Clinical Data Analysis using SQL

In today's data-driven healthcare landscape, analyzing clinical datasets is crucial for enhancing patient care, optimizing operations, and reducing costs. This project uses SQL in performing comprehensive data analysis on a clinical dataset of Massachusetts female patients, including conditions, encounters, immunizations, and patients. Leveraging SQL, we can extract meaningful insights that help healthcare providers make informed decisions, streamline processes, and improve service quality.

Analyzing clinical data provides hospitals with a deeper understanding of patient demographics, disease prevalence, treatment outcomes, and healthcare costs. These insights are essential for identifying trends, predicting future healthcare needs, and developing strategies to enhance patient care. For example, understanding common conditions and their associated costs enables hospitals to allocate resources effectively and implement preventive measures. Additionally, analyzing immunization data helps track vaccination rates and identify gaps in coverage, crucial for preventing disease outbreaks.

The dataset consists of four main tables: conditions (recording condition details and dates), encounters (capturing patient visit information, costs, and reasons), immunizations (logging vaccination dates and details), and patients (containing demographic information). By querying this dataset, we can uncover valuable insights into patient health trends, healthcare delivery efficiency, and areas for improvement, demonstrating the potential of SQL to transform raw clinical data into actionable intelligence for better healthcare outcomes and operational excellence.

Data source: Data Wizardry website.

For the analysis I will be answering the following questions. These questions can guide a comprehensive analysis of the clinical data to uncover valuable insights about patient health, healthcare utilization, costs, and outcomes.

Common conditions, encounter classes, and wait times

1. **What is the most common condition in the patient population?**
   The most frequently reported condition was **psychological or physical stress**, primarily occurring in the **ambulatory encounter class**.

```
SELECT
    COUNT(patient),
    description
FROM
    conditions_staging
GROUP BY
    description
ORDER BY
    COUNT(patient) Desc
LIMIT 5;
```

	count bigint	description character varying (200)
1	58962	Other psychological or physical stress, not elsewhere classifi...
2	9872	Pregnant state, incidental
3	5684	Acute bronchitis
4	5461	Body Mass Index 30.0-30.9, adult
5	4717	Unemployment

💡 This indicates a need for mental health resources and stress management programs, particularly in outpatient settings.

2. **What is the most common encounter class?**
   There are 10 encounter classes, and the highest number of encounters were recorded for the ambulatory encounter.

```
SELECT
    COUNT(*) AS encounter_count,
    encounterclass
FROM
    encounters_staging
GROUP BY
    encounterclass
ORDER BY
    encounter_count DESC;
```

	count bigint	encounterclass character varying (100)
1	244148	ambulatory
2	85849	outpatient
3	77849	wellness
4	21479	urgentcare
5	17903	emergency
6	3709	inpatient
7	2318	home
8	1221	snf
9	750	hospice
10	709	virtual

3. **Average time spent at the hospital for each encounter class**

```
SELECT
    ROUND(AVG(EXTRACT(EPOCH FROM (stop - start)) / 60),1) AS average_time_spent_minutes,
    ROUND(AVG(EXTRACT(EPOCH FROM (stop - start)) / 3600),1) AS average_time_spent_hours,
    encounterclass
FROM
    encounters_staging
GROUP BY encounterclass
ORDER BY average_time_spent_hours DESC;
```

	average_time_spent_minutes numeric	average_time_spent_hours numeric	encounterclass character varying (100)
1	31304.4	521.7	hospice
2	28740.9	479.0	snf
3	6943.9	115.7	inpatient
4	173.0	2.9	emergency
5	85.6	1.4	ambulatory
6	35.1	0.6	wellness
7	36.4	0.6	outpatient
8	35.0	0.6	urgentcare
9	24.8	0.4	virtual
10	15.0	0.3	home

4. **Find the percentages of patients for each encounter class who spend time higher than the average time.**

Based on this analysis we can see that more than **53%** of people who visit **urgent care spend more than the average recorded time** for urgent care patients.

💡 Operational efficiencies in urgent care settings can be improved to reduce patient wait times and enhance service delivery.

```
WITH avg_time_spent AS (
    SELECT encounterclass,
           ROUND(AVG(EXTRACT(EPOCH FROM (stop - start)) / 3600), 1) AS avg_time_spent_hours,
           COUNT(patient) AS total_patients
    FROM encounters_staging
    GROUP BY encounterclass
),
patient_time_spent AS (
    SELECT patient,
           encounterclass,
    EXTRACT(EPOCH FROM (stop - start)) / 3600 AS time_spent_hours
    FROM encounters_staging
),
patients_above_avg AS (
    SELECT pts.encounterclass,
           COUNT(pts.patient) AS count_of_patients_above_avg
    FROM patient_time_spent pts
    JOIN avg_time_spent ats
    ON pts.encounterclass = ats.encounterclass
    WHERE pts.time_spent_hours > ats.avg_time_spent_hours
    GROUP BY pts.encounterclass
)
SELECT
    p.encounterclass,
    ats.total_patients,
    p.count_of_patients_above_avg,
    ROUND((p.count_of_patients_above_avg::numeric / ats.total_patients) * 100, 2) AS percentage_above_avg
FROM patients_above_avg p
JOIN avg_time_spent ats
ON p.encounterclass = ats.encounterclass
ORDER BY percentage_above_avg DESC;
```

	encounterclass character varying (100)	total_patients bigint	count_of_patients_above_avg bigint	percentage_above_avg numeric
1	urgentcare	21479	11468	53.39
2	wellness	77849	40934	52.58
3	hospice	750	357	47.60
4	ambulatory	244148	96766	39.63
5	snf	1221	446	36.53
6	inpatient	3709	1303	35.13
7	virtual	709	242	34.13
8	outpatient	85849	21816	25.41
9	emergency	17903	1131	6.32

5. What is the most common reason code for encounters?

It was noticed that the reason code was missing for most of the encounters, and we see here that the reason code 585.4 associated with ambulatory encounter class has the highest encounters recorded.
(Using CTE)

```
WITH encounter_counts AS (
    SELECT encounterclass, reasoncode, COUNT(*) AS cnt
    FROM encounters_staging
    GROUP BY reasoncode, encounterclass
)
SELECT encounterclass, reasoncode, cnt
FROM encounter_counts
WHERE (encounterclass, cnt) IN (
    SELECT encounterclass, MAX(cnt)
    FROM encounter_counts
    GROUP BY encounterclass
);
```

encounterclass character varying (100)	reasoncode character varying (100)	cnt bigint
ambulatory	585.4	75822
emergency	[null]	11474
home	[null]	2318
hospice	[null]	293
inpatient	[null]	2215
outpatient	[null]	84228
snf	[null]	1221
urgentcare	[null]	21479
virtual	[null]	454
wellness	[null]	77622

6. What is the average duration of different conditions from start date to stop date?

The condition with the highest average duration was **'Septic Shock',** lasting approximately **36 years on average**.

```
SELECT
    avg(AGE(stop, start)) AS duration_hours,
    code,
    description
FROM conditions_staging
GROUP BY
    code,
    description
HAVING avg(AGE(stop, start)) IS NOT NULL
ORDER BY duration_hours DESC;
```

	duration_hours interval	code character varying (1000)	description character varying (200)
1	35 years 7 mons	785.52	Septic shock
2	19 years 3 mons 38 days 17:19:05.493818	585.1	Chronic kidney disease, Stage I
3	15 years 9 mons 37 days 09:09:53.391402	585.2	Chronic kidney disease, Stage II (mild)
4	14 years 8 mons 23 days 02:21:38.383318	691.8	Other atopic dermatitis and related conditions
5	10 years 5 mons 27 days 08:27:16.340073	585.3	Chronic kidney disease, Stage III (moderate)
6	8 years 3 mons 7 days	309.81	Posttraumatic stress disorder
7	4 years 4 mons 45 days 02:11:39.391543	V49.83	Awaiting organ transplant status
8	4 years 3 mons 23 days 01:21:30.58234	585.4	Chronic kidney disease, Stage IV (severe)
9	3 years 10 mons 23 days 10:39:59.9712	714	Rheumatoid arthritis
10	3 years 6 mons 42 days 18:57:37.110588	305.9	Other, mixed, or unspecified drug abuse, unspecified

💡 Long-term care strategies and monitoring programs are essential for managing chronic conditions like septic shock.

Conditions by demographics

7. How do the conditions vary across different patient demographics (e.g., age, gender)?

Age

```
WITH ranked_conditions AS (
    SELECT
        age_category,
        conditions_staging.description,
        COUNT(*) AS counts,
        ROW_NUMBER() OVER (PARTITION BY age_category ORDER BY COUNT(*) DESC) AS rank
    FROM (
        SELECT
            id,
            CASE
                WHEN DATE_PART('YEAR', AGE(current_date, birthdate)) <= 12 THEN 'Child'
                WHEN DATE_PART('YEAR', AGE(current_date, birthdate)) BETWEEN 13 AND 19 THEN 'Teen'
                WHEN DATE_PART('YEAR', AGE(current_date, birthdate)) BETWEEN 20 AND 35 THEN 'Young-Adult'
                WHEN DATE_PART('YEAR', AGE(current_date, birthdate)) BETWEEN 36 AND 55 THEN 'Middle-Aged'
                WHEN DATE_PART('YEAR', AGE(current_date, birthdate)) BETWEEN 56 AND 75 THEN 'Older-Adult'
                ELSE 'Senior'
            END AS age_category
        FROM
            patients_staging
    ) AS age_categories
    JOIN
        conditions_staging
    ON
        age_categories.id = conditions_staging.patient
    GROUP BY
        age_category,
        conditions_staging.description
)
SELECT
    age_category,
    description,
    counts
FROM
    ranked_conditions
WHERE
    rank <= 2
ORDER BY
    age_category,
    counts DESC;
```

	age_category text	description character varying (200)	counts bigint
1	Child	Unspecified otitis media	1047
2	Child	Acute bronchitis	442
3	Middle-Aged	Other psychological or physical stress, not elsewhere classifi...	12258
4	Middle-Aged	Pregnant state, incidental	4335
5	Older-Adult	Other psychological or physical stress, not elsewhere classifi...	23777
6	Older-Adult	Body Mass Index 30.0-30.9, adult	2239
7	Senior	Other psychological or physical stress, not elsewhere classifi...	18872
8	Senior	Body Mass Index 30.0-30.9, adult	1007
9	Teen	Acute bronchitis	548
10	Teen	Streptococcal sore throat	303
11	Young-Adult	Pregnant state, incidental	4629
12	Young-Adult	Other psychological or physical stress, not elsewhere classifi...	3999

💡This shows that tailored healthcare programs can be developed for different age groups to address their specific health issues effectively.

Ethnic Groups:

- The data set includes only two ethnic groups, **Hispanic and non-Hispanic**. Both groups most suffered from "**Other psychological or physical stress, not elsewhere classified**".

```
SELECT
    patients_staging.ethnicity,
    conditions_staging.description,
    count(*) as counts
FROM patients_staging
JOIN
    conditions_staging
ON
    patients_staging.id = conditions_staging.patient
Group by patients_staging.ethnicity,
    conditions_staging.description
order by patients_staging.ethnicity, counts desc;
```

- Regardless of ethnicity, the highest counts were recorded for the condition "Other psychological or physical stress, not elsewhere classified

```
WITH ranked_conditions AS (
    SELECT
        patients_staging.race,
        conditions_staging.description,
        COUNT(*) AS counts,
        ROW_NUMBER() OVER (PARTITION BY patients_staging.race ORDER BY COUNT(*) DESC) AS rank
    FROM
        patients_staging
    JOIN
        conditions_staging
    ON
        patients_staging.id = conditions_staging.patient
    GROUP BY
        patients_staging.race,
        conditions_staging.description
)
SELECT
    race,
    description,
    counts
FROM
    ranked_conditions
WHERE
    rank <= 2
ORDER BY
    race,
    counts DESC;
```

	race character varying (100)	description character varying (200)	counts bigint
1	asian	Other psychological or physical stress, not elsewhere classifi…	3612
2	asian	Pregnant state, incidental	714
3	black	Other psychological or physical stress, not elsewhere classifi…	6412
4	black	Pregnant state, incidental	872
5	hawaiian	Other psychological or physical stress, not elsewhere classifi…	1049
6	hawaiian	Pregnant state, incidental	128
7	native	Other psychological or physical stress, not elsewhere classifi…	245
8	native	Pregnant state, incidental	36
9	other	Other psychological or physical stress, not elsewhere classifi…	471
10	other	Pregnant state, incidental	93
11	white	Other psychological or physical stress, not elsewhere classifi…	47173
12	white	Pregnant state, incidental	8029

Location: For most counties, most patients suffered from "**Other psychological or physical stress, not elsewhere classified**".

```
SELECT
    patients_staging.county,
    conditions_staging.description,
    count(*) as counts
FROM patients_staging
JOIN
    conditions_staging
ON
    patients_staging.id = conditions_staging.patient
Group by patients_staging.county,
    conditions_staging.description
order by counts desc;
```

💡To better address these frequently occurring conditions, it is important to develop health interventions prioritizing mental health services and stress management initiatives.

Cost Analysis

8. What are the average base encounter costs and total claim costs?

```
SELECT
    ROUND(AVG(base_encounter_cost)::numeric, 2) AS avg_base_encounter_cost,
    ROUND(AVG(total_claim_cost)::numeric, 2) AS avg_total_claim_cost,
    ROUND(AVG(payer_coverage)::numeric, 2) AS avg_payer_coverage
FROM
    encounters_staging;
```

	avg_base_encounter_cost numeric	avg_total_claim_cost numeric	avg_payer_coverage numeric
1	117.03	3606.52	2813.75

💡Financial planning and budgeting can be optimized by understanding the cost structures and identifying areas to reduce unnecessary expenditures.

9. How do costs vary by reason code?

For each reason code Avg base cost varies between 84. 75 to 146.18 and avg total claim cost vary between 84.75 to 16317.57. Below is the top 10 highest costs and their reason codes.

```
SELECT
    reasoncode,
    AVG(base_encounter_cost) as avg_base_cost,
    AVG(total_claim_cost) as avg_total_claim_cost
FROM
    encounters_staging
GROUP BY
    reasoncode
ORDER BY
    avg_base_cost DESC,
    avg_total_claim_cost DESC;
```

	reasoncode character varying (100)	avg_base_cost double precision	avg_total_claim_cost double precision
1	995.3	146.1800000000002	2692.1559322033904
2	959.09	146.1800000000002	146.1800000000002
3	959.7	146.18000000000018	6374.722392857141
4	V49.83	146.18000000000018	4153.58255319149
5	V45.81	146.18000000000006	16317.571475054247
6	411.1	146.18	23355.31
7	575	146.17999999999998	12042.216666666665
8	434.91	144.87722222222226	19057.064722222225
9	633.1	144.37320754717	3600.1353962264166
10	850.9	143.6197245179041	143.82517906335866

10. **How do encounter costs differ by encounter class?**
    The highest average base cost was recorded for emergency class and highest average total claim cost was recorded for snf (skilled nursing facility).

```
SELECT
    COUNT(*) AS encounter_count,
    encounterclass,
    AVG(base_encounter_cost) AS avg_base_cost,
    AVG(total_claim_cost) AS avg_total_claim_cost
FROM
    encounters_staging
GROUP BY
    encounterclass
ORDER BY
    avg_base_cost DESC,
    avg_total_claim_cost DESC,
    encounter_count DESC;
```

	encounter_count bigint	encounterclass character varying (100)	avg_base_cost double precision	avg_total_claim_cost double precision
1	17903	emergency	144.77238284086246	3876.8564486398427
2	21479	urgentcare	142.57999999998663	1034.4758773685935
3	750	hospice	137.52999999999955	11124.481733333323
4	77849	wellness	136.7999999999432	1124.0083358809864
5	3709	inpatient	134.9715421946582	17915.35404152058
6	2318	home	128.53000000000242	692.5313028472748
7	85849	outpatient	123.71636874048036	1986.2282548436933
8	1221	snf	110.91999999999895	15394.501990171957
9	244148	ambulatory	103.72710368296994	4910.16058677515
10	709	virtual	97.8490832157969	981.3493511988783

11. **How do total claim costs vary across different payers / Which providers have the highest average total claim costs, and what conditions are they treating?**

```
SELECT
    ROUND(AVG(total_claim_cost)::numeric, 2) AS average_total_claim_cost,
    provider
FROM
    encounters_staging
GROUP BY
    provider
ORDER BY
    average_total_claim_cost DESC
LIMIT
    10;
```

	average_total_claim_cost numeric	provider character varying (100)
1	67866.32	4c33f3d8-c263-3c74-ad19-f55f34c2976c
2	63403.77	08a1fccf-31ab-3923-bc3f-8ea6eb3fb2dc
3	54266.09	097c46f7-27c7-3159-b5dc-1d80644543...
4	48556.08	7a927129-c0c6-300a-89e2-e6e93e9d84...
5	47854.07	a28669b2-f216-3b42-921b-9bdc26a9b5...
6	43259.79	45087313-0a63-3b4c-8370-c01b1944b...
7	43243.27	b64bb38c-ee4f-3c7a-833e-1b895002f8...
8	38587.78	5bdd543c-880c-36a4-a1b3-61451ac18...
9	38444.97	d03d6813-7926-3b5c-8728-7bdb9f62ce...
10	38006.68	b0f18e4f-f26e-3a90-8cc4-4bd07ca012c0

12. **For what conditions do people pay the highest total claim cost ?**
    - The condition with the highest total claim cost was "**Malignant neoplasm of bronchus and lung, unspecified**".
    - The second highest was "**High risk pregnancy**".

💡High-cost conditions should be the focus of preventive care and efficient management protocols to control healthcare expenses.

```sql
WITH condition_encounter AS (
    SELECT
        e.patient,
        c.description,
        e.payer,
        e.base_encounter_cost,
        e.total_claim_cost
    FROM
        conditions_staging c
    JOIN
        encounters_staging e
        ON c.patient = e.patient
)
SELECT
    description,
    ROUND(AVG(base_encounter_cost)::numeric, 2) AS avg_base_encounter_cost,
    ROUND(AVG(total_claim_cost)::numeric, 2) AS avg_total_claim_cost
FROM
    condition_encounter
GROUP BY
    description
ORDER BY
    avg_total_claim_cost DESC;
```

	description character varying (200)	avg_base_encounter_cost numeric	avg_total_claim_cost numeric
1	Malignant neoplasm of bronchus and lung, unspecified	101.76	19515.59
2	Supervision of unspecified high-risk pregnancy	137.80	10051.84
3	Other and unspecified coagulation defects	113.24	8987.82
4	Contact dermatitis and other eczema due to other specified agents	135.25	8878.94
5	Contact dermatitis and other eczema, unspecified cause	135.25	8878.94
6	Eclampsia, antepartum condition or complication	130.48	7869.63
7	Eclampsia, unspecified as to episode of care or not applicable	130.48	7869.63
8	Mild or unspecified pre-eclampsia, postpartum condition or complication	128.65	7213.24
9	Mild or unspecified pre-eclampsia, antepartum condition or complication	128.65	7213.24
10	Mild or unspecified pre-eclampsia, delivered, with mention of postpartum complication	128.65	7213.24

Provider and Payer Analysis

13. Which providers are associated with the highest number of encounters?

For each encounter class, the provider with the highest number of encounters was identified.

```
SELECT DISTINCT ON (encounterclass)
    encounterclass,
    provider,
    encounter_count
FROM (
    SELECT
        encounterclass,
        provider,
        COUNT(*) AS encounter_count
    FROM
        encounters_staging
    GROUP BY
        encounterclass,
        provider
    ORDER BY
        encounterclass,
        encounter_count DESC
) subquery
ORDER BY
    encounterclass,
    encounter_count DESC;
```

	encounterclass character varying (100)	provider character varying (100)	encounter_count bigint
1	ambulatory	4e98e792-2919-3258-9159-025edd33f9...	10141
2	emergency	dfae882b-944d-3245-b0d4-6472175409...	637
3	home	b6914532-b308-391c-9571-a1c6ef5b74...	124
4	hospice	508f4f73-67c1-3fed-a887-aa8985042450	41
5	inpatient	cca9f38e-e799-3c66-b6ba-c6796e6fa9e9	924
6	outpatient	4e98e792-2919-3258-9159-025edd33f9...	3216
7	snf	84b39cdc-f860-3d3a-afbe-967a330bf2c7	19
8	urgentcare	9804de49-b5de-3082-83e6-41acc2c2fef7	2494
9	virtual	cca9f38e-e799-3c66-b6ba-c6796e6fa9e9	143
10	wellness	22c34434-cef7-3e4f-b183-5eec594f3391	1175

The DISTINCT ON clause in PostgreSQL is used to return the first row of each set of rows where the specified column or columns have duplicate values. When using DISTINCT ON, you should also use ORDER BY to control which row of each set is returned. DISTINCT ON (encounterclass) selects only the first row for each unique encounterclass from the result set of the subquery.

💡 This can be helpful in resource allocation and partnership strategies to enhance patient care.

High Total Claim Costs:

14. How do total claim costs vary across different payers / Which providers have the highest average total claim costs, and what conditions are they treating?

```
SELECT
    ROUND(AVG(total_claim_cost)::numeric, 2) AS average_total_claim_cost,
    provider
FROM
    encounters_staging
GROUP BY
    provider
ORDER BY
    average_total_claim_cost DESC
LIMIT
    10;
```

	average_total_claim_cost numeric	provider character varying (100)
1	67866.32	4c33f3d8-c263-3c74-ad19-f55f34c2976c
2	63403.77	08a1fccf-31ab-3923-bc3f-8ea6eb3fb2dc
3	54266.09	097c46f7-27c7-3159-b5dc-1d80644543...
4	48556.08	7a927129-c0c6-300a-89e2-e6e93e9d84...
5	47854.07	a28669b2-f216-3b42-921b-9bdc26a9b5...
6	43259.79	45087313-0a63-3b4c-8370-c01b1944b...
7	43243.27	b64bb38c-ee4f-3c7a-833e-1b895002f8...
8	38587.78	5bdd543c-880c-36a4-a1b3-61451ac18...
9	38444.97	d03d6813-7926-3b5c-8728-7bdb9f62ce...
10	38006.68	b0f18e4f-f26e-3a90-8cc4-4bd07ca012c0

15. What provider has the highest payer coverage?

Highest Payer Coverage: The provider "**4c33f3d8-c263-3c74-ad19-f55f34c2976c**" paid the highest average payer coverage, amounting to **$67,866.32.**

```
SELECT
    provider,
    ROUND(AVG(payer_coverage)::numeric, 2) AS avg_payer_coverage
FROM
    encounters_staging
GROUP BY
    provider
ORDER BY
    avg_payer_coverage DESC;
```

	provider character varying (100)	avg_payer_coverage numeric
1	4c33f3d8-c263-3c74-ad19-f55f34c2976c	67866.32
2	08a1fccf-31ab-3923-bc3f-8ea6eb3fb2dc	50723.02
3	a28669b2-f216-3b42-921b-9bdc26a9b5b7	47804.07
4	097c46f7-27c7-3159-b5dc-1d8064454380	43412.88
5	7a927129-c0c6-300a-89e2-e6e93e9d841f	38751.67
6	45087313-0a63-3b4c-8370-c01b1944b9...	34607.84
7	b64bb38c-ee4f-3c7a-833e-1b895002f837	34594.62
8	7644d19d-6133-35ae-ab5a-da388c595bf2	32196.13
9	66c305cb-0440-3fde-b10d-ba52a01f26fa	32166.20
10	dbaaf106-195a-3546-a277-922734922510	31298.19

💡 Strategies can be developed to negotiate and manage payer contracts more effectively.

Immunization records

16. What are the 10 most common immunization types

- **Common Vaccines in 2023**: The most common vaccine type was **the seasonal flu vaccine**, followed by **"Five doses of tetanus toxoid, preservative-free and adsorbed, for adults."**
- Vaccine distribution and public health campaigns can be better planned to address the most demanded vaccines.

```
SELECT
    COUNT(*) AS counts,
    description
FROM
    immunizations_staging
GROUP BY
    description
ORDER BY
    counts DESC;
```

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

17. What are the immunization rates for different vaccines in last year?

```
SELECT
    EXTRACT(YEAR FROM date) AS year,
    description,
    COUNT(*) AS count
FROM
    immunizations_staging
WHERE
    EXTRACT(YEAR FROM date) = 2023
GROUP BY
    description, year
ORDER BY
    count(*) desc, description;
```

	year numeric	description character varying (500)	count bigint
1	2023	Seasonal Flu Vaccine	3918
2	2023	Five doses of tetanus toxoid, preservative-free and adsorbed, for adul…	342
3	2023	Diphtheria, Tetanus, and Pertussis Vaccine	243
4	2023	Pneumococcal Conjugate Vaccine 13	186
5	2023	Meningococcal Tetravalent Polysaccharide Vaccine (MCV4P)	165
6	2023	Human Papillomavirus (HPV) Four-strain Vaccine	152
7	2023	Inactivated Poliovirus Vaccine	152
8	2023	Herpes Zoster Vaccine (Live)	134
9	2023	Hepatitis B Vaccine in adolescents or children	130
10	2023	Adult Hepatitis B Vaccine	129

18. How do immunization rates vary across different patient demographics?

By age

To answer this question, I categorized patients into age groups and calculated the count of vaccine descriptions for each age group. Then using a window function I ranked the counts within each age group and filtered the results to get the top 3 counts for each age group.

```
WITH age_groups AS (
    SELECT
        id,
        CASE
            WHEN DATE_PART('YEAR', AGE(current_date, birthdate)) <= 12 THEN 'Child'
            WHEN DATE_PART('YEAR', AGE(current_date, birthdate)) >= 13 AND DATE_PART('YEAR', AGE(current_date, birthdate)) <= 19 THEN 'Teen'
            WHEN DATE_PART('YEAR', AGE(current_date, birthdate)) >= 20 AND DATE_PART('YEAR', AGE(current_date, birthdate)) <= 35 THEN 'Young-Adult'
            WHEN DATE_PART('YEAR', AGE(current_date, birthdate)) >= 36 AND DATE_PART('YEAR', AGE(current_date, birthdate)) <= 55 THEN 'Middle-Aged'
            WHEN DATE_PART('YEAR', AGE(current_date, birthdate)) >= 56 AND DATE_PART('YEAR', AGE(current_date, birthdate)) <= 75 THEN 'Older-Adult'
            ELSE 'Senior'
        END AS age_group
    FROM
        patients_staging
),
vaccine_counts AS (
    SELECT
        age_groups.age_group,
        immunizations.description,
        COUNT(*) AS count
    FROM
        immunizations
    JOIN
        age_groups ON immunizations.patient = age_groups.id
    GROUP BY
        age_groups.age_group,
        immunizations.description
),
ranked_vaccine_counts AS (
    SELECT
        age_group,
        description,
        count,
        ROW_NUMBER() OVER (PARTITION BY age_group ORDER BY count DESC) AS rank
    FROM
        vaccine_counts
)
SELECT
    age_group,
    description,
    count
FROM
    ranked_vaccine_counts
WHERE
    rank <= 3
ORDER BY
    age_group,
    rank;
```

	age_group text	description character varying (500)	count bigint
1	Child	Seasonal Flu Vaccine	6840
2	Child	Diphtheria, Tetanus, and Pertussis Vaccine	5005
3	Child	Pneumococcal Conjugate Vaccine 13	4192
4	Middle-Aged	Seasonal Flu Vaccine	21302
5	Middle-Aged	Five doses of tetanus toxoid, preservative-free and adsorbed, for adul...	2803
6	Middle-Aged	Adult Heptatitis A Vaccine	2764
7	Older-Adult	Seasonal Flu Vaccine	27302
8	Older-Adult	Five doses of tetanus toxoid, preservative-free and adsorbed, for adul...	2728
9	Older-Adult	Novel Coronavirus (COVID-19) mRNA Vaccine 30 mcg/0.3mL Dose	2110
10	Senior	Seasonal Flu Vaccine	12012
11	Senior	Five doses of tetanus toxoid, preservative-free and adsorbed, for adul...	1011
12	Senior	Novel Coronavirus (COVID-19) mRNA Vaccine 30 mcg/0.3mL Dose	682
13	Teen	Seasonal Flu Vaccine	8924
14	Teen	Human Papillomavirus (HPV) Four-strain Vaccine	2558
15	Teen	Diphtheria, Tetanus, and Pertussis Vaccine	1315
16	Young-Adult	Seasonal Flu Vaccine	16839
17	Young-Adult	Meningococcal Tetravalent Polysaccharide Vaccine (MCV4P)	2797
18	Young-Adult	Adult Hepatitis B Vaccine	2594

By race

To answer this question, I calculated the count of vaccine descriptions for each race, used a window function to rank the counts within each race and filtered the results to get the top 3 counts per race.

```
WITH vaccine_counts AS (
    SELECT patients_staging.race,
           immunizations.description,
           COUNT(*) AS count
    FROM immunizations
    JOIN patients_staging ON immunizations.patient = patients_staging.id
    GROUP BY patients_staging.race,
             immunizations.description
),
ranked_vaccine_counts AS (
    SELECT race,
           description,
           count,
           ROW_NUMBER() OVER (PARTITION BY race ORDER BY count DESC) AS rank
    FROM vaccine_counts
)
SELECT race,
       description,
       count
FROM ranked_vaccine_counts
WHERE rank <= 3
ORDER BY race,
         rank;
```

	race character varying (100)	description character varying (500)	count bigint
1	asian	Seasonal Flu Vaccine	5973
2	asian	Five doses of tetanus toxoid, preservative-free and adsorbed, for adul...	538
3	asian	Novel Coronavirus (COVID-19) mRNA Vaccine 30 mcg/0.3mL Dose	484
4	black	Seasonal Flu Vaccine	8430
5	black	Five doses of tetanus toxoid, preservative-free and adsorbed, for adul...	742
6	black	Novel Coronavirus (COVID-19) mRNA Vaccine 30 mcg/0.3mL Dose	660
7	hawaiian	Seasonal Flu Vaccine	1176
8	hawaiian	Five doses of tetanus toxoid, preservative-free and adsorbed, for adul...	112
9	hawaiian	Novel Coronavirus (COVID-19) mRNA Vaccine 30 mcg/0.3mL Dose	108
10	native	Seasonal Flu Vaccine	490
11	native	Five doses of tetanus toxoid, preservative-free and adsorbed, for adul...	39
12	native	Novel Coronavirus (COVID-19) mRNA Vaccine 30 mcg/0.3mL Dose	36
13	other	Seasonal Flu Vaccine	917
14	other	Novel Coronavirus (COVID-19) mRNA Vaccine 30 mcg/0.3mL Dose	96
15	other	Five doses of tetanus toxoid, preservative-free and adsorbed, for adul...	84
16	white	Seasonal Flu Vaccine	76233
17	white	Five doses of tetanus toxoid, preservative-free and adsorbed, for adul...	6919
18	white	Novel Coronavirus (COVID-19) mRNA Vaccine 30 mcg/0.3mL Dose	6179

19. Is there any correlation between certain immunizations and the occurrence or severity of specific conditions?

To answer this question, I joined the immunizations and conditions tables to identify patients who have both immunizations and conditions. Group the data by immunization and condition to get the count of occurrences.

Since majority of people got the seasonal flu vaccine, I removed the flu vaccine description from the list and looked at if there is any relation between the vaccine types of people take and the conditions they were treated on.

Based on the result and my understanding I see no relation between the conditions and the vaccines. However, a medical professional can identify any relation. For example, people with psychological and physical stress have got various types of vaccines. Do these vaccines have any relation with stress?

```
WITH immunizations_conditions AS (
    SELECT
        i.patient,
        i.code AS immunization_code,
        i.description AS immunization_description,
        c.code AS condition_code,
        c.description AS condition_description
    FROM
        immunizations_staging i
    JOIN
        conditions_staging c ON i.patient = c.patient
    WHERE
        i.description <> 'Seasonal Flu Vaccine'
)
SELECT
    immunization_code,
    immunization_description,
    condition_code,
    condition_description,
    COUNT(*) AS occurrence_count
FROM
    immunizations_conditions
GROUP BY
    immunization_code,
    immunization_description,
    condition_code,
    condition_description
ORDER BY
    occurrence_count DESC;
```

	immunization_code integer	immunization_description character varying (500)	condition_code character varying (1000)	condition_description character varying (200)	occurrence_count bigint
1	5303	Five doses of tetanus toxoid, preservative-free and adsorbed, for adults.	V62.89	Other psychological or physical stress, not elsewhere classified	55173
2	5309	Novel Coronavirus (COVID-19) mRNA Vaccine 30 mcg/0.3mL Dose	V62.89	Other psychological or physical stress, not elsewhere classified	43010
3	5304	Novel Coronavirus (COVID-19) mRNA Vaccine 100 mcg/0.5mL Dose	V62.89	Other psychological or physical stress, not elsewhere classified	35667
4	5301	Herpes Zoster Vaccine (Live)	V62.89	Other psychological or physical stress, not elsewhere classified	22211
5	5308	Adult Heptatitis A Vaccine	V62.89	Other psychological or physical stress, not elsewhere classified	13789
6	5321	23-Valent Pneumococcal Polysaccharide Vaccine	V62.89	Other psychological or physical stress, not elsewhere classified	10931
7	5315	Pneumococcal Conjugate Vaccine 13	V62.89	Other psychological or physical stress, not elsewhere classified	10039
8	5303	Five doses of tetanus toxoid, preservative-free and adsorbed, for adults.	V22.2	Pregnant state, incidental	9494
9	5309	Novel Coronavirus (COVID-19) mRNA Vaccine 30 mcg/0.3mL Dose	V22.2	Pregnant state, incidental	7214
10	5319	Adult Hepatitis B Vaccine	V22.2	Pregnant state, incidental	5818
11	5304	Novel Coronavirus (COVID-19) mRNA Vaccine 100 mcg/0.5mL Dose	V22.2	Pregnant state, incidental	5668
12	5319	Adult Hepatitis B Vaccine	V62.89	Other psychological or physical stress, not elsewhere classified	5603
13	5303	Five doses of tetanus toxoid, preservative-free and adsorbed, for adults.	V85.30	Body Mass Index 30.0-30.9, adult	5294
14	5310	Diphtheria, Tetanus, and Pertussis Vaccine	382.9	Unspecified otitis media	4705
15	5306	Meningococcal Tetravalent Polysaccharide Vaccine (MCV4P)	V22.2	Pregnant state, incidental	4580
16	5303	Five doses of tetanus toxoid, preservative-free and adsorbed, for adults.	V62.0	Unemployment	4565
17	5303	Five doses of tetanus toxoid, preservative-free and adsorbed, for adults.	466	Acute bronchitis	4270
18	5306	Meningococcal Tetravalent Polysaccharide Vaccine (MCV4P)	V62.89	Other psychological or physical stress, not elsewhere classified	4250