Ceny hoteli w Amsterdamie

Natalia Dyś

8 czerwca 2024

Spis treści

1 Wstep

1.1 Cel raportu

Raport ma na celu oszacowanie ceny per osoba jednonocnego pobytu w hotelu w Amsterdamie z użyciem metod regresji. Cena bedzie szacowana na podstawie parametrów takich jak:

- liczba osób
- data pobytu
- ocena hotelu (opinie użytkowników oraz ich ilość)
- liczba gwiazdek
- dystans od centrum

1.2 Zbieranie danych

Analizowane przeze mnie dane zostały zebrane poprzez generowanie losowych requestów do strony booking.com, a nastepnie wyodrebnienie poszczególnych ofert z html otrzymanego przy pomocy selenium i BeautifulSoup. Dane z każdego wyszukania sa zapisywane do odrebnego pliku csv (umożliwia to szybka analize mniejszego skrawka danych).

Losowo generowane parametry wyszukiwań to:

- liczba dorosłych osób - od 1 do 10
- data - od 30 do 400 dni od obecnej (w chwili wywołania scrapera) daty.

Ceny sa zbierane i zapisywane w euro.

Zakładamy pobyt jednodniowy, ponieważ długość pobytu nie ma bezpośrednio wpływu na cene, a dzieki temu maksymalizujemy ilość dostepnych obiektów.

Amsterdam jest arbitralnie wybrana lokalizacja. Jest on zapisywany w kolumnie 'city' aby działalność algorytmu można było potencjalnie rozszerzyć na inne miasta.

Nie uwzgledniam dat z najbliższego miesiaca, ponieważ wiekszość miejsc bedzie już prawdopodobnie zajeta, co może generować niemiarodajne dane.

Zdecydowałam sie też w moich wyszukaniach nie generować liczby dzieci (wynosi zawsze 0), jednak w kodzie istnieje taka możliwość.

2 Wstepna analiza danych

2.1 ilość zebranych rekordów

Ostatecznie dane zostały zebrane dla 400 requestów. Ilość listingów dla każdego wyszukania waha sie od 26 do 485 (booking wyświetla tylko obiekty dostepne w danym terminie dla podanej liczby osób) Łaczna ilość zebranych ofert to 42488, co daje średnio 106,22 per wyszukanie.

2.2 zbierane wartości

- name - nazwa hotelu
- city - miasto (parametr wyszukiwania)
- stars - liczba gwiazdek
- rating - ocena hotelu
- opinions - liczba ocen hotelu (czasami referowane jako reviews)
- distance_from_centre - odległość hotelu od centrum
- free_cancellation - czy istnieje możliwość darmowego odwołania rezerwacji (bool)
- breakfast - czy hotel oferuje śniadanie (bool)
- price - pełna cena rezerwacji
- price_per_person - cena rezerwacji podzielona przez ilość osób (dorośli + dzieci)
- date - data rezerwacji (parametr wyszukiwania)
- adults - liczba dorosłych (parametr wyszukiwania)
- children - liczba dzieci (parametr wyszukiwania)

2.3 usuniecie niepotrzebnych kolumn

Ponieważ dane zebrałam w tym przypadku tylko dla jednego miasta oraz nie uwzgledniam udziału dzieci, odpowiednie kolumny zostały usuniete.

Nazwa hotelu nie bedzie miała wpływu na jego cene. Z późniejszej analizy wynikło, że darmowe odwołanie jest we wszystkich rekordach oznaczone jako prawda (prawdopodobnie przez lekko przekłamany sposób wyświetlania bookingu), wiec możemy zignorować obie te kolumny.

W tym raporcie skupiam sie na cenie per osoba, dlatego kolumna z łaczna cena rezerwacji również powinna zostać usunieta. (inaczej wyliczalibyśmy cene na podstawie samej siebie)

2.4 brakujace dane

Collumn	Count Missing	Percentage Missing
rating	332	0.78
opinions	258	0.61

Brakujacych danych jest na szczeście niewiele i znajduja sie tylko w dwóch kolumnach, wiec nie powinny znaczaco wpłynać na wyniki.

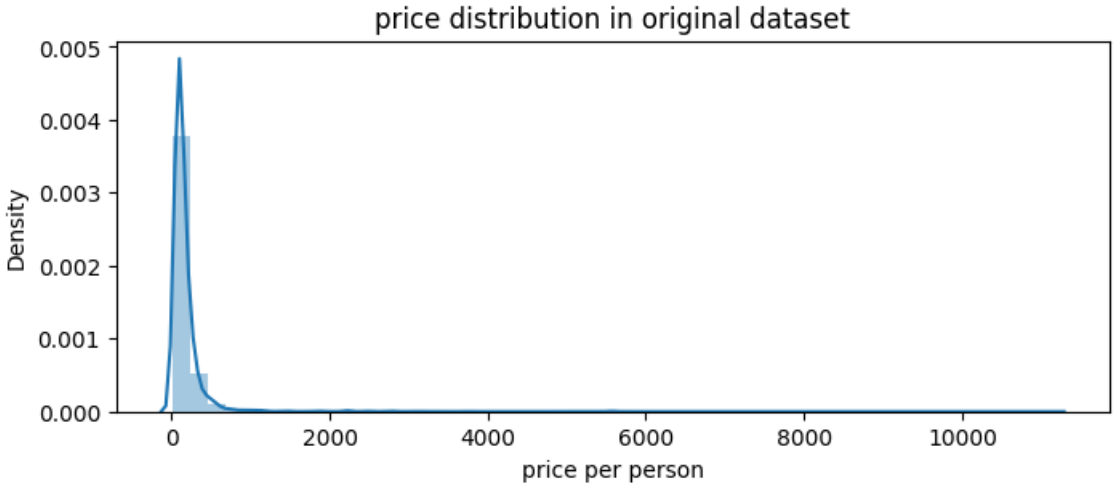
Według mojej analizy, liczby opinii bedzie brakować kiedy jest ona równa 0 (booking nie wyświetla wtedy liczby, dlatego parser jej nie wyłapuje). Zakładam, że ten fakt jest odpowiedzialny za wiekszość brakujacych wartości, dlatego zostana one wypełnione zerami.

Z ratingiem bedzie podobnie, jednak tutaj tutaj logiczniejsze wydaje sie uzupełnienie wartości losowa liczba z zakresu średnia $\pm$ odchylenie standardowe. (ma spore szanse być poprawna wartościa jeśli obiekt rzeczywiście miałby wystawione opinie)

2.5 analiza danych

2.5.1 cena

min	max	średnia	odchylenie std.
14.00	11147.00	181.15	400.67

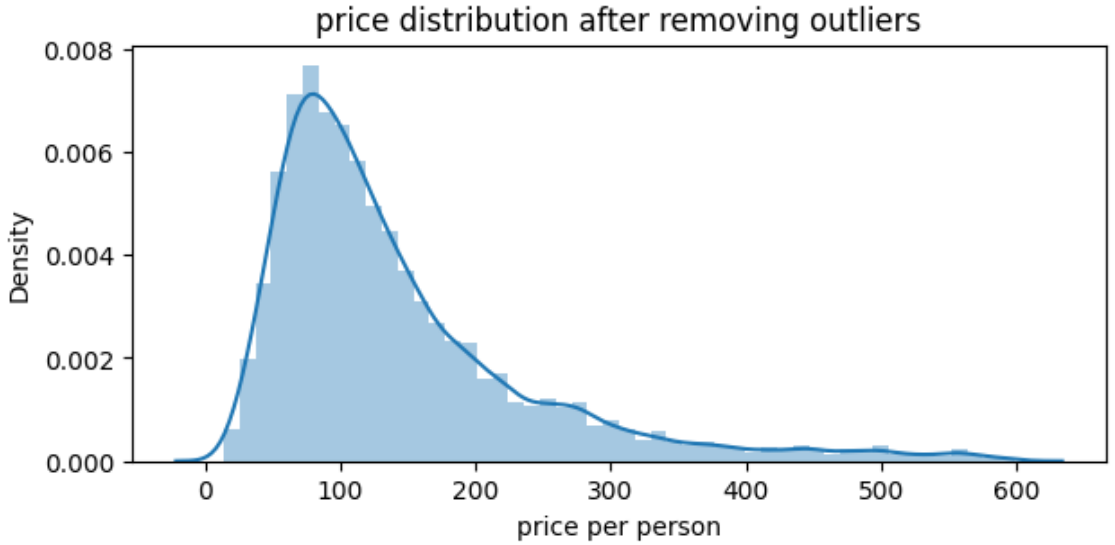


Rysunek 1: dystrybucja cen w oryginalnym zbiorze danych

Jak widać na wykresie, niewielka liczba skrajnych wartości jest odpowiedzialna spore odchylenie, które nie oddaje rzeczywistego rozkładu wiekszości danych. Z tego powodu zdecydowalam sie usunać cześć wyników. Za punkt odciecia uznałam cene 600 € per osoba.

W wyniku tej operacji tracimy 2,34 % oryginalnych danych, jednak odchylenie standardowe spada czterokrotnie (a zakres standardowych cen przestaje zawierać ujemne wartości).

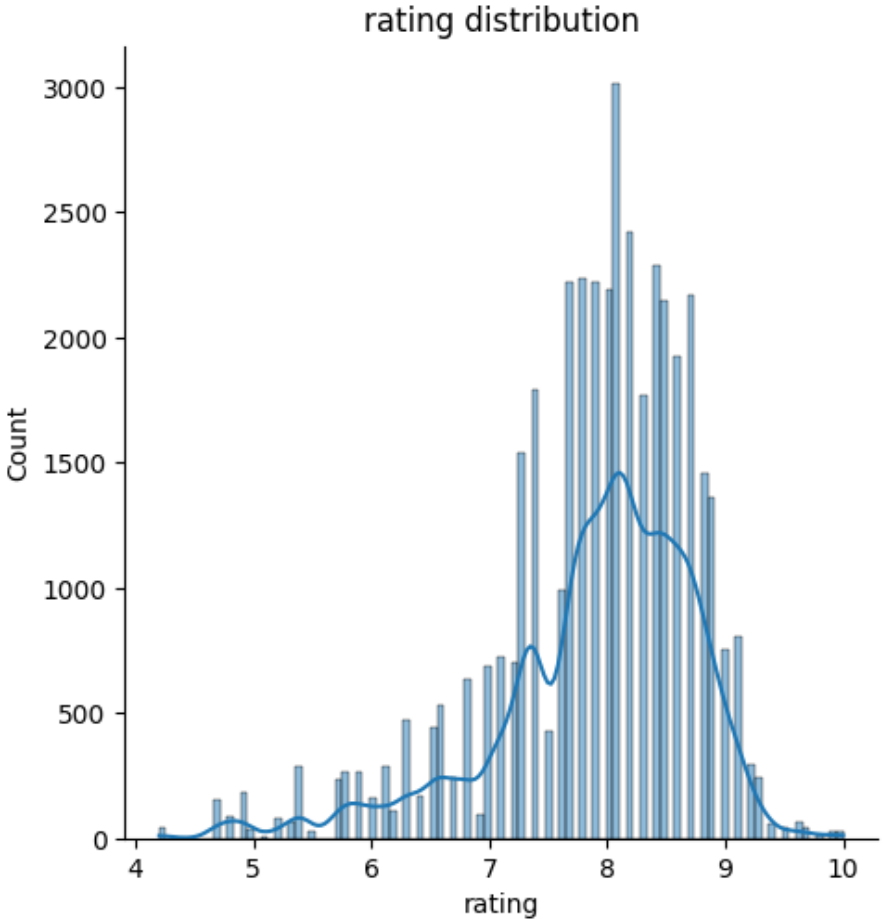
min	max	średnia	odchylenie std.
14.00	598.60	142.90	99.03



Rysunek 2: dystrybucja cen po usunieciu skrajnych wartości

2.5.2 ocena użytkowników

min	max	średnia	odchylenie std.
4.2	10	7.89	0.86



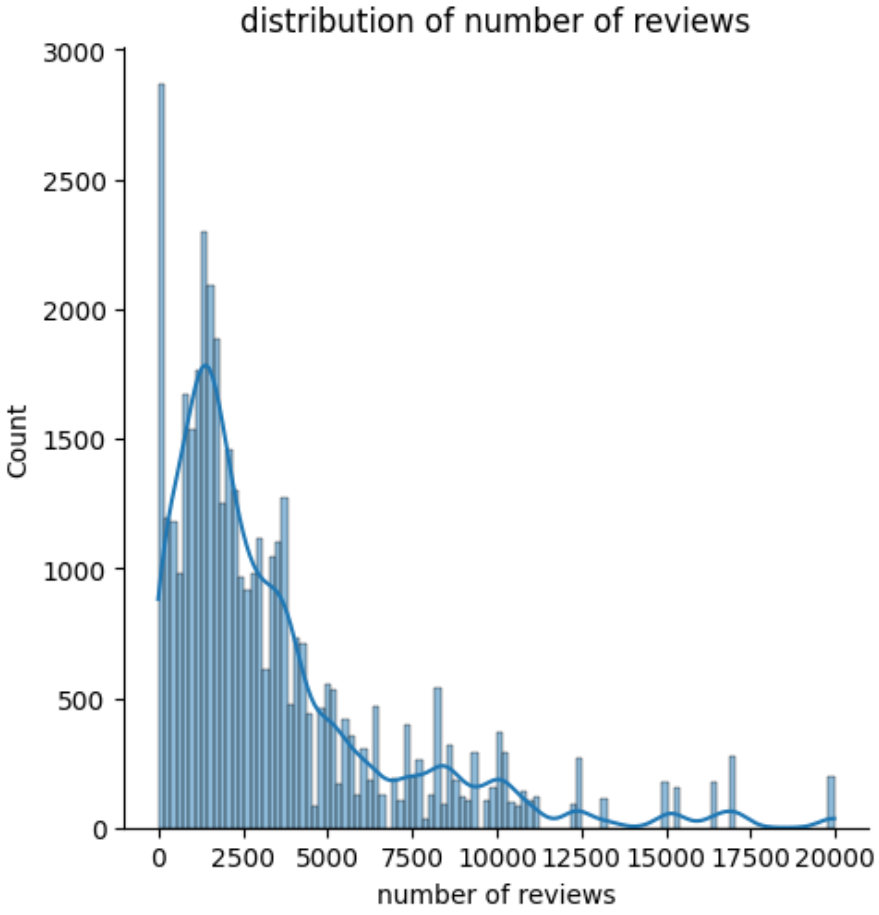
Rysunek 3: dystrybucja ocen

Rysunek 4: zależność ceny rezerwacji od oceny hotelu

Na wykresie można zobaczyć że istnieje pozytywna korelacja miedzy ocena hotelu a cena jaka musimy zapłacić za pobyt, jednak zależność nie jest liniowa i da sie ja zauważyć głównie przy ocenach powyżej 8.

2.5.3 liczba ocen

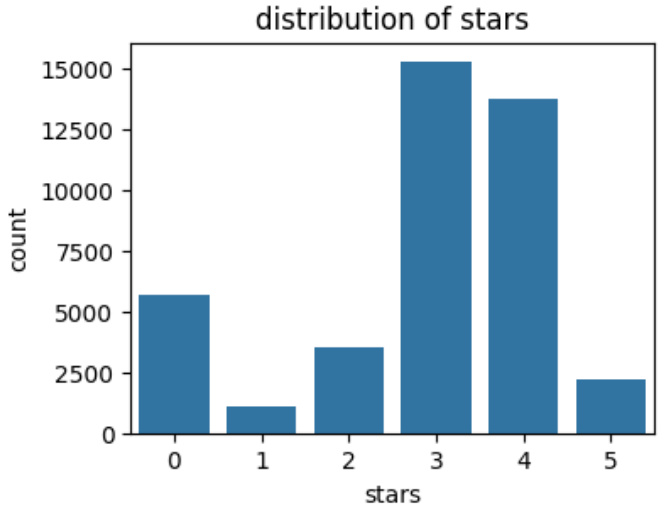
min	max	średnia	odchylenie std.
0.00	19967.00	3424.81	3474.33



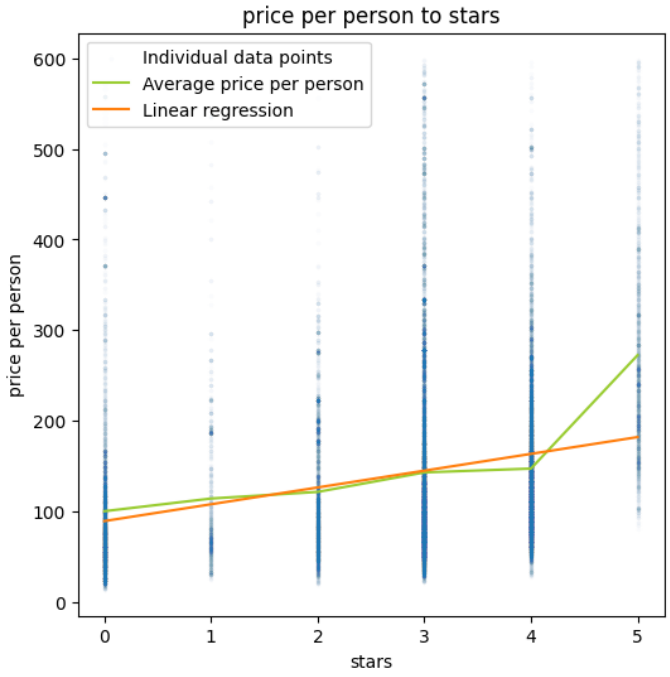
Rysunek 5: dystrybucja liczby ocen

2.5.4 liczba gwiazdek

min	max	średnia	odchylenie std.
0	5	2.89	1.39

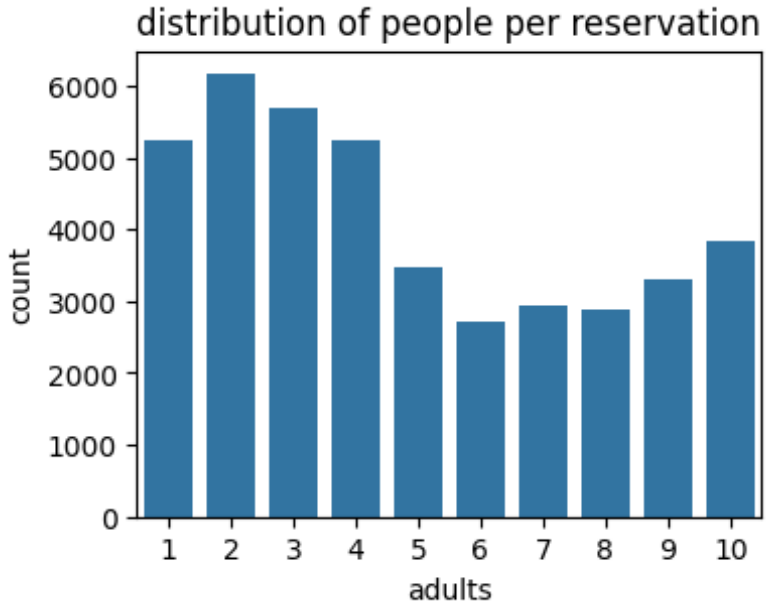


Rysunek 6: dystrybucja liczby gwiazdek

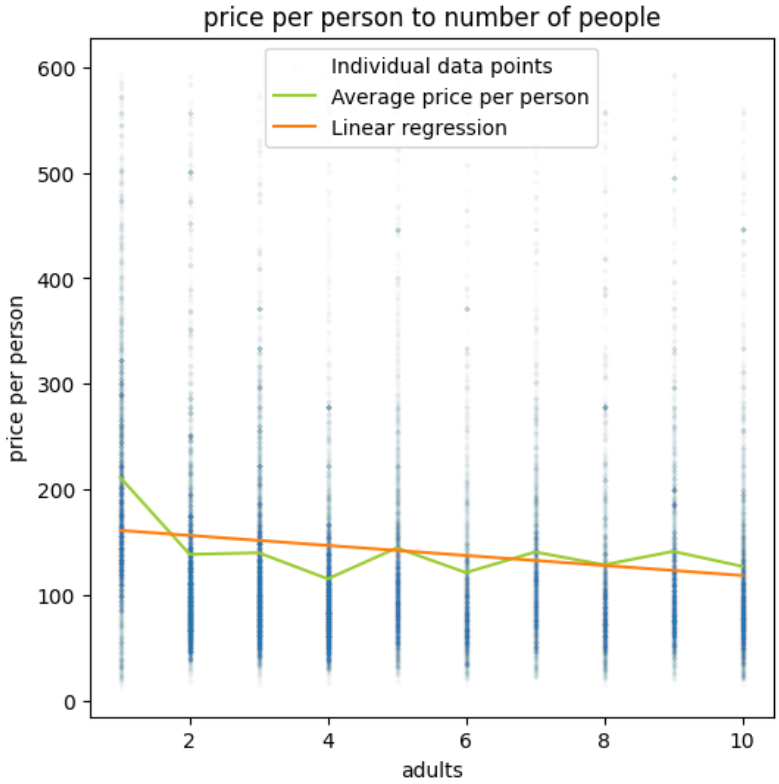


Rysunek 7: zależność ceny od liczby gwiazdek

2.5.5 liczba osób



Rysunek 8: dystrybucja liczby osób



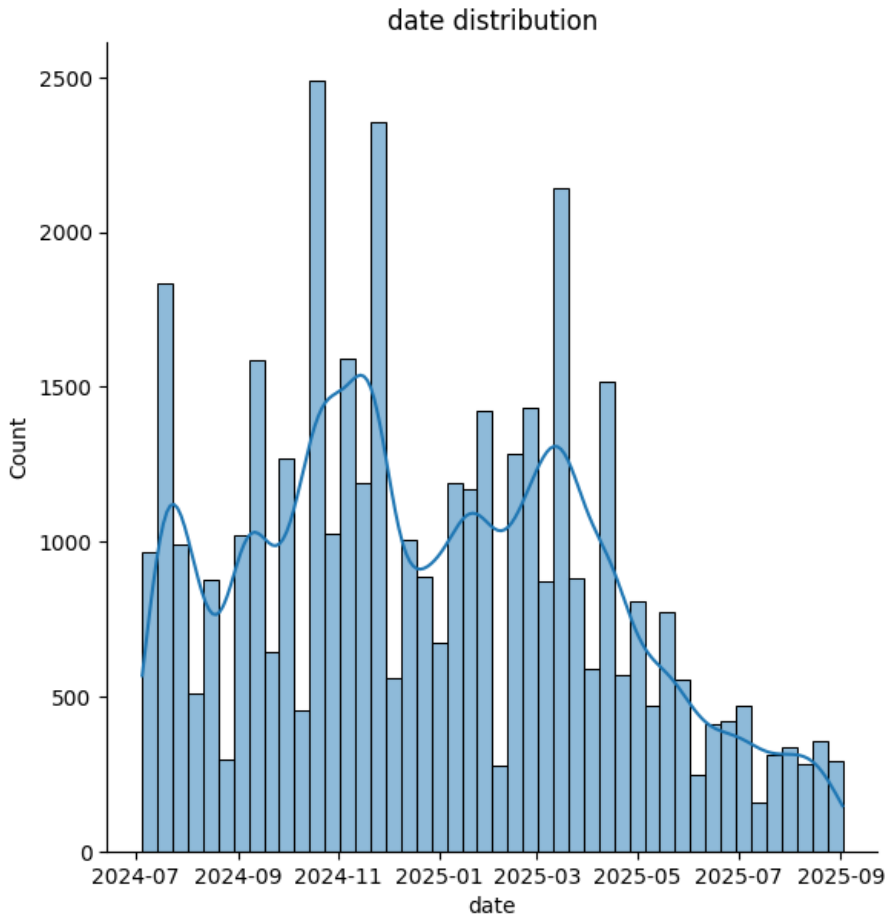
2.5.6 data

Daty mieszcza sie w zakresie od 2024-07-05 do 2025-09-02

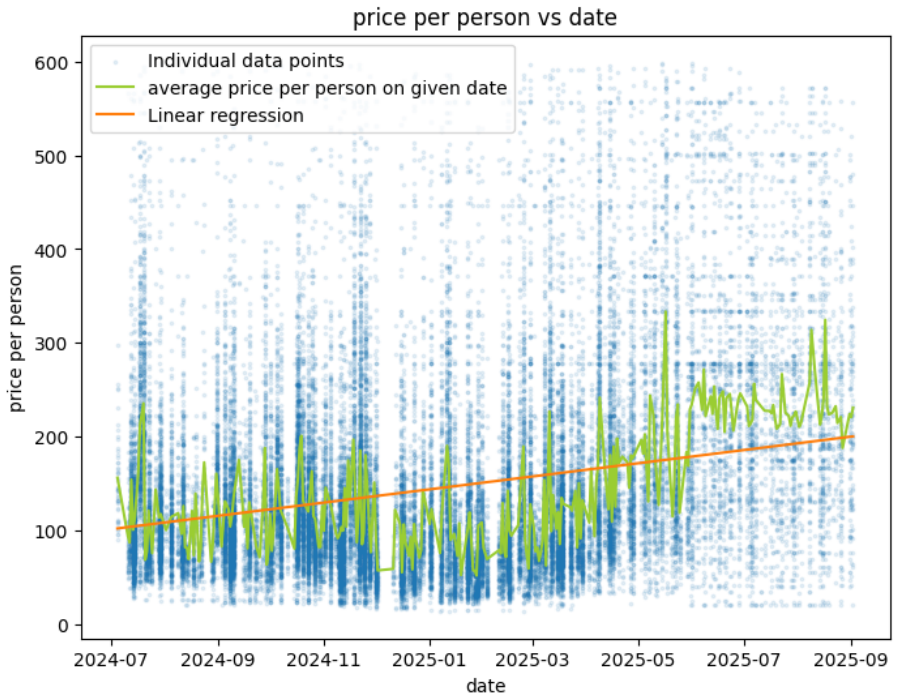
Rozkład wydaje sie dość losowy, prawdopodobnie jest on po cześci dyktowany zainteresowaniem (jeśli data jest oblegana pokaże sie mniej ofert). Potencjalnie można w ten sposób wytłumaczyć spadek liczby ofert w okolicy świat i sylwestra. Tutaj wpływ mogło mieć także odfiltrowanie rekordów ze skrajnymi cenami.

Liczba ofert zaczyna od pewnego momentu maleć wraz z czasem. Może to być powiazane zarówno z nakładajacym sie okresem wakacyjnym, jak i odległym terminem, przez co hotele moga jeszcze nie przyjmować rezerwacji.

W celu poprawnego szacowania ceny prawdopodobnie trzebaby uwzgledнić zarówno date pobytu, jak i date składania rezerwacji oraz analizować dane zebrane w wiekszym przedziale czasu.



Rysunek 9: dystrybucja dat



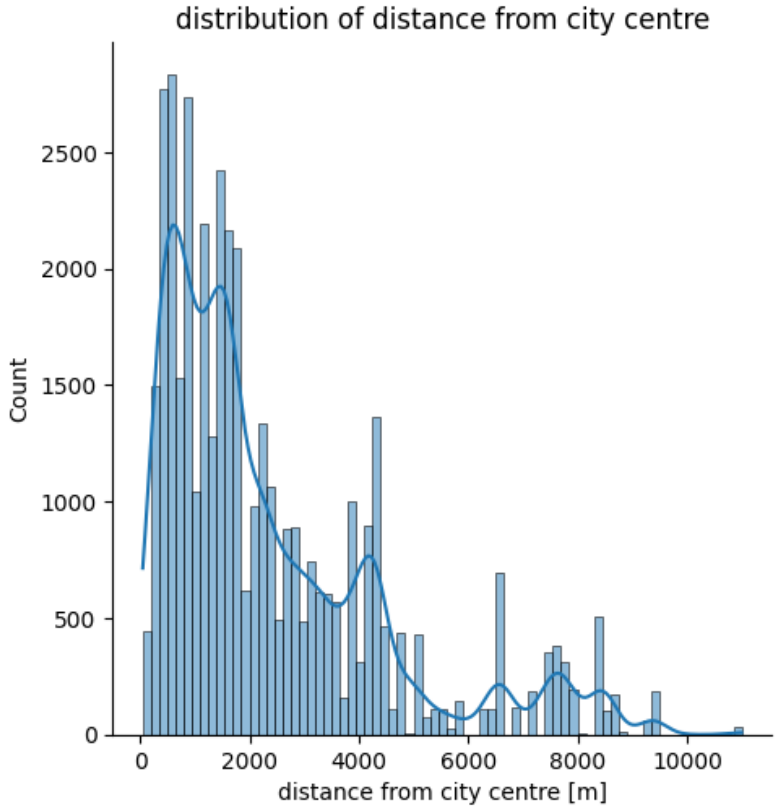
Rysunek 10: zależność ceny od daty rezerwacji

Na wykresie zależności ceny od daty możemy zobaczyć spory skok cen w okresie wakacyjnym. Potencjalnie na wyższa cene może na wpływać także mniejsza liczba dostepnych ofert. Aby jednoznacznie określić powód należałoby uwzgledniać dane zbierane na przestrzeni czasu i dodać kolumny odpowiadajace za date zebrania danych oraz liczbe dostepnych ofert dla określonych parametrów wyszukania.

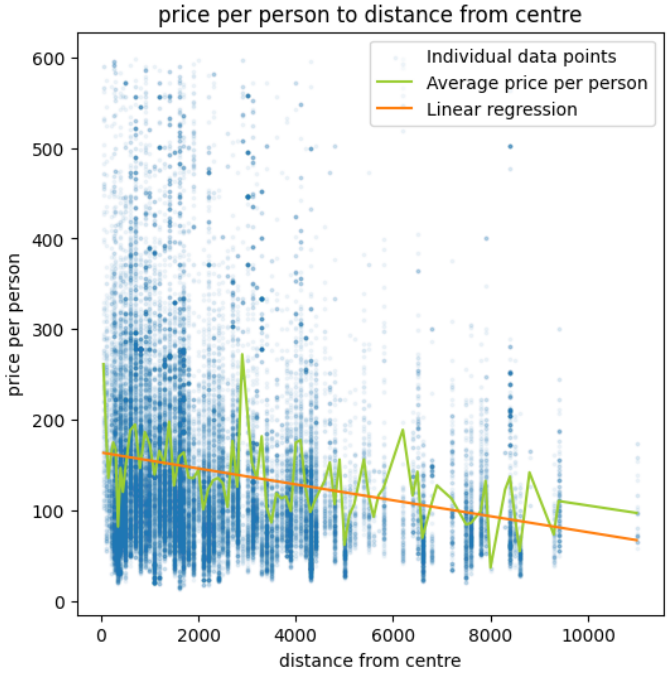
2.5.7 odległość od centrum

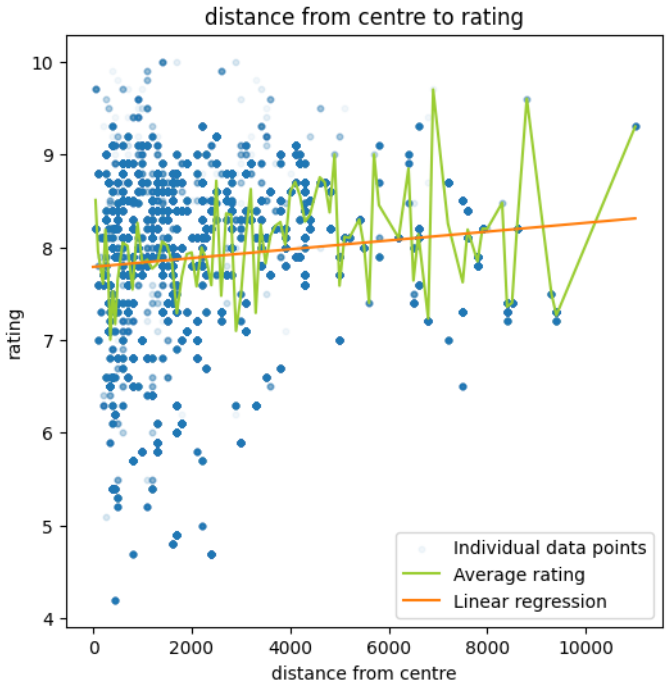
min	max	średnia	odchylenie std.
50.00	11000.00	2378.90	2079.47

Zdecydowanie najwiecej ofert znajduje sie w okolicy kilometra od centrum. Liczba ofert maleje wraz z dystansem. Wystepuje negatywna korelacja miedzy cena a dystansem. Wbrew moim podejrzeniom, odległość od centrum nie wydaje sie mieć zbyt dużego wpływu na ocene hotelu (zachowujemy obie kolumny).



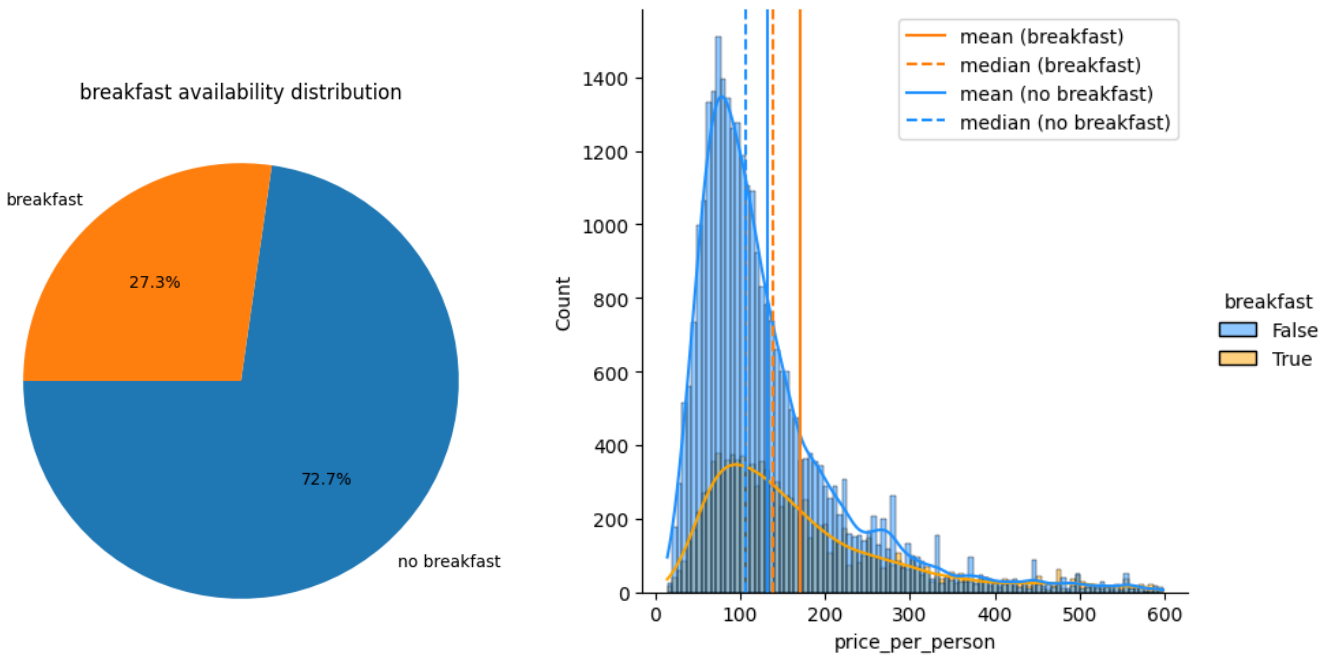
Rysunek 11: dystrybucja liczby osób


price per person to distance from centre
Individual data points
Average price per person
Linear regression
price per person
distance from centre
0
100
200
300
400
500
600
0
2000
4000
6000
8000
10000


distance from centre to rating
Individual data points
Average rating
Linear regression
rating
distance from centre
4
5
6
7
8
9
10
0
2000
4000
6000
8000
10000

2.5.8 śniadanie

Wraz z wzrostem ceny rośnie procent hoteli oferujacych śniadanie, co przekłada sie na wyższa średnia cene w tej grupie.



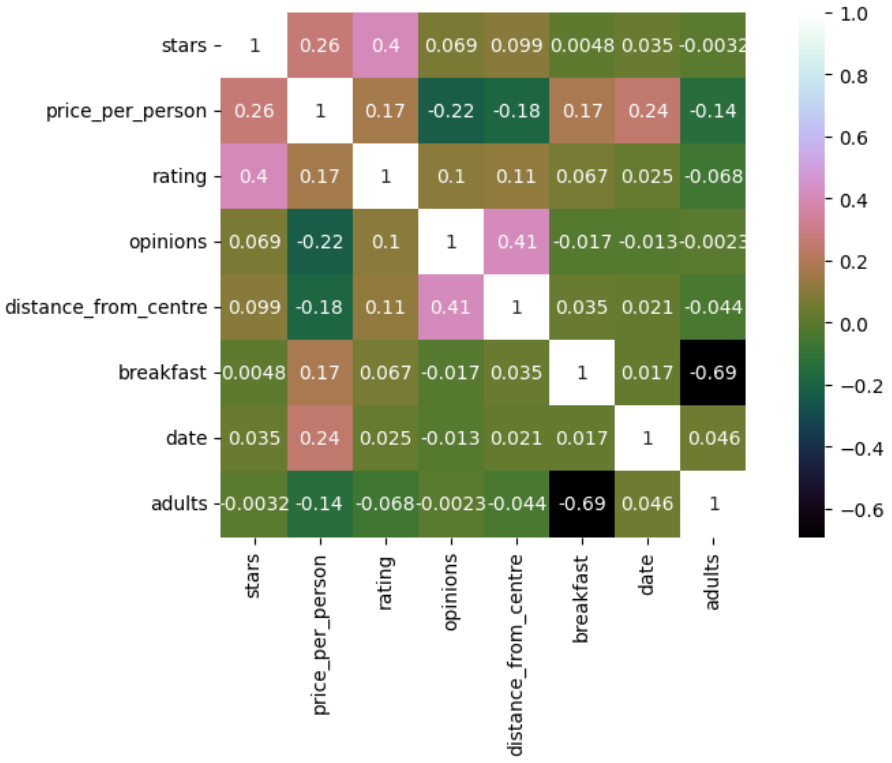
Rysunek 12: dystrybucja cen w zależności od dostepności śniadania

2.5.9 macierz korelacji

Na macierzy korelacji możemy zobaczyć, że wszystkie zmienne maja wpływ na cene. Wiekszość kolumn ma bardzo mała korelacje ze soba nawzajem. Wyjatkiem jest rating, na który reszta parametrów ma lekko wiekszy wpływ (co było dość spodziewane).

Widać też zależność miedzy liczba opinii a odległościa od centrum (prawdopodobnie wiekszy popyt na hotele w środku miasta). Nie jest ona jednak na tyle duża aby uznać któraś z kolumn za niepotrzebna.

Lekka anomalia jest silna ujemna korelacja miedzy dostepnościa śniadania a liczba osób. Trudno stwierdzić czym jest spowodowana.



Rysunek 13: macierz korelacji

3 Testowane modele

Dane zostały przeskalowane za pomoca StandardScaler (data wcześniej przekonwertowana na czas porzadkowy), a nastepnie losowo podzielone na zbiór treningowy i testowy w proporcji 70:30. Przetestowane zostały:

- 4 algorytmy regresji liniowej: least squares (lls), non-negative ls, ridge i lasso.
- least squares na wielomianach wiekszego stopnia
- random forest z różna liczba drzew decyzyjnych
- k-neighbours z różna liczba uwzglednianych sasiadów

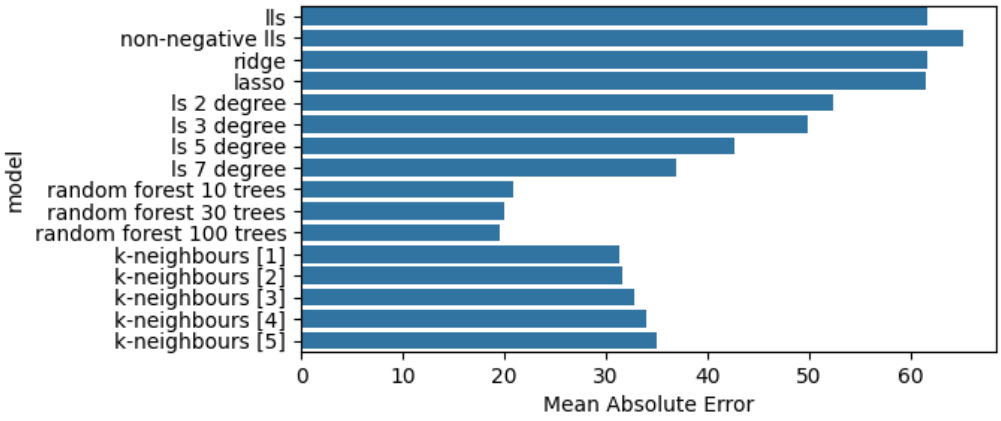
4 Wyniki i wnioski

Zdecydowanie najlepiej wypada alborytm random forest, nawet przy małej liczbie drzew decyzyjnych. Zwiekszenie liczby drzew powoduje niewielka poprawe kosztem szybko rosnacego czasu obliczeń.

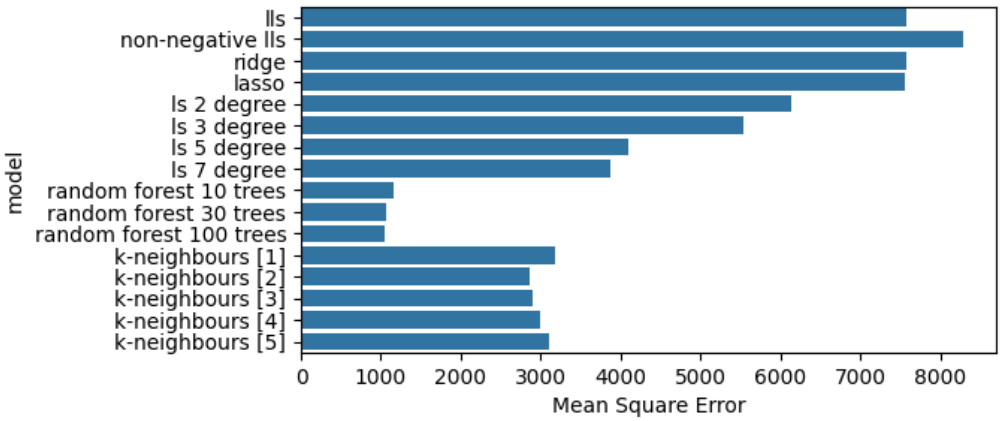
Drugi w kolejności jest algorytm k najbliższych sasiadów z różnica we współczynniku determinacji r2 równa około 0,2. Algorytm działa najlepiej przy uwzglednieniu 3 sasiadów jeśli za metryke uznamy współczynnik determinacji i MSE. Najmniejszy średni bład bezwzgledny (MAE) osiaga jednak dla jednego sasiada.

Algorytmy optymalizujace wielomian pierwszego stopnia wypadaja najgorzej i wszystkie osiagneły niemal identyczne wyniki (różnica przy zmianie parametrów również była prawie niezauważalna). Jedyny, który wypadł lekko gorzej od reszty to Non-negative LLS. Generalne niepowidzenie modeli liniowych ma sens, ponieważ nawet na wykresach możemy zobaczyć, że nie wszystkie zmienne maja liniowy (lub zbliżony do liniowego) wpływ na cene.

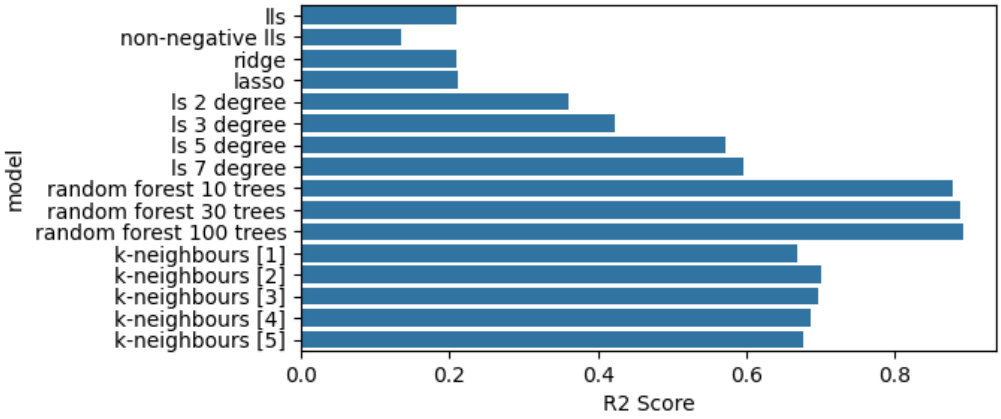
Zwiekszanie stopnia optymalizowanego wielomianu poprawia wyniki (zależność miedzy stopniem wielomiany a współczynnikien dopasowania bliska liniowej). Niestety, nawet drugiemu w kolejności algorytmowi k najbliższych sasiadów, ls zaczyna dorównywać dopiero przy wielomianie 7 stopnia, podczas gdy czas wywołania robi sie ponad 200-krotnie wiekszy



Rysunek 14: średni bład bezwzgledny



Rysunek 15: blad średniokwadratowy



Rysunek 16: współczynnik determinacji

4.1 propozycje rozwoju

Ponieważ booking pokazuje tylko obecnie dostepne listingi, a ceny moga sie dynamicznie zmieniać w zależności od czasu i dostepności miejsc, sadze że zbieranie danych na przestrzeni czasu oraz uwzglednienie zarówno daty rezerwacji jak i daty wyszukania mogłoby dawać bardziej użyteczne wyniki. (z obecnym zbiorem danych wiarygodnie da sie przewidzieć jedynie propozycje cen, które otrzymalibyśmy składajac rezerwacje około 30 maja 2024)

Aby rozszerzyć zakres eksperymentu w generowanych zapytaniach możnaby też uwzględnić różne miasta. (Przeprowadzić ten sam test na innym mieście i porównać wyniki poszczególnych algorytmów lub uwzględnić miasto jako zmienna regresji)