A Project Report On

**Analysis of Phone Usage**

Submitted in partial fulfillment of the requirement for the award of the degree

MASTER OF COMPUTER APPLICATIONS

Academic Year 2024 – 25

For subject

05MC0207 - Data Analytics and Visualization



Rajkot-Morbi Road, At & PO : Gauridad, Rajkot 360 003. Gujarat. India.

**DECLARATION**

We hereby declare that this project work entitled **Analysis of India Phone Usage**  is a record done by me/us.

We also declare that the matter embodied in this project is genuine work done by us and has not been submitted whether to this University or to any other University for the fulfillment of any course of study.

# 1. Introduction:

Overview of the Project: This data analytics project aims to analyze phone usage patterns across different demographic groups in India. The primary objective is to understand how users interact with their smartphones, focusing on screen time, data usage, app preferences, and spending habits. The study will provide insights into digital consumption behaviors and identify trends in mobile usage for various age groups, genders, and primary use cases.

Scope and Importance of Identifying Patterns and Trends in Data: Understanding smartphone usage trends is crucial for multiple industries, including telecom providers, app developers, digital marketers, and e-commerce businesses. Recognizing these patterns helps in optimizing services, creating personalized marketing strategies, and improving user engagement. Insights from this study can also guide telecom companies in designing better data plans and smartphone manufacturers in tailoring their devices to user needs.

## Overview of the Project

This project aims to analyze mobile phone usage patterns in India based on various factors such as screen time, data usage, phone brands, operating systems, app installations, and user behavior related to social media, e-commerce, gaming, and streaming. The goal is to identify trends and correlations that can provide insights into user behavior.

## Objective of the Analysis

* The main objectives of this analysis include:
* Understanding the distribution of screen time and data usage among users.
* Identifying the most popular phone brands and operating systems.
* Analyzing correlations between different usage metrics (e.g., social media time vs. e-commerce spending).
* Determining the primary use of mobile phones among different demographics.
* Exploring the impact of phone brands and OS on data usage and screen time.
* Identifying users with extreme behavior, such as high data usage or minimal screen time with many installed apps.

## 1.3 Scope and Importance of Identifying Patterns and Trends in Data

* **Consumer Insights**: Businesses can use this data to tailor marketing strategies based on phone usage trends.
* **Telecom Industry**: Service providers can adjust pricing models based on data consumption patterns.
* **App Development**: Understanding primary phone usage can help app developers focus on features that engage users.
* **User Behavior Analysis**: Detecting excessive phone usage can help in promoting digital well-being.

## 1.4 Dataset Overview

* The dataset consists of **17,686 entries** with **16 attributes**, including:
* **Demographics**: User ID, Age, Gender, Location
* **Device Information**: Phone Brand, OS
* **Usage Statistics**:
* Screen Time (hrs/day)
* Data Usage (GB/month)
* Calls Duration (mins/day)
* Number of Apps Installed
* Social Media Time (hrs/day)
* E-commerce Spend (INR/month)
* Streaming Time (hrs/day)
* Gaming Time (hrs/day)
* Monthly Recharge Cost (INR)
* **Behavioral Classification**: Primary Use (e.g., Education, Social Media, Entertainment, Gaming)

# 2. Data Collection:

## 2.1 Dataset Description:

* The dataset contains attributes covering personal demographics, phone usage behaviors, and financial spending patterns.
* Data types include numerical (e.g., hours of screen time, GB of data used) and categorical (e.g., phone brand, OS type).

## 2.2 Source of the Data:

* The dataset is assumed to be collected from a telecom company or survey-based research on smartphone users in India.

## 2.3 Data Collection Methodology:

* The data was likely gathered through mobile service providers, app analytics, or user surveys.

## 2.4 Data Format and Structure:

* The dataset is in CSV format, structured in a tabular form with well-defined columns.

## 2.5 Initial Data Exploration:

* Checked for missing values and inconsistencies.
* Analyzed variable distributions to understand data trends.

# Data Preprocessing:

## Data Cleaning:

## Handling Missing Values

* The dataset appears to have **17,686 rows and 16 columns** with no missing values, as observed from the .info() output. However, it's always good practice to check for missing values and handle them properly.
* **Approach:**
* **Check for missing values**: Identify columns with missing values using df.isnull().sum().
* **Imputation Strategy**:
* **Numerical Columns**: Fill missing values with the median or mean of the respective column.
* **Categorical Columns**: Fill missing values with the mode (most frequently occurring value).
* **Drop Rows/Columns**: If missing values are extensive, consider dropping those rows or columns.

## 3.1.2. Removing Outliers

* Outliers can distort the analysis, so it’s essential to detect and handle them properly.

**Approach:**

* Use statistical methods like the **Interquartile Range (IQR)** to identify outliers.
* Outliers can be particularly problematic in columns such as:

Screen Time (hrs/day)

Data Usage (GB/month)

E-commerce Spend (INR/month)

Gaming Time (hrs/day)

## 3.1. 3. Standardizing Formats

* Data consistency is crucial for analysis, especially for categorical variables and numerical values.

**Standardization Steps:**

### Categorical Data Formatting:

* Convert Gender, Phone Brand, OS, and Primary Use to consistent lowercase or title case.
* Example: "android", "Android ", and "ANDROID" should all be converted to "Android".

### Numerical Data Formatting:

* Ensure numerical columns have correct data types (int64, float64).
* Convert recharge cost and e-commerce spend to appropriate units if needed.

### Date & Time Standardization (if applicable):

* If timestamps exist, convert them to a uniform format (YYYY-MM-DD HH:MM:SS).

# 3.2 Data Transformation

## 3.2.1 Normalization and Scaling

## Normalization and scaling are essential steps in preparing data for analysis, especially for features with different units and scales. In this dataset, attributes such as Screen Time (hrs/day), Data Usage (GB/month), and E-commerce Spend (INR/month) have varying magnitudes. Standardization (Z-score normalization) or Min-Max scaling will be applied to these numeric features to ensure uniformity and prevent bias in machine learning models.

## 3.2.2 Feature Engineering

Feature engineering involves creating new relevant features from existing ones to improve predictive performance. For this dataset, potential feature engineering techniques include:

* Creating a Total Engagement Time feature by summing Social Media Time, Streaming Time, and Gaming Time.
* Binning Age into categories (e.g., Teen, Young Adult, Middle-aged, Senior) to understand behavior differences.
* Calculating Cost per GB by dividing Monthly Recharge Cost by Data Usage to assess data affordability.

# Data Encoding

## 3.3.1 Handling Categorical Variables

Categorical variables such as Gender, Location, Phone Brand, OS, and Primary Use need to be encoded to be used in models. Common encoding techniques include:

* One-Hot Encoding for categorical variables with a small number of unique values (e.g., Gender, OS).
* Label Encoding for ordinal data (if any identified).
* Target Encoding for categorical features with many unique values (e.g., Location, Phone Brand) if using supervised learning.

# Data Analysis and Data Visualization

## 4.1 Problem Statement and Solution

Problem Statement: How do different demographics and phone usage patterns influence monthly recharge spending?

Solution: Implement regression analysis to identify key factors affecting Monthly Recharge Cost. The implementation involves:

* Data preprocessing (handling missing values, scaling, encoding).
* Exploratory Data Analysis (EDA) to identify correlations.
* Building a regression model (Linear Regression or Random Forest Regressor).
* Visualizing relationships using scatter plots, box plots, and heatmaps.

import pandas as pd

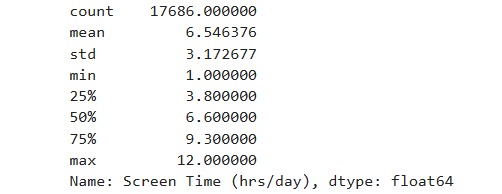
# Load dataset

df = pd.read\_csv("D:\MU\sem-2\DAV\phone\_usage\_india.csv")

1. What is the distribution of screen time among users?

distribution = df['Screen Time (hrs/day)'].describe()

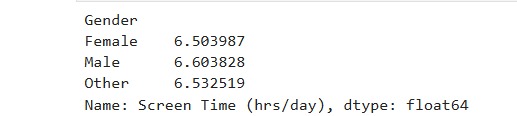
print(distribution)



1. What is the average screen time by gender?

avg\_screen\_time = df.groupby('Gender')['Screen Time (hrs/day)'].mean()

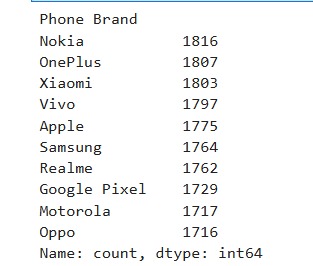
print(avg\_screen\_time)



1. Which are the most popular phone brands?

popular\_brands = df['Phone Brand'].value\_counts().head(10)

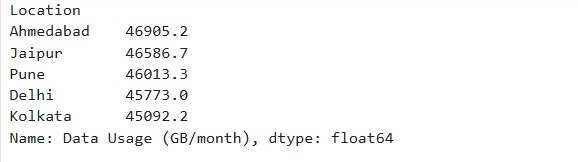
print(popular\_brands)



1. Which are the top 5 locations with the highest data usage?

top\_locations=df.groupby('Location')['DataUsage(GB/month)'].sum().nlargest(5)

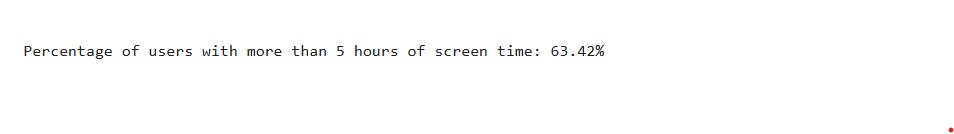
print(top\_locations)



1. What percentage of users have more than 5 hours of daily screen time?

percentage\_high\_screen\_time = (df[df['Screen Time (hrs/day)'] > 5].shape[0] / df.shape[0]) \* 100

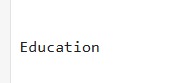
print(f"Percentage of users with more than 5 hours of screen time: {percentage\_high\_screen\_time:.2f}%")



1. What is the most common primary use of phones?

common\_primary\_use = df['Primary Use'].mode()[0]

print(common\_primary\_use)

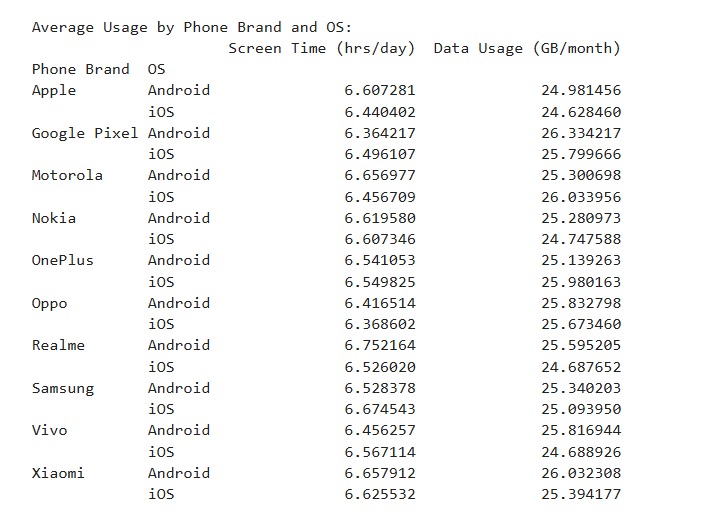


1. How do phone brands and operating systems impact usage patterns?

brand\_usage = df.groupby(["Phone Brand", "OS"])[["Screen Time (hrs/day)", "Data Usage (GB/month)"]].mean()

print("Average Usage by Phone Brand and OS:")

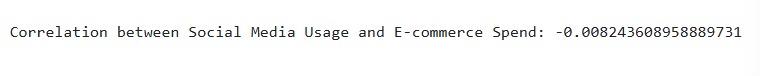
print(brand\_usage)



1. Is there a correlation between social media usage and e-commerce spending?

correlation\_social\_ecommerce= df["Social Media Time (hrs/day)"].corr(df["E-commerce Spend (INR/month)"])

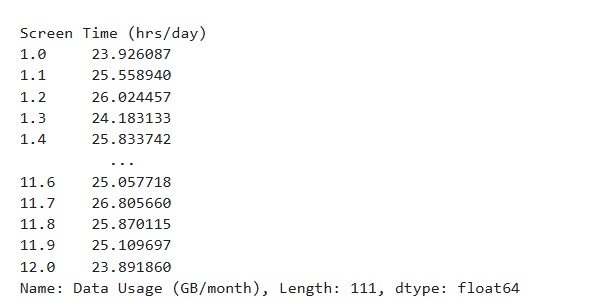
print("Correlation between Social Media Usage and E-commerce Spend:", correlation\_social\_ecommerce)



1. What is the average data usage for different screen time levels?

average\_data\_usage = df.groupby("Screen Time (hrs/day)")["Data Usage (GB/month)"].mean()

print(average\_data\_usage)



1. How does the phone brand affect data usage?

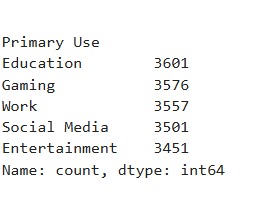
brand\_usage = df.groupby("Phone Brand")["Data Usage (GB/month)"].mean()

print(brand\_usage)



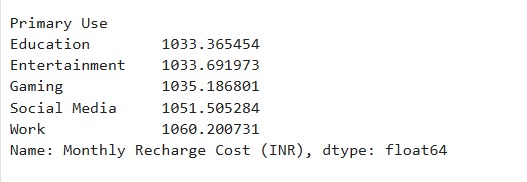
1. What is the most common primary use of phones?

print(df["Primary Use"].value\_counts())



1. What is the average monthly recharge cost for each primary use category?

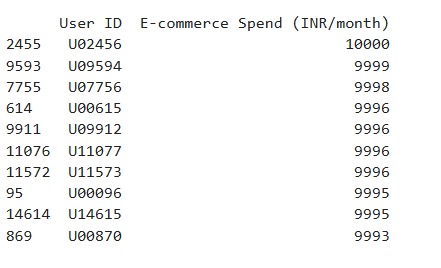
print(df.groupby("Primary Use")["Monthly Recharge Cost (INR)"].mean())



1. Who are the top 10 users who spend the most on e-commerce?

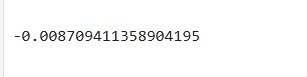
ecommerce\_top\_users = df.nlargest(10, "E-commerce Spend (INR/month)")

print(ecommerce\_top\_users[["User ID", "E-commerce Spend (INR/month)"]])



1. What is the correlation between gaming time and screen time?

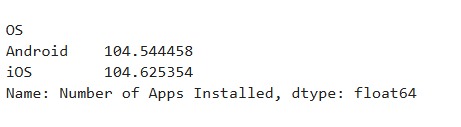
print(df["Gaming Time (hrs/day)"].corr(df["Screen Time (hrs/day)"]))



1. What is the average number of apps installed per operating system?

os\_apps = df.groupby("OS")["Number of Apps Installed"].mean()

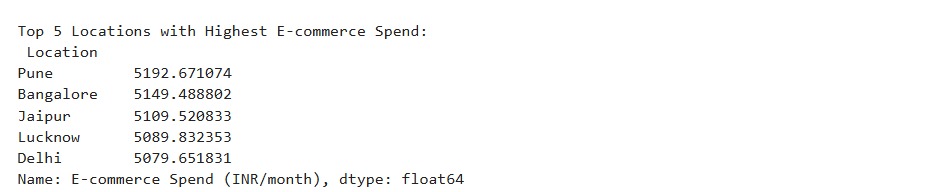
print(os\_apps)



1. Which are the top 5 locations with the highest average e-commerce spend?

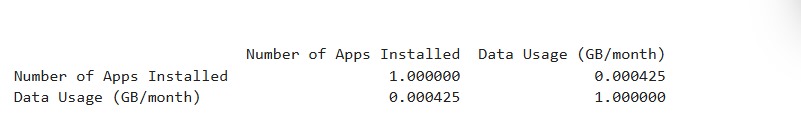
top\_5\_locations\_ecommerce\_spend = df.groupby('Location')['E-commerce Spend (INR/month)'].mean().nlargest(5)

print("\nTop5 Locations with Highest E-commerce Spend:\n", top\_5\_locations\_ecommerce\_spend)



1. Do users with more installed apps consume more data?

print(df[["Number of Apps Installed", "Data Usage (GB/month)"]].corr())



1. What are the 10 most common locations among users?

print(df["Location"].value\_counts().head(10))



1. How many users use the Android OS?

users\_android\_os = df[df['OS'] == 'Android'].shape[0]

print("\nNumber of Android Users:", users\_android\_os)

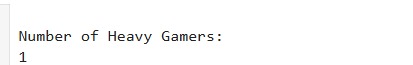


1. Find users who spend more than 5 hours daily on gaming

heavy\_gamers = df[df["Gaming Time (hrs/day)"] >5]

print("\nNumber of Heavy Gamers:")

print(len(heavy\_gamers))



### 📱 Phone Usage Patterns – Questions

# What is the distribution of daily screen time among users? *(Related to: Histogram of Screen Time)*

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

df = pd.read\_csv("phone\_usage\_india.csv")

sns.set\_palette("husl")

plt.figure(figsize=(8, 5))

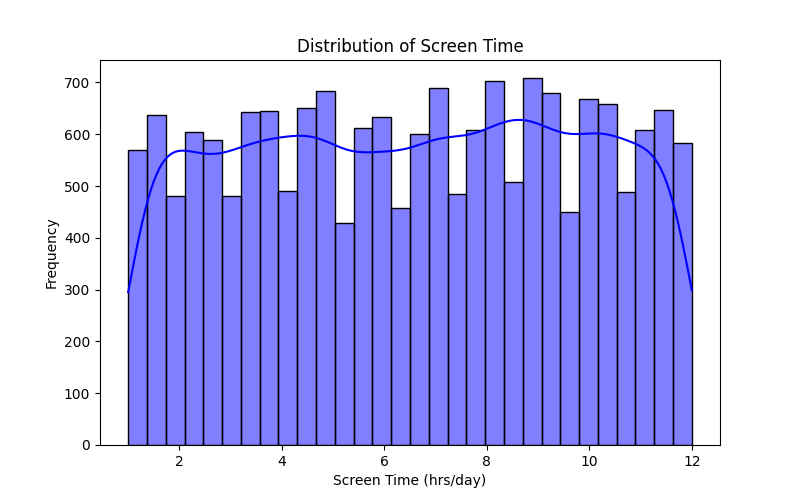
sns.histplot(df['Screen Time (hrs/day)'], bins=30, kde=True, color='blue')

plt.xlabel('Screen Time (hrs/day)')

plt.ylabel('Frequency')

plt.title('Distribution of Screen Time')

plt.show()



# How does the average screen time vary between different genders? *(Related to: Bar chart of Screen Time by Gender)*

gender\_screen\_time = df.groupby('Gender')['Screen Time (hrs/day)'].mean()

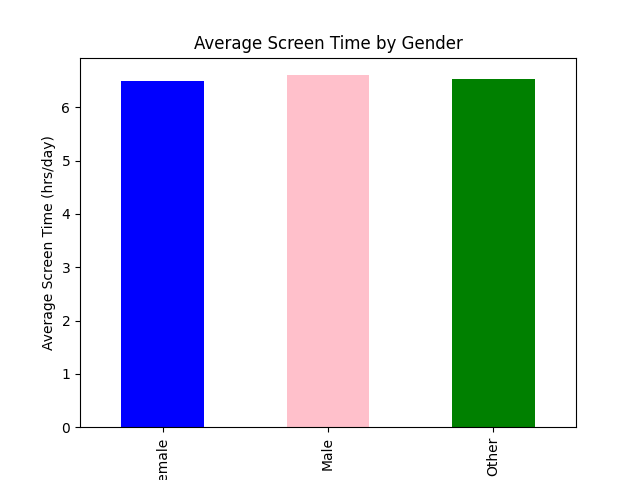
gender\_screen\_time.plot(kind='bar', color=['blue', 'pink', 'green'])

plt.xlabel('Gender')

plt.ylabel('Average Screen Time (hrs/day)')

plt.title('Average Screen Time by Gender')

plt.show()



# Which are the top 10 most popular phone brands among users in India? *(Related to: Bar chart of Phone Brand Counts)*

phone\_brand\_counts = df['Phone Brand'].value\_counts().head(10)

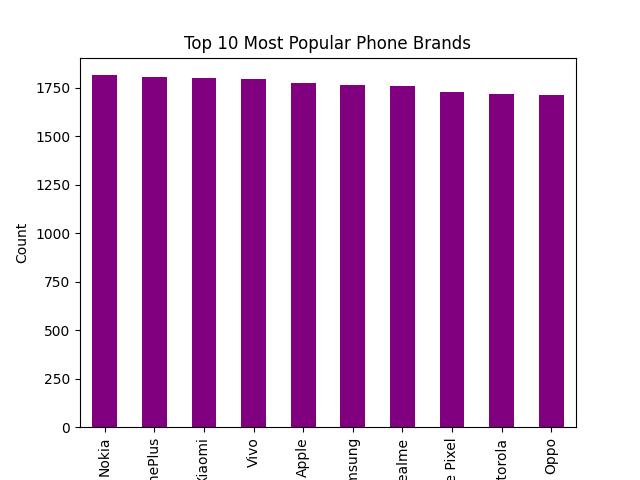
phone\_brand\_counts.plot(kind='bar', color='purple')

plt.xlabel('Phone Brand')

plt.ylabel('Count')

plt.title('Top 10 Most Popular Phone Brands')

plt.show()



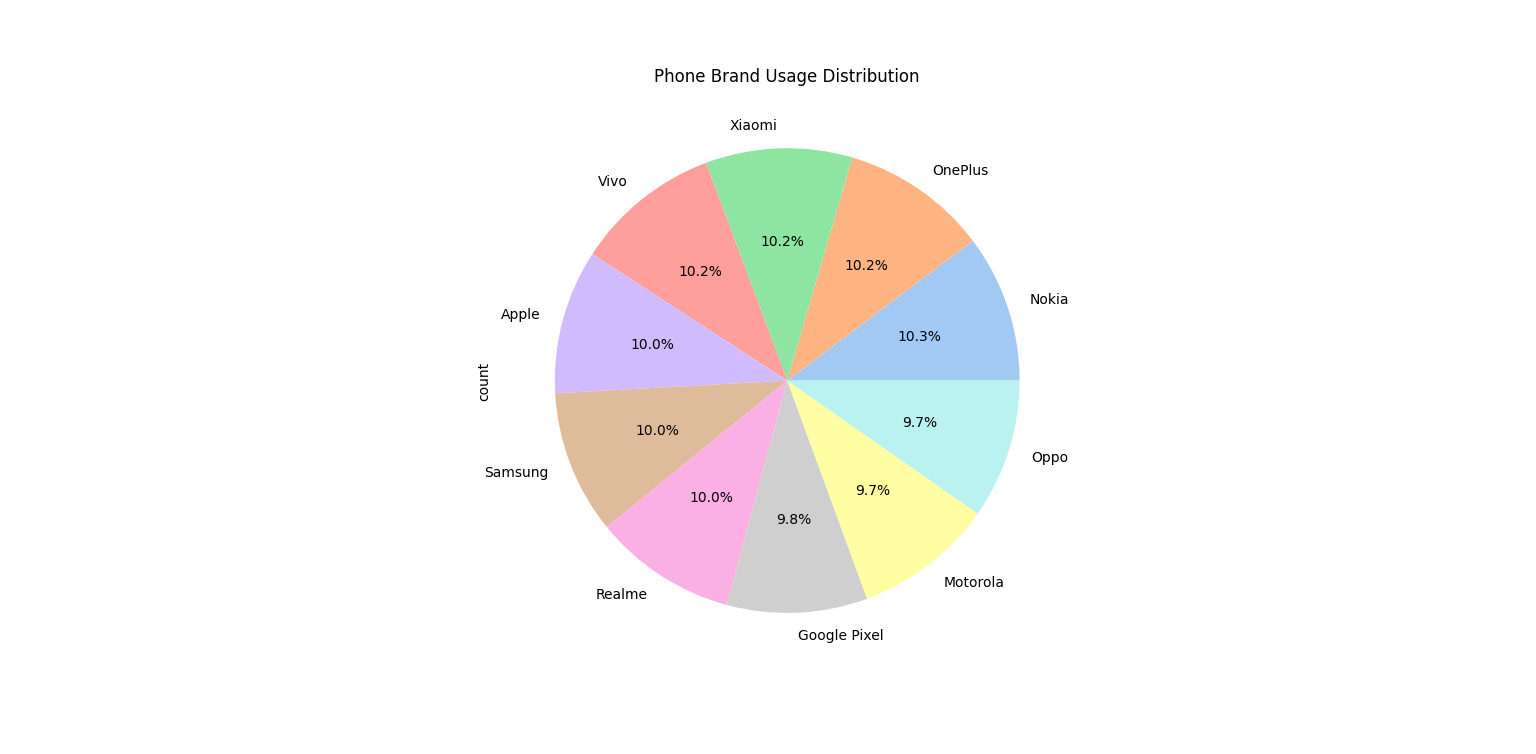
# What is the market share distribution of phone brands among users? *(Related to: Pie chart of Phone Brand Usage)*

plt.figure(figsize=(8, 8))

df["Phone Brand"].value\_counts().plot.pie(autopct="%1.1f%%", colors=sns.color\_palette("pastel"))

plt.title("Phone Brand Usage Distribution")

plt.show()



# What are the primary use cases for smartphones among users? *(Related to: Count Plot of Primary Use)*

plt.figure(figsize=(10, 5))

sns.countplot(

data=df,

y="Primary Use",

hue="Primary Use",

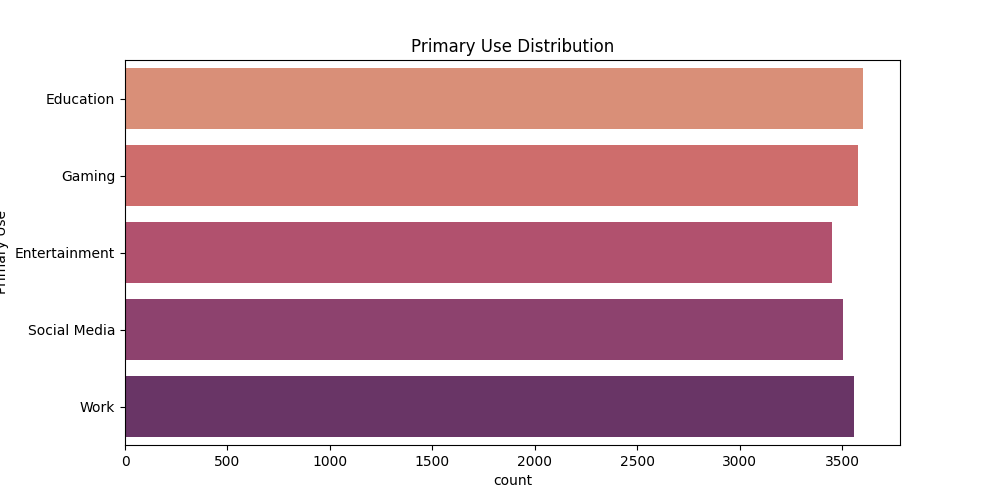
palette="flare",

legend=False

)

plt.title("Primary Use Distribution")

plt.show()



# What are the differences in streaming time between genders? *(Related to: Violin Plot of Streaming Time by Gender)*

plt.figure(figsize=(8, 5))

sns.violinplot(

data=df,

x="Gender",

y="Streaming Time (hrs/day)",

hue="Gender",

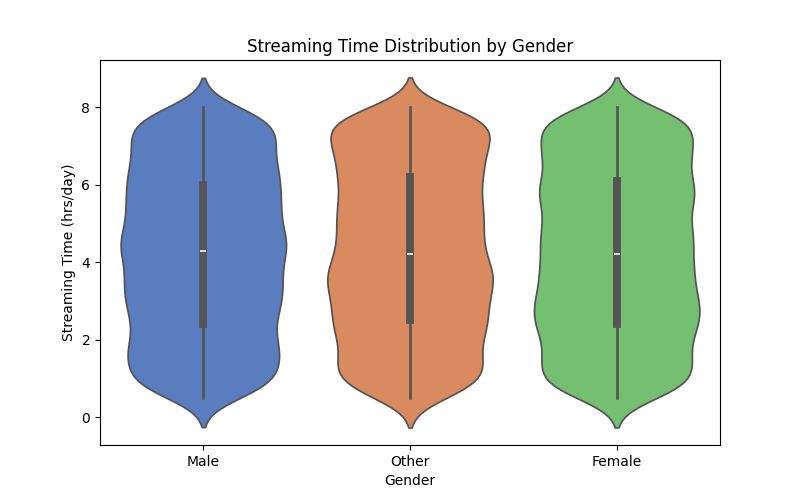
palette="muted",

legend=False

)

plt.title("Streaming Time Distribution by Gender")

plt.show()



# How does social media usage differ across age groups? *(Related to: Boxenplot of Social Media Time by Age)*

plt.figure(figsize=(10, 5))

sns.boxenplot(

data=df,

x="Age",

y="Social Media Time (hrs/day)",

hue="Age",

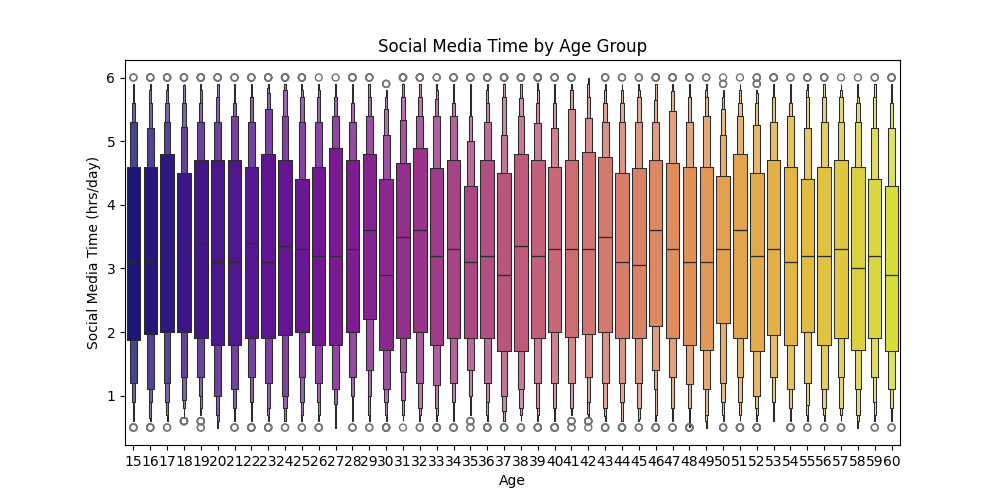
palette="plasma",

legend=False

)

plt.title("Social Media Time by Age Group")

plt.show()



### 👤 Demographic Insights – Questions

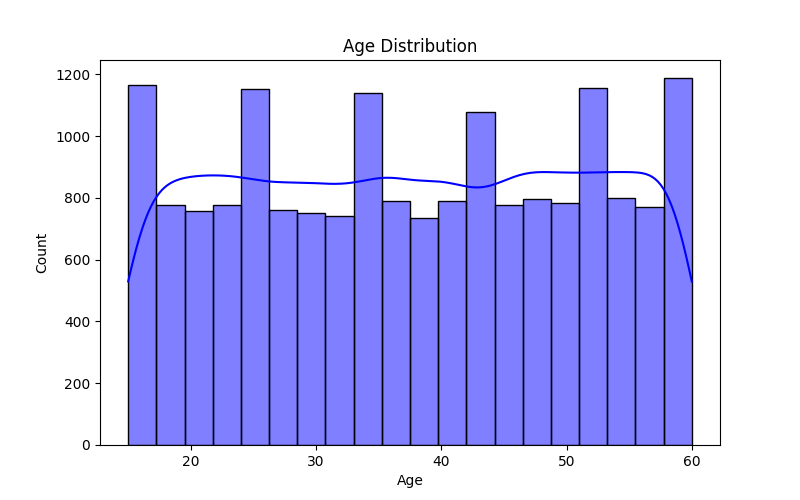
1. What is the age distribution of mobile phone users?  
   *(Related to: Histogram & KDE Plot of Age)*

plt.figure(figsize=(8, 5))

sns.histplot(df["Age"], bins=20, kde=True, color="blue")

plt.title("Age Distribution")

plt.show()



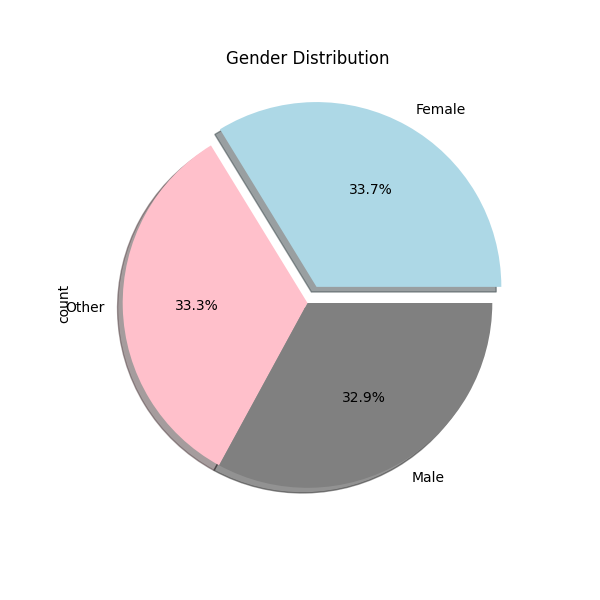
1. What is the gender distribution of the surveyed population?  
   *(Related to: Pie Chart of Gender Distribution)*

plt.figure(figsize=(6, 6))

df['Gender'].value\_counts().plot(kind='pie', autopct='%1.1f%%',explode=[0.1,0,0],shadow='True', colors=['lightblue', 'pink', 'gray'])

plt.title("Gender Distribution")

plt.show()



1. Which locations have the highest average data usage in India?  
   *(Related to: Bar Chart of Data Usage by Location)*

data\_usage\_location = df.groupby('Location')['Data Usage (GB/month)'].mean().nlargest(10)

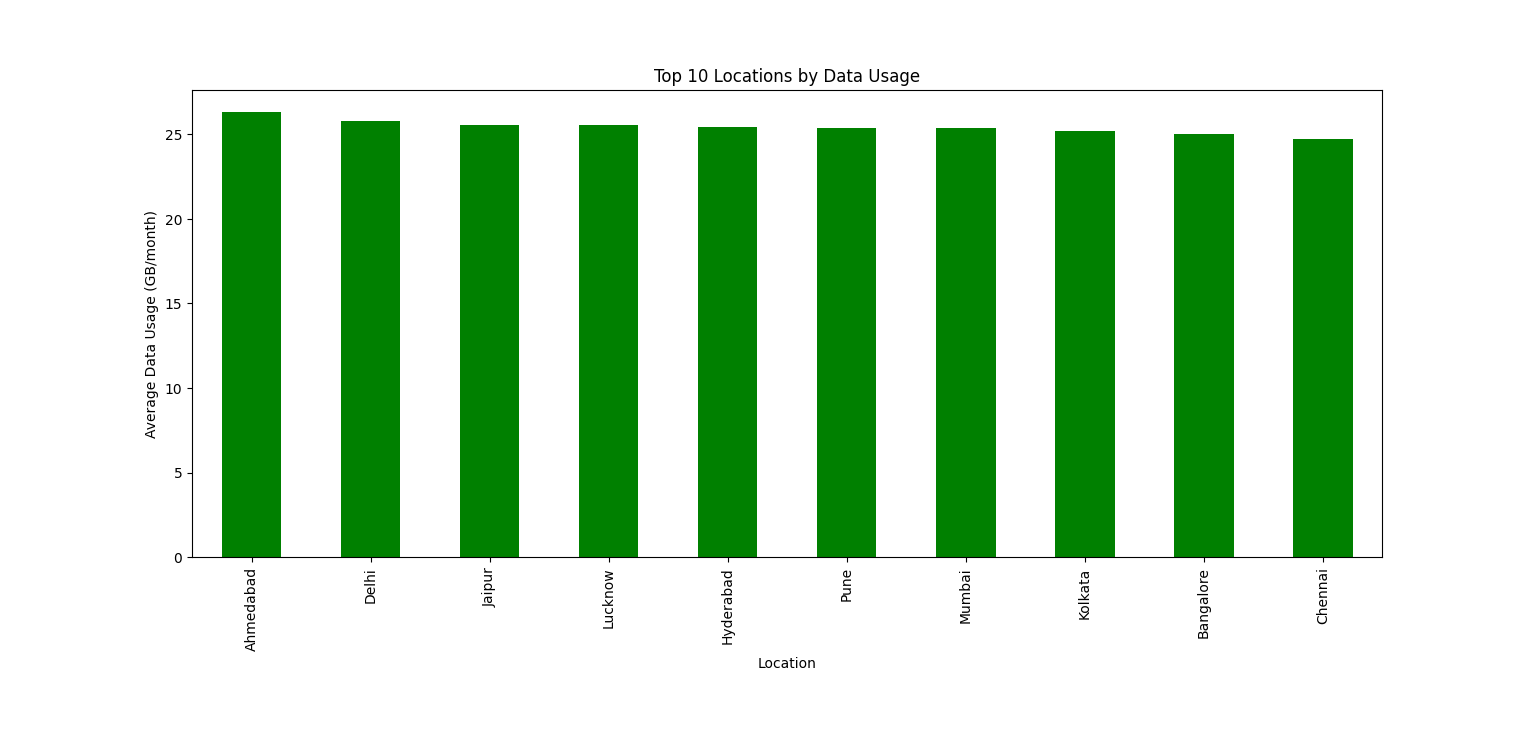
data\_usage\_location.plot(kind='bar', color='green')

plt.xlabel("Location")

plt.ylabel("Average Data Usage (GB/month)")

plt.title("Top 10 Locations by Data Usage")

plt.show()



### 📶 Data & Recharge Usage – Questions

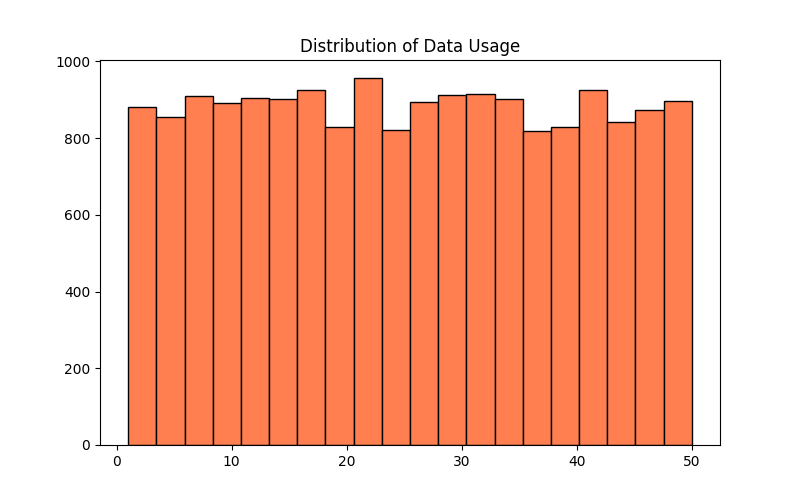
1. How is mobile data usage (GB/month) distributed across users?  
   *(Related to: Histogram of Data Usage)*

plt.figure(figsize=(8, 5))

plt.hist(df['Data Usage (GB/month)'], bins=20, color='coral', edgecolor='black')

plt.title('Distribution of Data Usage')

plt.show()



1. How does average data usage vary with age?  
   *(Related to: Bar Chart of Data Usage by Age)*

plt.figure(figsize=(10, 5))

sns.barplot(

data=df,

x="Age",

y="Data Usage (GB/month)",

hue="Age",

palette="viridis",

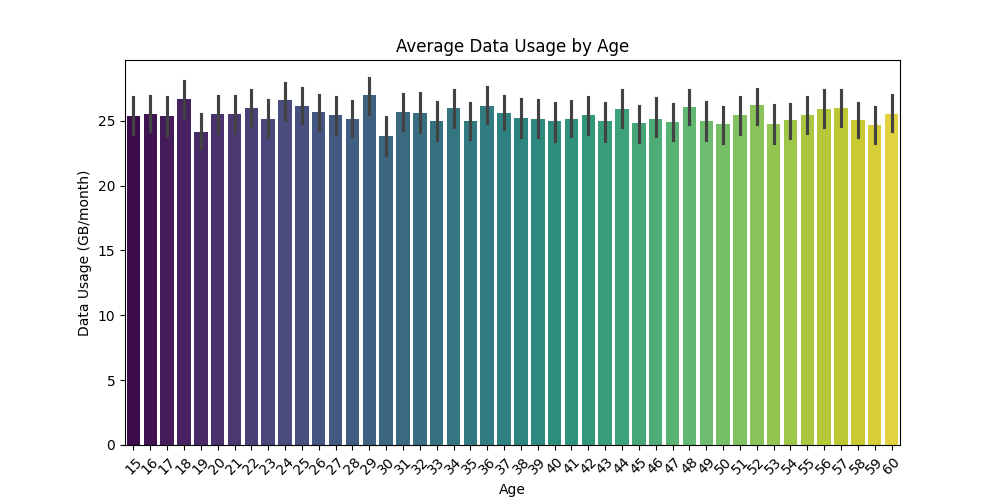
legend=False

)

plt.title("Average Data Usage by Age")

plt.xticks(rotation=45)

plt.show()



1. How does data usage vary across age groups (boxplot)?  
   *(Related to: Boxplot of Data Usage by Age)*

plt.figure(figsize=(10, 5))

sns.boxplot(

data=df,

x="Age",

y="Data Usage (GB/month)",

hue="Age",

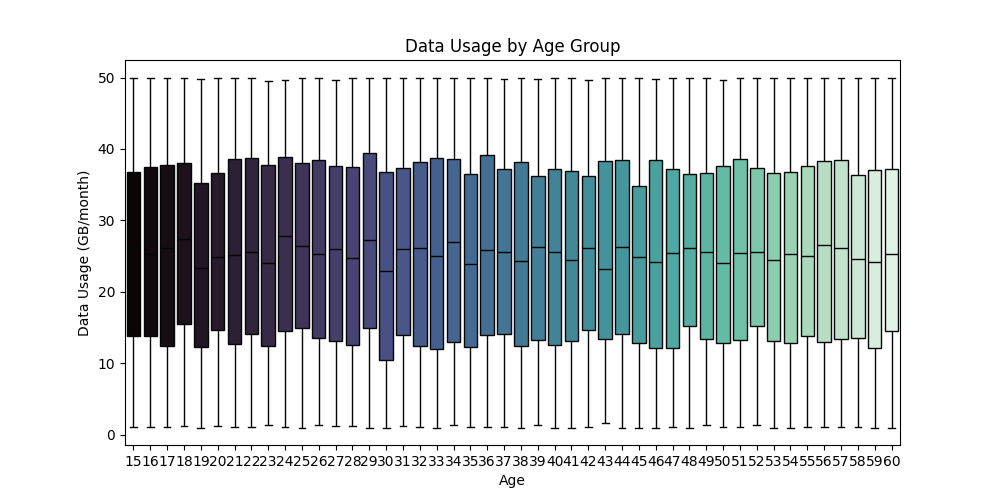
palette="mako",

legend=False

)

plt.title("Data Usage by Age Group")

plt.show()



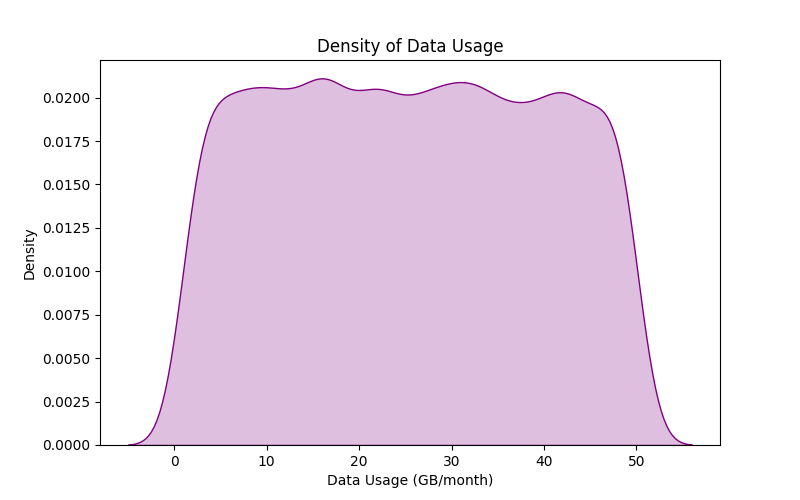
1. What does the density distribution of monthly data usage reveal?  
   *(Related to: KDE Density Plot of Data Usage)*

plt.figure(figsize=(8, 5))

sns.kdeplot(df["Data Usage (GB/month)"], fill=True, color="purple")

plt.title("Density of Data Usage")

plt.show()



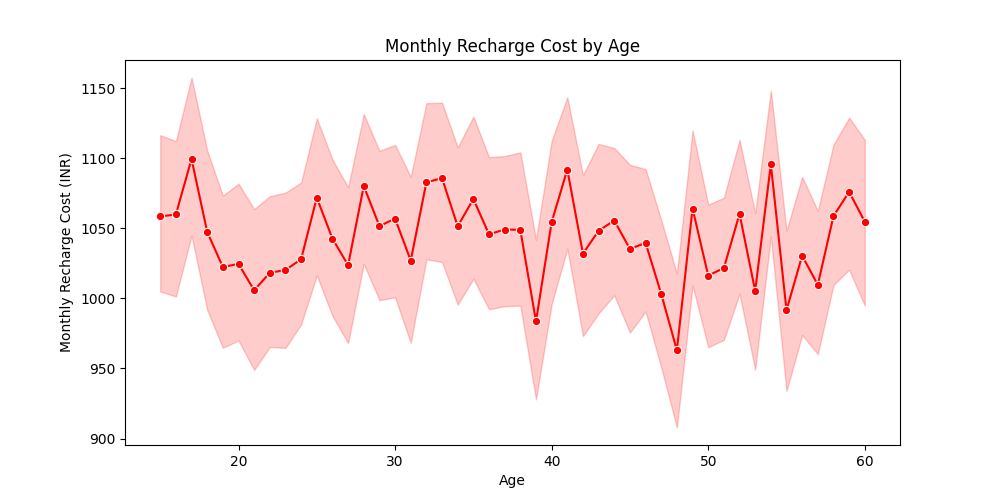
1. How does the monthly recharge cost change with age?  
   *(Related to: Line Chart of Recharge Cost by Age)*

plt.figure(figsize=(10, 5))

sns.lineplot(x=df["Age"], y=df["Monthly Recharge Cost (INR)"], marker="o", color="red")

plt.title("Monthly Recharge Cost by Age")

plt.show()



### 📈 Trends & Correlations – Questions

1. What is the relationship between age and average screen time?  
   *(Related to: Bar Chart of Age vs Screen Time)*

plt.figure(figsize=(10, 5))

sns.barplot(

data=df,

x="Age",

y="Screen Time (hrs/day)",

errorbar=None

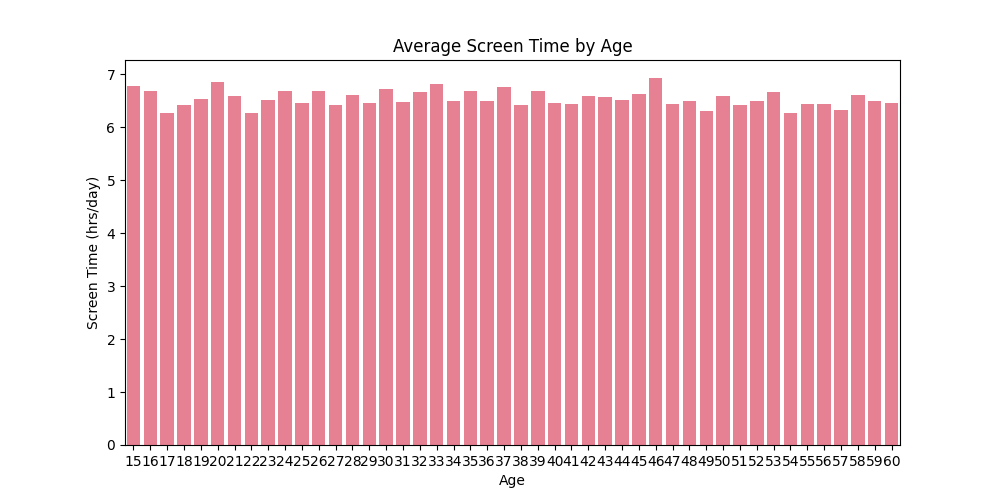
)

plt.xlabel("Age")

plt.ylabel("Screen Time (hrs/day)")

plt.title("Average Screen Time by Age")

plt.show()



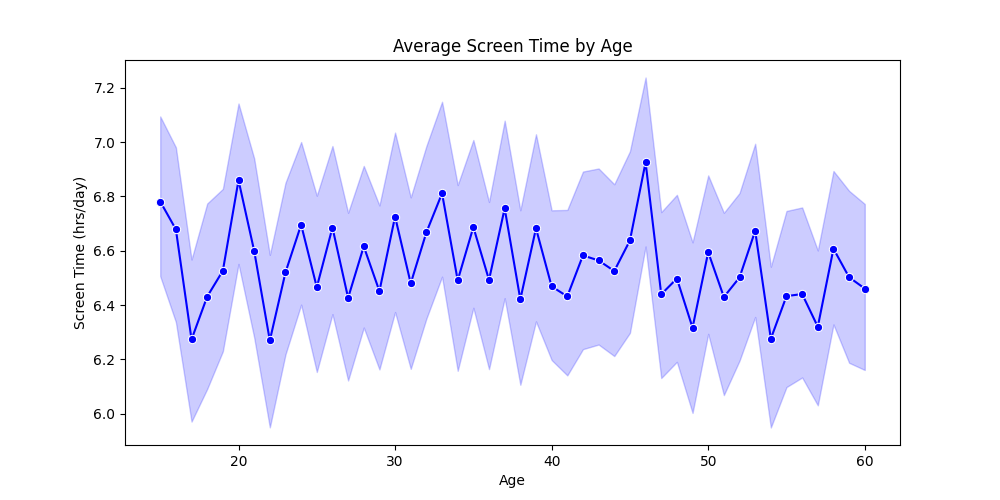
1. How does average screen time change with age (line chart)?  
   *(Related to: Line Chart of Screen Time by Age)*

plt.figure(figsize=(10, 5))

sns.lineplot(x=df["Age"], y=df["Screen Time (hrs/day)"], marker="o", color="blue")

plt.title("Average Screen Time by Age")

plt.show()



1. What is the combined trend of screen time and data usage across age groups?  
   *(Related to: Area Chart of Screen Time & Data Usage by Age)*

age\_grouped = df.groupby("Age")[["Screen Time (hrs/day)", "Data Usage (GB/month)"]].mean()

colors = ["#1f77b4", "#ff7f0e"]

plt.figure(figsize=(12, 6))

plt.stackplot(age\_grouped.index, age\_grouped["Screen Time (hrs/day)"], age\_grouped["Data Usage (GB/month)"],

labels=["Screen Time (hrs/day)", "Data Usage (GB/month)"], alpha=0.7, colors=colors)

plt.xlabel("Age", fontsize=14, fontweight='bold')

plt.ylabel("Average Usage", fontsize=14, fontweight='bold')

plt.title("Average Screen Time and Data Usage by Age", fontsize=16, fontweight='bold')

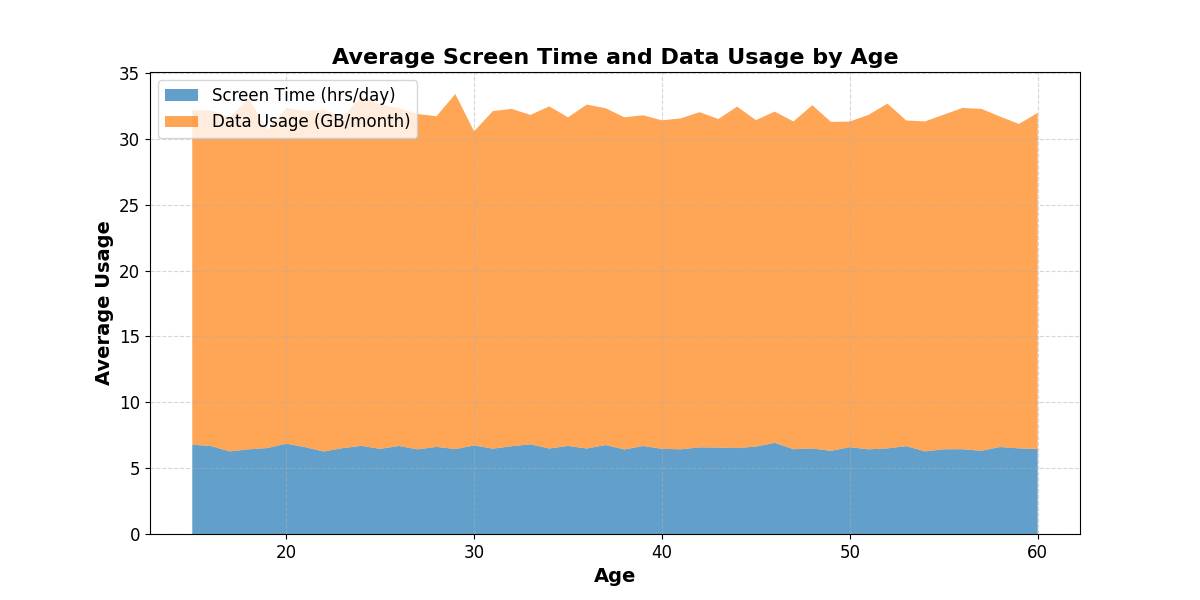
plt.legend(fontsize=12, loc="upper left")

plt.grid(True, linestyle="--", alpha=0.5)

plt.xticks(fontsize=12)

plt.yticks(fontsize=12)

plt.show()



1. What are the relationships between age, screen time, data usage, and recharge cost?  
   *(Related to: Pairplot of 100 Sampled Users)*

df\_sample = df[["Age", "Screen Time (hrs/day)", "Data Usage (GB/month)", "Monthly Recharge Cost (INR)"]].sample(100, random\_state=42)

pairplot = sns.pairplot(

df\_sample,

corner=True,

diag\_kind="kde",

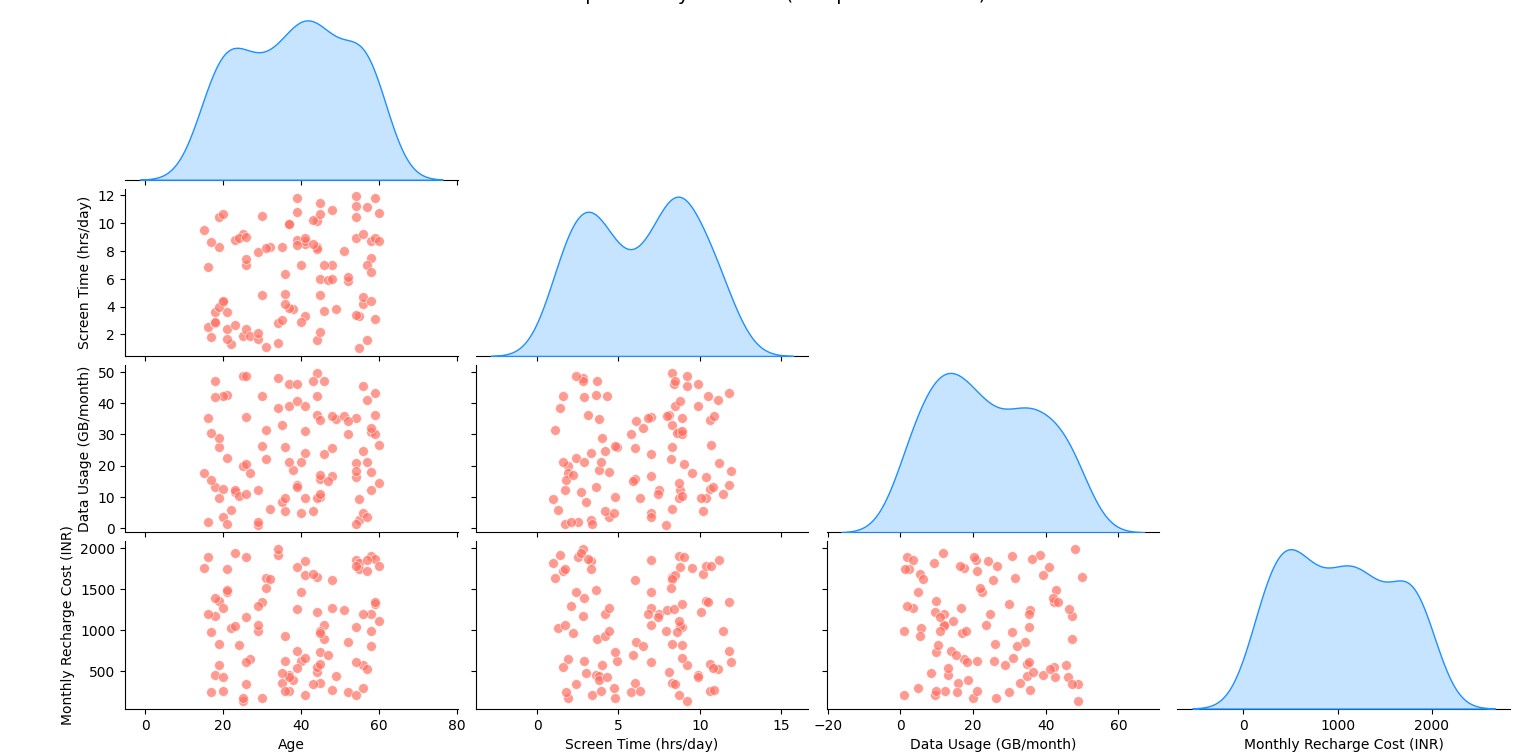
plot\_kws={"s": 50, "alpha": 0.7, "color": "#FF6F61"}, # Unique color for scatter plots

diag\_kws={"fill": True, "color": "#1E90FF"} # Unique color for KDE plots

)

plt.suptitle("Pairplot of Key Features (Sampled 100 Rows)", fontsize=14, y=1.02)

plt.show()



### 📋 Overall Usage Behavior – Question

1. "What are the most common primary activities users engage in on their phones?"  
   *(Perfect for count plot visualization)*

plt.figure(figsize=(10, 5))

sns.boxplot(

data=df.head(100),

x="Primary Use",

y="Screen Time (hrs/day)",

palette="Set3"

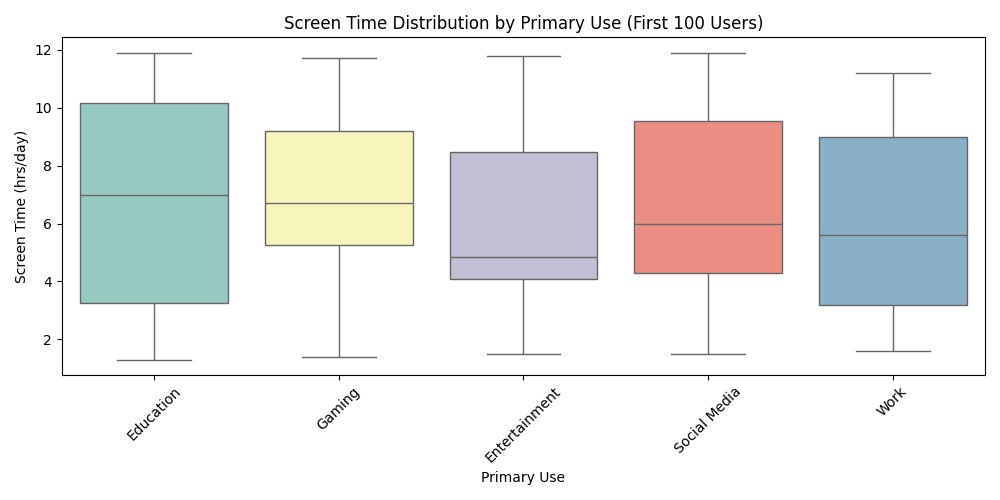
)

plt.title("Screen Time Distribution by Primary Use (First 100 Users)")

plt.xticks(rotation=45)

plt.tight\_layout()

plt.show()



# 5. Results and Insights

## 5.1 Key Patterns and Trends Identified

* Younger users (Teenagers and Young Adults) tend to have higher Gaming and Social Media Time.
* Higher E-commerce spending is correlated with more screen time and data usage.
* Android users tend to have higher data usage compared to iOS users.

## 5.2 Insights Derived from the Data Analysis

* Users with high data usage tend to have higher monthly recharge costs.
* The Primary Use category significantly impacts usage patterns (e.g., Entertainment users have the highest streaming time).
* Certain locations show higher spending trends, indicating potential regional differences in data affordability.

# 6. Conclusion

## 6.1 Summary of Findings

The analysis highlighted significant patterns in mobile usage and spending behavior. Younger users exhibit higher screen time, gaming engagement, and social media activity. Monthly recharge cost is influenced by data usage, app engagement, and e-commerce spending.

## 6.2 Implications of the Analysis

Understanding these insights can help telecom providers optimize pricing plans, target customer segments more effectively, and enhance user experience by offering personalized packages.

## 6.3 Limitations of the Project

* The dataset lacks information on internet speed and network quality, which may impact data usage.
* No temporal aspect (monthly trends) is considered, limiting trend analysis.
* Spending behaviors might be influenced by factors not captured in this dataset, such as income levels.

## 6.4 Future Work and Recommendations

* Incorporating additional data on network speed and pricing models.
* Analyzing temporal trends to understand seasonal variations in usage.
* Using clustering techniques to segment users based on their behavior for better customer profiling.

# 7. References

## 7.1 Data Sources

The dataset used for this analysis represents phone usage behavior across different demographics in India.

## 7.2 Tools and Libraries Used

* Python for data analysis and visualization.
* Pandas for data manipulation.
* Scikit-Learn for data preprocessing and modeling.
* Matplotlib & Seaborn for visualization.
* NumPy for numerical computations.