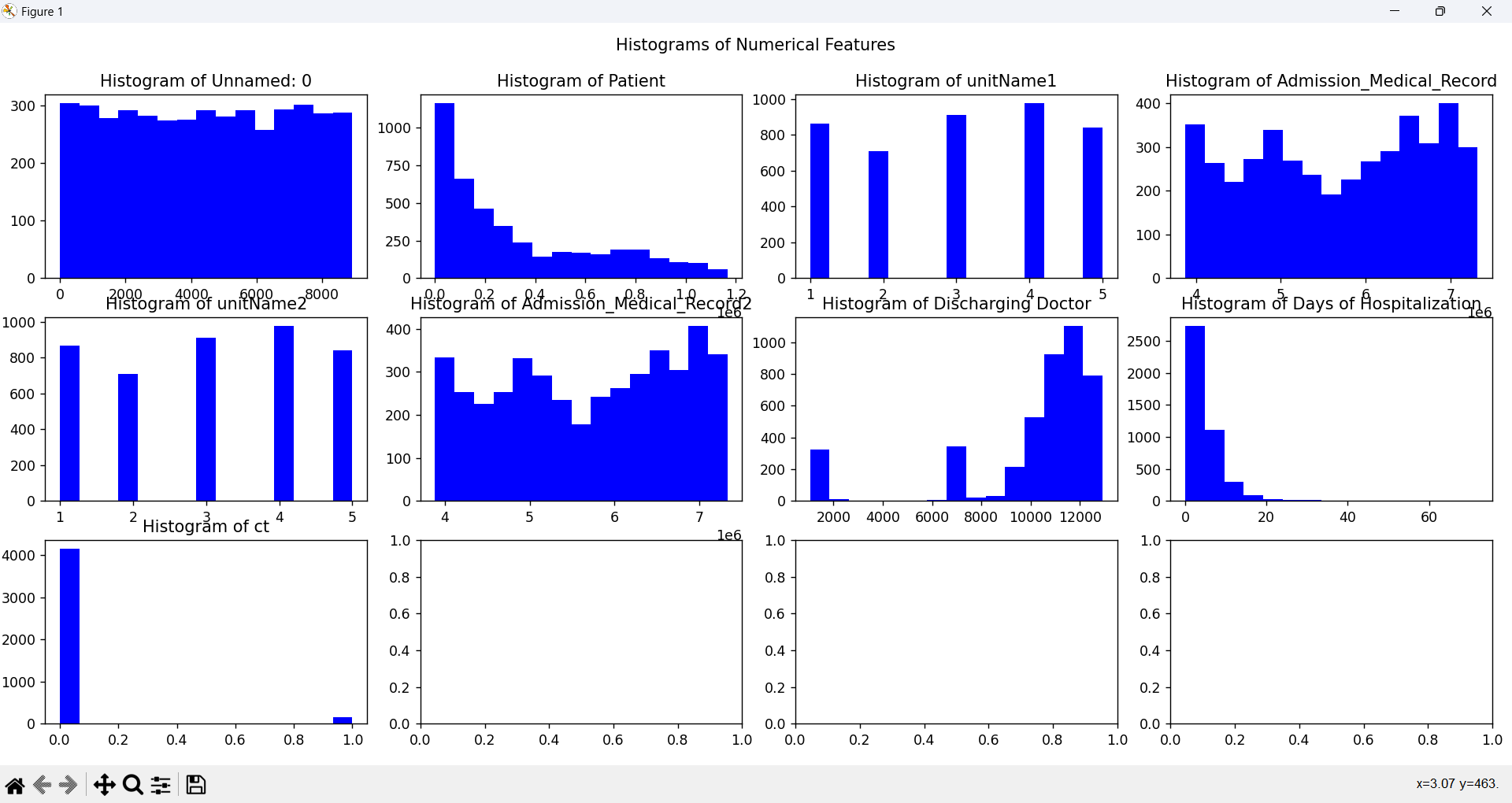
**Task 15: EDA of each parameter in the 'hospitalization2' table.**

The code 'Assignment 15.py' uses hospitalization2\_translated\_clean.csv'—an output from Task 7—as its input. The code generates the graphs shown below.

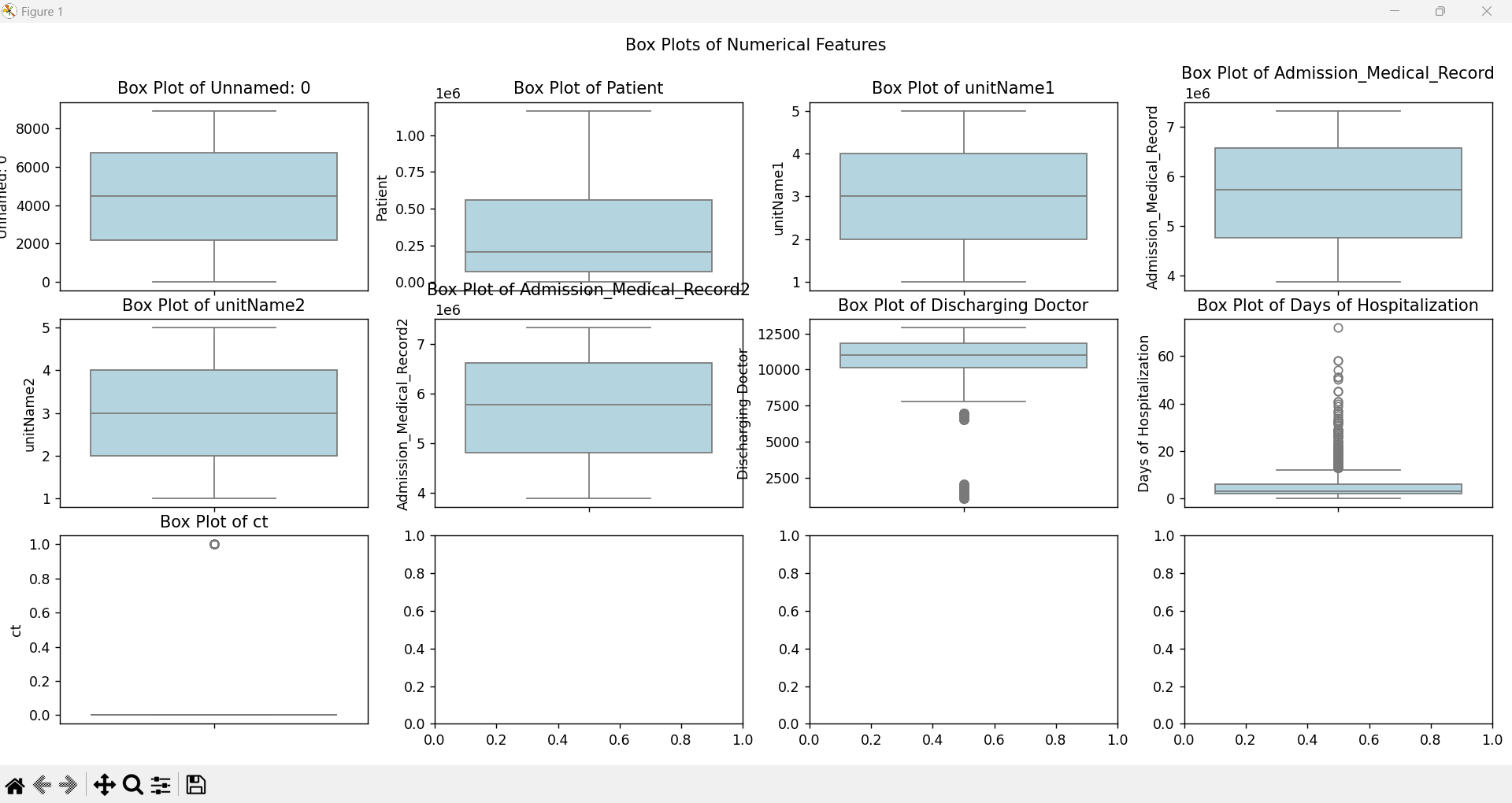
The dataset includes a mix of numerical, categorical, and date features:

1. **Numerical Features:** Features like 'Admission\_days2' may require normalization if they exhibit a wide range of values or are to be used in algorithms sensitive to feature scaling.
2. **Categorical Features and Dates:** These should be appropriately transformed:
   * **Categorical Features** (e.g., 'unitName1', 'Patient\_origin') should be encoded using techniques such as one-hot encoding or label encoding, depending on the algorithm requirements.
   * **Date Features** should be converted into a more analytically useful format. Options include converting dates to the number of days from a specific start date or decomposing dates into year, month, and day components, tailored to your analysis or modeling needs.
3. **Histograms for all numerical data**



The graphs generated from the dataset show the distribution of various features related to hospitalization records:

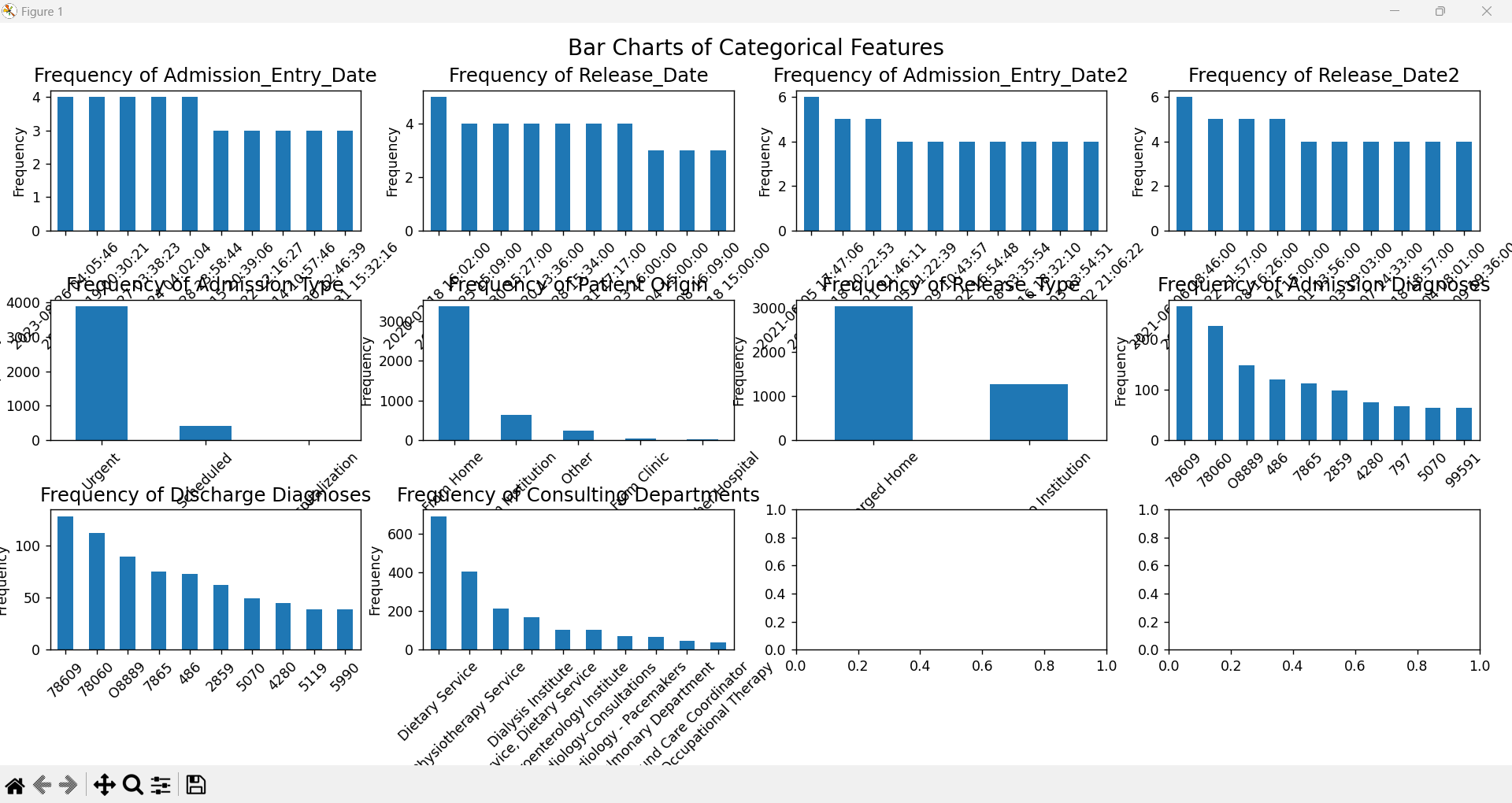
1. Unnamed: 0 - This histogram shows a nearly uniform distribution, suggesting that the data points are evenly spread across the range. This might indicate an index or identifier.
2. Patient - This histogram shows a right-skewed distribution, with a majority of data points concentrated towards the lower values. This could indicate that most patients have a lower value for this feature, possibly reflecting a frequency or count metric.
3. unitName1 - This histogram appears to have discrete values, indicating that this feature might represent categorical data that has been encoded as numbers, with a fairly even distribution across these categories.
4. Admission\_Medical\_Record - This shows a similar pattern to the first histogram with a uniform distribution, though it's not perfectly flat, indicating the values are spread across the range but with some slight variations in frequency.
5. unitName2 - Like unitName1, this feature appears to represent a categorical variable that is evenly distributed across several distinct categories.
6. Admission\_Medical\_Record2 - This histogram shows a roughly uniform distribution with slight peaks and troughs, suggesting that values are evenly spread across the range, though not perfectly.
7. Discharging Doctor - This feature seems to be heavily skewed with most data points concentrated towards the lower end, possibly indicating that most discharges are handled by a smaller number of doctors.
8. Days of Hospitalization - This histogram shows a significant skew towards lower values, suggesting that most patients have a short hospital stay, with very few staying for a long duration.
9. ct - The distribution here is highly skewed towards a single category, indicating that the majority of the data points have the same value for this feature. This could represent a binary or categorical feature where one category is overwhelmingly represented.
10. **Box plots for numeric columns**

****

The box plots offer a different perspective on the numerical features compared to the histograms, allowing us to see the spread, central tendency, and presence of outliers for each feature.

Here's an analysis of what each box plot might be indicating:

1. Unnamed: 0 - The box plot shows a relatively uniform distribution with no apparent outliers. The median is centered within the interquartile range (IQR), suggesting a balanced distribution of data points.
2. Patient - This feature has a wide range with the median slightly below the center of the IQR. There are no significant outliers, indicating the data is spread evenly with a slight skew towards the lower values.
3. unitName1 - This categorical feature, encoded as numerical values, shows that the data is fairly evenly distributed across the categories. The median is centered within the IQR, and there are no outliers.
4. Admission\_Medical\_Record - Similar to the "Unnamed: 0" feature, this plot shows a wide range of values with the median centrally located within the IQR, and no outliers are present.
5. unitName2 - This feature also represents categorical data, with the median centered and no outliers, indicating an even distribution across the categories.
6. Admission\_Medical\_Record2 - This box plot is similar to the previous ones, showing a centered median within the IQR and no outliers.
7. Discharging Doctor - This feature shows a significant number of outliers on the lower end, indicating that while most data points cluster around higher values, a few are much lower. This could indicate that while most doctors have a high identifier value, a few have much lower identifiers.
8. Days of Hospitalization - The box plot for this feature shows a large number of outliers, primarily on the higher end. This suggests that while most patients have a short hospital stay, a few have significantly longer stays. The median is low, further supporting the idea that the majority of patients are hospitalized for a short duration.
9. ct - This feature appears to be binary, with most values at 0 and only a few at 1, as indicated by the box plot. The lone outlier might be due to the sparse occurrence of the higher value.
10. **bar charts for categorical data**



1. Frequency of Admission\_Entry\_Date:

The data points appear evenly distributed across different dates with no obvious outliers, suggesting regular entries over time. This could indicate a steady flow of admissions.

1. Frequency of Release\_Date:

Similar to the Admission\_Entry\_Date, the release dates are also evenly spread, which aligns with a consistent patient flow through the hospital.

1. Frequency of Admission\_Type:

There is a clear dominance of 'Urgent' admissions over 'Scheduled' ones, indicating the hospital handles a significant number of emergency cases compared to planned admissions.

1. Frequency of Patient\_Origin:

Most patients are noted as coming from 'Home', with far fewer from other sources like 'Other' or 'Institution'. This might reflect on the primary catchment area or patient base of the hospital.

1. Frequency of Discharge Diagnoses:

This plot shows a variety of discharge diagnoses, with some being far more common than others, which might point to common conditions treated or specialized services offered by the hospital.

1. Frequency of Consulting Departments:

Certain departments like 'Dietary Service' appear more frequently, which might suggest a focus on nutritional support as part of patient care. Other services show less frequent involvement.

1. Frequency of Admission\_Entry\_Date2:

Appears similar to the first Admission\_Entry\_Date plot, likely another timestamp feature showing regular patient admissions.

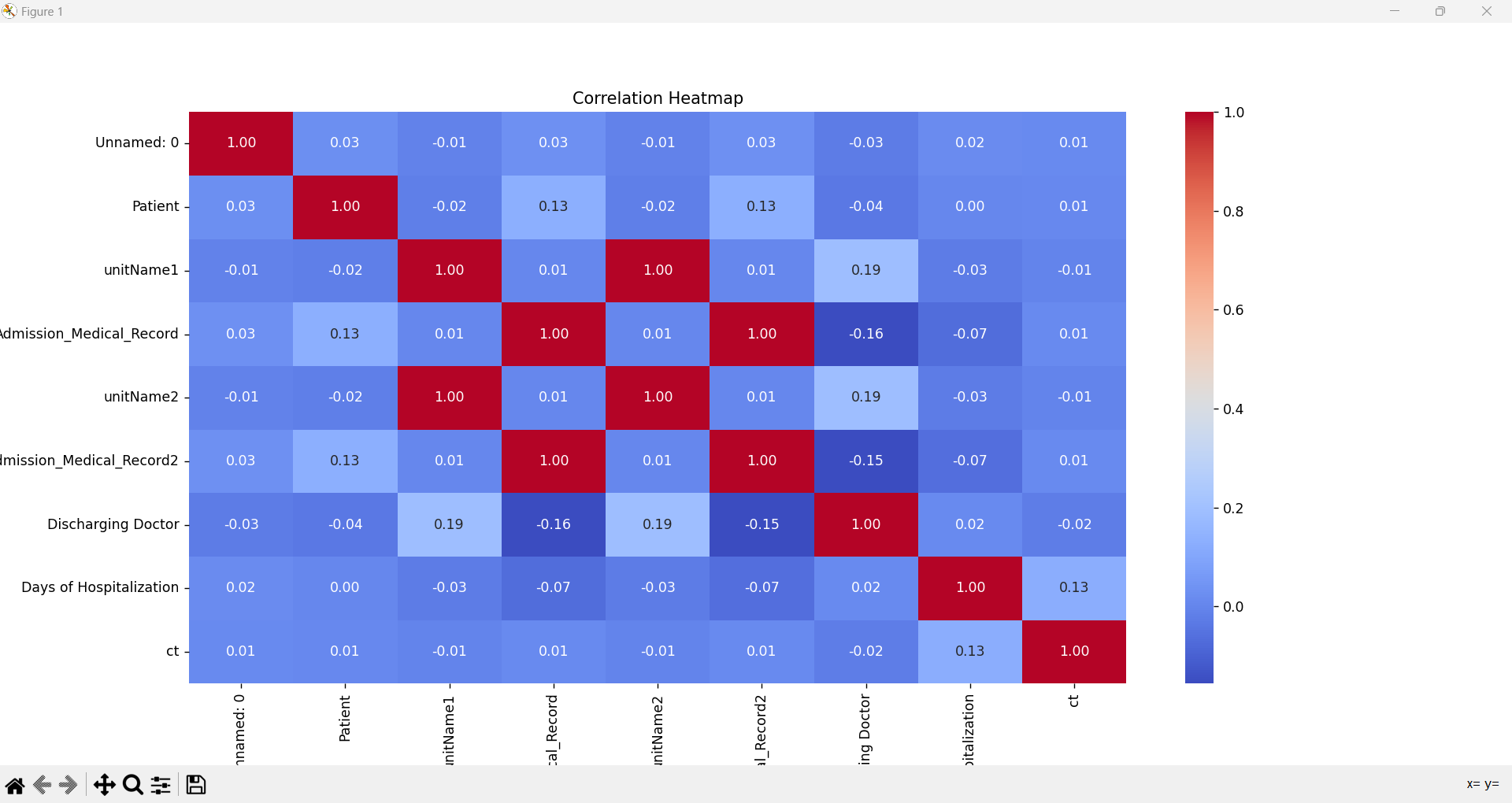
1. Frequency of Release\_Date2:

This mirrors the Release\_Date and shows a consistent pattern of patient releases, which might be used for assessing hospital stay durations or operational efficiency.

1. Frequency of Admission Diagnoses:

Shows a variety of admission reasons, with some being more prevalent. This distribution can help in understanding common ailments or conditions leading to hospitalization.

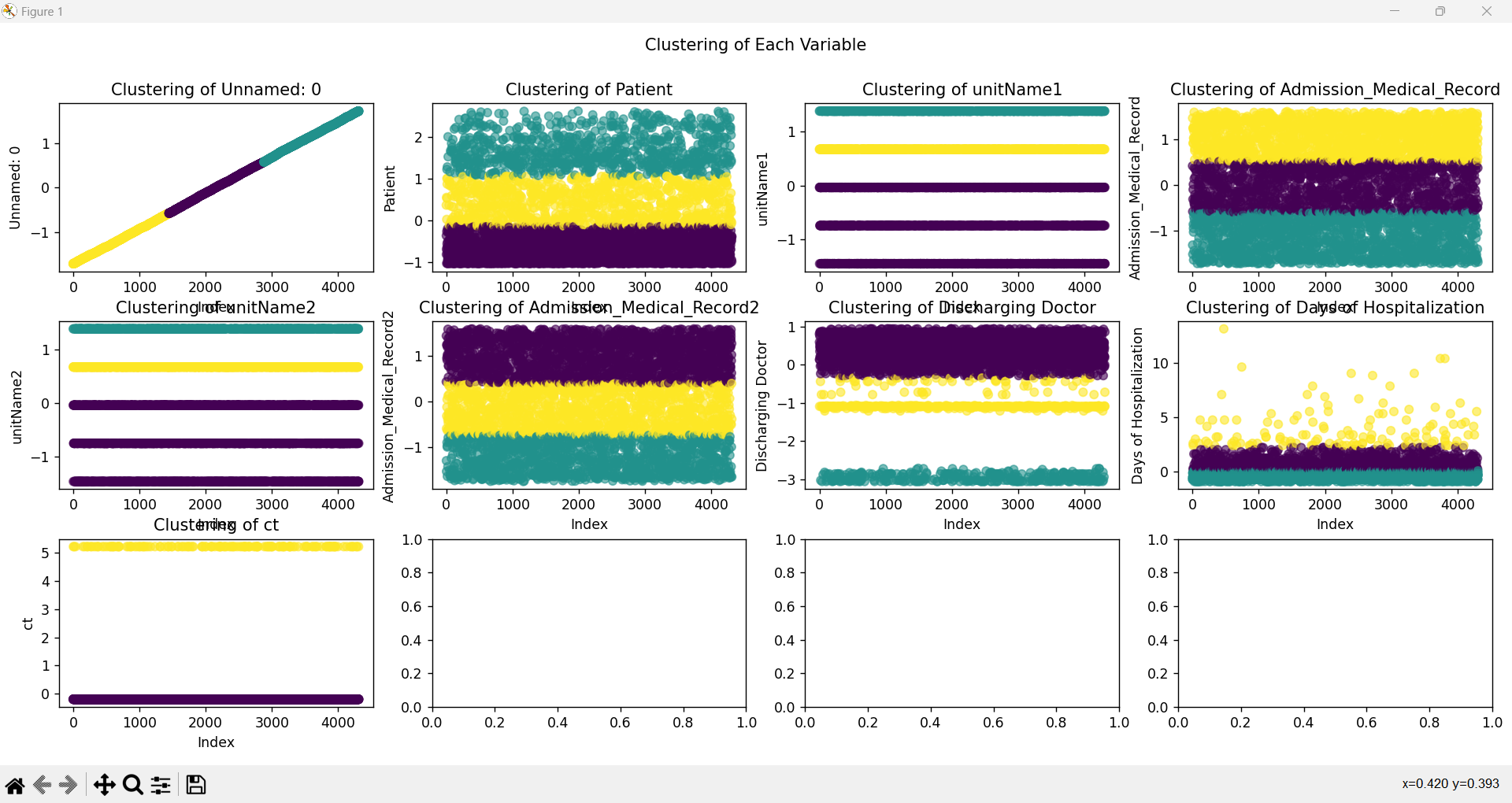
1. **Correlation heatmap for numerical data**



1. Unnamed: 0: This likely represents an index or an auto-generated identifier column. The correlation of this with any other variable being close to zero suggests it has no meaningful relationship with other data points, as expected for an index.
2. Patient: This could be a unique identifier for each patient. Any correlations involving this are minimal, which is typical for an ID column, indicating no linear dependency between the patient ID and other variables.
3. unitName1 and unitName2: These seem to be identifiers for different hospital units or departments. The positive correlation between them and specific medical records or doctors might suggest operational or administrative overlaps or specializations within the hospital.
4. Admission\_Medical\_Record and Admission\_Medical\_Record2: Both likely pertain to some form of patient admission documentation or codes. The fact that they correlate with each other suggests they might serve related administrative or clinical functions.
5. Discharging Doctor: This column’s correlations with unit names might indicate that certain doctors predominantly work in specific units, suggesting areas of specialization or departmental staffing structures.
6. Days of Hospitalization: This measures the length of hospital stays. Its correlation, albeit weak, with the feature 'ct' could hint at certain conditions or administrative practices influencing the length of a patient's stay.
7. ct: Without more context, it's hard to specify what 'ct' represents, but it could involve clinical treatment variables, counts, or categorizations relevant to hospital procedures. Its weak correlation with hospitalization days could suggest a slight influence on or consequence of the length of hospital stays.
8. There are cells not on the diagonal with a correlation coefficient of 1, it suggests that these variables are perfectly correlated with each other. This could imply several things:
   1. **Duplicate Data**: The variables might be duplicates of each other, possibly representing the same data under different column names. This is common when data is merged from different sources and column names are not standardized.
   2. **Derived Attributes**: One variable could be a derived attribute of the other, such as an encoded version of a categorical variable, or a scaled transformation where the relationship is linear and maintains the correlation.
   3. **Data Entry Error**: Sometimes, data entry errors or data collection methods can lead to different variables having the same values across all records inadvertently.

In practice, if two different variables exhibit a correlation of 1, it's often advisable to investigate why this is the case. If they convey the same information, like unitName1 and unitName2, one of them should be removed from the dataset to avoid issues with multicollinearity in any model training processes.

1. **Clustering**



This set of plots displays the clustering of each variable in the dataset, where each subplot represents a different variable and the points are colored based on their cluster assignment. Here's an observation for each plot:

1. Clustering of Unnamed: 0 - This appears to show a clear linear trend, possibly suggesting this variable might be an index or another type of sequential identifier.
2. Clustering of Patient - The data points show some overlapping cluster boundaries, suggesting variability in patient data that may require further normalization or a different number of clusters.
3. Clustering of unitName1 and unitName2 - These plots exhibit very rigid, non-overlapping clustering, which might indicate these variables have distinct, well-separated categories or values.
4. Clustering of Admission\_Medical\_Record and Admission\_Medical\_Record2 - These plots also show clearly defined clusters with distinct boundaries, suggesting that the values in these records can be segmented neatly into groups.
5. Clustering of Discharging Doctor - This plot shows several horizontal layers, indicating a categorical or ordinal variable with well-defined groups.
6. Clustering of Days of Hospitalization - The plot shows a spread of data across clusters with a focus on lower values, which could suggest that most hospitalizations are of shorter duration but there are outliers of longer stays.
7. Clustering of ct - Similar to some unit names, it shows clear segmentation into distinct groups, likely indicating distinct categories or states.

Each plot varies in the density and distribution of clusters, reflecting how data in each variable is structured and how well the K-means clustering algorithm can segment them into distinct groups. This visual overview can help you determine the appropriateness of the number of clusters chosen and whether additional preprocessing steps like scaling or transforming might be necessary for some variables to improve the clustering outcome.