Аналіз факторів вартості медичного страхування

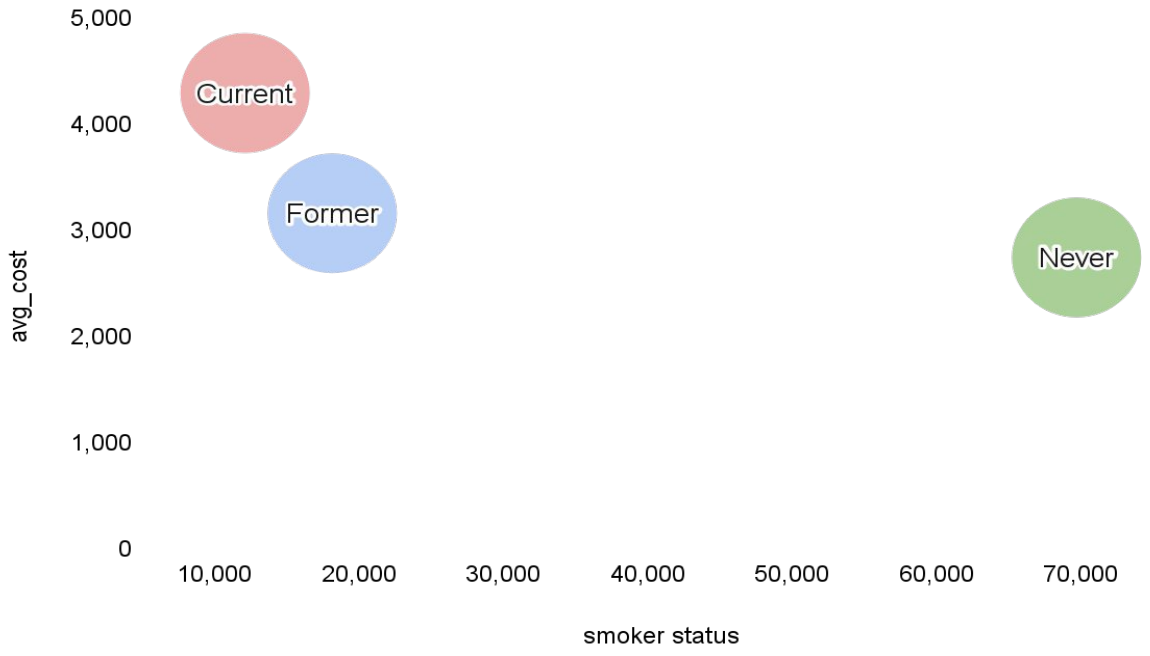
Катерина Майкова • 19.10.2025

Питання, що досліджуються

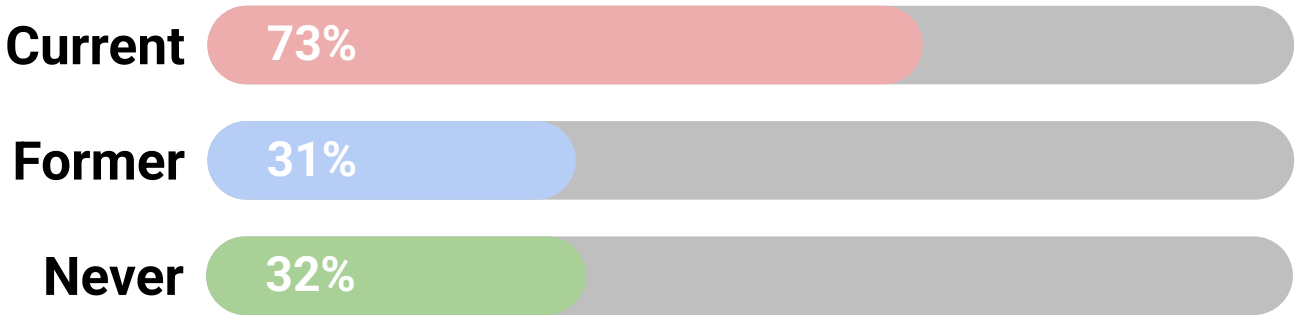
1. Наскільки статус курця впливає на середні медичні витрати порівняно з некурцями та колишніми курцями?
2. Чи існує статистично значуща різниця в поширеності хронічних захворювань між курцями та некурцями?
3. Яка комбінація призводить до найвищих середніх витрат?
4. Який профіль клієнта належить до 10% найдорожчих за витратами та чим він відрізняється від середнього клієнта?
5. Як вік клієнта корелює з витратами та чи відрізняється ця кореляція для курців та некурців?
6. Чи існує регіональна відмінність у “сімейному пакеті” (якщо є утриманці)?
7. Які три фактори є найбільш значущими для медичних витрат?
8. Як статус страхування впливає на середні витрати та чи є ця різниця послідовною в різних регіонах?

на основі даних
https://www.kaggle.com/datasets/mohankrishnathalla/medical-insurance-cost-prediction/data

К-ть клієнтів та їх середні витрати за статусом "Курця"



Мають “Високий ризик”



Чітка ієрархія витрат.
Поточні Курці платять найбільше, а ті, хто **ніколи не курив**, — найменше.

Середні витрати поточного курця **майже в 1.56 рази вищі** (4295.56 / 2746.01) за витрати того, хто ніколи не курив.

Витрати **колишніх курців** знаходяться між двома іншими групами, але вони **значно ближчі до тих, хто ніколи не курив** (різниця 3161 vs 2746), ніж до поточних (4295 vs 3161).

Статус курця як один з основних або навіть **вирішальних** факторів для присвоєння категорії "**високий ризик**".

Медичний та поведінковий портрети за статусом "Курця", %

smoker status	hypertension	diabetes	asthma	copd	cardio	cancer history	kidney disease	liver disease	arthritis	mental health
Current	20.6	8.6	5.7	3.5	4.9	2.2	1.4	1.4	10.6	13.6
Former	20.2	8.6	6.0	3.6	5.0	2.2	1.4	1.6	10.9	13.1
Never	20.3	8.6	5.9	3.6	5.2	2.1	1.5	1.5	10.9	12.9

smoker	alcohol daily	alcohol weekly	alcohol none
Current	4.8	20.1	29.9
Former	5.1	20.3	29.7
Never	5.0	19.6	30.2

Середні показники тиску та холестерину

smoker	avg_systolic_bp	avg_diastolic_bp	avg_ldl_cholesterol
Current	118.4	73.8	119.6
Former	118.2	73.7	120.2
Never	118.4	73.8	120.3

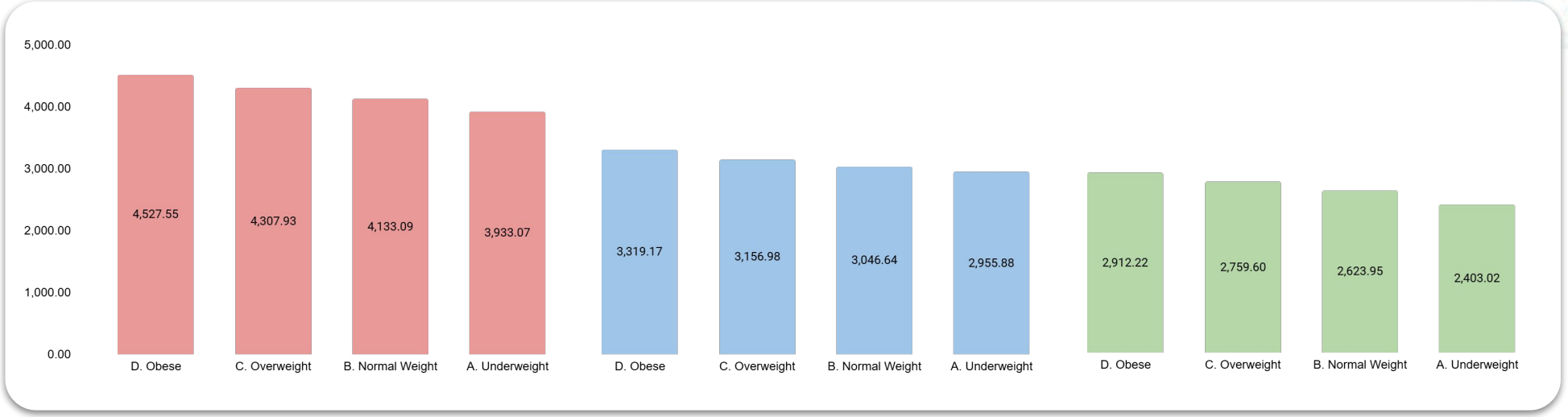
Вплив статусу курця на поширеність хронічних захворювань не впливає.

Єдина помітна відмінність:
Ментальне Здоров'я: Поточні курці мають найвищий відсоток (13.59%), що може вказувати на кореляцію куріння зі стресом або тривогою.

Рак (Cancer History): Хоча різниця мінімальна, колишні курці (Former, 2.23%) мають найвищу історію раку. Це наштовхує на думку, що ці люди кинули курити саме через діагноз.

Алкоголь не є предиктором ризику: Частота вживання алкоголю практично однакова у всіх трьох групах. Це **спростовує гіпотезу** про те, що **куріння йде пліч-о-пліч із високою частотою вживання алкоголю.**

Вплив ваги на вартість страхування



Всі чотири **найдорожчі** групи — це Поточні Курці (Current).

Навіть Поточний Курець з Нормальною Вагою платить $4,133.09, що вище, ніж будь-яка група Колишніх Курців чи тих, хто ніколи не курив, навіть якщо вони мають ожиріння (Obese).

Це доводить, що страховик оцінює поточне куріння як значно більший і безпосередній ризик, ніж хронічні наслідки ожиріння.

Висновки та рекомендації

1. **Статус Курця - найсильніший фактор**, що впливає на витрати, значно переважаючи BMI, стать та поточні біомаркери.

2. Найвища вартість страхування спостерігається у групі Current Smoker + Obese ($4,527).
Однак, куріння є настільки потужним ризиком, що Current Smoker з нормальною вагою ($4,133) коштує дорожче, ніж Never Smoker з ожирінням ($2,912).

3. **72% поточних курців** віднесені до **категорії високого страхового ризику**, але їхні середні біомаркери (тиск, холестерин) та поширеність хронічних хвороб практично **не відрізняються від некурців**. Це свідчить про те, що ціна ґрунтується на поведінковому, а не поточному медичному ризику

4. Клієнти, що належать до 10% найдорожчих за витратами ($6,268+), є старшими на 4 роки (Avg. 52.44) і в 2.18 рази частіше є поточними курцями, ніж клієнти з низькими витратами.

Переглянути політику формування страхових витрат, оцінюючи не лише статус курця, а й реальний медичний стан клієнта та його спосіб життя.

Зосередити програми зниження ризику на поточних курцях віком 45+

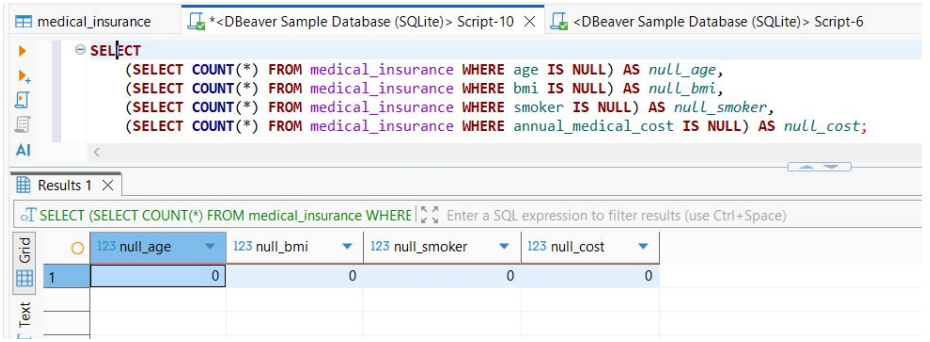
Створити більш значущі стимули для колишніх курців (Former Smoker).

Дашборд

Додатки

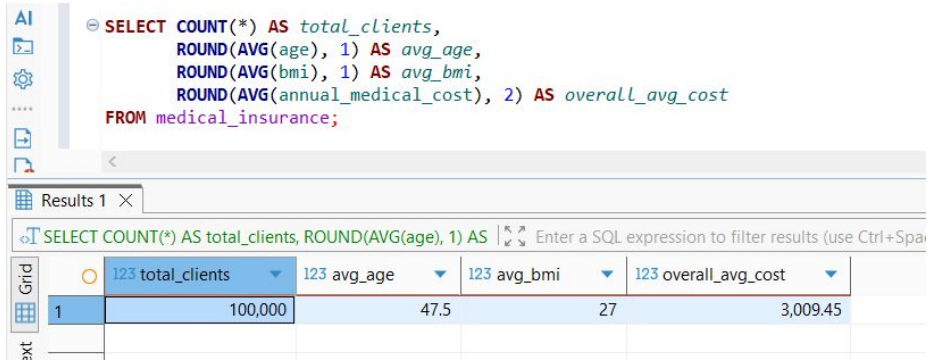
Перевірка даних на наявність відсутніх значень

```sql
SELECT
  (SELECT COUNT(*) FROM medical_insurance WHERE age IS NULL) AS null_age,
  (SELECT COUNT(*) FROM medical_insurance WHERE bmi IS NULL) AS null_bmi,
  (SELECT COUNT(*) FROM medical_insurance WHERE smoker IS NULL) AS null_smoker,
  (SELECT COUNT(*) FROM medical_insurance WHERE annual_medical_cost IS NULL) AS null_cost;
```



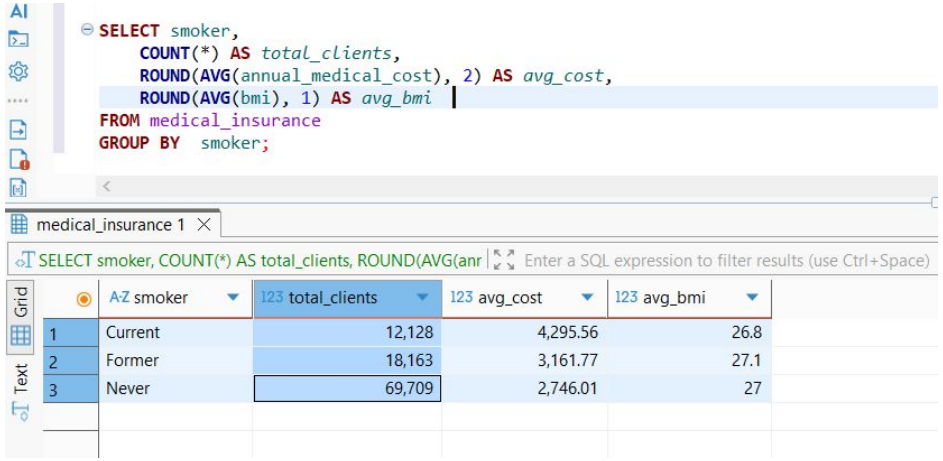
Розрахунок середніх значень основних показників

```sql
SELECT COUNT(*) AS total_clients,
    ROUND(AVG(age), 1) AS avg_age,
    ROUND(AVG(bmi), 1) AS avg_bmi,
    ROUND(AVG(annual_medical_cost), 2) AS overall_avg_cost
FROM medical_insurance;
```



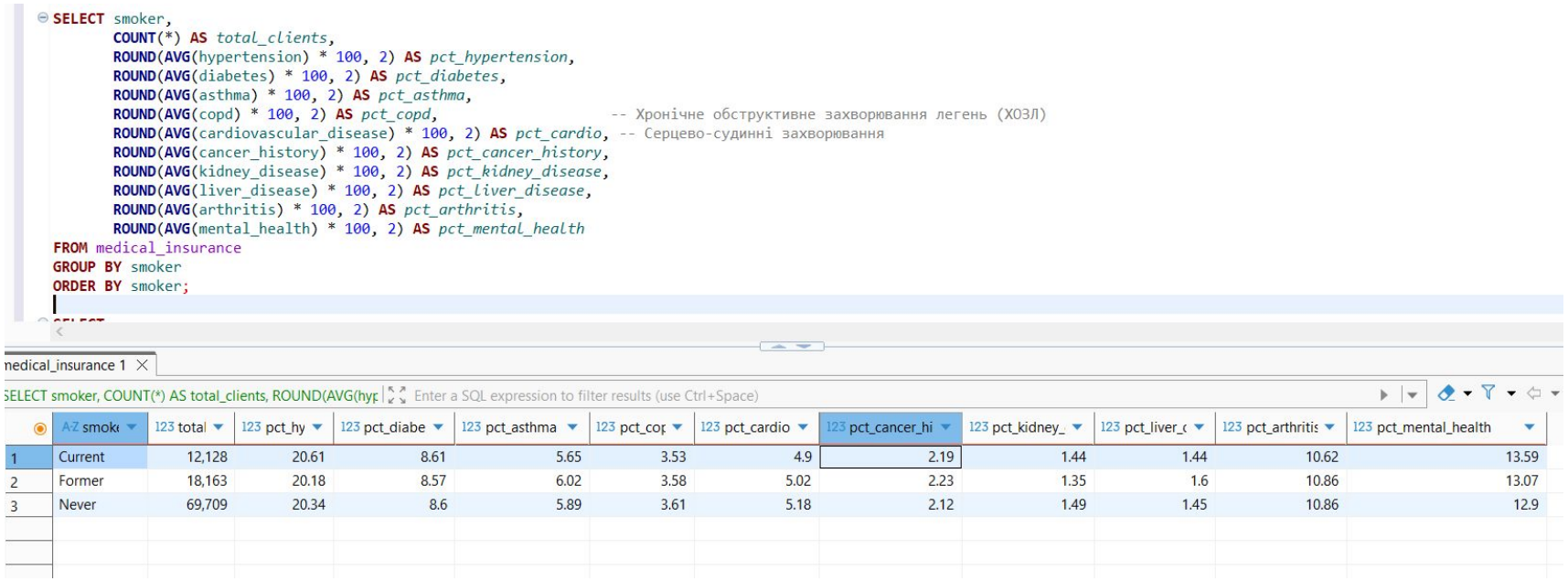
Про курця та його витрати

```sql
SELECT smoker,
  COUNT(*) AS total_clients,
  ROUND(AVG(annual_medical_cost), 2) AS avg_cost,
  ROUND(AVG(bmi), 1) AS avg_bmi
FROM medical_insurance
GROUP BY  smoker;
```



Хвороби курця

```sql
SELECT smoker,
    COUNT(*) AS total_clients,
    ROUND(AVG(hypertension) * 100, 2) AS pct_hypertension,
    ROUND(AVG(diabetes) * 100, 2) AS pct_diabetes,
    ROUND(AVG(asthma) * 100, 2) AS pct_asthma,
    ROUND(AVG(copd) * 100, 2) AS pct_copd,                -- Хронічне обструктивне захворювання легень (ХОЗЛ)
    ROUND(AVG(cardiovascular_disease) * 100, 2) AS pct_cardio, -- Серцево-судинні захворювання
    ROUND(AVG(cancer_history) * 100, 2) AS pct_cancer_history,
    ROUND(AVG(kidney_disease) * 100, 2) AS pct_kidney_disease,
    ROUND(AVG(liver_disease) * 100, 2) AS pct_liver_disease,
    ROUND(AVG(arthritis) * 100, 2) AS pct_arthritis,
    ROUND(AVG(mental_health) * 100, 2) AS pct_mental_health
FROM medical_insurance
GROUP BY smoker
ORDER BY smoker;
```



medical_insurance 1

SELECT smoker, COUNT(*) AS total_clients, ROUND(AVG(hyp

Enter a SQL expression to filter results (use Ctrl+Space)

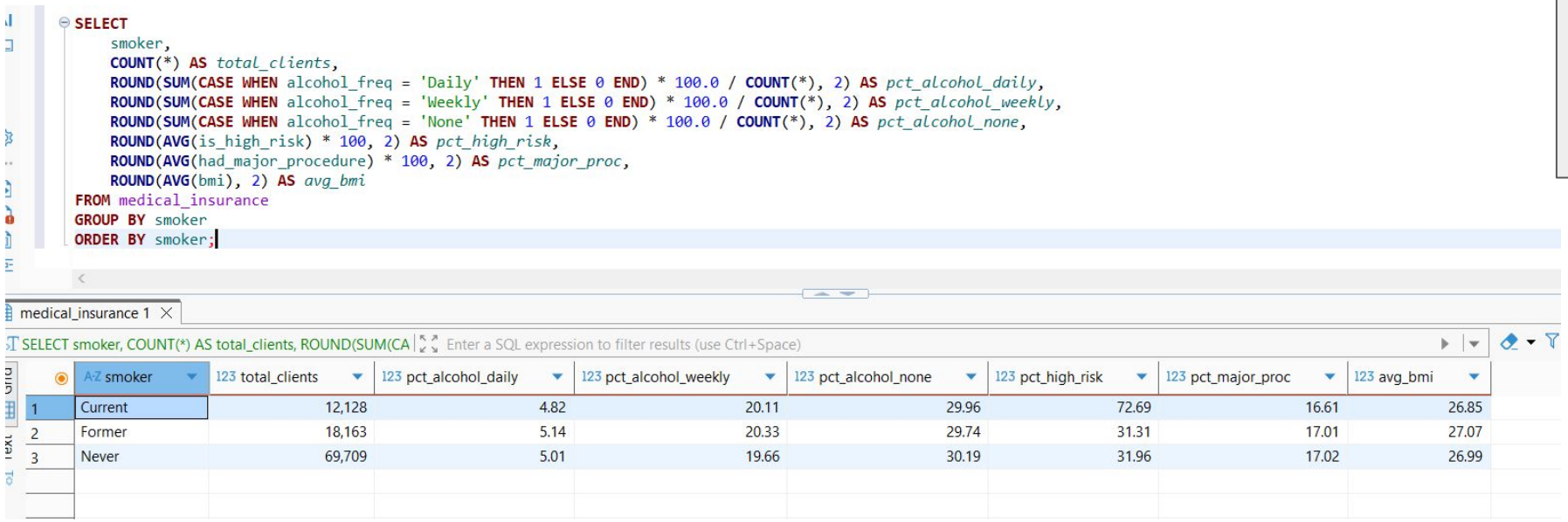
	A-Z smoke	123 total	123 pct_hy	123 pct_diabe	123 pct_asthma	123 pct_cop	123 pct_cardio	123 pct_cancer_hi	123 pct_kidney_	123 pct_liver_c	123 pct_arthritis	123 pct_mental_health
1	Current	12,128	20.61	8.61	5.65	3.53	4.9	2.19	1.44	1.44	10.62	13.59
2	Former	18,163	20.18	8.57	6.02	3.58	5.02	2.23	1.35	1.6	10.86	13.07
3	Never	69,709	20.34	8.6	5.89	3.61	5.18	2.12	1.49	1.45	10.86	12.9

Не допомогло вивести портрет
Підозрюю, що дані про хвороби не підтверджені медично, а проставлені клієнтом (особою, що страхується) самостійно

Хоча
Менталка: Поточні курці мають найвищий відсоток, що наштовхує на думку про **кореляцію куріння зі стресом або тривогою**.
Рак: різниця мінімальна, але колишні курці мають найвищу історію раку. Можливо ці люди кинули курити саме через діагноз.

Аналізую курця з іншої сторони

```sql
SELECT
  smoker,
  COUNT(*) AS total_clients,
  ROUND(SUM(CASE WHEN alcohol_freq = 'Daily' THEN 1 ELSE 0 END) * 100.0 / COUNT(*), 2) AS pct_alcohol_daily,
  ROUND(SUM(CASE WHEN alcohol_freq = 'Weekly' THEN 1 ELSE 0 END) * 100.0 / COUNT(*), 2) AS pct_alcohol_weekly,
  ROUND(SUM(CASE WHEN alcohol_freq = 'None' THEN 1 ELSE 0 END) * 100.0 / COUNT(*), 2) AS pct_alcohol_none,
  ROUND(AVG(is_high_risk) * 100, 2) AS pct_high_risk,
  ROUND(AVG(had_major_procedure) * 100, 2) AS pct_major_proc,
  ROUND(AVG(bmi), 2) AS avg_bmi
FROM medical_insurance
GROUP BY smoker
ORDER BY smoker;
```



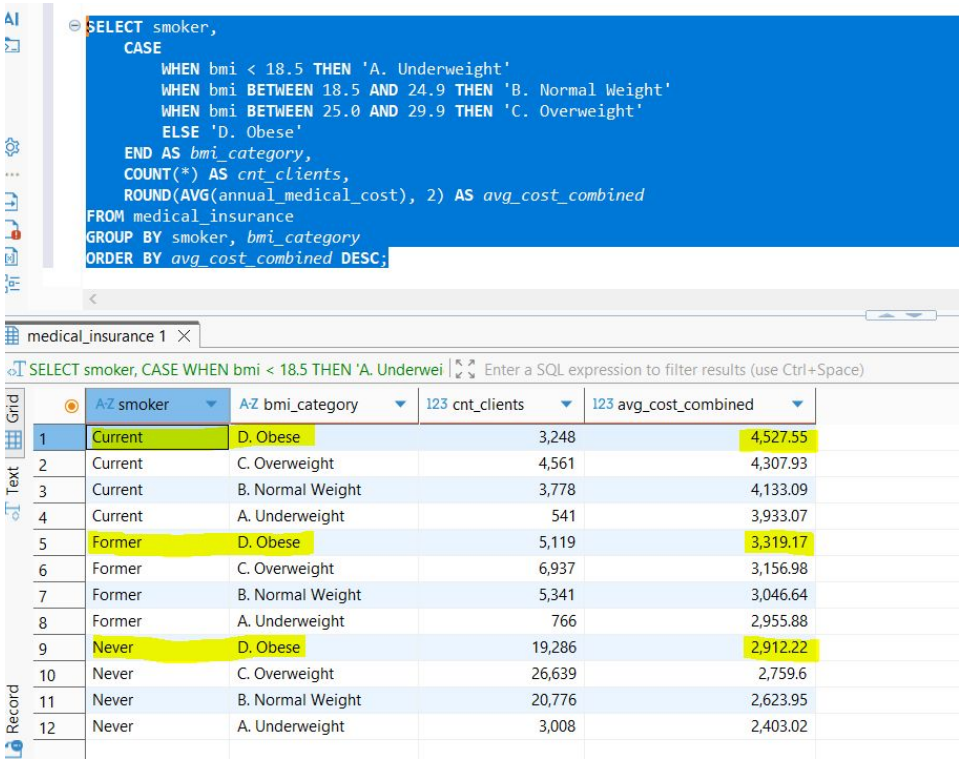
	smoker	total_clients	pct_alcohol_daily	pct_alcohol_weekly	pct_alcohol_none	pct_high_risk	pct_major_proc	avg_bmi
1	Current	12,128	4.82	20.11	29.96	72.69	16.61	26.85
2	Former	18,163	5.14	20.33	29.74	31.31	17.01	27.07
3	Never	69,709	5.01	19.66	30.19	31.96	17.02	26.99

Частота вживання алкоголю не впливає на рівень ризику клієнта, тому можна спокійно сьорбати вінішко щодня))
і як видно якісь серйозні втручання також не впливають, всі три категорії ведуть себе однаково

На рівень ризику клієнта впливає саме статус діючого курця
Підозрюю, що страховик впевнений шо куріння в майбутньому принесе “свої негативні плоди”

Аналізую чи впливає вага на вартість страхування (окрім статусу курця)

```sql
SELECT smoker,
  CASE
    WHEN bmi < 18.5 THEN 'A. Underweight'
    WHEN bmi BETWEEN 18.5 AND 24.9 THEN 'B. Normal Weight'
    WHEN bmi BETWEEN 25.0 AND 29.9 THEN 'C. Overweight'
    ELSE 'D. Obese'
  END AS bmi_category,
  COUNT(*) AS cnt_clients,
  ROUND(AVG(annual_medical_cost), 2) AS avg_cost_combined
FROM medical_insurance
GROUP BY smoker, bmi_category
ORDER BY avg_cost_combined DESC;
```



medical_insurance 1

SELECT smoker, CASE WHEN bmi < 18.5 THEN 'A. Underwei | Enter a SQL expression to filter results (use Ctrl+Space)

	smoker	bmi_category	cnt_clients	avg_cost_combined
1	Current	D. Obese	3,248	4,527.55
2	Current	C. Overweight	4,561	4,307.93
3	Current	B. Normal Weight	3,778	4,133.09
4	Current	A. Underweight	541	3,933.07
5	Former	D. Obese	5,119	3,319.17
6	Former	C. Overweight	6,937	3,156.98
7	Former	B. Normal Weight	5,341	3,046.64
8	Former	A. Underweight	766	2,955.88
9	Never	D. Obese	19,286	2,912.22
10	Never	C. Overweight	26,639	2,759.6
11	Never	B. Normal Weight	20,776	2,623.95
12	Never	A. Underweight	3,008	2,403.02

в будь-якій категорії вага є впливоим чинником на вартість страхування, що є логічним
Але це вже вторинний вплив. Першрчерговим є статус поточного курця

Аналізую вік/стать

```sql
SELECT
  smoker,
  sex,
  COUNT(*) AS total_clients,
  ROUND(AVG(age), 1) AS avg_age,
  ROUND(AVG(annual_medical_cost), 2) AS avg_cost
FROM medical_insurance
WHERE age >= 18
GROUP BY smoker, sex
ORDER BY smoker, avg_cost DESC;
```

```sql
SELECT
    smoker,
    sex,
    COUNT(*) AS total_clients,
    ROUND(AVG(age), 1) AS avg_age,
    ROUND(AVG(annual_medical_cost), 2) AS avg_cost
FROM medical_insurance
WHERE age >= 18
GROUP BY smoker, sex
ORDER BY smoker, avg_cost DESC;
```

dical_insurance 1

ECT smoker, sex, COUNT(*) AS total_clients, ROUND(AV(Enter a SQL expression to filter results (use Ctrl+Space)

smoker	sex	total_clients	avg_age	avg_cost
Current	Male	5,624	48.5	4,345.65
Current	Female	5,847	48.6	4,335.19
Current	Other	244	49.2	4,105.22
Former	Other	337	46.3	3,387.1
Former	Male	8,518	48.5	3,207.59
Former	Female	8,762	48.4	3,153.57
Never	Male	33,162	48.8	2,779.85
Never	Female	33,088	48.7	2,753.18
Never	Other	1,374	48.5	2,724.74

шось я вже починаю сумніватись, шо вибрала нормальний набір даних((

але вже всьо, путі назад нема…

аналіз знову показує, що ні вік ні стать не впливають на вартість

і навіть з тиском та холестерином у всіх все однаково

Аналізую тиск/ холестерин

```sql
SELECT
  smoker,
  COUNT(*) AS total_clients,
  ROUND(AVG(systolic_bp), 1) AS avg_systolic_bp,
  ROUND(AVG(diastolic_bp), 1) AS avg_diastolic_bp,
  ROUND(AVG(ldl), 1) AS avg_ldl_cholesterol
FROM medical_insurance
WHERE age >= 18
GROUP BY smoker
ORDER BY smoker;
```

```sql
SELECT
    smoker,
    COUNT(*) AS total_clients,
    ROUND(AVG(systolic_bp), 1) AS avg_systolic_bp,
    ROUND(AVG(diastolic_bp), 1) AS avg_diastolic_bp,
    ROUND(AVG(ldl), 1) AS avg_ldl_cholesterol
FROM medical_insurance
WHERE age >= 18
GROUP BY smoker
ORDER BY smoker;
```

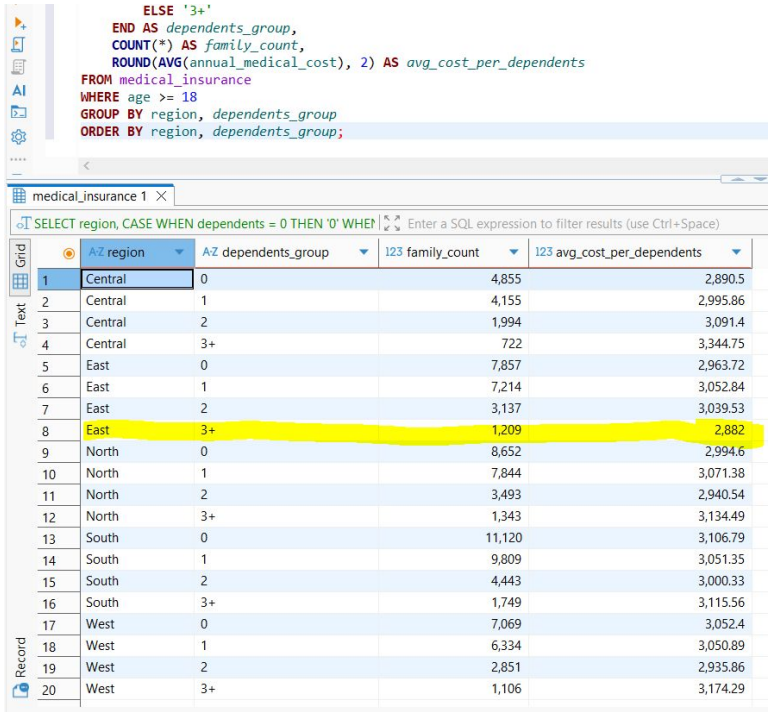
dical_insurance 1

ECT smoker, COUNT(*) AS total_clients, ROUND(AVG(sys Enter a SQL expression to filter results (use Ctrl+Space)

smoker	total_clients	avg_systolic_bp	avg_diastolic_bp	avg_ldl_cholesterol
Current	11,715	118.4	73.8	119.6
Former	17,617	118.2	73.7	120.2
Never	67,624	118.4	73.8	120.3

Залежність від к-ті дітей та регіону

```sql
SELECT region,
  CASE
    WHEN dependents = 0 THEN '0'
    WHEN dependents = 1 THEN '1'
    WHEN dependents = 2 THEN '2'
    ELSE '3+'
  END AS dependents_group,
  COUNT(*) AS family_count,
  ROUND(AVG(annual_medical_cost), 2) AS avg_cost_per_dependents
FROM medical_insurance
WHERE age >= 18
GROUP BY region, dependents_group
ORDER BY region, dependents_group;
```



```
    ELSE '3+'
  END AS dependents_group,
  COUNT(*) AS family_count,
  ROUND(AVG(annual_medical_cost), 2) AS avg_cost_per_dependents
FROM medical_insurance
WHERE age >= 18
GROUP BY region, dependents_group
ORDER BY region, dependents_group;
```

medical_insurance 1

SELECT region, CASE WHEN dependents = 0 THEN '0' WHEN ... Enter a SQL expression to filter results (use Ctrl+Space)

	region	dependents_group	family_count	avg_cost_per_dependents
1	Central	0	4,855	2,890.5
2	Central	1	4,155	2,995.86
3	Central	2	1,994	3,091.4
4	Central	3+	722	3,344.75
5	East	0	7,857	2,963.72
6	East	1	7,214	3,052.84
7	East	2	3,137	3,039.53
8	East	3+	1,209	2,882
9	North	0	8,652	2,994.6
10	North	1	7,844	3,071.38
11	North	2	3,493	2,940.54
12	North	3+	1,343	3,134.49
13	South	0	11,120	3,106.79
14	South	1	9,809	3,051.35
15	South	2	4,443	3,000.33
16	South	3+	1,749	3,115.56
17	West	0	7,069	3,052.4
18	West	1	6,334	3,050.89
19	West	2	2,851	2,935.86
20	West	3+	1,106	3,174.29

О! цікаво є анамалія у East регіоні, для сімей із 3+ дітьми вартість страхування найнижча