

***SCHOOL OF COMPUTER SCIENCE***



*THESIS REPORT ON*

**TRADE PROGRAM OPTIMIZATION (TPO) &**

**CUSTOMER TARGETING RECOMMENDATION**

*A Thesis in partial fulfillment of the requirements for the degree of*

*Bachelor’s Degree of Computer Science*

*Submitted By*

**Mr. SREY VOTANA**

*Under Advisory of*

**Mr. TENG CHANTO**

*Under Supervision of*

**Ms. Kanika MONTHA**

*Internship Period*

**15th January 2023 – 15th May 2023**



***School of Computer Science***

*A thesis submitted by*

**Mr. SREY VOTANA**

*Defense Date*

**7 June 2023**

*Thesis defense authorization*

President

**HE Dr. SENG SOPHEAP**

*Project Topic*

**Trade Program Optimization &**

**Customer Targeting Recommendation**

*Company:* **HEINEKEN CAMBODIA**

*Dean:* **H.E Dr. HEAN SOMBOUEN**

*Advisor:* **Mr. TENG CHANTO**

*Supervisor:* **Ms. Kanika MONTHA**

**PHNOM PENH**



**មហាវិទ្យាល័យវិទ្យាសាស្ដ្រកុំព្យូទ័រ**

**គម្រោងសញ្ញាបត្របរិញ្ញាបត្រ**

***និក្ខេបបទរបស់និស្សិត***

**លោក ស្រី វត្តនា**

*កាលបរិច្ឆេទការពារនិក្ខេបបទ*

**៧ មិថុនា ២០២៣**

*អនុញ្ញាតការពារនិក្ខេបបទ*

ប្រធានវិទ្យាស្ថានជាតិ **សេង សុភាព**

*ប្រធានបទនិក្ខេបបទ*

**Trade Program Optimization & Customer**

**Targeting Recommendation**

ក្រុមហ៊ុន៖ **ហ៊ិននីគែន ខេមបូឌា**

លោកគ្រូដឹកនាំ៖ **តេង ចាន់តូ**

អ្នកត្រួតពិនិត្យ៖ **កន្និកា**

**ACTKNOWLEDGEMENT**

First and foremost, I would like to express my sincere gratitude to **MY FAIMLY**, especially my parents for every support I have received including financial and emotional support. I am very glad to receive such support and contribution as otherwise, I would not have made it this far to the point of being able to pursue higher education at one of the most progressive colleges, winning numerous grants and even a prestigious exchange program.

Secondly, I would like to express my gratitude to His Excellency **Dr. SENG SOPHEAP**, president of the **Cambodia Academy of Digital Technology (CADT)**, and other respectful co-founders for essentially founding CADT with a spectacular vision of human resource development headed with the mix of entrepreneurship and technological innovation for a more digitally developed Cambodia.

Thirdly, I also would like to especially include my sincere gratitude to **Mr. CHANTO TENG**, a Data Science Research at CADT, who is my valuable internship advisor, for his tireless effort in directing and effectively guiding me throughout the internship process. Moreover, I was able to perform work smoothly in an organized manner thanks to the knowledge that he passionately imparted to me during the internship period.

Fourthly, I would like to express my gratitude to CADT for giving me such a great opportunity to experience the working environment despite my lack of experience in the field. I would like to greatly thank **Mr. VETTY CHIV**, my supervisor, for his invaluable supervising and mentoring me on what it takes to be professional research that succeeds in this competitive technology-driven world. With their help, I was able to complete the most required work in time and be equipped with hands-on knowledge and experience that are useful for my future career and prospects.

Last but not least, ‘thank you’ and greetings to my fellow **lecturers, classmates, and supporters** for your support and gift of knowledge that have shaped me into who I am today. I would not be able to work with the project with the learning that you have taught and in fact, I have benefited and leveraged that very learning into my experience. I can’t wait to hear more and connect with you. Thank you for everything.

**មូលន័យសង្ខេប**

របាយការណ៍នេះនឹងរៀបរាប់អំពីបទពិសោធន៍កម្មសិក្សាខែទីបួនរបស់ខ្ញុំនៅ HEINEKN CAMBODIA ក្នុងនាមជាអ្នកវិភាគទិន្នន័យ។ កម្មសិក្សាគឺចាប់ពីថ្ងៃទី 15 ខែមករា ឆ្នាំ 2023 ដល់ថ្ងៃទី 15 ខែឧសភា ឆ្នាំ 2023 ។ ចំណងជើងគម្រោងគឺ "ការបង្កើនប្រសិទ្ធភាពពាណិជ្ជកម្ម & អនុសាសន៍កំណត់គោលដៅអតិថិជន" គម្រោងនេះគឺដើម្បីគាំទ្រដល់ទីផ្សារពាណិជ្ជកម្មដើម្បីធ្វើកម្មវិធីដល់អតិថិជនត្រឹមត្រូវប្រកបដោយប្រសិទ្ធភាពនៅពេលពួកគេធ្វើពាណិជ្ជកម្ម។ កម្មវិធី ឬការផ្សព្វផ្សាយដល់អតិថិជន។ ជាងនេះទៅទៀត ក្រុមការងារទីផ្សារពាណិជ្ជកម្មនាពេលនេះគឺពិតជាអាក្រក់បំផុតលើការកំណត់គោលដៅកម្មវិធីអតិថិជន ហើយបញ្ហាគឺថាពួកគេមិនមានអនុសាសន៍ស្តង់ដារសម្រាប់ពួកគេដើម្បីកែលម្អការបង្កើនប្រសិទ្ធភាពពាណិជ្ជកម្មផងដែរ។ ដូច្នេះហើយ ក្រុមការងារ Data & Analytic បានបង្កើតគំនិតមួយដើម្បីជួយទីផ្សារពាណិជ្ជកម្ម និងក្រុមលក់ ដើម្បីដោះស្រាយបញ្ហានេះ ដើម្បីបង្កើនបរិមាណលក់ និងក្រុមហ៊ុន KPI ផងដែរ។

តាមរយៈការធ្វើដូចនេះ យើងបានប្រមូលទិន្នន័យប្រវត្តិសាស្រ្តពីកម្មវិធីពាណិជ្ជកម្ម ដើម្បីមើលការអនុវត្ត និងធ្វើការវិភាគលើព័ត៌មានលម្អិត។ លទ្ធផលនៃគម្រោងនេះ យើងនឹងមានគំរូដើម្បីកែលម្អកម្មវិធីពាណិជ្ជកម្ម រួមទាំងការអនុវត្ត កំណើនប្រាក់ចំណូល និងវិធីបង្កើនបរិមាណលក់ផងដែរ។ ជាការពិតណាស់ ក្នុងអំឡុងពេលកម្មសិក្សា ខ្ញុំបានជួបបញ្ហា និងបញ្ហាប្រឈមជាច្រើន។ ភាគច្រើននៅក្នុងការទំនាក់ទំនង និងវិធីនៃការធ្វើការជាមួយក្រុមគ្រប់គ្រង និងការរឹតបន្តឹងក្រុមហ៊ុនមួយចំនួនក្នុងការប្រើប្រាស់បច្ចេកវិទ្យាថ្មី និងប្រភពបើកចំហ។ ប៉ុន្តែបញ្ហាទាំងអស់នេះត្រូវដោះស្រាយដោយការជួយជ្រោមជ្រែងពីក្រុម រួមទាំងការទំនាក់ទំនង និងដោះស្រាយជាមួយក្រុមថ្នាក់ដឹកនាំផងដែរនូវបច្ចេកទេសមួយចំនួនផងដែរ។

ជាងនេះទៅទៀត ខ្ញុំបានទទួលនូវប្រាជ្ញា និងបទពិសោធន៍ជាច្រើនពីកម្មសិក្សានេះ រួមទាំង Hard-Skill និង Soft-Skill ផងដែរ ដូចជាការស្រាវជ្រាវបំពង់បង្ហូរប្រេង ការប្រមូលទិន្នន័យ និងស្ថាបត្យកម្ម ការប៉ះទង្គិចទិន្នន័យ ឬការសម្អាតទិន្នន័យ ការវិភាគទិន្នន័យរុករក ដំណើរការទិន្នន័យ បច្ចេកទេសរៀនម៉ាស៊ីន និង ការអភិវឌ្ឍន៍ ហើយជាពិសេសរឿងដែលសាលាមិនបង្រៀនគឺការប្រាស្រ័យទាក់ទង និងស្វែងយល់ក្នុងបរិយាកាសការងារពិត និងវិធីថ្មីក្នុងការធ្វើការតាមស្តង់ដារក្រុមហ៊ុនសកល។

**ABSTRACTION**

This report will describe my fourth month's internship experiences at **HEINEKN CAMBODIA** as a **DATA ANALYST INTERN**. The internship was from 15th January 2023 to 15th May 2023. The project title is **“TRADE PROGRAM OPTIMIZATION & CUSTOMER TARGETING RECOMMENDATION”**, this project is object to support the trade marketing to do the program to the right customers effectively when they are doing the trade program or promotion to customer. Moreover, Trade Marketing Team is currently quite well worst on the customer program targeting and the problem is they don’t have the standard recommendation to them to improve the trade optimization as well. Thus Data & Analytic Team has come up an idea to assist trade marketing and sale team to solve this out to increase the sell volume and company KPI as well.

By Doing this, we have collected the historical data from trade program to see the performance and do the analysis on the detail. The outcome of this project, we would have the model to improve the trade program including the performance, revenue growth, and the way to increase the sale volume as well.

Of course, during an internship, I met the issues and challenges a lot. Mostly in the communication and the way of working with management team and some of company restriction of using new technologies and open source. But all these problems have figure out by supporting from the team including communication and deal with leadership team also some of technical things either.

Moreover, I received a lot of wisdom and experience from this internship including Hard-Skill and Soft-Skill as well such as research pipeline, data collection and architecture, data wrangling or data cleansing, exploratory data analysis, data preprocessing, machine learning technique and development, and mostly especially thing that school doesn’t teach is communication and explore in real working environment and new way working in global company standard.

**ABBREVIATION**

|  |  |  |
| --- | --- | --- |
| No | Abbreviation | Meaning |
| 1 | TPO | Trade Program Optimization |
| 2 | HCAM | HEINEKN Cambodia |
| 3 | ML | Machine Learning |
| 4 | EDA | Exploratory Data Analysis |
| 5 | AA | Advance Analytic |
| 6 | KNN | K-Nearest Neighbors |
| 7 | DL | Deep Learning |
| 8 | TMK | Trade Marketing Team |
| 9 |  |  |
| 10 |  |  |
| 11 |  |  |
| 12 |  |  |
| 13 |  |  |

***TABLE OF CONTENT***

ACTKNOWLEDGEMENT.....................................................................................................................................

មូលន័យ.....................................................................................................................................

ABSTRACTION .....................................................................................................................................

ABBREVIATION............................................................................................................................

1. General presentation of the company .................................................................
   1. HEINEKEN Cambodia History ............................................................................
   2. Service ..................................................................................................................................
   3. Organization Chart .........................................................................................................
2. Introduction to the project ......................................................................................
   1. Presentation of the project ...........................................................................................
   2. Problematic ........................................................................................................................
   3. Objective ..............................................................................................................................
   4. Planning ...............................................................................................................................
   5. Development Methodology..........................................................................................
3. Literacture Review .........................................................................
4. Analysis and General Concept ...............................................................................
   1. Choice of technology .......................................................................................................
      1. Languages ..........................................................................................................................
      2. Libraries ..........................................................................................................................
      3. Tools ................................................................................................................................................................
   2. Methods ...............................................................................................................................
      1. Dataset ...............................................................................................................
      2. Data Collection .................................................................................................
      3. Data Cleansing..................................................................................................
      4. Exploratory Data Analysis...................................................................................................................
      5. Data Preprocessing ...............................................................................................................
      6. Model Training ...................................................................................................................................
      7. Model Evaluation ...................................................................................................................................
5. Conclusion .....................................................................................................................
   1. Conclusion ...........................................................................................................................
   2. Insight Result ....................................................................................................................................
   3. Difficulties ...........................................................................................................................
   4. Experiences ........................................................................................................................
   5. Future Perspective ...........................................................................................................
6. Reference .......................................................................................................................
7. Appendix ........................................................................................................................

***LIST OF FIGURES***

*Figure 1:*

*Figure 2:*

*Figure 3:*

***LIST OF TABLES***

*Table 1:*

*Table 2:*

*Table 3:*

1. **GENERAL PRESENTATION OF THE COMPANY** 
   1. **HEINEKN Cambodia History**

HEINEKEN Cambodia is a proud part of the HEINEKEN Company. We find our roots in Cambodia back in 1994. Born from the Joint Venture between Asia Pacific Breweries Limited and Progress Import-Export, we were known at that time and for many years after as Cambodia Breweries Limited (CBL). After joining the HEINEKEN family in 2014, we finally became HEINEKEN Cambodia in 2020.

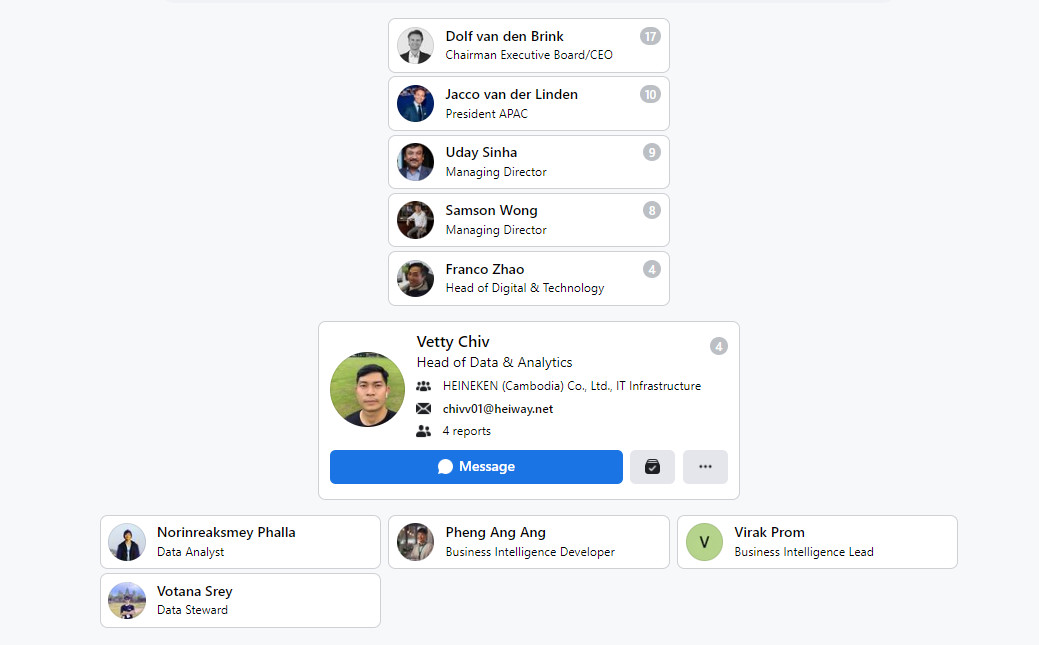
We are proud of our HEINEKEN history as a family-owned, independent brewer that stretches back 150 years in the Netherlands. We brew the highest quality beers, build globally loved brands and strive to do both things as sustainably as possible. Our master brewers produce our global, regional and local beers in our state-of-the-art brewery located in Phnom Penh. Made passionate by the art of brewing beers, they strictly follow the highest international standards and use ingredients and recipes identical to those applied in all HEINEKEN breweries across the globe. Every day, their magic brings to life our five jewels: Heineken®, Tiger, Anchor, ABC Extra Stout, and Gold Crown. We offer a world of opportunities where people are at the center. We value purpose, innovation, inclusion and diversity, and sustainability. Aiming to Brew a Better Cambodia, we are proud to be the top taxpayer across all industries, truly supporting the social and economic development of Cambodia

* 1. **HEINEKN Cambodia Service**

Figure

Every day, HEINEKEN’s magic brings to life our five jewels: Heineken®, Tiger, Anchor, ABC Extra Stout, and Gold Crown. We value purpose, innovation, inclusion and diversity, and sustainability. Aiming to Brew a Better Cambodia, we are proud to be the top taxpayer across all industries, truly supporting the social and economic development of Cambodia.

* Location: P.O. Box 1108, Phoum Robors Angkagne, Sangkat Prek Eng, Khan Chbar Ampov, Phnom Penh, Kingdom of Cambodia.
* Contact:
  + Phone Number: +855 96 555 1777 (Telephone and Telegram app)
  + Website: [HEINEKEN's Website](https://www.heinekencambodia.com/)
  + Facebook: [HEINEKEN's Facebook](https://www.facebook.com/HEINEKENCambodia)
  + LinkedIn: [HEINEKEN LinkedIn](https://www.linkedin.com/company/heinekencambodia/)
  1. **Organization Chart**

****

1. **INTRODUCTION OF THE PROJECT** 
   1. **Presentation of the Project**

Trade Program Optimization is the initiative project from Data and Analytic team to support the trade marketing and sale team to optimize the key performance indicator of them and increase the sale volume of the company as well. By doing this project, we have a model and analyze the root cause of decrease and increase reason of each trade program that implementing in the marketing and after that. Moreover, that’s not only one thing that our model helps to the stakeholder’s team and company, but it will also support the team to get the right target customers to fit their trade program scheme and KPI that they set. It’s generally involved the many teams such as Data & Analytic Team, Sale Team, Trade Marketing Team, Branding Team, Inventory Management Team, and Digital & Technology Team as well.

Thus, technically our model will aim to combine both the estimation of the target customer and the optimization of targeted promotions over the graph. For this purpose, we need to ensure that the target customer can be optimized over. Therefore, we construct the target customer recommendation model that can be represented by an interpretable graph model describing by how much one customer’s purchase increases the probability of another customer’s purchase, sale volume and make more profit to the company.

* 1. **Problematic**

The starting point of the root cause problem to raise this project up ais Trade Marketing team doesn’t know where, who, how should they target to the right customers that they want to achieve their KPI of their trade program. Thus, this is the big root cause that already happened. By doing that, we would our initiative project will support team to solve this problem out.

* 1. **Objectives**

The Trade Program Optimization has the object to support the trade marketing to do the program to the right customers effectively. Moreover, Trade Marketing Team is currently quite well doing on the customer program targeting but the problem is we don’t have the standard recommendation to them to improve the trade program targeting and optimization as well. Thus Data & Analytic Team has come up an idea to assist TM to solve this out to increase the sell volume and company KPI as well. The objectives would be:

* Improve the trade program targeting effectively
* Increase the sell volume and profit generate by trade program optimization model
* Understand the customer’s performance & behaviors
* Improve the master data and data model & what the need to push master data help the trade program targeting.

**Academic/Practical Relevance**: From an academic point of view, we want to develop the model to detect the customer targeting effect solely from transactional data and optimal program targeting as well.

* 1. **Project Planning**

The project has 4 months in duration. That’s starting from 15 January 2023 to 15 May 2023. This table below illustrate the project planning and listing the tasks from A to Z.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Task** | **Week** | | | | | | | | | | | | | | | |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** |
| Understand Business Pain Point |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Align with the Stakeholders |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Install an Environment |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Data Collection |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Data Cleansing |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Exploratory Data Analysis |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Machine Learning Training |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Machine Learning Evaluation |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Result Validation |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Prepare Slide & Report Thesis |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

* 1. **Development Methodology**

For this project management framework, HEINEKEN and the team is currently using the scrum framework, we 1 sprint has 2 weeks, and we can perform it and maintain the task by weekly following to what we have define in the jira project management tool. The Daily Scrum is a 15-minute time-boxed event for the Development Team. The Daily Scrum is held every day of the Sprint. At it, the Development Team plans work for the next 24 hours. This optimizes team collaboration and performance by inspecting the work since the last Daily Scrum and forecasting upcoming Sprint work. The Daily Scrum is held at the same time and place each day to reduce complexity.

The Development Team uses the Daily Scrum to inspect progress toward the Sprint Goal and to

inspect how progress is trending toward completing the work in the Sprint Backlog. The Daily

Scrum optimizes the probability that the Development Team will meet the Sprint Goal. Every

day, the Development Team should understand how it intends to work together as a self organizing team to accomplish the Sprint Goal and create the anticipated Increment by the end

of the Sprint.

The structure of the meeting is set by the Development Team and can be conducted in different

ways if it focuses on progress toward the Sprint Goal. Some Development Teams will use

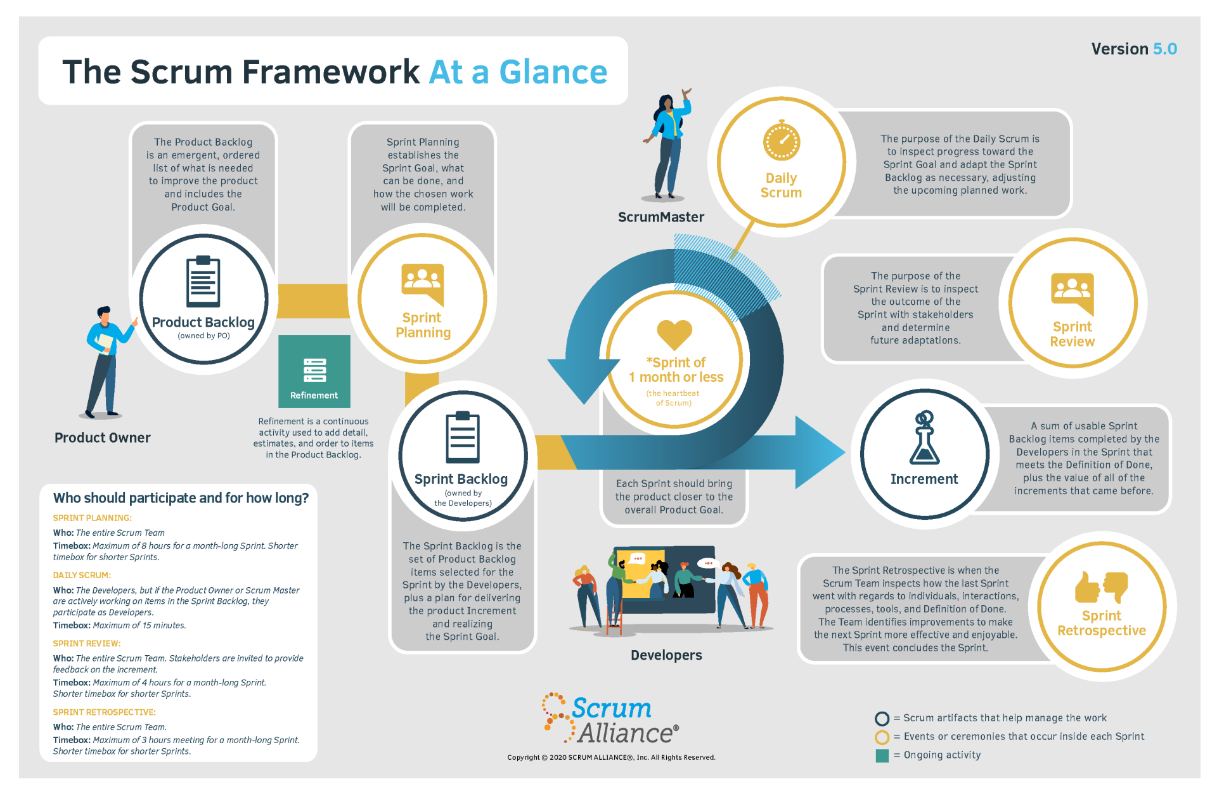
questions, some will be more discussion based. Here is an example of what might be used:

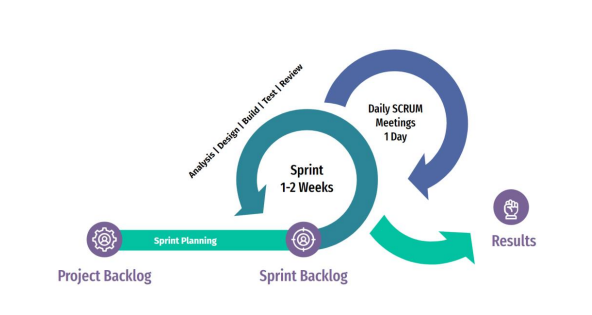
• What did I do yesterday that helped the Development Team meet the Sprint Goal?

• What will I do today to help the Development Team meet the Sprint Goal?

• Do I see any impediment that prevents me or the Development Team from meeting the

Sprint Goal?





1. **LITERACTURE REVIEW**

TPO stands for Trade Promotion Optimization, which is a process of designing and implementing effective trade promotion strategies to improve the efficiency and effectiveness of promotional activities. Trade promotion optimization involves analyzing historical data, market trends, and consumer behavior to identify the most effective promotional strategies, such as the timing, frequency, and level of discounts, among others. TPO also involves monitoring and evaluating the performance of promotional activities to make adjustments and improvements over time. The goal of TPO is to maximize the return on investment (ROI) of promotional activities and achieve marketing objectives, such as increasing sales, market share, and customer loyalty. Trade Promotion Optimization (TPO) is important for several reasons:

* Cost-effectiveness: Trade promotion activities can be expensive, and without proper optimization, firms may not see a significant return on their investment. TPO helps firms identify the most effective promotion strategies and allocate their resources accordingly, resulting in cost-effective promotion activities.
* Improved performance: TPO can help firms improve the performance of their promotional activities. By analyzing data and market trends, firms can identify the most effective promotion strategies, which can lead to increased sales, market share, and customer loyalty.
* Competitive advantage: TPO can provide firms with a competitive advantage. By optimizing their promotion activities, firms can better target their customers, differentiate their products from competitors, and improve their overall marketing effectiveness.
* Increased efficiency: TPO can also improve the efficiency of promotional activities. By streamlining processes and reducing waste, firms can achieve their marketing objectives with fewer resources and less time.

Trade promotion optimization (TPO) is a critical area of research in marketing and sales management. TPO involves the use of analytical tools and techniques to design, implement, and evaluate trade promotions that maximize the effectiveness of marketing investments. The literature on TPO includes a wide range of research studies and review articles that cover various aspects of trade promotion management. In this review, we summarize some of the most relevant research in this field. **Trade promotion management (TPM)** refers to the process of planning, executing, and evaluating trade promotions. A key aspect of TPM is the use of data analytics to optimize the effectiveness of promotions. Researchers have studied different aspects of TPM, including the role of data analytics, the effectiveness of different promotion types, and the impact of trade promotions on retailer and manufacturer profitability. **Promotion optimization models (POMs)** are mathematical models that use data analytics to optimize promotion design and execution. POMs can help manufacturers and retailers make data-driven decisions about promotion types, timing, and frequency. Researchers have developed various types of POMs, including linear and nonlinear programming models, decision trees, and simulation models. Studies have evaluated the effectiveness of different POMs and identified the factors that affect their accuracy and usefulness. **Trade promotions** are often used in conjunction with price discounts to attract customers and increase sales. Researchers have studied the relationship between pricing and promotions and identified the factors that affect the effectiveness of these strategies. Some studies have shown that promotions are more effective when they are combined with price discounts, while others have suggested that promotions can be effective even without price discounts. Researchers have studied the impact of trade promotions on various outcomes, including sales, profits, and customer behavior. Studies have shown that trade promotions can be effective in increasing sales and profits in the short term, but their long-term impact is less clear. Some studies have suggested that trade promotions can have a negative impact on profitability if they are not carefully managed. **Data analytics plays a critical role in TPO**, and researchers have studied the use of different data analytics techniques in promotion optimization. Some studies have focused on the use of machine learning algorithms to predict the effectiveness of promotions, while others have studied the use of data visualization tools to help managers make better decisions.

In conclusion, the literature on TPO is broad and diverse, covering various aspects of trade promotion management. Researchers have developed different models and analytical techniques to optimize promotion design and execution and have studied the impact of trade promotions on sales, profits, and customer behavior. The effectiveness of trade promotions depends on several factors, including promotion type, pricing strategy, and data analytics. Further research is needed to develop more accurate and effective promotion optimization models and to understand the long-term impact of trade promotions on business performance. "Optimizing trade promotion strategies in the presence of category captains" by Praveen K. Kopalle and Robert L. Steiner. Published in Marketing Science, this paper presents a model for optimizing trade promotion activities in the presence of category captains. The authors suggest that firms can increase their profitability by coordinating their promotion activities with category captains, who have a significant influence on consumer purchasing behavior. "Optimizing trade promotions: analysis of consumer and retailer promotion sensitivity" by Dipayan Biswas, David J. Lucking-Reiley, and Arun K. Jain. This paper, published in the Journal of Retailing, examines the impact of consumer and retailer promotion sensitivity on trade promotion effectiveness. The authors propose a model for optimizing trade promotion activities based on the degree of consumer and retailer sensitivity to promotions. "Optimizing trade promotions in the presence of forward buying and consumer stockpiling" by Shantanu Bhattacharya and Dipak C. Jain. This paper, published in Marketing Science, presents a model for optimizing trade promotion activities in the presence of forward buying and consumer stockpiling. The authors suggest that firms can increase their profits by carefully managing their promotion activities to account for these phenomenal. "A review of trade promotion optimization: from modeling to practice" by M. Berk Ataman and Carl Mela. Published in Marketing Science, this paper provides a comprehensive review of the literature on trade promotion optimization. The authors identify key trends and developments in the field and provide recommendations for future research. "Optimizing trade promotions: a category management approach" by Kusum L. Ailawadi and Kevin E. Lane. This paper, published in the Journal of Business Research, proposes a category management approach to optimizing trade promotions. The authors suggest that firms can increase their profitability by aligning their promotion activities with category management objectives.

1. **ANALYSIS AND GENERAL CONCEPT**
   1. **Choice of Technology**
      1. **Languages**

The following list is the languages that are used in the Project and Task Data Modelling and Machine Learning:

****

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built-in data structures, combined with dynamic typing and dynamic binding; make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python is the simple, easy-to-learn syntax that emphasizes readability and therefore reduces the cost of program maintenance. Python support modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms and can be freely distributed.



Structured Query Language, abbreviated as SQL, is a domain-specific language used in programming and designed for managing data held in a relational database management system, or for stream processing in a relational data stream management system. SQL is used to communicate with a database. According to ANSI (American National Standards Institute), it is the standard language for relational database management systems. SQL statements are used to perform tasks such as update data on a database or retrieve data from a database.

* + 1. **Libraries**

****

NumPy, short for Numerical Python, is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, 12 along with a large collection of high-level mathematical functions to operate on these arrays. It provides, among other things:

* A fast and efficient multidimensional array object and array
* Functions for performing element-wise computations with arrays or mathematical operations between arrays
* Tools for reading and writing array-based data sets to disk
* Linear algebra operations, Fourier transforms, and random number

Generation

* Tools for integrating connecting C, C++, and Fortran code to Python

Beyond the fast array-processing capabilities that NumPy adds to Python, one of its primary purposes with regards to data analysis is as the primary container for data to be passed between algorithms. For numerical data, NumPy arrays are a much more efficient way of storing and manipulating data than the other built-in Python data structures. In addition, libraries written in a lower-level language, such as C or Fortran, can operate on the data stored in a NumPy array without copying any data.

****

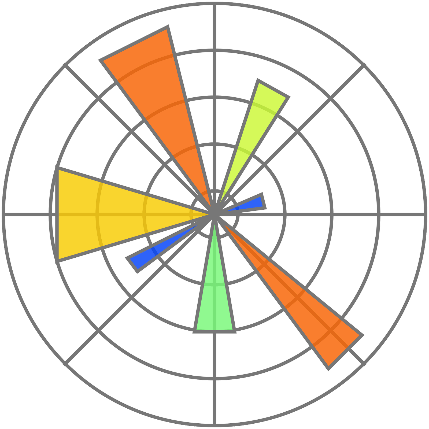
Pandas has always been great for data munging; however, it was not great for analysis compared with databases using SQL, Excel, or R data frames. Pandas is an open-source Python package providing fast, flexible, and expressive data structures designed to make working with “relational” or “labeled” data both easy and intuitive. Pandas were developed by Wes McKinney in 2008 while at AQR Capital Management, out of the need for a high-performance, flexible tool to perform quantitative analysis of financial data. Before leaving AQR, he was able to convince management to allow him to open source the library. Pandas is well suited for tabular data with heterogeneously typed columns, as in an SQL table or Excel spreadsheet.

****

Scikit-Learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means, DBSCAN, Linear Regression, Naive Bayes, Decision Tree, Logistic Regression, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy. Scikit-learn is a NumFOCUS fiscally sponsored project.

****

Seaborn is an open-source Python library built on top of matplotlib. It is used for data visualization and exploratory data analysis. Seaborn works easily with data frames and the Pandas library. The graphs created can also be customized easily.

****

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK. There is also a procedural "pylab" interface based on a state machine (like OpenGL), designed to closely resemble that of MATLAB, though its use is discouraged. SciPy makes use of Matplotlib.

****

Plotly's Python graphing library makes interactive, publication-quality graphs. Examples of how to make line plots, scatter plots, area charts, bar charts, error bars, box plots, histograms, heatmaps, subplots, multiple-axes, polar charts, and bubble charts. It’s used to build on top of the Plotly JavaScript library (plotly. js), plotly enables Python users to create beautiful interactive web-based visualizations that can be displayed in Jupyter notebooks, saved to standalone HTML files, or served as part of pure Python-built web applications using Dash.

* + 1. **Tools**

Tools is the software application that helps programmers develop software code efficiently. It increases developer productivity by combining capabilities such as software editing, building, testing, and packaging in an easy-to-use application. Just as writers use text editors and accountants use spreadsheets, software developers use IDEs to make their job easier.

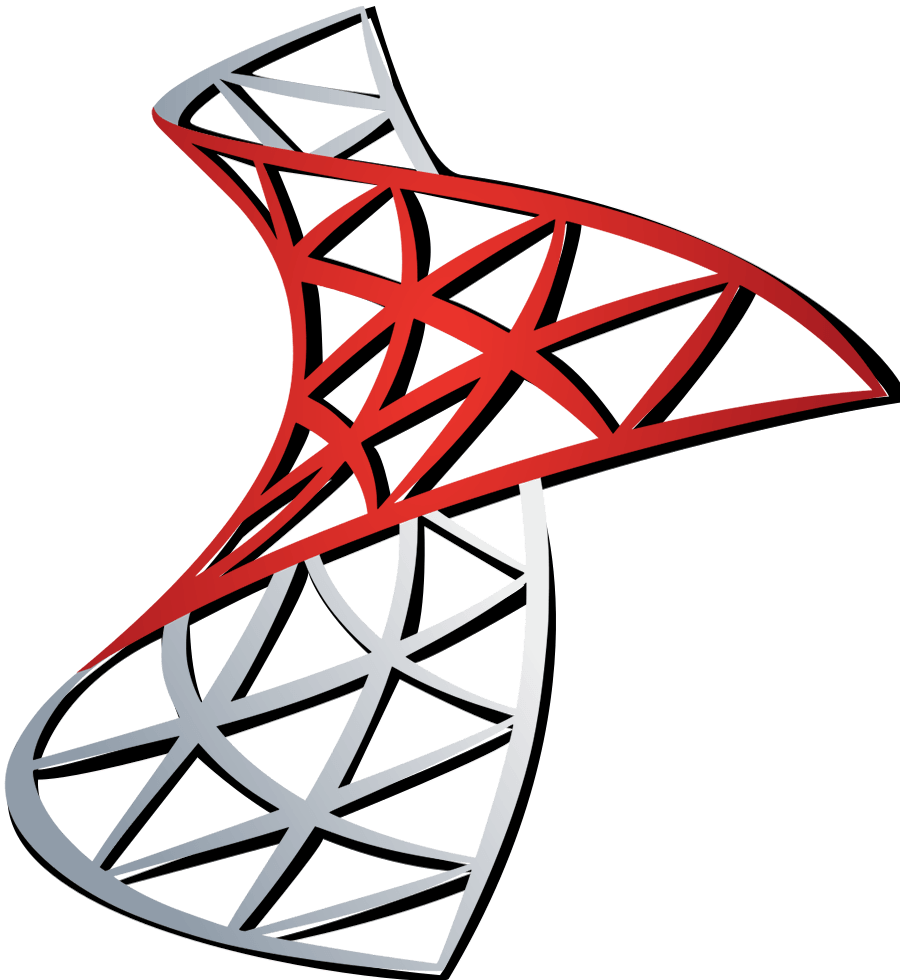
Here is the essential tools that we are using for perform our project:



Anaconda is a free and open-source distribution of the Python and R programming languages for data science and machine learning-related applications (large-scale data processing, predictive analytics, scientific computing), that aims to simplify package management and deployment. Package versions are managed by the package management system anaconda. The Anaconda distribution is used by over 6 million users, and it includes more than 250 popular data science packages suitable for Windows, Linux, and macOS.

****

Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations, and narrative text. Uses include data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.



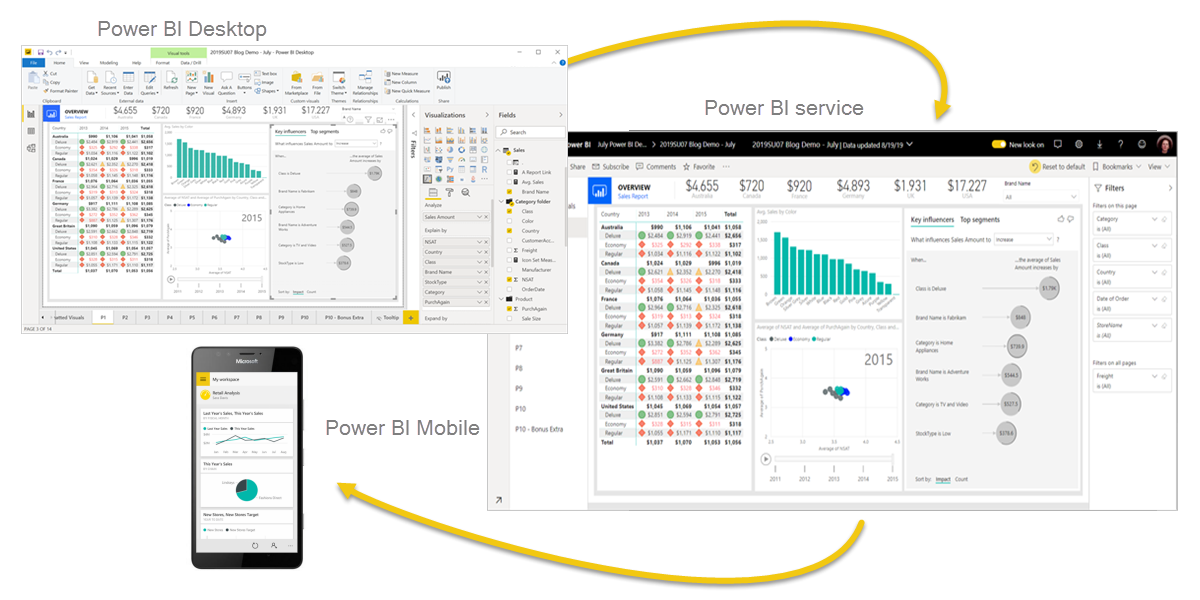
Microsoft SQL Server Management Studio is a software application developed by Microsoft that is used for configuring, managing, and administering all components within Microsoft SQL Server. First launched with Microsoft SQL Server 2005, it is the successor to the Enterprise Manager in SQL 2000 or before. Microsoft SQL Server is a relational database management system. As a database server that stores and retrieves data as requested by other software applications on the same computer or a remote computer using the client-server model. Microsoft provides APIs to access SQL Server over the internet as a web service.



Microsoft Power BI is an interactive data visualization software product developed by Microsoft with a primary focus on business intelligence. It is part of the Microsoft Power Platform. Power BI is a collection of software services, apps, and connectors that work together to turn your unrelated sources of data into coherent, visually immersive, and interactive insights. Your data might be an Excel spreadsheet, or a collection of cloud-based and on-premises hybrid data warehouses. Power BI lets you easily connect to your data sources, visualize and discover what's important, and share that with anyone or everyone you want.

Power BI consists of several elements that all work together, starting with these three basics:

* A Windows desktop application called Power BI Desktop.
* An online software as a service (SaaS) service called the Power BI service.
* Power BI Mobile apps for Windows, iOS, and Android devices.



* 1. **Method** 
     1. **Dataset**

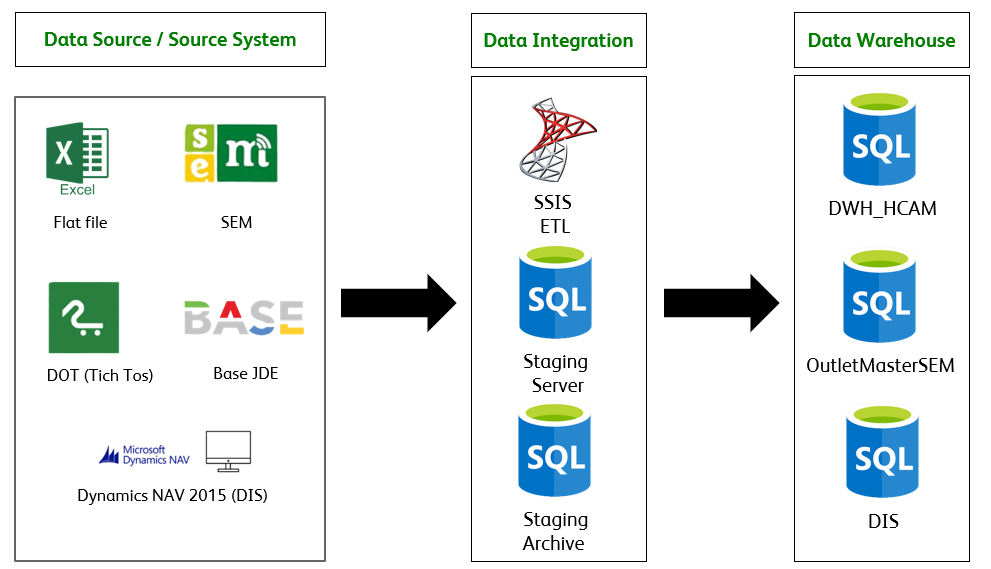
|  |  |  |
| --- | --- | --- |
| No | Name | Description |
| 1 | Agent\_Code | Agent ID Unique Code |
| 2 | Agent\_Name | Agent Name |
| 3 | Outlet\_Code | Customer Code |
| 4 | Outlet\_Name | Customer Name in Khmer |
| 5 | OutletNameEn | Customer Name in English |
| 6 | Region | Customer Region |
| 7 | Province | Customer Province |
| 8 | Account\_Type | Customer Type |
| 9 | Status | Customer Status |
| 10 | Grade | Customer Grade |
| 11 | City | Customer City |
| 12 | District | Customer District |
| 13 | Commune | Customer Commune |
| 14 | Village | Customer Village |
| 15 | Cluster\_Lead | Cluster Lead |
| 16 | Segment | Customer Segment |
| 17 | Business\_Org\_Segment | Business Organization Segment |
| 18 | Business\_Segment | Business Segment |
| 19 | Category | Category |
| 20 | ChannelName | Channel Name |
| 21 | OutletCreditStatus | Outlet Credit Status |
| 22 | AMS\_HCAM\_Premium\_Brand | AMS HCAM Premium Brand |
| 23 | AMS\_HCAM\_Mainstream\_Brand | AMS HCAM Mainstream Brand |
| 24 | AMS\_Competitors\_Premium\_Brand | AMS Competitors Premium Brand |
| 25 | AMS\_Competitors\_Mainstream\_Brand | AMS Competitors Mainstream Brand |
| 26 | OrderType | Order Type |
| 27 | Document No\_ | Document Number |
| 28 | DocType | Document Type |
| 29 | ContractType | Contract Type |
| 30 | OrderCategory | Order Category |
| 31 | ItemCode | Item Code |
| 32 | Quantity | Quantity |
| 33 | HL | HL |
| 34 | amount1 | Amount |
| 35 | Amount Including VAT | Amount Including VAT |
| 36 | itm\_code | Item Code |
| 37 | itm\_des | Item Description / Name |
| 38 | OUTLET\_LATITUDE | OUTLET LATITUDE |
| 39 | OUTLET\_LONGITUDE | OUTLET LONGITUDE |
| 40 | level1\_territory and Name | Sale Representative |
| 41 | level2\_territory and Name | Sale Supervisor |
| 42 | level3\_territory and Name | Sale Operation Manager |
| 43 | level4\_territory and Name | Channel Head |
| 44 | Order Date | Order Date |
| 45 | Posting Date | Posting Date |
| 46 | Posting Description | Posting Description |
| 47 | Shipment Date | Shipment Date |
| 48 | Due Date | Due Date |
| 49 | Document Date | Document Date |

* + 1. **Data Collection**

In ordering to collect the data. Data collection is the process of gathering information or data for research, analysis, or other purposes. The collected data can be qualitative or quantitative and can be gathered through various methods such as surveys, interviews, observations, and experiments.

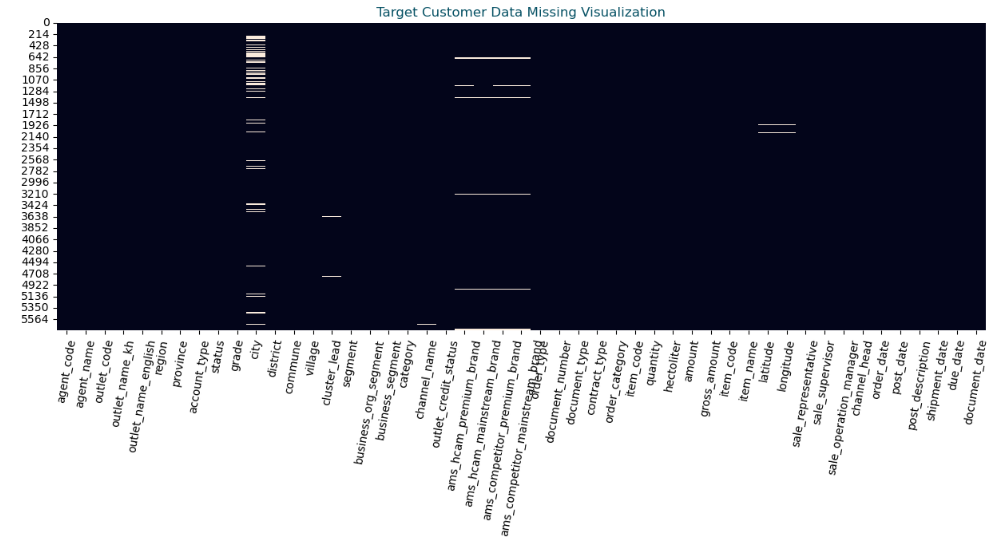
It is important to ensure that the data collected is reliable and valid. Reliability refers to the consistency and stability of the data over time and across different contexts, while validity refers to the accuracy and relevance of the data in relation to the research. To ensure reliability and validity, it is important to use appropriate data collection methods and instruments, and to carefully design and pilot-test data collection procedures.

Here is the process that we collect the dataset from to do the research and analytics. There are many different source systems. So we have only collect from 3 main database/data warehouse which support our analytics. Please kindly refer to the flow here.

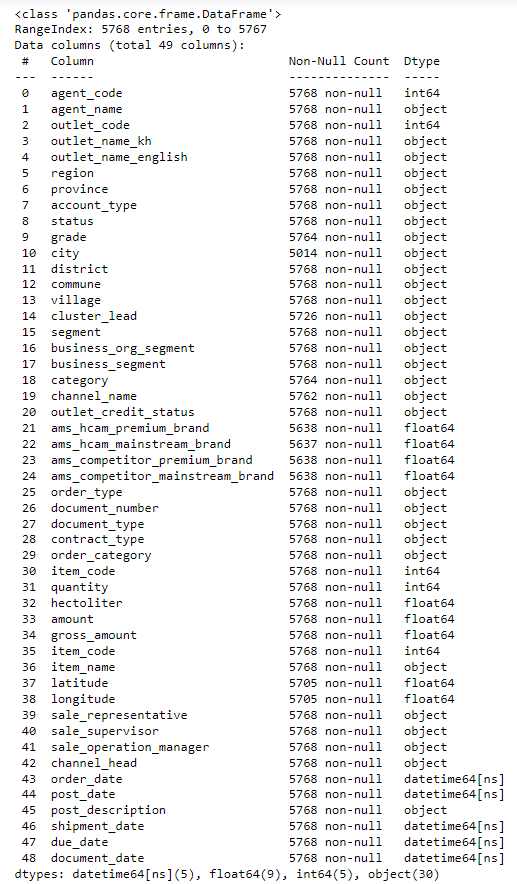
****

* + 1. **Data Cleansing**

Data cleansing, also known as data cleaning or data scrubbing, is the process of identifying and correcting or removing errors, inconsistencies, and inaccuracies in data. This process involves various techniques and tools to ensure that the data is accurate, consistent, and reliable. The primary objective of data cleansing is to improve the quality of the data and make it useful for analysis, reporting, and decision-making. So we have define the step to validate the data:

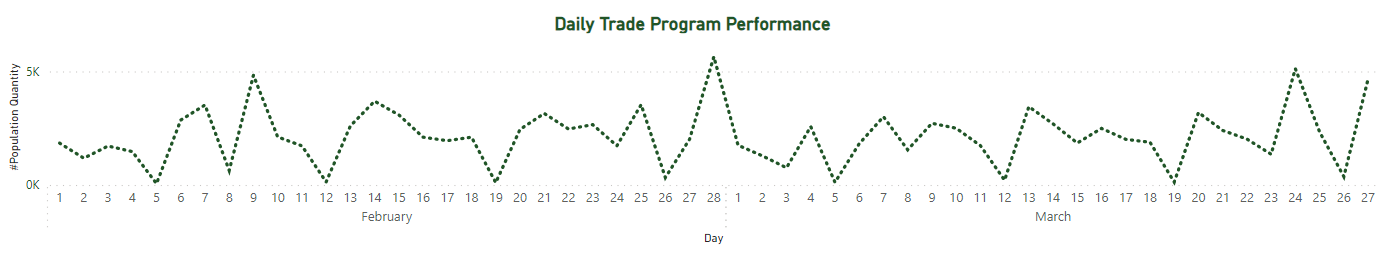
****

This plot, we used to check the availability of data records in each. Most of missing records is outlet grade, city and average monthly sale data (market share) that perform by sale team currently.

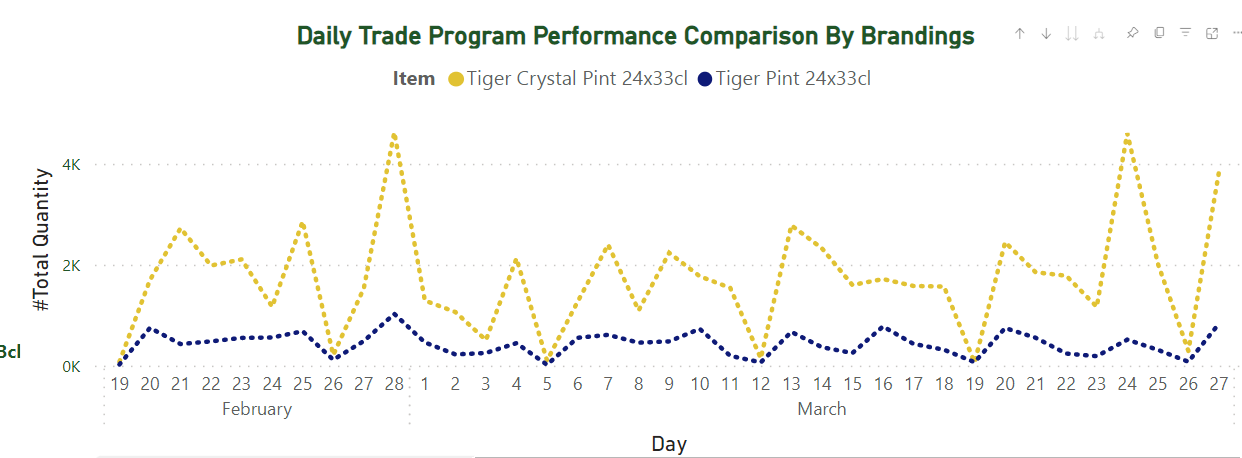
****

We already checked and validated on the number columns, column name and data type that will be transformed to be corrected one.

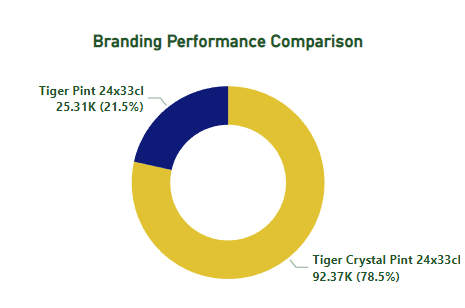
* + 1. **Exploratory Data Analysis**

****

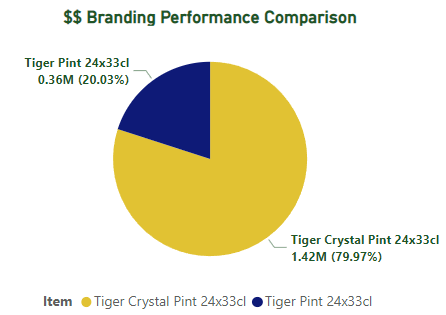
2 Months daily sale has been generated in one line plot to see the trend by day-to-day. This is the result of tiger trade program that we implement for on-trade outlet type. By doing this, we mostly see the sale increase mostly on weekend and Friday or any festival, day off as well. So that’s possible to do the promotion during that day to reduce the investment and also keep the profit rate the same.

****

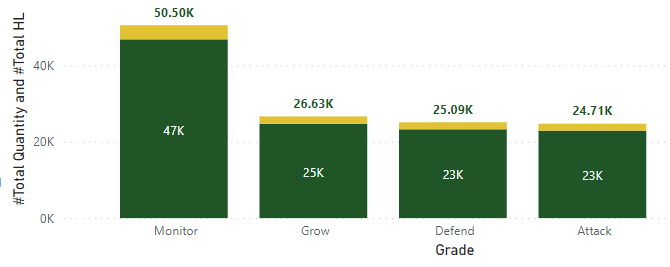
In the program, only 2 item are in the list which are tiger crystal and tiger Pint. We have compared what is the most performance brand for doing this trade credential program as well. So above here is the comparison plot to see the performance of tiger crystal and tiger pint perform on day-to-day as well. The most brand here we get is tiger crystal. The reason of this is modern on-trade outlet are mostly got top sale for tiger crystal.

****

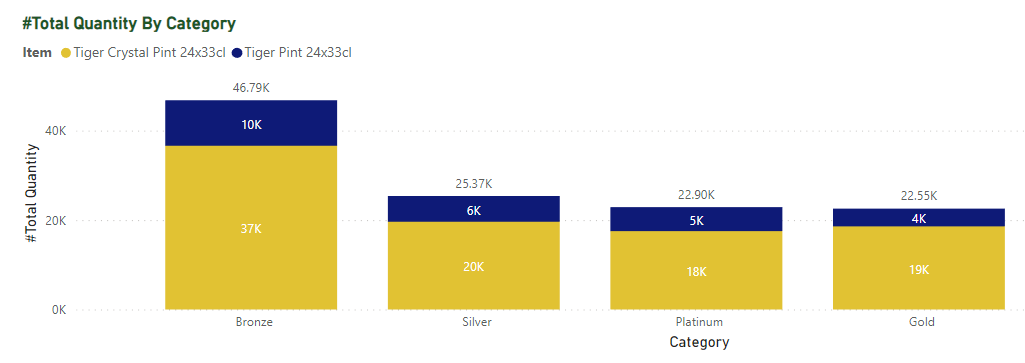
To see more detail number of branding in each, Tiger Crystal pint got 78.5% out of 100% and 21.5% of tiger pint. This is the pie plot that indicating each brand perform in whole program.

****

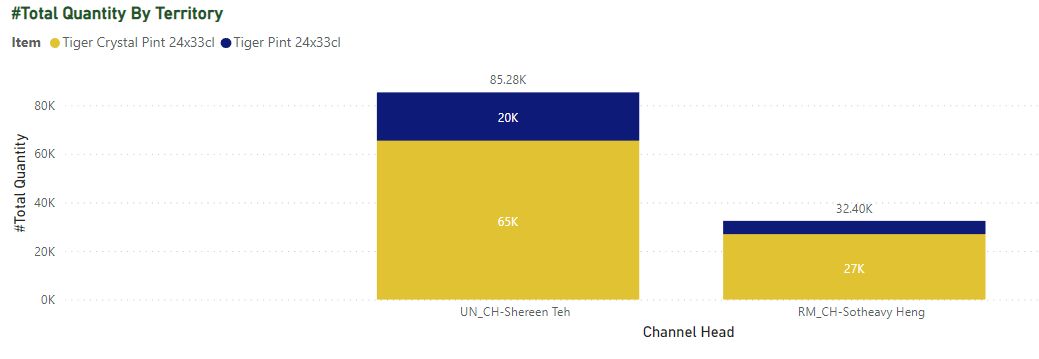
The pie plot above just indicating the distribution of profit we got from the trade program. The number of tiger crystal, we got $1.42M and also tiger Pint we only got $0.36M. this is also good number we got from the trade program as well if compared to the business case from finance team and this program is not working bad on ROI too. By the way, we can consider on Tiger Crystal for doing this next trade program if magmatism is the same like this one.

****

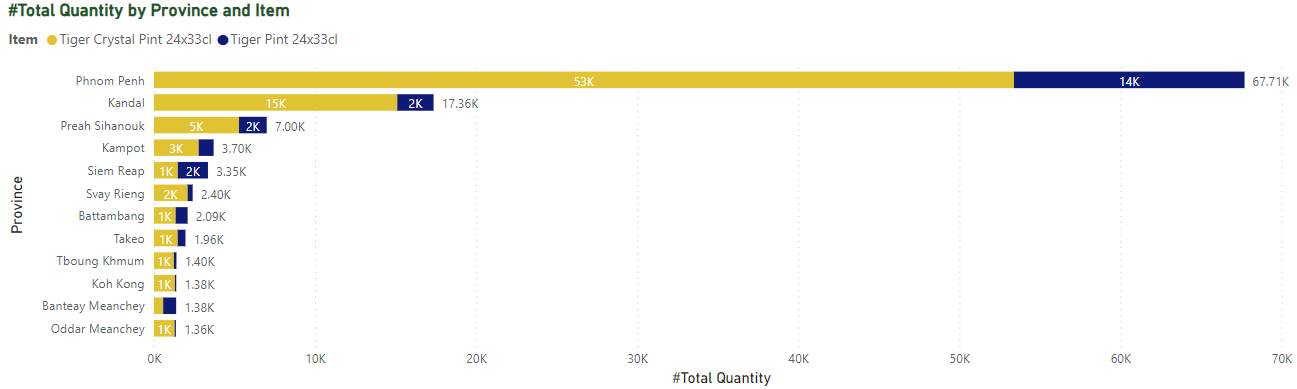
The distribution of outlet grade that perform during this promotion. The monitor outlet grade is really perform well and thank to the sale team as well that boost the outlet owner that got grade Monitor to increase their sale during the program. And next outlet grade Grow, Defend and Attack are perform quit similar to each other’s.

****

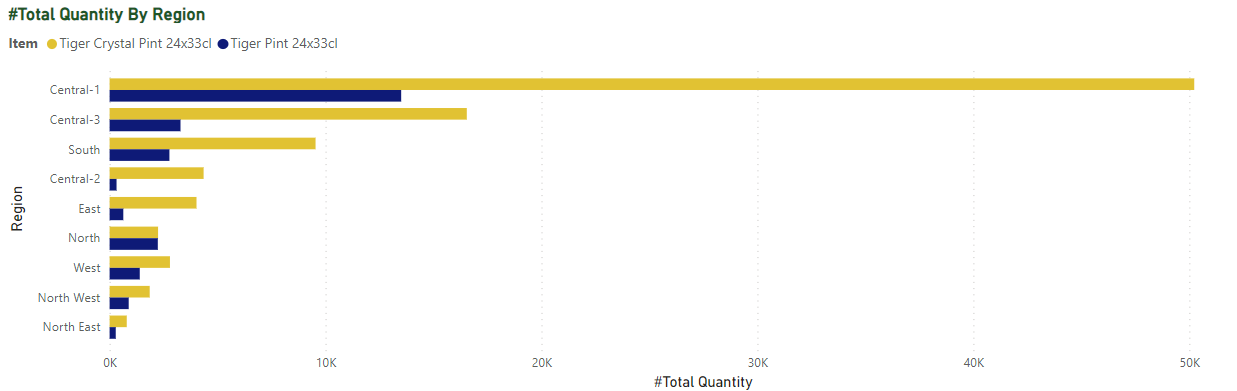
The distribution of outlet category that perform during this promotion. The Bronze category is really performed well and thank to the sale team as well.

****

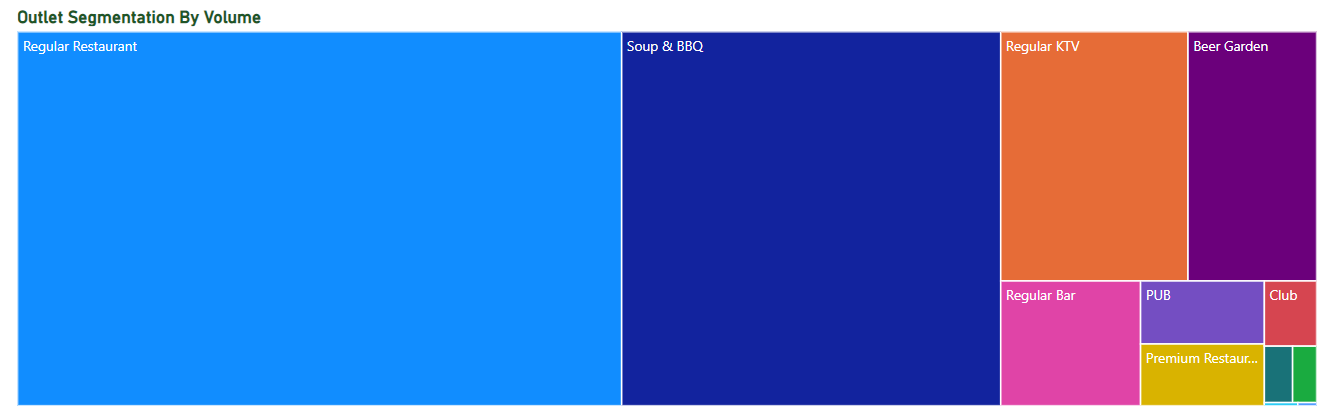
Since this program is implemented for modern on-trade so only 2 sale territory are in the listed. Mostly of the well-performed is Shereen The sale team. Since most of modern on-trade are under her supervision and their sale team is really active to support the daily sale.

****

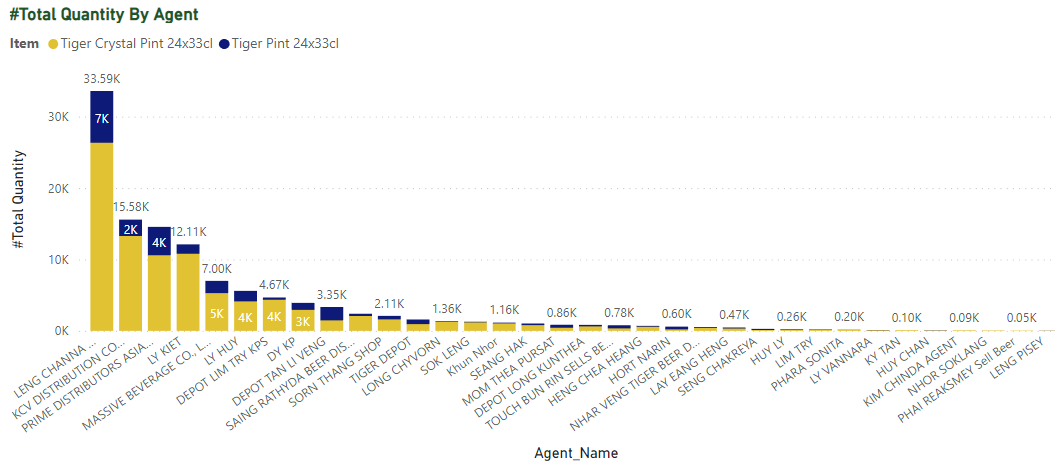
This is the performance by province. By seeing this plot, we can see Phnom Penh, Kandal, Preah Sihanouk and Kampot are the top sale during the program.

****

This is the performance by region. By seeing this plot Central-1 is perform really well and its correlation to the province as well since top performance sale of the provinces are in the central-1. So there is not weak correlation between 2 features here.

****

The tree plot here is indicating us the outlet segmentation sale by each segment. Regular Restaurant, Soup & BBQ. Regular KTV are really perform well during the program as well.

****

In the HEINEKEN Business Model, we don’t have the distributors. But we have agent as our distributor. So here is the top agents that perform well during the tiger credential here. There are LENG Channa and KCV Distribution Co LTD. It’s correlated to the province and region as well. Since we used region to allocate the agent location and zone of sale area as well.

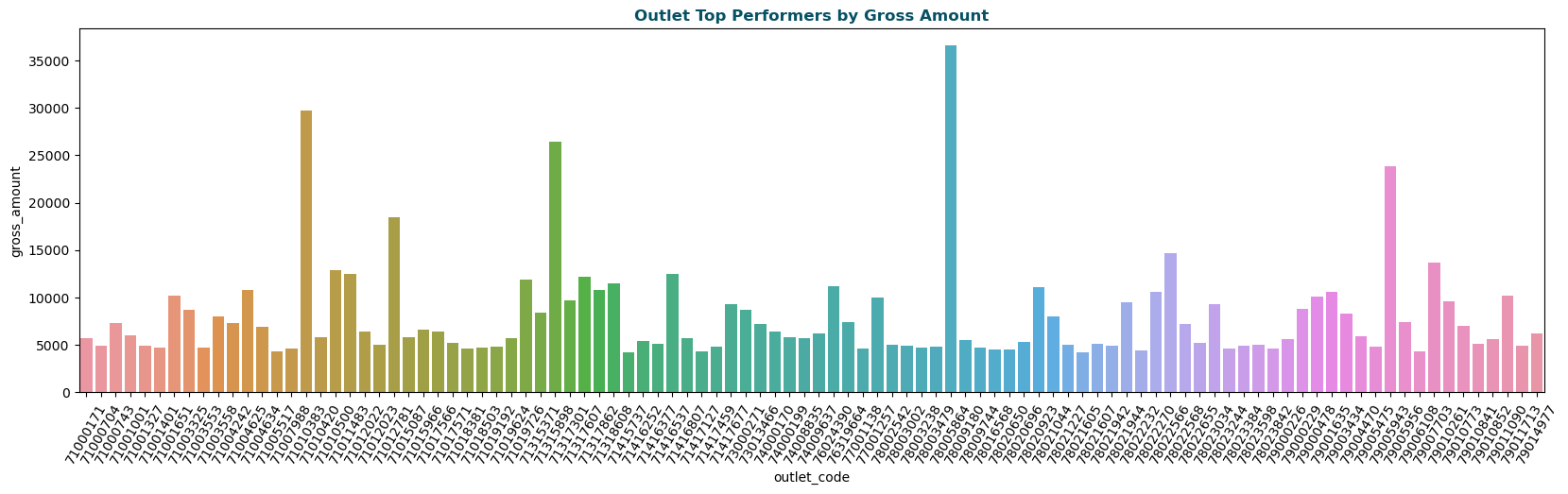
****

**DEEP ANALYTICS**

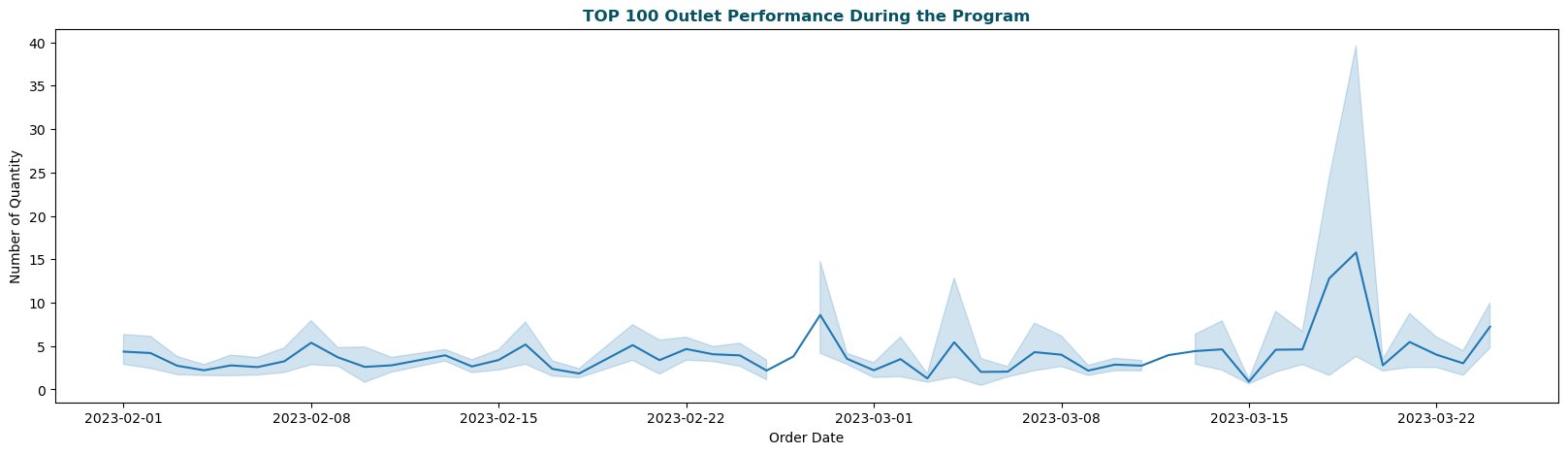
Often, driving execution at scale is the most difficult part of capturing the full value potential of analytics. To address this issue, the company developed a set of processes and tools to minimize complexity and workflow changes among them, guidelines for promotion calendars, a program to actively manage metrics around compliance and effectiveness, and a process that involved relevant functions (sales, for example) in insight generation and the development of promotion guidelines. The promotions team also worked with a revenue-growth-management center of excellence to develop and roll out a promotions playbook to select key-account teams, updated incentives for teams to use analytics to develop promotions, and communicated early success stories.

So on the performance analysis, we will do the comparison on the top performers and bottom performers around 100 outlets as the sample so see the customer behaviors and attribute and matrics that we will be used as the matrices for model recommendations developments

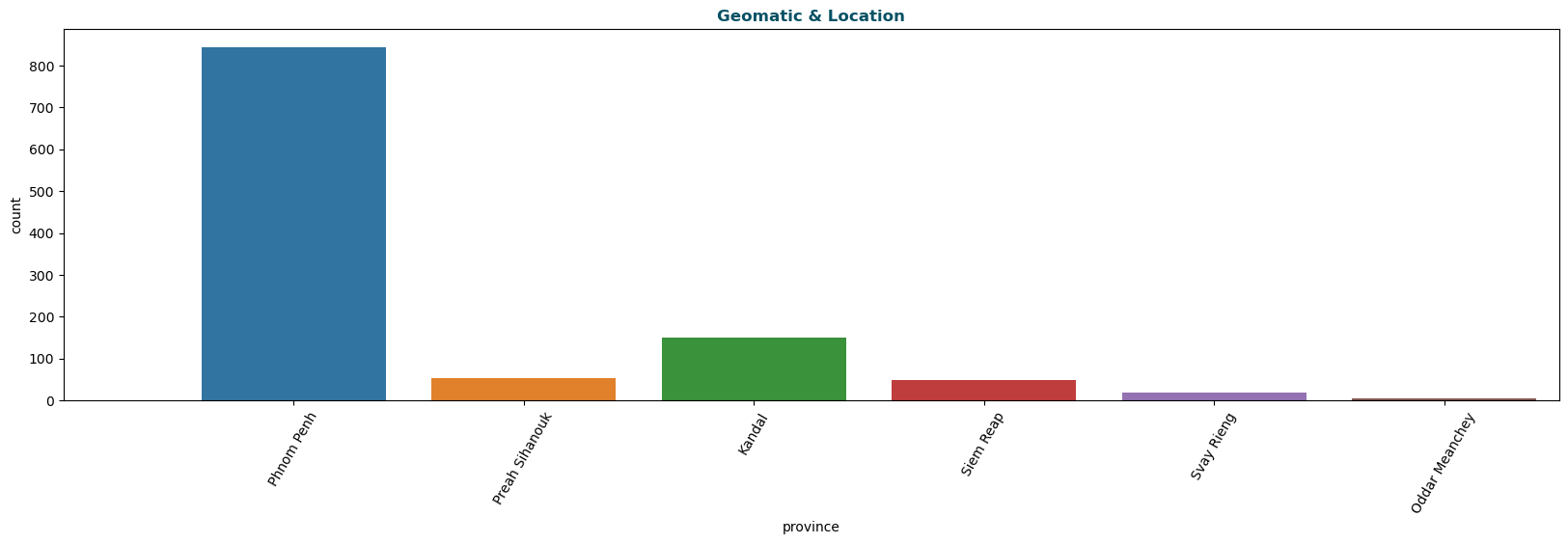
**TOP PERFORMERS ANALYSIS**

****

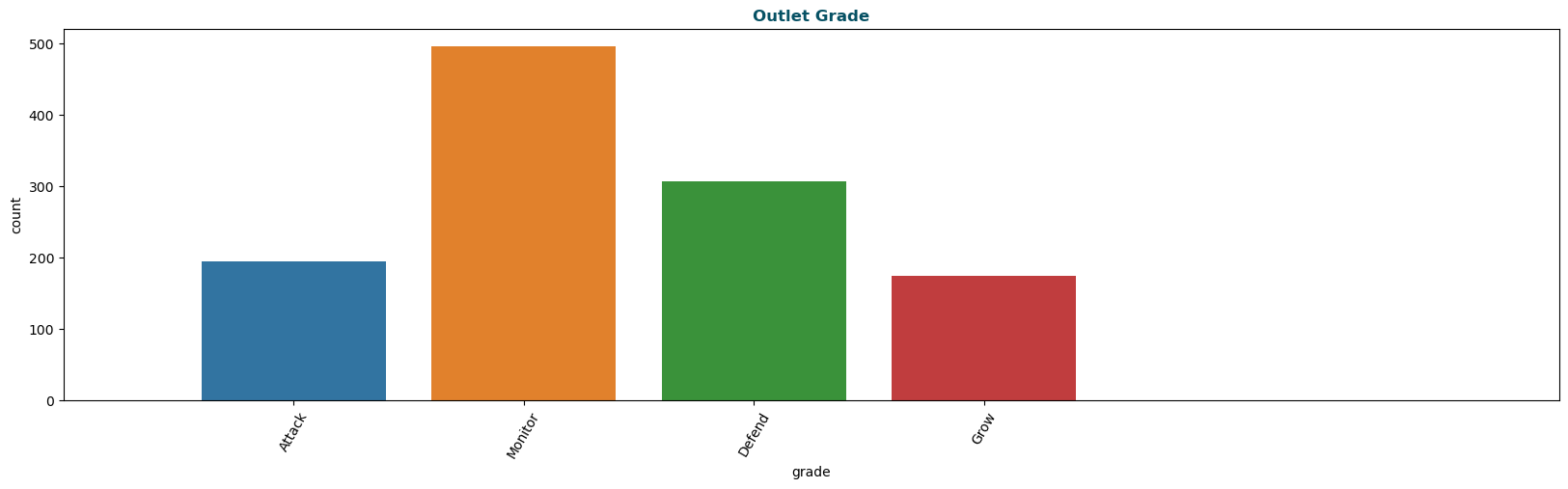
Top performers here is defined by gross amount values. With sales and marketing previously claiming all glory for revenue growth, customer centricity has now emerged from obscurity to claim its rightful place among the key drivers of revenue growth. It costs next to nothing. The process of improving customer service performance is largely iterative and involves taking an up-close look at the available data. Whether it’s gauging performance to cover bases in terms of customer retention or exploring new areas to upscale customer experience.

****

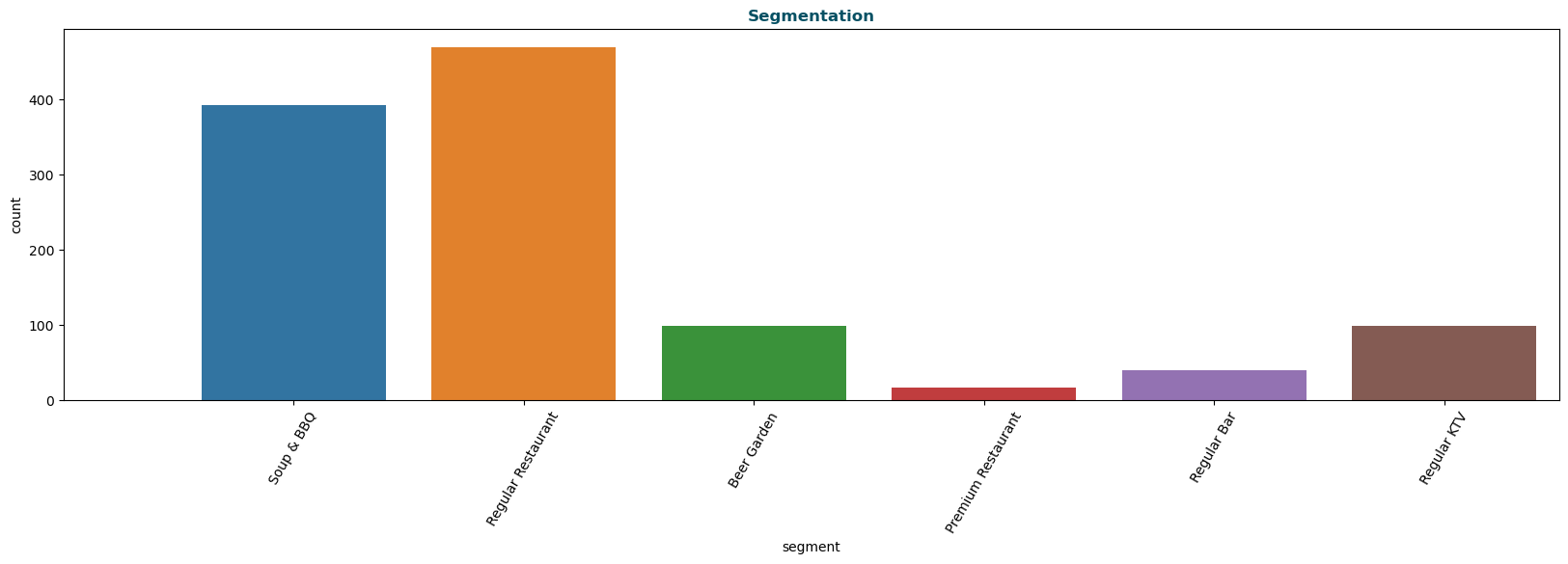
This is the order trade that top performers normally order with our application and from our sale team included. Around at the end of program, order trade mostly increased that most of outlet owners convinced by sale team to order and also, it’s running to finish the program as well. That’s why most of owner hurry buy in our product in from the program.

****

The location that also got top sale from top 100 outlet performers, that’s from Phnom Penh, Preah Sihanouk and Kandal. This is the location the performance well during the program.

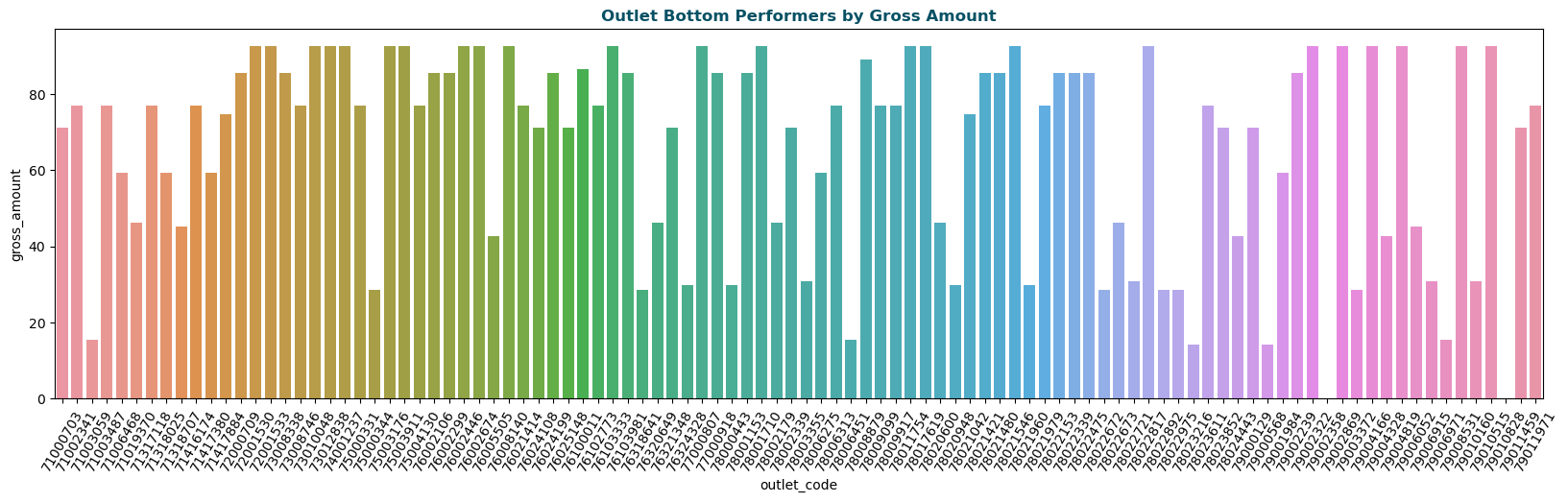
****

Monitor is still the best performance outlet grade of the population and here with top 100 outlet performers it's still well-perform too. This is came from the correlation too.

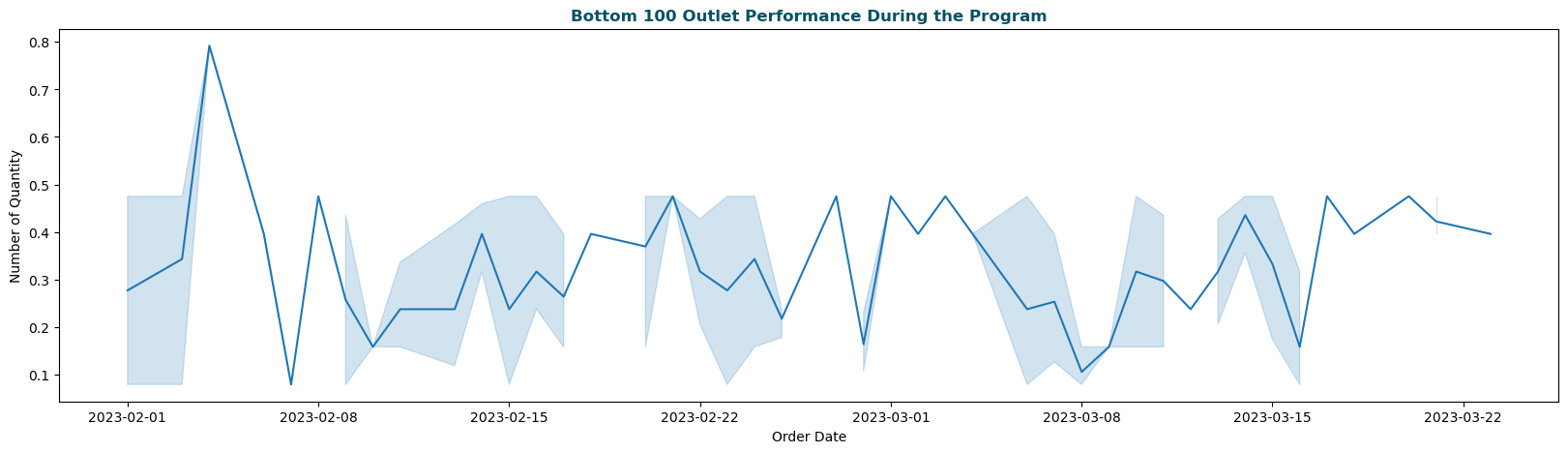
****

Here is indicating us the outlet segmentation sale by each segment. Regular Restaurant, Soup & BBQ. Regular KTV, and Beer Garden are really perform well during the program as well and top performers.

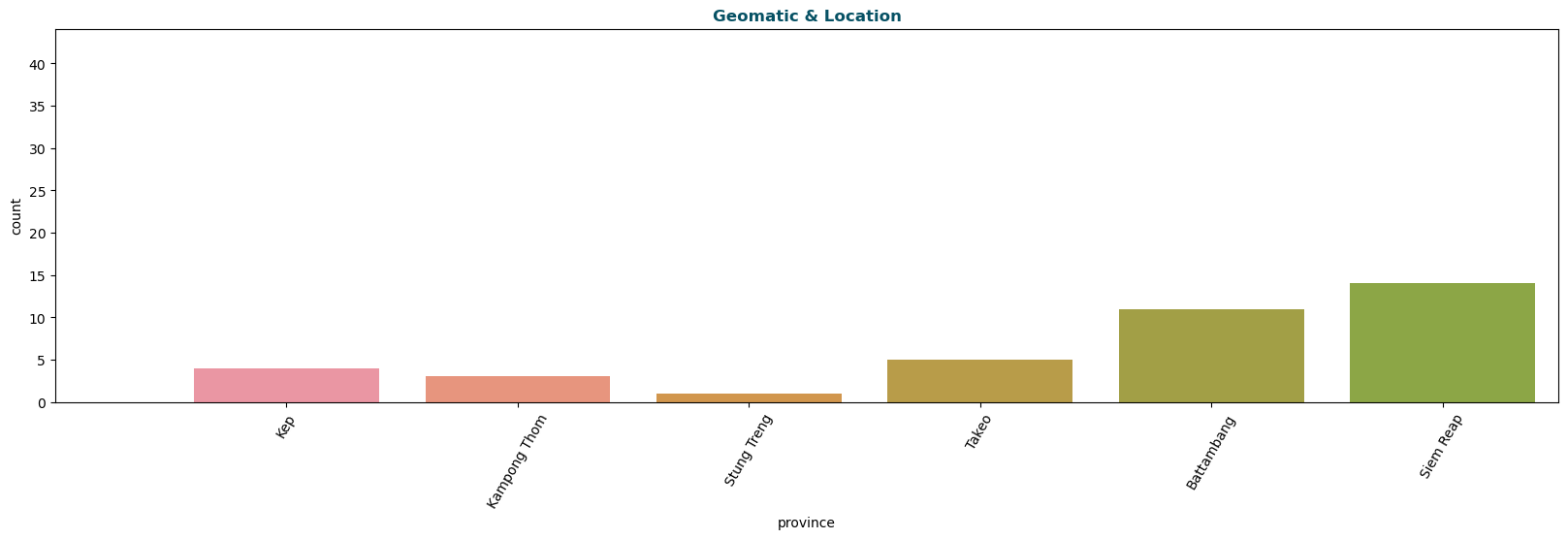
**BOTTOM PERFORMERS ANALYSIS**

****

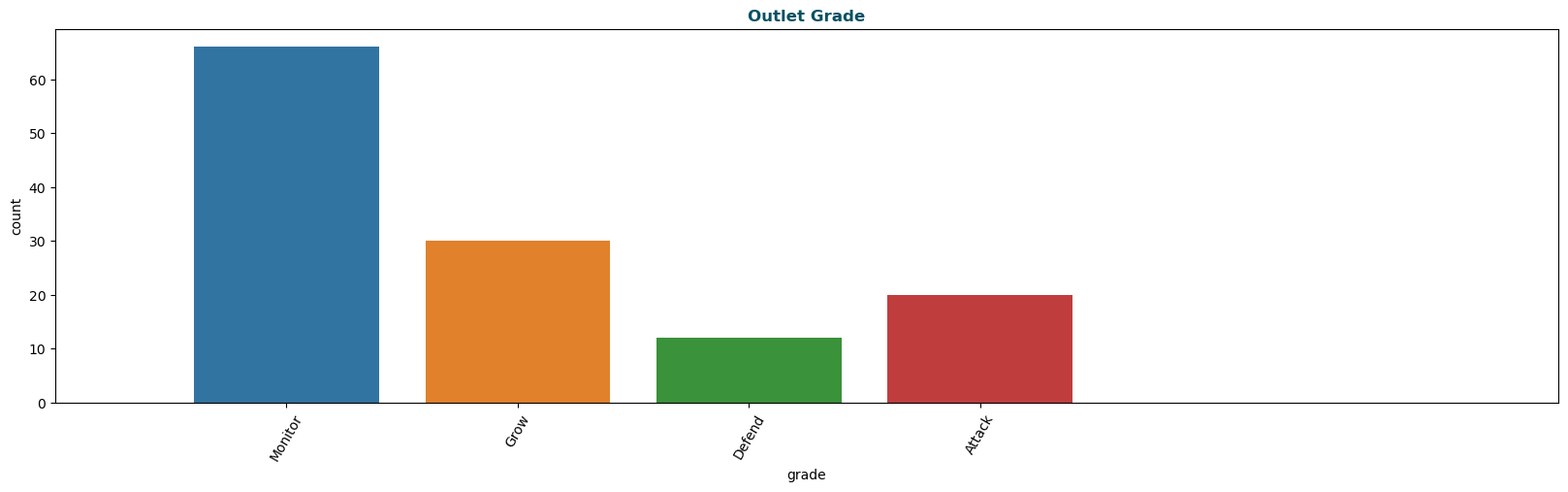
Bottom performers here is defined by low gross amount values. With sales and marketing previously claiming all glory for revenue growth, customer centricity has now emerged from obscurity to claim its rightful place among the key drivers of revenue growth. It costs next to nothing. The process of improving customer service performance is largely iterative and involves taking an up-close look at the available data. Whether it’s gauging performance to cover bases in terms of customer retention or exploring new areas to upscale customer experience.



This is the order trade that bottom performers normally order with our application and from our sale team included.

****

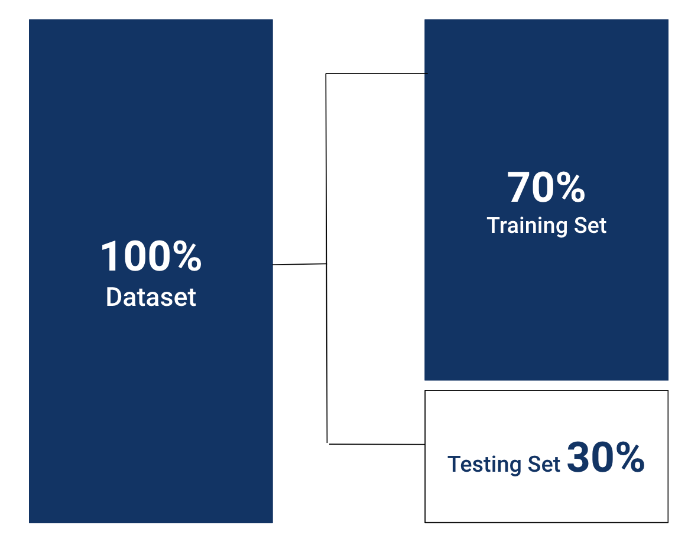
The big points different thing from top performers and bottom performers is the geo and location. What we see here in the top performance are from Phnom Penh, Preah Sihanouk and Kampot. But here mostly of bottom performers are from Siem Reap and Battambong. We are doing and finding the how and why this is happening and got different thing here. Is that behavior of sale team or internet that blocker the outlet owner hard to make order? Or they have limited knowledge to order by their own?

****

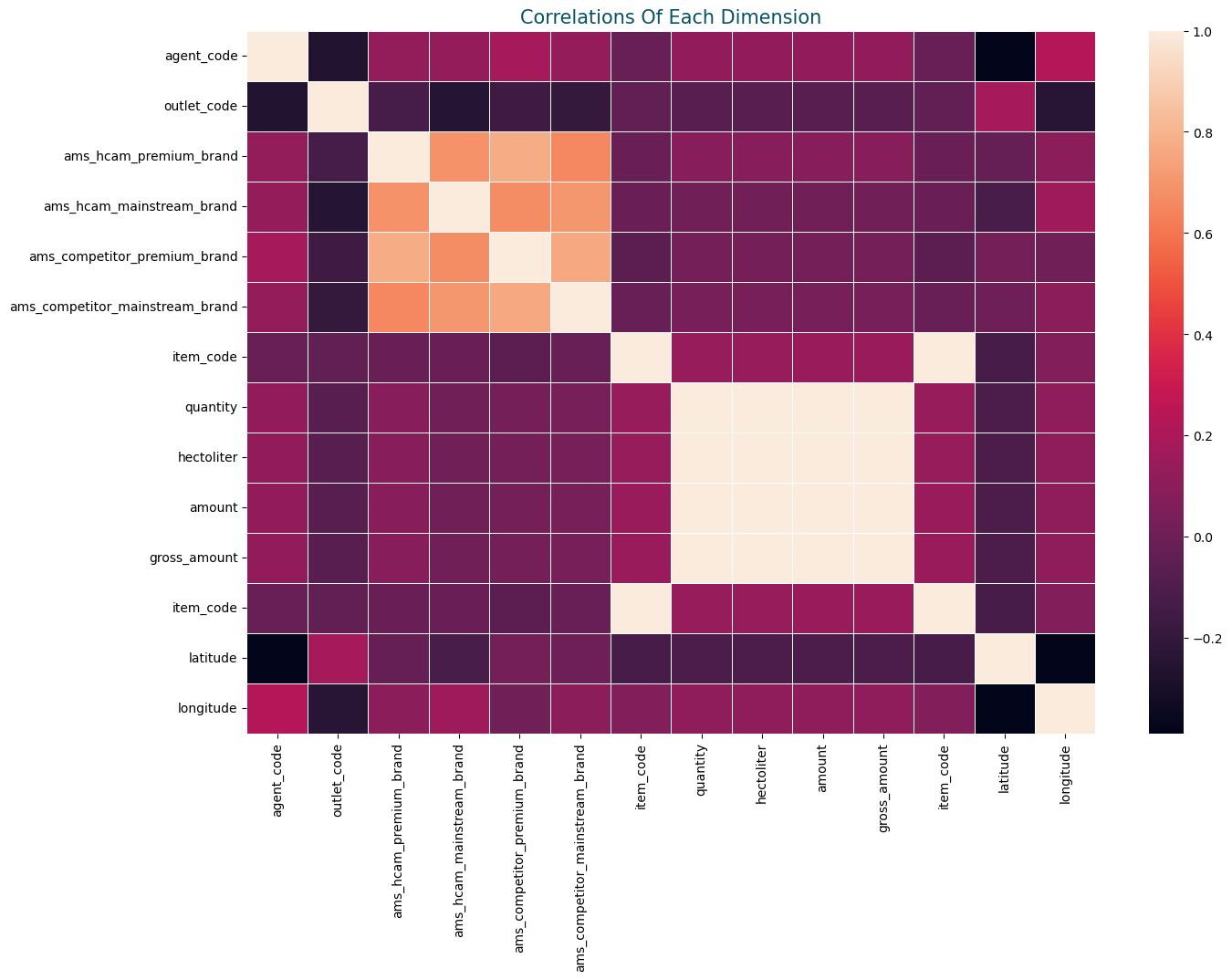
Monitor is still the best performance outlet grade of the population and here with top 100 outlet performers and bottom performers too. This is coming from the correlation too.

* 1. **Data Preprocessing**

First of all, I load libraries into the project. There are matplotlib.pyplot for plot configuration, numpy for numerical operations, pandas for database management and seaborn for data visualization. Second, I load the dataset into my project in order to start. I display the basic information by using the .info() function to see Range Index and column details. Then, I display statistical information using .describe() to see count, mean, std, min, 25%, 50%,75% and max value of the data. After that, I check if there is missing data and calculate its percentage. As I can see there are some features that have wrong data types, so I change those data types to the right one. Data preprocessing, a component of data preparation, describes any type of processing performed on raw data to prepare it for another data processing procedure. It has traditionally been an important preliminary step for the data mining process. Data preprocessing can refer to manipulation or dropping of data before it is used in order to ensure or enhance performance, and is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects. Before raw data could be sent through a machine learning model it has to undergo preprocessing. And it’s simply because data in the real world are generally Incomplete, Noisy and Inconsistent. And if this is fed into the machine learning model, results can come unexpectedly! And that’s not really what we want. Data preprocessing is a proven method for resolving such issues.

****

Data splitting is typically done to avoid overfitting. That is an instance where a machine learning model fits its training data too well and fails to fit additional data reliably. The original data in a machine learning model is typically taken and split into three or four sets. Here, we can split the dataset into 2 different training and testing (use in model evaluation).

****

Correlation coefficients measure the strength of the relationship between two variables. A correlation between variables indicates that as one variable changes in value, the other variable tends to change in a specific direction. Understanding that relationship is useful because we can use the value of one variable to predict the value of the other variable. For example, height and weight are correlated—as height increases, weight also tends to increase. Consequently, if we observe an individual who is unusually tall, we can predict that his weight is also above the average. In statistics, correlation coefficients are a quantitative assessment that measures both the direction and the strength of this tendency to vary together. There are different types of correlation coefficients that you can use for different kinds of data. In this post, I cover the most common type of correlation—Pearson’s correlation coefficient. Before we get into the numbers, let’s graph some data first so we can understand the concept behind what we are measuring.

To transform the categorical variable, we use the label encoder to make it out. Label Encoding is a technique that is used to convert categorical columns into numerical ones so that they can be fitted by machine learning models which only take numerical data. It is an important pre-processing step in a machine-learning project. Suppose we have a column Height in some dataset that has elements as Tall, Medium, and short. To convert this categorical column into a numerical column we will apply label encoding to this column. After applying label encoding, the Height column is converted into a numerical column having elements 0,1, and 2 where 0 is the label for tall, 1 is the label for medium, and 2 is the label for short height Limitation of label Encoding Label encoding converts the categorical data into numerical ones, but it assigns a unique number (starting from 0) to each class of data. This may lead to the generation of priority issues during model training of data sets. A label with a high value may be considered to have high priority than a label having a lower value. An attribute having output classes Mexico, Paris, Dubai. On Label Encoding, this column lets Mexico is replaced with 0, Paris is replaced with 1, and Dubai is replaced with 2. With this, it can be interpreted that Dubai has high priority than Mexico and Paris while training the model, but actually, there is no such priority relation between these cities here. Please kindly refer to the another real example below:

|  |  |
| --- | --- |
| Outlet Grade | After Label Encoder |
| Monitor | 0 |
| Grow | 1 |
| Defend | 2 |
| Attack | 3 |

* 1. **Model Training**

**K-MEAN Algorithm**

K-MEANS clustering is a 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 (cluster centers or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. k-means clustering minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult Weber problem: the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances. For instance, better Euclidean solutions can be found using k-medians and k-medoids. The problem is computationally difficult (NP-hard); however, efficient heuristic algorithms converge quickly to a local optimum. These are usually similar to the expectation-maximization algorithm for mixtures of Gaussian distributions via an iterative refinement approach employed by both k-means and Gaussian mixture modeling. They both use cluster centers to model the data; however, k-means clustering tends to find clusters of comparable spatial extent, while the Gaussian mixture model allows clusters to have different shapes. The unsupervised k-means algorithm has a loose relationship to the k-nearest neighbor classifier, a popular supervised machine learning technique for classification that is often confused with k-means due to the name. Applying the 1-nearest neighbor classifier to the cluster centers obtained by k-means classifies new data into the existing clusters. This is known as nearest centroid classifier or Rocchio algorithm.

**Kmeans** algorithm is an iterative algorithm that tries to partition the dataset into *K*pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to **only one group**. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster’s centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

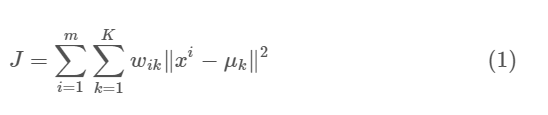
The way kmeans algorithm works is as follows:

1. Specify number of clusters *K*.
2. Initialize centroids by first shuffling the dataset and then randomly selecting *K*data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn’t changing.

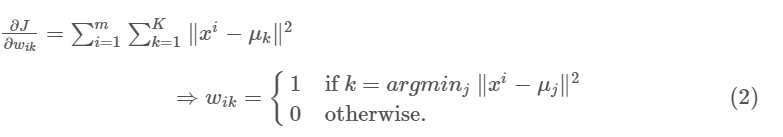
* Compute the sum of the squared distance between data points and all centroids.
* Assign each data point to the closest cluster (centroid).
* Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

The approach kmeans follows to solve the problem is called **Expectation-Maximization**. The E-step is assigning the data points to the closest cluster. The M-step is computing the centroid of each cluster. Below is a break down of how we can solve it mathematically (feel free to skip it).

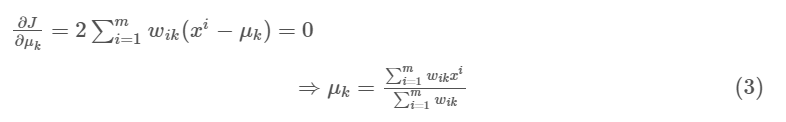
The objective function is:



where k=1 for data point xi if it belongs to cluster *k*; otherwise, wik=0. Also, μk is the centroid of xi’s cluster.



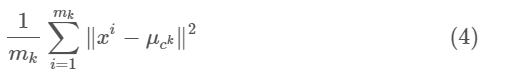
In other words, assign the data point xi to the closest cluster judged by its sum of squared distance from cluster’s centroid. And M-step is:



Which translates to recomputing the centroid of each cluster to reflect the new assignments.

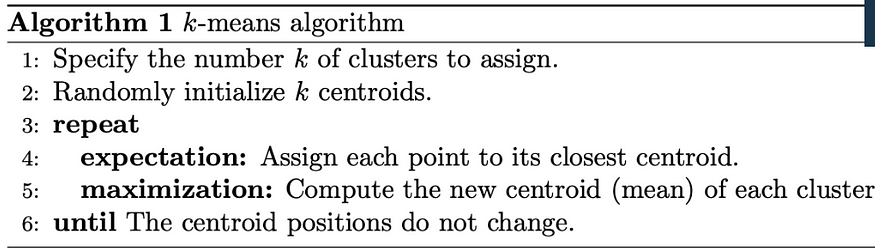
Few things to note here:

* Since clustering algorithms including kmeans use distance-based measurements to determine the similarity between data points, it’s recommended to standardize the data to have a mean of zero and a standard deviation of one since almost always the features in any dataset would have different units of measurements such as age vs income.
* Given kmeans iterative nature and the random initialization of centroids at the start of the algorithm, different initializations may lead to different clusters since kmeans algorithm may *stuck in a local optimum and may not converge to global optimum*. Therefore, it’s recommended to run the algorithm using different initializations of centroids and pick the results of the run that that yielded the lower sum of squared distance.
* Assignment of examples isn’t changing is the same thing as no change in within-cluster variation:



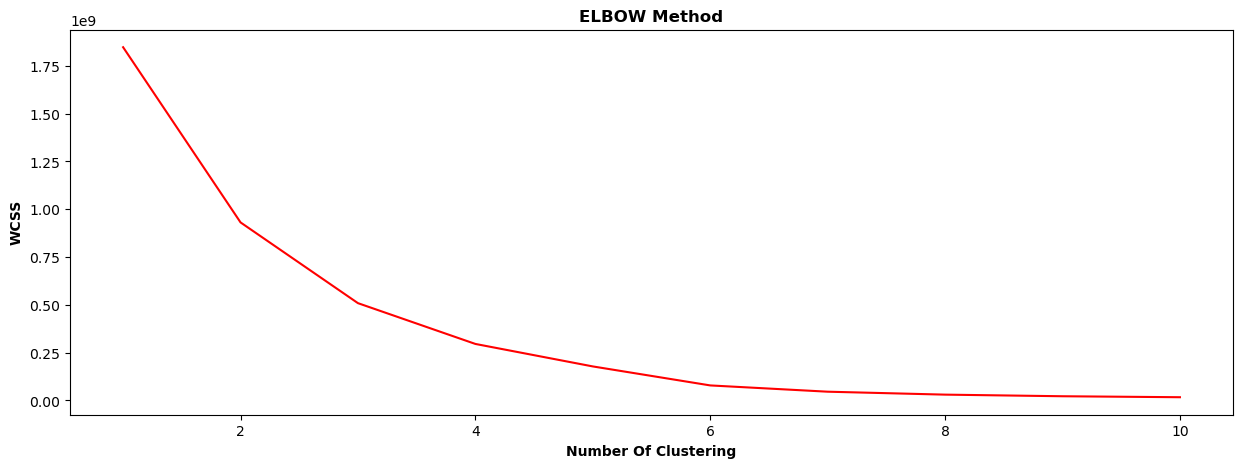
For this dataset, I’m assuming that we don’t have labels. Which is a general situation in real-life data. The best model for this scenario is K-means clustering. *Working of K-means clustering:*

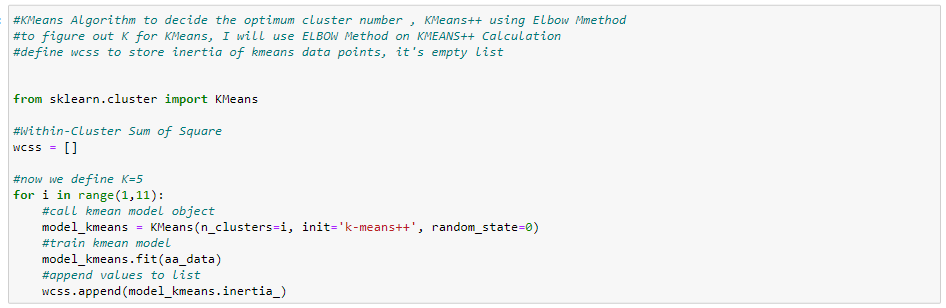
The K-means clustering is an Unsupervised Machine learning method that is used to identify clusters of data in the dataset. K-means clustering algorithms randomly select the k-number of the centroid (which we provide) which are used as the beginning point for every cluster and then perform this process iteratively to optimize the position of the centroids.

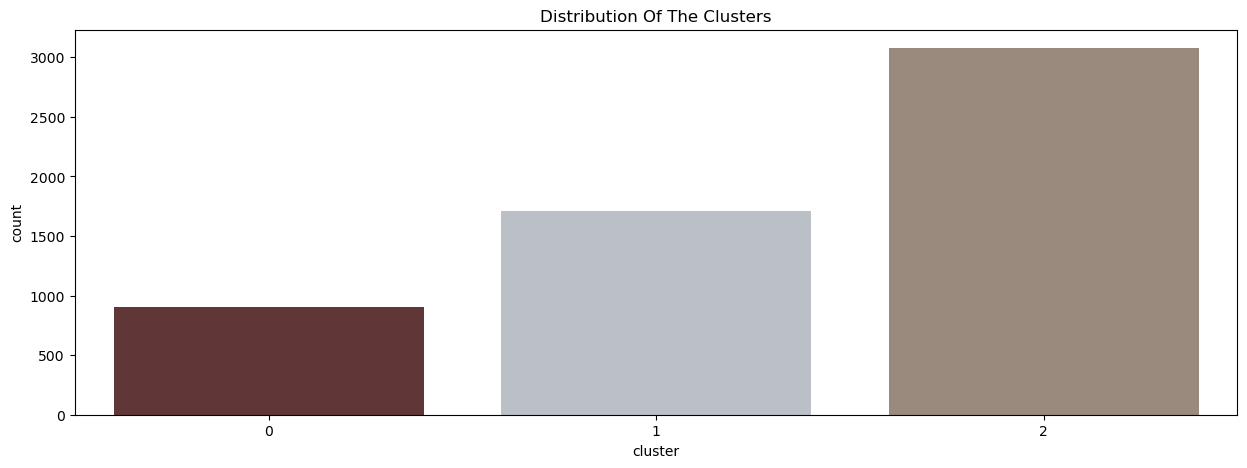


*Elbow Method:*

As mentioned above k-means algorithm initialize k centroids randomly how can we decide which is the optimal number for clustering for any dataset? Elbow method to rescue… Mathematically, the Elbow of the curve is a point where the curve visibly bends. The idea in segmentation, clusters will add much information since the data is consist of that many groups (actual cluster count), but once the number of clusters exceeds the actual number of groups in the data, the added information will drop sharply, because it is just subdividing the actual groups. Assuming this happens, there will be an elbow in the graph. Another important you need to learn before using any algorithm is the ‘Silhouette Score’. It is used for model evaluation.

****

****

****

**K Prototype clustering algorithm**

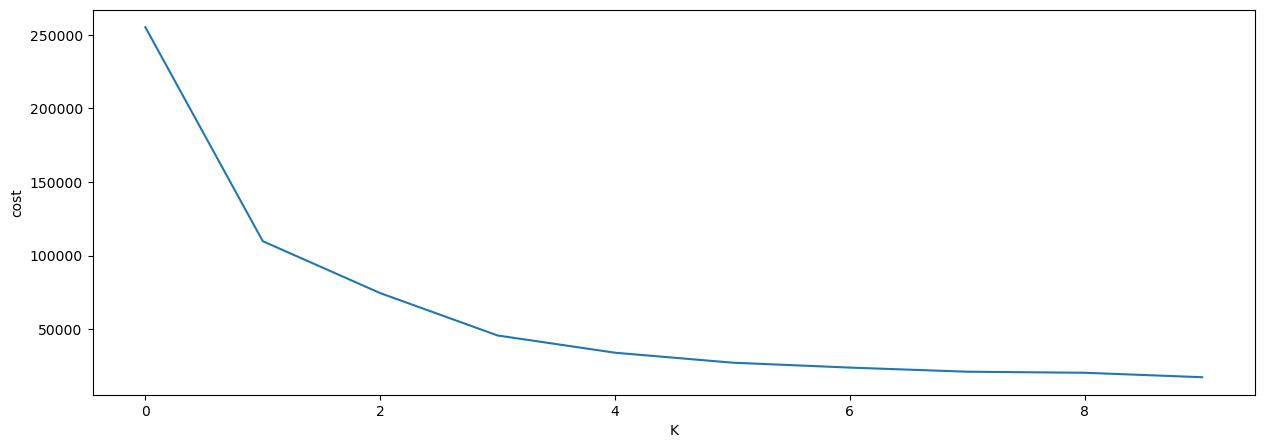
K-Prototypes clustering is a partitioning clustering algorithm. We use k-prototypes clustering to cluster datasets that have categorical as well as numerical attributes. The K-Prototypes clustering algorithm is an ensemble of k-means clustering and k-modes clustering algorithm. Hence, it can handle both numerical and categorical data.

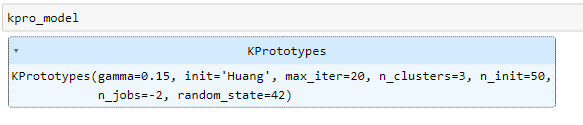
K-Prototypes clustering is a partitioning clustering algorithm. We use k-prototypes clustering to cluster datasets that have categorical as well as numerical attributes. The K-Prototypes clustering algorithm is an ensemble of k-means clustering and k-modes clustering algorithm. Hence, it can handle both numerical and categorical data. To understand the k-prototypes clustering in a better way, I would first suggest you read k-means clustering with a numerical example and k-modes clustering with a numerical example. In k-prototypes clustering, we select k-prototypes randomly at the start. After that, we calculate the distance between each data point and the prototypes. Accordingly, all the data points are assigned to clustering associated with different prototypes. After assigning data points to the clusters, we calculate the new prototype for the current cluster using the method discussed in the next sections. After that, we recalculate the distance of prototypes from the data points and reassign the clusters. This process is continued until the clusters converge. Before getting into the numerical example for k-prototypes clustering, Let us first discuss the distance measures and calculation of prototypes in the k-prototypes clustering algorithm.

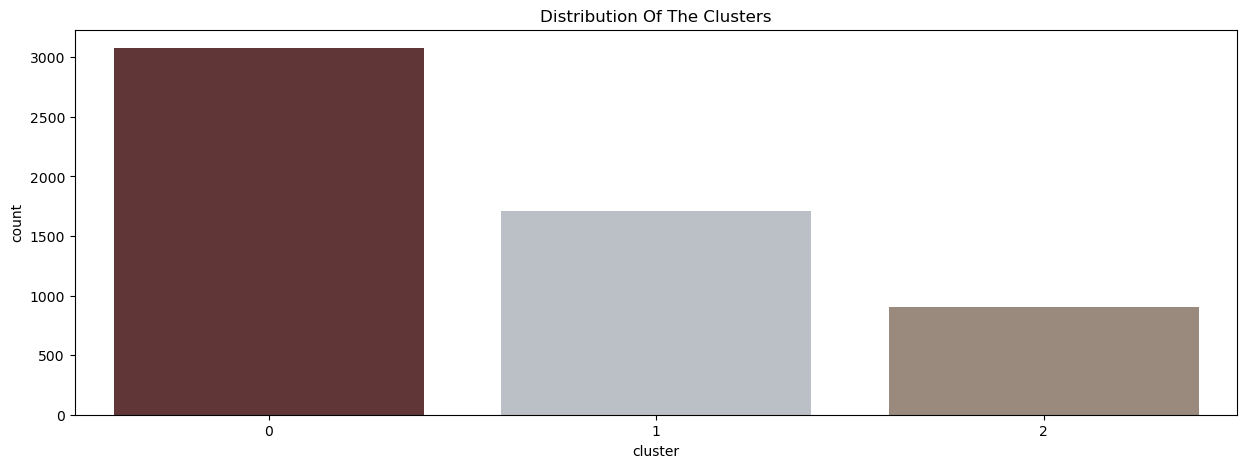
The K-prototypes clustering algorithm is similar to the k-means clustering algorithm. The difference lies in how we calculate the distance between two data points and how we select new prototypes for clusters. The steps in the k-prototypes clustering algorithm are discussed below.

* 1. First, we select K data points from the input dataset as initial prototypes.
  2. We then find the distance of each data point from the current prototypes. The distances are calculated as discussed in the previous sections.
  3. After finding the distance of each data point from the prototypes, we assign data points to clusters. Here, each data point is assigned to the cluster with the prototype nearest to the data point.
  4. After assigning data points to the clusters, we calculate new prototypes for each cluster. To calculate the prototypes, we take the mean of numeric attributes and the mode of categorical attributes as discussed previously.
  5. If the new prototypes are the same as the previous prototypes, we say that the algorithm has converged. Hence, the current clusters are finalized. Otherwise, we go to 2.

Now that we have introduced you to the k-prototypes clustering algorithm, let us discuss a numerical example of k-prototype clustering with a small sample dataset.

****

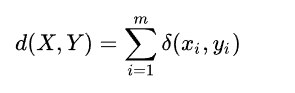
****

****

**KMode clustering algorithm**

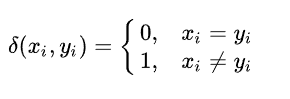
K-Modes clustering is a modified version of the standard k-means clustering process optimized to cluster categorical data. It does so by using the simple matching dissimilarity measure also referred to as the Hamming distance instead of the Euclidean distance to calculate the distance between two objects. Furthermore, it uses modes instead of means to represent the cluster centroids. While the k-means algorithm is a very popular choice when clustering numerical data, it performs poorly when applied to categorical data. The reason is that in order to cluster categorical data, the categorical values first have to be transformed into numerical values which distorts the clustering due to the usage of the Euclidean distance that leads the k-means algorithm to consider two distant values as close, simply based on the proximity of their numerical representations. One solution to this problem, although not suitable for high dimensional categorical data, is to convert the different categories to dummy binary variables and assign a 1 if the categorical value is present and a 0 else. A solution that works even with high dimensional categorical data is k-modes clustering.

The simple matching dissimilarity measure also referred to as the Hamming distance calculates the distance between two objects *X* and *Y* defined by *m* categorical attributes with

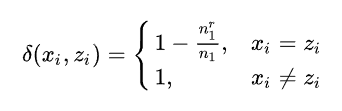
�(�,�)=∑�=1��(��,��) 

where

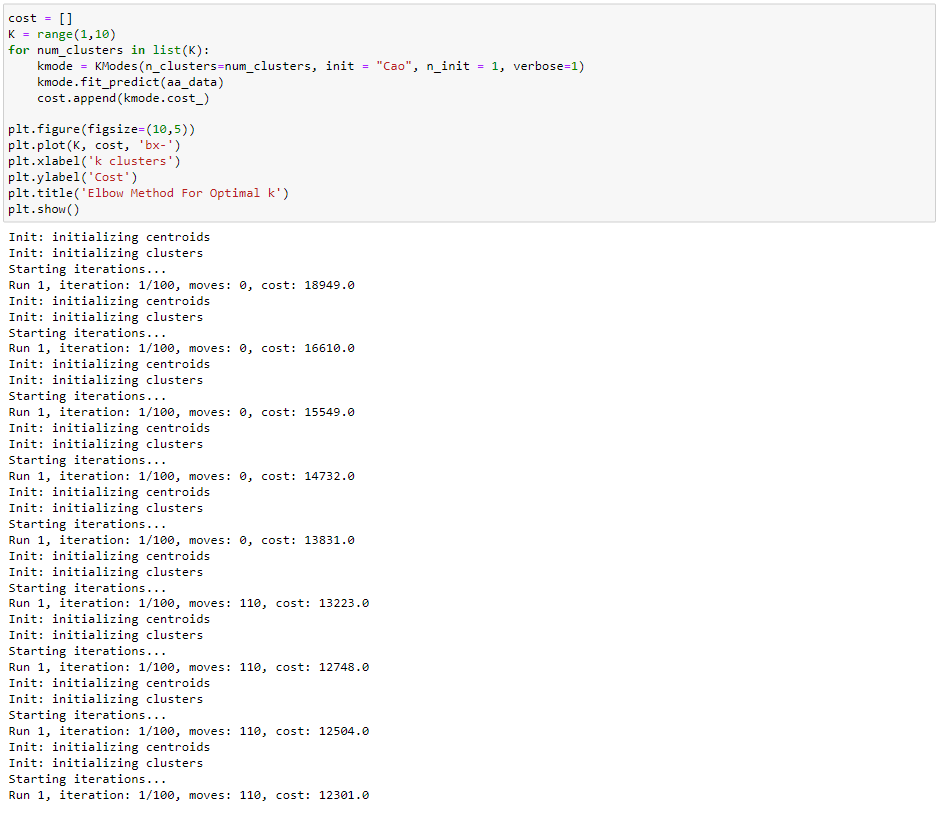
�(��,��)={0,��=��1,��≠��

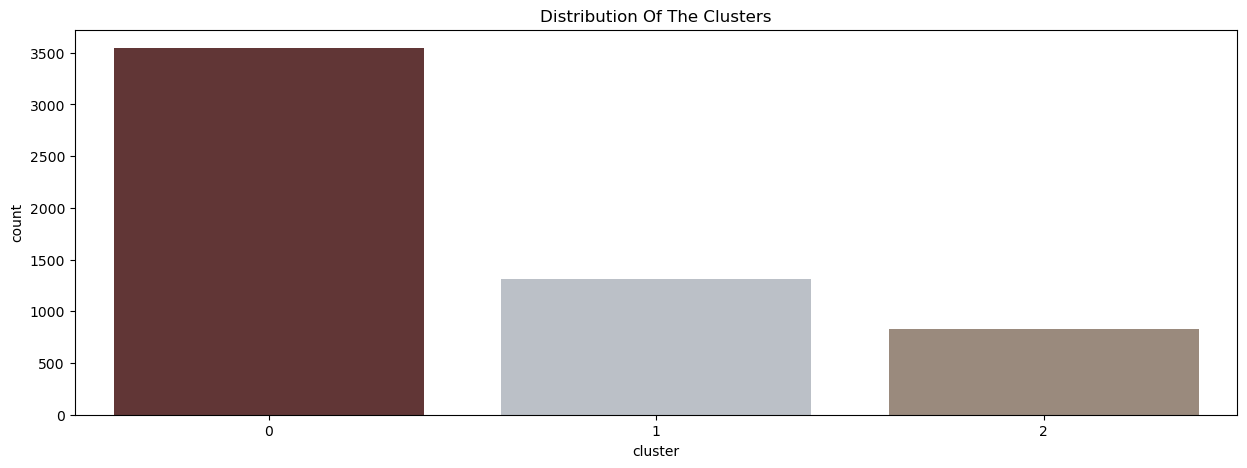


and *��Xi* and *��Yi Y*re the categorical values of attribute *i* in *X* and *Y*. The more mismatches there are between *X* and *Y*, the bigger the Hamming distance and thus the bigger the dissimilarity between the two objects. Applying this distance function to *k*-modes clustering, one obtains a new function that measures the distance and therefore the dissimilarity between an object *X* and a cluster

****

**Training the K-Modes**

****

****

1. **CONCLUSION** 
   1. **Conclusion**

Finally, I was very proud of myself that I can build this kind of ML model that is very useful when our world was meeting a pandemic virus that can be assisted to Cambodian students to improve their online learning. I would like to thank **HEINEKEN, Data & Analytic Team** which has contributed a lot to grow professionally and technically. Since the day I started to work with HEINEKEN, I know I’ve changed a lot that really helped me to become a better person inside and out. You are one of the reasons why I’m always trying my best. Thank you, my team! You know I appreciate you in so many ways. I feel very grateful to have had an opportunity to intern at this organization and for having you as my supervisor and mentor. You all were so patient with me. Thank you so much. I greatly appreciated that when training me you took the time to break things down to a level that I could understand and then build upon. Assigning work tasks to me that were sometimes quite difficult allowed me to have the chance to expand my knowledge and build confidence in my abilities. The valuable skills I’ve learned from you over the past year have given me the work experience. I appreciate everything and will remember my time here fondly.

Looking back, I want to give graduation thank so many heartfelt feelings of gratitude! You are awesome teachers in CADT. I have to pinch myself with excitement thinking about graduating from college. But there were still bittersweet moments as I cleared out my locker and helped you pack up your classroom supplies. I learned a lot during these past few years in school and you were an important part of my growth into a perfectional career.

Last but not least, I would like to thank once again **CADT** for giving me good hard skills and soft skills such as entrepreneurship, leadership, management that allow me easy to go to the real world.

* 1. **Difficulties**

The life of people is very hard and need to struggle sometimes, but the difficulty can make us stronger. Throughout my 4 months period of the internship, I am very needed to challenge big things in my life as you know I just finished year 2 and most I need to challenge other students to apply the job that they have more experience than me. Know much than me and stronger than me.

The main difficulty is technology is the very main point for this thesis that I have NOT implemented before. I have to do research and solve all the problems by myself with critical thinking and cleverness and also problem-solving. Adapting to the environment of the organization and learning new technology is what we have to do. As we know for CADT has NOT taught students about Machine Learning and Data Science before, I have only done the self-learning to improve and gain my skill in these things as well. Understanding algorithms of each machine learning model is the second obstacle that I must manage time to overcome by starting learning and asking from both google and seniors. Moreover, the second challenge is the data collection, it’s very hard to get the data from students, the data that we analyze is the primary data that we collected by ourselves and extract from some sources to increase the observation data to build the Machine Learning Model.

* 1. **Experience**

After finishing my internship, I have learned a lot about how to analyze data and insight data and wisdom to Machine Learning Building, Full Stack Data Science to build this project up and problem-solving skill either. In addition, I experienced working in a real environment and I feel employees that I have never experienced before. I feel so excited that I could do such an internship. Finally, I want to thank my advisor and everyone that supports and helps to complete this internship.

* 1. **Future Perspective**

For my future perspectives. I would like to apply new machine learning algorithms to do the experiment with the observation data, moreover, I will deploy the machine learning model that built and integrated these models with a mobile application and web application to predict the sentiment and emotion base on only their text. It takes less time to get someone's sentiment and emotion just inputting the text into the machine learning model and getting the output/result quickly and accurate result.

1. **REFERENCES**

Akhlaghpour, Hessameddin, Mohammad Ghodsi, Nima Haghpanah, Vahab S Mirrokni, Hamid Mahini, Afshin Nikzad. 2010. Optimal iterative pricing over social networks. International workshop on internet and network economics. Springer, 415–423

Lennart Baardman, Setareh Borjian Boroujeni, Tamar Cohen-Hillel, Kiran Panchamgam, Georgia Perakis, Detecting Customer Trends for Optimal Promotion Targeting

Evasaria M. Sipayung, Herastia Maharani, Benny A. Paskhadira, Designing Customer Target Recommendation System Using K-Means Clustering Method, IJITEE, Vol. 1, No. 1, March 2017

Sipayung E.M., Fiarni Cut, Tanudjaya R., “Decision Support System for Potential Sales Area of Product Marketing using Classification and Clustering Methods”, Proceeding International Seminar on Industrial Engineering and Management, 2015, pp. 33-39.

Sipayung E.M., Fiarni Cut, Tanudjaya R., “Modeling Data Mining Dynamic Code Attributes with Scheme Definition Technique”, Proceeding Electrical Engineering, Computer Science and Informatics, 2014, pp. 25-28.

Larosse, Daniel T., Discovering Knowledge in Data: An Introduction to Data Mining, Wiley-Interscience, 2005.

Kusrini and Lutfi E.T., Algoritma Data Mining, Yogyakarta, Indonesia: Andi, 2009.

Pavel, Berkhin, Survey on Clustering Data Mining Techniques. San Jose, CA: Accrue Software, 2002.

M. Kaur and U. Kaur, “Comparison Between K-Means and Hierarchical Algorithm Using Query Reduction,” International Journal of Advance Research in Computer Science and Software Engineering, Vol. 3, No. 7, pp. 1454-1459, 2013.

E. Turban, J. E. Aronson, and L. P. Ting, Decision Support System and Intelligent System, New Jersey: Pearson Education, Inc, 2005.

Davies, David L., Donald W. Bouldin, “Cluster Separation Measure”. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1(2), 95-104, 2013.