ANALYSIS OF PATTERNS AND TRENDS IN ADVERTISING PERFORMANCE

A FINAL PROJECT REPORT SUBMITTED
IN FULFILLMENT OF THE REQUIREMENTS FOR THE COURSE

STAT 112:
INTRODUCTION TO DATA PROCESSING AND VISUALIZATION

DEPARTMENT OF STATISTICS
MIDDLE EAST TECHNICAL UNIVERSITY

By

Muhammet Furkan Koç

Rabia Demircan

Ali Altuntaş

Mehmet Karaman

January 2024

1. Abstract

This research is about the number of sales and revenue generated by advertisements of certain types of products for each month between 2011-2022, region by region. Advertisements are generated for different products and create different data depending on the features they contain. These features are advertisement type, slot, the number of clicks and views. The number of purchases varies because of these features. Based on the outcome of the data cleaning process and data visualization to answer research questions, the data set became more understandable.

2. Introduction

The data set contains 12 years and the months of the years, and each month evaluated with 9 different criteria. The variable names shown here are the original values that are in the raw data. These names were changed in the data tidying steps.

Variable name	***Description***	***Scale***
YEAR and MONTH	Represent the time period of the observations.	Quantitative
**Product_Type (Product_Type)	The type of product advertised	Qualitative
Ad_Type	Nominal variable indicating the type of advertisement (e.g., Display, Video).	Qualitative
Ad_Slot	Nominal variable indicating the type of advertisement (e.g., Display, Video).	Qualitative
the number of times the ad was viewed (Impressions)	Continuous variable representing the number of times the ad was viewed.	Quantitative
Clicks	Continuous variable indicating the number of times users clicked on the ad.	Quantitative
CTR (Click_Rate)	Calculated as Clicks/Impressions * 100, represents the percentage of viewers who clicked on the ad.	Quantitative
SALESSS (Sales)	Discrete variable indicating the number of products sold as a result of the advertising.	Quantitative
Revenue	Continuous variable representing the revenue generated from the sales.	Quantitative
Region	Nominal variable specifying the geographic region of the advertising campaign (e.g., North, South, East, West).	Qualitative

As it is seen from the table, variable names have no standardization. Categorical columns involve capitalization mistakes, and some months are written as abbreviations, while others are written without using abbreviations; also some variables are written in quotes. The data contained spelling errors, too. Some mistakes were made while writing numerical variables and the data set contains empty columns. The process of tidying and cleaning the data set to address such issues can be found in the following sections. Exploratory data analysis gives further insight about the data.

2.1. Research Questions

5 research questions were prepared regarding to the data set:

1. Which advert type could be considered the most successful?
2. Is there a notable difference in revenue due to seasons?
3. Do regions affect sales, are there weak or strong correlations that can be established in this interaction?
4. Does the number of views have a direct impact on sales? If so, can the revenue earned from these sales show that a strategy to increase the number of views is profitable?Is there a significant change in the number of ad views over the years? If so, what could be the reason for this?
5. Which type of products' ads have generated more revenue?

3. Data Tidying and Cleaning

In the data pre-processing stage, the goal was to find and fix the inconsistencies in the data set. In this process; the dataset has been checked if it has any inconsistencies such as poorly named titles, columns that need tidying, wrong data types, negative values in numerical columns, missing values, values with wrong formatting, and duplicated rows.

First step was to import and inspect the data. .xlsx file was imported as a pandas dataframe. The data was inspected afterwards by using .describe(), .info(), .head(), and .tail() methods. Existing and expected data types were examined and compared. Values were checked if they were consistent within their columns. Unique, duplicate and NA values were examined to see the necessary cleaning steps There were a total of 4 NA values; 1 in "Clicks", 2 in Sales and 1 in "Revenue".

Next step was to correct the column names. Some of the titles needed manual renaming while some of them were corrected using the pandas method .str.title(). Column names were changed according to the data description:

- "**Product_Type" was renamed as "Product_Type"
- "the number of times the ad was viewed" was renamed as "Impressions"
- "SALESSS" was renamed as "Sales"
- "CTR" was renamed as "Click_Rate"

After that, object type columns were taken into processing. Value counts were taken by using .value_counts() method for each column. At this point it was necessary to handle the "Click_Rate" column as it was supposed to be a numerical variable, yet it was shown as an object column. The column contained string and float values together. Also, "Click_Rate" was constructed by "(Clicks/Impressions)*100" according to the data description, so it was convenient to use "Click_Rate" to reverse-calculate "Clicks" and fill the NA values there.

To accomplish that, "Click_Rate" values were converted to strings to apply .strip() methods, and stripped of any strings and whitespaces that occurred. Next, the decimal separator was set to dot instead of comma as Python uses dots to represent decimal values. Afterwards, the values were converted to floats and a demo "(Clicks/Impressions)*100" calculation was performed to see the correct range. It was determined that all values should be larger than one and values that are smaller (which were detected in the previous inspections) were multiplied by 100 to move the decimal place by 2 digits. A value count was taken to see

whether the values are in correct range, and “Clicks” were re-calculated by (“Click_Rate”*”Impressions”) / 100.

Before this step, “Impressions” was converted from int to float as description suggested it was continuous. Then, it was validated that there were no NA values in “Clicks” and the range was corrected as the negative values were gone.

At this stage it was time to tidy object type columns. There were no NA values in them so only string renaming was performed. A custom-made, basic cleaning function, str_basicclean() was called to automatically perform standard operations such as capitalizing and whitespace cleaning. A value count was taken and it was seen that “Product_Type” and “Ad_Type” columns were cleaned by str_basicclean(). “Month” values were capitalized and abbreviated manually as the majority of the values were that way, “Ad_Slot” had only one value to be manually replaced and “Region” was stripped of by some special characters and capitalized by again deploying str_basicclean(). Object columns were tidy afterwards.

Next step was to tidy numerical columns. Again, a value count was taken to see what was what. All columns were checked for negative values and it was found that there were none, and the only column that could have negative values by the logic of the data was “Revenue”. Regardless, the absolute value of all numerical values were taken with the .abs() method to make sure. Values were inspected again to see results. NA values were filled in “Sales” (2) and “Revenue” (1) with median. “Sales” was converted from float to int as data description suggested that the values were discrete.

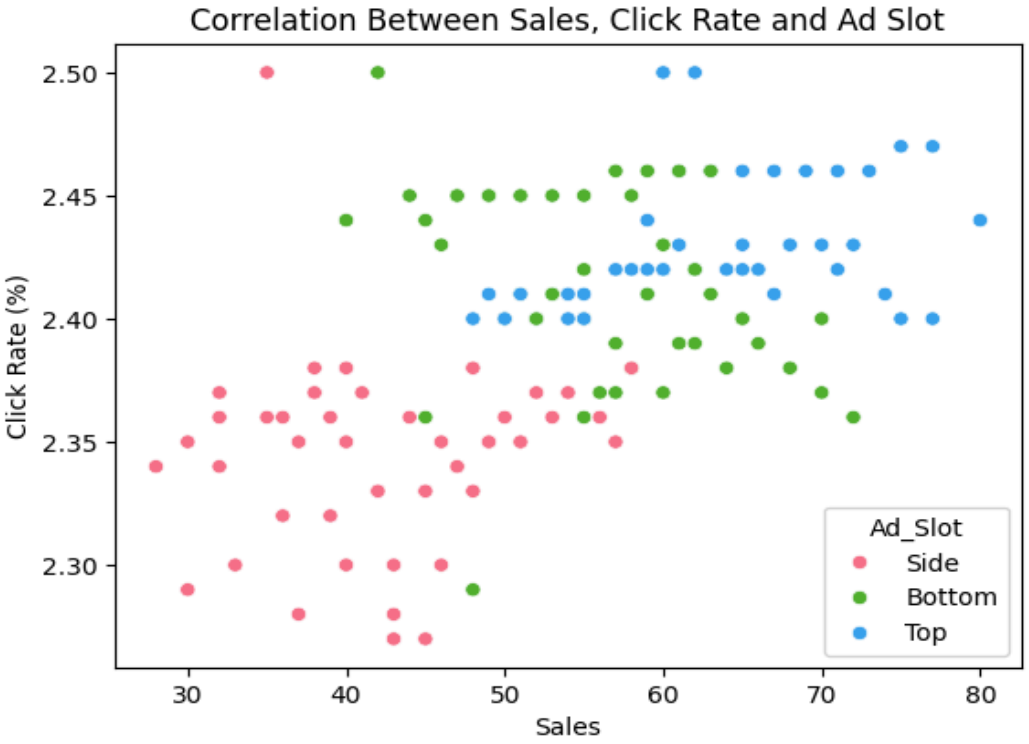
A “Date” variable was constructed from “Year” and “Month”, with datetime variable type, as it would make more sense to use it that way for a time series analysis. First, month abbreviations were swapped with integers that correspond to the month value. Then, column names were lowercased to “year” and “month” to comply with pandas’ .to_datetime() methods’ format, and the “Date” column was constructed. .dt.to_period(“M”) method was called on the column as there was no day variable in the dataframe, and values made more sense if they were monthly period-type. “Year” and “Month” were dropped; head, info, duplicates and NA values were checked once more, and the cleaning process was concluded. (Outliers were not inspected per instructor’s advice.)

	Product_Type	Ad_Type	Ad_Slot	Impressions	Clicks	Click_Rate	Sales	Revenue	Region	Date
0	Electronics	Display	Top	50000.0	1200.0	2.40	50	5000.0	North	2011-01
1	Clothing	Video	Side	35000.0	801.5	2.29	30	3000.0	West	2011-02
2	Beauty	Display	Bottom	45000.0	1098.0	2.44	40	4000.0	East	2011-03
3	Electronics	Video	Top	60000.0	1500.0	2.50	60	6000.0	South	2011-04
4	Clothing	Display	Side	40000.0	1000.0	2.50	35	3500.0	North	2011-05

4. Exploratory Data Analysis

	count	mean	std	min	25%	50%	75%	max
Impressions	144.0	54333.333	10939.120	32000.00	46000.00	55000.00	62000.00	80000.0
Clicks	144.0	1304.235	276.702	748.80	1098.00	1298.25	1500.60	1952.0
Click_Rate	144.0	2.395	0.051	2.27	2.36	2.40	2.43	2.5
Sales	144.0	53.799	11.559	28.00	45.75	54.00	62.00	80.0
Revenue	144.0	5389.583	1155.148	2800.00	4575.00	5500.00	6200.00	8000.0

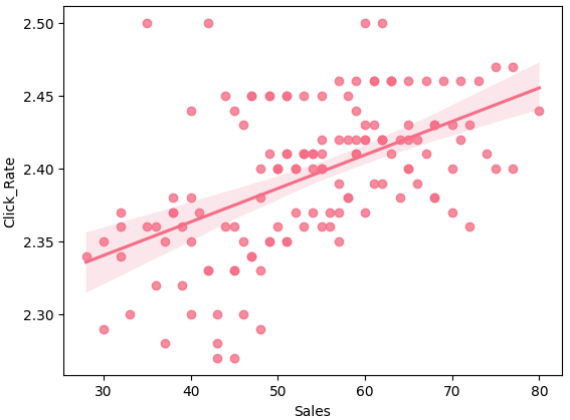
Research Question 1: *Which advert type could be considered the most successful?*

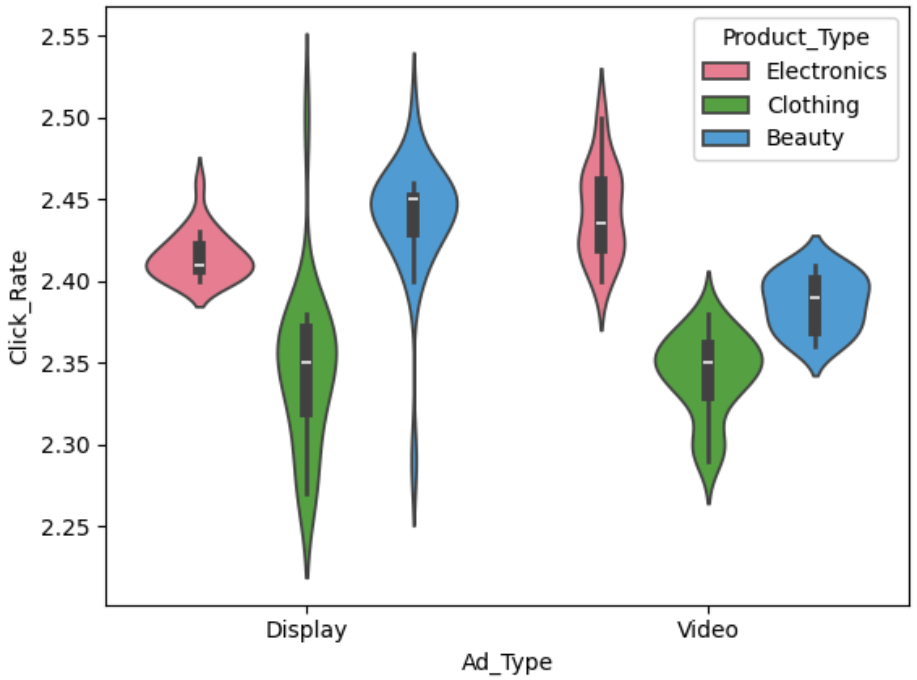


In this question there are multiple variables that are compared to determine the most successful advert type in terms of product type, ad placement and ad type. The number of sales and the click rates have an obvious relationship between the success of the ad so in the first step of the question the relationship between click rate and the number of sales should be compared according to place of advertisement.

Using a scatter plot is appropriate as both click rate and sales numbers are numerical variables. 3 different colors were chosen to represent the different ad slots so different types can easily be noticed without difficulty. The plot shows that there is a weak positive relationship between click rate and the number of sales. As the click rate increases, the number of sales increases too.

According to the scatter chart, ads at the top led to more sales. Consumers are more likely to click and watch ads if they are at the top, in which case it can be said that there is a slight positive correlation between the two. It can be also seen that the ads on the side have lower impressions and sales figures. Here it can be seen that there is a negative correlation between side ads and sales ads. There is a non-linear relationship between two variables when the ad is slotted at the bottom. As a result, it can be concluded that positioning ads at the top is the most advantageous situation for an ad.





It is determined that as click rate increases, the level of sales also increases and placing ads in top position is better to maximize click rate. There are 2 types of ads, these are display ads and video ads.

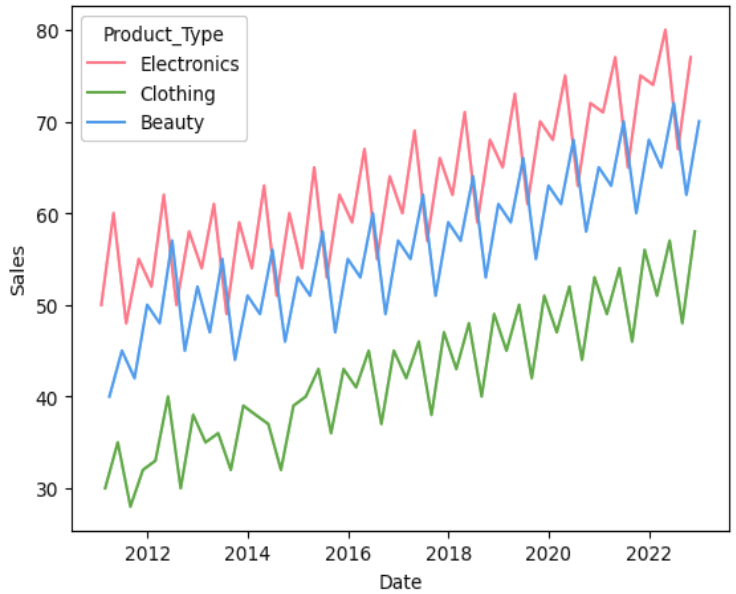
Violin plots are an appropriate way to compare categorical and numerical variables. Ad type and product type are categorical variables and click rate is a numerical variable. Different colors represent the different product types in this plot and usage of different colors make it easier to compare.

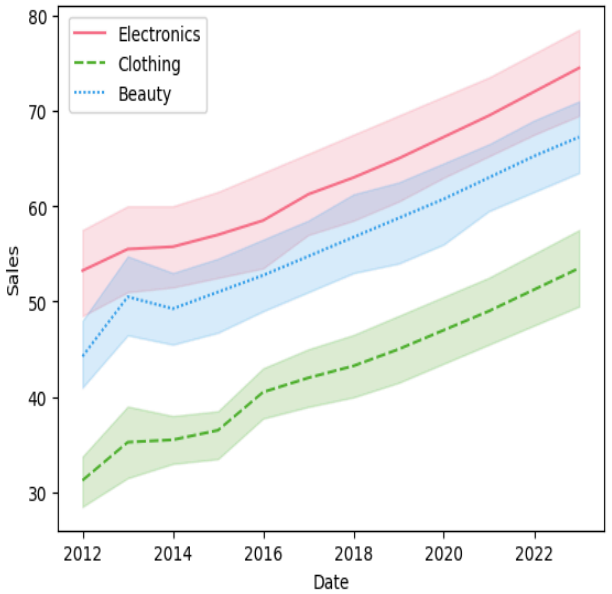
Display-type ads of clothing products' range is larger than the others and this type achieves the highest click rate. Also in other product types, display ads got more interaction than the video ads. Creating display ads for beauty products and clothing products generates more revenue and makes more sales. However, in electronic products usage of video ads will be more advantageous.

As a result of previous graphs, display advertisements of clothing products are the most successful advert type.

Research Question 2: *Are there any notable differences in revenue due to seasons?*

In this question, one of the variables is the date and the other one is the number of sales. Visualizing the variables with a time series graph would be convenient. Time series are divided into three-month periods on the line plot on the left and yearly periods on the right.



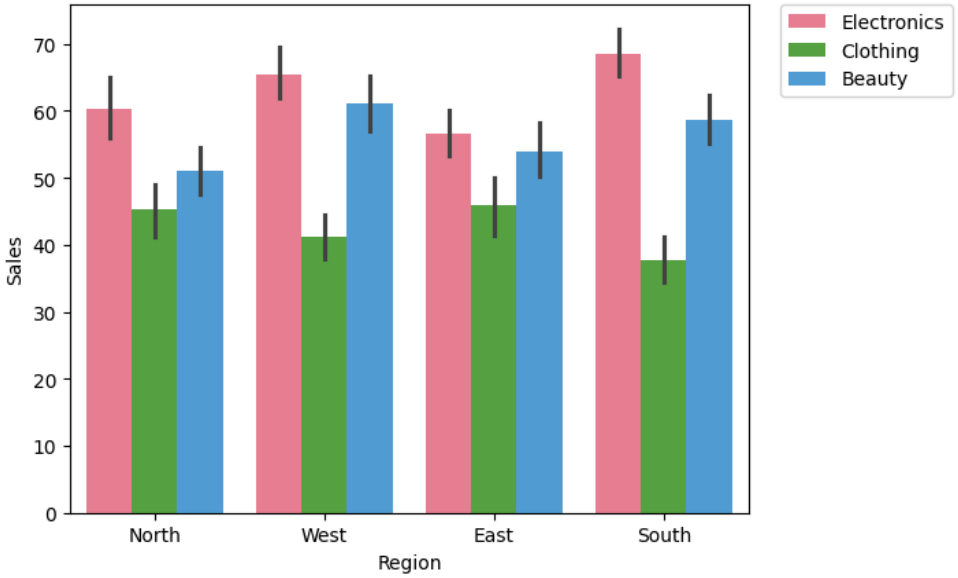


Representing different product types with different colors will help following each line without confusion. Analyzing each graph, it was found that:

- Electronics products were sold mostly in the second quarter of each year, with some exceptions. The number of electronics sales is increasing year by year. The electronics category had reached the highest number of sales when compared to the other categories.
- Beauty products were sold mostly in the second quarter, with a few exceptions. The change in the number of sales in each year has similar rates.
- Clothing products were sold less than the other categories and were mostly sold in the fourth quarter but there is no huge range between the fourth and second quarters. Even the highest number of sales of clothing products cannot reach the other categories' lowest number of sales.

To sum up, the second quarter revenues are higher than the other quarters in the beauty and electronic products categories with a huge number of variations.

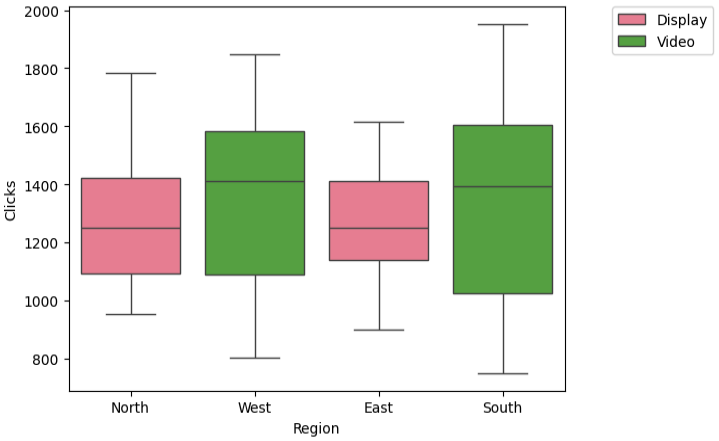
Research Question 3: *Does region affect sales, are there weak or strong correlations that can be established in this interaction?*



In this question, clustered bar charts were used to analyze each region and compare them with other regions. Representing each category with different colors makes the graph more understandable. The first step of the question is comparing each category and noticing the effect of the region.

The graph shows that, in all regions, electronic product ads generated higher sales than the other categories. For each region the descending order of the number of sales is electronics-beauty-clothing. Each products' sales numbers are close to each other. Almost the same shape can be seen for each region. In previous steps, there was a positive linear relationship between sales and interactions. Usage of box plot is appropriate here to assess the range of the clicks and compare region's effect on clicks.

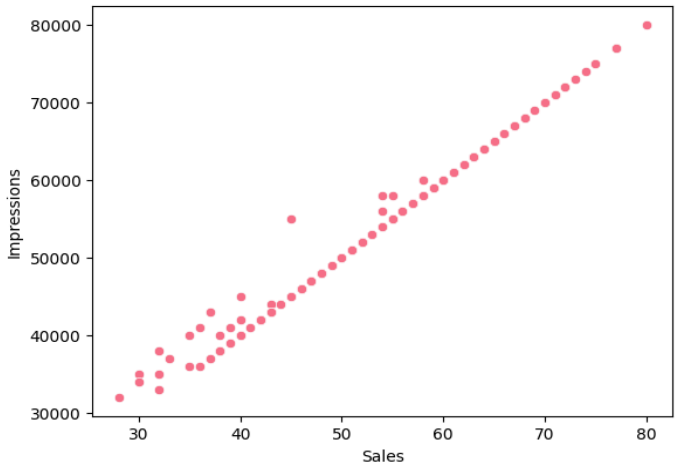
- The distribution of clicks in the North is slightly right skewed.
- Clicks for the ads that are shown in South and West graphs are negatively skewed. The median of the graphs is almost equal for North/East and West/South. The ranges of the graphs are different, and the range of clicks on south is higher than the others.
- Clicks for the ads which are placed in the East graph are positively skewed. The range of the values is smaller than others.

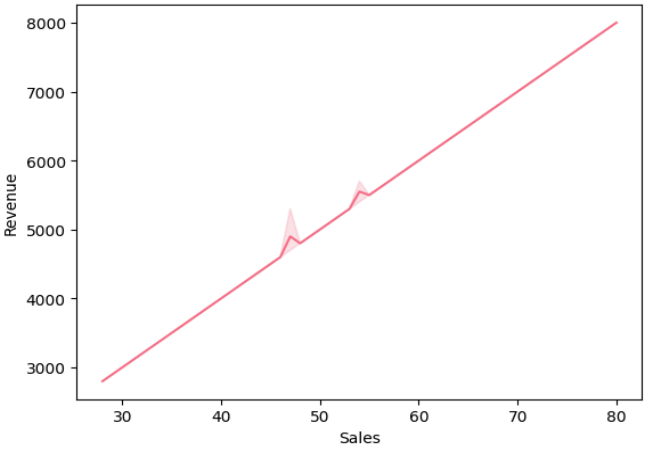


- Only display-type advertisements were used in North and East.
- Only video-type advertisements were used in the South and West.

As a result, it can be said that the region that ads are displayed affects the clicks, thus affecting sales figures.

Research Question 4: *Does the number of views have a direct impact on sales? If so, can the revenue earned from these sales show that a strategy to increase the number of views is profitable?*

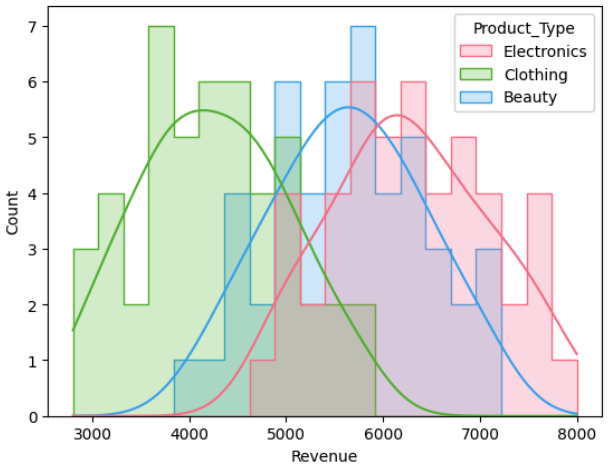




Impressions and sales are numerical variables, so using scatter plot will be appropriate to see the relationship between them.

Yes, there is a strongly positive relationship between impressions and sales figures. As the impressions of ads increases, the number of sales also increases with a few exceptions. If companies create interesting ads, they can use them as a strategic method and they can improve their sales figures. The following graph shows that there is a 100x difference between sales numbers and revenue, and it shows that there is a strongly positive relationship between these variables. Therefore, by using strategic advertising methods, companies can earn more money.

Research Question 5: *Which type of products ads have generated more revenue?*



Revenue is a numerical variable, and the frequencies of that variable were in question, so using a histogram is appropriate here. Different colors represent different categories, without different colors interpreting the graph would be difficult.

This histogram shows that the three product types exhibit a distribution shape close to the standard normal distribution. While it is seen that the highest revenue values are in electronics, the lowest revenue values are in clothing. All the three graphs' shapes are symmetric-like so their mean, median and mode values are equal or close within each product type.

To sum up, electronics products generated more revenue than the other types of products.

Conclusion

To summarize the project from the beginning; first, data had to be cleaned and the missing data was dealt with. This was done in Python using pandas and NumPy. 5 research questions were created immediately after. According to Figure 1 in the first question, it is seen that the number of sales increases when the click rate increases, but it can be said that there is a slightly positive correlation between them. It can be said that the most successful product segment among these advertisements is generally electronics. In the second question, second quarter revenues in the beauty and electronic products categories are higher than other quarters with a lot of diversity. In the third question, Figure 4 shows that regions affect sales figures. In addition, looking at Figure 5, there are radical differences in the type of advertising between regions. In the fourth question, it is seen that there is a direct effect between sales numbers and viewing numbers. In the fifth question, it is seen that the category that generates the most revenue is electronics.

GitHub Links:

Ali Altuntaş: https://github.com/ForxDeven/advertisement_analysis

Mehmet Karaman: https://github.com/memeth-my-hawk/advertisement-data-analysis

Muhammet Furkan Koç: https://github.com/Hamilcram/analysis-of-advertisement-data-visualization

Rabia Demircan : https://github.com/rabiademircan/analysisofadvertisement