Faculty of New Technologies of Information and Communication

Department of Computer Science and Information Technology

Artificial intelligence and Data science

Exploratory Data Analysis Project

Project Title : Analysis on Wikipedia (Toys and Games)

Prepared By

Mohammed taha Khamed

Salaheddine Khenfer

Supervised by

Bilel khaldi

Submission Date : 12 / 10 / 2024

Table of Contents

1. Project Overview

Project Title : **Analysis on Wikipedia (Toys and Games) .**

Goal of the project :

The aim of this project is to explore and analyze an in-depth dataset (texts and images) to identify patterns, relationships and trends. The goal is to conduct data cleaning and visualization and statistical analysis to obtain insights that will help in decision-making and build a search engine .

Dataset used : Wikipedia "Games and Toys" Search .

Analysis Date : 12 / 01 / 2024

Analyst Team :

Mohammed taha KHAMED

Salaheddine Khenfer

2. Data Collection and Initial Inspection

Source of the Data : advanced_toys_dataset.csv collected directly from Wikipedia after **scraping** pages related to **toys** and **games**

Data Size : **3.21 Mib**

Brief Description of the Data :

The dataset consists of articles and information gathered from Wikipedia's "Games and Toys" search. It includes :

- Textual data related to various toys, games, historical context, categories, and their significance .
- Visual content such as **images of toys and games** to enhance understanding and provide better visual representation .
- Screenshots and examples of relevant **Wikipedia pages** used for analysis .
- Statistical **charts and graphs** to illustrate trends and insights extracted from the data .

3. Data Cleaning and Transformation

Handling Missing Values

Rows with missing values in the image_urls column were dropped .

Missing or unspecified age groups were handled using heuristic rules .

Outlier Detection and Removal

Not explicitly mentioned in the code .

Data Transformation

Scaling : Not applied .

Encoding : Not applied (features are primarily numerical or text-based) .

Additional Transformations :

A heuristic rule was applied to predict the age_group based on topic_keywords where "unspecified" :

- Keywords like *"toy", "children"* → "children"
- Keywords like *"game", "teen"* → "teen"
- Default → "adult"

4. Feature and Sample Analysis

Univariate and Bivariate Analysis

Age Group Distribution : Pie chart of age_group_hint and predicted_age_group.

Central Tendency & Dispersion : Summary statistics for numerical features complexity_score, text_length, etc

Correlation Analysis

- Correlation Findings : A correlation heatmap was generated for numerical features . text_length, sentences_count, complexity_score, readability_score . and visualized distributions grouped by predicted_age_group .
- Strong or weak relationships between numerical features are highlighted in the heatmap.

5. Visualization of Key Patterns

Visualization Techniques

1. Pie Charts : Distribution of age groups (age_group_hint, predicted_age_group) .
2. Histograms : Distribution of text_length, sentences_count, and other numerical features by age group .
3. Heatmaps : Correlation matrix of selected numerical features .

Observed Patterns

1. Age group distribution highlights dominance of specific groups .
2. Text complexity and readability vary across age groups .
3. Features like text_length and complexity_score show significant correlations .

6. Summary Statistics and Key Insights

Summary of Central Tendency and Variability

Key numerical features (text_length, complexity_score) were analyzed for mean, median, and variance .

Interpretation of Statistics

Notable differences in text complexity and readability scores were observed across predicted age groups .

7. Conclusion

Summary of Findings :

- The dataset reveals significant patterns in age group distributions and text complexity .
- Heuristic predictions successfully classified unspecified age groups .
- Correlations between features provide insights for further analysis .

Next Steps:

- Explore machine learning models to automate age group predictions .
- Perform deeper analysis of outliers and anomalies .

Importance of Insights :

- Insights can help businesses and educators target specific age groups more effectively .

Notebook Link :

For further details on the analysis and code, please refer to the Colab Notebook at :
https://colab.research.google.com/drive/13N1m6NGFBv5S8y36uDEZbuhmys-3emAI