MACHINE LEARNING I: CLUSTERING PROPOSAL FOR A TRAVEL AGENCY

By Group 3, the Gaussian Gang

MBD S1 SEPT 2023

AGENDA FOR TODAY

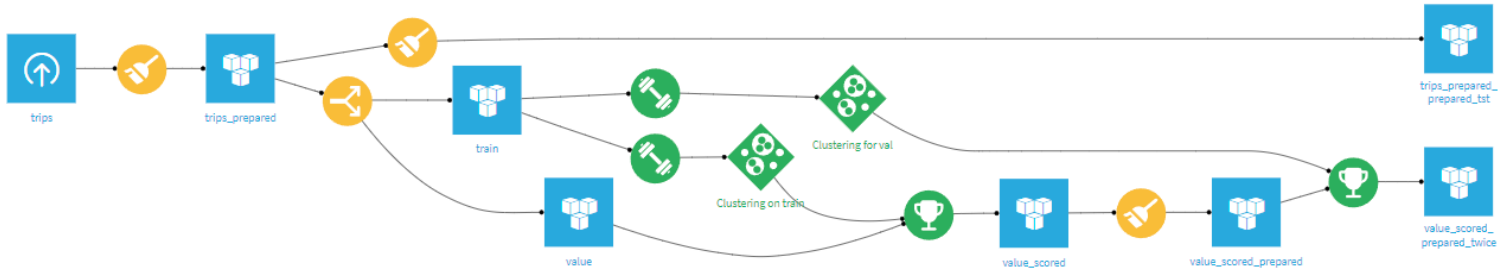
1. Introduction to the assessment
2. Business Analysis
3. Technical Analysis

AT A GLANCE: WE ORGANISED A COMPETITION TO GET THE BEST MODEL...

CESAR | IGNACIO | LEA

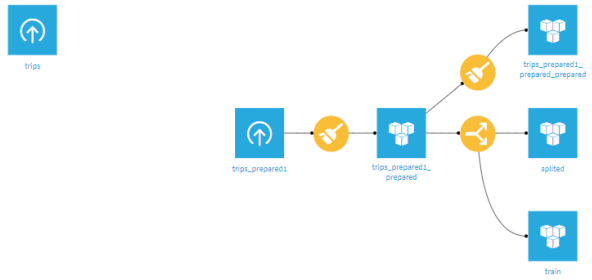
K-Means Clustering

KANGJIE



Agglomerative & Interactive Clustering

MORTEN



K-MEANS WAS SIGNIFICANTLY BETTER

...BUILDING MORE THAN 40 UNIQUE MODELS UNDER BOTH ALGORITHMS OVER 6 SERIES OF DECISIONS.

AFTER INITIAL TRIALS, WE STARTED TO COMPREHEND THE DATASET

We tried removing Age under the assumption of the buying person being not important, **but we were wrong!**
We tried encoding the office, but it was not an important feature!

We removed summer and created high season (jun., jul. and dec.) to see the effects of some months, **but we were wrong!**

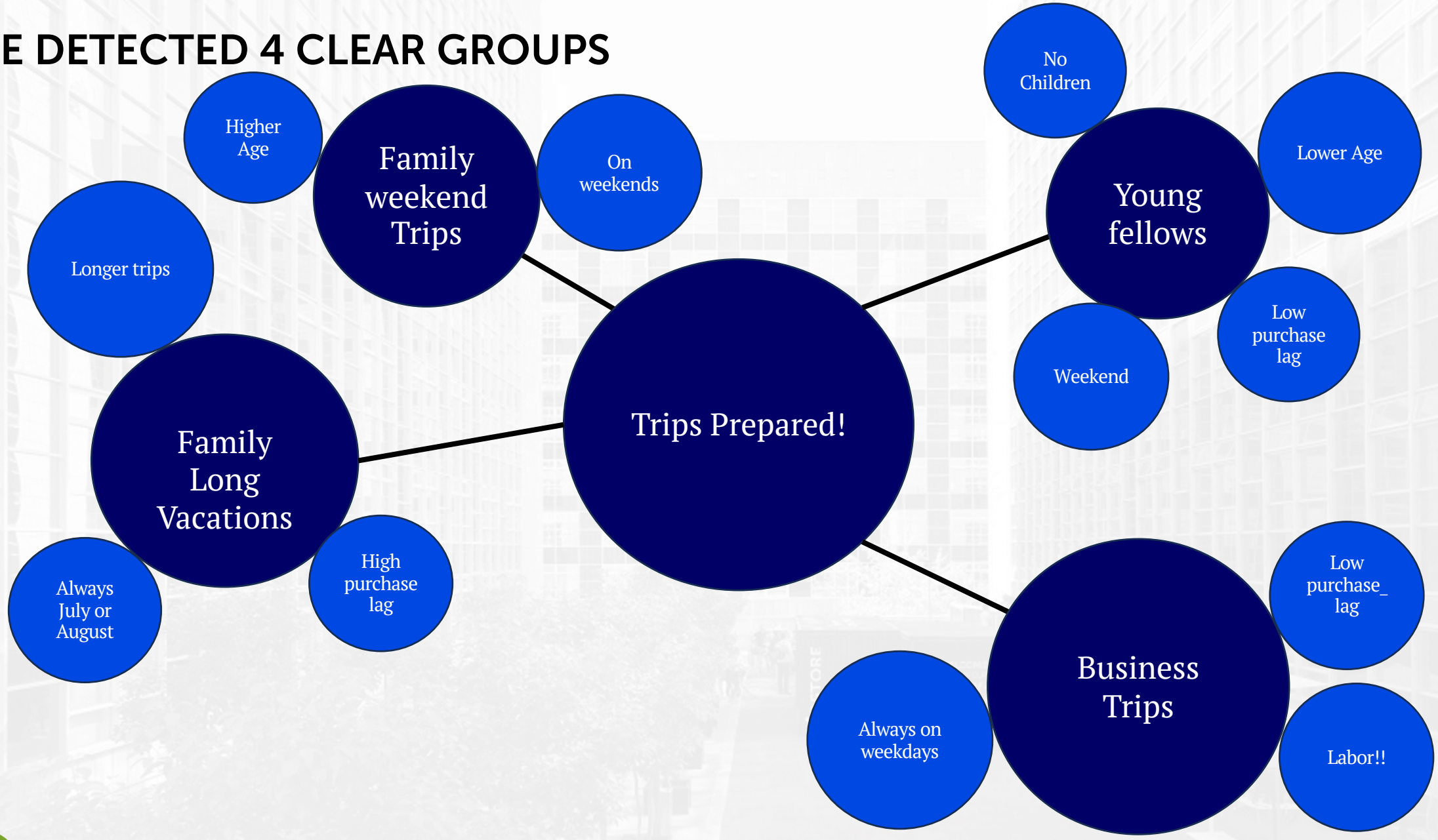
We created purchase_lag that turned out to be very important

We dropped variables based on correlations such as 'weekend' and 'alone_labor'.

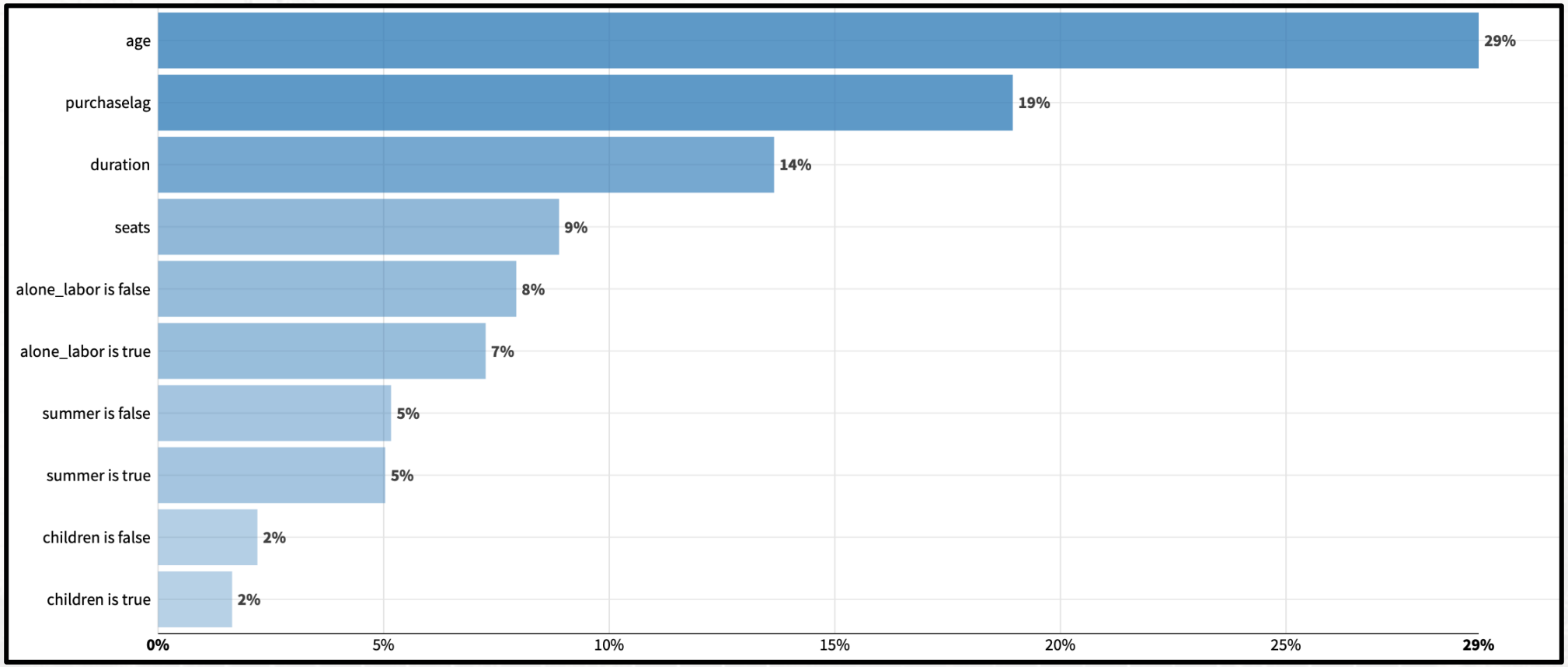
We preferred a silhouette score that was actually lower, but made more sense on the business side
Etc.

BUSINESS ANALYSIS

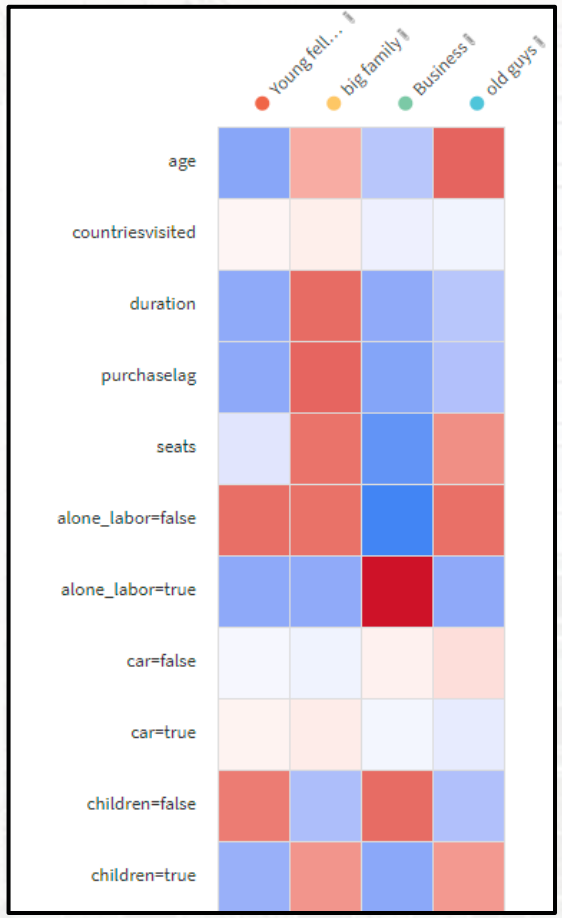
WE DETECTED 4 CLEAR GROUPS

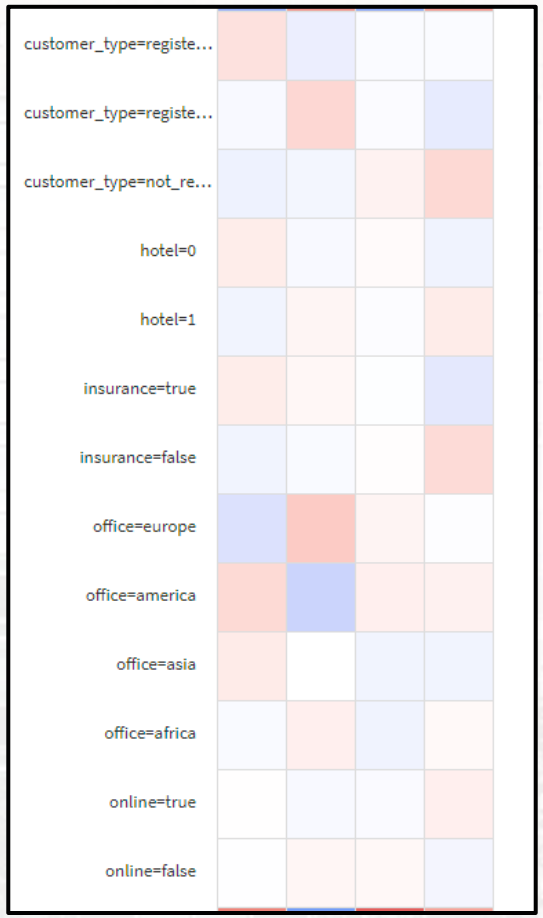


FEATURE IMPORTANCE OF OUR LAST MODEL



WHAT IS IMPORTANT? WHERE CAN WE IMPROVE?





BUSINESS RECOMMENDATIONS

1. Family Long Vacations:

Customized Family Packages: Offer packages that cater to families traveling in July or August, which could include longer trips with child-friendly activities and accommodations.

Early Bird Discounts: Encourage early booking with discounts for those who plan their vacations well in advance (indicated by "High purchase lag").

Loyalty Programs: Introduce a loyalty program that benefits repeat customers, as this group seems to have a pattern (always July or August).

2. Family Weekend Trips:

Relaxation and Comfort: Create travel options focused on relaxation and comfort, targeting older age groups who may prefer a slower-paced vacation.

Accessible Travel Services: Ensure that travel services are accessible, with options for those with limited mobility.

Cultural and Historical Tours: Offer cultural and historical tours that may appeal to an older demographic.

BUSINESS RECOMMENDATIONS

3. Young Fellows:

Adventure and Social Experiences: Design travel experiences that are adventurous and provide opportunities for socializing, as this group seems to be younger and without children.

Weekend Getaways: Since they travel on weekends, provide short, affordable weekend trips to nearby destinations.

Last-minute Deals: Offer last-minute deals to cater to those with "Low purchase lag," indicating spontaneous trip planning.

4. Business Trips:

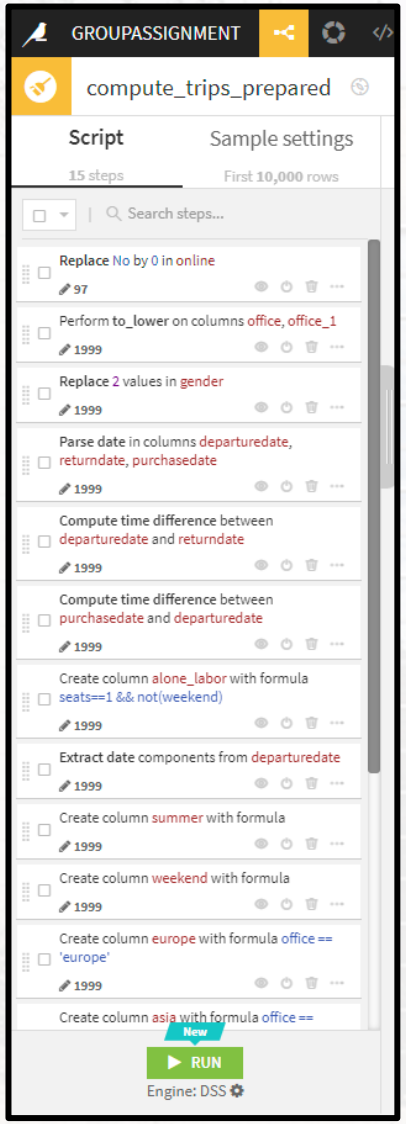
Efficient and Flexible Options: Provide business travelers with efficient travel options that include flexible booking and cancellation policies.

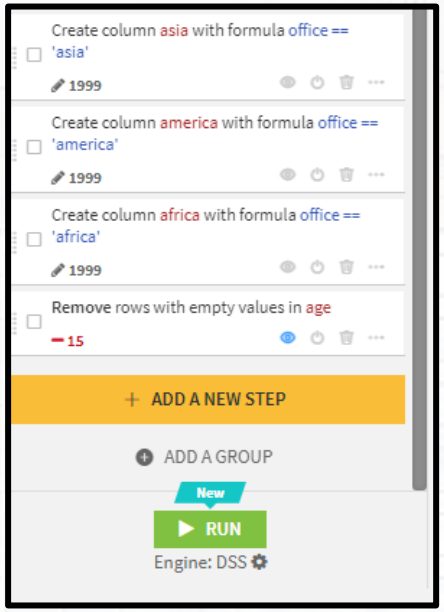
Travel and Accommodation Bundles: Offer bundles that combine travel, accommodation, and car rentals to simplify the booking process.

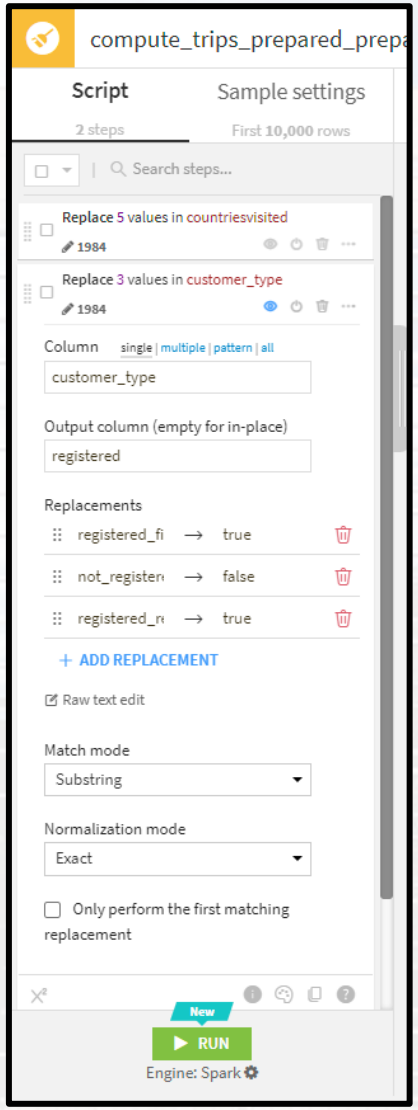
Rewards for Frequent Travelers: Implement a rewards program for frequent travelers to encourage brand loyalty and frequency.

TECHNICAL ANALYSIS

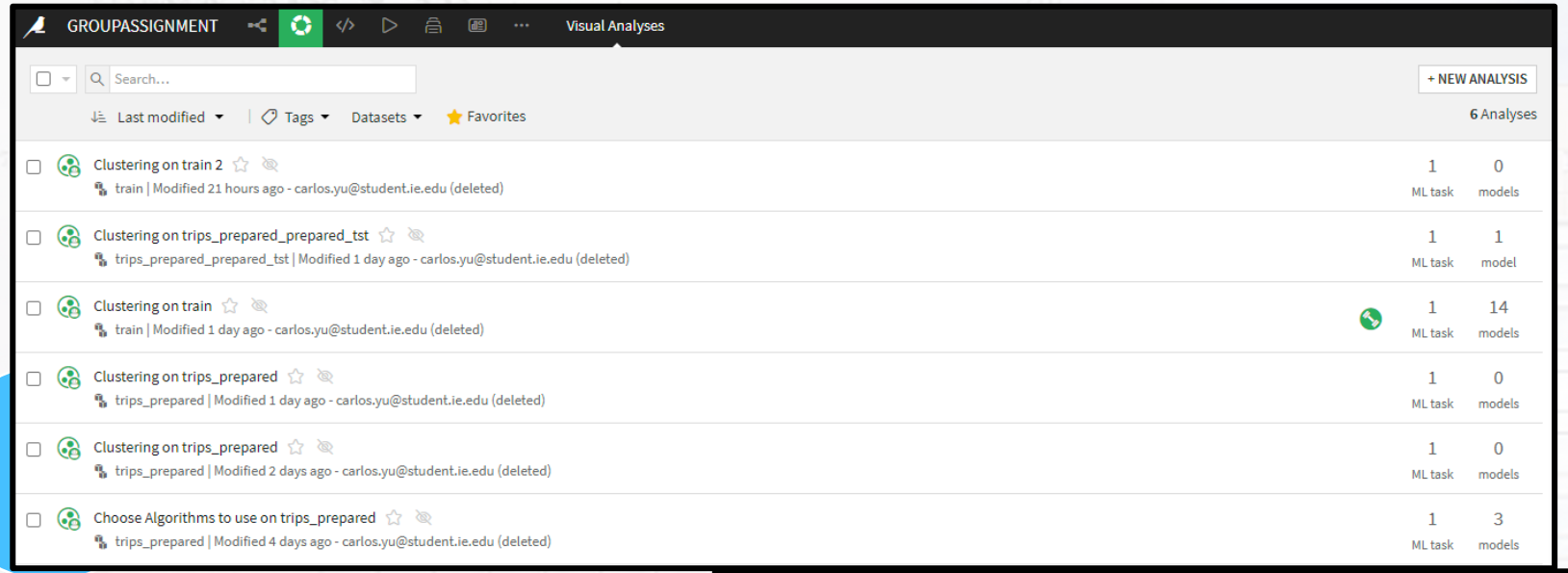
TECHNICAL: FEATURE ENGINEERING

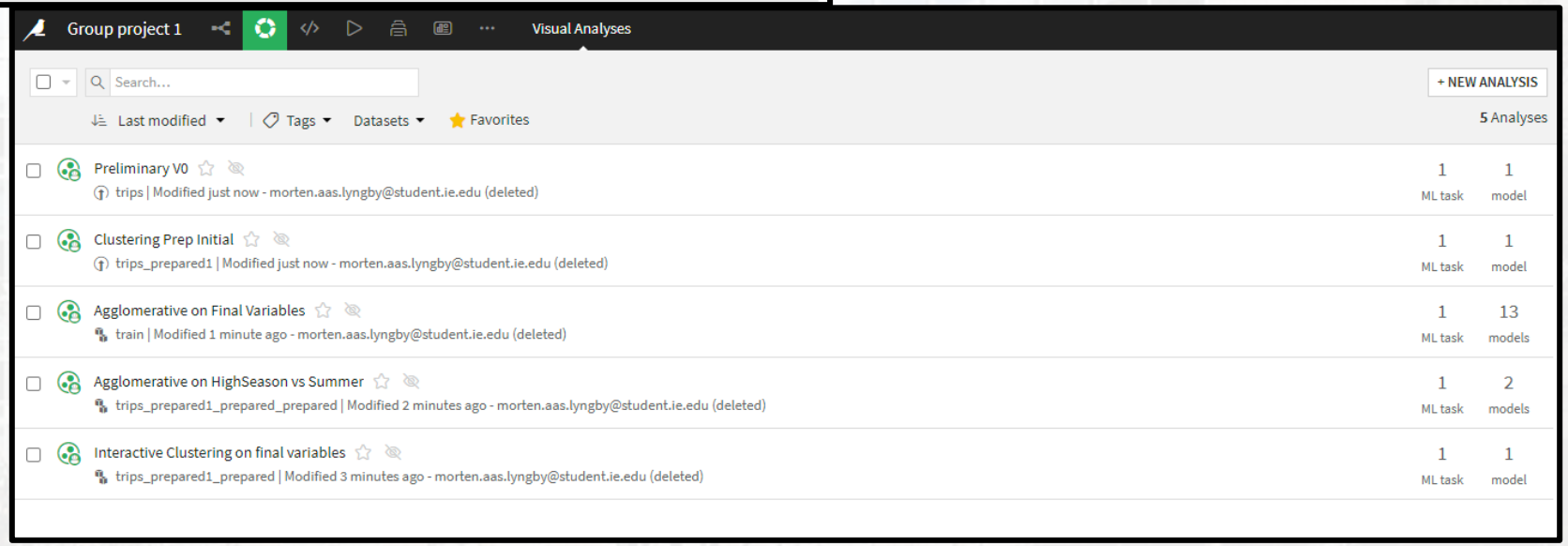






TECHNICAL: WE MIMIC BOTH PROJECTS





TECHNICAL: MODEL CONSTRUCTION I

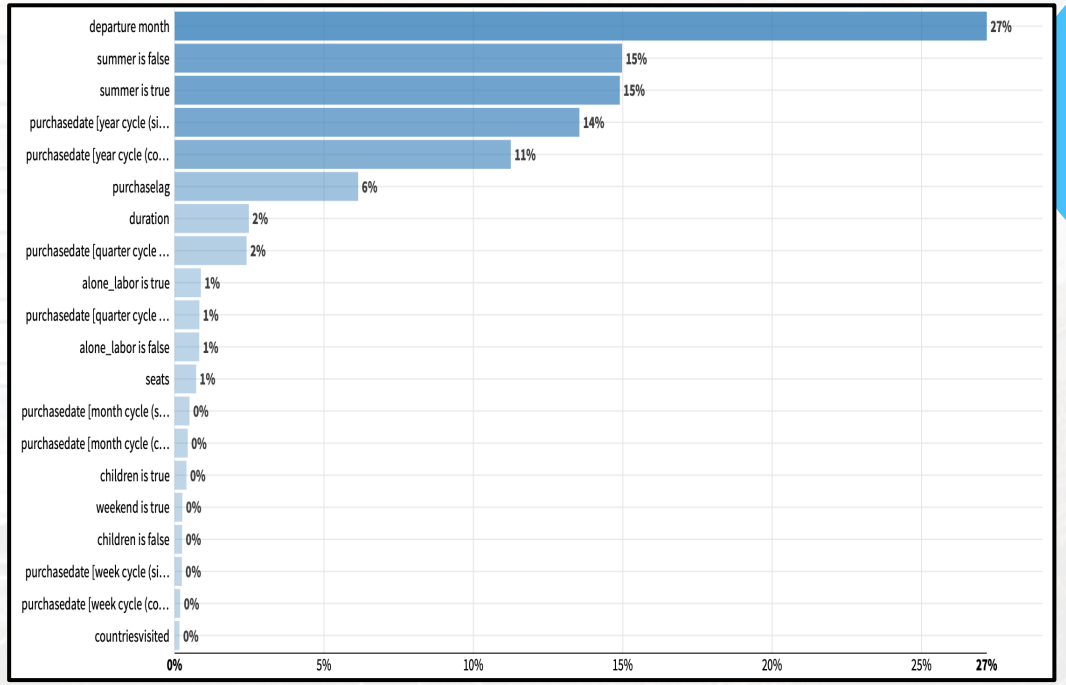
1st clustering heatmap

Features selected
Summer, departure month, duration, weekend, purchase date, purchase lag, children,
Car, hotel, insurance, sets, alone labor, customer_type, office, online, countries visited

Mini-cluster size threshold 15
(First we thought the age only shows the people who purchased but not everyone included in the trip, we thought it was inaccurate and did not take into account, which we changed at the third step.)

K-means & Agglomerative
Departure month -> categorical (from numerical), also we made a column 'high season' with month 7,8,12, but it decreased feature importance

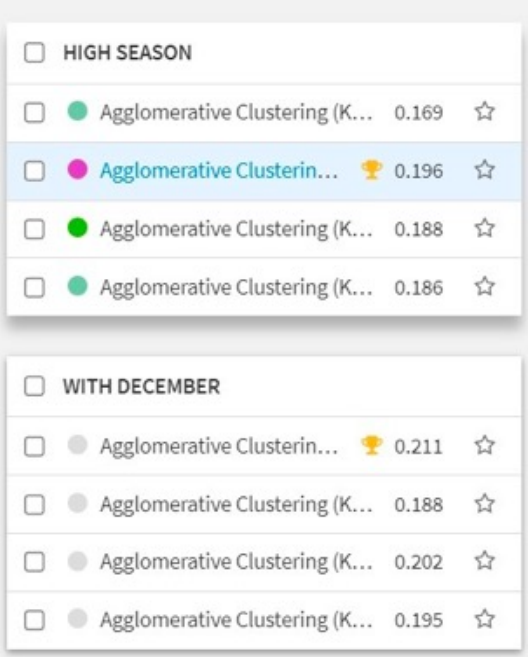
We decided to drop Purchasedate, too many values and we have the departuredate and duration that is adding extra importance to purchase date.

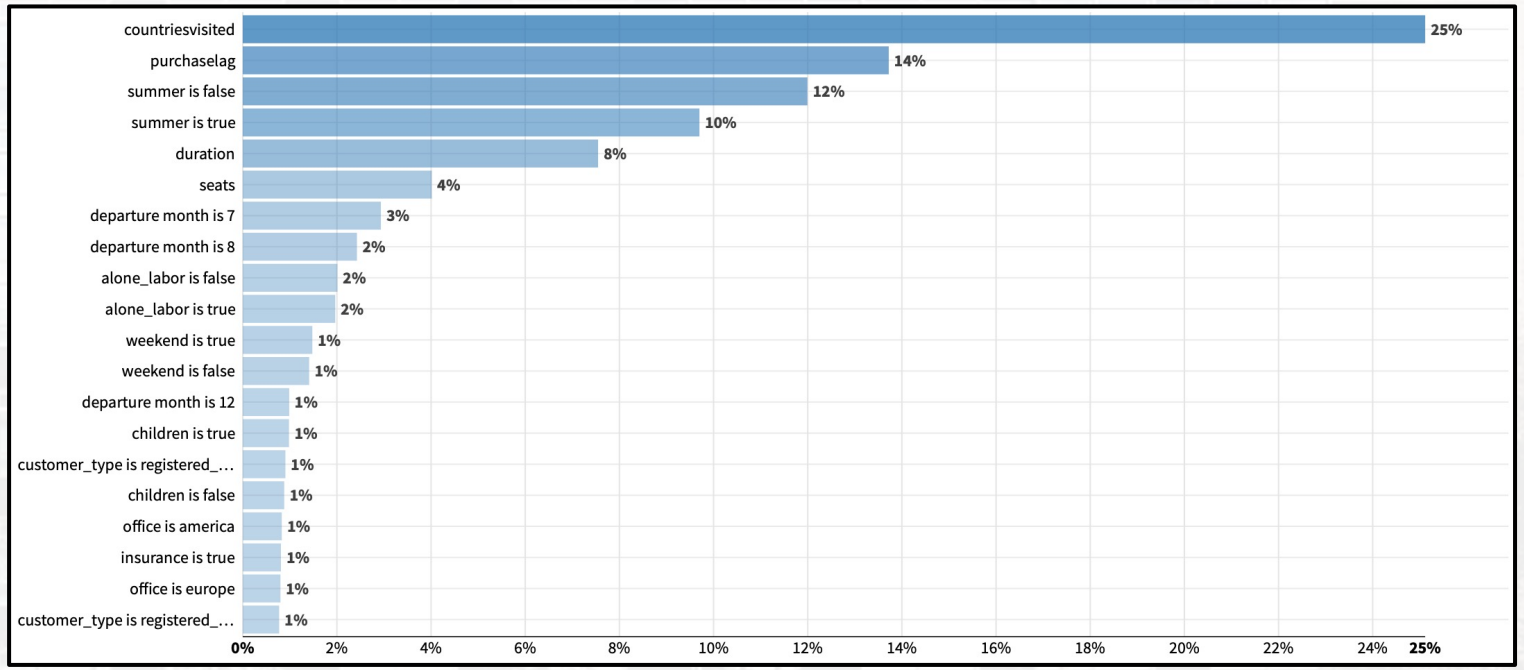


TECHNICAL: MODEL CONSTRUCTION II

2nd clusterings models

(1) Turn off departure month, departure date which shows common information with summer, leaving only the variable 'summer' (departuredate is in July and august).



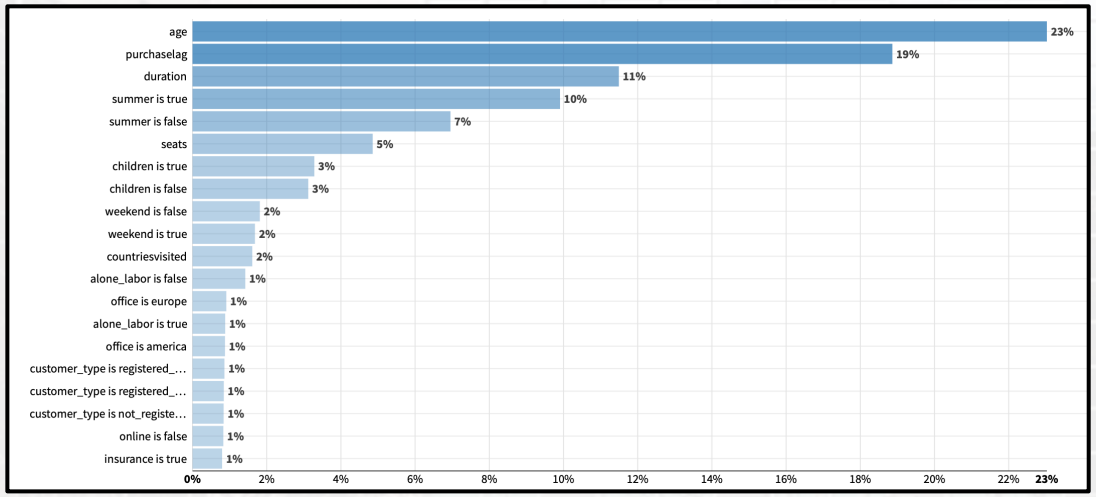


TECHNICAL: MODEL CONSTRUCTION III

3th clusterings model

We reconsider the age value since we think that the people who bought the tickets may have an information of the trip, like elder people buy tickets for family trips.

K-MEANS FOR AGE	
KMeans (k=5) (k-means for age)	0.107
KMeans (k=6) (k-means for age)	0.105
KMeans (k=4) (k-means for age)	0.111
KMeans (k=3) (k-means for…	0.125



4th clusterings heatmap

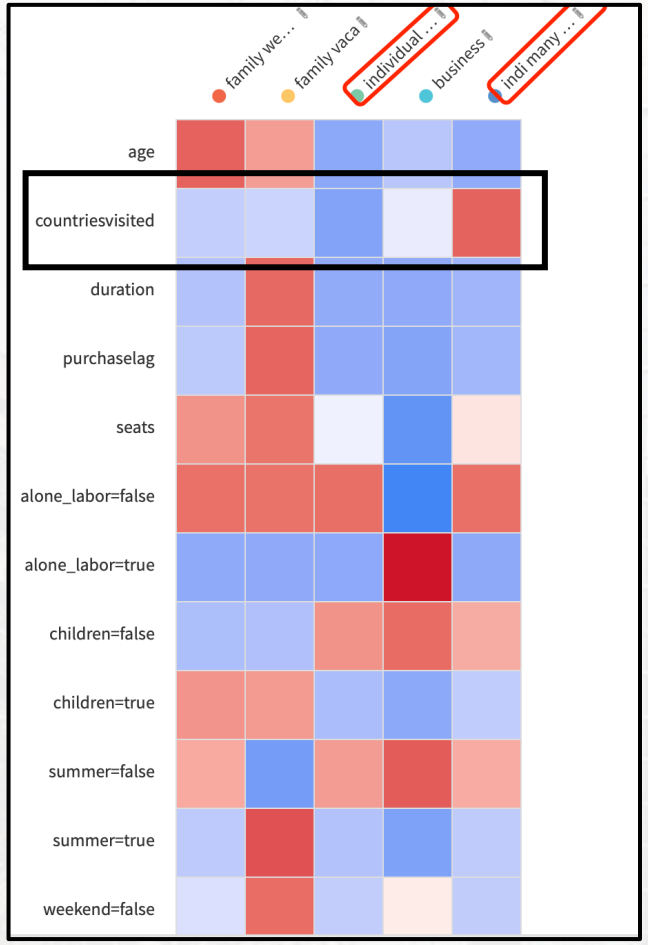
Features: insurance, customer_type, weekend, office, seats, summer, purchaselag, duration, countriesvisited, car, hotel, alone_labor, online,age.

(1) Hotel, office, Online, insurance, customer type, car, always been low importance and show little difference in the clustering in heat map.

(2) In this case, we have developed clusters that have obvious and stable characteristics.
(Family weekend/summer, individual weekend/business), however, in this case, we were expecting to see if we can cluster a cluster that might be individual summer long trips, however, it's very possible that young guys go with the family long summer trip.

WE CHOSE THE MOST HOLISTICAL MODEL IN K-MEANS

5th clusterings heatmap

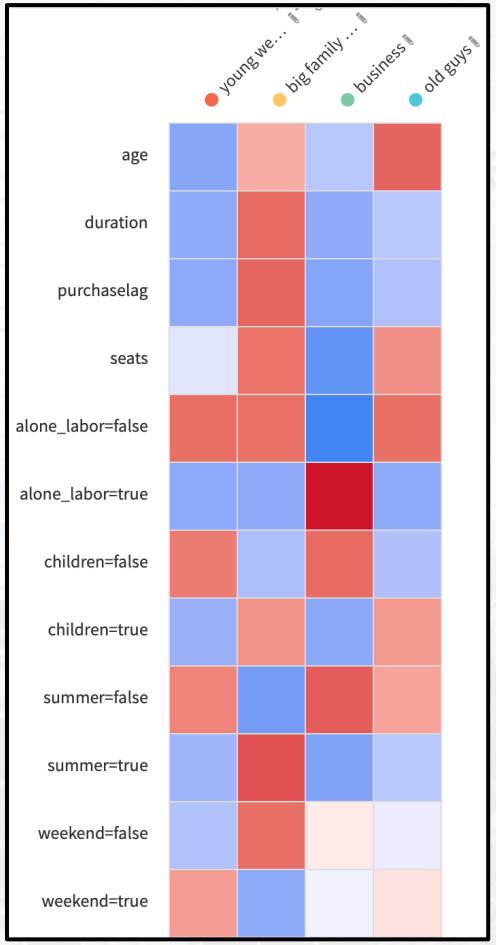


For cluster 3 and cluster 5, the features are basically the same **with only one difference in the countries visited**. Since other aspects are similar, differentiating between them does not help classify them from the whole group. **However, we can give some suggestions from here to the agency that young people tend to travel more to several countries.**

1. In this case, we can tell that **if we combine 3 and 5, there are basically 4 clusters with obvious characteristics,** so we'll continue clustering with 4 clusters.

2. After targeting 4 clusters, we found the perfect model by dropping the 'weekend' value since it's highly correlated with the value 'alone_labor'.

6th clusterings heatmap



Final Silhouete score: 0.275 where k = 4
Max achieved: 0.335 at k-means where k = 6

HOPE YOU LIKED IT,
SEE YOU ON THE NEXT ONE!