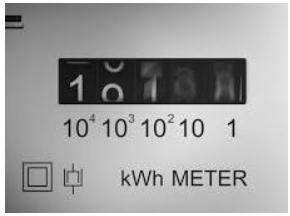
Household Electric Power Consumption

Cluster Analysis

Team members:
Hendro Lim – José Zepeda – Yiqing Hu



POLITECNICO DI MILANO
GRADUATE SCHOOL OF BUSINESS

CEFRIEL
DIGITAL INNOVATION

FT Executive Education Ranking 2016

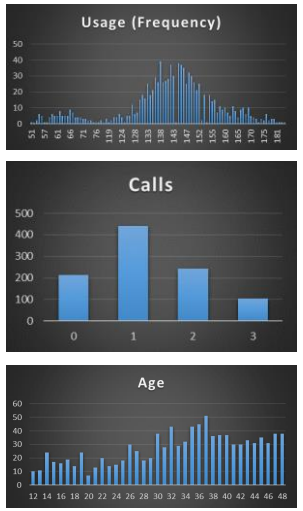
FT European Business Schools Ranking 2015

EFMD EQUIS ACCREDITED

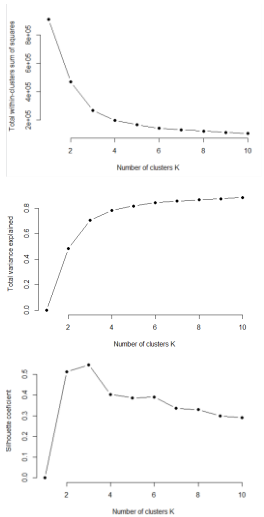
ASSOCIATION AMBA ACCREDITED

The methodology that we use to perform the analysis

Exploratory Analysis



Cluster Identifiication



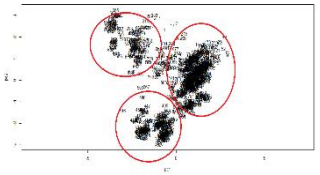
Attributes Identification and Transformations

- Group the customers' age; Assign binary attribute
- Split the total usage to day and night
- Standardize the data by considering its mean and standard deviation

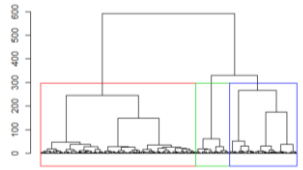
Model Development:

K-means/ Hierarchiecal

- K-means:



- Hirarchiecal



Model's Comparison

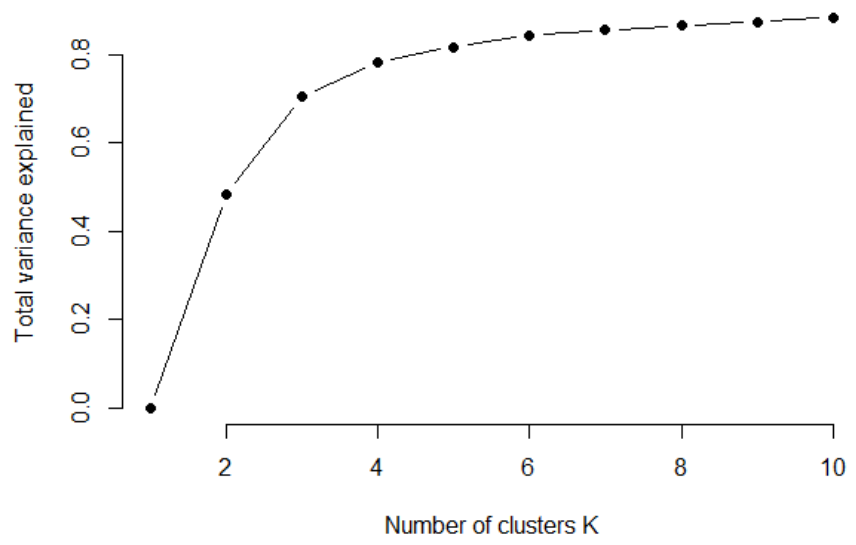
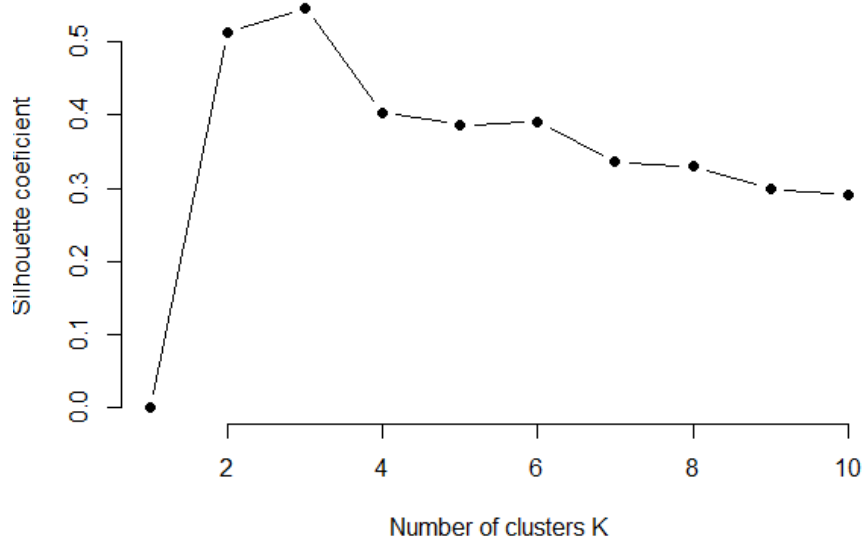
- Compare the cluster result from K-means and Hierarchiecal*

Results Interpretation

- Identify characteristics of clusters and strategies for each of them

** K-means is easier to be explained to business people compare to hierarchiecal*

Variance and silhouette coefficient indicate 3 clusters as the most optimal one

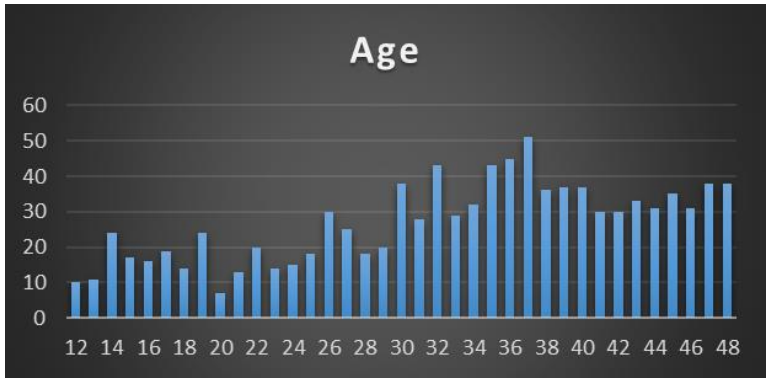
K-means Algorithm	Silhoutte Coefficient
	

- Based on the k-means approach, the first 3 clusters are able to explain **~80%** of total variance
- Moreover, silhouette coefficient also shows 3 clusters has the highest coefficient number
- Given that, the team will use 3 clusters for the analysis

Pre-processing is the critical process to follow prior to develop the model

Assign binary attribute to variable age:

- 2 years or less
- 2 to 3 years
- More than 3 years



Complete the data (on the usage):

- Calculate the day usage out of total EPC
- Extract the night usage

$$Night\ Usage = Total\ Usage\ - (\%\ Usage\ x\ Total\ Usage)$$

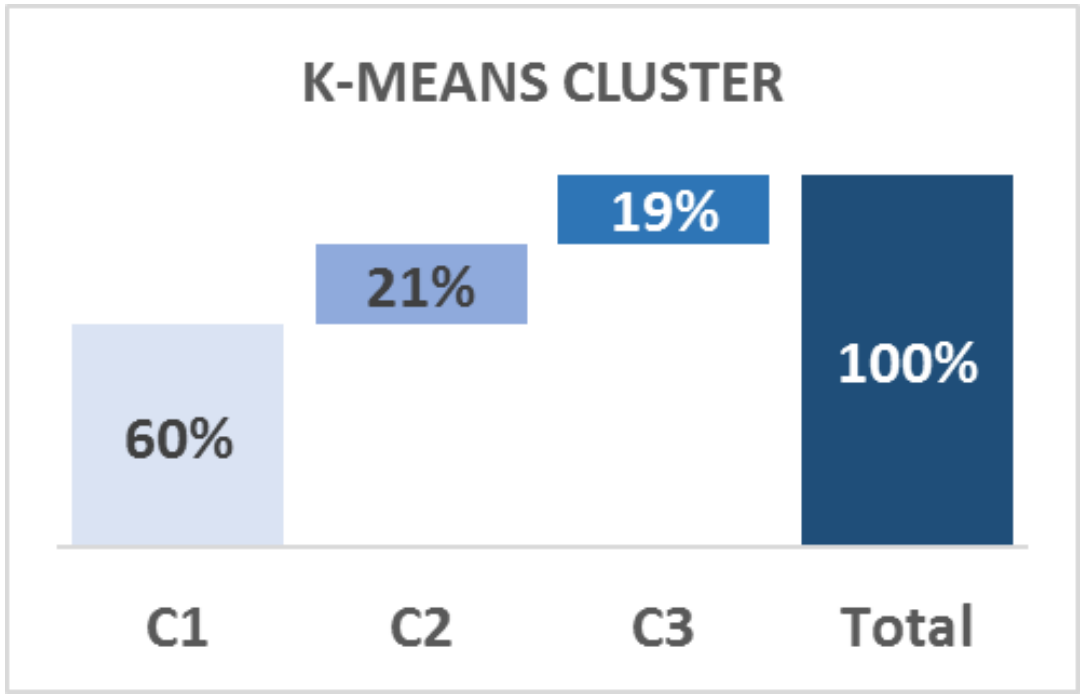
Standardize the data:

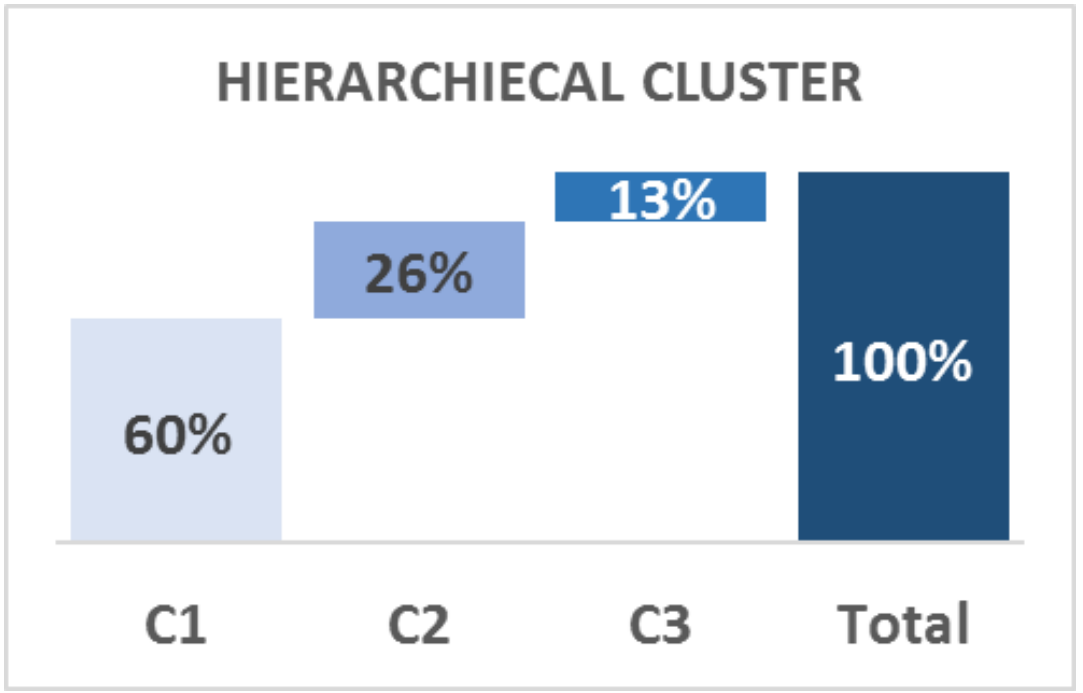
- To compare one with another variable

$$Standardize\ Value = (Value\ - Mean)/\ \text{Standard Deviation}$$

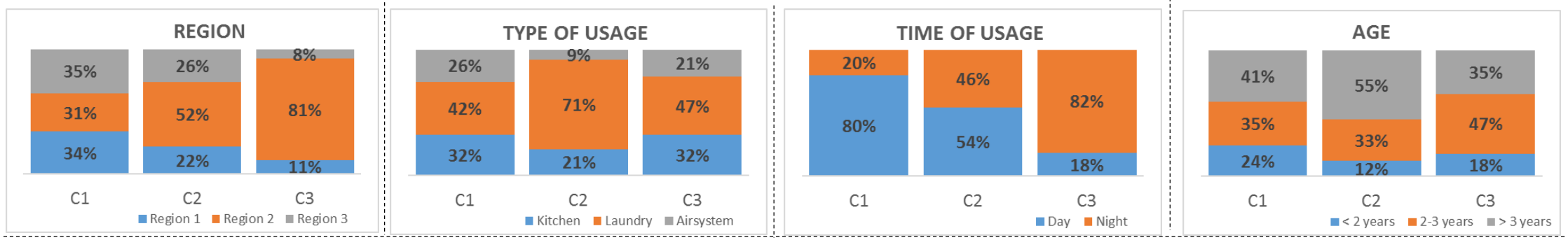
K-means and hierarchical clustering methodology give similar output

- Based on the clustering methodologies with 3 clusters, the team able to map the distribution of users per cluster
- Both methodologies show that there is a cluster which capture 60% of the users (population)





Based on the dimensions, we can categorize the clusters to 3 unique characteristics



C1 – FAMILY

Housewives who perform cleaning and cooking activities typically on day time

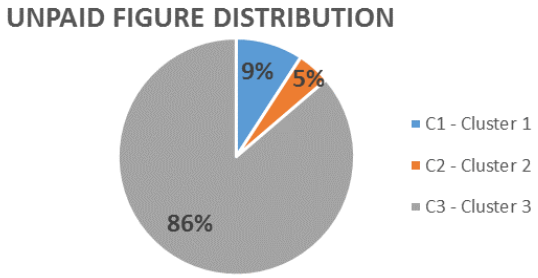
C2 – BOARDING HOUSE

Boarding house with several rooms rented to tenants, some of whom works on day and some on night

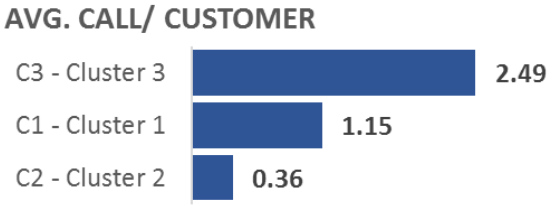
C3 – SINGLE WORKER

A person who rents or owns a house, but spends most of the day time at work

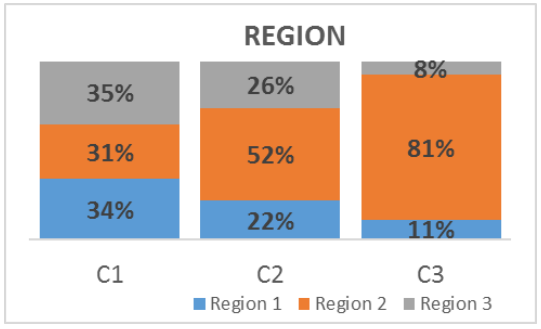
High level analysis shows that Cluster 3 has a very high unpaid figure



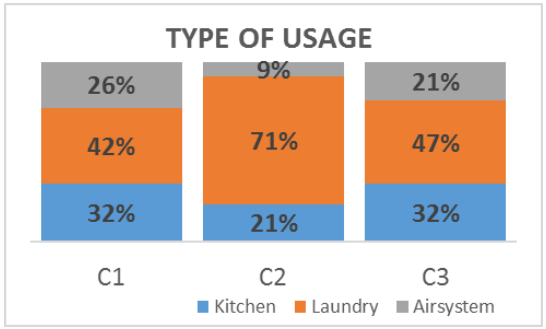
- Based on the unpaid figure distribution, we understand that **86%** of unpaid invoices in the last 3 months are belong to Cluster 3; Hence we need to focus on this segment since it will surely impact the company's bottom line



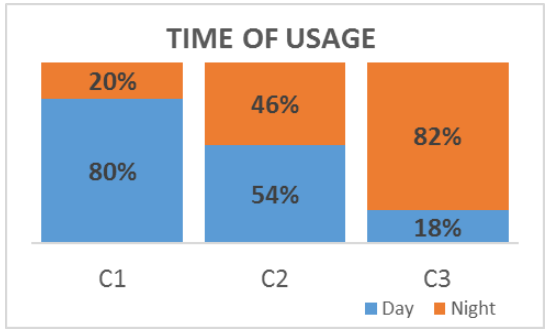
- The average call per customer figure strengthen indication that cluster 3 is the most interesting cluster to analyse, since the company's call center made quite high number of calls per each customer in that segment



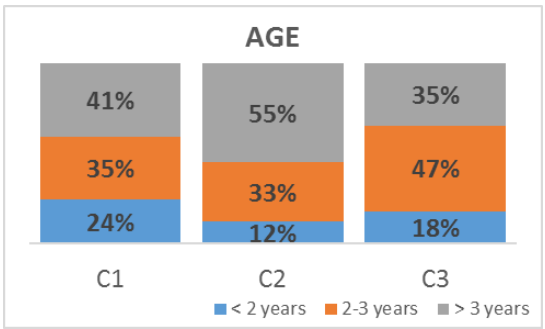
- Unlike C1 that has equal region distribution, **C2 and C3 distributed more to Region 2**



- From the type of usage dimension, we can see that only **C2 that has quite different EPC distribution, which heavily consumed on Laundry**



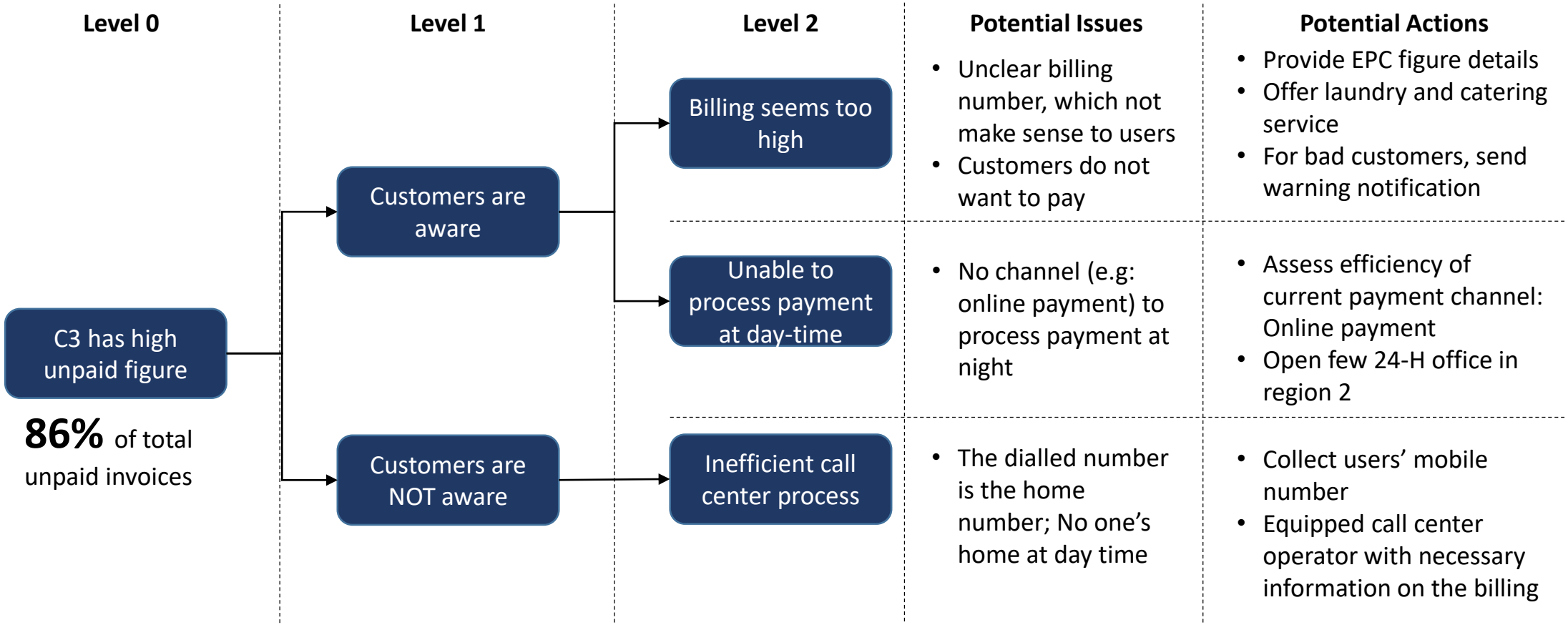
- There is a different time consumption among the clusters, where **C1 typically consumes on day while C3 consumes on night**



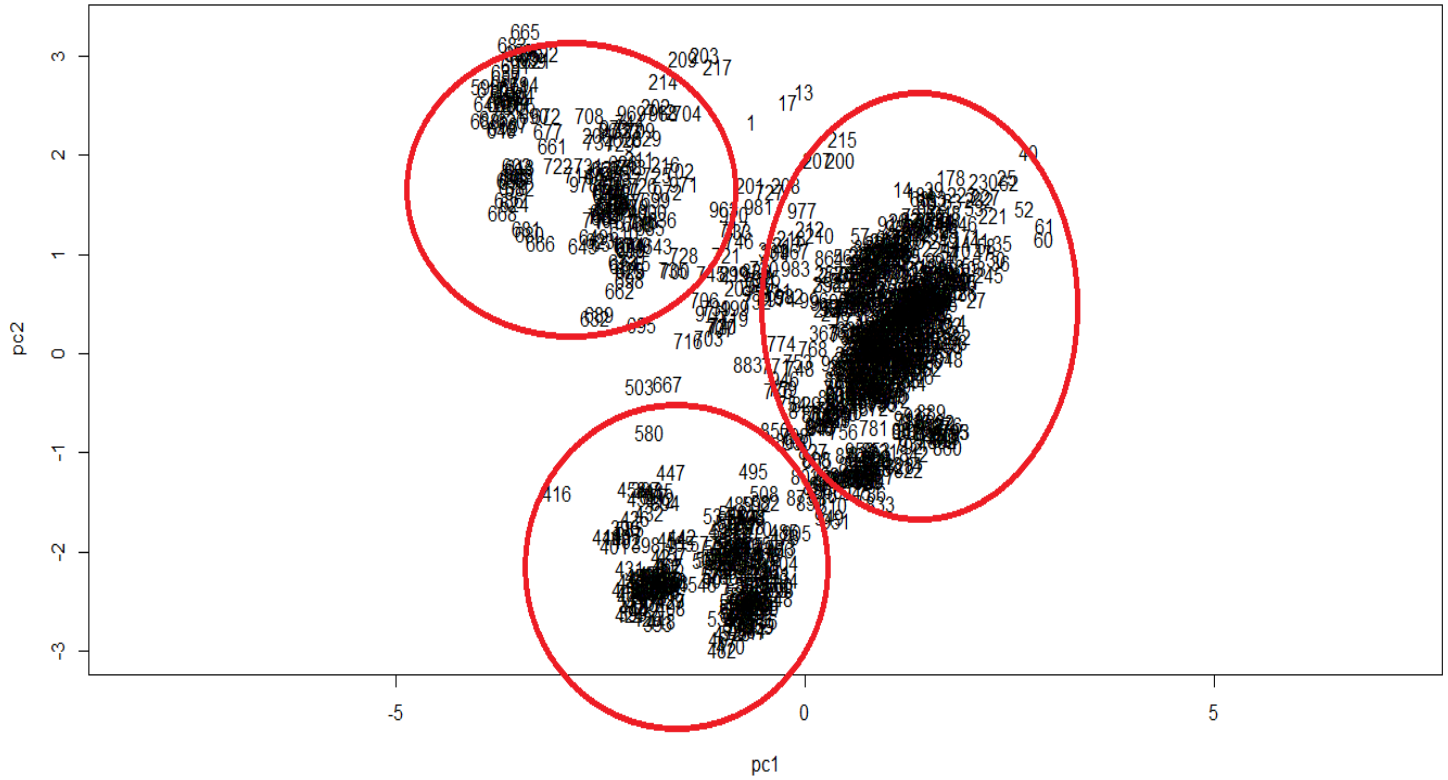
- Simply look at the distribution we can understand the customer's age figure, where across all clusters, old customer capture quite big portion

BABD

Post investigation, the team come up with some potential actions



Appendix: Principal Component Analysis to map the cluster into two dimensional figure



- Considering that the data have lot of dimensions that probably correlate each other, the team decided to perform **Principal Component Analysis**
- Based on the PCA, we will able to see the distribution of 3 clusters into two dimensional figure

BABD

Thank you!

Terima kasih!

Merci!

謝謝!