

SCREEN SENSE:-KIDS' SCREENTIME VISUALIZATION

WEEK 2 REPORT – PREPROCESSING AND FEATURE ENGINEERING

Objective:

The goal of Week 2 is to prepare the raw Indian Kids Screen Time dataset for analysis and modeling. This involves cleaning and standardizing the dataset, handling missing and inconsistent data, engineering new features, saving a reusable preprocessed version, and documenting all logic.

1. **Data Loading and Column Standardization** The dataset was read into pandas and column names were standardized to lowercase with underscores for consistency.

Code:

```
df = pd.read_csv("Indian_Kids_Screen_Time.csv")
df.columns = df.columns.str.strip().str.replace(" ", "_").str.replace("-", "_").str.lower()
```

Explanation:

- Reads the dataset into a pandas DataFrame.
- Column names are standardized to lowercase with underscores for consistency
 - Example: Primary Device → primary_device

Purpose: Ensures consistency across code references and avoids key errors in transformations.

2. Handling Missing Values

- Numeric columns → filled with **median**
- Categorical columns → filled with **mode**
- Any remaining unmapped text values default to "Unknown"

Code:

```
for col in df.columns:
```

```
    if df[col].dtype in ['int64', 'float64']:
```

```
        df[col].fillna(df[col].median(), inplace=True)
```

```
    else:
```

```
df[col].fillna(df[col].mode()[0] if not df[col].mode().empty else
"Unknown", inplace=True)
```

Purpose:

- Prevents data loss from row deletions.
- Median is robust to outliers in numeric data.
- Mode preserves most frequent categorical category.

After filling, all missing values were confirmed to be resolved.

3. Standardizing Categorical Variables

- Converts all text-based columns to lowercase and trims whitespace.
- Harmonizes inconsistent category names.

Code:

```
cat_cols = ['gender', 'primary_device', 'exceeded_recommended_limit',
'health_impacts', 'urban_or_rural']
```

```
for col in cat_cols:
```

```
    df[col] = df[col].astype(str).str.strip().str.lower()
```

```
df['gender'] = df['gender'].replace({'m': 'male', 'f': 'female'})
```

```
df['primary_device'] = df['primary_device'].replace({'mobile phone':
'mobile', 'cellphone': 'mobile', 'tv ': 'tv'})
```

Ensures consistency for grouping, one-hot encoding, and visualization.

"M" and "male" → **male**

"cellphone" and "mobile phone" → **mobile**

4. Derived Feature Creation

Age Banding

Logic: Ages are bucketed into developmental categories for meaningful aggregation.

Code:

```
bins = [0, 3, 7, 11, 15, 18]
```

```
labels = ['0-3', '4-7', '8-11', '12-15', '16-18']
```

```
df['age_band'] = pd.cut(df['age'], bins=bins,  
labels=labels,include_lowest=True)
```

- **Purpose:**

Enables analysis by age group rather than exact age, improving interpretability.

5. Final Preprocessed Dataset

Code:

```
df.to_csv("Cleaned_Indian_Kids_Screen_Time.csv",  
index=False)
```

It Saves the cleaned dataset for reuse in modeling or visualization without repeating preprocessing.

6. Feature Dictionary

Feature Name	Type	Description
age	Numeric	Age of the child (years)
gender	Categorical	Gender of the child (male/female)
avg_daily_screen_time_hr	Numeric	Average screen time (hours per day)
primary_device	Categorical	Most used device for screen usage
exceeded_recommended_limit	Binary	1 if exceeded WHO screen time limit, else 0

Feature Name	Type	Description
educational_to_recreational_ratio	Numeric (Derived)	Ratio of educational to recreational screen use
health_impacts	Categorical	Reported health issues (e.g., eye strain, headache)
urban_or_rural	Binary	1 for urban residence, 0 for rural
age_band	Categorical (Derived)	Categorized age ranges (0–3, 4–7, 8–11, 12–15, 16–18)
has_health_issue	Binary (Derived)	1 if any health issue reported, else 0
screen_time_zscore	Numeric (Derived)	Normalized screen time (Z-score)

Preprocessing Summary:

Total Rows: 9712

Total Columns: 13

Features Created: ['age_band', 'has_health_issue', 'screen_time_zscore']

Missing Values Remaining: 0

Conclusion:

The Week 2 preprocessing and feature engineering phase successfully transformed the raw Indian Kids Screen Time dataset into a clean, consistent, and analysis-ready format.

All missing values were handled appropriately using statistical imputation techniques, categorical inconsistencies were standardized, and several insightful derived features were created to enhance analytical depth.

Key engineered features — such as age bands, educational-to-recreational ratio, health issue indicators, and screen time z-scores — provide a solid foundation for understanding behavioral patterns and conducting meaningful statistical or predictive modeling in later phases.

The finalized dataset now exhibits:

- No missing or inconsistent entries
- Standardized and interpretable categories
- Rich, domain-relevant derived metrics
- Documented transformation logic ensuring full reproducibility

SUBMITTED BY:LAKSHMI REDDY KOTTAM