SPROCKET CENTRAL

KPMG DATA ANALYTICS VIRTUAL INTERNSHIP PROGRAM ON FORAGE

Final Draft Report

Task 1 – Data Quality Assessment

Task 2 – Data Insights

Task 3 – Data Visualization

Issued by: Nikita Tymoshenko

Document date: 19.08.2023

Contents

1. INTRODUCTION

The KPMG AU Data Analytics virtual internship offers an immersive learning experience designed to equip participants with practical skills in data analysis, visualization, and problem-solving. Through hands-on projects and real-world datasets, participants gain insights into the field of data analytics and its applications across various industries.

1.1. PROJECT BACKGROUND

Sprocket Central Pty Ltd[1], a medium size bikes & cycling accessories organization, has approached Tony Smith (Partner) in KPMG's Lighthouse & Innovation Team.

Relying on KPMG's expertise in its Analytics, Information & Modelling team, the Sprocket Central Pty Ltd needs help with its customer and transactions data. The organization has a large dataset relating to its customers, but their team is unsure how to effectively analyze it to help optimize its marketing strategy.

In order to support the analysis, the following stages are designed:

Phase 1 – Data Quality Assessment

The task requires to review the data quality to ensure that it is ready for our analysis in phase two. It is mandatory to issue the list of notes with any assumptions or issues that needed to go back to the client on. As well as recommendations going forward to mitigate current data quality concerns. The Data Quality Framework indicate a list of Data Quality Dimensions for evaluating the dataset.

Phase 2 – Exploratory Data Analysis

The stage aims to provide a comprehensive understanding of the dataset, offering valuable insights that can guide strategic decisions and drive business growth. The primary focus of this phase is to perform a detailed exploration of the dataset to identify correlations, distributions, and anomalies in the data.

Phase 3 – Data Insights and Visualization

The core analysis revolves around an RFM (Recency, Frequency, Monetary) analysis, a customer segmentation technique. This involves segmenting customers based on transaction behaviors and assigning scores to distinct categories.

Phase 4 – High-value customers prediction

By leveraging customer data, historical transaction data, and demographic information, the predictive models can be developed to identify and target high-value customers.

1.2. APPROACH AND TOOLS

Throughout this analysis, Python's pandas, seaborn, and matplotlib libraries play a pivotal role in manipulating, visualizing, and gaining insights from the dataset.

[1] Sprocket Central Pty Ltd is a fictional company named for the purposes of this virtual internship

2. DATA QUALITY ASSESSMENT

As part of the KPMG AU Data Analytics virtual internship, one of the tasks assigned was focused on data cleaning and preparation. The objective was to perform a comprehensive data quality assessment and cleansing process to ensure the dataset was accurate, complete, and ready for subsequent analysis.

2.1. DATASET AND DATA QUALITY DIMENSIONS

Sprocket Central Pty Ltd has a large dataset relating to its customers, but their team is unsure how to effectively analyze it to help optimize its marketing strategy.

The client provided KPMG with 4 datasets:

- Customer Demographic
- Customer Addresses
- Transaction's data
- NewCustomerList

The following list of the Data Quality dimensions has been used to evaluate dataset: Accuracy, Completeness, Consistency, Currency, Relevancy, Validity, Uniqueness.

The following sub-chapters describe outputs of the preliminary data exploration and identify ways to improve the quality of Sprocket Central Pty Ltd.'s data.

2.2. ASSESSMENT RESULTS

The main characteristics of given datasets are shown it the table below:

Table 1 – Data Profiling

Dataset	# of records	# of unique records	# of columns	percentage of missing values
Transactions	20,000	20,000	13	< 1%
NewCustomerList	1,000	1,000	23	1.4%
CustomerDemographic	4,000	4,000	13	3.4%
CustomerAddress	3,999	3,999	6	0%

Transactions table contained 20,000 records providing information on transactions made by 3,494 distinct customers for 101 distinct products and 6 brands from the year 2017. No duplicate transactions were found. The following data quality issues has been defined during assessment:

- 358 records (2% of transactions) had unspecified *online_status*, which can be filled as 'unspecified' to keep those transactions for further analysis.
- 197 records (1% of transactions) represent missing product attributes (brand, size, class, standard costs) and could be removed from the dataset.
- more information needed on what the column *product_first_sold_date* refers to.
- it should be mentioned that there is no column for quantity sold was observed.

CustomerDemographic table represents data related to 4,000 customers indicating names, genders, birthdates, job titles and other information. The following data quality issues has been defined during assessment:

- the *gender* column contains not allowable values (misspelling and different format) which can be replaced with F/M/U
- the following attributes contain missing values: *job_title* (497 customers), *job_industry_category* (656 customers), *last_name* (125 customers). Depending on analysis purposes it is possible to keep those customers replacing blank values with 'unspecified' category.
- DOB doesn't match the range constraint for 1 customer with 1843-12-21 indicated. Also, DOB is missing for 87 customers, for whom data in *default* and *tenure* columns are also missed.
- the column *default* is not interpretable thus cannot be used for further analysis.

CustomerAddress table provides data related to addresses, postcodes, states and countries for 3,999 customers referring to *customer_id* as foreign key. The following data quality issues has been defined during assessment:

- the *state* column doesn't meet the validity requirements due to different approaches in state naming. It`s possible to replace values and bring the data into one standardized format (e.g., NSW, QLD etc.)
- the following IDs are missing when *customer_id* used as foreign key and data has been merged with **CustomerDemographic** table: 3, 10, 22, 23, 4001, 4002, 4003.
- more information needed on what the column *property_valuation* refers to.

NewCustomers table expanded **CustomerDemographic** data with 1,000 new customers. The following data quality issues has been defined during assessment:

- there is no ID column for customers which could be used as primary key for further analysis.
- DOB data is missed for 17 customers.
- there are 5 unnamed columns contained numeric data with a lack of context.

The Data Quality Assessment results are shown in the table below:

Table 2 – Data Assessment Matrix

Dataset	Accuracy	Completeness	Consistency	Currency	Relevancy	Validity	Uniqueness
Transactions	✓	✗	✓	✓	✗	✓	✓
NewCustomerList	✓	✗	✓	✓	✗	✗	✓
CustomerDemographic	✗	✗	✗	✓	✗	✗	✓
CustomerAddress	✓	✓	✗	✓	✗	✗	✓

As a conclusion, the following mitigations can be applied to improve the accuracy of the underlying data:

- Empty values within core fields can be entirely filtered out from the resulting set if imputing is irrelevant;
- For categorical fields empty values can be replaced with 'unspecified' category depending on analysis purposes;
- Regular expressions can be used to replaced extended values into abbreviations to ensure consistency;
- Appropriate data transformation can be made to ensure consistent data types for a given field;

2.3. RECOMMENDATIONS FOR IMPROVING DATA QUALITY

The data quality assessment revealed notable issues in the dataset, prompting the implementation of effective strategies to address these inconsistencies. In light of these findings, a set of recommendations has been developed to proactively prevent future data quality issues and enhance the precision of the foundational data crucial for informed business choices.

- Root-cause analysis

Identifying the root causes of data inconsistencies is a critical step in ensuring the accuracy and reliability of the data. The following approaches can be used for determining root causes:

  - mapping the end-to-end data flow,
  - tracking the origin and transformations of data across different systems,
  - examining data sources,
  - thoroughly analyzing the data through data profiling techniques.

By understanding the underlying reasons for data inconsistencies, organization can implement targeted and effective solutions to address them.

- Data Quality Rules

Data quality rules and thresholds play a crucial role in assessing and maintaining the integrity of organizational data with the following key steps:

  - identifying the most critical data attributes for further decision-making process,
  - defining quantifiable quality metrics for each data attribute (data quality dimensions),
  - establishing clear thresholds or ranges for each quality metric,
  - developing validation rules to assess data quality,
  - automating data quality assessment process whenever possible.

In conclusion, establishing data quality rules and thresholds is a foundational step in maintaining trustworthy and valuable data assets.

- Maintain and update Issue log

An issue log serves as a centralized repository to document and manage identified data quality issues. It provides a structured approach to tracking, addressing, and resolving issues, ensuring that data remains accurate and reliable. The following components can be used to properly maintain the issue log:

  - definition of data quality issue,
  - identification of the underlying causes of the issue,
  - assessment of potential impact on business operations,
  - assigned responsibility for addressing and resolving the issue,
  - resolution steps and planned activities,
  - progress of issue resolution status
  - timestamp recorded for issue identification, assignment, resolution and closure.

By proactively addressing data quality challenges, organizations enhance their ability to make well-informed decisions and drive successful outcomes.

3. EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is a crucial step in the data analysis process, including when dealing with transactional data. EDA helps to gain insights, identify patterns, understand the structure of the data, and uncover potential issues.

3.1. DATASET SIZE AND STRUCTURE

The dataset is the result of comprehensive data cleaning and the integration of the transactions fact table with relevant dimension tables (demographic, address). This dataset is well-prepared for further analysis, insights extraction, and modeling. Several steps were taken to clean and prepare the data:

- Missing values were handled using appropriate methods such as imputation or dropping.
- Spelling mistakes, data inconsistencies, and anomalies were corrected.
- Irrelevant and not interpretable columns dropped out.
- Calculated fields added (profit, age, transaction months etc.)
- Merging data into a single dataset for further analysis

The final dataset consists of 19,319 records and 27 fields and contains data on approved transactions made by 3,411 distinct customers out of 3,494 customers given in the original transaction's dataset, which is 97.6%. Among transactional data the dataset represents a comprehensive collection of customer-related information. The following tables presents short description of dataset`s attributes:

Table 3 – Dataset

Attribute	Description	Comment
transaction_id, product_id, customer_id	Unique ID for each transaction, purchased product and customer associated with the transaction	19,319 transactions, 101 products and 3,411 customers
transaction_date	Date of the transaction	Year 2017
online_order	Whether the order was placed online	'1'/'0'/'unknown'
brand	Brand of the product	6 brands
product_line	Line of the product	'Standard', 'Road', 'Touring', 'Mountain'
product_class	Classification of the product	'medium', 'high', 'low'
product_size	Size category of the product	'medium', 'large', 'small'
profit	Profit generated from the transaction	Calculated as (list_price – standard_cost)
gender	Gender of the customer	'M'/'F'
past_3_years_bike_related_purchases	Number of purchases	Given within the original dataset
job_industry_category	Industry category of the customer's job	10 industry categories
wealth_segment	Customer's wealth segment	'Mass, 'Affluent, 'High Net Worth'
owns_car	Whether the customer owns a car	'Yes', 'No'
age	Age of the customer.	Calculated using given DOB
postcode	Postal code of the customer	Australian postcodes
state	State where the customer resides	'VIC', 'NSW', 'QLD'
property_valuation	Valuation of the customer's property	From 1 to 12

3.2. SUMMARY STATISTICS FOR NUMERIC COLUMNS

Basic descriptive statistics help to understand the distribution of values, identify outliers, and get a sense of the overall characteristics of the dataset. Summary statistics for numerical columns (mean, median, min, max, standard deviation) are given in the following table:

Table 4 – Basic descriptive statistics for numerical columns

descriptive statistics	count	mean	std	min	25%	50%	75%	max
list_price	19,319	1,106.29	582.86	12.01	575.27	1,163.89	1,635.30	2,091.47
standard_cost	19,319	555.76	405.69	7.21	215.14	507.58	795.10	1,759.85
profit	19,319	550.53	492.92	4.80	133.78	445.21	827.16	1,702.55
past_purchases	19,319	48.92	28.64	0.00	24.00	48.00	73.00	99.00
tenure	19,319	10.68	5.67	1.00	6.00	11.00	15.00	22.00
age	19,319	45.49	12.61	21.00	36.00	45.00	55.00	91.00
property_valuation	19,319	7.52	2.83	1.00	6.00	8.00	10.00	12.00

3.3. PRODUCTS AND BRANDS

Product analysis involves examining the characteristics and performance of the products within the dataset. The main focus maid on the popularity of products, their profitability and customer preferences. The overall product performance can be seen on the following figure:

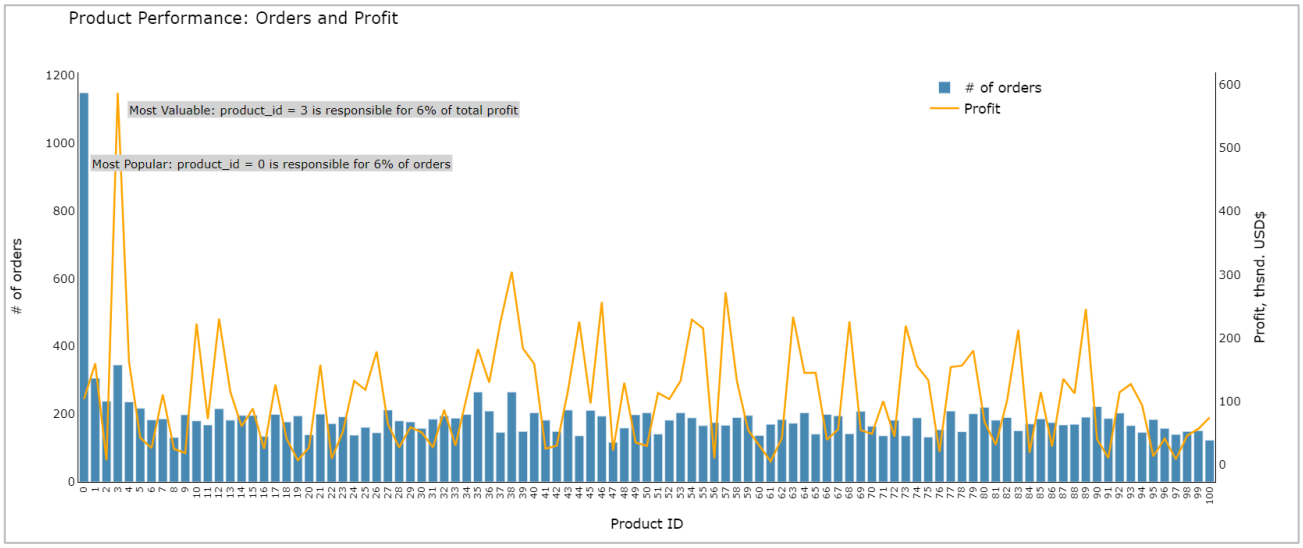

Figure 1 – Product Performance: Popularity and Profit

As can be seen on the figure above, product_id=0 stands out as the most popular item, contributing to a significant 6% of all orders. However, its contribution to the total income remains below 1%, emphasizing its wide customer appeal. On the other hand, product_id=3 emerges as the most valuable asset, accounting for an impressive 6% of the total income, reflecting its significant contribution to the organization's revenue stream.

The analysis reveals that the brand "Silex" is the most popular among customers, capturing the highest number of orders. On the other hand, "WeareA2B" emerges as the most valuable brand in terms of total profit generated. The Norco Bicycles considered to be the less popular brand with the least contribution to the profit.

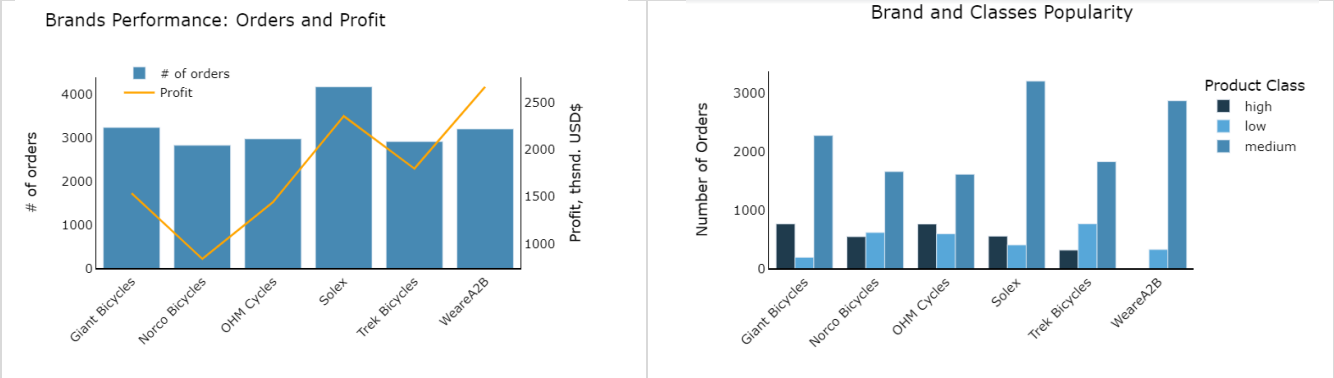

Figure 2 – Brands Performance: Popularity and Profit

Additionally, the data indicates that the "medium" product_class holds both the highest popularity and value across all brands, making it a noteworthy choice for potential marketing strategies.

3.4. SALES

Transactional sales analysis involves analyzing data related to individual sales transactions to extract valuable insights and make informed business decisions.

3.5. CUSTOMERS

Customer analysis involves studying customer behavior, preferences, and interactions to gain insights that can drive marketing, sales, and customer service strategies.

4. RFM ANALYSIS

RFM analysis is a powerful technique used by businesses to segment and understand customer behavior based on their transactional data. RFM stands for Recency, Frequency, and Monetary Value, which are three key dimensions used to evaluate and segment customers. It provides valuable insights into customer engagement and loyalty, and helps businesses identify and prioritize different customer segments for targeted marketing efforts, personalization, and retention strategies.

4.1. APPROACH

The analysis performed delving into three key dimensions of customer behavior:

- Recency (how recently a customer has made a purchase)

Recency for each customer is calculated by measuring the time between the most recent transaction date from the dataset and the date of the last customer`s transaction.

- Frequency (how frequently a customer makes purchases)

Frequency is determined by counting the total number of transactions each customer has made over a period of analysis.

- Monetary (how much monetary value a customer contributes)

Monetary value is calculated by summing the total spending of each customer.

Each described dimension is divided into quartiles to create segments based on distribution. Customers are ranked and assigned quartile scores from 1 to 4, with 4 being the highest score representing the most recent, frequent, or valuable customers accordingly. The individual R, F, and M scores are combined to create an RFM score.

By determining RFM scores through this approach, businesses gain a comprehensive understanding of customer engagement and value, enabling them to make informed decisions and implement strategies that drive customer loyalty, retention, and overall business success.

4.2. INSIGHTS AND RECOMMENDATIONS

Customers are segmented based on their RFM scores, resulting in distinct groups that represent varying levels of engagement and value. These segments provide insights into customer behavior and preferences, helping tailor marketing and engagement strategies.

The RFM matrix followed by RFM distributions are presented in the **Figure 3** below. The heatmap clearly highlights the distribution of customers across different Recency, Frequency, and Monetary score combinations. This segmentation provides a comprehensive view of customer behavior and engagement.

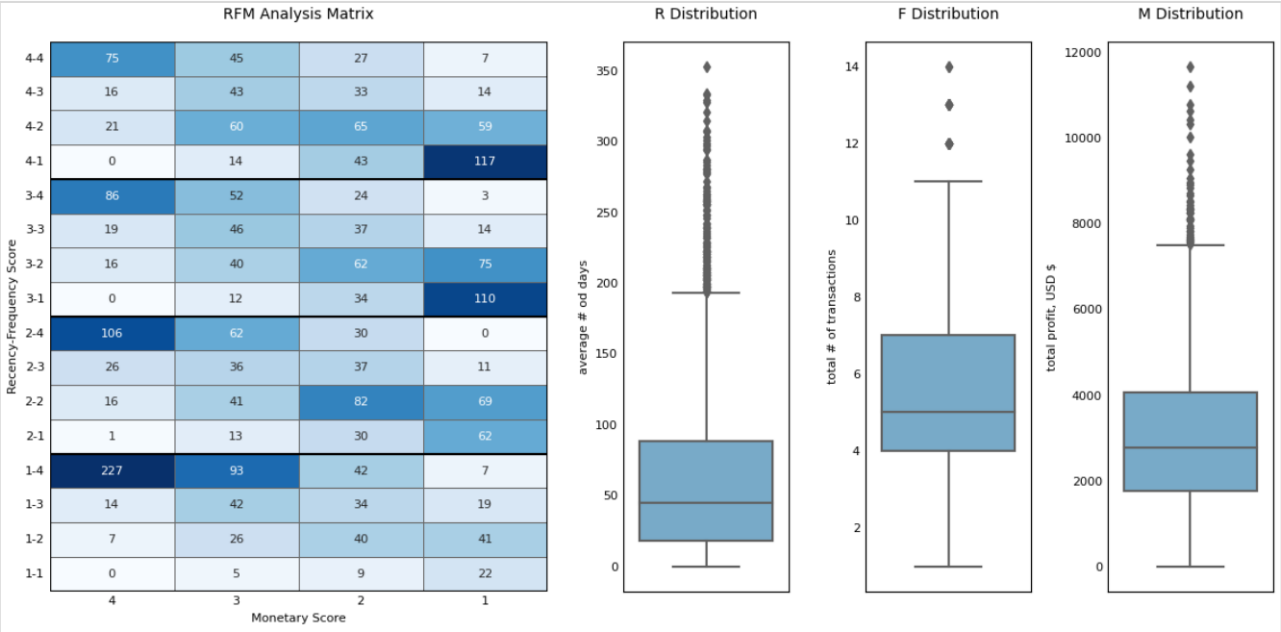

Figure 3 – RFM Matrix and Distribution

The following observations suggest interesting and valuable segments of customers:

- Segments 1-4-4 and 2-4-4 (13% of all customers)

These high-monetary customers might have made large purchases in the past on a regular-basis, indicating their loyalty and potential satisfaction with the products or services. However, their lack of recent transactions could indicate dormancy. It is highly recommended to develop a re-engagement strategy to reconnect with these dormant high-value customers based on the activities listed below.

Re-Engagement Strategy:
- sending personalized emails, offers, or promotions to encourage them to return and make new purchases;
- providing exclusive offers or discounts to entice these customers back;
- collecting feedback to identify any issues, concerns, or changes in customer`s preferences;
- designing win-back campaigns specifically targeted at this segment;

- Segments 3-1-1 and 4-1-1 (9% of all customers)

Customers in this segment have made recent transactions, but they do so infrequently and with relatively low monetary value. It is highly recommended to develop a strategy to encourage repeat transaction based on activities listed below:

Strategy to encourage repeat transaction:
- implementing personalized email campaigns to encourage making additional purchases sooner;
- offering loyalty rewards or discounts to incentivize these customers to make repeat purchases;
- providing tailored product recommendations that align with customers interests;
- introducing a membership program that offers benefits to members;
- creating time-sensitive promotions that encourage quick follow-up purchase;

It is recommended to develop targeted strategies based on the RFM segments to engage and retain customers effectively. The strategy must reflect directions for marketing campaigns, promotions, and personalized offers which are designed for each segment to optimize customer interactions.

5. COHORT RETENTION ANALYSIS

Cohort retention analysis is a powerful method used to understand and evaluate customer retention over time, providing insights into the long-term value of customers and the effectiveness of business strategies. It involves grouping customers based on a specific time period (the cohort), usually their first interaction or purchase, and then tracking their subsequent behavior, such as repeat purchases or engagement rates, over multiple time intervals.

5.1. APPROACH

Cohorts are created based on the month of customer`s first transaction. This groups customers who share a similar initial experience, allowing you to analyze their retention patterns. Retention rate is calculated by dividing the number of customers who make repeat purchases by the total number of customers in the initial cohort. Heatmaps and retention curves can be used to visually represent how cohorts evolve over time. The main task within the cohort retention analysis is to identify cohorts with the highest and lowest retention rates, as well as understand what factors contribute to corresponding behavior.

5.2. INSIGHTS AND RECOMMENDATIONS

The heatmap below presents the retention ratio for customer`s cohort.

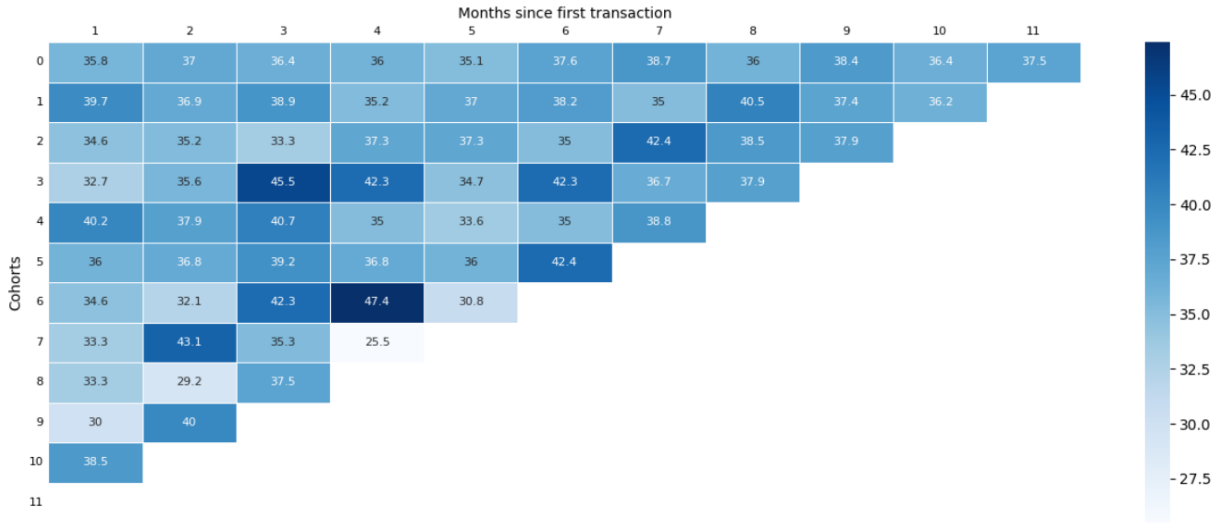

Figure 4 – Customer's retention

As can be seen in the figure above, the retention percentages are relatively consistent, with most cohorts having retention rates between 30% and 40%. There's one exception where the retention rate is less than 30%. Segmentation analysis and investigation of seasonal and geographical trends are planned to dive deeper into cohort segmentation results.

TBC