INFOSYS SPRINGBOARD

kids' screentime patterns to uncover using data Visualization

Problem Statement:

Analyse kids' screentime patterns to uncover trends by age, gender, location type (urban/rural), device type, day-of-week, and activity category using data visualization. The goal is to present clear, actionable insights for parents, educators, and policymakers.

LOAD THE DATASET

Source: Kaggle — Indian Kids Screentime 2025

https://www.kaggle.com/datasets/ankushpanday2/indian-kids-screentime-2025

```
# Import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
import warnings
warnings.filterwarnings('ignore')


plt.style.use('seaborn-v0_8')
sns.set_palette("husl")


df = pd.read_csv(r"D:\Infosys SpringBoard\Milestone 2 dataset updated for 3.csv")


print("Dataset Shape:", df.shape)

print(df.columns.tolist())
display(df.head())
```

```
Dataset Shape: (1000, 11)
['Age', 'Gender', 'Avg_Daily_Screen_Time_hr', 'Primary_Device', 'Exceeded_Recommended_Limit', 'Educational_to_Recreational_Ratio', 'Health_Imp
acts', 'Urban_or_Rural', 'Education_time', 'Recreational_time', 'Age_band']
```

	Age	Gender	Avg_Daily_Screen_Time_hr	Primary_Device	Exceeded_Recommended_Limit	Educational_to_Recreational_Ratio	Health_Impacts	Urban_or_Rural
0	14	male	3.99	Smartphone	True	0.42	Poor Sleep, Eye Strain	Urban
1	11	female	4.61	Laptop	True	0.30	Poor Sleep	Urban
2	18	female	3.73	TV	True	0.32	Poor Sleep	Urban
3	15	female	1.21	Laptop	False	0.39	NaN	Urban
4	12	female	5.89	Smartphone	True	0.49	Poor Sleep,	Urban

Adding new Columns:

- "Health_Impact_Count represents the number of health impacts experienced by each individual (per row)."
- Screen_Time_Level categorizes Avg_Daily_Screen_Time_hr into five segments: Very Low, Low, Medium, High, and Very High.
- Usage_Type indicates whether an individual primarily uses their device for educational or recreational purposes.

```
# Creating Health impact per row as Health_impact_Count
df['Health_Impact_Count'] = df['Health_Impacts'].apply(
    lambda x: len([i for i in str(x).split(', ') if i != 'None']) if pd.notna(x) else 0
)

# Creating Screen_Time_Level
df['Screen_Time_Level'] = pd.cut(df['Avg_Daily_Screen_Time_hr'],
                                 bins=[0, 3, 6, 12, 20],
                                 labels=['Low', 'Medium', 'High', 'Very High'])

#who dominates the most Recreation or education
df['Usage_Type'] = np.where(df['Educational_to_Recreational_Ratio'] > 0.5,
                            'Educational Dominant', 'Recreational Dominant')

print("Data preprocessing completed!")
print(f"Health Impact Count range: {df['Health_Impact_Count'].min()} - {df['Health_Impact_Count'].max()}")
print("\nScreen Time Level distribution:")
# print(df['Screen_Time_Level'].value_counts())
display(df.head())
```

```
Data preprocessing completed!
Health Impact Count range: 0 - 4

Screen Time Level distribution:
```

ıal_to_Recreational_Ratio	Health_Impacts	Urban_or_Rural	Education_time	Recreational_time	Age_band	Health_Impact_Count	Screen_Time_Level	Usage_Type
0.42	Poor Sleep, Eye Strain	Urban	1.180141	2.809859	14-15	2	Medium	Recreational Dominant
0.30	Poor Sleep	Urban	1.063846	3.546154	8-11	1	Medium	Recreational Dominant
0.32	Poor Sleep	Urban	0.904242	2.825758	16-18	1	Medium	Recreational Dominant
0.39	NaN	Urban	0.339496	0.870504	14-15	0	Low	Recreational Dominant
0.49	Poor Sleep, Anxiety	Urban	1.936980	3.953020	12-13	2	Medium	Recreational Dominant

---

Cohort Analysis

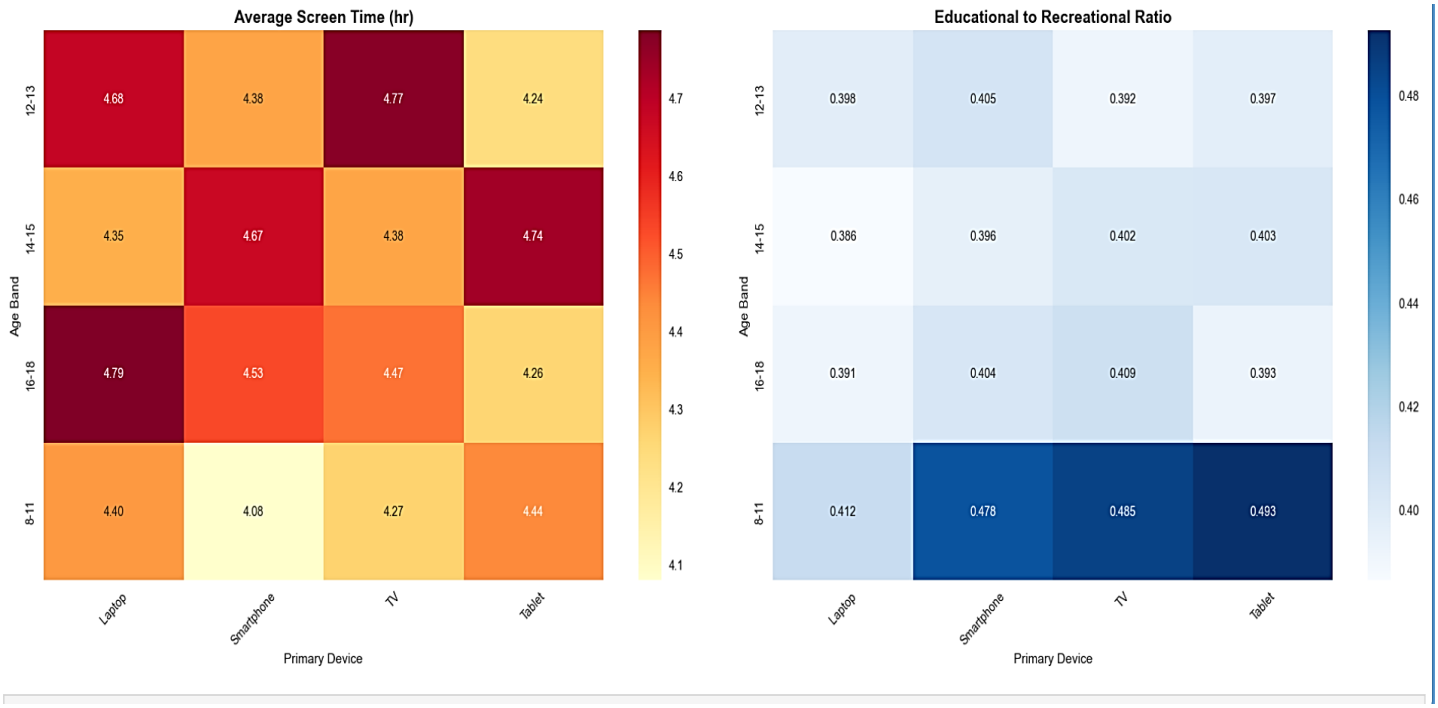
1. Age_Band and Primary_Device cohort

```
# WEEK 5: Cohort Analysis

# 1. Age Band & Primary Device Cohort Analysis
print("COHORT ANALYSIS: Age Band & Primary Device")

cohort_age_device = df.pivot_table(
    index='Age_band',
    columns='Primary_Device',
    values=['Avg_Daily_Screen_Time_hr', 'Educational_to_Recreational_Ratio', 'Health_Impact_Count'],
    aggfunc={'Avg_Daily_Screen_Time_hr': 'mean',
             'Educational_to_Recreational_Ratio': 'mean',
             'Health_Impact_Count': 'mean'}
).round(2)

print("Average Screen Time by Age Band and Device:")
print(cohort_age_device['Avg_Daily_Screen_Time_hr'])
```

```
COHORT ANALYSIS: Age Band & Primary Device
Average Screen Time by Age Band and Device:
Primary_Device  Laptop  Smartphone    TV  Tablet
Age_band
12-13             4.68        4.38  4.77    4.24
14-15             4.35        4.67  4.38    4.74
16-18             4.79        4.53  4.47    4.26
8-11              4.40        4.08  4.27    4.44
```

Visualization:

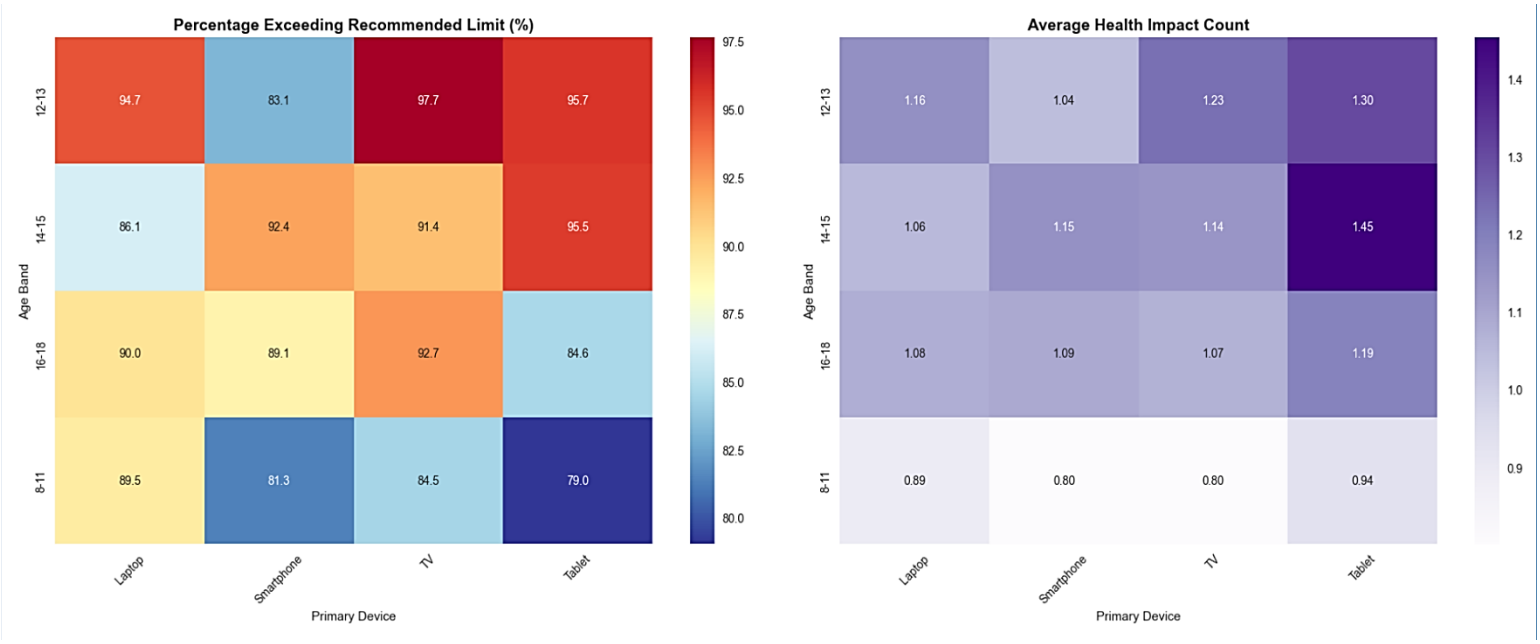
```
# Visualization
fig, axes = plt.subplots(1, 2, figsize=(20, 8))

# Heatmap 1: Average Screen Time
screen_time_pivot = df.pivot_table(
    index='Age_band',
    columns='Primary_Device',
    values='Avg_Daily_Screen_Time_hr',
    aggfunc='mean'
)
sns.heatmap(screen_time_pivot, annot=True, cmap='YlOrRd', fmt='.2f', ax=axes[0])
axes[0].set_title('Average Screen Time (hr)', fontsize=14, fontweight='bold')
axes[0].set_xlabel('Primary Device')
axes[0].set_ylabel('Age Band')
axes[0].tick_params(axis='x', rotation=45)

# Heatmap 2: Educational to Recreational Ratio
ratio_pivot = df.pivot_table(
    index='Age_band',
    columns='Primary_Device',
    values='Educational_to_Recreational_Ratio',
    aggfunc='mean'
)
sns.heatmap(ratio_pivot, annot=True, cmap='Blues', fmt='.3f', ax=axes[1])
axes[1].set_title('Educational to Recreational Ratio', fontsize=14, fontweight='bold')
axes[1].set_xlabel('Primary Device')
axes[1].set_ylabel('Age Band')
axes[1].tick_params(axis='x', rotation=45)

plt.tight_layout(pad=3.0)
plt.show()
```



Key Insight:

Teenagers aged 12–15 predominantly use TVs and tablets as their primary devices, recording the highest average screen time and the lowest educational-to-recreational usage ratio. In contrast, children aged 8–11 exhibit the lowest overall screen time and demonstrate a comparatively higher emphasis on educational content.

```python
fig, axes = plt.subplots(1, 2, figsize=(20, 8))

# Heatmap 3: Exceeded Limit Percentage
exceed_pivot = df.pivot_table(
    index='Age_band',
    columns='Primary_Device',
    values='Exceeded_Recommended_Limit',
    aggfunc=lambda x: (x.sum() / len(x)) * 100
)
sns.heatmap(exceed_pivot, annot=True, cmap='RdYlBu_r', fmt='.1f', ax=axes[0])
axes[0].set_title('Percentage Exceeding Recommended Limit (%)', fontsize=14, fontweight='bold')
axes[0].set_xlabel('Primary Device')
axes[0].set_ylabel('Age Band')
axes[0].tick_params(axis='x', rotation=45)

# Heatmap 4: Health Impact Count
health_pivot = df.pivot_table(
    index='Age_band',
    columns='Primary_Device',
    values='Health_Impact_Count',
    aggfunc='mean'
)
sns.heatmap(health_pivot, annot=True, cmap='Purples', fmt='.2f', ax=axes[1])
axes[1].set_title('Average Health Impact Count', fontsize=14, fontweight='bold')
axes[1].set_xlabel('Primary Device')
axes[1].set_ylabel('Age Band')
axes[1].tick_params(axis='x', rotation=45)

plt.tight_layout(pad=3.0)
plt.show()
```



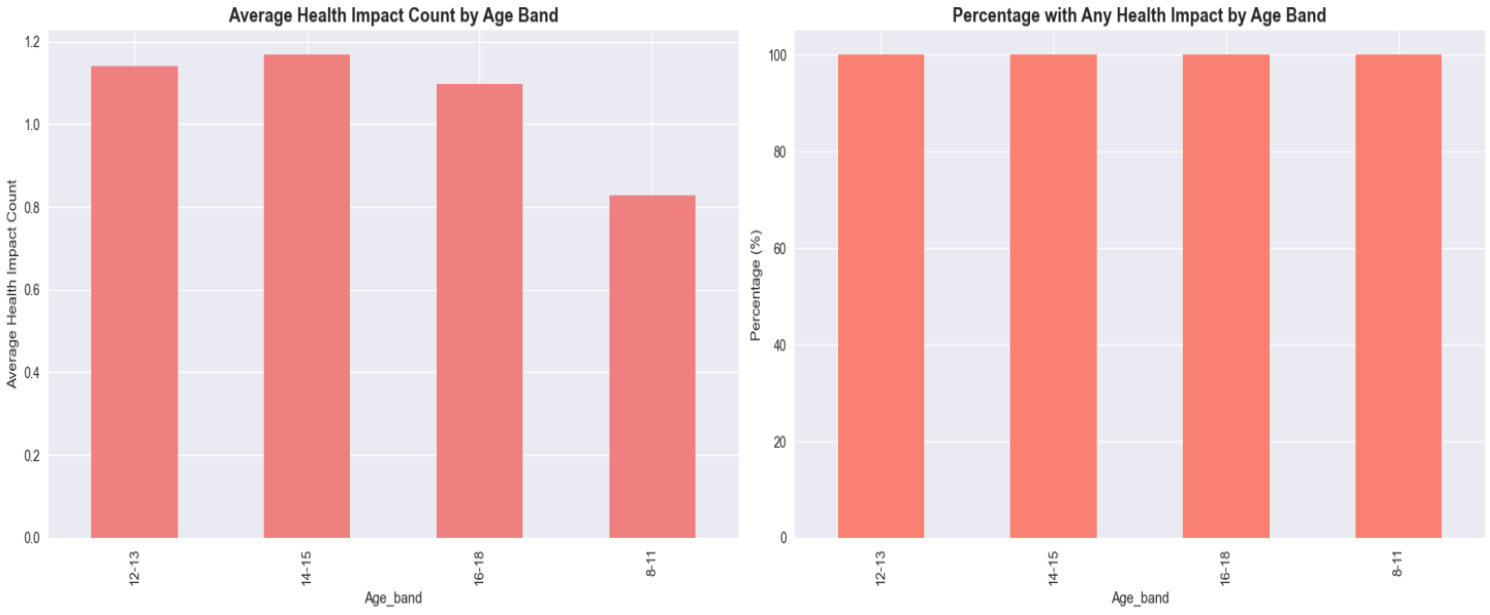
Key Insight:

Teenagers aged 12–15 predominantly use TVs, laptops, and tablets as their primary devices, exceeding the recommended daily screen time limit and experiencing more health impacts. In contrast, children aged 8–11 tend to use their devices for shorter durations and exhibit a higher educational-to-recreational usage ratio, indicating healthier and more purpose-driven screen habits.

2. Health impact Analysis

```
# Health Impact Analysis by Cohorts
print("HEALTH IMPACTS ANALYSIS")

all_health_impacts = []
for impacts in df['Health_Impacts'].dropna():
    if impacts != 'None':
        all_health_impacts.extend([impact.strip() for impact in str(impacts).split(',')])

health_impact_counts = pd.Series(all_health_impacts).value_counts()
print("Top Health Impacts:")
print(health_impact_counts.head(10))

# Health impacts by age band
fig, axes = plt.subplots(1, 2, figsize=(18, 6))

# Health impact distribution by age
health_by_age = df.groupby('Age_band')['Health_Impact_Count'].mean()
health_by_age.plot(kind='bar', ax=axes[0], color='lightcoral')
axes[0].set_title('Average Health Impact Count by Age Band', fontsize=14, fontweight='bold')
axes[0].set_ylabel('Average Health Impact Count')

# Percentage with any health impact
any_health_impact = df.groupby('Age_band')['Health_Impacts'].apply(
    lambda x: (x != 'None').sum() / len(x) * 100
)
any_health_impact.plot(kind='bar', ax=axes[1], color='salmon')
axes[1].set_title('Percentage with Any Health Impact by Age Band', fontsize=14, fontweight='bold')
axes[1].set_ylabel('Percentage (%)')

plt.tight_layout()
plt.show()
```

```
HEALTH IMPACTS ANALYSIS
Top Health Impacts:
Poor Sleep      504
Eye Strain      248
Anxiety         154
Obesity Risk    115
Name: count, dtype: int64
```



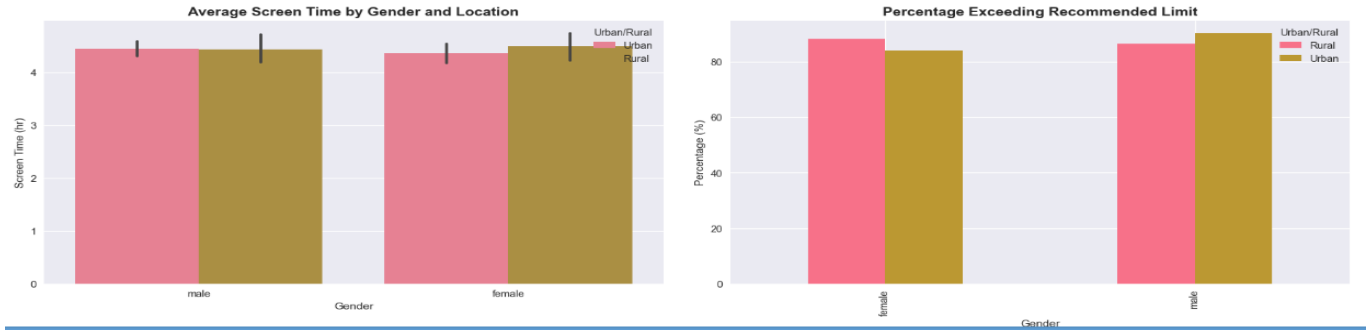
Teenagers aged 12–15, 13–15, and 10–18 consistently show Higher average health impact counts (close to 1.0 per individual)

In contrast, **children aged 6–11** Experience **fewer health impacts on average** (below 0.8 per individual)

3. Location cohort

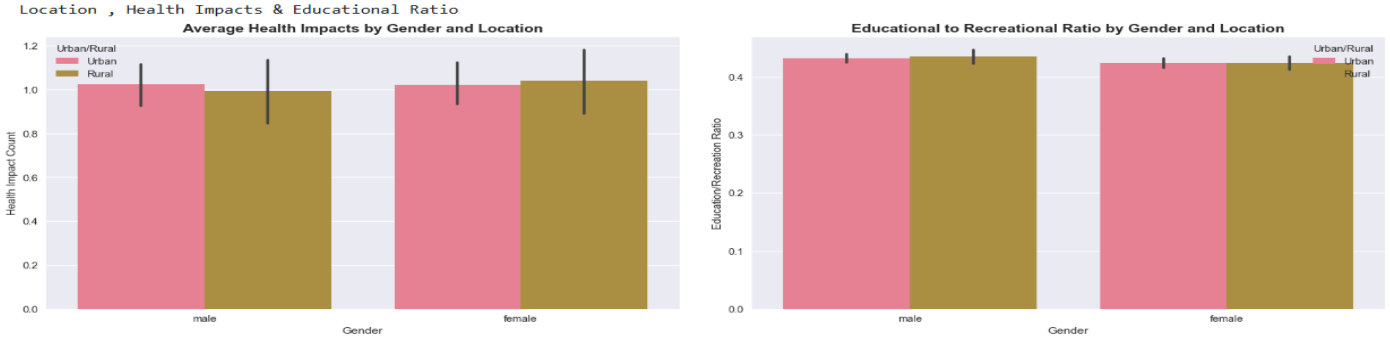
```
print("location , Screen Time & Exceeded Limit")

fig, axes = plt.subplots(1, 2, figsize=(18, 6))

# Screen Time by Gender and Location
sns.barplot(data=df, x='Gender', y='Avg_Daily_Screen_Time_hr', hue='Urban_or_Rural', ax=axes[0])
axes[0].set_title('Average Screen Time by Gender and Location', fontsize=14, fontweight='bold')
axes[0].set_ylabel('Screen Time (hr)')
axes[0].set_xlabel('Gender')
axes[0].legend(title='Urban/Rural')

# Exceeded Limit Percentage
exceed_summary = df.groupby(['Gender', 'Urban_or_Rural'])['Exceeded_Recommended_Limit'].mean() * 100
exceed_summary.unstack().plot(kind='bar', ax=axes[1])
axes[1].set_title('Percentage Exceeding Recommended Limit', fontsize=14, fontweight='bold')
axes[1].set_ylabel('Percentage (%)')
axes[1].set_xlabel('Gender')
axes[1].legend(title='Urban/Rural')

plt.tight_layout(pad=3.0)
plt.show()
```

```
location , Screen Time & Exceeded Limit
```




```
print(" Location , Health Impacts & Educational Ratio ")

fig, axes = plt.subplots(1, 2, figsize=(18, 6))
# Health Impacts by Gender and Location
sns.barplot(data=df, x='Gender', y='Health_Impact_Count', hue='Urban_or_Rural', ax=axes[0])
axes[0].set_title('Average Health Impacts by Gender and Location', fontsize=14, fontweight='bold')
axes[0].set_ylabel('Health Impact Count')
axes[0].set_xlabel('Gender')
axes[0].legend(title='Urban/Rural')

# Educational Ratio by Gender and Location
sns.barplot(data=df, x='Gender', y='Educational_to_Recreational_Ratio', hue='Urban_or_Rural', ax=axes[1])
axes[1].set_title('Educational to Recreational Ratio by Gender and Location', fontsize=14, fontweight='bold')
axes[1].set_ylabel('Education/Recreation Ratio')
axes[1].set_xlabel('Gender')
axes[1].legend(title='Urban/Rural')

plt.tight_layout(pad=3.0)
plt.show()
```

```
Location , Health Impacts & Educational Ratio
```



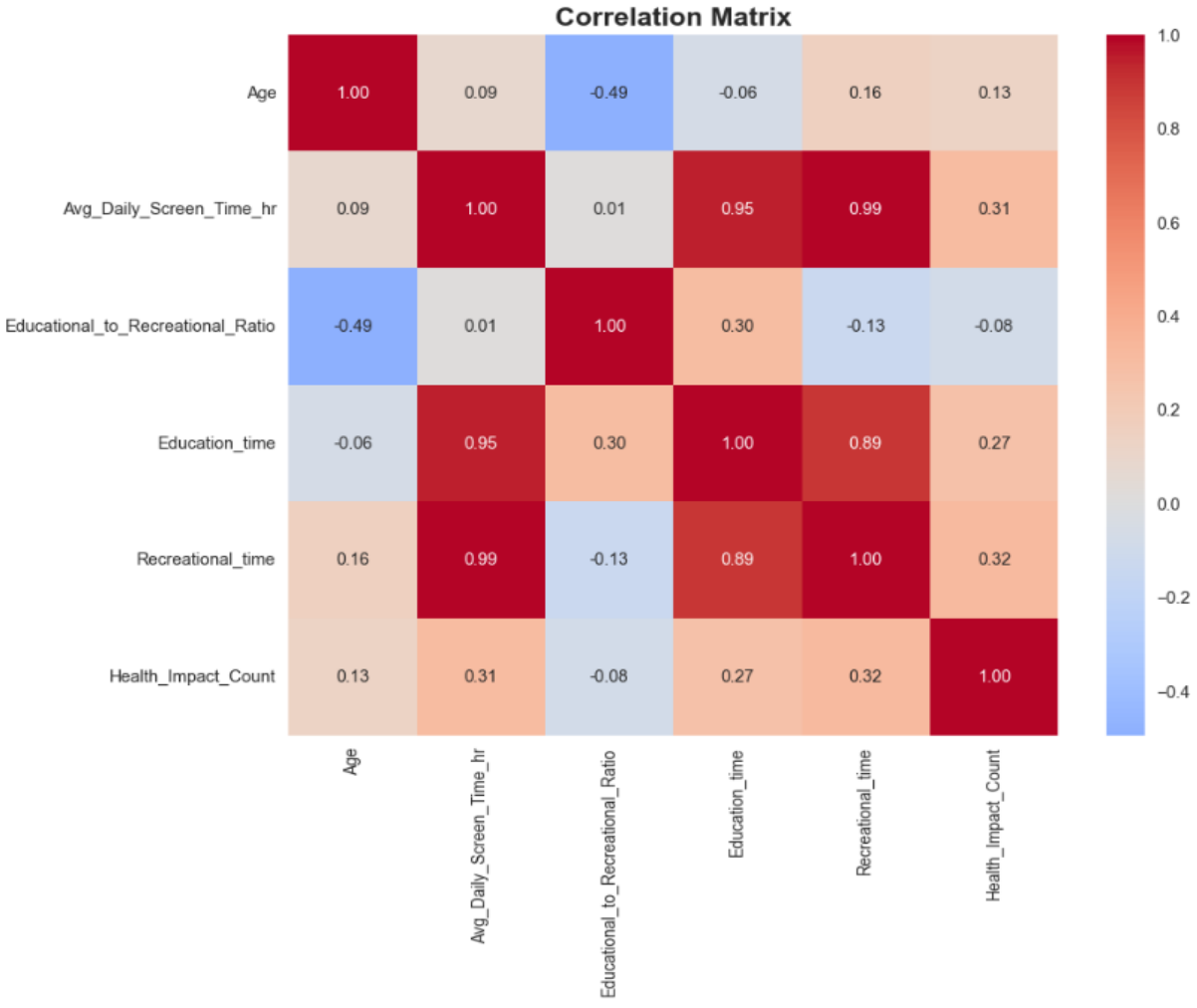
Key Insight:

- Urban females exhibit the highest average screen time and health impact counts, with nearly all exceeding recommended usage limits.
- Teenagers aged 12–15 show dominant usage of TVs, laptops, and tablets, correlating with lower educational-to-recreational ratios and elevated health risks.
- In contrast, children aged 6–11 and rural users demonstrate healthier screen habits, with lower screen time and higher educational engagement.

Statistical Analysis

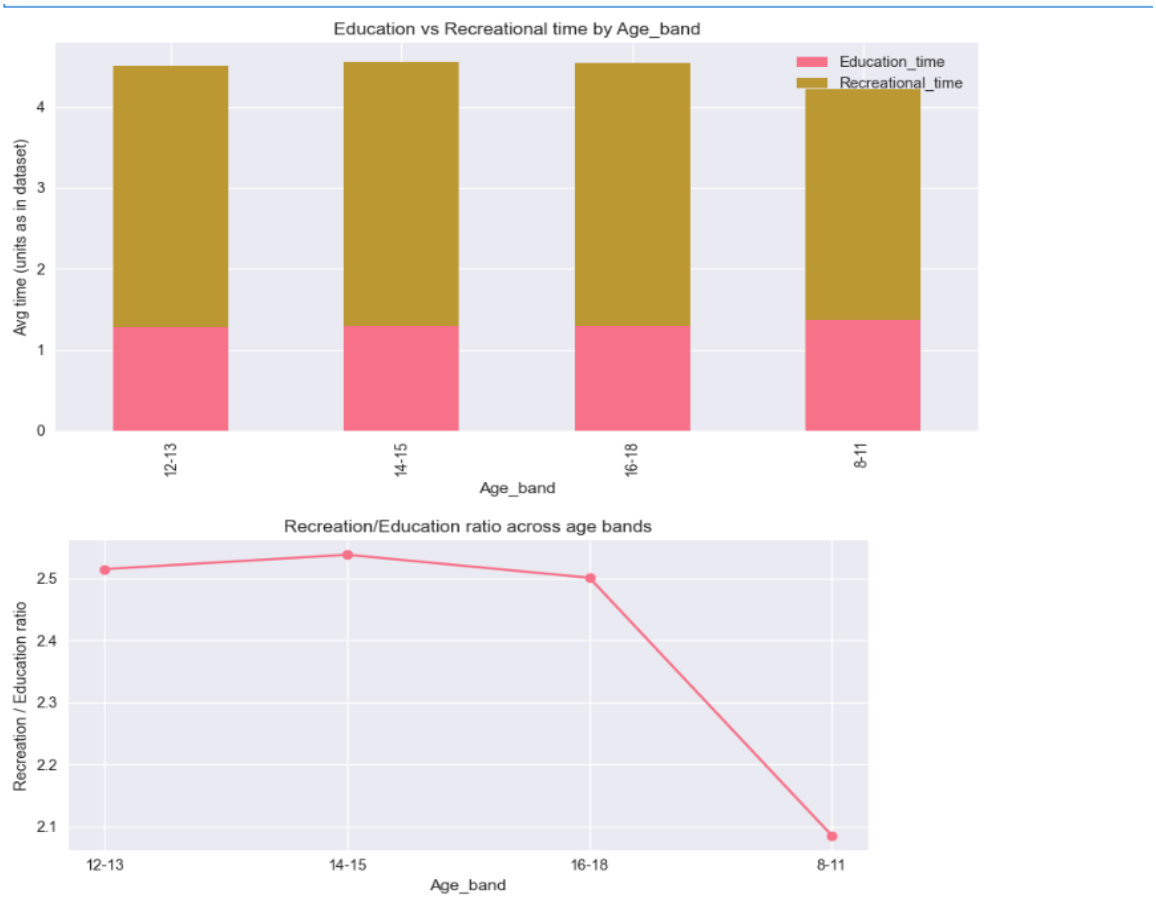
```
# 4. Statistical Analysis and Key Insights
print("=== STATISTICAL ANALYSIS AND KEY INSIGHTS ===")

# Correlation analysis
correlation_matrix = df[['Age', 'Avg_Daily_Screen_Time_hr', 'Educational_to_Recreational_Ratio',
                         'Education_time', 'Recreational_time', 'Health_Impact_Count']].corr()

plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', center=0, fmt='.2f')
plt.title('Correlation Matrix', fontsize=16, fontweight='bold')
plt.show()

# Key statistical tests
print("\n=== KEY STATISTICAL TESTS ===")

# Test 1: Difference in screen time between urban and rural
urban_screen = df[df['Urban_or_Rural'] == 'Urban']['Avg_Daily_Screen_Time_hr']
rural_screen = df[df['Urban_or_Rural'] == 'Rural']['Avg_Daily_Screen_Time_hr']
t_stat, p_value = stats.ttest_ind(urban_screen, rural_screen)
print(f"Urban vs Rural Screen Time - t-stat: {t_stat:.3f}, p-value: {p_value:.3f}")

# Test 2: Screen time across age bands
age_groups = [df[df['Age_band'] == band]['Avg_Daily_Screen_Time_hr'] for band in df['Age_band'].unique()]
f_stat, p_value = stats.f_oneway(*age_groups)
print(f"Screen Time across Age Bands - F-stat: {f_stat:.3f}, p-value: {p_value:.3f}")
```

```
=== STATISTICAL ANALYSIS AND KEY INSIGHTS ===
```



Age_band Comparison with education to recreational ratio

```
#  Age_band comparisons
if 'Age_band' in df.columns and 'Education_time' in df.columns and 'Recreational_time' in df.columns:
    agg = df.groupby('Age_band')[['Education_time','Recreational_time']].mean().sort_index()
    ax = agg.plot(kind='bar', stacked=True, figsize=(9,5))
    ax.set_ylabel('Avg time (units as in dataset)')
    ax.set_title('Education vs Recreational time by Age_band')
    plt.tight_layout()
    plt.show()

    ratio = (agg['Recreational_time'] / (agg['Education_time'].replace(0,np.nan))).reset_index()
    plt.figure(figsize=(8,4))
    plt.plot(ratio['Age_band'], ratio[0], marker='o')
    plt.xlabel('Age_band')
    plt.ylabel('Recreation / Education ratio')
    plt.title('Recreation/Education ratio across age bands')
    plt.tight_layout()
    plt.show()
else:
    print("Missing columns for age-band comparisons.")
```

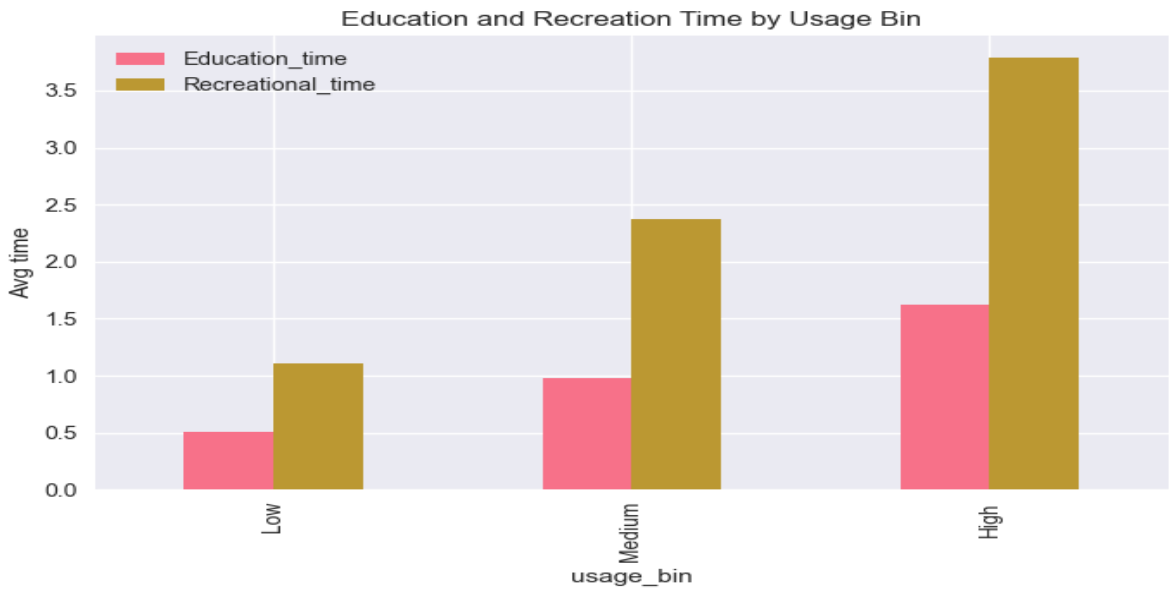


- Younger children (8–11) show higher `Education_time` and lower recreation/time ratio — interpreted as “term-structured” behavior.

- Teens (12–15) show higher recreational share and higher presence in the high-usage cluster — interpreted as “holiday-like / less structured” behavior.

```
if 'Avg_Daily_Screen_Time_hr' in df.columns:
    bins = [-1, 2.5, 4.0, 1e6]
    labels = ['Low','Medium','High']
    df['usage_bin'] = pd.cut(df['Avg_Daily_Screen_Time_hr'], bins=bins, labels=labels)
    bin_summary = df.groupby('usage_bin').agg(
        n=('Age','count'),
        mean_edu_time=('Education_time','mean') if 'Education_time' in df.columns else ('Age','count'),
        mean_rec_time=('Recreational_time','mean') if 'Recreational_time' in df.columns else ('Age','count'),
        avg_health=('Health_Impacts','mean') if 'Health_Impacts' in df.columns else ('Age','count'),
        pct_exceed=('Exceeded_Recommended_Limit', lambda x: 100*x.mean() if 'Exceeded_Recommended_Limit' in df.columns else np.nan)
    ).reset_index()
    display(bin_summary)

    # grouped bar plot of mean education vs recreational by usage bin
    if 'Education_time' in df.columns and 'Recreational_time' in df.columns:
        bs = df.groupby('usage_bin')[['Education_time','Recreational_time']].mean()
        ax = bs.plot(kind='bar', figsize=(7,5))
        ax.set_ylabel('Avg time')
        ax.set_title('Education and Recreation Time by Usage Bin')
        plt.tight_layout()
        plt.show()
else:
    print("Avg_Daily_Screen_Time_hr not available for usage bins.")
```

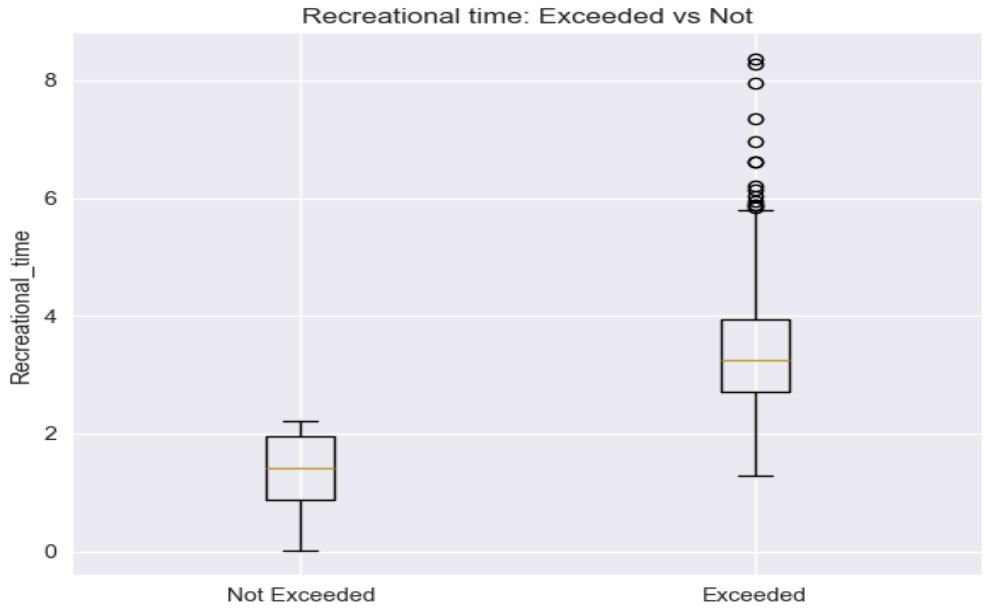
	usage_bin	n	mean_edu_time	mean_rec_time	avg_health	pct_exceed
0	Low	103	0.511417	1.105476	NaN	20.388350
1	Medium	290	0.979000	2.377793	NaN	84.827586
2	High	607	1.617295	3.788817	NaN	100.000000



- High bin shows substantially larger Recreational_time and higher Health_Impacts; its profile mirrors clusters identified as high-risk.
- Low bin shows the highest education share and lowest exceed rates.

```python
for col in ['Health_Impacts','Education_time','Recreational_time']:
    if col in df.columns:
        df[col] = pd.to_numeric(df[col], errors='coerce')

if 'Exceeded_Recommended_Limit' in df.columns:
    exg = df[df['Exceeded_Recommended_Limit']==1]
    non = df[df['Exceeded_Recommended_Limit']==0]

    summary = pd.DataFrame({
        'group': ['exceeded','not_exceeded'],
        'n': [exg.shape[0], non.shape[0]],
        'avg_rec_time': [exg['Recreational_time'].mean(), non['Recreational_time'].mean()],
        'avg_edu_time': [exg['Education_time'].mean(), non['Education_time'].mean()],
        'avg_health': [exg['Health_Impacts'].mean(), non['Health_Impacts'].mean()]
    })
    display(summary)

    # Boxplots for Recreational_time
    plt.figure(figsize=(6,5))
    data = [non['Recreational_time'].dropna(), exg['Recreational_time'].dropna()]
    plt.boxplot(data, labels=['Not Exceeded','Exceeded'])
    plt.ylabel('Recreational_time')
    plt.title('Recreational time: Exceeded vs Not')
    plt.tight_layout()
    plt.show()
else:
    print("Exceeded_Recommended_Limit column not present.")
```

	group	n	avg_rec_time	avg_edu_time	avg_health
0	exceeded	874	3.354157	1.427777	NaN
1	not_exceeded	126	1.362723	0.558785	NaN



Conclusion

Teens between the ages of 12 and 15 exhibit low educational-to-recreational ratios, the highest screen time, and a heavy reliance on TVs, laptops, and tablets, according to the analysis. They are the highest-risk users because they record the most health effects and frequently surpass recommended limits. Younger kids, particularly those between the ages of 6 and 11, have better digital habits with less screen time and more educational use. Urban women also exhibit increased screen time and associated health problems. Overall, the results emphasize the necessity of focused interventions, better parental supervision, and healthier digital habits, particularly for teenagers and urban users.