# Problem Formulation

## "Marketing Analytics Case Study"

### Practicing Exploratory and Statistical Analysis with Marketing Data

Jannik Seus (3199154), Michael Gilge (3251423)
Ihsan Tri Heldian (3528516), Zhixun Zhao (3544639)

## Problem:

In a business the recent marketing campaigns have not been as effective as they were expected to be. The Chief Marketing Officer (CMO) wants to know what can be done considering future marketing campaigns to increase the sales and the turnover again. We need to analyze the data set to understand the problem and to propose data-driven solutions.

## Goal:

The goal is to find suggestions to the CMO how to design the next campaign to maximise sales and turnover: We want to be able to suggest the top 10 target groups, we should make campaigns for and what product groups should be offered them through which channel, with or without discounts and in what country. We can estimate the turnover and the surety it will happen, too.

## The Data:

Source: *https://www.kaggle.com/jackdaoud/marketing-data*
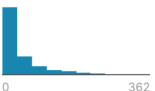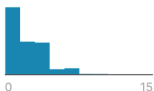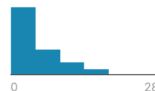
## About this file

The is a CSV file of 2240 observations (customers) with 28 variables related to marketing data. More specifically, the variables provide insights about:

- Customer profiles
- Products purchased
- Campaign success (or failure)
- Channel performance

| ∞ ID | # Year_Birth | ⚠ Education | ⚠ Marital_Status | ⚠ Income | # Kidhome | # Teenhome |
|---|---|---|---|---|---|---|
| Customer's unique identifier | Customer's birth year | Customer's education level | Customer's marital status | Customer's yearly household income | Number of children in customer's household | Number of teenagers in customer's household |

| ID | Year_Birth | Education | Marital_Status | Income | Kidhome | Teenhome |
|---|---|---|---|---|---|---|
| 0 ... 11.2k | 1893 ... 1996 | Graduation 50% / PhD 22% / Other (627) 28% | Married 39% / Together 26% / Other (796) 36% | [null] 1% / $7,500.00 1% / Other (2204) 98% | 0 ... 2 | 0 ... 2 |
| 1826 | 1970 | Graduation | Divorced | $84,835.00 | 0 | 0 |
| 1 | 1961 | Graduation | Single | $57,091.00 | 0 | 0 |
| 10476 | 1958 | Graduation | Married | $67,267.00 | 0 | 1 |
| 1386 | 1967 | Graduation | Together | $32,474.00 | 1 | 1 |

| 📅 Dt_Customer | # Recency | # MntWines | # MntFruits | # MntMeatProducts | # MntFishProducts | # MntSweetProducts |
|---|---|---|---|---|---|---|
| Date of customer's enrollment with the company | Number of days since customer's last purchase | Amount spent on wine in the last 2 years | Amount spent on fruits in the last 2 years | Amount spent on meat in the last 2 years | Amount spent on fish in the last 2 years | Amount spent on sweets in the last 2 years |
| 30Jul12 ... 29Jun14 | 0 ... 99 | 0 ... 1493 | 0 ... 199 | 0 ... 1725 | 0 ... 259 | 0 ... 263 |
| 6/16/14 | 0 | 189 | 104 | 379 | 111 | 189 |
| 6/15/14 | 0 | 464 | 5 | 64 | 7 | 0 |
| 5/13/14 | 0 | 134 | 11 | 59 | 15 | 2 |
| 5/11/14 | 0 | 10 | 1 | 1 | 0 | 0 |

| # MntGoldProds | # NumDealsPurcha... | # NumWebPurchases | # NumCatalogPurc... | # NumStorePurcha... | # NumWebVisitsMo... | # AcceptedCmp3 |
|---|---|---|---|---|---|---|
| Amount spent on gold in the last 2 years | Number of purchases made with a discount | Number of purchases made through the company's web site | Number of purchases made using a catalogue | Number of purchases made directly in stores | Number of visits to company's web site in the last month | 1 if customer accepted the offer in the 3rd campaign, 0 otherwise |
| 0 ... 362 | 0 ... 15 | 0 ... 27 | 0 ... 28 | 0 ... 13 | 0 ... 20 | 0 ... 1 |
| 218 | 1 | 4 | 4 | 6 | 1 | 0 |
| 37 | 1 | 7 | 3 | 7 | 5 | 0 |
| 30 | 1 | 3 | 2 | 5 | 2 | 0 |
| 0 | 1 | 1 | 2 | 2 | 7 | 0 |

| # AcceptedCmp4 | # AcceptedCmp5 | # AcceptedCmp1 | # AcceptedCmp2 | # Response | # Complain | ⚠ Country |
|---|---|---|---|---|---|---|
| 1 if customer accepted the offer in the 4th campaign, 0 otherwise | 1 if customer accepted the offer in the 5th campaign, 0 otherwise | 1 if customer accepted the offer in the 1st campaign, 0 otherwise | 1 if customer accepted the offer in the 2nd campaign, 0 otherwise | 1 if customer accepted the offer in the last campaign, 0 otherwise | 1 if customer complained in the last 2 years, 0 otherwise | Customer's location |
| 0 ... 1 | 0 ... 1 | 0 ... 1 | 0 ... 1 | 0 ... 1 | 0 ... 1 | SP 49% / SA 15% / Other (808) 36% |
| 0 | 0 | 0 | 0 | 1 | 0 | SP |
| 0 | 0 | 0 | 1 | 1 | 0 | CA |
| 0 | 0 | 0 | 0 | 0 | 0 | US |
| 0 | 0 | 0 | 0 | 0 | 0 | AUS |

## Software that we want to use:

- RapidMiner for visualizing, model building, etc.
- Tableau for visualization and other exploratory analysis.
- Microsoft Excel / Numbers / TextEdit for viewing the dataset.

## Planned Submission:

A poster with the following four sections (The purpose of the sections is to break down the problem into smaller problems or questions):

### Section 01: Exploratory Data Analysis (visualization, data cleaning, preprocessing)

Providing information from exploratory analysis using charts and diagrams on the dataset.

- Plotting and visualizing the data.
- Are there any null values or outliers? How will we handle them?
- Are there any variables that warrant transformations?
- Can we summarize data?
- Are there any useful variables that we can engineer with the given data?
- Do we notice any patterns or anomalies in the data? Can we plot them?

Data cleaning and preprocessing: For example:
- Delete $-symbol in the income value.
- Convert year of birth into age.
- Categorize the age into different groups (young, middle, old).
- Categorize income into different groups (high, middle, low).
- Convert date enrollment from customer into length of customer enrollment and categorize it into different groups.

## Section 02: Statistical Analysis (statistical tests in the form of regressions, modelling, data mining)

Finding what kind of attributes from the data have an impact on the success of a marketing campaign to increase revenue.

Doing some model building processes using RapidMiner for unsupervised learning (clustering) and supervised learning (labelling for different columns, especially the campaigns and Response) and evaluating performance for each process.

**General aspects to be considered (We may not answer everything.):**

People (Marketing concept: Target groups) (How much do customer-types impact the success?)
- What customer-types are there? (Clustering)
- What does the average customer look like for this company?
- What customer-types responded to deals most?
- Which customer-cluster spends most money?
- Are people, who buy gold, more conservative buying more in store purchases, spending an above average amount on gold in the last 2 years?
- Fish has Omega 3 fatty acids which are good for the brain. Accordingly, do "Married PhD candidates" have a significant relation with the amount spent on fish? What other factors are significantly related to the amount spent on fish? (Hint: use your knowledge of interaction variables/effects)
- What impacts customer (dis)satisfaction (accepted offers, purchases, complaints) -> Can we find any patterns?

Products (Marketing concept: Product portfolio)
- Which products are performing best?
- Which products are performing best for which customer-types?

Channels (Marketing concept: Channel management) (How much do channels impact the success?)
- Which channels are underperforming?
- Which channel is performing best?
- What factors are related to the number of store purchases?
- What factors are related to the number of catalogue purchases?

- What factors are related to the number of web purchases?
- For which customer type worked which channel best?
  - What customer-types buy over the store?
  - What customer-types buy over the catalogue?
  - What customer-types visit the web-site most?
  - What customer-types buy over the web-site?

Campaigns (How much do campaigns impact the success?)
- Which customer target group will respond to the next campaign, what will this customer group probably buy and what channel will they probably use?
- Which product group is more likely to be bought in general as a result of the next campaign?
- Which campaigns had an impact on purchases?
- Are there any differences between the campaigns? (Success, products, customer types, …)
- Which marketing campaign is most successful?
- Which marketing campaign is least successful?
- What customer-types responded to what campaign and how important were deals?
- What customer-types responded to what campaign and how important were deals measured by web-visits, purchases, turnover, ...?
- Predict the highest turnover in response to a campaign/discount for a product (train a predictive model).
- Can we find a development from campaign to campaign (time aspect)

Countries:
- How are the total purchases in each country?
- Are there substantial differences between the countries?
- In which country are most purchases?
- Is there a substantial relationship between geographical regions and the success of a campaign?

Turnover / Sales
- What factors are substantially related to the turnover?
- What product category provides the highest turnover?
- What campaign resulted with the highest turnover?
- Do complaints impact the turnover?

**Transformation of the attributes (detailed plan):**

Attributes in bold type are what we finally need. A transformation ist symbolized by the arrow „→" and a method is in brackets.


##############################################################
Job 1: Jannik: Clustering Target Groups:

Demographic customer data:
For all attributes it needs to be considered, if there are irrelevant data.

ID → Ø   (Can be ignored for clustering.)
Year_Birth –(Binning)→ {Child, Adult, Old_Person}  (There may be more groups.)
Education → Education (Seems to be irrelevant data.)
Marital_Status → Marital_Status  (Binning not necessary.)
Income –(Binning)→ {Low, Middle, High}  (There may be more groups.)
Kidhome → Kidhome  (Binning not necessary.)
Teenhome → Teenhome  (Binning not necessary.)

Year_Birth x Marital_Status x Income x Kidhome x Teenhome    –(Clustering)→
**TargetGroup**


##############################################################
Job 2: Michael: Aggregating Data:

Interaction data:
Dt_Customer → DaysEnrollment
DaysEnrollment x Recency  –((DaysEnrollment+1)/(Recency+1)) →  Loyalty
Loyalty  –(Average)→  aggLoyalty

Dt_Customer –(Binning)→ CustomerStatus = {Regular_Customer, New_Customer} (More?)

Year_Birth x Marital_Status x Income x Kidhome x Teenhome x CustomerStatus
–(Clustering)→ StatusTargetGroup
Loyalty  –(Average)→  aggStatusLoyalty

(Loyalty may need a different calculation.)

Product group data:
MntWines → aggMntWines

MntFruits → aggMntFruits
MntMeatProducts → aggMntMeatProducts
MntFishProducts → aggMntFishProducts
MntSweetProducts → aggMntSweetProducts
MntGoldProds → aggMntGoldProds

aggProductGroups = aggMntWines x aggMntFruits x aggMntMeatProducts x
aggMntFishProducts x aggMntSweetProducts x aggMntGoldProds
aggMntWines –(aggAmount/average)→ ImportanceMntWines
aggMntFruits –(aggAmount/average)→ ImportanceMntFruits
aggMntMeatProducts –(aggAmount/average)→ ImportanceMeatProducts
aggMntFishProducts –(aggAmount/average)→ ImportanceMntFishProducts
aggMntSweetProducts –(aggAmount/average)→ ImportanceSweetProducts
aggMntGoldProds –(aggAmount/average)→ ImportanceGoldProds

aggImportances = ImportanceMntWines x ImportanceMntFruits x
ImportanceMeatProducts x ImportanceMntFishProducts x
ImportanceMntSweetProducts x ImportanceMntGoldProds

aggProductGroups x aggImportances → **PreferredProducts** =
powerset(ProductGroups)

aggProductGroups → **aggTurnover**

Discount data:
NumDealsPurchases → aggDealsPurchases
aggDealsPurchases → **DealsSuccess**

Channel data:
NumWebPurchases → aggWebPurchases
NumCatalogPurchases → aggCatalogPurchases
NumStorePurchases → aggStorePurchases
NumWebVisitsMonth → aggWebVisitsMonth

Channels = NumWebPurchasesNum x CatalogPurchases x NumStorePurchases x
NumWebVisitsMonth

Channels → **aggSumPurchases**

Channels → **FavouriteChannels**


Campaign data:

AcceptedCmp1 → aggAcceptedCmp1
AcceptedCmp2 → aggAcceptedCmp2
AcceptedCmp3 → aggAcceptedCmp3
AcceptedCmp4 → aggAcceptedCmp4
AcceptedCmp5 → aggAcceptedCmp5

Response → aggResponse

Complain → aggComplain

###########################################################
## Job 3: Ihsan: Campaign Analysis (Prediction, Time Analysis):

aggCampaigns = aggAcceptedCmp1 x aggAcceptedCmp2 x aggAcceptedCmp3 x aggAcceptedCmp4 x aggAcceptedCmp5

aggCampaigns → SuretyAcceptance (How sure is it that target group will respond again?)

aggCampaigns –(AverageOfAcception)→ ExpectedValueAcception

InterestingFeatures = TargetGroup x aggLoyalty x PreferredProducts x DealsSuccess x FavouriteChannels x aggCampaigns

InterestingFeatures → ResponsePrediction


ResponsePrediction x aggResponse → **PredictionCorrectness**

(Campaigns 1-5 are the training sets and Response is the test set.)

InterestingFeatures x PredictionCorrectness → FavouriteCampaign1


###########################################################
## Job 4: Zhixun: Finalize:


<u>Complaint data:</u>

aggComplain → ThrowOut

FavouriteCampaign1 x TrowOut → **FavouriteCampaign2**


Country data:
Country → **TopCountries** = powerset(Country)  (TopCountries of TargetGroup. We may need an extra attribute with the numbers of customers per country.)


Final attributes:

TargetGroup x PreferredProducts x DealsSuccess x FavouriteChannels x FavouriteCampaign2 x TopCountries → **SuggestedCampaign** (percentage)

SuggestedCampaign → **TopTenCampaign**

FavouriteCampaign2 x aggTurnover  →  **ExpectedValueTurnover**

FavouriteCampaign2 x aggSumPurchases →  **ExpectedValueSales**

FavouriteCampaign2 x PredictionCorrectness  →  **SuretyCampaign**


## Section 03: Data Visualization
- Plotting and visualizing the results.
- We want to visualize the current condition from the dataset to show how the customer demographic data bring impact to the amount of spending in each group of products, the amounts of purchases by each channel, and how they react to each marketing campaign.


## Section 04: CMO Suggestions
- Bringing together everything from Sections 01 to 03 and provide data-driven suggestions to the CMO how to design the next campaign to maximize sales and turnover:
- We want to be able to suggest the top 10 target groups, we should make campaigns for and what product groups should be offered to them through which channel, with or without discounts and in what country.
- We can estimate the turnover, the sales and the surety it will happen for every target group, too.

## Schedule:

No exact dates, but chronological order:

1. Clustering target groups
2. Calculate aggregations where necessary
3. Calculate loyalty
4. Find favourite product groups
5. Calculate deals success
6. Find favourite channels
7. Find favourite campaigns (prediction)
8. Complaints-examination and evaluation
9. Country-Selection
10. Make CMO suggestions.

## Distribution of the tasks:

1. We want to do the data analysis (section 01) together.
2. Then we want to decide who will do which aspect in section 02.
3. Finally we want to bring our results together for sections 03 and 04.