**Marketing and Product Performance of Products by Continent(2010-2023)**

İbrahim Barakat   
 *Statistics*  
*Middle East Technical University  
Ankara,Turkey*  
e2656700 @metu.edu.tr

Oğuz Emre Üstün   
 *Statistics*  
*Middle East Technical University  
Ankara,Turkey*  
e256157@metu.edu.tr

Onur Demir  
*Statistics*  
*Middle East Technical University  
Ankara,Turkey*  
e251015 @metu.edu.tr

Tarık Yılmaz   
 *Statistics*  
*Middle East Technical University  
Ankara,Turkey*  
e2656700 @metu.edu.tr

Arda Öztürk  
Statistics  
*Middle East Technical University  
Ankara, Turkey*  
e2561488 @metu.edu.tr

İsmail Özdemirel  
 *Statistics*  
*Middle East Technical University  
Ankara,Turkey*  
e266671 @metu.edu.tr

*Abstract: In the world we live in, thousands of products are bought and sold and the data obtained touches the lives of humanity. It is up to us to understand this situation and affect it positively. In this project, many different sales methods were examined and supported with the necessary statistical and visual data.*

*Keywords—component, formatting, style, styling, insert (key words)*

# **Introduction**

In today's data-driven world, the accuracy and cleanliness of datasets are important to gain meaningful insights and making decisions. The process of tidying and cleaning data are fundamental to make correct inferences. This project aims to clean and organize the dataset using Python libraries which are Pandas and Numpy.

Understanding variables and data types, correcting inconsistencies, and standardizing formats are all necessary actions to prepare data for exploration part. This project follows a detailed checklist that addresses every aspect of data cleaning.

After the data cleaning process, the project moves into the Exploratory Data Analysis (EDA), where various insights are extracted to answer key research questions:

How do different marketing channels impact customer satisfaction, and which channels consistently perform better or worse?

How do revenue distributions vary across regions, and what factors might explain the differences in central tendencies and dispersion patterns among the regions?

How does the average revenue vary across regions and product categories, and which combinations yield the highest or lowest revenues?

How do the contributions of each region (Africa, Asia, Australia, Europe, North America, South America) to the total revenue vary over the years, and what trends can be identified in their growth or decline?

How does marketing efficiency vary throughout the year, and how is it related to the number of units sold each month?

The subsequent sections will detail the data cleaning steps, the exploratory data analysis approach, and the key findings derived from the cleaned dataset.

# **Data Descriptıon**

This dataset consists of 300 observations and 15 variables, collected to analyze marketing strategies and the performances of products. Initially the data contained issues such as typos, missing values, errors and format related issues which caused the data to be not usable for visualization until cleaning the data. The given table shows the dirty parts that were planned to be cleaned later on.

|  |  |
| --- | --- |
| **Untidy Version** | **Clean Data Planned to Be Made** |
| Year (In this case discrete) (float) | Date (integer) (Combine Year and Month and turn float into integer) |
| Month (In this case discrete) (float) | Date (integer) (Combine Year and Month and turn float into integer) |
| Product\_Category (object) (Inconsistencies such as !, \*, @) | Product Category, Removing the inconsistencies |
| Marketing\_Channel (Inconsistencies such as, !, \*, @) | Marketing Channel, Removing the inconsistencies |
| Discount\_Rate (values are 0-0.03) (float) | Discount\_Rate (values are between 0-30%) |
| Economic\_Conditions (Object) (Inconsistencies such as !, \*, @) | Economic\_Conditions, Removing the inconsistencies |
| Units\_Sold (discrete) (float) | Units\_Sold (discrete) (integer) |

# **DATA TIDYING AND CLEANING STEPS**

Data which is essential for research should be understood well and be clean as much as possible to make more accurate interpretations and inferences. Otherwise, these situations can create many consequences such as misinterpretation, misunderstanding and some logical mistakes in research. To avoid all these potential consequences, data cleaning should be detailed and well planned. Data cleaning and tidying were done completely using Python libraries. Pandas and Numpy are just Python libraries that were used in this project. All the cleaning was done according to the checklist.

Firstly, Pandas and Numpy imported to Python to use. After that the dataset loaded into Python. For understanding data, it is important to examine the variables and data datatypes. We used df.dtypes function through that step and we noticed that we need to change sama types to clean such as ‘Year’ and ‘Month’ column’s. For checking data’s tail and head we used df.head() and df.tail() functions. After that, we define dictionary to fix column names for example ‘REGION’ to ‘Region’. Also we used df.columns for defining that dictionary . To standardize string columns and fix typos such as ‘@‘ argument problem we used str.replace() function. Also, we followed similar steps to clean numeric columns to avoid unwanted symbols such as ‘$’ with using .replace() function. Moreover, for scaling and cleaning ‘Discount Rate’ according to the data information that rate should be between 0 and 0.3 we defined clean\_discount\_rate() function and in

that function we used apply() function to preserve valid values and .replace , np.nan to handle empty strings as Na. Also, we multiply by 100 to scaling ‘Discount Rate’ with using column\*100 at one of lines that near the end of the cleaning code. After this step, for converting numeric columns to appropriate types we used .astype function such as transforming. ‘Year’ type to ‘Int64’ type. Another important function that we used to fill ‘Na’ values with mean values related to checklist information that says this, we used .fillna() and .mean() functions. We used df.drop\_duplicates() for removing duplicates from datas. To clean categorical columns, we defined a function that named ‘normalize\_category’ and after definition we used this function to all categorical columns. Another function that we used to standardize region column values that named capitalize\_each\_word() works for if value is string capitalize the first letter of each word. ‘region\_mapping’ dictionary is used for fixing region names. After this lines, we used clean\_marketing\_channel() and similar functions that have same goal for other column’s to standardize all of the values and fix common typos. To check examine descriptive statistics to control our cleaning process we used df.describe() function. After analyzing these values, we step out to check outliers with using interquartile range. Therefore, we write Q1 and Q3 with using .quantile() function and we used to interquartile formula which is equals to Q3 - Q1 and this lines are in function that we defined as outlier\_columns() function. However, we defined a different function to use on ‘Discount\_rate’ because of the need to scale between 0-0.3. Also, we round numeric columns to better visualizations with using .round() function and for capitalize string values we used .map() function. After that we check every function to work properly, finally we combine ‘Year’ and ‘Month’ column into ‘Date’ column with using a string addition method to finish every step that given in the checklist.

After all of these steps the data is cleaned according to the checklist and become ready to EDA (Exploratory Data Analysis) step. The process fully checked against any logical mistakes and according to the checklist. Pandas and NumPy, both libraries we used well, and they give us a chance to clean data faster, efficiently and more accurately than using normal Python.

# **Exploratory Data Analysis**

It is very difficult, almost impossible for people to make sense of pure data. Various software should be used to eliminate this confusion. While this project was carried out, Matplotlib, Seaborn and Plotly libraries were used for data visualization, while the NumPy library was used to find important numerical statistical data. After this, we interpreted the obtained data with 6 research questions and concluded various findings.

# **Research questions**

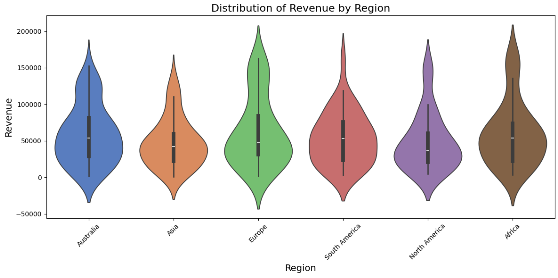
## How do different marketing channels impact customer satisfaction, and which channels consistently perform better or worse?(Oğuz Emre Üstün)

metin, ekran görüntüsü, diyagram, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

#### In this graph, the distribution of customer satisfaction according to different marketing types is shown as a box plot. When we look at general satisfaction, customer satisfaction varies between 1 and 5 for all marketing types. However, there are significant differences in their means and distributions. If we look at the average satisfaction, TV has the highest median value. This can be seen that marketing via TV was evaluated positively by customers. Email and radio are among the genres that receive positive reactions after TV. However, social media has a low median compared to other types. In addition, radio appears to have more consistent customer satisfaction with a narrower box plot, but social media and TV have a very wide distribution, which is a sign that they are not very consistent.

## How do revenue distributions vary across regions, and what factors might explain the differences in central tendencies and dispersion patterns among the regions?(Arda Öztürk)

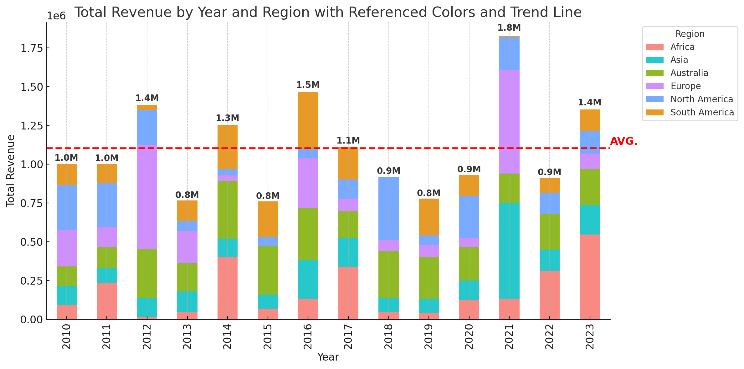


## The violin plot indicates revenue and region data including regions : Asia , Europe , Sout America , North America , Australia and Africa.The y-axis shows numbers between -50000 and 200000.Europe illustrates a a highest median revenue compared to others , recommending its enchanced performance.Also large amount of data in Europe under ≈40000.On the other hand , South America and present narrowest plot and lowest value of revenue .The Australian and Asian medians represent complete symmetry. Australia has wide distribution with greater variability with approximetaly range of 0 between 190000. Additionally, North America has the least diversity according to the violin chart. According to these trends, factors such as economic conditions, sizes of market and operational challenges may affect revenue.

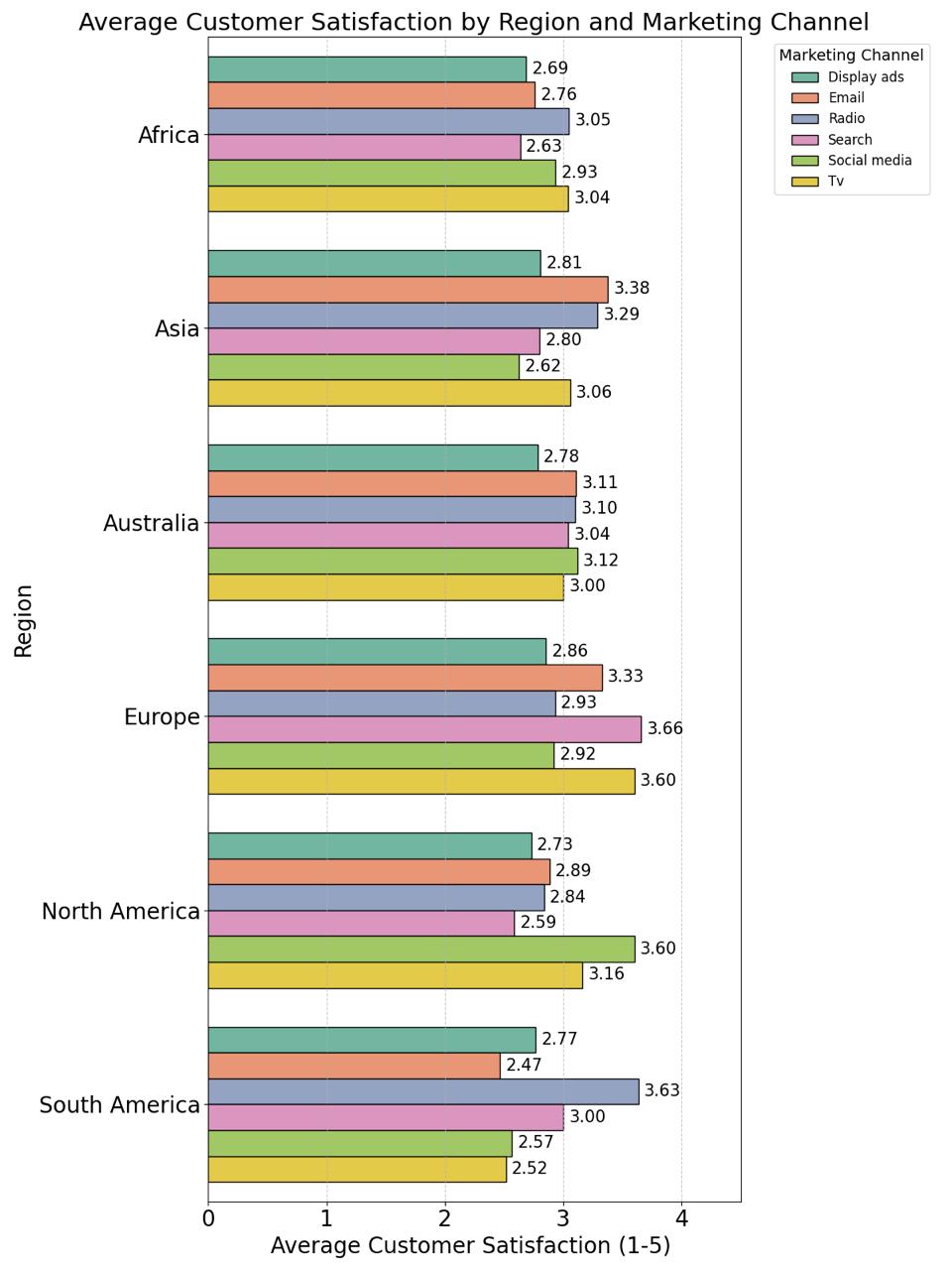
## How does the average revenue vary across regions and product categories, and which combinations yield the highest or lowest revenues? (İbrahim Barakat)

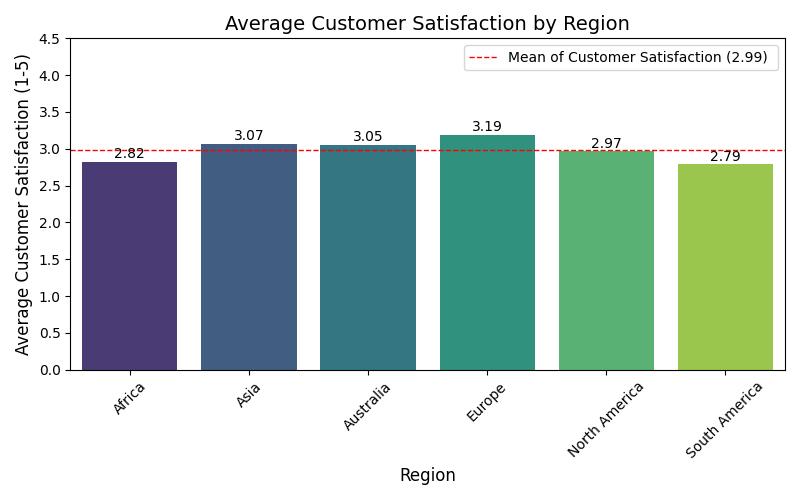
The heatmap shows significant variations in average revenue across regions and product cartegories. We can see that Africa is the best performing region, having generated the largest revenue in the Sports and Home Goods categories 93027 and 80644 respectively however the region lacks revenue in the Toys category. We can also see that Europe has a high average revenue in the Sports and Toys categories having average revenues values of 80280 and 77149 respectively. Looking at specific products of specific regions, we can see that Asia generated the lowest average revenue for the Toys cateegory. Further investigating the heatmap for lower revenue generating regions, we can see that North America is lacking in revenue from Electronics with an average revenue of 36659 and Europe is lacking at Home Goods revenue with an average revenue of 37475. This heatmap gives a concise and clear understanding of how average revenues vary across regions and product categories.

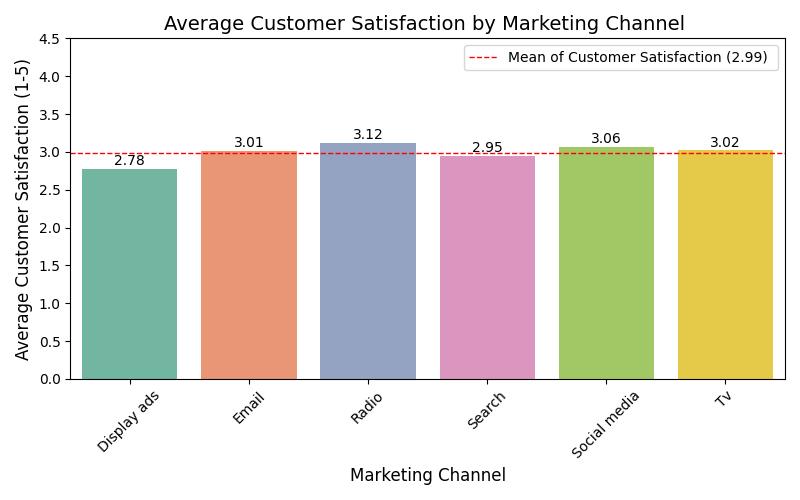
## How do the contributions of each region (Africa, Asia, Australia, Europe, North America, South America) to the total revenue vary over the years, and what trends can be identified in their growth or decline? (Onur Demir)

When we look at the data, the shares in total income remained constant in half of the regions and changed from time to time in the other half. The dominant region has been Europe most of the time. Despite this, it has been observed that its effect has decreased after 2020. If we consider the regions one by one, Africa has remained the lowest contribution. Its average contribution is around 0.1M every year. Its general trend is a decreasing contribution. When we look at the Asia region, its contribution is stable and generally between 0.1M-0.25M. When we examine Australia, while it was around 0.15M in 2010, this amount has decreased to around 0.1M by 2023. When we look at the Europe region, it has become the region with the largest share in total income. Its contribution has generally varied between 0.35M and 0.25M from 2010 to 2023. It has been the dominant region of the group. The contribution of the North America region has been quite low. It has not shown a significant change in general. Its contribution has mostly remained between 0.05M and 0.1M. When we examine South America, it has remained stable. Its contribution has generally fluctuated between 0.1M and 0.08M. If we interpret the trends, Asia, Africa and South America have followed a stable trend. The dominant region has been the European region, although it has slightly decreased after 2020.

## How do customer satisfaction levels vary across different regions and marketing channels?(İsmail Özdemirel)







The graph in Visualization 6 is a grouped bar chart that shows average customer satisfaction by region which includes all continents of world except Antarctica and by marketing channel that includes 6 types of different channels that combination of displaying ads, radio, email, searching, social media and tv. Each color represents a different type of marketing channel such as yellow shows tv in the graph. Also, this graph includes the average number for customer satisfaction, which is between 1-5 for every region and marketing channel and these numbers are written at the side of each bar. Search channels in Europe give the best results across all regions and marketing channels with 3.66 average customer satisfaction and radio channels in South America follow this with 3.63 average customer satisfaction. Oppositely, email channels in South America have the lowest customer satisfaction with 2.47 average customer satisfaction and tv channels in South America follows this with 2.52 average customer satisfaction. The mean of customer satisfaction in all areas and marketing channels is 2.99 and median is 3.0 which numbers that quite close to each other this means that the distribution of customer satisfaction is similar to normal distribution and balanced. Standard deviation of customer satisfaction data is 1.11 which shows many scores are likely close to the average, there are some regions or channels with notably higher or lower satisfaction. It’s important for getting closer into data to look separate to regions and marketing channels. According to Visualization 6.1 , Europe has the highest average customer satisfaction with 3.19 and it’s important to remember that the mean of all customer satisfaction was 2.99 so this means that Europe has far above from mean. Also, Asia and Australia have higher customer satisfaction than mean with respectively 3.07 and 3.05 average customer satisfactions. South America has the lowest average customer satisfaction with 2.79 and Africa follows them with 2.82 average customer satisfaction. However, North America is another region that has an average of customer satisfaction below mean with 2.97 which is very close to mean. According to Visualization 6.2, radio has the highest average customer satisfaction with 3.12 and social media follows them with 3.06 average customer satisfaction. Oppositely, displaying ads has lowest average customer satisfaction with 2.78 and searching has the second lowest average with 2.95. Email and Tv marketing channels which have quite above mean average customer satisfaction with respectively 3.01 and 3.02. That is important to notice that the distribution of average customer satisfaction in marketing channels is quite narrow and generally closer to the mean than regional distribution. In conclusion, customer satisfaction levels can change relative to region and used marketing channels. For instance, even radio has highest average customer satisfaction, social media in Europe gives the best results with 3.66 or another example is social media in South America which has 2.57 average customer satisfaction and far below mean even using marketing channel that has second highest average customer satisfaction. Therefore, marketing strategies should be planned to relate to project’s region. All of visualizations formatted for this question related to t IEEE format better such as Visualization 6.

## How does marketing efficiency vary throughout the year, and how is it related to the number of units sold each month? (Tarık Yılmaz)

metin, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, çizgi, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

The graph demonstrates a relationship between marketing efficiency and Units sold by month revealing certain trends. Marketing efficiency reaches its peak during the months of April and December reaching marketing efficiency values of 13.5 and 13.8 respectively. There is a big gap in marketing efficiency between the months of December and January, indicating that there might be a seasonal reasoning for the fluctuations. Additionally, there is major change rapidly between April and June. It decreases by approximately the value of 4.5. Moreover, marketing efficiency increased rapidly between October and December. Finally, it rose to its highest value in December with 13.8. Furthermore, units sold value has highest rank in January. Also, September and November have almost equal value in Units Sold value with approximately 39K.

# **Conclusıon**

The analysis showed that regional and marketing channel factors significantly affect revenue generation and satisfaction of customers. Europe consistently outperformed other regions in customer satisfaction and revenue, while social media showed lower and more variety in satisfaction levels compared to TV and radio. Revenue distributions varied significantly amongst regions and product categories. The findings show the necessity of how data could help increase revenue by optimizing the variables.