Hurtownie danych

Projekt zaliczeniowy

Sebastian Ptasznik

informatyka stosowana,

studia niestacjonarne magisterskie rok 1 grupa 1

Celem projektu było stworzenia web scrappera wybranej strony internetowej oraz wykonanie analizy zebranych danych. W projekcie dokonano analizy danych dotyczących mieszkań pochodzących z rynku wtórnego na terenie Krakowa pochodzących ze strony OtoDom. Do budowy scrappera użyto języka JavaScript a do analizy danych języka R.

1. Web Scraper

   Web Scraping to technika wyodrębniania danych ze stron internetowych, która zastępuje ręczne, powtarzalne wpisywanie lub kopiowanie i wklejanie. Technika ta pozwala zautomatyzować proces wyodrębniania informacji ze stron internetowych. Zwykle polega na wysyłaniu zapytań do serwera witryny internetowej, pobraniu znaczników HTML strony internetowej, a następnie analizowaniu wybranego fragmentu HTML w celu wyodrębnienia danych, którymi interesuje się użytkownik. Przewagą tego rozwiązania jest również fakt, że pobierane dane zapisywane są w wybranej przez nas ustrukturyzowanej formie.

   Do budowy scrapera wykorzystano środowisko Node.js. W projekcie użyto bibliotekę Puppeteer, która została opracowana przez Google i pozwala kontrolować bezgłową przeglądarkę Chrome za pośrednictwem protokołu DevTools.Bezgłowa oznacza przeglądarkę internetową bez interfejsu użytkownika. Przeglądarki bezgłowe to te, które faktycznie uzyskują dostęp do strony internetowej, ale GUI jest ukryte przed użytkownikiem. Puppeteer używa protokołu WebDriver do łączenia się z przeglądarką i symulowania interakcji użytkownika ze stronami HTML.

Opis programu

W tej części zostaną przedstawione najważniejsze funkcje programu.

```
const scrape = async () => {
    const browser = await puppeteer.launch( options: { headless: true });
    const page = await browser.newPage();

    await page.goto(url);

    let results = [];
    const  lastPageNumber = getNumberOfPages(page);

    for (let index = 0; index < lastPageNumber-1; index++) {
        await  waitForNetwork0(page);

        results =[...results, ...await resolveData(page)];

        if (index !== lastPageNumber - 1) {
            await Promise.all( values: [
                page.waitForNavigation( options: {timeout:4000}),
                page.click( selector: '[data-cy="pagination.next-page"]'),

            ])


        }
    }

    await browser.close();
    return results;
};
```

Na powyższym zrzucie ekranu widoczna jest główną funkcję programu. Na początku inicjalizowany jest obiekt przeglądarki, na którym następnie wykonywana jest metoda, która zwraca nową kartę. Następnie za pomocą metody ‘goTo’ zostaje przekazany adres URL do wcześniej otwartej karty. W kolejnym etapie inicjalizujemy pustą tablicę, w której będą przetrzymywane wyniki działania programu oraz zostanie wykonana funkcja ‘getNumberOfPage’ która zwraca liczbę podstron witryny.

Następnie została zadeklarowana główna pętla programu, która iteruje po podstronach witryny. Na początku każdego przejścia pętli wykonuje się funkcja ‘waitForNetwork’ która czeka zadaną liczbę sekund. W kolejnym kroku wykonana jest funkcja “resolveData’ odpowiedzialna za pobieranie danych ze storny, a jej wynik zapisywane jest do tablicy ‘result’. Następnie sprawdzany jest warunek, czy są jeszcze podstrony do odwiedzenia, jeżeli warunek jest spełniony zostaje wykonana metoda ‘waitForNavigation’ oraz metoda ‘click’ obiektu page odpowiedzialna za przejście na kolejną podstronę. Jeżeli warunek nie został spełniony,to znaczy wszystkie strony zostały odwiedzone przeglądarka zostaje zamknięta oraz funkcja zwraca tablicę wyników.

```
async function resolveData(page) {
    await autoScroll(page)
    return page.evaluate(() => {
        let data = [];
        const items = document.querySelectorAll( selectors: '.css-153eqh1.e1brl80i2');

        items.forEach( callbackfn: item => {
            if(item){
                const element = item.querySelectorAll( selectors: ".css-s8wpzb.e1brl80i1");

                if(element){
                    const apartment = {
                        price:Number(element[0].textContent.replace( searchValue: /[^0-9\.-]+/g, replaceValue: "")),
                        numberOfRooms: (element[2].textContent.split( separator: " ").shift()),
                        area:(element[3].textContent.split( separator: " ").shift()),
                    }
                    data.push(apartment)
                }
            }

        });
        return data;
    });
}
```

Funkcja resolveData to funkcja odpowiedzialna za wydobycie danych ze strony. Jako argument przyjmuje obiekt ‘page’ na którym wykonuje metodę ‘evaluate’. Metoda ‘evaulate’ wykonuje podaną funkcję w kontekście strony, dlatego w jej ciele mamy dostęp do obiektu ‘docuemnt’. Na początku funkcja wykonuje funkcję ‘autoScroll’. Za pomocą metody ‘querySelectorAll’ pobieramy ze storny wszystkie elementy które zawierają w sobie wybrane informacje na temat mieszkania. Następnie w pętli ‘forEach’ dla każdego elementu dane zostają pobrane, sformatowane i zapisane do tablicy. Po zakończeniu pętli funkcja zwraca wypełnioną tablicę danych.

Funkcja autoScroll.

```
async function autoScroll(page){
    await page.evaluate(async () => {
        await new Promise( executor: (resolve) => {
            let totalHeight = 0;
            const distance = 100;

            const timer = setInterval( handler: () => {
                const scrollHeight = document.body.scrollHeight;

                window.scrollBy( x: 0, distance);

                totalHeight += distance;

                if(totalHeight >= scrollHeight - window.innerHeight){
                    clearInterval(timer);
                    resolve();
                }
            }, timeout: 10);
        });
    });
}
```

Funkcja odpowiedzialna za przewijanie strony. W badanej witrynie zastosowano technikę optymalizacji ‘lazyLoading’, dlatego podczas początkowego załadowania strony mamy dostęp tylko do części informacji. Przewijając stronę na dół symulujemy zachowanie użytkownika, dzięki czemu informacje są doładowywane do pustych elementów.

Analiza zebranych danych

W celu analizy danych wykorzystano program RStudio. Przeanalizowano 3861 obserwacji. W zbiorze danych znajdują się cztery zmienne ilościowe:

- Price.zł. - cena mieszkania
- NumberOfRooms - liczba pokoi
- Area.m.2. - powierzchnia mieszkania w metrwach kwadratowych

Na początku utworzono nową kolumnę ‘priceSQM’, w której znajduję się cena za metr kwadratowy mieszkania. Następnie przystąpiono do oczyszczenia danych: usunięcia obserwacji odstających, usunięcia wierszy w których brakuje jakiejś wartości oraz sprawdzeniu, czy wszystkie dany mają poprawny format.
Zauważono, że zmienna opisująca liczbę pokoi została zapisana w formacie znaku. Zmieniono jest typ na liczbę. Następnie usunięto obserwacje odstające za pomocą reguły 1.5 wartości rozstępu międzykwartylowego.
U jej podstaw leży założenie, że wszystkie "typowe" obserwacje leżą pomiędzy punktami wyznaczonymi przez odległość 1.5 IQR (ang. *interquartile range*):

- "na lewo" od granicy pomiędzy pierwszym i drugim kwartylem,
- "na prawo" od granicy pomiędzy trzecim i czwartym kwartylem.

Statysty

Price.zł.	NumberOfRooms	Area.m.2.	priceSQM
Min. : 115000	Min. :1.000	Min. :10.50	Min. : 4933
1st Qu.: 449000	1st Qu.:2.000	1st Qu.:38.50	1st Qu.:10000
Median : 550000	Median :2.000	Median :49.30	Median :11486
Mean : 581981	Mean :2.359	Mean :49.81	Mean :11829
3rd Qu.: 699000	3rd Qu.:3.000	3rd Qu.:59.90	3rd Qu.:13199
Max. :1180000	Max. :6.000	Max. :93.40	Max. :19588

ki opisowe:

- Cena

Średnia cena mieszkania z drugiej ręki w Krakowie wynosi 581 000 zł. Minimalna cena to 115 000 zł a maksymalna 1 180 000 zł. Mediana wyniosła 550 000 zł oznacza to, *że* połowa badanych mieszkań kosztuje nie więcej niż 550 000 zł a połowa nie mniej niż 550 000 zł. Pierwszy kwartyl wynosi 449 000 zł oznaczo to, że 25% badanych mieszkań ma cenę niższa bądź równa 449 000 zł, a 75% badanych mieszkań ma cenę równą bądź większa niż 449 000 zł. Trzeci kwartyl wyniósł 699 000 zł oznacza to że 75% badanych mieszkań ma cenę niższa bądź równą 699 000 zł, a 25% badanych mieszkań ma cenę równą bądź większa niż 699 000 zł.

- Liczba pokoi

  Minimalna liczba pokoi wyniosła 1 a maksymalna 6 Mediana wyniosła 2 oznacza to, *że* połowa badanych mieszkań posiada nie więcej niż 2 pokoje a połowa nie mniej niż 2 pokoje. Pierwszy kwartyl wynosi 2 oznaczo to, że u 25% badanych mieszkań liczba pokoi jest niższa bądź równa 2, a w przypadku 75% badanych mieszkań liczba pokoi jest równą bądź większa niż 2. Trzeci kwartyl wyniósł 3 oznacza to że u 75% badanych mieszkań liczba pokoi jest niższa bądź równa 3 , a w przypadku 25% badanych mieszkań liczba pokoi jest równą bądź większa niż 3.

- Powierzchnia

  Średnia powierzchnia mieszkania z drugiej ręki w Krakowie wynosi 49.81m^2. Minimalna powierzchnia 10.5m^2 a maksymalna 93.4m^2. Mediana wyniosła 49.3m^2 oznacza to, *że* połowa badanych mieszkań ma powierzchnię nie większą niż 49.3m^2 a połowa nie mniejszą niż 49.3m^2. Pierwszy kwartyl wynosi 38.5m^2 oznaczo to, że 25% badanych mieszkań ma powierzchnię niższą bądź równa 38.5m^2 , a 75% badanych mieszkań ma powierzchnię równą bądź większa niż 38.5m^2. Trzeci kwartyl wyniósł 59.9m^2 oznacza to że 75% badanych mieszkań ma powierzchnię niższa bądź równą 59.9m^2, a 25% badanych mieszkań ma powierzchnię równą bądź większa niż 59.9m^2.

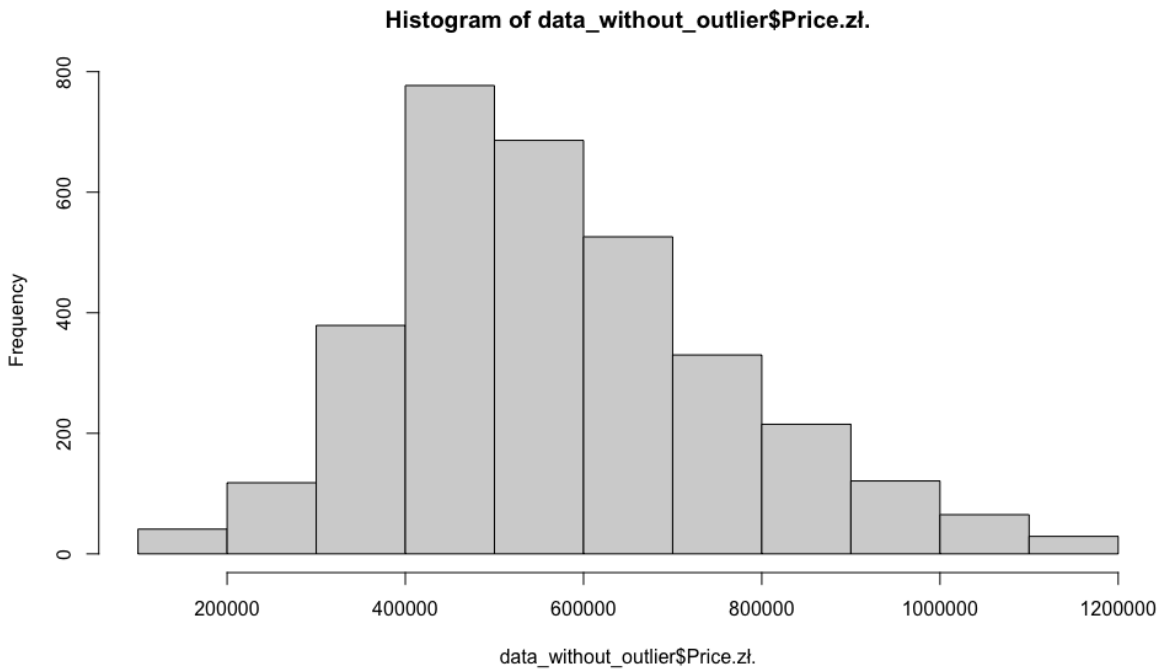
- Cena za metr kwadratowy

  Średnia cena metra kwadratowego wynosi 11 829 zł. Minimalna cena to 4933 zł a maksymalna 19588 zł. Mediana wyniosła 11 486 zł oznacza to, *że* połowa badanych mieszkań kosztuje nie więcej niż 11 486 zł za metr kwadratowy a połowa nie mniej niż 11 486 zł za metr kwadratowy. Pierwszy kwartyl wynosi 10 000 zł oznaczo to, że 25% badanych mieszkań ma cenę niższa bądź równa 10 000 zł za metr kwadratowy, a 75% badanych mieszkań ma cenę równą bądź większa niż 10 000 zł za metr kwadratowy. Trzeci kwartyl wyniósł 13 199 zł oznacza to że 75% badanych mieszkań ma cenę niższa bądź równą 13 199 zł za metr kwadratowy a 25%

badanych mieszkań ma cenę równą bądź większa niż 13 199 zł za metr kwadratowy

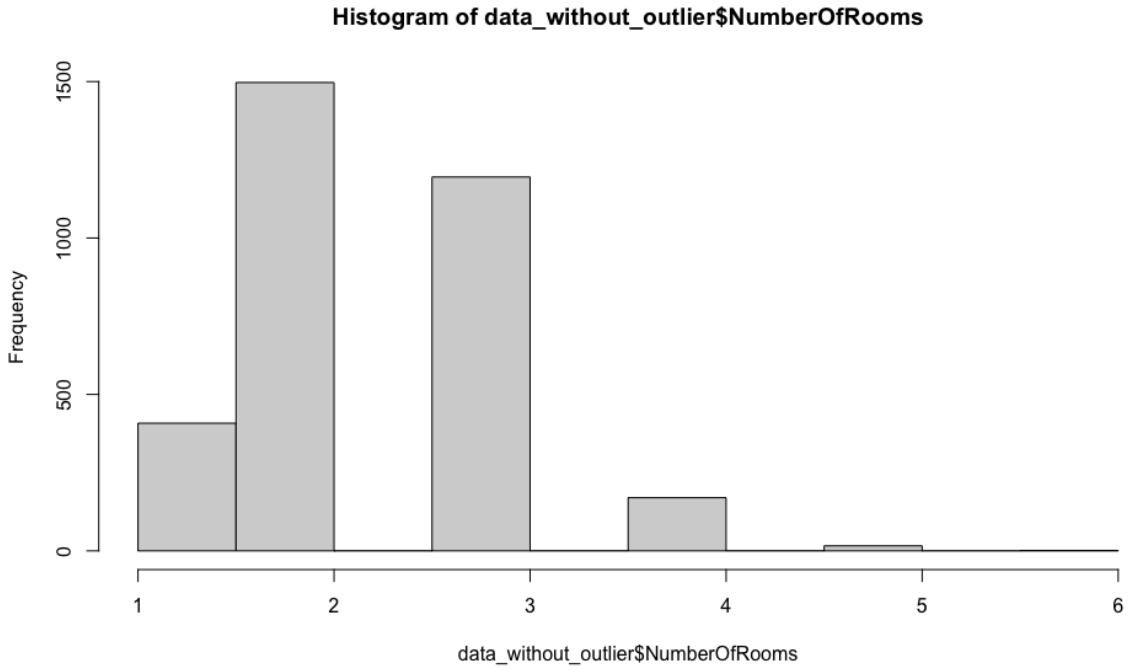
Histogram:

- Cena mieszkania



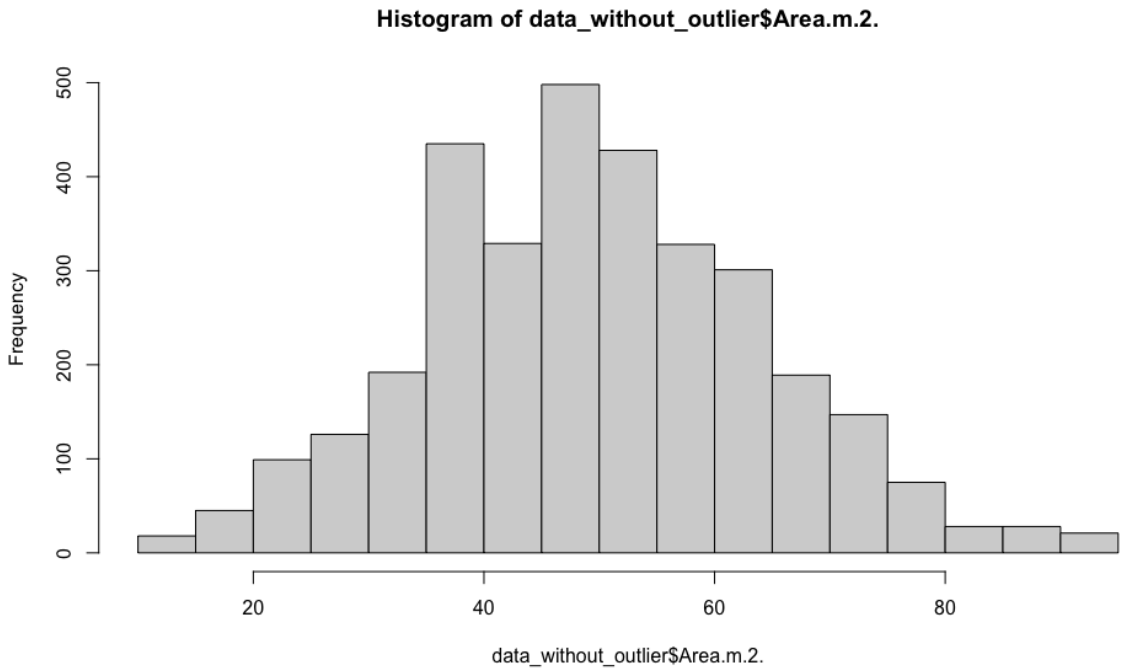
Dane rozłożone są w miarę symetrycznie wokół średniej.

- Liczba pokoi



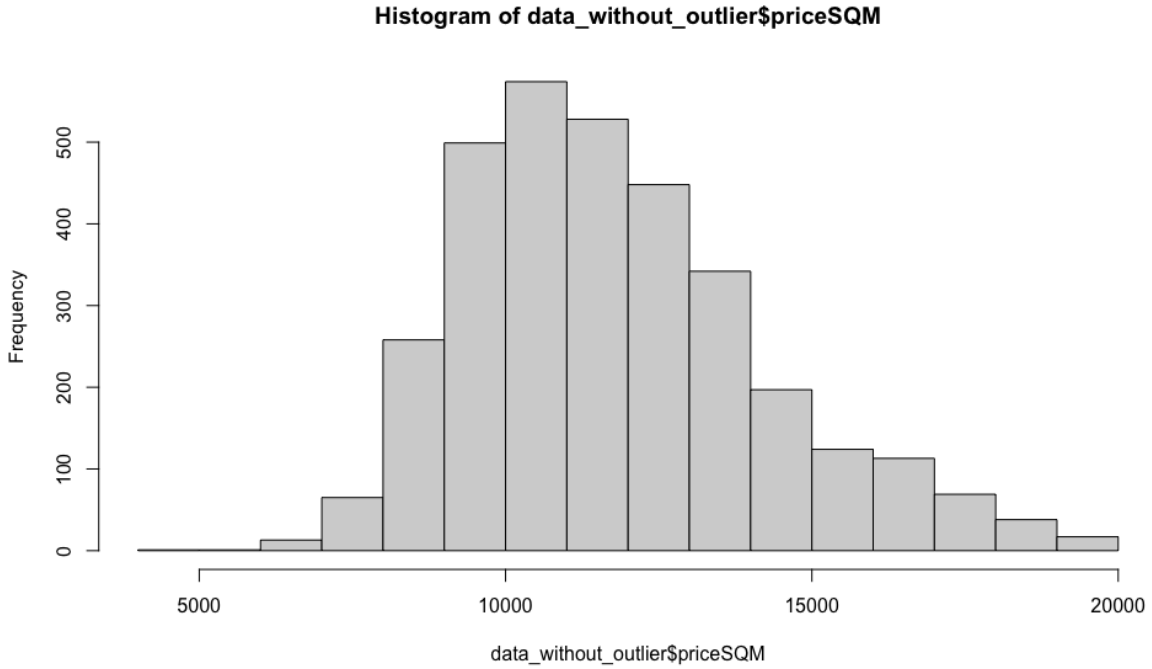
Na podstawie powyższego histogramu możemy stwierdzić, że w badanym zbiorze najwięcej występuje mieszkań z liczbą pokoi równą 2 i 3.

- Powierzchnia mieszkania



Obserwacje skoncentrowane są wokół średniej.

- Cena za metr kwadratowy



Histogram przypomina histogram danych dotyczących cen mieszkań, co jest zrozumiałe, gdyż cena za metr kwadratowy jest silnie skorelowana z całkowitą ceną mieszkania.

Podsumowanie

Rynek mieszkaniowy w Krakowie może się różnić w zależności od czynników takich jak lokalizacja, rodzaj nieruchomości i ogólna sytuacja ekonomiczna. Cena mieszkania uzależniona jest przede wszystkim od metrażu, liczby pokoi a przede wszystkim lokalizacji. W przedstawionej analizie najtańsze mieszkanie kosztuje 115 000zł ma 1 pokój i powierzchnię 12.5 metra kwadratowego