Data analysis Lab 1

2024-03-09

Назва команди - Команда №3

Перелік учасників колективу виконавців:

- Пономаренко Олександр (КМ-12)
- Земляний Даниїл (КМ-12)
- Борисенко Данило (КМ-11)
- Заіченко Дамир (КМ-13)
- Лук'яненко Василь (КМ-13)

Команду цікавили відповіді на наступні сформовані дослідницькі питання:

- Як змінюється кількість дорослих та дітей в залежності від типу номеру та плану харчування?
- Яким чином розподіляється кількість особливих побажань в залежності від типу номеру?
- Як впливає кількість дорослих і дітей на скасування бронювання?
- Що впливає на кількість проведених вихідних та робочих ночей у готелі?
- Як різні типи кімнат впливають на кількість попередніх скасувань/бронювань?
- Чи є різниця в кількості попередніх скасувань для клієнтів, які вимагають паркувальне місце і тих, хто його не потребує?

Очікувалося, що проведене дослідження також зможе підтвердити чи спростувати гіпотези подібного типу:

- у вихідні дні кімнати в отелі коштують дорожче, ніж у будні
- кількість дітей має вплив на кількість забронюваних ночей, відміну бронювання, потребу у паркувальному місці
- та інші "природні" речі

У цього датасету є більший "попередник", в описі якого вказано, що автор датасету брав натхнення з дослідження, яке використовувало дійсно існуючі дані, зібрані у двох готелях Португалії. Поточний датасет є очищеною версією, що не містить майже незадіяних змінних та купу missing data

на етапі очищення даних необхідно було переконатися, що відсутні неправильні кодування даних, що числові змінні, які стосуються, наприклад, часу, не є від'ємними і тп. В дійсності, як було зазначено вище, датасет вже з самого початку був очищений, що було вказано у його тегах на відповідному сайті kaggle, але додаткову перевірку все ж було проведено. Нижче наводиться деяка інформація, що підтверджує чистоту даних:

- перевірка даних на відсутність очевидних помилок, одруків, неправильного кодування тощо

```
describe(hotel)
```

```
## hotel
##
##  20  Variables      36275  Observations
## ---------------------------------------------------------------------------
## Booking_ID
##        n  missing distinct
##    36275        0    36275
##
## lowest : INN00001 INN00002 INN00003 INN00004 INN00005
## highest: INN36271 INN36272 INN36273 INN36274 INN36275
## ---------------------------------------------------------------------------
## no_of_adults
##        n  missing distinct     Info     Mean      Gmd
##    36275        0        5    0.617    1.845   0.4675
##
## Value          0     1     2     3     4
## Frequency    139  7695 26108  2317    16
## Proportion 0.004 0.212 0.720 0.064 0.000
##
## For the frequency table, variable is rounded to the nearest 0
## ---------------------------------------------------------------------------
## no_of_children
##        n  missing distinct     Info     Mean      Gmd
##    36275        0        6    0.207   0.1053   0.1977
##
## Value          0     1     2     3     9    10
## Frequency  33577  1618  1058    19     2     1
## Proportion 0.926 0.045 0.029 0.001 0.000 0.000
##
## For the frequency table, variable is rounded to the nearest 0
## ---------------------------------------------------------------------------
## no_of_weekend_nights
##        n  missing distinct     Info     Mean      Gmd
##    36275        0        8    0.863   0.8107   0.9145
##
## Value          0     1     2     3     4     5     6     7
```

```
## Value          0     1     2     3     4     5     6     7
## Frequency  16872  9995  9071   153   129    34    20     1
## Proportion 0.465 0.276 0.250 0.004 0.004 0.001 0.001 0.000
##
## For the frequency table, variable is rounded to the nearest 0
## ----------------------------------------------------------------------------
## no_of_week_nights
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    36275        0       18     0.94    2.204    1.456        0        1
##      .25      .50      .75      .90      .95
##        1        2        3        4        5
##
## Value          0     1     2     3     4     5     6     7     8     9    10
## Frequency   2387  9488 11444  7839  2990  1614   189   113    62    34    62
## Proportion 0.066 0.262 0.315 0.216 0.082 0.044 0.005 0.003 0.002 0.001 0.002
##
## Value         11    12    13    14    15    16    17
## Frequency     17     9     5     7    10     2     3
## Proportion 0.000 0.000 0.000 0.000 0.000 0.000 0.000
##
## For the frequency table, variable is rounded to the nearest 0
## ----------------------------------------------------------------------------
## type_of_meal_plan
##        n  missing distinct
##    36275        0        4
##
## Value       Meal Plan 1  Meal Plan 2  Meal Plan 3 Not Selected
## Frequency         27835         3305            5         5130
## Proportion        0.767        0.091        0.000        0.141
## ----------------------------------------------------------------------------
## required_car_parking_space
##        n  missing distinct
##    36275        0        2
##
## Value          0     1
## Frequency  35151  1124
## Proportion 0.969 0.031
## ----------------------------------------------------------------------------
## room_type_reserved
##        n  missing distinct
##    36275        0        7
##
## Value      Room_Type 1 Room_Type 2 Room_Type 3 Room_Type 4 Room_Type 5
## Frequency        28130         692           7        6057         265
## Proportion       0.775       0.019       0.000       0.167       0.007
##
## Value      Room_Type 6 Room_Type 7
## Frequency          966         158
## Proportion       0.027       0.004
## ----------------------------------------------------------------------------
## lead_time
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    36275        0      352        1    85.23    90.63        1        3
##      .25      .50      .75      .90      .95
##       17       57      126      213      273
##
## lowest :   0   1   2   3   4, highest: 381 386 418 433 443
## ----------------------------------------------------------------------------
## arrival_year
##        n  missing distinct     Info     Mean      Gmd
##    36275        0        2    0.442     2018   0.2947
##
## Value       2017  2018
## Frequency   6514 29761
## Proportion  0.18  0.82
## ----------------------------------------------------------------------------
## arrival_month
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    36275        0       12     0.99    7.424    3.497        2        3
##      .25      .50      .75      .90      .95
##        5        8       10       11       12
##
## Value          1     2     3     4     5     6     7     8     9    10    11
## Frequency   1014  1704  2358  2736  2598  3203  2920  3813  4611  5317  2980
## Proportion 0.028 0.047 0.065 0.075 0.072 0.088 0.080 0.105 0.127 0.147 0.082
##
## Value         12
## Frequency   3021
```

```
## Proportion 0.083
##
## For the frequency table, variable is rounded to the nearest 0
## ----------------------------------------------------------------------------
## arrival_date
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    36275        0       31    0.999     15.6    10.08        2        4
##      .25      .50      .75      .90      .95
##        8       16       23       28       29
##
## lowest :  1  2  3  4  5, highest: 27 28 29 30 31
## ----------------------------------------------------------------------------
## market_segment_type
##        n  missing distinct
##    36275        0        5
##
## Value          Aviation Complementary     Corporate       Offline
## Frequency            125           391          2017         10528
## Proportion         0.003         0.011         0.056         0.290
##
## Value             Online
## Frequency          23214
## Proportion         0.640
## ----------------------------------------------------------------------------
## repeated_guest
##        n  missing distinct
##    36275        0        2
##
## Value          0     1
## Frequency  35345   930
## Proportion 0.974 0.026
## ----------------------------------------------------------------------------
## no_of_previous_cancellations
##        n  missing distinct     Info     Mean      Gmd
##    36275        0        9    0.028  0.02335  0.04647
##
## Value          0     1     2     3     4     5     6    11    13
## Frequency  35937   198    46    43    10    11     1    25     4
## Proportion 0.991 0.005 0.001 0.001 0.000 0.000 0.000 0.001 0.000
##
## For the frequency table, variable is rounded to the nearest 0
## ----------------------------------------------------------------------------
## no_of_previous_bookings_not_canceled
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    36275        0       59    0.066   0.1534    0.304        0        0
##      .25      .50      .75      .90      .95
##        0        0        0        0        0
##
## lowest :  0  1  2  3  4, highest: 54 55 56 57 58
## ----------------------------------------------------------------------------
## avg_price_per_room
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    36275        0     3930        1    103.4    37.47    61.00    67.00
##      .25      .50      .75      .90      .95
##    80.30    99.45   120.00   147.60   165.00
##
## lowest : 0      0.5    1      1.48   1.6    , highest: 332.57 349.63 365    375.5  540
## ----------------------------------------------------------------------------
## no_of_special_requests
##        n  missing distinct     Info     Mean      Gmd
##    36275        0        6    0.805   0.6197   0.7848
##
## Value          0     1     2     3     4     5
## Frequency  19777 11373  4364   675    78     8
## Proportion 0.545 0.314 0.120 0.019 0.002 0.000
##
## For the frequency table, variable is rounded to the nearest 0
## ----------------------------------------------------------------------------
## booking_status
##        n  missing distinct
##    36275        0        2
##
## Value          Canceled Not_Canceled
## Frequency         11885        24390
## Proportion        0.328        0.672
## ----------------------------------------------------------------------------
## arrival_year_and_month
##        n  missing distinct
```

```
##   36275       0     550
##
## lowest : 2017-10-1  2017-10-10 2017-10-11 2017-10-12 2017-10-13
## highest: 2018-9-5   2018-9-6   2018-9-7   2018-9-8   2018-9-9
## ---------------------------------------------------------------------------
```

- перевірка на кількість пропущених значень для кожної змінної

```
hotel %>% summarise(across(everything(), ~ sum(is.na(.)))) %>%
  select(where(~ all(.) > 0))
```

```
0 rows
```

- перевірка на частку пропущених значень для кожної змінної

```
hotel %>% summarise(across(everything(), ~ mean(is.na(.)))) %>%
  select(where(~ all(.) > 0))
```

```
## Warning in all(.): coercing argument of type 'double' to logical

## Warning in all(.): coercing argument of type 'double' to logical

## Warning in all(.): coercing argument of type 'double' to logical

## Warning in all(.): coercing argument of type 'double' to logical

## Warning in all(.): coercing argument of type 'double' to logical

## Warning in all(.): coercing argument of type 'double' to logical

## Warning in all(.): coercing argument of type 'double' to logical

## Warning in all(.): coercing argument of type 'double' to logical

## Warning in all(.): coercing argument of type 'double' to logical

## Warning in all(.): coercing argument of type 'double' to logical

## Warning in all(.): coercing argument of type 'double' to logical

## Warning in all(.): coercing argument of type 'double' to logical

## Warning in all(.): coercing argument of type 'double' to logical

## Warning in all(.): coercing argument of type 'double' to logical

## Warning in all(.): coercing argument of type 'double' to logical

## Warning in all(.): coercing argument of type 'double' to logical

## Warning in all(.): coercing argument of type 'double' to logical

## Warning in all(.): coercing argument of type 'double' to logical

## Warning in all(.): coercing argument of type 'double' to logical
```

```
0 rows
```

- переведення відповідних змінних у фактор

```
hotel <- hotel %>% mutate(repeated_guest = as.factor(repeated_guest),
                          required_car_parking_space = as.factor(required_car_parking_space),
                          booking_status = as.factor(booking_status),
                          room_type_reserved = as.factor(room_type_reserved))
```

- Подивимось на розмірність датасету. Можемо побачити, що в ньому 36275 спостережень і 19 змінних

```
str(hotel, give.attr = FALSE)
```

```
## 'data.frame':    36275 obs. of  20 variables:
##  $ Booking_ID                  : chr  "INN00001" "INN00002" "INN00003" "INN00004" ...
##  $ no_of_adults                : int  2 2 1 2 2 2 2 2 3 2 ...
##  $ no_of_children              : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ no_of_weekend_nights        : int  1 2 2 0 1 0 1 1 0 0 ...
##  $ no_of_week_nights           : int  2 3 1 2 1 2 3 3 4 5 ...
##  $ type_of_meal_plan           : chr  "Meal Plan 1" "Not Selected" "Meal Plan 1" "Meal Plan 1" ...
##  $ required_car_parking_space  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ room_type_reserved          : Factor w/ 7 levels "Room_Type 1",..: 1 1 1 1 1 1 1 4 1 4 ...
##  $ lead_time                   : int  224 5 1 211 48 346 34 83 121 44 ...
##  $ arrival_year                : int  2017 2018 2018 2018 2018 2018 2017 2018 2018 2018 ...
##  $ arrival_month               : int  10 11 2 5 4 9 10 12 7 10 ...
##  $ arrival_date                : int  2 6 28 20 11 13 15 26 6 18 ...
##  $ market_segment_type         : chr  "Offline" "Online" "Online" "Online" ...
##  $ repeated_guest              : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ no_of_previous_cancellations : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ no_of_previous_bookings_not_canceled: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ avg_price_per_room          : num  65 106.7 60 100 94.5 ...
##  $ no_of_special_requests      : int  0 1 0 0 0 1 1 1 1 3 ...
##  $ booking_status              : Factor w/ 2 levels "Canceled","Not_Canceled": 2 2 1 1 1 1 2 2 2 2 ...
##  $ arrival_year_and_month      : chr  "2017-10-2" "2018-11-6" "2018-2-28" "2018-5-20" ...
```

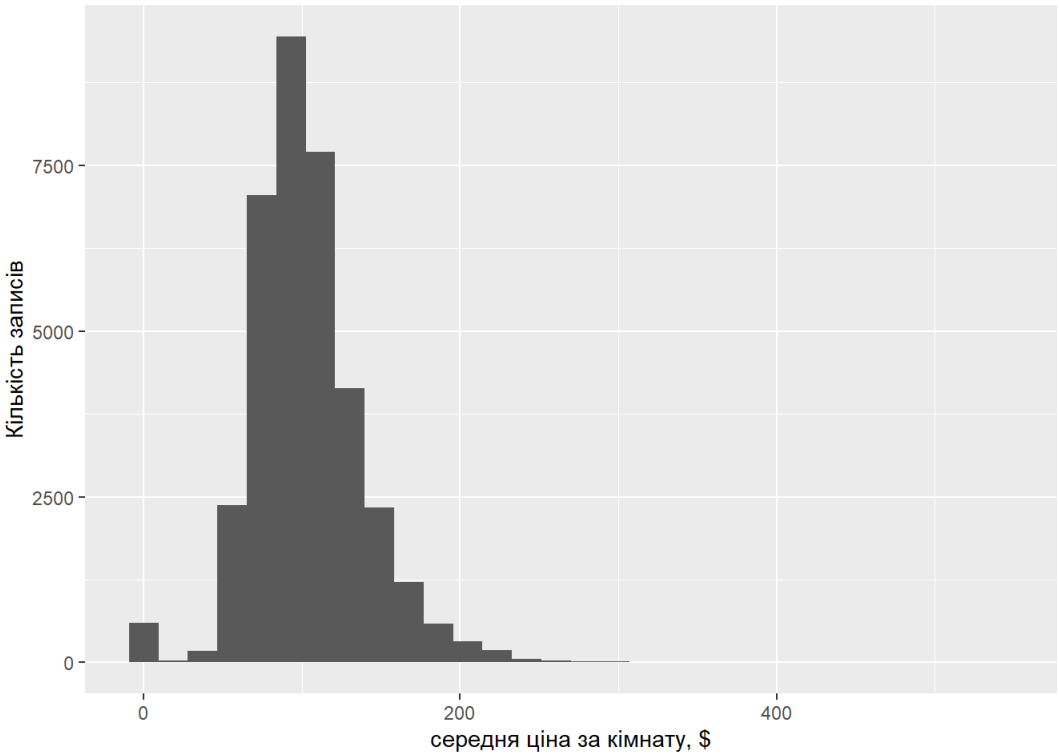
Викиди

розглянемо конкретні результати, які було отримано під час проведення EDA

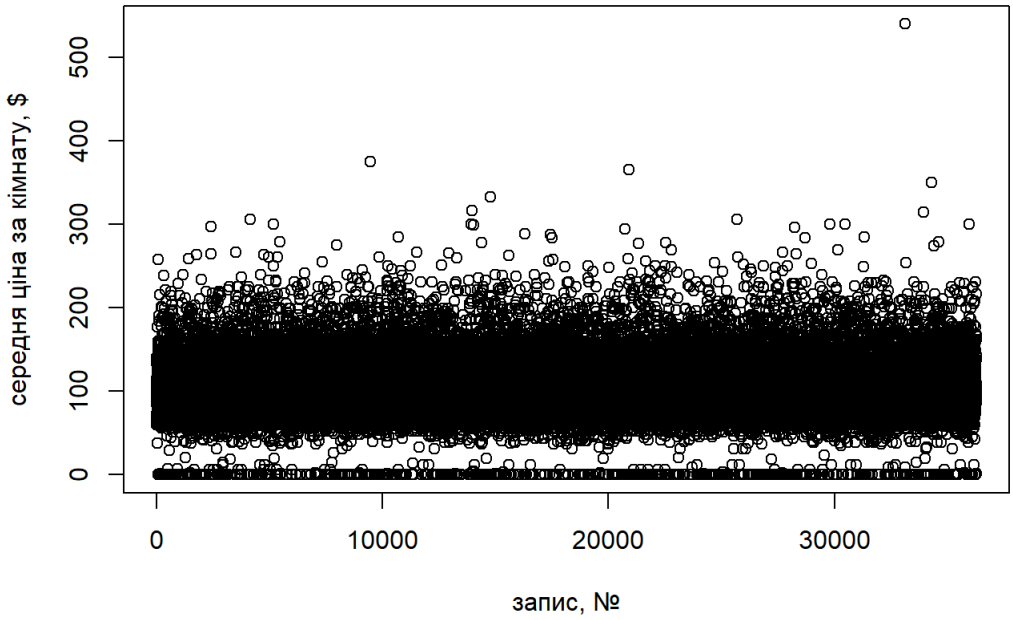
- спочатку переглянемо, яким чином виглядає графік для середньої ціни за кімнату

```
ggplot(hotel, aes(x = avg_price_per_room)) +
  geom_histogram() + labs(x = "середня ціна за кімнату, $", y = "Кількість записів")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
plot(hotel$avg_price_per_room, xlab = 'запис, №', ylab = 'середня ціна за кімнату, $')
```



перше зауваження стосується цін, які близькі до нуля і утворюють цей розрив на графіку. Візьмемо середнє значення ціни умовно 10$ і переглянемо усі резервації, ціна яких менше 10$. Обравши серед цього зрізу змінну market_segment_type, побачимо наступну картину:

```
segment_types <- hotel %>% filter(avg_price_per_room < 10) %>% select(market_segment_type)
table(segment_types)
```

```
## market_segment_type
## Complementary        Online
##           373           225
```

ці резервації можна вважати резервуаціями за “договірну” ціну, де ті, що Online, можна розглядати як домовленість по телефону чи по іншим засобам зв’язку

- скориставшись фільтром Гампеля було виявлено приблизні пороги цін, до яких і після яких теоретично можуть знаходитись викиди

```
hotel %>% filter(avg_price_per_room < median(avg_price_per_room) - 3*mad(avg_price_per_room) | avg_price_per_room >
median(avg_price_per_room) + 3*mad(avg_price_per_room)) %>%
  arrange(avg_price_per_room)
```

Booking_ID <chr>
INN00064
INN00146
INN00210
INN00267
INN00268
INN00289
INN00347
INN00416
INN00432
INN00541

1-10 of 1,344 rows | 1-1 of 20 columns Previous **1** 2 Next

подивимось на кількість резервацій з ціною у 0$

```
hotel %>% filter(avg_price_per_room == 0)
```

Booking_ID <chr>
INN00064
INN00146
INN00210
INN00267
INN00268
INN00289
INN00347
INN00416
INN00432
INN00541

1-10 of 545 rows | 1-1 of 20 columns Previous 1 2 Next

як бачимо, таких рядків 545. Враховуючи те, що записів із ціною резервації <10$ було 373 + 225 = 598, то нічим іншим як договірною ціною цей випадок не пояснюється

тепер, що стосується дорогих кімнат. Розглянемо кімнати із середньою ціною >300$, що перетинається з результатами фільтра Гампеля:

- перегляд інформації про дорогі кімнати

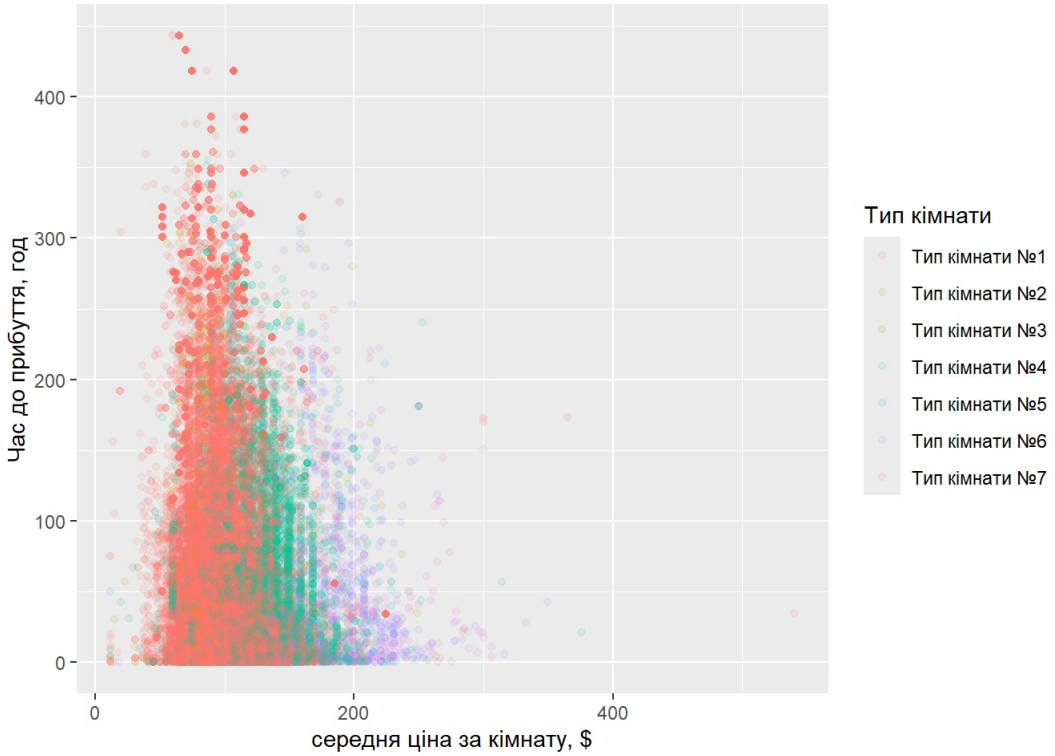
```
hotel %>% filter(avg_price_per_room > 300)
```

Booking_ID <chr>
INN04151
INN09462
INN13945
INN14774
INN20901
INN25671
INN33115
INN33956
INN34307

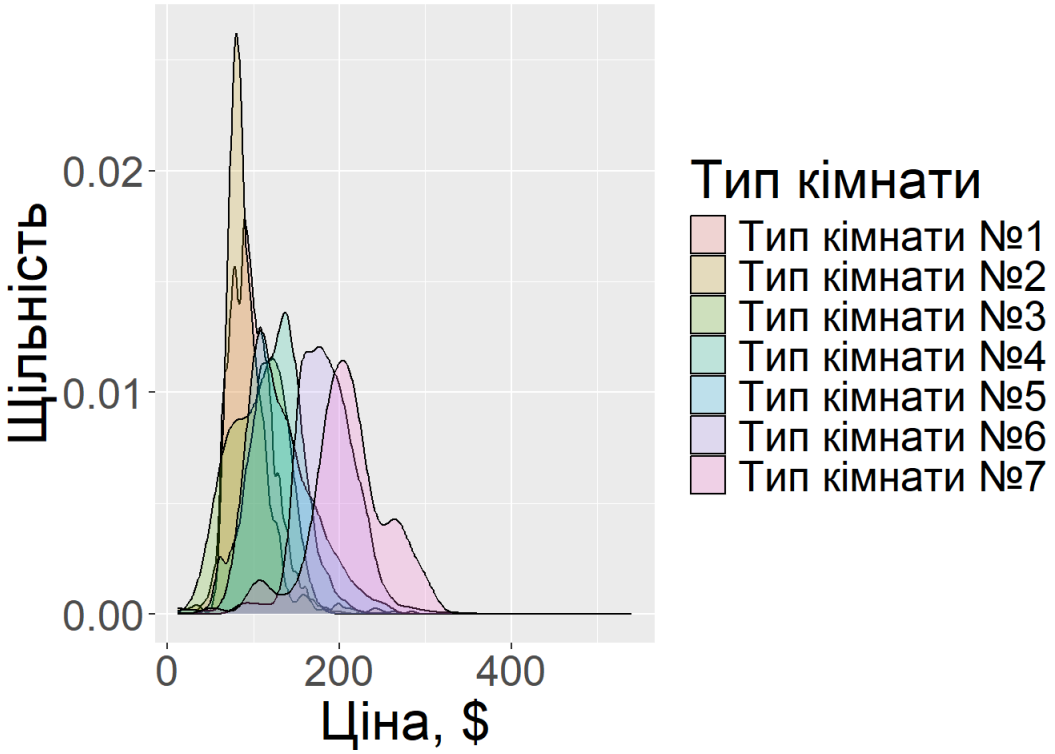
9 rows | 1-1 of 20 columns

- переглянемо чи усі ці резервації мають адекватну для свого типу кімнати ціну:

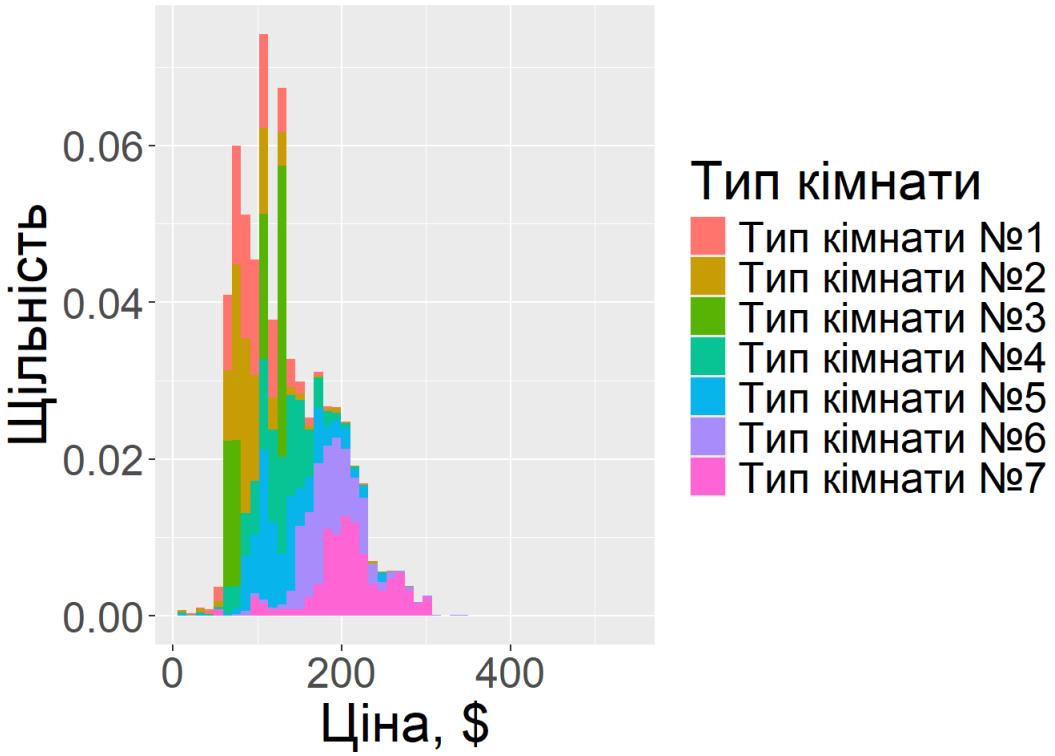
```
hotel_filtered <- subset(hotel, avg_price_per_room > 9)

ggplot(hotel_avg_prices, aes(x = avg_price_per_room, y = lead_time, color = as.factor(room_type_reserved))) +
  geom_point(alpha = 0.1) + labs (x = "середня ціна за кімнату, $", y = "Час до прибуття, год") +
  scale_color_manual(name = "Тип кімнати",
                     values = room_type_vector,
                     labels = room_label_vector)
```




```
ggplot(hotel_filtered ,
       aes(x = avg_price_per_room, y = after_stat(density),
           fill = room_type_reserved)) +
  geom_density(alpha = 0.2) +
  labs(x = "Ціна, $", y = "Щільність", fill = "Тип кімнати") +
  theme(axis.title = element_text(size = 25),
        axis.text = element_text(size = 20),
        legend.title = element_text(size = 25),
        legend.text = element_text(size = 20)) +
  scale_fill_manual(name = "Тип кімнати",
                    values = room_type_vector,
                    labels = room_label_vector)
```



```
ggplot(hotel_filtered ,
       aes(x = avg_price_per_room, y = after_stat(density),
           fill = room_type_reserved)) +
  geom_histogram(bins = 50) +
  labs(x = "Ціна, $", y = "Щільність", color = "Тип кімнати") +
  theme(axis.title = element_text(size = 25),
        axis.text = element_text(size = 20),
        legend.title = element_text(size = 25),
        legend.text = element_text(size = 20)) +
  scale_fill_manual(name = "Тип кімнати",
                    values = room_type_vector,
                    labels = room_label_vector)
```



- як бачимо, кімнати першого типу є відносно дешевими, що не стикується з тим, що резервація з найбільшою середньою ціною була якраз на Room Type 1:

```
hotel %>% filter(avg_price_per_room == max(avg_price_per_room))
```

Booking_ID <chr>
INN33115

1 row | 1-1 of 20 columns

- на нашу думку, цей запис можна вважати викидом. Ймовірно у ціну був записаний зайвий ‘0’ у кінці. Але в силу того, що такий запис лише один, і всі інші записи з ціною >300$ відносно нього та свого типу кімнати виглядають нормально, то ним можна знехтувати, адже дискриптивні статистики від цього особливо не спотворюються

Перейдемо безпосередньо до дослідницьких питань

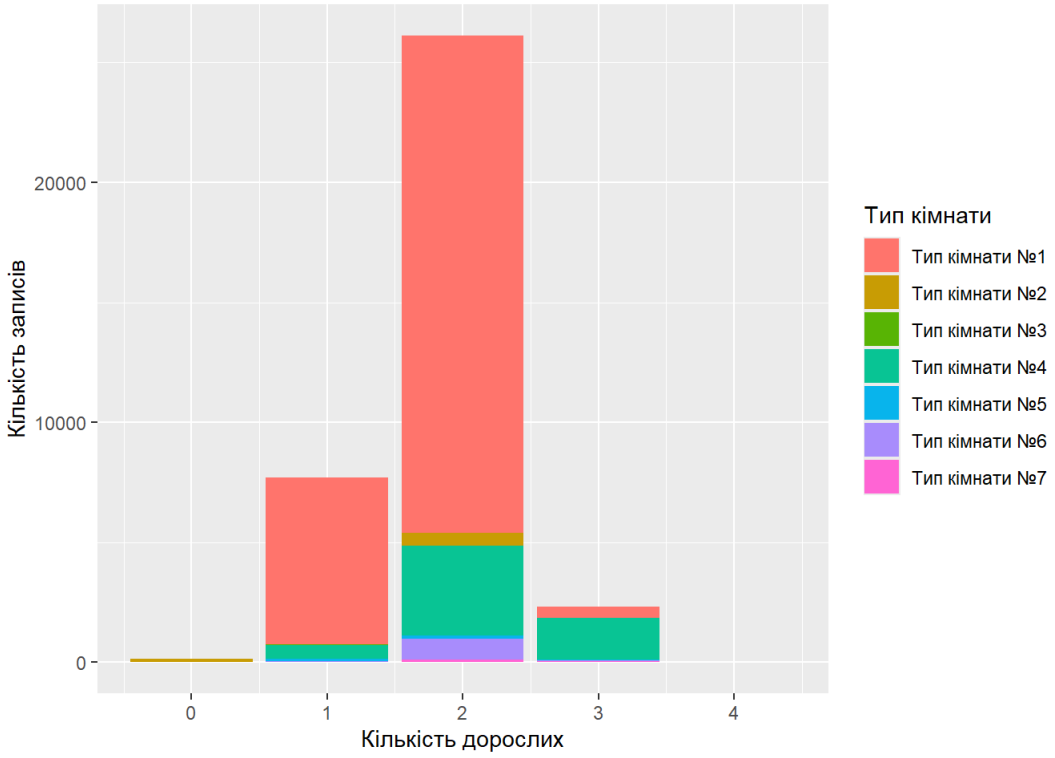
ПЕРШЕ ЗАПИТАННЯ

Формулювання:

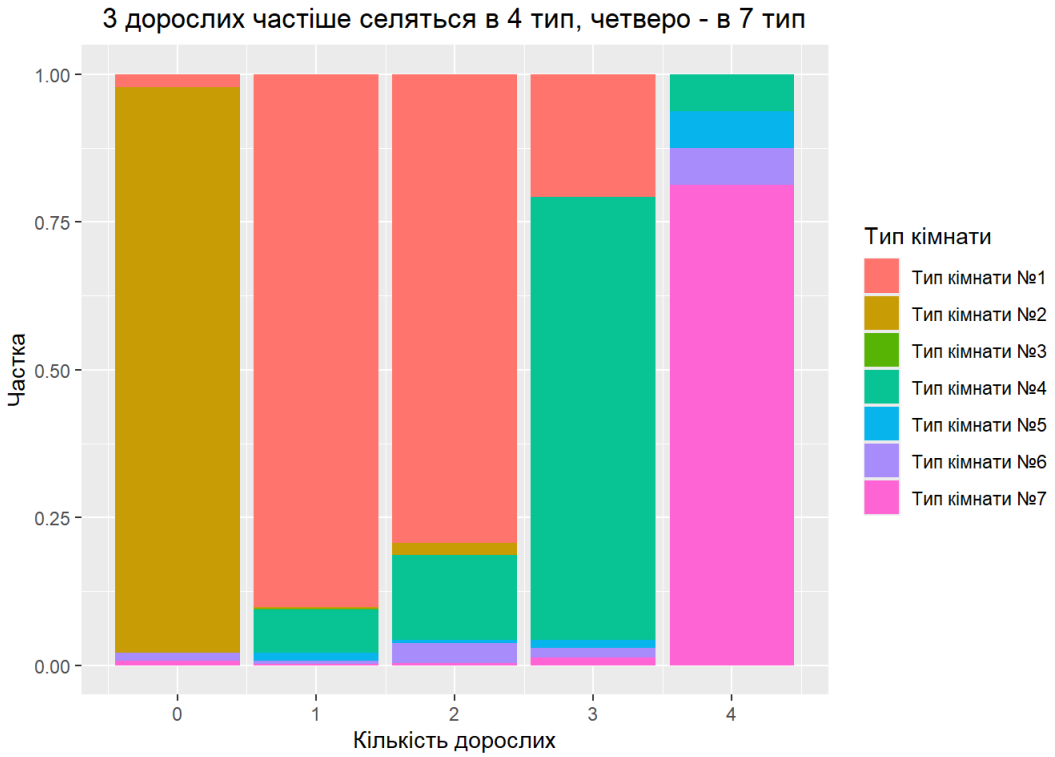
- Як змінюється кількість дорослих та дітей в залежності від типу номеру та плану харчування?

Про випадки, коли кількість дорослих дорівнює чотирьом або нулю, важко зробити якісь висновки через замалу частку таких записів. Бачимо, що один або двоє дорослих частіше бронюють номери першого типу, троє дорослих - номери четвертого типу.

```
ggplot(hotel, aes(x = no_of_adults, fill = as.factor(room_type_reserved))) + geom_bar() +
  labs(x = "Кількість дорослих", y = "Кількість записів", fill = "Тип Кімнати")+
  scale_fill_manual(name = "Тип кімнати",
                    values = room_type_vector,
                    labels = room_label_vector) + theme(plot.title = element_text(hjust = 0.5))
```

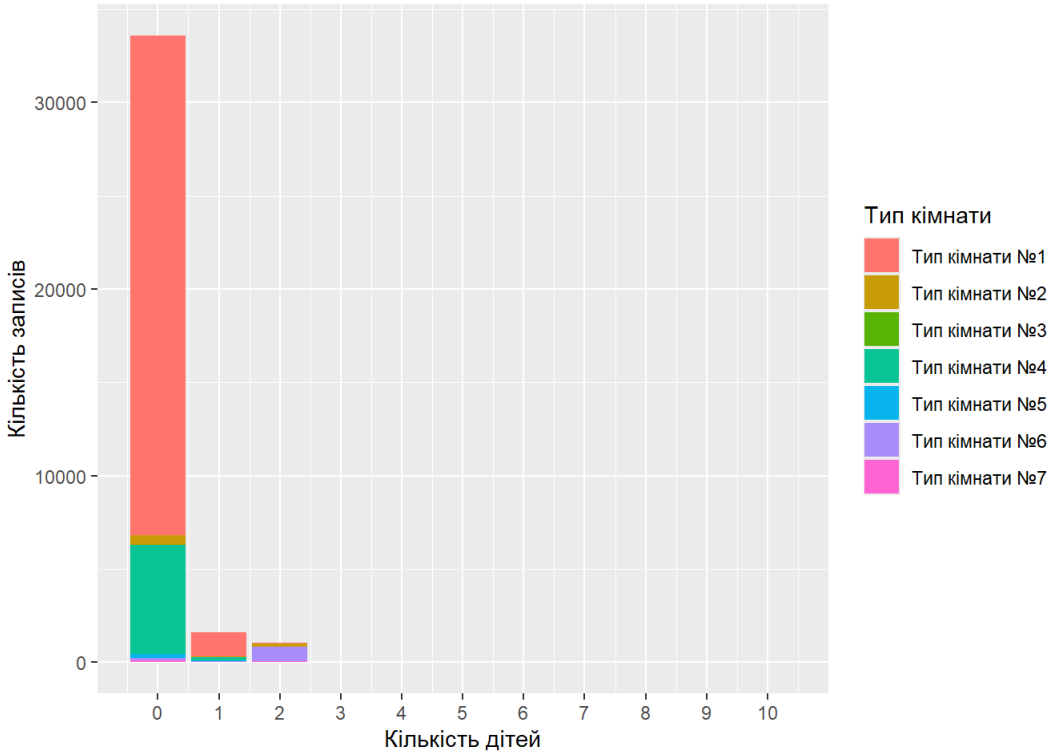



```
ggplot(hotel, aes(x = no_of_adults, fill = as.factor(room_type_reserved))) + geom_bar(position = "fill") +
  labs(title = "З дорослих частіше селяться в 4 тип, четверо - в 7 тип", x = "Кількість дорослих", y = "Частка", fill = "Тип Кімнати")+
  scale_fill_manual(name = "Тип кімнати",
                    values = room_type_vector,
                    labels = room_label_vector) + theme(plot.title = element_text(hjust = 0.5))
```

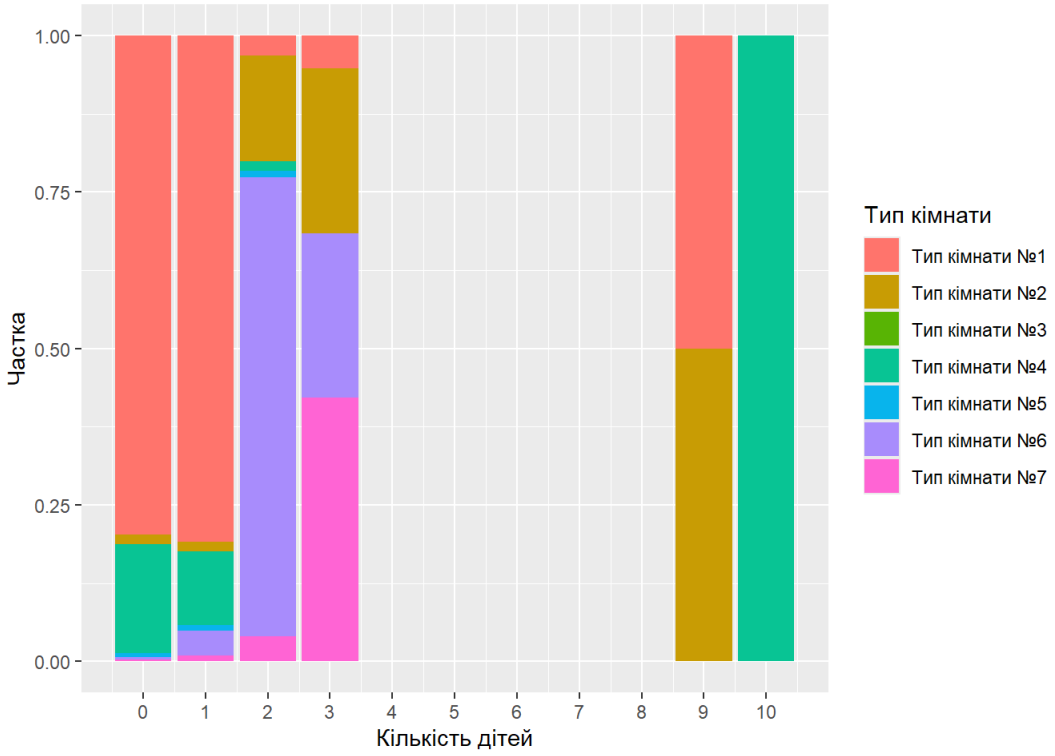


Про випадки, де кількість дітей більше або дорівнює трьом, важко щось сказати через замалу частку таких записів. Коли дитина одна або дітей немає, частіше всього броюнюють номер першого типу, для двох дітей - частіше номер шостого типу.

```
ggplot(hotel, aes(x = no_of_children, fill = as.factor(room_type_reserved))) + geom_bar() +
  labs(x = "Кількість дітей", y = "Кількість записів", fill = "Тип Кімнати") +
  scale_x_continuous(breaks = seq(0, max(hotel$no_of_children), by = 1))+
  scale_fill_manual(name = "Тип кімнати",
                    values = room_type_vector,
                    labels = room_label_vector) + theme(plot.title = element_text(hjust = 0.5))
```

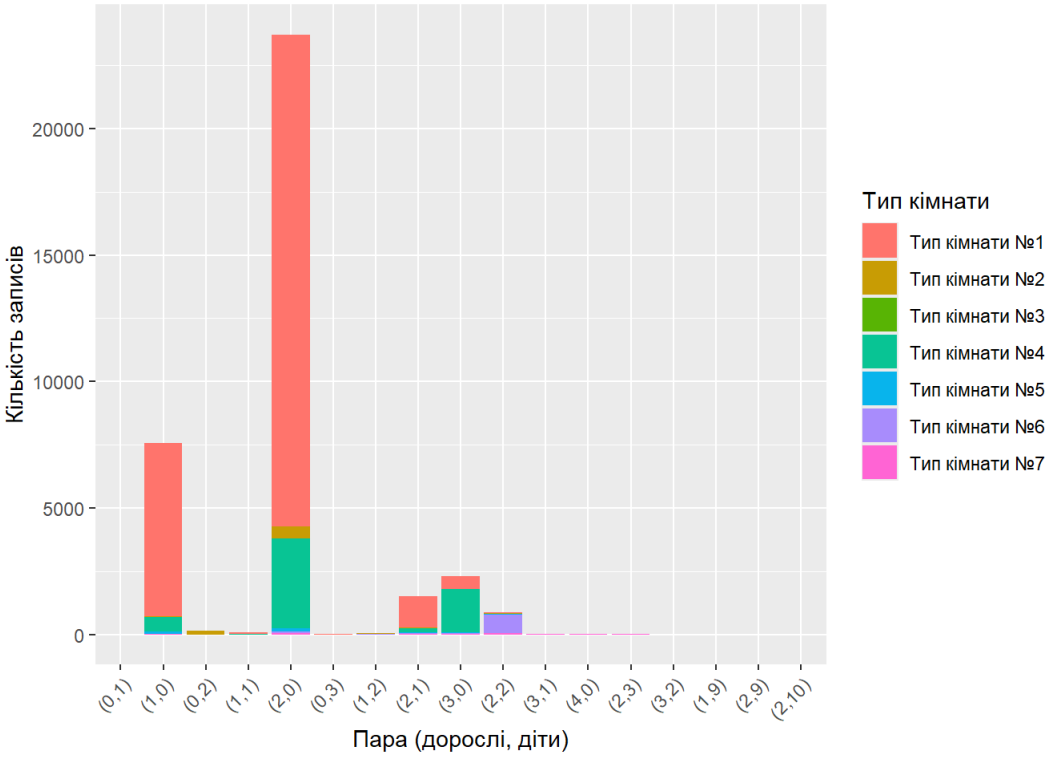



```
ggplot(hotel, aes(x = no_of_children, fill = as.factor(room_type_reserved))) + geom_bar(position = "fill") +
  labs(x = "Кількість дітей", y = "Частка", fill = "Тип Кімнати") +
  scale_x_continuous(breaks = seq(0, max(hotel$no_of_children), by = 1))+
  scale_fill_manual(name = "Тип кімнати",
                    values = room_type_vector,
                    labels = room_label_vector) + theme(plot.title = element_text(hjust = 0.5))
```

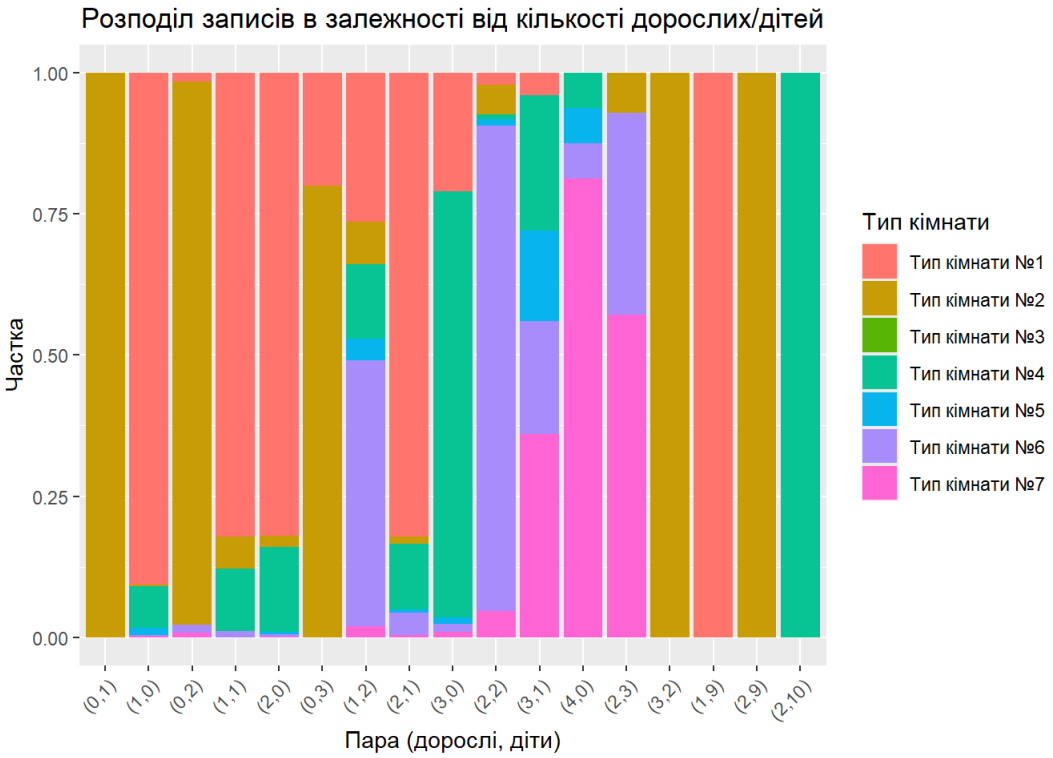


1-2 дорослих без дітей або з однією дитиною, частіше всього обирають перший тип кімнати, троє дорослих - четвертий, двоє дорослих і двоє дітей - шостий. Про інші випадки важко зробити висновок, через малу кількість записів.

```
ggplot(hotel_grouped_by_room, aes(x = reorder(pair, (no_of_adults + no_of_children)), y = total, fill = as.factor(room_type_reserved))) +
  geom_bar(stat = "identity") +
  labs(x = "Пара (дорослі, діти)", y = "Кількість записів", fill = "Тип кімнати") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_manual(name = "Тип кімнати",
                    values = room_type_vector,
                    labels = room_label_vector) + theme(plot.title = element_text(hjust = 0.5))
```

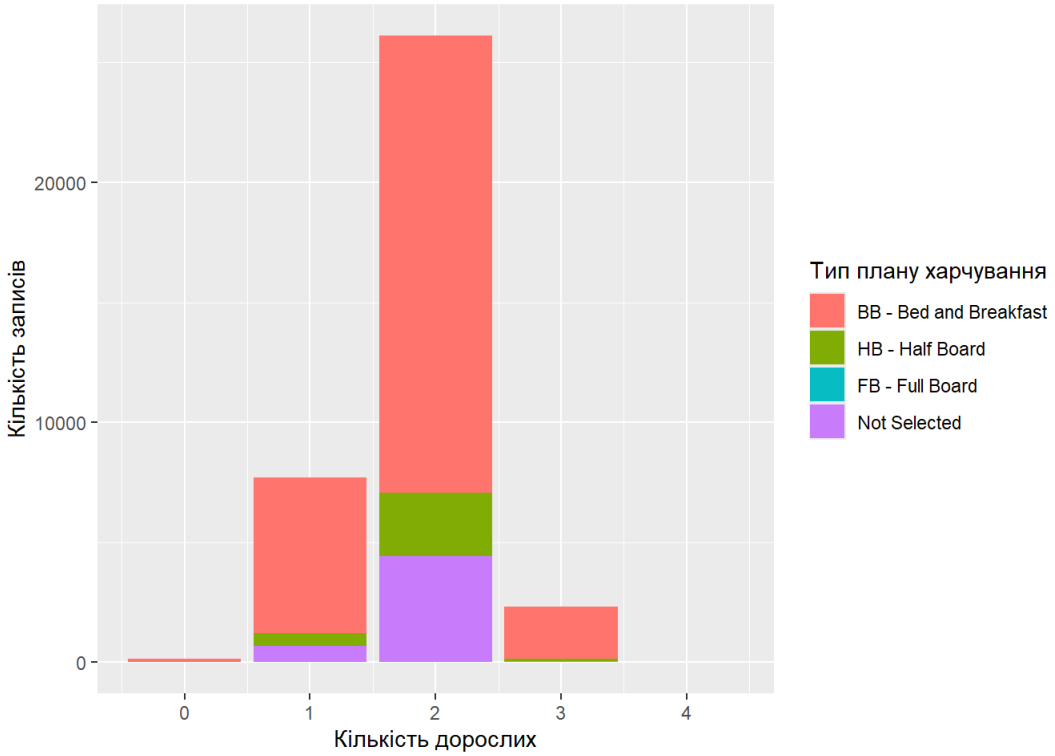



```
ggplot(hotel_grouped_by_room, aes(x = reorder(pair, (no_of_adults + no_of_children)), y = total, fill = as.factor(room_type_reserved))) +
  geom_bar(position = 'fill', stat = "identity") +
  labs(title = "Розподіл записів в залежності від кількості дорослих/дітей",x = "Пара (дорослі, діти)", y = "Частка", fill = "Тип кімнати") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_manual(name = "Тип кімнати",
                    values = room_type_vector,
                    labels = room_label_vector) + theme(plot.title = element_text(hjust = 0.5))
```

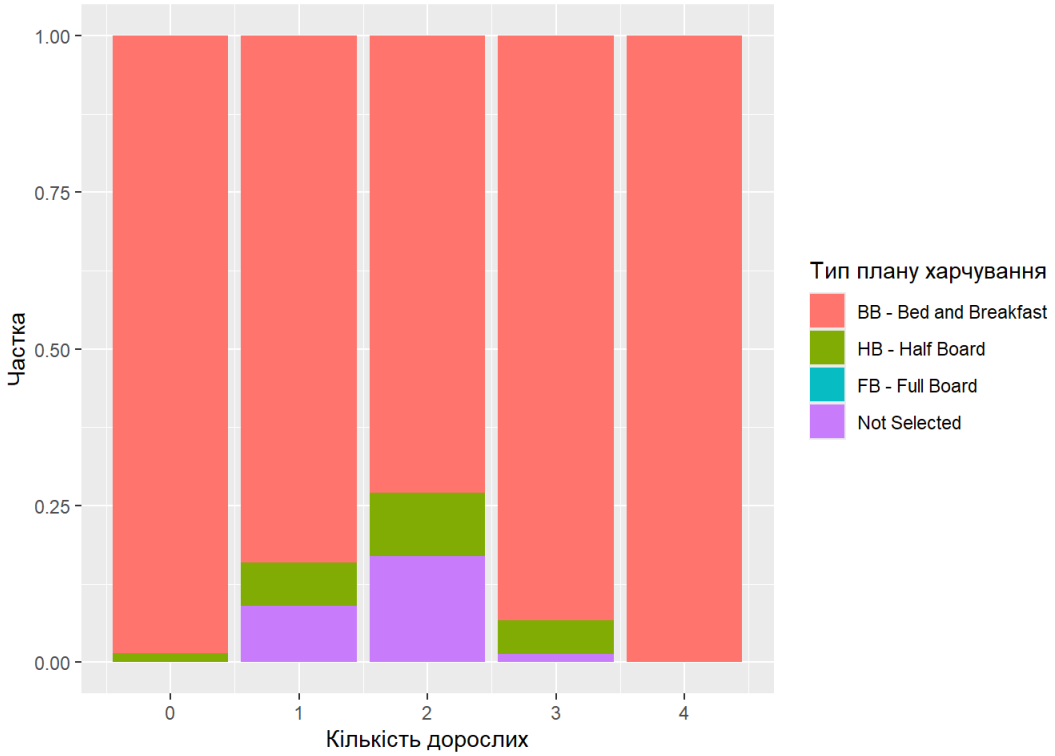


Бачимо, що незалежно від кількості дорослих, найбільшу перевагу віддають типу харчування BB, двоє дорослих частіше не обирають тип харчування, ніж один або троє.

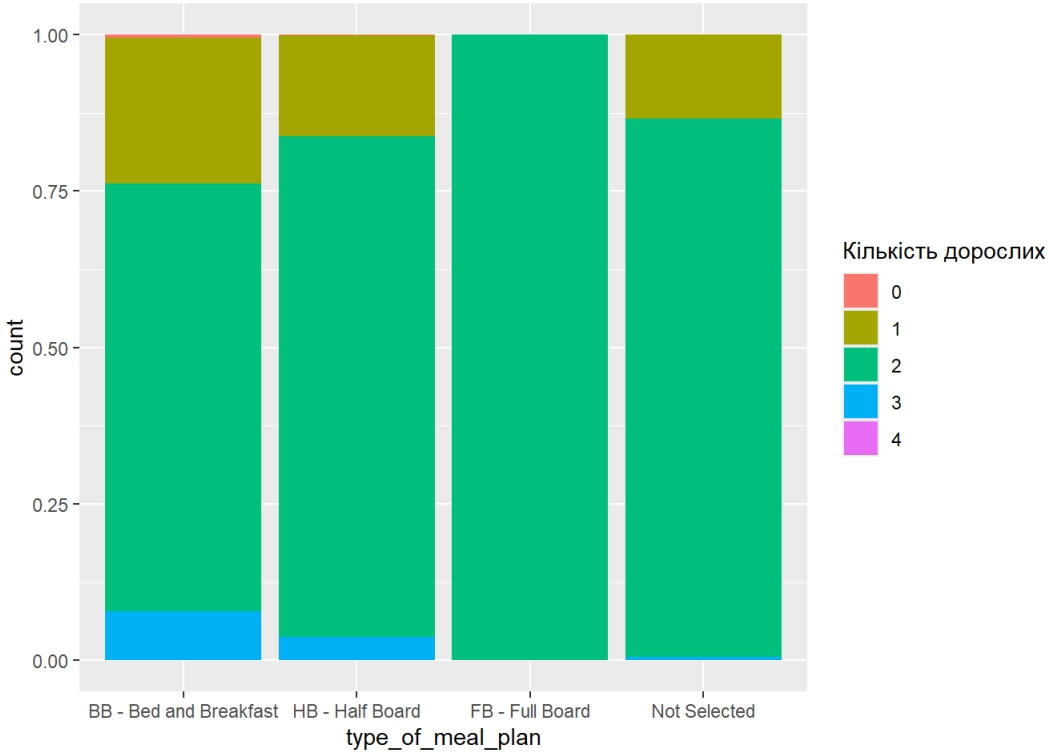
```
ggplot(hotel, aes(x = no_of_adults, fill = as.factor(type_of_meal_plan))) + geom_bar() +
  labs(x = "Кількість дорослих", y = "Кількість записів", fill = "Тип плану харчування") +
  scale_fill_manual(values = meal_type_vector,
    labels = meal_label_vector)
```




```
ggplot(hotel, aes(x = no_of_adults, fill = as.factor(type_of_meal_plan))) + geom_bar(position = "fill") +
  labs(x = "Кількість дорослих", y = "Частка", fill = "Тип плану харчування") +
  scale_fill_manual(values = meal_type_vector,
    labels = meal_label_vector)
```

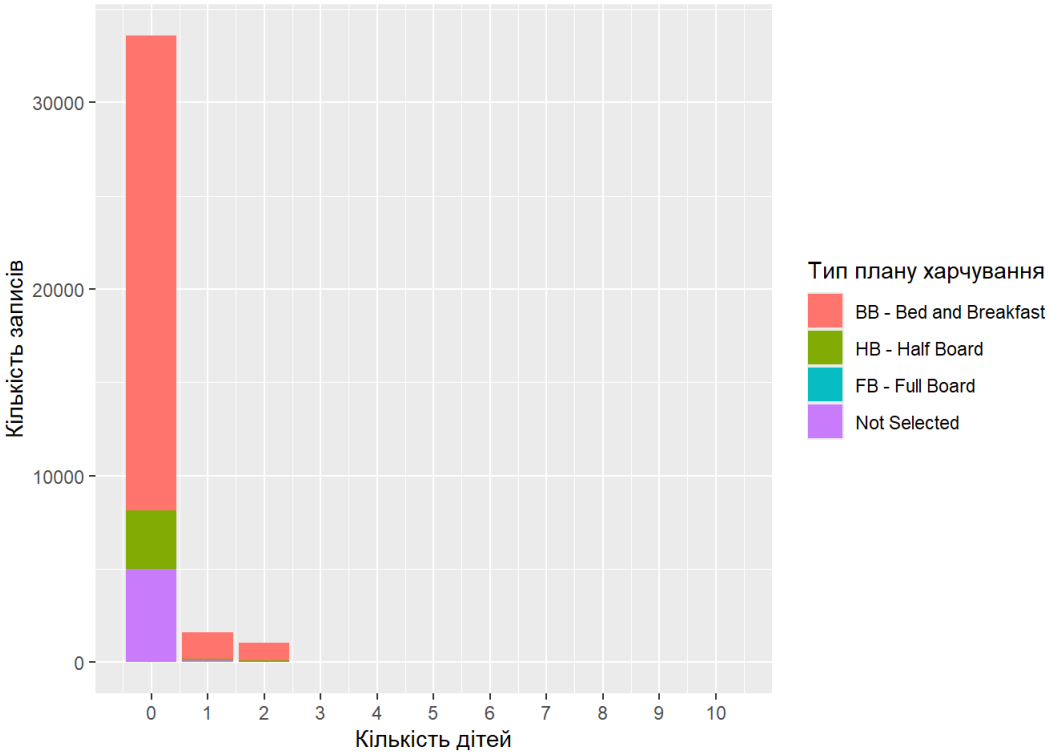



```
ggplot(hotel, aes(x = type_of_meal_plan, fill = as.factor(no_of_adults))) +
  geom_bar(position = "fill") +
  labs(fill = "Кількість дорослих") +
  scale_x_discrete(labels = meal_label_vector)
```

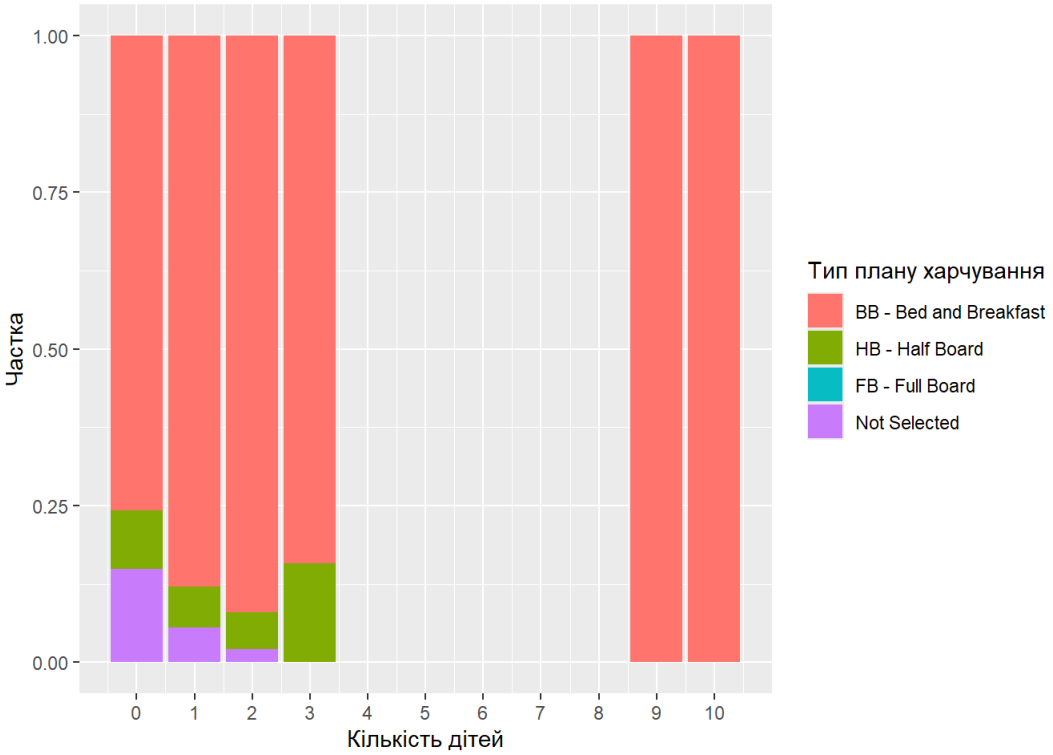


Видно, що зі збільшенням кількості дітей (від одного до трьох), зростає частка бронювань, де обрано тип харчування BB.

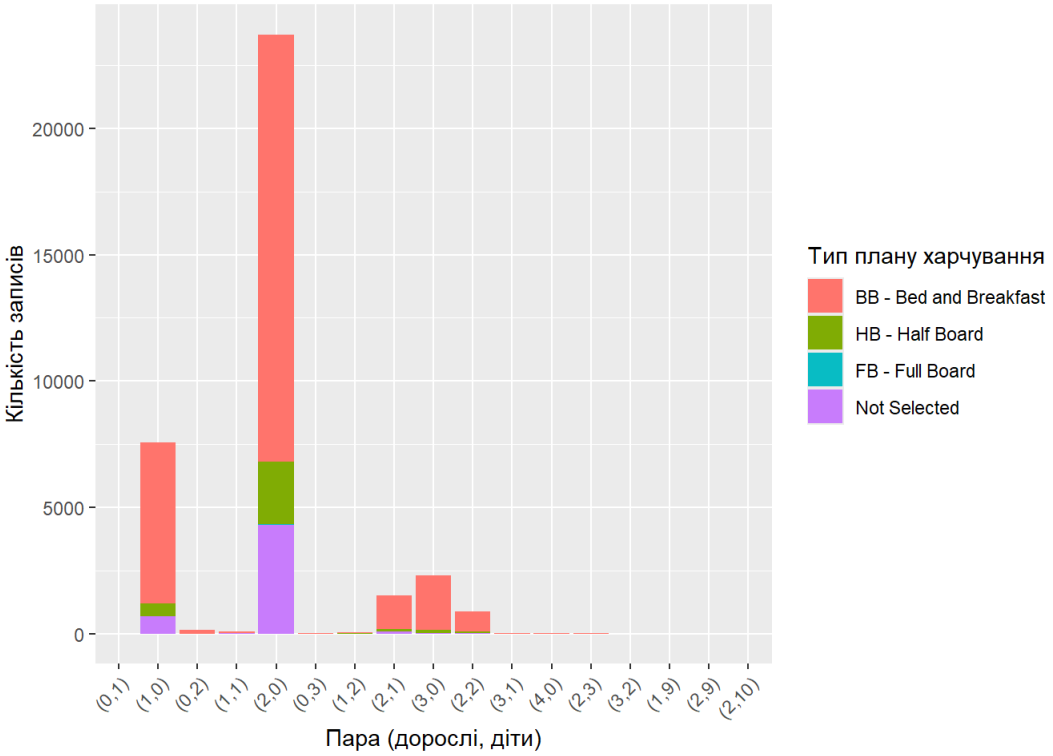
```
ggplot(hotel, aes(x = no_of_children, fill = as.factor(type_of_meal_plan))) + geom_bar() +
  labs(x = "Кількість дітей", y = "Кількість записів", fill = "Тип плану харчування") +
  scale_x_continuous(breaks = seq(0, max(hotel$no_of_children), by = 1)) +
  scale_fill_manual(values = meal_type_vector,
    labels = meal_label_vector)
```



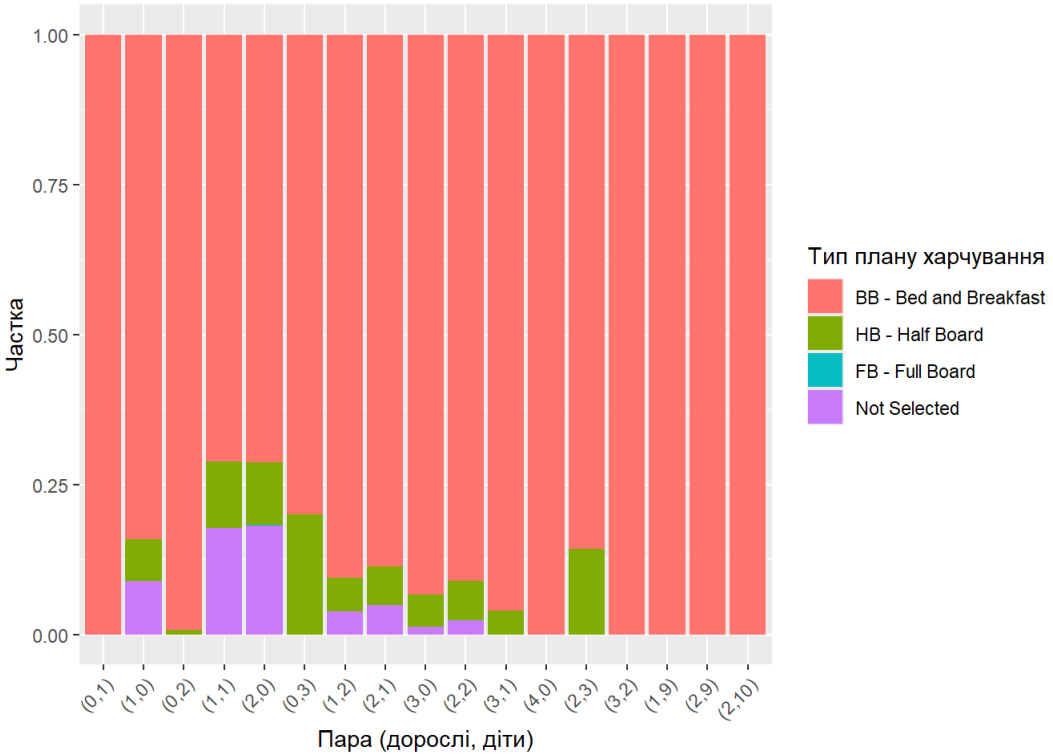

```
ggplot(hotel, aes(x = no_of_children, fill = as.factor(type_of_meal_plan))) + geom_bar(position = "fill") +
  labs(x = "Кількість дітей", y = "Частка", fill = "Тип плану харчування") +
  scale_x_continuous(breaks = seq(0, max(hotel$no_of_children), by = 1)) +
  scale_fill_manual(values = meal_type_vector,
    labels = meal_label_vector)
```




```
ggplot(hotel_grouped_by_meal, aes(x = reorder(pair, (no_of_adults + no_of_children)), y = total, fill = as.factor(type_of_meal_plan))) +
  geom_bar(stat = "identity") +
  labs(x = "Пара (дорослі, діти)", y = "Кількість записів", fill = "Тип плану харчування") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_manual(values = meal_type_vector,
    labels = meal_label_vector)
```




```
ggplot(hotel_grouped_by_meal, aes(x = reorder(pair, (no_of_adults + no_of_children)), y = total, fill = as.factor(type_of_meal_plan))) +
  geom_bar(position = 'fill', stat = "identity") +
  labs(x = "Пара (дорослі, діти)", y = "Частка", fill = "Тип плану харчування") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_manual(values = meal_type_vector,
    labels = meal_label_vector)
```



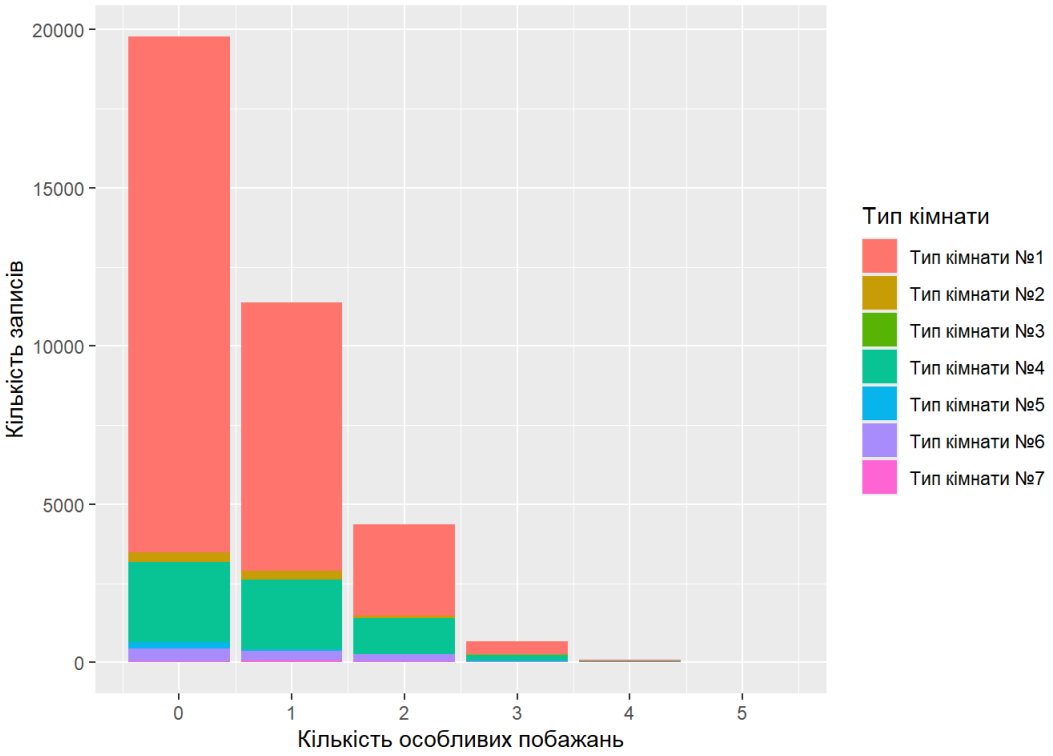
ДРУГЕ ЗАПИТАННЯ

Формулювання:

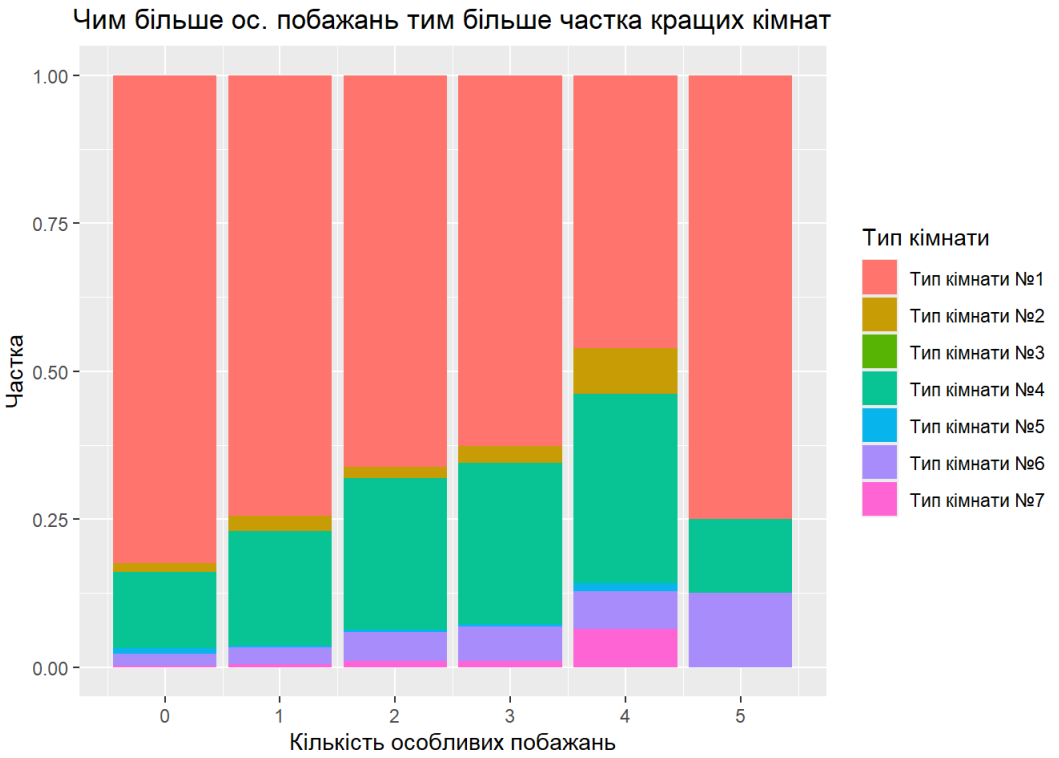
- Яким чином розподіляється кількість особливих побажань в залежності від типу номеру?

Побудуємо стовпчикову діаграму на основі даних про кількість особливих бажань, розфарбуємо в різні кольори відповідні типи кімнат. Якщо побудувати графік за кількістю записів, можна побачити, що для кожного числа особливих бажань приблизно однаковий розподіл типів кімнат. Якщо ж побудувати графік не по кількості особливих запитів, а по частці типів кімнат для кожної кількості особливих побажань (другий графік), спостерігаємо, що чим більше особливих бажань, тим більше частка кращих кімнат (кімнат 4,6 типу, частково 7 типу). Можливо, такий результат можна пов'язаний з тим, що у людей які мають гроші на кращі апартаменти, є гроші і на певні додаткові особливі побажання.

```
ggplot(hotel, aes(x = no_of_special_requests, fill = as.factor(room_type_reserved))) +
  geom_bar() + scale_x_continuous(breaks = seq(0, max(hotel$no_of_special_requests), by = 1)) +
  labs(x = "Кількість особливих побажань", y = "Кількість записів", fill = "Тип кімнати") +
  scale_fill_manual(name = "Тип кімнати",
                    values = room_type_vector,
                    labels = room_label_vector) + theme(plot.title = element_text(hjust = 0.5))
```



```
ggplot(hotel, aes(x = no_of_special_requests, fill = as.factor(room_type_reserved))) +
  geom_bar(position = "fill") + scale_x_continuous(breaks = seq(0, max(hotel$no_of_special_requests), by = 1)) +
  labs(title = "Чим більше ос. побажань тим більше частка кращих кімнат",x = "Кількість особливих побажань", y = "Частка", fill = "Тип кімнати") +
  scale_fill_manual(name = "Тип кімнати",
                    values = room_type_vector,
                    labels = room_label_vector) + theme(plot.title = element_text(hjust = 0.5))
```



ТРЕТЄ ЗАПИТАННЯ

Формулювання:

- Як впливає кількість дорослих і дітей на скасування бронювання?

Спочатку, переглянемо яка кількість яких типів сімей (кількість дорослих, кількість дітей) робить резервації у готель:

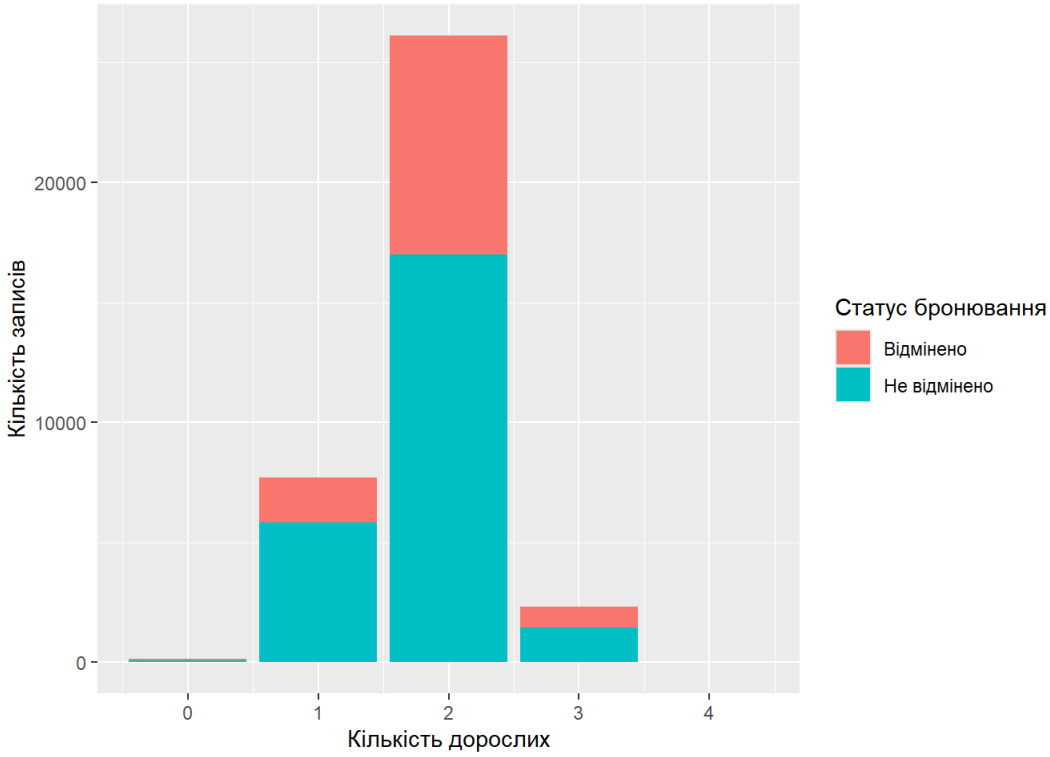
```
ftable(xtabs(~ no_of_adults + no_of_children + booking_status, data = hotel))
```

```
##                       booking_status Canceled Not_Canceled
## no_of_adults no_of_children
## 0            0                             0            0
##              1                             0            1
##              2                            44           89
##              3                             0            5
##              9                             0            0
##              10                            0            0
## 1            0                          1809         5742
##              1                            23           67
##              2                            24           29
##              3                             0            0
##              9                             0            1
##              10                            0            0
## 2            0                          8213        15506
##              1                           511          991
##              2                           389          482
##              3                             5            9
##              9                             1            0
##              10                            0            1
## 3            0                           857         1434
##              1                             6           19
##              2                             0            1
##              3                             0            0
##              9                             0            0
##              10                            0            0
## 4            0                             3           13
##              1                             0            0
##              2                             0            0
##              3                             0            0
##              9                             0            0
##              10                            0            0
```

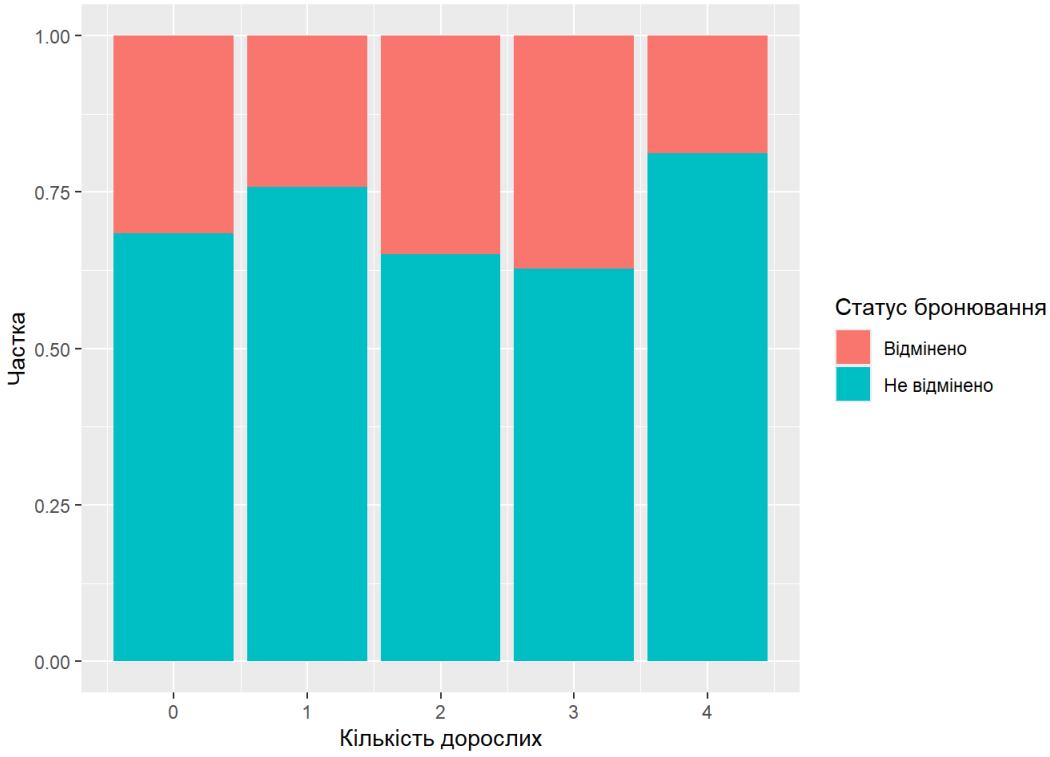
Нескладно помітити, що найпопулярнішими “типами” сімей є (1 дорослий без дітей), (2 дорослих без дітей), (2 дорослих з 1-2 дітьми) і (3 дорослих без дітей)

Побудуємо стовпчикові діаграми: кількість записів і кількість дорослих, позначивши кольорами статус бронювання записів, і таку ж, але частки замість кількості записів.

```
ggplot(hotel, aes(x = no_of_adults, fill = as.factor(booking_status))) + geom_bar() +
  labs(x = "Кількість дорослих", y = "Кількість записів", fill = "Статус бронювання") +
  scale_fill_manual(name = "Статус бронювання",
                    values = c("Not_Canceled" = "#00bfc4", "Canceled" = "#f8766d"),
                    labels = c("Відмінено", "Не відмінено"))
```



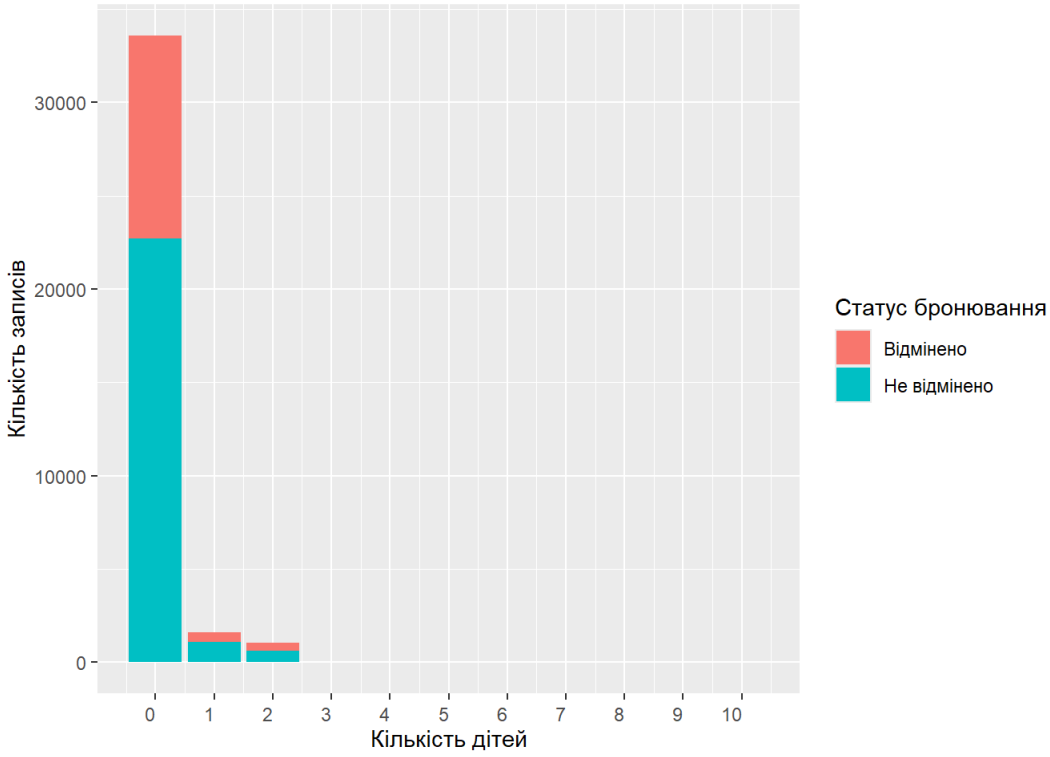

```
ggplot(hotel, aes(x = no_of_adults, fill = as.factor(booking_status))) + geom_bar(position = "fill") +
  labs(x = "Кількість дорослих", y = "Частка", fill = "Статус бронювання") +
  scale_fill_manual(name = "Статус бронювання",
                    values = c("Not_Canceled" = "#00bfc4", "Canceled" = "#f8766d"),
                    labels = c("Відмінено", "Не відмінено"))
```



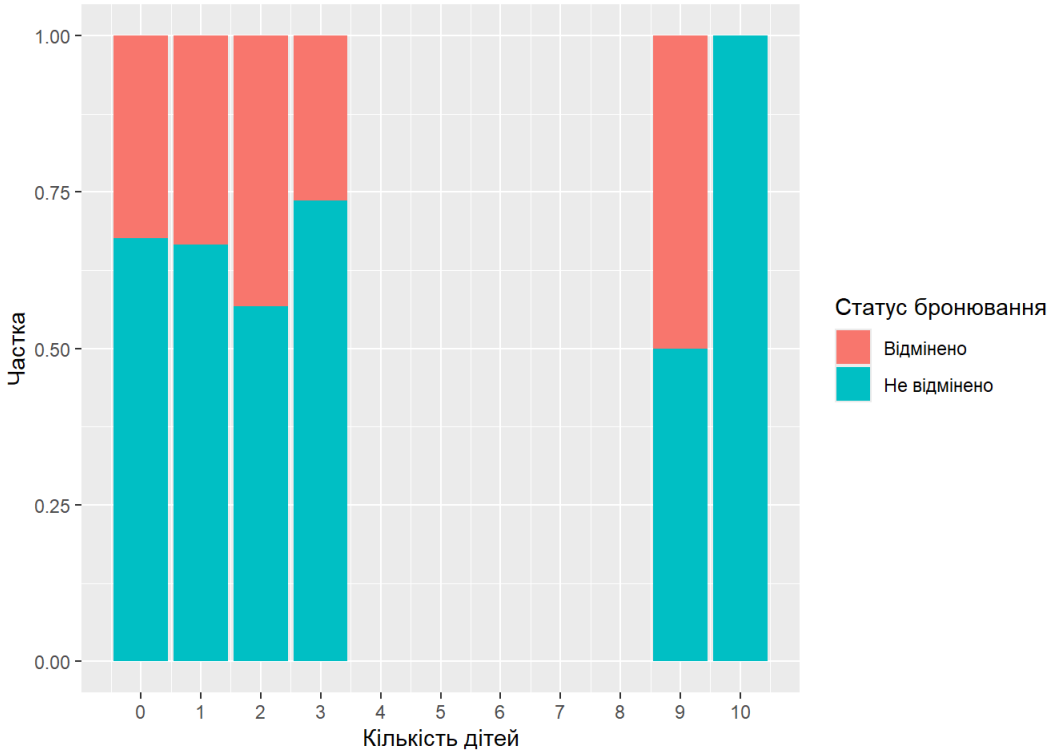
З другого графіку можемо побачити, що пропорційно статус бронювання практично не змінюється від кількості дорослих, тобто хоч дорослих 0 , хоч їх 4, відмінені будуть приблизно 30% записів.

Побудуємо аналогічні стовпчикові діаграми для кількості дітей:

```
ggplot(hotel, aes(x = no_of_children, fill = as.factor(booking_status))) + geom_bar() +
  labs(x = "Кількість дітей", y = "Кількість записів", fill = "Статус бронювання") +
  scale_x_continuous(breaks = seq(0, max(hotel$no_of_children), by = 1)) +
  theme(axis.text.x = element_text(angle = 0, vjust = 0.5, hjust = 1)) +
  scale_fill_manual(name = "Статус бронювання",
                    values = c("Not_Canceled" = "#00bfc4", "Canceled" = "#f8766d"),
                    labels = c("Відмінено", "Не відмінено"))
```




```
ggplot(hotel, aes(x = no_of_children, fill = as.factor(booking_status))) + geom_bar(position = "fill") +
  scale_x_continuous(breaks = seq(0, max(hotel$no_of_children), by = 1)) +
  labs(x = "Кількість дітей", y = "Частка", fill = "Статус бронювання") +
  scale_fill_manual(name = "Статус бронювання",
                    values = c("Not_Canceled" = "#00bfc4", "Canceled" = "#f8766d"),
                    labels = c("Відмінено", "Не відмінено"))
```



Можна побачити, що з дітьми приблизно така ж ситуація, якщо дітей до 3 включно. Тобто, якщо дітей до трьох включно, то статус бронювання пропорційно однаково розподілений - приблизно 35% записів буде відмінено. Для випадків коли дітей 9 чи 10 ситуація трохи відрізняється, але це дуже поодинокі випадки:

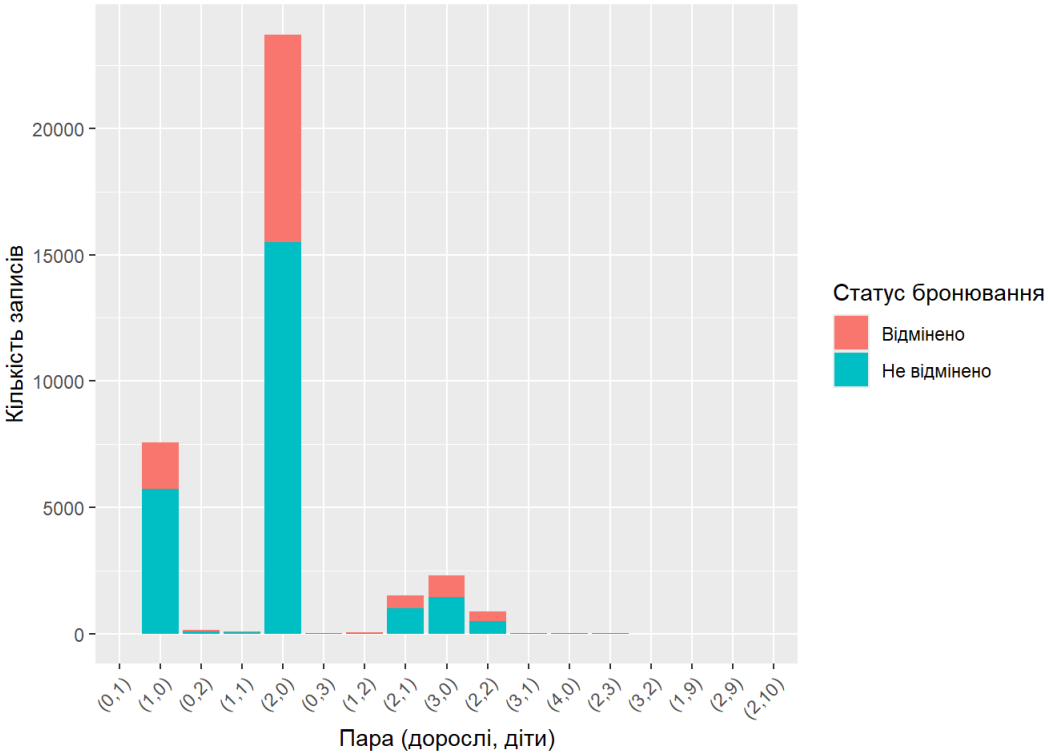
```
hotel %>% group_by(no_of_children) %>%
  summarise(total = n()) %>%
  filter(no_of_children >= 9)
```

no_of_children <int>
9
10

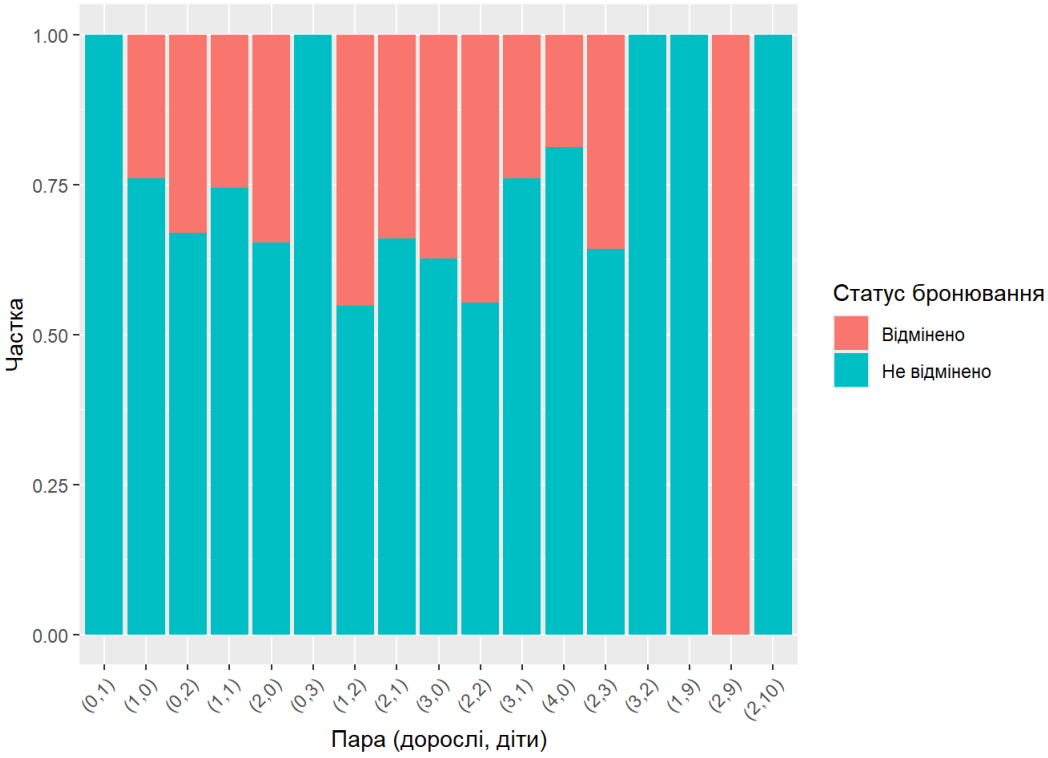
2 rows | 1-1 of 2 columns

Також зробимо аналогічні графіки для кількості “типів сімей” (кількість дорослих - кількість дітей), відзначивши кольором статус бронювання записів

```
ggplot(hotel_grouped_by_bs, aes(x = reorder(pair, (no_of_adults + no_of_children)), y = total, fill = booking_status)) +
  geom_bar(stat = "identity") +
  labs(x = "Пара (дорослі, діти)", y = "Кількість записів", fill = "Статус бронювання") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_manual(name = "Статус бронювання",
                    values = c("Not_Canceled" = "#00bfc4", "Canceled" = "#f8766d"),
                    labels = c("Відмінено", "Не відмінено"))
```




```
ggplot(hotel_grouped_by_bs, aes(x = reorder(pair, (no_of_adults + no_of_children)), y = total, fill = booking_status)) +
  geom_bar(stat = "identity", position = "fill") +
  labs(x = "Пара (дорослі, діти)", y = "Частка", fill = "Статус бронювання") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_manual(name = "Статус бронювання",
                    values = c("Not_Canceled" = "#00bfc4", "Canceled" = "#f8766d"),
                    labels = c("Відмінено", "Не відмінено"))
```



Спостерігаємо схожу ситуацію: для найпопулярніших “типів” сімей приблизно 25-35% записів скасовані. У поодиноких випадках ситуація не така стабільна.

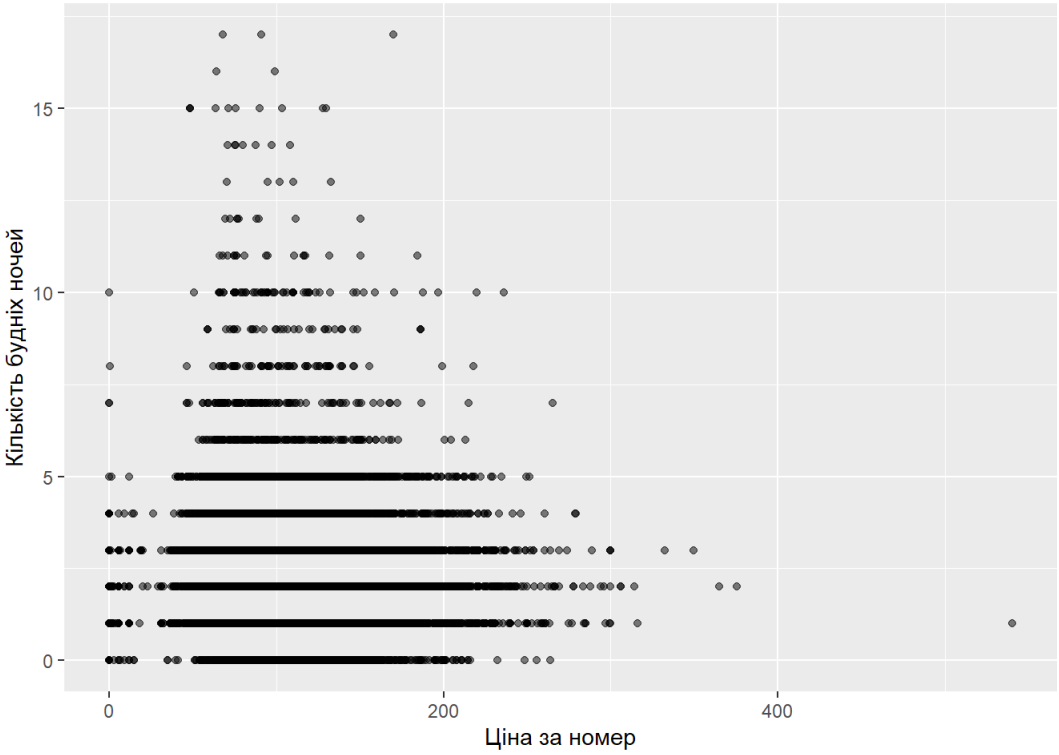
ЧЕТВЕРТЕ ЗАПИТАННЯ

Формулювання:

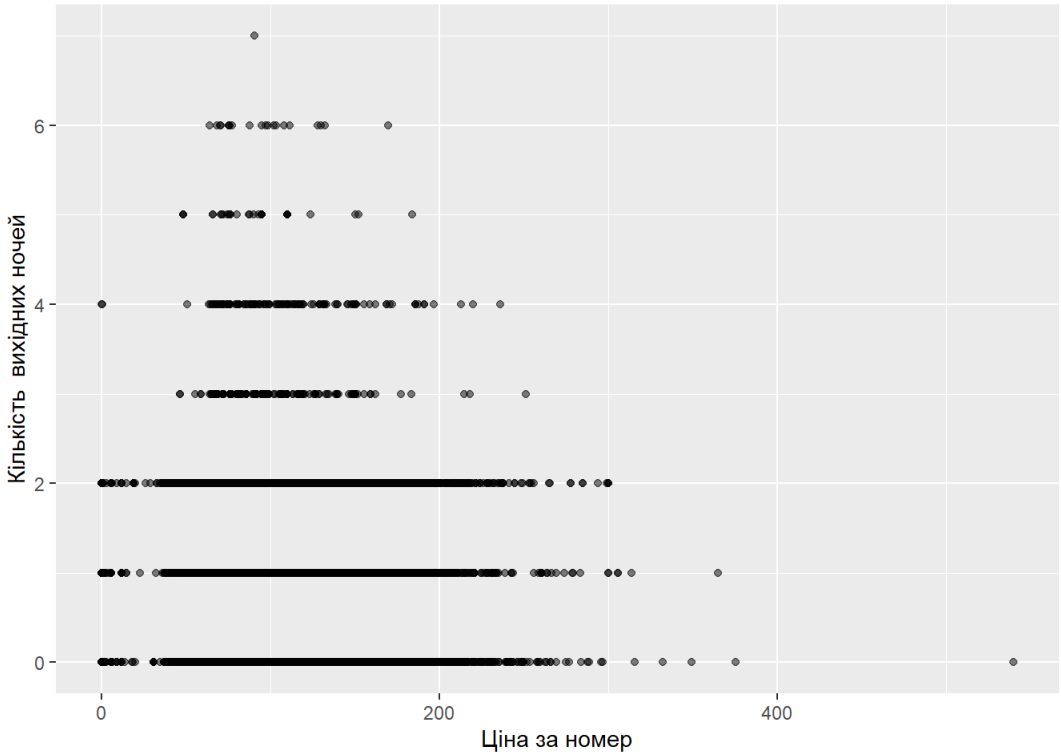
- Що впливає на кількість проведених вихідних та робочих ночей у готелі?

Перше, що спадає на думку перевірити, це чи не є у вихідні середня ціна більшою, ніж у будні:

```
ggplot(hotel, aes(x = avg_price_per_room, y = no_of_week_nights) ) + geom_point(alpha = 0.5) +
  labs(x = "Ціна за номер",
       y = "Кількість будніх ночей") + theme(plot.title = element_text(hjust = 0.5))
```



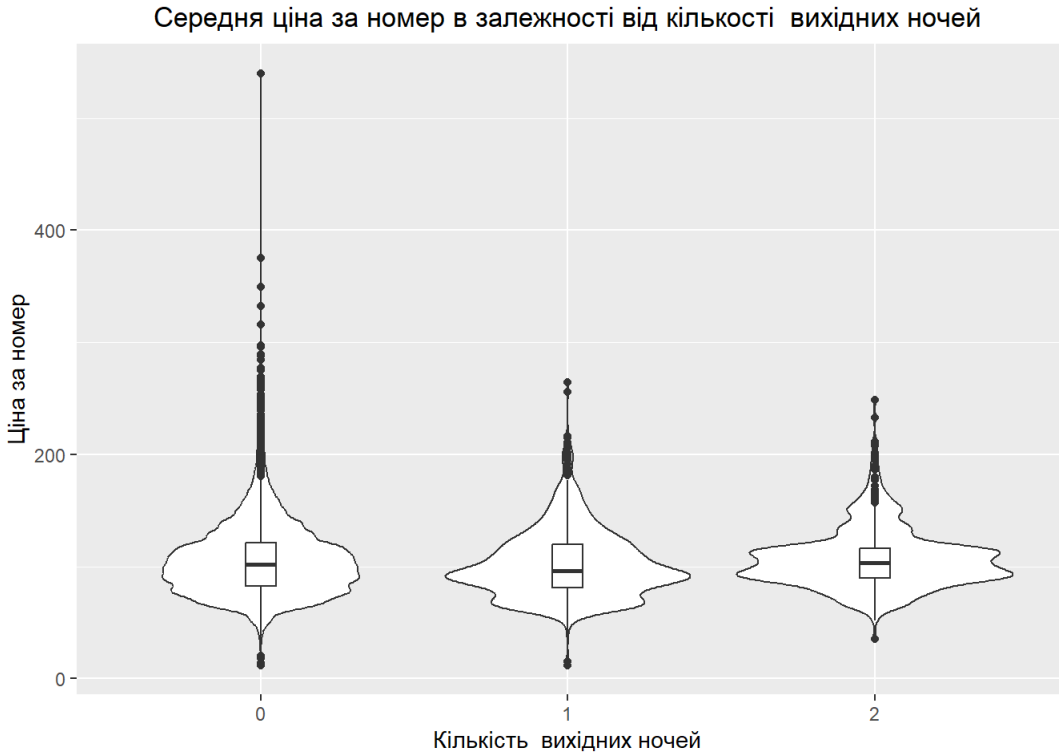

```
ggplot(hotel, aes(x = avg_price_per_room, y = no_of_weekend_nights) ) + geom_point(alpha = 0.5) +
  labs(x = "Ціна за номер",
       y = "Кількість  вихідних ночей") + theme(plot.title = element_text(hjust = 0.5))
```



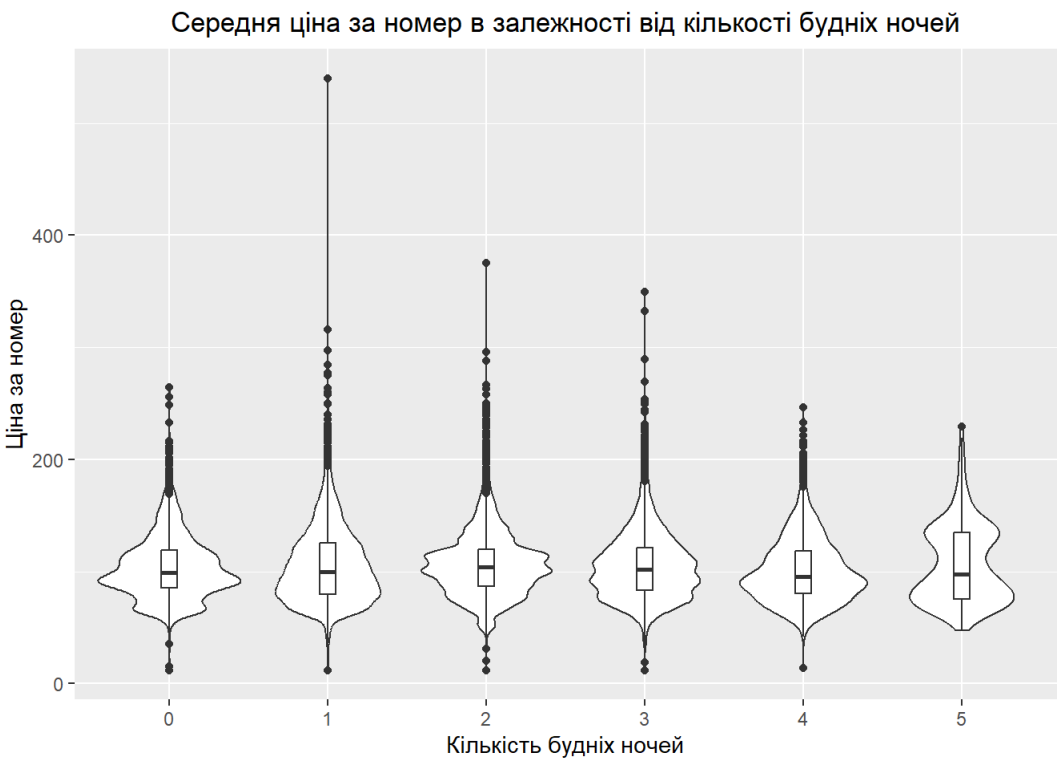
Як бачимо, графіки, грубо кажучи, є однаковими: графік для кількості будніх ночей є “плотнішим” (насиченішим), в силу того, що кількість будніх більша за кількість вихідних.

Для більш наглядної демонстрації цього висновку збудуємо відповідні вусаті скриньки:

```
ggplot(hotel_zero_nights, aes(x = as.factor(no_of_weekend_nights), y = avg_price_per_room)) + geom_violin() + stat_summary(fun = mean, geom="point"
, shape=23, size=2) + geom_boxplot(width=0.1) +
  labs(title = "Середня ціна за номер в залежності від кількості  вихідних ночей",
       x = "Кількість  вихідних ночей",
       y = "Ціна за номер") + theme(plot.title = element_text(hjust = 0.5))
```



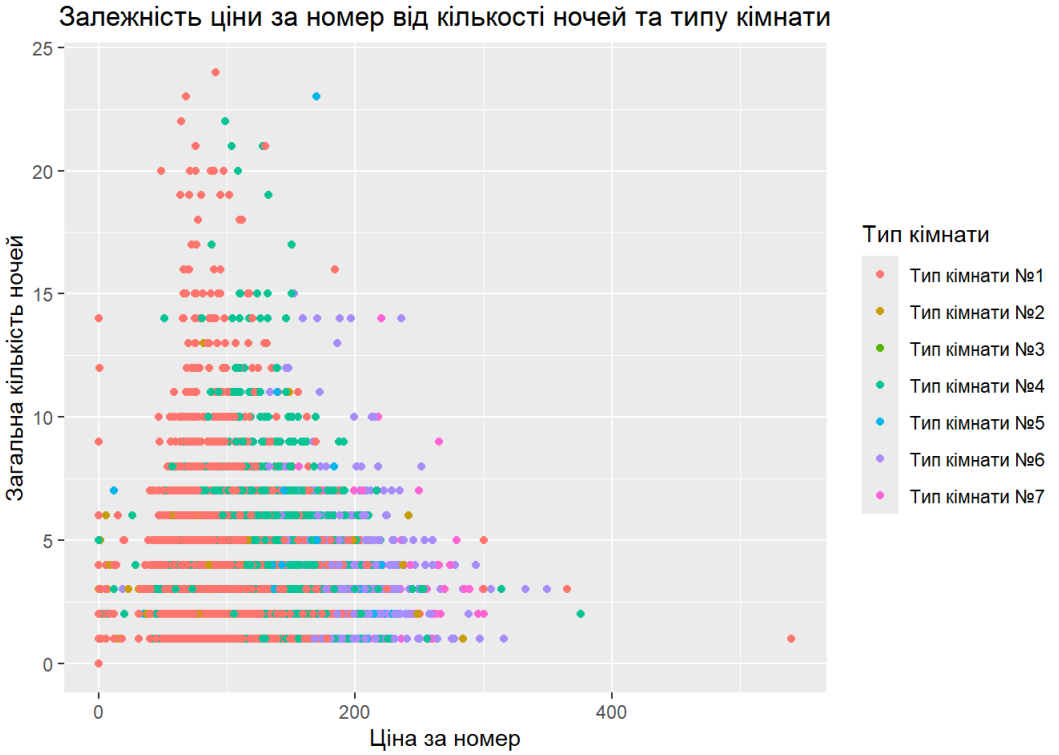

```
ggplot(hotel_zero_nights, aes(x = as.factor(no_of_week_nights), y = avg_price_per_room)) + geom_violin() + stat_summary(fun = mean, geom="point", sh
ape=23, size=2) + geom_boxplot(width=0.1) +
labs(title = "Середня ціна за номер в залежності від кількості будніх ночей",
     x = "Кількість будніх ночей",
     y = "Ціна за номер") + theme(plot.title = element_text(hjust = 0.5))
```



Можемо побачити, що за будь-якої кількості будніх/вихідних зарезервованих ночей, ціна є приблизно однаковою і не зазнає змін

Подивимось, чи залежить явним чином кількість заброньованих ночей (будніх+вихідних) від типу кімнати:

```
ggplot(hotel_with_nights, aes(x = avg_price_per_room, y = no_of_nights, color = as.factor(room_type_reserved))) +
  geom_point() +
  labs(title = "Залежність ціни за номер від кількості ночей та типу кімнати",
       x = "Ціна за номер",
       y = "Загальна кількість ночей",
       color = "Тип кімнати")+
  scale_color_manual(name = "Тип кімнати",
                     values = room_type_vector,
                     labels = room_label_vector) + theme(plot.title = element_text(hjust = 0.5))
```

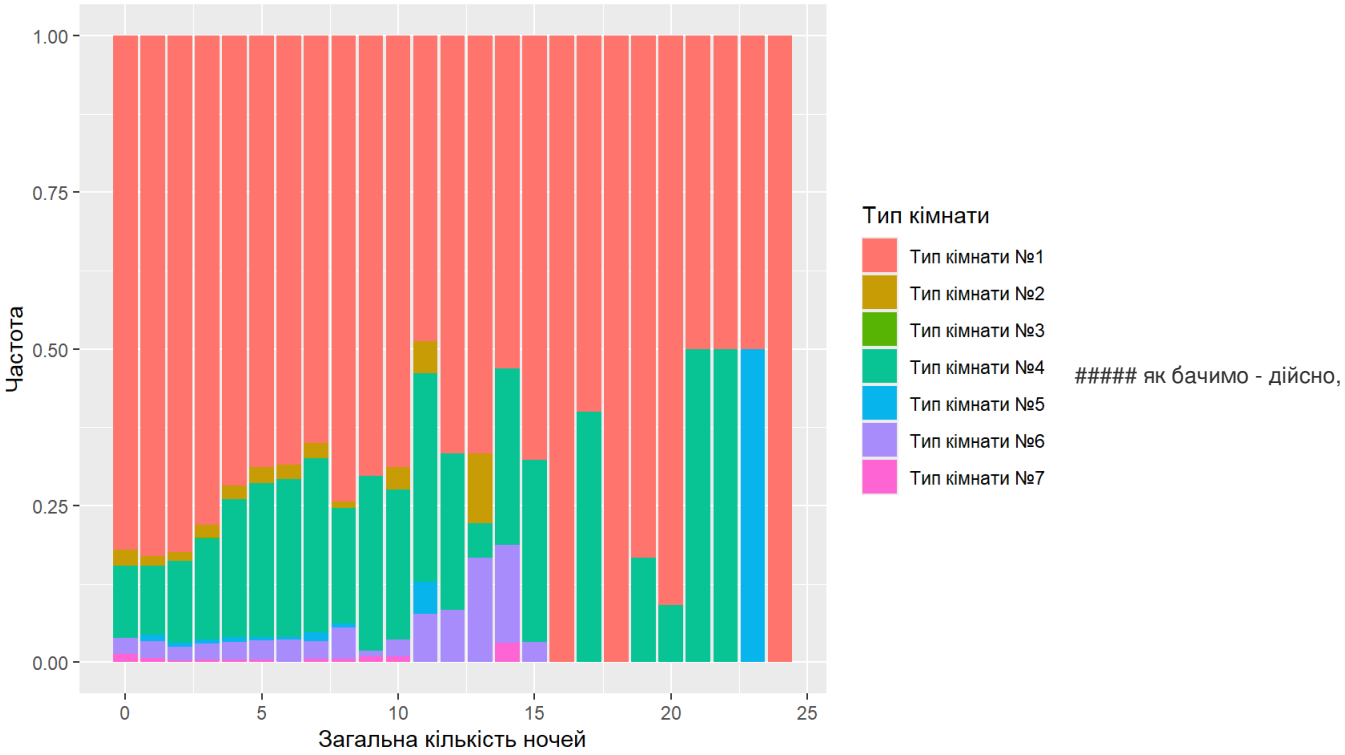


Як бачимо, можна відмітити, що

- від 15 до 25 днів, здебільшого резервування припадають на 1 та 4 типи кімнат.
- тип кімнати 6 обирають не більше, ніж на 2 тижні
- 7-ий тип кімнати обирають здебільшого на термін перебування ~тиждень

Розглянемо у більш зрозуміло вигляді яка частка яких типів кімнат припадає на кожне число загальної кількості заброньованих ночей:

```
ggplot(hotel_with_nights, aes(x = no_of_nights, fill = as.factor(room_type_reserved))) + geom_bar(position = "fill") +
  labs(
      x = "Загальна кількість ночей",
      y = "Частота",) +
  scale_fill_manual(name = "Тип кімнати",
                    values = room_type_vector,
                    labels = room_label_vector) + theme(plot.title = element_text(hjust = 0.5))
```



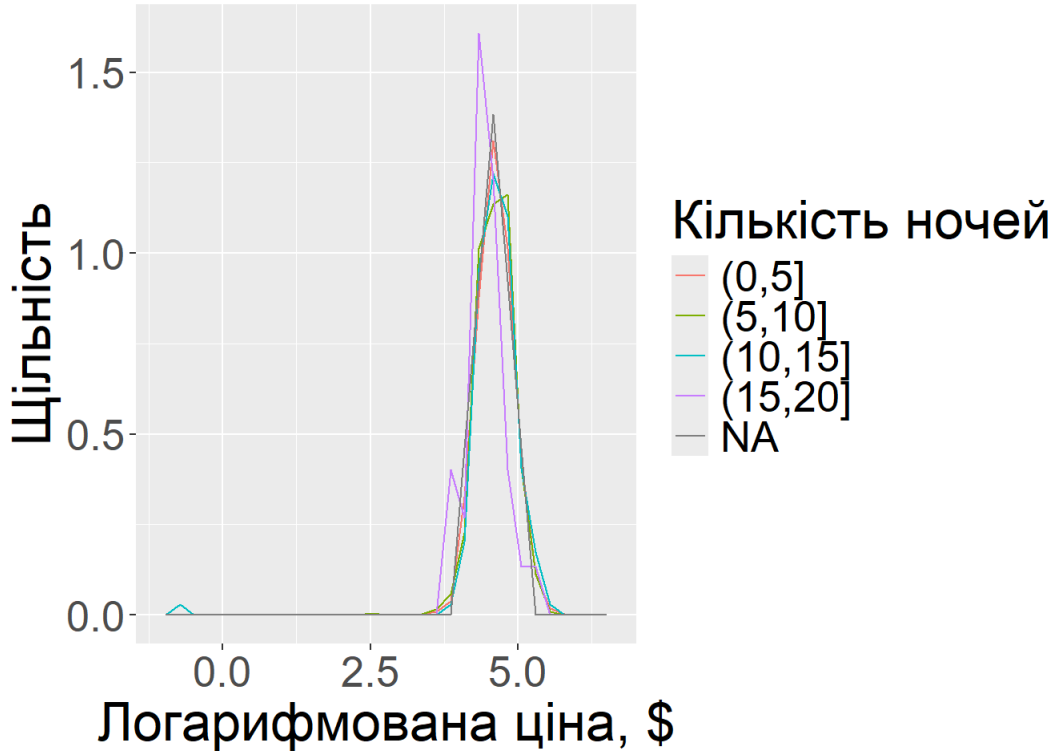
як бачимо - дійсно, попередні висновки підтвердилися

Цікаво було б дізнатися яким саме чином розподіляється середня ціна для різних діапазонів кількості забронькованих ночей

Відклавши на вісі іксів прологарифмовану ціну, а на вісі ігриків - щільність, та позначивши діапазони кількості ночей відповідними кольорами, можемо спостерігати досить цікавий результат:

```
ggplot(hotel_with_nights, aes(x = log(avg_price_per_room), y = after_stat(density),
                              color = cut(no_of_nights, breaks = seq(min(no_of_nights), max(no_of_nights), by = 5)))) +     geom_freqpoly(bins = 30) +
   labs(x = "Логарифмована ціна, $", y = "Щільність", color = "Кількість ночей") +
    theme(axis.title = element_text(size = 25),
          axis.text = element_text(size = 20),
          legend.title = element_text(size = 25),
          legend.text = element_text(size = 20)) + theme(plot.title = element_text(hjust = 0.5))
```

```
## Warning: Removed 545 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

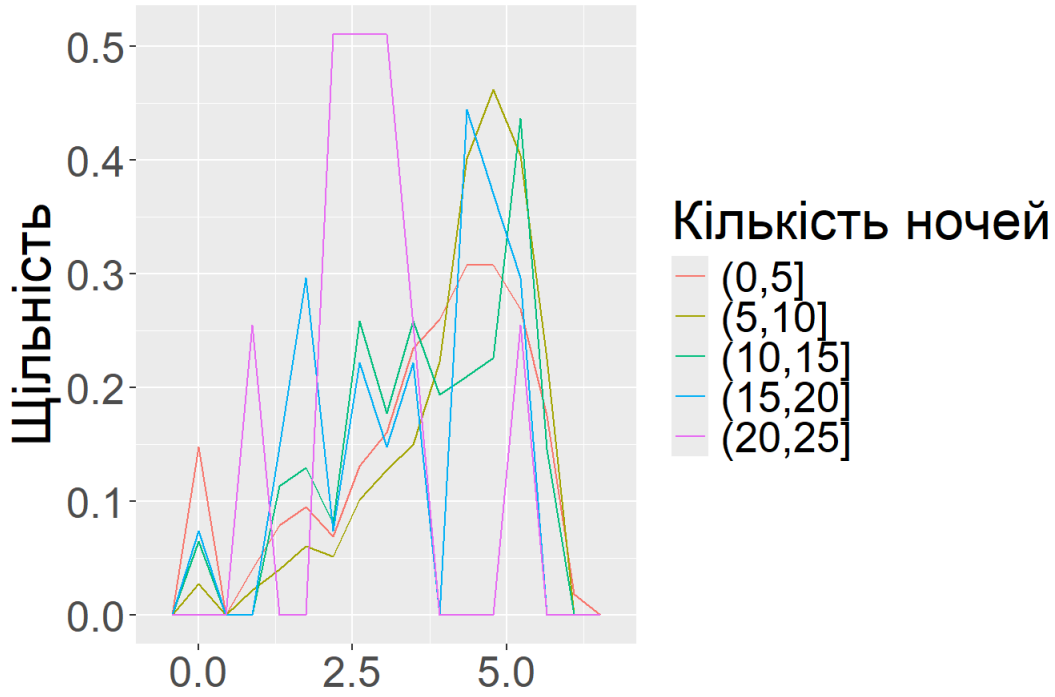


наведений графік дуже

нагадує нормальний розподіл, тож, можна припустити, що оригінальне значення ціни матиме логнормальний розподіл, але це припущення потребує подальшої перевірки

якщо побудувати аналогічний графік, взявши, наприклад, на вісі іксів прологарифмований час до прибуття, то нічого визначного побачити не вдасться:

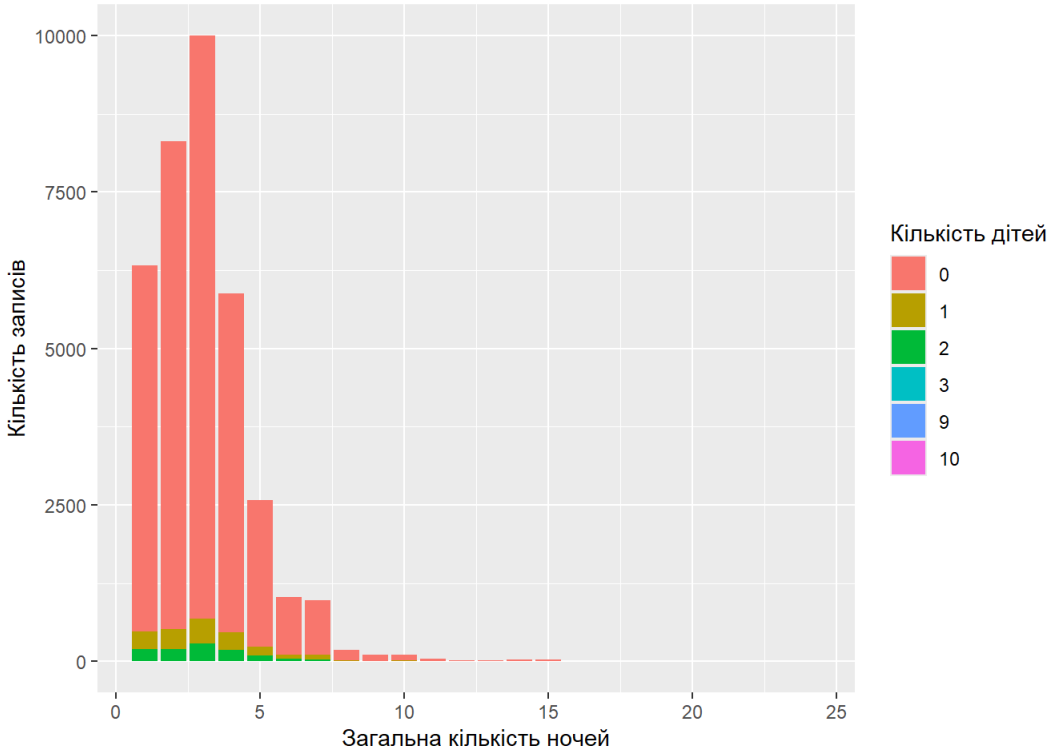
```
ggplot(hotel_lead_check, aes(x = log(lead_time), y = after_stat(density),
                             color = cut(no_of_nights, breaks = c(0, 5, 10, 15, 20, max(no_of_nights) + 1)
                                         ))) + geom_freqpoly(bins = 15) +
  labs(x = "Логарифм. час до прибуття, год", y = "Щільність", color = "Кількість ночей") +
     theme(axis.title = element_text(size = 25),
           axis.text = element_text(size = 20),
           legend.title = element_text(size = 25),
           legend.text = element_text(size = 20)) + theme(plot.title = element_text(hjust = 0.5))
```



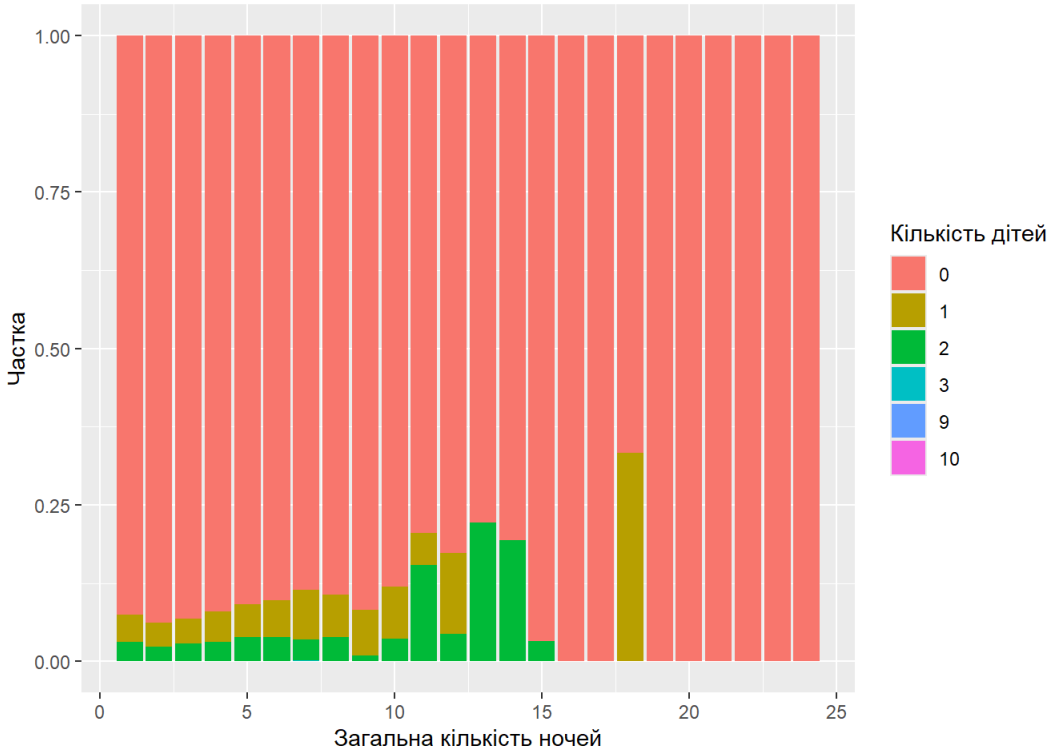
Можливо, вплив на кількість заброньованих ночей може мати кількість дітей: це може бути сімейна відпустка, яка триватиме, умовно, тиждень або трохи більше і т.п.

Побудуємо стовпчикові діаграми для кількості заброньованих ночей, відклавши на вісі ігриків кількості записів, що припадають на задану кількість заброньованих ночей, а кольорами позначимо записи з відповідною кількістю дітей:

```
ggplot(hotel_lead_check, aes(x = no_of_nights, fill = as.factor(no_of_children))) + geom_bar() +
  labs(x = "Загальна кількість ночей",
       y = "Кількість записів",
       fill = "Кількість дітей") + theme(plot.title = element_text(hjust = 0.5))
```




```
ggplot(hotel_lead_check, aes(x = no_of_nights, fill = as.factor(no_of_children))) + geom_bar(position = "fill") +
  labs(x = "Загальна кількість ночей",
       y = "Частка",
       fill = "Кількість дітей") + theme(plot.title = element_text(hjust = 0.5))
```



як бачимо, досить часто діти прибувають на (одно/дво)тижневі “відпустки” ##### також варто помітити кількість записів, що припадає на перші ~7 днів: серед великої сукупності записів дуже багато прибувших є виключно дорослими без дітей

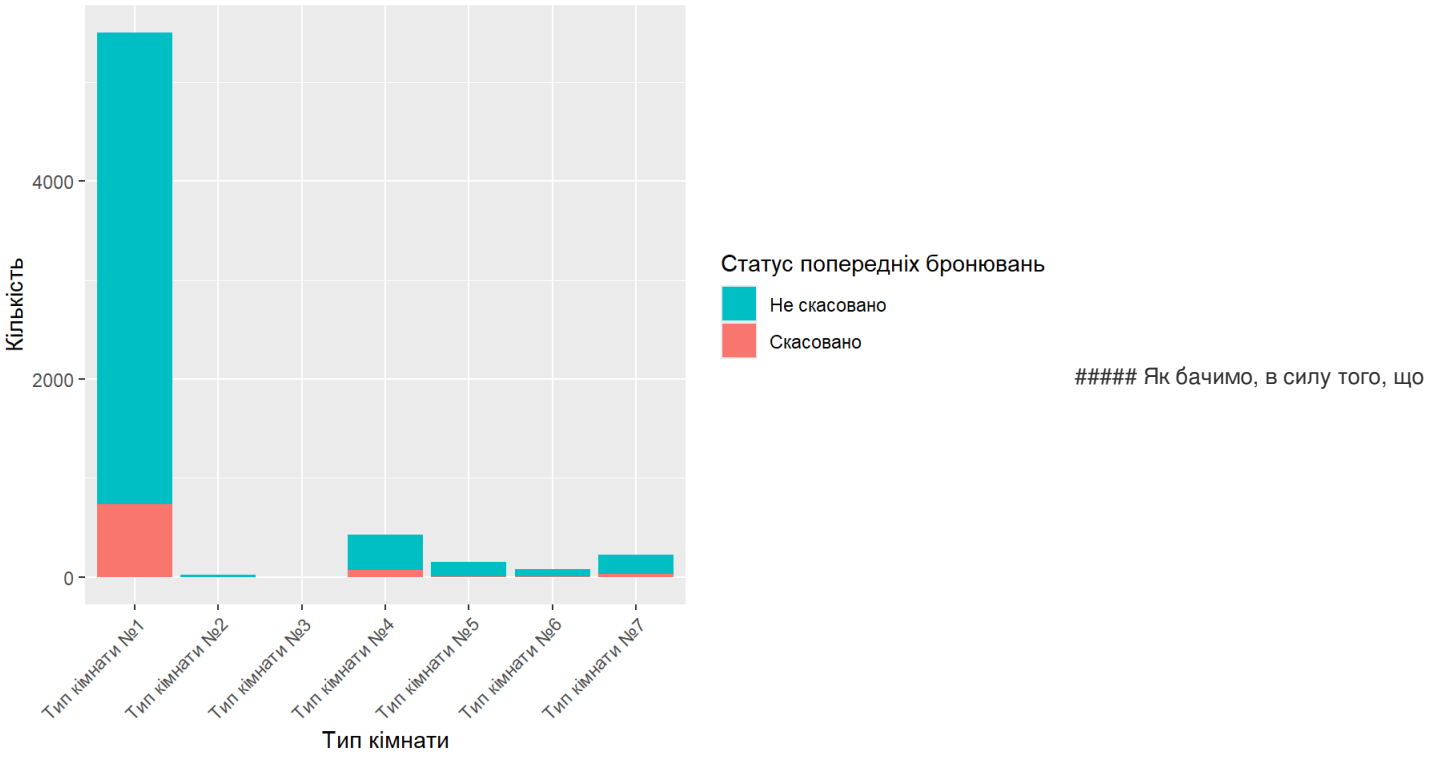
П’ЯТЕ ЗАПИТАННЯ

Формулювання:

- Як різні типи кімнат впливають на кількість попередніх скасувань/бронювань?

Побудуємо стовпчикову діаграму: на вісі іксів позначимо типи зарезервованих кімнат від 1-ої до 7-ої, а на вісі ігриків відкладатимемо значення суми (кількість попередніх скасувань + кількість попередніх НЕскасувань), відзначивши відповідні кількості синім і червоним кольорами:

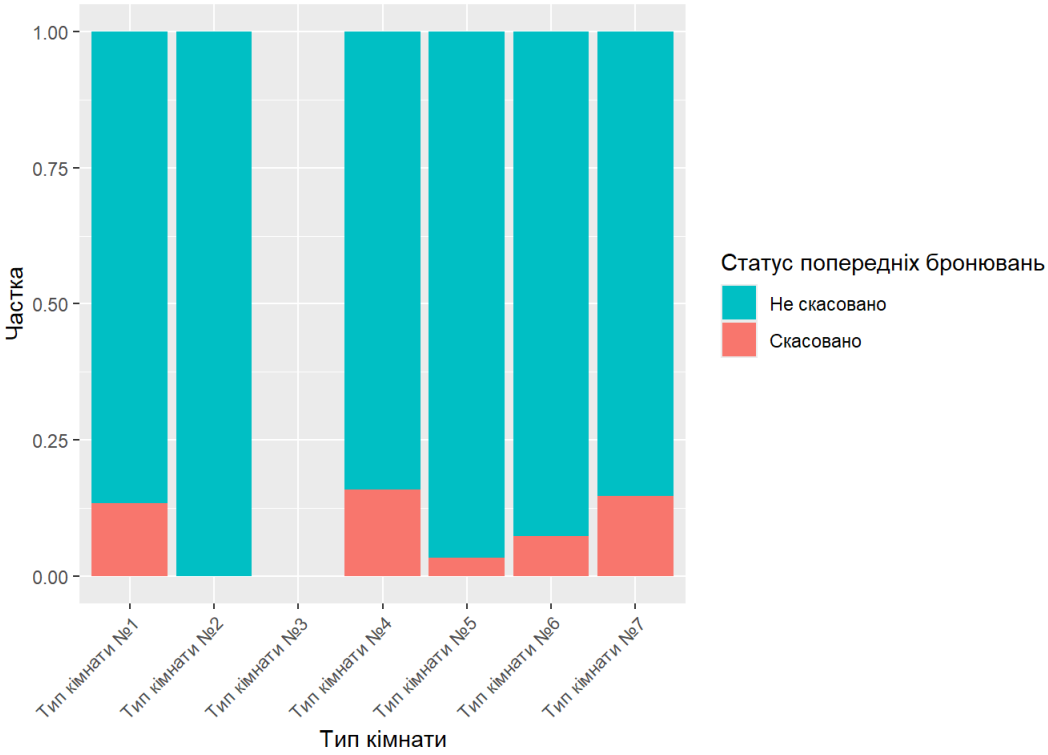
```
hotel_long <- pivot_longer(hotel, cols = c(no_of_previous_cancellations, no_of_previous_bookings_not_canceled), names_to = "status", values_to = "coun
t")

ggplot(hotel_long, aes(x=room_type_reserved, fill=status, y=count))+
  geom_bar(stat = "identity")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+
  labs(
      x = "Тип кімнати",
      y = "Кількість") +
  scale_fill_manual(name = "Статус попередніх бронювань",
                    values = c("no_of_previous_bookings_not_canceled" = "#00bfc4", "no_of_previous_cancellations" = "#f8766d"),
                    labels = c("Не скасовано", "Скасовано"))+
  scale_x_discrete(labels = room_label_vector) + theme(plot.title = element_text(hjust = 0.5))
```



Як бачимо, в силу того, що певних типів кімнат (здебільшого Room_Type 1) резервувалося значно більше, ніж інших, то у такому вигляді графік не є особливо інформативним, тому нормуємо значення ігриків таким чином, щоб для кожного типу кімнати стовпець відображав частки кількостей попередніх скасувань і НЕскасувань:

```
ggplot(hotel_long, aes(x=room_type_reserved, fill=status, y=count))+
  geom_bar(stat = "identity", position = "fill")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+
  labs(x = "Тип кімнати",
      y = "Частка") +
  scale_fill_manual(name = "Статус попередніх бронювань",
                    values = c("no_of_previous_bookings_not_canceled" = "#00bfc4", "no_of_previous_cancellations" = "#f8766d"),
                    labels = c("Не скасовано", "Скасовано"))+
  scale_x_discrete(labels = room_label_vector) + theme(plot.title = element_text(hjust = 0.5))
```

```
## Warning: Removed 14 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```



Як можна бачити з наведених двох графіків вище, типи кімнат, які бронювали частіше (1, 4, 7), мають більшу кількість попередніх скасувань ніж інші 4 типи.

Найкращою за кількістю попередніх НЕскасувань є кімната другого типу, яку забронювали відносно немалу кількість разів (а саме 692):

```
table(hotel$room_type_reserved)
```

```
##
## Room_Type 1 Room_Type 2 Room_Type 3 Room_Type 4 Room_Type 5 Room_Type 6
##       28130         692           7        6057         265         966
## Room_Type 7
##         158
```

ШОСТЕ ЗАПИТАННЯ

Формулювання:

- Чи є різниця в кількості попередніх скасувань для клієнтів, які вимагають паркувальне місце і тих, хто його не потребує?

Спочатку, подивимось на кількість скасованих/нескасованих бронювань в залежності від потреби у паркувальному місці у вигляді таблиці:

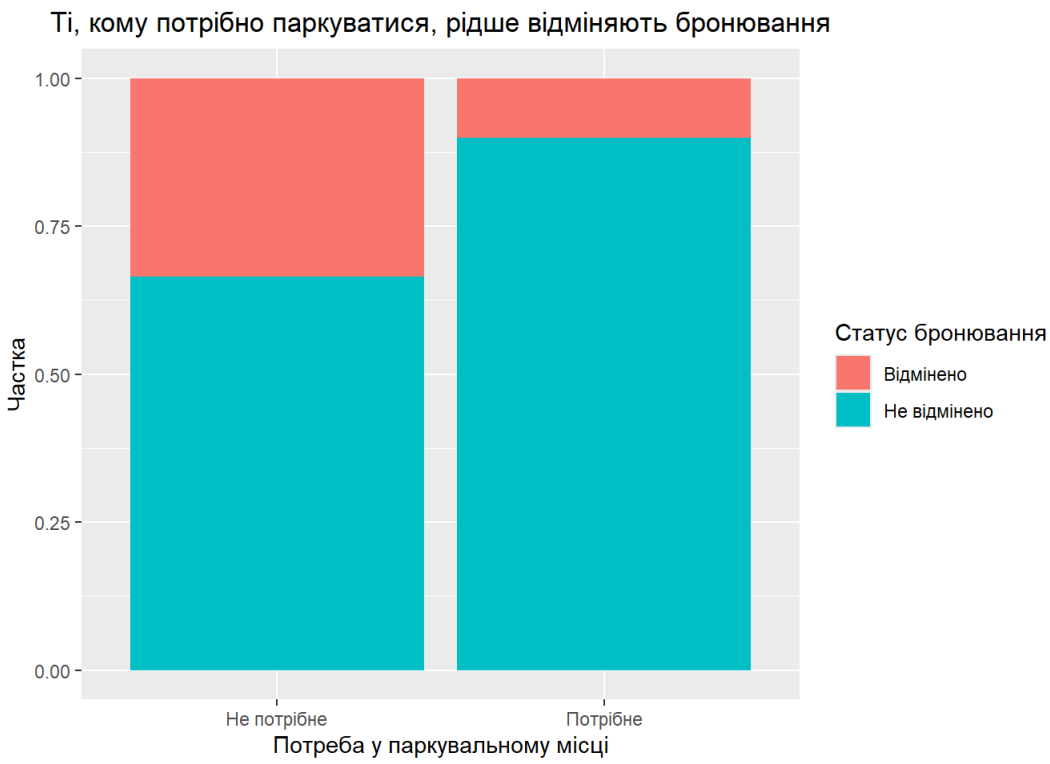
```
cont_tab <- xtabs(~ required_car_parking_space + booking_status, data = hotel)
cont_tab
```

```
##                         booking_status
## required_car_parking_space Canceled Not_Canceled
##                          0    11771        23380
##                          1      114         1010
```

Одразу кидається в очі те, що люди, яким необхідне паркувальне місце, рідше скасовують бронювання. Це питання потребує детальнішого вивчення: з одного боку це досить природньо - якщо людина завчасно говорить за паркувальне місце, вона, напевно, вже достатньо спланувала свій візит до готелю, з іншого - можливо є інші фактори які на це впливають

Подивимось на цей же результат у графічному вигляді: побудуємо стовпчикову діаграму на основі даних про потребу в паркувальному місці, розфарбуємо в різні кольори записи, які відповідно були скасовані і не скасовані.

```
ggplot(hotel, aes(x = as.factor(required_car_parking_space), fill = booking_status)) +
  geom_bar(position = "fill") +
  labs(title = "Ті, кому потрібно паркуватися, рідше відміняють бронювання",
       x = "Потреба у паркувальному місці",
       y = "Частка") +
  scale_fill_manual(name = "Статус бронювання",
                    values = c("Not_Canceled" = "#00bfc4", "Canceled" = "#f8766d"),
                    labels = c("Відмінено", "Не відмінено"))+
  scale_x_discrete(labels = c("Не потрібне", "Потрібне")) + theme(plot.title = element_text(hjust = 0.5))
```



інше

Таблиця статусу бронювання в залежності від його типу і обернена до неї. Найбільше записів зроблено онлайн, більша частина з них - скасовані; загалом скасовані записи складають приблизно 30% від всіх записів.

```
addmargins(cont_tab <- xtabs(~ market_segment_type + booking_status, data = hotel))
```

```
##                booking_status
## market_segment_type Canceled Not_Canceled   Sum
##       Aviation           37          88   125
##       Complementary       0         391   391
##       Corporate         220        1797  2017
##       Offline          3153        7375 10528
##       Online           8475       14739 23214
##       Sum             11885       24390 36275
```

```
addmargins(cont_tab <- xtabs(~ booking_status + market_segment_type, data = hotel))
```

```
##              market_segment_type
## booking_status Aviation Complementary Corporate Offline Online   Sum
##   Canceled          37            0       220    3153   8475 11885
##   Not_Canceled      88          391      1797    7375  14739 24390
##   Sum              125          391      2017   10528  23214 36275
```

Таблиця залежності кількості особливих побажань від кількості дітей. Можемо побачити, що кількість особливих побажань не залежить від кількості дітей явно.

```
addmargins(cont_tab <- xtabs(~ no_of_children + no_of_special_requests, data = hotel))
```

```
##              no_of_special_requests
## no_of_children     0     1     2     3     4     5   Sum
##          0   18831 10484  3655   538    61     8 33577
##          1     497   536   485    93     7     0  1618
##          2     442   347   215    44    10     0  1058
##          3       6     4     9     0     0     0    19
##          9       1     1     0     0     0     0     2
##          10      0     1     0     0     0     0     1
##          Sum 19777 11373  4364   675    78     8 36275
```

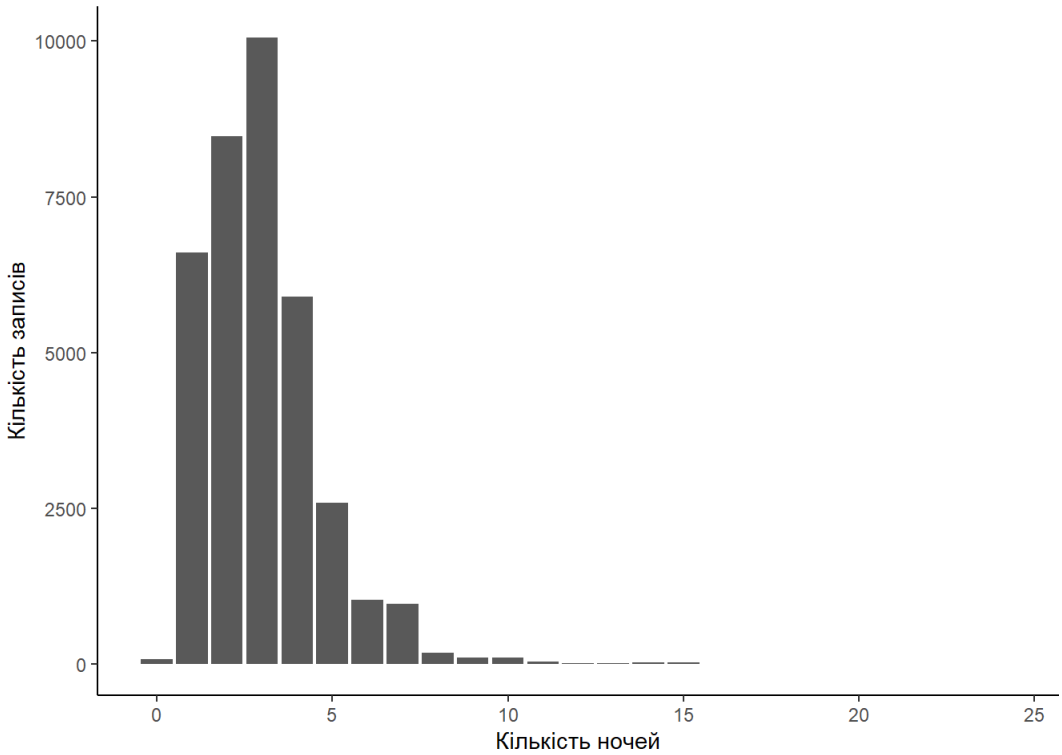
Таблиця залежності типу кімнати від типу плану харчування. Перший план харчування найпопулярніший, третій план майже не вибирали

```
addmargins(cont_tab <- xtabs(~ type_of_meal_plan + room_type_reserved, data = hotel))
```

```
##                 room_type_reserved
## type_of_meal_plan Room_Type 1 Room_Type 2 Room_Type 3 Room_Type 4 Room_Type 5
##      Meal Plan 1       20157        653          5       5748        242
##      Meal Plan 2        2934         16          0        273         14
##      Meal Plan 3           1          0          0          1          0
##      Not Selected       5038         23          2         35          9
##      Sum               28130        692          7       6057        265
##                 room_type_reserved
## type_of_meal_plan Room_Type 6 Room_Type 7   Sum
##      Meal Plan 1         878        152 27835
##      Meal Plan 2          66          2  3305
##      Meal Plan 3           0          3     5
##      Not Selected         22          1  5130
##      Sum                 966        158 36275
```

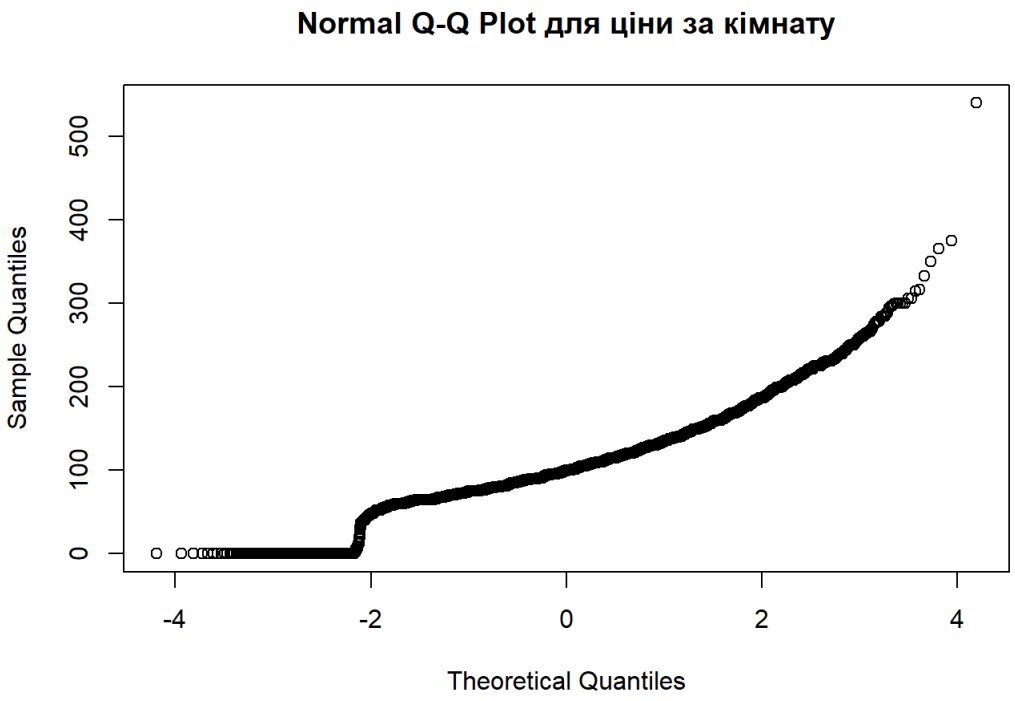
Гістограма кількості ночей, що гості планували провести в готелі. Як бачимо, люди частіше залишаються в готелі до 5 ночей, далі записів стає значно менше.

```
ggplot(hotel_with_nights, aes(x = no_of_nights)) +
  geom_bar() +
  theme_classic() + labs(
      x = "Кількість ночей",
      y = "Кількість записів")
```



Графік qqnorm для ціни за кімнату. Можна побачити викид зверху, вже було згадано, що це, ймовірно, одруківка, і багато записів з 0 значеннями - договірна ціна

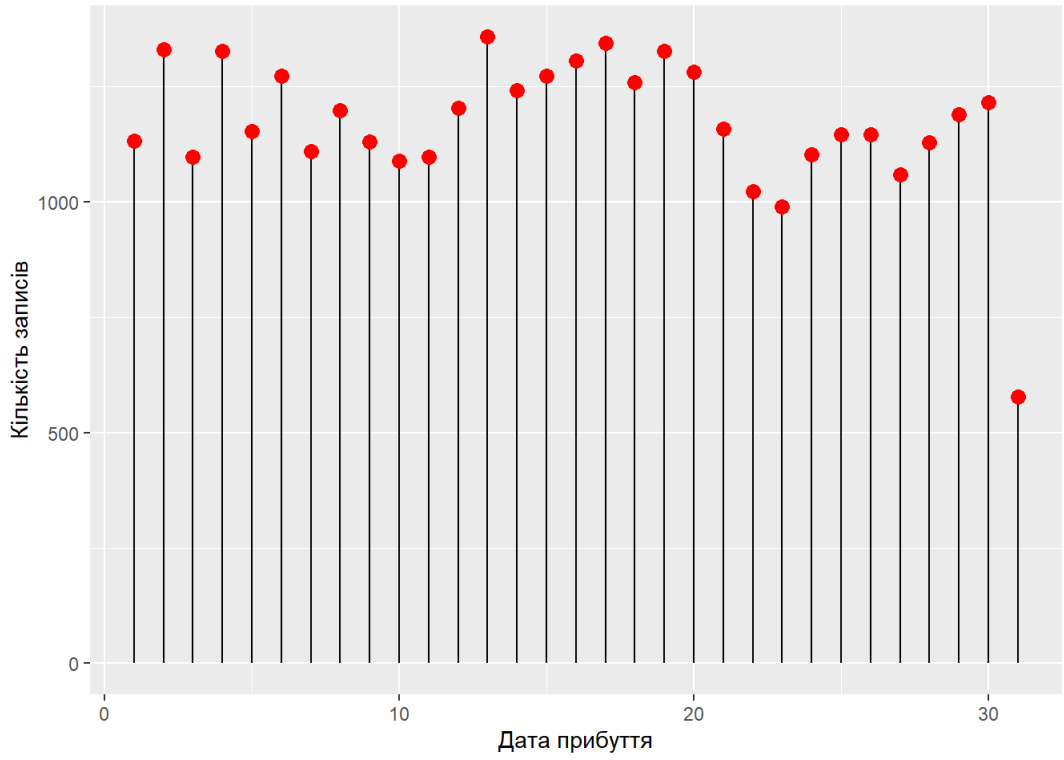
```
qqnorm(hotel$avg_price_per_room, main = 'Normal Q-Q Plot для ціни за кімнату')
```



гістограми

Гістограма для дати прибуття в готель. Очевидно, менше записів на 31 число - не кожен місяць має його.

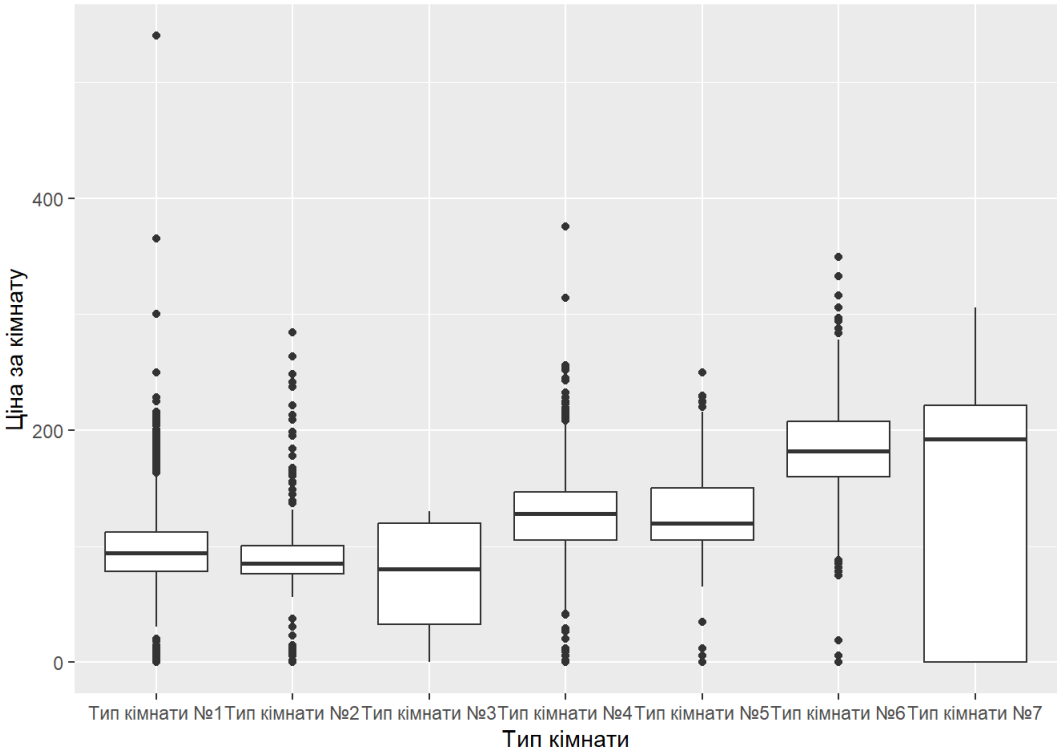
```
hotel_cusn <- hotel %>%
  group_by(arrival_date) %>%
  summarise(total = n())

ggplot(hotel_cusn, aes(x = arrival_date, y = total)) +
  geom_segment(aes(x = arrival_date, xend = arrival_date, y=total, yend=0)) +
  geom_point(color = 'red', size = 3) + labs(x = "Дата прибуття", y = "Кількість записів")
```



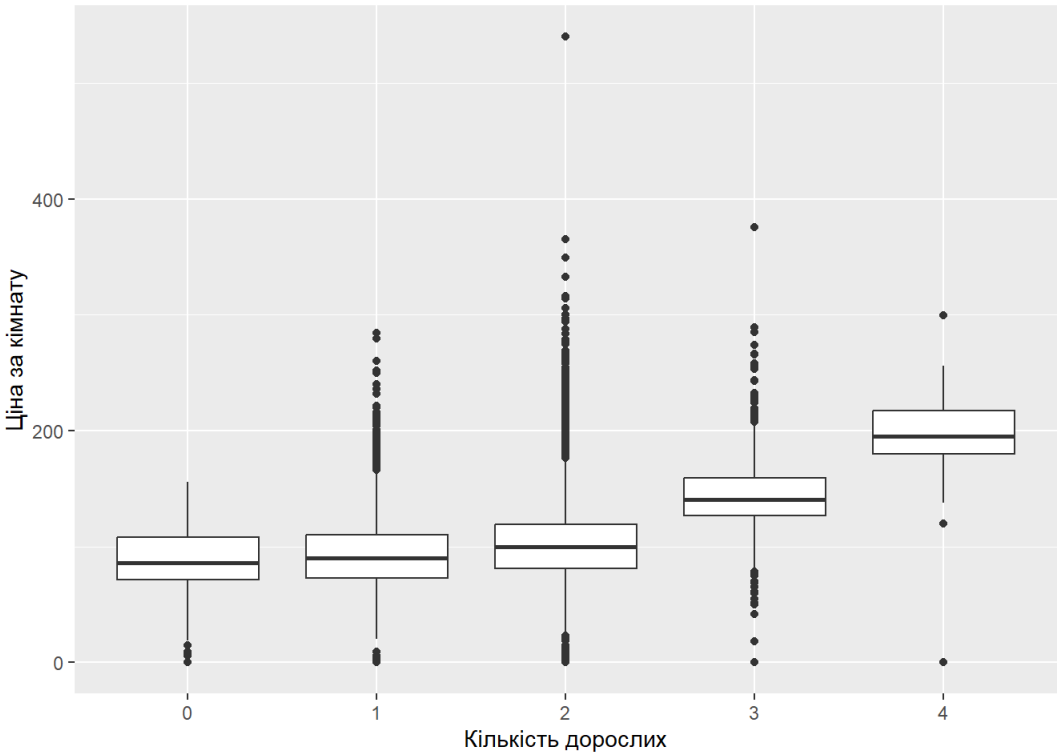
вусаті скриньки

Наведені вусаті скриньки залежності ціни від типів кімнати, плану харчування, кількості дорослих та дітей. Як бачимо, тип кімнати помітно впливає на ціну, так само і кількість дорослих і кількість дітей. Проте, важко сказати для плану харчування - не так багато записів для 2 і 3 типів.

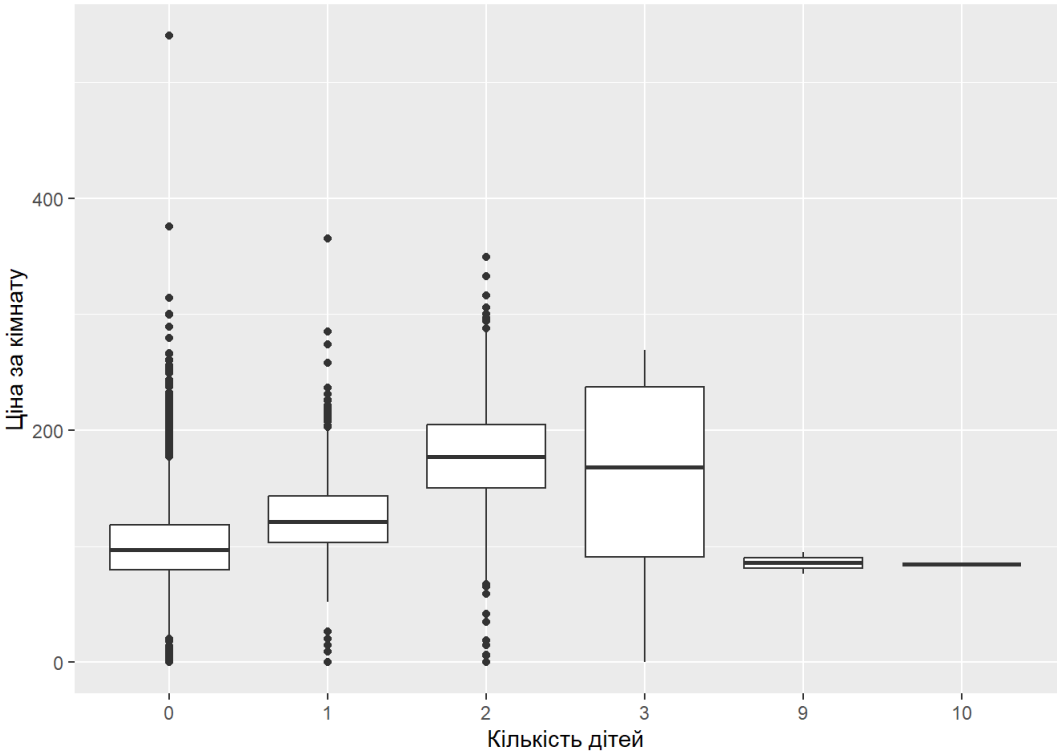
```
ggplot(hotel, aes(x = room_type_reserved, y = avg_price_per_room)) +
  geom_boxplot(varwidth = FALSE) + labs (x = "Тип кімнати", y = "Ціна за кімнату") +
  scale_x_discrete(labels = room_label_vector)
```



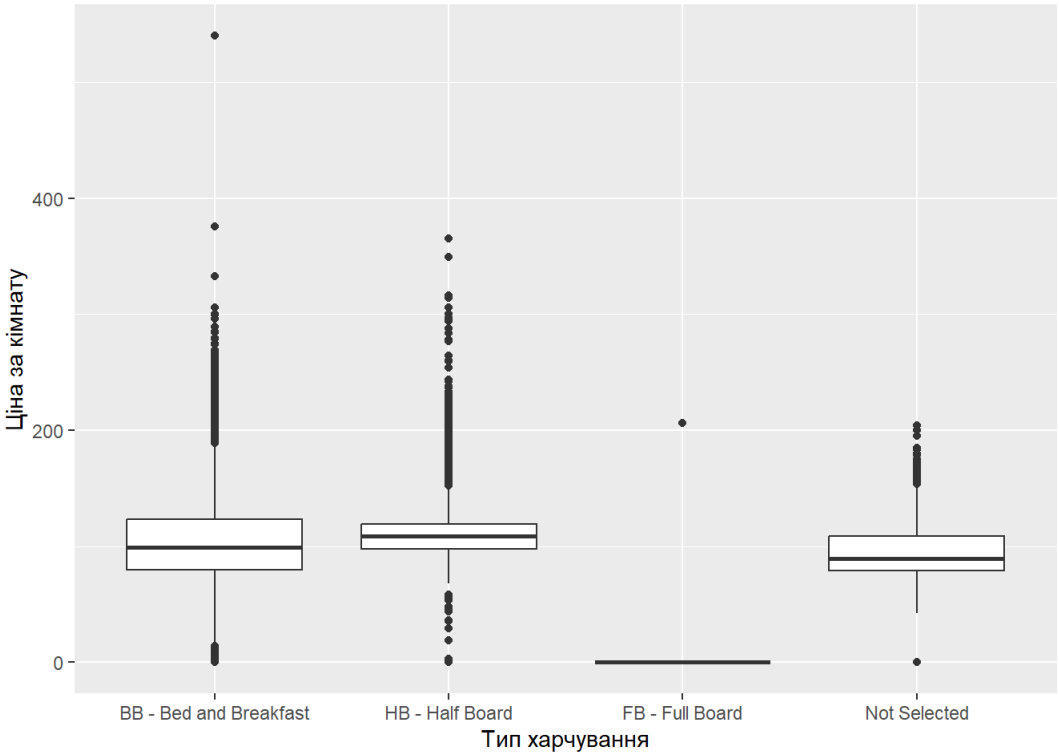

```
ggplot(hotel, aes(x = factor(no_of_adults), y = avg_price_per_room)) +
  geom_boxplot(varwidth = FALSE) + labs(x = "Кількість дорослих", y = "Ціна за кімнату")
```




```
ggplot(hotel, aes(x = factor(no_of_children), y = avg_price_per_room)) +
  geom_boxplot(varwidth = FALSE) + labs(x = "Кількість дітей", y = "Ціна за кімнату")
```

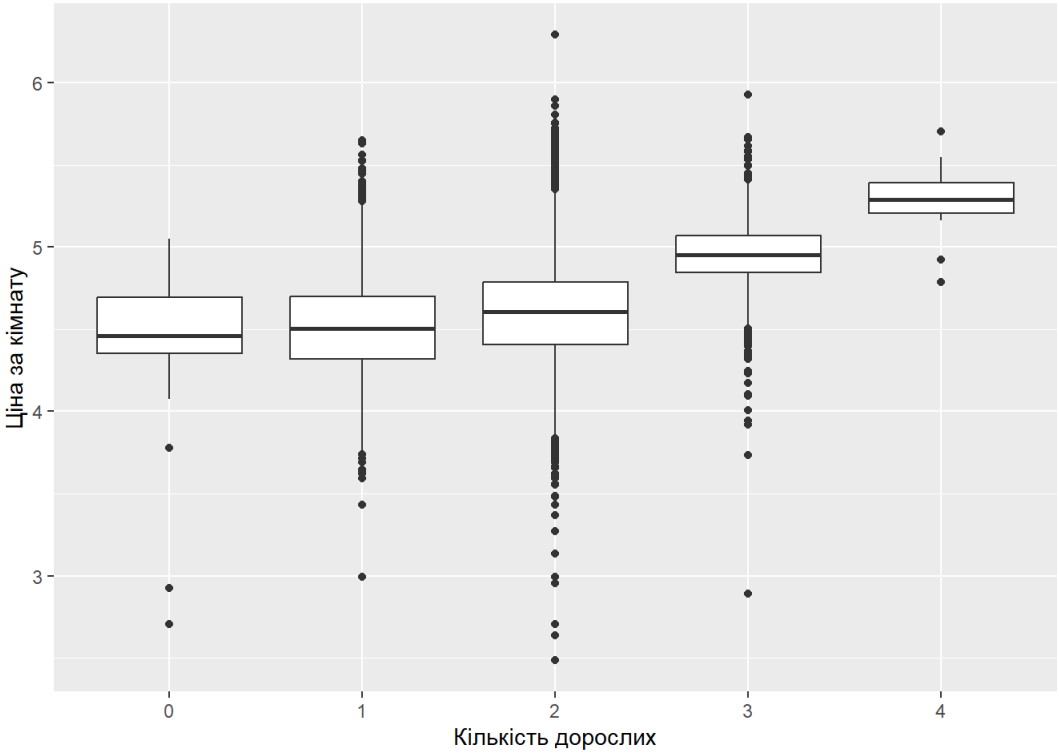



```
ggplot(hotel, aes(x = type_of_meal_plan, y = avg_price_per_room)) +
  geom_boxplot(varwidth = FALSE) + labs(x = "Тип харчування", y = "Ціна за кімнату") +
  scale_x_discrete(labels = meal_label_vector)
```



Прологаритмуємо ціну для вусатої скриньки залежності ціни від кількості дорослих, щоб краще побачити відмінності. Видно, що ціни для 3 і 4 дорослих значно відрізняються.

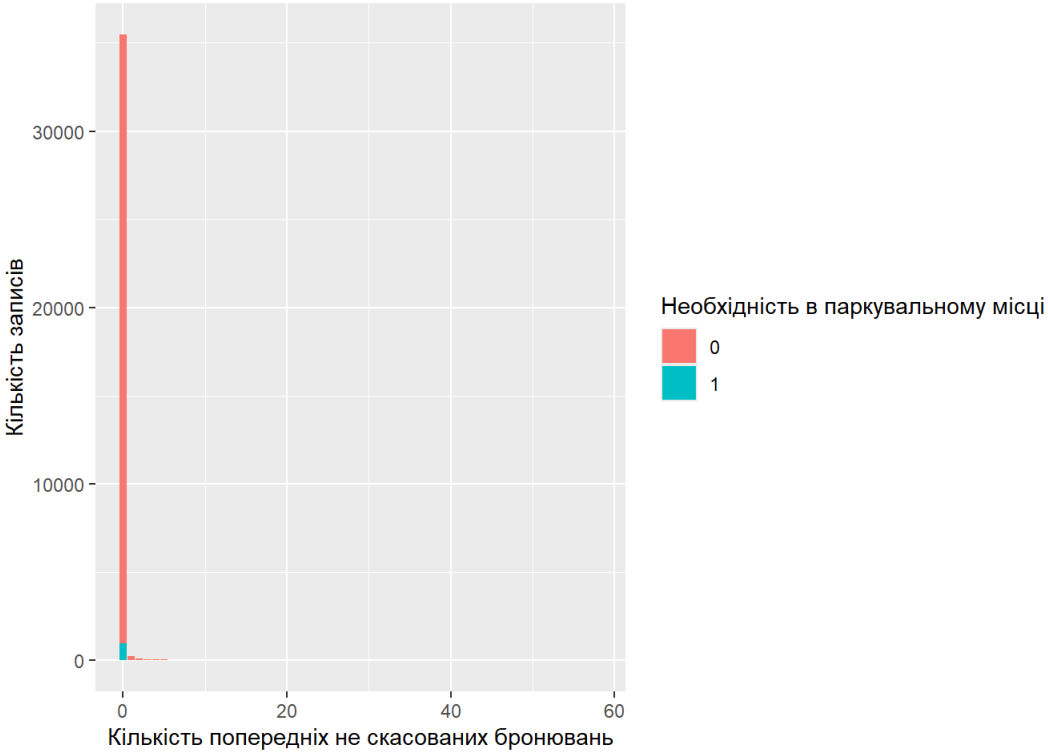
```
ggplot(hotel_lead_check, aes(x = factor(no_of_adults), y = log(avg_price_per_room))) +
  geom_boxplot(varwidth = FALSE) + labs(x = "Кількість дорослих", y = "Ціна за кімнату")
```



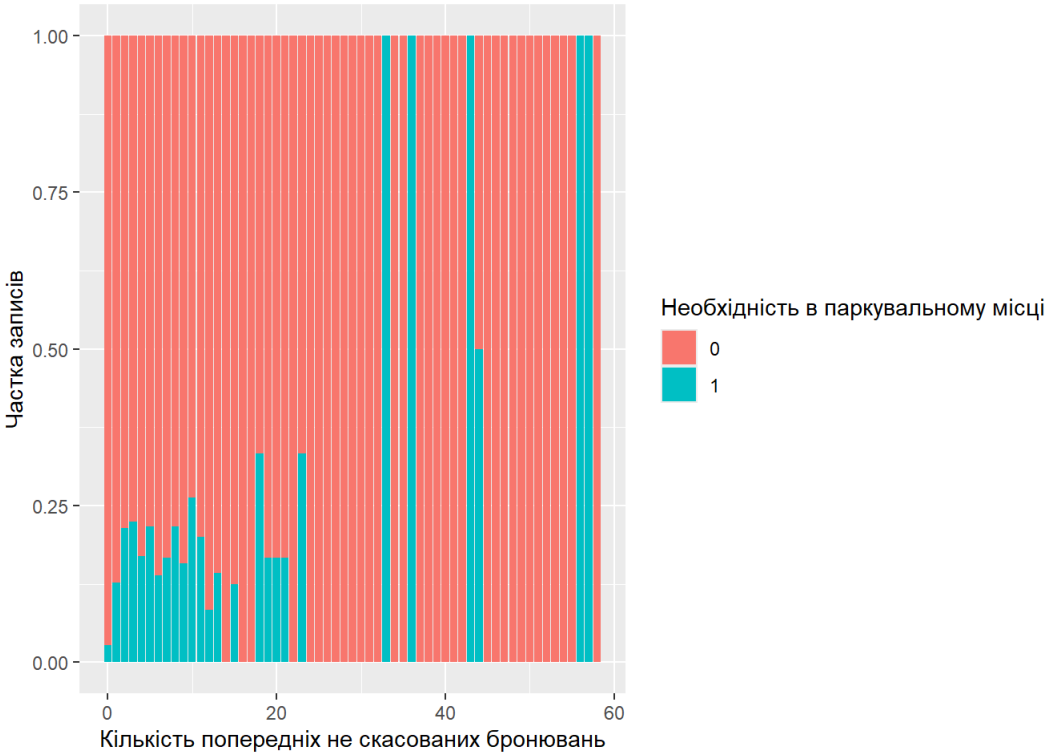
стовпчикові діаграми

Побудуємо стовпчикову діаграму для кількості попередніх НЕскасованих бронювань, відмітимо кольором необхідність у паркувальному місці. Складно побачити якісь закономірності, так як існує дуже мало записів, де кількість попередніх НЕскасувань більше 0

```
ggplot(hotel, aes(x = no_of_previous_bookings_not_canceled, fill = required_car_parking_space)) +
  geom_bar() + labs( x = "Кількість попередніх не скасованих бронювань", y = "Кількість записів", fill = "Необхідність в паркувальному місці")
```

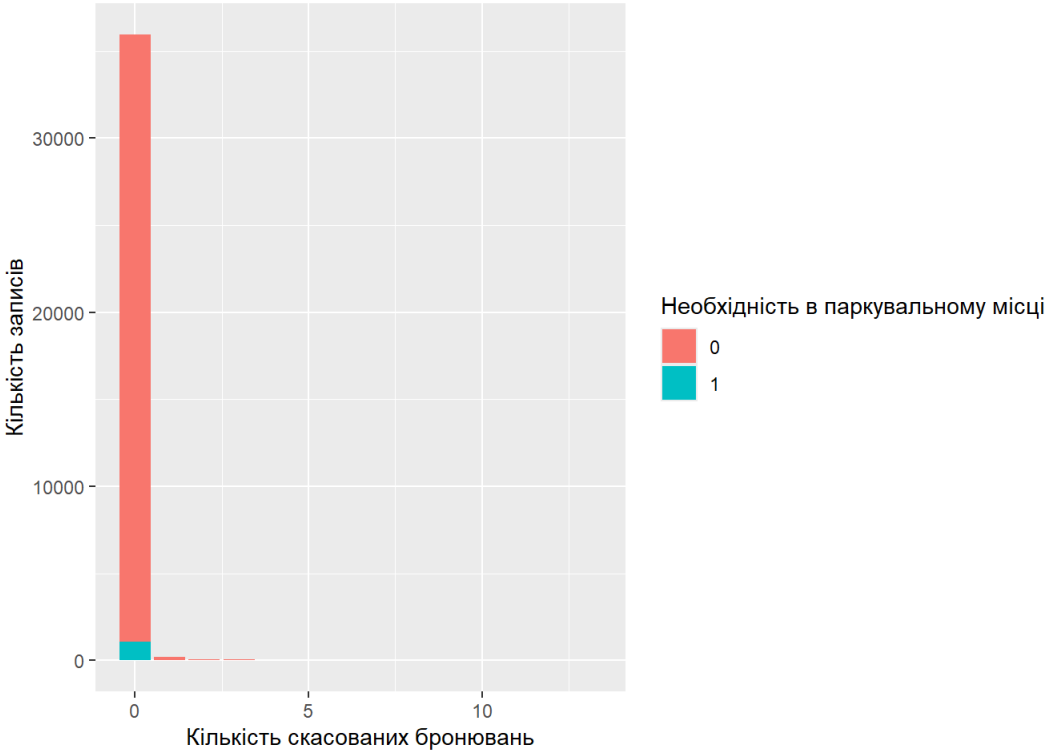



```
ggplot(hotel, aes(x = no_of_previous_bookings_not_canceled, fill = required_car_parking_space)) +
  geom_bar(position = 'fill') + labs( x = "Кількість попередніх не скасованих бронювань", y = "Частка записів", fill = "Необхідність в паркувальному місці")
```

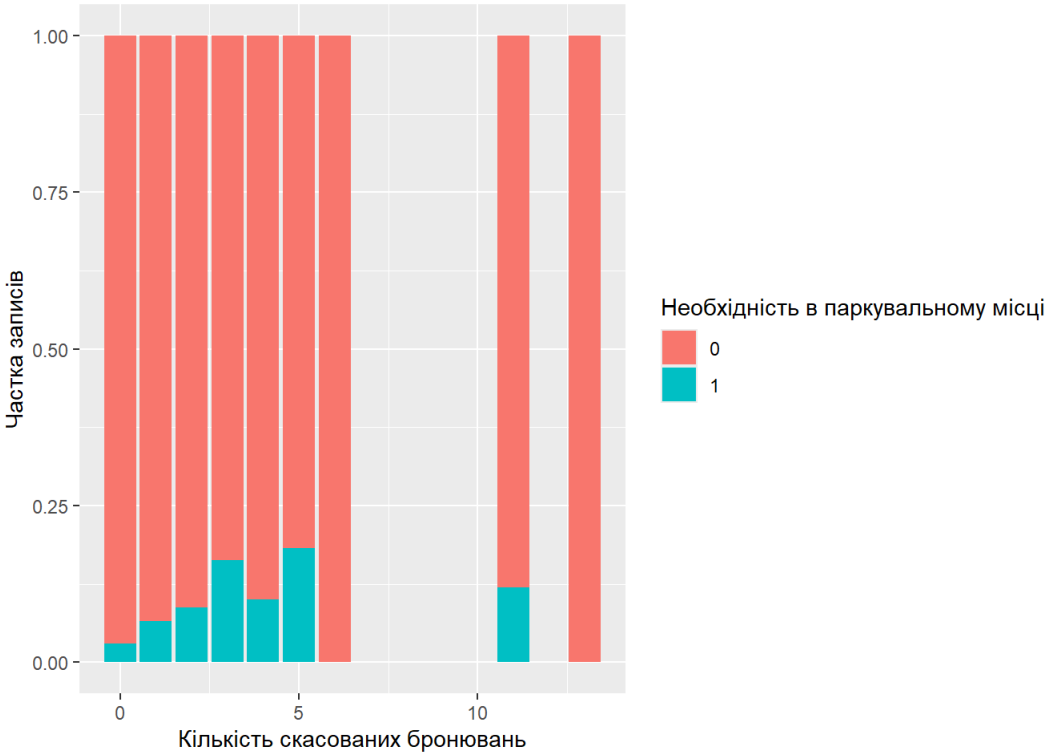


Окрім цього, побудуємо стовпчикову діаграму для кількості попередніх скасованих бронювань, відмітимо кольором необхідність у паркувальному місці. Складно побачити якісь закономірності з аналогічних причин

```
ggplot(hotel, aes(x = no_of_previous_cancellations, fill = required_car_parking_space)) +
  geom_bar() + labs( x = "Кількість скасованих бронювань", y = "Кількість записів", fill = "Необхідність в паркувальному місці")
```



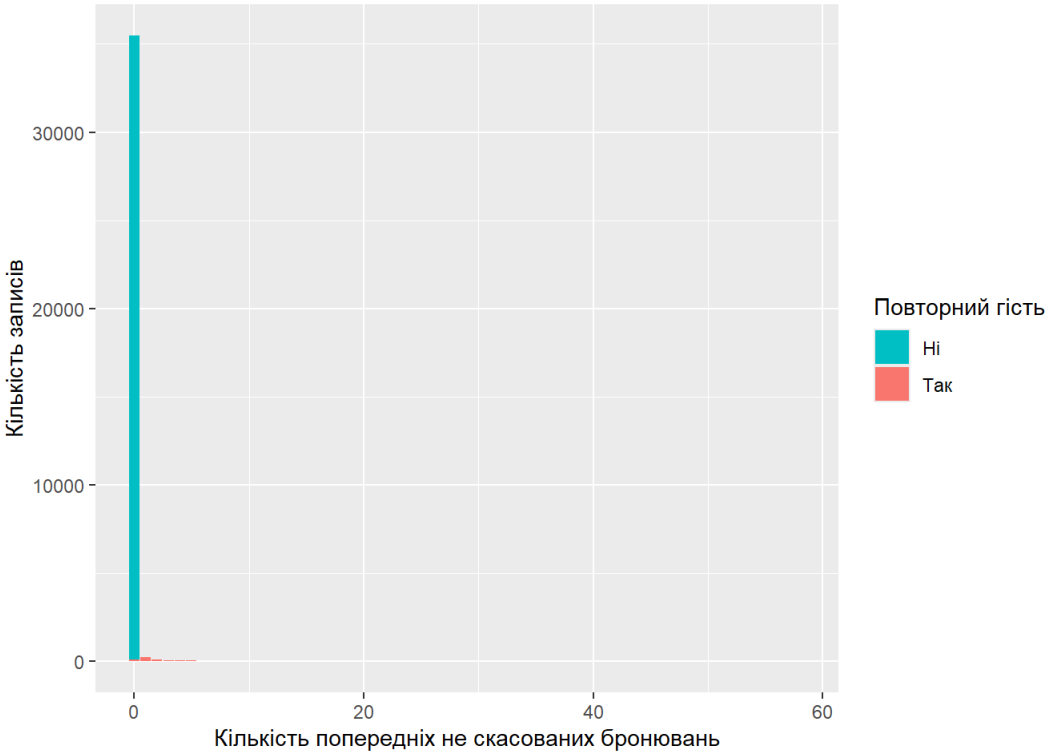

```
ggplot(hotel, aes(x = no_of_previous_cancellations, fill = required_car_parking_space)) +
  geom_bar(position = 'fill') + labs( x = "Кількість скасованих бронювань", y = "Частка записів", fill = "Необхідність в паркувальному місці")
```



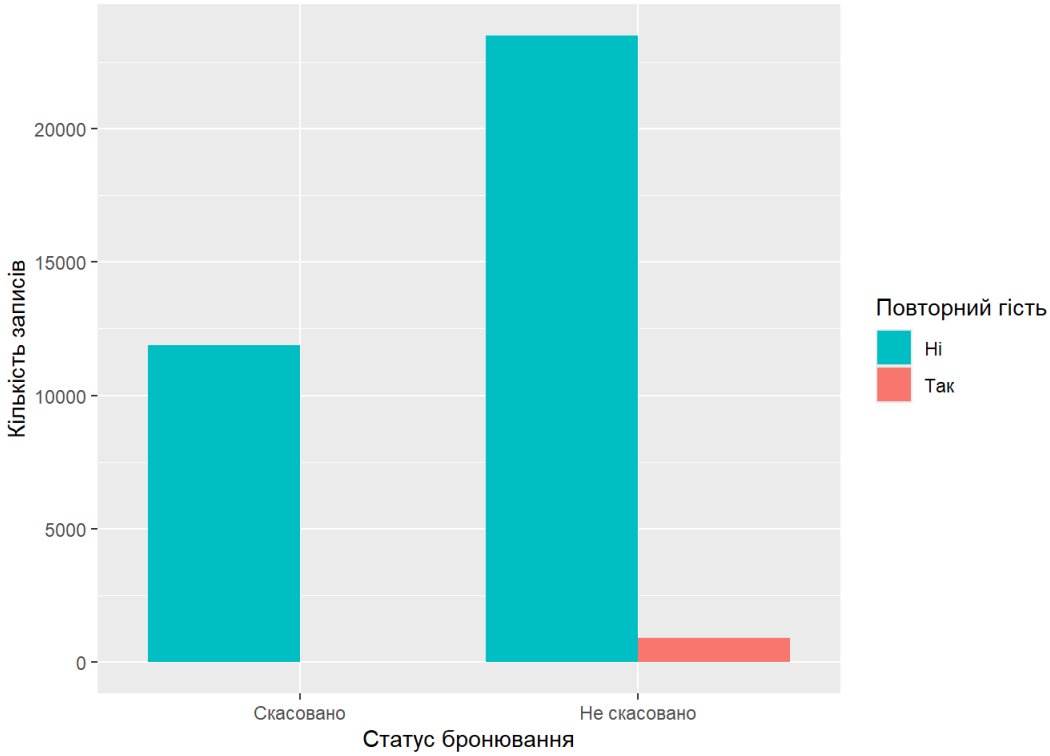
Побудуємо стовпчикові діаграми для кількості попередніх НЕскасованих бронювань і статусу бронювань, кольором відмітимо повторних гостей.

- Можемо побачити, що повторний гість має більше попередніх нескасованих бронювань, що не дивно
- Повторні гості частіше не скасовують бронювання

```
ggplot(hotel, aes(x = no_of_previous_bookings_not_canceled, fill = as.factor(repeated_guest))) +
  geom_bar() + labs( x = "Кількість попередніх не скасованих бронювань", y = "Кількість записів") + scale_fill_manual(name = "Повторний гість",
                     values = c("0" = "#00bfc4", "1" = "#f8766d"),
                     labels = c("Ні", "Так"))
```




```
ggplot(hotel, aes(x = booking_status, fill = repeated_guest)) +
  geom_bar(position = "dodge") + labs( x = "Статус бронювання", y = "Кількість записів") + scale_fill_manual(name = "Повторний гість",
                     values = c("0" = "#00bfc4", "1" = "#f8766d"),
                     labels = c("Ні", "Так")) +
  scale_x_discrete(labels = c("Canceled" = "Скасовано", "Not_Canceled" = "Не скасовано"))
```

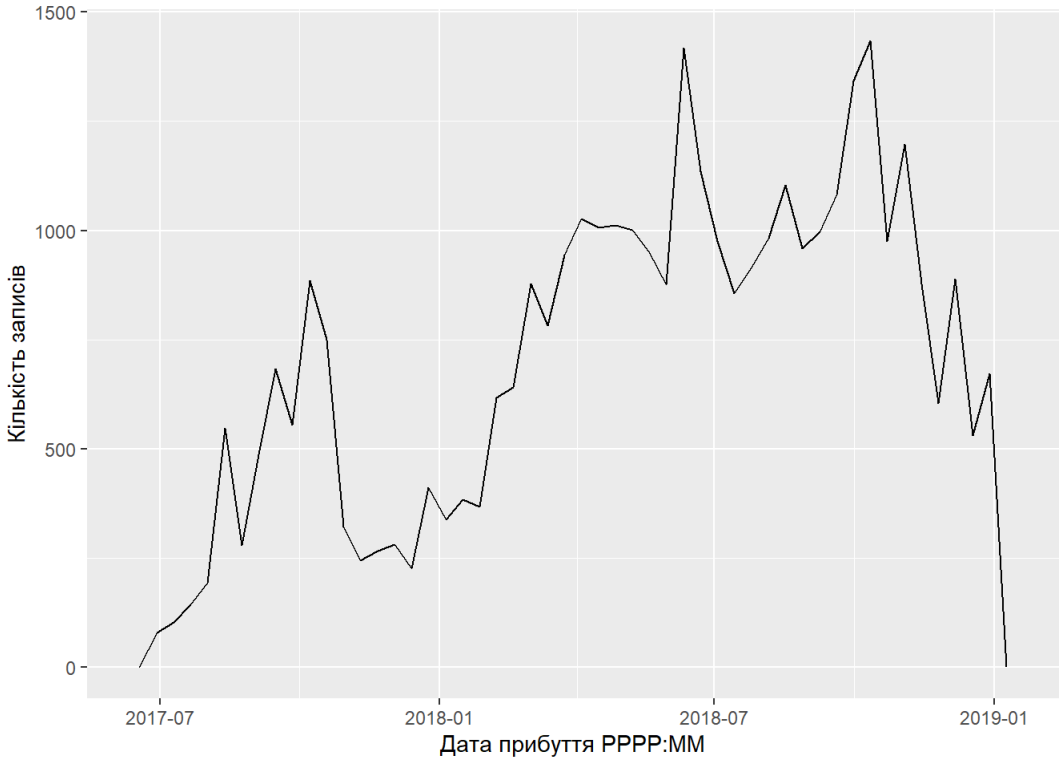


Побудуємо графік кількості записів в залежності від дати прибуття, і порівняємо його з графіком залежності ціни від дати прибуття.

Можна побачити схожі тенденції - взимку 2017-2018 року було менше записів, і як наслідок, бачимо падіння і ціни. І навпаки, коли записів стає більше - влітку 2018 року, ціна також зростає.

```
ggplot(hotel, aes(x = as.Date(arrival_year_and_month))) +
  geom_freqpoly(bins = 50) + labs( x = "Дата прибуття PPPP:MM", y = "Кількість записів")
```

```
## Warning: Removed 37 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

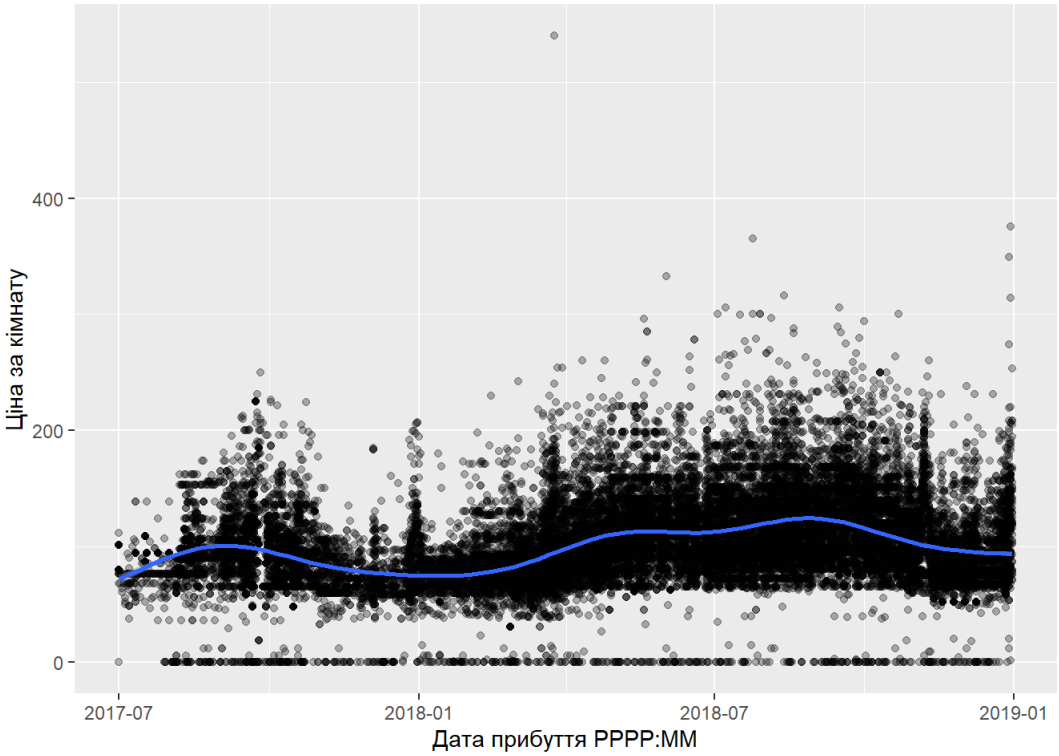



```
ggplot(hotel, aes(x = as.Date(arrival_year_and_month), y = avg_price_per_room)) +
  geom_point(alpha = 0.3) +
  geom_smooth() + labs( x = "Дата прибуття PPPP:MM", y = "Ціна за кімнату")
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 37 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 37 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

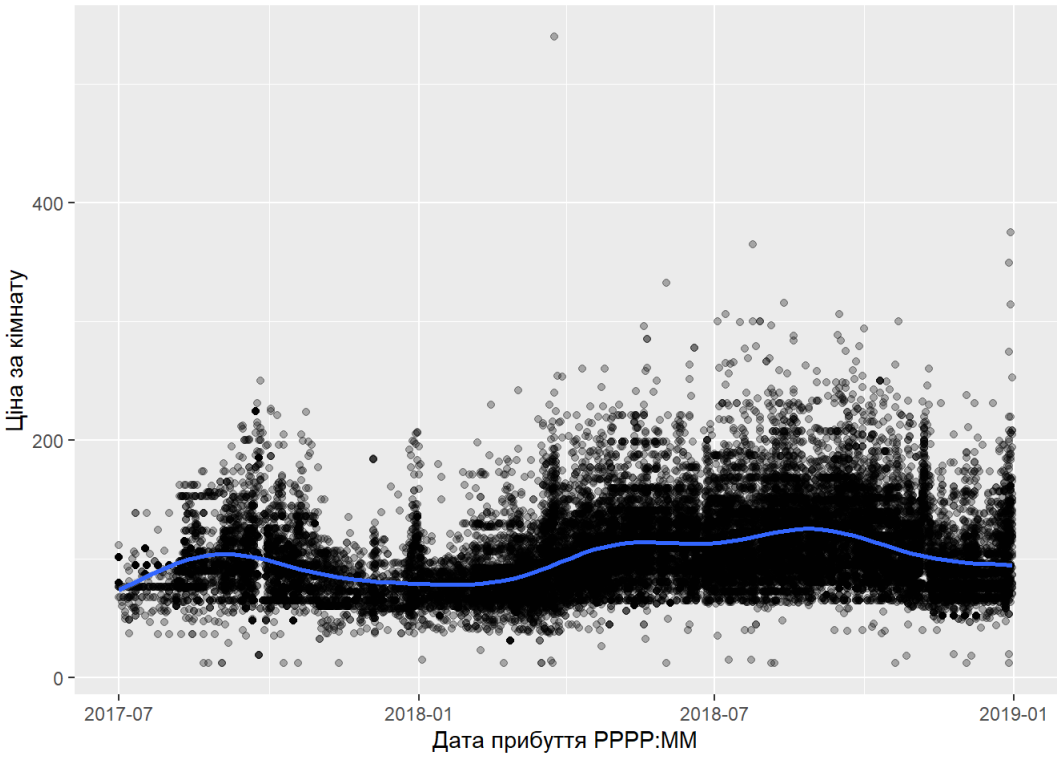



```
ggplot(hotel_avg_prices, aes(x = as.Date(arrival_year_and_month), y = avg_price_per_room)) +
  geom_point(alpha = 0.3) +
  geom_smooth() + labs( x = "Дата прибуття PPPP:MM", y = "Ціна за кімнату")
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

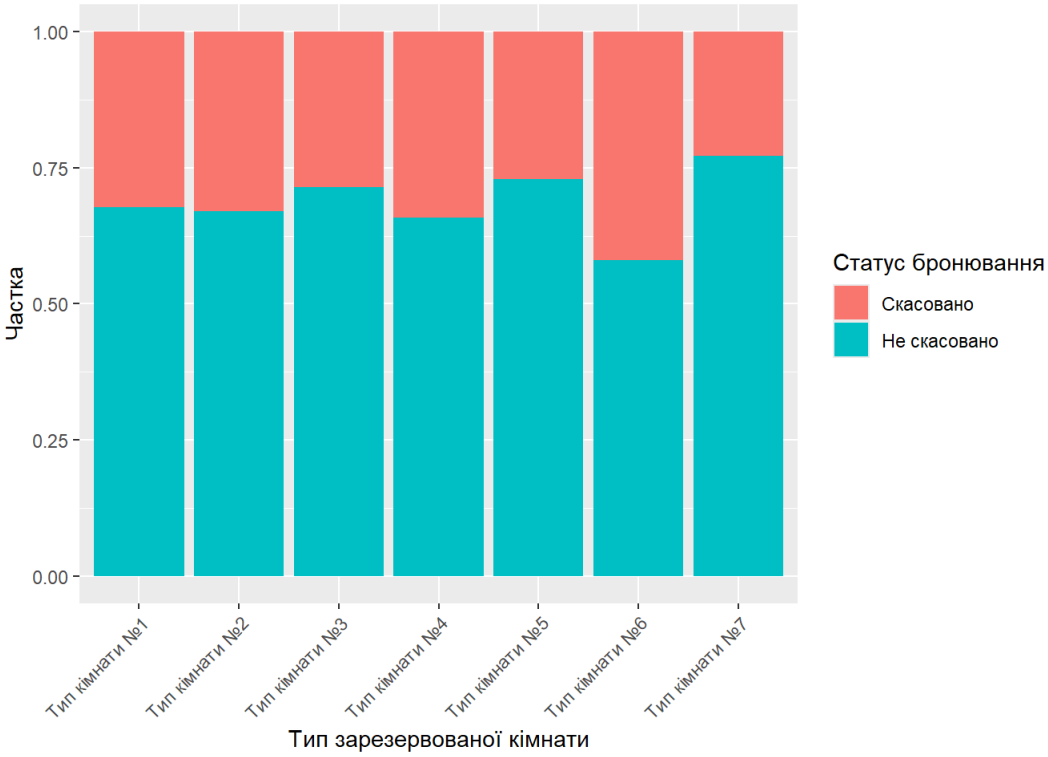
```
## Warning: Removed 36 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 36 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



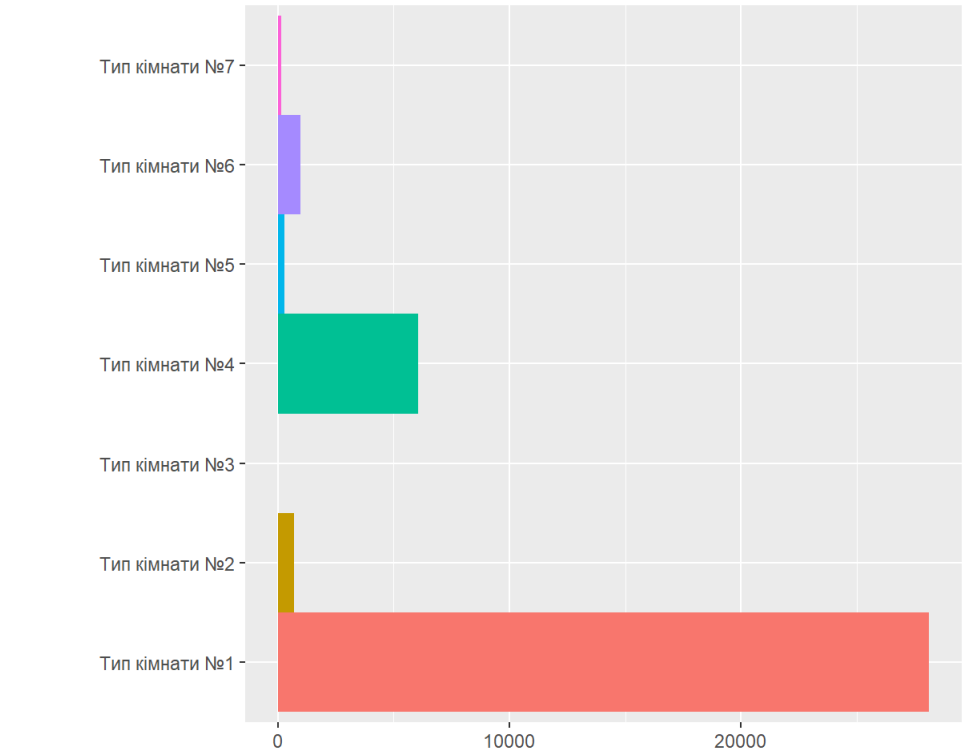
Побудуємо стовпчикову діаграму для типів зарезервованих кімнат, кольором відмітимо статус бронювання. Можна сказати, що статус бронювання розподілений приблизно рівномірно.

```
ggplot(hotel, aes(x = room_type_reserved, fill = booking_status)) + geom_bar(position = "fill") +
  labs(x = "Тип зарезервованої кімнати", y = "Частка", fill = "Статус бронювання") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_manual(name = "Статус бронювання",
                    values = c("Canceled" = "#f8766d", "Not_Canceled" = "#00bfc4"),
                    labels = c("Скасовано", "Не скасовано"))+
  scale_x_discrete(labels = room_label_vector) + theme(plot.title = element_text(hjust = 0.5))
```

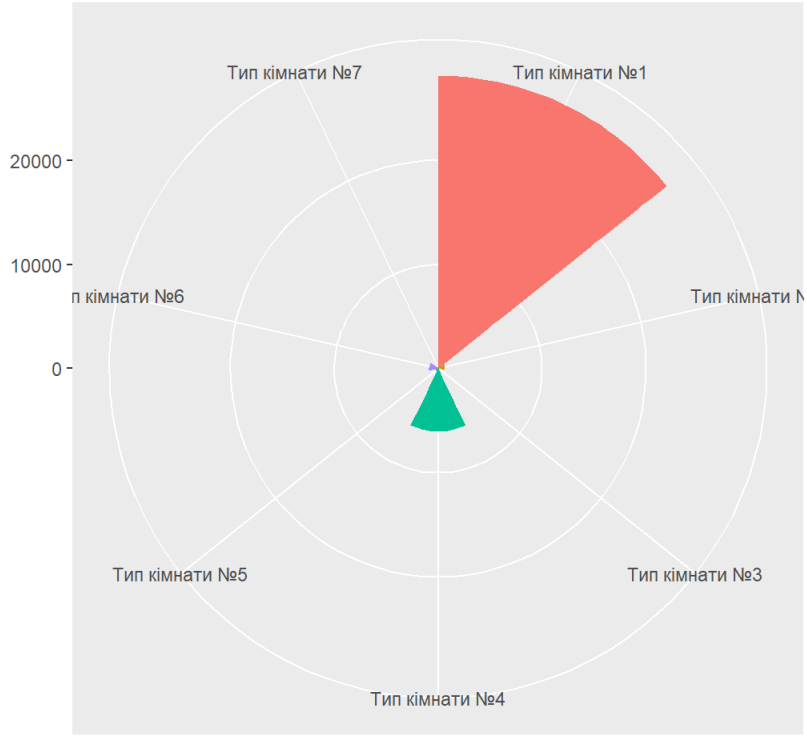


додаткові типи графіків) візуалізація кількості зарезервованих кімнат різних типів

```
bar <- ggplot(data = hotel) +
  geom_bar(
    mapping = aes(x = room_type_reserved, fill = room_type_reserved),
    show.legend = FALSE,
    width = 1
  ) +
  theme(aspect.ratio = 1) +
  labs(x = NULL, y = NULL) + scale_x_discrete(labels = room_label_vector)

bar + coord_flip()
```

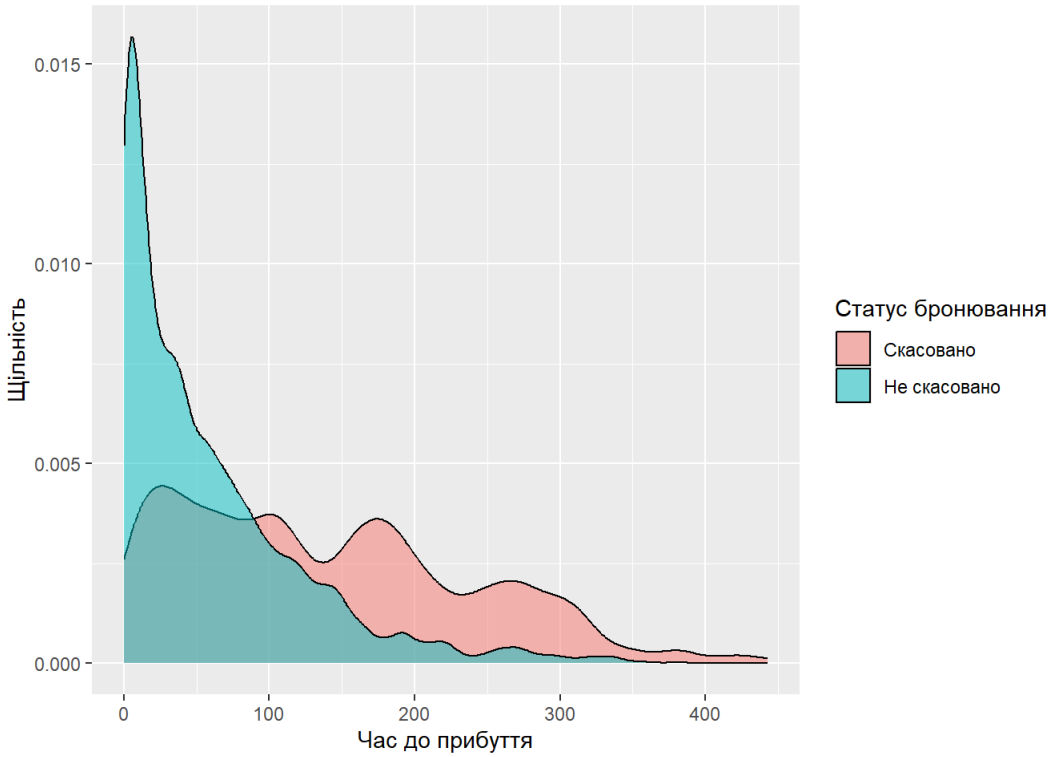


```
bar + coord_polar()
```



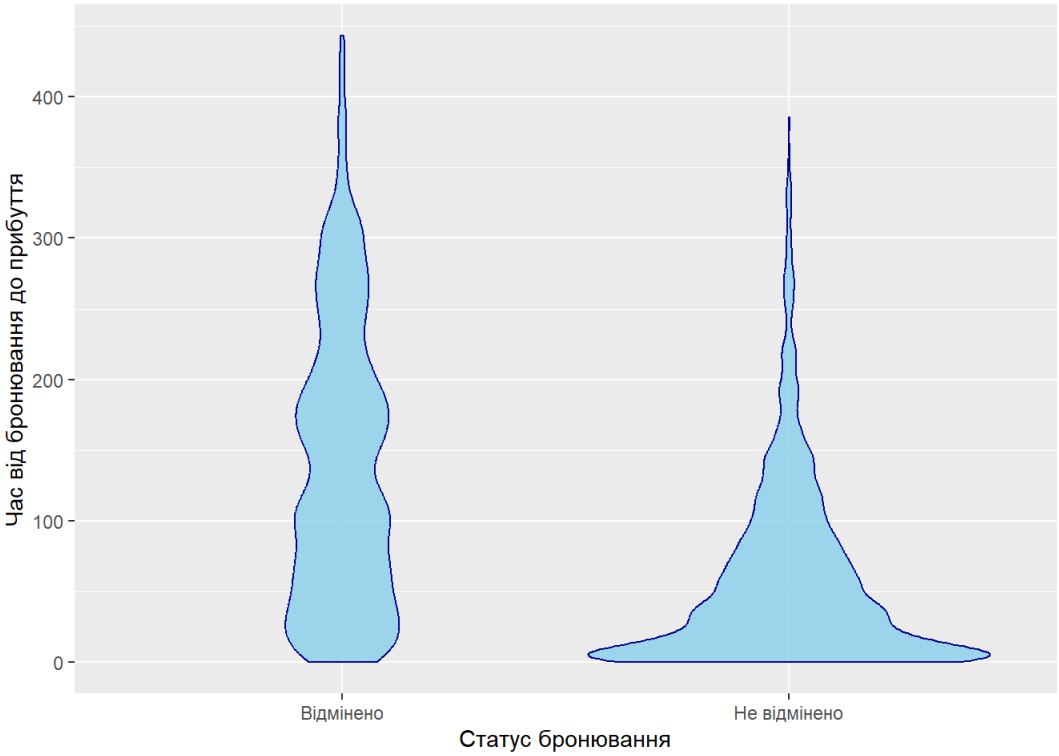
Графік, що показує як впливає час до прибуття на статус бронювання. Можемо побачити, що починаючи з часу приблизно в 90 годин, люди починають частіше скасовувати записи, ніж не скасовувати.

```
ggplot(hotel, aes(x=lead_time, fill=booking_status))+
  geom_density(alpha=0.5) + labs( x = "Час до прибуття", y = "Щільність") +
  scale_fill_manual(name = "Статус бронювання",
                    values = c("Canceled" ="#f8766d", "Not_Canceled" = "#00bfc4"),
                    labels = c( "Скасовано",  "Не скасовано"))
```



Violin-графік для тих же величин

```
ggplot(hotel, aes(x=booking_status, y=lead_time))+
  geom_violin(fill = "skyblue", color = "darkblue", alpha = 0.8)+
  scale_x_discrete(labels = c("Відмінено", "Не відмінено"))+
  labs(x="Статус бронювання",
       y="Час від бронювання до прибуття")
```



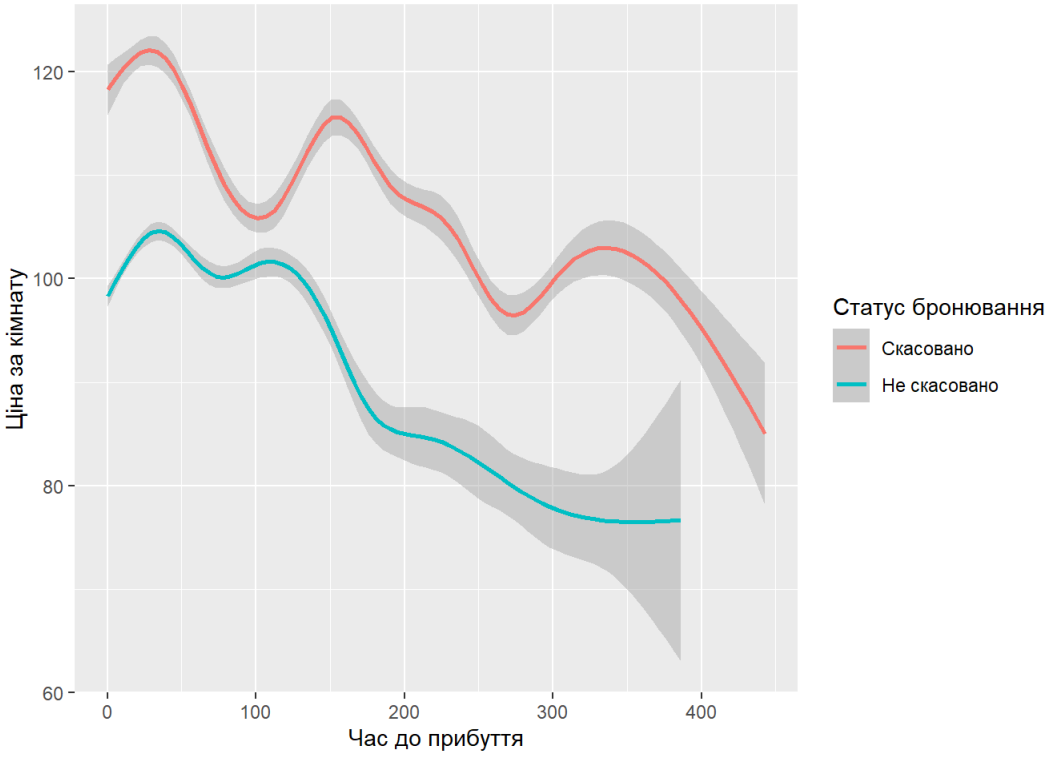
Графік залежності середньої ціни за кімнату в залежності від часу прибуття.

Можемо побачити, що

- скасовані записи зазвичай дорожче,
- чим більше час до прибуття - тим, в середньому, менше ціна.

```
ggplot(data = hotel, aes(x = lead_time, y = avg_price_per_room, color = booking_status)) +
  geom_smooth() + labs (x = "Час до прибуття", y = "Ціна за кімнату") +
  scale_color_manual(name = "Статус бронювання",
                     values = c("Canceled" ="#f8766d", "Not_Canceled" = "#00bfc4"),
                     labels = c( "Скасовано",  "Не скасовано"))
```

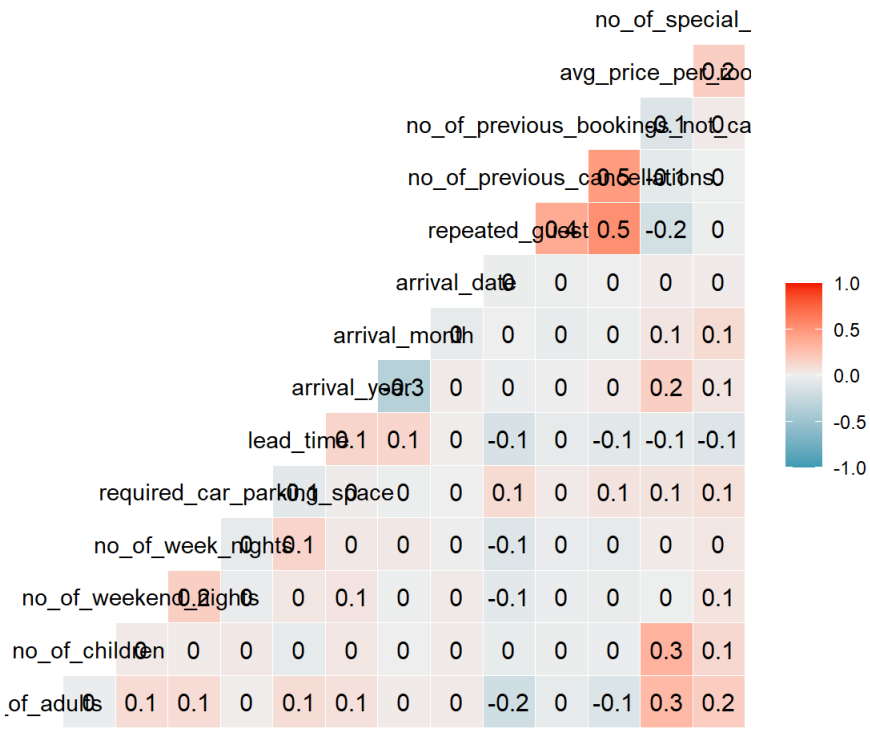
```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



кореляція

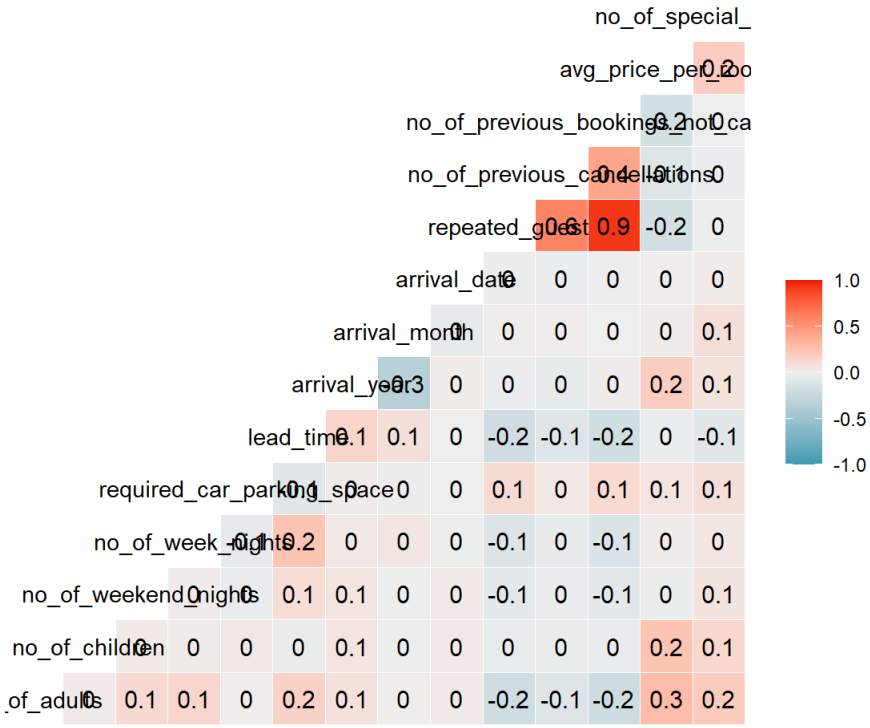
Як можна бачити, особливо сильної кореляції між змінними немає

```
ggcorr(hotel_corr %>% select(where(is.numeric)), label = TRUE)
```



Якщо і тому числі розглядати факторні змінні, можемо побачити дуже сильну кореляцію між тим, що попередні записи не були скасовані і це повторний гість. Але в даній кореляції немає нічого незвичайного, очевидно, якщо людина повторний гість - вона не відміняла попередні записи.

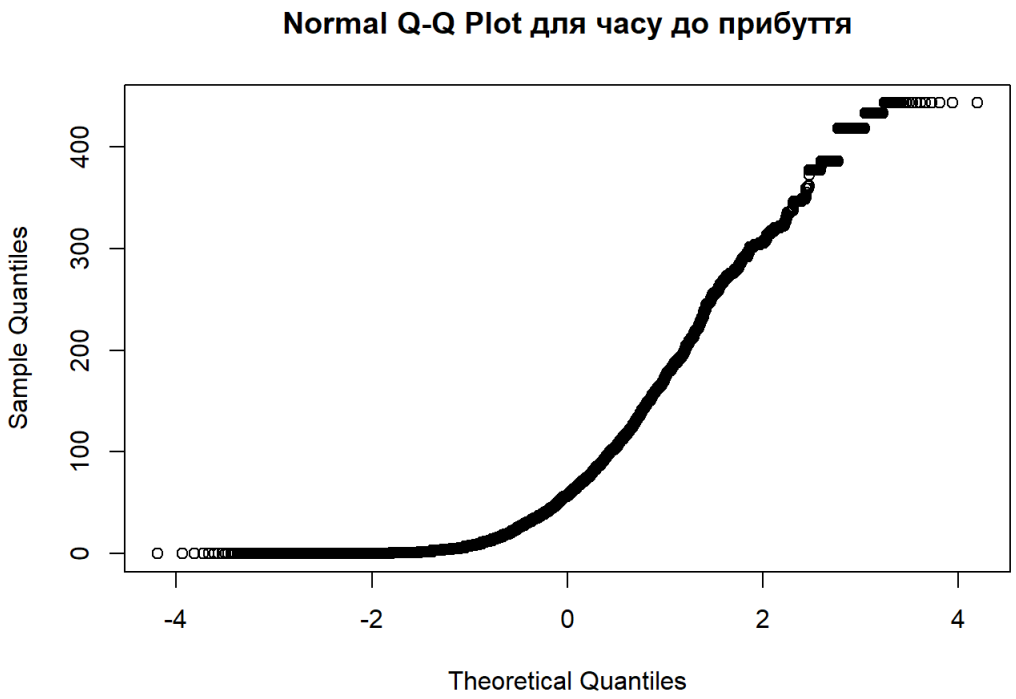
```
ggcorr(hotel_corr %>% select(where(is.numeric)), label = TRUE,
       method = c("pairwise", "spearman"))
```



побудова qq-графіків

Спершу побудуємо qqnorm-графік для часу до прибуття. Можемо побачити, що в нас дуже багато записів з 0 часом до прибуття. Такий викид тривіально пояснюється, точно є люди які як тільки прийшли одразу і поселилися.

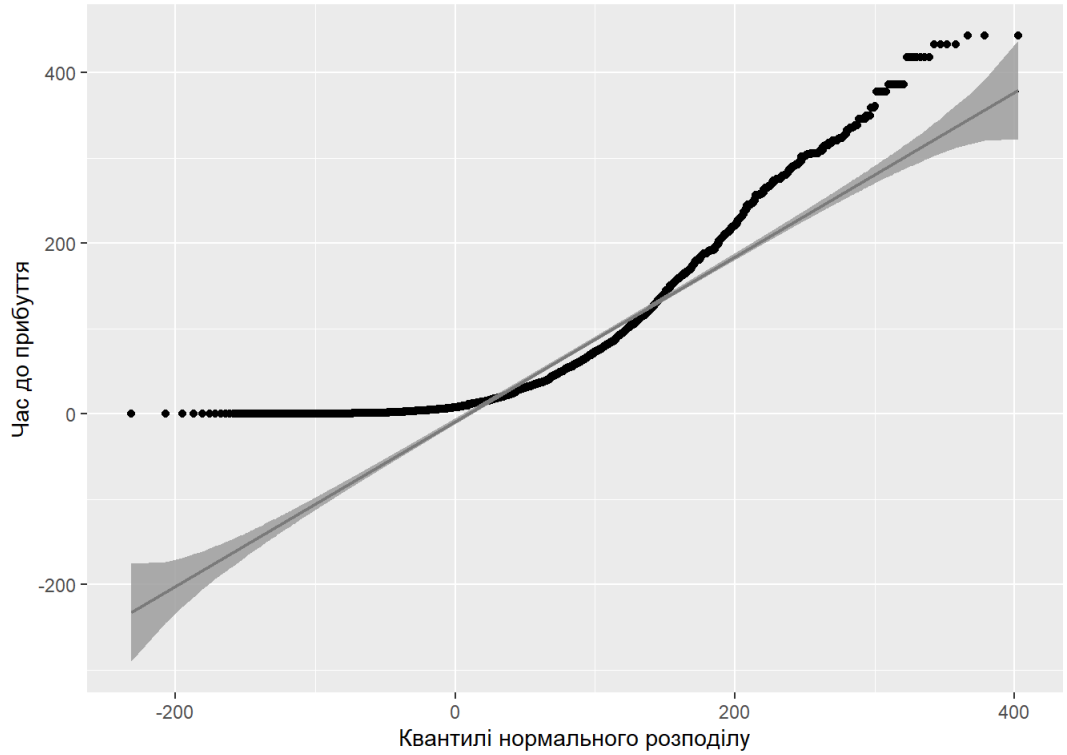
```
qqnorm(hotel$lead_time, main = 'Normal Q-Q Plot для часу до прибуття')
```



Порівняємо його з

нормальним розподілом. Бачимо, що розподіл часу не є нормальним

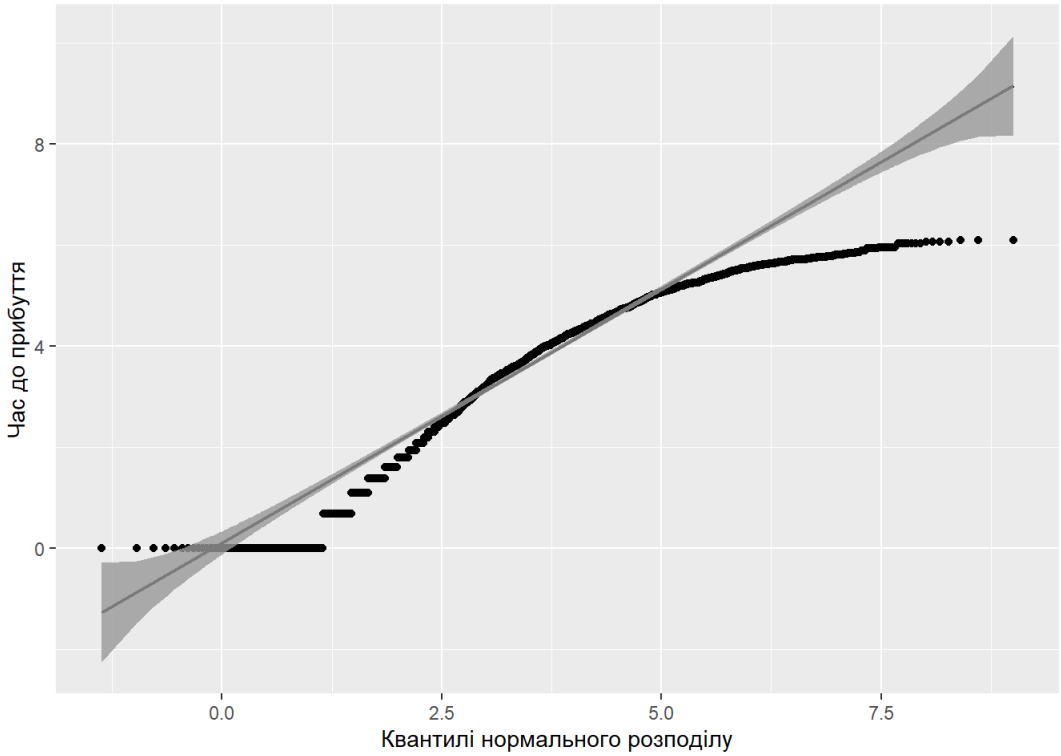
```
ggplot(hotel_sliced, aes(sample = lead_time)) +
  stat_qq_point() + stat_qq_line() + stat_qq_band() + labs (x = "Квантилі нормального розподілу", y = "Час до прибуття")
```



Логаритмування часу до

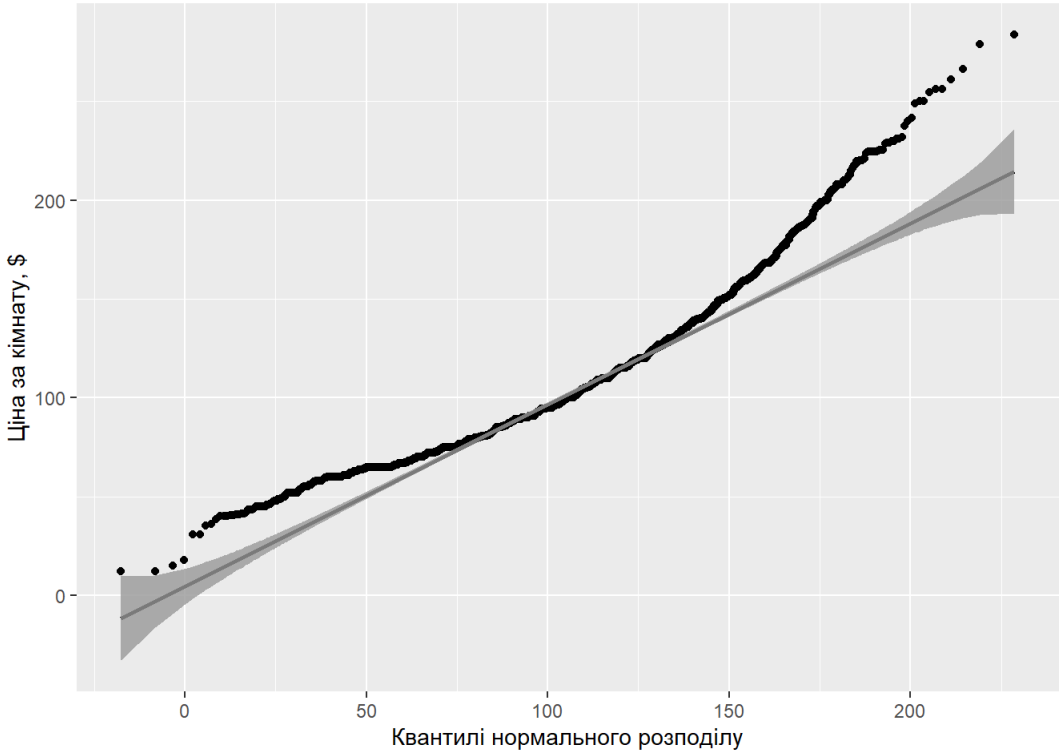
прибуття не сильно скрашує отриманий результат

```
ggplot(hotel_sliced, aes(sample = log(lead_time))) +
  stat_qq_point() + stat_qq_line() + stat_qq_band() + labs (x = "Квантилі нормального розподілу", y = "Час до прибуття")
```



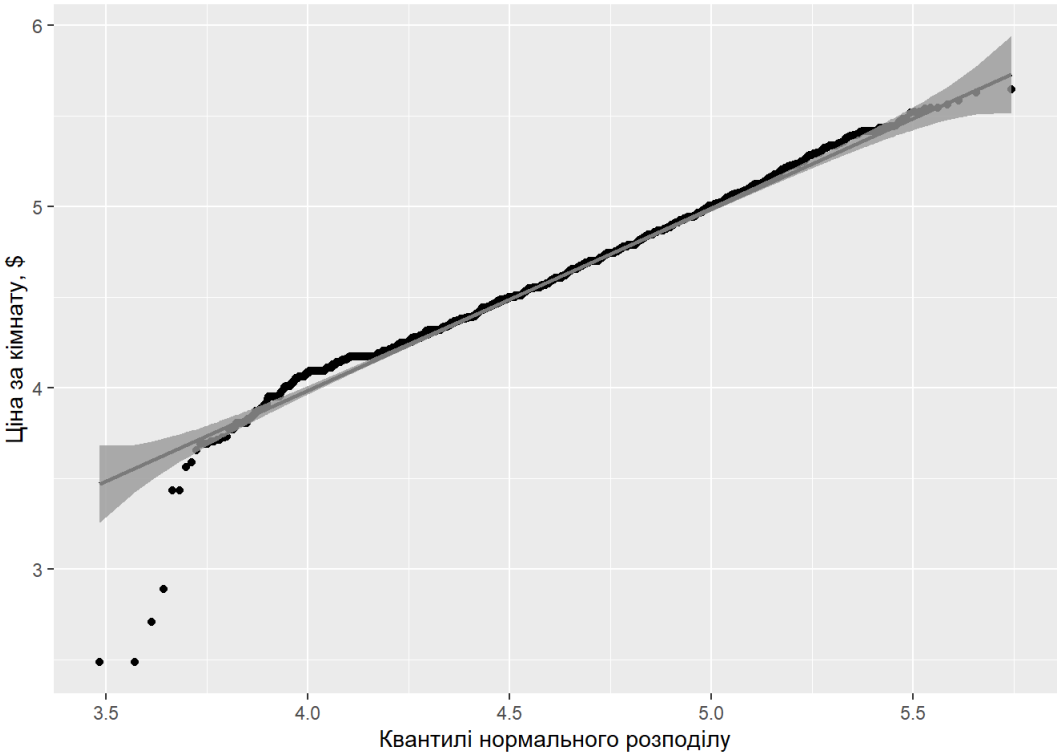
Побудуємо такий самий графік для ціни. Спостерігаємо викиди з дуже великою або низькою цінами, але, як було пояснено раніше, ці ціни цілком нормальні.

```
ggplot(hotel_sliced, aes(sample = avg_price_per_room)) +
  stat_qq_point() + stat_qq_line() + stat_qq_band() + labs (x = "Квантилі нормального розподілу", y = "Ціна за кімнату, $")
```



Якщо прологаритмуємо ціну - побачимо, що її розподіл стає досить схожим на нормальний, що є хорошим показником

```
ggplot(hotel_sliced, aes(sample = log(avg_price_per_room))) +
  stat_qq_point() + stat_qq_line() + stat_qq_band() + labs(x = "Квантилі нормального розподілу", y = "Ціна за кімнату, $")
```



Висновки

- Датасет був представлений в охайному вигляді, а знайдені викиди існують у незначній кількості та не мали впливу на подальші результати дослідження.

У підсумку команда дійшла до наступних висновків щодо дослідницьких питань:

- 1 - явний вплив типу кімнати на кількість дорослих/дітей - відсутній. Незалежно від кількості дорослих, найбільшу перевагу віддають типу харчування BB, двоє дорослих частіше не обирають тип харчування, ніж один або троє. Також видно, що зі збільшенням кількості дітей (від одного до трьох), зростає частка бронювань, де обрано тип харчування BB.
- 2 - зі збільшенням кількості особливих побажань, має тенденцію збільшуватися частка дорожчих типів кімнат. Можливо, такий результат можна пов'язаний з тим, що у людей які мають гроші на кращі апартаменти, є гроші і на певні додаткові особливі побажання.
- 3 - статус бронювання практично не змінюється від кількості дорослих. З кількістю дітей (до 3 включно) ситуація аналогічна. Випадки з кількістю дітей > 3 є поодинокими, через що на них не має сенсу загострювати увагу. Для найпопулярніших "типів" сімей ((1 дорослий без дітей), (2 дорослих без дітей), (2 дорослих з 1-2 дітьми) і (3 дорослих без дітей)) приблизно 25-35% записів скасовані, а у поодиноких випадках ситуація не така стабільна.
- 4 - ціна не має впливу на кількість проведених вихідних та робочих ночей у готелі, проте цей вплив мають тип забронюваної кімнати та кількість дітей: на різні терміни перебування (до 1 тижня, до 2 тижнів, більше ніж на 2 тижні) обирають 7, 6 та 1 типи кімнат, та заселяються з 1, 2 та 0 дітьми відповідно. Щільність ціни по діапазонам кількості заброньованих ночей візуально нагадує логнормальний розподіл.
- 5 - типи кімнат, які, незалежно від їх ціни, бронювали частіше (1, 4, 7), мають більшу кількість попередніх скасувань ніж інші 4 типи. Кімната другого типу може похизуватися кількістю попередніх НЕскасувань - 692 (що становить майже 100% випадків заселення)
- 6 - люди, яким необхідне паркувальне місце, рідше скасовують бронювання. Це питання потребує детальнішого вивчення: з одногу боку це досить природньо - якщо людина завчасно говорить за паркувальне місце, вона, напевно, вже достатньо спланувала свій візит до готелю, з іншого - можливо є інші фактори які на це впливають

Таким чином, деякі запитання пролили світло на ряд нових запитань і закономірностей, які потребують подальшої перевірки на спростування/підтвердження, а деякі - не мали помітних зв'язків між змінними і відповідно не дали очікуваних результатів.

З цього випливає необхідність проведення глибшого аналізу над цими даними.