Exploratory Data Analysis on Online Graphic Design Courses Dataset

By

Rekan Awat

Bwar Yasin

Zhiwar Mohammed

Supervised by

Dr. Miran Taha Abdullah

January, 2024 (English)

Table of Contents

1 Abstract

2 Introduction

3 Data Cleaning and Preprocessing

4 Descriptive Statistics

5 Key Findings

- **5.1** Pricing Analysis
- **5.2** Subscriber and Review Analysis
- **5.3** Lecture and Content Duration Analysis
- **5.4** Level of Courses Analysis
- **5.5** Published Time Analysis
- **5.6** Rating Analysis
- **5.7** Correlation Analysis

6 Exploratory Data Analysis (EDA)

- **6.1** Price Distribution
- **6.2** Subscriber Analysis
- **6.3** Reviews and Ratings
- **6.4** Course Content Analysis
- **6.5** Temporal Analysis
- **6.6** Correlation Analysis

7 Conclusion

1.Abstract

This comprehensive report delves into the intricate details of a dataset containing information on graphic design courses. The dataset, analyzed using Python and various data analysis libraries, explores key metrics such as pricing, subscribers, reviews, ratings, course content, temporal trends, and correlations among different features. The study provides valuable insights for educators, learners, and platform administrators, facilitating informed decision-making in the dynamic field of graphic design education.

2.Introduction

The report initiates with an overview of the dataset's structure and the tools employed for analysis. Python, coupled with Pandas, NumPy, Matplotlib, Seaborn, and SciPy, forms the backbone of the study. The dataset is loaded, and preliminary steps include data preprocessing to ensure quality and integrity. Subsequent sections delve into exploratory data analysis (EDA), offering an in-depth understanding of various aspects of graphic design courses.

3.Data Cleaning and Preprocessing:

The analysis begins with data cleaning and preprocessing steps to ensure the dataset's integrity and reliability. These steps include handling null values, removing duplicates, and converting data types. Notably, the 'price' column is transformed to numeric format, replacing 'Free' values with 0. The dataset is then explored to understand its size and structure.

4.Descriptive Statistics:

The report provides descriptive statistics for key columns, including 'price,' 'numSubscribers,' 'numReviews,' 'numPublishedLectures,' 'contentInfo,' 'instructionalLevel,' 'publishedTime,' and 'rating.' Histograms, pie charts, and bar plots are used to visualize the distributions of these attributes.

5.Key Findings:

5.1 Pricing Analysis:

The majority of courses are paid, with a clear distinction in the pricing distribution.

The analysis of price ranges and the distribution of prices provides insights into the affordability and diversity of courses.

5.2 Subscriber and Review Analysis:

The number of subscribers varies widely, with a focus on understanding the distribution within specific ranges.

Courses with fewer subscribers dominate, but a significant number falls within the 0-5000 range.

5.3 Lecture and Content Duration Analysis:

The distribution of the number of lectures provides an overview of the structure of these courses.

Content duration is explored, revealing the majority of courses have a duration of less than 10 hours.

5.4 Level of Courses Analysis:

The level of difficulty is categorized, showcasing the distribution of courses among different difficulty levels.

5.5 Published Time Analysis:

Courses' publication dates are analyzed, illustrating the growth in the number of courses over the years.

5.6 Rating Analysis:

The distribution of ratings provides insights into the overall satisfaction of students with these courses.

The correlation between ratings and other numeric columns, such as the number of subscribers and reviews, is explored.

5.7 Correlation Analysis:

A correlation matrix is presented, highlighting the relationships between different numeric variables.

Scatter plots and bar charts are utilized to visualize correlations, such as between price and level, and year-wise content duration.

6.Exploratory Data Analysis (EDA)

6.1 Price Distribution

The distribution of course prices is vividly presented through histograms, bar charts, and pie charts. The diverse pricing landscape, ranging from free courses to premium offerings, is captured. The visualizations offer a nuanced view of the distribution, enabling stakeholders to grasp the pricing dynamics within the graphic design courses dataset.

6.2 Subscriber Analysis

In-depth scrutiny of the 'numSubscribers' column unfolds the distribution of course popularity. Visualizations, including histograms and pie charts, provide a detailed perspective. Outliers are identified and filtered to enhance the analysis, leading to valuable insights into the factors influencing a course's popularity.

6.3 Reviews and Ratings

The distribution of reviews and ratings is explored to assess the quality and satisfaction levels of graphic design courses. Correlations between the number of reviews, ratings, and subscribers are visually represented, providing a comprehensive understanding of user engagement.

6.4 Course Content Analysis

The 'contentInfo' column, representing the duration of course content, undergoes preprocessing. The distribution of content hours is visualized, allowing for the identification of trends in course duration. This analysis aids educators and learners in making informed decisions about course engagement.

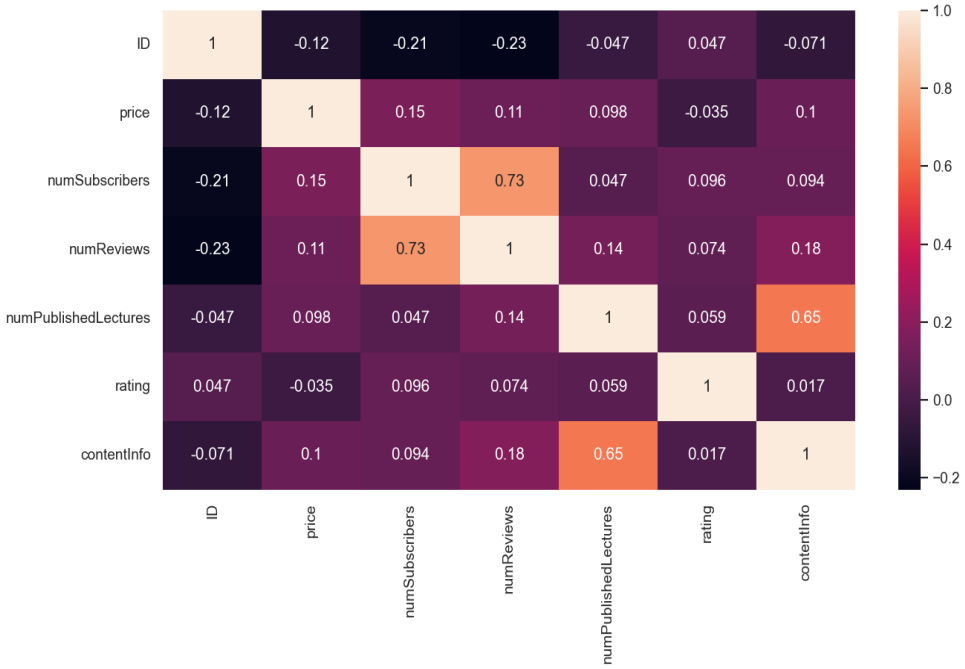
6.5 Temporal Analysis

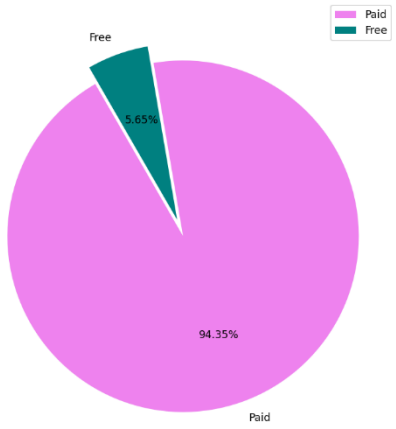
Temporal trends in course publication are uncovered through the 'publishedTime' column. The yearly distribution of course publications is visualized, offering insights into the growth and popularity of graphic design courses over time.

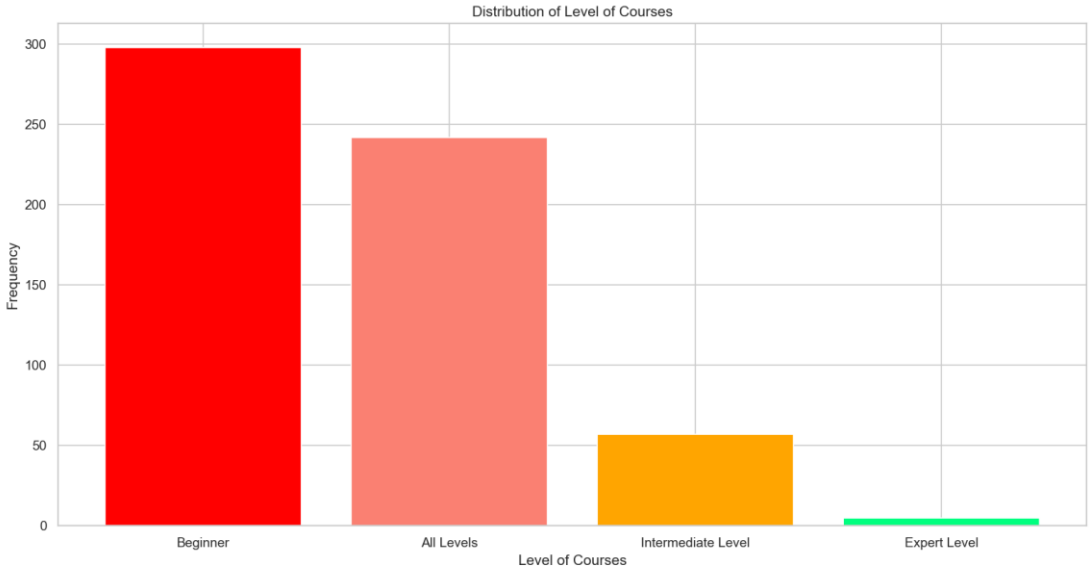
6.6 Correlation Analysis

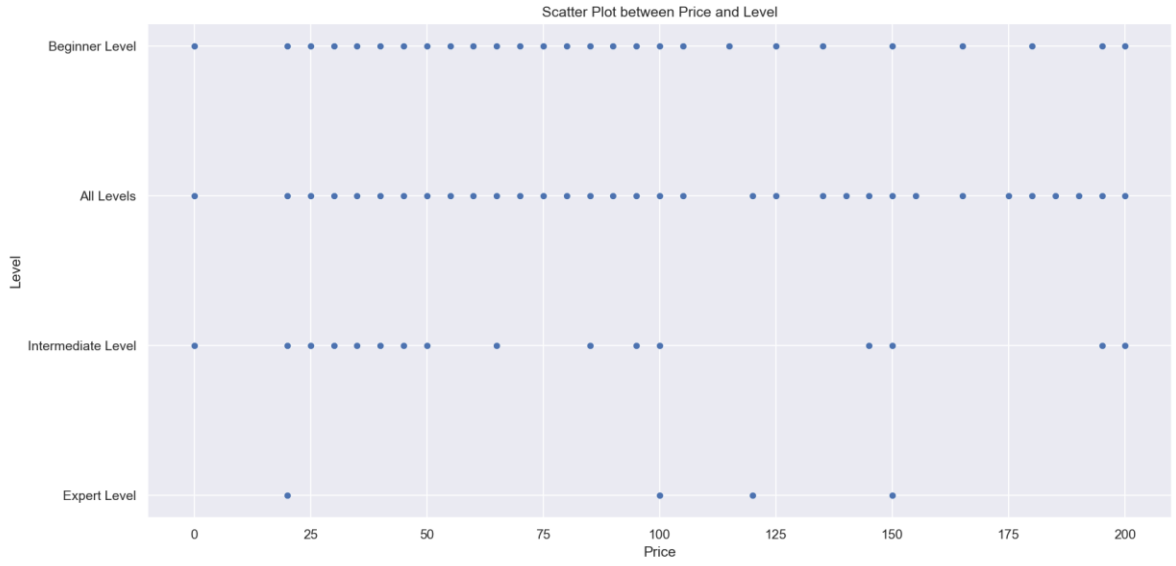
Correlation matrices and heatmaps are generated to quantify relationships between numeric columns. Scatter plots and bar plots highlight correlations between variables such as ratings, subscribers, and prices. Notable correlations are identified, contributing to a comprehensive understanding of the dataset.

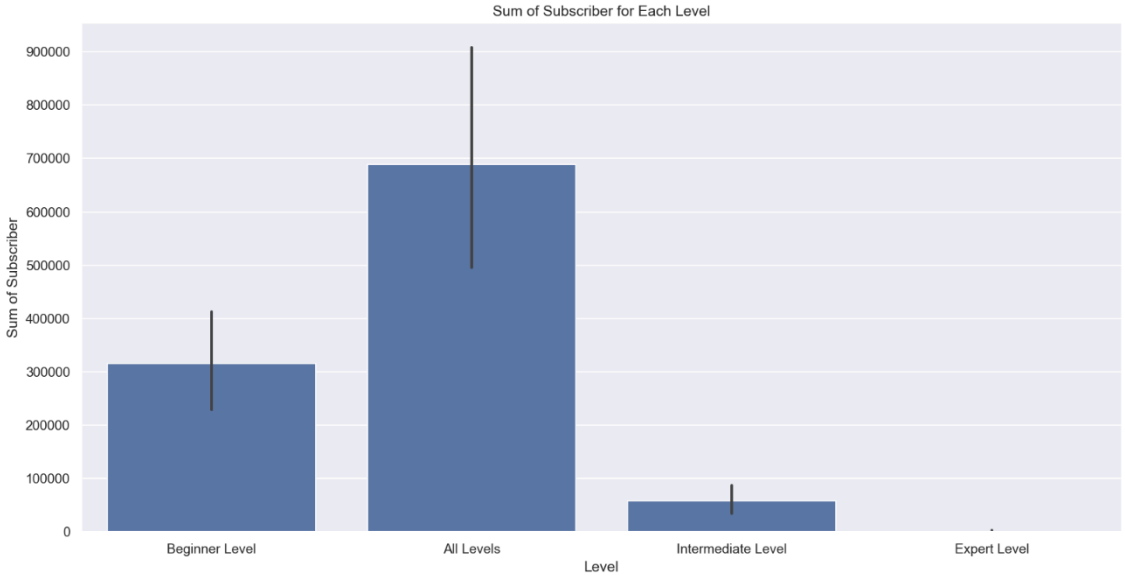
6.7 Some graphs and models






Distribution of Level of Courses
Frequency
300
250
200
150
100
50
0
Beginner
All Levels
Intermediate Level
Expert Level
Level of Courses


Scatter Plot between Price and Level
Beginner Level
All Levels
Intermediate Level
Expert Level
Level
0
25
50
75
100
125
150
175
200
Price



7.Conclusion

In conclusion, this scientific report provides a meticulous analysis of a graphic design courses dataset, employing a scientific approach and utilizing powerful data analysis libraries in Python. The insights gained from this analysis can guide decision-making processes for educators, learners, and platform administrators in the field of graphic design education.