Week 2: Preprocessing and Feature Engineering

1. Handling Missing Values and Inconsistent Categories

In this stage, we examined the dataset for missing or inconsistent values that could affect analysis accuracy. Missing numerical data was filled using appropriate measures like the median, while missing categorical values were replaced with default placeholders. Category names (like device types) were standardized to maintain consistency and avoid duplicates caused by spelling or spacing differences.

2. Creating Derived Fields (Feature Engineering)

New meaningful features were generated from the existing data to improve insights and model performance: - **Age Bands:** Categorized users into age groups (e.g., 5–8 years, 9–12 years) for better segmentation. - **Weekday/Weekend Flags:** Added a flag to differentiate between weekdays and weekends to identify behavioral patterns. - **Device/Activity Shares:** Calculated the proportion of total screen time spent on each device for better understanding of usage patterns.

3. Formatting Date and Time Fields

All date and time data were converted into a standard datetime format to enable easy extraction of components such as day, month, or year for time-based analysis.

4. Saving Preprocessed Data for Reuse

After cleaning and feature engineering, the processed dataset was saved as a new file for reuse in further analysis or modeling without repeating preprocessing steps.

5. Documenting the Logic

A preprocessing summary and feature dictionary were created to record each step of data transformation: - **Preprocessing Summary:** Explains each step and its purpose. - **Feature Dictionary:** Defines every field in the final dataset for future reference and clarity.

Deliverables:

- Cleaned Dataset – Fully processed and standardized data ready for analysis. - Preprocessing Summary – Detailed record of all data cleaning and transformation steps. - Feature Dictionary – Clear definitions of all variables and newly created features.

Coding and Execution:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# Sample dataset
data = {
```

```
    'AgeGroup': ['5-8', '5-8', '5-8', '5-8', '9-12', '9-12', '9-12', '9-12'],
    'DeviceType': ['Phone', 'Tablet', 'TV', 'Computer', 'Phone', 'Tablet', 'TV', 'Computer'],
    'AvgScreenTime_Min': [30, 45, 95, 50, 60, 75, 140, 90]
}

df = pd.DataFrame(data)
print("Original Dataset:\n", df)

# Check missing values
print("\n--- Missing Values Before Cleaning ---")
print(df.isnull().sum())

# Handle missing values
df = df.fillna({
    'AvgScreenTime_Min': df['AvgScreenTime_Min'].median(),
    'DeviceType': 'Unknown',
    'AgeGroup': 'Unknown'
})

# Standardize text format
df['DeviceType'] = df['DeviceType'].str.strip().str.capitalize()

print("\n--- Missing Values After Cleaning ---")
print(df.isnull().sum())

# Add date and time fields
df['Date'] = pd.date_range(start='2024-09-01', periods=len(df), freq='D')
df['Date'] = pd.to_datetime(df['Date'], errors='coerce')

df['DayOfWeek'] = df['Date'].dt.day_name()
df['IsWeekend'] = np.where(df['DayOfWeek'].isin(['Saturday', 'Sunday']), 1, 0)

# Calculate device share
total_time = df['AvgScreenTime_Min'].sum()
df['DeviceShare'] = round((df['AvgScreenTime_Min'] / total_time) * 100, 2)

# Save cleaned dataset
df.to_csv("cleaned_dataset.csv", index=False)
print("\nCleaned dataset saved as 'cleaned_dataset.csv'")

# Create preprocessing summary
summary = {
    'Step': [
        'Checked for missing values',
        'Handled missing data',
        'Standardized device names',
        'Created date & weekday/weekend flags',
        'Calculated device share (%)'
    ],
    'Explanation': [
        'Used isnull() to detect blanks',
        'Filled numeric with median, text with default values',
        'Made all device names consistent',
        'Added Date, DayOfWeek, and IsWeekend fields',
        'Computed share of screen time by each device'
    ]
}
```

```
pd.DataFrame(summary).to_csv("preprocessing_summary.csv", index=False)

# Create feature dictionary
features = {
    'AgeGroup': 'Categorical age range (5-8, 9-12)',
    'DeviceType': 'Type of device used',
    'AvgScreenTime_Min': 'Average daily screen time in minutes',
    'Date': 'Date of data record',
    'DayOfWeek': 'Day extracted from Date',
    'IsWeekend': '1 for weekend, 0 for weekday',
    'DeviceShare': 'Percentage of total screen time'
}

pd.DataFrame(list(features.items()), columns=['Feature', 'Description']).to_csv("feature_diction

print("Preprocessing summary and feature dictionary saved.")

# Visualization
plt.figure(figsize=(8,5))
plt.bar(df['DeviceType'], df['AvgScreenTime_Min'], color='skyblue', edgecolor='black')
plt.title('Average Screen Time by Device Type', fontsize=14, fontweight='bold')
plt.xlabel('Device Type', fontsize=12)
plt.ylabel('Avg Screen Time (Minutes)', fontsize=12)
plt.grid(axis='y', linestyle='--', alpha=0.6)
plt.show()

plt.figure(figsize=(6,6))
plt.pie(df['DeviceShare'], labels=df['DeviceType'], autopct='%1.1f%%', startangle=90, shadow=Tru
plt.title('Device Share of Total Screen Time', fontsize=14, fontweight='bold')
plt.show()

print("\nFinal Cleaned Dataset:\n", df)
```

Results:

- The dataset is now cleaned, consistent, and enriched with derived fields. - The preprocessing summary and feature dictionary were successfully generated. - Two visualizations (bar chart and pie chart) display screen time distribution by device type.

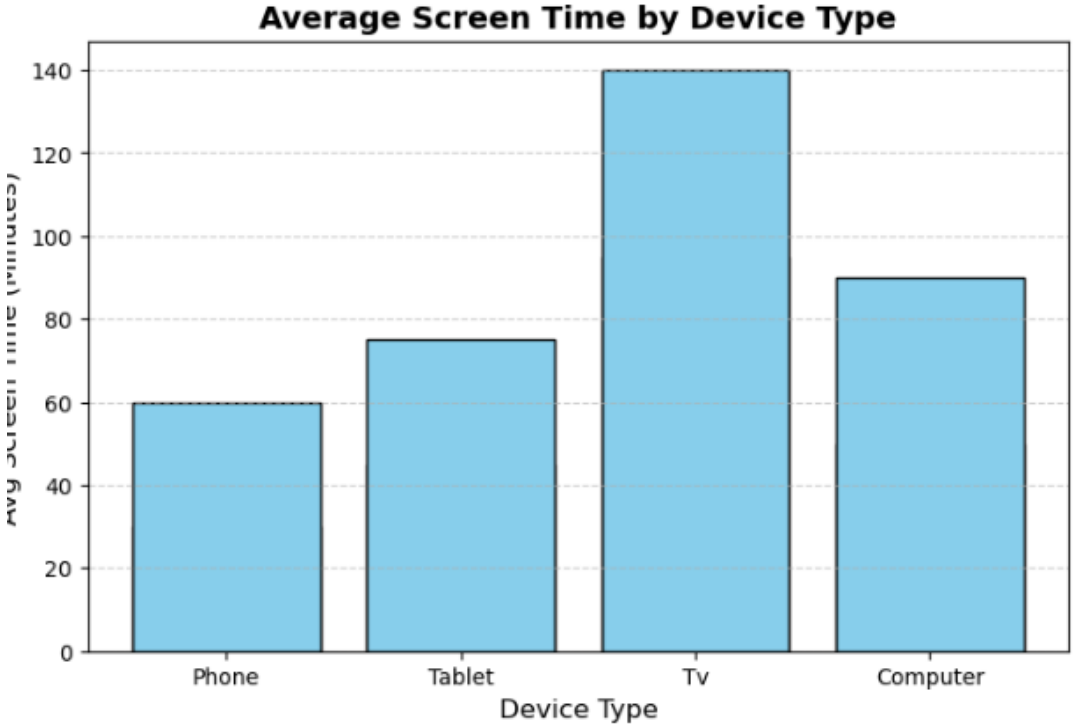
OUTPUT :

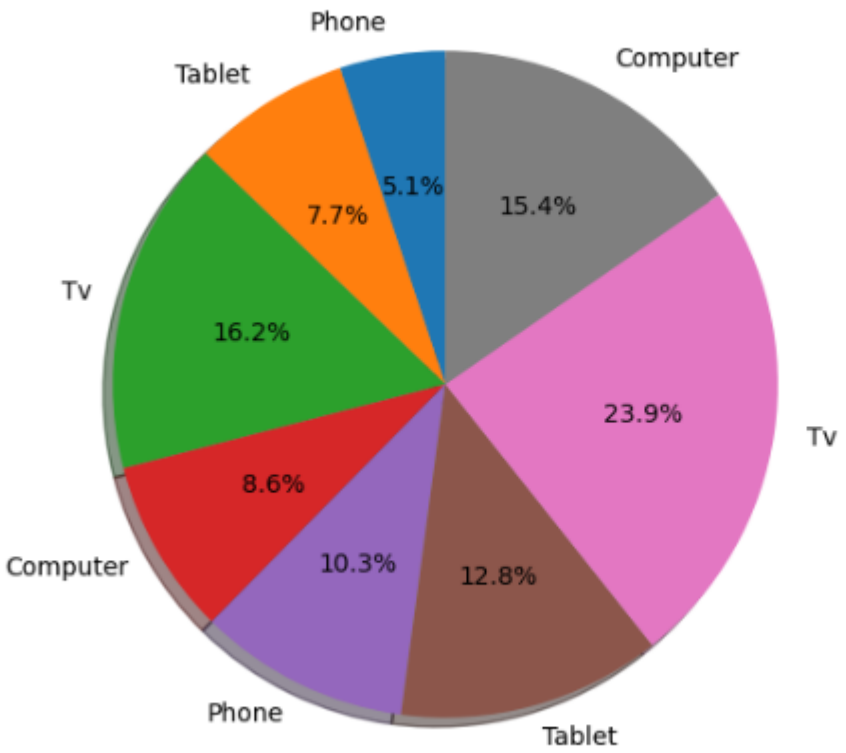
```
Original Dataset:
   AgeGroup DeviceType  AvgScreenTime_Min
0       5-8      Phone                 30
1       5-8     Tablet                 45
2       5-8         TV                 95
3       5-8   Computer                 50
4      9-12      Phone                 60
5      9-12     Tablet                 75
6      9-12         TV                140
7      9-12   Computer                 90

--- Missing Values Before Cleaning ---
AgeGroup             0
DeviceType           0
AvgScreenTime_Min    0
dtype: int64

--- Missing Values After Cleaning ---
AgeGroup             0
DeviceType           0
AvgScreenTime_Min    0
dtype: int64
```

✅ Cleaned dataset saved as 'cleaned_dataset.csv'
📝 Preprocessing summary and feature dictionary saved.





Final Cleaned Dataset:

	AgeGroup	DeviceType	AvgScreenTime_Min	Date	DayOfWeek	IsWeekend	\
0	5-8	Phone	30	2024-09-01	Sunday	1	
1	5-8	Tablet	45	2024-09-02	Monday	0	
2	5-8	Tv	95	2024-09-03	Tuesday	0	
3	5-8	Computer	50	2024-09-04	Wednesday	0	
4	9-12	Phone	60	2024-09-05	Thursday	0	
5	9-12	Tablet	75	2024-09-06	Friday	0	
6	9-12	Tv	140	2024-09-07	Saturday	1	
7	9-12	Computer	90	2024-09-08	Sunday	1	

	DeviceShare
0	5.13
1	7.69
2	16.24
3	8.55
4	10.26
5	12.82
6	23.93
7	15.38