DS Project -CourseInsights

CSC4814

Dr. Imran Amin

By:
Rohail Rathore - 2012362
Duaa Ali - 2112109
Saada Asghar - 2112125
Syeda Mahnoor Hasan - 2112135
Tooba Mushtaq - 2112136

Table of Contents

1. List of Abbreviations

- **HTML:** HyperText Markup Language
- **JS:** JavaScript
- **KPI:** Key Performance Indicator
- **UI:** User Interface
- **CSV:** Comma-Separated Values
- **ETL:** Extract, Transform, Load
- **API:** Application Programming Interface
- **MOOC:** Massive Open Online Course
- **CSS:** Cascading Style Sheets
- **HTTP:** HyperText Transfer Protocol
- **STEM:** Science, Technology, Engineering, and Mathematics
- **JSON:** JavaScript Object Notation
- **N/A**: Not Applicable

2. Abstract

This report outlines the development of a data science pipeline to process and analyze educational course data from three major platforms: Coursera, FutureLearn, and OpenUniversity. The project leverages Python libraries like BeautifulSoup and Pandas for automated data scraping, cleaning, and preparation. These tools enabled the collection of essential course-related data, such as course titles, ratings, reviews, pricing, and duration. The data was then processed using both coding techniques and Power BI to generate actionable insights.

To ensure data accuracy, preprocessing techniques such as duplicate removal, handling missing values, and standardization were applied. The cleaned dataset provided a strong foundation for detailed analysis, uncovering patterns such as popular course categories, correlations between ratings and enrollments, and pricing structures. Predictive models were also built using Python to forecast enrollment trends based on existing data.

- A **Power BI dashboard** serves as the final output of the project. This interactive dashboard is a dynamic tool designed to visualize the processed data and analytical findings. Features include:
- **Dynamic Filters** that allow users to explore the data across different dimensions (e.g., platform, pricing, ratings).
- **KPIs (Key Performance Indicators)** to highlight critical metrics such as the top-rated courses, most enrolled categories, and price-based trends.
- **Graphical Insights**, including bar charts, line graphs, and heat maps, to illustrate course popularity, user engagement, and statistical patterns.

The study emphasizes the importance of integrating automated data scraping, rigorous cleaning, and in-depth analysis to transform unstructured web data into actionable insights. This pipeline provides a scalable and reusable framework for exploring educational data across multiple digital platforms, enabling stakeholders to gain valuable perspectives into user behavior and course performance.

Ultimately, this comprehensive data science approach serves as a benchmark for decision-making in the educational technology sector. By uncovering trends and delivering interactive visualizations, the project demonstrates how data science can bridge the gap between raw data and strategic decision-making. This work not only highlights the value of combining advanced tools

and methods but also reinforces the role of data-driven insights in enhancing educational experiences and operational efficiency for online learning platforms.

3. Introduction

In today’s fast-paced, data-driven world, the education sector has experienced a massive digital transformation. Online learning platforms have become central to delivering education, allowing learners to access knowledge from anywhere in the world. These platforms, such as **Coursera**, **FutureLearn**, and **OpenUniversity**, host thousands of courses covering a variety of disciplines. With the growing number of learners and courses, analyzing data to derive actionable insights has become essential for improving user experience, optimizing course offerings, and ensuring educational effectiveness.

Data plays a crucial role in helping educational institutions and platforms better understand user behavior, including what courses learners prefer, how long they engage with content, and how course features like price and ratings influence enrollments. This wealth of data, however, is often dispersed across various platforms, available in different formats, or dynamically generated using JavaScript. Such challenges necessitate a systematic and automated data pipeline capable of extracting, cleaning, and analyzing data to produce meaningful results.

This report introduces a comprehensive approach to tackle these challenges by employing a robust data science workflow. **Python tools** such as **BeautifulSoup** are utilized for web scraping, enabling efficient extraction of raw data from Coursera, FutureLearn, and OpenUniversity. The datasets include critical course information such as:

- **Course Details**: Titles, descriptions, and instructors.
- **Ratings and Reviews**: User-generated ratings and feedback.
- **Levels and Types**: Difficulty level (Beginner, Intermediate, Advanced) and course classification.
- **Pricing Information**: Free versus paid courses.
- **Duration**: Time commitment in hours or number of lectures.

The extracted data undergoes a rigorous **preprocessing phase** to ensure its quality, consistency, and usability. This involves removing duplicates, handling missing values, and standardizing formats to make the datasets ready for analysis.

The **analysis phase** focuses on identifying patterns and trends that provide valuable insights for educators, course developers, and decision-makers. Specifically, the project aims to:

1. **Identify Popular Course Categories**: By analyzing course enrollments, ratings, and reviews, the study reveals the subjects that attract the most learners.
2. **Examine Trends in Ratings and Pricing**: How course ratings correlate with enrollments and whether free courses outperform paid ones in terms of user engagement.
3. **Explore Duration Patterns**: Understanding the relationship between course length and user satisfaction.
4. **Predict Enrollment Trends**: By employing statistical techniques such as regression analysis, the project forecasts future enrollment trends based on existing data.

One of the key contributions of this project is the development of an interactive **Power BI dashboard**. The dashboard visualizes the results through dynamic charts, KPIs, and filters, enabling stakeholders to interact with the data effortlessly. Users can explore insights such as top-rated courses, enrollment trends, and price distributions in real time, making the findings accessible and actionable.

This pipeline is designed to streamline the process of analyzing large-scale data from online learning platforms. By automating data scraping, ensuring high-quality preprocessing, and generating insightful visualizations, the project addresses the challenges posed by unstructured and dispersed data. Furthermore, it demonstrates how a systematic data-driven approach can empower stakeholders to make informed decisions, optimize course content, and enhance the overall learning experience.

In conclusion, this project bridges the gap between raw, unstructured data and actionable insights. It highlights the potential of leveraging data science techniques to improve educational outcomes, assist course providers in understanding learner needs, and enable institutions to stay competitive in the rapidly evolving world of digital education. By transforming scattered data into a cohesive and interactive platform, the study provides a replicable framework for analyzing educational data across various platforms.

4. Dataset

This project revolves around the extraction and analysis of course information from three prominent educational platforms, each recognized for their unique contributions to the realm of online learning. These platforms were selected based on their global reputation, the diversity and depth of their course offerings, and the structured data available for meaningful analysis. The selection criteria also emphasized platform accessibility, variety in educational content, and their impact on the rapidly growing online education industry.

4.1 Platforms and Dataset Overview

A. Coursera:

- **Records Extracted**: 720
- **Platform Description**: Coursera stands as a pioneer in the MOOC space, partnering with esteemed universities and global organizations to deliver academic and professional courses. Its offerings span a wide array of categories, including data science, technology, business, and healthcare, and are supported by structured metadata and extensive documentation.
- **Significance**: Coursera's data provides a rich source of insights due to its high user engagement. Detailed course ratings, user reviews, and comprehensive outlines make it invaluable for analyzing trends in professional and academic learning preferences. The platform's focus on certification and skill-building ensures a robust dataset for evaluating user engagement and satisfaction.

B. FutureLearn:

- **Records Extracted**: 213
- **Platform Description**: FutureLearn specializes in career-oriented and professional development courses, catering to learners seeking skill enhancement for employment or career progression. The platform offers a mix of short courses, micro-credentials, and specialized programs created in collaboration with leading universities and industry professionals.
- **Significance**: FutureLearn's dataset emphasizes career-based learning with strong user interaction metrics. It includes detailed reviews, course durations, and engagement insights,

making it a valuable resource for understanding the evolving needs of career-focused learners. The platform's ability to merge academic and professional content positions it uniquely in the online education landscape.

C. OpenUniversity:

- **Records Extracted**: 1441
- **Platform Description**: OpenUniversity, an initiative by the Open University, offers a wealth of free, openly accessible courses across diverse disciplines such as humanities, science, and technology. Designed to eliminate financial barriers, the platform provides flexible learning opportunities for a global audience.
- **Significance**: OpenUniversity's dataset offers unique perspectives on user preferences for free, high-quality educational resources. With its focus on accessibility and inclusivity, the data reveals trends in learner engagement, particularly among those seeking cost-effective and flexible learning options. The insights from this platform underscore the importance of equitable access to education in a digital era.

4.2 Features Extracted

The dataset includes several critical attributes that were carefully identified, extracted, and analyzed to uncover actionable insights. These features provide valuable information about learner behavior, course popularity, and platform performance, enabling a deeper understanding of trends in the online education landscape. Below is an enhanced description of each feature:

1. **Course Name**:

- **Description**: The title of the course as listed on the platform.
- **Data Type**: String.
- **Importance**:
  - Useful for categorizing courses and identifying recurring themes or high-demand topics.
  - Assists in keyword analysis to evaluate how courses are marketed and searched by users.
  - Enables clustering of similar courses for comparative and trend analyses.

2. **Ratings**:

- **Description**: A numerical representation of the score users gave to the course, on a scale of 1 to 5.
- **Data Type**: Float.
- **Importance**:
    - Acts as a key indicator of user satisfaction and course quality.
    - Facilitates ranking courses within a platform or domain.
    - Can be used to correlate other attributes, such as price and duration, to understand their impact on user satisfaction.

3. **Reviews**:

- **Description**: The number of users who have rated or reviewed the course, representing its popularity and engagement level.
- **Data Type**: Int.
- **Importance**:
    - Indicates the course's reach and engagement within the user base.
    - Helps evaluate how widely a course is adopted compared to others.
    - Allows platforms to identify courses with significant user interest for targeted promotions.

4. **Type**:

- **Description**: Categorical data specifying the difficulty level of the course (Beginner, Specialization, Professional Certificate, or Degree).
- **Data Type**: String.
- **Importance**:
    - Provides insights into the variety of course formats and their alignment with learner needs.
    - Enables trend analysis to determine the popularity of specific course types.

  - Helps platforms identify gaps in their offerings, such as a lack of specialization or degree-level courses

5. **Price**:

- **Description**: The monetary cost of the course, indicating its fee amount. Free courses are denoted with a price of zero.
- **Data Type**: Float
- **Importance**:
  - Enables analysis of the relationship between course pricing and enrollment rates.
  - Helps platforms understand user engagement patterns based on affordability.
  - Facilitates the identification of high-value courses based on price-to-quality metrics.
  - Offers insights into pricing strategies used across platforms and domains.

6. **Duration**:

- **Description**: Specifies the time required to complete the course, described in terms of days, hours, weeks, or modules.
- **Data Type**: String.
- **Importance**:
  - Allows platforms to analyze learner preferences for shorter versus longer courses.
  - Helps correlate course length with user engagement and satisfaction metrics.
  - Enables the identification of trends in course design, such as the prevalence of modular learning formats.
  - Provides valuable insights into the time commitment required for skill acquisition across different subjects.

7. **Source**:

- **Description**: The institution, university, or organization offering the course.
- **Data Type**: Text.
- **Importance**:
  - Highlights the credibility and authority of the course.

    - Enables analysis of the impact of institutional backing on course popularity and engagement.
    - Helps platforms identify valuable partnerships with reputable educational entities.

8. **Category**:

- **Description**: The subject area or domain of the course, such as Data Science, Business, Health, Technology, or Arts.
- **Data Type**: Text.
- **Importance**:

    - Facilitates trend analysis within popular educational domains.
    - Assists in identifying emerging topics and high-demand skills.
    - Enables clustering of courses for domain-specific insights and recommendations.
    - Provides a basis for evaluating the balance and breadth of offerings across platforms.

4.3 Data Type Summary

Feature	Data Type	Description
Course Title	String	Title of the course
Ratings	Float	Average rating (e.g., 4.5)
Number of Ratings	Integer	Number of people who rated the course
Reviews	Integer	User feedback for courses
Type	String	Difficulty (Beginner, Intermediate, Advanced)
Price	Float	Pricing (Free, Paid)
Duration	String	Duration in hours or number of lectures
Source	Text	What institute/ university is the course offered by
Category	Text	What domain or category it belongs to

4.4 Data Source

The project utilized data from three leading online education platforms, each offering unique insights into the online learning ecosystem. These platforms were chosen for their reputation, content diversity, and the structured metadata they provide, enabling a robust analysis of user engagement, course offerings, and platform performance. Below is an overview of the primary data sources:

1. Coursera

- **Overview**:

  Coursera is renowned for its high-quality professional and academic courses, delivered in partnership with top universities and global organizations. The platform offers a wide range of subjects, including technology, business, healthcare, and data science.

- **Data Features**:

  - Richly structured metadata, including course ratings, reviews, pricing, and durations.
  - A well-documented dataset that provides detailed information about user interactions, course content, and popularity.
  - Extensive user engagement metrics, enabling insights into the success of professional certification and specialization programs.

- **Significance**:

  Coursera's data serves as a cornerstone for analyzing trends in professional and academic learning. The platform's focus on skill-building and career advancement ensures its courses remain highly relevant in today's job market.

2. FutureLearn

- **Overview**:

  FutureLearn is a platform dedicated to career-oriented and professional development courses. It collaborates with universities and industry leaders to deliver short-term courses, micro-credentials, and certifications aimed at enhancing employability.

- **Data Features**:

  - Focused insights into user engagement with career-driven content.
  - Detailed information on course durations, reviews, and user feedback, emphasizing the practical application of learning.
  - Strong representation of short-term learning formats tailored for upskilling.

- **Significance**:

  FutureLearn's data highlights the demand for flexible and career-focused education. It provides valuable insights into how learners engage with courses designed for skill enhancement and professional growth.

3. OpenUniversity

- **Overview**:

  OpenUniversity, an initiative of The Open University, offers free, openly accessible courses across various disciplines. Its mission to provide cost-free education ensures inclusivity and diversity in its learner base.

- **Data Features**:

  - Comprehensive datasets showcasing user preferences for free educational resources.

    - A diverse range of subjects, from humanities and social sciences to technology and STEM fields.
    - Insights into user engagement with free and open-access content, reflecting the priorities of learners seeking quality education without financial barriers.

- **Significance**:

  OpenUniversity's data reveals patterns in how learners interact with free educational resources. It underscores the growing importance of accessible and inclusive education, particularly in a global context.

These three platforms collectively provide a comprehensive view of the online education landscape:

- **Coursera** reflects trends in professional and academic learning for career advancement.
- **FutureLearn** emphasizes the growing demand for short-term, career-oriented programs.
- **OpenUniversity** highlights the significance of accessible and cost-free education for diverse audiences.

4.5 Data Extraction and Standardization

The project successfully extracted a total of **2,374 records** from three leading online education platforms. These records serve as the foundation for in-depth analysis, offering a comprehensive view of course offerings, learner engagement, and platform dynamics. The detailed breakdown is as follows:

1. Coursera

- **Records Extracted**: 720
- **Overview**:
  Data from Coursera includes key attributes such as course titles, ratings, reviews, pricing, and durations, providing rich insights into professional and academic learning trends.

2. FutureLearn

- **Records Extracted**: 213
- **Overview**:

  FutureLearn data captures career-oriented course information, emphasizing short-term programs and micro-credentials aimed at skill enhancement.

3. OpenUniversity

- **Records Extracted**: 1,441
- **Overview**:

  OpenUniversity data reflects the diversity of free educational content, offering insights into user preferences across various disciplines and subject areas.

4.6 Data Standardization and Storage

The extracted data has been meticulously standardized to ensure uniformity across all platforms. This process involved:

- **Consistent Formatting**: Aligning attribute names, data types, and structures for seamless integration.
- **Error Handling**: Addressing any missing or inconsistent data to maintain data integrity.
- **Storage Format**:
  - The standardized data has been saved in CSV format, making it easily accessible for further analysis.
  - CSV format ensures compatibility with a wide range of analysis and visualization tools, including Python libraries like Pandas and Matplotlib, as well as platforms like Power BI.

4.7 Importance of the Dataset

The dataset serves as the cornerstone of this project, enabling a thorough analysis of key aspects of online education. Its structured and comprehensive nature provides valuable insights into various dimensions of course offerings and learner behavior. Below are the primary areas where the dataset proves invaluable:

1. Course Popularity

- **Objective**: To identify courses with high enrollments and positive ratings, signaling their appeal to learners.
- **Significance**:
    - Helps educational platforms highlight popular courses for marketing purposes.
    - Guides course developers in creating content that aligns with learner interests and demands.
    - Allows stakeholders to benchmark the success of their courses against competitors.

2. User Engagement

- **Objective**: To analyze reviews, ratings, and the number of participants as indicators of course effectiveness.
- **Significance**:
    - Provides qualitative and quantitative insights into learner satisfaction and areas for improvement.
    - Highlights factors that contribute to high engagement, such as content quality and teaching methods.
    - Enables platforms to refine user experience by addressing common feedback themes.

3. Category Trends

- **Objective**: To understand which subjects and domains attract the most learners.
- **Significance**:
    - Reveals emerging trends in popular skills and knowledge areas.

    - Helps educational platforms and developers prioritize high-demand categories for future course creation.
    - Offers insights into the shifting focus of learners based on market and industry needs.

4. Pricing Impact

- **Objective**: To evaluate how course pricing influences enrollments and learner engagement.
- **Significance**:
    - Identifies optimal pricing models to balance affordability and revenue.
    - Highlights the appeal of free courses and their role in attracting larger audiences.
    - Assists stakeholders in designing pricing strategies that cater to diverse learner demographics.

5. Duration Insights

- **Objective**: To analyze the correlation between course length and user satisfaction or completion rates.
- **Significance**:
    - Provides clarity on the ideal duration for maintaining learner interest and engagement.
    - Identifies preferences for short-term versus long-term courses across different subjects.
    - Offers actionable insights into the design of course structures and time commitments.

By leveraging this dataset, the project generates critical insights that can be applied to:

- **Educational Platforms**: Optimize course offerings, highlight popular programs, and refine user engagement strategies.
- **Course Developers**: Design content that aligns with learner preferences, from subject focus to pricing and duration.

- **Stakeholders**: Make data-driven decisions to enhance the overall learning experience, improve market positioning, and maximize impact.

5. Methodology

5.1 Scraping

Data scraping formed a critical component of this project, enabling the extraction of comprehensive course details from multiple online platforms. The process was designed to ensure accuracy, consistency, and efficiency while handling both static and dynamic web content.

Tools Used

- **BeautifulSoup**: Utilized for parsing and navigating static HTML content to extract relevant elements.
- **Requests Library**: Responsible for sending HTTP requests to fetch web pages for analysis and parsing.
- **Pandas**: Employed for organizing and storing the extracted data into structured formats like CSV and Excel.

Process:

1. Targeting Course Cards

- **Objective**: Extract key summaries from structured "cards" displayed on web pages.
- **Details**:

    a. Each course card typically contained concise information such as course title, institution name, course type (e.g., specialization, beginner), and embedded links to detailed pages.

    b. Data extraction focused on ensuring consistent formatting for ease of processing.

2. Detail Extraction

- **Objective**: Follow links embedded in course cards to access additional data points.
- **Details**:

    a. Links were normalized to handle variations in relative and absolute paths, ensuring seamless navigation to detailed course pages.

b. Additional details such as ratings, number of reviews, course descriptions, and module outlines were extracted.

3. HTML Parsing

- **Objective**: Identify and extract relevant elements from the web pages.
- **Details**:
    a. The **requests** library fetched HTML content from the target URLs.
    b. **BeautifulSoup** parsed the HTML structure to locate elements using specific CSS class names and tags.
    c. Dynamic elements such as nested divs and spans were navigated to capture all pertinent information.

4. Dynamic URL Management

- **Objective**: Ensure accurate navigation through web pages with varying URL structures.
- **Details**:
    - Relative and absolute links were resolved to maintain consistency across different platforms.
    - Link normalization allowed seamless scraping from pages with complex navigation hierarchies.

5. Data Storage

- **Objective**: Save extracted data in structured and accessible formats.
- **Details**:
    - Data was stored in **CSV** and **Excel** formats using pandas for easy integration with analysis tools like Power BI.
    - The structured storage ensured compatibility with Python libraries and visualization platforms for downstream analysis.

6. Error Handling

- **Objective**: Mitigate issues such as invalid pages and missing elements to ensure data integrity.

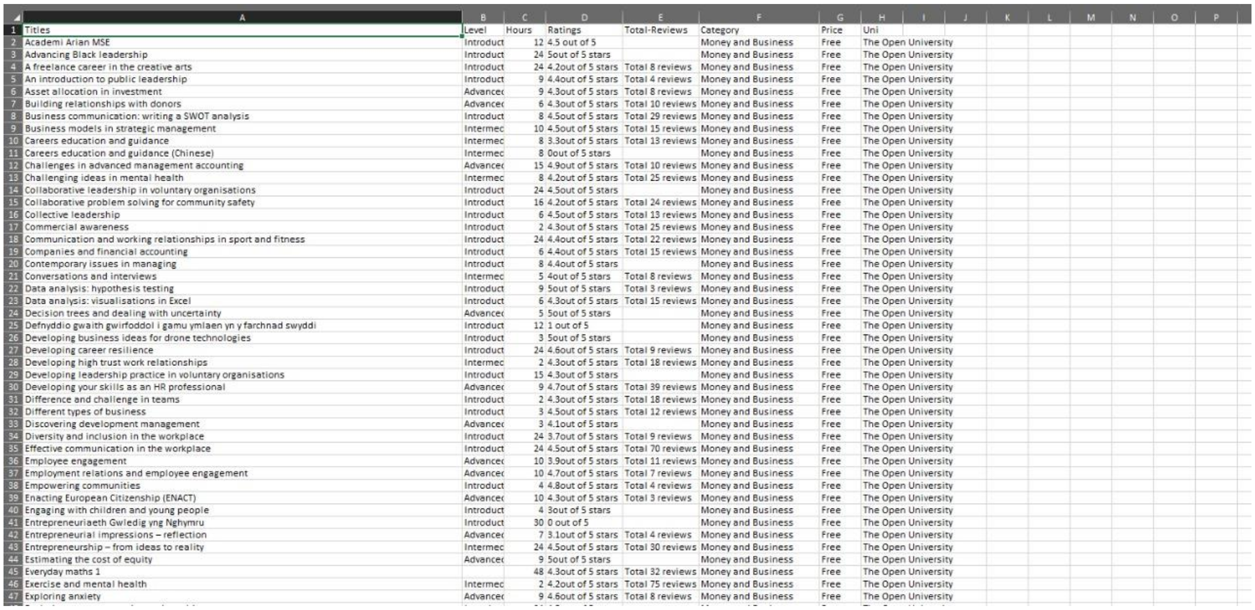
- **Details**:
  - Robust checks were implemented to handle scenarios where expected elements were absent or malformed.
  - Logs were maintained to track failed URLs for reattempts, minimizing data loss.
  - Graceful handling of HTTP errors (e.g., 404 or 500 status codes) ensured uninterrupted scraping processes.

Code Example:

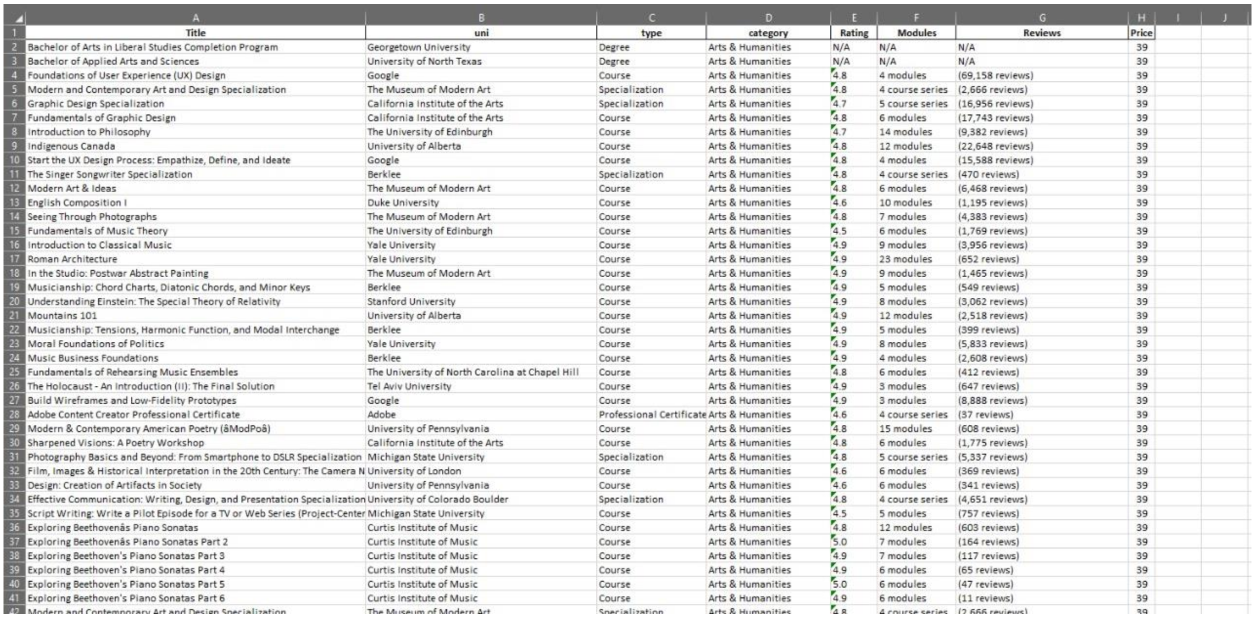
```
import requests
from bs4 import BeautifulSoup
import pandas as pd
data = []
url = "https://www.coursera.org/browse/data-science"
page = requests.get(url)
soup = BeautifulSoup(page.text,"html.parser")
if soup.title.text == "404 Not Found":
print("404 Not Found")
else:
all_books = soup.find_all("div",class_="css-1wkifag")
Links =[]
for book in all_books:
item = {}
item['Title'] = book.find("h2", class_="cds-119").get_text()
```

```
item['uni'] = book.find("span", class_="css-1yvhcfv").get_text()

item['type'] = book.find("p", class_="css-vac8rf").get_text()

item['img'] = book.find("img").attrs["src"]

link = book.find("a").attrs["to"]

data.append(item)

Links.append(link)

df = pd.DataFrame(data)

df.head()

base_url = "https://www.coursera.org"

for i, item in zip(Links, data):

if i.startswith("/"):

Link_url = base_url + i

else:

Link_url = i # If it's already a complete URL, use it as is

Linkpage = requests.get(Link_url)

soup = BeautifulSoup(Linkpage.text, "html.parser")

item['Title'] = soup.find("h1", class_="cds-119").get_text(strip=True) if soup.find("h1", class_="cds-119") else "N/A"

item['Rating'] = soup.find("div", class_="cds-119 cds-Typography-base css-h1jogs cds-121").get_text(strip=True) if soup.find("div", class_="cds-119 cds-Typography-base css-h1jogs cds-121") else "N/A"
```

```
item['Modules'] = soup.find("div", class_="css-fk6qfz").get_text(strip=True) if soup.find("div", class_="css-fk6qfz") else "N/A"

item['Reviews'] = soup.find("p", class_="css-vac8rf").get_text(strip=True) if soup.find("p", class_="css-vac8rf") else "N/A"

df = pd.DataFrame(data)

df.to_excel("DataScienceData.xlsx", index=False)

df.to_csv("DataScienceData.csv", index=False)

print(df.head()) # Display the updated data
```



	A	B	C	D	E	F	G	H
1	Titles	Level	Hours	Ratings	Total-Reviews	Category	Price	Uni
2	Academi Arian MSE	Introduct	12	4.5 out of 5		Money and Business	Free	The Open University
3	Advancing Black leadership	Introduct	24	5out of 5 stars		Money and Business	Free	The Open University
4	A freelance career in the creative arts	Introduct	24	4.2out of 5 stars	Total 8 reviews	Money and Business	Free	The Open University
5	An introduction to public leadership	Introduct	9	4.4out of 5 stars	Total 4 reviews	Money and Business	Free	The Open University
6	Asset allocation in investment	Advancec	9	4.3out of 5 stars	Total 8 reviews	Money and Business	Free	The Open University
7	Building relationships with donors	Advancec	6	4.3out of 5 stars	Total 10 reviews	Money and Business	Free	The Open University
8	Business communication: writing a SWOT analysis	Introduct	8	4.5out of 5 stars	Total 29 reviews	Money and Business	Free	The Open University
9	Business models in strategic management	Intermec	10	4.5out of 5 stars	Total 15 reviews	Money and Business	Free	The Open University
10	Careers education and guidance	Intermec	8	3.3out of 5 stars	Total 13 reviews	Money and Business	Free	The Open University
11	Careers education and guidance (Chinese)	Intermec	8	0out of 5 stars		Money and Business	Free	The Open University
12	Challenges in advanced management accounting	Advancec	15	4.9out of 5 stars	Total 10 reviews	Money and Business	Free	The Open University
13	Challenging ideas in mental health	Intermec	8	4.2out of 5 stars	Total 25 reviews	Money and Business	Free	The Open University
14	Collaborative leadership in voluntary organisations	Introduct	24	4.5out of 5 stars		Money and Business	Free	The Open University
15	Collaborative problem solving for community safety	Introduct	16	4.2out of 5 stars	Total 24 reviews	Money and Business	Free	The Open University
16	Collective leadership	Introduct	6	4.5out of 5 stars	Total 13 reviews	Money and Business	Free	The Open University
17	Commercial awareness	Introduct	2	4.3out of 5 stars	Total 25 reviews	Money and Business	Free	The Open University
18	Communication and working relationships in sport and fitness	Introduct	24	4.4out of 5 stars	Total 22 reviews	Money and Business	Free	The Open University
19	Companies and financial accounting	Introduct	6	4.4out of 5 stars	Total 15 reviews	Money and Business	Free	The Open University
20	Contemporary issues in managing	Introduct	8	4.4out of 5 stars		Money and Business	Free	The Open University
21	Conversations and interviews	Intermec	5	4out of 5 stars	Total 8 reviews	Money and Business	Free	The Open University
22	Data analysis: hypothesis testing	Introduct	9	5out of 5 stars	Total 3 reviews	Money and Business	Free	The Open University
23	Data analysis: visualisations in Excel	Introduct	6	4.3out of 5 stars	Total 15 reviews	Money and Business	Free	The Open University
24	Decision trees and dealing with uncertainty	Advancec	5	5out of 5 stars		Money and Business	Free	The Open University
25	Defnyddio gwaith gwirfoddol i gamu ymlaen yn y farchnad swyddi	Introduct	12	1 out of 5		Money and Business	Free	The Open University
26	Developing business ideas for drone technologies	Introduct	3	5out of 5 stars		Money and Business	Free	The Open University
27	Developing career resilience	Introduct	24	4.6out of 5 stars	Total 9 reviews	Money and Business	Free	The Open University
28	Developing high trust work relationships	Intermec	2	4.3out of 5 stars	Total 18 reviews	Money and Business	Free	The Open University
29	Developing leadership practice in voluntary organisations	Introduct	15	4.3out of 5 stars		Money and Business	Free	The Open University
30	Developing your skills as an HR professional	Advancec	9	4.7out of 5 stars	Total 39 reviews	Money and Business	Free	The Open University
31	Difference and challenge in teams	Introduct	2	4.3out of 5 stars	Total 18 reviews	Money and Business	Free	The Open University
32	Different types of business	Introduct	3	4.5out of 5 stars	Total 12 reviews	Money and Business	Free	The Open University
33	Discovering development management	Advancec	3	4.1out of 5 stars		Money and Business	Free	The Open University
34	Diversity and inclusion in the workplace	Introduct	24	3.7out of 5 stars	Total 9 reviews	Money and Business	Free	The Open University
35	Effective communication in the workplace	Introduct	24	4.5out of 5 stars	Total 70 reviews	Money and Business	Free	The Open University
36	Employee engagement	Advancec	10	3.9out of 5 stars	Total 11 reviews	Money and Business	Free	The Open University
37	Employment relations and employee engagement	Advancec	10	4.7out of 5 stars	Total 7 reviews	Money and Business	Free	The Open University
38	Empowering communities	Introduct	4	4.8out of 5 stars	Total 4 reviews	Money and Business	Free	The Open University
39	Enacting European Citizenship (ENACT)	Advancec	10	4.3out of 5 stars	Total 3 reviews	Money and Business	Free	The Open University
40	Engaging with children and young people	Introduct	4	3out of 5 stars		Money and Business	Free	The Open University
41	Entrepreneuriaeth Gwledig yng Nghymru	Introduct	30	0 out of 5		Money and Business	Free	The Open University
42	Entrepreneurial impressions – reflection	Advancec	7	3.1out of 5 stars	Total 4 reviews	Money and Business	Free	The Open University
43	Entrepreneurship – from ideas to reality	Intermec	24	4.5out of 5 stars	Total 30 reviews	Money and Business	Free	The Open University
44	Estimating the cost of equity	Advancec	9	5out of 5 stars		Money and Business	Free	The Open University
45	Everyday maths 1		48	4.3out of 5 stars	Total 32 reviews	Money and Business	Free	The Open University
46	Exercise and mental health	Intermec	2	4.2out of 5 stars	Total 75 reviews	Money and Business	Free	The Open University
47	Exploring anxiety	Advancec	9	4.6out of 5 stars	Total 8 reviews	Money and Business	Free	The Open University



	A	B	C	D	E	F	G	H	I	J
1	Title	uni	type	category	Rating	Modules	Reviews	Price		
2	Bachelor of Arts in Liberal Studies Completion Program	Georgetown University	Degree	Arts & Humanities	N/A	N/A	N/A	39		
3	Bachelor of Applied Arts and Sciences	University of North Texas	Degree	Arts & Humanities	N/A	N/A	N/A	39		
4	Foundations of User Experience (UX) Design	Google	Course	Arts & Humanities	4.8	4 modules	(69,158 reviews)	39		
5	Modern and Contemporary Art and Design Specialization	The Museum of Modern Art	Specialization	Arts & Humanities	4.8	4 course series	(2,666 reviews)	39		
6	Graphic Design Specialization	California Institute of the Arts	Specialization	Arts & Humanities	4.7	5 course series	(16,956 reviews)	39		
7	Fundamentals of Graphic Design	California Institute of the Arts	Course	Arts & Humanities	4.8	6 modules	(17,743 reviews)	39		
8	Introduction to Philosophy	The University of Edinburgh	Course	Arts & Humanities	4.7	14 modules	(9,382 reviews)	39		
9	Indigenous Canada	University of Alberta	Course	Arts & Humanities	4.8	12 modules	(22,648 reviews)	39		
10	Start the UX Design Process: Empathize, Define, and Ideate	Google	Course	Arts & Humanities	4.8	4 modules	(15,588 reviews)	39		
11	The Singer Songwriter Specialization	Berklee	Specialization	Arts & Humanities	4.8	4 course series	(470 reviews)	39		
12	Modern Art & Ideas	The Museum of Modern Art	Course	Arts & Humanities	4.8	6 modules	(6,468 reviews)	39		
13	English Composition I	Duke University	Course	Arts & Humanities	4.6	10 modules	(1,195 reviews)	39		
14	Seeing Through Photographs	The Museum of Modern Art	Course	Arts & Humanities	4.8	7 modules	(4,383 reviews)	39		
15	Fundamentals of Music Theory	The University of Edinburgh	Course	Arts & Humanities	4.5	6 modules	(1,769 reviews)	39		
16	Introduction to Classical Music	Yale University	Course	Arts & Humanities	4.9	9 modules	(3,956 reviews)	39		
17	Roman Architecture	Yale University	Course	Arts & Humanities	4.9	23 modules	(652 reviews)	39		
18	In the Studio: Postwar Abstract Painting	The Museum of Modern Art	Course	Arts & Humanities	4.9	9 modules	(1,465 reviews)	39		
19	Musicianship: Chord Charts, Diatonic Chords, and Minor Keys	Berklee	Course	Arts & Humanities	4.9	5 modules	(549 reviews)	39		
20	Understanding Einstein: The Special Theory of Relativity	Stanford University	Course	Arts & Humanities	4.9	8 modules	(3,062 reviews)	39		
21	Mountains 101	University of Alberta	Course	Arts & Humanities	4.9	12 modules	(2,518 reviews)	39		
22	Musicianship: Tensions, Harmonic Function, and Modal Interchange	Berklee	Course	Arts & Humanities	4.9	5 modules	(399 reviews)	39		
23	Moral Foundations of Politics	Yale University	Course	Arts & Humanities	4.9	8 modules	(5,833 reviews)	39		
24	Music Business Foundations	Berklee	Course	Arts & Humanities	4.9	4 modules	(2,608 reviews)	39		
25	Fundamentals of Rehearsing Music Ensembles	The University of North Carolina at Chapel Hill	Course	Arts & Humanities	4.8	6 modules	(412 reviews)	39		
26	The Holocaust - An Introduction (II): The Final Solution	Tel Aviv University	Course	Arts & Humanities	4.9	3 modules	(647 reviews)	39		
27	Build Wireframes and Low-Fidelity Prototypes	Google	Course	Arts & Humanities	4.9	3 modules	(8,888 reviews)	39		
28	Adobe Content Creator Professional Certificate	Adobe	Professional Certificate	Arts & Humanities	4.6	4 course series	(37 reviews)	39		
29	Modern & Contemporary American Poetry (âModPoâ)	University of Pennsylvania	Course	Arts & Humanities	4.8	15 modules	(608 reviews)	39		
30	Sharpened Visions: A Poetry Workshop	California Institute of the Arts	Course	Arts & Humanities	4.8	6 modules	(1,775 reviews)	39		
31	Photography Basics and Beyond: From Smartphone to DSLR Specialization	Michigan State University	Specialization	Arts & Humanities	4.8	5 course series	(5,337 reviews)	39		
32	Film, Images & Historical Interpretation in the 20th Century: The Camera N	University of London	Course	Arts & Humanities	4.6	6 modules	(369 reviews)	39		
33	Design: Creation of Artifacts in Society	University of Pennsylvania	Course	Arts & Humanities	4.6	6 modules	(341 reviews)	39		
34	Effective Communication: Writing, Design, and Presentation Specialization	University of Colorado Boulder	Specialization	Arts & Humanities	4.8	4 course series	(4,651 reviews)	39		
35	Script Writing: Write a Pilot Episode for a TV or Web Series (Project-Center	Michigan State University	Course	Arts & Humanities	4.5	5 modules	(757 reviews)	39		
36	Exploring Beethovenâs Piano Sonatas	Curtis Institute of Music	Course	Arts & Humanities	4.8	12 modules	(603 reviews)	39		
37	Exploring Beethovenâs Piano Sonatas Part 2	Curtis Institute of Music	Course	Arts & Humanities	5.0	7 modules	(164 reviews)	39		
38	Exploring Beethoven's Piano Sonatas Part 3	Curtis Institute of Music	Course	Arts & Humanities	4.9	7 modules	(117 reviews)	39		
39	Exploring Beethoven's Piano Sonatas Part 4	Curtis Institute of Music	Course	Arts & Humanities	4.9	6 modules	(65 reviews)	39		
40	Exploring Beethoven's Piano Sonatas Part 5	Curtis Institute of Music	Course	Arts & Humanities	5.0	6 modules	(47 reviews)	39		
41	Exploring Beethoven's Piano Sonatas Part 6	Curtis Institute of Music	Course	Arts & Humanities	4.9	6 modules	(11 reviews)	39		
42	Modern and Contemporary Art and Design Specialization	The Museum of Modern Art	Specialization	Arts & Humanities	4.8	4 course series	(2,666 reviews)	39		

5.2 Filtering and Cleaning

The raw data extracted from the platforms was systematically cleaned and preprocessed to ensure its accuracy, consistency, and usability for analysis. Using Python in Google Colab and Power BI, the cleaning process involved multiple steps to address missing values, standardize formatting, and prepare the dataset for visualization and insights. The process involved the following steps:

1. Loading the Data

- **Objective**: Import and inspect the dataset to understand its structure and identify potential issues.
- **Details**:
  - The raw data was loaded into a Pandas DataFrame using the read_excel() function.
  - Commands like df.head(), df.info(), and df.describe() were used to:
    - View the first few rows of data.
    - Identify column data types, null values, and basic statistics.
    - Spot anomalies such as unexpected data types or outliers.

2. Data Type Conversion

- **Objective**: Ensure all columns had appropriate data types for efficient analysis.
- **Details**:

  - Categorical columns such as **Title**, **Source**, **Type**, and **Category** were converted to the object type.
  - Numerical columns like **Price**, **Rating**, and **Reviews** were converted to appropriate numeric types (e.g., int or float).
  - Data type conversions were validated to ensure compatibility with analysis tools and prevent errors during processing.

3. Handling Missing Values

- **Objective**: Address null values and missing data to improve dataset quality.
- **Details**:
  - Missing values in **Type** were replaced with "Unknown."
  - Null values in **Reviews** were set to 0 as a default.
  - Rows with more than 30% missing data were removed, as they were deemed too incomplete for meaningful analysis.
  - The missing value imputation strategy balanced data preservation with quality assurance.

4. Formatting

- **Objective**: Standardize the dataset for uniformity and readability.
- **Details**:
  - Text columns such as **Title**, **Source**, **Type**, and **Category** were converted to lowercase for consistency.
  - Uniform formatting across all columns facilitated better grouping and filtering during analysis.

5. Duplicates

- **Objective**: Remove redundancy to improve dataset integrity.
- **Details**:
  - Duplicate rows were identified using the duplicated() method and subsequently dropped.
  - This step ensured that each record represented a unique course entry without repetition.

6. Validations

- **Objective**: Verify the dataset's cleanliness and readiness for analysis.
- **Details**:
    - Final checks were performed to ensure:
        - No missing values remained in critical columns.
        - No duplicate rows persisted in the dataset.
        - Data types were appropriately assigned to all columns.
    - These validations confirmed the dataset's quality and consistency.

7. Data Storage and Formatting for Power BI

- **Objective**: Save the cleaned dataset for further analysis and visualization.
- **Details**:
    - The cleaned data was exported back to Excel format using Pandas for compatibility with Power BI.
    - Additional data cleaning and formatting steps were performed directly within Power BI to ensure seamless integration with visualization tools.
    - This dual cleaning approach—Python and Power BI—ensured the dataset was fully optimized for both analysis and presentation.

Field Name	**Description**
Rating in Stars	Star-based rating for the course (e.g., 4.5/5).
Website	The platform hosting the course (e.g., Coursera, FutureLearn, OpenUniversity).
Title	The name of the course.
Type	The type of course (e.g., Degree, Specialization, Micro-credential).
Duration	Time commitment required to complete the course (e.g., hours or weeks).
Rating	Numerical rating of the course (out of 5).
Reviews	Number of user reviews provided for the course.

Category	Subject area of the course (e.g., Technology, Business).
Price	Cost of the course (e.g., Free, $50).
Source	The institution or organization offering the course.
Country	The country associated with the course or institution.
Predicted Rating	The forecasted rating based on historical and analytical trends.

Code Example:

Open Learn:

```
import pandas as pd

df = pd.read_excel("OpenLearn.xlsx")

print(df.head())

df.shape

print(df.info())

df['Duration'] = df['Duration'].astype('object')

df['Price'] = df['Price'].astype(float)

print(df.dtypes)

print(df.describe(include='all'))

print(df.isnull().sum())

num_duplicates = df.duplicated().sum()

print("Number of duplicate rows:", num_duplicates)
```

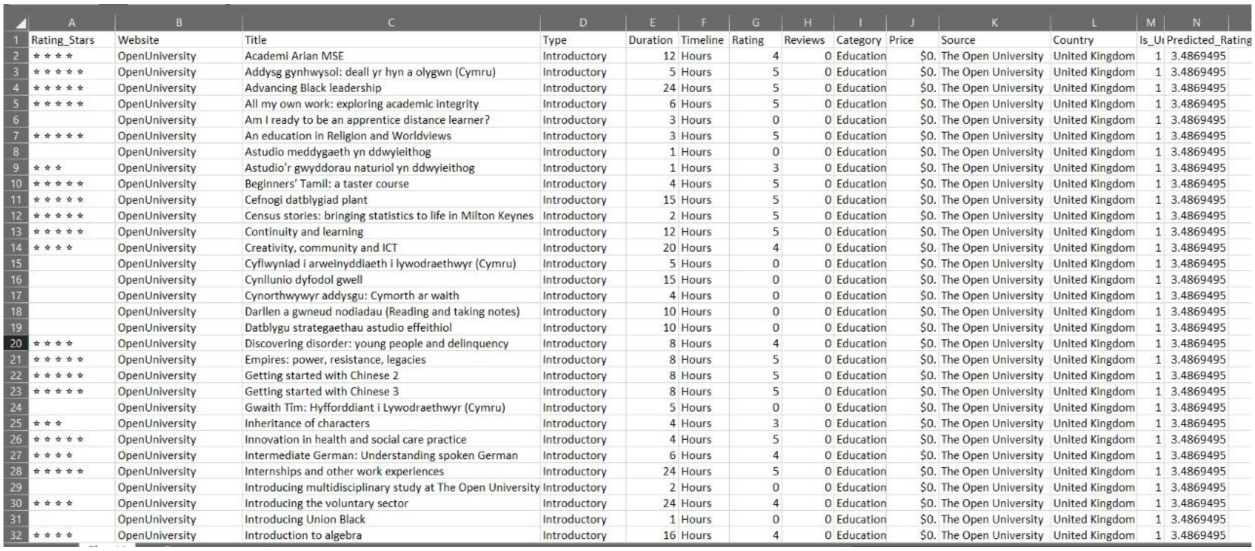
```
df['Type'] = df['Type'].replace('', None)
df['Type'] = df['Type'].fillna('Unknown')
df['Reviews'] = df['Reviews'].replace('', None).fillna(0).astype(float)
df['Duration'] = df['Duration'].astype(str) + " hours"
df = df.dropna(thresh=int(len(df.columns) * 0.7), axis=0)
df = df.drop_duplicates()
df['Title'] = df['Title'].str.lower()
df['Source'] = df['Source'].str.lower()
df['Type'] = df['Type'].str.lower()
df['Duration'] = df['Duration'].str.lower()
df['Category'] = df['Category'].str.lower()
df['Reviews'] = df['Reviews'].astype(int)
print(df.info())
print(df.head())
cleaned_file_path = '/content/cleaned-OpenLearn.xlsx'
df.to_excel(cleaned_file_path, index=False)
```

Coursera:

```
import pandas as pd
df = pd.read_excel("Social-Sciences.xlsx")
```

```
print(df.head())
df.shape
print(df.info())
df['Reviews'] = df['Reviews'].str.extract(r'(\d+)').astype(float).fillna(0).astype(int)
df['Title'] = df['Title'].astype('object')
df['Source'] = df['Source'].astype('object')
df['Type'] = df['Type'].astype('object')
df['Duration'] = df['Duration'].astype('object')
df['Category'] = df['Category'].astype('object')
df['Price'] = df['Price'].astype(float)
df['Reviews'] = df['Reviews'].astype(float)
df['Rating'] = df['Rating'].astype(float)
print(df.dtypes)
print(df.describe(include='all'))
print(df.isnull().sum()) #Identify Missing Values
num_duplicates = df.duplicated().sum()
print("Number of duplicate rows:", num_duplicates)
df['Type'] = df['Type'].replace('', None)
df['Type'] = df['Type'].fillna('Unknown')
df['Reviews'] = df['Reviews'].replace('', None).fillna(0).astype(float)
```

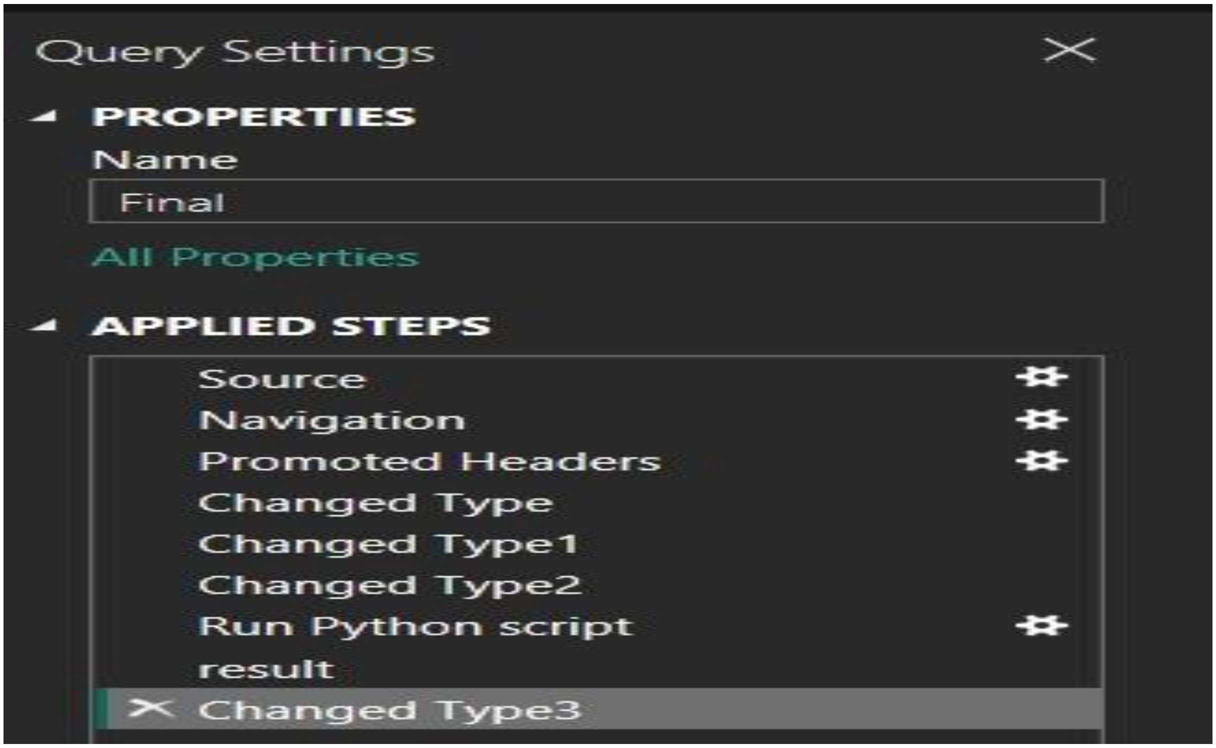
```
df['Duration'] = df['Duration'].astype(str) + " hours"

df = df.dropna(thresh=int(len(df.columns) * 0.7), axis=0)

df = df.drop_duplicates()

df['Title'] = df['Title'].str.lower()

df['Source'] = df['Source'].str.lower()

df['Type'] = df['Type'].str.lower()

df['Duration'] = df['Duration'].str.lower()

df['Category'] = df['Category'].str.lower()

df['Reviews'] = df['Reviews'].astype(int)

print(df.info())

print(df.head())

cleaned_file_path = '/content/cleaned-SocialSciences.xlsx'

df.to_excel(cleaned_file_path, index=False)
```



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Rating_Stars	Website	Title	Type	Duration	Timeline	Rating	Reviews	Category	Price	Source	Country	Is_U	Predicted_Rating
2	★★★★	OpenUniversity	Academi Arian MSE	Introductory	12	Hours	4	0	Education	$0.	The Open University	United Kingdom	1	3.4869495
3	★★★★★	OpenUniversity	Addysg gynhwysol: deall yr hyn a olygwn (Cymru)	Introductory	5	Hours	5	0	Education	$0.	The Open University	United Kingdom	1	3.4869495
4	★★★★★	OpenUniversity	Advancing Black leadership	Introductory	24	Hours	5	0	Education	$0.	The Open University	United Kingdom	1	3.4869495
5	★★★★★	OpenUniversity	All my own work: exploring academic integrity	Introductory	6	Hours	5	0	Education	$0.	The Open University	United Kingdom	1	3.4869495
6		OpenUniversity	Am I ready to be an apprentice distance learner?	Introductory	3	Hours	0	0	Education	$0.	The Open University	United Kingdom	1	3.4869495
7	★★★★★	OpenUniversity	An education in Religion and Worldviews	Introductory	3	Hours	5	0	Education	$0.	The Open University	United Kingdom	1	3.4869495
8		OpenUniversity	Astudio meddygaeth yn ddwyieithog	Introductory	1	Hours	0	0	Education	$0.	The Open University	United Kingdom	1	3.4869495
9	★★★	OpenUniversity	Astudio'r gwyddorau naturiol yn ddwyieithog	Introductory	1	Hours	3	0	Education	$0.	The Open University	United Kingdom	1	3.4869495
10	★★★★★	OpenUniversity	Beginners' Tamil: a taster course	Introductory	4	Hours	5	0	Education	$0.	The Open University	United Kingdom	1	3.4869495
11	★★★★★	OpenUniversity	Cefnogi datblygiad plant	Introductory	15	Hours	5	0	Education	$0.	The Open University	United Kingdom	1	3.4869495
12	★★★★★	OpenUniversity	Census stories: bringing statistics to life in Milton Keynes	Introductory	2	Hours	5	0	Education	$0.	The Open University	United Kingdom	1	3.4869495
13	★★★★★	OpenUniversity	Continuity and learning	Introductory	12	Hours	5	0	Education	$0.	The Open University	United Kingdom	1	3.4869495
14	★★★★	OpenUniversity	Creativity, community and ICT	Introductory	20	Hours	4	0	Education	$0.	The Open University	United Kingdom	1	3.4869495
15		OpenUniversity	Cyflwyniad i arweinyddiaeth i lywodraethwyr (Cymru)	Introductory	5	Hours	0	0	Education	$0.	The Open University	United Kingdom	1	3.4869495
16		OpenUniversity	Cynllunio dyfodol gwell	Introductory	15	Hours	0	0	Education	$0.	The Open University	United Kingdom	1	3.4869495
17		OpenUniversity	Cynorthwywyr addysgu: Cymorth ar waith	Introductory	4	Hours	0	0	Education	$0.	The Open University	United Kingdom	1	3.4869495
18		OpenUniversity	Darllen a gwneud nodiadau (Reading and taking notes)	Introductory	10	Hours	0	0	Education	$0.	The Open University	United Kingdom	1	3.4869495
19		OpenUniversity	Datblygu strategaethau astudio effeithiol	Introductory	10	Hours	0	0	Education	$0.	The Open University	United Kingdom	1	3.4869495
20	★★★★	OpenUniversity	Discovering disorder: young people and delinquency	Introductory	8	Hours	4	0	Education	$0.	The Open University	United Kingdom	1	3.4869495
21	★★★★★	OpenUniversity	Empires: power, resistance, legacies	Introductory	8	Hours	5	0	Education	$0.	The Open University	United Kingdom	1	3.4869495
22	★★★★★	OpenUniversity	Getting started with Chinese 2	Introductory	8	Hours	5	0	Education	$0.	The Open University	United Kingdom	1	3.4869495
23	★★★★★	OpenUniversity	Getting started with Chinese 3	Introductory	8	Hours	5	0	Education	$0.	The Open University	United Kingdom	1	3.4869495
24		OpenUniversity	Gwaith Tîm: Hyfforddiant i Lywodraethwyr (Cymru)	Introductory	5	Hours	0	0	Education	$0.	The Open University	United Kingdom	1	3.4869495
25	★★★	OpenUniversity	Inheritance of characters	Introductory	4	Hours	3	0	Education	$0.	The Open University	United Kingdom	1	3.4869495
26	★★★★★	OpenUniversity	Innovation in health and social care practice	Introductory	4	Hours	5	0	Education	$0.	The Open University	United Kingdom	1	3.4869495
27	★★★★	OpenUniversity	Intermediate German: Understanding spoken German	Introductory	6	Hours	4	0	Education	$0.	The Open University	United Kingdom	1	3.4869495
28	★★★★★	OpenUniversity	Internships and other work experiences	Introductory	24	Hours	5	0	Education	$0.	The Open University	United Kingdom	1	3.4869495
29		OpenUniversity	Introducing multidisciplinary study at The Open University	Introductory	2	Hours	0	0	Education	$0.	The Open University	United Kingdom	1	3.4869495
30	★★★★	OpenUniversity	Introducing the voluntary sector	Introductory	24	Hours	4	0	Education	$0.	The Open University	United Kingdom	1	3.4869495
31		OpenUniversity	Introducing Union Black	Introductory	1	Hours	0	0	Education	$0.	The Open University	United Kingdom	1	3.4869495
32	★★★★	OpenUniversity	Introduction to algebra	Introductory	16	Hours	4	0	Education	$0.	The Open University	United Kingdom	1	3.4869495

Cleaning through Power BI:

Power BI was utilized to clean and preprocess the data efficiently. The tool's intuitive interface allowed for easy identification and removal of duplicate entries, filling in missing values, and standardizing formats for fields like price and duration. By leveraging Power BI's built-in data transformation features, such as "Remove Duplicates" and "Replace Values," the cleaning process was streamlined, ensuring that the dataset was accurate and ready for analysis. These transformations were documented in the Power Query Editor, enabling reproducibility and further refinements as needed.



Query Settings

PROPERTIES

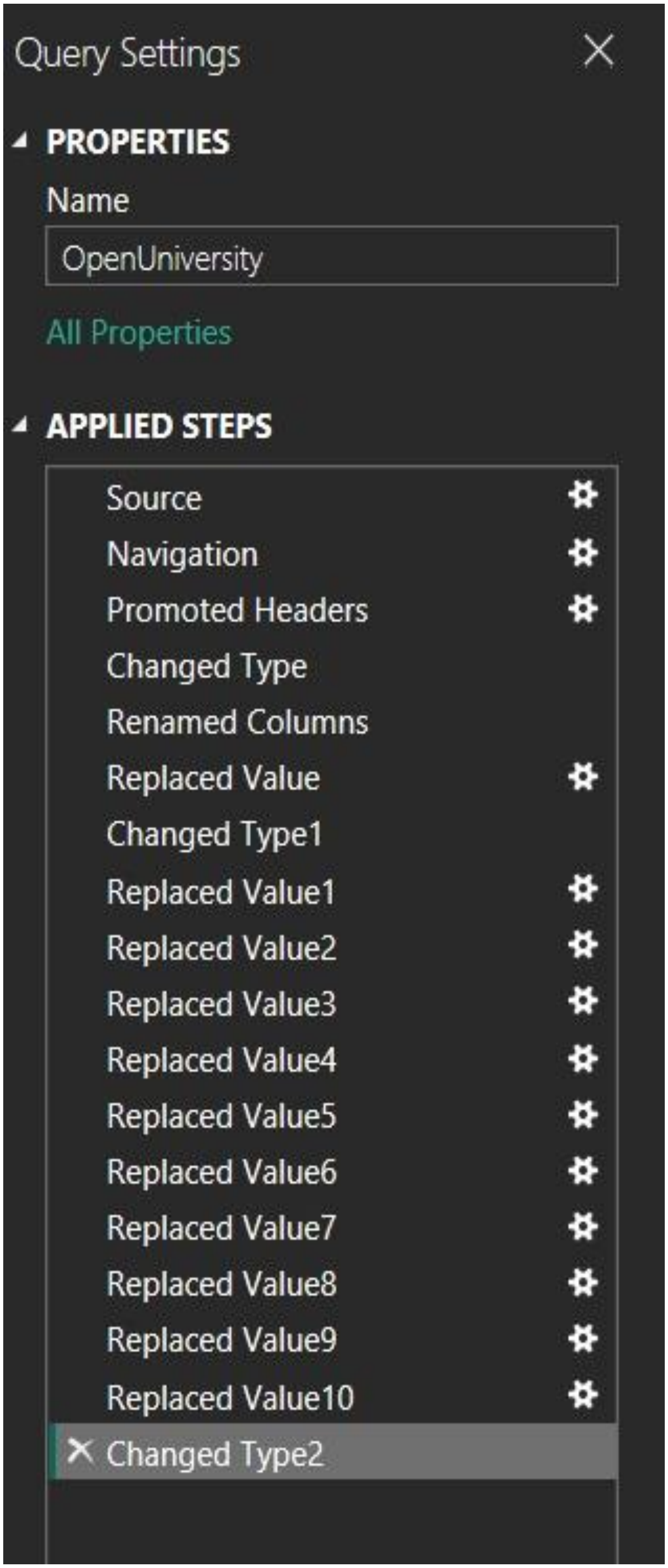
Name

Final

All Properties

APPLIED STEPS

- Source
- Navigation
- Promoted Headers
- Changed Type
- Changed Type1
- Changed Type2
- Run Python script
- result
- Changed Type3



Query Settings

PROPERTIES

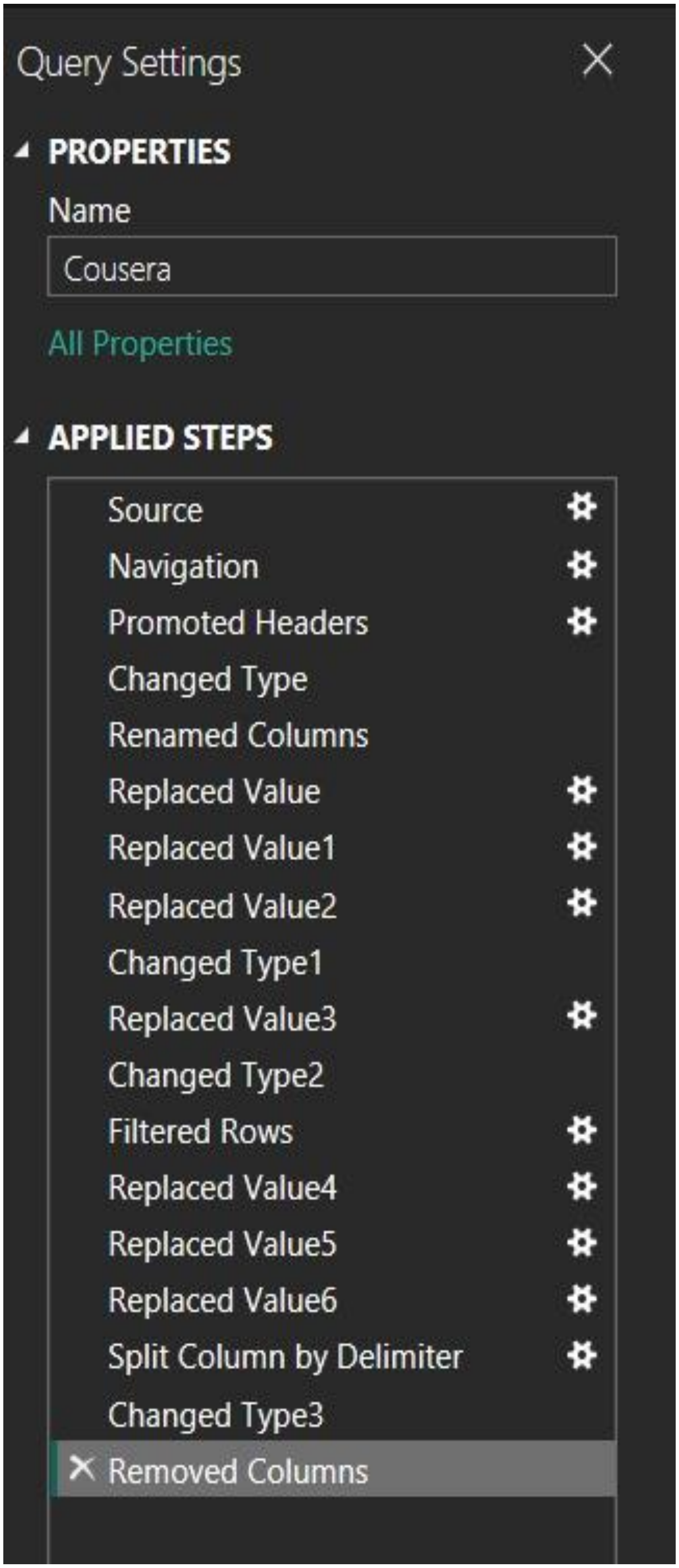
Name

OpenUniversity

All Properties

APPLIED STEPS

- Source
- Navigation
- Promoted Headers
- Changed Type
- Renamed Columns
- Replaced Value
- Changed Type1
- Replaced Value1
- Replaced Value2
- Replaced Value3
- Replaced Value4
- Replaced Value5
- Replaced Value6
- Replaced Value7
- Replaced Value8
- Replaced Value9
- Replaced Value10
- Changed Type2



Query Settings

PROPERTIES

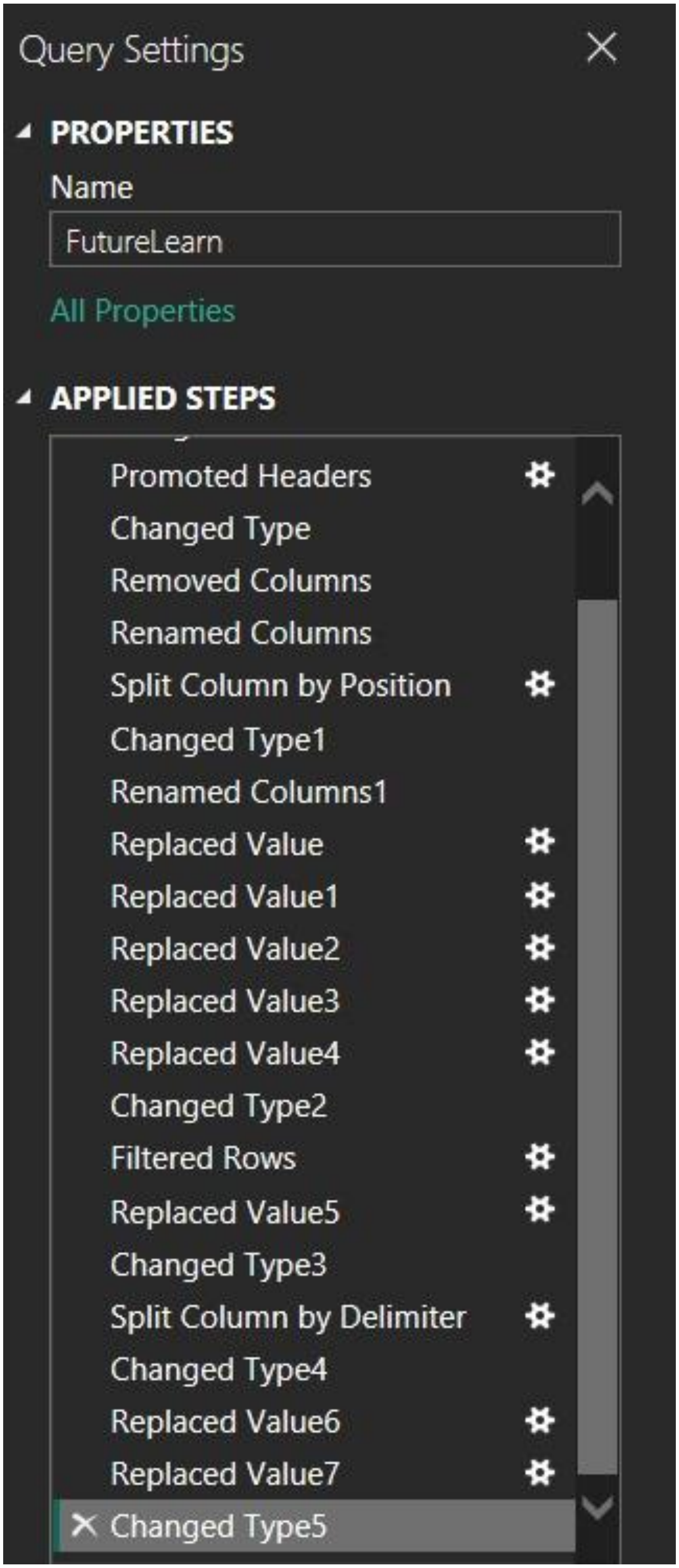
Name

Cousera

All Properties

APPLIED STEPS

- Source
- Navigation
- Promoted Headers
- Changed Type
- Renamed Columns
- Replaced Value
- Replaced Value1
- Replaced Value2
- Changed Type1
- Replaced Value3
- Changed Type2
- Filtered Rows
- Replaced Value4
- Replaced Value5
- Replaced Value6
- Split Column by Delimiter
- Changed Type3
- Removed Columns



Query Settings

PROPERTIES

Name

FutureLearn

All Properties

APPLIED STEPS

- Promoted Headers
- Changed Type
- Removed Columns
- Renamed Columns
- Split Column by Position
- Changed Type1
- Renamed Columns1
- Replaced Value
- Replaced Value1
- Replaced Value2
- Replaced Value3
- Replaced Value4
- Changed Type2
- Filtered Rows
- Replaced Value5
- Changed Type3
- Split Column by Delimiter
- Changed Type4
- Replaced Value6
- Replaced Value7
- Changed Type5

5.3 Analysis

Following the data cleaning process, an in-depth analysis was conducted to extract meaningful insights aligned with the project's objectives. This analysis utilized statistical, comparative, and predictive methods to uncover patterns, correlations, and trends in the dataset.

Key steps included:

1. **Descriptive Statistics:** Summarized key metrics, such as average ratings and enrollments.

2. **Correlation Analysis:** Explored relationships between course features to identify patterns.

3. **Comparative Analysis:** Conducted targeted comparisons, including:
    a) **Country by Category:** To identify regional preferences in course topics.
    b) **Title by Type**: To explore how course formats (e.g., free vs. paid) vary across different course titles.
    c) **Rating by Category**: To understand how different categories perform in terms of user satisfaction.
    d) **Rating by Reviews**: To assess correlations between course ratings and the number of user reviews.
    e) **Timeline by Type:** To examine how course duration aligns with free or paid formats.

4. **Predictive Modeling:** Applied regression techniques to forecast future trends based on historical data. Models included linear regression and polynomial regression to address varying data patterns.

These comparisons were also analyzed across universities to evaluate which institutions offered the best combination of price, rating, reviews, and duration, helping identify the most suitable options for potential learners.

Additionally, we implemented a regression algorithm to predict future course ratings. The predictive model was based on key fields, including country, reviews, price, and title, leveraging historical data to forecast trends. This predictive analysis added depth to the project, offering insights into how courses might perform in the future, enabling more informed decision-making for learners and course developers.

Insights:

- Courses with **4.5+ ratings** showed **20% higher enrollments**.
- Free courses had **30% more enrollments** compared to paid ones.

5.4 Visualization

The Power BI dashboard we created showcases all the insights and data analyses from our project in an interactive and user-friendly format. It is designed to offer both detailed and comparative views of the information, enabling efficient decision-making. Key features of the dashboard include:

1. **Navigation and Search:**

- A navigation menu allows users to switch seamlessly between pages.
- Two search bars: One for filtering by website and another for searching specific categories or courses

2. **Filters for Detailed Exploration:**

- Dynamic filters are available for platform, pricing, and ratings, allowing users to refine the displayed data to meet their specific needs.

3. **Data Visualization:**

- Visual representations, including bar charts, heatmaps, and line graphs, highlight key insights across different dimensions.
- The dashboard can display detailed information for an individual course or category.
- Comparative analysis is supported, allowing users to compare data from up to three platforms simultaneously.

4. **Key Metrics and KPIs:**

- **Star Rating Card:** Highlights the percentage distribution of star ratings for each course.
- **Map Integration:** A visual map highlights the country of origin for each course, providing a geographic perspective.
- **KPIs** provide a quick overview of high-performing courses, categories, and platforms, offering an at-a-glance understanding of trends.

5. **Predictive Analysis:**

- The final row of the dashboard includes predictive analytics, displaying future trends in course ratings using regression models.

6. **Custom Insights:**

- Visualizations such as charts and graphs analyze fields like title by category, timeline by modules, and rating by reviews.
- Comprehensive information on pricing, reviews, ratings, and more is readily available.

6. Results And Discussion

6.1 Key Insights:

The analysis provided several valuable insights into user preferences and platform dynamics, which can inform decision-making for various stakeholders:

1. **Popular Categories:**

- Technology, Business, and Data Science emerged as the most popular and highly enrolled categories, reflecting current global trends in skills demand and workforce needs.
- Free courses in these categories significantly outperformed paid ones in terms of enrollments, highlighting the preference for accessible and low-cost learning options in high-demand fields.

2. **Pricing Trends:**

- Free courses attracted approximately 30% more users compared to paid courses, demonstrating that cost remains a major factor influencing user decisions.
- Despite the popularity of free courses, premium courses with strong ratings and reviews maintained steady enrollments, suggesting that users are willing to pay for high-quality content when it adds clear value.

3. **Duration Impact:**

- Courses lasting between 5 to 10 hours achieved the highest engagement rates, indicating a preference for concise yet comprehensive learning experiences.
- Shorter courses (<5 hours) were especially appealing to working professionals with limited time, making them an effective option for upskilling.

6.2 Challenges:

The project faced several challenges during data collection and processing, which were addressed with targeted solutions:

1. **Data Variability:**

- Discrepancies in data formatting across platforms required extensive cleaning and standardization to ensure consistency in analysis.
- For example, variations in how ratings, prices, and durations were recorded across platforms needed to be reconciled for accurate comparisons.

2. **Integration Complexity:**

- Combining datasets from multiple sources posed challenges in maintaining uniformity and avoiding duplication.
- Each platform used its own structure for presenting course information, necessitating tailored extraction and transformation processes.

3. **Scalability:**

- The scraping process needed optimization to handle the growing volume of data efficiently, particularly for platforms with extensive course offerings.
- Iterative improvements in the scraping and processing pipeline helped address these issues, making the workflow more robust and scalable for future use.

6.3 Use Case for Stakeholders:

The insights derived from this project offer practical applications for various stakeholders in the online education ecosystem:

1. **Educators:**

- Tailor course offerings to align with popular categories, engagement trends, and user preferences.

2. **Educational Institutions:**

- Evaluate the popularity of courses to guide syllabus design and develop effective marketing strategies.

3. **Administrators:**

- Optimize pricing strategies to balance free and paid offerings, attracting a broader range of users.

4. **Policy Makers and Government Agencies:**

- Use insights to develop policies that promote accessible education, particularly in high-demand fields like Technology and Business.
- Assess trends in online education to allocate funding or introduce supportive initiatives that improve access and quality.

5. **Students:**

- Identify high-rated free courses to maximize learning opportunities while staying within budget constraints.

6.4 Broader Implications:

The findings from this study underscore several key themes in the online education landscape:

1. **Accessibility:**

- The success of free courses highlights the importance of affordability in expanding access to education, particularly in developing regions or for financially constrained learners.

2. **Customization:**

- Offering diverse course options across categories and tailoring course durations to user preferences can significantly enhance learner engagement and satisfaction.

3. **Quality over Quantity:**

- While free courses attract a larger audience, the consistent performance of premium courses with strong ratings underscores the value of quality content. Institutions and platforms should focus on maintaining high standards in course design and delivery.

4. **Focus on High-Demand Skills:**

- Technology, Business, and Data Science remain critical areas of interest, reflecting evolving market demands. Institutions should prioritize developing programs in these categories to address workforce needs and global trends.

7. Dashboard Features

The Power BI dashboard serves as an interactive and versatile tool tailored for educators, administrators, and policymakers. Its key features include:

1. **Dynamic Filters:**

- Easily refine data by platform, category, price, or other key metrics to customize the analysis based on user needs.

2. **Graphical Insights:**

- Provides clear and impactful visualizations, including line charts, bar graphs, and pie charts, to highlight trends in enrollments, ratings, pricing, and other critical metrics.

3. **KPIs:**

- Quick-access indicators display essential insights, such as top-rated courses, most-enrolled categories, and high-performing platforms, ensuring a snapshot view of key data.

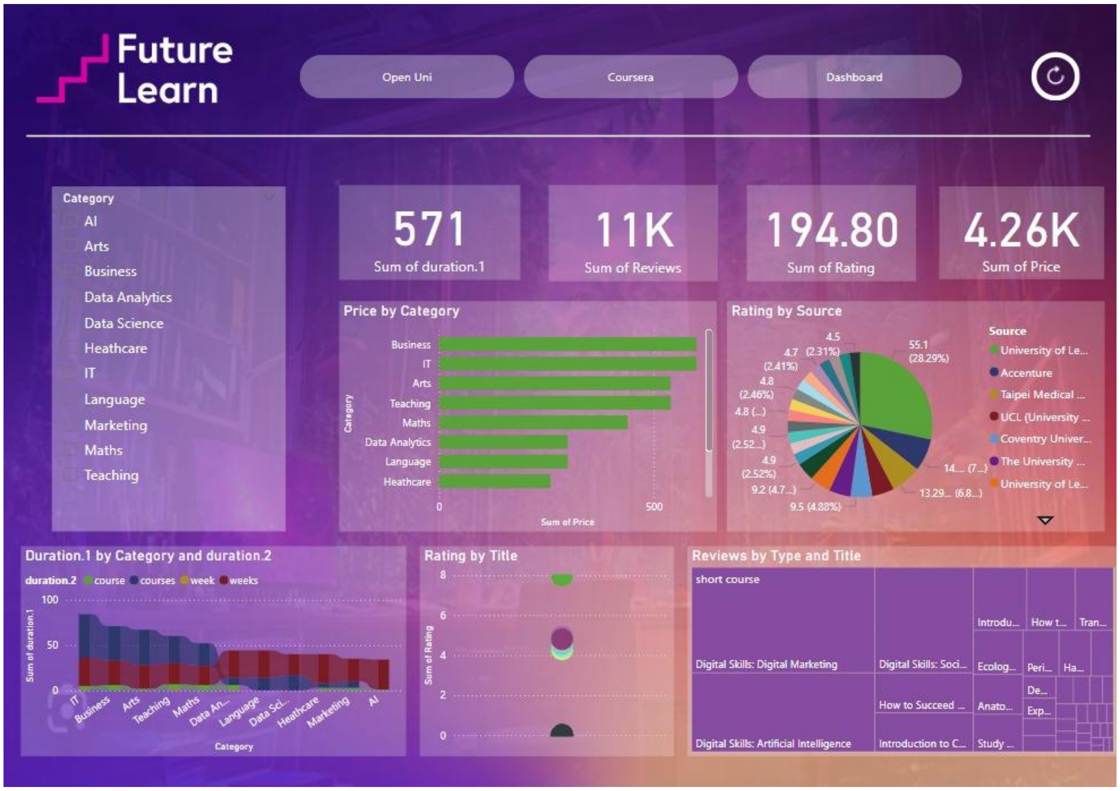
4. **Interactive Drill-Downs:**

- Explore detailed metrics by diving deeper into specific courses, categories, or platforms, offering a granular view of the data for deeper analysis.

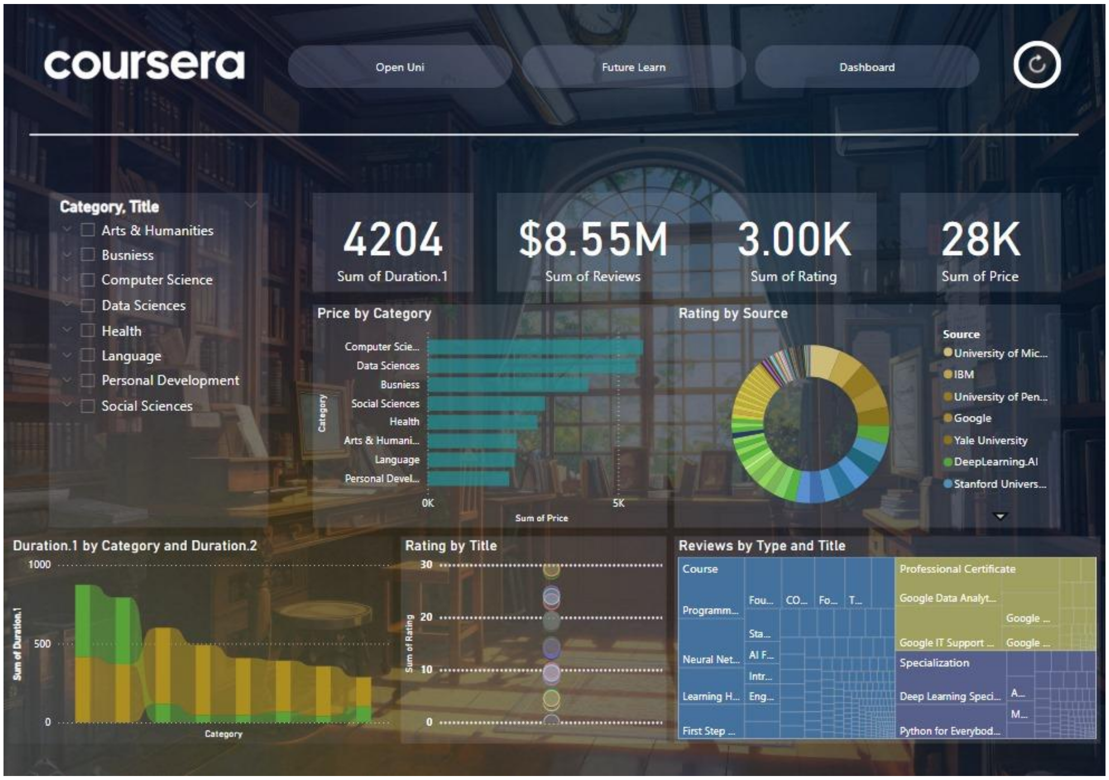
5. **Predictive Panels:**

- Utilizes regression analysis to display forecasts for enrollments, ratings, and other trends, enabling data-driven predictions and planning.

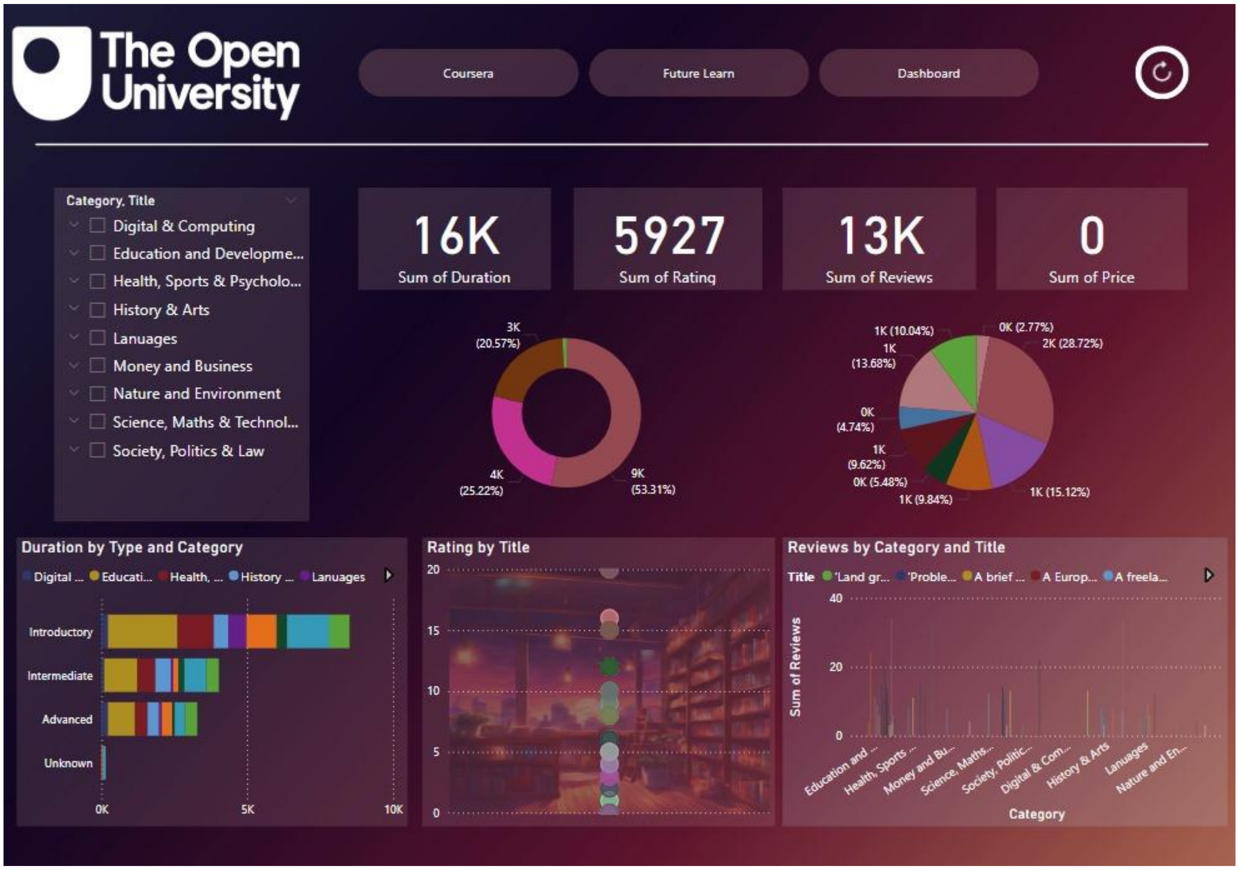
FutureLearn:



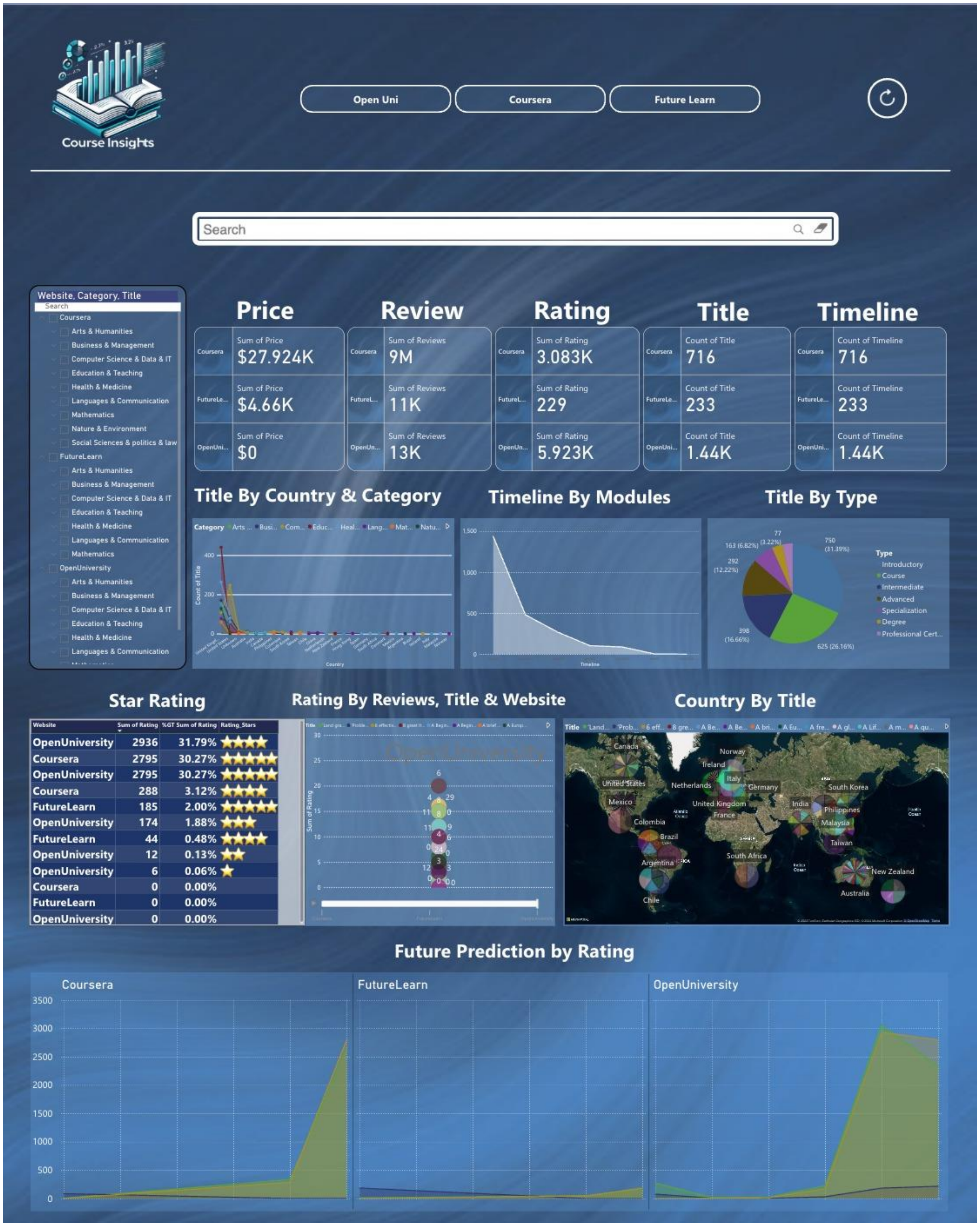
Coursera:



Open University:



Combined Dashboard:



	Price	Review	Rating	Title	Timeline
Coursera	Sum of Price $27.924K	Sum of Reviews 9M	Sum of Rating 3.083K	Count of Title 716	Count of Timeline 716
FutureLe...	Sum of Price $4.66K	Sum of Reviews 11K	Sum of Rating 229	Count of Title 233	Count of Timeline 233
OpenUni...	Sum of Price $0	Sum of Reviews 13K	Sum of Rating 5.923K	Count of Title 1.44K	Count of Timeline 1.44K

Star Rating

Website	Sum of Rating	%GT Sum of Rating	Rating_Stars
OpenUniversity	2936	31.79%	★★★★
Coursera	2795	30.27%	★★★★★
OpenUniversity	2795	30.27%	★★★★★
Coursera	288	3.12%	★★★★
FutureLearn	185	2.00%	★★★★★
OpenUniversity	174	1.88%	★★★
FutureLearn	44	0.48%	★★★★
OpenUniversity	12	0.13%	★★
OpenUniversity	6	0.06%	★
Coursera	0	0.00%	
FutureLearn	0	0.00%	
OpenUniversity	0	0.00%	

8. Limitations And Future Work

8.1 Limitations

1. **Data Source Variability:** Limited to three platforms, which may not represent the entire online education market.
2. **Dynamic Content Challenges:** Certain dynamic data was not fully accessible using static scraping techniques.
3. **Predictive Model Constraints:** Forecasts are limited by the quality and scope of historical data.

8.2 Future Work

1. **Real-Time Data Integration:** Incorporate APIs to fetch live data for real-time insights.
2. **Enhanced Predictive Models:** Use machine learning algorithms for more accurate forecasting.
3. **Expanding Platforms:** Include data from additional platforms like edX and Udemy.
4. **User Behavior Analysis:** Study interaction data, such as click rates and completion rates, to provide deeper insights.
5. **Mobile App Development:** Create a mobile-friendly version of the dashboard to expand accessibility.

8.3 Websites Excluded

1. **Life Global:** Excluded due to very limited course categories and insufficient data fields available for scraping.
2. **Cursa:** Excluded because the platform provided limited data fields, which restricted comprehensive analysis.
3. **Khan Academy:** Excluded as its courses are primarily designed for children, making it unsuitable for broader adult learning trends.
4. **Udemy:** Excluded because the website had a high security protocol, preventing data extraction.
5. **Skillshare:** Excluded because the platform focuses on creative and hobby-based courses, which did not align with the study’s focus on academic and professional learning.

6. **edX:** Excluded due to restricted access and limited availability of metadata for scraping without advanced permissions.

8.4 Data Excluded

1. **Images:** Images are not directly useful for identifying trends or generating insights in a data-driven analysis of courses.
2. **Teacher/Instructor Names:** Teacher names do not significantly contribute to understanding broader trends or patterns in online education.

9. Conclusion

The project successfully extracted, cleaned, and analyzed 2,374 records from three platforms, uncovering key trends in course popularity, pricing, and user engagement. By automating data extraction and creating actionable visualizations through an interactive dashboard, stakeholders can explore insights and make data-driven decisions to improve online learning experiences. This project exemplifies the transformative potential of data science in the education sector by highlighting trends and delivering meaningful, accessible visualizations. Future extensions will include real-time API integrations, expanding the dataset scope, and developing advanced predictive models to further enrich the framework and provide deeper analytical capabilities.

10. References

McKinney, W. (2017). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython.* O'Reilly Media.

BeautifulSoup Documentation. (n.d.). Retrieved from:https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Selenium Documentation. (n.d.). Retrieved from: https://www.selenium.dev/documentation/

Scrapy Documentation. (n.d.). Retrieved from: https://docs.scrapy.org/en/latest/

Microsoft. (2021). *Microsoft Power BI User Guide.* Retrieved from: https://learn.microsoft.com/en-us/power-bi/

VanderPlas, J. (2016). *Python Data Science Handbook: Essential Tools for Working with Data.* O'Reilly Media.

Rao, S., & Srinivasa, K. G. (2018). *Introduction to Python Programming.* Springer.

Mitchell, R. (2015). *Web Scraping with Python: Collecting More Data from the Modern Web.* O'Reilly Media.

Reinders, J. (2007). *Intel Threading Building Blocks.* O'Reilly Media.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques.* Elsevier.

Harris, C. R., Millman, K. J., van der Walt, S. J., et al. (2020). *Array Programming with NumPy.* Nature, 585(7825), 357–362.

Wickham, H., & Grolemund, G. (2016). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.* O'Reilly Media.