Best-Selling Manga Dataset Analysis

1. Data Cleaning

Loading and Inspecting Data

- The dataset was loaded and inspected for its structure, including numerical and categorical variables.
- Key columns: **Manga Series, Author(s), Publisher, Demographic, No. of Collected Volumes, Serialized Year, Approximate Sales, Average Sales per Volume.**

Handling Missing Values

- Checked for missing values across all columns.
- Applied **imputation or removal techniques** where necessary.

Identifying and Removing Duplicates

- Checked for duplicate records in the dataset.
- Any duplicate entries were removed to ensure data integrity.

Detecting and Treating Outliers

- Used **boxplots and statistical methods** to detect extreme values.
- Outliers in **Approximate Sales and No. of Volumes** were examined and addressed accordingly.

Standardizing Categorical Values

- Fixed inconsistencies in categorical values (e.g., correcting typos in **Demographics, Publisher Names**).

---

2. Exploratory Data Analysis (EDA)

Univariate Analysis

Summary Statistics:

- Calculated **mean, median, variance, skewness**, etc., for numerical variables.

Frequency Distributions:

- Analysed **Demographic and Publisher distributions** to understand data composition.
- **Shonen demographic dominated**, followed by **Seinen**.

Histograms and Box Plots:

- **Histograms** visualized the distributions of numerical variables like **Approximate Sales, Number of Volumes, and Average Sales per Volume**.
- **Box plots** were used to detect **skewness and outliers** in sales and volumes.

Bivariate Analysis

Correlation Matrix:

- Identified relationships among numerical variables.
- **Approximate Sales and Number of Collected Volumes showed a positive correlation**.

Scatter Plots:

- **Serialization Year vs. Approximate Sales**: Older manga tend to have higher total sales.
- **Average Sales per Volume vs. Number of Collected Volumes**: Shorter series tend to have higher per-volume sales.

Box Plots & Violin Plots:

- **Demographic vs. Sales**: Shonen had the highest median sales.
- **Publisher vs. Sales**: Shueisha and Kodansha dominated the best-selling manga list.

Multivariate Analysis

Pair Plots:

- Explored multiple relationships between **Sales, Volumes, and Serialized Year**.
- Highlighted clusters based on **Demographics and Publishers**.

Heatmaps:

- Showed correlation strengths among **Sales, Volumes, and Serialization Periods**.

Grouped Comparisons:

- Analysed the impact of multiple factors (e.g., **Sales vs. Serialization Year, grouped by Demographic**).

---

3. Key Findings & Insights

- **Shonen manga dominate in terms of sales**, proving their widespread popularity.
- **Long-running series tend to accumulate higher total sales**, while **shorter series achieve higher per-volume sales**.
- **Publishers like Shueisha and Kodansha contribute to most of the top-selling manga**.
- **Serialization period plays a role in success**, with **older manga having an advantage in total sales**.
- **Demographics affect sales distribution**, with **Seinen and Shonen leading, while Shoujo and Kodomo have smaller markets**.

Conclusion

This analysis provided deep insights into the best-selling manga dataset, revealing key trends and success factors. Understanding **serialization trends, publisher dominance, and demographic influences** helps explain what makes a manga commercially successful.