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.

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%)

Додаткова робота з даними

Була створена змінна area, яка поділяє країни на 'North', 'South' і 'Centre'.

```
north_countries <- c('ATA', 'DNK', 'EST', 'FIN', 'FRO', 'GBR', 'IMN', 'IRL', 'ISL',
                     'LTU', 'LVA', 'NOR', 'SWE')
south_countries <- c('ABW', 'AGO', 'AIA', 'ALB', 'AND', 'ARE', 'ARG', 'ARM', 'ASM',
                     'ATF', 'AUS', 'AZE', 'BDI', 'BEN', 'BFA', 'BGD', 'BGR', 'BHR',
                     'BHS', 'BIH', 'BOL', 'BRA', 'BRB', 'BWA', 'CAF', 'CHL', 'CIV',
                     'CMR', 'COL', 'COM', 'CPV', 'CRI', 'CUB', 'CYM', 'CYP', 'DMA',
                     'DOM', 'DZA', 'ECU', 'EGY', 'ESP', 'ETH', 'FJI', 'GAB', 'GEO',
                     'GHA', 'GIB', 'GLP', 'GNB', 'GRC', 'GTM', 'GUY', 'HND', 'HRV',
                     'IDN', 'IND', 'IRQ', 'ISR', 'ITA', 'JAM', 'JOR', 'KEN', 'KHM',
                     'KIR', 'KNA', 'KWT', 'LAO', 'LBN', 'LBY', 'LCA', 'LKA', 'MAR',
                     'MCO', 'MDG', 'MDV', 'MEX', 'MKD', 'MLI', 'MLT', 'MMR', 'MNE',
                     'MOZ', 'MRT', 'MUS', 'MWI', 'MYS', 'MYT', 'NAM', 'NCL', 'NGA',
                     'NIC', 'NZL', 'OMN', 'PAK', 'PAN', 'PER', 'PHL', 'PLW', 'PRI',
                     'PRT', 'PRY', 'PYF', 'QAT', 'RWA', 'SAU', 'SDN', 'SEN', 'SGP',
```

'SLE', 'SLV', 'SMR', 'STP', 'SUR', 'SYC', 'SYR', 'TGO', 'THA',
'TJK', 'TMP', 'TUN', 'TUR', 'TZA', 'UGA', 'UMI', 'URY', 'VEN',
'VGB', 'VNM', 'ZAF', 'ZMB', 'ZWE', 'IRN')

```
centre_countries <- c('AUT', 'BEL', 'BLR', 'CHE', 'CHN', 'CN', 'CZE', 'DEU', 'DJI',
                      'FRA', 'GGY', 'HKG', 'HUN', 'JEY', 'JPN', 'KAZ', 'KOR', 'LIE',
                      'LUX', 'MAC', 'NLD', 'NPL', 'POL', 'ROU', 'RUS', 'SRB', 'SVK',
                      'SVN', 'TWN', 'UKR', 'USA', 'UZB')
```

Була створена зміна season, яка показує, у який сезон прибули відвідувачі (теплий чи холодний).
Була створена зміна with_children, яка показує, чи були з відвідувачами діти.
Була створена зміна lead_time_case, яка показує за скільки часу було зроблене бронювання.

Довірчі інтервали

1) Середні значення:

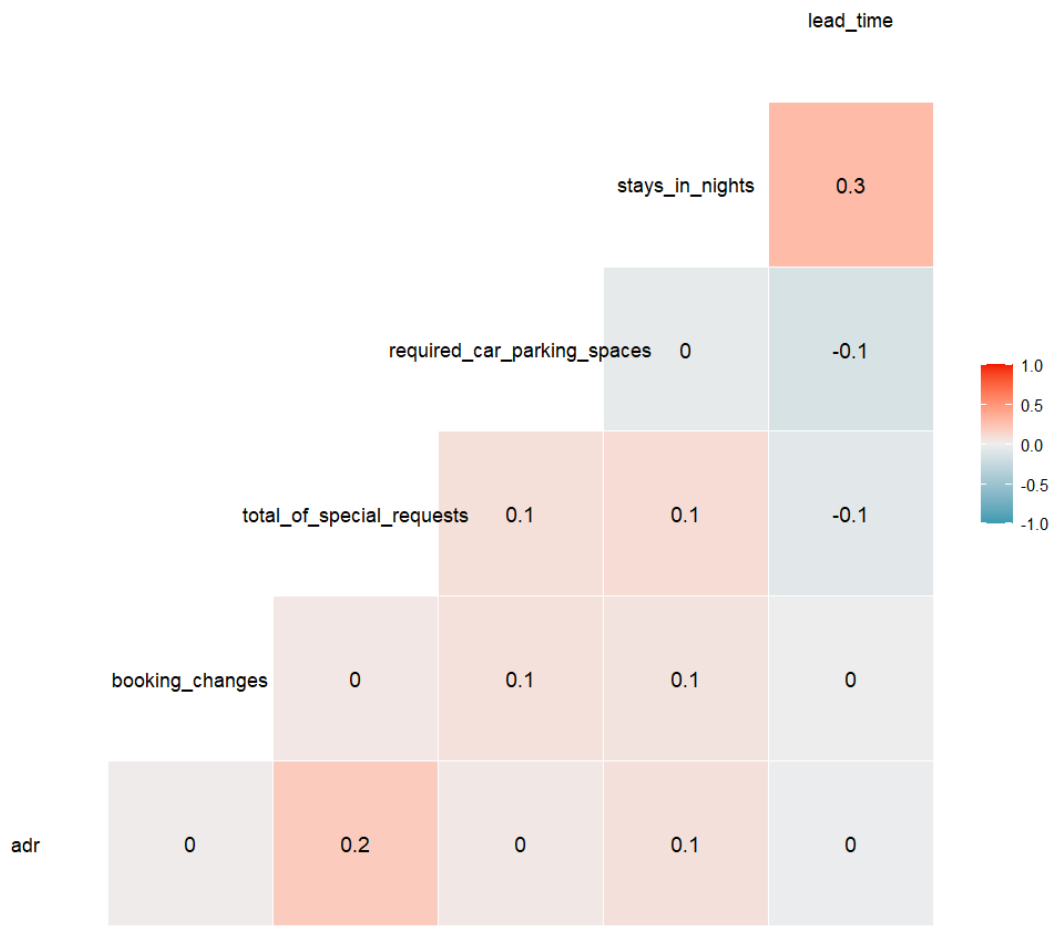
	mean	sd	n	a	b	a_t	b_t
lead_time	105.3809	106.9364	116920	104.7679	105.9939	104.7679	105.9939
stays_in_weekend_nights	0.9371964	0.9918537	116920	0.9315111	0.9428816	0.931511	0.9428817
stays_in_week_nights	2.521887	1.880868	116920	2.511106	2.532668	2.511106	2.532668
adults	1.86294	0.4802359	116920	1.860188	1.865693	1.860188	1.865693
children	0.1047126	0.3991863	116920	0.1024245	0.1070007	0.1024245	0.1070008
babies	0.007706124	0.08880477	116920	0.007197098	0.00821515	0.007197093	0.008215155
all_children	0.1124187	0.4111695	116920	0.1100619	0.1147756	0.1100619	0.1147756
previous_cancellations	0.08698255	0.850319	116920	0.08210855	0.09185655	0.0821085	0.0918566
bookings_not_canceled	0.1197314	1.432424	116920	0.1115208	0.127942	0.1115207	0.1279421
booking_changes	0.2157886	0.6312376	116920	0.2121703	0.2194068	0.2121703	0.2194068
days_in_waiting_list	2.348135	17.71489	116920	2.246594	2.449677	2.246593	2.449678
adr	103.6372	46.56286	116920	103.3703	103.9041	103.3703	103.9041
required_car_parking_spaces	0.06184571	0.242045	116920	0.06045831	0.0632331	0.0604583	0.06323312
total_of_special_requests	0.5712282	0.791426	116920	0.5666918	0.5757646	0.5666917	0.5757647

2) Медіани:

	median	lower_lim	upper_lim
lead_time	71	70.38703	71.61297
stays_in_weekend_nights	1	0.9943146	1.005685
stays_in_week_nights	2	1.989219	2.010781
adults	2	1.997247	2.002753
children	0	-0.002288165	0.002288165
babies	0	-0.0005090354	0.0005090354
all_children	0	-0.002356853	0.002356853
previous_cancellations	0	-0.00487409	0.00487409
bookings_not_canceled	0	-0.008210758	0.008210758
booking_changes	0	-0.0036183	0.0036183
days_in_waiting_list	0	-0.101543	0.101543

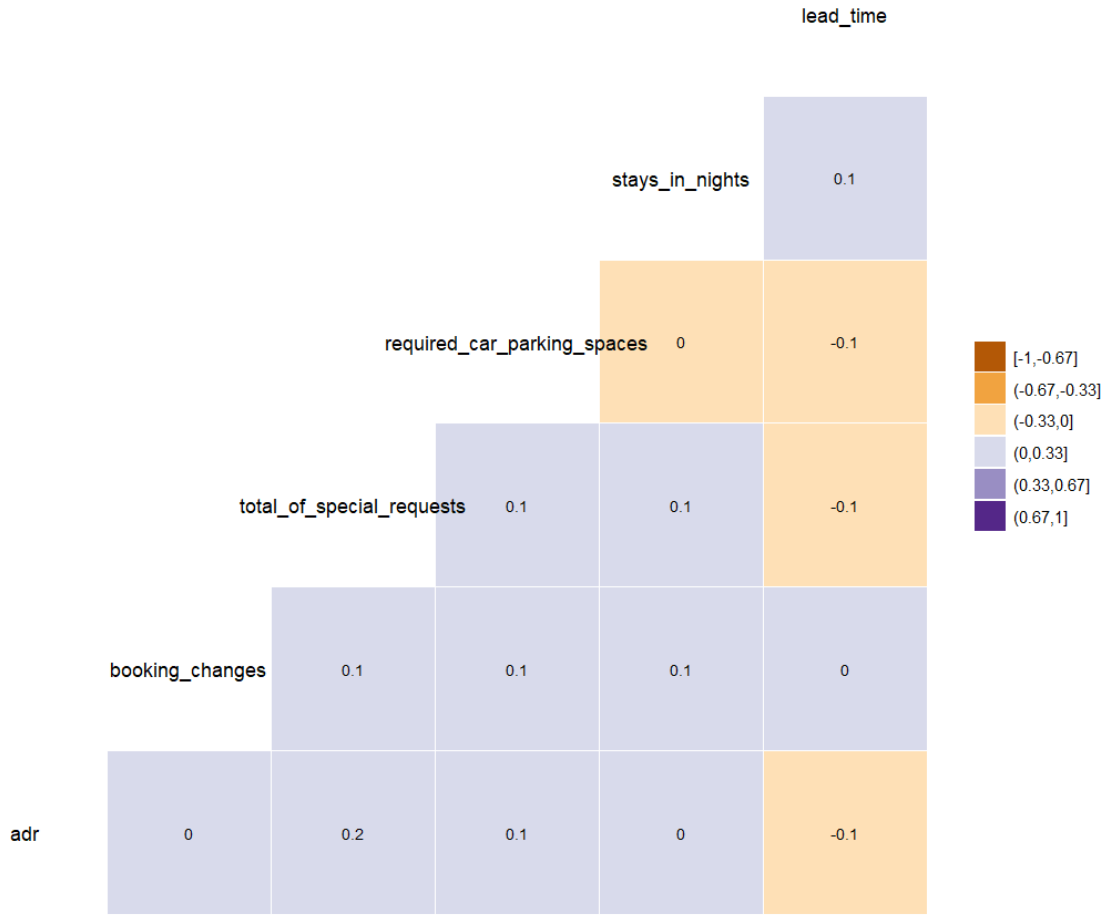
adr	95	94.7331	95.2669
required_car_parking_spaces	0	-0.00138742	0.00138742
total_of_special_requests	0	-0.004536511	0.004536511

3) Кореляція Спірмена(adr):



	rho	p_value	Normal	Basic	Percentile
booking_changes	0.01629069	2.537e-08	(0.0780, 0.4367)	(0.0859, 0.4544)	(0.0675, 0.4359)
total_of_special_requests	0.202532	< 2.2e-16	(0.2692, 0.5820)	(0.2795, 0.6019)	(0.2546, 0.5770)
required_car_parking_spaces	0.03503525	< 2.2e-16	(0.0976, 0.4525)	(0.1059, 0.4701)	(0.0869, 0.4510)
stays_in_nights	0.07815288	< 2.2e-16	(0.1423, 0.4876)	(0.1513, 0.5050)	(0.1314, 0.4851)
lead_time	-0.01770009	1.424e-09	(-0.4379, -0.0795)	(-0.4556, -0.0874)	(-0.4371, -0.0689)

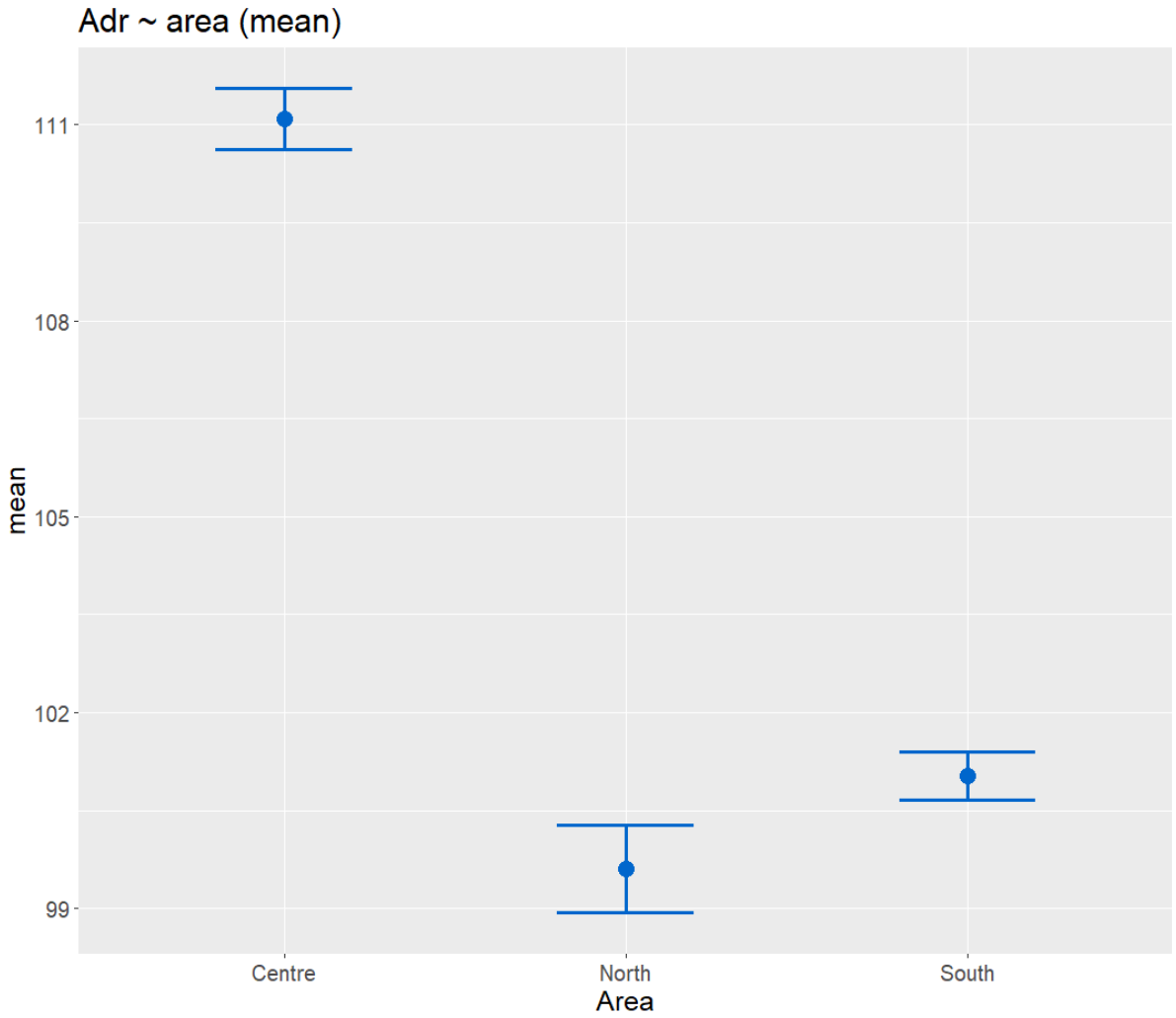
4) Кореляція Пірсона (adr):

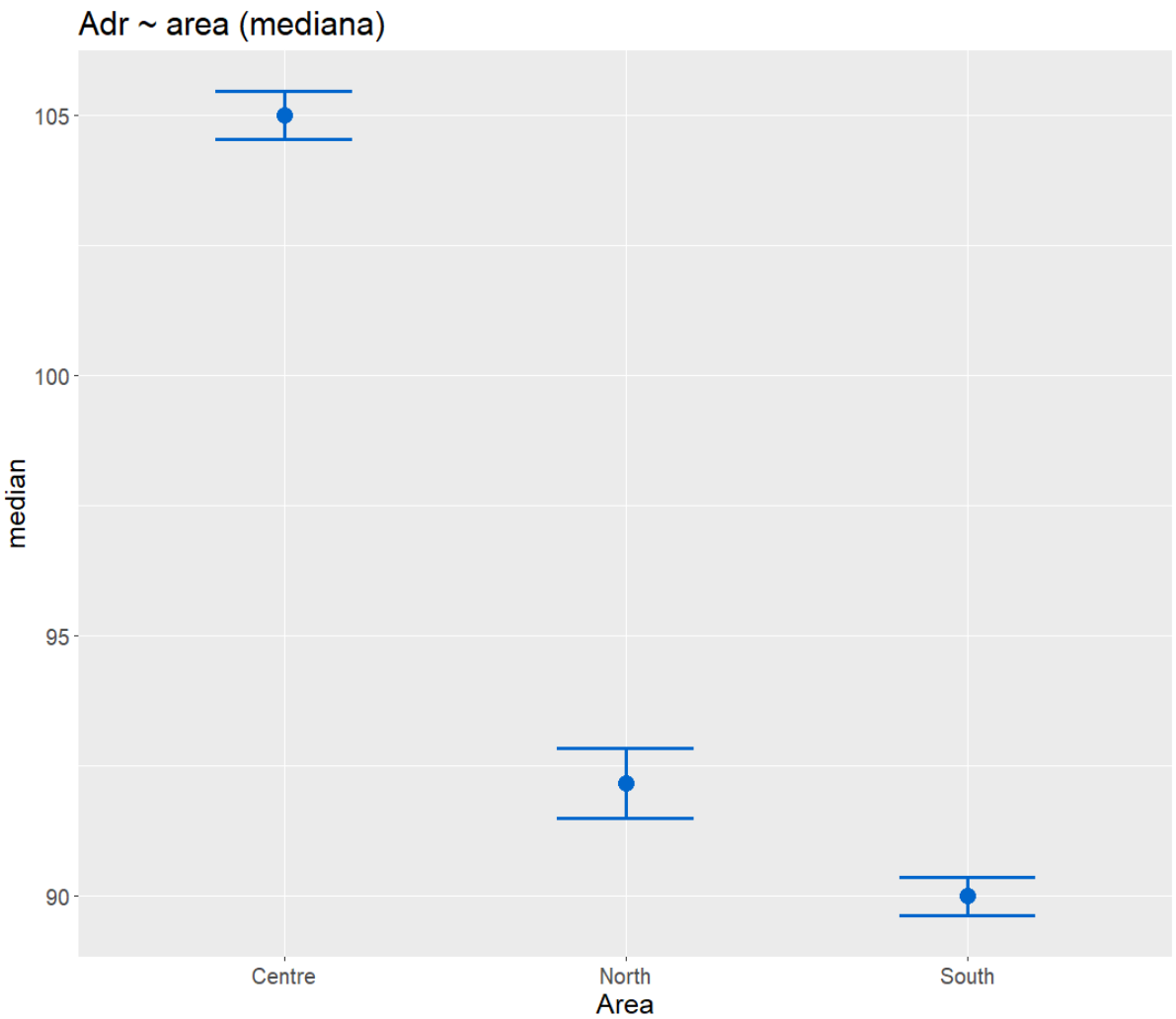


	coef	p_value	t-test statistic	conf.int	Normal	Basic	Percentile
booking_changes	0.0382 7124	< 2.2e-16	13.096	(0.03254640; 0.04399357)	(0.1009, 0.4551)	(0.1094, 0.4727)	(0.0902, 0.4536)
total_of_special_requests	0.1909 512	< 2.2e-16	66.516	(0.1854222; 0.1964681)	(0.2575, 0.5736)	(0.2677, 0.5933)	(0.2432, 0.5688)
required_car_parking_spaces	0.0641 3214	< 2.2e-16	21.974	(0.05842164; 0.06983844)	(0.1301, 0.4744)	(0.1371, 0.4787)	(0.1321, 0.4736)
stays_in_nights	0.0490 0504	< 2.2e-16	16.777	(0.04328522; 0.05472165)	(0.1121, 0.4640)	(0.1207, 0.4815)	(0.1013, 0.4621)
lead_time	-0.0946 2525	< 2.2e-16	-32.501	(-0.10030282; -0.08894151)	(-0.5007, -0.1593)	(-0.5182, -0.1685)	(-0.4978, -0.1481)

Дослідження

Питання 1: Гості з яких країн приносять найбільший прибуток?


Adr ~ area (mean)
111
108
105
102
99
mean
Centre
North
South
Area


Adr ~ area (mediana)
105
100
95
90
median
Centre
North
South
Area

Як бачимо, інтервали не перетинаються, найбільший прибуток приносять гості з центральних країн.

1) H_0: з північних країн гості приносять менший/такий самий прибуток як і з південних (mean(adr_N) - mean(adr_S) ≤ 0).

H_1: з північних країн гостей приносять більший прибуток ніж з південних (mean(adr_N) - mean(adr_S) > 0).

Тест Волда:

```
> mean_hat_s
[1] 99.60703
> p_value
[1] 0.9998732
> conf.int
[1] -2.066662       Inf
```

Як бачимо, p_value >>0.05, отже, в нас немає підстав відхилити H_0,

2) H_0: з північних країн гості приносять менший/такий самий прибуток як і з центральних (mean(adr_N) - mean(adr_C) ≤ 0).

H_1: з північних країн гостей приносять більший прибуток ніж з центральних (mean(adr_N) - mean(adr_C) > 0).

Тест Волда:

```
> mean_hat_s
[1] 99.60703
> p_value
[1] 1
> conf.int
[1] -12.16375       Inf
```

Знову маємо p_value >>0.05, отже в нас немає підстав відхилити H_0,

3) H_0: з центральних країн гості приносять менший/такий самий прибуток як і з південних (mean(adr_C) - mean(adr_S) ≤ 0).

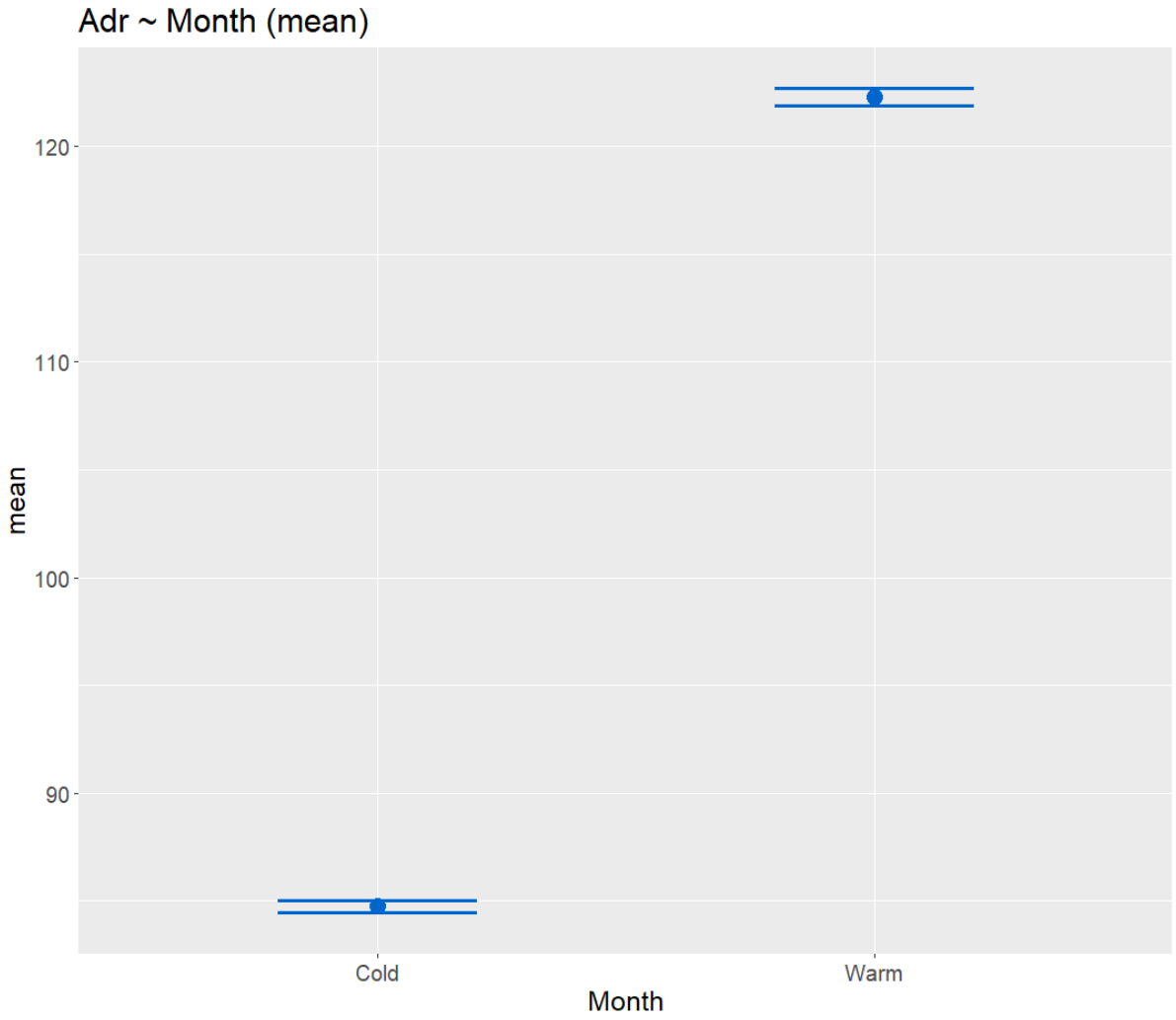
H_1: з центральних країн гостей приносять більший прибуток ніж з південних (mean(adr_C) - mean(adr_S) > 0).

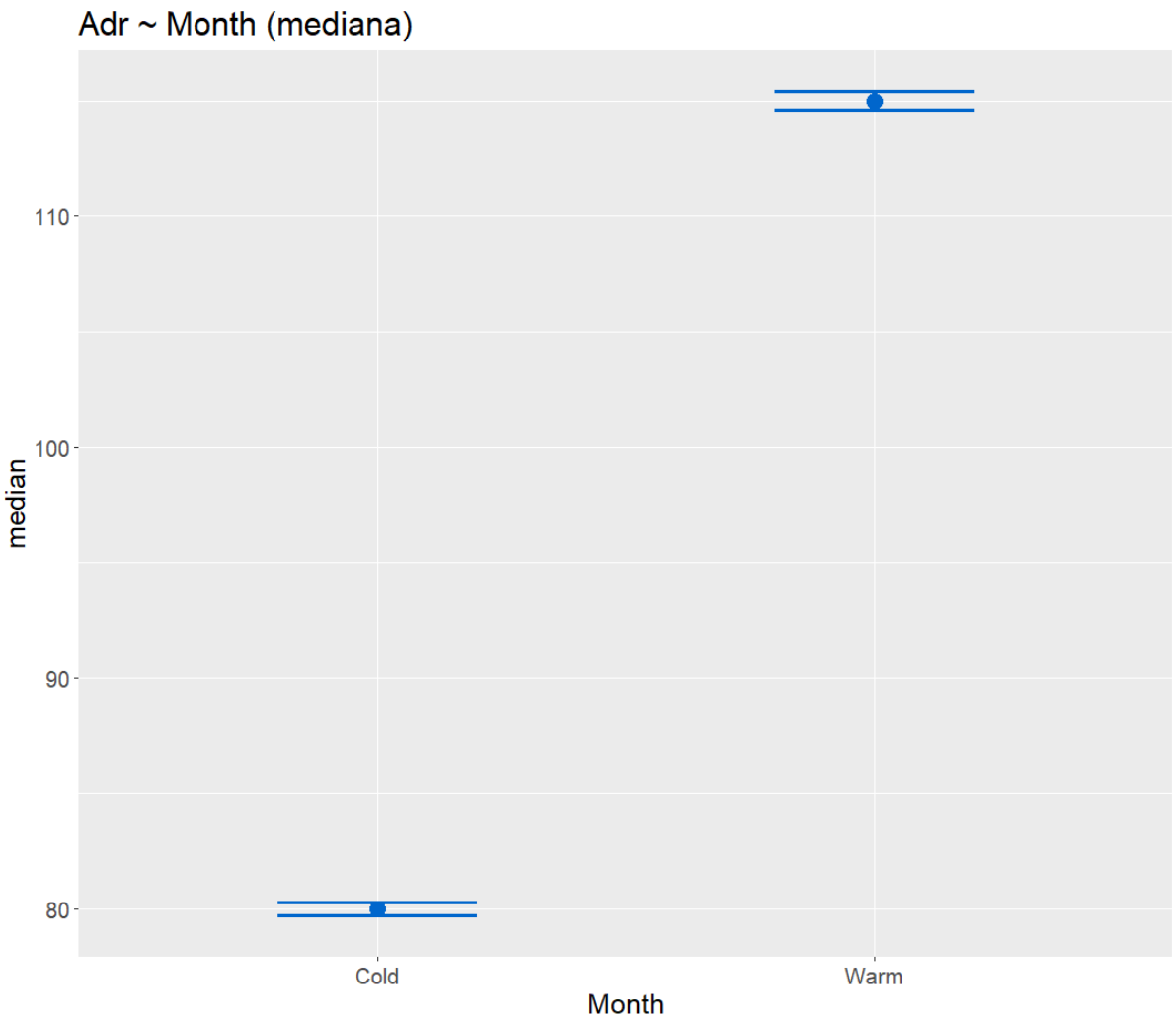
Тест Волда:

```
> mean_hat_s
[1] 111.0883
> p_value
[1] 5.154471e-245
> conf.int
[1] 9.560456      Inf
```

Як бачимо, p_value << 0.05, отже, в нас є підстави відхилити H_0.

Питання 2: Коли прибутковий сезон для готелів?





Як бачимо, інтервали не перетинаються, найменший прибуток у холодний сезон.

H_0: у теплі місяці (травень, червень, липень, серпень, вересень) прибуток менший/ такий самий, ніж в інші(mean(adr_W) - mean(adr_O) ≤ 0).

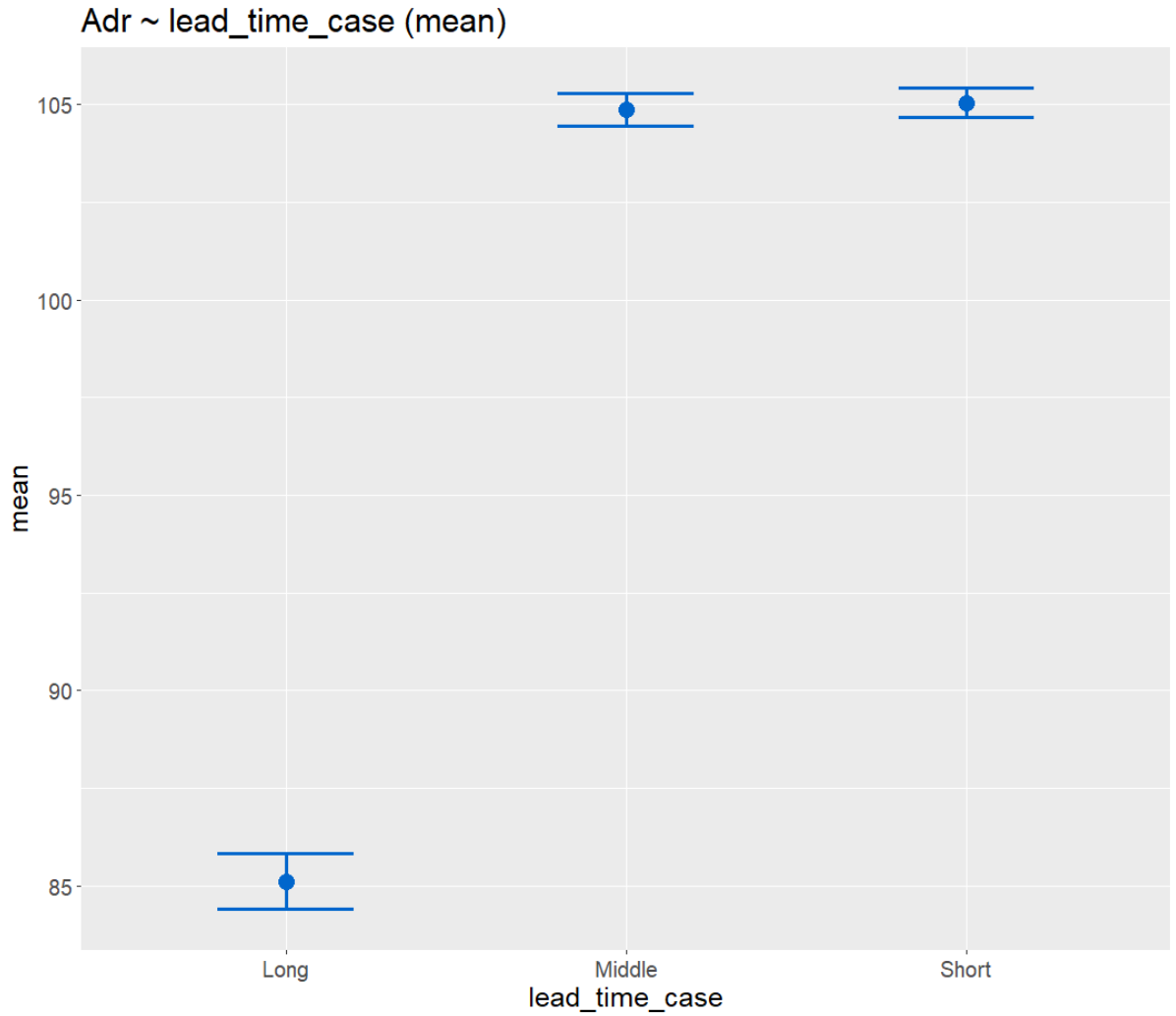
H_1: у теплі місяці (травень, червень, липень, серпень, вересень) прибуток більший, ніж в інші(mean(adr_W) - mean(adr_O) > 0).

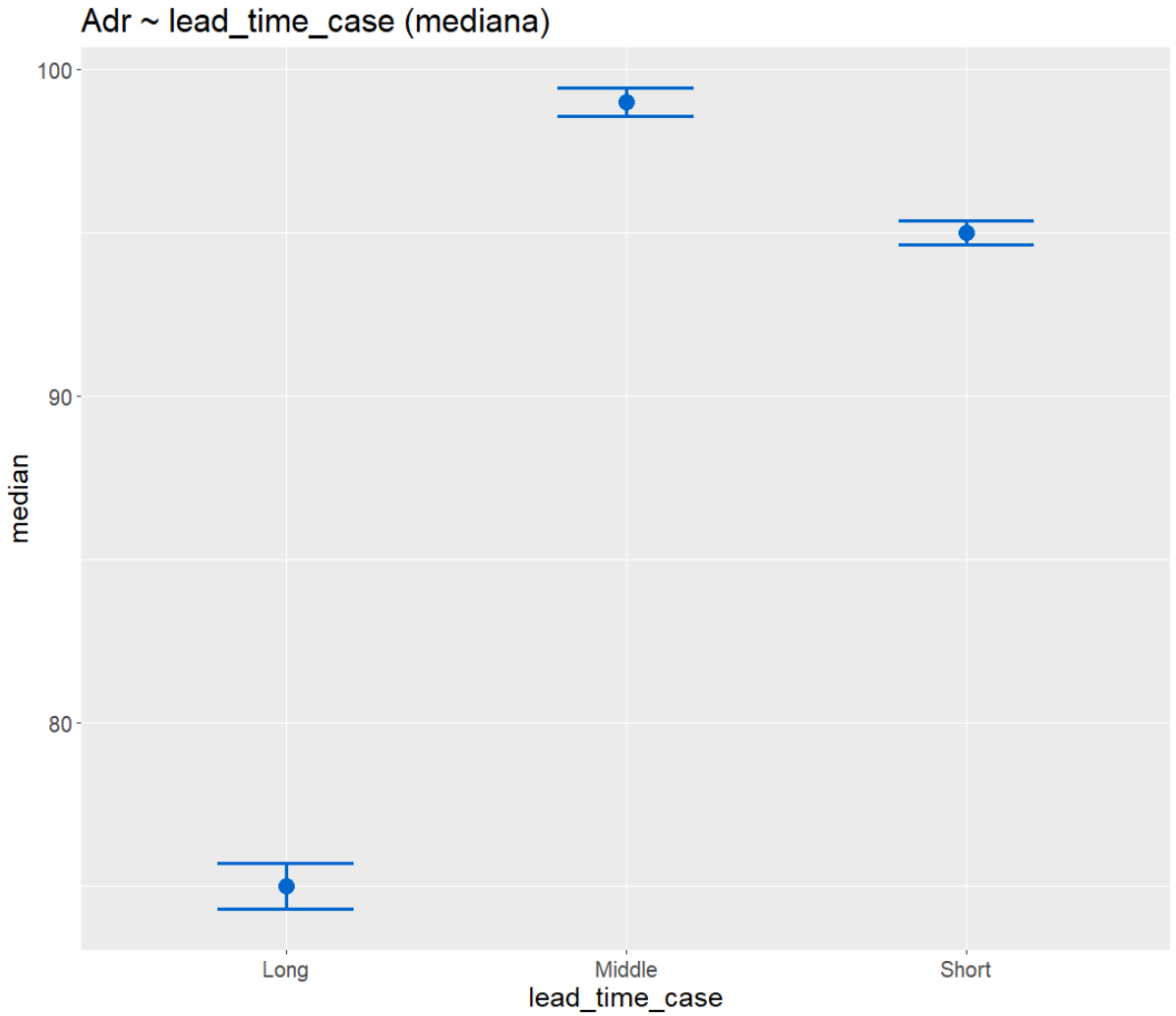
Тест Волда:

```
> mean_hat_s
[1] 111.0883
> p_value
[1] 5.154471e-245
> conf.int
[1] 9.560456      Inf
```

Як бачимо, p_value << 0.05, отже, в нас є підстави відхилити H_0.

Питання 3: За скільки часу до візиту вигідно планувати відпочинок?





1) H_0: вигідніше / так само за ціною планувати відпочинок за короткий час до візиту (mean(adr_S) - mean(adr_L) ≤ 0).

   H_1: вигідніше планувати відпочинок за довгий час до візиту (mean(adr_S) - mean(adr_L) > 0).

Тест Волда:

```
> mean_hat_s
[1] 111.0883
> p_value
[1] 5.154471e-245
> conf.int
[1] 9.560456      Inf
```

Як бачимо, p_value << 0.05, отже, в нас є підстави відхилити H_0.

2) H_0: вигідніше / так само за ціною планувати відпочинок за середній час до візиту (mean(adr_M) - mean(adr_L) ≤ 0).

   H_1: вигідніше планувати відпочинок за довгий час до візиту (mean(adr_M) - mean(adr_L) > 0).

Тест Волда:

```
> mean_hat_s
[1] 111.0883
> p_value
[1] 5.154471e-245
> conf.int
[1] 9.560456      Inf
```

Як бачимо, p_value << 0.05, отже, в нас є підстави відхилити H_0.

Питання 4: Чи пов'язані тип відвідувачів (дорослі/(+діти)) з типом харчування?

H_0: тип відвідувачів (дорослі/(+діти)) пов'язан із типом харчування.

H_1: такої залежності немає.

Проведемо тест χ^2:

```
> pchisq(q = T, df = 4, lower.tail = FALSE)
[1] 4.146018e-138
```

тест χ^2 – Пірсона:

```
> chisq.test(cont_tab1, correct = FALSE)

        Pearson's Chi-squared test

data:  cont_tab1
X-squared = 644.23, df = 4, p-value < 2.2e-16
```

Як бачимо, p_value << 0.05, отже, в нас є підстави відхилити H_0.

Питання 5: Чи пов'язані тип відвідувачів (дорослі (+ діти)) з типом номеру?

H_0: тип відвідувачів (дорослі/(+діти)) пов'язан із типом номеру.

H_1: такої залежності немає.

Проведемо тест χ^2:

```
> pchisq(q = T, df = 8, lower.tail = FALSE)
[1] 0
```

тест χ^2 – Пірсона:

```
> chisq.test(cont_tab2, correct = FALSE)

        Pearson's Chi-squared test

data:  cont_tab2
X-squared = 33587, df = 8, p-value < 2.2e-16
```

Як бачимо, p_value << 0.05, отже, в нас є підстави відхилити H_0.

Питання 6: Чи пов'язані тип відвідувачів (дорослі (+ діти)) з типом бронювання?

H_0: тип відвідувачів (дорослі/(+діти)) пов'язан із типом бронювання.

H_1: такої залежності немає.

Проведемо тест χ^2:

```
> pchisq(q = T, df = 4, lower.tail = FALSE)
[1] 1.141757e-196
```

тест χ^2 – Пірсона:

```
        Pearson's Chi-squared test

data:  cont_tab3
X-squared = 914.6, df = 4, p-value < 2.2e-16
```

Як бачимо, p_value << 0.05, отже, в нас є підстави відхилити H_0.

Але всі ці висновки мають дуже обмежену користь.

Висновки

1) adr має найбільшу кореляцію зі змінною lead_time, але навіть вона не є значущою.
2) Клієнти з північних країн виявилися не головним джерелом прибутку. Натомість можна сказати, що гості із центральних країн принесли найбільший прибуток.
3) Ми відхилили гіпотезу про те, що холодні місяці приносять такий самий прибуток як і теплі. Тобто прибуткові місяці теплі
4) Ми дізналися, що готелям найменш вигідно, коли відпочинок планується заздалегідь.
5) Також ми з'ясували, що тип відвідувачів (з дітьми/без них) не впливає на вибір харчування, типу номеру та способу бронювання.