CIRRELT

Centre interuniversitaire de recherche
sur les réseaux d'entreprise, la logistique et le transport

Interuniversity Research Centre
on Enterprise Networks, Logistics and Transportation

A Typology of Carsharing Customers in Montreal Based on Large-Scale Behavioural Dataset

Hanieh Baradaran Kashani
Martin Trépanier

March 2018

CIRRELT-2018-16

Bureaux de Montréal :
Université de Montréal
Pavillon André-Aisenstadt
C.P. 6128, succursale Centre-ville
Montréal (Québec)
Canada H3C 3J7
Téléphone : 514 343-7575
Télécopie : 514 343-7121

Bureaux de Québec :
Université Laval
Pavillon Palasis-Prince
2325, de la Terrasse, bureau 2642
Québec (Québec)
Canada G1V 0A6
Téléphone : 418 656-2073
Télécopie : 418 656-2624

www.cirrelt.ca

A Typology of Carsharing Customers in Montreal Based on Large-Scale Behavioural Dataset

Hanieh Baradaran Kashani, Martin Trépanier*

Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT) and Department of Mathematics and Industrial Engineering, Polytechnique Montréal, P.O. Box 6079, Station Centre-ville, Montréal, Canada H3C 3A7

Abstract. The emerging modes of transportation in the recent years, are a pointer to the fact that owning a private car is not all we need for the sake of traveling. Confirmed by the researchers, many features might be the reasons of choosing a mode of transportation. The cost, parking issues and environmental effects are some of the major features. Carsharing or in general vehicle sharing is one of these recent modes whose mission is providing the population with a travelling facility that is cheaper, easier and more environmentally friendly. Communauto, founded in 1994 in Quebec City, Canada, is now one of the major carsharing companies in Montreal, Canada. Using k-means clustering and Principal Component Analysis, the purpose of this paper is to study the behaviour of the Communauto regular-service carsharing users in Montreal over a year (2014) and find the usage patterns in each cluster. Also, by having the Communauto customer features available, the characteristics of the customers in each cluster will be defined. While previously several works in this field are performed utilising k-means clustering on the original data, our methodology emphasizes on implementing Principal Component Analysis (PCA) before k-means. Although, it always depends on the data characteristics, for the Communauto available data the assumption of k-means, i.e. the distribution of the data is spherical, could not be met without PCA transformation. However, without considering this assumption the accuracy of clustering results on the original data was always less than on the PCA transformed data. The k-means clustering results show that the Communauto regular-service carsharing users, are divided into nine different clusters. Some of the clusters patterns are similar during the weeks and some are similar during a year, but none of them are similar during week and year. So, each cluster has a unique usage pattern over the weeks and the year. Furthermore, the resulting clusters are ordered from high intensive users to the occasional ones based on the frequency of carsharing usage over one year.

Keywords: Carsharing, customer segmentation, principal component analysis.

Acknowledgements. This work was possible thanks to the support and collaboration of Communauto. The research project is also funded by the Natural Sciences and Engineering Research Council of Canada (NSERC RDCPJ # 474642-14).

---

* Corresponding author: Martin.Trepanier@cirrelt.ca

Dépôt légal – Bibliothèque et Archives nationales du Québec
Bibliothèque et Archives Canada, 2018

1 INTRODUCTION

Carsharing (Autopartage in French) sits within the emerging class of 'mobility services' that draw on modern technology to enable access to car-based mobility without the consumer owning the physical asset (a car) (Le Vine, Zolfaghari et al. 2014). As described by the carsharing association in USA, the mission of carsharing is to reduce car ownership, provide easy access to automobiles for publics, reduce vehicle distant travelled and so on. Easy access to carsharing services has a wide sense with many aspects such as affordability, little or no paper work, 24/7 access.

An important advantage of carsharing could be the fact that the users enjoy this mode of transportation like a private one, without having to take the responsibilities of owning a car. This might be one of the main reasons that people are more and more attracted to this mode of transportation. Consequently, it is more important to study about carsharing users' behaviours, to better understand their needs as well as carsharing role in the urban transportation system. This way the carsharing enterprises would find the best ways of improving their services to their customers.

Communauto, a carsharing company based in Montreal, Quebec, Canada, also operates in three other cities in Quebec: Quebec City, Gatineau and Sherbrook. This company which is founded in Quebec City in 1994, by Benoît Robert, offers two types of services: 1) Regular service, which offers station-based vehicles that should be reserved up to a month in advance and be brought back to the same station it is picked up. 2) Free-floating which offers auto-mobile vehicles, that needs no reservation and can be released anywhere in the service area.

By means of K-means clustering, the aim of this paper is to explore the Communauto regular service reservation dataset of the year 2014, to uncover the usage patterns of the customers. Because of the nature of the data and the k-means assumptions, Principal Component Analysis or PCA is applied on the data before k-means. Therefore, the k-means clustering is implemented on the principal components. Since k-means is an unsupervised learning which attempts to cluster the observations, the original data can adopt the resulting clusters from PCA, so that the results would be interpretable.

In this paper, we will first review some of the related studies on carsharing systems, the customers' usage behaviours as well as the studies which relate PCA to k-means clustering. Next, we will focus on the methodology of this study by introducing the datasets and the adopted statistical methods, the results of this study will come afterwards and the paper concludes with section five which is devoted to the discussion part of the study.

2 RELATED WORKS

In this section, some studies on carsharing systems are reviewed in three parts. The first part, reviews the studies on carsharing system in general, while the second part provides a brief look on more specific studies about carsharing users' behaviours using statistical methods such as k-means clustering. The third part reviews the studies on k-means clustering methods.

2.1 Carsharing in general

Even until the late 20's, no one could imagine sharing a vehicle could be happening among all the fellow-citizens. Nowadays, this has been turned to a popular mode of transportation. People now support such a mode of transportation that takes away the concerns of car ownership and at the same time provides them with the comfort of driving a private car that is also environmentally

friendly. According to the carsharing missions, various studies have been conducted that can show to what extent these missions have been achieved so far.

As the car ownership reduction is one of the missions of the carsharing companies, a very recent study in London, UK, (Le Vine and Polak 2017) established the early stage impact of free-floating carsharing on private car ownership. The results of this study showed, 37% of the users revealed that free-floating carsharing impacted their ownership of private vehicle. This study also exposed that the frequent service-users, had a higher level of education and income than the average of the population.

A study in Canada, (Klincevicius, Morency et al. 2014) assessed the reduction of car ownership in an area in Montreal served by Communauto regular-service carsharing. Using the historical data from the population and the users' behaviours, the authors examined the relation between the people car ownership and exposure to carsharing. The results by a linear regression model showed that car ownership had a reverse correlation with the number of carsharing vehicles in a 500-m radius of the local households.

Carsharing aims to reduce the greenhouse gas by decreasing the number of vehicles in use in the cities. A survey in Montreal, Canada, examined the contribution of Communauto carsharing, in reducing greenhouse gas emission. They examined the quantity of CO_2, as the main source of GHS, emitted by Communauto vehicles. Considering the usage habits of carsharing users, their results confirmed that each carsharing vehicle replaces ten to fourteen private cars (). Therefore, each carsharing user produces 1160 kg of CO_2 less than when they were not subscribed to this service. Also a case study, (Sioui, Morency et al. 2013) based on two comparative surveys showed that carsharing members did not reached the level of car use of typical residents owing one or more car. This is another indication that the carsharing users' contribution in greenhouse gas is less than other people owning a car.

2.2 Vehicle-sharing users' behaviour studies

Getting to know the customers' behaviours and requirements is essential as it helps the companies to better improve their services and adopt their strategies to the customers' needs. A study in California US, (Shaheen and Cohen 2008) based on 33 carsharing international surveys, claimed that cost saving, convenient locations and the guaranteed parking are the main motivations of using the carsharing services worldwide.

The customers' behaviour and their frequency of usage have been attractive to the researchers and the carsharing companies. A study identified typical patterns of carsharing use, by k-means clustering (Morency, Trépanier et al. 2007). Eight clusters were found, in each of which the users had some favourite weekdays of using the carsharing system. They also classified the carsharing users from different aspects. For instance, based on frequency of use, the users were classified to the occasional and frequent users, and from the aspect of trip length to the short-distance users and long-run users.

Like carsharing, the usage behaviour of bikesharing users have also been studied worldwide. A research in Lyon, France, (Vogel, Hamon et al. 2014) was carried out based on a large-scale behavioural dataset of bicycle sharing users. Exploiting cluster analysis, they produced user typology based on annual weekly, monthly and daily patterns. They found nine cluster of users with a unique profile for each of them, using the characteristics of the customers.

Another study on Montreal's Bixi bikesharing members (Morency, Trepanier et al. 2017), indicated that people living near carsharing system have a higher possibility of being a recurrent user. They also showed that the weather conditions influence the users' behaviours in using the bikesharing system. Some other studies are conducted to compare the usage behaviours in

carsharing and bikesharing systems. The results of a study in Montreal, Canada (Wielinski, Trépanier et al. 2017), confirmed that the bikesharing users are mostly men and younger and have higher income, whereas most of the carsharing users have more children and fewer cars. Using a multinomial logit model, they found that the carsharing-only users have the lowest income among the others and tend to use public transport systems more than the others, while the bikesharing-only users have the highest income and tend to use private cars more. They also identified two-system users whose income and transportation usage behaviour are in-between these two groups.

2.3 K-means clustering and PCA

Cluster or segmentation analysis is a kind of exploratory analysis that seeks to find some structure in the data. It divides the data points into the clusters in which they are similar, but dissimilar with the other data points in the other clusters. Because of its nature of exploratory, it has a wide application in different fields of studies.

PCA or Principal Component Analysis, sometimes is referred by factor analysis, is a statistical procedure that shrinks the high dimensional data to a lower dimension by maintaining most of the information of the data. In many applications, the data analysts prefer to lower the dimension of data by using a shrinkage method like PCA, specially for k-means clustering. The main reasons to do so will be discussed more in detail in the methodology section.

One of the issues that is sometimes discussed about PCA is that, applying other statistical techniques on the principal components instead of the original variables, cause the loss of interpretability of the results. However, this is not always true and depends on the methods that we desire to apply on the components. A research in Atlanta, USA, (Liang, Balcan et al. 2013) introduced a distributed PCA algorithm, and theoretically proved that any good approximation solution on the projected data by distributed PCA for k-means clustering, would be also a good approximation on the original data.

Some studies in this field are devoted to the connections of PCA and k-means clustering. A study in California, (Ding and He 2004) indicated that PCA as an unsupervised dimension reduction is very close to k-means clustering as an unsupervised learning. They showed that principal components are relaxed (without constraint) solutions of the cluster indicator vector in k-means clustering. They also proved that the cluster centroid subspace is spanned by the first *k*-1 principal components directions.

Since one of the problems in k-means is that a change in the initial values of the centroids changes the results of the clustering, some articles attempt to solve this initialization problem. For instance, a study in Boston, (Su and Dy 2004) focused on this and suggested that by utilising the eigenvectors of the covariance matrix in Principal Component Analysis as the k-means' centroid initialization, the clustering results produce smaller Sum of Squared Errors (SSE), also they showed that k-means converges faster by this approach. However, at the end they confirmed that, because of some limitations, this initialization approach might sometimes fail.

For the problem of initialization, a study whose results are very popular now, (Arthur and Vassilvitskii 2007) introduced a method of initialization named "k-means++". In this study, the authors propose a method by which the initial centroids are first chosen randomly, but to choose the next centroids, the data points are weighed according to their squared distance from the closest previously chosen centroid. Finally, they empirically show that k-means++ initialization, is very often faster and gives more accurate results than the random initialization. Similarly, in this study we utilize k-means++ initialization for k-means.

3 METHODOLOGY

In this section, the aim is to describe the adopted methods for this study. As discussed, the objective is to cluster the customers based on the frequency of their usage (number of reservations), and in each cluster, find the usage patterns. Figure 3.1. gives a summary of the methods we went through to achieve this objective.

First, the relevant variables in Communauto reservation dataset are selected and the data is cleaned. Then, the vector of attributes is produced using the pivot tables counting the number of reservations daily, weekly and monthly. Afterwards, the k-means assumptions and PCA prerequisites are verified, so that at the next step the data would be pre-processed according to them. When the data is appropriately pre-processed, PCA projects the data into a smaller subspace. Next, K-means clusters the new projected data and consequently, the usage patterns in each cluster are defined.

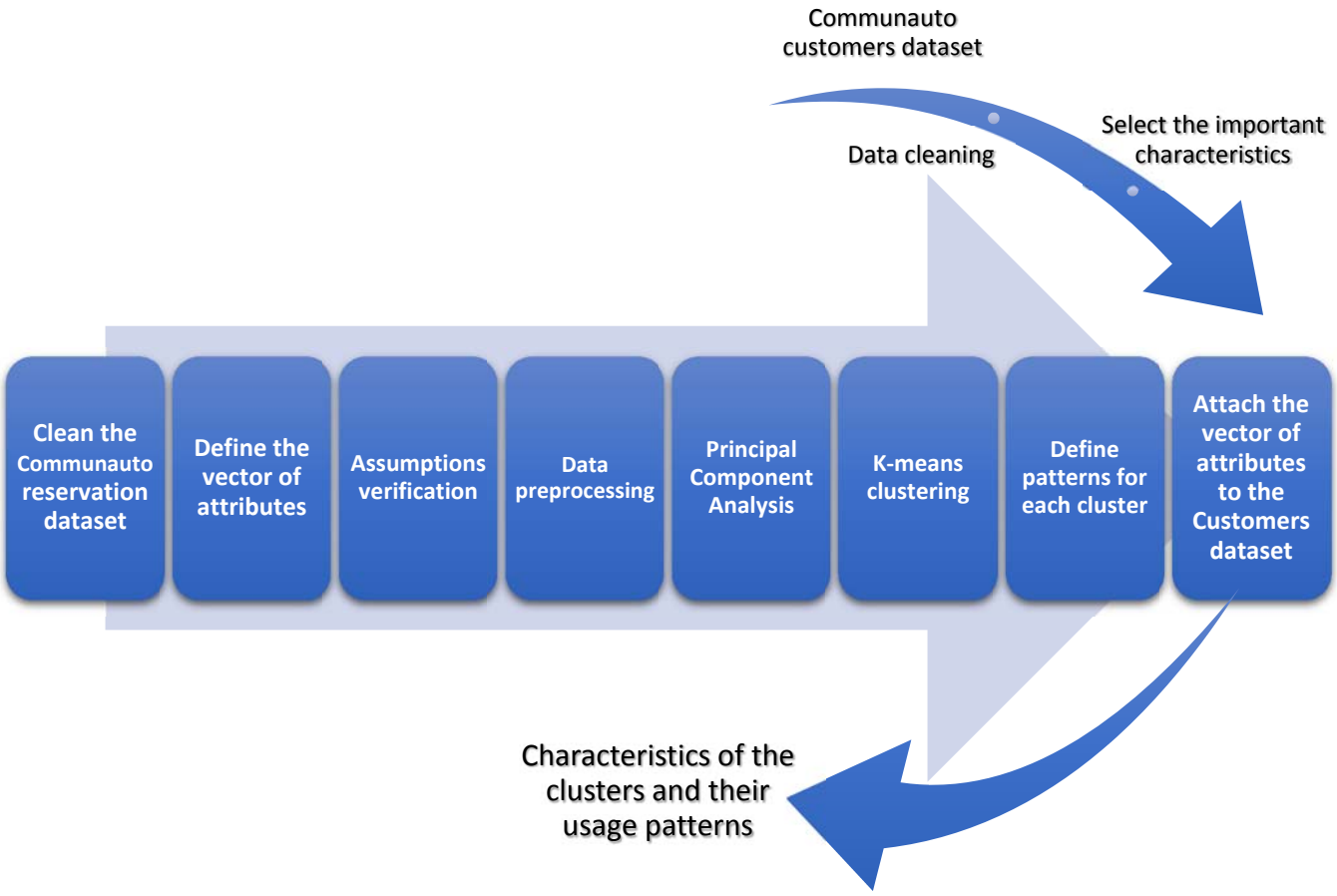

Figure 3-1: *Diagram of the study methodology*

At the next step, the relevant variables of the Communauto customers' dataset are selected and the data is cleaned. Then it is attached to the vector of attributes. Finally using the produced clusters and the customers' dataset, the characteristics of each carsharing usage pattern is defined. Each of these steps are explained in detail as follows.

3.1 Communauto Datasets

According to the two types of Communauto carsharing services, there are also two types of datasets of the carsharing transactions, regular (station-based) and Auto-mobile (free-floating), each of which contains its features related to the type of service. The available data in this study are the transactions happened only in the year 2014, and we chose to work only on the regular-service carsharing as for the other, Auto-mobile, the approach would be the same.

Because these datasets contain several columns that are not relevant to our objective, Table 3.1, describes the relevant variables.

Variables	Description	Values' Range
CustomerID	The identity number for the customers, which is unique for each customer	[4, 82636]
datDateDebutReservation	The date and time that the customer reserves the vehicle	2014-01-01 00:00:00 to 2014-12-31 23:45:00
intDebutKilometrage	The mileage of the car, at the time of reservation	[0, 234747]
intFinKilometrage	The mileage of the car, after the reservation is ended	[0, 234837]

Table 3-1: *Descriptions of the relevant variables in the regular-service reservation dataset*

In this dataset, each row belongs to each reservation, so we might encounter a customer ID repeating several times for several reservations during the year.

The reservation date helps to find out about the time of the carsharing usages. Which months have the most frequent usage? Are the customers using the carsharing system regularly or irregularly? This variable can disclose the usage pattern of the customers by answering to such questions.

The car's mileage, uncovers the reservations without usage, in which we are not interested and must be removed from the data.

Also, certain customers' specifications like gender, age, language, etc. were available for some of the customers. Tables 3.2, provides a summary of them. These specifications helped us at the end, to better describe each cluster and the usage patterns.

As described in the table 3.2, the two variables 'DateNaissance' (birthdate), and 'DateAbonnement' (Subscription date), contain some dates that are not logical.

For instance, there were 424 observations whose birthdate was 1900-01-01, which reminds us the Microsoft Excel default year. Besides, their corresponding subscription dates were all after 1994, which is not consistent with the birthdate. Some of the birthdates were happening in the future like 2032-01-29. On the other hand, there were some observations, for which the subscription date was 1901-01-01, i.e. before the foundation of Communauto of Montreal (1994).

Thus, such examples were replaced by 'Null' in the dataset. So, after these corrections, the two variables changed as described in the table 3.3 and the major errors were removed from these variables.

Variables	Description	Values' Range
CustomerID	The identity number for the customers, which is unique for each customer	[1, 92536]
IngSexe	Customers' gender	127: Female 128: Male
LanguageID	Customers' language	1: French 2: English
DateNaissance	Customers' birthdate	1900-01-01 to 2032-01-29
DateAbonnement	Customers' Subscription date	1901-01-01 00:00:00 To 2016-02-17 18:22:04

Table 3-2: *Descriptions of the relevant variables in the Customers' dataset*

Variables	Description	Values' Range
DateNaissance	Customers' birthdate	1914-04-10 to 1998-11-30
DateAbonnement	Customers' Subscription date	1994-01-08 00:00:00 To 2016-02-17 18:22:04

Table 3-3: *Descriptions of the birthdate and the Subscription date, after making corrections*

3.2 Vector of attributes

For our objective to be met, the number of reservations, per weekdays, weeks and months for each customer must be counted. To do so, the year (2014), month (1, …,12), weekday (Mon, …, Sun) and the number of the week in the year (1, …,53) are extracted from one variable, "reservation date". Next, a pivot table for each of them is made to count the number of reservations per customer. Therefore, in the new data frame the number of rows equals the number of

customers in regular service reservation dataset. In this section, the vector of attributes i.e. the variables of the new dataset are described:

- X_1, …, X_{12}: Average monthly use, i.e. the average number of trips per month.

Month	1	2	3	4	5	6	7	8	9	10	11	12	Sum
CustomerID													
4.0	0.11	0.17	0.00	0.00	0.00	0.06	0.06	0.17	0.06	0.06	0.11	0.22	1.0
5.0	0.03	0.03	0.03	0.05	0.08	0.08	0.21	0.13	0.08	0.10	0.08	0.13	1.0
6.0	0.00	0.00	0.00	0.00	0.11	0.00	0.11	0.33	0.00	0.00	0.33	0.11	1.0
14.0	0.15	0.10	0.00	0.20	0.10	0.00	0.05	0.15	0.00	0.05	0.00	0.20	1.0
31.0	0.07	0.09	0.07	0.14	0.16	0.02	0.07	0.02	0.09	0.07	0.12	0.07	1.0

Table 3-4 : *X_1, …, X_{12}*

- X_{13}, …, X_{19}: Average daily use, i.e. the average number of trips per days of the week, Monday to Sunday.

WeekDay	Fri	Mon	Sat	Sun	Thu	Tue	Wed	Sum
CustomerID								
4.0	0.22	0.17	0.06	0.06	0.11	0.22	0.17	1.0
5.0	0.08	0.13	0.21	0.18	0.10	0.10	0.21	1.0
6.0	0.11	0.11	0.00	0.00	0.33	0.11	0.33	1.0
14.0	0.20	0.00	0.25	0.25	0.05	0.10	0.15	1.0
31.0	0.07	0.12	0.09	0.02	0.35	0.26	0.09	1.0

Table 3-5 : *X_{13}, …, X_{19}*

- X_{20}: Averaged weekly use, i.e. the average number of trips per weeks of the year 2014, calculated over all the weeks during which user travels at least once. Divided by seven to make it consistent with the previous attributes.

WeekNumber	00	01	02	03	04	05	06	07	08	09	...	44	45	46	47	48	49	50	51	52	Averaged_weekly_NonZero
CustomerID																					
4.0	0	0	1	0	1	2	0	0	1	0	...	1	0	1	0	1	1	2	0	0	0.160714
5.0	0	0	1	0	0	0	0	1	0	0	...	0	1	1	1	0	0	1	3	1	0.192118
6.0	0	0	0	0	0	0	0	0	0	0	...	0	0	3	0	0	1	0	0	0	0.214286
14.0	0	0	1	1	1	0	1	1	0	0	...	0	0	0	0	0	0	2	1	1	0.178571
31.0	0	2	0	1	0	0	1	1	3	1	...	3	0	2	0	0	1	1	1	0	0.191964

Table 3-6 : *X_{20} is the rightest column*

▪ X_{21}: Normalized total trips, i.e. Total number of trips made over the year 2014, summed over all weeks, normalized dividing by 1.5 times its interquartile range of the distribution for all users. The interquartile range is used as a robust measure of scale. That is, it is an alternative to the standard deviation and it is less effected by extremes than the standard deviation.

WeekNumber	00	01	02	03	04	05	06	07	08	09	...	45	46	47	48	49	50	51	52	TotalTrips	Normalized-TotalTrips
CustomerID																					
4.0	0	0	1	0	1	2	0	0	1	0	...	0	1	0	1	1	2	0	0	18	0.545455
5.0	0	0	1	0	0	0	0	1	0	0	...	1	1	1	0	0	1	3	1	39	1.181818
6.0	0	0	0	0	0	0	0	0	0	0	...	0	3	0	0	1	0	0	0	9	0.272727
14.0	0	0	1	1	1	0	1	1	0	0	...	0	0	0	0	0	2	1	1	20	0.606061
31.0	0	2	0	1	0	0	1	1	3	1	...	0	2	0	0	1	1	1	0	43	1.303030

Table 3-7 : *X_{21} is the rightest column*

This way, out of only one variable: “Reservation date”, 21 variables are created. Then they are attached to build the new dataset containing these 21 variables as the columns and the customer ID’s as the indexes.

Now that the data is ready, they must be verified if they need to be pre-processed before any analysis. For this reason, the assumptions of k-means clustering must be verified.

3.3 Assumptions verification

Assumptions play an important role in the statistical methods. Without verifying the assumptions, any result from the analysis would not be reliable.

K-means clustering assumes that the distribution of the data is spherical. This means when looking at the scatter plot we should not observe any ellipse suggesting some correlation between two variables. In other words, sphericity means the variables are uncorrelated (covariance = 0), and they all have an equal variance of one, which means the covariance matrix is equal to the identity matrix. Bartlett Sphericity test, also suggests the same approach for verifying the sphericity:

H_0: The correlation matrix of the data is equal to the identity matrix

H_1: The correlation matrix of the data is different from the identity matrix

If we reject the null hypothesis (P-Value<0.05), then the correlation matrix is different from the identity matrix and the data is not spherical.

The above-mentioned test is highly sensitive to the number of samples, n, i.e. if n is very large, it rejects the null hypothesis even if the correlations are very close to zero, but not zero. This test also assumes that the multivariate distribution of the data is normal. (Sarmento and Costa 2017)

Despite these restrictions, we considered this test to verify the sphericity of the data, we also kept observing the scatter plot of the variables at every step of the pre-processing the data.

On the other hand, Principal Component Analysis or PCA which will be described in the section 3.5, is sensitive to the noises. Besides, it needs the variables to follow the same measurements. The data pre-processing section will go through the pre-processing steps to make the data ready for the analysis.

3.4 Data pre-processing

According to the previous section, the data should be verified if the assumptions are already met. To verify this, figure 3.2 illustrates how the pairwise scatter plot of our data looks like. Also, the diagonal line plot shows how each variable is distributed. According to this figure there are big outliers in the data.

The outliers cause many deficiencies in the data. They are one of the reasons that the data distribution is not spherical and it is extremely skewed. Also, big outliers distract k-means clustering in finding the appropriate centroids.

Thus, the first step in pre-processing the data would be the best to remove particularly the big outliers to study them later. So, we need to keep them apart and verify the assumptions again in the rest of the data.

Figure 3.2. reveals that the distribution of the variables is strongly right-skewed. The scatter plot of the last two variables, i.e. X_{20}: Averaged weekly use and X_{21}: Normalized total trips, are not visible, and this is due to the very large outliers that exist in these two variables.

3.4.1 Outlier removal using Mahalanobis distance

An outlier, or noise, is an observation that is distant from other observations (Maddala, G. S. 1992). It may be due to variability in the measurement or an experimental error (Grubbs 1969). The outliers might also exist because some of the observations show different behaviours.

In our data, beside the fact that some of the customers showed to have a totally different usage pattern, the major cause of the outliers is that, the measuring of the last two variables is different from the other 19 variables. The other variables are the averages that vary between 0 and 1, but the last two are positive variables ranging in an interval of (0, ∞).

Table 3.8, which is the descriptive statistics of the variables X_{20}: Averaged-weekly, X_{21}: Normalized-TotalTrips, From the top, it shows the total number of observations, i.e. the customers, the average amount and the standard deviation for each of these attributes, the minimum, the 1st, 2nd and 3rd quantiles and finally the maximum value of each variable. It also highlights the 3rd quantiles and the maximum values for these two variables.

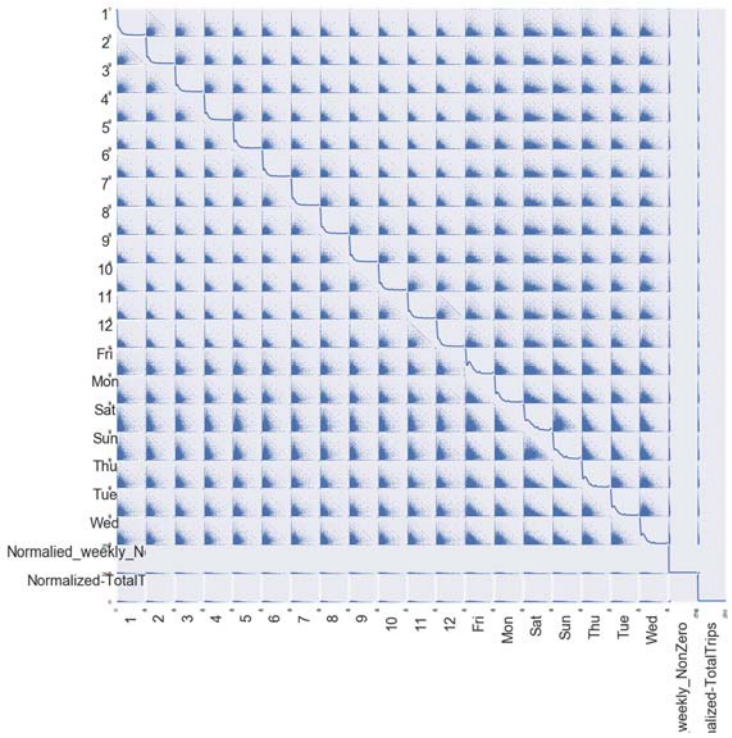

Figure 3-2 : *The pairwise scatter plots for the vector of variables created from Communauto reservation dataset (2014)*

	Averaged-weekly	Normalized-TotalTrips
count	28616	28616
mean	0.19	0.63
std	0.15	1.71
min	0.14	0.03
0.25	0.14	0.12
0.50	0.17	0.33
0.75	**0.20**	**0.79**
max	**21.92**	**246.45**

Table 3-8 : *Descriptive statistics of the variables X_{20} and X_{21}*

One could clearly observe in this table that there is a high difference between the third quantiles and the maximum values, which can be assumed as a confirmation for the existence of outliers in these two variables. However, when removing the outliers, all the attributes should be considered in a multivariate method.

Considering the two sources of outliers, 1) different usage patterns and 2) different measurements, there are different solutions.

The first source of the noises, can be resolved by keeping apart those customers whose behaviour is far from the others, next observe the rest of the data. The second source of the noises suggests that some transformation would help to smooth the noises. Thus first, to better observe the scatter plots, it would be best to temporarily remove some outliers, not to throw them out, but to better comprehend the rest of the data.

The outliers might be univariate or multivariate. The former can be treated separately for each variable. This happens when the variables vary independently. On the other hand, multivariate outliers are the ones which cannot be treated separately, to remove them one must consider the whole structure of the data and examine the relationship of the variables using the multivariate methods, such as Mahalanobis distance.

Using an estimate of the location of each observation, Mahalanobis distance, locates the data points, that are significantly distant from the rest of the data. (Franklin, Thomas et al. 2000) Here we present a simple definition of Mahalanobis distance:

Let $\vec{x} = \{x_1, x_2, \dots, x_n)^T$, represent a set of data points with the mean $\vec{\mu} = (\mu_1, \mu_2, \dots, \mu_n)^T$, and the covariance S. The Mahalanobis distance would be as follows: (De Maesschalck, Jouan-Rimbaud et al. 2000)

$$D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T S^{-1} (\vec{x} - \vec{\mu})} \quad (3.1)$$

3.4.2 Log transformation

Log transformation, replaces all the data points by $z_i = f(x_i)$, where f is a logarithm function usually with the base of 2, 10 or e=2.718. By maintaining the order of the data points, log transformation has the biggest effect on the largest values, i.e. it reduces the distances between the large numbers more. This causes that log transformation to play many roles in pre-processing the data:

1) making the highly-skewed data, less skewed and closer to the normal distribution;
2) smoothing the noises to some extents
3) because of the previous ones, it also helps on making the data more spherical

The skewness of our data is already shown in figure 3.2. Therefore, at the same time of resolving the skewness of the data, log transformation, smooths the noises and makes the distribution of the data more spherical. This helps the k-means assumptions and the PCA requirements to be met.

3.4.3 standardization

Standardization, is a process that transforms the mean and the standard deviation of the data to zero and one, respectively. This is always done by subtracting the mean from the data and dividing them by the standard deviation of the data:

$$Z_i = \frac{X_i - \mu}{\sigma} \quad (3.2)$$

Depending on the data, most of the times standardization is a crucial step prior to PCA. For PCA, it is very important to ensure that the variables follow the same measurements.

3.5 Principal Component Analysis (PCA)

Principal Component analysis or PCA is an orthogonal linear transformation of the data which projects the variables to a lower dimensional subspace by which the first new variable (or the first principal component) holds the highest variance of the data, the second variable holds the second highest variance and so on. (Jolliffe 2002) In other words the new coordinate system is a combination of all the original variables in a way that it can capture the maximum variance it can from the data. For example, the first principal component is a normalized combination of all the variables:

$$Z_1 = \varphi_{11}X_1 + \varphi_{12}X_2 + \cdots + \varphi_{P1}X_P \quad (3.3)$$

Normalized here means that $\sum_{j=1}^{P} \varphi_{j1}^2 = 1$, and $\varphi_{11}, \varphi_{12}, \ldots, \varphi_{P1}$ are called the loadings of the first principal components. (Friedman, Hastie et al. 2001)

Principal component analysis like many other statistical procedures has both its pros and cons. However, there are strong reasons to apply PCA on the data, before k-means clustering:

- As the dimensionality increases, the accuracy of k-means decreases. This is called the "curse of the dimensionality". PCA, projects all the variables of a dataset to a lower dimensional subspace.
- At the same time of reducing the dimension, PCA is building new variables which are not redundant or correlated.
- K-means is sensitive to the noises and outliers, PCA helps improving the clustering accuracy by smoothing the noises.

3.5.1 *K* selection in PCA

To implement the principal component analysis, the eigenvalues and eigenvectors should be calculated and to choose the top- *K* subspace, the value of *K* must be selected, depending on the variance of the data that we choose to keep.

Recall that PCA tries to minimize the average squared projection error:

$$\frac{1}{m}\sum_{i=1}^{m}\left|x^{(i)} - x_{approx}^{(i)}\right|^2 \quad (3.4)$$

Which means it tries to minimize the squared distance between *x* and its projection onto that lower-dimensional surface. The total variation in the data is given by:

$$\frac{1}{m}\sum_{i=1}^{m}\left|x^{(i)}\right|^2 \quad (3.5)$$

This second formulation defines how far the training examples are from the vector, i.e. from being all zeros. To choose a *K,* a common rule of thumb is to calculate the ratio between (3.4) and (3.5) to be less than a certain value, *c*:

$$\frac{\frac{1}{m}\sum_{i=1}^{m}\left|x^{(i)} - x_{approx}^{(i)}\right|^2}{\frac{1}{m}\sum_{i=1}^{m}\left|x^{(i)}\right|^2} < c \quad (3.6)$$

This ratio defines the amount that the data varies. For example, if the *c* value is 0.01, we retain 99% of the variance of the data. Depending on the data, the analyst might choose another percentage for the variance of the data to be retained. Therefore, the value of *K* is chosen in a way that the ratio (3.6) is satisfied.

A more efficient way to do this, is to calculate the singular value decomposition (*SVD*(Σ)=*USV**) of the covariance matrix. Considering the elements of the decomposed diagonal matrix *S*, the following equation, which is an equivalent to formula (3.6), can be calculated:

$$1 - \frac{\sum_{i=1}^{k} S_{ii}}{\sum_{i=1}^{m} S_{ii}} < c \qquad (3.7)$$

Where S_{ii} is the i'th diagonal element of the matrix *S*. Equation (3.7) ensures that (1- c) *%* of the variance of the data is retained. We used the latter equation to decide about the number of principal components, which will be discussed in results.

3.6 K-means

K-means clustering is a method for finding clusters and cluster centres in a set of unlabelled data. One chooses the desired number of cluster centres, say *K*, and the K-means procedure iteratively moves the centres to minimize the total within cluster variance. Given an initial set of centres, the K- means algorithm alternates the two steps:

- for each centre, we identify the subset of training points (its cluster) that is closer to it than any other centre;
- the means of each feature for the data points in each cluster are computed, and this mean vector becomes the new centre for that cluster.

These two steps are iterated until convergence.

The K-means algorithm is one of the most popular iterative descent clustering methods. It is intended for situations in which all variables are of the quantitative type, and squared Euclidean distance:

$$d(x_i, x_{i'}) = \sum_{j=1}^{P} \left(x_{ij} - x_{i'j}\right)^2 = \|x_i - x_{i'}\|^2 \qquad (3.8)$$

is chosen as the dissimilarity function (Friedman, Hastie et al. 2001). K-means method uses K prototypes, the centroids of clusters, to characterize the data. They are determined by minimizing the sum of squared errors:

$$J_k = \sum_{k=1}^{K} \sum_{i \in C_k} (x_i - m_k)^2 \qquad (3.9)$$

where $(x_1, \ldots, x_n) = X$ is the data matrix and $m_k = \sum_{i \in C_k} x_i / n_k$ is the centroid of the cluster C_k and n_k is the number of points in C_k. (Ding and He 2004)

3.6.1 K-means++ initialization

Among different methods of initialization, k-means++ proved to be one of the most popular, fast and accurate methods. To better describe this method, we provide its algorithm which is already presented by its authors in k-means++ paper. (Arthur and Vassilvitskii 2007)

K-means++ Algorithm:

Let *D(x)* be the shortest distance from a data point to the closet previously chosen centroid. Then k-means++ algorithm will be as follows:

1a. Take one centre c_1, chosen uniformly at random from X.

1b. Take a new centre c_i, choosing x $\in$ X with probability $\frac{D(x)^2}{\sum_{x \in X} D(x)^2}$.

1c. Repeat Step 1b. until we have taken k centres altogether.

2. Proceed as with the standard k-means algorithm.

3.6.2 *K* selection in K-Means

One of the draw-backs of k-means clustering is that, it requires a priori specification of the number of clusters, *k.* There are several methods to find the proper number of clusters. We have selected some of them to find *k,* at the same time verifying the accuracy of the clustering.

Silhouette refers to a method of interpretation and validation of consistency within clusters of data (Rousseeuw 1987). Let b(i) be the lowest average dissimilarity of *i* to any other cluster, of which *i* is not a member. The cluster with this lowest average dissimilarity is the "neighbouring cluster" of *i,* because it is the next best fit cluster for point *i.* We now define a silhouette:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \qquad (3.10)$$

From the above definition:

$$-1 \leq s(i) \leq 1 \qquad (3.11)$$

For *s(i)* to be close to 1 we require that *a(i)* ≤ *b(i).* As *a(i)* is a measure of how dissimilar *i* is to its own cluster, a small value means it is well matched. Furthermore, a large *b(i)* implies that *i* is badly matched to its neighbouring cluster. Thus, an *s(i)* close to one means that the data is appropriately clustered. (de Amorim and Hennig 2015)

Cross-validation is perhaps the simplest and the most commonly practiced model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. In *k*-fold cross-validation, the original sample is randomly split into *k* almost equal-sized subsets. One of the *k* subsets is kept as the validation set for testing the model that is already fit on the remaining *k* - 1 subsets as the training set.

Therefore, k-means is first fit on the *k* – 1 subsets, then the model is tested on the validation data and the cluster labels are predicted for them. This process is repeated *k = 1,2, …, K* times, in a way that each of the *k* subsets is used as the validation set only once. The average of the *k* estimations of prediction error, resulting from the *k* folds can be considered as the overall estimated error. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once.

4 RESULTS

The described methods were applied on the Communauto datasets. Therefore, the results are described here.

As discussed, the first step after building the vector of attributes is to pre-process the data, according to the data and the assumptions of the methods we are adopting. The pre-processing steps are already discussed, so here we only go through the results of applying those methods on the data.

4.1 Data pre-processing

As mentioned before, the data we are dealing with, contains very big outliers that we preferred to remove and keep them apart for further analysis. Also, we saw in the figure 3.2 that the data was strongly right-skewed and the scatter plots were not spherical. Thus, the following pre-processing steps we must take to meet the prerequisites of PCA and k-means:

1. Remove the outliers using a classic multivariate method named Mahalanobis distance
2. Log-transform the data to balance the right-skewness
3. Standardize the data to keep the measurement balanced for PCA

Here, we go through the results of the pre-processing the data at each step.

4.1.1 Outlier removal using Mahalanobis distance

We chose to remove only 0.05 percent of the data as the outliers, i.e. 0.005 of the data having the biggest Mahalanobis distance were removed, to be studied further. Figure 4.1, shows that the scatter plots of the two variables, X_{20} and X_{21}, are now visible, but still right-skewed. The strong correlation between these two variables is also visible after removing the outliers. As already discussed, as an assumption of k-means clustering, the distribution of the variables should be spherical. Therefore, this strong correlation is against this assumption.

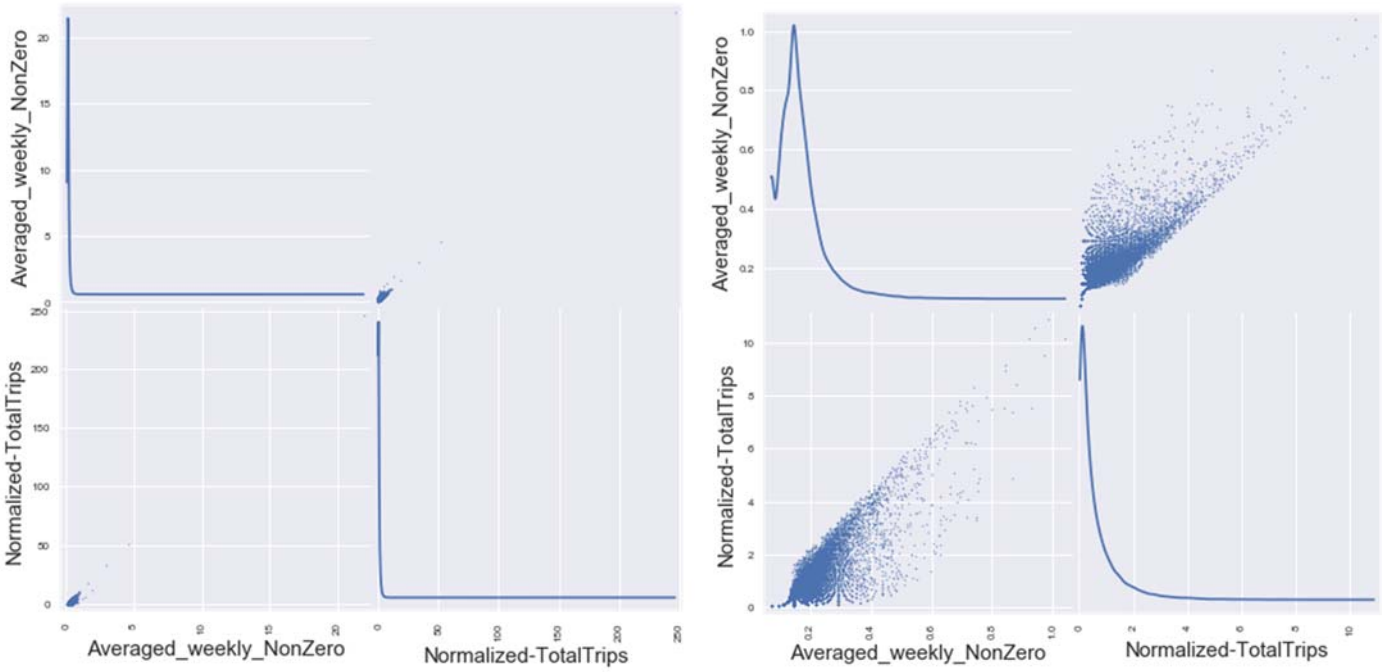

Figure 4-1 : *The two variables, X_{20} and X_{21}, before (up) and after (down) removing 0.05 percent of the data as the outliers*

4.1.2 Log-transformation

As the distribution of all the variables is right-skewed, log-transformations helps on resolving it. The default logarithm in python is the natural logarithm, which is suitable for our case. Logarithm with the base 10, is not suitable here because it makes the data too small.

Since the first 19 variables include zeros, it is impossible to log transform them directly. We solved the problem by adding a half of the non-zero minimum value of the whole dataset, to the data, and then applied the logarithm. Figure 4.2, illustrates the pairwise scatter plots of the variables after outlier removal and log-transformation. A comparison between figure 4.2 and figure 3.2, proves that the data is less skewed after this transformation.

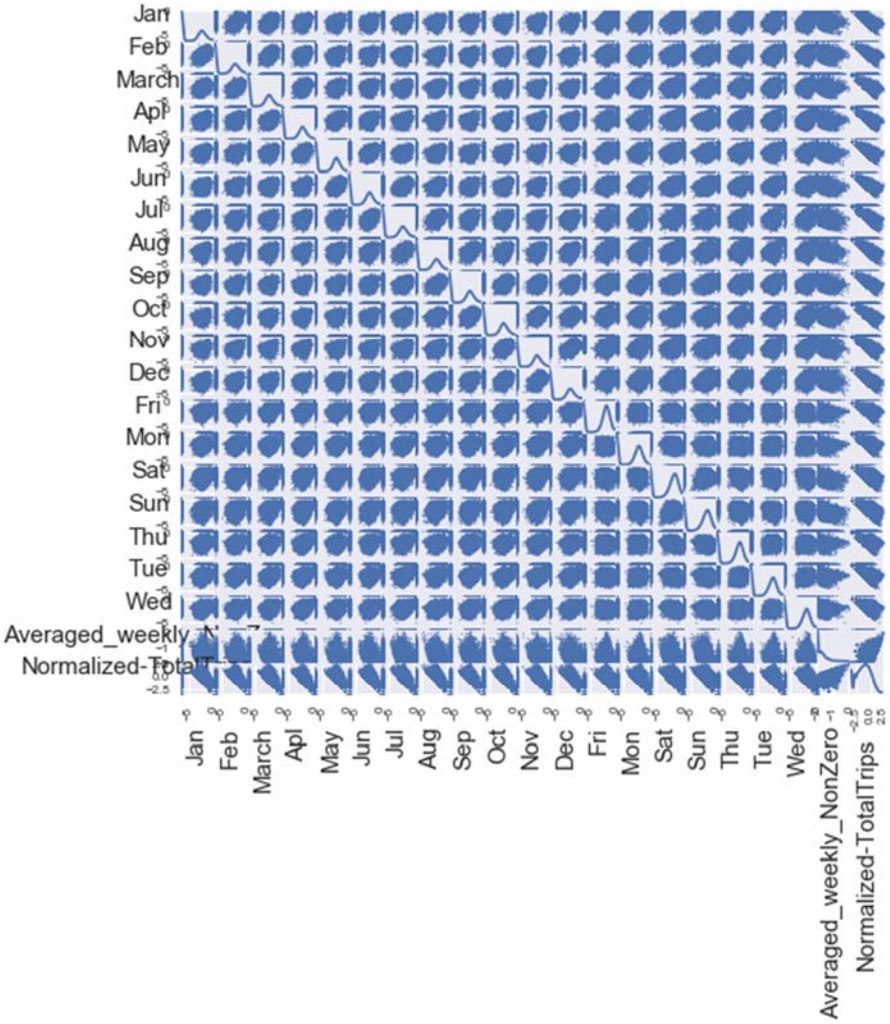

Figure 4-2 : *The pairwise scatter plots for all the variables after log-transformation*

However, some of the variables are still strongly correlated. We expect PCA to resolve this problem.

But before that, an important prerequisite for PCA is to ensure, all the variables follow the same measurement. For this reason, standardization is essential before PCA. We might standardize the data first and then perform PCA, or use the correlation matrix instead of covariance matrix. Both would lead to the same results.

4.2 Principal Component Analysis

As already calculated, if we retain 72% of the data's variance, the number of components would be *K*=10, which empirically proves to be small enough to treat the "curse of dimensionality" for k-means clustering. On the other hand, preserving 72% of the whole variance of the data is reasonable for the percentage of information we desire to maintain. Figure 4.3 shows the accumulated explained variance by the principal components.

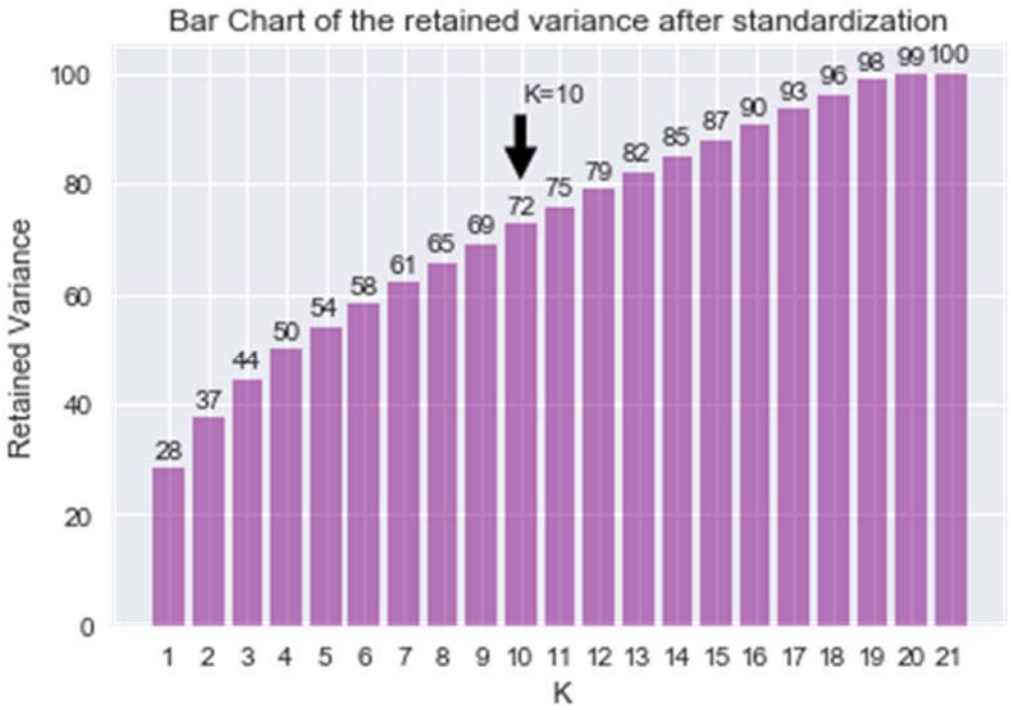

Figure 4-3 : *The accumulated retained variance for each number of principal components*

Here we look back to the assumptions. We would like to observe if the distributions of the variables are spherical now. Figure 4.4 shows that the pairwise scatter plots of the variables is spherical. In comparison with the figures 4.2 and 3.2, the histograms of the variables are now closer to the normal distribution and less skewed.

Recall that because of sphericity, the principal components must have a correlation of zero between them and variance of one.

Figure 4.5, also graphically shows that the correlation matrix of the principal components is very equal to the identity matrix. So, we conclude that the variance of the principal components is one and the covariance between them is zero, which is a proof that the distributions of the variables are spherical.

Now that the assumptions of k-means are met, the k-means clustering can be applied on the principal components.

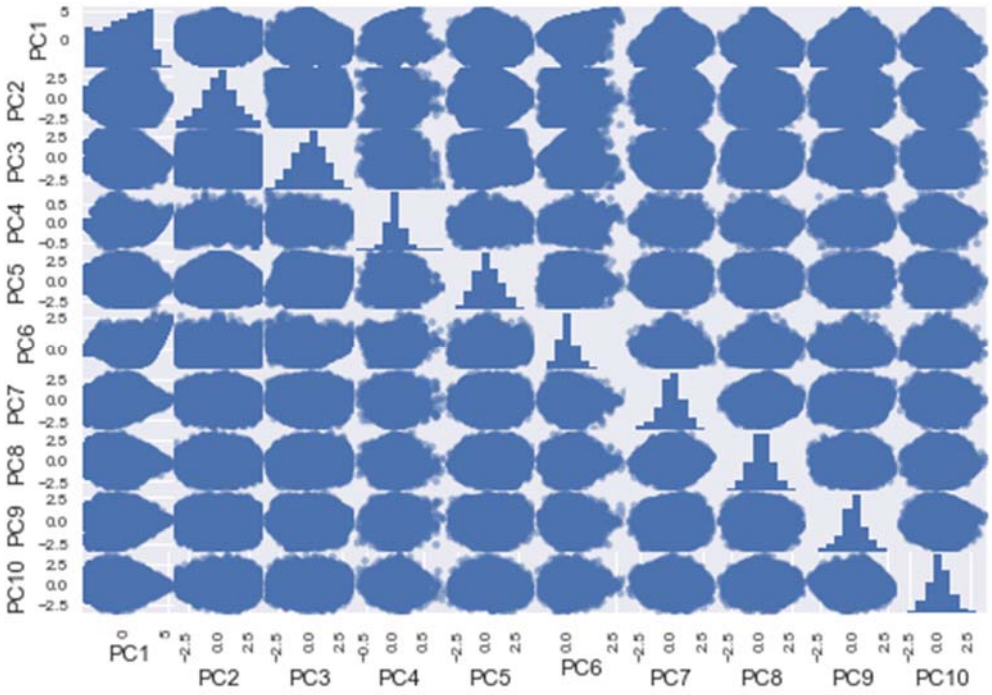

Figure 4-4: *The scatter matrix of the principal components*

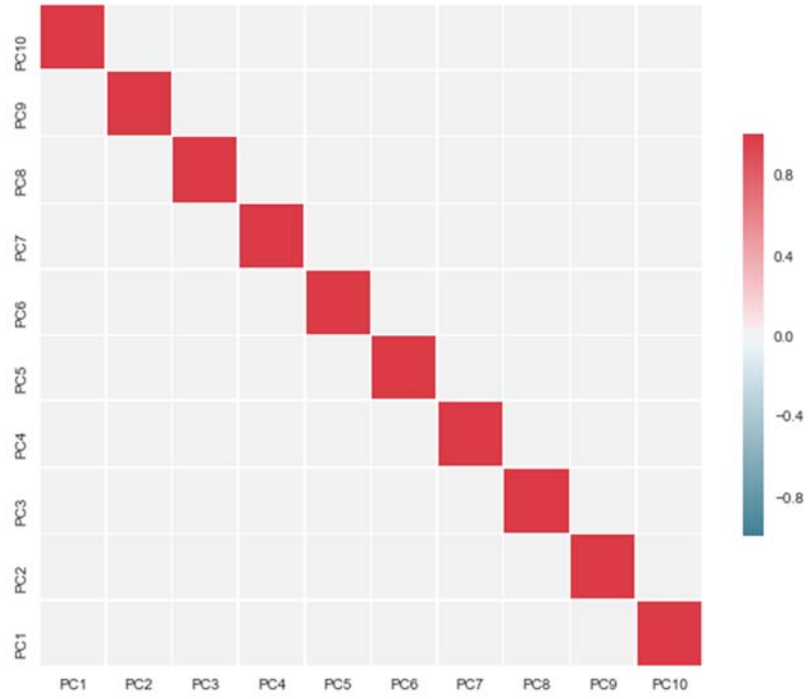

Figure 4-5: *The principal components correlation matrix*

4.3 K-means Clustering

Selecting the number of clusters prior to k-means clustering, is essential and Silhouette score is introduced as one of the methods of k selection in k-means. Figure 4.5, shows the silhouette score for each number of clusters, which is an average of all the silhouette scores of the observations.

Since the closer to 1 the silhouette score is, the better the clustering is supposed to be, the candidates for the number of clusters according to Figure 4.5, could be *K*=5, 7, 9. At these three points, this score drops afterwards. Specially at the points 5 and 9.

Although the silhouette score is higher for 2 or 3 clusters, they do not seem appropriate numbers of clusters, because they tend to have higher errors.

Cross-validation is another method of defining the number of clusters in k-means. Since k-means clustering is trying to minimize the sum of squared errors (SSE), explained in (3.9), we calculated the Mean Squared Error or MSE, in a 5-fold cross-validation for each number of clusters. Figure 4.6, shows how the average amount of MSE for each corresponding number of clusters changes.

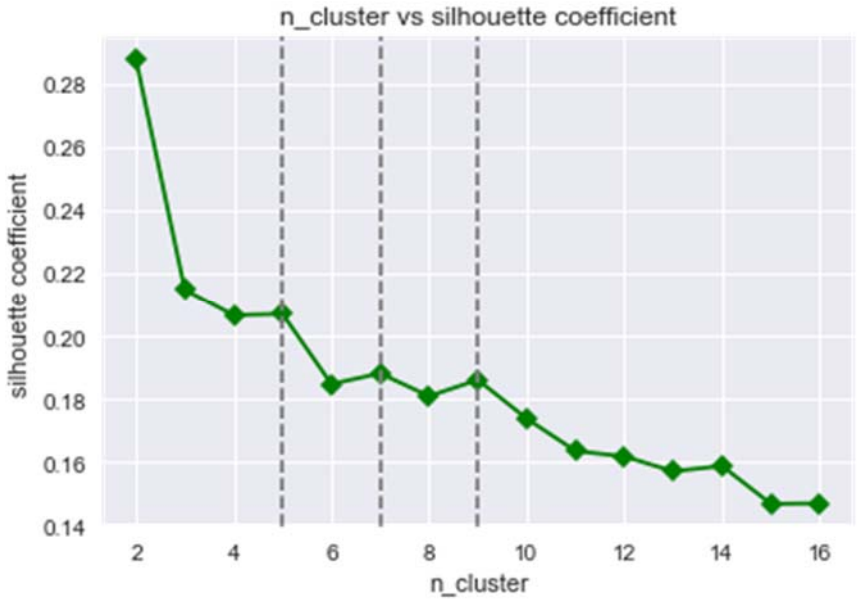

Figure 4-6: *Silhouette score for each number of clusters*

According to the figures 4.6 and 4.7, we may accept a range of K in which the amount of MSE drops moderately. Meaning that it should not drop significantly and it should not drop very slightly. Because, as the number of clusters increases, MSE decreases and this is favourable for k-means, while the silhouette score is dropping for the bigger K's, which is not favourable.

Consequently, we start the interval of the possible K's at a point from which MSE drops less significantly (K=7) and end it at a point from which MSE drops more slightly (K=12). The cross-validation to select *K* for K-means is implemented in python. (See appendix 1)

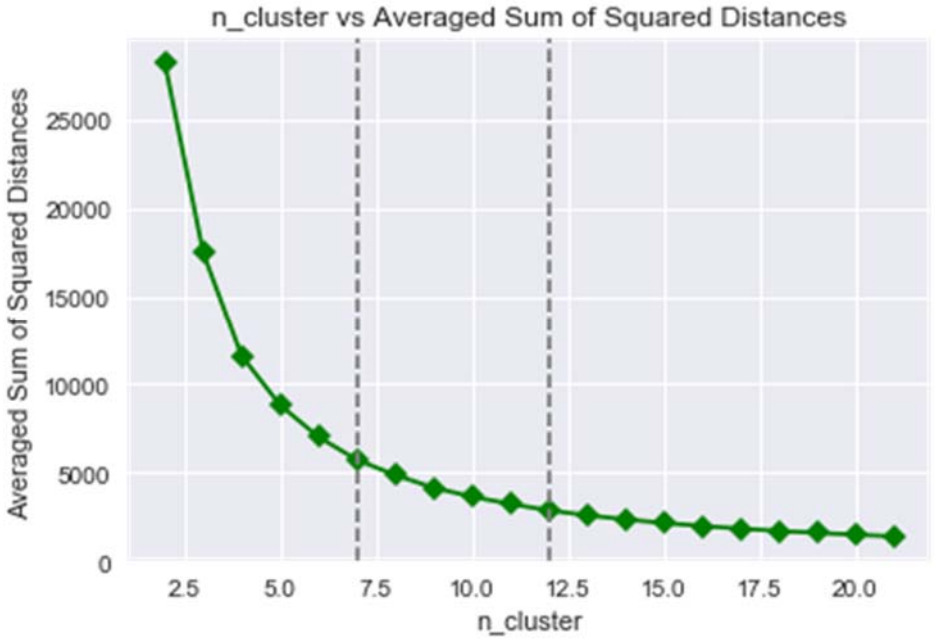

Figure 4-7: *Mean squared errors for each number of clusters*

4.3.1 Which *K*?

To finally resolve the number of clusters, among the range of possible *K*'s from cross-validation results (*K*= 7,8,9,10,11,12), we pick the ones with the highest silhouette scores which are *K*=7 and *K*=9. We also examine *K*=12 because of the lowest MSE.

Conversely, as illustrated in figure 4.7 we do not accept *K*=5 because of the high MSE, but because of its high silhouette score we verify it to compare with the other number of clusters.

Figure 4.8, illustrates customer monthly usage patterns in each cluster for different number of clusters in k-means, i.e. for *K*=5,7,9,12.

This figure shows that clustering with 5 and 7 clusters, build three distinct look-alike clusters and the other clusters seem to follow close patterns. To be exact, clusters "0, 2, 3" in *K*=5, are the same as clusters "1, 2, 4", in K=7. Whereas, clustering with 9 clusters, reveals one more distinct cluster, cluster 4 – yellow, and the other five clusters follow very close patterns. Clustering with 12 clusters, does not add anything to the previous ones. In contrary, it shows less distinct clusters.

Because there is a trade-off between the mean square error and the silhouette score, we must be careful about the number of clusters, to keep the balance. As a result, ***K*=9** can prove to be the best, because it produces less error than *K*=5,7 and it has the same silhouette score as *K*=7. Besides, the clusters in clustering with 9 clusters, are more distinct than in clustering with 12 clusters.

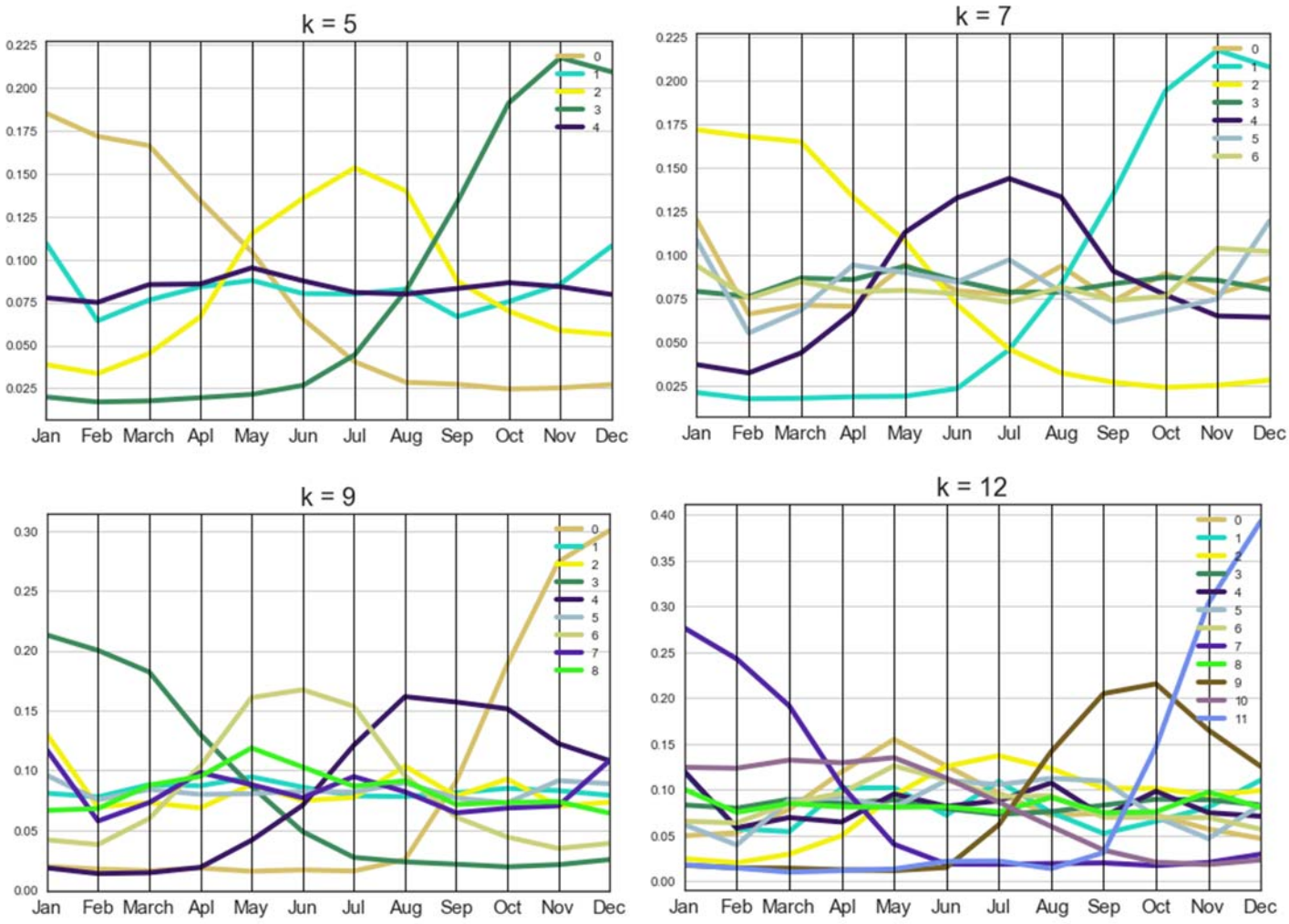

Figure 4-8: *Customers monthly usage patterns for different numbers of clusters.*

4.3.2 K-means with nine clusters

Now that we are confident about the number of clusters, we study more in deep the behaviours of the customers. As already illustrated in figure 4.7, the customers' usage variation in different clusters is different.

Figure 4.9 helps to observe better these variations in months. In this illustration, figures from A to D, are separated for clusters from higher variations to the lower variations respectively.

The usage patterns in (A), are extremely clear and there is no sudden jump. The users in (A) choose some consecutive months to use carsharing system, each cluster shows a different season, and the average usage goes down to near zero in the other months. Appendix A, displays the colour spectral table for the monthly usage patterns in each cluster. In this table, the users of different seasons can be distinguished.

The usage pattern in (C), is also showing some months as the most preferable ones, which are happening in spring and summer. The users in (B) and (D) happen to show up in any season. However, they also have some preferable months. Table 4.1 in the next section, describes these usage patterns in detail.

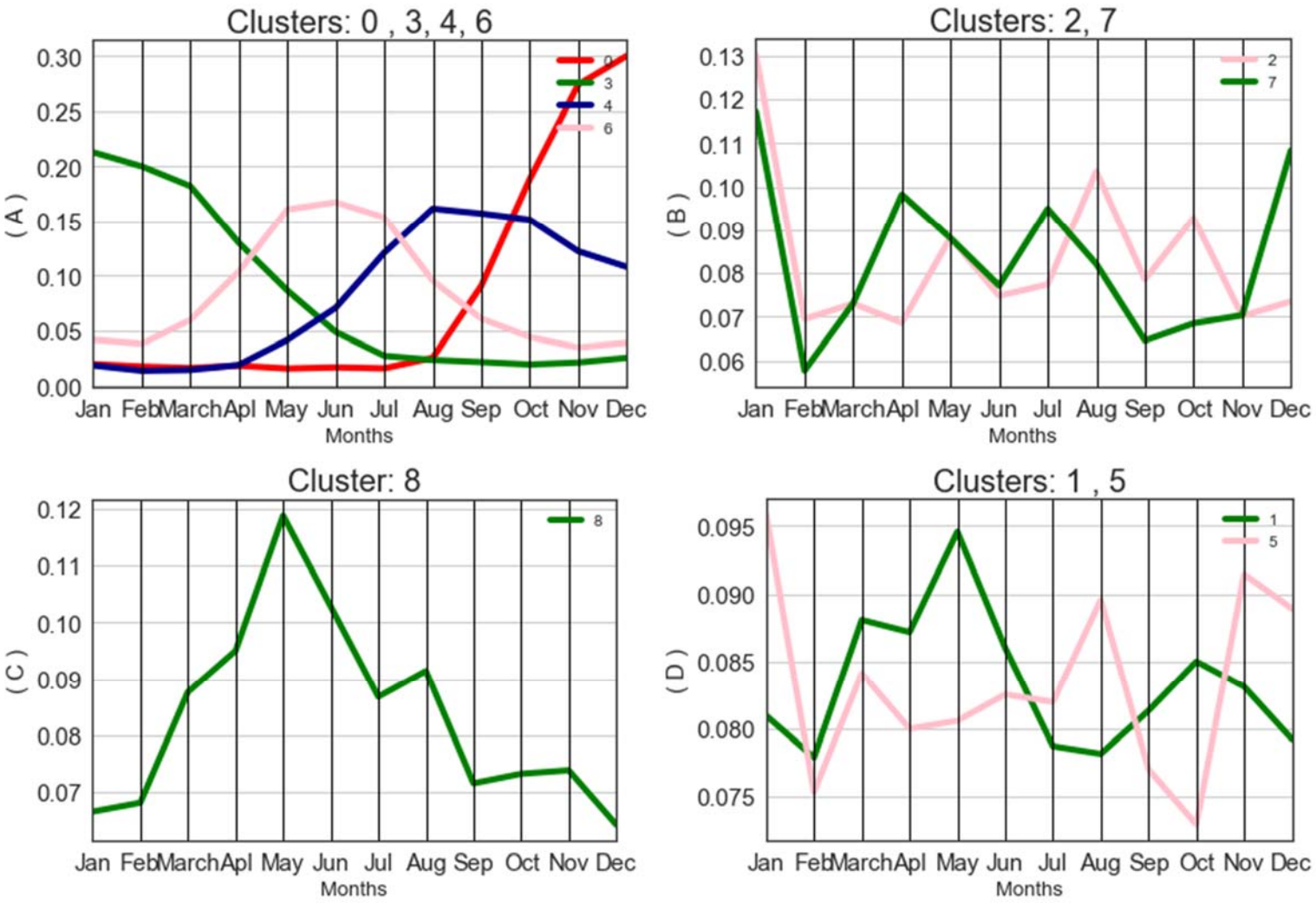

Figure 4-9: *Customer monthly usage pattern for k-means clustering with nine clusters*

Figure 4.10, displays such variations in weekdays. The five illustrations in this figure are separated according to the range of variations and the similarity of patterns.

In this figure, clusters in A, are mainly Saturday and Sunday users, while in B they are almost only-Friday users. The users who are less interested in weekends are grouped in C, whereas users in D, are mostly Saturday users, but also, they tend to show up on Fridays and Sundays. We must note that the variations from the illustrations A to D, decrease. So, Although the users in D (clusters 1 and 4) seem to be the weekend users, their usage frequency over the week does not change as much as it does for the clusters in A (clusters 5 and 8).

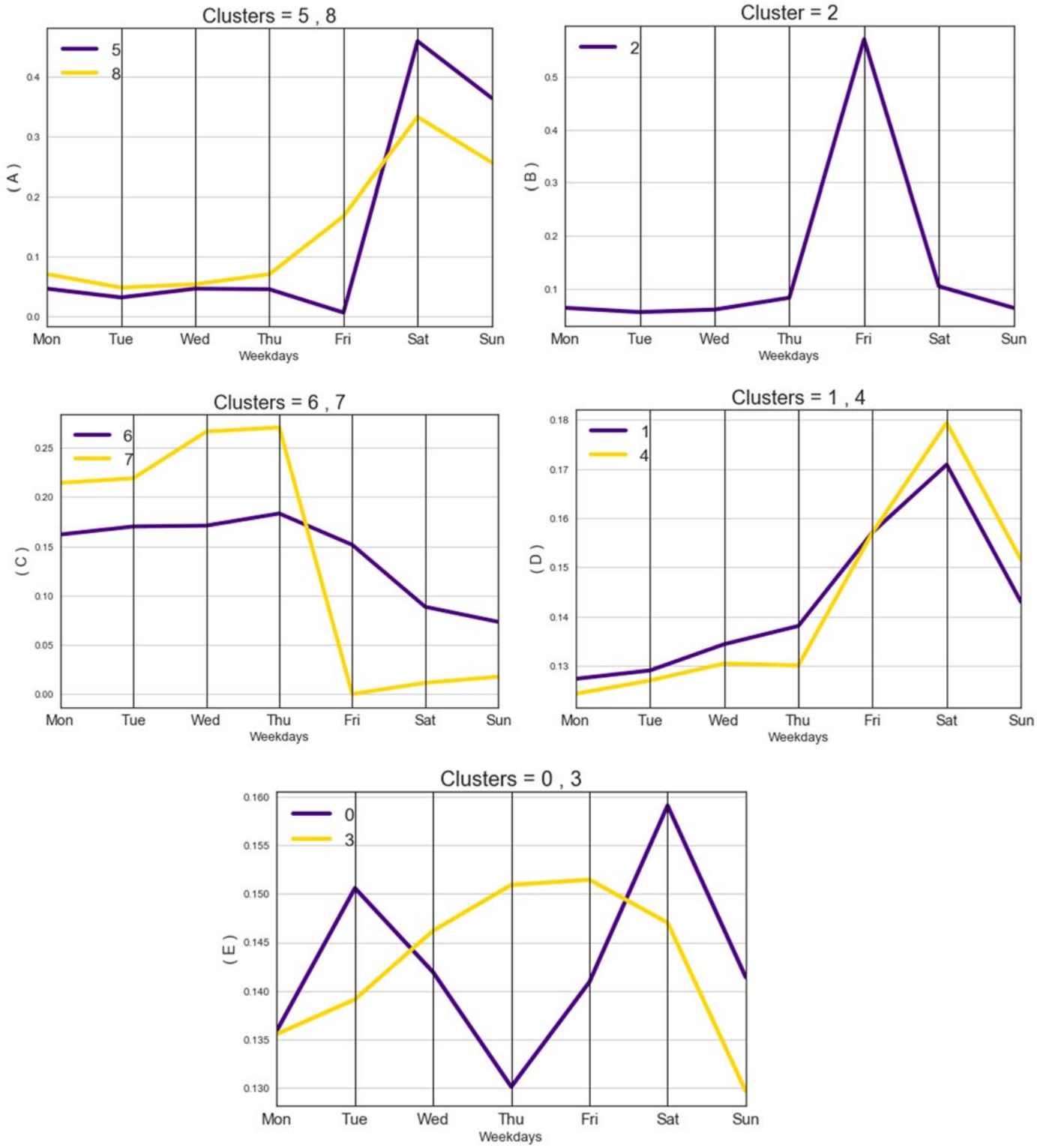

Figure 4-10: *Customer weekday usage patterns for k-means clustering with nine clusters*

Users in E are likely to use the carsharing system any day of the week, but almost never on Thursdays and Sundays in cluster 0 and 3, respectively. Appendix A, displays the colour spectral table for the weekday usage patterns in each cluster.

Figure 4.11 shows the customer usages as a total in the year 2014 by the attribute “Normalized-Total-Trip”, and an average over the weeks in the same year that the customer travelled at least once, by the attribute “Averaged-weekly-Nonzero”. These attributes, especially the former, help to recognize the intensity of usage during one year. This way, we would find out which cluster of customers are intensive users of carsharing system, and which ones are just occasional users. As illustrated in figure 4.10, cluster 1, has the most intensity of usage during a year which is in distant with the other clusters.

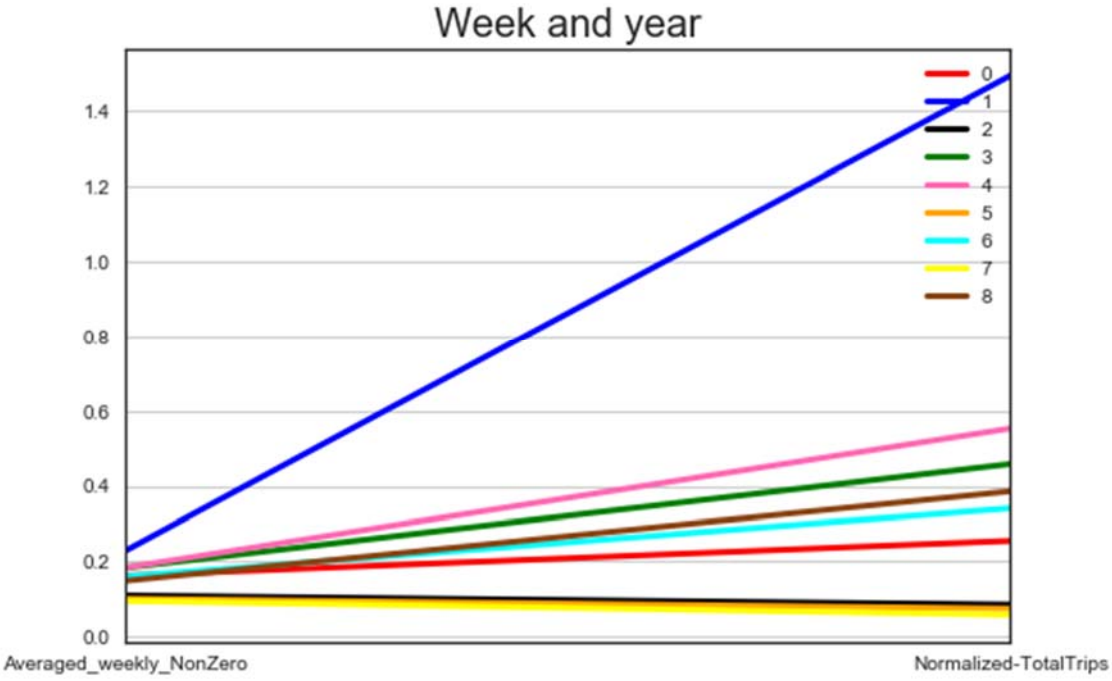

Figure 4-11: *Customer total usage in a year and customer averaged usage over all the weeks in a year that they travelled at least once*

4.4 Customers

As explained in methodology section, customers’ dataset is attached to the vector of attributes, to better describe each cluster according to the customers’ characteristics. Out of 28,464 customers only 19,309 of them could be found in the customers’ dataset and the personal specifications of the others were not available. The feature “Age”, which is calculated according to the birthdates and the year which this dataset relates to, 2014, contained some missing values. Since one of the treatments for missing values is to replace them by a value like median, the median age is calculated over the available ages, so that it can be a representative of the missing ones as well. Table 4.1, describes the characteristics of each cluster, according to the intensity and patterns of usage, as well as the available customers’ features.

In this table, the user class is built according to the usage intensity over a year, and the monthly and weekdays usage patterns, shown in the figures 4.10, 4.8 and 4.9 respectively, as well as in the appendix A and B.

Clusters	*User Class*	*Number of users*	*Average number of trips per user*	*Gender Ratio (W/M)*	*Distribution of Women(%)*	*Distribution of Men(%)*	*Median Age*	*Average years of membership*	*Language Ratio (Fr/En)*
1 = A	Extreme - Regular	6528 (34%)	49.53	1.07	34.00	33.61	41	5.19	3.76
4 = B	Intensive – summer, fall	1208 (6%)	15.99	0.98	6.01	6.51	40	4.88	4.25
3 = C	Intensive – winter, Spring	2170 (11%)	15.19	1.17	11.80	10.65	38	4.25	3.84
8 = D	Very Frequent – weekends	2254 (12%)	12.85	1.12	12.01	11.32	39	4.52	5.85
6 = E	Very Frequent – spring, summer	1588 (8%)	10.88	0.83	7.24	9.26	43	4.94	4.84
0 = F	Frequent – Fall	790 (4%)	7.65	1.06	4.10	4.08	42	4.73	5.75
2 = G	Occasional – Fridays	1340 (7%)	2.87	1.04	6.88	7.00	40	4.29	4.88
5 = H	Occasional– weekends	1938 (10%)	2.50	1.21	10.71	9.33	41	4.12	5.61
7 = I	Occasional – Mon to Thu	1493 (8%)	1.95	0.93	7.25	8.24	42	4.73	5.07
Total		19309 (100%)	13.27	1.05	100.00	100.00	41.00	4.63	4.87

Table 4-1: *Cluster characteristics based on the Communauto regular-service customers' dataset and the k-means clustering patterns*

The clusters are ordered with respect to the usage intensity in each cluster. The "average number of trips per user" as an indicator displays this order. Also, the intensity of usage per user is named by "Extreme" users to the "Occasional" users. Regular in the user class column, means that the users in this cluster are using the carsharing system almost all the months and all the days of the week. Whereas an indicated month or day means that the users are showing up in some specific days or months more than the other times. For instance, the customers' usage behaviour in cluster B, is very intensive in summer and fall, but they don't show a very specific pattern for the weekdays. In contrary, the users in cluster G who are occasional ones, tend to show up on Fridays much more than the other days, however they might be using the regular carsharing system in any month of the year.

Gender ratio indicates that in all the clusters, except for B, E and I, women are the dominant customers of Communauto carsharing regular-service. The median age of all the users is around forty-one and the average years of membership in all the clusters is more than four years. Plus, the French speakers are very dominant in all the clusters which is quite expected in Quebec, Canada

4.5 What about the outliers?

In the sections, 3.4.1 and 4.1.1, we explained about the outliers and the reasons why we needed to remove them as the first step of pre-processing the data. The outliers were removed temporarily to be studied separately. Since only 152 customers were put aside as the outliers, we would assume them as one cluster which is isolated from the data, not by k-means clustering, but by Mahalanobis distance as a multivariate outlier detector.

Figure 4.11, shows the carsharing usage patterns of the users detected as the outliers. Among them, some users were discovered to had travelled only once during the year 2014. On the other hand, one user with the customer id: 108, had 8,133 trips during only one year, which was so far from any other customer records. Therefore, the outliers were found to be divided in three classes: 1) One-trip users, 2) The customer id: 108, 3) Others.

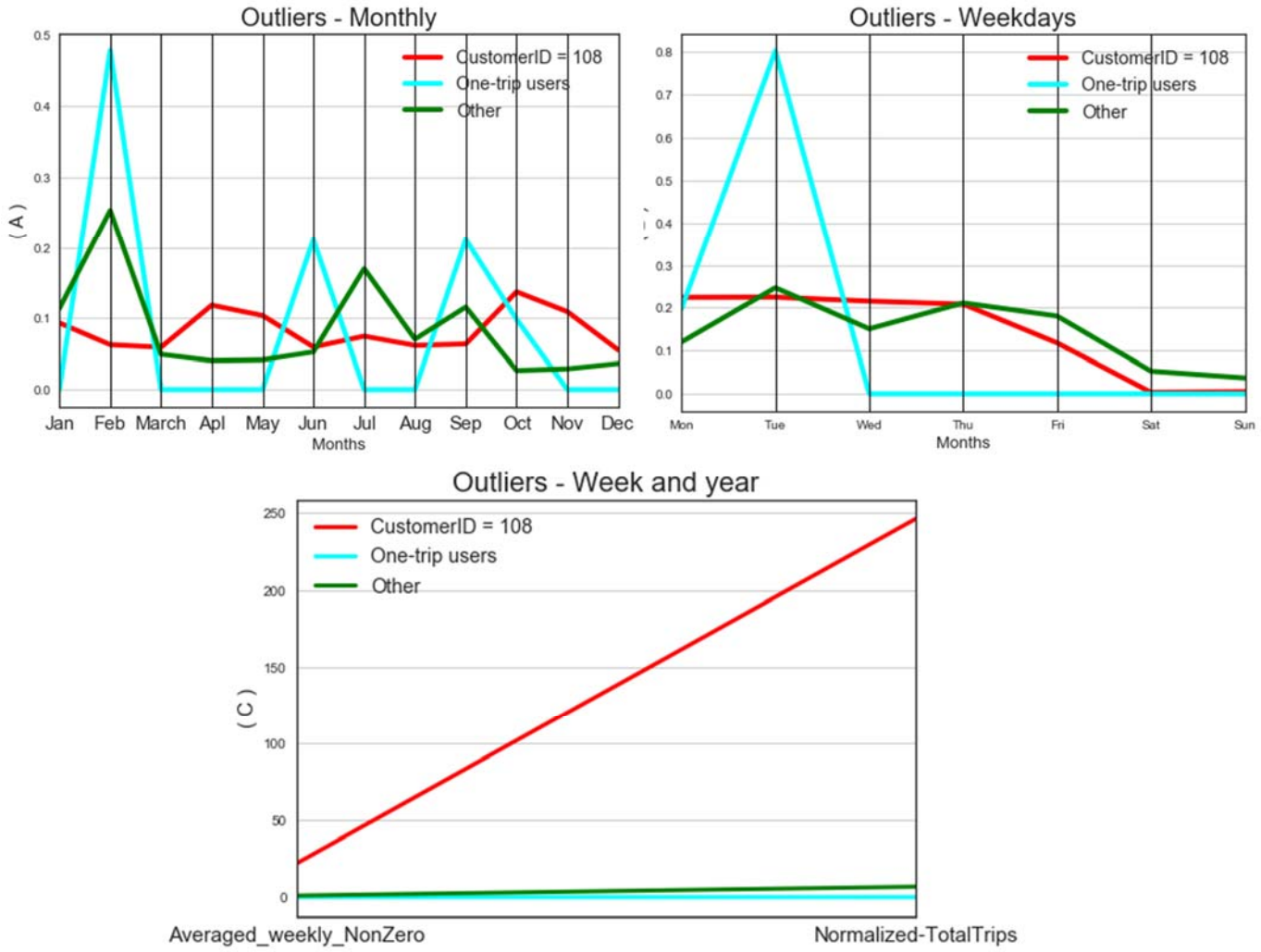

Figure 4-12: *Carsharing usage patterns for the users identified as the outliers*

As shown in figure 4.12 (A), most of the one-trip users were the customers of February, and some of them were the users in June, September and October. No one-trip user detected as the outliers, can be seen in the other months. The “other” users with the green line, were also mostly interested in using the carsharing in February. Customer with the id 108, seem to be a user of almost every month. Illustration (B) shows that the one-trip class of outliers were mostly Tuesday users and Monday afterwards. But zero use for them in the other days. Other users and customer id 108, have similar behaviours, except the customer id 108, is not a weekend user at all. However, the “other” users are also less interested in weekends but their usage is not zero. Illustration (C), also shows that the customer id 108 total number of trips and average weekly usage, is quite bigger than the other users in the class of outliers.

A bit of investigation could possibly reveal that the customer id 108, could be an employee whose specifications is among the customers by mistake. However, this is only a deduction by the fact that he uses carsharing a lot and as shown in figure 4.12 (B), he is absolutely absent during the weekends.

Table 4.2 describes the three classes of the outliers in accordance with the customers' dataset. As the customers' specifications were not available for all the customers, these descriptions are available only for 92 customers out of 152 customers. Also, for some of these 92 customers, "Age" feature was missing, so the median age is calculated out of the available ones.

Class	*User Class*	*Number of users*	*Average number of trips per user*	*Gender Ratio (W/M)*	*Distribution of Women(%)*	*Distribution of Men(%)*	*Median Age*	*Median of years in Communauto*	*Language Ratio (Fr/En)*
1	Customer ID = 108	1	8133.00	Male	----	----	NA	13.00	French
2	One-Trip users	71	1.00	0.97	0.49	0.51	39	2.00	4.07
3	Other users	20	224.00	0.67	0.40	0.60	41.5	5.50	3

Table 4-2: *The characteristics of the identified classes in the outliers*

As described in the table 4.2, the number of trips for the customer id: 108 is significantly distant from the average of the users that have travelled more than once in the "other users" class. As already discussed, this specific user's behaviour is different from all the other users in the whole dataset. As shown in the table 4.2, his specifications were available in the customers' dataset, but not all of them. For example, the age of this person was not available.

A comparison between the average number of trips in table 4.1 and the outlier class "other users" in table 4.2, provides another evidence that these customers' behaviours might be different from rest of the data.

Nevertheless, the multivariate outlier detection methods like Mahalanobis distance, consider all the variables dependently to detect the outliers, not only one variable such as "the total trips". This might justify the "One-Trip" class of users in the outliers. i.e. not only the total number of trips, but the whole variables together might put a data point farther from the rest.

However, Mahalanobis distance's results might have a percentage of error like every other statistical method. Thus, one might concludes that a possible treatment is to put the group of one-trip users back to the data to be clustered by k-means.

4.6 K-means' MSE with or without PCA

Previously, the main reasons of applying PCA transformation on the data before k-means clustering were discussed. But, how about clustering the data on the original variables. In this section, we empirically prove that PCA lower the Mean Squared Error of k-means clustering very significantly.

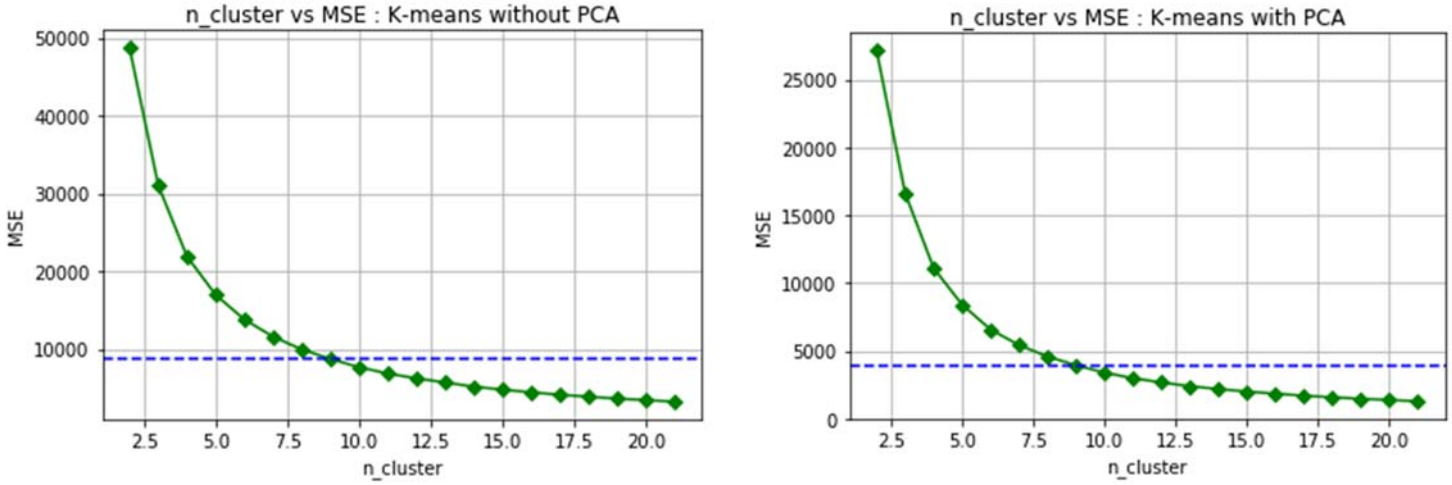

Figure 4-13: *Number of clusters vs Mean Squared Error of K-means clustering on the original variables and on the PCA transformed data, from left to right respectively*

Figure 4.13 illustrates the difference between MSEs in K-means clustering without PCA and with PCA. The mean squared errors of clustering on the original variables is about twice bigger than the ones on the PCA transformed data. For instance, the MSE of clustering with nine clusters is about 4000 whereas without PCA it is close to 10000.

To know what PCA changes on the data that causes lessening the clustering error, we refer to what we explained earlier. Since the original variables are correlated, the distribution of the data is non-spherical and diagonal, which is against the assumptions of K-means. Especially the two variables X_{20}*: Averaged weekly trips* and X_{21}*: Normalized total trips*, are strongly correlated together, about 68%, and X_{21}*: Normalized total trips* is remarkably correlated with the other variables, 40% to 50%. Figure 4.14, shows the K-means' MSEs if we put aside these two variables.

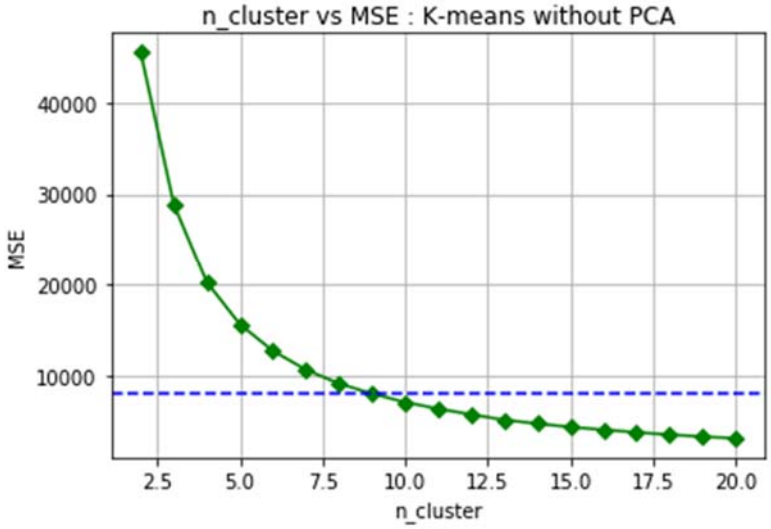

Figure 4-14: *The number of clusters vs MSE of clustering on the original variables if there were no X_{20} and X_{21}*

In comparison with the left-side graph in figure 4.13, figure 4.14 shows lower MSEs. This is due to removing two high correlated variables. However, the MSEs are still much higher than the right-side graph in figure 4.13. These results support the fact that the correlation among the variables and their non-spherical distribution are not the only issues for K-means.

K-means clustering suffers from the curse of dimensionality, especially when the effective number of variables is less than the actual number of them. Perhaps “21 variables” should not be a big number, but how many of those are effective. Principal Component Analysis, builds a new subspace with the effective number of variables, which are uncorrelated most of the times. In our case the 21 variables were transformed into 10 uncorrelated variables by PCA, explaining 72% of the variance of the original data.

5 CONCLUSION

The objective of this study was to find the usage patterns of Communauto carsharing regular-service customers. Using k-means clustering, nine unique user profiles were found. These profiles were ordered from the most frequent users to the most occasional ones. Each cluster or user profile is identified with the most favourite season or days of the week for using the service. The low value of Silhouette score was not a major issue, as the main purpose of this study was to exploit the k-means clustering via PCA and discuss the methodology.

5.1 Contribution

K-means is one of the most popular clustering methods, because of its speed and simplicity. However, it has some assumptions and limitations on the data. It assumes that the distribution of the data to be clustered is spherical and consequently the variables are uncorrelated and have a variance of one. This assumption was not met on our raw data. On the contrary, the data contained big outliers and the distribution of the variables was strongly right-skewed. The cloud of the data had no sphericity, but it was more diagonal, meaning that there were correlations among the variables. This issue needed to be resolved before clustering.

Therefore, the big outliers were kept apart from the data using Mahalanobis distance and were analysed separately. This was a very important task to be done before everything, since the very big outliers could distract any other task on the data. But still the data distribution was strongly right-skewed, so, log-transformation helped on this issue and made the distribution of the variables closer to normal and even the cloud of the data closer to spherical but not for all the variables.

Since some of the variables were correlated, Principal Component Analysis was chosen to be applied on the data, to have uncorrelated variables. At the same time, this reduced the noise and the number of variables to which k-means clustering was sensitive. However, PCA is also sensitive to the data measurements and had to be performed on the standardized data. So, the data was transformed in three steps: first log-transformation, then standardization and afterwards PCA. Subsequently, k-means clustering was performed on the transformed data.

Several works have been done on the vehicle sharing datasets, but none talked about the k-means clustering issues. In this study, we examined and found that k-means clustering on the PCA transformed data had smaller mean-squared error than the k-means clustering on the original data.

Some of the similar works that preferred not using PCA transformation for k-means, addressed the interpretability of the results of k-means on PCA transformed data, as an issue. Whereas interpretability should not be an issue when thinking of k-means clustering as an unsupervised learning.

To discuss this thought recall that, k-means attempts to cluster the observations of the unlabelled data, and PCA transforms the data according to the variables (columns). Consequently, the observations (rows) remain the same in the new transformed data. Thus, the original data would adopt the resulting cluster labels, and they could be simply interpreted according to the original variables.

5.2 Future works

Statistical methods always have difficulties dealing with most types of data. There are always assumptions to be met before analysing. However, there are robust methods that are less sensitive and can handle the data conditions. For instance, in *"Robust and sparse k-means clustering"* (Xu, Han et al. 2016), a k-means approach has been proposed that can treat the outliers. There are also some other works that propose alternative methods to handle the data that is not spherical. Like in *"What to do when K-means clustering fails: a simple yet principled alternative algorithm"* (Raykov, Boukouvalas et al. 2016). Since data transformations can alter the accuracy of the results, one of the advantages of utilizing the robust methods is that the original data would be clustered without any transformation. As a future work on the similar data, the analyst could consider the robust methods to improve the results accuracy.

BIBLIOGRAPHY

Arthur, D. and S. Vassilvitskii (2007). k-means++: The advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics.

de Amorim, R. C. and C. Hennig (2015). "Recovering the number of clusters in data sets with noise features using feature rescaling factors." Information Sciences **324**: 126-145.

De Maesschalck, R., D. Jouan-Rimbaud and D. L. Massart (2000). "The mahalanobis distance." Chemometrics and intelligent laboratory systems **50**(1): 1-18.

Ding, C. and X. He (2004). K-means clustering via principal component analysis. Proceedings of the twenty-first international conference on Machine learning, ACM.

Franklin, S., S. Thomas and M. Brodeur (2000). Robust multivariate outlier detection using Mahalanobis' distance and modified Stahel-Donoho estimators. Proceedings of the Second International Conference on Establishment Surveys, American Statistical Association Buffalo, NY.

Friedman, J., T. Hastie and R. Tibshirani (2001). The elements of statistical learning, Springer series in statistics New York.

Jolliffe, I. T. (2002). "Springer series in statistics." Principal component analysis **29**.

Klincevicius, M., C. Morency and M. Trépanier (2014). "Assessing impact of carsharing on household car ownership in Montreal, Quebec, Canada." Transportation Research Record: Journal of the Transportation Research Board(2416): 48-55.

Le Vine, S. and J. Polak (2017). "The impact of free-floating carsharing on car ownership: Early-stage findings from London."

Le Vine, S., A. Zolfaghari and J. Polak (2014). "Carsharing: evolution, challenges and opportunities." Scientific advisory group report **22**.

Liang, Y., M.-F. Balcan and V. Kanchanapally (2013). Distributed PCA and k-means clustering. The Big Learning Workshop at NIPS.

Morency, C., M. Trépanier, B. Agard, B. Martin and J. Quashie (2007). Car sharing system: what transaction datasets reveal on users' behaviors. Intelligent Transportation Systems Conference, 2007. ITSC 2007. IEEE, IEEE.

Morency, C., M. Trepanier, A. Frappier and J.-S. Bourdeau (2017). Longitudinal Analysis of Bikesharing Usage in Montreal, Canada.

Raykov, Y. P., A. Boukouvalas, F. Baig and M. A. Little (2016). "What to do when K-means clustering fails: a simple yet principled alternative algorithm." PloS one **11**(9): e0162259.

Rousseeuw, P. J. (1987). "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." Journal of computational and applied mathematics **20**: 53-65.

Sarmento, R. and V. Costa (2017). Comparative Approaches to Using R and Python for Statistical Data Analysis, IGI Global.

Shaheen, S. A. and A. P. Cohen (2008). "Worldwide carsharing growth: An international comparison." Transportation Research Record Journal of the Transportation Research Board 1992 (458718).

Sioui, L., C. Morency and M. Trépanier (2013). "How carsharing affects the travel behavior of households: a case study of montréal, Canada." International Journal of Sustainable Transportation **7**(1): 52-69.

Su, T. and J. Dy (2004). A deterministic method for initializing k-means clustering. Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on, IEEE.

Trépanier, M., Morency, C., Nouri P., Braham A. (2013), Impacts of carsharing on urban mobility: environmental and behaviourial evidences, 13th World Conference on Transport Research, Rio de Janeiro, Brésil, 15-18 juillet

Vogel, M., R. Hamon, G. Lozenguez, L. Merchez, P. Abry, J. Barnier, P. Borgnat, P. Flandrin, I. Mallon and C. Robardet (2014). "From bicycle sharing system movements to users: a typology of Vélo'v cyclists in Lyon based on large-scale behavioural dataset." Journal of Transport Geography **41**: 280-291.

Wielinski, G., M. Trépanier and C. Morency (2017). Carsharing vs Bikesharing: Comparing Mobility Behaviors.

Xu, J., J. Han, K. Xiong and F. Nie (2016). Robust and Sparse Fuzzy K-Means Clustering. IJCAI.

APPENDIX A – TABLE OF CLUSTERS' PATTERNS 1

Cluster	Jan	Feb	March	April	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
A = 1	0.08	0.08	0.09	0.09	0.09	0.09	0.08	0.08	0.08	0.08	0.08	0.08
B = 4	0.02	0.01	0.01	0.02	0.04	0.07	0.12	0.16	0.16	0.15	0.12	0.11
C = 3	0.21	0.20	0.18	0.13	0.09	0.05	0.03	0.02	0.02	0.02	0.02	0.03
D = 8	0.07	0.07	0.09	0.09	0.12	0.10	0.09	0.09	0.07	0.07	0.07	0.06
E = 6	0.04	0.04	0.06	0.10	0.16	0.17	0.15	0.10	0.06	0.04	0.03	0.04
F = 0	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.09	0.19	0.27	0.30
G = 2	0.13	0.07	0.07	0.07	0.09	0.07	0.08	0.10	0.08	0.09	0.07	0.07
H =5	0.10	0.08	0.08	0.08	0.08	0.08	0.08	0.09	0.08	0.07	0.09	0.09
I = 7	0.12	0.06	0.07	0.10	0.09	0.08	0.09	0.08	0.06	0.07	0.07	0.11

Cluster	Mon	Tue	Wed	Thu	Fri	Sat	Sun
A = 1	0.13	0.13	0.13	0.14	0.16	0.17	0.14
B = 4	0.12	0.13	0.13	0.13	0.16	0.18	0.15
C = 3	0.14	0.14	0.15	0.15	0.15	0.15	0.13
D = 8	0.07	0.05	0.05	0.07	0.17	0.33	0.26
E = 6	0.16	0.17	0.17	0.18	0.15	0.09	0.07
F = 0	0.14	0.15	0.14	0.13	0.14	0.16	0.14
G = 2	0.06	0.06	0.06	0.08	0.57	0.10	0.06
H = 5	0.05	0.03	0.05	0.05	0.01	0.46	0.36
I = 7	0.21	0.22	0.27	0.27	0.00	0.01	0.02

APPENDIX B - TABLE OF CLUSTERS' PATTERNS 2

intensity	Cluster	weekend users	Friday	Weekday users	All-day users	Winter users	spring users	summer users	Fall users	all-seasons users
Extreme	A = 1				1					1
Intensive	B = 4				1			1	1	
Intensive	C = 3				1	1	1			
Very Frequent	D = 8	1								1
Very Frequent	E = 6			1			1	1		
Frequent	F = 0				1				1	
Occasional	G = 2		1							1
Occasional	H = 5	1								1
Occasional	I = 7			1						1