**Case Study Project: Unlocking Content Intelligence, A Deep Dive into Netflix’s Global Library**

**Client Brief:**

Netflix's **Content Strategy & Analytics Team** wants to make data-driven decisions on licensing, regional content expansion, and user personalization. As a Junior Data Analyst intern, you have been asked to prepare a detailed exploratory report using their internal dataset of titles.

The leadership team is particularly interested in understanding patterns in the catalog and how those might inform strategic initiatives, such as producing local-language content or acquiring longer or shorter titles.

**Your Role:**

You will explore, clean, transform, and analyze the data using advanced exploratory techniques. Your report should include visual insights, statistical summaries, feature insights, and practical takeaways, all using **Python, Pandas, NumPy, and Seaborn/Matplotlib**.

Data Link: [netflix\_titles.csv](https://github.com/prasertcbs/basic-dataset/blob/master/netflix_titles.csv)

**Section 1: Data Familiarization and Cleaning**

1. Load the Netflix dataset and display basic info (shape, columns, datatypes).
2. Identify and handle missing values:
   * Use dropna() for critical fields.
   * Use fillna() where appropriate (e.g., fill missing country or rating).
   * Experiment with **IterativeImputer** for fields with numerical implications (like release\_year).
3. Parse and clean columns:
   * Extract numeric values from duration.
   * Convert date\_added to datetime format.
   * Create a new column for **content age** (current year – release year).

**Section 2: Feature Engineering**

1. **Normalize** and **Standardize** the numeric features:
   * Use StandardScaler() and Normalizer() from sklearn.preprocessing.
   * Apply on duration (converted to minutes or seasons), content\_age, etc.
2. Bin content into release year ranges (e.g., Classic, 90s, 2000s, 2010s, Modern).
3. Create a binary column indicating whether content was added in the last 2 years.

**Section 3: Univariate Analysis**

1. What’s the distribution of:
   * Content type (TV Show vs Movie)?
   * Duration?
   * Genres (listed\_in)?
   * Content age?
2. Use:
   * Histogram
   * Countplot
   * Boxplot
   * Violinplot
3. Compute and interpret:
   * Mean, Median, Mode, Stdev, IQR

**Section 4: Bivariate Analysis**

1. Analyze:
   * Duration vs. Genre
   * Type vs. Content Age
   * Content Age vs. Country
   * Genre vs. Country
2. Use:
   * Scatterplot
   * Boxplot
   * Barplot
   * Jointplot

Use .groupby(), .agg(), sns.scatterplot, sns.jointplot, pd.crosstab, np.select()

**Section 5: Multivariate Analysis**

1. Analyze:
   * How has genre distribution changed over time? (Genre × Year × Type)
   * Does country × type × duration reveal regional content patterns?
2. Visualizations:
   * Heatmap (Correlation matrix)
   * Grouped barplots or stacked bars
   * Pivot tables and .crosstab()

**Section 6: Outlier Analysis**

1. Identify outliers in duration and content\_age:
   * Using **Boxplot**, **Z-score**, **Modified Z-score**, and **IQR**.
2. List top 5 most unusual records and explain why they are outliers.

Use filtering with thresholds, np.abs(zscore) > 3, boxplot, etc.

**Section 7: Advanced Filtering and Discovery**

1. Find:
   * All content from Pakistan or in Urdu.
   * All shows longer than 3 seasons.
   * Movies with durations > 2.5 hours.
   * Top 5 most frequent actors or directors.
2. Use:
   * str.contains(), .isin(), groupby().count(), np.select()