DATA SCIENCE LAB 2

Final project

Team 8 - Niv Cohen

Contents

Task 6 - Data cleaning for table: erBeforeHospitalization2

The data was missing many values, where most of the null values were in the same row:

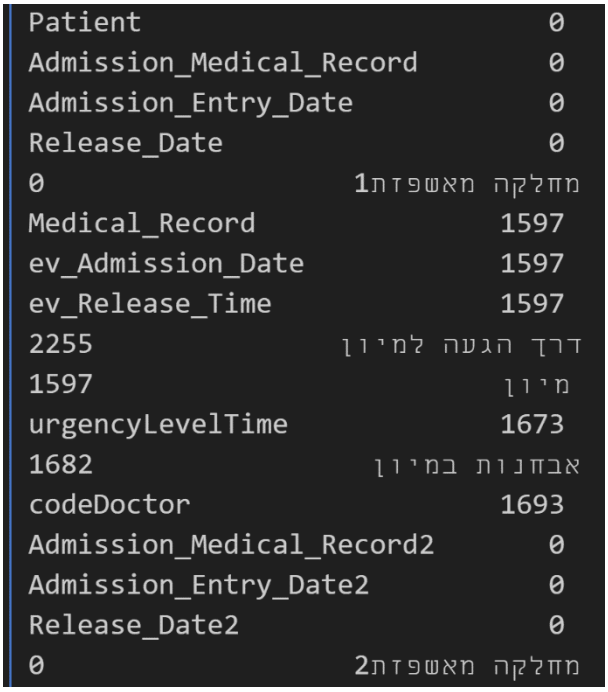

Figure 1 : null values in the erBeforeHospitalization2 dataset

For that reason, rows that had multiple null values were discarded. In order to fill the values for "urgencyLevelTime" and "דרך הגעה למיון", I observed the relation between Urgency vs Way of Arrival, Urgency vs ER department and ER department vs Way of Arrival, in order to fill the values in the most accurate way, because of the relation these features have.

Urgency vs Way of Arrival

	דרך הגעה למיון_אחר	דרך הגעה למיון_בן משפחה	דרך הגעה למיון_לבד	דרך הגעה למיון_נהג אמבולנס
3.000000	16	75	1444	283
2.000000	5	6	163	18
4.000000	54	304	3006	1056
5.000000	6	44	156	156
1.000000	0	0	45	1

Figure 2 : Urgency VS Way of arrival

Urgency vs ER department

	מיון_המחלקה לרפואה דחופה	מיון_מיון אורתופדי	מיון_מיון כירורגי	מיון_מיון מהלכים	מיון_מיון נשים	מיון_מיון עיניים	מיון_מיון פנימי	מיון_רפואה דחופה זיהומים
3.000000	1	2	40	8	0	0	1790	120
2.000000	0	0	2	0	0	0	189	17
4.000000	4	17	103	162	1	0	4009	523
5.000000	0	1	11	34	0	1	335	14
1.000000	0	0	0	0	0	0	49	0

Figure 3 : Urgency VS ER department

ER department vs Way of Arrival

	דרך הגעה למיון_אחר	דרך הגעה למיון_בן משפחה	דרך הגעה למיון_לבד	דרך הגעה למיון_נהג אמבולנס
מיון_המחלקה לרפואה דחופה	0	1	4	0
מיון_מיון אורתופדי	1	0	15	2
מיון_מיון כירורגי	1	5	90	42
מיון_מיון מהלכים	2	63	8	109
מיון_מיון נשים	0	0	0	0
מיון_מיון עיניים	0	0	0	2
מיון_מיון פנימי	67	337	4169	1304
מיון_רפואה דחופה זיהומים	10	23	531	57

Figure 4 : ER department VS Way of arrival

We can see from the previous tables that most of the ways of arrival to the hospital and most of the ER departments relate mainly to urgency level 4, and some of them to level 3. In addition, we can see that most of the ways of arrival relates to "מיון פנימי", Arriving alone relates also to "רפואה דחופה וזיהומים" and arriving with family member or with ambulance relates as well to "מיון מהלכים".

Thus, I replaced every column where there is no urgency level with urgency level 4. In addition, every column without ER department replaced with "מיון פנימי".

We can see from the first table that most of the patients arrived alone or in ambulance or with a family member, thus, I put a random choice for the computer to decide the way of arrival.

The final shape of the dataset after cleaning null values and columns that were not usable such as 'Admission_Medical_Record' as 'Medical_Record' is:
(7415, 15)

Task 15 – EDA of Hospitalization2 sheet

The EDA made on the cleaned hospitalization sheet by team 10.

Types in the dataset:

unitName1	int64
Admission_Entry_Date	object
Release_Date	object
unitName2	int64
Admission_Entry_Date2	object
Release_Date2	object
Entry_Type	object
Patient_Origin	object
Release_Type	object
Releasing_Doctor	int64
Admission_Days2	int64
Diagnosis_In_Reception	object
Diagnosis_In_Release	object
ct	int64
Admission_Days	int64
Period_Between_Admissions	object

Figure 5: Data types in hospitalization2 sheet

Distribution of rehospitalizations over the different hospital units (1-5):

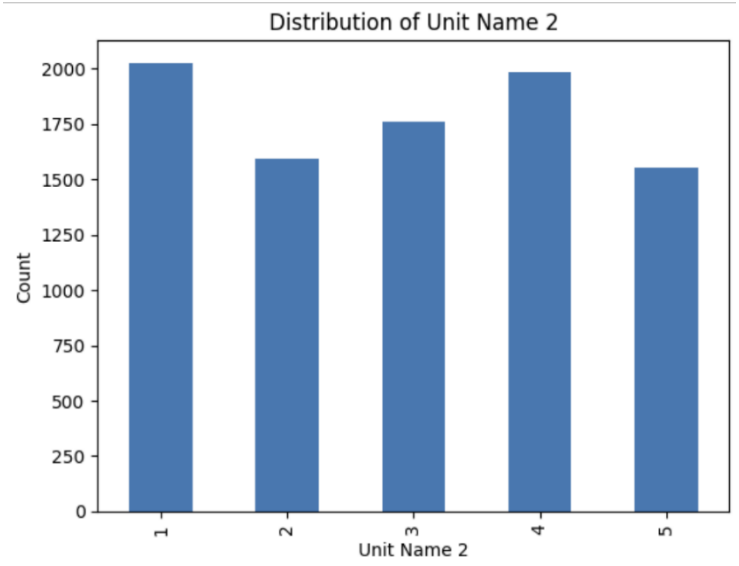

Figure 6: Distribution of rehospitalization in each unit

It is evident that hospital unit 1 experienced a higher number of rehospitalizations compared to the other units. Following this unit, units 4 and 3 show the next highest counts of rehospitalizations.

Most frequent hospitalization duration in the first and second hospitalization:

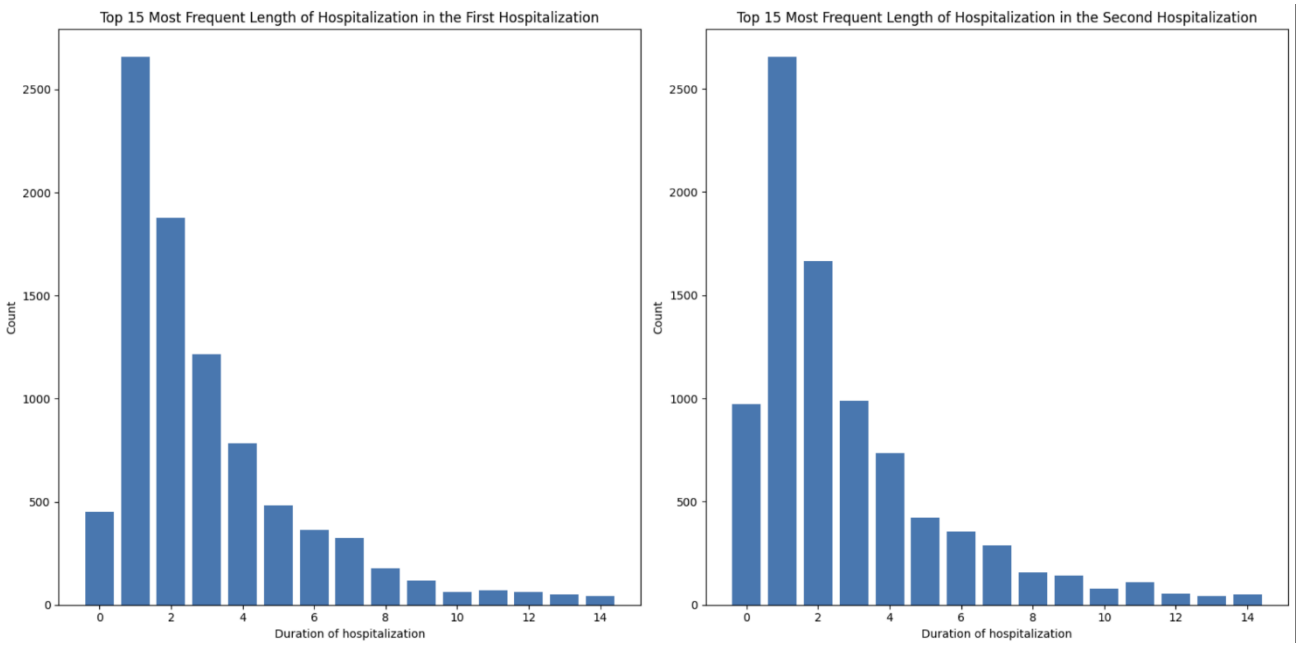

Figure 7: Frequent length of hospitalizations

In these two graphs we can see the duration of the hospitalization in each time. We can see that the duration of the second hospitalization generally is shorter than the first hospitalization.

Distribution of hospitalization in different time frames:

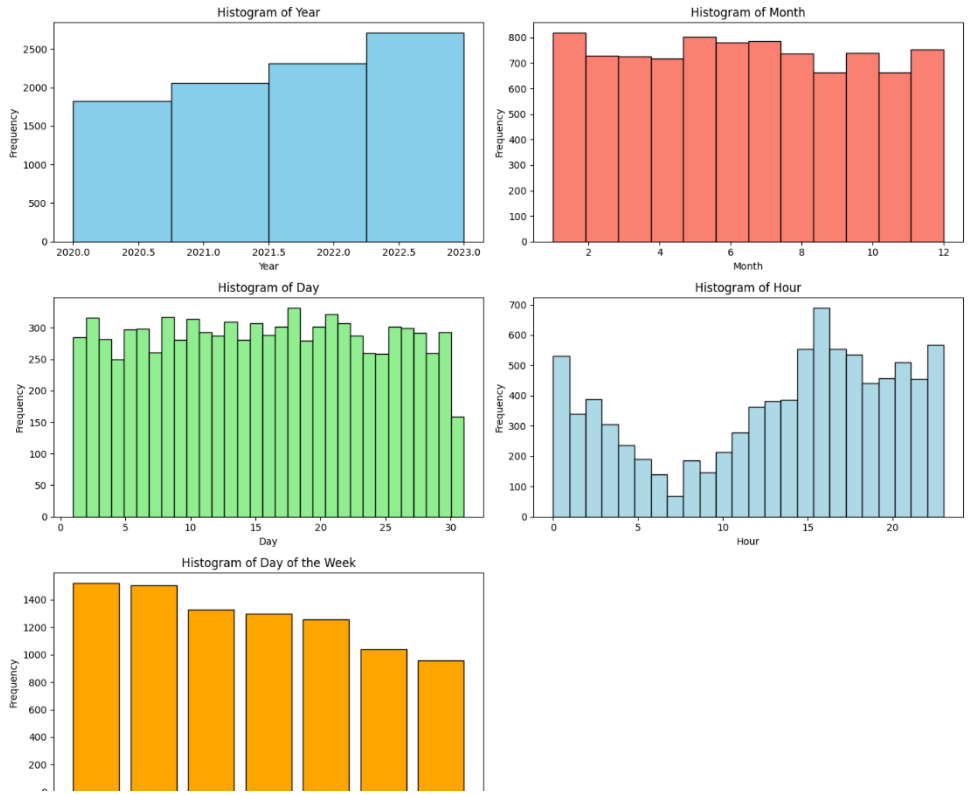

Figure 8: Distribution of rehospitalization in different time periods

The graphs reveal that rehospitalizations are fairly evenly distributed across each month and day of the month. However, there is a noticeable increase in rehospitalizations over the years. Additionally, rehospitalizations tend to decrease as the week progresses, with the majority occurring at the beginning of the week, particularly after 15:00 and extending into the evening.

Most frequent diagnosis in rehospitalization

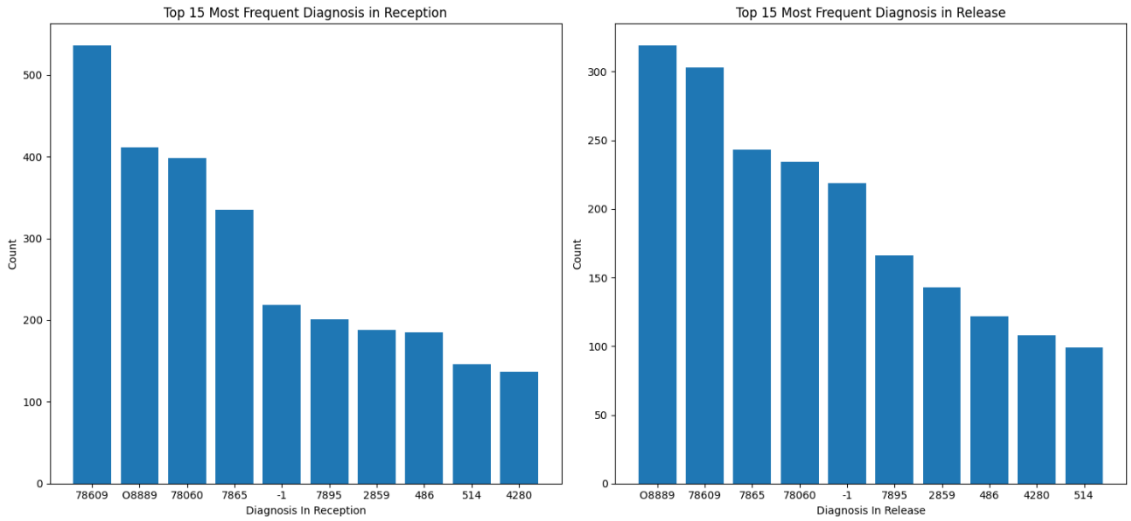

Figure 9: Frequent diagnosis in reception and release

In the data collected, the three most frequent diagnoses upon reception were: 78609 (Other Dyspnea and Respiratory Abnormality), O8889 (Diagnosis with No Code), and 78060 (Fever). Upon release, the top three diagnoses were: O8889 (Diagnosis with No Code), 78609 (Dyspnea), and 7865 (Non-Cardiac Chest Pain).

These findings indicate that the most common diagnoses were related to shortness of breath (Dyspnea), fever, and chest pain. Notably, the data spans a period largely influenced by the COVID-19 pandemic, which began in 2019 and continued until the end of 2022. This context may explain the high occurrence of diagnoses related to Dyspnea and Fever, as these symptoms are commonly associated with COVID-19.

Task 19 – Indication for same diagnosis in first and second hospitalization

In the sheet “hospitalization2” the count of matches and mismatches are:

Matches: 1311

Mismatches: 7415

The matches are only 15% of all cases of rehospitalization. While the number of matches is only 15%, it is non-negligible part of the patients that were rehospitalized.

Task 31– Time analysis by dates where there were more rehospitalizations

As we saw in fig. 8 , the rehospitalization rates appear to be fairly uniform across the months of the year and throughout the days of each month. However, significant variations are observed in relation to the "Year," "Hour," and "Week." Notably, the number of rehospitalizations shows an approximately linear increase year over year. Therefore, my focus will be on examining the relationship between the day of the week and the time of day. This analysis will consider that re-hospitalization rates are distributed relatively evenly across months and days, while the overall frequency of re-hospitalizations is expected to rise each year.

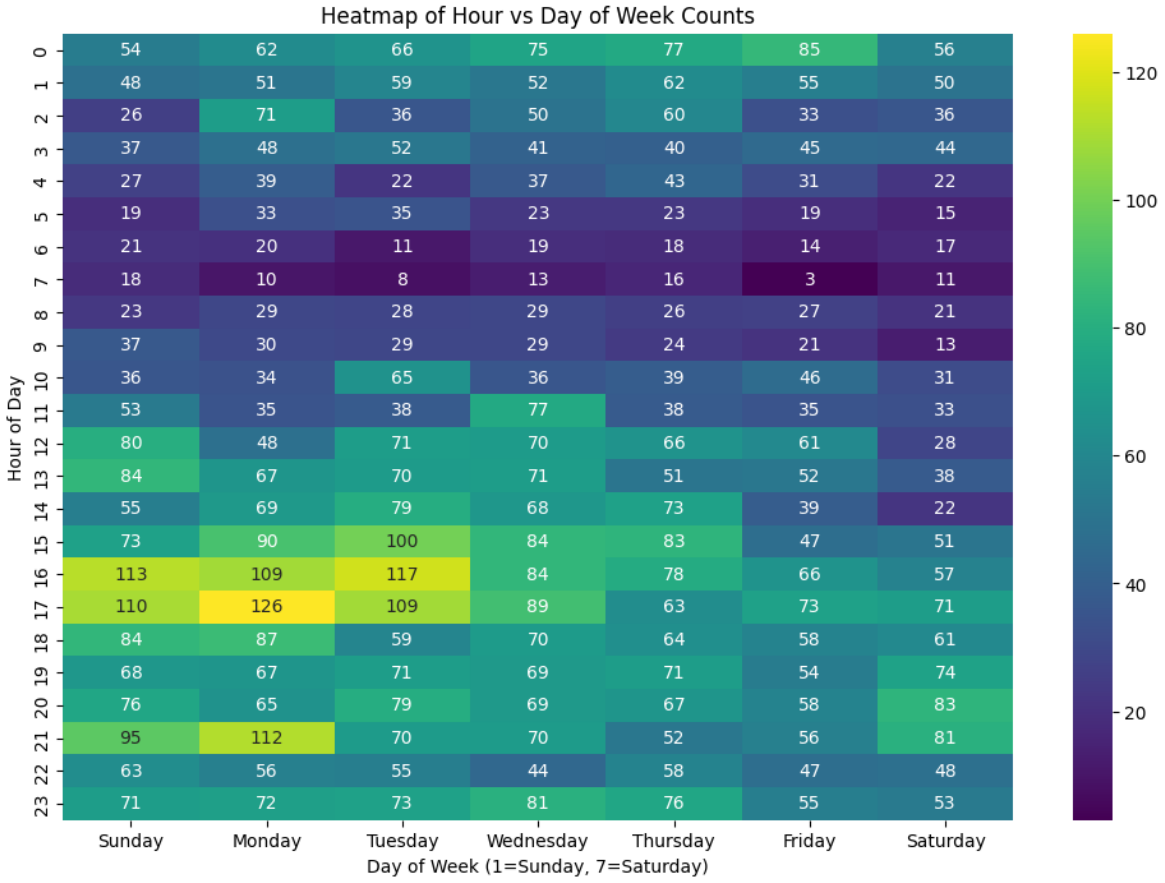

Figure 10: Heatmap of Hour VS Day of week counts

We can see some distribution of the two features together (Day of week and hour of the day) which is concentrated around the beginning of the week at around 17:00. I would like to find a distribution that will fit the data in order to predict the relative volume of rehospitalizations that can occur on each day and in what time.

The first obvious distribution is a multivariate Gaussian distribution:

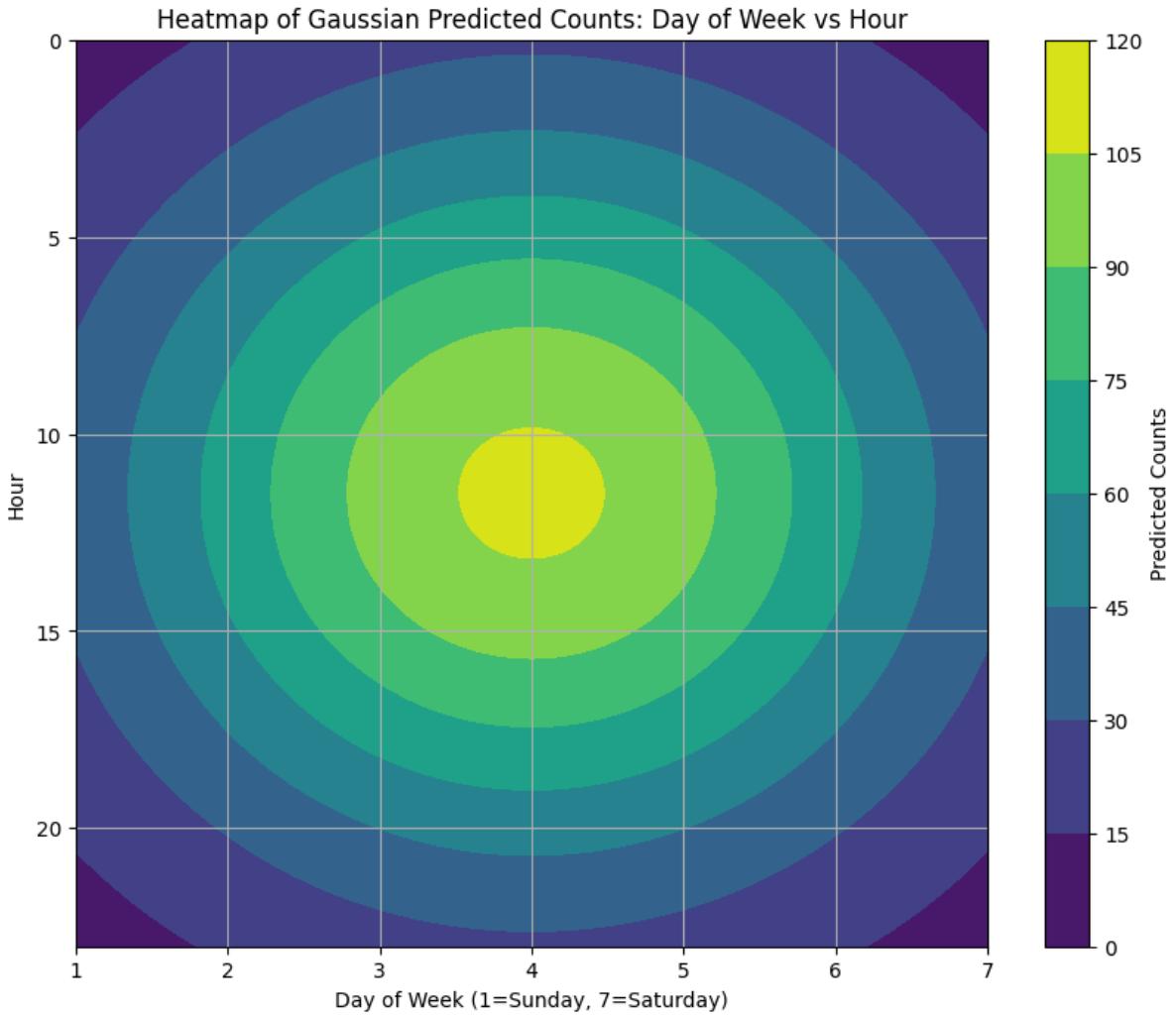

Figure 11: Heatmap of Gaussiam distribution prediction count: Hour VS Day of the week

The Gaussian distribution fit gave the following results:

Mean: [4.0, 11.5]

Covariance Matrix: $\begin{pmatrix} 4.023 & 0 \\ 0 & 48.203 \end{pmatrix}$

Mean Squared Error (MSE): 3429.58

Root Mean Squared Error (RMSE): 58.56

It is possible to see that the Gaussian distribution did not fit well as expected to the right place as expected from observing the heap map of the actual data.

Thus, another distribution was used to fit the data, the Poisson distribution:

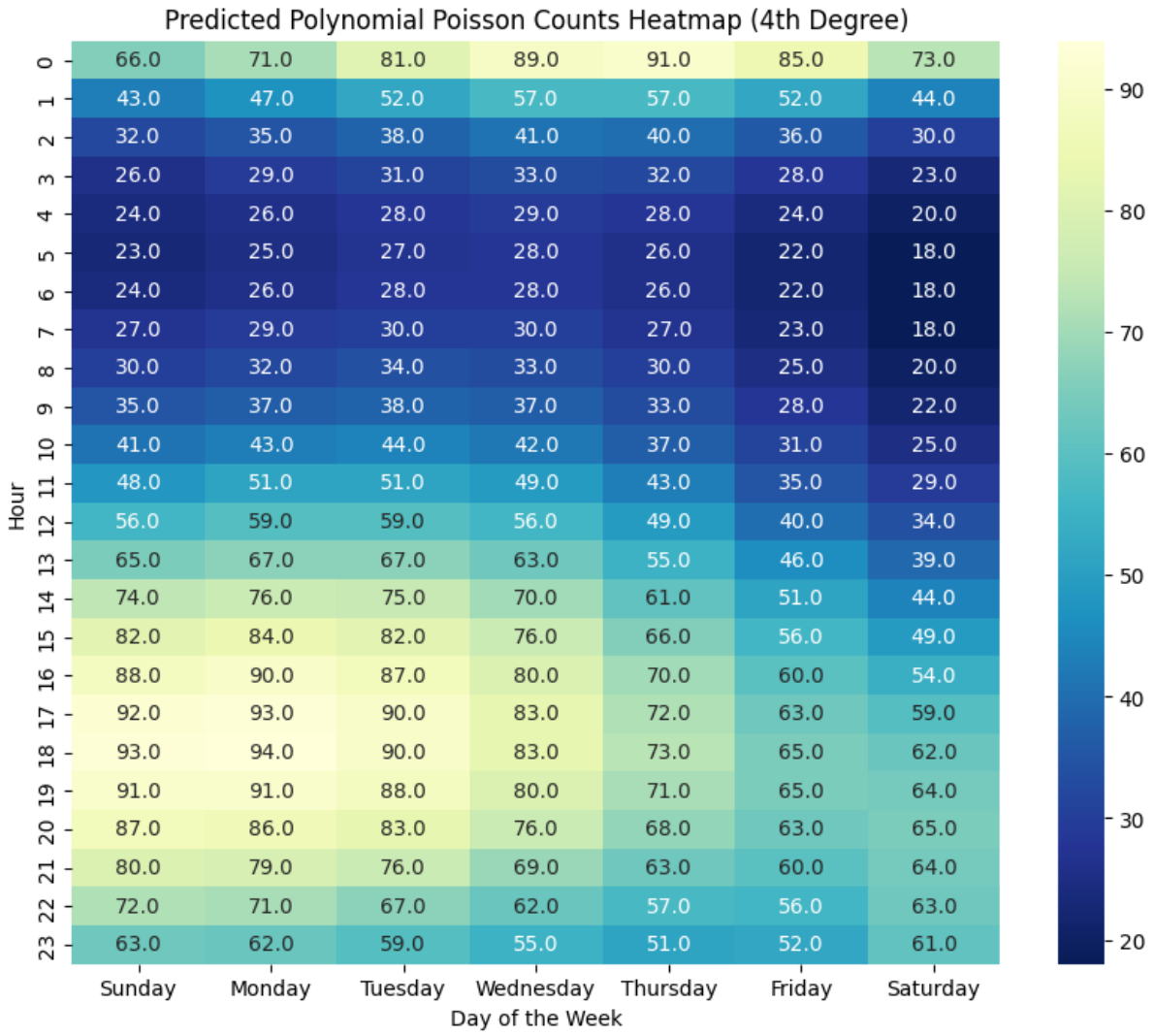

Figure 12: Heatmap of predicted polynomial Poisson count of Hour VS Day of the week

Poisson measure of fit:

Mean Squared Error: 166.0

Root Mean Squared Error: 12.9

We can see that the poisson distribution fitted the data much more accurately and can predict the relative volume of re-admission patients over each day and time of the day.

Conclusions from task 31

- The rehospitalization rates appear to be fairly uniform across the months of the year and throughout the days of each month.
- The number of rehospitalizations shows an approximately linear increase year over year.

- The "Hour" and "Day of Week" variables exhibit notable variations throughout their respective intervals.
- The Poisson model proved to be reasonably reliable in predicting the relative volume of rehospitalized patients across different days of the week and times of the day.

Task 33– Time analysis of each hospital unit in the rehospitalization

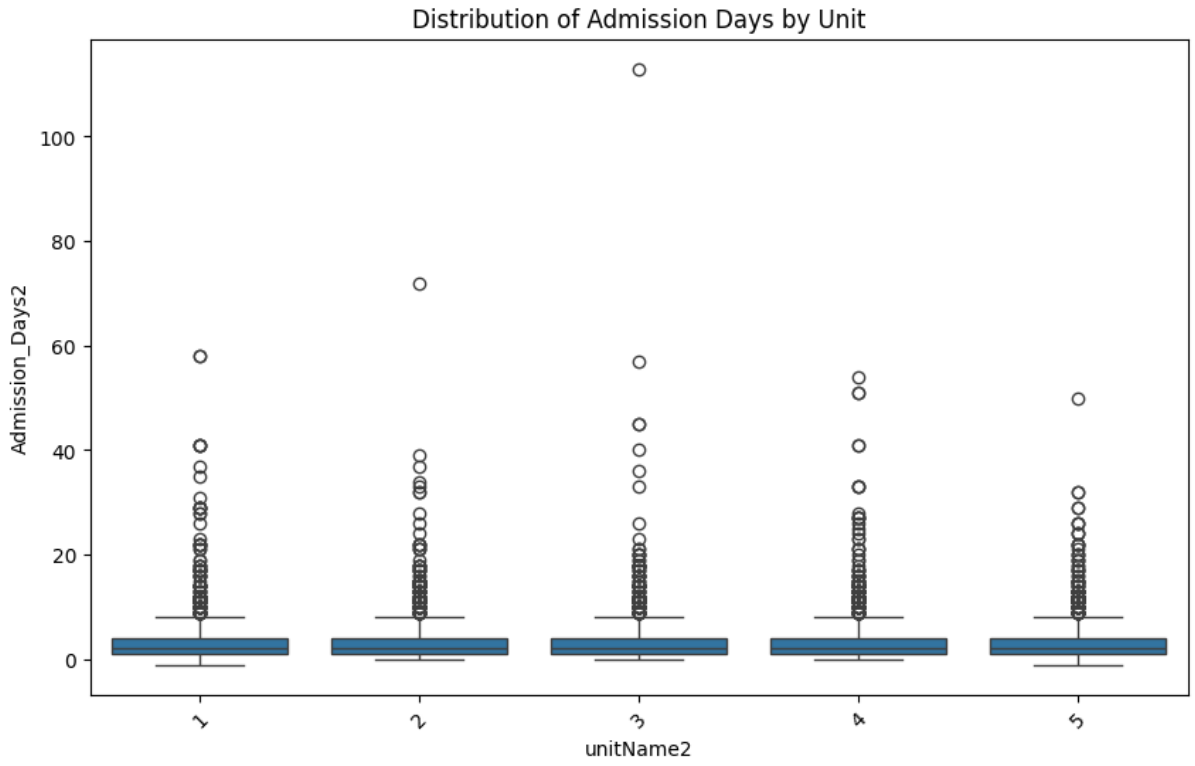

Figure 13: Distribution of duration of hospitalization by unit

We can see from fig.13 That the duration of the second hospitalization is about the same for each hospital unit.

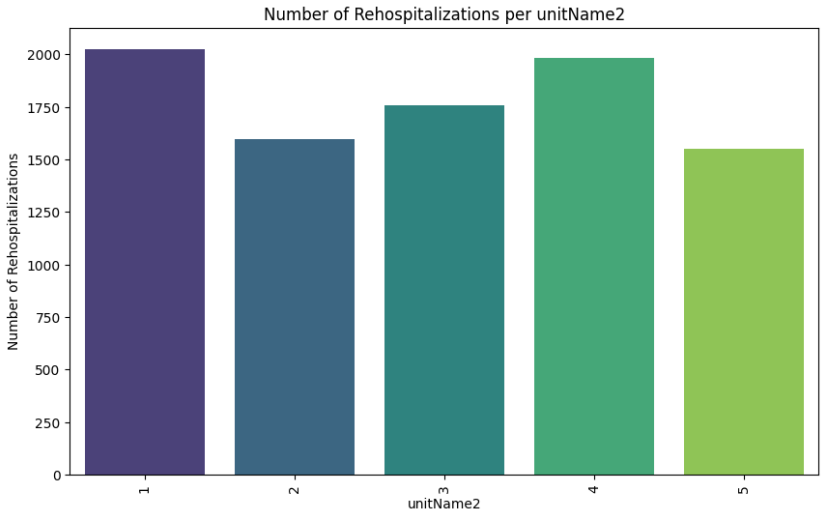

Figure 14: Number of rehospitalization per unit

In fig. 14 we can see from the graph that units 1 and 4 had the most rehospitalizations, while the other units are not far behind.

	Feature	Test	Statistic	p-value
0	unitName1	ANOVA	inf	0.000000e+00
1	Admission_Entry_Date	Chi-Square	3.566000e+04	2.099399e-189
2	Release_Date	Chi-Square	3.529397e+04	9.279979e-195
3	Admission_Entry_Date2	Chi-Square	3.509740e+04	5.206423e-93
4	Release_Date2	Chi-Square	3.479540e+04	3.313554e-95
5	Entry_Type	Chi-Square	3.309282e+01	5.926105e-05
6	Patient_Origin	Chi-Square	6.096996e+01	3.589562e-07
7	Release_Type	Chi-Square	1.185976e+01	1.842514e-02
8	Releasing_Doctor	ANOVA	1.795715e+02	3.291472e-148
9	Admission_Days2	ANOVA	2.738152e+00	2.716491e-02
10	Diagnosis_In_Reception	Chi-Square	9.875884e+03	4.606143e-73
11	Diagnosis_In_Release	Chi-Square	1.593389e+04	4.345298e-89
12	ct	ANOVA	3.748142e+00	4.739432e-03
13	Admission_Days	ANOVA	2.539185e+01	6.012354e-21
14	Period_Between_Admissions	Chi-Square	1.779567e+01	2.281165e-02

Figure 15: Statistical test for the impact of each feature on the hospital unit

It seems that all the parameters have a great impact relating to the unit that received the patient in the rehospitalization observing the p-value.

Let us observe the seasonality of the rehospitalization for each unit:

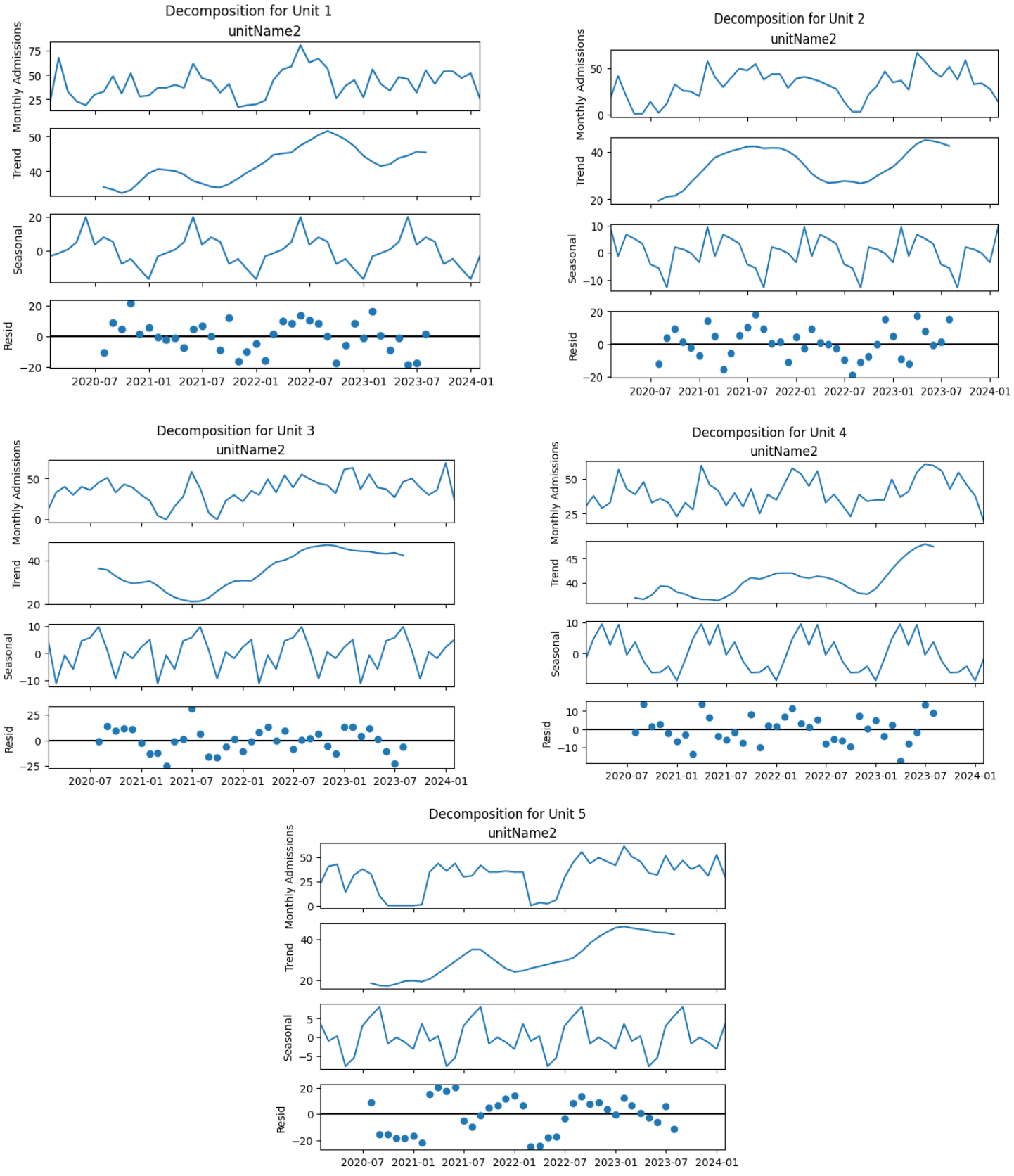

Figure 16: Seasonalty and trends of each hospital unit

In the last five graphs, we observe the seasonality patterns of the five different hospital units over a period of almost four years. The Trend graphs reveal that the number of admissions increases year over year in each unit. Although this increase is not constant throughout the period, the overall trend is upward.

In the Seasonality graphs, we see that each unit experiences different periods of heightened rehospitalizations:

- Unit 1 shows a significant peak in rehospitalizations around May and June.
- Unit 2 experiences most rehospitalizations approximately between September and May.
- Unit 3 has peaks in July, with a smaller peak in January.
- Unit 4 sees a sustained period of high rehospitalizations from March to August.
- Unit 5 has peaks in February and August.

Each unit displays its own unique seasonality, with varying peak times throughout the year.

Task 43 – Literature review

Rehospitalization is a significant concern in healthcare systems worldwide, reflecting not only the recurring needs of patients but also the efficiency of healthcare delivery and follow-up care. Defined as the unplanned return of a patient to a hospital within a specified period after discharge, rehospitalization rates can indicate various underlying issues, including the severity of initial conditions, the quality of discharge planning, and adherence to post-discharge care. The high rates of rehospitalization have attracted considerable focus in the academic literature, with studies revealing a range of predictors that influence the likelihood of re-admission. This literature review aims to synthesize the current research on rehospitalization, focusing on key predictors, theoretical frameworks, and existing gaps in the field.

The study by Elizabeth Mayfield Arnold, *Rates and Predictors of Rehospitalization Among Formerly Hospitalized Adolescents*, provides significant insights into factors that influence adolescent psychiatric rehospitalizations. The research followed 180 adolescents for over a decade after their discharge from an inpatient psychiatric unit, examining various demographic and psychiatric predictors. The study found that 44% of the adolescents experienced at least one rehospitalization during the follow-up period, with 19% re-hospitalized within six months. In this study, gender, race, age, psychiatric variables, including diagnoses, prehospitalization suicide attempts, and previous hospitalizations, were examined as potential predictors of rehospitalization. Key predictors identified were younger age and the presence of an affective disorder, emphasizing the importance of closely assessing these factors before discharge. The study's findings underline the need for targeted interventions for younger patients and those with depressive disorders to reduce the risk of rehospitalization.

The study by John Billings, *Development of a Predictive Model to Identify Inpatients at Risk of Re-admission Within 30 Days of Discharge (PARR-30)*, presents a predictive logistic regression model designed to identify patients at high risk of rehospitalization within 30 days of discharge in the National Health Service (NHS) hospitals in England. Using multivariate statistical analysis and logistic regression, the model was developed from a sample of hospital episode statistics (HES) data covering over half a million admissions from 2008 to 2009. The predictors used in the study are: number admissions to hospital by type (emergency versus non-emergency) according to a time interval prior to current admission (90, 180, 365, 730 and 1095 days); the number of episodes per spell in prior admissions (a proxy measure of complex health problems); number of different types of specialists consulted in the last 12 months; a range of diagnostic categories and hierarchical diagnostic groups;36 characteristics of the area of residence and length of stay. The result of this study is a model that outputs a risk score for each patient, with a threshold score of 0.5 yielding a prediction accuracy of 59.2%.

The study by Alessandro Morandi, *Predictors of Rehospitalization among Elderly Patients Admitted to a Rehabilitation Hospital*, the authors examined factors contributing to rehospitalization among elderly patients above the age 65. The study specifically explored the impact of polypharmacy, functional status, and length of stay in rehabilitation hospitals as potential predictors of rehospitalization. The researchers found that patients with multiple medications (polypharmacy) and lower functional status at discharge were at higher risk of rehospitalization. Additionally, a longer stay in the rehabilitation hospital was associated with a increased likelihood of rehospitalization. The findings suggest that monitoring these factors can help healthcare providers identify elderly patients at higher risk and implement interventions to reduce their chances of readmission. This study highlights the importance of personalized care and careful management of medication and rehabilitation outcomes in preventing rehospitalization.

In conclusion, various factors contribute to the likelihood of rehospitalization. The most significant predictors, as highlighted by multiple studies, include age, psychological disorders, the number of medications a patient consumes, functional status (particularly in elderly patients), and the length of the initial hospitalization.