# Capstone Project 1 – In Depth Analysis

## **FEATURE ENGINEERING**

Our original dataset is a time series data with 5 features; smsIn, smsOut, callIn, callOut and internet volumes. For clustering our 10,000 grids into different groups, we will convert this time series data into grid-wise data. Creating a total of smsIn, smsOut, callIn, callout and internet volumes for each grid will have very minimal information about the behavior patterns of a grid’s telecommunication activities. Hence, we will perform Feature Engineering which is nothing but extracting more information from the existing time series data that helps Clustering algorithm to understand each grid better.

We have created a total of 83 features that are indexed by grid id.

|  |  |
| --- | --- |
| Features | *Description* |
| Weekend Hourly: |  |
| hourlysmsMax\_WE  hourlycallMax\_WE hourlyinternetMax\_WE | Maximum hourly volume of SMS sent and received, call made and received, internet accessed during weekend |
| hourlysmsMin\_WE  hourlycallMin\_WE  hourlyinternetMin\_WE | Minimum hourly volume of SMS sent and received, call made and received, internet accessed during weekend |
| hourlysmsAvg\_WE  hourlycallAvg\_WE  hourlyinternetAvg\_WE | Average hourly volume of SMS sent and received, call made and received, internet accessed during weekend |
| Weekend Daily: |  |
| smsMax\_WE  callMax\_WE  internetMax\_WE | Maximum daily volume of SMS, Call & Internet during weekend |
| smsMin\_WE  callMin\_WE  internetMin\_WE | Minimum daily volume of SMS, Call & Internet during weekend |
| smsAvg\_WE  callAvg\_WE  internetAvg\_WE | Average daily volume of SMS, Call & Internet during weekend |
| totalSmsDay\_WE  totalCallDay\_WE  totalInternetDay\_WE | Total SMS, Calls & Internet from 8AM till 10PM on weekends |
| totalSmsNight\_WE  totalCallNight\_WE  totalInternetNight\_WE | Total SMS, Calls & Internet from midnight till 8AM on weekends |
| *WeekDay Hourly:* |  |
| hourlysmsMax\_WD  hourlycallMax\_WD hourlyinternetMax\_WD | Maximum hourly volume of SMS sent and received, call made and received, internet accessed during weekday |
| hourlysmsMin\_WD  hourlycallMin\_WD  hourlyinternetMin\_WD | Minimum hourly volume of SMS sent and received, call made and received, internet accessed during weekday |
| hourlysmsAvg\_WD  hourlycallAvg\_WD  hourlyinternetAvg\_WD | Average hourly volume of SMS sent and received, call made and received, internet accessed during weekday |
| *WeekDay Daily:* |  |
| smsMax\_WD  callMax\_WD  internetMax\_WD | Maximum daily volume of SMS, Call & Internet during weekday |
| smsMin\_WD  callMin\_WD  internetMin\_WD | Minimum daily volume of SMS, Call & Internet during weekday |
| smsAvg\_WD  callAvg\_WD  internetAvg\_WD | Average daily volume of SMS, Call & Internet during weekday |
| totalSmsDay\_WD  totalCallDay\_WD  totalInternetDay\_WD | Total SMS, Calls & Internet from 8AM till 10PM on weekdays |
| totalSmsNight\_WD totalCallNight\_WD  totalInternetNight\_WD' | Total SMS, Calls & Internet from midnight till 8AM on weekdays |
| Daily: |  |
| dailySmsIn/dailySmsOut | Ratio of SMS received to SMS sent daily |
| dailyCallIn/dailyCallOut | Ratio of Calls received to Calls made daily |
| dailySms/dailyCall | Ratio of daily SMS to daily Call volumes |
| dailyInternet/dailySmsCall | Ratio of daily Internet to daily SMS & Call volumes |
| 'totalSmsDay\_WD', 'totalCallDay\_WD', 'totalInternetDay\_WD' | Total SMS, Calls & Internet from Midnight to 8AM |
| Weekly: |  |
| smsAvgdiff\_weekly callAvgdiff\_weekly internetAvgdiff\_weekly | Average of difference in the volume of SMS, Calls & Internet from one week to another |
| smsMax\_weekly  callMax\_weekly  internetMax\_weekly | Maximum volume of weekly SMS, Calls & Internet |
| smsMin\_weekly  callMin\_weekly  internetMin\_weekly | Minimum volume of weekly SMS, Calls & Internet |
| smsAvg\_weekly  callAvg\_weekly  internetAvg\_weekly | Average volume of weekly SMS, Calls & Internet |
| Monthly: |  |
| monthlyAvg\_sms  monthlyAvg\_call  monthlyAvg\_internet | Average volume of monthly SMS, Calls & Internet |
| smsAvg\_Nov  callAvg\_Nov  internetAvg\_Nov | Average volume of November month SMS, Calls & Internet |
| smsAvg\_Dec  callAvg\_Dec  internetAvg\_Dec | Average volume of December SMS, Calls & Internet |
| smsMax\_Nov  callMax\_Nov  internetMax\_Nov | Maximum volume of November month SMS, Calls & Internet |
| smsMax\_Dec  callMax\_Dec  internetMax\_Dec | Maximum volume of December month SMS, Calls & Internet |
| smsMin\_Nov  callMin\_Nov  internetMin\_Nov | Minimum volume of November month SMS, Calls & Internet |
| smsMin\_Dec  callMin\_Dec  internetMin\_Dec | Minimum volume of December month SMS, Calls & Internet |
| *Christmas & New Year* |  |
| totalSms\_xMas  totalCall\_xMas  totalInternet\_xMas | Total SMS, Calls & Internet volumes on Christmas day |
| totalSms\_NewYear  totalCall\_NewYear totalInternet\_NewYear | Total SMS, Calls & Internet volumes on New Year day |
| totalSms\_NewYearEve totalCall\_NewYearEve totalInternet\_NewYearEve | Total SMS, Calls & Internet volumes on New Year Eve [Dec 31st 6Pm to 1AM] |
| *Totals* |  |
| totalSmsIn  totalSmsOut  totalCallIn  totalCallOut  totalSMS  totalCall  totalInternet | Grid-wise total SMS-In, SMS-Out, Call-In, Call-Out, SMS, Calls & Internet |

## **K-MEANS CLUSTERING**

We will apply K-Means algorithm from Sci-kit learn package for clustering the grids. K-Means iteratively partitions the dataset into K subgroups, such that each data point belongs to only one group (no overlapping). Data points are assigned to a cluster such that its sum of the squared distance from the cluster’s centroid is at the minimum.

There are few steps to follow in order to prepare the dataset for K-Means model,

1. Remove all NAN values from the dataset after creating new features.
2. Standardization of the data: Since clustering algorithms use distance-based measurements to determine the similarity between data points, it’s recommended to standardize the data to have a mean of zero and a standard deviation of one since almost always the features in any dataset would have different units of measurements.

But our dataset has features with same unit of measurement, which is volumes of telecommunication activities, thus, we do not do any standardization.

1. All column values are converted to NumPy array, which is the input format for Sci-kit learn K-Means algorithm.
2. Number of subgroups must be pre-determined from the dataset.

### **Finding the Optimal K value**

#### Elbow Method:

It is a plot of sum of squared distance (SSE) between data points and their assigned clusters’ centroids for a range of K values. We pick K at the spot where SSE starts to flatten out and forming an elbow.

From our plot we have two optimal candidates for K, K=6 & K=7 beyond which the plot plateaus.

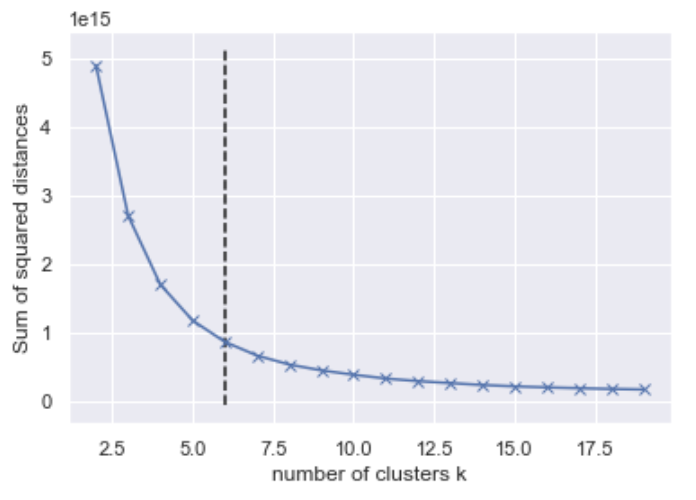
## 

*Fig 1: Elbow method to find the optimal K value*

#### Kneed package:

Visually inspecting the plot to identify the Knee/Elbow point could be confusing as in our case. We will take help of Kneed package the mathematically computes the Knee/Elbow point.

For our dataset, Kneed package has returned K=6 as the knee point.

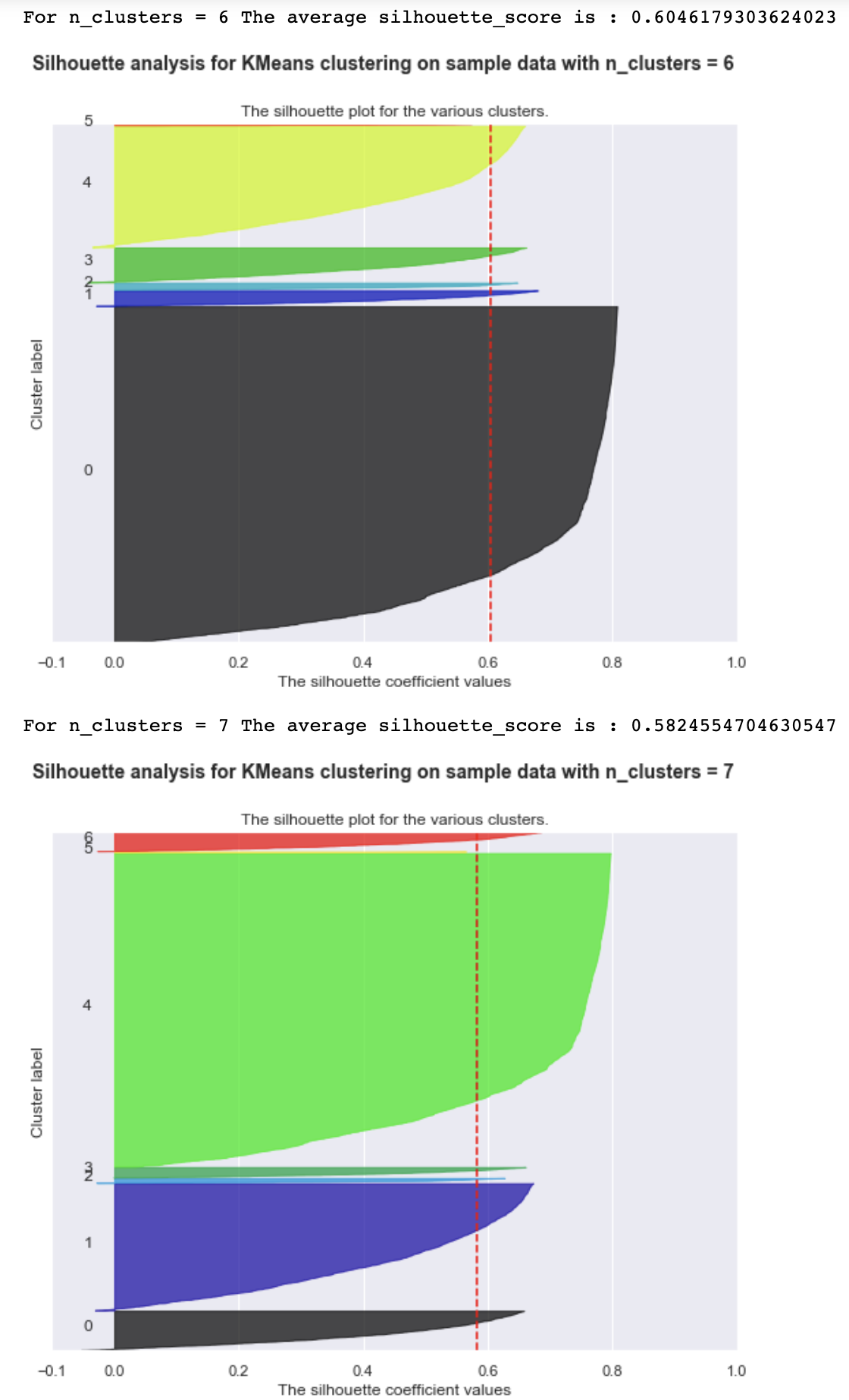


*Fig 2: Kneed package output shows K=6*

#### Silhouette coefficient plot:

The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. Silhouette coefficients (as these values are referred to as) near +1 indicate that the sample is far away from the neighboring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster.

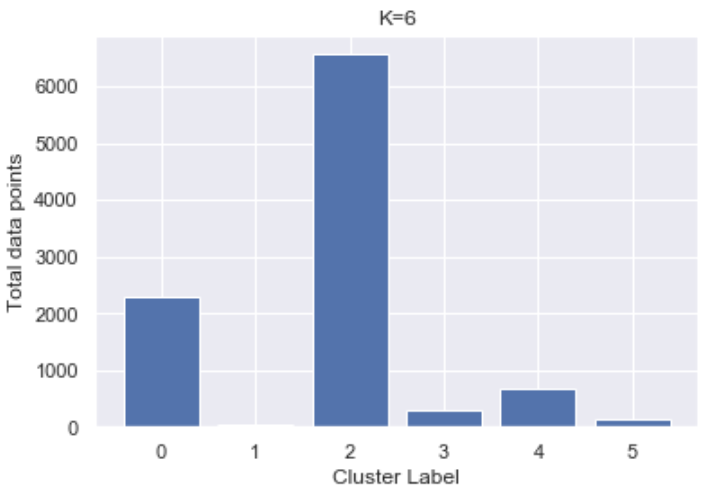
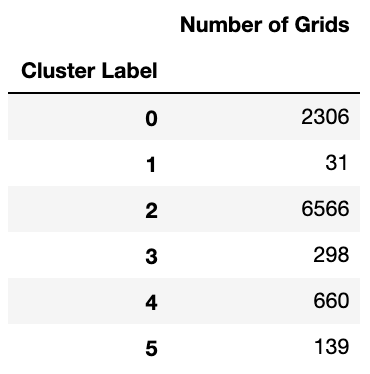
For our dataset we will pick K=6, as Silhouette coefficient plot shows a smaller number of datapoints assigned to the wrong cluster for K=7.



*Fig 3: Plot of Silhouette coefficients of all the data points for K=6 & K=7*

### **K-Means Clustering with K = 6**

Applying the model with K=6, results in 6 subgroups with distribution of grids as shown below,

*Fig 4: Subgroups after K-Means Clustering with K=6*

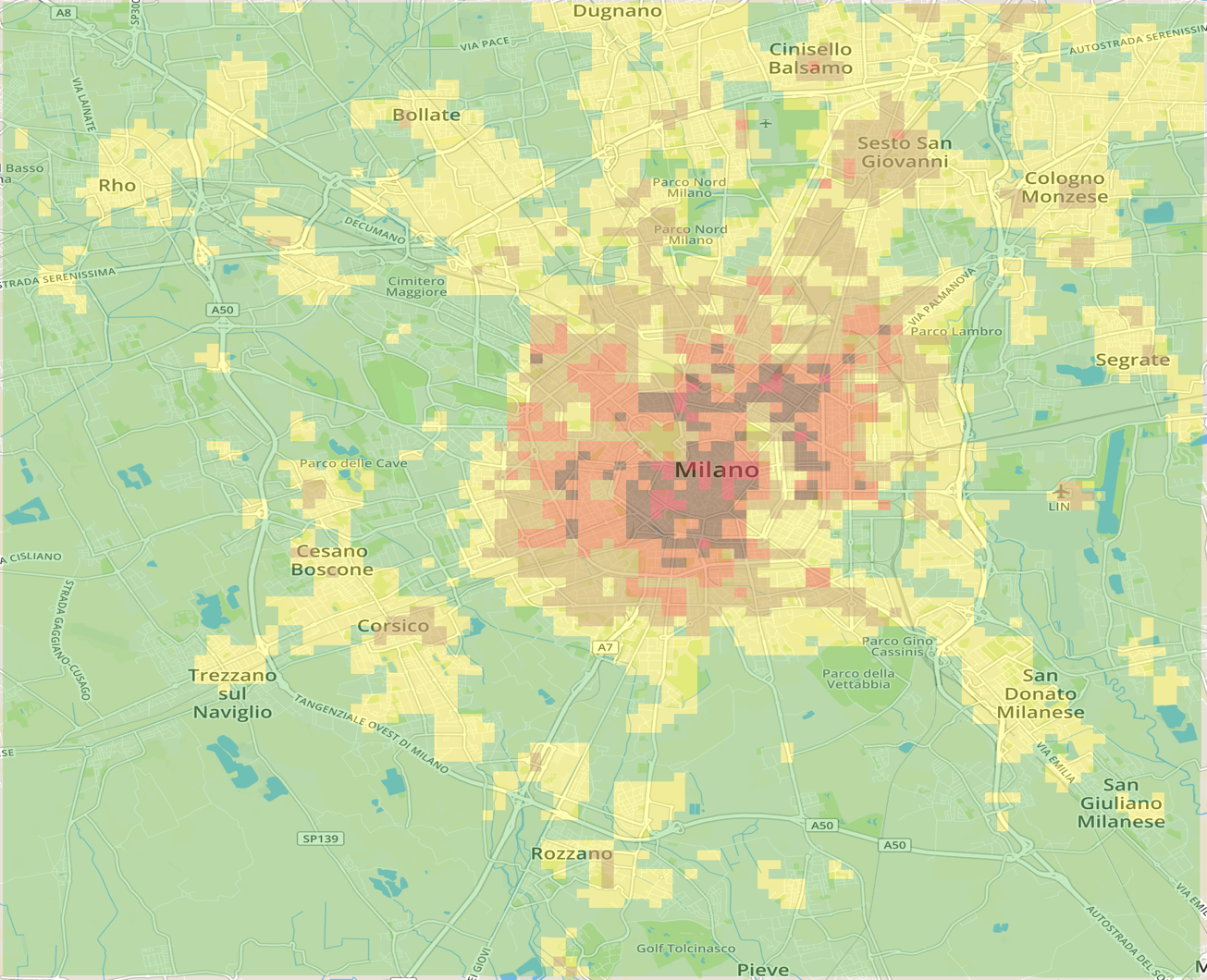
### **Principal Component Analysis**

Displaying the clusters in subgroups in 92 dimensions is not possible. We will reduce the dimensions to 2 dimensions using PCA in order to visualize the datapoints in subgroups.



*Fig 5: Visualization of the datapoints in its subgroups*

Subgroup labels and grid id are extracted and converted to geojson format with color properties added for each subgroup. This geojson is then displayed on the map. It appears that the subgroups created by the volumes of telecommunication activities is closely related to the population distribution of the region.



*Fig 6: Visualization of 10000 grids clustered into 6 subgroups*

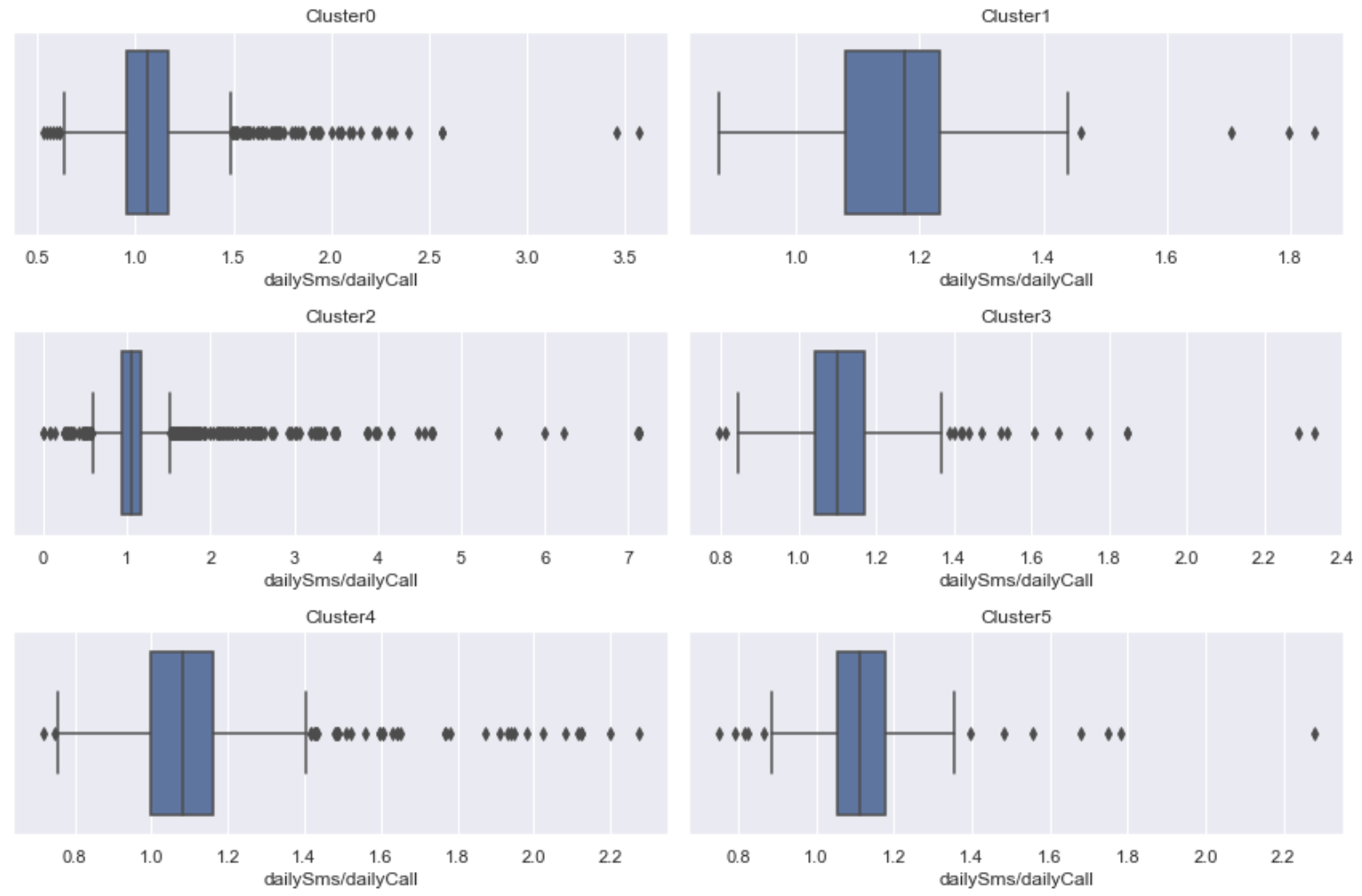


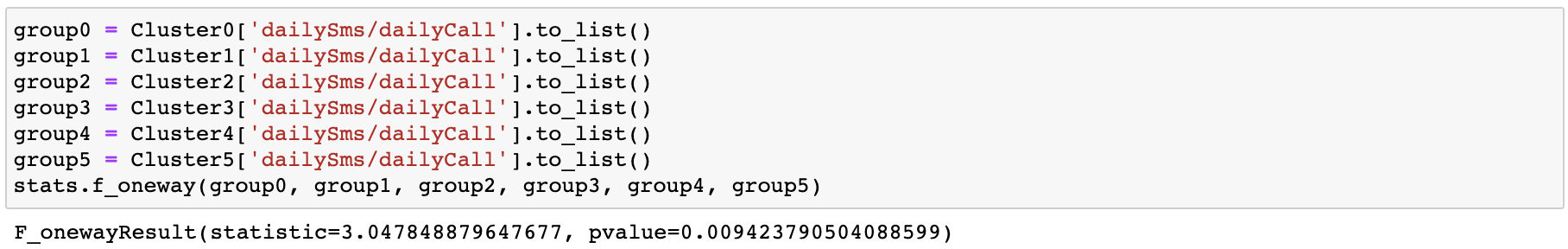
*Fig 7: Largest municipalities by population of Milan [sourced from Wikipedia]*

## **ANALYSIS OF THE CLUSTERS**

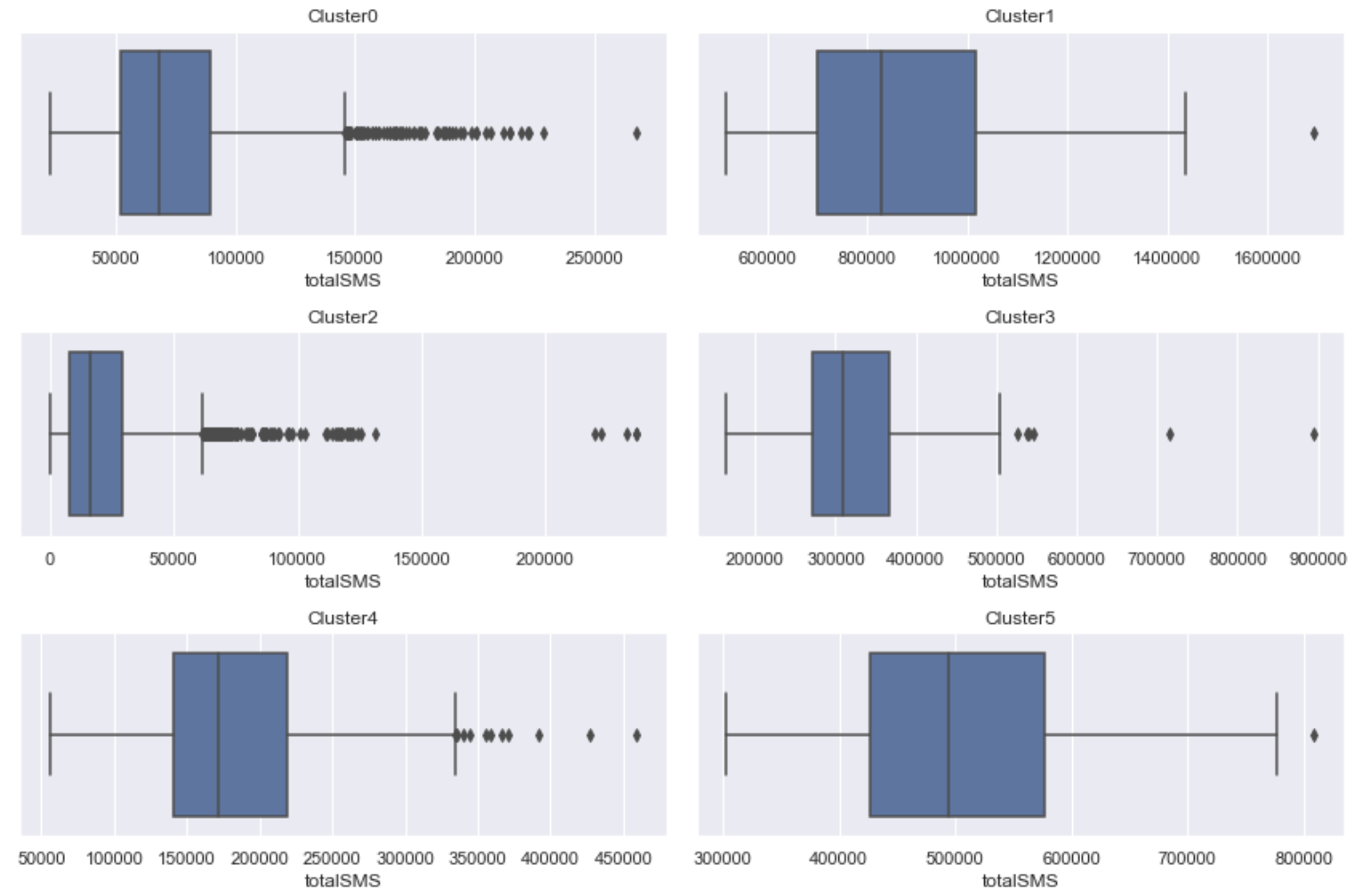
Clusters are plotted against different dimensions. Subgroups from the clustering model shows defined volume ranges for each dimension. Presence of outliers indicates wrong assignment of the data points to a subgroup with respect to that dimension alone. In general, this clustering model has done well for Internet volumes. In case of SMS & Calls, Cluster1, Cluster5 & Cluster3 are clearly formed with few outliers only.

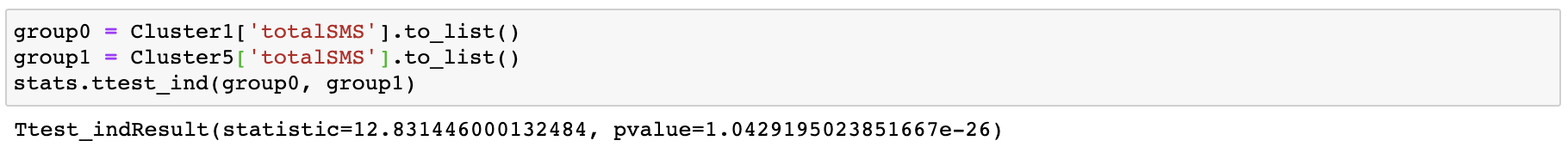
T-tests and one-way ANOVA tests can be performed on the subgroups for different dimensions in order to understand the similarities and dissimilarities in their behavioral patterns.

****

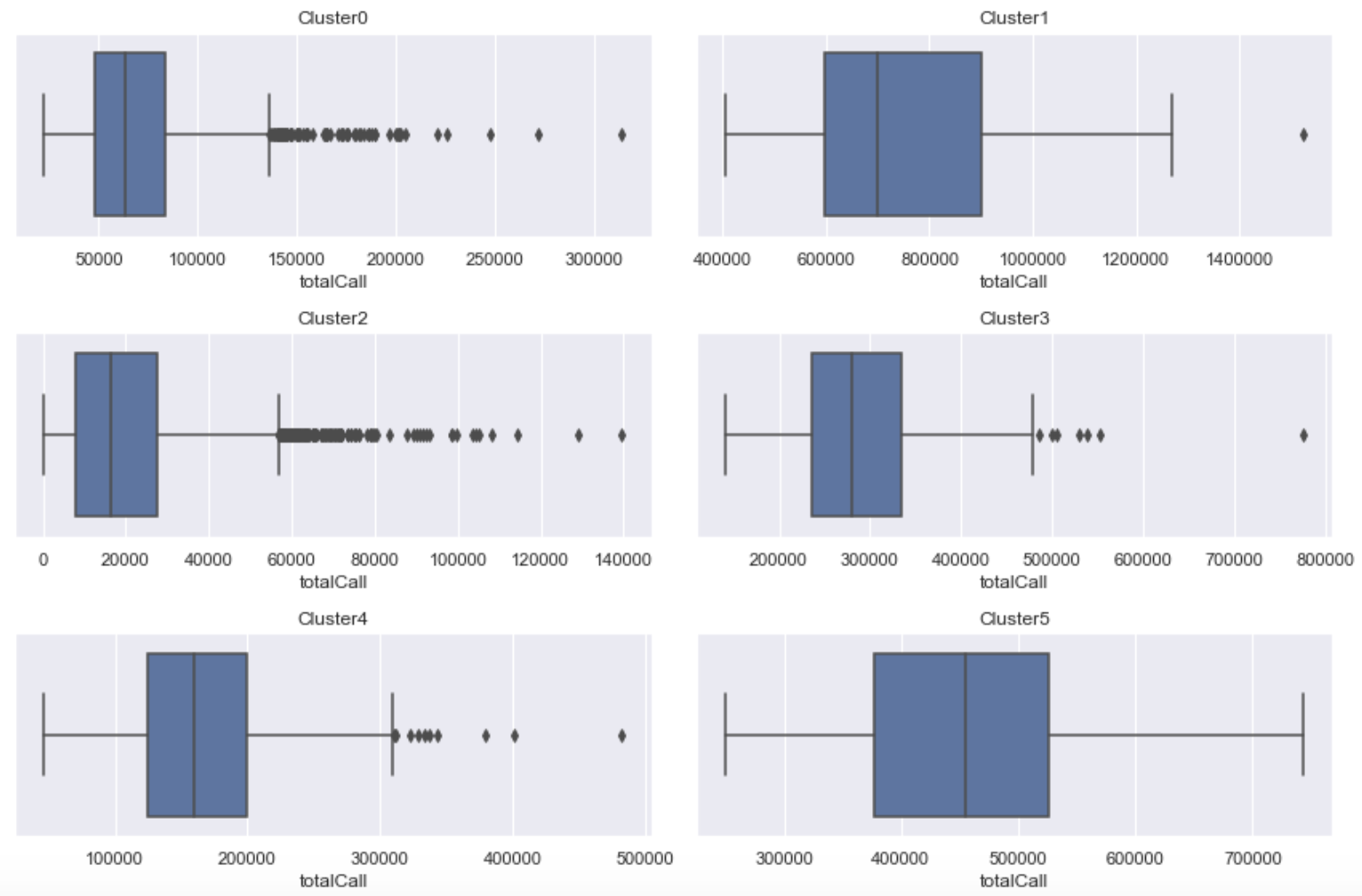
****

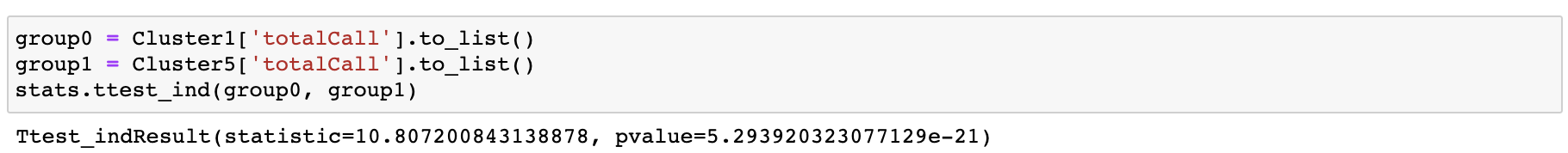
*Fig 8: Box-plot of 6 subgroups and its dailySms/dailyCall feature. One-Way ANOVA test shows approximately same mean value for all the groups*

****

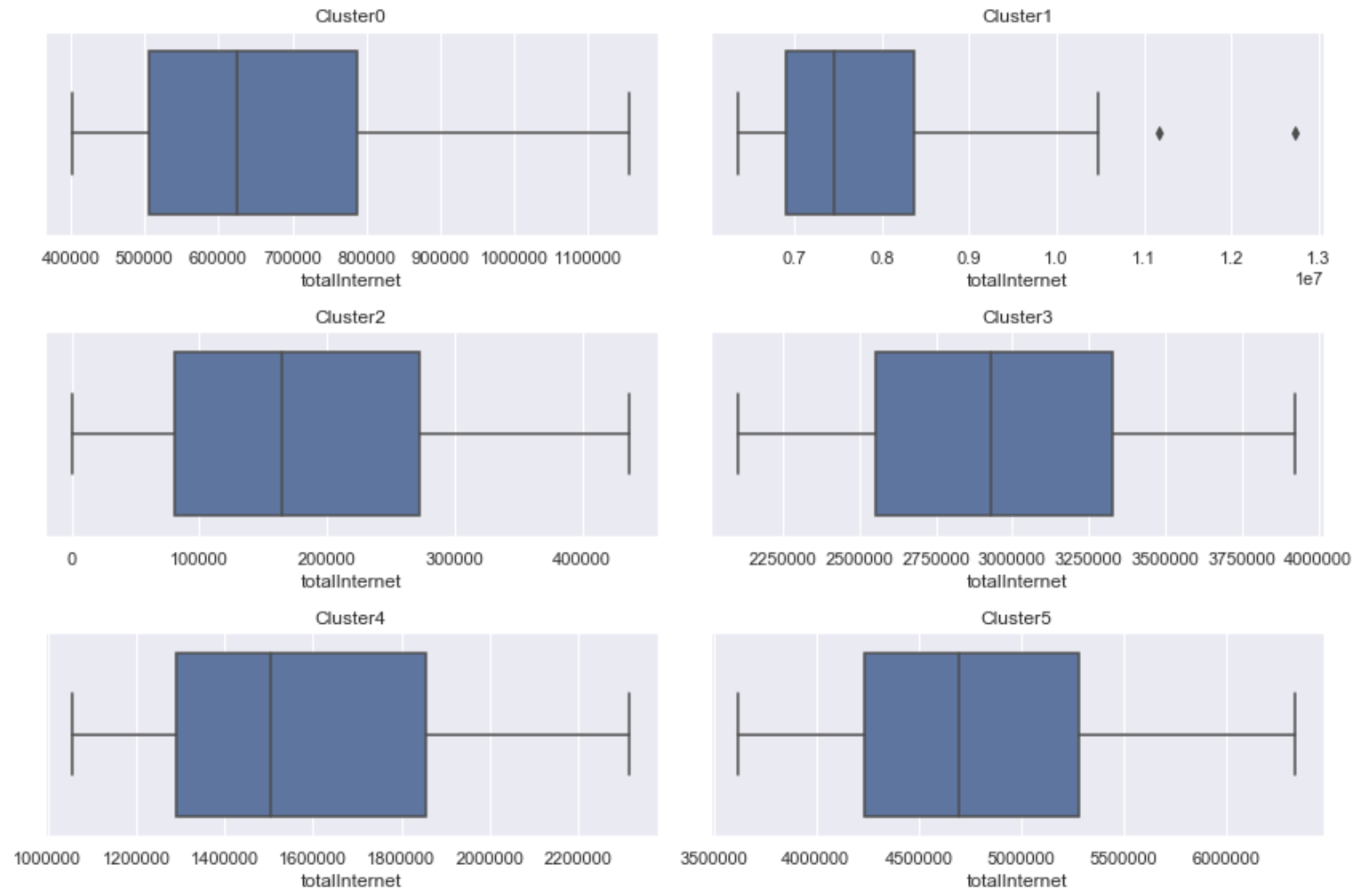
****

*Fig 9: Box-plot of 6 subgroups and its totalSMS feature. T-test shows approximately same mean value for subgroup 1&5*

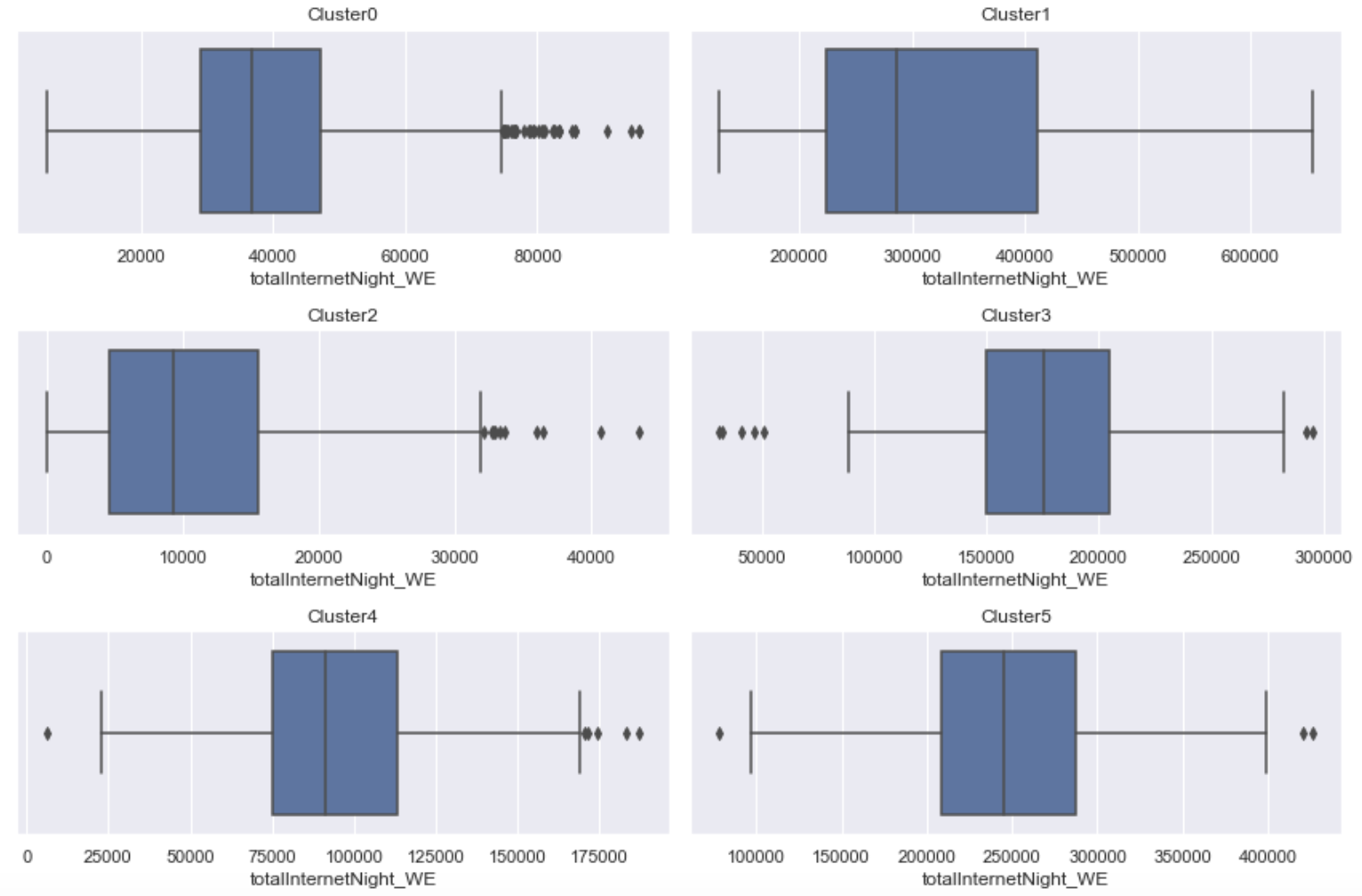
****

****

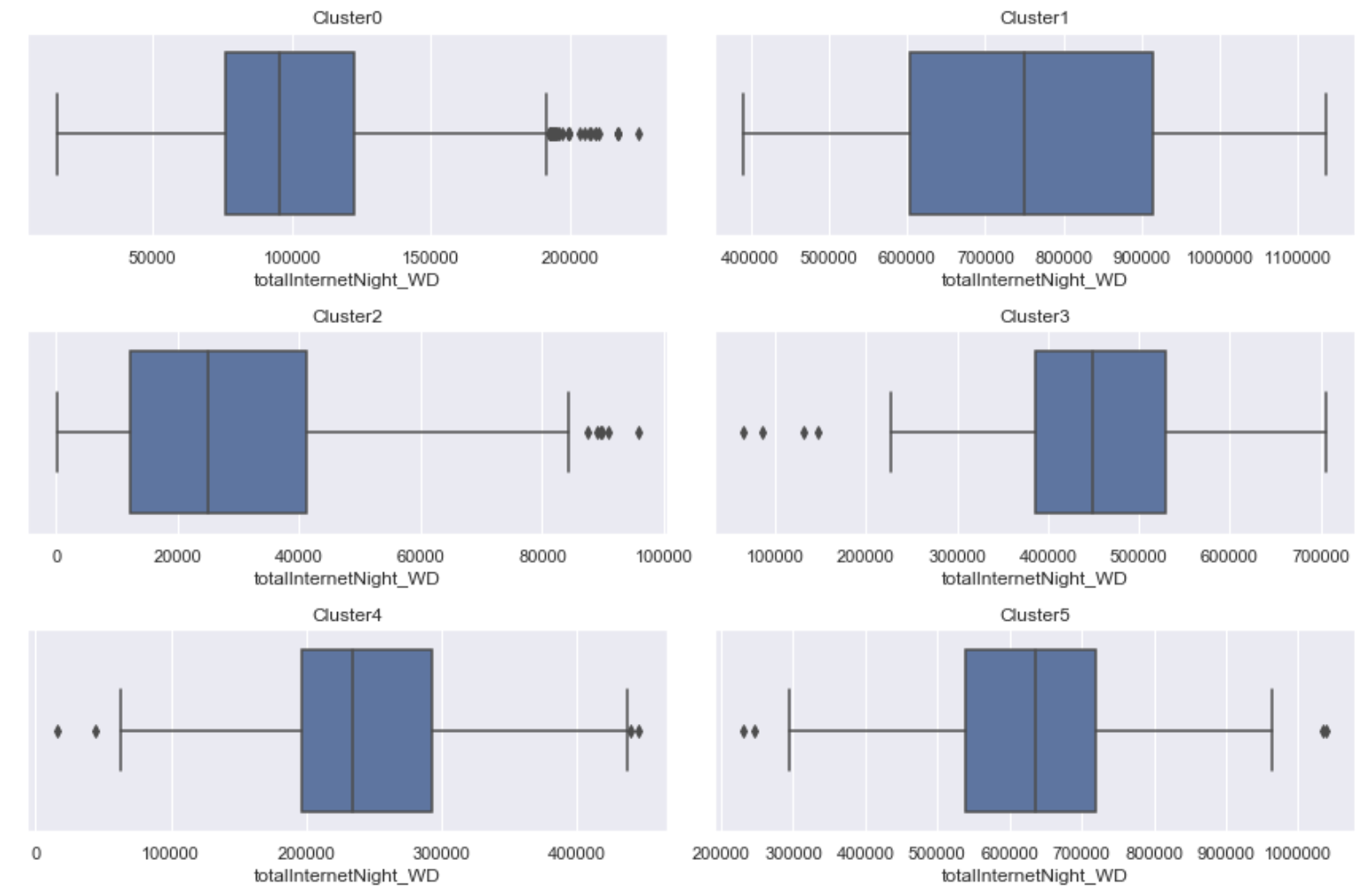
*Fig 10: Box-plot of 6 subgroups and its totalCall feature. T-test shows approximately same mean value for subgroup 1&5*

****

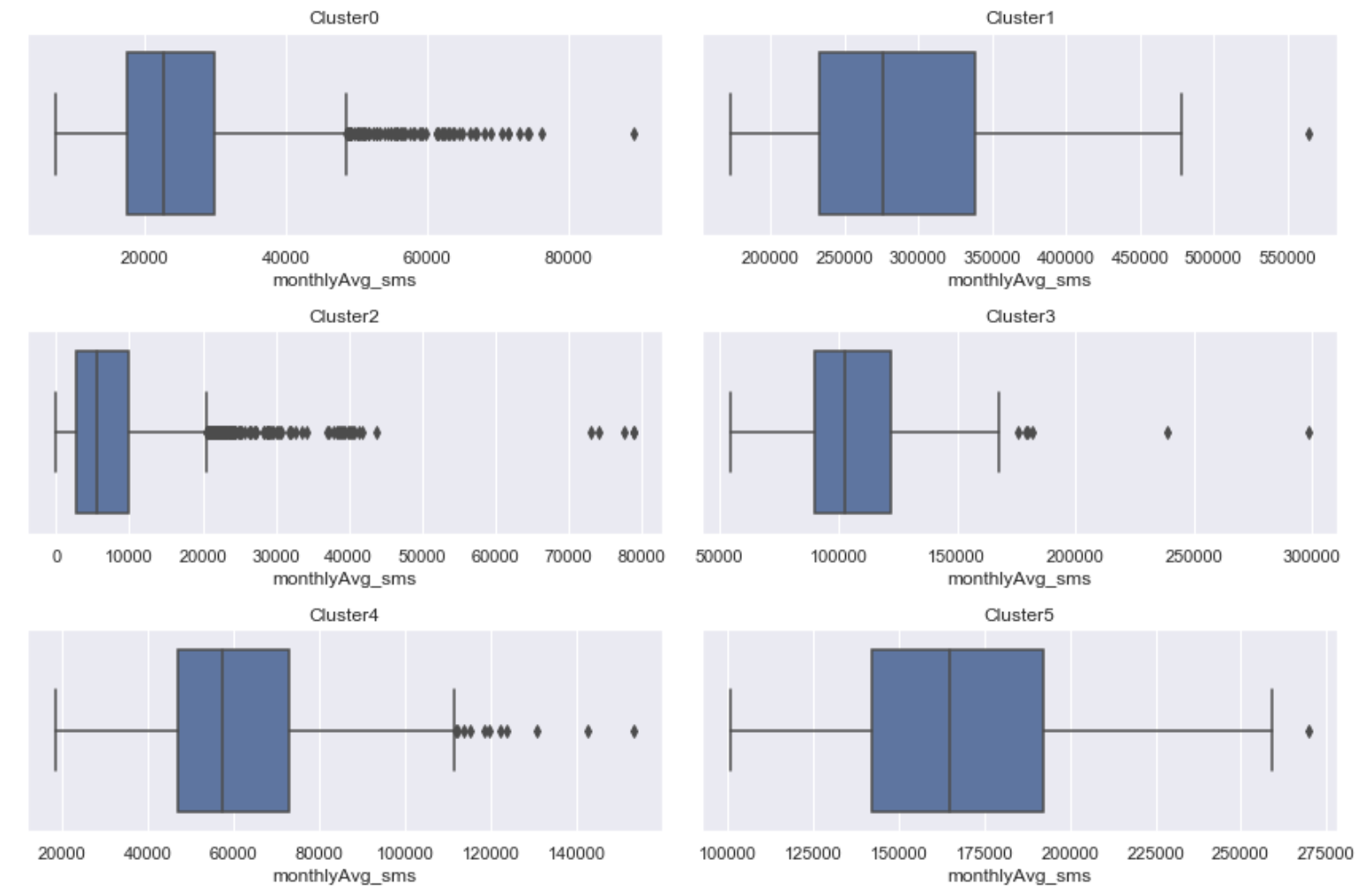
*Fig 11: Box-plot of 6 subgroups and its totalInternet feature*

****

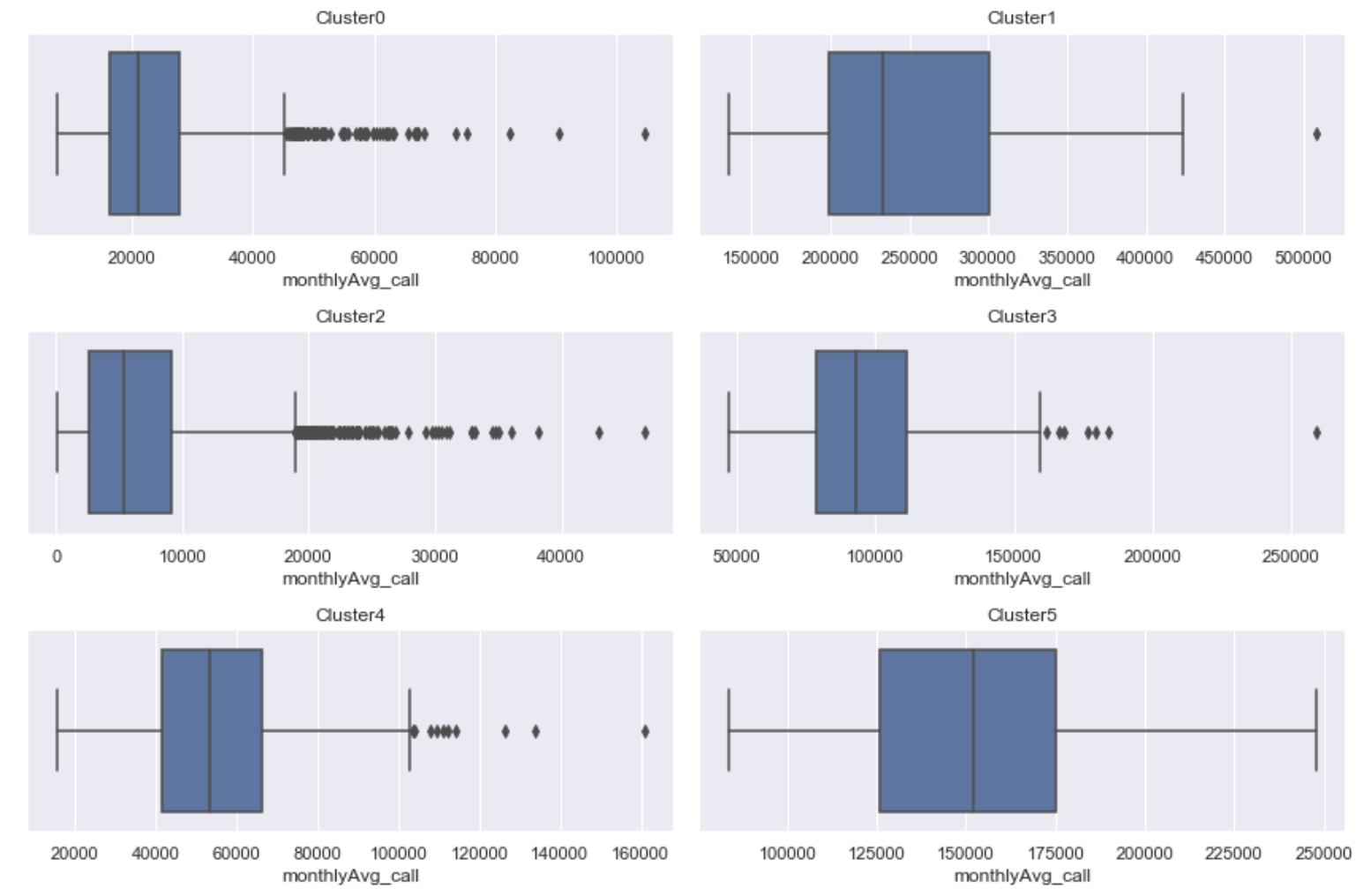
*Fig 12: Box-plot of 6 subgroups and its totalInternetNight\_WE feature*

****

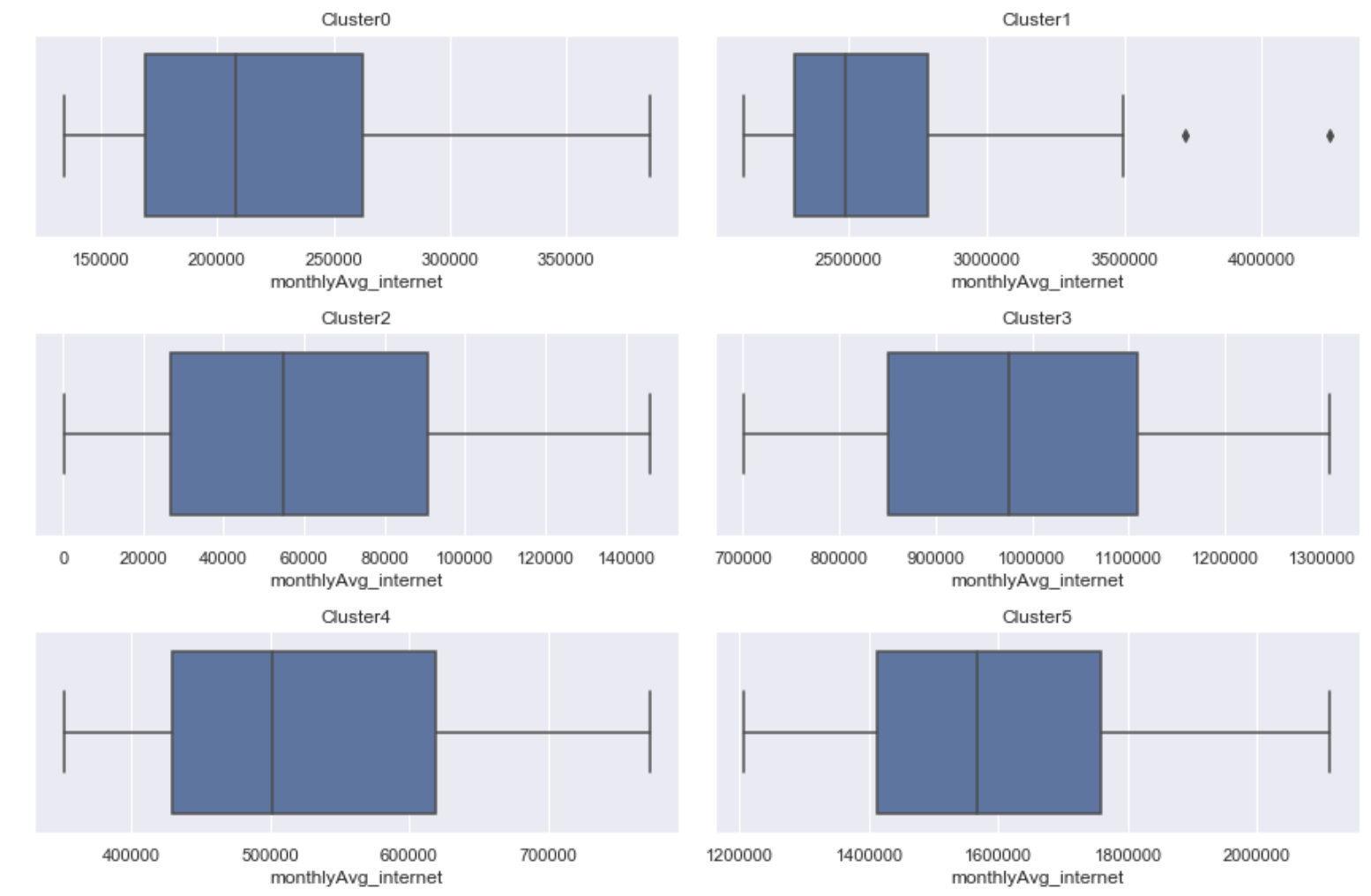
*Fig 13: Box-plot of 6 subgroups and its totalInternetNight\_WD feature*

****

*Fig 14: Box-plot of 6 subgroups and its monthlyAvg\_sms feature*

****

*Fig 15: Box-plot of 6 subgroups and its monthlyAvg\_call feature*

****

*Fig 16: Box-plot of 6 subgroups and its monthlyAvg\_internet feature*