Ekonometria – zajęcia laboratoryjne

Badanie, jakie czynniki wpływają na cenę zakwaterowania w serwisie Airbnb

Projekt końcowy

10.06.2019

Zapalski Mikołaj

II rok IiE 2018/19

nr indeksu #298102

Spis treści

1. Wstęp

Opis

Airbnb Inc., jest amerykańską firmą zajmującą się pośrednictwem na rynku internetowym i usługami brokerskimi z siedzibą w San Francisco. Członkowie mogą korzystać z usług w celu znalezienia lub zaoferowania zakwaterowania. Firma nie jest właścicielem żadnej z ofert nieruchomości. Działa jako pośrednik, otrzymując prowizje od każdej rezerwacji.[1]

Hipoteza

W mojej pracy postaram się zbadać, jaki wpływ na cenę wynajmu mają poszczególne czynniki, takie jak na przykład: typ zakwaterowania, lokalizacja, ilość łóżek, opinie o gospodarzu itp. Postaram się również sprawdzić czy istnieją jakieś nieoczywiste zależności oraz czy logiczne założenia, takie jak np. wzrost ceny przy wzroście ilości pomieszczeń zostaną potwierdzone.

Źródło danych

Dane zostały zebrane przez użytkownika o pseudonimie *SteveZheng* i są przez niego publicznie udostępnione na platformie kaggle.com.[2] Niestety nie wykorzystałem ich w całości ze względu na ograniczenia mojego komputera podczas pracy w programie gretl.

2. Przygotowanie i przedstawienie danych

Oczyszczenie danych

Oryginalne dane zawierały 29 kolumn i ponad 74 tyś. wierszy oraz zajmowały 32 MB. Miałem problem z zaimportowaniem ich do programu gretl w całości.

Zdecydowałem się na pozbycie się nieprzydatnych kolumn ze względu na ich unikalność, nieprzydatność w modelu lub brak spójności w formacie (w przypadku zmiennej zip-code). W tym celu wykorzystałem bibliotekę pandas[3] dla języka python (Rysunek 1).

```
In [1]:
import pandas as pd
import numpy as np

In [47]:
src = 'C:\\Users\\Mikolaj\\Desktop\\airbnb-price-prediction\\train.csv'
df = pd.read_csv(src)
df.columns

Out[47]: Index(['id', 'log_price', 'property_type', 'room_type', 'amenities',
       'accommodates', 'bathrooms', 'bed_type', 'cancellation_policy',
       'cleaning_fee', 'city', 'description', 'first_review',
       'host_has_profile_pic', 'host_identity_verified', 'host_response_rate',
       'host_since', 'instant_bookable', 'last_review', 'latitude',
       'longitude', 'name', 'neighbourhood', 'number_of_reviews',
       'review_scores_rating', 'thumbnail_url', 'zipcode', 'bedrooms', 'beds'],
      dtype='object')

In [48]:
df = df.drop(['id','name','description','first_review','last_review','thumbnail_url','zipcode'], axis=1)
df.to_csv('data.csv',index=False)
```

Rysunek 1

Zaimportowane dane miały 22 kolumn i 74 111 wierszy i prezentowały się następująco (Rysunek 2). Spora część zmiennych była w postaci tekstowej, więc zdecydowałem się na przekonwertowanie ich na zmienne binarne oraz usunąłem zmienne, które okazały się być zbędne. (Rysunek 3).

ID #	Nazwa zmiennej	Peł
0	const	
1	log_price	
2	property_type	
3	room_type	
4	amenities	
5	accommodates	
6	bathrooms	
7	bed_type	
8	cancellation_policy	
9	cleaning_fee	
10	city	
11	host_has_profile_pic	
12	host_identity_verified	
13	host_response_rate	
14	host_since	
15	instant_bookable	
16	latitude	
17	longitude	
18	neighbourhood	
19	number_of_reviews	
20	review_scores_rating	
21	bedrooms	
22	beds	

Rysunek 2

ID #	Nazwa zmiennej	Pełny opis zmiennej
0	const	
1	log_price	
2	property_type	
19	Dproperty_type_1	sztuczna zm. property_type = 'Apartment'
20	Dproperty_type_2	sztuczna zm. property_type = 'House'
21	Dproperty_type_3	sztuczna zm. property_type = 'Condominium'
3	room_type	
22	Droom_type_1	sztuczna zm. room_type = 'Entire home/apt'
23	Droom_type_2	sztuczna zm. room_type = 'Private room'
24	Droom_type_3	sztuczna zm. room_type = 'Shared room'
4	accommodates	
5	bathrooms	
6	bed_type	
25	Dbed_type_1	sztuczna zm. bed_type = 'Real Bed'
26	Dbed_type_2	sztuczna zm. bed_type = 'Futon'
27	Dbed_type_3	sztuczna zm. bed_type = 'Pull-out Sofa'
28	Dbed_type_4	sztuczna zm. bed_type = 'Couch'
29	Dbed_type_5	sztuczna zm. bed_type = 'Airbed'
7	cancellation_policy	
30	Dcancellation_policy_1	sztuczna zm. cancellation_policy = 'strict'
31	Dcancellation_policy_3	sztuczna zm. cancellation_policy = 'flexible'
8	city	
32	Dcity_1	sztuczna zm. city = 'NYC'
33	Dcity_2	sztuczna zm. city = 'SF'
34	Dcity_3	sztuczna zm. city = 'DC'
35	Dcity_4	sztuczna zm. city = 'LA'
36	Dcity_5	sztuczna zm. city = 'Chicago'
37	Dcity_6	sztuczna zm. city = 'Boston'
9	host_has_profile_pic	
38	Dhost_has_profile_pic_1	sztuczna zm. host_has_profile_pic = 't'
39	Dhost_has_profile_pic_2	sztuczna zm. host_has_profile_pic = 'f'
10	host_identity_verified	
40	Dhost_identity_verified_1	sztuczna zm. host_identity_verified = 't'
41	Dhost_identity_verified_2	sztuczna zm. host_identity_verified = 'f'
11	instant_bookable	
42	Dinstant_bookable_1	sztuczna zm. instant_bookable = 'f'
43	Dinstant_bookable_2	sztuczna zm. instant_bookable = 't'
12	latitude	
13	longitude	
14	neighbourhood	
15	number_of_reviews	
16	review_scores_rating	
17	bedrooms	
18	beds	

Rysunek 3

Opis zmiennych

log_price – logarytm ceny wynajmu zakwaterowania, rozkład tej zmiennej po zastosowaniu logarytmu wydaję się przystępniejszy do modelowania, w tym przypadku będzie to zmienna objaśniana

property_type – typ posiadłości taki jak apartament, łódka, zamek, kamper lub nawet wyspa. Unikalnych wartości było ponad 35, dlatego postanowiłem uwzględnić zmienne binarne jedynie dla tych najpopularniejszych, czyli **apartament**, **dom rodzinny** oraz **mieszkanie w bloku**

room_type – typ wynajmu, jaki dana posiadłość oferuję, czyli **całość** do naszej dyspozycji, **prywatny pokój** lub **pokój dzielony**

amenities – udogodnienia, zdecydowałem się na pominięcie tej zmiennej ze względu na problematyczność jej formatu. Przykładowa wartość wyglądała następująco:

{"Wireless Internet","Air conditioning",Kitchen,Heating,"Family/kid friendly",Essentials,"Hair dryer",Iron,"translation missing: en.hosting_amenity_50"}

accommodates – dozwolona maksymalna ilość osób

bathrooms – liczba łazienek w/przy obiekcie

bed_type – rodzaj łóżka, wśród obserwacji występowało 5 różnych wartości: zwykłe łóżko, **futon** (tradycyjne japońskie łóżko), **rozkładana sofa**, **kanapa** oraz **dmuchany materac**

cancellation_policy – ustalenia dot. anulowania rezerwacji, w modelu znaczenie miały jedynie wartości skrajne – **rygorystyczne** oraz **elastyczne**

cleaning_fee – opłata za sprzątanie, zmienna została pominięta ze powodu braku istotności, co jest logiczne ze względu na to, że serwis airbnb nie wlicza tej opłaty do ceny, tylko podaje ją oddzielnie

city – miasto, w którym posiadłość się znajduję, dane zebrane zostały dla 6 miast (Nowy Jork, San Francisco, Washington DC, Los Angeles, Chicago, Boston). W przypadku, gdy chcielibyśmy wykorzystać nasz model do predykcji w innym mieście dane mógłby być przekłamane, ze względu na to, że model *uczył* się na tych miastach

host_has_profile_pic – zmienna binarna, informuję czy gospodarz miał ustawiony awatar

host_identity_verified – zmienna binarna, informuję czy gospodarz ma status potwierdzonego użytkownika

host_response_rate – informacja o tym jak szybko gospodarz odpisywał na zapytania

host_since – data wystawienia pierwszej oferty przez gospodarza

instant_bookable – zmienna binarna, informuję o możliwości natychmiastowej rezerwacji

latitude – szerokość geograficzna

longitude – wysokość geograficzna

neighbourhood – nazwa dzielnicy, liczba unikalnych dzielnic przekroczyła 600, więc nie zdecydowałem się na konwersję na zmienne binarne

numer_of_reviews – liczba recenzji

review_scores_rating – średnia ocen od recenzentów

bedrooms – liczba sypialni w obiekcie

beds – liczba łóżek w posiadłości

4. Model

Pierwsza wersja modelu

Przy wyborze zmiennych do finalnego modelu kierowałem się metodą krokową wsteczną. Pierwszy model z wszystkimi zmiennymi miał następującą postać (Rysunek 4). Następnie wykluczyłem zmienne nieistotne, oraz te, które wykazywały na współliniowość.

	współczynnik	błąd standardowy	t-Studenta	wartość p	
const	-73,8052	1,79163	-41,19	0,0000	***
Dproperty_type_1	-0,0258777	0,00686066	-3,772	0,0002	***
Dproperty_type_2	-0,0541767	0,00742281	-7,299	2,95e-013	***
Dproperty_type_3	0,0821789	0,0116927	7,028	2,12e-012	***
Droom_type_1	1,12915	0,0126547	89,23	0,0000	***
Droom_type_2	0,523713	0,0124526	42,06	0,0000	***
accommodates	0,0792423	0,00172819	45,85	0,0000	***
bathrooms	0,135268	0,00408618	33,10	1,29e-237	***
Dbed_type_1	0,0720718	0,0257847	2,795	0,0052	***
Dbed_type_2	-0,00316373	0,0314551	-0,1006	0,9199	
Dbed_type_3	0,0701586	0,0325233	2,157	0,0310	**
Dbed_type_4	0,0909081	0,0439524	2,068	0,0386	**
Dcancellation_~_1	0,0294323	0,00438728	6,709	1,99e-011	***
Dcancellation_~_3	-0,0117202	0,00562902	-2,082	0,0373	**
Dcleaning_fee_1	-0,0142097	0,00515680	-2,756	0,0059	***
Dcity_1	-2,77848	0,0755512	-36,78	5,44e-292	***
Dcity_2	-51,5658	1,06024	-48,64	0,0000	***
Dcity_3	-5,82219	0,157248	-37,03	7,04e-296	***
Dcity_4	-47,4395	1,00117	-47,38	0,0000	***
Dcity_5	-17,1733	0,338653	-50,71	0,0000	***
Dhost_has_prof~_1	-0,130475	0,0477546	-2,732	0,0063	***
Dhost_identity~_1	0,0189699	0,00432821	4,383	1,17e-05	***
host_response_ra~	2,12407e-05	0,000223020	0,09524	0,9241	
Dinstant_booka~_1	0,0430833	0,00417559	10,32	6,21e-025	***
latitude	0,102030	0,0271129	3,763	0,0002	***
longitude	-1,01836	0,0203572	-50,02	0,0000	***
number_of_reviews	-0,000316188	4,43769e-05	-7,125	1,06e-012	***
review_scores_ra~	0,00604730	0,000263564	22,94	7,13e-116	***
bedrooms	0,152677	0,00352320	43,33	0,0000	***
beds	-0,0474695	0,00264949	-17,92	1,51e-071	***

Średn.aryt.zm.zależnej	4,751302	Odch.stand.zm.zależnej	0,675706
Suma kwadratów reszt	7899,910	Błąd standardowy reszt	0,406717
Wsp. determ. R-kwadrat	0,637919	Skorygowany R-kwadrat	0,637699
F(29, 47757)	2901,343	Wartość p dla testu F	0,000000
Logarytm wiarygodności	-24800,85	Kryt. inform. Akaike'a	49661,70
Kryt. bayes. Schwarza	49924,94	Kryt. Hannana-Quinna	49744,33

Wyłączając stałą, największa wartość p jest dla zmiennej 12 (host_response_rate)

Rysunek 4

Metoda krokowa wsteczna

Po usunięciu wszystkich nieistotnych zmiennych z modelu, uzyskałem następujący wynik (Rysunek 5). Ilość wykorzystanych obserwacji wynosi 56 989. Ze względu na niekompletność obserwacji pominięto 17 122 obserwacji (~ 21%).

	współczynnik	błąd standardowy	t-Studenta	wartość p	
const	-76,7282	1,73218	-44,30	0,0000	***
Dproperty_type_1	-0,0367130	0,00661487	-5,550	2,87e-08	***
Dproperty_type_2	-0,0538455	0,00719137	-7,488	7,12e-014	***
Dproperty_type_3	0,0788692	0,0111780	7,056	1,74e-012	***
Droom_type_1	1,07649	0,0118469	90,87	0,0000	***
Droom_type_2	0,479445	0,0117082	40,95	0,0000	***
accommodates	0,0817970	0,00165601	49,39	0,0000	***
bathrooms	0,133634	0,00390263	34,24	2,23e-254	***
Dbed_type_1	0,0699381	0,0136626	5,119	3,08e-07	***
Dbed_type_3	0,0660846	0,0229306	2,882	0,0040	***
Dbed_type_4	0,106827	0,0356721	2,995	0,0027	***
Dcancellation_~_1	0,0294547	0,00410694	7,172	7,48e-013	***
Dcancellation_~_3	-0,0164116	0,00492131	-3,335	0,0009	***
Dcity_1	-2,76006	0,0725838	-38,03	0,0000	***
Dcity_2	-52,1748	1,02839	-50,73	0,0000	***
Dcity_3	-5,75645	0,151020	-38,12	0,0000	***
Dcity_4	-47,8218	0,969600	-49,32	0,0000	***
Dcity_5	-17,4096	0,328777	-52,95	0,0000	***
Dhost_has_prof~_1	-0,0983354	0,0427414	-2,301	0,0214	**
Dhost_identity~_1	0,0170472	0,00397242	4,291	1,78e-05	***
Dinstant_booka~_1	0,0439100	0,00396415	11,08	1,74e-028	***
latitude	0,146071	0,0257792	5,666	1,47e-08	***
longitude	-1,03406	0,0197701	-52,30	0,0000	***
number_of_reviews	-0,000302436	4,35387e-05	-6,946	3,79e-012	***
review_scores_ra~	0,00560491	0,000222573	25,18	3,62e-139	***
bedrooms	0,146779	0,00332720	44,11	0,0000	***
beds	-0,0459529	0,00255799	-17,96	5,85e-072	***

Średn.aryt.zm.zależnej	4,750101	Odch.stand.zm.zależnej	0,669091
Suma kwadratów reszt	9619,533	Błąd standardowy reszt	0,410946
Wsp. determ. R-kwadrat	0,622949	Skorygowany R-kwadrat	0,622777
F(26, 56962)	3619,629	Wartość p dla testu F	0,000000
Logarytm wiarygodności	-30170,39	Kryt. inform. Akaike'a	60394,78
Kryt. bayes. Schwarza	60636,45	Kryt. Hannana-Quinna	60470,03

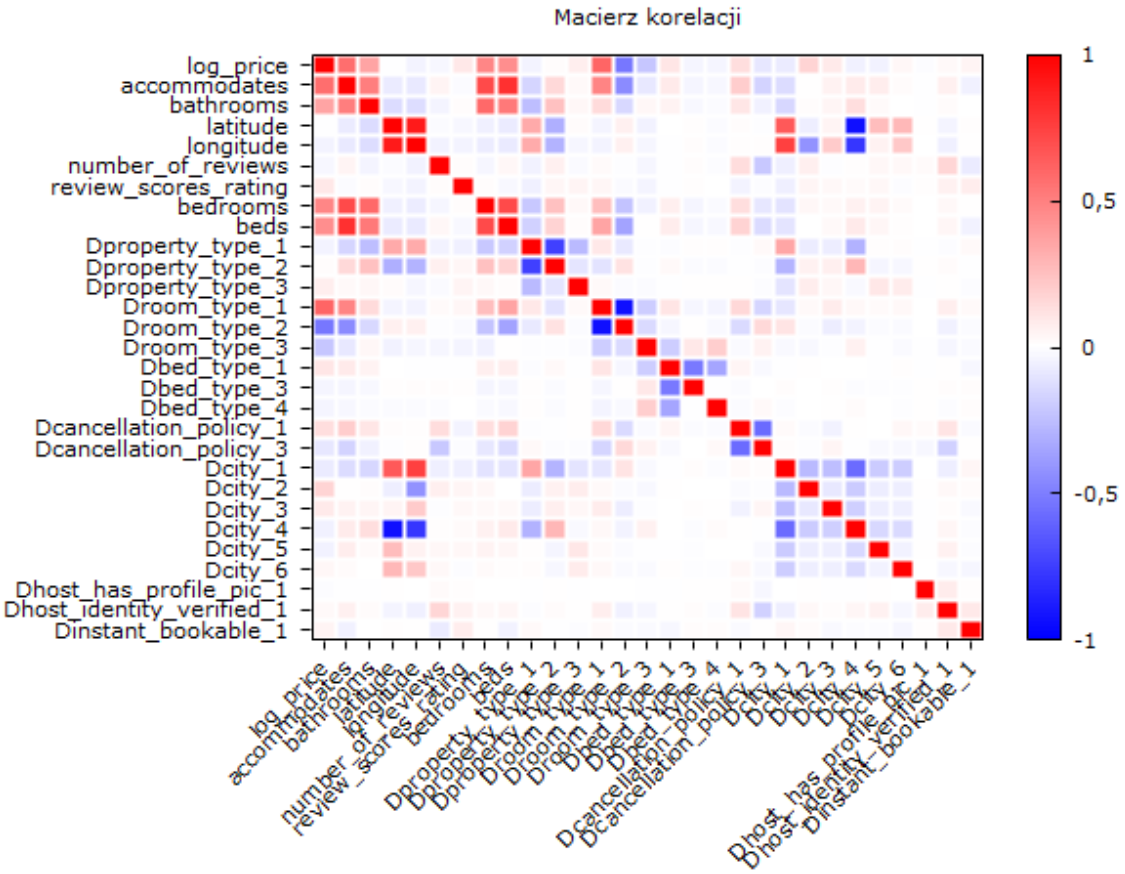
Rysunek 5

Metoda Helwiga

Przy wyborze zmiennych do modelu zastosowałem metodę Helwiga w celu odnalezienia potencjalnej lepszej kombinacji niż ta wybrana przeze mnie. Jednakże po upływie 2 dni przerwałem wykonywanie skryptu, gdyż zajmowało to za dużo czasu. Przez ten czas przeanalizowano 117 696 kombinacji, co stanowi zaledwie około 0,02% wszystkich możliwych kombinacji. (2^n - 1 w przypadku 29 zmiennych wynosi ponad 536 mln).

Korelacja

Mapa cieplna korelacji wszystkich zmiennych ze sobą wygląda następująco (Rysunek 6).



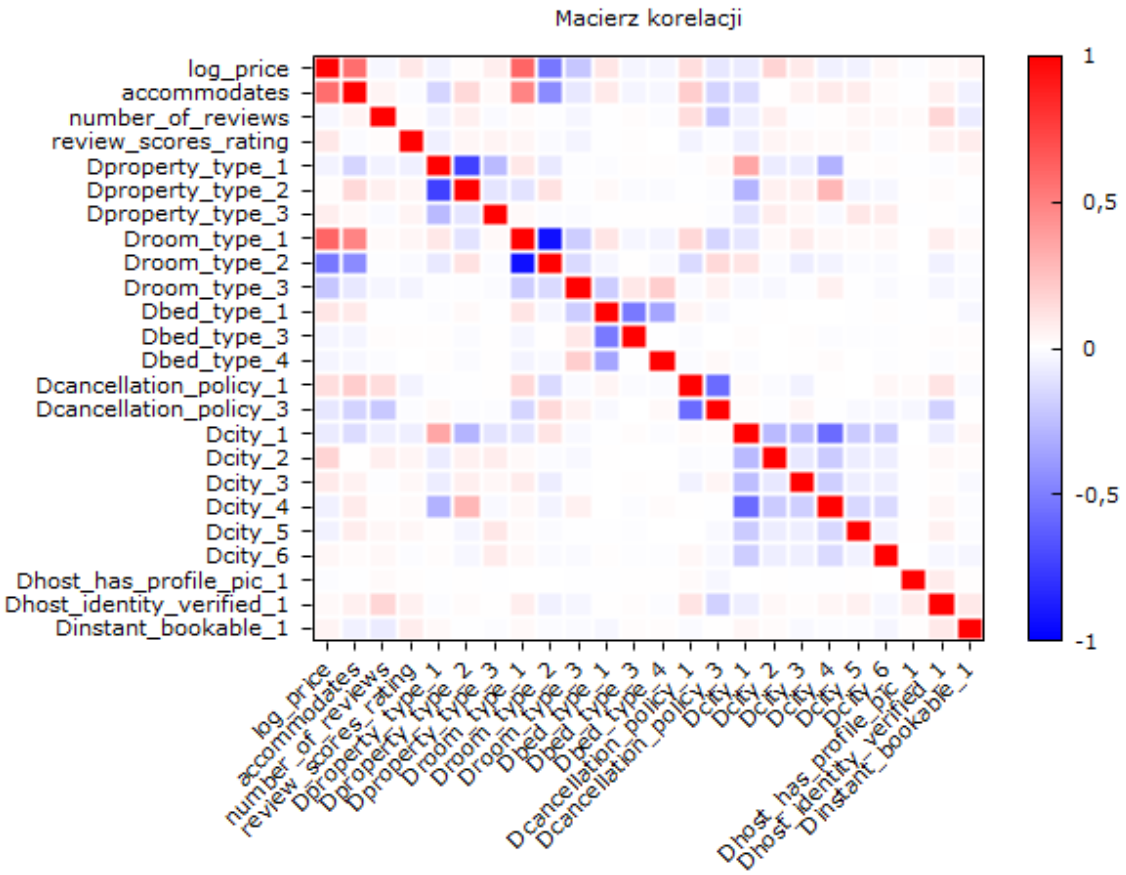
Rysunek 6

Możemy zauważyć, że duża korelacja występuje dla zmiennych określających miasto oraz długość i szerokość geograficzną. Na tej podstawie zdecydowałem się na usunięcie zmiennych **latitude** oraz **longitude** z modelu.

Teraz popatrzmy na sytuację ze zmiennymi **accommodates, bathrooms, bedrooms oraz beds**. Wszystkie te zmienne informują o tym ile osób może być zakwaterowana w danym miejscu, aby mieszkało się komfortowo. Zdecydowałem się na pozostawienie jedynie zmiennej **accommodates**, ponieważ jest ona najbardziej szczegółowa z całej czwórki oraz ma najwyższą korelację ze zmienną objaśnianą **log_price** (0,5676).

Widać też zależność zmiennej accommodates z typem zakwaterowania. Jest to wytłumaczalne w praktyce, tym, że jeśli szukamy noclegu dla 10 osób, mało prawdopodobnym będzie znalezienie pokoju, a wynajmem całego domku jest samo nasuwającym się logicznym rozwiązaniem. Nie zdecydowałem się na wyeliminowanie tych zmiennych, ze względu na odmienny typ problemu, który starają się wskazać.

Po zastosowaniu wyżej wymienionych operacji mapa cieplna wygląda następująco (Rysunek 7). Nadal można zauważyć wysoką korelację przy zmiennych, które zostały przerobione na kategoryczne. Prawdopodobnie jest to efekt tego, że przy tworzeniu nie została usunięta najczęściej/najmniej występująca wartość.



Rysunek 7

Teraz popatrzmy na korelację zmiennych X z Y. Informacja o tym, że gospodarz posiada zdjęcie profilowe nie ma związku na cenę, dlatego usuwam ją z modelu. Pozostałe zmienne również mają niską korelację, ale uznałem, że je zachowam, ze względu na już i tak stosunkowo niski współczynnik R^2 modelu.

Ostateczna wersja modelu

	współczynnik	błąd standardowy	t-Studenta	wartość p	
const	2,92007	0,0300541	97,16	0,0000	***
Droom_type_1	1,04898	0,0123501	84,94	0,0000	***
Droom_type_2	0,477886	0,0122419	39,04	0,0000	***
Dcity_1	-0,0423684	0,00877427	-4,829	1,38e-06	***
Dcity_2	0,259653	0,0103322	25,13	1,31e-138	***
Dcity_3	-0,128998	0,0107733	-11,97	5,33e-033	***
Dproperty_type_3	0,105467	0,0118657	8,888	6,37e-019	***
accommodates	0,120814	0,00101635	118,9	0,0000	***
number_of_reviews	-0,000560191	4,58488e-05	-12,22	2,74e-034	***
review_scores_ra~	0,00639447	0,000235741	27,12	5,25e-161	***
Dproperty_type_1	-0,0330653	0,00699013	-4,730	2,25e-06	***
Dproperty_type_2	-0,0127899	0,00757331	-1,689	0,0913	*
Dbed_type_1	0,0724192	0,0144666	5,006	5,58e-07	***
Dbed_type_3	0,0509039	0,0243227	2,093	0,0364	**
Dbed_type_4	0,106758	0,0378174	2,823	0,0048	***
Dcancellation_~_1	0,0454008	0,00435243	10,43	1,88e-025	***
Dcancellation_~_3	-0,0111018	0,00522429	-2,125	0,0336	**
Dcity_4	-0,181821	0,00901914	-20,16	4,57e-090	***
Dcity_5	-0,333340	0,0113365	-29,40	1,25e-188	***
Dhost_identity~_1	0,0208397	0,00420752	4,953	7,33e-07	***
Dinstant_booka~_1	0,0717117	0,00419426	17,10	2,25e-065	***

Średn.aryt.zm.zależnej	4,749340	Odch.stand.zm.zależnej	0,668891
Suma kwadratów reszt	10936,64	Błąd standardowy reszt	0,437161
Wsp. determ. R-kwadrat	0,573007	Skorygowany R-kwadrat	0,572858
F(20, 57227)	3839,820	Wartość p dla testu F	0,000000
Logarytm wiarygodności	-33850,84	Kryt. inform. Akaike'a	67743,68
Kryt. bayes. Schwarza	67931,74	Kryt. Hannana-Quinna	67802,22

Rysunek 8

Opis

Współczynnik determinacji R^2 wynosi 57,3%, co oznacza, że model wyjaśnia 57,3% zmienności badanego zjawiska.

Wartość p dla testu F (H_0: Wszystkie współczynniki równe 0) wynosi 0. Odrzucam hipotezę o zerowości wszystkich współczynników.

Średnia wartość logarytmu ceny wynosi 4,7493, a odchylenie standardowe wynosi 0,6689. Z tego wynika, że średnia cena wynajmu wynosi $e^{4,7493}$, czyli 115.50 Dolarów.

Przy predykcji model średnio pomyli się o 0,4371 (logarytmu ceny).

Test normalności rozkładu reszt

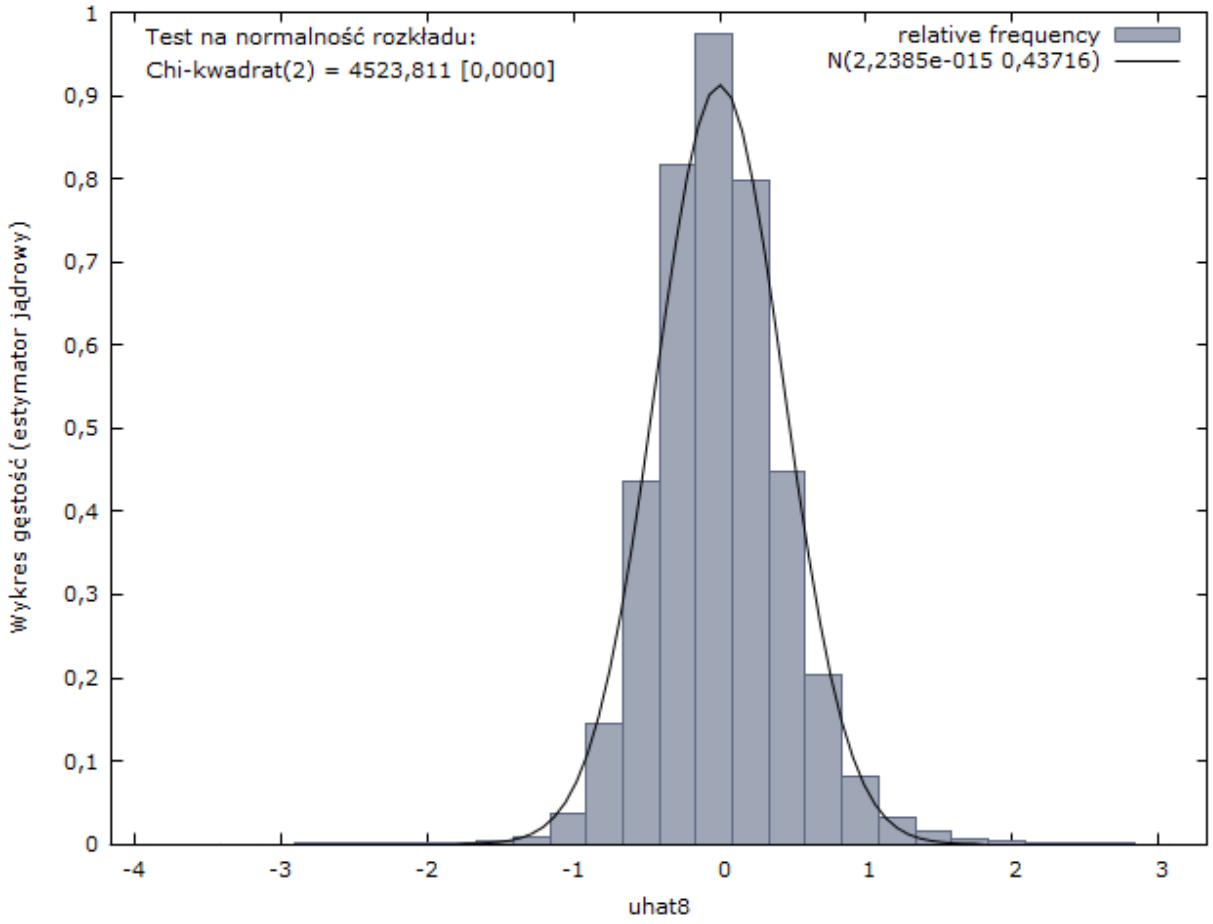
H_0: dystrybuanta empiryczna posiada rozkład normalny (Test Doornika-Hansena)

```
Test na normalność rozkładu reszt -
  Hipoteza zerowa: składnik losowy ma rozkład normalny
  Statystyka testu: Chi-kwadrat(2) = 4523,81
  z wartością p = 0
```

Rysunek 9

p-value wynosi 0 => odrzucam hipotezę zerową o rozkładzie normalnym reszt

Biorąc pod uwagę bardzo dużą liczbę obserwacji n=57 248 i powołując się na twierdzenia asymptotyczne, mogę stwierdzić, że reszty mają rozkład normalny. Rozkład możemy zaobserwować na ilustracji (Rysunek 10).



Rysunek 10

5. Diagnostyka modelu

Interpretacja parametrów

Podczas interpretacji należy pamiętać o tym, że zmienna objaśniana jest w postaci logarytmu ceny. W takim wypadku parametry interpretujemy w sposób przedstawiony na prezentacji pana Jakuba Mućka „Ekonometria - Model nieliniowe i funkcja produkcji" (strona 13)[4] .

ln y = α + βx.

Wzrost X o jednostkę odpowiada wzrostowi y o 100β % jednostek.

Nazwa zmiennej	**Współczynnik**	**Interpretacja**
Droom_type_1	1,04898	Jeśli wynajmujemy cały dom, to cena wzrasta o 104,90%
Droom_type_2	0,477886	Jeśli wynajmujemy pokój, to cena wzrasta o 47,79%
Dcity_1	−0,0423684	Jeśli lokalizacja to Nowy Jork, cena spada o 4,24%
Dcity_2	0,259653	Jeśli lokalizacja to San Francisco, cena wzrasta o 25,97%
Dcity_3	−0,128998	Jeśli lokalizacja to Washington DC, cena spada o 12,90%
Dcity_4	−0,181821	Jeśli lokalizacja to Los Angeles, cena spada o 18,18%
Dcity_5	−0,333340	Jeśli lokalizacja to Chicago, cena spada o 33,33%
Dproperty_type_1	−0,0330653	Jeśli wynajmujemy mieszkanie, to cena spada o 3,33%
Dproperty_type_2	−0,0127899	Jeśli wynajmujemy dom, to cena spada o 1,29%
Dproperty_type_3	0,105467	Jeśli wynajmujemy apartament, to cena wzrasta o 10,55%

accommodates	0,120814	Wzrost liczby gości o jednostkę odpowiada wzrostowi ceny o 12,08%
number_of_reviews	−0,000560191	Wzrost liczby ocen o jednostkę odpowiada spadkowi ceny o 0,06%
review_scores_rating	0,00639447	Wzrost średniej oceny o jednostkę odpowiada wzrostowi ceny o 0,64%
Dbed_type_1	0,0724192	Jeśli łóżko jest zwykłego typu, cena wzrasta o 7,24%
Dbed_type_3	0,0509039	Jeśli łóżko jest rozkładaną sofą, cena wzrasta o 5,09%
Dbed_type_4	0,106758	Jeśli oferowany nocleg jest na kanapie, to cena wzrasta o 10,67%
Dcancellation_policy_1	0,0454008	Jeśli ustalenie dot. anulowania rezerwacji są rygorystyczne, cena wzrasta o 4,54%
Dcancellation_policy_3	−0,0111018	Jeśli ustalenie dot. anulowania rezerwacji są elastyczne, cena spada o 1,11%
Dhost_identity_verified_1	0,0208397	Jeśli gospodarz jest zatwierdzonym użytkownikiem, cena wzrasta o 2,08%
Dinstant_bookable_1	0,0717117	Jeśli istnieje możliwość natychmiastowej rezerwacji, cena wzrasta o 7,17%

Analiza

Największy wpływ na cenę ma to, czy wynajmujemy dom lub pokój. Sytuacja standardowa, czyli taka gdzie wynajmujemy pokój dzielony odpowiada cenie 0 $. Zaskakującym może być fakt, że gdy jako lokalizację przyjmiemy NYC, to cena spada. Może to wynikać z tego, że powierzchnie mieszkalne w Nowym Jorku są bardzo małe i dlatego ich średnia cena jest niższa, niż w porównywanych miastach. Niestety portal airbnb

nie publikuje danych na temat powierzchni użytkowych, co mogłoby w tym modelu być dość istotnym czynnikiem. Każdy kolejny gość to podwyżka ceny o ok. 12%, co również wydaję się być znaczącym składnikiem. Typ wynajmowanej nieruchomości również zdaję się odzwierciedlać logiczne hipotezy. Apartament kosztuje nas 11% więcej, z kolei za zwykłe mieszkanie zapłacimy o 3% mniej, a za zwykły dom cena spadnie o trochę ponad 1%. Zaskakującym faktem jest to, że pośród zmiennych opisujących typ łóżka, najwięcej zapłacimy za rozkładaną sofę (cena wzrasta o 10%!). Wytłumaczenie, jakie przyszło mi do głowy jest takie, że oferty zawierające rozkładaną sofę, nie zawierają w sobie wielu innych udogodnień i z reguły łączą się z liczbą gości równą 1 oraz pokojem dzielonym, przez co model musi „nadrobić" wartość ceny. Interpretacja zmiennych dotyczących polityki rezerwacji wydaję się poprawna, jeśli gospodarz ma doświadczenie, duży ruch w jego nieruchomości lub jest ona wysokiej klasy – zastosuje rygorystyczną politykę rezerwacji (np. nie zostanie zwrócona zaliczka), przez co cena wzrośnie, oraz w przeciwnym przypadku – cena spadnie przy bardziej łagodnej polityce. Zatwierdzenie profilu w portalu również zwiększa cenę, co było spodziewane, podobnie jak możliwość natychmiastowej rezerwacji. Jeżeli chodzi o oceny użytkowników, to wzrost oceny o 1 punkt powoduję wzrost zaledwie, o 0,65%, lecz należy mieć na uwadze, że skala ocen mieści się między 0-100. Interesujący jest również przypadek „Wzrost liczby ocen o jednostkę odpowiada spadkowi ceny o 0,06%", może to wynikać z tego, że są przypadki, w których oferta ma 1-2 oceny bardzo pozytywne, co daje bardzo zawyżony obraz sytuacji.

Efekt katalizy

Macierz pomocnicza wygenerowana przez skrypt (1- gdy zmienna jest katalizatorem) (Rysunek 11).

Niefortunnie, wynikiem skryptu była informacja, o tym, że wszystkie zmienne są katalizatorami (Rysunek 12). Nie wykluczam, że może być to błąd w moim skrypcie.

```
? kat
kat (20 x 20)

  0   0   0   1   1   1   1   1   1   1   1   0   0   1   0   1   0   1   1   1
  1   1   1   1   0   1   0   1   1   1   0   1   1   0   1   1   1   1   1   1
  0   0   0   0   1   1   1   0   0   1   1   1   0   1   1   0   0   1   0   0
  0   1   1   1   0   1   1   0   0   1   1   1   1   1   1   0   1   1   0   0
  1   1   1   1   0   0   0   1   0   0   1   0   0   0   1   1   1   1   1   1
  1   0   1   0   0   0   0   1   1   1   0   1   0   0   1   1   0   1   1   0
  1   1   1   0   0   0   0   0   1   1   0   1   1   0   1   1   1   1   1   1
  0   1   1   0   0   0   1   1   0   1   0   0   1   0   0   0   0   0   1   1
  0   0   0   1   0   1   1   1   0   1   1   1   0   1   0   0   1   1   0   0
  0   1   0   1   1   0   0   1   0   1   0   1   1   0   0   1   1   0   1   1
  0   0   0   1   0   1   1   0   0   0   1   0   0   1   0   1   0   1   1   0
  0   0   1   0   0   1   1   1   1   0   1   0   0   0   1   1   1   1   0   0
  1   1   1   1   1   1   0   1   1   1   0   1   1   0   0   0   1   0   0   1
  0   0   0   1   1   1   1   0   0   0   1   1   0   1   0   0   0   1   1   1
  0   0   1   1   1   1   1   1   0   1   1   1   1   1   0   0   0   1   0   0
  1   0   0   1   1   1   1   1   0   0   0   1   1   1   0   0   1   1   0   0
  1   1   1   1   0   1   0   0   0   1   0   0   1   0   1   0   1   0   1   1
  1   0   0   1   0   0   0   0   0   0   1   0   0   1   0   0   0   1   0   1
  0   0   0   1   1   1   1   0   0   0   1   0   1   1   0   0   0   1   0   0
  1   0   0   1   1   0   0   0   0   0   1   0   0   1   0   0   0   1   0   0
```

Rysunek 11

```
zmienna X1 jest katalizatorem
zmienna X2 jest katalizatorem
zmienna X3 jest katalizatorem
zmienna X4 jest katalizatorem
zmienna X5 jest katalizatorem
zmienna X6 jest katalizatorem
zmienna X7 jest katalizatorem
zmienna X8 jest katalizatorem
zmienna X9 jest katalizatorem
zmienna X10 jest katalizatorem
```

```
zmienna X11 jest katalizatorem
zmienna X12 jest katalizatorem
zmienna X13 jest katalizatorem
zmienna X14 jest katalizatorem
zmienna X15 jest katalizatorem
zmienna X16 jest katalizatorem
zmienna X17 jest katalizatorem
zmienna X18 jest katalizatorem
zmienna X19 jest katalizatorem
zmienna X20 jest katalizatorem
```

Rysunek 12

Współliniowość zmiennych

Z testu (Rysunek 13) wynika, że w modelu możliwa jest współliniowość. Biorąc pod uwagę, to, że zmienne, u których występuje to podejrzenie są zmiennymi binarnymi, stwierdzam, że efekt ten wystąpił, dlatego, że podczas tworzenia zmiennych nie została usunięta najpopularniejsza obserwacja i zdecydowałem, że zostawię je w modelu.

```
Ocena współliniowości VIF(j) - czynnik rozdęcia wariancji
VIF (Variance Inflation Factors) - minimalna możliwa wartość = 1.0
Wartości > 10.0 mogą wskazywać na problem współliniowości - rozdęcia wariancji

             accommodates    1,424
        number_of_reviews    1,061
     review_scores_rating    1,023
         Dproperty_type_1    3,298
         Dproperty_type_2    3,022
         Dproperty_type_3    1,419
             Droom_type_1   11,161
             Droom_type_2   10,767
              Dbed_type_1    1,702
              Dbed_type_3    1,507
              Dbed_type_4    1,187
   Dcancellation_policy_1    1,417
   Dcancellation_policy_3    1,431
                Dcity_1    5,668
                Dcity_2    2,570
                Dcity_3    2,311
                Dcity_4    5,118
                Dcity_5    2,036
Dhost_identity_verified_1    1,048
      Dinstant_bookable_1    1,034
```

Rysunek 13

Koincydencja

Weryfikacja polega na zbadaniu w modelu zachowania się danej zmiennej. Analizie poddany jest, zatem znak parametru strukturalnego w porównaniu do znaku współczynnika korelacji zmiennej objaśnianej i

zmiennej objaśniającej, przy której stoi badany. Mówimy, że model jest incydenty, jeśli dla każdej zmiennej objaśniającej modelu spełniony jest warunek:

$$\mathbf{sgn(r_i) = sgna(a_i)}$$

Definicja zaczerpnięta z podręcznika pana Eligiusza W. Nowakowskiego (strona 61) [5]. Poniżej przedstawiam zestawienie korelacji zmiennych z Y (Rysunek 13) oraz wektor współczynników (Rysunek 14).

accommodates	0,120814
number_of_reviews	-0,000560191
review_scores_ra~	0,00639447
Dproperty_type_1	-0,0330653
Dproperty_type_2	-0,0127899
Dproperty_type_3	0,105467
Droom_type_1	1,04898
Droom_type_2	0,477886
Dbed_type_1	0,0724192
Dbed_type_3	0,0509039
Dbed_type_4	0,106758
Dcancellation_~_1	0,0454008
Dcancellation_~_3	-0,0111018
Dcity_1	-0,0423684
Dcity_2	0,259653
Dcity_3	-0,128998
Dcity_4	-0,181821
Dcity_5	-0,333340
Dhost_identity~_1	0,0208397
Dinstant_booka~_1	0,0717117

Rysunek 14

0,5676	accommodates
-0,0325	number_of_revi~
0,0912	review_scores_~
-0,0452	Dproperty_type~
0,0112	Dproperty_type~
0,0657	Dproperty_type~
0,6025	Droom_type_1
-0,5316	Droom_type_2
0,0992	Dbed_type_1
-0,0395	Dbed_type_3
-0,0400	Dbed_type_4
0,1288	Dcancellation_~
-0,0959	Dcancellation_~
-0,0770	Dcity_1
0,1667	Dcity_2
0,0823	Dcity_3
-0,0566	Dcity_4
-0,0519	Dcity_5
0,0242	Dhost_identity~
0,0443	Dinstant_booka~

Rysunek 15

Koincydencja występuje w przypadku 5 zmiennych. Należałoby zmienić postać analityczną modelu lub zmienić zestaw zmiennych. W moim przypadku próba zmiany zmiennych, kończyła się drastycznym spadkiem współczynnika determinacji R^2.

Istotność zmiennych

Zakładając poziom istotności na poziomie 10%, wszystkie zmienne są istotne. Dla każdej przeprowadzony został test t-studenta (Rysunek 16).

Wszystkie zmienne zawarte w modelu są statystycznie stotne.

	t-Studenta	wartość p	
accommodates	118,9	0,0000	***
number_of_reviews	-12,22	2,74e-034	***
review_scores_ra~	27,12	5,25e-161	***
Dproperty_type_1	-4,730	2,25e-06	***
Dproperty_type_2	-1,689	0,0913	*
Dproperty_type_3	8,888	6,37e-019	***
Droom_type_1	84,94	0,0000	***
Droom_type_2	39,04	0,0000	***
Dbed_type_1	5,006	5,58e-07	***
Dbed_type_3	2,093	0,0364	**
Dbed_type_4	2,823	0,0048	***
Dcancellation_~_1	10,43	1,88e-025	***
Dcancellation_~_3	-2,125	0,0336	**
Dcity_1	-4,829	1,38e-06	***
Dcity_2	25,13	1,31e-138	***
Dcity_3	-11,97	5,33e-033	***
Dcity_4	-20,16	4,57e-090	***
Dcity_5	-29,40	1,25e-188	***
Dhost_identity~_1	4,953	7,33e-07	***
Dinstant_booka~_1	17,10	2,25e-065	***

Rysunek 16

Test RESET (REgression Specification Error Test)

Sprawdzenie czy model regresji liniowej, jest najodpowiedniejszą formą. Jeżeli oryginalny model regresji dobrany MNK jest odpowiedni, to dodanie do otrzymanego równania innych funkcji nie liniowych, nie będzie istotne (strona 306)[6]. Przeprowadzam test RESET (Rysunek 17).

H_0: model regresji liniowej jest dobrą formą

```
Test RESET na specyfikację (kwadrat i sześcian zmiennej)
Statystyka testu: F = 227,784816,
z wartością p = P(F(2,57225) > 227,785) = 2,92e-099

Test RESET na specyfikację (tylko kwadrat zmiennej)
Statystyka testu: F = 4,436439,
z wartością p = P(F(1,57226) > 4,43644) = 0,0352

Test RESET na specyfikację (tylko sześcian zmiennej)
Statystyka testu: F = 0,342103,
z wartością p = P(F(1,57226) > 0,342103) = 0,559
```

Rysunek 17

Dla przypadku modelu ze zmienną podniesioną do sześcianu, odrzucam hipotezę H_0, co oznacza że powinienem podnieść zmienną Y do sześcianu.

Test Chowa

H_0: po podzieleniu danych w okolicach wartości środkowej, parametry modelu dla dwóch podzbiorów będą równe

Przeprowadzam test Chowa (Rysunek 18).

```
Test Chowa na zmiany strukturalne przy podziale próby w obserwacji 37056
  F(21, 57206) = 0,968087 z wartością p 0,5005
```

Rysunek 18

p-value = 0,5005 => odrzucam hipotezę H_0

Należałoby dobrać inna formę modelu.

Heteroskedastyczność składnika losowego

Przeprowadzam test Test Breuscha-Pagana (Rysunek 19).

H_0: występuje zjawisko homoskedastyczności

```
Statystyka testu: LM = 4480,188262,
z wartością p = P(Chi-kwadrat(20) > 4480,188262) = 0,000000
```

Rysunek 19

p-value = 0 => nie mam podstaw do odrzucenia hipotezy o braku heteroskedastyczności

Ostateczna forma modelu

Y = 2,92007 + 1,04898 * **X1** + 0,477886 * **X2** − 0,0423684 * **X3** + 0,259653 * **X4** − 0,128998 * **X5** − 0,181821 * **X6** − 0,333340 * **X7** − 0,0330653 * **X8** − 0,0127899 * **X9** + 0,105467 * **X10** + 0,120814 * **X11** − 0,000560191 * **X12** + 0,00639447 * **X13** + 0,0724192 * **X14** + 0,0509039 * **X15** + 0,106758 * **X16** + 0,0454008 * **X17** − 0,0111018 * **X18** + 0,0208397 * **X19** + 0,0717117 * **X20**,

Gdzie:

Y- log ceny wynajmu mieszkania
X1 – czy wynajmujemy cały dom dla siebie
X2 – czy wynajmujemy własny pokój
X3 – czy lokalizacja to NYC

X4 – czy lokalizacja to SF
X5 – czy lokalizacja to DC
X6 – czy lokalizacja to LA
X7 – czy lokalizacja to Chicago
X8 – czy wynajmujemy mieszkanie
X9 – czy wynajmujemy dom rodzinny
X10 – czy wynajmujemy apartament
X11 – maksymalna liczba gości
X12 – liczba recenzji
X13 – średnia z recenzji
X14 – czy łózko jest zwykłego typu
X15 – czy łóżko jest rozkładaną sofą
X16 – czy łóżko jest kanapą
X17 – czy jest anulowanie rezerwacji jest rygorystyczne
X18 – czy anulowanie rezerwacji jest elastyczne
X19 – czy gospodarz jest zatwierdzonym użytkownikiem
X20 – czy istnieje możliwość natychmiastowej rezerwacji

6. Wnioski płynące z badań w kontekście postawionych hipotez badawczych

Podsumowując, największy wpływ na kształtowanie się ceny wynajmu noclegu ma liczba gości oraz to czy współdzielimy pomieszczenie z właścicielem, czy oddaje on nam całą nieruchomość dla siebie. Znaczący wpływ ma ocena wystawiana przez użytkowników, oraz miasto, w którym chcemy nocować. Ciekawym faktem, może być to, że Nowy Jork, uznawany za jedno z najdroższych miejsc do życia na ziemi, posiada ujemny współczynnik w modelu. Może to wynikać z bardzo małej powierzchni użytkowej tamtejszych mieszkań lub bardzo dużego popytu na krótkie noclegi tworzonego przez turystów, a co z tym idzie dużą konkurencję w ofertach. Zaskakiwać, może również to, że typ łóżka najbardziej wpływający na cenę to kanapa. Czynniki takie jak zdjęcie gospodarza, to czy jego profil jest zweryfikowany czy chociażby podejście do anulowania rezerwacji, mają znikomy wpływ na kształtowanie się ceny, lecz nadal w pewnym stopniu istotne.

7. Bibliografia oraz źródła

[1] https://en.wikipedia.org/wiki/Airbnb
[2] https://www.kaggle.com/stevezhenghp/airbnb-price-prediction
[3] https://pandas.pydata.org/
[4] Jakub Mućk - „Ekonometria - Model nieliniowe i funkcja produkcji” , http://web.sgh.waw.pl/~jmuck/Ekonometria/EkonometriaPrezentacja7.pdf
[5] Eligiusz W. Nowakowski - „PODSTAWY EKONOMETRII z elementami algebry liniowej”, https://liceum.wszechnicapolska.edu.pl/dokumenty/biblioteka/publikacje-cyfrowe/E-Nowakowski-Podstawy-ekonometrii.pdf
[6] Jeffrey Wooldridge, „Introductory Econometrics: A Modern Approach” (2012)