

**INTERNSHIP PROJECT 2023**

CİHAT KUCUKCAPRAZ

DİLEK KURT

EKİN BİÇEN

KEREM İMAL

LEVENT BAHADIR İBİCİOĞLU

MUHAMMED MEMON

MUSTAFA B

NACİYE MÜLAYİM

YUSUF KENAN BAYRAK

The Rugs UK is a company that specializes in providing high-quality and unique rugs for every room in homes. This project aims to examine the data provided by the company, which consists of five data files:

1. Satışlar File,
2. İadeler File,
3. Ürün Yorumları File,
4. Amazon kampanyaları File,
5. Diğer kampanyalar File.

The goals of the project are to conduct various analyses, including Exploratory Data Analysis, Income Analysis, Product Analysis, Advertisement and Social Media Analysis, and Customer Analysis. The aim is to draw valuable insights and conclusions from these analyses. To achieve these goals, the project will utilize various tools and techniques, including Python, Excel, SQL, Tableau, PowerBI, and machine and deep learning algorithms.

Further analysis will be made to draw valuable insights and conclusions.

**First and Second Week Explanation of the Actions**

**On 12 December 2023,** it was explained internship project information and given some directions about the program schedule by the external stakeholders.

**On 14 December 2023,** raw data was introduced and it was answered questions asked by the data analytics team.

**On 15 December 2023,** the raw data was delivered to the internship team and it was determined.

It is given the general project steps in below it is planned by the authorities:

| Raw data definitions and first overview | General meeting and updates |
| --- | --- |
| Raw data and explanation |
| Raw data hand over |
| Explaratory data analysis together with PostgreSQL | Exporting the data to SQL |
| Feature Engineering |
| PostgreSQL table works |
| Python Explaratory Data Analysis overview |
| Deciding on the features and analysis for the project |
| Determining the External Featurelerin and importation |
| Feature Importance |
| Income Analysis | Time Series Analysis on the basis of monthly gross income |
| Determining the features that will be used and the coefficients and power of the model that will be made |
| Detecting the deviations on the monthly gross income rates |
| Causality analysis  The effects of war, Inflation of the export country,exchange rates on sales |
| In accordance with the times series analysis, forecasting the sales and maximum profit that will be made in the next 6 months. |
| A/B testing on the consistency of monthly average sales that is made on different sales platforms |
| The similarity of sales amounts in different sales platforms (Effect Size analysis) |
| The changes of sales made in different cities across UK |
| The trends on monthly sales (Line Graphic) |
| Analysis of the relationship between advertisment spendings and income |
| Data visualization on Tableau and Power BI platforms |
| Product Analysis | The seasonal sales of different products (Season analysis) |
| Analysis of the sales of each product on different platforms |
| Basket Analysis |
| Return products and return reasons analysis |
| Carpet pattern recognition analysis with CV (Computer Recognition) |
| Data visualization through Tableau and Power BI |
| Advertisement and social media analysis | Advertisement analysis (A-B Testing) |
| Google Analytics analysis (Customer age and city) |
| CTR analysis (click rate) |
| Customer Analysis | Customer Loyalty Analysis (Churn Prediction) |
| The relationship between the loyal customers and what city they live in |
| What customers are more loyal according to what they buy. |
| Customer re-purchase duration |
| Loyal customer list |
| The common characteristics of loyal customers |
| What sales platforms have more loyal customers |
| The effects of features on customer loyalty |
| Customer segmentation analysis (Kohort) |
| LTV (Lifetimevalue) analysis |
| Customer location analysis |
| Sentiment analysis with NLP |
| Data visualization through Tableau and Power BI |
|  | General Deployment |

**Table1:** Internship project schedule

**On 26 and 27 December 2023**, PostgreSQL Program was installed by team members and the data was transferred to this program and five tables occurred.

**On 28 December 2023**,The team members began the process of evaluating the tables. They started by examining the "Satislar (Purchase)" table as it contained the most data. Over the next two days, the rest of the tables were evaluated. During this process, several discrepancies and problems were identified, including missing and incorrect explanations, as well as problematic data and missing explanations. To address these issues, the team formed a set of questions to ask the authorities and also worked to identify the column names and define the relationships between all of the data.

In below, there are column names of the tables given for analysis:

| **Urun Yorumlari** | **Iadeler** | **Satislar** | **Amazon Kampanyalari** | **Diger Kampanyalar** |
| --- | --- | --- | --- | --- |
| order\_id | Type | nOrderId | State | ReportingStarts |
| review\_title | nOrderId | Company | Campaigns | ReportingEnds |
| comments | cPostCode | CustomerID | Status | CampaingName |
| rating | CustomerID | cPostCode | Type | CampaingDelivery |
| status | ItemNumber | dReceivedDate | Targeting | AdSetBudget |
| date\_created | ItemTitle | Country | Campaignbiddingstrategy | AttributionSetting |
| sku | dReceievedDate | Status | Startdate | Results |
| Customer\_ID | cCountry | Proceed | Enddate | ResultIndicator |
| Address | cCountryCode | ProcessedDate | Portfolio | Reach |
| product\_sku | cCurrency | Source | Budget(GBP) | Impressions |
| product\_name | Source | Currency | Top-of-search IS | Cost\_per\_results |
| product\_link | Subsource | Subtotal | Costtype | AmountSpent |
| video\_review\_prompt\_id | ReturnDate | Tax | Impressions | Ends |
| Tags | ReturnQty | Total | Clicks | Frequency |
| Reply | Category | OrderItemSKU | CTR | Unique\_link\_clicks |
| reply\_private | ResendOrExchangeQty | OrderItemTitle | Spend | Landing\_page\_views |
| reply\_date | RMAActioned | ItemCategory | CPC | Link\_clicks |
| published\_images | RefundAmount | OrderItemQuantity | Orders | Cost\_per\_landing\_page |
| unpublished\_images | ReturnReason | DispatchStockUnit | Sales | Adds\_to\_cart |
| published\_videos |  | TotalWeight | ACOS | Website\_adds\_to\_cart |
| unpublished\_videos |  | Purchase | ROAS | Meta\_add\_to\_cart |
| Source |  | TrackingNumber | NTBorders | Checkouts\_initiated |
| Location |  | PostalService | %ofordersNTB | Meta\_checkouts\_initiated |
| Timeago |  |  | NTBsales | Purchases |
| video\_first\_campaign |  |  | Viewableimpressions | Meta\_purchases |
|  |  |  | VCPM | Purchases Conversion |
|  |  |  |  | Value |
|  |  |  |  | Website purchases |
|  |  |  |  | conversion value |
|  |  |  |  | Meta purchase conversion value |
|  |  |  |  | Purchase ROAS (return on ad spend) |
|  |  |  |  | Website purchase ROAS (return on advertising spend) |

**Table 2:** illustration of file names and columns

**Excel Files Data Overview Before Analysis**

It was prepared title explanations for each table's column names and it is noted for each column evaluations. Incomprehensible data was identified and it is detected data that needs to be deleted or corrected.

In below, each tables title explanation and table explanation was given:

**File name: Satislar**

**i) Satislar Titles Explanation**

**nOrderId:** Order Number

**Company:** It is the column that denotes the name of the company. However, the system name did not reflect the actual company name. Instead, the Amazon platform reflects the name "Redacted." The purpose of this is to close the data after a certain period of time for data security reasons.

**Customer ID:** Customer identification number

**cPostCode:** Customer's Address Zip Code.

**dReceievedDate:** The date and time the order was placed

**Country:** The country where the customer is located

**status:** Displays payment status

**Processed:** Indicates the completion status of the order-sending process to the customer.

**ProcessedDate:** Indicates the date the order was processed.

**Source:** Shows the platform where the sale is made

**Currency:** Shows the currency in which the sale was made

**Subtotal:** It shows the amount of the product, excluding tax, if the product was purchased. For example, let's say that the price of an item is 5 Pounds, if one is purchased it will appear as 5, if 2 units are purchased it will appear as 10.

**Tax:** Shows the tax amount if the product has been taxed

**Total:** Subtotal+Tax

**OrderItemSKU:** Unique product code

**OrderItemTitle:** Product name

**ItemCategory:** Product category name

**OrderItemQuantity:** Order Item Quantity

**DispatchStockUnitCost:** Product purchase price

**TotalWeight:** Product weight

**PurchasePrice:** Product unit cost

**TrackingNumber:** Shipment Number

**PostalService:** Postal service to which the shipment was made

**ii) Satıslar Table Evaluations**

The file; purchase (Satislar) was evaluated and some explanations are added to column names. It fixed some problems and questions about some columns data values. All informations and explanations are in below:

* There are duplicate values in A Column in the file. Solve duplicate problems in A column.
* Remove Column B related to the customer data (column has to be deleted due to customer privacy). The column can be deleted in Python.
* Customer ID column (Column C) has different data under unique customer ID. The issue has to be confirmed by asking to data management team. The Column labeled as **Important**.
* cPostCode column (Column D) is clear. There are columns named as redacted. Column D can be cleaned through Python for further analysis.
* dReceivedDate column (Column E) is clean. No change required. Date formats will be changed in Python.
* Country column (Column F) is clean. No change required.
* Status column(Column G) is clean.No changes required.
* Processed column (Column H) can be deleted. There is only one unique value on all rows.
* Processed\_Date column (Column I) has 3465 null values. it is required filling the significant values. It has issues with the date format. The format needs to be aligned with the same format through Python.
* Source column (Column J) is clean. No changes required.
* Currency column (Column K) is clean. No changes required.
* Subtotal column (Column L) has different currencies in the data. Column L has to be converted to GBP currency (The currency will be converted by referencing the central bank). Column labeled as **Important**.
* Tax column (Column M) is clean. No change required.
* Total column (Column N) has to be changed to GBP currency (by referencing the Central Bank website). Column labeled as **Important**.
* OrderItemSKU (Column O) is clean. **Note:** The SKU codes might show differences according to the platform that the sale has been made. It is thought that dividing the columns by the platforms that the sale made is offered during the meeting.
* OrderItemTitle (Column P) is clean. No change required.
* ItemCategorycolumn (Column Q) is clean. No change required. Column labeled as **Important**.
* OrderItemQuantity (Column R) is clean . No changes required.
* DispatchStockUnitCost column (Column S) is unknown. **Question**: The column DispatchStockUnitCost is vague. Is the column S the value of stock cost or the unit cost for sale?
* TotalWeight column (Column T) is vague. The column might be deleted since it has 0 kg/lbs values.
* PurchasePrice column (Column U) is vague. **Question**: There are columns that have 0 purchasePrice values. How accurate is the data?
* TrackingNumber (Column V) can be deleted. No possibility for further analysis. **Question**: How are the tracking numbers assigned? Default values have different meanings for each postal service?
* PostalService column (Column W) has default values. Data can be analyzed after cleaning through Python. **Question**: What are the meanings of **Default** rows?

During the internship meeting , it was occurred some questions

**iii) Questions to be asked**

1. Customer ID column (Column C) has different data under unique customer ID. The issue has to be confirmed by asking to data management team. **Question**: Is the customer ID accurate?
2. DispatchStockUnitCost column (Column S) is unknown. **Question**: The column DispatchStockUnitCost is vague. Is the column S the value of stock cost or the unit cost for sale?
3. PurchasePrice column (Column U) is vague. **Question**: There are columns that have 0 purchasePrice values. How accurate is the data?
4. TrackingNumber (Column V) can be deleted. No possibility for further analysis. **Question**: How are the tracking numbers assigned? Default values have different meanings for each postal service? (-)
5. PostalService column (Column W) has default values. Data can be analyzed after cleaning through Python. **Question**: What are the meanings of **Default** rows?

**2)File name: Iadeler**

It contains 7941 raws

**i)Iadeler title explanation**

**Type:** Reason for return

**nOrderId:** Number given by the system

**cPostCode**: Postal code of the returning customer

**Customer ID:**

**ItemNumber :** Product code

**ItemTitle:** Product name

**dReceievedDate:** Request date and time

**cCountry:** Customer country

**cCountryCode:** Customer country code

**cCurrency:** Currency unit

**source:** Return Request sales platform

**subsource:** Subsource sales platform

**Return Date:** Return date

**ReturnQty:**  Amount of returned product

**Category:** Category of returned product

**ResendOrExchangeQty:**  Amount of resent or exchange

**RMA Actioned:** Returns Status

**Refund Amount:** Mount if a refund is made

**Return Reason:** Return explanations

**ii) Iadeler Table Evaluations**

* **Type** (Column A) has 7941 null values.It contains five subtitles
* **nOrderId** column (B column) is clean. It is offered to make the B column the **primary key.**
* **cPostCode** column (C column) is clean. No change is required. Further analysis can be made through Python.
* **CustomerID** column (D column) is the same as the **satislar file:** Customer ID column. The column is labeled as **Important**.
* **ItemNumber** (E column). **Question**: Can the same column be added from another column to Iadeler file? The data is cleaner in other Excel files.
* **ItemTitle** column (F column). **Question**: Can the same column be added from another column to Iadeler file? The data is cleaner in other Excel files.
* **dReceievedDate** (G column) has 7941 null rows
* **cCountry** (H column) is clean. No change is required.
* **cCountryCode** (I column) is clean. No change is required. Further analysis can be made through Python.
* **cCurrency** (J column) is clean. No change is required.
* **Source** column (K column) is clean. No change is required. More analysis will be made.
* **Subsource** column (L column) is clean. Either the **source column** or the **subsource column** can be dropped due to the similarity of data rows.
* **ReturnDate** column (M column) is mostly empty. The column can be deleted.
* **ReturnQty** column (N column). **Question**: Out of 7941 rows, there are 5321 zero values in the column. What are the meanings of zero values?
* **Category** column (O column) has 5000 empty rows. There are 105 empty rows. The empty rows are considered and planned to be filled by project members in the meeting. The filling process will be based on ReturnReason (S column).
* **ResendOrExchaneQty** column (P column) is mostly empty. The column will be deleted for the accuracy of the analysis.
* **RMMActioned** column (Q column) can be deleted. Most of the rows are empty
* RefundAmount (R column) has 930 empty rows. Further analysis can be done after cleaning.
* Return Reason column ( S Column ) contains 5462 rows.

**3) File name: Urun Yorumlari**

## **i)Urun Yorumları Title Explanation**

**order\_id :** The number given by the system

**review\_title: ?** (absent any data)

**comments:** Comments

**rating:** rating for the product

**status:** system position

**date\_created:** The date the comment was created

**sku:** Product code

**Customer ID:** Customer Identification number

**address:** Customer Address

**product\_sku:**Product code

**product\_name:** Product name

**product\_link:** Product link

**video\_review\_prompt\_id: ?**

**tags:** ?

**reply:** absent any data

**reply\_private:** absent any data

**reply\_date:** absent any data

**published\_images:?**

**unpublished\_images:?**

**published\_videos:?**

**unpublished\_videos:?**

**source:** The platform where the comment was made

**location:** Country where the comment was made

**timeago:?**

**video\_first\_campaign:?**

**ii) Urun Yorumları Table Evaluations**

This file contains 562 (411) rows and 25(24) columns.

* **Order ID** has 46 null values and the column (A column) and the satislar excel file's comments are not the same. It contains same values
* **Review\_title** column (B Column)has not any data
* **Comments** column (C Column**)** has 105 null values
* **Rating** column **(**D Column**)** has labeled values from 1 to 5 and there is no absent rows
* **Status** column **(**E Column**)** has labeled active and inactive there is no absent rows.
* **date\_created (**F column**)** has not null values
* **Sku column** (G column) and product\_sku column (J column) have the same values. One of the columns can be dropped. There is a format problem in the columns. The issue needs to be fixed before analysis on Excel.
* **Customer ID (H column):** The customers who have C0000001 values mean redacted in the **SatişlarExcel file** and **İadeler Excel files**. When analyzing, the redacted values (in this case, C0000001) can be dropped and continued with linking the tables to the **Urun Yorumlari Excel file**.
* **address** column (I Column)has not any data.
* **product\_sku column (J column)** has same values Sku column it can be deleted
* **Product\_name** column (K column) needs to be split into size, color and types of rugs after cleaning on Excel. It has three null rows.
* **product\_link** column(L column) contains only “199” values and some rows are null(369)
* **video\_review\_prompt\_id** column ( M column)
* **tags** column( N column)
* **reply** column ( O column) absent any data
* **reply\_private** column( P column) absent any data
* **reply\_date** column( R column) absent any data
* **published\_images** column( S column) almost all rows are null and it was decided to be deleted
* **unpublished\_images** column( T column) almost all rows are null and it was decided to be deleted
* **published\_videos** column( U column) almost all rows are null and it was decided to be deleted
* **unpublished\_videos** column( V column) almost all rows are null and it was decided to be deleted
* **source** column( W column) The platform where the comment was made. There are 89 null values
* **location** column ( X column) Country where the comment was made.There are 206 null values.
* **timeago** column( Y column) has no any value
* **video\_first\_campaign** column ( Z column) has no any value

**iii) Questions to be asked**

* status column (E column) is vague. **Question**: What is the meaning of active and inactive data rows?
* Order ID column (A column) and the satislar excel file's comments are not the same. **Question**: How are the tables can be linked? Should we even connect the tables?

**4) File name: Amazon Kampanyalari**

This file contains 127 rows and 27 columns.

**i) Amazon Kampanyalari Title Explanation**

**State:** Campaign position

**Campaigns:** Campaign Name

**Status:** Campaign position

**Type:** N/A

**Targeting:** N/A(Amazon campaign preferences: manual-automatic)

**Campaign bidding strategy:** Campaign Strategy

**Start date:** Campaign start date

**End date:** Campaign ending date

**Portfolio:** N/A

**Budget(GBP):** Campaign Budget

**Top-of-search IS:** Top-of-search impression share

**Cost type:** CPS :Clicks Per Searc**h**

**Impressions:** Impressions measure the number of times Amazon shows shoppers your Ad, regardless of whether they clicked on it or not. It is one of the key metrics of Amazon Advertising and can be used to check at a glance how well an Ad is doing.

**Clicks:** The number of clicks visitors have made on your Associates links.

**CTR:** Click Through Rate CTR= Clicks/Impression

**Spend(GBP):**

**CPC(GBP):**Cost-per-click, or CPC, is a common Amazon PPC advertising term that refers to the price you pay each time your ad receives a click. This metric is relevant for your Amazon PPC strategy. CPC Amazon is an important metric to track. If it’s too high, achieving a return on your investment for your advertising campaign can be difficult.

**Orders:** Product orders values

**Sales(GBP):** Money from sales

**ACOS:** Advertisement Cost of Sales.Amazon advertising cost of sales (ACOS) is a metric used to measure Amazon pay-per-click (PPC) advertising campaigns. It compares the amount spent on PPC campaigns to the amount earned, and it helps determine if your brand generated campaigns that were cost-efficient. Amazon ACOS helps measure the performance of Sponsored Products ads on Amazon

**ROAS:** Return on Advertising Spend. It's a metric that measures how much revenue you've made in sales for each dollar you've spent on ads

**NTB orders:** New-to-brand (NTB) orders: The number of first-time orders for products on Amazon within the brand over a one-year lookback window

**% of orders NTB: percent of Orders:**

**NTB sales(GBP):**New-to-brand (NTB) sales: The total sales (in local currency) of new-to-brand orders. For Sponsored Brands only.

**% of sales NTB:**

**Viewable impressions:** A viewable impression is a standard measure of ad viewability defined by the International Advertising Bureau (IAB) to be an ad, which appears at least 50% on screen for more than one second. Viewable Impressions are the metric that sellers can use to quantify the percentage of ads that are actually viewed by real people.

**VCPM(GBP):** Viewable CPM—stands for cost per thousand viewable impressions. It's a metric used to determine how many people actually see ads on a web page, instead of simply how many users see the website.

**ii) Amazon Kampanyaları Table Evaluations**

* **The State** column (A column) and Status column (C column) have similar values. This **needs to be considered** before starting the analysis.
* **Campaigns** column (B column) contains campaign names. it can be given an ID instead of campaign names
* The type column (D column) is vague. The meanings of SB, SP, and SBV are unknown. **Question**: The type column (D column) in the Amazon Kampanyalari Excel file is not understood. What are the meanings of SB, SP, and SBV?
* The targeting column (E column) has no column explanation. **Question**: What does the Targeting column (E column) mean?
* The campaign bidding strategy column (F column) is vague. More research is needed before the analysis.
* The start date column (G column) is clean and has the data, but the End date column (H column) is totally empty. It is not advised for further analysis.
* The portfolio column (I column) is mostly empty and vague. The column can be deleted for further analysis.
* The budget(GBP) column (J column) is clean. It is advised in the team meeting to compare profit/ loss rates in the campaign.
* Top-of-search IS column (K column)
* Cost type column (L column)
* Impressions Column (M column)
* Clicks Column (N column)
* Explanation of CTR column (O column) in detail:
  + Click/impression=CTR

A ratio showing how often people who see your ad or free product listing end up clicking it. Clickthrough rate (CTR) can be used to gauge how well your keywords and ads, and free listings, are performing.

CTR is the number of clicks that your ad receives divided by the number of times your ad is shown: clicks ÷ impressions = CTR. For example, if you had 5 clicks and 100 impressions, then your CTR would be 5%.

* Spend(GBP) column (P column)
* CPC(GBP) column (Q column)
* ROAS column (U column) has a formula behind it.
  + How to calculate ROAS

This is how you can calculate the ROAS:

ROAS = (profit/advertising costs) \* 100

**iii) Questions to be asked**

* + - The type column (D column) is vague. The meanings of SB, SP, and SBV are unknown. **Question**: The type column (D column) in the Amazon Kampanyalari Excel file is not understood. What are the meanings of SB, SP, and SBV?
    - The targeting column (E column) has no column explanation. **Question**: What does the Targeting column (E column) mean?

**5) File name: Diğer kampanyalar**

## **i)Diger Kampanyalar Title Explanation**

* **Reporting starts:** Campaign Starting
* **Reporting ends:** Campaign ending
* **Campaign name:** Campaign name
* **Campaign delivery:** Campaign status
* **Ad set budget:** Campaign budget
* **Ad set budget type:** Campaign budget type: daily, weekly etc.
* **Attribution setting:** Association setting
* **Results:** The results metric shows you how well your ad campaign is performing, based on your chosen business goals and settings. Using this metric you can compare the performance of similar campaigns and identify areas of opportunity for better results.
* **Result indicator:** Results indicators can be quantitative, qualitative, or more commonly, a mixture of both. Moreover, results indicators show how much or how well outcome-based objectives are being or have been achieved; and verify thus the accomplishment of the project's underlying overall goal.
* **Reach:** Reach is the total number of people who see your content
* **Impressions:** Impressions are the number of times your content is displayed, no matter if it was clicked or not.
* **Cost per results:** Cost per impression
* **Amount spent (GBP):** Total budget spent
* **Ends:** Campaign ending date
* **Frequency:** Frequency Capping or Impression Frequency basically refers to how often an ad will be shown to an internet user.
* **Unique link clicks:** The number of people who performed a link click.This is an estimated measurement.This measurement counts people, not actions. This metric uses sampled data.
* **Landing page views:** Landing page views let you know how many times people loaded your website, after clicking your ad. You can compare landing page views to link clicks to understand how many people clicked on your ad but left before your website loaded. To report on landing page views, you must have created a Facebook pixel."
* The most important takeaway, is that of all the metrics above, Landing Page Views is the only one that counts when a user loads a webpage after clicking an ad, since the pixel fires any of these events: PageView, PixelInitialized or ViewContent (PageView and PixelInitialized are automatically captured by the Facebook Pixel).
* **Link clicks:** This is the total number of clicks on your ads. Seems simple by definition, however, the main callout is that it also counts any actions on the ad itself, including:

1. Link clicks
2. Clicks to the associated business Page profile or profile picture
3. Post reactions (such as likes or loves)
4. Comments or shares
5. Clicks to expand media (such as photos) to full screen
6. Clicks to take actions identified as your campaign objective (such as liking your Page for a Page engagement campaign)

* **Cost per landing page view (GBP):** Average cost for each landing page view. This is a metric in development.
* Calculation Format:This metric is calculated as total amount spent divided by landing page views. Related Measurements

1. Landing Page Views
2. Link Click

* **Adds to cart:** The number of add-to-cart events attributed to your ads, based on information from one or more of your Linked Meta Business Tools. In some cases, this metric is estimated and may have been reported by a third-party API.

Calculation Format

This metric calculates add-to-cart events tracked by one of your Meta Business Tools (such as Meta pixel, Conversions API, app SDK, or offline event set) and attributed to your ads. Events can be counted differently depending on the attribution setting selected. In some cases where events cannot be counted directly due to partial or missing data, statistical modeling can be used to account for some events.

In some cases, such as reported results for iOS 14 campaigns, this metric may not be calculated.

Related Measurements

Fee per Add to Cart

* **Website adds to cart:** The number of add to cart events tracked by the pixel or Conversions API on your website and attributed to your ads.
* **Meta add to cart:** Same Add to chart column
* **Checkouts initiated:** The number of payment initiation events attributed to your ads, based on information from one or more of your Linked Meta Business Tools.

In some cases, this metric is estimated and may have been reported by a third-party API.

**Calculation Format**

This metric calculates purchase launch events tracked by one of your Meta Business Tools (such as Meta Pixel, Conversions API, app SDK, or offline event set) and attributed to your ads. Events can be counted differently depending on the attribution setting selected. In some cases where events cannot be counted directly due to partial or missing data, statistical modeling can be used to account for some events.

In some cases, such as reported results for iOS 14 campaigns, this metric may not be calculated.

Related Measurements

Fee per Initiated Purchase

* **Website checkouts initiated:**
* **Meta checkouts initiated:** Number of purchase launch events tracked by the pixel or Conversions API on your website and attributed to your ads
* **Purchases:**
* **Website purchases:**
* **Meta purchases:** The number of purchases made on meta technologies (e.g. Pages or Messenger) and attributed to your ads.
* **Purchases Conversion Value:**
* **Website purchases conversion value**
* **Meta purchase conversion value:** The total value of purchases.
* **Purchase ROAS (return on ad spend)**
* **Website purchase ROAS (return on advertising spend)**

**ii) Diğer Kampanyalar Table Evaluation**

* The campaign delivery column (D column) is vague. The results column (H column) is vague.
* All the columns that include **META word** will be deleted. There is no data available for these columns.

**Exploratory Data Analysis**

**Data Import and Cleaning**

After carefully evaluating the available data sources, the data team has determined that two files are the most suitable for further analysis using Excel. These files are "Diger Kampanyalar.xlsx" and "Amazon Kampanyalari.xlsx." The team plans to utilize the Excel Data Query Tool for the cleaning process.

The remaining three files, "Satislar.xlsx," "Iadeler listesi.xlsx," and "Urun yorumlari.xlsx," will be analyzed using Python due to their size and functional requirements. These files are too large or complex to be effectively analyzed in Excel, and Python offers the necessary tools and capabilities to handle the data.

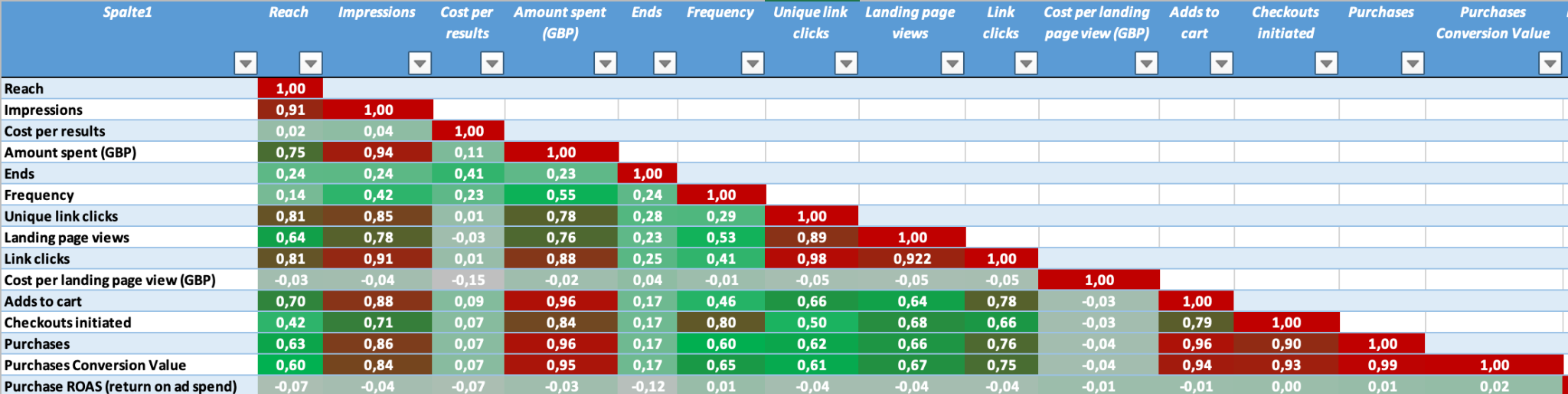
Overall, the data team has carefully considered the strengths and limitations of each tool and chosen the most appropriate ones for the task at hand. By utilizing both Excel and Python, they will be able to effectively and efficiently analyze the data and derive valuable insights.

Firstly, Amazon Kampanyaları was chosen to clean operation after evaluation all column explanations

* **Diger Kampanyalar.xlsx**

Diger Kampanyalar.xlsx file has a data frame of 152 rows and 30 columns which was considered suitable for further analysis on Excel. During the cleaning process, several actions were taken.

* Adds to cart column (**T column**) had a minor amount of missing values which was replaced with 0 as an integer for the data evaluation.
* Unique link clicks column (**P column**) had 19 blank rows. Therefore, blank rows were filled after taking the average of the P column which resulted in no change on the average of unique clicks.
* Link clicks column (**R column**) had the same issue with having 20 blank rows. The method deployed to fill the link clicks was by dividing the impression column to unique clicks (formula = impression / unique link clicks). It is important to mention that the use of this method was deployed due to the correlation ratios of inter-related columns. The correlation table is provided with the table below.



**Table 2:** Correlation ratios of each column (red: high corr, green: low corr)

* The correlation levels were sustained after filling blank rows by the formula applied. Additionally, Amount Spent (GBP) (**Column M**) and Ends (**Column N**) columns had empty rows that was filled based on their correlations.

Ad set budget column (E column) is labeled as belows:

* Using ad set budget=0
* daily = 1
* Lifetime = 2

Attribution setting column(G column) is labeled as belows:

* 7-day click or 1-day view=0
* 1-day click=1
* 7-day click=2
* 28-day click or 1-day view=3
* Multiple attribution settings = 4

In CPR\_LandingPageView column was generated by using formula as below:

CPR=cost per result =amount spend / purchases

formulas.

In CPR\_LINKCLICKS column was generated by using formula as below:

CPR=cost per result=Amount spend / Link clicks

In CPR\_ADDTOCART column was generated by using formula as below:

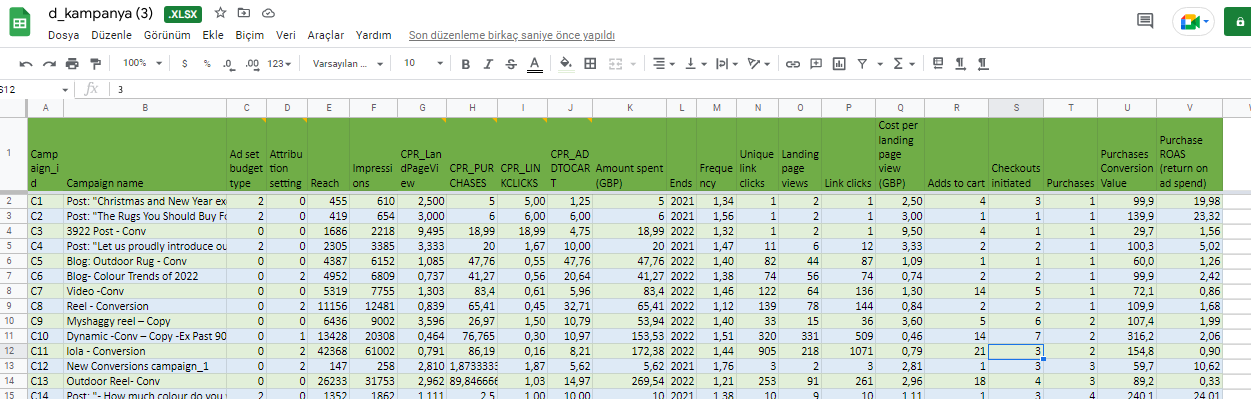
CPR=cost per result =amount spend / add to cart

Thus, it is generated new columns CPR\_LandPageView, CPR\_PURCHASES, CPR\_LINKCLICKS, CPR\_ADDTOCART

The table Diger Kampanyalar.xlsx also had empty columns that had to be deleted for integrity and enhanced analysis. These columns were generally META columns which was associated with the Facebook ad campaigns.

At the end of the data cleaning process for Diger Kampanyalar.xlsx file, rows decreased from 152 to 146 and columns decreased from 30 to 22. The remaining data frame is 22x146.

1. Campaign\_id
2. Campaign name
3. Ad set budget
4. type
5. Attribution setting
6. Reach Impressions
7. CPR\_LandPageView
8. CPR\_PURCHASES
9. CPR\_LINKCLICKS
10. CPR\_ADDTOCART
11. Amount spent (GBP)
12. Ends
13. Frequency
14. Unique link clicks
15. Landing page views
16. Link clicks
17. Cost per landing page view (GBP)
18. Adds to cart
19. Checkouts initiated
20. Purchases
21. Purchases Conversion Value
22. Purchase ROAS (return on ad spend)

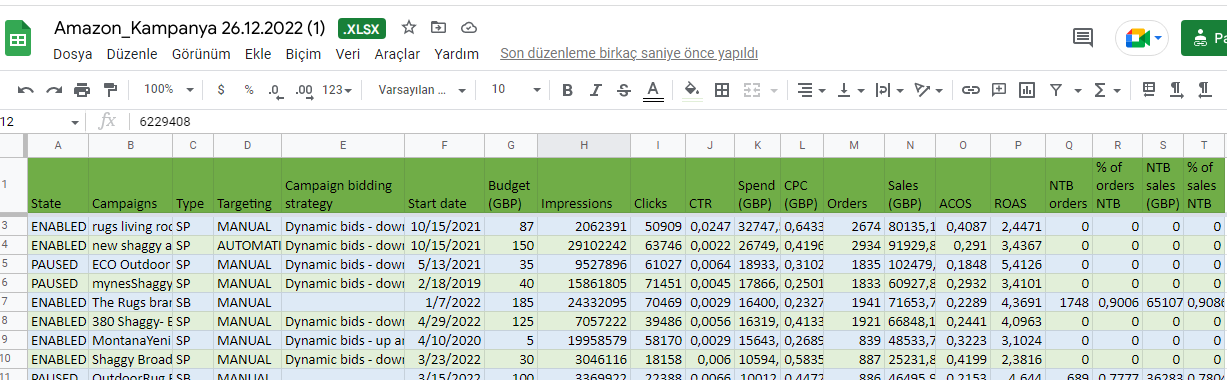
 **Table 3**. Diğer kampanyalar excel file .

**Amazon Kampanyalari.xlsx**

This file has 27 columns and 128 rows. All columns are evaluated separately after that it is decided to drop some columns because it is indicated explanation as in below:

* The "State" and "Status" columns contain the same values that is why status column was deleted.
* The "End date" and "Portfolio" columns were dropped because they had a large number of empty rows according to the raw data.
* The "Top-of-search IS" and "Cost type" columns were also deleted due to a high number of null values.
* The "Viewable impressions" and "VCPM(GBP)" columns were also deleted for the same reasons.
* There are 4 rows containing 0 values, so because it is thought that it will not fill meaningful values, they are dropped. Thus it has 124 rows now.
* It was thought that it is not necessary to generate new features.If it is necessary in next steps in the analysis parts, it will be planned.

New columns is below after cleaning the data:

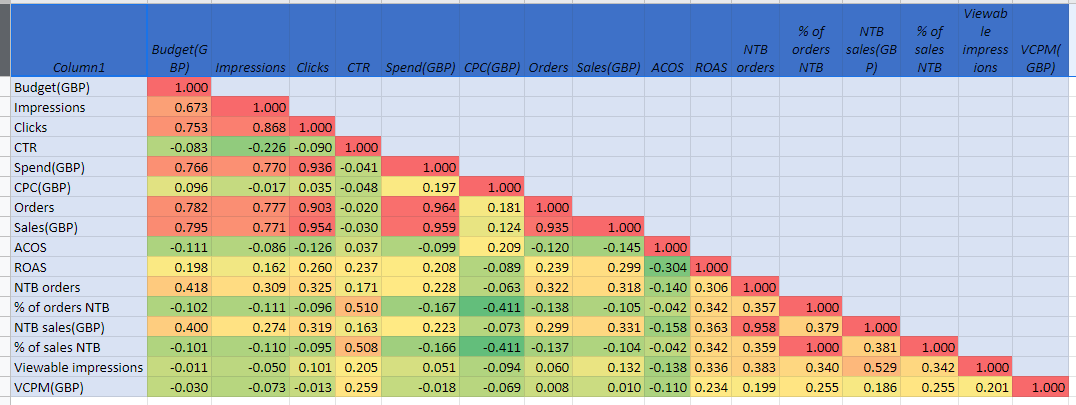


**Table 4.** “Amazon kampanyaları” excel file after cleaning

After cleaning operations, obtained column names of “Amazon Kampanyaları" is below :

1. State
2. Campaigns
3. Type
4. Targeting
5. Campaign bidding strategy
6. Start date
7. Budget(GBP)
8. Impressions
9. Clicks
10. CTR
11. Spend(GBP)
12. CPC (GBP)
13. Orders
14. Sales (GBP)
15. ACOS
16. ROAS
17. NTB orders
18. % of orders NTB
19. NTB sales (GBP)
20. % of sales NTB

Amazon kampanyaları correlation table is given below:



**Table 5:** Correlation ratios of each column (red: high corr, green: low corr)

**Satislar.ipynb**

The sales file is the file with the most data and it is expected that the most insights for the company's operations will be obtained from this file. Therefore, before performing data analysis, all columns have been thoroughly examined and attempts have been made to become aware of empty rows that need to be filled in.

There are column names of the data as in below :

'nOrderId', 'Company', 'Customer ID', 'cPostCode', 'dReceievedDate' , 'Country', 'status', 'Processed', 'ProcessedDate', 'Source', 'Currency','Subtotal', 'Tax', 'Total', 'OrderItemSKU', 'OrderItemTitle', 'ItemCategory', 'OrderItemQuantity', 'DispatchStockUnitCost', 'TotalWeight', 'PurchasePrice', 'TrackingNumber', 'PostalService'

Some of the title of the columns are changed as follows:

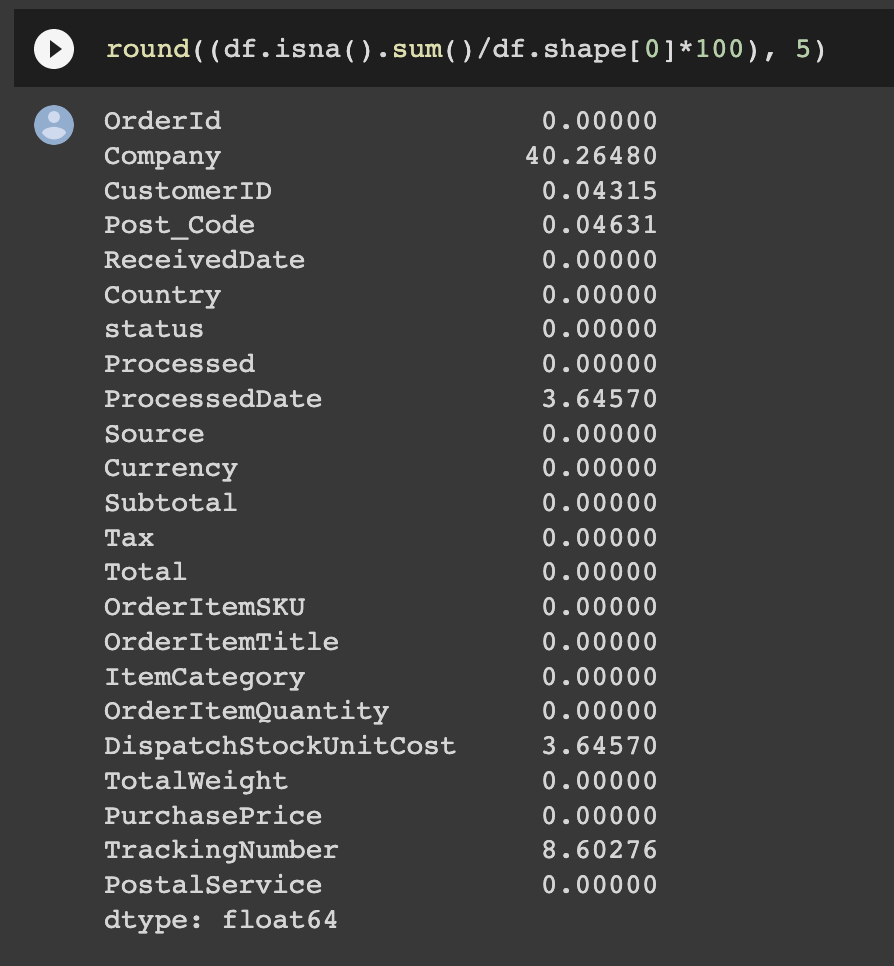
"nOrderId":"OrderId",

"Customer ID":"CustomerID",

"cPostCode":"Post\_Code",

"dReceievedDate":"ReceivedDate",

Shape of the raw data is (190031, 23) after cleaning the rows contain null values based on the OrderId column and after 101 pieces duplicate values is cleaned, the shape of the data is (94915, 23)



**Table 6:** code line for rounding the null columns

Through the code shown above, the common missing values of the data are detected. The company, Processed date and DispatchStockUnitCost columns are the ones that show a high number of missing values in the data frame.

ReceivedDate and ProcessDate columns have incompatible values with each other. For instance, there are instances where the ReceivedDate is a later date than the ProcessDate. However, this should not be possible, as the ProcessDate should always be later than the ReceivedDate. To fix this issue, coding steps were applied and the problem was successfully resolved.

We updated the values in the Total column according to the currencies listed in the Currency column and created a new column called 'Adjust\_total'.

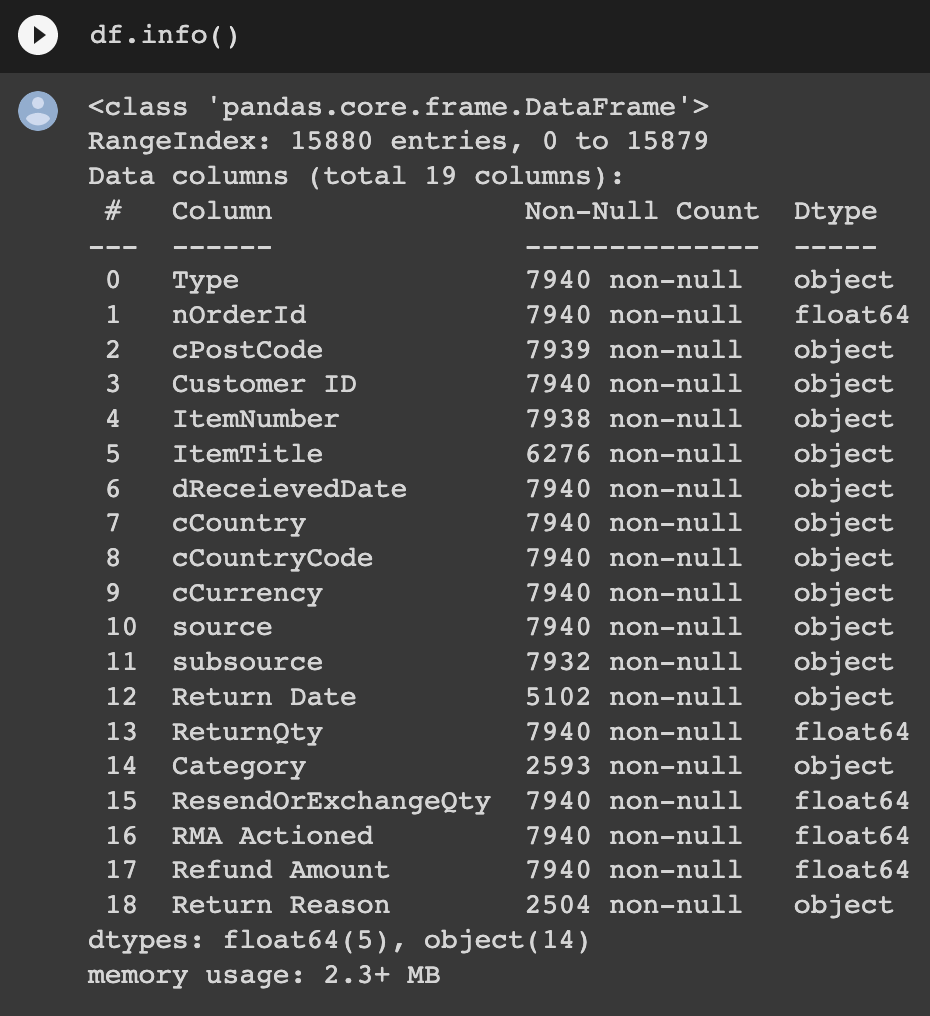
"Company", "TrackingNumber", "Processed", "TotalWeight", "DispatchStockUnitCost" Columns are dropped.

In Satışlar file, there are more sales based on amazon source and some information about the customer is not given by amazon source.for this reason, it is thought that the file is more suitable to separate two files as amazon\_satışlar and other\_satışalar files.

Explaratory Data Analysis on Satislar File end date: **January, 21, 2023**

**Iadeler EDA**

Python is considered for the exploratory analysis on Iadeler file with Pandas and Numpy libraries. The Iadeler file has a shape of 15580 rows and 19 columns which contains mainly the object and float64 data types. The file size is 2.3 Megabytes.



.

**Table 7:** information about the Iadeler dataset

After acquiring the information, team has decided to change column names in order to avoid potential problems to connect the clean data to the PostgreSQL database. The column names consisted of complex names, therefore, the c letter at the beginning of the column names were taken off. The c letter indicates "customer" in the column names. The spaces between multiple words in the column names were taken off to avoid connection problems in database.

Therefore the column names has changed in the table through "df.rename(colums={}, inplace=True) as shown below.

df.rename(columns={"nOrderId":"Order\_id",

"Customer ID":"CustomerID",

"cPostCode":"PostCode",

"dReceievedDate":"ReceivedDate",

"cCountryCode":"CountryCode",

"cCountry" : "Country",

"cCurrency" : "Currency",

'Return Date' : 'ReturnDate',

'Return Reason' : 'ReturnReason',

'RMA Actioned': 'RMAActioned',

'Refund Amount':'RefundAmount',

}, inplace=True)

**Table 8**: renaming the column names for Iadeler file

Therefore the column names has changed in the table through "df.rename(colums={}, inplace=True) as shown above.

**Table 9:** Identification of null values on Iadeler File

The dataset contained null values in certain rows. These rows had only null values matching with every column of the dataset. Therefore, 7940 null rows were deleted in the report.

On the other hand, some of the columns contained essential data values together with the null values. It is considered better to fill the NaN values with "no info". In this way, category and country columns were filled with missing data for the unity of the dataset. Additionally, the item title column null values were replaced with unknown as an object data type. The replacement of missing data are deployed through the code given below.

df.Category = df.Category.replace(np.nan, 'No info') # replacing missing values with the label "No info"

**Table 10:** Replacing NaN values with "no info" object type

df['ItemTitle'].fillna('unknown', inplace=True) # labelling missing values in ItemTitle column as unknown

**Table 11**: Replacing NaN values with "unkown" object type in ItemTitle

After the cleaning, the remaining dataset had the shape of 7938 rows and 15 columns. The initial data had 19 columns. After careful consideration, it is decided not to continue with the following columns that is shown below in the table.

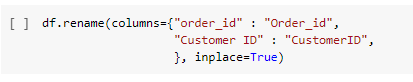
df = df.drop(["CountryCode", 'Currency', 'ResendOrExchangeQty', 'RMAActioned', 'subsource'], axis = 1)

**Table 11:** Columns that will be deleted in the "Iadeler" file.

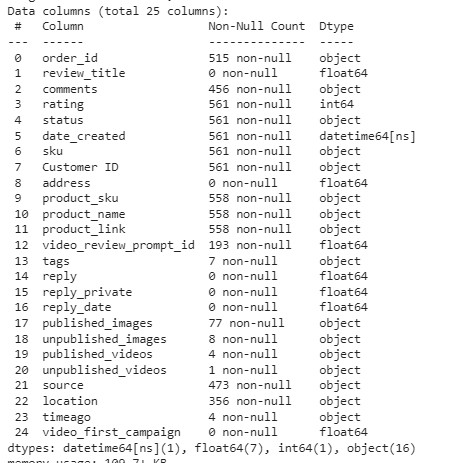
Explaratory Data Analysis on Iadeler File end date: **January, 4, 2023**

**Urun Yorumları EDA**

Shape of the data is (561, 25). There are difference type in the data columns. It is checked all types that are the correct type. after that, because it is recognized some names of the column names has gaps,they were renamed.

****

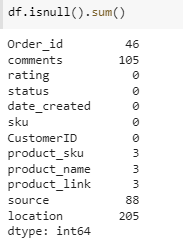
**Table 12.** Renamed columns

****

**Table 13.**The columns data types

When the columns were evaluated, it was thought that some columns could not used for analysis and some columns are the repetitive values. So, these columns are dropped thereby one of them. After this dropping method, it is the data shape is (561,12)

Then, by checking the missing values, it was started to fill them using meaning strategies.

****

**Table 14.** Null values in the columns

The location column had 205 null values, these values were filled with “unknown” words because it is not guessing real values. At the same time this column had country and city knowledge of the customers. So, by separating values in the column, it was created new two columns “ country” and “city”.

Comments column has 105 null values too, so it was filled with “No comments” . Similarly source column null values filled with “unknown” .