A REPORT

ON

Data Science project for the end semester

By

Name of the Student ID No.

AT

DECEMBER,2023

Key Words: Dashboard, Keywords, Data analysis

Project Areas: Data analysis, Machine learning, Data science

Abstract: This report discusses the key achievements made for the internship which include several Mini-projects undertaken and analyzes how they are an integral part of the overall vision of *****. It also discusses some of the major problem statements that the company tackles using data science and the role of the Mini-projects to aid a part of the solution pipeline.**

Signature(s) of Student(s)

Date:18/12/2023

TABLE OF CONTENTS

1. Introduction **7**
1.1. About the PS Station: 7
1.2. Domain in which the PS station works: 7
1.3. Organisational structure of PS station 8
1.4. Software Process 8
2. Project1: **10**
2.1. Problem Statement: 10
2.2. Motivation: 10
2.3. Objectives: 10
2.4. Solution: 10
3. Project 2: Campaign-Keyword dashboard **13**
3.1. Problem statement: 13
3.2. Motivation: 13
3.3. Objectives: 13
3.4. Solution: 13
4. Project 3: SELECTION DASHBOARD **16**
4.1. Problem statement: 16
4.2. Motivation: 16
4.3. Objectives: 16
4.4. Solution: 16
5. Project 4: Keyword harvesting **19**
5.1. Problem statement: 19
5.2. Motivation: 19
5.3. Objectives: 19
5.4. Solution: 19
6. Project 5: Anomaly detection **23**
6.1. Problem statement: 23
6.2. Motivation: 23
6.3. Objectives: 23
6.4. Solution: 23
7. Project 6: Budget Optimizer **25**
7.1. Problem statement: 25
7.2. Motivation: 25

7.3. Objectives: 26
7.4. Solution: 26
8. Project 7: Campaign potential analysis 28
8.1. Problem statement: 28
8.2. Motivation: 28
8.3. Solution: 28
9. Project 8: Spend threshold calculation 31
9.1. Problem statement: 31
9.2. Motivation: 31
9.3. Objectives: 31
9.4. Solution: 31
10. Project 9: Exploratory data analysis on Experiment data 35
10.1. Problem statement: 35
10.2. Motivation: 36
10.3. Objectives: 36
10.4. Solution: 36
11. Conclusion 39
11.1. Learning Outcome 39
12. References 40
13. Glossary 40

1. Introduction

1.1. Organisational structure of PS station

The company is divided into several branches located in the USA, India and the UK and is expanding to other countries. While the development is done mostly in India that too is subdivided into different teams like RMM, DS etc. In the data science team, there are again two separate teams, the DS sales team and the DS advertising team. I was included in the DS advertising team which has 6 other team members. Each member works on a different area that either deals with the budget optimizer, bid optimizer or various other sub-categories that fall in the pipeline to analyze and release new versions or changes in the two. Members often work in collaboration to solve a particular problem. The work done by each member works towards optimizing the advertising pipeline for different clients and ensuring their objectives are met.

1.2. Software Process

The team communicates developments and discusses improvements daily through a scrum. Here every member has to provide updates on the work done and clarify questions as well as seek assistance or collaboration from other team members. Members of the team can also coordinate separately with each other or with members of other teams personally or through Slack. There are also bi-monthly standups for all teams with members of management for communication of the vision of the company. The databases are primarily stored in Snowflake from which required data can be queried using SQL. The local code can be hosted on Amazon AWS ec2 using bitbucket. Atlassian Jira is also used for project tracking and communication in some cases. Atlassian Jira is also being used to create tickets for deliverables and track the progress of the side projects, and ad-hoc tasks as well as major products of all members of the team.

2. Project1:

2.1. Problem Statement:

Build a dashboard using Streamlit and Python that selects a data frame from Snowflake from a database between any two dates, aggregates certain parameters on daily, weekly and monthly levels and displays the resulting as well as original data frames with a provision to download the same using CSV. Display a chart showing the trend of aggregates concerning the time frame selected.

2.2. Motivation:

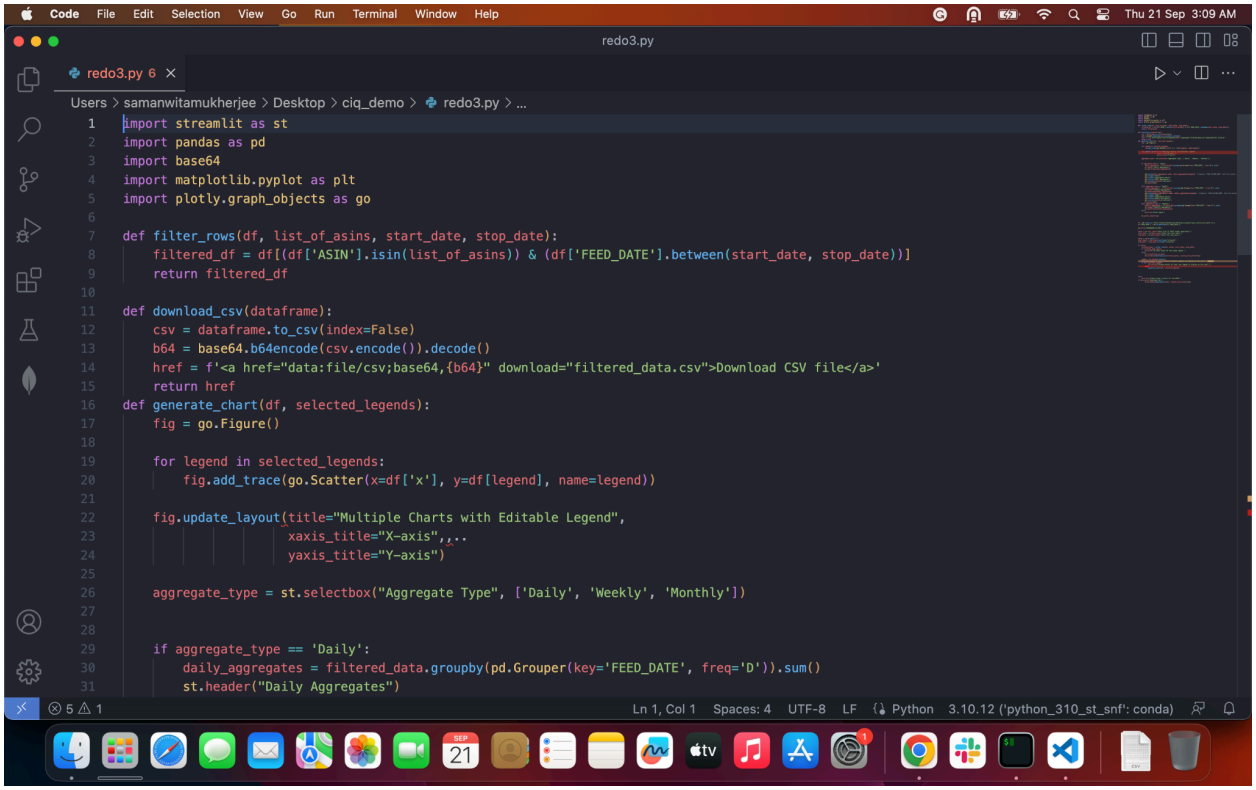
1. Learn to create and launch streamlit dashboards
2. Work with manipulating and processing data frames in Python
3. Learn the various functionalities involved in analyzing and querying data
4. Create visualizations for better representation of data

2.3. Objectives:

1. Provide a visual representation of metrics
2. Observe trends in metrics over a given period
3. Identify outliers and anomalies from analyzing data
4. Provide a clearer understanding of metric trends

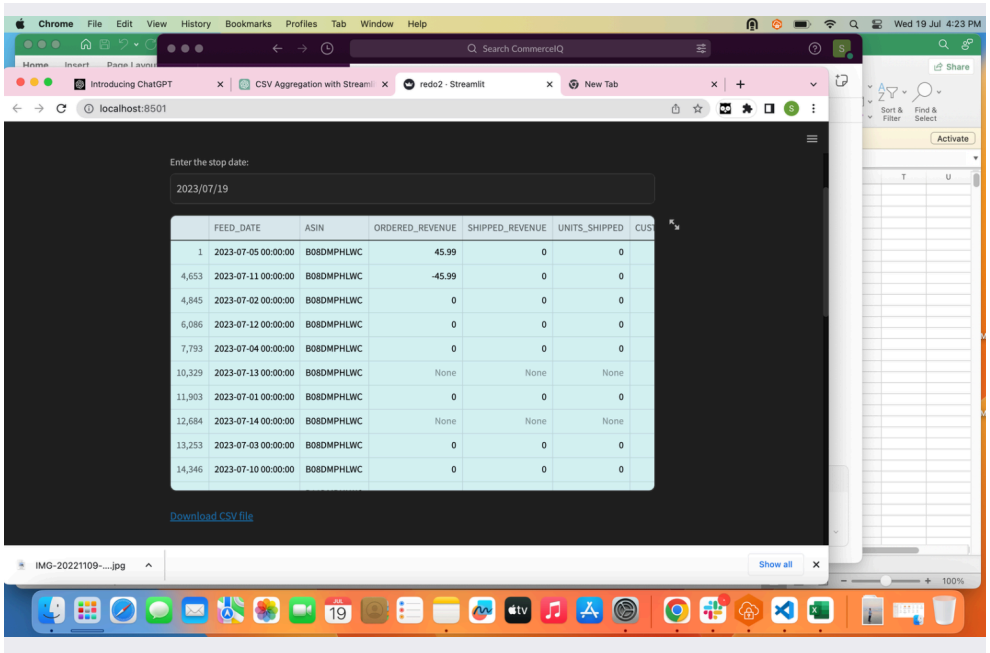
2.4. Solution:

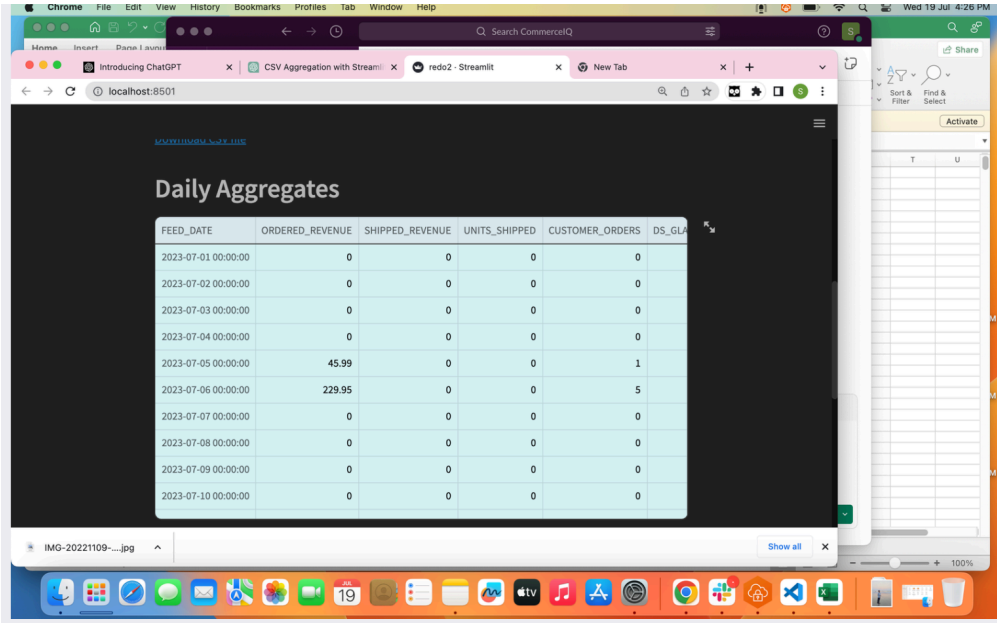
The given problem was solved by writing code in Python and hosting it on Streamlit. The code uses SQL queries to obtain the data of desired parameters from Snowflake and then filter data between 2 given dates. The chosen parameters are then aggregated and printed using pandas. An option is made available to download the data frame. Using matplotlib. pyplot the data is displayed as a chart.

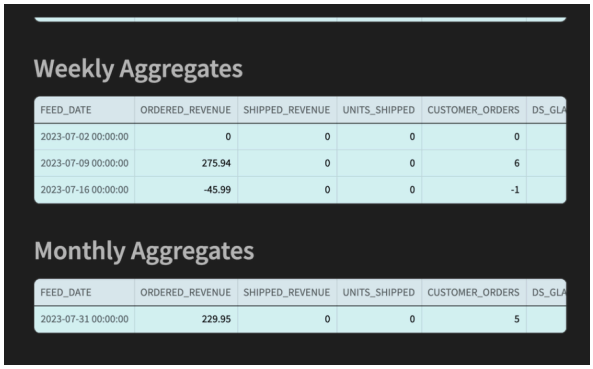


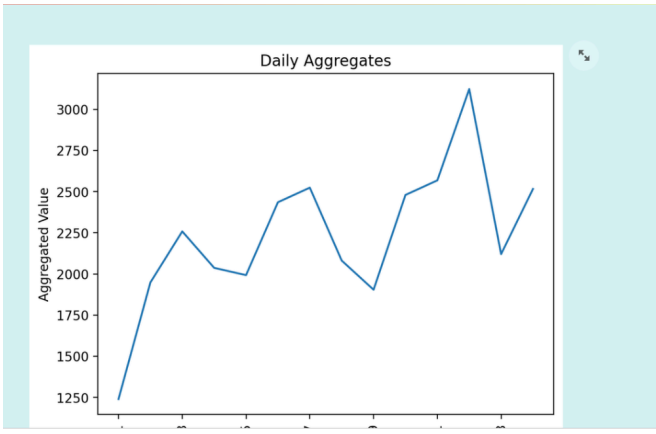
A snippet of code

Dashboard




Daily Aggregates
FEED_DATE | ORDERED_REVENUE | SHIPPED_REVENUE | UNITS_SHIPPED | CUSTOMER_ORDERS | DS_GLA
2023-07-01 00:00:00 | 0 | 0 | 0 | 0
2023-07-02 00:00:00 | 0 | 0 | 0 | 0
2023-07-03 00:00:00 | 0 | 0 | 0 | 0
2023-07-04 00:00:00 | 0 | 0 | 0 | 0
2023-07-05 00:00:00 | 45.99 | 0 | 0 | 1
2023-07-06 00:00:00 | 229.95 | 0 | 0 | 5
2023-07-07 00:00:00 | 0 | 0 | 0 | 0
2023-07-08 00:00:00 | 0 | 0 | 0 | 0
2023-07-09 00:00:00 | 0 | 0 | 0 | 0
2023-07-10 00:00:00 | 0 | 0 | 0 | 0


Weekly Aggregates
FEED_DATE | ORDERED_REVENUE | SHIPPED_REVENUE | UNITS_SHIPPED | CUSTOMER_ORDERS | DS_GLA
2023-07-02 00:00:00 | 0 | 0 | 0 | 0
2023-07-09 00:00:00 | 275.94 | 0 | 0 | 6
2023-07-16 00:00:00 | -45.99 | 0 | 0 | -1
Monthly Aggregates
FEED_DATE | ORDERED_REVENUE | SHIPPED_REVENUE | UNITS_SHIPPED | CUSTOMER_ORDERS | DS_GLA
2023-07-31 00:00:00 | 229.95 | 0 | 0 | 5


Daily Aggregates
Aggregated Value
3000
2750
2500
2250
2000
1750
1500
1250

3. Project 2: Campaign-Keyword dashboard

3.1. Problem statement:

Build a dashboard that extracts data from a database of users's choice and enables filtering of data at profile, portfolio and campaign levels. It also provides a search bar for individual campaigns. If no filter is selected display all the existing and suggested keywords for all campaigns for a client else display the filtered data.

3.2. Motivation:

1. Understand the structure of an advertising campaign
2. Understand how to filter data frames at different levels.
3. Learn to choose particular databases from Snowflake programmatically.
4. Learn how to design, modify and perform other functions on tables in Python

3.3. Objectives:

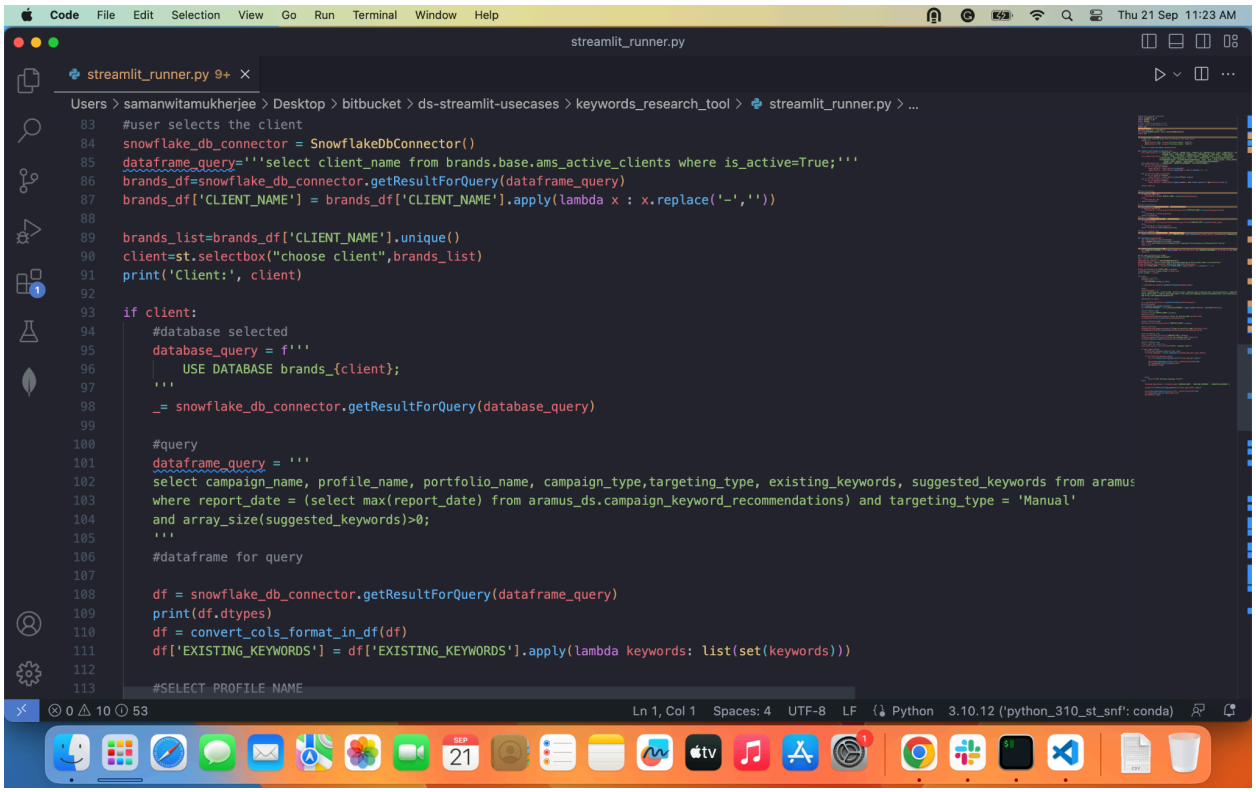
1. Aid in the Bid optimizer process by displaying the existing as well as suggested keywords that can be bid on for advertising
2. Provide a framework for comparison for new clients onboarded regarding relevant keywords.
3. Provide an easy way to filter out keywords at different levels of advertising to enable better targeting of keywords at any level

3.4. Solution:

Select all the available clients from a meta-data table and provide a list of clients to choose from for the user and obtain the original data frame. Create separate functions for filtering by profile, portfolio and campaign which takes a data frame as input. If no option for the filter is chosen then it returns the input data frame. In the main function call the functions for the client, profile, portfolio and campaign filtering concurrently. Also, add a provision to filter according to specific campaigns. Print the resulting data frame as a table and provide an option to

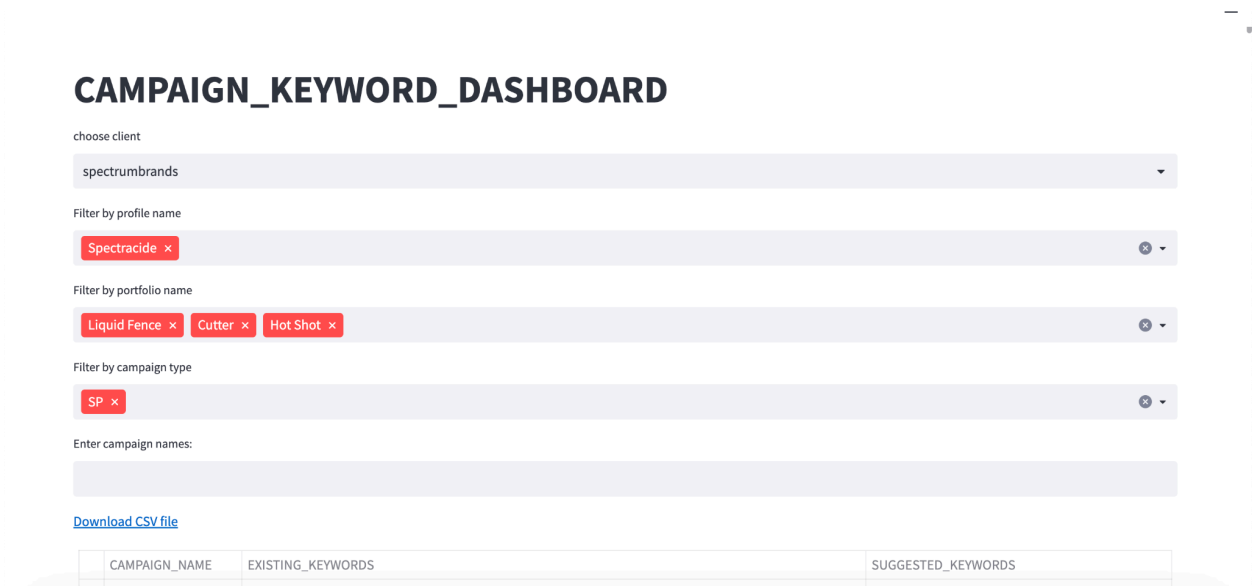
download it as a CSV. Push the code to a branch in Bitbucket and pull it on AWS to host it in EC2.

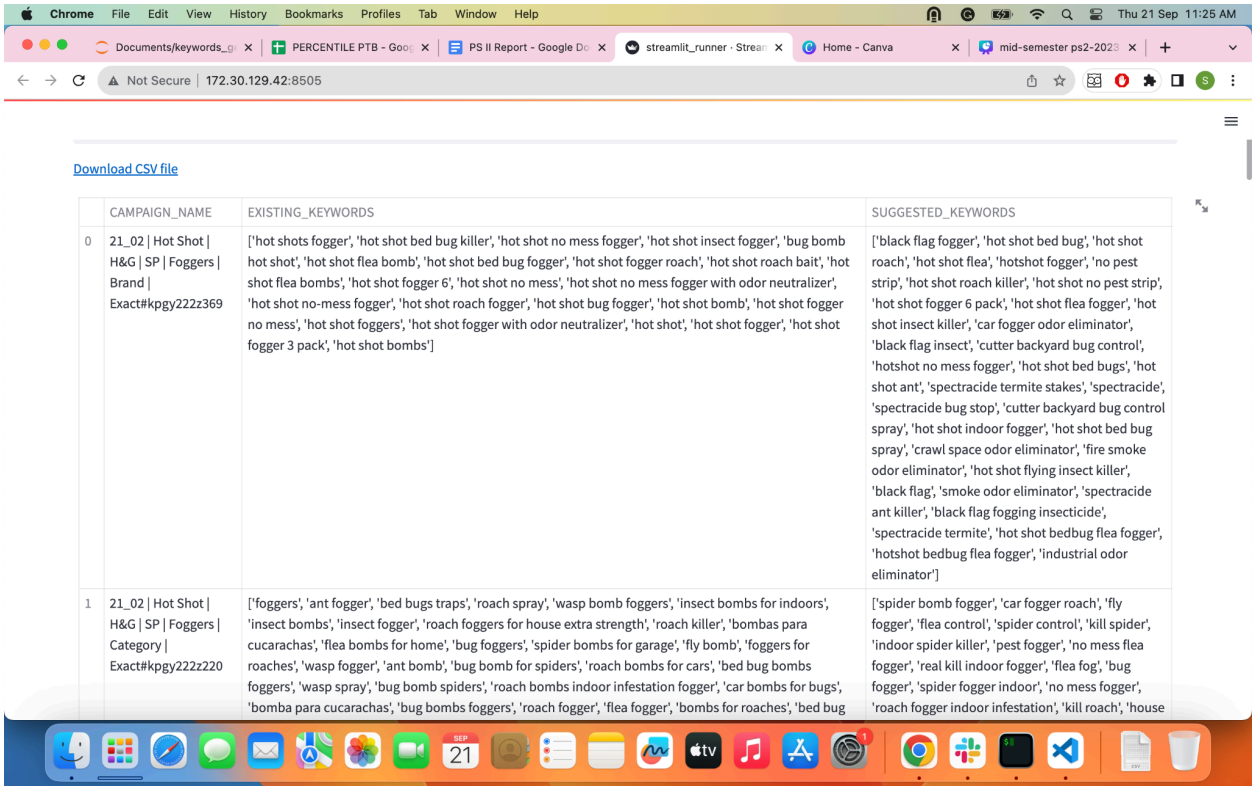
Link for the dashboard:http://172.30.129.42:8505/



A snippet of the code

Dashboard:





	CAMPAIGN_NAME	EXISTING_KEYWORDS	SUGGESTED_KEYWORDS
0	21_02 | Hot Shot | H&G | SP | Foggers | Brand | Exact#kpgy222z369	['hot shots fogger', 'hot shot bed bug killer', 'hot shot no mess fogger', 'hot shot insect fogger', 'bug bomb hot shot', 'hot shot flea bomb', 'hot shot bed bug fogger', 'hot shot fogger roach', 'hot shot roach bait', 'hot shot flea bombs', 'hot shot fogger 6', 'hot shot no mess', 'hot shot no mess fogger with odor neutralizer', 'hot shot no-mess fogger', 'hot shot roach fogger', 'hot shot bug fogger', 'hot shot bomb', 'hot shot fogger no mess', 'hot shot foggers', 'hot shot fogger with odor neutralizer', 'hot shot', 'hot shot fogger', 'hot shot fogger 3 pack', 'hot shot bombs']	['black flag fogger', 'hot shot bed bug', 'hot shot roach', 'hot shot flea', 'hotshot fogger', 'no pest strip', 'hot shot roach killer', 'hot shot no pest strip', 'hot shot fogger 6 pack', 'hot shot flea fogger', 'hot shot insect killer', 'car fogger odor eliminator', 'black flag insect', 'cutter backyard bug control', 'hotshot no mess fogger', 'hot shot bed bugs', 'hot shot ant', 'spectracide termite stakes', 'spectracide', 'spectracide bug stop', 'cutter backyard bug control spray', 'hot shot indoor fogger', 'hot shot bed bug spray', 'crawl space odor eliminator', 'fire smoke odor eliminator', 'hot shot flying insect killer', 'black flag', 'smoke odor eliminator', 'spectracide ant killer', 'black flag fogging insecticide', 'spectracide termite', 'hot shot bedbug flea fogger', 'hotshot bedbug flea fogger', 'industrial odor eliminator']
1	21_02 | Hot Shot | H&G | SP | Foggers | Category | Exact#kpgy222z220	['foggers', 'ant fogger', 'bed bugs traps', 'roach spray', 'wasp bomb foggers', 'insect bombs for indoors', 'insect bombs', 'insect fogger', 'roach foggers for house extra strength', 'roach killer', 'bombas para cucarachas', 'flea bombs for home', 'bug foggers', 'spider bombs for garage', 'fly bomb', 'foggers for roaches', 'wasp fogger', 'ant bomb', 'bug bomb for spiders', 'roach bombs for cars', 'bed bug bombs foggers', 'wasp spray', 'bug bomb spiders', 'roach bombs indoor infestation fogger', 'car bombs for bugs', 'bomba para cucarachas', 'bug bombs foggers', 'roach fogger', 'flea fogger', 'bombs for roaches', 'bed bug	['spider bomb fogger', 'car fogger roach', 'fly fogger', 'flea control', 'spider control', 'kill spider', 'indoor spider killer', 'pest fogger', 'no mess flea fogger', 'real kill indoor fogger', 'flea fog', 'bug fogger', 'spider fogger indoor', 'no mess fogger', 'roach fogger indoor infestation', 'kill roach', 'house

4. Project 3: SELECTION DASHBOARD

4.1. Problem statement:

Create a dashboard that takes a start and stop date as input and then asks the user to view by either selecting data filtered by category or according to a list of ASINs provided by the user. In the category filter, it analyzed the data to create subcategories that enable the user to further filter according to the unique subcategory value. A provision to select parameters associated with the data frame is also provided and the parameters can be aggregated on a daily, weekly or monthly level and the aggregation of all subcategories are displayed separately in the form of a chart alongside the filtered data frame. If the user provides a list of ASINs, then the subcategories to which they belong by analyzing its data and providing an option to choose parameters as well as aggregation type.

4.2. Motivation:

1. Learn to create subcategories by analysing data programmatically
2. Learn to create nested filters
3. Learn to filter by parameters
4. Learn how to create charts with editable legends that are chosen by a user and display them in the same frame.

4.3. Objectives:

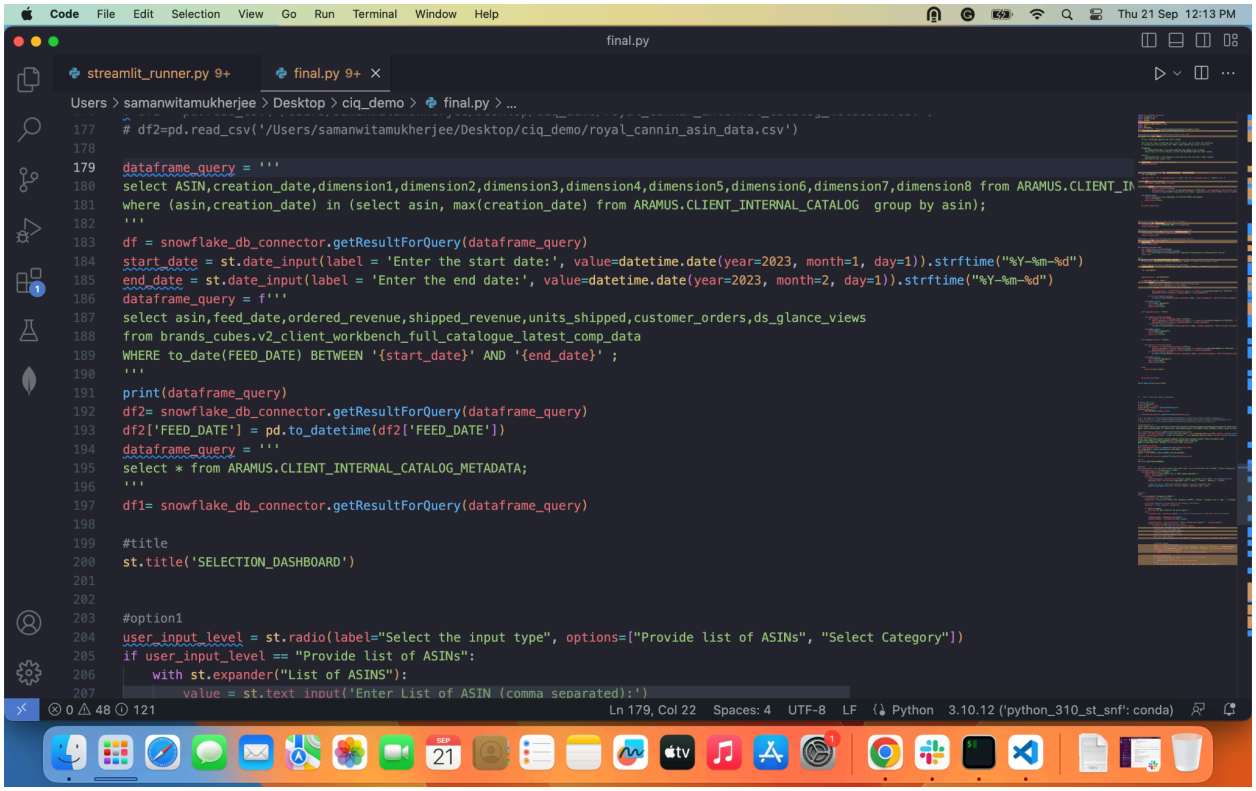
1. Enable easy visualization of data
2. Help to analyze a large database by creating multiple subcategories
3. Enable the user to only choose relevant data and legends

4.4. Solution:

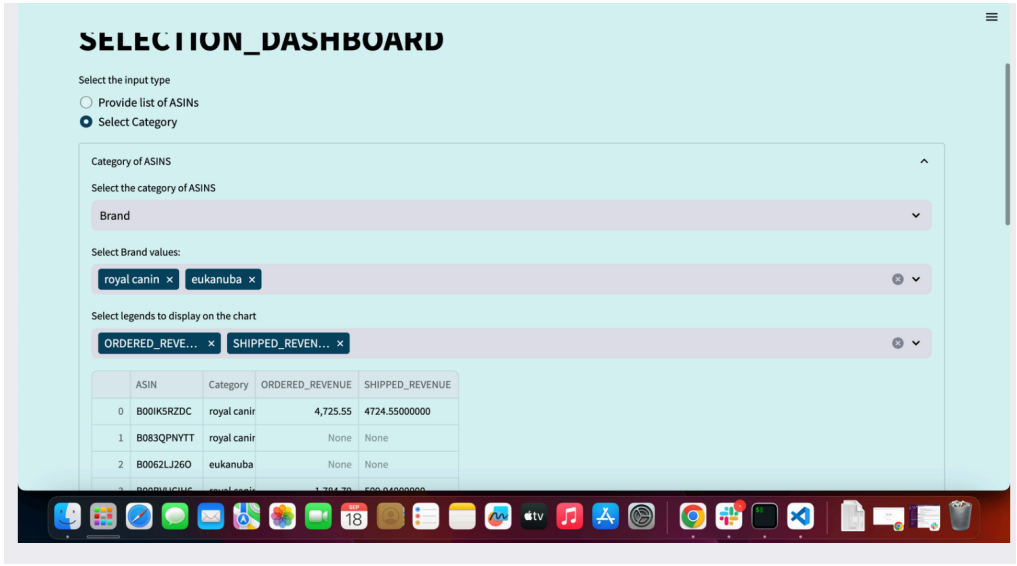
Obtain the data from 3 different databases, one having the different categories that are there for all the different columns of the original data, a database having relevant metadata and the original data. Filter the original data by start and stop data and then use the categories database to obtain major subcategories. For selected subcategories, analyze the columns programmatically to further obtain categories for filtering by linking them with the metadata. Then choose the parameters of interest and aggregation type. Based on the options selected,

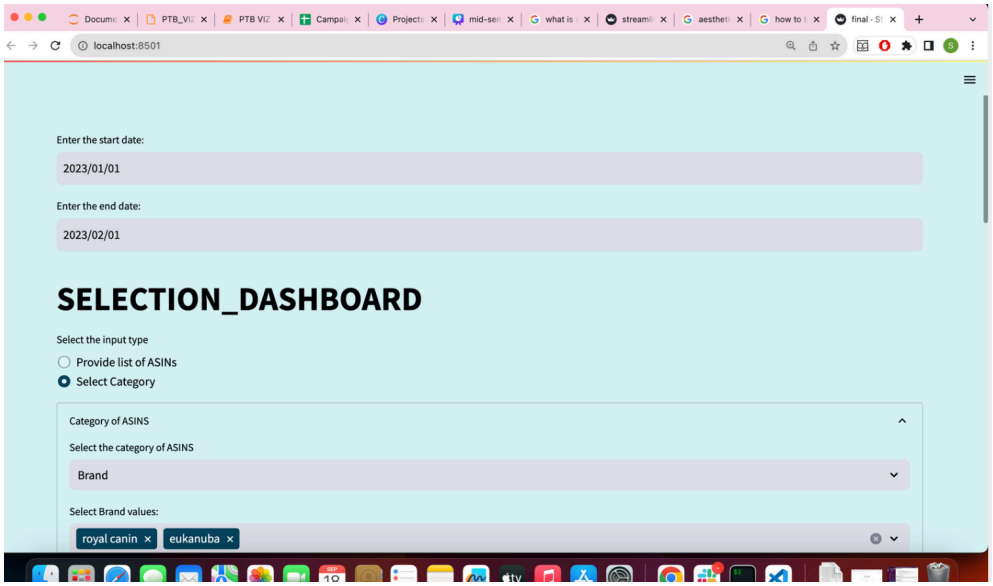
aggregate the metrics and plot them separately on a chart. There should be a button to choose between providing a list of ASINs or selecting from categories. For each option, the above process should be run and an option to download the resulting CSV should be provided.

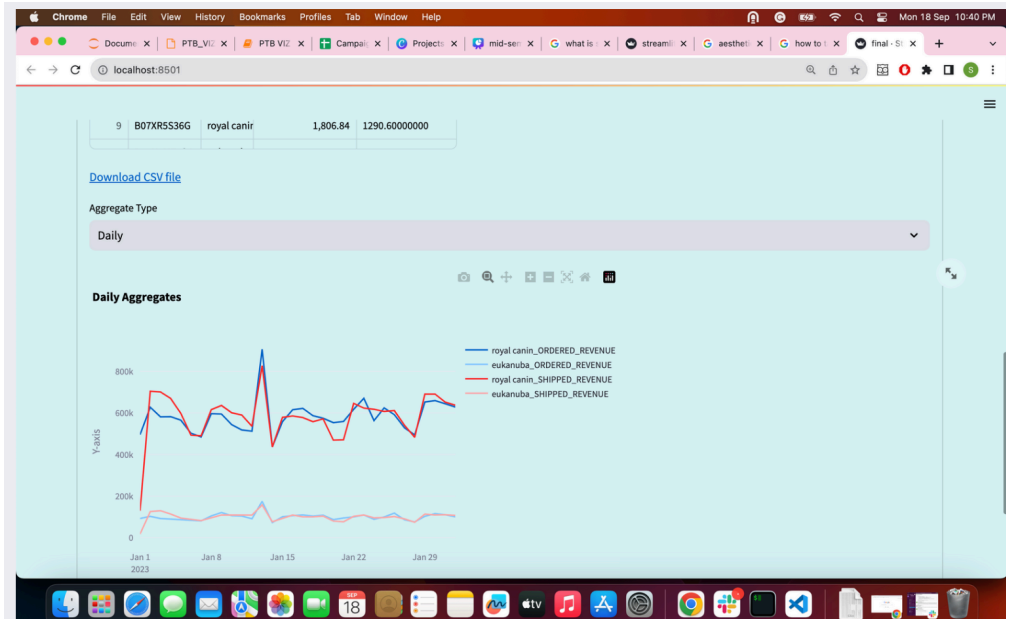
A snippet of code:



Dashboard:




localhost:8501
Enter the start date:
2023/01/01
Enter the end date:
2023/02/01
SELECTION_DASHBOARD
Select the input type
Provide list of ASINs
Select Category
Category of ASINS
Select the category of ASINS
Brand
Select Brand values:
royal canin
eukanuba


Chrome File Edit View History Bookmarks Profiles Tab Window Help
Mon 18 Sep 10:40 PM
localhost:8501
9 B07XR5S36G royal canir 1,806.84 1290.60000000
Download CSV file
Aggregate Type
Daily
Daily Aggregates
royal canin_ORDERED_REVENUE
eukanuba_ORDERED_REVENUE
royal canin_SHIPPED_REVENUE
eukanuba_SHIPPED_REVENUE
Y-axis
800k
600k
400k
200k
0
Jan 1
2023
Jan 8
Jan 15
Jan 22
Jan 29

5. Project 4: Keyword harvesting

5.1. Problem statement:

From a large language model obtain 60% generic, 20% client-randed and 20%competitor keywords for all products sold by a particular client. Verify the validity of the data obtained and suggest its usability for existing as well as new clients.

5.2. Motivation:

1. Learn prompt engineering to extract proper information from a large language model
2. Obtain data for all ASINs of a client programmatically by running the prompt iteratively
3. Learn to remove noise and irregularities from LLM output
4. Learn how to implement a validation pipeline
5. Learn to automate process

5.3. Objectives:

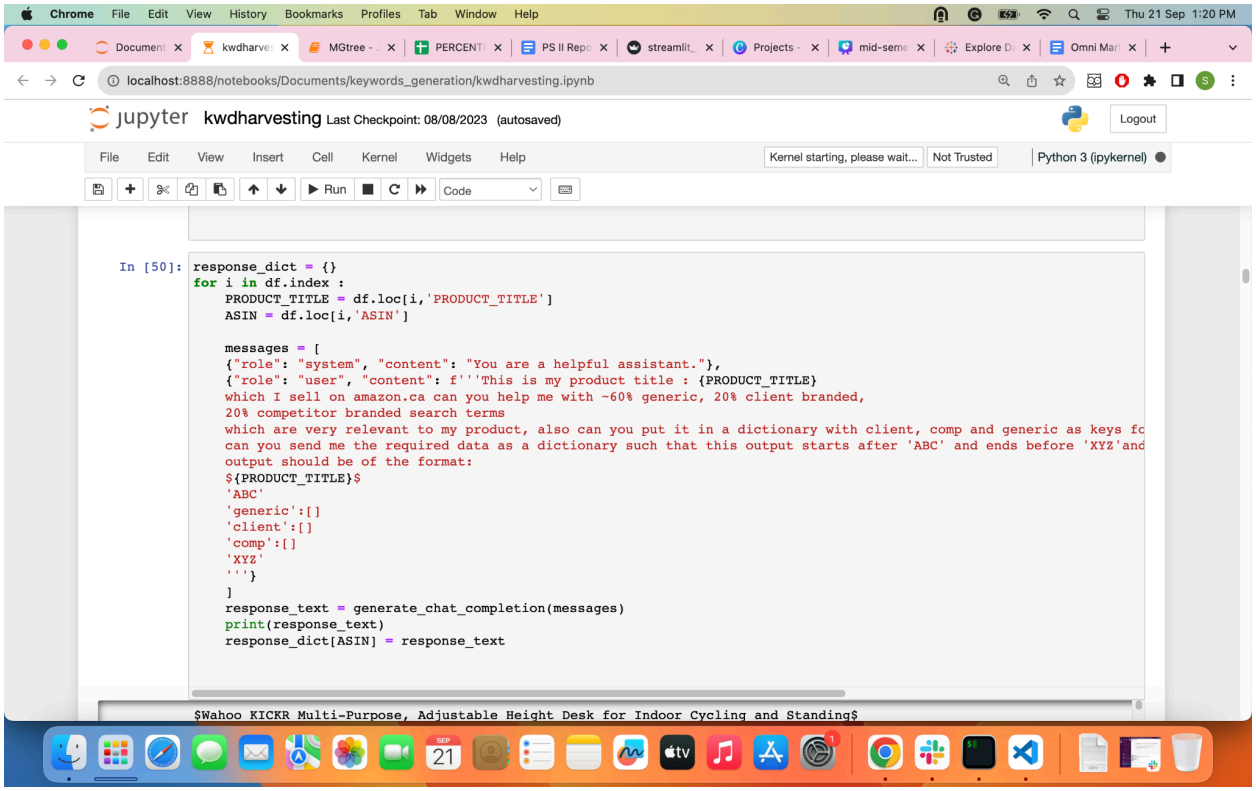
1. Create a repository of a mix of different types of keywords for a client that can be used for onboarding a new client.
2. Verify the efficacy of using an LLM for keyword generation
3. Aid bid optimizer by creating a list of suggested keywords for existing clients
4. Compare with existing databases as well as a database containing keywords of the highest search volume to predict the performance of keywords on Amazon

5.4. Solution:

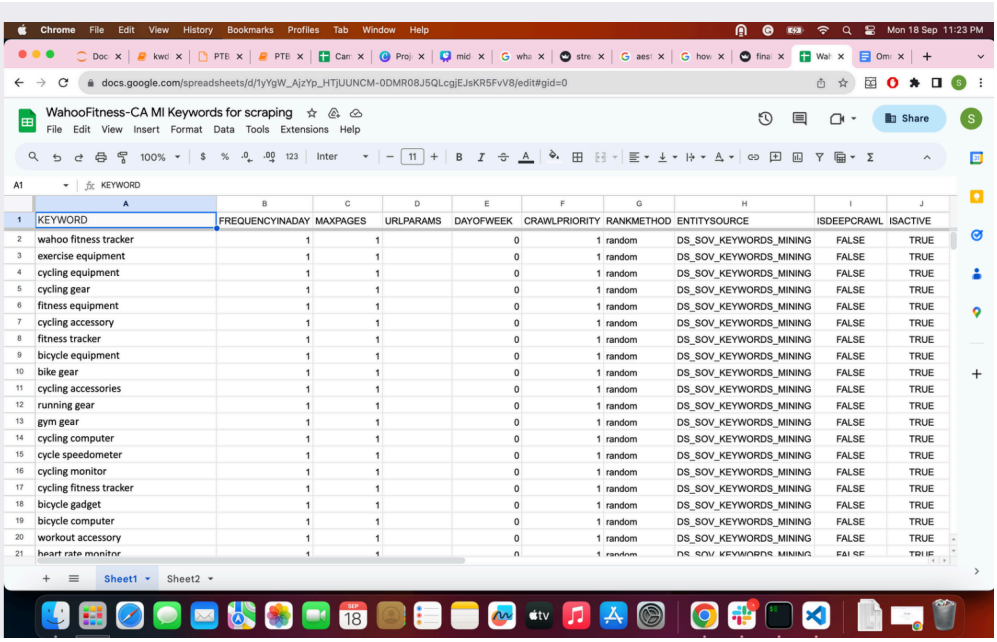
1. Keyword generation: Obtain a list of all ASINs and product titles for a particular client. Create a loop that runs a prompt to obtain 60% generic, 20% client-branded and 20% competitor keywords for each client and store the resulting output in a dictionary. Process the dictionary to remove noise from the generated data to extract the desired keywords and store it in a CSV.

2. Frequency sorting and filtration: Sort the CSV of keywords according to the frequency of each keyword for all of the products. Sort the keywords in descending order of frequency and filter out the top 10,000 keywords.
3. Validation: Run the above 2 steps for a client with already existing keywords. Find all the common keywords between the generated keywords and the ones that already exist. Also, compare the top 10,000 keywords with keywords having the greatest search volume and find the common keywords. Compare the results to obtain those keywords that are relevant and have a higher search volume. If a majority of the keywords generated are included in this result it can be concluded that the above outlined method can be used to replace the current database. If the common keywords form a considerable percentage of generated keywords but not a majority then it can be concluded that this method should be implemented only while launching new clients with no keyword database and should not be used for existing clients.

A snippet of code used for keyword generation



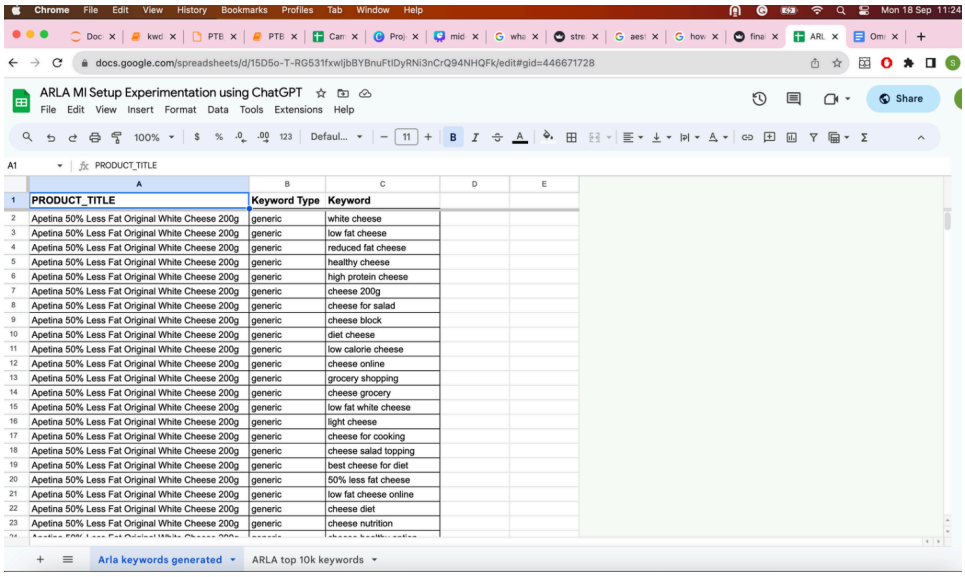
Results:



WahooFitness-CA MI Keywords for scraping

docs.google.com/spreadsheets/d/1yYgW_AjzYp_HTjUUNCM-0DMR08J5QLcgjEJsKR5FvV8/edit#gid=0

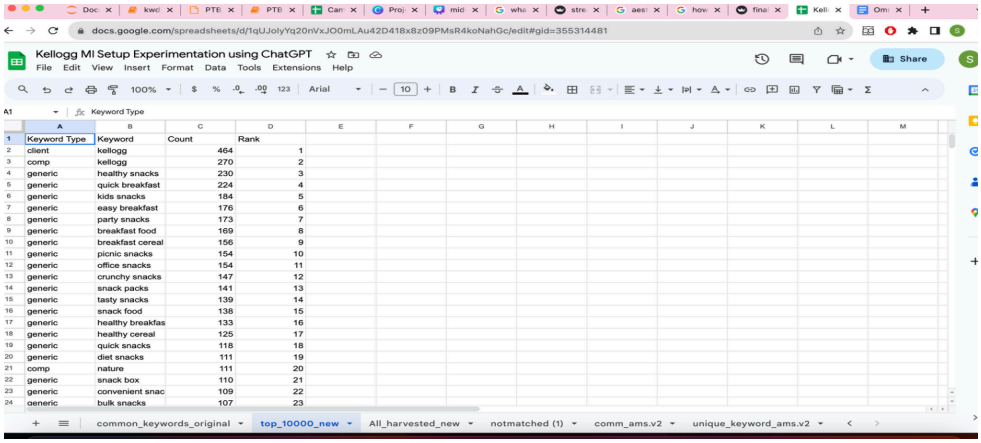
KEYWORD	FREQUENCYINADAY	MAXPAGES	URLPARAMS	DAYOFWEEK	CRAWLPRIORITY	RANKMETHOD	ENTITYSOURCE	ISDEEPCRAWL	ISACTIVE
wahoo fitness tracker	1	1		0	1	random	DS_SOV_KEYWORDS_MINING	FALSE	TRUE
exercise equipment	1	1		0	1	random	DS_SOV_KEYWORDS_MINING	FALSE	TRUE
cycling equipment	1	1		0	1	random	DS_SOV_KEYWORDS_MINING	FALSE	TRUE
cycling gear	1	1		0	1	random	DS_SOV_KEYWORDS_MINING	FALSE	TRUE
fitness equipment	1	1		0	1	random	DS_SOV_KEYWORDS_MINING	FALSE	TRUE
cycling accessory	1	1		0	1	random	DS_SOV_KEYWORDS_MINING	FALSE	TRUE
fitness tracker	1	1		0	1	random	DS_SOV_KEYWORDS_MINING	FALSE	TRUE
bicycle equipment	1	1		0	1	random	DS_SOV_KEYWORDS_MINING	FALSE	TRUE
bike gear	1	1		0	1	random	DS_SOV_KEYWORDS_MINING	FALSE	TRUE
cycling accessories	1	1		0	1	random	DS_SOV_KEYWORDS_MINING	FALSE	TRUE
running gear	1	1		0	1	random	DS_SOV_KEYWORDS_MINING	FALSE	TRUE
gym gear	1	1		0	1	random	DS_SOV_KEYWORDS_MINING	FALSE	TRUE
cycling computer	1	1		0	1	random	DS_SOV_KEYWORDS_MINING	FALSE	TRUE
cycle speedometer	1	1		0	1	random	DS_SOV_KEYWORDS_MINING	FALSE	TRUE
cycling monitor	1	1		0	1	random	DS_SOV_KEYWORDS_MINING	FALSE	TRUE
cycling fitness tracker	1	1		0	1	random	DS_SOV_KEYWORDS_MINING	FALSE	TRUE
bicycle gadget	1	1		0	1	random	DS_SOV_KEYWORDS_MINING	FALSE	TRUE
bicycle computer	1	1		0	1	random	DS_SOV_KEYWORDS_MINING	FALSE	TRUE
workout accessory	1	1		0	1	random	DS_SOV_KEYWORDS_MINING	FALSE	TRUE
heart rate monitor	1	1		0	1	random	DS_SOV_KEYWORDS_MINING	FALSE	TRUE



ARLA MI Setup Experimentation using ChatGPT

docs.google.com/spreadsheets/d/15D5o-T-RG531fxwIjbBYBnuFtIDyRNi3nCrQ94NHQFk/edit#gid=446671728

PRODUCT_TITLE	Keyword Type	Keyword
Apetina 50% Less Fat Original White Cheese 200g	generic	white cheese
Apetina 50% Less Fat Original White Cheese 200g	generic	low fat cheese
Apetina 50% Less Fat Original White Cheese 200g	generic	reduced fat cheese
Apetina 50% Less Fat Original White Cheese 200g	generic	healthy cheese
Apetina 50% Less Fat Original White Cheese 200g	generic	high protein cheese
Apetina 50% Less Fat Original White Cheese 200g	generic	cheese 200g
Apetina 50% Less Fat Original White Cheese 200g	generic	cheese for salad
Apetina 50% Less Fat Original White Cheese 200g	generic	cheese block
Apetina 50% Less Fat Original White Cheese 200g	generic	diet cheese
Apetina 50% Less Fat Original White Cheese 200g	generic	low calorie cheese
Apetina 50% Less Fat Original White Cheese 200g	generic	cheese online
Apetina 50% Less Fat Original White Cheese 200g	generic	grocery shopping
Apetina 50% Less Fat Original White Cheese 200g	generic	cheese grocery
Apetina 50% Less Fat Original White Cheese 200g	generic	low fat white cheese
Apetina 50% Less Fat Original White Cheese 200g	generic	light cheese
Apetina 50% Less Fat Original White Cheese 200g	generic	cheese for cooking
Apetina 50% Less Fat Original White Cheese 200g	generic	cheese salad topping
Apetina 50% Less Fat Original White Cheese 200g	generic	best cheese for diet
Apetina 50% Less Fat Original White Cheese 200g	generic	50% less fat cheese
Apetina 50% Less Fat Original White Cheese 200g	generic	low fat cheese online
Apetina 50% Less Fat Original White Cheese 200g	generic	cheese diet
Apetina 50% Less Fat Original White Cheese 200g	generic	cheese nutrition

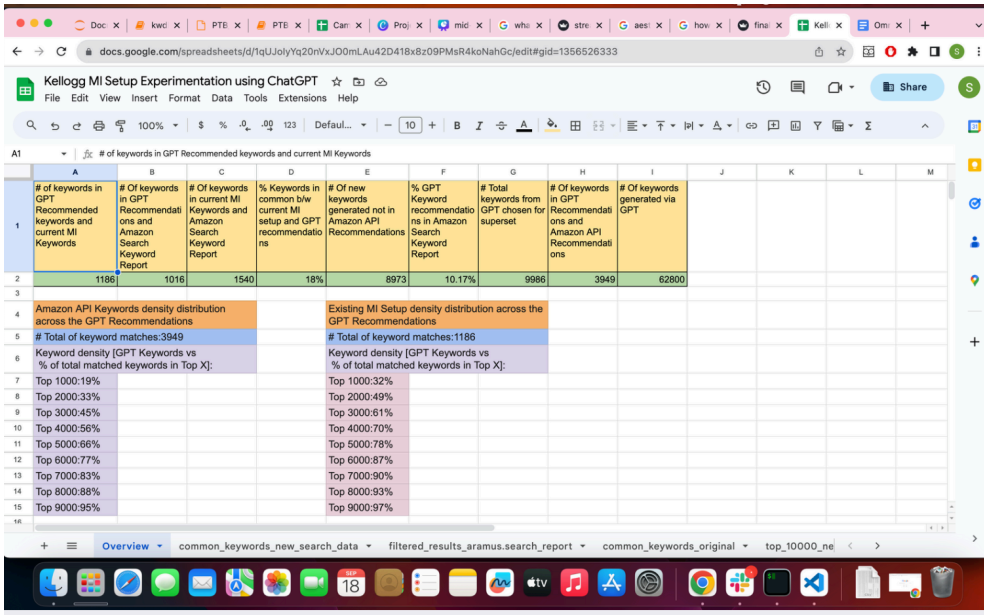


Kellogg MI Setup Experimentation using ChatGPT

docs.google.com/spreadsheets/d/1qUJolyYq20nVxJO0mLAu42D418x8z09PMsR4koNahGc/edit#gid=355314481

Keyword Type	Keyword	Count	Rank
client	kellogg	464	1
comp	kellogg	270	2
generic	healthy snacks	230	3
generic	quick breakfast	224	4
generic	kids snacks	184	5
generic	easy breakfast	176	6
generic	party snacks	173	7
generic	breakfast food	169	8
generic	breakfast cereal	156	9
generic	picnic snacks	154	10
generic	office snacks	154	11
generic	crunchy snacks	147	12
generic	snack packs	141	13
generic	tasty snacks	139	14
generic	snack food	138	15
generic	healthy breakfas	133	16
generic	healthy cereal	125	17
generic	quick snacks	118	18
generic	diet snacks	111	19
comp	nature	111	20
generic	snack box	110	21
generic	convenient snac	109	22
generic	bulk snacks	107	23

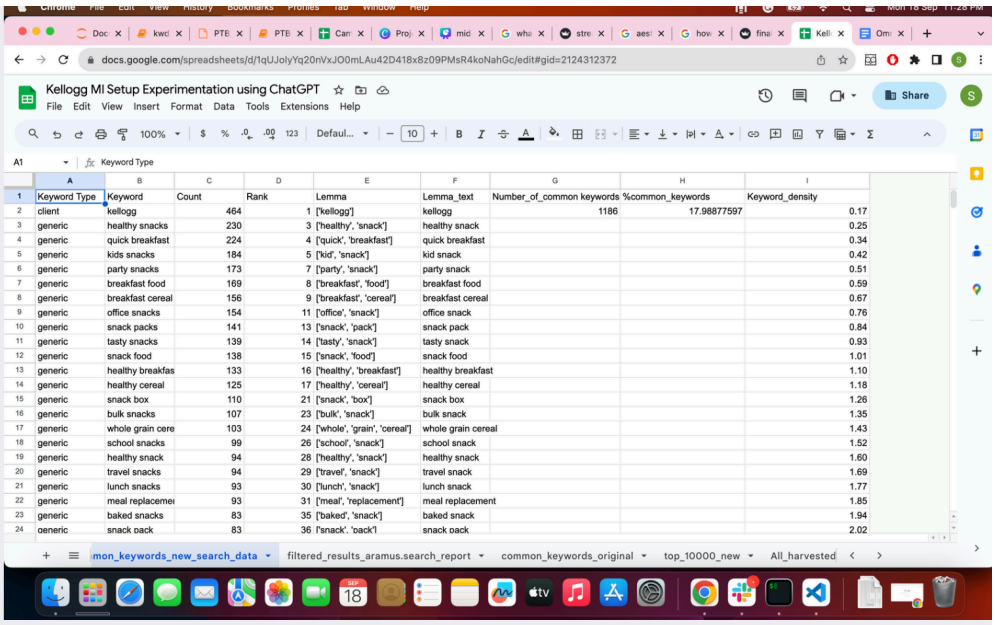
Validation



Kellogg MI Setup Experimentation using ChatGPT

# of keywords in GPT Recommended keywords and current MI Keywords	# Of keywords in GPT Recommendations and Amazon Search Keyword Report	# Of keywords in current MI Keywords and Amazon Search Keyword Report	% Keywords in common b/w current MI setup and GPT recommendations	# Of new keywords generated not in Amazon API Recommendations	% GPT Keyword recommendations in Amazon Search Keyword Report	# Total keywords from GPT chosen for superset	# Of keywords in GPT Recommendations and Amazon API Recommendations	# Of keywords generated via GPT
1186	1016	1540	18%	8973	10.17%	9986	3949	62800

Amazon API Keywords density distribution across the GPT Recommendations	Existing MI Setup density distribution across the GPT Recommendations
# Total of keyword matches:3949	# Total of keyword matches:1186
Keyword density [GPT Keywords vs % of total matched keywords in Top X]:	Keyword density [GPT Keywords vs % of total matched keywords in Top X]:
Top 1000:19%	Top 1000:32%
Top 2000:33%	Top 2000:49%
Top 3000:45%	Top 3000:61%
Top 4000:56%	Top 4000:70%
Top 5000:66%	Top 5000:78%
Top 6000:77%	Top 6000:87%
Top 7000:83%	Top 7000:90%
Top 8000:88%	Top 8000:93%
Top 9000:95%	Top 9000:97%



Kellogg MI Setup Experimentation using ChatGPT

Keyword Type	Keyword	Count	Rank	Lemma	Lemma_text	Number_of_common keywords	%common_keywords	Keyword_density
client	kellogg	464	1	['kellogg']	kellogg	1186	17.98877597	0.17
generic	healthy snacks	230	3	['healthy', 'snack']	healthy snack			0.25
generic	quick breakfast	224	4	['quick', 'breakfast']	quick breakfast			0.34
generic	kids snacks	184	5	['kid', 'snack']	kid snack			0.42
generic	party snacks	173	7	['party', 'snack']	party snack			0.51
generic	breakfast food	169	8	['breakfast', 'food']	breakfast food			0.59
generic	breakfast cereal	156	9	['breakfast', 'cereal']	breakfast cereal			0.67
generic	office snacks	154	11	['office', 'snack']	office snack			0.76
generic	snack packs	141	13	['snack', 'pack']	snack pack			0.84
generic	tasty snacks	139	14	['tasty', 'snack']	tasty snack			0.93
generic	snack food	138	15	['snack', 'food']	snack food			1.01
generic	healthy breakfas	133	16	['healthy', 'breakfast']	healthy breakfast			1.10
generic	healthy cereal	125	17	['healthy', 'cereal']	healthy cereal			1.18
generic	snack box	110	21	['snack', 'box']	snack box			1.26
generic	bulk snacks	107	23	['bulk', 'snack']	bulk snack			1.35
generic	whole grain cere	103	24	['whole', 'grain', 'cereal']	whole grain cereal			1.43
generic	school snacks	99	26	['school', 'snack']	school snack			1.52
generic	healthy snack	94	28	['healthy', 'snack']	healthy snack			1.60
generic	travel snacks	94	29	['travel', 'snack']	travel snack			1.69
generic	lunch snacks	93	30	['lunch', 'snack']	lunch snack			1.77
generic	meal replacemer	93	31	['meal', 'replacement']	meal replacement			1.85
generic	baked snacks	83	35	['baked', 'snack']	baked snack			1.94
generic	snack pack	83	36	['snack', 'pack']	snack pack			2.02

6. Project 5: Anomaly detection

6.1. Problem statement:

Identify outliers for the keywords generated using different anomaly detection algorithms and check the efficacy of each algorithm by verifying whether the anomalies obtained are truly anomalies. Also, identify outliers for numerical data in a metric called per cent time in budget present in the budget optimizer.

6.2. Motivation:

1. Learn about different anomaly detection algorithms
2. Learn how to apply anomaly detection algorithms for text data
3. Learn how to automate the validation process

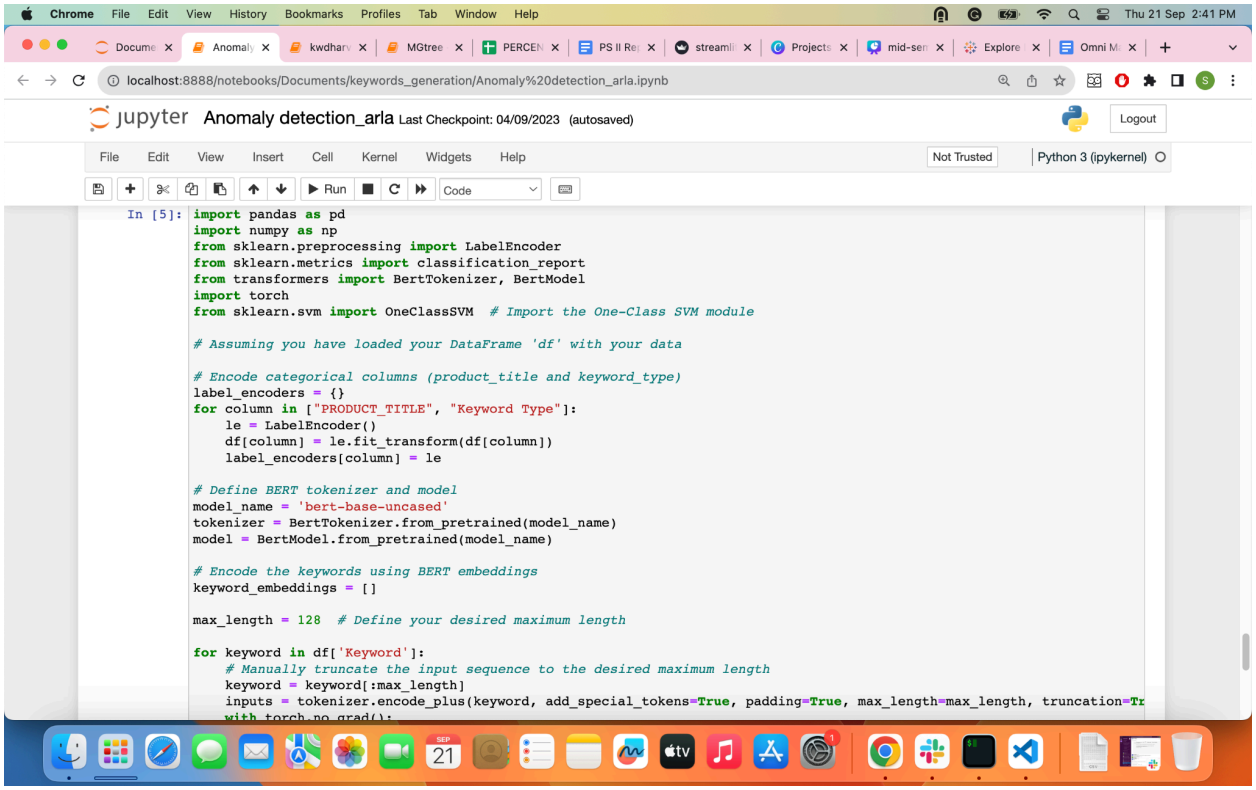
6.3. Objectives:

1. Identify the efficacy of keywords generated
2. Identify irregularities in the budget by analyzing the percentage of time in the budget
3. Help identify better algorithms to filter out invalid data

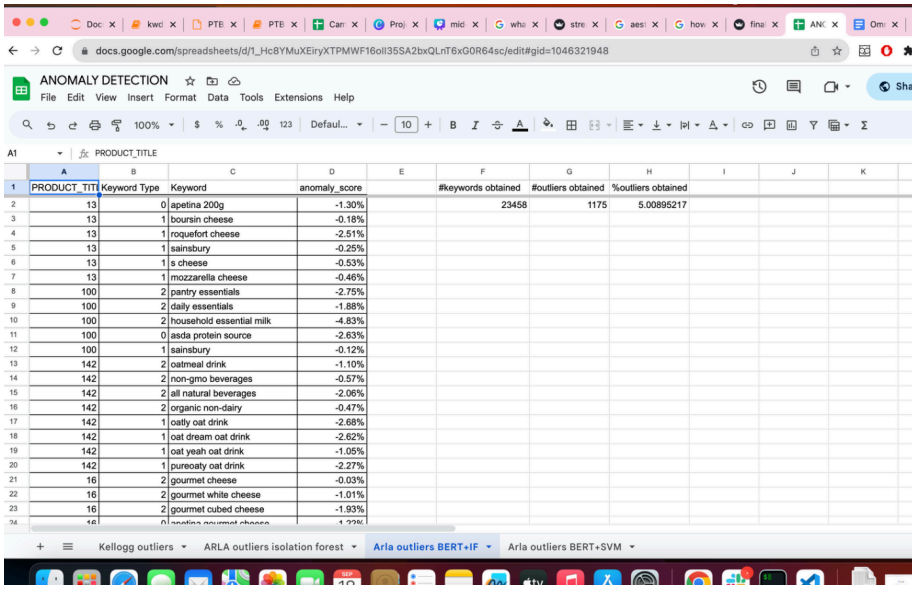
6.4. Solution:

To check the goodness of the data scraped as well as its relevance anomaly detection technique was applied first by using isolation forest algorithm, BERT encoding+ isolation forest and BERT encoding+ support vector machine. The obtained anomalies are then manually checked for relevance and based on the results it was concluded that better anomaly detection techniques need to be obtained.

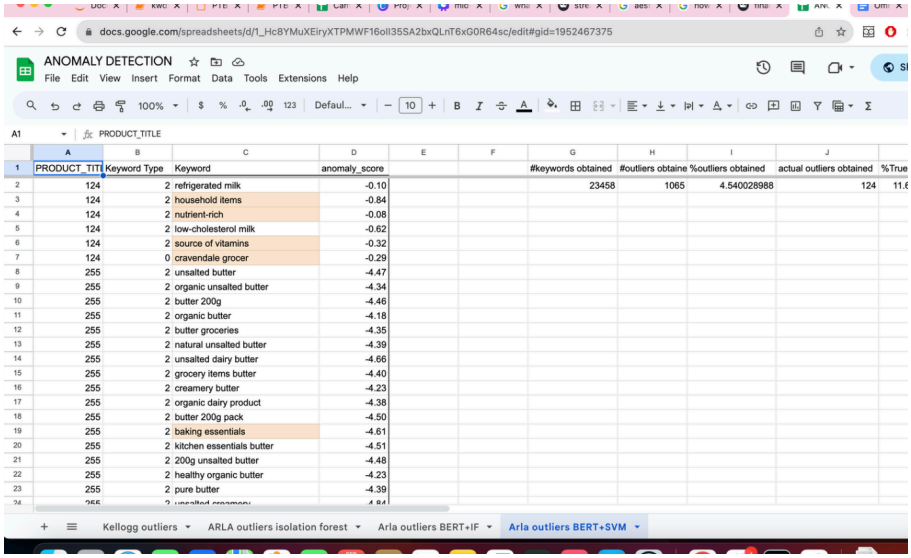
Code snippet



Results:



PRODUCT_TITL	Keyword Type	Keyword	anomaly_score	#keywords obtained	#outliers obtained	%outliers obtained
13	0	apetina 200g	-1.30%	23458	1175	5.00895217
13	1	boursin cheese	-0.18%			
13	1	roquefort cheese	-2.51%			
13	1	sainsbury	-0.25%			
13	1	s cheese	-0.53%			
13	1	mozzarella cheese	-0.46%			
100	2	pantry essentials	-2.75%			
100	2	daily essentials	-1.88%			
100	2	household essential milk	-4.83%			
100	0	asda protein source	-2.63%			
100	1	sainsbury	-0.12%			
142	2	oatmeal drink	-1.10%			
142	2	non-gmo beverages	-0.57%			
142	2	all natural beverages	-2.06%			
142	2	organic non-dairy	-0.47%			
142	1	oatly oat drink	-2.68%			
142	1	oat dream oat drink	-2.62%			
142	1	oat yeah oat drink	-1.05%			
142	1	pureoaty oat drink	-2.27%			
16	2	gourmet cheese	-0.03%			
16	2	gourmet white cheese	-1.01%			
16	2	gourmet cubed cheese	-1.93%			



7. Project 6: Budget Optimizer

7.1. Problem statement:

Analyze the per cent time in the budget metric present in budget optimizer data and calculate the percentiles present in that. Identify those campaigns with a value<0.2 and classify them as anomalies.

7.2. Motivation:

1. Study the various functions in budget optimizer
2. Understand how the budget is allocated between different advertising levels and use cases
3. Obtain real-time data using Mongo
4. Understand automatic input to Google Sheets

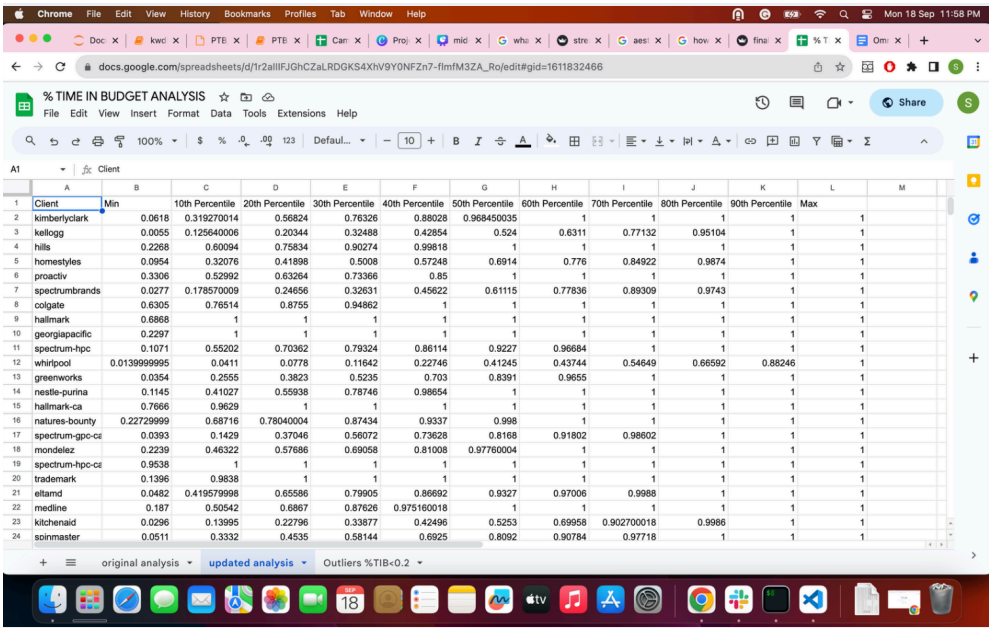
7.3. Objectives:

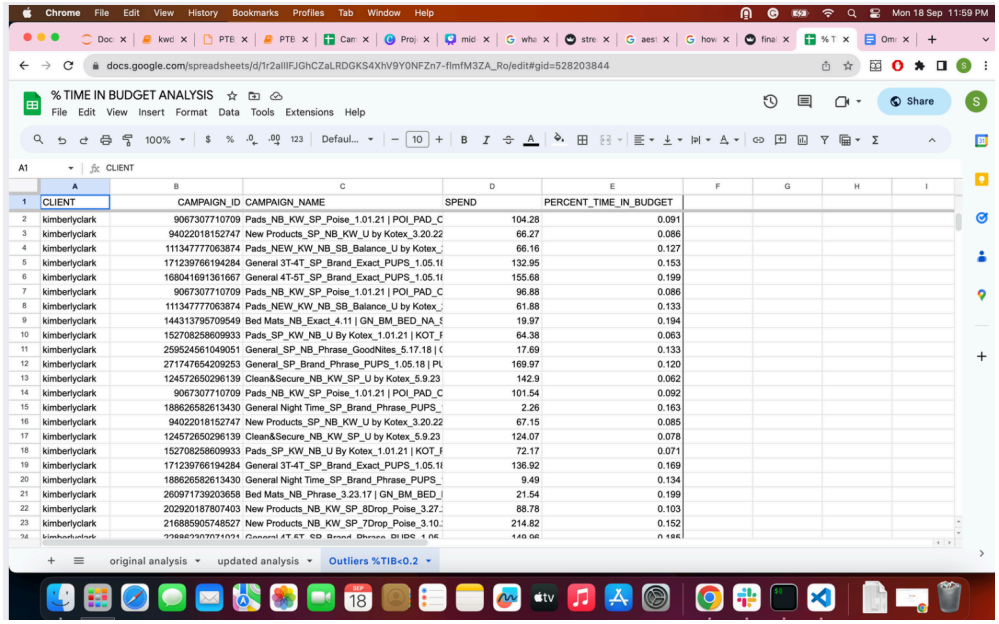
Budget optimizer is one of the services provided by Commerce-iq that programmatically allocates budget to different campaigns for greatest utilization and output.

For budget optimizer, percent_time_in_budget is an important metric that denotes the amount of time for which the budget lasted throughout the day. For B.O. analysis of the P.T.B was done to detect gaps in the algorithm and detect anomalies.

7.4. Solution:

Real-time data was obtained using Mongo and then the per cent time in budget metric was filtered out. For every unique client and campaign, the metric was sorted by per cent time in budget and the minimum, 10th percentile to max was calculated. The campaigns with <0.2 value were filtered out and the corresponding budget was noted. If the budget is considerable yet the utilization time is low, then the campaign is classified as an anomaly.





	A	B	C	D	E	F	G	H	I
1	CLIENT	CAMPAIGN_ID	CAMPAIGN_NAME	SPEND	PERCENT_TIME_IN_BUDGET				
2	kimberlyclark	9067307710709	Pads_NB_KW_SP_Poise_1.01.21 | POI_PAD_C	104.28	0.091				
3	kimberlyclark	94022018152747	New Products_SP_NB_KW_U by Kotex_3.20.22	66.27	0.086				
4	kimberlyclark	111347777063874	Pads_NEW_KW_NB_SB_Balance_U by Kotex_	66.16	0.127				
5	kimberlyclark	171239766194284	General 3T-4T_SP_Brand_Exact_PUPS_1.05.1	132.95	0.153				
6	kimberlyclark	168041691361667	General 4T-5T_SP_Brand_Exact_PUPS_1.05.1	155.68	0.199				
7	kimberlyclark	9067307710709	Pads_NB_KW_SP_Poise_1.01.21 | POI_PAD_C	96.88	0.086				
8	kimberlyclark	111347777063874	Pads_NEW_KW_NB_SB_Balance_U by Kotex_	61.88	0.133				
9	kimberlyclark	144313795709549	Bed Mats_NB_Exact_4.11 | GN_BM_BED_NA_	19.97	0.194				
10	kimberlyclark	152708258609933	Pads_SP_KW_NB_U By Kotex_1.01.21 | KOT_F	64.38	0.063				
11	kimberlyclark	259524561049051	General_SP_NB_Phrase_GoodNites_5.17.18 | C	17.69	0.133				
12	kimberlyclark	271747654209253	General_SP_Brand_Phrase_PUPS_1.05.18 | PU	169.97	0.120				
13	kimberlyclark	124572650296139	Clean&Secure_NB_KW_SP_U by Kotex_5.9.23	142.9	0.062				
14	kimberlyclark	9067307710709	Pads_NB_KW_SP_Poise_1.01.21 | POI_PAD_C	101.54	0.092				
15	kimberlyclark	188626582613430	General Night Time_SP_Brand_Phrase_PUPS_	2.26	0.163				
16	kimberlyclark	94022018152747	New Products_SP_NB_KW_U by Kotex_3.20.22	67.15	0.085				
17	kimberlyclark	124572650296139	Clean&Secure_NB_KW_SP_U by Kotex_5.9.23	124.07	0.078				
18	kimberlyclark	152708258609933	Pads_SP_KW_NB_U By Kotex_1.01.21 | KOT_F	72.17	0.071				
19	kimberlyclark	171239766194284	General 3T-4T_SP_Brand_Exact_PUPS_1.05.1	136.92	0.169				
20	kimberlyclark	188626582613430	General Night Time_SP_Brand_Phrase_PUPS_	9.49	0.134				
21	kimberlyclark	260971739203658	Bed Mats_NB_Phrase_3.23.17 | GN_BM_BED_I	21.54	0.199				
22	kimberlyclark	202920187807403	New Products_NB_KW_SP_8Drop_Poise_3.27.2	88.78	0.103				
23	kimberlyclark	216885905748527	New Products_NB_KW_SP_7Drop_Poise_3.10.2	214.82	0.152				

8. Project 7: Campaign potential analysis

8.1. Problem statement:

Calculate the difference between the campaign potential obtained using old and new algorithms from historical data present in the databases and compare both the results with Amazon data to verify if the error is reduced. Analyse the error obtained to find a formula that minimises the error.

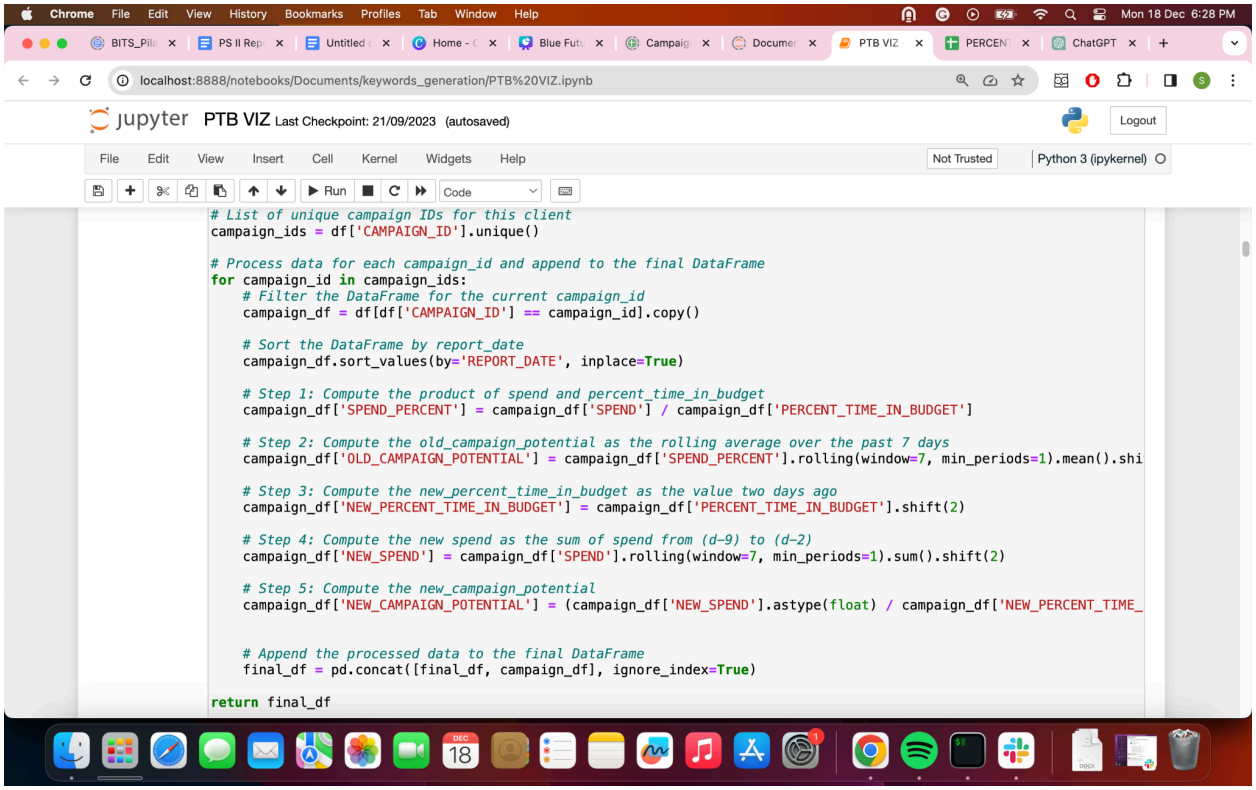
8.2. Motivation:

1. Understand the Amazon method of campaign potential calculation
2. Verify the efficacy of the new method of potential calculation for bid optimization
3. Find a trend in error to identify the reason for anomalies

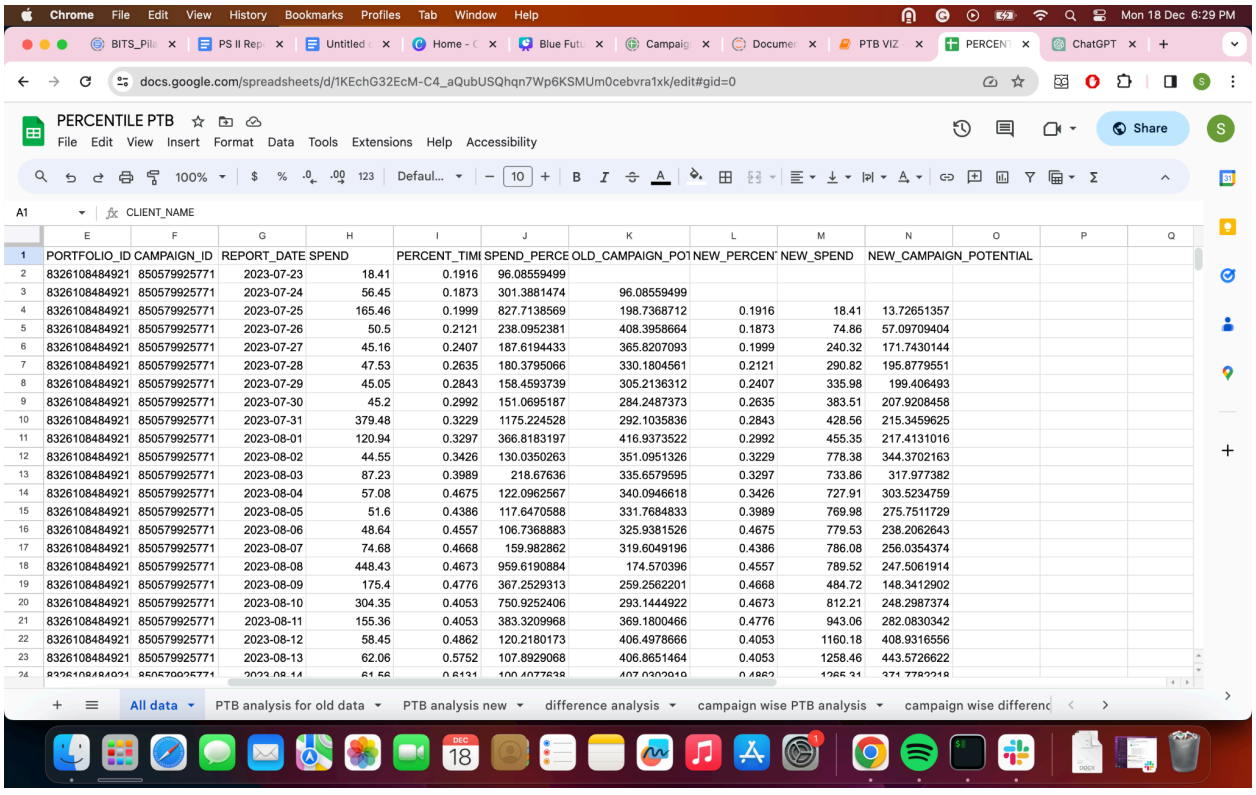
8.3. Solution:

Obtain data from Snowflake across multiple clients. Calculate the old potential by calculating the average spend and dividing it by the average per cent time in budget rolling over the past 7 days. The new potential is calculated by summing the spend and dividing it by the per cent time in the budget of t-2, which is assumed to be the rolling average of t-2 to t-9 and then dividing the overall result by 7. This process is run for each date for each campaign for all clients. The difference between the old potential, new potential and potential obtained from Amazon is obtained and these errors are plotted against their respective spend and budget. The trend is observed to discern where the error is less. The trend is analysed to find reasons for anomalies and obtain a correlation between the parameters involved to formulate an equation that denotes closely the actual calculation pattern being followed in Amazon. A percentile-wise analysis is also done and analysed.

Code Snippet:

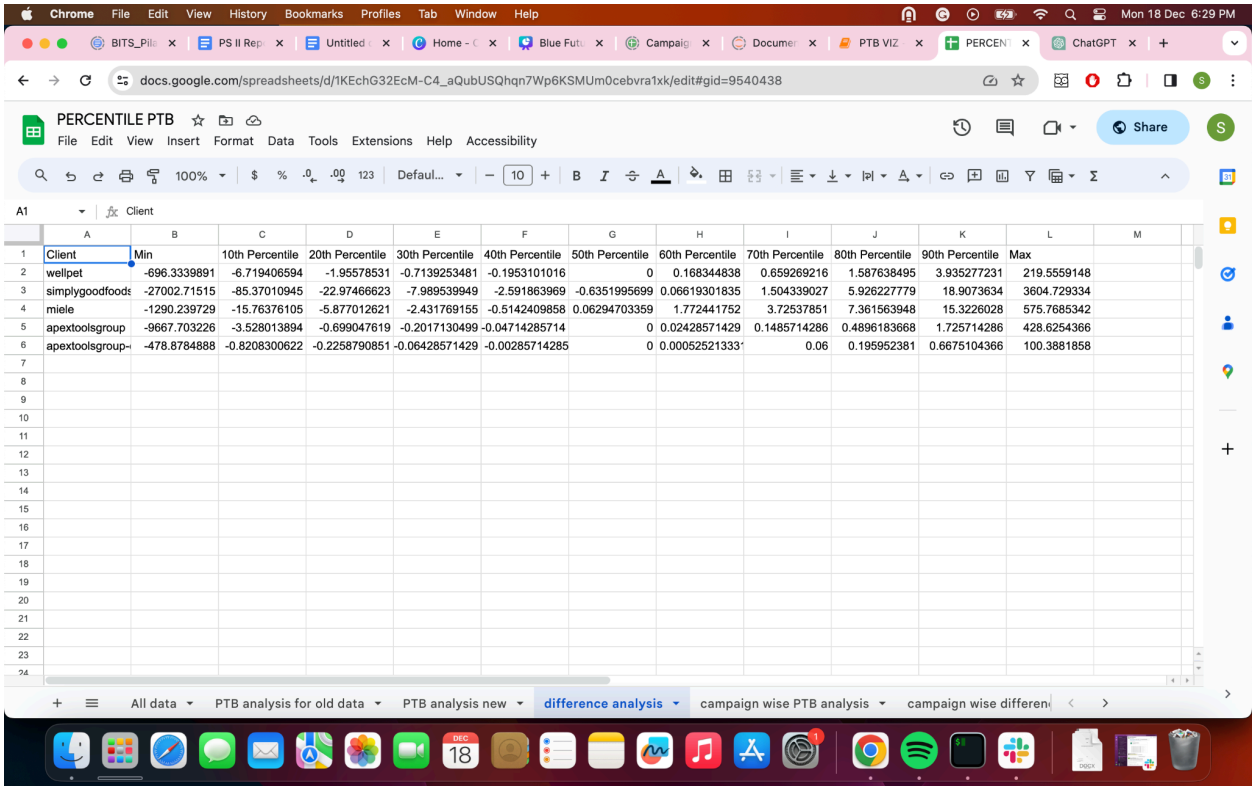


```
# List of unique campaign IDs for this client
campaign_ids = df['CAMPAIGN_ID'].unique()

# Process data for each campaign_id and append to the final DataFrame
for campaign_id in campaign_ids:
    # Filter the DataFrame for the current campaign_id
    campaign_df = df[df['CAMPAIGN_ID'] == campaign_id].copy()

    # Sort the DataFrame by report_date
    campaign_df.sort_values(by='REPORT_DATE', inplace=True)

    # Step 1: Compute the product of spend and percent_time_in_budget
    campaign_df['SPEND_PERCENT'] = campaign_df['SPEND'] / campaign_df['PERCENT_TIME_IN_BUDGET']

    # Step 2: Compute the old_campaign_potential as the rolling average over the past 7 days
    campaign_df['OLD_CAMPAIGN_POTENTIAL'] = campaign_df['SPEND_PERCENT'].rolling(window=7, min_periods=1).mean().shi

    # Step 3: Compute the new_percent_time_in_budget as the value two days ago
    campaign_df['NEW_PERCENT_TIME_IN_BUDGET'] = campaign_df['PERCENT_TIME_IN_BUDGET'].shift(2)

    # Step 4: Compute the new spend as the sum of spend from (d-9) to (d-2)
    campaign_df['NEW_SPEND'] = campaign_df['SPEND'].rolling(window=7, min_periods=1).sum().shift(2)

    # Step 5: Compute the new_campaign_potential
    campaign_df['NEW_CAMPAIGN_POTENTIAL'] = (campaign_df['NEW_SPEND'].astype(float) / campaign_df['NEW_PERCENT_TIME_

    # Append the processed data to the final DataFrame
    final_df = pd.concat([final_df, campaign_df], ignore_index=True)

return final_df
```



	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	PORTFOLIO_ID	CAMPAIGN_ID	REPORT_DATE	SPEND	PERCENT_TIMI	SPEND_PERCE	OLD_CAMPAIGN_POT	NEW_PERCEN	NEW_SPEND	NEW_CAMPAIGN_POTENTIAL			
2	8326108484921	850579925771	2023-07-23	18.41	0.1916	96.08559499							
3	8326108484921	850579925771	2023-07-24	56.45	0.1873	301.3881474	96.08559499						
4	8326108484921	850579925771	2023-07-25	165.46	0.1999	827.7138569	198.7368712	0.1916	18.41	13.72651357			
5	8326108484921	850579925771	2023-07-26	50.5	0.2121	238.0952381	408.3958664	0.1873	74.86	57.09709404			
6	8326108484921	850579925771	2023-07-27	45.16	0.2407	187.6194433	365.8207093	0.1999	240.32	171.7430144			
7	8326108484921	850579925771	2023-07-28	47.53	0.2635	180.3795066	330.1804561	0.2121	290.82	195.8779551			
8	8326108484921	850579925771	2023-07-29	45.05	0.2843	158.4593739	305.2136312	0.2407	335.98	199.406493			
9	8326108484921	850579925771	2023-07-30	45.2	0.2992	151.0695187	284.2487373	0.2635	383.51	207.9208458			
10	8326108484921	850579925771	2023-07-31	379.48	0.3229	1175.224528	292.1035836	0.2843	428.56	215.3459625			
11	8326108484921	850579925771	2023-08-01	120.94	0.3297	366.8183197	416.9373522	0.2992	455.35	217.4131016			
12	8326108484921	850579925771	2023-08-02	44.55	0.3426	130.0350263	351.0951326	0.3229	778.38	344.3702163			
13	8326108484921	850579925771	2023-08-03	87.23	0.3989	218.67636	335.6579595	0.3297	733.86	317.977382			
14	8326108484921	850579925771	2023-08-04	57.08	0.4675	122.0962567	340.0946618	0.3426	727.91	303.5234759			
15	8326108484921	850579925771	2023-08-05	51.6	0.4386	117.6470588	331.7684833	0.3989	769.98	275.7511729			
16	8326108484921	850579925771	2023-08-06	48.64	0.4557	106.7368883	325.9381526	0.4675	779.53	238.2062643			
17	8326108484921	850579925771	2023-08-07	74.68	0.4668	159.982862	319.6049196	0.4386	786.08	256.0354374			
18	8326108484921	850579925771	2023-08-08	448.43	0.4673	959.6190884	174.570396	0.4557	789.52	247.5061914			
19	8326108484921	850579925771	2023-08-09	175.4	0.4776	367.2529313	259.2562201	0.4668	484.72	148.3412902			
20	8326108484921	850579925771	2023-08-10	304.35	0.4053	750.9252406	293.1444922	0.4673	812.21	248.2987374			
21	8326108484921	850579925771	2023-08-11	155.36	0.4053	383.3209968	369.1800466	0.4776	943.06	282.0830342			
22	8326108484921	850579925771	2023-08-12	58.45	0.4862	120.2180173	406.4978666	0.4053	1160.18	408.9316556			
23	8326108484921	850579925771	2023-08-13	62.06	0.5752	107.8929068	406.8651464	0.4053	1258.46	443.5726622			
24	8326108484921	850579925771	2023-08-14	61.56	0.6131	100.4077638	407.0302919	0.4862	1265.31	371.7782218			

All data | PTB analysis for old data | PTB analysis new | difference analysis | campaign wise PTB analysis | campaign wise differenc

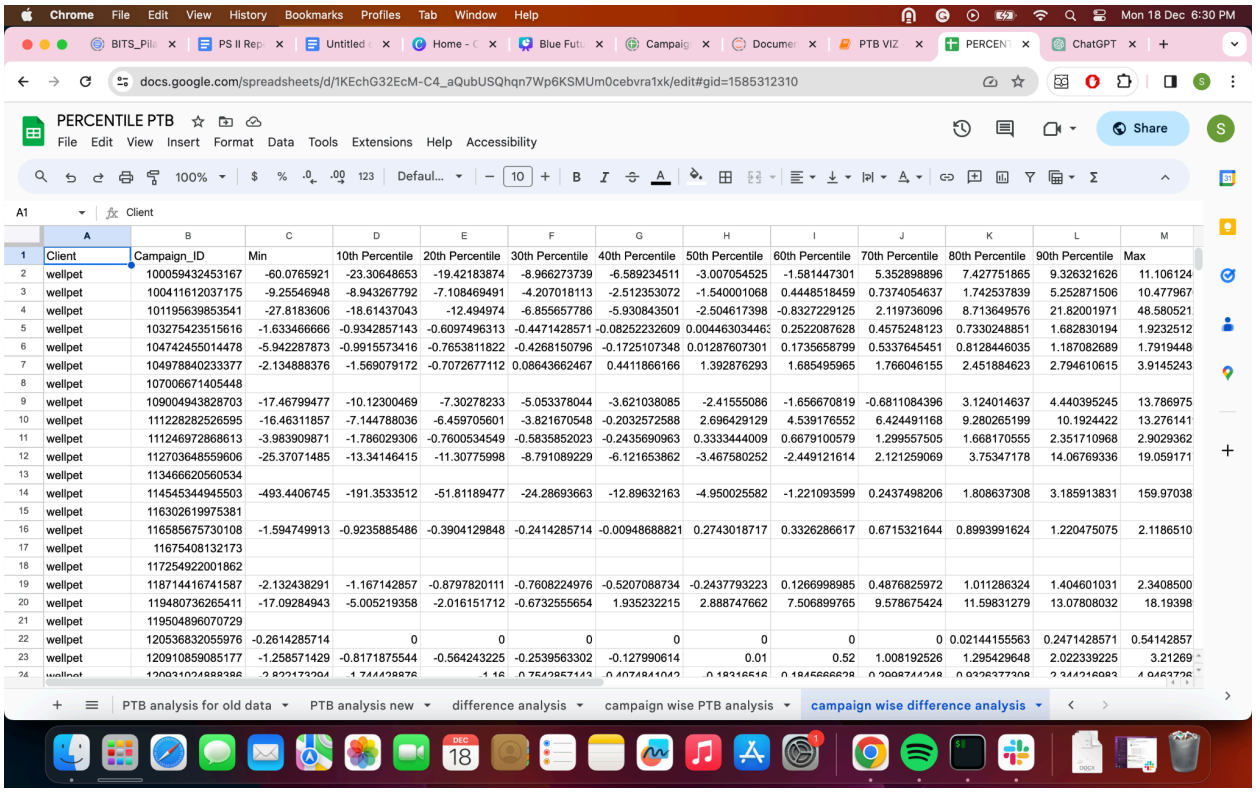


PERCENTILE PTB

docs.google.com/spreadsheets/d/1KEchG32EcM-C4_aQubUSQhqn7Wp6KSMUm0cebvra1xk/edit#gid=9540438

Client	Min	10th Percentile	20th Percentile	30th Percentile	40th Percentile	50th Percentile	60th Percentile	70th Percentile	80th Percentile	90th Percentile	Max
wellpet	-696.3339891	-6.719406594	-1.95578531	-0.7139253481	-0.1953101016	0	0.168344838	0.659269216	1.587638495	3.935277231	219.5559148
simplygoodfoods	-27002.71515	-85.37010945	-22.97466623	-7.989539949	-2.591863969	-0.6351995699	0.06619301835	1.504339027	5.926227779	18.9073634	3604.729334
miele	-1290.239729	-15.76376105	-5.877012621	-2.431769155	-0.5142409858	0.06294703359	1.772441752	3.72537851	7.361563948	15.3226028	575.7685342
apextoolsgroup	-9667.703226	-3.528013894	-0.699047619	-0.2017130499	-0.04714285714	0	0.02428571429	0.1485714286	0.4896183668	1.725714286	428.6254366
apextoolsgroup-	-478.8784888	-0.8208300622	-0.2258790851	-0.06428571429	-0.00285714285	0	0.00052521333	0.06	0.195952381	0.6675104366	100.3881858

All data | PTB analysis for old data | PTB analysis new | difference analysis | campaign wise PTB analysis | campaign wise differen



PERCENTILE PTB

docs.google.com/spreadsheets/d/1KEchG32EcM-C4_aQubUSQhqn7Wp6KSMUm0cebvra1xk/edit#gid=1585312310

Client	Campaign_ID	Min	10th Percentile	20th Percentile	30th Percentile	40th Percentile	50th Percentile	60th Percentile	70th Percentile	80th Percentile	90th Percentile	Max
wellpet	100059432453167	-60.0765921	-23.30648653	-19.42183874	-8.966273739	-6.589234511	-3.007054525	-1.581447301	5.352898896	7.427751865	9.326321626	11.106124
wellpet	100411612037175	-9.25546948	-8.943267792	-7.108469491	-4.207018113	-2.512353072	-1.540001068	0.4448518459	0.7374054637	1.742537839	5.252871506	10.477967
wellpet	101195639853541	-27.8183606	-18.61437043	-12.494974	-6.855657786	-5.930843501	-2.504617398	-0.8327229125	2.119736096	8.713649576	21.82001971	48.580521
wellpet	103275423515616	-1.633466666	-0.9342857143	-0.6097496313	-0.4471428571	-0.08252232609	0.004463034463	0.2522087628	0.4575248123	0.7330248851	1.682830194	1.9232512
wellpet	104742455014478	-5.942287873	-0.9915573416	-0.7653811822	-0.4268150796	-0.1725107348	0.01287607301	0.1735658799	0.5337645451	0.8128446035	1.187082689	1.7919448
wellpet	104978840233377	-2.134888376	-1.569079172	-0.7072677112	0.08643662467	0.4411866166	1.392876293	1.685495965	1.766046155	2.451884623	2.794610615	3.9145243
wellpet	107006671405448											
wellpet	109004943828703	-17.46799477	-10.12300469	-7.30278233	-5.053378044	-3.621038085	-2.41555086	-1.656670819	-0.6811084396	3.124014637	4.440395245	13.786975
wellpet	111228282526595	-16.46311857	-7.144788036	-6.459705601	-3.821670548	-0.2032572588	2.696429129	4.539176552	6.424491168	9.280265199	10.1924422	13.276141
wellpet	111246972868613	-3.983909871	-1.786029306	-0.7600534549	-0.5835852023	-0.2435690963	0.3333444009	0.6679100579	1.299557505	1.668170555	2.351710968	2.9029362
wellpet	112703648559606	-25.37071485	-13.34146415	-11.30775998	-8.791089229	-6.121653862	-3.467580252	-2.449121614	2.121259069	3.75347178	14.06769336	19.059171
wellpet	113466620560534											
wellpet	114545344945503	-493.4406745	-191.3533512	-51.81189477	-24.28693663	-12.89632163	-4.950025582	-1.221093599	0.2437498206	1.808637308	3.185913831	159.97038
wellpet	116302619975381											
wellpet	116585675730108	-1.594749913	-0.9235885486	-0.3904129848	-0.2414285714	-0.00948688821	0.2743018717	0.3326286617	0.6715321644	0.8993991624	1.220475075	2.1186510
wellpet	11675408132173											
wellpet	117254922001862											
wellpet	118714416741587	-2.132438291	-1.167142857	-0.8797820111	-0.7608224976	-0.5207088734	-0.2437793223	0.1266998985	0.4876825972	1.011286324	1.404601031	2.3408500
wellpet	119480736265411	-17.09284943	-5.005219358	-2.016151712	-0.6732555654	1.935232215	2.888747662	7.506899765	9.578675424	11.59831279	13.07808032	18.19398
wellpet	119504896070729											
wellpet	120536832055976	-0.2614285714	0	0	0	0	0	0	0	0.02144155563	0.2471428571	0.54142857
wellpet	120910859085177	-1.258571429	-0.8171875544	-0.564243225	-0.2539563302	-0.127990614	0.01	0.52	1.008192526	1.295429648	2.022339225	3.21269

PTB analysis for old data | PTB analysis new | difference analysis | campaign wise PTB analysis | campaign wise difference analysis

9. Project 8: Spend threshold calculation

9.1. Problem statement:

Obtain the error trend against Spend or Budget. Analyze the error to determine a threshold that could be used to determine the greatest concentration of error. This should denote a range in which anomalies can be ignored.

9.2. Motivation:

1. The primary motivation is to understand and visualize the trend of errors in relation to Spend or Budget. By obtaining and analyzing this trend, the stakeholders aim to gain insights into how errors vary over time in the context of financial expenditure. Identifying the patterns in error occurrences helps in comprehending the overall quality and reliability of the data, highlighting potential areas of concern.
2. Setting a threshold for discerning the concentration of errors that can be considered acceptable or within a normal range
3. Denote a specific range of errors that can be safely ignored as normal fluctuations.

9.3. Objectives:

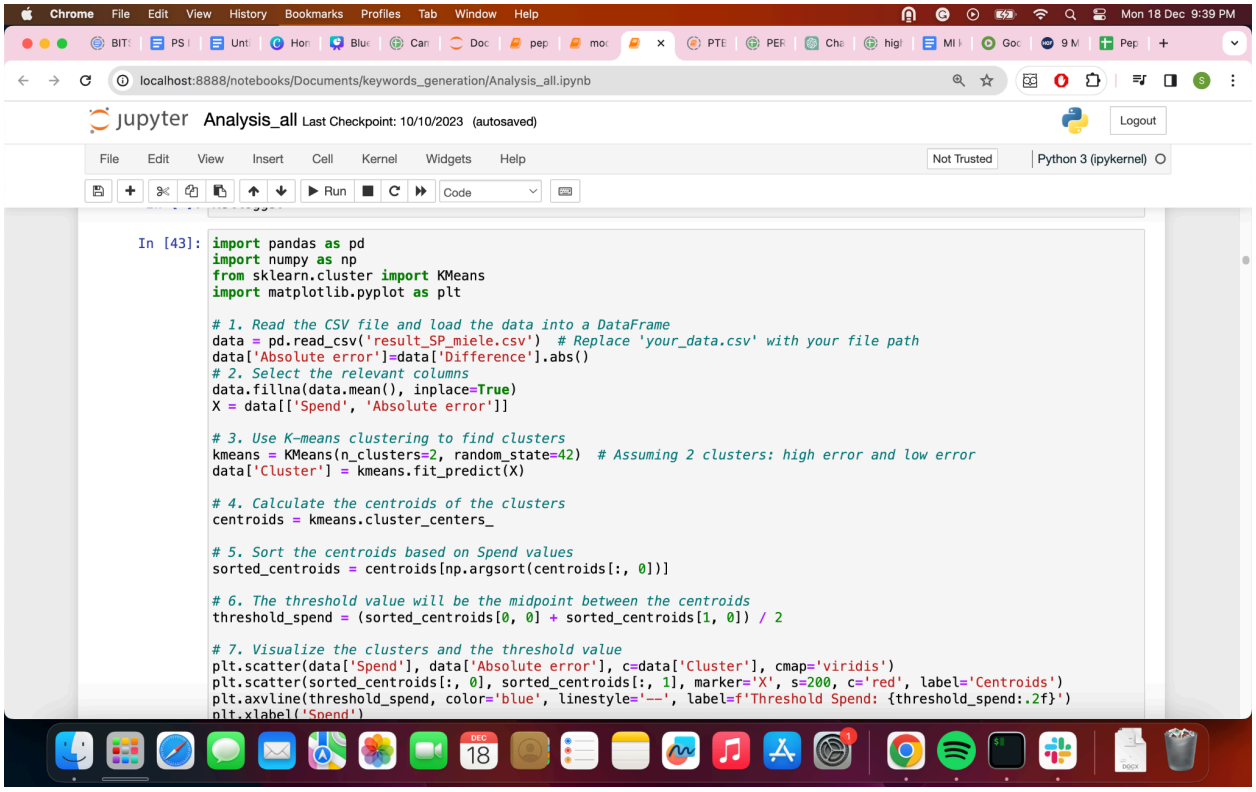
1. Obtain a comprehensive understanding of how errors in the data relate to Spend or Budget over a specified period.
2. Establish a clear, actionable threshold that aligns with the organization's tolerance for deviations in financial data.
3. Identification and exclusion of anomalies, streamlining data interpretation and allowing stakeholders to focus on addressing significant issues that fall outside the established range.

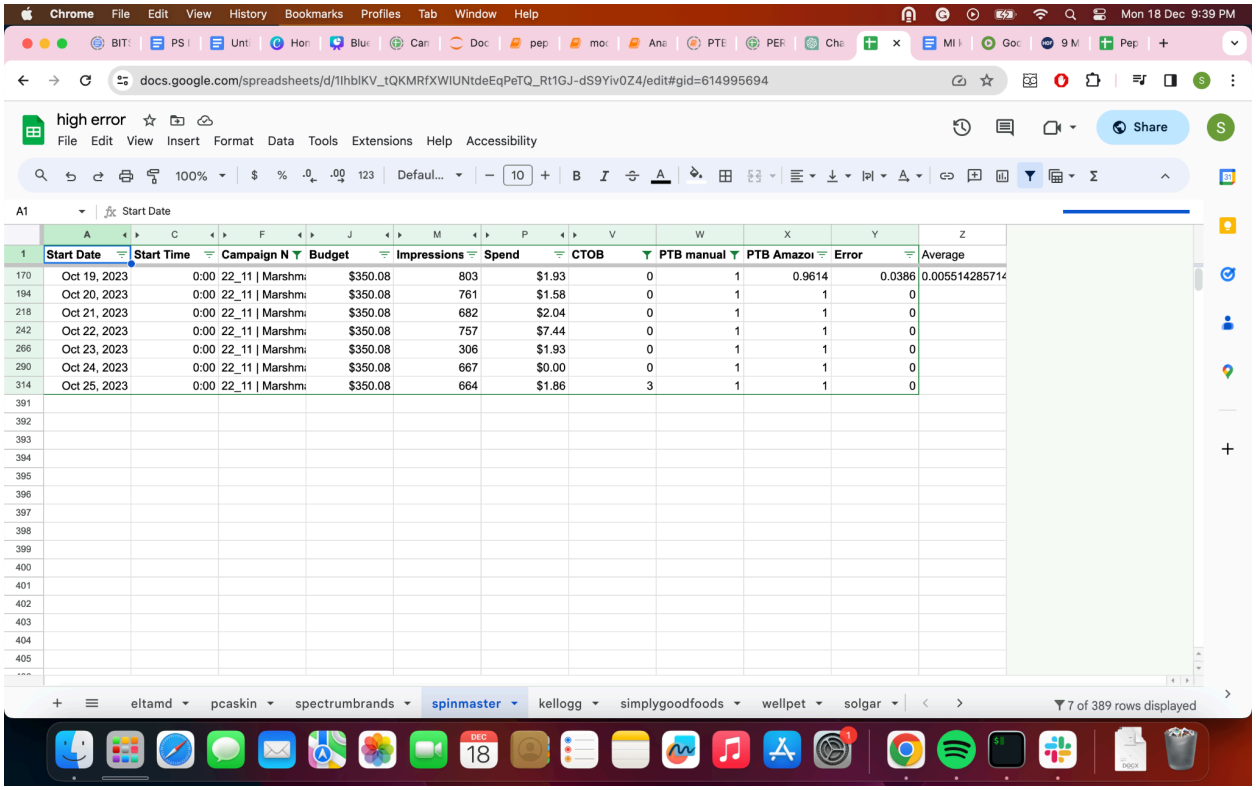
9.4. Solution:

Obtain Amazon hourly campaign report for 10 clients and use the report to calculate the cumulative time out of budget by analysing the time stamp, impression and spend for each entry. After the cumulative time out of budget for every day of every campaign is calculated for the client, obtain the per cent time in the budget by obtaining a rolling average of the cumulative time out of budget by dividing it by(24*7). After the per cent time out of the budget has been computed for all campaigns, the data is compared for the given time frame with the existing database and the absolute error is computed. The error vs spend and error vs budget graphs are computed for each client. To obtain the spend threshold 4 methods are applied:

1. K-means clustering: The absolute errors are clustered based on their deviation from their respective median and a boundary between the 2 regions is obtained as the threshold
2. DB Scan clustering: By clustering concerning error density, a boundary is achieved that denotes the threshold
3. Percentile-based classification: The errors are arranged from min to max and the mean absolute error corresponding to each percentile is computed. The data is then arranged in decreasing order of mean absolute error and the spend or budget value corresponding to the highest error is considered as the threshold.
4. Interval-based classification: The entire range of budget or spending is subdivided into intervals based on the minimum and maximum value. For each of these intervals, the mean absolute error is calculated

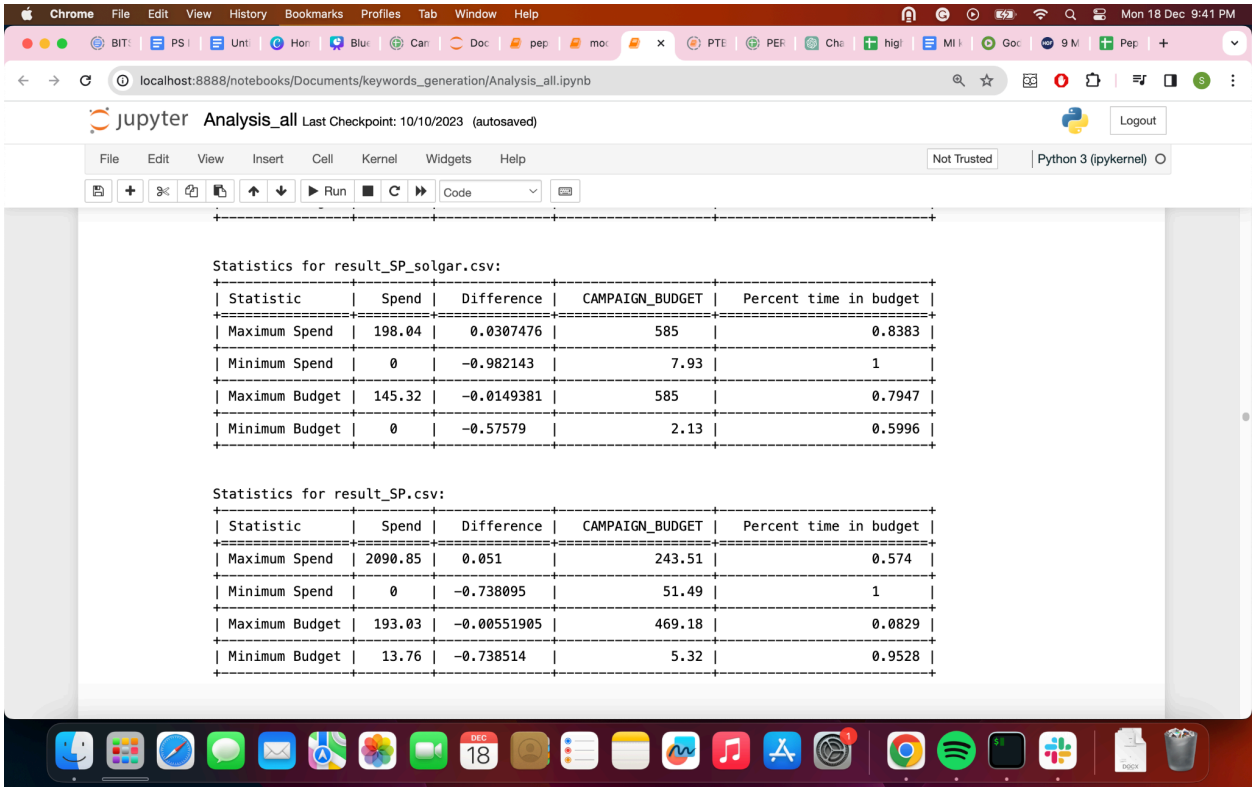
Code snippet:





high error

Start Date	Start Time	Campaign N	Budget	Impressions	Spend	CTOB	PTB manual	PTB Amazo	Error	Average
Oct 19, 2023	0:00	22_11 | Marshma	$350.08	803	$1.93	0	1	0.9614	0.0386	0.005514285714
Oct 20, 2023	0:00	22_11 | Marshma	$350.08	761	$1.58	0	1	1	0	
Oct 21, 2023	0:00	22_11 | Marshma	$350.08	682	$2.04	0	1	1	0	
Oct 22, 2023	0:00	22_11 | Marshma	$350.08	757	$7.44	0	1	1	0	
Oct 23, 2023	0:00	22_11 | Marshma	$350.08	306	$1.93	0	1	1	0	
Oct 24, 2023	0:00	22_11 | Marshma	$350.08	667	$0.00	0	1	1	0	
Oct 25, 2023	0:00	22_11 | Marshma	$350.08	664	$1.86	3	1	1	0	

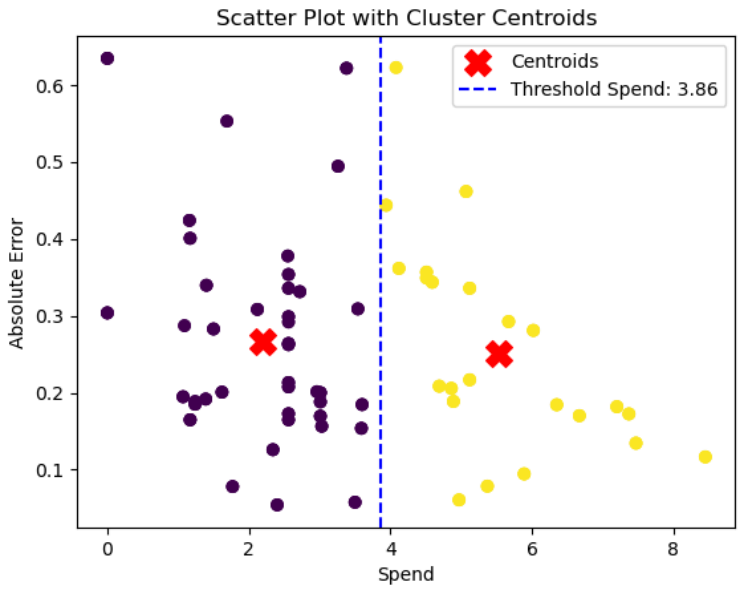


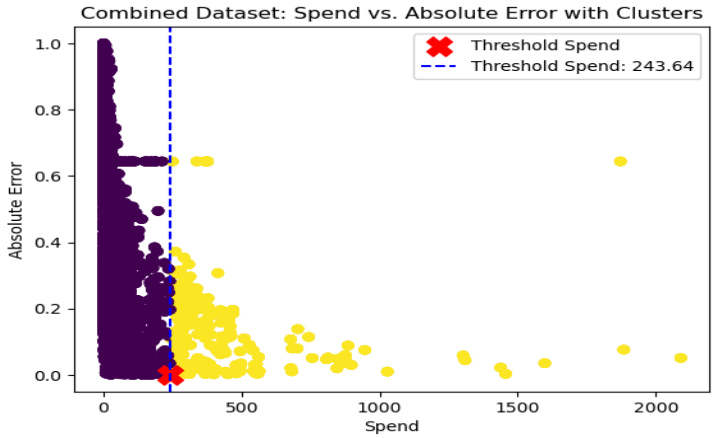
Statistics for result_SP_solgar.csv:

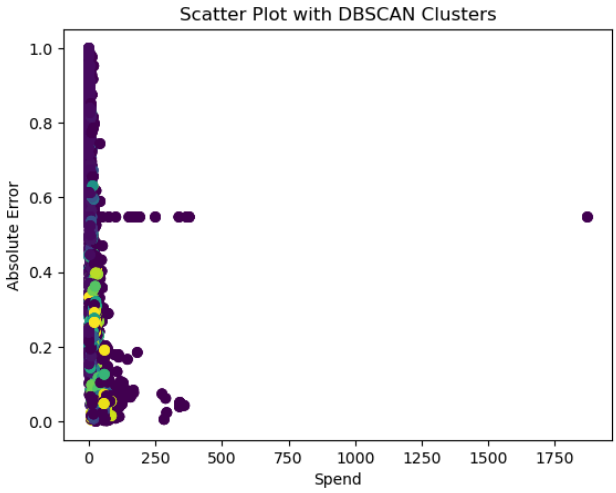
Statistic	Spend	Difference	CAMPAIGN_BUDGET	Percent time in budget
Maximum Spend	198.04	0.0307476	585	0.8383
Minimum Spend	0	-0.982143	7.93	1
Maximum Budget	145.32	-0.0149381	585	0.7947
Minimum Budget	0	-0.57579	2.13	0.5996

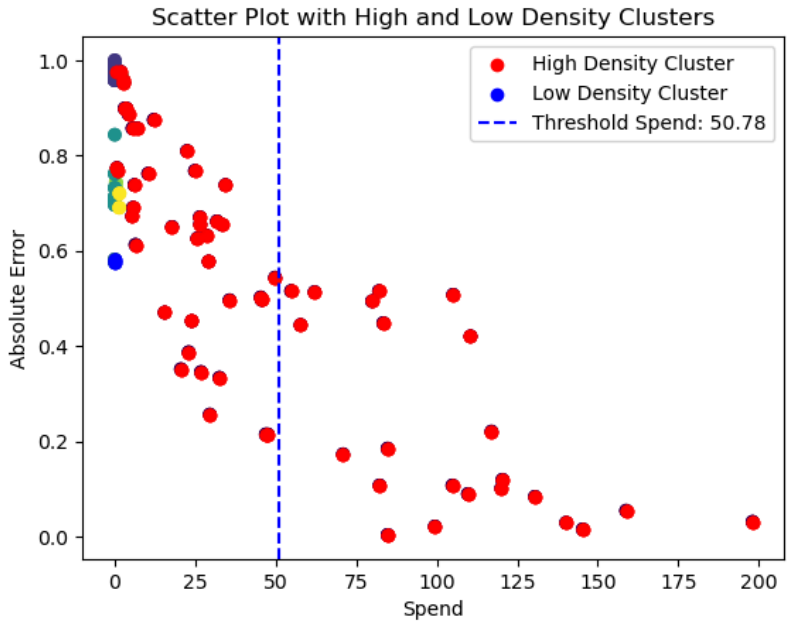
Statistics for result_SP.csv:

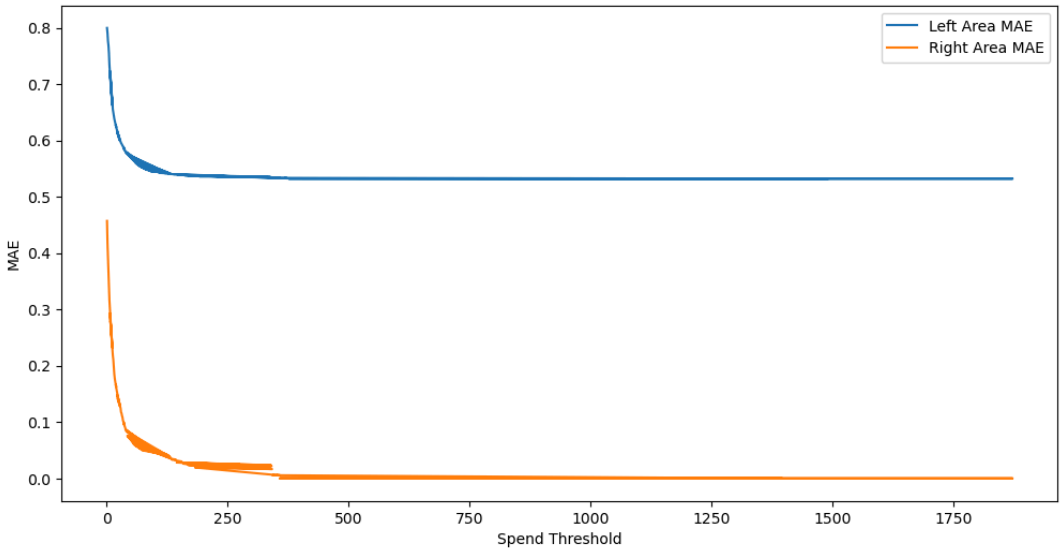
Statistic	Spend	Difference	CAMPAIGN_BUDGET	Percent time in budget
Maximum Spend	2090.85	0.051	243.51	0.574
Minimum Spend	0	-0.738095	51.49	1
Maximum Budget	193.03	-0.00551905	469.18	0.0829
Minimum Budget	13.76	-0.738514	5.32	0.9528


Scatter Plot with Cluster Centroids
Centroids
Threshold Spend: 3.86
Absolute Error
Spend
0.1
0.2
0.3
0.4
0.5
0.6
0
2
4
6
8


Combined Dataset: Spend vs. Absolute Error with Clusters
Threshold Spend
Threshold Spend: 243.64
Absolute Error
Spend
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000


Scatter Plot with DBSCAN Clusters
Absolute Error
Spend
0.0
0.2
0.4
0.6
0.8
1.0
0
250
500
750
1000
1250
1500
1750





10. Project 9: Exploratory data analysis on Experiment data

10.1. Problem statement:

Obtain reason for error in experiment data where the test day incremental fraction is greater than the control date.

10.2. Motivation:

1. Learn about the classification of data by test and control days
2. Learn how to obtain incremental fractions for each group from raw data
3. Learn how to analyse the anomalies to obtain a pattern

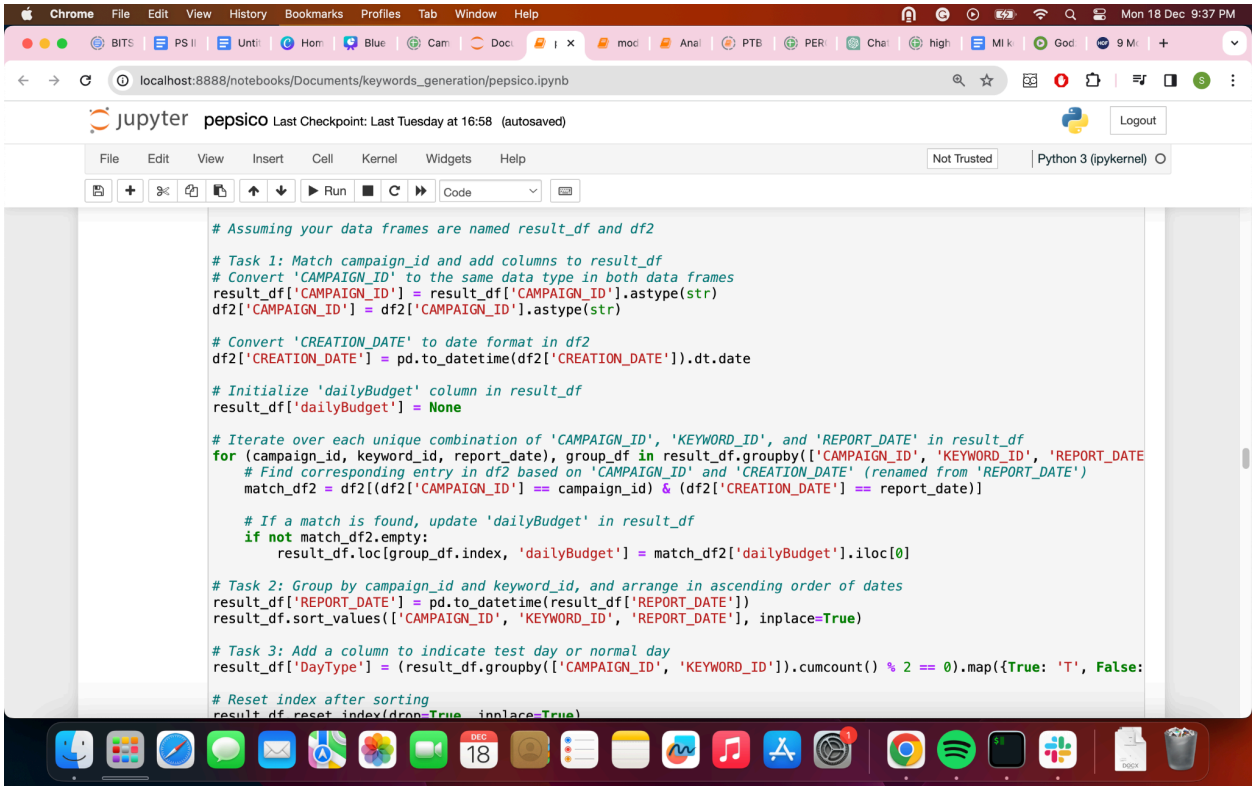
10.3. Objectives:

1. Obtain keyword-level and campaign-level data from databases and engineer parameters for the same.
2. Observe the pattern across the data to obtain a connecting parameter to denote the key findings
3. Analyse the overall anomalies and formulate an explanation for it.

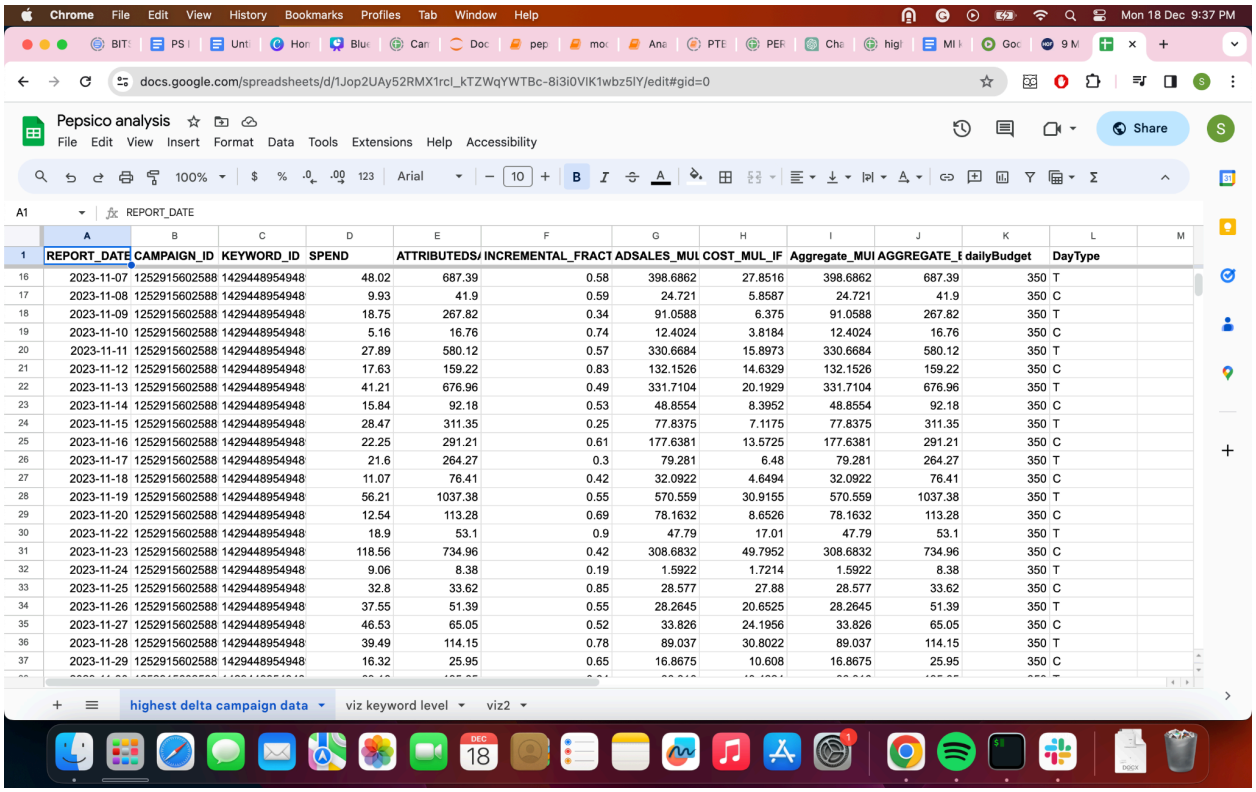
10.4. Solution:

Create a mapping that considers all report days and classifies them as either test or control data based on the start date. This process is run for all keywords present in all the campaigns of a particular client. After classification group the keywords of each campaign by test or control day and analyse the trend of attributed sales and spending of control vs test days for each campaign. Obtain the incremental fraction from attributed sales and spending and observe its trends across keywords to observe keyword-level anomalies. Obtain the campaign that has the highest difference between the control day and test day incremental fraction and observe its behaviour at the keyword level for attributed sales and spending to spot any irregularities in the pattern.

Code snippet:



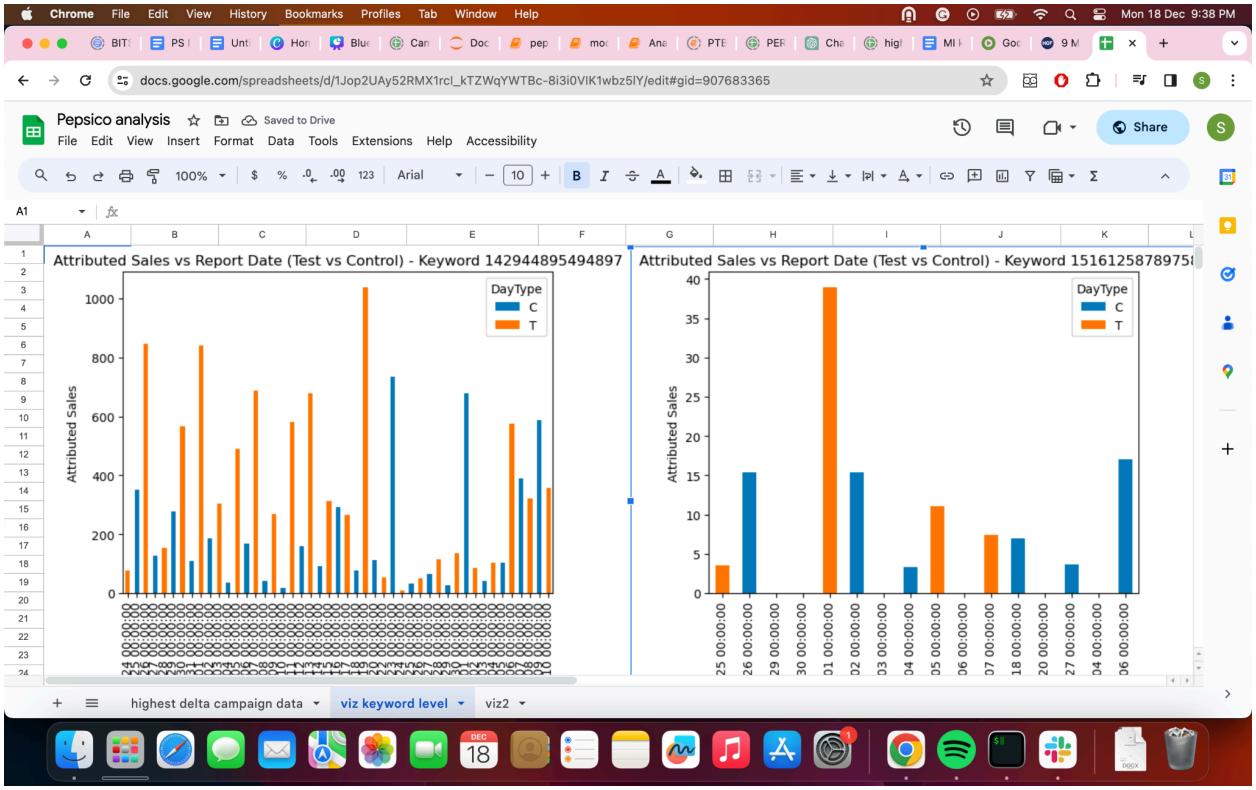
```
# Assuming your data frames are named result_df and df2

# Task 1: Match campaign_id and add columns to result_df
# Convert 'CAMPAIGN_ID' to the same data type in both data frames
result_df['CAMPAIGN_ID'] = result_df['CAMPAIGN_ID'].astype(str)
df2['CAMPAIGN_ID'] = df2['CAMPAIGN_ID'].astype(str)

# Convert 'CREATION_DATE' to date format in df2
df2['CREATION_DATE'] = pd.to_datetime(df2['CREATION_DATE']).dt.date

# Initialize 'dailyBudget' column in result_df
result_df['dailyBudget'] = None

# Iterate over each unique combination of 'CAMPAIGN_ID', 'KEYWORD_ID', and 'REPORT_DATE' in result_df
for (campaign_id, keyword_id, report_date), group_df in result_df.groupby(['CAMPAIGN_ID', 'KEYWORD_ID', 'REPORT_DATE
    # Find corresponding entry in df2 based on 'CAMPAIGN_ID' and 'CREATION_DATE' (renamed from 'REPORT_DATE')
    match_df2 = df2[(df2['CAMPAIGN_ID'] == campaign_id) & (df2['CREATION_DATE'] == report_date)]

    # If a match is found, update 'dailyBudget' in result_df
    if not match_df2.empty:
        result_df.loc[group_df.index, 'dailyBudget'] = match_df2['dailyBudget'].iloc[0]

# Task 2: Group by campaign_id and keyword_id, and arrange in ascending order of dates
result_df['REPORT_DATE'] = pd.to_datetime(result_df['REPORT_DATE'])
result_df.sort_values(['CAMPAIGN_ID', 'KEYWORD_ID', 'REPORT_DATE'], inplace=True)

# Task 3: Add a column to indicate test day or normal day
result_df['DayType'] = (result_df.groupby(['CAMPAIGN_ID', 'KEYWORD_ID']).cumcount() % 2 == 0).map({True: 'T', False:

# Reset index after sorting
result_df.reset_index(drop=True, inplace=True)
```

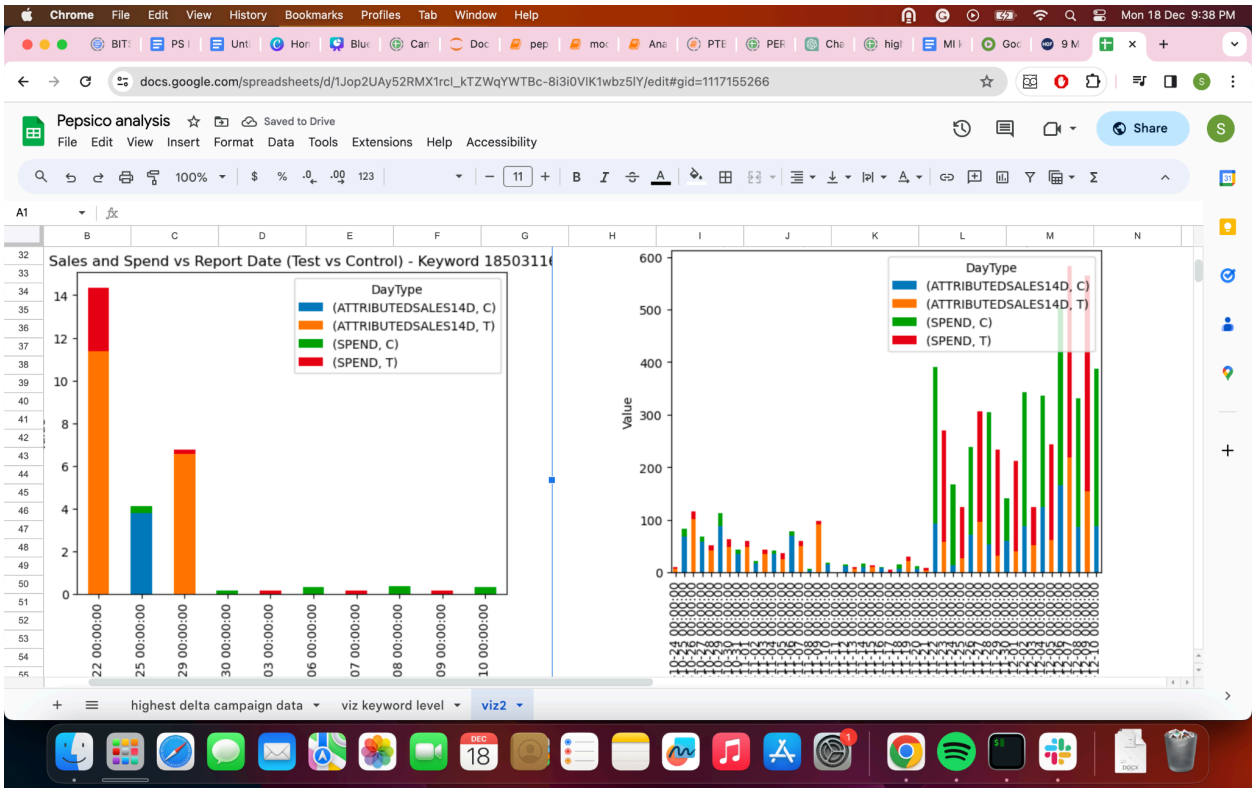


Pepsico analysis

	REPORT_DATE	CAMPAIGN_ID	KEYWORD_ID	SPEND	ATTRIBUTEDS/	INCREMENTAL_FRACT	ADSALES_MUL	COST_MUL_IF	Aggregate_MUI	AGGREGATE_I	dailyBudget	DayType
16	2023-11-07	1252915602588	1429448954948	48.02	687.39	0.58	398.6862	27.8516	398.6862	687.39	350	T
17	2023-11-08	1252915602588	1429448954948	9.93	41.9	0.59	24.721	5.8587	24.721	41.9	350	C
18	2023-11-09	1252915602588	1429448954948	18.75	267.82	0.34	91.0588	6.375	91.0588	267.82	350	T
19	2023-11-10	1252915602588	1429448954948	5.16	16.76	0.74	12.4024	3.8184	12.4024	16.76	350	C
20	2023-11-11	1252915602588	1429448954948	27.89	580.12	0.57	330.6684	15.8973	330.6684	580.12	350	T
21	2023-11-12	1252915602588	1429448954948	17.63	159.22	0.83	132.1526	14.6329	132.1526	159.22	350	C
22	2023-11-13	1252915602588	1429448954948	41.21	676.96	0.49	331.7104	20.1929	331.7104	676.96	350	T
23	2023-11-14	1252915602588	1429448954948	15.84	92.18	0.53	48.8554	8.3952	48.8554	92.18	350	C
24	2023-11-15	1252915602588	1429448954948	28.47	311.35	0.25	77.8375	7.1175	77.8375	311.35	350	T
25	2023-11-16	1252915602588	1429448954948	22.25	291.21	0.61	177.6381	13.5725	177.6381	291.21	350	C
26	2023-11-17	1252915602588	1429448954948	21.6	264.27	0.3	79.281	6.48	79.281	264.27	350	T
27	2023-11-18	1252915602588	1429448954948	11.07	76.41	0.42	32.0922	4.6494	32.0922	76.41	350	C
28	2023-11-19	1252915602588	1429448954948	56.21	1037.38	0.55	570.559	30.9155	570.559	1037.38	350	T
29	2023-11-20	1252915602588	1429448954948	12.54	113.28	0.69	78.1632	8.6526	78.1632	113.28	350	C
30	2023-11-22	1252915602588	1429448954948	18.9	53.1	0.9	47.79	17.01	47.79	53.1	350	T
31	2023-11-23	1252915602588	1429448954948	118.56	734.96	0.42	308.6832	49.7952	308.6832	734.96	350	C
32	2023-11-24	1252915602588	1429448954948	9.06	8.38	0.19	1.5922	1.7214	1.5922	8.38	350	T
33	2023-11-25	1252915602588	1429448954948	32.8	33.62	0.85	28.577	27.88	28.577	33.62	350	C
34	2023-11-26	1252915602588	1429448954948	37.55	51.39	0.55	28.2645	20.6525	28.2645	51.39	350	T
35	2023-11-27	1252915602588	1429448954948	46.53	65.05	0.52	33.826	24.1956	33.826	65.05	350	C
36	2023-11-28	1252915602588	1429448954948	39.49	114.15	0.78	89.037	30.8022	89.037	114.15	350	T
37	2023-11-29	1252915602588	1429448954948	16.32	25.95	0.65	16.8675	10.608	16.8675	25.95	350	C

highest delta campaign data | viz keyword level | viz2


Chrome File Edit View History Bookmarks Profiles Tab Window Help
Mon 18 Dec 9:38 PM
docs.google.com/spreadsheets/d/1Jop2UAy52RMX1rcl_kTZWqYWTBc-8i3i0VIK1wbz5IY/edit#gid=907683365
Pepsico analysis
Saved to Drive
File Edit View Insert Format Data Tools Extensions Help Accessibility
Share
Attributed Sales vs Report Date (Test vs Control) - Keyword 142944895494897
Attributed Sales vs Report Date (Test vs Control) - Keyword 15161258789758
DayType
C
T
Attributed Sales
highest delta campaign data
viz keyword level
viz2


Chrome File Edit View History Bookmarks Profiles Tab Window Help
Mon 18 Dec 9:38 PM
docs.google.com/spreadsheets/d/1Jop2UAy52RMX1rcl_kTZWqYWTBc-8i3i0VIK1wbz5IY/edit#gid=1117155266
Pepsico analysis
Saved to Drive
File Edit View Insert Format Data Tools Extensions Help Accessibility
Share
Sales and Spend vs Report Date (Test vs Control) - Keyword 1850311(
DayType
(ATTRIBUTEDSALES14D, C)
(ATTRIBUTEDSALES14D, T)
(SPEND, C)
(SPEND, T)
Value
highest delta campaign data
viz keyword level
viz2

11. Conclusion

11.1. Learning Outcome

This internship has provided me with the opportunity to work on several mini-projects which has enhanced my overall knowledge of the field and enabled me to understand not only the structure of an advertising campaign but also the application of data science to solve e-commerce problems.

Here is a list of things learnt:

- Important terms for advertisement on e-commerce platforms and corresponding advertising structures.
- Building dashboards and launching them using GIT
- Using LLMs for obtaining useful data using prompt engineering
- Anomaly detection techniques
- Data analysis, summarization and presentation
- Using Snowflake and other databases for retrieving and querying data
- Exploratory data analysis
- Formulation of solution statement for anomalies
- Data visualization and presentation

12. References

1. Wilke, C. O. (2019). *Fundamentals of data visualization: A primer on making informative and compelling figures*. O'Reilly Media. https://clauswilke.com/dataviz/
2. Wilkinson, L. (2005). *The grammar of graphics* (2nd edition). Springer.
3. Wu, K., Petersen, E., Ahmad, T., Burlinson, D., Tanis, S., & Szafir, D. A. (2021). Understanding data accessibility for people with intellectual and developmental disabilities. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3411764.3445743
4. Xie, Y. (2023). *knit A general-purpose package for dynamic report generation in R*. https://yihui.org/knitr/
5. Yau, N. (2011). *Visualize this: The FlowingData guide to design, visualization, and statistics*. John Wiley & Sons.
6. Yau, N. (2013). *Data points: Visualization that means something*. John Wiley & Sons.
7. Yu, A. Z., Ronen, S., Hu, K., Lu, T., & Hidalgo, C. A. (2016). Pantheon 1.0, a manually verified dataset of globally famous biographies. *Scientific Data*, *3*(1), 1–16. https://doi.org/10.1038/sdata.2015.75

13. Glossary

Budget	Financial plan that serves as an estimate of future cost, revenues or both.
Equity	Equity is the holding or stake that shareholders have in a company. Equity capital is raised by the issue of new shares or by retaining profit.
Experience	A term used to describe the relationship, usually expressed as a percent or ratio, of premiums to claims for a plan, coverage, or benefits for a stated period.
Interest	Scheduled payments made to a creditor in return for the use of borrowed money. The size of the payments will be determined by the interest rate, the amount borrowed or principal and the duration of the loan.
Leverage	With regard to corporate analysis, leverage (or gearing) refers to the extent to which a company is funded by debt.
Market	An assessment of the property value, with the value being compared to similar properties in the area.
Option	An option gives the buyer or holder the right, but not the obligation, to buy or sell an underlying financial asset at a pre-determined price.

Portfolio	A collection of investments held by an individual investor or financial institution. They may include stocks, bonds, futures contracts, options, real estate investments or any item that the holder believes will retain its value.
Provision	The amount set aside or deducted from operating income to cover expected or identified loan losses.
Recovery	The action or process of regaining possession or control of something lost. To recoup losses.
Release	An agreement between the creditor and debtor, in terms of which the creditor releases the debtor from its obligations.
Attributed Sales	the connection between the revenue generated and the marketing channel that brought the customer
DB Scan clustering	Density-based spatial clustering of applications with noise is a data clustering algorithm
K means clustering	method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean
LLM	A large language model is a large-scale language model notable for its ability to achieve general-purpose language understanding and generation
Anomaly detection	the identification of rare items, events or observations which deviate significantly from the majority of the data and do not conform to a well-defined notion of normal behaviour.