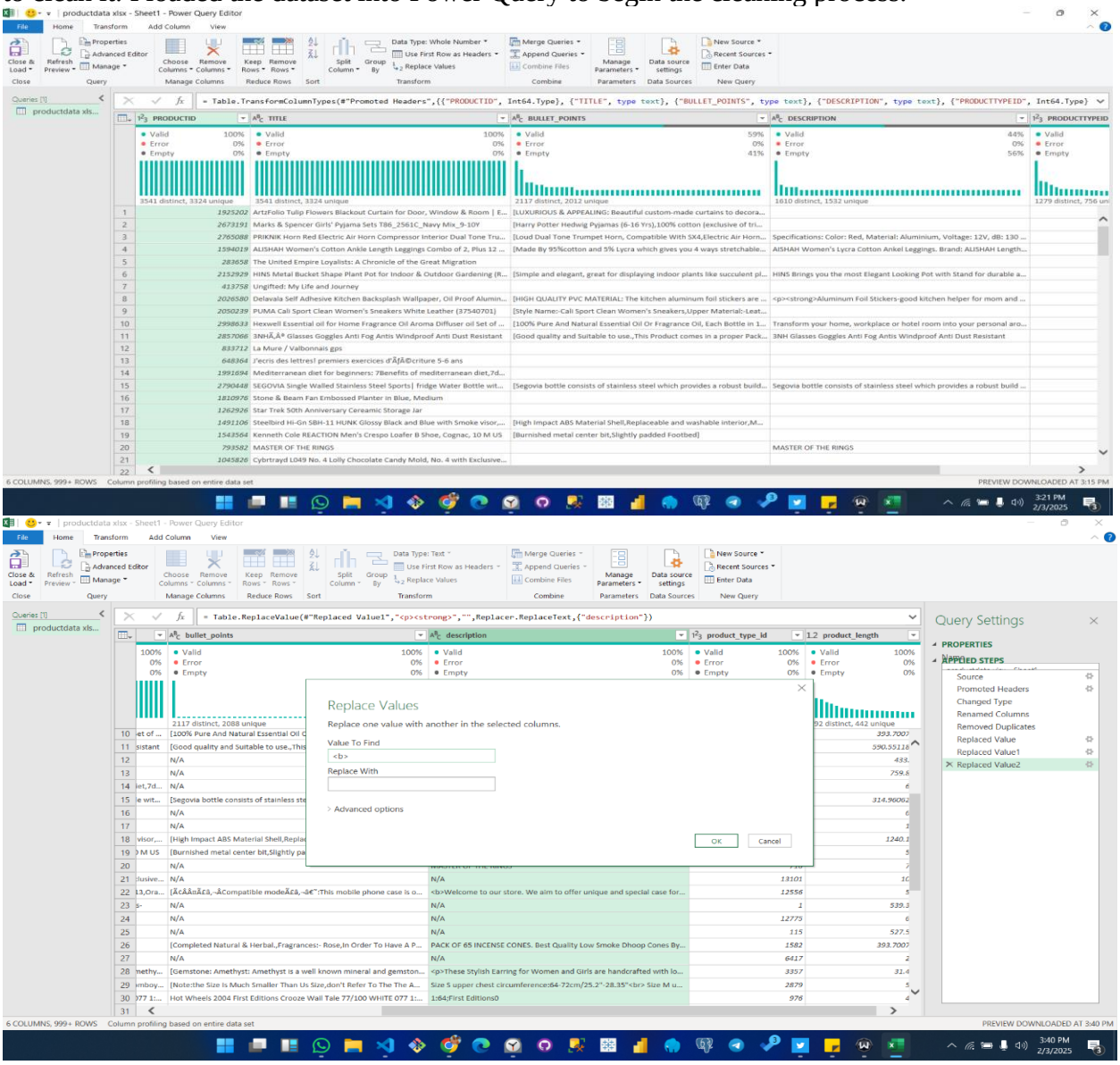
Technical Report: Data Cleaning & Title Optimization By Adordev

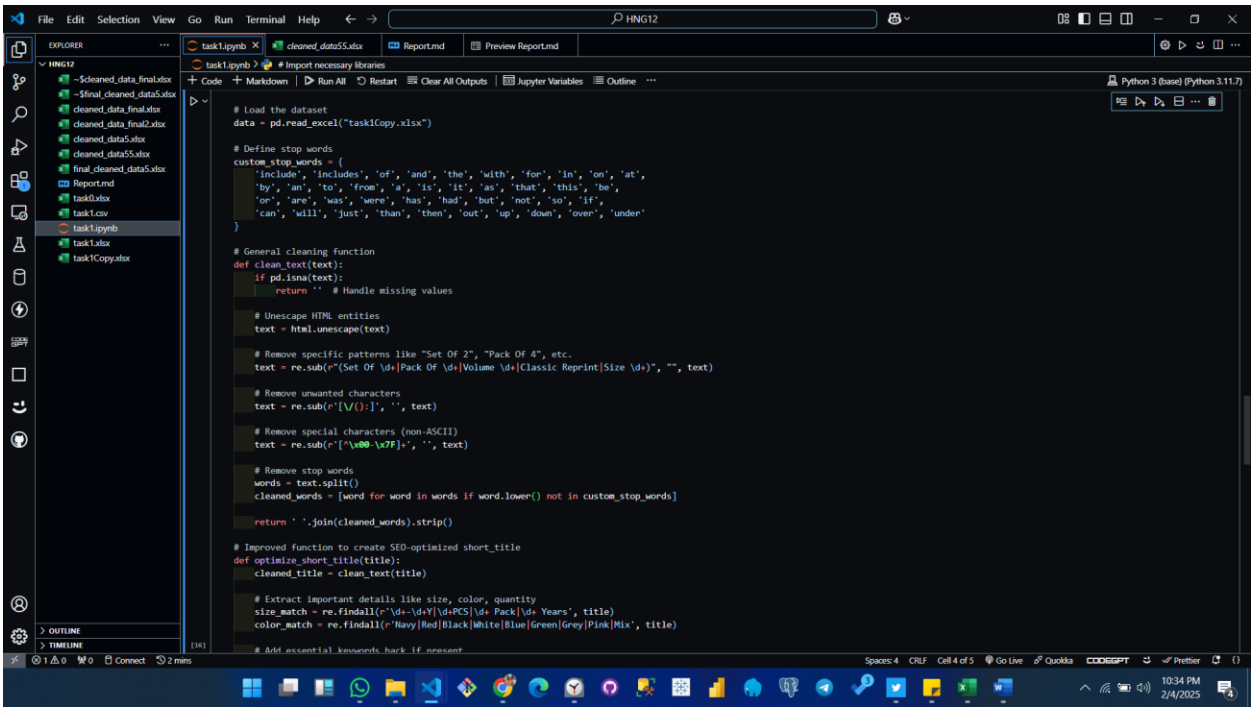
Introduction

The goal of this project was to clean raw marketing data, address data quality issues, and create a concise, SEO-optimized `short_title` feature to improve product discoverability.

I initially downloaded the dataset with the intention of using **Excel**, specifically **Power Query**, to clean it. I loaded the dataset into Power Query to begin the cleaning process.

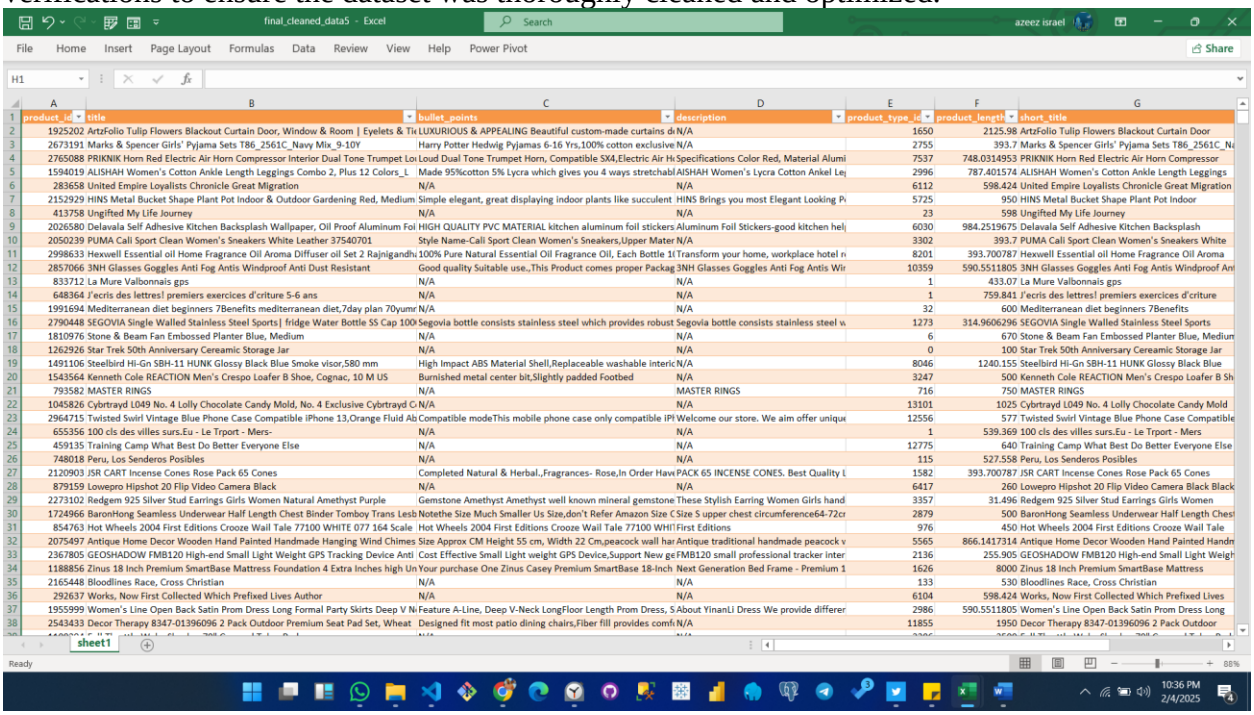


However, as the dataset grew larger and the cleaning process became more complex, it became tedious to handle solely in Excel. I realized that using **Python** would be more efficient for automating repetitive tasks. I wrote a Python script to handle the bulk of the data cleaning process, which significantly improved both speed and accuracy.



```
# Load the dataset
data = pd.read_excel("task1Copy.xlsx")

# Define stop words
custom_stop_words = {
    'include', 'includes', 'of', 'and', 'the', 'with', 'for', 'in', 'on', 'at',
    'by', 'an', 'to', 'from', 'a', 'is', 'it', 'as', 'that', 'this', 'be',
    'or', 'are', 'was', 'were', 'has', 'had', 'but', 'not', 'so', 'if',
    'can', 'will', 'just', 'than', 'then', 'out', 'up', 'down', 'over', 'under'
}

# General cleaning function
def clean_text(text):
    if pd.isna(text):
        return ''  # Handle missing values

    # Unescape HTML entities
    text = html.unescape(text)

    # Remove specific patterns like "Set Of 2", "Pack Of 4", etc.
    text = re.sub(r"(Set Of \d+|Pack Of \d+|Volume \d+|Classic Reprint|Size \d+)", "", text)

    # Remove unwanted characters
    text = re.sub(r'[\/():]', '', text)

    # Remove special characters (non-ASCII)
    text = re.sub(r'[^\x00-\x7F]+', '', text)

    # Remove stop words
    words = text.split()
    cleaned_words = [word for word in words if word.lower() not in custom_stop_words]

    return ' '.join(cleaned_words).strip()

# Improved function to create SEO-optimized short_title
def optimize_short_title(title):
    cleaned_title = clean_text(title)

    # Extract important details like size, color, quantity
    size_match = re.findall(r'\d+-\d+Y|\d+PCS|\d+ Pack|\d+ Years', title)
    color_match = re.findall(r'Navy|Red|Black|White|Blue|Green|Grey|Pink|Mix', title)

    # Add essential keywords back if present
```

After cleaning the data with Python, I used Power Query again for some final adjustments and verifications to ensure the dataset was thoroughly cleaned and optimized.



product_id	title	bullet_points	description	product_type_id	product_length	short_title
1925202	ArtzFolio Tulip Flowers Blackout Curtain Door, Window & Room | Eyelets & Ti	LUXURIOUS & APPEALING Beautiful custom-made curtains de	N/A	1650	2125.98	ArtzFolio Tulip Flowers Blackout Curtain Door
2673191	Marks & Spencer Girls' Pyjama Sets T86_2561C_Navy Mix_9-10Y	Harry Potter Hedwig Pyjamas 6-16 Yrs,100% cotton exclusive	N/A	2755	393.7	Marks & Spencer Girls' Pyjama Sets T86_2561C_Na
2765088	PRINKNK Horn Red Electric Air Horn Compressor Interior Dual Tone Trumpet Loi	Loud Dual Tone Trumpet Horn, Compatible SX4,Electric Air Ho	Specifications Color Red, Material Alumi	7537	748.0314953	PRINKNK Horn Red Electric Air Horn Compressor
1594019	ALISHAH Women's Cotton Ankle Length Leggings Combo 2, Plus 12 Colors_L	Made 95%cotton 5% Lycra which gives you 4 ways stretchabl	ALISHAH Women's Lycra Cotton Ankel Le	2996	787.401574	ALISHAH Women's Cotton Ankle Length Leggings
283658	United Empire Loyalists Chronicle Great Migration	N/A	N/A	6112	598.424	United Empire Loyalists Chronicle Great Migration
2152929	HINS Metal Bucket Shape Plant Pot Indoor & Outdoor Gardening Red, Medium	Simple elegant, great displaying indoor plants like succulent	HINS Brings you most Elegant Looking P	5725	950	HINS Metal Bucket Shape Plant Pot Indoor
413758	Ungifted My Life Journey	N/A	N/A	23	598	Ungifted My Life Journey
2026580	Delavala Self Adhesive Kitchen Backsplash Wallpaper, Oil Proof Aluminum Foi	HIGH QUALITY PVC MATERIAL kitchen aluminum foil stickers	Aluminum Foil Stickers-good kitchen hel	6030	984.2519675	Delavala Self Adhesive Kitchen Backsplash
2050239	PUMA Cali Sport Clean Women's Sneakers White Leather 37540701	Style Name-Cali Sport Clean Women's Sneakers,Upper Mater	N/A	3302	393.7	PUMA Cali Sport Clean Women's Sneakers White
2998633	Hexwell Essential oil Home Fragrance Oil Aroma Diffuser oil Set 2 Rajnigandhi	100% Pure Natural Essential Oil Fragrance Oil, Each Bottle 1(	Transform your home, workplace hotel r	8201	393.700787	Hexwell Essential oil Home Fragrance Oil Aroma
2857066	3NH Glasses Goggles Anti Fog Antis Windproof Anti Dust Resistant	Good quality Suitable use.,This Product comes proper Packag	3NH Glasses Goggles Anti Fog Antis Wir	10359	590.5511805	3NH Glasses Goggles Anti Fog Antis Windproof An
833712	La Mure Valbonnais gps	N/A	N/A	1	433.07	La Mure Valbonnais gps
648364	J'ecris des lettres! premiers exercices d'criture 5-6 ans	N/A	N/A	1	759.841	J'ecris des lettres! premiers exercices d'criture
1991694	Mediterranean diet beginners 7Benefits mediterranean diet,7day plan 70yumr	N/A	N/A	32	600	Mediterranean diet beginners 7Benefits
2790448	SEGOVIA Single Walled Stainless Steel Sports| fridge Water Bottle SS Cap 100	Segovia bottle consists stainless steel which provides robust	Segovia bottle consists stainless steel w	1273	314.9606296	SEGOVIA Single Walled Stainless Steel Sports
1810976	Stone & Beam Fan Embossed Planter Blue, Medium	N/A	N/A	6	670	Stone & Beam Fan Embossed Planter Blue, Medium
1262926	Star Trek 50th Anniversary Cereamic Storage Jar	N/A	N/A	0	100	Star Trek 50th Anniversary Cereamic Storage Jar
1491106	Steelbird Hi-Gn SBH-11 HUNK Glossy Black Blue Smoke visor,580 mm	High Impact ABS Material Shell,Replaceable washable interic	N/A	8046	1240.155	Steelbird Hi-Gn SBH-11 HUNK Glossy Black Blue
1543564	Kenneth Cole REACTION Men's Crespo Loafer B Shoe, Cognac, 10 M US	Burnished metal center bit,Slightly padded Footbed	N/A	3247	500	Kenneth Cole REACTION Men's Crespo Loafer B Sh
793582	MASTER RINGS	N/A	MASTER RINGS	716	750	MASTER RINGS
1045826	Cybrtrayd L049 No. 4 Lolly Chocolate Candy Mold, No. 4 Exclusive Cybrtrayd C	N/A	N/A	13101	1025	Cybrtrayd L049 No. 4 Lolly Chocolate Candy Mold
2964715	Twisted Swirl Vintage Blue Phone Case Compatible iPhone 13,Orange Fluid Ab	Compatible modeThis mobile phone case only compatible iPh	Welcome our store. We aim offer unique	12556	577	Twisted Swirl Vintage Blue Phone Case Compatible
655356	100 cls des villes surs.Eu - Le Trport - Mers-	N/A	N/A	1	539.369	100 cls des villes surs.Eu - Le Trport - Mers
459135	Training Camp What Best Do Better Everyone Else	N/A	N/A	12775	640	Training Camp What Best Do Better Everyone Else
748018	Peru, Los Senderos Posibles	N/A	N/A	115	527.558	Peru, Los Senderos Posibles
2120903	JSR CART Incense Cones Rose Pack 65 Cones	Completed Natural & Herbal.,Fragrances- Rose,In Order Have	PACK 65 INCENSE CONES. Best Quality L	1582	393.700787	JSR CART Incense Cones Rose Pack 65 Cones
879159	Lowepro Hipshot 20 Flip Video Camera Black	N/A	N/A	6417	260	Lowepro Hipshot 20 Flip Video Camera Black Black
2273102	Redgem 925 Silver Stud Earrings Girls Women Natural Amethyst Purple	Gemstone Amethyst Amethyst well known mineral gemstone	These Stylish Earring Women Girls hand	3357	31.496	Redgem 925 Silver Stud Earrings Girls Women
1724966	BaronHong Seamless Underwear Half Length Chest Binder Tomboy Trans Lesb	Notethe Size Much Smaller Us Size,don't Refer Amazon Size C	Size S upper chest circumference64-72cr	2879	500	BaronHong Seamless Underwear Half Length Chest
854763	Hot Wheels 2004 First Editions Crooze Wall Tale 77100 WHITE 077 164 Scale	Hot Wheels 2004 First Editions Crooze Wall Tale 77100 WHI	First Editions	976	450	Hot Wheels 2004 First Editions Crooze Wall Tale
2075497	Antique Home Decor Wooden Hand Painted Handmade Hanging Wind Chimes	Size Approx CM Height 55 cm, Width 22 Cm,peacock wall har	Antique traditional handmade peacock v	5565	866.1417314	Antique Home Decor Wooden Hand Painted Handm
2367805	GEOSHADOW FMB120 High-end Small Light Weight GPS Tracking Device Anti	Cost Effective Small Light weight GPS Device,Support New ge	FMB120 small professional tracker inter	2136	255.905	GEOSHADOW FMB120 High-end Small Light Weigh
1188856	Zinus 18 Inch Premium SmartBase Mattress Foundation 4 Extra Inches high Un	Your purchase One Zinus Casey Premium SmartBase 18-Inch	Next Generation Bed Frame - Premium 1	1626	8000	Zinus 18 Inch Premium SmartBase Mattress
2165448	Bloodlines Race, Cross Christian	N/A	N/A	133	530	Bloodlines Race, Cross Christian
292637	Works, Now First Collected Which Prefixed Lives Author	N/A	N/A	6104	598.424	Works, Now First Collected Which Prefixed Lives
1955999	Women's Line Open Back Satin Prom Dress Long Formal Party Skirts Deep V Ni	Feature A-Line, Deep V-Neck LongFloor Length Prom Dress, S	About YinanLi Dress We provide differer	2986	590.5511805	Women's Line Open Back Satin Prom Dress Long
2543433	Decor Therapy 8347-01396096 2 Pack Outdoor Premium Seat Pad Set, Wheat	Designed fit most patio dining chairs,Fiber fill provides comfi	N/A	11855	1950	Decor Therapy 8347-01396096 2 Pack Outdoor

Dataset Overview

The dataset contained the following key columns:

- **product_id**
- **title**
- **bullet_points**

- **description**
- **product_type_id**
- **product_length**

The primary focus was on cleaning the text-based columns (`title`, `bullet_points`, and `description`) and optimizing product titles for marketing purposes.

Data Cleaning Process

1. Handling Missing Values

- **Issue Identified:** Missing values were present in the `bullet_points` and `description` columns.
- **Solution:** Replaced missing values (`NaN`) with empty strings to ensure smooth processing.
- **Code Used:**

```
data.fillna('', inplace=True)
```

---

2. Removing Duplicates

- **Issue Identified:** Potential duplicate product entries were identified.
- **Solution:** Removed all duplicate records to maintain data integrity.
- **Code Used:**

```
data.drop_duplicates(inplace=True)
```

---

3. Standardizing Data Formats

- **Actions Taken:**
  - Removed unwanted characters such as `/`, `(`, `:`, and `)`.
  - Removed special characters like `é`, `ü`, etc., instead of converting them to ASCII.
  - Unescaped HTML entities to clean up the text.
- **Code Used:**

```
text = re.sub(r'[\\/():]', '', text)              # Remove unwanted
characters
text = re.sub(r'[^\x00-\x7F]+', '', text)         # Remove non-ASCII
characters
```

Creating the `short_title` Feature

Objective:

Generate concise product titles (30–50 characters) that retain essential information for SEO and readability.

Methodology:

1. **Removed Redundant Phrases:**
   - Phrases like `"Set Of 2"`, `"Pack Of 4"`, and `"Classic Reprint"` were removed.
2. **Extracted Key Attributes:**
   - Attributes such as product size (e.g., `9-10Y`), color (e.g., `Navy`), and quantity (e.g., `2 PCS`) were retained.
3. **Intelligent Truncation:**
   - Applied smart truncation to ensure titles did **not cut off mid-word**.
4. **Fallback Mechanism:**
   - If the cleaned title became too short, the script reverted to the original title (up to 50 characters).

Example Transformations

Original Title	Short Title
ALISHAH Women's Cotton Ankle Length Leggings Combo 2, Plus 12 Colors_L	ALISHAH Women's Cotton Ankle Length Leggings
PosterHub Pink Floyd Wall Poster Matte Finish Paper Print 12 x18 Inch Multicolor HS - P076	PosterHub Pink Floyd Wall Poster Matte Finish
Oxza Universal Foldable Stand Holder Mount Bracket Tablet, Cell, Mobile Phone Table Stand Mobile Holder	Oxza Universal Foldable Stand Holder Mount

Clean Dataset Overview

Metric	Before Cleaning	After Cleaning
Total Rows	3,847	3,541 (duplicates removed)
Missing Values	Present	Handled (replaced with empty strings)
Duplicate Entries	Present	Removed
Unwanted Characters	Present	Cleaned
`short_title` Feature	Not Present	Added

Conclusion

In conclusion, the data cleaning process involved:

- **Addressing missing values**
- **Removing duplicates**
- **Standardizing data formats**
- **Optimizing product titles for SEO**

The `short_title` feature improves readability and search performance while retaining essential product information.