НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
ІМЕНІ ІГОРЯ СІКОРСЬКОГО»
Факультет прикладної математики
Кафедра прикладної математики

Лабораторна робота №1

із дисципліни «Аналіз даних»

Розвідковий аналіз даних (EDA)

Виконали:

Долинний Денис (КМ-01)

Ганушке́вич Євгеній (КМ-02)

Рижко́ва Дар'я (КМ-02)

Гри́нів Юрій (КМ-02)

Голи́нський Денис (КМ-02)

Керівник:

Тавров Д.Ю.

Київ — 2023

Команда

Долинний Денис (КМ-01)
Ганушке́вич Євгеній (КМ-02)
Рижко́ва Дар'я (КМ-02)
Гри́нів Юрій (КМ-02)
Голи́нський Денис (КМ-02)

Про датасет

Датасет створений шляхом SQL-запитів до бази даних Hotel property management systems. Дані відображають готелі у Португалії за 2015-2017 роки. Датасет має 119390 спостережень і 32 змінні.

Питання для дослідження

1. З якої країни люблять подорожувати, і в який саме тип готелю?
2. Для яких країн які місяці туристичні (+ прибуток (adr) від місяцю)?
3. На які місяці вигідніше бронювати готелі?
4. Чи пов'язані тип відвідувачів (дорослі/(+діти)) з типом харчування?
5. Чи пов'язані тип відвідувачів (дорослі (+ діти)) з типом номеру?
6. Чи пов'язані тип відвідувачів (дорослі (+ діти)) з типом бронювання?
7. Залежність прибутку (adr) від різних факторів (booking_changes, total_of_special_requests, required_car_parking_spaces, lead_time).

Змінні

Факторні змінні
Числові змінні

hotel (chr) - тип готелю (H1 = Resort Hotel, H2 = City Hotel).
is_canceled (int) - показує чи була бронь скасована (1), чи ні (0).
lead_time (int) - кількість днів між бронюванням та прибуттям до готелю.
arrival_date_year - рік прибуття.
arrival_date_month(chr) - місяць прибуття.

arrival_date_week_number (int) - номер тижня прибуття .

arrival_date_day_of_month (int) - день прибуття.

stays_in_weekend_nights (int) - кількість вихідних (субота й неділя), які гість перебував або забронював у готелі.

stays_in_week_nights (int) - кількість будніх днів (понеділок - п'ятниця), які гість перебував або забронював у готелі.

adults (int) - кількість дорослих.

children (int) - кількість дітей.

babies (int)- кількість немовлят.

meal (chr) - тип замовленого харчування. Категорії: Undefined/SC – не подають харчування; BB – лише сніданок; HB – сніданок і ще один прийом їжі (зазвичай вечеря); FB – повне харчування (сніданок, обід і вечеря).
country (chr) - країна замовника.Закодовано в ISO 3155–3:2013 форматі.
market_segment (chr) - сегмент ринку. Категорії: “Aviation”, “Complementary”, “Corporate”, “Direct”, “Groups”, “Offline TA/TO”, “Online TA”, “Undefined” (“TA” – “Travel Agents”, “TO” – “Tour Operators”).
distribution_channel (chr)- розподіл бронювання. Категорії: “TA” – “Travel Agents”/“TO” means “Tour Operators”, “Corporate”, “Direct”, “GDS” – Global Distribution System.
is_repeated_guest (int) - бронювання від “старого” гостя (1) чи ні (0).
previous_cancellations (int) - кількість бронювань, які клієнт скасував до поточного бронювання.
previous_bookings_not_canceled (int) - кількість бронювань, які клієнт не скасував до поточного бронювання.
reserved_room_type (chr)- код номеру, який забронювали. Замість позначення наводиться код з міркувань анонімності.
assigned_room_type (chr)- код номеру, призначеного для бронювання. Іноді призначений тип номера відрізняється від типу зарезервованого номера через причини роботи готелю (наприклад, надмірне бронювання) або за запитом клієнта. Замість позначення наводиться код з міркувань анонімності.
booking_changes (int) - Кількість змін/доповнень, внесених до бронювання з моменту введення бронювання до моменту заселення або скасування замовлення.
deposit_type (chr) -зазначає чи вніс клієнт депозит, і якщо так – то який. Категорії: No Deposit, Non Refund, Refundable.
agent (chr)- ID туристичної агенції, що зробила замовлення.
company (chr)- ID компанії, що зробила замовлення або відповідальна за оплату.
days_in_waiting_list (int) - кількість днів, які бронювання було в списку очікування, перш ніж його було підтверджено
customer_type (chr) - тип бронювання. Категорії: Contract, Group, Transient, Transient-party.
adr (num) - середня добова ставка (статистична одиниця, яка показує дохід за номер за окремий період часу). Визначається як сума всіх операцій поділена на кількість ночей.
required_car_parking_spaces (int) - кількість паркувальних місць, які забронював клієнт.
total_of_special_requests (int)- кількість спеціальних запитів клієнта (наприклад двоспальне ліжко або високий поверх)
reservation_status (chr) - останній статус бронювання. Категорії: Canceled, Check-Ou, No-Show.
reservation_status_date (chr) - дата останнього оновлення статусу.

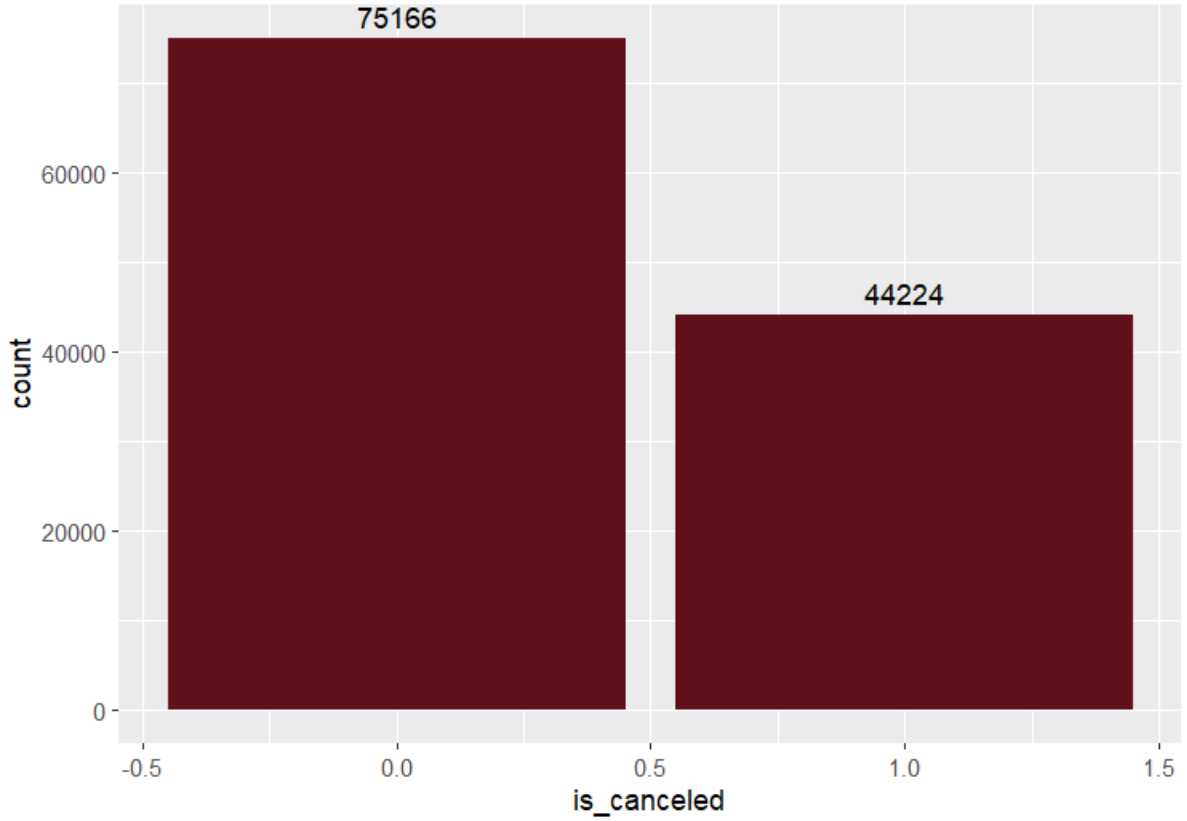
Перевірка даних

Лише змінна children має 4 значення NA.
Почнемо з факторних змінних:

1) hotel

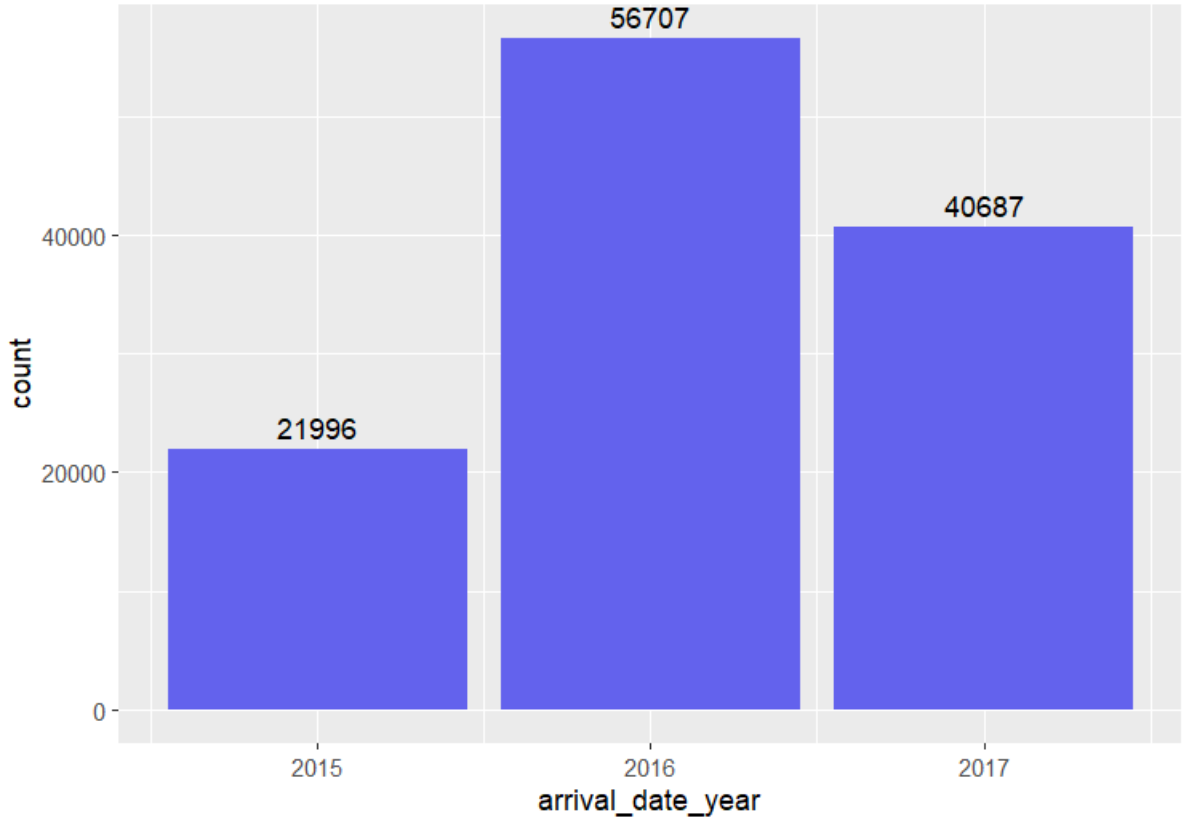
Як бачимо на графіку, кодування правильне, але датасет не є збалансованим. Отже будемо порівнювати не абсолютні, а відсоткові значення.

2) is_canceled:



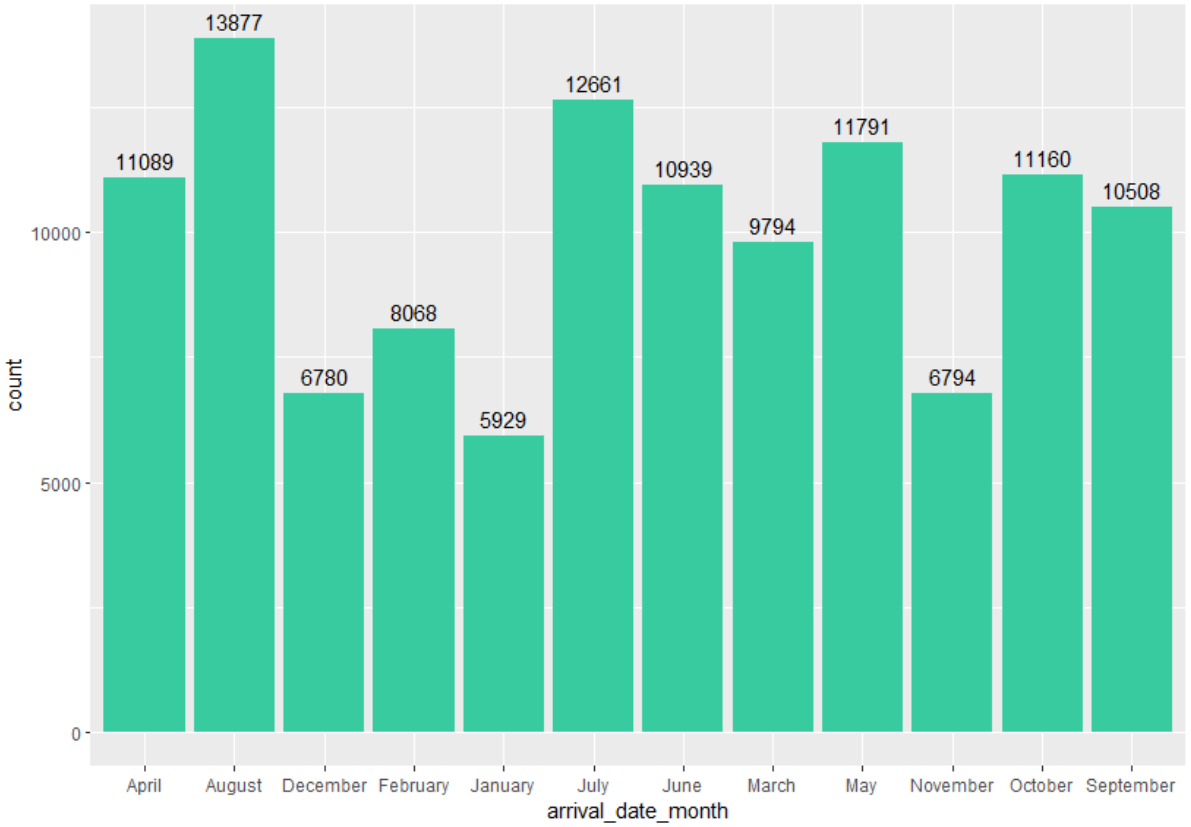
Немає значень, крім 0 і 1.

3) arrival_date_year:



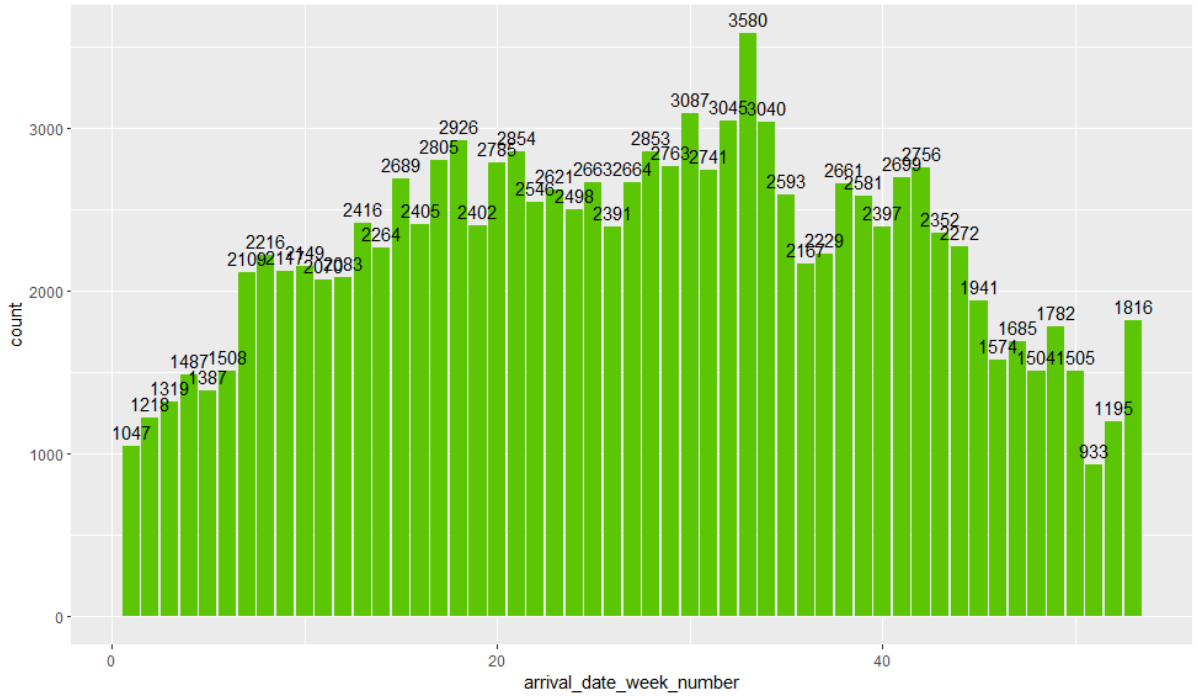
Немає значень, крім 2015, 2016, 2017

4) arrival_date_month:



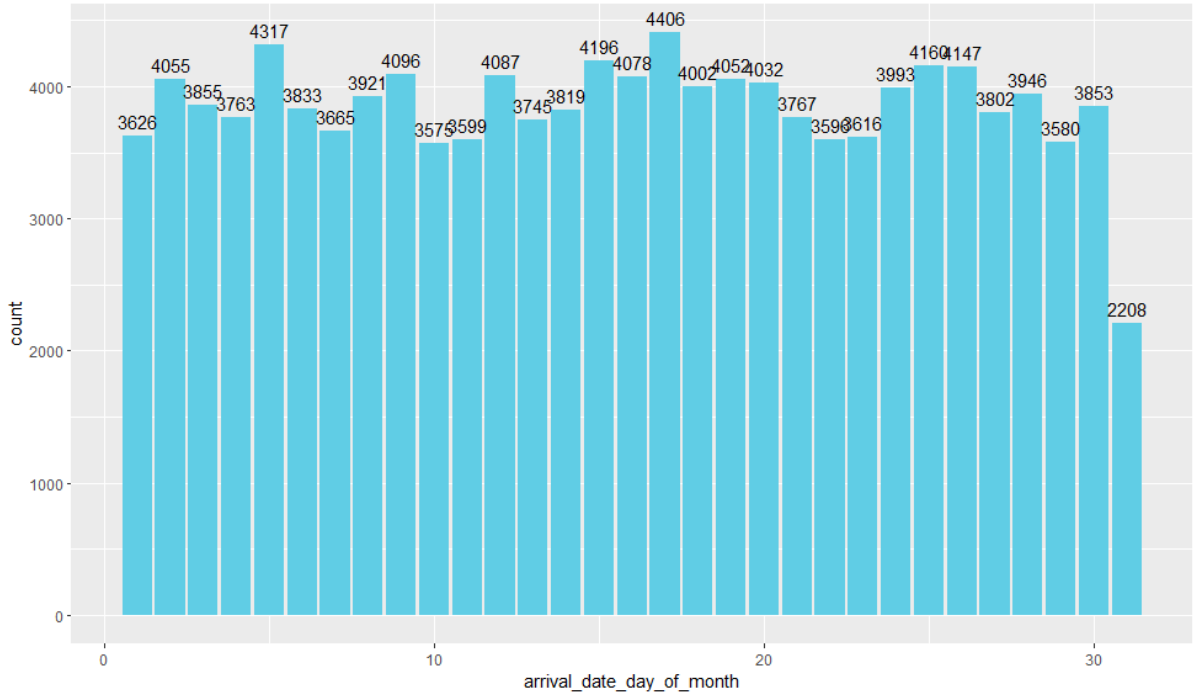
Немає недопустимих значень.

5) arrival_date_week_number:



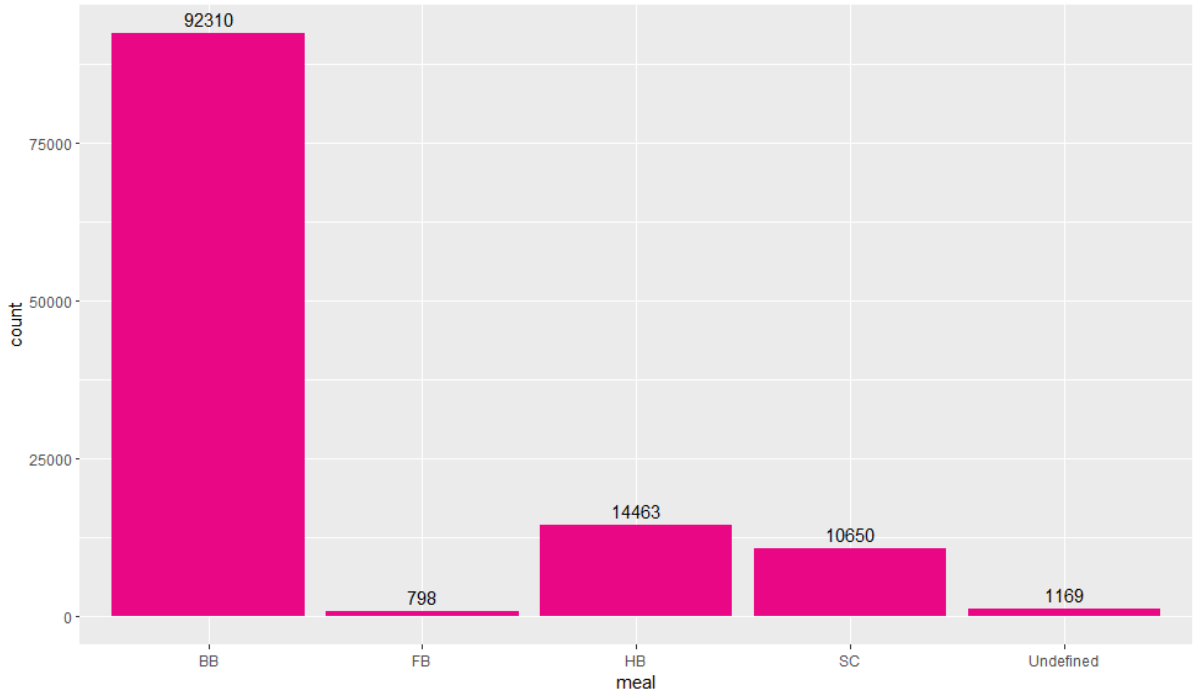
Значення в межах [1; 53].

6) arrival_date_day_of_month:



Значення в межах [1; 31]

7) meal:



Немає недопустимих значень.

8) country:

ABW	AGO	AIA	ALB	AND	ARE	ARG	ARM	ASM	ATA
2	362	1	12	7	51	214	8	1	2
ATF	AUS	AUT	AZE	BDI	BEL	BEN	BFA	BGD	BGR
1	426	1263	17	1	2342	3	1	12	75
BHR	BHS	BIH	BLR	BOL	BRA	BRB	BWA	CAF	CHE
5	1	13	26	10	2224	4	1	5	1730
CHL	CHN	CIV	CMR	CN	COL	COM	CPV	CRI	CUB
65	999	6	10	1279	71	2	24	19	8
CYM	CYP	CZE	DEU	DJI	DMA	DNK	DOM	DZA	ECU
1	51	171	7287	1	1	435	14	103	27
EGY	ESP	EST	ETH	FIN	FJI	FRA	FRO	GAB	GBR
32	8568	83	3	447	1	10415	5	4	12129
GEO	GGY	GHA	GIB	GLP	GNB	GRC	GTM	GUY	HKG
22	3	4	18	2	9	128	4	1	29
HND	HRV	HUN	IDN	IMN	IND	IRL	IRN	IRQ	ISL
1	100	230	35	2	152	3375	83	14	57
ISR	ITA	JAM	JEY	JOR	JPN	KAZ	KEN	KHM	KIR
669	3766	6	8	21	197	19	6	2	1
KNA	KOR	KWT	LAO	LBN	LBY	LCA	LIE	LKA	LTU
2	133	16	2	31	8	1	3	7	81
LUX	LVA	MAC	MAR	MCO	MDG	MDV	MEX	MKD	MLI
287	55	16	259	4	1	12	85	10	1
MLT	MMR	MNE	MOZ	MRT	MUS	MWI	MYS	MYT	NAM
18	1	5	67	1	7	2	28	2	1
NCL	NGA	NIC	NLD	NOR	NPL	NULL	NZL	OMN	PAK
1	34	1	2104	607	1	488	74	18	14
PAN	PER	PHL	PLW	POL	PRI	PRT	PRY	PYF	QAT
9	29	40	1	919	12	48590	4	1	15
ROU	RUS	RWA	SAU	SDN	SEN	SGP	SLE	SLV	SMR
500	632	2	48	1	11	39	1	2	1
SRB	STP	SUR	SVK	SVN	SWE	SYC	SYR	TGO	THA
101	2	5	65	57	1024	2	3	2	59
TJK	TMP	TUN	TUR	TWN	TZA	UGA	UKR	UMI	URY
9	3	39	248	51	5	2	68	1	32
USA	UZB	VEN	VGB	VNM	ZAF	ZMB	ZWE		
2097	4	26	1	8	80	2	4		

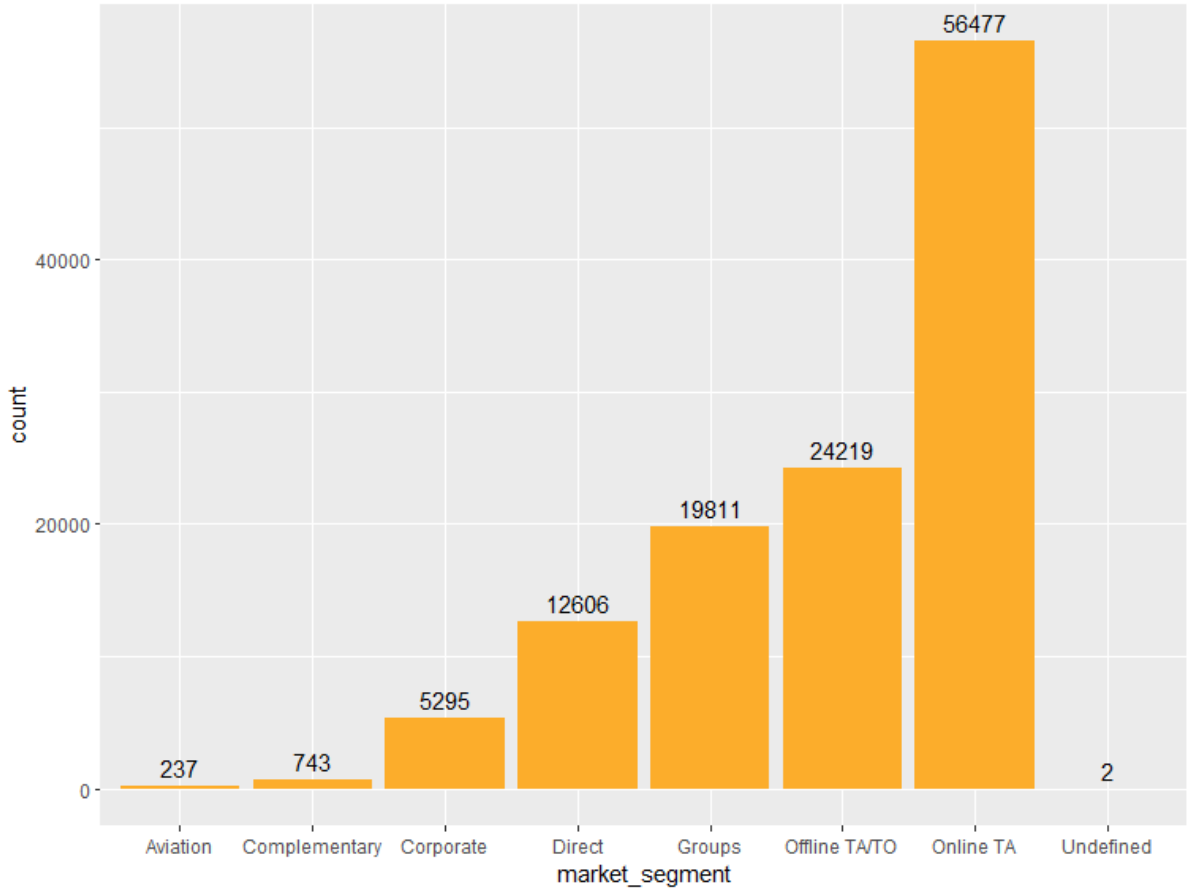
З’являється значення NULL.

Топ-10 країн:

1	PRT	19691
2	GBR	9599
3	FRA	8427
4	ESP	6311
5	DEU	6028
6	IRL	2537
7	ITA	2416
8	BEL	1862
9	NLD	1712
10	USA	1585

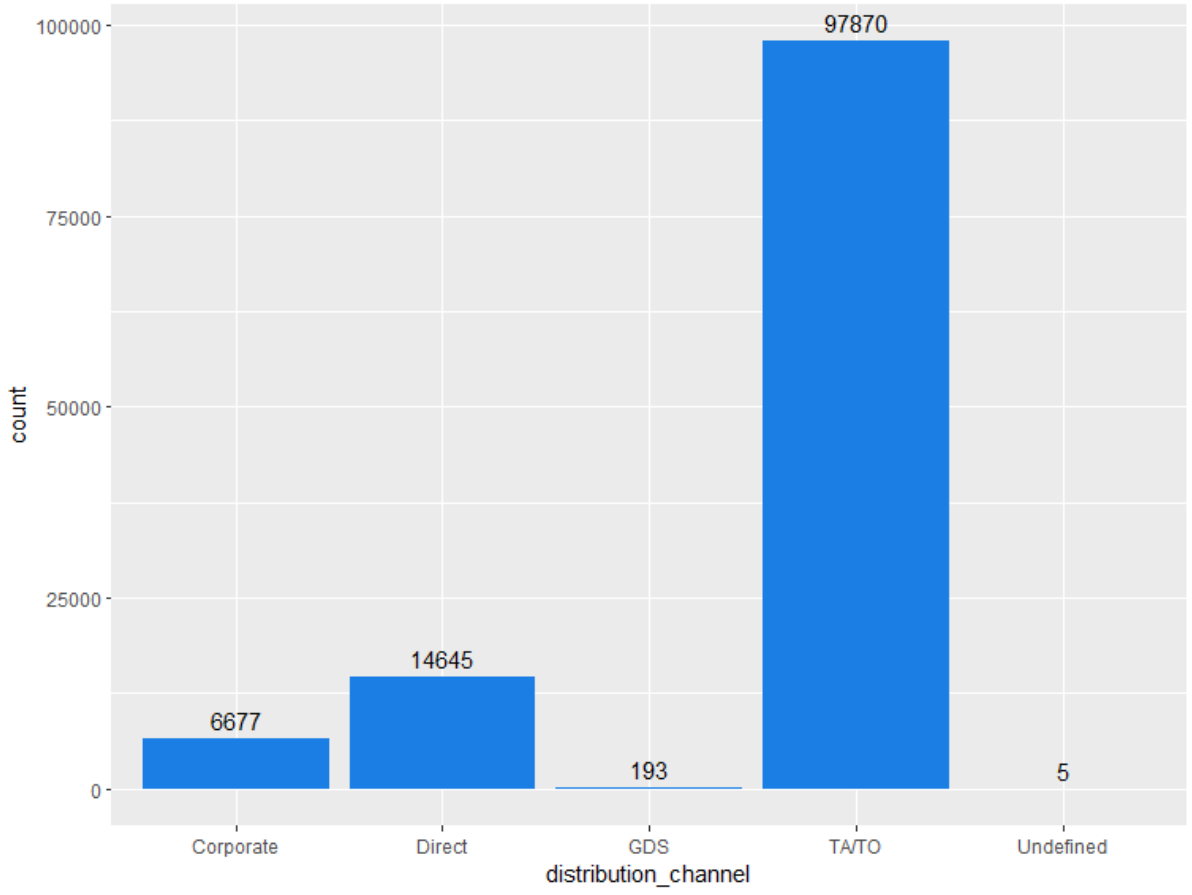
Як бачимо Португалія має суттєву перевагу над іншими країнами, що буде у певній мірі спотворювати результати.

9) market_segment:



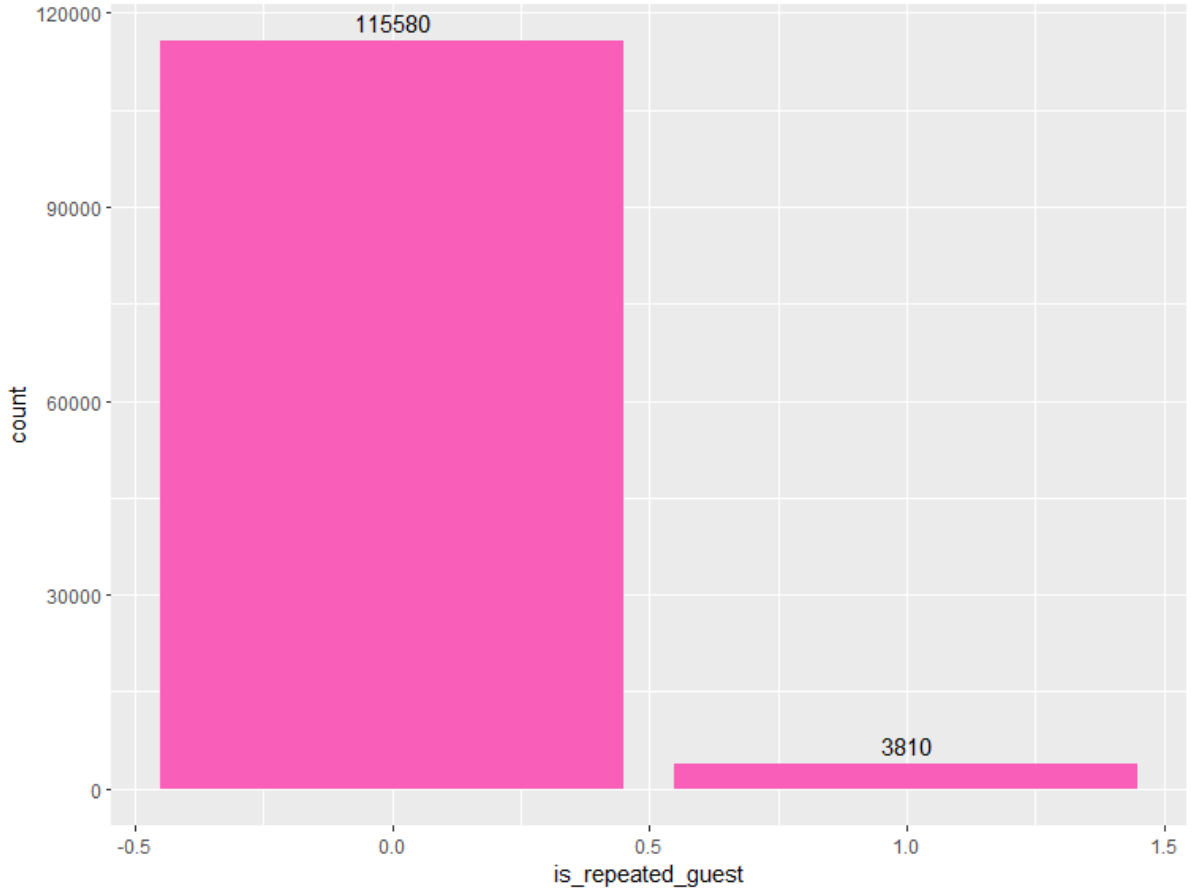
Немає недопустимих значень.

10) distribution_channel:



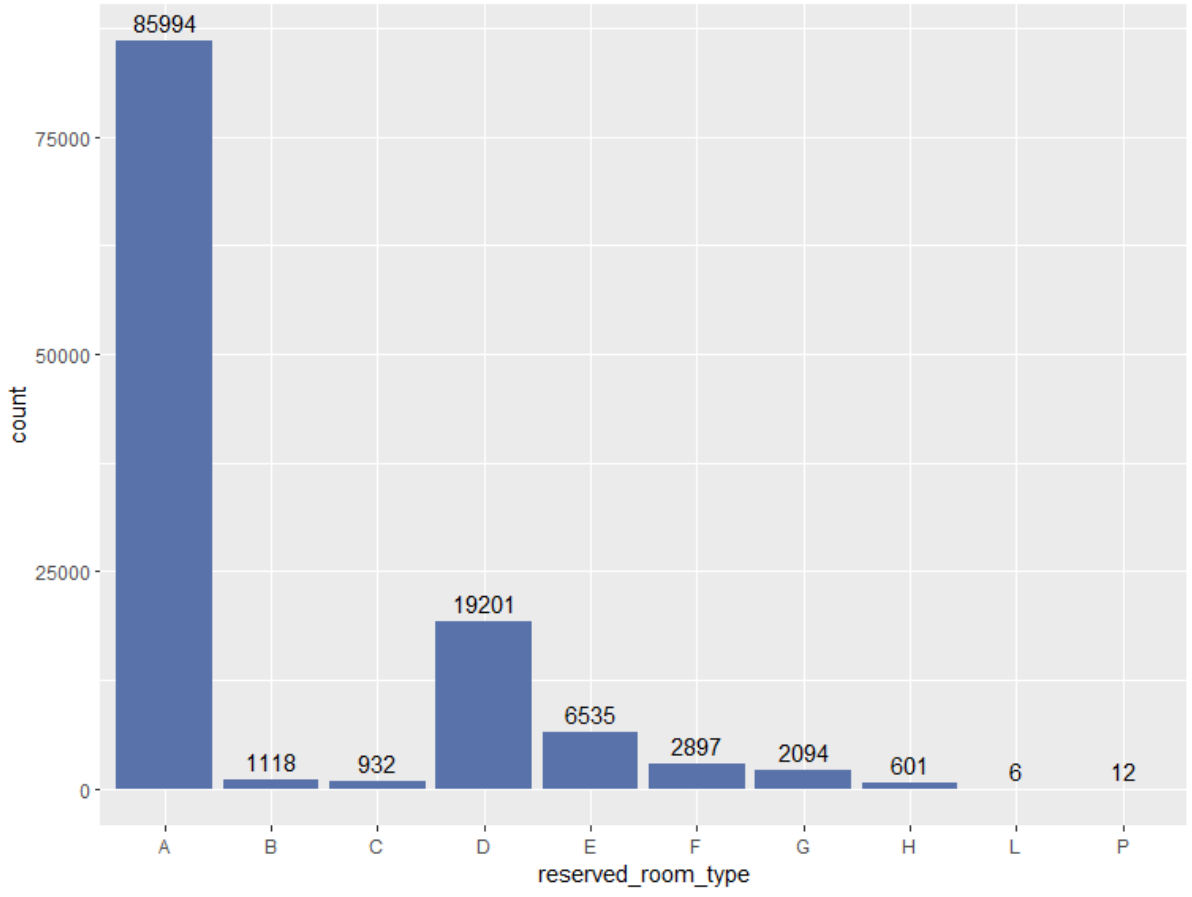
Немає недопустимих значень.

11) is_repeated_guest:



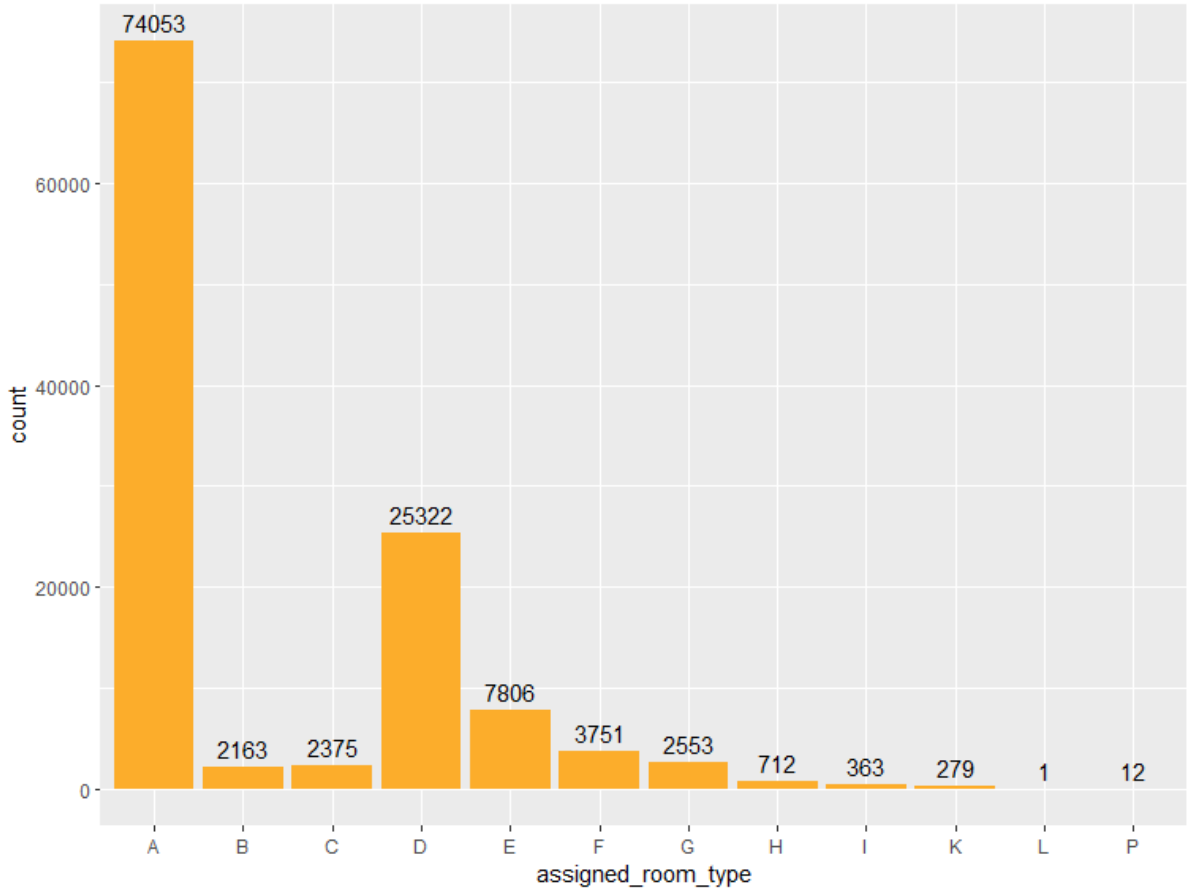
Немає значень, крім 0 і 1.

12) reserved_room_type:



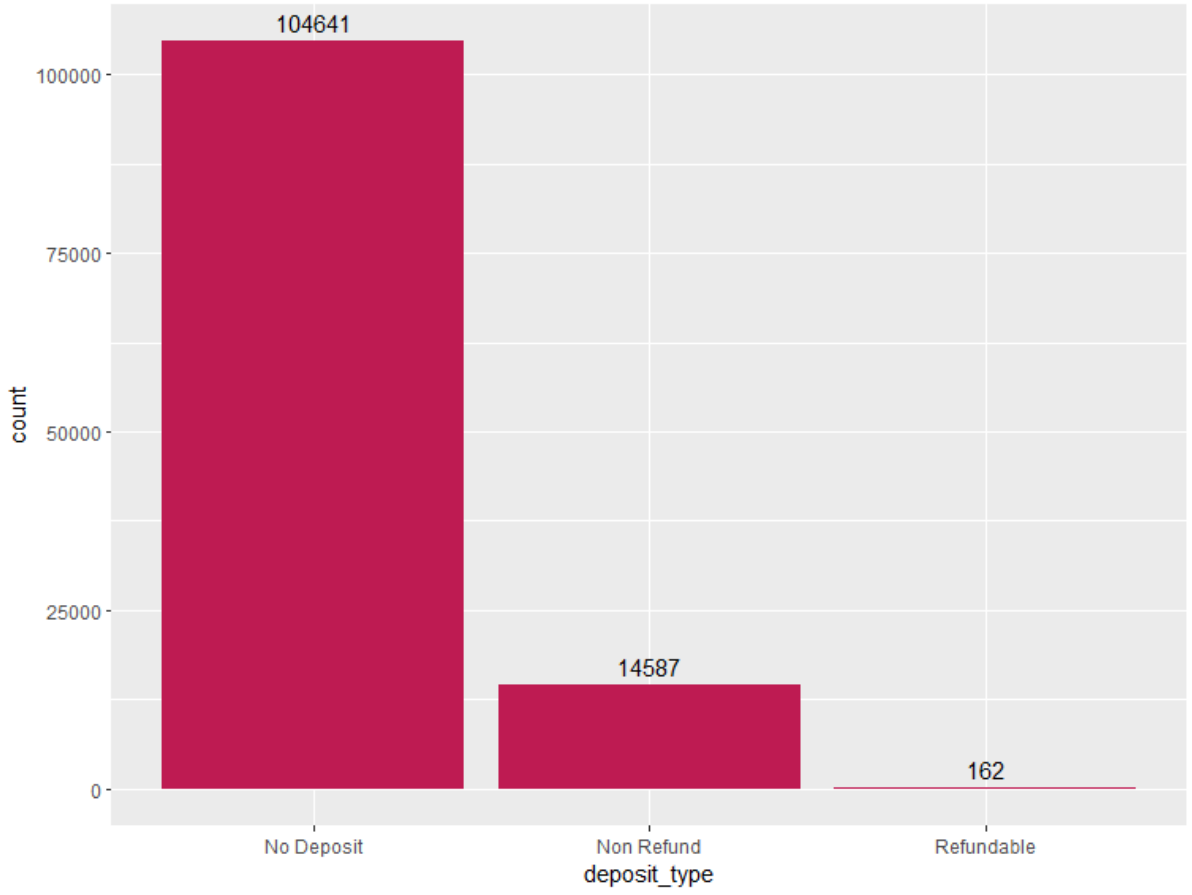
Немає недопустимих значень.

13) assigned_room_type:



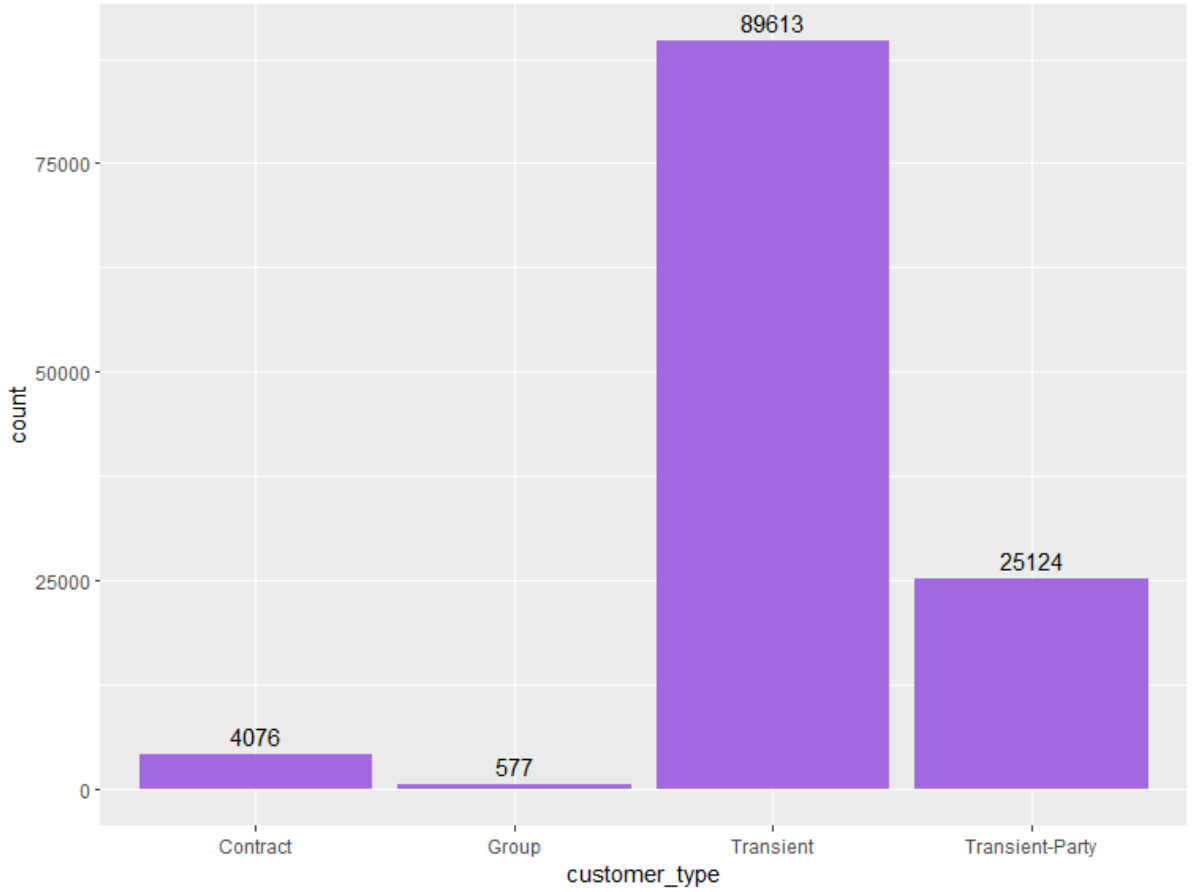
Немає недопустимих значень, але з’явилися нові значення порівняно з попередніми

14) deposit_type:



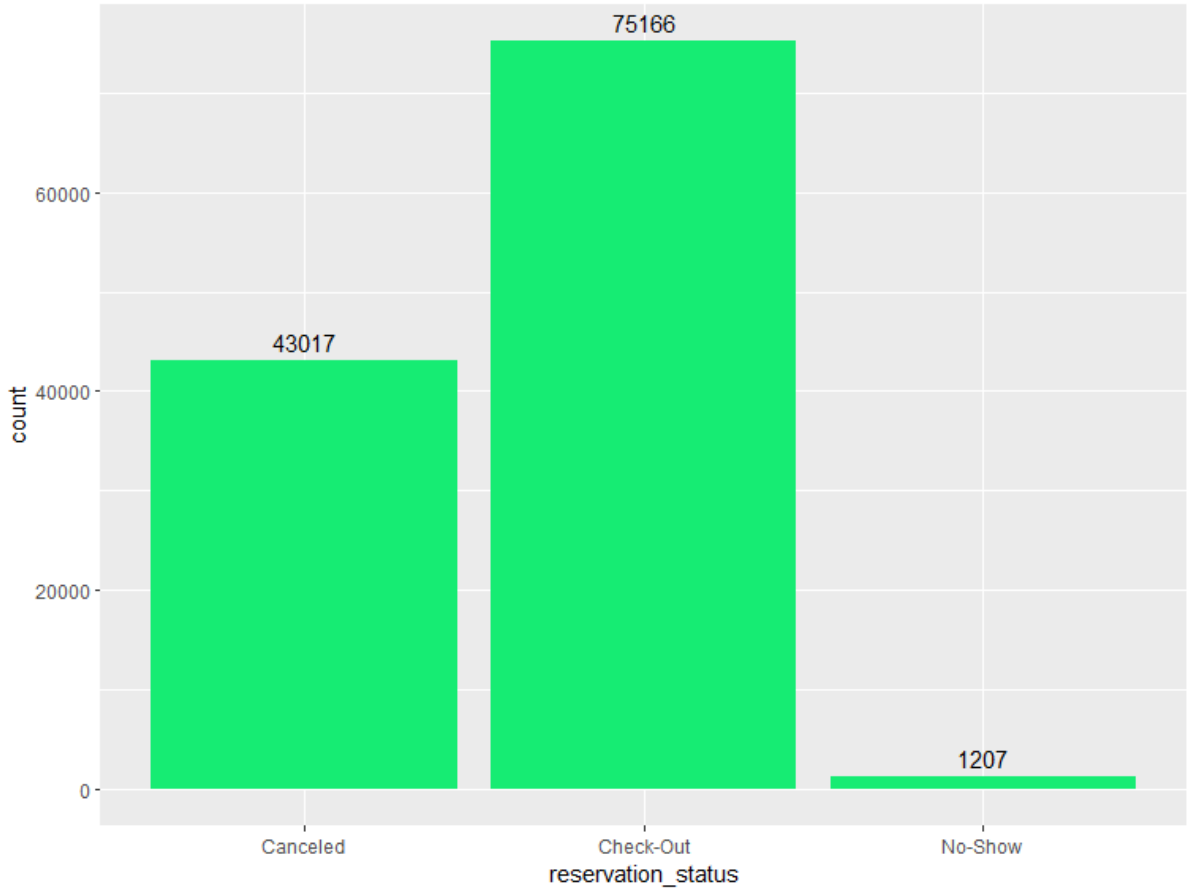
Немає недопустимих значень.

15) customer_type:



Немає недопустимих значень.

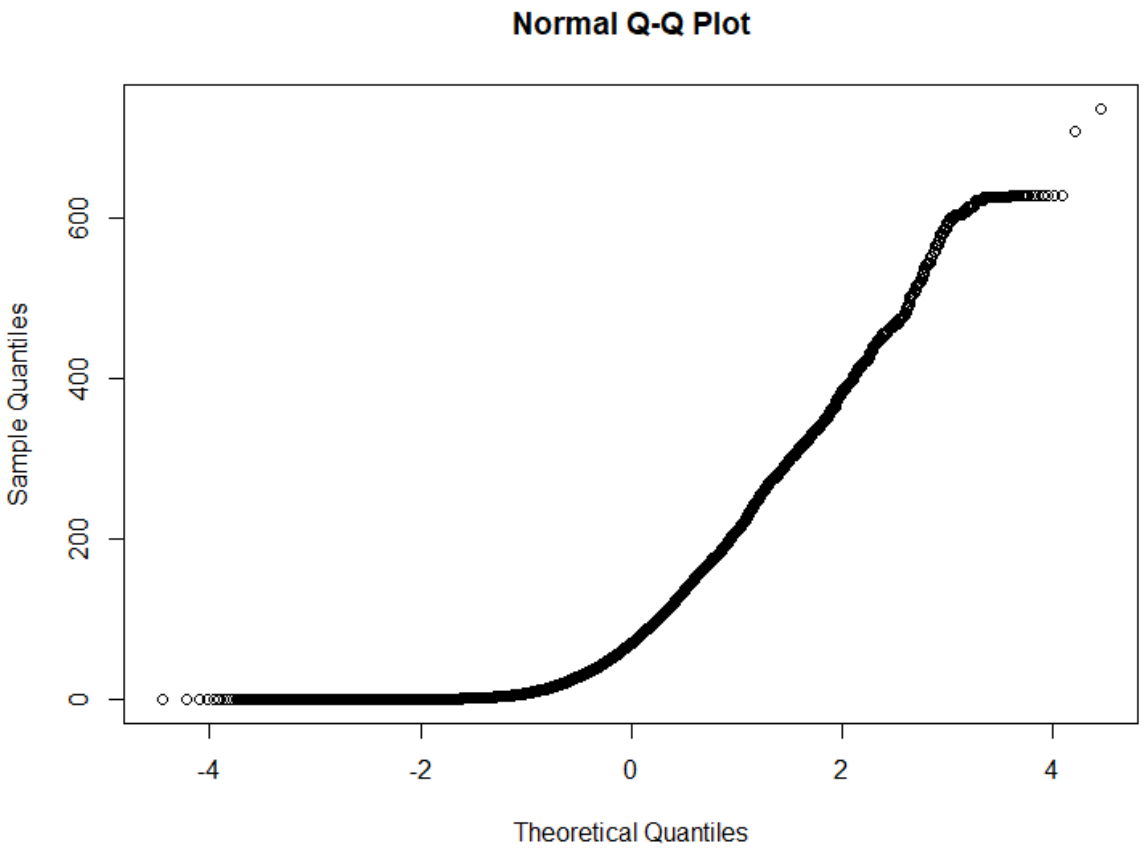
16) reservation_status:



Немає недопустимих значень.

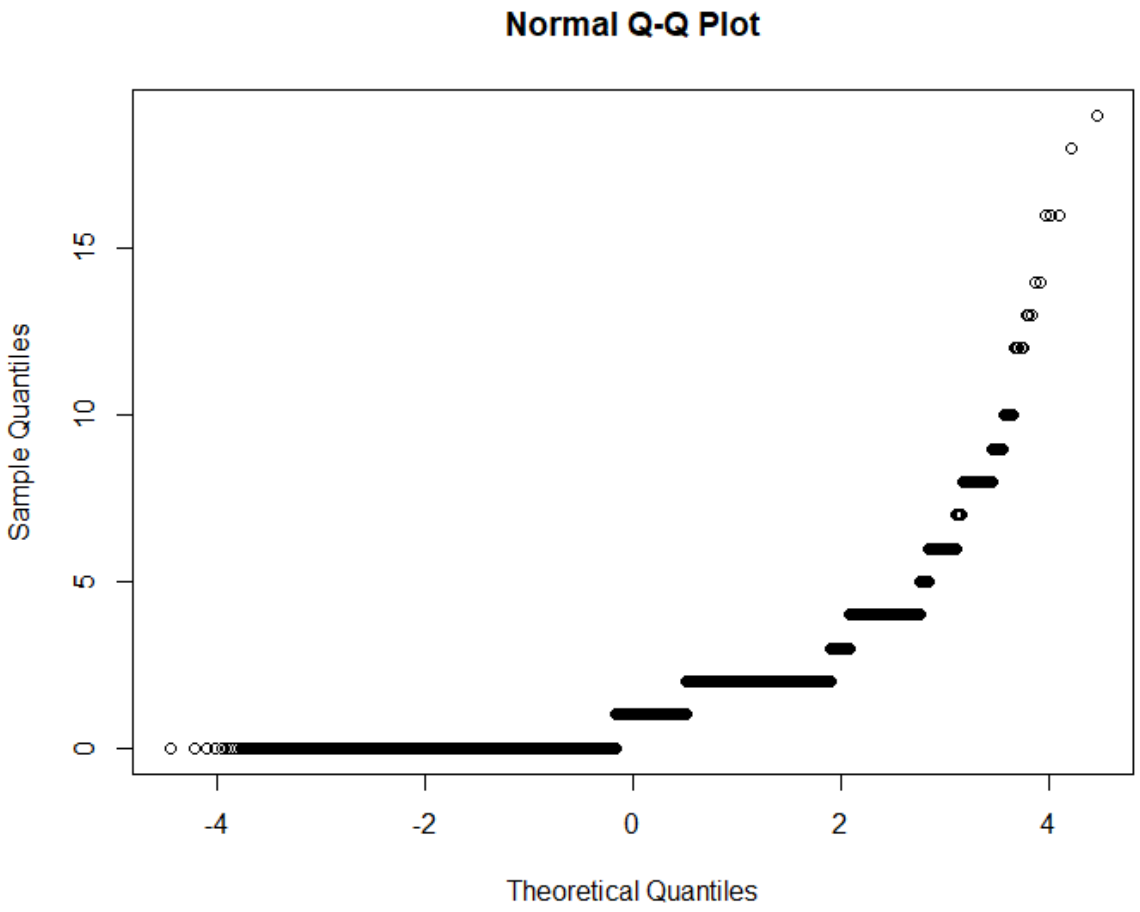
Числові змінні:

1) lead_time:



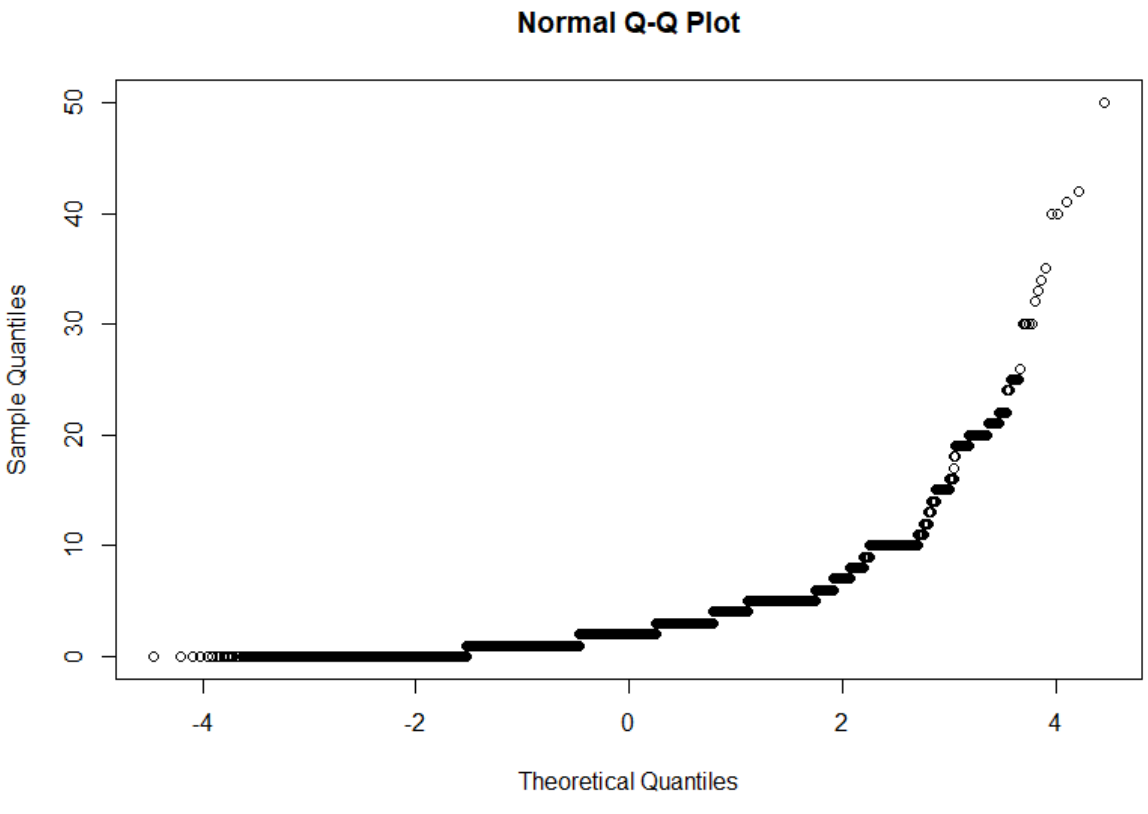
Бачимо дивні викиди, які значно більші за 600, та значення, які рівні нулю. Значень, які дорівнюють 0 більше 6000, отже, не можемо стверджувати, що це помилка. Значення, які вибиваються (> 700), але ніяких інших дивних значень у цих спостереженнях немає.

2) stays_in_weekend_nights:



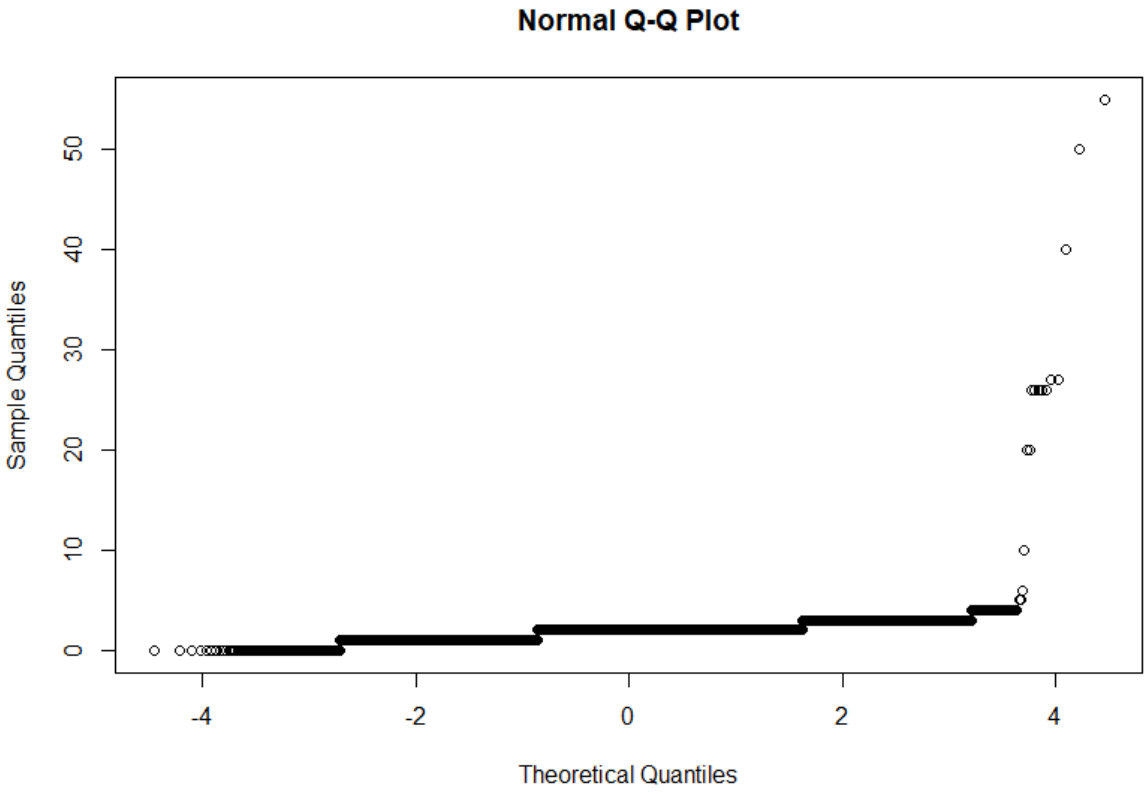
У цій змінній викидів значно більше. Значень більше 10 – 15.

3) stays_in_week_nights:



Значень більше 30 – 9.

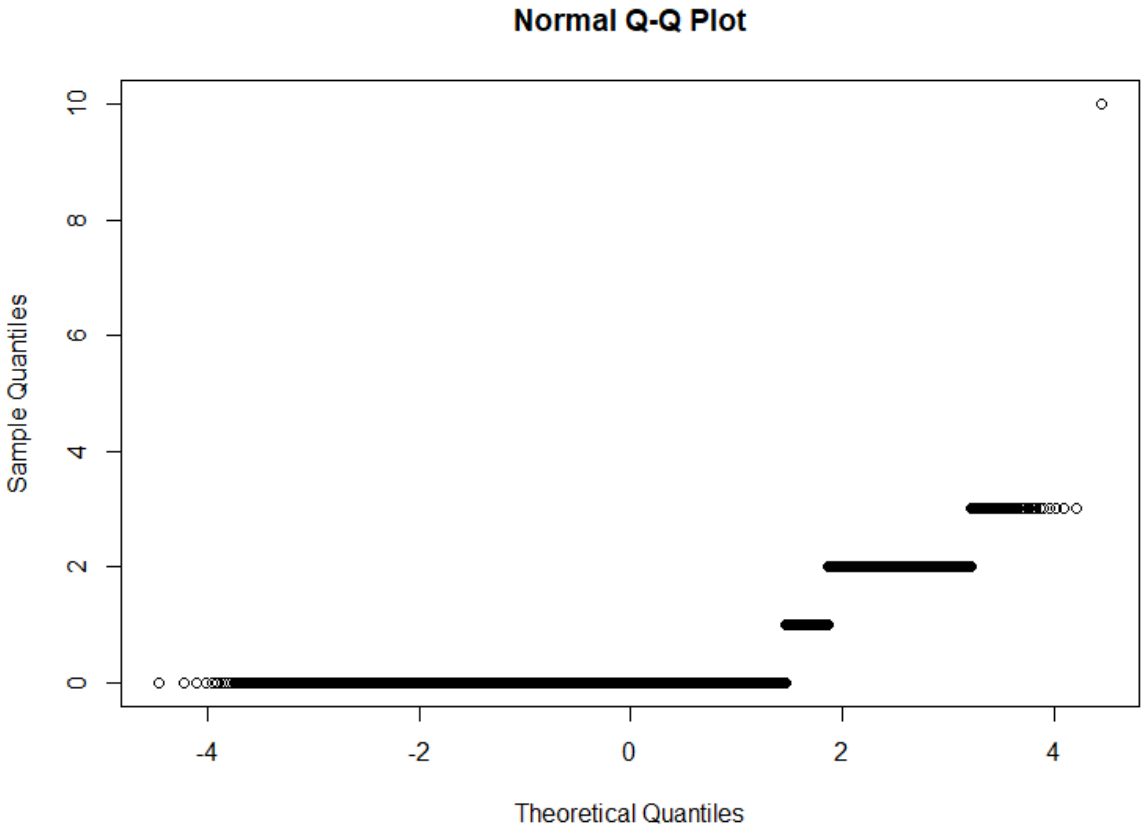
4) adults:



Є значення, які рівні 0, що маловірогідно. При детальному розгляді бачимо, що є 180 спостережень, де змінні adults, children і babies одночасно рівні 0. Також є 16 спостережень, де adults > 4, які будемо вважати викидами.

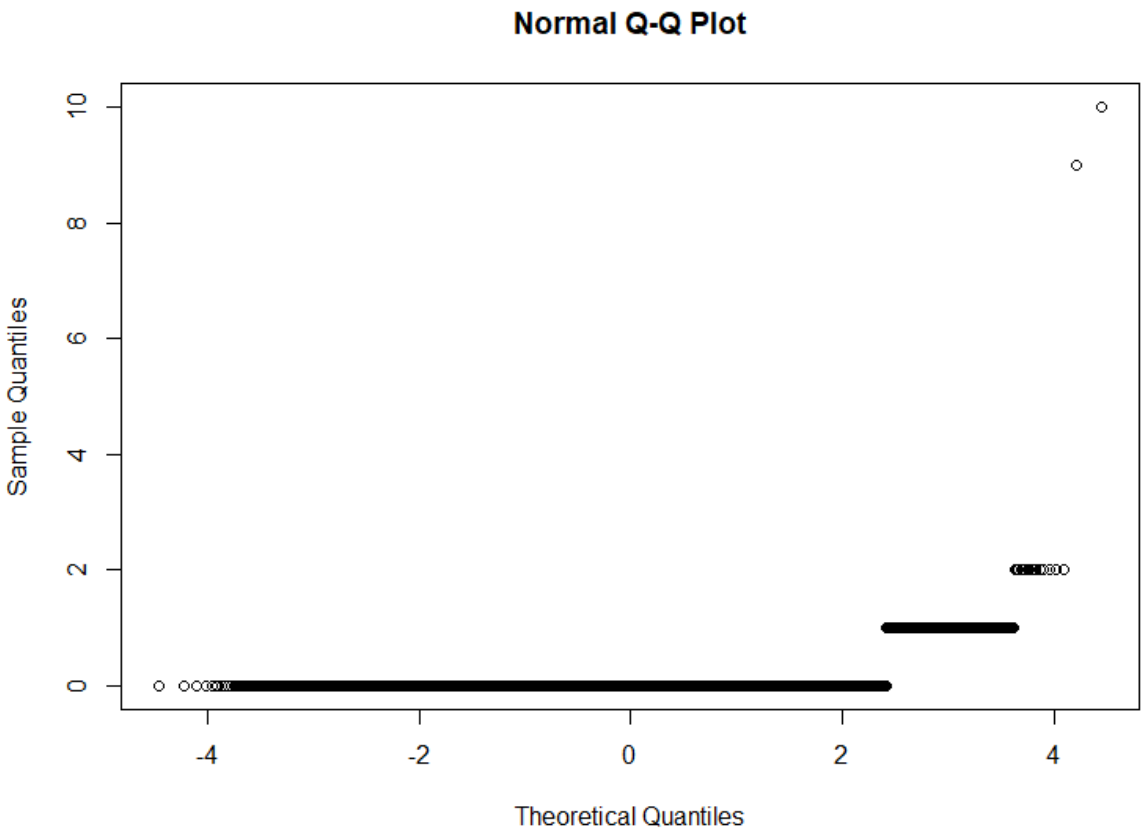
5) children:

Спочатку замінимо значення NA у children на 0.



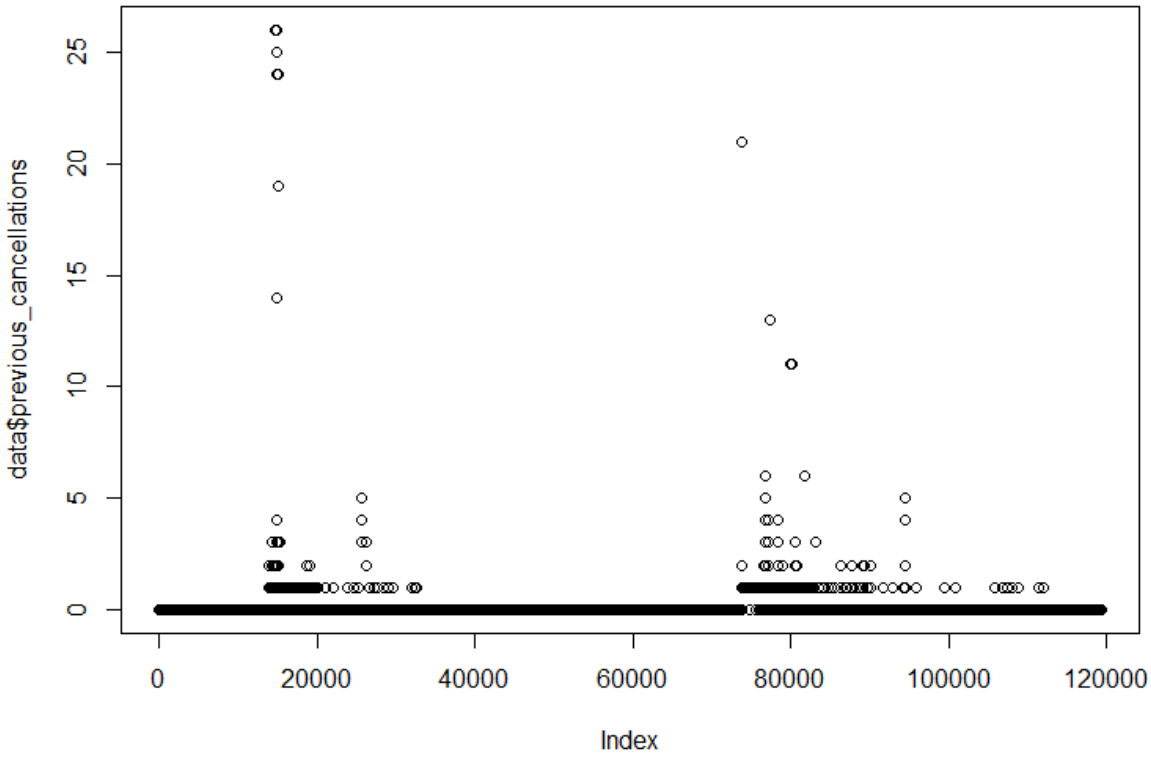
Бачимо 1 викид.

6) babies:



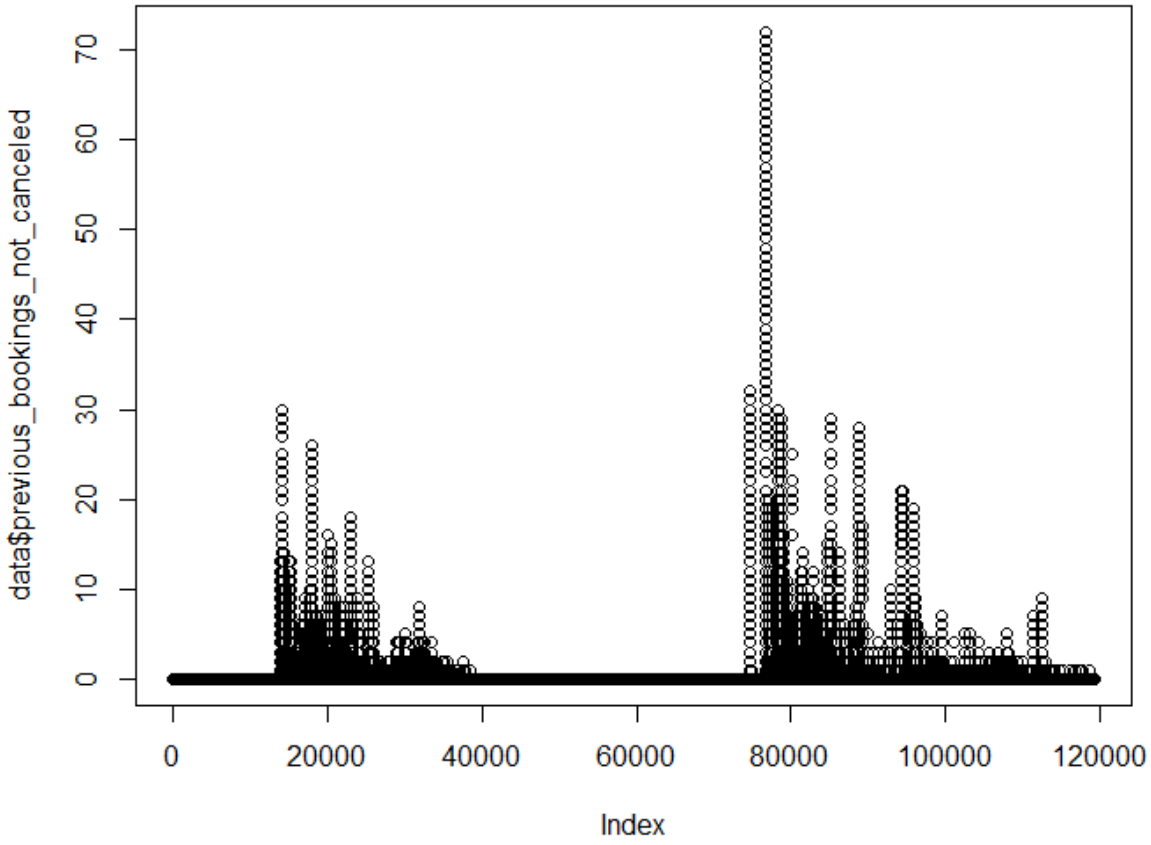
Бачимо 2 викиди.

7) previous_cancellations:



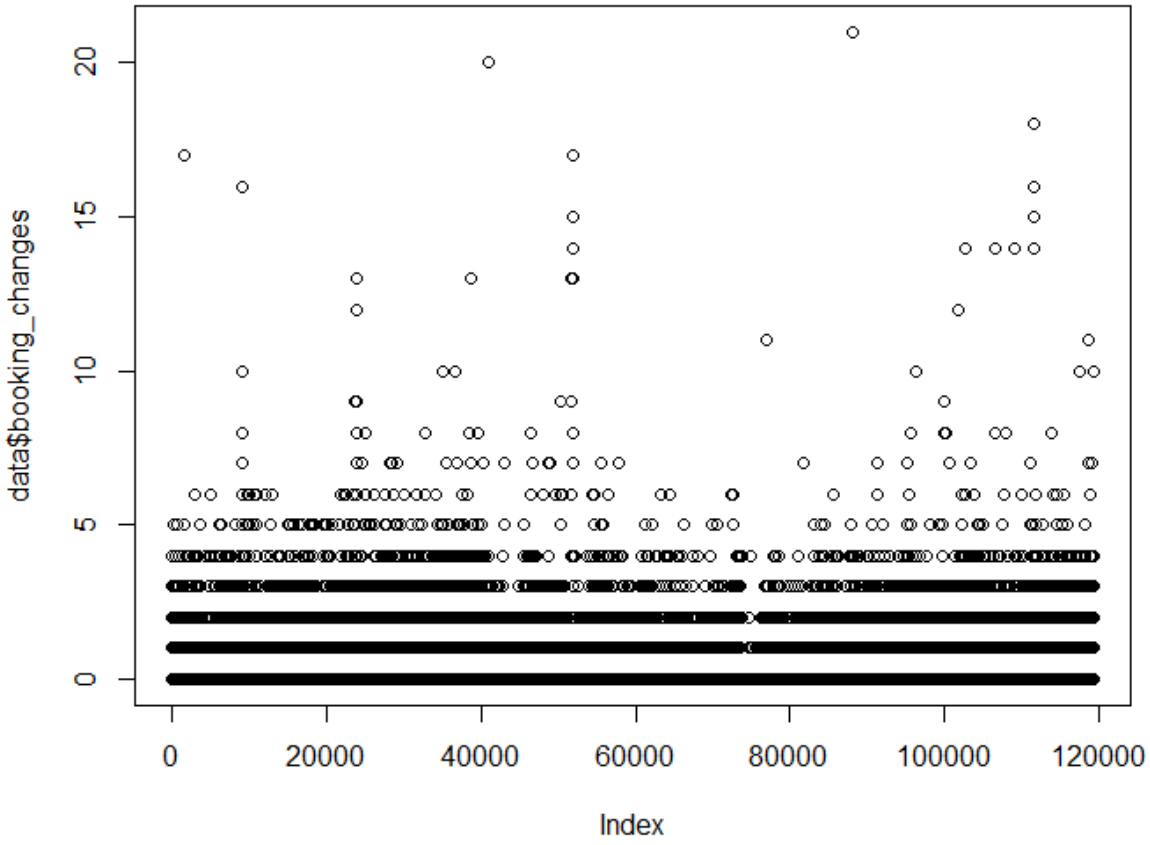
Бачимо 180 значень, які більше 10

8) previous_bookings_not_canceled:



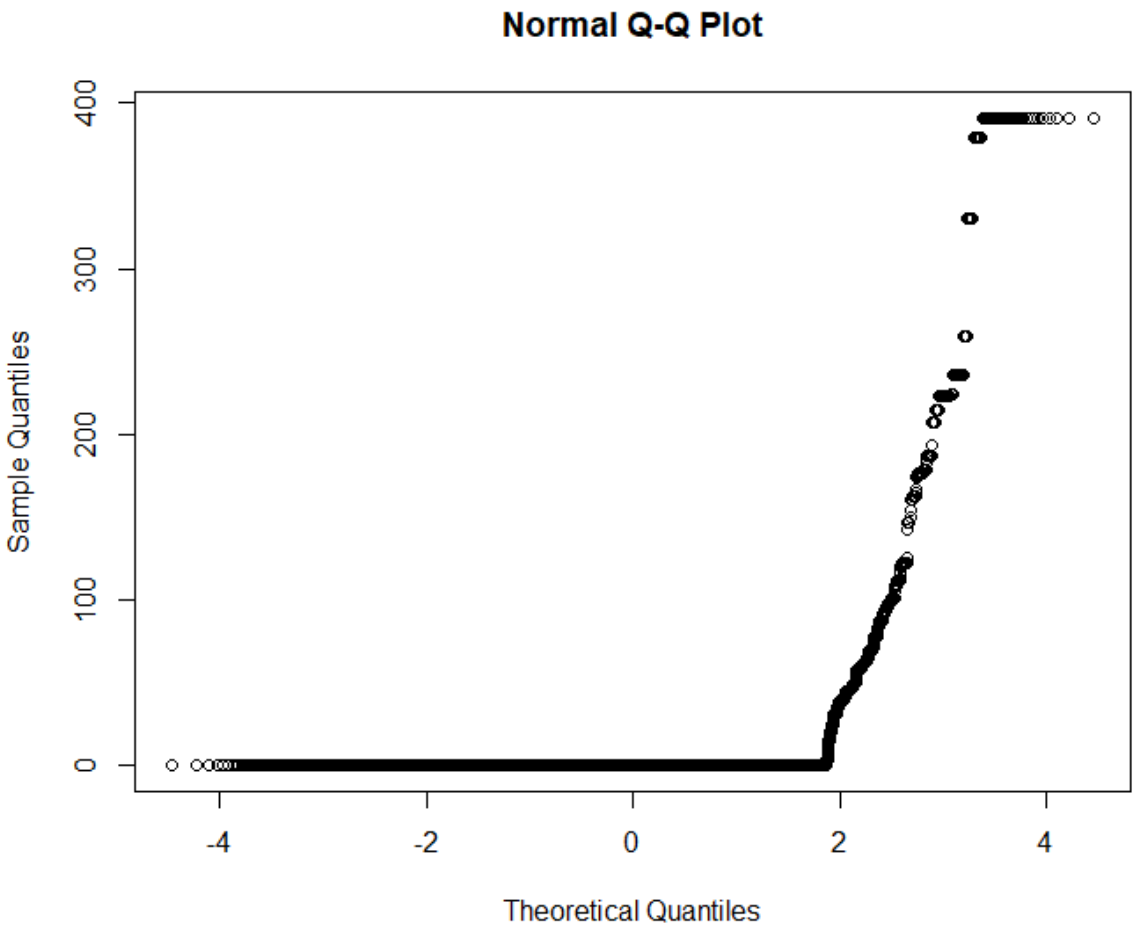
Бачимо 135 значень, які більше 20

9) booking_changes:



Бачимо 149 значень, які більше 5.

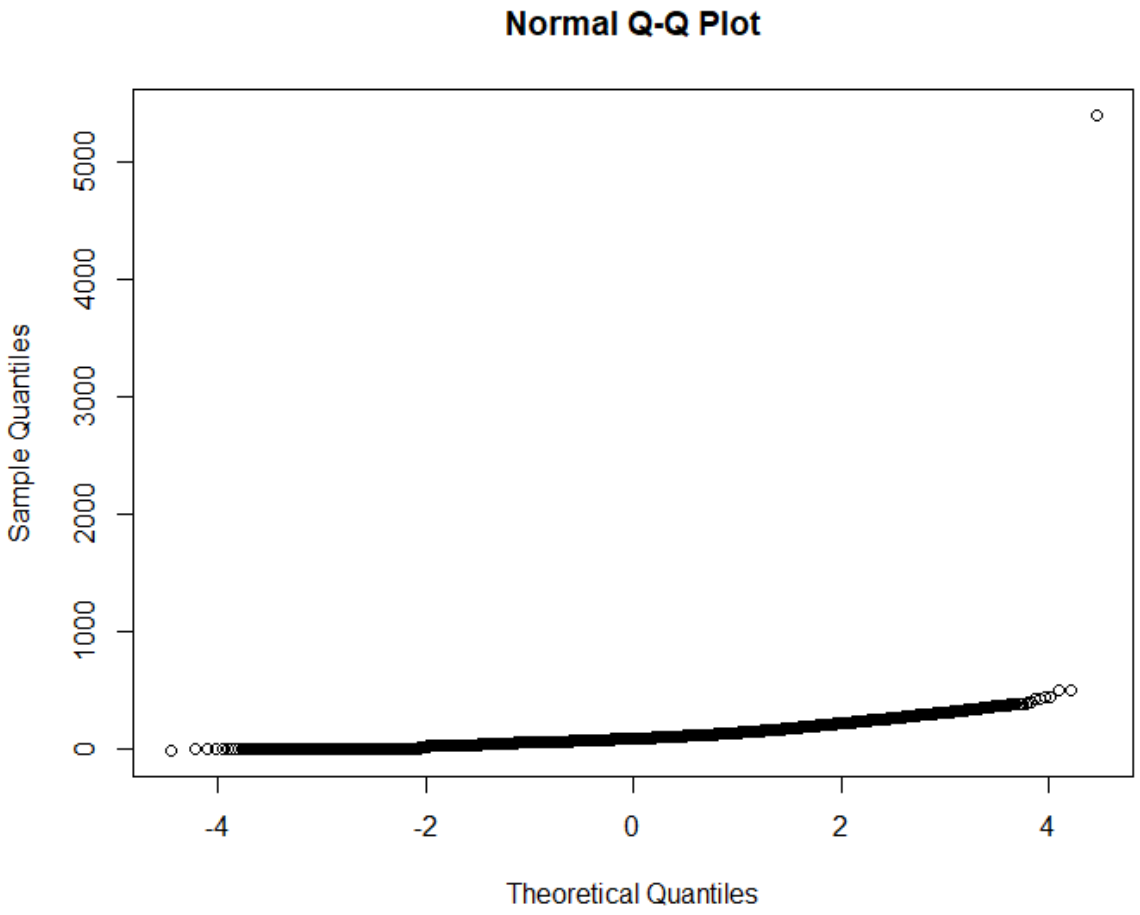
10) days_in_waiting_list:



adr

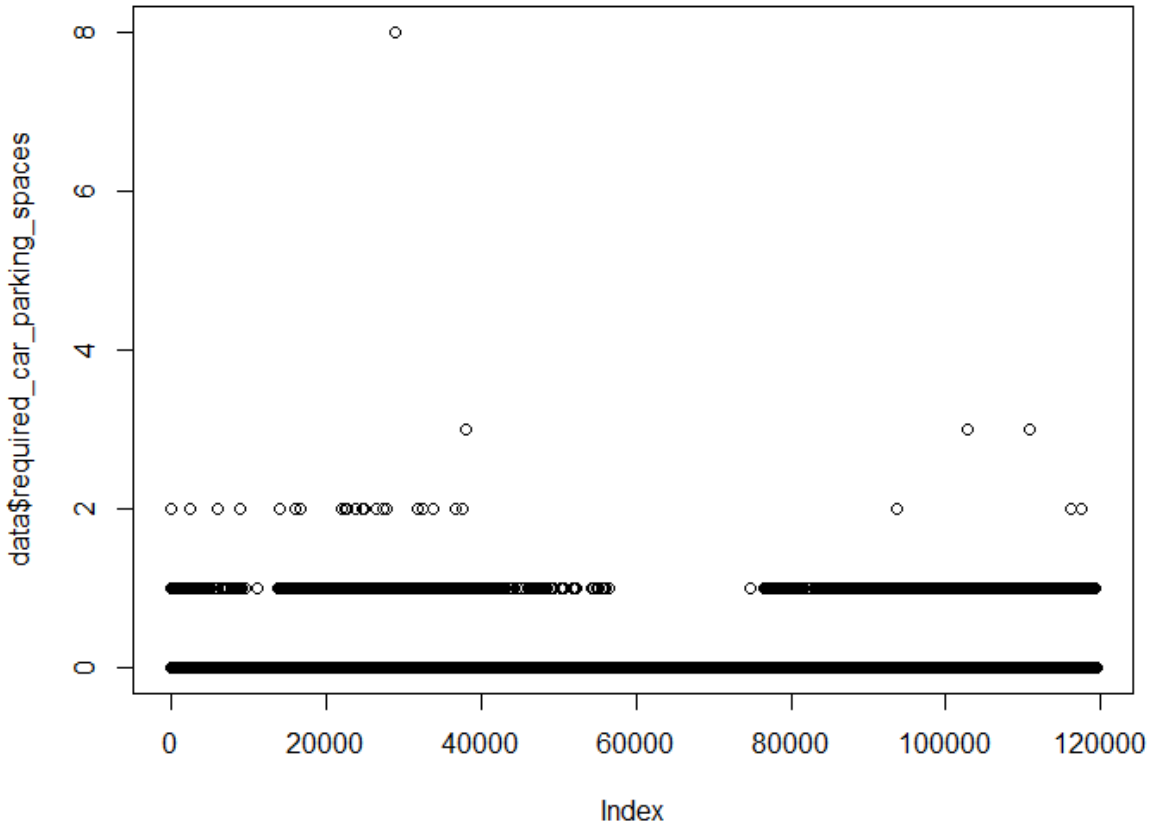
Бачимо 227 значень, які більше 200.

11) adr:



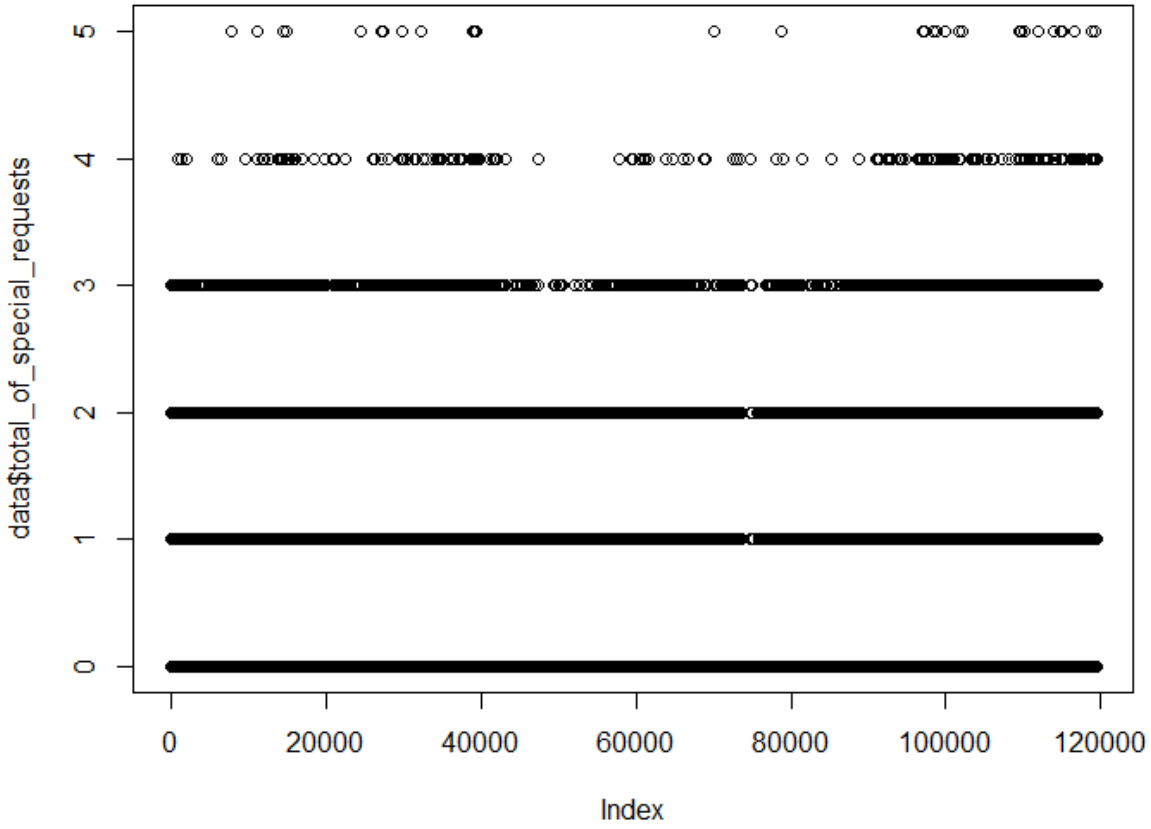
Бачимо 1 значення менше 0 та 1 значення більше 550 (5400).

12) required_car_parking_spaces:



Бачимо 5 значень, які більше 2. У них бачимо, що максимальна кількість людей 2. Тобто очевидна помилка.

13) total_of_special_requests:



Бачимо 380 значень, які більше 3.

Інші змінні:

1) agent: значеннями є цілі числа, факторна змінна. Є 16340 значень NULL.
2) company: значеннями є цілі числа, факторна змінна. Є 112593 значень NULL.
3) reservation_status_date: значеннями є рядки у форматі “РР-ММ-ДД”. Регулярні вирази показують, що всі дані правильні

Дескриптивні характеристики:

	Min.	1st Qu.	Median	Mean	Mean 3rd Qu.	Max.
lead_time	0	18	69	104	160	737
stays_in_weekend_nights	0	0	1	0,9276	2	19
stays_in_week_nights	0	1	2	2,5	3	50
adults	0	2	2	1,856	2	55
children	0	0	0	0,1039	0	10
babies	0	0	0	0,007949	0	10
previous_cancellations	0	0	0	0,08712	0	26
previous_bookings_not_canceled	0	0	0	0,1371	0	72
booking_changes	0	0	0	0,2211	0	21
days_in_waiting_list	0	0	0	2,321	0	391
adr	-6,38	69,29	94,58	101,83	126	5400

required_car_parking_spaces	0	0	0	0,06252	0	8
total_of_special_requests	0	0	0	0,5714	1	5

Очистка даних

Створимо нову змінну (all_guests), що включає всіх гостей.
Видалимо 180 рядків, де змінні adults, children і babies одночасно рівні 0 (тобто залишаться спостереження, де all_guests != 0).
Видалемо змінні company та agent, бо вони не несуть ніякої корисної інформації.
Додамо змінну stays_in_nights, яка показуватиме загальну кількість ночей. Бачимо 645 рядків, де stays_in_nights = 0 (тобто stays_in_weekend_nights і stays_in_week_nights одночасно дорівнюють 0). Видаляємо ці рядки.
Додамо змінну all_children, яка показуватиме загальну кількість дітей.
У змінній required_car_parking_spaces є 5 значень, які більше 2. У них бачимо, що максимальна кількість людей 2. Тобто очевидна помилка. Замінимо їх на 2.
У змінній adr є 1 значення, що менше 0, та 1 значення, що більше 550 (5400). Видалемо їх, оскільки їх мало. Також видалимо ті, що більше 400 (їх 7).
У змінних babies та children сумарно 3 викиди, які ми видаляємо.Також видаляємо 16 рядків, де adults > 4.
З lead_time видаляємо значення, які більше 700.
Видаляємо спостереження, де country має значення NULL.
Усі інші змінні мають не такі явні викиди (тобто ми не можемо сказати це помилки чи ні). Отже, доцільно розглядати випадки з ними і без них, і подивитися, який вони мали вплив.
Після очистки дескриптивні характеристики мають вигляд:

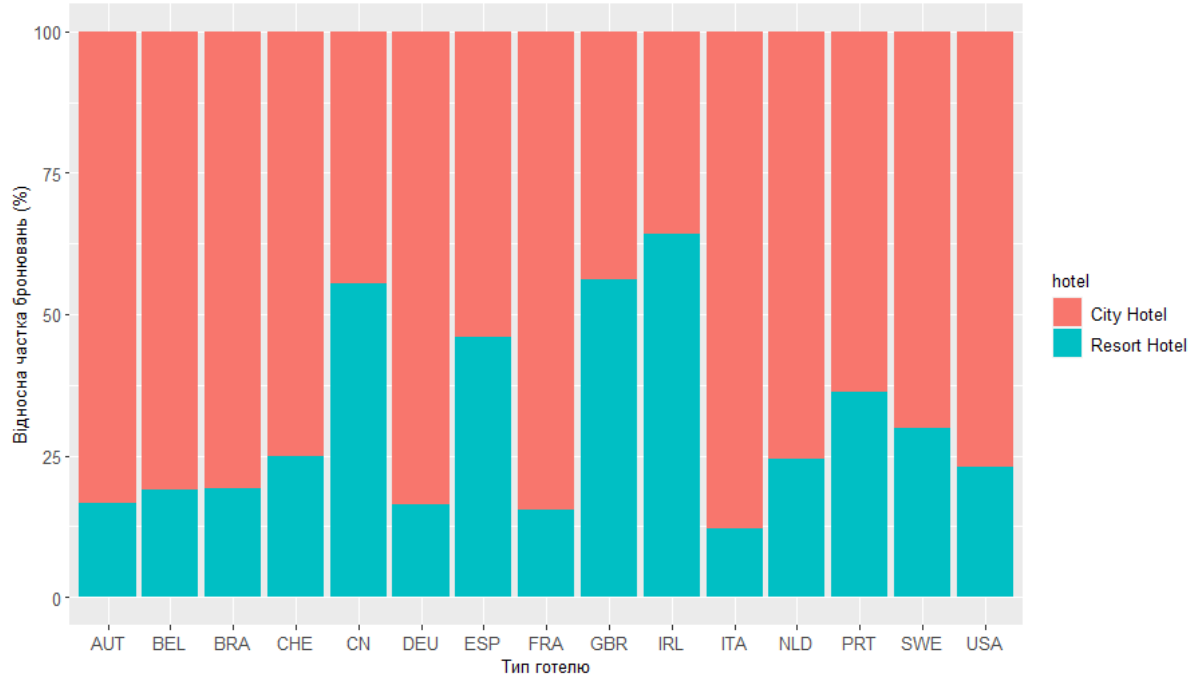
	Min.	1st Qu.	Median	Mean	Mean 3rd Qu.	Max.
lead_time	0	19	71	105,4	162	629
stays_in_weekend_nights	0	1	1	0,9372	2	16
stays_in_week_nights	0	1	2	2,522	3	40
adults	0	2	2	1,863	2	4
children	0	0	0	0,1047	0	3
babies	0	0	0	0,007706	0	2
previous_cancellations	0	0	0	0,08698	0	26
previous_bookings_not_canceled	0	0	0	0,01197	0	72
booking_changes	0	0	0	0,2158	0	18
days_in_waiting_list	0	0	0	2,348	0	391
adr	0,26	71	95	103,64	126	397,38
required_car_parking_spaces	0	0	0	0,06185	0	2
total_of_special_requests	0	0	0	0,5712	1	5

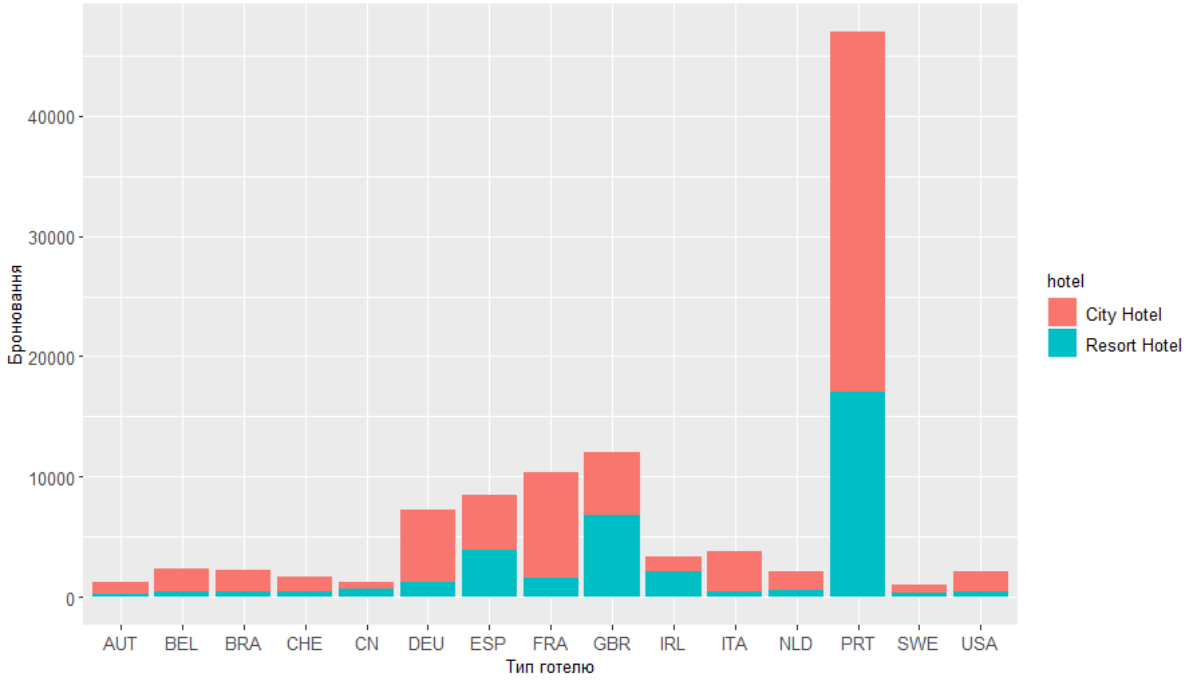
Отже, зміни в датасеті: 119390 → 116920 спостережень (видалили 2470 ~ 2,1%) і 32 → 33 змінні.
Також, у деяких стовпцях є дані, які можуть спотворювати результати, тому дослідження, у яких вони мають взяти участь, будуть використовуватися датасети без них (файл **conditions.R**).

lead_time	data1
stays_in_weekend_nights	data2
stays_in_week_nights	data3
previous_cancellations	data4
previous_bookings_not_canceled	data5
booking_changes	data6
days_in_waiting_list	data7
total_of_special_requests	data8
country	data9

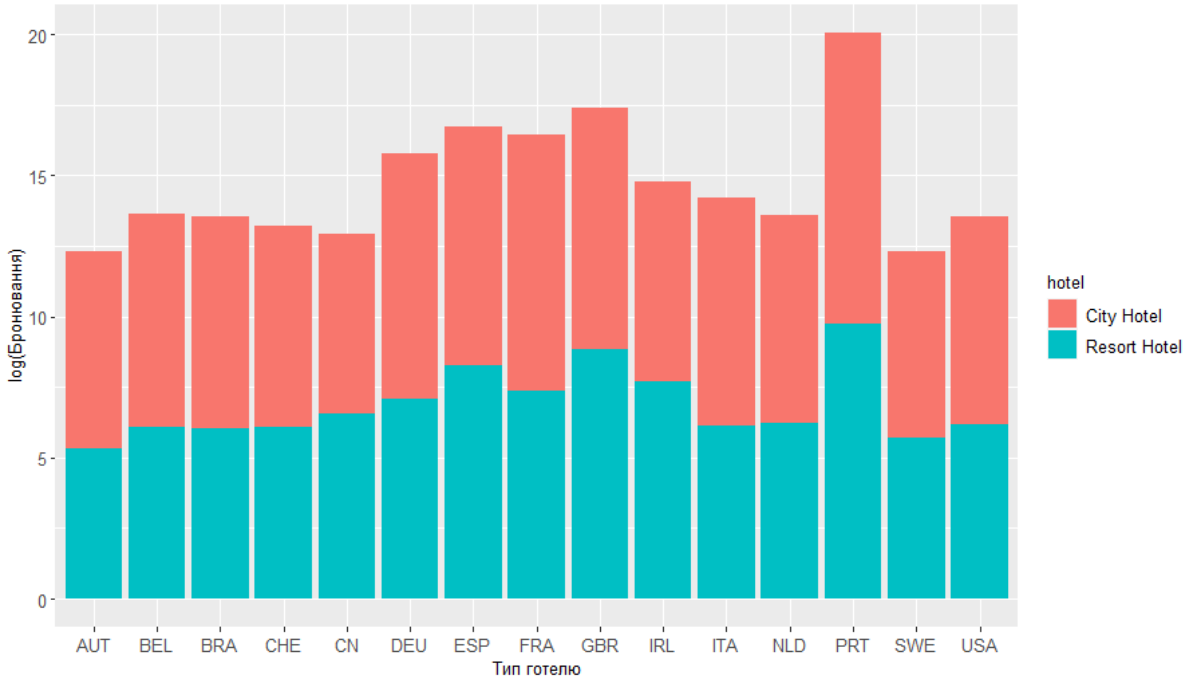
Дослідження

Питання: з якої країни люблять подорожувати, і в який саме тип готелю?
Гіпотеза: найбільш часті гості з північних країн, перевага надається City Hotel
Абсолютні значення:

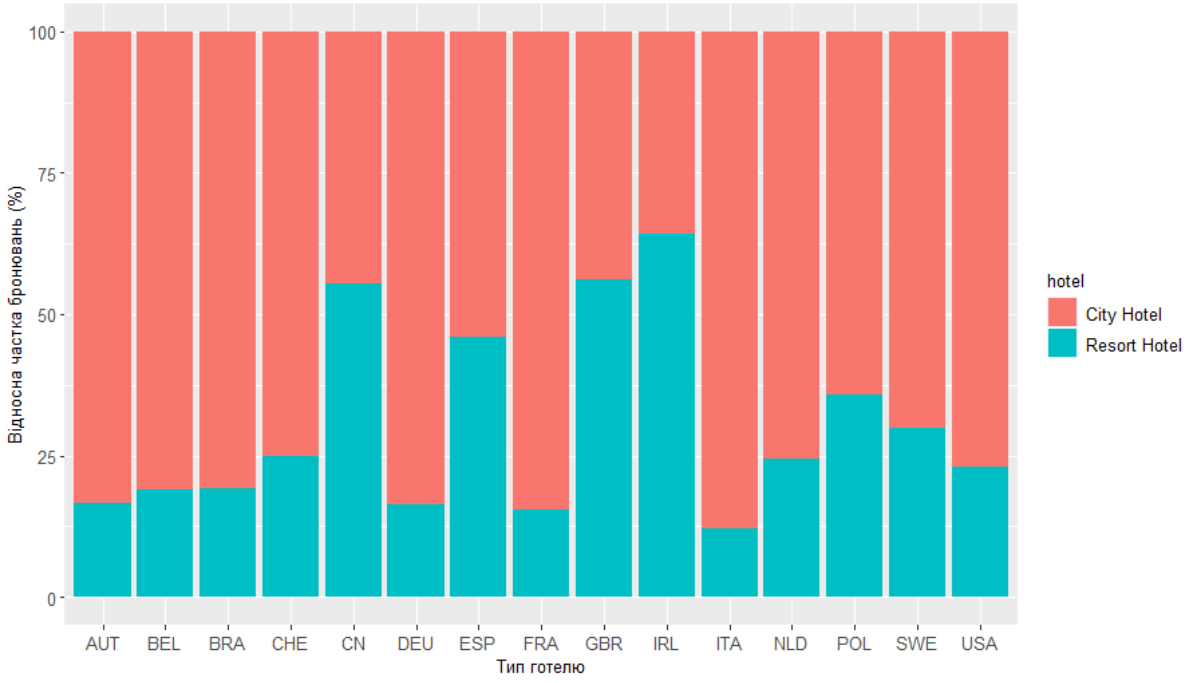


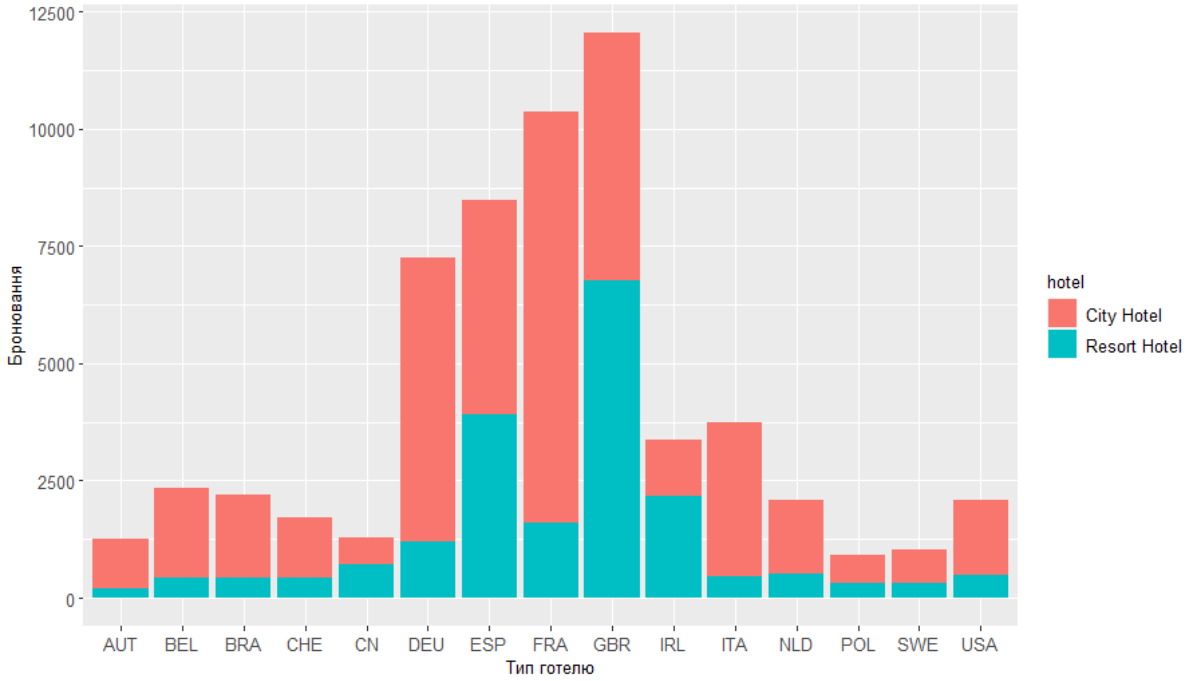


Графік після логарифмування:

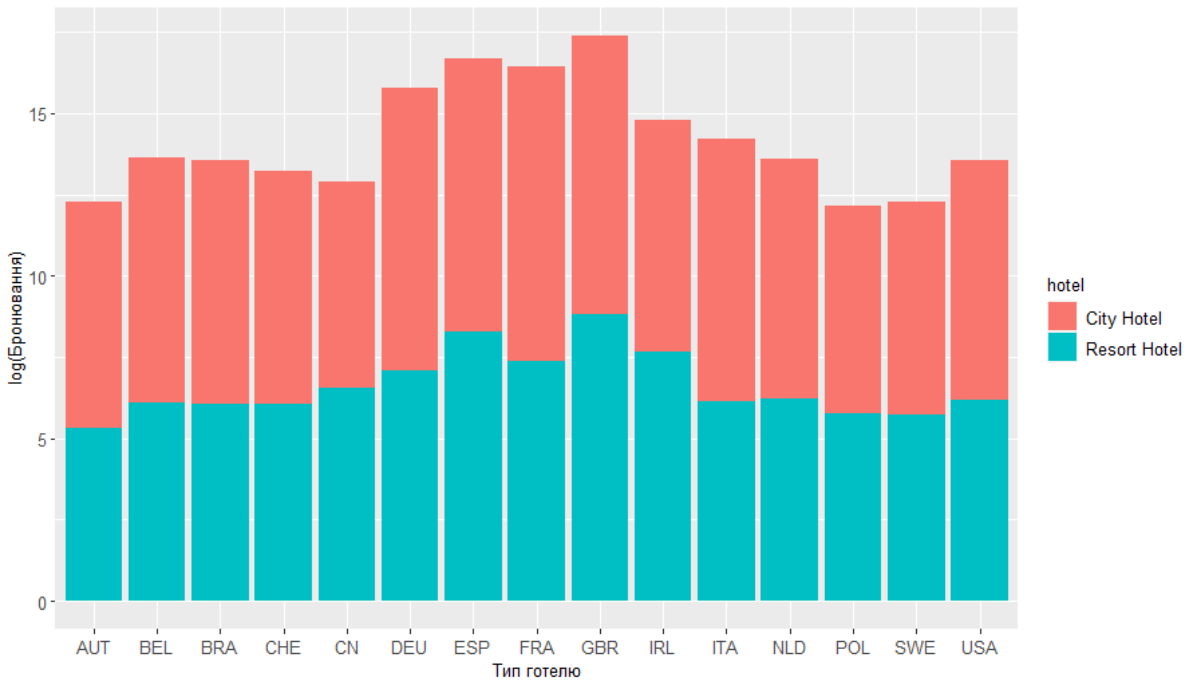


Як бачимо, сама Португалія має суттєву перевагу над іншими країнами, тому проведемо дослідження без неї.
Абсолютні значення:





Графік після логарифмування:



Відповідь: гіпотезу про те, що Португалія, як країна для відпочинку популярна серед північних країн спростовано. Кількість City Hotel у спостереженнях переважають над Resort Hotel приблизно в 2 рази. Але у відносних значеннях перевага в популярності City Hotel незначна.

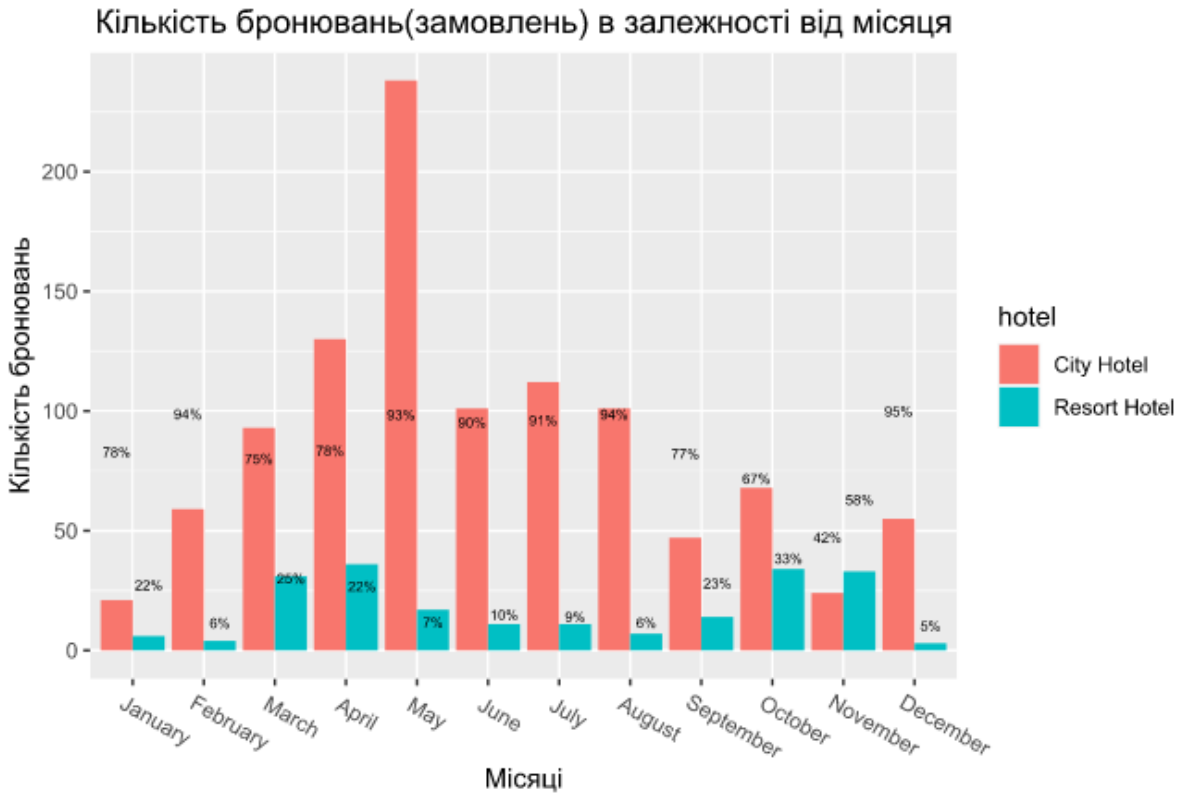
Питання: для яких країн які місяці туристичні?

Гіпотеза: від найбільш популярних до найменш: літні, осінні, весняні, зимові місяці

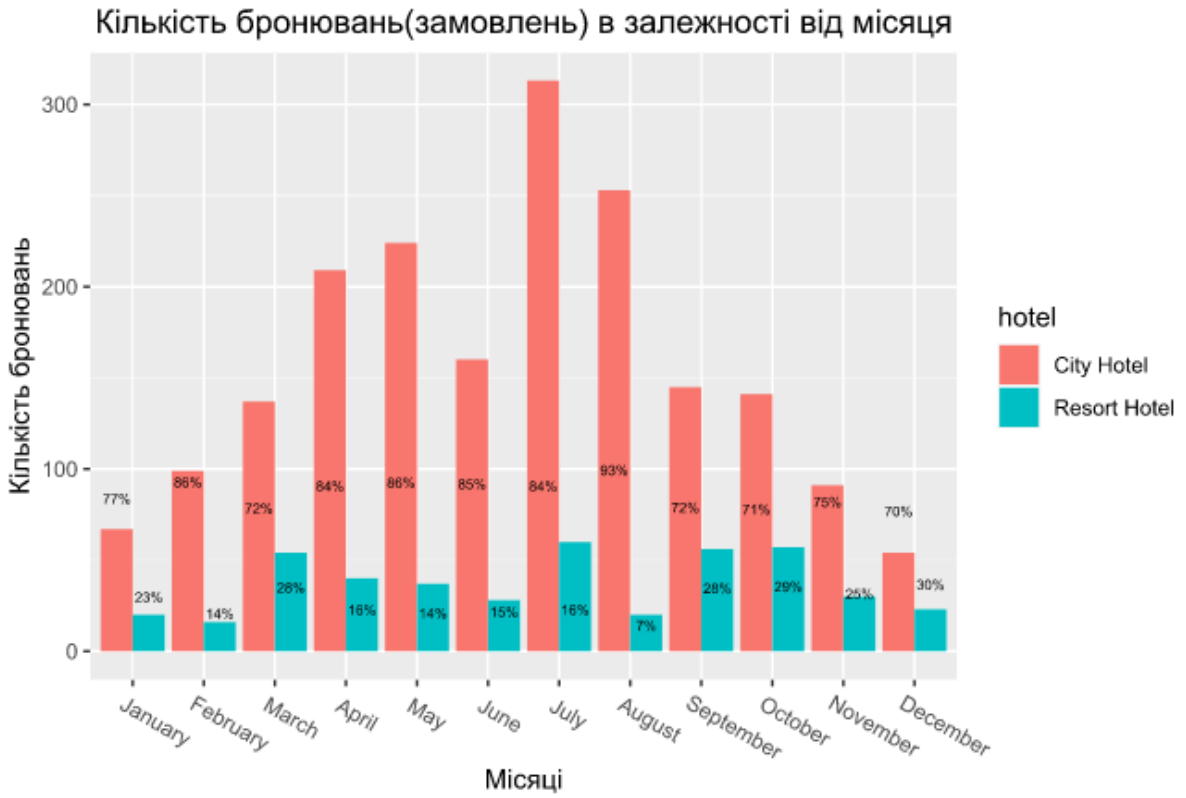
Як бачимо Португалія суттєво спотворила дослідження. Проведемо дослідження без неї.

Окремо для різних країн:

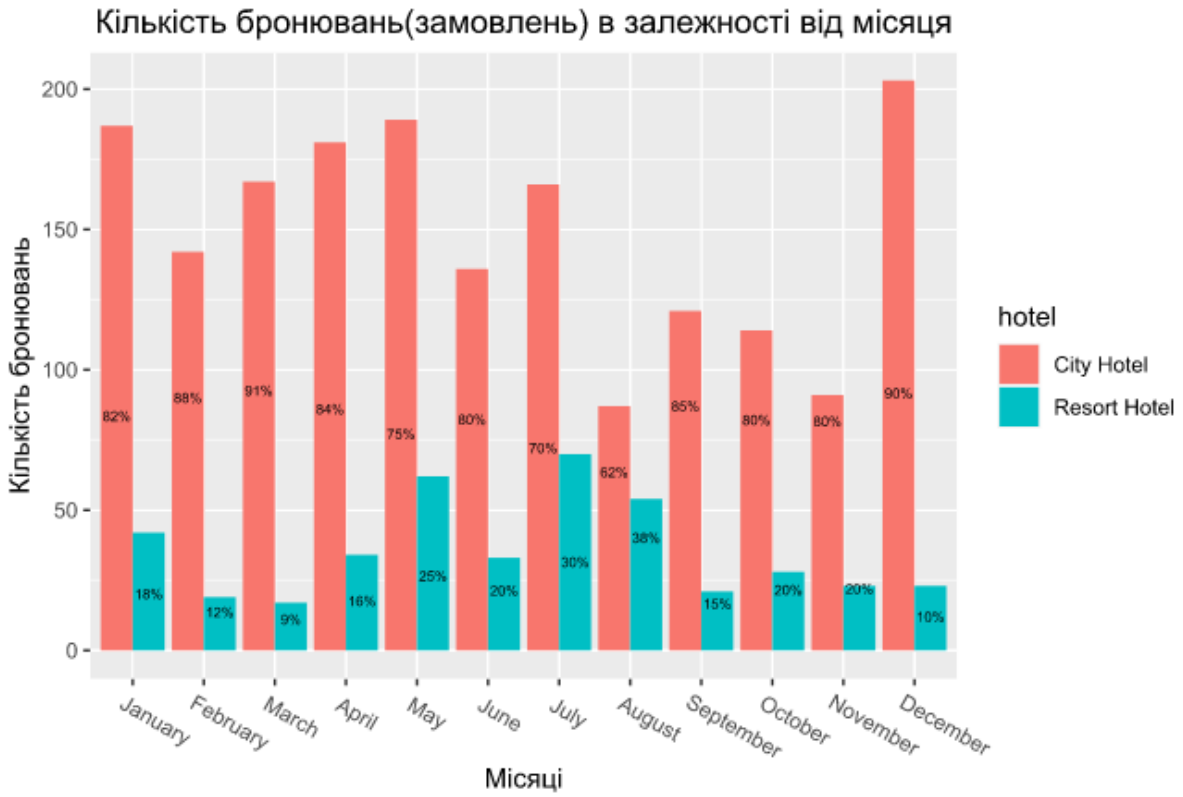
Австрія



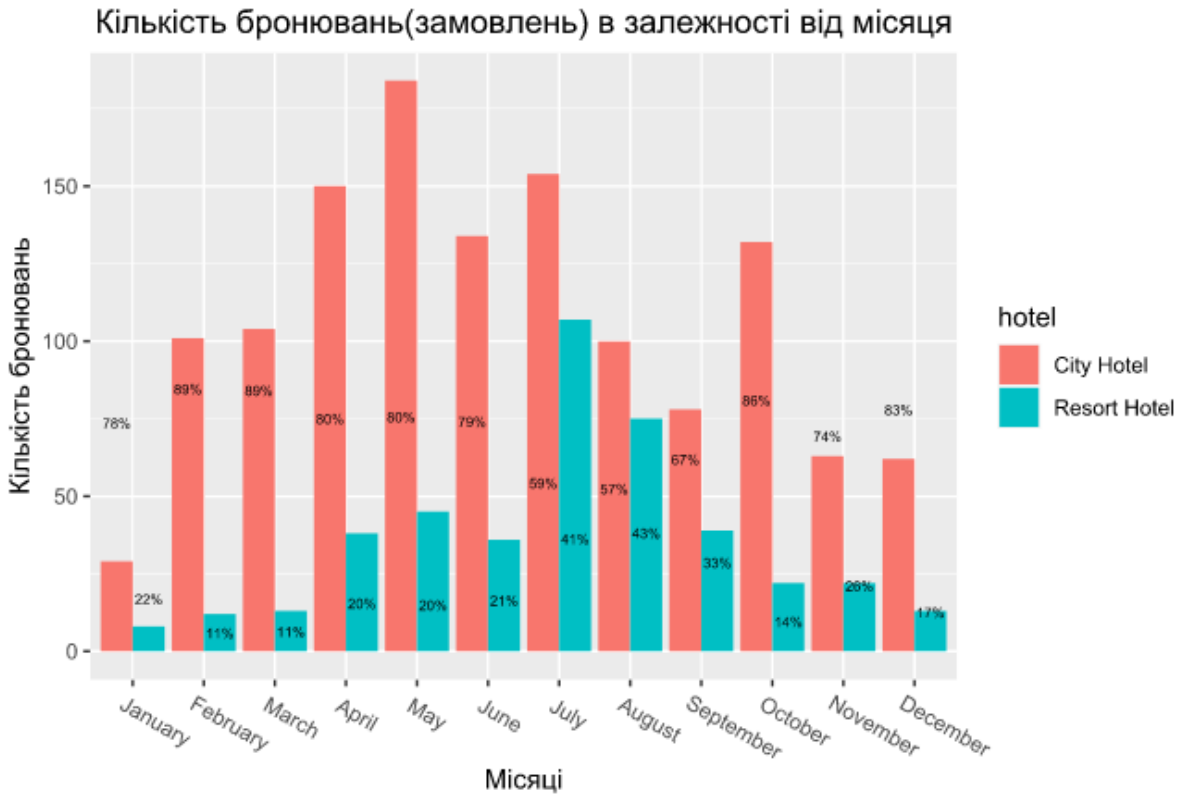
Бельгія



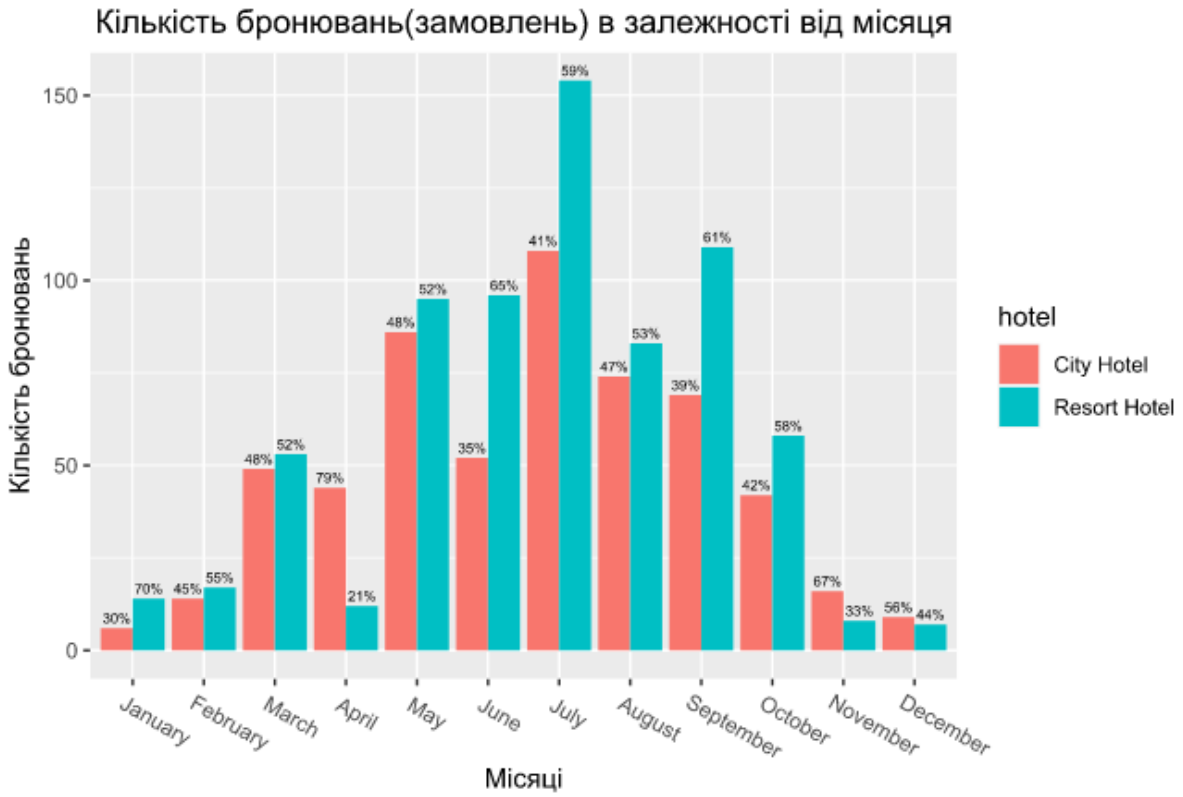
Бразилія



Швейцарія

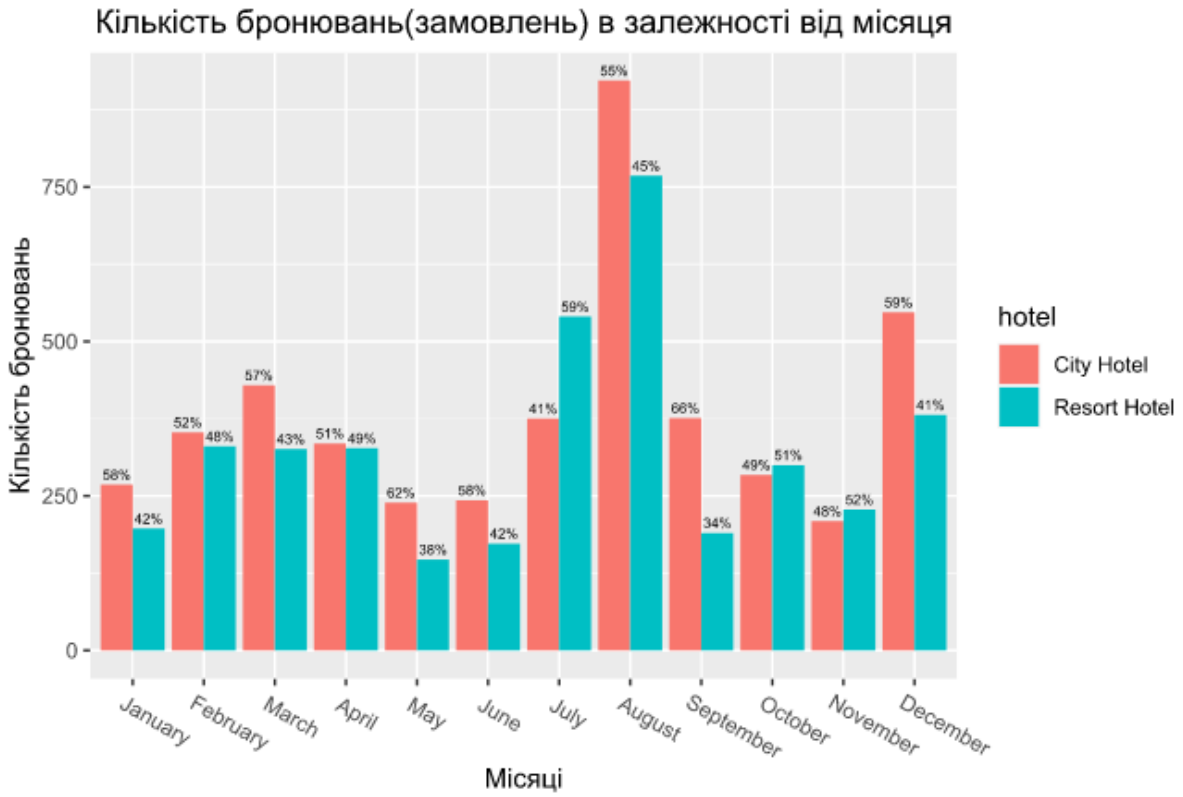


Китай

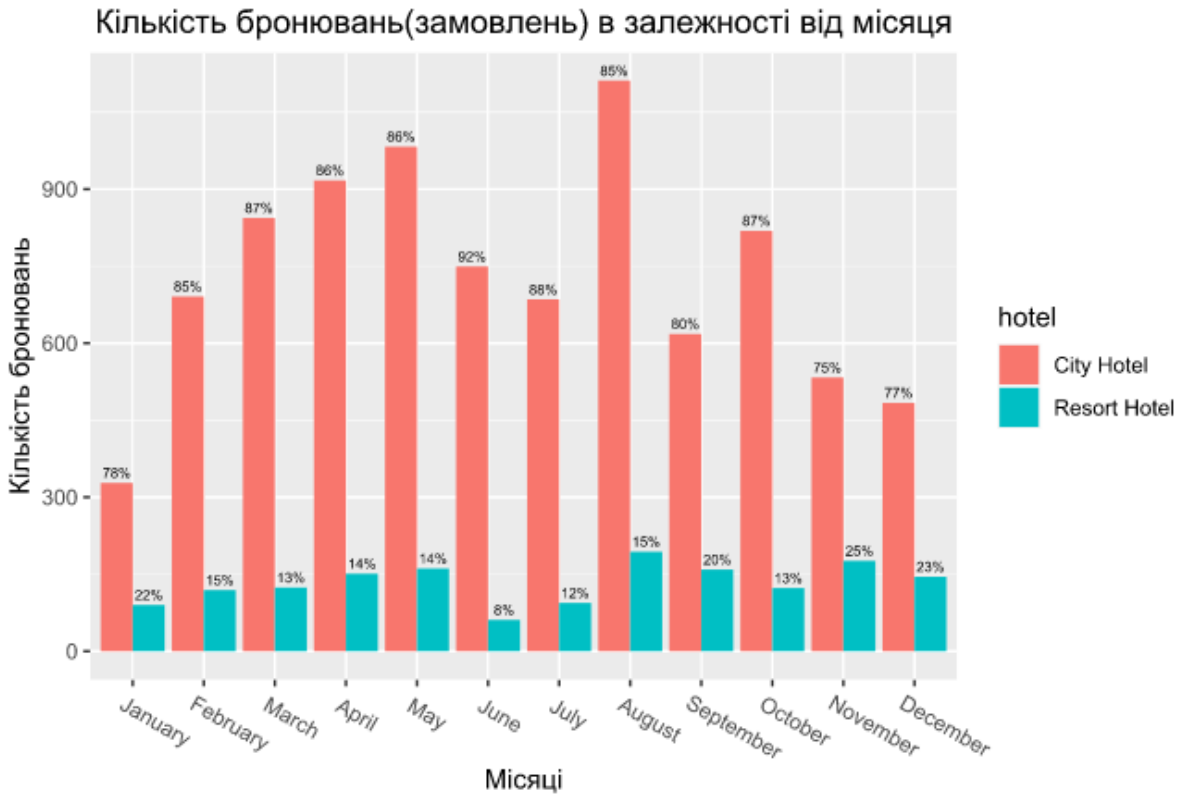


Німеччина

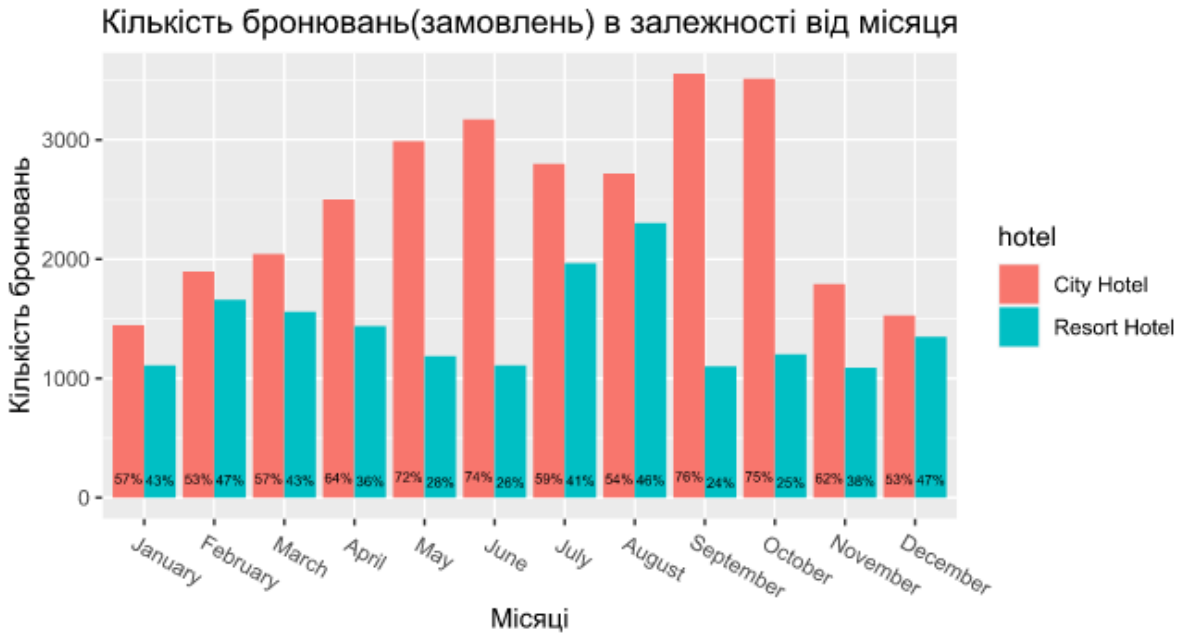
Іспанія



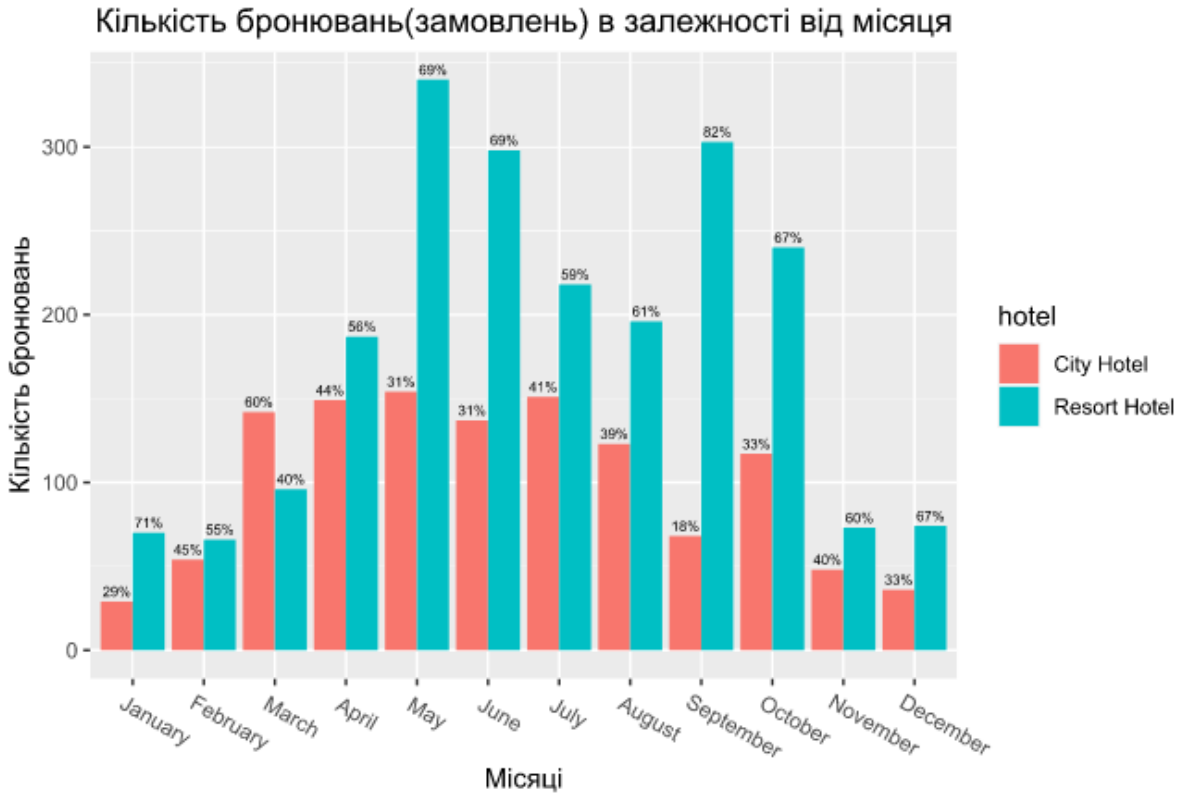
Франція



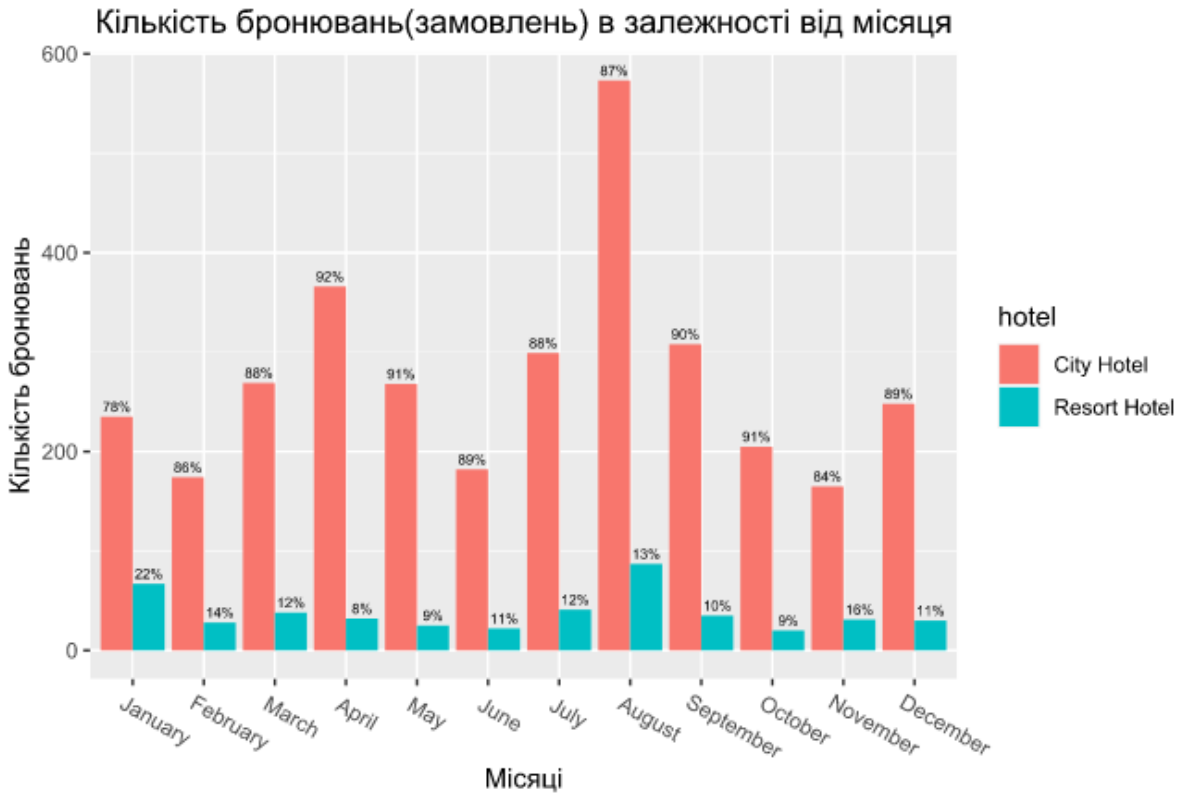
Велика Британія



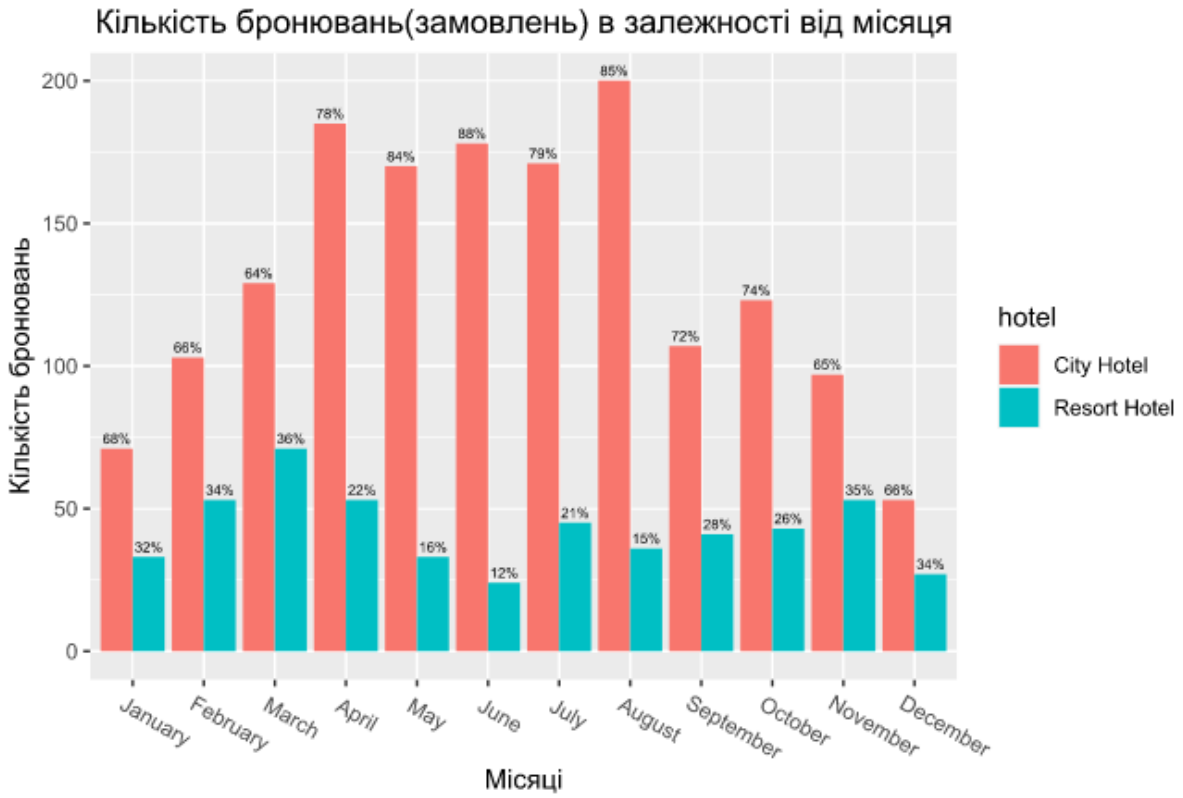
Ірландія



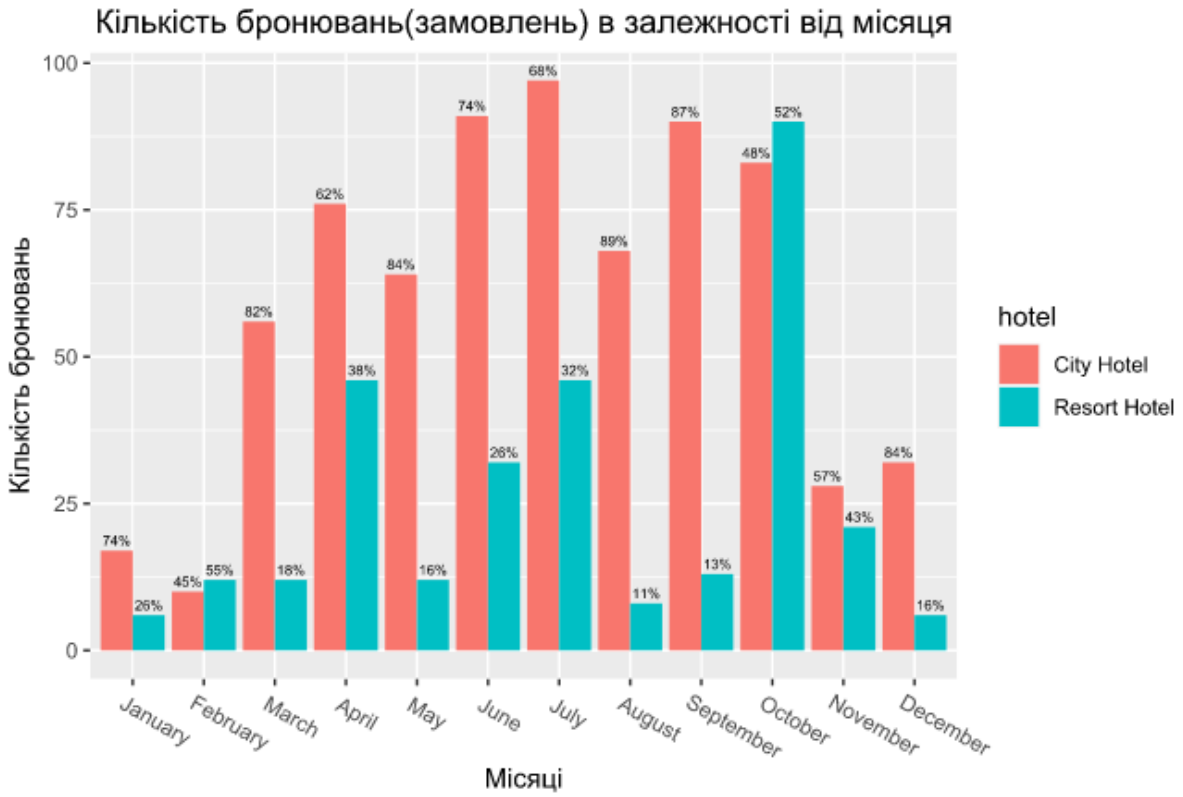
Італія



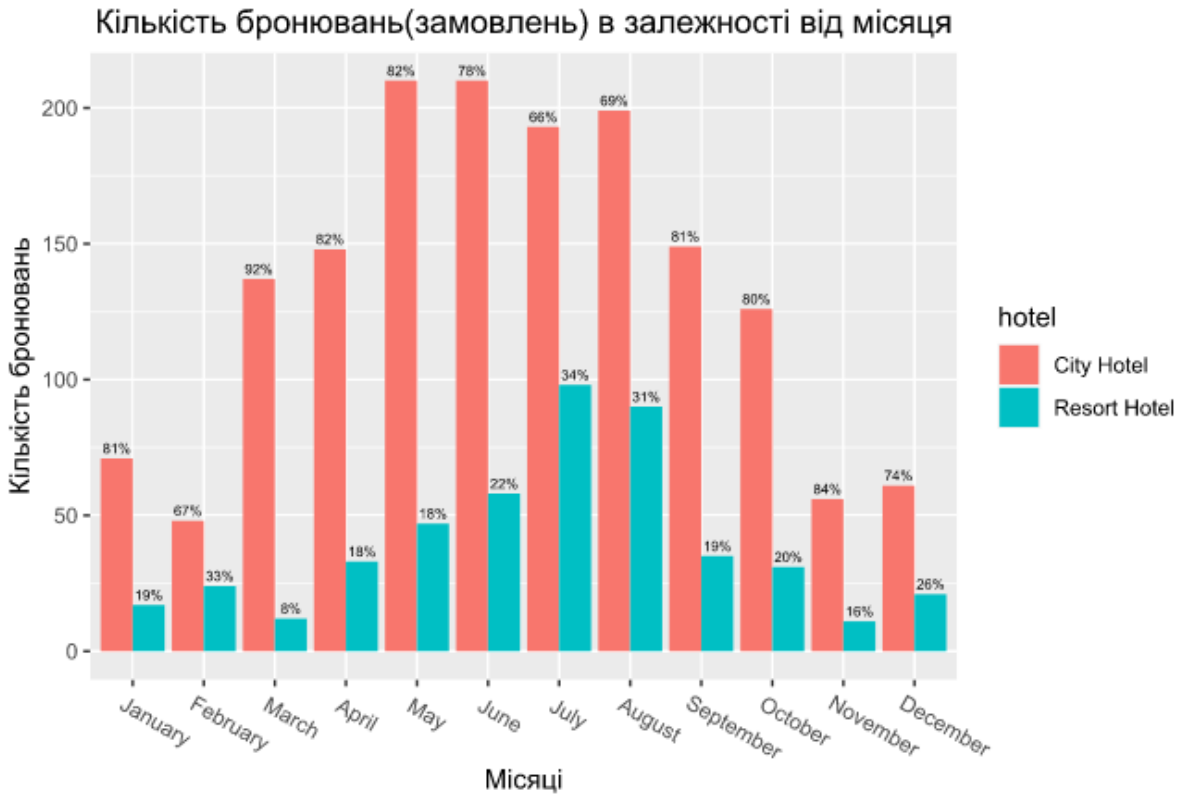
Нідерланди



Швеція



Сполучені Штати Америки

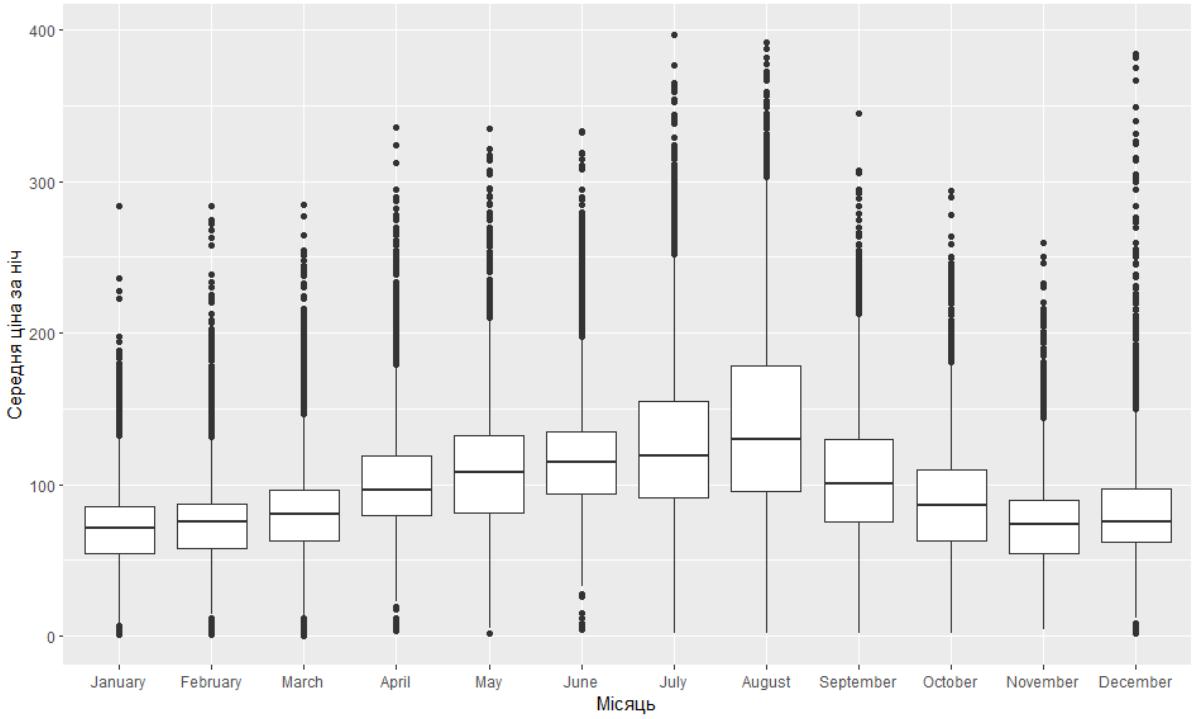


Відповідь: як бачимо найбільш популярні місяці - квітень, травень, червень, серпень, липень, вересень → весняні, літні, осінні, зимові місяці. Отже, гіпотезу спростовано.

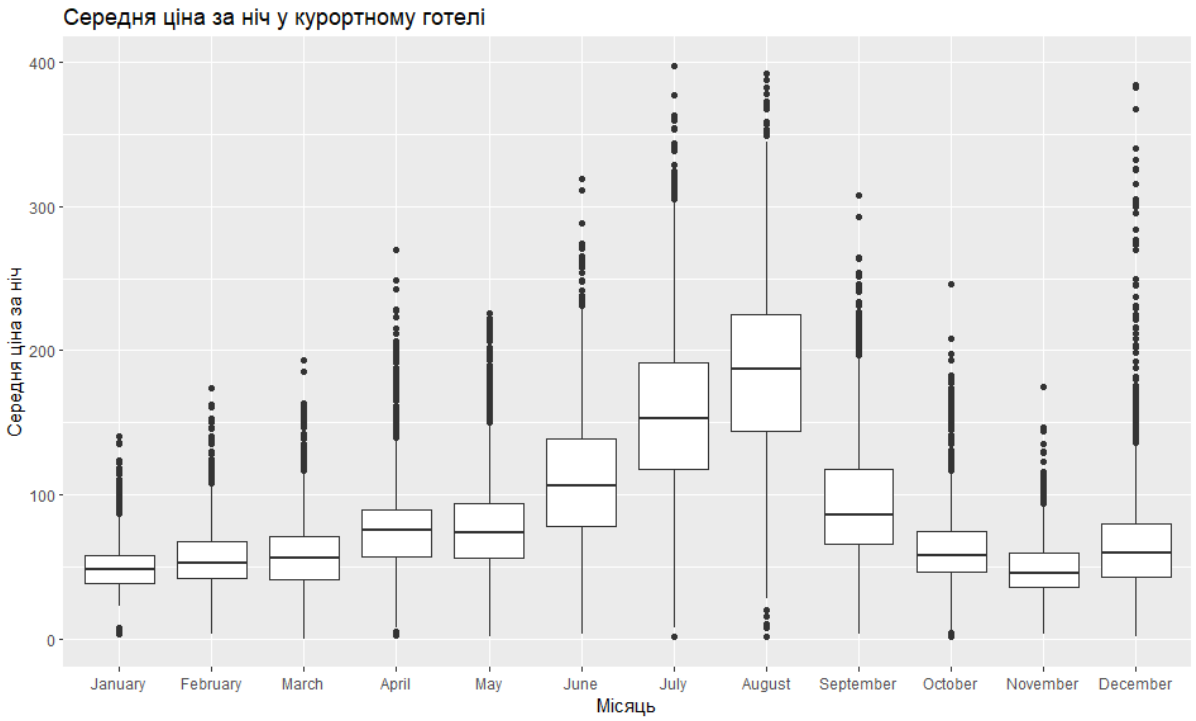
Питання: на які місяці вигідніше бронювати готель, тобто у які місяці середня ціна за ніч нижча?

Гіпотеза: у холодні місяці(кінець осені початок весни) ціни будуть нижчі

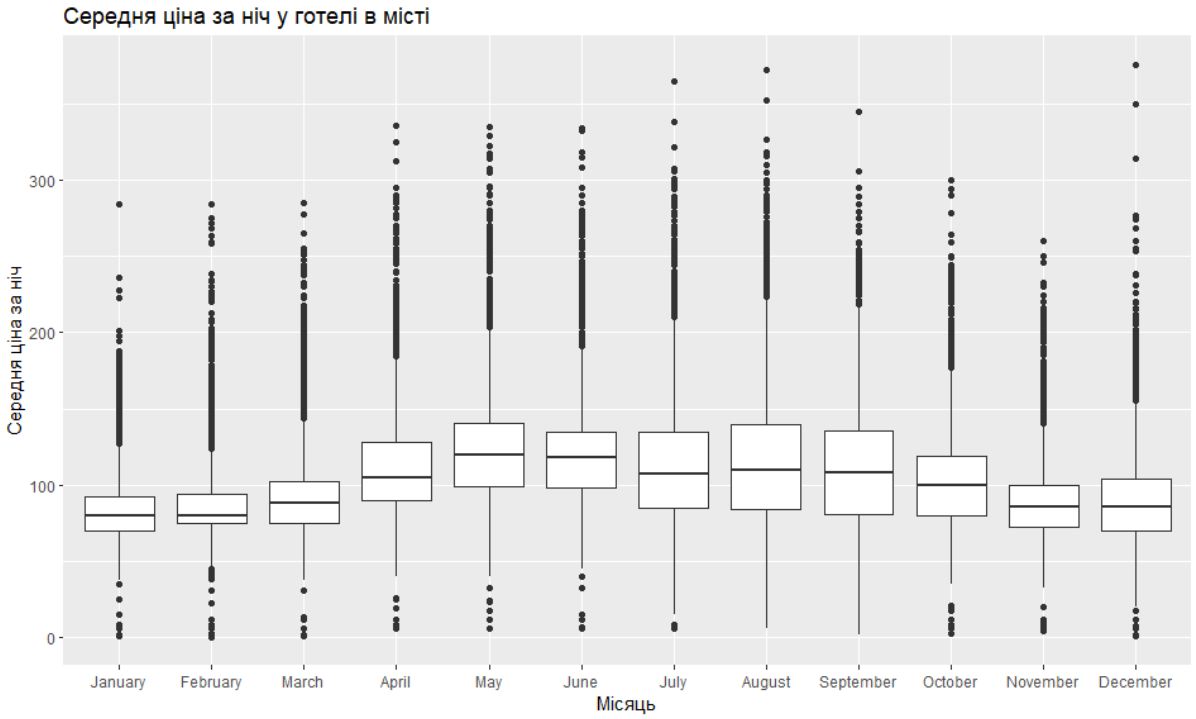
Розглянемо ціну на усі типи готелів



За отриманим результатом бачимо, що ціна для зимових місяців справді менша ніж у курортних місяців. Тепер переглянемо ціни на курортні та місцеві готелі.

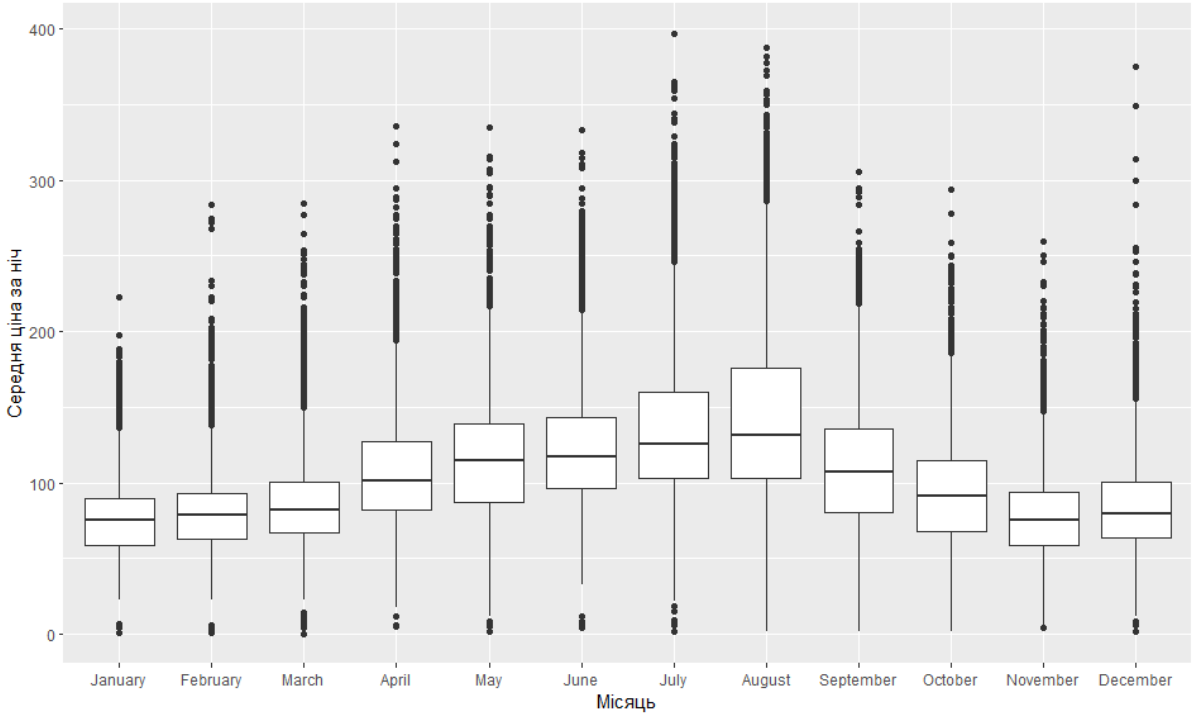


Висновок такий ж, але варто підмітити, великий ріст середньої ціни у курортні місяці та її спад на початку осені.

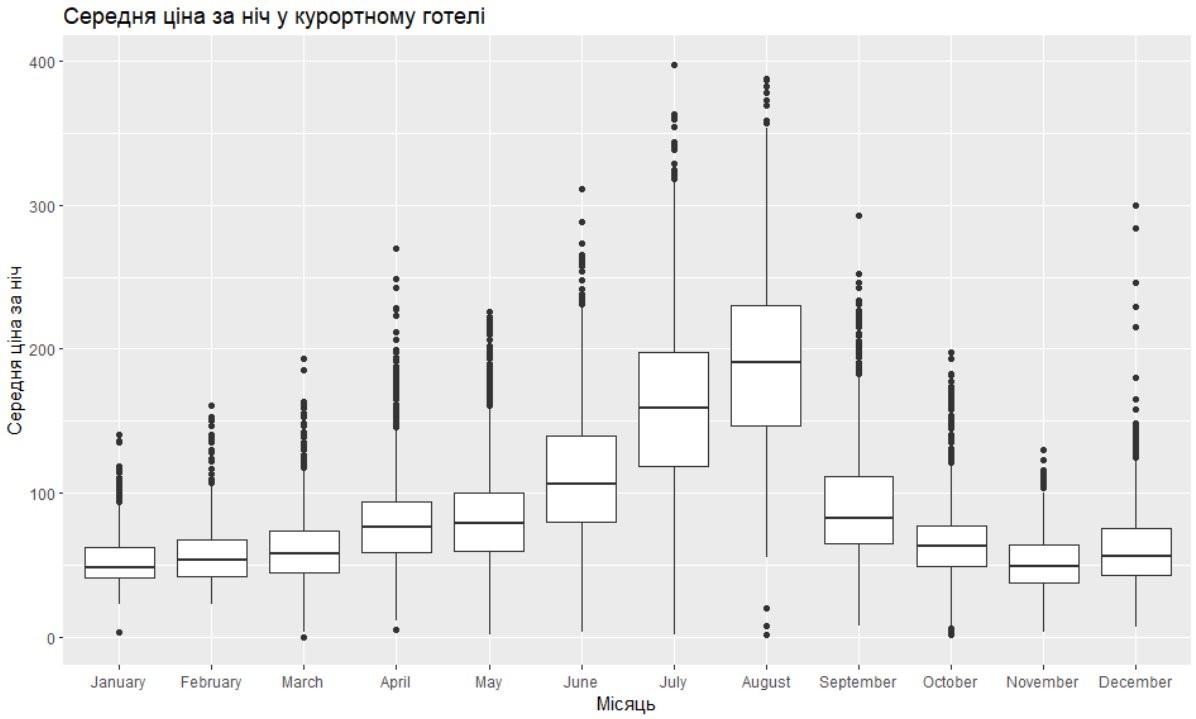


Щодо місцевих готелів, то ціна майже однакова, але період кінця осені та початку весни має нижчі ціни.

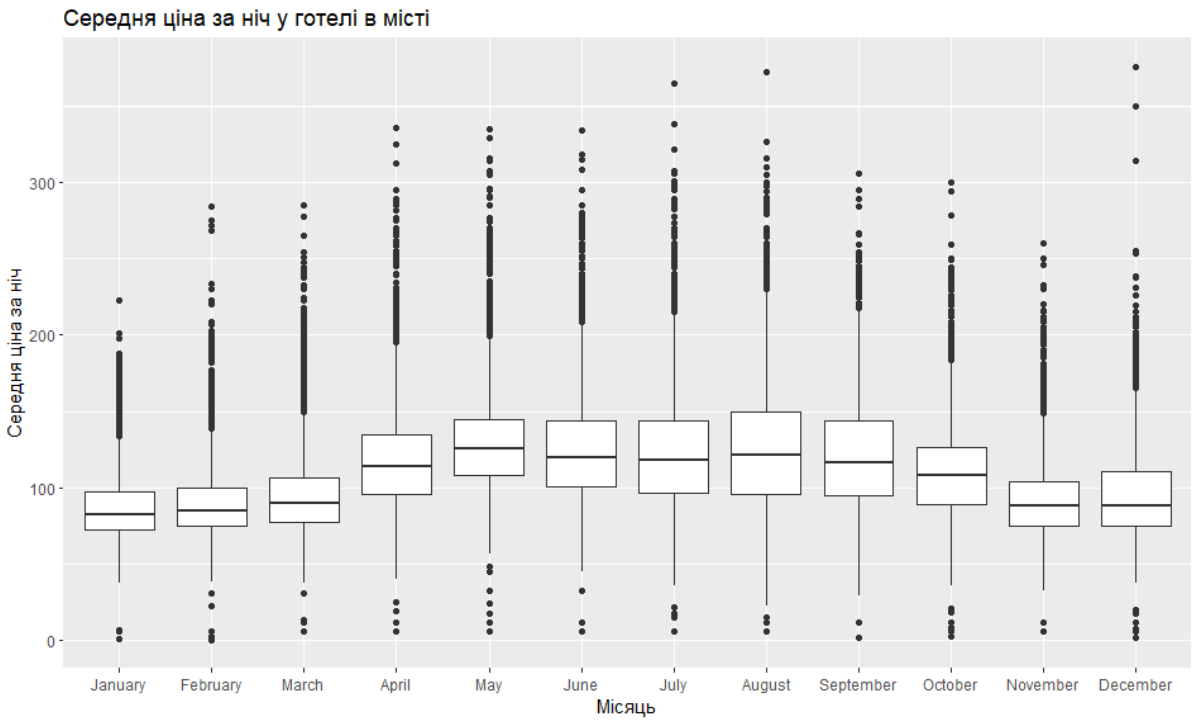
Тепер порівняємо результати не враховуючи замовників із Португалії.



Результати кардинально не змінилися, але переглянемо ситуацію для різних типів готелів.

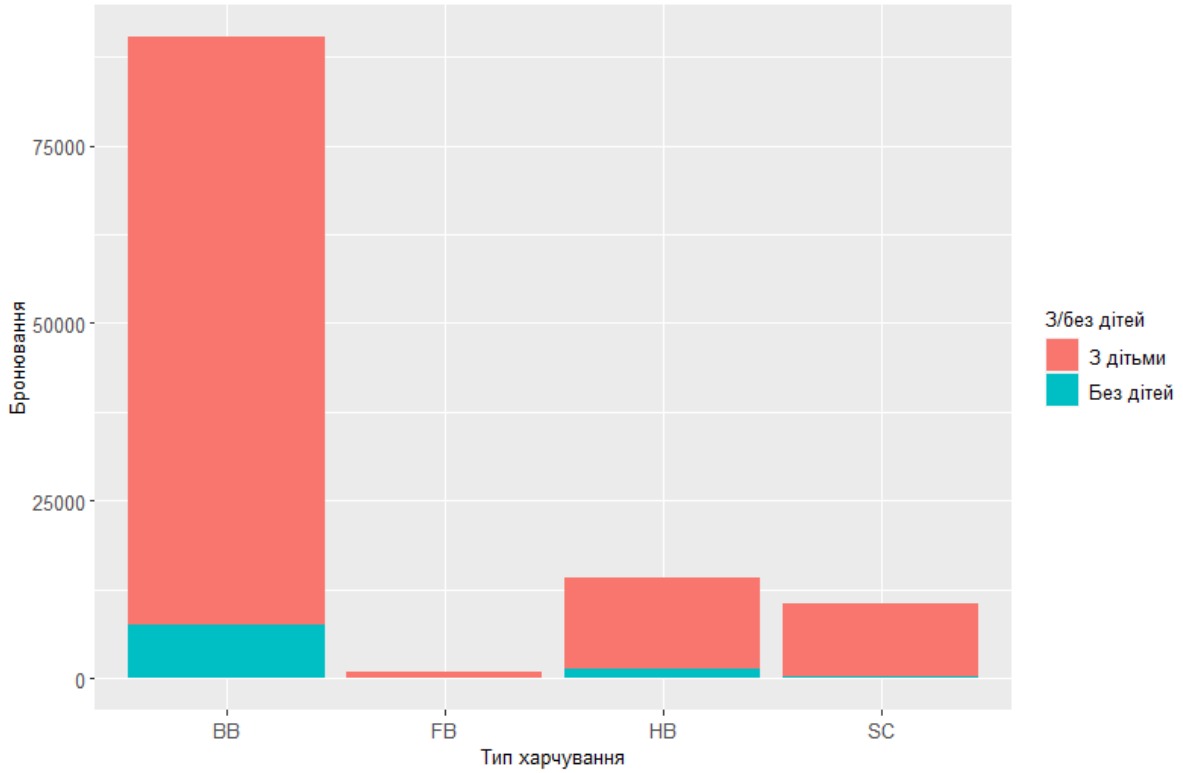


Як бачимо результати майже не відрізняються: та сама динаміка росту та спаду ціни.

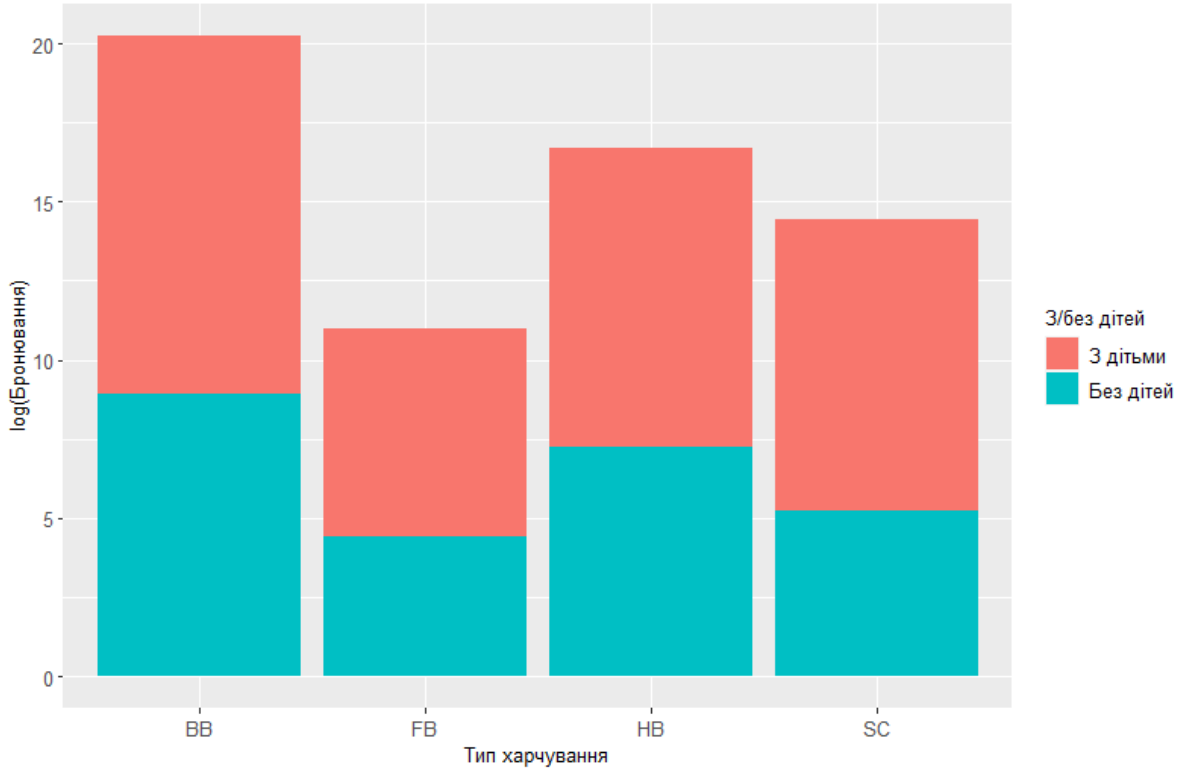


Відповідь: можна підтвердити гіпотезу. Середня ціна за ніч нижча у період з кінця осені до початку весни.

Питання:чи пов'язані тип відвідувачів (дорослі/(+діти)) з типом харчування?
Гіпотеза: відвідувачі з дітьми будуть замовляти готель з FB (повне харчування).
Абсолютні значення:



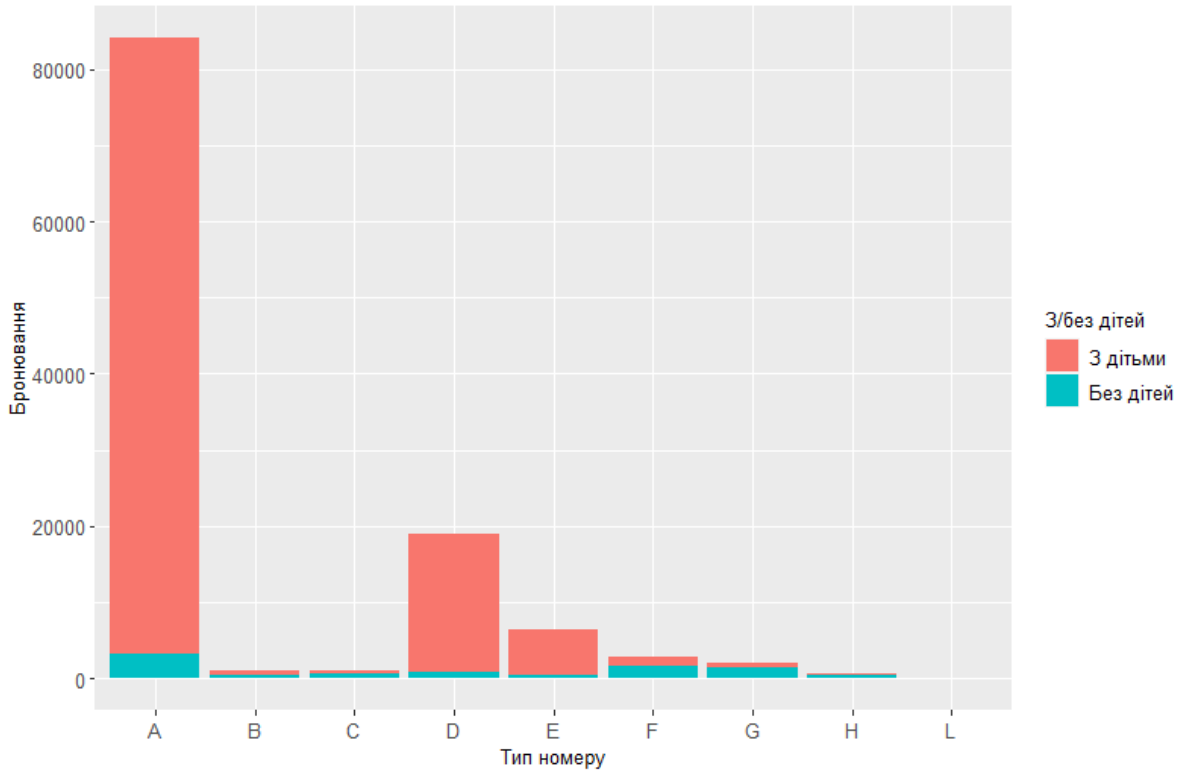
Графік після логарифмування:



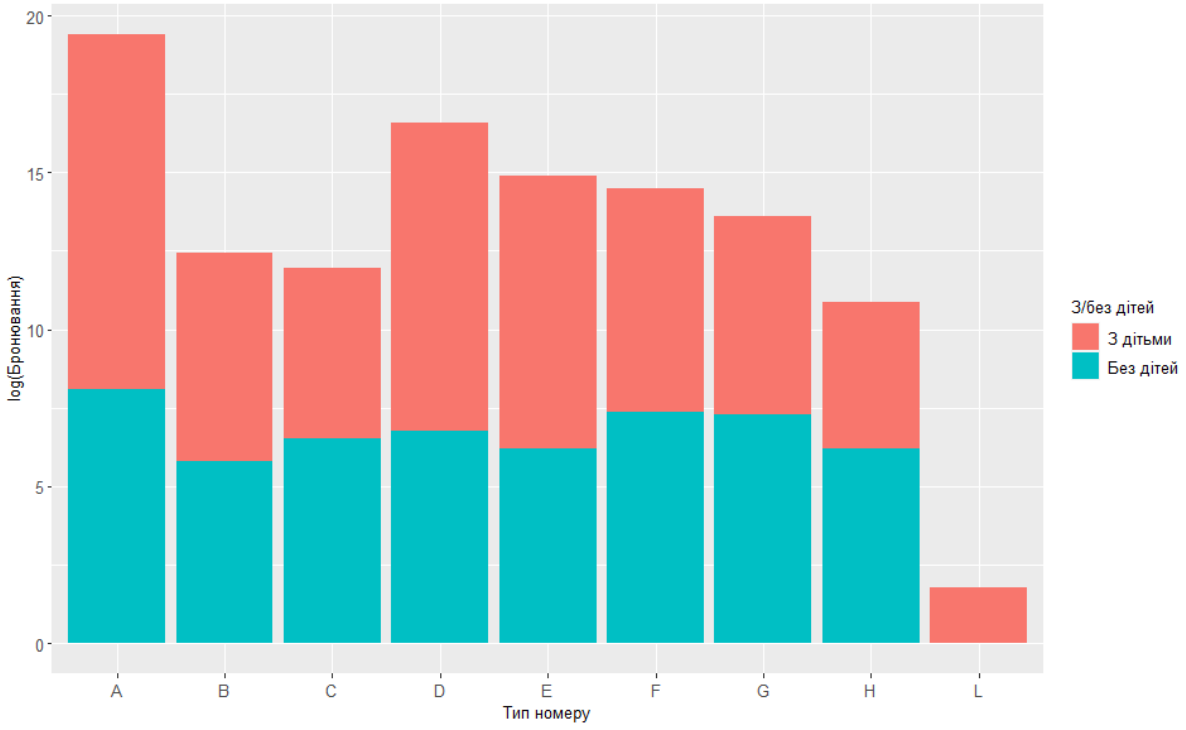
Відповідь: відвідувачі з дітьми замовляють готелі з типом харчування BB (лише сніданок). Гіпотезу спростовано.

Питання: чи пов'язані тип відвідувачів (дорослі (+ діти)) з типом номеру?
Гіпотеза: вибір типу номера не залежить від того чи є відвідувачі з дітьми, чи без них.

Абсолютні значення:



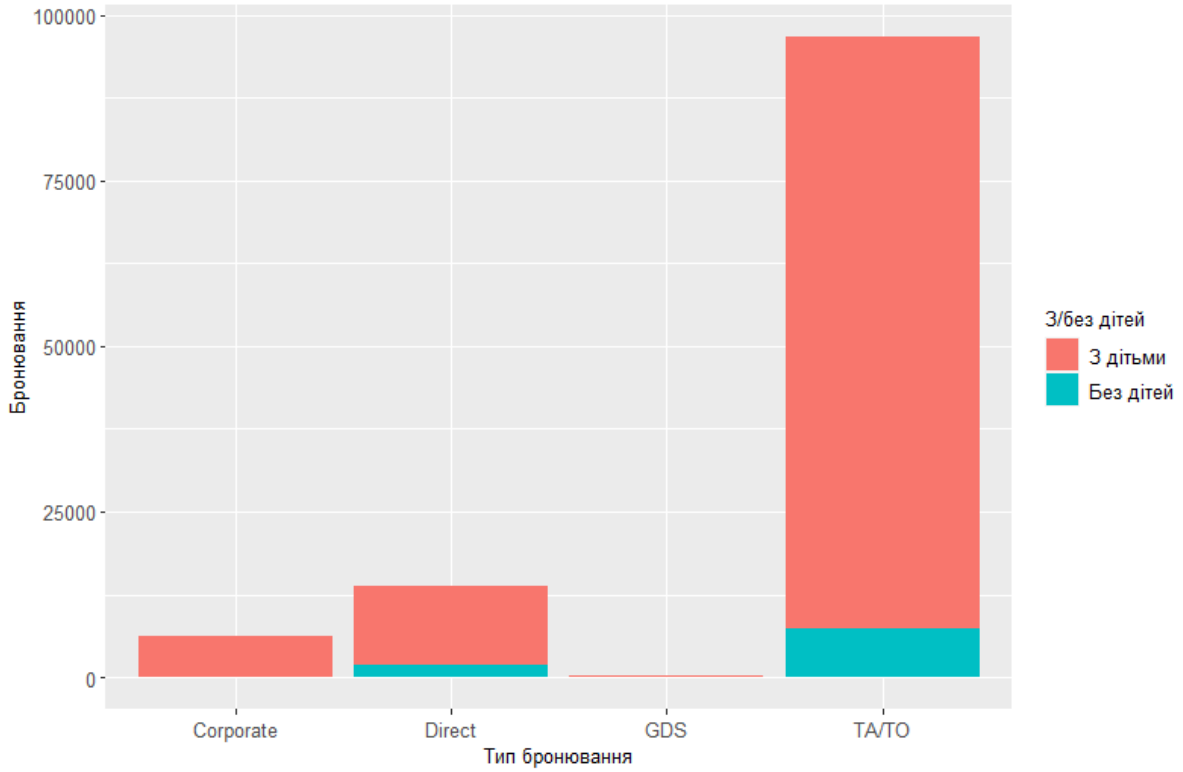
Графік після логарифмування:



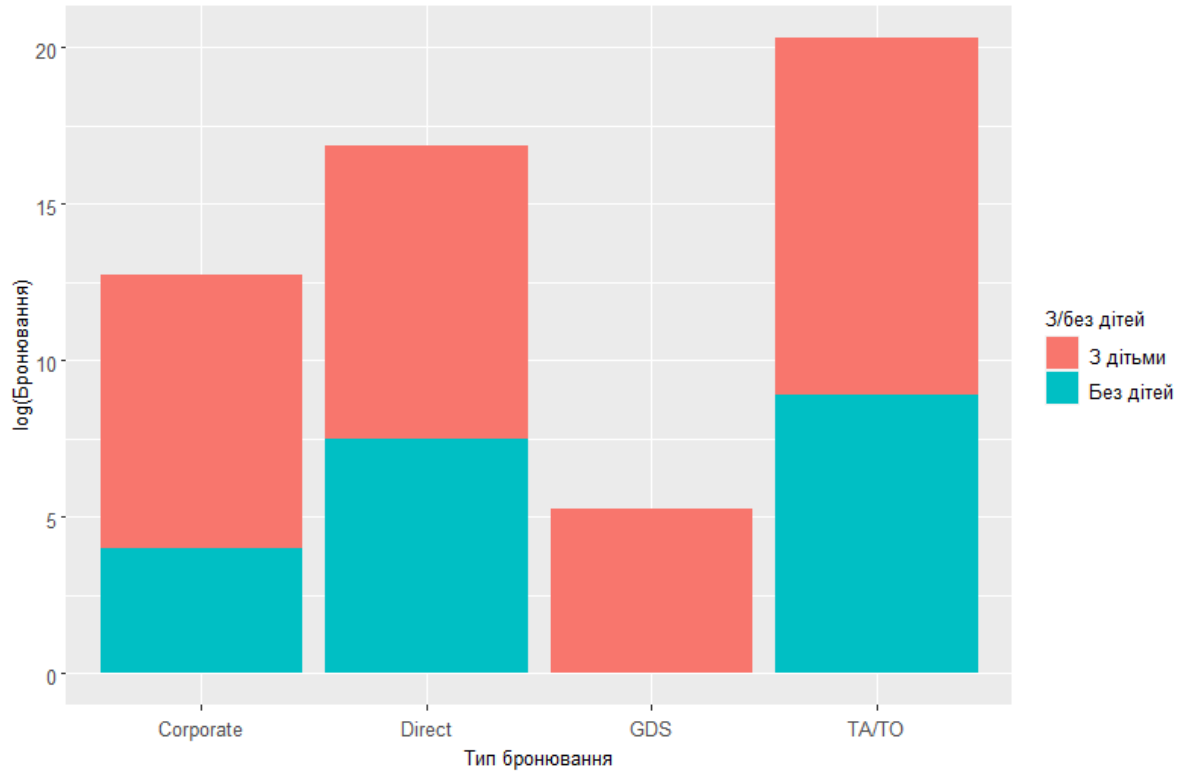
Відповідь: тип відвідувачів більш-менш рівномірно розподілений з типом номеру. Гіпотезу підтверджено.

Питання: чи пов'язані тип відвідувачів (дорослі (+ діти)) з типом бронювання?
Гіпотеза:

Абсолютні значення:

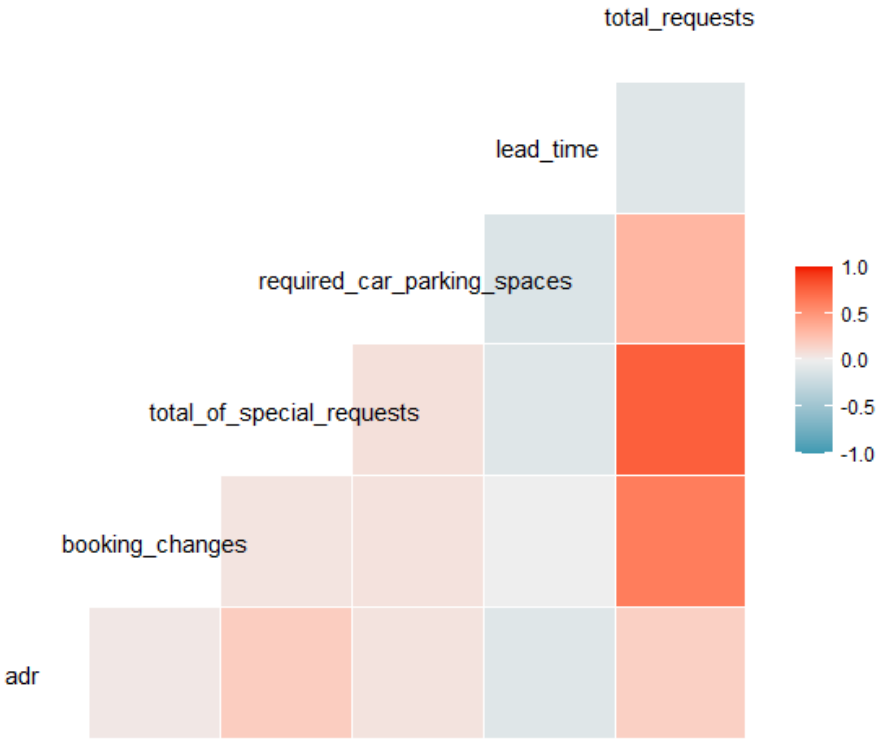


Графік після логарифмування:

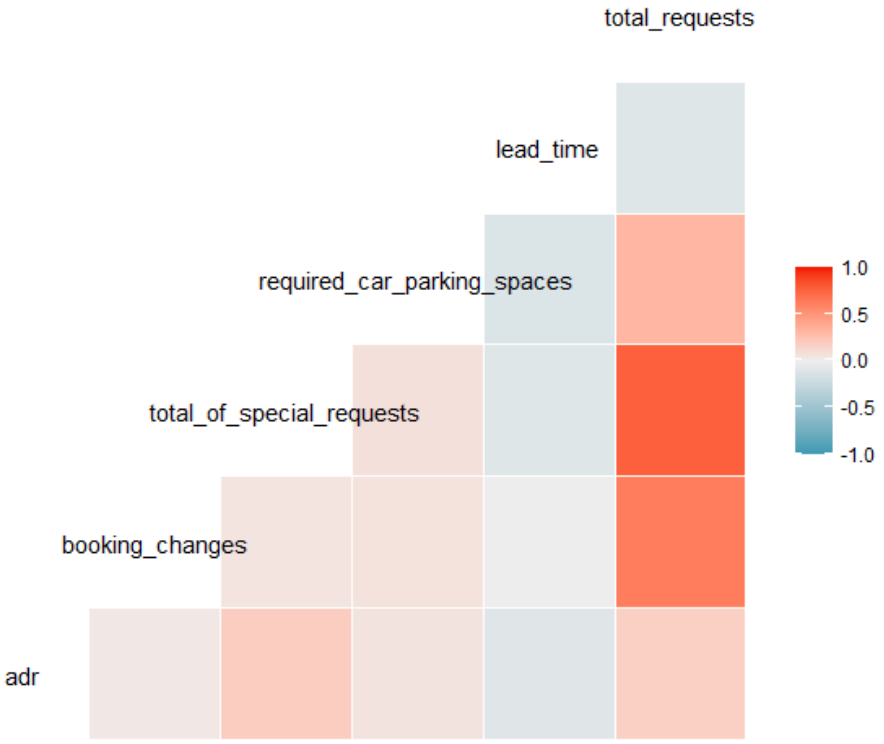


Відповідь: гіпотезу підтверджено.

Питання: залежність прибутку (adr) від різних факторів (booking_changes, total_of_special_requests, required_car_parking_spaces, lead_time).
Гіпотеза: на adr найбільше впливають: кількість спеціальних запитів клієнта (total_of_special_requests), кількість днів між бронюванням та прибуттям до готелю (lead_time); менше впливають: кількість змін/доповнень (booking_changes), кількість паркувальних місць (required_car_parking_spaces).



З data1(без підозрілих значень lead_time):



Як бачимо, підозрілі значення lead_time майже не впливають на дослідження.
Відповідь: найбільше впливає кількість спеціальних запитів клієнта та кількість днів між бронюванням та прибуттям до готелю, інші значення менше. Отже, гіпотезу підтверджено.

Висновки

Питання: з якої країни люблять подорожувати, і в який саме тип готелю?
Гіпотеза: найбільш часті гості з північних країн, перевага надається City Hotel
Відповідь: гіпотезу спростовано.

Питання: для яких країн які місяці туристичні?
Гіпотеза: від найбільш популярних до найменш: літні, осінні, весняні, зимові місяці
Відповідь: гіпотезу спростовано.

Питання: чи пов'язані тип відвідувачів (дорослі/(+діти)) з типом харчування?
Гіпотеза: відвідувачі з дітьми будуть замовляти готель з FB (повне харчування).
Відповідь: гіпотезу спростовано.

Питання: чи пов'язані тип відвідувачів (дорослі (+ діти)) з типом номеру?
Гіпотеза: вибір типу номера не залежить від того чи є відвідувачі з дітьми, чи без них.
Відповідь: гіпотезу підтверджено.

Питання: чи пов'язані тип відвідувачів (дорослі (+ діти)) з типом бронювання?
Гіпотеза: відвідувачі з дітьми будуть замовляти готель через “TA” (“Travel Agents”) або “TO” (“Tour Operators”).
Відповідь: гіпотезу підтверджено.

Питання: залежність прибутку (adr) від різних факторів (booking_changes, total_of_special_requests, required_car_parking_spaces, lead_time).
Гіпотеза: на adr найбільше впливають: кількість спеціальних запитів клієнта (total_of_special_requests), кількість днів між бронюванням та прибуттям до готелю (lead_time); менше впливають: кількість змін/доповнень (booking_changes), кількість паркувальних місць (required_car_parking_spaces).
Відповідь: гіпотезу підтверджено.

Питання: на які місяці вигідніше бронювати готель, тобто у які місяці середня ціна за ніч нижча?
Гіпотеза: у холодні місяці(кінець осені початок весни) ціни будуть нижчі
Відповідь: гіпотезу підтверджено