Online Retail Sales

By Carolina Salas

Project introduction:

The primary objective of this project is to comprehensively assess sales performance over different periods and across various regions, including country, state, and city. This involves identifying the top-selling products and categories to highlight customer preferences and market demand. Additionally, the analysis will examine sales trends over time to uncover patterns and fluctuations in sales performance. By evaluating the impact of different campaign schemas on sales, the project aims to determine the effectiveness of marketing strategies and their influence on customer purchasing behavior. This multi-faceted approach will provide valuable insights into regional and temporal sales dynamics, aiding data-driven decision-making and strategic planning.

Primary objective:

The primary objective of this analysis is to comprehensively assess the sales performance over different periods and across various regions, including country, state, and city. This involves identifying the top-selling products and categories, thereby highlighting customer preferences and market demand. Additionally, the analysis will examine sales trends over time, to uncover patterns and fluctuations in sales performance. By evaluating the impact of different campaign schemas on sales, the analysis aims to determine the effectiveness of marketing strategies and their influence on customer purchasing behavior. This multi-faceted approach will provide valuable insights into regional and temporal sales dynamics, helping to inform data-driven decision-making and strategic planning.

Overview of Dataset:

The dataset provides a comprehensive overview of customer orders, encompassing various aspects of the purchasing process and customer demographics. Each order is uniquely identified by order_id and is associated with a specific customer identified by customer_id. Customer details include name, gender, age, credit score, and monthly income, along with geographic information such as country, state, and city.

Marketing and product-related details include the campaign schema linked to the order, the Category of the product, and the specific Product ordered. Financial aspects of the order are captured through variables like Cost, Price, and quantity of the product ordered. The dataset also tracks the progress and status of each order with order confirmation status, timestamps for cart addition time and order confirmation time, the payment method used, and order return status. Finally, the dataset calculates the Total Revenue generated from each order, providing valuable insights into the financial performance and customer behavior.

Variables:

- order_id: A unique identifier for each order.
- customer_id: A unique identifier for each customer.
- name: The name of the customer.
- gender: The gender of the customer.
- age: The age of the customer.
- credit score: The credit score of the customer.
- monthly income: The monthly income of the customer.
- country: The country where the customer resides.
- state: The state where the customer resides.
- city: The city where the customer resides.
- campaign schema: The marketing campaign associated with the order.
- Category: The category to which the ordered product belongs.
- Product: The specific product ordered by the customer.
- Cost: The cost price of the product.
- Price: The selling price of the product.
- quantity: The number of units of the product ordered.
- order confirmation: The status of order confirmation.
- cart addition time: The timestamp when the product was added to the cart.
- order confirmation time: The timestamp when the order was confirmed.
- payment method: The method used by the customer to pay for the order.
- order return: The status indicating whether the order was returned.
- Total Revenue: The total revenue generated from the order.

Analysis & Visualization

Question 1: Include summary statistics for price, monthly income and credit score. Create a box plot for each variables and analyze your findings

TABLE 1: Summary Statistics for Price, Monthly Income & Credit Score

Price		*monthly_income*		*credit_score*	
Mean	$ 204.59	Mean	$ 50,230.57	Mean	709.2297872
Standard Error	$ 11.86	Standard Error	$ 415.26	Standard Error	1.334896312
Median	$ 50.00	Median	$ 57,602.50	Median	713.5
Mode	$ 50.00	Mode	$ 58,533.00	Mode	757
Standard Deviation	$ 374.94	Standard Deviation	$ 12,731.78	Standard Deviation	40.92714593
Sample Variance	$ 140,580.07	Sample Variance	$ 162,098,097.47	Sample Variance	1675.031274
Kurtosis	$ 4.95	Kurtosis	$ 0.48	Kurtosis	-0.155831883
Skewness	$ 2.47	Skewness	$ (1.30)	Skewness	-0.768294576
Range	$ 1,492.00	Range	$ 45,485.00	Range	160
Minimum	$ 8.00	Minimum	$ 15,407.00	Minimum	600
Maximum	$ 1,500.00	Maximum	$ 60,892.00	Maximum	760
Sum	$ 204,593.00	Sum	$ 47,216,735.00	Sum	666676
Count	$ 1,000.00	Count	$ 940.00	Count	940

TABLE 2: Price Box Plot

TABLE 3: Monthly Income Box Plot

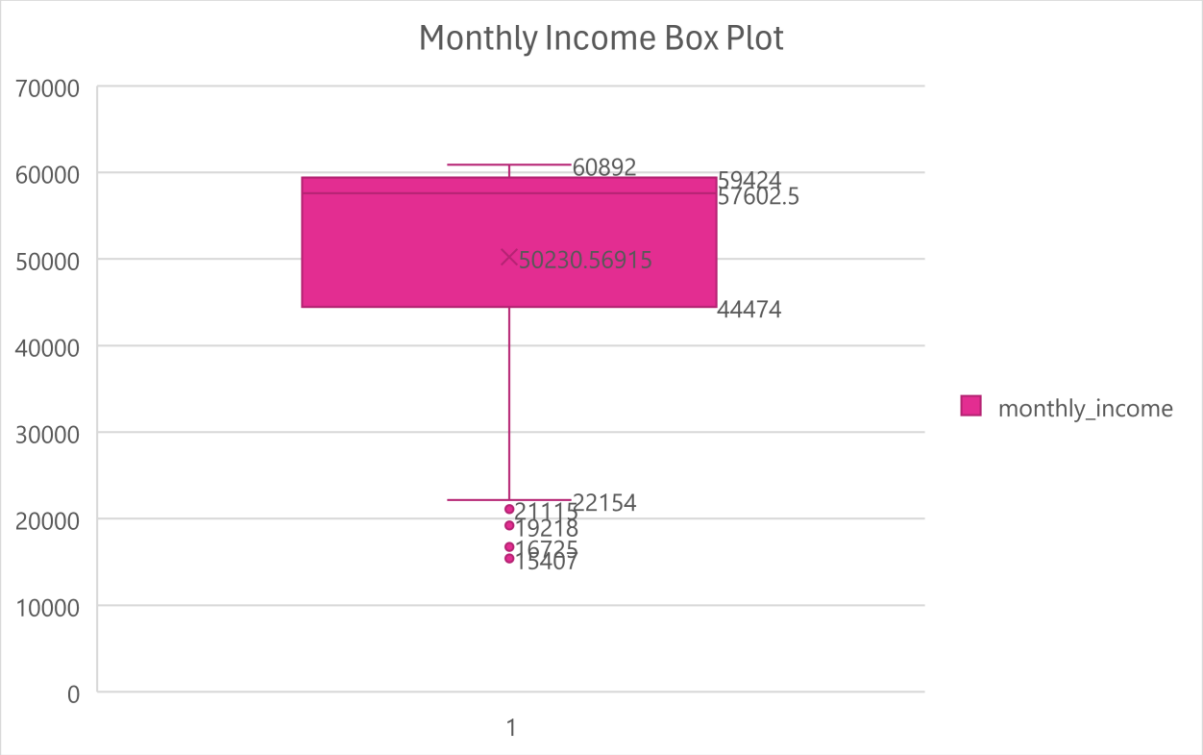
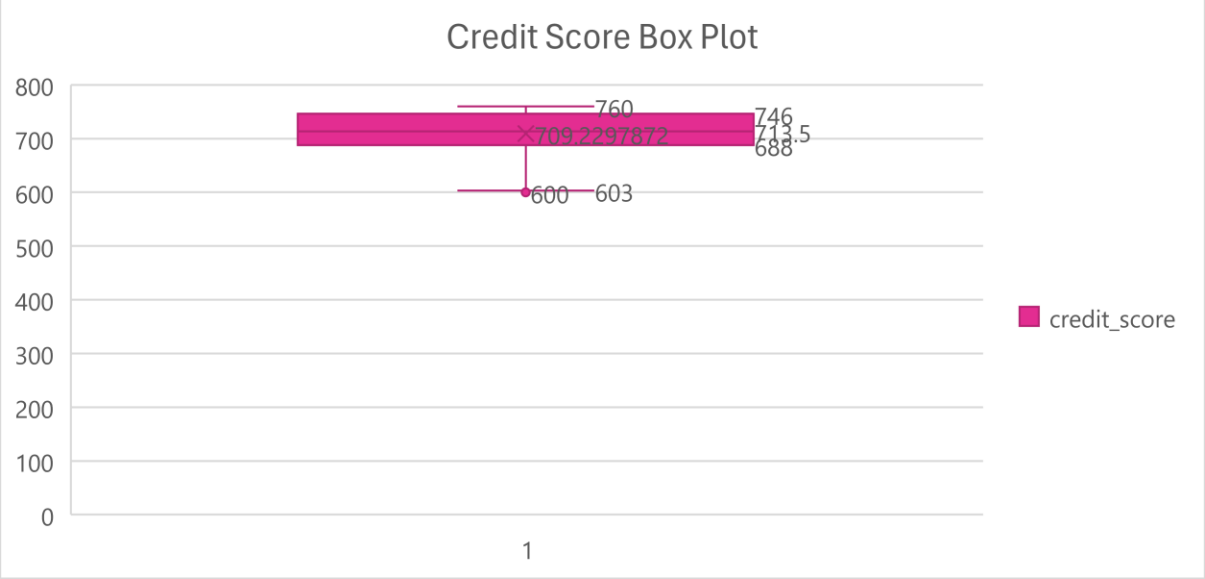


TABLE 4: Credit Score Box Plot



Analysis –

The analysis begins with Price, where the dataset reveals a mean of $204.59 and a standard deviation of $374.94, indicating a wide range of product pricing. The median and mode are both $50.00, reflecting a frequent price point among orders. The price distribution exhibits positive skewness (2.47) and high kurtosis (4.95), suggesting a tendency towards higher values with some extreme outliers, such as one at $1,500.00. The box plot illustrates this distribution: the lower quartile (Q1) is $25.00, the median (Q2) is $50.00, and the upper quartile (Q3) is $97.50, with the outlier noted at $1,500.00.

Moving to Monthly Income, the dataset shows a mean income of $50,230.57 and a standard deviation of $12,731.78, with a median income of $57,602.50 and a mode of $58,533.00. Income ranges from $15,407.00 to $60,892.00, with negative skewness (-1.30) and a slight left-skewed distribution (kurtosis = 0.48). The box plot outlines this distribution: Q1 stands at

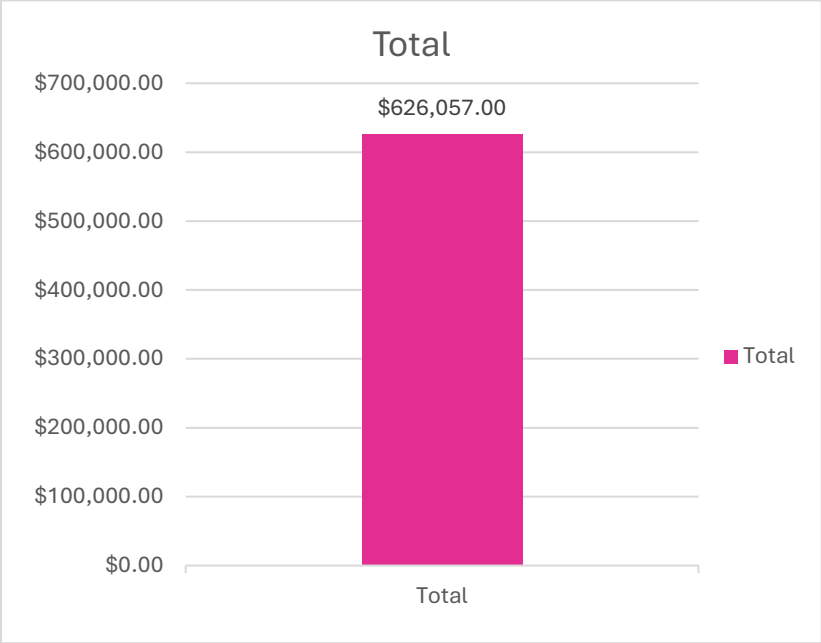
$44,474.00, the median at $57,602.50, and Q3 at $59,424.00, with an outlier observed at $15,407.00, indicating a significantly lower income value.

Lastly, examining Credit Score reveals a mean score of 709.23 and a standard deviation of 40.93, with a median of 713.5 and a mode of 757. Scores range from 600 to 760, displaying a relatively tight distribution without significant outliers. The skewness is negative (-0.77), and kurtosis is close to zero (-0.16), indicating a near-normal distribution. The box plot details this distribution: Q1 is 688, the median is 713.5, and Q3 is 746, illustrating the consistency and central tendency of credit scores within the dataset.

Question 2: What is the total sales revenue generated during the analysis period?

TABLE 5: Total Sales of Revenue

Sum of Total Revenue
$626,057.00



Analysis –

The total sales revenue generated during the analysis period amounts to $626,057.00. The pivot chart represents a cumulative revenue from customer order processed within dataset. A bar graph illustrating this total revenue shows a single bar labeled at $626,057.00, providing a clear visual representation of the overall sales performance for the period analyzed.

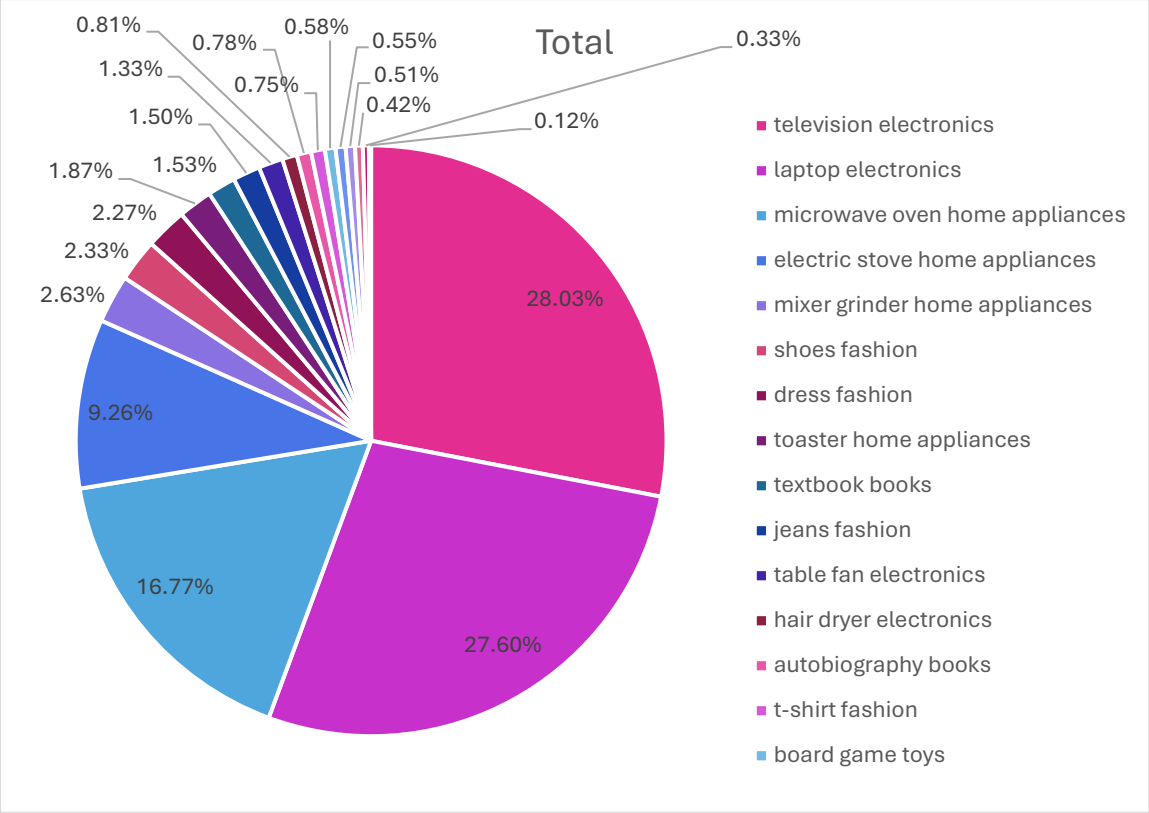
Question 3. Which products are the top sellers by revenue? Which categories are the top sellers by revenue?

TABLE 6: Product Category Distribution and Total Revenue Share

Row Labels	Sum of Total Revenue
television	**28.03%**
electronics	28.03%
laptop	**27.60%**
electronics	27.60%
microwave oven	**16.77%**
home appliances	16.77%
electric stove	**9.26%**
home appliances	9.26%
mixer grinder	**2.63%**
home appliances	2.63%
shoes	**2.33%**
fashion	2.33%
dress	**2.27%**
fashion	2.27%

toaster	**1.87%**
home appliances	1.87%
textbook	**1.53%**
books	1.53%
jeans	**1.50%**
fashion	1.50%
table fan	**1.33%**
electronics	1.33%
hair dryer	**0.81%**
electronics	0.81%
autobiography	**0.78%**
books	0.78%
t-shirt	**0.75%**
fashion	0.75%
board game	**0.58%**
toys	0.58%
novel	**0.55%**
books	0.55%
puzzle	**0.51%**
toys	0.51%
plush toy	**0.42%**
toys	0.42%
action figure	**0.33%**
toys	0.33%
magazine	**0.12%**
books	0.12%
Grand Total	**100.00%**

TABLE 7: Revenue Distribution by Product Category



Analysis-

The analysis reveals that televisions are the highest revenue-generating product, contributing 28.03% to the total sales. Following closely are laptops, which account for 27.60% of the revenue. Microwave ovens and electric stoves also make notable contributions, generating 16.77% and 9.26% of the total revenue, respectively. Additionally, smaller items like toasters contribute 1.87% to the overall revenue share.

In terms of categories, electronics emerge as the leading category by revenue, collectively accounting for 58.77% of the total sales. This is primarily driven by strong sales in televisions and laptops. Home appliances, encompassing microwave ovens, electric stoves, and toasters, follow with a combined revenue share of 28.90%. Fashion items, including shoes, dresses, jeans, and t-shirts, contribute 6.85% to total revenue. Books, covering textbooks, novels, autobiographies, and magazines, account for 3.98% of sales. Toys, such as plush toys, action figures, board games, and puzzles, make up 1.84% of the revenue.

The pie chart visually represents the distribution of total revenue across various product categories, revealing key insights into top-selling items based on their respective revenue shares. At the forefront, televisions lead with the largest slice, accounting for 28.03% of total revenue. Following closely, laptops occupy the second significant position with 27.60% of the revenue share. In the home appliances category, microwave ovens constitute the third-largest segment, contributing 16.77% to the total revenue. Electric stoves follow with 9.26% of the revenue share. Together, these top four categories dominate the pie chart, illustrating strong consumer demand for electronics and essential home appliances

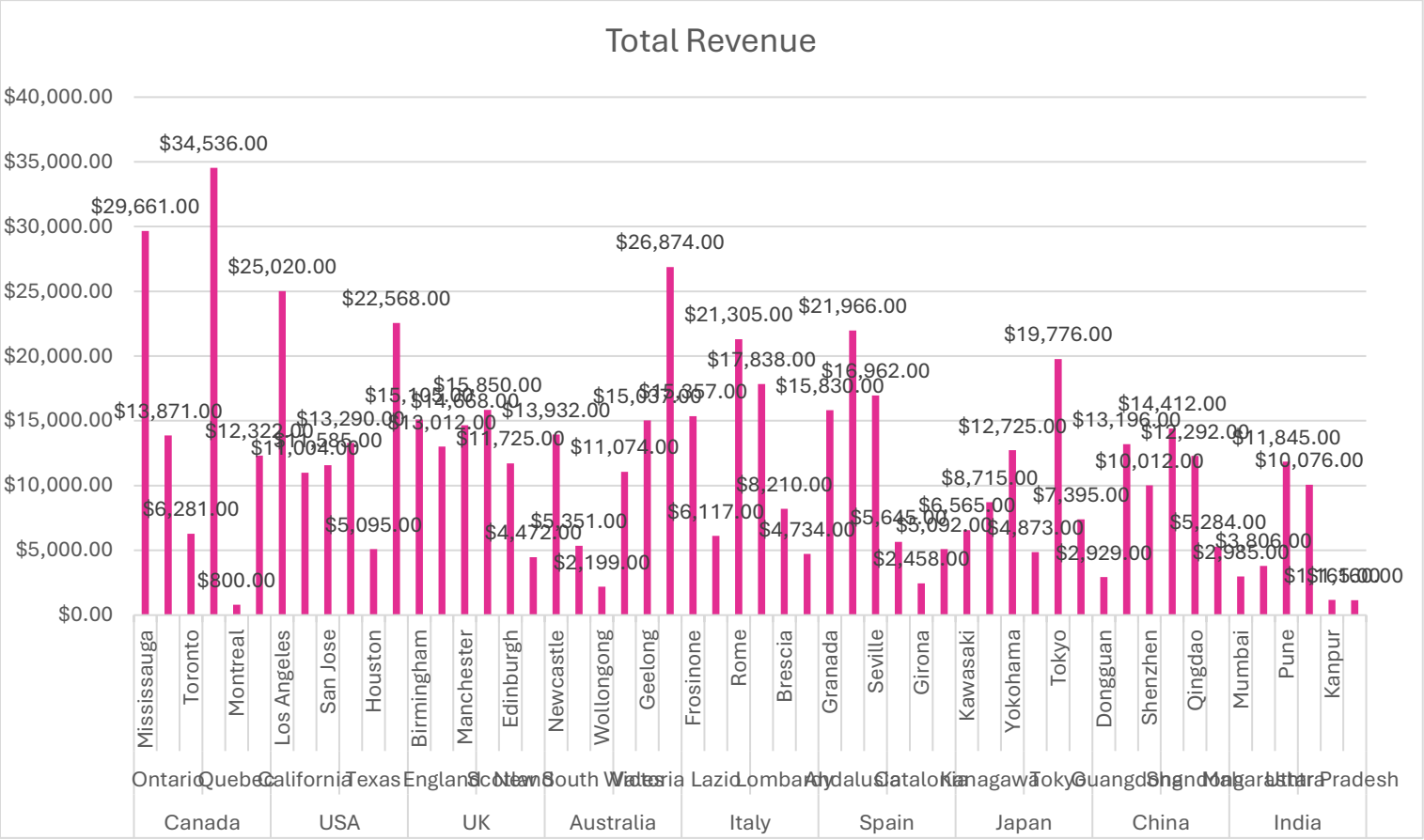
Question 4. How do sales differ across various regions (country, state, city)?

TABLE 8: Regional Sales

Row Labels	Sum of Total Revenue
Canada	**$97,471.00**
Ontario	**$49,813.00**
Mississauga	$29,661.00
Ottawa	$13,871.00
Toronto	$6,281.00
Quebec	**$47,658.00**
Laval	$34,536.00
Montreal	$800.00
Quebec City	$12,322.00
USA	**$88,562.00**
California	**$47,609.00**
Los Angeles	$25,020.00
San Diego	$11,004.00
San Jose	$11,585.00
Texas	**$40,953.00**
Dallas	$13,290.00
Houston	$5,095.00
San Antonio	$22,568.00
UK	**$74,832.00**
England	**$42,785.00**
Birmingham	$15,105.00
London	$13,012.00
Manchester	$14,668.00
Scotland	**$32,047.00**
Aberdeen	$15,850.00
Edinburgh	$11,725.00
Glasgow	$4,472.00
Australia	**$74,467.00**
New South Wales	**$21,482.00**
Newcastle	$13,932.00
Sydney	$5,351.00
Wollongong	$2,199.00
Victoria	**$52,985.00**
Ballarat	$11,074.00
Geelong	$15,037.00
Melbourne	$26,874.00
Italy	**$73,561.00**
Lazio	**$42,779.00**
Frosinone	$15,357.00
Latina	$6,117.00
Rome	$21,305.00
Lombardy	**$30,782.00**
Bergamo	$17,838.00
Brescia	$8,210.00
Milan	$4,734.00
Spain	**$67,953.00**
Andalusia	**$54,758.00**
Granada	$15,830.00
Malaga	$21,966.00
Seville	$16,962.00
Catalonia	**$13,195.00**
Barcelona	$5,645.00
Girona	$2,458.00
Tarragona	$5,092.00
Japan	**$60,049.00**
Kanagawa	**$28,005.00**
Kawasaki	$6,565.00
Sagamihara	$8,715.00
Yokohama	$12,725.00
Tokyo	**$32,044.00**
Kawasaki	$4,873.00
Tokyo	$19,776.00
Yokohama	$7,395.00
China	**$58,125.00**
Guangdong	**$26,137.00**
Dongguan	$2,929.00
Guangzhou	$13,196.00
Shenzhen	$10,012.00
Shandong	**$31,988.00**
Jinan	$14,412.00
Qingdao	$12,292.00
Yantai	$5,284.00

India	**$31,037.00**
Maharashtra	**$18,636.00**
Mumbai	$2,985.00
Nagpur	$3,806.00
Pune	$11,845.00
Uttar Pradesh	**$12,401.00**
Ghaziabad	$10,076.00
Kanpur	$1,165.00
Lucknow	$1,160.00
Grand Total	**$626,057.00**

TABLE 9: Regional Revenue Distribution Overview



Analysis

The "Regional Revenue Distribution Overview" chart offers a detailed insight into how sales vary across different regions, including countries, states, and cities. It provides a comprehensive summary of total revenue generated from various geographical locations, highlighting key regions where sales are particularly strong. In Canada, Ontario emerges as a significant contributor with $49,813.00, followed by Mississauga at $29,661.00 and Ottawa at $13,871.00. In the USA, California leads with $47,609.00 in revenue, driven by major cities like Los Angeles with $25,020.00 and San Diego with $11,004.00. Texas follows closely with $40,953.00, including notable contributions from San Antonio at $22,568.00. Across the Atlantic, the UK shows robust performance, with England generating $42,785.00, bolstered by Birmingham at $15,105.00 and Manchester at $14,668.00. Scotland contributes significantly with $32,047.00, including notable figures from Aberdeen at $15,850.00. Australia exhibits strong sales, particularly in Victoria with $52,985.00, driven by Melbourne at $26,874.00, and New South Wales with $21,482.00, including Newcastle at $13,932.00. In Europe, Italy demonstrates solid performance, with Lazio at $42,779.00 and Lombardy at $30,782.00, including significant contributions from Rome at $21,305.00. Spain showcases notable revenue from Andalusia with $54,758.00, featuring Granada at $15,830.00 and Malaga at $21,966.00. In Asia, Japan's sales are strong, with Kanagawa at $28,005.00 and Tokyo at $32,044.00, including Yokohama at $12,725.00. Lastly, China contributes substantially, with Guangdong at $26,137.00, including Guangzhou at $13,196.00 and Shenzhen at $10,012.00.

In my analysis of the regional revenue distribution overview pivot table, several key findings stand out across different countries and regions. In Canada, the city of Laval in Quebec shows the highest revenue of $34,536. Moving to the USA, Los

Angeles in California leads with the highest revenue of $25,020. In the UK, Aberdeen in Scotland generates a significant total revenue of $15,850. Australia sees Victoria, specifically Melbourne, contributing the highest revenue of $26,874. In Italy, Rome in Lazio achieves the highest total revenue at $21,305. Meanwhile, in Spain, Malaga in Andalusia stands out with the highest revenue of $21,966. In Japan, Tokyo records the highest revenue of $19,776. Shandong in China, specifically Jinan, sees the highest total revenue of $14,412. Lastly, in India, Maharashtra, with Pune as its city, achieves a total revenue of $11,845. These insights underscore the regional variations in sales performance, highlighting specific cities and regions that drive significant revenue within their respective countries.

Question 5.

How do different campaign schemas impact sales revenue? Which campaign schema results in the highest average order value?

Table 10: Campaign Impact on Total Revenue Distribution

Row Labels	Sum of Total Revenue
Billboard-QR code	14.98%
E-mails	15.61%
Facebook-ads	14.28%
Google-ads	17.92%
Instagram-ads	19.81%
Twitter-ads	17.40%
Grand Total	**100.00%**

Table 11: Campaign Impact on Total Revenue Distribution

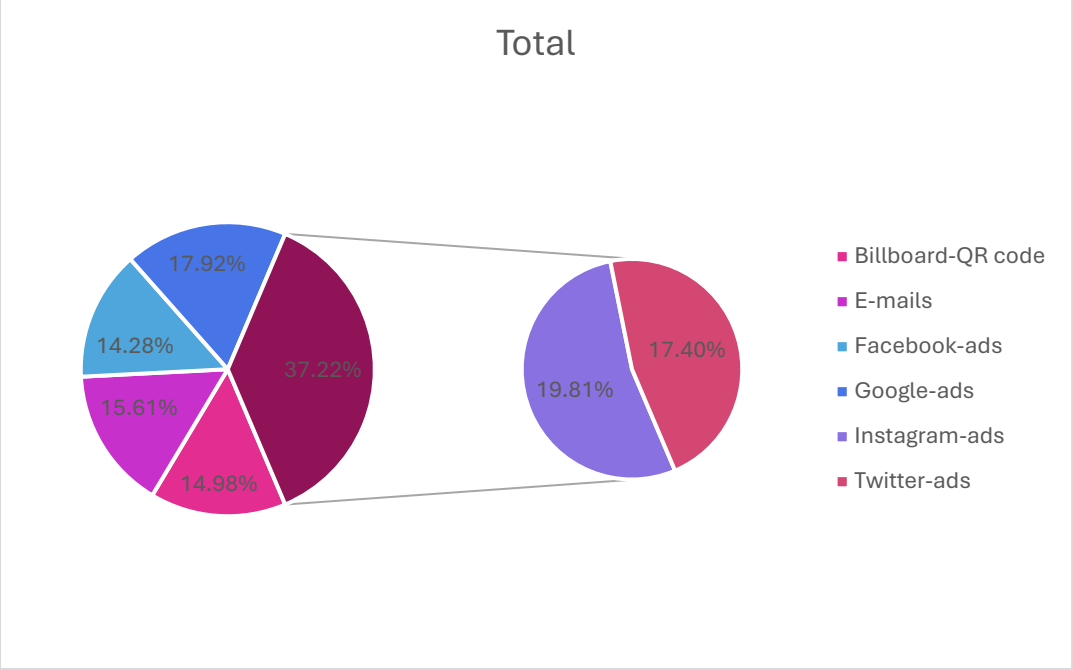
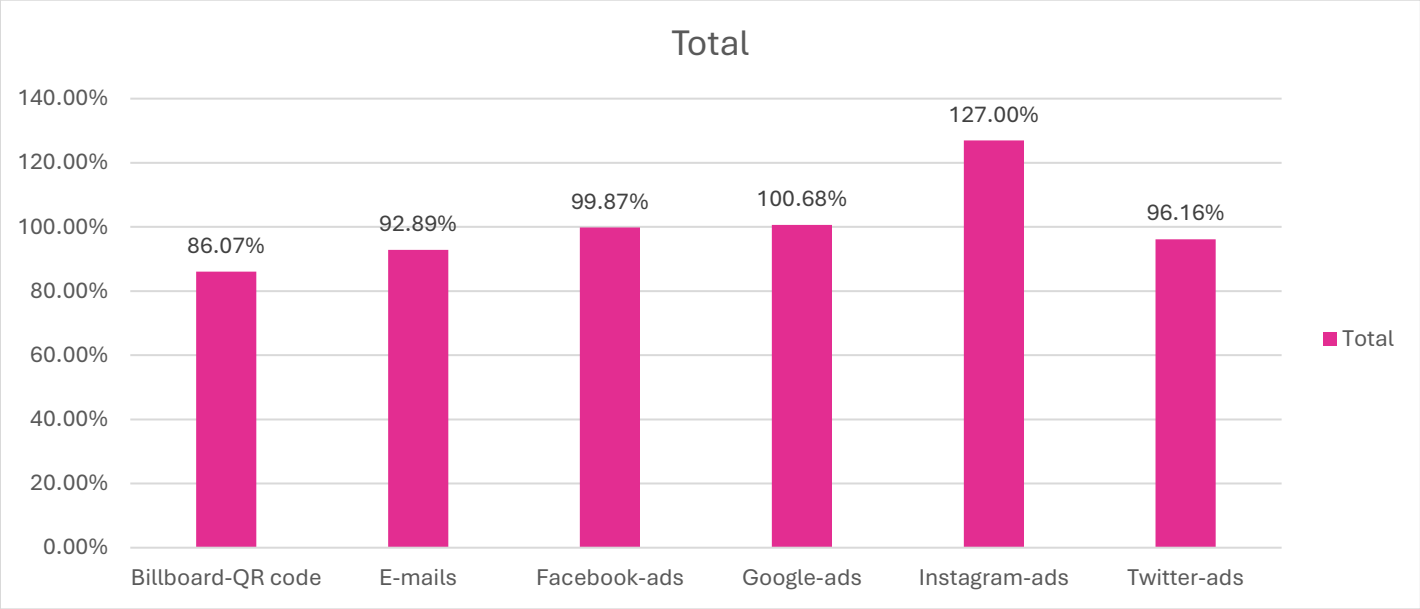


Table 12: Campaign Impact on Average Order Values

Row Labels	Average of Total Revenue
Billboard-QR code	86.07%
E-mails	92.89%
Facebook-ads	99.87%
Google-ads	100.68%
Instagram-ads	127.00%

Twitter-ads	96.16%
Grand Total	**100.00%**

Table 13: Campaign Impact on Average Order Values



Analysis-

The analysis in table 10 of different campaign schemas reveals varying impacts on sales revenue and average order values across marketing channels. Instagram-ads lead in total revenue share with 19.81%, indicating strong sales volume, followed closely by Google-ads at 17.92% and Twitter-ads at 17.40%. Facebook-ads contribute significantly with 14.28%, while E-mails account for 15.61% of total revenue. In contrast, Billboard-QR code campaigns generate 14.98% of revenue, showing the lowest impact among the listed channels.

Instagram-ads lead with an average of 127.00%, indicating higher spending per order compared to other channels. Google-ads follow with 100.68%, maintaining robust revenue per transaction. Facebook-ads achieve 99.87%, demonstrating effective monetization of user engagement. Twitter-ads record 96.16%, indicating solid performance in revenue per order. E-mails achieve 92.89%, showing moderate effectiveness in driving transaction values. Billboard-QR code campaigns have the lowest average order value at 86.07%, highlighting lower spending per transaction despite contributing to overall revenue.

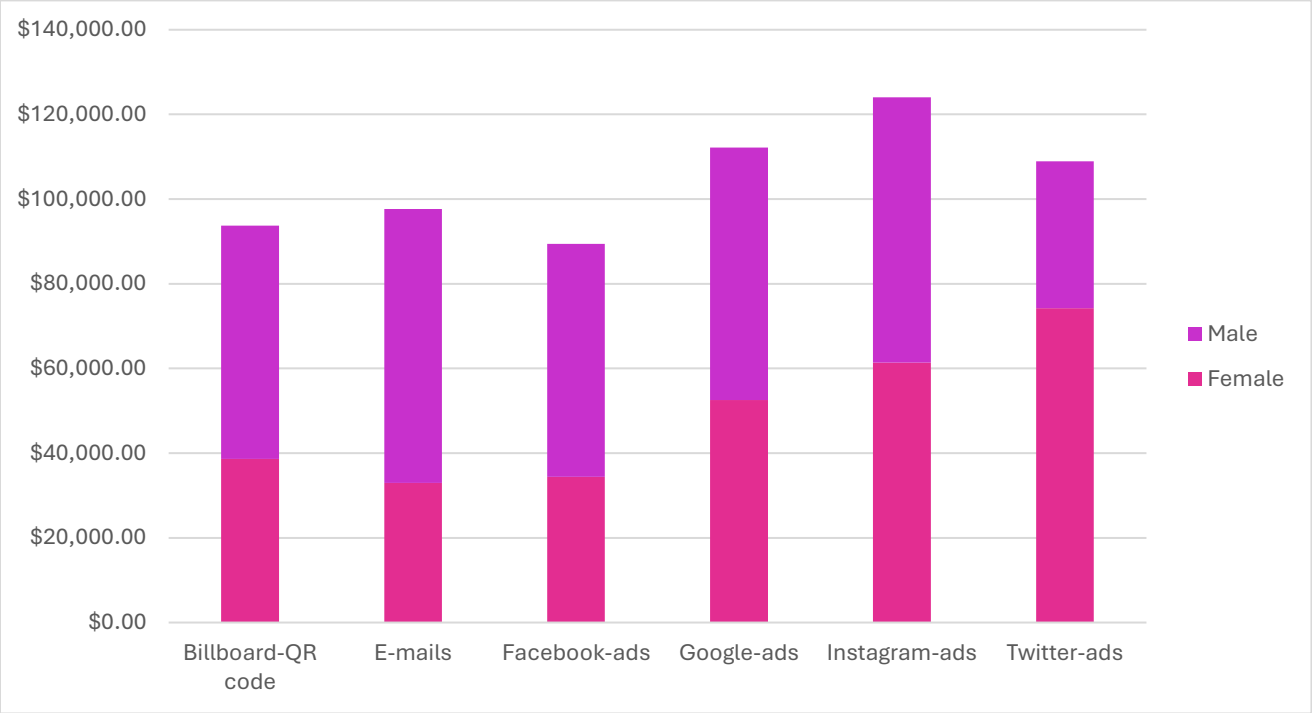
Question 6. How do customer demographics (age, gender) influence sales under different campaign schemas?

TABLE 14: Revenue Distribution by Campaign Schema and Gender

age	(All)

Sum of Total Revenue	**Column Labels**		
Row Labels	**Female**	**Male**	**Grand Total**
Billboard-QR code	$38,662.00	$55,093.00	$93,755.00
E-mails	$32,939.00	$64,764.00	$97,703.00
Facebook-ads	$34,446.00	$54,965.00	$89,411.00
Google-ads	$52,552.00	$59,646.00	$112,198.00
Instagram-ads	$61,413.00	$62,617.00	$124,030.00
Twitter-ads	$74,203.00	$34,757.00	$108,960.00
Grand Total	**$294,215.00**	**$331,842.00**	**$626,057.00**

TABLE 15: Comparison of Revenue by Campaign Schema and Gender



Analysis-

Across all campaign schemas, Instagram-ads stand out with the highest total revenue of $124,030.00, attracting a balanced contribution from both genders. This suggests that Instagram campaigns effectively appeal to a broad audience regardless of gender, potentially due to their visual nature and broad user engagement.

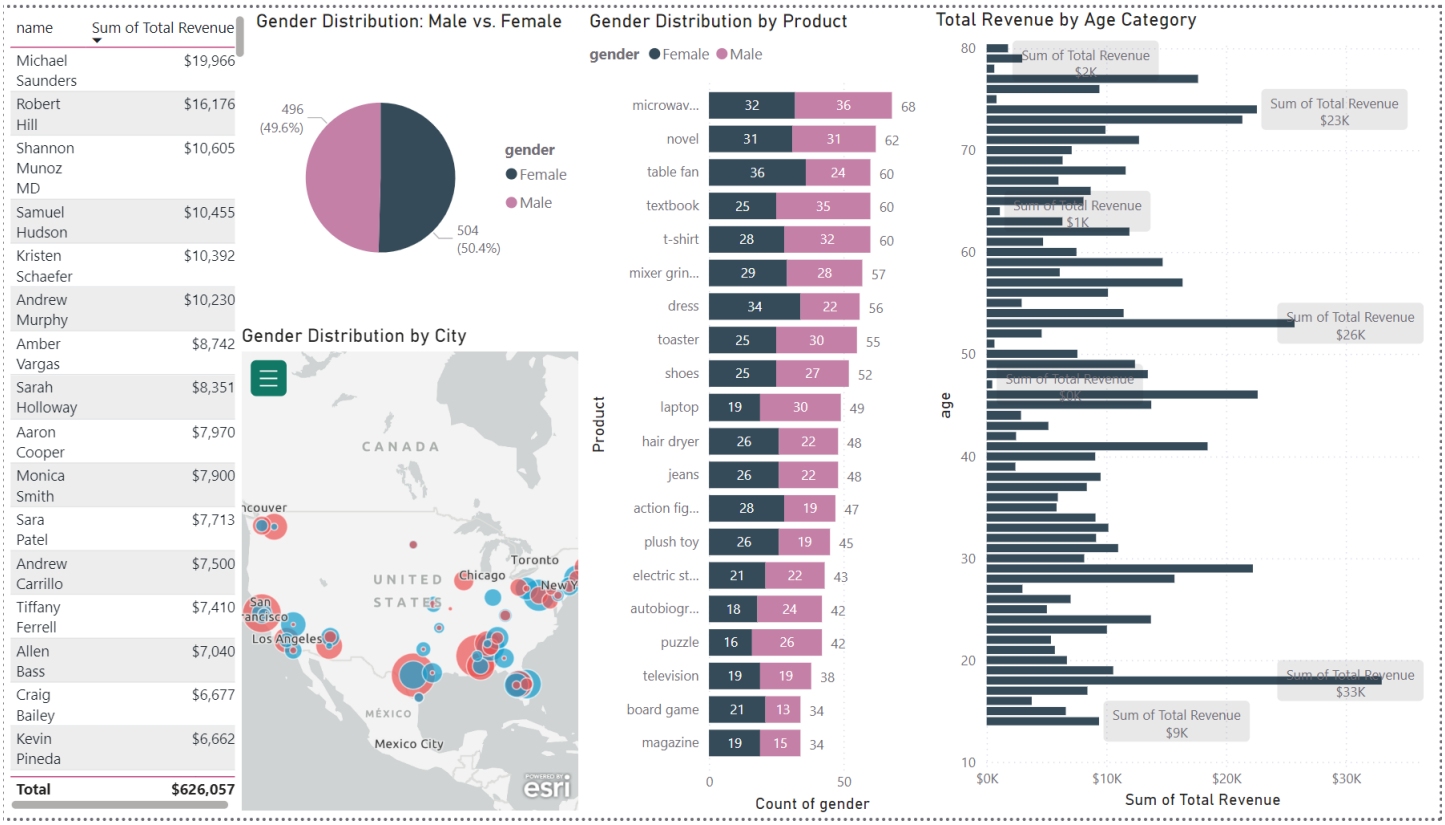
Interestingly, Twitter-ads show a stark contrast in revenue distribution between genders, with $74,203.00 from females compared to $34,757.00 from males, totaling $108,960.00. This indicates a significant preference among females for Twitter campaigns, possibly influenced by content relevance or promotional strategies tailored to female demographics.

The stacked bar graph visually reinforces these findings, illustrating the total revenue for each campaign schema where male contributions are represented in purple and female contributions in pink. This visualization effectively highlights how each campaign type performs across gender demographics, providing a clear picture of where marketing efforts may be more effective based on gender-specific targeting strategies.

Cleaning Using R Code

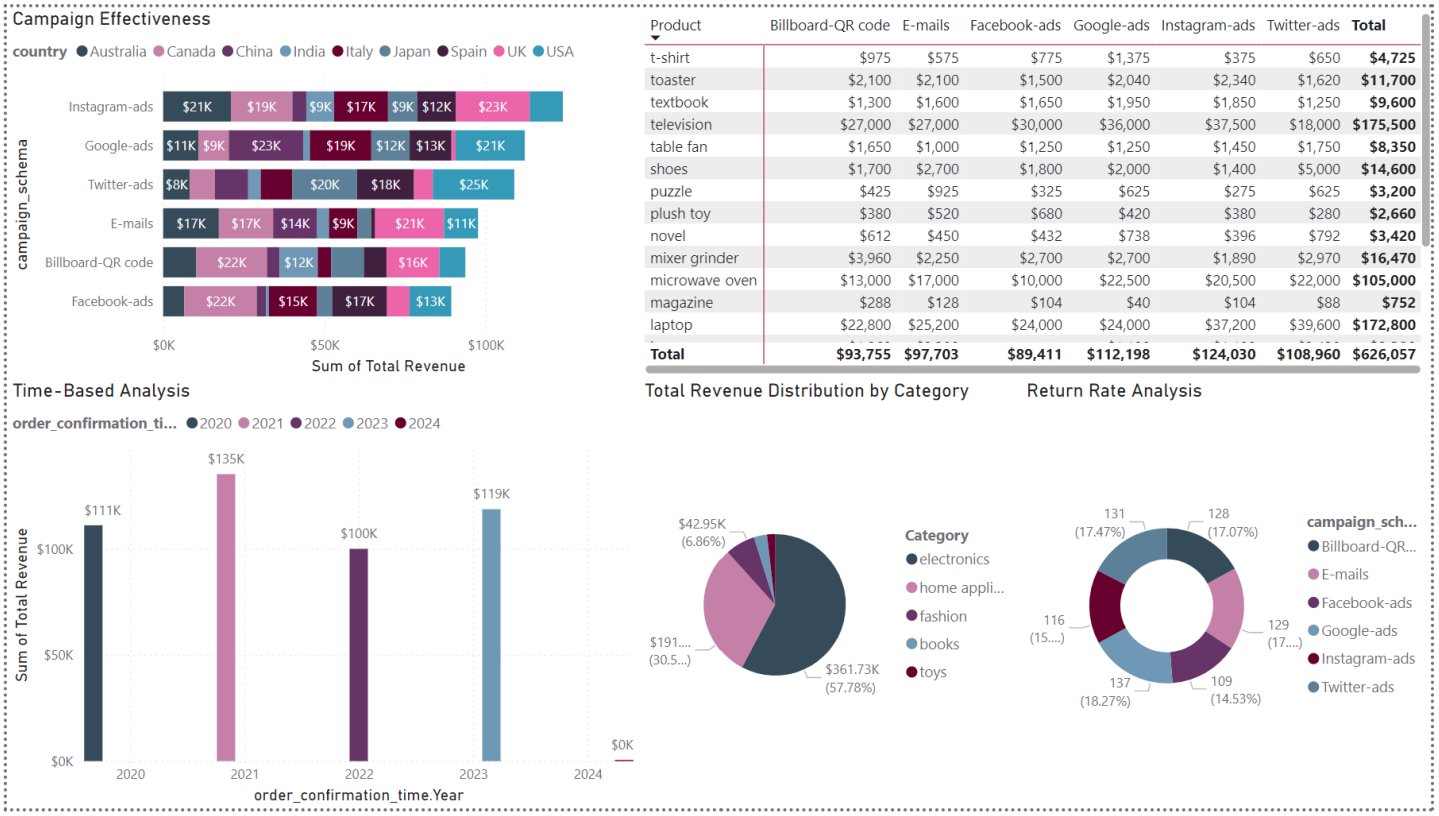
```
#Install necessary packages if not already installed
install.packages("data.table")
install.packages("lubridate")

# Load the packages
library(data.table)
library(lubridate)

# Convert sales_data to data.table for efficient processing
setDT(sales_data)

# Parse and clean the order_confirmation_time column in-place
sales_data[, order_confirmation_time := parse_date_time(order_confirmation_time,
                                                        orders = c("ymd HMS", "dmy HM", "Ymd HMS", "dmY HMS", "ymd HM", "dmy
HMS"))]

# Parse and clean the cart_addition_time column in-place
sales_data[, cart_addition_time := parse_date_time(cart_addition_time,
                                                   orders = c("ymd HMS", "dmy HM", "Ymd HMS", "dmY HMS", "ymd HM", "dmy HMS"))]

# Format the cleaned date-time columns to the desired format
sales_data[, order_confirmation_time := format(order_confirmation_time, "%d-%m-%Y %H:%M")]
sales_data[, cart_addition_time := format(cart_addition_time, "%d-%m-%Y %H:%M")]

# Display the cleaned data
print(sales_data)



> sales_data <- sales_data %>%
+     mutate(Gender_Num = ifelse(gender == "Female", 1, 0))
> head(sales_data)
```

TABLE 16: Customer Profile Analysis



The first visual highlights our top spenders, with Michael Saunders leading at $19,966, followed by Robert Hill with $26,176, and Shannon Munoz MD at $10,606, showcasing their significant contributions to our revenue. The second visual illustrates gender distribution, revealing a nearly equal split with males comprising 49.6% and females 50.4% of the dataset. Moving to the third visual, a map representation displays gender distribution by city. For instance, San Antonio shows a higher female population (31) compared to males (19), providing insights into regional gender dynamics. The fourth visual, a stacked bar graph titled "Gender Distribution by Product," highlights purchasing trends. Microwaves emerge as the top-selling item, with males purchasing 36 units compared to females' 32. Conversely, magazines show lower sales, with 19 purchases by females and 15 by males. In the final visual, "Total Revenue by Age Category," 18-year-olds lead in spending with $33,000, followed by 53-year-olds at $26,000 and 74-year-olds at $23,000, indicating varied spending patterns across different age groups. These visuals collectively offer valuable insights into customer demographics, spending patterns, and product preferences, guiding strategic decisions to enhance customer engagement and optimize revenue growth.

TABLE 17: Revenue Analysis Dashboard Overview



Product	Billboard-QR code	E-mails	Facebook-ads	Google-ads	Instagram-ads	Twitter-ads	Total
t-shirt	$975	$575	$775	$1,375	$375	$650	**$4,725**
toaster	$2,100	$2,100	$1,500	$2,040	$2,340	$1,620	**$11,700**
textbook	$1,300	$1,600	$1,650	$1,950	$1,850	$1,250	**$9,600**
television	$27,000	$27,000	$30,000	$36,000	$37,500	$18,000	**$175,500**
table fan	$1,650	$1,000	$1,250	$1,250	$1,450	$1,750	**$8,350**
shoes	$1,700	$2,700	$1,800	$2,000	$1,400	$5,000	**$14,600**
puzzle	$425	$925	$325	$625	$275	$625	**$3,200**
plush toy	$380	$520	$680	$420	$380	$280	**$2,660**
novel	$612	$450	$432	$738	$396	$792	**$3,420**
mixer grinder	$3,960	$2,250	$2,700	$2,700	$1,890	$2,970	**$16,470**
microwave oven	$13,000	$17,000	$10,000	$22,500	$20,500	$22,000	**$105,000**
magazine	$288	$128	$104	$40	$104	$88	**$752**
laptop	$22,800	$25,200	$24,000	$24,000	$37,200	$39,600	**$172,800**
Total	**$93,755**	**$97,703**	**$89,411**	**$112,198**	**$124,030**	**$108,960**	**$626,057**

The "Campaign Effectiveness" visual presents a stacked bar graph highlighting spending across various advertising channels. Notably, the UK leads with $23K expenditure on Instagram ads, China tops Google ads with $25K, and the USA spends $25K on Twitter ads. Email campaigns are dominated by the UK with $21K, while Canada spends $22K each on Billboard-QR codes and Facebook ads. The "Time-Based Analysis" visual displays revenue trends over years: $111K in 2020, $135K in 2021, $100K in 2022, and $119K in 2023. "Total Revenue Distribution by Category" indicates electronics as the top category at 57.78%, followed by home appliances at 30.5%, and fashion at 6.86%. Lastly, the "Return Rate Analysis" reveals Google ads with the highest return rate at 18.2%, followed by Twitter ads at 17.47%, and email campaigns at 17.2%. These insights provide a comprehensive view of revenue patterns and campaign effectiveness across different dimensions.

Conclusion

The analysis reveals significant insights across pricing, income, credit scores, and revenue distribution. Product prices vary widely, with a mean of $204.59 and outliers influencing skewness (2.47) and kurtosis (4.95). Monthly incomes average $50,230.57, skewed left with an outlier at $15,407.00. Credit scores cluster tightly around a mean of 709.23, indicating a stable distribution. Total revenue amounts to $626,057.00, driven by electronics (58.77%) and home appliances (28.90%). Regional sales show notable contributions from various countries, underscoring diverse market dynamics. Campaign effectiveness highlights Instagram ads as the top performer (19.81%), with notable gender-based revenue variations observed in Twitter ads. This analysis provides strategic insights for optimizing marketing efforts and enhancing revenue growth across different market segments and regions.

Source: [Online Retail Sales Data (kaggle.com)]