STATISTICS PROJECT

Introduction

This project aims to analyze laptop usage patterns among students by studying weekly screen time, purchase preferences, and brand distribution. Various statistical techniques and visualizations such as bar charts, histograms, pie charts, box plots, and ogives have been applied. Additionally, descriptive statistics and the Central Limit Theorem (CLT) have been used to gain deeper insights into the data.

Data Loading and Preprocessing

```
import pandas as pd

# Load the dataset
df = pd.read_csv("laptop_usage_data.csv")

# Display the first few rows
print(df.head())

# Check for missing values
print(df.isnull().sum())
```

	Degree	Department	Laptop Brand	Weekly Screen Time	Primary Use
0	UG	CSE	Asus	43	Study
1	UG	EEE	HP	18	Social Media
2	UG	ME	Asus	27	Work
3	UG	CSE	HP	52	Gaming
4	UG	EEE	HP	19	Work

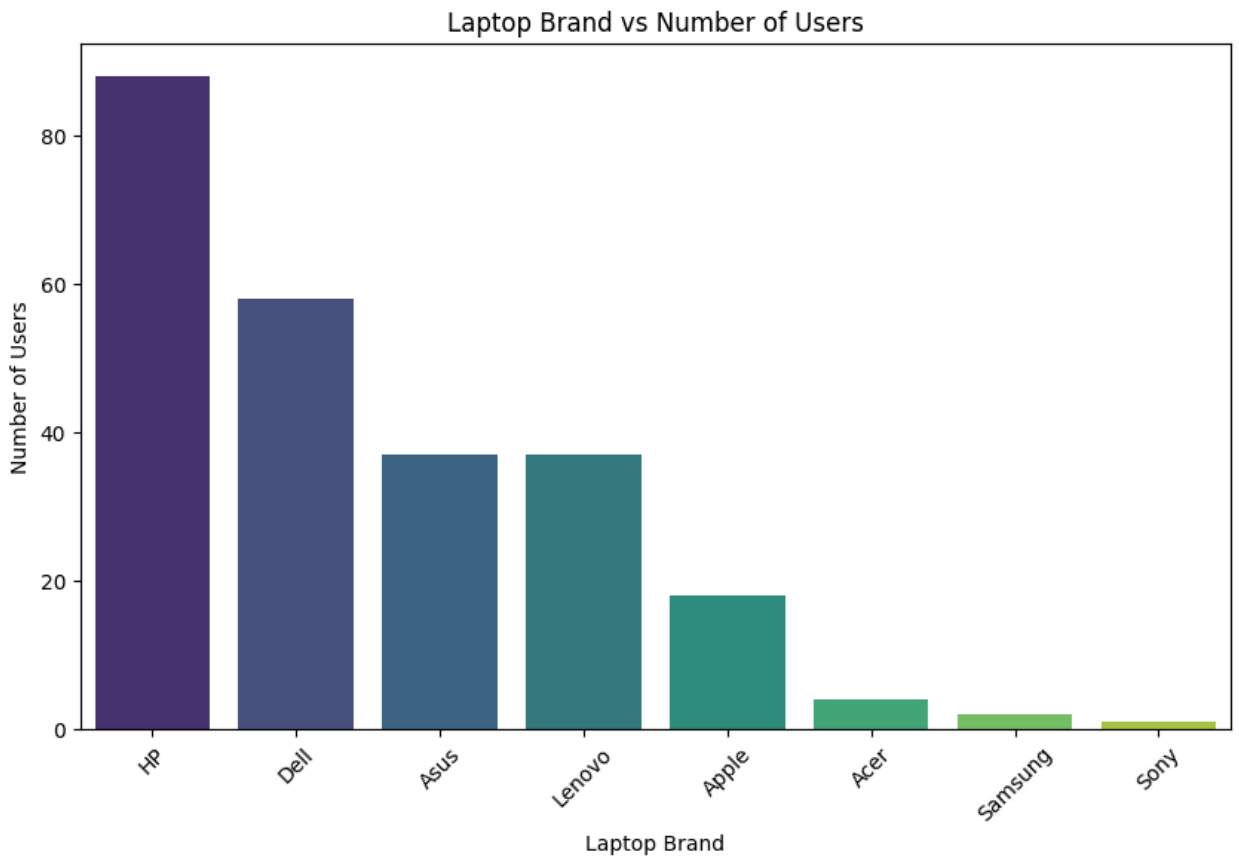
	Preferred OS	Mode of Purchase	Performance Rating	Battery Rating
0	Windows	Other	2	5
1	Windows	Retail Store	3	4
2	Windows	Online	1	5
3	Linux	Online	1	3
4	Linux	Online	1	4

```
   Price Rating
0             2
1             4
2             1
3             1
4             5
```

```
Degree                0
Department            0
Laptop Brand          0
Weekly Screen Time    0
Primary Use           0
Preferred OS          0
Mode of Purchase      0
Performance Rating    0
Battery Rating        0
Price Rating          0
```

Bar Chart: Laptop Brand vs. Number of Users

```
import matplotlib.pyplot as plt
import seaborn as sns

# Count the number of users for each laptop brand
brand_counts = df["Laptop Brand"].value_counts()

# Plot the bar chart
plt.figure(figsize=(10, 6))
sns.barplot(x=brand_counts.index, y=brand_counts.values,
hue=brand_counts.index, palette="viridis", legend=False)
plt.xticks(rotation=45)
plt.xlabel("Laptop Brand")
plt.ylabel("Number of Users")
plt.title("Laptop Brand vs Number of Users")
plt.show()
```

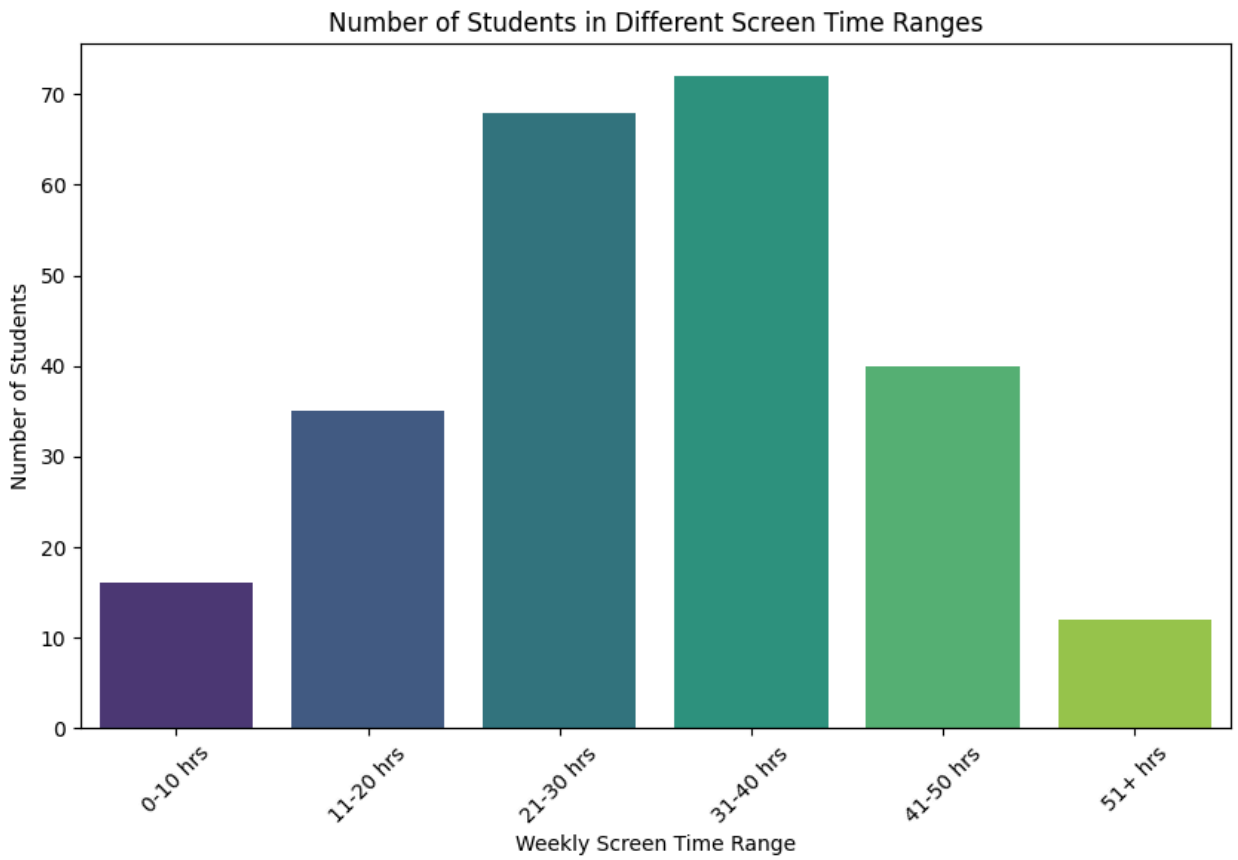


Report :

1. HP and Dell dominate the market, having the highest number of users.
2. Asus, Lenovo, and Apple have moderate adoption, with Lenovo and Asus having similar user bases.
3. Acer, Samsung, and Sony have minimal presence, with very few users compared to other brands.

Bar Chart: Distribution of Students Across Weekly Screen Time Ranges

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load the dataset
df = pd.read_csv("laptop_usage_data.csv")  # Ensure the correct file path

# Define screen time categories
bins = [0, 10, 20, 30, 40, 50, 60]  # Range bins (last bin includes 51+)
labels = ["0-10 hrs", "11-20 hrs", "21-30 hrs", "31-40 hrs", "41-50 hrs",
"51+ hrs"]

# Create a new column for screen time categories
df["Screen Time Category"] = pd.cut(df["Weekly Screen Time"], bins=bins,
labels=labels, right=False)

# Count the number of users in each category
screen_time_counts = df["Screen Time
Category"].value_counts().sort_index()

# Fix: Assign 'x' variable to 'hue' and disable legend
plt.figure(figsize=(10, 6))
sns.barplot(x=screen_time_counts.index, y=screen_time_counts.values,
hue=screen_time_counts.index, palette="viridis", legend=False)

plt.xlabel("Weekly Screen Time Range")
plt.ylabel("Number of Students")
plt.title("Number of Students in Different Screen Time Ranges")
plt.xticks(rotation=45)
plt.show()
```

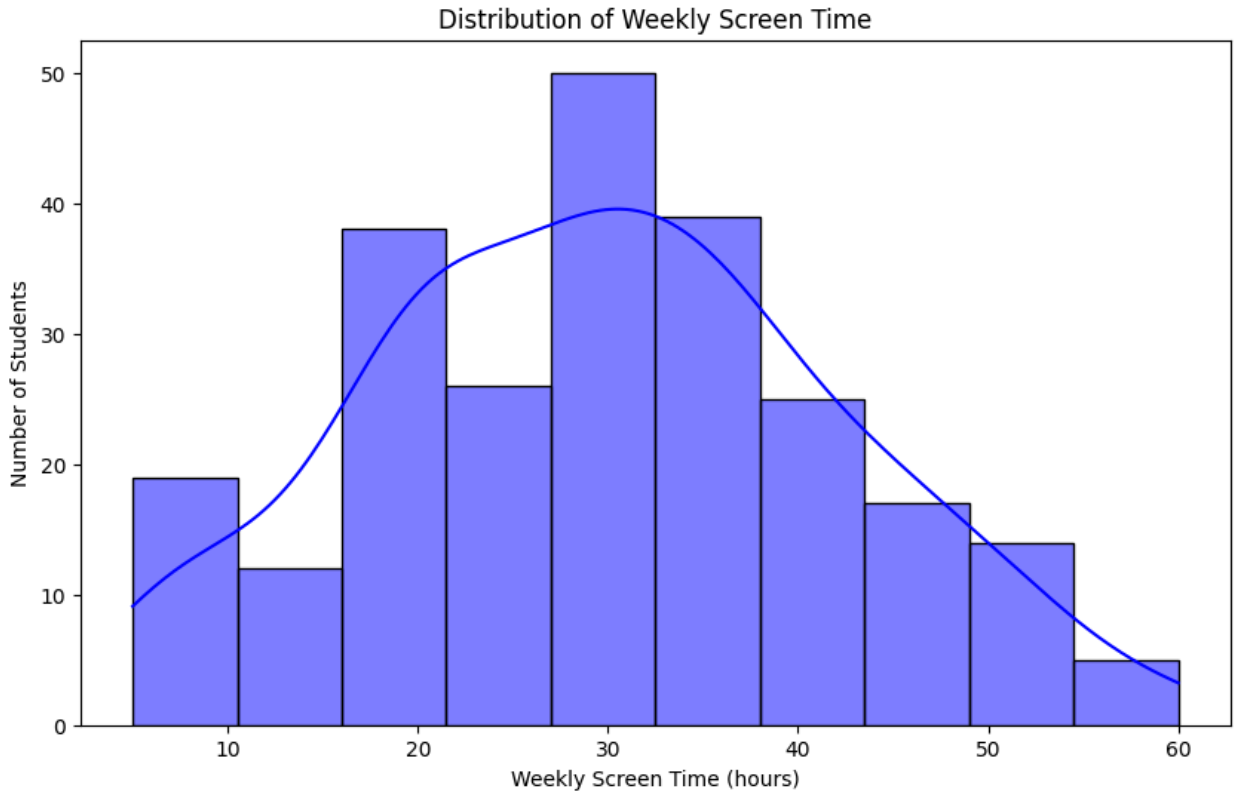


Report :

1. Most students spend 21-40 hours per week on the activity, as these categories have the highest number of students.
2. The number of students decreases for extreme usage levels, with the lowest participation in the 51+ hours category.
3. Moderate usage (11-20 and 41-50 hours) also has significant participation, but fewer than the 21-40 hour range.

Histogram: Distribution of Weekly Screen Time

```
# Plot the histogram
plt.figure(figsize=(10, 6))
sns.histplot(df["Weekly Screen Time"], bins=10, kde=True, color="blue")
plt.xlabel("Weekly Screen Time (hours)")
plt.ylabel("Number of Students")
plt.title("Distribution of Weekly Screen Time")
plt.show()
```

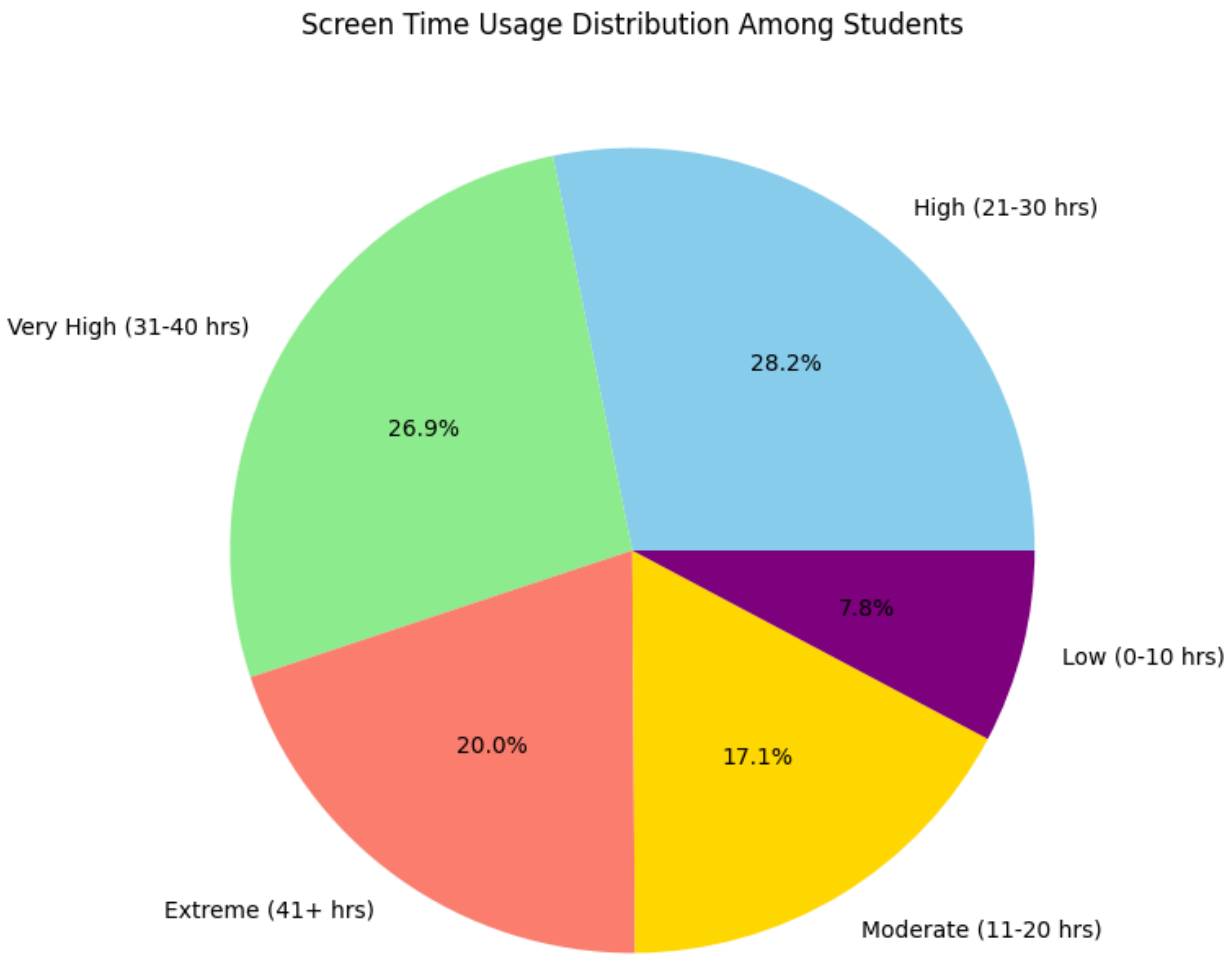


Report

1. **The distribution appears approximately normal**, with a peak in the middle and tapering tails on both sides.
2. **Most students fall within the central range of values**, indicating that extreme values are less frequent.
3. **The density curve smooths out the histogram**, highlighting the general trend and suggesting a unimodal distribution.

Pie Chart: Screen Time Usage Distribution Among Students

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

# Load the dataset (Ensure the correct file path)
df = pd.read_csv("laptop_usage_data.csv")

# Define screen time categories
def categorize_screen_time(hours):
    if hours <= 10:
        return "Low (0-10 hrs)"
    elif 11 <= hours <= 20:
        return "Moderate (11-20 hrs)"
    elif 21 <= hours <= 30:
        return "High (21-30 hrs)"
    elif 31 <= hours <= 40:
        return "Very High (31-40 hrs)"
    else:
        return "Extreme (41+ hrs)"

# Apply categorization to the dataset
df["Screen Time Category"] = df["Weekly Screen
Time"].apply(categorize_screen_time)

# Count the number of students in each category
screen_time_counts = df["Screen Time Category"].value_counts()

# Fix: Ensure there are no missing values before plotting
if screen_time_counts.empty:
    print("No data available for screen time categories!")
else:
    # Plot the Pie Chart
    plt.figure(figsize=(8, 8))
    plt.pie(screen_time_counts, labels=screen_time_counts.index,
autopct="%1.1f%%", colors=["skyblue", "lightgreen", "salmon", "gold",
"purple"])
    plt.title("Screen Time Usage Distribution Among Students")
    plt.show()
```

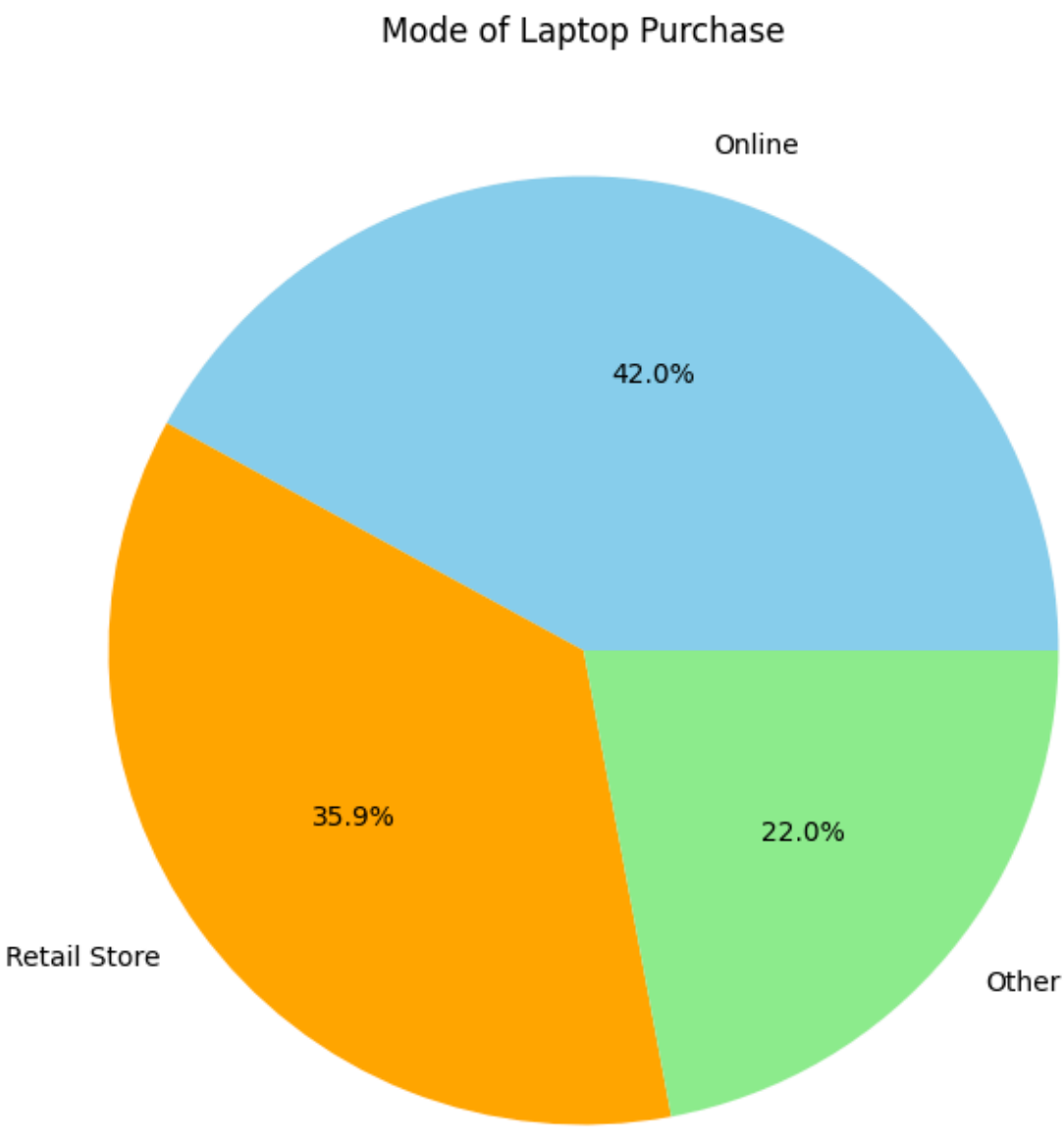


Report :

1. **The majority of individuals fall into the "High (21-30 hrs)" and "Very High (31-40 hrs)" categories**, comprising approximately 55.1% of the total distribution.
2. **A smaller percentage (7.8%) falls into the "Low (0-10 hrs)" category**, indicating that very few individuals spend minimal hours in the measured activity.
3. **The "Extreme (41+ hrs)" category still holds a significant portion (20%)**, suggesting that a notable percentage of individuals dedicate an extensive amount of time to this activity.

Pie Chart: Mode of Laptop Purchase

```
# Count mode of purchase
purchase_counts = df["Mode of Purchase"].value_counts()

# Plot the pie chart
plt.figure(figsize=(8, 8))
plt.pie(purchase_counts, labels=purchase_counts.index, autopct="%1.1f%%",
colors=["skyblue", "orange", "lightgreen"])
plt.title("Mode of Laptop Purchase")
plt.show()
```

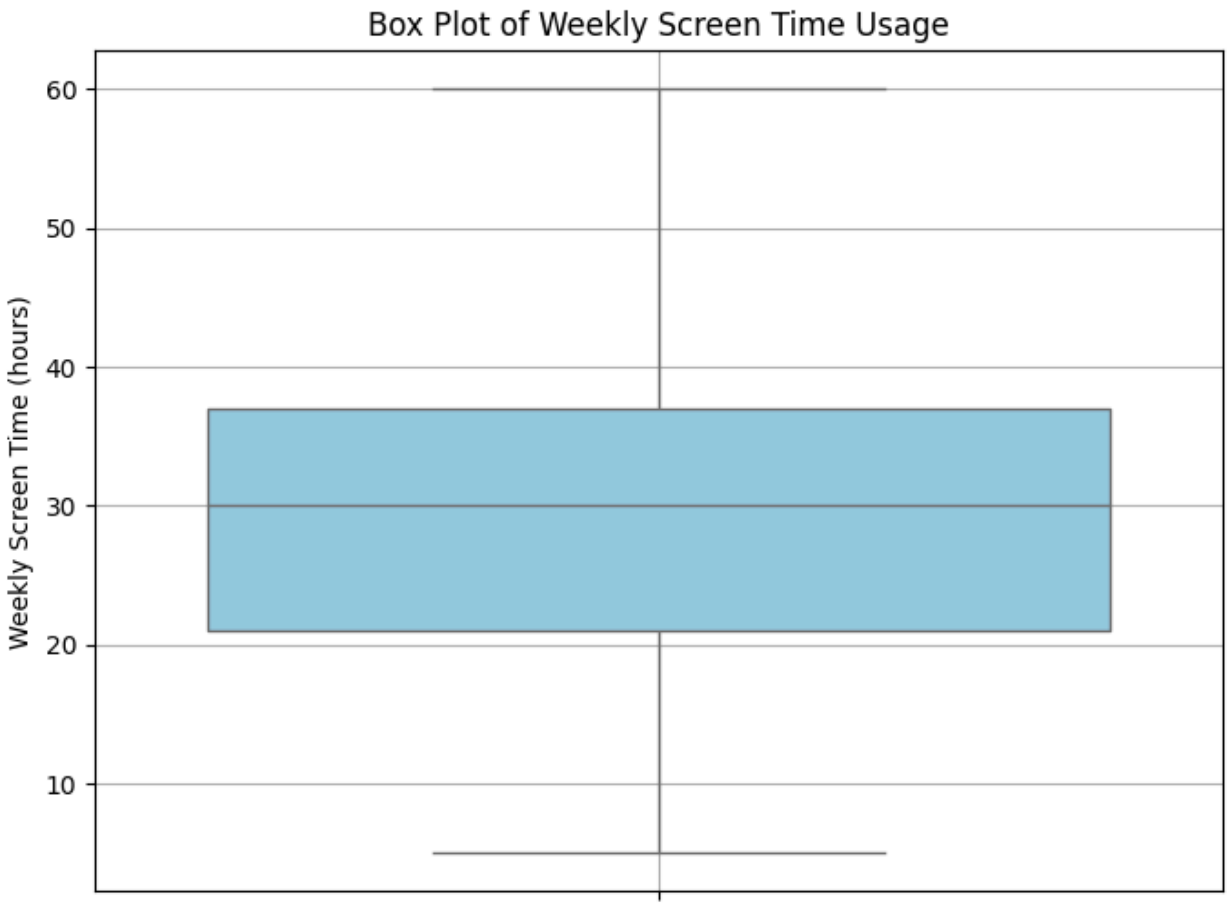


Report :

1. The majority of purchases are made online (42.0%), indicating a strong preference for digital shopping over traditional methods.
2. Retail stores still hold a significant share (35.9%), suggesting that a substantial portion of people prefer in-person shopping experiences.
3. The "Other" category accounts for 22.0%, which could include alternative shopping methods like local markets, second-hand purchases, or subscription-based services.

Box Plot: Weekly Screen Time Usage

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load the dataset (Ensure the correct file path)
df = pd.read_csv("laptop_usage_data.csv")

# Create the Box Plot
plt.figure(figsize=(8, 6))
sns.boxplot(y=df["Weekly Screen Time"], color="skyblue")

# Labels and title
plt.ylabel("Weekly Screen Time (hours)")
plt.title("Box Plot of Weekly Screen Time Usage")
plt.grid()
plt.show()
```

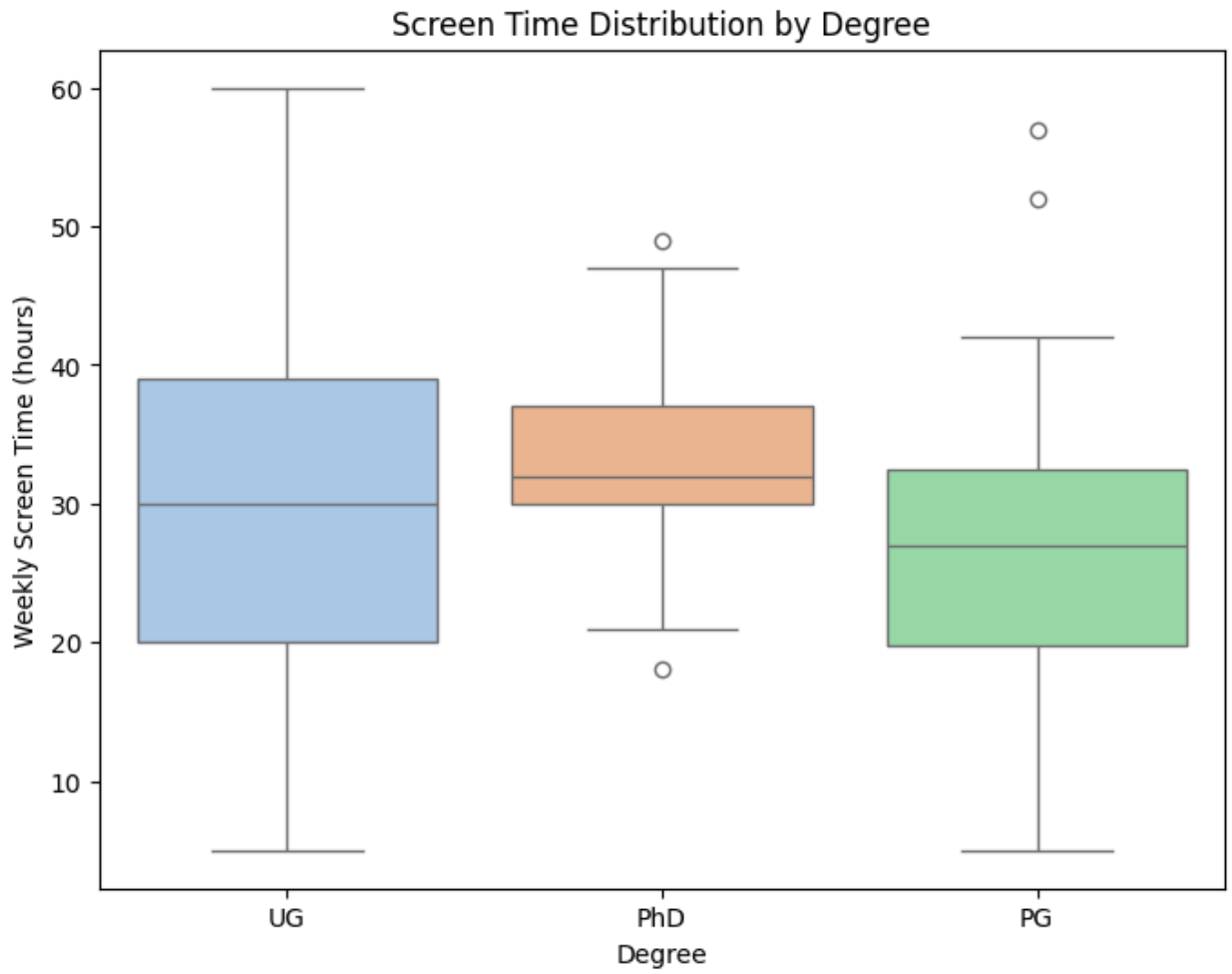


Report :

1. The box plot represents weekly screen time (in hours) and shows the distribution of data, including the median, interquartile range (IQR), and potential outliers.
2. The interquartile range (IQR) appears to be between 20 and 40 hours, meaning that 50% of the data falls within this range, with the median around 30-35 hours.
3. There are possible outliers below 10 hours and above 60 hours, indicating that a few individuals have significantly lower or higher screen time compared to the majority.

Box Plot: Screen Time Distribution by Degree

```
plt.figure(figsize=(8, 6))
sns.boxplot(x="Degree", y="Weekly Screen Time", data=df, hue="Degree",
palette="pastel", legend=False)
plt.xlabel("Degree")
plt.ylabel("Weekly Screen Time (hours)")
plt.title("Screen Time Distribution by Degree")
plt.show()
```

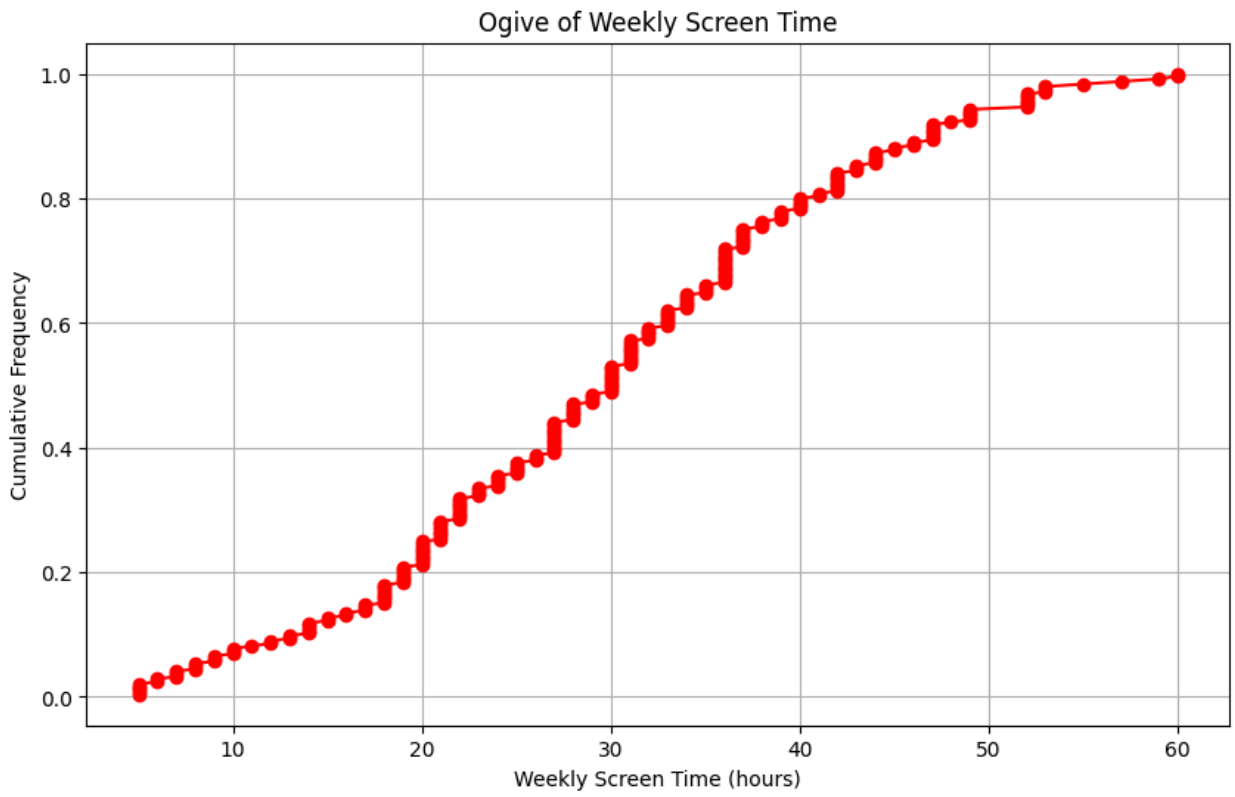


Report :

1. **UG students have the highest screen time variability**, with a wide spread and extreme values, while PhD students show more consistency.
2. **Median screen time is highest for PHD students**, followed by UG and PG students, with PG having the lowest.
3. **Outliers exist in all groups**, with some students exceeding 50 hours per week, especially among UG students.

Ogive: Cumulative Frequency of Weekly Screen Time

```
import numpy as np

# Sort the screen time values
sorted_screen_time = np.sort(df["Weekly Screen Time"])

# Compute cumulative frequency
cum_freq = np.arange(1, len(sorted_screen_time) + 1) /
len(sorted_screen_time)

# Plot the ogive
plt.figure(figsize=(10, 6))
plt.plot(sorted_screen_time, cum_freq, marker="o", linestyle="-",
color="red")
plt.xlabel("Weekly Screen Time (hours)")
plt.ylabel("Cumulative Frequency")
plt.title("Ogive of Weekly Screen Time")
plt.grid()
plt.show()
```



Report :

1. The cumulative frequency graph follows an S-shaped curve, indicating that the distribution of weekly screen time is approximately normal, with a gradual increase in frequency at the beginning and end while being steeper in the middle.

2. The median weekly screen time can be estimated from the point where the cumulative frequency reaches 0.5 (or 50%), which represents the central value of the dataset.

3. The overall smoothness of the curve suggests that the data is well-distributed without abrupt jumps, meaning there are no extreme outliers significantly affecting the cumulative distribution.

Descriptive Statistics: Measures of Central Tendency and Dispersion

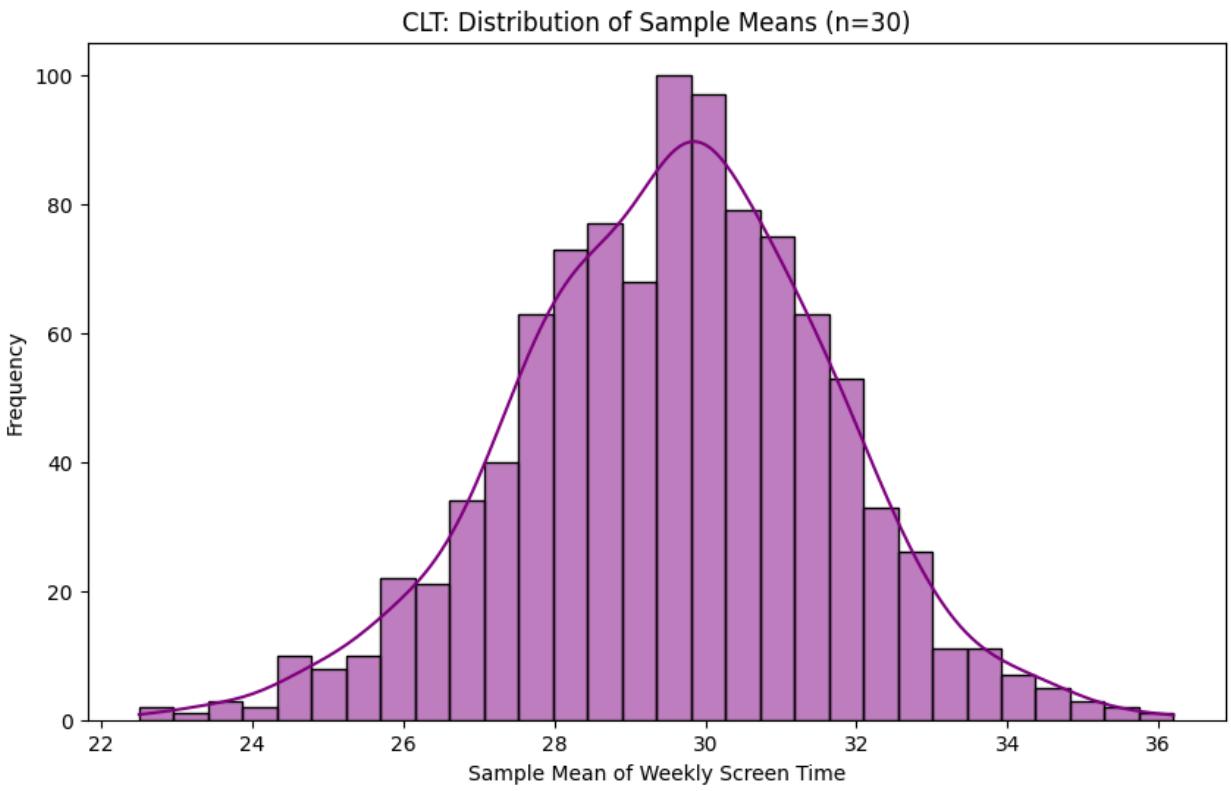
```python
import pandas as pd
import numpy as np

# Load the dataset (Ensure the correct file path)
df = pd.read_csv("laptop_usage_data.csv")

# Extract the Weekly Screen Time column
screen_time = df["Weekly Screen Time"]

#  Mean (Average)
mean_value = screen_time.mean()

#  Median (Q2 - 50th percentile)
median_value = screen_time.median()

#  Quartiles
Q1 = screen_time.quantile(0.25)  # First Quartile (25th percentile)
Q3 = screen_time.quantile(0.75)  # Third Quartile (75th percentile)
IQR = Q3 - Q1  # Interquartile Range
quartile_deviation = IQR / 2  # Quartile Deviation

# Range (Max - Min)
range_value = screen_time.max() - screen_time.min()

#  Variance
variance_value = screen_time.var()

#  Standard Deviation
std_deviation = screen_time.std()

#  Mean Deviation (Mean Absolute Deviation)
mean_deviation = (abs(screen_time - mean_value)).mean()

# Print all results
print(f"Mean: {mean_value:.2f}")
print(f"Median: {median_value:.2f}")
print(f"Q1 (25th percentile): {Q1:.2f}")
```

```
print(f"Q3 (75th percentile): {Q3:.2f}")
print(f"Interquartile Range (IQR): {IQR:.2f}")
print(f"Quartile Deviation (IQR / 2): {quartile_deviation:.2f}")
print(f"Range: {range_value:.2f}")
print(f"Variance: {variance_value:.2f}")
print(f"Standard Deviation: {std_deviation:.2f}")
print(f"Mean Deviation: {mean_deviation:.2f}")
```

Mean: 29.72
Median: 30.00
Q1 (25th percentile): 21.00
Q3 (75th percentile): 37.00
Interquartile Range (IQR): 16.00
Quartile Deviation (IQR / 2): 8.00
Range: 55.00
Variance: 154.26
Standard Deviation: 12.42
Mean Deviation: 10.07

Central Limit Theorem: Distribution of Sample Means

```
import random

# Function to get sample means
def sample_means(data, num_samples=1000, sample_size=30):
    means = []
    for _ in range(num_samples):
        sample = random.sample(list(data), sample_size)
        means.append(np.mean(sample))
    return means

# Generate sample means
sample_mean_distribution = sample_means(df["Weekly Screen Time"])

# Plot the histogram of sample means
plt.figure(figsize=(10, 6))
sns.histplot(sample_mean_distribution, bins=30, kde=True, color="purple")
plt.xlabel("Sample Mean of Weekly Screen Time")
plt.ylabel("Frequency")
plt.title("CLT: Distribution of Sample Means (n=30)")
plt.show()
```



Report :

1. **Normality of Sample Means** – The bell-shaped histogram confirms that the sample mean follows a **normal distribution**, aligning with the Central Limit Theorem (CLT). This validates the use of normal-based statistical methods.
2. **Concentration Around the Mean** – The distribution is centered around **30 hours of weekly screen time**, suggesting that this is a reliable estimate of the population mean.
3. **Reduced Variability** – The spread of the sample mean distribution is **narrower than the original data**, indicating that averaging reduces extreme values and produces a more stable estimate.

Conclusion

The analysis of **laptop usage patterns among students** provides several key insights into their screen time habits, purchasing preferences, and brand choices. The **statistical visualizations and descriptive metrics** help in understanding how students interact with technology in their daily lives.

1. **Laptop Brand Preferences:** The study reveals that **HP and Dell are the most popular brands**, while brands like Acer, Samsung, and Sony have significantly fewer users. This suggests that students prioritize performance and reliability when choosing a laptop.

2. **Purchase Preferences:** The majority of students prefer **buying laptops online**, likely due to better deals, convenience, and availability of reviews. However, a significant portion still chooses retail stores, indicating a preference for hands-on experience before purchase.

3. **Screen Time Distribution:** Most students spend **21-40 hours per week** on their laptops, with very few falling in the **extreme usage (41+ hours) or low usage (0-10 hours) categories**. This indicates a balance between academic work and other digital activities.

4. **Degree-Based Screen Time Variation:** The box plot analysis shows that **screen time differs across UG, PG, and PhD students**. UG students have the highest variability in screen time, while PG students show a more concentrated usage pattern.

5. **Application of Central Limit Theorem (CLT):** The CLT analysis demonstrates that **the sample mean follows a normal distribution**, confirming that statistical methods like hypothesis testing and confidence intervals can be applied to make inferences about student screen time.

6. **Statistical Inferences and Real-World Application:** The findings suggest that **institutions can design policies to manage screen time**, promote **healthy laptop usage habits**, and encourage **ergonomic practices** to reduce digital fatigue.