Dokumentacja Modelu Uczenia Maszynowego

**Autor:** Szymon Szczurowski

**Data:** 2023-12-10

# 1. Wstęp

**Cel Projektu:**  
Celem tego projektu jest stworzenie modelu uczenia maszynowego, którego głównym zadaniem będzie predykcja cen lokali mieszkaniowych na rynku nieruchomości na terenie miasta Morąg. Model ten będzie wykorzystywał różnorodne dane dotyczące nieruchomości, takie jak data sprzedaży, cena, liczba izb, kondygnacja, powierzchnia użytkowa, powierzchni użytkowa pomieszczeń przynależnych, aby dokładnie przewidzieć cenę sprzedaży mieszkania. Zastosowanie takiego modelu ma na celu wspieranie decyzji kupujących, sprzedających oraz inwestorów nieruchomości, dostarczając im wartościowych wskazówek i prognoz cenowych..

**Zakres:**  
Projekt obejmuje kilka kluczowych etapów: pozyskiwanie i przetwarzanie danych dotyczących nieruchomości, analizę eksploracyjną tych danych, wybór i trenowanie odpowiedniego modelu uczenia maszynowego oraz ocenę jego skuteczności. Model będzie skoncentrowany na przewidywaniu cen, biorąc pod uwagę specyfikacje poszczególnych lokali mieszkaniowych. Ostatecznym celem jest stworzenie narzędzia, które będzie mogło być wykorzystane do generowania wiarygodnych i precyzyjnych predykcji cenowych w dynamicznie zmieniającym się środowisku rynku nieruchomości.

# 2. Pozyskiwanie Danych

**Źródła Danych:**  
Dane, wykorzystane w tym projekcie, zostały pozyskane z Urzędu Gminy Morąg. Zbiór danych zawiera szeroki zakres informacji dotyczących nieruchomości mieszkalnych w tej lokalizacji, w tym szczegółowe dane o charakterystykach każdej z nieruchomości. Dane te zostały zebrane przez urząd w ramach normalnej działalności administracyjnej i są regularnie aktualizowane, co zapewnia ich aktualność i wiarygodność. Wykorzystanie danych z urzędu gwarantuje, że model będzie opierał się na rzetelnych i reprezentatywnych informacjach, co jest kluczowe dla dokładności predykcji cen.

**Proces Zbierania Danych:**  
Proces zbierania danych w Urzędzie Gminy Morąg jest ściśle związany z bieżącą działalnością urzędniczą. Dane o nieruchomościach mieszkalnych są gromadzone systematycznie w ramach różnych procedur urzędowych, takich jak transakcje kupna-sprzedaży. Każda z tych interakcji generuje dane, które są następnie wprowadzane do systemu informatycznego urzędu.

**Wstępne Czyszczenie Danych:**  
Podczas wstępnej analizy zauważono, że zbiór danych zawierał liczne wartości nullowe oraz cechy, które nie wnosiły istotnych informacji do modelu. W pierwszym etapie wstępnego czyszczenia danych, podjęto decyzję o usunięciu tych nieistotnych atrybutów, co pozwoliło na skupienie analizy na najbardziej relewantnych cechach. Kolejnym krokiem było dokładne przeglądanie całego zbioru w celu identyfikacji i usunięcia wierszy z niekompletnymi danymi, które mogłyby negatywnie wpłynąć na jakość modelu.

W przypadku wartości nullowych zastosowano metodę imputacji opartą na percentylach, co pozwoliło na zastąpienie brakujących danych wartościami, które najlepiej odzwierciedlają rozkład danego atrybutu w całej populacji. Taki sposób postępowania zwiększa wiarygodność danych bez wprowadzania znaczących zniekształceń.

Ostatnim etapem wstępnego czyszczenia danych było ustawienie poprawnych typów danych dla każdej z cech. Zapewnienie, że każdy atrybut ma odpowiedni typ danych, jest niezbędne do prawidłowego działania algorytmów uczenia maszynowego i pozwala uniknąć błędów wynikających z niezgodności typów. Przykładowo, cechy kategoryczne zostały przekształcone w odpowiednie formaty, a dane liczbowe zoptymalizowano pod kątem ich skali i formatu.

Te staranne działania wstępnego czyszczenia danych zapewniają solidną podstawę do dalszych etapów analizy i modelowania, a także znacząco zwiększają szanse na uzyskanie wiarygodnych wyników predykcyjnych.

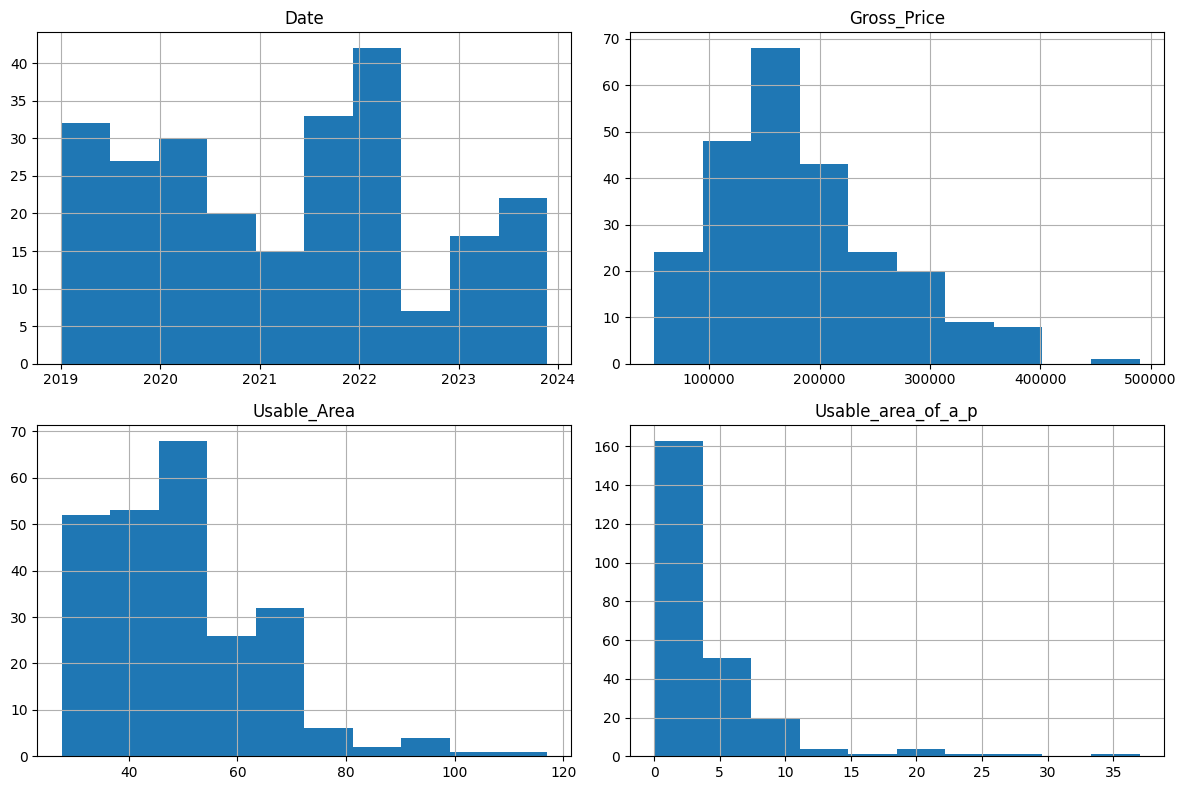
# 3. Analiza Eksploracyjna Danych (EDA)

**Statystyki Opisowe:**

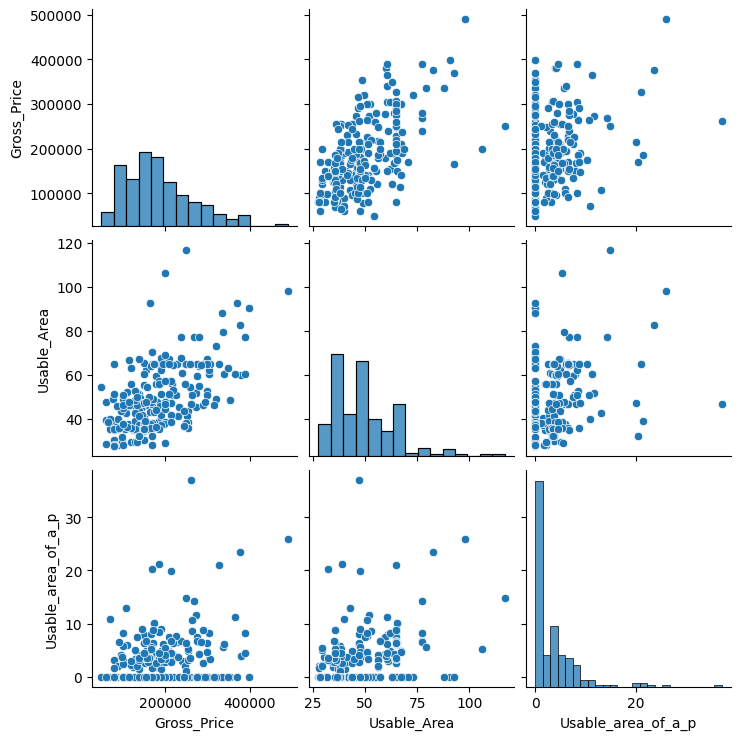
* Średnia cena brutto nieruchomości wynosi około 183,722.
* Średnia powierzchnia użytkowa to około 49 m².
* Średnia powierzchnia pomieszczeń przynależnych to około 3.2 m².
* Standardowe odchylenie cen brutto to około 78,995, powierzchni użytkowej to około 14.7 m², a powierzchni pomieszczeń przynależnych to około 4.91 m².
* Ceny brutto mieszczą się w zakresie od 50,000 do 490,000.
* Maksymalna powierzchnia użytkowa to 117 m², a pomieszczeń przynależnych to 37 m².
* Dane obejmują nieruchomości na rynku pierwotnym i wtórnym, z przewagą rynku wtórnego.
* Najczęstsza liczba pokoi w nieruchomościach to 2 lub 3, co może odzwierciedlać preferencje nabywców lub dostępność na rynku.
* Nieruchomości są zazwyczaj umiejscowione na niższych piętrach, co może sugerować preferencje dla łatwiejszego dostępu lub niższych cen.

**Wizualizacja Danych:**

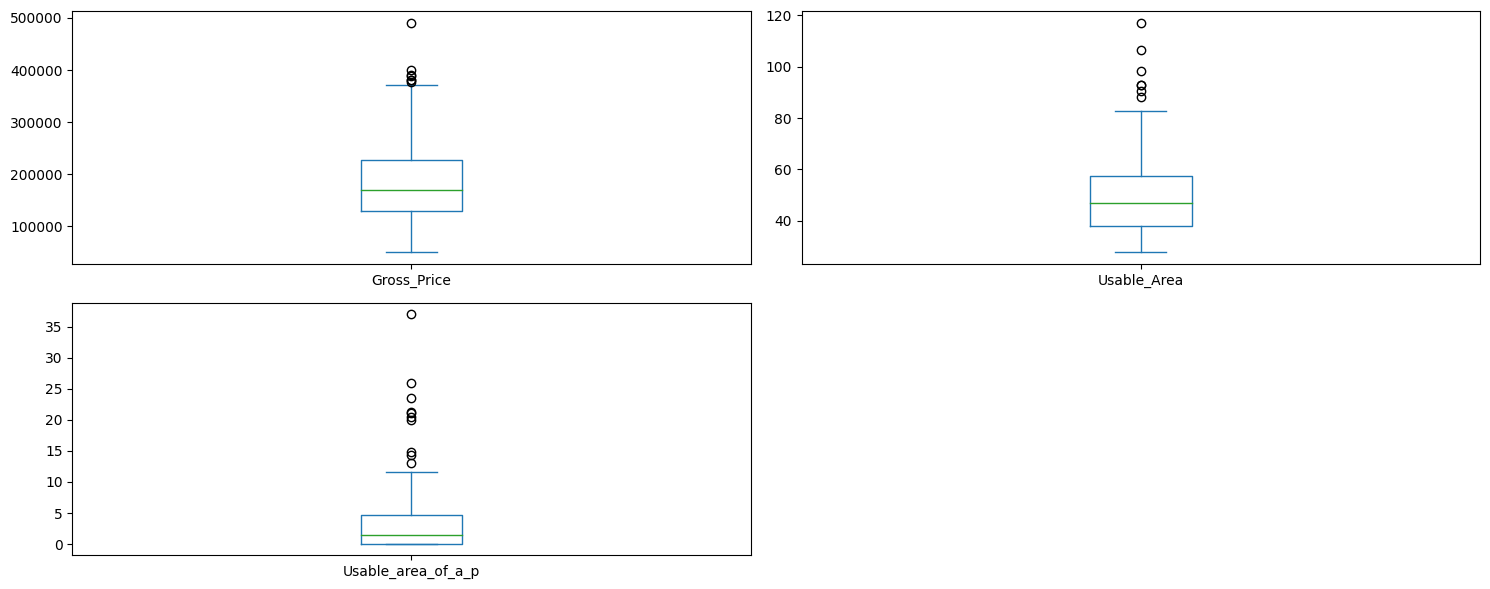
1. Histograms



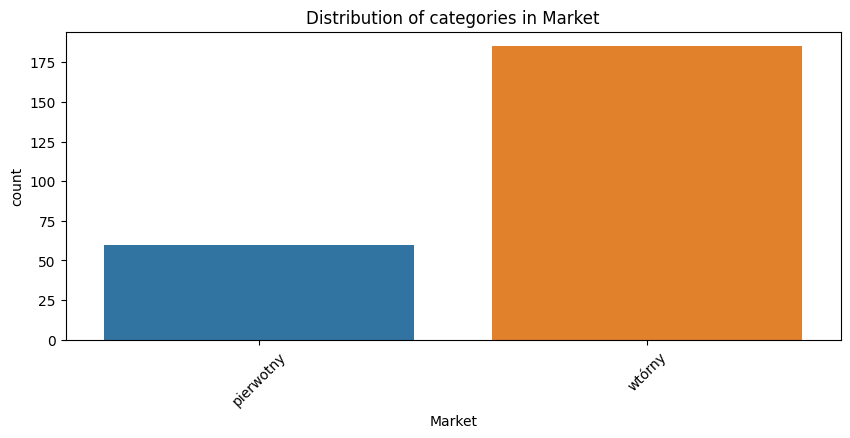
1. Par charts

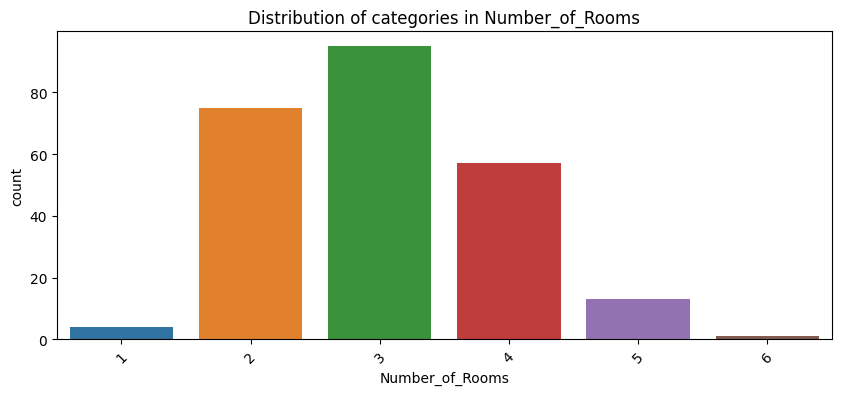


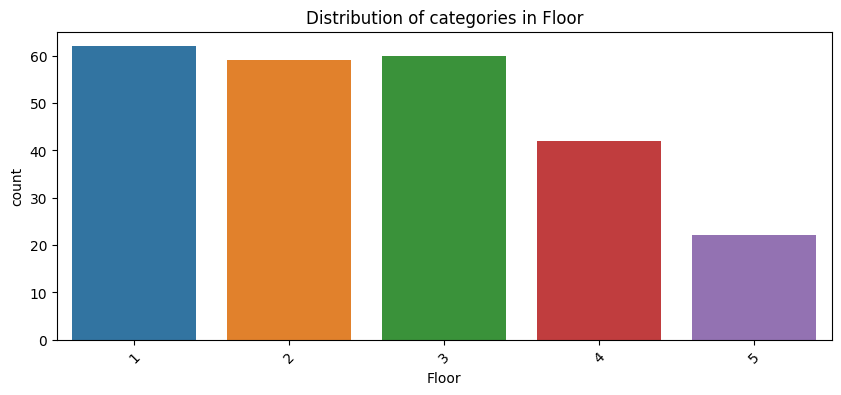
1. Box charts



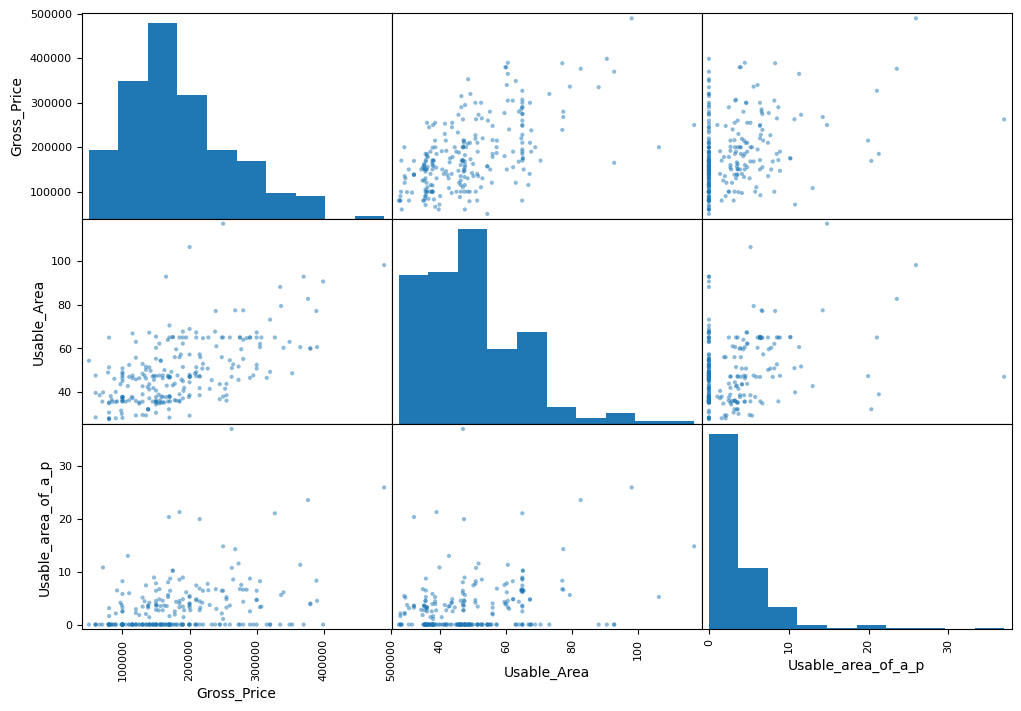
1. Bar charts for Categorical Variables



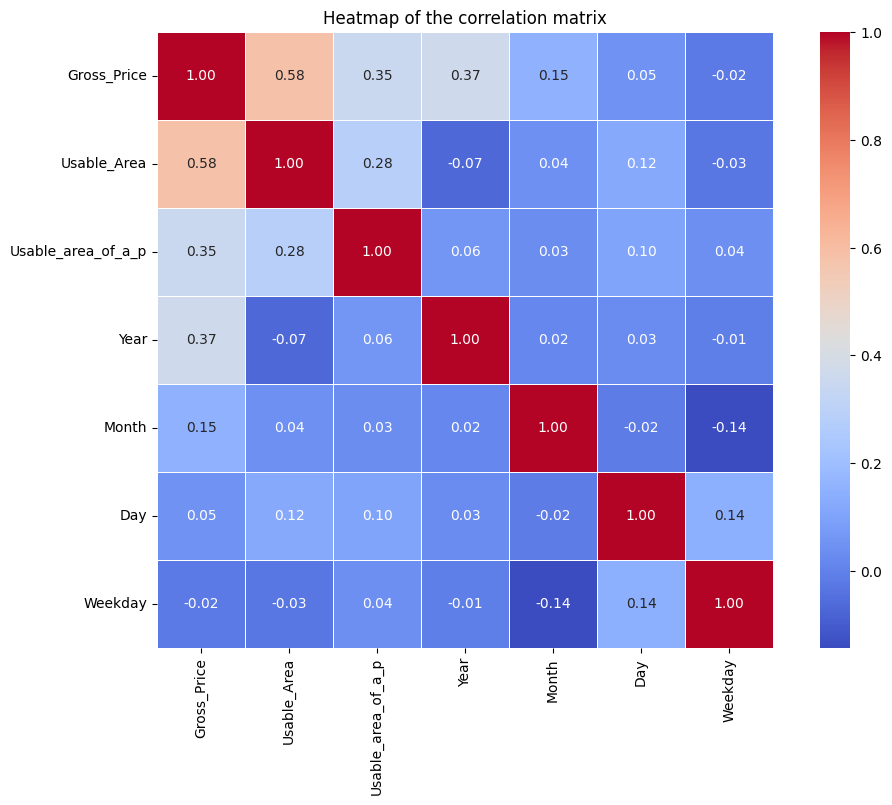




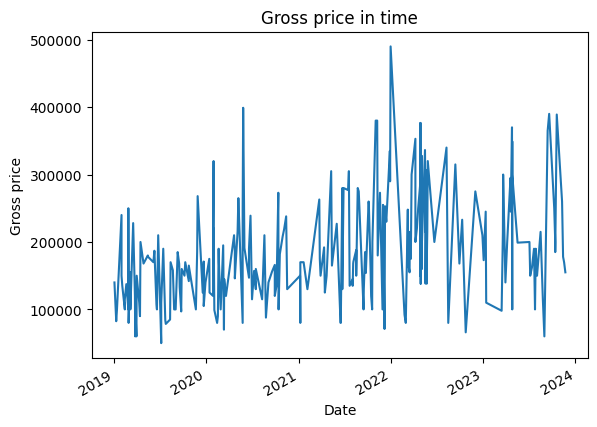
1. Scatterplot matrix



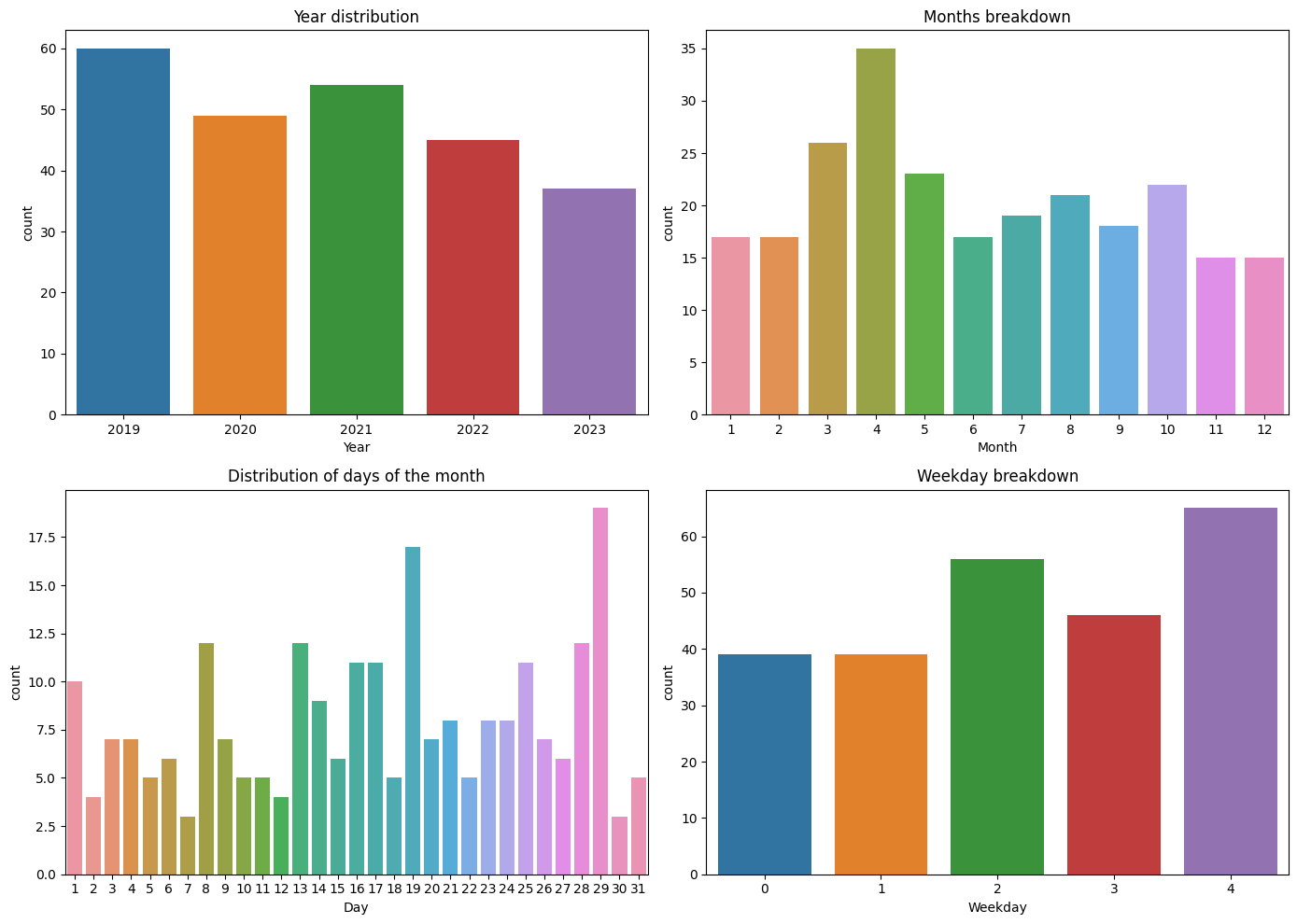
1. Heatmap of correlations



1. Time trends charts



1. Timetables



**Wnioski z EDA:**

* Umiarkowana korelacja między ceną a powierzchnią użytkową wskazuje, że większe mieszkania mają tendencję do bycia droższymi.
* Istnieje sezonowość w aktywności rynkowej, z wiosną i latem jako szczytowymi okresami.
* Wydarzenia globalne mogą mieć wpływ na ceny nieruchomości, co widać po wzrostach cen w okresach takich jak początek pandemii COVID-19 i wojna na Ukrainie.
* Dane pokazują preferencję dla nieruchomości z mniejszą liczbą pokoi, co może odzwierciedlać demograficzne lub ekonomiczne tendencje na rynku..
* Dane wskazują na to, że rynek wtórny jest bardziej aktywny niż pierwotny, co może mieć wpływ na strategie marketingowe, cenę i dostępność nieruchomości.
* Preferencje dotyczące liczby pokoi i piętra, a także rozkład transakcji w ciągu roku, miesiąca i tygodnia, mogą odzwierciedlać wzorce zachowań nabywców oraz wpływ czynników takich jak praktyczność, dostępność usług, czy sezonowość.
* Analiza wartości kategorialnych jest równie ważna jak analiza zmiennych ilościowych, ponieważ może pomóc zrozumieć preferencje konsumentów oraz zmiany na rynku nieruchomości.

# 3. Przetwarzanie i Przygotowanie Danych

Metody Przetwarzania:  
Opis technik przetwarzania danych.

Podział Danych:  
Opis, jak podzielono dane.

# 5. Wybór Modelu

Kandydaci do Modelu:  
Lista rozważanych modeli uczenia maszynowego.

Kryteria Wyboru:  
Opis, na podstawie jakich kryteriów wybrano model.

# 6. Trenowanie Modelu

Parametry Modelu:  
Szczegóły konfiguracji modelu.

Proces Trenowania:  
Opis procesu trenowania.

# 7. Ocena Modelu

Metryki Oceny:  
Wykorzystane metryki.

Wyniki:  
Prezentacja wyników modelu.

# 8. Wdrażanie

Środowisko Wdrożenia:  
Opis środowiska wdrożenia.

Proces Wdrożenia:  
Kroki potrzebne do wdrożenia modelu.

# 9. Wnioski i Dalsze Kierunki

Podsumowanie:  
Końcowe wnioski dotyczące projektu.

Propozycje Dalszego Rozwoju:  
Sugestie dotyczące poprawek, rozszerzeń lub nowych kierunków badań.

# Załączniki

Kody Źródłowe:  
Linki do repozytoriów z kodem.

Dokumentacja API:  
Jeśli model jest udostępniony jako usługa API.

Dodatkowe Materiały:  
Inne materiały, takie jak prezentacje, pliki danych itp.