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.

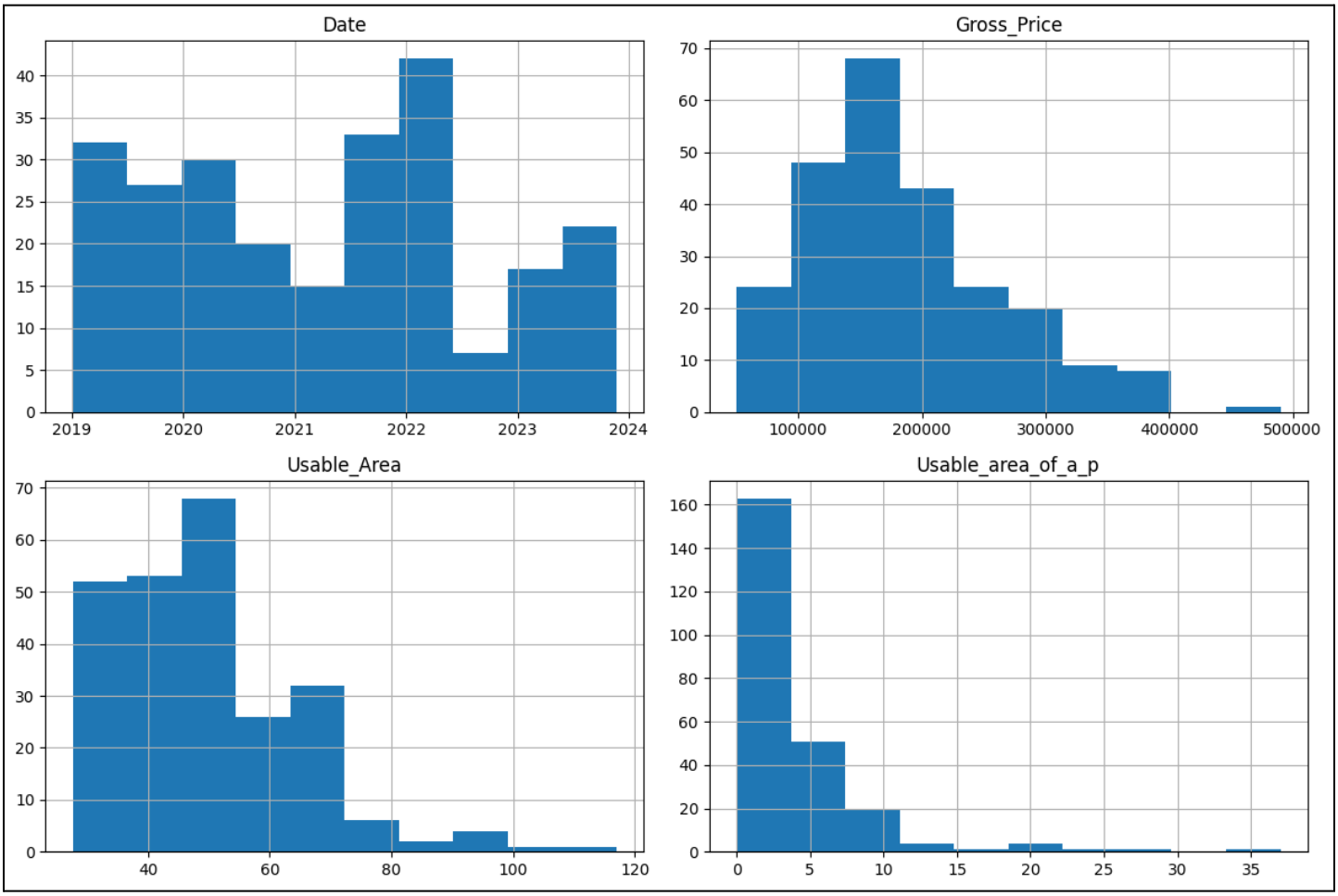
3. Analiza Eksploracyjna Danych (EDA)

Statystyki Opisowe:

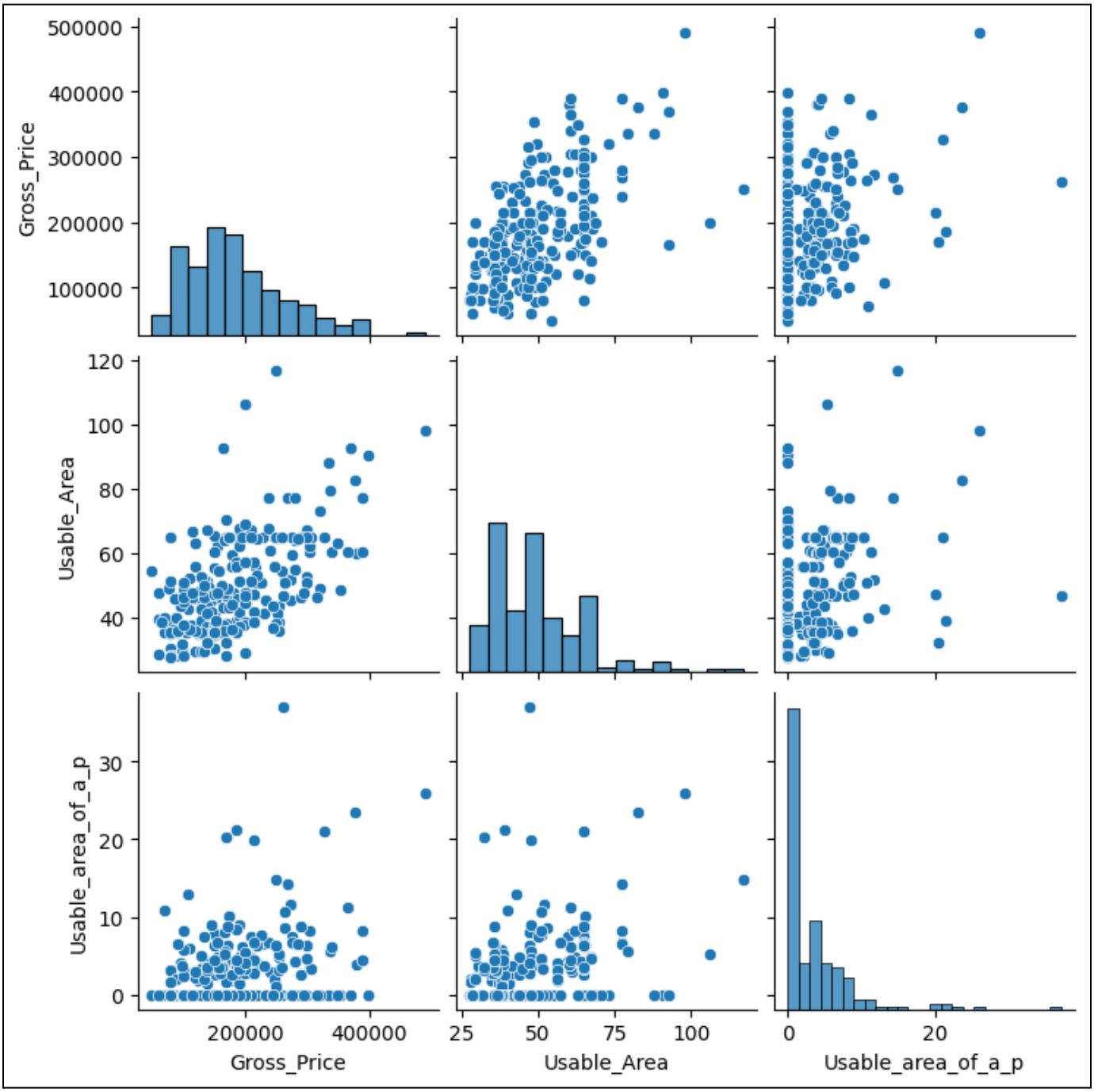
- Średnia cena brutto nieruchomości wynosi około 183,722.
- Średnia powierzchnia użytkowa to około 49 m^2.
- Średnia powierzchnia pomieszczeń przynależnych to około 3.2 m^2.
- Standardowe odchylenie cen brutto to około 78,995, powierzchni użytkowej to około 14.7 m^2, a powierzchni pomieszczeń przynależnych to około 4.91 m^2.
- Ceny brutto mieszczą się w zakresie od 50,000 do 490,000.
- Maksymalna powierzchnia użytkowa to 117 m^2, a pomieszczeń przynależnych to 37 m^2.
- 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:

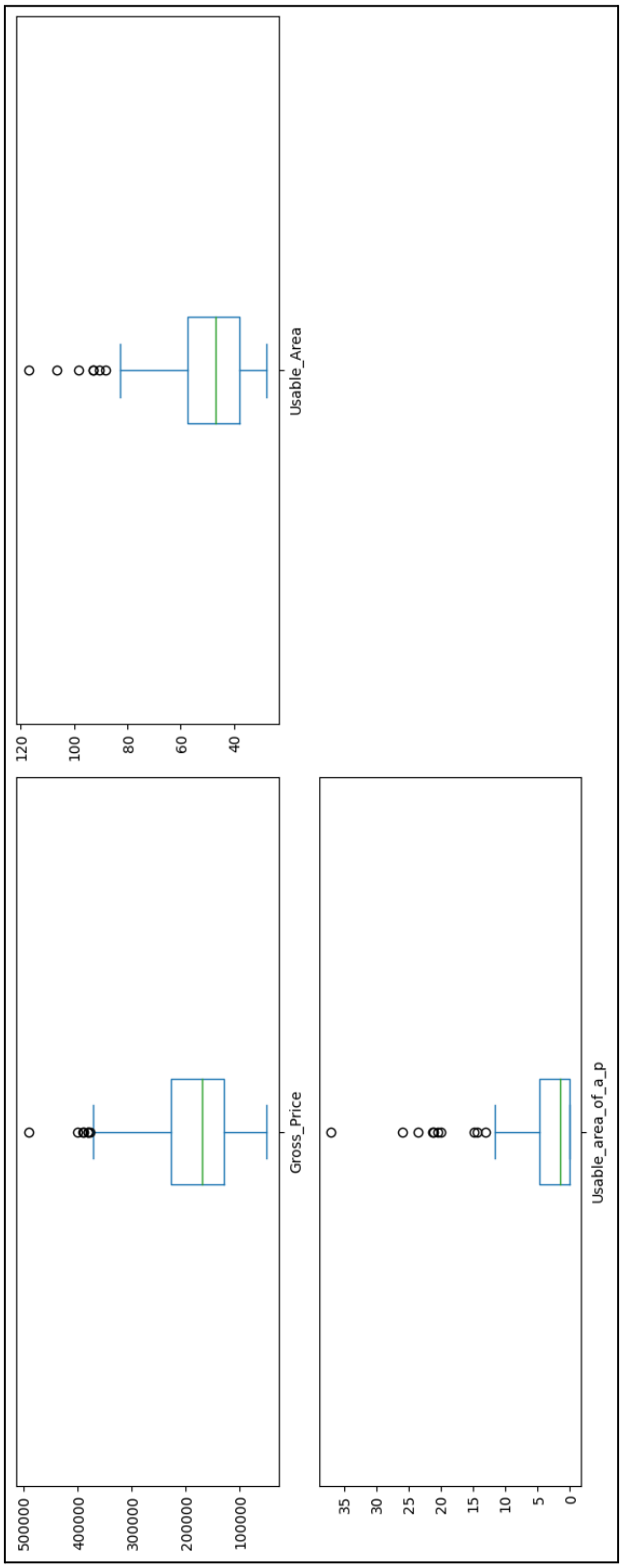
1. Histograms



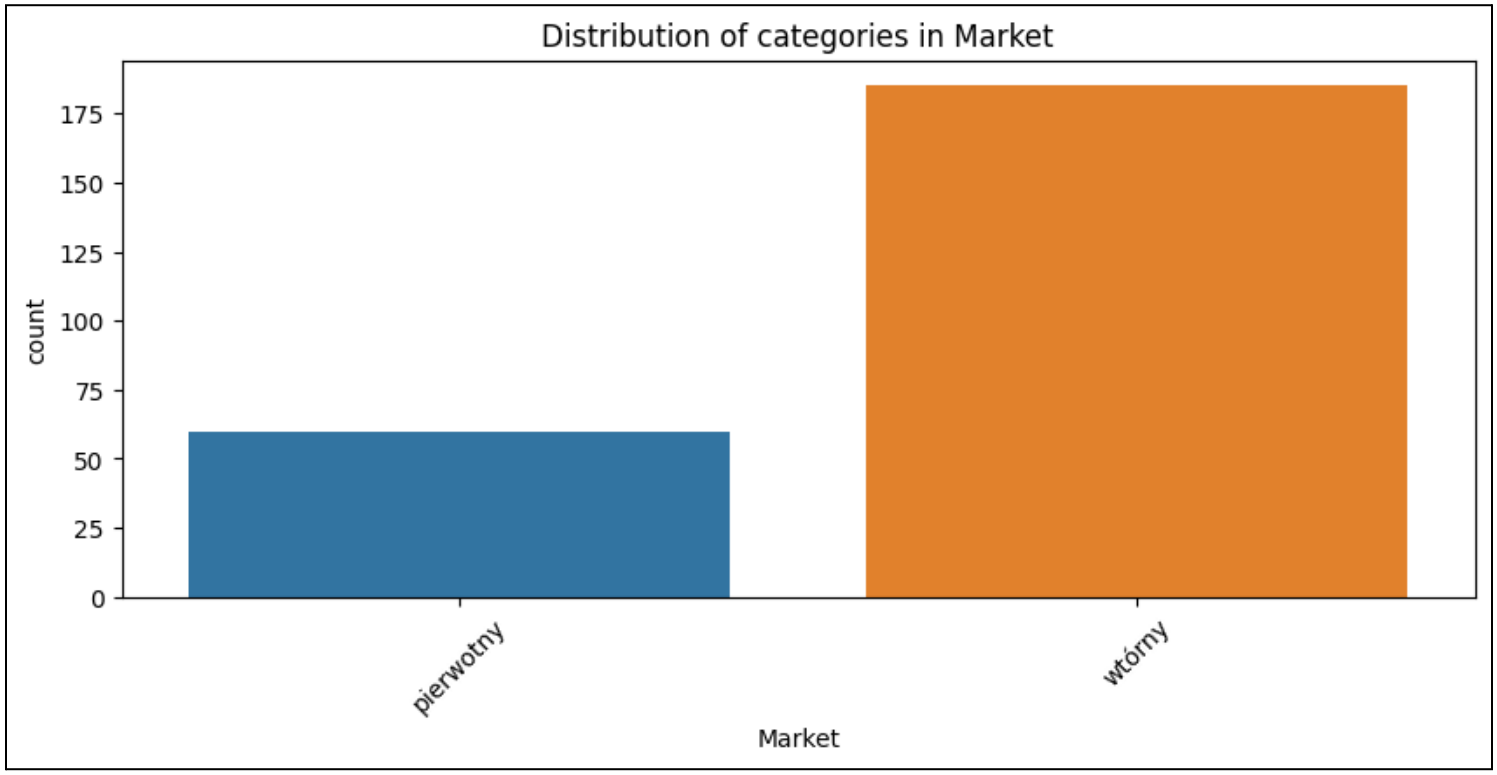
2. Par charts

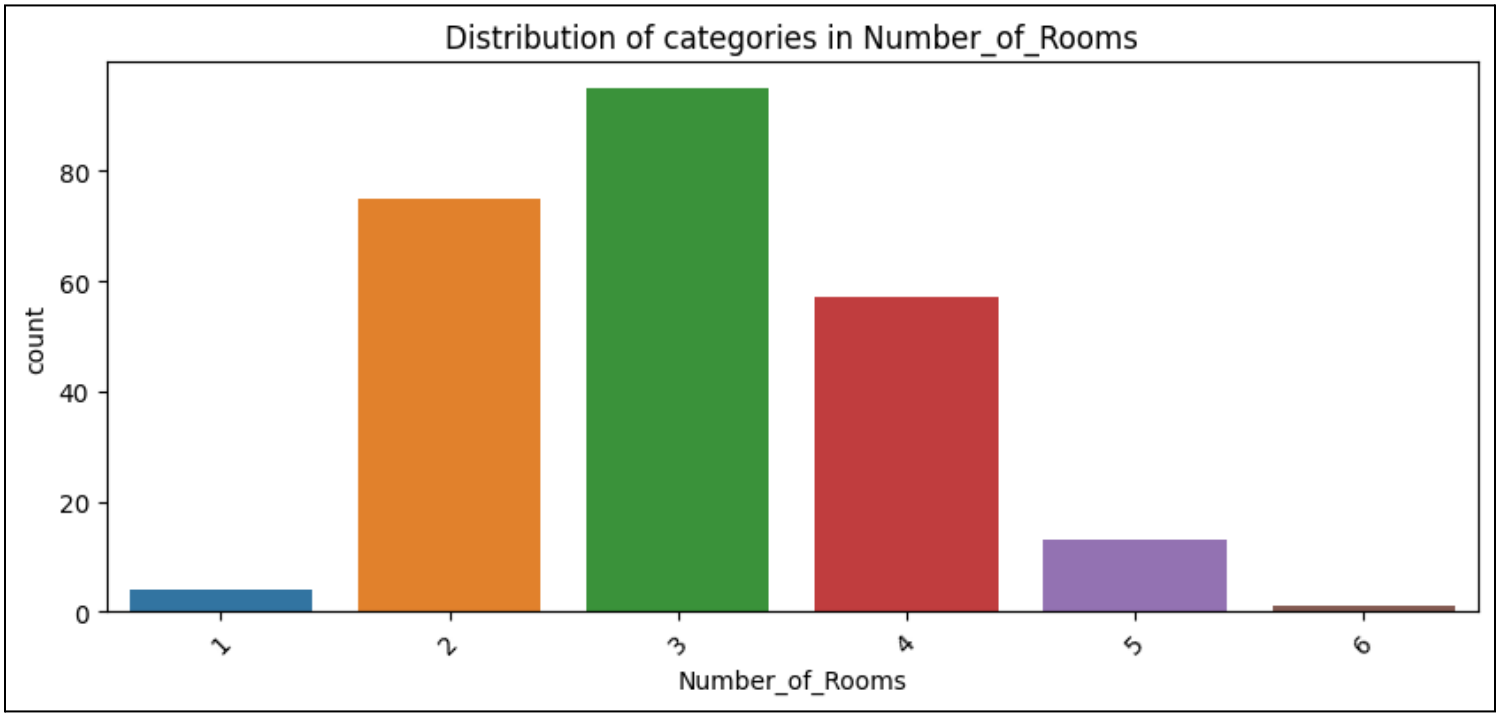

Gross_Price
Usable_Area
Usable_area_of_a_p
500000
400000
300000
200000
100000
120
100
80
60
40
30
20
10
0
200000
400000
25
50
75
100
0
20
Gross_Price
Usable_Area
Usable_area_of_a_p

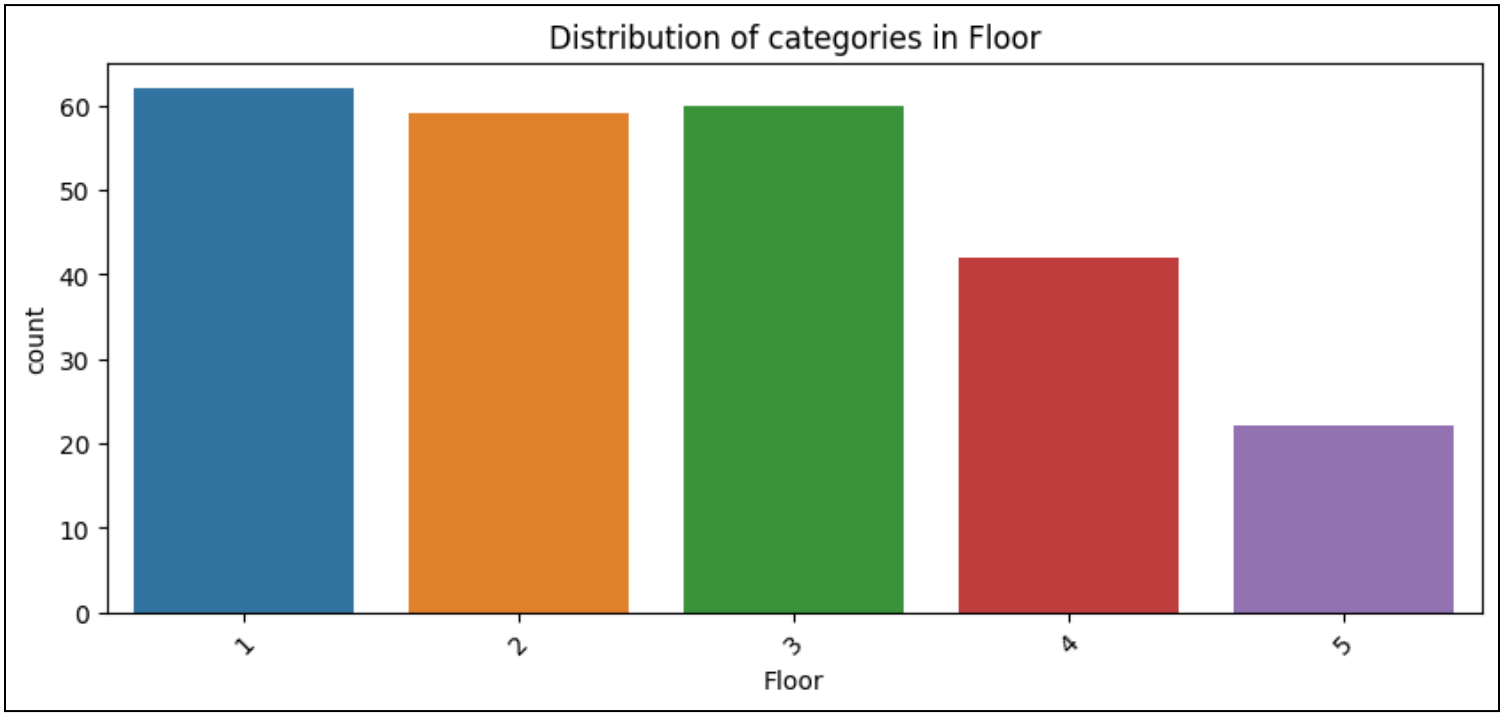
3. Box charts



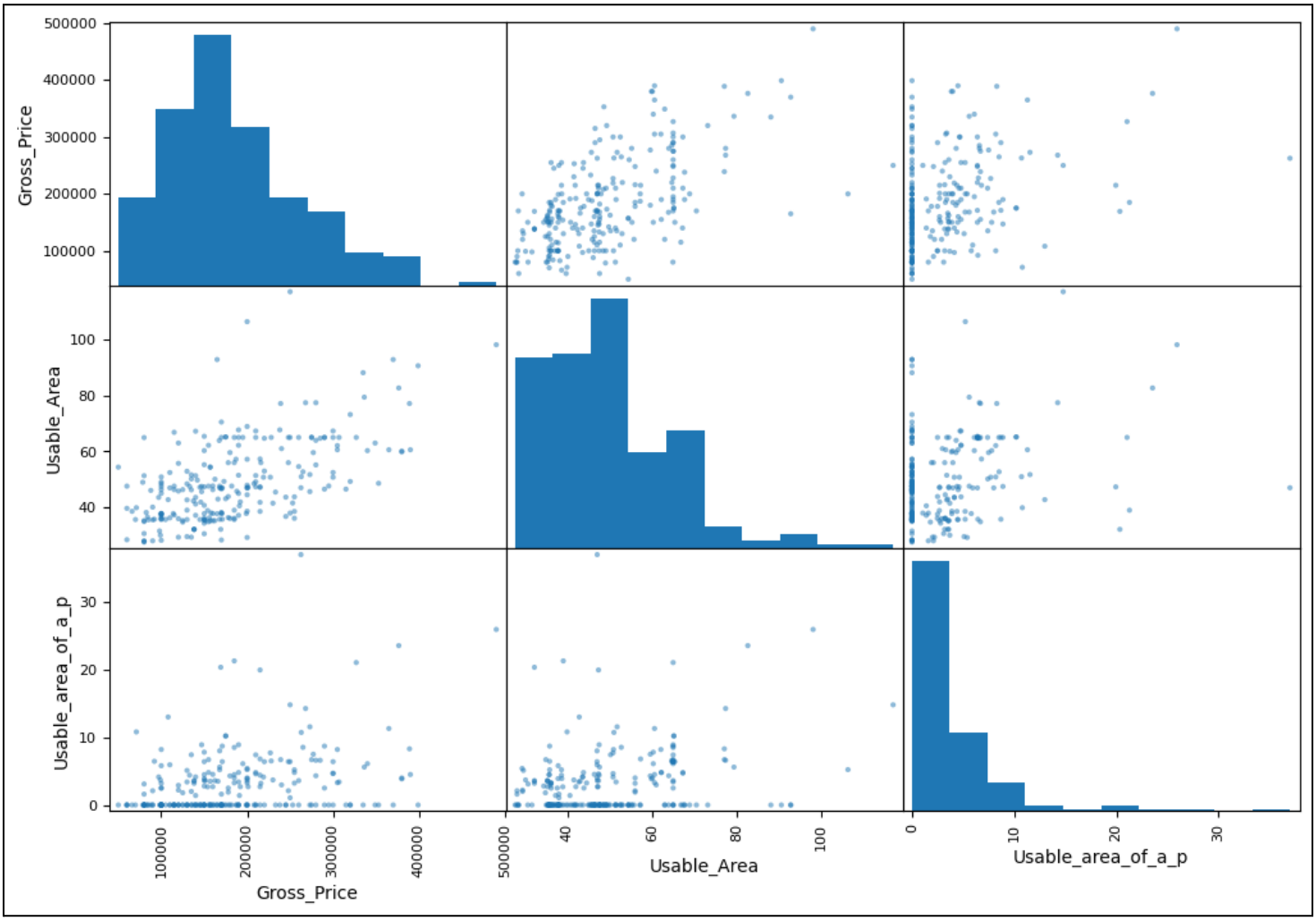
4. Bar charts for Categorical Variables



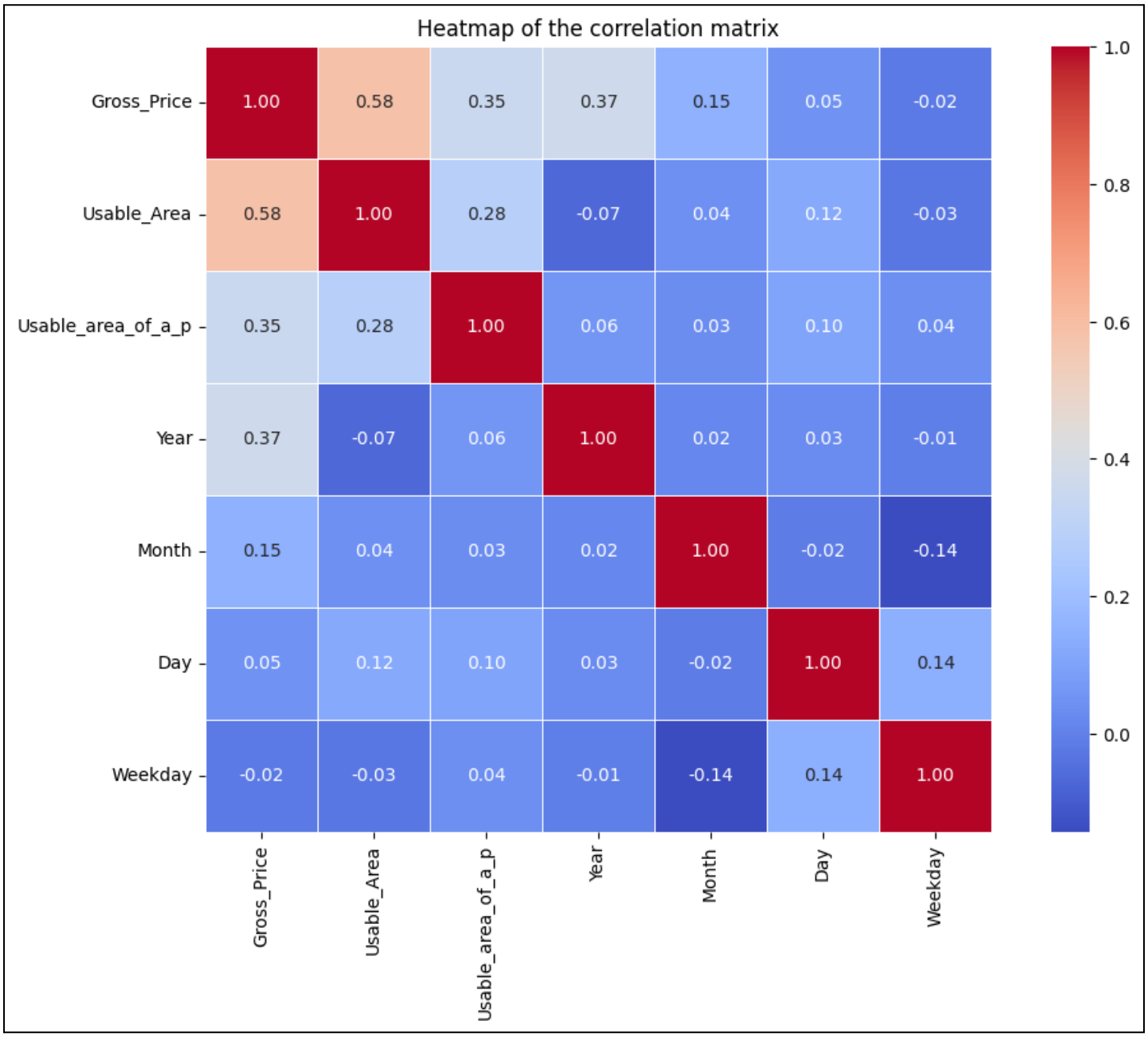



Distribution of categories in Floor
count
0
10
20
30
40
50
60
1
2
3
4
5
Floor

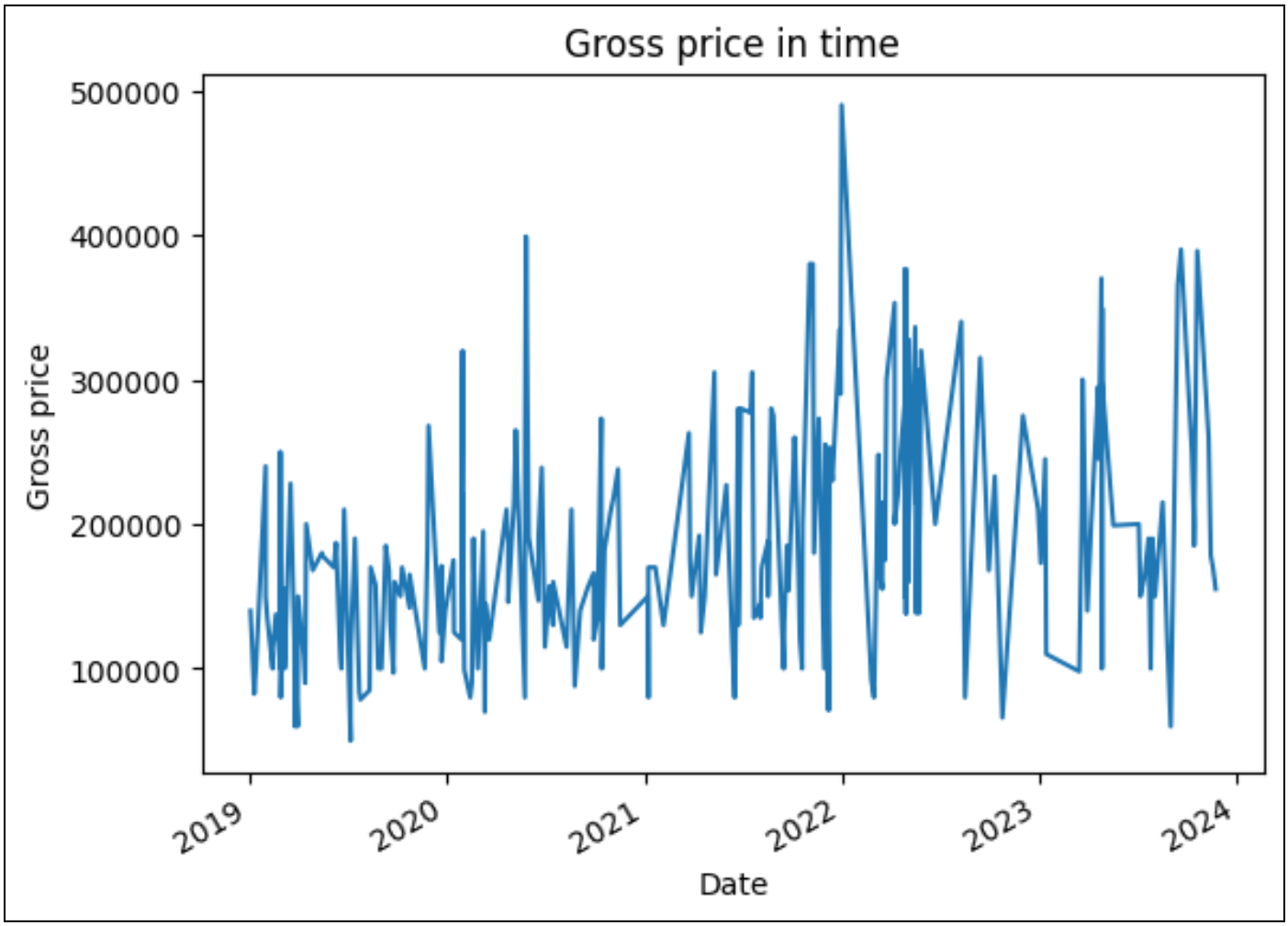
5. Scatterplot matrix



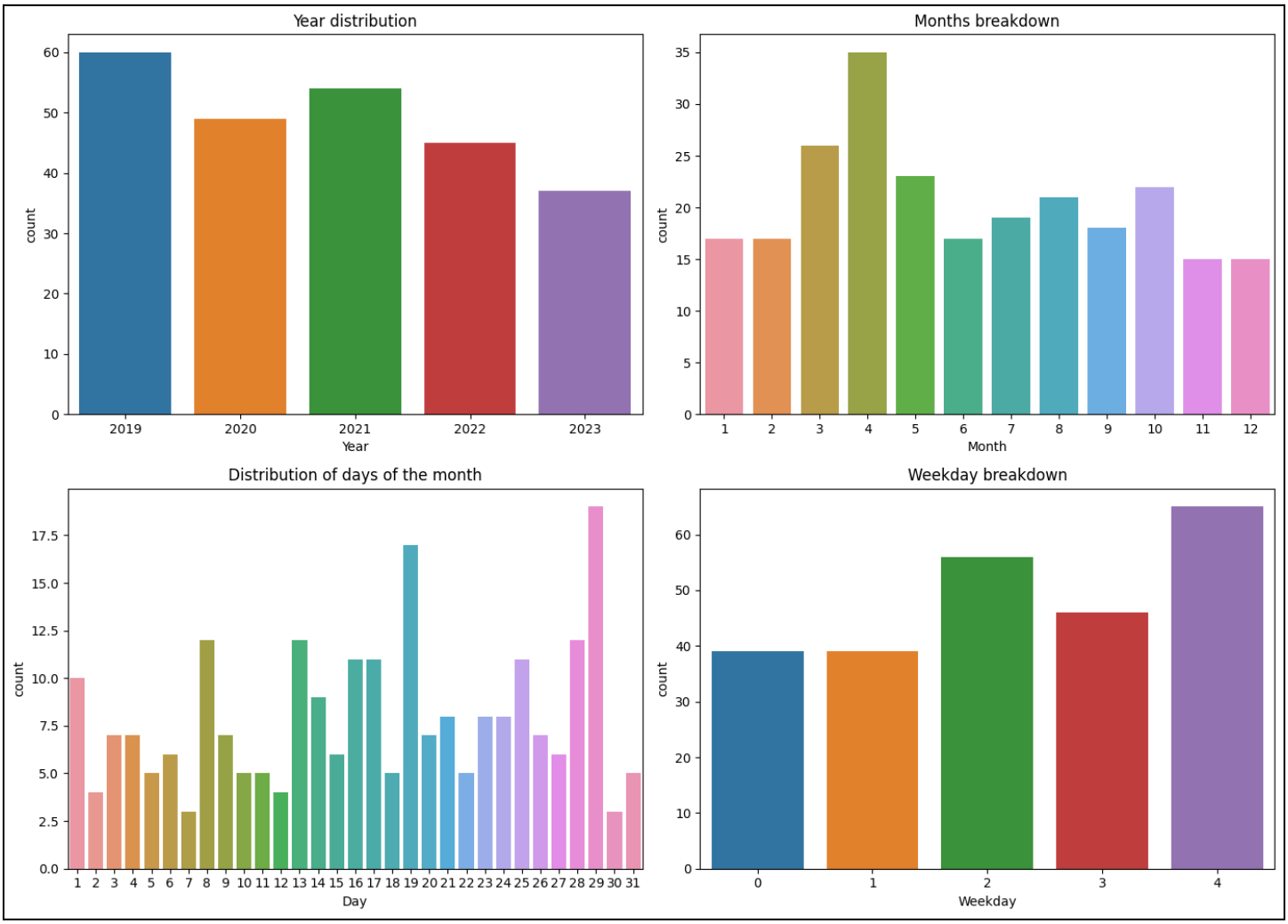
6. Heatmap of correlations



7. Time trends charts



8. 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

Pierwszy Model (tradycyjne metody uczenia maszynowego):

- Skalowanie Danych: Użycie StandardScaler do normalizacji zmiennych numerycznych, co pozwala na ujednolicenie skali i poprawę wydajności modeli.
- Kodowanie Zmiennych Kategorycznych: Zastosowanie OneHotEncoder do przekształcenia zmiennych kategorycznych w formę, którą mogą łatwo przetwarzać algorytmy uczenia maszynowego.
- Zastosowanie Potoków (Pipelines): Użycie Pipeline w scikit-learn do efektywnego łączenia etapów przetwarzania danych, co pozwala na uproszczenie kodu i uniknięcie błędów związanych z przetwarzaniem danych.

Drugi Model (sieci neuronowe):

- Normalizacja Danych: Zastosowanie warstwy normalizującej z TensorFlow (tf.keras.layers.Normalization) do automatycznego skalowania danych wejściowych, co jest kluczowe dla efektywnego trenowania sieci neuronowych.
- Przetwarzanie Zmiennych Kategorycznych: Tak jak w przypadku pierwszego modelu, zastosowanie OneHotEncoder do kodowania zmiennych kategorycznych.

Podział Danych

Dla Obu Modeli:

- Użycie funkcji train_test_split z biblioteki scikit-learn do podziału danych na zbiory treningowe i testowe.
- Typowy podział to 80% danych na trening i 20% na testowanie, choć ten stosunek może być dostosowany w zależności od wielkości i charakterystyki zbioru danych.
- Podział ten jest kluczowy dla weryfikacji wydajności modelu na nieznanych danych i zapobiegania problemowi nadmiernego dopasowania.
- Te metody przetwarzania i podziału danych są kluczowe dla zapewnienia, że modele są trenowane i testowane na odpowiednio przygotowanych i reprezentatywnych danych

5. Wybór Modelu

Kandydaci do Modelu:

Pierwszy Model (tradycyjne metody uczenia maszynowego):

- Linear Regression
- Decision Tree Regressor
- Random Forest Regressor
- Ridge Regression
- Lasso Regression
- Elastic Net

Drugi Model (sieci neuronowe):

- Sieć neuronowa z wykorzystaniem TensorFlow

Kryteria Wyboru:

Wybrano modele na podstawie ich zdolności do modelowania złożonych zależności w danych oraz elastyczności w dostosowywaniu do różnych typów danych.

Sieć neuronowa została wybrana ze względu na jej zdolność do modelowania nieliniowych zależności.

6. Trenowanie Modelu

Parametry Modelu:

Pierwszy Model (tradycyjne metody uczenia maszynowego):

- Różne konfiguracje hiperparametrów dla każdego modelu, zoptymalizowane za pomocą GridSearchCV i RandomizedSearchCV.

Drugi Model (sieci neuronowe):

- Architektura: Kilka warstw gęstych, każda z 500 neuronami.
- Funkcja aktywacji: ReLU.
- Optymalizator: Adam.
- Funkcja straty: MSE.

Proces Trenowania:

Pierwszy Model (tradycyjne metody uczenia maszynowego):

- Trenowanie za pomocą standardowych technik uczenia maszynowego z użyciem scikit-learn.

Drugi Model (sieci neuronowe):

- Trenowanie modelu przez 95 epok.
- Użycie TensorBoard do monitorowania postępów.

7. Ocena Modeli

Metryki Oceny:

- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)

Wyniki:
Pierwszy Model (tradycyjne metody uczenia maszynowego):

Wyniki modelu powinny być prezentowane w formie błędu średniokwadratowego dla różnych modeli, podkreślając ich wydajność na danych testowych.

Drugi Model (sieci neuronowe):

Wyniki oceny modelu sieci neuronowej na zbiorze testowym, prezentując wartości MSE i RMSE.

Wizualizacje porównujące rzeczywiste wartości z przewidywaniami modelu.

9. Wnioski i Dalsze Kierunki

Podsumowanie

Pierwszy Model (tradycyjne metody uczenia maszynowego):

- Modele wykazały zdolność do przewidywania cen mieszkań z zadowalającą dokładnością.
- Różne techniki modelowania pozwoliły na identyfikację najlepszego podejścia dla danego zestawu danych.
- Optymalizacja hiperparametrów za pomocą GridSearchCV i RandomizedSearchCV pozwoliła na znalezienie optymalnych konfiguracji modeli.

Drugi Model (sieci neuronowe):

- Najskuteczniejszy z najlepszymi wynikami
- Sieć neuronowa zapewniła możliwość modelowania bardziej złożonych wzorców w danych.
- Wysoka liczba epok trenowania i zastosowanie warstw o dużej liczbie neuronów umożliwiły dokładniejsze dopasowanie modelu do danych.

Propozycje Dalszego Rozwoju

1. Eksploracja Zaawansowanych Technik Modelowania:

Próba zastosowania innych architektur sieci neuronowych, takich jak konwolucyjne sieci neuronowe (CNN) lub rekurencyjne sieci neuronowe (RNN), które mogą lepiej radzić sobie z pewnymi rodzajami danych.

2. Głębsza Analiza Hiperparametrów:

Dalsze badania nad optymalizacją hiperparametrów, zwłaszcza dla sieci neuronowych, aby zwiększyć dokładność modelu.

3. Zastosowanie Modeli w Innych Kontekstach Danych:

Przetestowanie modeli na innych zbiorach danych dotyczących cen nieruchomości lub podobnych problemów regresyjnych, aby ocenić ich uniwersalność i skuteczność.

4. Integracja Modeli z Aplikacjami:

Rozwój interfejsów API lub aplikacji webowych, które mogą wykorzystywać te modele do przewidywania cen nieruchomości w czasie rzeczywistym.

Załączniki

Kody Źródłowe:
https://github.com/szymonszczurowski/Housing-Market-Analysis

Dodatkowe Materiały:
https://drive.google.com/drive/folders/1LjIEUIL-VV5nLi2gqCjrsW9BDzNZS3f2?usp=sharing