Name: Harshit Pandey
Email: pandeyharshit7277@gmail.com

Week 2 Deliverables — Preprocessing and Feature Engineering

1. Objective

The main objective of Week 2 was to clean and prepare the dataset for further analysis and visualization. This included handling missing values, fixing inconsistent data entries, converting incorrect data types, and creating new derived features such as Age_Band, Usage_Level, and Is_Urban ,Educational_share and Recreational_Share to make the dataset ready for analysis in the next stage.

2. Tasks Performed

a. Handled Missing Values

- Checked for missing values in all columns.
- Since numeric columns were complete, only categorical columns with missing entries were cleaned.
- Filled missing categorical values using the **most frequent (mode)** value for each column.

b. Cleaned Text Data

- Standardized capitalization and removed unnecessary spaces.
- Ensured text consistency in columns such as *Primary_Device*, *Health_Impacts*, and *Urban_or_Rural*.

c. Created Derived Columns

- **Age_Band:** Divided ages (8–18) into four categories — *8_to_10*, *11_to_13*, *14_to_15*, *16_to_18*.
- **Usage_Level:** Classified screen time hours into five categories — *Low (0–2 hrs)*, *Moderate (2–5 hrs)*, *High (5–8 hrs)*, *Very High (8–12 hrs)*, *Extreme (12–15 hrs)*.
- **Is_Urban:** Added a flag where Urban = 1 and Rural = 0, created without using lambda functions.
- **Activity Share(educational and recreational percentages):**To understand the screentime share for educational and entertainment purposes.

d. Saved Cleaned Dataset

The final cleaned dataset was exported as **Cleaned_Kids_ScreenTime.csv** for reuse in Week 3.

Name: Harshit Pandey
Email: pandeyharshit7277@gmail.com

3. Code (Implementation)

[108]:
```
import pandas as pd
import numpy as np
```

[109]:
```
#Load dataset
df = pd.read_csv("Indian_Kids_Screen_Time.csv")
```

[110]:
```
#checking for missing values
print("Missing values before cleaning:\n",df.isnull().sum())
```

```
Missing values before cleaning:
 Age                                0
Gender                              0
Avg_Daily_Screen_Time_hr            0
Primary_Device                      0
Exceeded_Recommended_Limit          0
Educational_to_Recreational_Ratio   0
Health_Impacts                   3218
Urban_or_Rural                      0
dtype: int64
```

[111]:
```
#Fill missing values with column mode in categorical columns
cat_cols = df.select_dtypes(include='object').columns
for col in cat_cols:
    df.fillna({col:df[col].mode()[0]},inplace=True)
```

[112]:
```
df['Health_Impacts']=df.apply(
    lambda row: 'No Impact' if row['Avg_Daily_Screen_Time_hr']==0 else row['Health_Impacts'],axis=1
)
```

[113]:
```
print("Missing Values after cleaning:\n",df.isnull().sum())
```

```
Missing Values after cleaning:
 Age                                0
Gender                              0
Avg_Daily_Screen_Time_hr            0
Primary_Device                      0
Exceeded_Recommended_Limit          0
Educational_to_Recreational_Ratio   0
Health_Impacts                      0
Urban_or_Rural                      0
dtype: int64
```

[114]:
```
#Handle inconsistent text data (fixing capitalization and remove unwanted spaces)
text_cols = ['Primary_Device','Exceeded_Recommended_Limit','Health_Impacts','Urban_or_Rural']
for col in text_cols:
    df[col] = df[col].astype(str).str.strip().str.title()
```

[115]:
```
#Creating derived features
#1)Age Band
bins=[7,10,13,15,18]
labels=['8_to_10','11_to_13','14_to_15','16_to_18']
df['Age_Band'] = pd.cut(df['Age'],bins=bins,labels=labels,right=True)

#2)Screen time level
df['Usage_Level'] = pd.cut(
    df['Avg_Daily_Screen_Time_hr'],
    bins=[-0.1,2,5,8,12,15],
    labels=['Low','Moderate','High','Very High','Extreme']
)

#3)Urban/Rural
df['Is_Urban']=df['Urban_or_Rural'].apply(lambda x: 1 if x.lower() == 'urban' else 0)

#4)Activity shares
# Recreational screen time
df['Recreational_Screen_Time'] = df['Avg_Daily_Screen_Time_hr'] / (1 + df['Educational_to_Recreational_Ratio'])

# Educational screen time
df['Educational_Screen_Time'] = df['Avg_Daily_Screen_Time_hr'] - df['Recreational_Screen_Time']

df['Educational_Share'] = df.apply(
    lambda row: (row['Educational_Screen_Time'] / row['Avg_Daily_Screen_Time_hr']) * 100
    if row['Educational_Screen_Time'] != 0 and row['Recreational_Screen_Time'] != 0 else 0,
    axis=1
)

df['Recreational_Share'] = df.apply(
    lambda row: (row['Recreational_Screen_Time'] / row['Avg_Daily_Screen_Time_hr']) * 100
    if row['Educational_Screen_Time'] != 0 and row['Recreational_Screen_Time'] != 0 else 0,
    axis=1
)
```

Name: Harshit Pandey
Email: pandeyharshit7277@gmail.com

```
[116]: #Preview
       print("Preview of derived columns:")
       print(df[['Age','Age_Band','Avg_Daily_Screen_Time_hr','Usage_Level','Urban_or_Rural','Is_Urban']].head(7))

Preview of derived columns:
   Age  Age_Band  Avg_Daily_Screen_Time_hr Usage_Level Urban_or_Rural  \
0   14  14_to_15                      3.99    Moderate          Urban
1   11  11_to_13                      4.61    Moderate          Urban
2   18  16_to_18                      3.73    Moderate          Urban
3   15  14_to_15                      1.21         Low          Urban
4   12  11_to_13                      5.89        High          Urban
5   14  14_to_15                      4.88    Moderate          Urban
6   17  16_to_18                      2.97    Moderate          Rural

   Is_Urban
0         1
1         1
2         1
3         1
4         1
5         1
6         0

[117]: #saving cleaned dataset
       df.to_csv('Cleaned_Kids_ScreenTime.csv',index=False)
       print("cleaned dataset saved succesfully")

cleaned dataset saved succesfully
```

4. Feature Dictionary

Feature Name	Description	Type
Age	Age of the child (cleaned numeric column)	Numeric
Gender	Gender of the child	Categorical
Avg_Daily_Screen_Time_hr	Average daily screen time (in hours)	Numeric
Primary_Device	Device most used by the child	Categorical
Exceeded_Recommended_Limit	Whether the child exceeded recommended screen time	Categorical
Educational_to_Recreational_Ratio	Ratio of educational use to recreational use	Numeric
Educational_Share	Percentage of total screen time spent on educational activities	Numeric
Recreational_Share	Percentage of total screen time spent on recreational activities	Numeric
Health_Impacts	Reported health effects (e.g., Eye Strain, Poor Sleep)	Categorical
Urban_or_Rural	Indicates whether the child lives in an urban or rural area	Categorical
Age_Band	Derived age group (8_to_10, 11_to_13, 14_to_15, 16_to_18)	Categorical
Usage_Level	Categorized screen time level (Low–Extreme)	Categorical
Is_Urban	Binary flag (1 = Urban, 0 = Rural)	Numeric

Name: Harshit Pandey
Email: pandeyharshit7277@gmail.com

5. Conclusion

In Week 2, the dataset was cleaned, standardized, and enhanced with new calculated features.
All missing and inconsistent entries were resolved, invalid age and screen-time values corrected, and new columns such as Age_Band, Usage_Level, Is_Urban, Educational_Share, and Recreational_Share were created.
The cleaned and feature-rich dataset is now structured, reliable, and ready for detailed visualization and insight generation in Week 3.