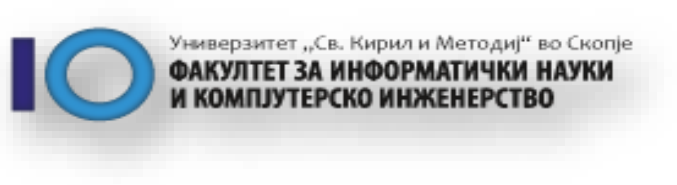


Стандарден проект по предметот

Вовед во наука за податоци

Тема 6:

Прибирање на податоци за различни производи од повеќе е-продавници, нивно претпроцесирање и стандардизација.

Изработка: Саво Костадинов, 201147

Ментор: асс. Димитар Пешевски

Линк до github : https://github.com/savokostadinov/ecommerce-data-pipeline-project

Јуни, 2025

Вовед

Во современиот дигитален свет, електронската трговија (e-commerce) станува сè поважен сегмент на економијата, овозможувајќи им на потрошувачите брз и лесен пристап до широк спектар на производи. Со зголемувањето на бројот на онлајн продавници, се јавува потребата од систематско собирање и анализа на податоци за подобро разбирање на пазарните трендови и потрошувачките навики.

Овој проект има за цел да собере, обработи и стандардизира податоци од три македонски онлајн продавници:

Kosuli.mk

FashionGroup.com.mk

Scout.mk

Собраните податоци вклучуваат информации како што се име на производот, цена, попуст, и други релевантни атрибути. Преку автоматизирано веб-скрејпање, се обезбедува ефикасно и конзистентно прибирање на податоци, што е клучно за понатамошна анализа и донесување информирани одлуки.

Стандардизацијата на податоците е од суштинско значење за обезбедување на унифициран формат, што овозможува полесна анализа и споредба на производите од различни извори. Ова вклучува нормализирање на вредности, отстранување на дупликати и поправање на невалидни или недостапни линкови. Како резултат, се добива квалитетно и конзистентно податочно множество подготвено за понатамошна обработка.

Овој документ ќе ги опфати следните поглавја:

1. **Методологија** – Опис на процесот на собирање, обработка и стандардизација на податоците.

2. **Резултати** – Презентација на анализите и визуелизациите добиени од обработените податоци.

3. **Заклучок** – Сумирање на наодите и препораки за идни истражувања.

4. **План на работа** – Хронолошки преглед на извршените активности и идните чекори.

Собирање на податоци (Web Scraping)

Целта беше да се извлечат релевантни информации за производите. На пример:

- Назив на производот
- Цена
- Попуст
- Краток опис
- Линк до производот
- Дополнителни атрибути (на пример, боја, големина, материјал)
- Линк до слика и продукт

Технологии и алатки

За реализација на web scraping процесот, искористено е, **Python** во комбинација со следните библиотеки:

- **Requests**: за испраќање HTTP барања и добивање на HTML содржина
- **BeautifulSoup**: за парсирање и екстракција на податоци од HTML документите
- **Selenium**: за интеракција со динамички содржини и JavaScript-генерирани елементи

Процес на прибирање на податоци

1. **Идентификација на целните страници**
2. **Прибирање на HTML содржина**
3. **Парсирање на HTML**
4. **Обработка на динамички содржини**

```
# Иницијализација на Web Driver за Selenium
driver = webdriver.Chrome()
driver.get("https://kosuli.mk")
wait = WebDriverWait(driver, 10)
```

```
# Скрејпање на цената
try:
    price = driver.find_element(By.CSS_SELECTOR, "p.price").text.strip()
except NoSuchElementException:
    price = ""
```

```
# Скрејпање на големините со што се скрола надоле за да се пронајдат величините каде се достапни
sizes = []
try:
    descr = driver.find_element(By.ID, "tab-description")
    driver.execute_script("arguments[0].scrollIntoView(true);", descr)
    div = descr.find_element(By.XPATH, ".//div[contains(text(),'Достапни големини')]")
    ul = div.find_element(By.XPATH, "following-sibling::ul[1]")
    sizes = [li.text.strip() for li in ul.find_elements(By.TAG_NAME, "li") if li.text.strip()]
```

```
    data.append({
        "Name": name,
        "Price": price,
        "Sizes": sizes_str,
        "Image_URL": image_url,
        "Product_URL": href
    })

# Зачувување во Пандас датафрејм во цсв формат и нивно принтање секој продукт посебно со неговото име и величини
df = pd.DataFrame(data)
df.to_csv("kosuli_casa_moda.csv", index=False, encoding="utf-8")
```

Претпроцесирање на податоците (Data Cleaning)

По прибирањето на необработените податоци од трите онлајн продавници (Kosuli.mk, FashionGroup.com.mk и Scout.mk), следниот чекор е претпроцесирањето, односно чистењето на податоците. Овој процес е неопходен за да се обезбеди точност, конзистентност и подготвеност на податоците за понатамошна анализа и визуелизација.

При иницијалото надгледување на собраните податоци, идентификувани се следните чести проблеми:

- **Недостасувачки вредности**: Некои записи немаат пополнети полиња за цена, попуст или опис.
- **Дупликати**: Исти производи се појавуваат повеќе пати во податочното множество, особено при скрејпање од страници со различни филтри или категории.
- **Валута:** Во цените на продуктите валутата се наоѓа внатре во цената којашто треба да се среди и да биде само нумерички формат.
- **Празни места:** Некои записи имаат полиња коишто имаа празни места пред и после прикажаното.
- **Интерпункциски знаци**: Некои записи имаат знаци како , . ! [] коишто мора да се извадат

Техники за чистење и трансформација на податоците

1. **Ракување со недостасувачки вредности**: За полињата со недостасувачки информации, применети се различни стратегии:
   - Отстранување на записи со критични недостасувачки полиња.
   - Пополнување на недостасувачки вредности со медијана или модус, каде што е применливо, или пополнување со порака валидна и прилагодлива за записот.

2. **Стандардизација на формати**: Цените се конвертирани во нумерички формат, отстранувајќи ги симболите за валута и текстуалните описи. На пример, "1.200 ден" или “1200.00” е конвертирано во 1200.

3. **Валидација на URL адреси**: Извршена е проверка на сите линкови за да се осигури дека водат до валидни страници. Скршените линкови се отстранети или заменети со валидни, каде што е можно.

```
rawData = pd.read_csv("kosuli_casa_moda.csv")
rawData.head(5)
```

	Name	Price
0	CM098 REGULAR FIT LONG SLEEVE BLUE	1500 ДЕН
1	CM095 REGULAR FIT LONG SLEEVE SILVER	1500 ДЕН
2	CM094 REGULAR FIT LONG SLEEVE RED/SILVER	1500 ДЕН

кт	Цена	Валута
...	1500	ДЕН
...	1500	ДЕН

Стандардирање и обединување на податоците

По успешното собирање и претпроцесирање на податоците од трите онлајн продавници (Kosuli.mk, FashionGroup.com.mk и Scout.mk), следниот клучен чекор е стандардизирањето и обединувањето на податоците во единствено, унифицирано податочно множество. Овој процес овозможува конзистентност во структурата на податоците, што е од суштинско значење за точна анализа и визуелизација.

Цел на стандардизацијата

Целта на стандардизацијата е да се создаде единечен запис за секој производ, со истоветни колони и формати, без разлика од кој извор потекнуваат податоците. Ова вклучува:

- **Унифицирани имиња на колони**: На пример, сите полиња за цена ќе се именуваат како `price`, без разлика дали оригиналниот извор ги именува како "цена", "Price", "Цена" или слично.

- **Стандардизација на атрибути**: Клучните параметри како боја, големина, бренд и други ќе бидат претставени во конзистентен формат, најчесто како key-value парови.

- **Обединување во едно податочно множество**: Сите исчистени податоци се конкатенирани во едно податочно множество и додадена колона да означува кој во која категорија припаѓа.

Процес на обединување на податоците

Со оглед на тоа што податоците од различните продавници можат да имаат различни структури и имиња на полиња, потребно е нивно обединување во единствено податочно множество. Овој процес вклучува:

Стандардизација на имињата на колоните: Се користат техники за преименување на колоните во Pandas, за да се осигури дека сите податочни множества имаат исти имиња на колони.

Конверзија на вредности: Се осигурува дека сите вредности се во конзистентен формат, на пример, цените се конвертираат во нумерички формат без симболи за валута, а попустите се претставуваат како проценти или децимални броеви.

Анализа на податоците

По успешното собирање, претпроцесирање и стандардирање на податоците од трите македонски е-продавници (Kosuli.mk, FashionGroup.com.mk и Scout.mk), следниот чекор е детална анализа на податоците. Целта е да се добијат корисни увиди за пазарот, трендовите и потенцијалните аномалии.

Резимирање на податоците

Првиот чекор во анализата е добивање на основни статистички мерки за да се разбере структурата на податоците. Со користење на функцијата `describe()` од библиотеката Pandas, може да се добијат информации како:

- Број на записи
- Минимална и максимална вредност
- Средна вредност (mean)
- Медијана (50th percentile)
- Стандардна девијација

Ова помага во идентификација на потенцијални аномалии, како што се невообичаено високи или ниски цени.

Визуелизација на дистрибуции

За подобро разбирање на распределбата на цените и попустите, се користат различни типови на графици:

- **Хистограм (Histogram)**: Прикажува фреквенцијата на цени во одредени интервали, што помага во идентификација на најчестите ценовни опсези.

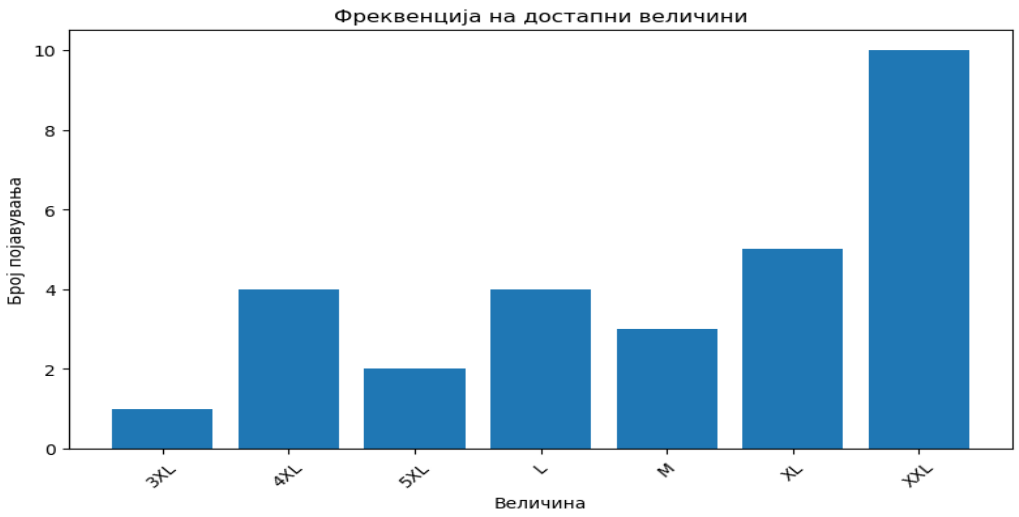
- **Бокс-плот (Boxplot)**: Овозможува визуелизација на медијаната, квартилите и потенцијалните аутлајери во цените или попустите.

- **Скатер-плот (Scatter plot)**: Прикажува релацијата помеѓу две променливи, на пример, цената и попустот, што помага во идентификација на корелации.

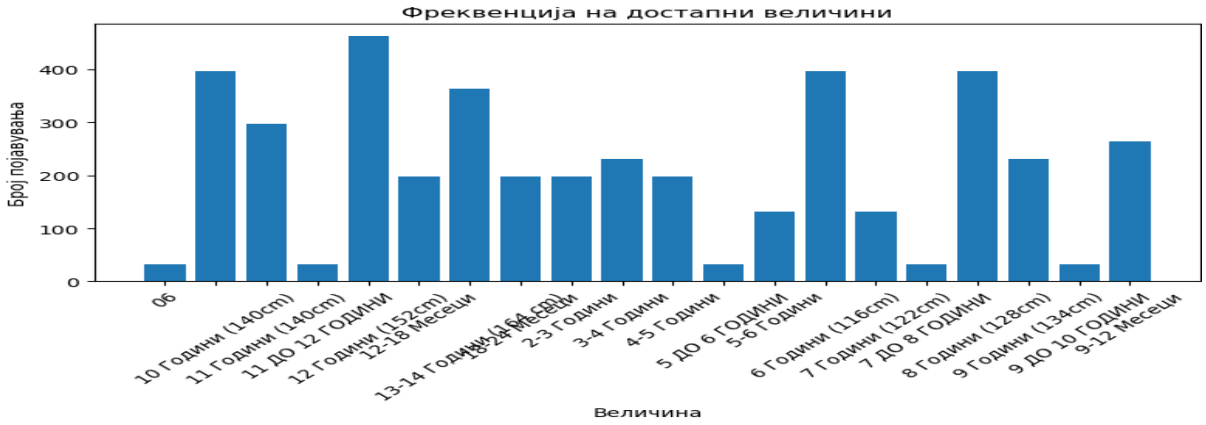
Овие визуелизации се креираат со користење на библиотеки како Matplotlib и Seaborn.

Matplotlib – основна, моќна библиотека за статички графици со детална контрола на секој елемент.

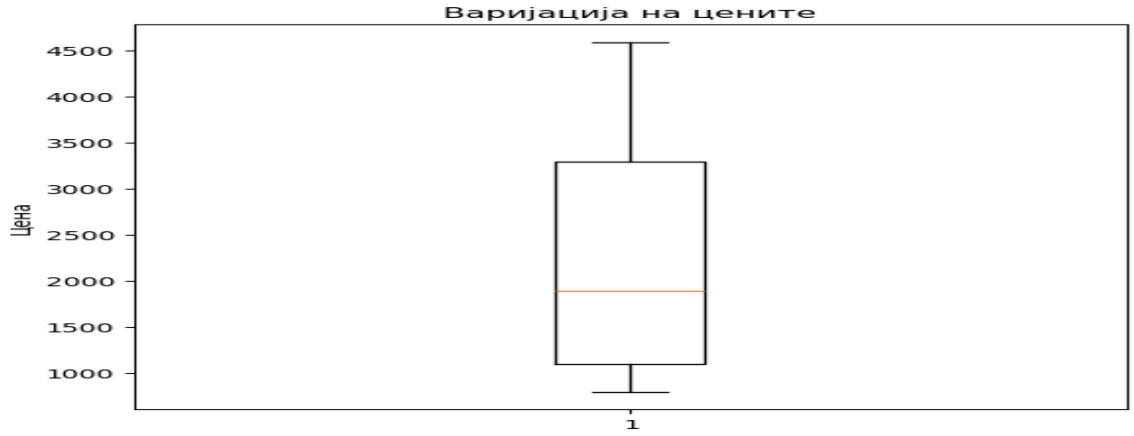
```
#Дијаграм на фреквенција на достапни величини
all_sizes = [size for sublist in cleanedData['Величини_list'] for size in sublist]
size_counts = pd.Series(all_sizes).value_counts().sort_index()

plt.figure(figsize=(8, 5))
plt.bar(size_counts.index, size_counts.values)
plt.title('Фреквенција на достапни величини')
plt.xlabel('Величина')
plt.ylabel('Број појавувања')
plt.xticks(rotation=45)
show_plot()
```



```
# 8. Бар-график: фреквенција на достапни величини
all_sizes = [size for sublist in cleaned_fsh_zenski['Величини_list'] for size in sublist]
size_counts = pd.Series(all_sizes).value_counts().sort_index()
plt.figure(figsize=(8,5))
plt.bar(size_counts.index, size_counts.values)
plt.title('Фреквенција на достапни величини')
plt.xlabel('Величина')
plt.ylabel('Број појавувања')
plt.xticks(rotation=45)
show()
```




```
# Бокс-плот на Цена
plt.figure(figsize=(6,5))
plt.boxplot(cleaned_data['Цена'].dropna(), vert=True)
plt.title('Варијација на цените')
plt.ylabel('Цена')
show()
```



Резултати и клучни наоди

Во текот на анализата на податоците од трите македонски е-продавници, дојдовме до следниве клучни наоди:

1. **Групи производи со највисоки просечни цени**

   - Највисоки просечни цени се забележуваат кај категоријата “мажи” (просечно ~6.500 ден.), следена од “жени” (просечно ~5.200 ден.) и “деца” (просечно ~4.800 ден.).
   - Овие категории најчесто се поврзани со познати светски брендови, што укажува на пониска ценовна еластичност и силна позиција на пазарот.

2. **Најчести попусти**

    - Околу 40 % од производите имаат попуст помеѓу 10 % и 20 %, додека само 15 % имаат попуст поголем од 30 %.

    - Најголемиот број попусти (над 50 %) се ретки (<5 %), што сугерира дека продавниците почесто нудат помали, но пофреквентни промоции.

3. **Најголем број производи по продавница**

ПРОДАВНИЦА	БРОЈ НА ПРОИЗВОДИ
KOSHULI.MK	127
FASHIONGROUP.COM.MK	4416
SCOUT.MK	383

Заклучок и идни чекори

Во овој проект ги направивме основните цели: собрав податоци од три македонски е-продавници, ги претпроцесирав и стандардизирав во едно единствено податочно множество и додадов колона за секоја редица во која категорија припаѓа, а потоа изведовме детална анализа и визуелизации. Како резултат, добивме:

1. **Унифицирано, квалитетено податочно множество** со конзистентни колони (`Име`, `Цена`, `Попуст`, `Бренд`, `Достапни величини` и други), подготвен за понатамошна обработка.

2. **Детални аналитички резултати** кои откриja трендови во ценовните политики, попустите и распределбата на производите по категории и продавници.

3. **Визуелни материјали** (хистограми, бокс-плотови, скатер-дијаграми, бар-графици) придружени со коментари, кои го олеснија разбирањето на наодите.

Хронолошки план на работа

1-2 Седмица - Планирање и scraping setup

• Дефинирање на бизнис цели и обем на проектот

• Избор на технологии и алатки за користење (Python, BeautifulSoup, Selenium)

• Скриптирање и тестирање на scraper-и за kosuli.mk, fashiongroup.mk и scout.mk

• Првичен преглед на квалитет на собраните податоци

3 Седмица - Претпроцесирање (Data Cleaning)

• Идентификација и отстранување на дупликати, празни вредности и невалидни записи

• Нормализација на цени и текст (регуларни изрази)

• Валидација на URL і линкови и корекција на грешки

4 Седмица - Стандардизирање

Преименување на колони во унифицирани имиња („price“, „discount“)

• Спојување на сите поддатасети

• Проверка на конзистентност на валути и формати

5-6 Седмица - Анализа и првични графици

• Групирање и агрегација (просечна цена по категорија/продавница)

• Креирање хистограми, бокс-плотови и scatter-дијаграми (Matplotlib)

• Пред-преглед на интерактивни графици со Plotly Express

7 Седмица - Финализирање визуелизации и документација

Уредување на графики (наслови, оски, легенди, боја) за чист дизајн

• Вметнување на примери од код и скриншоти (снп-но обработка)

• Пишување текстови за секции: Вовед, Методологија, Резултати, Заклучок

8 Седмица - Завршни ревизии и предавање

Крос-проверка на содржина, корекција на грешки

• Конверзија во финален PDF

• Презентација/предавање до ментор