# Notes for master thesis

## General idea of the study

Main goal of the study is to apply marketing segmentation on customers of e-commerce site.

Profits of such analysis for the business include:

* one is able to get to know the customers better
* better planning of advertising strategies - possible usage of microtargeting, optimization of expenses on campaigns because of focus on the most promising customers
* one can find the most profitable products categories - this can guide future decisions about direction of further growth of the shop
* creating recommendation system is possible - because most of the shop customers are first-timers, a problem of cold start is very important. Recommendation can be done by assigning the customer to the segment of previous customers, based on the items he already has in the cart. This is called cross-selling.
* **It is possible to create strategies for customer retention.** In the e-commerce analysed, almost 97% of the customers never return to the site to buy again. This has a huge potential for improvement in revenue. It is usually said that it is easier to retain a customer than to obtain a new one (because of the cost). Customer segmentation offers a solution for this - explained in the next section.

The goal of the strategy I would like to implement is to find the customers that bought in the shop only one time and find the ones that are most likely to buy next time. After obtaining such information, it is possible to “give them a push” - for example target them with a suited advertisement, offer special discounts etc. The method is based on the fact that we have the data for the customers that bought more than one time, and we can check what are their characteristics. Specifically, this can include price and rating of the first purchase, bought categories of products, customer location, sociodemographic variables (based on the region of the customer) - wealth, probable ethnicity, religion etc. Also, one can obtain specific information about what is the perfect time to give an advertisement for the customer - based on the time that has passed from first to second buy for previous customers. (The method described here is largely inspired by an analysis available under this link: <https://www.optimove.com/blog/how-to-convert-one-timers> . )

The approach looks like this:

1. Split customers into 2 groups based on whether they have bought second time.
2. Assess “loyal customers” group. Run clustering algorithm using variables present in the dataset and generated by other means (described in detail later). Use some method for assessment of variable importance to determine irrelevant variables. Important remark is that clustering should be done for the moment **just after** the first order. This way, population of one-timers and loyal customers can be compared meaningfully. Also, PAM sounds like a good choice for the algorithm, as it gives real representation of the cluster centers.
3. Use clusters obtained on loyal customers group to “classify” customers that made just one purchase.

It would be valuable to do all analyses in a rolling manner - to reflect the real-life situation, in which data is sequentially from longer period. This approach can also help to assess the cluster stability and how the model performs. First run of the analysis on 6-month period, then run next iteration every month. (Dataset spans almost 2 years).

Method described above helps to get to know one-timers better. This can be used to prepare tailor-made recommendations and advertisements for them. However, an unanswered task is to judge, for which customers it is the most reasonable to spend money and effort on advertisements. It is probably the case that some customers are more likely to buy next time than the others. A challenge is to select the ones that are the most likely but still didn’t do it and focus advertising on them. To that advertising, results of clustering from the previous step can be used.

One simple heuristic to assess attractiveness of one-timer is checking days since the first purchase. From EDA, ~60% of second orders are made in time span of 2 months after first order. Other, more sophisticated method would be to run some classification algorithm - make “if\_bought\_second\_time” a dependent variable and check if it can be predicted. However, I have not tested such models, and I am not sure if I should extend my study for this.

## Potential variables to use

The dataset is very rich in variables. Some of them are ready to use as-is in clustering (like order value), but others require quite a processing (text review of the order). I think that this is a main strength of the article – most of the studies I have found on customers segmentation focus only on small subset of basic variables. Main topics of the variables include:

* **Order information** - order value, number of items bought, date, day of the week of the purchase, date of the delivery, delay in the delivery time, if the order was made in instalments or by card (in Brazil instalments payments account for majority of transactions online). This information is straightforward to obtain from the dataset.
* **Items bought in an order** - category of the product, unit value of the product, if the price was lower than previously. This information is available in the dataset. However, it is tricky to add this to main dataset, because there can be multiple items per one order. The easiest (but far from perfect) solution would be to select one category of the product to main dataset, for example category of the costliest product. However, in this way lot of information is lost. Other solution is more sophisticated, but worth trying. It involves clustering of the baskets; the details are described below.
* **Location info for the customer** - exact geographic location, if rural area, mean income in particular area, population density, ethnicity structure of the region. Except first one (present in the dataset), others have to be obtained from external web service (SIBRA, Brazilian statistical office). Joining of the dataset will not be straightforward. In the e-commerce dataset, assignment to particular area is done by 5-digit ZIP code. In SIBRA, resolution is available to “Microregions”. These two don’t match at all - level of aggregation is similar, but the boundaries not. Also, specific polygons for zip codes (SHP files) are available only as a paid service, and in the dataset, there are only a few points present for each ZIP.
* **Review of the order** - 1-5 rating of the purchase, text review. 1-5 rating is easy to join to the main dataset, however text review is more complicated. One of the possibilities is topic modelling - as a mean to assess what the review is actually about, and what exactly did the customer rate positively/negatively. This can be done via classic Latent Dirichlet Allocation algorithm, but also newer ones that rely on word2vec idea can be used. A tricky part is that the reviews are in Portuguese, but it should be quite easy to translate them via Google Translate API or other service.

## Market Basket Analysis - clustering of the baskets

A challenge in adding information about bought items lies in the fact that number of items is not constant for every order. Because of that, I cannot add the information to the dataset directly - adding a dummy variable column for very possible item is impossible. However, this information can be compressed using clustering. This idea is inspired by document clustering technique from text mining. In original algorithm, first frequencies of words are computed per one document. This gives a matrix where observation is a document, and variables are words. On that dataset, clustering is performed. This procedure gives a cluster assignment for every entity in the dataset, so it is possible to reduce a document into single number.

Translation of the procedure to the basket case could be as follows. One observation is one order, and in columns there are all product categories. Weights can be a simple indicator (1 if an item of particular category was in that order), number of items from category in an order, or monetary value of all items from such order.

## Work done so far

* Research about variable selection in clustering
* General research about customer segmentation in marketing
* Analysis of the data available for loyal customers (ones that bought more than one time) - in separate notebook
* Research about external datasets providers - getting to know SIBRA website. Basic dataset in good resolution is provided, including income per capita, population density and mean years of schooling.
* Basic analysis of spatial distribution of customers - in separate notebook
* First clustering with usage of Recency, Frequency, Monetary value approach - in separate notebook

# Notes from papers

Notes from the most promising resources and papers I have found, potentially useful later in literature review.

### General resources on segmentation

**Kotler – Principles of Marketing (“bible of marketing”):**

Lots of business knowledge regarding customer segmentation, but just 1 paragraph about usage of quantitative methods in market segmentation. Standard approach is to choose couple of variables and create cross-tables, with every cell being one segment. Very inflexible and basic approach, but also advantage is explainability.

**https://www.optimove.com/blog/how-to-convert-one-timers**

Business-centric article. Standard approach in marketing segmentation is to put all customers that bought the product only one time into one big segment. The authors here state that this approach is invalid. They propose to analyse loyal customers and segment them first. After that, check what were characteristics of the first purchase made by each of these loyal customers. For each of the one-timer, find the cluster it belongs (clusters are created based on loyal customer's first purchase). Such approach is interesting in the case of this dataset, as only 3% of the customers bought more than one time - there is big potential for improvement. Also, there is a consensus among marketing researchers that it is way easier to retain customer than to obtain new one.

**https://link.springer.com/article/10.1057/dbm.2012.17**

**"Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining"** - very basic analysis, but the problems in the dataset are very similar - also big skew is present in frequency and monetary value.

**https://carto.com/blog/new-approach-customer-segmentation-location-data/**

A business-oriented case study for using location data in marketing segmentation. In the dataset there is big spatial resolution provided. Also, Brazilian “GUS” provides some information on a very detailed level - it is probably worth exploring

**https://inseaddataanalytics.github.io/INSEADAnalytics/CourseSessions/Sessions45/ClusterAnalysisReading.html**

Other case study for segmentation based on RFM - good description of steps needed to assess accuracy of the solution.

**https://appsilon.com/customer-segmentation-leads-to-goodies/**

A case study of customer segmentation in R - with codes and nice visualisations.

### Resources for the particular dataset

**https://www.kaggle.com/andresionek/geospatial-analysis-of-brazilian-e-commerce**

Great geospatial EDA of the dataset created by the data provider. Some interesting information can be drawn from here:

- Region of the customer largely influences profitability (better in urban areas)

- Micro segmentation can be crucial for good clustering - for example in bigger cities, profitability depends largely if the customer is located downtown or in poorer favelas.

**https://github.com/DavidJulienMillet/MscP2\_Olist-customer-segmentation/blob/master/Olist\_modelization.ipynb**

Case study performed on the same dataset - also customer segmentation. However, more focus is put on assessment clustering stability, not validity of the approach.

### Feature selection in clustering

**"A survey on feature selection approaches for clustering" Salem Alelyani, Jiliang Tang and Huan Liu -** very good and recent (2020) summary of existing methods for feature selection. Also, importance of feature selection in clustering is very well explained.

**http://papers.neurips.cc/paper/2486-feature-selection-in-clustering-problems.pdf**

*"It combines a Gaussian mixture model with a Bayesian feature selection principle"* - model-based clustering. However, as in the dataset there are lots of categorical variables, using gaussian model is impossible.

[**https://dollar.biz.uiowa.edu/~nstreet/research/ida02.pdf**](https://dollar.biz.uiowa.edu/~nstreet/research/ida02.pdf) **-"Evolutionary model selection in unsupervised learning"**

Algorithm for finding the best features using genetic algorithm. There are 4 metrics used instead of one - for example one similar to "total within-cluster sum of squares". There is also penalty employed for selecting too many features and too many clusters. Main advantage of this method over approach with maximising one metric is ability to obtain parsimonious model. Genetic algorithm helps to optimize individual metrics without worsening the others. Approach with using multiple metrics to assess clustering quality is promising. Also, genetic algorithm sounds interesting.

**"Improving Performance of Similarity-Based Clustering by Feature Weight Learning"**

Main idea is to optimize weights of each feature (with potentially 0 weight - excluding the feature completely). This method is meant for fuzzy clustering, in which similarity matrix is used (e.g. PAM). Described algorithm is gradient descent, with "fuzziness" of the clustering as the objective function (defined in the article). Simplicity of gradient descent is desirable, but on top of that there is complexity of fuzzy clustering - overall complicated method.

**https://stats.stackexchange.com/questions/21807/evaluation-measures-of-goodness-or-validity-of-clustering-without-having-truth**

Interesting general discussion about assessing goodness of fit for clustering.

**https://arxiv.org/pdf/1007.1075v1.pdf**

"Clustering stability: an overview"

One solution for getting the best model for clustering is stability assessment. In this paper it is claimed that non-stable solutions should not be considered at all. However, when having few stable solutions, we cannot say that the most stable solution is "the best".

**https://github.com/o1iv3r/FeatureImpCluster**

Nice package for testing relative importance of variables. It is based on permutations of each variable and measuring the "misclassification rate". Also, it is model-agnostic, so different algorithms are possible to use side-by-side. It looks very promising and simple, however the algorithm is not presented in any published paper - only this package and GitHub page.