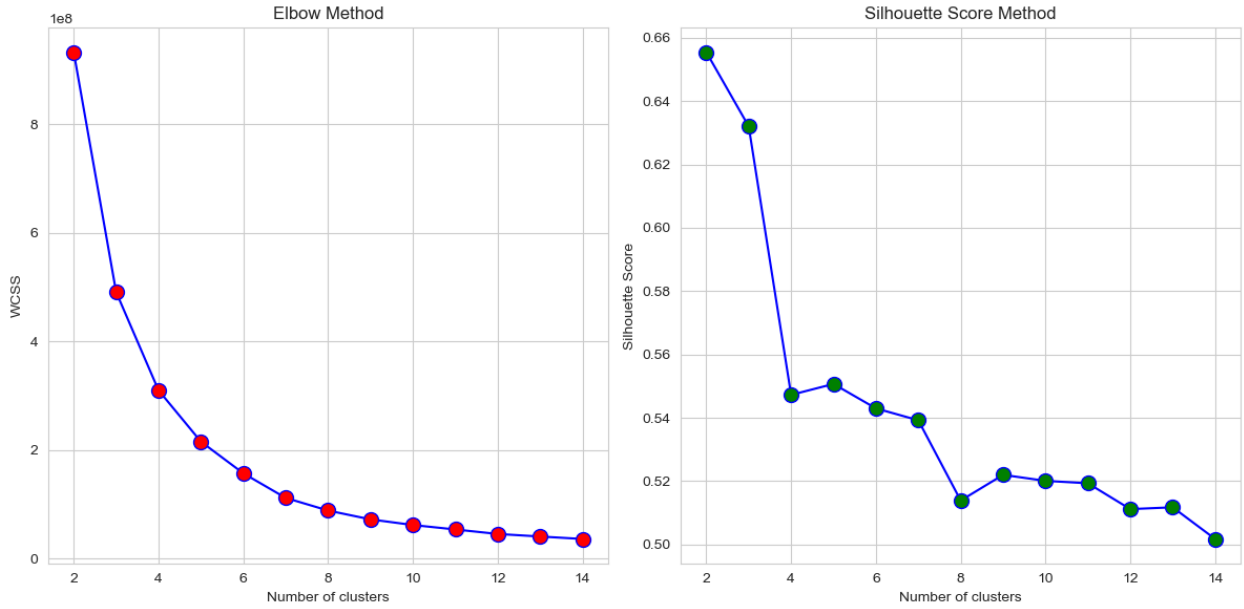
NumDiagnosis	NumProcs	0.874762
	NumTransfers	0.781260
NumProcs	NumDiagnosis	0.874762
	NumTransfers	0.852218
NumInput	NumOutput	0.623768
	NumChartEvents	0.651933
	TotalNumInteract	0.686398
NumOutput	NumInput	0.623768
	NumChartEvents	0.680269
	TotalNumInteract	0.665836
NumTransfers	NumDiagnosis	0.781260
	NumProcs	0.852218
NumChartEvents	NumInput	0.651933
	NumOutput	0.680269
	TotalNumInteract	0.977825
TotalNumInteract	NumInput	0.686398
	NumOutput	0.665836
	NumChartEvents	0.977825
	NumLabs_min_max	0.662861
NumLabs_min_max	TotalNumInteract	0.662861

Overall, there does not appear to be any significant difference between the correlations of our attributes before and after preprocessing. Therefore, the preprocessing didn't notably alter the inherent relationships between the attributes.

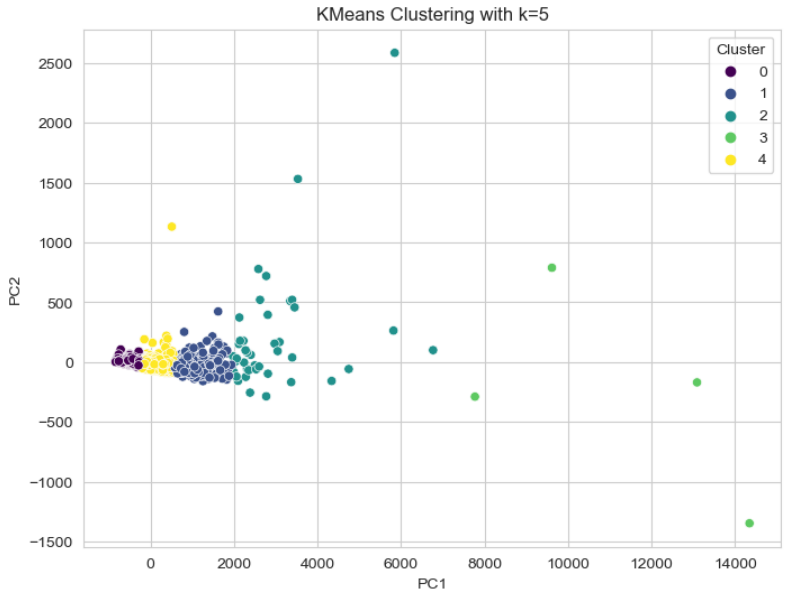
Cluster Analysis

To further explore the nature of our dataset and the correlations between values, K-means clustering was performed in an effort to gain insight on the predominant groups of patients.
A quick overview of our method involved:

- Utilising one-hot encoding for the following categorical columns: gender, admit_location, AdmitDiagnosis, insurance, AdmitProcedure
- Applying the Elbow Method and Silhouette Score Method to determine the amount of K-values



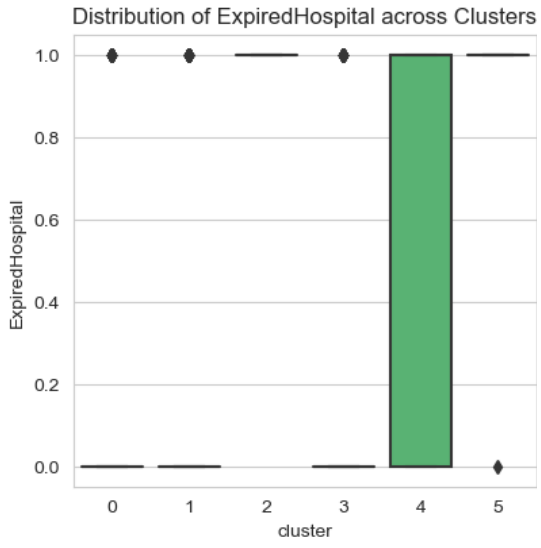
- Elbow method indicated K = 4 to be optimal whilst the Silhouette score indicated K =2 to be optimal. Due to the heterogeneity of our dataset, we decided to use the K = 3 value from the Silhouette score. This left us with K = 4 and K = 3 as our optimal K-values from these two methods.
- Upon visualising what the clusters would look like as K =4, K = 3 and for the sake of exploration, K = 2, K =5 and K =6, the following was observed.
- Given the level of overlap from the other K-values, K = 5 was determined to be the most appropriate value.



Having established 5 clusters within our dataset, approximately 4 clusters do appear to be prominent with cluster #3 being composed of still present outlier values.
Nevertheless, clusters 0, 1, 2 and 4 seem to offer reasonable levels of overlap and separation.
A series of plots were generated (full view in Appendix 2.1) visualising the distribution of attributes across each of our 5 clusters.

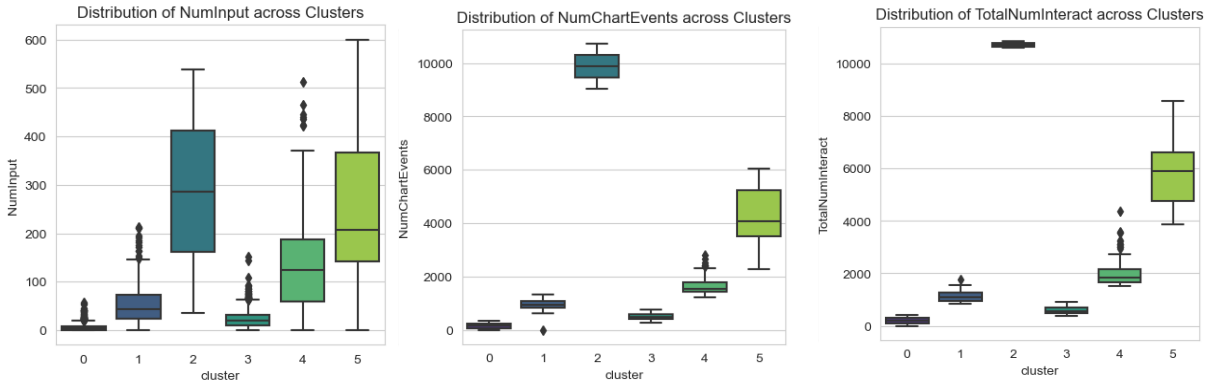
Cluster Analysis of ExpiredHospital Attribute

What stood out immediately was for patient deaths (ExpiredHospital attribute), a very large majority was distributed into cluster 4.



From our earlier heatmap visual, ExpiredHospital did not have any significantly strong correlations, however it did have two moderate correlations which was 0.45 with *NumInput*, 0.4 with *numChartevents* and 0.42 with *TotalNumInteract*.

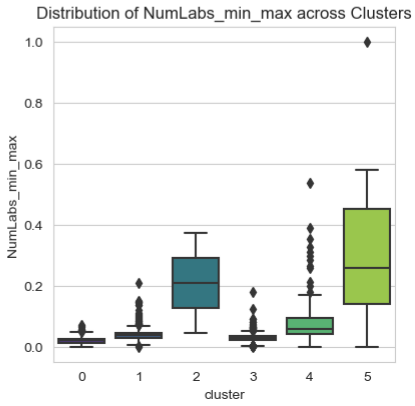
Exploring the distribution of the above attributes with cluster 4 shows:



The distribution of NumINput, NumChartEvents and TotalNumInteract between the clusters do not exactly show Cluster 4 as significant above other clusters or as an outlier. So it is not possible to infer from the correlation and cluster distributions alone what factors may be contributing to Cluster 4 having the largest distribution of deaths.

It may very well be possible that our K-means clustering algorithm created a cluster specifically to describe patient deaths, thus meaning Cluster 4 may be a specific “expired” cluster.

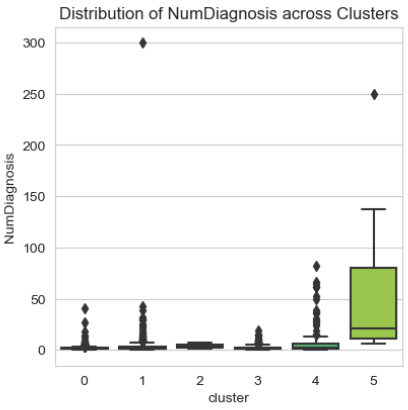
Cluster Analysis of NumLabs Attribute



The NumLabs attribute has shown interesting activity with its outliers from earlier distribution and heatmap analysis, which have led to its marked positive skew.

Our cluster analysis of NumLabs shows that Clusters 5 and 4 are the two groups of patients with the largest counts of laboratory tests, which may indicate these two groups to be responsible for the massive skew observed earlier.

To glean deeper insights from these clusters, a focused analysis on NumDiagnosis is done below.



When considering NumDiagnosis, one may discern a potential correlation between the volume of lab tests and the breadth of diagnosed conditions. A closer examination of Clusters 5 and 4 in the context of NumDiagnosis has shown some distinct patterns. Cluster 5 is characterised by a pronounced concentration within its interquartile range (IQR). Meanwhile, Cluster 4 is distinguished by a significant presence of outliers. When benchmarked against other clusters, both Clusters 5 and 4 show elevated counts for the previously mentioned attributes.

Appendices

1B. Data Preprocessing

1B1. Binning the “Age” Attribute

Analysis of Age Distribution prior to Binning

	Value
count	2359.0
mean	53.11
std	25.98
min	0.0
25%	42.0
50%	59.0
75%	73.0
max	88.0

Descriptive Statistics of Age prior to Binning

Comment:

- Age centralises around the age of 53 years old.
- The dataset is divided in two at 59 years old. Half are older and half are younger.
- Most patients are between the ages of 42 and 73 years old.

Who is our largest age group?

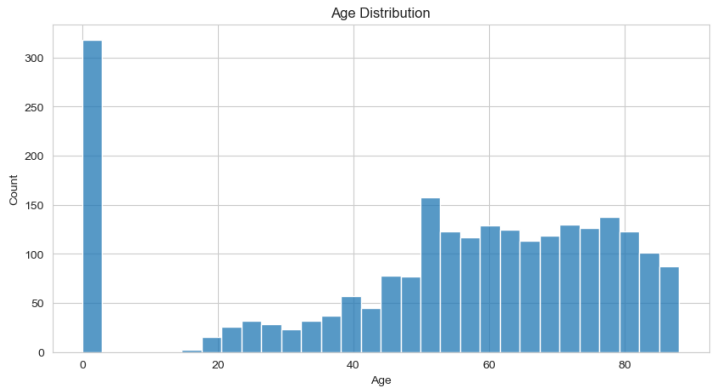
Mode: 0

Comment:

Newborns are our largest age group!

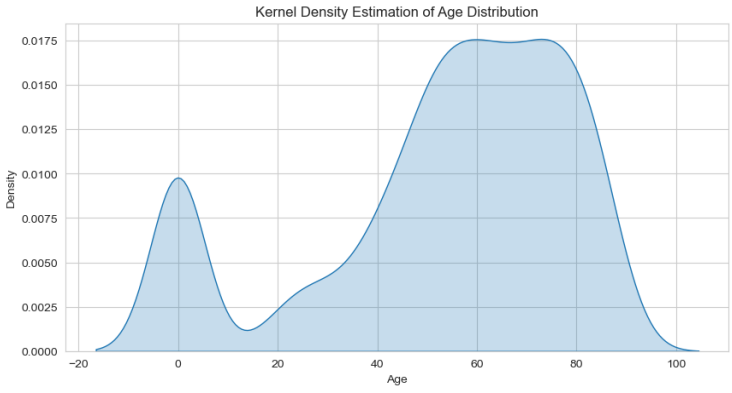
Outliers

Number of outliers detected: 0



Age Distribution

- Histogram: Newborns make up our largest number of people. However the remainder of our population is bulked as elder/old people. Meanwhile teenagers and older adults make up the minority of our population.
- KDE: Patients at this hospital are mostly newborns. middle aged and elderly people. Young adults make up the minority of the patients in the hospital.



Analysis of Age after Equi-width Binning

A quick note on justifying our binning

- There was a clear peak at age 0, didn't need binning to interpret that
- We do know there is a concentration also among older people, however we don't exactly have a clear interpretation on which "sub-group(s)" of older people has the largest concentration(s)

So let's bin our ages by decades so we can get answers such as (for example) "people in 70's are our common elder population" as opposed to age intervals.

Creating our Age bins

```
# Smoothing with Equi-wdith binning

# Bins are by decades, the range is from our minimum age of 0 to 88
(rounding up to 90 for cleanliness)
bins = list(range(0, 91, 10))  # Creates a list [0, 10, 20, ..., 90]

labels = ['0-9', '10-19', '20-29', '30-39', '40-49', '50-59', '60-69',
'70-79', '80-89']

# Creating a binned attribute into our dataframe
df['age_bin'] = pd.cut(df['age'], bins=bins, labels=labels, right=False)
# right=False makes the interval left-closed

# Displaying the original age and its corresponding bin
print(df[['age', 'age_bin']].head())
```

Output:

```
   age age_bin
0   65   60-69
1   84   80-89
2   38   30-39
3   78   70-79
4   58   50-59
```

Checking distribution among bins

```
# Lets see how many people are in each bin
age_bin_counts = df['age_bin'].value_counts().sort_index().reset_index()
age_bin_counts.columns = ['Age Bin', 'Count']
print(age_bin_counts)
```

Output:

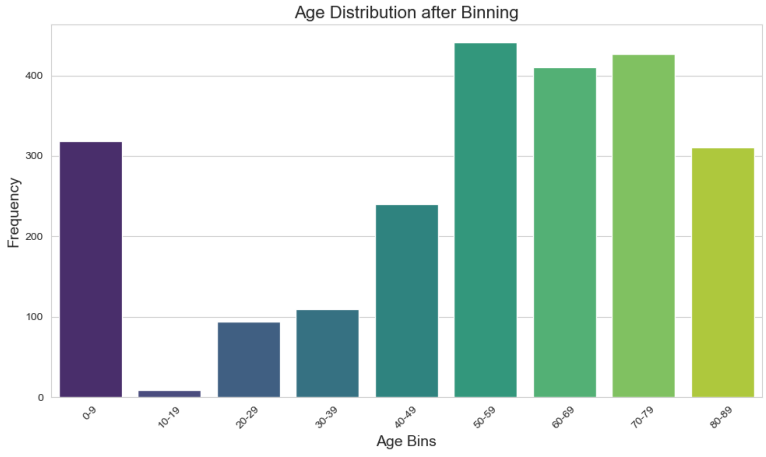
	Age Bin	Count
0	0-9	318
1	10-19	9
2	20-29	94
3	30-39	109
4	40-49	240
5	50-59	441
6	60-69	410
7	70-79	427
8	80-89	311

Now that we've established our equi-width binning rationale, let us now see how the graphs look now.

Comments:

So after binning and making another histogram (i.e. bar graph), we can see the following:

- People in their 50's (i.e. the bin that says 50-59), their 60's and 70's make up the majority of the people in our dataset
- Followed closely by newborns and people in their 80's



Analysis of Age after Equi-depth Binning

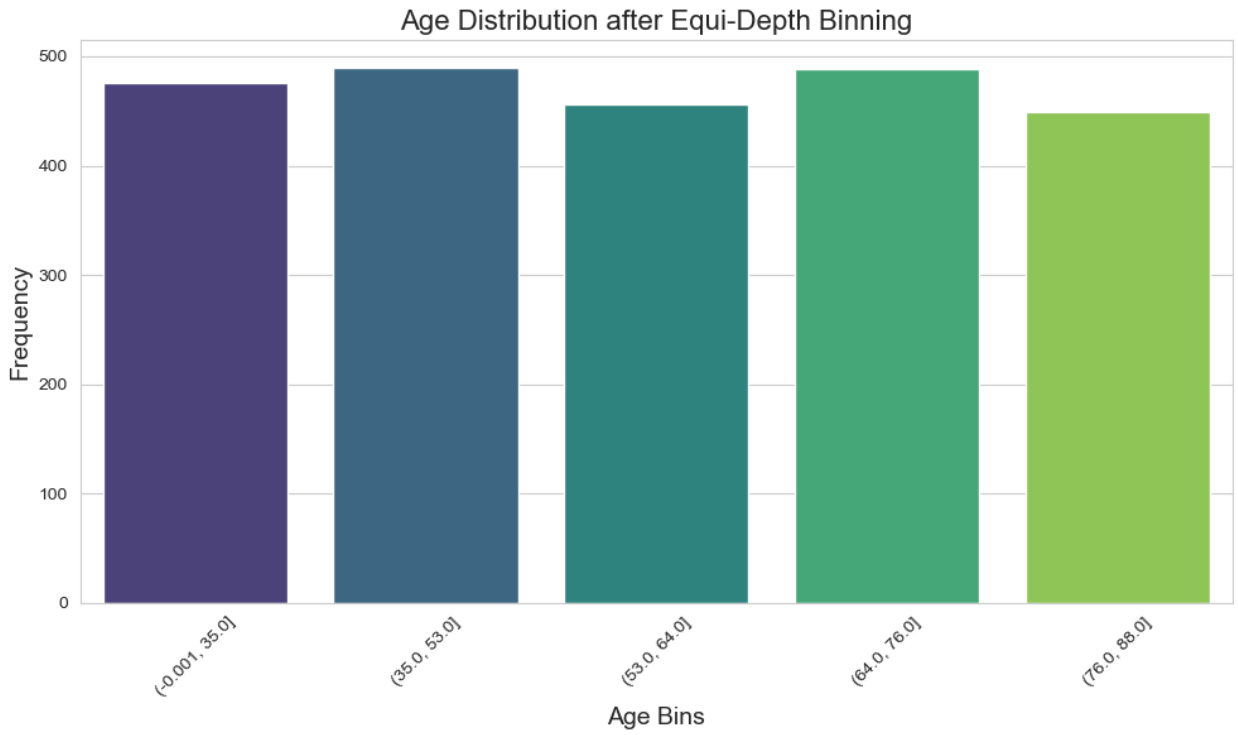
Code for Equi-depth Binning

```
# Smoothing with Equi-depth binning

# Creates equi-depth binned columns
df['age_equidepth_bin'] = pd.qcut(df['age'], q=5)

# Aggregate and sort the data
age_equidepth_counts =
df['age_equidepth_bin'].value_counts().sort_index().reset_index()
age_equidepth_counts.columns = ['Age Bin', 'Count']
```

Result was not as helpful as compared to Equi-width Binning:



1B2. “NumLabs” Attribute Normalisation

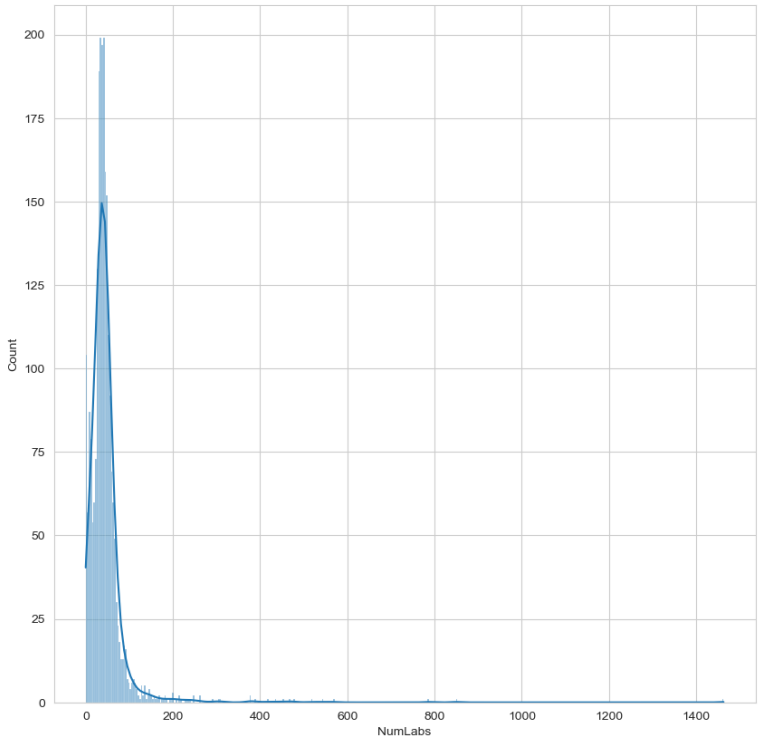
Before we normalise NumLabs, let's have a look at the scale we are dealing with.

```
# Summary stats
description = df['NumLabs'].describe()
print(description)
```

```
count    2359.000000
mean       45.438588
std        55.862703
min         0.000000
25%        27.410000
50%        38.880000
75%        51.490000
max      1462.500000
Name: NumLabs, dtype: float64
```

Visualisation of our raw NumLabs distribution

Massive positive skew here! Our max value is 1462 which must be an outlier. Lets further check if there are any other outliers present.



```python
# Checking for outliers using IQR
Q1 = df['NumLabs'].quantile(0.25)
Q3 = df['NumLabs'].quantile(0.75)
IQR = Q3 - Q1

# Bounds for the outliers
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

# Outliers
outliers = df[(df['NumLabs'] < lower_bound) | (df['NumLabs'] >
upper_bound)]

# Count the lower and upper outliers
lower_outliers_count = df[df['NumLabs'] < lower_bound].shape[0]
upper_outliers_count = df[df['NumLabs'] > upper_bound].shape[0]
```

```
Q1 (25th Percentile): 27.41
Q3 (75th Percentile): 51.489999999999995
IQR: 24.079999999999995
Lower Bound for Outliers: -8.709999999999999
Upper Bound for Outliers: 87.60999999999999

Outliers:
Number of Lower Outliers: 0
Number of Upper Outliers: 139
6         99.64
16       112.33
23        90.48
26       155.56
31        88.83
           ...
2243     114.93
2276     104.91
2301     109.17
2316     116.17
2323     109.02
Name: NumLabs, Length: 139, dtype: float64
```

So we have 139 records that are above the upper IQR range.
Due to the nature of our data, our 139 outliers are likely representative of patients having medical cases with complex needs and thus requiring more diagnostic tests.
However it can also be indicative of miscommunication between healthcare professionals resulting in over-ordering of tests.

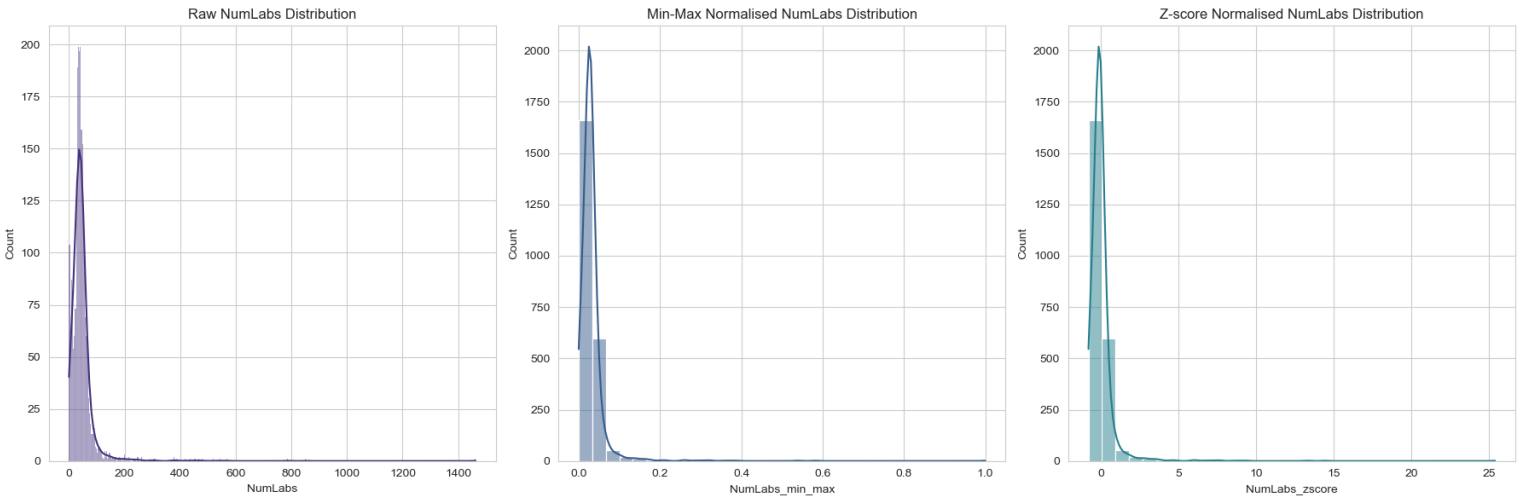
Given we have 139 outliers above the upper boundary of our IQR yet the majority of our 2000+ patients are within the IQR, this would affect any clustering investigations we do, thus we will have to normalise numLabs prior to clustering.
If we visualise data with attributes of differing magnitudes in a scatter plot, the axis with the larger magnitude might stretch out, making clusters or patterns in the other attributes harder to discern.

Normalising with via Min-Max and Z-Scores

Here we will use the respective Sklearn scalers to conduct our normalisations. Whilst we could normalise manually with pandas, using the scalers from Sklearn will be beneficial for when we want to use other functions within Sklearn such as for Clustering.

```
from sklearn.preprocessing import MinMaxScaler, StandardScaler
# SkLearn Min-Max Scaler
min_max_scaler = MinMaxScaler()
df['NumLabs_min_max'] = min_max_scaler.fit_transform(df[['NumLabs']])
#SkLearn Z score scaler
standard_scaler = StandardScaler()
df['NumLabs_zscore'] = standard_scaler.fit_transform(df[['NumLabs']])
```

Raw numLabs distribution vs Normalised by Min-Max vs Normalised by Z-Score



1B3. “LOSdays” Attribute Discretisation

```
bins = [0, 5, 15, 50, float('inf')]
labels = ['shorter periods', 'medium periods', 'longer periods', 'very
longer periods']

df['LOSdays_category'] = pd.cut(df['LOSdays'], bins=bins, labels=labels,
right=False)

frequency = df['LOSdays_category'].value_counts()
```

Discretisation is fairly simple, once the number of bins was decided (as specified in the task specifications), *pandas.cut* function can be applied to the attribute of interest, resulting in values being labeled based on the bins we’ve set.

1B.4 “Marital Status” Attribute Binarisation

```
# Binarise the 'marital status' column
df['binarised_marital_status'] = df['marital status'].apply(lambda x: 1 if
x == "MARRIED" else 0)
```

Binarisation is also fairly simple, where we use a if statement-esque comparison operator with the help of the *pandas.DataFrame.apply* function applied to our entire feature of interest and set all values to 1 if the value in a given record is “MARRIED " and 0 if not.

2.1. Full Cluster Distribution