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:

Laptop Price Comparison and Aggregation from Setec and Tehnomarket

Линк до кодот:

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

https://github.com/tomislavpavloski/VNP/blob/main/DS_23_24_standardni_seminarski_raboti_tema_10_201055.ipynb

Линк до видеото:

Laptop Price Comparison and Aggregation from Setec and Tehnomarket

https://youtu.be/pbivlage45c

Изработи:
Томислав Павлоски, СИИС - 201055

Одговорен:
асс. Ана Тодоровска

Преглед на проектот:

Проектот има за цел да ја анализира и спореди ценовната понуда на лаптопи од две значајни продавници во Македонија: Setec и Tehnomarket. Основната задача е да се соберат и анализираат податоци за лаптопи кои се продаваат на овие веб страници, со цел да се утврдат разликите во цените за идентични производи и да се добијат корисни информации за потрошувачите.

Проектот главно се состои од следниве клучни чекори - фази:

1. Собирање на податоци:

Извлекување на информации од веб страниците на Setec и Tehnomarket. Ова вклучува собирање на податоци за лаптопи како што се имиња, шифри, цени итн. Податоците ќе се извлечат со помош на техники за веб скрејпинг.

2. Агрегација на идентични производи:

По собирањето на податоците, следниот чекор е агрегација на идентичните производи. Ова значи дека ќе се идентификуваат, спојат и анализираат лаптопи со исти или многу слични спецификации од двете веб страници, за да се овозможи директна споредба на цените.

3. Споредба на цените:

Анализирањето на споредбата на цените се фокусира на утврдување на разликите во цените помеѓу Setec и Tehnomarket за исти модели. Оваа споредба ќе помогне да се види дали постојат значајни разлики во цените за истите лаптопи.

4. Визуелизација и анализа:

Ова ќе овозможи подобро разбирање на трендовите и разликите во цените. Дополнително, ќе се извршат статистички анализи за да се оценат и интерпретираат резултатите.

Собирање податоци

Податоците ќе се собираат со помош на техники за веб скрејпинг со Python библиотеки како requests, BeautifulSoup, pandas итн.

Опис на податоците (Setec)

1. **Име на лаптоп:**
    - **Тип на податок:** String
    - **Опис:** Името на моделот на лаптопот.
2. **Регуларна цена:**
    - **Тип на податок:** String
    - **Опис:** Цената на лаптопот пред попуст.
3. **Попуст цена:**
    - **Тип на податок:** String
    - **Опис:** Цената на лаптопот по попуст.
4. **Код на лаптоп:**
    - **Тип на податок:** String
    - **Опис:** Кодот на лаптопот за идентификација.

Опис на податоците (Tehnomarket)

1. **Име на лаптоп:**
    - **Тип на податок:** String
    - **Опис:** Името на моделот на лаптопот.
2. **Регуларна цена:**
    - **Тип на податок:** String
    - **Опис:** Цената на лаптопот пред попуст.
3. **Попуст цена:**
    - **Тип на податок:** String
    - **Опис:** Цената на лаптопот по попуст.
4. **URL:**
    - **Тип на податок:** String
    - **Опис:** Линк до страницата со производот на Tehnomarket.

Секој податок е извлечен од веб страниците на Setec и Tehnomarket и зачуван во посебен CSV фајл за понатамошна анализа. Описот на податоците е сличен и за Tehnomarket со мали разлики кои секако ќе бидат елиминирани, бидејќи во случајов нам од интерес ни се имињата на лаптопите и нивната цена. Подоцна настанува спојување на податочните множества во едно податочно множество за негова агрегација, визуелизација, споредба и анализа.

Агрегација на идентични производи

➔ **Форматирање на податоците:**

Конвертирање на цените од string (оние кои треба да бидат конвертирани) во нумерички за понатамошна анализа. Имињата на лаптопите ќе бидат нормализирани за да се осигураме дека не постојат непотребни празни места или специјални карактери.

➔ **Справување со missing values, чистење на податоци и спојување во едно финално податочно множество за обработка:**

При справувањето со вредности што недостасуваат, сите редови со пропуштени или NaN вредности се отстранети за да се овозможи конзистентност на податоците. Дополнително, сите дупликат редови ќе бидат елиминирани за да се осигураме дека во финалната анализа ќе бидат задржани само уникатни записи. Ова е многу важен чекор за обезбедување на точност и прецизност во понатамошната обработка на податоците. Во процесот на наоѓање и споредување на слични лаптопи, користен е напреден алгоритам за чистење на податоците, при што имињата на лаптопите се нормализираат преку отстранување на непотребните симболи и се користи **token_set_ratio** за подобра прецизност. Сите записи со 80% или повисока сличност се вклучени во финалната табела. На овој начин знаеме дека се избрани токму оние лаптопи кои се идентични и се продаваат на двете веб страници со цел подоцна да правиме споредба на нивната цена во секоја од продавниците.

- **Инсталација на потребните библиотеки:**

  fuzzywuzzy, rapidfuzz, и re, за манипулација со стрингови.

- **Екстракција на имињата на лаптопите:**

  Конверзија на колоните „Laptop Name“ од двете податочни множества на Setec и Tehnomarket во листи за полесна обработка.

- **Чистење и стандардизација на имињата на лаптопите:**

  Чистење на секое име на моделот на лаптопот преку отстранување на неалфанумерички знаци и конвертирање на текстот во мали букви за да се осигура конзистентна споредба.

- **Наоѓање на најдобрите совпаѓања:**

  Користење на token_set_ratio од fuzzywuzzy за наоѓање на најдобрите совпаѓања на имињата на лаптопите помеѓу Setec и Tehnomarket, земајќи ги предвид само оние со 80% или поголема сличност.

- **Преземање на цени на совпаднатите лаптопи:**

  За секој совпаднат лаптоп, ги преземаме соодветните цени од оригиналните колони на DataFrame и ги потоа ги чуваме во нов DataFrame.

- **Чистење на совпаднатите податоци:**

  Отстранување на сите редови со missing или NaN вредности и елиминација на дупликат редовите за да се осигураме дека ќе се задржат само уникатните записи.

- **Преуредување и ресетирање на индексот:**

  Преуредување на редовите на финалниот DataFrame за подобра презентација и ресетирање на индексот да биде секвенцијален.

- **Прикажување на финалниот DataFrame:**

  Приказ на исчистениот DataFrame, подготвен за понатамошни анализи и визуелизации.

Споредба на цените

Од споредбата на цените на лаптопите меѓу Setec и Tehnomarket можат да се извлечат неколку важни заклучоци. Прво, може да се забележат разлики во цените за исти модели на лаптопи, што укажува на различни пазарни стратегии на овие две продавници. Овие разлики може да бидат резултат на различни набавни политики, промоции, или попусти кои се применуваат на одредени модели. Дополнително, анализата може да открие дали одредени брендови или модели систематски се поскапи или поевтини кај едниот продавач во споредба со другиот. Ова може да им помогне на потрошувачите да донесат поинформирани одлуки при купување, земајќи ја предвид не само цената, туку и дополнителните карактеристики и вредности што ги нудат продавачите. Оваа информација е од голема важност не само за потрошувачите, туку и за производителите и дистрибутерите, кои можат да ја користат за да ги оптимизираат своите понуди и да ја подобрат својата пазарна позиција.

Визуелизации и анализа

Визуелизации

Визуелизацијата на податоците е значаен дел од процесот на анализа бидејќи овозможува полесно разбирање на трендовите, разликите и моделите што произлегуваат од собраните податоци. Преку графички прикази, може да се добие подобра перспектива за тоа како цените на лаптопите варираат меѓу двете продавници и различните модели, што овозможува појасна интерпретација на податоците. Дополнително, статистичките анализи служат како основа за подлабоко истражување на добиените резултати. Овие анализи вклучуваат различни статистички тестови и анализи, кои ќе ни помогнат да ги разбереме и оцениме разликите во цените, влијанието на попустите, како и општите трендови на пазарот. Ова резултира со подетални заклучоци и препораки, кои би биле од големо значење за потенцијалните купувачи и продавачи. Во продолжение ќе разгледаме краток опис на визуелизациите (кои може да се видат во самиот проект), нивно кратко објаснување и зошто токму таа визуелизација е значајна.

❖ **Scatter Plot of Setec vs. Tehnomarket Laptop Prices**

Оваа scatter plot визуелизација прикажува споредба на цените на идентични лаптопи помеѓу двата продавачи: Setec и Tehnomarket. На графикот, цените на лаптопите од Setec се прикажани на X-оската, додека цените од Tehnomarket се прикажани на Y-оската.

Точките на графикот соодветствуваат на цените на истите модели на лаптопи од двата продавачи. Оваа визуелизација ни помага да идентификуваме разлики во цените помеѓу Setec и Tehnomarket. Ако точките се групираат околу дијагонална линија, тоа укажува на конзистентни цени помеѓу двата продавачи. Најголемите расфрлања на точките можат да укажат на значителни разлики во цените, што може да биде резултат на различни ценовни стратегии, попусти или промоции. Оваа визуелизација е корисна за брзо согледување на конкурентноста на цените и може да помогне во донесувањето одлуки при купување или анализа на пазарната конкуренција.

❖ **Histogram of Price Differences Between Setec and Tehnomarket**

Оваа визуелизација прикажува распределба на разликите во цените помеѓу лаптопите од Setec и Tehnomarket. На графикот, X-оската ја прикажува разликата во цените помеѓу Tehnomarket и Setec, додека Y-оската ја прикажува фреквенцијата на тие разлики.

Хистограмот помага да се разбере како се распределуваат разликите во цените. Ако распределбата е центрирана околу нулата, тоа укажува на слични цени помеѓу двата продавачи. Ако има значителни врвови на позитивната или негативната страна, тоа означува значителни разлики во цените. Оваа визуелизација е корисна за идентификување на ценовните разлики и може да помогне во анализа на ценовните стратегии и пазарната динамика.

❖ **Interactive Bubble Chart of Laptop Prices and Differences**

Оваа интерактивна визуелизација со bubbles претставува визуелна авантура во светот на лаптоп цените од Setec и Tehnomarket. На X-оската се сместени цените на лаптопите од Setec, додека Y-оската ги покажува цените од Tehnomarket.

Секој bubble на графикот носи приказ на разликата во цените, со големината на меурчињата која го одразува степенот на разликата: поголемите bubbles укажуваат на поголеми разлики.

Бојата на bubbles ја открива насоката на ценовната разлика, со топли тонови за поголеми разлики и ладни тонови за помали. Кога поминуваме со курсорот над кое било меурче, добиваме детални информации за конкретниот лаптоп.

- **Correlation Heatmap of Laptop Prices and Differences**

Heat мапата нуди длабок увид во корелацијата помеѓу цените на лаптопите од Setec, цените од Tehnomarket и разликата во цените помеѓу овие два продавачи. На графикот, црвените нијанси укажуваат на негативна корелација, што значи дека кога цената на лаптопите од Setec расте, цената на истите лаптопи од Tehnomarket обично опаѓа, и обратно, сините нијанси означуваат позитивна корелација, што значи дека зголемувањето на цената на лаптопите од Setec се совпаѓа со зголемувањето на цената на истите лаптопи од Tehnomarket.

Correlation heatmap ги прикажува конкретните нивоа на корелација со броеви директно на графикот, што ни овозможува прецизно да ја разбереме врската помеѓу ценовните варијации и разликата во цените.

Анализа

Матрицата на корелација ги прикажува врските помеѓу цените на лаптопите од Setec, цените од Tehnomarket и разликата во цените помеѓу нив.

Корелацијата помеѓу цените на лаптопите од Setec и Tehnomarket изнесува -0.36, што укажува на умерено негативна корелација. Ова значи дека кога цената на лаптопите од Setec расте, цената на истите лаптопи од Tehnomarket има тенденција да опаѓа, и обратно, но оваа корелација не е силна.

Корелацијата помеѓу цените на лаптопите од Setec и разликата во цените изнесува -0.95, што укажува на силна негативна корелација. Ова значи дека поголемите цени на лаптопите од Setec обично се поврзуваат со поголеми разлики во цените помеѓу Setec и Tehnomarket, со тенденција за поголеми разлики кога цените од Setec се повисоки.

Корелацијата помеѓу цените на лаптопите од Tehnomarket и разликата во цените изнесува 0.62, што укажува на умерена позитивна корелација. Ова значи дека кога цената на лаптопите од Tehnomarket расте, разликата помеѓу цените на лаптопите од Setec и Tehnomarket има тенденција да расте, и обратно.

Статистичките мерења нудат детална слика за ценовните опсези на лаптопите од Setec и Tehnomarket:

Цените на лаптопите од **Setec** имаат просечна вредност од 20,163.44 денари, со значителна варијација која се мери со стандардна девијација од 3,951.94 денари. Ова укажува на тоа дека цените на лаптопите од Setec имаат поширок распон, со најниска цена од 12,999 денари и највисока цена од 26,995 денари. Варијансата од 15,617,840,000 денари ја потврдува оваа варијабилност. Skewness на цените е -0.10, што укажува на мала лево наклонета распределба, додека kurtosis од -0.11 сугерира распределба која е близу до нормалниот облик, со умерено фокусирање околу просекот.

Од друга страна, лаптопите од **Tehnomarket** имаат просечна цена од 23,110.11 денари со помала варијација, која се мери со стандардна девијација од 1,516.12 денари. Овие цени се повеќе концентрирани, со минимална цена од 20,499 денари и максимална цена од 26,499 денари. Варијансата од 2,298,611,111 денари покажува дека цените на Tehnomarket имаат помала флуктуација во споредба со Setec. Skewness на цените од Tehnomarket е 0.80, што укажува на десно наклонета распределба, со повеќе високи вредности. Kurtosis од 1.75 покажува дека цените имаат поголеми екстремни вредности и се повеќе концентрирани околу медијаната.

Овие статистички мерења помагаат да се разберат разликите во ценовната структура помеѓу двата продавачи.

Краток преглед на резултатите од анализата на цените на лаптопите

Во рамките на оваа анализа на цените на лаптопите понудени од Setec и Tehnomarket, беа идентификувани неколку значајни работи кои ни откриваат многу за пазарната динамика и стратегиите на овие продавачи. Со користење на различни аналитички техники и визуелизации, успеавме да ги идентификуваме цените на лаптопите кај овие продавачи и да ја споредиме цената за ист модел. Со помош на корелациски анализи, откриваме дека постои одредена позитивна врска помеѓу цените на двата продавачи, но таа врска не е многу силна. На пример, еден и ист модел на лаптоп може да биде значително поскап во Tehnomarket отколку во Setec, или обратно. Овој тренд укажува дека Setec и Tehnomarket не секогаш следат една иста ценовна политика, туку нивните цени можат да варираат значително врз основа на различни фактори како што се маркетинг стратегиите, залихите, па дури и перцепциите на клиентите. Понатаму, анализата на разликите во цените меѓу двата продавачи откри уште еден интересен тренд. Во голем број на случаи, лаптопите од Setec беа поевтини од истите модели понудени од Tehnomarket. Ова сугерира дека Setec можеби применува поразлична ценовна политика, со цел да привлече поширок круг на клиенти и да ги натера да се одлучат за купување токму од нив. Tehnomarket, од друга страна, можеби се потпира на својата репутација, квалитетот на услугите, или дури и на додатните понуди и промоции кои ги нуди, за да ја оправда повисоката цена. Оваа разлика во цените е особено значајна за потрошувачите кои сакаат да го добијат најдоброто за своите пари, и укажува на потребата за внимателно споредување на цените пред донесувањето на конечната одлука за купување. Во анализата, хистограмот даде уште подлабок увид во тоа како варираат цените меѓу двата продавачи. Поголемиот дел од разликите се концентрирани околу нултата вредност, што значи дека цените на многу модели се релативно слични кај двата продавачи. Ова укажува на тоа дека Setec и Tehnomarket се во директна конкуренција за многу модели, и дека цените на тие модели се прилагодени за да бидат конкурентни на пазарот. Сепак, постојат и отстапувања од ова правило, што може да биде резултат на различните стратегиски одлуки кои ги применуваат овие два продавачи, како што се различни набавни цени, различни трошоци за маркетинг и слично. Дополнителна анализа на дистрибуцијата на цените откри уште неколку интересни појави. На пример, дистрибуцијата на цените во Tehnomarket покажа десно наклонета распределба, што значи дека повеќето од лаптопите кои се понудени од овој продавач имаат повисоки цени во споредба со просечната цена. Од друга страна пак, цените на лаптопите во Setec се распределени попропорционално околу просечната цена.

Заклучок

Оваа семинарска работа ни овозможи да добиеме подлабок увид во динамиката на пазарот и стратегиите на формирањето на цени што ги применуваат овие две големи продавници во Македонија. Еден од главните заклучоци што произлегоа од оваа анализа е дека постојат разлики во цените на истите модели на лаптопи помеѓу Setec и Tehnomarket. Овие разлики укажуваат на тоа дека купувачите имаат можност да заштедат со внимателна споредба на понудите пред да купат лаптоп. Како што може да се забележи, Setec генерално нуди пониски цени во споредба со Tehnomarket. Ова е од големо значење за купувачите, посебно за оние кои се во потрага да најдат добра зделка и да заштедат пари. Преку примената на разни методи за анализа на податоци, визуелизациите, корелациските анализи и слично, добиваме значајни резултати кои можат да послужат како основа за други понатамошни истражувања. Преку анализата на основните дескриптивни статистики, како што се средната вредност, медијаната и стандардната девијација, успеавме да создадеме јасна слика за распределбата на цените на лаптопите во Setec и Tehnomarket. Средната вредност на цените ни даде генерална претстава за нивото на цени на двата продавачи, при што беше забележано дека Tehnomarket има тенденција да нуди малку повисоки цени во споредба со Setec. Skewness ни даде увид во тоа дали распределбата на цените е наклонета кон пониски или повисоки вредности. Накратко, статистичките мерења беа навистина значајни за оваа семинарска работа, бидејќи овозможија не само да ги разбереме моменталните услови и цени на пазарот, туку и да создадеме основа за идни истражувања.