

# Czy da się przewidzieć ruch turystyczny w polskich powiatach?(ML)

Analiza danych BDL GUS  
(2013–2024)  
+ model predykcyjny (ML)



**Projekt grupowy / raport**  
na podstawie danych  
statystycznych



**Zakres:** powiaty w Polsce  
(w tym miasta na prawach  
powiatu)



**Wykorzystane narzędzia:**  
Python, Jupyter, Pandas,  
scikit-learn



**Cel prezentacji:** pokazać  
trend, różnice przestrzenne  
oraz wyniki prognoz

# Dlaczego ten temat jest ważny?



**Turystyka realnie wpływa na gospodarkę lokalną:** miejsca pracy, usługi, budżety gmin i powiatów



**Samorządy muszą planować:** infrastrukturę, komunikację, promocję regionu, bazę noclegową



**Prognoza ruchu turystycznego pomaga:** lepiej zarządzać sezonem, podejmować decyzje inwestycyjne, przygotować służby



**Pytanie: czy da się to zrobić “z danych”, bez zgadywania?**

# Cel projektu



**Główny cel:** sprawdzić, czy można przewidzieć ruch turystyczny w powiatach, używając danych BDL GUS



**Zmienna docelowa (Y):** liczba udzielonych noclegów (noclegi w obiektach noclegowych)

## Cele szczegółowe:



1. analiza trendu w czasie (2013–2024)
2. wskazanie liderów i zmian w TOP powiatach
3. sprawdzenie zależności pomiędzy cechami (korelacje)
4. porównanie modeli ML i wybór najlepszego
5. prognoza wartości dla roku 2024 i interpretacja błędów



## Pytania badawcze

- Czy w skali kraju widać stabilny wzrost noclegów?



- Jak pandemia 2020 wpłynęła na dane i czy nastąpiła odbudowa?



- Które powiaty generują największą liczbę noclegów i dlaczego?



- Które czynniki są najsilniej powiązane z liczbą noclegów: popyt czy podaż?



- Który model regresji daje najlepszą jakość predykcji i dlaczego?



- Gdzie model myli się najbardziej i z jakich powodów?



# Źródło danych: BDL GUS



Dane pochodzą z Banku  
Danych Lokalnych GUS –  
oficjalne statystyki publiczne



Zalety BDL: wiarygodność,  
spójność definicji,  
dostępność dla wielu lat



Jednostka terytorialna:  
powiat (w tym miasta na  
prawach powiatu)



Dane roczne → brak pełnej  
sezonowości, ale dobry obraz  
trendów długoterminowych



Dane obejmują zarówno  
turystykę miejską, biznesową,  
jak i wypoczynkową

# Co dokładnie mierzamy? (zmienna docelowa)

**Liczba udzielonych noclegów** – kluczowa miara ruchu turystycznego:

- pokazuje realne wykorzystanie bazy noclegowej
- jest bardziej „twarda” (bo uwzględnia długość pobytu)
- Zależy od:
  - liczby odwiedzających
  - czasu pobytu
  - dostępnej infrastruktury noclegowej



Wysokie noclegi = duży popyt, obciążenie usług i potencjalne zyski regionu



Pokazuje **realne wykorzystanie bazy** noclegowej



**Jest bardziej „twarda”** niż liczba turystów (uwzględnia długość pobytu)

Zależy od:



Liczba odwiedzających



Czas pobytu



Dostępna infrastruktura

# Zmienne objaśniające (co może wpływać na noclegi)

Przykładowe cechy (X), użyte w analizie:



**liczba turystów**  
bezpośredni  
popyt



**miejscia noclegowe** –  
pojemność,  
„ile osób można przyjąć”



**liczba obiektów  
noclegowych** –  
struktura podaży  
(hotele, pensjonaty itd.)



**ludność / gęstość  
zaludnienia** –  
potencjał usług, transport,  
wielkość ośrodka



**lesistość** –  
czynnik  
środowiskowy  
(atrakcyjność naturalna)



Zmienna docelowa i cechy  
występują w układzie powiat–rok  
(panel danych)

# Przygotowanie danych (pipeline)



**Budowa zbioru panelowego:**  
każdy rekord = konkretny powiat w danym roku

# Cechy opóźnione (lag features) – po co?



To zbliża ML do logiki prognoz czasowych, ale nadal w ujęciu regresji

# Narzędzia i technologia

## Narzędzia i technologia



- **Python + Jupyter Notebook** – praca analityczna krok po kroku
- **Pandas** – czyszczenie, łączenie, transformacje danych
- **Matplotlib/Seaborn** – wykresy trendów, rankingi, korelacje
- **scikit-learn** – trenowanie modeli, podział danych, metryki
- Podejście: porównujemy kilka modeli i wybieramy najlepszy na testach



**Pandas** –  
czyszczenie,  
łączenie,  
transformacje  
danych



**Matplotlib/Seaborn** –  
wykresy trendów,  
rankingi, korelacje



**scikit-learn** –  
trenowanie  
modeli, podział  
danych, metryki



Podejście: porównujemy kilka modeli i  
wybieramy najlepszy na testach



## Trend w Polsce 2013–2019 (wzrost)

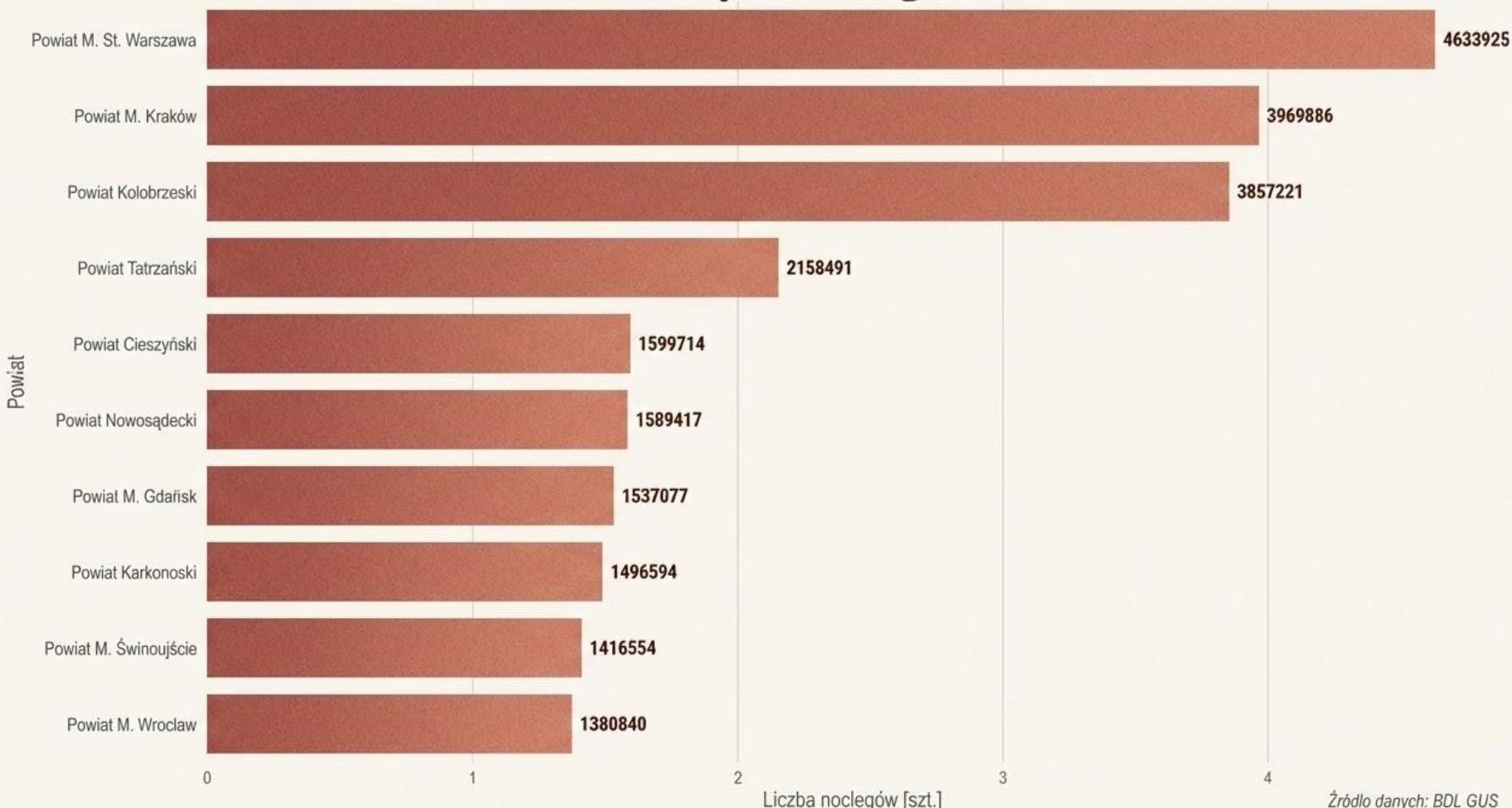
- W latach 2013–2019 widać regularny wzrost noclegów w Polsce
- Oznacza to rosnącą aktywność turystyczną oraz rozwój infrastruktury

### Możliwe przyczyny:

- ✈️ rozwój tanich linii i transportu
- 💰 rosnące dochody i mobilność ludzi
- 📢 większa promocja miast i regionów
- 🏨 inwestycje w hotele, apartamenty, atrakcje

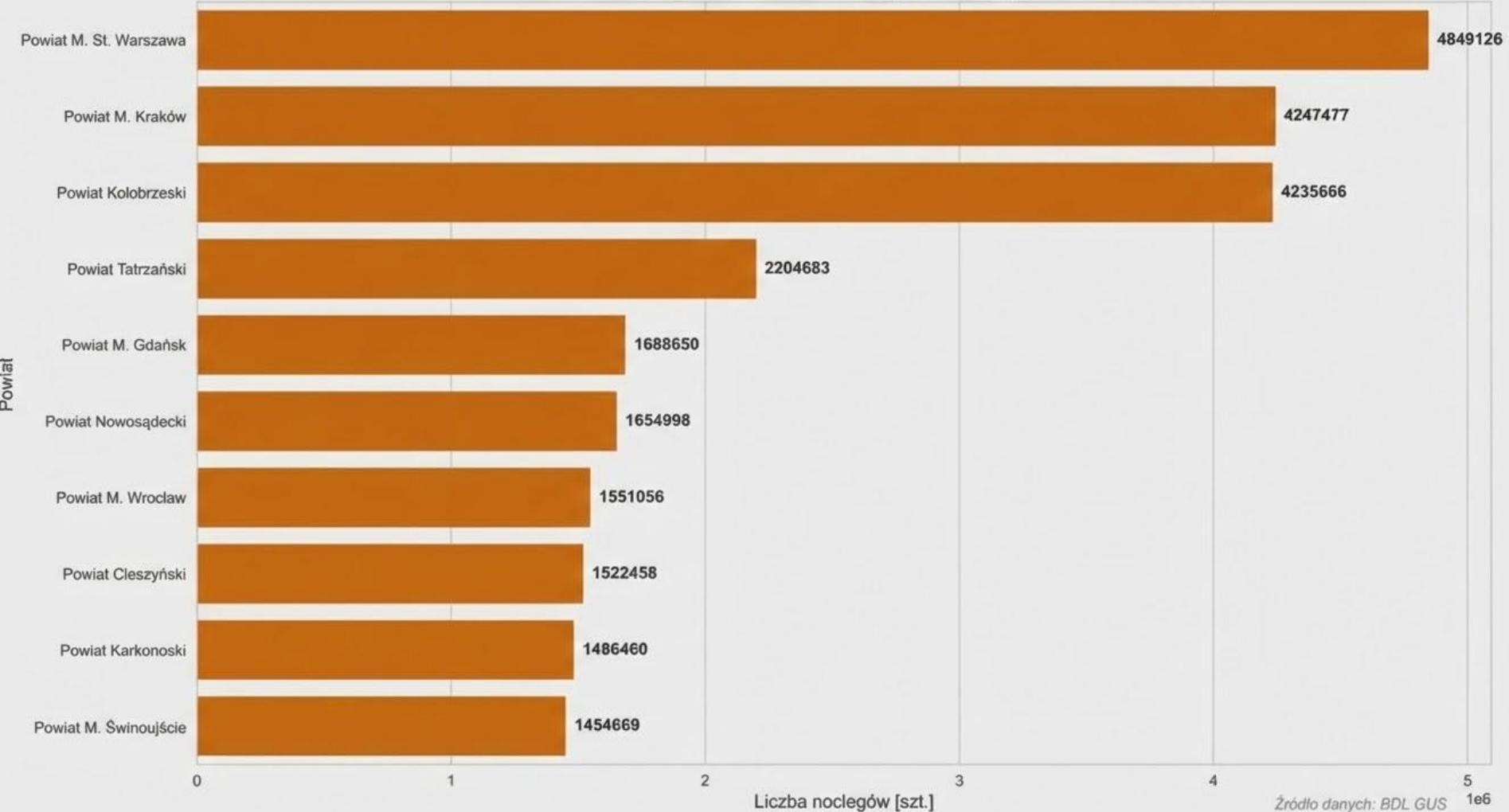
**Wniosek:** przed pandemią trend był stabilnie rosnący

# TOP 10 Powiatów: Liczba udzielonych noclegów w 2013 roku



Źródło danych: BDL GUS

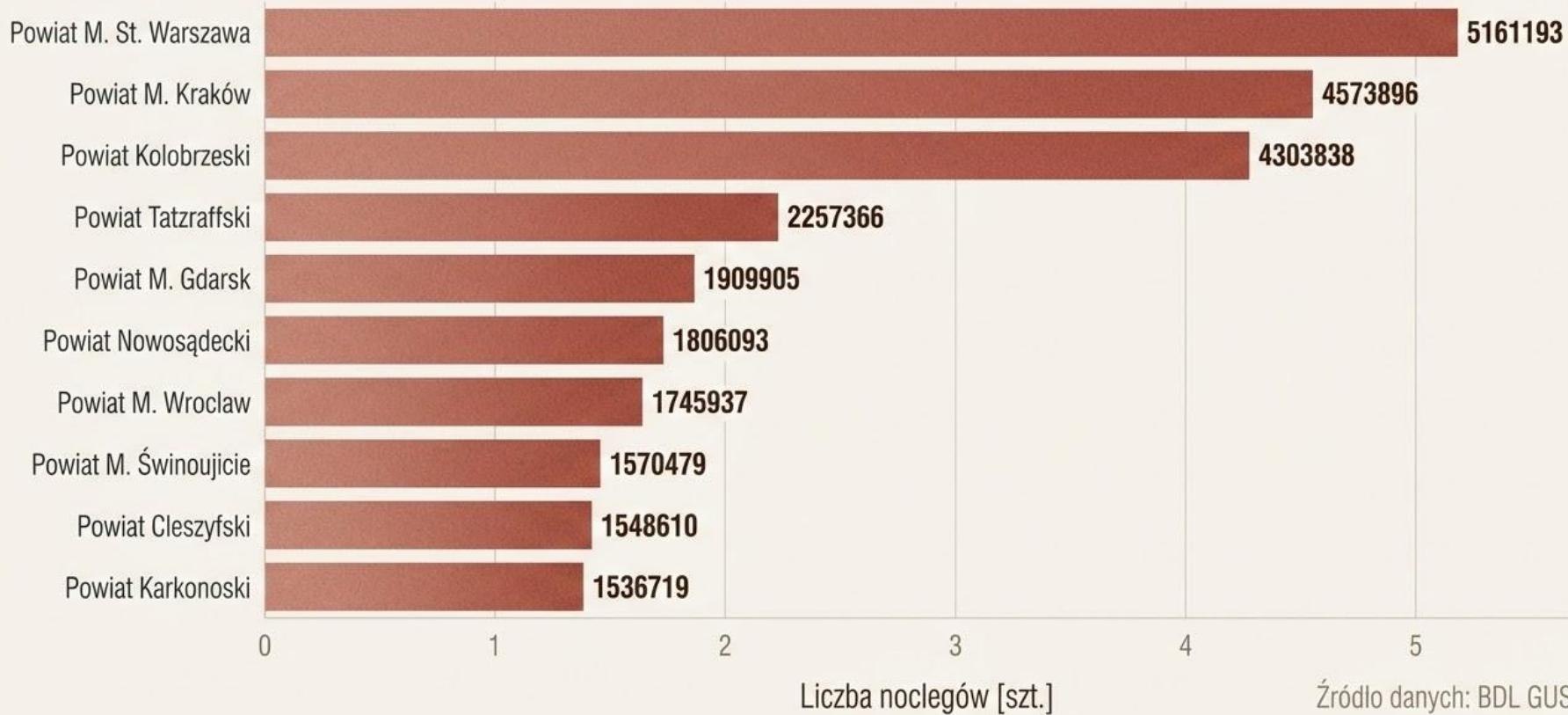
## TOP 10 Powiatów: Liczba udzielonych noclegów w 2014 roku



Źródło danych: BDL GUS

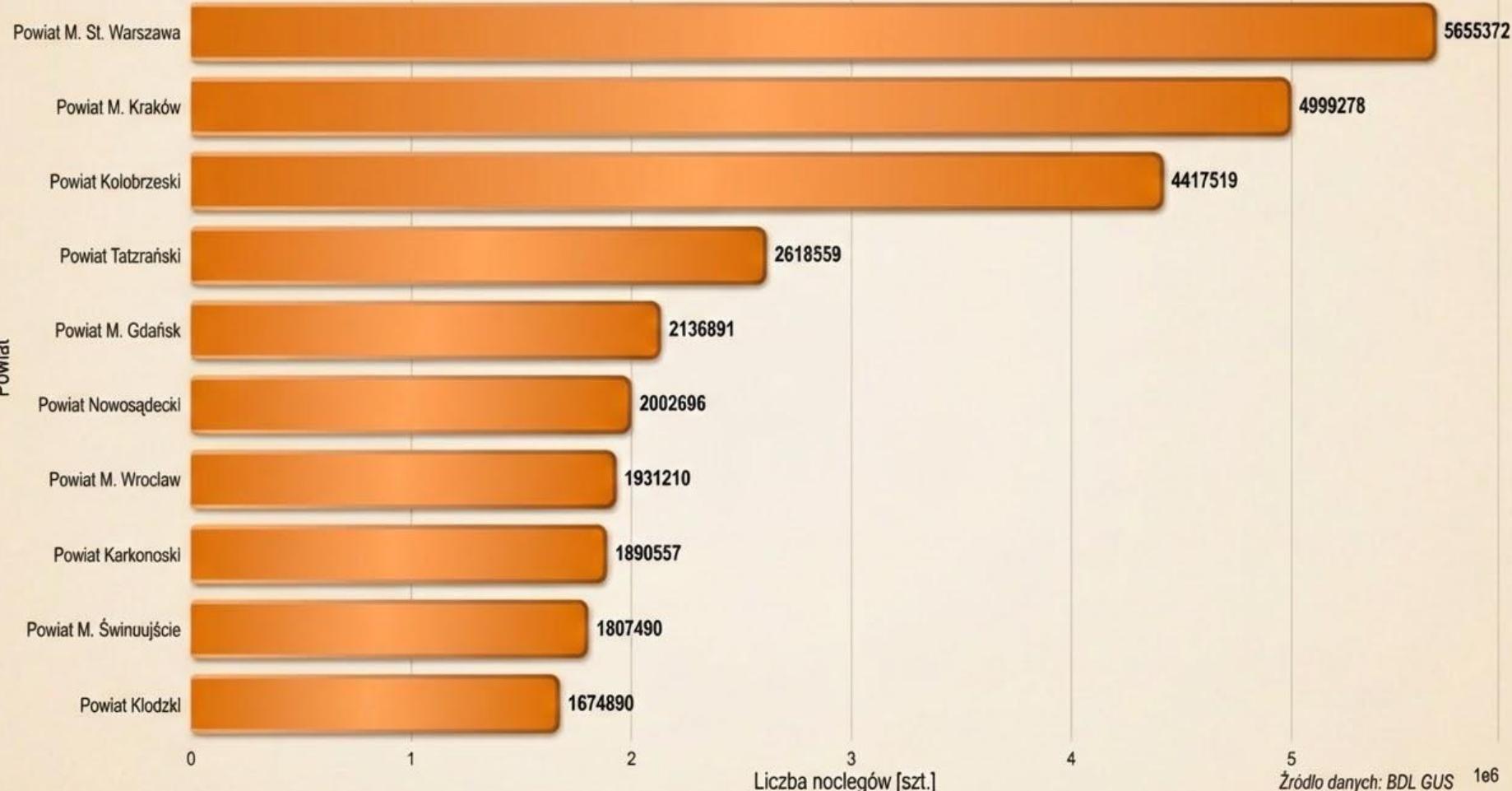
5  
1e6

# TOP 10 Powiatów: Liczba udzielonych noclegów w 2015 roku



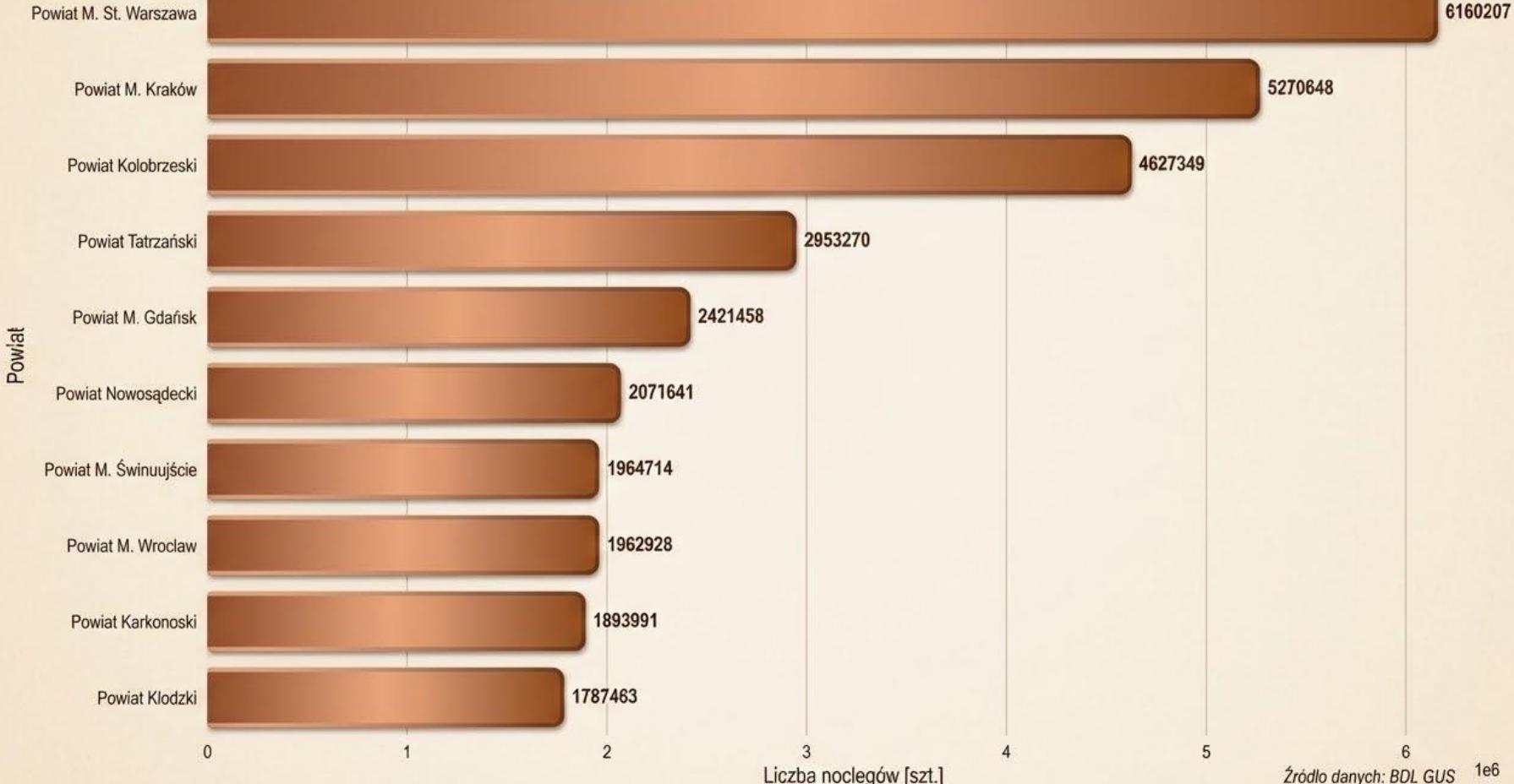
Źródło danych: BDL GUS

## TOP 10 Powiatów: Liczba udzielonych noclegów w 2016 roku

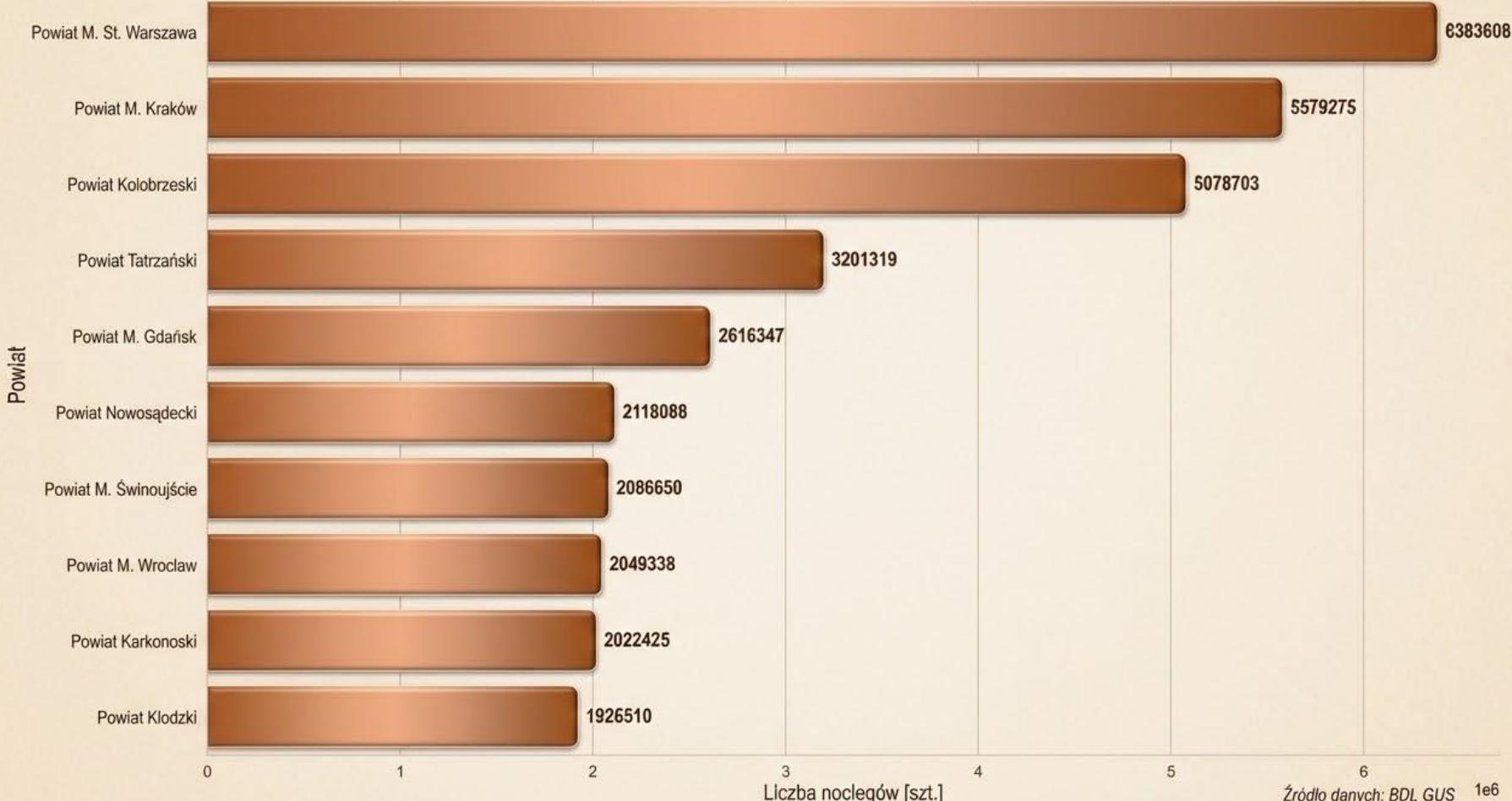


5  
Źródło danych: BDL GUS 1e6

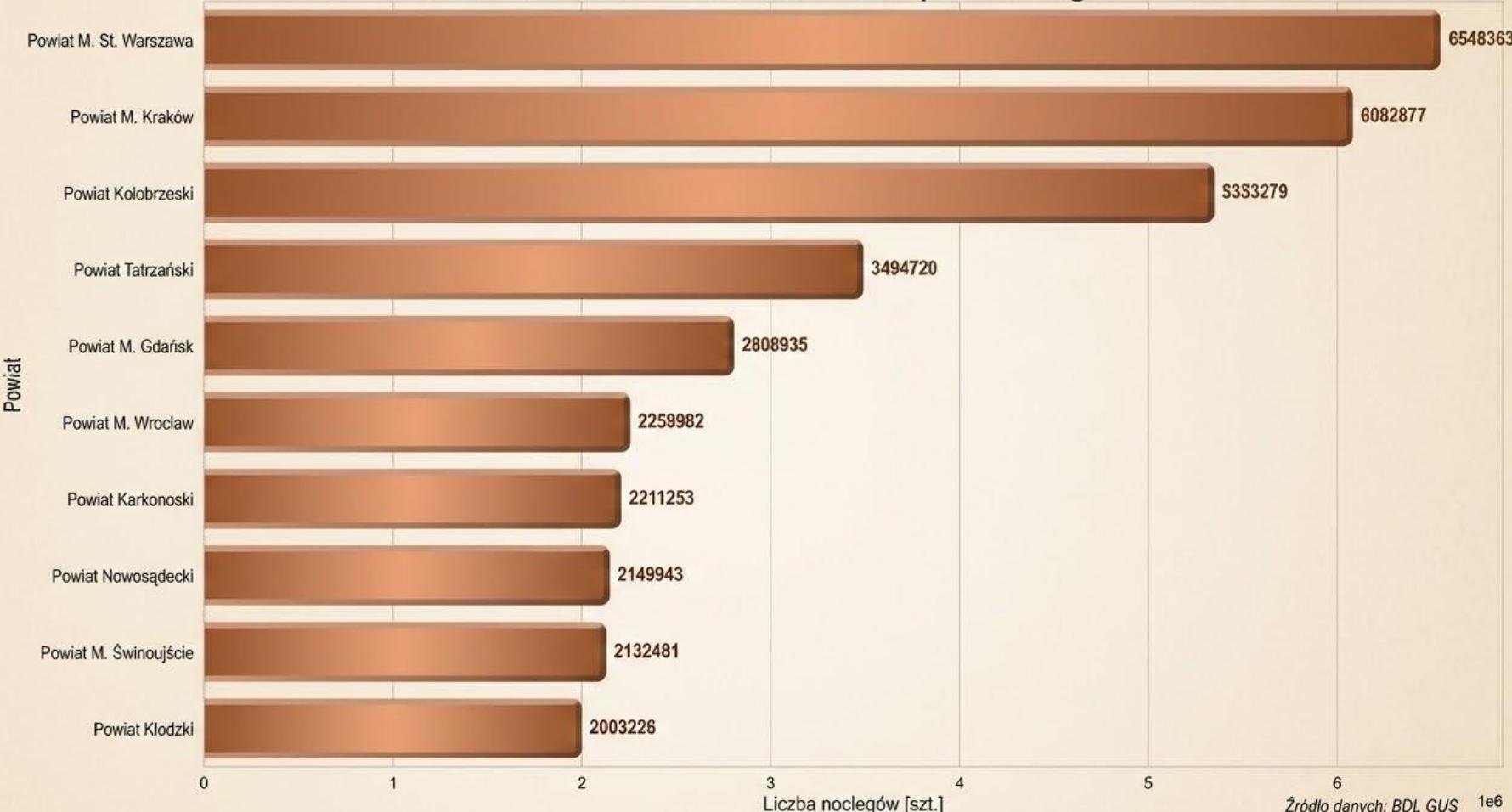
## TOP 10 Powiatów: Liczba udzielonych noclegów w 2017 roku



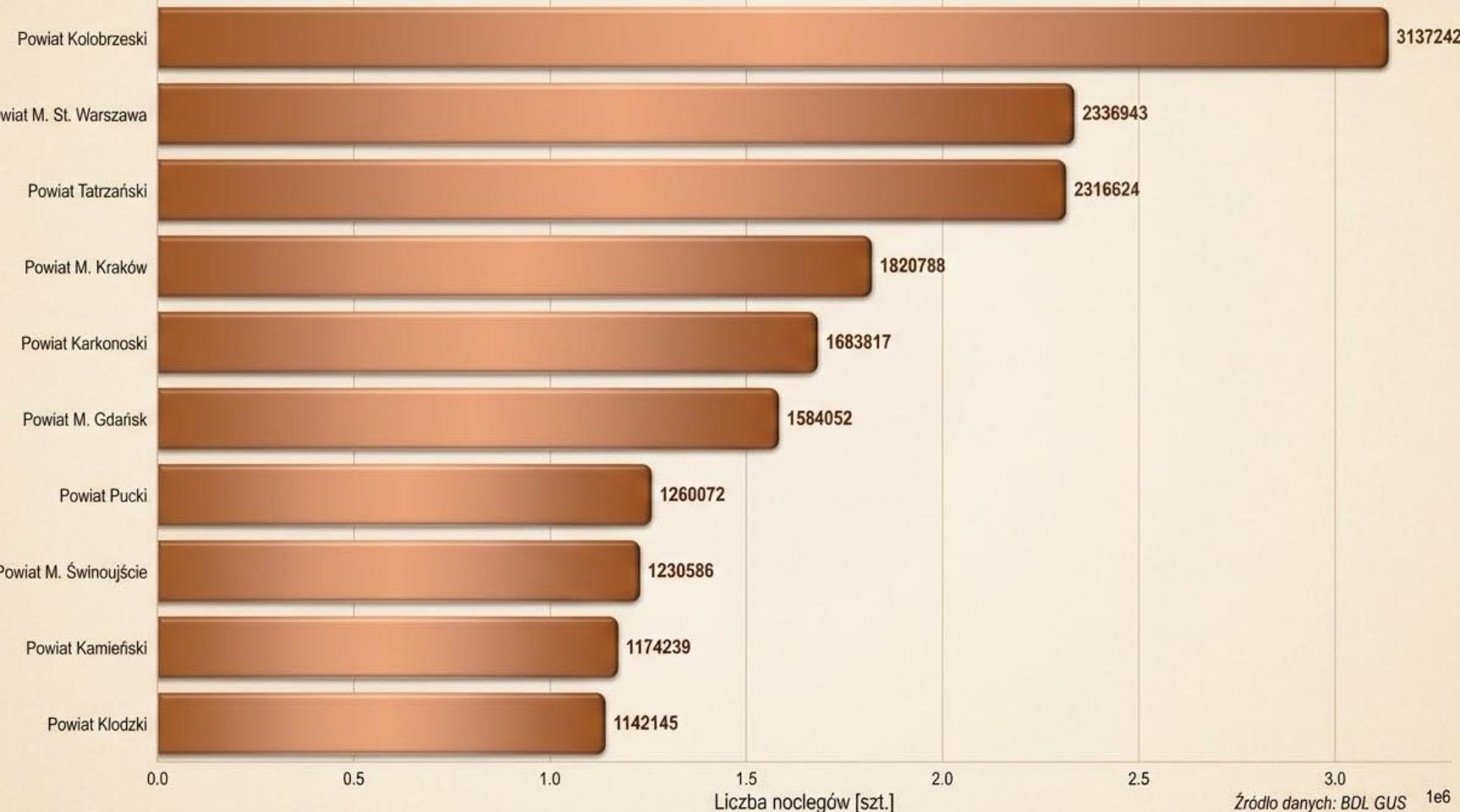
## TOP 10 Powiatów: Liczba udzielonych noclegów w 2018 roku



## TOP 10 Powiatów: Liczba udzielonych noclegów w 2019 roku



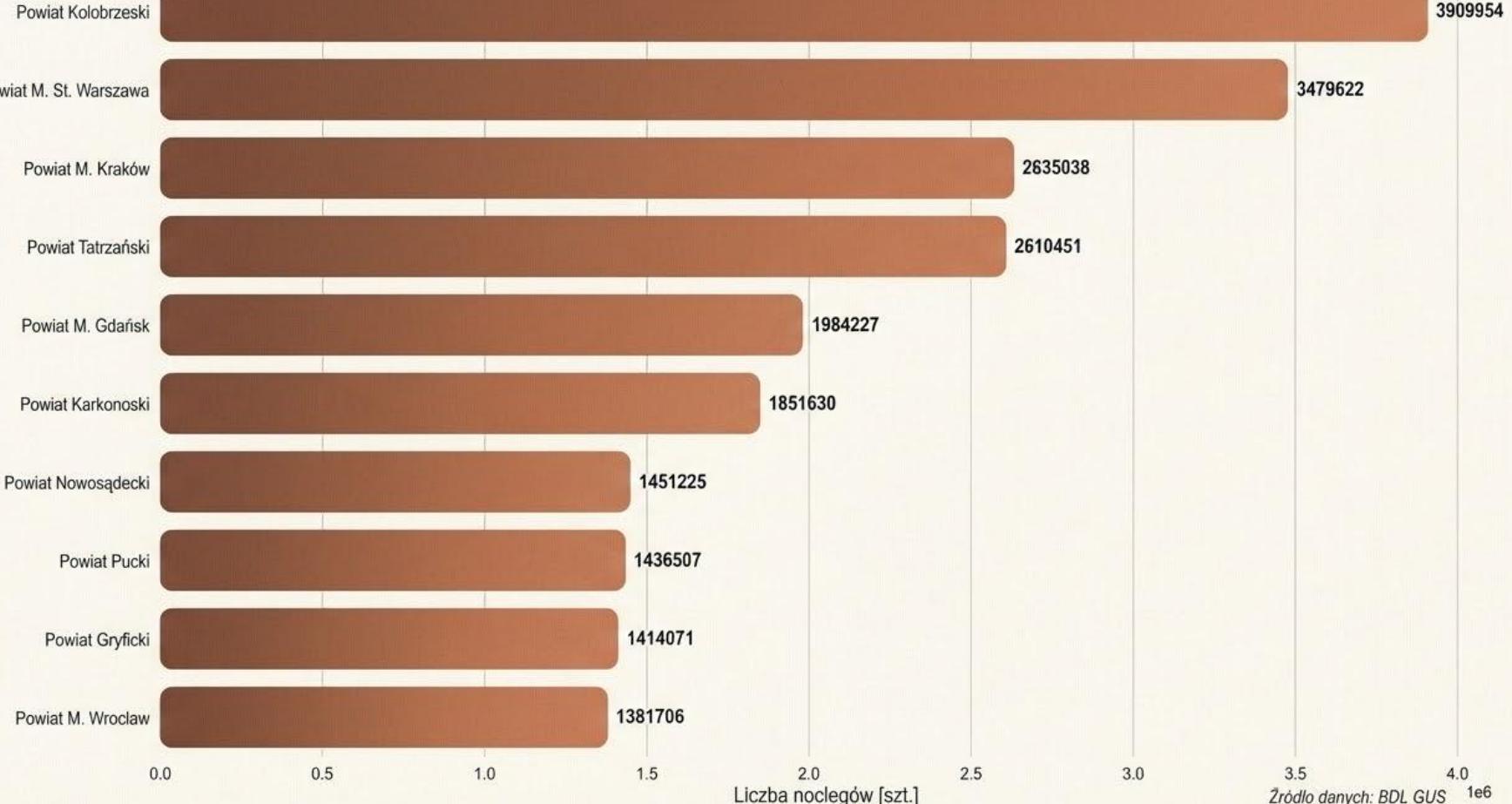
## TOP 10 Powiatów: Liczba udzielonych noclegów w 2020 roku



Liczba noclegów [szt.]

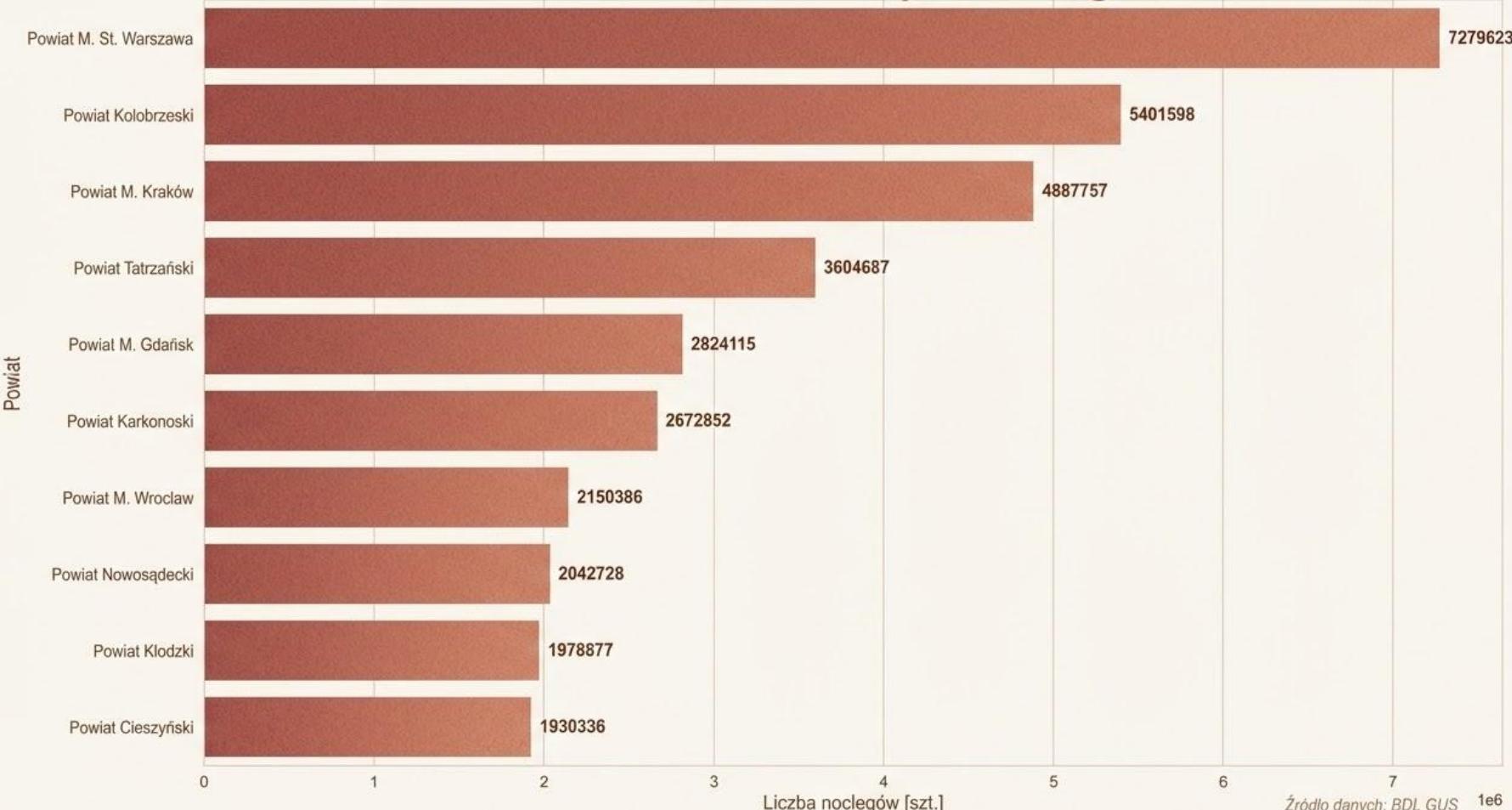
Źródło danych: BDL GUS 1e6

## TOP 10 Powiatów: Liczba udzielonych noclegów w 2021 roku



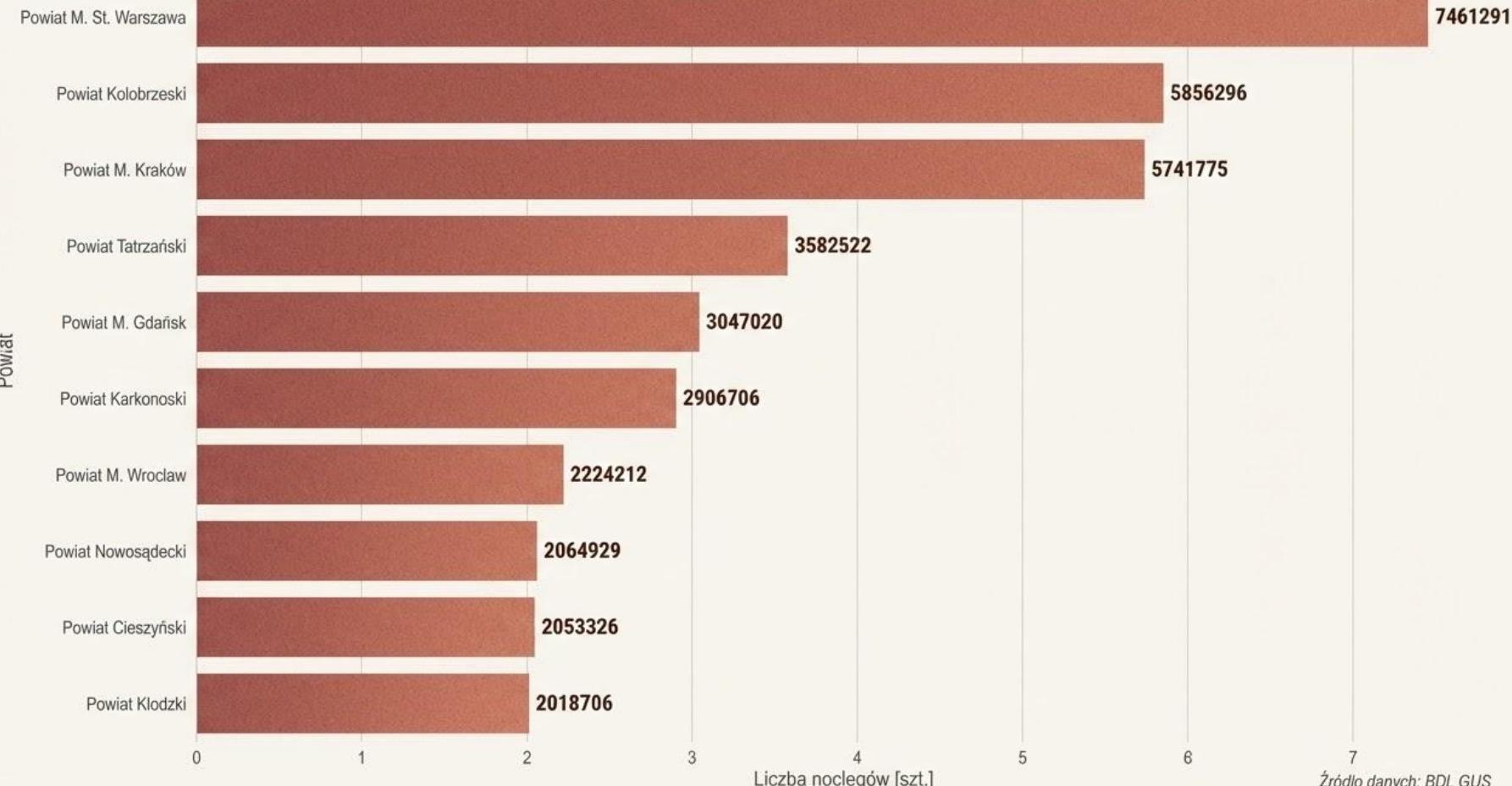
Źródło danych: BDL GUS 1e6

# TOP 10 Powiatów: Liczba udzielonych noclegów w 2022 roku



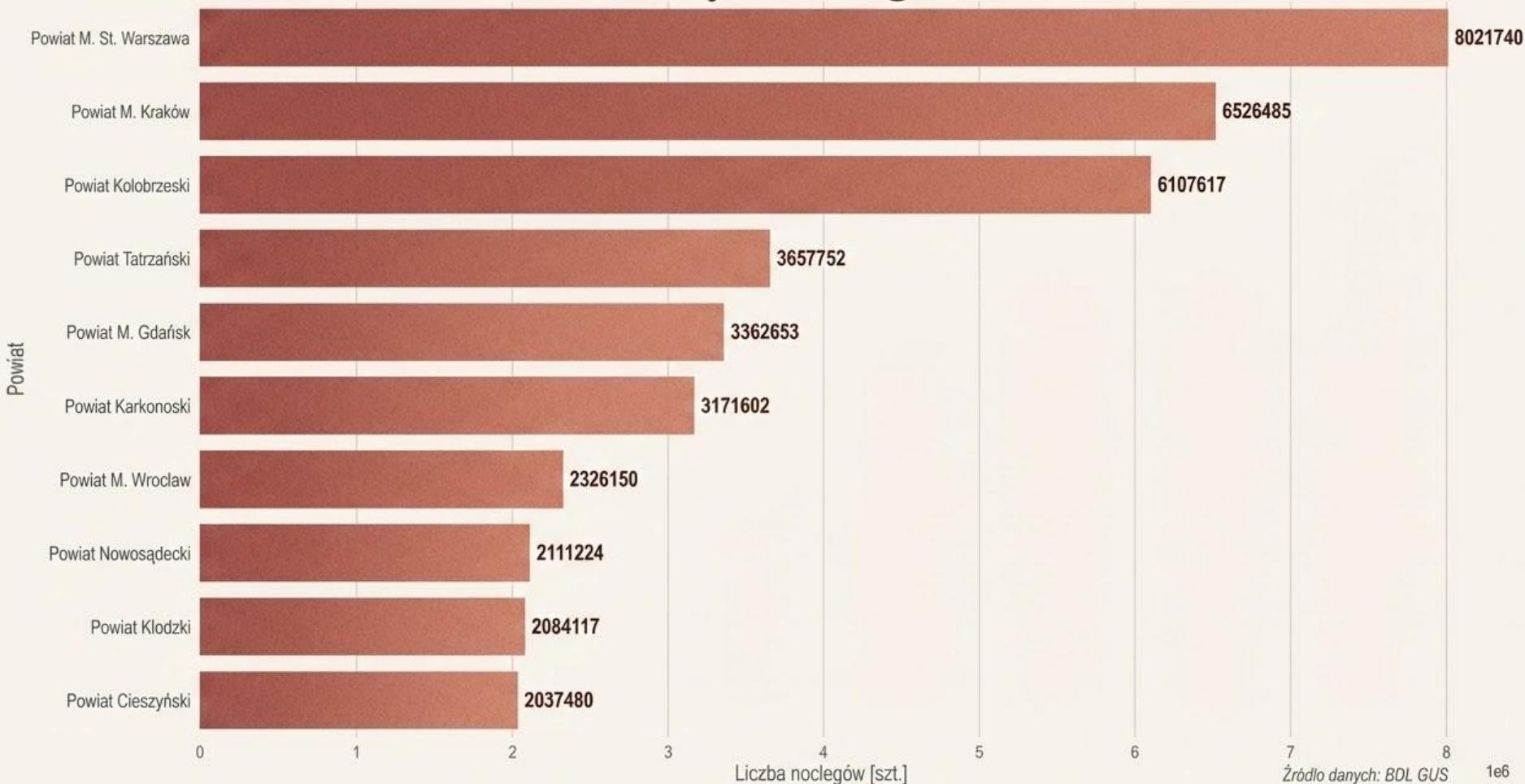
Źródło danych: BDL GUS 1e6

# TOP 10 Powiatów: Liczba udzielonych noclegów w 2023 roku



Źródło danych: BDL GUS

# TOP 10 Powiatów: Liczba udzielonych noclegów w 2024 roku



Źródło danych: BDL GUS

1e6

# Załamanie w 2020 – efekt COVID-19

## Szok i Przyczyny



2020 to silny szok w danych: spadek noclegów w całej Polsce



Przyczyny: lockdownny, ograniczenia podróży, niepewność

## Wyzwanie i Rozwiążanie



To trudny moment dla modelowania, bo:



zachowanie danych jest nienaturalne



modele mogą “nie rozumieć” takiego załamania bez dodatkowych cech



Dlatego ważne było użycie historii (lagów) i test porównawczy modeli

# Odbudowa 2021-2024



Po 2020 nastąpiło stopniowe odbicie rynku turystycznego

Powrót ruchu oznacza:



- zniesienie ograniczeń



- odłożony popyt (ludzie chcieli podróżować po przerwie)



- rozwój turystyki krajowej



2024 to poziom porównywalny  
lub wyższy niż przed pandemią



**Wniosek: turystyka w Polsce  
wróciła na ścieżkę wzrostu**

# TOP powiaty 2013 – gdzie było najwięcej noclegów?



Największy ruch koncentrował się w:



- **dużych miastach**  
(turystyka miejska i biznesowa)



- **regionach wypoczynkowych**  
(morze, góry)

Przykładowi liderzy:



Warszawa



Kraków



Kołobrzeg



Tatry



Pokazuje to, że turystyka jest nierównomiernie rozłożona:



## TOP powiaty 2024 – co się zmieniło?



- Liderzy nadal dominują, ale widać wzrost wartości w wielu powiatach



- Warszawa i Kraków zwiększają przewagę (miasta o silnej ofercie)



- Regiony wypoczynkowe pozostają bardzo mocne (sezonowe, ale stabilne)



- Wzrost znaczenia Trójmiasta (np. Gdańsk) pokazuje:
  - rosnącą popularność turystyki nadmorskiej miejskiej
  - rozwój bazy noclegowej i atrakcji



**Wniosek: ranking jest dość stabilny, ale skala rośnie**

## Dlaczego liderzy wygrywają? (interpretacja przestrzenna)



**Warszawa:**  
biznes, konferencje,  
administracja, wydarzenia  
kulturalne, lotnisko



**Kraków:**  
turystyka kulturowa,  
zabytki, rozpoznawalna  
marka miasta



**Kołobrzeg:**  
morze + uzdrowisko +  
sanatoria → długи pobyt  
i dużo noclegów

**Tatry (tatrzański):**  
góry, sporty zimowe,  
całoroczna  
atrakcyjność



**Klucz:** połączenie popytu + infrastruktury + dostępności + marki regionu

# Analiza korelacji – po co?

Korelacja pokazuje, które zmienne "idą razem" z liczbą noclegów



Korelacja nie oznacza 100% przyczynowości, ale daje mocną wskazówkę

## Najsilniejsze zależności (wyniki korelacji)



**Miejsca noclegowe**  
bardzo silna korelacja z noclegami



**Liczba turystów**  
także bardzo silnie związana z noclegami



→ logika: im większa pojemność, tym więcej noclegów da się obsłużyć

→ logika: więcej przyjazdów → więcej noclegów

→ **ma mocną zależność**  
większa oferta i konkurencja → więcej możliwości dla turystów



**Wniosek:** kluczowe są zmienne bezpośrednio związane z rynkiem turystycznym

# Słabsze zależności (i co to znaczy)



Gęstość zaludnienia ma słabszą korelację

- bo nie każdy gęsty obszar jest turystyczny
- ale miasta często mają większą bazę i usługi



Lesistość

ma prawie **zerową** korelację globalnie

- może działać lokalnie (np. turystyka przyrodnicza)
- ale w skali kraju nie jest prostym predyktorem noclegów



Wniosek: "natura" nie zawsze odbija się w danych tak wyraźnie jak infrastruktura

# Modele w projekcie – dlaczego akurat te?

- **Linear Regression:** prosta baza porównawcza, łatwa interpretacja
- **Decision Tree:** uczy się reguł nieliniowych (np. „jeśli miejsc dużo → noclegi rosną”)
- **Random Forest:** wiele drzew → stabilniejszy model, mniej przeuczenia
- **Gradient Boosting:** model sekwencyjny, poprawia błędy krok po kroku
- Dzięki porównaniu widać, czy zależności są liniowe czy bardziej złożone



Dzięki porównaniu widać, czy zależności są liniowe czy bardziej złożone

# Metryki oceny: $R^2$ , RMSE, MAE



## $R^2$ (Współczynnik determinacji)

jak dobrze model wyjaśnia zmienność danych (1 = idealnie)



## RMSE (Błąd średniokwadratowy)

kara za duże błędy (gdy model bardzo się myli w liderach)



## MAE (Średni błąd absolutny)

średni błąd absolutny – najbardziej intuicyjna metryka

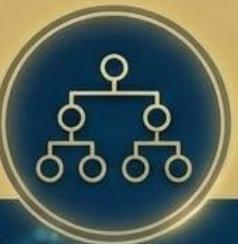
W projekcie ważne było, żeby model działał dobrze "na powiatach ogólnie", więc MAE jest bardzo użyteczne do oceny jakości.



# Wyniki porównania modeli (wniosek)



**Linear Regression:**  
może mieć bardzo  
dobre dopasowanie,  
ale większy średni  
w części przypadków



**Decision Tree:**  
często ma ryzyko  
przeuczenia  
(zbyt dopasowany  
do treningu)



**Random Forest:**  
stabilniejszy, ale  
nie zawsze najlepszy  
w średnim błędzie



**Gradient Boosting:**  
(najniższy MAE)



Najlepszy kompromis w projekcie dał **Gradient Boosting** (najniższy MAE).  
Modele osiągnęły wysokie **R<sup>2</sup>**, czyli ogólnie dobrze przewidują noclegi.

# Walidacja prognozy 2024 (predykcja vs rzeczywistość)

Predykcja

Rzeczywistość

$y=x$  (Linia idealna)

Liderzy  
(ogromne  
wartości)



Wykres ' $y = x$ ' pokazuje, że większość punktów jest blisko linii idealnej



To znaczy, że model trafia w skalę wartości dla wielu powiatów



Największe odchylenia są zwykle tam, gdzie wartości są ogromne (liderzy)

Wniosek: model jest praktyczny do prognozowania na poziomie powiatów



# SKĄD BIORĄ SIĘ BŁĘDY PROGNOZY?

- Powody, dla których model może się mylić:



## Brak Sezonowości w Danych Rocznych:

(miesiące/kwartały dałyby więcej)



**lokalne wydarzenia:** duże imprezy, inwestycje, remonty, nowe połączenia

- **zmiany po pandemii:** inny styl podróżowania, wzrost turystyki krajowej



**część ruchu** może być niewidoczna (np. prywatne najmy, jednodniowe wyjazdy)

- To normalne w danych społeczno-gospodarczych



## Lokalne Wydarzenia i Inwestycje:

dane roczne nie pokazują sezonowości



## Niewidoczny Ruch (Szara Strefa):

część ruchu może być niewidoczna (np. prywatne najmy, jednodniowe wyjazdy)

To normalne w danych społeczno-gospodarczych

# Podsumowanie i wnioski końcowe



Tak – da się przewidywać ruch turystyczny w powiatach na podstawie danych BDL GUS

## Najśilniejsze czynniki



liczba turystów + baza noclegowa  
(miejscia/obiekty)

## Trend 2013–2024



wzrost + silny spadek w 2020  
+ odbudowa

## Najlepszy model



Gradient Boosting Regressor  
(dobry kompromis jakości)

## Zastosowania:



planowanie infrastruktury  
i promocji regionu



lepsze zarządzanie sezonem  
i obciążeniem usług



wsparcie decyzji inwestycyjnych  
i strategii lokalnych

Dziękujemy za uwagę!