POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Systems and Methods for Big and Unstructured Data Project

Author(s): **Marco Minaudo**

Academic Year: 2024-2025

Contents

1 | Introduction

This project analyzes a dataset containing various udemy courses (roughly 98000) to understand the current online course market and better understand its characteristics. To analyze the dataset, for this project, Elasticsearch was chosen, a search engine optimized for full-text search and complex queries, making it ideal for efficiently managing and analyzing such a large dataset. In addition, using the entire ELK stack, it is possible to easily manage large amounts of data and create effective dashboards to visualize query results. The entire project was run on docker to simplify management and sharing, while Postman was used to execute requests to Elasticsearch.

2 | Data Wrangling

Logstash has been used to prepare the dataset contained in a CSV file. The process consists of multiple steps, including loading, transforming, and cleaning the data to ensure that the dataset is structured and ready to be analyzed.

2.1. Data loading

In this process the dataset in csv is chosen and uploaded to Logstash The following code snippet is part of the Logstash configuration file responsible for importing the file

```
input {
    file {
        path => "/usr/share/logstash/data/udemy_courses.csv"
        start_position => "beginning"
    }
}
```

2.2. Data Transformation and Cleaning

In Logstash, all transformations are defined within the `filter {..}` block. Within this block, the `csv` and `mutate` filters are used to process and transform the data.

The `csv` filter is applied to parse the data from a CSV file. The `columns` parameter defines the names of the columns to be used, such as "id", "title", "url", and others, which correspond to the fields in the CSV.

```
csv {
    skip_header => true
    columns => ["id", "title", "url", "is_paid",
       "instructor_names", "category", "headline",
       "num_subscribers", "rating", "num_reviews",
       "instructional_level", "objectives", "curriculum"]
    separator => ","
}
```

The `mutate` filter is applied to modify and clean the data. such removes unnecessary fields and logstash metadata. Furthemore, to ensure the correctness of data type conversions are performed on specific fields

```
    mutate {
        remove_field => ["id", "url", "@timestamp", "@version", "event", "host", "log"]
        lowercase => ["is_paid"] // True and False value in elastic are in lowercase
        convert => { "id" => "integer" }
        convert => { "num_subscribers" => "integer" }
        convert => { "rating" => "float" }
        convert => { "num_reviews" => "integer" }
    }
```

The result is a dataset that is consistent, relevant, and ready for processing.

3 | Dataset

The dataset provides detailed information on courses offered by Udemy. It contains 98 107 courses, and for each course are available important information such as titles, instructors, categories, ratings, number of students enrolled, and reviews.

3.1. Attributes and Description

The following table shows all the fields in the dataset and their description

Field	**Type**	**Description**
id	Long	A unique identifier for each course. It is used to ensure that each course can be uniquely identified in the system.
title	String	The name or title of the course.
url	String	The web link to access the course on Udemy. This field provides a direct link to the course page on the platform.
is_paid	Boolean	Indicates whether the course is paid (true) or free (false).
instructor_names	String	The names of the instructors teaching the course. This field can contain one or more names separated by commas, indicating all the instructors involved in the course.
category	String	The main category or topic of the course, such as "Development", "Business", "Design", etc. This attribute helps classify the course based on its content and refine searches by topic.
headline	String	Briefly describes its main content. This field serves to provide an engaging summary of the course to attract potential learners.
num_subscribers	Integer	The number of users enrolled in the course. It indicates the course's popularity.
rating	Float	The average rating of the course, based on user reviews. The score ranges from 1 to 5.
num_reviews	Integer	The total number of reviews submitted by users for the course. This field helps determine the reliability of the course's average rating.
instructional_level	String	The difficulty level of the course, which can be "Beginner", "Intermediate", "Advanced", or "All Levels". This helps describe the target audience of the course.
objectives	String	The learning objectives or outcomes the course aims to achieve. This field describes what students will learn upon completing the course.
curriculum	String	The course content or structure, outlining the modules and specific lessons covered.

3.2. Schema

In the schema defined for Elasticsearch, the `dynamic` parameter was set to `false` to ensure that new fields are not automatically added to the index when documents are ingested. Additionally, the `id` and `url` fields were excluded from the schema because they are not relevant to the analyses and queries performed during the search process.

```
{
  "dynamic": "false",
  "properties": {
    "title": {
      "type": "text"
      },
    "is_paid": {
      "type": "boolean"
    },
    "instructor_names": {
      "type": "text"
    },
    "category": {
      "type": "keyword"
    },
    "headline": {
      "type": "text"
    },
    "num_subscribers": {
      "type": "integer"
    },
    "rating": {
      "type": "float"
    },
    "num_reviews": {
      "type": "integer"
    },
    "instructional_level": {
      "type": "keyword"
    },
    "objectives": {
      "type": "text"
    },
    "curriculum": {
      "type": "text"
    }
  }
}
```

4 | Queries

4.1. Top free 5 python courses

4.1.1. Description

This Elasticsearch query is designed to get the top five free Python courses that have a significant number of reviews (greater than or equal to 100), sorted by their rating and popularity. Furthermore, the query uses `filter` instead of `must` to improve efficiency. Since the goal is to retrieve the top 5 Python courses sorted by rating and number of reviews, it is not necessary to compute a score for each document.

4.1.2. Query

```
{
  "query": {
    "bool": {
      "filter": [
        {"term": {"is_paid": false}},
        {"match": {"title": "python"}},
        {"range": {"num_reviews": {"gte": 100}}}
      ]
    }
  },
  "sort":[
      { "rating":"desc"},
      { "num_reviews":"desc"}
  ],
  "size":5
}
```

4.1.3. Output

To improve readability, only the first result will be displayed

```
{
    "_index": "udemy_courses",
    "_id": "5n2GmJMBkz9uf3b-iTkr",
    "_score": null,
    "_source": {
        "curriculum": "Introduction, Introduction, What We Will Cover?,
           What is Programming?, Introduction to Python, Tools to Write
           Code (Text Editors), Introduction to Notebook, Data Types
           and Variables, Data Structures, Variables, Keywords, Data
           Types - String, Data Types - Numbers, Python Operators-
           Introduction, Data Types - Boolen, Basic Syntax, Arithmetic
           Operators, Assignment Operators, Logical Operators,
           Membership Operators, String Concatenation, Functions and
           Methods, Basic Syntax, Python Data Structures, Lists, List
           Methods, Dictionaries",
        "rating": 4.6141534,
        "num_reviews": 132,
        "is_paid": "false",
        "instructor_names": "Rizwan Ahmed",
        "objectives": "Students will learn python programming
           language.",
        "title": "Python For Accountants I",
        "headline": "A Journey From Excel to Python",
        "category": "Finance & Accounting",
        "instructional_level": "All Levels",
        "num_subscribers": 3534
    },
    "sort": [
        4.6141534,
        132
    ]
}
```

4.2. Average Subscribers per Course Category

4.2.1. Description

This query performs an aggregation on the data to calculate the average number of enrollees for each course category; in addition, the categories are sorted according to their frequency.

4.2.2. Query

```
{
    "size": 0,
    "aggs":{
        "category": {
            "terms": {"field":"category", "order": { "_count": "desc" }},
            "aggs":{"avarage_subscribers":{
                "avg":{"field":"num_subscribers"}}
            }
        }
    }
}
```

4.2.3. Output

To improve readability, the last categories have been removed.

```
{
    "aggregations": {
        "category": {
            "doc_count_error_upper_bound": 0,
            "sum_other_doc_count": 10055,
            "buckets": [
                {
                    "key": "Development",
                    "doc_count": 9945,
                    "avarage_subscribers": {
                        "value": 16531.432378079437
                    }
                },
                {
                    "key": "Business",
                    "doc_count": 9912,
                    "avarage_subscribers": {
                        "value": 7204.035613397901
                    }
                },
                {
                    "key": "IT & Software",
                    "doc_count": 9888,
                    "avarage_subscribers": {
                        "value": 9432.101031553399
                    }
                },
                {
                    "key": "Teaching & Academics",
                    "doc_count": 9762,
                    "avarage_subscribers": {
                        "value": 2793.945912722803
                    }
                },
                {
                    "key": "Personal Development",
                    "doc_count": 9692,
                    "avarage_subscribers": {
                        "value": 3122.5803755674783
                    }
                },
                {
                    "key": "Design",
                    "doc_count": 9263,
                    "avarage_subscribers": {
                        "value": 4407.886429882327
                    }
                }
            ]
        }
    }
}
```

4.3. Top 5 most popular courses in Business

4.3.1. Description

The goal of the query is to retrieve the 5 most popular courses within the Business category, specifically based on the number of enrollees in each course. As in query 4.1, `filter` is used instead of `must` to improve efficiency.

4.3.2. Query

```
{
  "query": {
    "bool": {
      "filter": [
        {"term": {"category": "Business"}}
      ]
    }
  },
  "sort":[
      { "num_subscribers":"desc"}
  ],
  "size":5
}
```

4.3.3. Output

To improve readability, only the first two results will be displayed, and long texts have been truncated.

```
{
    "_index": "udemy_courses",
    "_id": "PHyGmJMBkz9uf3b-Vfhz",
    "_score": null,
    "_source": {
        "curriculum": "Course Introduction, Welcome Message,
           Introduction, Course Curriculum.....",
        "rating": 4.6756315,
        "num_reviews": 217617,
        "is_paid": "true",
        "instructor_names": "Jose Portilla, Pierian Training",
        "objectives": "Use SQL to query a database Use SQL to perform
           data analysis Be comfortable putting SQL and PostgreSQL on
           their resume",
        "title": "The Complete SQL Bootcamp: Go from Zero to Hero",
        "headline": "Become an expert at SQL!",
        "category": "Business",
        "instructional_level": "All Levels",
        "num_subscribers": 866379
    },
    "sort": [
        866379
    ]
},
{
    "_index": "udemy_courses",
    "_id": "PXyGmJMBkz9uf3b-Vfhz",
    "_score": null,
    "_source": {
        "curriculum": "Getting Started, READ ME: Important Notes for
           New Students....",
        "rating": 4.6441402,
        "num_reviews": 145802,
        "is_paid": "true",
        "instructor_names": "Maven Analytics, Chris Dutton, Aaron
           Parry",
        "objectives": "Build professional-quality business intelligence
           reports from the ground up Blend and transform raw data into
           beautiful interactive visuals & dashboards Design and
           implement the same tools used by professional data analysts
           and data scientists",
        "title": "Microsoft Power BI Desktop for Business Intelligence",
        "headline": "Master Power BI Desktop for data prep, data
           analysis, data visualization &amp; dashboard design w/ top
           Power BI instructors!",
        "category": "Business",
        "instructional_level": "All Levels",
        "num_subscribers": 578138
    },
    "sort": [
        578138
    ]
}
```

4.4. Comparison of ratings between Free and Paid courses by Category

4.4.1. Description

The query shows the average rating for paid and free courses within each course category.

4.4.2. Query

```
{
    "size": 0,
    "aggs":{
        "category": {
            "terms": {"field":"category"},
            "aggs":{
                "paid_courses":{
                    "filter":{"term":{"is_paid":true}},
                    "aggs":{"average_rating":{"avg":{"field":"rating"}}}
                },
                "free_courses":{
                    "filter":{"term":{"is_paid":false}},
                    "aggs":{"average_rating":{"avg":{"field":"rating"}}}
                }
            }

        }
    }
}
```

4.4.3. Output

To improve readability, the last categories have been removed.

```
{
    "buckets": [
        {
            "key": "Development",
            "doc_count": 9945,
            "paid_courses": {
                "doc_count": 9945,
                "average_rating": {
                    "value": 4.265429405437276
                }
            },
            "free_courses": {
                "doc_count": 0,
                "average_rating": {
                    "value": null
                }
            }
        },
        {
            "key": "Business",
            "doc_count": 9912,
            "paid_courses": {
                "doc_count": 9912,
                "average_rating": {
                    "value": 4.284277572674651
                }
            },
            "free_courses": {
                "doc_count": 0,
                "average_rating": {
                    "value": null
                }
            }
        },
        {
            "key": "IT & Software",
            "doc_count": 9888,
            "paid_courses": {
                "doc_count": 9888,
                "average_rating": {
                    "value": 4.270185320400015
                }
            },
            "free_courses": {
                "doc_count": 0,
                "average_rating": {
                    "value": null
                }
            }
        },
        {
            "key": "Teaching & Academics",
            "doc_count": 9762,
            "paid_courses": {
                "doc_count": 9762,
                "average_rating": {
                    "value": 4.326908442383539
                }
            },
            "free_courses": {
                "doc_count": 0,
                "average_rating": {
                    "value": null
                }
            }
        },
        {
            "key": "Personal Development",
            "doc_count": 9692,
            "paid_courses": {
                "doc_count": 9692,
                "average_rating": {
                    "value": 4.413791444593848
                }
            },
            "free_courses": {
                "doc_count": 0,
                "average_rating": {
                    "value": null
                }
            }
        },
        {
            "key": "Design",
            "doc_count": 9263,
            "paid_courses": {
                "doc_count": 8502,
                "average_rating": {
                    "value": 3.8843617413190357
                }
            },
            "free_courses": {
                "doc_count": 761,
                "average_rating": {
                    "value": 4.262028957948422
                }
            }
        }
    }
}
```

4.5. Free Python and Machine Learning Courses

4.5.1. Description

The query retrieves free courses related to Python or Machine Learning. It uses the `match` query on the `title` and `objectives` fields, and `match_phrase` is applied to search for multi-word phrases. In addition, a `minimum_should_match` parameter was specified to ensure that the query returns valid results.

4.5.2. Query

```
{
  "query": {
    "bool": {
        "minimum_should_match": 1,
      "should": [
        { "match": { "title": "Python" } },
        { "match": { "objectives": "Python" } },
        { "match_phrase": { "title": "Machine Learning" } },
        { "match_phrase": { "objectives": "Machine Learning" } }
      ],
      "filter": [
        {"term": {"is_paid": false}}
        ]
    }
  }
}
```

4.5.3. Output

To improve readability, only the first two results will be shown, long texts have been truncated, and only the relevant fields will be displayed

```
{
    "_index": "udemy_courses",
    "_id": "in2GmJMBkz9uf3b-iTo2",
    "_score": 19.446255,
    "_source": {
        "curriculum": "Content, Introduction, Concepts, Data
           Acquisition, Technical Analysis...",
        "is_paid": "false",
        "instructor_names": "Genbox Trading",
        "objectives": "How to apply Rule Induction Algorithms in
           Python...",
        "title": "Hands-on Machine Learning - Cryptocurrency Trading
           [Python]",
        "headline": "Learn how to use Machine Learning and Intermarket
           Analysis to trade Crypto",
    }
},
{
    "_index": "udemy_courses",
    "_id": "V36GmJMBkz9uf3b-_xuK",
    "_score": 17.96323,
    "_source": {
        "curriculum": "Obesity Prediction Using Machine Learning,
           Introduction Obesity...",
        "is_paid": "false",
        "instructor_names": "Soubraylu Sivakumar",
        "objectives": "Understanding of Obesity and Its Implications -
           Gain comprehensive...",
        "title": "Obesity Prediction Using Machine Learning Practical
           Approach",
        "headline": "Machine Learning models for Health Care (Obesity)",
    }
},
{
    "_index": "udemy_courses",
    "_id": "P32GmJMBkz9uf3b-1Lra",
    "_score": 17.37134,
    "_source": {
        "curriculum": "Essential Foundations, Must Watch: Discover
           Essential Resources...",
        "is_paid": "false",
        "instructor_names": "Corbin Brown",
        "objectives": "Fundamentals of digital image & video
           processing....",
        "title": "AI and Machine Learning for Image and Video Editing",
        "headline": "Empowering Visual Perception: Techniques and
           Applications in AI-driven Image and Video Analysis",
    }
}
```

4.6. Course Distribution by Instructional Level

4.6.1. Description

The query returns how many courses exist for each level. These values are useful for understanding how courses are distributed according to their difficulty.

4.6.2. Query

```
{
  "size": 0,
  "aggs": {
    "courses_per_experience_level": {
      "terms": {"field": "instructional_level"}
    }
  }
}
```

4.6.3. Output

```
{
    "aggregations": {
        "courses_per_experience_level": {
            "doc_count_error_upper_bound": 0,
            "sum_other_doc_count": 0,
            "buckets": [
                {
                    "key": "All Levels",
                    "doc_count": 53354
                },
                {
                    "key": "Beginner Level",
                    "doc_count": 31550
                },
                {
                    "key": "Intermediate Level",
                    "doc_count": 11455
                },
                {
                    "key": "Expert Level",
                    "doc_count": 1744
                }
            ]
        }
    }
}
```

4.7. Business Courses with Preference for High Rating or Popularity

4.7.1. Description

The query retrieves all Business courses, giving preference, through the `should` clause, to courses with a high rating (greater than or equal to 4.5) or a significant number of subscribers (greater than or equal to 10,000).

4.7.2. Query

```
{
  "query": {
    "bool": {
      "should": [ // Preference
        { "range": { "rating": { "gte": 4.5 } } },
        { "range": { "num_subscribers": { "gte": 10000 } } }
      ],
      "filter": [
        {"term": {"category": "Business"}}
        ]
    }
  }
}
```

4.7.3. Output

To improve readability, only the first 3 results will be displayed, and only relevant fields will be displayed

```
{
    "_index": "udemy_courses",
    "_id": "eHyGmJMBkz9uf3b-Vvl0",
    "_score": 2.0,
    "_source": {
        "rating": 4.8133125,
        "title": "Advanced SQL: MySQL for Ecommerce Data Analysis",
        "num_subscribers": 45792
    }
},
{
    "_index": "udemy_courses",
    "_id": "e3yGmJMBkz9uf3b-Vvl0",
    "_score": 2.0,
    "_source": {
        "rating": 4.5931826,
        "title": "Marketing Customer Analytics, Segmentation, and
           Targeting",
        "num_subscribers": 13750
    }
},
{
    "_index": "udemy_courses",
    "_id": "RnyGmJMBkz9uf3b-Vfh0",
    "_score": 2.0,
    "_source": {
        "rating": 4.602804,
        "title": "Microsoft Power BI - The Practical Guide 2024",
        "num_subscribers": 293769
    }
}
```

4.8. Statistical Insights on Development Courses

This query focuses on courses in the Development category, performing aggregations to provide key statistical information. It calculates the average rating of all Development courses, identifies the maximum number of enrollees among these courses, and determines the minimum number of enrollees.

4.8.1. Query

```
{
  "query": {
    "term": {
      "category": "Development"
    }
  },
  "aggs": {
    "average_rating": { "avg": { "field": "rating" } },
    "max_subscribers": { "max": { "field": "num_subscribers" } },
    "min_subscribers": { "min": { "field": "num_subscribers" } }
  },
  "size": 0
}
```

4.8.2. Output

```
{
    "aggregations": {
        "average_rating": {
            "value": 4.265429405437276
        },
        "max_subscribers": {
            "value": 1976866.0
        },
        "min_subscribers": {
            "value": 69.0
        }
    }
}
```

4.9. Highly Rated Courses in Business and Finance

4.9.1. Description

This query retrieves courses from the Business and Finance & Accounting categories that have a rating of 4.0 or higher

4.9.2. Query

```
{
  "query": {
    "bool": {
      "must": [
        { "terms": { "category": ["Business", "Finance &
            Accounting"] } },
        { "range": { "rating": { "gte": 4.0 } } }
      ]
    }
  }
}
```

4.9.3. Output

To improve readability, only the first 3 results will be displayed, and only relevant fields will be displayed

```
{
    "_index": "udemy_courses",
    "_id": "uVKa2ZMBhaD-1vHmHwqK",
    "_score": 2.0,
    "_source": {
        "category": "Business",
        "title": "Presentation Skills: Master Confident Presentations",
        "rating": 4.5968637
    }
},
{
    "_index": "udemy_courses",
    "_id": "ulKa2ZMBhaD-1vHmHwqK",
    "_score": 2.0,
    "_source": {
        "category": "Business",
        "title": "Business Fundamentals: Corporate Strategy",
        "rating": 4.569851
    }
},
{
    "_index": "udemy_courses",
    "_id": "u1Ka2ZMBhaD-1vHmHwqK",
    "_score": 2.0,
    "_source": {
        "category": "Business",
        "title": "Robotic Process Automation - RPA Overview",
        "rating": 4.547368
    }
}
```

4.10. Top Business courses for all levels with high ratings and reviews

4.10.1. Description

The query retrieves Business courses designed for students of all levels. Courses are filtered to include only those with a rating of 3.0 or higher and at least 50 reviews. Results are sorted in descending order, first by rating and then by number of reviews, to ensure that courses with the highest ratings and reviews appear at the top.

4.10.2. Query

```
{
  "query": {
    "bool": {
      "filter": [
        { "term": { "instructional_level": "All Levels" } },
        { "range": { "rating": { "gte": 3.0} } },
        { "range": { "num_reviews": { "gte": 50 } } },
        {"term": {"category": "Business"}}
      ]
    }
  },
  "sort":[
      { "rating":"desc"},
      { "num_reviews":"desc"}
  ]
}
```

4.10.3. Output

To improve readability, only the first 3 results will be displayed, and only relevant fields will be displayed

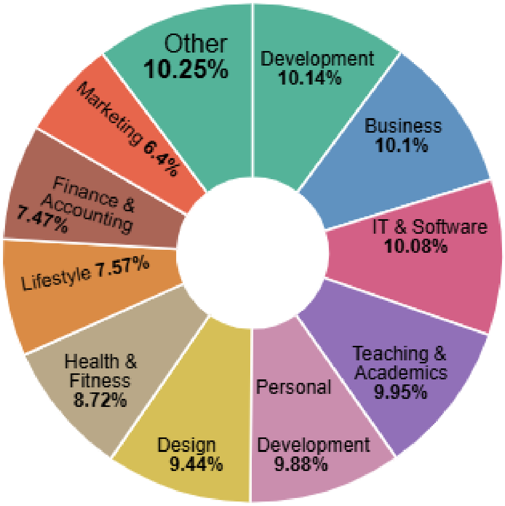
```
{
    "_index": "udemy_courses",
    "_id": "g3yGmJMBkz9uf3b-W_9V",
    "_score": null,
    "_source": {
        "rating": 5.0,
        "num_reviews": 692,
        "title": "How a Building is Designed and Built - Part 1 of 6",
        "category": "Business",
        "instructional_level": "All Levels"
    },
    "sort": [
        5.0,
        692
    ]
},
{
    "_index": "udemy_courses",
    "_id": "V32GmJMBkz9uf3b-YgMt",
    "_score": null,
    "_source": {
        "rating": 5.0,
        "num_reviews": 244,
        "title": "How To Analyze Passive Real Estate Investment
           Opportunities",
        "category": "Business",
        "instructional_level": "All Levels"
    },
    "sort": [
        5.0,
        244
    ]
},
{
    "_index": "udemy_courses",
    "_id": "PH2GmJMBkz9uf3b-ZwfM",
    "_score": null,
    "_source": {
        "rating": 5.0,
        "num_reviews": 133,
        "title": "Generative AI Startup Strategy Case Studies | Sramana
           Mitra",
        "category": "Business",
        "instructional_level": "All Levels"
    },
    "sort": [
        5.0,
        133
    ]
}
```

5 | Extra

To fully visualize and understand the dataset, Kibana was used to create graphs to facilitate data analysis.

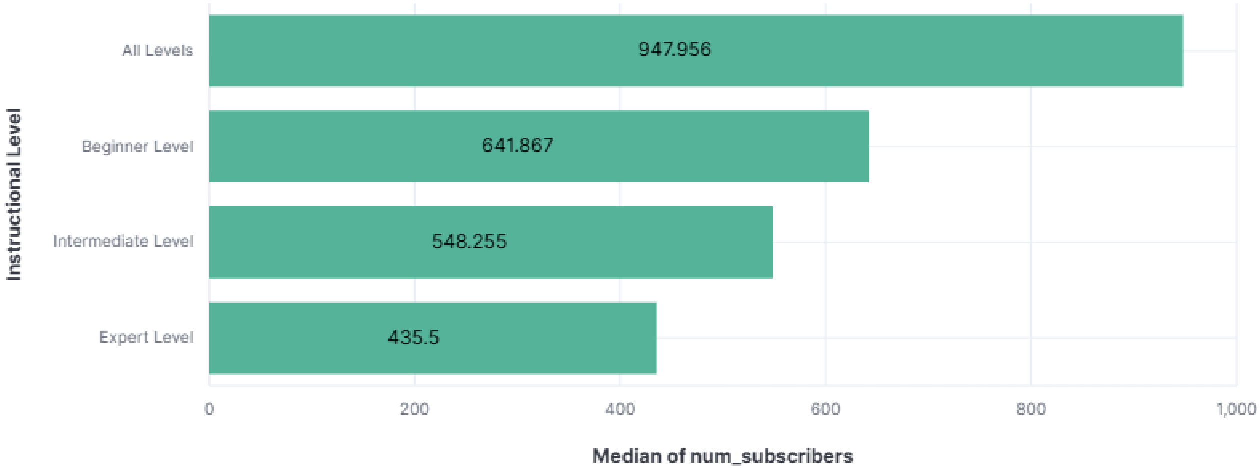
5.1. Category distribution

The pie chart shows the distribution of course categories, providing a visual representation of the relative proportions of each category within the dataset. This visualization makes it easy to identify which categories are predominant and which are less represented.



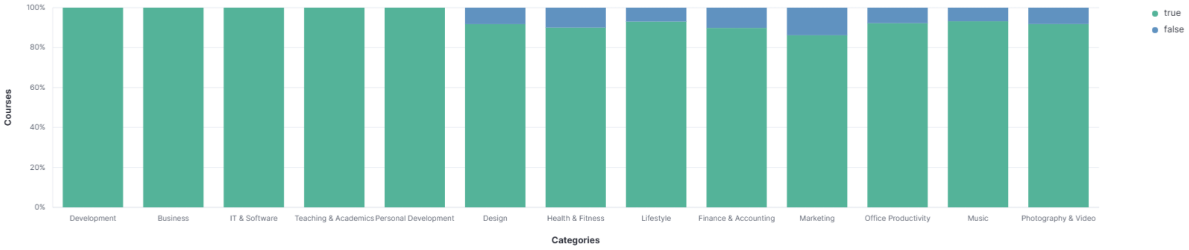
5.2. Subscribers by Instructional Level

The bar graph shows the median number of enrollees for each course belonging to a specific educational level. This visualization helps to identify trends in the number of enrollments among different educational levels.



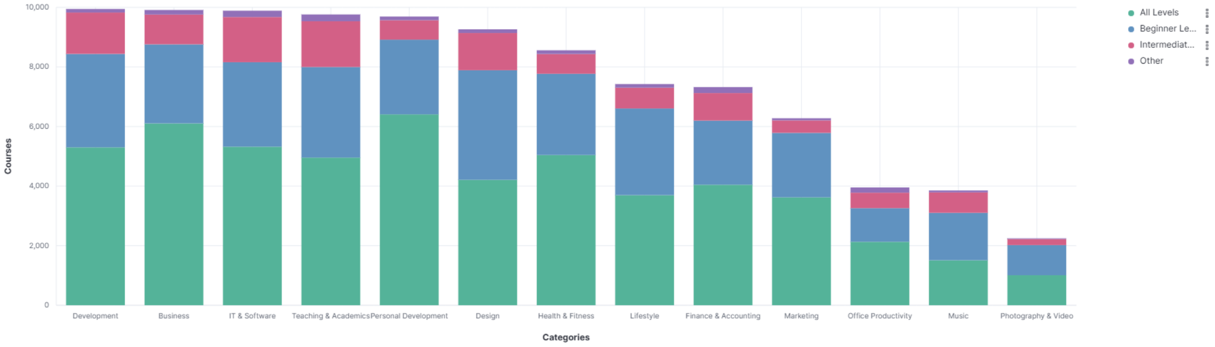
5.3. Free vs paid courses by category

The chart compares the percentage of free (blue) and paid (green) courses across categories, visually highlighting the percentage of each type within their respective categories.



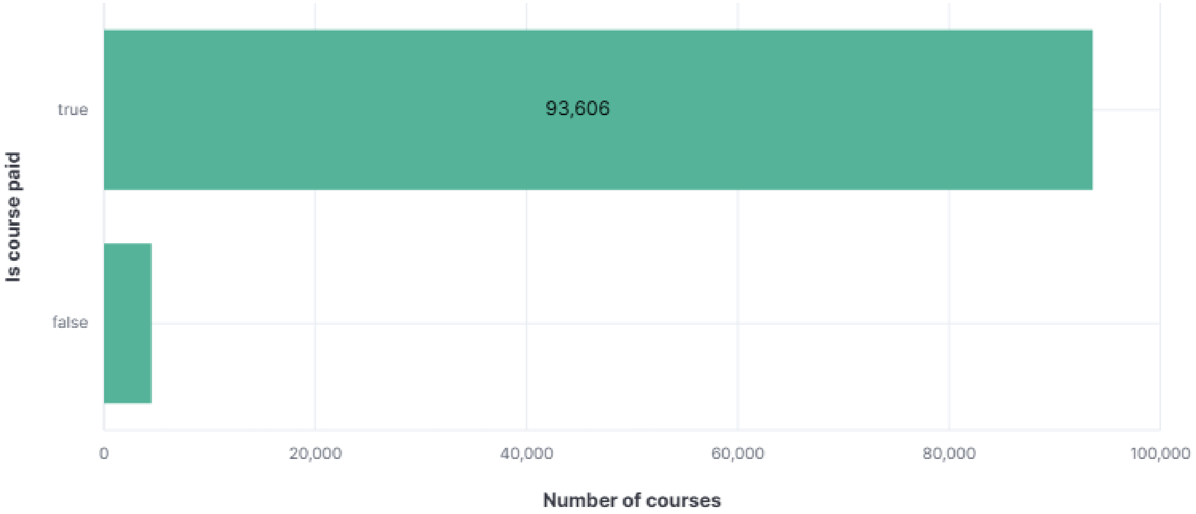
5.4. Number of courses by instructional level for each category

The bar graph shows the number of courses by education level for each category, with the categories listed on the horizontal axis and the corresponding number of courses on the vertical axis. This visualization allows for a clear comparison of the distribution of courses among the different levels of education within each category.



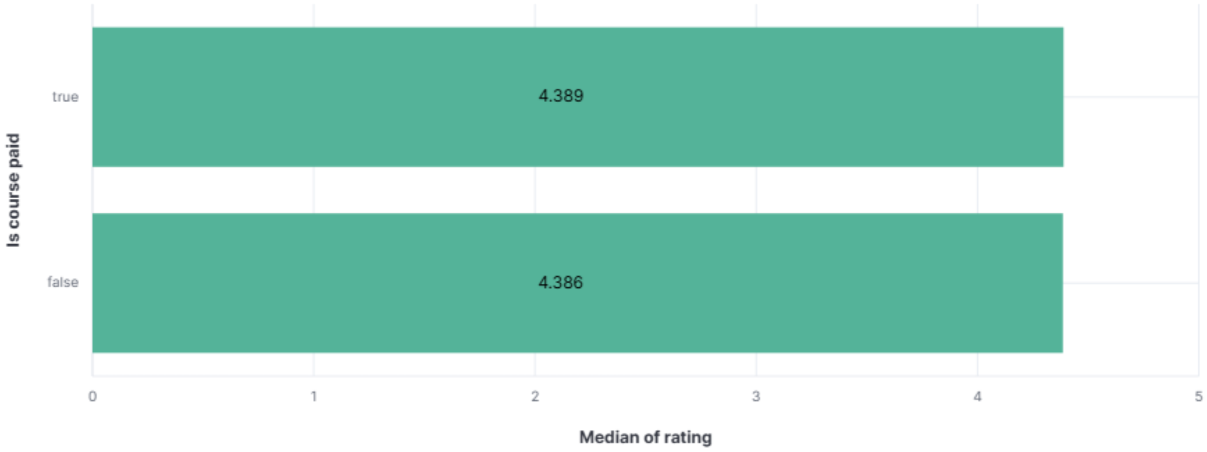
5.5. Comparison of the number of free and paid courses

The bar chart compares the number of free and paid courses. The y-axis represents whether a course is paid ("true") or free ("false"), while the x-axis represents the number of courses.



5.6. Comparison of median ratings for free and paid courses

This bar chart compares the median ratings of free and paid courses. The y-axis indicates whether the course is paid ("true") or free ("false"), while the x-axis represents the median rating


true
false
Is course paid
4.389
4.386
0
1
2
3
4
5
Median of rating