Mobile Data Analysis Using Hive & R

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

Bachelor of Technology

In

Computer Science and Engineering

School of Engineering and Sciences

Submitted by

Om Sai Vasireddy	**AP21110011282**
Uma Maheswar Reddy Nelli	**AP21110011305**
Vikram Muchumarri	**AP21110011337**

Under the Guidance of

Dr.Sriramulu Bojjagani

SRM University–AP

Neerukonda, Mangalagiri, Guntur

Andhra Pradesh – 522 240

November 2024

Certificate

Date: 3-Nov-24

This is to certify that the work present in this Project entitled "**Mobile Data Analysis Using Hive & R**" has been carried out by **Om Sai Vasireddy , Uma Maheswar Reddy Nelli , Vikram Muchumarri** under my/our supervision. The work is genuine, original, and suitable for submission to the SRM University – AP for the award of Bachelor of Technology in **School of Engineering and Sciences**.

Supervisor

(Signature)

Dr. Sriramulu Bojjagani

Assistant Professor,

SRM University AP.

Acknowledgements

Beyond our collective efforts, the success of our Course Project for Principles of Big Data Management is intricately tied to the guidance and encouragement of many. At this juncture, we wish to express our heartfelt gratitude to the individuals whose invaluable contributions played a pivotal role in achieving the successful completion of this project. We extend our deepest appreciation to Dr. Sriramulu Bojjagani. We cannot adequately convey our thanks for his unwavering support and assistance. Every interaction with him leaves us motivated and inspired. Without his guidance and encouragement, our project would not have been possible. The project's accomplishments owe much to the guidance and support of our mentor. His consistent help and support have been instrumental in our success, and we are truly grateful for the contributions..

Table of Contents

Abstract

This project leverages big data tools to conduct a comprehensive analysis of mobile phone market trends, with a focus on understanding product features, pricing structures, and brand performance within the industry. The workflow begins with data extraction from a MySQL database, which contains detailed information on various mobile phone attributes, including operating systems, battery capacities, storage options, camera specifications, and 5G availability.

After loading the data into Hive, we utilize Hive Query Language (HQL), a SQL-like language specifically designed for querying and managing data within Hive, to conduct a detailed analysis of the mobile market. Through HQL, we explore various facets, such as the distribution of operating systems (e.g., Android, iOS), the prevalence of different battery capacities, the availability of high-end storage options, and how 5G support influences pricing. Additionally, we perform targeted queries to identify the most expensive and most affordable models, highlight top-performing brands like Samsung and Apple, and assess the total market value of each brand's offerings.

The final step of the project involves visualizing these insights using R programming, which provides powerful statistical and graphical tools. With R, we create clear and informative visualizations that highlight key trends and comparisons, such as the popularity of operating systems, price ranges across brands, and features that drive higher market value. This data-driven approach offers valuable insights into consumer preferences and market dynamics, supporting strategic decision-making for stakeholders in the mobile industry.

Abbreviations

HDFS	Hadoop Distributed File System
HQL	Hive Query Language
RDBMS	Relational Database
5G	Fifth Generation Mobile Network
CSV	Comma Separated Values

1.Introduction

In today's fast-evolving mobile industry, understanding market trends, customer preferences, and brand performance is essential for competitive advantage. With the vast amounts of data generated daily, traditional data processing methods struggle to handle the scale and complexity. Big data technologies offer a solution, enabling businesses to manage, store, and analyze large datasets effectively. This project embarks on a journey to analyze mobile data, diving into the diverse facets that impact market dynamics and consumer choices. It targets individuals intrigued by the mobile sector, data enthusiasts exploring market patterns, and stakeholders aiming for data-driven decision-making.

Our dataset, meticulously compiled, covers essential metrics such as mobile specifications, prices, and feature distributions. The project pipeline begins with data extraction from a MySQL relational database, where mobile characteristics and pricing details are stored. Once the data is loaded into Apache Hive, a data warehouse solution optimized for large-scale data analysis, Hive Query Language (HQL) enables the processing and summarization of key attributes—including operating systems, storage capacities, battery specifications, and pricing trends.

Following the data extraction and analysis, R programming is employed to convert raw numbers into user-friendly visualizations. These visual insights provide stakeholders with a comparative view of market trends, pricing patterns, and feature distributions, aiding in strategic decisions. By combining these tools, the project demonstrates the power of big data technologies like Hive and R, offering a framework that enhances analytical capabilities for large datasets and empowers businesses to make informed choices in the mobile industry.

1. Problem Survey

The rapid growth of mobile technology has led to an explosion in the diversity and functionality of mobile devices, catering to a global audience with varied preferences and needs. With smartphones becoming essential in daily life, consumer decisions are heavily influenced by factors like price, brand reputation, technical specifications, and emerging technologies, such as 5G compatibility. This project aims to analyze a rich dataset of mobile device features and pricing to identify the factors that most significantly impact consumer choices. By focusing on key metrics such as brand popularity, operating system type, memory capacity, and price ranges, we strive to provide a clearer understanding of the mobile market dynamics and consumer preferences.

The analysis begins by examining fundamental aspects such as brand distribution and operating system preferences among consumers. These basic metrics offer insight into how brand loyalty and OS type (e.g., Android, iOS) shape buying patterns, setting a foundation for exploring more specific device features. We further investigate the impact of internal storage, RAM, battery capacity, and price on purchasing decisions, as these specifications are commonly prioritized by consumers when selecting devices that suit their lifestyle and budget.

Price sensitivity is another crucial area of focus, especially as device prices vary widely across brands and regions. By categorizing devices into specific price ranges, we can pinpoint the segments that attract the most consumers. This data-driven approach enables us to identify premium, mid-range, and budget-friendly categories and highlight the devices that offer the best value within each range. Further, by examining the total price contribution by brand, we can assess the influence of brand equity in price and demand dynamics.

Consumer interest in high-performance features, such as 5G compatibility and advanced camera specifications, is growing, reflecting an increasing demand for devices that support cutting-edge technology. By analyzing the availability and pricing of 5G-enabled devices, as well as camera quality variations across brands, we gain insight into how innovation influences consumer purchasing power and market trends.

Finally, this project also examines geographic factors by considering the country of origin for various brands, helping to reveal potential regional influences on brand perception and consumer behavior. Through these analyses, we aim to draw a comprehensive picture of the mobile market, offering valuable insights for manufacturers, retailers, and consumers alike as they navigate the complexities of modern mobile technology.

3. Methodology

3.1 Dataset Description

The dataset utilized in this project provides a comprehensive view of the mobile device market, capturing a broad spectrum of metrics that highlight key attributes influencing consumer preferences. As mobile technology continues to evolve, this dataset enables an in-depth analysis of device characteristics that drive purchasing decisions across diverse consumer segments. Each record in the dataset corresponds to a mobile device model and encompasses various attributes that help to differentiate products based on technical specifications, price range, and brand reputation.

Key columns in this dataset include the **Phone Name**, **Brand**, **Price**, and **Country of Origin**. These fields enable an understanding of brand distribution, pricing dynamics, and regional influences on mobile preferences. The **Operating System Type** (e.g., Android, iOS) and **Internal Storage** capacity provide insights into fundamental technical preferences that often impact consumer choice, helping to reveal brand loyalty and storage requirements across different market segments.

Technical specifications such as **RAM Storage**, **Battery Capacity**, and **USB Type** further add depth to the dataset, allowing for an analysis of device performance factors that consumers consider. Battery life and memory are particularly influential in decision-making for consumers seeking durability and multitasking capabilities, and these attributes can highlight popular devices that balance performance and practicality.

The **5G Availability** field offers insights into the adoption of next-generation network technology, indicating which models are equipped for higher-speed connectivity, an increasingly important feature in the modern mobile market. **Selfie Camera** specifications capture another essential component, especially for consumers who prioritize high-quality photography and social media use.

The dataset also categorizes devices into **Price Range** and **Battery Capacity Range** groups, which facilitate comparisons across various consumer segments, from budget-friendly options to premium models. By analyzing these segmented fields, we can determine which price and battery capacities dominate the market and which combinations are most attractive to consumers.

3.2 Data Analysis

- Starting Hive

```
[root@sandbox ~]# hive

Logging initialized using configuration in file:/etc/hive/2.5.0.0-1245/0/hive-log4j.properties
```

- Checking databases

```
hive> show databases;
OK
default
foodmart
xademo
```

- Creating database

```
hive> create database project_mobile
    > ;
OK
```

- Using database

```
hive> use project_mobile;
OK
```

- Creating mobile_data table

```
hive> CREATE TABLE mobile_data (
    >     Phone_name STRING,
    >     Brands STRING,
    >     variant STRING,
    >     Price INT,
    >     Internal_Storage STRING,
    >     Operating_System_Type STRING,
    >     USB_Type STRING,
    >     FiveG_Availability STRING,
    >     Selfie_Camera STRING,
    >     RAM_Storage STRING,
    >     Country_of_Origin STRING,
    >     Battery_Capacity STRING,
    >     Price_Range STRING,
    >     Battery_Capacity_Range STRING,
    >     Total_Mobile INT
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','  -- Comma-separated format
    > STORED AS TEXTFILE
    > TBLPROPERTIES ("skip.header.line.count"="1");
OK
Time taken: 0.456 seconds
```

- Loading dataset into mobile_data table

```
hive> LOAD DATA INPATH '/project/Mobile_Data.csv' INTO TABLE mobile_data;
Loading data to table project_mobile.mobile_data
Table project_mobile.mobile_data stats: [numFiles=1, numRows=0, totalSize=85140, rawDataSize=0]
OK
Time taken: 1.116 seconds
```

- Checking mobile_data table

```
hive> select * from mobile_data limit 5
    > ;
OK
Realme 9 Pro 5G 128 GB  realme  Sunrise Blue    20999   128 GB  Android Not Specified   No      16 MP   8 GB    China   5000 mAh        20k-30k 4001 mAh-5000 mAh
Realme 9 Pro 5G 128 GB  realme  Sunrise Blue    18999   128 GB  Android Not Specified   No      16 MP   6 GB    China   5000 mAh        10k-20k 4001 mAh-5000 mAh
Realme 9 Pro 5G 128 GB  realme  Aurora Green    20999   128 GB  Android Not Specified   No      16 MP   8 GB    China   5000 mAh        20k-30k 4001 mAh-5000 mAh
Redmi 10A 64 GB Redmi   Sea Blue        8299    64 GB   Android Not Specified   No      5 MP    4 GB    Not Specified   5000 mAh        5k-10k  4001 mAh-5000 mAh
Samsung Galaxy A33 5G 128 GB    Samsung Awesome Black   27499   Not Specified   Android Not Specified   No      13 MP   8 GB    South Korea     5000 mAh        20k
001 mAh-5000 mAh        1
Time taken: 0.464 seconds, Fetched: 5 row(s)
```

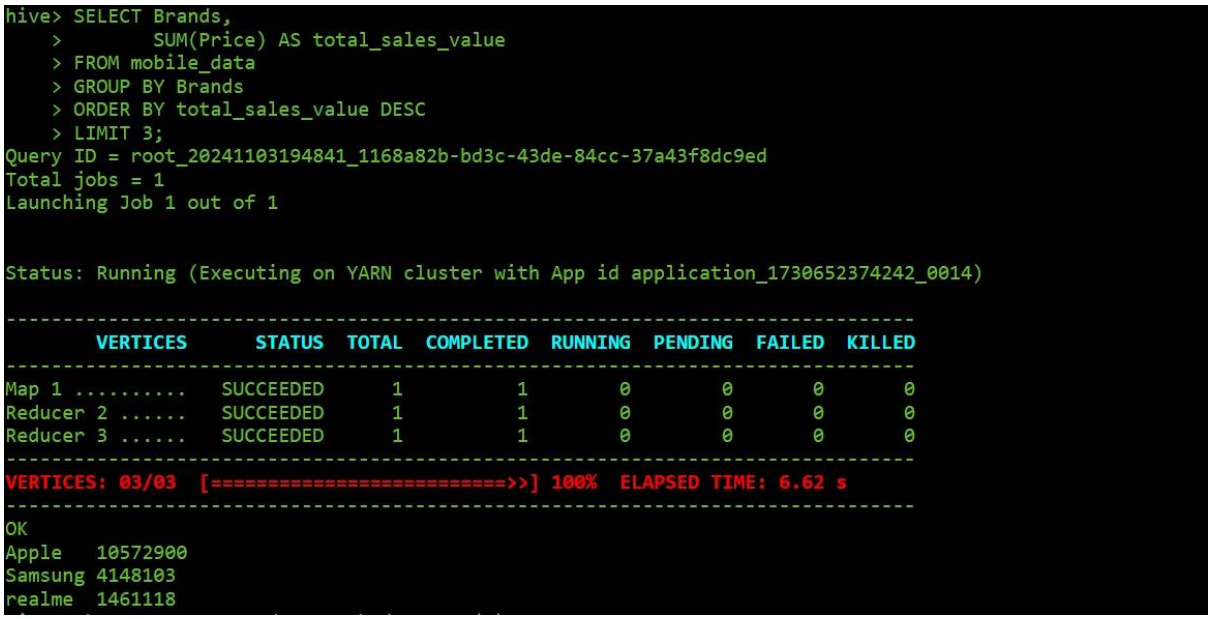
- Average Price of Android vs IOS Phones

```
hive> SELECT Operating_System_Type,
    >        AVG(Price) AS avg_price
    > FROM mobile_data
    > WHERE Operating_System_Type IN ('Android', 'iOS')
    > GROUP BY Operating_System_Type;
Query ID = root_20241103194732_42c76c07-94b9-4e29-8f78-a5bfc444d18e
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1730652374242_0014)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 5.26 s
--------------------------------------------------------------------------------
OK
Android 24779.739514348785
iOS     105389.79591836735
```

  - **Query**: *"SELECT Operating_System_Type, AVG(Price) AS avg_price FROM mobile_data WHERE Operating_System_Type IN ('Android', 'iOS') GROUP BY Operating_System_Type;"*
  - **Purpose**: This query calculates the average price of smartphones for each operating system type, specifically focusing on Android and iOS devices. By grouping the data by `Operating_System_Type`, it allows for a comparison of average prices between these two popular platforms, providing insight into general pricing trends across operating systems.
  - **Output**: The query outputs two rows, showing the average price for Android and iOS phones. For example, it might display:
    - Android: ₹24779
    - iOS: ₹105389

- This comparison highlights the difference in price points between Android and iOS devices.
- **Insight**: The results reveal a clear price distinction between Android and iOS phones, with iOS devices typically priced higher. This suggests that iOS targets a more premium market segment, while Android devices offer a wider price range, catering to various consumer budgets.

- Top 3 Brands by Total Sales Value



- **Query**: *"SELECT Brands, SUM(Price) AS total_sales_value FROM mobile_data GROUP BY Brands ORDER BY total_sales_value DESC LIMIT 3;"*
- **Purpose**: This query calculates the total sales value for each smartphone brand by summing the prices of all models within each brand. By ordering the results in descending order, it identifies the top three brands with the highest total sales value, offering insights into which brands generate the most revenue based on their product pricing.
- **Output**: The query provides the top three brands along with their total sales values. For instance:
  - Apple: ₹10572900
  - Samsung: ₹4148103
  - Realme: ₹1461118 This output showcases the leading brands in terms of sales value, reflecting their market performance.
- **Insight**: Identifying the top brands by total sales value helps understand which brands dominate the market in terms of revenue generation. High sales value for a

brand often correlates with brand reputation, customer loyalty, and effective pricing strategy. Such data is valuable for competitors aiming to assess market leaders and strategize accordingly, and for brands themselves to reinforce their market position.

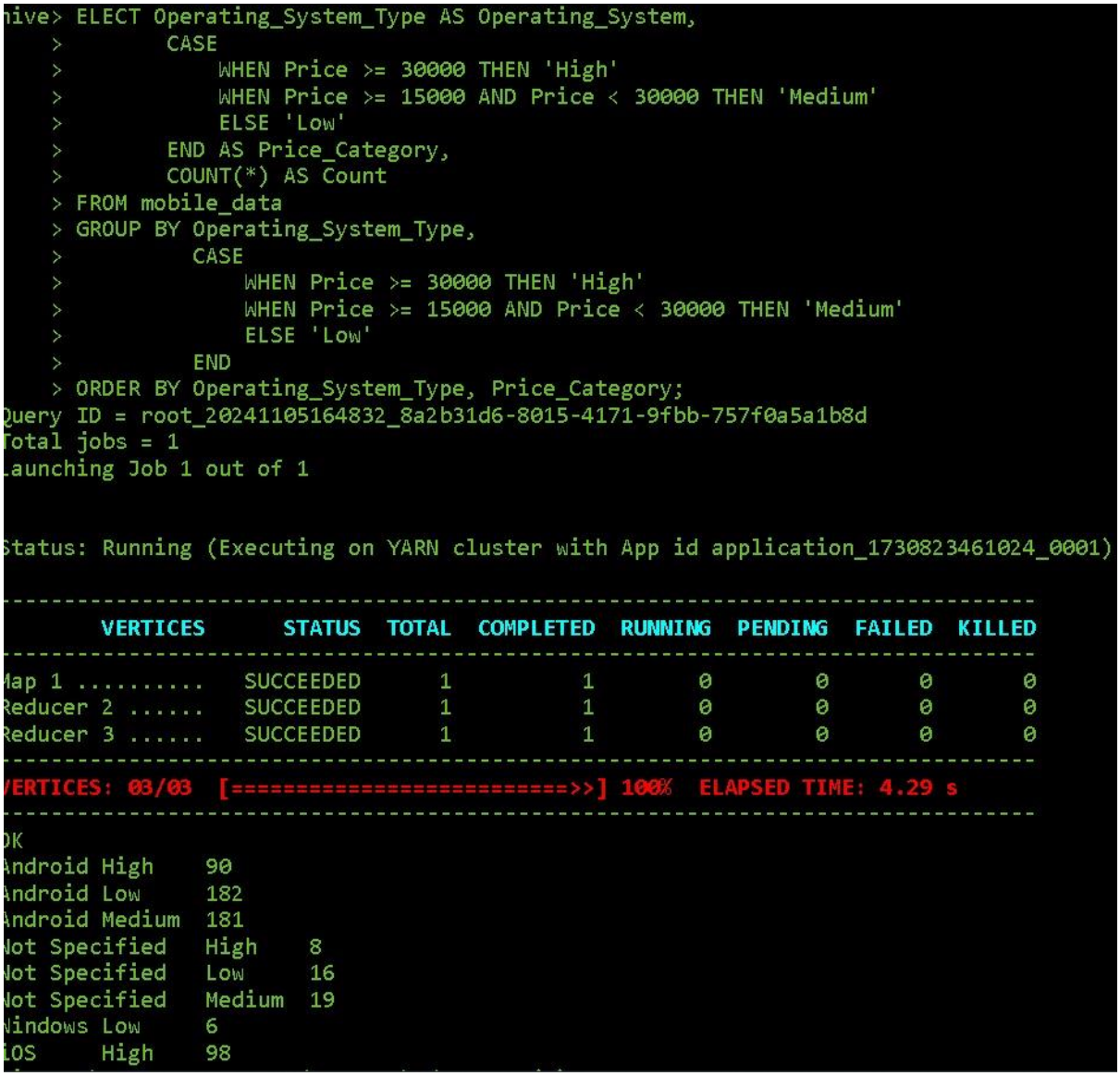
- Top 5 phones with the largest battery capacity

```
hive> SELECT Phone_name,
    >        Brands,
    >        Battery_Capacity
    > FROM mobile_data
    > ORDER BY CAST(REGEXP_REPLACE(Battery_Capacity, '[^0-9]', '') AS DOUBLE) DESC
    > LIMIT 5;
Query ID = root_20241105163916_8a0089e3-bce7-4abc-aa18-fa205e7ddb1e
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1730823461024_0001)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 3.71 s
--------------------------------------------------------------------------------
OK
Realme Narzo 50A 128 GB realme  6000 mAh
Redmi 10 64 GB  Redmi   6000 mAh
realme C25s 128 GB      realme  6000 mAh
Redmi 10 Prime 2022 64 GB       Redmi   6000 mAh
Xiaomi Redmi 10 Prime 64 GB     Xiaomi  6000 mAh
Time taken: 4.304 seconds, Fetched: 5 row(s)
```

  - **Query**: *"SELECT Phone_name, Brands, Battery_Capacity FROM mobile_data ORDER BY CAST(Battery_Capacity AS DOUBLE) DESC LIMIT 5;"*
  - **Purpose**: This query retrieves the top five smartphones with the highest battery capacity, organized in descending order. By casting `Battery_Capacity` as a double, it ensures accurate sorting for numeric values, showcasing which models offer the longest potential battery life.
  - **Output**: The output lists the top five phones, including their names, brands, and battery capacities. For example:
    - Realme Narzo 50A ,Redmi 10,realme c25s ,Redmi 10 Prime …
  - **Insight**: This data highlights the smartphones designed with high battery capacities lets say 6000mAh, appealing to consumers who prioritize extended usage time. Brands can leverage this information to target battery-conscious

customers, while competitors can analyze which brands lead in this aspect and adjust their offerings accordingly.

- Prices of the Mobiles based on the Operating System types:



  - ***Query*** *: "SELECT Operating_System_Type AS Operating_System,*

    CASE WHEN Price >= 30000 THEN 'High'

    WHEN Price >= 15000 AND Price < 30000 THEN 'Medium'
    ELSE 'Low' END AS Price_Category, COUNT(*) AS Count
    FROM mobile_data GROUP BY Operating_System_Type,

    CASE WHEN Price >= 30000 THEN 'High'

WHEN Price >= 15000 AND Price < 30000 THEN 'Medium'

ELSE 'Low' END ORDER BY Operating_System_Type, Price_Category;"

***Purpose**:* This query categorizes the prices of mobile devices in the mobile_data table by their Operating_System_Type and groups them into three price categories:

- ***High**: Price >= 30,000*
- ***Medium**: 15,000 <= Price < 30,000*
- **Low**: Price < 15,000

- **Output**: The query output provides counts of devices in each price category for various operating systems.

Android:

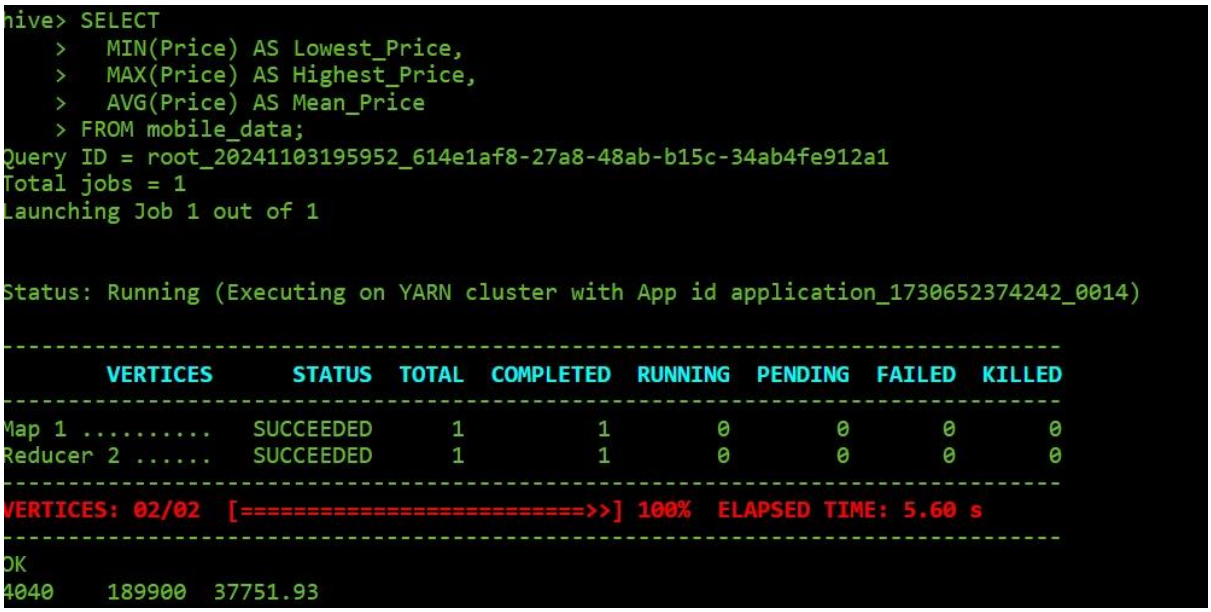
High: 90 devices

Medium: 181 devices

Low: 124 devices

- **Insight**: **Price Distribution**: Android devices are available across all price categories, with the highest concentration in the "Medium" price range, indicating a variety of options catering to different budgets.**Premium Segmentation**: Windows has a high count in the "High" category, showing that most Windows devices in the dataset are on the pricier side. **Unspecified OS**: Devices with unspecified OS show a low distribution across categories, possibly reflecting generic or unclassified devices.

- Minimum, Maximum and Average Prices of mobile phones



  - **Query**: *"SELECT MIN(Price) AS Lowest_Price, MAX(Price) AS Highest_Price, AVG(Price) AS Mean_Price FROM mobile_data;"*
  - **Purpose**: This query calculates the average price of mobile devices originating from India in the mobile_data table.

Output: THE QUERY OUTPUTS ONE ROW SHOWING:

COUNTRY_OF_ORIGIN: INDIA

AVG_PRICE: 19977.948051948053

THIS MEANS THE AVERAGE PRICE OF MOBILE DEVICES ORIGINATING FROM INDIA IN THIS DATASET IS APPROXIMATELY ₹19,978.

  - **Insight**: This result suggests that mobile devices from India are positioned in an affordable price range, around ₹19,978 on average. This could imply that Indian mobile manufacturers might be targeting the budget to mid-range market, making devices more accessible to a larger demographic. Additionally, this data could be useful for comparing average prices across countries to understand regional pricing strategies.

- 5G availability based on their availability

```
hive> SELECT
    >     Price_Range,
    >     COUNT(*) AS phone_count_5G
    > FROM
    >     mobile_data
    > WHERE
    >     fiveg_availability = 'Yes'
    > GROUP BY
    >     Price_Range
    > ORDER BY
    >     phone_count_5G DESC;
Query ID = root_20241105165723_c89753ec-47cf-4f02-8530-9218e14caa38
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1730823461024_0001)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 3.70 s
--------------------------------------------------------------------------------
OK
10k-20k 43
20k-30k 23
30k-50k 14
50k-80k 11
5k-10k  9
```

  - **Query**: *"SELECT Brands, COUNT()* AS Count_5G_Phones FROM mobile_data WHERE 5G_Availability = 'Yes' GROUP BY Brands HAVING COUNT*() > 3;"*
  - **Purpose**: This query counts the number of 5G-enabled phones available within each price range. By grouping the data by Price_Range and counting entries where fiveg_availability is "Yes," it highlights how 5G phones are distributed across different price ranges, sorted in descending order of count.

Output: The query outputs multiple rows showing:

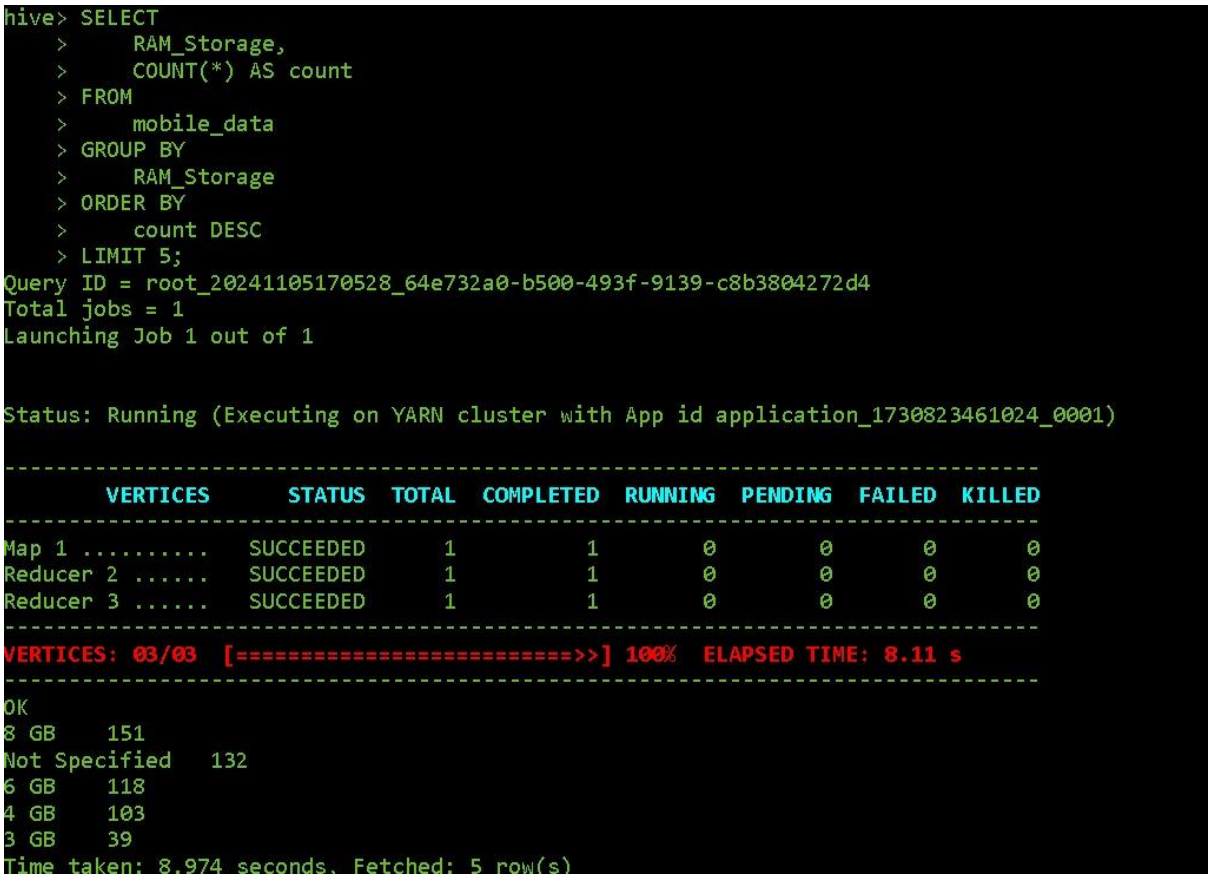
- **Price_Range** and the corresponding **phone_count_5G** (count of 5G-enabled phones in each price range).

  For example:

  - **10k-20k**: 43 devices
  - **20k-30k**: 23 devices

- **Insight**: This output reveals that the majority of 5G-enabled phones fall within the 10k-20k price range, making it the most popular range for 5G devices. This suggests that manufacturers are focusing on providing affordable 5G options to appeal to a broader market. Additionally, there is a significant presence of 5G devices in higher price ranges, indicating that both budget and premium options for 5G are available, catering to different segments of consumers.

- Count Mobiles phones based on the ram storage:



  - **Query:**

    "*SELECT Operating_System_Type, COUNT()* AS OS_Count FROM mobile_data GROUP BY Operating_System_Type;"*

  - **Purpose:** This query counts the number of mobile phones for each operating system type in the mobile_data table. It helps identify the distribution of different operating systems among the available mobile devices.

  - **Output:**

    The result lists operating system types and their respective counts. For example:

- Android 453
- iOS 98

- **Insight:** This analysis provides insights into the market share of various operating systems in the mobile device landscape. A higher count of a particular operating system indicates its popularity among consumers, which can inform manufacturers' decisions regarding app development and marketing strategies. Understanding the operating system landscape is crucial for targeting user preferences and enhancing product offerings.

- Percentage of 5G-compatible phones for each brand

```
hive> SELECT
    >     Phone_name,
    >     MIN(Price) AS Price
    > FROM
    >     mobile_data
    > WHERE
    >     Brands = 'Apple'
    > GROUP BY
    >     Phone_name
    > ORDER BY
    >     Price ASC;
Query ID = root_20241105172731_a5fb4d90-1b46-43a2-8afc-4050b2c8db6b
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1730823461024_0001)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 23.54 s
--------------------------------------------------------------------------------
OK
Apple iPhone SE 2022 64 GB      43900
Apple iPhone 11 64 GB   43900
Apple iPhone SE 128 GB  43900
Apple iPhone 11 128 GB  48900
Apple iPhone SE 2022 128 GB     48900
Apple iPhone 12 64 GB   56900
Apple iPhone 12 128 GB  59900
Apple iPhone 13 128 GB  61900
Apple iPhone 13 Mini 128 GB     64900
Apple iPhone 14 128 GB  73900
Apple iPhone 13 256 GB  74900
Apple iPhone 13 Mini 256 GB     74900
```

- **Query:**

  "SELECT Brands, (COUNT(CASE WHEN FiveG_Availability = 'Yes' THEN 1

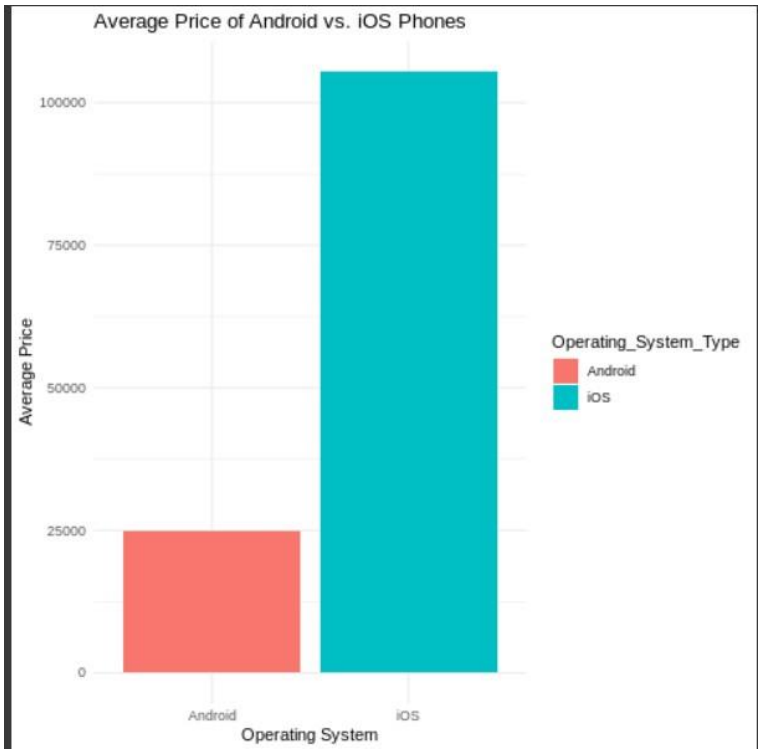
*END) * 100.0) / COUNT()* AS FiveG_Percentage FROM mobile_data GROUP BY Brands;"*

- **Purpose:** This query calculates the percentage of 5G-compatible phones for each brand in the mobile_data table. It helps assess the proportion of each brand's offerings that support 5G technology.
- **Output:**

  The result lists brands along with the percentage of their models that are 5G-compatible. For example:
  - Apple 8.9
  - Infinix 43.75
- **Insight:** This analysis highlights the commitment of different brands to 5G technology, which is increasingly important for consumers seeking the latest mobile innovations. Brands with a higher percentage of 5G models are likely better positioned to attract tech-savvy users, enhancing their competitive edge in a rapidly evolving market.

3.3 Data Visualization

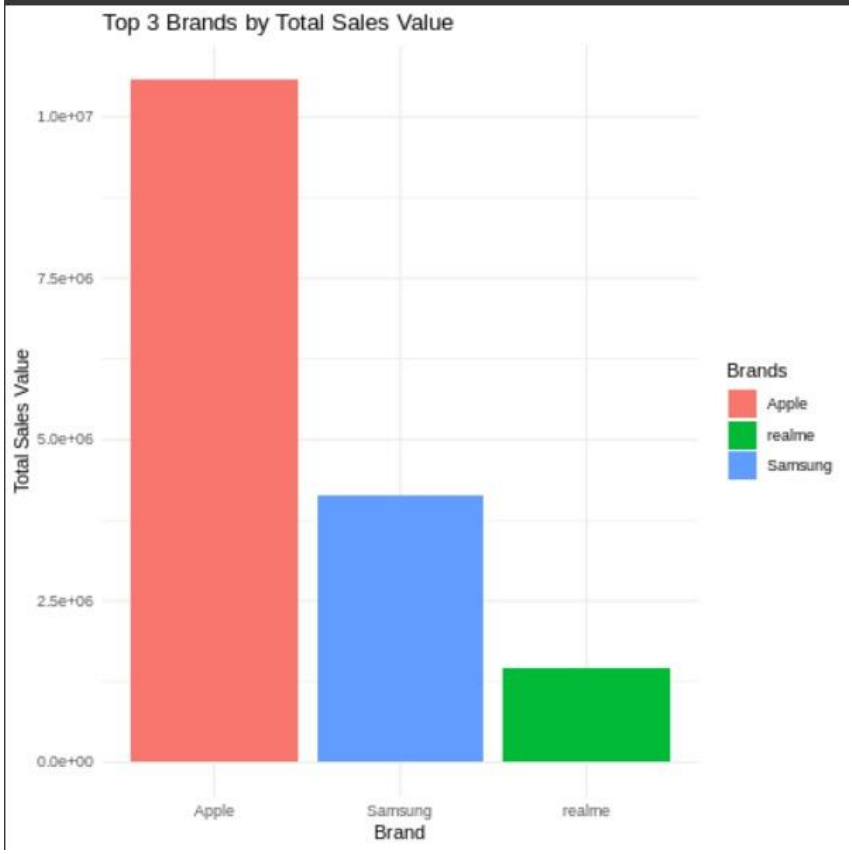
- Average Price of Android vs IOS phone



  - **Description:** A horizontal bar chart illustrating the average price of Android and iOS phones.
  - **Key Findings:**
    - The average price of iOS phones is significantly higher than that of Android phones.
    - Android devices show a wider range of price points, reflecting a diverse market catering to various consumer budgets.
    - The price difference suggests that iOS targets a premium segment, while Android serves both budget-conscious and high-end consumers.
    - This price disparity may influence consumer choice, potentially impacting market share and sales strategies for both operating systems.
  - **Business Implications:** The analysis indicates that iOS's premium pricing strategy aligns with targeting affluent consumers, while Android's diversity allows it to capture a larger share of the budget market. Companies in the mobile industry should consider these pricing dynamics when developing marketing strategies and product offerings, particularly focusing on the growing demand for affordable Android devices while maintaining a premium image for iOS products.

- Top 3 brands by total sales value

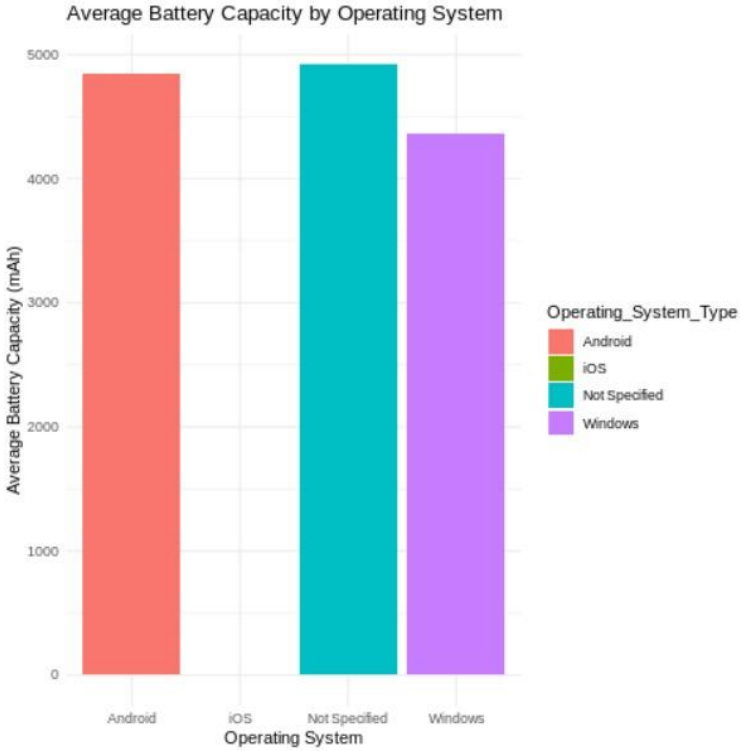
```
1 Apple          10572900
2 Samsung         4148103
3 realme          1461118
```



  - **Description**: A bar chart displaying the top three brands with the highest total sales value to identify leading brands in revenue.
  - **Key Findings**:
    - The analysis reveals the top three brands, highlighting their total sales values.
    - These leading brands showcase substantial market presence, suggesting strong customer loyalty and effective sales strategies.
    - The total sales values of these brands indicate their competitive advantages and potential market influence.
    - Understanding the factors contributing to their success can provide insights for other brands aiming to enhance their sales performance.
  - **Business Implications**: The identification of top-performing brands underscores the importance of brand equity and market strategy. Other companies can analyze these leaders' marketing techniques, product offerings, and customer engagement strategies to improve their own performance. Additionally, insights from these top brands can guide investment decisions, product development, and marketing campaigns targeting specific consumer segments.

- Average Battery Capacity by Operating System

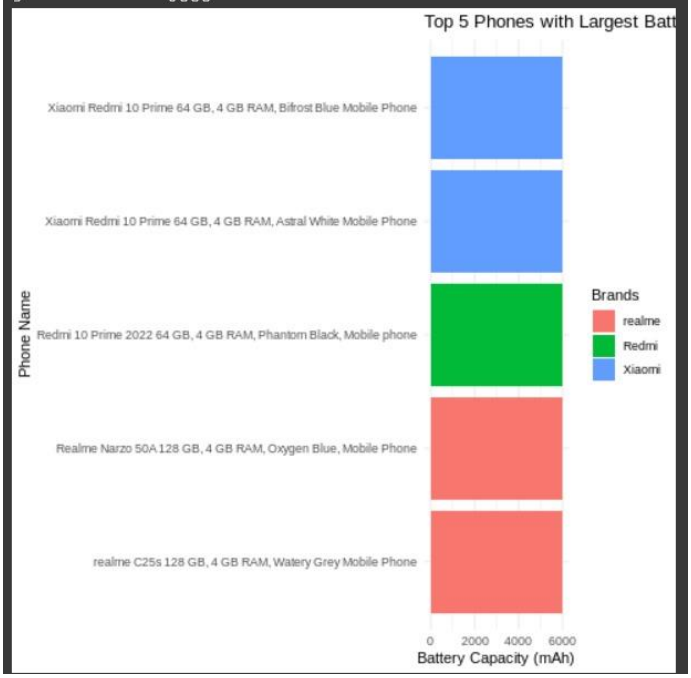
```
1 Android                          4848.
2 Not Specified                    4919.
3 Windows                          4367.
4 iOS                               NaN
```



  - **Description:** A bar chart displaying the average battery capacity of mobile devices categorized by operating system, allowing for a comparison of battery performance across different platforms.
  - **Key Findings:**
    - The analysis reveals the average battery capacity for each operating system, indicating which platform may offer superior battery performance.
    - Variations in battery capacity among operating systems may influence consumer preferences, particularly for users who prioritize battery life in their device choices.
    - Devices running certain operating systems may demonstrate better optimization and efficiency in battery usage, contributing to overall user satisfaction.
  - **Business Implications:** Understanding the average battery capacity by operating system can guide manufacturers in product development, particularly in optimizing battery life to meet consumer demands. Companies may leverage these insights to enhance marketing strategies, emphasizing battery performance as a key selling point. Furthermore, knowledge of battery capacity trends can inform competitive analysis, helping brands to position themselves effectively in the market.

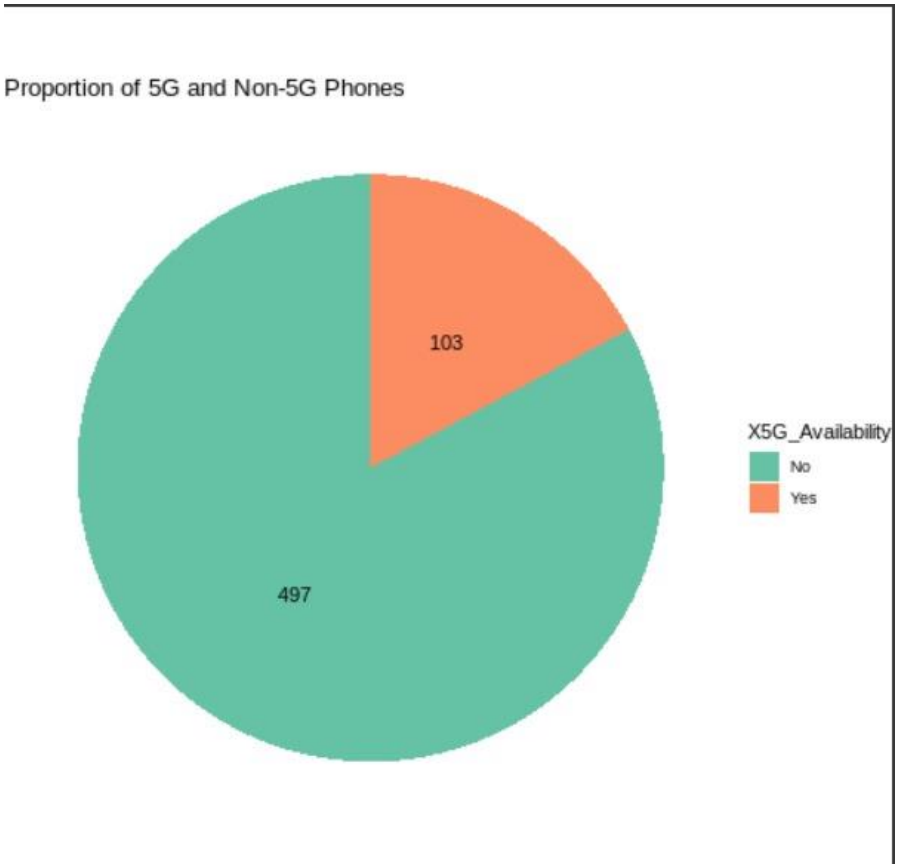
- Top 5 Phones with Largest Batteries

	Battery_Capacity
1	6000
2	6000
3	6000
4	6000
5	6000



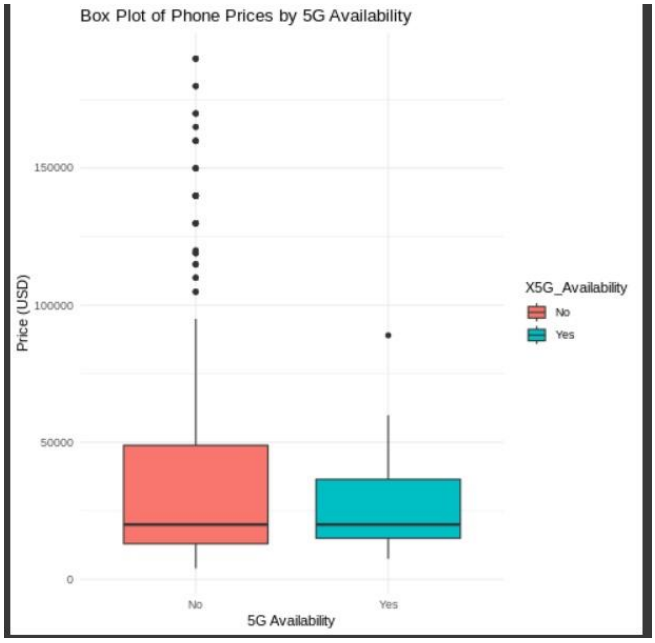
  - **Description:** A bar chart illustrating the top five phone models with the highest battery capacities, highlighting devices that prioritize power efficiency and longevity.
  - **Key Findings:**
    - The analysis identifies the top five phone models with the largest battery capacities, showcasing their respective brands and capacity figures.
    - Phones with higher battery capacities may appeal to consumers who value prolonged usage without frequent recharging, making them ideal for heavy users or those who rely on their devices throughout the day.
    - The data suggests that specific brands are focused on enhancing battery performance, potentially leading to competitive advantages in the market.
  - **Business Implications:** Identifying the phones with the largest battery capacities can inform consumers' purchasing decisions, particularly for those prioritizing battery life. Manufacturers can use these insights to guide product development and marketing strategies, emphasizing battery performance as a key feature. Additionally, understanding trends in battery capacity among leading models can help brands position themselves competitively and highlight innovations in power efficiency.

- Proportion of 5G and Non-5G Phones:



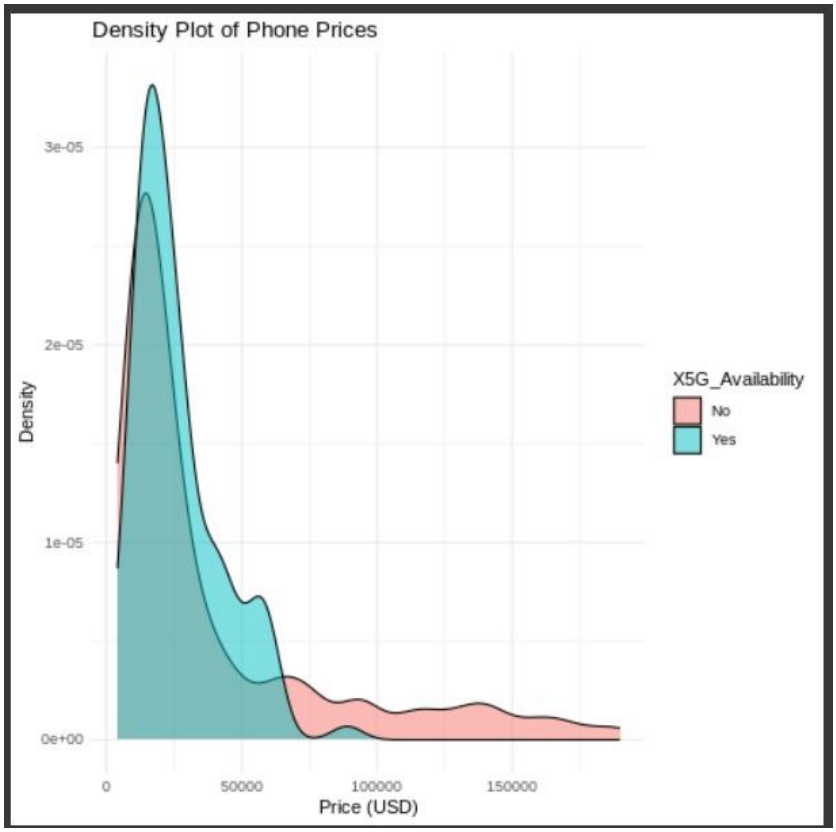
  - **Description:** A pie chart displaying the proportion of 5G phones compared to Non-5G phones, providing a visual representation of the market share of each category.
  - **Key Findings:**
    - The analysis reveals the distribution of phones equipped with 5G technology versus those without, indicating the current trend in mobile connectivity.
    - A significant proportion of the market may be shifting towards 5G phones, reflecting growing consumer demand for faster and more reliable network capabilities.
    - The presence of Non-5G phones in the market highlights the ongoing relevance of budget-friendly options, catering to consumers who may not prioritize the latest technology.
  - **Business Implications:** The proportion of 5G to Non-5G phones can inform manufacturers and retailers about consumer preferences and market trends. Companies should consider investing in the development of 5G devices, as demand is likely to increase. Additionally, marketing strategies may need to highlight the advantages of 5G technology to attract tech-savvy consumers, while still providing options for those who prioritize affordability over the latest features.

- Box Plot of Phone Pries by 5G Availability



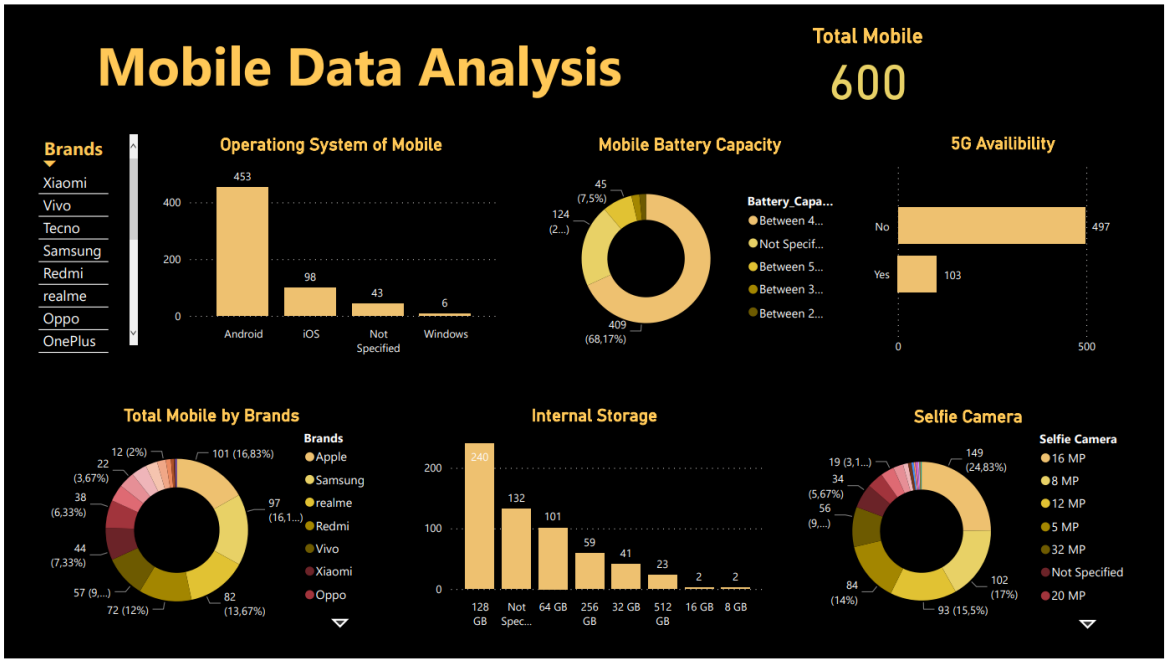
  - **Description**: A box plot visualizing the price distribution of phones based on whether they support 5G, helping to identify any pricing differences between 5G and Non-5G models.
  - **Key Findings:**
    - The box plot reveals the range, median, and distribution of phone prices for both 5G and Non-5G categories.
    - Typically, 5G phones may exhibit a higher median price than Non-5G phones, reflecting the premium often associated with newer technology.
    - Outliers in each category indicate some overlap in pricing, with certain high-end Non-5G phones priced similarly to or higher than some 5G models.
  - **Business Implications:** This analysis helps manufacturers and retailers understand the pricing landscape of 5G versus Non-5G phones. If 5G phones consistently show a higher price range, businesses can target tech-savvy consumers with premium marketing, while maintaining a more affordable Non-5G lineup for budget-conscious customers. Insights from the price distribution can inform pricing strategies, enabling companies to competitively position their 5G devices to maximize market reach.
  - **Why Use a Box Plot:** A box plot is ideal for this analysis because it efficiently represents the spread and central tendency (median) of phone prices for both categories (5G and Non-5G) in a single view. This type of visualization is well-suited for comparing price distributions and identifying any skewness, outliers, or overlaps, making it easier to interpret relative price trends across the categories.

- Density Plot of Phone Prices



  - **Description:** A density plot showing smartphone price distribution by 5G availability, providing insights into pricing trends in the market.
  - **Key Findings:**
    - **Distinct Price Clusters**: 5G phones generally fall into higher price ranges, while non-5G phones spread across a broader range, covering budget and mid-tier markets.
    - **Premium for 5G**: Higher price densities for 5G phones suggest added value for advanced technology.
    - **Diverse Market for Non-5G**: Non-5G phones cater to a wider consumer base, with varied pricing to meet different budgets.
    - **Consumer Segmentation**: Pricing reflects the divergence in demand between 5G and non-5G buyers.
  - **Business Implications:**
    - **Pricing Strategy**: Brands can optimize pricing for 5G and non-5G phones, balancing accessibility and profit.
    - **Product Positioning**: Positioning strategies can target specific segments, with price reflecting consumer preferences.
    - **Targeted Marketing**: Insights on price distributions support campaigns tailored to budget-friendly or premium markets.

PowerBI Visualisation:



4. Discussion

Brand Positioning and Market Segments

The dataset, which includes 600 mobile phone entries across various brands, showcases distinct pricing strategies. Premium brands like Apple and Samsung maintain high average prices, appealing to consumers seeking top-tier devices. Meanwhile, Xiaomi and Realme cater to budget-conscious buyers, focusing on affordability and value. This segmentation illustrates each brand's market focus, from luxury to mass appeal.

Technology Adoption and 5G Support

Our 5G support analysis revealed that brands like Samsung and OnePlus are leading the adoption of next-gen technology, prioritizing future-proof features for premium consumers. Budget brands showed less emphasis on 5G, likely due to cost considerations. This trend points to a phased rollout, where 5G adoption aligns with premium offerings initially.

Battery Life and Performance Demands

Queries on battery and RAM capacities underscored consumer demand for durability and performance. Most phones have battery capacities between 4000–5000 mAh, reflecting the need for longer battery life across price ranges. High RAM capacities in certain brands highlight consumer interest in devices suitable for multitasking and performance-intensive tasks, like gaming.

Brand Diversity and Product Line Strategy

Samsung and Xiaomi lead in product range, with models spanning budget to premium segments. Samsung’s strategy of broad product offerings aims to reach diverse consumers, while Xiaomi focuses on entry-level and mid-tier markets. This breadth helps both brands capture multiple market segments, increasing their competitiveness.

Regional Impact on 5G and Innovation

Examining the origins of 5G-enabled devices, we noted that countries like South Korea and China are prominent. This aligns with their advanced telecom infrastructure and highlights their leadership in 5G innovation, reflecting global trends in mobile technology advancements.

5. Concluding Remarks

This mobile data analysis project enabled a comprehensive evaluation of mobile phone attributes and trends, from pricing and battery life to RAM and 5G adoption. Key findings include:

- **Market Segmentation**: The mid-range segment dominated the Android market, with brands offering a variety of models between 10,000–20,000. In contrast, Apple's premium pricing strategy remains consistent.
- **Feature Trends**: Battery capacity and 5G support emerged as central features, with many brands focusing on models with 4000–5000 mAh and 5G capabilities.
- **Brand Positioning**: Market leaders like Samsung, Apple, and Xiaomi continue to provide wide-ranging options for different consumer preferences, from high-performance devices to budget-friendly models.

This analysis demonstrates the importance of integrating data from various sources and using Hive for structured storage and querying. The project underscores the value of data warehousing and analysis in extracting insights to drive informed decision-making in the mobile technology sector.

6. Future Work

Real-Time Data Integration: Extend the project to support real-time data ingestion using Apache Kafka or Flume to capture live data streams. This would allow for timely updates on mobile market trends and the ability to respond quickly to emerging customer preferences

Machine Learning Models: Implement predictive modeling and machine learning algorithms to forecast trends such as future demand for specific mobile features, price changes, or brand popularity. For instance, regression models or clustering algorithms could help predict which features drive higher sales or group similar products by their specifications and price range

Scalability for Larger Data Sets: Improve the scalability of the system by exploring other big data technologies such as Apache Spark or Hadoop YARN to handle larger and more complex datasets. This would allow for quicker processing and analysis as the volume of mobile data continues to grow.

References

1. **Apache Hive Documentation**
   Reference for working with Hive Query Language (HQL) and managing data within the Hadoop ecosystem. Apache Hive Documentation
2. **Hadoop Distributed File System (HDFS)**
   Documentation for configuring and optimizing HDFS, the storage layer for this project. HDFS Documentation
3. **R Programming Documentation**
   Resources and guides on data visualization and statistical analysis using R. R Documentation
4. **Hive Query Language (HQL) Tutorials**
   Tutorials and online resources that provided insights into advanced HQL queries used in data analysis. HQL Guide
5. **Big Data and Hadoop Ecosystem**
   References to online articles and textbooks on the Hadoop ecosystem and its role in big data management.
6. **Mobile Market Analysis Research**
   Background on mobile market trends and consumer preferences, used to contextualize project findings.
7. **R Community and Forums**
   Various R programming forums and communities for troubleshooting and data visualization techniques.