# TURIN MUSEUMS

## A business analysis

BRUNAZZI ALICE,
MAT. 864566

# The business problem

*CHURNERS*

The Torino Museum Card Association offers an annual “Museum Card”, granting free access to all city museums.
Despite its value proposition, a substantial share of card-holders fails to renew annually (churn), undermining recurring revenue and the association’s cultural outreach.

1

2

3

4

5

6

7

8

9

# The Objectives

*CHURNERS*

- Understand churn drivers
- Quantify economic impact
- Segment customers: Discover clusters  of high-risk churners and high-lifetime-value users.
- Measure causal effects of gender on churn
- Develop predictive models
- Optimize a retention campaign

1

2

3

4

5

6

7

8

9

# Meet the costumers

## Demographic and Card Purchase

- Codcliente: Unique customer ID.
- data_inizio: Card start date (valid for one year).
- importo, sconto, Riduzione: Price paid and discount type (e.g. student, senior, NGO partner).
- Tipo_pag, Agenzia, Agenzia_tipo: Payment method and sales channel (online, newsstand, museum, CRAL, etc.).
- Sesso, data_di_nascita, professione, CAP: Gender, birth year, occupation, and postal code.
- Nuovo_abbonato: Flag for “new subscriber” (no card in 2012).

# The cleaning

## Demographic and Card Purchase

- data_inizio: the month was kept, coded from 1 to 12
- importo, sconto: importo was kept, sconto was binarized as Present or Absence
- Riduzione: codified to encode only the informative classes (86% of records), "Musei Torino", "Musei Ridotto", "Offerta su quantitativo" and "Other"
- Tipo_pag, Agenzia: were not used in the analysis

- Agenzia_tipo: only the most representative labels were kept, encoded as "canale diretto (museo o online)", "info point", "retail fisico", "attraverso partner" or "other (missing value, not observed but still informative)"
- Sesso, 2566 records removed, NULL values
- data_di_nascita, records within the interval 1925-2007 were kept, to ensure that the ages were significant
- professione, only had NULL value, the variable was removed
- CAP and Comuni: not used in analysis
- Nuovo_abbonato: under-repersented class, non informative. Not used in analysis

# Meet the costumers

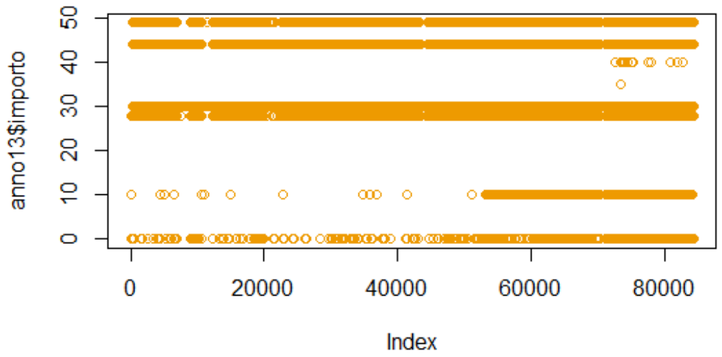
*87052 CUSTOMER ANALYZED*

*GENDER*

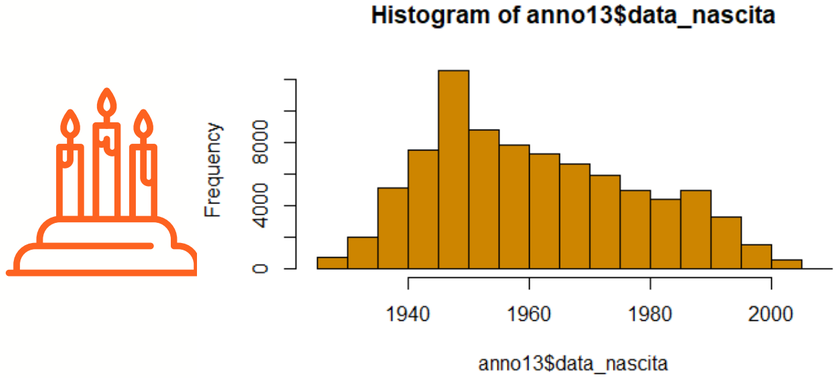
*47655 FEMALES*

*36687 MALES*

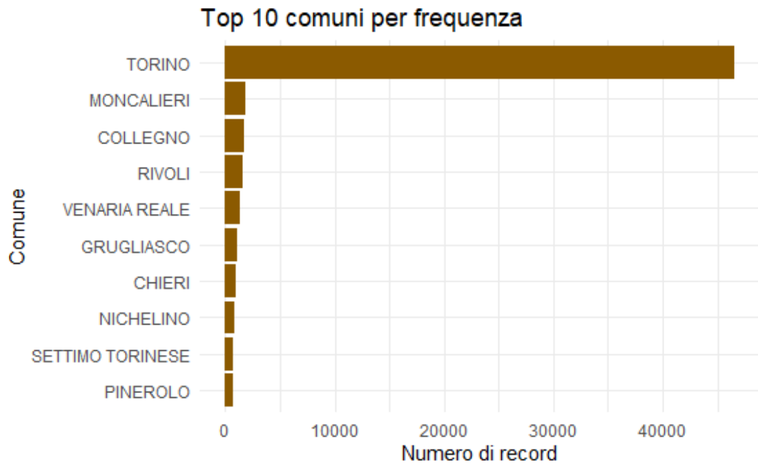
*CARD COST*




*YEAR OF BIRTH*

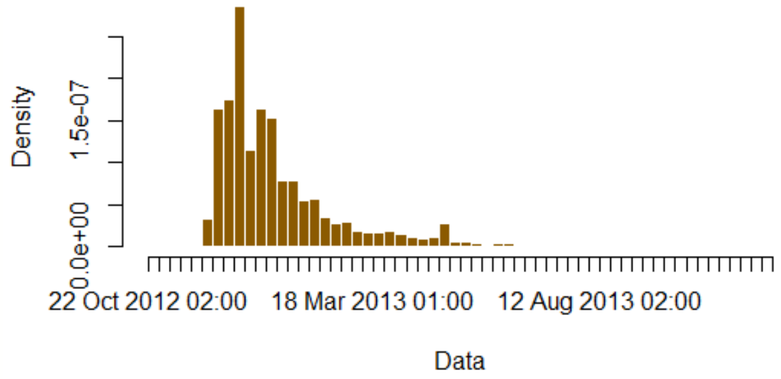



*TOWN OF RESIDENCE*

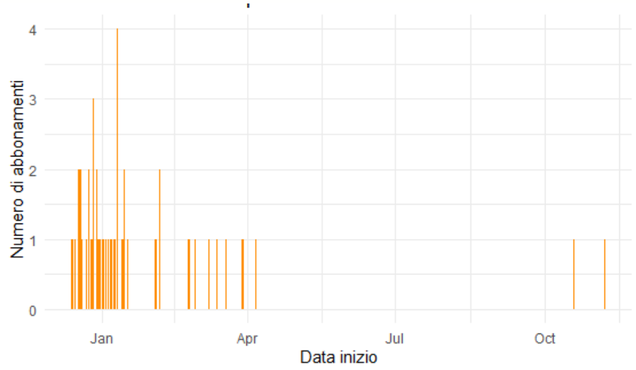


# Meet the costumers

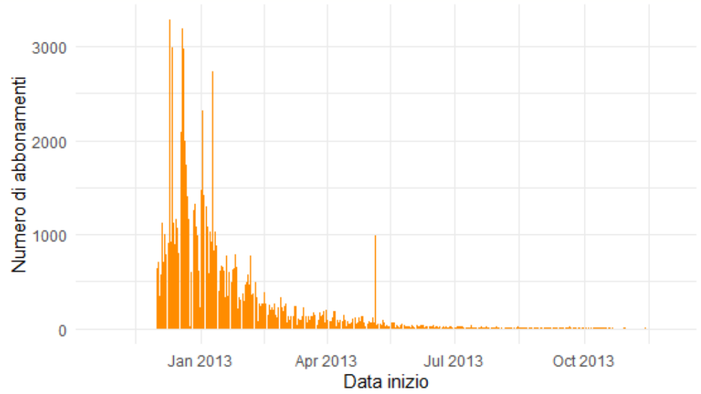
When did our customer  buy their card?




When did our **OLD** customer  buy their card?




**43 CUSTOMERS**

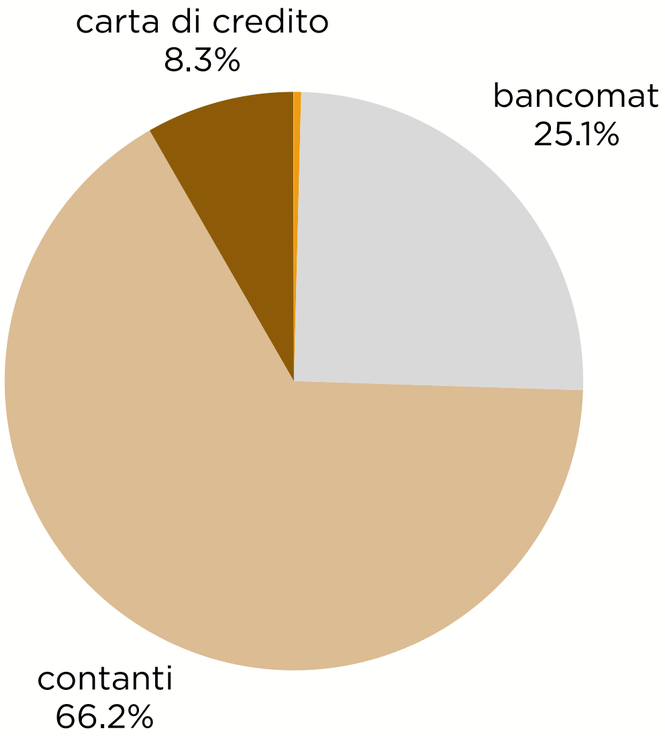



**81005 CUSTOMERS**

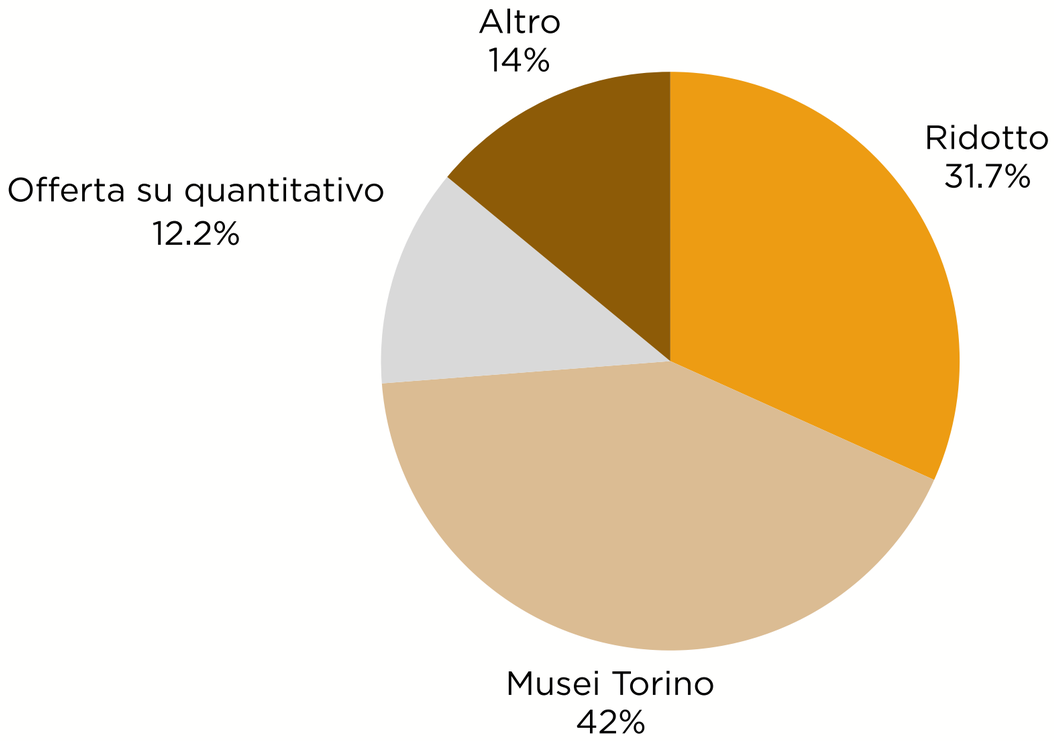
When did our **NEW** customer  buy their card?

# Meet the costumers

**How did they pay?**

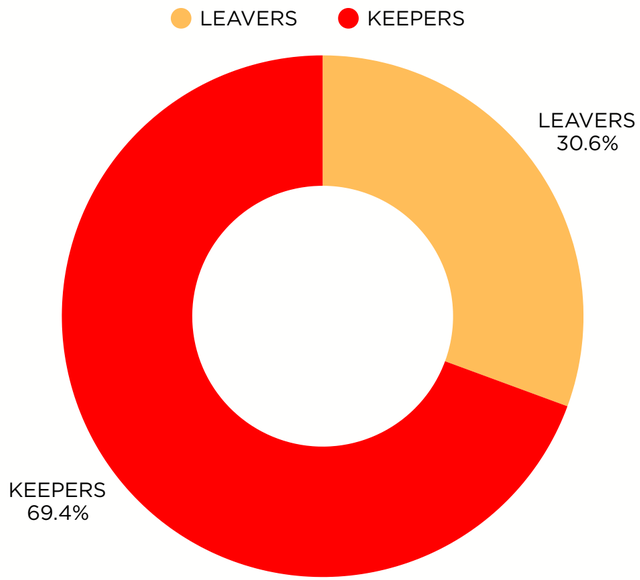



Did they have a price reduction? Which one?



# Renewal Status & Visit History

- Si2014: Churn label (0 = did not renew in 2014).
- abb13, abb14: Subscription date in 2013 and renewal date in 2014 (if renewed).
- ultimo_ing.x: Date of the customer's last museum visit during the 2013–2014 period.



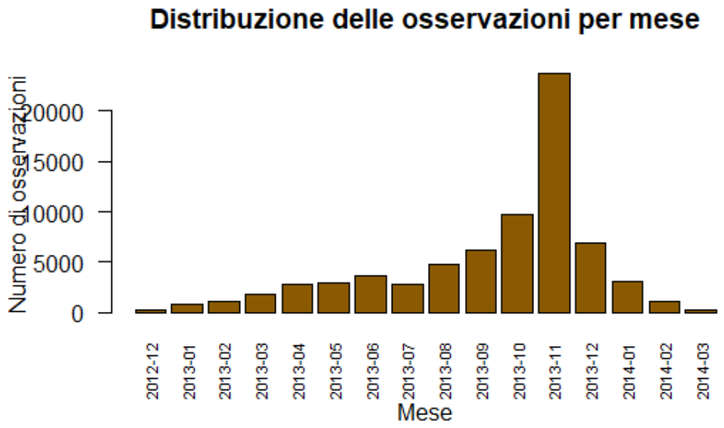
# The Cleaning

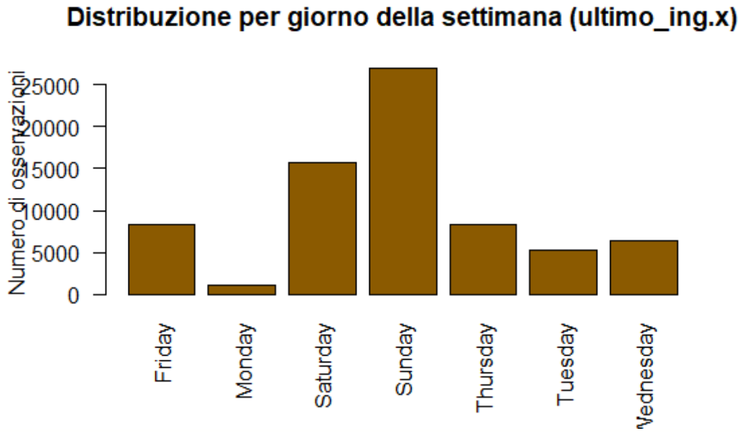
- Si2014: didn't require pre-processing or cleaning
- abb13, abb14: converted in months, coded from 1 to 12, with 0 meaning the absence. Abb14 was removed from the analysis, being a leakage variable in terms of churners
- ultimo_ing.x: converted in months, coded from 1 to 12

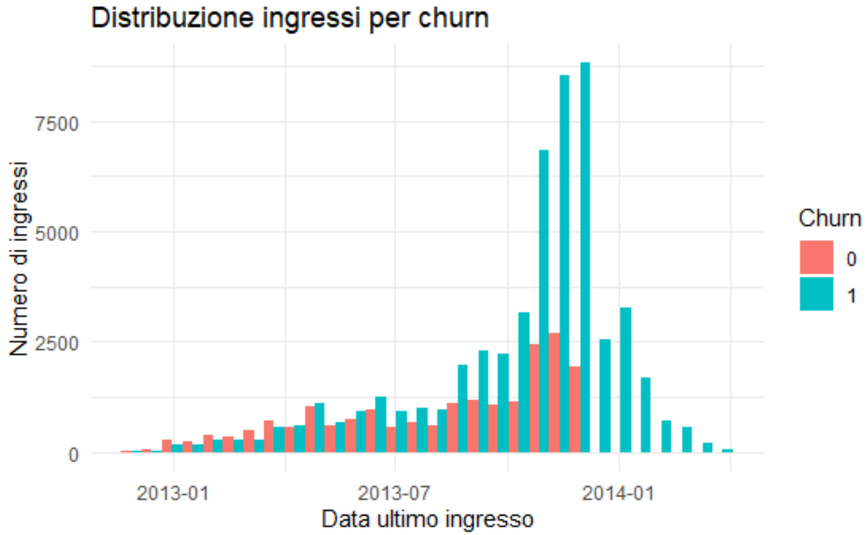
# The entries and the churners

There is a peak of last-entrances in November (23657) and october (9732)

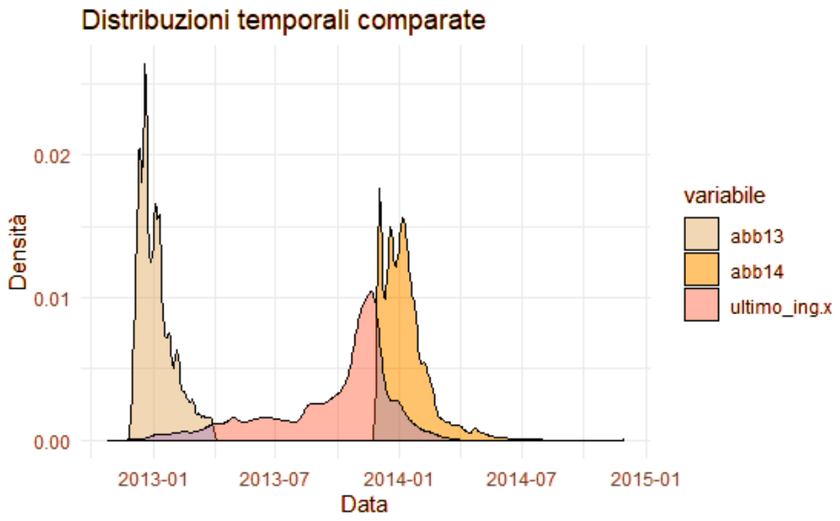
Usually a saturday or a sunday

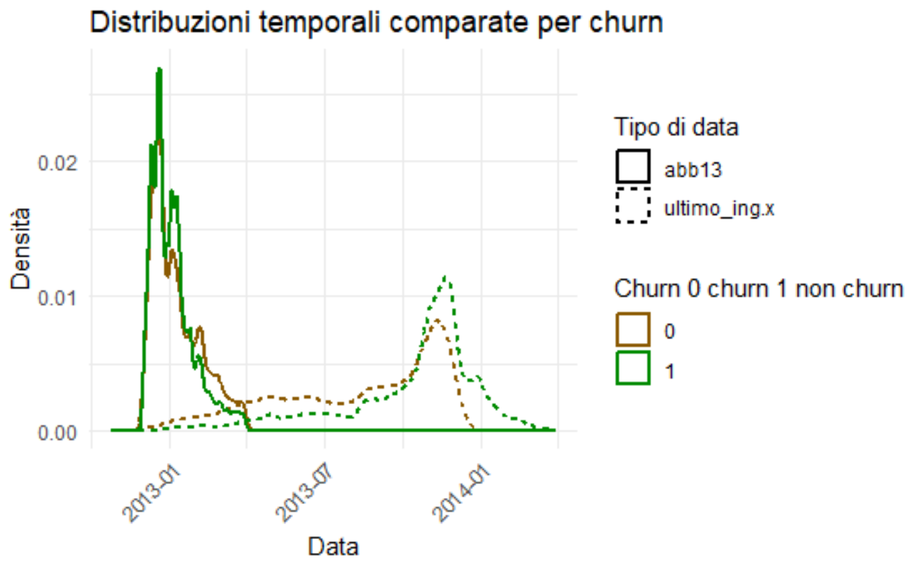


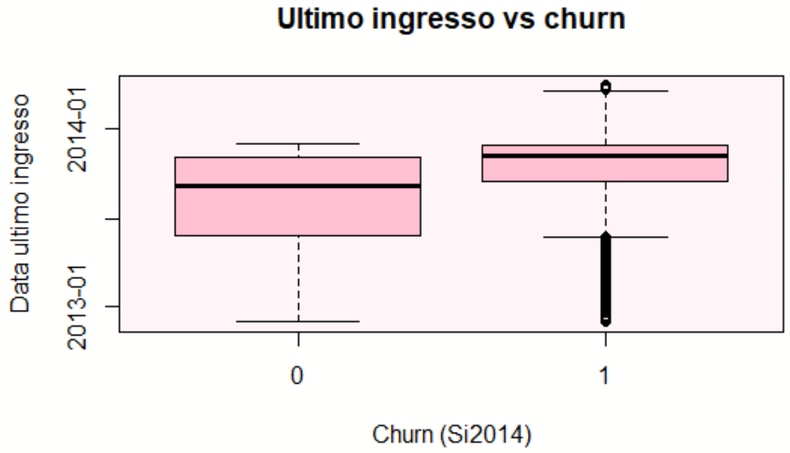



# The entries and the churners




There is a significant difference between the last usage of the card between churners (usually around the summer) and non churners (fall or early winter)

1
2
3
4
5
6
7
8
9

# Singular Museum Visit

- Codcliente, datai, orai: Customer ID, visit date, and time.
- museo, prov_museo, com_museo: Museum name and location (province, city).
- importo: Ticket price (for tracking purposes, though not paid by cardholders).

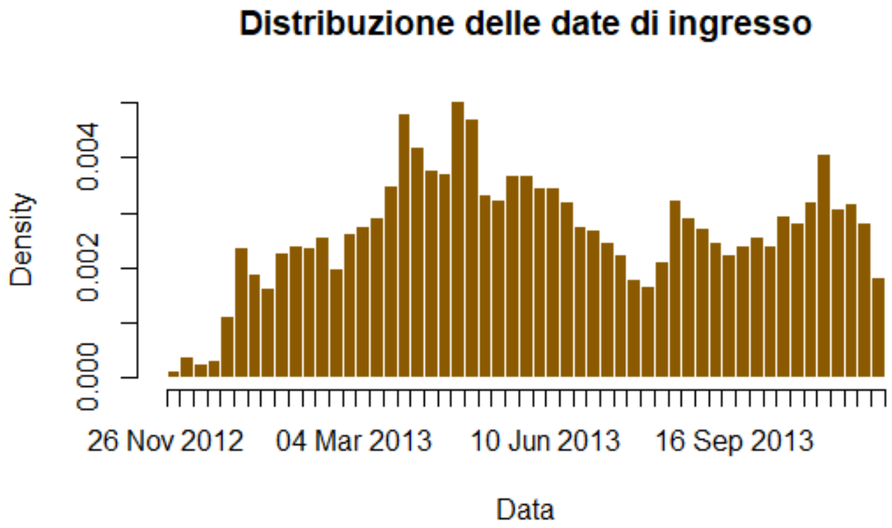
1
2
3
4
5
6
7
8
9

# The cleaning

- datai: did not require pre processing, only kept for network analysis
- orai: some museums had late-nigth openings, but the entries coded at midnigth were considered mistakes in data entry, and eliminated. The variable was only kept for network analysis.
- museo, used as an aggregated variable later on
- prov_museo, not kept
- com_museo: used as an aggregated variable later on
- importo: used as an aggregated variable later on

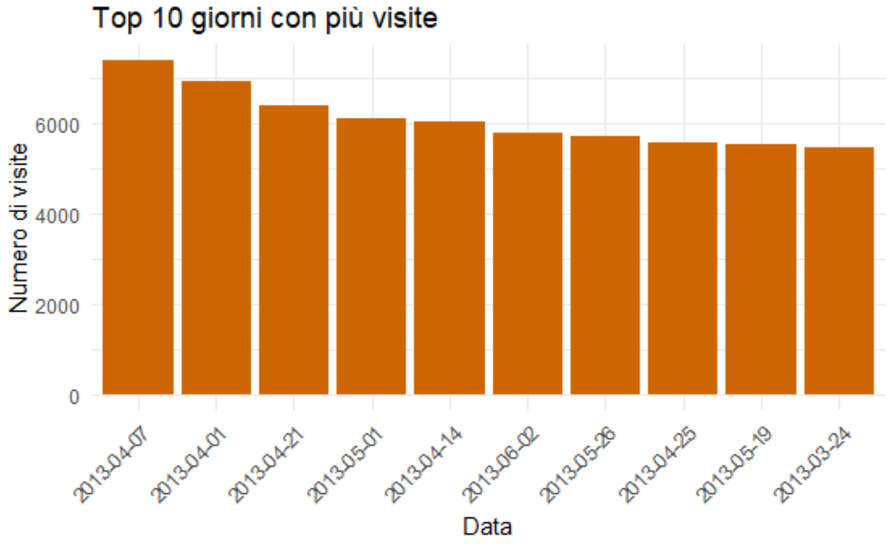
# The visits



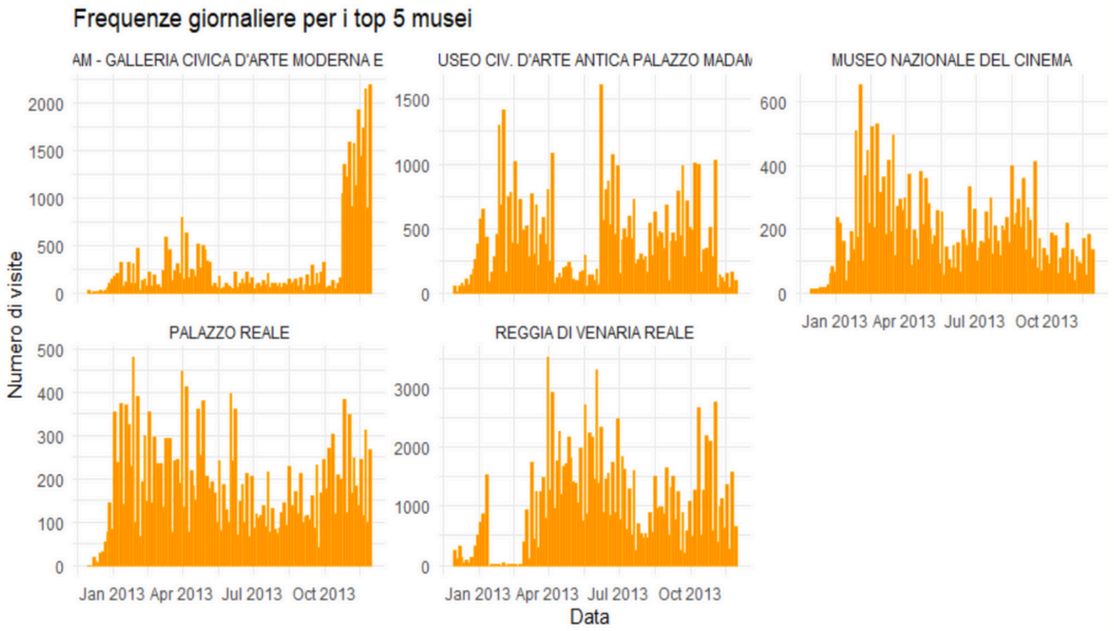


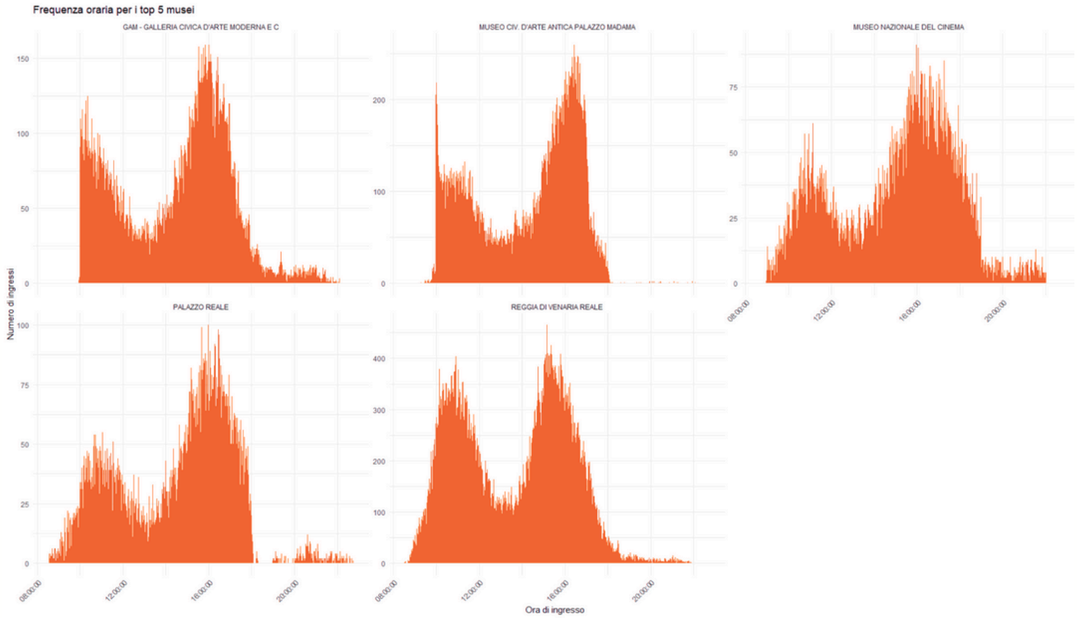


The distribution of museum visits peaks during the spring, and is usually more frequent in the afternoon



# The visits




The «top 5 museums», in term of visits, present a really different distribution of visits during the year, and even during the day. Palazzo reale peaks during the end of the year, while Venaria Reale has a more equal distribution of visits during the mornings and the afternoons.

# The final dataset

Records that  an13 and data1 had in common: 74433

The variables kept were: CodCliente, total paid for the card, gender , year of birth , presence or absence of a discount , type of card, type of channel through which the card was aquired, churn, month of card acquisition. The other variables were used in the EDA, but not kept. The decision was made compromizing between infromation loss and sparcity of the dataset

Infomation provided from in13 were included, aggregating variables:

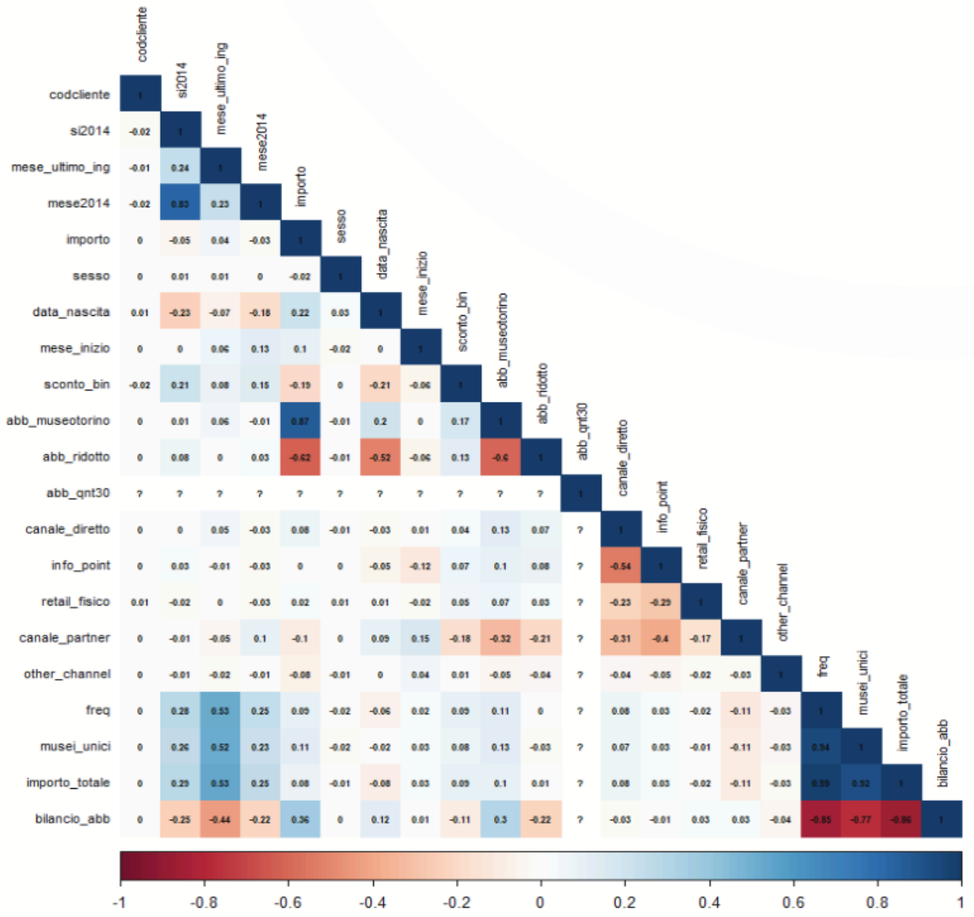
- Visit frequency, intended as total visit for each client
- Museums visited for each client
- Total cost for each client, intended as the total value of the visits
- Cost balance for each client ( cost of  the card - cost of each entrance, if negative, the client saved money).

# The correlation

**~ Spearman correlation**

**si2014 (renewal indicator) is correlated with:**

- bilancio_abb ( - 0.44): customers with a better balance (i.e. their ticket costs exceed the card price) are more likely to renew.
- freq (0.53) and musei_unici (0.52): more visits and museum variety are linked to renewal.
- importo_totale (0.53): higher total ticket value is linked to renewal.
- abb_ridotto (-0.52): reduced-rate subscriptions are linked to higher churn risk.
- info_point (-0.54): buying through info points is linked to higher churn.

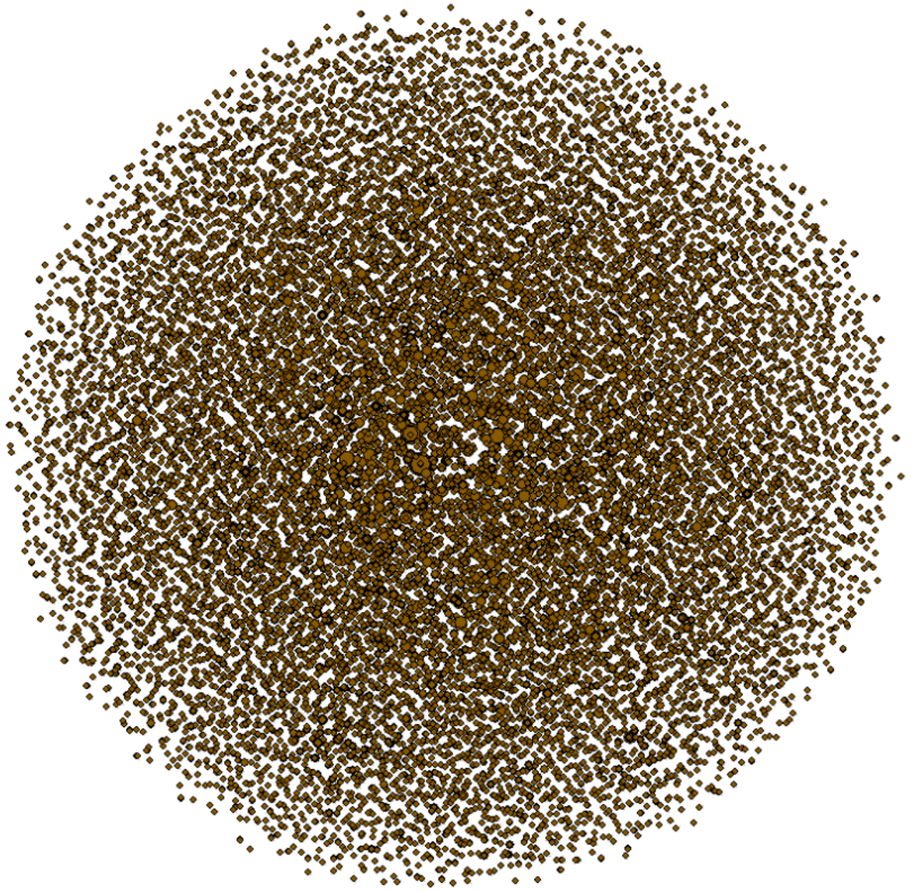


1
2
3
4
5
6
7
8
9

# Network

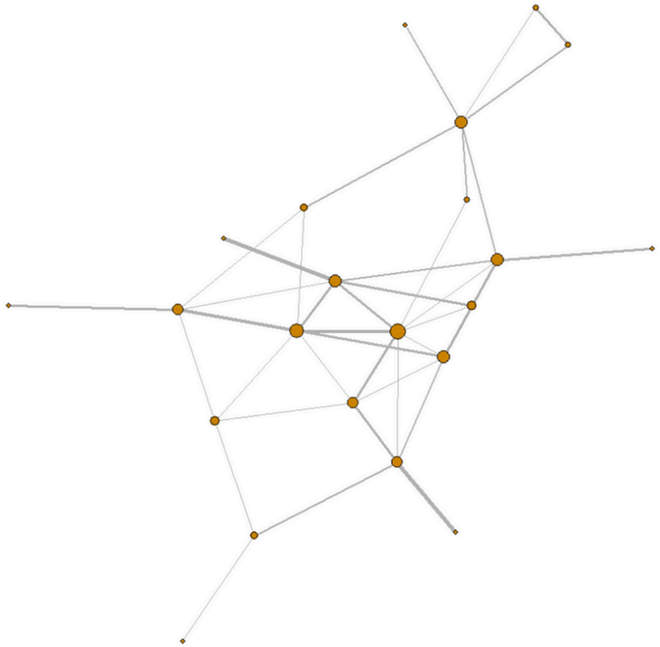
*Considering as connected, customers who visited the same museum at the same time more than twice.*

AT LEAST THREE  SHARED VISIT



**Number of nodes: 30337**
**Number of edges: 17550**
**Number of connected: 13887**
**Network density: 3.81396e-05**
**Giant component dimension: 25**

CENTER



the network is rather sparse (not all customers are connected to each other).

# Most central nodes



In terms of BETWEENNESS, the top 10 profiles presented this characteristics:

| sesso | data_nascita | importo_totale | freq | musei_unici | sconto_bin | nuovo_abb |
|---|---|---|---|---|---|---|
| 1.0 | 1976.0 | 269.0 | 55.0 | 6.0 | 0.0 | 1.0 |
| 1.0 | 1978.0 | 228.75 | 54.0 | 21.0 | 1.0 | 1.0 |
| 1.0 | 1980.0 | 227.25 | 46.0 | 5.0 | 1.0 | 1.0 |
| 0.0 | 1961.0 | 182.95 | 46.0 | 21.0 | 1.0 | 1.0 |
| 0.0 | 1982.0 | 182.25 | 43.0 | 18.0 | 1.0 | 1.0 |
| 0.0 | 1979.0 | 177.5 | 44.0 | 19.0 | 1.0 | 1.0 |
| 1.0 | 1980.0 | 155.5 | 43.0 | 18.0 | 1.0 | 1.0 |
| 0.0 | 1980.0 | 146.0 | 36.0 | 19.0 | 1.0 | 1.0 |
| 1.0 | 1978.0 | 73.75 | 18.0 | 8.0 | 1.0 | 1.0 |

*Co-visit Strength*

- **5+ shared visits:** 5 068 pairs
- **10+ shared visits:** 751 pairs
- **20+ shared visits:** 53 pairs
- **40+ shared visits:** 1 pair

The only pair that scored  43 visits togheter was:

| sesso | data_nascita | freq | musei_unici | importo_totale | si2014 |
|---|---|---|---|---|---|
| 1.0 | 1974.0 | 73.0 | 28.0 | 294.35 | 1.0 |
| 0.0 | 1935.0 | 70.0 | 29.0 | 279.85 | 1.0 |

# Most central nodes

In terms of BETWEENNESS, the top 10 profiles presented this characteristics:

| CLIENT | DEGREE | BETWEENNESS | BETWEENNESS |
|---|---|---|---|
| 258274 | 6 | 61, 00000 | 61, 00000 |
| 104201 | 5 | 53, 00000 | 53, 00000 |
| 177058 | 8 | 49, 33333 | 49, 33333 |
| 104647 | 9 | 45, 33333 | 45, 33333 |
| 167295 | 4 | 44, 00000 | 44, 00000 |
| 163155 | 5 | 39, 50000 | 39, 50000 |
| 76136 | 6 | 39, 33333 | 39, 33333 |
| 58986 | 6 | 38, 00000 | 38, 00000 |
| 167259 | 3 | 38, 00000 | 38, 00000 |
| 24093 | 2 | 36, 00000 | 36, 00000 |

Degree: Nodes like 104647 (9) and 177058 (8) are among the most connected: they shared visits with the highest number of other customers.

Betweenness: 258274 has the highest betweenness

Closeness: Customers like 104647 and 104201 have higher closeness → they are, on average, closer to all other customers in the network, making them easier to reach.

Eigenvector: there are no nodes that dominate the network in terms of connections to other “important” nodes

# Network Takeaways

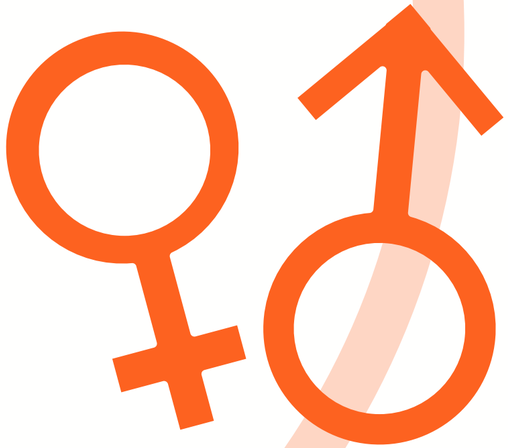
- Ambassador Program: Recruit the handful of high-betweenness customers as social “bridges.” Even if the number of highly connected nodel is low, is still possible to leverage it.
- 1:1 Retention: For the majority (median centralities = 0), deploy personalized outreach (vouchers, reminders).
- Referral Campaigns: Leverage top connectors to amplify word-of-mouth across otherwise disconnected subgroups.

The logic is: it is probable that the most central customers are tied to local groups/ small circles of users, this user can be encouraged and used as a bridge

1
2
3
4
5
6
7
8
9

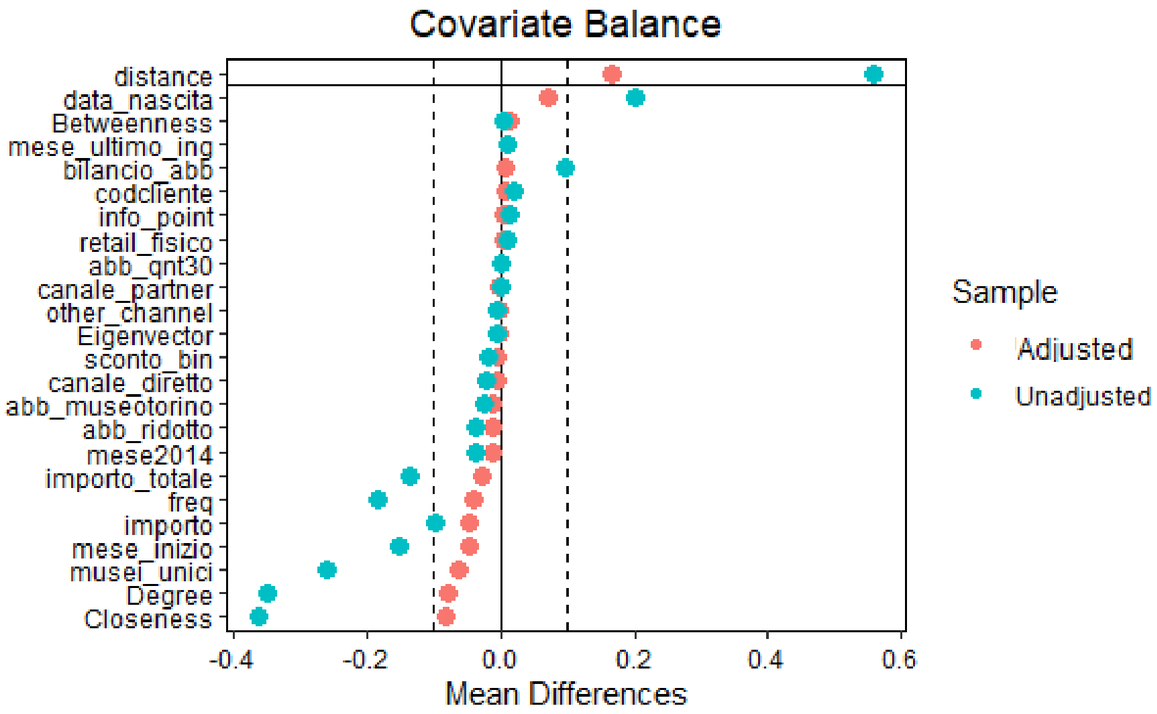
# Causality

*Does GENDER drive churn?*
*Matching based on Propensity Scores calcolated with a glm*



# Causality

*After achieving covariates balance, matching male vs female customer in age, price paid, visit frequencies, discounts and so on, we obtained a fair «treated vs. control » comparison. All covariates have a Standardized Mean Difference that is lower than 0.1, otherwise removed.*




| | Means Treated | Means Control |
|---|---|---|
| distance | 5.697 | 5.630 |
| codcliente | 1.429.123.024 | 1.423.804.895 |
| mese_ultimo_ing | 77.605 | 77.162 |
| mese2014 | 43.262 | 43.904 |
| importo | 350.840 | 355.435 |
| data_nascita | 1.962 | 1960, 6 |
| mese_inizio | 6, 6 | 7 |
| sconto_bin | 6.592 | 6.635 |
| abb_museotorino | 4.257 | 4.357 |
| abb_ridotto | 3.205 | 3.327 |

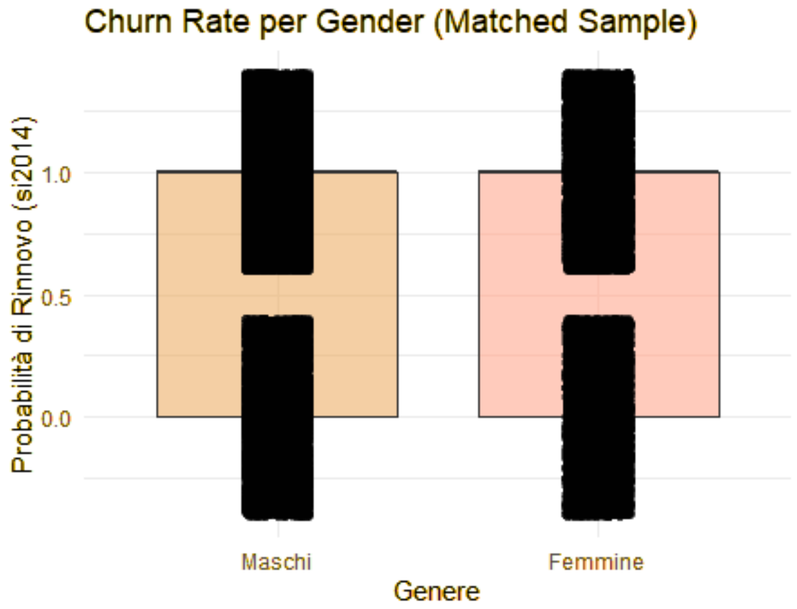
# Controlled comparison

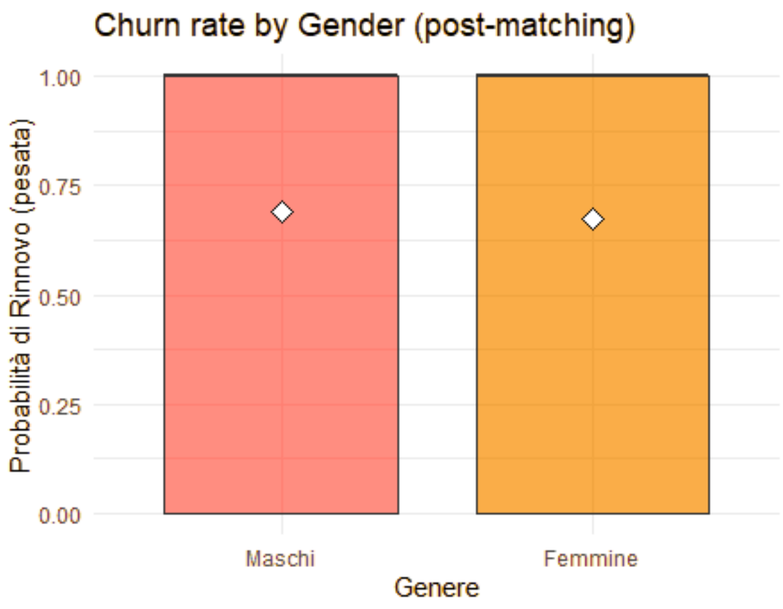
| Method | Effect Size (F = 1 vs. M = 0) | Statistical Significance |
|---|---|---|
| **Matched Linear Model** | –0.9 pp renewal probability | p = 0.015 (*) |
| **Matched Logistic Model** | 0.99 odds of renewal | p = 0.015 (*) |
| **IPTW (Population-Weighted)** | –0.0035 pp | p ≈ 1.00 (n.s.) |
| **PS-Stratified (Quintiles)** | ~0 pp | — |
| **Sensitivity (Rosenbaum)** | **No discordant pairs → no effect** | — |

UNCONTROLLED COMPARISON : Simple logistic regression on full data (no covariate control) shows acoefficient of 1, p - value at 0. If explained by just gender, +4 pp advantage for women (p=0.013). But effect reverses or vanishes once we account for all confounders.

# Causality

*Estimates hover around zero and shift sign depending on method.*
*No robust, consistent causal impact of gender on churn.*



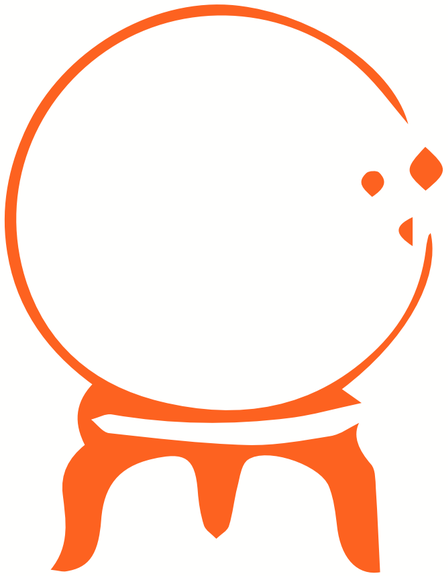
# Causality Takeaways

- **Do not** allocate separate retention budgets by gender.
- Focus efforts on segments with **clear drivers** (e.g. low-frequency visitors, late-renewers, specific discount types).
- **Personalize** outreach using proven predictors (visit recency, total visits, network centrality), not gender.
- Monitor through **A/B tests** rather than relying on simple demographic splits.

1
2
3
4
5
6
7
8
9

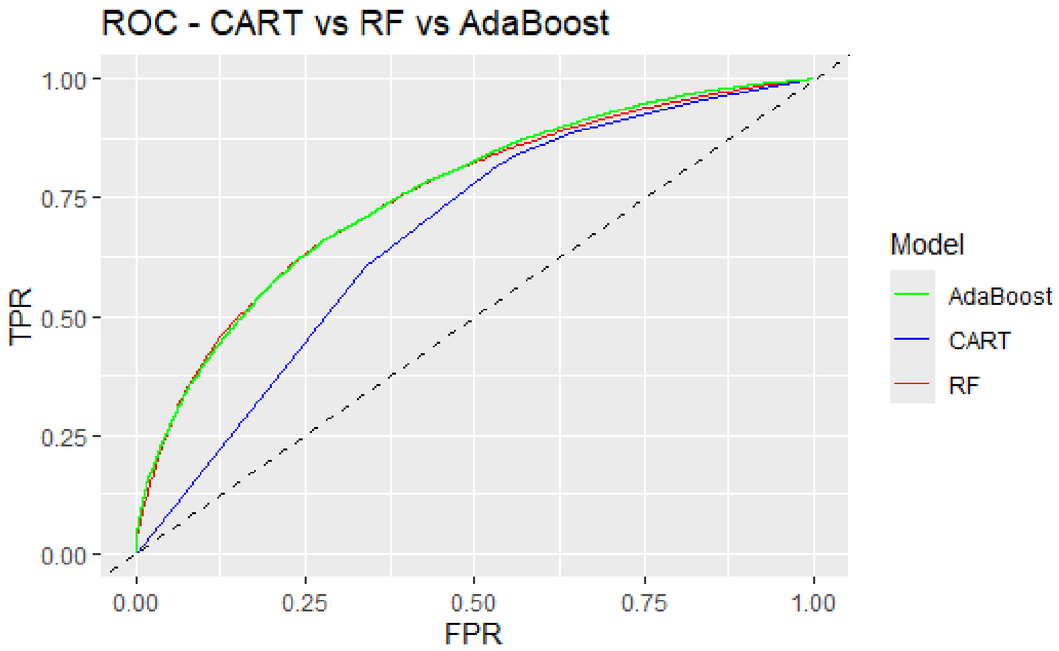
# PREDICTION

*Models trained on available data to predict churners*



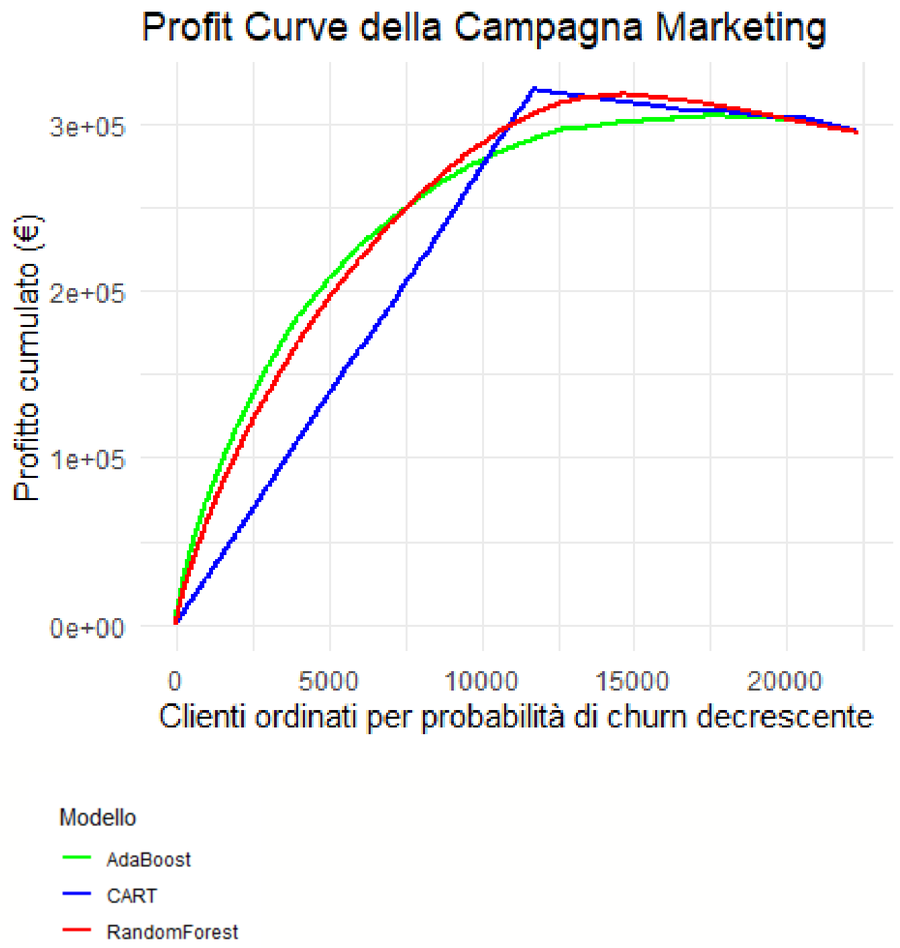
# PREDICTION

| Metric | CART | Random Forest | AdaBoost | Best Model |
|---|---|---|---|---|
| **AUC (approx.)** | 78 | 85 | 86 | **AdaBoost** / RF (tied) |
| **Balanced Accuracy** | 0 , 637 | 0 , 685 | 0, 6461 | **Random Forest** |
| **Sensitivity (Recall of churners)** | 0 , 41 | 0, 66 | 0. 8997 | **Ada Boost** |
| **Specificity** | 0, 86 | 0, 71 | 0, 39 | **CART** |
| **Max Campaign Profit (€)** | ~€325 K | ~€320 K | ~€310 K | **CART** (slightly) |
| **Early-win Profit** | weakest | strong early lift | strongest early lift | **AdaBoost** |




*The models were trained on the complete dataset, assuming that CHURN variable's unbalanceness is still manageble*

# PREDICTION



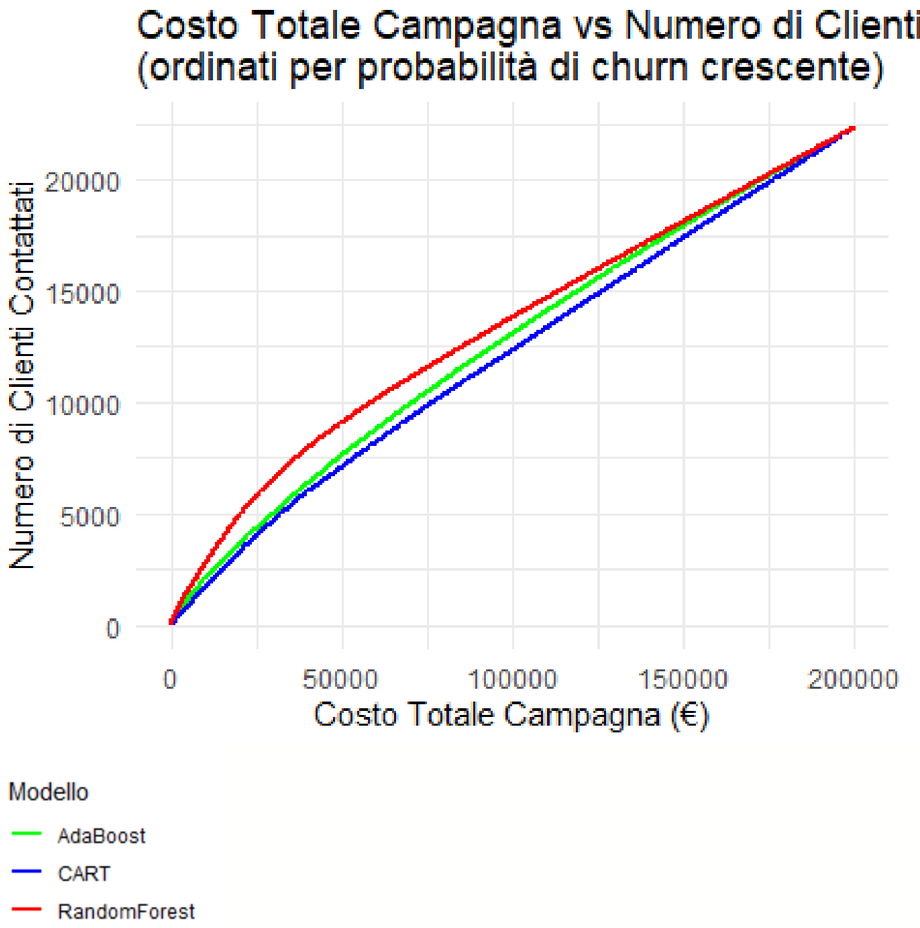

*Each contacted custumer costs 2 euros, but every churner converted brings a profit of 10 euros.*

All three models generate significant cumulative profit when you target customers in descending churn risk. (over random line)

CART slightly edges out on total peak profit (€325 K), but only after contacting 12 000 customers.

AdaBoost and Random Forest achieve faster early payoff (steeper initial profit curves), reaching €200 K with fewer contacts, making them more critical if contact budget is tight.
If the campaing contacts less than 7500 customers, AdaBoost is the best, after that treshold, Random forest is better.

# PREDICTION




When ordering customers by increasing churn probability, the cost curve shows that AdaBoost and Random Forest require fewer contacts (hence lower spend) to neutralize churn for the same number of saved customers. Random forest performs better.

CART is the least efficient early on, and never rises.

# MARKETING TAKEAWAY

| Id | Churner | Prediction | Customer value |
|---|---|---|---|
| 14478 | 1 | 1 | 163, 35 |
| 10554 | 1 | 1 | 153, 85 |
| 5257 | 1 | 1 | 150, 85 |
| 15447 | 1 | 1 | 150, 60 |
| 19670 | 1 | 1 | 147, 06 |
| 15149 | 1 | 1 | 138, 35 |
| 19632 | 1 | 1 | 137, 85 |
| 11025 | 1 | 1 | 136, 35 |
| 2905 | 1 | 1 | 130, 60 |
| 20340 | 1 | 1 | 128, 50 |

*BEST Model: Random Forest*
Offers the best trade-off of AUC, balanced accuracy, and churn **recall**, ensuring to catch the majority of at-risk customers without overspending on safe ones.

*Tactical Use of AdaBoost:*
If the campaign budget is extremely constrained and the goal is to maximize **early gains**, AdaBoost's strong early-lift makes it a solid runner-up.

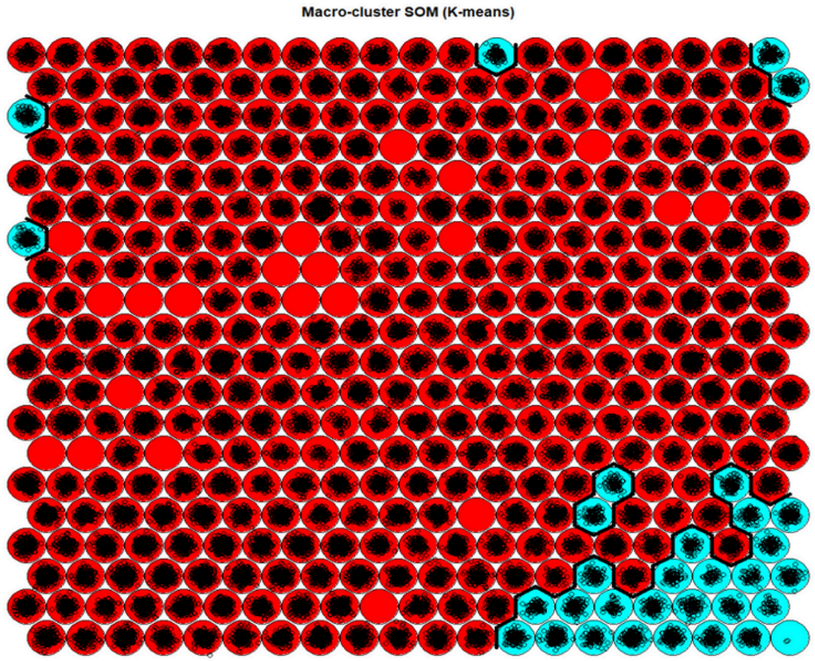
*CART for Peak Profit:*
If there is a large outreach budget and can afford to contact up to ~12 000 highest-risk customers, CART delivers the very **highest total profit**

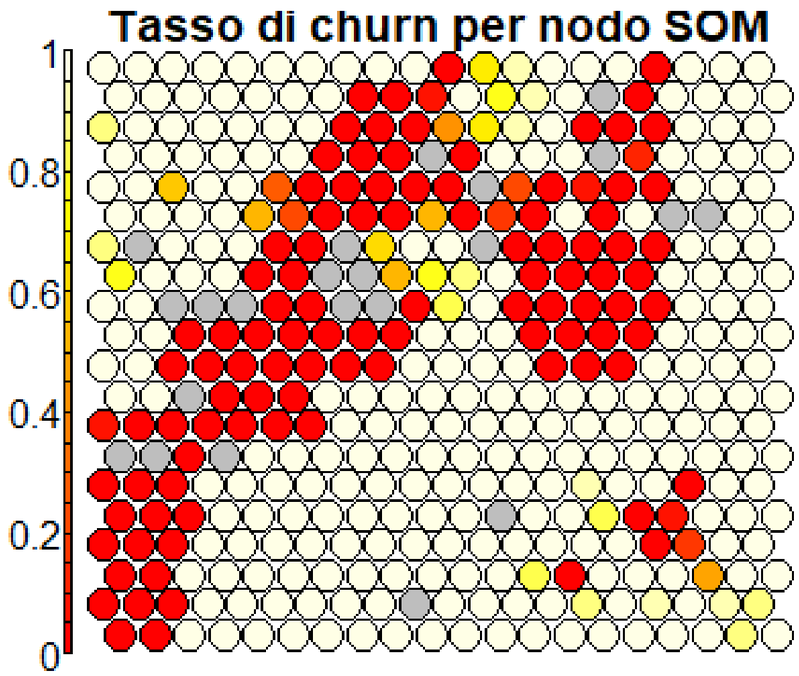
# CLUSTERING

*Does the data have some underlying structure?*
*Is it possible to find some natural customer groups?*

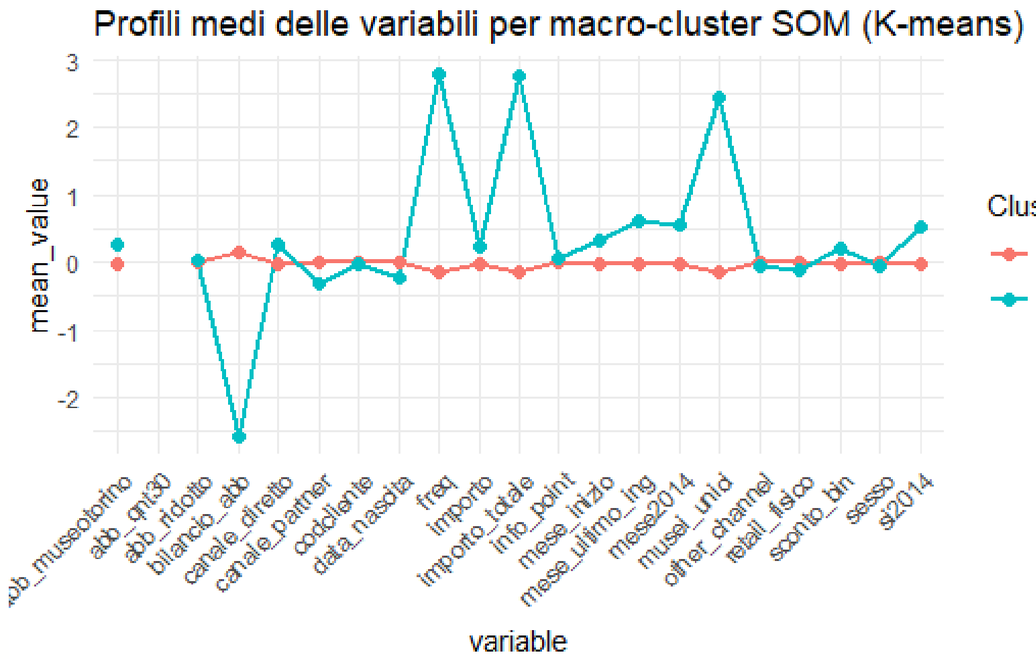
# CLUSTERING

*SOM of size 20x20 with a hexagonal topology and a bubble neighbourhood function.*
*Distance measure(s) used: euclidean. Training data included: 74433 objects. Mean distance to the closest unit in the map: 1.696.*



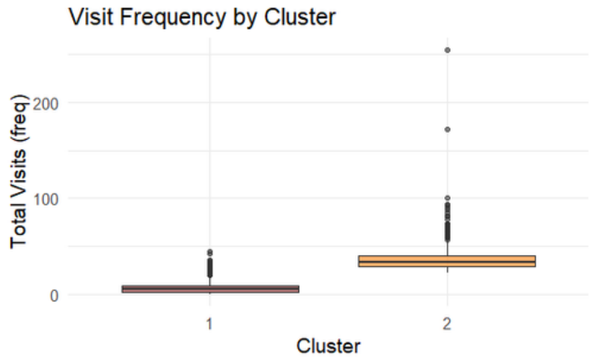
# CLUSTERING

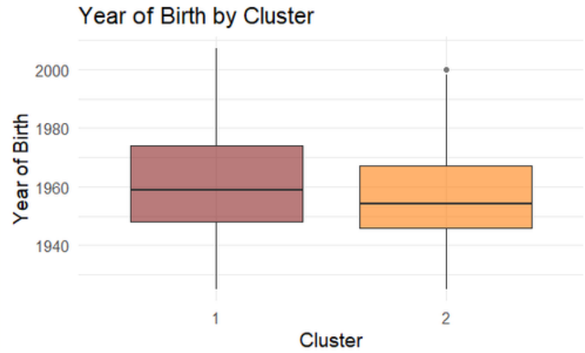


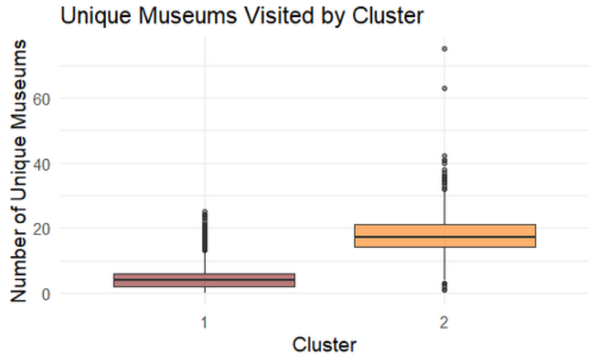

The algorithm identified 2 different clusters, and the main drivers of these differencies were: "bilancio abbonato" (lower values means that the user saved a lot with their card), "freq", the frequencies, total import and number of total museums visited

# CLUSTERING

The two clusters present relevant differences in term of churn rate, identifying the possible drivers of churn.
Other variables present significant differences.

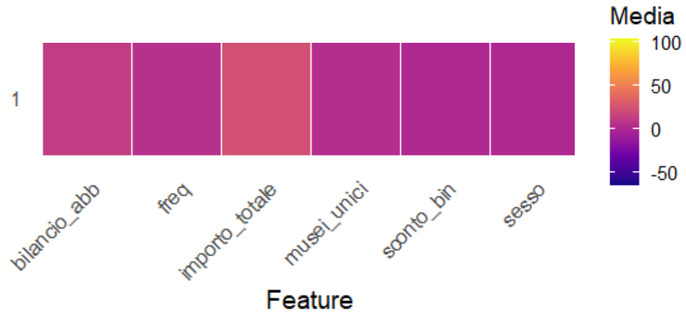



# CLUSTER 1



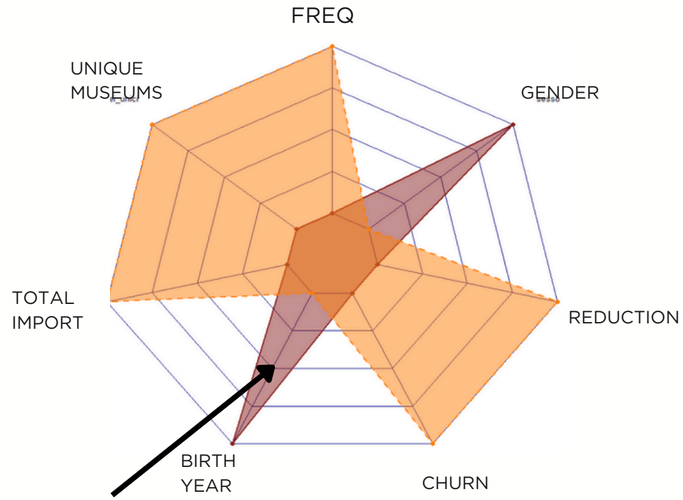



**Size and Churn**
- n = 71`553 customers
- Churn rate = 31.8 % (highest abandonment risk)

**Demographics & Acquisition**
- Birth year 1964: slightly younger than average
- Subscription start: early in the season
- Balanced gender mix (male/female ≈ average)

**Purchase & Discount**
- Price reduction uptake: in line with overall sample
- Card types (“Museo Torino”, “Ridotto”, “Quantitativo 30”): no strong skew
- Sales channels evenly distributed (online, museum, newsstands)

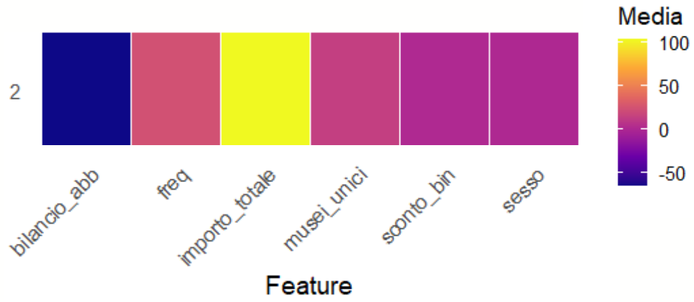
**Usage & Value**
- Visit frequency: below average
- Unique museums visited: below average
- Total paid: slightly below average
- Balance (subscription cost minus pay-per-entry cost): slightly positive → low realized savings




*THIS CLUSTER IS FILLED WITH NON- FIDELIZED CUSTOMERS, who*

# CLUSTER 2



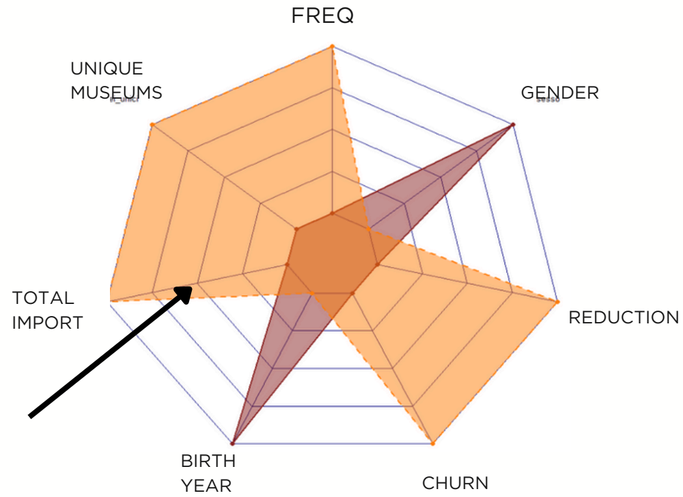



**Size & Churn**
- n = 2`880 customers
- Churn rate = 5.9 % (very loyal)

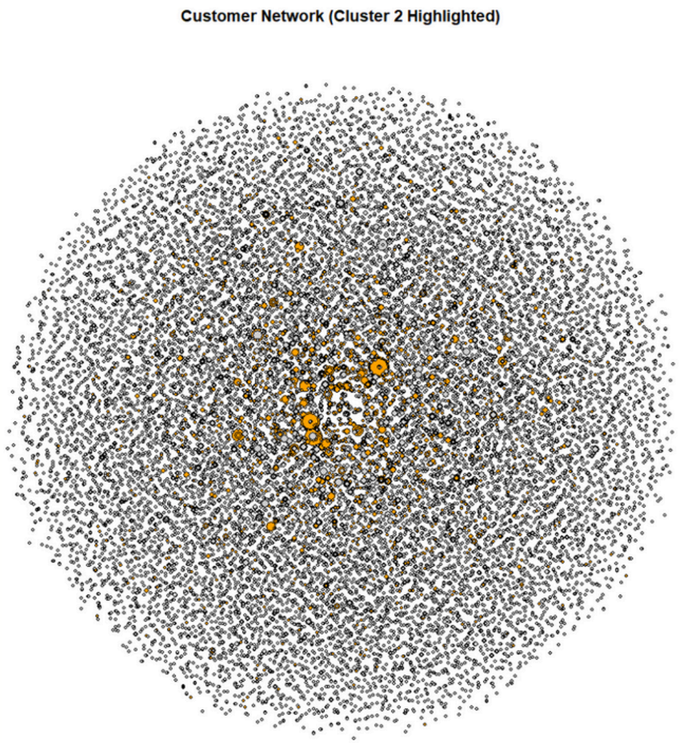
**Demographics & Acquisition**
- Birth year ~1960 (z≈–0.21): older than average
- Tend to subscribe later in the season (z≈+0.31)
- Slightly more male (z≈–0.05)

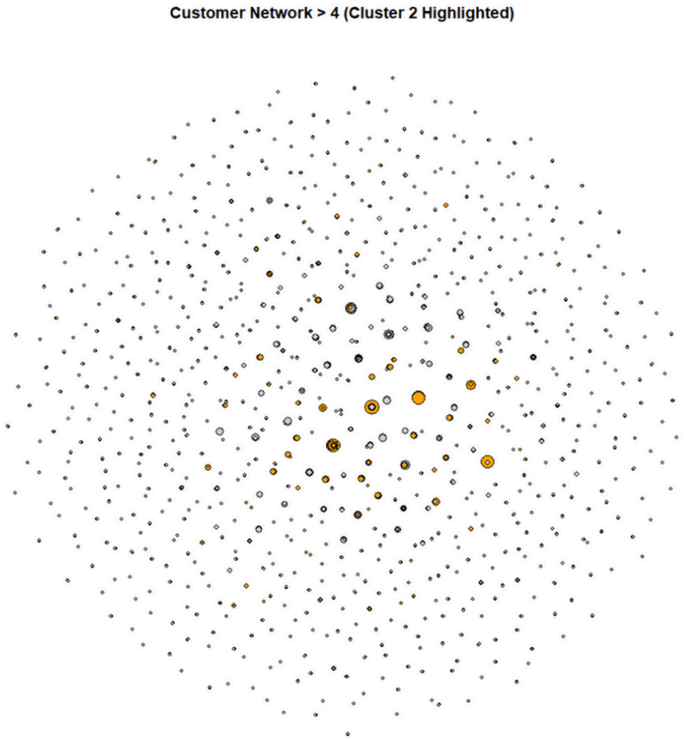
**Purchase & Discount**
- High uptake of “Museo Torino” (z≈+0.27) and “Quantitativo 30” (z≈+0.03) cards
- Above-average discount use (z≈+0.21)
- Predominantly online/direct purchases (z≈+0.28); low partner/retail channel use

**Usage & Value**
- Visit frequency: extremely high (z≈+2.78)
- Unique museums visited: extremely high (z≈+2.44)
- Total paid: extremely high (z≈+2.75)
- Balance: strongly negative (z≈–2.57) → they maximize per-visit savings

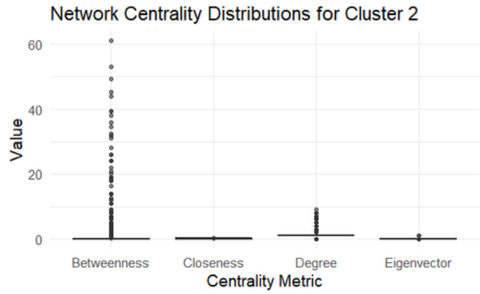


# CLUSTERING




On average each Cluster 2 customer shares ⩾3 visits with just 1 other person. Half of them have only a single such connection.



# MARKETING IMPLICATIONS

**Reactivation Campaigns for Cluster 1:**

- Automated email/SMS reminders before card expiry
- Limited-time reactivation vouchers (e.g. €5 off renewal)

**Loyalty & Upsell for Cluster 2:**

- Premium-tier passes, backstage tours, early-access events
- “Bring-a-Friend” referral incentives using network centrality
- Predictive Modeling & Churn Scoring

- Reward referrals with exclusive perks
- Personalization & Dynamic Offers
- A/B test voucher sizes and timing per segment
- Use SOM-based profiles to tailor communications (e.g. “Your Top 5 Museums” digest)
- Implement real-time geofenced push notifications for “nearby museum” alerts

*Thank you for your attention*