

Studium Licencjackie

Kierunek: Metody Ilościowe w Ekonomii i Systemy Informacyjne

Specjalność: ……………………………

Imię i nazwisko autora

Nr albumu 000000

**Tytuł tytuł tytuł tytuł tytuł tytuł tytuł tytuł tytuł** **tytuł tytuł tytuł tytuł tytuł tytuł**

Praca licencjacka

napisana w Katedrze/Instytucie

……………………………………………

pod kierunkiem naukowym

……………………...……………………

Warszawa 20…..

Spis treści

[Wstęp 5](#_Toc134996466)

[Rozdział I. Teoria i koncepcje związane z procesem ETL, web scrapowaniem oraz wizualizacją danych 6](#_Toc134996467)

[I.1 Pojęcie ETL 6](#_Toc134996468)

[I.2 Przykłady zastosowań procesu ETL 9](#_Toc134996469)

[I.3 Web scraping jako sposób pozyskiwania danych 10](#_Toc134996470)

[I.4 Wizualizacja danych jako narzędzie do analiz informacji 13](#_Toc134996471)

[Rozdział II. Analiza serwisu otodom.pl 20](#_Toc134996472)

[II.1 Charakterystyka serwisu 20](#_Toc134996473)

[II.2 Funkcjonalności oraz dane 21](#_Toc134996474)

[Rozdział III. Projekt procesu ETL 28](#_Toc134996475)

[III.1 Koncepcja i założenia 28](#_Toc134996476)

[III.2 Narzędzia i technologie 29](#_Toc134996477)

[III.2.1.1 Python 29](#_Toc134996478)

[III.2.1.2 Visual Code Studio (IDE) 30](#_Toc134996479)

[III.2.1.3 Selenium 31](#_Toc134996480)

[III.2.1.4 SQLite 32](#_Toc134996481)

[III.2.1.5 Apache Airflow 33](#_Toc134996482)

[III.2.1.6 Power BI 34](#_Toc134996483)

[Rozdział IV. Implementacja 37](#_Toc134996484)

[IV.1 Pozyskanie danych 37](#_Toc134996485)

[Zakończenie 37](#_Toc134996486)

[LITERATURA 39](#_Toc134996487)

[Spis tabel 42](#_Toc134996488)

[Spis rysunków 42](#_Toc134996489)

[Streszczenie 43](#_Toc134996490)

[Summary 43](#_Toc134996491)

Wstęp

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

# Teoria i koncepcje związane z procesem ETL, web scrapowaniem oraz wizualizacją danych

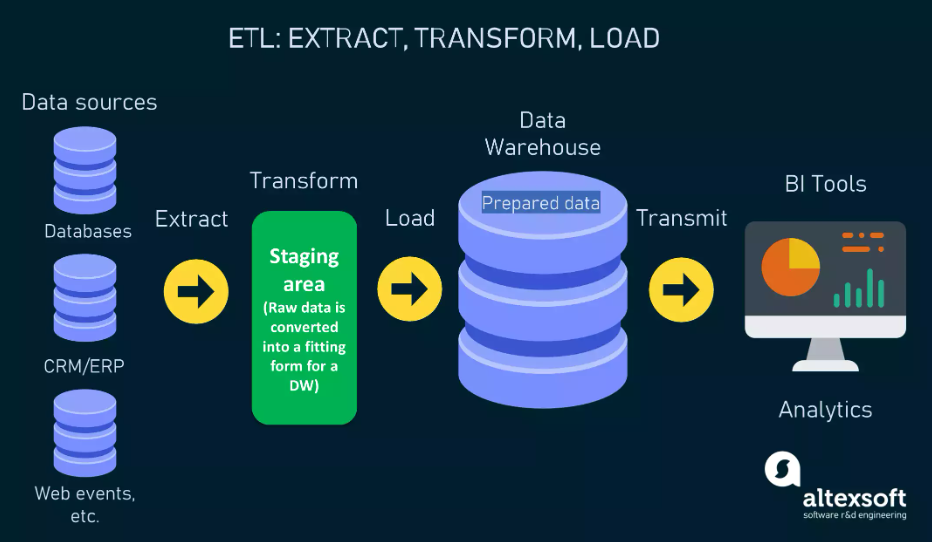
Poniższy rozdział skupia się na trzech bardzo ważnych aspektach analizy danych a co za tym idzie również opracowanego projektu, który przedstawiony zostanie w dalszej części pracy: procesie ETL, webscrapingu i wizualizacji. W pierwszym podrozdziale omówione zostanie, czym dokładnie jest proces ETL oraz jakie korzyści niesie za sobą jego stosowanie w kontekście analizy danych. Następnie, w drugim podrozdziale, zostaną przedstawione przykłady zastosowań procesu ETL w różnych dziedzinach biznesowych. Kolejnym ważnym aspektem, który zostanie omówiony w tej części, jest webscraping jako sposób pozyskania danych. Przedstawione zostaną tutaj narzędzia i techniki, które umożliwiają automatyczne pobieranie danych z różnych źródeł internetowych. Ostatni podrozdział skupi się na roli wizualizacji w procesie analizy danych. Omówione zostaną techniki i narzędzia do wizualizacyji, które pozwalają na skuteczną prezentację informacji i zrozumienie złożonych zbiorów danych.

## Pojęcie ETL

Skrót ETL (z języka angielskiego extraction, transformation, loading) oznacza wyodrębnianie, przekształcanie i ładowanie, to proces integracji danych, który łączy dane z wielu źródeł danych w jeden, spójny magazyn danych, który jest ładowany do hurtowni danych lub innego systemu docelowego co przestawia poniższa grafika. Wraz z rosnącą popularnością baz danych w latach 70-tych, proces ETL został wprowadzony jako sposób na integrację i ładowanie danych do obliczeń i analiz, stając się ostatecznie główną metodą przetwarzania danych dla projektów hurtowni danych. ETL stanowi podstawę dla procesów analitycznych i uczenia maszynowego. Poprzez serię reguł biznesowych, process ten oczyszcza i organizuje dane w sposób, który odpowiada określonym potrzebom dotyczącym business intelligence, takim jak miesięczne raportowanie, ale może także zajmować się bardziej zaawansowaną analizą, co może poprawić procesy backendowe lub doświadczenia użytkowników końcowych. ETL jest często wykorzystywany przez organizacje w celu[[1]](#footnote-1):

* Pobieranie danych z dotychczasowych systemów
* Oczyszczanie danych w celu poprawy ich jakości i ustalenia spójności
* Załadowanie danych do bazy danych

Rysunek 1. Przedstawienie procesu ETL.

**Źródło:** ETL overview <https://www.altexsoft.com/blog/etl-vs-elt/> [19-03-2023].

Najprostrzym sposobem aby zrozumień czym jest ETL jest zrozumienie co następuje podczas każdego z trzech etapów.

Podczas wyodrębniania, dane surowe są kopiowane lub eksportowane z lokalziacji źródłowych do obszaru etapowego. Zespoły zarządzające danymi mogą pobierać je z różnych źródeł, które mogą być ustruktyzowane lub nieustruktyzowane. Źródła te mogą obejmować np[[2]](#footnote-2):

* Serwery SQL lub NoSQL
* Systemu CRM lub ERP
* Emaile
* Strony internetowe
* API
* Pliki

Podczas przekształcania danych, kiedy są one obszarze etapowym, następuje ich transformacja oraz konsolidowanie z myślą o ich zamierzonej analizie. Ta faza może obejmować wiele działań takich jak[[3]](#footnote-3):

* Filtrowanie, oczyszczanie, usuwanie duplikatów, weryfikacja i uwierzytelnianie danych,
* Wykonywanie obliczeń, tłumaczeń lub podsumowań na podstawie surowych danych. Może to obejmować zmianę nagłówków wierszy i kolumn dla spójności, konwersję walut lub innych jednostek miary, edycję ciągów tekstowych i wiele innych,
* Przeprowadzanie audytów w celu zapewnienia jakości danych i zgodności.
* Usuwanie, szyfrowanie lub ochrona danych objętych regulacjami branżowymi lub rządowymi.
* Formatowanie danych do tabel lub tabel połączonych w celu dopasowania schematu docelowej hurtowni danych.

Podczas ładowania danych, które jest ostatnim etapem ETL, przekształcone dane są przenoszone z obszaru etapowego do docelowej lokalizacji. Zazwyczaj obejmuje to początkowe wczytanie wszystkich danych, a następnie okresowe wczytywanie zmian danych inkrementalnych i rzadziej, pełne odświeżanie w celu usunięcia i zastąpienia danych w magazynie danych. Dla większości organizacji korzystających z ETL, proces jest zautomatyzowany, dobrze zdefiniowany, ciągły i sterowany partiami. Zwykle ETL odbywa się w godzinach poza szczytem, gdy ruch na systemach źródłowych i magazynie danych jest najniższy.[[4]](#footnote-4)

Można wyróżnić wiele rodzajów ETL jednak potrzeby biznesowe zwykle powodują, że najczęściej używanymi są[[5]](#footnote-5):

* Przetwarzanie partiami (batch processing): Dane są pobierane i przetwarzane w określonych partiach, zazwyczaj w określonych odstępach czasu, takich jak co godzinę lub co dzień,
* Przetwarzanie w czasie rzeczywistym (real-time processing): Dane są przetwarzane natychmiast po ich pobraniu ze źródłowego systemu, umożliwiając szybsze i bardziej aktualne wyniki przetwarzania.

Podsumowująć, proces etl składa się z trzech etapów jakimi są pozyskanie, przekształcenie oraz importowanie danych do jednej lokalizacji. Każdy z tych etapów może być przeprowadzony w różny sposób. Podobnie do jego podprocesów, ETL można przeprowadzić na wiele sposobów w zależności od potrzeb biznesowych.

## Przykłady zastosowań procesu ETL

Proces ETL jest powszechnie stosowany w różnych obszarach biznesowych do integrowania, przekształcania i konsolidowania danych z różnych źródeł w jedną, ujednoliconą bazę danych lub hurtownię danych. Poniżej przedstawiono 5 przykładów.

Pierwszą dziedziną biznesową w której powszechnym jest budowanie procesów ETL to finanse i bankowość. W tej branży ETL służy do konsolidacji danych z różnych źródeł, takich jak banki, firmy obsługujące karty kredytowe i instytucje finansowe. Dane te mogą być wykorzystywane do analizowania zachowań klientów, identyfikowania potencjalnych zagrożeń i wykrywania oszustw.

ETL jest również używany w branży medycznej do integracji danych ze źródeł, takich jak elektroniczna dokumentacja medyczna, roszczenia ubezpieczeniowe i badania kliniczne. Dane te można wykorzystać do monitorowania wyników pacjentów, analizowania skuteczności leczenia i poprawy świadczenia opieki zdrowotnej.

Kolejnym obszarem jest sprzedaż. W branży detalicznej ETL służy do konsolidacji danych ze źródeł takich jak systemy punktów sprzedaży, programów lojalnościowych dla klientów i mediów społecznościowych. Dane te można wykorzystać do analizy zachowań klientów, śledzenia trendów sprzedaży i optymalizacji zarządzania zapasami.

Marketing: W branży marketingowej ETL służy do konsolidacji danych z kampanii reklamowych, ankiet konsumenckich i mediów społecznościowych. Dane te mogą być wykorzystywane do analizowania zachowań klientów, śledzenia wyników działań marketingowych i optymalizacji kampanii marketingowych.

Ostatnim przykładem jest produkcja: ETL jest używany w przemyśle wytwórczym do integracji danych ze źródeł, takich jak systemy produkcyjne, systemy łańcucha dostaw i systemy kontroli jakości. Dane te można wykorzystać do monitorowania wydajności produkcji, identyfikowania wąskich gardeł i optymalizacji procesów produkcyjnych.

Podsumowując, proces ETL jest używany w różnych obszarach biznesowych do integrowania, przekształcania i konsolidowania danych z różnych źródeł oraz zapewniania wglądu w świadome podejmowanie decyzji.

## Web scraping jako sposób pozyskiwania danych

Każdy proces ETL rozpoczyna się od pozyskania danych z różnego rodzaju źródeł, które zostały wcześniej wymienione. W tym podrozdziale zostanie przedstawione pozyskiwanie danych ze stron internetowych, które jest nazywane web scrapingiem.

Web scraping to technika używana do automatycznego wydobywania danych ze stron internetowych. Wyodrębnione dane można wykorzystywać do różnych celów, takich jak analiza danych czy uczenie maszynowe. Dane są wyodrębniane za pomocą programowanych narzędzi zwanych scraperami, które mogą przeszukiwać strony internetowe i przechowywać je w ustrukturyzowanym formacie, takim jak np. baza danych lub arkusz kalkulacyjny. „Skrobanie” sieci można wykonać przy użyciu różnych języków programowania i bibliotek, takich jak Python, R, Scrapy, Selenium, Requests czy BS4. Wykorzystanie „skrobania” stron internetowych wzrosło w ostatnich latach ze względu na rozwój Internetu i dostępność danych. Jednak web scraping może być złożonym i trudnym zadaniem, ponieważ strony internetowe mogą mieć skomplikowane do odczytu struktury, a pobieranie danych może być utrudnione przez właścicieli witryn[[6]](#footnote-6).

Proces Web scrapowania składa się z dwóch etapów jakimi są pozyskanie oraz ich przetworzenia do pożądanej formy. Dane najczęściej są pozyskiwane w dwóch formach jakimi są json lub kod html.

Json to popularny format do wymiany danych, używany przez różne API (interfejsy programowania aplikacji). Json pozwala na łączenie dużych ilości danych w pojedynczym pakiecie tekstu i wysyłanie ich do innych serwisów. Każda wartość danych ma swoją nazwę, czyli klucz. W przypadku tego formatu danych pozyskanie danych jest stosunkowo proste i można je uzyskać za pomocą biblioteki Requests (przykład poniżej)[[7]](#footnote-7).

Rysunek 2. Przykładowy kod pozyskania danych w formie json przy użyciu biblioteki Requests.

Obraz zawierający tekst

Opis wygenerowany automatycznie

**Źródło:** Opracowanie własne.

W powyższym przykładzie najpierw definiowany jestpunkt końcowy interfejsu API i parametry jako zmienne. Następnie używane zostają metody request.get() do wysłania żądania GET do API, przekazując adres URL punktu końcowego i parametry jako argumenty. Odpowiedź jest przechowana w zmiennej o nazwie response. Następnie sprawdzany jest kod statusu odpowiedzi za pomocą atrybutu response.status\_code. Jeśli kod statusu to 200, można założuć, że żądanie zakończyło się pomyślnie, więc dane są wyodrębniane z odpowiedzi za pomocą metody response.json(), która zwraca słownik Pythona. Następnie można coś zrobić z danymi, na przykład wydrukować je w konsoli lub utworzyć z nich ramkę danych. Jeśli kod stanu nie jest równy 200, można założyć, że żądanie nie powiodło się, więc pokazany zostanie komunikat o błędzie przy użyciu atrybutu response.text, który zawiera informację o błędzie zwróconym przez interfejs API.

Kolejnym formatem jest kod HTML (Hypertext Markup Language) odgrywa kluczową rolę w scrapowaniu stron internetowych, ponieważ zapewnia strukturę i zawartość stron internetowych. HTML jest językiem znaczników używanym do tworzenia stron internetowych i składa się z zestawu znaczników definiujących elementy i zawartość strony internetowej, takie jak tekst, obrazy, łącza, formularze i inne. Aby zeskrobać dane ze strony internetowej, skrobak musi zrozumieć strukturę i układ kodu HTML tej witryny. Kod HTML strony internetowej zawiera informacje o strukturze i zawartości strony internetowej, takie jak położenie poszczególnych elementów, hierarchia dokumentu i relacje między różnymi elementami. Poniżej znajduję się przykładowy kod, który pozwala na wydobycie potrzebnych danych z kodu HTML[[8]](#footnote-8).

W tym przykładzie najpierw importowane są biblioteki Requests i BeautifulSoup. Następnie określamy adres URL strony, z której chcemy pobrać dane, przypisując go do zmiennej url. Następnie zostaje wysyłane żądanie GET na adres URL za pomocą request.get() i odpowiedź serwera zostaje zachowana w zmiennej response. Następnie zostaje dokonana analiza zawartości odpwiedzi HTML za pomocą Beautiful Soup, tworząc obiekt BeautifulSoup z treścią odpowiedzi i określonym parserem. Na koniec zostaje wyświetlony tytuł strony za pomocą soup.title.text i wszystkie linki na stronie za pomocą pętli for, która przegląda wszystkie znaczniki <a> na stronie i drukuje wartość atrybutu href dla każdego znacznika za pomocą link.get('href').

Rysunek 3. Przykładowy kod pozyskania danych w formacie kodu HTML przy użyciu bibliotek BeatifulSoup i Requests.

Obraz zawierający tekst

Opis wygenerowany automatycznie

**Źródło:** Opracowanie własne.

Podsumowując, web scraping jest potężnym narzędziem do uzyskiwania dostępu i analizowania danych ze stron internetowych, a jego zalety sprawiają, że jest to niezbędna technika dla firm i osób prywatnych, które polegają na danych przy podejmowaniu decyzji i analizie.

## Wizualizacja danych jako narzędzie do analiz informacji

Wizualizacja danych jest niezbędnym narzędziem do analizy danych w wielu dziedzinach biznesowych oraz naukowych. Pozwala na przedstawienie dużych i złożonych zbiorów danych w łatwym do zrozumienia formacie, pomagając decydentom w szybkim identyfikowaniu wzorców, trendów i spostrzeżeń, które mogą być trudne do zidentyfikowania za pomocą tradycyjnych metod analitycznych[[9]](#footnote-9).

Istnieje kilka typów technik wizualizacji danych, które są powszechnie stosowane do reprezentowania danych w zrozumiały sposób.

Wykresy słupkowe służą do przedstawiania danych liczbowych za pomocą prostokątnych słupków, przy czym wysokość lub długość każdego słupka reprezentują wartość danych. Wykresów słupkowych można używać do porównywania różnych kategorii lub do śledzenia zmian w czasie[[10]](#footnote-10).

Rysunek 4. Przykładowy wykres słupkowy.

Obraz zawierający wykres

Opis wygenerowany automatycznie

**Źródło:** Opracowanie własne.

Wykres liniowy służy do wyświetlania wartości ilościowych w ciągłym przedziale lub okresie. Wykres ten jest najczęściej używany do przedstawiania trendów i analizowania zmian danych w czasie. Wykresy liniowe są rysowane przez wykreślenie punktów danych na kartezjańskiej siatce współrzędnych, a następnie połączenie linii między wszystkimi tymi punktami. Zazwyczaj oś Y ma wartość ilościową, podczas gdy oś X to skala czasu lub sekwencja interwałów. Wartości ujemne mogą być wyświetlane poniżej osi X. Kierunek linii na wykresie jest dobrą metaforą danych: nachylenie w górę wskazuje, gdzie wartości wzrosły, a nachylenie w dół wskazuje, gdzie wartości spadły. Podróż linii po wykresie może tworzyć wzorce, które ujawniają trendy w zbiorze danych[[11]](#footnote-11).

**Rysunek 5. Przykładowy wykres liniowy.**

Obraz zawierający wykres

Opis wygenerowany automatycznie

**Źródło:** Opracowanie własne.

Mapy cieplne wizualizują dane za pomocą różnic w kolorystyce. W przypadku formatu tabelarycznego są przydatne do analizy danych wielowymiarowych poprzez umieszczanie zmiennych w wierszach i kolumnach oraz kolorowanie komórek w tabeli. Mapy cieplne są dobre do pokazywania wariancji między wieloma zmiennymi, ujawniania wszelkich wzorców, wyświetlania, czy jakieś zmienne są do siebie podobne, oraz do wykrywania, czy istnieją jakiekolwiek korelacje. Dane liczbowe wymagają skali kolorów, która łączy jeden kolor w drugi, aby przedstawić różnicę w wysokich i niskich wartościach[[12]](#footnote-12).

**Rysunek 6. Przykładowy wykres cieplny.**

Obraz zawierający stół

Opis wygenerowany automatycznie

**Źródło:** <https://datavizcatalogue.com/methods/heatmap.html>, [06-04-2022].

Wykres punktowy umieszcza punkty w systemie współrzędnych kartezjańskich, aby wyświetlić wszystkie wartości między dwiema zmiennymi. Za jego pomocą można wykryć czy istnieje związek lub korelacja między nimi. Rodzaj korelacji można zinterpretować na podstawie wzorców ujawnionych na wykresie rozrzutu. Są to: dodatnie (wartości rosną razem), ujemne (jedna wartość maleje wraz ze wzrostem drugiej) lub zerowa (brak korelacji). Siłę korelacji można określić na podstawie tego, jak blisko siebie rozmieszczone są punkty na wykresie[[13]](#footnote-13).

**Rysunek 7. Przykładowe wykresy punktowe.**

Obraz zawierający wykres

Opis wygenerowany automatycznie

**Źródło:** Opracowanie własne.

Mapy geograficzne to popularny rodzaj wizualizacji danych używany do przedstawiania danych na podstawie lokalizacji. Można ich używać do wizualizacji wzorców i trendów w różnych regionach, co ułatwia zrozumienie rozkładu danych w różnych obszarach. Mapy geograficzne są często używane w dziedzinach takich jak epidemiologia, biznes i nauki społeczne do identyfikowania wzorców i trendów, które mogą nie być oczywiste w danych tabelarycznych lub tekstowych. Jedną z kluczowych zalet map geograficznych jest to, że mogą one zapewniać wizualną reprezentację danych, która może być łatwo zrozumiana przez osoby nieobeznane ze statystycznym lub technicznym żargonem. Na przykład mapa przedstawiająca rozpowszechnienie określonej choroby w różnych województwach[[14]](#footnote-14).

**Rysunek 8. Przykładowy wykres mapy.**

Obraz zawierający mapa

Opis wygenerowany automatycznie

**Źródło:** <https://datavizcatalogue.com/methods/heatmap.html>, [06-04-2022].

Powyżej zostały przedstawione wizualizacje które między innymi zostały przedstawione w projekcie, jednak istnieje wiele innych rodzajów technik wizualizacji danych, a wybór której należy użyć zależy od rodzaju analizowanych danych i potrzeb analizy. Istotym elementem wizualizacji jest również dobór narzędzi.

Obecnie osoby zajmujące się wizualizacją danych mają dostęp do wielu narzędzi, dzięki którym stosunkowo małym nakładem pracy mogą budować złożone wizualizację. W pojekcie został użyty program Microsoft Power BI. Zostanie on opisany w dalszej części pracy. Jest on jednym z najbardziej popularnych programów tego typu na rynku. Innymi programami, które są powszechnie wykorzystywane są np.: Tabelau, Google Data Studio, Looker, Qlik oraz wiele innych. Narzędzia te cechują się wieloma wspólnymi aspektami takimi jak[[15]](#footnote-15):

* Łączenie i integracja danych: Wszystkie te narzędzia umożliwiają użytkownikom łączenie się z różnymi źródłami danych, takimi jak arkusze kalkulacyjne, bazy danych, usługi w chmurze i inne aplikacje, umożliwiając użytkownikom tworzenie wizualizacji i pulpitów nawigacyjnych przy użyciu najbardziej aktualnych informacji.
* Interfejs „drag and drop”: te narzędzia wykorzystują interfejs „drag and drop”, aby umożliwić użytkownikom szybkie i łatwe tworzenie wykresów, wykresów i innych wizualizacji.
* Możliwości dowolnych wizualizacji: narzędzia te oferują szeroki zakres opcji wizualizacji, aby umożliwić użytkownikom tworzenie znaczących i wnikliwych raportów i pulpitów nawigacyjnych. Obejmują one wykresy kołowe, wykresy słupkowe, wykresy liniowe, wykresy punktowe, mapy cieplne i inne.
* Współpraca i udostępnianie: Wszystkie te narzędzia umożliwiają użytkownikom udostępnianie raportów, wizualizacji i pulpitów nawigacyjnych innym członkom zespołu lub interesariuszom w celu współpracy i podejmowania decyzji.
* Personalizacja: te narzędzia oferują różne opcje dostosowywania, takie jak kolory, style, czcionki i inne, aby umożliwić użytkownikom tworzenie spersonalizowanych i profesjonalnie wyglądających raportów i pulpitów nawigacyjnych.

Ogólnie rzecz biorąc, narzędzia i techniki do wizualizacji danych mają na celu pomóc organizacjom w przekształcaniu dużych i złożonych zbiorów danych w praktyczne spostrzeżenia i umożliwiać podejmowanie świadomych decyzji.

# Analiza serwisu otodom.pl

Poniższy rozdział skupia się na opisie oraz analizie serwisu internetowego jakim jest otodom.pl. Stanowi on źródło danych opracowanego projektu. W pierwszym pod rozdziale zostanie przedstawiona ogólna charakterystyka tejże platformy, czyli przedstawienie czym ona jest oraz do czego służy. W drugim podrozdziale zostaną przedstawione jej funkcjonalności wraz z danymi które można z niej pozyskać.

## Charakterystyka serwisu

Otodom.pl to jeden z największych i najpopularniejszych serwisów internetowych oferujących nieruchomości w Polsce. Serwis ten stanowi platformę dla sprzedających, wynajmujących oraz poszukujących mieszkań, domów, działek, lokali użytkowych oraz komercyjnych. Otodom.pl został założony w 2006 roku i od tamtej pory stał się jednym z najważniejszych graczy na rynku nieruchomości w Polsce[[16]](#footnote-16).

Serwis otodom.pl działa na zasadzie portalu ogłoszeniowego, na którym prywatni oraz profesjonalni użytkownicy mogą umieszczać ogłoszenia dotyczące nieruchomości. Ogłoszenia te są w pełni darmowe i w łatwy sposób można je przeglądać dzięki intuicyjnemu interfejsowi użytkownika. Dostępne są filtry wyszukiwania, które pozwalają na precyzyjne określenie kryteriów poszukiwanej nieruchomości, takich jak lokalizacja, metraż, liczba pokoi, cena czy typ nieruchomości[[17]](#footnote-17).

Otodom jest obecnie najbardziej popularnym serwisem nieruchomości w Polsce. Według najnowszego raportu PBI Gemius, w samym tylko wrześniu (rok 2020) odwiedziło go ponad 4 miliony rzeczywistych użytkowników co przełożyło się na ponad 18 milionów wizyt. Serwis aktywnie współpracuje z 975 firmami deweloperskimi. Ponadto otodom charakteryzuję się największą liczbą ofert mieszkań. Poniższy wykres przedstawia liczbę ofert w warszawie, które oferuję konkurencja. Pod tym względem inne serwisu cechują się mniej okazałą ofertą[[18]](#footnote-18).

**Wykres 1. Porównanie liczby ogłoszeń mieszkań w Warszawie.**

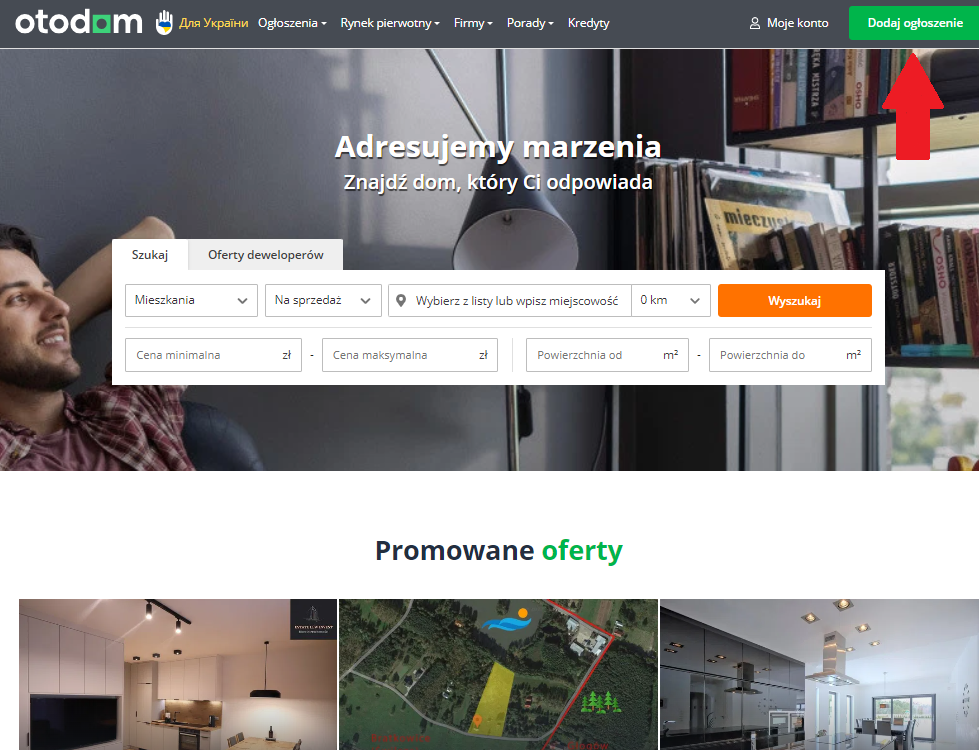
**Źródło**: Opracowanie własne na podstawie <https://www.otodom.pl/pl/oferty/sprzedaz/mieszkanie/warszawa>, <https://warszawa.nieruchomosci-online.pl/mieszkania,sprzedaz/>, <https://mieszkanie.pl/ogloszenia/szukaj/mieszkanie/sprzedaz/warszawa>, <https://www.morizon.pl/mieszkania/warszawa/>, <https://gratka.pl/nieruchomosci/mieszkania/warszawa>, <https://www.domiporta.pl/mieszkanie/sprzedam?Localizations%5B0%5D.Name=warszawa>, [02-05-2023].

## Funkcjonalności oraz dane

Otodom oferuje szereg fukcnonalności, które umożliwiają użytkownikom łatwe i efektwyne koszystanie z serwisu. Główną funkcjonalnością na której bazuje serwis otodom.pl jest zamieszacznie ogłoszeń przez sprzadających. Proces ten jest stosunkowo intuicyjny dla użytkowników. Składa się on z trzech etapów, podczas których zbierane są dane o ofercie. Celem tego działania jest zbieranie parametrów które pomagają w wyszukiwaniu jak najbardziej spersonalizowanych ogłoszeń. Etapami tymi są:

* Pierwszym etapem jest użycie przycisku “Dodaj ogłoszenie”, który widnieje w prawym górnym rogu strony głównej. Jest on w kolorze jasno zielonym, dzięki czemu wyróżnia się on na tle innych elementów na stronie, które są w odcieniach bardziej stonowanych. Jest to przedstawione na grafice poniżej (czerwona strzałka).

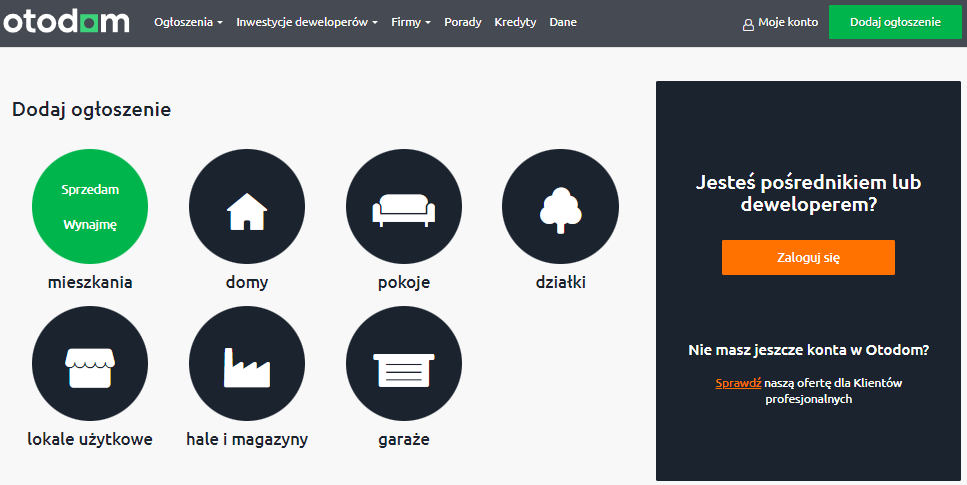
**Rysunek 9. Strona główna serwisu otodom.**



**Źródło**: <https://www.otodom.pl/>, [02-05-2023].

* Drugim etapem jest wybranie rodzaju ogłoszenia. Tutaj następuję pierwsze wprowadzenie danych przez sprzedającego. Użytkownik wybiera spośród takich możliwości jak: mieszkania, domy, pokoje, działki,locale użytkowe, hale i magazyny oraz garaże. Każda z typu ofert może obrać dwa stany jakimi są “na sprzedaż” lub“na wynajem”. Widoczne to jest w grafice poniżej. W ninijszej pracy analizowane zostaną danę dotyczące ofert mieszkań na sprzedaż w warszawie więc dalsza ścieżka dodawania ogłoszenia będzie z tym zgodna.

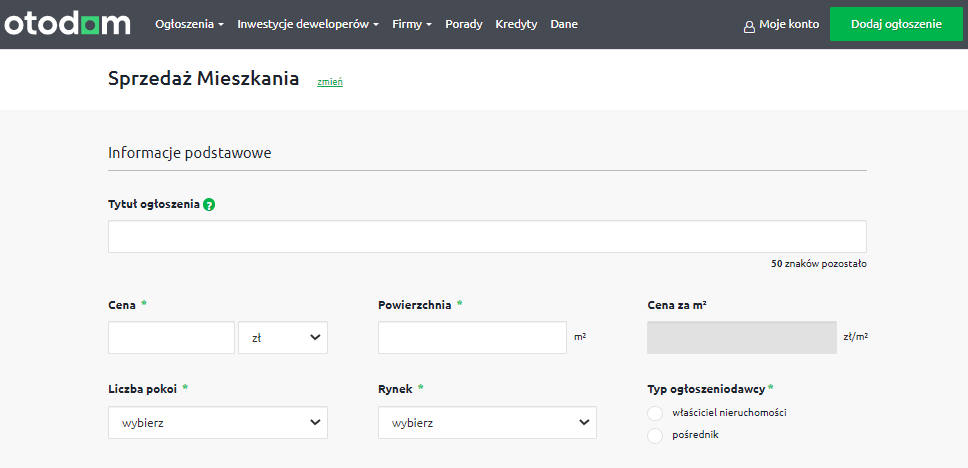
**Rysunek 10. Strona z wyborem typu ogłoszenia serwisu otodom.**



**Źródło**: <https://www.otodom.pl/nowe-ogloszenie/kategoria/>, [02-05-2023].

* Podczas trzeciego etapu następuje wprowadzenie szczegółowych informacji o nieruchomości. Są nimi między innymi powierzchnia, cena, liczba pokoi oraz wiele innych parametrów które zostaną omówione w dalszej części rozdziału. Wprowadzenie danych polega na uzupełnieniu formularza, którego część można zauważyć na poniższej grafice.

**Rysunek 11. Przedstawienie formularza z wyborem typu ogłoszenia serwisu otodom.**



**Źródło**: <https://www.otodom.pl/sprzedaz/mieszkanie/nowe-ogloszenie/>, [02-05-2023].

Następnie, na podstawie wprowadzonych danych, generowana jest strona z ogłoszeniem, która jest jednym z tysięcy źródeł danych opracowanego projektu. Dane, które są pozyskiwane dzielą się na cztery sekcję jakimi są:

* Lokalizacja, jest pozyskiwana z górnego sektora strony. Widnieją tam informację na temat województwa, miasta, dzielnicy, poddzielnicy oraz ulicy (Rysunek 13., oznaczenie **1**).
* Cena, jest umieszczona pod zdjęciami i wyrażona jest w złotówkach. Poniżej znajduję się również cena za metr kwadratowy. Jest to jedna z kluczowych parametrów każdej z ofert. (Rysunek 13., oznaczenie **2**).
* Dane szczegółowe, jest to sekcja w której zawiera się 10 parametrów oferty (Rysunek 13., oznaczenie **3**). Można tam znaleźć informację na temat:
* Powierzchni, wyrażonej w metrach w kwadratowych,
* Liczby pokoi,
* Piętra na którym znajduję się nieruchomość,
* Ceny czynszu,
* Formy własności,
* Stanie wykończenia,
* Informacji o balkonie/ogrodzie/tarasie.
* Miejsca parkingowego,
* Typu ogrzewania
* Informacji o obsłudze zdalnej
* Informację dodatkowe (Rysunek 12.) znajdują się w dolnych rejonach strony. Jest to sekcja w której znajdują się informacje takie jak:
* Rodzaj rynku (wtórny lub pierwotny),
* Typ ogłoszeniodawcy
* Data od której nieruchomość jest dostępna,
* Rok budowy,
* Rodzaj zabudowy,
* Rodzaj okien,
* Dostęp do windy,
* Media,
* Zabezpieczenia,
* Wyposażenie,
* Informacje dodatkowe,
* Materiał budyku.

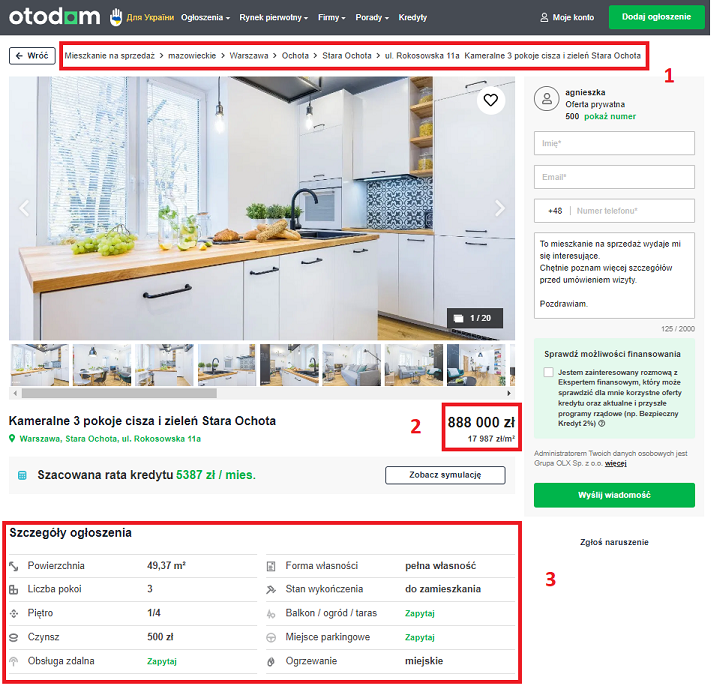
**Rysunek 12. Element strony ogłoszenia z informacjami dodatkowymi.**

Obraz zawierający tekst

Opis wygenerowany automatycznie

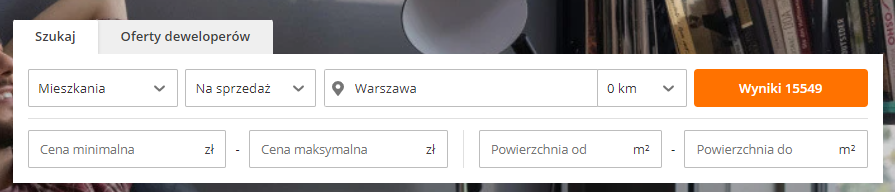
**Źródło**: <https://www.otodom.pl/pl/oferta/kameralne-3-pokoje-cisza-i-zielen-stara-ochota-ID4l96L>, [02-05-2023].

**Rysunek 13. Przykładowa oferty z podziałem na sekcję danych.**

**Źródło**: <https://www.otodom.pl/pl/oferta/kameralne-3-pokoje-cisza-i-zielen-stara-ochota-ID4l96L>, [02-05-2023].

Na podstawie przedstawionych danych, które wprowadził sprzedający, kupujący mogą z łatwością znaleźć nieruchomości o wymaganych parametrach. Jest to możliwe również dzięki panelowi na stronie głownej (Rysunek 14.). Użytkownik może w prosty sposób uzupełnić formularz i w oparciu o wporwadzone dane, zaprezentowane będą oferty zgodne z jego oczekiwaniami. Jest to istnotny element projektu, ponieważ dzięki tej fuknconalności możliwe było wygenerowanie adresu url do stron z ogłoszeniami mieszkań na sprzedaż w warszawie.

**Rysunek 14. Panel ułatwiający szukanie porządanych ogłoszeń.**



**Źródło**: <https://www.otodom.pl/> , [02-05-2023].

Przedstawione funkcjonalności oraz dane, które posiada serwis otodom sprawiają, że jest on w pewnym sensie obszerną „bazą danych” o rynku nieruchomości w Polsce. Dzieki web scrapowaniu, można darmowo czerpać informacje, przetwarzać je oraz dokonywać analiz na ich podstawie. W kolejnym rozdziale zostanie szczegółowo opisany tenże proces jak i narzędzia oraz cały projekt.

# Projekt procesu ETL

Poniższy rozdział opisuje koncepcje, założenia a także narzędzia, które zostały użyte do wykonania projektu.

## Koncepcja i założenia

Celem projektu jest opracowanie procesu ETL (Extract, Transform, Load) umożliwiającego pobieranie, transformowanie,magazynowanie i wizualizację danych z serwisu Otodom.pl. Projekt zakłada automatyczne pozyskiwanie informacji o nieruchomościach dostępnych na serwisie Otodom.pl, przetwarzanie tych danych w celu uzyskania spójnej struktury oraz przechowywanie ich w dedykowanej bazie danych. Następnie, dane te będą wizualizowane w formie interaktywnych wykresów i dashboardów, umożliwiających łatwą analizę i eksploatację zgromadzonych informacji.

Projekt opiera się czterech głównych etapach jakimi są:

1. Pozyskanie danych dzięki technikom web scrapowania przy użyciu takich narzędzi jak Visual Studio Code, Airflow, język programowania Python oraz biblioteka Selenium.
2. Przekształcenie danych, dzięki wykorzystaniu języka Python oraz biblioteki Pandas.
3. Ładowania danych, dzięki któremy magazynujemy zdobyte dane do bazy danych. W przedstawionym projekcie została wykorzystana baza danych SQLite.
4. Wizualizacja danych, dzięki której można lepiej zrozumieć dane, została wykonana za pomocą programu Microsoft Power Bi.

Założeniami projektu są następujące:

* Pobieranie danych następuję w raz w tygodniu (sobota),
* Dane dotyczą nieruchomości wyłącznie z Warszawy.

## Narzędzia i technologie

#### Python

Python to wysokopoziomowy język programowania, który został stworzony przez Guido van Rossuma i po raz pierwszy został wydany w 1991 roku. Jest to język interpretowany, co oznacza, że kod tworzony w Pythonie jest tłumaczony na bieżąco w trakcie jego wykonywania, a nie kompilowany do kodu maszynowego przed uruchomieniem.

Python jest językiem dynamicznym, co oznacza, że nie trzeba deklarować typów zmiennych przed ich użyciem. W związku z tym jest elastyczny i umożliwia programistom pisanie bardziej zwięzłego i czytelnego kodu. Python obsługuje zarówno programowanie proceduralne, jak i obiektowe, a także posiada cechy programowania funkcyjnego.

Python jest znany ze swojej czytelnej i przejrzystej składni, która ma na celu zapewnienie czytelności kodu. Zamiast stosowania nawiasów klamrowych do bloków kodu, Python używa wcięć. To oznacza, że odpowiednie wcięcia są niezbędne do oznaczenia bloków kodu, co wpływa na czytelność i spójność programu.

**Rysunek 15. Przykładowy kod w języku python.**

Obraz zawierający tekst

Opis wygenerowany automatycznie

**Źródło**: Opracowanie własne.

Język Python ma szerokie zastosowanie i jest wykorzystywany w wielu dziedzinach, takich jak web development, analiza danych, sztuczna inteligencja, uczenie maszynowe, automatyka, robotyka i wiele innych.

Python jest głównym narzędziem służącym do realizacji projektu. Jest on jednym z najbardziej popularnych technologii wykorzystywanych do web scrapowania[[19]](#footnote-19). Jest to spowodowane dostępem do wielu bibliotek, które znaczącą ułatwiają ten proces. Właśnie dlatego został on wybrany do stworzenia projektu.

Podsumowując, Python jest językiem programowania ogólnego przeznaczenia, który cechuje się czytelną składnią, szerokimi możliwościami zastosowania, bogatą pulą bibliotek i modułów.

#### Visual Code Studio (IDE)

Visual Studio Code (VS Code) to bezpłatne, środowisko programistyczne stworzone przez firmę Microsoft. Jest dostępne na platformy Windows, macOS i Linux. VS Code zostało zaprojektowane z myślą o zapewnieniu wygodnego i efektywnego środowiska programistycznego dla wielu języków programowania i technologii.

Visual Studio Code oferuje wiele funkcji i narzędzi, które pomagają programistom w ich codziennej pracy. Oferuje edytor z podświetlaniem składni i inteligentnym podpowiadaniem, które ułatwiają pisanie kodu. VS Code umożliwia także debugowanie kodu bezpośrednio w edytorze, co ułatwia identyfikację i naprawę błędów.

**Rysunek 16. Widok edytora kodu w Visual Code Studio.**

Obraz zawierający tekst, zrzut ekranu, monitor, elektronika

Opis wygenerowany automatycznie

**Źródło**: Opracowanie własne.

Środowisko Visual Studio Code integruje się również z systemami kontroli wersji, takimi jak Git, umożliwiając programistom monitorowanie zmian w kodzie i zarządzanie gałęziami. Dzięki wbudowanemu terminalowi, użytkownicy mogą wykonywać polecenia systemowe i uruchamiać skrypty bez potrzeby opuszczania edytora.

Jedną z najważniejszych cech VS Code jest jego elastyczność i rozszerzalność poprzez bibliotekę dostępnych rozszerzeń. Programiści mogą dostosować środowisko za pomocą rozszerzeń, takich jak wsparcie dla konkretnych frameworków, narzędzia do analizy statycznej, formatowania kodu i inne.

Dodatkowo, VS Code oferuje funkcję synchronizacji ustawień, która umożliwia przenoszenie preferencji, ustawień, rozszerzeń i motywów na różne komputery, co ułatwia zmianianie środowiska pracy.

Przedstawione cechy i właściwości spowowodały, że to właśnie IDE jakim jest Visual Code Studio zostało wykrozystane do wykoanania projektu.

#### Selenium

Selenium jest narzędziem do automatyzacji testów aplikacji internetowych, ale może być również używane do web scrapowania. Jest to otwarte narzędzie, które umożliwia interakcję z przeglądarkami internetowymi w sposób programatyczny. Selenium obsługuje wiele przeglądarek, takich jak Chrome, Firefox, Safari, Edge, co daje użytkownikom szeroki zakres możliwości w automatyzacji działań w przeglądarce.

W przypadku web scrapowania, Selenium może być używane do przechodzenia przez strony internetowe, wypełniania formularzy, klikania w elementy, pobierania danych, a także do zarządzania plikami cookies i sesjami użytkownika. Działa to poprzez symulowanie akcji użytkownika w przeglądarce, umożliwiając programiście przechwytywanie informacji i manipulowanie stroną internetową.

Główną rolą Selenium w web scrapowaniu jest możliwość interakcji z dynamicznymi stronami internetowymi. Wiele nowoczesnych stron internetowych korzysta z technologii takich jak JavaScript, które dynamicznie aktualizują zawartość strony lub generują ją na podstawie interakcji użytkownika. Tradycyjne metody web scrapowania, takie jak pobieranie źródła HTML, mogą nie uwzględniać tych dynamicznych zmian. Selenium pozwala na symulację interakcji użytkownika, dzięki czemu możliwe jest pobieranie aktualnych danych ze stron internetowych, które wykorzystują technologie oparte o JavaScript.

Dodatkowo, Selenium oferuje również funkcje takie jak wyszukiwanie elementów na stronie za pomocą selektorów CSS lub XPath, co ułatwia znajdowanie konkretnych danych do scrapowania. Możliwość zarządzania sesjami użytkownika i cookies jest również przydatna w przypadku stron, które wymagają logowania lub utrzymywania stanu sesji.

Ważnym aspektem web scrapowania za pomocą Selenium jest to, że symuluje on rzeczywiste zachowanie użytkownika w przeglądarce internetowej. Może to pomóc w uniknięciu mechanizmów obronnych stosowanych przez strony internetowe, które próbują blokować lub utrudniać dostęp do danych. Jednak warto pamiętać, że niektóre strony internetowe mogą posiadać zabezpieczenia przeciwko automatycznemu scrapowaniu, takie jak Captcha, co może utrudnić lub uniemożliwić użycie Selenium w takich przypadkach.

Podsumowując, Selenium jest narzędziem do automatyzacji testów aplikacji internetowych, które można również wykorzystać do web scrapowania. Zapewnia możliwość interakcji z dynamicznymi stronami internetowymi i symuluje zachowanie użytkownika w przeglądarce, co umożliwia pobieranie aktualnych danych i manipulowanie stroną za pomocą kodu dlatego też zostało one wybrane do wykonania projektu.

#### SQLite

SQLite to silnik baz danych SQL. Jest to oprogramowanie o otwartym kodzie źródłowym, które umożliwia przechowywanie i zarządzanie danymi w formie baz danych. Co wyróżnia SQLite spośród innych systemów zarządzania bazami danych, to fakt, że baza danych SQLite jest przechowywana jako pojedynczy plik, co sprawia, że jest łatwa w implementacji i nie wymaga konfiguracji serwera czy instalacji odrębnych komponentów.

SQLite obsługuje standardowy język zapytań SQL (Structured Query Language), co czyni go zgodnym z większością systemów bazodanowych. Pozwala na tworzenie tabel, wstawianie, aktualizację, usuwanie i wyszukiwanie danych za pomocą zapytań SQL. Baza danych SQLite obsługuje również wiele typów danych, takich jak tekst, liczby, daty, wartości logiczne, co pozwala na przechowywanie różnorodnych informacji.

Ze względu na swoje cechy, takie jak łatwość użycia, przenośność, niezawodność i efektywność, SQLite jest szeroko stosowane w różnych aplikacjach, takich jak systemy zarządzania treścią, aplikacje mobilne, systemy wbudowane, narzędzia analityczne i wiele innych. Z tychże powodów silnik ten został wybrany do stworzenia projektu.

#### Apache Airflow

Apache Airflow to otwarte narzędzie do zarządzania przepływami pracy i harmonogramów, które pomaga w przetwarzaniu danych w sposób efektywny i automatyczny. Airflow został zaprojektowany z myślą o pracy z procesami ETL (Extract, Transform, Load) oraz innymi procesami przetwarzania danych, takimi jak przetwarzanie strumieniowe, uczenie maszynowe i analiza danych.

Airflow działa na zasadzie definiowania zadań i procesów za pomocą kodu w Pythonie, a następnie harmonogramowania ich do wykonania w odpowiednim czasie. Wszystkie zadania są zapisywane w formie skryptów w języku Python, co umożliwia programistom łatwe dodawanie niestandardowych zadań do procesów.

Airflow posiada pojęcie DAG (Directed Acyclic Graph), czyli skierowanego acyklicznego grafu. DAG jest grafem skierowanym, w którym każde zadanie jest reprezentowane przez węzeł, a zależności między zadaniami są reprezentowane przez krawędzie. DAG określa kolejność wykonywania zadań i zależności między nimi.

Airflow oferuje również wbudowany interfejs użytkownika, który umożliwia wizualizację postępów pracy i monitorowanie wykonywania zadań. Interfejs użytkownika Airflow umożliwia również wykonywanie ręczne poszczególnych zadań oraz ich ponowne uruchomienie w przypadku błędów (Rysunek 16.).

Airflow jest łatwo konfigurowalny i skalowalny. Można go dostosować do różnych potrzeb i wymagań, a także zintegrować z innymi narzędziami do przetwarzania danych, takimi jak Hadoop, Spark, SQL, NoSQL i wiele innych.

Podsumowując, Apache Airflow działa na zasadzie definiowania zadań i procesów za pomocą kodu w języku Python, a następnie harmonogramowania ich do wykonania w odpowiednim czasie. Airflow korzysta z DAG-ów, aby określić kolejność wykonywania zadań i ich zależności między nimi. Użytkownik Airflow może zarządzać harmonogramem wykonywania zadań, a wbudowany interfejs użytkownika umożliwia monitorowanie postępów pracy i ręczne wykonywanie zadań.

**Rysunek 17. Widok interfejsu Apache Airflow.**

Obraz zawierający tekst

Opis wygenerowany automatycznie

**Źródło**: <https://github.com/apache/airflow>, [06-05-2023]

#### Power BI

Power BI to program do analizy i wizualizacji danych opracowany przez firmę Microsoft. Jest to narzędzie, które umożliwia użytkownikom biznesowym w łatwy i intuicyjny sposób tworzenie interaktywnych raportów, dashboardów oraz analizy danych.

Power BI składa się z kilku elementów, w tym:

* Power BI Desktop - jest to aplikacja desktopowa, która pozwala na tworzenie raportów i wizualizacji. Użytkownicy mogą importować dane z różnych źródeł, łączyć je i przetwarzać w celu stworzenia analiz i raportów (Rysunek 18.)
* Power BI Service - jest to platforma w chmurze, która umożliwia udostępnianie raportów i dashboardów w organizacji. Użytkownicy mogą udostępniać swoje raporty innym użytkownikom, tworzyć grupy robocze oraz kontrolować dostęp do danych.
* Power BI Mobile - jest to aplikacja mobilna, która pozwala na przeglądanie i interaktywne korzystanie z raportów i dashboardów na urządzeniach mobilnych.

**Rysunek 18. Widok raportu stworzonego w Power BI**

Obraz zawierający tekst, zrzut ekranu, Czcionka, diagram

Opis wygenerowany automatycznie

**Źródło**: Opracowanie własne.

Power BI oferuje wiele narzędzi i funkcjonalności, takich jak:

* Wizualizacje danych - Power BI oferuje wiele różnych wizualizacji, takich jak wykresy, tabele, mapy, diagramy i wiele innych. Użytkownicy mogą dostosowywać wizualizacje, aby lepiej oddawały ich potrzeby biznesowe.
* Pozyskiwanie danych – Power BI umożliwia pozyskiwanie danych z różnych źródeł takich jak pliki xlsx, json, hurtownie danych, data lake, strony internetowe , pliki csv.
* Transformacje danych - użytkownicy mogą łączyć, przetwarzać i korygować dane, aby uzyskać dokładne i spójne wyniki. Jest to możliwe również dzięki dodatkowi Power Querry.
* Automatyzacja raportów - Power BI pozwala na automatyczne generowanie raportów i wysyłanie ich do wybranych użytkowników.
* Integracja z innymi narzędziami - Power BI integruje się z innymi narzędziami firmy Microsoft, takimi jak Excel, SharePoint i Teams, co umożliwia łatwe udostępnianie danych i raportów.

Power BI to wszechstronne narzędzie, które pozwala na szybkie i skuteczne analizowanie danych oraz prezentowanie wyników w formie interaktywnych raportów i dashboardów. Jest to popularne narzędzie wśród użytkowników biznesowych i analityków danych, którzy potrzebują szybkiego i łatwego sposobu na analizowanie danych[[20]](#footnote-20). Wymieniona wcześniej lista funkcjonalności oraz doświadczenie autora w wykorzystaniu tego narzędzia zadecydowało, że to właśnie Power BI został wybrany do wizualizacji pozyskanych danych w opracowanym projekcie.

# Implementacja

Poniższy rozdział skupia się na implementacji projektu. Zostanie w nim przedstawiony i opisany każdy jego etap poprzez prezentacje stworzonych rozwiązań przy użyciu wcześniej opisanych narzędzi.

## Pozyskanie danych

Proces pozyskania danych o ofertach mieszkań w warszawie z serwisu otodom.pl składa się z podprocesów, które zostaną poniżej opisane. Wynika to między innymi ze struktury serwisu, który charakteryzuję się paginacją stron z ofertami oraz koniecznością użycia odpowiednich filtrów.

Pierwszym etapem jest pozyskanie „startowego” adresu url, który pozwala na zawężenie ofert tylko do takich, które zostaną są celem analizy. Jest to możliwe dzięki panelowi na stronie głównej, który został przedstawiony we wcześniejszym etapie pracy. W celu pozyskania tegoż adresu, zostają wprowadzone konieczne parametry takie jak rodzaj nieruchomości („mieszkania”), rodzaj oferty („na sprzedaż”) oraz lokalizacja („Warszawa”). Jest to widoczne na poniższej grafice.

**Rysunek 19. Panel wyszukiwarki z uzupełnionymi parametrami w serwisie otodom.pl**

Obraz zawierający tekst, zrzut ekranu, Ludzka twarz, ubrania

Opis wygenerowany automatycznie

**Źródło**: <https://www.otodom.pl/>, [20-05-2023].

Po zatwierdzeniu parametrów, serwis przenosi użytkownika do pierwszej z wielu podstron z ofertami, której adres url jest wcześniej wymienionym „startowym”. Jak widać na poniższej grafice, otodom.pl stosuję technikę paginacji stron. Jest to istotna kwestia, która będzie uwzględniana w dalszych pracach.

**Rysunek 20. Widok serwisu otodom.pl po zatwierdzeniu wymaganych parametrów w panelu wyszukiwania na stronie głównej.**

Obraz zawierający tekst, zrzut ekranu, Strona internetowa, oprogramowanie

Opis wygenerowany automatycznie

**Źródło**:<https://www.otodom.pl/pl/oferty/sprzedaz/mieszkanie/warszawa?distanceRadius=0&locations=%5Bcities_6-26%5D&viewType=listing> , [20-05-2023].

Kolejnym etapem procesu, który jest konieczny aby skorzystać z pozyskanego adresu url jest stworzenie obiektu w języku Python, który będzie pełnił rolę przeglądarki. Jest to możliwe dzięki Selenium. Poniżej znajduję się grafika z kodem, który właśnie to wykonuje.

**Rysunek 21. Tworzenie obiektu przęglądarki.**

Obraz zawierający tekst, zrzut ekranu, Czcionka

Opis wygenerowany automatycznie

**Źródło**: Opracowanie własne.

Ten kod definiuje funkcję „configure\_driver()”, która konfiguruje sterownik przeglądarki Chrome. Funkcja tworzy obiekt „chrome\_options” reprezentujący opcje konfiguracyjne dla przeglądarki. Następnie dodaje trzy argumenty do chrome\_options: „--headless” (tryb bezinterakcyjny), „--no-sandbox” (tryb bezpieczny) i „--disable-dev-shm-usage” (mniejsze użycie pamięci dzielonej).

Kolejnym krokiem jest utworzenie obiektu „driver” klasy „webdriver.Chrome”, który reprezentuje sterownik przeglądarki Chrome. Sterownik ten jest skonfigurowany z użyciem ścieżki do pliku wykonywalnego „chromedriver.exe” oraz opcji „chrome\_options”.

Na koniec, zmienna „driver” otrzymuje wartość zwróconą przez funkcję „configure\_driver()”, co oznacza, że jest gotowa do użycia jako sterownik do przeprowadzania działań w przeglądarce Chrome.

Aby pozyskać adresy url z każdej z ofert, należy pozyskać informację o ich liczbie. Jest to możliwe dzięki temu, że jak wcześniej zostało wspomniane, strony te objęte są paginacją. Poniższy kod przedstawia pozyskanie informacji o liczbie stron z ofertami.

**Rysunek 22. Pozyskanie informacji o liczbie stron z ofertami.**

Obraz zawierający tekst, zrzut ekranu, Czcionka

Opis wygenerowany automatycznie

**Źródło**: Opracowanie własne.

Powyższy kod, najpierw otwiera przeglądarkę i wczytuje stronę o podanym „linku startowym”. Następnie przewija stronę do końca za pomocą skryptu JavaScript, aby załadować całą treść. Tworzy obiekt BeautifulSoup, który analizuje źródło strony, przechowywane w „driver.page\_source”. To pozwala na ekstrakcję danych ze strony. Potem znajduje wszystkie elementy na stronie, które mają określoną klasę, i zapisuje je w liście „liczby stron”. Następnie pobiera tekst z czwartego elementu z listy „liczby\_stron” i przypisuje go do zmiennej „liczba\_stron” w której zawarta jest liczba wszystkich stron z ofertami.

Zakończenie

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur.

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur.

LITERATURA

**Książki**

1. Bradley K., *Podstawy metodyki PRINCE 2*, Warszawa 2002.
2. Bukowski M., *Korzyści z wdrożenia metodyki PRINCE2 w wybranej jednostce administracji publicznej*, op. naukowy dr inż. W. Dąbrowski, Warszawa, czerwiec 20.
3. Flasiński M., *Zarządzanie projektami informatycznymi,* Wydawnictwo Naukowe PWN, Warszawa 2006.
4. Frączkowski K., *Zarządzanie projektem informatycznym. Projekty w środowisku wirtualnym. Czynniki sukcesu i niepowodzeń projektów*, Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław 2003, s.125-126.
5. Gunia G., *Implementacja zintegrowanych systemów informatycznych w małych i średnich przedsiębiorstwach*, „Zarządzanie Przedsiębiorstwem”, 2009.
6. Gunia G., *Wdrażanie zintegrowanych systemów informatycznych,* Wydawnictwo Fundacji Centrum Nowych Technologii, Bielsko Biała 2009.
7. Januszewski A., *Funkcjonalność informatycznych systemów zarządzania. Tom 1*, Wydawnictwo Naukowe PWN, Warszawa 2008.
8. Kisielnicki J., *Zintegrowane systemy informatyczne, dobre praktyki wdrożeń systemów klasy ERP*, Wydawnictwo Naukowe PWN
9. Kisielnicki J., *Systemy informacyjne biznesu,* Placet, Warszawa 1999.
10. Koszlajda A.*, Zarządzanie projektami IT*, Gliwice 2010.
11. Lech P., Zintegrowane Systemy Zarządzania ERP/ERP2. *Wykorzystanie w biznesie wdrażanie,* Difin, Warszawa 2003.
12. Miłosz M. (red.), *Wdrażanie i eksploatacja systemów informatycznych. Wybrane problemy,* Polskie Towarzystwo Informatyczne, Lublin 2002.
13. Office of Government Commerce, *Skuteczne zarządzanie projektami PRINCE2*, London: TSO, wydanie 2005.
14. Rokicka-Broniatowska A., *Wstęp do informatyki gospodarczej*, Szkoła Główna Handlowa, Warszawa 2006.
15. Stefanowicz B., *Informacyjne systemy zarządzania. Przewodnik,* Szkoła Główna Handlowa, Warszawa 2007.
16. Szyjewski Z., *Metodyki zarządzania projektami informatycznymi,* PLACET, Warszawa 2004.
17. Trocki M., *Metodyki zarządzania projektami*, Biblioteka Project Managera, Warszawa 2011.
18. Waćkowski K., Chmielewski J., *Wspomaganie zarządzania projektami informatycznymi. Poradnik dla menadżera*, Helion, Gliwice 2007.

**Artykuły i studia**

1. Dębowski L.,*ZASTOSOWANIE KOMPUTERÓW W NAUCE I TECHNICE’ 2006*, Zeszyty Naukowe Wydziału Elektrotechniki i Automatyki Politechniki Gdańskiej Nr 22
2. IE. 1991. Competition in manufacturing leads to MRP II. 23 (July) 10-13
3. WJ Hopp, ML Spearman Commissioned Paper To Pull or Not to Pull: What Is the Question, Manufacturing & Service Operations Management, 2004

**Strony internetowe**

1. Poksiński P., *PRINCE2*, materiał dostępny w wersji elektronicznej pod adresem internetowym: http://www.poksinski.com/pdf\_files/PRINCE2.pdf [18-09-2012]
2. *Słownik APICS*, http://www.apics.org/dictionary/dictionary-information?ID=2399 [4-01-2013]
3. Schmidt P., *PRINCE2 i techniki planowania*, materiał dostępny w wersji elektronicznej pod adresem internetowym: http://ww.4pm.pl/artykul/prince2\_i\_techniki\_planowania-56-1247.html [17-10-2012]
4. *Cloud Computing*, materiał zaczerpnięty ze strony internetowej: http://www.it.integro.pl/pl/aplikacje-w-chmurze-cloud-computing [4-01-2013]
5. Trąbka J., *Zarządzanie projektem wdrożeniowym systemu klasy ERP – autorska metodyka*, materiał dostępny w wersji elektronicznej pod adresem internetowym: http://kkio2012.agh.edu.pl/presentations/KKIO2012-C1\_3.pdf [10-11-2012]
6. *Skuteczne wdrożenia systemu ERP dzięki metodyce Microsoft Sure Step;* http://www.it.integro.pl/pl/metodyka-wdrozeniowa. [28-09-2012]
7. *Nasza metodyka na wdrożenie ERP w twojej firmie;* http://www.profidata.com.pl/erp/metodyka-pit-stop.html. [28-09-2012]
8. *Manifesto for Agile Software Development*; Cytat ze strony http://www.agilemanifesto.org/ [10-10-2012]
9. *Metodyka wdrożenia QlikView- Business Intelligence;* , http://www.businessintelligence.pl/pl/metodyka-wdrozenia. [28-09-2012]
10. *QlikView przełamuje bariery rynku – Business Intelligence;* http://www.businessintelligence.pl/pl/a/QlikView-przelamuje-bariery-rynku [11-10-2012]
11. *ERP- centrum wiedzy o systemach ERP na decyzje-IT.pl;* http://decyzje-it.pl/centrum-wiedzy/erp.html [10-10-2012]
12. *The history of Prince2- Project Smart*; http://www.projectsmart.co.uk/history-of-prince2.html [18-09-2012]
13. *Oracle- Hardware and Software, Enginnered to work together*; http://www.oracle.com/index.html [23-09-2012]
14. CRP for Oracle R12- Welcom to Oracle ERP http://www.oracleerp4u.com/2010/06/crp-for-oracle-r12.html [189-10-2012]
15. *Przykłady zastosowań technologii RFID w magazynach* http://rfid-lab.pl/przyklady-zastosowan-technologii-rfid-w-magazynach [07-01-2013]

Spis tabel

[Tabela 1. Lorem ipsum dolor sit amet **Błąd! Nie zdefiniowano zakładki.**](#_Toc355852181)

[Tabela 2. Lorem ipsum dolor sit amet, consectetur adipisicing elit **Błąd! Nie zdefiniowano zakładki.**](#_Toc355852182)

Spis rysunków

[Rysunek 1. Lorem ipsum dolor sit amet Lorem ipsum dolor sit amet **Błąd! Nie zdefiniowano zakładki.**](#_Toc355852186)

[Rysunek 2. Lorem ipsum dolor sit amet, consectetur **Błąd! Nie zdefiniowano zakładki.**](#_Toc355852187)

Streszczenie

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Summary

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

1. What is ETL? <https://www.ibm.com/topics/etl> [19-03-2023]. [↑](#footnote-ref-1)
2. Tamże. [↑](#footnote-ref-2)
3. Tamże. [↑](#footnote-ref-3)
4. Tamże. [↑](#footnote-ref-4)
5. What is ETL (extract transform load)? <https://www.informatica.com/resources/articles/what-is-etl.html> [19-03-2023]. [↑](#footnote-ref-5)
6. R. Mitchell, *Web Scraping with Python*, O’Reilly, Sebastopol 2015, str. 9. [↑](#footnote-ref-6)
7. R. Mitchell, *Web Scraping with Python*, O’Reilly, Sebastopol 2015, str. 23. [↑](#footnote-ref-7)
8. R. Mitchell, *Web Scraping with Python*, O’Reilly, Sebastopol 2015, str. 24. [↑](#footnote-ref-8)
9. C. O. Wilke, *Podstawy wizualizacji danych. Zasady tworzenia atrakcyjnych wykresów*, Helion, Gliwice 2015, str. 16. [↑](#footnote-ref-9)
10. C. O. Wilke, *Podstawy wizualizacji danych. Zasady tworzenia atrakcyjnych wykresów*, Helion, Gliwice 2015, str. 28. [↑](#footnote-ref-10)
11. Tamże. [↑](#footnote-ref-11)
12. J. Steele, N. Iliinsky, *Beautiful Visualization: Looking at Data through the Eyes of Experts (Theory in Practice)*, O'Reilly, 2010, str. 42. [↑](#footnote-ref-12)
13. J. Steele, N. Iliinsky, *Beautiful Visualization: Looking at Data through the Eyes of Experts (Theory in Practice)*, O'Reilly, 2010, str. 28. [↑](#footnote-ref-13)
14. C. Nussbaumer Knafic, *Storytelling with Data*, John Wiley & Sons, 2015, str. 62. [↑](#footnote-ref-14)
15. The Best Data Visualization Tools Of 2023, <https://www.forbes.com/advisor/business/software/best-data-visualization-tools/>, [07-04-2023]. [↑](#footnote-ref-15)
16. Ranking portali sprzedających nieruchomości, <https://ceo.com.pl/ranking-portali-sprzedajacych-nieruchomosci-53865#OTODOM>, [02-05-2023]. [↑](#footnote-ref-16)
17. Tamże. [↑](#footnote-ref-17)
18. Tamże. [↑](#footnote-ref-18)
19. *Best Programming Languages for Effective Web Scraping*, <https://oxylabs.io/blog/best-web-scraping-language>, [05-05-2023] [↑](#footnote-ref-19)
20. *Power BI Vs Tableau: Difference and Comparison*, <https://www.simplilearn.com/tutorials/power-bi-tutorial/power-bi-vs-tableau#features_of_power_bi>, [14-05-2023]. [↑](#footnote-ref-20)