SQL Project: Exploratory Analysis of Synthea

Background: For this project, I utilized SQL to delve into the Synthea dataset [(1)], an open-source synthetic patient population simulator. It generates realistic synthetic patient data for research, testing, and educational purposes, encompassing a wide range of health-related information such as patient demographics, medical conditions, medications, and healthcare encounters.

The Data: The dataset used in this analysis includes 4 tables:

- **Conditions:** 156,945 rows
- **Encounters:** 455,935 rows
- **Immunizations:** 165,493 rows
- **Patients:** 11,363 rows

Once the schema and tables were created and filled with records, it was time to start analyzing.

Analysis and Insights

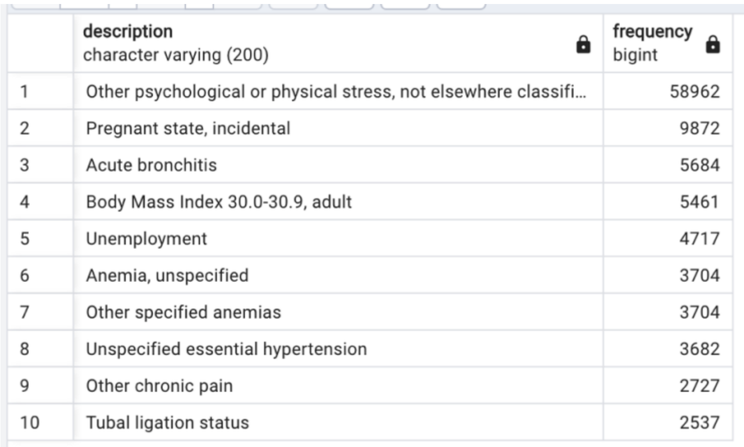
Question 1: What are the most common conditions among patients?

Why is this important? Knowing the most common conditions helps healthcare providers prioritize resources, allocate beds efficiently, and plan for treatment protocols.

Query:

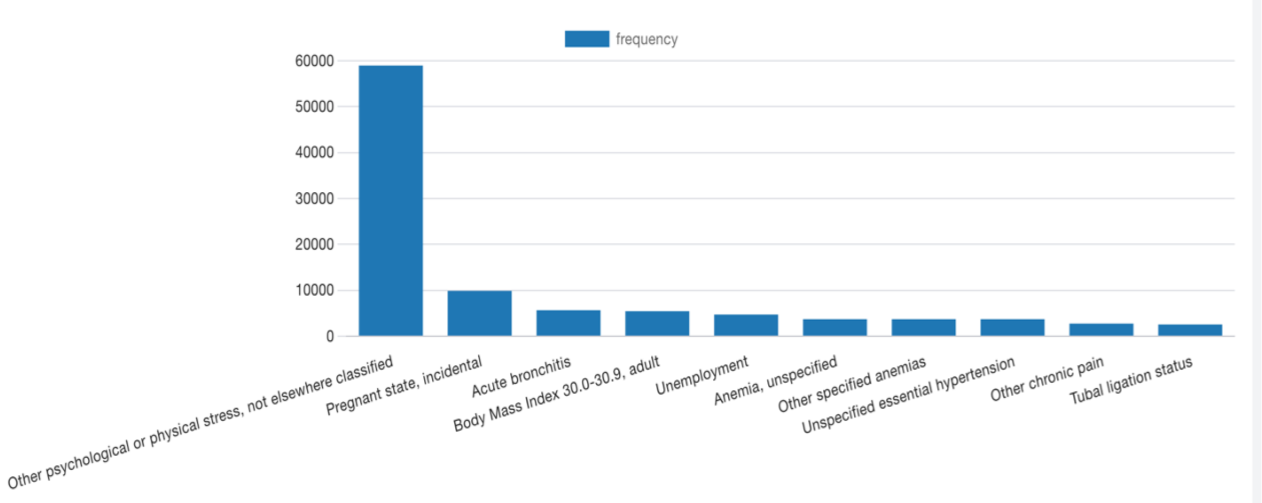
```
SELECT DESCRIPTION, COUNT(*) AS frequency
FROM conditions
GROUP BY DESCRIPTION
ORDER BY frequency DESC LIMIT 10;
```

Results:



	description character varying (200)	frequency bigint
1	Other psychological or physical stress, not elsewhere classifi…	58962
2	Pregnant state, incidental	9872
3	Acute bronchitis	5684
4	Body Mass Index 30.0-30.9, adult	5461
5	Unemployment	4717
6	Anemia, unspecified	3704
7	Other specified anemias	3704
8	Unspecified essential hypertension	3682
9	Other chronic pain	2727
10	Tubal ligation status	2537

Visualizations:



Significance and Implications:

- **Resource Allocation:** Conditions like acute bronchitis and anemia, which appear frequently, may require dedicated resources such as specific treatment protocols or medication stocks.
- **Health Management Focus:** Identifying prevalent conditions helps in focusing preventive healthcare efforts, such as stress management programs or hypertension screening.

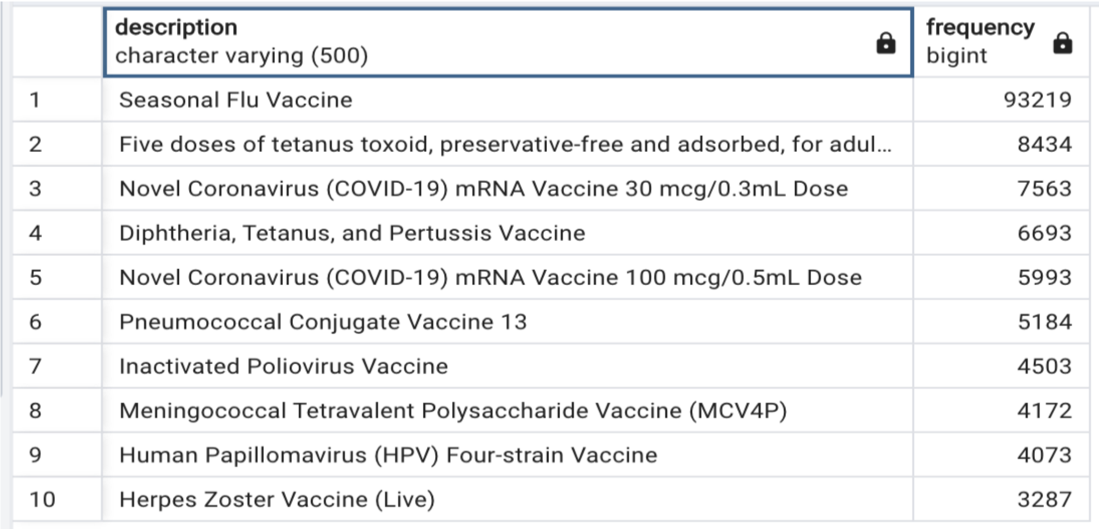
Question 2: What are the most common immunizations administered to patients in the dataset?

Why is this important? Understanding the most frequently administered immunizations provides insights into public health efforts and disease prevention strategies. This information can help in evaluating vaccination coverage, planning future immunization programs, and allocating resources for vaccines.

Query:

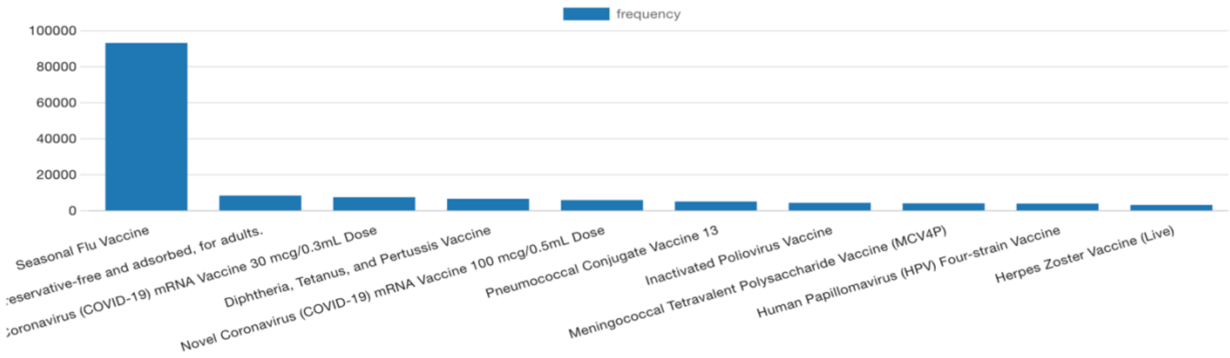
```
SELECT DESCRIPTION, COUNT(*) AS frequency
FROM immunizations
GROUP BY DESCRIPTION
ORDER BY frequency DESC LIMIT 10;
```

Results:



	description character varying (500) 🔒	frequency bigint 🔒
1	Seasonal Flu Vaccine	93219
2	Five doses of tetanus toxoid, preservative-free and adsorbed, for adul...	8434
3	Novel Coronavirus (COVID-19) mRNA Vaccine 30 mcg/0.3mL Dose	7563
4	Diphtheria, Tetanus, and Pertussis Vaccine	6693
5	Novel Coronavirus (COVID-19) mRNA Vaccine 100 mcg/0.5mL Dose	5993
6	Pneumococcal Conjugate Vaccine 13	5184
7	Inactivated Poliovirus Vaccine	4503
8	Meningococcal Tetravalent Polysaccharide Vaccine (MCV4P)	4172
9	Human Papillomavirus (HPV) Four-strain Vaccine	4073
10	Herpes Zoster Vaccine (Live)	3287

Visualization:



Significance and Implications:

- **Vaccination Coverage:** The high numbers of seasonal flu vaccines (93,219) and COVID-19 vaccines (13,556 combined for both doses) demonstrate the success of aggressive public health campaigns in achieving widespread vaccine acceptance and preventing outbreaks.
- **Public Health Priorities:** The administration of vaccines such as the Diphtheria, Tetanus, and Pertussis Vaccine (6,693) and Pneumococcal Conjugate Vaccine 13 (5,184) reflects a strategic focus on reducing childhood and adult morbidity and mortality from severe infections.
- **Resource Allocation:** The significant number of vaccines administered allows healthcare providers to better plan and allocate supplies, ensuring availability and avoiding shortages.

Question 3: What are the most common conditions for different age groups?

Why is this important? This helps in understanding how certain health conditions are distributed across different age groups, which can inform age-specific healthcare strategies.

Query:

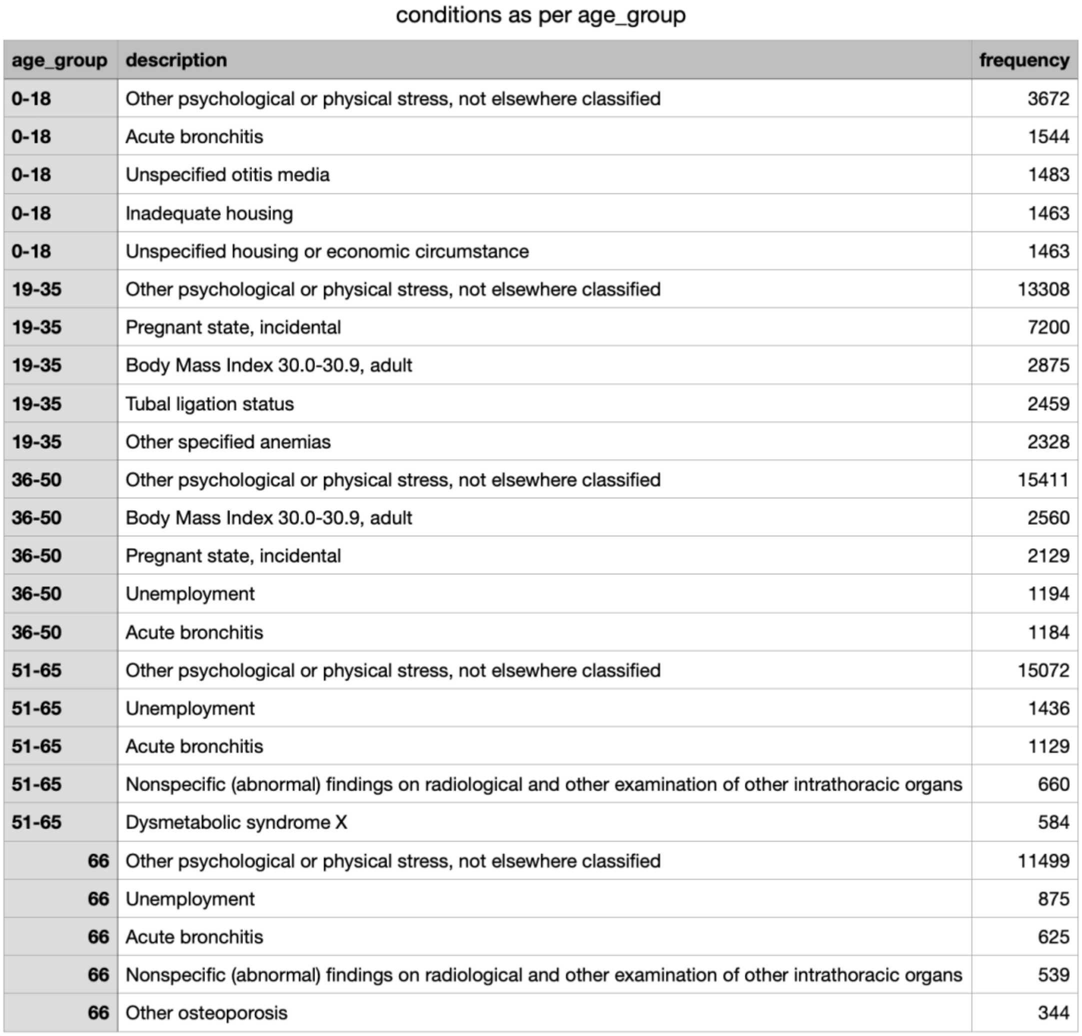
```
-- Add new columns
ALTER TABLE conditions
ADD age_at_condition integer,
ADD age_group varchar(30);

-- Update age at condition
UPDATE conditions AS t1
SET age_at_condition = EXTRACT(YEAR FROM age(t1.start,
t2.birthdate))
FROM patients AS t2
WHERE t2.id = t1.patient;

-- Categorize age groups
UPDATE conditions
SET age_group = CASE
    WHEN age_at_condition <= 18 THEN '0-18'
    WHEN age_at_condition BETWEEN 19 AND 35 THEN '19-
35'
    WHEN age_at_condition BETWEEN 36 AND 50 THEN '36-
50'
    WHEN age_at_condition BETWEEN 51 AND 65 THEN '51-
65'
    ELSE '66+'
END;

-- Create a CTE for ranking and filter top 5 conditions
WITH rank_conditions AS (
    SELECT t2.age_group,
           t2.description,
           COUNT(*) AS frequency,
           ROW_NUMBER() OVER (PARTITION BY t2.age_group
ORDER BY COUNT(*) DESC) AS rank
    FROM conditions AS t2
    JOIN patients AS t1 ON t1.id = t2.patient
    GROUP BY t2.description, t2.age_group
)
SELECT age_group, description, frequency
FROM rank_conditions
WHERE rank <= 5
ORDER BY age_group, rank;
```

Results:



conditions as per age_group

age_group	description	frequency
0-18	Other psychological or physical stress, not elsewhere classified	3672
0-18	Acute bronchitis	1544
0-18	Unspecified otitis media	1483
0-18	Inadequate housing	1463
0-18	Unspecified housing or economic circumstance	1463
19-35	Other psychological or physical stress, not elsewhere classified	13308
19-35	Pregnant state, incidental	7200
19-35	Body Mass Index 30.0-30.9, adult	2875
19-35	Tubal ligation status	2459
19-35	Other specified anemias	2328
36-50	Other psychological or physical stress, not elsewhere classified	15411
36-50	Body Mass Index 30.0-30.9, adult	2560
36-50	Pregnant state, incidental	2129
36-50	Unemployment	1194
36-50	Acute bronchitis	1184
51-65	Other psychological or physical stress, not elsewhere classified	15072
51-65	Unemployment	1436
51-65	Acute bronchitis	1129
51-65	Nonspecific (abnormal) findings on radiological and other examination of other intrathoracic organs	660
51-65	Dysmetabolic syndrome X	584
66	Other psychological or physical stress, not elsewhere classified	11499
66	Unemployment	875
66	Acute bronchitis	625
66	Nonspecific (abnormal) findings on radiological and other examination of other intrathoracic organs	539
66	Other osteoporosis	344

Significance and Implications:

- **"0-18"**: Health conditions tend to be more related to specific issues like otitis media and housing circumstances.
- **"19-35"**: Includes conditions related to pregnancy and body mass index.
- **"36-50"**: Shows an increase in psychological stress and adult obesity conditions.
- **"51-65"**: Significant presence of unemployment and psychological stress, alongside continuing issues like acute bronchitis.
- **"66+"**: Psychological stress remains a top condition, with additional focus on unemployment and osteoporosis.

Question 4: What are the most common health conditions in different geographical regions?

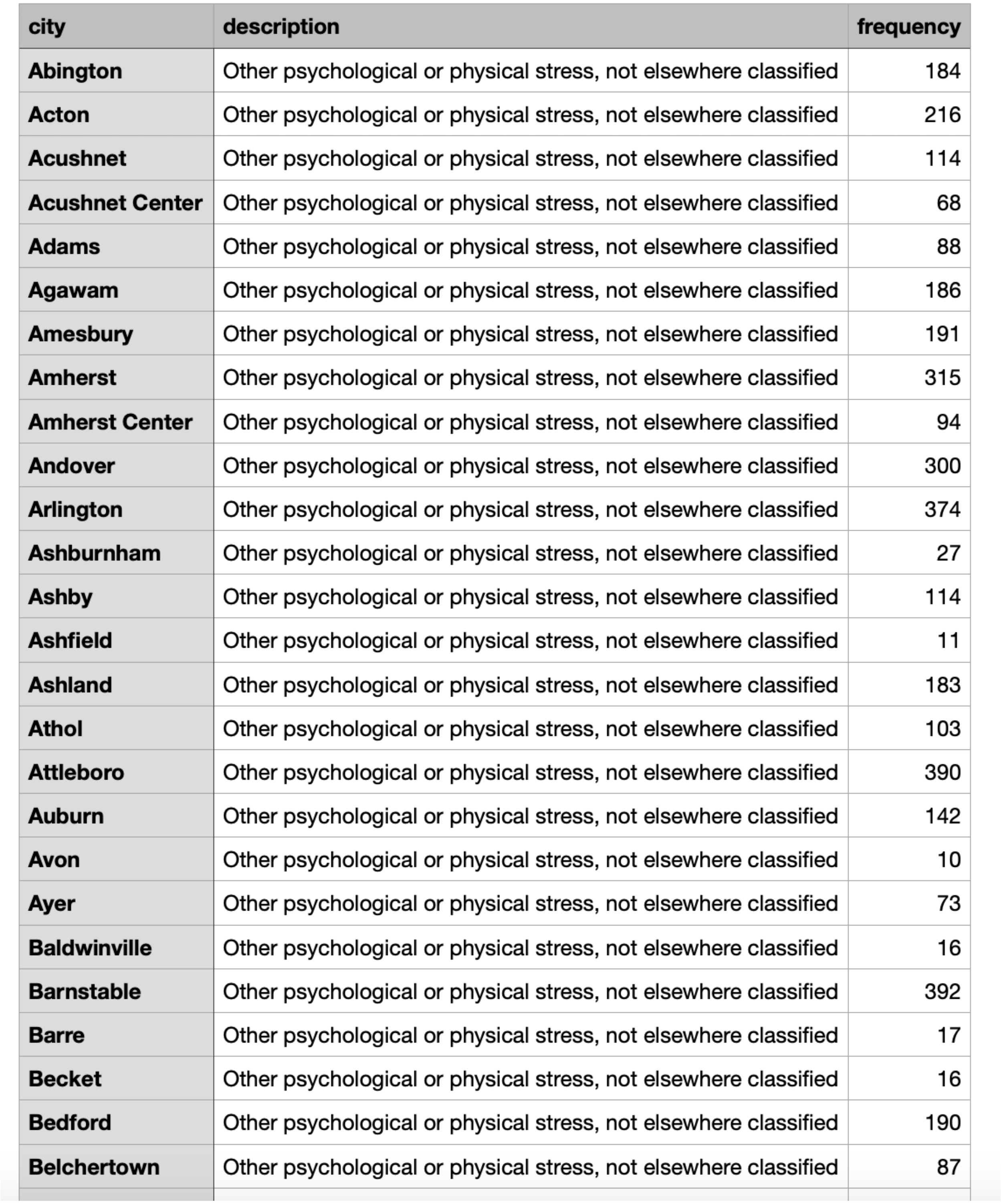
Why is this important? Understanding the geographical distribution of health conditions can help in allocating resources more effectively and developing region-specific healthcare programs.

Query:

```
WITH ranked_condition AS (
    SELECT t1.city, t2.description, COUNT(*) AS
frequency,
           ROW_NUMBER() OVER(PARTITION BY t1.city ORDER
BY COUNT(*) DESC) AS rank
    FROM patients AS t1
    JOIN conditions AS t2 ON t1.id = t2.patient
    GROUP BY t1.city, t2.description
)
SELECT city, description, frequency
FROM ranked_condition
WHERE rank = 1
ORDER BY city;
```

Results:

Snapshot of output:



city	description	frequency
Abington	Other psychological or physical stress, not elsewhere classified	184
Acton	Other psychological or physical stress, not elsewhere classified	216
Acushnet	Other psychological or physical stress, not elsewhere classified	114
Acushnet Center	Other psychological or physical stress, not elsewhere classified	68
Adams	Other psychological or physical stress, not elsewhere classified	88
Agawam	Other psychological or physical stress, not elsewhere classified	186
Amesbury	Other psychological or physical stress, not elsewhere classified	191
Amherst	Other psychological or physical stress, not elsewhere classified	315
Amherst Center	Other psychological or physical stress, not elsewhere classified	94
Andover	Other psychological or physical stress, not elsewhere classified	300
Arlington	Other psychological or physical stress, not elsewhere classified	374
Ashburnham	Other psychological or physical stress, not elsewhere classified	27
Ashby	Other psychological or physical stress, not elsewhere classified	114
Ashfield	Other psychological or physical stress, not elsewhere classified	11
Ashland	Other psychological or physical stress, not elsewhere classified	183
Athol	Other psychological or physical stress, not elsewhere classified	103
Attleboro	Other psychological or physical stress, not elsewhere classified	390
Auburn	Other psychological or physical stress, not elsewhere classified	142
Avon	Other psychological or physical stress, not elsewhere classified	10
Ayer	Other psychological or physical stress, not elsewhere classified	73
Baldwinville	Other psychological or physical stress, not elsewhere classified	16
Barnstable	Other psychological or physical stress, not elsewhere classified	392
Barre	Other psychological or physical stress, not elsewhere classified	17
Becket	Other psychological or physical stress, not elsewhere classified	16
Bedford	Other psychological or physical stress, not elsewhere classified	190
Belchertown	Other psychological or physical stress, not elsewhere classified	87

Significance and Implications:

- **Localized Healthcare Strategies**: The results reveal specific health conditions prevalent in each city, enabling targeted healthcare interventions and resource allocation to address the most common issues.
- **Preventive Measures**: Cities like East Douglas with high rates of Acute bronchitis can benefit from targeted respiratory health programs and air quality improvements.

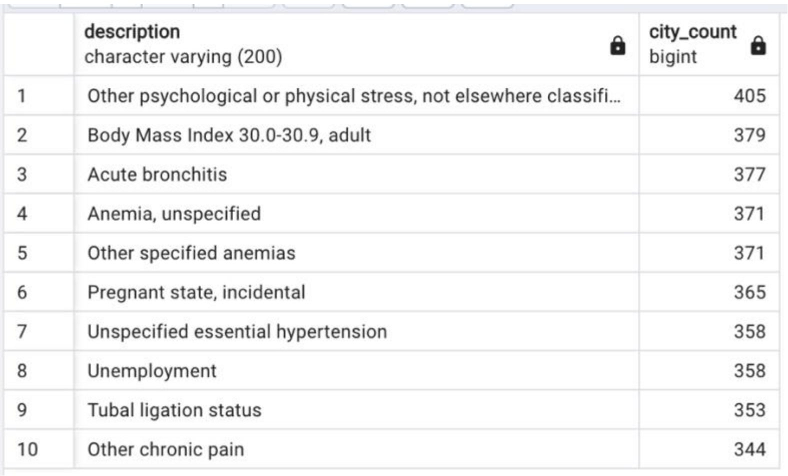
Question 5: Which Health Conditions Are Most Common in Different Cities?

Why This Matters: We've noticed that a lot of cities have 'Other psychological or physical stress, not elsewhere classified' as their top health condition. But we want to dive deeper: How many cities have this as their top condition? And what about other health issues?

Query:

```
SELECT description, COUNT(*) AS city_count
FROM (SELECT
            t1.city,
            t2.description,
            COUNT(*) AS frequency
        FROM patients AS t1
        JOIN conditions AS t2 ON t1.id = t2.patient
        GROUP BY t1.city, t2.description
       ORDER BY t1.city, count(*)
    ) AS top_conditions
GROUP BY description
ORDER BY city_count DESC LIMIT 10;
```

Results:



	description character varying (200)	city_count bigint
1	Other psychological or physical stress, not elsewhere classifi…	405
2	Body Mass Index 30.0-30.9, adult	379
3	Acute bronchitis	377
4	Anemia, unspecified	371
5	Other specified anemias	371
6	Pregnant state, incidental	365
7	Unspecified essential hypertension	358
8	Unemployment	358
9	Tubal ligation status	353
10	Other chronic pain	344

Significance and Implications:

- Mental Health Needs: 'Other psychological or physical stress, not elsewhere classified' is the top health issue in 405 cities, highlighting the urgent need for widespread mental health services.
- Obesity Concerns: 'Body Mass Index 30.0-30.9, adult' ranks highest in 379 cities, pointing to obesity as a significant concern and emphasizing the need for effective nutritional and fitness programs.
- Respiratory Health: 'Acute bronchitis' is the leading condition in 377 cities, suggesting a need for improved respiratory health programs and air quality initiatives.

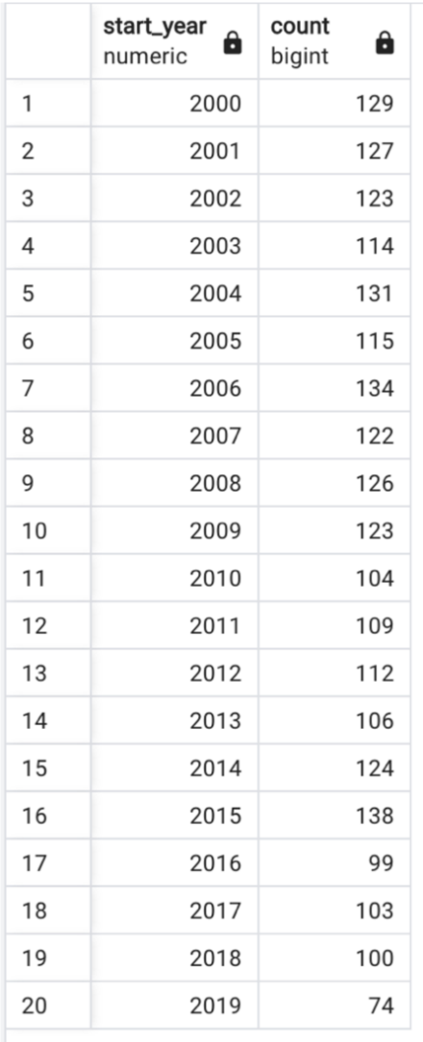
Question 6: How has the frequency of 'Body Mass Index 30.0-30.9, adult' changed over the years (2000 to 2020)?

Why This Matters: Understanding the trend in the frequency of 'Body Mass Index 30.0-30.9, adult' over the years can provide insights into the effectiveness of public health initiatives related to obesity, and help in planning future interventions.

Query:

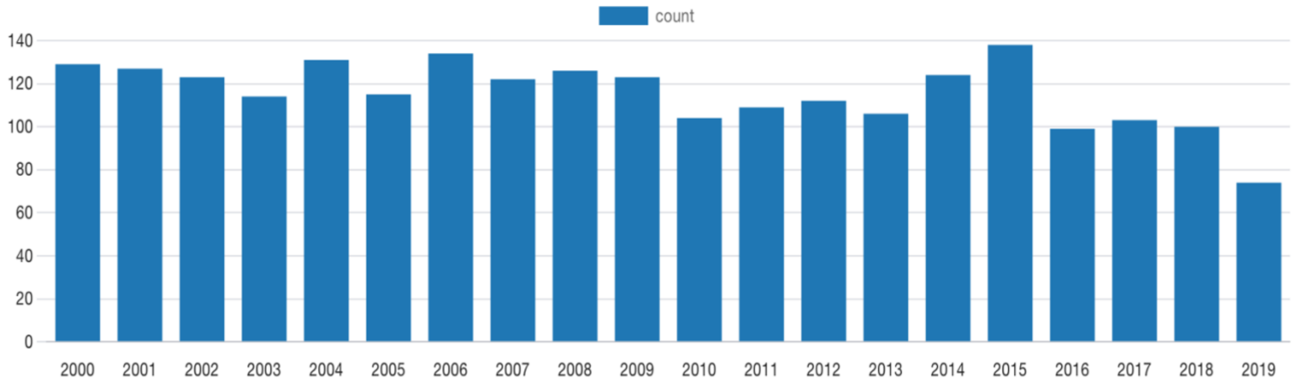
```
SELECT extract(year FROM start) AS start_year, COUNT(*)
FROM conditions AS c
WHERE c.start BETWEEN '2000-01-01' AND '2020-01-01' AND
    c.description = 'Body Mass Index 30.0-30.9, adult'
GROUP BY start_year
ORDER BY start_year;
```

Results:



	start_year numeric 🔒	count bigint 🔒
1	2000	129
2	2001	127
3	2002	123
4	2003	114
5	2004	131
6	2005	115
7	2006	134
8	2007	122
9	2008	126
10	2009	123
11	2010	104
12	2011	109
13	2012	112
14	2013	106
15	2014	124
16	2015	138
17	2016	99
18	2017	103
19	2018	100
20	2019	74

Visualization:



Significance and Implications:

- **Trend Analysis:** The data reveals fluctuations in the frequency of 'Body Mass Index 30.0-30.9, adult' from 2000 to 2019. Notably, there is no clear upward or downward trend over the years, suggesting periodic variations.
- **Peak Years:** The years 2004 and 2015 show relatively higher frequencies (131 and 138 cases, respectively). These peak years may correlate with certain public health events, policy changes, or societal trends affecting obesity rates.

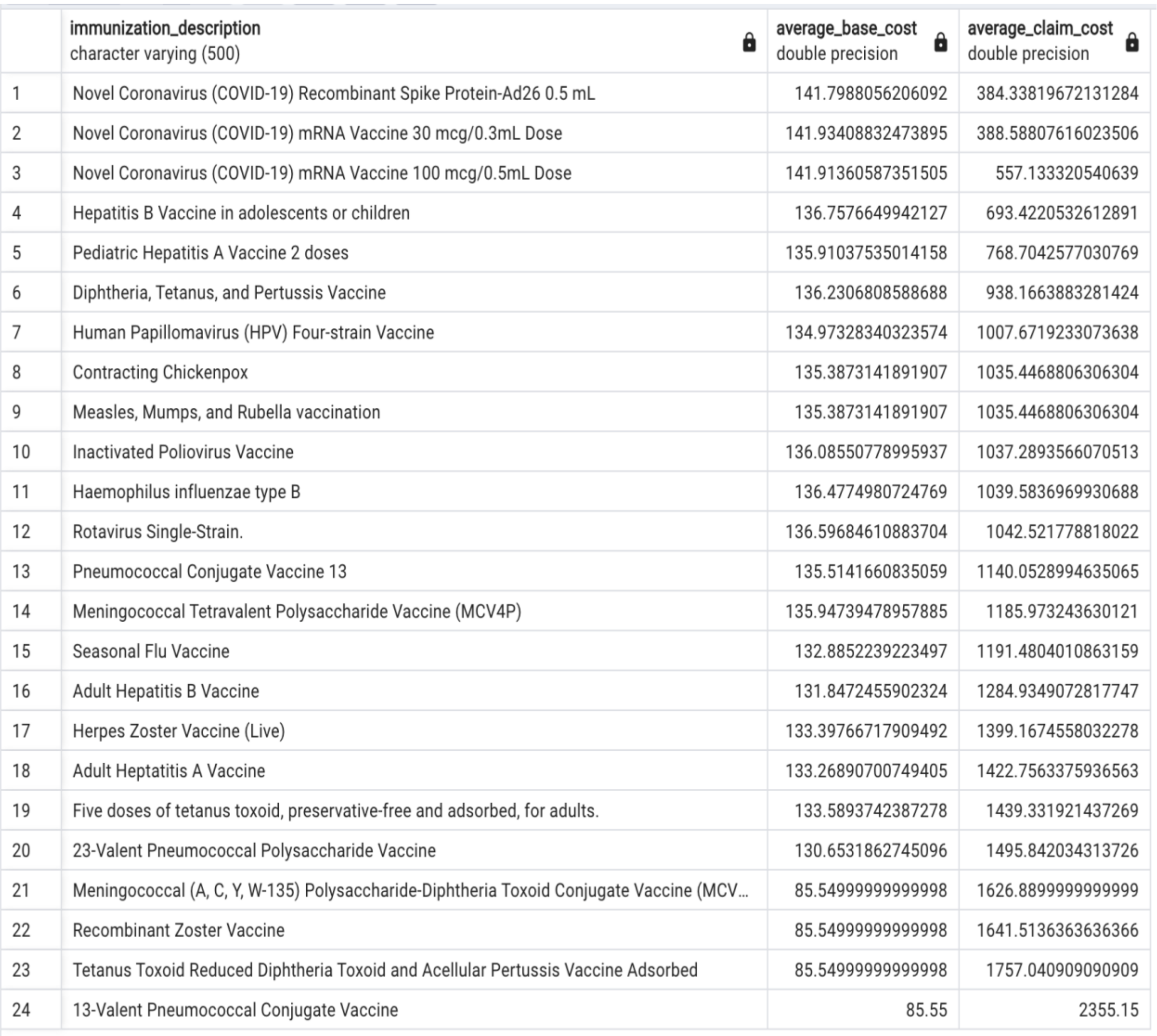
Question 7: What is the cost of different immunizations.

Why This Matters: Tracking the cost of different immunizations helps us understand their financial impact and aids in budget planning. Knowing which vaccines are more expensive can guide smarter decisions about resource allocation and cost management.

Query:

```
SELECT
    t1.description AS Immunization_Description,
    AVG(t2.base_encounter_cost) AS Average_Base_Cost,
    AVG(t2.total_claim_cost) AS Average_Claim_Cost
FROM
    immunizations AS t1
JOIN
    encounters AS t2
ON
    t2.id = t1.encounter
GROUP BY
    t1.description
ORDER BY
    AVG(t2.total_claim_cost)
```

Results:



	immunization_description character varying (500)	average_base_cost double precision	average_claim_cost double precision
1	Novel Coronavirus (COVID-19) Recombinant Spike Protein-Ad26 0.5 mL	141.7988056206092	384.33819672131284
2	Novel Coronavirus (COVID-19) mRNA Vaccine 30 mcg/0.3mL Dose	141.93408832473895	388.58807616023506
3	Novel Coronavirus (COVID-19) mRNA Vaccine 100 mcg/0.5mL Dose	141.91360587351505	557.133320540639
4	Hepatitis B Vaccine in adolescents or children	136.7576649942127	693.4220532612891
5	Pediatric Hepatitis A Vaccine 2 doses	135.91037535014158	768.7042577030769
6	Diphtheria, Tetanus, and Pertussis Vaccine	136.2306808588688	938.1663883281424
7	Human Papillomavirus (HPV) Four-strain Vaccine	134.97328340323574	1007.6719233073638
8	Contracting Chickenpox	135.3873141891907	1035.4468806306304
9	Measles, Mumps, and Rubella vaccination	135.3873141891907	1035.4468806306304
10	Inactivated Poliovirus Vaccine	136.08550778995937	1037.2893566070513
11	Haemophilus influenzae type B	136.4774980724769	1039.5836969930688
12	Rotavirus Single-Strain.	136.59684610883704	1042.521778818022
13	Pneumococcal Conjugate Vaccine 13	135.5141660835059	1140.0528994635065
14	Meningococcal Tetravalent Polysaccharide Vaccine (MCV4P)	135.94739478957885	1185.973243630121
15	Seasonal Flu Vaccine	132.8852239223497	1191.4804010863159
16	Adult Hepatitis B Vaccine	131.8472455902324	1284.9349072817747
17	Herpes Zoster Vaccine (Live)	133.39766717909492	1399.1674558032278
18	Adult Heptatitis A Vaccine	133.26890700749405	1422.7563375936563
19	Five doses of tetanus toxoid, preservative-free and adsorbed, for adults.	133.5893742387278	1439.331921437269
20	23-Valent Pneumococcal Polysaccharide Vaccine	130.6531862745096	1495.842034313726
21	Meningococcal (A, C, Y, W-135) Polysaccharide-Diphtheria Toxoid Conjugate Vaccine (MCV…	85.54999999999998	1626.8899999999999
22	Recombinant Zoster Vaccine	85.54999999999998	1641.5136363636366
23	Tetanus Toxoid Reduced Diphtheria Toxoid and Acellular Pertussis Vaccine Adsorbed	85.54999999999998	1757.040909090909
24	13-Valent Pneumococcal Conjugate Vaccine	85.55	2355.15

Significance and Implications:

- **13-Valent Pneumococcal Conjugate Vaccine** stands out with the highest average claim cost at $2,355.15. This indicates a significant financial impact, potentially due to its importance in preventing severe infections.
- **COVID-19 Vaccines** have a notably lower claim cost compared to other high-impact vaccines, with average costs ranging between $384 and $557. This lower cost could be attributed to mass production and government subsidies, ensuring accessibility during the pandemic.

Question 8: Finding out average length of stay as per age groups

Why this matters: Understanding the average length of hospital stays by age groups can help healthcare providers allocate resources more efficiently, tailor patient care strategies, and improve overall hospital management.

Query:

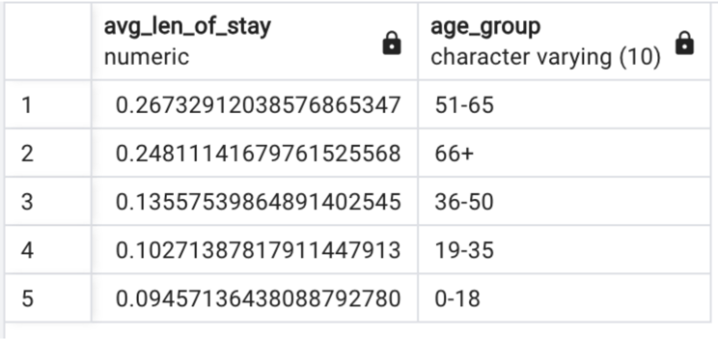
```
-- Alter table to add columns
ALTER TABLE encounters
ADD COLUMN age_at_encounter INT,
ADD COLUMN age_group VARCHAR(10);

-- Update age_at_encounter based on patient's birthdate
UPDATE encounters
SET age_at_encounter = EXTRACT(YEAR FROM
AGE(encounters.start, patients.birthdate))
FROM patients
WHERE encounters.patient = patients.id;

-- Update age_group based on age_at_encounter
UPDATE encounters
SET age_group = CASE
    WHEN age_at_encounter < 19 THEN '0-18'
    WHEN age_at_encounter BETWEEN 19 AND 35 THEN '19-
35'
    WHEN age_at_encounter BETWEEN 36 AND 50 THEN '36-
50'
    WHEN age_at_encounter BETWEEN 51 AND 65 THEN '51-
65'
    ELSE '66+'
END;

-- Calculate average length of stay by age_group
SELECT AVG(EXTRACT(EPOCH FROM (stop - start)) / 86400)
AS Avg_len_of_stay, age_group
FROM encounters
GROUP BY age_group
ORDER BY Avg_len_of_stay DESC;
```

Results:



	avg_len_of_stay numeric	age_group character varying (10)
1	0.26732912038576865347	51-65
2	0.24811141679761525568	66+
3	0.13557539864891402545	36-50
4	0.10271387817911447913	19-35
5	0.09457136438088792780	0-18

Significance and Implications:

The data reveals fascinating trends in hospital stays across different age groups:

- **Ages 51-65:** Patients in this age range have the longest average stays at 0.27 days, reflecting the increased medical attention typically required by this group.
- **Ages 66+:** Elderly patients follow closely with an average stay of 0.25 days, highlighting their need for more intensive care.
- **Ages 36-50:** This group shows a significant drop, with an average stay of 0.14 days, suggesting fewer severe health issues.
- **Ages 19-35:** Young adults have shorter stays at 0.10 days, indicative of quicker recovery times.
- **Ages 0-18:** Children and adolescents have the briefest stays, averaging 0.09 days, showcasing their resilience and faster recovery.

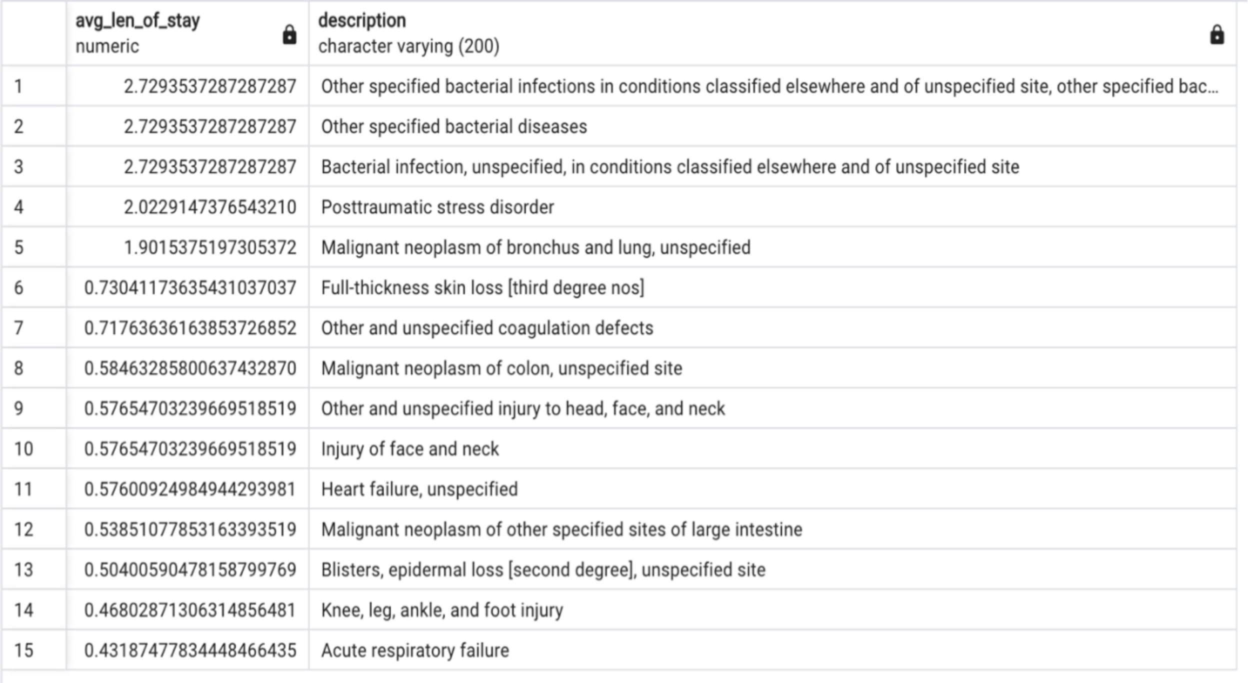
9. Average hospital stay as per different conditions.

Why this matters: Understanding the average length of hospital stays for various conditions helps in resource planning and highlights areas where treatment protocols could be improved.

Query:

```
SELECT AVG(EXTRACT(EPOCH FROM (e.stop - e.start)) /
86400) AS avg_len_of_stay, c.description
FROM encounters AS e
JOIN conditions AS c ON e.patient = c.patient
GROUP BY c.description
ORDER BY avg_len_of_stay DESC LIMIT 15;
```

Results:



	avg_len_of_stay numeric	description character varying (200)
1	2.7293537287287287	Other specified bacterial infections in conditions classified elsewhere and of unspecified site, other specified bac…
2	2.7293537287287287	Other specified bacterial diseases
3	2.7293537287287287	Bacterial infection, unspecified, in conditions classified elsewhere and of unspecified site
4	2.0229147376543210	Posttraumatic stress disorder
5	1.9015375197305372	Malignant neoplasm of bronchus and lung, unspecified
6	0.73041173635431037037	Full-thickness skin loss [third degree nos]
7	0.71763636163853726852	Other and unspecified coagulation defects
8	0.58463285800637432870	Malignant neoplasm of colon, unspecified site
9	0.57654703239669518519	Other and unspecified injury to head, face, and neck
10	0.57654703239669518519	Injury of face and neck
11	0.57600924984944293981	Heart failure, unspecified
12	0.53851077853163393519	Malignant neoplasm of other specified sites of large intestine
13	0.50400590478158799769	Blisters, epidermal loss [second degree], unspecified site
14	0.46802871306314856481	Knee, leg, ankle, and foot injury
15	0.43187477834448466435	Acute respiratory failure

Significance and Implications:

- **Extended Stays for Bacterial Infections**

Conditions like "Other specified bacterial infections" and "Bacterial infection, unspecified" have the longest average stays of around 2.7 days, indicating complex, long-term treatment needs.

- **Mental Health Care Demands More Time**

"Posttraumatic stress disorder" requires an average hospital stay of over 2 days, reflecting the intensive care and support needed for effective mental health treatment.

References:

1. Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, Scott McLachlan, Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record, Journal of the American Medical Informatics Association, Volume 25, Issue 3, March 2018, Pages 230–238, https://doi.org/10.1093/jamia/ocx079