Discipline: Data Collection & Preparation – SIS 1

Team name – EduAnalyzer

Team members - Amreyeva Alina, Beketay Symbat

The goal of this project - to analyze the online courses data obtained from two different sources: the Coursera API and web scraping from a test site. By gathering data from these sources, we aim to investigate various aspects such as course pricing, ratings, and correlations between them. This analysis will provide valuable insights into the characteristics of online courses in terms of their price and ratings.

Data Collection

1. **Coursera API**: We used the Coursera API to gather information about the courses available on the platform. The API provided details such as course names, course types, and course IDs. We limited the number of courses to 50 for analysis.
2. **Web Scraping**: We also collected data using web scraping from the webscraper.io test site, specifically from a page listing laptops. Using BeautifulSoup, we scraped information about product titles, prices, and ratings. This provided us with another set of course-related data.

Both sources of data were then combined to create a single dataset for analysis. The data was merged based on the course names, which allowed us to combine information from both APIs and web scraping sources.

Data Cleaning

1. We handled missing values by filling in any missing entries in the `price` and `rating` columns with their respective mean values. This ensured that we had no missing data for analysis.
2. We removed any duplicate rows to ensure the integrity of the data.
3. We added a new column for the course category, assigning a default category of 'Uncategorized' for any missing values in that column.
4. The price values were cleaned by removing any currency symbols and commas to ensure that they were in a proper numerical format for analysis.

Data Analysis

We performed various analyses on the cleaned data to derive insights into the online courses.

1. **Average Price**: We calculated the average price of the courses in the dataset.
2. **Average Rating**: We also calculated the average rating for the courses.
3. **Correlation**: We analyzed the correlation between price and rating. This gave us an understanding of whether higher-priced courses tend to have higher ratings.

The results of these analyses were as follows:

- The average price of the courses was approximately $756.16.
- The average rating of the courses was 4.67.
- The correlation between price and rating was 0.62, indicating a moderate positive relationship.

Data Visualization

To better understand the data, we created several visualizations using Matplotlib and Seaborn.

1. **Boxplots**: We created boxplots to analyze the distribution of prices and ratings. These visualizations helped us identify potential outliers in the data.
2. **KDE Plots**: We also created Kernel Density Estimate (KDE) plots to visualize the density of prices and ratings. These plots provided insight into the overall distribution of the data and highlighted areas with the highest concentrations of courses.
3. **Correlation Heatmap**: Finally, we created a heatmap to visualize the correlation between numerical columns in the dataset. The heatmap confirmed the positive correlation between price and rating, showing that as the price of a course increases, so does its rating.

Conclusion

In conclusion, this analysis provided valuable insights into the relationship between the prices and ratings of online courses. The key findings include:

- The average price of the courses in the dataset is $756.16, while the average rating is 4.67.
- There is a moderate positive correlation (0.62) between price and rating, suggesting that higher-priced courses tend to have higher ratings.
- The visualizations helped identify outliers and provided a deeper understanding of the data distribution.

By using both API data and web scraping, we were able to create a comprehensive dataset for analysis. This analysis can help inform decisions about the pricing and quality of online courses, providing valuable insights for educators, students, and platforms offering online courses.

Recommendations

Based on the findings, here are some recommendations:

- Courses with higher prices tend to have better ratings, but this correlation is moderate. Further analysis could explore whether the quality of a course is directly linked to its pricing strategy.
- The visualizations suggest that there are outliers in the dataset. Investigating these outliers could provide insights into the reasons for extreme pricing or ratings.

Overall, this project demonstrated how to combine data from different sources, clean and preprocess the data, perform analysis, and visualize the results. This approach can be applied to other datasets to gain similar insights into various domains.