Usecase Delaware: Retail

De Vroey, Bram
Mendoza, Daniel

About this project

- Customer data for fashion retail
- Customers are leaving, why?
- Can we identify which customers are leaving?

Who are the customers?

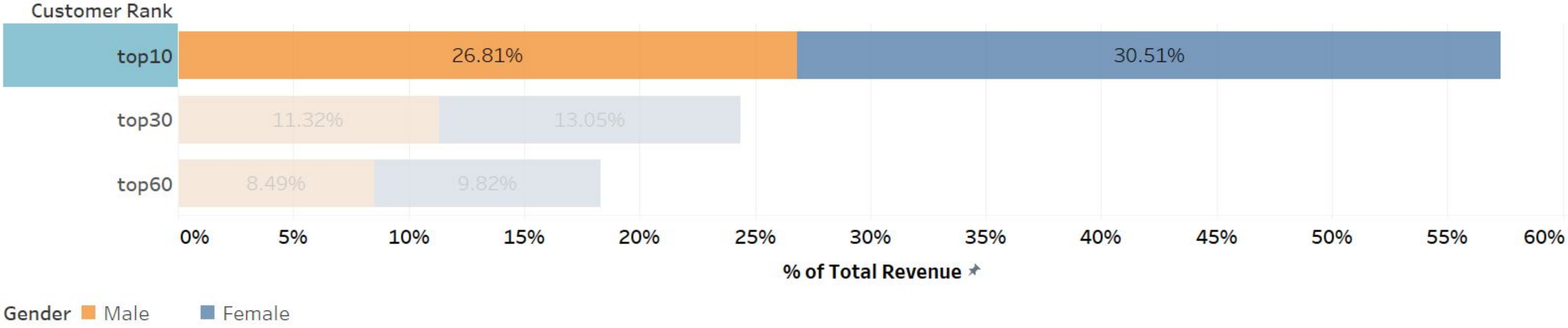
Who are the customers?

- We split the customers in 3 ranks by total revenue
    - 10% highest paying customers
    - 30% middle segment
    - 60% lowest paying

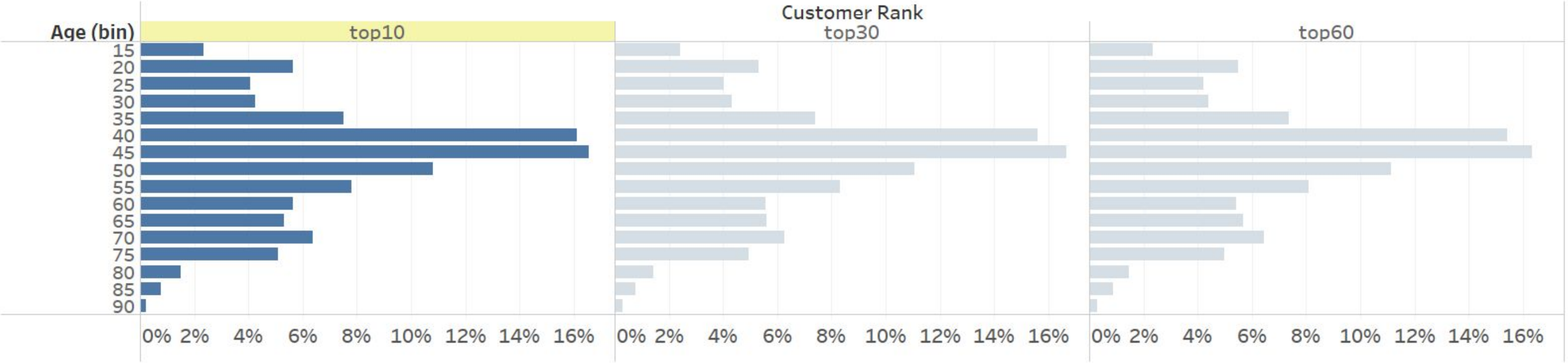
- We only consider registered customers
- 212530 unique customers *(after cleaning)*
- Mostly from Belgium

Some insights about the ranks

The **Top 10** is responsible for **more than 50%** of the total revenue

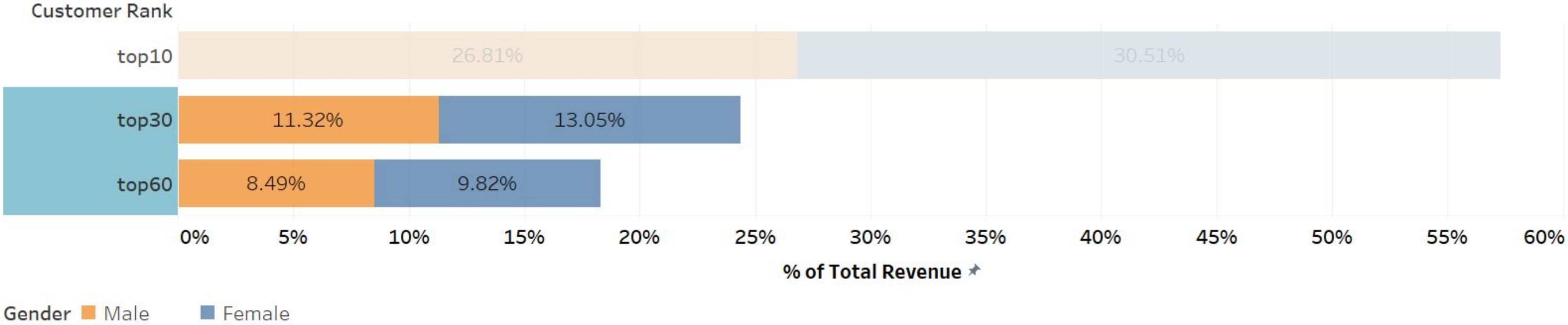


All **3 Ranks** show a similar **Age Distribution.**

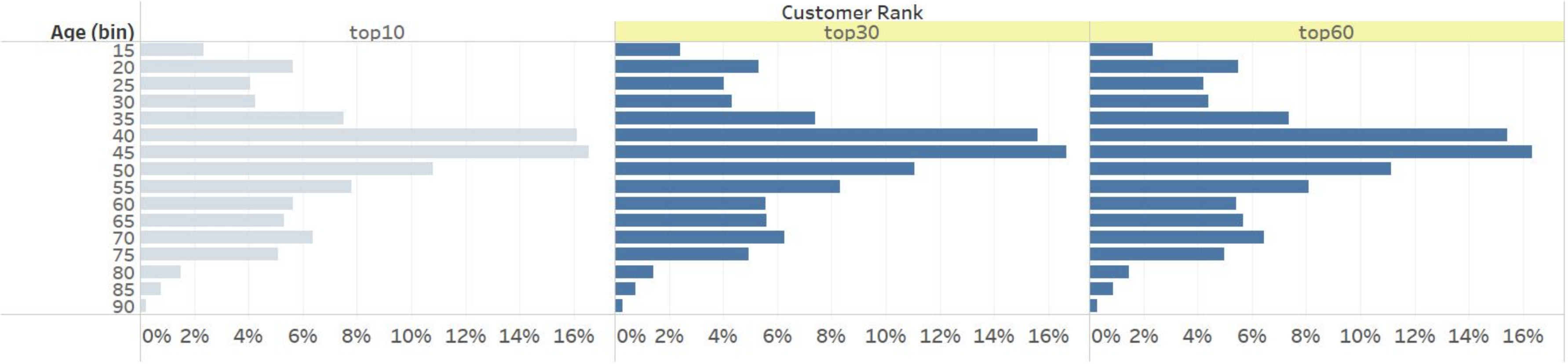


Some insights about the ranks

The **Top 10** is responsible for **more than 50%** of the total revenue



All **3 Ranks** show a similar **Age Distribution.**



Machine learning

Identifying customer clusters

- **Goal**: Identify similar customers based on the data

- Which clusters are likely to churn?
- Target marketing solutions for these customers

Predicting customer churn

- **Goal**: predict whether a customer will **churn** soon, based on easily obtainable features

- **Train** on early data
- **Predict** churn for most recent new customers *(last 6 months)*

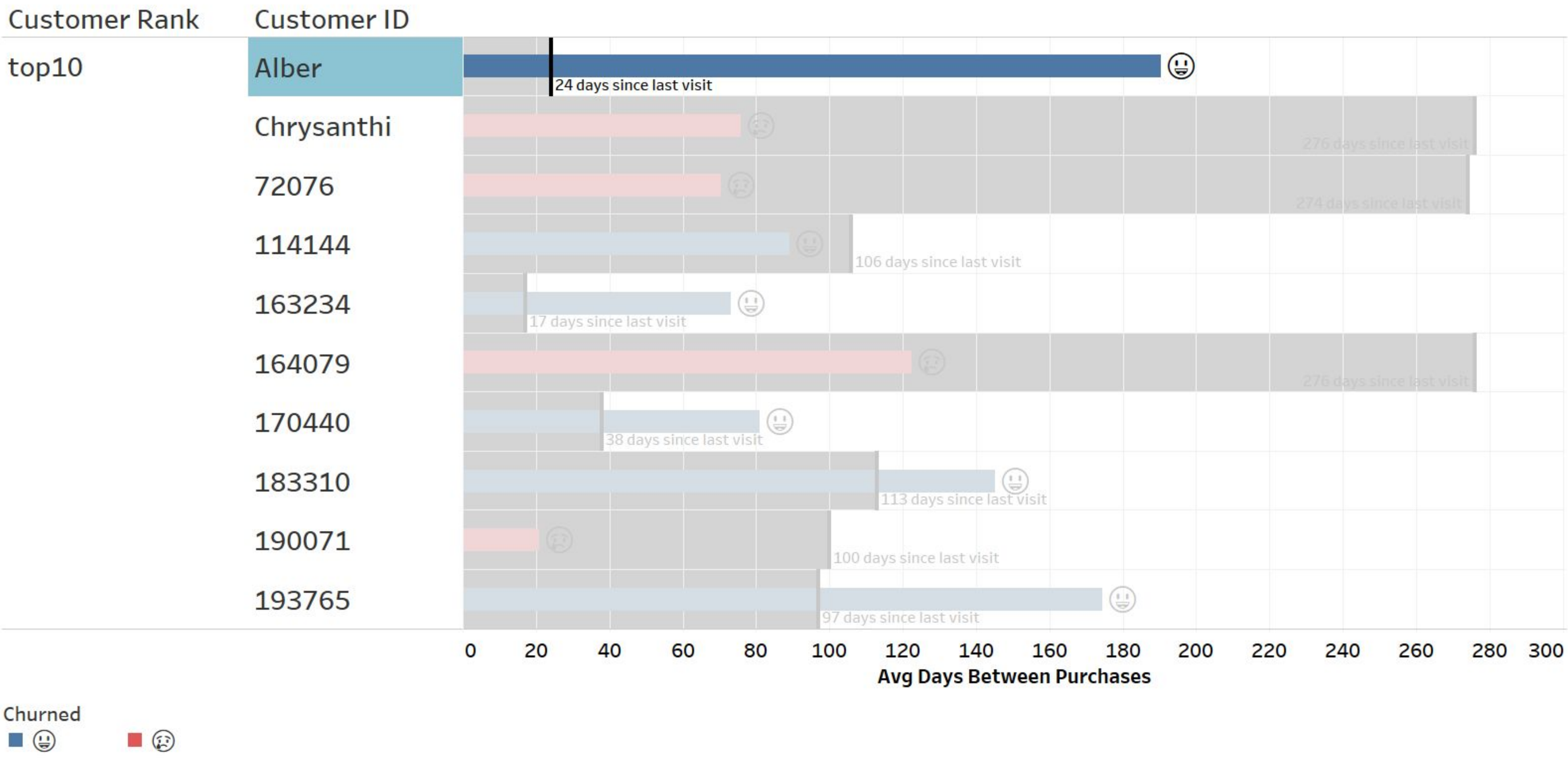
How to label churned?

- How to know if a customer won’t come back?
- We look at their history:
  - Median time between purchases
  - How far behind are they at the end of the dataset?
  - Set some tolerance with an arbitrary factor

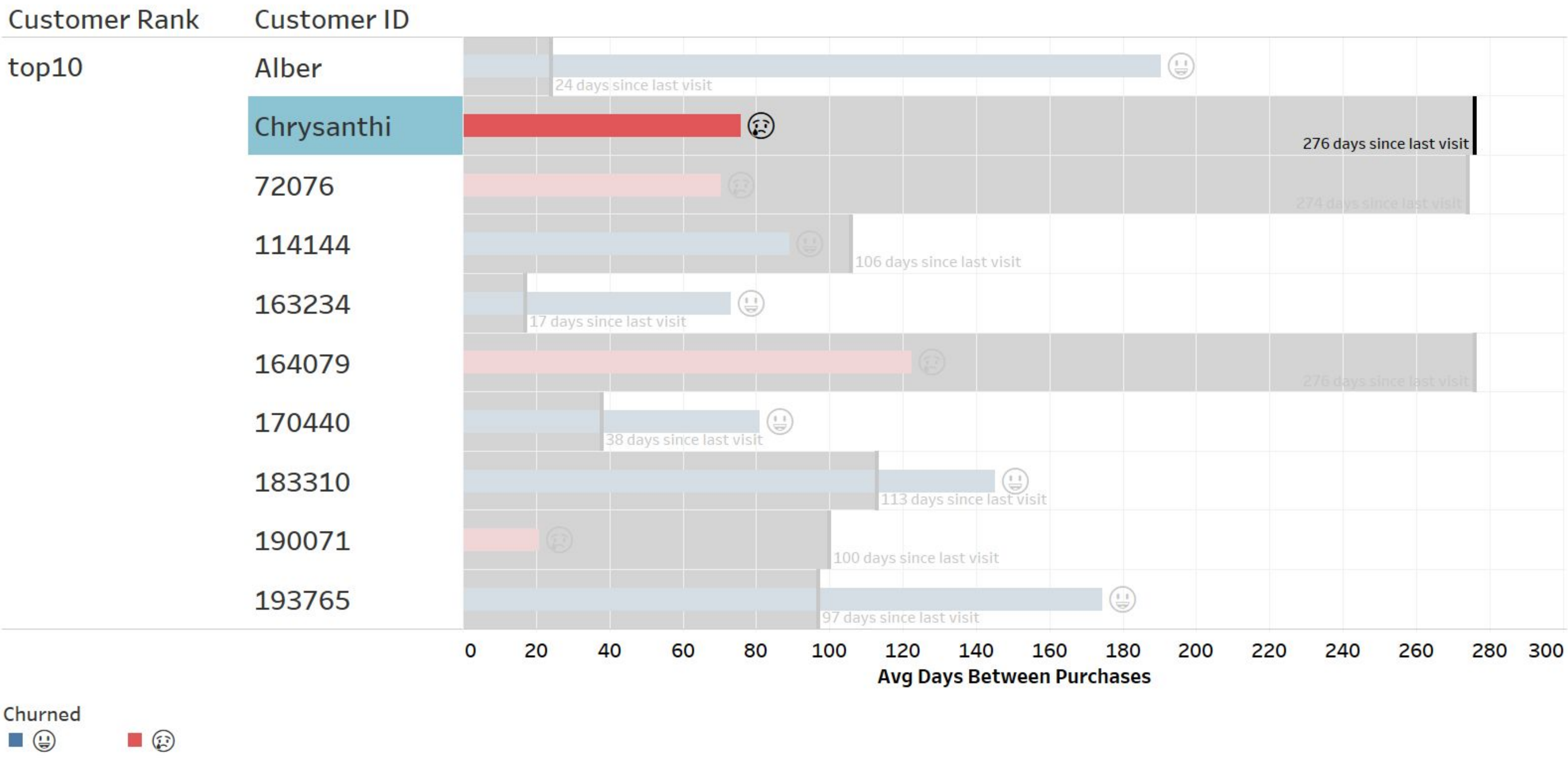
`1.5 × median days between purchases < days since last`
`⇒ churned`

- Exception for recent (new) customers

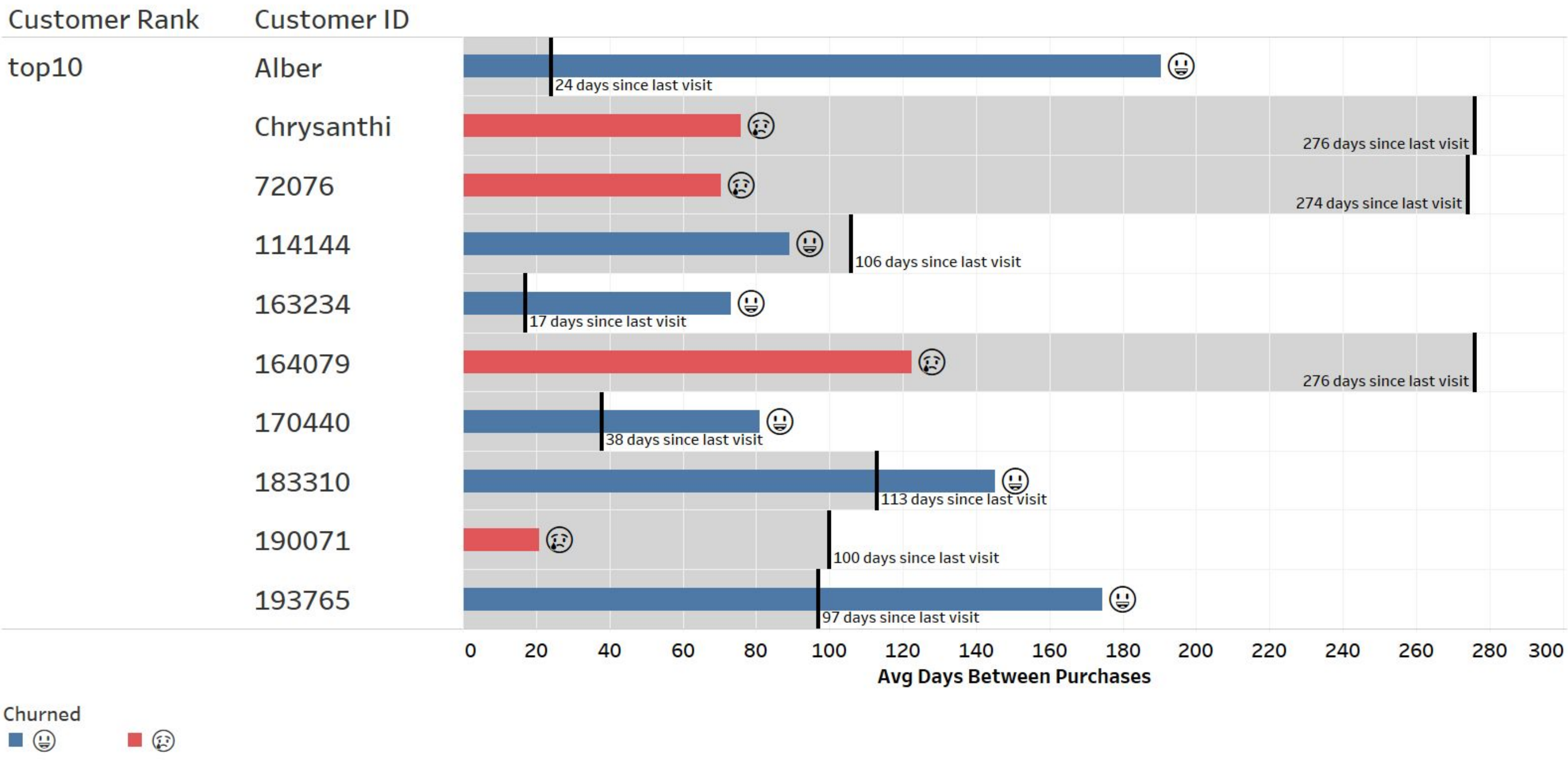
the **Top 10** of the **Top 10**

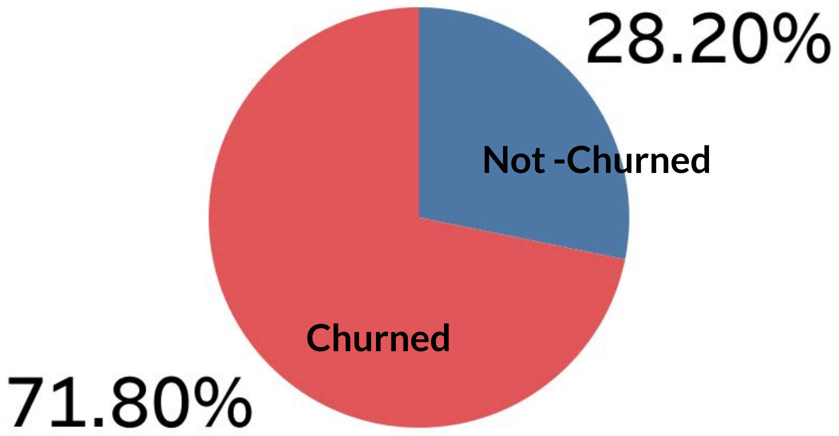


the **Top 10** of the **Top 10**



the **Top 10** of the **Top 10**




28.20%
Not -Churned
Churned
71.80%

Technical details

- Train with customers that are at least 6 months
- Rebalance the data with SMOTE upsampling
- Random Forest Classification
- 12 features used like gender, age,…
  but also total expense, number of days, amount of discounted purchases…

Results for predicting churn

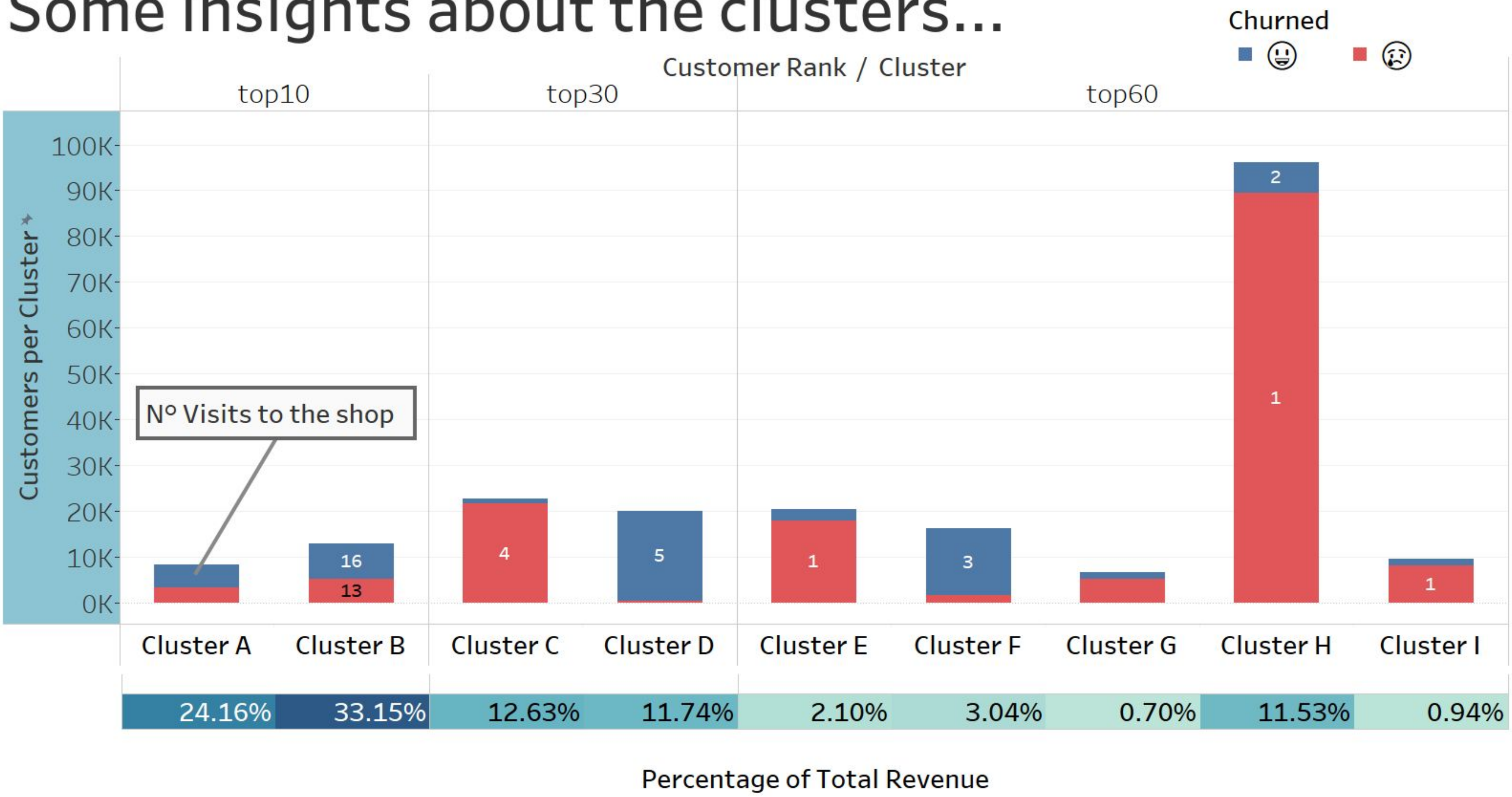
	precision	recall
loyal	86%	96%
churned	96%	85%

- If model predicts churn:
  - 4% chance that it will not churn
  - 15% of churns are missed

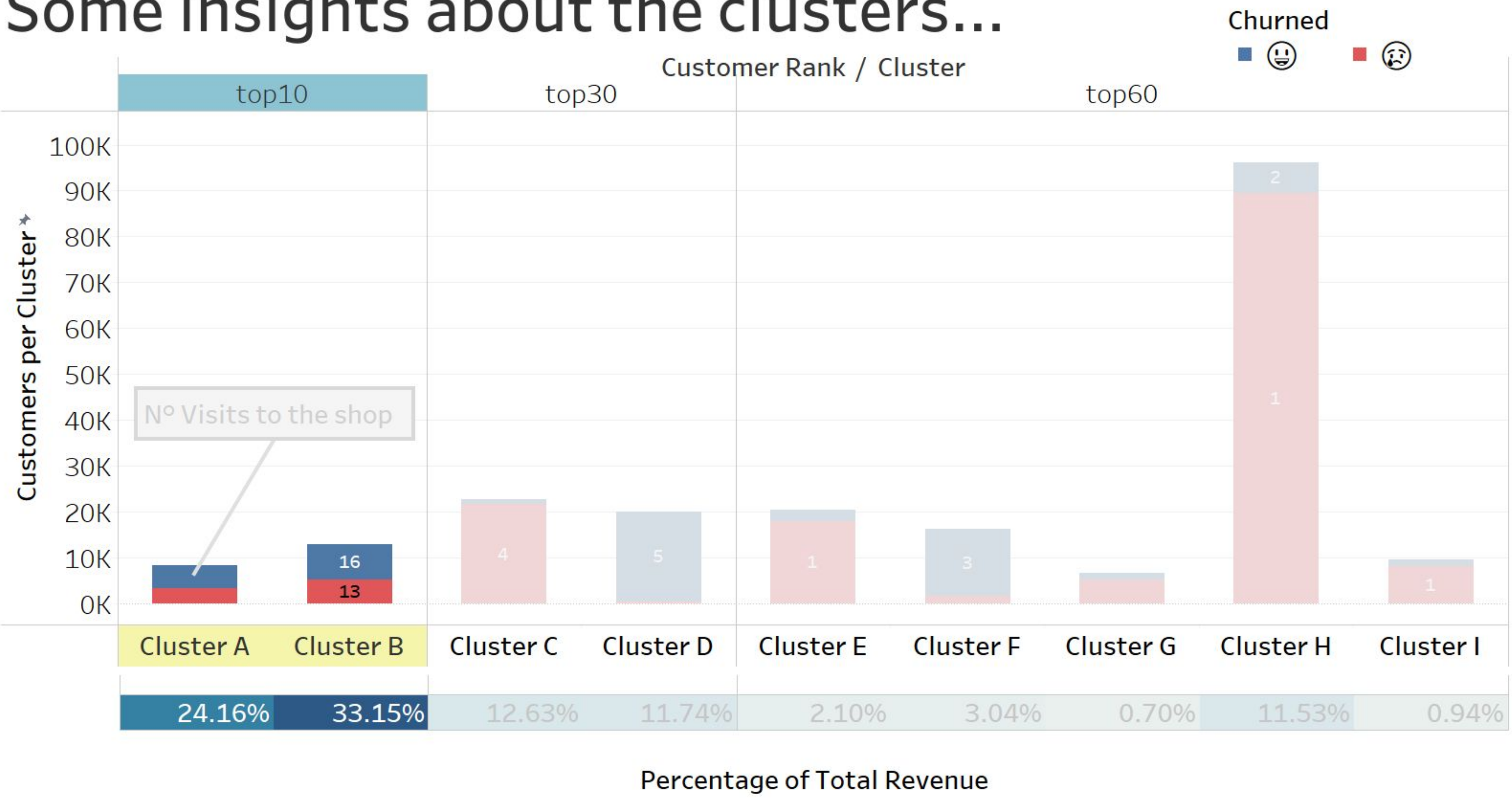
How to “clustering”?

- Which data to use?:
  - Grouped by “Customer ID”
- Explore the right number of cluster and the features.
- Clusters per customer ranks (Top 10/30/60)
- We found 2 - 2 - 5 clusters
- Cluster → K-means algorithm

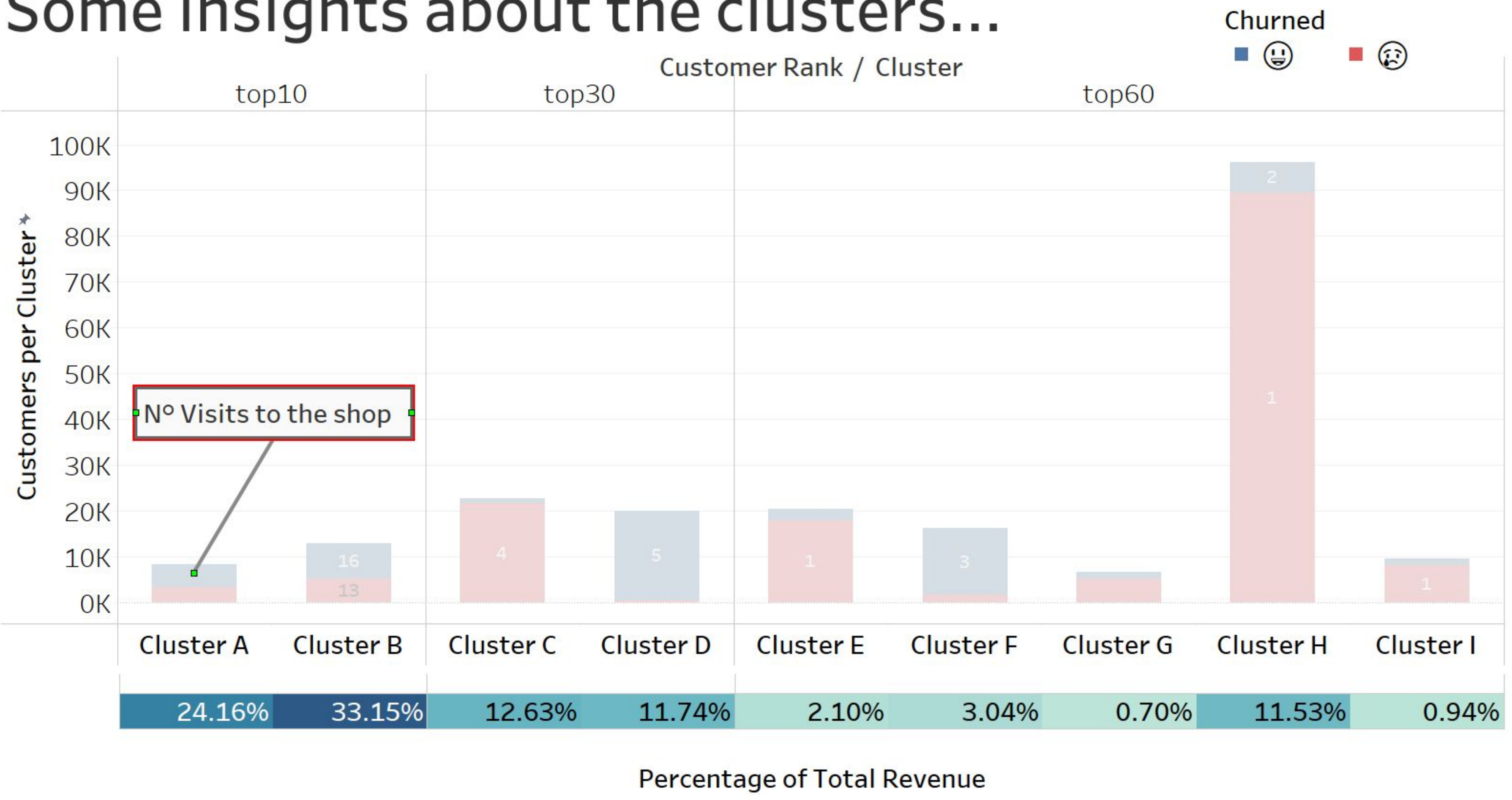
Some insights about the clusters...



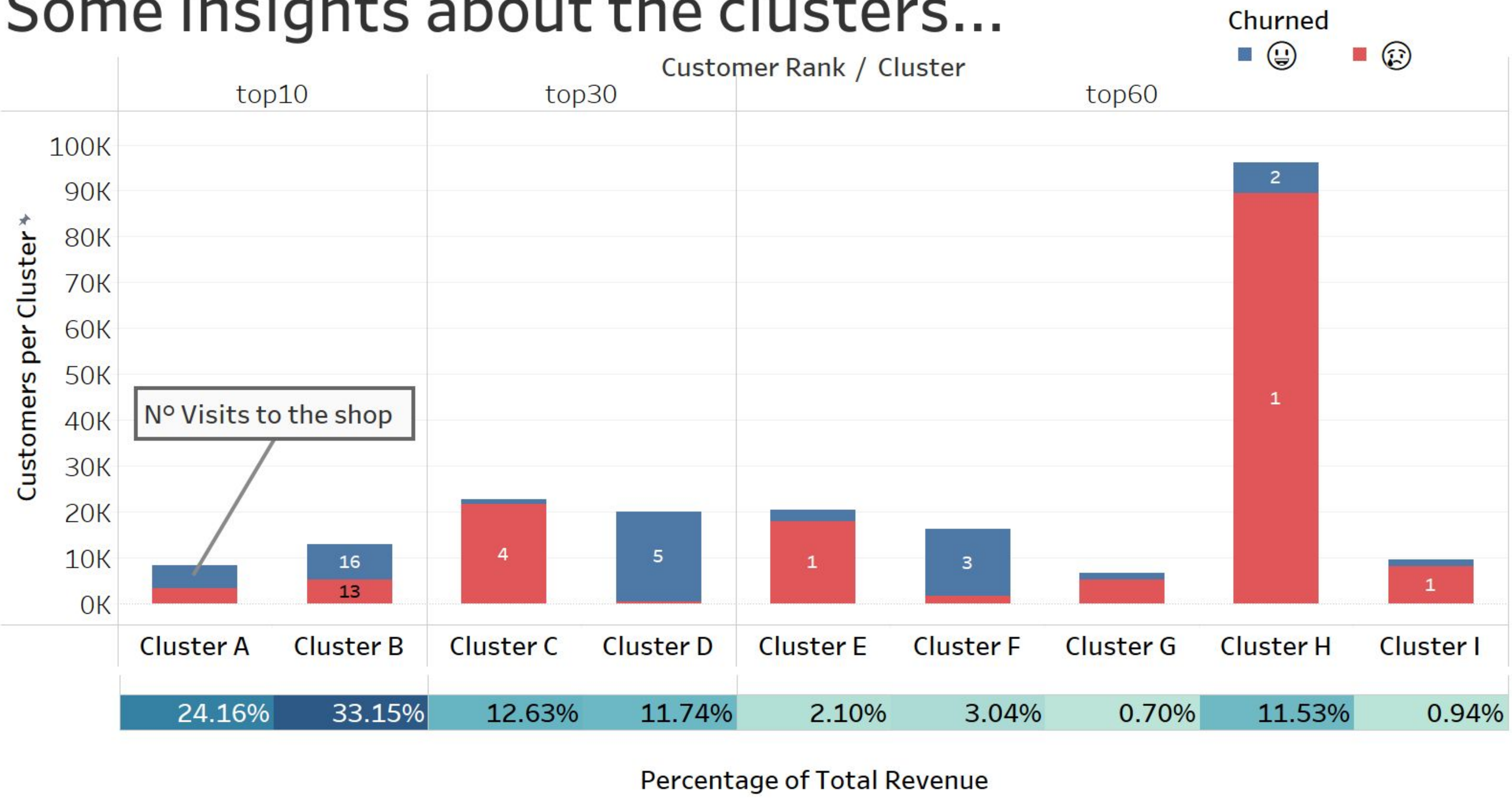
Some insights about the clusters...



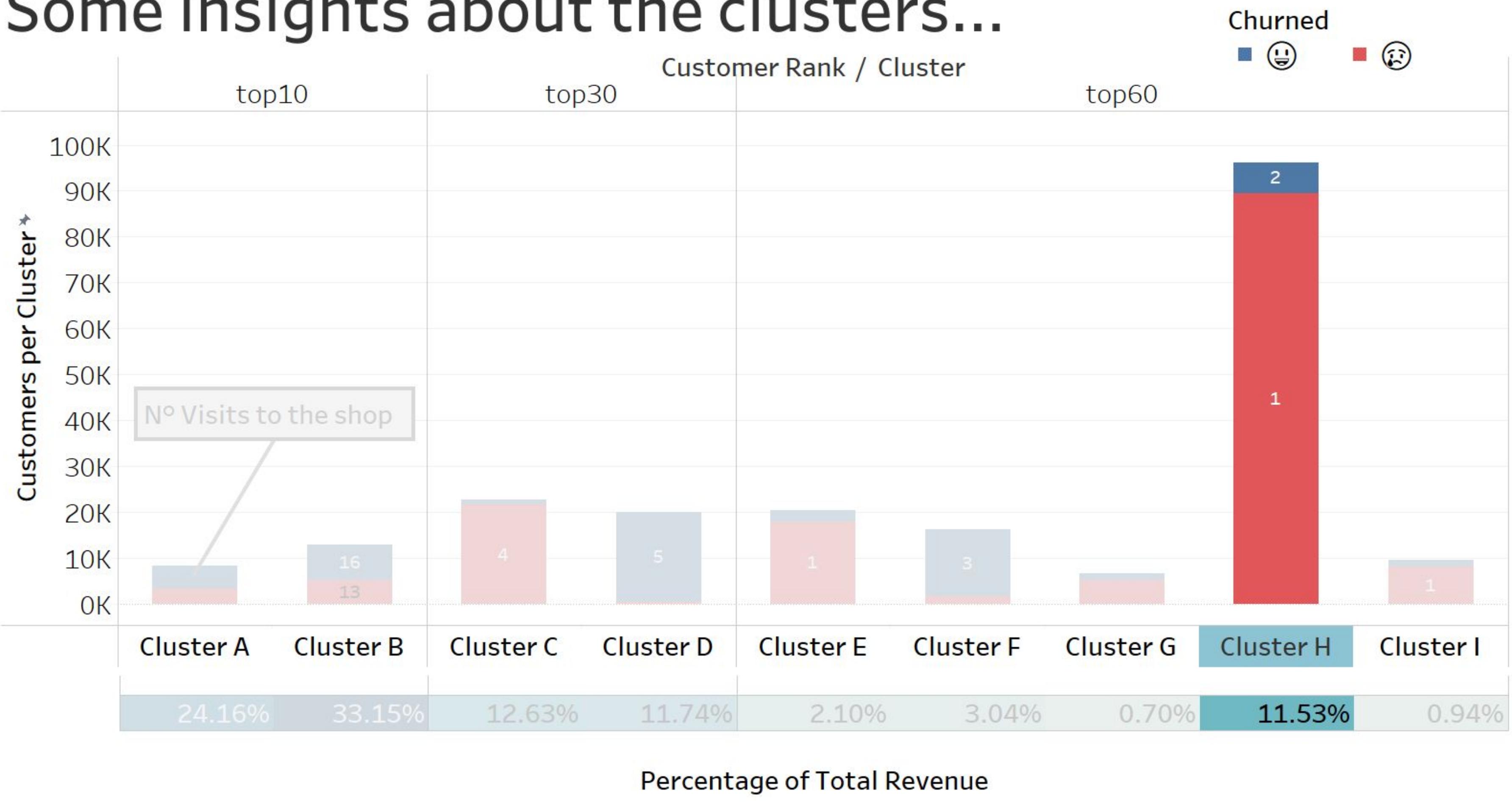
Some insights about the clusters...



Some insights about the clusters...



Some insights about the clusters...



What did we learn from the data?

About Churn prediction

- Quite reliable churn prediction if enough data available
  - Important to have more than 1 transaction for the customer
  - Trust more in “churned” result
- Real validation needed?

About Clustering

If we focus on the **“cool”** customer, who brings the highest amount of revenue **(TOP10):**

Cluster A:

- Main product type:
  - 1° -> KIDS
  - 2° -> WOMENS (winter)
- Avg. Price per Product ≈ 50€

Cluster B:

- Main product type:
  - 1° -> ONLY KIDS
- Avg. Price per Product ≈ 35€

About Clustering

If we focus on the “**regular”** customer” **(TOP30):**

Cluster D:

- Main product type:
  - 1° -> KIDS
  - 2° -> WOMENS (winter)
- Avg. Price per Product ≈ 35€

About Clustering

If we focus on the **“passing by”** customer” **(TOP60):**

Cluster H:

- Main product type:
    - 1° -> KIDS
    - 2° -> WOMENS (summer & winter)
- Avg. Price per Product ≈ 40€

Improvements

How useful can ML be?

- **Classification**
  Depends on the quality of the data
  Time-dependant properties, so need full customer history to make good predictions

- **Clustering**
  Found cluster-dependant results, clear patterns are debatable.

To do better

- Model deployment
- Explore alternative classification and cluster algorithms
- It would be interesting to compare the *churn rate* with other companies of the same sector.
- Try more varied conditions for the *churn rate*
- Customers profiles can be improved with more data
    - Increase the level of trust in the data collected (add zip code, product categories ....)

Conclusions

Conclusions & our advice

- Churn rate is very high
  - Biggest spending customers are relatively loyal
- A lot of one time customers
  - Not necessarily bad (10% of revenue)

Questions?