HEALTHCARE DATA EXPLORATORY ANALYSIS

SUMMARY

- This report presents the findings of an exploratory data analysis (EDA) project as we dive into an entirely synthetic dataset.
- Healthcare data obtained from https://www.kaggle.com/datasets/prasad22/healthcare-dataset
- Encompassing patient demographics, admission specifics, medical conditions, and healthcare services, this dataset provides a rich landscape for analysis.
- The overarching objective is to explore, cleanse, and visualize the data, thereby unraveling healthcare trends and understanding the distribution of medical conditions and the role they play in hospitalizations.

CLEANING STEPS

- Imported healthcare_dataset.csv file into Excel
- Adjusted the Billing Amount column from text to currency
- Split Name column into 4 columns for name1, name2, name3, name4
- Removed prefixes and postnominals
    - =IF(OR(B2="Mr.", B2="Mrs."), C2, B2)
    - =IF(A2<>C2, C2, D2)
        - If name1 is not equal to name2, name2, name3
    - =IF(OR(D2="Jr.", D2="II“, D2=“III”, D2=“IV”), B2 & " " & D2, B2)
        - Combined suffix with name2
    - Deleted unneeded columns
- Renamed column name1 to first_name and column name2 to last_name
- Added column to calculate number of days admitted

INITIAL ANALYSIS

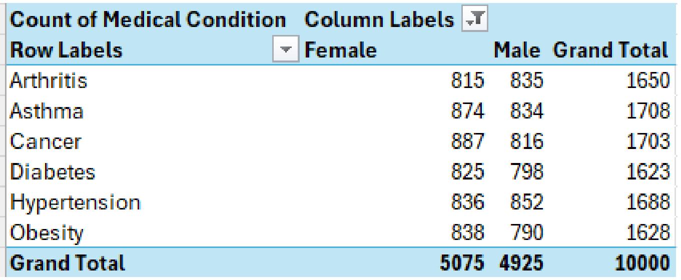
EXCEL – PIVOT TABLES

- The sample file consists of 51% female and 49% male patients, with 6 different medical conditions, spanning 5 years.

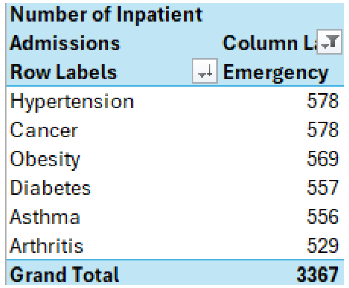
- The leading causes of Emergency inpatient admissions were Hypertension and Cancer.

- The average number of inpatient days for all medical conditions combined was 15.56.

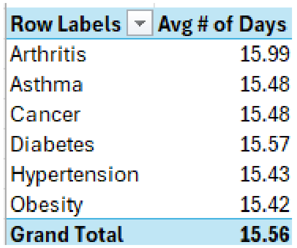
- The month of October has the highest number of inpatient admissions, however this is helped by a higher than average number of admissions that occurred in October 2022 (207 vs 166 on average for all months excluding Oct 2022 and Oct 2018 where there were only 8 admissions). That has not been realized since, as the average monthly inpatient rate following October 2022 are down slightly at 160 per month.



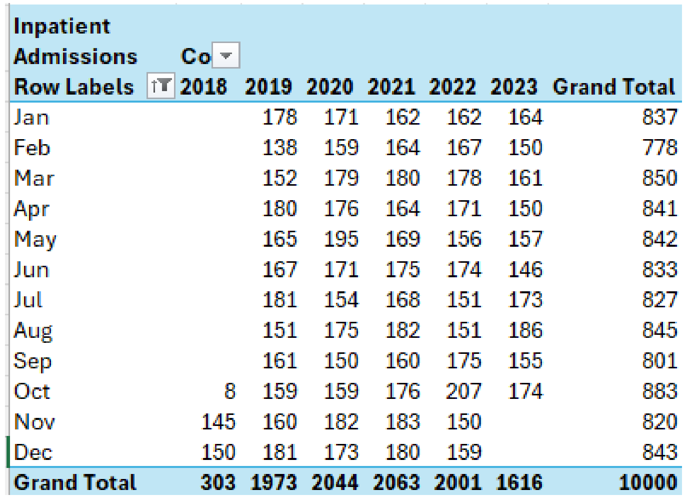
Count of Medical Condition	Column Labels		
Row Labels	Female	Male	Grand Total
Arthritis	815	835	1650
Asthma	874	834	1708
Cancer	887	816	1703
Diabetes	825	798	1623
Hypertension	836	852	1688
Obesity	838	790	1628
Grand Total	5075	4925	10000



Number of Inpatient Admissions	Column L
Row Labels	Emergency
Hypertension	578
Cancer	578
Obesity	569
Diabetes	557
Asthma	556
Arthritis	529
Grand Total	3367



Row Labels	Avg # of Days
Arthritis	15.99
Asthma	15.48
Cancer	15.48
Diabetes	15.57
Hypertension	15.43
Obesity	15.42
Grand Total	15.56

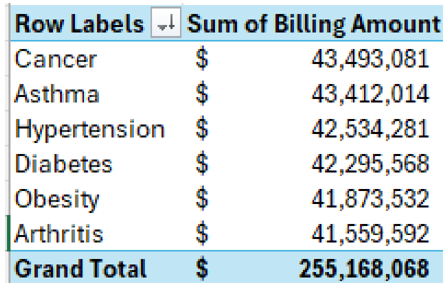


Inpatient Admissions	Co						
Row Labels	2018	2019	2020	2021	2022	2023	Grand Total
Jan		178	171	162	162	164	837
Feb		138	159	164	167	150	778
Mar		152	179	180	178	161	850
Apr		180	176	164	171	150	841
May		165	195	169	156	157	842
Jun		167	171	175	174	146	833
Jul		181	154	168	151	173	827
Aug		151	175	182	151	186	845
Sep		161	150	160	175	155	801
Oct	8	159	159	176	207	174	883
Nov	145	160	182	183	150		820
Dec	150	181	173	180	159		843
Grand Total	303	1973	2044	2063	2001	1616	10000

INITIAL ANALYSIS

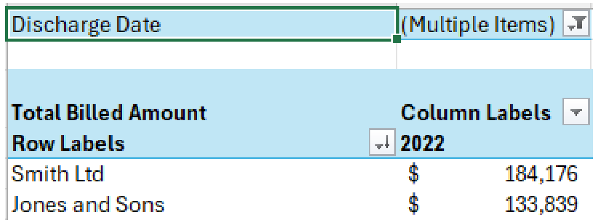
EXCEL – PIVOT TABLES

- Billed amounts for Cancer were the highest at just under $43.5M and attributed to 17.04% of the total billed amounts for all medical conditions.



Row Labels	Sum of Billing Amount
Cancer	$ 43,493,081
Asthma	$ 43,412,014
Hypertension	$ 42,534,281
Diabetes	$ 42,295,568
Obesity	$ 41,873,532
Arthritis	$ 41,559,592
Grand Total	**$ 255,168,068**

- Smith Ltd hospital billed the most in 2022 at $184,176. This was 11% higher than the second highest biller for that year which was Jones and Sons hospital.



Discharge Date	(Multiple Items)
Total Billed Amount	**Column Labels**
Row Labels	**2022**
Smith Ltd	$ 184,176
Jones and Sons	$ 133,839

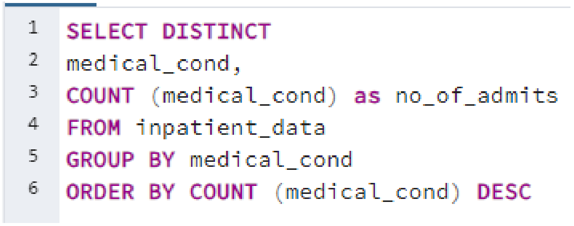
ANALYSIS PREPARATION

POSTGRESQL

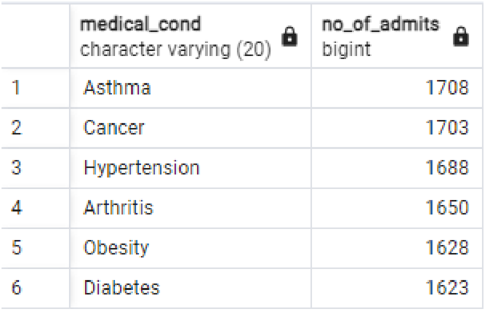
1. Created database 'healthcare_data' in pgAdmin
2. Created table 'inpatient_data'
3. Imported healthcare_dataset.csv file into the 'inpatient_data' table

INITIAL FINDINGS

- Asthma has the highest number of impatient admissions overall, followed closely by cancer. Diabetes has the lowest number of impatient admissions.

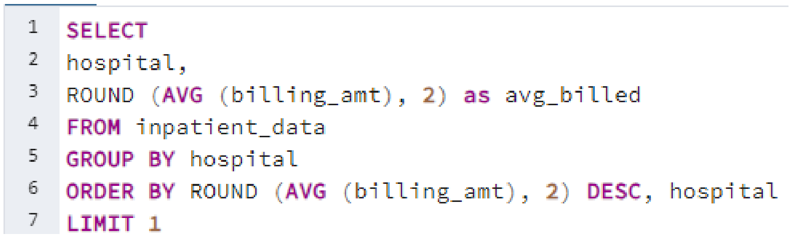


```
SELECT DISTINCT
medical_cond,
COUNT (medical_cond) as no_of_admits
FROM inpatient_data
GROUP BY medical_cond
ORDER BY COUNT (medical_cond) DESC
```

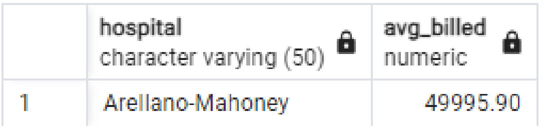


	medical_cond character varying (20)	no_of_admits bigint
1	Asthma	1708
2	Cancer	1703
3	Hypertension	1688
4	Arthritis	1650
5	Obesity	1628
6	Diabetes	1623

- The hospital with the highest average billed amount is Arellano-Mahoney at just under 50K.



```
SELECT
hospital,
ROUND (AVG (billing_amt), 2) as avg_billed
FROM inpatient_data
GROUP BY hospital
ORDER BY ROUND (AVG (billing_amt), 2) DESC, hospital
LIMIT 1
```



	hospital character varying (50)	avg_billed numeric
1	Arellano-Mahoney	49995.90

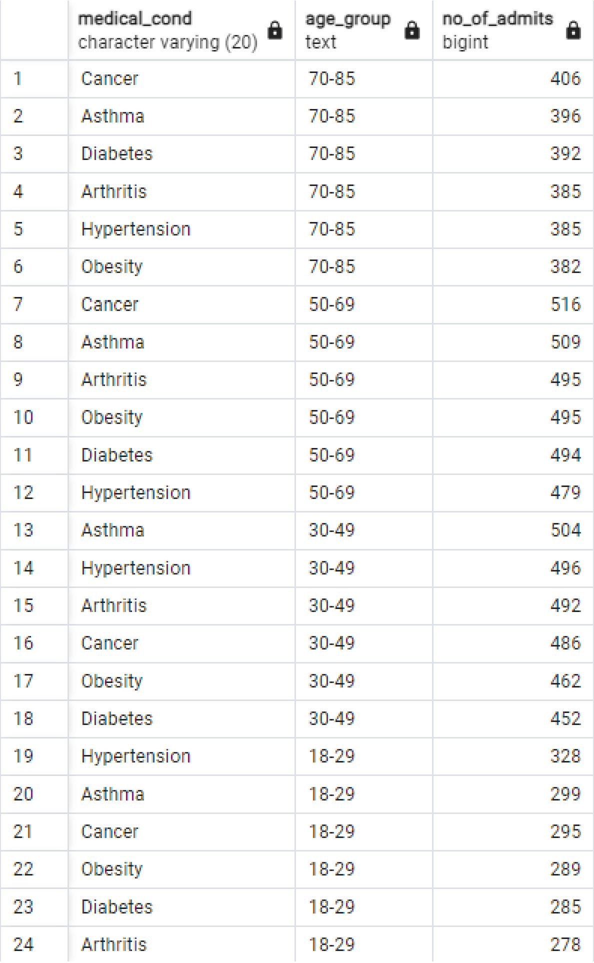
ANALYSIS FINDINGS CONT’D

POSTGRESQL

- When grouping by age ranges, Cancer has the highest rate of inpatient admissions for 70-85 year-olds, and 50-69 year-olds.

- Asthma is the highest for 30-49 year-olds, and Hypertension is the highest for 18-29 year-olds.

```
SELECT
medical_cond,
CASE
        WHEN age< 30 THEN '18-29'
        WHEN age < 50 THEN '30-49'
        WHEN age < 70 THEN '50-69’
        ELSE '70-85’
END as age_group,
COUNT (medical_cond) as no_of_admits
FROM inpatient_data
GROUP BY age_group, medical_cond
ORDER BY age_group DESC, COUNT (medical_cond) DESC
```



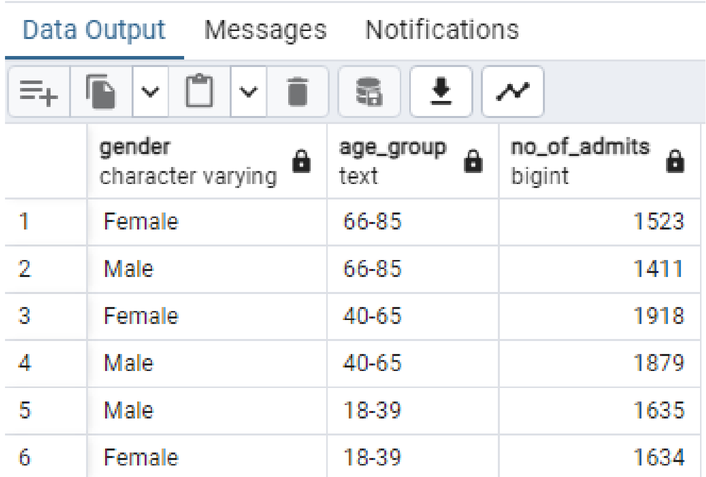
	medical_cond character varying (20)	age_group text	no_of_admits bigint
1	Cancer	70-85	406
2	Asthma	70-85	396
3	Diabetes	70-85	392
4	Arthritis	70-85	385
5	Hypertension	70-85	385
6	Obesity	70-85	382
7	Cancer	50-69	516
8	Asthma	50-69	509
9	Arthritis	50-69	495
10	Obesity	50-69	495
11	Diabetes	50-69	494
12	Hypertension	50-69	479
13	Asthma	30-49	504
14	Hypertension	30-49	496
15	Arthritis	30-49	492
16	Cancer	30-49	486
17	Obesity	30-49	462
18	Diabetes	30-49	452
19	Hypertension	18-29	328
20	Asthma	18-29	299
21	Cancer	18-29	295
22	Obesity	18-29	289
23	Diabetes	18-29	285
24	Arthritis	18-29	278

ANALYSIS FINDINGS CONT’D

POSTGRESQL

- Adjusting to 3 age bands and diving into gender differences, women have higher inpatient rates than men between the ages of 66-85 and 40-65 (7.9% and 2.1% respectively).
- Inpatient rates for both men and women between the ages of 18-39 are virtually even.

```
SELECT

gender,

CASE

        WHEN age < 40 THEN '18-39’

        WHEN age < 66 THEN '40-65’

        ELSE '66-85'

END as age_group,

COUNT (medical_cond) as no_of_admits

FROM inpatient_data

GROUP BY age_group, gender

ORDER BY age_group DESC, COUNT (medical_cond) DESC
```



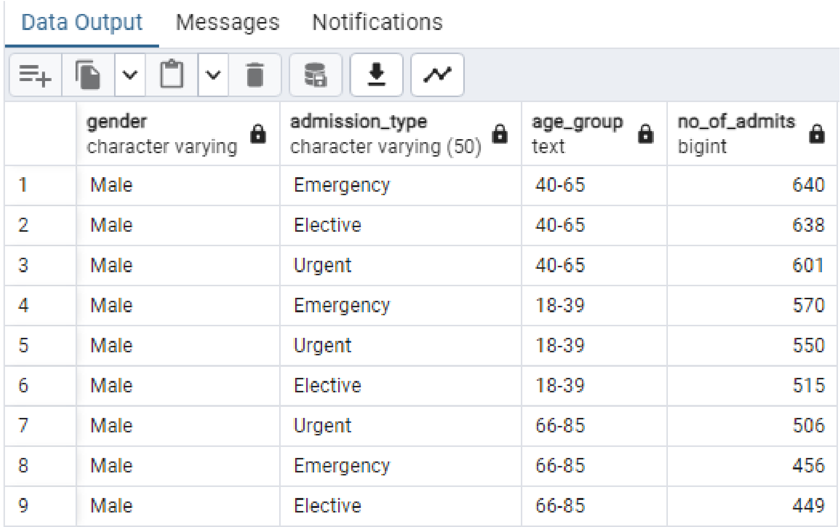
	gender character varying	age_group text	no_of_admits bigint
1	Female	66-85	1523
2	Male	66-85	1411
3	Female	40-65	1918
4	Male	40-65	1879
5	Male	18-39	1635
6	Female	18-39	1634

ANALYSIS FINDINGS CONT’D

POSTGRESQL

- Men between the ages of 40-65 had a 40% higher emergency inpatient rate than those between the ages of 66-85.

```
SELECT

gender,

admission_type,

CASE

        WHEN age< 40 THEN '18-39’

        WHEN age < 66 THEN '40-65’

        ELSE '66-85’

        END as age_group,

COUNT (medical_cond) as no_of_admits

FROM inpatient_data

WHERE gender = 'Male'

GROUP BY age_group, gender, admission_type

ORDER BY COUNT (medical_cond) DESC
```



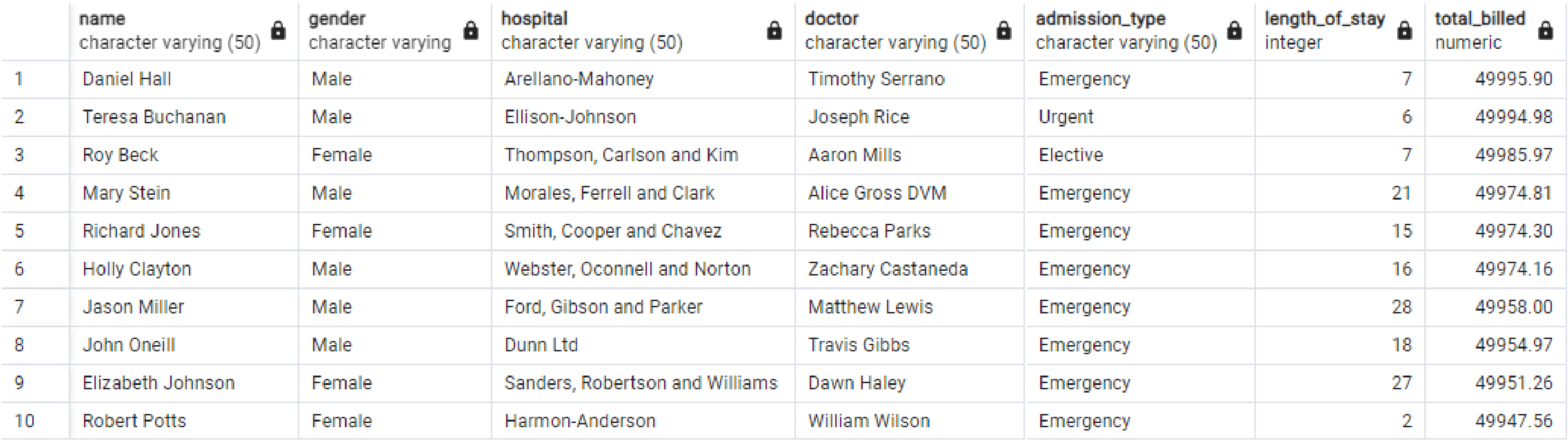
	gender character varying	admission_type character varying (50)	age_group text	no_of_admits bigint
1	Male	Emergency	40-65	640
2	Male	Elective	40-65	638
3	Male	Urgent	40-65	601
4	Male	Emergency	18-39	570
5	Male	Urgent	18-39	550
6	Male	Elective	18-39	515
7	Male	Urgent	66-85	506
8	Male	Emergency	66-85	456
9	Male	Elective	66-85	449

ANALYSIS FINDINGS CONT'D

POSTGRESQL

- Of the top 10 highest billed amounts, 60% were male and 40% were female patients.
- 80% were emergency admissions
- The length of stay ranged from 2 days to 28 days and was 14.7 days on average.
- None of the patients went to the same hospital or had the same doctor.

```
SELECT

name, gender, hospital, doctor, admission_type,

discharge_date - admission_date as length_of_stay,

ROUND (SUM (billing_amt), 2) as total_billed

FROM inpatient_data

GROUP BY name, gender, hospital, doctor, admission_type, discharge_date,  admission_date

ORDER BY ROUND (SUM (billing_amt), 2) DESC, hospital

LIMIT 10
```

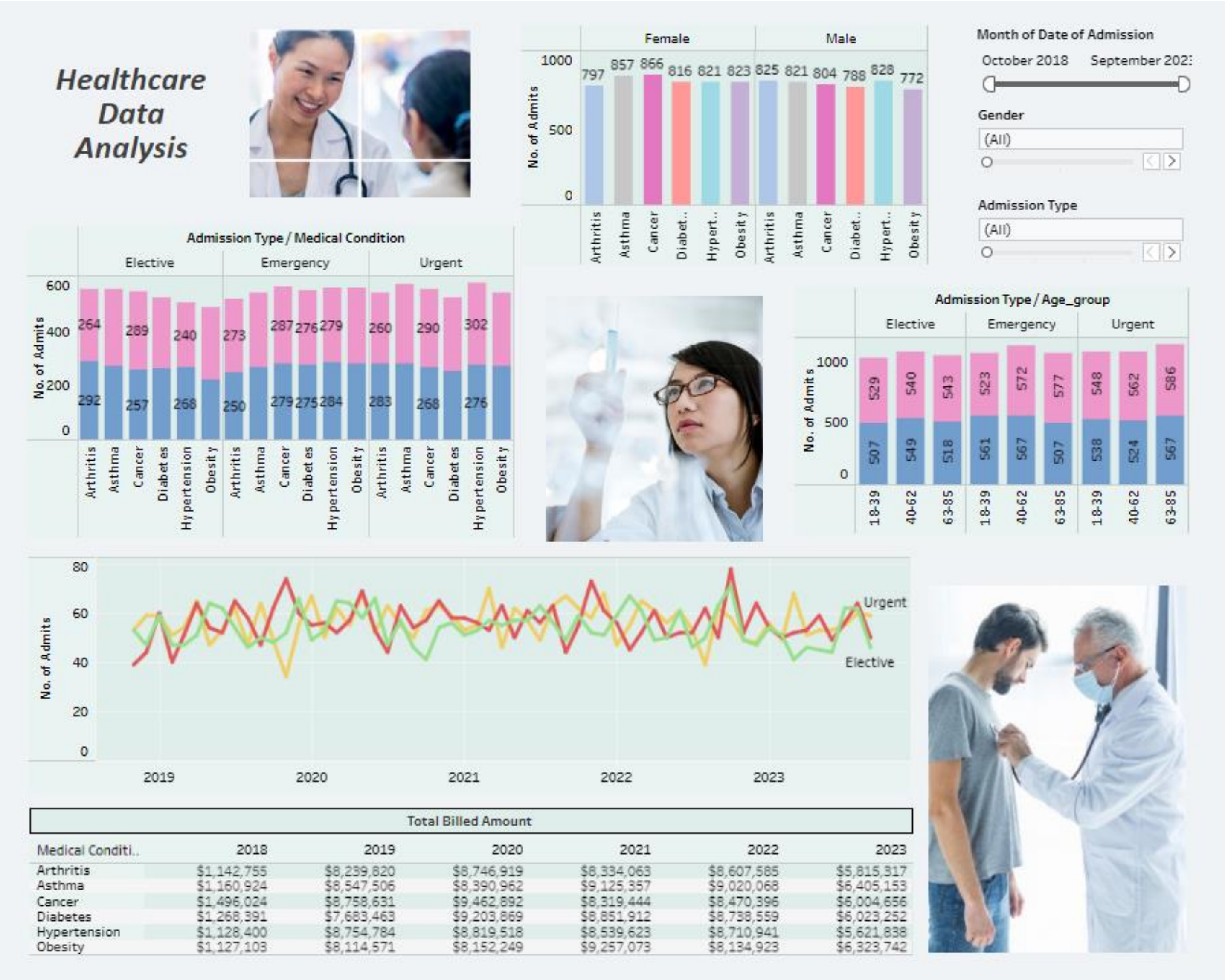


	name character varying (50)	gender character varying	hospital character varying (50)	doctor character varying (50)	admission_type character varying (50)	length_of_stay integer	total_billed numeric
1	Daniel Hall	Male	Arellano-Mahoney	Timothy Serrano	Emergency	7	49995.90
2	Teresa Buchanan	Male	Ellison-Johnson	Joseph Rice	Urgent	6	49994.98
3	Roy Beck	Female	Thompson, Carlson and Kim	Aaron Mills	Elective	7	49985.97
4	Mary Stein	Male	Morales, Ferrell and Clark	Alice Gross DVM	Emergency	21	49974.81
5	Richard Jones	Female	Smith, Cooper and Chavez	Rebecca Parks	Emergency	15	49974.30
6	Holly Clayton	Male	Webster, Oconnell and Norton	Zachary Castaneda	Emergency	16	49974.16
7	Jason Miller	Male	Ford, Gibson and Parker	Matthew Lewis	Emergency	28	49958.00
8	John Oneill	Male	Dunn Ltd	Travis Gibbs	Emergency	18	49954.97
9	Elizabeth Johnson	Female	Sanders, Robertson and Williams	Dawn Haley	Emergency	27	49951.26
10	Robert Potts	Female	Harmon-Anderson	William Wilson	Emergency	2	49947.56

ANALYSIS & VISUALIZATION

TABLEAU

HTTPS://PUBLIC.TABLEAU.COM/SHARED/3P5DCH6P9?:DISPLAY_COUNT=N&:ORIGIN=VIZ_SHARE_LINK



Healthcare Data Analysis

Month of Date of Admission
October 2018 September 202:

Gender
(All)

Admission Type
(All)

Total Billed Amount						
Medical Conditi..	2018	2019	2020	2021	2022	2023
Arthritis	$1,142,755	$8,239,820	$8,746,919	$8,334,063	$8,607,585	$5,815,317
Asthma	$1,160,924	$8,547,506	$8,390,962	$9,125,357	$9,020,068	$6,405,153
Cancer	$1,496,024	$8,758,631	$9,462,892	$8,319,444	$8,470,396	$6,004,656
Diabetes	$1,268,391	$7,683,463	$9,203,869	$8,851,912	$8,738,559	$6,023,252
Hypertension	$1,128,400	$8,754,784	$8,819,518	$8,539,623	$8,710,941	$5,621,838
Obesity	$1,127,103	$8,114,571	$8,152,249	$9,257,073	$8,134,923	$6,323,742