

Amazon UK

Data Mining Project

amazon.co.uk

Student ID: 23206188
Student Name: Dinesh Thapa

Student ID: 22227184
Student Name: Arun Rajput

Student ID: 23111092
Student Name: Ashwani Kumar

Part of a
Data Mining – CMP7206

BIRMINGHAM CITY UNIVERSITY
FACULTY OF COMPUTING ENGINEERING AND THE BUILT
ENVIRONMENT



Table of Contents

1. Domain Description.....	1
2. Problem Definition	2
3. Literature Review	3
4. Dataset Description.....	5
4.1 Code – Dataset Description using R Programming	5
5. Dataset Preprocessing.....	11
5.1 Our Processes	11
6. Experiments	14
6.1 Exploratory Data Analysis (EDA) and Time Series Analysis.....	14
6.2 Regression Analysis	16
6.3 K-means Clustering.....	17
7. Analysis and Results.....	19
8. Conclusion	26
9. References.....	27
10. Appendix: Complete R Code.....	28

TABLE OF FIGURES

FIGURE 1: SUMMARY OF THE DATASET	6
FIGURE 2: STRUCTURE OF THE DATASET	7
FIGURE 3: SUMMARY STATISTICS OF NUMERICAL ATTRIBUTES.....	8
FIGURE 4: MISSING VALUES AND FREQUENCY IN VARIABLES	8
FIGURE 5: DENSITY PLOT OF STARS	9
FIGURE 6: CORRELATION MATRIX FOR STARS, REVIEWS, PRICE AND BOUGHTINLASTMONTH	10
FIGURE 7: PRICE BY BESTSELLER STATUS	10
FIGURE 8: REVIEWS BY BESTSELLER STATUS.....	10
FIGURE 9: TOP 10 TRENDING PRODUCT CATEGORIES BY TOTAL SALES.....	19
FIGURE 10: STAR VS BOUGHTINLASTMONTH - REGRESSION ANALYSIS.....	21
FIGURE 11: LINEAR REGRESSION: REVIEWS VS BOUGHTINLASTMONTH	21
FIGURE 12: K-MEANS CLUSTERING SHOWING CLUSTERED CATEGORIES BASED ON SALES PERFORMANCE	23
TABLE 1:DATASET DESCRIPTION	5
TABLE 2: TOP 10 PRODUCT CATEGORIES WITH THEIR TOTAL SALES VOLUME	19

1. Domain Description

Amazon UK is the biggest global E-commerce brand. It is one of the best e-commerce in the UK that offers a vast range of products including Electronics, Clothing, Books, Household Items, Beauty Products, and many more.

Amazon UK Dataset helps us to get in-depth insights and visualizations of what products sell best, what are the trending product categories and their sales performance, customer ratings, and best-selling products, identifying the niche categories for high sales and the best price range for a product category.

The primary goal of this project is to extract valuable insights that can significantly impact the Amazon platform's sales performance and enhance customer satisfaction. In addition, Data Mining helps us to study customer behavior, preferences, trends, and patterns in the product information and its variables. Furthermore, its analysis will help them in offering personalized marketing and product recommendations.

Good data mining allows for forecasting customer demand, inventory management, and creating the right pricing strategies based on historical sales data and trends. Data Mining optimizes sales and marketing strategies by analyzing data supports in targeting specific audiences and improving conversion rates.

In summary, these insights will drive data-driven strategies, optimizing marketing approaches, inventory management, and overall sales performance within the e-commerce platform.

2. Problem Definition

The primary aim of this Data Mining Project centers around the Amazon UK dataset to extract valuable insights that drive growth in sales performance and enhance customer satisfaction within the e-commerce platform. We believe that it will revolutionize the marketing and sales strategies of Amazon UK.

The major problem statements and our objectives include:

a. Finding Trending Product Categories and Sales Performance

We will find what types of products are becoming more popular and selling better within the Amazon UK e-commerce platform. This involves looking at how well different types of products are selling over time. Through visual tools like time-series analysis, we'll create graphs or charts to see the sales trends of various product categories. This will help us understand which types of items are in high demand and performing well in the market right now.

b. Analyzing Customer Ratings for Best/Top Seller Products

We're examining which products customers like the most by looking at their star ratings. Our focus is on finding the items with the best ratings and figuring out why they're so popular. We're also studying whether these highly rated products tend to have lots of reviews and if they sell really well. We want to see if there's a connection between high ratings, the number of reviews, and how much these products are selling. This helps us understand if products with great ratings are more likely to be bought by customers.

c. Identifying niche categories for high sales

We're looking into specific product categories that are doing exceptionally well in sales. By analyzing how different types of products are selling, we can find the ones that have a lot of potential for high sales. Our goal is to identify categories where there's less competition but still a strong demand from customers. This insight helps us focus on areas where sales can be improved, ultimately boosting overall sales performance.

In conclusion, our Data Mining Project for the Amazon UK dataset aims to extract insights on trending product categories, top-rated items, and niche sales areas to empower Amazon UK with data-driven strategies, enhancing marketing, inventory management, and overall sales performance in e-commerce.

3. Literature Review

Our goal is to mine the Amazon UK Products dataset and get valuable data-driven decisions and insights that help us in creating better marketing and sales strategies for Amazon UK to grow and boost its sales performance.

So, we did a literature review to understand what has been already done in the e-commerce industry data mining and what are the problems and challenges, what are the techniques that we can use, and many more. Several scholars and researchers provided insightful methodologies including different approaches, and techniques with predictive accuracy in their report that they created for data mining for e-commerce analysis and prediction.

Orie Abu Alghanam, Sumaya N. Al-Khatib, and Mohammad O. Hiari from Al-Ahliyya Amman University have introduced us to predicting customer purchase behavior in an e-commerce context. Their research employed K-Means clustering, decision tree classification, and the Apriori PT association rule algorithm, achieving a predictive accuracy of 86.5%. These methods illustrated the way to understand customer behavior and preferences in the online retail domain.

Anurag Bejj, from the Birla Institute of Technology & Science, conducted a comprehensive sales analysis of e-commerce websites using data mining techniques. Using the ID3 algorithm for decision tree classification, their research yielded a predictive accuracy of 86.4789%. This study focused on implementing decision trees to predict sales patterns within e-commerce platforms.

Jing Zhang and Juan Li contributed valuable insights in their study on Retail Commodity Sale Forecast Models using data mining. Through innovative techniques such as season analysis models and Markov models, they achieved predictive accuracies of 79.1% and 79.7% respectively. These models provide effective tools for forecasting sales in retail commodities, aiding in inventory management and sales projection strategies.

Yuan Chen and Feifei Wang explored the domain of dynamic pricing models for e-commerce based on data mining. Their work incorporated association rule analysis, cluster analysis, classification analysis, and sequential pattern analysis. These methodologies offer diverse perspectives to establish effective dynamic pricing strategies, enabling businesses to adjust prices in response to market dynamics and customer behavior.

These academic studies have collectively contributed to a deeper understanding of leveraging data mining techniques for e-commerce, spanning from predicting customer behavior to sales analysis and dynamic pricing.

They have offered invaluable methodologies with varying predictive accuracies, providing the foundation for effective data-driven strategies that help Amazon UK grow based on data-driven decisions.

Case Study and Research Papers

In addition, our extensive research also includes a collection of case studies and research papers that significantly contribute to the understanding and application of data mining techniques within the e-commerce domain.

The work by Jon Kepa Gerrikagoitiaa, Iñigo Castandera, Fidel Rebóna, and Aurkene Alzua-Sorzabal provided meaningful insight into the new trends of intelligent e-marketing based on web mining for e-shops. This research provides the implementation of different techniques such as classification, regression, clustering, and affinity analysis, shedding light on their applications and effectiveness in enhancing e-commerce strategies.

Bei-Ni Yan, Tian-Shyug Lee, and Tsung-Pei Lee conducted a comprehensive analysis of research papers in e-commerce from 2000 to 2013, utilizing text mining approaches. Their study extracted high-frequency keywords, density analysis, centrality analysis, and multidimensional scaling (MDS) analysis, offering an in-depth understanding of the evolution of e-commerce research during that period.

David L. Banks and Yasmin H. Said explored the statistical challenges and opportunities in electronic commerce. Their work addressed the unique statistical challenges that arise in the e-commerce industry, providing insights into the opportunities that data-driven methodologies offer in enhancing business strategies within this domain.

Furthermore, the study conducted by Bhumika Pahwa, Dr. S. Taruna, and Dr. Neeti Kasliwal focused on the important role of data mining in analyzing consumers' online buying behavior.

In conclusion, this research work and case studies collectively helped us in getting the data mining applications, techniques, and methodologies within the electronic commerce sector.

4. Dataset Description

The dataset is publicly available on the Kaggle website, and it is from product information from Amazon UK Products Dataset. It has 2.2 million rows and 10 columns.

Table 1:Dataset Description

Attribute	Description	Data Type
asin	Product ID unique to Amazon	string
title	Product title	string
imgUrl	URL of the product image	string
productURL	URL of the product on Amazon	string
stars	Product rating; 0 indicates no ratings available	float
reviews	Number of reviews; 0 signifies no reviews available	integer
price	Buy now price of the product in GBP; 0 denotes unavailable pricing info	float
isBestSeller	Indicates whether the product achieved Amazon BestSeller status	boolean
boughtInLastMonth	Number of products sold in the last month according to Amazon	integer
categoryName	Category name to which the product belongs	string

4.1 Code – Dataset Description using R Programming

We have implemented various operations using R programming to analyze our dataset, its structures, summary, and attributes. In addition, we have also performed some visualizations to understand the overall data trends and figures well-using histogram, correlogram, and boxplot on our dataset.

a. Loading dataset into variable using R

```
getwd()

setwd("D:/MSc Big Data/Data Mining/Dataset")

library(tidyverse)

amazon_data <- read.csv("D:/MSc Big Data/Data
Mining/Dataset/Amazon_UK_Products_Dataset_2023.csv")
```

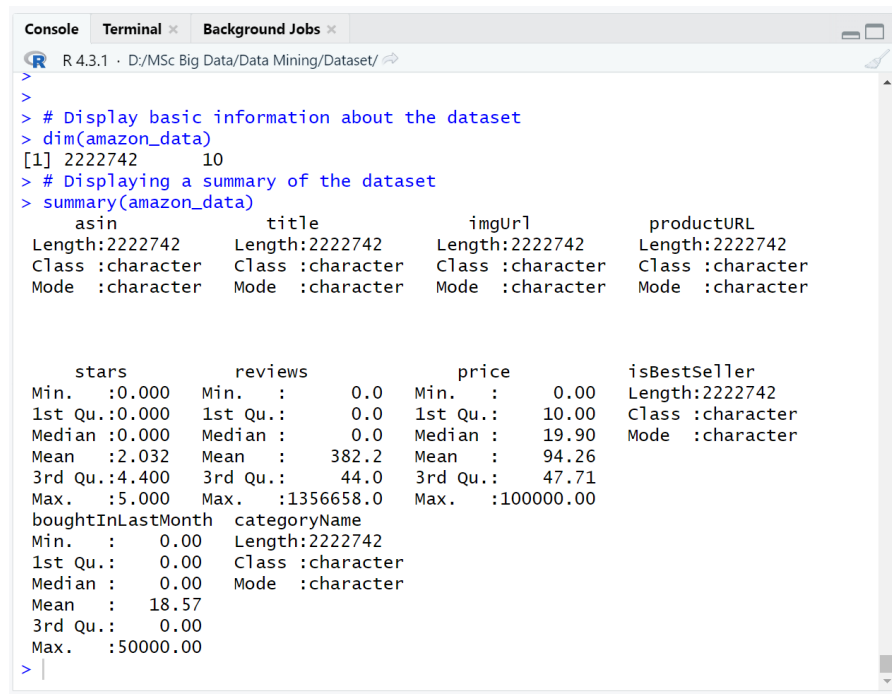
b. Display basic information and summary of the dataset

```
dim(amazon_data)

summary(amazon_data)
```

Output:

Figure 1: Summary of the dataset



c. Structure of a dataset and generating summary of numerical variables

```
str(amazon_data)

unique(amazon_data$isBestSeller)

unique(amazon_data$categoryName)

head(amazon_data)

amazon_data[c("asin", "title", "stars", "price", "categoryName")]

summary(amazon_data$stars)

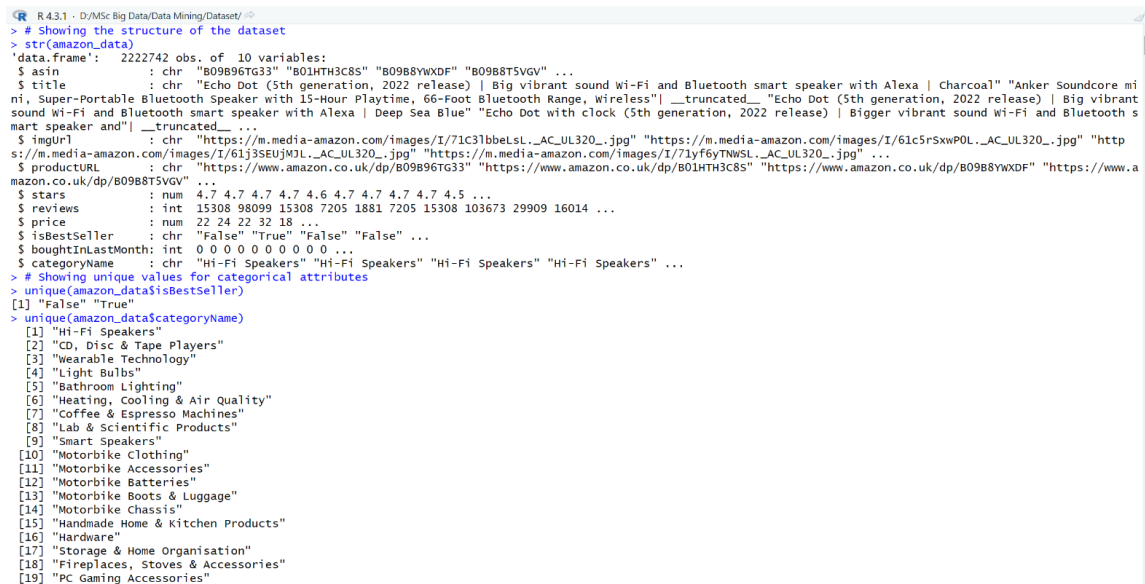
summary(amazon_data$price)

summary(amazon_data$reviews)

summary(amazon_data$boughtInLastMonth)
```

Output:

Figure 2: Structure of the dataset



```
R 4.3.1 - D:/MSc Big Data/Data Mining/Dataset/
> # Showing the structure of the dataset
> str(amazon_data)
'data.frame': 2222742 obs. of  10 variables:
 $ asin      : chr  "B09B96TG33" "B01HTH3C8S" "B09B8YWXDF" "B09B8T5GVG" ...
 $ title     : chr  "Echo Dot (5th generation, 2022 release) | Big vibrant sound Wi-Fi and Bluetooth smart speaker with Alexa | Charcoal" "Anker Soundcore mi
ni, Super-Portable Bluetooth Speaker with 15-Hour Playtime, 66-Foot Bluetooth Range, Wireless" | __truncated__ "Echo Dot (5th generation, 2022 release) | Big vibrant
sound Wi-Fi and Bluetooth smart speaker with Alexa | Deep Sea Blue" "Echo Dot with clock (5th generation, 2022 release) | Bigger vibrant sound Wi-Fi and Bluetooth s
mart speaker and" | __truncated__ ...
 $ imageUrl  : chr  "https://m.media-amazon.com/images/I/71C3lbbELsL._AC_UL320_.jpg" "https://m.media-amazon.com/images/I/61c5rSxwPOL._AC_UL320_.jpg" "http
s://m.media-amazon.com/images/I/61j3SEUjMjL._AC_UL320_.jpg" "https://m.media-amazon.com/images/I/71yfeYTNWSL._AC_UL320_.jpg" ...
 $ productURL: chr  "https://www.amazon.co.uk/dp/B09B96TG33" "https://www.amazon.co.uk/dp/B01HTH3C8S" "https://www.amazon.co.uk/dp/B09B8YWXDF" "https://www.a
mazon.co.uk/dp/B09B8T5GVG" ...
 $ stars     : num  4.7 4.7 4.7 4.7 4.6 4.7 4.7 4.7 4.7 4.5 ...
 $ reviews  : int   15308 98099 15308 7205 1881 7205 15308 103673 29909 16014 ...
 $ price     : num   22 24 22 32 18 ...
 $ isBestSeller: chr   "False" "True" "False" "False" ...
 $ boughtInLastMonth: int   0 0 0 0 0 0 0 0 0 ...
 $ categoryName: chr   "Hi-Fi Speakers" "Hi-Fi Speakers" "Hi-Fi Speakers" "Hi-Fi Speakers" ...
> # Showing unique values for categorical attributes
> unique(amazon_data$isBestSeller)
[1] "False" "True"
> unique(amazon_data$categoryName)
[1] "Hi-Fi Speakers"
[2] "CD, Disc & Tape Players"
[3] "Wearable Technology"
[4] "Light Bulbs"
[5] "Bathroom Lighting"
[6] "Heating, Cooling & Air Quality"
[7] "Coffee & Espresso Machines"
[8] "Lab & Scientific Products"
[9] "Smart Speakers"
[10] "Motorbike Clothing"
[11] "Motorbike Accessories"
[12] "Motorbike Batteries"
[13] "Motorbike Boots & Luggage"
[14] "Motorbike Chassis"
[15] "Handmade Home & Kitchen Products"
[16] "Hardware"
[17] "Storage & Home Organisation"
[18] "Fireplaces, Stoves & Accessories"
[19] "PC Gaming Accessories"
```

Figure 3: Summary Statistics of numerical attributes

```
200 4.3 415.00 Hi-Fi Speakers
[ reached 'max' / getOption("max.print") -- omitted 222542 rows ]
> # Generating summary statistics for numerical attributes
> summary(amazon_data$stars)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.000  0.000   0.000   2.032   4.400   5.000
> summary(amazon_data$price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.00   10.00   19.90   94.26   47.71 100000.00
> summary(amazon_data$reviews)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0    0.0    0.0    382.2   44.0 1356658.0
> summary(amazon_data$boughtInLastMonth)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.00   0.00   0.00   18.57   0.00 50000.00
> |
```

d. Missing Values, Unique Products and Frequency in Variables

```
colSums(is.na(amazon_data))
length(unique(amazon_data$asin))
table(amazon_data$categoryName)
table(amazon_data$isBestSeller)
```

Output:

Figure 4: Missing Values and Frequency in variables

```
R 4.3.1 - D:/MSc Big Data/Data Mining/Dataset/
> # Checking missing values in the dataset
> colSums(is.na(amazon_data))
   asin      title    imgURL    productURL      stars    reviews      price    isBestSeller
   0         0         0         0         0         0         0         0
boughtInLastMonth    categoryName
   0         0
> # Counting the number of unique products in the dataset
> length(unique(amazon_data$asin))
[1] 2222742
> # Frequency of Categorical Variable
> # Counting Frequency for 'categoryName'
> table(amazon_data$categoryName)
   3D Printers      248      3D Printing & Scanning      4022      Abrasive & Finishing Products      250
   Action Cameras    1014      Adapters      251      Agricultural Equipment & Supplies      8036
   Alexa Built-In Devices    69      Art & Craft Supplies      200      Arts & Crafts      5756
   Baby      10833      Baby & Toddler Toys      9058      Bakeware      236
   Ballet & Dancing Footwear    2982      Barebone PCs      9477      Basketball Footwear      6697
   Bass Guitars & Gear    519      Bath & Body      14880      Bathroom Furniture      2627
   Bathroom Lighting    254      Bathroom Linen      404      Beauty      5872
   Bedding & Linen    3115      Bedding Accessories      203      Bedding Collections      6298
   Bedroom Furniture    4759      Beer, Wine & Spirits      9414      Billiard, Snooker & Pool      249
   Binoculars, Telescopes & Optics      9031      Bird & Wildlife Care      8984      Birthday Gifts      7413
   Blank Media Cases & Wallets    243      Boating Footwear      139      Bowling      242
   Boxes & Organisers    847      Boxing Shoes      4069      Boys      11784
   Building & Construction Toys    9125      Building Supplies      5681      Cables & Accessories      9029
   Calendars & Personal Organisers    250      Camcorders      470      Camera & Photo Accessories      7664
   Cameras      470      Candles & Holders      7664
```

e. Exploring Dataset by using some basic Visualization

```
plot(density(amazon_data$stars), main = "Density Plot of Stars", xlab =  
"Stars")  
  
correlation_matrix <- cor(select(amazon_data, stars, reviews, price,  
boughtInLastMonth))  
  
print(correlation_matrix)  
  
corrplot::corrplot(correlation_matrix, method = "circle")  
  
boxplot(price ~ isBestSeller, data = amazon_data, main = "Price by BestSeller  
Status", xlab = "BestSeller Status", ylab = "Price in GBP")  
  
boxplot(reviews ~ isBestSeller, data = amazon_data, main = "Reviews by  
BestSeller Status", xlab = "BestSeller Status", ylab = "Number of Reviews")
```

Output:

Figure 5: Density Plot of Stars

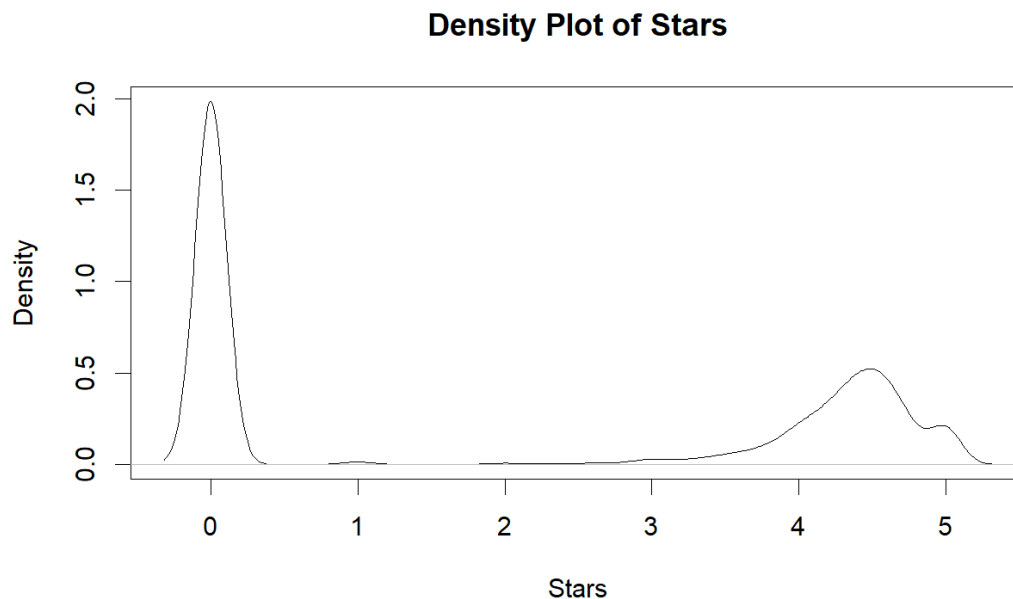


Figure 6: Correlation Matrix for Stars, Reviews, Price and boughtInLastMonth

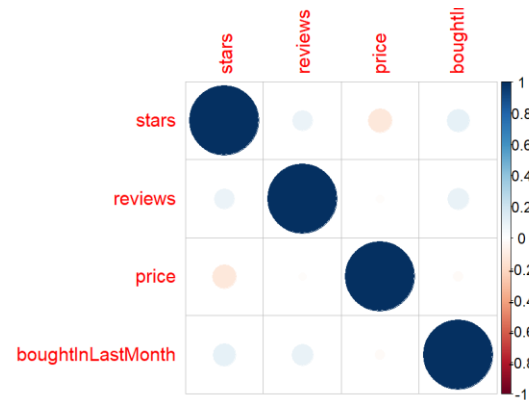
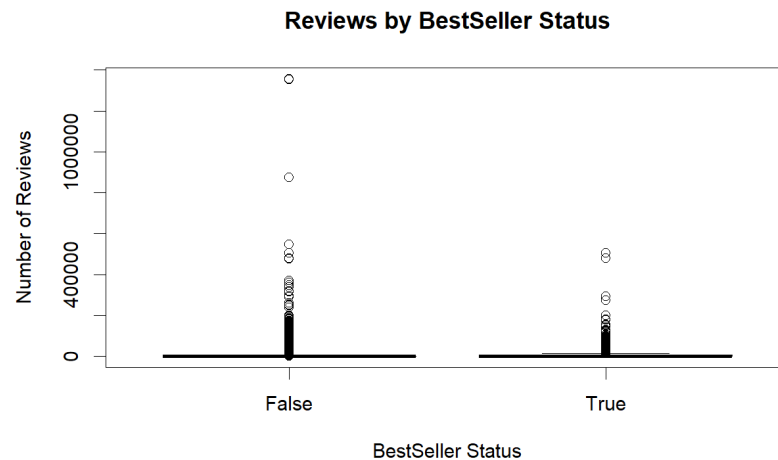


Figure 7: Price by BestSeller Status



Figure 8: Reviews by BestSeller Status



5. Dataset Preprocessing

Data Preprocessing for the Amazon UK dataset involves several steps to clean, organize, and prepare the data before analysis. It involves refining and organizing data to enhance its quality and suitability for analysis.

Data preprocessing helps us to get accurate results and decisions. For example: we convert categorical or nominal data type so, we convert that data into factor data type to improve the accuracy of our model.

5.1 Our Processes

We follow the step-by-step processes while preprocessing data using different techniques available in R programming.

a. Handling Missing Values

```
# Summary Statistics by Group
# Summary statistics of 'price' by 'categoryName'
aggregate(price ~ categoryName, data = amazon_data, FUN = summary)

# Summary statistics of 'reviews' by 'categoryName'
aggregate(reviews ~ categoryName, data = amazon_data, FUN = summary)
```

b. Handling Outliers

```
# Handling Outliers
# Detect and remove outliers in 'price' using IQR method
Q1 <- quantile(amazon_data$price, 0.25, na.rm = TRUE)
Q3 <- quantile(amazon_data$price, 0.75, na.rm = TRUE)
IQR_value <- Q3 - Q1
upper_bound <- Q3 + 1.5 * IQR_value
amazon_data <- filter(amazon_data, price <= upper_bound)
```

c. Feature Engineering

```
# Feature Engineering / Adding new column in dataset  
# Creating a new column 'review_sentiment' based on review counts  
amazon_data$review_sentiment <- ifelse(amazon_data$reviews > 50, "Positive",  
"Neutral")
```

d. Data Transformation

```
# Data Transformation  
# Log transformation of 'boughtInLastMonth' column  
amazon_data$boughtInLastMonth <- log(amazon_data$boughtInLastMonth + 1)  
  
# Normalization (min-max scaling) of 'boughtInLastMonth' column  
min_value <- min(amazon_data$boughtInLastMonth)  
max_value <- max(amazon_data$boughtInLastMonth)  
amazon_data$boughtInLastMonth_normalized <- (amazon_data$boughtInLastMonth -  
min_value) / (max_value - min_value)
```

e. Converting Category to Factors

```
# Convert 'categoryName' to factors for analysis  
amazon_data$categoryName <- as.factor(amazon_data$categoryName)
```

f. Dropping redundant and unnecessary columns

```
# Dropping redundant or unnecessary columns  
amazon_data <- select(amazon_data, -c(imgUrl, productURL))
```

**g. Saving the preprocessed data into new file as
“preprocessed_amazon_data.csv”**

```
# Save preprocessed dataset to a new file  
write.csv(amazon_data, file = "preprocessed_amazon_data.csv", row.names =  
FALSE)
```


6. Experiments

6.1 Exploratory Data Analysis (EDA) and Time Series Analysis

Finding Trending Product Categories and Sales Performance (Done by Ashwani Kumar)

We can use Exploratory Data Analysis which is the initial step that we can perform to understand our dataset patterns based on our task objective. Here, we will find the top 10 trending product categories in terms of sales using exploratory data analysis techniques. Exploratory Data Analysis allows for quick and insightful assessment of the high-performing categories with their total sales volume.

On the other hand, Time Series Analysis provides a deeper understanding of the sales trends over time for each category. For this, we will implement the ARIMA (Autoregressive Integrated Moving Average) modeling technique to forecast and identify possible trends or seasonal variations in sales.

The dataset does not have explicit dates or time intervals but sales trends over this can be examined by creating time series objects with a frequency of 12 (monthly data) and using the `ts()` function to create them. This analysis allows us for a deeper understanding of how sales volume fluctuates over time within each category providing valuable insights into seasonal trends and potential forecasting opportunities for these high-performing categories.

R Code:

Calculate total sales for each product category

```
sales_by_category <- amazon_data %>%  
  group_by(categoryName) %>%  
  summarise(total_sales = sum(boughtInLastMonth))
```

Identify top 10 trending product categories based on total sales

```
top_10_trending_categories <- sales_by_category %>%  
  top_n(10, wt = total_sales) %>%  
  arrange(desc(total_sales))
```

Visualize top 10 trending product categories

```
ggplot(top_10_trending_categories, aes(x = reorder(categoryName,
total_sales), y = total_sales)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(
    title = "Top 10 Trending Product Categories by Total Sales",
    x = "Category",
    y = "Total Sales"
  ) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Now, we perform time-series analysis on the sales trend for each category among the top 10 trending categories in a dataset.

As we do not have date and time interval variables in a dataset, we create a time series object using the `ts()` function specifying a frequency of 12 which could imply monthly data in this context.

R Code:

Time-series analysis for each category using ARIMA

```
for (category in top_10_trending_categories$categoryName) {
  category_data <- amazon_data %>%
    filter(categoryName == category) %>%
    select(boughtInLastMonth)
```

Create a time series object

```
ts_category <- ts(category_data$boughtInLastMonth, frequency = 12)
```

ARIMA modeling

```
arima_model <- auto.arima(ts_category)
```

Plotting ARIMA forecast

```
plot(forecast(arima_model), main = paste("Time-Series Analysis for",
category))
}
```

6.2 Regression Analysis

Analyzing Customer Ratings for Best Seller Products (Done by Arun Rajput)

Here, we are analyzing customer ratings for best seller products using Regression Technique. Using Regression, we can understand the relationships between variables and how one variable impact another. The main task of the analysis is to conduct using a linear regression method that models the relationship between the target variable (boughtInLastMonth) and the predictor variables (stars, reviews, and price). The regression technique helps in quantifying the impact of customer ratings (stars and reviews) on product sales. The regression analysis's coefficients will show how much and in which direction each variable affects the total number of products sold.

R Code:

Filtering only Best Seller Products

```
best_sellers_data <- amazon_data %>%  
  filter(iffelse(isBestSeller, "TRUE", "FALSE") == "TRUE")
```

Selecting only Relevant Columns

```
regression_data <- best_sellers_data %>%  
  select(stars, reviews, price, boughtInLastMonth)
```

Handling Missing Values

```
sum(is.na(regression_data))  
  
regression_data$price[is.na(regression_data$price)] <-  
mean(regression_data$price, na.rm = TRUE)
```

Performing Linear Regression

```
sales_prediction_model <- lm(boughtInLastMonth ~ stars + reviews + price,  
data = regression_data)  
  
summary(sales_prediction_model)
```

Scatter plots for stars vs boughtInLastMonth

```
ggplot(regression_data, aes(x = stars, y = boughtInLastMonth)) +  
  geom_point() +  
  labs(title = "Stars vs BoughtInLastMonth", x = "Stars", y = "Bought In  
Last Month")
```

Scatter plots for reviews vs boughtInLastMonth

```
ggplot(regression_data, aes(x = reviews, y = boughtInLastMonth)) +  
  geom_point() +  
  geom_smooth(method = "lm", se = FALSE, color = "blue") +  
  labs(title = "Linear Regression: Reviews vs BoughtInLastMonth", x =  
    "Reviews", y = "Bought In Last Month")
```

6.3 K-means Clustering

Identifying niche categories for high sales (Done by Dinesh Thapa)

By using K-means clustering, we are grouping a dataset into a set of k groups. We are clustering them based on their category and relation with sales volume. The K-means technique looks for patterns or similarities in the attributes of the dataset with a focus on the relationship between product categories and the sales volumes that correspond with them. After the clustering process, the clustered categories will be displayed according to their sales performance using visualization techniques. This analysis involves analyzing the characteristics of each cluster identifying top product categories within each cluster based on their sales volume understanding the similarities and differences between clusters. This technique provides valuable insights into the distinct market segments or categories that offer high sales performance.

R Code:

Performing K-means clustering based on sales performance

```
num_clusters <- 3  
sales_data <- sales_by_category[, -1]  
scaled_sales_data <- scale(sales_data)  
kmeans_model <- kmeans(scaled_sales_data, centers = num_clusters)  
sales_by_category$cluster <- as.factor(kmeans_model$cluster)
```

Visualize clustered categories based on sales performance

```
ggplot(sales_by_category, aes(x = reorder(categoryName, total_sales), y =  
total_sales, fill = cluster)) +  
  geom_bar(stat = "identity") +  
  labs(  
    title = "Clustered Categories based on Sales Performance",  
    x = "Category",  
    y = "Total Sales",  
    fill = "Cluster"  
  ) +  
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Analyzing Clustering

```
cluster_sizes <- table(sales_by_category$cluster)  
print(cluster_sizes)  
  
cluster_centers <- data.frame(Cluster = 1:num_clusters,  
kmeans_model$centers)  
print(cluster_centers)  
  
top_products_by_cluster <- sales_by_category %>%  
  group_by(cluster, categoryName) %>%  
  summarise(total_sales = sum(total_sales)) %>%  
  arrange(cluster, desc(total_sales)) %>%  
  top_n(5, total_sales)  
print(top_products_by_cluster)
```

7. Analysis and Results

Exploratory Data Analysis and Time Series Analysis to find Trending Product Categories and Sales Performance.

Figure 9: Top 10 Trending Product Categories by Total Sales

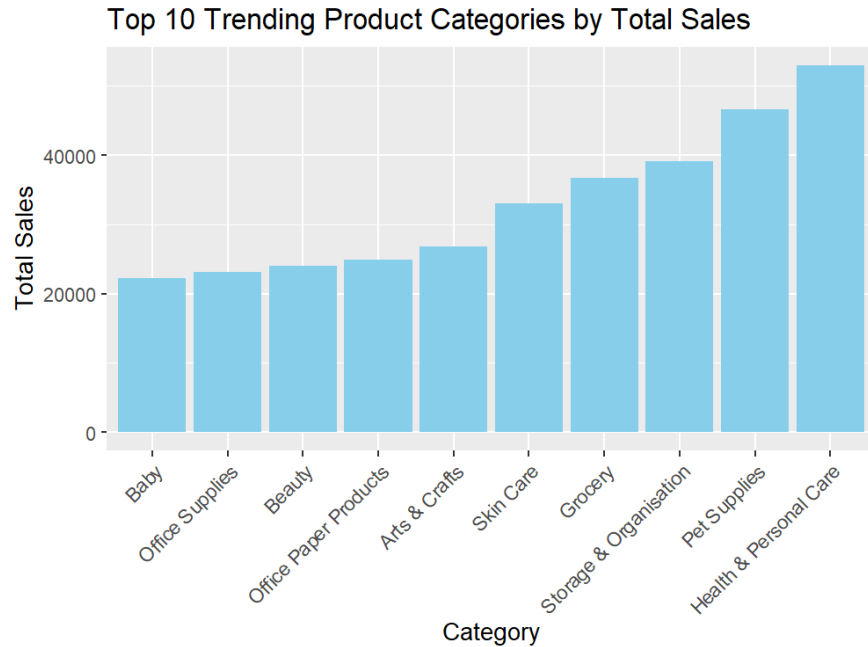


Table 2: Top 10 Product Categories with their Total Sales Volume

categoryName	total sales
Health & Personal Care	52988.
Pet Supplies	46587.
Storage & Organisation	39115.
Grocery	36756.
Skin Care	33001.
Arts & Crafts	26756.
Office Paper Products	24952.
Beauty	23969.
Office Supplies	23173.
Baby	22214.

The category with the highest sales seems to be "Health & Personal Care," which is followed by "Office Paper Products", "Beauty", "Pet Supplies", "Storage & Organization", "Grocery", "Skin Care", "Arts & Crafts", and "Baby".

In the above bar diagram clearly shows the Top 10 Trending Categories with their total sales volume.

This analysis can guide marketing and strategies to focus on these high-performing categories for marketing and sales campaigns, inventory management, and product recommendations to maximize sales of the company.

Regression Analysis to analyze Customer Ratings for Best Seller Products

Figure 10: Star Vs BoughtInLastMonth - Regression Analysis

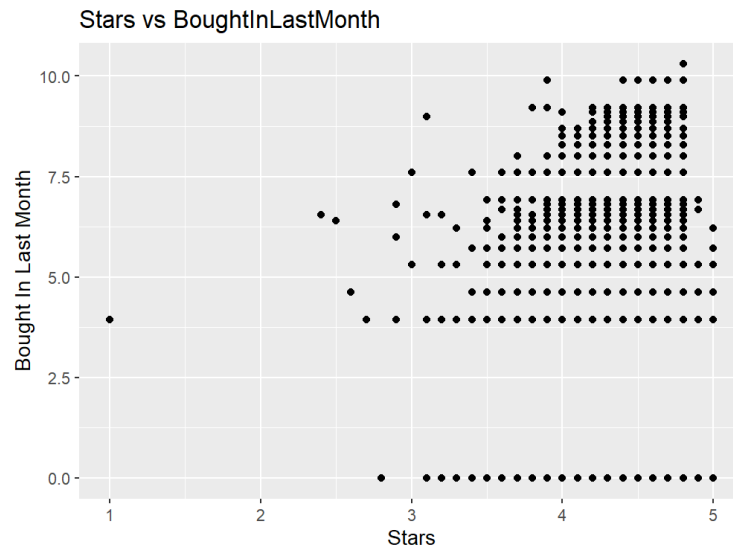
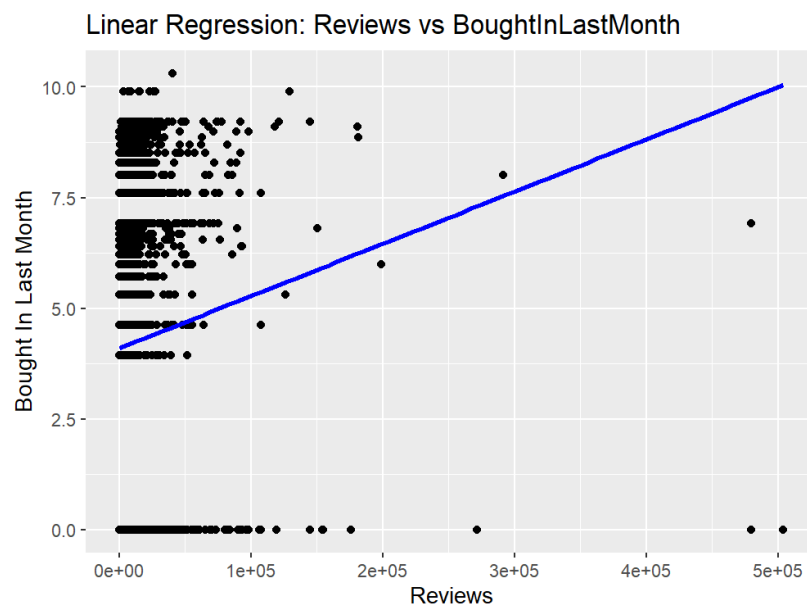


Figure 11: Linear Regression: Reviews Vs BoughtInLastMonth



Coefficients:

Intercept (0.6006): The intercept represents the expected value of 'boughtInLastMonth' when all other predictors (stars, reviews, price) are zero.

stars (0.9537): For every unit increase in stars, 'boughtInLastMonth' is expected to increase by approximately 0.9537 units, holding other variables constant.

reviews (0.00001054): For every additional review, 'boughtInLastMonth' is expected to increase by approximately 0.00001054 units, holding other variables constant.

price (-0.04104): With every unit increase in price, 'boughtInLastMonth' is expected to decrease by approximately 0.04104 units, holding other variables constant.

Residuals and Error Metrics:

Residuals: The differences between observed and predicted values of 'boughtInLastMonth'. These differences range from -10.04 to 7.83 units, indicating the model's variability in predicting sales.

Residual Standard Error (3.031): Measures the average amount that the model's predictions deviate from the actual values.

Multiple R-squared (0.05703): Indicates the proportion of variance in 'boughtInLastMonth' that is explained by the predictors (stars, reviews, price). Only about 5.7% of the variability in sales can be explained by these predictors.

Adjusted R-squared (0.05654): A corrected version of R-squared that adjusts for the number of predictors in the model.

Significance:

The p-values for stars, reviews, and price are all < 0.001 , indicating that these predictors are statistically significant in explaining the variation in sales.

Discussion:

Impact of Predictors: The analysis suggests that customer ratings (stars and reviews) have a positive impact on sales, with higher ratings likely to lead to higher sales. Conversely, the price has a negative impact, where higher prices may lead to decreased sales.

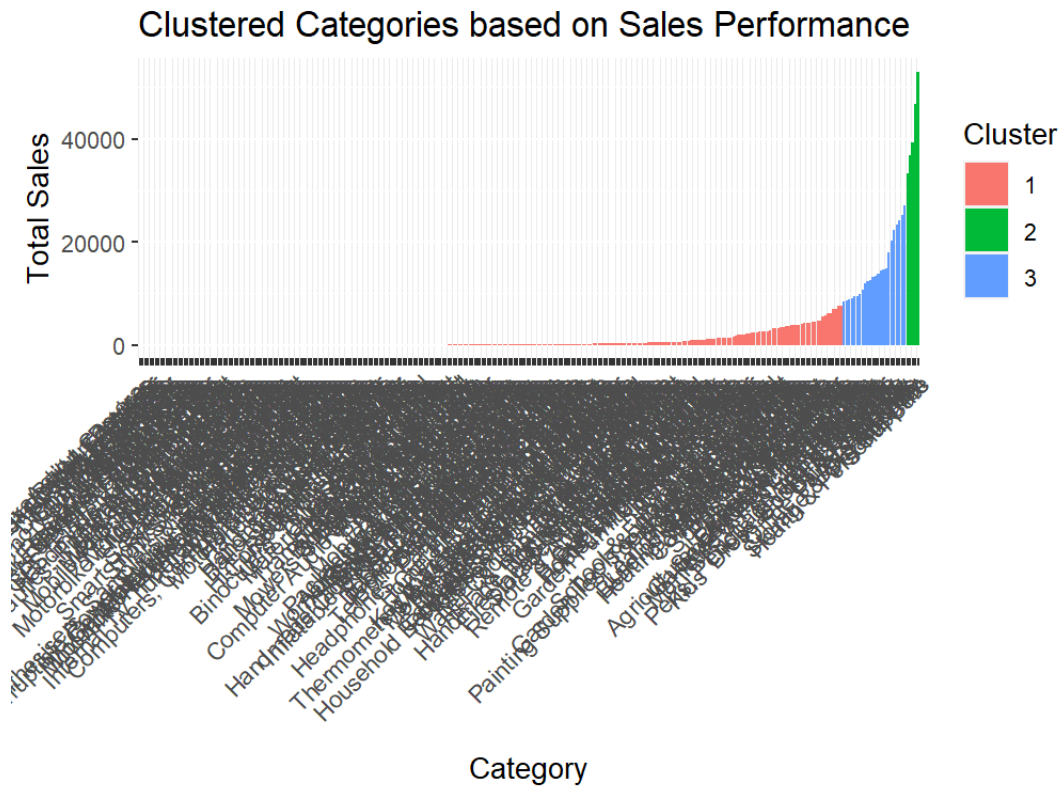
Model Fit: However, the model's overall fit is relatively low (R-squared is around 0.05703), indicating that these predictors alone might not sufficiently explain the variation in sales performance.

Further Considerations: Other factors not included in the model could significantly impact sales, such as marketing strategies, product uniqueness, market demand, etc.

K-means Clustering to identify niche categories for high sales

Visualization

Figure 12: K-Means Clustering showing clustered categories based on sales performance



Cluster Sizes:

Cluster 1: 267 categories

Cluster 2: 5 categories

Cluster 3: 24 categories

Cluster Centers (Average Sales):

Cluster 1: -0.2678984

Cluster 2: 5.7298052

Cluster 3: 1.7866600

Top 5 Product Categories by Sales in Clusters:

Cluster 1:

Light Bulbs - 7671 units

Baby & Toddler Toys - 7557 units

Fragrances - 7071 units

Indoor Lighting - 7034 units

Sports Toys & Outdoor - 6163 units

Cluster 2:

Health & Personal Care - 53027 units

Pet Supplies - 46772 units

Storage & Organization - 39368 units

Grocery - 36870 units

Skin Care - 33214 units

Cluster 3:

Arts & Crafts - 27015 units

Office Paper Products - 25234 units

Beauty - 24114 units

Office Supplies - 23317 units

Baby - 22372 units

Insights and Discussions:

Cluster 1 is the largest cluster containing a significant number of categories compared to clusters 2 and 3.

Cluster 2 contains significantly higher average sales compared to the other clusters.

Different market segments and preferences are indicated within each cluster by distinct top-selling product categories. Health & Personal Care, Pet Supplies, and Grocery are included in Cluster 2 indicating that these categories have a major impact on the higher sales in this cluster.

Cluster 3 shows interest in categories like Arts and crafts, Beauty, and Baby products, reflecting another set of preferences.

Finally, using K-means clustering, we identified niche categories for high sales. This gives meaningful insights to E-commerce businesses so that they can focus on high-selling categories for their marketing campaigns to increase their sales. Furthermore, this will also help them in managing their inventory, customer engagement, and their preferences.

8. Conclusion

The Amazon UK Data Mining Project has provided a vast understanding of the e-commerce industry through insightful data mining and analytics to revolutionize its marketing and sales strategies. The exploration of the Amazon products dataset uncovered valuable insights into trending product categories, customer preferences, and sales trends and patterns. Through exploratory data analysis, time-series analysis, regression modeling, and k-means clustering, we examined trends, identified top-performing product categories, and revealed customer behavior patterns.

The analysis highlighted the Health & Personal Care category as the leader in the sales volume closely followed by Office Paper Products, Beauty, Pet Supplies, and Storage & Organization. In addition, regression analysis showed the impact of product ratings, reviews, and pricing on sales demonstrating the significance of customer satisfaction and competitive pricing strategies in driving sales performance.

The k-means clustering approach divided categories into distinct clusters showcasing varying sales patterns and consumer preferences. Cluster 3 represented preferences for products related to arts and crafts, beauty, and babies, while Cluster 2 included categories like groceries, pet supplies, and health and personal care.

Our findings show the importance of data mining in helping to create more data-driven and effective marketing and sales strategies, inventory management, and personalized customer experience through targeting audiences for their niche product categories. By using these insights, e-commerce platforms can optimize their operations, tailor their marketing campaigns, and enhance their customer engagement to boost sales performance.

In conclusion, this transformative potential of data mining analysis provides a roadmap for Amazon UK and similar platforms a path to success through informed data-driven decisions, targeted marketing, and enhanced customer experiences.

9. References

Alghanam, O.A., Al-Khatib, S.N. and Hiari, M.O. (2022). Data Mining Model for Predicting Customer Purchase Behavior in E-Commerce Context. *International Journal of Advanced Computer Science and Applications*, 13(2). doi:<https://doi.org/10.14569/ijacsa.2022.0130249>.

Bejju, A. (2016). Sales Analysis of E-Commerce Websites using Data Mining Techniques. *International Journal of Computer Applications*, 133(5), pp.36–40. doi:<https://doi.org/10.5120/ijca2016907812>.

Zhang, J. and Li, J. (2016). Retail Commodity Sale Forecast Model Based on Data Mining. [online] IEEE Xplore. doi:<https://doi.org/10.1109/INCoS.2016.42>.

Chen, Y. and Wang, F. (2009). A Dynamic Pricing Model for E-commerce Based on Data Mining. 2009 Second International Symposium on Computational Intelligence and Design. doi:<https://doi.org/10.1109/iscid.2009.99>.

10. Appendix: Complete R Code

Data Description Code:

```
getwd()

setwd("D:/MSc Big Data/Data Mining/Dataset")

library(tidyverse)

amazon_data <- read.csv("D:/MSc Big Data/Data
Mining/Dataset/Amazon_UK_Products_Dataset_2023.csv")

dim(amazon_data)

summary(amazon_data)

str(amazon_data)

unique(amazon_data$isBestSeller)

unique(amazon_data$categoryName)

head(amazon_data)

amazon_data[c("asin", "title", "stars", "price", "categoryName")]

summary(amazon_data$stars)

summary(amazon_data$price)

summary(amazon_data$reviews)

summary(amazon_data$boughtInLastMonth)

colSums(is.na(amazon_data))

length(unique(amazon_data$asin))

table(amazon_data$categoryName)

table(amazon_data$isBestSeller)

plot(density(amazon_data$stars), main = "Density Plot of Stars", xlab =
"Stars")

correlation_matrix <- cor(select(amazon_data, stars, reviews, price,
boughtInLastMonth))

print(correlation_matrix)

corrplot::corrplot(correlation_matrix, method = "circle")

boxplot(price ~ isBestSeller, data = amazon_data, main = "Price by BestSeller
Status", xlab = "BestSeller Status", ylab = "Price in GBP")
```

```

boxplot(reviews ~ isBestSeller, data = amazon_data, main = "Reviews by
BestSeller Status", xlab = "BestSeller Status", ylab = "Number of Reviews")

aggregate(price ~ categoryName, data = amazon_data, FUN = summary)

aggregate(reviews ~ categoryName, data = amazon_data, FUN = summary)

```

Data Preprocessing Code:

```

getwd()

setwd("D:/MSc Big Data/Data Mining/Dataset")

library(dplyr)

library(tidyr)

amazon_data <- read.csv("D:/MSc Big Data/Data
Mining/Dataset/Amazon_UK_Products_Dataset_2023.csv")

amazon_data$stars[amazon_data$stars == 0] <- NA
amazon_data$reviews[amazon_data$reviews == 0] <- NA
amazon_data$price[amazon_data$price == 0] <- NA

Q1 <- quantile(amazon_data$price, 0.25, na.rm = TRUE)
Q3 <- quantile(amazon_data$price, 0.75, na.rm = TRUE)
IQR_value <- Q3 - Q1
upper_bound <- Q3 + 1.5 * IQR_value
amazon_data <- filter(amazon_data, price <= upper_bound)

amazon_data$review_sentiment <- ifelse(amazon_data$reviews > 50, "Positive",
"Neutral")

amazon_data$boughtInLastMonth <- log(amazon_data$boughtInLastMonth + 1)

min_value <- min(amazon_data$boughtInLastMonth)

```



```

max_value <- max(amazon_data$boughtInLastMonth)

amazon_data$boughtInLastMonth_normalized <- (amazon_data$boughtInLastMonth -
min_value) / (max_value - min_value)

amazon_data$categoryName <- as.factor(amazon_data$categoryName)

amazon_data <- select(amazon_data, -c(imgUrl, productURL))

write.csv(amazon_data, file = "preprocessed_amazon_data.csv", row.names =
FALSE)

```

Exploratory Data Analysis and Time Series Analysis

```

getwd()

setwd("D:/MSc Big Data/Data Mining/Dataset")

library(tidyverse)

library(ggplot2)

library(forecast)

amazon_data <- read.csv("D:/MSc Big Data/Data
Mining/Dataset/preprocessed_amazon_data.csv")

amazon_data <- amazon_data %>%
  na.omit()

sales_by_category <- amazon_data %>%
  group_by(categoryName) %>%
  summarise(total_sales = sum(boughtInLastMonth))

top_10_trending_categories <- sales_by_category %>%
  top_n(10, wt = total_sales) %>%

```

```

    arrange(desc(total_sales))

ggplot(top_10_trending_categories, aes(x = reorder(categoryName,
total_sales), y = total_sales)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(
    title = "Top 10 Trending Product Categories by Total Sales",
    x = "Category",
    y = "Total Sales"
  ) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

top_10_trending_categories

for (category in top_10_trending_categories$categoryName) {
  category_data <- amazon_data %>%
    filter(categoryName == category) %>%
    select(boughtInLastMonth)

  ts_category <- ts(category_data$boughtInLastMonth, frequency = 12)

  arima_model <- auto.arima(ts_category)

  plot(forecast(arima_model), main = paste("Time-Series Analysis for",
category))
}

```

Regression Analysis

```
getwd()

setwd("D:/MSc Big Data/Data Mining/Dataset")

amazon_data <- read.csv("D:/MSc Big Data/Data
Mining/Dataset/preprocessed_amazon_data.csv")

library(tidyverse)
library(ggplot2)

best_sellers_data <- amazon_data %>%
  filter(ifelse(isBestSeller, "TRUE", "FALSE") == "TRUE")

regression_data <- best_sellers_data %>%
  select(stars, reviews, price, boughtInLastMonth)

sum(is.na(regression_data))

regression_data$price[is.na(regression_data$price)] <-
mean(regression_data$price, na.rm = TRUE)

sales_prediction_model <- lm(boughtInLastMonth ~ stars + reviews + price,
data = regression_data)

summary(sales_prediction_model)

ggplot(regression_data, aes(x = stars, y = boughtInLastMonth)) +
  geom_point() +
```

```
labs(title = "Stars vs BoughtInLastMonth", x = "Stars", y = "Bought In Last Month")
```

```
ggplot(regression_data, aes(x = reviews, y = boughtInLastMonth)) +  
  geom_point() +  
  geom_smooth(method = "lm", se = FALSE, color = "blue") +  
  labs(title = "Linear Regression: Reviews vs BoughtInLastMonth", x =  
"Reviews", y = "Bought In Last Month")
```

```
predicted_sales <- predict(sales_prediction_model, newdata = regression_data)
```

```
rsquared <- summary(sales_prediction_model)$r.squared  
cat("R-squared ( $R^2$ ):", rsquared, "\n")
```

```
rmse <- sqrt(mean((regression_data$boughtInLastMonth - predicted_sales)^2))  
cat("Root Mean Squared Error (RMSE):", rmse, "\n")
```

```
mae <- mean(abs(regression_data$boughtInLastMonth - predicted_sales))  
cat("Mean Absolute Error (MAE):", mae, "\n")
```

K-means Clustering

```
getwd()  
setwd("D:/MSc Big Data/Data Mining/Dataset")  
amazon_data <- read.csv("D:/MSc Big Data/Data  
Mining/Dataset/preprocessed_amazon_data.csv")  
amazon_data
```

```
library(tidyverse)  
library(cluster)  
library(arules)
```

```

cleaned_data <- amazon_data %>%
  select(categoryName, boughtInLastMonth) %>%
  na.omit()

sales_by_category <- cleaned_data %>%
  group_by(categoryName) %>%
  summarise(total_sales = sum(boughtInLastMonth))

num_clusters <- 3
sales_data <- sales_by_category[, -1]

scaled_sales_data <- scale(sales_data)

kmeans_model <- kmeans(scaled_sales_data, centers = num_clusters)

sales_by_category$cluster <- as.factor(kmeans_model$cluster)

ggplot(sales_by_category, aes(x = reorder(categoryName, total_sales), y =
total_sales, fill = cluster)) +
  geom_bar(stat = "identity") +
  labs(
    title = "Clustered Categories based on Sales Performance",
    x = "Category",
    y = "Total Sales",
    fill = "Cluster"
  ) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

cluster_sizes <- table(sales_by_category$cluster)

```

```
print(cluster_sizes)

cluster_centers <- data.frame(Cluster = 1:num_clusters, kmeans_model$centers)
print(cluster_centers)

top_products_by_cluster <- sales_by_category %>%
  group_by(cluster, categoryName) %>%
  summarise(total_sales = sum(total_sales)) %>%
  arrange(cluster, desc(total_sales)) %>%
  top_n(5, total_sales)
print(top_products_by_cluster)
```