НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ

імені Ігоря Сікорського»

Факультет прикладної математики

Кафедра прикладної математики

Лабораторна робота №1

Розвідковий аналіз даних

Виконав:

студенти групи КМ-12

Якімов Я. О., Лазебний О. А.,

Петров В.

Керівник:

Тавров Д. Ю.

Київ — 2024

ЗМІСТ

ВСТУП

В рамках даної лабораторної роботи було проведено розвідковий аналіз даних за набором “Diabetes Health Indicators Dataset”, головною задачею якого було дати відповіді на дослідницькі питання:

1. залежність наявності діабету від факторів ризику (підвищений рівень холестерину, наявність ожиріння, проблеми з серцево-судинною системою, вік);
2. залежність наявності діабету від соціального статусу (раціон харчування, паління, зарплата, освіта, страховка);
3. залежність частоти обстежень рівня холестерину від соціального статусу (рівня освіти, зарплати та медичного страхування);
4. залежність наявності діабету від суб'єктивної оцінки стану здоров'я. Порівняння із залежністю наявності діабету від об'єктивних факторів: імт, тиску, чи є хвороби серця і того, чи важко людині ходити (чи завищують/занижують люди оцінку власного здоров'я);
5. чи впливає поганий ментальний стан на низький рівень фізичного здоров'я;
6. залежність ментального здоров'я, оцінки фізичного здоров'я, від віку серед людей, які хворіють на діабет.

Поставлена задача перевірити гіпотези та очікування щодо даного набору даних, зокрема, чи є сильним вплив серцево-судинних захворювань, звичок, ментального стану, рівня освіти, зарплати, віку, статі та використанням медичних послуг на наявність діабету, а також вплив між цим факторами.

ОПИС ДАНИХ ТА КРОКІВ ІЗ ЇХ ПІДГОТОВКИ

Дані зібрані у рамках щорічного телефонного опитування Центрами контролю та профілактики захворювань США (Centers for Disease Control and Prevention) у 2015 році. Було опитано 441 455 осіб, на основі чого було створено початковий набір даних із 330 змінними: це безпосередні результати опитування кожного та проміжні дані.

У загальному доступі наявний набір даних із 22 змінними, який і використовується для досліджень у цій роботі.

З метою очищення даних було проведено перетворення деяких змінних, зокрема категоріальних та цензурованих. Значення цих змінних приведено у відповідність до назв у кодовій книжці (codebook). Задля цього було використано скрипт на мові “R”, який зчитує, перетворює та зберігає у файлі “.csv” результат роботи.

Файл очищених даних містить 253 680 записів із 22 змінними, пропущених даних немає.

Дескриптивні статистики набору даних такі

Назва змінної	Дескриптивні статистики

Diabetes	Логічна змінна
	Таблиця спряженості
	`FALSE  TRUE`
	`213703  39977`

HighBP	Логічна змінна
	Таблиця спряженості
	`FALSE  TRUE`
	`144851 108829`

HighChol	Логічна змінна
	Таблиця спряженості
	`FALSE  TRUE`
	`146089 107591`

CholCheck	Логічна змінна
	Таблиця спряженості
	`FALSE  TRUE`
	`  9470 244210`

BMI	Неперервна змінна Середнє: 28.38236 Середньоквадратичне відхилення: 6.608694

	Підсумкові характеристики
	`Min. 1st Qu. Median Mean 3rd Qu. Max.`
	`12.00 24.00 27.00 28.38 31.00 98.00`
	За правилом трьох сигм виявлено 2963 викидів, інтервал - [12; 48]
	Фільтр Гампеля показує 11506 викидів, залишаючи інтервал [14; 40]
	На qqplot не виявлено істотних викидів (див. додаток А)
	На діаграмі розкиду - так само, за винятком кількох значень 98 і 96

Smoker	Логічна змінна
	Таблиця спряженості
	`FALSE TRUE`
	`141257 112423`

Stroke	Логічна змінна
	Таблиця спряженості
	`FALSE TRUE`
	`243388 10292`

HeartDiseaseorAttack	Логічна змінна
	Таблиця спряженості
	`FALSE TRUE`
	`229787 23893`

PhysActivity	Логічна змінна
	Таблиця спряженості
	`FALSE TRUE`
	`61760 191920`

Fruits	Логічна змінна
	Таблиця спряженості
	`FALSE TRUE`
	`92782 160898`

Veggies	Логічна змінна
	Таблиця спряженості
	`FALSE TRUE`
	`47839 205841`

HvyAlcoholConsump	Логічна змінна
	Таблиця спряженості
	`FALSE TRUE`
	`239424 14256`

AnyHealthcare	Логічна змінна

	Таблиця спряженості FALSE TRUE 12417 241263
NoDocbcCost	Логічна змінна Таблиця спряженості FALSE TRUE 232326 21354
GenHlth	Порядкова змінна Таблиця спряженості excellent fair good poor 45299 31570 75646 12081 very good 89084 SD = 1.0684774 Var = 1.1416439 Min. 1st Qu. Median Mean 3rd Qu. Max. 1.000 3.000 4.000 3.489 4.000 5.000
MentHlth	Порядкова змінна Таблиця спряженості 0 1 2 3 4 175680 8538 13054 7381 3789 5 6 7 8 9 9030 988 3100 639 91 10 11 12 13 14 6373 41 398 41 1167 15 16 17 18 19 5505 88 54 97 16 20 21 22 23 24 3364 227 63 38 33 25 26 27 28 29 1188 45 79 327 158 30 12088 SD = 7.4128467 Var = 54.9502961 Min. 1st Qu. Median Mean 3rd Qu. Max. 0.000 0.000 0.000 3.185 2.000 30.000
PhysHlth	Порядкова змінна Таблиця спряженості 0 1 2 3 4

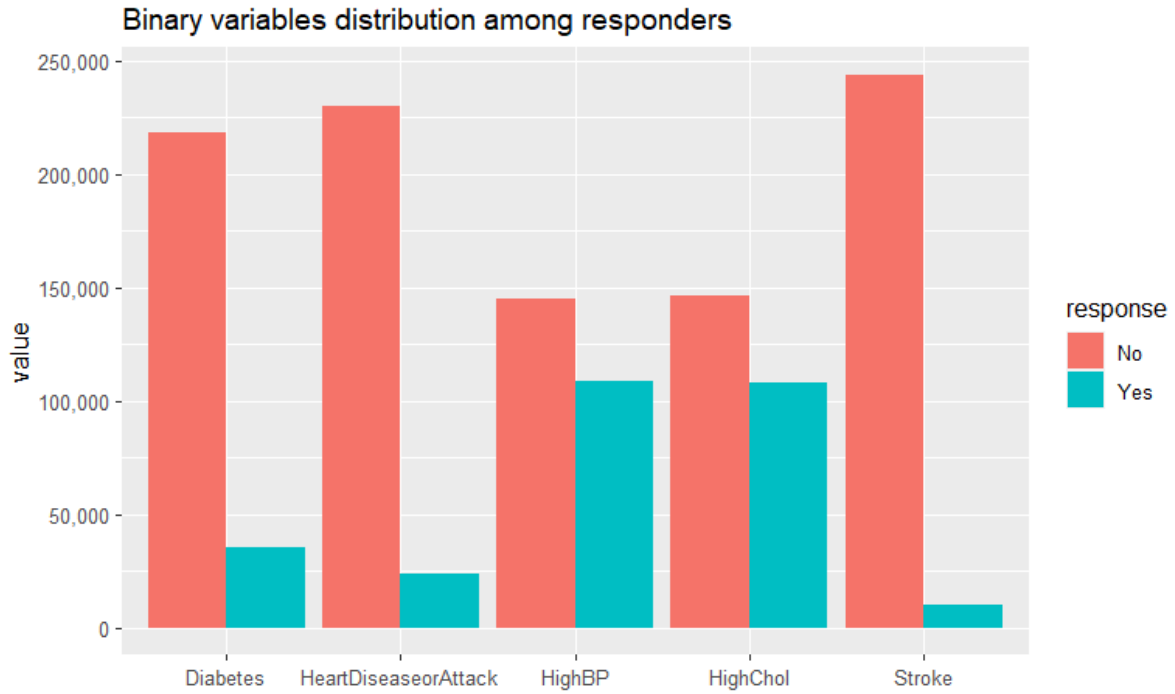
	160052 11388 14764 8495 4542 5 6 7 8 9 7622 1330 4538 809 179 10 11 12 13 14 5595 60 578 68 2587 15 16 17 18 19 4916 112 96 152 22 20 21 22 23 24 3273 663 70 56 72 25 26 27 28 29 1336 69 99 522 215 30 19400 SD = 8.7179513 Var = 76.0026750 Min. 1st Qu. Median Mean 3rd Qu. Max. 0.000 0.000 0.000 4.242 3.000 30.000
DiffWalk	Логічна змінна Таблиця спряженості FALSE TRUE 211005 42675
Sex	Логічна змінна (закодована рядком) Таблиця спряженості female male 141974 111706
Age	Порядкова змінна Таблиця спряженості 18-24 25-29 30-34 35-39 40-44 45-49 5700 7598 11123 13823 16157 19819 50-54 55-59 60-64 65-69 70-74 75-79 26314 30832 33244 32194 23533 15980 80+ 17363 SD = 3.0542204 Var = 9.3282625 Min. 1st Qu. Median Mean 3rd Qu. Max. 1.000 6.000 8.000 8.032 10.000 13.000
Education	Порядкова змінна

	Таблиця спряженості College graduate 107325 Elementary 4043 High school graduate 62750 No education 174 Some college or tech. school 69910 Some high school 9478 SD = 0.9857742 Var = 0.9717507 Min. 1st Qu. Median Mean 3rd Qu. Max. 1.00 4.00 5.00 5.05 6.00 6.00
Income	Порядкова змінна Таблиця спряженості $10,000-$15,000 $15,000-$20,000 11783 15994 $20,000-$25,000 $25,000-$35,000 20135 25883 $35,000-$50,000 $50,000-$75,000 36470 43219 $75,000 or more Less than $10,000 90385 9811 SD = 2.0711476 Var = 4.2896522 Min. 1st Qu. Median Mean 3rd Qu. Max. 1.000 5.000 7.000 6.054 8.000 8.000

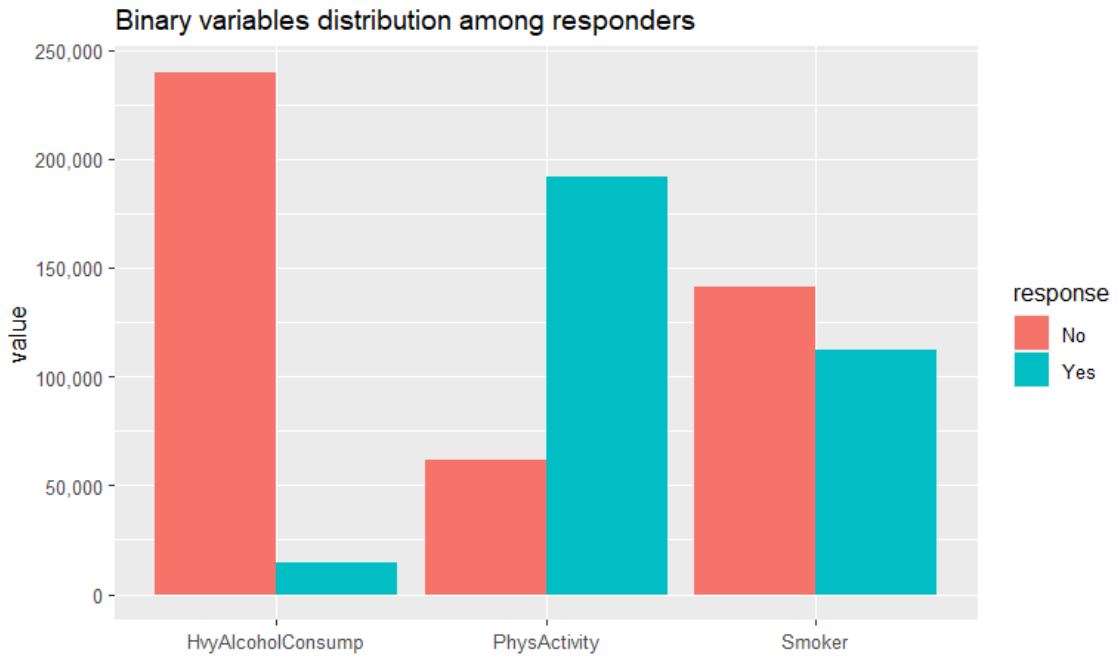
ОСОБЛИВОСТІ РОЗПОДІЛІВ ОКРЕМИХ ЗМІННИХ

1. Бінарні змінні

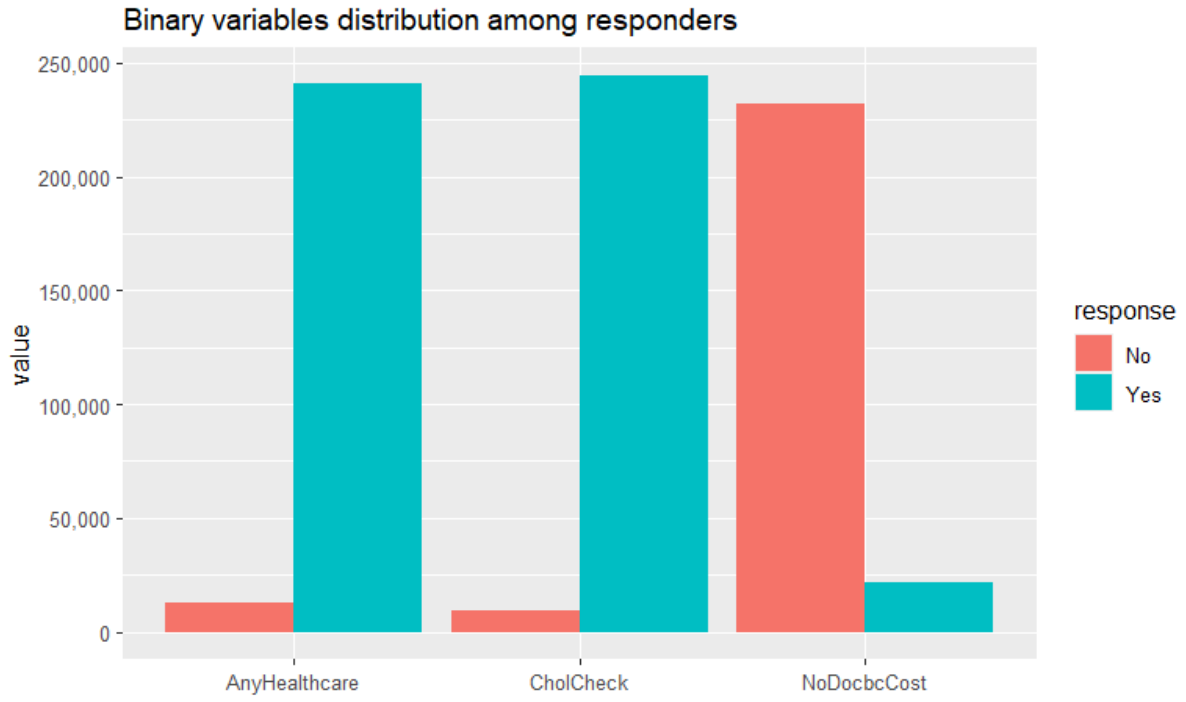
1.1 Змінні, що стосуються факторів ризику



1.2 Змінні, що стосуються способу життя

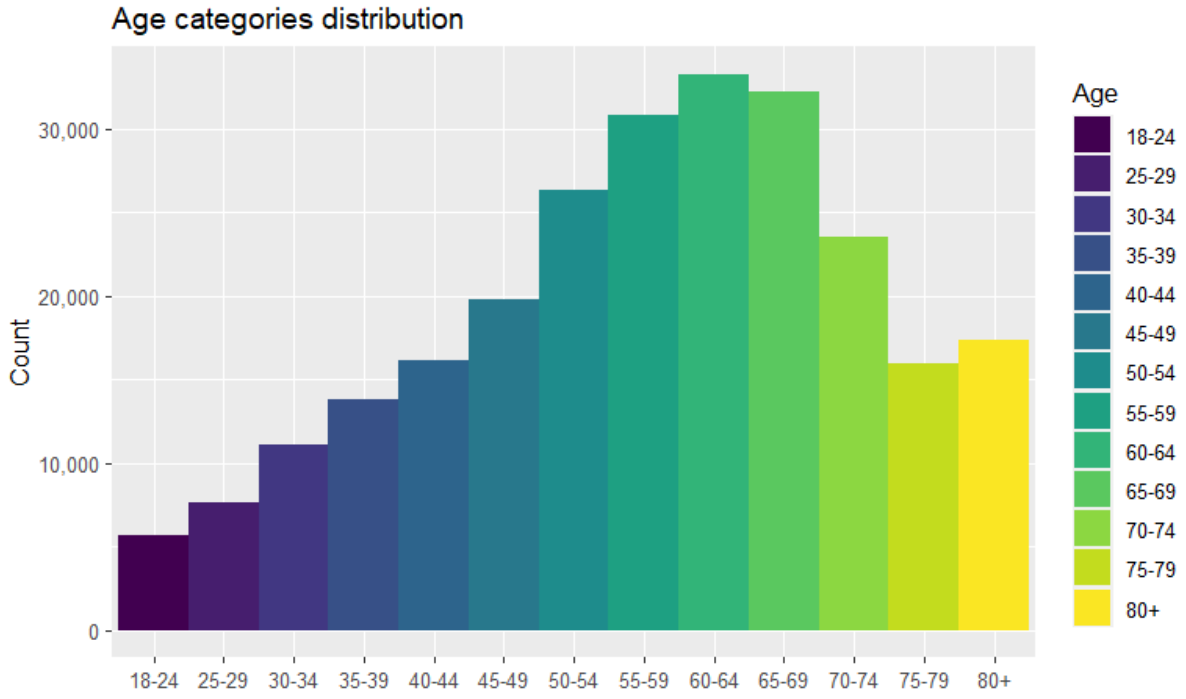


1.3 Змінні, що стосуються рівня заробітку та інше

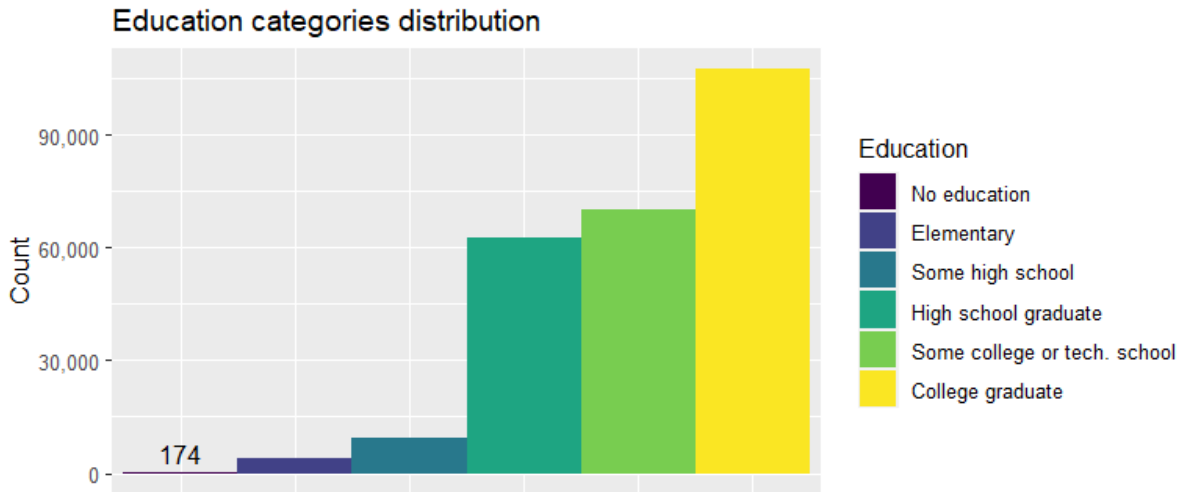


2. Впорядковані категорійні змінні

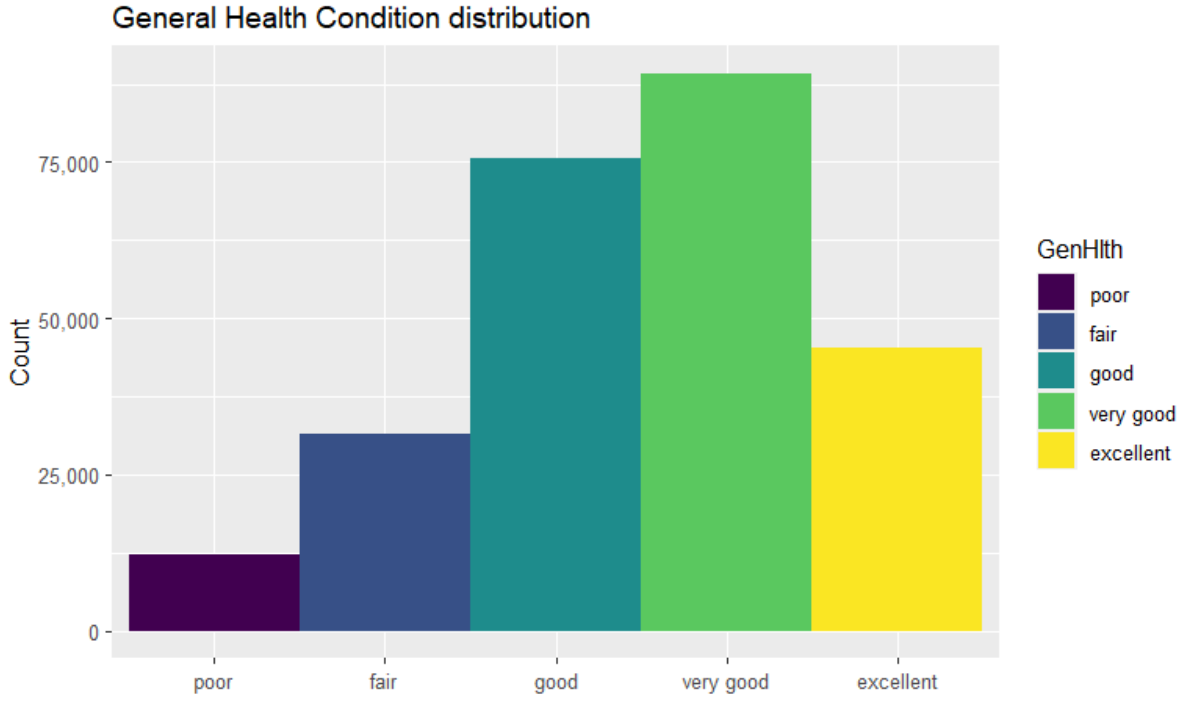
2.1 Вікова категорія



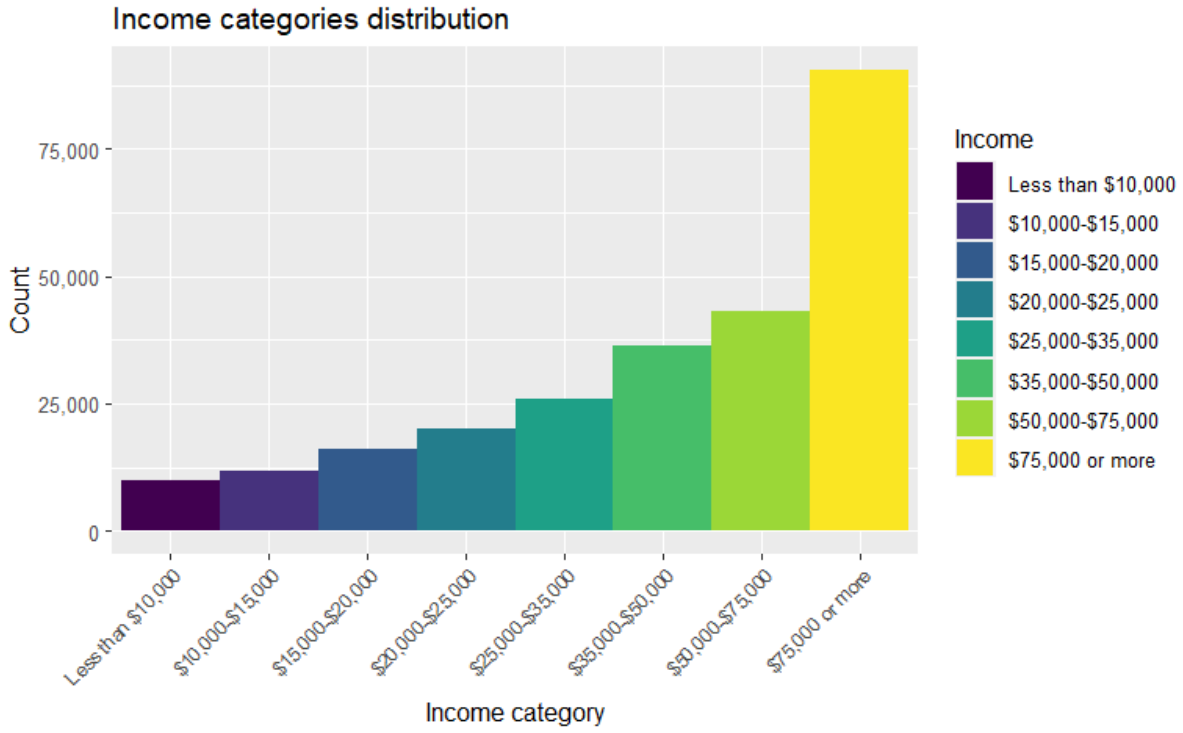
2.2 Рівень освіти



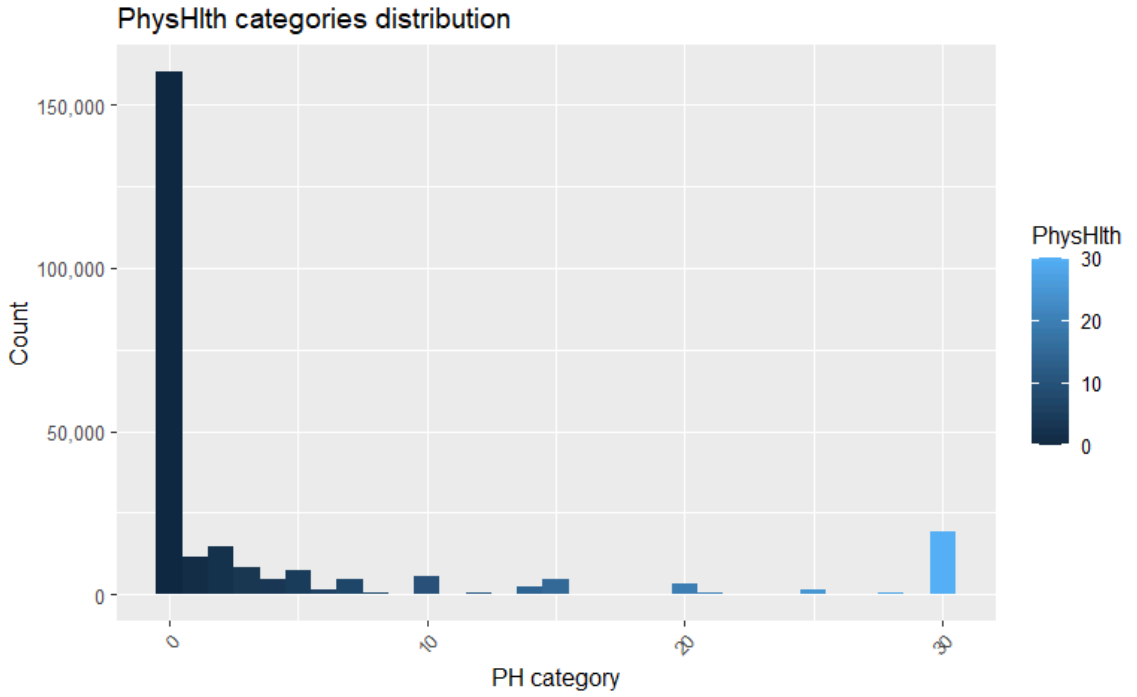
2.3 Загальний стан здоров’я

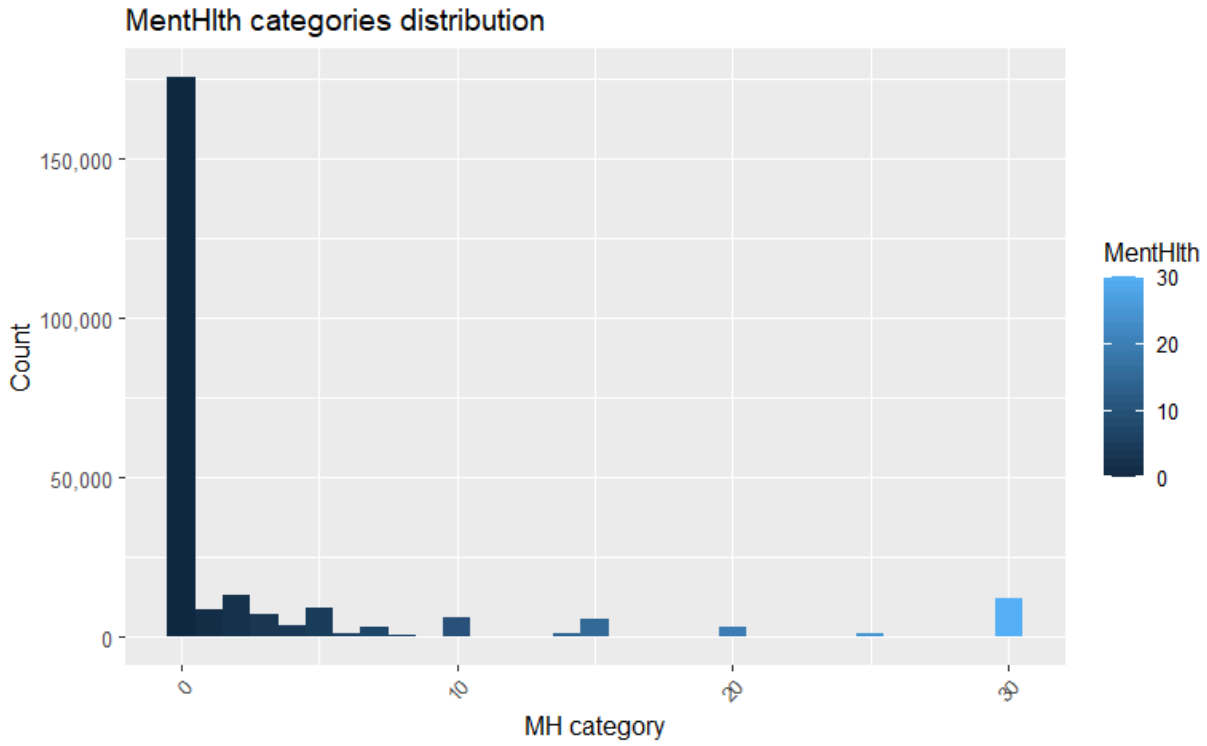


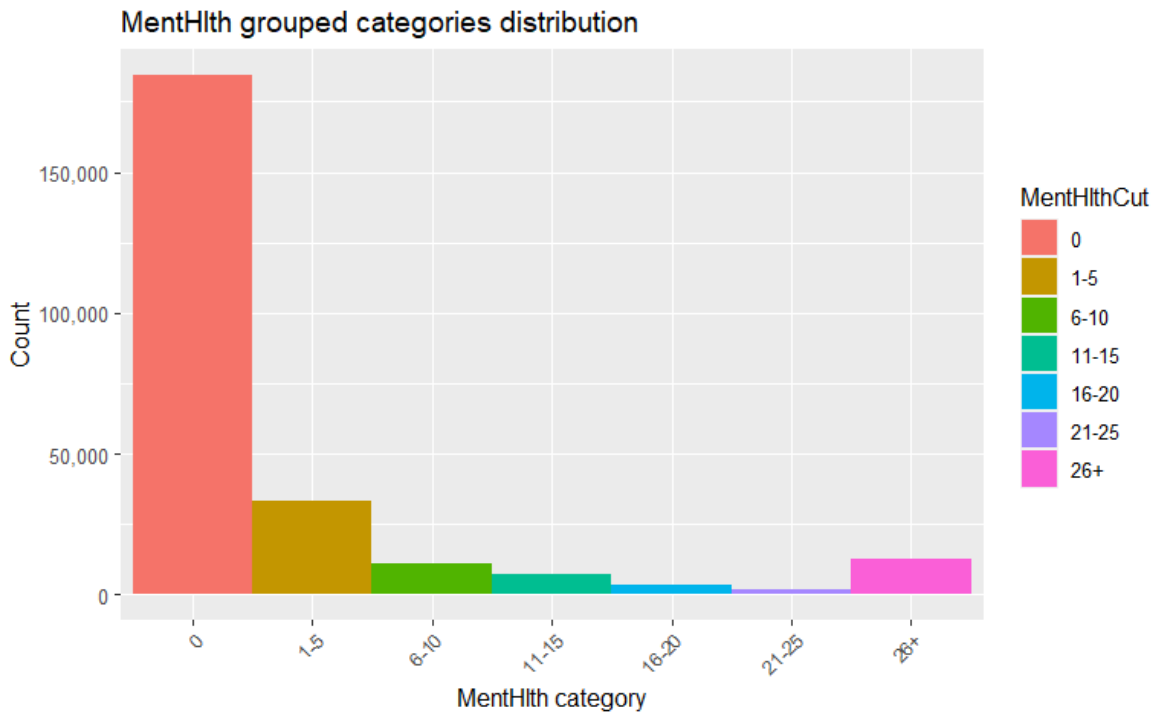
2.4 Рівень заробітку

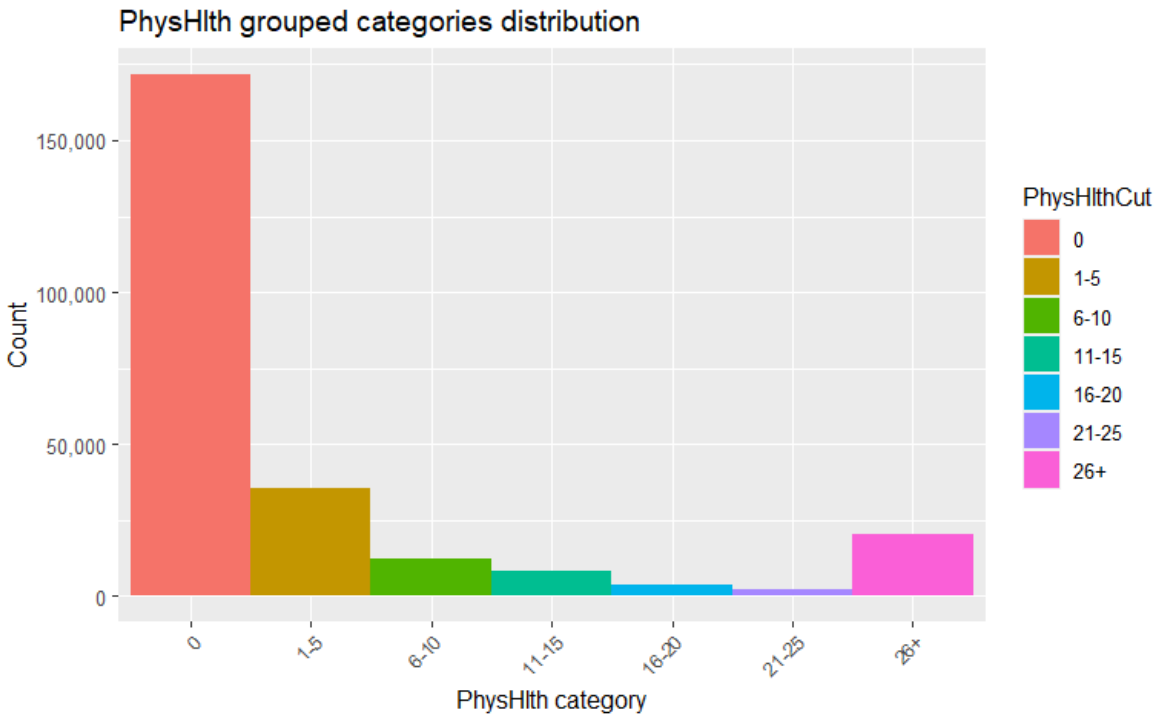


2.5 MentHlth, PhysHlth



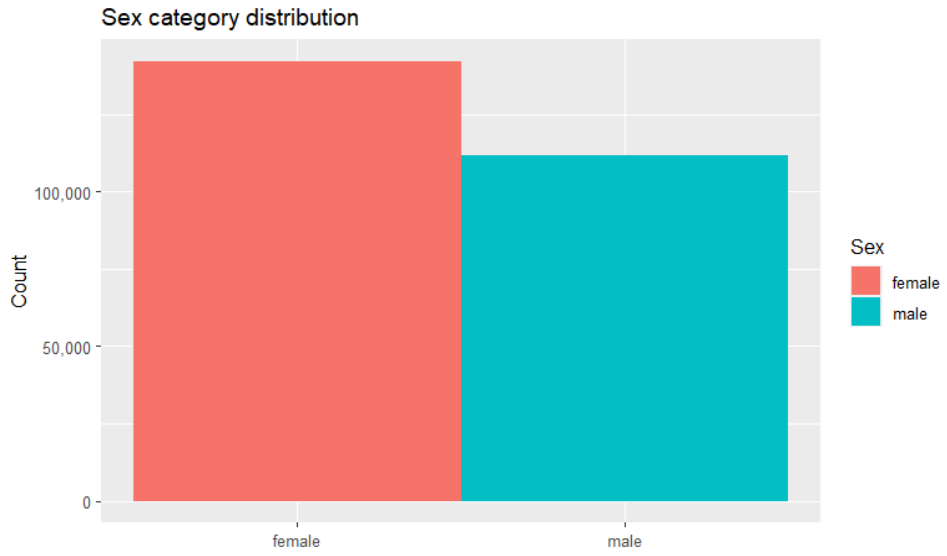



MentHlth grouped categories distribution
Count
150,000
100,000
50,000
0
0
1-5
6-10
11-15
16-20
21-25
26+
MentHlth category
MentHlthCut
0
1-5
6-10
11-15
16-20
21-25
26+


PhysHlth grouped categories distribution
Count
150,000
100,000
50,000
0
0
1-5
6-10
11-15
16-20
21-25
26+
PhysHlth category
PhysHlthCut
0
1-5
6-10
11-15
16-20
21-25
26+

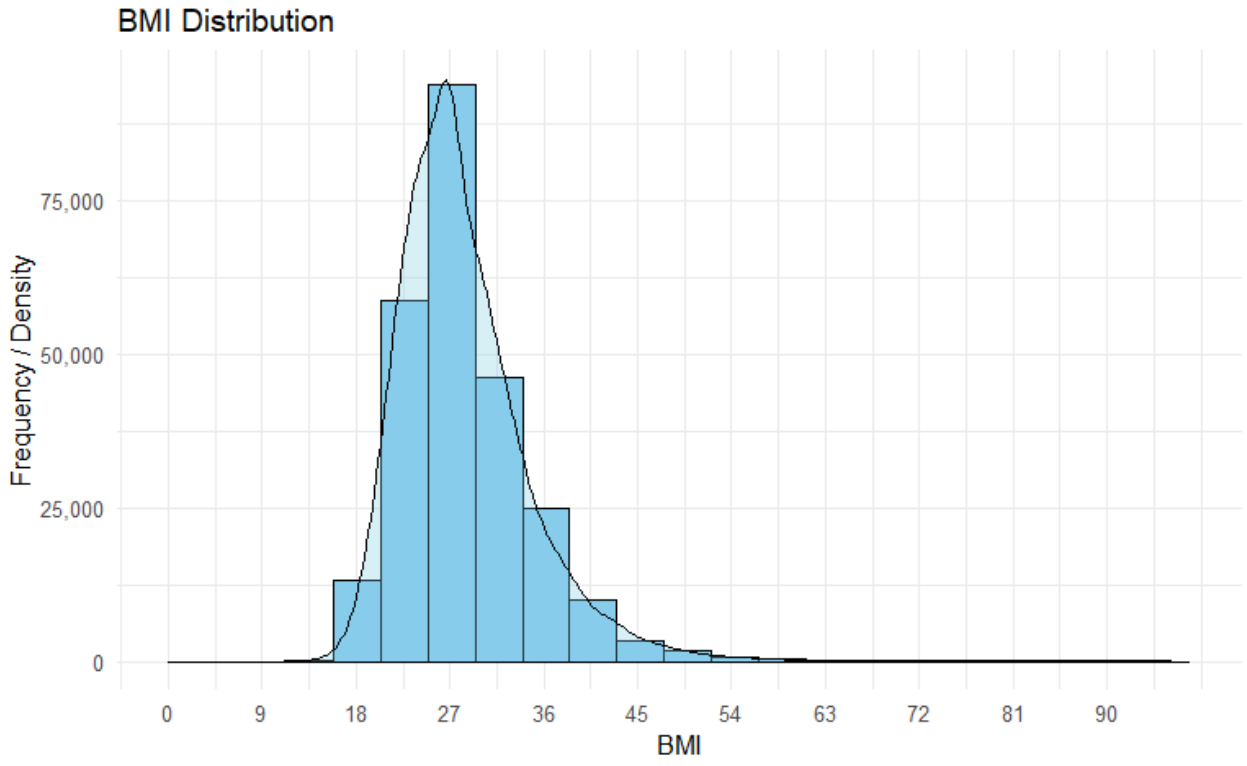
3. Невпорядковані категорійні змінні

3.1 Стать

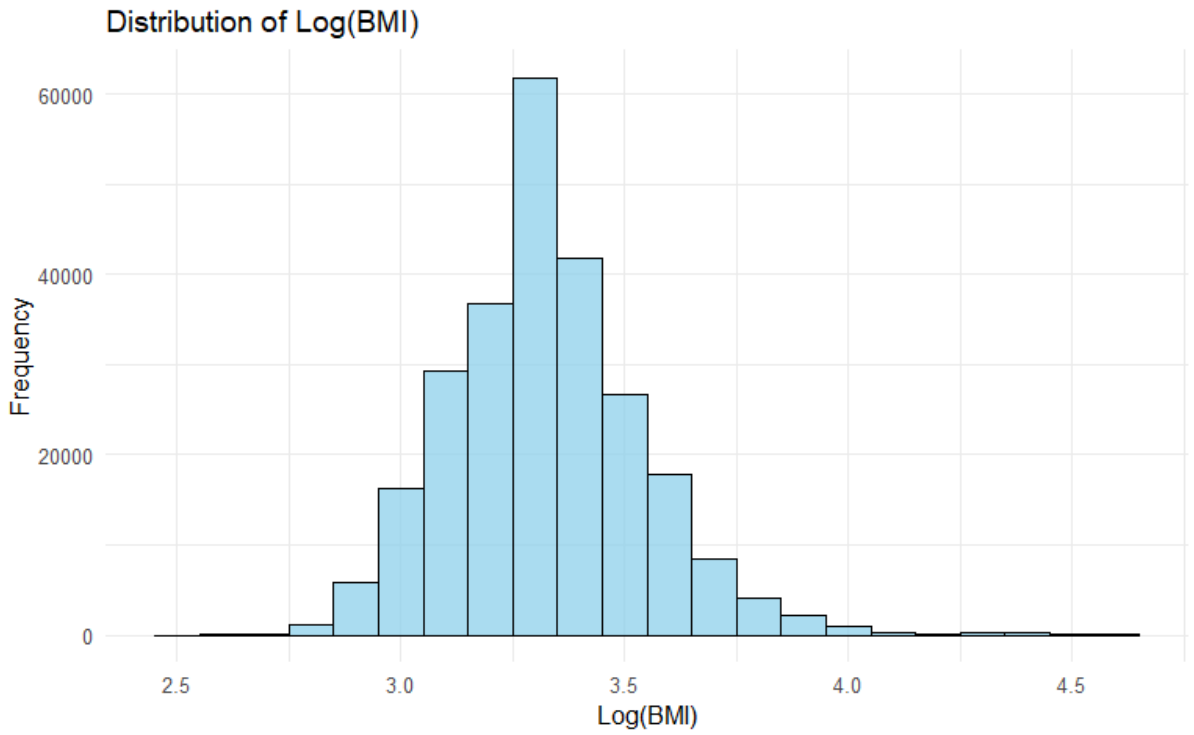


4. Неперервні змінні

4.1 ІМТ

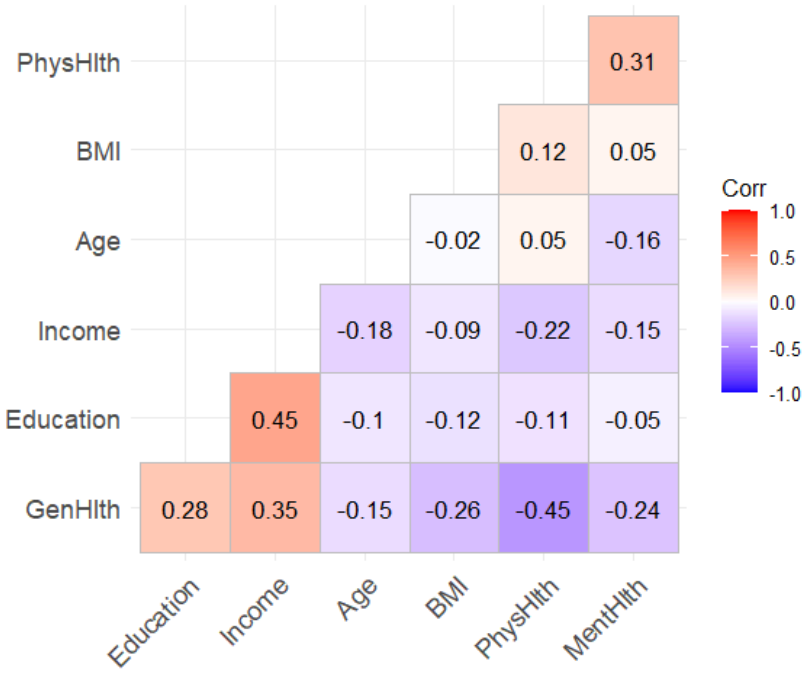


4.2 Логарифм від ІМТ



КОРЕЛЯЦІЇ

Кореляції Спірмена



З теплової карти можна побачити, що найбільш суттєвими є обернена кореляція між змінними **`PhysHlth`** та **`GenHlth`**, кореляція між змінними **`Education`** та **`Income`**.

Отже, щодо першої пари - чим вища категорія PhysHlth (PhysHlth : дні фізичної хвороби або травми за останні 30 днів за шкалою від 1 до 30), тим нижчий показник оцінки респондентом власного стану здоров'я.

Щодо другої пари, чим вищий рівень освіти - тим вищий рівень заробітної плати.

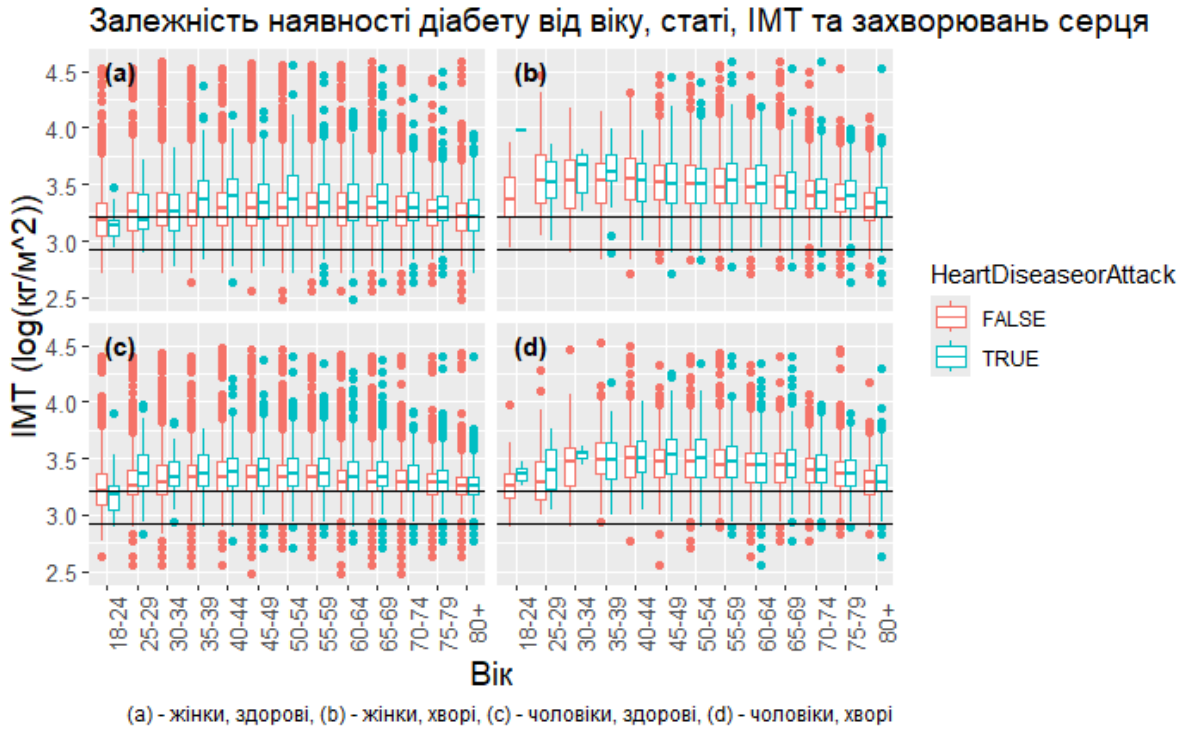
Серед інших пар можна виділити також позитивну кореляцію між змінними **`Income`** та **`GenHlth`** (0.31), **`MentHlth`** та **`PhysHlth`** (0.35).

РЕЗУЛЬТАТИ EDA

1. Перше питання

“Залежність наявності діабету від факторів ризику (підвищений рівень холестерину, наявність ожиріння, проблеми з серцево-судинною системою, вік)”

Графік №1. Залежність наявності діабету від віку, статі, ІМТ та захворювань серця



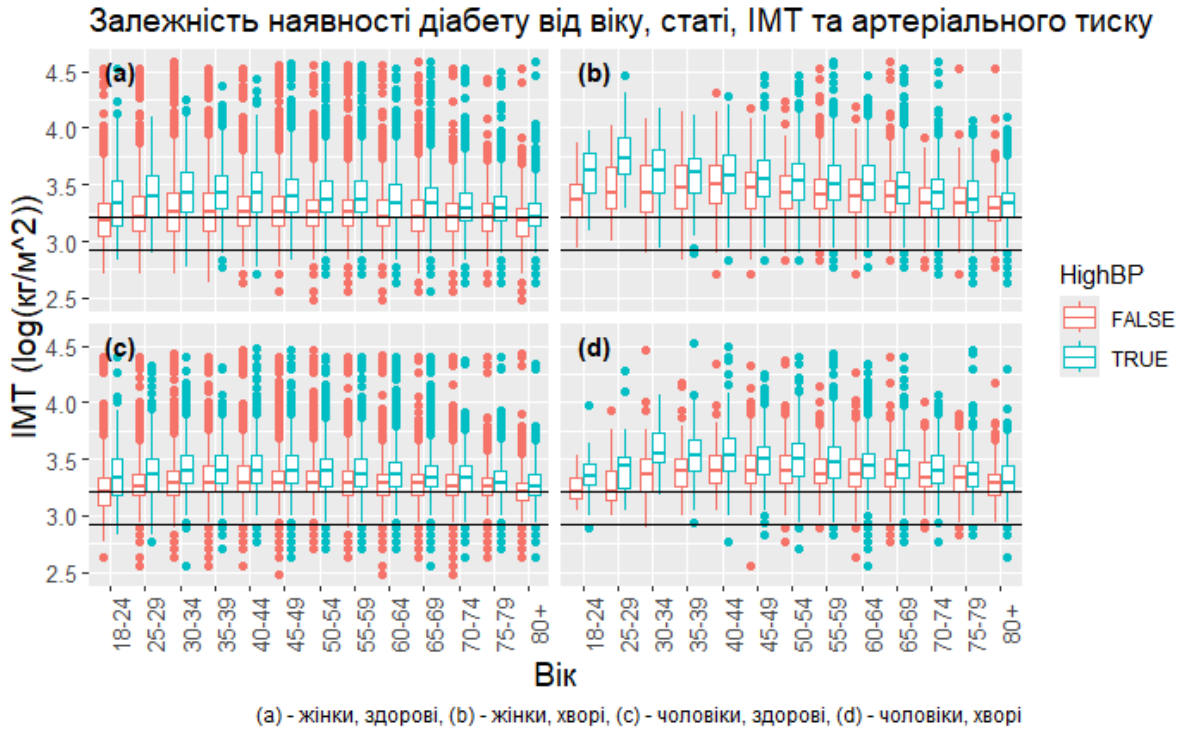
Таблиця співвідношень кількості людей за наявністю хвороб

```
, , Diabetes = FALSE, Sex = female

                     Age
HeartDiseaseorAttack 18-24 25-29 30-34 35-39 40-44 45-49 50-54 55-59 60-64 65-69 70-74 75-79   80+
               FALSE  2689  3879  5807  7271  8436  9807 12727 14572 14491 13539  9603  6653  7242
               TRUE     11    23    64    88   124   219   436   642   812   954  1003   931  1540

, , Diabetes = TRUE, Sex = female

                     Age
HeartDiseaseorAttack 18-24 25-29 30-34 35-39 40-44 45-49 50-54 55-59 60-64 65-69 70-74 75-79   80+
               FALSE    44    83   183   348   540   784  1419  1914  2445  2664  1992  1415  1222
               TRUE      1     6     8    18    36   118   223   341   523   586   561   419   518
```

```
, , Diabetes = FALSE, Sex = male

                    Age
HeartDiseaseorAttack 18-24 25-29 30-34 35-39 40-44 45-49 50-54 55-59 60-64 65-69 70-74 75-79   80+
               FALSE  2907  3533  4888  5771  6402  7792  9551 10494 10886  9438  6278  3821  3842
               TRUE     15    23    50    67   144   259   512   861  1322  1705  1508  1172  1530

, , Diabetes = TRUE, Sex = male

                    Age
HeartDiseaseorAttack 18-24 25-29 30-34 35-39 40-44 45-49 50-54 55-59 60-64 65-69 70-74 75-79   80+
               FALSE    31    49   119   240   428   724  1192  1599  2064  2360  1713   998   898
               TRUE      2     2     4    20    47   116   254   409   701   948   875   571   571
```

На графіку видно, що для хворих на діабет загалом спостерігаються більші значення основних підсумкових характеристик. Здорові респонденти жіночої статі мають вищі показники ІМТ, ніж чоловіки, тоді як показники хворих між обома статями істотно не відрізняються (за винятком вікових груп 18-24, 25-29 і 30-34). Крім того, медіанне значення ІМТ часто перевищує здорові межі, і майже всі хворі на діабет мають надмірну вагу. У людей із захворюваннями серця та без діабету ІМТ вище, ніж у людей без таких захворювань, втім аналогічні графіки для людей з діабетом показують незначну різницю. Бачимо, що люди старшого віку більш схильні до наявності захворювань серця та діабету, ніж молоді, так само бачимо, що чоловіки більш схильні до хвороб, ніж жінки. Загалом співвідношення людей із захворюваннями серця та без них трохи більше для перших при діабеті, проте більшість людей із захворюваннями серця не має діабету.

Графік №2. Залежність наявності діабету від віку, статі, ІМТ та артеріального тиску



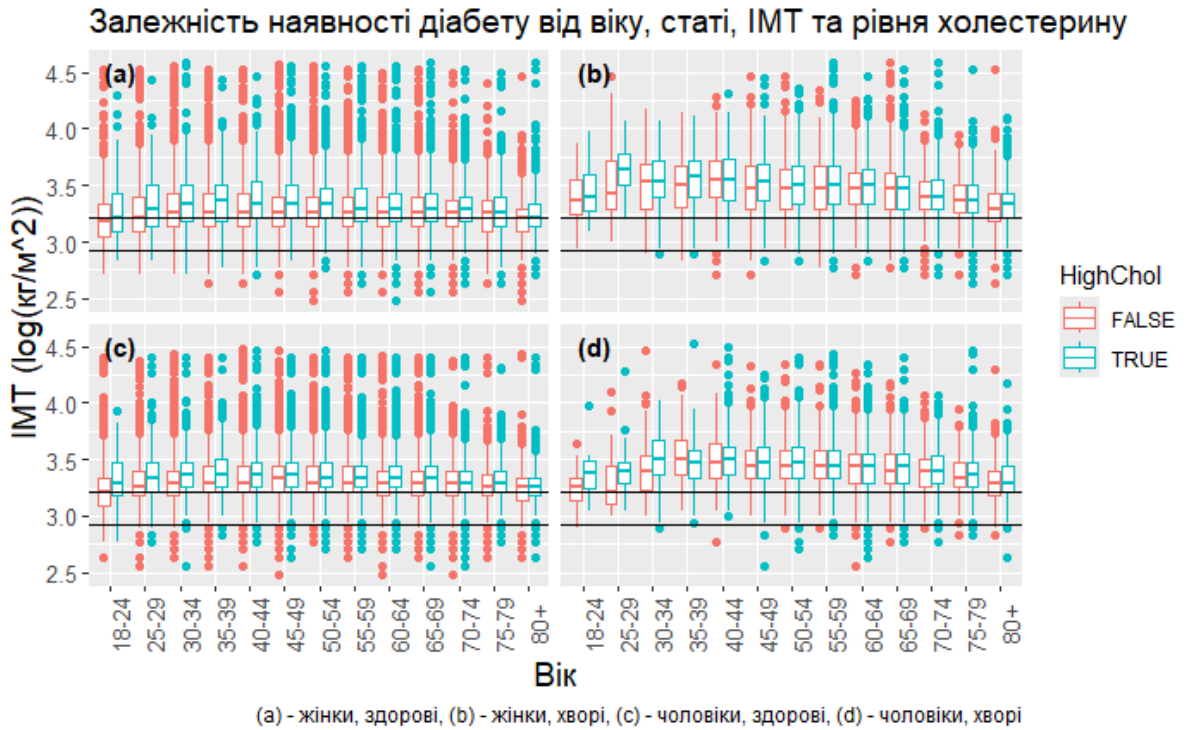
Таблиця співвідношень кількості людей за наявністю хвороб

```
, , Diabetes = FALSE, Sex = female

       Age
HighBP   18-24  25-29  30-34  35-39  40-44  45-49  50-54  55-59  60-64  65-69  70-74  75-79
80+
   FALSE   2556   3632   5274   6438   7118   7837   9526 10036   9129   7497   4743   2905
3124
   TRUE     144    270    597    921   1442   2189   3637   5178   6174   6996   5863   4679
5658

, , Diabetes = TRUE, Sex = female

       Age
HighBP   18-24  25-29  30-34  35-39  40-44  45-49  50-54  55-59  60-64  65-69  70-74  75-79
80+
   FALSE     35     63    117    189    251    331    529    590    664    637    447    320
333
   TRUE      10     26     74    177    325    571   1113   1665   2304   2613   2106   1514
1407

, , Diabetes = FALSE, Sex = male

       Age
HighBP   18-24  25-29  30-34  35-39  40-44  45-49  50-54  55-59  60-64  65-69  70-74  75-79
80+
   FALSE   2594   2994   4019   4522   4850   5615   6373   6446   6313   5128   3218   1975
2247
   TRUE     328    562    919   1316   1696   2436   3690   4909   5895   6015   4568   3018
3125

, , Diabetes = TRUE, Sex = male
```

```
       Age
HighBP   18-24  25-29  30-34  35-39  40-44  45-49  50-54  55-59  60-64  65-69  70-74  75-79
80+
  FALSE     23     28     77    105    197    301    427    511    620    680    548    348
371
  TRUE      10     23     46    155    278    539   1019   1497   2145   2628   2040   1221
1098
```

На відміну від попереднього графіку, тут спостерігається істотна різниця між ІМТ у тих, хто має високий артеріальний тиск, і тих, хто його не має: в перших він значно більший. Водночас бачимо, що на цю різницю майже не впливає наявність діабету, але у молодших вікових груп видно значний відрив між людьми з артеріальним тиском та без нього. Загалом є разюча різниця у співвідношенні людей із тиском та без нього при діабеті та за його відсутності. Але, аналогічно із попереднім випадком, тиск навряд може бути маркером діабету: навіть при ньому більшість не мають діабету.

Графік №3. Залежність наявності діабету від віку, статі, ІМТ та рівня холестерину



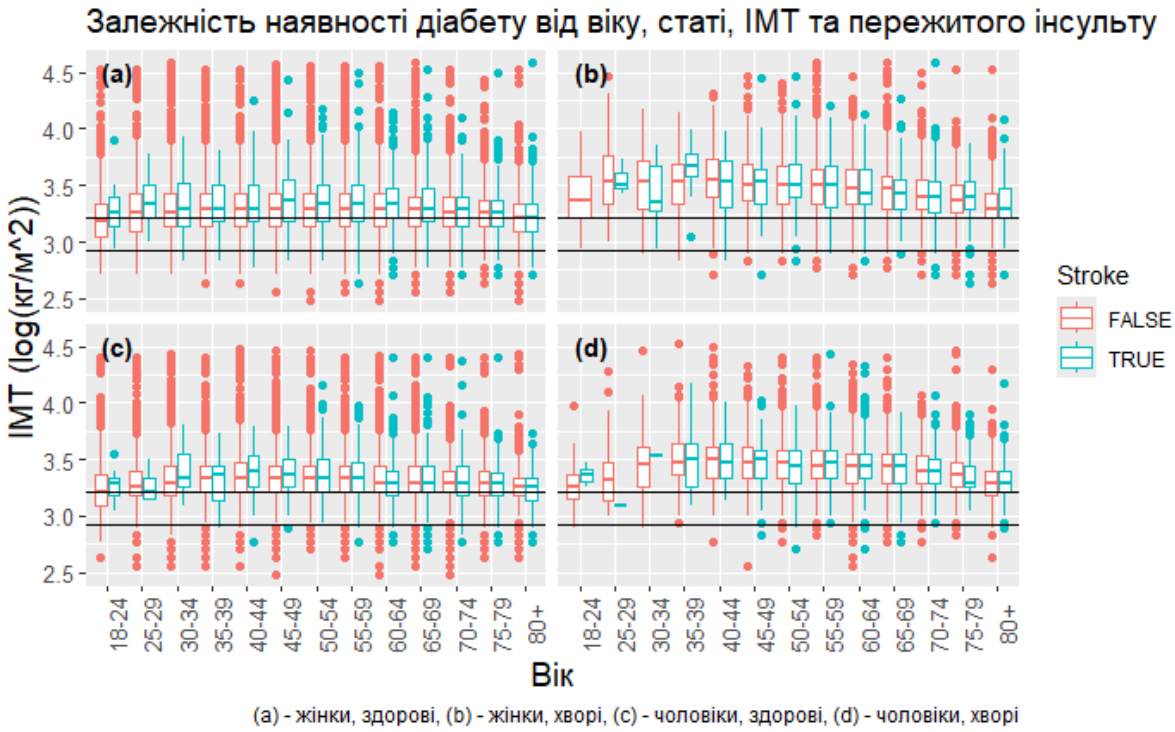
Таблиця співвідношень кількості людей за наявністю хвороб

```
, , Diabetes = FALSE, Sex = female

        Age
HighChol 18-24 25-29 30-34 35-39 40-44 45-49 50-54 55-59 60-64 65-69 70-74 75-79
80+
   FALSE   2456   3445   5077   6146   6837   7431   8947   9212   8327   7171   4900   3566
4297
   TRUE     244    457    794   1213   1723   2595   4216   6002   6976   7322   5706   4018
4485

, , Diabetes = TRUE, Sex = female

        Age
HighChol 18-24 25-29 30-34 35-39 40-44 45-49 50-54 55-59 60-64 65-69 70-74 75-79
80+
   FALSE     30     59    109    186    240    347    548    683    850    908    735    577
618
   TRUE      15     30     82    180    336    555   1094   1572   2118   2342   1818   1257
1122

, , Diabetes = FALSE, Sex = male

        Age
HighChol 18-24 25-29 30-34 35-39 40-44 45-49 50-54 55-59 60-64 65-69 70-74 75-79
80+
```

```
   FALSE  2687  3114  4039  4438  4598  5160  6060  6301  6287  5257  3493  2394
2789
   TRUE    235   442   899  1400  1948  2891  4003  5054  5921  5886  4293  2599
2583

, , Diabetes = TRUE, Sex = male

        Age
HighChol 18-24 25-29 30-34 35-39 40-44 45-49 50-54 55-59 60-64 65-69 70-74 75-79
80+
   FALSE    25    27    70   106   197   316   482   646   891  1015   853   563
579
   TRUE      8    24    53   154   278   524   964  1362  1874  2293  1735  1006
890
```

На графіку можна побачити, що ті, хто не має діабету, але має високий рівень холестерину, має також і вищий ІМТ порівняно з тими, у кого він не є високим. Водночас коробкові графіки у хворих на діабет приблизно однакові як при високому рівню холестерину, так і без нього (за винятком деяких молодих груп). Знову спостерігаємо, що ті, хто має діабет, найімовірніше має і стороннє захворювання (у даному випадку - високий рівень холестерину), але не навпаки.

Графік №4. Залежність наявності діабету від віку, статі, ІМТ та пережитого інсульту



Таблиця співвідношень кількості людей за наявністю хвороб

```
, , Diabetes = FALSE, Sex = female

       Age
Stroke   18-24  25-29  30-34  35-39  40-44  45-49  50-54  55-59  60-64  65-69  70-74  75-79
80+
  FALSE   2691   3883   5820   7290   8436   9870 12865 14785 14851 13971 10046   7083
8017
  TRUE       9     19     51     69    124    156    298    429    452    522    560    501
765

, , Diabetes = TRUE, Sex = female

       Age
Stroke   18-24  25-29  30-34  35-39  40-44  45-49  50-54  55-59  60-64  65-69  70-74  75-79
80+
  FALSE     45     86    185    353    546    844   1518   2041   2696   2959   2278   1622
1507
  TRUE       0      3      6     13     30     58    124    214    272    291    275    212
233

, , Diabetes = FALSE, Sex = male

       Age
Stroke   18-24  25-29  30-34  35-39  40-44  45-49  50-54  55-59  60-64  65-69  70-74  75-79
80+
  FALSE   2912   3550   4915   5792   6480   7940   9859 11069 11779 10657   7318   4593
4838
  TRUE      10      6     23     46     66    111    204    286    429    486    468    400
534

, , Diabetes = TRUE, Sex = male

       Age
```

```
Stroke   18-24  25-29  30-34  35-39  40-44  45-49  50-54  55-59  60-64  65-69  70-74  75-79
80+
  FALSE     31     50    120    251    466    797   1350   1852   2504   2987   2334   1391
1265
  TRUE       2      1      3      9      9     43     96    156    261    321    254    178
204
```

Даний графік показує малу інформативність для молодих людей, але для решти можемо зробити висновки, що пережитий інсульт не є гарантією діабету, і навіть при діабеті більшість не переживали його, хоча він дещо збільшує ймовірність діабету.

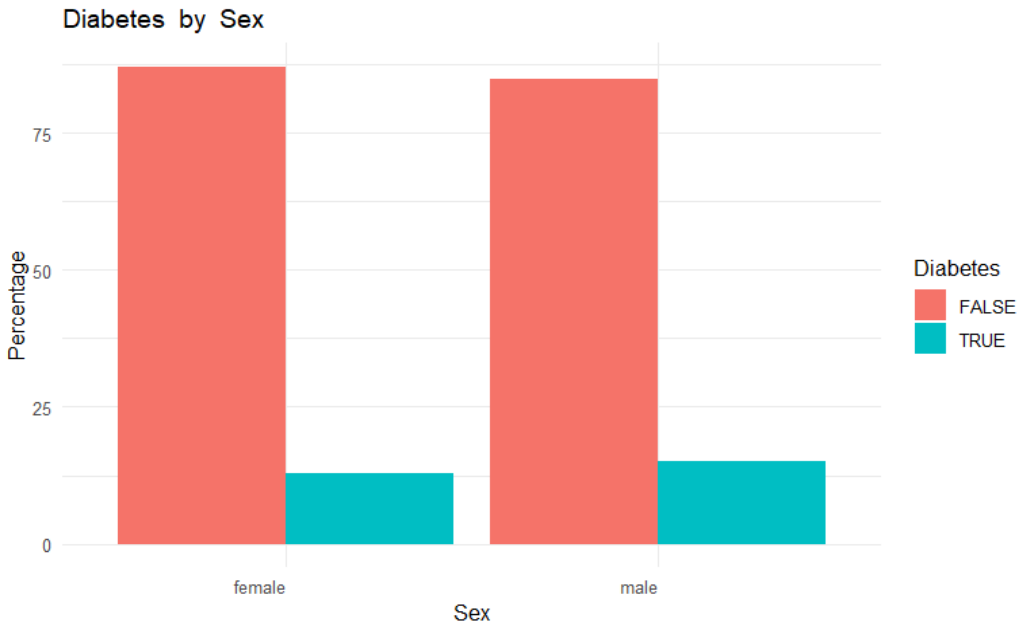
На завершення проаналізуємо детальніше співвідношення діабету та інших факторів окремо.

Ізольоване дослідження залежностей між наявністю діабету й окремими факторами:

Стать

```
       Sex
Diabetes    female      male
  FALSE  0.8703213 0.8483967
  TRUE   0.1296787 0.1516033
```

Візуалізація №1. Залежність діабету від статі.

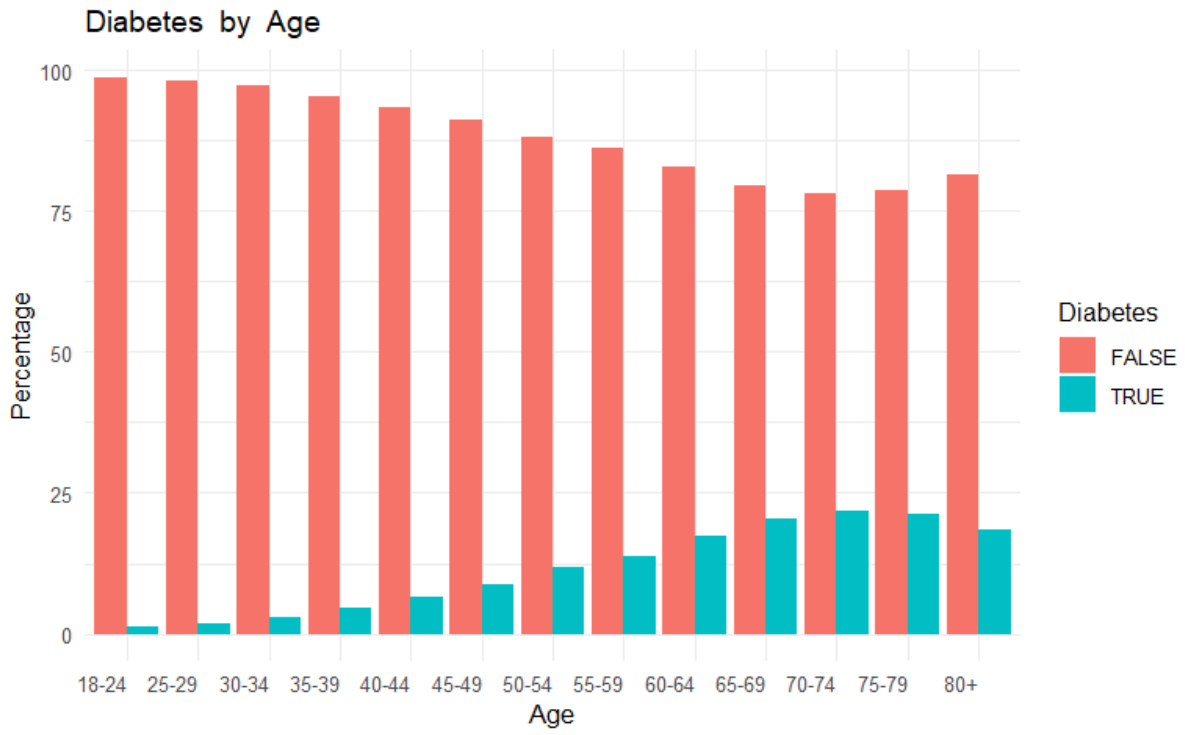


Залежності немає.

Вік

```
       Age
Diabetes      18-24      25-29      30-34      35-39      40-44      45-49
   FALSE 0.98631579 0.98157410 0.97177021 0.95471316 0.93495080 0.91210455
   TRUE  0.01368421 0.01842590 0.02822979 0.04528684 0.06504920 0.08789545
       Age
Diabetes      50-54      55-59      60-64      65-69      70-74      75-79
   FALSE 0.88264802 0.86173456 0.82754783 0.79629745 0.78154082 0.78704631
   TRUE  0.11735198 0.13826544 0.17245217 0.20370255 0.21845918 0.21295369
       Age
Diabetes        80+
   FALSE 0.81518171
   TRUE  0.18481829
```

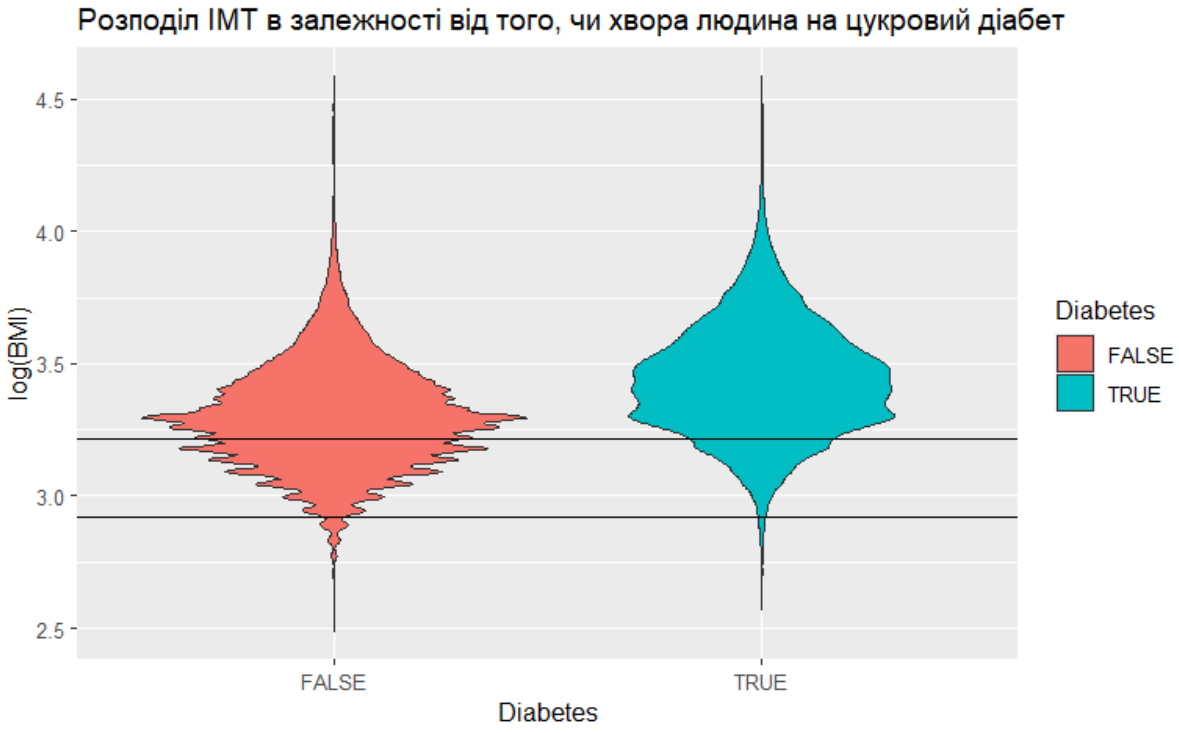
Візуалізація №2. Залежність діабету від віку.

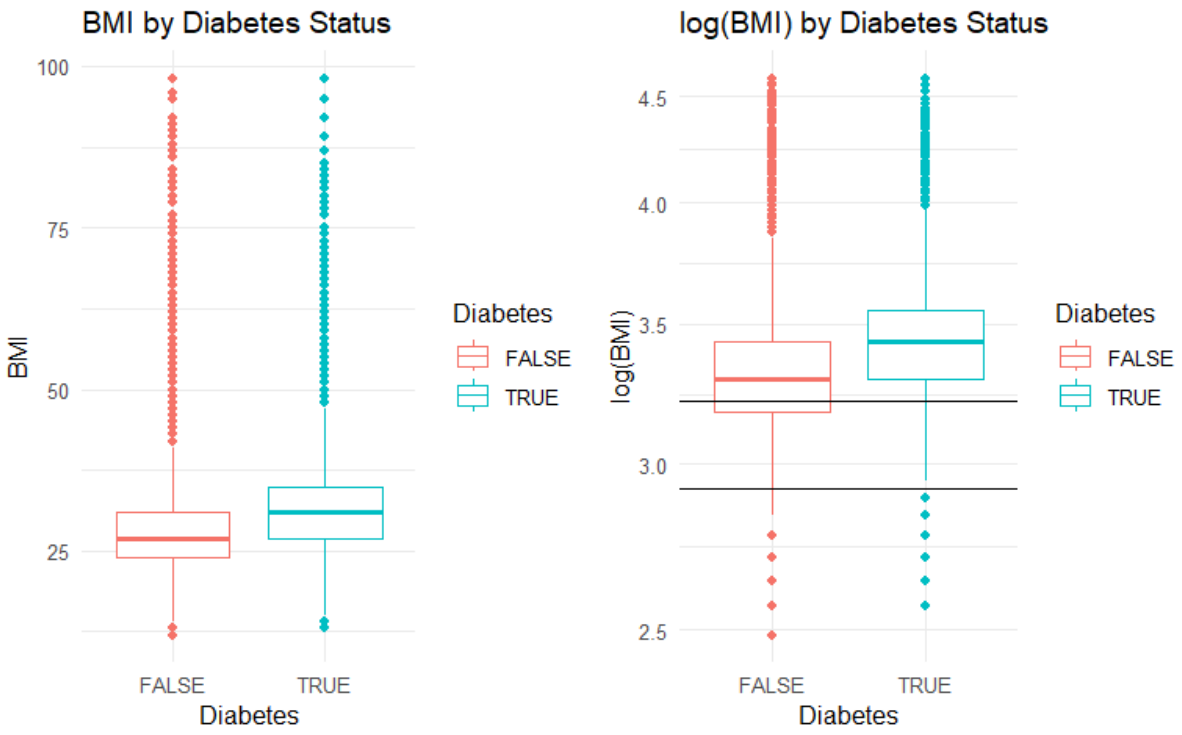


Найбільша частка діабетиків серед вікової категорії 70-74 роки, найменша ж серед молодих людей 18-24 років. Як можна бачити з графіку - чим вища вікова категорія (чим старші люди), тим вища частка хворих на діабет в цій категорії і нижча частка здорових. Має місце залежність.

Візуалізація №3. Залежність діабету від ІМТ (особливості розподілу ІМТ серед групи людей не хворих на діабет та хворих на діабет).

ІМТ





Захворювання серця

```
       HeartDiseaseorAttack
Diabetes     FALSE      TRUE
   FALSE 0.8804632 0.6702800
   TRUE  0.1195368 0.3297200
```

Високий тиск

```
       HighBP
Diabetes      FALSE       TRUE
   FALSE 0.93964833 0.75554310
   TRUE  0.06035167 0.24445690
```

Високий рівень холестерину

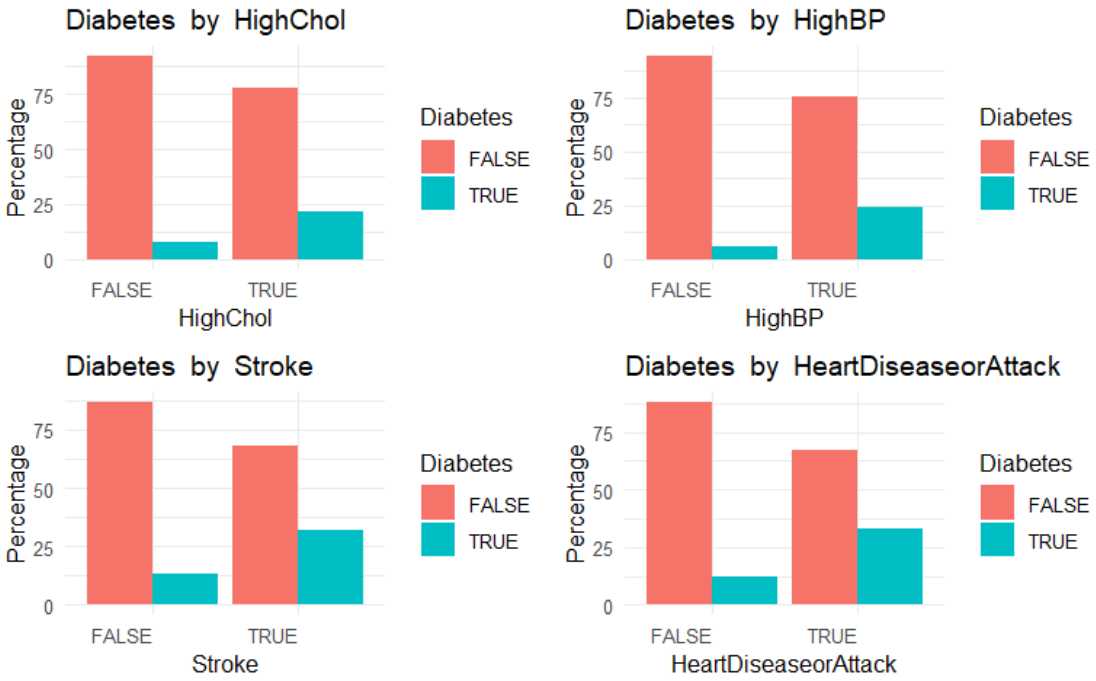
```
       HighChol
Diabetes      FALSE       TRUE
   FALSE 0.92018564 0.77985147
   TRUE  0.07981436 0.22014853
```

Пережитий інсульт

```
       Stroke
Diabetes     FALSE      TRUE
   FALSE 0.8682022 0.6824718
   TRUE  0.1317978 0.3175282
```

Візуалізація відносно відповідних категорій:

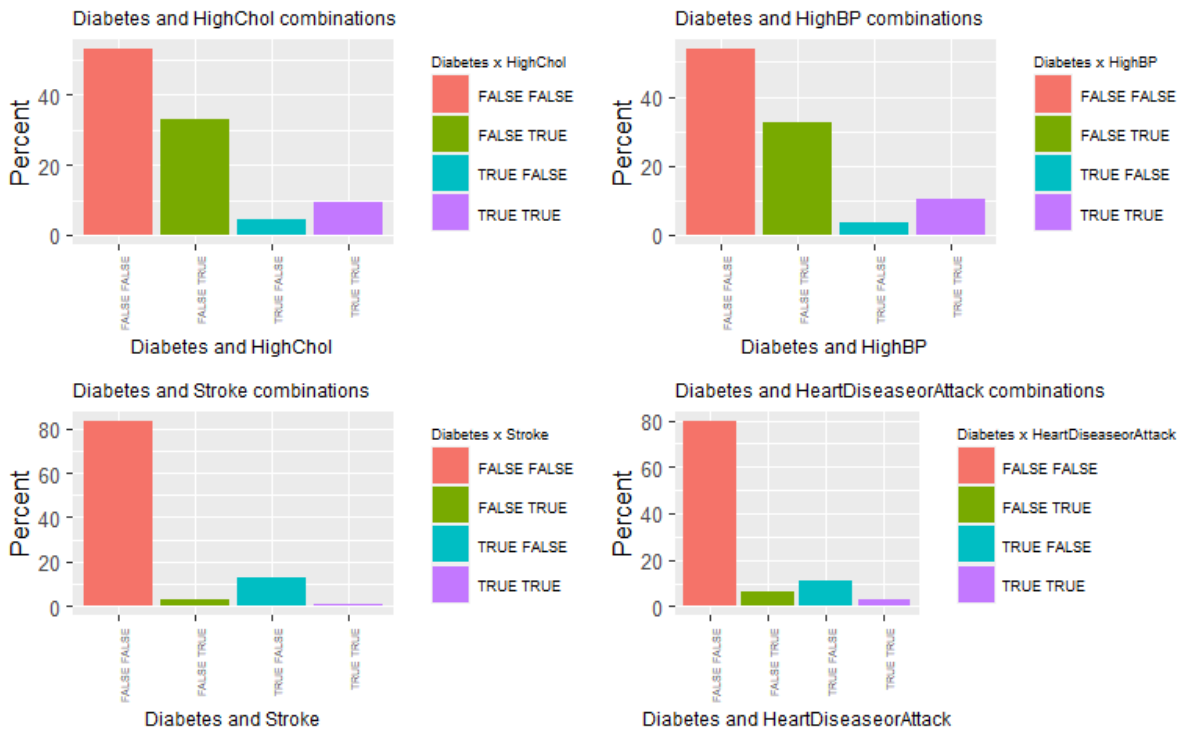
Візуалізація №4. Залежність діабету від факторів ризику



Нумерація: зверху вниз, зліва направо.

1. Графік #1. 22% проти 8% - різниця 14%.
2. Графік #2. 24% проти 6% - різниця 18%. Різниця більша на 4% в порівнянні з високим холестерином. Можна перевірити вплив HighChol x HighBP на наявність діабету.
3. Графік #3. 32% проти 13% - різниця 19%.
4. Графік #4. 32% проти 12% - різниця 20%.

Візуалізація відносно всієї вибірки:

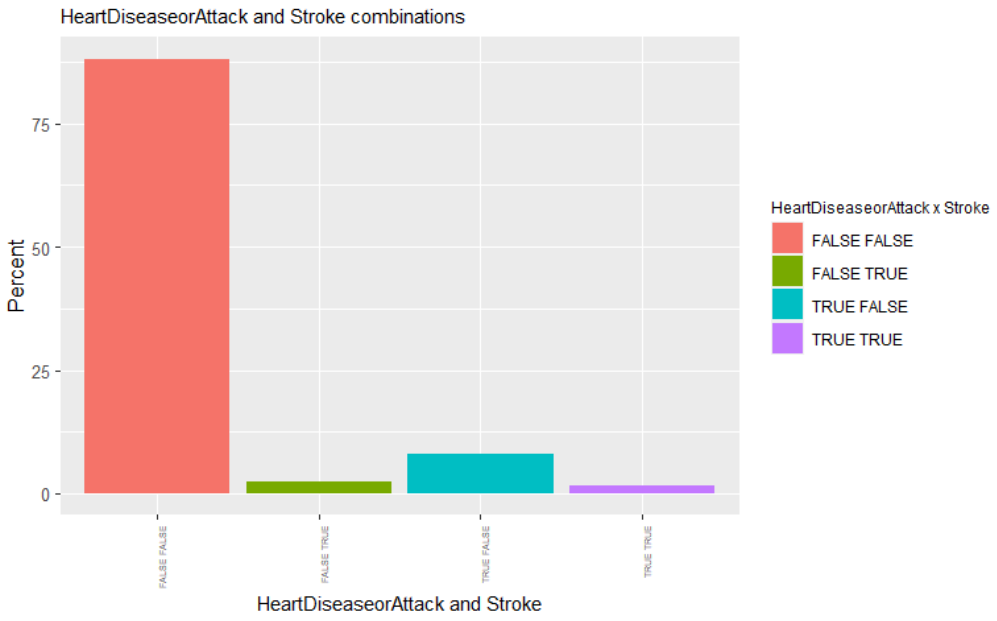


Помічаємо, що візуалізації таблиць для змінних **Stroke** та **HeartDiseaseorAttack**, **HighBP** та **HighChol** є дещо схожими, тому є сенс перевірити залежність між цими двома змінними окремо.

Зв’язок між схожими змінними.

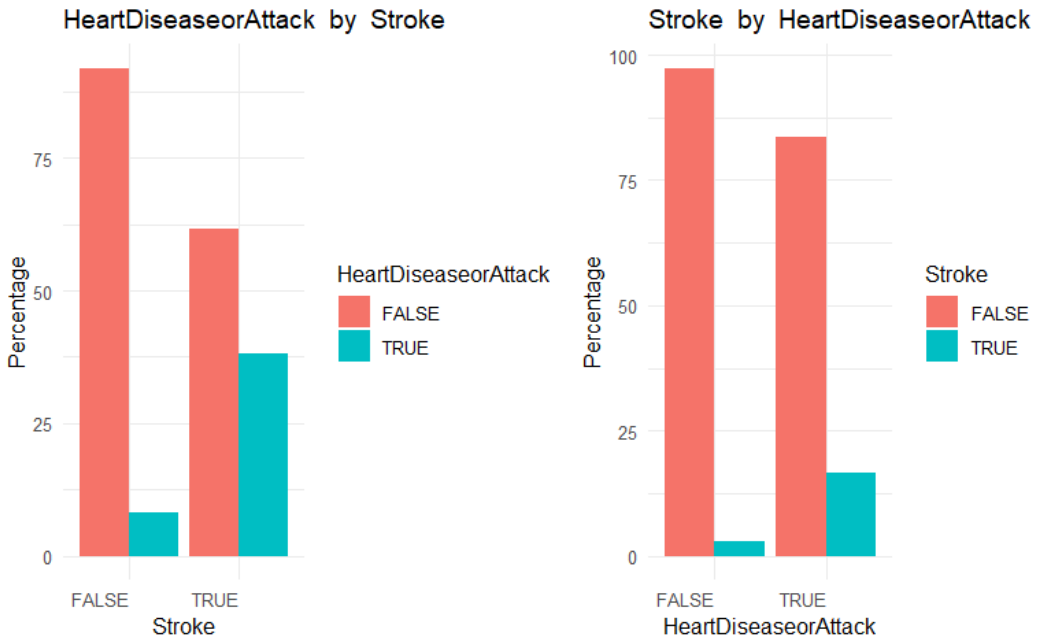
Зв'язок між змінними HeartDiseaseorAttack і Stroke.

Комбінації змінних **Stroke** та **HeartDiseaseorAttack**:



В датасеті найбільша частка людей не хворих. На другому місці частка людей які хворі на серцево-судинні захворювання, або пережили інфаркт і не мали інсульт (менше 10%).

Візуалізація залежності цих змінних.



Таблиця до першого графіку:

Stroke <lgl>	HeartDiseaseorAttack <lgl>	Count <int>	Percent <dbl>
FALSE	FALSE	223432	91.800746
FALSE	TRUE	19956	8.199254
TRUE	FALSE	6355	61.746988
TRUE	TRUE	3937	38.253012

Серед людей в яких не було інсульту, частка людей які пережили серцевий напад або мають серцево-судинні захворювання менша ніж серед людей, які пережили інсульт. (8% проти 38%)

Таблиця до другого графіку:

HeartDiseaseorAttack <lgl>	Stroke <lgl>	Count <int>	Percent <dbl>
FALSE	FALSE	223432	97.234395
FALSE	TRUE	6355	2.765605
TRUE	FALSE	19956	83.522371
TRUE	TRUE	3937	16.477629

Зворотне порівняння. Серед людей в яких не було серцевого нападу/в яких немає серцево-судинних захворювань, частка людей в яких був інсульт менша ніж серед протилежної категорії (2,7% проти 16,5%).

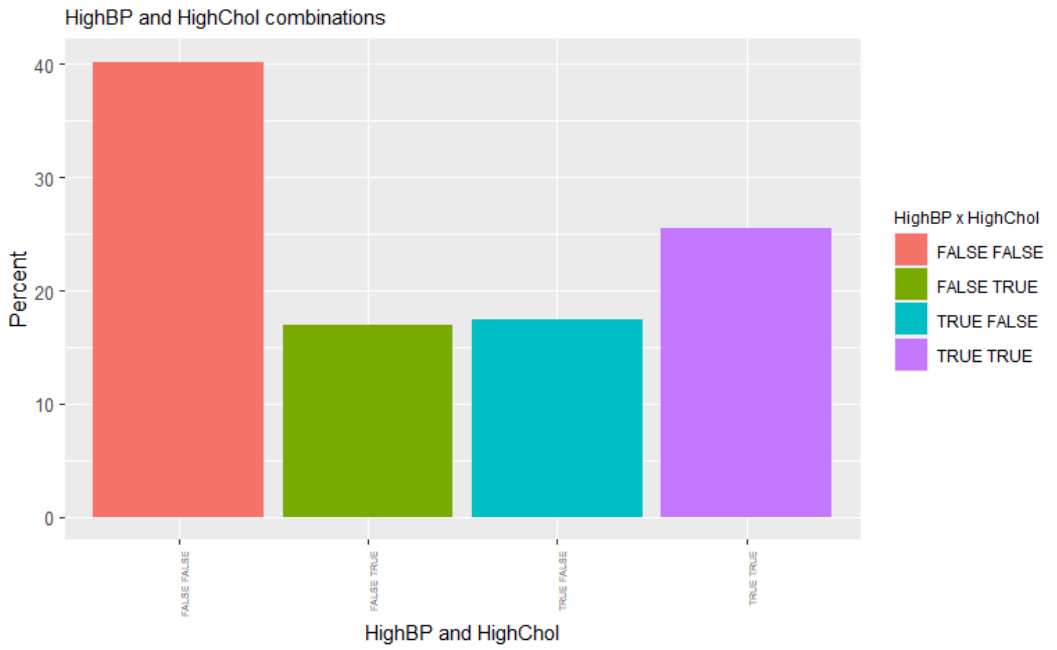
Варто додати, що кількість людей, в яких був інсульт (але немає серцево судинних захворювань - 6355) менша за кількість людей, які мають серцево судинні захворювання, але не мали інсульту (їх 19956).

Має місце зв'язок між цими двома бінарними змінними.

Зв'язок між змінними **HighBP** і **HighChol**.

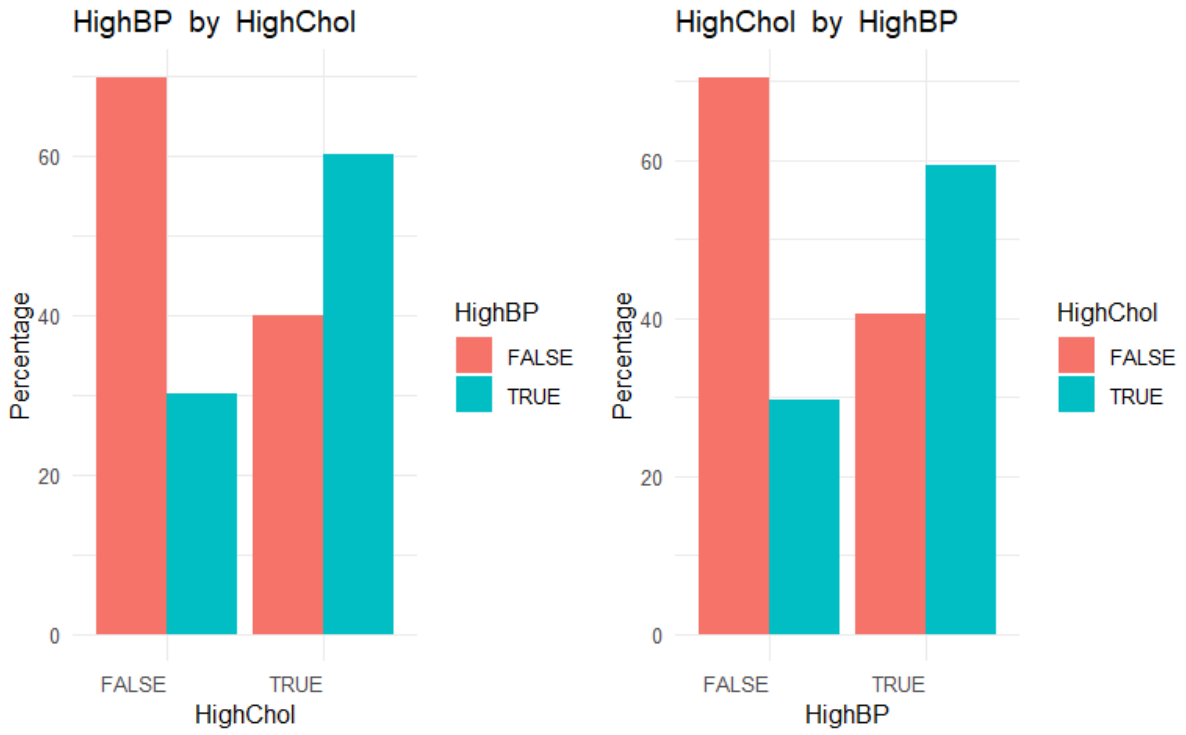
Розглянемо пару змінних **HighBP** та **HighChol**:

Візуалізація комбінацій цих змінних:



Як бачимо з датасету, найбільшу частку складають люди в яких немає проблем ні з високим тиском, ні з високим рівнем холестирину. Але якщо взяти частку людей в яких є хоч одна з цих проблем (що складає 60% серед всіх респондентів в датасеті), вона перевищить частку здорових людей на 20%.

Візуалізація залежності цих змінних:



HighChol <lgl>	HighBP <lgl>	Count <int>	Percent <dbl>
FALSE	FALSE	101920	69.76569
FALSE	TRUE	44169	30.23431
TRUE	FALSE	42931	39.90204
TRUE	TRUE	64660	60.09796

Серед людей з високим рівнем холестерину, частка людей з високим тиском значно перевищує частку людей, в яких немає проблем з високим тиском (на 20%). А в порівнянні з часткою людей з високим тиском серед людей з низьким рівнем холестерину різниця складає 30%. Має місце залежність.

HighBP <lgl>	HighChol <lgl>	Count <int>	Percent <dbl>
FALSE	FALSE	101920	70.36196
FALSE	TRUE	42931	29.63804
TRUE	FALSE	44169	40.58569
TRUE	TRUE	64660	59.41431

Аналогічну картину бачимо і в оберненому напрямку. Отже має місце залежність в “обидві сторони”.

Отже, щодо цього питання можна підбити попередні підсумки:

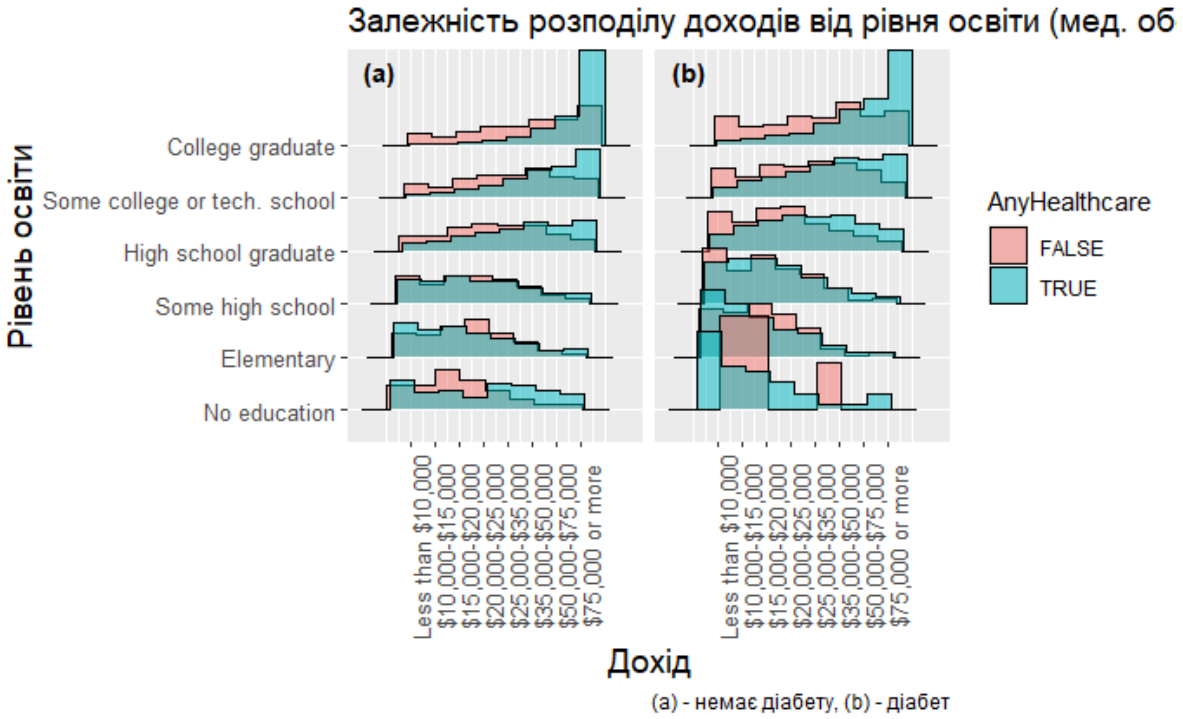
- вік (по мірі збільшення) та надмірність ваги є основними факторами наявності діабету;
- у жодному випадку побічне захворювання не може однозначно свідчити про діабет;
- деякі побічні захворювання (фактори) є тісно пов’язаними одне з одними, тому розглядаючи ізольований вплив цих факторів на наявність у людини діабету (для прогнозування наявності у людини діабету) можна дійти до хибних висновків.

- побічні захворювання дещо збільшують ймовірність діабету, проте більшість все одно його не мають;
- хворі на діабет найчастіше мають високий рівень тиску та холестерину, особливо, - люди з віком 40+, водночас ці захворювання більш поширені серед людей без діабету;
- при дослідженні з урахуванням багатьох факторів на одному графіку, помітно, що чоловіки більш схильні до захворювання на діабет, але при ізольованому дослідженні ця залежність майже не прослідковувалася.

2. Друге питання

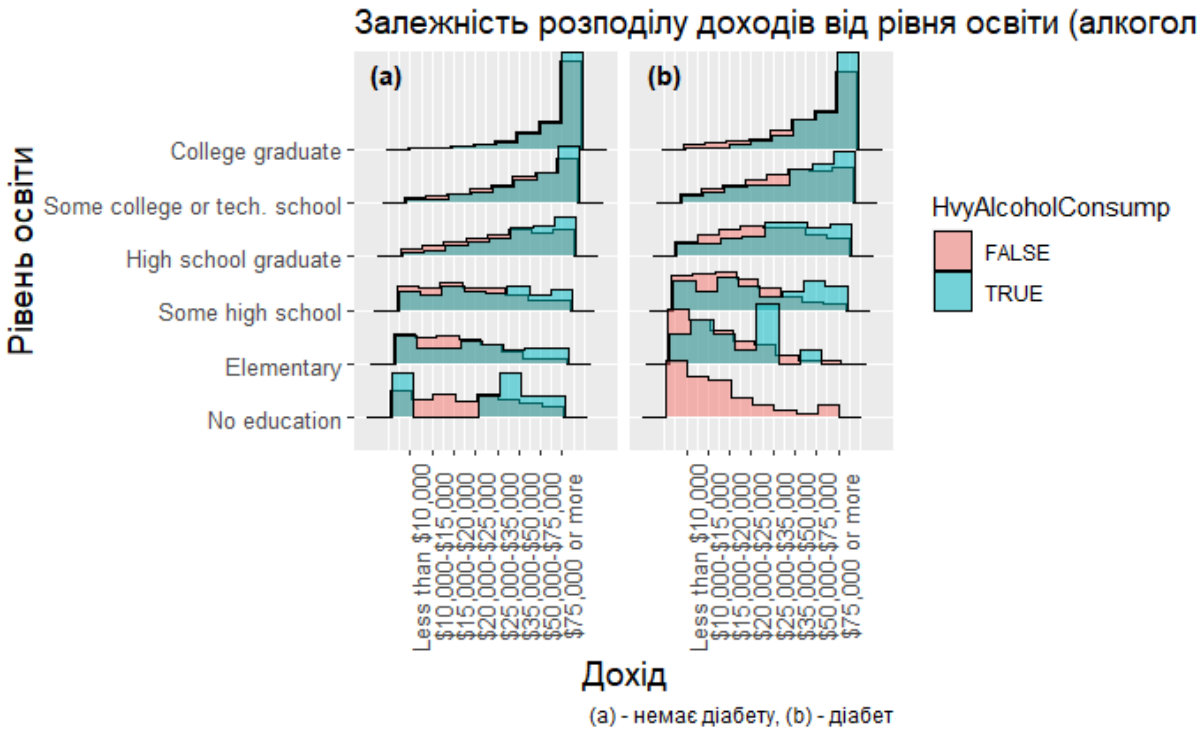
“Залежність наявності діабету від соціального статусу (раціон харчування, паління, зарплата, освіта, страховка)”

Графік №1. Залежність розподілу доходів від рівня освіти (мед. обстеження)



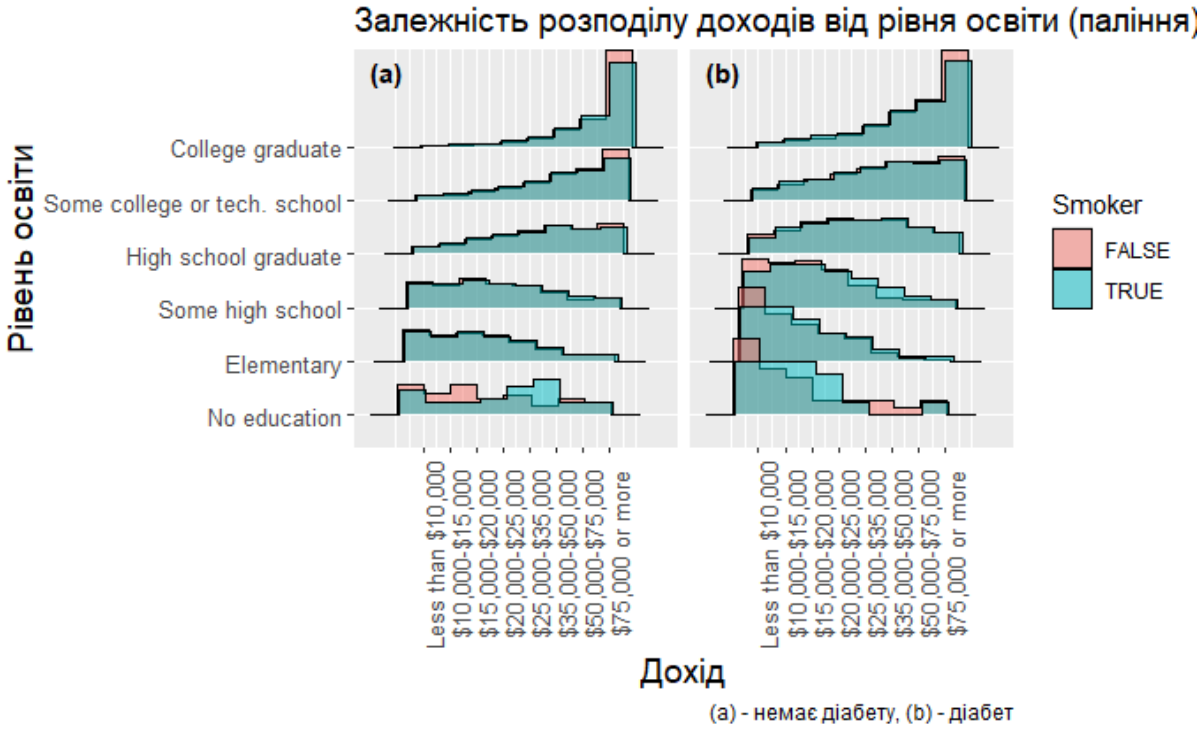
На графіку видно велику кількість бідних людей без освіти, що мають діабет, тоді як така сама група без діабету майже рівномірно розподілена по доходах (можливо, грає роль того, що багатих та неосвічених людей дуже мало, і навпаки - мало освічених людей із низькими доходами, тому за абсолютними характеристиками робити висновки недоцільно). Також видно, що у людей з технічною і вищою освітою розподіл за доходами наближається до рівномірного, якщо вони не користуються медичними послугами, тоді як менш освічені по мірі зменшення доходів схильні не користуватися ними. Судячи із розподілів для людей із вищою освітою, діабет може бути пов’язаним із низькими доходами та відсутністю медичного обслуговування, проте в більш освічених не спостерігається сильний вплив цих факторів: на обох графіках бачимо подібні розподіли.

Графік №2. Залежність розподілу доходів від рівня освіти (алкоголь)



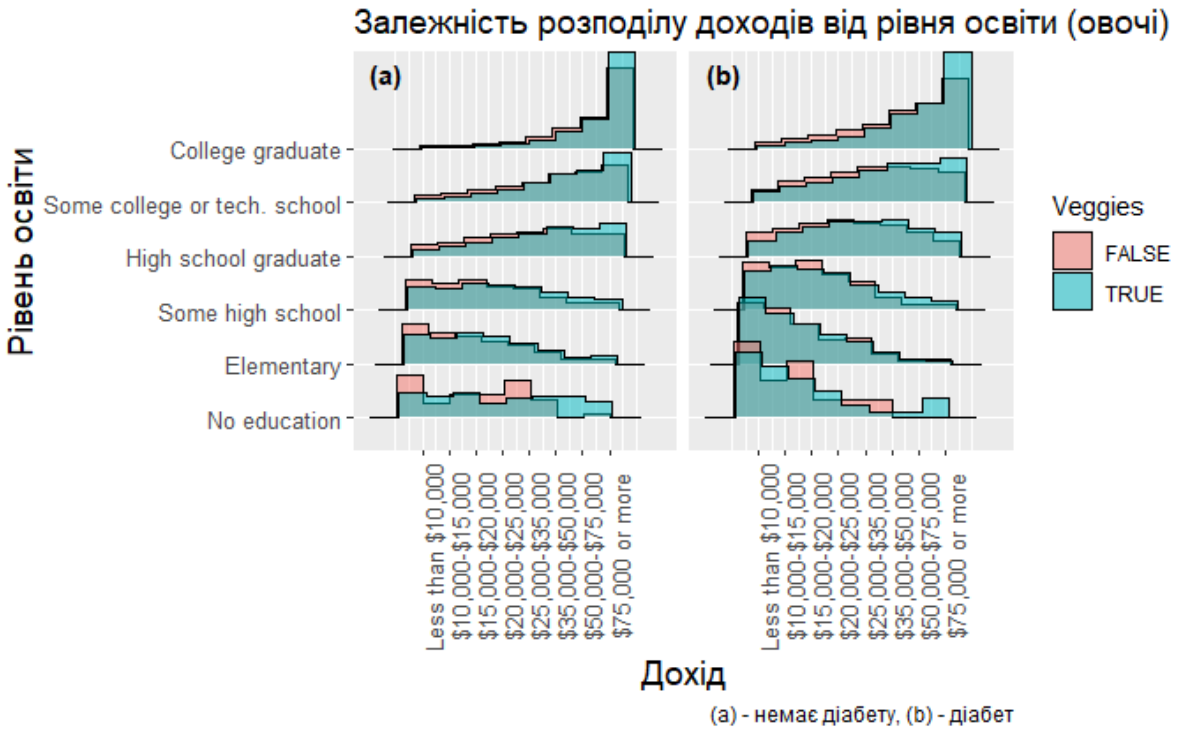
На даному графіку бачимо, що розподіли подібні як у випадку відсутності діабету, так і за його наявності. Для людей з діабетом розподіли при відсутності алкоголізму та наявності базової освіти виходять на плато, досягаючи рівня доходів $20000-25000, тоді як в людей без діабету вони схильний рости. Це може свідчити, що при діабеті для багатих людей із середньою освітою з деякого рівня доходів перестає відчуватися дія фактору діабету на відсутність алкоголізму, тобто можливо у багатих менша схильність до діабету, і цьому дещо сприяє відсутність алкоголізму. Дані стосовно неосвічених людей вочевидь недостатні, але показують, що серед них немає таких, що вживають алкоголь та хворіють на діабет. Водночас є певна кількість хворих без алкоголізму.

Графік №3. Залежність розподілу доходів від рівня освіти (паління)



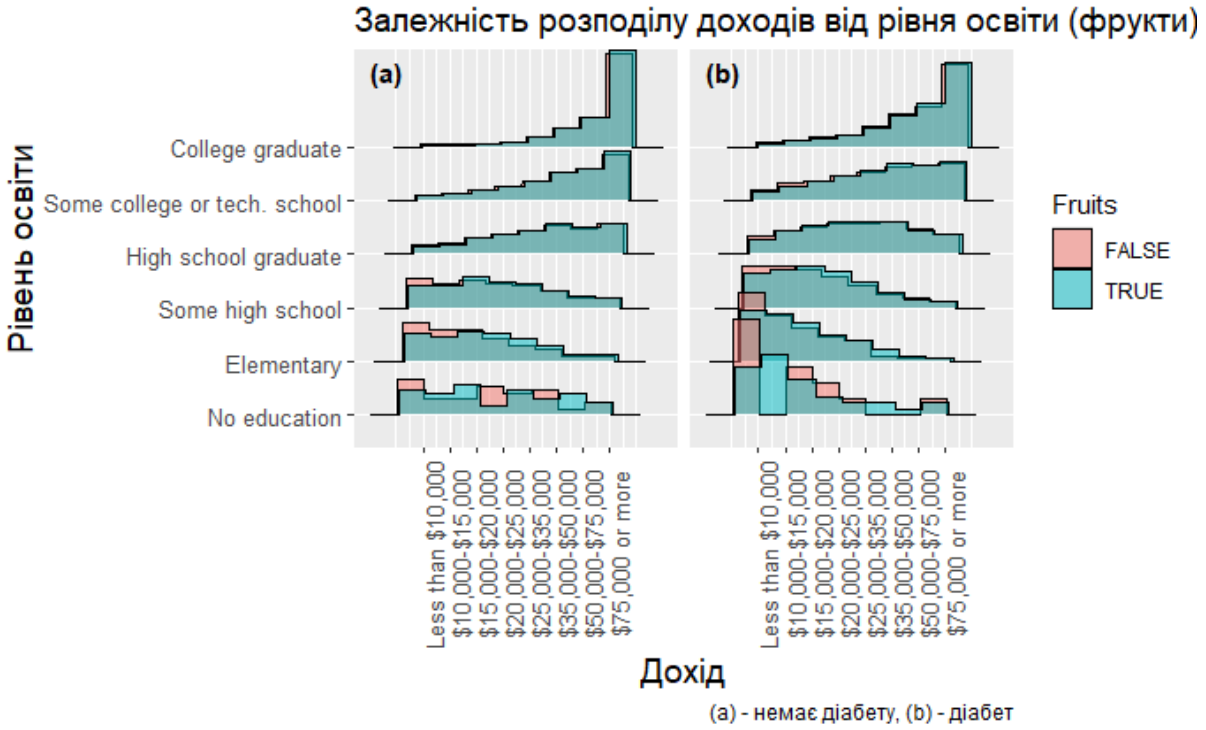
Тут ми бачимо, що розподіли, за винятком людей без освіти, майже ідентичні за критерієм паління. У декількох груп із низькими доходами спостерігається ріст частки тих, хто не палить. Також бачимо серед неосвічених збільшення частки тих, хто палить і не має діабету при доходах $25000-50000. Знову можна помітити, що подекуди розподіли прагнуть до вирівнення при досягненні певного рівня доходів.

Графік №4. Залежність розподілу доходів від рівня освіти (овочі)



Знову можна помітити подібність розподілів за фактором вживання овочів і зменшення приросту при діабеті незалежно від харчування. Бачимо тільки, що більше багатих схильні вживати овочі, тоді як бідні частіше не вживають їх.

Графік №5. Залежність розподілу доходів від рівня освіти (фрукти)



Бачимо цікаву ситуацію: розподіли при вживанні фруктів у обох графіках майже не відрізняються від розподілів без їх вживання для тих, хто має базову освіту або вище. Проте для груп з нижчим рівнем освіти спостерігається сильний шум в розподілах, ймовірно, через недостатню кількість даних.

На завершення проаналізуємо детальніше співвідношення діабету та інших факторів окремо.

Ізольоване дослідження залежностей між наявністю діабету й окремими факторами:

Освіта (абсолютні значення)

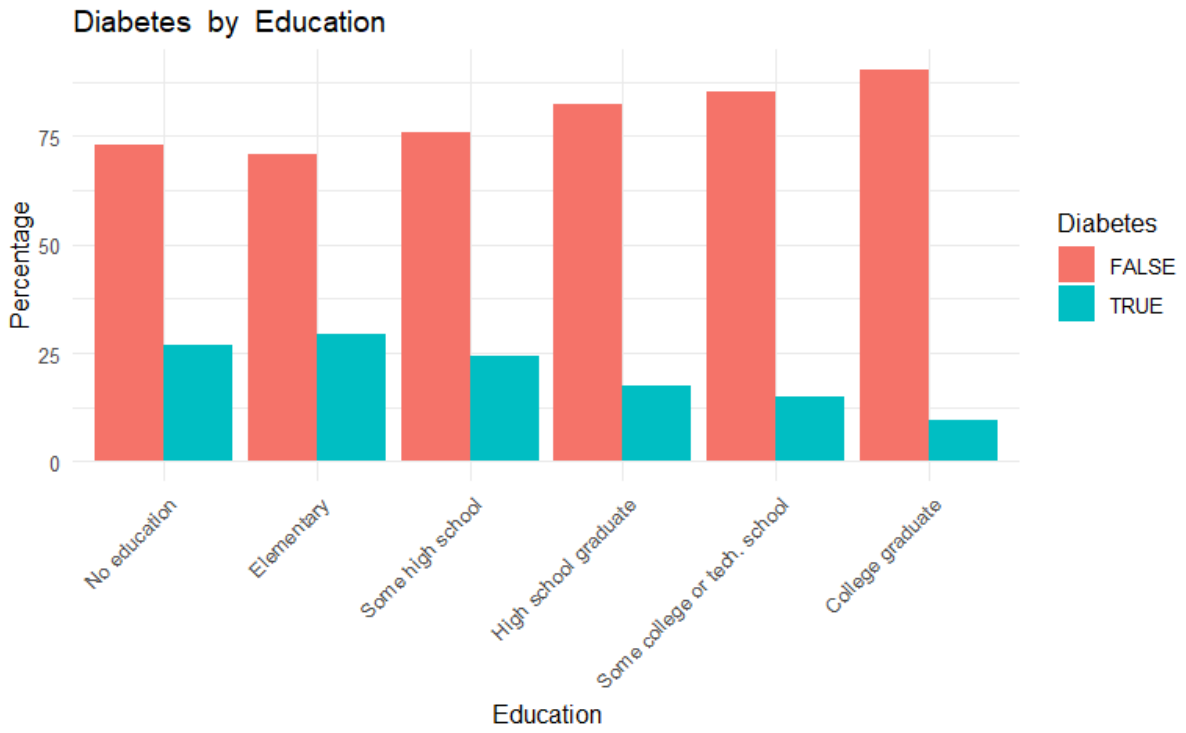
```
       Education
Diabetes No education Elementary Some high school High school graduate
   FALSE          127       2860             7182                51684
   TRUE            47       1183             2296                11066
        Education
Diabetes Some college or tech. school College graduate
   FALSE                        59556            96925
   TRUE                         10354            10400
```

Освіта

```
       Education
```

```
Diabetes No education Elementary Some high school High school graduate
   FALSE    0.72988506 0.70739550       0.75775480           0.82364940
   TRUE     0.27011494 0.29260450       0.24224520           0.17635060
        Education
Diabetes Some college or tech. school College graduate
   FALSE                   0.85189529       0.90309807
   TRUE                    0.14810471       0.09690193
```

Візуалізація №1. Залежність діабету від рівня освіти.



Найбільша частка діабетиків серед людей з низьким рівнем освіти. Чим нижчий рівень освіти - тим більша кількість діабетиків. Проте, рівень освіти пов'язаний з рівнем доходів, а як вже було показано вище серед людей з меншим рівнем доходу - більше діабетиків.

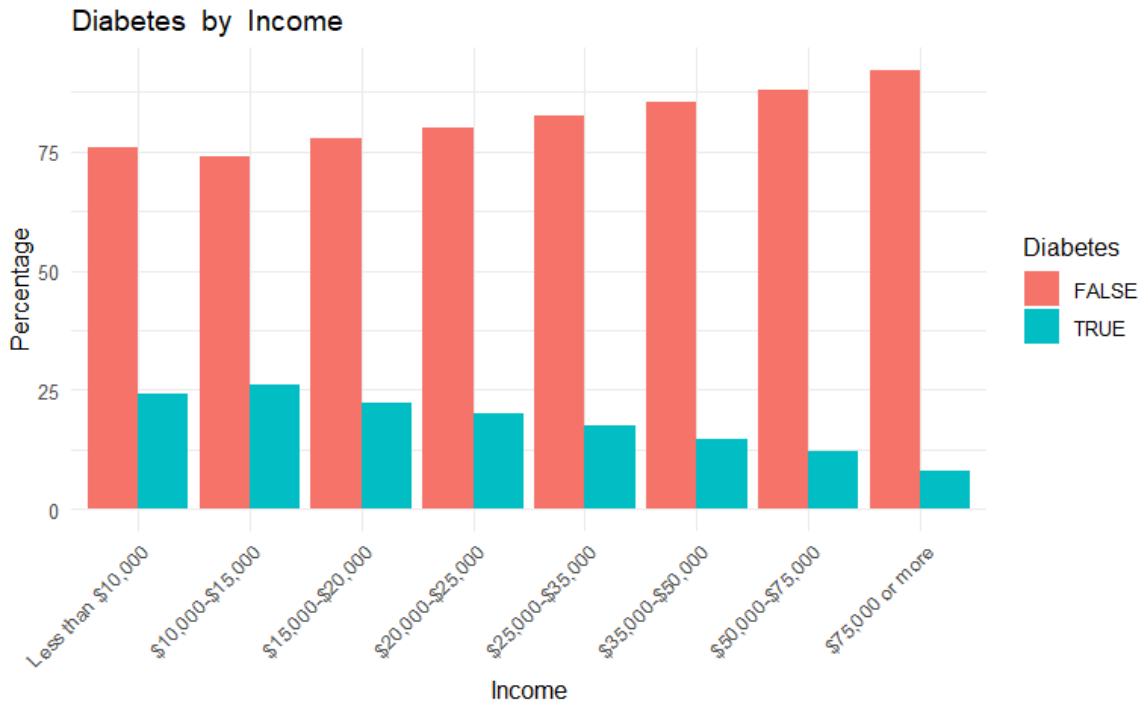
Доходи (абсолютні значення)

```
       Income
Diabetes Less than $10,000 $10,000-$15,000 $15,000-$20,000 $20,000-$25,000
   FALSE               7428             8697            12426            16081
   TRUE                2383             3086             3568             4054
        Income
Diabetes $25,000-$35,000 $35,000-$50,000 $50,000-$75,000 $75,000 or more
   FALSE            21379            31179            37954            83190
   TRUE              4504             5291             5265             7195
```

Доходи

```
        Income
Diabetes Less than $10,000 $10,000-$15,000 $15,000-$20,000 $20,000-$25,000
   FALSE         0.75710937      0.73809726      0.77691634      0.79865905
   TRUE          0.24289063      0.26190274      0.22308366      0.20134095
         Income
Diabetes $25,000-$35,000 $35,000-$50,000 $50,000-$75,000 $75,000 or more
   FALSE       0.82598617      0.85492185      0.87817858      0.92039608
   TRUE        0.17401383      0.14507815      0.12182142      0.07960392
```

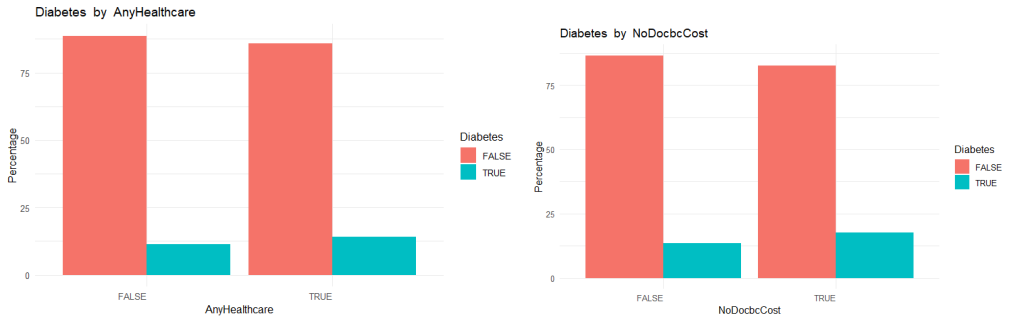
Візуалізація №2. Залежність діабету від рівня доходів.



Серед бідніших більша частка діабетиків ніж серед багатших. Чим вища категорія заробітної плати тим менша частка діабетиків серед людей цієї категорії.

Візуалізація №3. Залежність діабету від наявності доступу до медичних послуг.

```
        AnyHealthcare
Diabetes     FALSE      TRUE
   FALSE 0.8854796 0.8593900
   TRUE  0.1145204 0.1406100
```



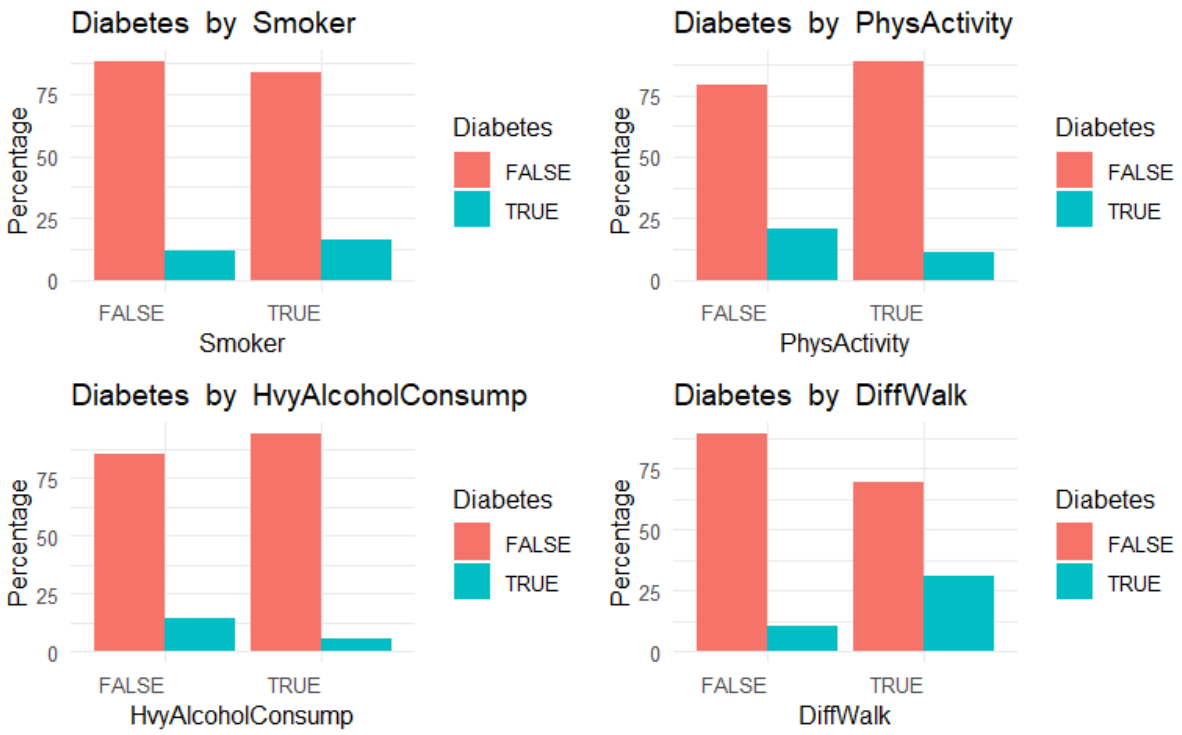
Вживання алкоголю

```
          HvyAlcoholConsump
Diabetes          FALSE       TRUE
   FALSE    0.85584570 0.94163861
   TRUE     0.14415430 0.05836139
```

Паління

```
          Smoker
Diabetes        FALSE      TRUE
   FALSE    0.8794467 0.8370707
   TRUE     0.1205533 0.1629293
```

Візуалізація №4. Залежність наявності діабету, від способу життя. (Чи є курцем, чи займається фізичними активностями)



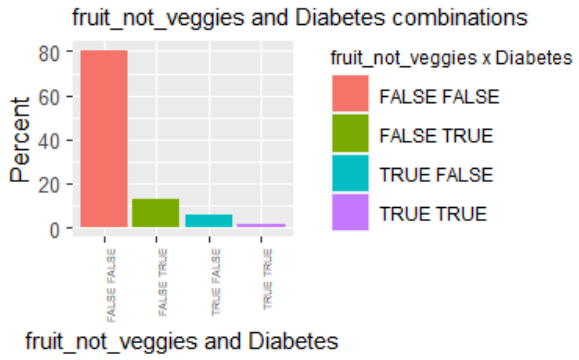
Вживання овочів

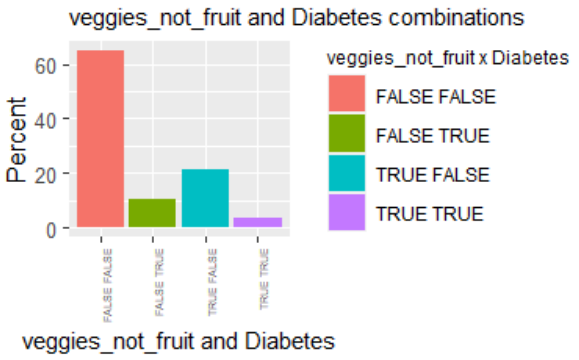
```
          Veggies
Diabetes      FALSE      TRUE
   FALSE 0.8200213 0.8701133
   TRUE  0.1799787 0.1298867
```

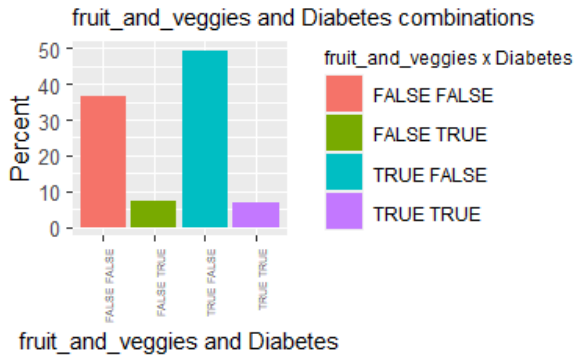
Вживання фруктів

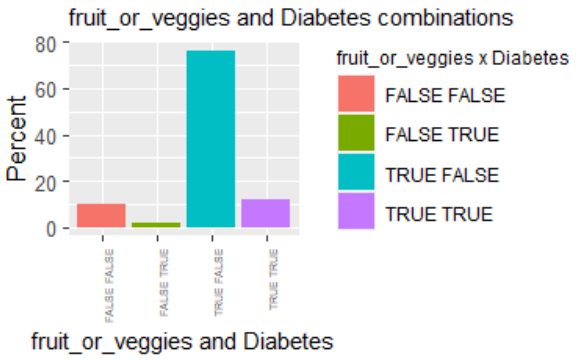
```
          Fruits
Diabetes      FALSE      TRUE
   FALSE 0.8420707 0.8713906
   TRUE  0.1579293 0.1286094
```

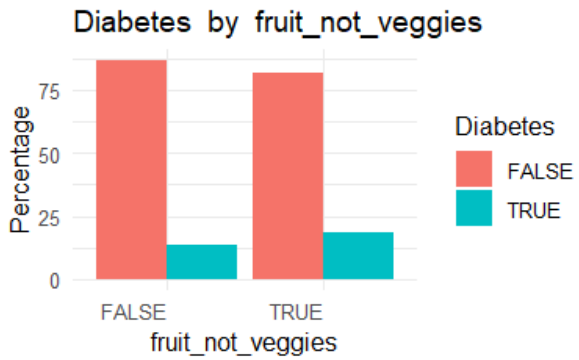
Візуалізація №5. Залежність наявності діабету від наявності в раціоні фруктів або овочей.

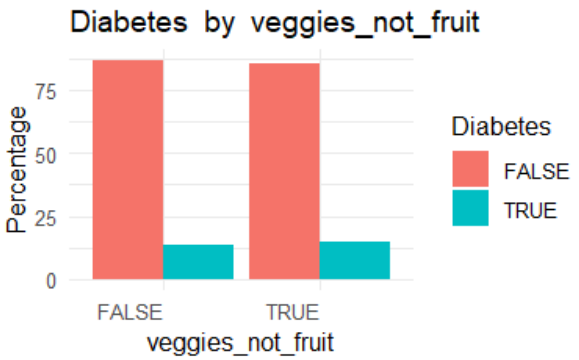


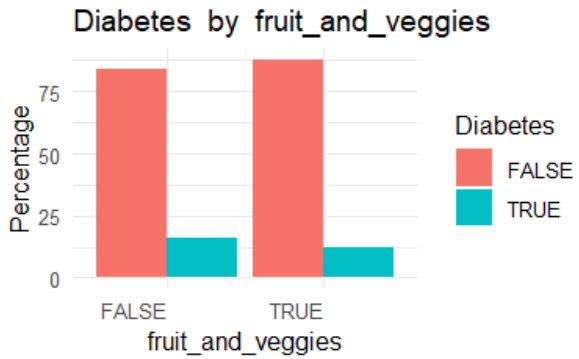


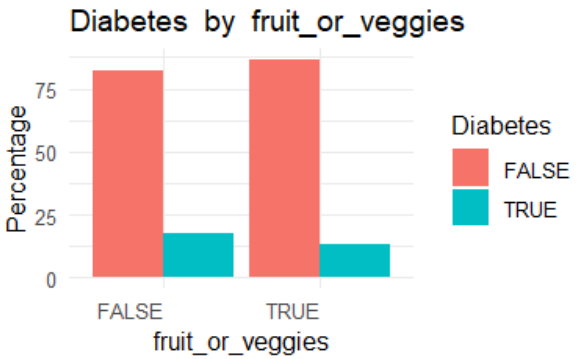












Як можна побачити значну частку датасету складають люди, які мають в своєму раціоні і фрукти і овочі (більше 50%), а людей які не їдять ні фруктів ні овочей насправді доволі мало (10%). Щодо другого графіку. Частка діабетиків більша серед людей які не мають в своєму раціоні ні фруктів ні овочей (приблизно 16%) в порівнянні з часткою людей які їдять

і овощі і фрукти, але ця частка є доволі малою, тому немає сенсу далі розглядати залежність між ціма змінними, їх вплив не є суттєвим.

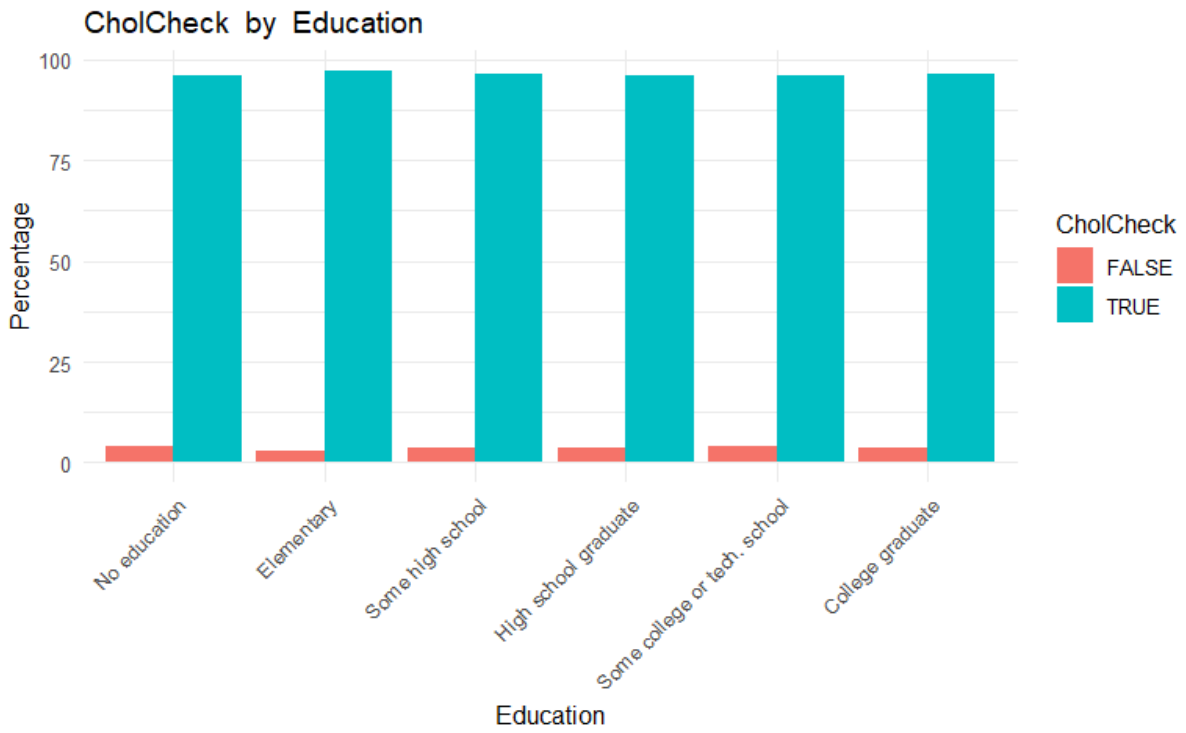
Отже, щодо цього питання можна підбити попередні підсумки:

- спостерігаються менші частки хворих на діабет у багатших людей та людей з вищим рівнем освіти;
- зі збільшенням доходів деякі із негативних факторів (алкоголь, паління, відсутність в раціоні овочів та фруктів) можуть сприяти діабету;
- схоже, що люди без освіти та з низькими доходами частіше мають діабет, але може бути вплив інших факторів (напр. неможливість купити продукти), може даватися в знаки мала кількість таких людей.

3. Третє питання

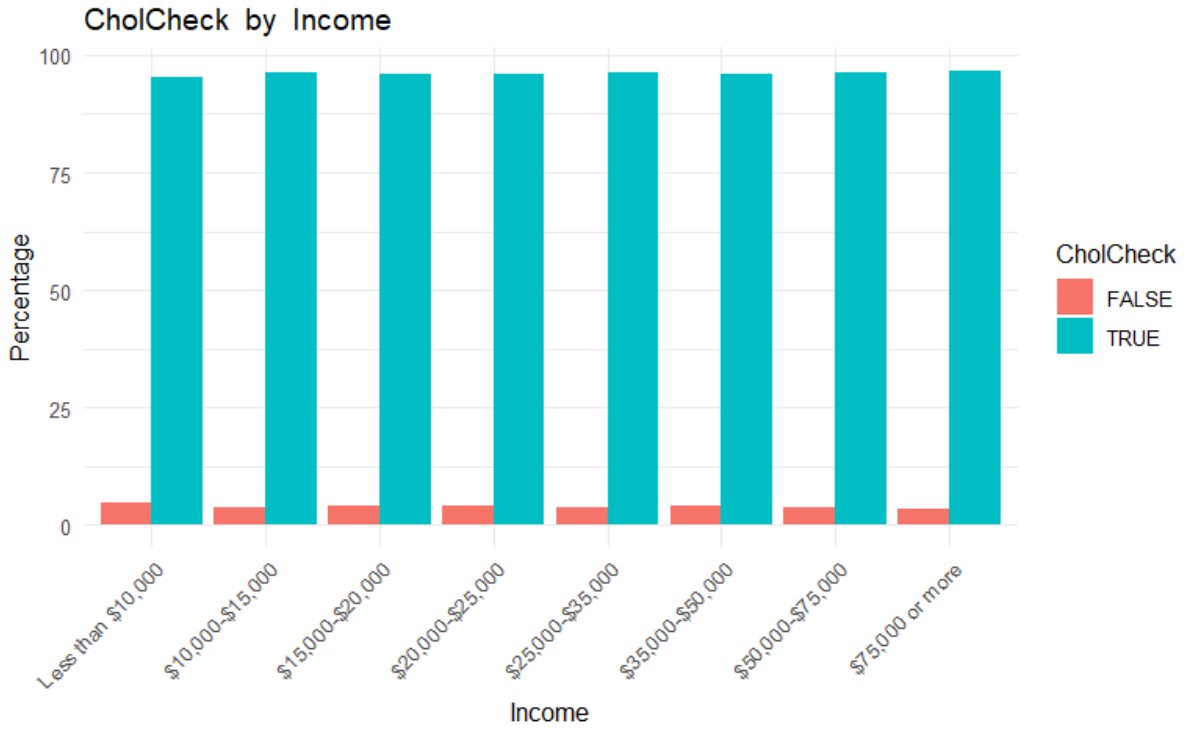
“Залежність частоти обстежень рівня холестерину від соціального статусу (рівня освіти, зарплати та медичного страхування);”

Графік №1. Залежність частоти обстежень рівня холестерину від рівня освіти



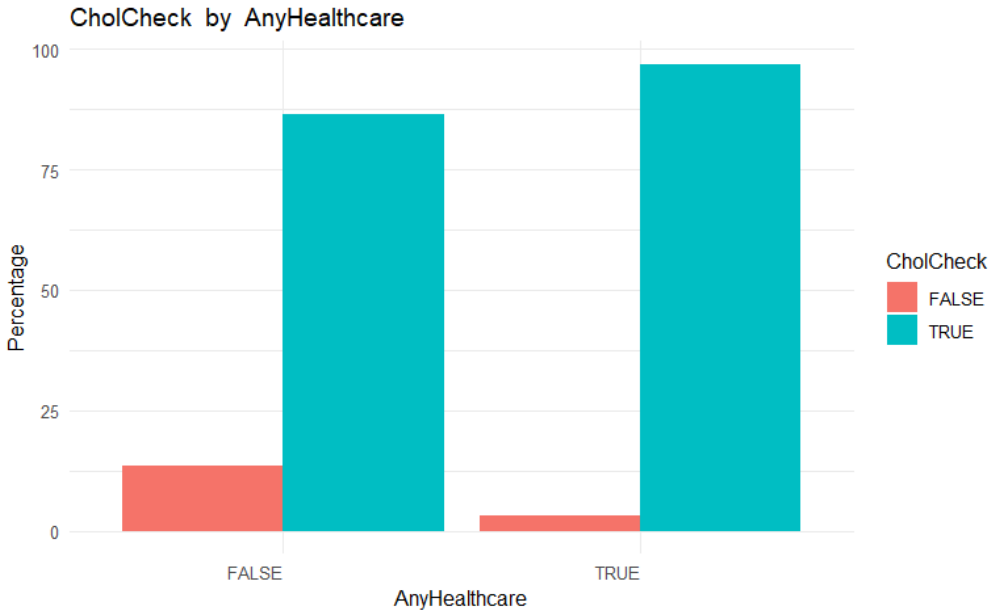
Залежності немає.

Графік №2. Залежність частоти обстежень рівня холестерину від рівня заробітної плати.



Залежності немає.

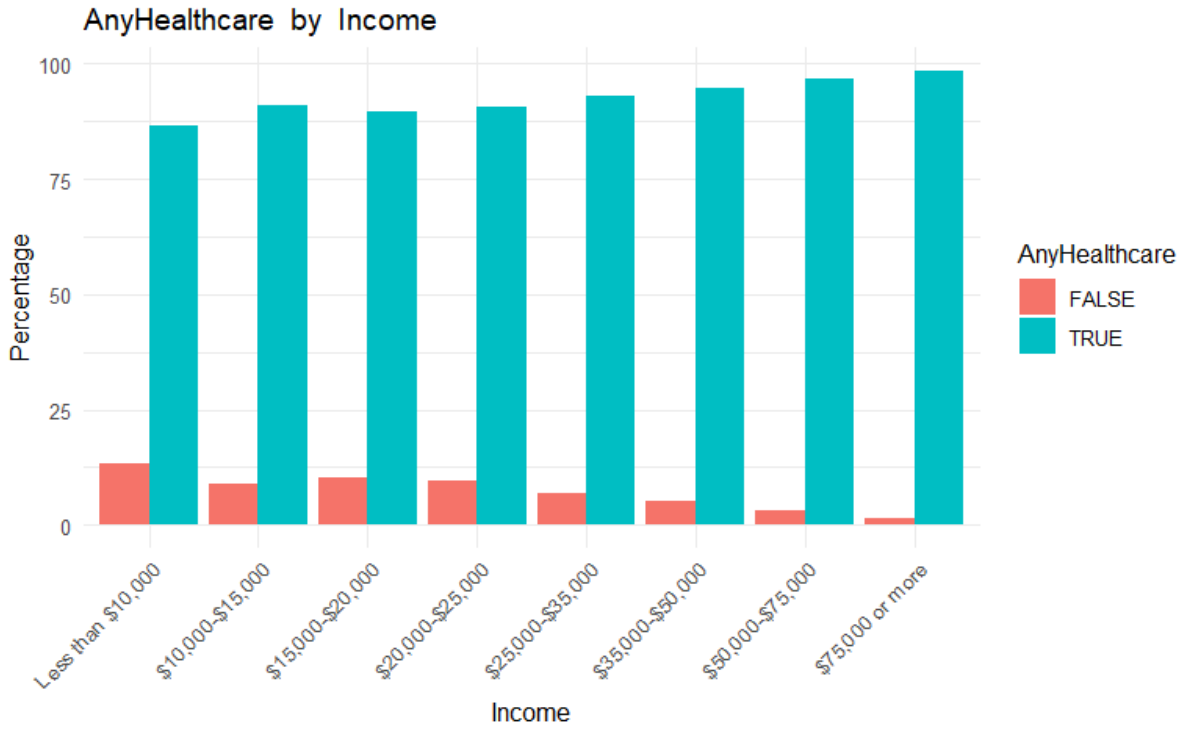
Графік №3. Залежність частоти обстежень рівня холестерину від того чи є медичне страхування.



AnyHealthcare <lgl>	CholCheck <lgl>	Count <int>	Percent <dbl>
FALSE	FALSE	1684	13.562052
FALSE	TRUE	10733	86.437948
TRUE	FALSE	7786	3.227184
TRUE	TRUE	233477	96.772816

Серед людей без медичного страхування частка людей, які не зробили жодної перевірки рівня холестерину за останні 5 років, більша ніж серед людей з медичним страхуванням. (13.5% проти 3%). Деяка залежність прослідковується.

* Додаткове питання 1. “Чи є залежність наявності страховки від заробітної плати?”



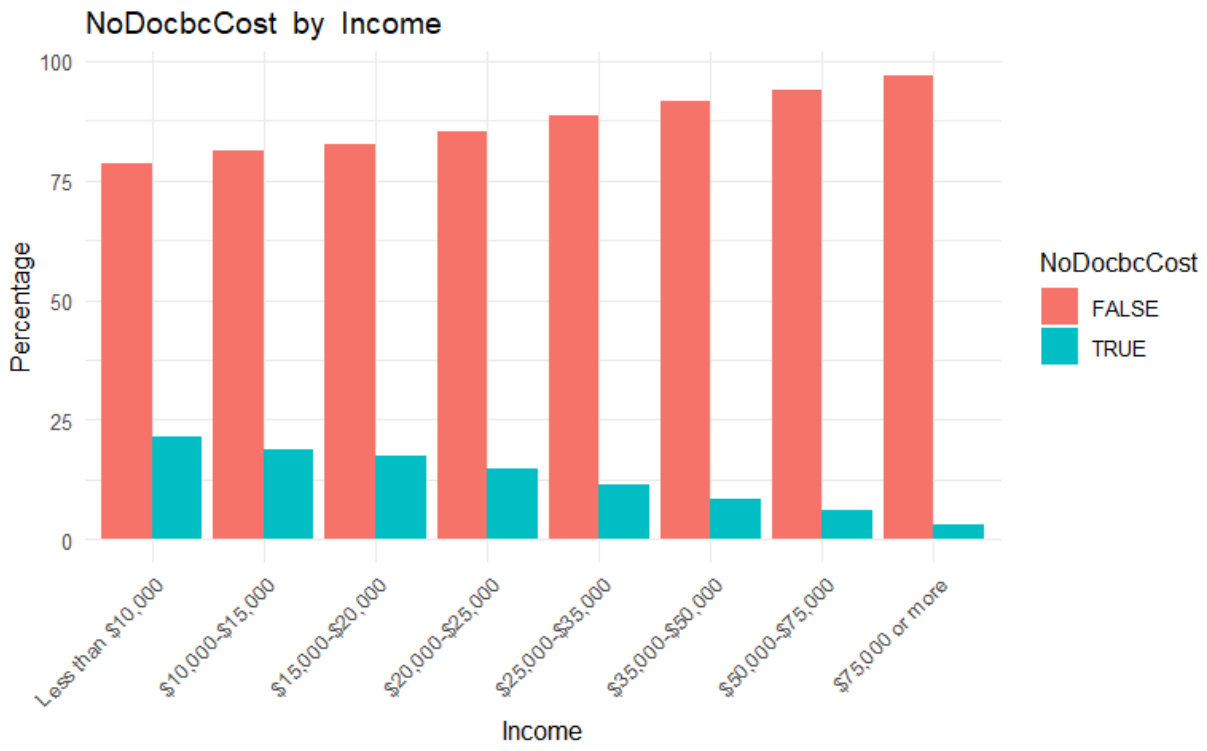
Income <ord>	AnyHealthcare <lgl>	Count <int>	Percent <dbl>
Less than $10,000	FALSE	1320	13.454286
Less than $10,000	TRUE	8491	86.545714

$10,000-$15,000	FALSE	1061	9.004498
$10,000-$15,000	TRUE	10722	90.995502
$15,000-$20,000	FALSE	1667	10.422658
$15,000-$20,000	TRUE	14327	89.577342
$20,000-$25,000	FALSE	1901	9.441271
$20,000-$25,000	TRUE	18234	90.558729
$25,000-$35,000	FALSE	1784	6.892555
$25,000-$35,000	TRUE	24099	93.107445
$35,000-$50,000	FALSE	1899	5.207019
$35,000-$50,000	TRUE	34571	94.792981
$50,000-$75,000	FALSE	1381	3.195354
$50,000-$75,000	TRUE	41838	96.804646
$75,000 or more	FALSE	1404	1.553355
$75,000 or more	TRUE	88981	98.446645

Так, залежність прослідковується, категорії людей, які заробляють менше мають частку людей без страховки більшу - ніж категорії, які заробляють більше.

* Додаткове питання 2. “Чи є залежність змінної NoDocbcCost від заробітної плати?”

Цікавить саме положення людей, які заробляють мало, зрозуміло що в людей з рівнем заробітку більше 70 тисяч в рік, є можливість виділити гроші на послуги висококваліфікованого спеціаліста чи дорогу операцію.

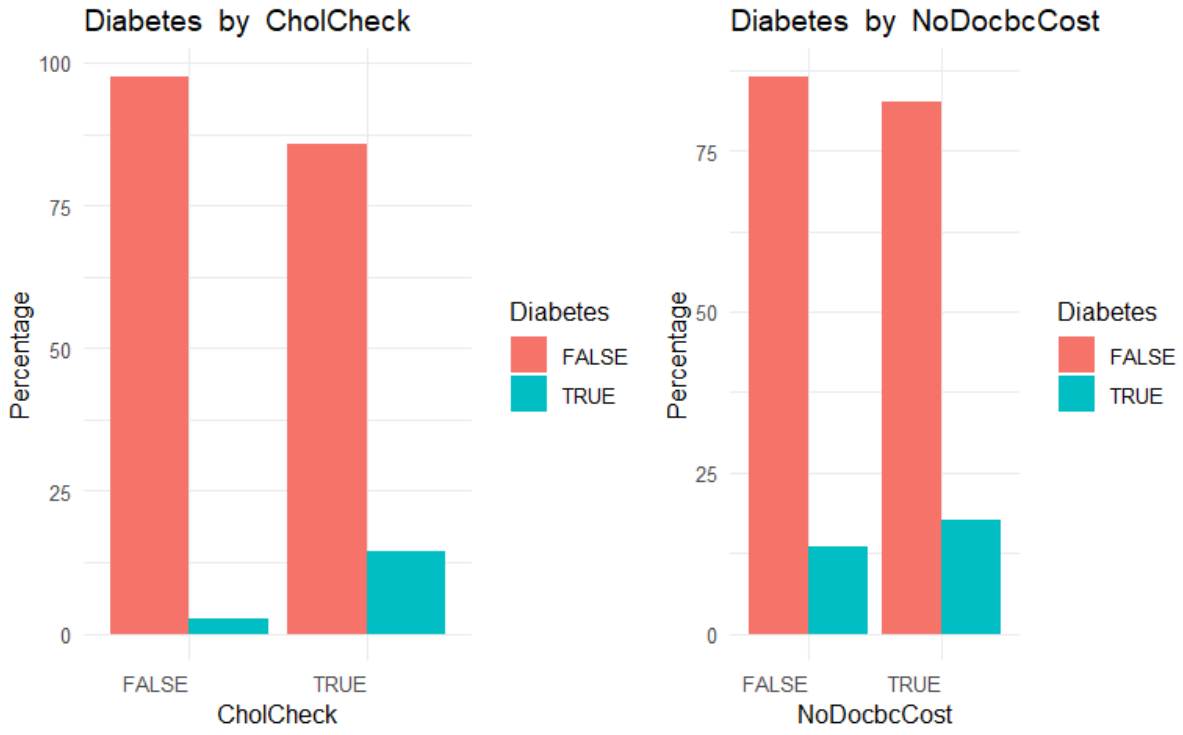


Income <ord>	NoDocbcCost <lgl>	Count <int>	Percent <dbl>
Less than $10,000	FALSE	7714	78.626032
Less than $10,000	TRUE	2097	21.373968
$10,000-$15,000	FALSE	9578	81.286599
$10,000-$15,000	TRUE	2205	18.713401
$15,000-$20,000	FALSE	13206	82.568463
$15,000-$20,000	TRUE	2788	17.431537
$20,000-$25,000	FALSE	17153	85.189968
$20,000-$25,000	TRUE	2982	14.810032
$25,000-$35,000	FALSE	22961	88.710737
$25,000-$35,000	TRUE	2922	11.289263
$35,000-$50,000	FALSE	33452	91.724705
$35,000-$50,000	TRUE	3018	8.275295
$50,000-$75,000	FALSE	40600	93.940165
$50,000-$75,000	TRUE	2619	6.059835

$75,000 or more	FALSE	87662	96.987332
$75,000 or more	TRUE	2723	3.012668

Як і передбачалось, серед менш заможних респондентів кількість людей, які мали труднощі з вибором лікаря, через вартість його послуг, значно більша. Серед найбіднішої категорії - 21%, серед найбагатшої - 3%.

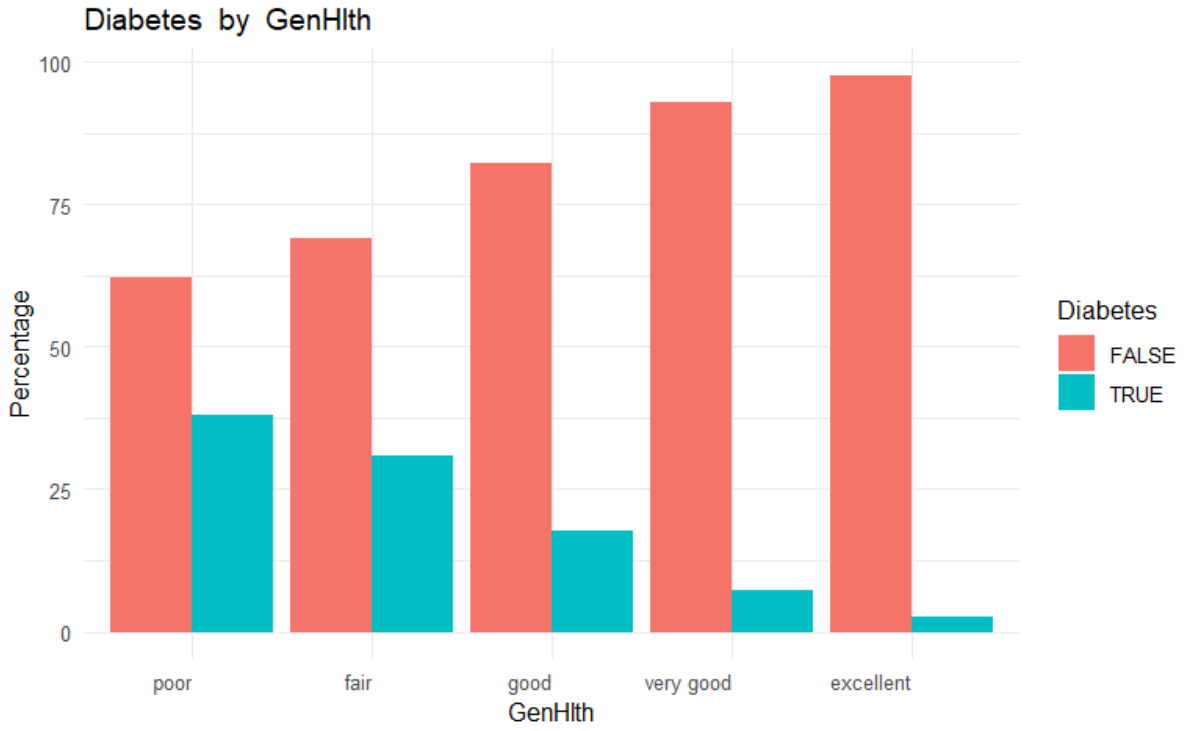
* Додаткове питання 3. Залежність наявності діабету від змінних **CholCheck**, **NoDocbcCost**.



4. Четверте питання

“Залежність наявності діабету від суб'єктивної оцінки стану здоров'я. Порівняння із залежністю наявності діабету від об'єктивних факторів: імт, тиску, чи є хвороби серця і того, чи важко людині ходити (чи завищують / занижують люди оцінку власного здоров'я).”

Графік №1. Залежність наявності діабету від суб’єктивної оцінки стану здоров’я



GenHlth	Diabetes	Count	Percent
<ord>	<lgl>	<int>	<dbl>
poor	FALSE	7503	62.105786
poor	TRUE	4578	37.894214
fair	FALSE	21780	68.989547
fair	TRUE	9790	31.010453
good	FALSE	62189	82.210560
good	TRUE	13457	17.789440

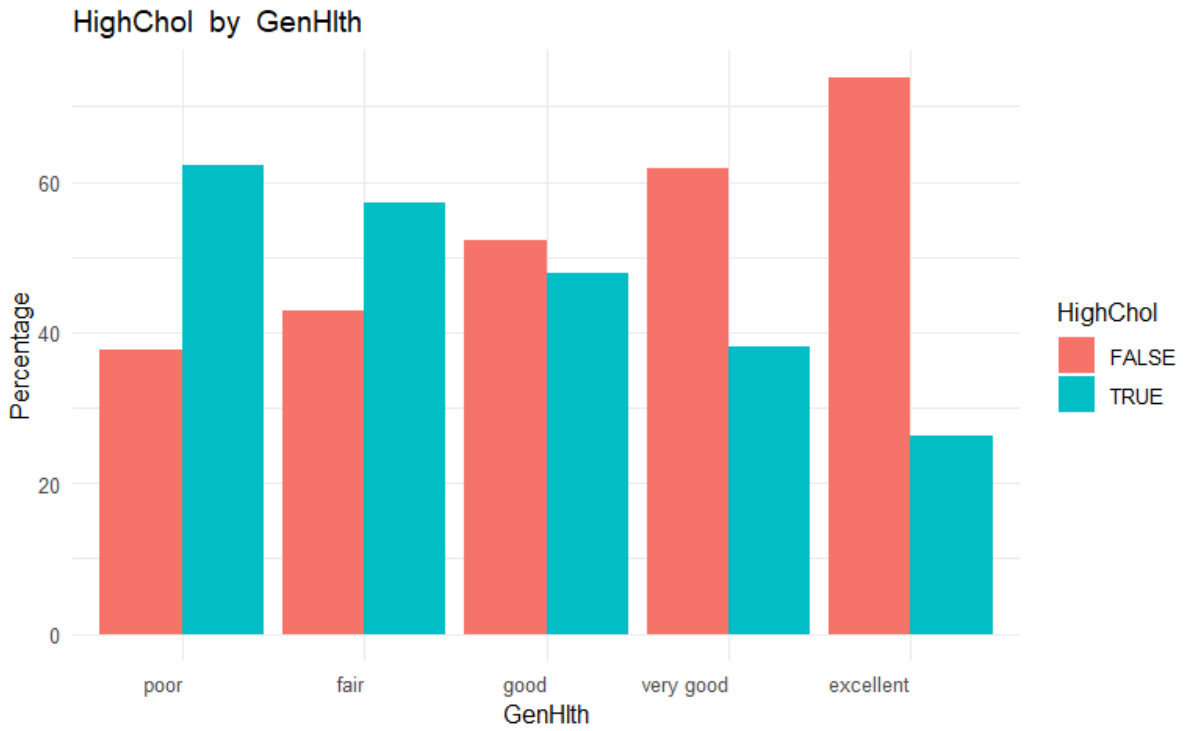
very good	FALSE	82703	92.837098
very good	TRUE	6381	7.162902
excellent	FALSE	44159	97.483388
excellent	TRUE	1140	2.516612

Частка діабетиків серед людей, які оцінили свій стан здоров'я як поганий найбільша, якщо порівнювати з іншими категоріями. І прослідковується залежність - чим вища оцінка, тим менша частка діабетиків серед категорії респондентів. Різниця між часткою діабетиків в poor і в excellent: 37,9% - 2,5% = 35,4%. При переході до вищої категорії, різниці складають: 7% - 14% - 10% - 5%. Залежність має місце.

Чи завищують / занижують люди оцінку власного здоров'я?

Для відповіді на це запитання перевіримо частку людей з високим рівнем холестирину, високим тиском, серцево-судинними захворюваннями, в яких був інсульт, яким важко ходити, в яких низький рівень активності, імт поза межами норми в залежності від категорії GenHlth.

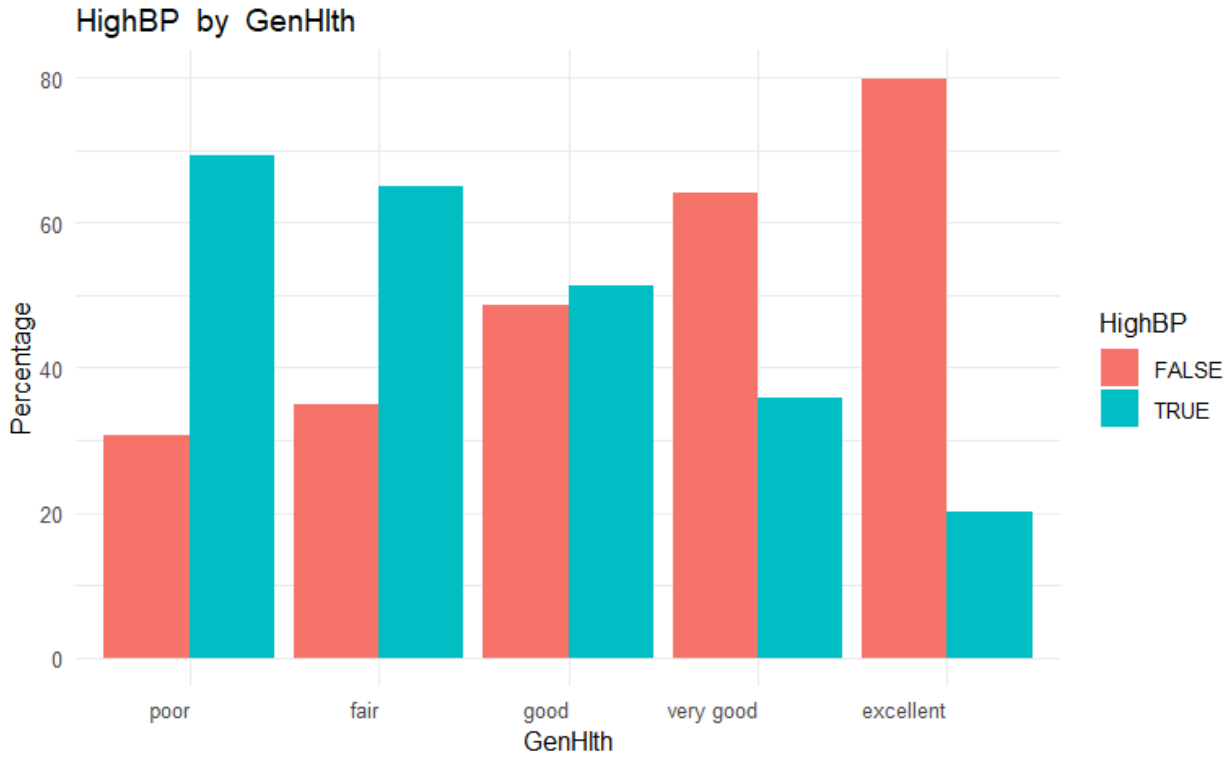
Графік №2. Залежність високого рівня холестерину від оцінки респондентом стану власного здоров’я.



За критерієм високого рівня холестирину в крові, люди завищують оцінку свого стану здоров'я. Якщо інтерпретувати цей графік з іншої точки зору, тенденція зберігається - чим вища оцінка, тим менше людей з високим рівнем холестирину.

GenHlth <ord>	HighChol <lgl>	Count <int>	Percent <dbl>
poor	FALSE	4557	37.72039
poor	TRUE	7524	62.27961
fair	FALSE	13510	42.79379
fair	TRUE	18060	57.20621
good	FALSE	39514	52.23541
good	TRUE	36132	47.76459
very good	FALSE	55092	61.84276
very good	TRUE	33992	38.15724
excellent	FALSE	33416	73.76763
excellent	TRUE	11883	26.23237

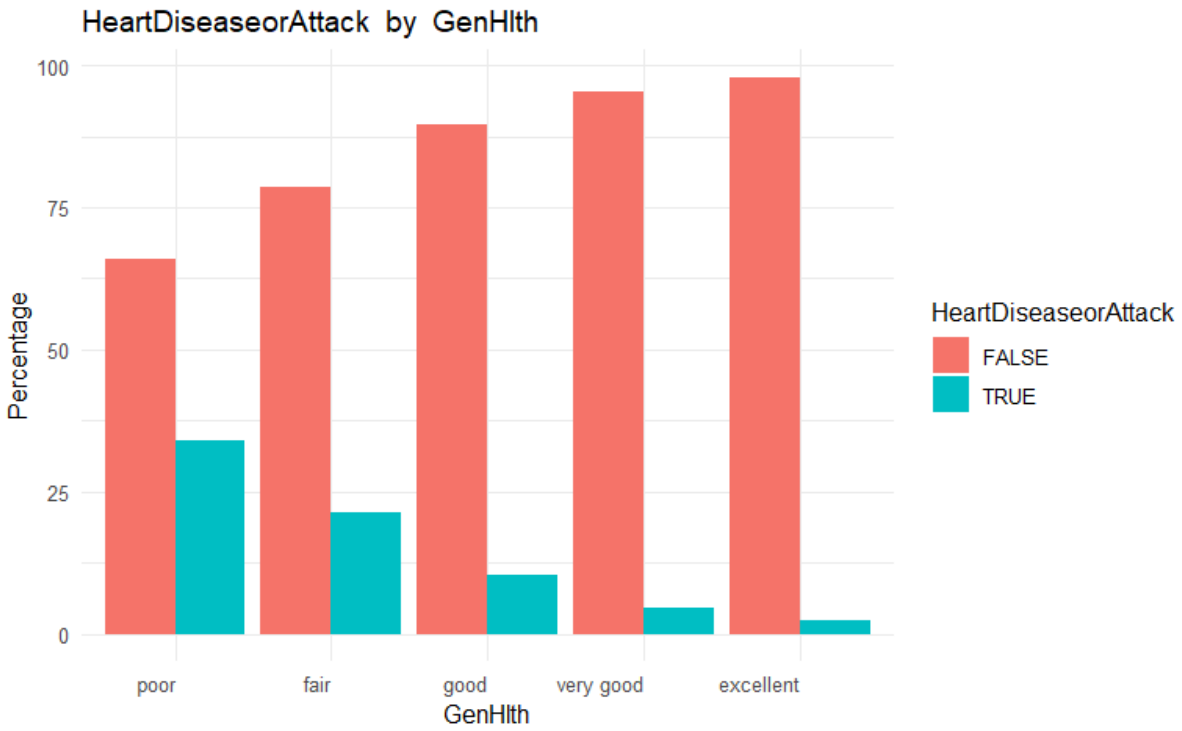
Графік №3. Залежність високого артеріального тиску від оцінки респондентом стану власного здоров’я.



GenHlth <ord>	HighBP <lgl>	Count <int>	Percent <dbl>
poor	FALSE	3707	30.68455
poor	TRUE	8374	69.31545
fair	FALSE	11027	34.92873
fair	TRUE	20543	65.07127
good	FALSE	36734	48.56040
good	TRUE	38912	51.43960
very good	FALSE	57233	64.24610
very good	TRUE	31851	35.75390
excellent	FALSE	36150	79.80309
excellent	TRUE	9149	20.19691

За критерієм високого тиску, завищують, але не так як на графіку №2. Тенденція зберігається.

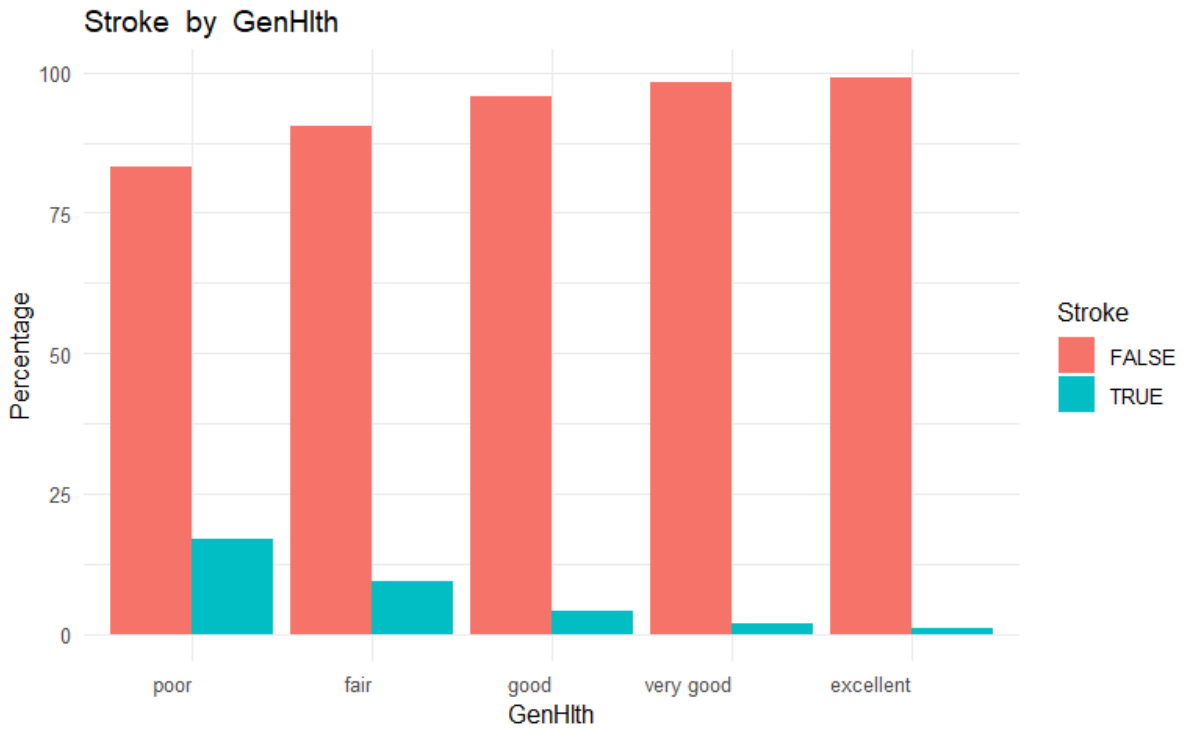
Графік №4. Залежність наявності серцево-судинних захворювань від оцінки респондентом стану власного здоров'я.



GenHlth <ord>	HeartDiseaseorAttack <lgl>	Count <int>	Percent <dbl>
poor	FALSE	7974	66.004470
poor	TRUE	4107	33.995530
fair	FALSE	24842	78.688628
fair	TRUE	6728	21.311372
good	FALSE	67732	89.538112
good	TRUE	7914	10.461888
very good	FALSE	84956	95.366171
very good	TRUE	4128	4.633829
excellent	FALSE	44283	97.757125
excellent	TRUE	1016	2.242875

За критерієм наявності серцево-судинних захворювань/чи був серцевий напад, не завищують. Тенденція зберігається (червоний стовпчик підіймається, зелений спадає)

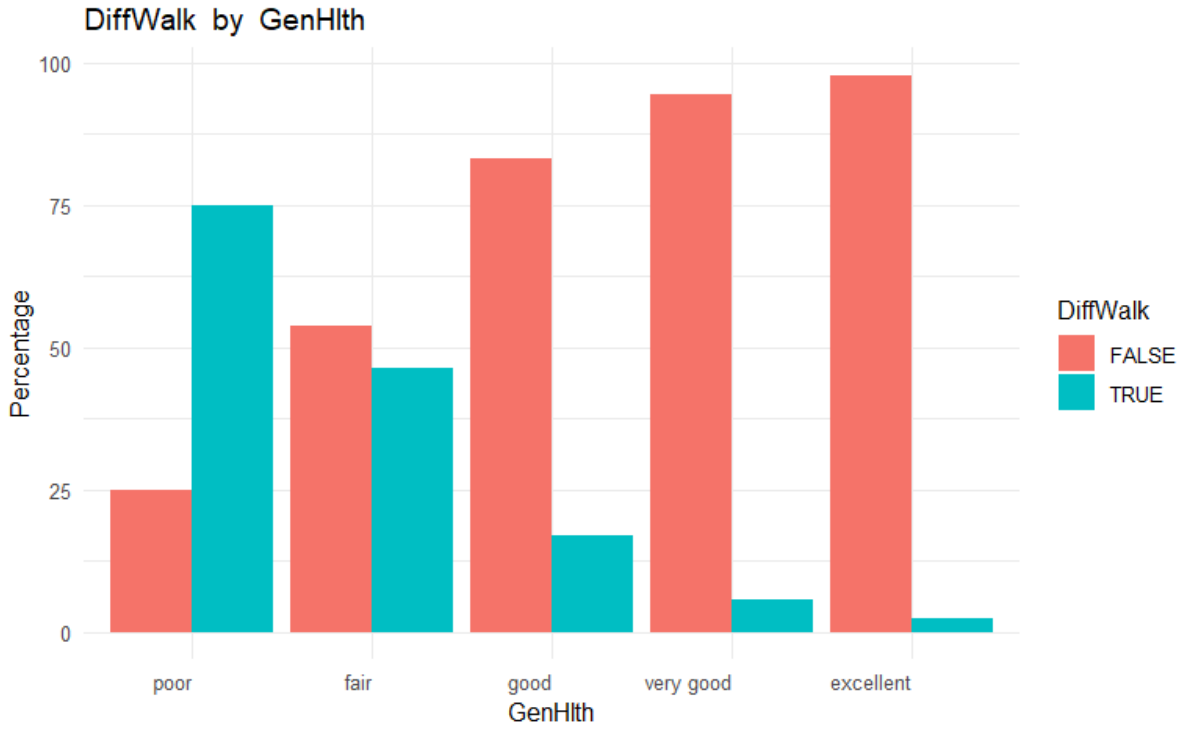
Графік №5. Залежність змінної “Чи діагнозували у вас інсульт” від оцінки респондентом стану власного здоров’я.



GenHlth <ord>	Stroke <lgl>	Count <int>	Percent <dbl>
poor	FALSE	10050	83.1884778
poor	TRUE	2031	16.8115222
fair	FALSE	28591	90.5638264
fair	TRUE	2979	9.4361736
good	FALSE	72473	95.8054623
good	TRUE	3173	4.1945377
very good	FALSE	87420	98.1321000
very good	TRUE	1664	1.8679000
excellent	FALSE	44854	99.0176384
excellent	TRUE	445	0.9823616

Не інформативно, оскільки, в вибірці відносно мало людей в яких Stroke = TRUE. Тенденція - зберігається.

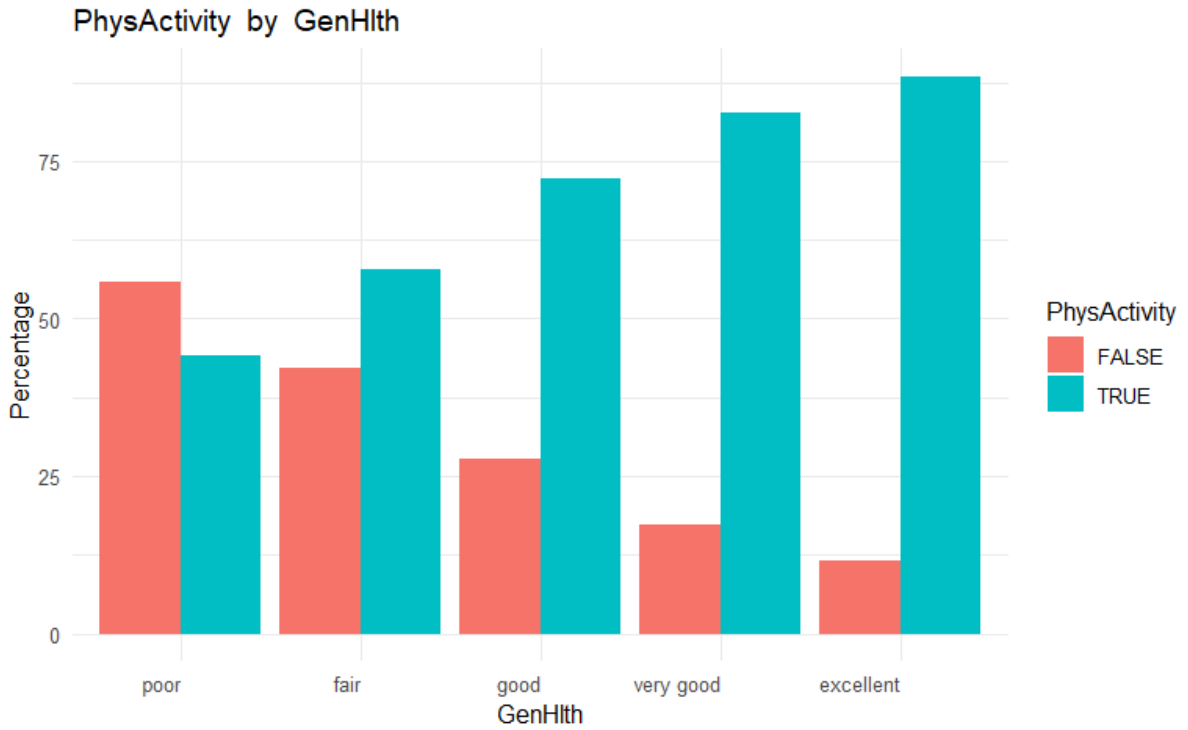
Графік №6. Залежність змінної “Чи відчуваєте Ви серйозні труднощі при ходьбі або підйомі сходами? ” від оцінки респондентом стану власного здоров’я.



GenHlth <ord>	DiffWalk <lgl>	Count <int>	Percent <dbl>
poor	FALSE	3020	24.997931
poor	TRUE	9061	75.002069
fair	FALSE	16976	53.772569
fair	TRUE	14594	46.227431
good	FALSE	62794	83.010338
good	TRUE	12852	16.989662
very good	FALSE	83988	94.279556
very good	TRUE	5096	5.720444
excellent	FALSE	44227	97.633502
excellent	TRUE	1072	2.366498

В категорії poor частка людей, яким важко ходити складає більшість - 75%. Якщо дивитися на графік, то за цим критерієм люди дають доволі справедливу оцінку.

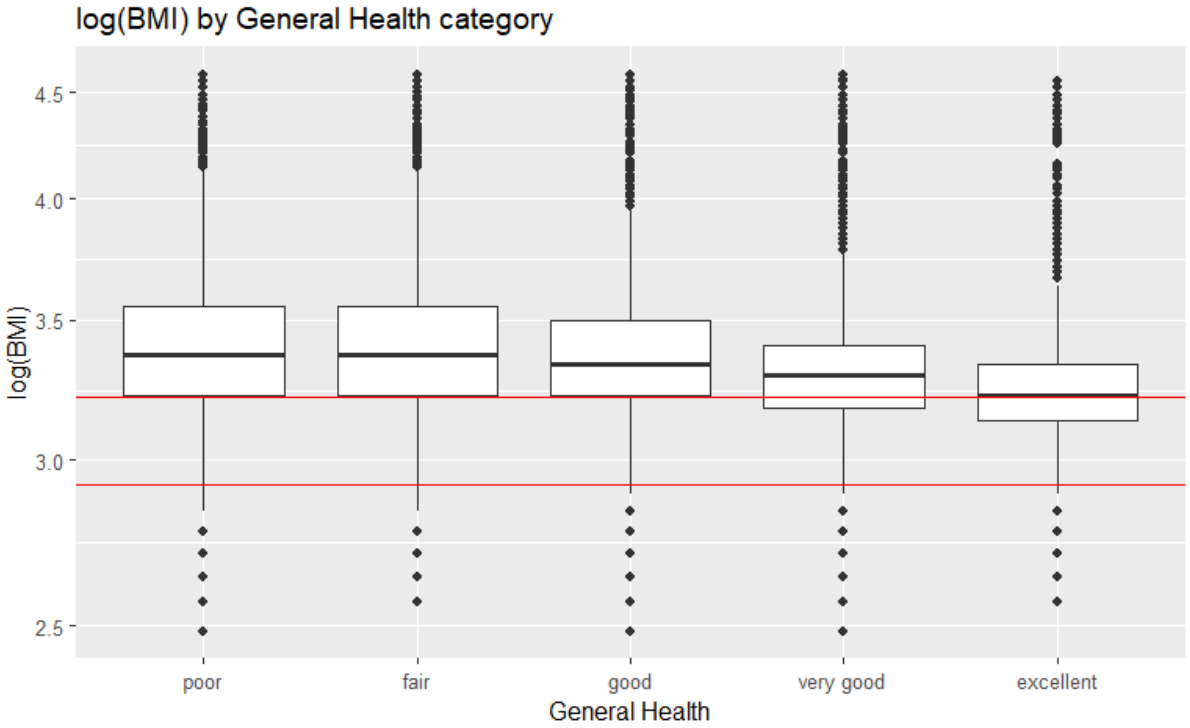
Графік №7. Залежність змінної “Чи займалися фізичною активністю за останні 30 днів - не враховуючи роботу ?” від оцінки респондентом стану власного здоров’я.



GenHlth <ord>	PhysActivity <lgl>	Count <int>	Percent <dbl>
poor	FALSE	6742	55.80664
poor	TRUE	5339	44.19336
fair	FALSE	13310	42.16028
fair	TRUE	18260	57.83972
good	FALSE	20989	27.74634
good	TRUE	54657	72.25366
very good	FALSE	15432	17.32298
very good	TRUE	73652	82.67702
excellent	FALSE	5287	11.67134
excellent	TRUE	40012	88.32866

За цим критерієм люди частіше занижують оцінку власного здоров'я. Так, частка людей, які займаються фізичними активностями серед людей, які оцінили свій стан здоров'я як поганий, не суттєва в порівнянні з тими, хто займається (майже 50 на 50).

Графік №8. Залежність ІМТ від оцінки респондентом стану власного здоров’я.



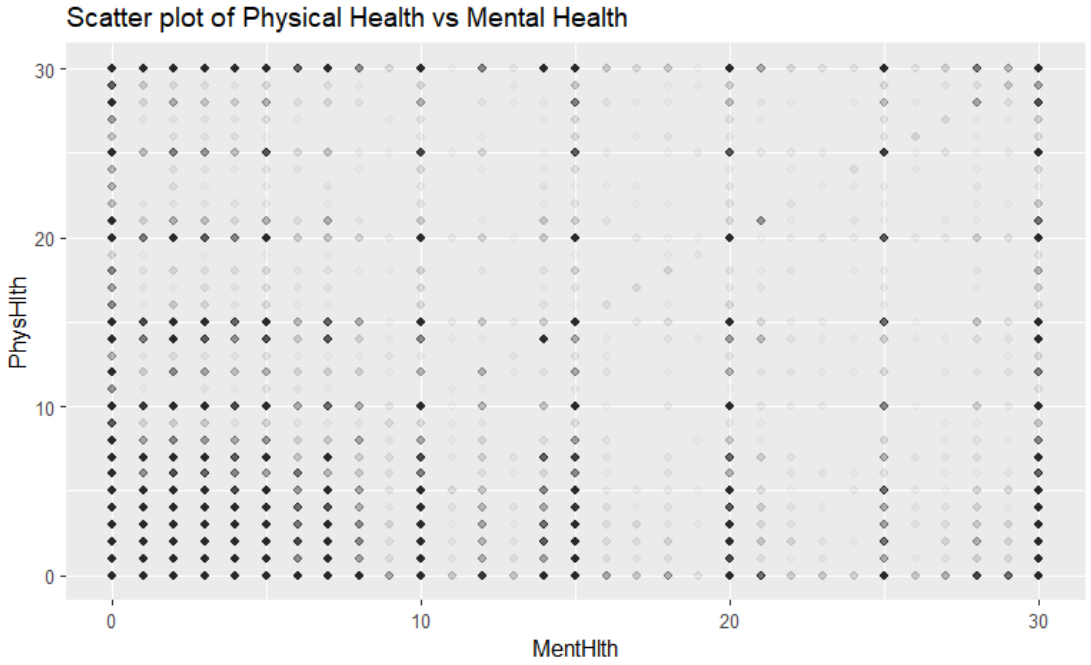
В poor і fair категоріях розподіли майже не відрізняються. Але можна прослідкувати, що чим вища категорія - тим нижчий імт. А медіанне ІМТ для poor відрізняється від медіанного ІМТ для excellent на 5.

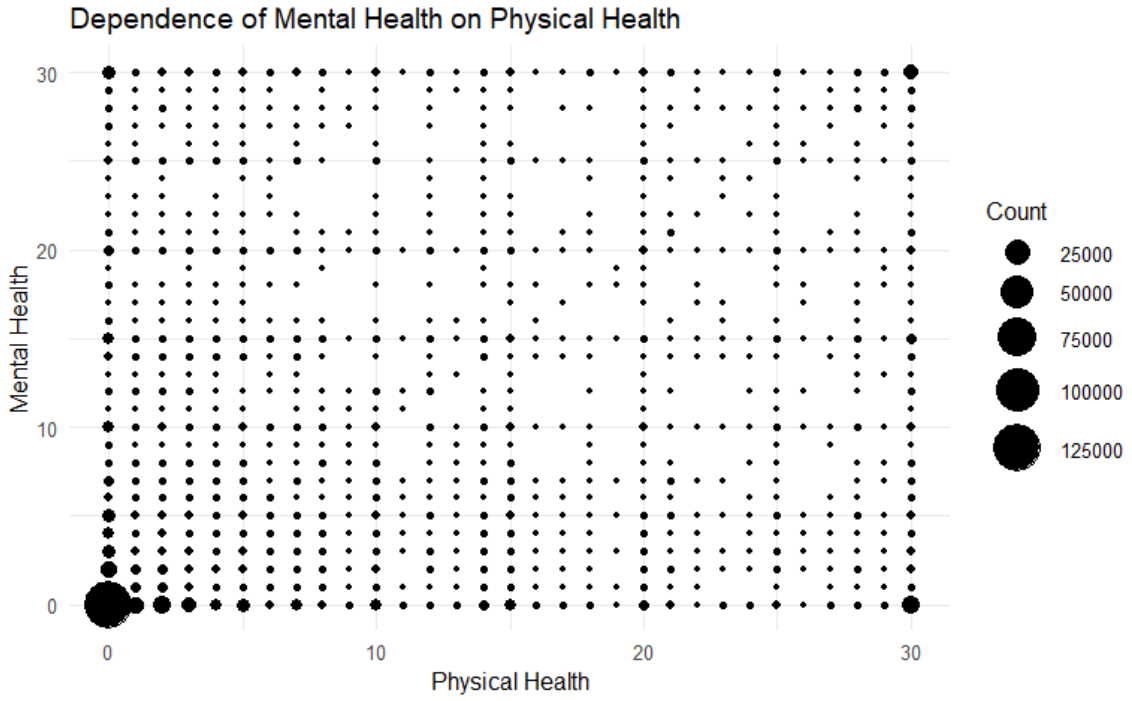
Як можна помітити наполовину в межах норми знаходиться розподіл людей, які вказали свій стан здоров’я як відмінний, а всі інші майже повністю за межами норми, а саме перевищують норму.

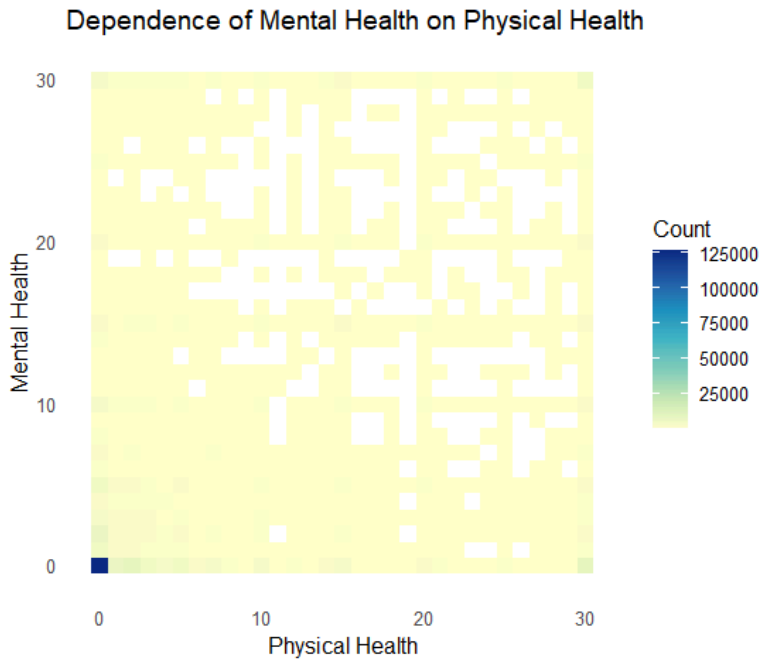
5. П’яте питання

“Чи впливає поганий ментальний стан на низький рівень фізичного здоров'я.”

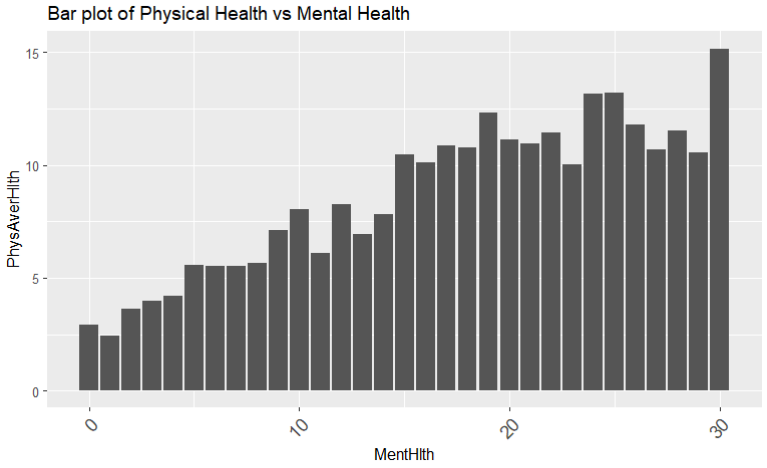
Графіки 1/2/3. Залежність PhysHlth від MentHlht. Діаграми розсіювання, теплова карта.

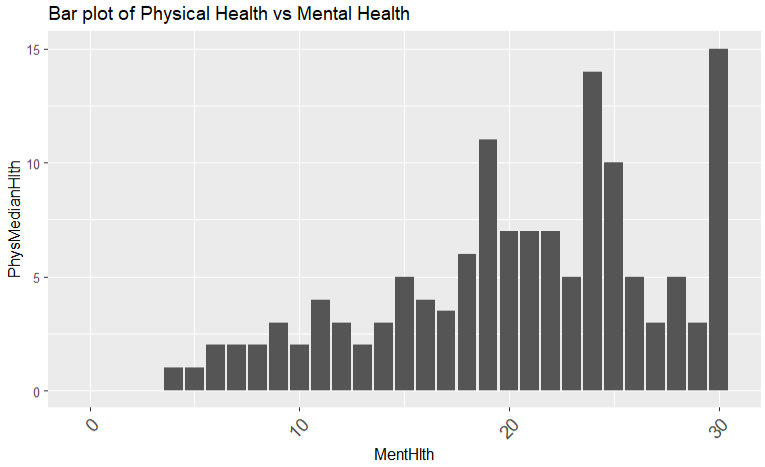






Графіки 4/5. Залежність PhysHlth від MentHlth. (Середні вибіркові\медіани в залежності від категорії MentHlth)



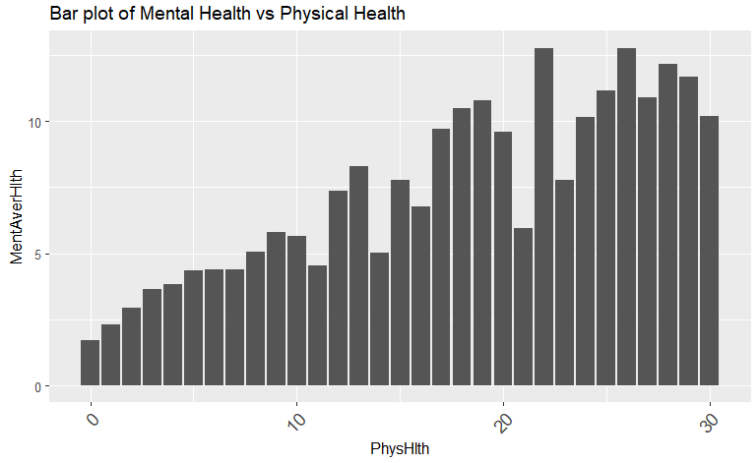


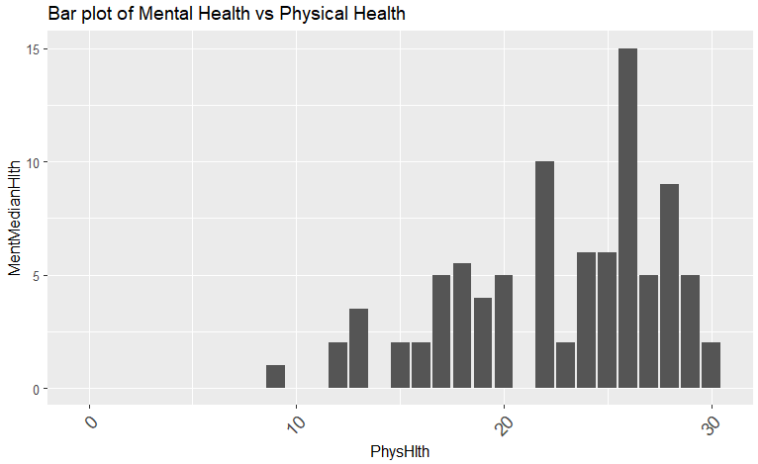
Щодо вибіркового середнього:

Чим вище значення ментального здоров'я (чим більше днів людина мала погане ментальне здоров'я), тим вищий показник середнього PhysMeanHlth (кількість днів, коли у людини була травма, або людина була хвора, або почувала себе фізично погано).

Розглянемо обернений випадок

Графіки 6/7. Залежність MentHlth від PhysHlth. (Середні вибіркові\медіани в залежності від категорії PhysHlth)

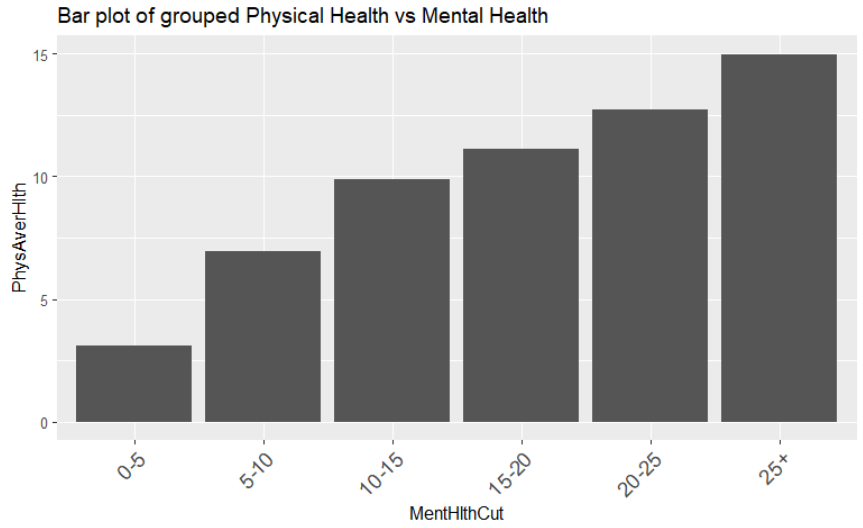


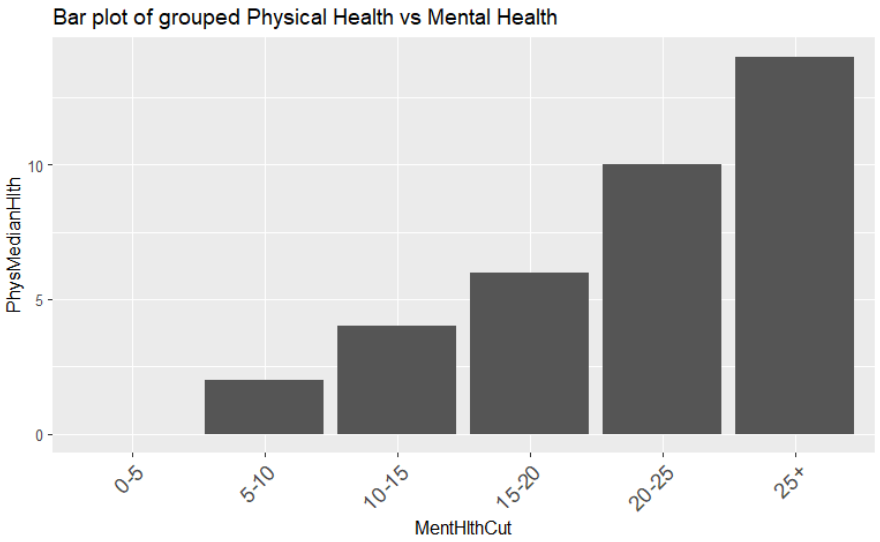


Щодо вибіркового середнього:

В інший бік залежність не така очевидна, спробуємо розбити на групи.

Графіки 8/9. Залежність PhysHltht від MentHlthCu. (Середні вибіркові\медіани в залежності від категорії MentHlth)

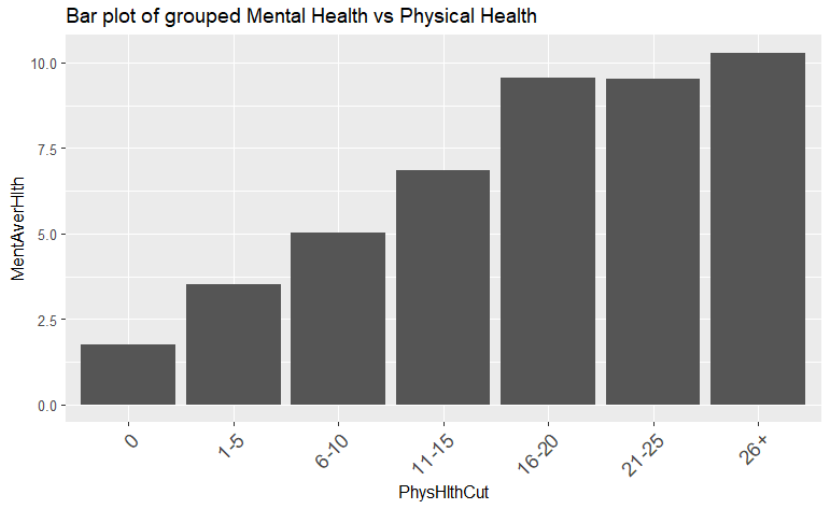


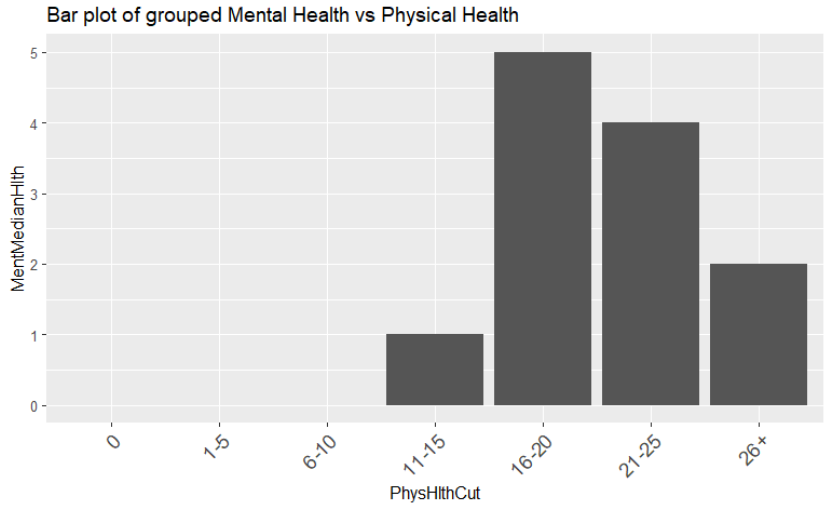


Має місце залежність. Чим вища категорія MentHlthCut - тим більше середнє\медіанне значення PhysHlth. Тобто чим гірше ментальне здоров’я - тим гірше фізичне.

Обернений випадок (залежність PhysHlthCut від MentHlth):

Графіки 10/11. Залежність PhysHlthCut від MentHlth. (Середні вибіркові\медіани в залежності від категорії PhysHlth)

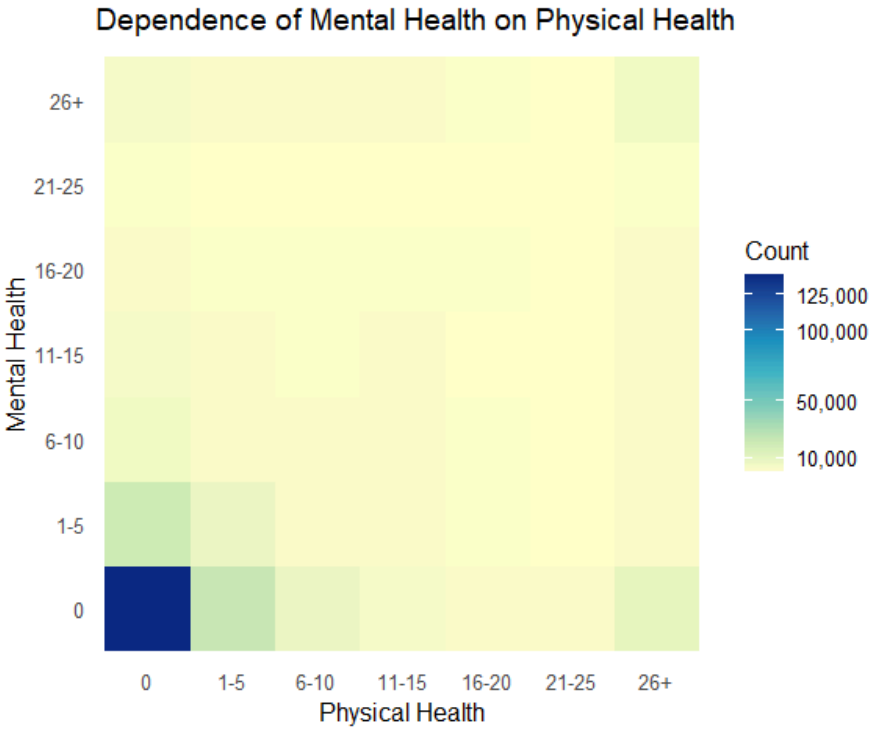




Як можна побачити, якщо судити по вибіркових середніх - залежність має місце. Якщо судити по медіанах - складно сказати.

Тепер побудуємо теплову карту для згрупованих MentHlth і PhysHlth

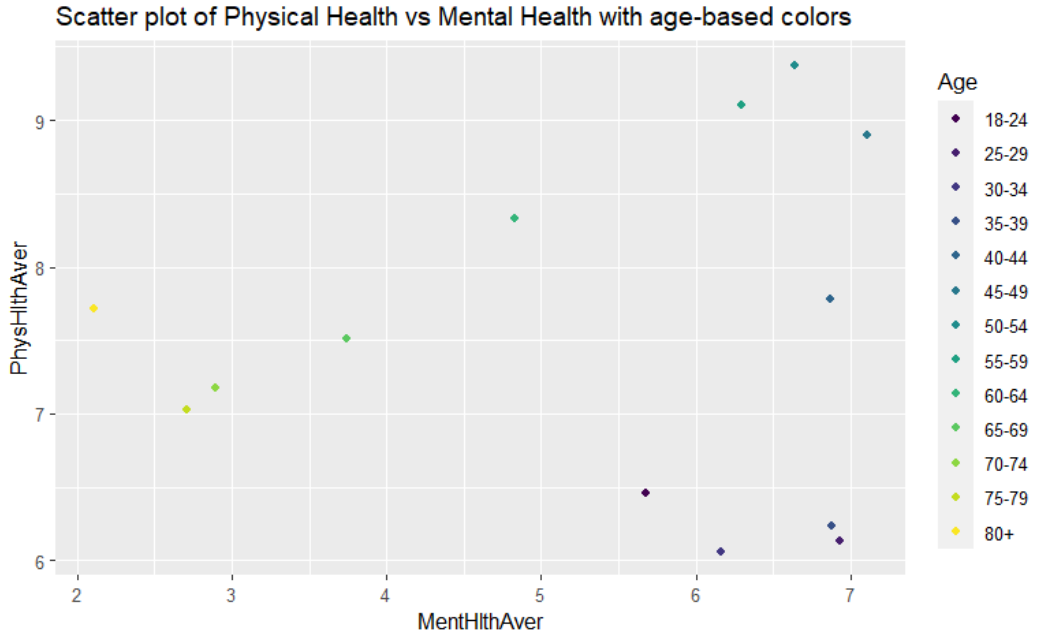
Графік 12. Теплова карта для згрупованих MentHlth і PhysHlth

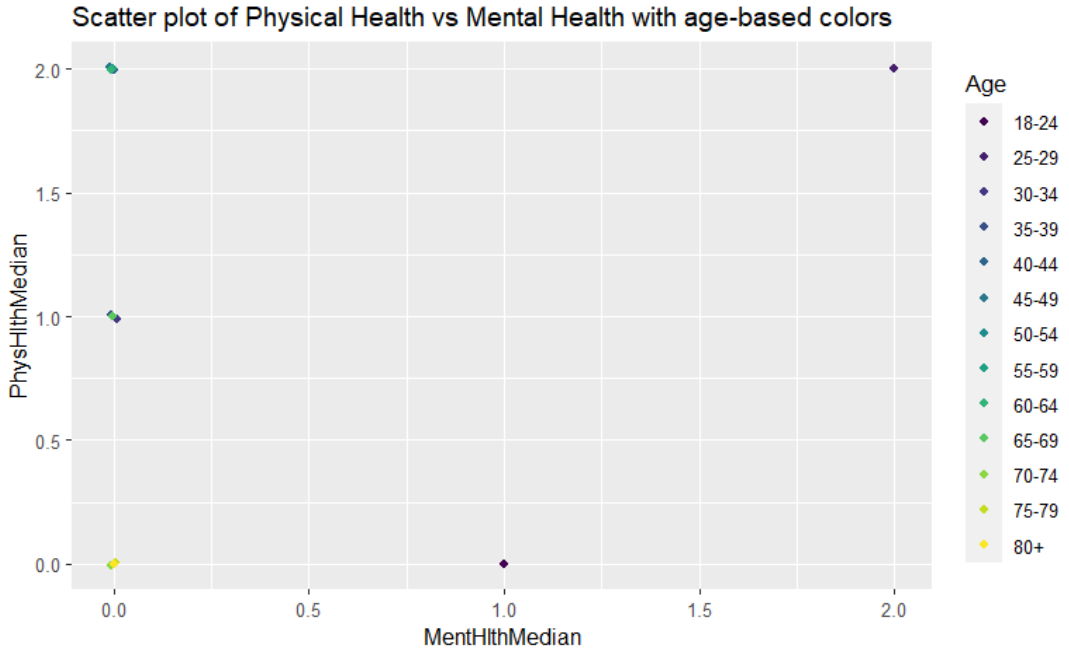


6. Шосте питання

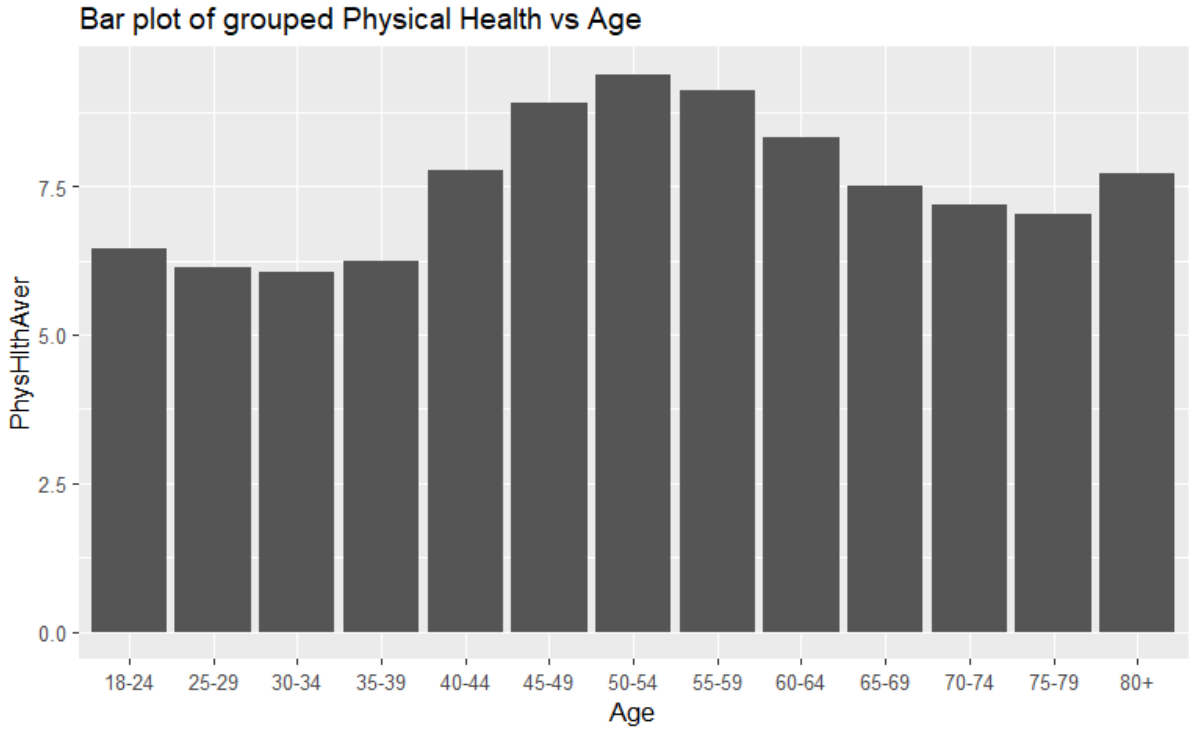
“Залежність ментального здоров'я, оцінки фізичного здоров'я, від віку серед людей, які хворіють на діабет.”

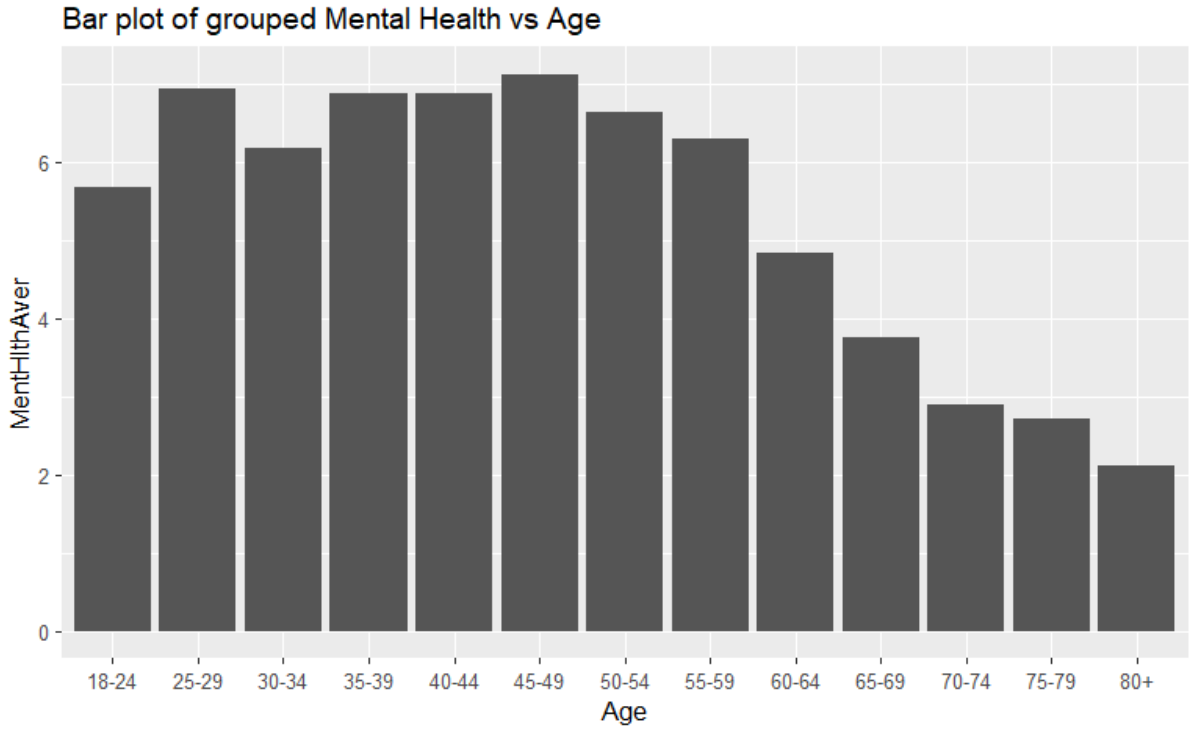
Графік 1. Діаграми розсіювання в залежності від вікової категорії для виб.середніх/медіан змінних PhysHlth і MentHlth серед людей хворих на діабет.



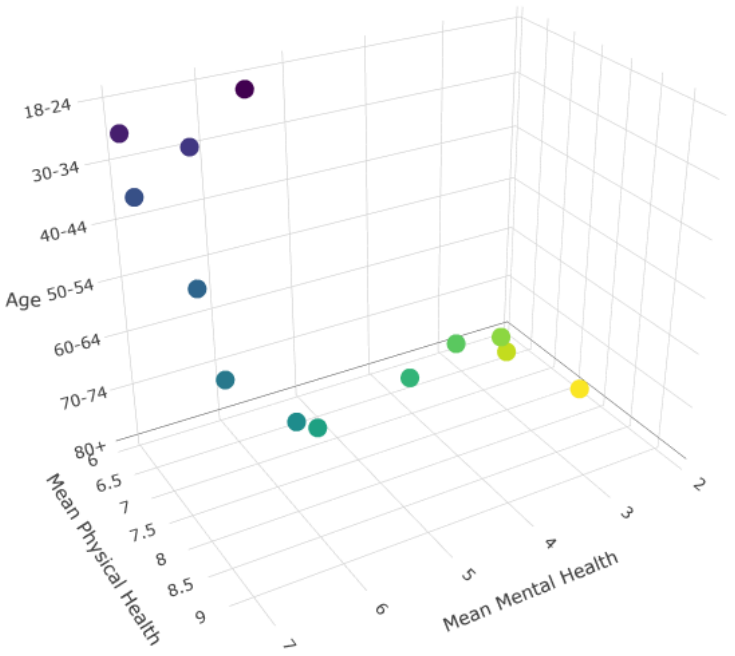


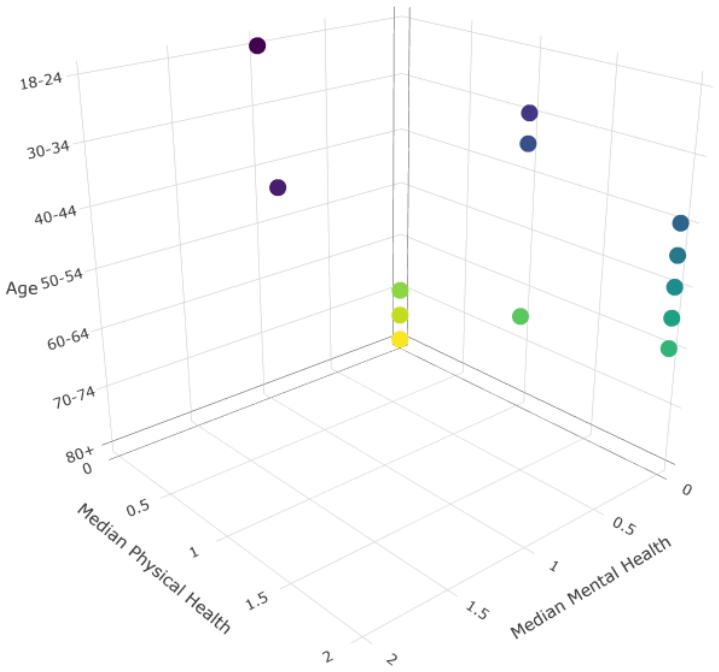
Графіки 2/3. Стовпчикові діаграми в залежності від вікової категорії для виб.середніх змінних PhysHlth і MentHlth серед людей хворих на діабет.





Графіки 4/5. 3-вимірні графіки діаграм розсіювання в залежності від вікової категорії для виб.середніх/медіан змінних PhysHlth і MentHlth серед людей хворих на діабет.





ВИСНОВКИ

Згідно із виконаними дослідженнями, було зроблено ряд висновків щодо поставлених на початку питань. Так було з’ясовано, що на наявність діабету можуть впливати вік, ІМТ, деякі інші хвороби (найбільшою мірою - артеріальний тиск та високий рівень холестерину) і навіть стать (чоловіки частіше мають діабет, ніж жінки). Стосовно соціальних факторів, має місце вплив доходів та рівня освіти (багаті та освічені рідко хворіють на діабет), а також видно, що діабетові сприяють вживання алкоголю, паління та погане харчування.

Дослідження третього питання показує, що люди без медичного страхування рідше обстежуються на рівень холестерину. В процесі цього виникли додаткові питання: чи є залежність наявності страховки від заробітної плати, чи є залежність змінної NoDocbcCost від заробітної плати та залежність наявності діабету від змінних CholCheck, NoDocbcCost. Видно залежність страховки від доходів, і залежність наявності діабету від обстеження холестерину та труднощів із вибором лікаря (більша частка людей з діабетом - при обстеженні холестерину та неможливості забезпечити собі лікаря).

Результати четвертого питання показують зменшення частки людей з діабетом по мірі покращення оцінки свого стану. Бачимо покращення рівня здоров’я при підвищенні оцінки свого здоров’я (найбільшу різницю показують графіки за рівнем холестерину, тиску, важкістю ходьби та фізичною активністю), а також можемо побачити покращення ІМТ у людей, які вище оцінюють своє здоров’я.

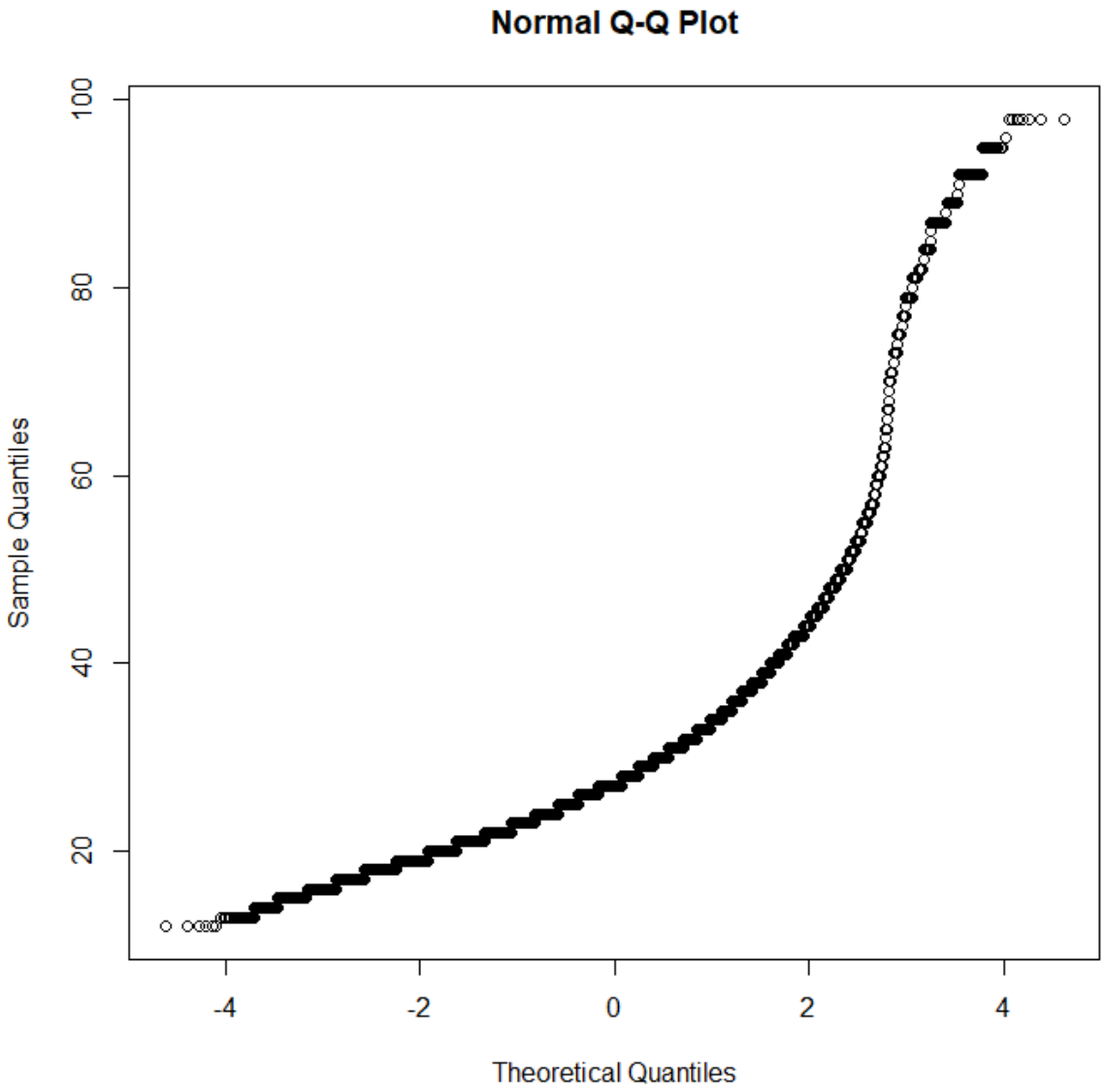
Результати п’ятого питання показують певний зв’язок середніх показників ментального та фізичного здоров’я, однак важко стверджувати, що по одному параметру можна судити про інший, а шостого - погіршення середнього показника ментального по мірі збільшення віку опитуваних, тоді як фізичне здоров’я майже не залежить від віку.

ДЖЕРЕЛА

1. Introduction to ggridges. (дата звернення: 28.05.2024)
2. 3D Scatter Plots in R. URL.
3. Ідеї для побудови групованих стовпчикових діаграм та “Violin” графіків були взяті з галереї графіків в R.
4. Матеріали лекції 2.

ДОДАТОК А. ГРАФІКИ ДЕСКРИПТИВНИХ ХАРАКТЕРИСТИК ЗМІННИХ

BMI
qqplot



Діаграма розкиду (Scatter Plot)