Northeastern University

ALY-6040
Data Mining Applications

BY INSTRUCTOR
Appari, Ajit

GROUP PROJECT REPORT - DATASET PROPOSAL

SUBMITTED BY
Group-6

ITI ROHILLA
SHARON ARIGELA
GOUTHAM YERRAMADA

On
April-15-2024

Introduction

In the fast-paced and highly competitive mobile phone industry, understanding consumer sentiment is paramount for manufacturers and marketers seeking to stay ahead of the curve. With the advent of social media and online review platforms, consumers have more power than ever to influence purchasing decisions and shape brand perceptions. Consequently, analyzing user reviews has become an indispensable tool for businesses aiming to optimize their product offerings, marketing strategies, and customer engagement initiatives.

Our group embarks on a comprehensive exploratory data analysis (EDA) of a rich dataset containing mobile phone user reviews sourced from diverse platforms and countries. As the landscape of consumer preferences continues to evolve, our research seeks to unearth valuable insights that can guide strategic decision-making processes and drive business success in this dynamic industry.

Business Context:

The mobile phone market is characterized by rapid technological advancements, intense competition, and ever-changing consumer preferences. Manufacturers invest heavily in research and development to introduce innovative features and functionalities that captivate consumers' attention and drive sales. "As per the insights from Gartner, smartphone sales saw a 26% increase in the first quarter of 2021, reflecting the dynamic nature of the mobile phone market with its swift technological progress, intense competitive landscape, and evolving consumer preferences 【Gartner】. Meanwhile, marketers navigate a complex landscape of digital channels and social media platforms to engage with customers and amplify brand visibility.

In this context, understanding the factors that influence user ratings and perceptions of mobile phones is crucial for companies vying for market share and brand loyalty. By leveraging data-driven insights, businesses can gain a competitive edge by tailoring their products and marketing efforts to resonate with target audiences across different demographics, regions, and online platforms.

AIM:

The principal aim of this project is to understand the factors that influence user ratings and perceptions of mobile phones—a pivotal determinant of success in the highly competitive mobile phone market. With consumer preferences shifting rapidly amidst technological evolution, the most crucial variable in this project appears to be user ratings.

In our analysis, we'll use **user ratings** as the response variable, aiming to predict how different factors influence customer satisfaction. The predictors will include categorical variables like **country, review source**, and **brand**, alongside derived metrics from customer review text such as sentiment scores. For more granular analysis, we'll also incorporate technical specs and price points as either continuous or categorical variables, depending on their nature. Decision trees, random forests, and generalized linear models (GLM) will parse through this mix of predictors to uncover patterns and drive actionable insights.

RESEARCH QUESTIONS

1. Impact of Country on User Ratings:

Research Question: Does the country of the user significantly influence their ratings of mobile phones?

Significance: With the mobile phone market spanning across the globe, understanding how user ratings vary across different countries is paramount for manufacturers and marketers. By discerning regional preferences and cultural nuances, companies can adapt their product offerings, marketing strategies, and distribution channels to better resonate with diverse international audiences.

2. Association between Review Source and User Ratings:

Research Question: Are there significant differences in user ratings based on the source of the review (e.g., review websites, social media, forums)?

Significance: In an era dominated by digital communication, the source of user reviews plays a pivotal role in shaping consumer perceptions and purchase decisions. By analyzing the association between **review sources** and **user ratings**, businesses can identify key platforms that wield the most influence and allocate resources effectively to engage with customers and manage brand reputation.

3. Predicting Review Scores:

Research Question: Can we predict review scores using customer review text as input features?

Significance: Predictive modeling of **review scores** based on customer review text can provide businesses with insights into the underlying sentiments and keywords that influence scores. This

model can inform product improvements and highlight areas that resonate well with users, ultimately leading to enhanced customer satisfaction and loyalty.

4. Sentiment Classification:

Research Question: Can we classify user sentiments (positive, neutral, negative) based on review text?

Significance: Sentiment classification enables businesses to quickly gauge user sentiment without the need to read through every review manually. Automatic categorization of reviews can prioritize responses to negative feedback, allowing companies to manage their online reputation effectively and address customer concerns promptly.

5. Review Helpfulness:

Research Question: What factors influence the helpfulness of user reviews?

Significance: Understanding the determinants of review helpfulness can guide companies in generating templates or tips for users to write more helpful reviews. This, in turn, can aid future customers in their purchasing decisions, improve the visibility and credibility of user-generated content, and ultimately enhance the overall customer experience.

As we delve into our analysis, our aim is not only to uncover actionable insights but also to empower mobile phone manufacturers and marketers with the knowledge needed to navigate the ever-changing landscape of consumer preferences and online interactions. Through rigorous data analysis and interpretation, we endeavor to provide strategic guidance that drives business growth and fosters lasting connections with customers worldwide.

METHODOLOGY

Dataset Description:

The dataset comprises user reviews for mobile phones sourced from various platforms and countries. Each observation represents a unique user review, providing valuable insights into consumer sentiment and preferences regarding mobile phone products.

S.NO	Column_Name	Description
1	phone_url	URL of the phone's detail page.
2	date	Date of the review.
3	lang	Language of the review.
4	country	Country of the reviewer.
5	source	Website where the review was posted.
6	domain	Domain of the website where the review was posted.
7	score	Rating given by the reviewer.
8	score_max	Maximum possible rating.
9	extract	Extract of the review text.
10	author	Author of the review.
11	product	Product name and specifications.

Table: Dataset Overview

Descriptive Statistics:

Descriptive statistics will be provided for:

- **Numerical variables:** score, score_max, year.

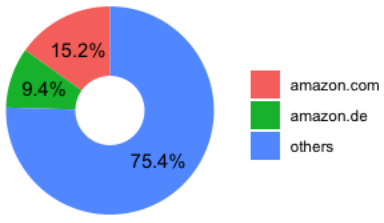
Characteristic	N = 1,415,133
score	
Mean	8.06
Median	9.20
SD	2.57
Minimum	0.20
Maximum	10.00

Table: Descriptive_Statistics_Numerical_Variable

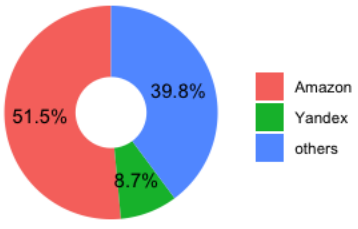
- **Categorical variables** (lang, country, source, domain):

Overview Of Categorical Variable

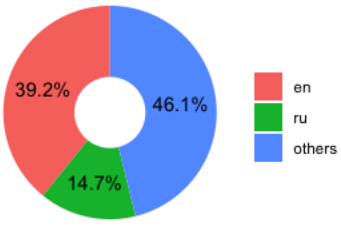
Domain



Sources



language



Summary of Categorical Data	
Summary Statistics	Count
Total Entries	1415133
Unique Phone URLs	5556
Unique Languages	22
Unique Countries	42
Unique Sources	331
Unique Domains	384
Unique Authors	801107
Unique Products	61314

Table: Descriptive_Statistics_Categorical_Variable

Data Preparation:

Before conducting exploratory data analysis (EDA), we perform data preparation steps to ensure the dataset is clean, structured, and ready for analysis. This includes:

1. Reading and combining individual CSV files: We utilize the readr package to read each CSV file containing user reviews and combine them into a single data frame using the bind_rows() function.

2. Normalizing variables: We normalize relevant variables such as the product names to ensure consistency and accuracy in our analysis.

3. Handling missing values: We handle missing values and convert relevant variables to appropriate data types (e.g., converting date to a Date object, score to numeric) to facilitate analysis.

EXPLORATORY DATA ANALYSIS (EDA)

Our EDA aims to uncover patterns, trends, and relationships within the dataset to inform strategic decision-making. Key analyses include:

1. Impact of Country on User Ratings:

Dataset Description: The dataset contains user reviews for mobile phones from various countries, along with information about the source of the review, the product, and the user's rating.

Variables: We will focus on the user's rating (score), the country of the user, and possibly other relevant variables such as the source of the review.

2. Association between Review Source and User Ratings:

Dataset Description: Similar to the previous dataset, this dataset contains user reviews for mobile phones, but with additional information about the source of the review (e.g., review websites, social media, forums).

Variables: We will focus on the user's rating (score) and the source of the review.

3. Predicting Review Scores:

Dataset Description: We will utilize the same dataset containing user reviews for mobile phones, but we will focus on the review text in addition to the numerical rating.

Variables: The response variable will be the numerical review score, and the predictor variables will be features extracted from the review text using techniques such as TF-IDF or word embeddings.

4. Sentiment Classification:

Dataset Description: Similar to the previous dataset, we will use the dataset containing user reviews for mobile phones, with a focus on the review text.

Variables: The response variable will be the sentiment category (positive, neutral, negative), and the predictor variables will be features extracted from the review text.

5. Review Helpfulness:

Dataset Description: Again, we will use the dataset containing user reviews for mobile phones, with a focus on the review text.

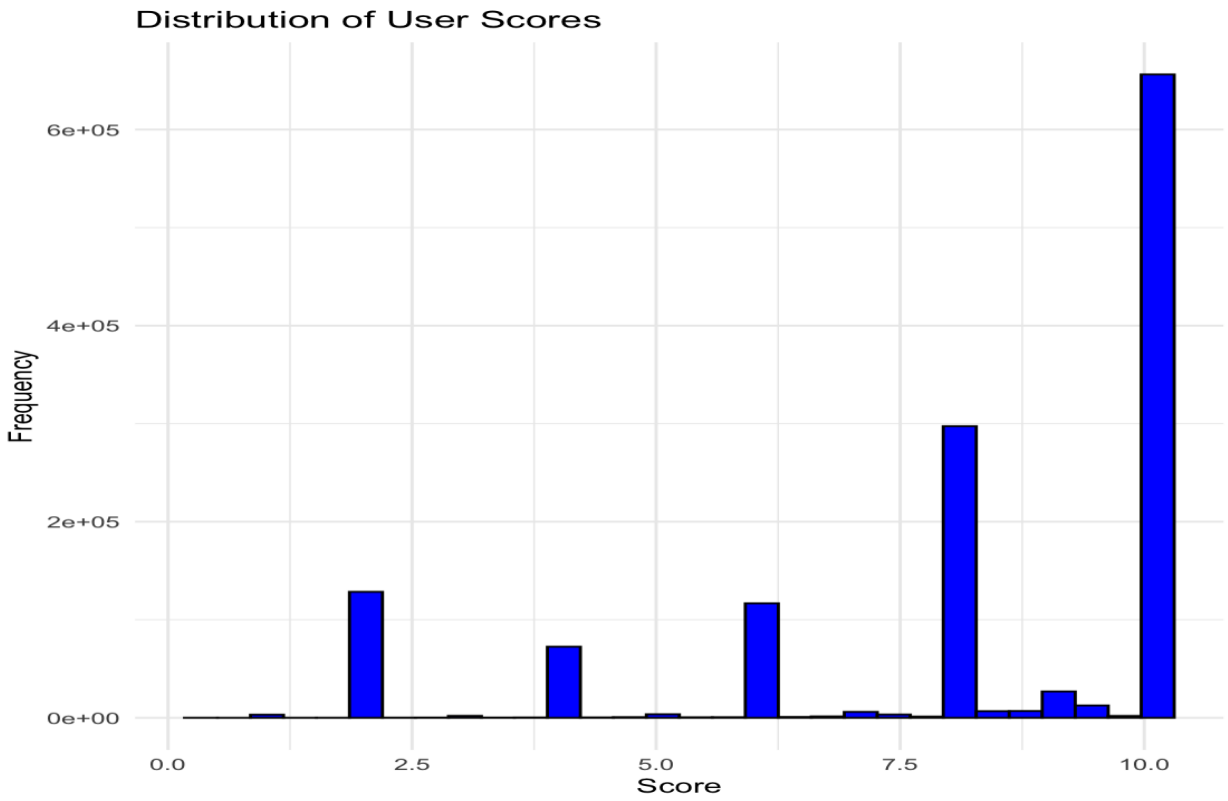
Variables: The response variable will be a binary indicator of whether a review was marked as helpful, and the predictor variables will include features extracted from the review text.

Results & Interpretations:

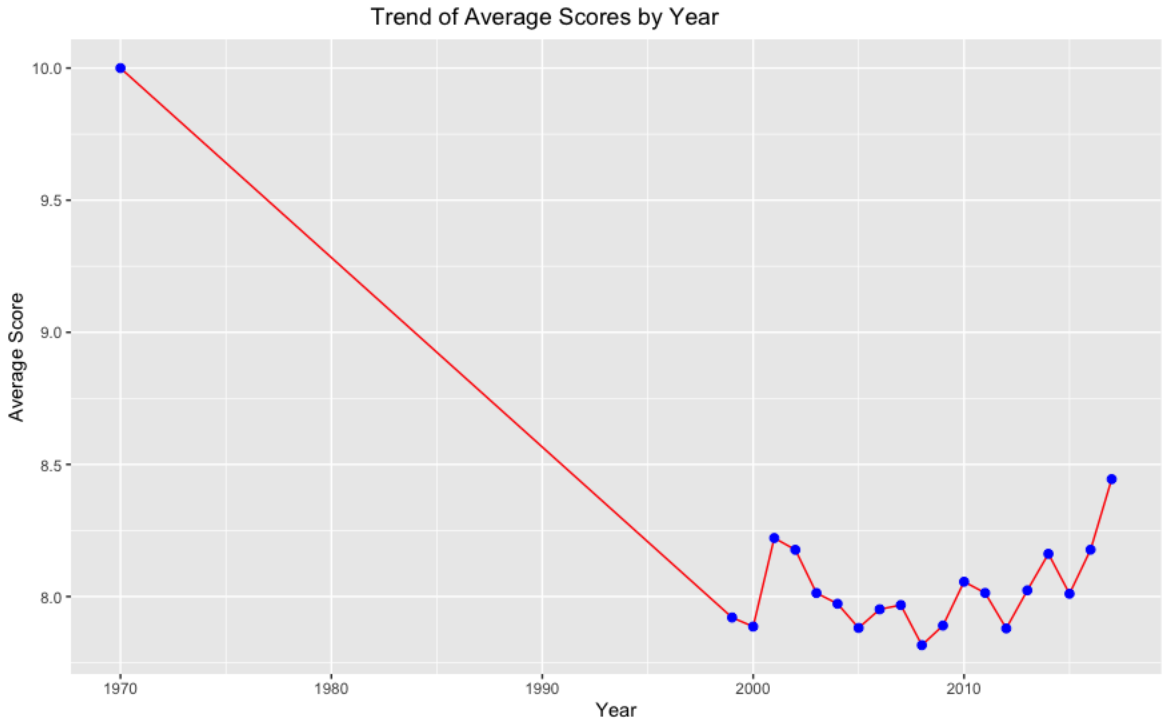
Exploratory Data Analysis (EDA)

Histogram of User Scores:

The histogram illustrates the distribution of user scores, indicating the frequency of different rating values. The distribution may exhibit characteristics such as skewness or multimodality, providing insights into the overall sentiment distribution among users.



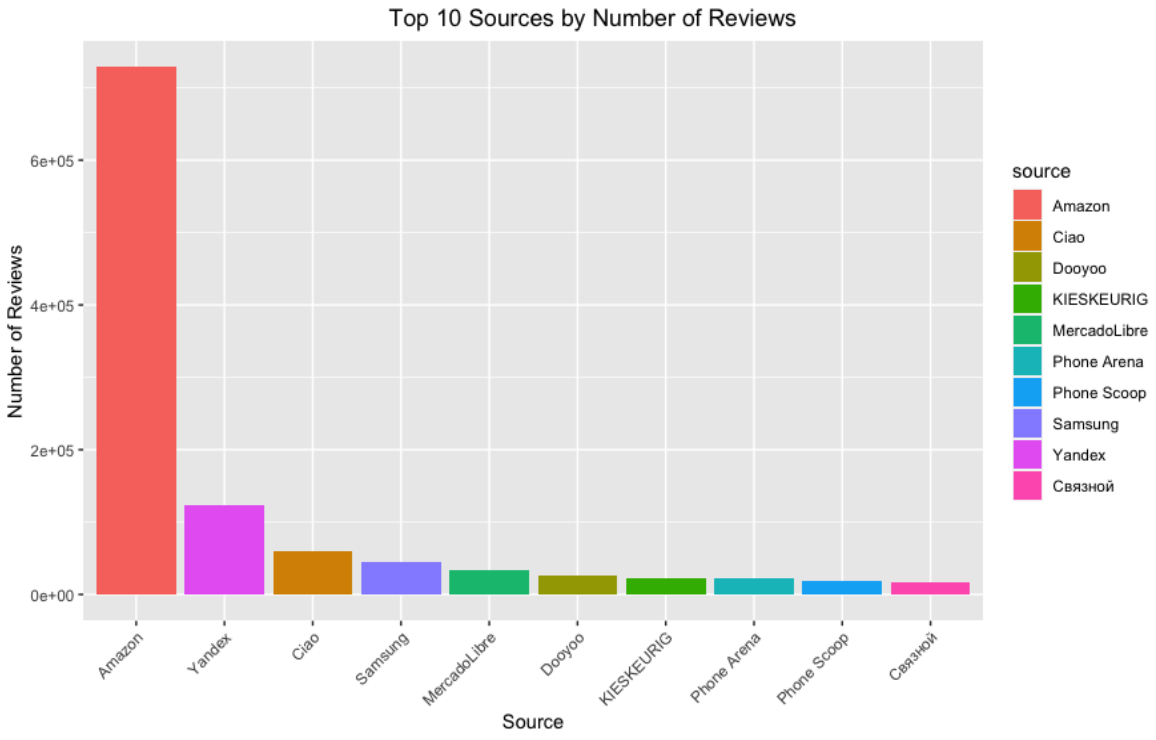
Line Chart of Average Scores:



The line graph shows a notable decline in average scores from the initial peak of nearly 10, leading to a nadir around the late 1990s or early 2000s, as marked by blue points linked by a red line. Following this low, the scores exhibit fluctuations, with a general, albeit slight, increase in more recent years.

This pattern suggests a period of recovery or improvement in the measured attributes, albeit with inconsistencies. To fully understand the causes behind these trends, further analysis is needed, considering possible changes in measurement systems, demographic shifts, or external variables.

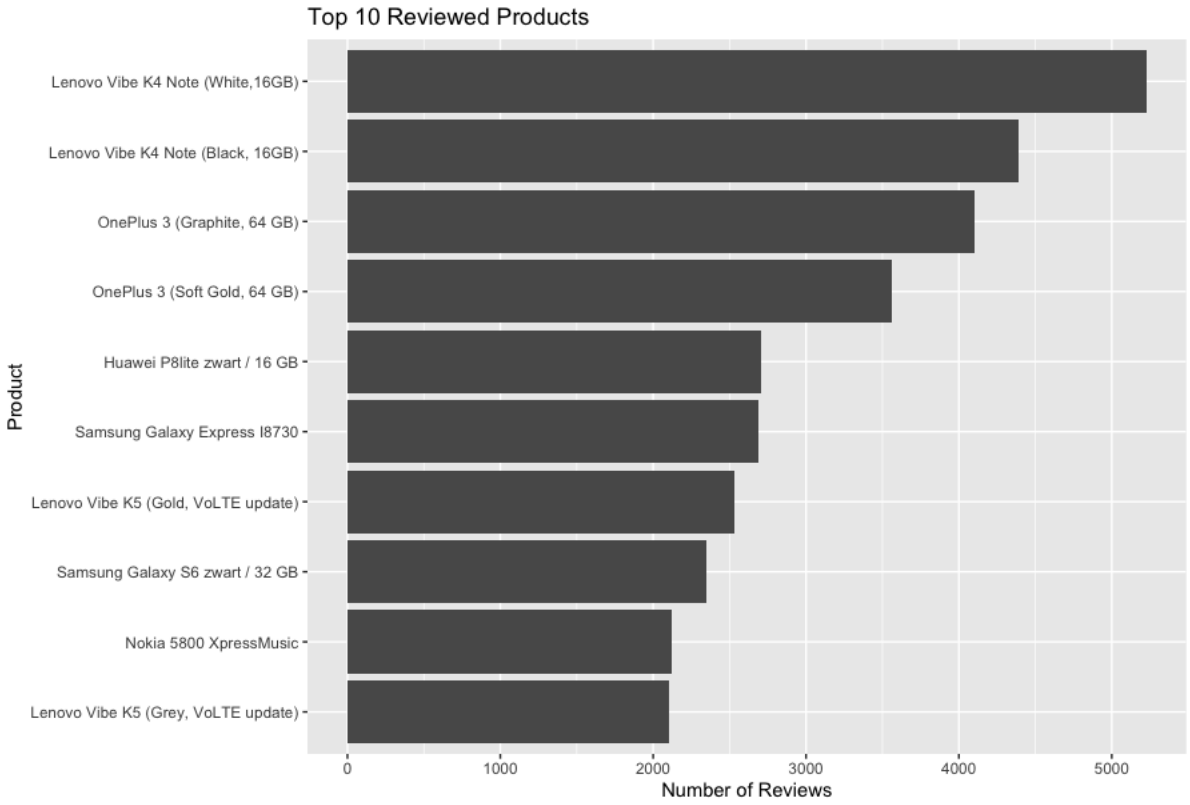
Bar Chart for Top Sources:



In the bar chart, Amazon stands out with the highest number of reviews, represented by a red bar, followed by Yandex and Ciao with fewer reviews. Other platforms like Samsung and MercadoLibre show moderate review counts. The significant gap between Amazon and the rest points to a potential collection bias towards Amazon. Understanding this imbalance is crucial when using the data for analysis or decision-making, as it may reflect varying popularity or customer engagement levels.

Count of Reviews by Product:

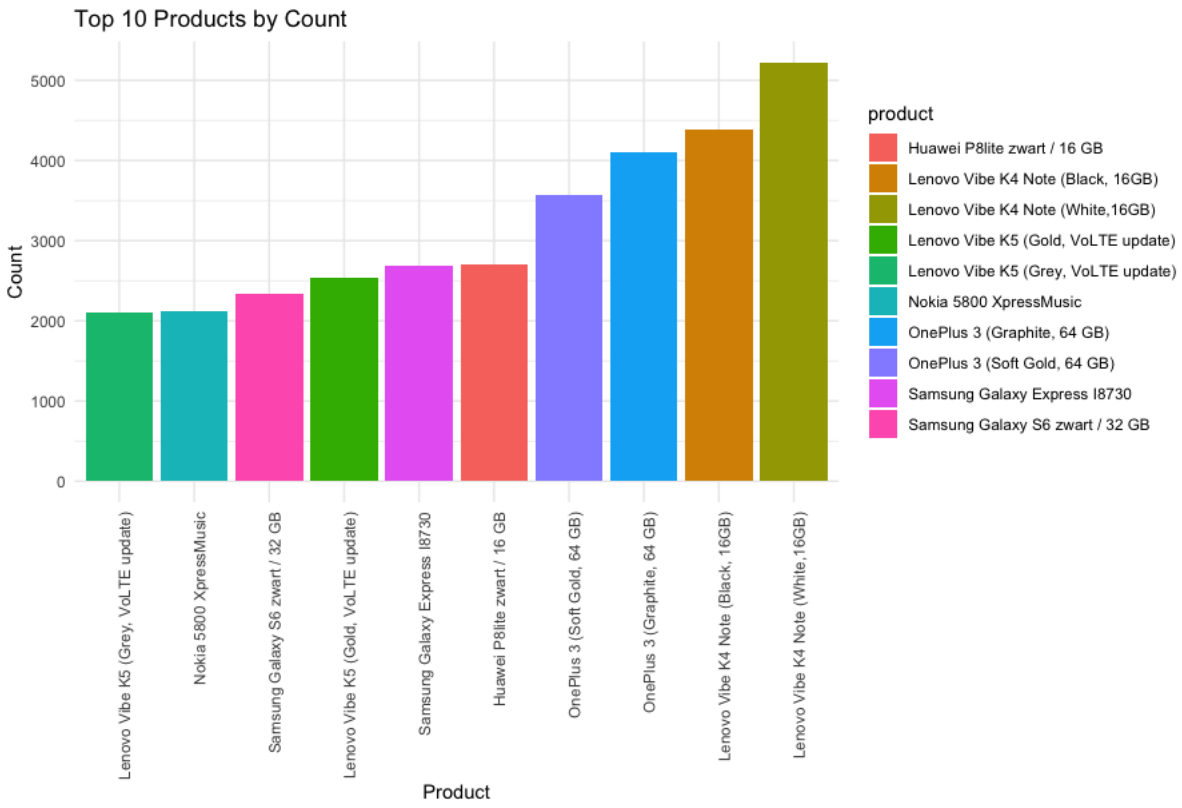
The bar plot presents the top 10 most reviewed products, highlighting the products with the highest review volume. This analysis identifies popular products that attract substantial user attention and can inform product development strategies and marketing campaigns.



Overall, the results of these analyses provide actionable insights for businesses to optimize their products, marketing strategies, and customer support services, ultimately enhancing customer satisfaction and driving business performance.

Product Popularity:

The chart highlights which products are the most popular or most frequently discussed. This could be useful for understanding market trends or consumer preferences. the chart may give insight into which products are generating the most buzz or consumer interest, which could be a proxy for consumer satisfaction.



CONCLUSION

In conclusion, our exploratory data analysis (EDA) provided valuable insights into user ratings of mobile phones and the factors influencing them. Here's a summary of our EDA tasks and the next steps:

1. Exploratory Data Analysis (EDA) Tasks Undertaken:

We analyzed the distribution of user scores through histograms and boxplots, revealing patterns and variations in ratings. By examining scores across different countries and review sources, we identified significant differences, indicating the influence of these factors on user ratings. Additionally, we explored the most reviewed products to understand user preferences and engagement levels.

2. Next Steps:

In the next phase, we plan to delve deeper into the data by performing sentiment analysis and predictive modeling. Sentimental_Analysis will allow us to categorize user reviews based on their sentiment (positive, neutral, negative) and understand overall user sentiment trends. Predictive modeling, including regression analysis and machine learning algorithms, will enable us to predict review scores, classify sentiment, and identify factors influencing review helpfulness. Furthermore, we aim to refine our models and extract actionable insights to help businesses optimize their products, marketing strategies, and customer support services.

By leveraging advanced analytics techniques, we aim to provide actionable recommendations to enhance customer satisfaction, improve brand reputation, and drive business performance in the highly competitive mobile phone market.

CITATIONS

Sentimental Analysis on Mobile Users

https://www.researchgate.net/publication/273532029_Sentiment_Analysis_on_Reviews_of_Mobile_Users

Artificial Intelligence and Sentiment Analysis: A Review in Competitive Research

https://www.mdpi.com/2073-431X/12/2/37

Predicting mobile application breakout using sentiment analysis of Facebook posts

https://journals.sagepub.com/doi/abs/10.1177/0165551520917099

Identifying Significance of Product Features on Customer Satisfaction Recognizing Public Sentiment Polarity: Analysis of Smart Phone Industry Using Machine-Learning Approaches

https://www.tandfonline.com/doi/full/10.1080/08839514.2020.1787676