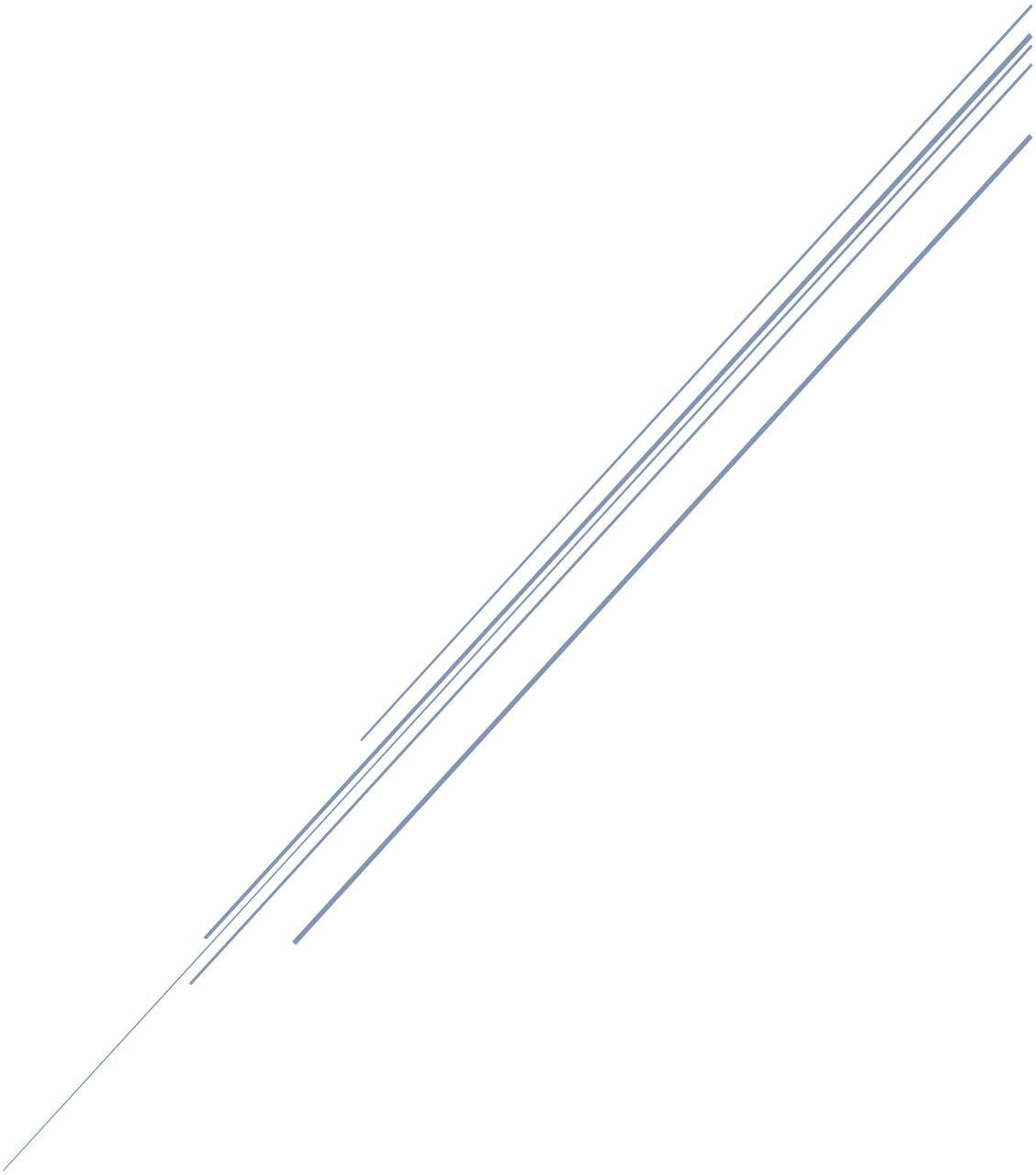
ASSESSMENT TASK 2

Data Exploration and Preparation



Zhitan Wu - University of Technology, Sydney

31250 – Intro to Data Analytics

Table of Contents

1A INITIAL DATA EXPLORATION (ALL COMPLETED WITH PYTHON) 2

ATTRIBUTE TYPES 2
1. GENDER 2
2. AGE 3
3. LOSDAYS 5
4. ADMIT_LOCATION 6
5. ADMITDIAGNOSIS 7
6. INSURANCE 9
7. NUMCALLOUTS 10
8. NUMDIAGNOSIS 13
9. NUMPROCS 15
10. ADMITPROCEDURE 17
11. NUMCPTEVENTS 18
12. NUMINPUT 21
13. NUMLABS 23
14. NUMMICROLABS 25
15. NUMOUTPUT 28
16. NUMTRANSFERS 30
17. NUMCHARTEVENTS 32
18. EXPIREDHOSPITAL 35
19. TOTALNUMINTERACT 36
20. MARITAL STATUS 38

EXPLORING DATASET 40
Outliers *40*
Clusters of Similar Instances *47*
Interesting Attributes and Relations *50*

1B DATA PRE-PROCESSING (ALL COMPLETED WITH PYTHON) 53

BINNING TECHNIQUES 53
Equi-Width Binning *53*
Equi-Depth Binning *55*
NORMALISING ATTRIBUTES 57
Min-Max *57*
Z-Score(s) *59*
DISCRETISING ATTRIBUTES 61
BINARIZING ATTRIBUTES 63

1C SUMMARY 65

SUMMARISATION OF FINDINGS 65

1A Initial Data Exploration (All Completed with Python)

Attribute Types

This is the full list of attribute types that are present in the given dataset, with their type and justifications that relate to their specific scenario:

1. **Gender**

   Attribute Type: Nominal (categorical).

   Justifications: The gender of patients can be matched to the characteristics of nominal or categorical attributes. Although patients can be classified as being from one of two genders, male and female, and no numerical values are involved, there is no set order for how this will occur. Information such as this allows doctors to monitor how certain diseases may be more prevalent in one gender over the other as a result of biological differences that can alter how a disease manifests and treatments work from hormone levels.

   Summarising Properties and Visualisations:

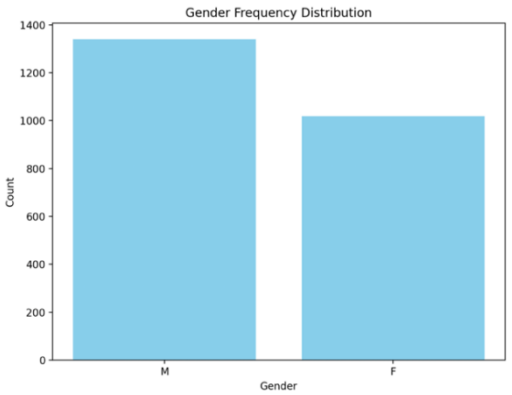


Figure 1: Frequency Distribution (Bar Plot)

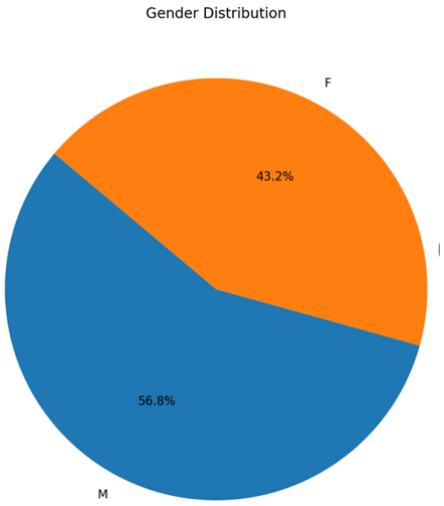


Figure 2: Percentage Distribution (Pie Chart)

Title	Values
Frequency Distribution	Male: 1340, Female: 1019 – As seen in Figure 1
Percentage Distribution	Male: 56.8% (1.d.p), Female: 43.2% (1.d.p) – As seen in Figure 2
Mode	Male – As seen in Figures 1 & 2
Count of Unique Categories	2; ‘Male’ and ‘Female’

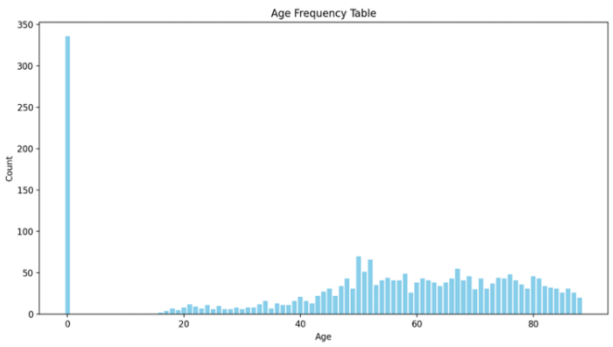
Table 1: Summary Properties (Based on the ‘Gender’ attribute)

2. **Age**

Attribute Type: Ratio.

Justifications: The age of patients in the hospital database is generally regarded as a ratio data type. This is because there's the presence of meaningful intervals, for example, the difference between 20 and 30 years old is conceptually the same as the difference between 40 and 50 years old. If someone is 0 years old in a hospital database for healthcare, it means that they are a newborn and haven't quite reached their first birthday. In this case, there is a true 'zero' point although it alters with time and introduces complexities when comparing values across different time periods or dealing with historical data.

Summarising Properties With Visualisations:



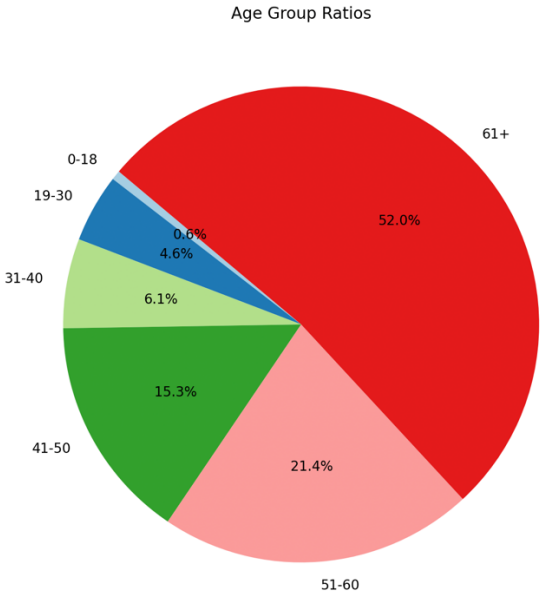


Figure 3: Frequency Distribution (Histogram) *Figure 4: Percentage Distribution (Pie Chart)*

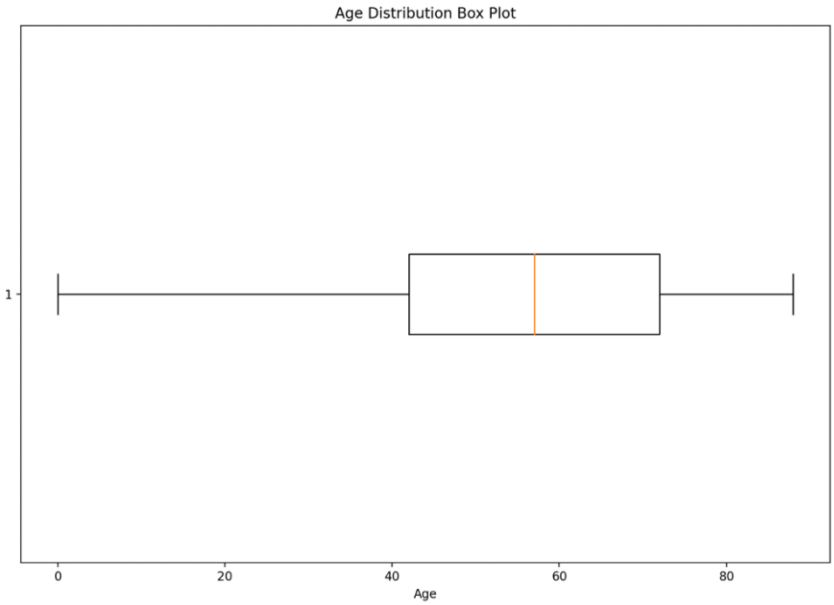


Figure 5: Age Distribution (Box-and-Whisker Plot)

0	336	14.243323
16	2	0.084782
17	4	0.169563
18	7	0.296736
19	5	0.211954
20	8	0.339127
21	12	0.508690
22	9	0.381518
23	7	0.296736
24	11	0.466299
25	6	0.254345
26	10	0.423908
27	6	0.254345
28	6	0.254345
29	8	0.339127
30	6	0.254345
31	8	0.339127
32	8	0.339127
33	12	0.508690

Figure 6: Part of The Output From Python – First Column (on the left): Ages, Middle Column: Frequencies, Last Column: Percentages

Age	Freq.	Percentage (%)
0.0	336.0	14.243323442136498
16.0	2.0	0.0847816871555744
17.0	4.0	0.1695633743111488
18.0	7.0	0.2967359050445104
19.0	5.0	0.21195421788893598

Figure 7: Sample Frequency/Percentage Distribution (based on the diagrams as seen in Figures 3, 4 & 6) – Refer To 'Full Dataset 1' For Complete Records

Measures	Values
Mean	51.978 (3.d.p) – As seen in Figure 5
Median	57 – As seen in Figure 5
Mode	0 – As seen in Figures 3, 4, 6 & 7
Range	88 (Min: 0; max: 88) – As seen in Figure 5
Variance	679.319 (3.d.p)
Standard Deviation (S.D)	26.064 (3.d.p)
25th Percentile	42 – As seen in Figure 5
75th Percentile	72 – As seen in Figure 5

Table 2: Summary Properties (Based on the 'age' attribute)

3. **LOSdays**

Attribute Type: Interval.

Justifications: The length of stay in days ('LOSdays') for patients in a hospital is generally regarded as an interval attribute. A property of this attribute is how it has meaningful intervals, with the difference between 5 and 10 days the same as that of 20 and 25 days. They lack a 'true zero' as records of 0 days does not mean that the patient was never in the hospital due to it perhaps being a very short stay or same-day discharge. It is possible to perform arithmetic operations on the attribute, but it is not sensible to assume that a LOS of 10 days is 'twice as long' as 5 days.

Summarising Properties With Visualisations:

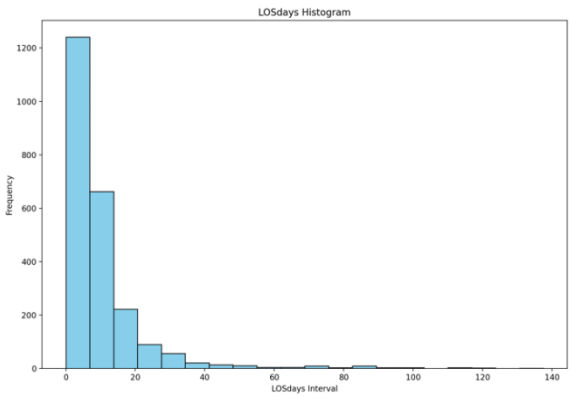


Figure 8 – Frequency Distribution (Histogram)

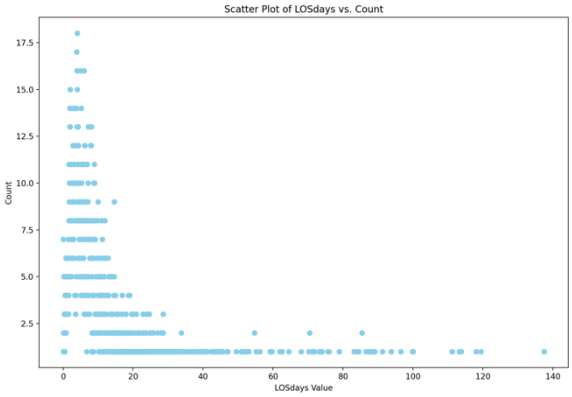


Figure 9 – Scatter Plot Distribution

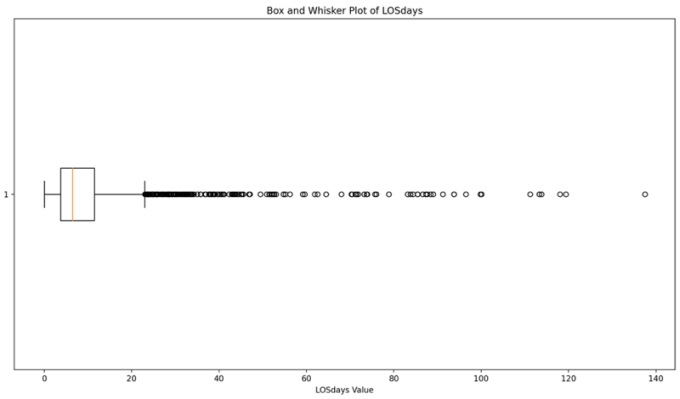


Figure 10: Data with Outliers (Box-and-Whisker Plot)

Measures	Values
Mean	10.14 (2.d.p)
Median (50th Percentile)	6.46 (2.d.p) – As seen in Figure 10
Mode	4 – Within the first and second 'bins'; As seen in Figures 8 & 9
Range/Maximum Value	137.5 (1.d.p) – As seen in Figure 10
Variance	170.93 (2.d.p)
Standard Deviation (S.D)	13.07 (2.d.p)
Minimum Value	0 – As seen in Figure 10
25th Percentile	3.71 (2.d.p) – As seen in Figure 10
75th Percentile	11.44 (2.d.p) – As seen in Figure 10
Upper Bound Value (Excl. Outliers)	23.04 (2.d.p) – As seen in Figure 10

Table 3: Summary Properties (Based on the 'LOSdays' attribute)

4. **Admit_location**

Attribute Type: Nominal (categorical).

Justifications: The 'Admit_location' of patients can be regarded as a nominal or categorical attribute. For the hospital, this is used to locate where patients are and many patients can be grouped in one location, for example the 'emergency' room or its associated counterparts. These locations are discrete and have no inherent numerical order or value that is associated with them. Arithmetic operations are also invalid as it is not possible to perform arithmetic operations on categorical data such as the specific location of people in a hospital.

Summarising Properties and Visualisations:

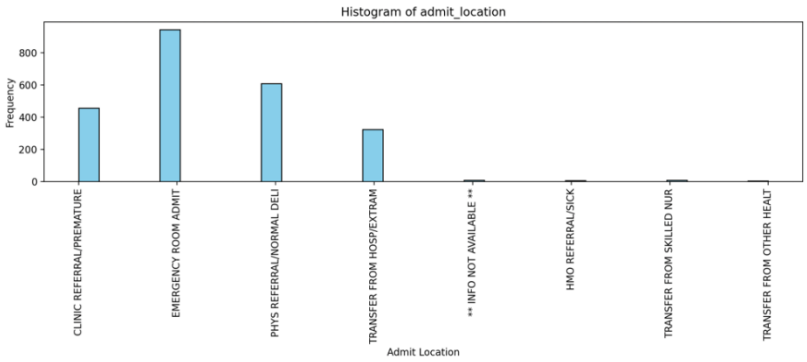


Figure 11: Frequency Distribution (Histogram)

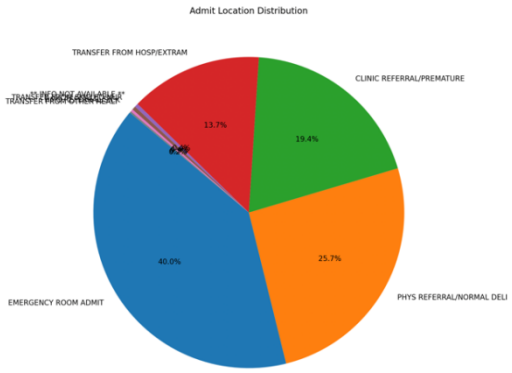


Figure 12: Percentage Distribution (Pie Chart)

Title	Values
Frequency Distribution	Emergency Room Admit: 943; Phys Referral/Normal Deli: 607; Clinic Referral/Premature: 457; Transfer From Hosp/Extram: 324; **Info Not Available**: 9; Transfer From Skilled Nur: 9; HMO Referral/Sick: 6; Transfer From Other Healt: 4 – As seen in Figure 11
Percentage Distribution	Emergency Room Admit: 40% (1.d.p); Phys Referral/Normal Deli: 25.7% (1.d.p); Clinic Referral/Premature: 19.4% (1.d.p); Transfer From Hosp/Extram: 13.7% (1.d.p); **Info Not Available**: 0.4% (1.d.p); Transfer From Skilled Nur: 0.4% (1.d.p); HMO Referral/Sick: 0.3% (1.d.p); Transfer From Other Healt: 0.1% (1.d.p) – As seen in Figure 12
Mode	Emergency Room Admit – As seen in Figures 11 & 12
Count of Unique Categories	8; including '**info not available**'

Table 4: Summary Properties (Based on the 'Admit_location' attribute)

5. **AdmitDiagnosis**

Attribute Type: Nominal (categorical).

Justifications: The 'AdmitDiagnosis' of patients in a hospital can be regarded as a nominal or categorical attribute. This refers to the medical diagnosis that a patient has for them to be admitted. It is possible for multiple people to have the same condition or be there for similar reasons, which also includes people going to the hospital because they have a newborn. Overall, these are discretely grouped while having no order or value, and arithmetic cannot be performed on them.

Summarising Properties With Visualisations:

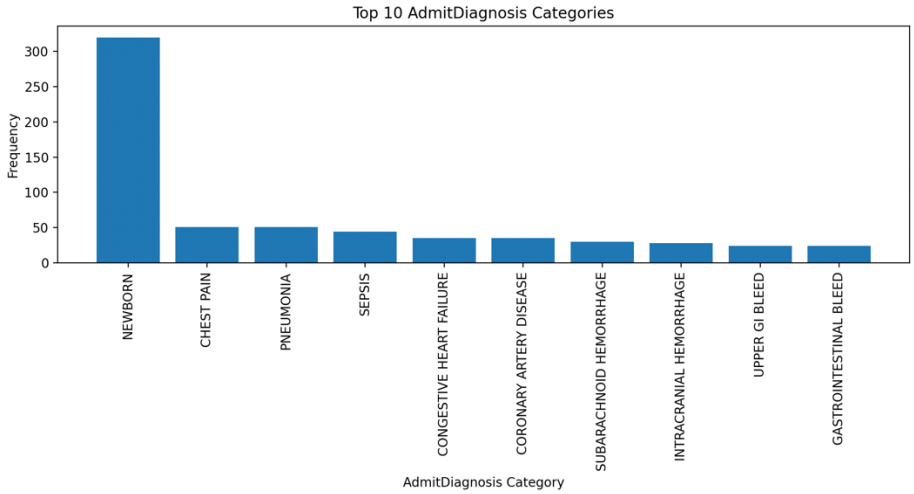


Figure 13: Frequency Distribution (Histogram)

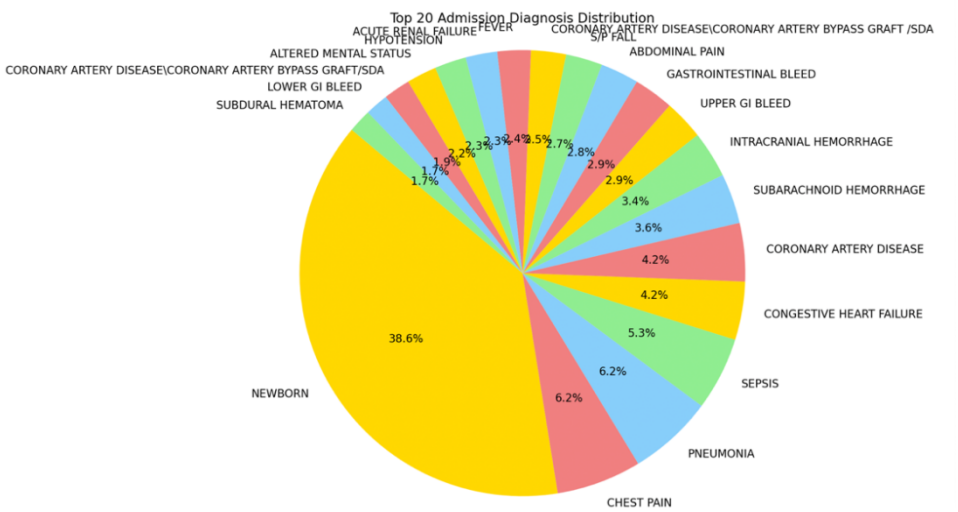


Figure 14: Percentage Distribution (Pie Chart)

Title	**Values**
Frequency Distribution	Newborn: 320; Pneumonia: 51; Chest pain: 51; Sepsis: 44; Congestive Heart Failure: 35 – As seen in Figure 13
Percentage Distribution	Sample Values - Newborn: 13.6% (1.d.p); Pneumonia: 2.2% (1.d.p), Chest pain: 2.2% (1.d.p); Sepsis: 1.9% (1.d.p); Congestive Heart Failure: 1.5% (1.d.p) – As seen in Figure 14
Mode	Newborn – As seen in Figure 13 & 14
Count of Unique Categories	1098

Table 5: Summary Properties (Based on the 'AdmitDiagnosis' Attribute - Only Included 5 Categories for Frequency and Percentage as There's Too Many To Analyse)

6. Insurance

Attribute Type: Nominal (categorical).

Justifications: The 'Insurance' status of patients in a hospital can be regarded as a nominal or categorical attribute. This reflects the status of everyone's healthcare cover, with it being possible that multiple people have the same cover, for example 'medicare', 'private', and 'medicaid'. People are grouped based on such details, however, there does not seem to be any apparent natural order that makes them ordered based on the importance of one over the other or alphabetically. These values are typically also not numerical, so a '0' value will not make sense, and it will need to be replaced by 'N/A'.

Summarising Properties and Visualisations:

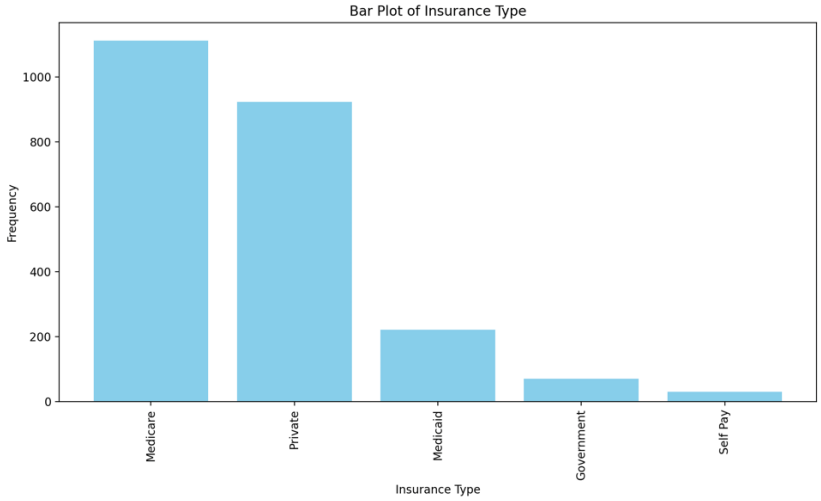


Figure 15: Frequency Distribution (Histogram)

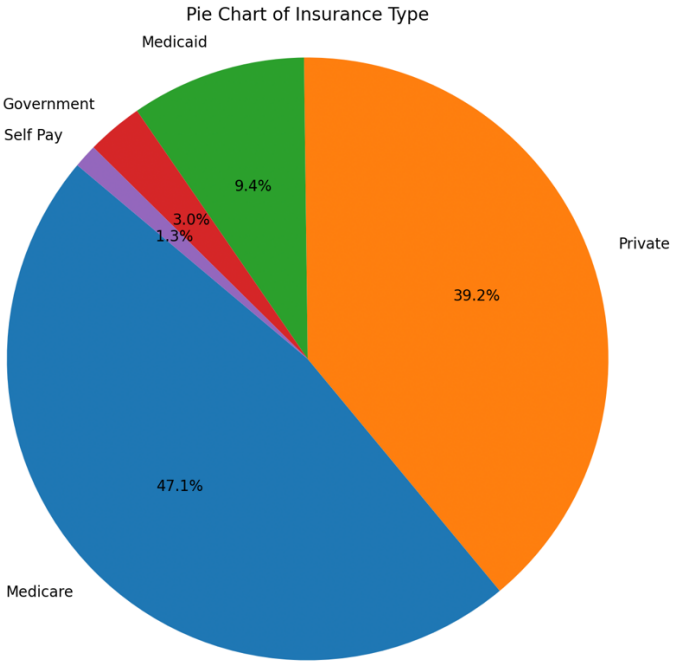


Figure 16: Percentage Distribution (Pie Chart)

Title	Values
Frequency Distribution	Medicare: 1112; Private: 924; Medicaid: 222; Government: 71; Self Pay: 30 – As seen in Figure 15
Percentage Distribution	Medicare: 47.1% (1.d.p); Private: 39.2% (1.d.p); Medicaid: 9.4% (1.d.p); Government: 3%; Self Pay: 1.3% (1.d.p) – As seen Figure 16
Mode	Medicare – As seen in Figures 15 & 16
Count of Unique Categories	5

Table 6: Summary Properties (based on the 'Insurance' attribute)

7. NumCallouts

Attribute Type: Ratio.

Justifications: The attribute 'NumCallouts' matches the definition as a ratio attribute as it has a 'true zero' point and is not just a placeholder. In terms of a hospital, it likely refers to the number of times healthcare professionals are called or contacted. The 'true zero' point means that the patient has not had any contact with any healthcare professionals for issues such as a routine check-up, follow-up appointment, or medical emergency. 'NumCallouts' may be a useful tool for a hospital to track a patient's medical history and identify those who require more frequent care as they are at risk of developing chronic health conditions.

Summarising Properties With Visualisations:

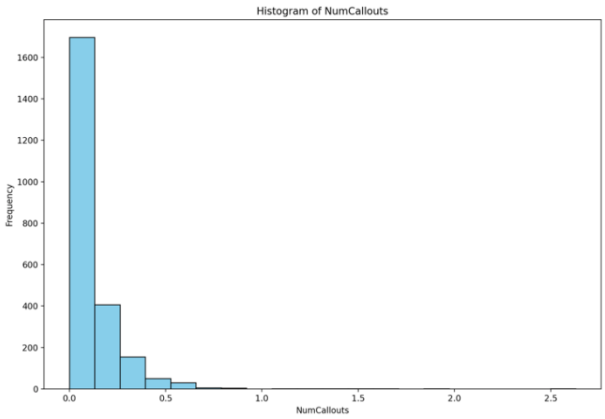


Figure 17: Frequency Distribution (Histogram)

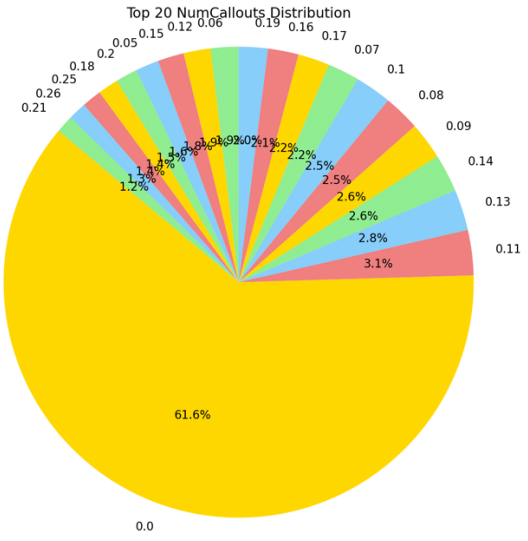


Figure 18: Percentage Distribution (Pie Chart)

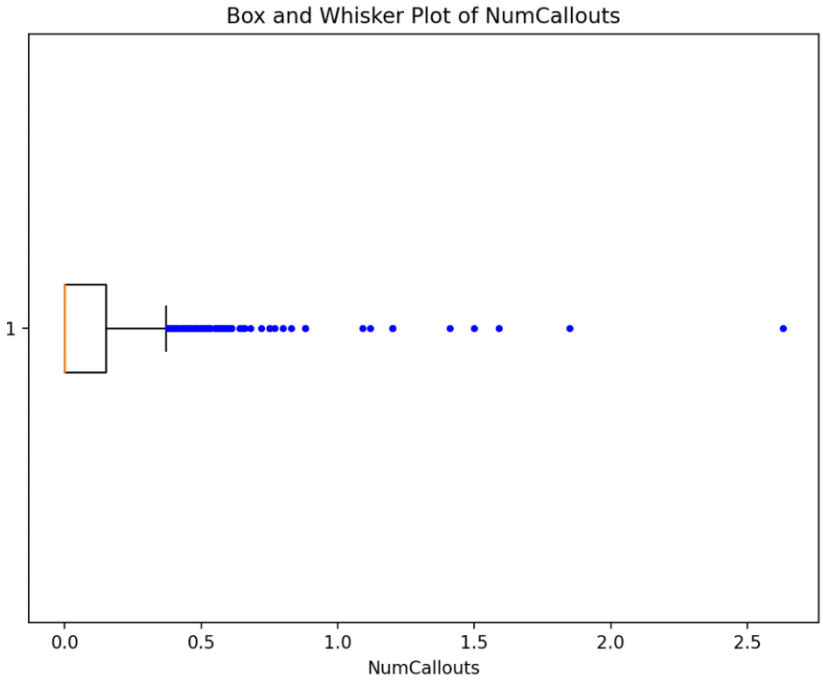
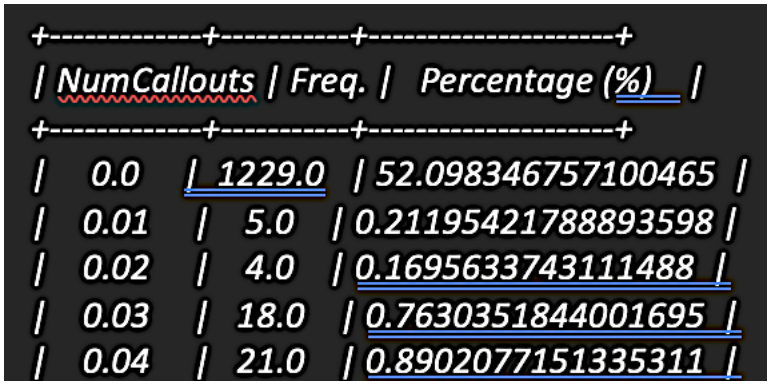


Figure 19: 'NumCallouts' Distribution (Box-and-Whisker Plot with Outliers)

NumCallouts	Frequency	Percentage
0.00	1229	52.098347
0.01	5	0.211954
0.02	4	0.169563
0.03	18	0.763035
0.04	21	0.890208
0.05	32	1.356507
0.06	38	1.610852
0.07	44	1.865197
0.08	50	2.119542
0.09	51	2.161933
0.10	50	2.119542
0.11	62	2.628232
0.12	37	1.568461
0.13	56	2.373887
0.14	52	2.204324
0.15	36	1.526070
0.16	42	1.780415
0.17	43	1.822806
0.18	28	1.186944
0.19	40	1.695634
0.20	30	1.271725
0.21	24	1.017380
0.22	18	0.763035
0.23	23	0.974989
0.24	19	0.805426
0.25	27	1.144553
0.26	25	1.059771
0.27	18	0.763035
0.28	11	0.466299

Figure 20: Part of The Output From Python



NumCallouts	Freq.	Percentage (%)
0.0	1229.0	52.098346757100465
0.01	5.0	0.21195421788893598
0.02	4.0	0.1695633743111488
0.03	18.0	0.7630351844001695
0.04	21.0	0.8902077151335311

Figure 21: Sample Frequency/Percentage Distribution (based on the diagrams as seen in Figures 17, 18 & 20) – Refer To 'Full Dataset 2' For Complete Records

Measures	**Values**
Mean	0.098 (3.d.p) – As seen in Figure 19
Median	0.0 – As seen in Figure 19
Mode	0.0 – As seen in Figures 17, 18, 20 & 21
Range	2.63 – As seen in Figure 19
Variance	0.026 (3.d.p)
Standard Deviation (S.D)	0.163 (3.d.p)
25th Percentile	0.0 – As seen in Figure 19
75th Percentile	0.15 – As seen in Figure 19

Table 7: Summary Properties (Based on the 'NumCallouts' attribute)

8. NumDiagnosis

Attribute Type: Ratio.

Justifications: The attribute 'NumDiagnosis' matches the characteristics of a ratio attribute as it does have a definitive zero point which is not just a placeholder in the dataset. In a healthcare scenario, it likely refers to the number of medical conditions or events that a patient has been diagnosed with. The 'zero point' in this case means that the patient has not been diagnosed with any medical conditions, making it meaningful while not being arbitrary, although there are officially no records of such patients in the given dataset. The values do have an order and distances between values are meaningful.

Summarising Properties With Visualisations:

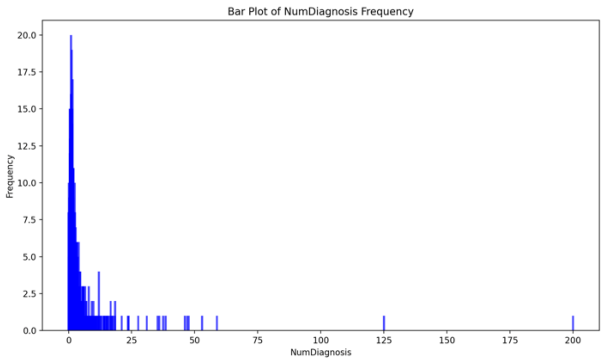


Figure 22: Frequency Distribution (Histogram)

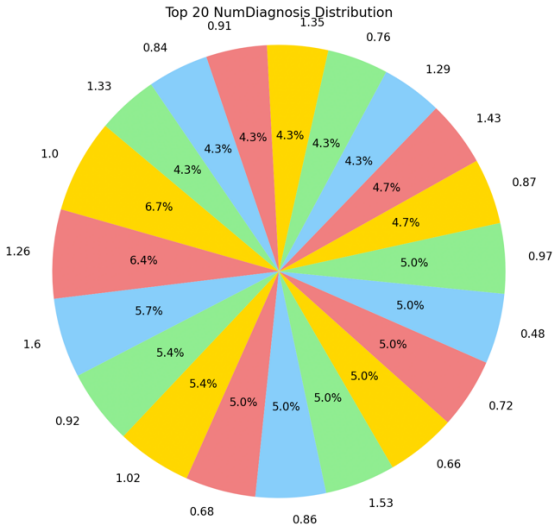


Figure 23: Percentage Distribution (Pie Chart)

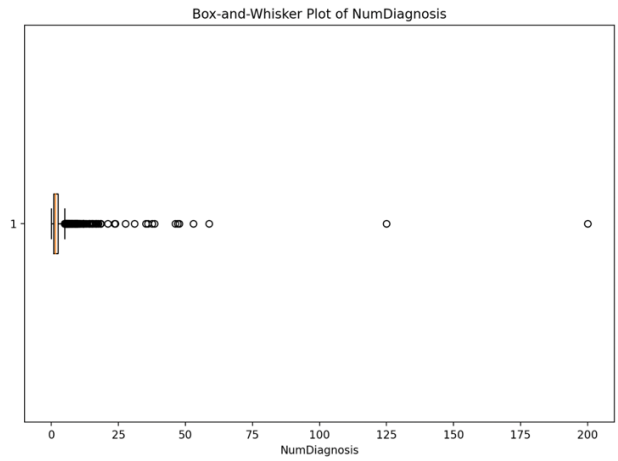


Figure 24: 'NumDiagnosis' Distribution (Box-and-Whisker Plot with Outliers)

NumDiagnosis	Frequency	Percentage (%)
0.00	8	0.339127
0.07	1	0.042391
0.08	3	0.127173
0.09	1	0.042391
0.10	10	0.423908
0.13	5	0.211954
0.14	2	0.084782
0.15	4	0.169563
0.16	5	0.211954
0.17	4	0.169563
0.18	2	0.084782
0.19	5	0.211954
0.20	4	0.169563
0.21	5	0.211954
0.22	8	0.339127
0.23	7	0.296736
0.24	4	0.169563

Figure 25: Part of The Output From Python

NumDiag	Freq.	Percentage (%)
0.0	8.0	0.3391267486222976
0.07	1.0	0.0423908435777872
0.08	3.0	0.1271725307333616
0.09	1.0	0.0423908435777872
0.1	10.0	0.42390843577787196

Figure 26: Sample Frequency/Percentage Distribution (based on the diagrams as seen in Figures 22, 23 & 25) – Refer To 'Full Dataset 3' For Complete Records

Measures	**Values**
Mean	2.38 (2.d.p)
Median	1.42 – As seen in Figure 24
Mode	1 – As seen in Figures 22, 23 & 26
Range	200 – As seen in Figure 24
Variance	35.92 (2.d.p)
Standard Deviation (S.D)	5.99 (2.d.p)
25th Percentile	0.845 – As seen in Figure 24
75th Percentile	2.45 – As seen in Figure 24

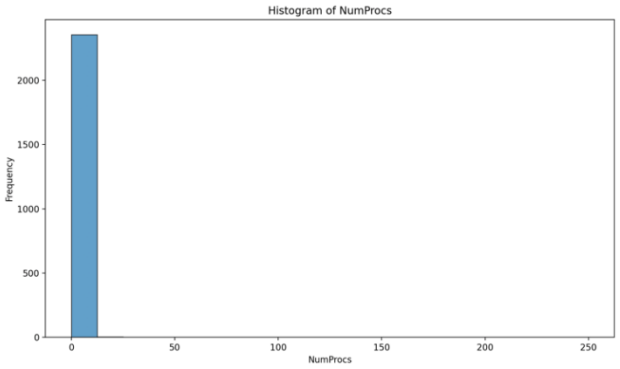
Table 8: Summary Properties (Based on the 'NumDiagnosis' attribute)

9. NumProcs

Attribute Type: Ratio.

Justifications: The attribute 'NumProcs' matches the traits of a ratio attribute as there is a definitive zero point which is not just a placeholder and makes it not an arbitrary value. These values have an order and distances between the values are quite meaningful. In a healthcare setting, it likely identifies the number of medical procedures that the patients have gone through. The records of people in the dataset who have '0' means that they have not undergone any cumulative medical procedures.

Summarising Properties With Visualisations:



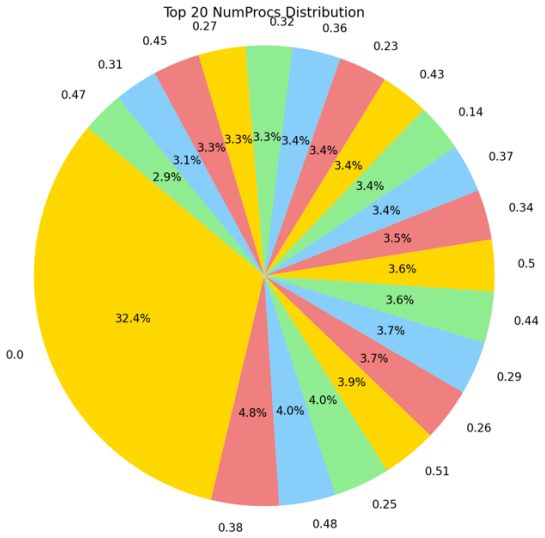


Figure 27: Frequency Distribution (Histogram) *Figure 28: Percentage Distribution (Pie Chart)*

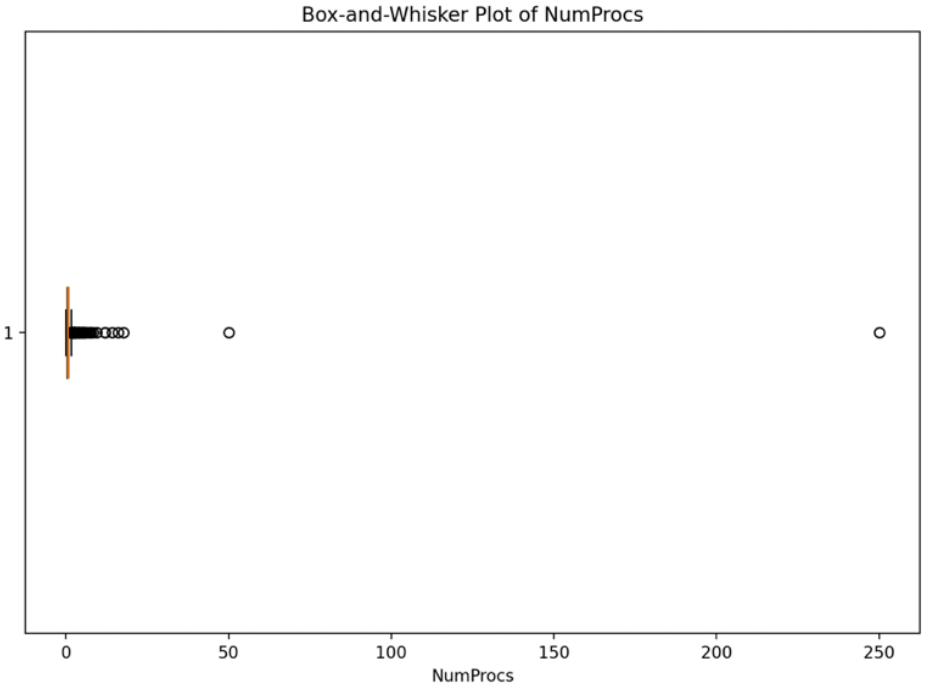


Figure 29: 'NumProcs' Distribution (Box-and-Whisker Plot with Outliers)

NumProcs	Frequency	Percentage (%)
0.00	286	12.123781
0.02	1	0.042391
0.03	3	0.127173
0.04	3	0.127173
0.05	6	0.254345
0.06	4	0.169563
0.07	12	0.508690
0.08	15	0.635863
0.09	11	0.466299
0.10	19	0.805426
0.11	26	1.102162
0.12	16	0.678253
0.13	24	1.017380
0.14	30	1.271725
0.15	23	0.974989
0.16	25	1.059771
0.17	24	1.017380
0.18	20	0.847817
0.19	19	0.805426
0.20	19	0.805426
0.21	18	0.763035
0.22	26	1.102162
0.23	30	1.271725
0.24	21	0.890208
0.25	35	1.483680
0.26	33	1.398898
0.27	29	1.229334
0.28	23	0.974989

Figure 30: Part of The Output From Python

NumProcs	Freq.	Percentage (%)
0.0	286.0	12.123781263247139
0.02	1.0	0.0423908435777872
0.03	3.0	0.1271725307333616
0.04	3.0	0.1271725307333616
0.05	6.0	0.2543450614667232

Figure 31: Frequency/Percentage Distribution (based on the diagrams as seen in Figures 16, 17 & 19) – Refer To 'Full Dataset 4' For Complete Records

Measures	**Values**
Mean	0.76 (2.d.p)
Median	0.43 – As seen in Figure 29
Mode	0 – As seen in Figures 27, 28, 30 & 31
Range	250 – As seen in Figure 29
Variance	28.43 (2.d.p)
Standard Deviation (S.D)	5.33 (2.d.p)
25th Percentile	0.21 – As seen in Figure 29
75th Percentile	0.74 – As seen in Figure 29

Table 9: Summary Properties (Based on the 'NumProcs' attribute)

10. AdmitProcedure

Attribute Type: Nominal (categorical).

Justifications: The attribute 'AdmitProcedure' can be considered as a nominal and discrete attribute as there is no inherent order, although they do classify people in the hospital into groups by using names or labels. The distances between the values provided are not meaningful and the zero point is arbitrary. In a hospital context, it refers to the process of how patients are assisted by the hospital based on their 'AdmitDiagnosis'. Those who have a 'N/A' record means that they have not gone to the hospital because of certain treatments, techniques, or conditions, and therefore the hospital does not need to come up with procedures to remedy them.

Summarising Properties With Visualisations:

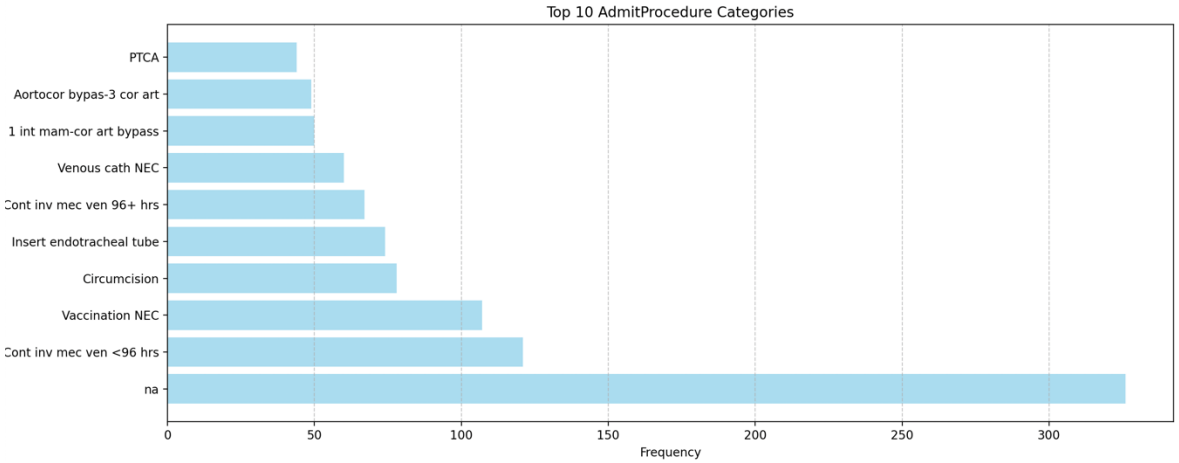


Figure 32: Frequency Distribution (Bar Plot)

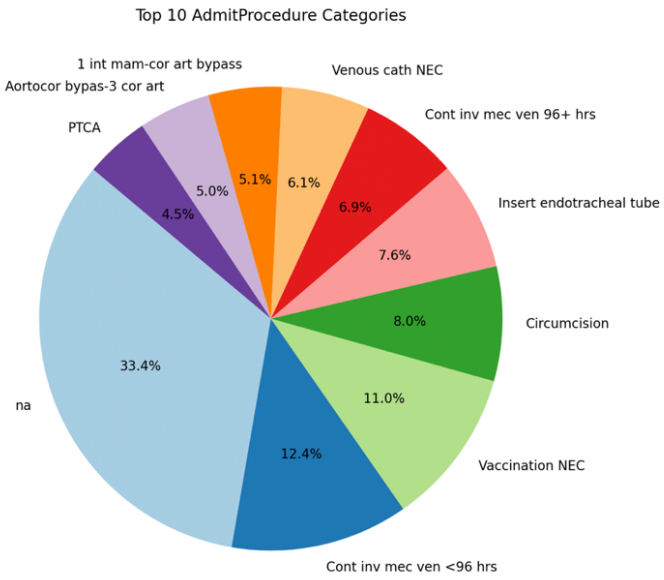


Figure 33: Percentage Distribution (Pie Chart)

Title	Values
Frequency Distribution	Top 5: N/A (326), Cont inv mec ven <96 hrs (121), Vaccination NEC (107), Circumcision (78), Insert endotracheal tube (74) – As seen in Figure 32
Percentage Distribution	Top 5 (based on the total): N/A (13.8%: 1.d.p), Cont inv mec ven <96 hrs (5.1%: 1.d.p), Vaccination NEC (4.5%: 1.d.p), Circumcision (3.3%: 1.d.p), Insert endotracheal tube (3.1%: 1.d.p) – As seen in Figure 33
Mode	N/A
Count of Unique Categories	377

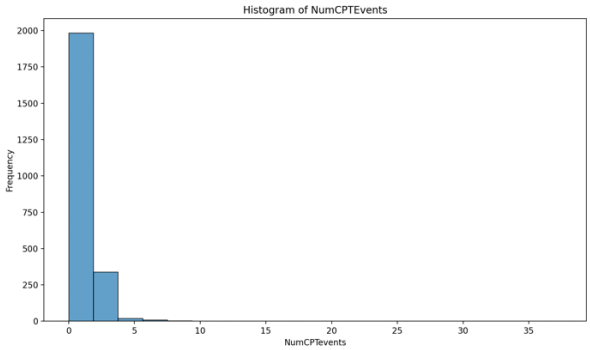
Table 10: Summary Properties (Based on the 'AdmitProcedure' attribute)

11. NumCPTevents

Attribute Type: Ratio.

Justifications: The attribute 'NumCPTevents' matches the definition as a ratio attribute. It does have a definitive and natural 'zero point'. Since it probably displays the number of times a patient has undergone a particular medical procedure, it can be likened to a ratio attribute. This can also be measured on a continuous scale, while it is also possible to perform arithmetic on the values. A value of '0' for a patient means that they have not went through any specific medical procedures that are concerning.

Summarising Properties With Visualisations:



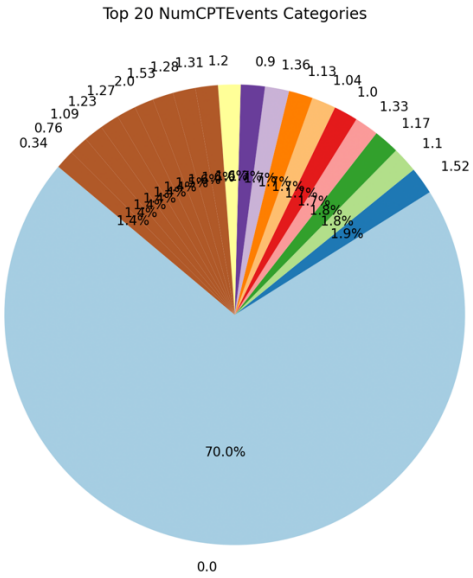


Figure 34: Frequency Distribution (Histogram) *Figure 35: Percentage Distribution (Pie Chart)*

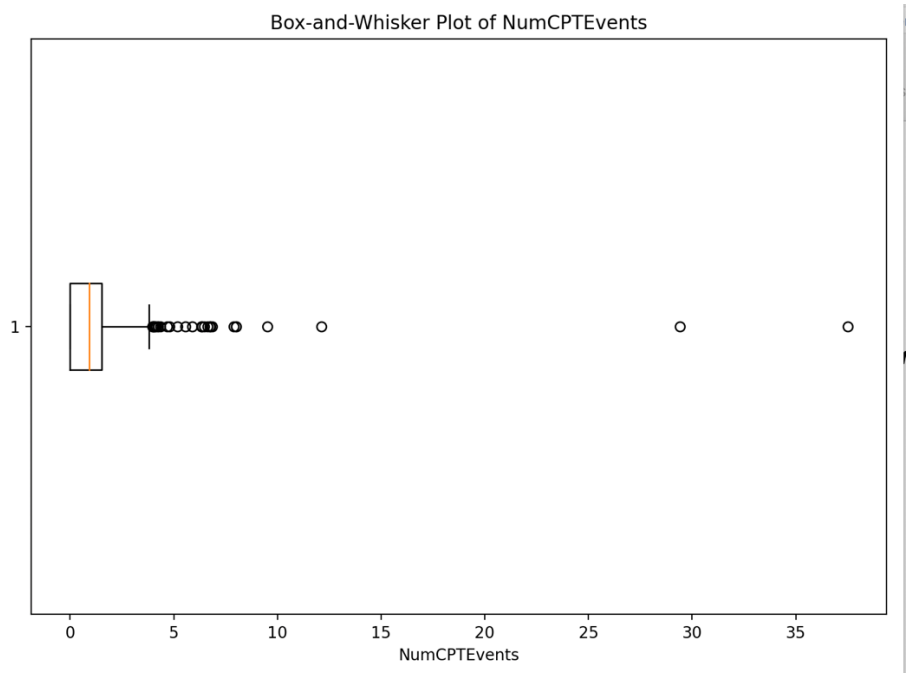


Figure 36: 'NumCPTevents' Distribution (Box-and-Whisker Plot with Outliers)

NumCPTEvents	Frequency	Percentage (%)
0.00	622	26.367105
0.01	1	0.042391
0.04	2	0.084782
0.05	2	0.084782
0.06	1	0.042391
0.07	3	0.127173
0.08	3	0.127173
0.09	6	0.254345
0.10	4	0.169563
0.11	2	0.084782
0.12	9	0.381518
0.13	11	0.466299
0.14	8	0.339127
0.15	5	0.211954
0.16	11	0.466299
0.17	7	0.296736
0.18	5	0.211954
0.19	8	0.339127
0.20	5	0.211954
0.21	7	0.296736
0.22	5	0.211954
0.23	8	0.339127
0.24	7	0.296736
0.25	10	0.423908
0.26	6	0.254345
0.27	7	0.296736
0.28	5	0.211954
0.29	7	0.296736

Figure 37: Part of The Output from Python

NumCPTEvents	Frequency	Percentage (%)
0.0	622.0	26.36710470538364
0.01	1.0	0.0423908435777872
0.04	2.0	0.0847816871555744
0.05	2.0	0.0847816871555744
0.06	1.0	0.0423908435777872
0.07	3.0	0.1271725307333616
0.08	3.0	0.1271725307333616
0.09	6.0	0.2543450614667232
0.1	4.0	0.1695633743111488
0.11	2.0	0.0847816871555744
0.12	9.0	0.38151759220008474
0.13	11.0	0.4662992793556591
0.14	8.0	0.3391267486222976

Figure 38: Frequency/Percentage Distribution (based on the diagrams as seen in Figures 16, 17 & 19) – Refer To 'Full Dataset 5' For Complete Records

Measures	**Values**
Mean	1.03 (2.d.p)
Median	0.92 – As seen in Figure 36
Mode	0 – As seen in Figures 34, 35, 37 & 38
Range	37.5 – As seen in Figure 36
Variance	1.99 (2.d.p)
Standard Deviation (S.D)	1.41 (2.d.p)
25th Percentile	0.0 – As seen in Figure 36
75th Percentile	1.53 – As seen in Figure 36

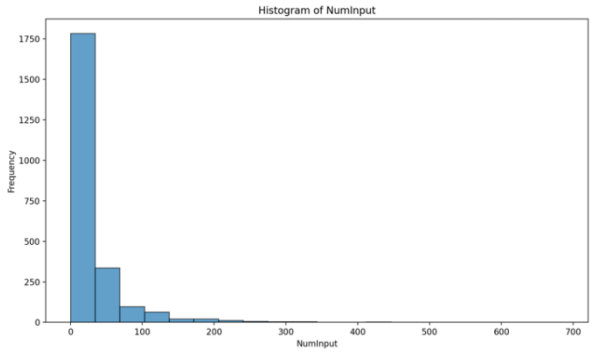
Table 11: Summary Properties (Based on the 'NumCPTevents' attribute)

12. NumInput

Attribute Type: Ratio.

Justifications: The attribute 'NumInput' matches the characteristics of a ratio attribute. It is a continuous attribute with a natural zero point that has inherent meaning, is not arbitrary, and is not just a placeholder. Such values have an order and the distances between them are meaningful. In a hospital or healthcare scenario, the attribute may refer to the number of units of a substance that has been received by patients. A value of '0' for the patients mean that they have not received any substances.

Summarising Properties With Visualisations:



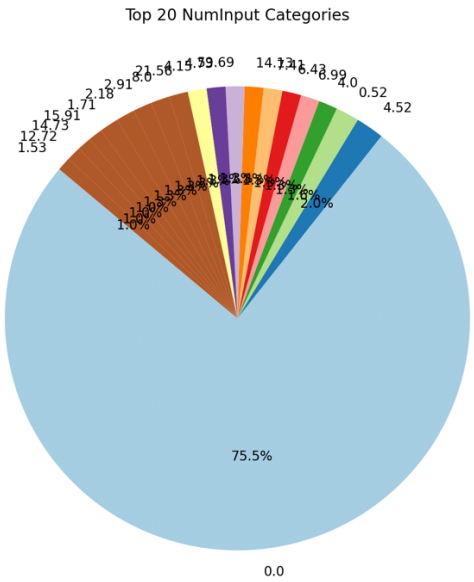


Figure 39: Frequency Distribution (Histogram) *Figure 40: Percentage Distribution (Pie Chart)*

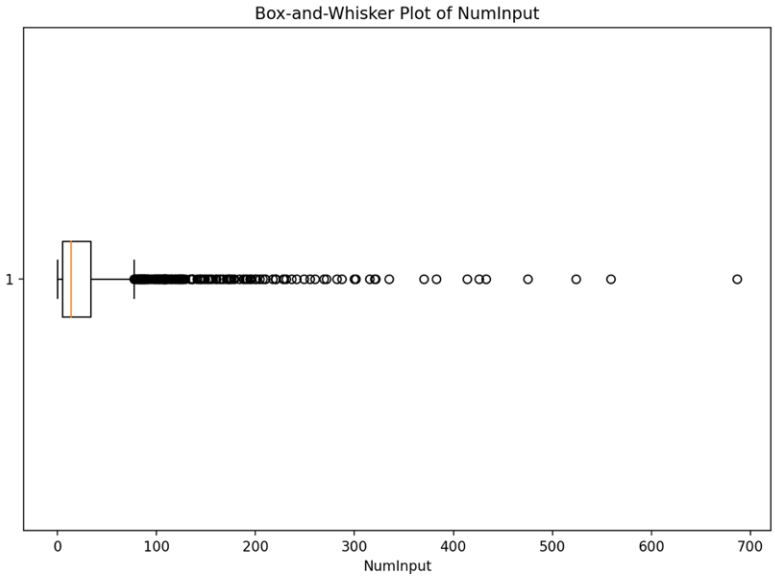


Figure 41: 'NumInput' Distribution (Box-and-Whisker Plot with Outliers)

NumInput	Frequency	Percentage (%)
0.00	231	9.792285
0.04	1	0.042391
0.14	1	0.042391
0.18	1	0.042391
0.22	1	0.042391
0.24	1	0.042391
0.25	2	0.084782
0.26	2	0.084782
0.28	1	0.042391
0.29	1	0.042391
0.37	1	0.042391
0.39	2	0.084782
0.40	1	0.042391
0.44	3	0.127173
0.45	1	0.042391
0.46	2	0.084782
0.50	1	0.042391
0.51	1	0.042391
0.52	5	0.211954
0.53	1	0.042391
0.55	1	0.042391
0.57	1	0.042391
0.58	1	0.042391
0.60	1	0.042391
0.63	1	0.042391
0.64	2	0.084782
0.65	1	0.042391
0.66	1	0.042391

Figure 42: Part of The Output from Python

NumInput	Freq.	Percentage (%)
0.0	231.0	9.792284866468842
0.04	1.0	0.0423908435777872
0.14	1.0	0.0423908435777872
0.18	1.0	0.0423908435777872
0.22	1.0	0.0423908435777872

Figure 43: Frequency/Percentage Distribution (based on the diagrams as seen in Figures 39, 40, 41 & 42) – Refer To 'Full Dataset 6' For Complete Records

Measures	**Values**
Mean	29.19 (2.d.p)
Median	13.21 – As seen in Figure 41
Mode	0 – As seen in Figures 39, 40, 42 & 43
Range	686.48 – As seen in Figure 41
Variance	2465.20 (2.d.p)
Standard Deviation (S.D)	49.65 (2.d.p)
25th Percentile	4.6 – As seen in Figure 41
75th Percentile	33.64 (2.d.p) – As seen in Figure 41

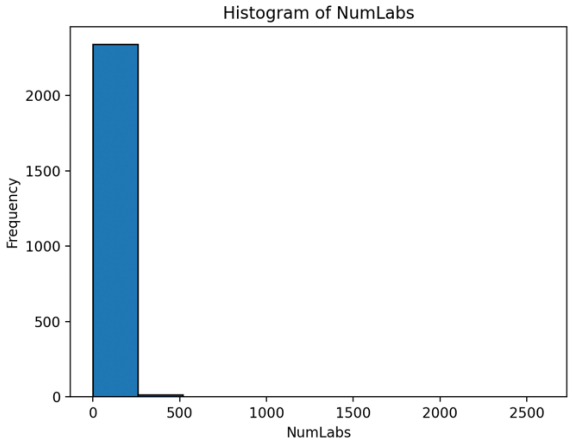
Table 12: Summary Properties (Based on the 'NumInput' attribute)

13. NumLabs

Attribute Type: Ratio.

Justifications: The attribute 'NumLabs' closely adopts the traits for a ratio attribute. In this case, the values are all provided as decimals with natural non-arbitrary '0' values which makes it continuous. There is order to the values while the distances between them are meaningful. Within a hospital, this can refer to the number of laboratory tests that a patient has undergone. The hospital may record the number of blood, urine, stool, or imaging tests, whereby a value of '0' means that the patient has not had any laboratory tests recorded.

Summarising Properties With Visualisations:



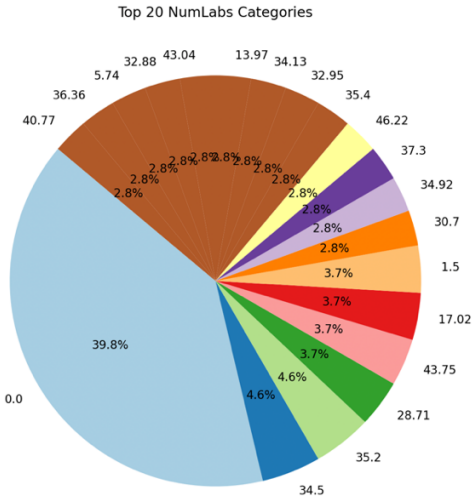


Figure 44: Frequency Distribution (Histogram) *Figure 45: Percent. Distribution (Pie Chart)*

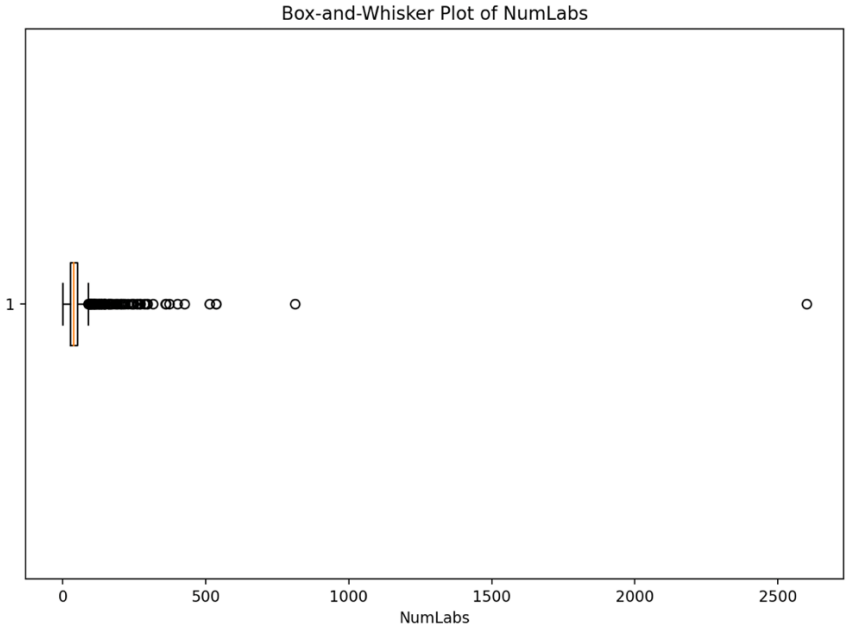


Figure 46: 'NumLabs' Distribution (Box-and-Whisker Plot with Outliers)

NumLabs	Frequency	Percentage (%)
0.00	43	1.822806
0.25	1	0.042391
0.55	1	0.042391
0.64	1	0.042391
0.74	3	0.127173
0.75	1	0.042391
0.76	1	0.042391
0.77	3	0.127173
0.78	1	0.042391
0.80	1	0.042391
0.95	1	0.042391
0.96	1	0.042391
1.00	1	0.042391
1.06	1	0.042391
1.16	2	0.084782
1.20	1	0.042391
1.22	2	0.084782
1.29	1	0.042391
1.31	2	0.084782
1.33	1	0.042391
1.36	2	0.084782
1.38	1	0.042391
1.41	2	0.084782
1.50	4	0.169563
1.53	1	0.042391
1.54	1	0.042391
1.57	1	0.042391
1.65	1	0.042391

Figure 47: Part of The Output from Python

NumLabs	Frequency	Percentage (%)
0.0	43.0	1.8228062738448496
0.25	1.0	0.0423908435777872
0.55	1.0	0.0423908435777872
0.64	1.0	0.0423908435777872
0.74	3.0	0.1271725307333616
0.75	1.0	0.0423908435777872
0.76	1.0	0.0423908435777872
0.77	3.0	0.1271725307333616

Figure 48: Sample Frequency/Percentage Distribution (based on the diagrams as seen in Figures 44, 45 & 47) – Refer To 'Full Dataset 7' For Complete Records

Measures	Values
Mean	44.73 (2.d.p)
Median	37.77 – As seen in Figure 46
Mode	0 – As seen in Figures 45, 47 & 48
Range	2600 – As seen in Figure 46
Variance	4782.64 (2.d.p)
Standard Deviation (S.D)	68.58 (2.d.p)
25th Percentile	25.53 – As seen in Figure 46
75th Percentile	51 – As seen in Figure 46

Table 13: Summary Properties (Based on the 'NumLabs' attribute)

14. NumMicroLabs

Attribute Type: Ratio.

Justifications: The attribute 'NumMicroLabs' inherits the traits of a ratio attribute. It is continuous with a natural, non-arbitrary zero point that makes it meaningful. Such values do have an inherent order, while meaningful distances are maintained between each other. In a hospital sense, it refers to the number of microscopic laboratory tests that a patient has undergone. It measures the number of cells in samples of stool, blood, or urine. The '0' values mean that the patient has had no medical microscopic laboratory tests conducted.

Summarising Properties With Visualisations:

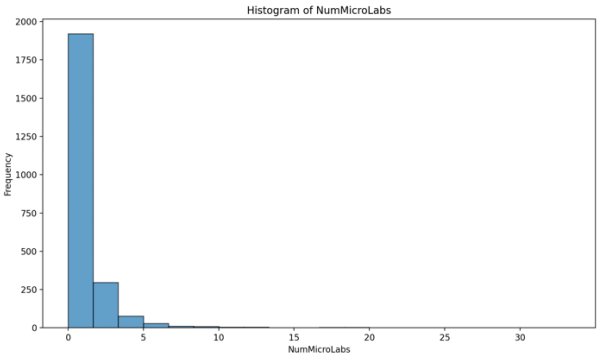


Figure 49: Frequency Distribution (Histogram)

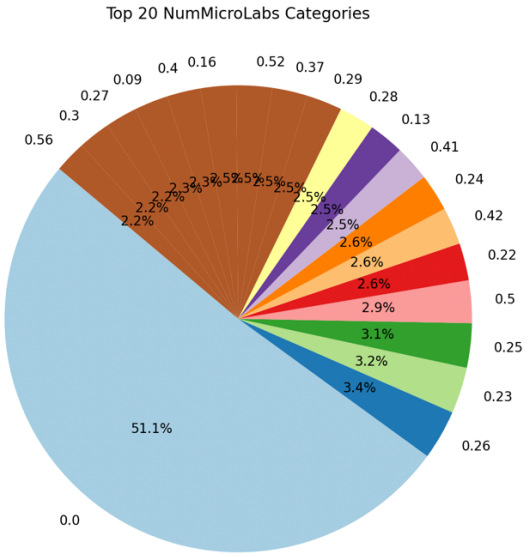


Figure 50: Percentage Distribution (Pie Chart)

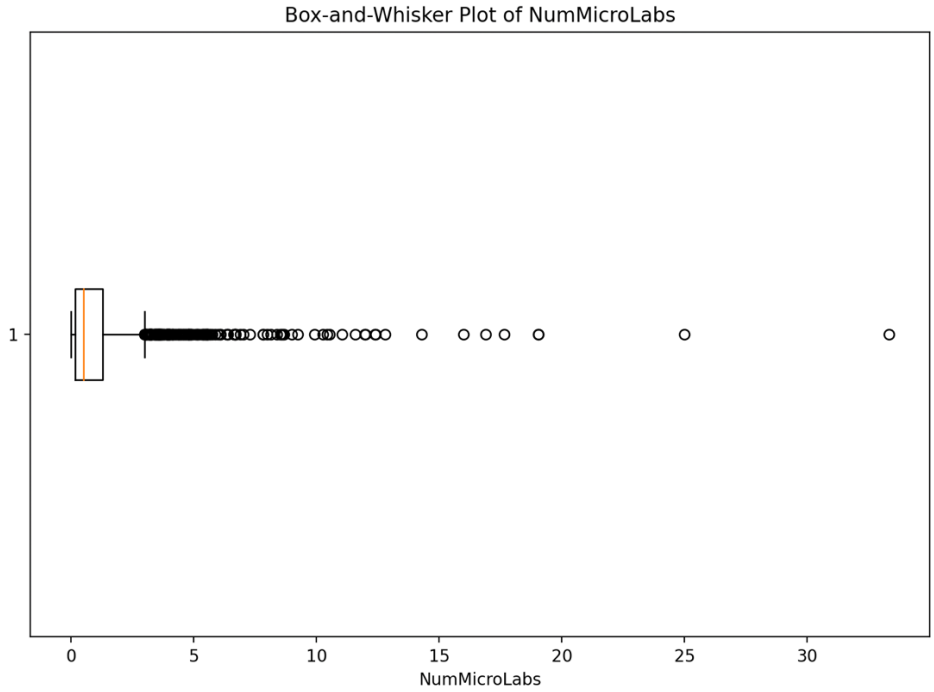


Figure 51: 'NumCallouts' Distribution (Box-and-Whisker Plot with Outliers)

NumMicroLabs	Frequency	Percentage (%)
0.00	416	17.634591
0.01	2	0.084782
0.02	11	0.466299
0.03	9	0.381518
0.04	13	0.551081
0.05	9	0.381518
0.06	6	0.254345
0.07	9	0.381518
0.08	9	0.381518
0.09	19	0.805426
0.10	13	0.551081
0.11	13	0.551081
0.12	15	0.635863
0.13	20	0.847817
0.14	12	0.508690
0.15	13	0.551081
0.16	20	0.847817
0.17	14	0.593472
0.18	18	0.763035
0.19	12	0.508690
0.20	12	0.508690
0.21	14	0.593472
0.22	21	0.890208
0.23	26	1.102162

Figure 52: Part of The Output From Python

```
+-------------+----------+--------------------+
| NumMicroLabs | Freq. |   Percentage (%)   |
+-------------+----------+--------------------+
|    0.0     |   416.0   | 17.634590928359476 |
|    0.01    |    2.0    | 0.0847816871555744 |
|    0.02    |   11.0    | 0.4662992793556591 |
|    0.03    |    9.0    | 0.38151759220008474 |
|    0.04    |   13.0    | 0.5510809665112336 |
```

Figure 53: Sample Frequency/Percentage Distribution (based on the diagrams as seen in Figures 49. 50 & 52) – Refer To 'Full Dataset 8' For Complete Records

Measures	**Values**
Mean	1.06 (2.d.p)
Median	0.5 – As seen in Figure 51
Mode	0 – As seen in Figures 49, 50, 52 & 53
Range	33.33 – As seen in Figure 51
Variance	3.41 (2.d.p)
Standard Deviation (S.D)	1.85 (2.d.p)
25th Percentile	0.16 – As seen in Figure 51
75th Percentile	1.3 (1.d.p) – As seen in Figure 51

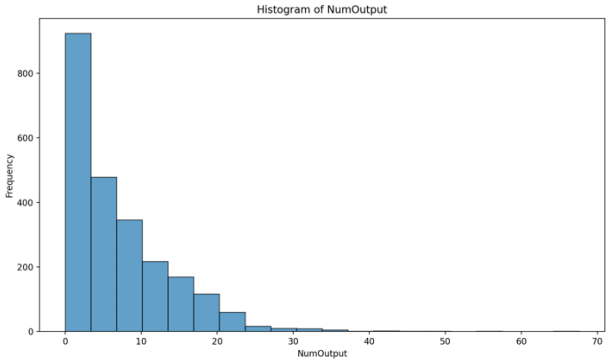
Table 14: Summary Properties (Based on the 'NumMicroLabs' attribute)

15. NumOutput

Attribute Type: Ratio.

Justifications: The attribute 'NumOutput' reflects the characteristics of a ratio attribute. It is continuous with a natural, non-arbitrary zero point with an order that makes distances between the values meaningful. In a hospital and healthcare context, the attribute refers to the number of units of a substance that a patient has produced, which may include their blood, urine, or stool. The '0' values mean that the patient has had no output of a particular substance.

Summarising Properties With Visualisations:



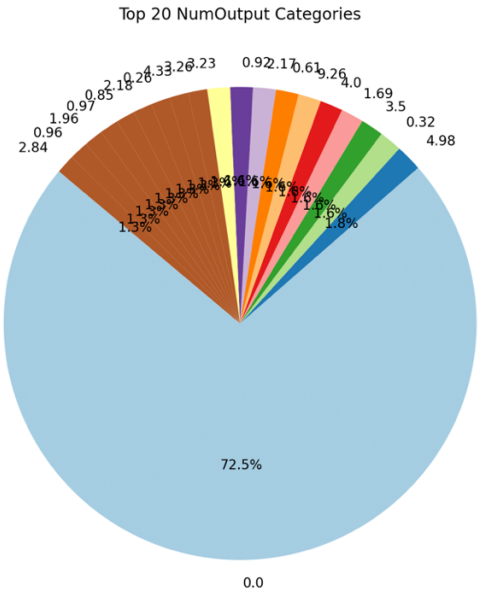


Figure 54: Frequency Distribution (Histogram) *Figure 55: Percent. Distribution (Pie Chart)*

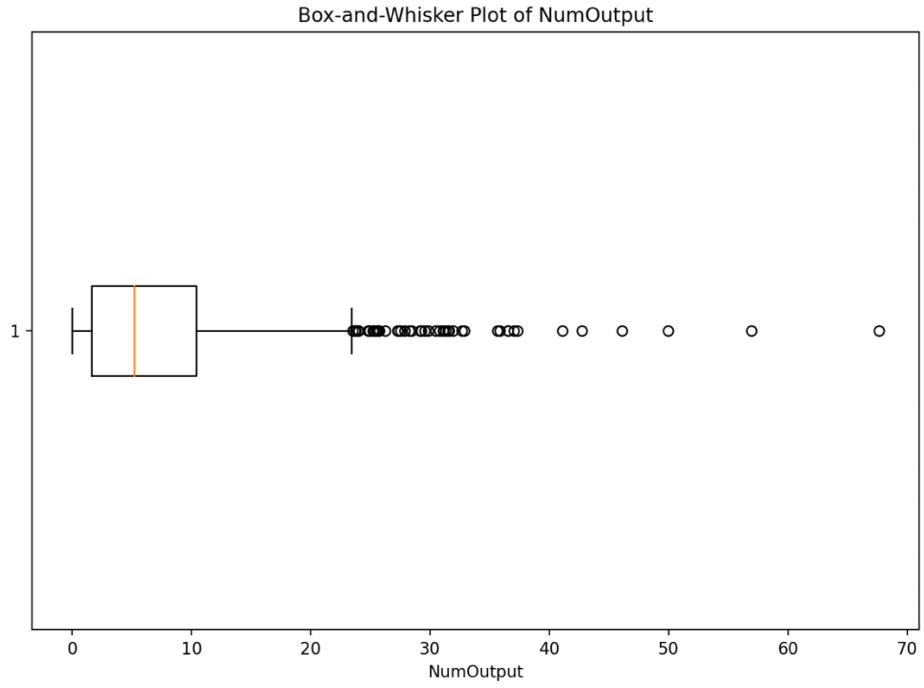


Figure 56: 'NumOutput' Distribution (Box-and-Whisker Plot with Outliers)

NumOutput	Frequency	Percentage (%)
0.00	280	11.869436
0.03	1	0.042391
0.04	1	0.042391
0.07	1	0.042391
0.08	1	0.042391
0.10	1	0.042391
0.11	1	0.042391
0.14	2	0.084782
0.18	3	0.127173
0.19	2	0.084782
0.20	1	0.042391
0.21	3	0.127173
0.22	2	0.084782
0.23	1	0.042391
0.24	3	0.127173
0.25	3	0.127173
0.26	5	0.211954
0.27	5	0.211954
0.28	2	0.084782
0.29	4	0.169563
0.30	5	0.211954
0.31	2	0.084782
0.32	6	0.254345
0.33	2	0.084782
0.34	2	0.084782
0.35	2	0.084782

Figure 57: Part of The Output From Python

NumOutput	Frequency	Percentage (%)
0.0	280.0	11.869436201780417
0.03	1.0	0.0423908435777872
0.04	1.0	0.0423908435777872
0.07	1.0	0.0423908435777872
0.08	1.0	0.0423908435777872
0.1	1.0	0.0423908435777872
0.11	1.0	0.0423908435777872
0.14	2.0	0.0847816871555744
0.18	3.0	0.1271725307333616

Figure 58: Sample Frequency/Percentage Distribution (based on the diagrams as seen in Figures 54, 55, & 57) – Refer To 'Full Dataset 9' For Complete Records

Measures	**Values**
Mean	6.96 (2.d.p)
Median	5.17 – As seen in Figure 56
Mode	0 – As seen in Figures 54, 55, 57 & 58
Range	67.61 – As seen in Figure 56
Variance	47.72 (2.d.p)
Standard Deviation (S.D)	6.91 (2.d.p)
25th Percentile	1.63 (2.d.p) – As seen in Figure 56
75th Percentile	10.38 (2.d.p) – As seen in Figure 56

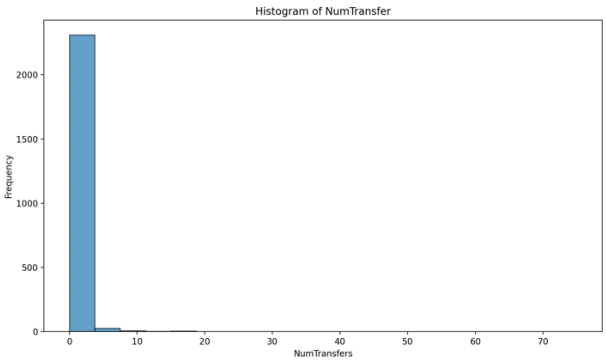
Table 15: Summary Properties (Based on the 'NumOutput' attribute)

16. NumTransfers

Attribute Type: Ratio.

Justifications: The attribute 'NumTransfer' resembles a ratio attribute. It fits the definition as a continuous attribute and has a natural non-arbitrary zero point with real meaning. The values do have an order, while the distances between the values are meaningful. In a hospital scenario, the attribute likely refers to the number of times that a patient has been transferred to a different ward or medical department. The data provided suggests that a patient may be transferred zero or any positive number of times.

Summarising Properties With Visualisations:



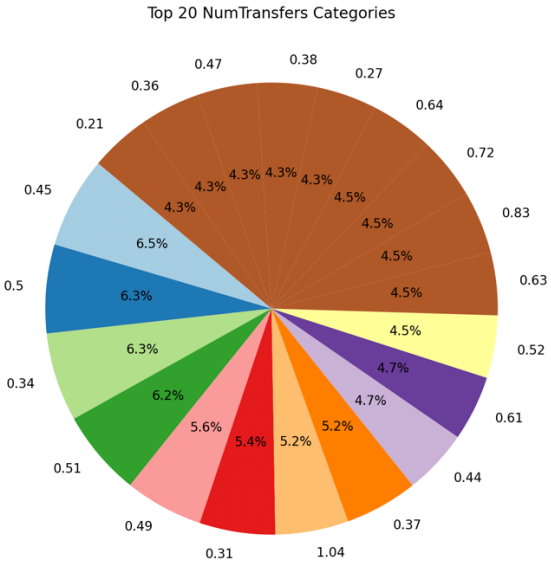


Figure 59: Frequency Distribution (Histogram) *Figure 60: Percent. Distribution (Pie Chart)*

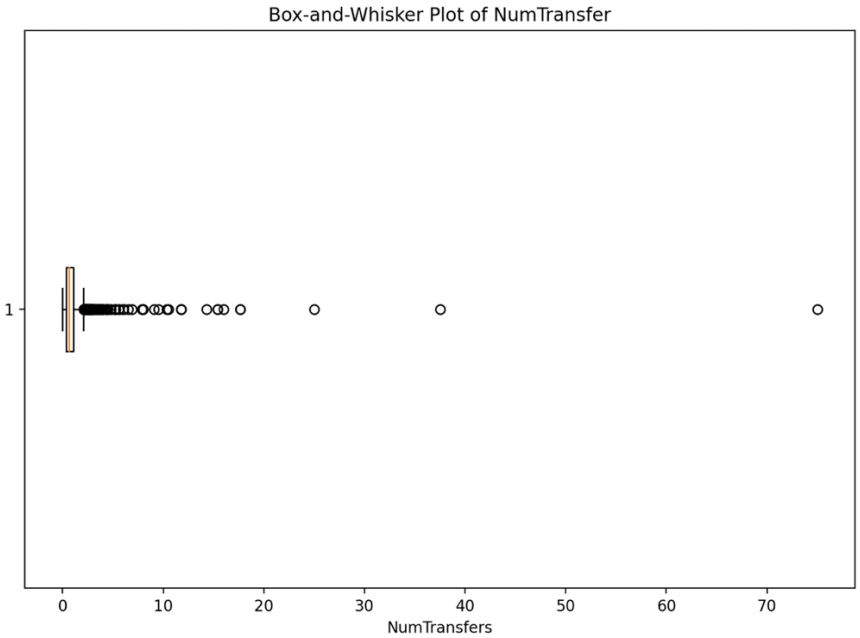


Figure 61: 'NumTransfers' Distribution (Box-and-Whisker Plot with Outliers)

NumTransfers	Frequency	Percentage (%)
0.00	7	0.296736
0.03	2	0.084782
0.04	4	0.169563
0.05	6	0.254345
0.06	7	0.296736
0.07	12	0.508690
0.08	5	0.211954
0.09	13	0.551081
0.10	11	0.466299
0.11	13	0.551081
0.12	18	0.763035
0.13	16	0.678253
0.14	13	0.551081
0.15	15	0.635863
0.16	15	0.635863
0.17	17	0.720644
0.18	15	0.635863
0.19	11	0.466299
0.20	19	0.805426
0.21	23	0.974989
0.22	18	0.763035
0.23	17	0.720644
0.24	18	0.763035
0.25	17	0.720644
0.26	17	0.720644

Figure 62: Part of The Output from Python

NumTransfers	Freq.	Percentage (%)
0.0	7.0	0.2967359050445104
0.03	2.0	0.0847816871555744
0.04	4.0	0.1695633743111488
0.05	6.0	0.2543450614667232
0.06	7.0	0.2967359050445104
0.07	12.0	0.5086901229334464
0.08	5.0	0.21195421788893598
0.09	13.0	0.5510809665112336
0.1	11.0	0.4662992793556591
0.11	13.0	0.5510809665112336
0.12	18.0	0.7630351844001695

Figure 63: Sample Frequency/Percentage Distribution (based on the diagrams as seen in Figures 59, 60 & 62) – Refer To 'Full Dataset 10' For Complete Records

Measures	Values
Mean	1.00 (2.d.p)
Median	0.66 – As seen in Figure 61
Mode	0.45 – As seen in Figures 59, 60 & 63
Range	75 – As seen in Figure 61
Variance	4.67 (2.d.p)
Standard Deviation (S.D)	2.16 (2.d.p)
25th Percentile	0.39 – As seen in Figure 61
75th Percentile	1.07 – As seen in Figure 61

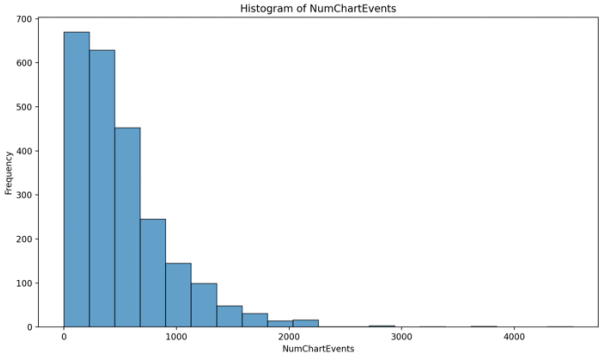
Table 16: Descriptive Statistics (based on the 'NumTransfers' attribute)

17. NumChartEvents

Attribute Type: Ratio.

Justifications: The attribute 'NumChartEvents' is akin to a ratio attribute. It behaves as a continuous attribute with a natural non-arbitrary zero point with inherent meaning. The values do have an order and the distances between values in the dataset are meaningful. In a hospital scenario, the attribute likely refers to the number of chart events that has been recorded for a patient. Such events may include things like vital signs, lab results, and medications. It is possible for a patient to have zero or any positive number of chart events.

Summarising Properties With Visualisations:



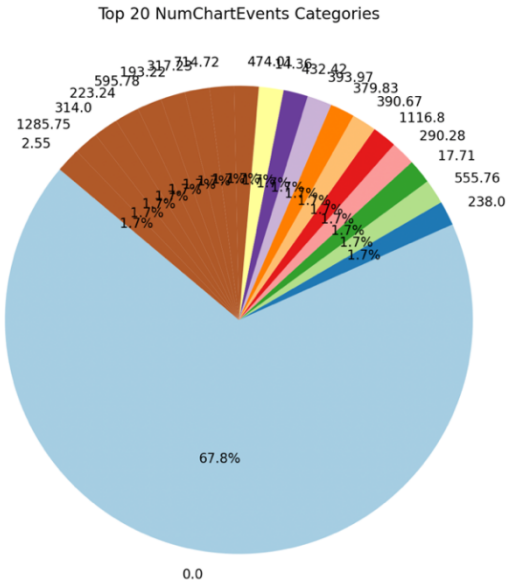


Figure 64: Frequency Distribution (Histogram) Figure 65: Percent. Distribution (Pie Chart)

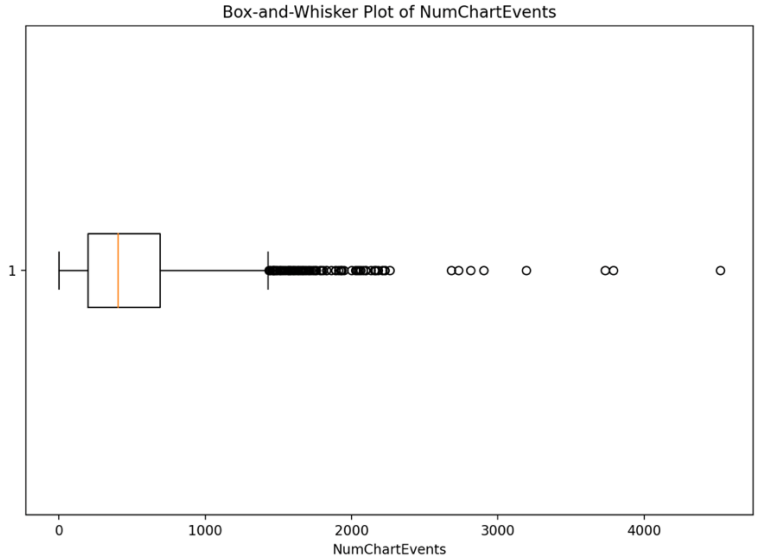


Figure 66: Box-and-Whisker Plot

NumChartEvents	Frequency	Percentage (%)
0.00	80	3.391267
1.15	1	0.042391
2.55	2	0.084782
3.79	1	0.042391
3.96	1	0.042391
4.29	1	0.042391
4.65	1	0.042391
4.80	1	0.042391
5.00	1	0.042391
5.45	1	0.042391
5.65	1	0.042391
5.75	1	0.042391
5.80	1	0.042391
5.83	1	0.042391
6.04	1	0.042391
6.06	1	0.042391
6.38	1	0.042391
7.25	1	0.042391
7.29	1	0.042391
7.31	1	0.042391
7.79	1	0.042391
7.92	1	0.042391
8.00	1	0.042391
8.25	1	0.042391
8.36	1	0.042391
8.37	1	0.042391

Figure 67: Part of The Output from Python

```
+---------------+-----------+--------------------+
| NumChartEvents | Freq. |   Percentage (%)   |
+---------------+-----------+--------------------+
|    0.0     |   80.0    | 3.3912674862229757 |
|    1.15    |    1.0    | 0.0423908435777872 |
|    2.55    |    2.0    | 0.0847816871555744 |
|    3.79    |    1.0    | 0.0423908435777872 |
|    3.96    |    1.0    | 0.0423908435777872 |
|    4.29    |    1.0    | 0.0423908435777872 |
|    4.65    |    1.0    | 0.0423908435777872 |
|    4.8     |    1.0    | 0.0423908435777872 |
|    5.0     |    1.0    | 0.0423908435777872 |
|    5.45    |    1.0    | 0.0423908435777872 |
```

Figure 67: Sample Frequency/Percentage Distribution (based on the diagrams as seen in Figures 64, 65 & 67) – Refer To 'Full Dataset 11' For Complete Records

Measures	**Values**
Mean	508.93 (2.d.p)
Median	403.42 (2.d.p) – As seen in Figure 66
Mode	0 – As seen in Figures 64, 65 & 67
Range	4518.77 – As seen in Figure 66
Variance	202225.76 (2.d.p)
Standard Deviation (S.D)	449.7 (1.d.p)
25th Percentile	198.04 (2.d.p) – As seen in Figure 66
75th Percentile	691.01 – As seen in Figure 66

Table 17: Summary Properties (Based on the 'NumChartEvents' attribute)

18. ExpiredHospital

Attribute Type: Nominal (categorical).

Justifications: The attribute 'ExpiredHospital' can be identified as a nominal (categorical) attribute that is also discrete. The distances between values are not meaningful, and the '0' point is arbitrary. In a hospital context, a value of '0' means that the patient did not die and a '1' means that the patient has died while in the hospital.

Summarising Properties With Visualisations:

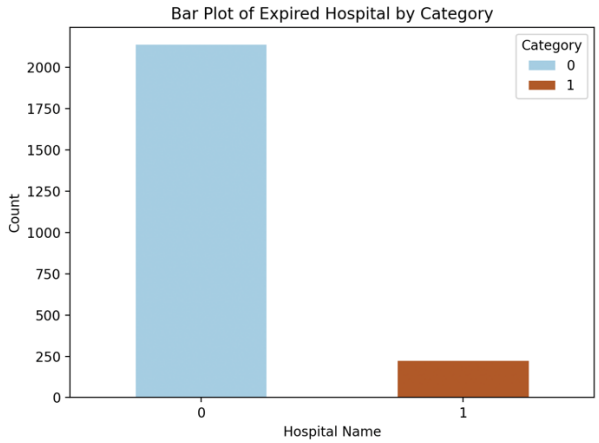


Figure 68: Frequency Distribution (Bar Plot)

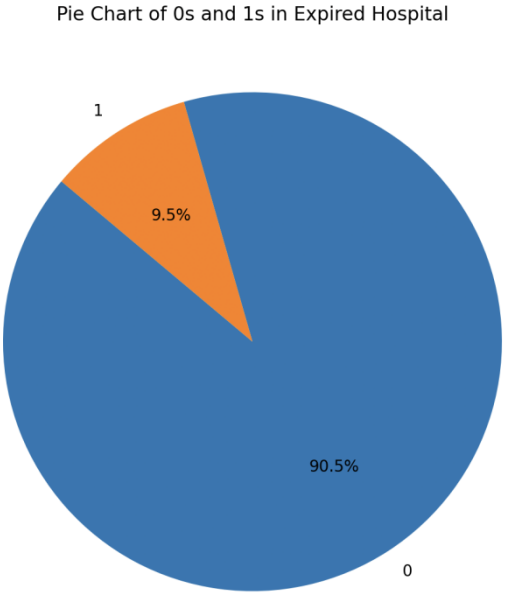


Figure 69: Percentage Distribution (Pie Chart)

Title	**Values**
Frequency Distribution	0 (2136), 1 (223) – As seen in Figure 68
Percentage Distribution	0 (90.5%), 1 (9.5%) – As seen in Figure 69
Mode	0
Count of Unique Categories	2

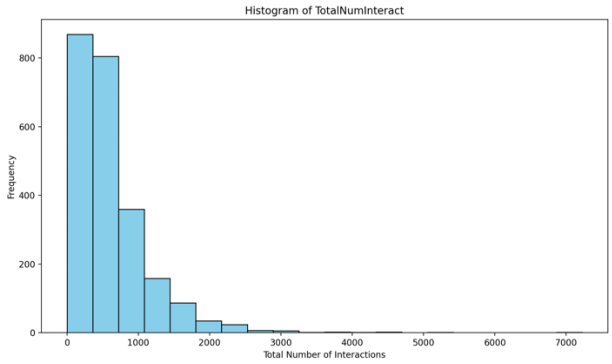
Figure 18: Summary Properties (Based on the 'ExpiredHospital' attribute)

19. TotalNumInteract

Attribute Type: Ratio.

Justifications: The attribute 'TotalNumInteract' correlates to a ratio attribute. It is continuous albeit having a natural and non-arbitrary zero point, with the values displayed as decimals. Such values have an order and the distances between them are meaningful. In a hospital context, 'TotalNumInteract' may reference the total number of interactions that a patient may have had with healthcare providers. Such interactions may include doctor and nurse visits alongside medication administration. It is possible for a patient to have zero or any positive number of interactions.

Summarising Properties With Visualisations:



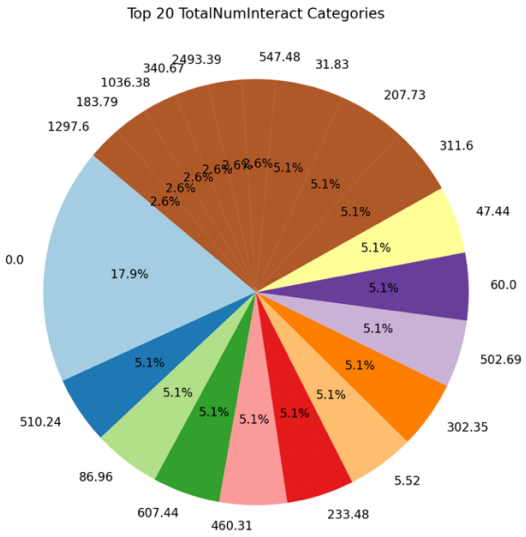


Figure 70: Frequency Distribution (Histogram) *Figure 71: Percent. Distribution (Pie Chart)*

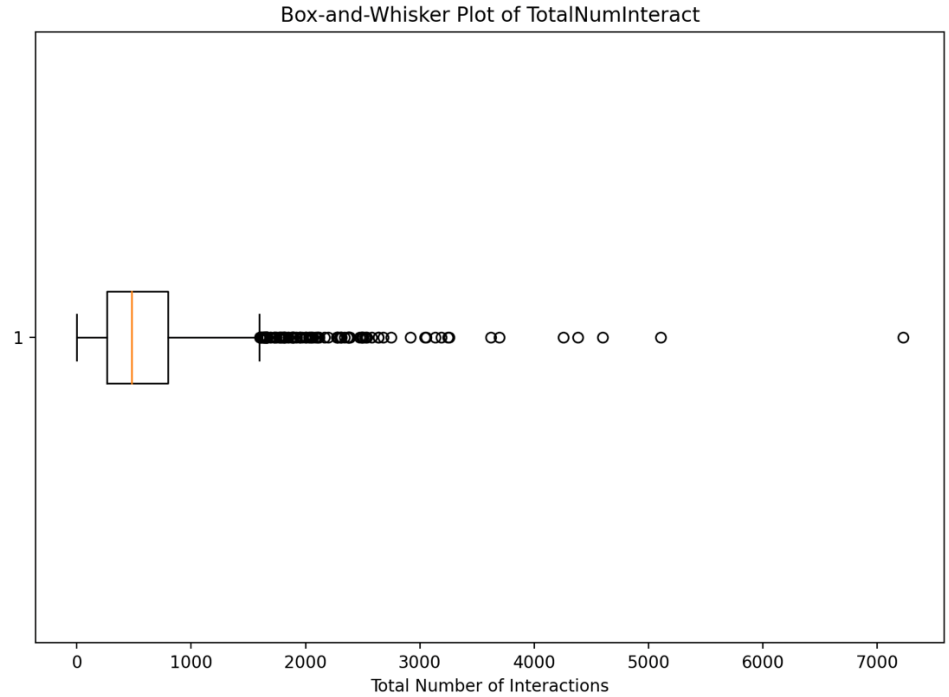


Figure 72: 'TotalNumInteract' Distribution (Box-and-Whisker Plot with Outliers)

TotalNumInteract	Frequency	Percentage (%)
0.00	7	0.296736
1.39	1	0.042391
1.73	1	0.042391
2.82	1	0.042391
2.95	1	0.042391

Figure 73: Part of The Output from Python

TotalNumInteract	Frequency	Percentage (%)
0.0	7.0	0.2967359050445104
1.39	1.0	0.0423908435777872
1.73	1.0	0.0423908435777872
2.82	1.0	0.0423908435777872
2.95	1.0	0.0423908435777872
3.12	1.0	0.0423908435777872
3.49	1.0	0.0423908435777872
3.55	1.0	0.0423908435777872
3.75	1.0	0.0423908435777872
3.76	1.0	0.0423908435777872
3.8	1.0	0.0423908435777872
4.56	1.0	0.0423908435777872
4.68	1.0	0.0423908435777872

Figure 74: Sample Frequency/Percentage Distribution (based on the diagrams as seen in Figures 70, 71 & 73) – Refer To 'Full Dataset 12' For Complete Records

Measures	**Values**
Mean	609.1 (1.d.p)
Median	480.34 – As seen in Figure 72
Mode	0 – As seen in 70, 71, 73 & 74
Range	7225 – As seen in Figure 72
Variance	299772.8 (1.d.p)
Standard Deviation (S.D)	547.52 (2.d.p)
25th Percentile	260.75 (2.d.p) – As seen in Figure 72
75th Percentile	795.08 (2.d.p) – As seen in Figure 72

Table 19: Summary Properties (Based on the 'TotalNumInteract' attribute)

20. Marital Status

Attribute Type: Nominal (categorical).

Justifications: The attribute 'Marital Status' fits the definition of a nominal attribute. It is a discrete attribute with an arbitrary zero point, whereby the values are simply names or labels. In terms of the context, the possible values include 'life partner', 'single', 'widowed', 'married', and 'divorced', which are also classifications of patients. Having 'Marital Status' as an attribute for a hospital database is crucial for healthcare professionals to provide social support, made decisions on treatment, collect data, and comply with regulations. The distances between these values are not meaningful, for instance, it cannot be concluded that 'married' is greater than 'single'.

Summarising Properties With Visualisations:

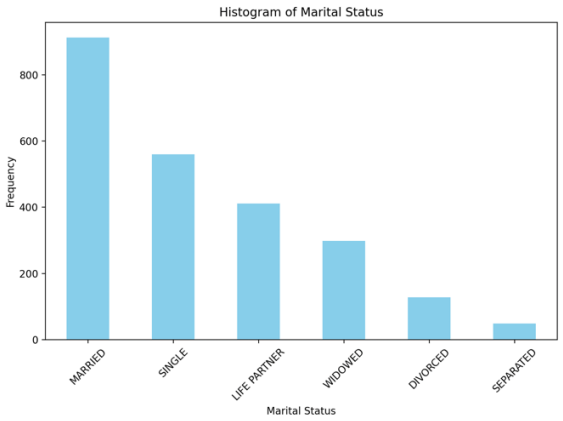


Figure 75: Frequency Distribution (Bar Plot)

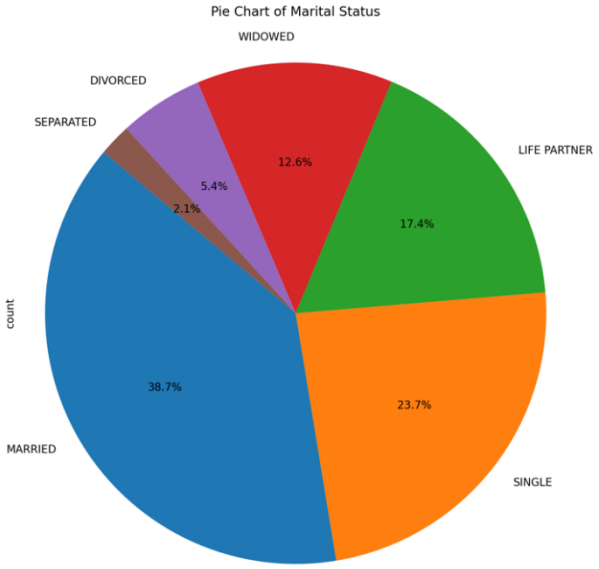


Figure 76: Percentage Distribution (Pie Chart)

marital status	Frequency	Percentage (%)
DIVORCED	128	5.426028
LIFE PARTNER	411	17.422637
MARRIED	913	38.702840
SEPARATED	49	2.077151
SINGLE	560	23.738872
WIDOWED	298	12.632471

Figure 77: Python Output

Title	Values
Frequency Distribution	Divorced (128), life partner (411), married (913), separated (49), single (560), widowed (298) – As seen in Figure 75
Percentage Distribution	Divorced (5.4%), life partner (17.4%), married (38.7%), separated (2.1%), single (23.7%), widowed (12.6%) – As seen in Figure 76
Mode	Married – As seen in Figures 75, 76 & 77
Count of Unique Categories	6

Table 20: Summary Properties (Based on the 'MaritalStatus' attribute)

Exploring Dataset

Outliers

The dataset is centred around hospital records that document patients and their treatment histories. Due to the context being potentially life-threatening, data integrity is vital. However, discrepancies may still arise due to human error or system glitches which may have subtle influence on the quality and accuracy of data. This may lead to the emergence of outliers. The outliers that stem from data collection errors may cause inaccuracies in terms of statistical analysis, so it is imperative to identify and implement proactive measures to minimise them from occurring initially.

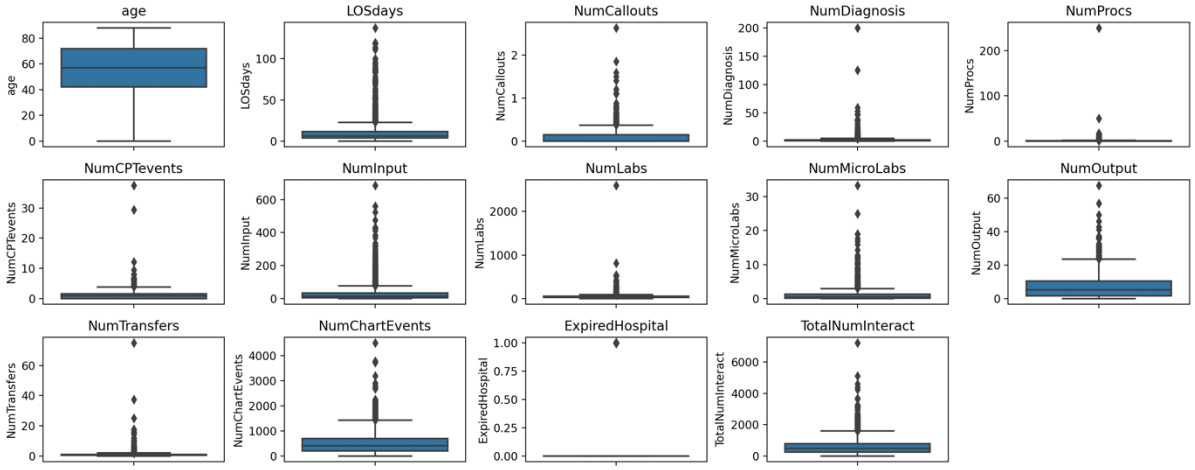


Figure 78: Box-and-Whisker Plots for Outlier Detection

For my personal dataset, all outlier detection processes are applied to numerical attributes only, including those with decimal values. This will enhance the consistency, quantifiability and comparability of the outliers by avoiding multiple data types. The method that was chosen to detect outliers include z-scores, utilising the default value of the z-score threshold ('3') and Python. The application of the technique revealed outliers for 13 out of the 20 attributes in the dataset. This is visually apparent in Figure 78, where the black data points outside of the 'main plot' represents outliers that are present in all attributes except for 'age'.

Here is a summary of the attributes and outliers that was determined:

Attribute Name	**No. Of Outliers**	**Example Values**	**Interpretation**
LOSdays	53	99.79, 54.71, 56.25, 83.17, 137.5	Some interpretations for these high values may be: - People who may have severe medical situations due to an emergency and the hospital need to dedicate ample time to tend to them. - People who require more care due to factors like genetics, age or gender.

			- Such values are considered relatively high; some people are probably there for medical emergencies while others may be there to give birth (examples).
NumCallouts	32	0.61, 2.63, 1.2, 0.75, 0.88	These numbers highlight: - How severe the patients' concerns are, if the hospital need to consult other parties to assist. - The highest values are around 2 to 3 with the average being around 0 to 0.5, meaning that most people have mild conditions that the hospital can resolve by themselves.
NumDiagnosis	16	31.03, 58.82, 46.15, 125, 21.05	Such values provide a snapshot that: - There are people who require more treatment or care than others due to them having higher values. - The hospital needs to prioritise care to these individuals as they may have potentially harmful diseases or conditions that are prevalent.
NumProcs	3	17.65, 50, 250	These values provide a snapshot that: - Since they have higher values than the others in the dataset, they have more medical

			procedures conducted. - It can be said that they are the most at risk of developing chronic and life-threatening conditions, due to the volume of procedures conducted.
NumCPTevents	18	8, 6.77, 12.12, 6.46, 6.75	This data can offer these insights: - There are still some patients who has had many individual services/procedures given to them by the healthcare provider. - The hospital can leverage this data to identify patients who are at risk of readmission due to high numbers of CPT events with a healthcare provider within the last few days.
NumInput	49	190.76, 425.74, 432.92, 187.95, 301.3	The values of NumInput in the dataset may suggest: - If a patient has or has not been infected with a specific bacteria or condition; hospitals can enact effective control measures to mitigate the impact. - If a patient's heart rate, blood pressure and temperature have sudden huge lapses in their values which may mean that more care needs to be dedicated to them.

			- Looking at the values, they are mostly clustered together. - If any becomes a concern, the hospital staff should allocate more time to attend to them.
NumLabs	21	2600, 269.59, 560, 535.29, 271.43	These values of laboratory tests that has been performed on patients suggest: - There's a couple of patients who has had high levels of laboratory tests, they may be identified by the hospital as potential people who may return in the next few days. - Such values can vary based on the condition of the patient and care received, also the healthcare setting.
NumMicroLabs	38	8.62, 25, 7.02, 9.91, 19.05	The values here suggest: - There is a bit of a variation in the microbiological laboratory tests conducted by patients. - A sizeable portion of patients have engaged with microbiological laboratory tests, tracking their infections. - The hospital can use the data to identify who is at the most risk to develop sepsis.
NumOutput	28	29.13, 36.53, 37.04, 31.58, 56.95	The values here suggest:

			- There is a portion of patients who has produced substances like urine. - These substances are presumably for medical testing and analysis. - They may be used to draw the line between those requiring or not requiring more care, particularly those who has higher amounts recorded.
NumTransfers	20	16, 75, 8, 10.34, 7.94	The values here suggest that: - There is a wide variation in how often or frequently a patient is transferred from one medical department to another. - This may associate the patients with increased costs, risks, and disruptions to care. - Hospitals can use the data of patients being transferred during their course of treatment to identify those who may be readmitted in the next few days.
NumChartEvents	37	1924.14, 1932.48, 2094.01, 2210.67, 1889.47	The values here suggest: - The values that are listed as outliers are quite close and clustered together. - Each value correlate to an individual patient, whereby patients with high values of

			'NumChartEvents' may require longer stays in the hospital due to their complex medical issues. - It can be used to monitor the quality of care to patients, as if the average is increasing, physicians or nurses should spend more time with their patients to reduce it as much as possible.
ExpiredHospital	223	1	The values here suggest: - The patient has died while being in the hospital and receiving their treatment. - Even though it may seem like 223 out of 2359 patients has perished is a small number, it has the potential to be more life threatening. - Hospitals can use information like this to identify areas of care where they can improve, conduct research on possible treatment options, while educating their staff to prevent similar complications occurring in the future.
TotalNumInteract	44	3043.04, 2375, 2277.23, 4384.11, 2297.45	The values here imply: - The patients may have had this many interactions with healthcare providers due to their respective conditions which

			are complex problems to solve. - Although such values are high, it still varies a bit as there is a difference between those with chronic conditions and those with acute illnesses. This also holds true with the type of hospital, whether they have integrated or fragmented care. - Hospitals can effectively filter out the patients who have high rates of 'TotalNumInteract' to determine their risk of readmission to the hospital.

Table 21: Attributes and Outlier Information Summary

Outliers, in patient healthcare records, often signal that the data does not have quality and that it is tainted. They should be promptly rectified to support patient safety and clinical decision making. The data may lead to biased estimates since statistics such as the mean and standard deviation are altered. The data may also have severe implications that arises from erroneous or incomplete conclusions.

In healthcare, outliers suggest that there are patients with extreme diagnosis, and it can occur at either end of the spectrum. It is imperative for the hospital or healthcare institution to be alerted and customise treatment options to address a patient's unique circumstances. Those attributes include **'TotalNumInteract'**, **'NumTransfers'** and **'NumOutput'**, which may raise concern during the assessment of patient conditions. The hospital can gauge the severity of a patient's condition based on metrics related to the substances they produce. Furthermore, the medical facility can assess patients' conditions by considering their interactions with healthcare professionals and how they are transferred. In situations where patients exhibit values that are outliers for some of the attributes in table 21 such as **'NumChartEvents'**, **'NumLabs'**, and **'NumMicroLabs'**, despite undergoing treatment for weeks or months, the hospital can consider enhancing their protocols to minimise such occurrences in the future.

Clusters of Similar Instances

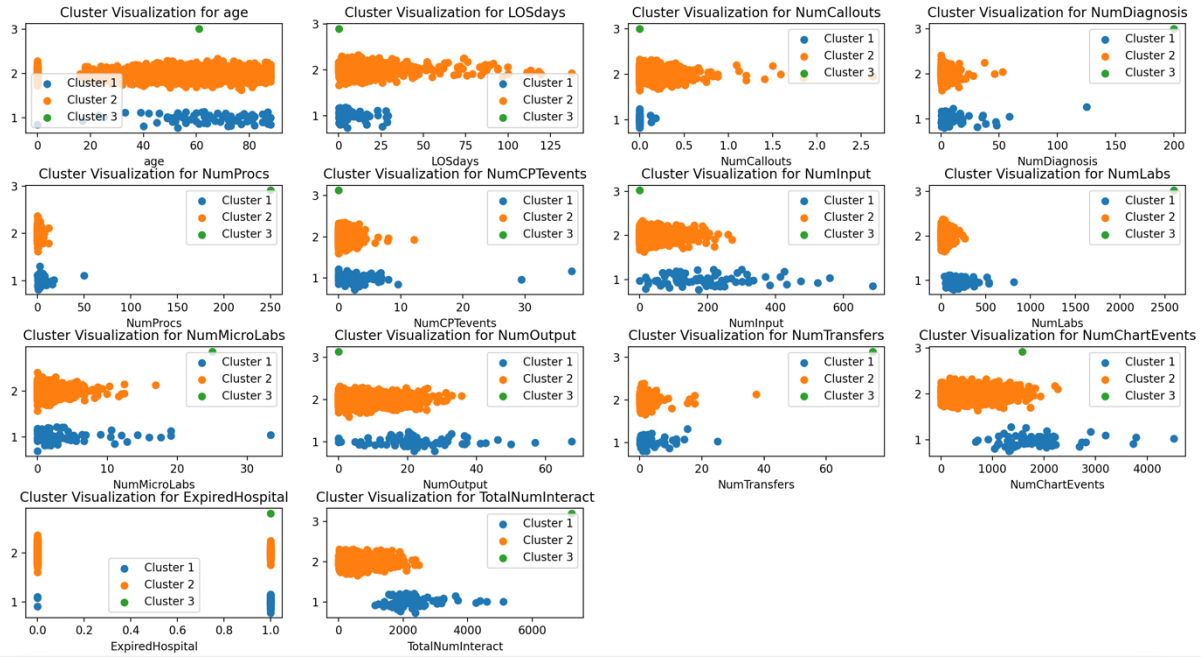


Figure 79: Scatter Plot Matrix (Hierarchical Clustering – Numerical Attributes Only)

As patients are recorded in healthcare databases, there may be multiple who share common traits or are undergoing similar conditions or situations. Specifically, they should have similar medical profiles. The constant accumulation will create clusters, which has the potential impact the statistical analysis by the hospital later.

Clustering can assist hospitals as it means that common traits are identified for them to tailor their treatment plans and options, and the patients can opt for anything there without additional requests being considered. Hospitals can allocate equipment, beds, staff, and the type of therapy, for example those pertaining to mental health. The information can be used to reduce adverse drug reactions and improve medication adherence. It can be used to identify geographic disease outbreaks to guide public health responses after recognising subpopulations with similar genetic profiles for developing targeted treatments. It is also helpful in fraud detection as unusual claim patterns can be found when a patient applies for healthcare insurance. By grouping patients who have similar symptoms, researchers and clinicians can gain insights into disease progression and treatment strategies.

Figure 79 illustrates a hierarchical scatter plot matrix that are of 14 numerical attributes: **'Age', 'LOSdays', 'NumCallouts', 'NumDiagnosis', 'NumProcs', 'NumCPTevents', 'NumInput', 'NumLabs', 'NumMicroLabs', 'NumOutput', 'NumTransfers', 'NumChartEvents', 'ExpiredHospital' and 'TotalNumInteract'.** The other attributes that include categorical are not quantifiable and hence, they are not represented. One-hot encoding can be implemented to rectify the problem, however that may get confusing if many categories are involved. The clusters have been labelled and graphed so that they are able to be clearly seen.

For Figure 79, there are 3 clusters for each attribute that has been generated, however the third cluster is often regarded as an outlier due to the minimal values that belongs to it. From the scatter plots generated, the **'Age'** attribute is the most consistent with the data points creating a positive nor negative correlation that is balanced around the centre of the attribute while

extending past 80. Balanced data points mean that there are both young and older patients in the hospital who are seeking treatment or support. The other attributes all have a positive skew, with data frequency reducing as it approaches extreme values. In terms of **'LOSdays'**, there is a bit of clustering between 0 to 100 on a scale that extends past 125, which means that there are patients with extreme values at either end of the spectrum. Since the range almost covers the entire scale, it can be said that there is a mixture of patients who must stay in the hospital due to the varying nature of their medical conditions.

For **'NumCallouts'**, there is clustering between 0 and 1 on a scale that extends past 2.5, meaning that on average, most people do not need to contact healthcare providers over the phone too many times as their conditions are mild. For '**NumDiagnosis'**, there is clustering between 0 and 50 on a scale that extends past 200, meaning that patients do not have too many diagnosed medical conditions or diseases during their hospital stay on average. In terms of **'NumProcs'**, there is clustering between 0 and 25 on a scale that extends past 250, meaning that on average, the patients have undergone a low number of medical procedures during a predefined timeframe or hospital stays that may be cumulative. There is clustering between 0 and 10 on a scale that extends to 40 for **'NumCPTevents'**, indicating that on average, the patients in the hospital have had less complex treatment plans that comprise their medical history.

For **'NumInput'**, there is clustering from 0 to 300 on a scale that extends past 600. This means that on average, the patients have had relatively low to moderate levels of medical resource utilization like medical tests, medications, or treatments from inputs due to their mild illnesses. There is clustering for **'NumLabs'** between 0 and 500 on a scale that extends past 2500, meaning that many patients have had a relatively low number of laboratory tests on average. It may occur due to a patient's medications, natural variation, diet, nutrition, age, genetics, or simply human error when inputting lab details that led to inaccurate results. For **'NumMicroLabs'**, there is clustering between 0 and 10 on a scale that extends past 30, meaning that on average, the number of microbiological tests that is carried out to patients is relatively low. It can signal the fact that the patients' conditions are less complex, that they are undergoing routine health checks where limited numbers of such tests need to be performed, that they are in their early diagnosis of health conditions and more testing may be needed later, or that they are managing their health well to the point where microbiological testing is not needed as much.

There is clustering between 0 and 40 for **'NumOutput'** on a scale that extends past 60, indicating that on average, there are patients who has had relatively low to medium volumes of substances produced. This may be because their condition is stable, the period is short, or it is a result of the hospital's clinical protocols for documentation. In terms of **'NumTransfers'**, clustering is prevalent between 0 and 20 on a scale that extends past 60, meaning that the patients do not require frequent transfers between healthcare facilities during their treatment. This may be because of their mild medical conditions that can be remedied in few hospital departments with frequent transferrals unnecessary, hospital size, shorter stays, advanced equipment that is available for use, or due to the hospital specialising in providing care for a few conditions that reduces the need for many transferrals. For **'NumChartEvents'**, clustering occurs between 0 and 2000 on a scale that extends past 4000, meaning that a significant portion of patients may not have had high frequencies of medical records documented. This can also highlight variability in data which may occur because of the care obtained, severity of medical conditions, or diligence of healthcare professionals.

The **'ExpiredHospital'** attribute has clustering at either end of the spectrum, '0' and '1', with most people having a '0' recorded. This means that the mortality rate in the hospital is low, which may be because of high-quality care, robust patient safety programmes, specialised services, advanced medical technology, early intervention of patient conditions, or that they have allocated resources efficiently. There is clustering between 0 and 2000 for **'TotalNumInteract'** on a scale that extends past 6000, signifying that there are patients in the hospital who has had both few and moderate interactions with staff. This may either be because of the complexity of their medical conditions or anomalies with the data that has led to inaccurate recordings of such information.

Overall, the information about the typical length of stay, age of patients, and the number of times healthcare providers needs to be contacted, for instance, can be analysed by the hospital to improve their operations in society. They can offer more support if they find a patient who are exceeding the average values of each attribute analysed although their treatment has only recently commenced. The deep analysis of the clusters within the numerical attributes in the dataset provided (as seen in Figure 79) can enable the hospital to enhance their equipment, facilities, and staff. They can also make their treatment options work better, reducing as many side effects as possible for many patients.

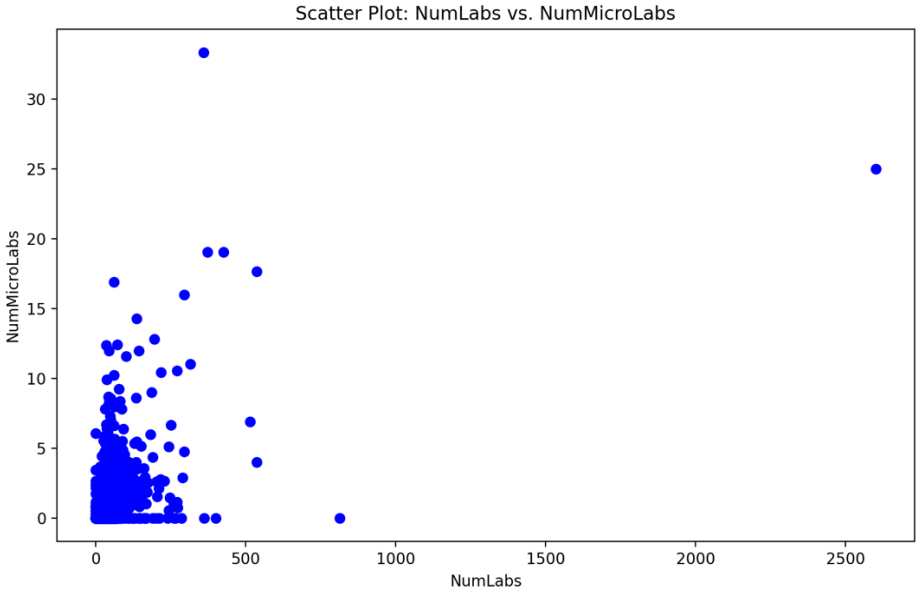


Figure 80: Scatter Plot ('NumLabs' VS 'NumMicroLabs')

The grouping of data points in Figure 80 presents clustering around 0 to 500 marks for **'NumLabs'** and around 0 to 10 for **'NumMicroLabs'**. This means that those patients have similar laboratory and microbiological testing patterns. Such details can assist valuable insights in the identification of patient profiles with stratification, which is beneficial in anomaly detection for a hospital if they do not properly fit into a certain cluster. On average, most people will have higher **'NumLabs'** than **'NumMicroLabs'**, so the hospital may opt to increase the pricing of **'NumLabs'** to remain competitive in the market while complying with government restrictions. The analysis also means that resource allocation by the hospital should prioritise overall laboratory testing rather than microbiological testing for its patients as part of its clinical decision-making protocols.

Interesting Attributes and Relations

The analysis of the given dataset poses peculiar attributes, which are also represented in Figure 78. The attribute of **'NumLabs'** is interesting as the median is close to 0. Due to the context of it being a hospital, it would be expected that there will be a high number of patients who has undergone laboratory testing, but it does not seem to be the case. **'TotalNumInteract'** is interesting, as the data provided does not seem realistic at first glance due to the upper bound being nearly 2000, which is almost an extreme value of its own although there are still higher values than that in the attribute. **'TotalOutput'** is also interesting, as due to the context and nature of the data, it seems wise to assume that the substance output by a patient will be high and the median is not even at the first '20' mark on the y-axis. **'NumProcs'** and **'NumLabs'** all have abnormally high values that stray from the majority, and in general, all the attributes apart from 'age' has data points that are outside the actual box-and-whisker plot. The high values in **'NumProcs'** and '**NumLabs'** may suggest the prevalence of outliers, and the data points outside the box-and-whisker plot may aid in outlier detection processes. Attributes such as **'TotalNumInteract'**, **'NumChartEvents'** and **'NumLabs'** are interesting as their variance is higher than the maximum value that may be stipulated by the range.

The relation of **'NumLabs'** and **'NumMicroLabs'** in Figure 80 seems to be interesting due to how it can be perceived that the higher the value of **'NumLabs'**, the higher **'NumMicroLabs'** will be due to more specialised microbiological tests from the overall laboratory tests being performed on a patient. However, in the figure, there is a value of **'NumLabs'** that is nearly 500 and the corresponding **'NumMicroLabs'** is higher than the data point where **'NumLabs'** is higher than 2500. As a result, it can be said that the prior is only an assumption while being inconclusive as subtle variations can occur across any value. Patients can have underlying reasons for undergoing laboratory testing and they may not always include microbiology testing.

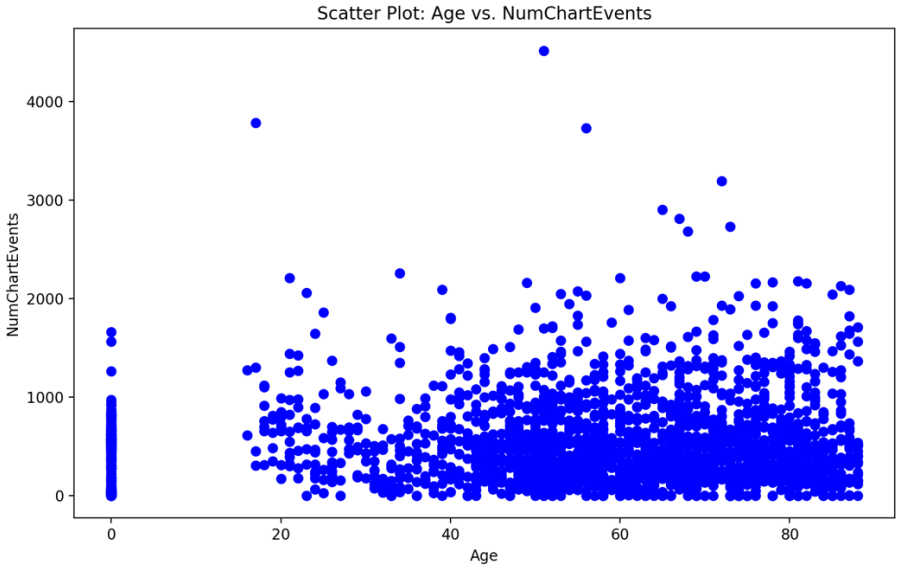


Figure 81: Scatter Plot ('Age' VS 'NumChartEvents')

Figure 81 further presents an interesting relation between **'Age'** and **'NumChartEvents'**. A way of seeing the correlation is that older patients typically have more chart events as they have a greater chance of developing severe chronic conditions, but it is not always true and is therefore not a universal rule. Looking at the data displayed at first glance, the method of inspecting the relationship between the attributes seems accurate as there is clustering around patients who are 40 to 88-year-old as the data points are more concentrated. However, a closer

inspection of the data unveiled insights that it is equally likely that younger people will have the same number of chart events as older people. In fact, the outlier in the young people category is very close to the other ones around the 50 to 60 range. This can occur due to the patients having longer hospital stays and their specific circumstances that can exacerbate medical processes, while variations in data collection systems and protocols can further influence such values regardless of the patient age. Also, there are no patients with **'NumChartEvents'** who are around 0 to approximately 17 years old, which is a bit odd in a hospital and healthcare sense.

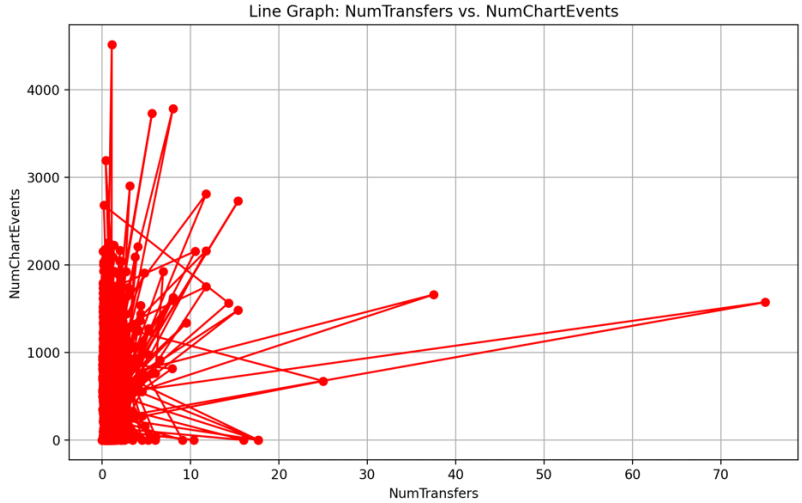


Figure 82: Line Graph Of 'NumTransfers' VS 'NumChartEvents'

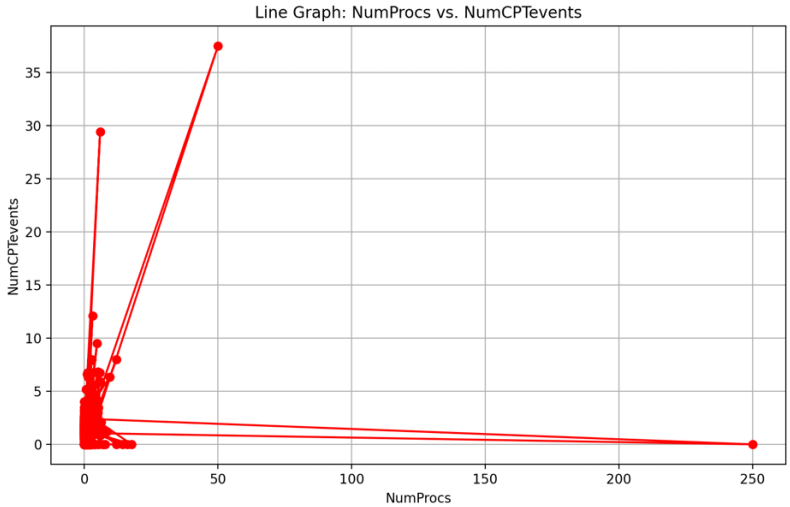


Figure 83: Line Graph Of 'NumProcs' VS 'NumCPTevents'

The data patterns that can be determined from figures 82 and 83 are intriguing.

In Figure 82, it can be noted that there is clustering around 0 to 5 for **'NumTransfers'** and within the range of 0 to 2000 for **'NumChartEvents'**. Normally, it is expected that high values of high **'NumTransfers'** have a direct relationship with high values for the corresponding attribute of **'NumChartEvents'**, suggesting that the number of patient transfers correlates with the number of chart events recorded. However, this relationship can vary based on the factors of patient population, hospital protocols and data collection methods that has been utilised.

They should also be able to explain instances where lower **'NumTransfers'** are associated with higher **'NumChartEvents'**.

Similarly, clustering can be observed at values of around 0 and 25 **'NumProcs'** and around 0 and 7 for **'NumCPTevents'** in Figure 83. A noteworthy anomaly that has been observed is when **'NumProcs'** is at a maximum of 250, **'NumCPTevents'** is at a minimum of 0. The pattern that has been illustrated can be attributed to data entry errors or missing values as it seems unusual to have a substantial number of procedures without corresponding current procedural terminology events. Alternatively, this may highlight how certain clinical procedures may not typically be associated with CPT events and further investigation by the analytics unit of the hospital may be necessary for clarification.

1B Data Pre-Processing (All Completed with Python)

Binning Techniques

Equi-Width Binning

To do this using Python, here are the steps followed:

1. Import the necessary Python libraries; Pandas and NumPy for numerical data manipulation and operations.

```
import pandas as pd
import numpy as np
```

Figure 84: Importing Necessary Python Libraries

2. Load the Excel dataset into a Pandas dataframe, whereby the dataset is the excel file '24747735 – Intro To Data Analytics Dataset.xlsx' and the sheet name is '24747735 – Original Dataset'.

```
# Load the Excel file into a DataFrame
df = pd.read_excel("24747735 - Intro To Data Analytics Dataset.xlsx", sheet_name="24747735 - Original Dataset")
```

Figure 85: How the Excel Dataset Is Loaded

3. Decide on the number of bins required for the respective dataset. Since the intention is to do 'Equi-Width Binning' on the age attribute that has a range from 0 to 88, the square root of 88 is what was used for the number of bins, equating to around 9 bins. However, there are other ways to calculate bins in a dataset, which depends on how the data is supposed to be visualised and understood. The step also involves filtering out the **'Age'** attribute in the dataset and dividing the range of values into intervals that have the same width with the 'pd.cut()' function. The goal is to make the data consistent and meaningful in the respective given context.

```
# Define the number of bins
num_bins = 9

# Perform equi-width binning
df['Equi-Width Bins'] = pd.cut(df['age'], bins=num_bins)
```

Figure 86: How the Bins Are Defined and Equi-Width Binning Is Performed

4. Once that has been completed, the results can be saved back into an Excel file that has the binned column included. The new file has been named 'binned_data.xlsx' and the sheet name is 'Equi-Width Binning (Age)'. The index has been set to false for simpler data processing per data row in the respective column and reduction of memory.

```
df.to_excel("binned_data.xlsx", sheet_name="Equi-Width Binning (Age)", index=False)
```

Figure 87: How the Binned Results Are Saved Back to An Excel Sheet

ui-Width Bins
(68.444, 78.222]
(78.222, 88.0]
(68.444, 78.222]
(48.889, 58.667]
(58.667, 68.444]
(68.444, 78.222]
(58.667, 68.444]
(68.444, 78.222]
(48.889, 58.667]
(48.889, 58.667]
(-0.088, 9.778]
(58.667, 68.444]
(48.889, 58.667]
(78.222, 88.0]
(78.222, 88.0]
(78.222, 88.0]
(68.444, 78.222]
(48.889, 58.667]
(48.889, 58.667]
(48.889, 58.667]
(-0.088, 9.778]
(58.667, 68.444]
(58.667, 68.444]
(48.889, 58.667]
(58.667, 68.444]
(-0.088, 9.778]

Figure 88: Result Of 'Equi-Width Bins' - Part of The Binned Column in An Excel Sheet

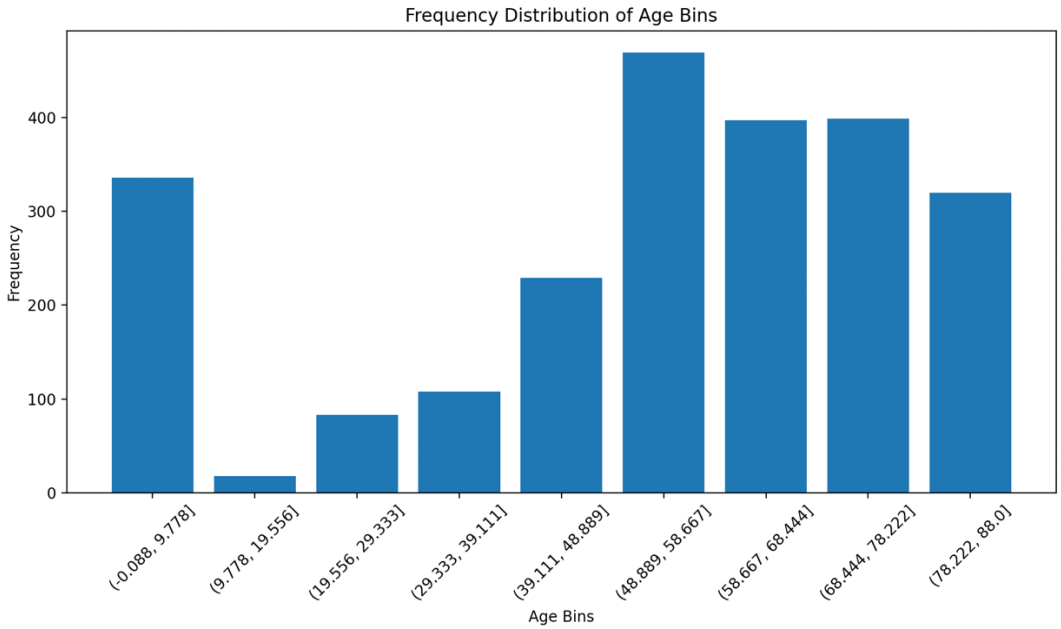


Figure 89: Result of 'Equi-Width Bins' - Frequency Distribution

Equi-Depth Binning

To do this using Python, here are the steps followed:

1. Import the essential Pandas and NumPy libraries for data manipulation and numerical operations.

```
import pandas as pd
import numpy as np
```

Figure 90: Importing Necessary Python Libraries

2. Load the Excel dataset into a Pandas DataFrame, with the file name as '24747735 – Intro To Data Analytics.xlsx' and sheet name as '24747735 – Original Dataset'.

```
# Load the Excel file into a DataFrame
df = pd.read_excel("24747735 - Intro To Data Analytics Dataset.xlsx", sheet_name="24747735 - Original Dataset")
```

Figure 91: How the Excel Dataset Is Loaded

3. Decide on the number of bins that are suitable for the attribute. To ensure conformity across the binning techniques, 9 bins were chosen for exploration. Once that has been done, equi-depth binning should be done after extracting the **'Age'** attribute from the dataset. It involves dividing the data into individual bins with approximately similar data points. The python function 'pd.qcut()' divides the data into quartiles with roughly similar numbers of data points, while 'duplicates = 'drop'' ensures that bins with the same boundaries are merged into one.

```
# Define the number of bins
num_bins = 9

# Perform equi-depth binning
df['Equi-Depth Bins'] = pd.qcut(df['age'], q=num_bins, duplicates='drop')
```

Figure 92: How the Bins Are Defined and Equi-Depth Binning Is Performed

5. Once that has been completed, the results can be saved back into an Excel file that has the binned column included. For clarity, the file name has been altered to 'binned_data1.xlsx', and the sheet_name is 'Equi-Depth Binning (Age)'. The index has been set to false for simplicity as indexing may make the whole process more complex than it should be, and memory reduction.

```
df.to_excel("binned_data1.xlsx", sheet_name="Equi-Depth Binning (Age)", index=False)
```

Figure 93: How the Binned Results Are Saved Back to An Excel Sheet

ui-Depth Bins
(74.0, 80.0]
(74.0, 80.0]
(74.0, 80.0]
(54.0, 61.0]
(67.0, 74.0]
(67.0, 74.0]
(61.0, 67.0]
(67.0, 74.0]
(49.0, 54.0]
(49.0, 54.0]
(-0.001, 38.0]
(61.0, 67.0]
(54.0, 61.0]
(80.0, 88.0]
(80.0, 88.0]
(80.0, 88.0]
(67.0, 74.0]
(54.0, 61.0]
(38.0, 49.0]
(49.0, 54.0]
(-0.001, 38.0]
(61.0, 67.0]
(61.0, 67.0]
(54.0, 61.0]
(54.0, 61.0]

Figure 94: Result Of 'Equi-Depth Bins' - Part of The Binned Column in An Excel Sheet

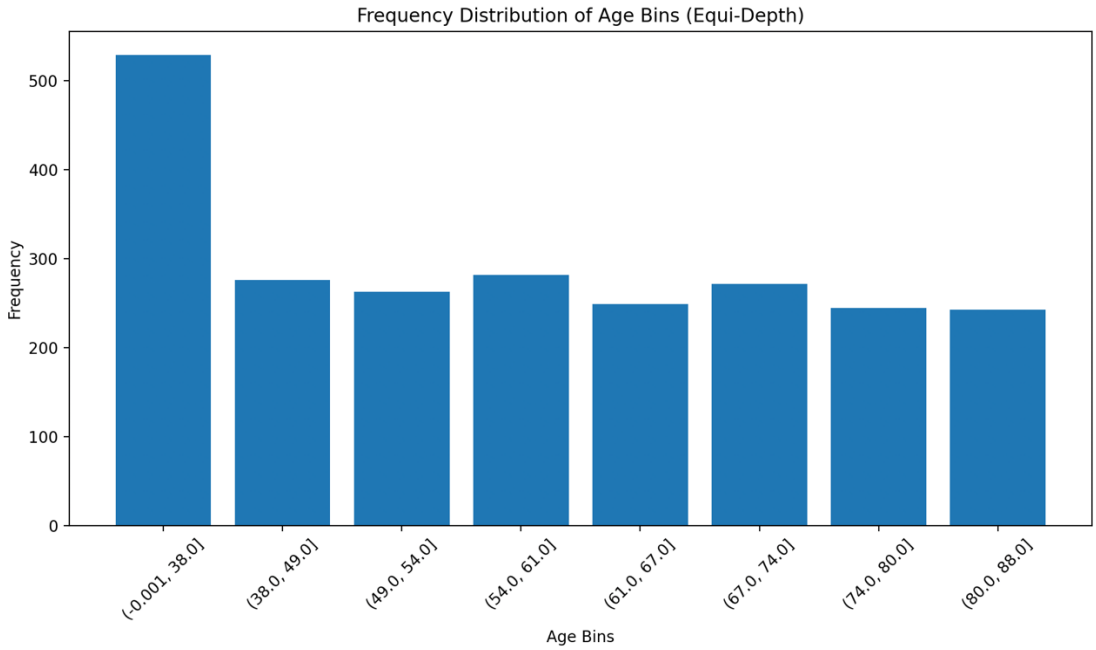


Figure 95: Result Of 'Equi-Depth Bins' – Frequency Distribution

Normalising Attributes

Min-Max

During the normalisation process that uses min-max, Python can prove to be beneficial to automate it. When utilising Python, here are the typical steps:

1. Import the essential Python libraries of Pandas and NumPy for data manipulation and maths operations. This can enable normalisation to occur later.

```
import pandas as pd
import numpy as np
```

Figure 96: Importing Necessary Python Libraries

2. Load the excel dataset using Python into a Pandas DataFrame. The Excel file should be called '24747735 – Intro To Data Analytics Dataset.xlsx' and the sheet name should be called '24747735 – Original Dataset'.

```
# Load the Excel file into a DataFrame
df = pd.read_excel("24747735 - Intro To Data Analytics Dataset.xlsx", sheet_name="24747735 - Original Dataset")
```

Figure 97: How the Excel Dataset Is Loaded

3. Perform the min-max normalization process after extracting the **'NumLabs'** attribute. This process scales the data down to values between 0 and 1 typically, by extracting the attribute and locating the smallest and largest values in the attribute.

```
# Define the range for min-max normalization
min_value = df['NumLabs'].min()
max_value = df['NumLabs'].max()

# Perform min-max normalization
df['Min-Max Norm'] = (df['NumLabs'] - min_value) / (max_value - min_value)
```

Figure 98: Main Process to Perform the Min-Max Normalisation

4. The result with the normalised columns can later be saved back into an Excel file, after specifying the file and sheet name. The index remains false for simplicity and to avoid confusion.

```
df.to_excel("Min-Max (Normalisation).xlsx", sheet_name="Min-Max Normalisation (NumLabs)", index=False)
```

Figure 99: How the Normalised Results Are Saved Back to An Excel Sheet

in-Max Nor
0.006854
0.016531
0.014769
0.022565
0.012308
0.012281
0.015631
0.018362
0.075077
0.021042
0.0047
0.007404
0.012096
0.0214
0.022081
0.011673
0.017885
0.015931
0.008223
0.047065
0.0027
0.014312
0.197612
0.056154
0.014004

Figure 100: Result Of 'Min-Max' - Part of The Normalised Column in An Excel Sheet

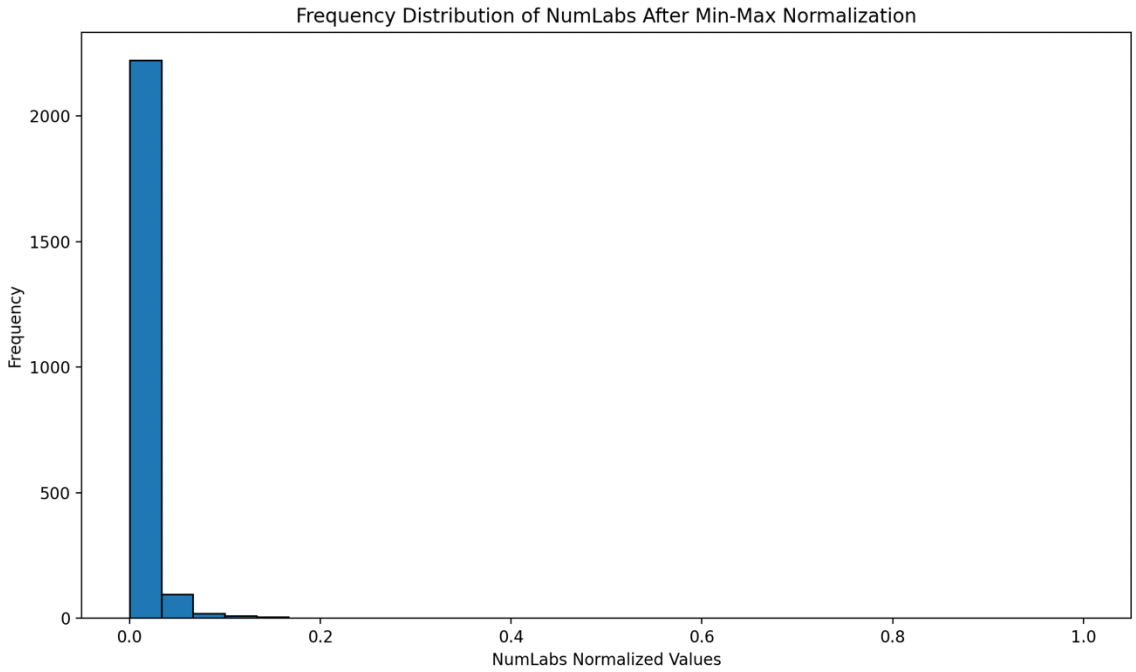


Figure 101: Result Of 'Min-Max' – Frequency Distribution

Z-Score(s)

During the normalisation process that uses the statistical measure of z-scores, Python can prove to be beneficial to automate it. When utilising Python, here are the typical steps:

1. Import the essential Python libraries of Pandas and NumPy for data manipulation and maths operations. This can later enable normalisation to occur.

```
import pandas as pd
import numpy as np
```

Figure 102: Importing Necessary Python Libraries

2. Load the excel dataset using Python into a Pandas DataFrame. The Excel file should be called '24747735 – Intro To Data Analytics Dataset.xlsx' and the sheet name should be called '24747735 – Original Dataset'.

```
# Load the Excel file into a DataFrame
df = pd.read_excel("24747735 - Intro To Data Analytics Dataset.xlsx", sheet_name="24747735 - Original Dataset")
```

Figure 103: How the Excel Dataset Is Loaded

3. Perform the z-score normalization process after extracting the **'NumLabs'** attribute. This process scales the data to have a mean of 0 and a standard deviation of 1. The mean and standard deviation values are used in the calculation to transform the data in the **'NumLabs'** attribute.

```
# Calculate the mean and standard deviation of the 'NumLabs' column
mean_value = df['NumLabs'].mean()
std_deviation = df['NumLabs'].std()

# Perform z-score normalization
df['Z-Score Norm'] = (df['NumLabs'] - mean_value) / std_deviation
```

Figure 104: Main Process to Perform The Z-Score Normalisation

4. The result of the transformed **'NumLabs'** after implementing z-scores can be saved into an Excel file with no indexing and a custom file and sheet name.

```
df.to_excel("Z-Score (Normalisation).xlsx", sheet_name="Z-Score Normalisation (NumLabs)", index=False)
```

Figure 105: How the Normalised Results Are Saved Back to An Excel Sheet

Score Norm
-0.39235
-0.02545
-0.09224
0.20334
-0.18557
-0.18659
-0.05958
0.04396
2.19428
0.1456
-0.47401
-0.37149
-0.19359
0.15916
0.18497
-0.20963
0.02588
-0.0482
-0.34043
1.13224
-0.54984
-0.10959
6.84009
1.47683
-0.12126
-0.52388
0.81653

Figure 106: Result Of 'Z-Scores' - Part of The Normalised Column in An Excel Sheet

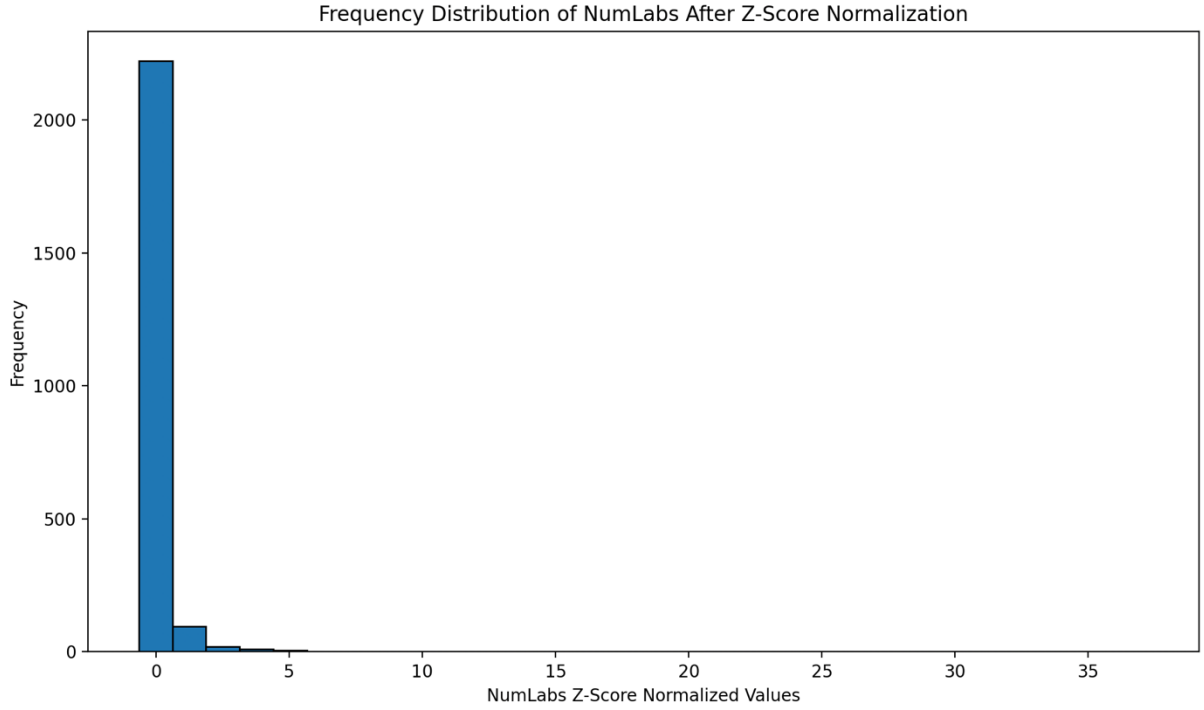


Figure 107: Result Of 'Z-Scores' – Frequency Distribution

Discretising Attributes

During the process to discretise the attribute **'LOSdays'**, Python can assist to automate it. Here is the series of steps that should be implemented:

1. Import the Python Pandas library for data manipulation to occur.

```
import pandas as pd
```

Figure 108: How The Pandas Library Is Imported in Python

2. Load the Excel file with data and its respective sheet as a Pandas DataFrame.

```
# Load the Excel file into a DataFrame
df = pd.read_excel("24747735 - Intro To Data Analytics Dataset.xlsx", sheet_name="24747735 - Original Dataset")
```

Figure 109: How the Excel Dataset Is Loaded

3. Define the bins and labels of the **'LOSdays'** attribute. The bins have been defined as '0-5' (shorter periods), '5-15' (medium periods), '15-50' (longer periods) and '50+' (very longer periods). The python function of 'pd.cut()' is used to discretise it, while 'right = False' means that there would not be repetition of values due to 'end values' not being included as frequencies are calculated for each category.

```
# Define the bins for discretization
bins = [0, 5, 15, 50, float('inf')]  # Use 'float('inf')' to represent 50+

# Define labels for the categories
labels = ['Shorter Periods', 'Medium Periods', 'Longer Periods', 'Very Longer Periods']

# Use 'pd.cut' to discretize the 'LOSdays' attribute
df['LOS Categories'] = pd.cut(df['LOSdays'], bins=bins, labels=labels, right=False)
```

Figure 110: The Main Discretization Process Of 'LOSdays'

4. Calculate the frequencies for each of the four categories that are inherent in the data. The 'value_counts()' function really can assist in this process, and it is now 'LOS Categories' to match the context.

```
# Calculate the frequency of each category
category_counts = df['LOS Categories'].value_counts()
```

Figure 111: Frequency Calculation by Python

5. The frequency counts can then be saved before making it into an Excel file format with custom file and sheet names that has no indexing for simplicity.

```
# Create a DataFrame for the frequency counts
frequency_df = pd.DataFrame({'Category': category_counts.index, 'Frequency': category_counts.values})

# Save the results to an Excel file
frequency_df.to_excel("frequency_of_LOS_categories.xlsx", sheet_name="FrequencyData", index=False)
```

Figure 112: Saving Frequency Counts and Making the Excel File

Category	Frequency
Medium Pe	1058
Shorter Per	913
Longer Peri	336
Very Longe	52

Figure 113: Discretisation Of 'LOSdays' Into Categories and Frequencies

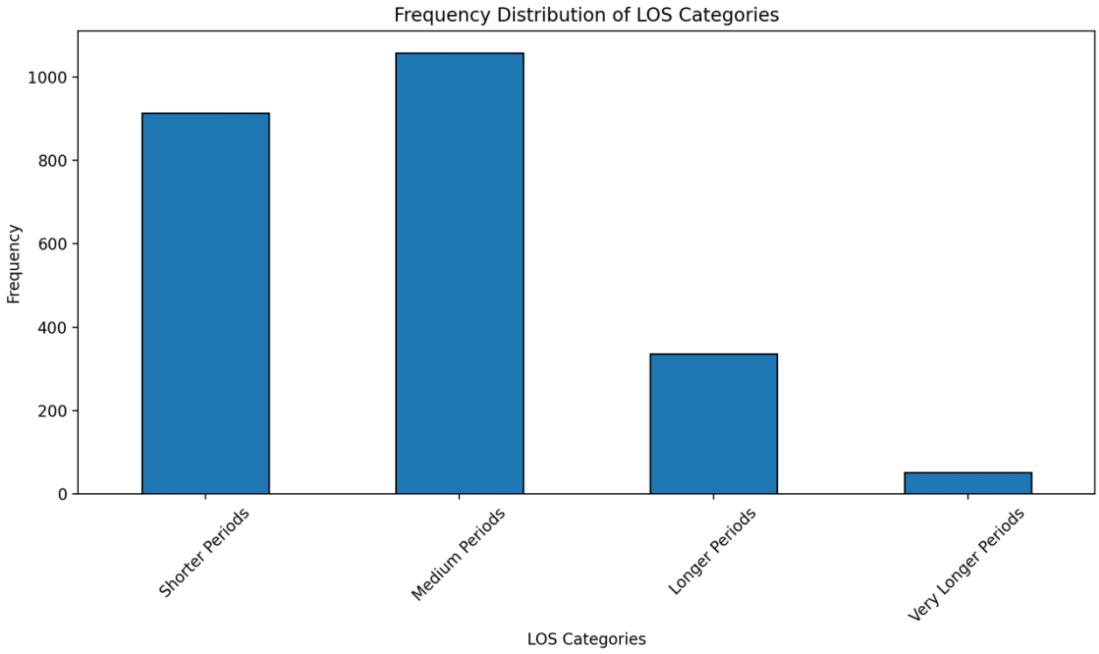


Figure 114: Discretisation Of 'LOSdays' – Frequency Distribution

Binarizing Attributes

Python is an essential tool that can assist with binarizing the **'Marital status'** attribute in a dataset. These are the steps that must be followed to turn the categorical values ('MARRIED', 'LIFE PARTNER', 'SINGLE', etc.,) into binary values ('0' and '1'):

1. Import the necessary Python library of 'pandas' to manipulate data values.

```
import pandas as pd
```

Figure 115: The Necessary Python Library Being Imported

2. Load the Excel dataset file and its related sheet into a Pandas DataFrame.

```
# Load the Excel file into a DataFrame
df = pd.read_excel("24747735 - Intro To Data Analytics Dataset.xlsx", sheet_name="24747735 - Original Dataset")
```

Figure 116: How the Excel Dataset Is Loaded

3. Define and articulate a mapping pattern. This part varies as it depends on which category the dataset should provide greater insights on (i.e., the type of patient the hospital wants to track), and that category will be '1' in binary while all the others will remain as '0'. For the given dataset, it has been decided that the objective is to track the number of 'married' patients, so that has been made to values of '1'.

```
# Define a mapping for binarization
binarization_mapping = {"MARRIED": 1, "LIFE PARTNER": 0, "SINGLE": 0, "WIDOWED": 0, "SEPARATED": 0}
```

Figure 117: The Mapping Pattern That Has Been Determined

4. The 'map()' function can then be used, to transform the 'married' category to 1 while the others remain at 0 within the specific DataFrame.

```
# Apply binarization to the 'marital status' column
df['Binarized Marital Status'] = df['marital status'].map(binarization_mapping)
```

Figure 118: The 'map()' Function And Transforming Categories To Binary Values

5. Save the results in the DataFrame to an Excel file with a custom file and sheet name, but no index for simplicity of data organisation.

```
df.to_excel("binarized_data.xlsx", sheet_name="Marital Status (Binarised)", index=False)
```

Figure 119: Saving the Binarized Results and Making the Excel File

Status (Bina
0
0
0
1
1
1
1
0
0
0
0
1
1
0
1
0
1
1
1
1
0
1
0
1

Figure 120: Binarized 'Marital Status' Attribute (Part of The Column)

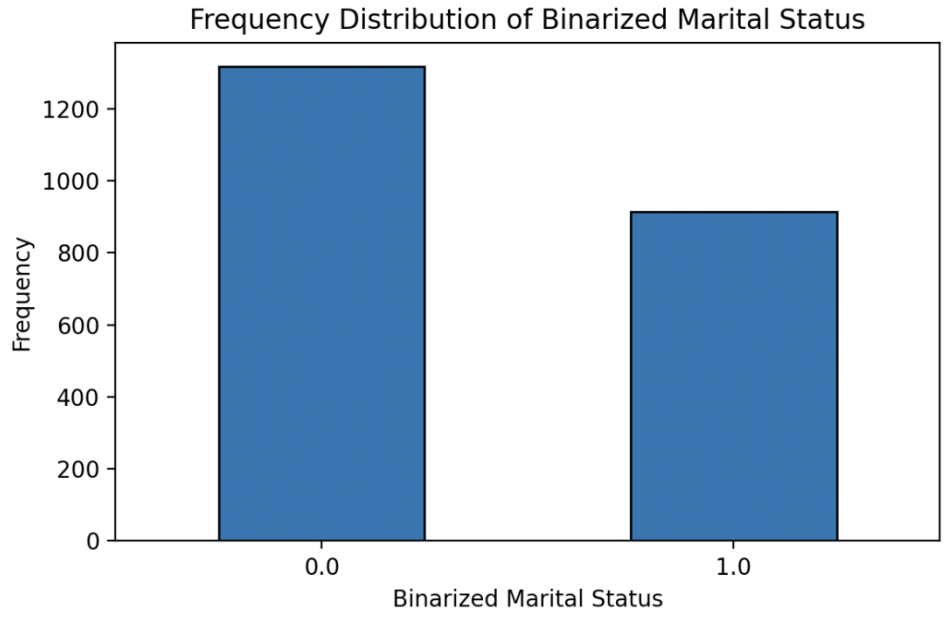


Figure 121: Binarized 'Marital Status' Attribute – Frequency Distribution

1C Summary

Summarisation Of Findings

The report presents an examination of the prescribed dataset and its associated attributes, which has led to two key insights. They are:

- Patient demographics: This is a comprehensive analysis that includes their age, admitted location, length of stay, marital status, admit procedure, and status within the hospital with crucial information such as if they are deceased or not.
- Data characteristics: This delves into correlations between attributes, such as how they are calculated and dealing with 'missing values'.

The general patient demographics has been uncovered from the analysis. With regards to age, there are several conclusions that can be drawn. In Figure 78, most of the patients in the hospital are around 40 to 70 years old. This forms the interquartile range of the prescribed dataset and may alert the hospital to effectively plan their treatment with consent. The devised treatment plan should aim to prevent disease when considering risk and associated medical procedures. The attribute of **'Age'** should also be investigated further by the analytics unit as there are negative values after the binning process has been performed that suggest possible outliers or the skewing of data.

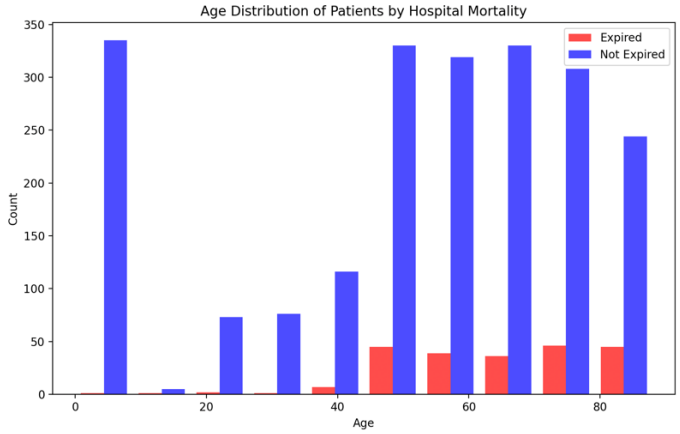


Figure 122: Patient Age Distribution and Mortality

The first part of data characteristics in the analysis has been demystified through figure 116. Figure 122 represents a histogram that illustrates the correlation between ages and whether a patient has died or not in the hospital. This figure supports the assumption that with increases in age, the mortality rate is higher. Such information regarding mortality should be analysed more rigorously by the hospital as is essential for accurately tracking patient outcomes, complying with regulations, conduct research, and providing closure to families. If the hospital finds that patients in a certain age bracket is dying frequently, they can try to reduce and prevent future deaths from occurring.

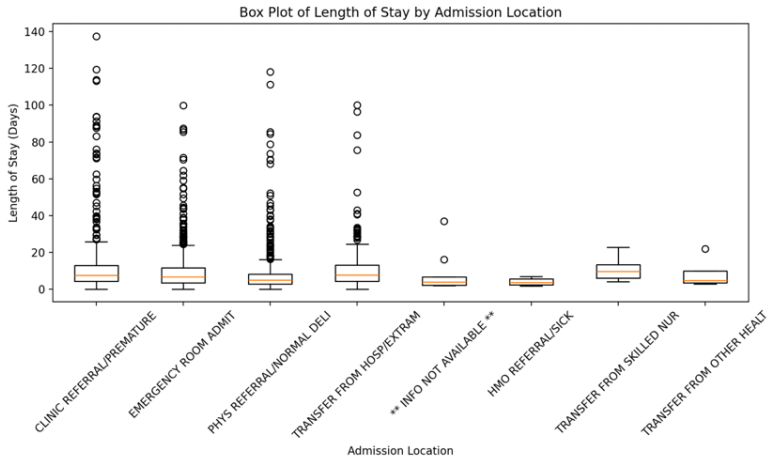


Figure 123: Admission Location VS Length of Stay

The next part of data traits has been presented in figure 123. Figure 123 presents a series of box-and-whisker plots that highlights the length of stay of patients in various hospital locations. This figure suggests that a patient will stay the longest in 'Clinic referral/premature', 'Emergency room admit', 'Phys referral/normal deli' and 'Transfer from hosp/extram'. Those are also the locations with the greatest number of outliers and significant amounts of clustering outside of the 'last whisker'. The analytics unit should rigorously analyse and use this information accordingly for resource allocation as an influx of patients mean that timely care is essential from nurses, doctors, and other healthcare professionals. From the knowledge of the admitted location of patients, they can also protect others by implementing infection control measures if a patient becomes or is diagnosed with a severe disease in the hospital.

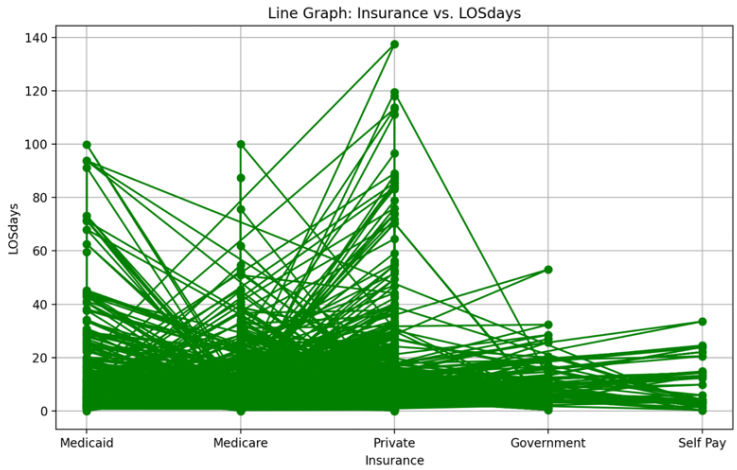


Figure 124: Insurance VS LOSdays

Following on, Figure 124 presents the relationship between **'insurance'** and **'LOSdays'** as a line graph. According to the data points, patients with 'private insurance' tend to stay for the most days in the hospital, followed by 'medicare' and 'medicaid'. There is a lot of clustering between 0 and 100, while the patients with 'government' and 'self-pay' as their insurance provider has clustering around 0 and 20. This is pertinent for rigorous analysis as hospitals need to ensure that claims for billing and reimbursement of patients are handled correctly with insurance providers. It may include preauthorisation for some hospital services by the insurance company, verifying if services are covered by a plan and selecting an efficient method of treatment for the patient.

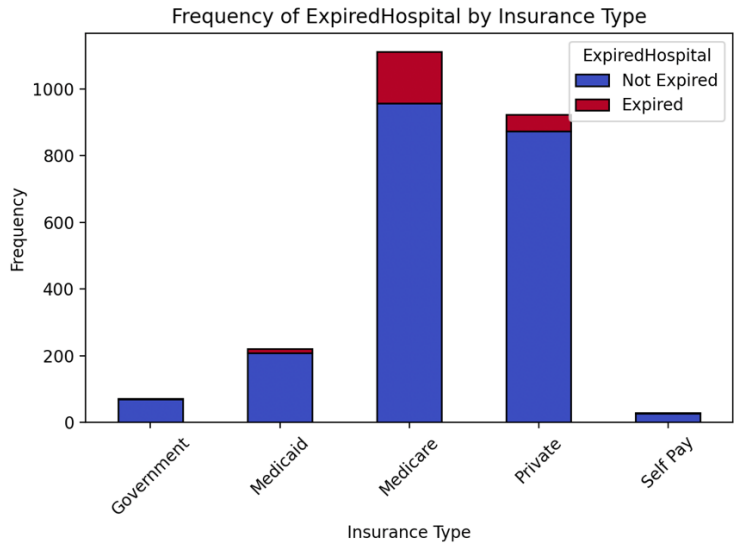


Figure 125: Insurance VS ExpiredHospital

Lastly, Figure 125 presents a stacked bar graph that illustrates the correlation between **'insurance'** and **'ExpiredHospital'**. A key finding from this visualisation is that medicare is the insurance type with the highest frequency followed by 'private'. However, they are also the top two insurance types for most of the patients to die in the hospital. Generally, hospitals need to know such information in a healthcare setting for financial and billing purposes, treatment decisions, resource allocation that includes the quality of care for patients and planning their healthcare policies. Based on the frequency and the number of patients who has died, the hospital can prioritise their analysis of the healthcare coverage of patients with 'Medicare' and 'Private' for research and reporting, which may include offering support and treatment subsides if needed.

For some of the other attributes, key insights gained include the prevalence of 'missing values' with records of 'Info not available' or 'N/A' and data collection protocols. In terms of key ratio attributes such as **'Age'** and **'NumCallouts'**, it was simply not enough to collect the data by rounding each percentage to 1 decimal place upon the creation of a frequency table with categories, frequencies, and the percentage of each category from the whole. Rounding each percentage to 1 decimal place means that the total will not equal to 100%. A much better approach that has been implemented was to use Python to make the entire frequency table for me, but that will need to be in a separate file to reduce cluttering and confusion. The **'NumLabs'** attribute should be analysed further by the analytics team due to negative values appearing in the dataset after 'z-score normalisation' has been performed. This is essential as it can occur due to skewing or data points being a few standard deviations below the mean, and healthcare facilities has legal obligations to track the number of laboratory tests conducted on their patients. The attributes such as **'NumLabs'**, **'NumCallouts'**, **'NumMicroLabs'** and **'NumTransfers'** should be investigated further as they are presented as decimal numbers which should be represented as whole numbers. It does not make sense to have patients who has made 0.31 calls or have patients who has had 0.2 lab tests performed. **'NumProcs'** should particularly be further examined as figure 79 suggests that there is clustering at low values (between 0 and 500 on a scale that extends past 2500), so it is possible that human error occurred when the data was being inputted or it may be a result of protocols at certain healthcare facilities. Attributes like **'TotalNumInteract'**, **'NumChartEvents'** and **'NumLabs'** should also be examined as the descriptive statistic of variance is higher than the

range which incorporates the maximum value. This may be because the data points are not evenly distributed around the range, and they exhibit significant variability around the mean.

Overall, a large-scale analysis of the attributes like **'LOSdays'**, **'Age'**, **'admit_location'**, **'AdmitDiagnosis'**, and **'AdmitProcedure'** has led to enticing findings in patient profiles and conditions for being admitted to the hospital. One of them is the fact that most of the patients in the hospital are married, followed by 'single' and 'life partner' (as seen in Figures 75, 76 & 77). They are aged at around 40 to 70 years old, with most of the patients recording a 0 with healthcare professionals that comprises 17.9% of the total (as seen in Figure 71). 32.4% of the patients in the healthcare records has 0 recorded for the number of procedures that has been carried out, while 4.8% has had 0.38 procedures performed (as seen in Figures 27, 28 & 29). The patients who constitute 13.8% of the total has 'N/A' in 'AdmitProcedure' while 5.1% has been admitted with 'Cont inv mec ven <96 hrs' (as seen in Figures 32 & 33).

In the real world, the information is complex, meaning that analysis by the hospital unit is crucial on a routine basis. The head of analytics for the hospital should also analyse the relations illustrated in Figures 82 and 83, particularly with regards to the anomalies mentioned. This is essential for the hospital to improve their activities in terms of treatment methods and the amount of care delivered to patients, while aiming to contain disease outbreaks from patients if they do eventuate.