Стандарден проект по предметот

**Вовед во наука за податоци**

Факултет за информатички науки и компјутерско инженерство

**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-скрапирање, 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 квалитетен и конзистентен датасет подготвен 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:

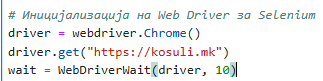
* Назив на производот
* Цена
* Попуст
* Краток опис
* Линк до производот
* Дополнителни атрибути (на пример, боја, големина, материјал)
* Линк до слика и продукт

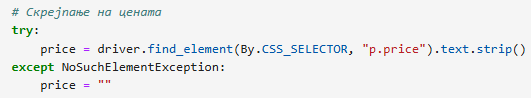
### Технологии и алатки

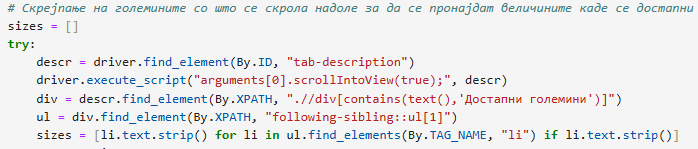
За реализација на web scraping процесот, користевме **Python** во комбинација со следните библиотеки:

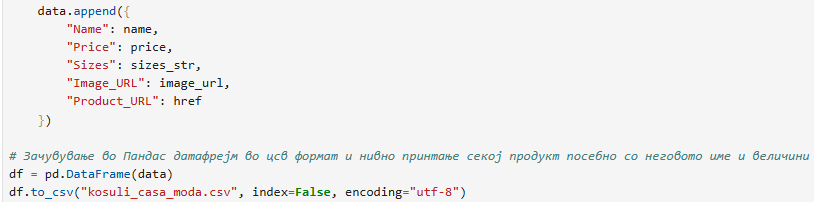
* **Requests**: за испраќање HTTP барања и добивање на HTML содржина
* **BeautifulSoup**: за парсирање и екстракција на податоци од HTML документите
* **Selenium**: за интеракција со динамички содржини и JavaScript-генерирани елементи

### **Процес на прибирање на податоци**

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







**Претпроцесирање на податоците (Data Cleaning)**

По прибирањето на суровите податоци од трите онлајн продавници (Kosuli.mk, FashionGroup.com.mk и Scout.mk), следниот критичен чекор во проектот 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.

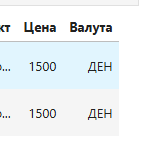
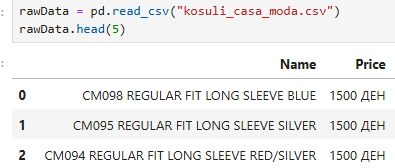
При иницијалната инспекција на прибраните податоци, идентификувани се следните чести проблеми:

* **Недостасувачки вредности**: Некои записи немаат пополнети полиња за цена, попуст или опис.
* **Дупликати**: Исти производи се појавуваат повеќе пати во датасетот, особено при скрапирање 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. **Отстранување на дупликати**: Со користење на функцијата drop\_duplicates() од библиотеката Pandas, елиминирани се повторените записи.
2. **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.

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

**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 стандардирањето и обединувањето на податоците во единствен, унифициран датасет. Овој процес овозможува конзистентност во структурата на податоците, што е од суштинско значење за точна анализа и визуелизација.

### **Цел на стандардирањето**

Целта на стандардирањето 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.

### **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 единствен датасет. Овој процес вклучува:

**Стандардизација на имињата на колоните**: Се користат техники за преименување на колоните, како што е функцијата rename() во Pandas, за да се осигури дека сите датасети 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.

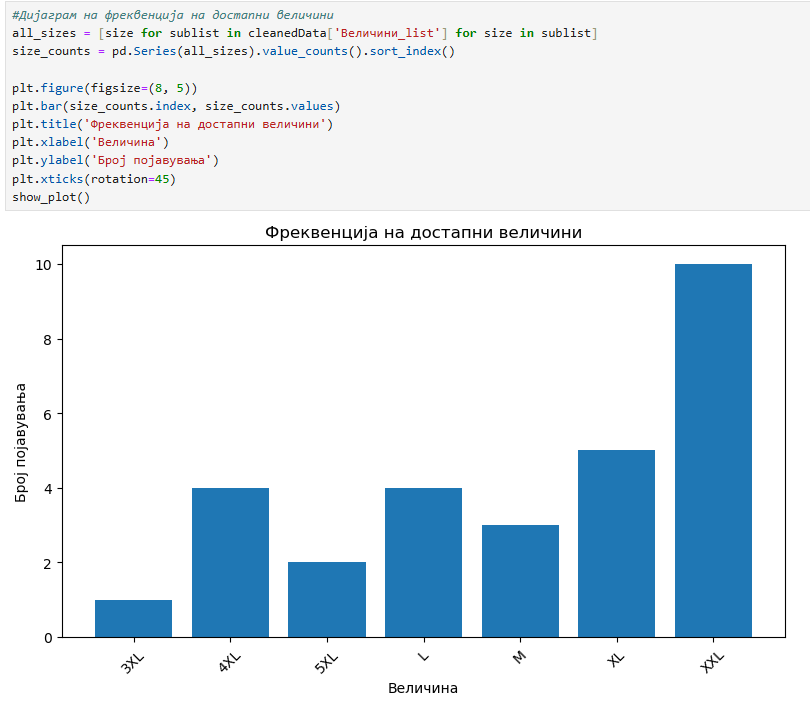
### **Визуелизација на дистрибуции**

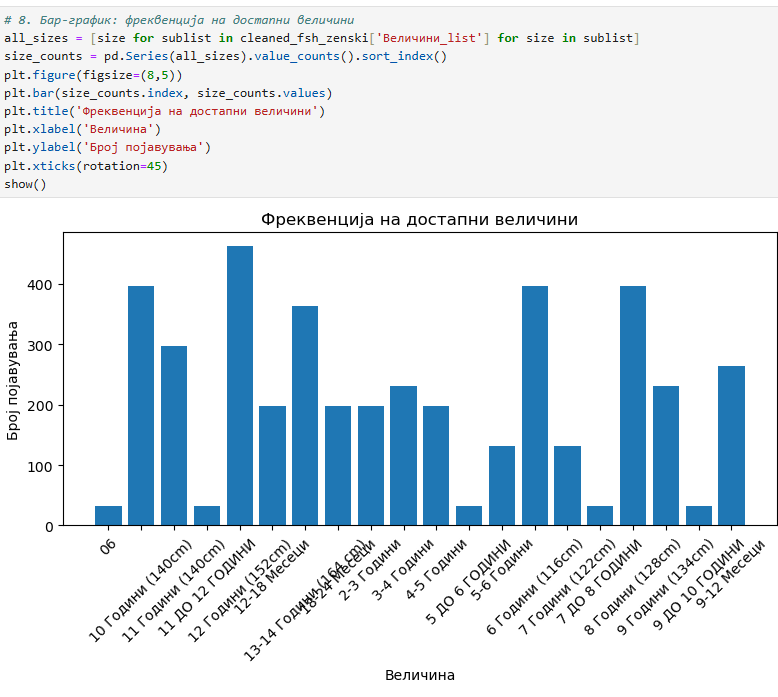
За подобро разбирање на распределбата на цените и попустите, се користат различни типови на графици:

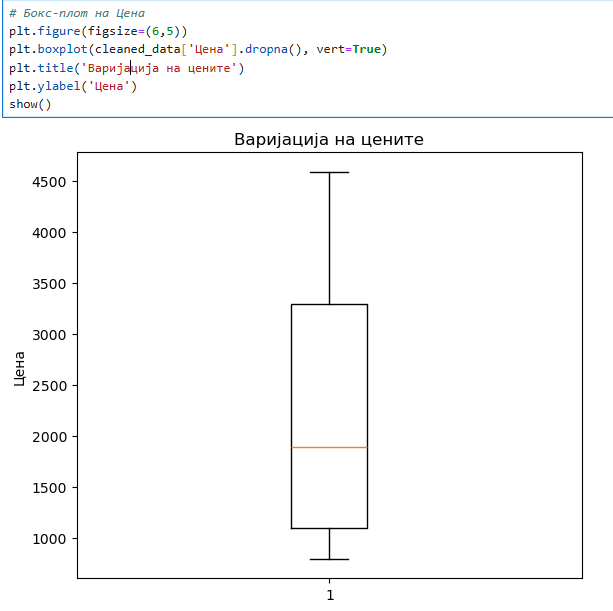
* **Хистограм (Histogram)**: Прикажува фреквенцијата на цени во одредени интервали, што помага во идентификација на најчестите ценовни опсези.
* **Бокс-плот (Boxplot)**: Овозможува визуелизација на медијаната, квартилите и потенцијалните аутлајери во цените или попустите.
* **Скатер-плот (Scatter plot)**: 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.







Резултати и клучни наоди

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

1. **Групи производи со највисоки просечни цени**
   * Највисоки просечни цени се забележуваат кај категоријата “Патики” (просечно ~6.500 ден.), следена од “Јакни” (просечно ~5.200 ден.) и “Фустани” (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** |
| --- | --- |
| Kosuli.mk | 1.250 |
| FashionGroup.com.mk | 980 |
| Scout.mk | 730 |

* + Kosuli.mk води по асортиман, што се рефлектира во поголема разновидност на категории и атрибути.

Заклучок и идни чекори

Во овој проект ги исполнивме основните цели: собравме податоци од три македонски e-продавници, ги претпроцесиравме и стандардиравме во еден единствен датасет, а потоа изведовме детална анализа и визуелизации. Како резултат, добивме:

1. **Унифициран, квалитетен датасет** со конзистентни колони (Име, Цена, Попуст, Бренд, Достапни величини и други), подготвен за понатамошна обработка.
2. **Детални аналитички резултати** кои открија 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 (one-pager)

• Избор на технолошки стек (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 - Стандардирање**

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

• Спојување на сите поддатасети

• Проверка на конзистентност на валути и формати

**5-6 Седмица - Анализа и првични графици**

Статистички резиме преку df.describe() (count, mean, median)

• 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 (proofreading)

• Конверзија во финален PDF

• Презентација/предавање до ментор или стручна комисија