Data Cleaning and Preprocessing on Netflix Originals Data

Objective: Students will tackle the data with different cleaning and preprocessing techniques. They will apply which categorical encoding and feature scaling technique to apply to make the data ML ready.

- Perform the initial analysis on the dataset. Print the first 10 rows of the data to see how the data looks like.
- Print the shape of the data to find the number of rows and columns.
- The genre column has multiple values in each row. As a part of cleaning, extract only the first genre from each row.
- Find the number of missing values in the data. If there are missing values, what method will you choose for each of the columns? Choose between mean/median/mode imputation, ffill, bfill, dropping the missing values. Please provide clear justification for choosing a particular approach.
- Plot the histograms for numerical data.
- Plot the bar chart for top 10 genres.
- Plot the line chart for the number of releases per year.
- Plot the sunburst chart for the distribution of genres by content type. (Use plotly library).
- Perform encoding on the categorical data. Choose between label encoding, one hot encoding or ordinal encoding and provide a clear justification for choosing a particular encoding method.
- Identify if there are any outliers in the data. If so, handle the outliers (Refer to the lecture notebook).
- If given an option to replace the outliers with 0, would you do that? Why or why not? Provide clear justification.
- Perform all feature scaling techniques using StandardScaler, MinMaxScaler and RobustScaler. Which method is the most appropriate for this data and why? Please provide a clear explanation.

Submission:
Please provide a clear code with comments and justification for all the areas.
Submit a .ipynb notebook in classrooms.