Carlsberg Group A1 - Executive Summary

Alston Zhuang, Julien Bajot, Kushal Bajpai, Sumin Lee, Victor Feng, Xingyu Guo

---

I. Introduction

In 2010, Carlsberg developed a revolutionary draught beer system, DraughtMaster (DM), aimed at significantly improving its on-trade business. By using compressed air instead of adding CO_2, DM benefits from its shelf life (keeping the beer fresh up to six times longer) and portability, as it is both lighter and takes less space. Traditionally, the steel keg restricted that the beer must be consumed within a week, so restaurants and bars tended to only offer beers that are in high demand to avoid waste. By using DM, outlets are able to serve a wider range of beer brands and manage stock levels more effectively. This in turn benefits Carlsberg further, as the newly added selections of beers are often craft and specialties, with higher profit margins. Therefore, DM creates a win-win situation for outlets and Carlsberg, ultimately providing consumers with a wider beer portfolio and a premium draught experience.

This technology was first rolled out in Italy and enabled Carlsberg to set up digital DraughtMaster. By using sensors linked to the DraughtMaster technology, Carlsberg can produce real-time data on hourly beer consumption across its outlets and help monitor the stock level of each outlet and replenish promptly. More importantly, real-time data makes it easy and accurate to analyze the operating status of each customer, based on which investment and promotional decisions are made.

The main objective of this project is to produce a more complex customer segmentation analysis based on the sensor data as well as the customers and materials data. We will be leveraging external data sources to enrich our segmentation analysis. In addition, we would like to identify the most promising customers and their characteristics such that the sales representatives can easily determine potential high value customers when expanding their business. Last but not least, we explored additional analytical methods, namely designing beer portfolio recommendations and weekly demand predictions of beer consumption.

II. Data

a. Internal

The Carlsberg team provided us three different sources of data from their sell-out (B2B). They are the following:

1. *'customer.csv'*

This dataset contains information on the outlets that DM sells beer to. It includes the outlet type, name and address, geographical information, the sales representatives in charge, etc. Most importantly, it tells us whether an outlet is linked to the sensor data, which contains data on its hourly beer consumption.

2. *'materials.csv'*

This dataset provides us with information about each alcoholic beverage that is offered. It contains data on the beer name, beer type, selling package, etc. Of the outlets that contain sensor data, Carlsberg offers 47 beers precisely.

3. *'time series data on beer consumption' (221 items)*

The final source contains data about the hourly beer consumption of over 200 outlets, spanning up to 18 months from September 2018 to February 2020 depending on the outlet. It is important to note that the timeframe for this consumption data varies across outlets. For some outlets we have more than a year of data, whilst some others contain less than two months.

b. External

In addition to the internal data provided, we reach out to several different external data sources in order to enrich the dimension of our segmentation analysis and increase the reliability of our result. For the time series data, we added information on the fixtures of the Italian football league, Serie A, and Champions League. In addition, we also added region-based historical daily weather data in Italy, which includes average/max/min temperature, wind speed and humidity. With regards to the static customer dataset, we added data from the Italian National Institute of Statistics, iStat. This covers demographic data such as household income and population, as well as behavior regarding alcohol consumption such as the frequency of beer consumption per week. The final data sources are firmographic and come directly from Google API, which entails specific features of the outlet such as rating, main keywords to describe the outlets, outlet type, opening hours, central location and proximity to different "public spaces" such as universities, theatres, football stadiums. A complete description of the external columns we gathered can be found in the appendix.

c. Limitations

It is important to note the possible limitations in our datasets and subsequent analysis. Firstly, whilst Carlsberg has 2708 customers in Italy, only 198 of them actually possess the key sensor data that produces real-time hourly beer consumption (see Exhibit 1 in appendix). Not only is this a relatively small sample, many of these customers are concentrated in northern Italy, which could lead to misleading and biased results that may not be representative of Carlsberg's entire customer base in Italy. Secondly, the results of our weekly beer consumption prediction are limited by the insufficient timeframe (less than two years) needed to perform a forecast of statistical rigor. Typically, we would require several years of data in order to build a very accurate predictive model – in this case, many of the customers have less than a years' worth of consumption data. Last but not least, we identified some inconsistencies in the internal data provided. During the data exploration, we noticed the same restaurant chain to be classified differently (i.e. a restaurant vs. bar). As the classification type was originally Carlsberg's main attribute to segment its customers, we believed that it was necessary to undertake further data collection to correctly identify their classification by using Google search's outlet classification and refining it at our discretion. Additionally, the presence of NAs in some columns such as 'channels' meant that we could not include potentially important features in our analysis.

III. Exploratory Data Analysis

At the EDA stage, we produced several visualizations and simple regressions to gain a basic understanding of consumption behavior over time and different characteristics that define Carlsberg's customers. Our customer data included a detailed laundry list of all 198 customers and its aggregate consumption for each beer as well as external sources such as geographical features (i.e. city center or not), population and income data, whether it contains outdoor seating, etc. Our time series data included the **daily** beer consumption of each customer and each beer it offered. The external columns we incorporated include temperature, humidity, whether a Serie A or Champions League fixture was played, etc. Our rationale behind using 'daily' rather than 'hourly' consumption is due to the computational demand of hourly data that would have made our machine learning methods difficult to do. We believe using 'daily' data would still capture the segmentation analysis effectively and provide us enough time to explore the data further such as building beer recommendations and demand forecasts.

From graphs attached at the end of this executive summary, several critical points can be drawn as shown below:

- Amongst all 2708 clients in Italy, only 198 clients' consumption data was provided. Most data concentrated in Milan and Turin in the North, dispersed in the middle of Italy and sparse in the South, and no data from Sardinia Island in the dataset.
- Among the 198 clients with data, one third of customers are located in Milan (63), followed by Turin (13). For most cities, we only have one or two customers there.
- Lombardy is leading the beer consumption in Italy such that its total consumption is 3.6 times higher than the second highest region. Also, Lombardy takes up 43% of total clients and 48% of total consumption in the dataset.
- From the consumption records of each client, two types of consumption patterns can be observed. For most clients, their consumption level is relatively stable and low throughout the year while other clients see high levels as well as high variations in monthly consumption level.
- More than half of all records in the dataset are classified as 'restaurants'. However, of all outlet types, 'pubs' have the highest average consumption with approximately 2800L per month.
- Beer consumption is directly related to seasons and among all beers, Craft beer and Lager beer are much popular than Specialties beers
- After the location visualization, the client with the highest monthly consumption is near cinema, shopping center, theatre, tourist attraction and university concurrently, where people gather the most

IV. Data Analysis Methodology 1 – AdaBoost Decision Tree

Apart from the segmentation analysis, there were two key issues that Carlsberg encouraged us to explore: identifying what features exist across profitable customers and what types of beers should be offered to the different customers and segments. Having gathered a rich variety of external sources to supplement our existing data, we proceeded with an AdaBoost ensemble method to determine the most important features that affect varying levels of beer consumption. This method works particularly well as it uses regression trees as a base and takes into account the weight of each classifier. Thus, we can easily identify which attributes have the most influence on the outcome we want to predict, **average beer consumption**. With this algorithm, we provided a ranking of the most important features as shown in Exhibit 2. For example, we see that the city population size and the type of beer offered (lager or craft) are some of the most important features that influence beer consumption level. Based on this first step, we are able to utilize these important characteristics and build them into our clustering analysis subsequently – this would increase the precision of our segmentation and enable us to drive better business insights from our analysis.

It is important to note that our takeaway from this algorithm is not a prediction of future consumption but rather an indication of which features may possess the greatest explanatory power for the variability of beer consumption between different customers. As noted before, as we only have 198 customers, the training model is learning from a relatively small subset of data – thus, we should be aware of the accuracy of the model. All in all, the algorithm still provides us with a list of important outlet characteristics that relate to consumption. These features will be used in our clustering analysis in the next part of this report.

V. Data Analysis Methodology 2 – PAM Segmentation

Having identified the 23 important features from our decision tree model, we pursued a "partitioning around medoids" clustering algorithm, or PAM. The use of PAM is based on the fact that our customer dataset contains both categorical and continuous variables, which means that the popular k-means clustering method is not suitable in this case as it uses the Euclidean distance to partition the observations into different clusters. The PAM method, on the other hand, uses Gower distance, which measures the partial dissimilarity across the observations and takes both numerical and categorical features into account. Just like k-means, PAM can partition data into a pre-specified number of clusters, by minimizing the distances between the points that are labelled to be in the same cluster and the point designated to be the center of that cluster. To determine the optimal number of clusters, we consider two aspects. First, we ensure that the silhouette width for the corresponding number of clusters is high enough, as shown in Exhibit 3.3. A high silhouette coefficient suggests high similarity between a single observation and its own cluster compared to neighboring clusters. The general rule of thumb is that the higher the silhouette coefficient, the better the cluster configuration. Second, we establish that the number of clusters can provide actual business implications and allow Carlsberg to develop sales and marketing strategies around the analysis. Therefore, we chose **five** (k=5) as our initial number of clusters (see Exhibit 4).

Based on the result of the clustering, descriptions of the five segments are summarized below:

Segment 1	The stores (mostly not city centered restaurants) selling late night foods with casual and cozy atmosphere all over Italy with **moderate average monthly beer consumption**
Segment 2	The stores (mostly a pub with outdoor seating) selling great cocktails with casual atmosphere near universities and basketball courts in Milan with the **highest monthly beer consumption**

Segment 3	The stores (mostly restaurants with outdoor seating and the highest google ratings) selling late night foods and great cocktails with cozy atmosphere in all over Italy with **high average monthly beer consumption**
Segment 4	The stores (mostly restaurants in the city-center) selling late night foods and pizzas with casual and cozy atmosphere in touristic spots in the big cities with the **second highest monthly beer consumption**
Segment 5	The stores (95% restaurants) serving 70% of craft beers with the **lowest average monthly beer consumption** Note: 80% of the outlets in this segment are an American diner chain called 'America Graffiti' which was originally founded in Italy. They are mostly non city-center and dispersed all over Italy.

VI. Performance Clustering

a. Promising clients

In this section, we aim to find the characteristics of high performing outlets in order to help Carlsberg direct their efforts to find profitable future clients. To achieve this, we combined two performance clustering methods. First, similar to before, we used the PAM method but decided on two distinct clusters because not only did this give us the highest silhouette width, but it also allows us to clearly distinguish between "high performing" and "low performing" segments Through this method, we were able to identify key features that define our high performing segment (see Exhibit 6) For instance, they were more likely to be located in the city center of large cities, typically within close proximity to neighborhoods, cinemas, touristic spots and universities. The higher performing segment also offered more beer brands, on average, and were more likely to provide outdoor seating. Although craft beer was the most popular beer type in terms of the offerings, lager beer consumption accounted for over half of all beers consumed.

Second, we decided on four distinct segments based on its percentile in terms of average monthly consumption. They were split into four segments:

Segments	# of customers	Avg. monthly consumption (L)
Top 5% - Ultra high	10	2,964
Next 20% - High	40	1,207
Interquartile Range – Medium	98	530
Bottom 25%	50	191

Whilst this method was a further addition to the PAM segmentation, we aimed to strengthen the analysis by isolating 'average consumption' as the proxy for measuring profitable segments. Similar to the findings in our previous method, the location of the high performing segments is a key feature: pubs centrally located in big cities typically had higher average consumption. Carlsberg lager beer was the most popular for both top two segments and they provided disproportionately more beer brands than the low performing segments. One key thing to note is that across both methods, we assume 'average monthly consumption' to be the defining variable for a 'profitable' client.

b. Mini Case: A Qualitative Comparison

In addition to the quantitative analysis of the features affecting beer consumption, we also decided to explore some qualitative features that could also explain this phenomenon such as the opening hours or the atmosphere of the outlet. In order to investigate the effects, we conducted a control experiment because qualitative factors cannot be included in regressions. We found several outlets that are highly similar to each other, but only differ in the opening hours or furnishing styles. To control for the outlets' similarity, we took a few features into consideration such that they must:

- Be located within 200 meters of each other
- Have the same outlet classification (i.e. bar or restaurant)
- Offer a similar number of beer brands

Restricted by these constraints, we found four groups to compare.

1. *Next* and *Tortuga*

Both are pubs and centrally located in Milan with the same zip code, 20135. Although both have similar opening time at 6pm in the evening, Next opens until 3.30 am while Tortuga closes at 2am. Consequently, we also see that Next has nearly 60% higher average consumption.

2. *ParcoMilano* and *Duomo Dal 1952*

Both are restaurants in the city center of Milan, located adjacently within a few meters of each other. The difference in consumption level suggests that longer opening hours are associated with higher average consumption. This is in line with intuition that longer opening hours brings more footfall thus boosts consumption. In terms of the elegance of interior design and furnishing, we also found it to be a significant factor.

3. *Il Cestino* and *Quore Italiano*

Both are restaurants within close proximity with similar opening hours. With everything else being equal, Il Cestino has a more decent furnishing and a cozier atmosphere (as described by Google) – we see Il Cestino's average beer consumption levels more than double that of Quore Italiano's.

4. *Bar Ristornate Pizzeria Otivm Lunch Café* and *Bar Canaja*

We find similar results when comparing Bar Ristorante Pizzeria Otivm Lunch Cafe and Bar Canaja. Although the latter has longer opening hours, its quality of furnishing is far more basic and less appealing, consequently attracting fewer customers. Bar Canaja's average monthly consumption is less than half of its counterpart.

VII. Beer Recommendations

Carlsberg offers a very wide range of beers for outlets to choose from. Beer popularity will differ across cities, outlet type, or even geographical proximity to universities or to the city center. This section will focus on our methodology to give beer recommendations for every outlet. The data we will use for this task is the customer data, which is the static data mentioned in part II of this report.

The recommendation system has two applications. The first application is for customers for which Carlsberg has consumption data. The application is for customers for which Carlsberg does not have consumption data. Our approach to this problem can be summarized in four steps:

1. **Find similar outlets**: As specified earlier, the mix of categorical and numerical variables in our data means we will use the Gower distance. We calculate the distance from each user to every other user resulting in a distance matrix based on Gower distance. In the first application, beer consumption is taken into account when calculating user similarity. In the second application, only "visible" outlet characteristics are used. This effectively shows us how "close" or how "far away"

some outlets are from each other. From this distance matrix, we can find the top 'n' closest outlets for each outlet.

2. **Find popular beers**: We need to find the popular beers among the similar users only. Out of the top n similar users from step one, we find the most popular beers by average monthly consumption (the mean consumption by beer across the similar outlets).
3. **Find popular beers for specific outlet**: The third step is to find the most popular beers for the outlet in question. Subsequently we will have the most popular beers for our outlet and the most popular beers for its top 10 similar outlets.
4. **Recommend beers**: Last but not least, we identify which beers to recommend to each outlet. When recommending beers, we need to take into account both the popular beers of the outlet and the popular beers of similar outlets. Essentially, we combine the popular beers for similar users and the popular beers for our outlet to give a 'top five' beers which constitutes our recommendation for that outlet.

While this model is quite simple, we believe it can be effective and could constitute a first step towards building an elaborate recommendation system.

VIII. Beer Consumption Prediction

Considering that weekly consumption prediction is very important for Carlsberg to arrange the weekly delivery and improve customer service, we tested forecasting models by predicting the next week's consumption based on historical data through the LSTM (Long Short Term Memory networks) deep learning process.

Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. Not only can it process single data points (such as images), but also entire sequences of data (such as speech or video). Because the recurrent network has feedback connections, new neurons remember "important" information from previous neurons. This type of structure is therefore for time series prediction.

Our predictions are by outlet, so the beer consumption data is split to match the outlets. It is important to note that the timeframe for which data is available varies a lot across the different outlets. Some outlets have more than a year of data, while some only have a couple of months. For the purpose of the prediction, we decided to use total daily consumption as the input data. We had to use daily data because weekly data had too few data points for some outlets. Our algorithm predicts consumption for the next seven days.

The intuition for the prediction is the following: to predict day T, one will use all data until day T-1. The data given to the algorithm to train it is a combination of consumption for day T (what we want to predict [y]) and consumption until day T (what we use to predict [x]) for all possible days T through a moving window.

One characteristic of the algorithm to keep in mind: the algorithm predicts multiple steps into the future, therefore, the further away we move from the last data point we have, the larger the uncertainty. Also, LSTM has its own limitation technically speaking, but moreover, we should also consider the categorical variables for certain clients, if possible, to do a better prediction instead of looking at the historical consumption data only.

I. Summary and Results

According to our analysis, we have successfully identified **five customer segments** for Carlsberg, based on which we could suggest different promotional and investment strategies.

The results from performance clustering showcase the potential high value customers that Carlsberg should target. Specifically, Carlsberg should seek outlets located in the city center of large cities, which are close to neighborhoods, cinemas, touristic spots and universities. Outdoor seating should also be preferred, and if the outlets offer a wide range of beer selection, it has a higher potential to perform well.

Throughout the project, we also identified several aspects that we encourage Carlsberg to pursue in the future. Carlsberg could accelerate their digitization process, for example, to disseminate sensors to all the customers as soon as possible, thereby assisting small retailers in predicting their sales on a weekly basis. In addition, Carlsberg should also collaborate with sales representatives on database renewal and data management, which includes updating customer demographic information and standardizing measurement matrix such as classifications and channels.

Appendix

A. Column descriptions gathered from external data

customer.csv

avg_con_hourly:
the average hourly consumption of all beers for the customer

avg_con_daily:
the average daily consumption of all beers for the customer

avg_con_monthly:
the average monthly consumption of all beers for the customer;
This was calculated as: total_consumption / months

months:
the number of months where beer is consumed for the customer (based on the beer consumption data)

total_consumption:
the total consumption of all beers for the customer

center_or_not:
Cities are described as located in the city center or non-center. For Milan and Napoli, stores located within the radius of 1.75km from the "central point (decided by google, most likely CBD)" are classified as "center". For all other cities, the radius is modified to 0.87km based on a lower population density.

classification_new:
Since the customer's name is the companies' name rather than a restaurants' name, Google API cannot locate the proper object. There is no good way to identify clients' type. For now, this column is done manually, based on personal understanding.

total_pop_in_province:
the total population in the province that the customer is located in 2019

total_pop_in_city:
the total population in the city that the customer is located

male_pop:
the total male population in the city that the customer is located

female_pop:
total female population in the city that the customer is located

male_perc:
the percentage of male in the city that the customer is located

female_perc:
the percentage of females in the city that the customer is located

never_social_per_100:
proportion of people in the region that never socialize (%)

everyday_social_per_100:
proportion of people in the region that socialize every day (%)

persons.aged.11.years.and.over.who.consume.wine:
proportion of people in the region that consume wine at all (%)

persons.aged.11.years.and.over..who.consume.beer:
proportion of people in the region that consume beer at all (%)

more.of.half.a.liter.of.wine.a.day:
proportion of people in the region that consume over half a liter of wine a day (%)

he.she.consumes.wine.more.rarely:
proportion of people in the region that consume wine on rare occasions (%)

he.she.consumes.beer.only.seasonally:
proportion of people in the region that consume beer on rare occasions (%)

he.she.consumes.beer.more.rarely:
proportion of people in the region that consume beer less often than every day (%)

he.she.consumes.beer.everyday:
proportion of people in the region that consumer beer every day (%)

X1.2.glasses.of.wine.a.day:
proportion of people in the region that consume 1-2 glasses of wine per day (%)
population

sd:
the variability of consumption for each customer, measured with standard deviation of consumption

med_income:
the median household income by region in 2017

avg_income:
the average household income by region in 2017

ci_craft:
the total consumption of craft beers for the customer

ci_lager:
the total consumption of lager beers for the customer

ci_specialties:
the total consumption of specialty beers for the customer

ci_craft_prop_beers:
the proportion of total beers offered by the customer that are **craft beers**

ci_lager_prop_beers:
the proportion of total beers offered by the customer that are **lager beers**

ci_specialties_prop_beers:
the proportion of total beers offered by the customer that are **specialty beers**

ci_craft_prop_con:
the proportion of total beer consumption that is **craft beer** for each customer

ci_lager_prop_con:
the proportion of total beer consumption that is **lager beer** for each customer

ci_specialties_prop_con:
the proportion of total beer consumption that is **specialty beer** for each customer

number_of_brands:
the number of different types of beers offered by each customer

near_basketball:
whether the customer is located within a 1km radius of a basketball court

near_cinema:
whether the customer is located within a 1km radius of a cinema

near_football:
whether the customer is located within a 1km radius of a football field

near_neighborhood:
whether the customer is located within a 1km radius of a neighborhood

near_shopping:
whether the customer is located within a 1km radius of a shopping mall

near_theatre:
whether the customer is located within a 1km radius of a theatre

near_tourist:
whether the customer is located within a 1km radius of a tourist attraction

near_university
whether the customer is located within a 1km radius of a university

real_name:
name of the customer based on Google Search

google_address:
name of the address based on Google Search

google_classification:
customer classification based on Google Search

rating:
customer rating (scale of 1-5) based on Google Search

keyword_1:
first keyword that describes the customer based on Google Search

keyword_2:
second keyword that describes the customer based on Google Search

keyword_3:
third keyword that describes the customer based on Google Search

outdoor_seatting:
whether the customer provides outdoor seating based on Google Search (Y/N)

time_series_data.csv

serie:
0: The local team did not play in the Serie A league on that day
1: The local team played in the Serie A league
2: The two local teams played against each other in the Serie A league

cl:
0: The Champions League did not happen on that day
1: The Champions League happened on that day
2: Two Italian teams played the game in the Champions League on that day
3: Four Italian teams played the game in the Champions League on that day
(There was no such occasion that one or three Italian teams played on a single day)

weekend:
whether the date occurs on a weekday (0) or a weekend (1)

max_temp:
the city's maximum temperature during that day, in Celsius

avg_temp:
the city's average temperature during that day, in Celsius

min_temp:
the city's minimum temperature during that day, in Celsius

max_humidity_perc:
the city's maximum humidity level during that day, in percentage

avg_humidity_perc:
the city's average humidity level during that day, in percentage

min_humidity_perc:
the city's minimum humidity level during that day, in percentage

max_wind_speed_mph:
the city's maximum wind speed during that day, in miles per hour

avg_wind_speed_mph:
the city's average wind speed during that day, in miles per hour

min_wind_speed_mph:
the city's minimum wind speed during that day, in miles per hour

Exhibit 1: Map of Italy with 2708 total clients vs. 198 clients with beer consumption data

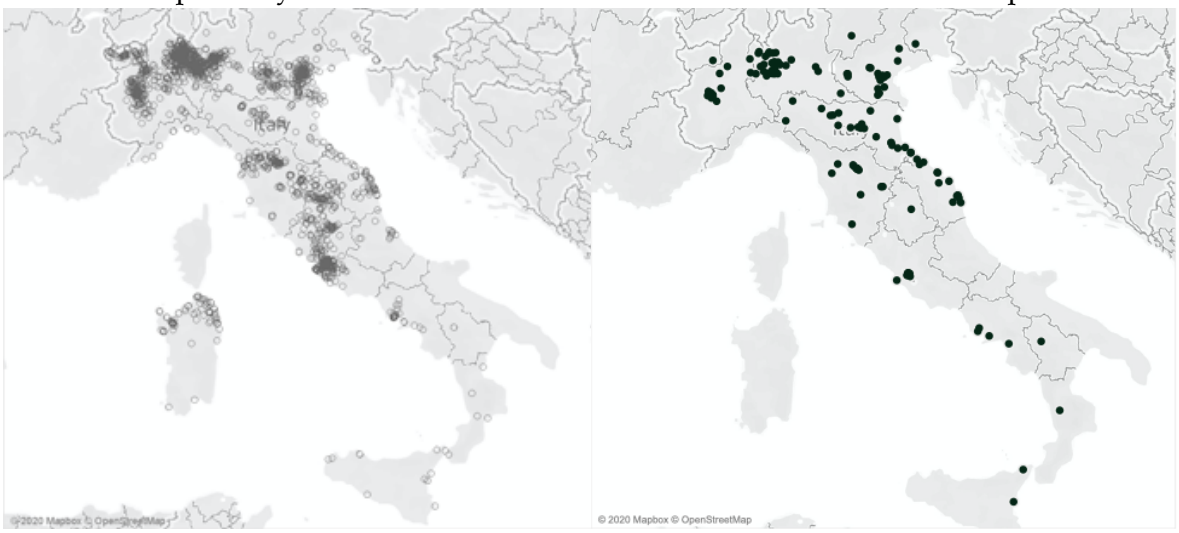


Exhibit 2: Feature importance from AdaBoost algorithm

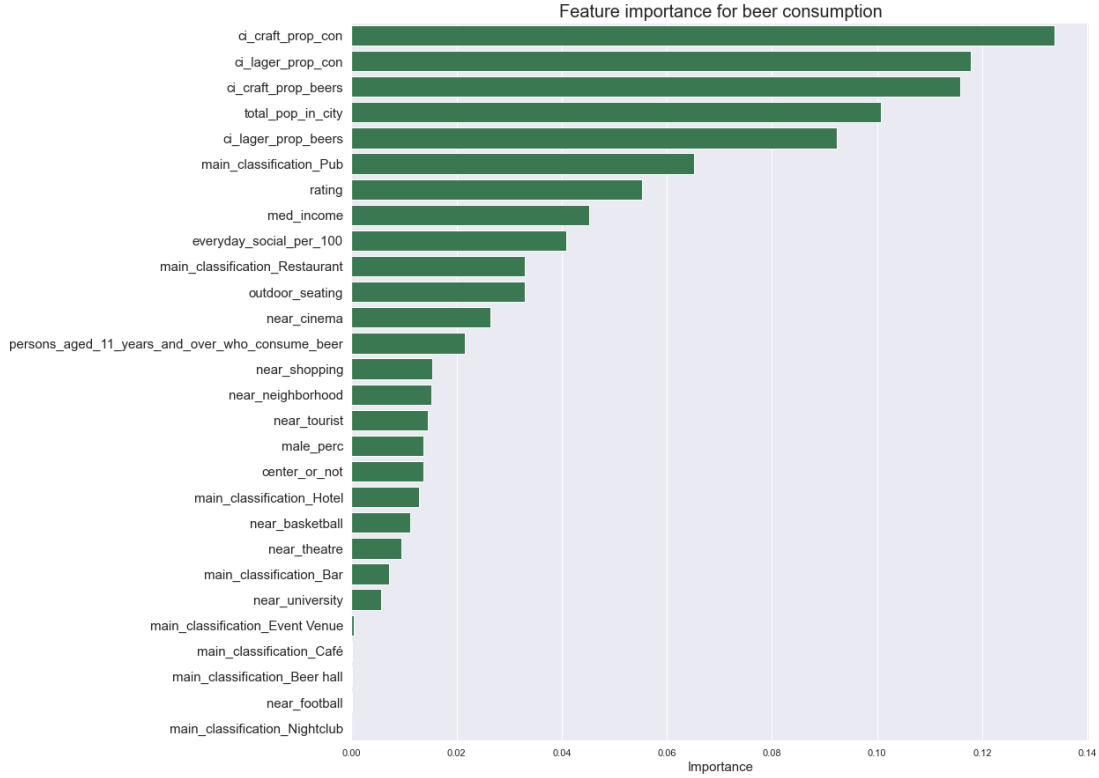
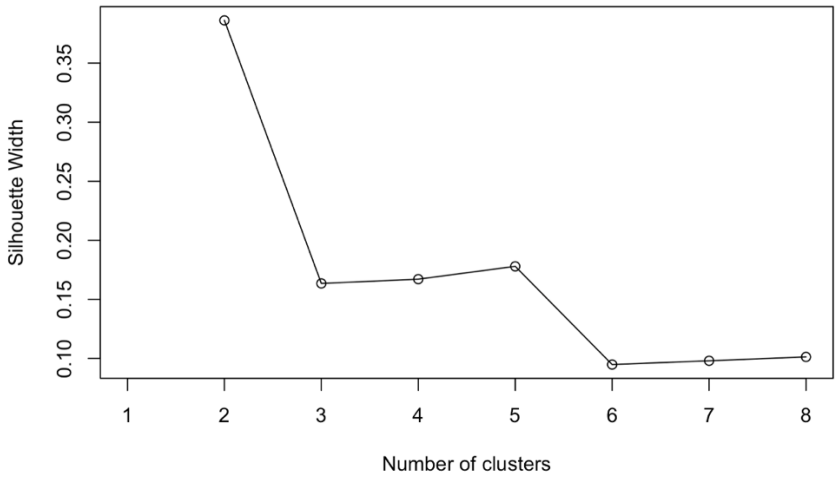
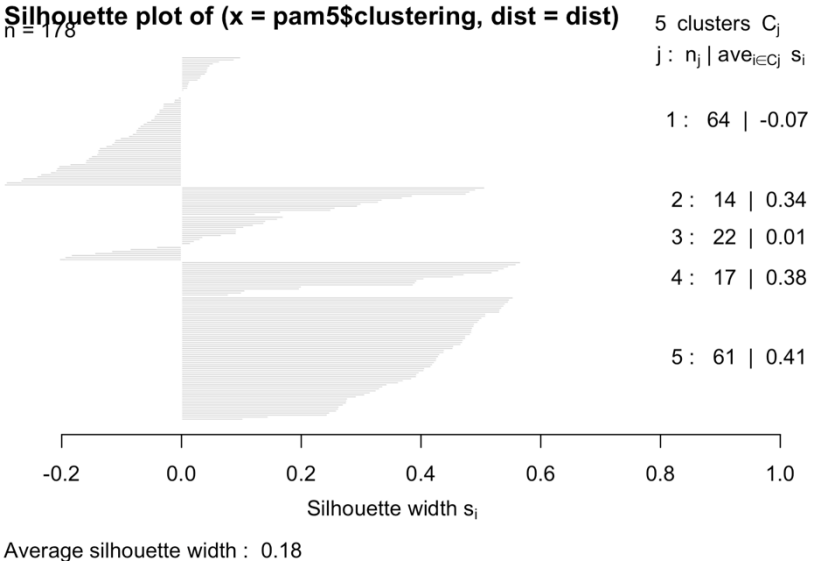


Exhibit 3: Silhouette Analysis with the master customer dataset

3.1 Average silhouette width for K equals to number 2 to 8



3.2 Silhouette plot for K = 2



3.3 Silhouette plot for K = 5

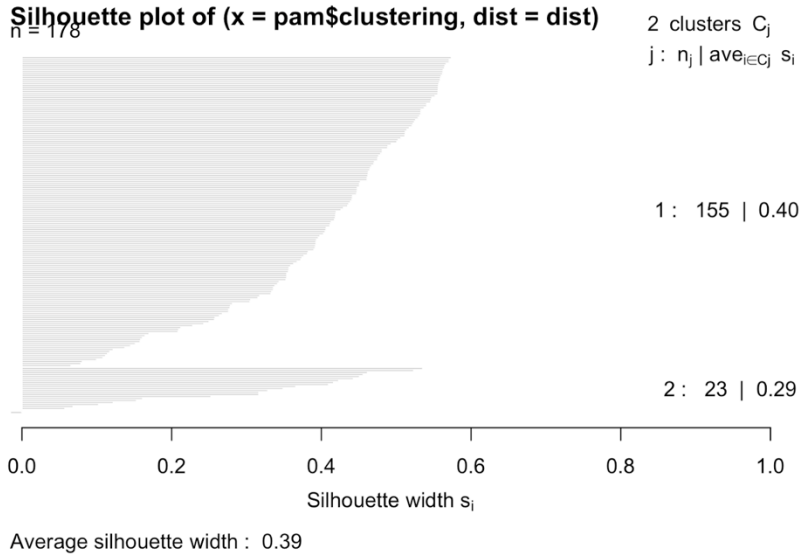


Exhibit 4: Segmentation with k = 5 results

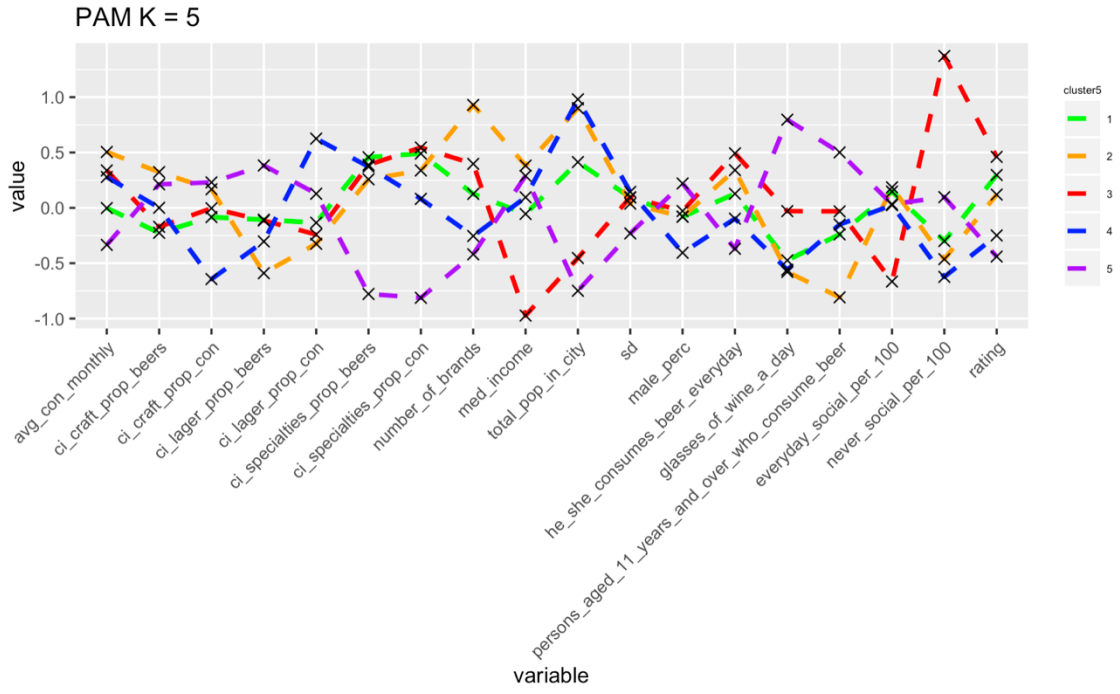


Exhibit 5: Performance Clustering Result

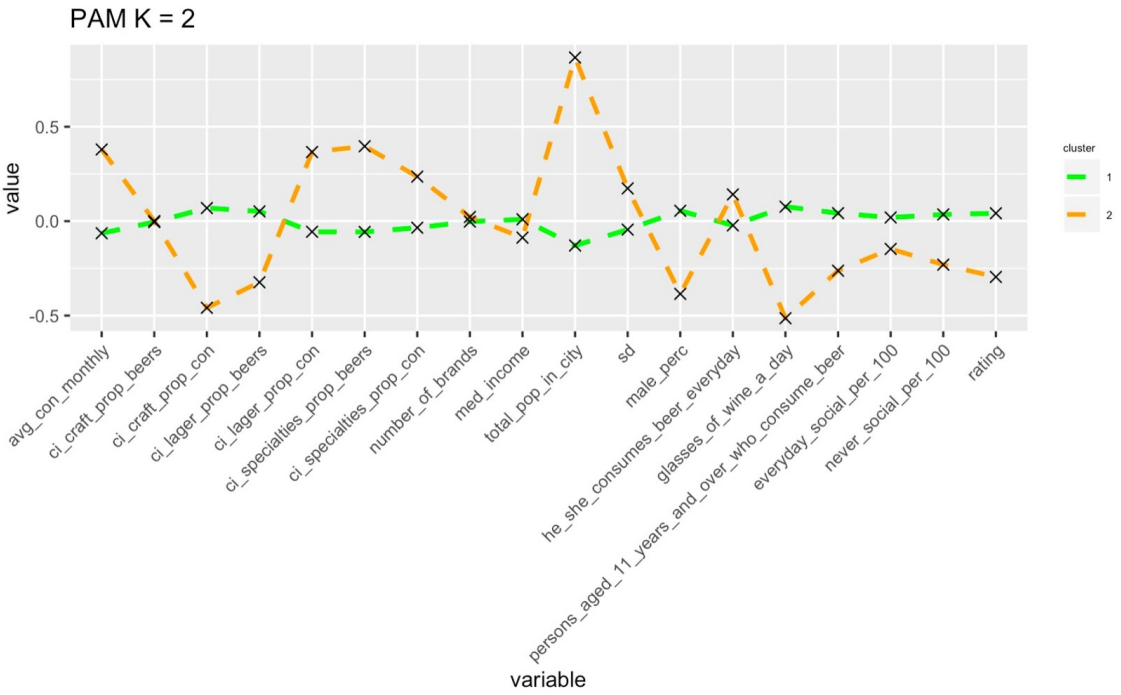


Exhibit 6: Performance Clustering Method 2 - Percentile Range

	Segment 1: Ultra High	Segment 2: High	Segment 3: Moderate	Segment 4: Low
# Customers	10 (5%)	40 (20%)	**98 (50%)**	50 (25%)
Avg. monthly consumption (L)	**2,964**	1,207	530	191
Number of brands	**15.2**	10.9	8.3	6.0
% Outdoor seating	**80%**	48%	36%	42%
% City center	**40%**	28%	33%	38%
Avg. population	**1.24m**	989k	600k	483k

Exhibit 7: Mini case information summary

Pub/20135	Next	Tortuga
Avg daily consumption	5.89L	3.72L
Number of brand	7	10
Opening hours	6pm – 3.3am	6pm – 2am

Gastropub/20145	ParcoMilano	Duomo dal 1952
Avg daily consumption	0.96L	2.3L
Number of brand	12	12
Opening hours	12pm – 11pm	8am – 11pm

Restaurant/20123	Il Cestino	Quore Italiano
Avg daily cons.	2.63L	1.25L
Number of brand	5	6
Opening hours	6pm – 10am	7pm – 12am

Bar/35122	Bar Ristorante Pizzeria Otivm Lunch Cafe	Bar Canaja
Avg daily cons.	2.75L	1.08L
Number of brand	8	9
Opening hours	10am – 10pm	8am – 12am