# 🏥 ReadmitGuard: Healthcare Patient Readmission Analysis Guide

## 🧩 Section 1. 📊 Data Exploration in Google Sheets

### a. Descriptive Stats

* Load dataset into Google Sheets
* Use:
* =AVERAGE(), =MEDIAN(), =MIN(), =MAX(), =STDEVP()
* Apply to each numerical column

### 🧾 Summary Metrics

| Metric | Lab Procedures | Procedures | Medications | Outpatient | Emergency | Inpatient | Hospital Stay |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Average** | 43.10 | 1.34 | 16.02 | 0.37 | 0.20 | 0.64 | 4.40 |
| **Min** | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| **Max** | 132 | 6 | 81 | 42 | 76 | 21 | 14 |
| **Median** | 44 | 1 | 15 | 0 | 0 | 0 | 4 |
| **Std. Dev.** | 19.67 | 1.71 | 8.13 | 1.27 | 0.93 | 1.26 | 2.99 |

### b. Re-admissions by Gender and Age (Pivot Table)

* Create a Pivot Table:
  + **Rows:** gender, age
  + **Columns:** readmitted
  + **Values:** COUNTA(encounter\_id)

| gender | age | <30 | >30 | NO | Grand Total |
| --- | --- | --- | --- | --- | --- |
| Female | [0-10) | 1 | 13 | 69 | 83 |
|  | [10-20) | 24 | 147 | 231 | 402 |
|  | [20-30) | 177 | 346 | 591 | 1114 |
|  | [30-40) | 242 | 647 | 1273 | 2162 |
|  | [40-50) | 511 | 1685 | 2615 | 4811 |
|  | [50-60) | 851 | 3105 | 4616 | 8572 |
|  | [60-70) | 1206 | 4012 | 5843 | 11061 |
|  | [70-80) | 1635 | 5108 | 7242 | 13985 |
|  | [80-90) | 1282 | 3855 | 5378 | 10515 |
|  | [90-100) | 223 | 600 | 1180 | 2003 |
| **Female Total** |  | 6152 | 19518 | 29038 | 54708 |
| Male | [0-10) | 2 | 13 | 63 | 78 |
|  | [10-20) | 16 | 77 | 196 | 289 |
|  | [20-30) | 59 | 164 | 320 | 543 |
|  | [30-40) | 182 | 540 | 891 | 1613 |
|  | [40-50) | 516 | 1593 | 2765 | 4874 |
|  | [50-60) | 817 | 2812 | 5055 | 8684 |
|  | [60-70) | 1296 | 3885 | 6240 | 11421 |
|  | [70-80) | 1434 | 4367 | 6280 | 12081 |
|  | [80-90) | 796 | 2368 | 3518 | 6682 |
|  | [90-100) | 87 | 208 | 495 | 790 |
| **Male Total** |  | 5205 | 16027 | 25823 | 47055 |
| Unknown/Invalid | [60-70) |  |  | 1 | 1 |
|  | [70-80) |  |  | 2 | 2 |
| **Unknown/Invalid Total** |  |  |  | 3 | 3 |
| **Grand Total** |  | 11357 | 35545 | 54864 | 101766 |

### c. Race Distribution Chart

* Steps:
  + Use =UNIQUE(race\_range) for distinct categories
  + Use =COUNTIF(race\_range, value) for frequency
  + Replace ? with "Unknown"

#### Race Frequency Table

| Race | Frequency |
| --- | --- |
| Caucasian | 76099 |
| AfricanAmerican | 19210 |
| Unknown | 2273 |
| Hispanic | 2037 |
| Other | 1506 |
| Asian | 641 |

### d. Re-admission Status Distribution

| Status | Count |
| --- | --- |
| NO | 54,864 |
| >30 | 35,545 |
| <30 | 11,357 |

* Visualize using a bar chart
* Formulas:
* =COUNTIFS(ReadmittedRange, "\*<30\*")  
  =COUNTIFS(ReadmittedRange, "\*>30\*")

## 🗄️ Section 2. Data Loading & SQL Tasks (MySQL Workbench)

### a. Create Database and Table

-- Create the database if it doesn't exist  
CREATE DATABASE IF NOT EXISTS healthcare;  
USE healthcare;  
  
-- Create the main table for diabetic patient records  
CREATE TABLE diabetic\_data (  
 encounter\_id BIGINT PRIMARY KEY,  
 patient\_nbr BIGINT,  
 race VARCHAR(50),  
 gender VARCHAR(15),  
 age VARCHAR(20),  
 weight VARCHAR(10),  
 admission\_type\_id INT,  
 discharge\_disposition\_id INT,  
 admission\_source\_id INT,  
 time\_in\_hospital INT,  
 payer\_code VARCHAR(20),  
 medical\_specialty VARCHAR(100),  
 num\_lab\_procedures INT,  
 num\_procedures INT,  
 num\_medications INT,  
 number\_outpatient INT,  
 number\_emergency INT,  
 number\_inpatient INT,  
 diag\_1 VARCHAR(10),  
 diag\_2 VARCHAR(10),  
 diag\_3 VARCHAR(10),  
 number\_diagnoses INT,  
 max\_glu\_serum VARCHAR(20),  
 A1Cresult VARCHAR(20),  
 metformin VARCHAR(20),  
 repaglinide VARCHAR(20),  
 nateglinide VARCHAR(20),  
 chlorpropamide VARCHAR(20),  
 glimepiride VARCHAR(20),  
 acetohexamide VARCHAR(20),  
 glipizide VARCHAR(20),  
 glyburide VARCHAR(20),  
 tolbutamide VARCHAR(20),  
 pioglitazone VARCHAR(20),  
 rosiglitazone VARCHAR(20),  
 acarbose VARCHAR(20),  
 miglitol VARCHAR(20),  
 troglitazone VARCHAR(20),  
 tolazamide VARCHAR(20),  
 examide VARCHAR(20),  
 citoglipton VARCHAR(20),  
 insulin VARCHAR(20),  
 glyburide\_metformin VARCHAR(20),  
 glipizide\_metformin VARCHAR(20),  
 glimepiride\_pioglitazone VARCHAR(20),  
 metformin\_rosiglitazone VARCHAR(20),  
 metformin\_pioglitazone VARCHAR(20),  
 `change` VARCHAR(10),  
 diabetesMed VARCHAR(10),  
 readmitted VARCHAR(20)  
);

### b. Data Cleaning

UPDATE diabetic\_data  
SET race = 'Unknown'  
WHERE race = '?';  
  
SELECT race, COUNT(\*) FROM diabetic\_data GROUP BY race;

### c. Dataset Overview

-- Encounters  
SELECT COUNT(\*) FROM diabetic\_data;  
  
-- Age-wise distribution  
SELECT age, COUNT(\*) FROM diabetic\_data GROUP BY age ORDER BY age;

### d. Readmission Analysis

-- Total readmitted  
SELECT COUNT(\*) FROM diabetic\_data WHERE readmitted IN ('<30', '>30');  
  
-- Readmission percentage  
SELECT ROUND(  
 (SELECT COUNT(\*) FROM diabetic\_data WHERE readmitted IN ('<30', '>30')) \* 100.0 /  
 (SELECT COUNT(\*) FROM diabetic\_data),  
 2  
);

### e. Readmission by Payer Code

SELECT payer\_code,  
 COUNT(\*) AS total\_cases,  
 SUM(CASE WHEN readmitted IN ('<30', '>30') THEN 1 ELSE 0 END) AS readmitted\_cases,  
 ROUND(SUM(CASE WHEN readmitted IN ('<30', '>30') THEN 1 ELSE 0 END) \* 100.0 / COUNT(\*), 2) AS rate\_pct  
FROM diabetic\_data  
GROUP BY payer\_code  
ORDER BY rate\_pct DESC;

### f. Resource Use Metrics

-- Hospital stay by admission type  
SELECT admission\_type\_id, AVG(time\_in\_hospital) FROM diabetic\_data GROUP BY admission\_type\_id;  
  
-- Drugs vs Age  
SELECT age, AVG(num\_medications) FROM diabetic\_data GROUP BY age;

### g. Top 10 Diagnoses

SELECT diag\_1, COUNT(\*) FROM diabetic\_data GROUP BY diag\_1 ORDER BY COUNT(\*) DESC LIMIT 10;

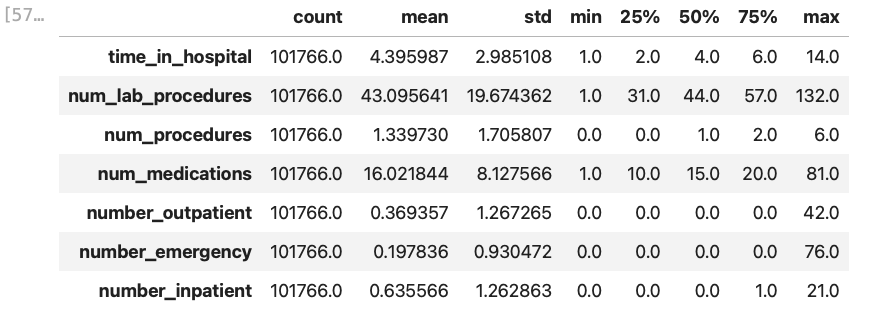
## 🐍 Section 3. Data Analysis in Jupyter (Python)

### a. Load Data and Clean

import pandas as pd  
import seaborn as sns  
import matplotlib.pyplot as plt  
  
df = pd.read\_csv("diabetic\_data.csv")  
df['race'] = df['race'].replace('?', 'Unknown')  
df['readmitted'] = df['readmitted'].replace({'&gt;30': '>30', '&lt;30': '<30'})

### b. Describe & Summary Stats

numeric\_cols = [  
 'time\_in\_hospital',  
 'num\_lab\_procedures',  
 'num\_procedures',  
 'num\_medications',  
 'number\_outpatient',  
 'number\_emergency',  
 'number\_inpatient'  
]  
  
# Get summary statistics  
df[numeric\_cols].describe()  
  
# Display mean values  
display(df[numeric\_cols].mean())  
  
# Transpose the result if you prefer vertical view  
df[numeric\_cols].describe().T

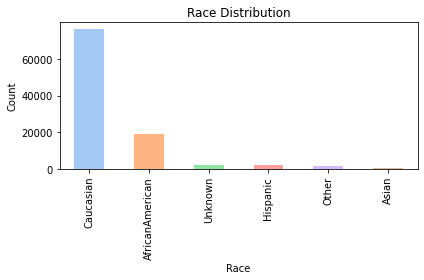


Screenshot 2025-07-31 at 10.42.01.png

### c. Feature Distribution Plots

#### i. Race Distribution

df['race'].value\_counts().plot(kind='bar', title='Race Distribution', color=sns.color\_palette('pastel'))  
plt.xlabel('Race')  
plt.ylabel('Count')  
plt.tight\_layout()  
plt.show()

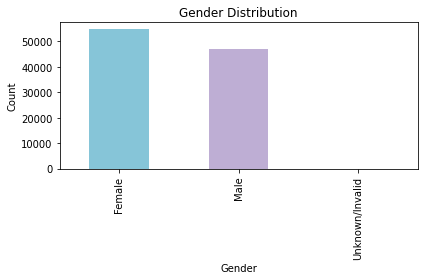


output\_8\_0.png

**Findings:**  
The cohort is predominantly Caucasian, with African American as the next largest group. About 2,000 records are “Unknown”, which can affect population health equity analyses.

#### ii. Gender Distribution

df['gender'].value\_counts().plot(kind='bar', title='Gender Distribution', color=['#86c5d8', '#beaed4', '#fdc086'])  
plt.xlabel('Gender')  
plt.ylabel('Count')  
plt.tight\_layout()  
plt.show()

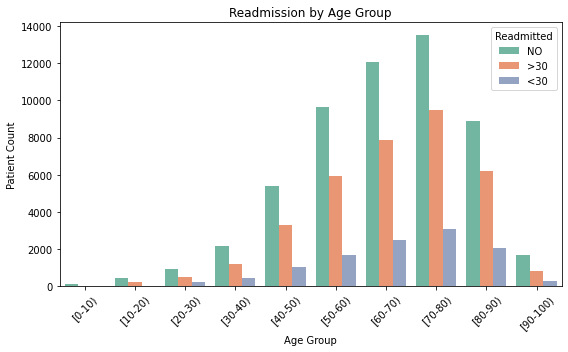


output\_10\_0.png

**Interpretation:**  
The gender split is balanced, with a slight female predominance. A very small number of “Unknown/Invalid” entries are present, which were also seen in race.

#### iii. Readmission by Age Group

plt.figure(figsize=(8,5))  
sns.countplot(x='age', hue='readmitted', data=df, palette='Set2')  
plt.title('Readmission by Age Group')  
plt.xlabel('Age Group')  
plt.ylabel('Patient Count')  
plt.xticks(rotation=45)  
plt.legend(title='Readmitted')  
plt.tight\_layout()  
plt.show()

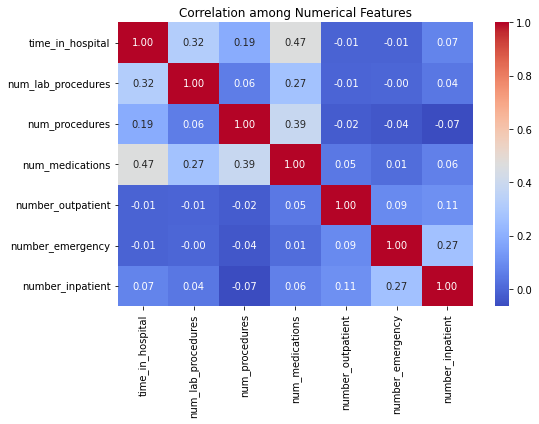


output\_12\_0.png

**Insights:**  
- Re-admissions are more frequent in older age groups (≥60). - Both the “>30” and “<30” day re-admissions are highest among elderly patients, highlighting a vulnerable population for targeted intervention.

#### iv. Correlation among Numerical Features

num\_cols = ['time\_in\_hospital','num\_lab\_procedures','num\_procedures','num\_medications','number\_outpatient','number\_emergency','number\_inpatient']  
plt.figure(figsize=(8,6))  
  
# Select these columns from df, then call .corr()  
sns.heatmap(df[num\_cols].corr(), annot=True, cmap='coolwarm', fmt='.2f')  
plt.title('Correlation among Numerical Features')  
plt.tight\_layout()  
plt.show()

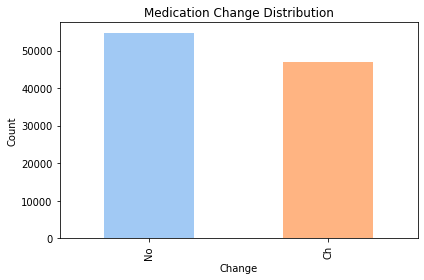


output\_14\_0.png

**Key Points:**  
- time\_in\_hospital is most strongly correlated with num\_medications (r ≈ 0.47) and num\_lab\_procedures (r ≈ 0.32). - Previous encounters (number\_inpatient, etc.) are not strongly correlated with resource usage, indicating other drivers for hospital resource consumption. - No multi-collinearity concern found.

#### v. Medication Change Distribution

df['change'].value\_counts().plot(kind='bar', title='Medication Change Distribution', color=sns.color\_palette('pastel'))  
plt.xlabel('Change')  
plt.ylabel('Count')  
plt.tight\_layout()  
plt.show()

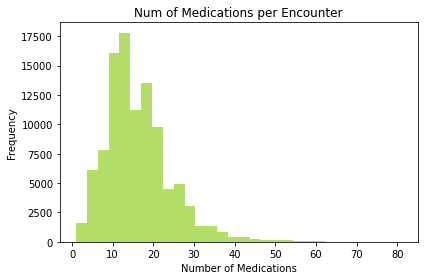


output\_16\_0.png

**Observation:**  
About half the patients had their medication regimens changed, supporting CAP\_healthcare suggestion to analyze therapy management as a potential influence on readmission.

#### vi. Number of Medications per Encounter

df['num\_medications'].plot(kind='hist', bins=30, title='Num of Medications per Encounter', color='#b3de69')  
plt.xlabel('Number of Medications')  
plt.ylabel('Frequency')  
plt.tight\_layout()  
plt.show()

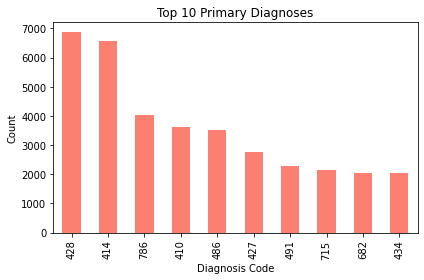


output\_18\_0.png

**Insight:**  
Most encounters involved 10-20 medications. High medication count may be a proxy for complex comorbidity or polypharmacy, which deserves attention for risk management.

#### vi. Top 10 Primary Diagnoses

df['diag\_1'].value\_counts().head(10).plot(kind='bar', color='#fb8072')  
plt.title('Top 10 Primary Diagnoses')  
plt.xlabel('Diagnosis Code')  
plt.ylabel('Count')  
plt.tight\_layout()  
plt.show()

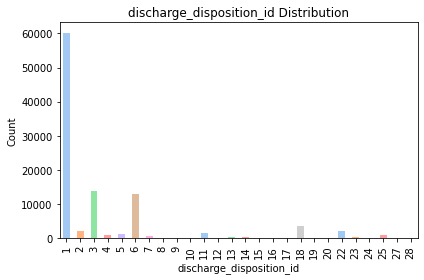
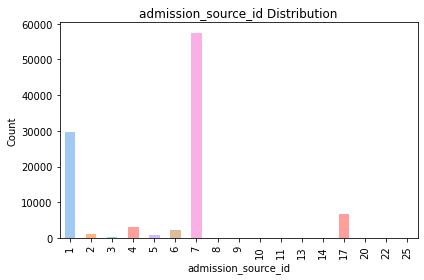
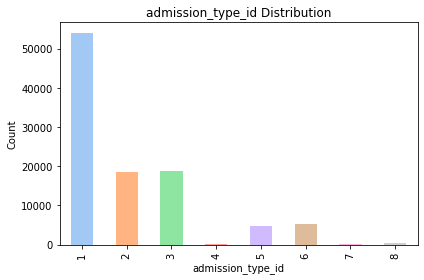


output\_20\_0.png

**Finding:**  
The most frequent diagnoses (e.g., 428, 414) typically correspond to chronic complications of diabetes, such as heart failure and ischemic heart disease, which aligns with the CAP\_healthcare clinical recommendations for prioritizing cardiovascular care.

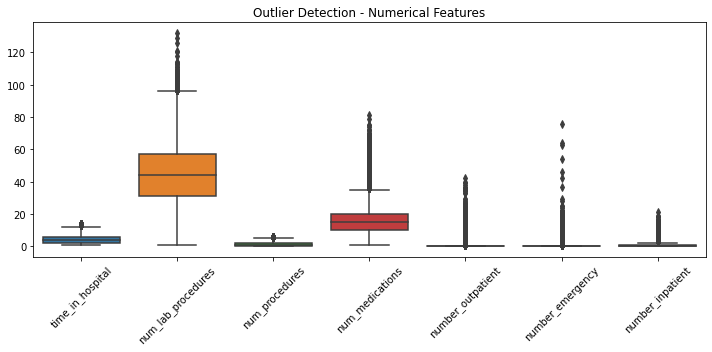
#### vii. Distribution of Admission/Source/Discharge

for col in ['admission\_type\_id', 'admission\_source\_id', 'discharge\_disposition\_id']:  
 df[col].value\_counts().sort\_index().plot(  
 kind='bar',  
 title=f'{col} Distribution',  
 color=sns.color\_palette('pastel')  
 )  
 plt.xlabel(col)  
 plt.ylabel('Count')  
 plt.tight\_layout()  
 plt.show()

 > **Commentary:**  
> - Most patients are admitted as emergencies or from outpatient referrals. > - Discharge dispositions show the majority returning home, but non-home discharges could be explored for their relationship to readmission risk.

#### viii. Outlier Detection – Numerical Features

plt.figure(figsize=(10,5))  
sns.boxplot(data=df[num\_cols])  
plt.title('Outlier Detection - Numerical Features')  
plt.xticks(rotation=45)  
plt.tight\_layout()  
plt.show()



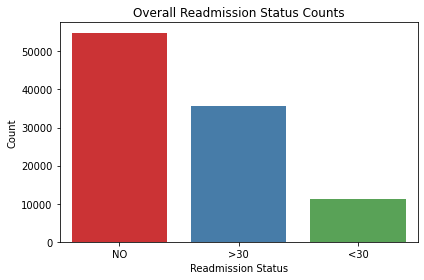
output\_24\_0.png

**Interpretation:**  
The box plot highlights outliers in procedures and previous visits. These may represent super-utilizers or possible data entry anomalies, and should be carefully checked before predictive modeling.

#### ix. Enhancement: Additional Visualizations

**A. Readmission Counts**

sns.countplot(x= 'readmitted', data=df, palette='Set1')  
plt.title('Overall Readmission Status Counts')  
plt.xlabel('Readmission Status')  
plt.ylabel('Count')  
plt.tight\_layout()  
plt.show()

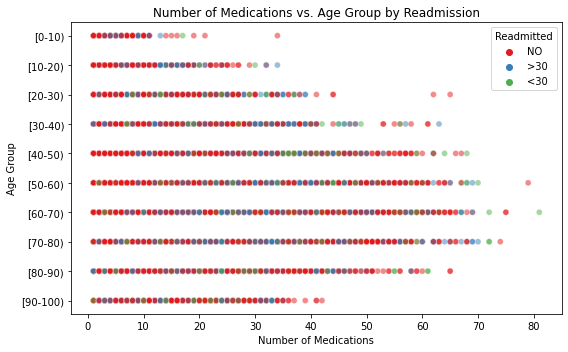


output\_26\_0.png

**Takeaway:**  
The majority of patients are not readmitted, but nearly half experience some form of readmission, highlighting the clinical and operational importance of this issue.

**B. Age vs. Number of Medications (Scatter for Complexity)**

plt.figure(figsize=(8,5))  
sns.scatterplot(data=df,x='num\_medications',y='age',hue='readmitted',alpha=0.5,palette='Set1')  
plt.title('Number of Medications vs. Age Group by Readmission')  
plt.xlabel('Number of Medications')  
plt.ylabel('Age Group')  
plt.legend(title='Readmitted')  
plt.tight\_layout()  
plt.show()



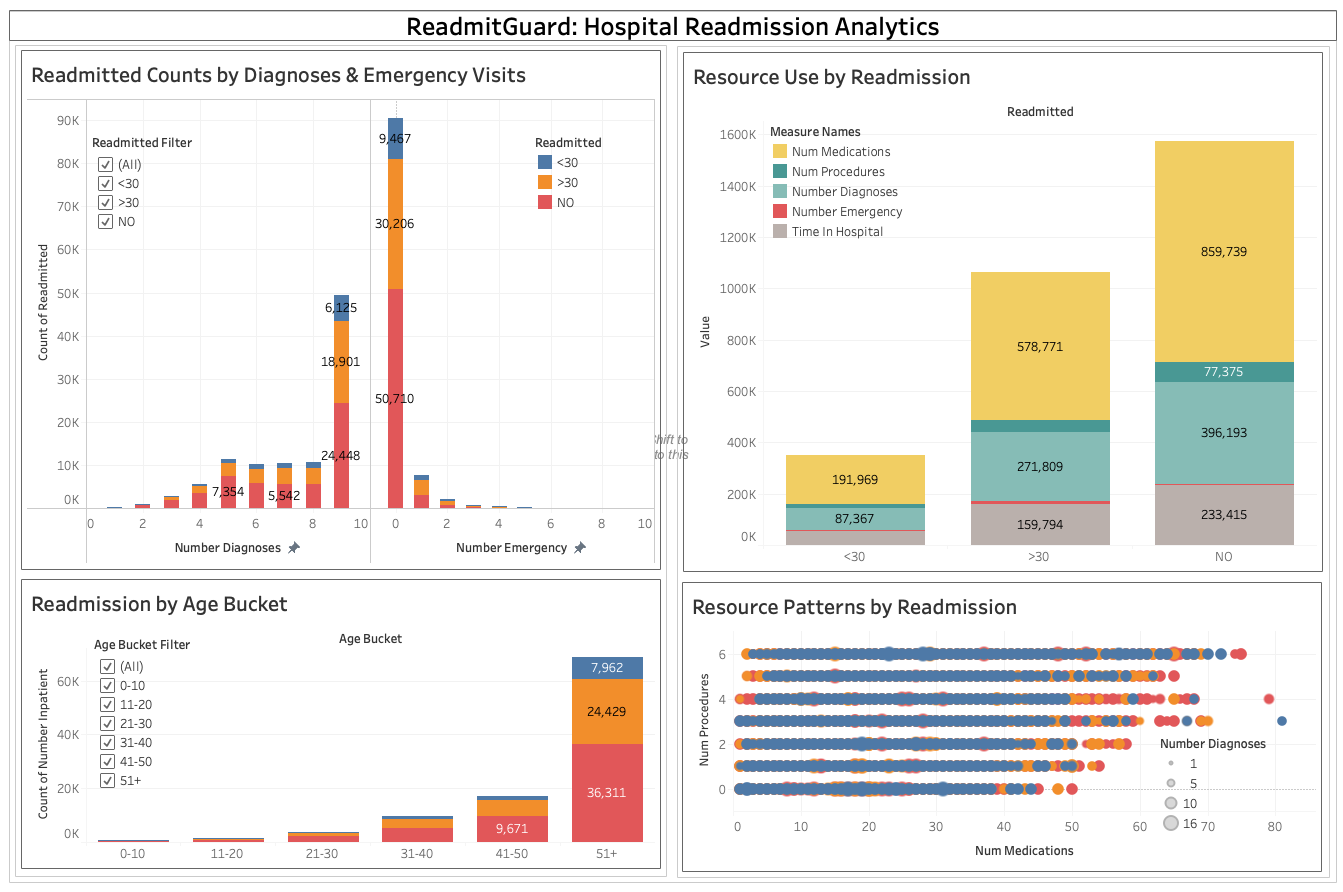
output\_28\_0.png

**Note:**  
This plot reveals patterns between age, poly-pharmacy, and readmission. Considered alongside earlier findings, multi-morbidity in older patients is visually evident.

## 📊 Section 4. Tableau Public Dashboard (Interactive)

### a. Layout Design

* **Fixed size:** 1366 x 900 px
* **Containers:**
  + 1 Vertical ➝ 1 Horizontal ➝ 2 Vertical (Left & Right)
* Charts:
  + Readmitted Counts (Diagnosis/Emergency)
  + Age vs Readmitted
  + Resource Usage by Readmitted
  + Bubble Chart: Medications vs Diagnoses



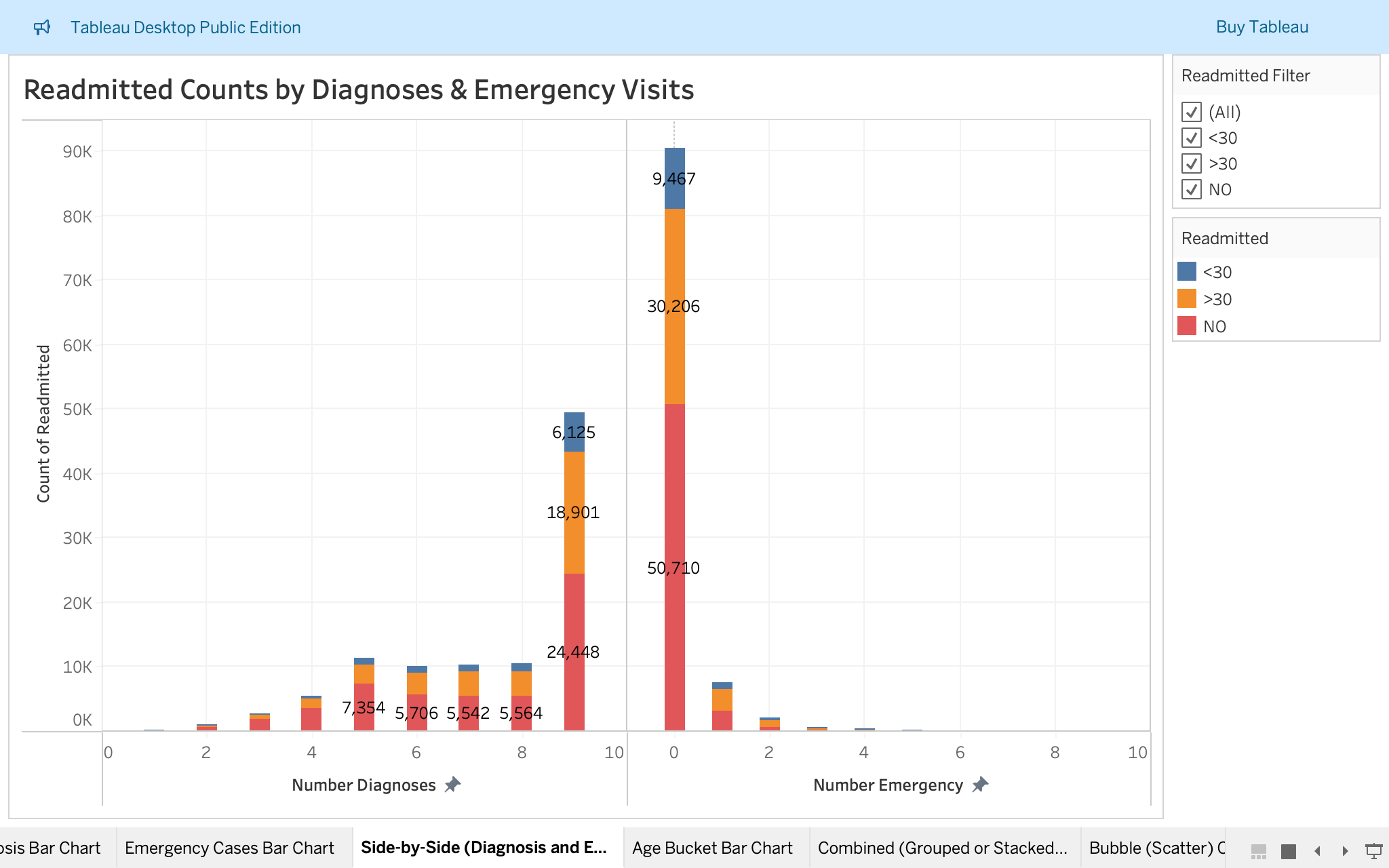
Screenshot 2025-08-03 at 11.15.04.png

### b. Fields & Calculated Columns

* Age Bucket (calculated from age string)
* **Calculated Fields:**
  + Age Numeric:
  + INT(LEFT(REPLACE(REPLACE([age], "[", ""), ")", ""), 2))
  + Age Bucket:
  + IF [Age Numeric] <= 10 THEN "0-10"  
    ELSEIF [Age Numeric] <= 20 THEN "11-20"  
    ELSEIF [Age Numeric] <= 30 THEN "21-30"  
    ELSEIF [Age Numeric] <= 40 THEN "31-40"  
    ELSEIF [Age Numeric] <= 50 THEN "41-50"  
    ELSE "51+"  
    END
* Custom filters: readmitted, Age Bucket

#### i. Readmitted Counts by Diagnoses and Emergencies

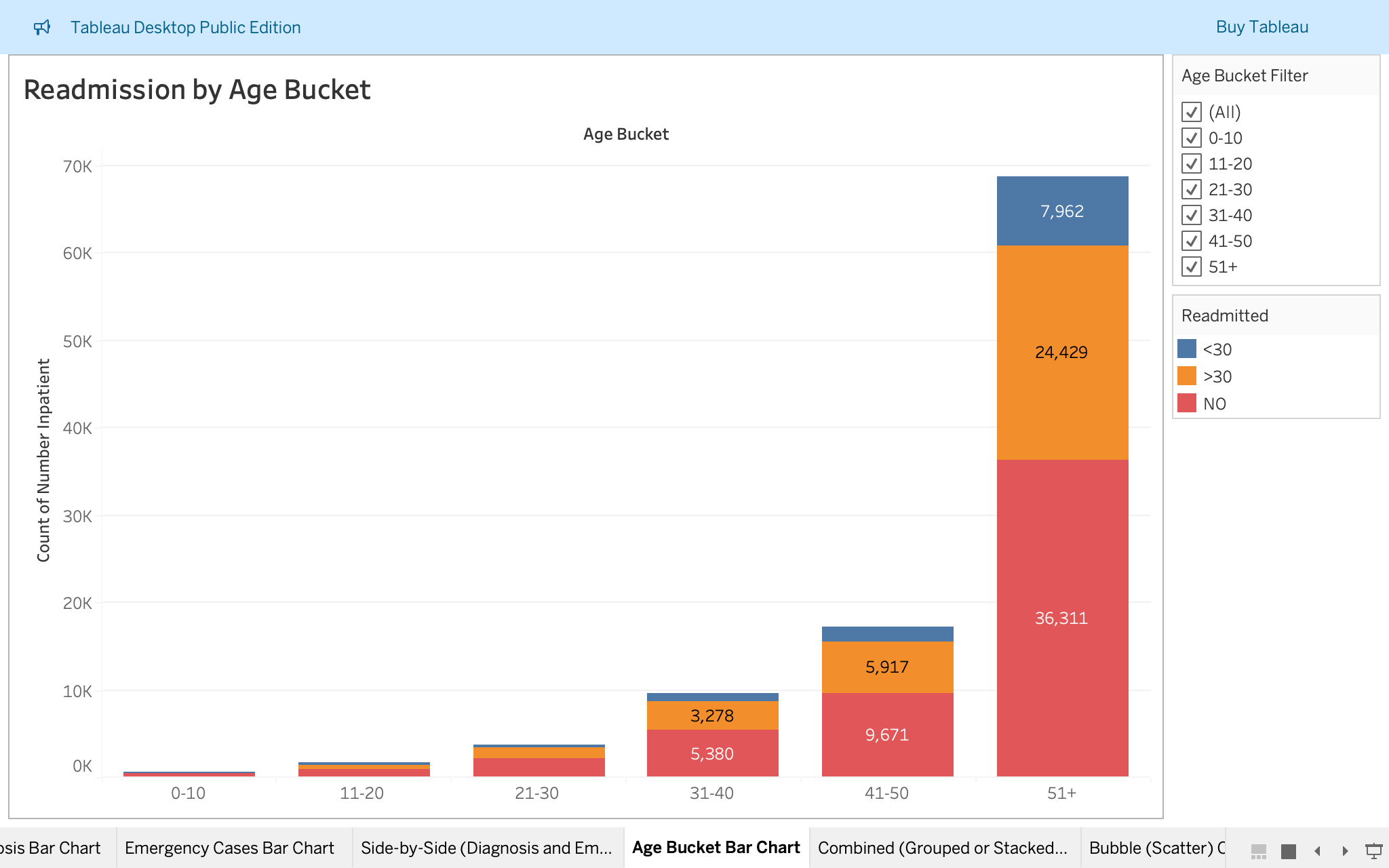
* **Type:** Stacked/side-by-side bar chart
* **X-Axis:** number\_diagnoses (or number\_emergency)
* **Y-Axis:** Count (by readmitted)
* **Color:** readmitted status (NO, <30, >30)
* **Worksheet Name:** Readmitted Counts by Diagnoses & Emergencies
* **Purpose:** Bar: Diagnoses/Emergencies by Readmission



Screenshot 2025-08-03 at 11.28.02.png

#### ii. Readmission by Age Bucket

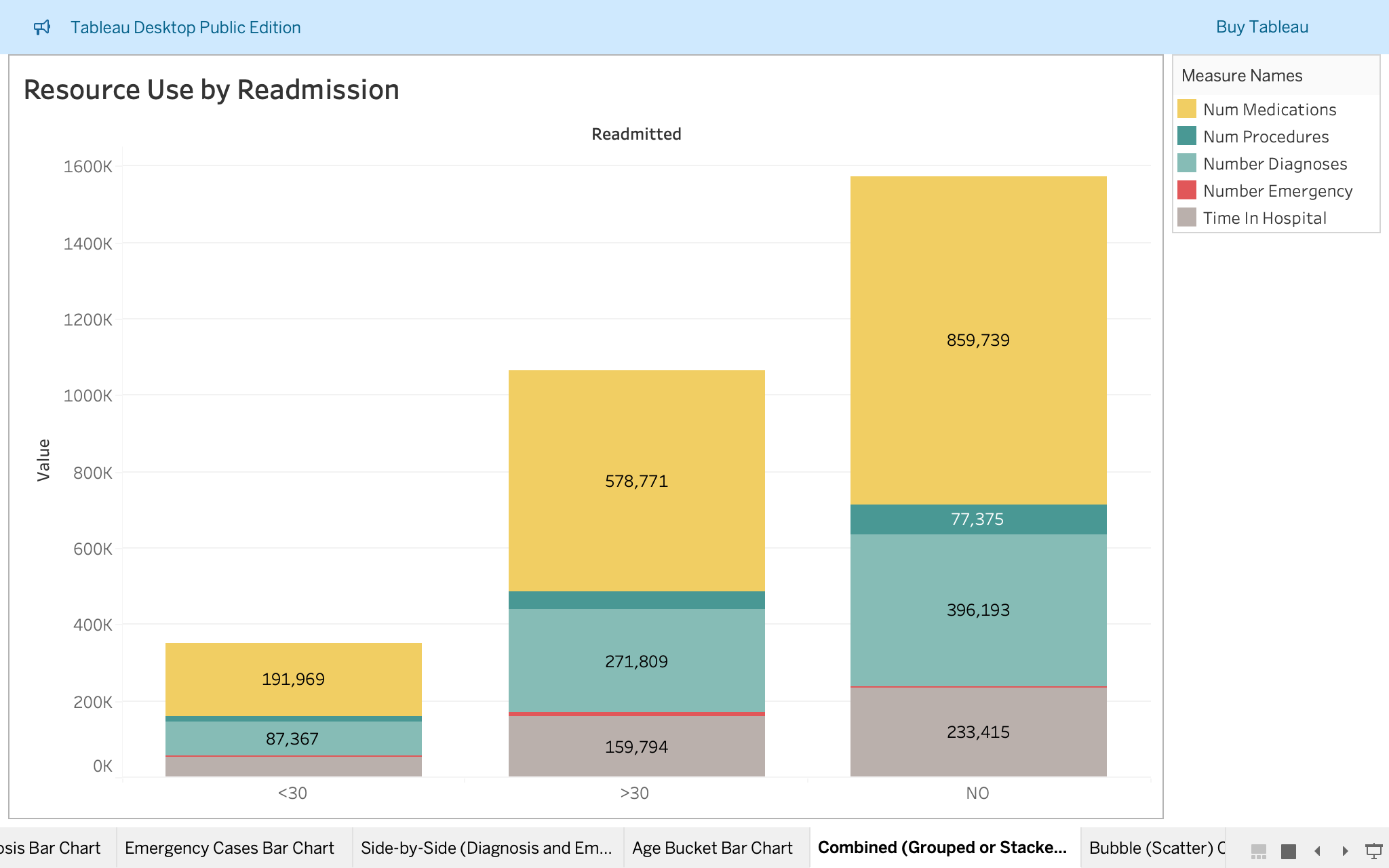
* **Type:** Stacked/clustered bar chart
* **X-Axis:** Age Bucket
* **Y-Axis:** Count of number\_inpatient
* **Color:** readmitted
* **Labels:** Show mark labels
* **Worksheet Name:** Readmission by Age Bucket
* **Purpose:** Bar: Age group distribution across readmission



Screenshot 2025-08-03 at 11.28.14.png

#### iii. Resource Use by Readmission

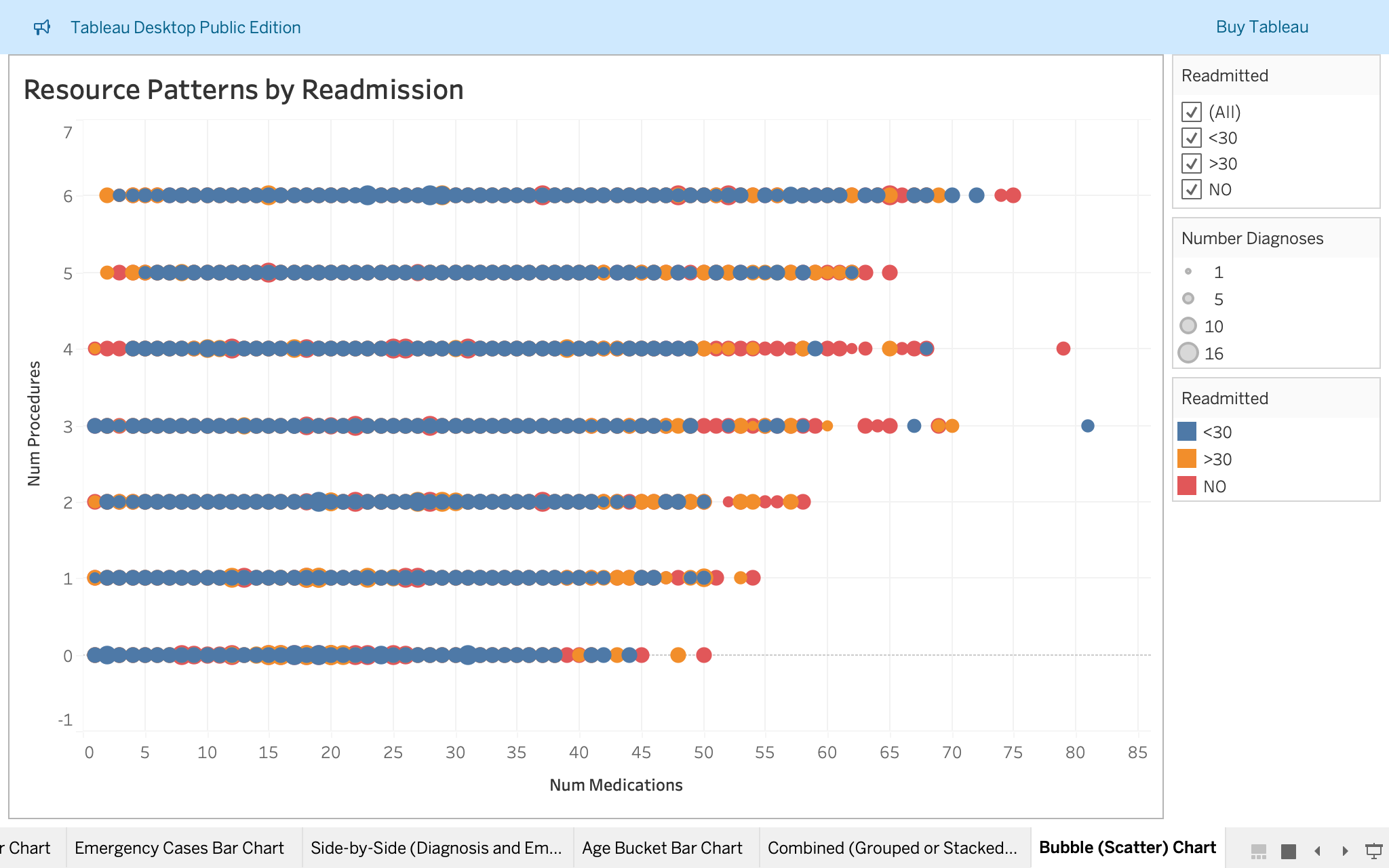
* **Type:** Grouped or stacked bar chart using “Measure Names and Measure Values.”
* **Measures Visualized:**
  + num\_medications, num\_procedures, number\_diagnoses, number\_emergency, time\_in\_hospital
* **Worksheet Name:** Resource Use by Readmission
* **Interactivity:** Use color or label for Measure Names; change aggregation as needed (sum/avg)
* **Purpose:** Bar: Compare clinical metric averages



Screenshot 2025-08-03 at 11.28.21.png

#### iv. Resource Patterns by Readmission (Bubble Chart)

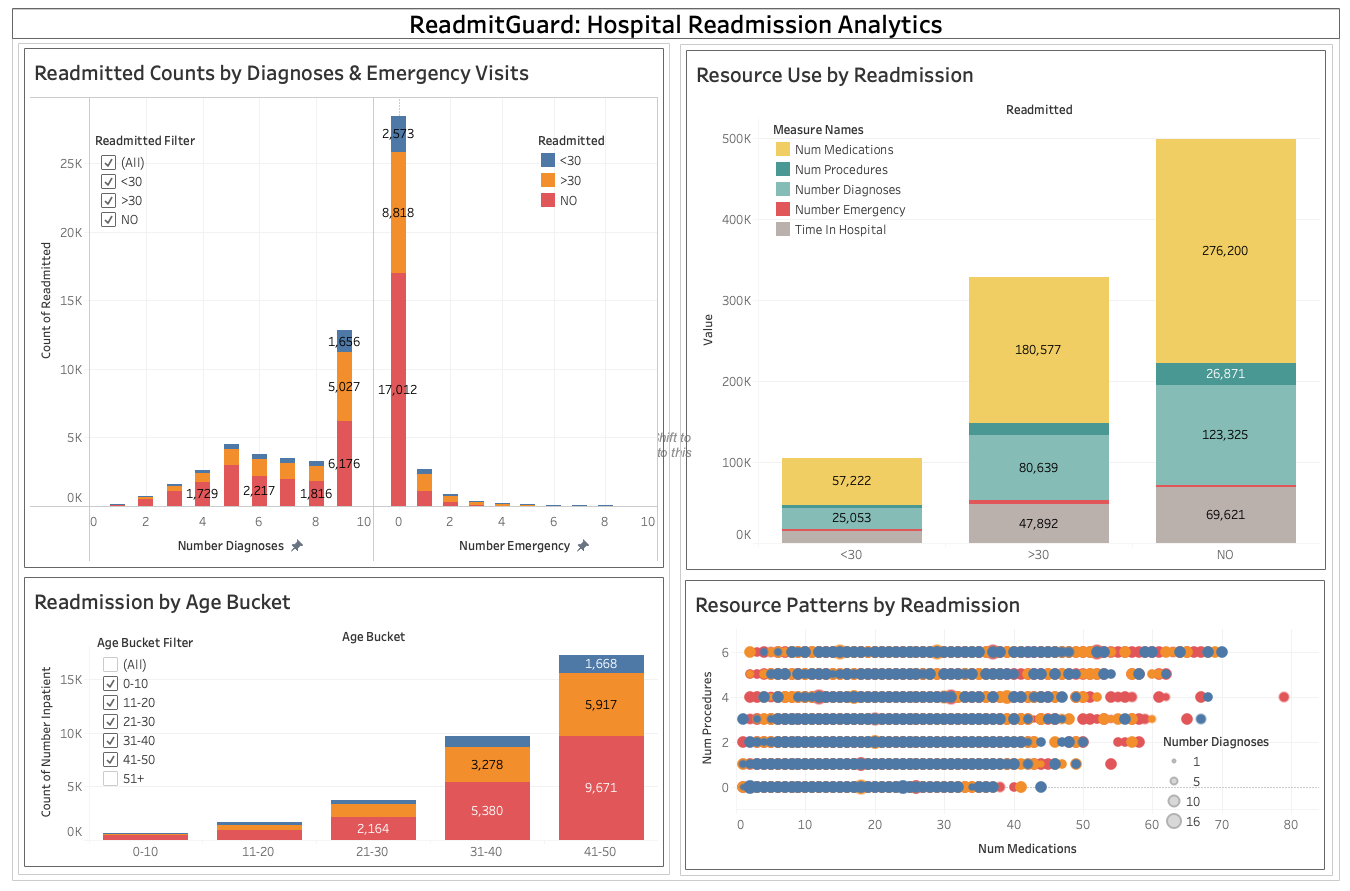
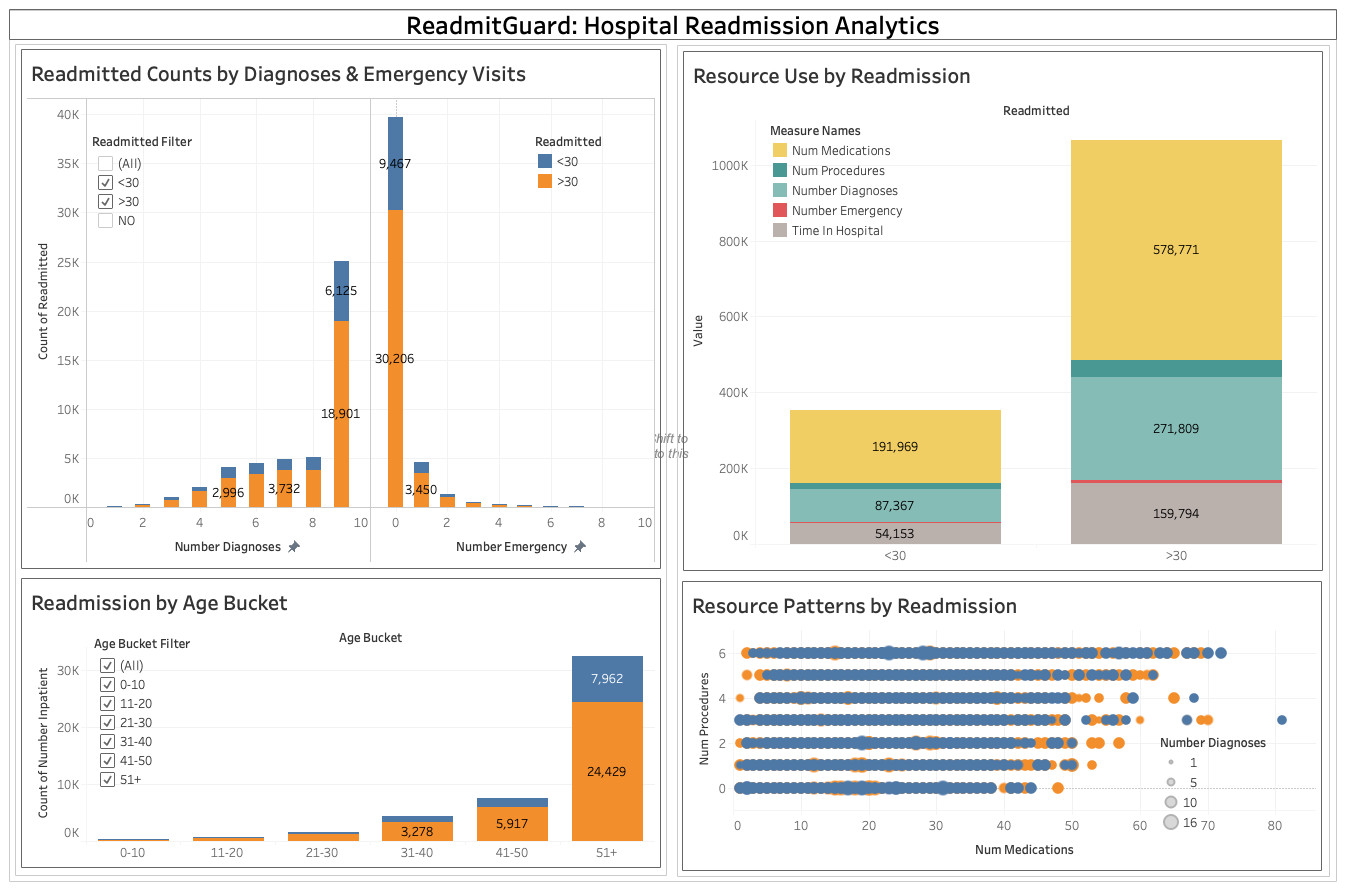
* **Type:** Scatter/Bubble Chart
* **X-Axis:** num\_medications
* **Y-Axis:** num\_procedures
* **Size:** number\_diagnoses
* **Color:** readmitted
* **Shape/Filter:** Optionally add readmitted
* **Worksheet Name:** Resource Patterns by Readmission
* **Purpose:** Bubble: Medications, Labs, Diagnoses pattern



Screenshot 2025-08-03 at 11.28.27.png

### **C. Interactivity via Filtering**

* **Add readmitted and Age Bucket filters** (as drop-down or list) to the dashboard
* **Make charts “interactive filters”** so clicking a bar or segment filters all other charts
* **Set filters as multi-select for group comparison, or single-select for detailed drill-down**
* **Synchronize filters across all worksheets** by right-clicking filter > “Apply to Worksheets” > “All Using This Data Source”



## ✅ Summary Conclusion: ReadmitGuard Project

The **ReadmitGuard healthcare analytics guide** provides a streamlined, end-to-end workflow for analyzing diabetic patient re-admissions using:

* **Excel/Sheets** for quick summary statistics and pivot tables.
* **SQL** to structure, cleanse, and summarize healthcare metrics.
* **Python (Pandas + Seaborn)** for deeper exploratory data analysis.
* **Tableau Public** for interactive and presentation-ready dashboards.

This integrated approach:

* Reveals **high-risk patient segments** (e.g., older poly-pharmacy cases).
* Highlights **patterns in diagnoses, procedures, and discharge outcomes.**
* Empowers both **technical and non-technical users** to gain actionable insights.
* Delivers a **reproducible, modular pipeline** that’s easy to adapt for other healthcare datasets.