Etap 1 : Poprawki

Zmiany wprowadzone odnośnie założeń do klasyfikacji

Wraz z pojawieniem się nowych danych, postanowiliśmy zmienić “granicę”, jaką przyjmujemy za krótki pobyt. Zwiększyliśmy ją do 7 dni włącznie, gdyż wraz z nowymi danymi pojawiły się dla nas oferty wynajmowane z zakresu 1- 14 dni, gdzie w poprzedniej wersji danych nasze dane obejmowały rezerwacje w zakresie 1- 7 dni.

Zmienne wejściowe a zmienna celu

> Czy jesteście Państwo pewni, że zmienne wejściowe niosą jakąś informację o zmiennej celu?

W celu weryfikacji, czy dostarczone dane niosą jakieś informację dokonaliśmy następującej analizy, bazującej na nowo dostarczonej wersji danych, która :

- zawiera więcej danych dotyczących dłuższych pobytów
- zawierają mniej brakujących danych

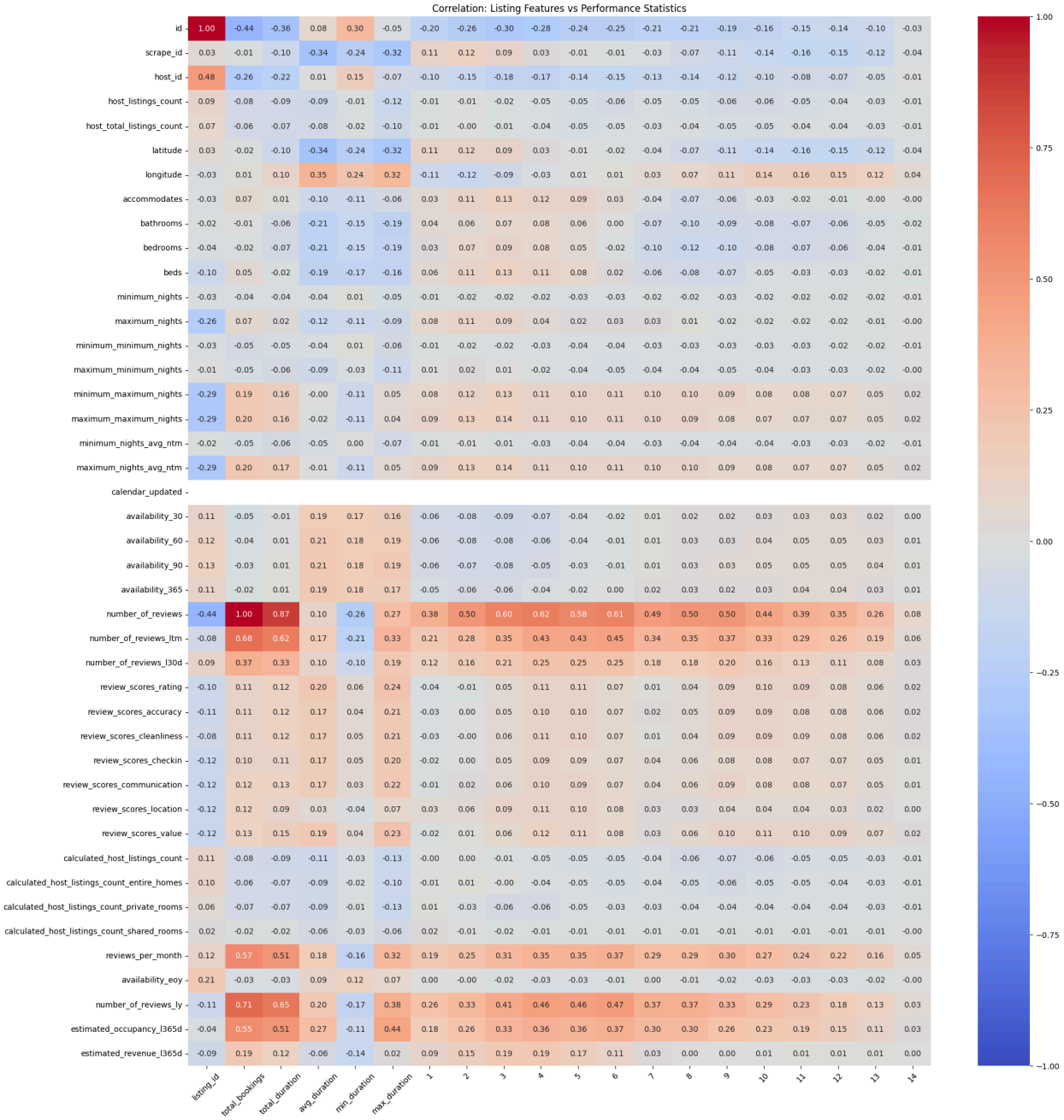
Sprawdziliśmy następujące rzeczy :

1. Na podstawie pliku sessions.csv byliśmy w stanie obliczyć dane dotyczące długości pobytów :
    a. ile razy dany listing został wynajęty
    b. ile razy zosatał wynajęty na X dni z przedziału 1 - 14 dni
    c. jakie były najkrótsze i najdłuższe pobyty tam
    d. średnia długość wynajęcia

	listing_id	total_bookings	total_duration	avg_duration	min_duration	max_duration	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	21853.0	33	48	1.454545	1	4	21	10	1	1	0	0	0	0	0	0	0	0	0	0
1	27262.0	29	144	4.965517	2	7	0	1	2	6	9	10	1	0	0	0	0	0	0	0
2	30320.0	172	360	2.093023	1	4	43	76	47	6	0	0	0	0	0	0	0	0	0	0
3	30959.0	8	12	1.500000	1	3	5	2	1	0	0	0	0	0	0	0	0	0	0	0
4	33945.0	78	540	6.923077	3	10	0	0	2	1	5	17	29	18	4	2	0	0	0	0

Na podstawie powyższej tabeli tabeli a dokładnie wartości avg_duration mogliśmy sprawdzić wpływ konkretnych zmiennych na wartość celu, która w naszym wypadku polega na określeniu długości pobytu (średnia nam tutaj dużo powie)

2. Stworzyliśmy macierz korelacji między zmiennymi numerycznymi i danymi statystycznymi, w celu powiązania ważnych wartości numerycznych z naszą zmienną celu

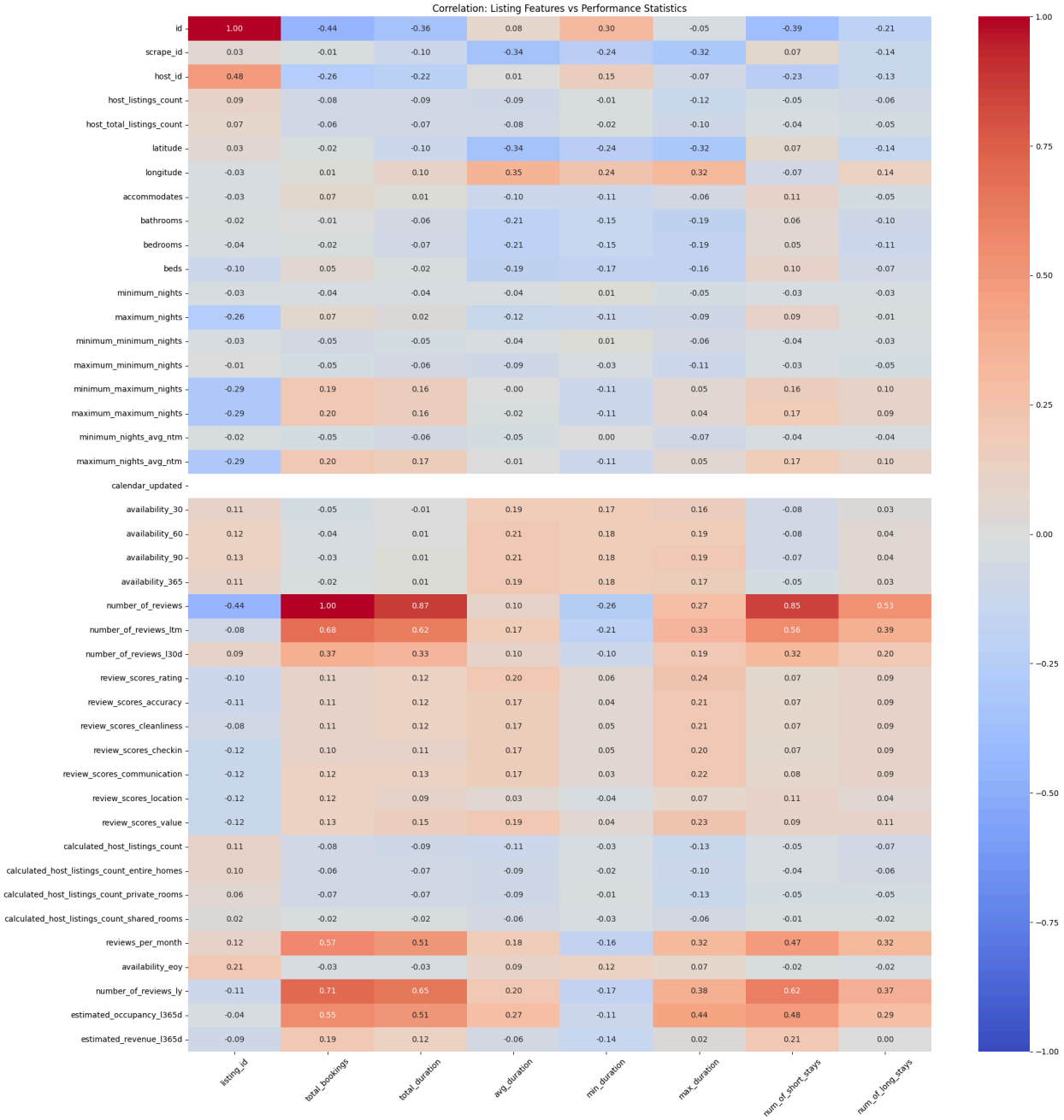


- Na podstawie macierzy korelacji można zauważyć, że średnia długość pobytu oraz konkretne wartości długości pobytu są silnie powiązane z ocenami dotyczącymi wynajmowanego miejsca. Współczynnik korelacji dla tych zmiennych osiąga wysokie wartości, sięgające nawet około 0,8. Kluczowym atrybutem w tym kontekście jest liczba opinii.
- Mając to na uwadze, w kolejnej fazie planujemy przeanalizować także sentyment opinii i jego wpływ na naszą zmienną celu. W tym celu zamierzamy

wytrenować model językowy, co będzie zadaniem realizowanym w kolejnym etapie projektu.

- Warto również zauważyć, że niektóre cechy wykazują ujemny współczynnik korelacji na poziomie około -0,3. Choć jest to wartość umiarkowana, wyróżnia się na tle innych i uznajemy ją za istotną dla naszej zmiennej celu. Przykładowo, takie kolumny to m.in. „bathrooms", „bedrooms" oraz „beds"

Tutaj również macierz korelacji dla klasyfikacji binarnej pobytów :



3. Oprócz współczynnika korelacji liniowej sprawdziliśmy także współczynnik MI (Mutual Information), który przydatny jest w celach wykrycia zależności innych niż liniowe. Również tam najbardziej wyróżniał się atrybuty dotyczące głównie recenzji. (Wykresy znajdują się w notatniku data_v2_eda.ipnyb)
4. Przyjrzeliśmy się także, danym kategorycznym :
    a. cena

b. typ posiadłości
c. typ pokoju
d. opis łazienki

Zrobiliśmy porównanie, uwzględniające każdą wartość z danego typu i jej wpływ na średnią długość pobytu i wyniki utwierdził nas, że niektóre zmienne kategoryczne mają duży wpływ na średnią długość wynajmowania :

```
Sorting by column: property_type
```

	property_type	avg_duration
7	Entire cottage	9.009615
25	Private room in cave	7.319088
62	Yurt	6.824859
6	Entire condo	6.280447
61	Tiny home	6.154276
9	Entire guesthouse	5.935102
26	Private room in chalet	5.722426
8	Entire guest suite	5.197302
10	Entire home	5.103420
16	Entire townhouse	5.076395

	property_type	avg_duration
57	Shared room in hostel	2.066929
60	Shared room in rental unit	1.814029
58	Shared room in hotel	1.394467
54	Shared room in bed and breakfast	1.375697
28	Private room in dome	1.266667
36	Private room in pension	1.166667
47	Room in bed and breakfast	1.090909
59	Shared room in loft	1.045977
49	Room in casa particular	1.035156
38	Private room in ryokan	1.000000

```
Sorting by column: room_type
```

	room_type	avg_duration
0	Entire home/apt	5.126607
2	Private room	4.139962
1	Hotel room	2.254003
3	Shared room	2.019045

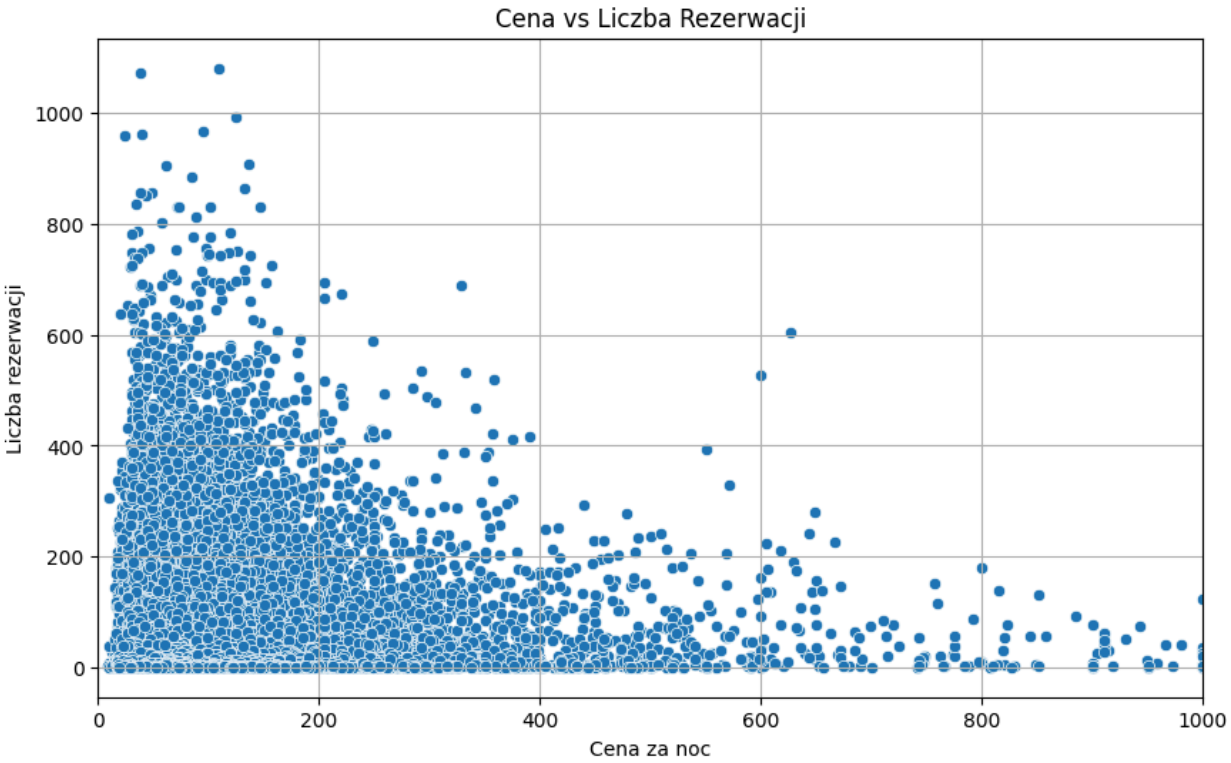
Sorting by column: bathrooms_text

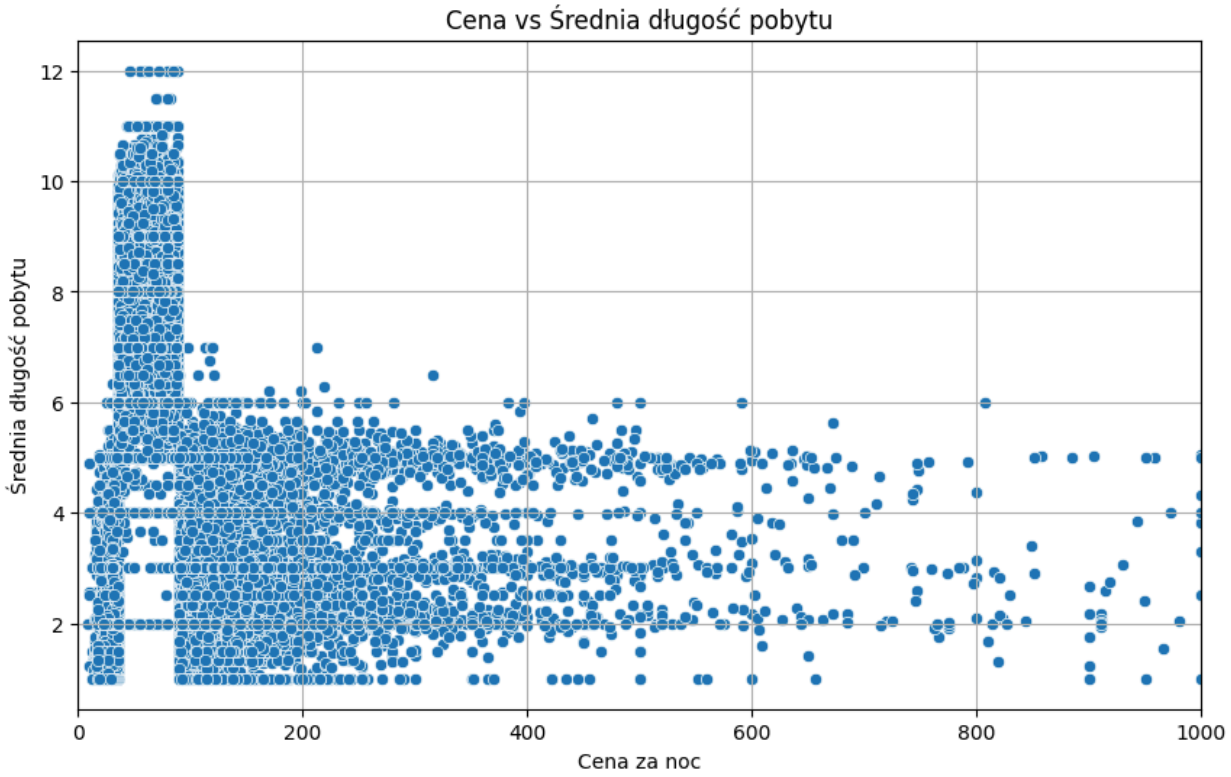
	bathrooms_text	avg_duration
9	15 baths	5.968750
7	10 baths	5.684211
2	1 bath	5.413667
35	Private half-bath	5.218742
5	1.5 baths	4.982354
34	Half-bath	4.973205
23	5 shared baths	4.741040
0	0 baths	4.705494
30	8 baths	4.616667
8	12 baths	4.428571

	bathrooms_text	avg_duration
19	4 shared baths	3.179154
15	3 shared baths	3.175817
28	7 baths	3.133405
25	6 baths	3.129481
32	8.5 baths	2.884615
33	9 baths	2.571429
17	3.5 shared baths	1.512082
21	4.5 shared baths	1.469681
26	6 shared baths	1.200000
29	7 shared baths	1.000000

Co więcej porównując ceny ofert, zauważyliśmy zależność, że :

- tańsze noclegi są częściej wynajmowane
- średnia długość tańszych noclegów jest większa niż tych droższych - mamy więcej długich rezerwacji dla ofert znacznie tańszych (do 100$ za noc)


Cena vs Liczba Rezerwacji
Liczba rezerwacji
Cena za noc
0
200
400
600
800
1000


Cena vs Średnia długość pobytu
Średnia długość pobytu
Cena za noc
2
4
6
8
10
12
0
200
400
600
800
1000

Na podstawie tego, również uważamy, że dane opisujące dokładnie wyposażenie mieszkań, mogą wpływać na podejmowanie decyzji przez użytkowników na jak długo wynająć dany lokal, aczkolwiek na tym etapie jest to “educated guess”

Analityczne kryterium sukcesu a dostępne dane

> Ustalone analityczne kryterium sukcesu powinno być jakoś oparte o dostępne dane

Dzięki uzyskaniu nowej wersji danych byliśmy w stanie powiązać listingi z rezerwacjami, co pozwoliło nam na obliczenie, jak wspomniałem wcześniej:

- ile razy dany lokal został wynajęty na konkretną liczbę dni,
- jaka jest średnia długość wynajmu danego lokalu.

Dzięki temu możemy:

- porównywać nasze przewidywania ze średnią długością rezerwacji,
- stworzyć binarną klasyfikację lokali na te wynajmowane zazwyczaj na długi oraz na krótkoterminowy okres.

W poprzedniej wersji danych nie byliśmy w stanie powiązać rezerwacji z konkretnym lokalem, przez co takie analizy nie były możliwe. Sytuacja ta uległa zmianie wraz z otrzymaniem bardziej szczegółowych i kompletnych danych.

Kryterium sukcesu a model naiwny

> Jak porównuje się zaproponowane kryterium sukcesu do modelu naiwnego (zwracającego zawsze taki sam wynik)?

Klasyfikator (naiwny)

- W analizie klasyfikacyjnej przyjmujemy podział rezerwacji na:

Regresja Liniowa (naiwna)

- W przypadku regresji, model naiwny zwraca **średnią długość**

    - **krótkoterminowe** (≤ 7 dni),
    - **długoterminowe** (> 7 dni).
- Naiwny klasyfikator zawsze zwraca klasę większościową, czyli w tym przypadku **krótkoterminową rezerwację**.
- Taki model może osiągać **dokładność na poziomie ~74%**, ale **kompletnie nie radzi sobie z wykrywaniem rezerwacji długoterminowych**.
- Choć wysoka dokładność może sugerować dobry wynik, w rzeczywistości model ten **ignoruje klasę mniejszościową**, co czyni go bezużytecznym z perspektywy naszego celu – identyfikacji ofert prowadzących do długich rezerwacji.

rezerwacji dla danego ogłoszenia (wyliczoną na podstawie danych historycznych).

- Stanowi to prosty punkt odniesienia (baseline), względem którego oceniamy skuteczność bardziej złożonych modeli.
- **Błąd MAE** uzyskany przez model naiwny pozwala oszacować, czy opracowany model rzeczywiście wnosi wartość dodaną.
- Celem jest zbudowanie modelu, który **przewiduje długość pobytu z dokładnością do około 1 dnia (MAE ≈ 1)**. Wartość MAE modelu naiwnego służy tu jako dolny próg referencyjny.

	listing_id	total_bookings	total_duration	avg_duration
0	21853.0	33	48	1.454545
1	27262.0	29	144	4.965517
2	30320.0	172	360	2.093023
3	30959.0	8	12	1.500000
4	33945.0	78	540	6.923077

```python
short_listings = 0
long_listings = 0

short_listings = len(listings_statistics[listings_statistics['num_of_short_stays'] - listings_statistics['num_of_long_stays'] > 0])
long_listings = len(listings_statistics[listings_statistics['num_of_long_stays'] - listings_statistics['num_of_short_stays'] > 0])

print(f"Short Listings Amount : {short_listings}")
print(f"Long Listings Amount : {long_listings}")
print("Total Listings Amount : ", short_listings + long_listings)
print("Short Listings Percentage : ", short_listings / (short_listings + long_listings) * 100)
print("Long Listings Percentage : ", long_listings / (short_listings + long_listings) * 100)
```

✓ 0.0s

```
Short Listings Amount : 25226
Long Listings Amount : 6746
Total Listings Amount :  31972
Short Listings Percentage :  78.9002877517828
Long Listings Percentage :  21.09971224821719
```

```python
short_listings_total = listings_statistics['num_of_short_stays'].sum()
long_listings_total = listings_statistics['num_of_long_stays'].sum()
print(f"Short Listings Total : {short_listings_total}")
print(f"Long Listings Total : {long_listings_total}")
print("Total Listings Amount : ", short_listings_total + long_listings_total)
print("Short Listings Percentage : ", short_listings_total / (short_listings_total + long_listings_total) * 100)
print("Long Listings Percentage : ", long_listings_total / (short_listings_total + long_listings_total) * 100)
```

✓ 0.0s

```
Short Listings Total : 1471075
Long Listings Total : 511747
Total Listings Amount :  1982822
Short Listings Percentage :  74.1909762954012
Long Listings Percentage :  25.8090237045988
```

Zmiana metryk ewaluacyjnych – ROC AUC i AP (dla klasyfikatora)

> W przypadku klasyfikacji binarnej warto korzystać z metryk odpornych na niezbalansowanie klas, takich jak ROC AUC oraz AP.

Ze względu na silnie niezbalansowany rozkład klas (**74% krótkoterminowych, 26% długoterminowych**), stosowanie klasycznych metryk, takich jak dokładność (accuracy), może prowadzić do błędnych wniosków. Model zawsze przewidujący klasę dominującą osiąga wysoką accuracy, ale nie spełnia rzeczywistego celu analitycznego.

Dlatego jako główne metryki ewaluacyjne przyjęto:

- **ROC AUC (Area Under the ROC Curve)**

Metryka mierząca zdolność modelu do odróżniania klas, niezależnie od progu decyzyjnego. Szczególnie przydatna przy nierównomiernym rozkładzie klas.

- **Average Precision (AP)**

  Uśredniona wartość precyzji przy różnych poziomach recallu. Szczególnie istotna, gdy naszym celem jest wykrywanie klasy mniejszościowej (długoterminowych rezerwacji).

Wracając do modelu naiwnego klasyfikatora. Bierze on zawsze klasę większościową, więc :

- Osiąga wysoką dokładność (~74%), **ale ma zerowy recall i precision dla klasy „długi pobyt”**, przez co jest **bezużyteczny w praktyce**.
- ROC AUC = 0.5 (brak zdolności rozróżniania klas).
- AP (Average Precision) też będzie bardzo niskie.

Użycie tych metryk (ROC i AP) pozwala na bardziej rzetelną i sprawiedliwą ocenę skuteczności modelu, szczególnie w kontekście praktycznych zastosowań biznesowych – tj. **identyfikacji ofert, które mają większy potencjał generowania długich pobytów**.

Proste modele i ich wyniki

> Przy wstępnym modelowaniu warto "wrzucić" dane w prosty model i zobaczyć jakie wyniki osiągamy - pozwala to na określenie sensownego baseline'u dla problemu

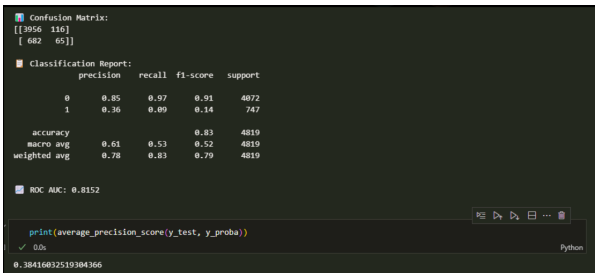
Postanowiliśmy stworzyć proste modele, żaby sprawdzić jakiego rodzaju wyników możemy się spodziewać. Wykorzystaliśmy modele z biblioteki sklearn :

- `LogisticRegression`
- `LinearRegression`

Oraz użyliśmy kolumn : ["number_of_reviews", "number_of_reviews_ltm", "number_of_reviews_l30d", "review_scores_rating", "reviews_per_month", "number_of_reviews_ly", "estimated_occupancy_l365d", "estimated_revenue_l365d", "bathrooms", "bedrooms", "beds"]

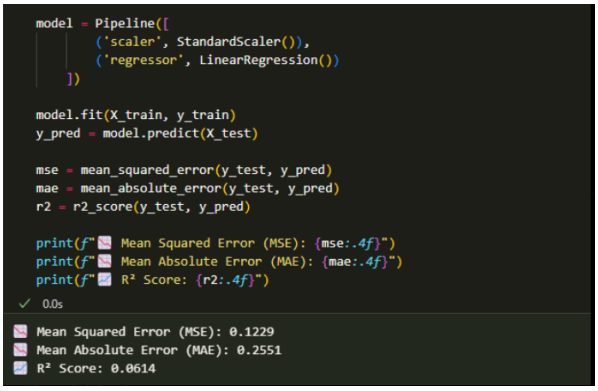
Klasyfikacja

- *Acc = 0.83*
- *ROC_SCORE = 0.81*
- *AP = 0.38*



Regresja

- *MAE = 0.2551*
- *MSE = 0.1229*



Nowe kryteria sukcesu

Biorąc pod uwagę, osiągnięte wyniki, oraz ilość zastosowanych danych i atrybutów oceniamy, że odpowiednimi kryteriami analitycznymi do celowania będą:

1. **Dla zadania regresji:**
    a. MAE na poziomie ~0.2 dnia
2. **Dla zadania klasyfikacji:**
    a. Accuracy nie będzie najlepszym wyznacznikiem jakości modelu, jednak modele osiągające dokładność wyższą niż 74% będą działały lepiej niż model naiwny.
    b. ROC na poziomie 0.85 - 0.9
    c. AP na poziomie 0.5 - 0.6

Dodatkowo, chcemy zbadać Feature Importance przy użyciu XGBoost’a, ale to w kolejnychiteracjach