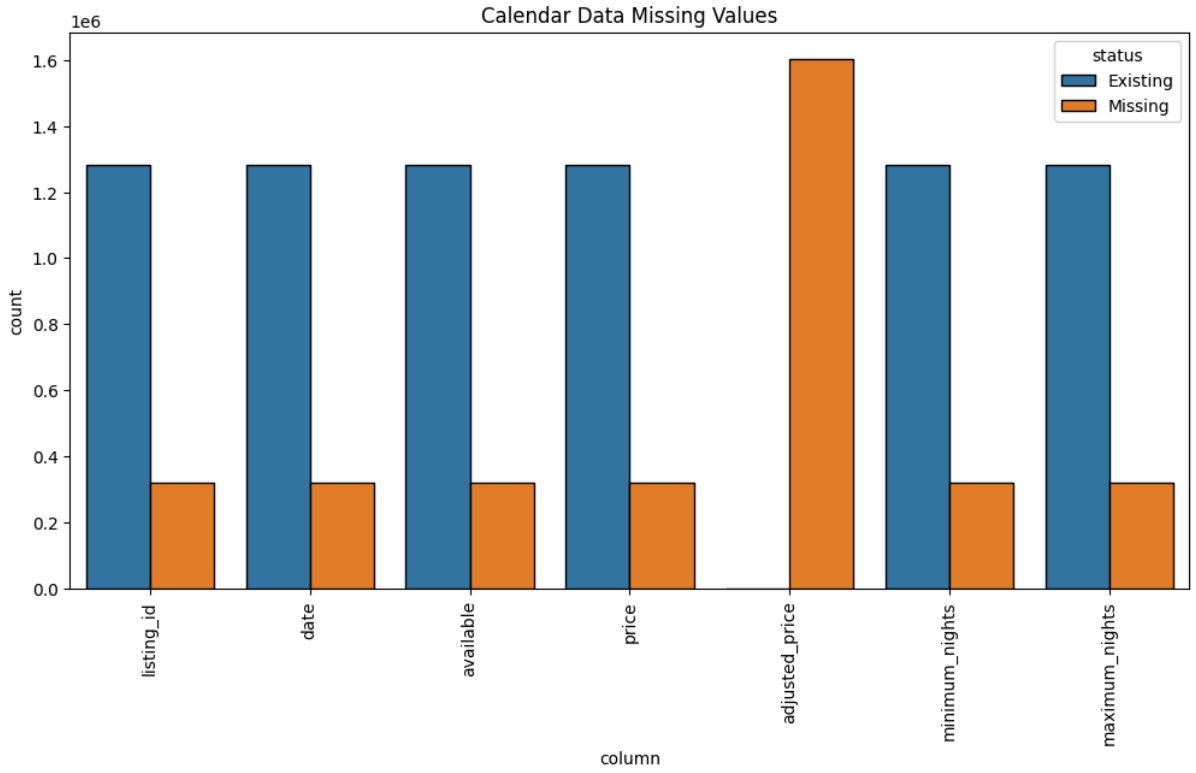
# Raport ze wstępnej analizy danych v1

## Zbiór danych *Calendar*

### Kolumny

- id listingu
- data
- czy dostępne - t / f
- cena - w dolarach
- dostosowana cena
- minimalna liczba nocy jaką trzeba zarezerwować
- maksymalna liczba nocy jaką można zarezerwować

### Brakujące wartości

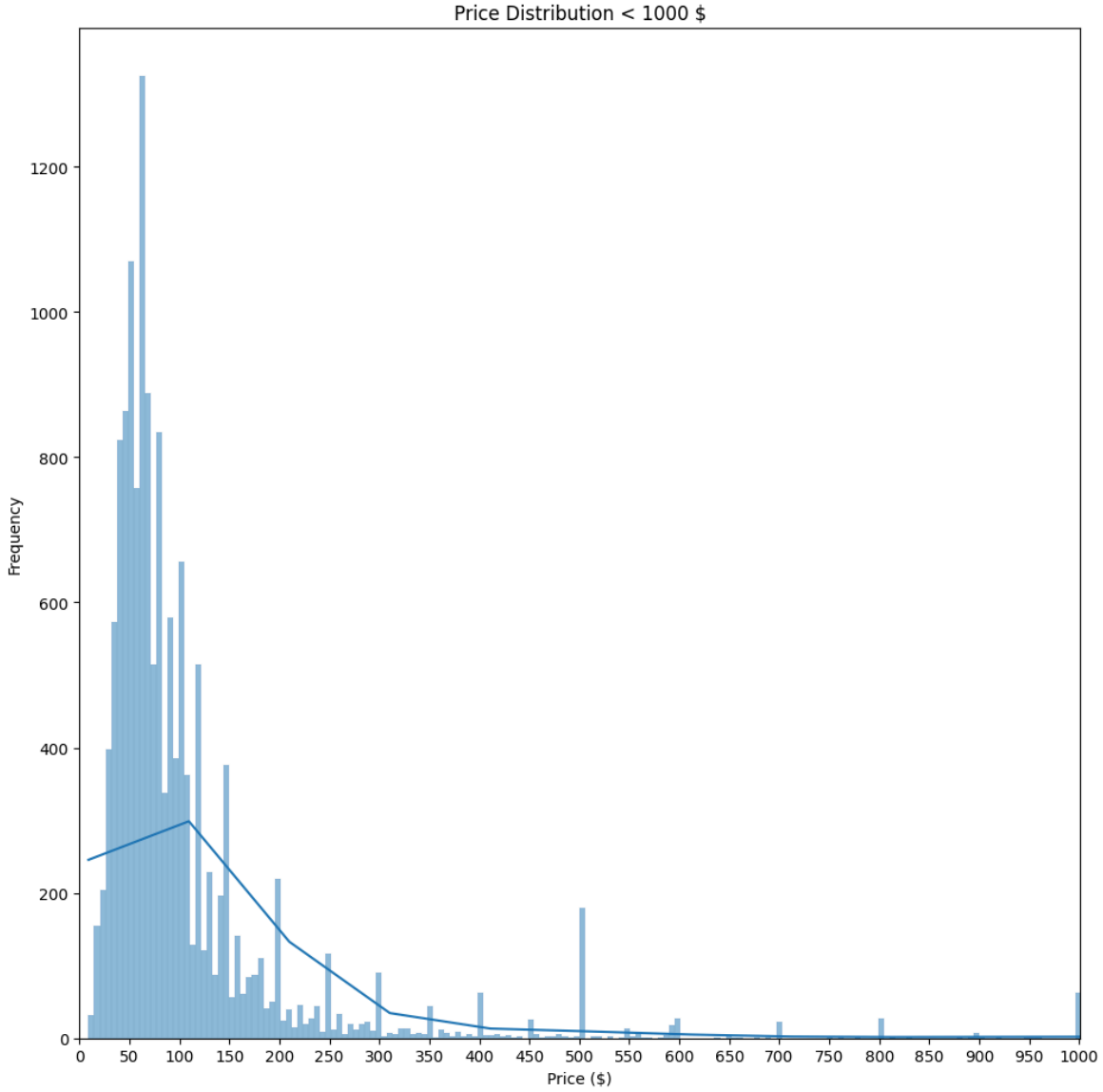


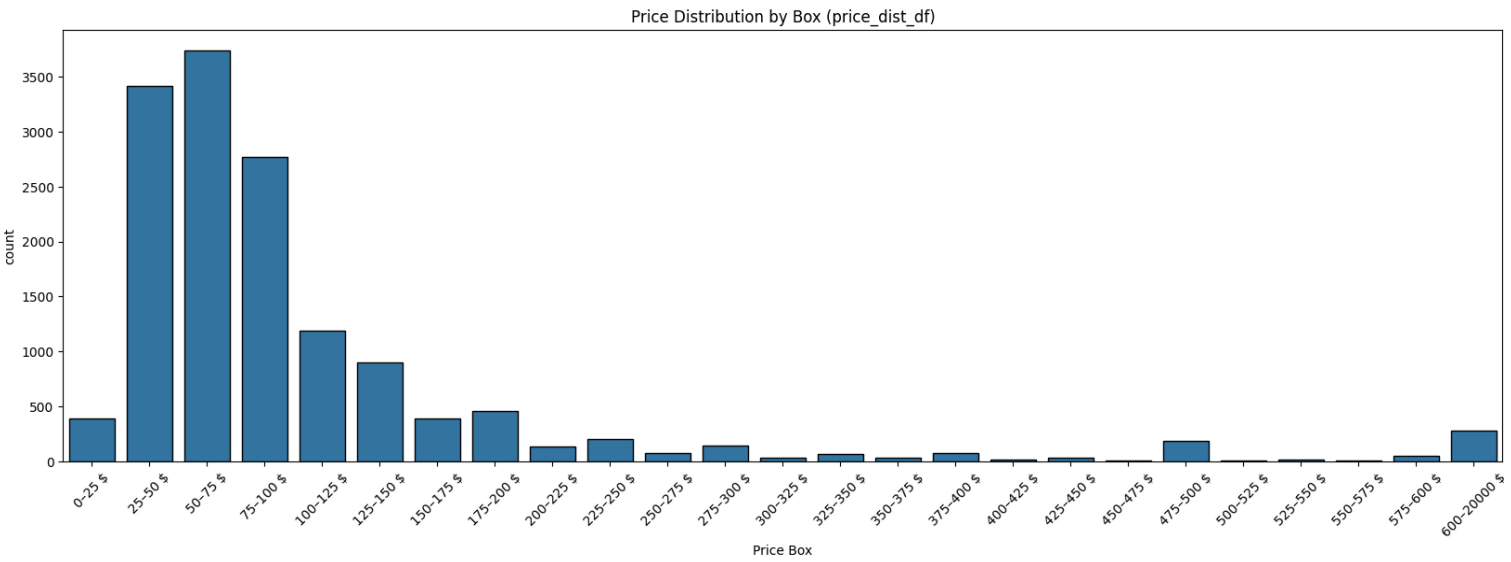

- brakuje około 20% danych z każdej kolumny

### Cena

- średnia - $153.71
- odchylenie standardowe - $588.44
- wartość najmniejsza - $9
- wartość największa - $ 20 000
- mediana - $ 50

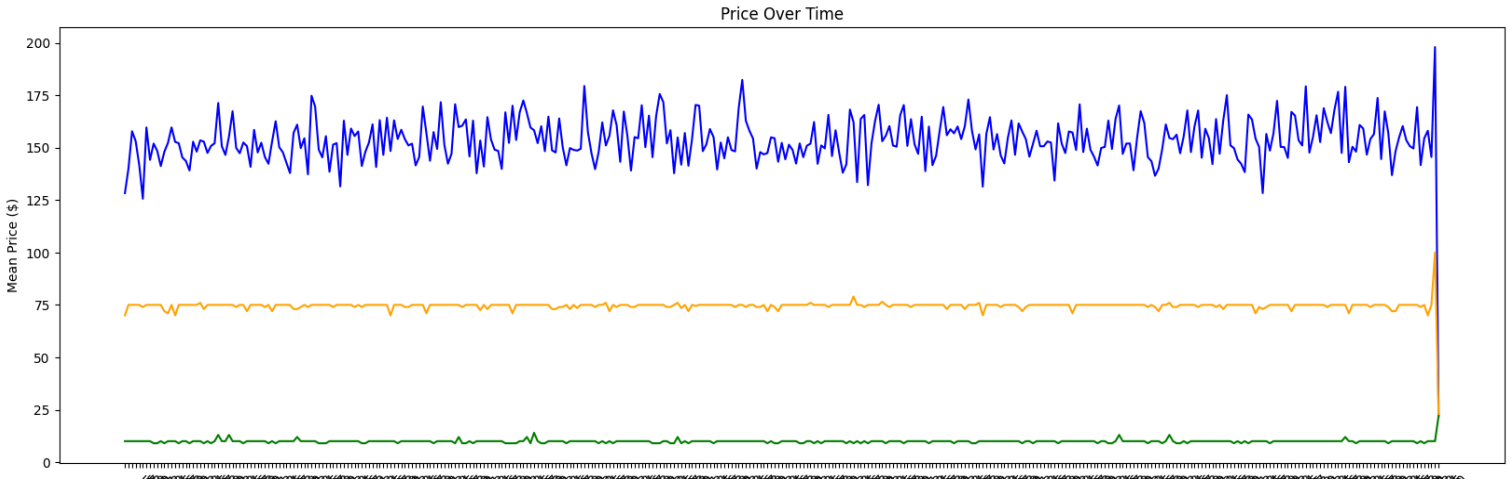
## Rozkład


Price Distribution < 1000 $
Frequency
Price ($)
0
200
400
600
800
1000
1200
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750
800
850
900
950
1000

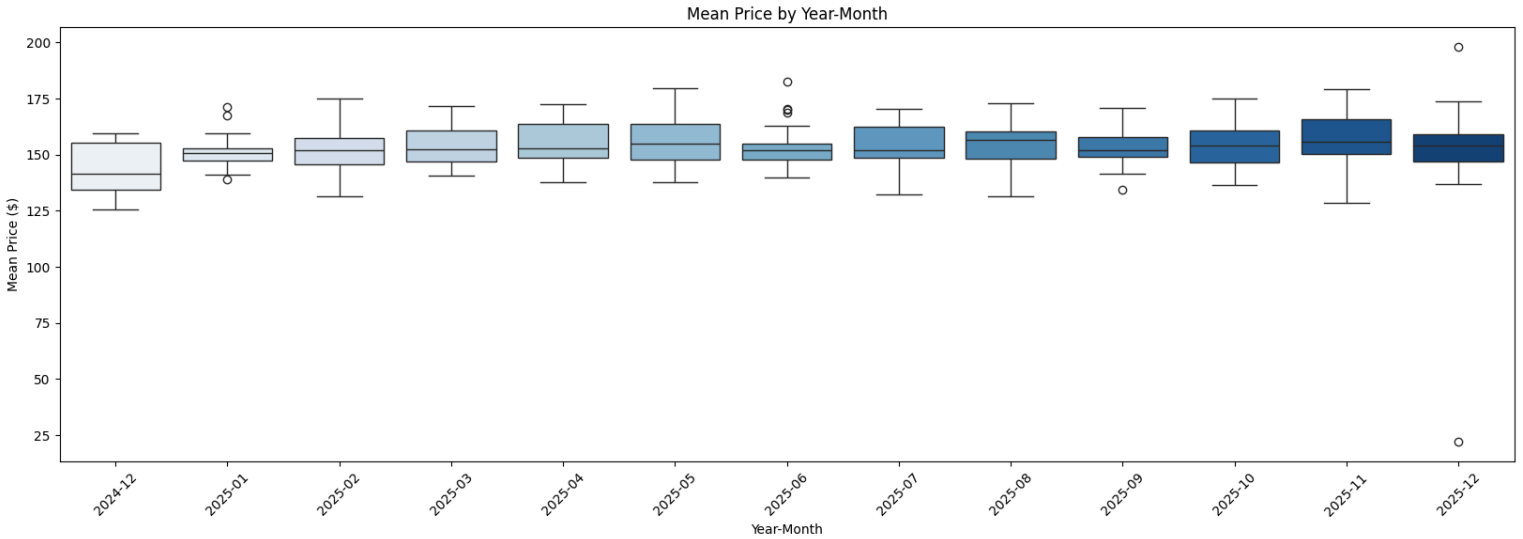

Price Distribution by Box (price_dist_df)
count
Price Box
0
500
1000
1500
2000
2500
3000
3500
0–25 $
25–50 $
50–75 $
75–100 $
100–125 $
125–150 $
150–175 $
175–200 $
200–225 $
225–250 $
250–275 $
275–300 $
300–325 $
325–350 $
350–375 $
375–400 $
400–425 $
425–450 $
450–475 $
475–500 $
500–525 $
525–550 $
550–575 $
575–600 $
600–20000 $

- ceny skupiają się głównie w niższych przedziałach, szczególnie poniżej 200$

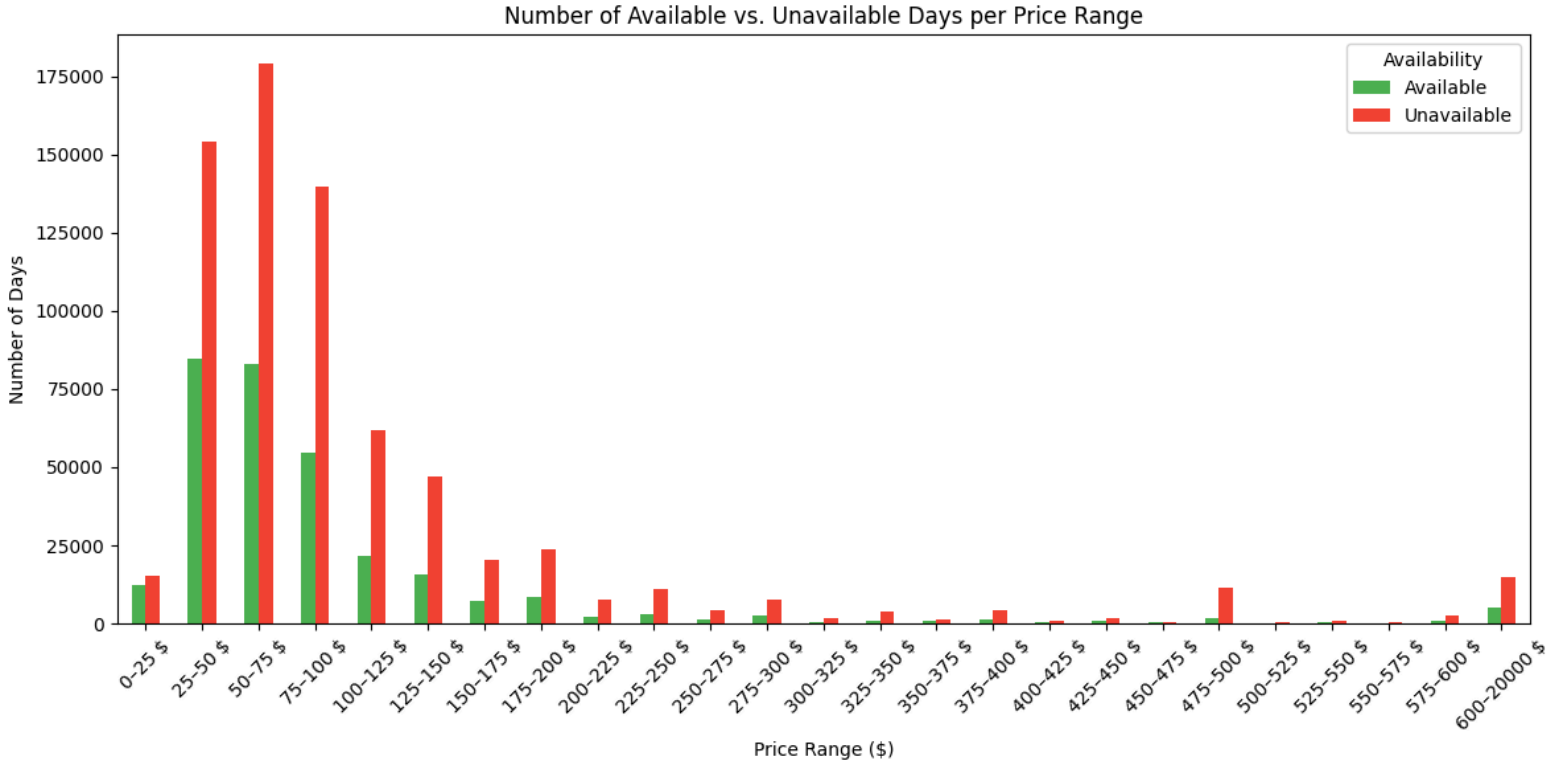
## Cena w czasie

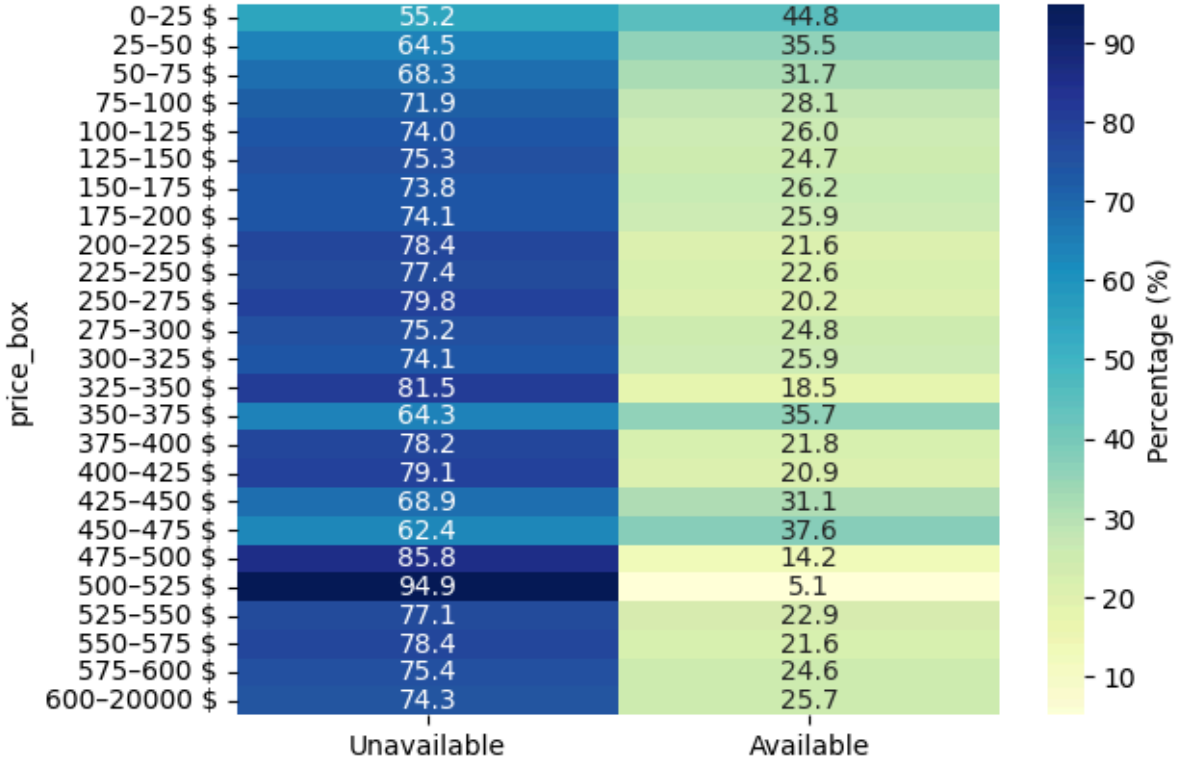



dane od: 25.12.2024, do: 30.12.2025




## Dostępność ofert






- dla cen z przedziału 0 - 25 $ procent wolnych i zajętych lokali jest zbliżony
- najmniej dostępnych lokali w przedziale od 500$ do 525$, potem procent dostępnych lokali rośnie z drobnymi wahaniami

### Wnioski

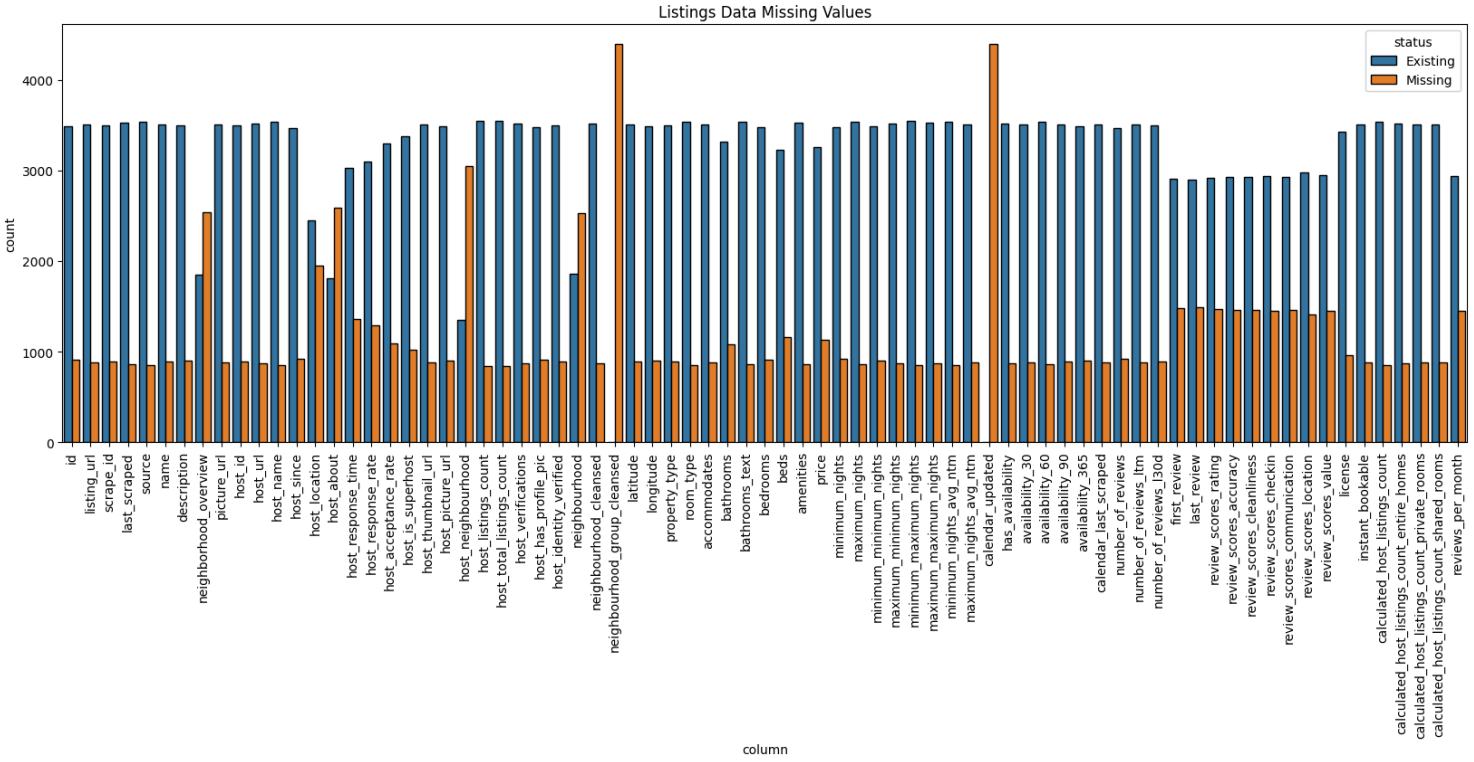
- brakuje danych wcześniejszych niż grudzień 2024
- analiza może bazować praktycznie na kilkunastu miesiącach, z czego połowa to dane rezerwacji, które jeszcze się nie odbyły i mogą ulec zmianie

## Zbiór danych *Listings*

- duża liczba kolumn - 75
- kolumny szczegółowo opisują ofertę zamieszczoną na portalu, m. in.:
    - nazwa i opis oferty
    - dane właściciela
        - imię
        - od kiedy jest hostem
        - jak szybko odpowiada
        - zdjęcie
        - liczba jego wszystkich listingów
        - czy jest zweryfikowany
        - dane pozostałych ofert tego hosta
    - dane okolicy
    - opis mieszkania:
        - rodzaj pokoju
        - ile osób pomieści

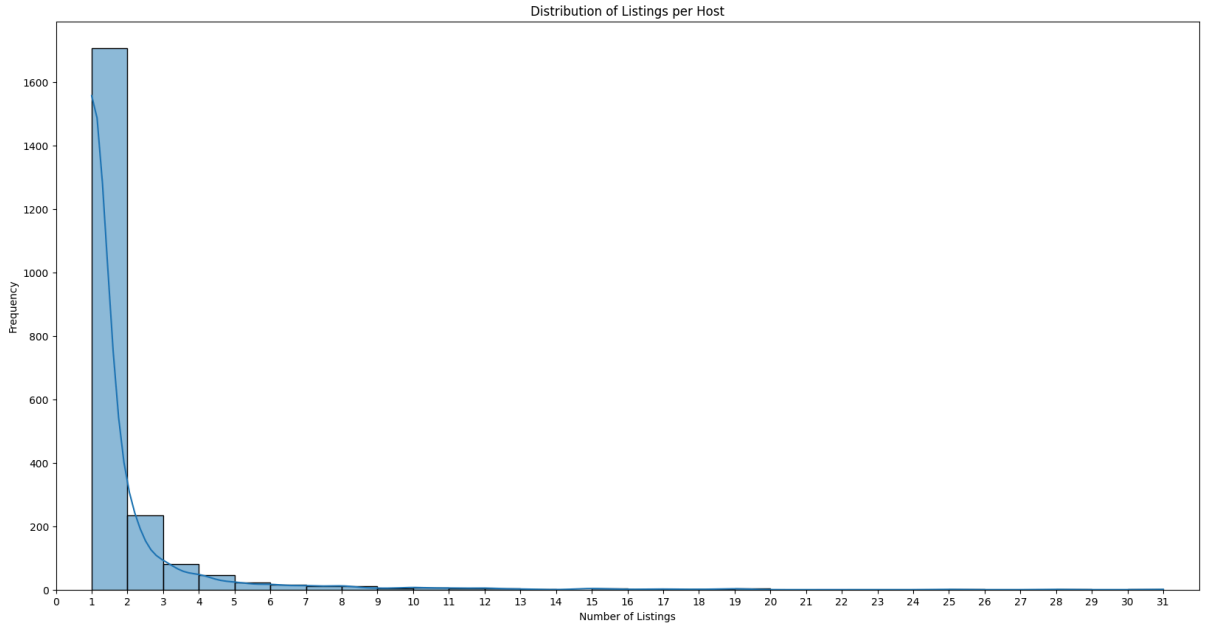
        - liczba łazienek, sypialni, łóżek
        - udogodnienia, w tym zaopatrzenie
        - cena
    - minimalna i maksymalna liczba nocy jaką można zarezerwować
    - opinie:
        - liczba opinii
        - data pierwszej i ostatniej opinii
        - ocena - ogólna oraz za czystość, check-in, komunikację z hostem, lokalizację
        - liczba ocen na miesiąc

## Brakujące wartości



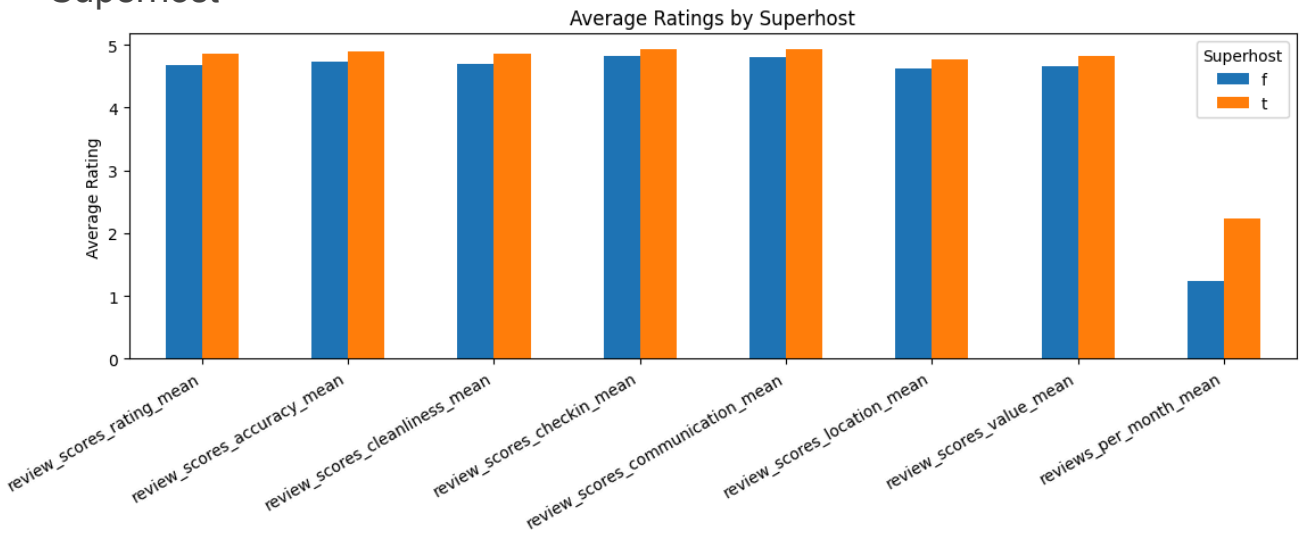

- sytuacja podobna jak w przypadku dany ze zbioru *Calendar*,  w każdej kolumnie jest minimum około 20% danych brakujących
- jeszcze mniej danych w kolumnach:
    - neighborhood_overview
    - host_about
    - host_location
    - host_neighborhood
- około 50 procent danych brakujących dotyczących opinii

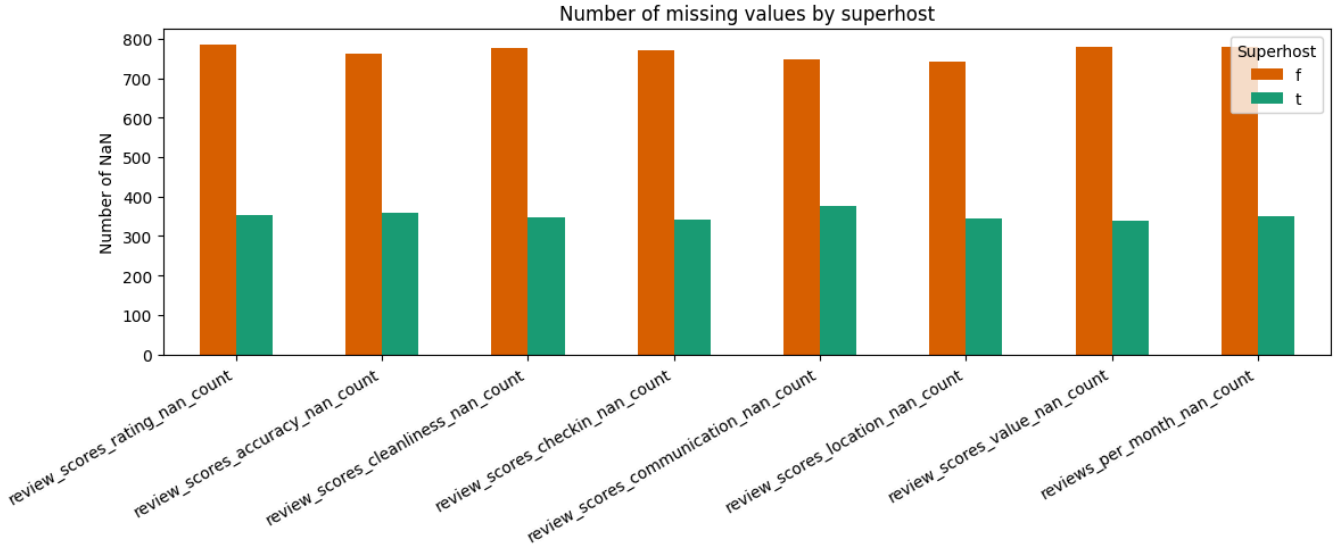
## Liczba ogłoszeń wystawianych przez hosta




- w znacznej większości przypadków jeden host wystawia jedno ogłoszenie

## Superhost



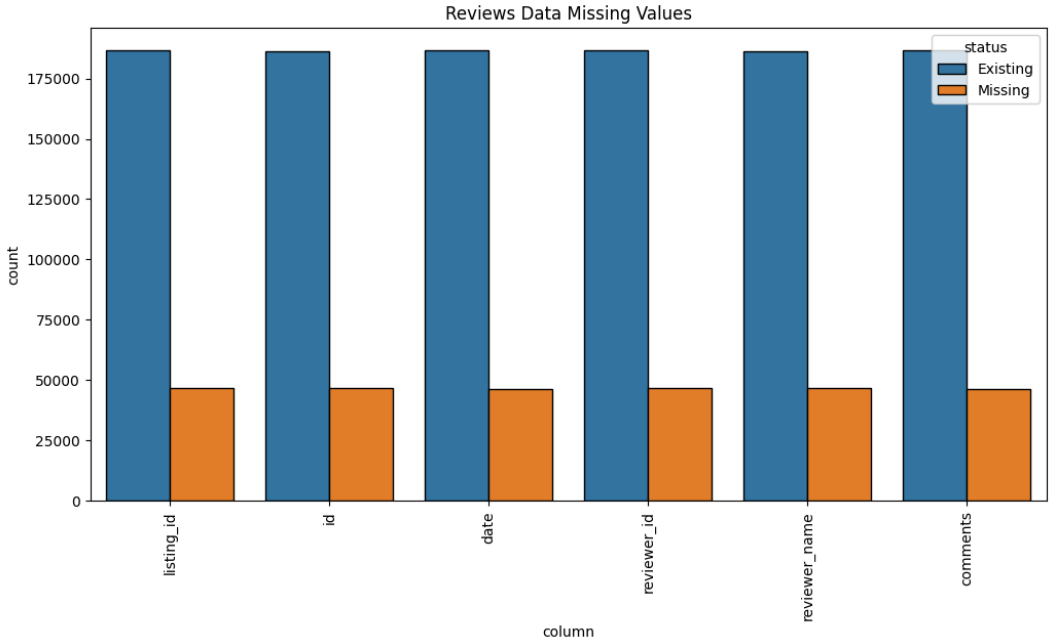



- superhost ma średnio wyższe oceny i mniej brakujących danych

# Zbiór danych *Reviews*

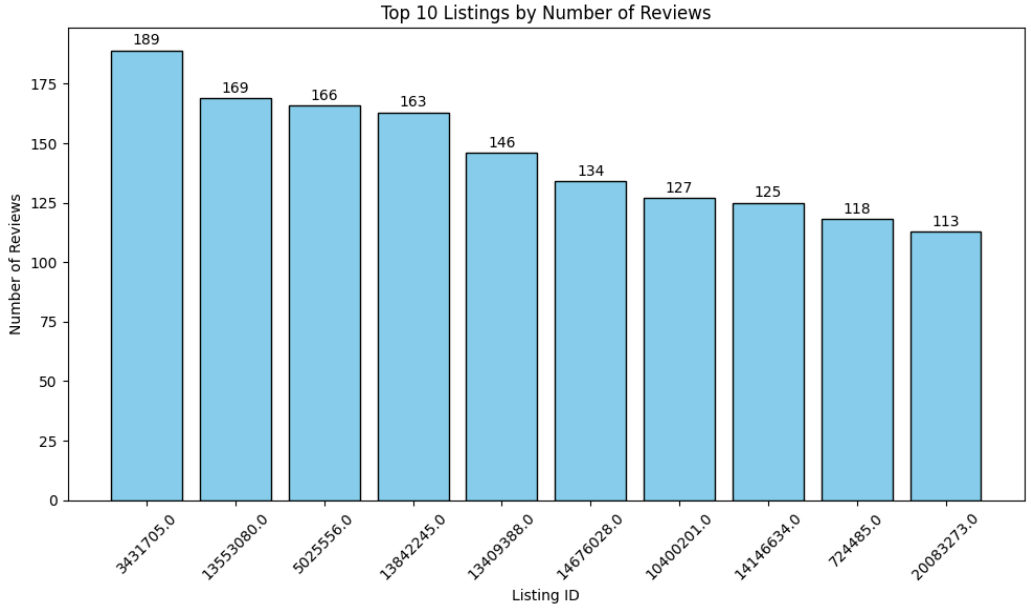
- kolumny:
  - id listingu
  - id
  - data
  - id wystawiającego
  - nazwa wystawiającego
  - zawartość

## Brakujące wartości

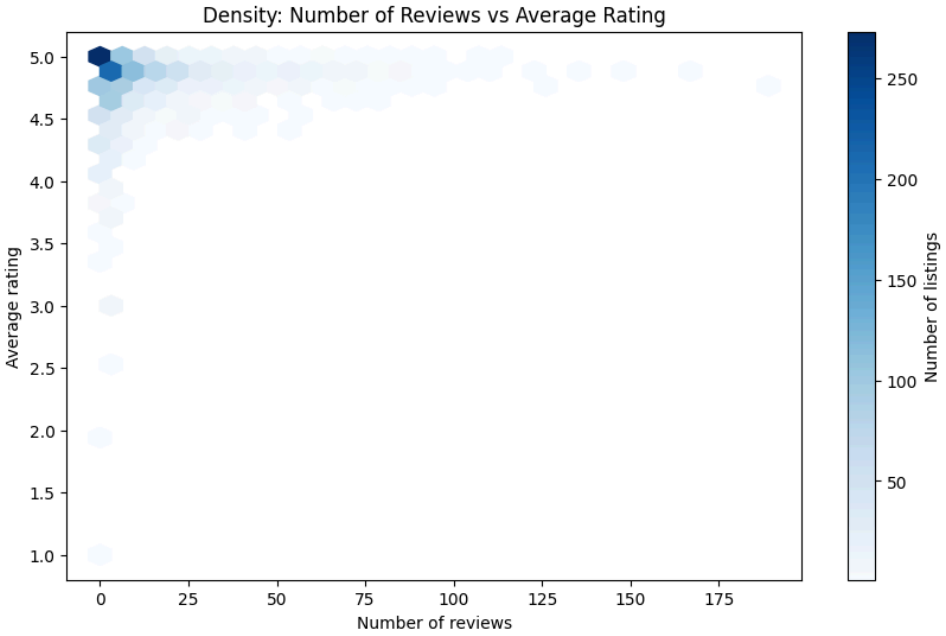



- około 20% dla każdej kolumny

## Listingi z największą liczbą opinii



## Ilość ocen a średnia ocena



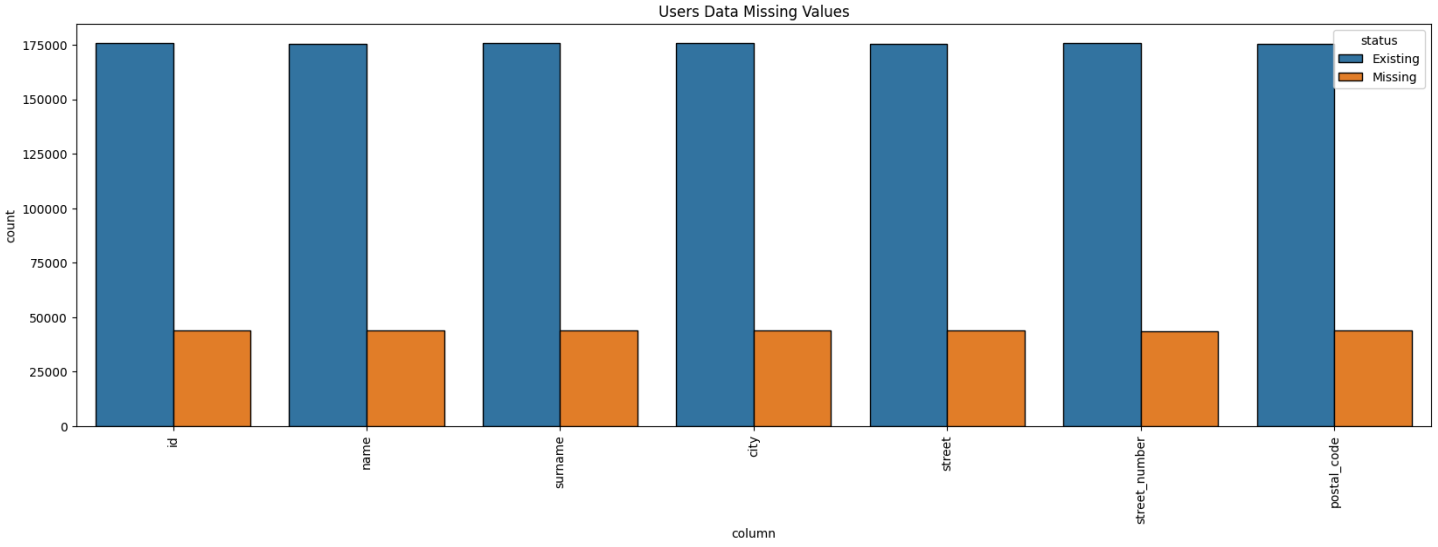

- oceny w większości są bardzo dobre
- największe zagęszczenie przy liczbie recenzji < 20 → wiele ofert jest raczej świeżych lub rzadko wynajmowanych

# Zbiór danych *Users*

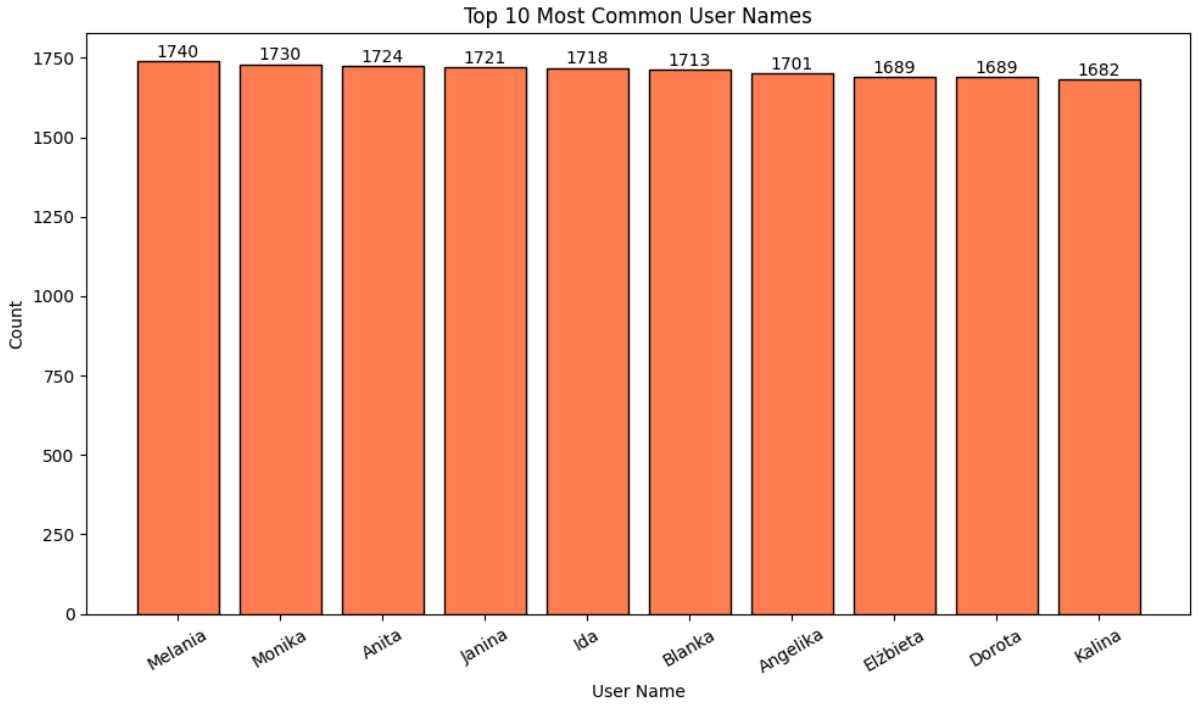
- kolumny:
  - id użytkownika
  - imię
  - nazwisko
  - miasto
  - ulica
  - numer domu
  - kod pocztowy
- liczba wierszy - 219 488

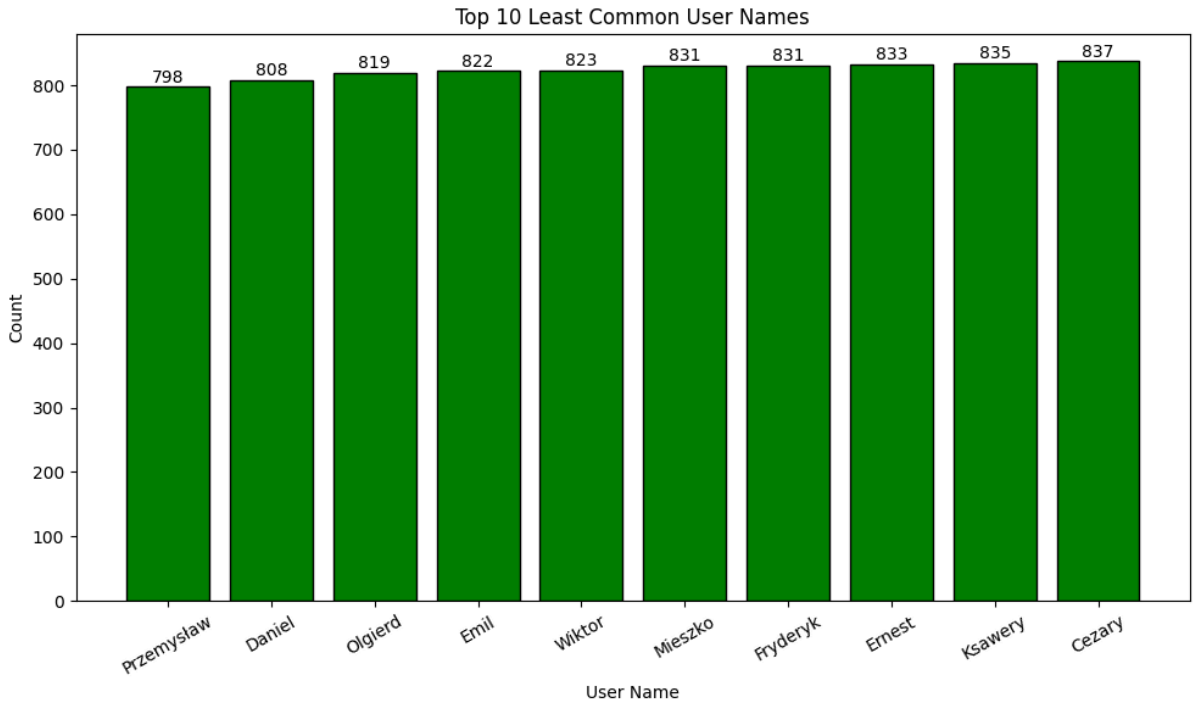
## Brakujące wartości

- około 20 % dla każdej kolumny



## Imiona użytkowników



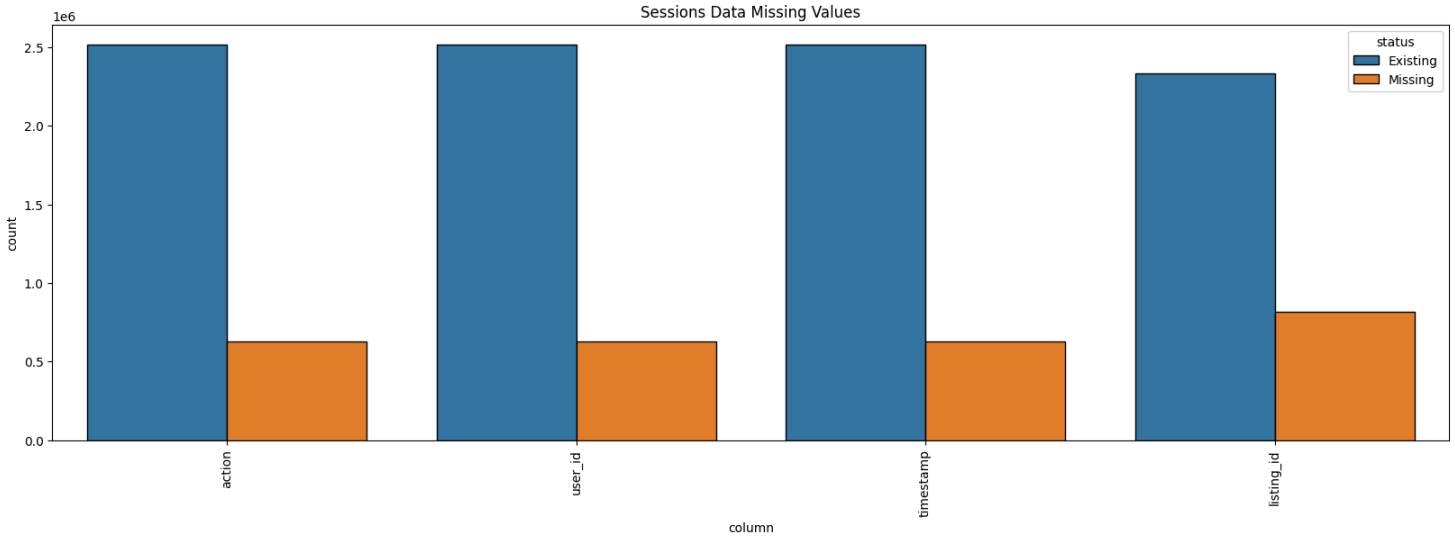

- na pierwszy rzut oka wygląda, jakby to kobiety w większej ilości korzystały z portalu nocarz

# Zbiór danych *Sessions*

- kolumny

    - akcja - browse_listing / view_listing / book_listing
    - id użytkownika
    - timestamp
    - id listingu
- liczba wierszy: 3 147 014

## Brakujące wartości




- około 20 procent dla każdej kolumny

### Dane w których brakuje *user_id*

| | action | user_id | timestamp | listing_id |
|---|---|---|---|---|
| 0 | browse_listings | NaN | 2024-06-30T14:44:43.340297 | NaN |
| 2 | view_listing | NaN | 2024-06-30T15:13:38.340297 | 902309243550043904.000 |
| 6 | view_listing | NaN | 2024-06-30T17:35:18.340297 | 1131235801642140928.000 |
| 8 | NaN | NaN | 2024-06-30T18:17:40.340297 | NaN |
| 9 | view_listing | NaN | 2024-06-30T19:02:14.340297 | 1168687047912305920.000 |
| ... | ... | ... | ... | ... |
| 3146979 | view_listing | NaN | 2024-04-14T16:40:47.135884 | NaN |
| 3146988 | browse_listings | NaN | 2024-04-14T12:56:38.300786 | NaN |
| 3146997 | view_listing | NaN | NaN | NaN |
| 3147005 | view_listing | NaN | 2024-04-14T17:45:42.371619 | 629945264982240128.000 |
| 3147010 | view_listing | NaN | 2024-04-14T13:25:21.065014 | 1095767923297763584.000 |

628980 rows × 4 columns

Dane, w których brakuje *listing_id*

| | action | user_id | timestamp | listing_id |
|---|---|---|---|---|
| 0 | browse_listings | NaN | 2024-06-30T14:44:43.340297 | NaN |
| 4 | NaN | 24106857.000 | NaN | NaN |
| 8 | NaN | NaN | 2024-06-30T18:17:40.340297 | NaN |
| 12 | view_listing | NaN | 2024-06-30T20:52:33.340297 | NaN |
| 14 | browse_listings | 501218607.000 | 2024-04-14T12:56:30.475311 | NaN |
| ... | ... | ... | ... | ... |
| 3146996 | browse_listings | 67727859.000 | 2024-04-14T12:56:38.371619 | NaN |
| 3146997 | view_listing | NaN | NaN | NaN |
| 3147003 | view_listing | 67727859.000 | 2024-04-14T16:52:42.371619 | NaN |
| 3147008 | NaN | 351649506.000 | 2024-04-14T12:56:38.065014 | NaN |
| 3147009 | NaN | 351649506.000 | 2024-04-14T13:04:37.065014 | NaN |

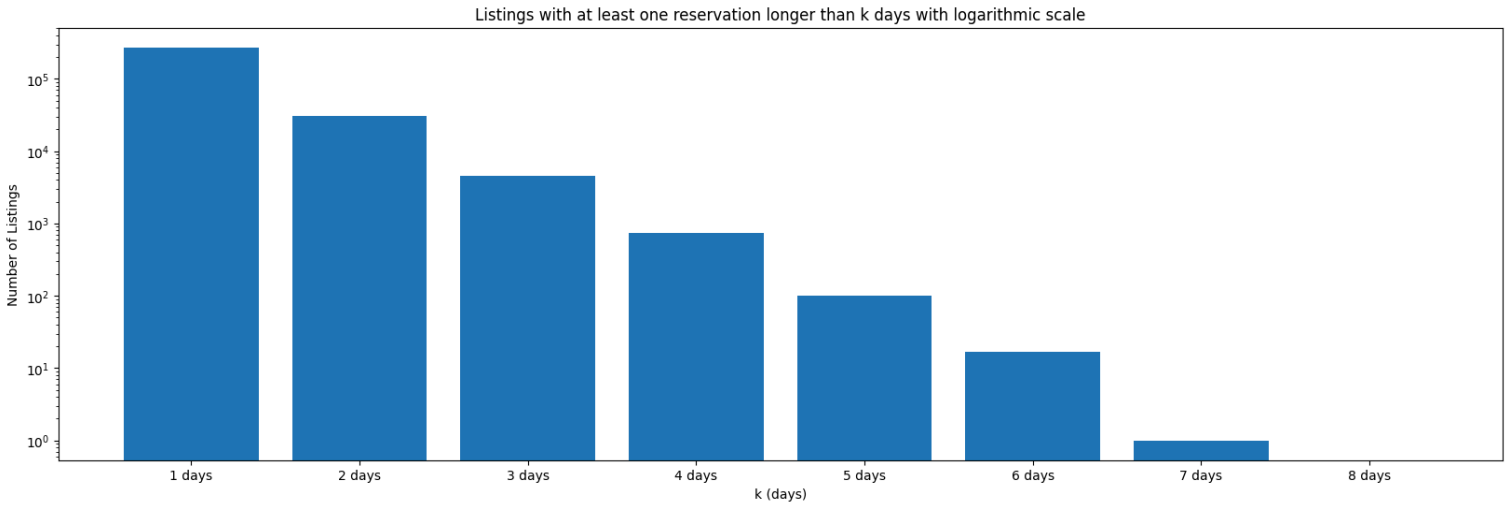
815573 rows × 4 columns

- około 1.5 miliona wierszy z tej tabeli jest dla nas mało informatywnych, bo bez *user_id* albo *listing_id* ciężko będzie połączyć akcję z konkretnym użytkownikiem bądź listingiem

## Analiza pod kątem dłuższych pobytów

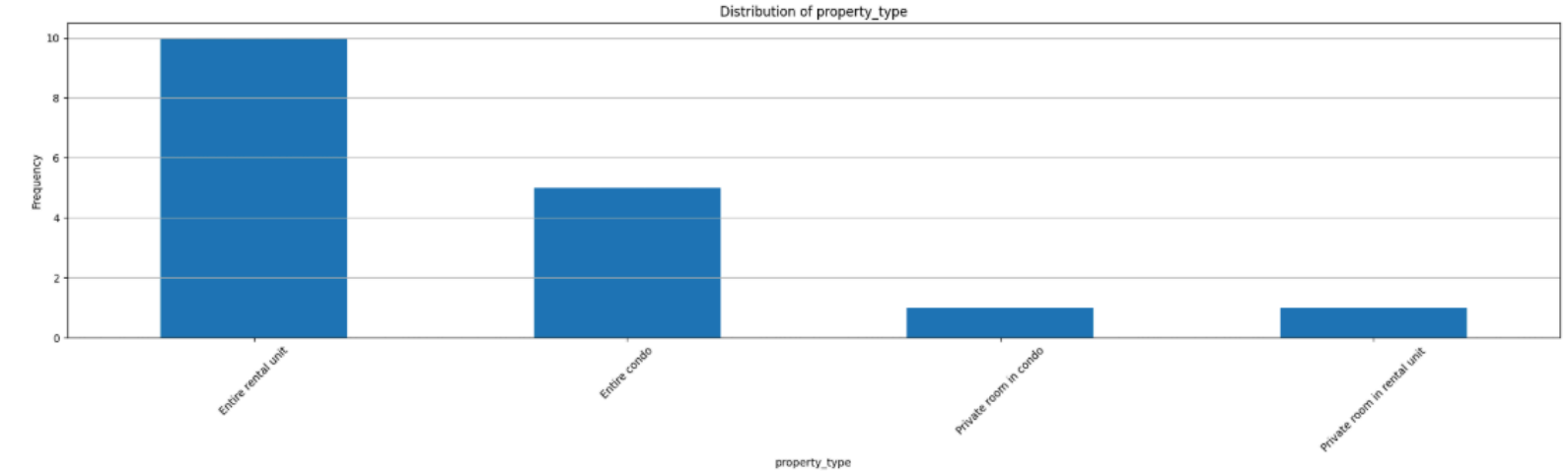
- przyjmujemy, że dłuższy pobyt oznacza pobyt zarezerwowany na więcej niż 5 dni

### Liczba rezerwacji dłuższych niż k dni

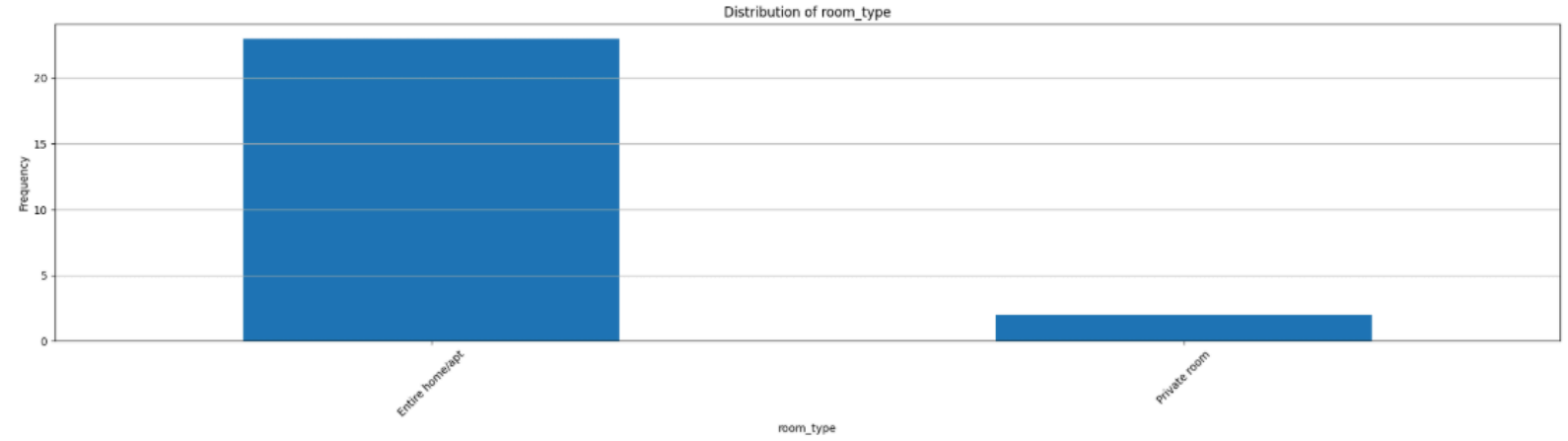



- mamy stosunkowo mało danych dla dłuższych pobytów - 118

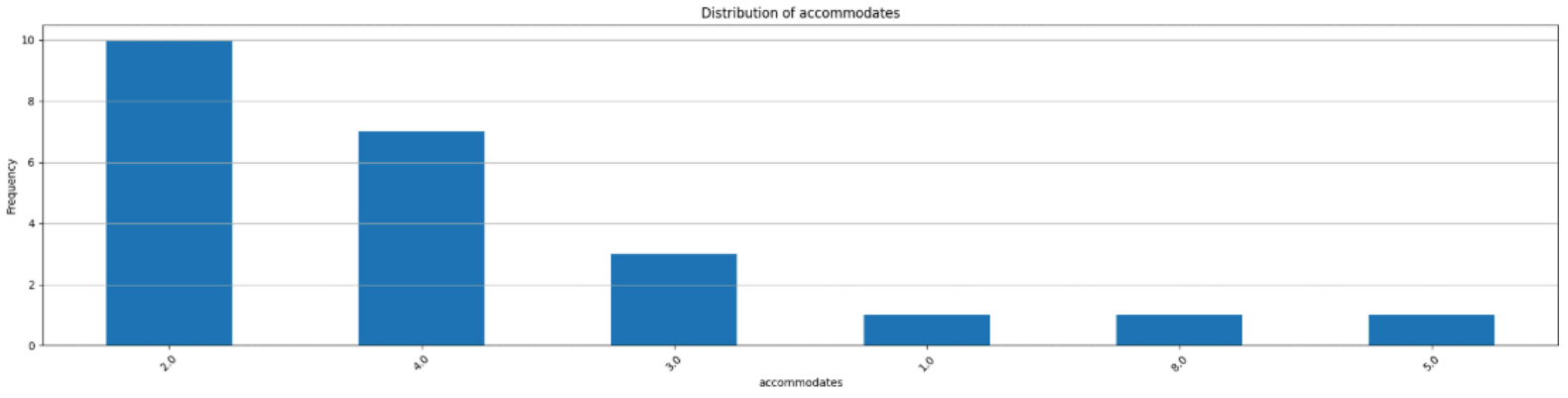
## Rodzaj zakwaterowania



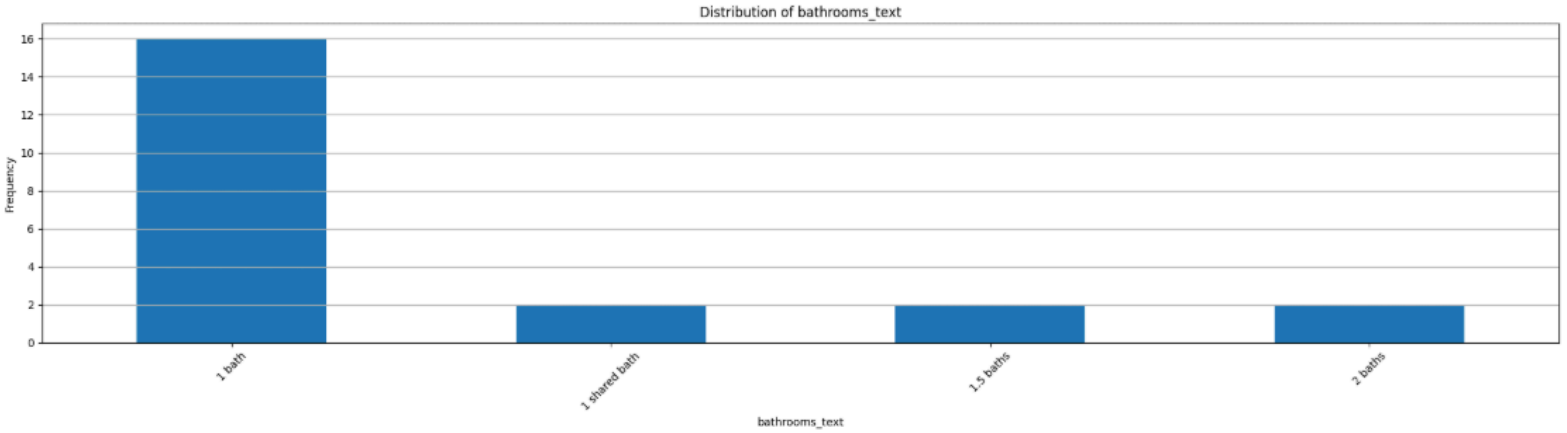

## Rodzaju pokoju



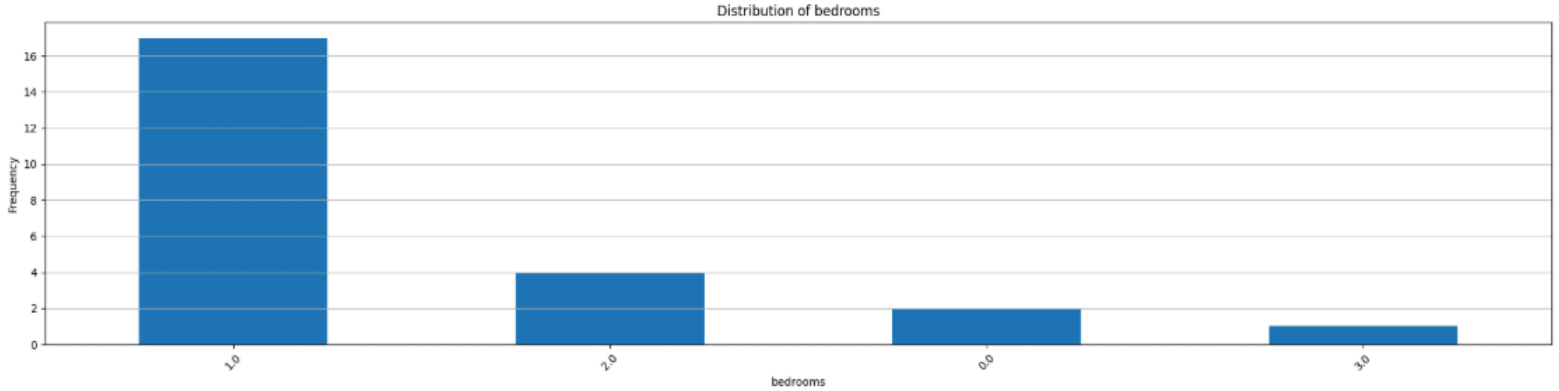

## Dla ilu osób przeznaczona jest oferta



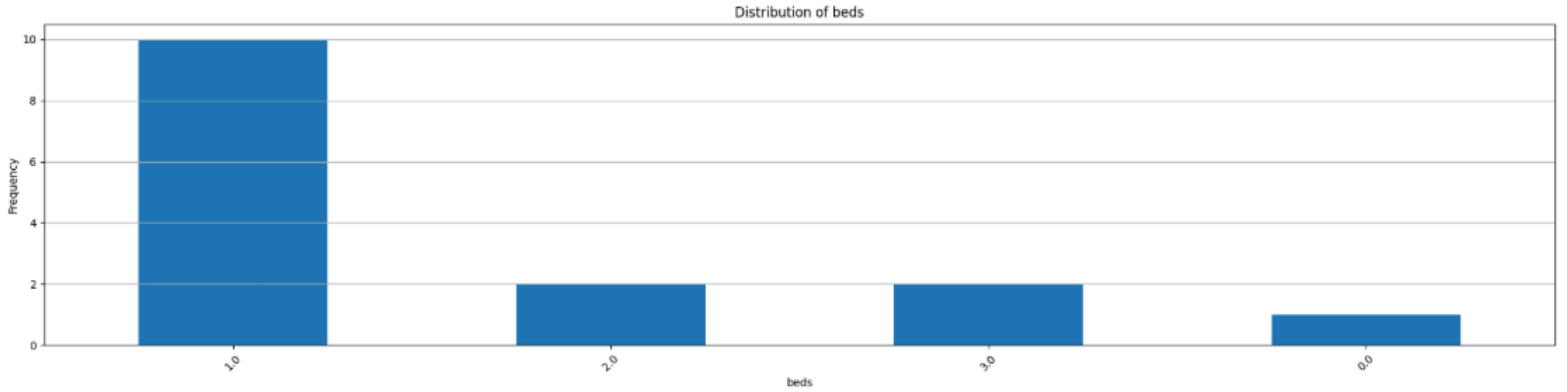
## Ilość łazienek




## Ilość sypialni




## Ilość łóżek




## Cena