BUSINESS CASES WITH DATA SCIENCE

MASTER'S DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS – MAJOR IN BUSINESS ANALYTICS

Wonderful wines of the world customer segmentation

Group S – Business Case #1

António Carvalho, 20200642

Frederico Rodrigues, 20200583

Gonçalo Carvalho, 20200664

Manuel Borges, 20200596

February 2021

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Índex

1 Introduction

This report was written to analyze and cluster the dataset provided by the wonderful Wines of the World. While reaching this goal, the data handling was processed to extract the maximum value and minimum redundancy.

Wonderful Wines of the World is in the wine industry for seven years. Four years ago, they started building a database with all their customers' purchases and various characteristics, like income, age, frequency, and some others. The wine industry has much competition between producers, distribution, and retails - Thousands of different wines around the world are produced and consumed every day, and according to wine specialists - it's a very sensitive business once its customers should have a very personalized and detailed approach in the sense of being able to know the right information about what place, price, product, and promotion to use to each customer. This strategy, known as Marketing mix, is the one WWW wants us, as Data Scientist consultants, to prepare to improve their relationship with the clients, reach new ones, and distinguish which ones to prioritize.

Marketing mix refers to the set of actions an organization uses to promote its brand. As explained above, this marketing methodology is commonly grounded in 4Ps of marketing – Price, Product, Place, and Promotion – the best way to *put the right product in the right place at the right price with the proper promotion*. This way becomes easier for wonderful wines of the world to meet their customer's needs and demands.

With this in mind, our project was established to support Wonderful Wines of the world in creating a customer segmentation with the commitment of distinguishing and characterize clients. Understand how WWW can reach its groups of clients through these four principles is our priority in this report.

For the process of mining the data in order to achieve what was established, we have used a process called CRISP-DM – *Cross-industry standard process for data mining* – in which the raw data goes through a step-by-step process beginning in understanding the business – in this case, what does WWW need and what is required. Alongside this step, there is the need to understand the data – collecting, identifying, and analyzing all WWW resources. After this step is complete, following CRISP-DM methodology, it is the phase of Data Preparation where, after understanding not only the data but also the business, we clean, select and format the data. After the data is prepared, the next step is modeling – some techniques are used in producing the best results. After the products are created, the following steps are evaluation and deployment. Using this methodology (CRISP-DM) as a guideline in our project is a great way to provide a uniform framework in a well-known and standardized approach.

2 Business Understanding

2.1 Determine business goals

2.1.1 Background

To properly understand the project and its business, it's crucial that the first phase starts by understanding the background. Wonderful wines of the world is a seven-year-old international company, also known as WWW, that focuses on selling wines from all over the world—characterized by distributing from the producer to the consumer different and interesting wines through several channels worldwide.

With a robust website, ten small stores in major cities around the USA, new catalogs sent to its customers every six weeks, and a telephone retail channel, WWW has a wide variety of approaching its clients. Additionally, the company has pursued aggressive promotions actions in wine and food magazines which enabled it to have a large customer base of 350 000 customers.

Therefore, one of this company's priorities is to understand better their client's behavior with lots and different clients.

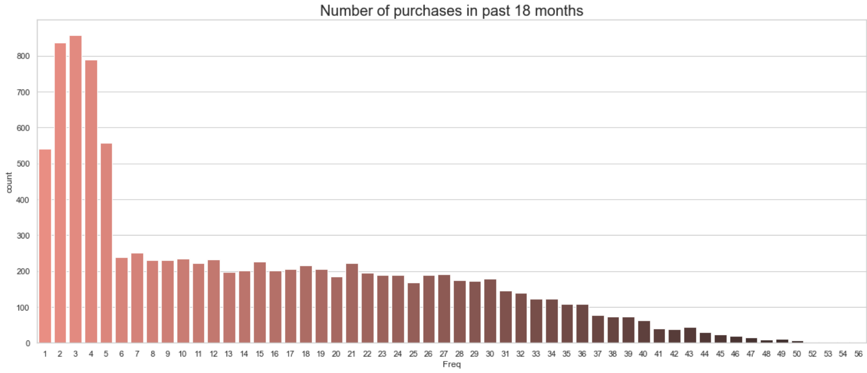

Figure 1- number of purchases in the past 18 months

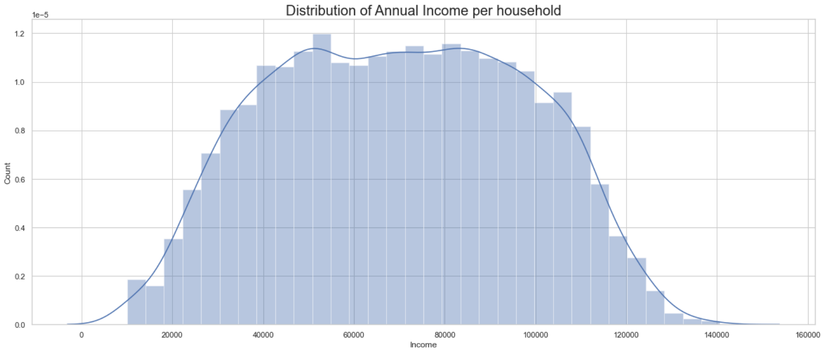

Figure 2- Income per household

Throw-out these two figures, we can quickly understand a bit more about customers' income and purchases. In figure 2.1. we see a considerable drop of purchases from the fifth item to the six – passing from more than 500 clients purchasing five items in the last 18 months to just more or less 200 clients buying six items. Regarding figure 2.2. there is a regular income per household from WWW's customer's sample.

2.1.2 Business objectives

WWW pretends to know more about the different types of customers and their behavior to improve their relationship. According to the company - "*Our customers are highly involved in the wine industry and have sufficient money to indulge their passion for wine*" - however, they don't have specific knowledge about customers, and up until now all marketing moves were based on feedback from salespeople and intuition.

Our objective as a business consultant is to improve and understand how WWW's customers behave and its clients' different characteristics to design a better and efficient marketing approach, for this analysis was provided by the company a dataset with almost 3% of all clients with several variables/characteristics about them.

2.1.3 Business Success criteria

Our main criteria to evaluate the success of the project in a more overall and broadway is to provide the most useful information about each type of customer to reduce uncertainty surrounding costumers, including everything about them, from purchases to wine preferences, which is crucial for better business performance.

In a more objective and specific method, we also measure success from three parameters: 1. decrease the churn rate. 2. Revenue increase;

2.2 Situation assessment

The company provided the team the following documents:

1. An excel with all data from WWW's costumers with 10.000 observations (purchases from clients in the last 18 months) and 30 variables (containing income, purchases, wine types, between lot more). This excel was extracted from a dataset with a time span of 4 years with around 350.000 customers' purchases.
2. 2. A metadata file with the general context and all variables explained.

The constraints of this project that need to be respected are that there are not clients in the analyzed data that haven't bought anything in the last 18 months (inactive customers) as well as underaged clients (less than 18 years old).

Terminology:

- Market segmentation: The process of identfying groups of customers based on their purchase behavior.
- Pricing: The process of determining the value of the product.
- Churn rate: Rate of customers that have stopped buying the product.

2.2.1 Risks and Contingencies

Missing data: Fill missing data with median, mode, mean, or KNN Imputer.

Outliers: Find them manually, IQR or Z-score and delete them. Afterwards, use the decision tree classifier to find the right cluster for each outlier.

Redundant variables: Use correlation matrix to detect redundant features and not use them in the clustering process.

2.2.2 Costs and Benefits

Benefits:

1. No cost of Mass-Marketing

2. Decrease of the churn Rate

3. Increase of the revenue

Costs:

1. Cost of a Differentiated-Marketing

2.3 Determine Data Mining goals

Data mining goals have to be in line with the business goals. Therefore, since the business objective is to get valuable insights into the customer segmentation, the data mining one is to have good explanatory data analysis through clustering and other visualization tools, such as pairwise table or correlation matrix.

2.4 Produce Project Plan

2.4.1 Project Plan

The project was mainly divided into two perspective the data mining part (step 1-10) and the business perspective (step 11-12)

1. Explore the data: The team starts off with the analysis of the different types of data in each variable the statistical information of each variable and the relation between.

2. Acquire insights from the data visualization.

3. Coherence test: Check if any information that did not respect the rules, for example, check if all the clients bought something from the shop during the last 18 months (active clients) or were underaged.

4. Data cleanup: The variables “Custid” and “Rand” are deleted as they are useless for the analysis.

5. Outliers: Since the clustering analysis is based on the distance between points, outliers can distort the analysis, as such they had to be eliminated. The best method for this was the IQR, as it was the one that detected the most reasonable amount of outliers (4.3%)

6. Normalization: Besides outliers, the different scales of the variables also distort the distance-based clustering process. Therefore the data has to be put on the same scale. The method used was normalize() from the sklearn library.

7. Redundancy: A correlation matrix was constructed to find variables that do not significantly contribute to the analysis. The only found considered redundant was ‘Monetary.’

8. Clustering Process: Several clustering methods were deployed (Kmeans, GMM,HC, K-means + SOM, HC+ SOM). Inside each clustering process, it was required to determine *a priori*, the number of clusters through several techniques such as K-elbow Plots, Silhouette Analysis, Dendrogram, the type of Hierarchical Clustering (by using the R2 Plots) and, finally, determining the type and number of components of GMM using as metrics the AIC and BIC scores. For the final steps, to choose the best clustering methodology, we plotted R^2 plot for the five different clustering methods to compare each cluster rating.

9. Interpretation of the categorical variables: The categorical variables were then compared to the clusters using Multiple Correspondence Analysis. The only categorical variables that are useful for our task are ‘Kidhome’ and ‘Teenhome’.

10. Classification of the outliers: Using a decision tree classifier, the outliers were then assigned to a cluster. The classifier had a reasonable accuracy of 84.44%.

11. Interpretation of the clusters: The team proceeded to analyse all 4 clusters through the variations in the variables and then, in a more business view, created a persona for each cluster.

12. Deployment: Presentation of several marketing solutions.

2.4.2 Initial Assessment of tools and Techniques

1.Explore the data: info(),describe(),Pairwise,Boxplot and Correlation matrix

5.Outliers: Manual, IQR, Z-score method and the combination of both.

6.Normalization: StandardScaler() and then normalize()

7.Redundancy Correlation matrix

8.Cluster: Kmeans,GMM,HC, Kmeans+SOM, HC+SOM

9.Interpretation of the categorical variables: MCA from prince library.

3. MODELING ANALYTICS PROCESS

Describe only the significant steps involved in the process. Do not replicate what is already described in the Notebook. If necessary, reference the reader to the Notebook.

3.1 Data understanding

Initially, the dataset contains 10000 clients, with at least one purchase in the previous 18 months, with 29 variables having 20 metric and nine non-metrics (categorical).

Overall, this dataset was well maintained and did not have many needs of being treated, which led to a visual exploration of the variables through the many visualizations shown below.

Since it was clearly understood, from the beginning, that this was a clustering problem, our team followed the necessary steps to allow our models to perform better.

3.2 Data Preparation

When dealing with a customer's segmentation problem, that will most likely require a clustering approach, the data preparation is a vital step for the performance of the clusters. This phase started with the data cleanup (getting rid of 'Custid' and 'Rand'), detecting outliers with IQR (it had a detection rate of around 4.3 %), Normalizing the scale for all variables and at last but not least, eliminating redundant variables, more specifically.

3.3 Modelling

Following the Data Preparation and having set the variables to use on a clean data frame, it was time to initiate the clustering process.

Due to its simplicity and easy implementation, the K-Means algorithm was chosen as the first one to be applied. One of the downsides of this method is that it requires the number o k (clusters) to be set à priori for k centroids to be created and stored providing a distance comparison to each point, forming a cluster.

Knowing that our first method has serious problems when confronted with clusters of irregular shapes, it was decided to try Gaussian Mixture Model or GMM, a distribution-based model. The GMM assumes that the datapoints belong to a certain gaussian distribution which represents a cluster.

To try another approach perspective and open ours views about the clusters, we decided to use Hierarchical clustering that is a powerful technique that allows you to build tree structures from data similarities. To perform that best Hierarchical clustering we plotted R^2 plot for various hierarchical methods and chose the "Ward" method (the best one figure…).

Additionally, the team decided to use a mix of clustering methods, namely SOM (Self Organizing Maps) with K means, SOM with Hierarchical Clustering. In the two methods with SOM, the datapoints from multidimensional space were laid out onto a 2-dimensional grid, which enabled us to visualize better the data and also to use on top of it the K Means and Hierarchical Clustering

(Roux & Rouanet, 2004)

3.4 Evaluation

The teams chose the R^2 as a metric to pick the best clustering method. And according to the graph the best one seems to be K-means followed by GMM.

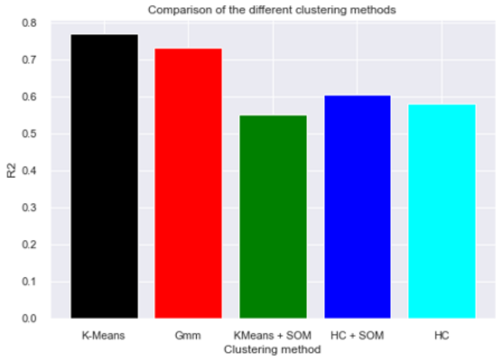

Figure 3- Comparison of the different clustering methods

4. RESULTS EVALUATION

Cluster 0: This group generally presents an above average education, and their tastes focus mainly on Dry Red wines, with clear lack of interest on Dry White wines.

Cluster 1: These clients are mainly characterized by being the youngest group but also the most underpaid and lowest level of education, they also do not shop WWW frequently. Taste wise, this cluster tends to prefer every type of wine except Dry Red.

Cluster 2: is also a formed by young people, with the difference on pursuit of a higher education, resulting on a predicted rise of income. This cluster shows great attention to discounts and seems to be well informed, consuming all wines types.

Cluster 3: This group is formed by older people with a high level of income and wealth, leading to a frequent purchase of WWW items, mainly on the store.

Once there is no possibility of exactly knowing real-world consumers, potential clients and how customers' segmentation will be treated and approach, our group has prepared a fictitious, but precise, way of describing each type of client in a more interesting method using Personas – personas is a collective image of a segment of a company target, each persona doesn´t represent the entire target but represents a fraction of a target audience.

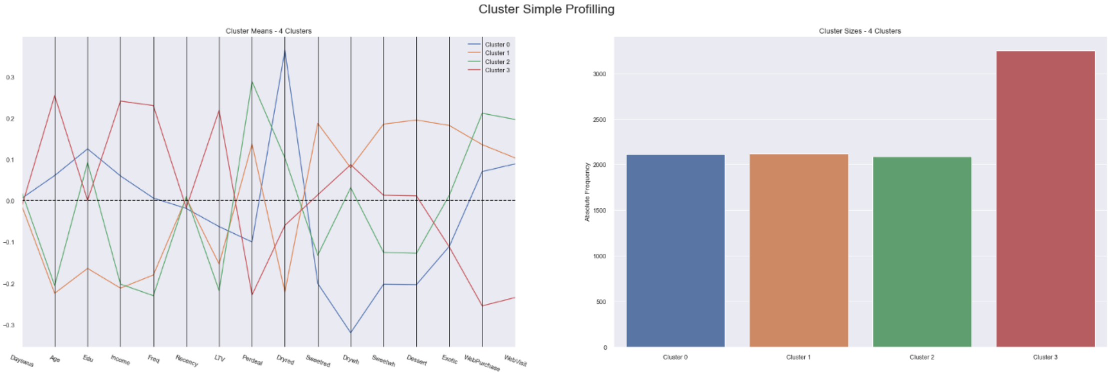

Figure 4- K-Means clustering profiling

As a result of the segmentation made and explained above, we can visualize different clients with singular characteristics and tastes who have bought from WWW in the past 18 months. Having this into account, the following personas have been created:

Cluster 0 - Mark Smith, 53 years old, has a Ph.D. in International Law and works in a Lawyer's firm in San Francisco, California. He has a strong preference for Dry Red wines, being practically the only type of wine he consumes. Mark is enthusiastic about buying wines online and visiting WWW Website. Sometimes he enjoys taking advantage of discounts. He has two teens living with him.

Cluster 1 - Henry Jones, 24 years old. He is currently trying to finish high school with difficulty while working part-time with a very low income, Henry can´t buy wines frequently, but when he buys, he really cares if the wine is on discount. He loves all types of wines except Dry Red. Henry almost always buys it online, probably because he needs to be at home with his two-year-old child.

Cluster 2 - Marry Lee, 28 years, Professor in Atlanta with a low salary, Georgia, has a bachelor's degree in management. She buys wines very rarely, but when she buys, it's almost always at a discount. Marry likes every type of wine but especially Dry Red and she loves online sails and visiting the WWW website. She had a six-year-old daughter and ten years, old boy.

CLUSTER 3 - Willian Gates, 77 years-old, from Jacksonville, Florida. William has a usual education but with incredibly high income. He likes to buy wines very frequently from WWW, and therefore he has much value for the company. He doesn´t pay attention to promotions, and he doesn´t even like to buy wines online. Its favorite wine type is Dry white, and he dislikes unusual wines.

4.2 Mca

In order to have a good perspective from each cluster, the cluster analysis was made only with the metric features, as including the categorical features into distance-based clustering methods would not make logical sense. Therefore, they were discarded before in the Pre-Process phase. However, now that the cluster analysis has been made, the team is able to find relations between the clusters and those categorical variables through a method named Multiple Correspondence Analysis. These methods consist of plotting all the clusters and variables into a two-dimensional plot and see the linkage between the points. The closer a categorical variable is from a cluster, the more import is for that cluster.

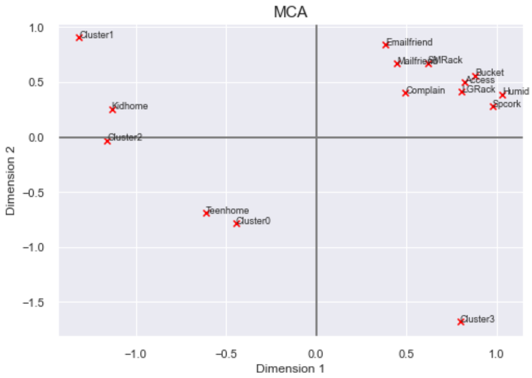

Figure 5- Mca plot

From the analysis of the plot, we can see that cluster 0 is constituted by buyers that have a high probability of having teens living in their home, the cluster 2 are constituted by buyers that have a high probability of having a child living in their home and the cluster 1 have a big probability of having a child living in their home, but that probability is lower than the probability of buyers from the cluster 2.

4.3 Classification of Outliers

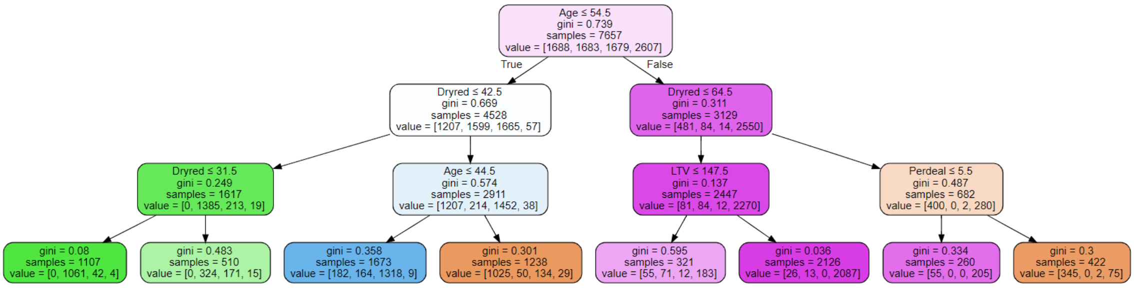

Figure 6- Decision tree for outliers' classification

For the final part of the modeling, the team proceeded to classify the outliers. In earlier stages, the outliers were left out of the analysis due to the fact that their inclusion would distort the clustering process. However, now the outliers can be included with the help of of a decision tree that has an accuracy of around 84.44%. Moreover, since decision trees are a white box classifier, the company can manually assign new clients to the clusters.

5. DEPLOYMENT AND MAINTENANCE PLANS

Following the cluster's characterization, two strategies for WWW were defined. One more extensive and widespread (Macro), and the second detailed to the clusters found previously (Micro).

On a general approach, our team focused on the information shown below.

Knowing that our online sales offer us a greater margin of revenue and given the fact that there seems to be an equilibrium between online vs in-store purchases our proposals are the following:

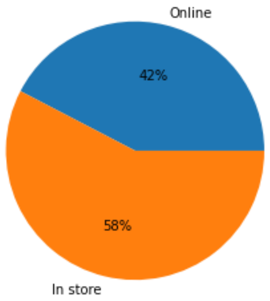

Figure 7- Pie chart with the percentage of purchases in store and online sales

i) Creation of a global network for our client base. This strategy should be applied in collaboration with one of the major wine rating apps, Vivino or Delectable. The main goal is to capture the client's interests in the app, allowing them to interactively demonstrate their personal tastes for a future predictive model to be applied. This kind of strategy is proven to develop further interest in the maintenance of an updated profile as well as a greater chance of recommendation of our services.

ii) Development of two subscription plans with the intention of increasing the loyalty of the customers. This subscription, besides having countless advantages in terms of the core business, it is also a great way to fascinate the clients. A subscription plan can have different options and targets. One option is the creation of a “wine club” in which WWW regularly sends to its members a variety of different types of wines – turning the act of buying and drinking wine into a completely new experience. The second option is a “premium” subscription.

These two proposals cut off one of the main problems in supply chain management, the volatility of the demand. When working at full steam, this strategy will allow us to not only control the predicted demand but also to avoid costs of store maintenance while bringing our clients together in a sense of community (to be implemented).

For our micro strategy, we focused on each one of the clusters found previously and tried to make the best use of our new channels created in the macro strategy.

Cluster 0: As this seems to be a group of people with a broad dexterity for usage of our online platforms and presents a greater interest in Dry Red Wine, we suggest that these users adopt our global network, where we promote gatherings for the people on this cluster through taste events or dinner parties.

Cluster 1: A group which stands out from the rest due to its lack of education. These people present a low level of income, so it is no surprise that they are responsive to discounts. Through our global network app, we will offer specific promotions and reward points based on the customer’s interaction with the company.

Cluster 2: this type of clients despite having a low income, have in its majority a good education level, likes all type of wines and online sails. Having this into account we suggest to approach then via both macro marketing methods – both connecting them to WWW global network and also suggesting a subscription plan receiving at home different types of wines

Cluster 3: these customers are characterized for not giving attention to discounts and online store. Having already a certain age but incredible high income, it is a great opportunity to implement a Premium subscription where the client receives top wines regularly without the trouble of moving to the store.

6. CONCLUSIONS

To conclude this project, is important to stress that the group attempted to make a complete Business case project with all required steps following the CRISP-DM methodology. The dataset provided by the company (Wonderful wines of the world) was in a very good shape and did not have many needs in terms of treatment.

After exploring the data with different types of visualizations, we started the clustering process for all metric variables and it was applied five different clusters techniques (K-means, Gmm, Hierarchical Clustering, SOM + K-Means, K-Means + HC). For the final steps, in order to choose the best clustering methodology, we plotted R^2 plot for the five different clustering methods to compare each cluster rating.

In the final part of the project, the team proceeded to brainstorm marketing ideas that could be in line with the customer segments that were found earlier. For each customer segment, there was a suggestion of a marketing tool that can tackle it in an efficient manner, such as providing a better app experience for younger customer or creating premium subscription for higher-income customers.

7. REFERENCES

Chengqi Zhang, Q. Y. (2013, May). Data Preparation for Data Mining. *Applied Artificial Intelligence*.

Roux, L., & Rouanet, B. a. (2004). Geometric Data Analysis, From Correspondence Analysis to Structured Data Analysis. *Dordrecht. Kluwer:*, p. 180.

pypi.org (n.d.). prince 0.7.1 Retrieved from: https:// https://pypi.org/project/prince/

wineintelligence.com (n.d.). GLOBAL WINE TREND PREDICTIONS FOR 2020 – MID YEAR UPDATE Retrieved from: https://www.wineintelligence.com/global-wine-trend-predictions-for-2020-mid-year-update/.

Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (Vol. 1). London, UK: Springer-Verlag.

The CRISP-DM process model (1999), http://www.crisp-dm.org/

ZAGORULKO, D. Digital Marketing for Wine Companies: An Innovative Approach. *STRATEGICA*, 343.

8. FIGURE INDEX

Link to our github: https://github.com/fredericojpr/BC1_GroupS