Smartphone Market Analysis and Pricing Trends Using Python:

In the **SMARTPRIX SMARTPHONES** project, smartphone data was gathered and analyzed from the Smartprix website using web scraping. This process was followed by data cleaning and **Exploratory Data Analysis (EDA)** to derive insights into smartphone features and pricing trends. The project also involved applying **statistical tests** to validate findings and creating data visualizations to better communicate insights. A detailed 102-page document was created.

Web Scraping:

- **Tools Used**: Employed **Selenium** for automating browser interaction and **BeautifulSoup** to parse the HTML content.
- **Data Extracted**: Collected information including model names, prices, operating systems, SIM card types, processors, RAM, battery capacity, display details, camera specifications, and memory card support.
- **Challenge**: Managed dynamic content on the Smartprix website by automating clicks with Selenium.

Data Cleaning:

- **Issues Identified**:
  - **Inconsistent brand names** (e.g., "SAMSUNG" vs. "Samsung").
  - **Misplaced data**: For example, information like operating system, Bluetooth, and FM radio appeared in columns like memory card.
  - **Outliers**: Extreme values, such as luxury phones made of gold and diamond, were removed as they didn't represent standard smartphone pricing.
- **Steps Taken**:
  - Converted object columns (e.g., price, RAM, internal memory) into appropriate data types for analysis.
  - Filled missing values using advanced techniques like **KNNImputer** for numerical data and **SimpleImputer** for categorical data.
  - Created new columns for features such as **5G**, **NFC**, and **IR Blasters** based on SIM information.

Splitting Columns:

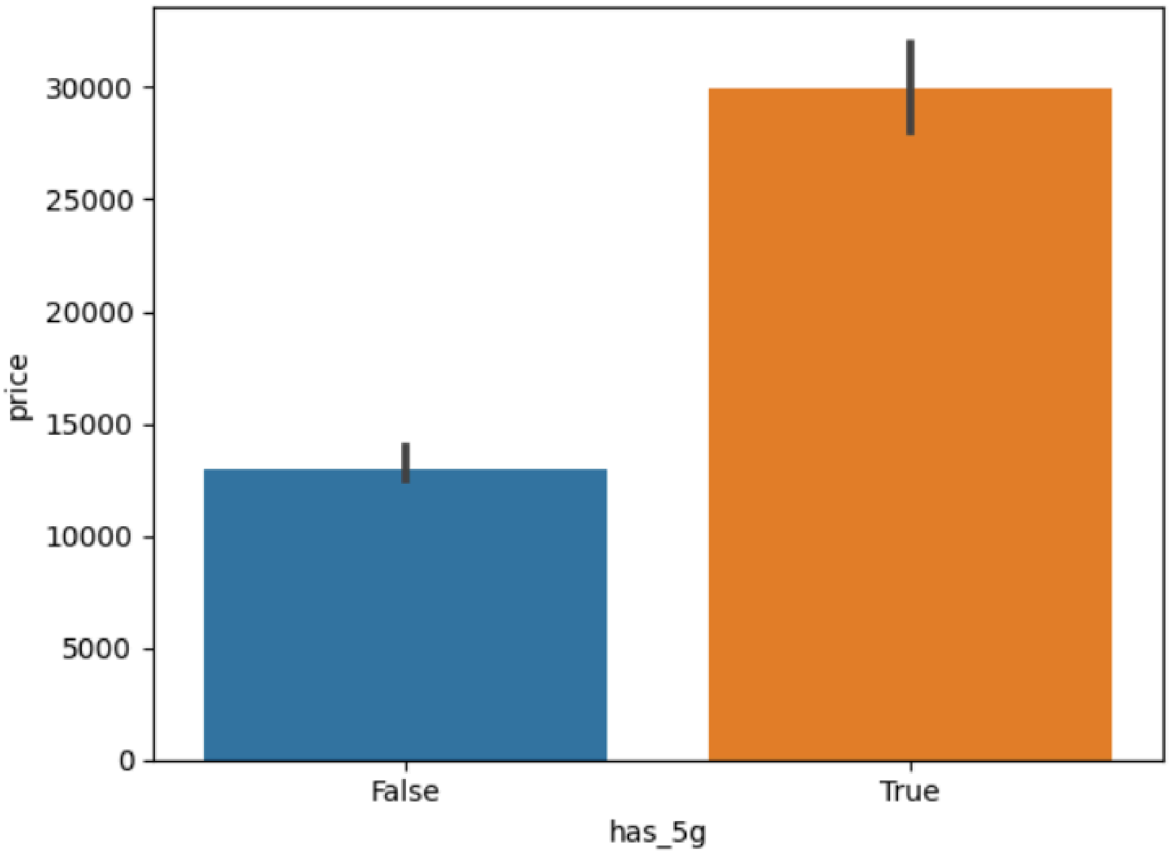
- **Multi-Value Columns**: Separated data such as battery, processor, display, and camera specifications into individual columns for better analysis.

Outliers and Inconsistent Data:

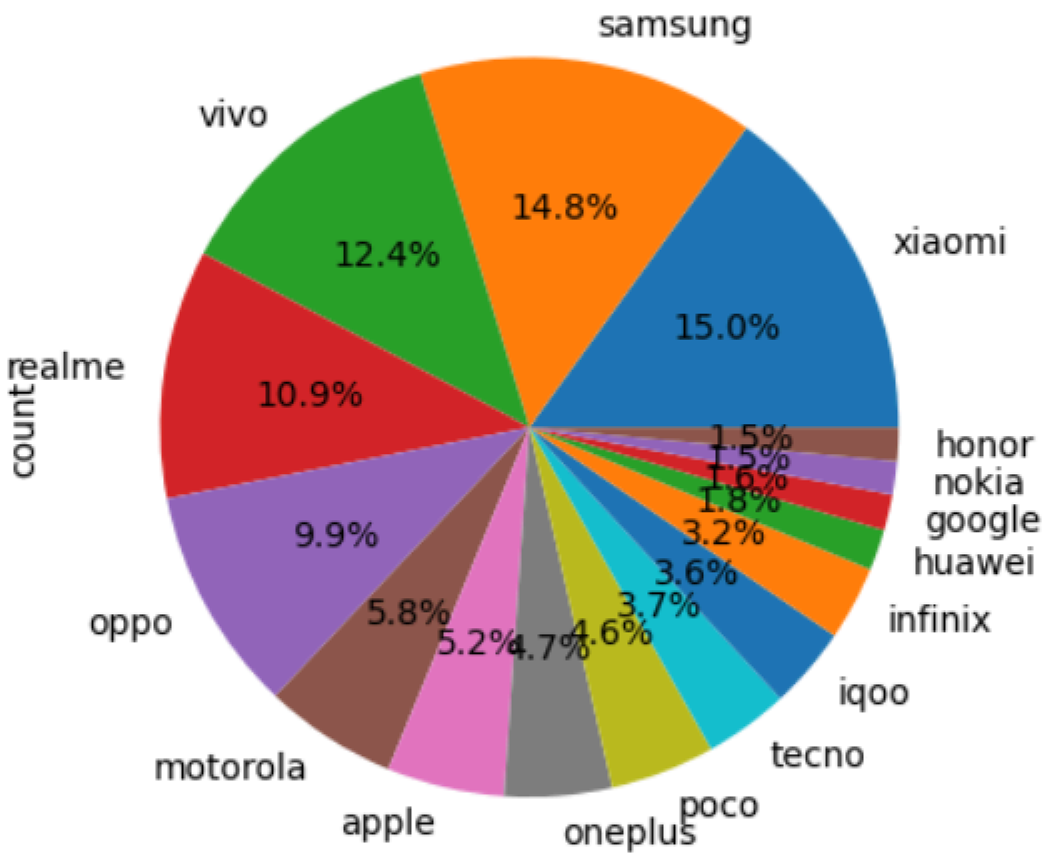
- Outliers like phones made of gold were removed.
- Misplaced data (e.g., battery details in the wrong column) were shifted into the correct columns.

Exploratory Data Analysis (EDA):

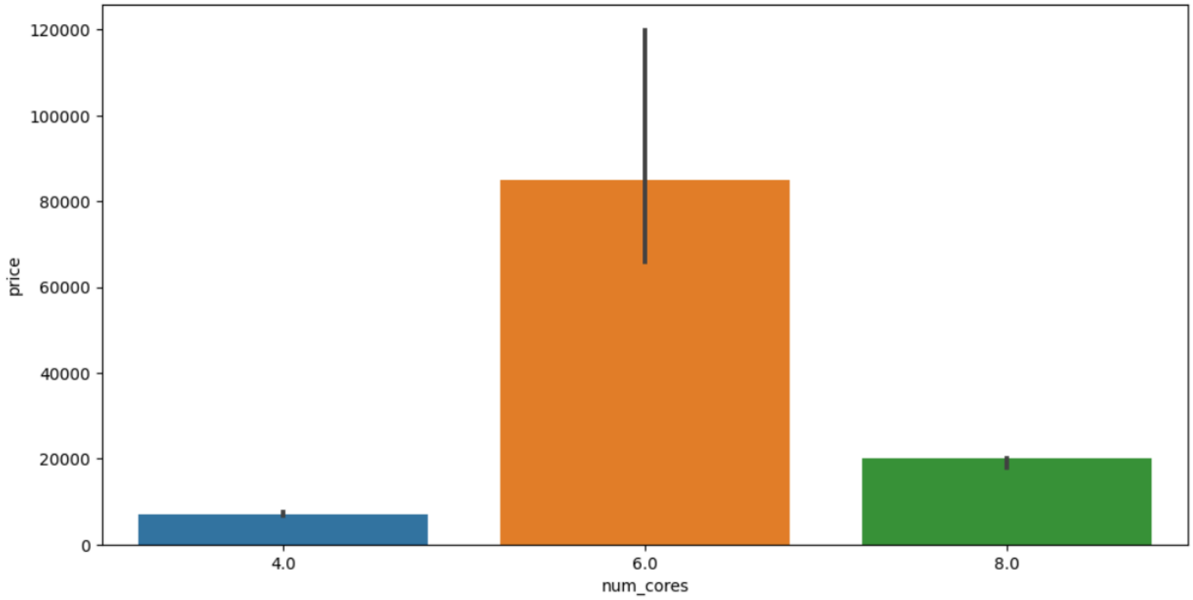
- **Key Insights**:
    - **Price and 5G**: Smartphones with 5G are priced 130% higher than those without 5G. Over 56% of smartphones in the dataset support 5G.



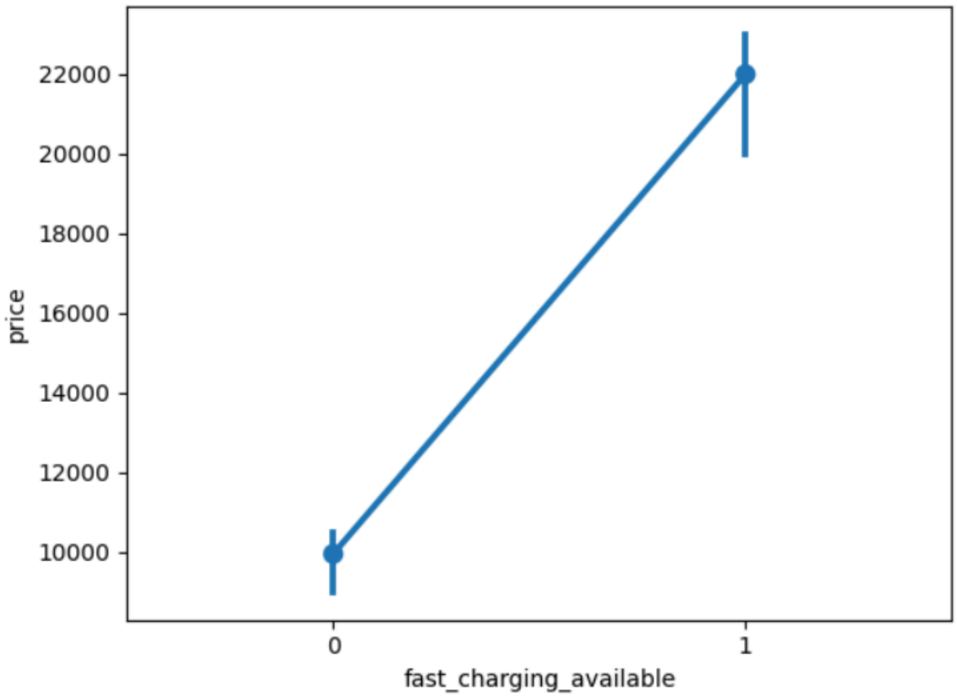
    - **Brand Market Share**: **Xiaomi** and **Samsung** dominate the market, accounting for nearly 30% of the smartphone models. 75% of brands offer more 5G models than non-5G models.



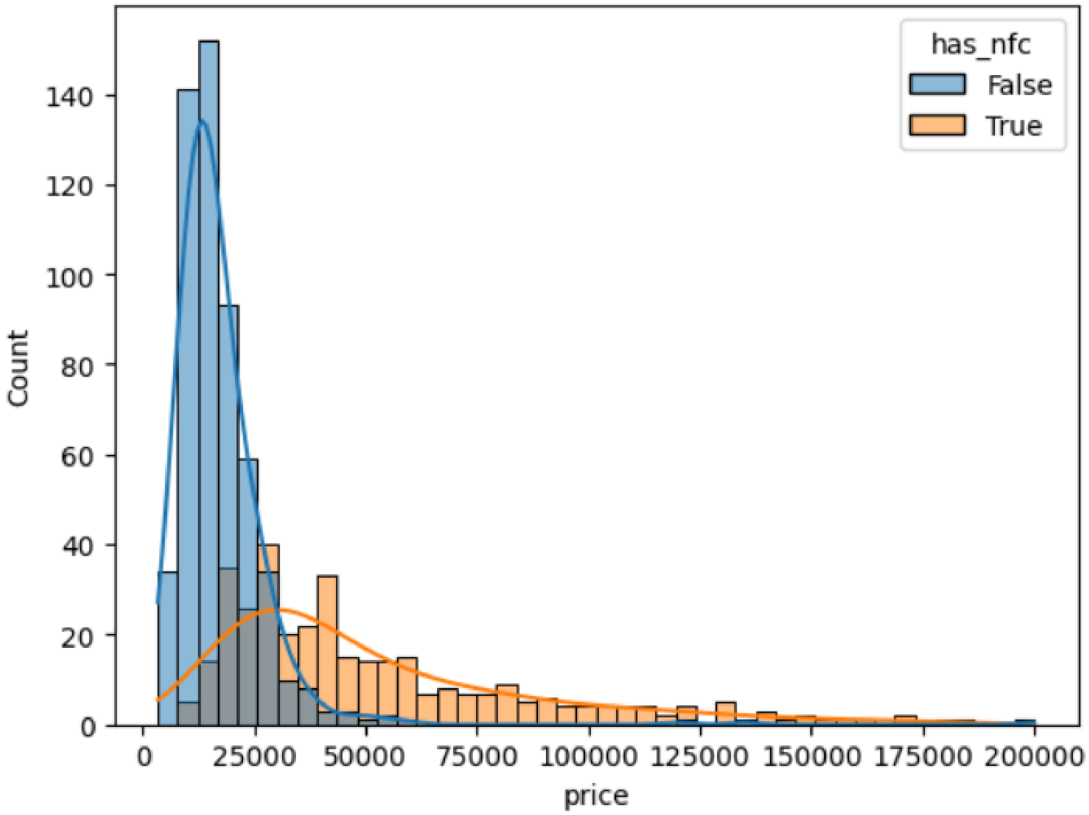
- ○ **Processor and Price**: **Apple's hexa-core processors**, used in 95% of iPhones, lead to prices that are 325% higher than those of octa-core Android devices.



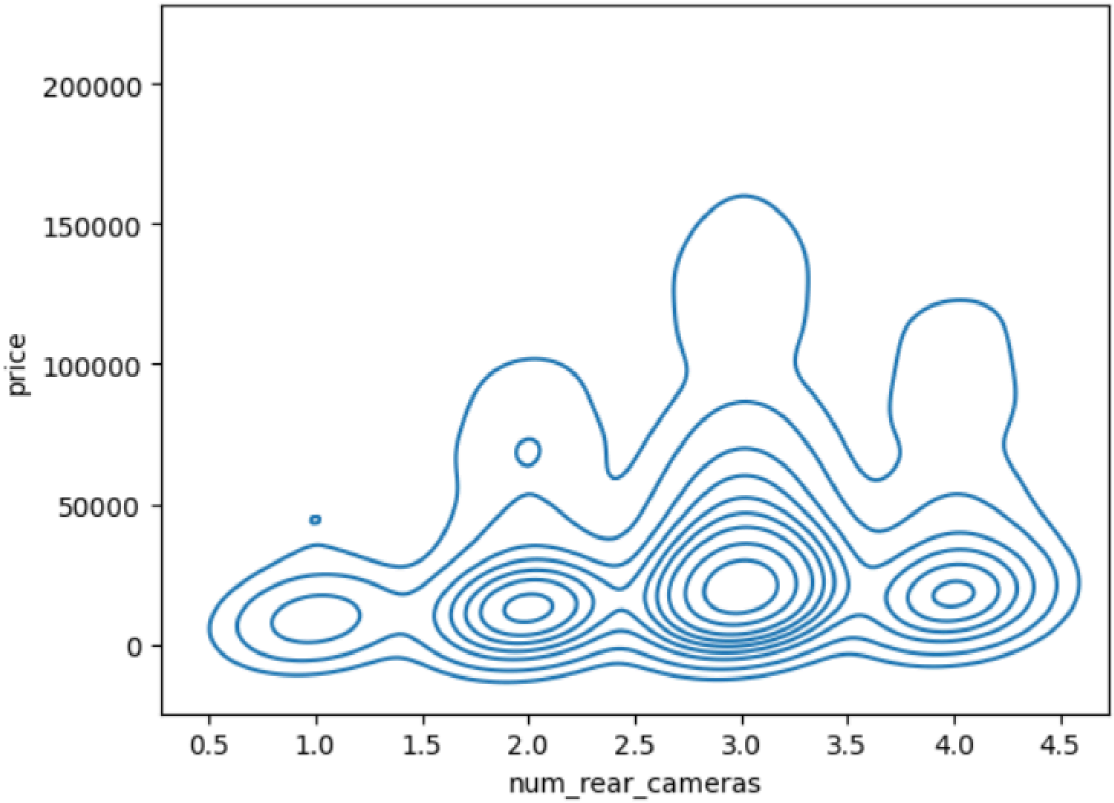
- **Battery Capacity**: Over 50% of smartphones have a 5000mAh battery. Fast-charging phones are priced 121% higher than those without fast charging.



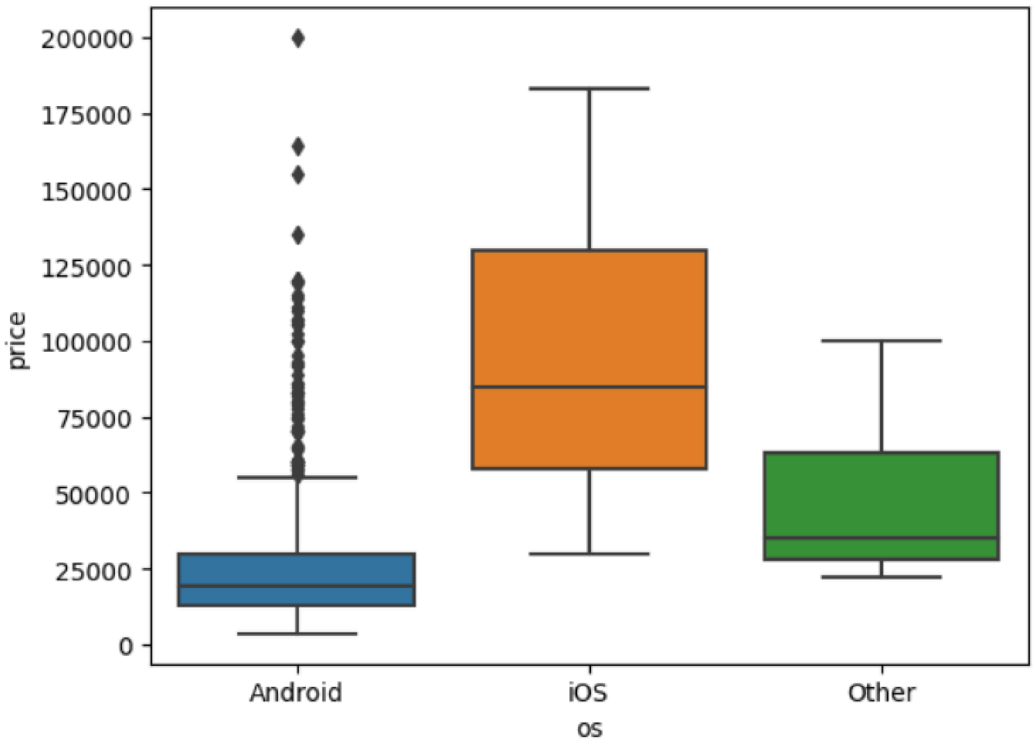
- **NFC and Price**: NFC-enabled phones are 166% more expensive. Nearly all (97%) of Apple models include NFC, compared to only 35% of Android models.



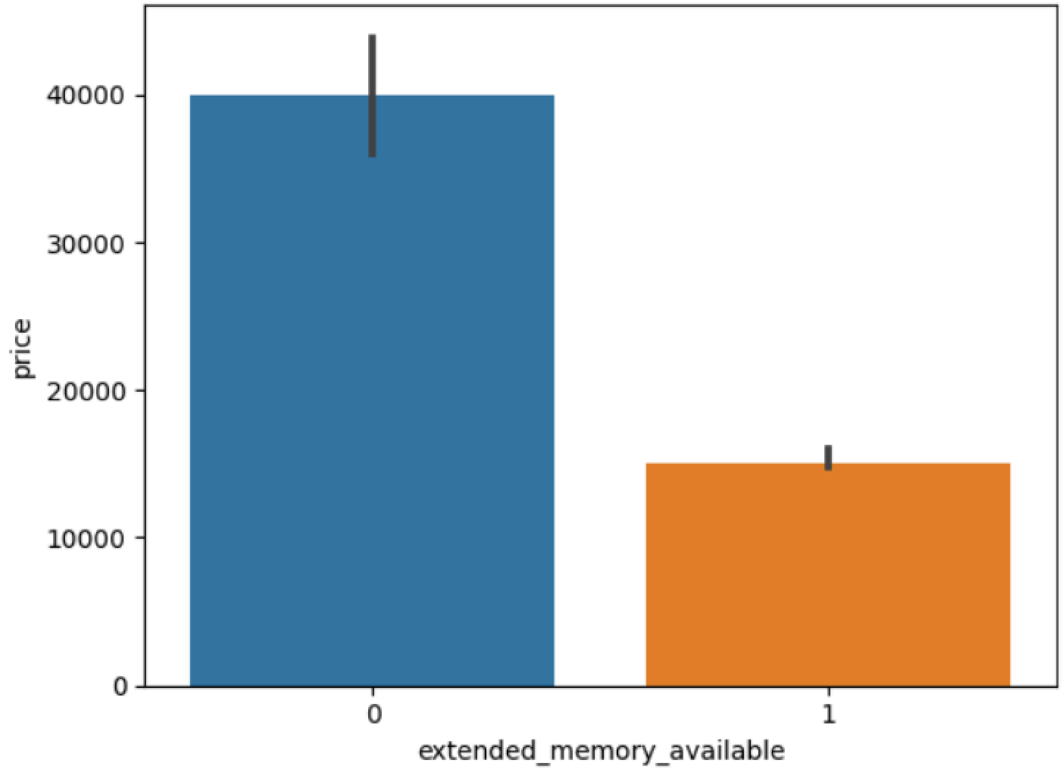
- **Camera and Price**: 57% of smartphones have 3 rear cameras. Prices increase with more cameras, but phones with 4 cameras tend to be slightly cheaper than those with 3 cameras.


price
200000
150000
100000
50000
0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
num_rear_cameras

- **Operating System**: **Android** powers 93.9% of smartphones, while **iOS** makes up 5.2%. iOS phones are priced 347% higher on average.



- **Memory**: 63.8% of smartphones offer expandable memory, but phones without this option are 166% more expensive.



Statistical Analysis (Explained Simply):

Kendall's Tau: Measures the relationship between the number of cores and price.

Spearman's Rho: Analyzes the relationship between the presence of an IR blaster and price, and the presence of NFC and price.

Point-Biserial Correlation: Examines the relationship between the presence of 5G and price, and the presence of NFC and price.

Shapiro-Wilk Test: Tests normality for variables: NFC, IR blaster, price, rating, and 5G.

Kruskal-Wallis Test: Assesses relationships between categorical and continuous variables, such as:

- Brand name and processor speed
- Screen size and price
- Fast charging and price
- Battery capacity and price
- Refresh rate and price
- Brand name and refresh rate
- Resolution and price

- Number of rear cameras and price
- Number of front cameras and price
- Primary rear camera and price
- Processor brand and price
- Rating and operating system (OS)
- Rating and brand name

Dunn's Test: Post-hoc test for multiple comparisons, such as:

- Number of rear cameras
- Number of front cameras
- Types of primary rear cameras
- Types of primary front cameras
- Memory expansion capacity and price
- Price and processor brand
- Screen size and price
- Price and RAM capacity

Bootstrapping: Used for estimating confidence intervals, applied to internal memory and price.

Cramer's V: Measures association between categorical variables, including:

- Presence of NFC and brand name
- Internal memory and brand name
- Resolution and brand name
- Number of rear cameras and brand name
- Primary rear cameras and brand name
- Number of cores and brand name
- Fast charging and brand name

Chi-Square Test: Assesses relationships between categorical variables, such as:

- Number of rear cameras and brand name
- Primary rear cameras and brand name
- Primary front cameras and brand name
- Memory expansion capacities and brand name
- Processor brand and brand name
- Presence of 5G and brand name
- Presence of 5G and OS
- NFC and brand name
- Number of cores and brand name
- Processor speed and OS
- Screen size and brand name
- Fast charging and brand name
- RAM capacity and brand name
- Internal memory and brand name
- Battery capacity and brand name
- Resolution and brand name

Challenges and Solutions:

- **Missing Values**: Used advanced imputation techniques like **KNNImputer** and **SimpleImputer** to handle missing data.
- **Inconsistent Data**: Standardized and corrected data fields to ensure accuracy.
- **Impact of Data Cleaning**: Improved the dataset's reliability, accuracy, and consistency, making it ready for further analysis.

Final Dataset:
The final cleaned dataset was structured, corrected for errors, and had reduced inconsistencies, making it suitable for advanced analysis and modeling.

For more details, please refer to the document **"Smartprix Smartphone Data Analysis – Web Scraping, Cleaning, Code, and Insights,"** along with Python code files on my GitHub: https://github.com/lajhwanthi/Smartprix-Smartphone-analysis.git