UNIVERSITÉ DE GENÈVE
GENEVA SCHOOL OF ECONOMICS AND MANAGEMENT

Customer Segmentation

Telecommunications

Huber Alejo Pablo
Camela David
Wieland Oscar

April 22th, 2024

Executive summary

This paper reveals, after analysis, four distinct customer segments within the telecommunications user base, offering vital insights for refining marketing strategies and tailoring product offerings.

Contents

1 Introduction

The primary objective of this paper is to define distinct customer segments. These segments offer valuable insights for the marketing team, enabling them to refine marketing strategies, tailor product offerings, and enhance overall customer engagement.

To achieve this goal, we have opted for the widely employed k-means clustering method. Clustering facilitates the identification of similarities among observations and the subsequent categorization based on these similarities or dissimilarities. At its core, k-means clustering aims to minimize the total intra-cluster variation, thereby defining clusters where objects within each cluster share significant similarities. Through this method, we aim to partition our customer base into meaningful groups for targeted marketing efforts and improved customer understanding.

Dataset Overview

We were provided with a large dataset made of 24 variables that can be separated in 5 main categories: demographic, service duration, usage metrics, average duration metrics, and cost-related variables.

Variables Description

Demographic Variables: The dataset records the *Age* of the subscriber.

Service Duration Variable: It includes *Length of Service (L_O_S)*, representing the duration in months that the subscriber has been using the service.

Usage Metrics: This category encapsulates data on service usage such as the number of *Dropped Calls*, counts of calls and minutes during *Peak*, *Off-Peak*, and *Weekend* periods. It also includes the total minutes of *International Calls* and both the total number and minutes of *National Calls*.

Average Duration Metrics: These metrics indicate the average duration of calls during *Peak*, *Off-Peak* and *Weekend* periods. It also includes the average duration of *National* calls.

Cost-Related Variables: The dataset details various cost metrics such as *Total National Call Cost*, which aggregates the total cost of national calls across different times, *Net National Billable Minutes* (number of billable national call minutes). It also provides the costs after applying free minutes (*Net National call cost per minute* and the *Net National Call Cost*). Overall costs including *Total Call Cost*, *Total Cost*, and *Average Cost Per Min* which combines all call costs and divides by the total number of call minutes.

2 Exploratory Data Analysis

As a preliminary step, we analyze the dataset and assess its quality. After reviewing the dataset, we were pleased to report that the dataset was already cleaned and is free of any missing values (NAs), which ensures the integrity and reliability of our subsequent analyses.

We chose to use k-means as our approach to tackle the problem because it's a straightforward and widely-used method for clustering similar data points together. With k-means, we aim to group our data into clusters based on their similarities. It's like putting things that are alike into the same box. This method is efficient and works well with large datasets, making it a popular choice for many data analysis tasks. Plus, it's relatively easy to understand and implement, which is great for getting quick insights into our data.

We wanted to see if the k-means method was a good fit for our model, so we checked if it followed some basic rules. With k-means clustering, it assumes that all the things we're looking at have about the same amount of variation. If they don't, the clustering might not work well, and we might end up with groups that don't make sense.

We can address this by standardizing our data. That means we make sure everything is on the same scale, so one type of data doesn't overshadow the others. This helps us figure out if k-means is the right approach for our model.

After implementing the standardization, we used various test to determine the correct number of clusters to use for the analysis like the elbow (Section 2.1) and the Average silhouette method (Section 2.2).

2.1 Elbow method

As a first approach, we used the Elbow method which evaluates the total within-cluster sum of squares for various numbers of clusters. The goal is to look for a point where the curve bends, resembling an elbow. This point indicates that adding more clusters no longer offers a significant improvement. However, it's important to understand that the primary role of the elbow plot is to offer insights, and its interpretation can often be ambiguous and subjective. In figure 1 we can see that the correct number of cluster lies somewhere between 3 and 6. After that point adding clusters doesn't seem to strongly impact the total within sum of square.

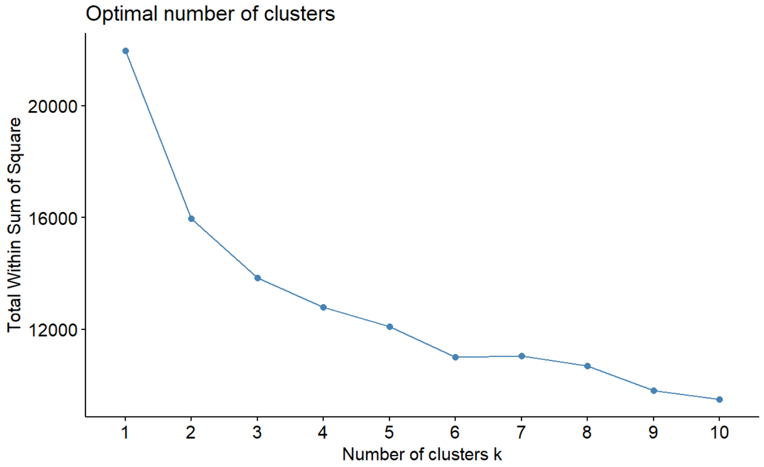

Figure 1: Graphical representation of the Elbow Method

2.2 Silhouette method

The average silhouette method assesses the coherence of each data point within its assigned cluster across different values of k. The ideal number of clusters, k, is determined by maximizing the average silhouette score computed over a range of possible k values.

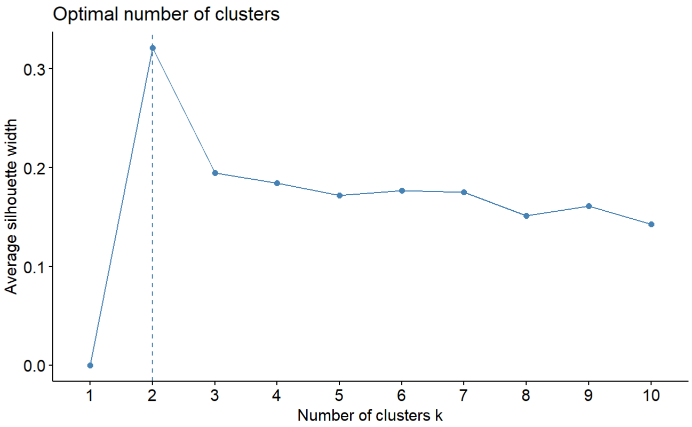

Figure 2: Graphical representation of the Silhouette method

The assessment of the optimal number of clusters involved multiple methods including the elbow and silhouette techniques, along with visual inspection, as depicted in Figures 2 and 4. Initially, the results from Figure 2 indicated that using just two clusters might be sufficient; however, discrepancies between the two methods prompted a further visual assessment of different cluster numbers ranging from two to six. This examination was detailed in Figure 4, where it became evident that employing four clusters optimally separated a small group of points in the bottom-left corner, enhancing cluster distinction.

Upon integrating these three approaches (elbow, silhouette, and visual inspection), we analyzed the homogeneity of cluster compositions by evaluating the mean of the variables that constitute each cluster. This comprehensive evaluation confirmed that a configuration of four clusters yielded the greatest homogeneity among the clusters, establishing it as the best choice for our clustering model.

3 Results

In our study, we conducted a k-means clustering analysis to create four distinct customer segments. Subsequently, we conducted an in-depth examination of the mean values of each variable within each cluster to reveal the unique characteristics defining each segment. This analysis provided valuable insights into the distinguishing traits of each cluster, facilitating a deeper understanding of the clustering outcomes.

During our investigation, we identified several variables, including Age and Length of Stay, where the means across clusters were similar. Understanding the potential impact of these variables on the interpretability of the clustering results, we considered removing them from the analysis to refine the clustering process. To determine whether their exclusion was necessary, we examined the variances of these variables across clusters using scaled data.

After analyzing Table 1 and Figure 6, we observed that variables with variances around 0.05 were less significant in identifying distinct customer segments. Therefore, we chose to consider variables with values around 0.05 for potential removal. Removing these identified variables helped streamline the clustering process and improve the interpretability of the results. With a refined set of variables, we conducted a follow-up k-means clustering analysis with four clusters, aiming to achieve clearer distinctions between customer segments. The results are illustrated in Figure (3).

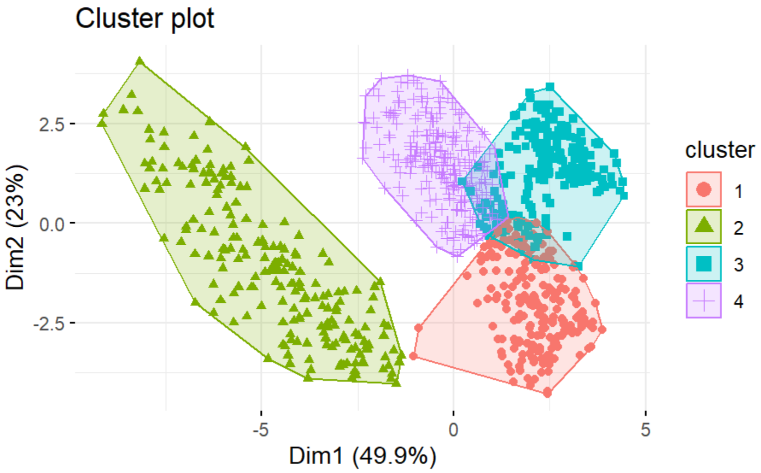

Figure 3: K-means clustering

Upon completion of the refined clustering analysis, we once again examined the mean values of each variable within each cluster to gain insights into the characteristics defining each segment.

3.1 Detailed Cluster Analysis

Based on Figure 5 and Table 2, we can derive conclusions and elucidate the distinctions among the various clusters.

Cluster 1: Economic Users

Cluster 1 has the lowest average for total peak calls, with about 60 calls and 193 minutes, indicating minimal usage during common periods. However, their off-peak usage is relatively higher, with about 152 calls and 364 minutes. Weekend activity is modest, with 18 calls and 56 minutes. International call minutes are moderate at 152. This cluster has the second lowest total call cost, approximately $62.55, and the lowest total cost (including fixed charges). These individuals are likely economic consumers who despite not being the ones with the lowest usage, are the one who pay the less in total.

- **Interpretation**: This segment likely includes price-sensitive individuals who strategically use services during off-peak hours to minimize costs.
- **Business Insight**: Target this group with off-peak promotions, such as reduced-rate bundles or incentives for prepaid top-ups during non-peak hours. Since they are sensitive to pricing, retaining them through loyalty programs or customized discounts could prevent churn.

Cluster 2: High-Usage Premium

The users in Cluster 2 exhibit the highest usage levels, particularly during peak hours, with an average of 411 calls and 944 minutes. Their off-peak usage is also substantial, with 167 calls and 425 minutes. Despite their high activity levels, their weekend call minutes are quite similar to the other groups. They make significant international calls, the most among all clusters, at 317 minutes which could indicate that they are business users. This cluster incurs the highest total call and total costs, about $113.92 and $263.92, respectively, reflecting their premium usage. However, they manage costs effectively, with a lower average cost per minute of $0.152.

- **Interpretation**: Likely composed of professionals or business users who depend on telecommunication services for work, including international communication.
- **Business Insight**: Offer premium services such as international call packages, priority customer service, or unlimited peak-time plans. These users are willing to spend for convenience, so upselling advanced services (e.g., faster internet, business-level packages) would cater to their needs.

Cluster 3: Low usage users

This cluster represents users with low call volumes especially during off peak period with 48 call only and also during peak period with 119 calls. They are slightly more active over the weekends, with the highest usage at 23 calls and 65 minutes. International usage is the lowest among clusters with 113 minutes only. However, despite having the lowest national call usage they pay a high price in national call cost and their average cost per minute is the highest among clusters regarding calls. Additionally these users pay a total cost of $139 which make them the third most expensive cluster despite having the lowest phone usage.

- **Interpretation**: This segment might include individuals who use their phones sporadically but do not optimize their plans, potentially paying more for limited usage.
- **Business Insight**: Introduce low-cost, usage-based plans or flexible pay-as-you-go options. Educate this group on optimizing their plan selection to match their usage patterns. Additionally, highlight how they could reduce their costs by shifting to better-suited plans to improve satisfaction and retention.

Cluster 4: Balanced Medium Users

Users in Cluster 4 have moderate activity levels with 205 peak calls and 545 minutes, and similar patterns in off-peak times with 83 calls and 209 minutes. Their weekend usage is the lowest, with only 13 calls and 42 minutes, but their international minutes are moderately high at 174. The total call costs and overall costs are mid-range, approximately $76.81 and $179.78, respectively. This cluster strikes a medium balance between usage and cost, reflected in an average cost per minute of $0.18.

- **Interpretation**: Likely a mix of personal and professional users who balance phone use between work and personal life.

- **Business Insight**: This group could be swayed by family plans or multi-line discounts, as they likely look for value in both personal and business usage. You could cross-sell complementary services like data packages or entertainment subscriptions, as they are neither price-sensitive nor premium spenders.

4 Conclusion

5 Conclusion

In conclusion, the application of k-means clustering allowed us to effectively segment the telecommunications customer base into four distinct clusters, each with unique usage patterns and cost behaviors. These insights enable businesses to:

- **Cluster 1 (Economic Users)**: Provide tailored off-peak promotions to price-sensitive customers.
- **Cluster 2 (High-Usage Premium)**: Develop premium packages for heavy business or professional users with high international call volumes.
- **Cluster 3 (Low-Usage Users)**: Offer flexible, low-usage plans or educate users on optimizing their plans.
- **Cluster 4 (Balanced Medium Users)**: Introduce family plans, multi-line discounts, and cross-sell complementary services like data or entertainment packages.

By leveraging these actionable insights, businesses can enhance customer satisfaction, increase retention, and optimize marketing strategies. These segments also provide opportunities to up-sell and cross-sell targeted services, maximizing the lifetime value of each customer group.

A Supplementary Figures

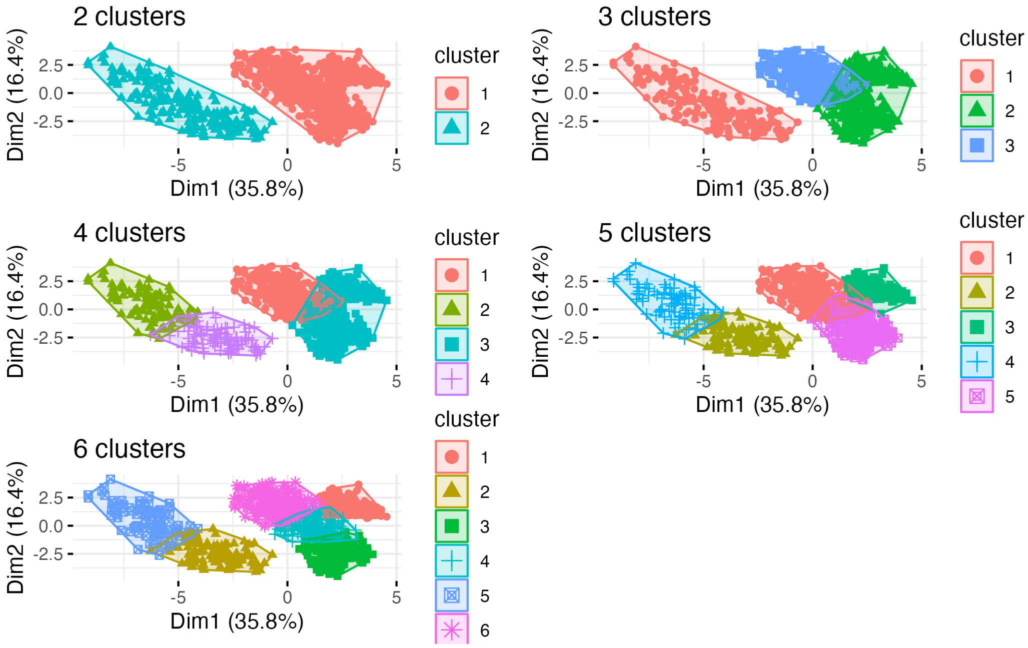

Figure 4: Visual inspection depending on the number of clusters

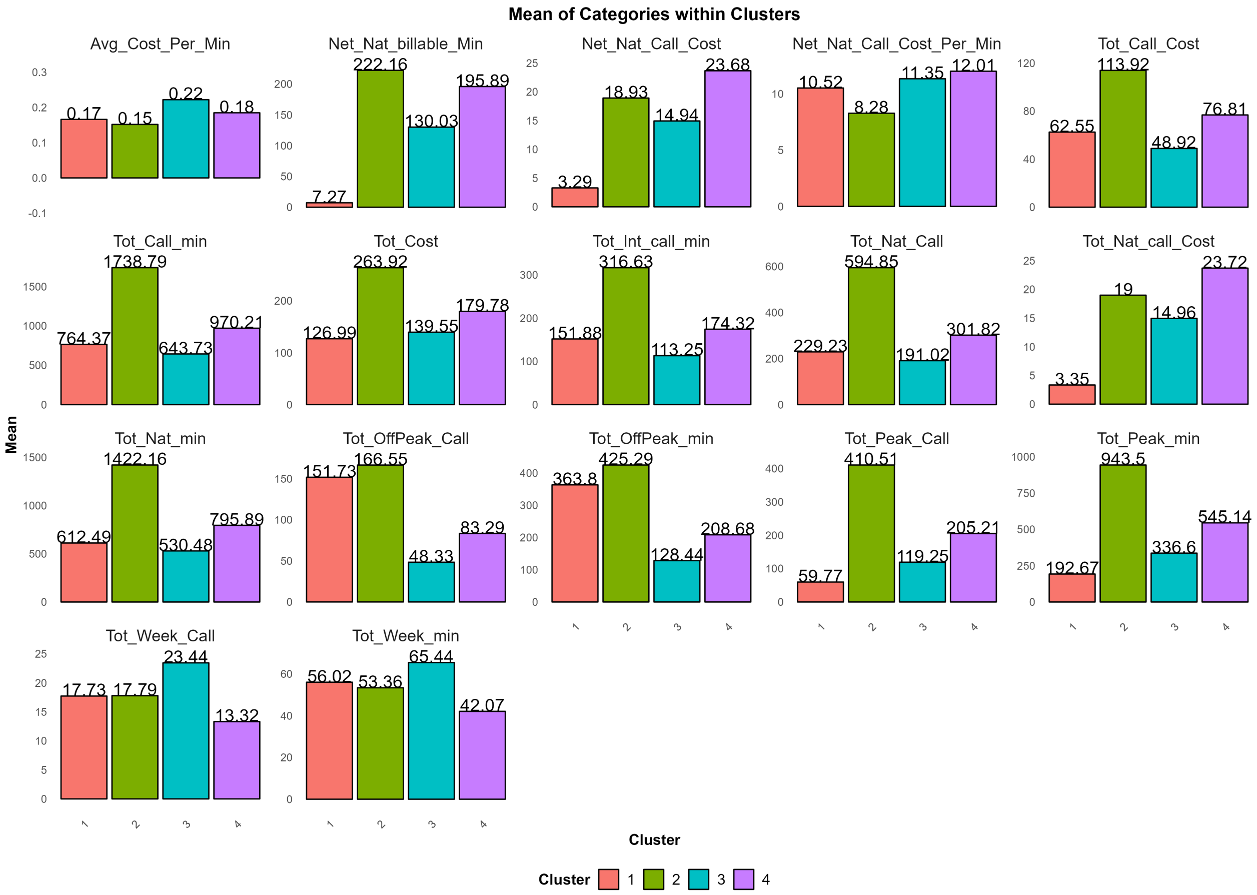

Figure 5: Mean of categories within clusters

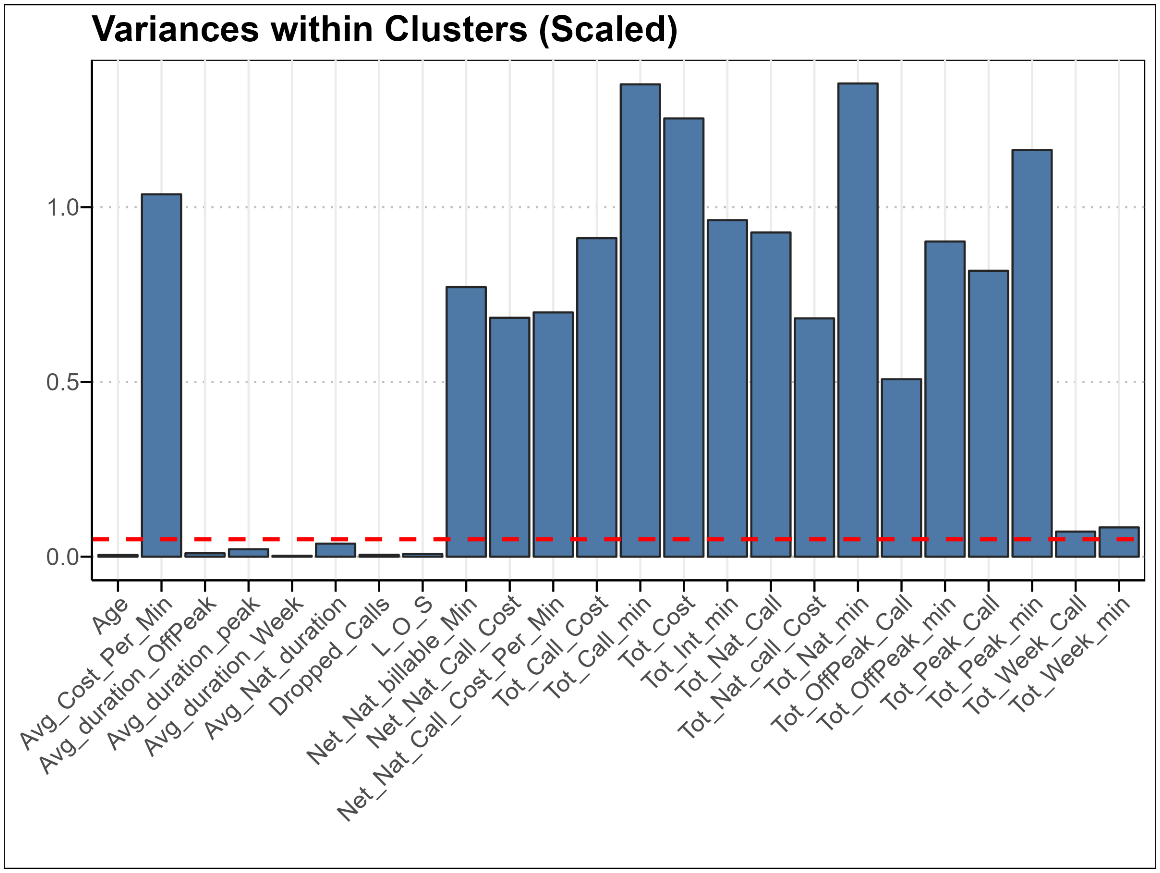

Figure 6: Visualization of variances between clusters

B Additional Tables

Table 1: Cluster Variances of Scaled Variables

Cluster Variances			
Variable	**Value**	**Variable**	**Value**
Age	1.666667	L_O_S	0.004914511
Dropped_Calls	0.005627958	Tot_Peak_Call	0.8180269
Tot_Peak_min	1.163501	Tot_OffPeak_Call	0.507684
Tot_OffPeak_min	0.9017362	Tot_Week_Call	0.07184939
Tot_Week_min	0.0837561	Tot_Int_min	0.9628666
Tot_Nat_call_Cost	0.6817291	Avg_duration_peak	0.02133707
Avg_duration_OffPeak	0.009727948	Avg_duration_Week	0.00266792
Tot_Nat_Call	0.9274528	Tot_Nat_min	1.353829
Avg_Nat_duration	0.03736951	Tot_Call_min	1.351304
Net_Nat_billable_Min	0.7711961	Net_Nat_Call_Cost_Per_Min	0.698902
Net_Nat_Call_Cost	0.6835523	Tot_Call_Cost	0.9111978
Tot_Cost	1.253856	Avg_Cost_Per_Min	1.036797

Table 2: Variable Means per Cluster

Cluster 1			
Variable	**Mean**	**Variable**	**Mean**
Tot_Peak_Call	59.77391	Tot_Peak_min	192.6652
Tot_OffPeak_Call	151.73043	Tot_OffPeak_min	363.8048
Tot_Week_Call	17.72609	Tot_Week_min	56.01870
Tot_Int_min	151.8837	Tot_Nat_call_Cost	3.349289
Tot_Nat_Call	229.2304	Tot_Nat_min	612.4887
Tot_Call_min	764.3724	Net_Nat_billable_Min	7.271304
Net_Nat_Call_Cost_Per_Min	10.520039	Net_Nat_Call_Cost	3.289944
Tot_Call_Cost	62.54599	Tot_Cost	126.9915
Cluster 2			
Tot_Peak_Call	410.51256	Tot_Peak_min	943.5045
Tot_OffPeak_Call	166.54774	Tot_OffPeak_min	425.2945
Tot_Week_Call	17.78894	Tot_Week_min	53.36131
Tot_Int_min	316.6281	Tot_Nat_call_Cost	18.999744
Tot_Nat_Call	594.8492	Tot_Nat_min	1422.1603
Tot_Call_min	1738.7884	Net_Nat_billable_Min	222.160302
Net_Nat_Call_Cost_Per_Min	8.279563	Net_Nat_Call_Cost	18.930997
Tot_Call_Cost	113.91943	Tot_Cost	263.9194
Cluster 3			
Tot_Peak_Call	119.25339	Tot_Peak_min	336.6000
Tot_OffPeak_Call	48.32579	Tot_OffPeak_min	128.4407
Tot_Week_Call	23.44344	Tot_Week_min	65.44344
Tot_Int_min	113.2490	Tot_Nat_call_Cost	14.955765
Tot_Nat_Call	191.0226	Tot_Nat_min	530.4842
Tot_Call_min	643.7331	Net_Nat_billable_Min	130.031674
Net_Nat_Call_Cost_Per_Min	11.346386	Net_Nat_Call_Cost	14.941177
Tot_Call_Cost	48.91587	Tot_Cost	139.5484
Cluster 4			
Tot_Peak_Call	205.20677	Tot_Peak_min	545.1406
Tot_OffPeak_Call	83.29323	Tot_OffPeak_min	208.6782
Tot_Week_Call	13.31955	Tot_Week_min	42.07256
Tot_Int_min	174.3209	Tot_Nat_call_Cost	23.721520
Tot_Nat_Call	301.8195	Tot_Nat_min	795.8914
Tot_Call_min	970.2122	Net_Nat_billable_Min	195.891353
Net_Nat_Call_Cost_Per_Min	12.014201	Net_Nat_Call_Cost	23.676505
Tot_Call_Cost	76.81040	Tot_Cost	179.7776