Стандардна семинарска работа по предметот Вовед во науката за податоци

Опис на проектот:

Да се соберат податоци за цените на еден ист продукт од продавница налик ananas и да се направи агрегација на продукти кои се исти со цел споредба на цена

Наслов и тема на проектот:

Phone Price Comparison and Aggregation from Setec and Tehnomarket

Линк до кодот :

Pepi134/VNP (github.com)

Линк до видеото:

https://youtu.be/GMsENWLQxGs

Изработи:

Петрина Станковска, СИИС-201073

Одговорен:

асс. Ана Тодоровска

Вовед

1.Цел на проектот

Во денешното динамично опкружување, цените на производите и услугите постојано се менуваат, а потрошувачите се соочуваат со предизвикот на правење информирани одлуки при купување. Споредбата на цени од различни продавници станува клучен аспект на рационалното управување со личните финансии и оптимизирање на потрошувачките навики. Важноста на оваа споредба лежи во можноста за откривање на најдобри понуди, идентификување на трендови и воспоставување на стратегии за заштеда на трошоци.

Проектот "Phone Price Comparison and Aggregation from Setec and Tehnomarket" има за цел да истражи и анализира ценовните разлики на мобилни телефони помеѓу две водечки продавници во Македонија: Setec и Tehnomarket. Се стреми да обезбеди детално разбирање на разликите во цените на исти или слични производи помеѓу овие две продавници. Оваа анализа ќе овозможи потрошувачите да направат поинформирани избори при купување на мобилни телефони и да го максимизираат своето финансиско планирање.

Во рамките на проектот, ќе се користат техники за веб-скрејпинг за автоматско собирање на податоци од веб-страниците на Setec и Tehnomarket. Со помош на напредни алатки за анализа на податоци, ќе се обработат и упатат податоците за да се идентификуваат и агрегираат слични продукти, а потоа ќе се направи споредба на цените. Овој пристап ќе овозможи создавање на прегледни и компаративни извештаи за ценовните разлики, кои ќе бидат корисни за потрошувачите кои сакаат да направат најдобри избори за своите купувачки потреби.

Целта на проектот е да ги осветли тековните трендови за образување на цени, да ги открие можностите за заштеда и да ги информира потрошувачите за ценовната конкурентност помеѓу двете значајни продавници. Оваа информација ќе биде корисна не само за потрошувачите, туку и за продавниците кои можат да ги користат резултатите за прилагодување на своите стратегии за ценообразување и маркетинг.

2.Преглед на методологија

Методологијата на проектот "Phone Price Comparison and Aggregation from Setec and Tehnomarket" е конципирана со цел да обезбеди систематски и точен пристап за собирање, обработка и анализа на податоци за цените на мобилни телефони од двете продавници. Овој дел ќе го опфати краткиот преглед на чекорите кои ќе се применат за успешно завршување на проектот, од собираењето на податоците до финалната споредба и анализа.

1. **Дефинирање на целите и објектите на анализата**:
    - Првиот чекор во методологијата е дефинирање на конкретните цели на проектот и идентификација на објектите на анализа. Во овој случај, целта е споредба на цените на мобилни телефони помеѓу Setec и Tehnomarket. Ова вклучува одредување на кои мобилни телефони ќе се анализираат, како и дефинирање на критериумите за споредба.
2. **Собирање на податоци преку веб-скрејпинг**:
    - За да се добијат податоци за цените на мобилни телефони од Setec и Tehnomarket, ќе се користат техники на веб-скрејпинг. Овој метод вклучува автоматско собирање на информации од веб-страниците на продавниците преку специфични Python библиотеки како BeautifulSoup и Scrapy. Скрејпинг процесот ќе се фокусира на собирање на информации како што се името на производот, URL на страницата на производот и цената.
3. **Чистење и обработка на податоци**:
    - Собраните податоци често содржат некомплетни или неконзистентни информации. Следниот чекор е чистење на податоците, што вклучува идентификација и отстранување на дупликатите, исправка на грешки и унифицирање на форматите на податоците. Оваа фаза исто така вклучува конверзија на податоците во структурирани форми погодни за анализа.
4. **Агрегација и мапирање на производи**:
    - Потоа, ќе се спроведе агрегација на податоците за да се идентификуваат слични или идентични производи меѓу Setec и Tehnomarket. За оваа цел ќе се користат техники на fuzzy matching, кои овозможуваат споредба на имиња на производи со различни варијации и формати. Овој процес ќе овозможи комбинирање на цените за истите производи од различни продавници.
5. **Анализа и споредба на цените**:
    - Откако податоците ќе бидат агрегирани, следи анализа на ценовните разлики. Овој чекор вклучува споредба на цените на слични или идентични производи и изработка на статистички анализи кои ќе ги откријат значајните разлики и трендови. Анализата ќе се претстави преку графички и табеларни прикажувања.

Овој систематски пристап за собирање, обработка и анализа на податоци ќе овозможи детална и точна споредба на цените на мобилни телефони помеѓу Setec и Tehnomarket, и ќе обезбеди вредни информации кои можат да помогнат на потрошувачите при донесување на информирани купувачки одлуки.

Опис на проблемот

1. Преглед на избраните продавници

Во овој проект, Setec и Tehnomarket се избрани како две од најпопуларните продавници за електроника и технолошки производи во Македонија. Овие продавници играат клучна улога во пазарот на мобилни телефони и нудат широка палета на производи, што ги прави идеални кандидати за споредба на цените. Описот на овие продавници ќе помогне да се разбере нивната пазарна позиција, понудата на производи, како и важноста на споредбата на цените меѓу нив.

Значење на споредбата на цените

Изборот на Setec и Tehnomarket како субјекти на овој проект е логичен поради нивната голема популарност и важност на македонскиот пазар за електроника. И двете продавници нудат слични производи, но разликите во цените и промоциите може да бидат значајни. Споредбата на цените помеѓу овие два ритеилери ќе помогне да се идентификуваат најдобрите понуди и да се разберат факторите кои влијаат на ценовните разлики, што е од огромна важност за информирано донесување одлуки од страна на потрошувачите.

2.Предизвици во споредба на цени

Споредбата на цени помеѓу различни продавници, како што се Setec и Tehnomarket, претставува сложен процес поради различните начини на кои продавниците ги именуваат и форматираат производите во своите онлајн каталози. Овие разлики во насловите и форматите на продуктите создаваат сериозни предизвици при обидот за прецизна споредба на истите производи, што може да доведе до неточни заклучоци ако не се адресираат соодветно.

Разлики во насловите на производите

Еден од главните предизвици е различното именување на производите во различните продавници. И покрај тоа што се работи за ист производ, насловите често се разликуваат во зависност од продавницата. На пример, еден мобилен телефон може да биде именуван како „iPhone 13 Pro Max 256GB Blue“ во една продавница, додека во друга може да се појави како „Apple iPhone 13 Pro Max 256 GB (Сина)“. Овие суптилни разлики во насловите може да го направат идентификувањето на истите производи тешко, особено при автоматизирано споредување на податоците.

Некои од овие разлики можат да вклучуваат:

- Употреба на различни формати за капацитет на меморија („256GB“ наспроти „256 GB“).
- Присуство или отсуство на информации за боја или модел.
- Употреба на различни јазици или симболи (на пример, „Blue“ наспроти „Сина“).
- Разлики во начинот на кои се наведуваат брендови и модели (на пр. „Apple iPhone“ наспроти само „iPhone“).

Овие варијации значат дека директно споредување на насловите може да доведе до неправилно идентификување на производите како различни, иако всушност се исти. За да се надминат овие предизвици, неопходно е да се применат техники за нормализација на податоците и методи за „fuzzy matching“, кои овозможуваат прецизна споредба и агрегација на податоците од различни извори.

Разлики во форматите на податоците

Покрај разликите во насловите, форматите на податоците исто така можат да претставуваат предизвик при споредбата на цените. На пример, цените може да бидат изразени во различни формати, со или без валута, со различни разделници за илјадници и децимални точки, или со различни формати за попусти и промоции. Ова може да предизвика проблеми при обидот за споредба на цените ако не се стандардизираат сите податоци на унифициран формат.

Пример за вакви разлики може да биде:

- Цената во една продавница да биде наведена како „24,999 ден.“, додека во друга да биде „24999 MKD“.
- Присуство на попусти кои се изразени на различни начини, како процентуален попуст или како апсолутен износ.
- Разлики во форматот на датумите кога се применуваат промоциите или попустите.

Овие разлики можат да ги комплицираат пресметките и анализата, што ја прави потребна дополнителна обработка на податоците за да се обезбеди точност.

Методологија

1. Собирање податоци: Опис на веб-скрејпинг процесот

Во овој проект, целта беше да се соберат податоци за цените на мобилните телефони од веб-страниците на Setec и Tehnomarket за понатамошна споредба. За реализација на оваа задача, беше применета техниката на веб-скрејпинг, која овозможи автоматско

извлекување на податоците од веб-страниците со помош на Python и неколку специфични библиотеки, како што се Requests и BeautifulSoup.

За собирање на податоци од Setec и Tehnomarket, беше изработен скрипт кој ги презема податоците од секоја страница на категоријата мобилни телефони. Процесот започна со дефинирање на базната URL адреса, по што беше направена итерација низ сите страници (од 1 до 10). За секоја страница се испраќаше HTTP барање со употреба на requests библиотеката и се добиваше HTML документот на таа страница.

Потоа, со користење на BeautifulSoup, HTML документот беше парсиран за да се извлечат релевантните податоци, вклучувајќи ги името на телефонот, шифрата, редовната цена, и акциската цена. Секој производ беше идентификуван преку специфични HTML класи и атрибути, овозможувајќи систематско и прецизно извлекување на податоците. Овие податоци потоа се организираа во структуриран формат кој беше лесен за понатамошна обработка и анализа.

2.Обработка и зачувување на податоците

Откако податоците беа успешно извлечени од двете веб-страници, тие беа организирани и зачувани во табели. Оваа организација овозможи понатамошно чистење и нормализација на податоците, со цел да се отстранат дупликатите, да се усогласат форматите, и да се подготват податоците за директна споредба. Така, се обезбедија точни и конзистентни информации кои беа клучни за понатамошната анализа на цените и за постигнување на целите на овој проект.

Веб-скрејпингот ни овозможи ефикасно и систематско прибирање на голема количина на податоци, кои ќе бидат основа за споредба на цените и ќе помогнат во добивање на значајни увиди за потрошувачите при изборот на производи.

3. Чистење и обработка на податоци

Откако податоците беа собрани од Setec и Tehnomarket, следниот важен чекор беше нивното чистење и обработка. Овој процес е клучен за осигурување на точноста и конзистентноста на податоците кои ќе бидат користени за споредба на цените. Во продолжение, ќе бидат опишани главните чекори кои беа преземени во оваа фаза на проектот.

Отстранување на несакани елементи од цените

Податоците за цените од Setec беа собрани со ознаката "Ден." што го означува македонскиот денар. За да се олесни понатамошната анализа и споредба, беше неопходно да се отстрани овој збор и да се отстранат точките кои служат како илјадници сепаратори. Овој чекор беше изведен со користење на функцијата str.replace, која ги

отстрани несаканите елементи од колоните за редовна и акциска цена. Со оваа трансформација, цените беа претворени во чисти бројки, што овозможува нивно директно споредување и анализа.

Бришење на несуштински колони

Во податоците собрани од Tehnomarket, секој производ имаше поврзан URL кој водеше до страницата на производот. Оваа колона, иако корисна за проверка, не беше релевантна за понатамошната анализа на цените, па затоа беше избришана. Откако оваа колона беше отстранета, податоците беа полесни за управување и обработка.

Нормализација и усогласување на имињата на производите

Во процесот на чистење на податоците, беше извршена нормализација на имињата на производите, бидејќи некои од нив содржеа варијации во форматот и правописот. Се применија техники за автоматско чистење и нормализација на имињата, со што се осигура дека секоја верзија на производот е точно споредена меѓу двата продавници.

Конверзија на цените во соодветен формат

Дополнително, беше извршена конверзија на вредностите на цените од текстуален формат во нумерички, со цел да се обезбеди точност при понатамошната анализа. Ова беше особено важно за да се осигури дека сите вредности се правилно споредени и анализирани. Со оваа конверзија, сите текстуални вредности на цените беа успешно претворени во нумерички формат, овозможувајќи поефикасна и прецизна анализа и споредба.Сите овие чекори доведоа до комплетно подготвени и чисти податоци, кои беа подготвени за понатамошна анализа и извлекување на релевантни заклучоци за цените на мобилните телефони меѓу Setec и Tehnomarket.

4.Агрегација на продукти

Со цел да се обезбеди точна и релевантна споредба, беше потребно да се идентификуваат истите или многу слични продукти од Setec и Tehnomarket и да се споредат нивните цени. Ова беше постигнато преку користење на напредни техники за споредба на текст и агрегација на податоци.

Процес на агрегација:

1. **Користење на Fuzzy Matching:** За да се идентификуваат слични продукти од двете продавници, беше применет методот на "fuzzy matching," кој овозможува споредба на текстуални податоци кои не се идентични, но се доволно слични за да бидат разгледани како исти продукти. За оваа цел, беа користени библиотеките fuzzywuzzy и rapidfuzz, кои овозможуваат споредба на текстови базирана на различни алгоритми за мерење на сличност.

2. **Чистење на текстуалните податоци:** За да се зголеми точноста на споредбата, се примени агресивно чистење на текстуалните податоци. Ова вклучуваше отстранување на бројки, специјални знаци, и специфични клучни зборови кои би можеле да го отежнат процесот на споредба, како што се "black," "silver," "ultra," "pro," "lite," итн. Целта беше да се остане само со основниот модел на телефонот, што би овозможило попрецизно споредување.
3. **Примена на fuzz.token_set_ratio:** Откако текстовите беа исчистени, за споредба се користеше функцијата fuzz.token_set_ratio, која ги споредува два текста базирани на сличноста на нивните токени. Овој метод овозможува да се идентификуваат најсличните производи од двете листи, дури и ако нивните имиња не се идентични.
4. **Филтрирање и зачувување на резултатите:** Само производи кои достигнаа праг од 90% сличност беа разгледани како потенцијални натпревари. Потоа, се создаде нов DataFrame во кој беа вклучени само производите кои беа пронајдени и во Setec и во Tehnomarket, заедно со нивните соодветни цени. Дополнително, сите дупликати и некомплетни записи беа отстранети за да се обезбеди чиста и консолидирана база на податоци за понатамошна анализа.

Овој процес овозможи успешно да се агрегираат и споредат цените на мобилните телефони од Setec и Tehnomarket. На крајот, со овој пристап се доби база на податоци која ги содржи само оние продукти кои се достапни во двете продавници, со нивните соодветни цени, што овозможи понатамошна анализа и споредба на понудените цени.

Резултати и Визуелизации

Во овој дел од проектот, ќе ги прикажеме резултатите од анализата преку визуелизации. Визуелизациите се креирани за да овозможат појасен увид во податоците и да помогнат во разбирањето на разликите и сличностите помеѓу цените на истите производи од Setec и Tehnomarket. Преку овие графички претстави, ќе ги истакнеме клучните наоди и трендови кои произлегуваат од собраните и обработените податоци.

Визуелизација 1: Distribution of Price Differences

Опис: Оваа визуелизација е хистограм кој ја прикажува распределбата на разликите во цените на телефони помеѓу Setec и Tehnomarket. На x-оската е прикажана разликата во цените, додека на y-оската е прикажана фреквенцијата на тие разлики.

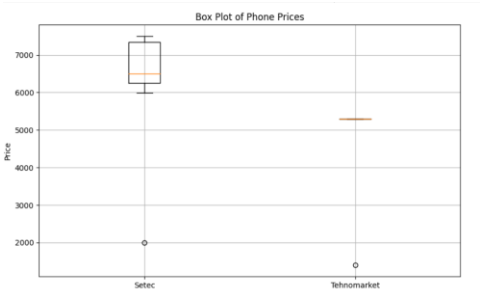
Анализа: Од хистограмот може да се забележи дека најголем дел од разликите во цените се концентрирани околу две точки. Ова укажува дека постојат две главни групи на телефони каде разликата во цените е најчеста. Ова може да биде резултат на различни фактори како што се различни модели, спецификации или промоции во двете продавници.

Визуелизација 2: Scatter Plot-Setec vs. Tehnomarket Phone Prices

Опис: Оваа визуелизација е scatter plot кој ги прикажува цените на телефони во Setec во споредба со цените на истите телефони во Tehnomarket. На x-оската се прикажани цените во Setec, додека на y-оската се прикажани цените во Tehnomarket.

Анализа: Од scatter plot-от може да се забележи како цените на телефони во Setec се споредуваат со цените во Tehnomarket. Ако точките се блиску до дијагоналната линија (y = x), тоа значи дека цените се слични во двете продавници. Ако точките се под или над линијата, тоа укажува на разлики во цените. Оваа визуелизација помага да се идентификуваат телефони кои имаат значителни разлики во цените помеѓу двете продавници.

Визуелизација 3: Box Plot of Phone Prices



Опис: Оваа визуелизација е box plot кој ги прикажува цените на телефони во Setec и Tehnomarket. На x-оската се прикажани двете продавници, додека на y-оската се прикажани цените на телефони.

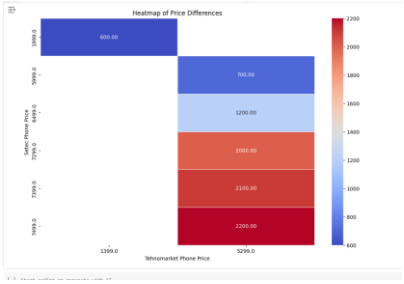
Анализа: Од box plot-от може да се забележи дека цените на телефони во Setec имаат поголема варијација во споредба со цените во Tehnomarket. Во Setec, цените се распространети во поширок опсег, додека во Tehnomarket, цените се повеќе концентрирани околу една вредност. Ова укажува дека Setec нуди поширок спектар на телефони со различни ценовни категории, додека Tehnomarket има помала варијација во цените.

Визуелизација 4: Interactive Bubble Chart of Phone Prices and Differences

Опис: Оваа визуелизација е интерактивен bubble chart кој ги прикажува цените на телефони во Setec во споредба со цените на истите телефони во Tehnomarket. Големината на секој балон ја претставува апсолутната вредност на разликата во цените, додека бојата ја претставува вистинската разлика во цените. При поминување со глувчето над балоните, се прикажува името на телефонот.

Анализа: Од bubble chart-от може да се забележи како големите балони укажуваат на значителни разлики во цените помеѓу двете продавници. Оваа визуелизација овозможува лесно идентификување на телефони со најголеми разлики во цените, што може да биде корисно за потрошувачите кои сакаат да ги најдат најдобрите понуди.

Визуелизација 5: Heatmap of Price Differences



Опис: Оваа визуелизација е heatmap кој ги прикажува разликите во цените на телефони помеѓу Setec и Tehnomarket. На x-оската се прикажани цените во Tehnomarket, додека на y-оската се прикажани цените во Setec. Боите ја претставуваат големината на разликите во цените, каде што топлите бои (црвени) укажуваат на поголеми разлики, а ладните бои (сини) на помали разлики.

Анализа: Оваа визуелизација овозможува лесно идентификување на ценовните разлики помеѓу двете продавници. Поголемите разлики се прикажани со потопли бои, што укажува на значителни разлики во цените за одредени модели на телефони.

Статистика на податоците

	Редовна цена во Сетек	Редовна цена во Техномаркет
count	7.000000	7.000000
mean	6170.428571	4741.857143
std	1923.290831	1474.061445
min	1999.000000	1399.000000
25%	6249.000000	5299.000000
50%	6499.000000	5299.000000
75%	7349.000000	5299.000000
max	7499.000000	5299.000000

Интерпретација на резултатите

Број на податоци: Бројот на податоци за Setec е значително поголем од оној за Technomarket, што може да укаже на поголема разновидност на продукти во Setec.

Просек: Просечната цена во Setec е значително повисока од онаа во Technomarket, што може да укаже на различни ценовни стратегии или различен квалитет на продуктите.

Стандардна девијација: Стандардната девијација е повисока во Setec, што укажува на поголема варијација во цените на продуктите.

Минимална и максимална вредност: Минималната и максималната вредност во Setec се повисоки од оние во Technomarket, што може да укаже на поширок опсег на цени во Setec.

Дескриптивни статистики

```
                          count         mean          std     min     25%  \
Редовна цена во Сетек       7.0  6170.428571  1923.290831  1999.0  6249.0
Редовна цена во Техномаркет 7.0  4741.857143  1474.061445  1399.0  5299.0

                            50%     75%     max      variance  skewness  \
Редовна цена во Сетек      6499.0  7349.0  7499.0  3.699048e+06 -1.692412
Редовна цена во Техномаркет 5299.0  5299.0  5299.0  2.172857e+06 -2.041241

                           kurtosis
Редовна цена во Сетек      1.402884
Редовна цена во Техномаркет 2.166667
```

	count	mean	std	min	25%	50%	75%	max	variance	skewness	kurtosis
Редовна цена во Сетек	7.0	6170.428571	1923.290831	1999.0	6249.0	6499.0	7349.0	7499.0	3.699048e+06	-1.692412	1.402884
Редовна цена во Техномаркет	7.0	4741.857143	1474.061445	1399.0	5299.0	5299.0	5299.0	5299.0	2.172857e+06	-2.041241	2.166667

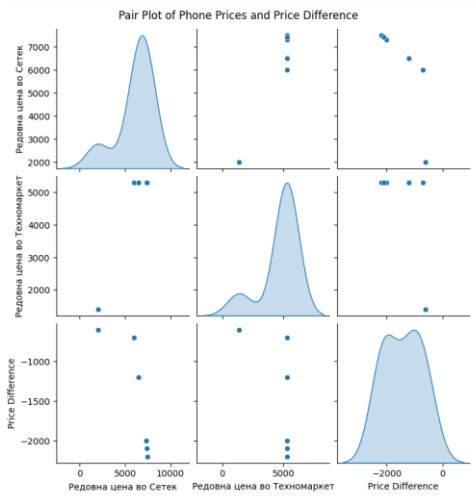
•**Варијанса:** Варијансата е повисока во Setec, што укажува на поголема варијација во цените на продуктите.

•**Скју:** Негативната скју вредност укажува на тоа дека дистрибуцијата на цените е асиметрична и дека има повеќе цени кои се под просекот.

•**Куртоза:** Вредностите на куртозата укажуваат на тоа дека дистрибуцијата на цените во Technomarket има подолги опашки во споредба со Setec.

Овие статистички податоци ни даваат подлабоко разбирање за дистрибуцијата на цените во двете продавници и ни помагаат да ги идентификуваме клучните разлики во нивните ценовни стратегии.

Визуелизација 6: Pair Plot- Phone Prices and Price Difference



Овој pair plot ги прикажува односите помеѓу редовните цени на телефони во Setec и Technomarket, како и разликата во цените помеѓу двете продавници. Дијагоналните графици се Kernel Density Estimate (KDE) графици, кои даваат мазна проценка на дистрибуцијата за секоја променлива. Останатите графици се scatter plots кои ги прикажуваат односите помеѓу паровите променливи.

- **KDE графици**: Овие графици на дијагоналата ни покажуваат како се распределени цените во Setec и Technomarket, како и разликата во цените. На пример, KDE графикот за ‘Price Difference’ покажува дека разликата во цените е асиметрична, што укажува дека една продавница има тенденција да има повисоки цени од другата.
- **Scatter plots**: Овие графици ни покажуваат како се поврзани цените во Setec и Technomarket. На пример, scatter plot-от помеѓу ‘Редовна цена во Сетек’ и ‘Редовна цена во Техномаркет’ ни покажува како цените во Setec се споредуваат со цените во Technomarket за различни продукти.

Модел за предвидување на цените во Техномаркет врз основа на цените на Сетек

Во овој дел, користиме модел за линеарна регресија за да предвидиме цени на телефони во Technomarket врз основа на цените во Setec. Процесот вклучува следниве чекори:

1. **Подготовка на податоците**: Користиме цените на телефони од Setec како карактеристика (feature) и цените од Technomarket како цел (target).
2. **Поделба на податоците**: Податоците ги делиме на тренинг сет и тест сет со сооднос 80:20.
3. **Тренирање на моделот**: Користиме линеарна регресија за да го тренираме моделот врз основа на тренинг сетот.
4. **Предвидување и евалуација**: Го користиме моделот за да предвидиме цени на тест сетот и ја пресметуваме средната квадратна грешка (Mean Squared Error - MSE) за да ја оцениме точноста на моделот.

Оваа анализа може да биде корисна за разбирање на врската помеѓу цените во двете продавници и за идентификување на потенцијални разлики во ценовните стратегии.

Заклучок

Во рамките на овој проект беше спроведена детална анализа на цените на мобилни телефони помеѓу Setec и Tehnomarket, со цел да се овозможи поголем увид во разликите и сличностите на понудите на овие две продавници. Примената на веб-скрејпинг техники и напредни методи за споредба на текстуални податоци резултираше со ефикасно собирање и обработка на податоци, што дозволи точно и детално споредување на цените на истите производи.

Анализата откри значајни разлики во цените помеѓу двете продавници, што укажува на потенцијални варијации во ценовните стратегии и промоции. Исто така, методите за чистење и усогласување на податоците, како и примената на fuzzy matching, овозможија точна идентификација на слични производи и појасна споредба на цените.

Основни заклучоци:

1. **Разлики во Цените:** Се откриј значителни разлики во цените помеѓу двете продавници, што може да биде од корист за потрошувачите при донесување на одлуки за купување. Односно цените во Сетек се повисоки од цените во Техномаркет за голем број од производите.
2. **Ефикасност на Методите:** Напредните методи за обработка на податоци и споредба на текстуални податоци ја зголемија точноста на анализата и обезбедија вредни увиди.

3. **Подобрена Ажурираност:** Редовното ажурирање на податоците и примена на напредни техники за веб-скрејпинг се клучни за добивање на точни и актуелни информации.

Проектот не само што обезбеди корисни резултати за потрошувачите, туку и истакна значењето на точните и ажурирани податоци во процесот на споредба на цените. Со продолжување на примената на овие техники и методи, можно е понатамошно усовршување на процесот на собирање и анализа на податоци, што ќе доведе до подобри резултати и поголема задоволство на корисниците.