Data Scientist Case Study

Author:

Navid Yamini

Email:

navid.yamini@studenti.polito.it

26/2/2023

Contents

Summary

This document provides an overview of the steps taken in pre-analysis, analysis, and data manipulation of a case study. In pre-analysis, it is important to check for data quality issues, understand the data structure and correlations, and consider adding more information about the products and customer demographics. The analysis phase involves visualizing the data using different plots and exploring the potential impact of a new marketing channel (TikTok). Data manipulation includes calculating the total money spent across all channels, aggregating daily transactions, merging tables, and creating new data frames. A heat map is used to study the correlation between different attributes. Weekly and monthly data frames are created to uncover patterns or insights that may not be apparent when looking at daily records alone. Following these steps, line graphs are used to present the daily data effectively, including the total number of transactions per day, revenue, and advertising expenses Furthermore, data analysis shows that Facebook, Google Ads (GADS), Amazon (lees than others), and TikTok have a significant correlation with the number of transactions. Finally, some Regression models were applied on the data, and Linear Regression on daily data yielded the best result. These steps help to ensure that the data is clean, relevant, and accurate for building the model and informing decision-making regarding resource allocation and marketing strategy.

1- Pre Analysis

Before starting any analysis, it is crucial to ensure the accuracy, completeness, and relevance of the data to the problem at hand. To achieve this, there are several steps that should be taken, including understanding the problem and defining the objective, checking for data quality issues such as missing values, outliers, and inconsistencies, and spending time understanding the data structure and correlations between different entities and attributes.

When analyzing a specific case study, it is essential to first load the data into a Pandas data frame in Jupyter and check the column names and first rows to understand the data types and advertising channels used by the company. Additionally, it is important to check for missing values, duplicates, and outliers. In this case study it was not needed.

To enrich this dataset, we can consider adding more information about the products being sold, such as their category, price, and popularity. This additional information can help provide context and insights into the sales data, and allow us to identify patterns and trends in customer behavior. Additionally, we could include data on customer demographics, such as age, gender, and location, to better understand the target audience and tailor our marketing strategies accordingly.

To gain a better understanding of the company's sales performance and the effectiveness of its marketing efforts, it would be useful to determine the daily and weekly sales targets set by the company. Additionally, by comparing the revenue generated from sales to the company's marketing expenses, we can assess the return on investment (ROI) of its advertising campaigns.

Furthermore, we might consider looking for external data sources that can provide additional insights into market trends or competitor activity. For example, we could gather data on consumer spending patterns and economic indicators to better understand how the overall market is performing. We could also analyze data from social media and online reviews to identify customer sentiment and preferences, and use this information to improve our products and services.

2- Analysis

One of the steps that can be taken in data analysis is to visualize the data, this can also be useful in gaining a deeper understanding of the data, and several types of plots such as histograms, scatter plots, box plots, and heat maps can be used to identify correlations between different attributes.

In Figure 1, a box plot is displayed that shows the amount of money spent by the company on each online channel. From the plot, it is clear that the most frequently used channels are Google Ads (GADS) and Facebook, followed by Amazon. On the other hand, Bing and Criteo are the least utilized channels. Interestingly, there is no data related to TikTok on this plot.

Upon further analysis, it becomes apparent that the company started using TikTok at a later stage than the other channels, and the first recorded data for this channel is from June 12th, 2022. By identifying this trend, we can gain a better understanding of the company's marketing strategy and adjust our analysis and recommendations accordingly.

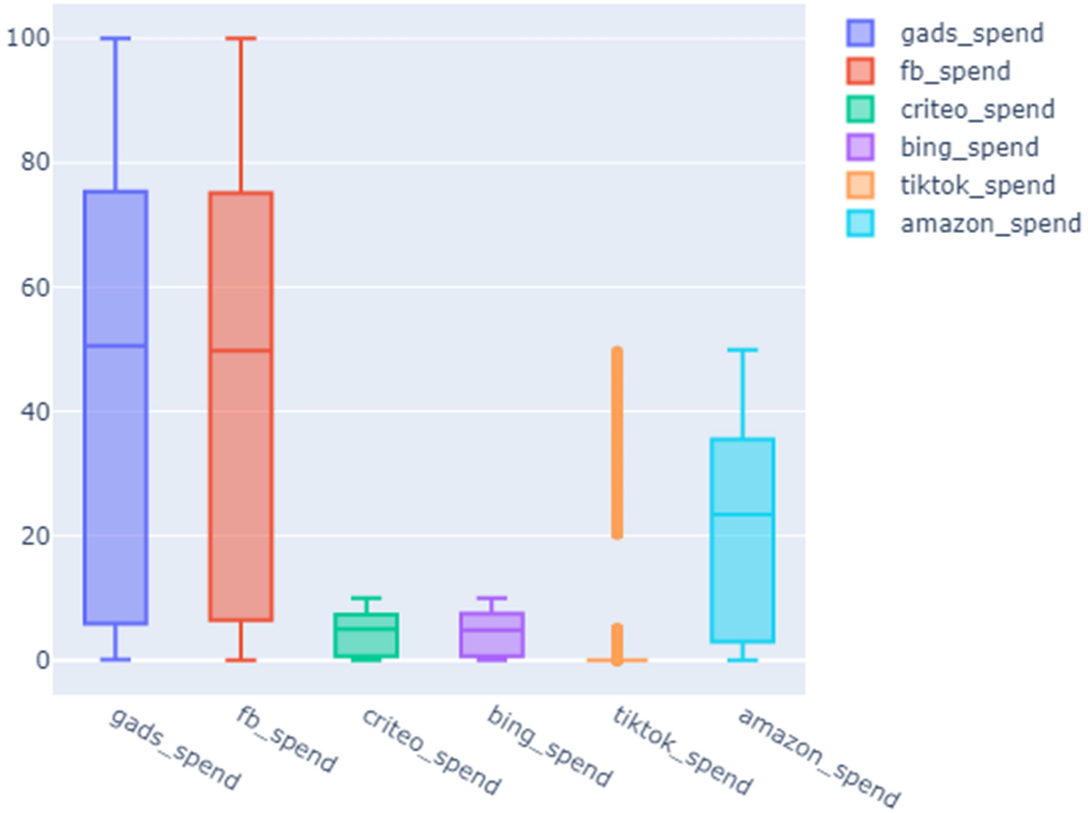


Figure 1: Daily Spend by Channel Box Plot

In Figure 2, a histogram graph is displayed, which highlights a notable trend in the data. Specifically, we can observe that the amount of money spent on TikTok is higher than that spent on Bing and Criteo, even though TikTok has been used for a shorter duration. This suggests that TikTok is a promising advertising channel for the company, and further analysis is warranted to explore its potential impact on the company's overall marketing strategy.

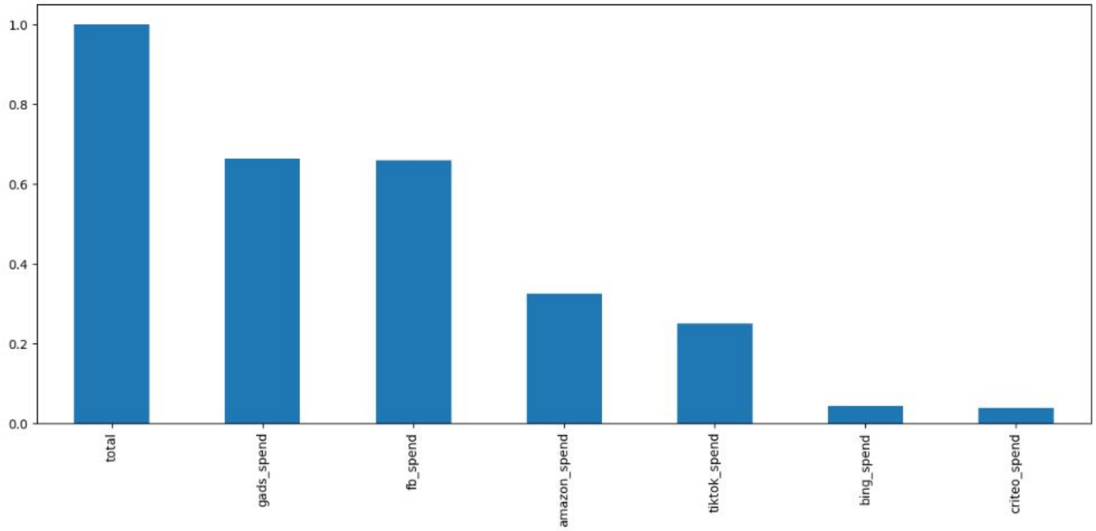


Figure 2: Daily Spend By Channel Histogram

2-1- Data Manipulation

To prepare the data for model creation, several steps have been taken. Firstly, the total money spent across all channels was calculated, followed by aggregating daily transactions to determine the total number of transactions per day. Next, three tables were merged on a common date to create a new data frame. Additionally, the old and new strings in the website version were replaced by 0 and 1 respectively. Finally, the date that TikTok was added to other channels and the date of migration from the old to the new website were extracted. These steps help to ensure that the data is clean, relevant, and accurate for building the model.

After exploring the data structure and verifying its quality, a heat map was used to study the correlation between different attributes. This approach allows for the identification of variables that may be strongly correlated, enabling a better understanding of the underlying relationships within the data.

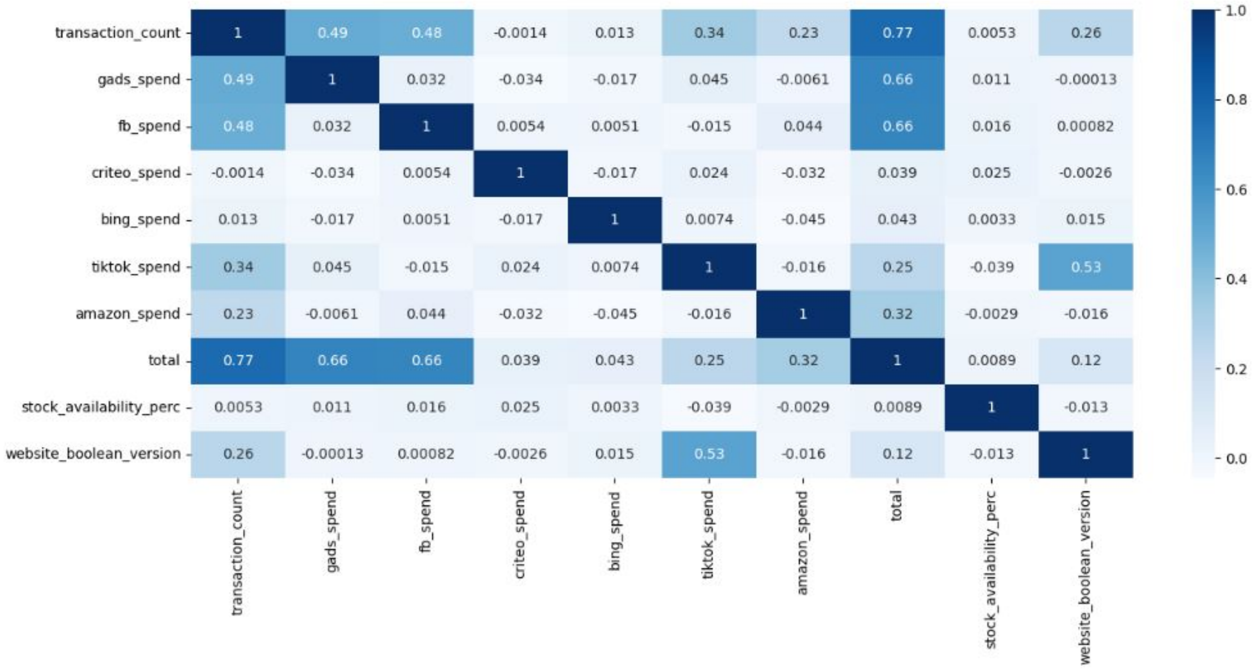


Figure 3: Final Data Heat Map

In Figure 3, the final heat map of the data is presented, providing a visual representation of the correlation between various attributes. The plot highlights a strong positive correlation between the number of daily transactions and the total amount of money spent across different channels. Specifically, Facebook and Google Ads (GADS) channels demonstrate a particularly strong relationship with the number of daily transactions, while Amazon and TikTok also display significant correlations, albeit to a lesser extent. Conversely, the heat map indicates a weak or negative correlation between the number of daily transactions and Bing and Criteo. These findings suggest that certain channels may be more effective than others in generating sales, and may help inform decision-making regarding resource allocation and marketing strategy.

Now that the daily records have been prepared for running the model, we have taken it a step further by creating four additional data frames to work with. In order to get a bigger picture of the trends in the data, we have aggregated the data into weekly and monthly data frames. For each one of these time frames, we have used both the sum and mean aggregation methods to get a more comprehensive view of the data. By doing this, we hope to uncover any patterns or insights that may not have been apparent when looking at the daily records alone.

Plotting a line graph for each of these data frames can provide us with valuable insights.

To present the daily data effectively, a line graph has been created to show the total number of transactions per day (Figure 4). Due to the large amount of daily data, only the total number of transactions has been included in this graph. Two vertical lines have been added to indicate the first day of TikTok's usage (2022-01-02) and the launch of the new website (2022-06-12). This line graph has been divided into three parts based on these vertical lines. By comparing the data from these periods with the corresponding days in previous years, we can observe an improvement in sales. Therefore, we can conclude that these two changes have had a significant impact on sales. For more details please check the Python script.

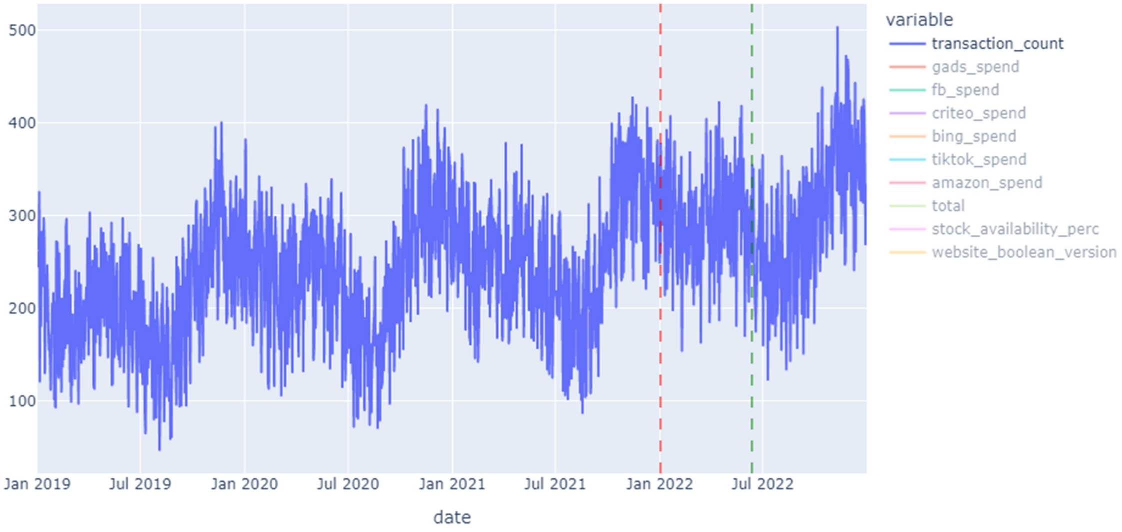


Figure 4: Daily Line Graph

Analyzing weekly and monthly aggregations can also yield valuable insights. By looking at the data aggregated weekly or yearly, we can observe some seasonal trends. Figure 5 and 6 displays a notable increase in the number of transactions during November, followed by a sharp drop from December to the first week of January. While it may not be as useful to sum up stock percentages in weekly aggregation, we can still see that the total spent, available stock, and number of transactions all experience a decline in the first week of the year. This information is illustrated in figures 5 through 7.

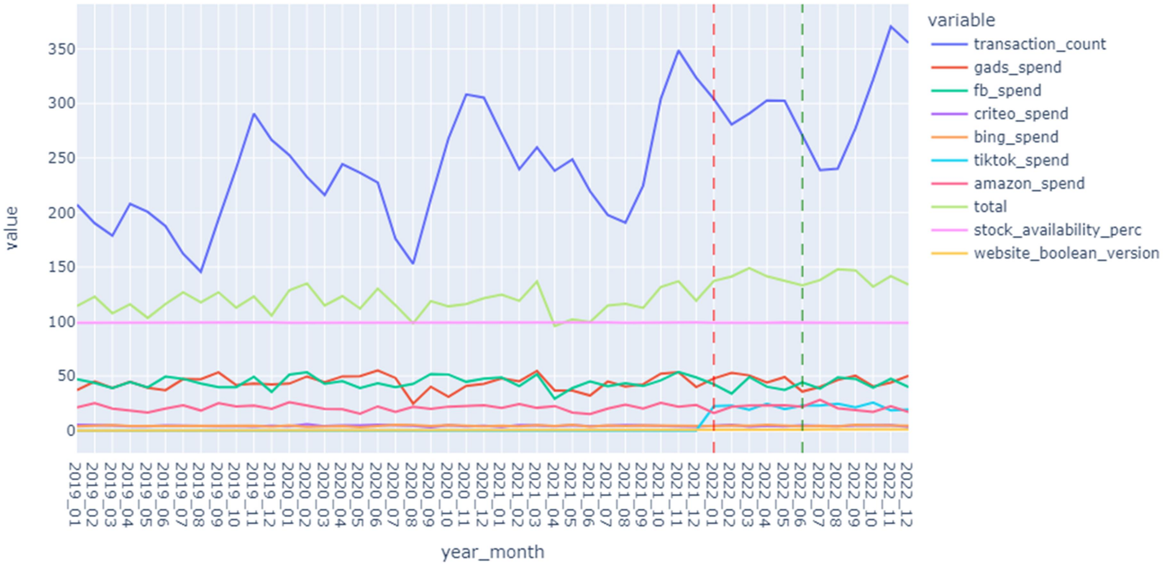


Figure 5: Average per Month

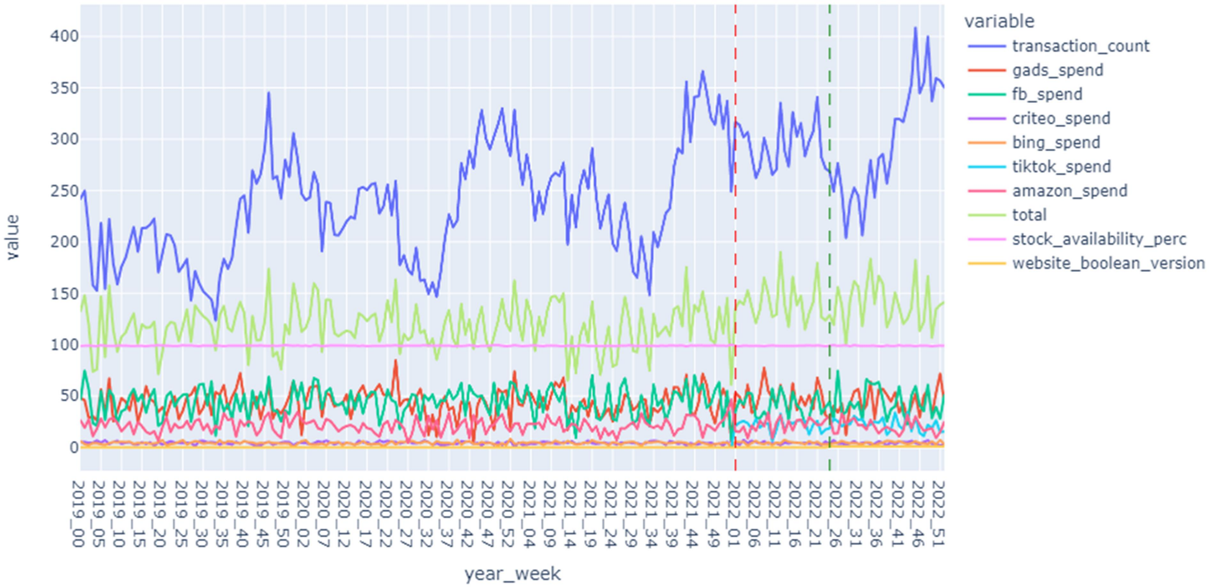


Figure 6: Average per Week

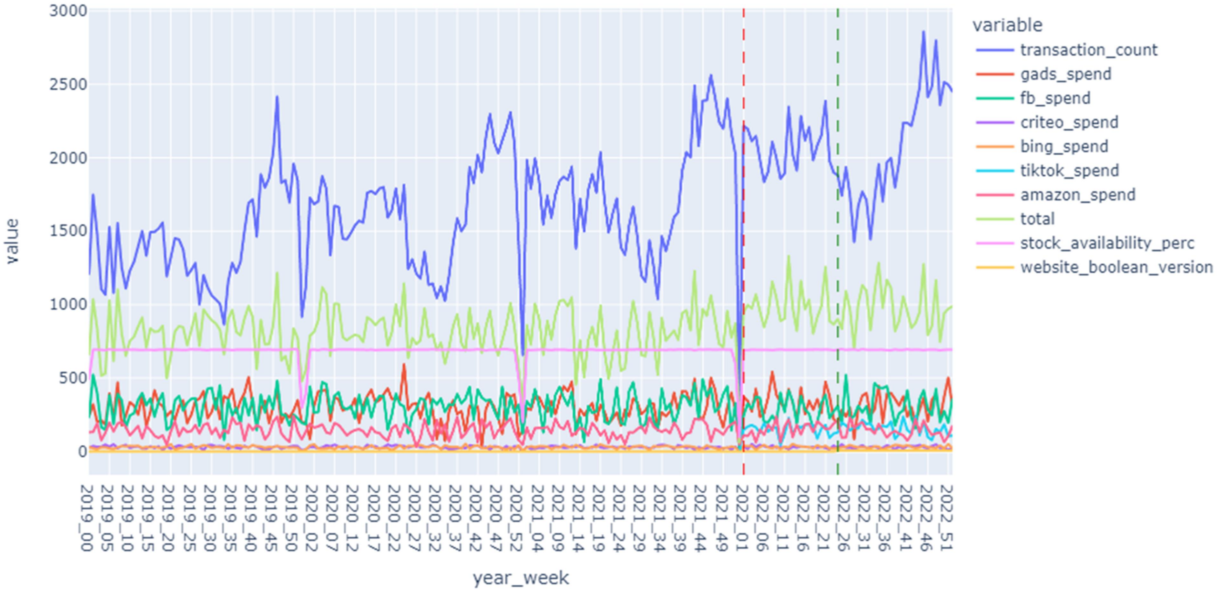


Figure 7: Sum per Week

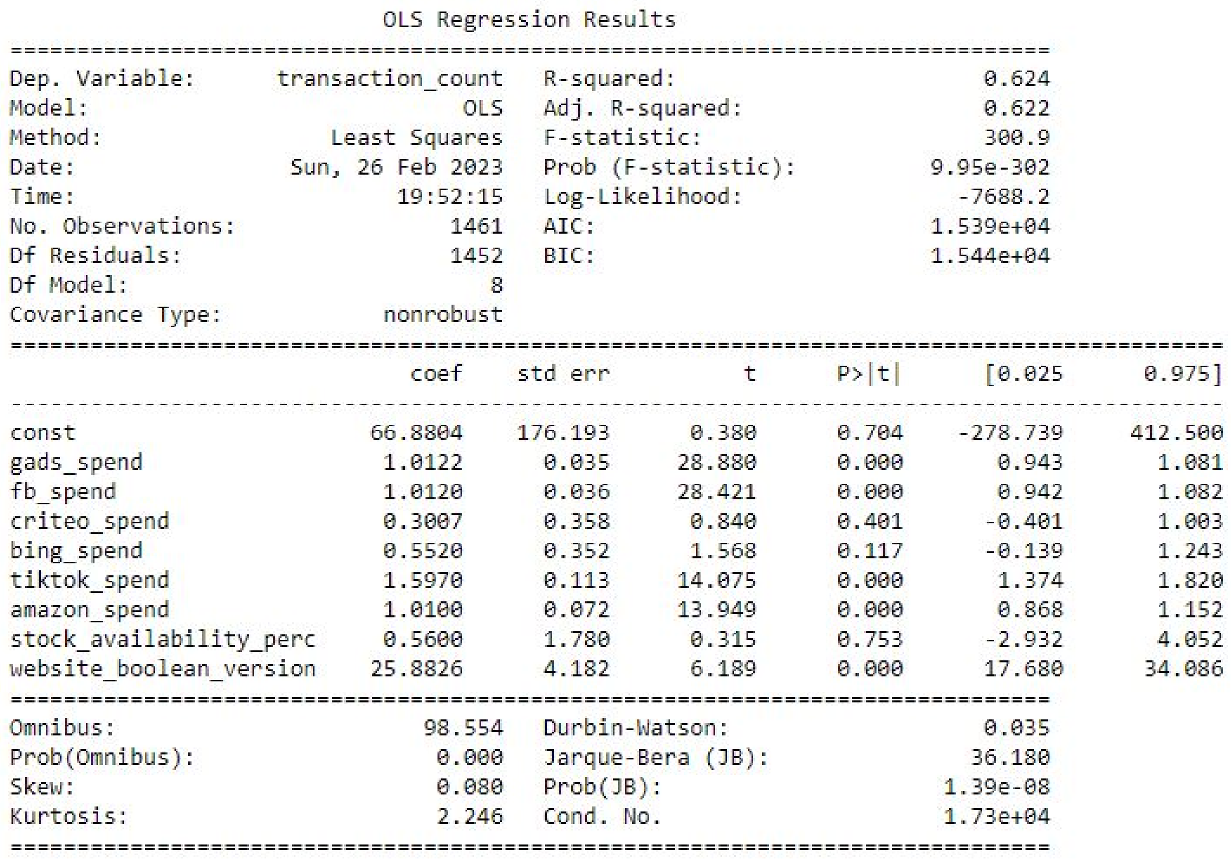
3- Creating Model

In this case study, the goal is to predict the number of transactions using regression models, as is typical in Marketing Mix Model analysis. To identify the most relevant attributes that correlate with the Y value, the Ordinary Least Squares (OLS) method was run on all generated data frames to calculate F-tests and

P-values. The p-value helps to select the most relevant attributes that are necessary to create the X set for the model. By looking at the p-value, we can identify attributes that have a statistically significant correlation with the number of transactions, which means that they have a strong impact on the outcome variable. This is important because including irrelevant variables in the model can lead to overfitting, which reduces the model's predictive power. Therefore, selecting only the relevant variables based on their p-values helps to improve the model's accuracy and predictive power. The attributes with strong correlations were then selected for the X set. Next, the data set was split into train and test sets, and various models were applied to the train data. The model with the lowest Root Mean Squared Error was chosen to run on the test data, and final scores were calculated. By following these steps, the most accurate prediction model for the number of transactions was created.

4- Results

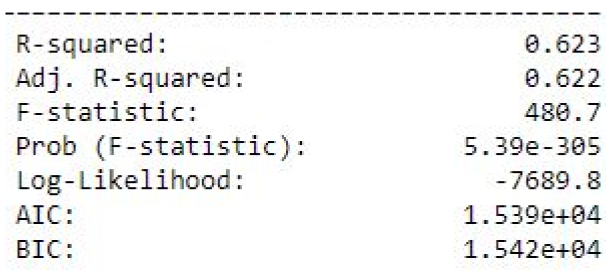
First step was to run the OLS on the data. Table 1 shows the OLS summery and reveals that the p-values for Bing, Criteo, and stock availability are greater than 0.05. As a result, it is better to exclude them from the X set.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:      transaction_count   R-squared:                       0.624
Model:                            OLS   Adj. R-squared:                  0.622
Method:                 Least Squares   F-statistic:                     300.9
Date:                Sun, 26 Feb 2023   Prob (F-statistic):          9.95e-302
Time:                        19:52:15   Log-Likelihood:                -7688.2
No. Observations:                1461   AIC:                         1.539e+04
Df Residuals:                    1452   BIC:                         1.544e+04
Df Model:                           8
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
const                      66.8804    176.193      0.380      0.704    -278.739     412.500
gads_spend                  1.0122      0.035     28.880      0.000       0.943       1.081
fb_spend                    1.0120      0.036     28.421      0.000       0.942       1.082
criteo_spend                0.3007      0.358      0.840      0.401      -0.401       1.003
bing_spend                  0.5520      0.352      1.568      0.117      -0.139       1.243
tiktok_spend                1.5970      0.113     14.075      0.000       1.374       1.820
amazon_spend                1.0100      0.072     13.949      0.000       0.868       1.152
stock_availability_perc     0.5600      1.780      0.315      0.753      -2.932       4.052
website_boolean_version    25.8826      4.182      6.189      0.000      17.680      34.086
==============================================================================
Omnibus:                       98.554   Durbin-Watson:                   0.035
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               36.180
Skew:                           0.080   Prob(JB):                     1.39e-08
Kurtosis:                       2.246   Cond. No.                     1.73e+04
==============================================================================
```

Table 1: OLS on Daily Data Frame

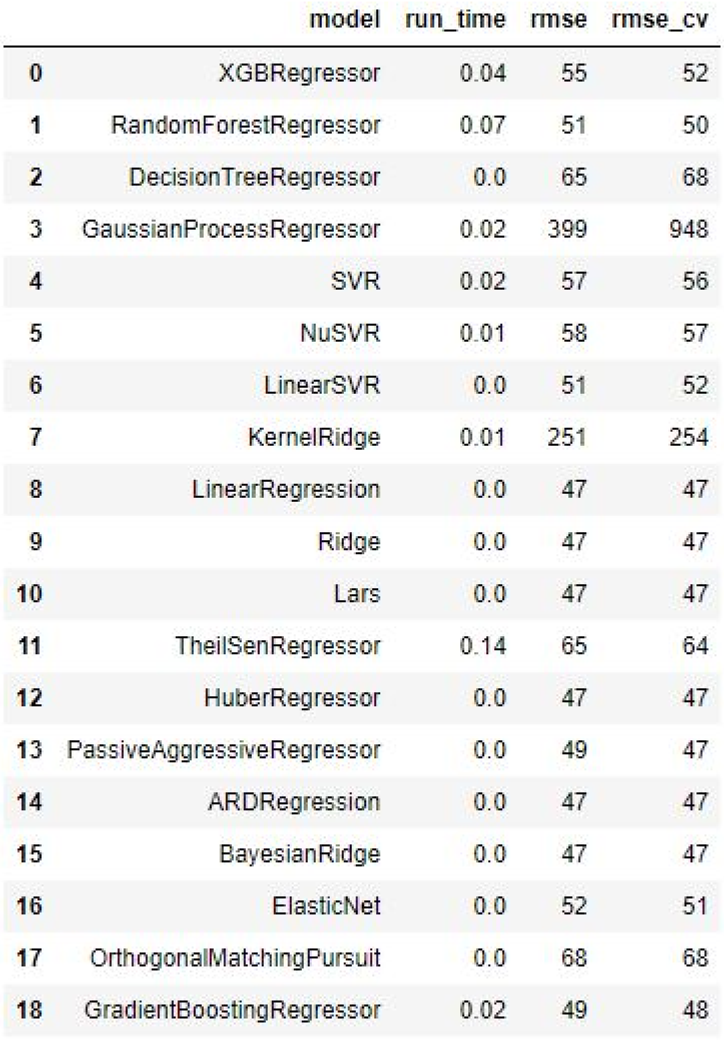
After removing the unwanted attributes, the OLS has been run again, and as a result the F- statistic value increases from 300.9 to 480.7 therefore the model's overall fit has improved. This is shown in table 2.



```
========================================
R-squared:                        0.623
Adj. R-squared:                   0.622
F-statistic:                      480.7
Prob (F-statistic):           5.39e-305
Log-Likelihood:                 -7689.8
AIC:                          1.539e+04
BIC:                          1.542e+04
```

Table 2: New F- statistic for Daily Data

To find the best model, various models have been tested on the data. The results of the models are summarized in Table 3, which shows the Root Mean Squared Error (RMSE) and the Cross-Validated Root Mean Squared Error (RMSE CV). Based on this result Linear Regression has been chosen.



	model	run_time	rmse	rmse_cv
0	XGBRegressor	0.04	55	52
1	RandomForestRegressor	0.07	51	50
2	DecisionTreeRegressor	0.0	65	68
3	GaussianProcessRegressor	0.02	399	948
4	SVR	0.02	57	56
5	NuSVR	0.01	58	57
6	LinearSVR	0.0	51	52
7	KernelRidge	0.01	251	254
8	LinearRegression	0.0	47	47
9	Ridge	0.0	47	47
10	Lars	0.0	47	47
11	TheilSenRegressor	0.14	65	64
12	HuberRegressor	0.0	47	47
13	PassiveAggressiveRegressor	0.0	49	47
14	ARDRegression	0.0	47	47
15	BayesianRidge	0.0	47	47
16	ElasticNet	0.0	52	51
17	OrthogonalMatchingPursuit	0.0	68	68
18	GradientBoostingRegressor	0.02	49	48

Table 3: RMSE and RMSE CV Evaluation

Table 4 summarizes the training and testing results of the model, and Figure 8 displays the actual and predicted line graph for cc per day.

Set	Score
Train	0.6191031854242777
Test	0.6286979173558918

Table 4: Linear Regression Mode Score

The results are not spectacular but are acceptable with these data sets.

All the model implementation steps discussed above have been applied to other data frames as well. However, to avoid making this report too long, the script output has not been included here. Instead, the results have been summarized in Table 5. Since the only significant value for Monthly data was TikTok, it is decided to ignore the running model on the monthly data set. For a more detailed analysis, please refer to the Python script.

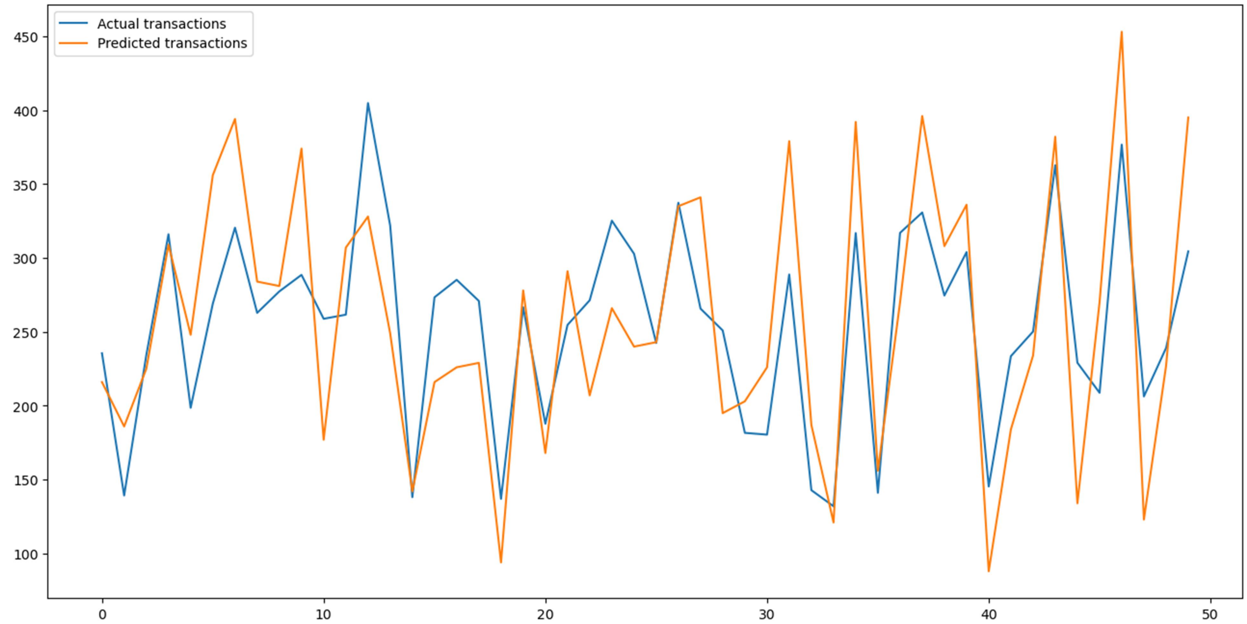


Figure 8: Prediction Result On Daily Set

	Daily Set	Weekly_Avg set	Weekly_Sum set	Monthly_Avg Set	Monthly_Sum Set
P-Value gads_spen	0	0	0	0.222	0.237
P-Value fb_spend	0	0	0	0.229	0.216
P-Value criteo_spend	0.401	0.072	0.072	0.113	0.138
P-Value bing_spend	0.117	0.273	0.273	0.274	0.286
P-Value tiktok_spend	0	0	0	0.014	0.015
P-Value amazon_spend	0	0.053	0.053	0.667	0.708
P-Value stock_availability_perc	0.753	0.684	0.684	0.946	0.491
P-Valuewebsite_boolean_version	0	0.368	0.368	0.823	0.825
First F-statistic	300.9	15.24	15.24	3.049	3.129
Second F-statistic	480.7	28.75	38.53		
Model	LinearRegression	LinearRegression	PassiveAggressiveRegressor		
Train Score	0.619103185	0.401294315	0.457211709		
Test Score	0.628697917	0.232804231	0.330090405		

Table 5: Summary of all Models

5- Conclusion

Based on our analysis, we found that Facebook, Google Ads (GADS), Amazon (although to a lesser extent), and TikTok have a significant impact on the number of transactions. It's worth noting that TikTok is a newer channel, and since its implementation, we've seen an increase in transactions. Additionally, the use of the new website has also impacted sales. While aggregating data can lead to some loss of accuracy, we found that Amazon had a good p-value in weekly aggregation, and TikTok still showed a positive relationship in monthly aggregations. Furthermore, our data showed a seasonal pattern where the number of transactions tends to increase significantly in November. In conclusion, we recommend investing more in TikTok and reevaluating the investments in Criteo and Bing to determine whether there is a correlation between their use and sales or if it's due to low usage.

6- Possible Improvement

To enrich this dataset, we can consider adding more information about the products being sold, such as their category, price, and popularity. This additional information can help provide context and insights into the sales data, and allow us to identify patterns and trends in customer behavior. Additionally, we could include data on customer demographics, such as age, gender, and location, to better understand the target audience and tailor our marketing strategies accordingly.

To gain a better understanding of the company's sales performance and the effectiveness of its marketing efforts, it would be useful to determine the daily and weekly sales targets set by the company. Additionally, by comparing the revenue generated from sales to the company's marketing expenses, we can assess the return on investment (ROI) of its advertising campaigns.

Since some seasonal behavior has been detected in data it is possible to use Time-series models such as ARIMA, SARIMA, and Prophet can be used to model the sales data over time. These models can capture the seasonality, trends, and other patterns in the data that may affect sales.

And at end the, in case of having large data set it is possible to use neural networks. Neural networks are powerful algorithms that can model complex relationships between the marketing spend and the sales. They are particularly useful when the data is high-dimensional and non-linear.

7- Long Story Short

- Which actions would you take before starting any type of analysis?
    - check the size of the data, understand the data, understand each entity and each attribute, and ensure the accuracy, completeness, and relevance of the data to the problem
- What do we want to make sure of?
    - No missing value, no duplicate record, no outliers, and completeness of entities
- Which data/info do you think could enrich this dataset?
    - sales data, product category, price, information about customers, weekly targets,
- Could you please develop a Marketing Mix Model (Extra Awesome: develop the model yourself)?
    - Yes, it is done, you can find the results in this documents or you can run the Python script
- Can you think of any other analysis which would be helpful with this data?
    - As an alternative to the Marketing Mix Model (MMM) is the Multi-Touch Attribution (MTA) or Digital Attribution. These models can be used these days however there not enough data in this dataset to implement these models.
    - MMM typically looks at historical sales data and tries to identify the relationship between various marketing channels (such as TV, radio, print, online ads, etc.) and sales. MMM uses statistical techniques to model the impact of each channel on sales, taking into account factors such as seasonality, pricing, promotions, and other variables that may affect sales.
    - MTA, on the other hand, tracks customer journeys across multiple touchpoints (such as email, social media, search, display ads, etc.) and assigns credit to each touchpoint based on its contribution to the conversion. MTA is typically more granular and can provide insights into how each touchpoint contributes to the overall customer journey, as well as how different combinations of touchpoints affect conversion rates.
    - Digital Attribution: This model focuses specifically on measuring the impact of digital channels (such as social media, search, and display ads) on sales or other business outcomes.
- How do the different variables affect sales?
    - It is difficult to say with this data however, it is possible to understand adding TikTok, using new website, time periods like holidays or before holidays affects the sale
- Which are the most profitable channels?
    - Facebook, Google Ads
- If budgets are doubled in each channel, would the same channels still be the most profitable ones?
    - It's hard to say for sure without analyzing the data further, but doubling the budgets in each channel could potentially lead to a shift in the most profitable channels. It's possible that the channels that were previously less profitable, such as Amazon, could become more profitable with a larger budget. On the other hand, the channels that

    were previously more profitable, such as Facebook , Google Ads and TikTok, may not see as large of an increase in profitability with a doubled budget. It would be worth conducting further analysis and running some simulations to determine the most effective allocation of the increased budget.

    - To do this, you would need to update the values for the investment in each channel in the data set and then apply the prediction model to the updated data. This would give you a new set of predicted sales values for each channel. You could then compare the predicted sales values with the original sales values to see how much of an increase in sales you would expect to see if the investment is doubled in each channel. It's important to note that prediction models like linear regression are not perfect, and there may be other factors that affect sales that are not captured by the model.

- If the client gives us a total budget for Q1 2023 of +20% with respect to Q4 2022, how would you suggest using this budget?
    - if we consider the budget of Q4 2022 equal to 100,000 euros, a 20% increase in budget means 20000 euros more, from this 20000euros more, my suggestion is to invest 30% on TikTok, 20% for each one of Facebook and Google Ads, 15% on Amazon, and 10% invest in new marketing initiatives such as influencer marketing, email marketing, or content marketing.