Insights and Exploratory Analysis of Digital Advertising Campaigns

- **Author:** Matias MU
- **Tool Used:** Jupyter Notebook. Python (Pandas, Matplotlib, Seaborn, Scipy) for data cleaning, analysis, visualization, and statistical testing.
- **Techniques:** Data cleaning and preprocessing, Exploratory Data Analysis (EDA) , Statistical testing , Correlation analysis, Grouping and filtering data
- **Date:** 31 December, 2024

Table of Contents

1. Introduction

The dataset used in this analysis contains information about digital video advertising campaigns from 26 advertiser accounts over a one-year period, between 2021 and 2022. With a total of 237,409 records and 51 columns, which are provided in Spanish. The dataset covers a wide range of key metrics related to campaign performance. These include costs, impressions, clicks, conversions, and return on advertising spend, among others.

URL of the dataset:

- https://zenodo.org/records/7965793

The dataset columns include detailed information such as date, device, campaign execution country, product category, and economic metrics associated with the campaign, such as cost (`Coste_i` and `Coste_f`) and revenue (`Revenue`). However, due to the lack of specific documentation on the columns, some have been inferred from the names, and new columns have been created to calculate ROAS.

The dataset, while valuable, presents some challenges related to its quality, as it consists of synthetic data with an uneven distribution. However, efforts have been made to normalize the data and make it suitable for analysis.

This project aims to draw relevant conclusions from this big data analysis, with the goal of providing practical insights to advertisers, helping them optimize the performance of their advertising campaigns through a better understanding of the data.

2. Objective of the analysis

The main objective of this analysis is to identify key patterns and relationships within the digital advertising campaign dataset, in order to provide practical recommendations for improving campaign performance. Through an exploratory approach and utilizing statistical techniques and visualizations, we aim to answer strategic questions that can help advertisers optimize their investments and maximize their return based on the data from the dataset.

The specific questions that guided this analysis are:

1. **Which days of the week generate the highest ROAS, and are the results statistically different?**
   Analyze if there are significant differences in campaign performance based on the day of the week and statistically compare the results.
2. **What is the relationship between the final cost and the ROAS of a campaign? The more is invested, the more ROAS the campaign generates?**
   Evaluate if a higher investment in campaigns leads to better results in terms of ROAS.
3. **Which devices generate the most conversions and how do they impact ROAS?**
   Determine which device contributes the most to conversions and its impact on the overall performance of the campaigns.
4. **Are there product categories that perform better in certain countries?**
   Identify whether some product categories perform better depending on the country.

3. Data Analysis Methodology

3.1. Dataset preparation

- **Initial Inspection:** The analysis process began by inspecting the first rows and the structure of the dataset using the methods `df.info()` and `df.describe()`. This allowed us to understand the number of rows and columns, the data types of each column, and the presence of any null values.
- **Data Cleaning:**
  - **Elimination of Irrelevant Columns:** Irrelevant columns for the analysis were identified and removed, such as the column `TQ_SourceTag`, which contained null values and did not contribute any value to the analysis.
  - **Data Type Conversion:** Incorrect or inconsistent data types were changed:
    - **Dates:** The `Fecha` column was converted to `datetime` type to facilitate future temporal analysis.
    - **Percentages:** Variables with percentage values (such as `ROAS_i`, `ROAS_f`, etc.) were converted from `object` to `float` format and divided by 100.
  - **Categorical Variables:** Some numeric columns (such as `Cuenta_id`, `Campaña_id`, `Grupo,id`) were converted to `object` type, as they represented unique campaign identifiers, not continuous numeric values.

3.2. Grouping and Filtering Variables:

- **Variable segmentation:**
  - **Categorical variables:** All `object` variables were extracted to enable analysis as categorical variables.
  - **Numerical Variables:** Numerical variables were identified and isolated for further trend and correlation analysis, such as `Coste_f`, `Revenue`, `ROAS_calculated`, among others.
- **Creation of new variables:**
  - **Calculated ROAS:** To address the lack of clarity in the calculation of ROAS in the original dataset, a new `ROAS_calculated` column was calculated, obtained by dividing `Revenue` by `Coste_f` (final cost). This variable provided a key index for assessing campaign performance.

3.3. Outlier filtering:

- **Elimination of Outliers:**
  - Outliers on key variables were identified and removed as `ROAS_calculated` using the 1% and 99% percentiles, in order to get a more representative picture of the overall performance of the campaigns.
  - This allowed a better visualisation of the distributions and prevented outliers from distorting the results and the interpretation of the analysis.

3.4. Selection of Relevant Variables:

- **Column Filtering:**
  - The most relevant columns were selected for the performance analysis of the campaigns. These include the variables `Fecha`, `Dia`, `Dia_sem`, `Coste_f`, `Revenue`, `Conversiones_f`, `CPC_f`, `ROAS_calculated` , among others.
  - The column selection process allowed the dataset to be reduced to the variables that were considered fundamental to the analysis questions posed, ensuring efficiency in the subsequent analysis.

4. Table of translated and described columns

This table provides the English translations and descriptions of the columns from the dataset used in the analysis. The dataset contains 51 columns, but only the relevant columns used for this analysis are included here. These columns represent the key metrics and attributes necessary for understanding the performance of the digital advertising campaigns analyzed.

Column (Spanish)	Column (English)	Description
Fecha	Date	Date of the campaign
Dia	Day	Day of the month
Dia_sem	Weekday	Day of the week
Mes	Month	Month of the campaign
Cuenta_id	Account_id	Account identifier
Campaña_id	Campaign_id	Campaign identifier
Grupo_id	Group_id	Group identifier
Anuncio_id	Ad_id	Ad identifier
Dispositivo_revenue	Device_revenue	Device used for revenue
Categoria	Category	Category of the product or service
Keyword	Keyword	Keyword used for targeting
Impresiones_f	Impressions_f	Impressions of the campaign
Coste_f	Final_Cost	Final cost of the campaign
Revenue	Revenue	Revenue generated by the campaign
Profit_f	Final_Profit	Final profit of the campaign
CPC_f	CPC	Cost per click
RPC	RPC	Revenue per click
Conversiones_f	Conversions_f	Final conversions
Clicks_f	Clicks_f	Final clicks
CR_G_f	CR_G_f	Final conversion rate
ROAS_calculated	ROAS_calculated	Calculated ROAS

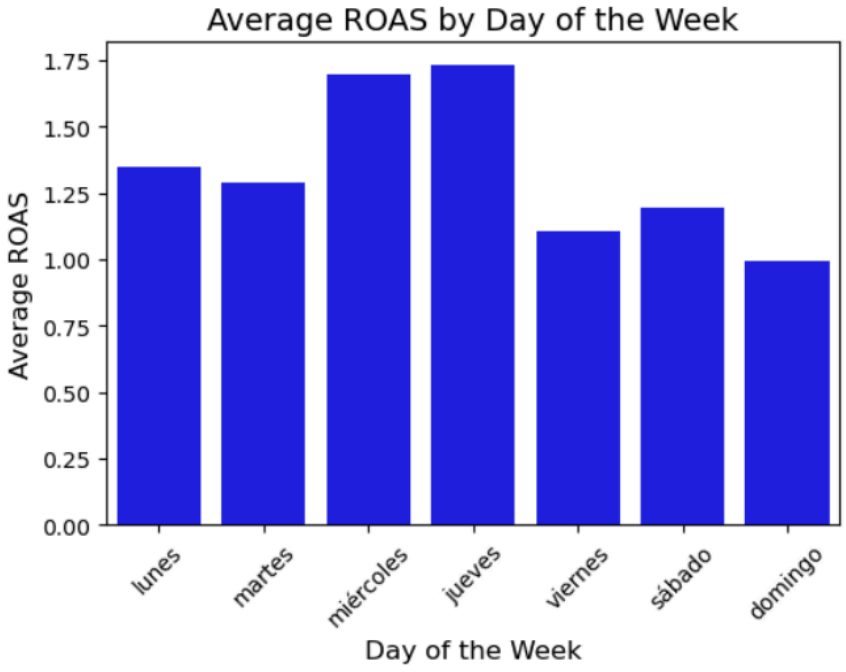
5. Analysis and Results

5.1. Insight 1

- Which days of the week generate the highest ROAS and are the results statistically different?

The analysis of the average ROAS by day of the week revealed that Tuesday and Wednesday generated the best results in terms of ROAS, with average values of 1.35 and 1.69, respectively. In comparison, Saturday and Sunday showed the lowest values, with an average ROAS of 1.19 and 1.10.

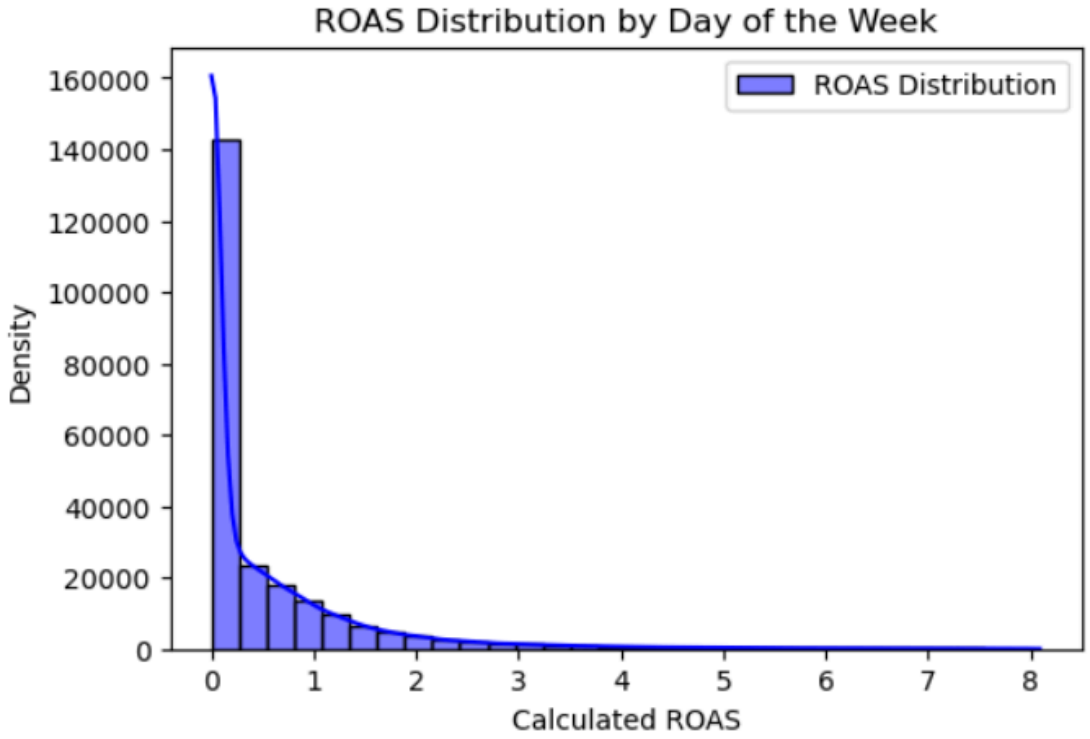
This pattern suggests that weekdays, particularly Tuesday and Wednesday, appear to be more effective in maximizing return on advertising investment, while weekends show lower performance.



Graph 1: Average ROAS by day of the week. Source: Author's elaboration.

ROAS distribution

By observing the distribution of ROAS through a histogram, it is noticeable that there is a concentration of values close to 0, with some outliers representing campaigns with exceptionally high ROAS. This distribution suggests that most campaigns have a low ROAS, but there are exceptional cases that achieve high returns. By removing the outliers (extreme values), we gain a more accurate view of the data, focusing on the most representative results.



Graph 2: ROAS Distribution by day of the week. Source: Author's elaboration.

Statistical Analysis - ANOVA Test

To validate whether the observed differences in average ROAS between days of the week are statistically significant, an ANOVA test was performed. The results of the ANOVA were as follows:

- **F**: 2.00
- **p-value**: 0.0621

```python
# Perform ANOVA test

from scipy.stats import f_oneway
from statsmodels.stats.multicomp import pairwise_tukeyhsd

anova_result = f_oneway(*roas_values_by_day)
print(f"ANOVA Result - F: {anova_result.statistic:.2f}, p-value: {anova_result.pvalue:.4f}")
```

```
ANOVA Result - F: 2.00, p-value: 0.0621
```

```python
# Interpretation of ANOVA

if anova_result.pvalue < 0.05:
    print("There are statistically significant differences in ROAS between the days of the week.")
else:
    print("No statistically significant differences in ROAS were found between the days of the week.")
```

```
No statistically significant differences in ROAS were found between the days of the week.
```

Code extract 1: ANOVA Test. Source: Author's elaboration.

The p-value of 0.0621 is slightly higher than the commonly used significance threshold (0.05), suggesting that no statistically significant differences were found between the days of the week in terms of ROAS. Although visually Tuesday and Wednesday seem to perform better, the statistical analysis does not support the hypothesis that these days generate a significantly higher ROAS than the other days of the week.

Conclusion

Although the visual and descriptive analysis indicates that Tuesday and Wednesday perform better in terms of ROAS, the ANOVA statistical test shows that the differences are not strong enough to be considered significant. This result suggests that, while certain days appear to offer a better return, the variations in ROAS performance between the days of the week are not consistent enough to establish a clear statistical relationship.

This finding is important for campaign planning strategies, as it indicates that other factors may influence the success of campaigns more than the day of the week.

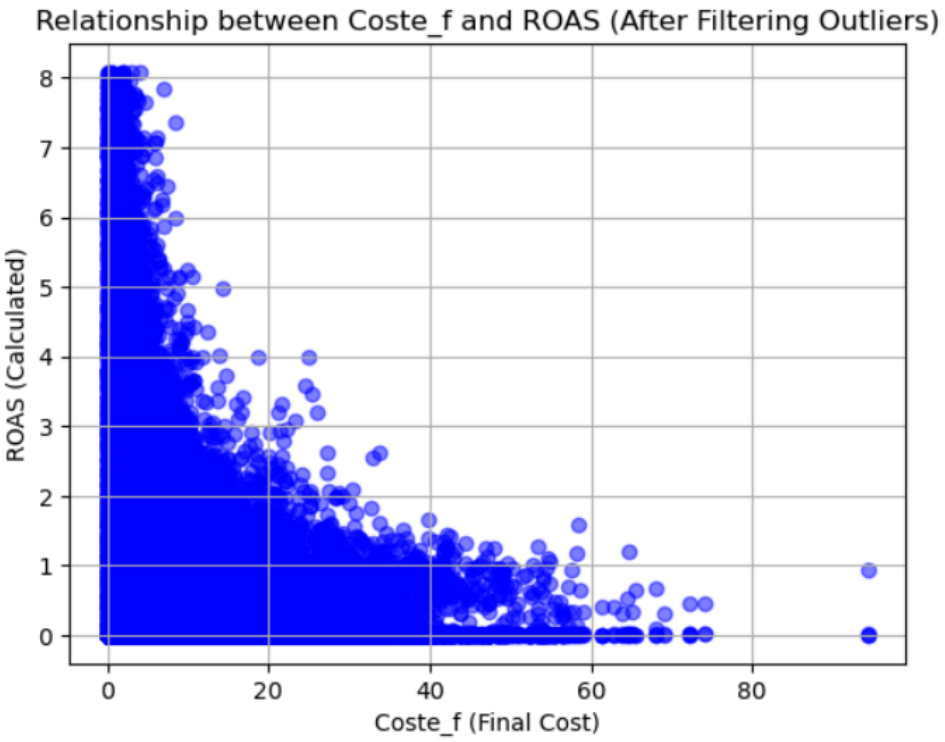
5.2. Insight 2

- What is the relationship between the final cost and the ROAS of a campaign? The more is invested, the more ROAS the campaign generates?

Descriptive Analysis

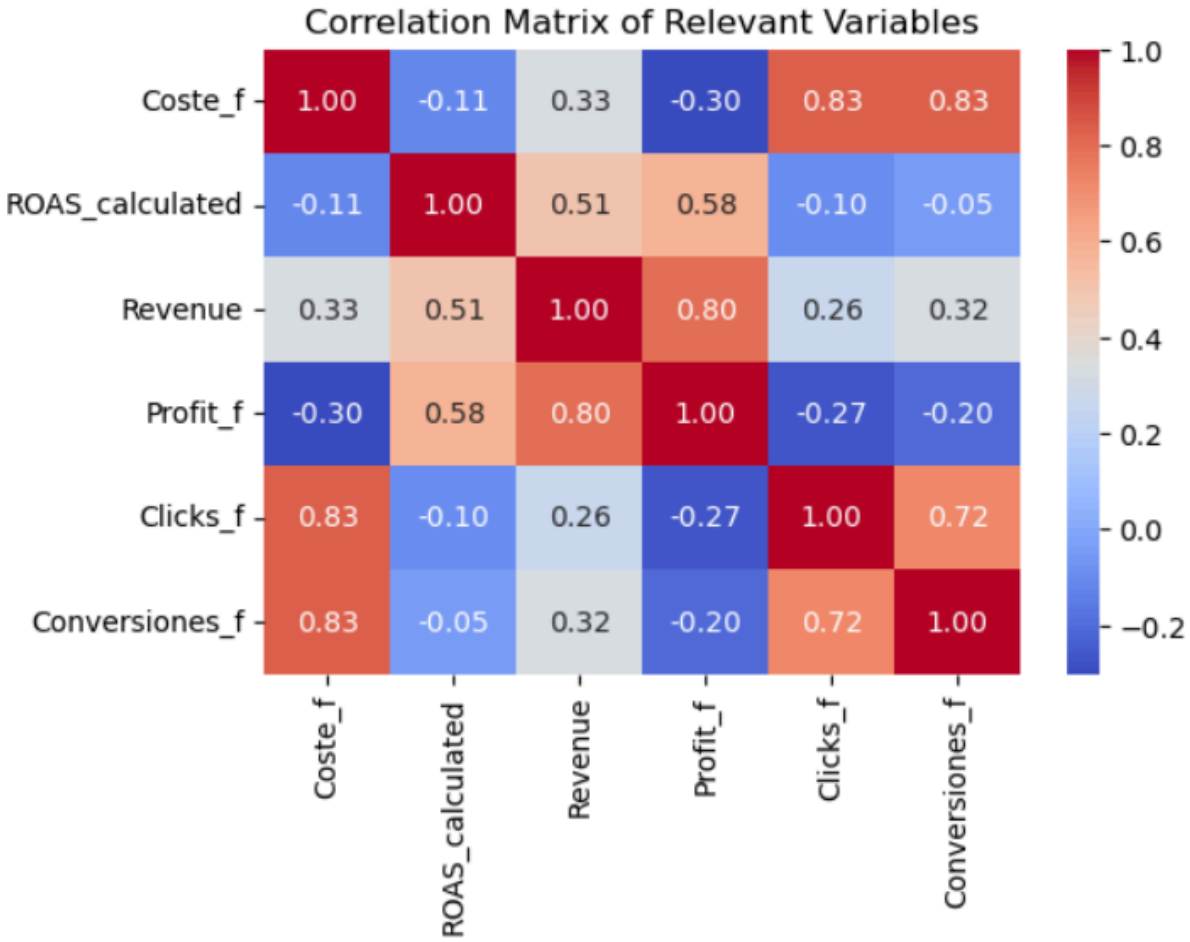
In the scatter plot where the Final Cost (`Coste_f`) and the Calculated ROAS (`ROAS_calculated`) have been analyzed, a clear inverse relationship between these two variables can be observed. That is, as the final cost of the campaign increases, the ROAS tends to decrease. This may suggest that, although campaigns with higher investments tend to generate higher revenues, it is not always proportional to the return on advertising investment.

In the scatter plot between `Coste_f` and `ROAS_calculated`, it can be observed that there is a high concentration of points in the low-cost range, where some campaigns achieve a high ROAS. However, at higher values of `Coste_f`, the `ROAS_calculated` begins to stabilize at lower levels, suggesting that increasing spending does not always guarantee a better return.



Graph 3: Relationship between Coste_f and ROAS (After filtering outliers). Source: Author's elaboration.

The correlation matrix shows that `Coste_f` has a weak negative correlation of -0.11 with `ROAS_calculated`, reinforcing the observation that, on average, higher costs do not necessarily result in a higher ROAS. However, the analysis also indicates that other variables, such as `Conversiones_f` and `Clicks_f`, have a stronger positive correlation with `ROAS_calculated`, suggesting that not only cost influences ROAS, but also the quality of conversions and clicks generated.



Matrix 1: Correlation Matrix of Relevant Variables. Source: Author's own elaboration.

The observed relationship between Final Cost and ROAS suggests that increasing spending does not necessarily guarantee an improvement in return on investment. In fact, some campaigns with low spending achieve high returns, which could be related to better segmentation strategies or a stronger focus on the right target audience.

Although intuitively one might expect that a higher investment would generate a higher return (ROAS), this data suggests that other variables play an important role.

5.3. Insight 3

- Which devices generate the most conversions and how do they impact ROAS?

Descriptive Analysis

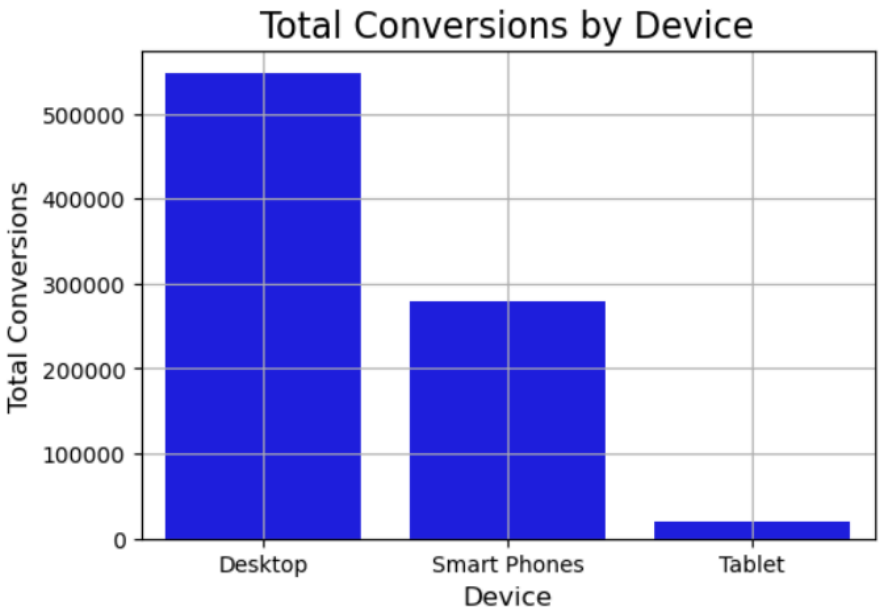
Regarding the conversions generated by device, the results show that Desktop is the leading device with a significant difference in terms of total conversions. With 547,021 conversions, Desktop is the most used device in terms of interaction, followed by Smart Phones with 278,433 conversions, and finally Tablets, which generate the least with only 20,519 conversions.

```
# Group data by 'Dispositivo_revenue' and calculate total conversions by device
conversions_by_device = df.groupby('Dispositivo_revenue')['Conversiones_f'].sum().reset_index()

# Verify the results of the grouping
conversions_by_device
```

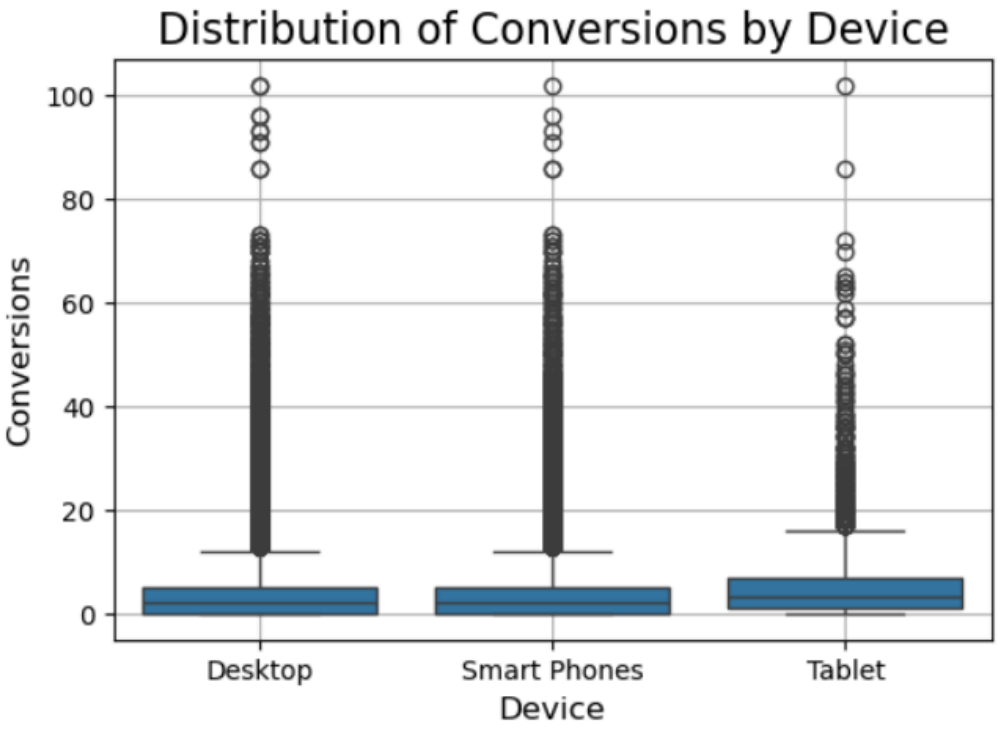
	Dispositivo_revenue	Conversiones_f
0	Desktop	547021
1	Smart Phones	278433
2	Tablet	20519

Code extract 2: Table of conversions per device. Source: Author's elaboration.



Graph 4: Total Conversions by Device. Source: Author's elaboration.

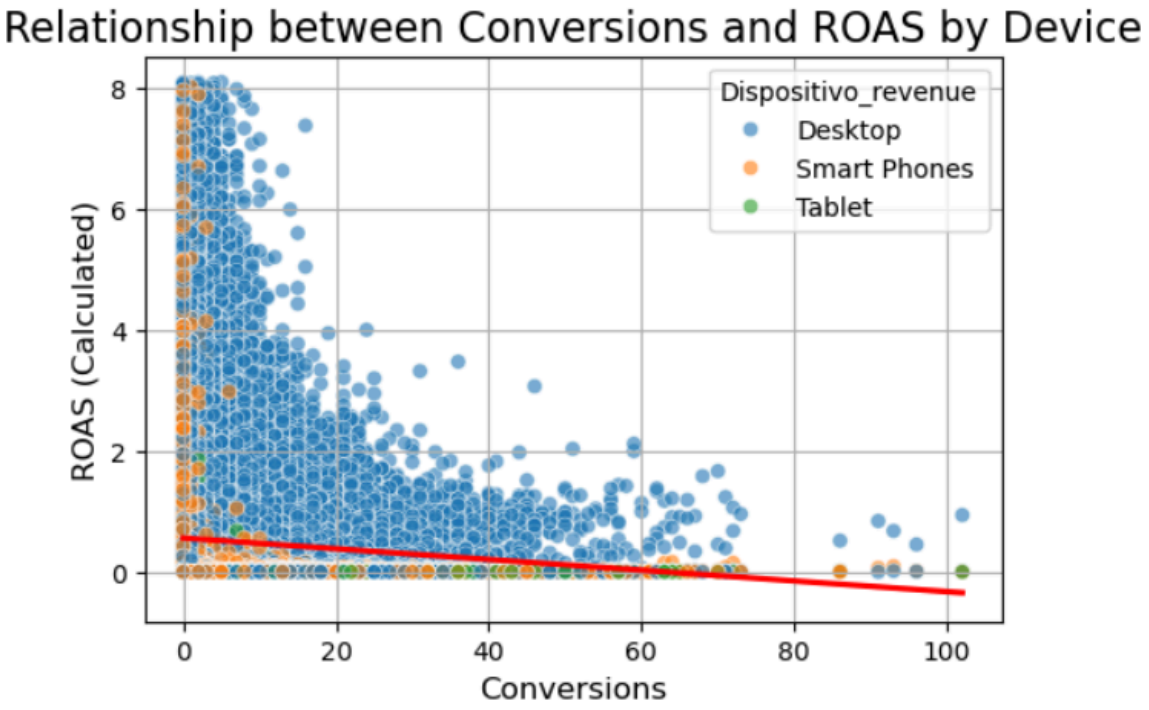
In the distribution of conversions by device, we observe that, although Desktop and Smart Phones have a high concentration of low conversions, Tablet shows a greater dispersion, suggesting it has higher variability in conversions, but in smaller amounts. This could indicate that, although Tablet generates fewer conversions, some campaigns may have been more effective on this device, but they are not representative of the total.



Graph 5: Distribution of Conversions by Device. Source: Author's elaboration.

In the visualization of the relationship between Conversions and ROAS, it stands out that Desktop generates many conversions, but these do not always translate into a high ROAS. The dispersion in the graph shows that, although conversions by Desktop are high, the ROAS is concentrated around lower values for most conversions. This implies that more conversions do not necessarily result in better performance in terms of return on investment (ROAS).

On the other hand, both Smart Phones and Tablets generate more limited conversions, and the graph shows that their conversions do not always have a high ROAS performance. In fact, it is observed that although mobile devices are generating conversions, the performance remains low, indicating that these devices may require further optimization.



Graph 6: Relationship between Conversions and ROAS by Device. Source: Author's elaboration.

Conclusion

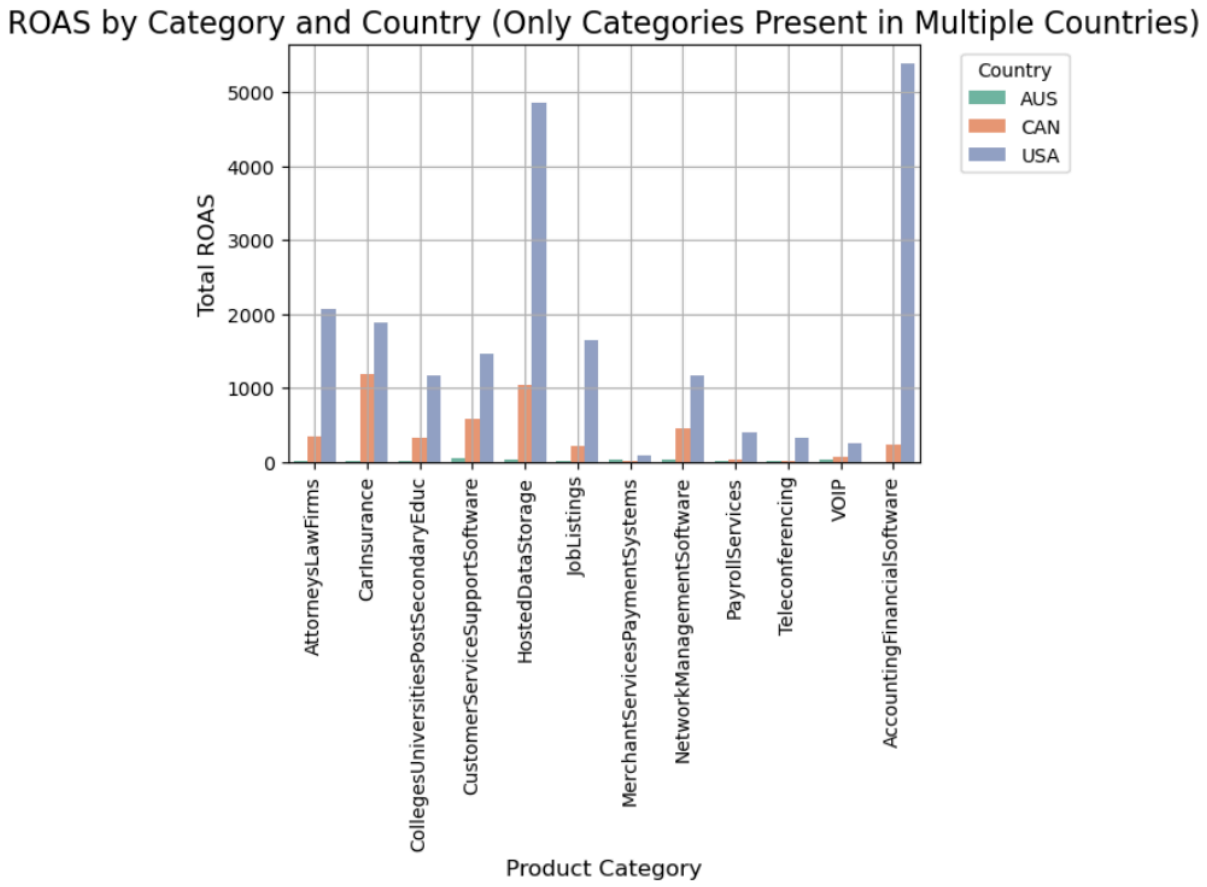
Based on the analyzed data, in this dataset, we can observe that Desktop is the device with the highest number of conversions, but these conversions do not always yield a good return (ROAS). Although Smart Phones and Tablets generate fewer conversions, their low ROAS suggests that these conversions are not profitable. This highlights the importance of focusing not only on the quantity of conversions but also on their quality and profitability.

5.4. Insight 4

- Are there product categories that perform better in certain countries?

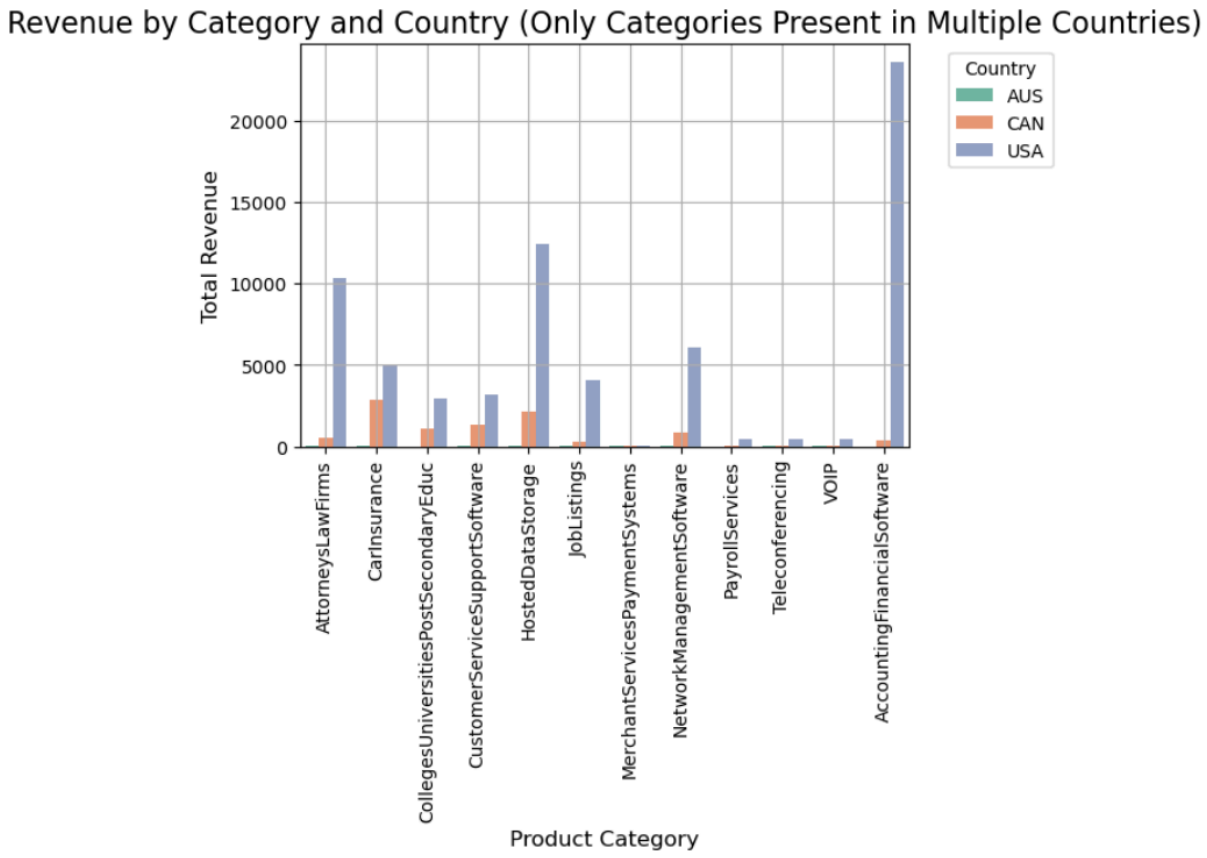
Descriptive Analysis

The analysis reveals that the performance of categories varies significantly across different countries. For example, in the USA, categories such as 'HostedDataStorage' and 'AttorneysLawFirms' have a considerably high ROAS, surpassing other countries like Canada and Australia, where the results are more balanced.



Graph 7: ROAS by Category and Country (Only Categories Present in MultipleCountries). Source: Author's elaboration.

On the other hand, categories such as 'AccountingFinancialSoftware' and 'HostedDataStorage' also stand out in terms of Revenue in the USA, showing that some categories perform much stronger in certain countries.



Graph 8: Revenue by Category and Country (Only Categories Present in Multiple Countries). Source: Author's elaboration.

Conclusion

This behavior suggests that the selection of categories in advertising campaigns should consider the geographical context, as some products or services may generate higher returns in specific markets. Therefore, identifying and prioritizing the best-performing categories in each country can significantly improve campaign efficiency. This approach of personalization by country and category can be crucial to maximize return on investment globally.

6. Challenges

6.1. Conversion of Percentage Values to Decimals:

The dataset contained several columns with percentage values represented as `objects` (e.g., `ROAS_i`, `ROAS_f`). To perform accurate calculations and analyses, these percentage values needed to be converted to decimals. The conversion was done by removing the percentage symbol and dividing the values by 100 to standardise them for numerical analysis.

```
object_to_numeric = ['Cost_diff', 'ROAS_i', 'ROAS_f', 'CR_JOT_i', 'CR_JOT_f',
                     'CR_G_i', 'CR_G_f', 'Conv_diff_i_%', 'Conv_diff_f _%']

for temp in object_to_numeric:
    df[temp] = df[temp].str.replace('%', '').astype(float) / 100
```

This process ensured that all relevant columns were in a consistent numerical format, allowing for proper analysis and calculations.

6.2. Data Inconsistency:

During the initial analysis, it was identified that some values in the `ROAS_f` column were incorrect. When comparing the results obtained by calculating `Coste_f` divided by `Revenue`, it was observed that the values in `ROAS_f` did not match the expected calculation. This could have been caused by data entry errors or issues during the data collection process. To address this issue, a new column called `ROAS_calculated` was created, which was directly calculated from the `Coste_f` and `Revenue` columns as follows:

```
# Creating the ROAS_calculated column as float type
df['ROAS_calculated'] = 0.0

# Calculating ROAS for rows where Cost_f is not 0 (allowing negative ROAS)
df.loc[df['Coste_f'] != 0, 'ROAS_calculated'] = (df['Revenue'] / df['Coste_f'])

# Checking the first values to make sure they were calculated correctly
df[['Campaña_id', 'Coste_f', 'Revenue', 'ROAS_calculated']].head()
```

This adjustment helped correct the inconsistency and provided a more accurate measure of the return on advertising investment (ROAS) for the campaigns analysed.

6.3. Identification of Outliers:

During the exploratory analysis, it was observed that the ROAS distribution contained several extreme values, which could distort the overall insights. Outliers can often skew the results, making it harder to identify trends and patterns within the main data set. To address this issue, outliers were identified using percentiles (1st and 99th) for `ROAS_calculated` , which allowed for more accurate visualisation and analysis.

The data was filtered to remove values outside the 1st and 99th percentiles, as shown below:

```
# Filtering the dataset to remove outliers using percentiles (1% and 99%).
lower_percentile = df['ROAS_calculated'].quantile(0.01)
upper_percentile = df['ROAS_calculated'].quantile(0.99)

df_filtered_percentiles = df[(df['ROAS_calculated'] >= lower_percentile) & (df['ROAS_calculated'] <= upper_percentile)]

# Checking the number of records after filtering
df_filtered_percentiles.shape[0]
```

This process helped to refine the dataset, allowing for a clearer analysis of the more typical campaign performance. The removal of extreme values enabled a more meaningful interpretation of the data, particularly in terms of general trends and average ROAS.

7. References

Dataset: "Data set and classification method for low quality web traffic identification in video marketing campaigns," Zenodo, 2023. [Online]. Available: https://zenodo.org/records/7965793.

Symeon Charalabides, Lecture 10 - Data analysis techniques, National College of Ireland, 2024.

Symeon Charalabides, Lecture 11 - Data visualization, National College of Ireland, 2024.

Symeon Charalabides, Lecture 12 - Data Graphing - I, National College of Ireland, 2024.

Symeon Charalabides, Lecture 12 - Data Graphing - I, National College of Ireland, 2024.

Symeon Charalabides, Lecture 12 - Data Graphing and Visualization - II, National College of Ireland, 2024.

Dholakia, R. R. & K. M. Bagozzi. "The Role of Digital Marketing in Customer Engagement and Advertising." Journal of Interactive Marketing, 2020.

Google, (n.d.) ‘Understand customer needs’, Think with Google, Available from: https://www.thinkwithgoogle.com/consumer-insights/consumer-journey/understand-customer-needs/.