Assessment Task II

Niklas Bieda

September 2023

Contents

List of Figures

List of Tables

1 Attributes

1.1 Attribute types

Attribute Name: gender
Attribute Type: Nominal
Justification: The Attribute type of gender is 'Nominal' as it cannot be ordered, it represents discrete units and is used to label variables, that have no quantitative value. These values are male and female.

Attribute Name: age
Attribute Type: Ratio
Justification: The Attribute type of age is 'Ratio' it can be ordered, all mathematical operations are possible and it has clearly defined zero point. Age logically starts with 0 at birth and from there is measured continously in years, months, days and so on.

Attribute Name: LOSdays
Attribute Type: Ratio
Justification: The Attribute type of LOSdays is 'Ratio' it can be ordered, all mathematical operations are possible and it has clearly defined zero point. The shortest possible time a patient can stay is zero days, from there on length of stay is measured similar to age.

Attribute Name: admit_location
Attribute Type: Nominal
Justification: The Attribute type of admit_location is 'Nominal' as it cannot be ordered, it represents discrete units and is used to label variables, that have no quantitative value.

Attribute Name: AdmitDiagnosis
Attribute Type: Nominal
Justification: The Attribute type of AdmitDiagnosis is 'Nominal' as it cannot be ordered, it represents discrete units and is used to label variables, that have no quantitative value.

Attribute Name: insurance
Attribute Type: Nominal
Justification: The Attribute type of insurance is 'Nominal' as it cannot be ordered, it represents discrete units and is used to label variables, that have no quantitative value.

Attribute Name: NumCallouts
Attribute Type: Ratio
Justification: The Attribute type of NumCallouts is 'Ratio' it can be ordered, all mathematical operations are possible and it has clearly defined zero point.

Minimal possible callouts is zero, all other numbers are possible afterwards.

Attribute Name: NumDiagnosis
Attribute Type: Ratio
Justification: The Attribute type of NumDiagnosis is 'Ratio' it can be ordered, all mathematical operations are possible and it has clearly defined zero point. Similar to NumCallouts.

Attribute Name: AdmitProcedure
Attribute Type: Nominal
Justification: The Attribute type of AdmitProcedure is 'Nominal' as it cannot be ordered, it represents discrete units and is used to label variables, that have no quantitative value.

Attribute Name: NumCPTevents
Attribute Type: Ratio
Justification: The Attribute type of NumCPTevents is 'Ratio' it can be ordered, all mathematical operations are possible and it has clearly defined zero point. Similar to NumCallouts.

Attribute Name: NumInput
Attribute Type: Ratio
Justification: The Attribute type of NumInput is 'Ratio' it can be ordered, all mathematical operations are possible and it has clearly defined zero point. Similar to NumCallouts.

Attribute Name: NumLabs
Attribute Type: Ratio
Justification: The Attribute type of NumLabs is 'Ratio' it can be ordered, all mathematical operations are possible and it has clearly defined zero point. Similar to NumCallouts.

Attribute Name: NumMicroLabs
Attribute Type: Ratio
Justification: The Attribute type of NumMicroLabs is 'Ratio' it can be ordered, all mathematical operations are possible and it has clearly defined zero point. Similar to NumCallouts.

Attribute Name: NumOutput
Attribute Type: Ratio
Justification: The Attribute type of NumOutput is 'Ratio' it can be ordered, all mathematical operations are possible and it has clearly defined zero point. Similar to NumCallouts.

Attribute Name: NumTransfers
Attribute Type: Ratio
Justification: The Attribute type of NumTransfers is 'Ratio' it can be ordered, all mathematical operations are possible and it has clearly defined zero point. Similar to NumCallouts.

Attribute Name: NumChartEvents
Attribute Type: The Attribute type of NumChartEvents is 'Ratio' it can be ordered, all mathematical operations are possible and it has clearly defined zero point. Similar to NumCallouts.

Attribute Name: ExpiredHospital
Attribute Type: Nominal
Justification: The Attribute type of ExpiredHospital is 'Nominal' as it cannot be ordered, it represents discrete units and is used to label variables, that have no quantitative value.

Attribute Name: TotalNumInteract
Attribute Type: Ratio
Justification: The Attribute type of TotalNumInteract is 'Ratio' it can be ordered, all mathematical operations are possible and it has clearly defined zero point. Similar to NumCallouts.

Attribute Name: marital status
Attribute Type: Nominal
Justification: The Attribute type of marital status is 'Nominal' as it cannot be ordered, it represents discrete units and is used to label variables, that have no quantitative value.

1.2 Summary or Attribute Properties

1.2.1 Nominal

Mode: The mode of all nominal properties was determined.

	mode
gender	M
admit_location	EMERGENCY ROOM ADMIT
AdmitDiagnosis	NEWBORN
insurance	Medicare
AdmitProcedure	na
ExpiredHospital	False
marital status	MARRIED

Table 1: Mode

To visualize the frequencies of each nominal property a table is not suitable. These are presented here as pie charts or histograms.

Gender The pie chart below (figure 1) shows the absolute and percentage values of the sex of all admitted patients. Female is abbreviated to F and Males are represented through the letter M. 1052 women were admitted to the hospital, which is only 44.6% of all patients, resulting in men being disproportional often admitted to the hospital, with 1307 men being admitted in total (55.4%)

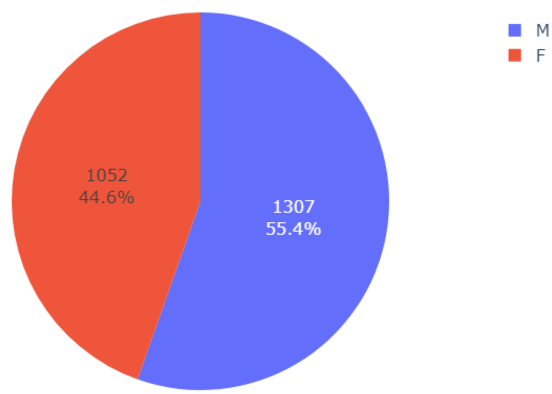


Figure 1: Gender

admit_location the property admit_location describes the location from which the patient was admitted. 39.3% of all patients are admitted in the emergency room, 24.6% are referrals from physicians, 20.9% are referrals from other, another 14% are transfers. The remaining locations of admittance were grouped into other. These have the following tags:

- **INFO NOT AVAILABLE ** - 0.382%
- TRANSFER FROM SKILLED NUR - 0.382%
- HMO REFERRAL/SICK - 0.17%
- TRANSFER FROM OTHER HEALT - 0.127%

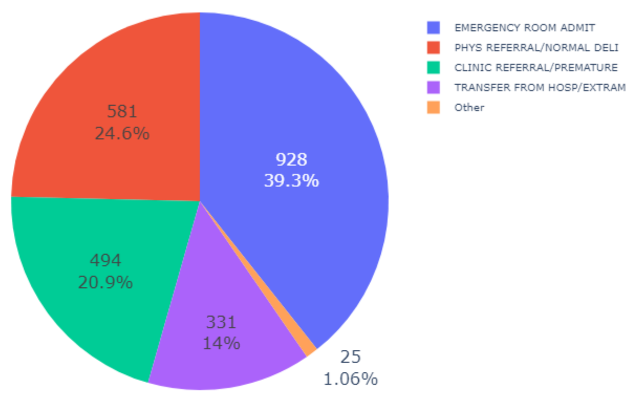


Figure 2: Admit Location

AdmitDiagnosis The property AdmitDiagnosis specifies the diagnosis given upon admission to the clinic. There are a total of 1091 different diagnosis, for only 2359 patients. This makes proper visualization difficult, because a lot of diagnosis only occur once. To give an overview of the data, in the following pie chart only diagnosis with a share of equal or greater than 1% are displayed in the pie chart, all diagnosis smaller 1% are grouped into other.

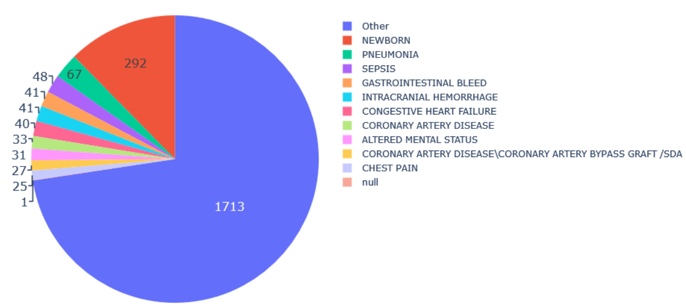


Figure 3: Admit Diagnosis

insurance With 48.2 %, nearly half of all patients are insured via Medicare. The second largest group of patients is privately insured, making up 37.8%. Medicaid is providing insurance to 10% of patients and only a small fraction of all patients are government or self pay (2.76% and 1.19% respectivly).

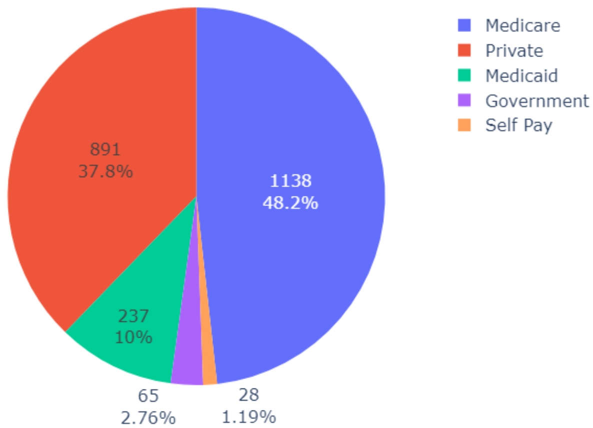


Figure 4: Insurance

AdmitProcedure The procedure upon admittance. This suffers from a similar problem as the property AdmitDiagnosis. In total 369 different procedures for only 2359 patients. Again all procedures occuring less than 1% of all procedures are grouped into other. This still results in 21 different procedures, making a histogram better suited to display the data than a pie chart.

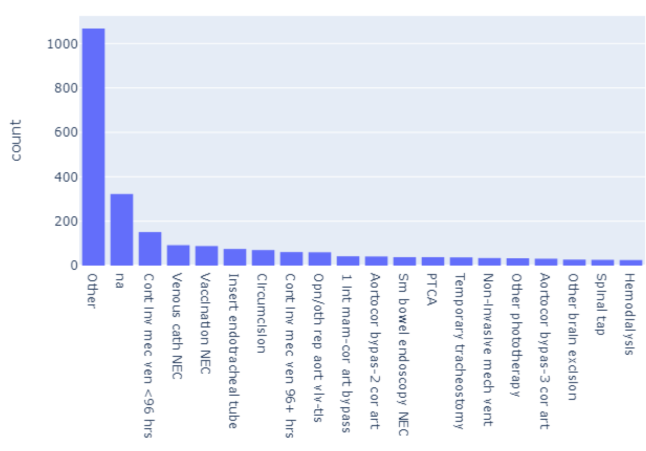


Figure 5: Admit Procedures

ExpiredHospital Describes if the patient died in the hospital or not. In the original data represented by zeros and ones, replaced by True and False, to display the boolen value properly. Zero or false means the patient did not die during his stay in the hospital, one or true means the patient died.
The vast majority of patients did not die, with false making up for 90.7% for all patients. However 9.28% of patients did die during their hospital saty.

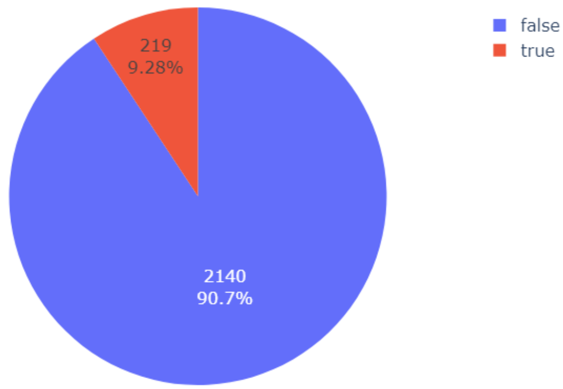


Figure 6: Expired in Hospital

marital status The property marital status is split into six categories:

- married
- single
- life partner
- widowed
- divorced
- separated

No indication is given how these were determined. The biggest group is married, with 41.1% of patients being married. Another 16.5% have a life partner, resulting in 57.6% of all patients being in a relationship. Of the people not being in a relationship, most of these are categorised as single (23.5%). 12.3% are widowed, 5.13% are divorced and 1.44% separated.

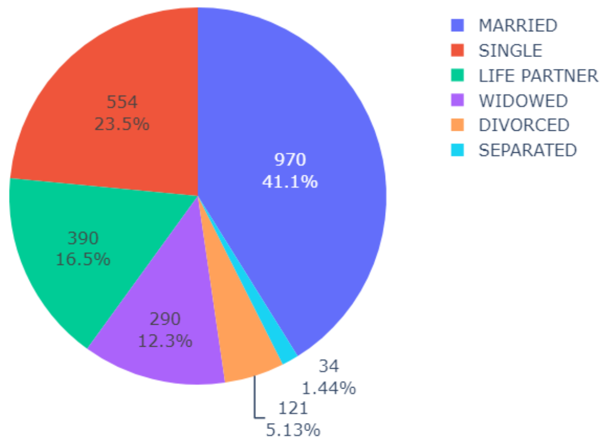


Figure 7: Martial Status

1.2.2 Ratio

All the remaining properties are of the type ratio. This allows for a wide array of statistical analysis. For all properties of the ratio type, mean, median, standard diviation, variance, minimum, maximum and the quartiles were calculated. All calculations and visualisation were performed on a cleaned dataset as described in the outlier section later. This was done to make trends in the data more obvious and figures more readable, since big outliers distorted the scale of the plots. The

	count	mean	median	std	var	min	25%	50%	75%	max
age	2359.00	53.43	58.00	25.74	662.36	0.00	43.00	58.00	73.00	88.00
LOSdays	2315.00	8.89	6.38	7.97	63.50	0.00	3.71	6.38	11.25	46.42
NumCallouts	2328.00	0.09	0.00	0.12	0.02	0.00	0.00	0.00	0.15	0.58
NumDiagnosis	2331.00	2.04	1.43	2.25	5.07	0.00	0.81	1.43	2.42	21.43
NumProcs	2346.00	0.60	0.41	0.79	0.62	0.00	0.21	0.41	0.69	8.00
NumCPTevents	2356.00	1.04	0.98	1.02	1.04	0.00	0.07	0.98	1.58	9.52
NumInput	2308.00	23.52	12.85	28.58	817.06	0.00	4.36	12.85	31.00	157.99
NumLabs	2331.00	39.62	38.01	24.00	576.17	0.00	26.88	38.01	49.33	207.69
NumMicroLabs	2320.00	0.90	0.49	1.09	1.18	0.00	0.14	0.49	1.25	5.99
NumOutput	2341.00	6.61	5.07	6.12	37.47	0.00	1.56	5.07	9.86	27.79
NumTransfers	2344.00	0.88	0.65	0.98	0.96	0.00	0.38	0.65	1.04	11.76
NumChartEvents	2347.00	499.27	402.24	415.56	172687.80	0.00	197.63	402.24	692.85	2356.66
TotalNumInteract	2339.00	587.85	480.38	469.00	219961.44	0.00	254.46	480.38	791.52	2727.65

Table 2: Attribute properties ratio

age The age of the patient is specified through the attribute of the same denominator. A sharp spike can be observed for newborns at the age of zero, followed by no patients younger 16. The amount of patients per age bracket rises until the age of 48, hitting a plateau there and only dropping of for the ages 84 and older. The oldest patient in the data set is 88 years old.

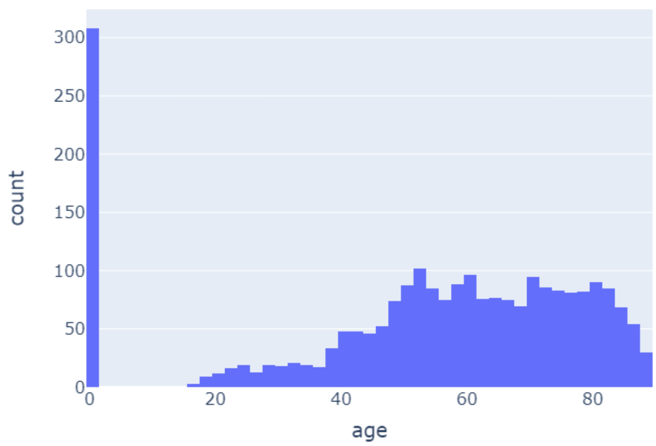


Figure 8: Age Distribution

LOSdays The attribute LOSdays represents the amount of days the patient spent in the hospital. 75% of all patients stay shorter than 12 days. Most patients stay for 5 days. After the 5 day maximum the amount of days stayed seems to drop exponentially. All the following "Num" attributes are given as a daily fraction over the length of stay of a patient.

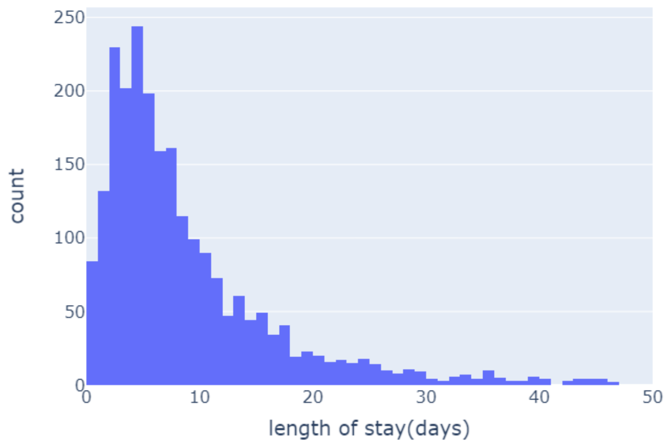


Figure 9: Length of Stay

NumCallouts NumCallouts describes the amount of call outs over the whole length of stay for each patient. Over 50% of all patients didn't have a single call out.

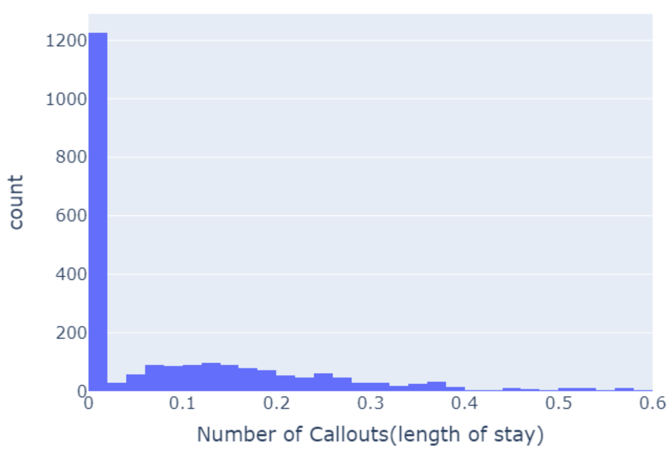


Figure 10: Number of Callouts

NumDiagnosis Similar to NumCallouts, NumDiagnosis describes all diagnosis issued over the duration of stay in the hospital. Reaches it maximum at 0.5

diagnosis for each day per length of stay and decreases exponentially afterwards. The most amount of diagnosis per day were 21.21.

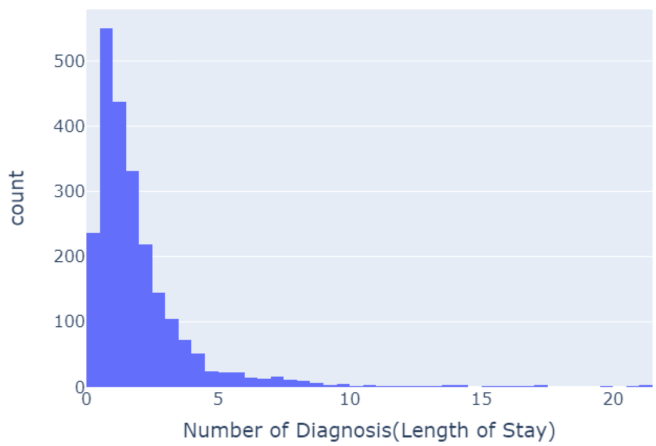


Figure 11: Number of Diagnosis

NumProcs The amount of procedures administered to a patient over his length of stay.

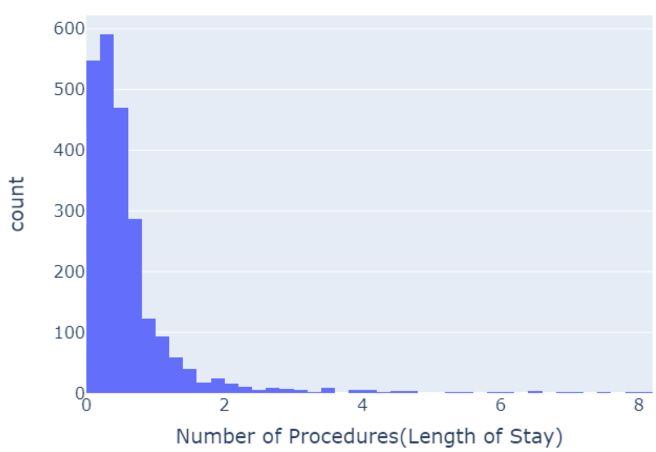


Figure 12: Number of Procedures

NumCPTevents Number of CPT events per patient, again as a fraction over LOSdays. CPT stands for Current Procedural Terminology and is used to describe which treatments were administered to a patient.

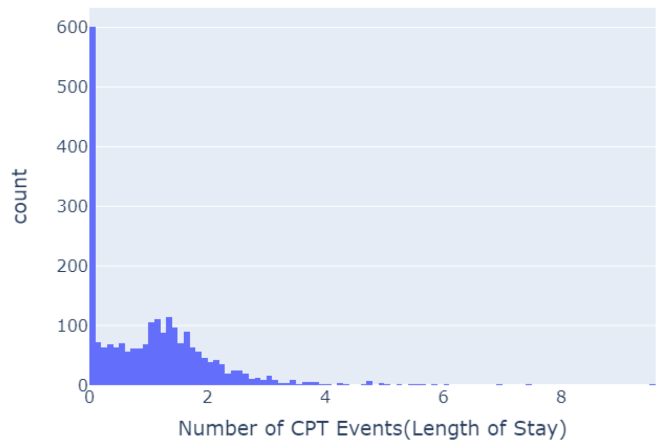


Figure 13: Number of CPT Events

NumInput Number of Inputs per patient, again as a fraction over LOSdays. Inputs refer to fluids, medication and nutrients given to a patient.

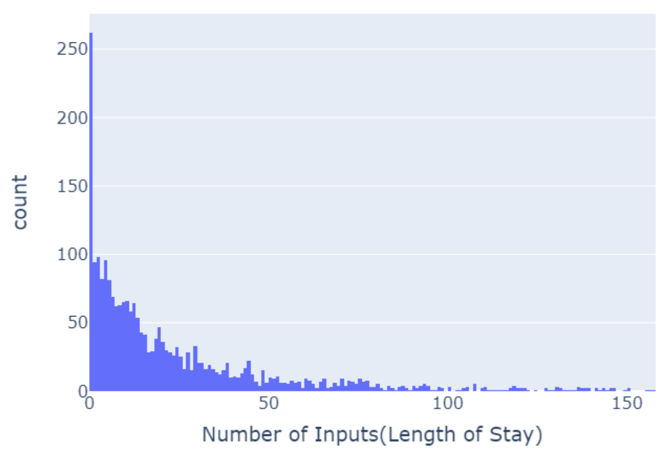


Figure 14: Number of Inputs

NumLabs Number of laboratory tests per patient, again as a fraction over LOSdays.

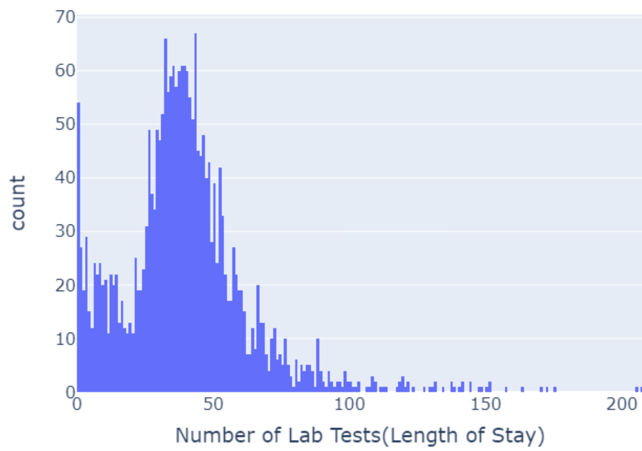


Figure 15: Number of Labs

NumMicroLabs Number of micro laboratory tests per patient, again as a fraction over LOSdays.

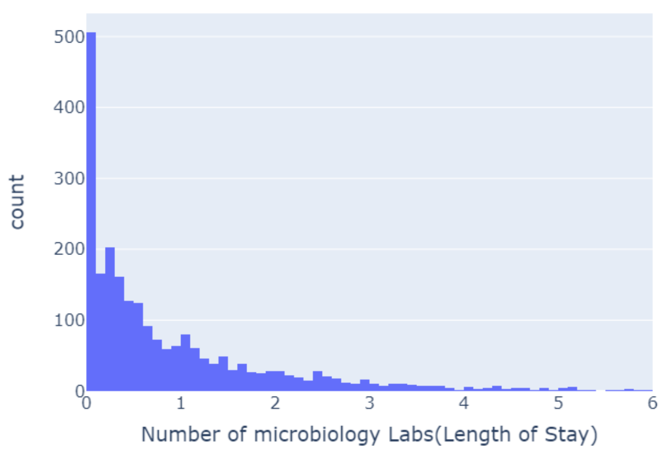


Figure 16: Number of Microbiology Labs

NumOutput Similar to inputs, again as a fraction over LOSdays. Outputs refer to fluids that leave the patient.

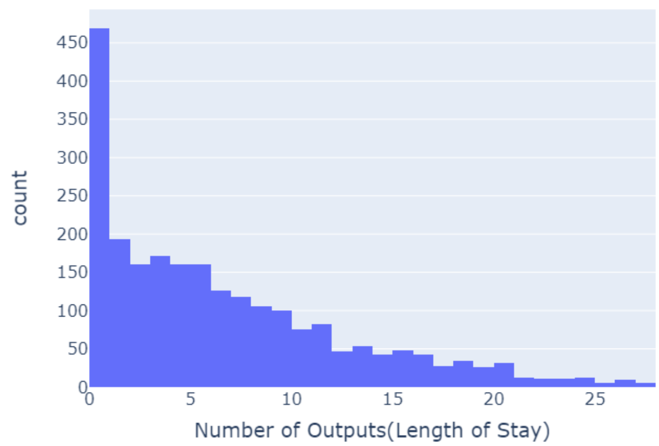


Figure 17: Num of Outputs

NumTransfers Number of transfers per patient, again as a fraction over LOS-days. A transfer is if a patient gets moved from one ward or facility to another.

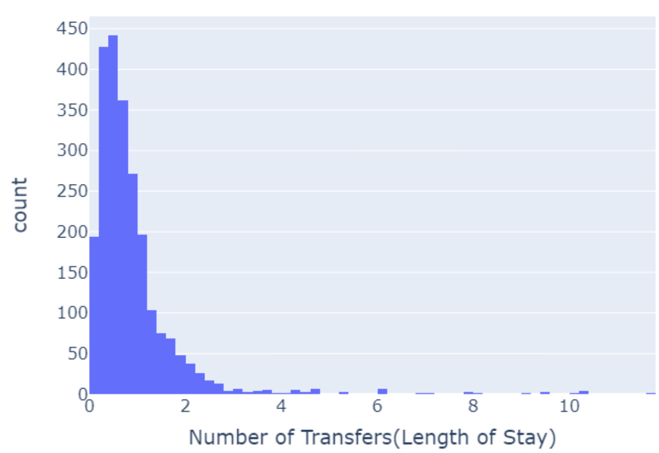


Figure 18: Number of Transfers

NumChartEvents Number of chart events, again given over LOSdays. No clarification is available what chart events are.

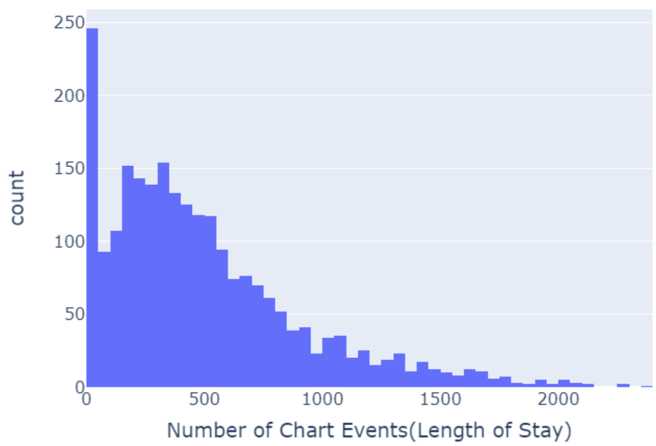


Figure 19: Number of Chart Events

TotalNumInteract Number of total interactions with a patient, again as a fraction over LOSdays.

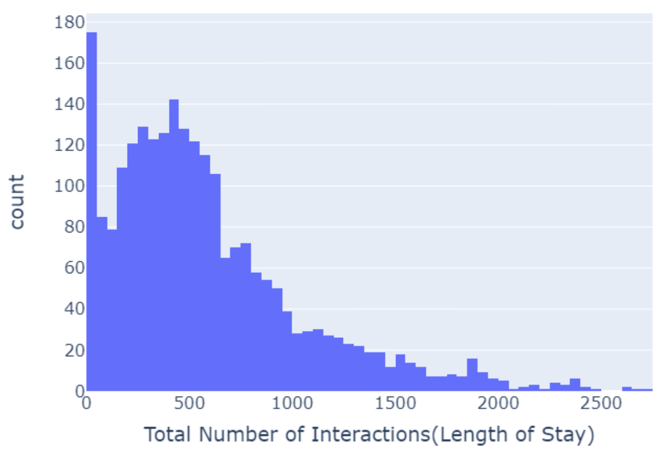


Figure 20: Total Number of Interactions

2 Data Exploration

2.1 Outliers

To get a better understanding of the data and to be able to better identify trends in the data, outliers were determined. To determine outliers, a approach using z-score, to find values outside a certain threshold of standard deviations was used.

To calculate z-score:

$$z = (value - mean)/std \tag{1}$$

By choosing the range of standard deviations, outliers below a certain threshold can be determined. According to the 68-95-99.7 rule, for normal distributed data sets, says that 99.7% of all data points lie within three standard deviations. To only eliminate roughly 0.3% of data points, which contain the biggest outliers, the data set was cleaned by choosing std = 3. This approach is of course only usable for the properties of the ratio type, for nominal values no outliers could be determined.

- 198 outliers over 12 properties with the type ratio were found
- age was the only property of the type ratio without outliers
- with 51 outliers, NumInput had the most outliers, followed by LOSdays with 43 outliers

The following table shows the amount of outliers of each property of the type ratio.

property	number of outliers
age	0
LOSdays	43
NumCallouts	18
NumDiagnosis	23
NumProcs	10
NumCPTevents	3
NumInput	51
NumLabs	28
NumMicroLabs	36
NumOutput	17
NumTransfers	7
NumChartEvents	12
TotalNumInteract	20

Table 3: Outliers

The individual outliers are visualized in the following scatter plot, using a logarithmic y-axis to better show grouping of properties.

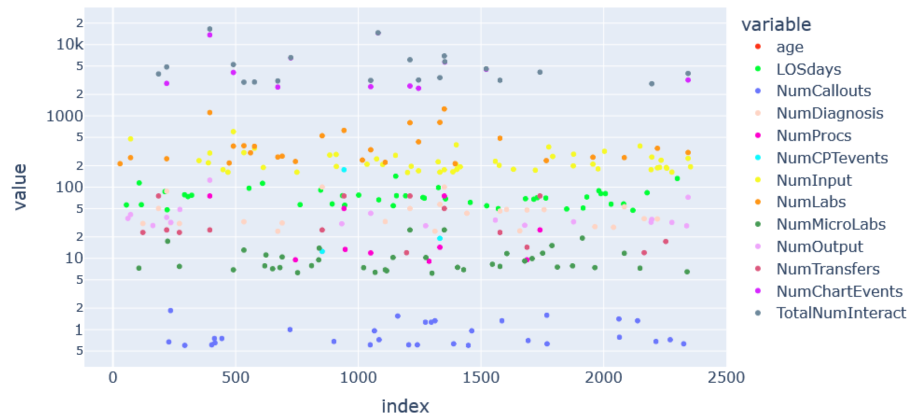


Figure 21: Outliers

The code to do this is presented in the following figure.

```
df_ratio = df.select_dtypes(include=['float64', 'int64'])
df_nominal = df.select_dtypes(exclude=['float64', 'int64'])

outliers = df_ratio[(np.abs(stats.zscore(df_ratio)) > 3)].dropna(thresh=1)
properties_with_outliers = df_ratio.T[(np.abs(stats.zscore(df_ratio)) > 3)].dropna(thresh=1)

df_no_outliers = df_ratio[(np.abs(stats.zscore(df_ratio)) < 3)]
df_no_outliers = pd.concat([df_nominal, df_no_outliers.reindex(df_no_outliers.index)], axis=1)
outliers['age'].unique()
```

Figure 22: Outliers code example

2.2 Correlation

To get an overview of correlation and ”interesting” attributes, a correlation matrix was generated. To be able to correlate the nominal attributes with the ratio attributes, the former got encoded, using discrete numbers (starting at 1 to number of nominal values).

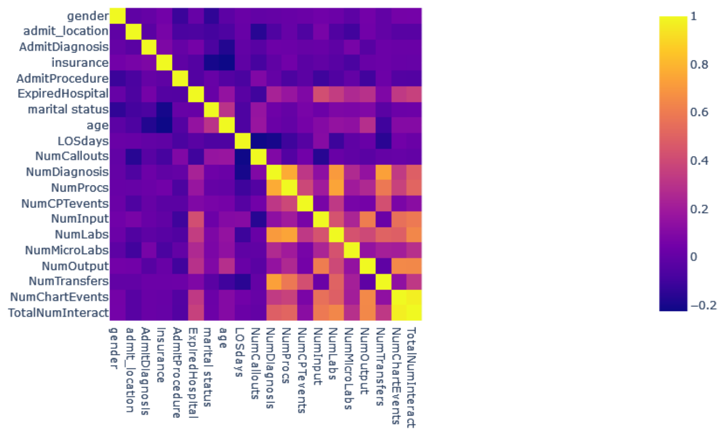


Figure 23: Correlation Matrix

There is a strong correlation for all the "Num" attributes besides "NumCallouts". "ExpiredHospital" is also correlated stronger to the "Num" attributes and to the "age" attribute. Besides these pattern, no interesting or surprising correlations could be found through the matrix.
The most interesting attributes for my analysis were "AdmitDiagnosis", "AdmitProcedure".

2.3 Interessting properties

2.3.1 AdmitDiagnosis

Age effects In Figure 23, all diagnosis upon admittance which were diagnosed at least five times are displayed as a bar-diagram, with age used to color code the individual bars. This threshold includes roughly 43% of all patients.
All patients with the diagnosis "NEWBORN" were zero years old, which is not surprising. This pattern is also true for the diagnosis "HYPERBILIRUBINEMIA". Patients between 20 and 55 were more likely to be diagnosed with a "DIABETIC KETOACIDOSIS" and "OVERDOSE" than older patients. For all other diagnosis, higher age resulted in a higher likely hood to be diagnosed.

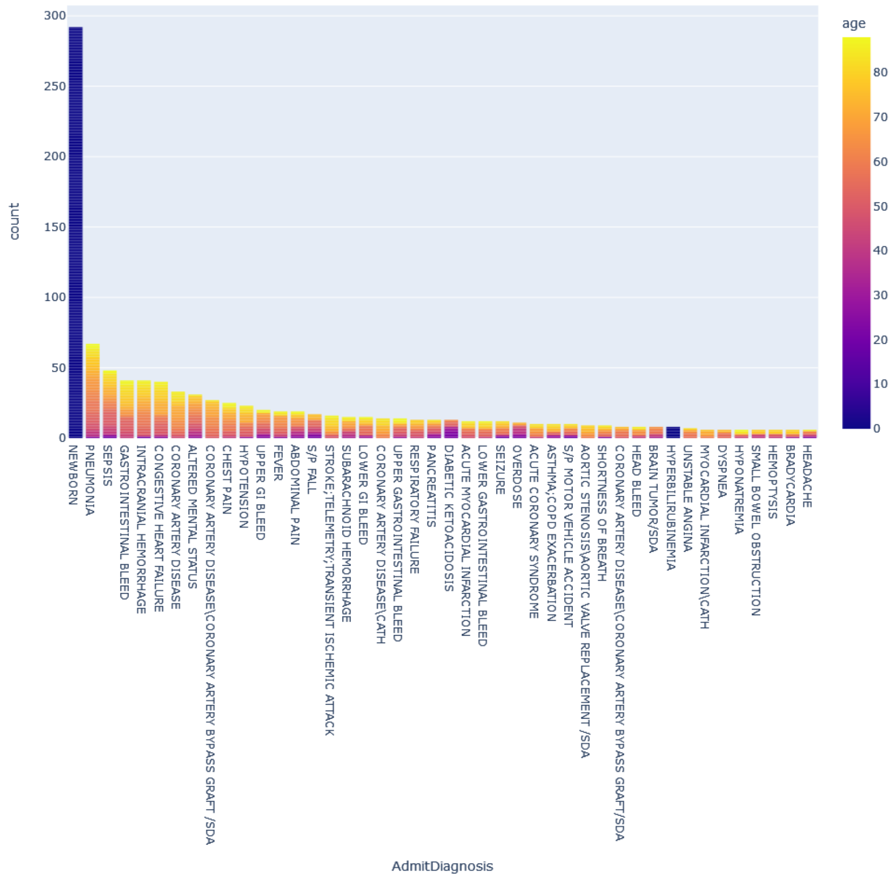


Figure 24: Distribution of diagnosis over age for all diagnosis occurring more than 5 times

Fatality rate To be able to better understand which diagnosis are responsible for the highest fatality rates, the number of deaths per "AdmitDiagnosis" were summed up and a percentage was calculated. Not only the fatality rate is interesting though, but also the total amount of fatalities, to better understand which diagnosis are responsible for the highest amount of fatalities. All these findings are visualised in Fig. 24.

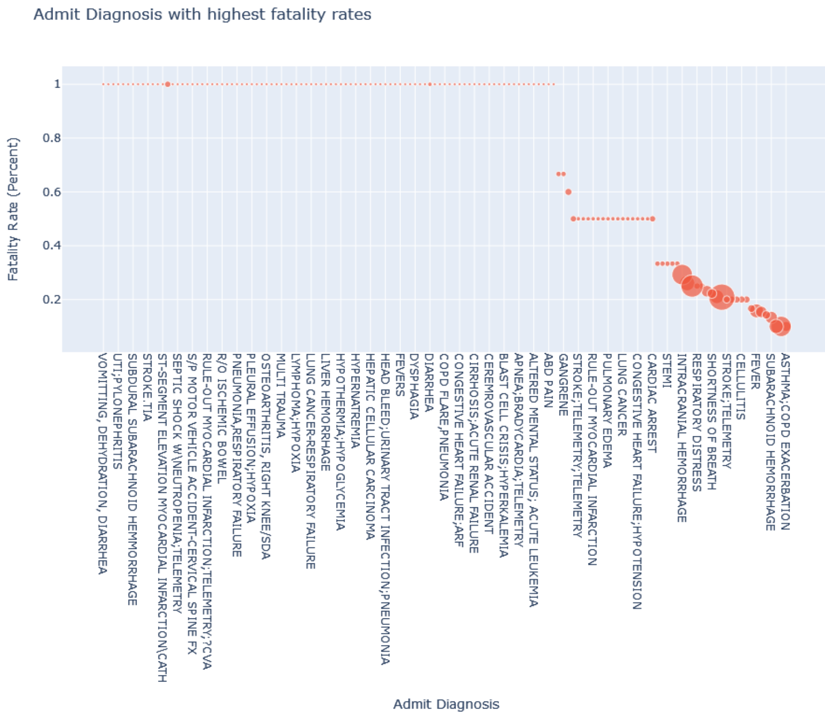


Figure 25: Fatality rate

2.3.2 AdmitProcedure

Deaths per Procedure To get a better overview how often the hospital has to conduct certain procedures and if there were any fatalities these were visualised in Fig. 25. Just because a patient died in the hospital had a certain procedure administered during his admission, this doesn't mean he died during the procedure. I still believe analyzing these properties is interesting, because it enables the hospital to conduct a more thorough analysis, to see if fatality is in some case connected to the procedure administered. All procedures with a smaller occurrence than 0.5% in the dataset were grouped into other. For roughly 2400 entries in the dataset, this means that all "AdmitProcedures" which occur less than 12 times are grouped into other. As shown in Fig. 6, the average fatality rate for the hospital is 9.28%. Of the 833 procedures grouped in other, 71 died, making the fatality rate of other 0.85%. This is lower than the average, resulting in a higher fatality rate for the remaining procedures.

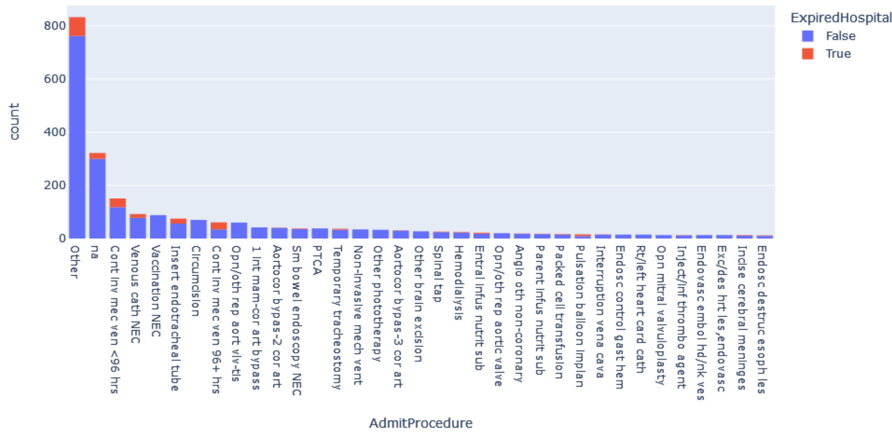


Figure 26: Admit diagnosis and count of fatalities

Death rates vs Procedure For each of the filtered procedures from above the patient fatality rate was calculated. Again this doesn't mean the patients died because of the procedure, it just shows that a patient that was administered these procedures was more likely to die during his stay in the hospital. These are displayed in Fig. 26. Only procedures with a higher "fatality rate" than the average are displayed. These fatalaty rates are up to 5 times higher than the average and offer a great chance to the hospital to review these processes.

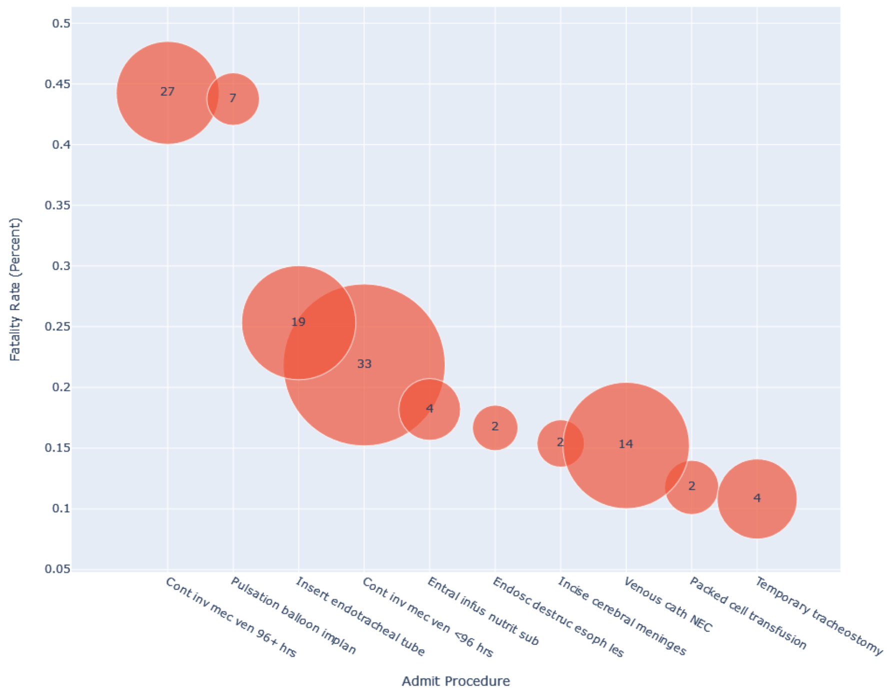


Figure 27: Fatality rate per procedure

3 Data Preprocessing

3.1 Binning

To do the binning, for both cases python with the pandas library was used. Pandas comes with two build-in binning functions. The code example in Fig. 28 includes also a labler function, which is used to generate the labels for the equi-width bins.

```
df_: pd.DataFrame = pd.DataFrame()
age: pd.Series = df_no_outliers['age']
df_['age'] = age

w: float = (max(df_['age'])-0) / (9)

def labler(size: int, step: float) -> list[str]:
    labels: list[str] = []
    i: int
    for i in range(0,int(size)):
        bin_range_low: int = int(min(df_['age']) + i*step)
        bin_range_high: int = int(min(df_['age']) + (i+1)*step)
        labels.append(str(bin_range_low) + '-' + str(bin_range_high))
    return labels

labels_width: list[str] = labler(9, w)

df_['equi-width'] = pd.cut(age, 9, labels=labels_width)
df_['equi-depth'] = pd.qcut(age.rank(method='first'),7, labels=range(0,7))
```

Figure 28: Code example binning

3.1.1 Equi-Width

Equi-width binning, sorts the values into bins of equal width. For the equi-width binning, nine bins were chosen. This results roughly in one bar per decade. At first all labels were calculated using the following formula.

$$w = (max – min)/(bins) \tag{2}$$

This results in bins that follow the following pattern.

$$[min + w], [min + 2w], ..., [min + nw] \tag{3}$$

The pandas dataframe was then binned using the pd.cut(age, 9, labels=labels_width) function to create nine equi-width bins with the previous calculated labels.

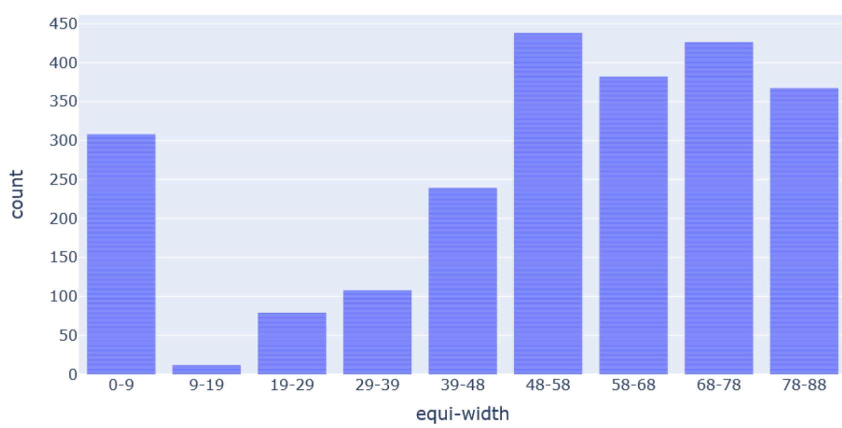


Figure 29: Equi-width bins

3.1.2 Equi-Depth

For equi-depth, all bins should roughly be of the same size. All patients of one age group should be distributed in the same bin, as the edges of each bin should ideally be unique. To achieve this, a single bin must be able to hold at least the largest age group. In this case newborns are by far the largest group of a single age, with 308 members. Dividing the length of the dataset by six results in a bin size of 294.88, which is too small. Dividing by seven results in a bin size of 337, which is just big enough.

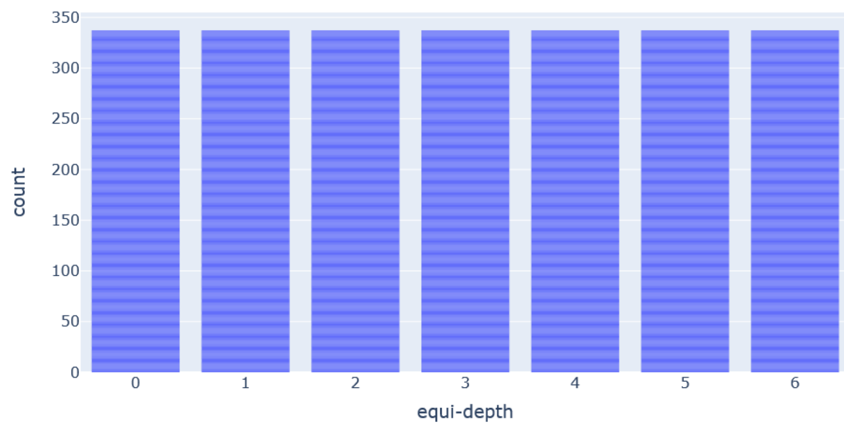


Figure 30: Equi-depth bins

3.2 Normalization

Both normalization methods were done with python using the equations given in the lectures. For min-max normalization the following formula was used.

$$A' = \frac{A - oldMin}{oldMax - oldMin} - (newMax - newMin) + newMin \tag{4}$$

Since in this case newMax and newMin are 1 and 0 respectivly, the latter part of the equation can be ignored.

For z-score normalization the following formula was used.

$$A' = \frac{A - mean}{std} \tag{5}$$

```
num_labs = df_no_outliers['NumLabs']
num_labs_normalized_min_max = num_labs.copy()

minimum = min(num_labs)
maximum = max(num_labs)
for i in range(len(num_labs)):
    num_labs_normalized_min_max[i] = (num_labs[i] - minimum) / (maximum - minimum)

num_labs_normalized_zscore = num_labs.copy()
mean = np.mean(num_labs)
std = np.std(num_labs)
for i in range(len(num_labs)):
    num_labs_normalized_zscore[i] = ((num_labs[i]-mean)/std)

df_ = pd.DataFrame()
df_['min_max'] = num_labs_normalized_min_max
df_['zscore'] = num_labs_normalized_zscore
```

Figure 31: Code example normalization

For the min-max normalization this resulted in all values being in the range of 0 to 1, while z-score normalization, the values are between -1.65 and 7.00. All values are displayed in Fig. 32.

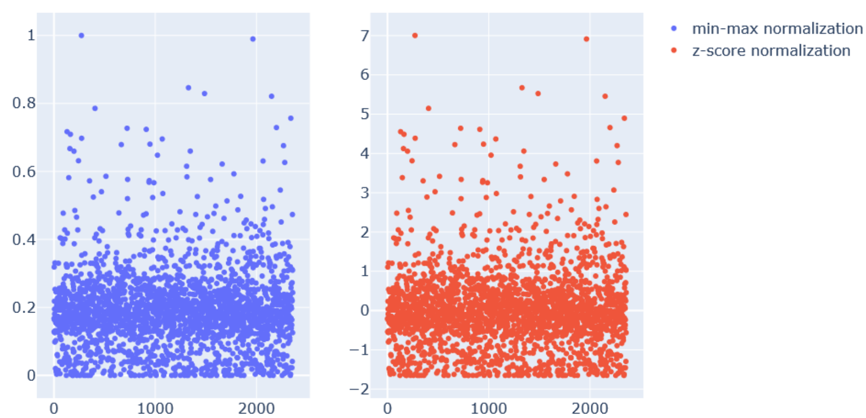


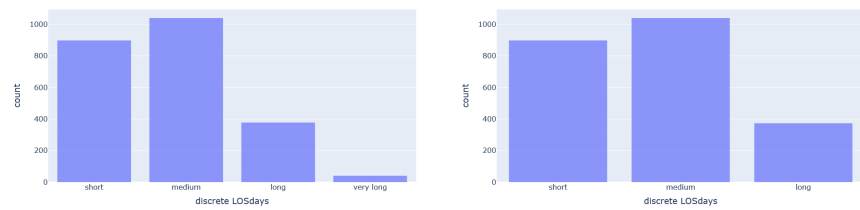
Figure 32: Code example normalization

3.3 Discretization

To transform the "LOSdays" attribute into discrete values, the same pandas function as for the equi-width binning was used. This time the discretization was done one time on the data including the outliers and one time without, as there are no patients in the very long bracket in the cleaned data set.

```
df_ = pd.DataFrame()
los_days = df['LOSdays']
discrete_los_days =pd.cut(los_days, bins=[0, 5.0, 15.0, 50.0, max(los_days)],labels=['short','medium','long','very long'])
df_['discrete LOSdays'] = discrete_los_days
df_
```

Figure 33: Code example discretization



(a) Raw data set

(b) Cleaned data set

Figure 34: Three simple graphs

3.4 Binarization

To binarize the property marital status, the dataset needs extra columns for each possible unique value in marital status. As there are six unique values the dataset needs to be expanded by six columns. These are named after the six unique values. If a patient is for example married, the value in married is set to one and in all other columns to zero.
Pandas again provides a function to do this efficiently.

```
df_ = df_no_outliers.copy()
df_ = pd.get_dummies(pd.DataFrame(df_.pop('marital status').values.tolist()), prefix='', prefix_sep='')
df_ = df_.replace({True: 1, False: 0})
df_
```

Figure 35: Code example binarization

4 Summary

This report shows following insights derived from an analysis of medical data. First Outliers were identified through z-score analysis, with 198 outliers detected across 12 ratio-type properties. Notably, "NumInput" and "LOSdays" had the highest numbers of outliers, while age was the only attribute of the type ratio without outliers. No outliers could be determined for nominal values, since these consist of fixed value ranges.

Strong correlations were observed among the "Num" attributes, except for "NumCallouts." Additionally, "ExpiredHospital" exhibited stronger correlations with both "Num" attributes and "age."

Gender and age play a crucial role in determining the chances to be hospitalised. Escpecially age has a crucial influence on the diagnosis. Conversely, diagnoses such as "DIABETIC KETOACIDOSIS" and "OVERDOSE" were more prevalent in patients aged between 20 and 55. While all other diagnosis seem to be correlated with older age.
Most of the diagnosis with a 100% fatality rate only occur once, but elevated myocardial infarction, as well as diarrhea occur 4 and respectively 2 times. More relevant in total numbers in respect to the fatality rate were diagnosis in the 20 to 30% fatality rate which accounted for a total of 65 of the 219 fatalities.

A few of the admit procedures were followed in more patient deaths than others. Even these might be not direct results of the procedures, the analysis of this data could be helpful to review processes in the hospital and maybe improve care for its patients.
Patients that were target of the procedure "cont inv mec ven 96+ hrs" were especially in danger of a death during the stay in the hospital. Nearly 45% of these died during their stay.
The same procedure but in another time frame ("cont inv mec ven ¡96 hrs") also shows higher patient fatality rate, with 21% of patients receiving this procedure dying during their hospital stay. In total these two procedures have 60 patient deaths connected to them which is nearly a third of all patient deaths.