AREADUM REPORT

DAMA - UPC

REPETITIVE ROUTE ANAYSIS

April 27, 2017

1 Introduction

In this report, the intention is to analyze the repeated daily routes of the users who are responsible with delivery of goods in Barcelona. The number of users is 49172. Thus, it is not easy to extract the routes without an algorithm which serves our need.

As the first step of our analysis, we detected *One-time Users* who are the users with one single delivery in our dataset. This users and their corresponding deliveries are removed since having them is not suitable for repetitive pattern analysis.

For the filtered data set, a common Bioinformatics approach, *Longest Common Subsequence* algorithm is used as baseline. We improved the algorithm until it becomes suitable and sufficient for our problem, which is the level of *Multiple Longest Common Subsequence*. This new improved version of the algorithm is used to detect the loading/unloading areas which are repeated exactly everyday by each user in order to see how the users adhere strictly to the fixed routes.

As a complementary repetitive routes' analysis, first we did clustering based on the number of loading/unloading areas visited in a day in order to detect the target groups. Then, we do ratio analyses to see the proportions of repetitiveness. The activity type effects on the repetitive routes are examined for each cluster to understand the behaviors if they are similar. These analyses and examinations are repeated for weekly and monthly routes as well.

The rest of this paper is structured as follows. In Section 2, we describe the data used. In Section 3, we give an overview of methods used. In Section 4, the experiments and results for the data does not include disallowed repeated check-ins are detailed for daily, weekly and monthly routes. Finally, in Section 5, we conclude and make remarks about future work.

2 Data Description

The same data we used for previous reports is used for repetitive route analyses. **We could not import the new data came for the dates from mid July 2016 to January 2017 is not used since there are some missing attributes in the new data.**

Since we do analysis for repetitive patterns, we removed the *One-time*

Users who are the users with only one single delivery in the data. With this data cleansing technique 4784 of 49172 users (9.73%) are filtered out.

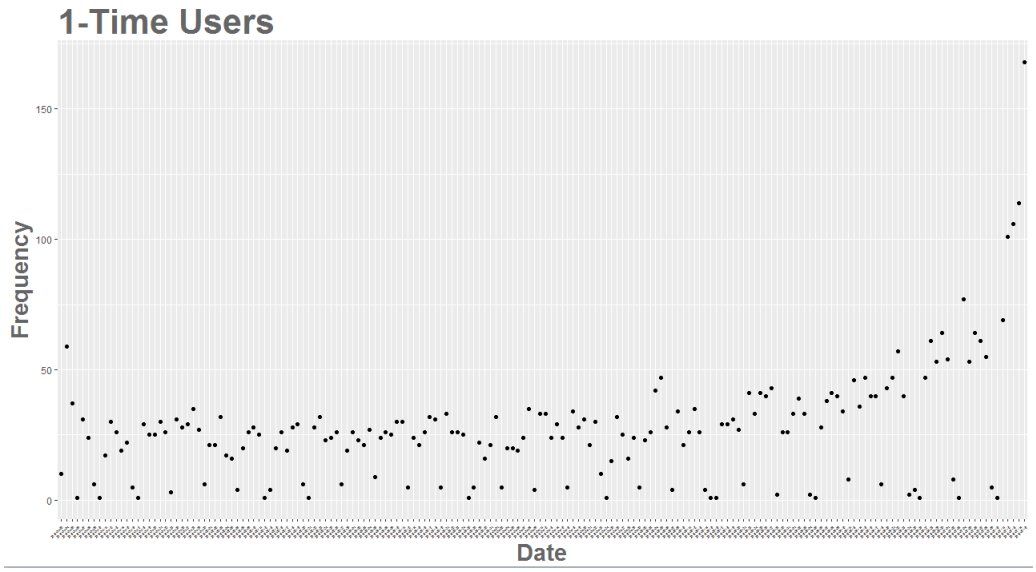
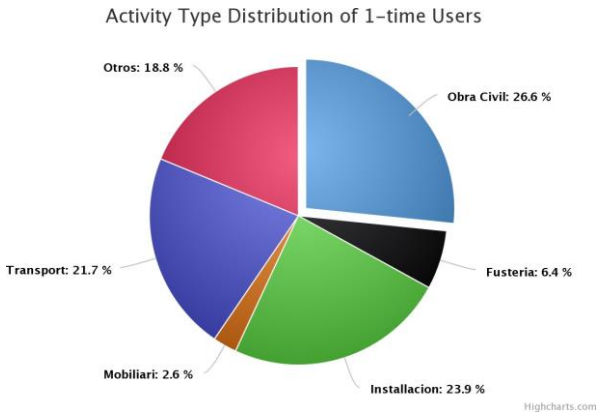


Figure 1: The number of one-time users by date

For the interpretation of Figure 1, we also need to consider the fact that the right side of the plot can be misleading. This increase can be related to the new users who would have more check-ins in the new data. However, we filtered out these users for now. Furthermore, we also plotted the distribution of activity types for one-time users.



3 Methods

In order to detect the delivery areas which are visited every day by a user, we need a specific algorithm which meets the need of multiple days capability. Thus, we first made the improvements that are needed to reach the level of multiple strings from the level of two strings. After the algorithm, we also give a brief information about the clustering technique we used.

3.1 Multiple Longest Common Subsequence

Longest Common Subsequence is a common algorithm which is used to detect the common subsequences for two sequences. It is a classic computer science problem, the basis of *diff* (a file comparison program that outputs the differences between two files) and has applications in Bioinformatics.

This algorithm finds the longest subsequence and its length present in both given two sequences. The output of the algorithm is a subsequence which is a sequence that appears in the same relative order, but not necessarily contiguous.

3.1.1 Implementation into Our Problem: Step 1

The LCS (Longest Common Subsequence) compares the characters from two strings in order to find the longest substring. In our case, we used the *Delivery Area IDs* as unique characters instead of letters. Each string consists of *Delivery Area ID* s represent a list of delivery areas that are visited in one day by a specific user. Nested list structure holds the information of each user's daily routes.

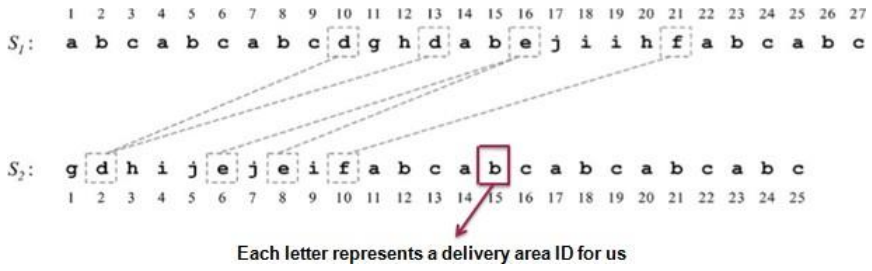


Figure 2: Longest Common Subsequence for Two Strings

In Figure 2, the letters are the representatives of delivery areas. S_1 and S_2 are are the lists of delivery areas which hold the information for two different days.

3.1.2 Implementation into Our Problem: Step 2

We have already pointed out that the subsequence is a sequence that in the same order. Because of this "same" order concept, we reordered the *Delivery Area IDs* (i.e. consist of 4 or 5 digits, integer) in an ascending sort.

3.1.3 Implementation into Our Problem: Step 3

The original version of LCS algorithm is only for two strings. Since the strings represent the daily routes of one user in our case, we can simply conclude that having only two days for a user is a restriction for us since it is easy to guess that there would be some users with more than two days record. Thus, we need the second implementation step here.

With the purpose of finding the longest common subsequence between any N strings, one might think there is a necessity to have a general way to iterate an array of N dimensions. However, we solved the problem recursively, instead of iterating the $N - dimensional$ array.

3.2 Model-Based Clustering

The traditional clustering methods such as *hierarchical clustering* and *partitioning algorithms* (k-means and others) are heuristic and are not based on formal models. An alternative is to use model-based clustering, in which, the data are considered as coming from a distribution that is mixture of two or more components (i.e. clusters). Each component k (i.e. group or cluster) is modeled by the normal or Gaussian distribution which is characterized by the parameters:

- μ_k: mean vector
- Σ_k : covariance matrix
- An associated probability in the mixture. Each point has a probability of belonging to each cluster.

The model parameters can be estimated using the EM (Expectation - Maximization) algorithm initialized by hierarchical model-based clustering. Each cluster k is centered at the means μ_k, with increased density for points near the mean. Geometric features (shape, volume, orientation) of each cluster are determined by the covariance matrix Σ_k. Different possible parameterizations of Σ_k are available in the R package mclust. The key advantage of model-based approach, compared to the standard clustering methods (e.g. k-means, hierarchical clustering etc.), is the suggestion of the number of clusters and an appropriate model.

4 Experiments & Results for The Data Without Disallowed Repeated Check-ins From January to Mid July 2016

4.1 Based on Daily Delivery

In this section, we analyze the routes of users in order to extract the subroutes which are repeated **everyday**. The main purpose is to see if the users are assigned to some specific loading/unloading areas in a part of the city. On the other hand, we get the information if they move around the city in different patterns without following any fixed route.

4.1.1 Clustering

There are two main axes we consider in order to locate the target groups.

- The number of days the users had delivery,
- The number of loading/unloading areas visited in a day.

In Figure 3, x-axis represents the number of delivery areas visited in a day by users, whereas y-axis represents the total number of days users had delivery. White color in the heatmap is for low values, and dark red is for high values. It does not make sense to take all data as one piece to analyze. For this reason, we cluster the frequency data which is located into cells comes from the number of delivery areas visited in a day in the heatmap.

We compared the output and the accuracy of the models created using

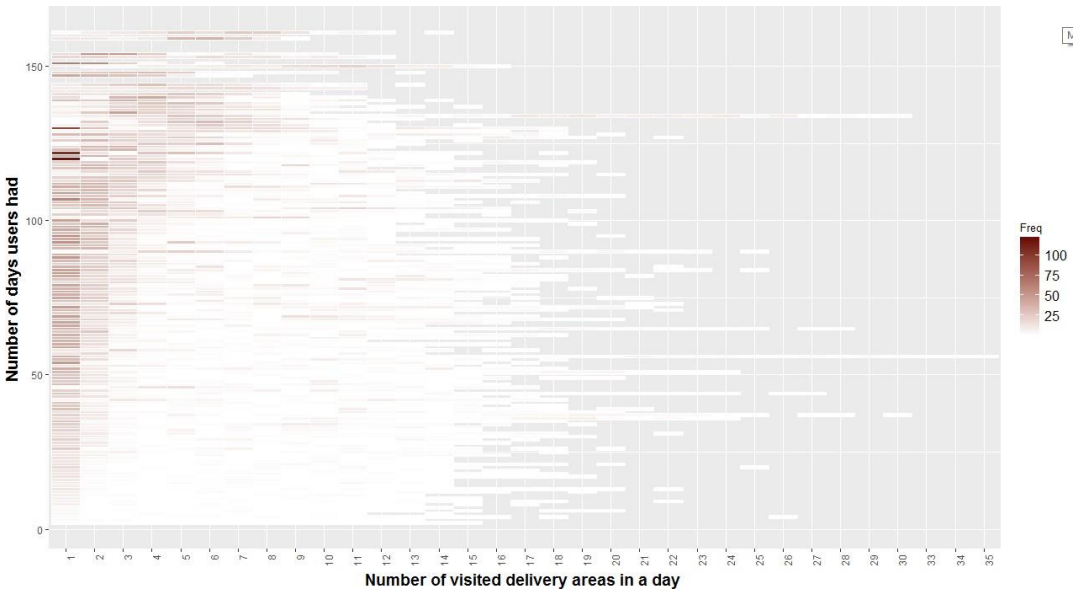


Figure 3: Distribution of check-ins frequencies by the number of delivery areas visited a day and the number of days the user had delivery

different clustering methods. Because of the reasons we pointed out in Section 3.1, model based clustering is the chosen clustering technique. The main difference among the different models was the breaking points for the intervals. *Mclust* is the only method it creates the intervals in a way that we want. Other clustering methods are so sensitive to the outliers. Thus, they put the data points with high frequency into different clusters alone.

```
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust V (univariate, unequal variance) model with 5 components:

 log.likelihood  n df       BIC      ICL
      -287.9585 33 14 -624.8681 -629.807

Clustering table:
1 2 3 4 5
7 7 6 6 7
```

Figure 4: Summary of Mclust Model

Model based clustering technique created 5 clusters as follows:

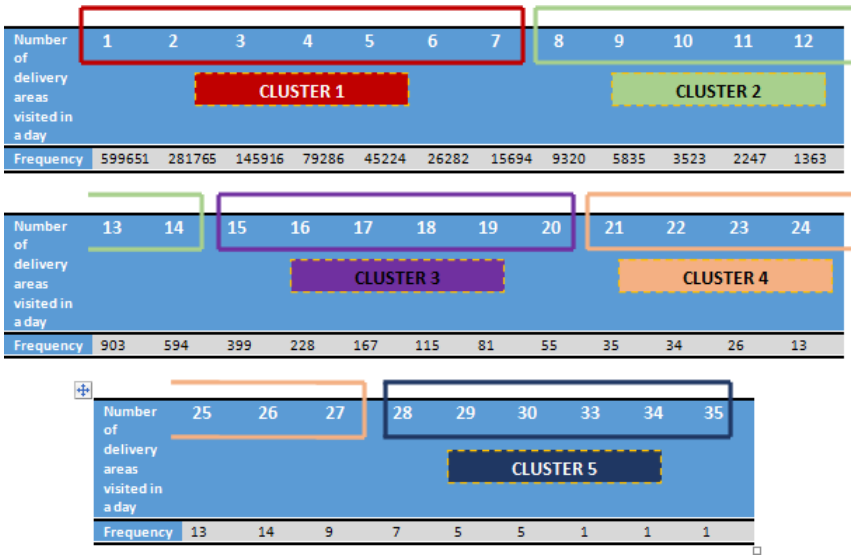


Figure 5: Clustered Number of Delivery Areas Visited in One Day

After this clustering results, we did one modification manually. We divided *Cluster 1* into two clusters, and the new version of clusters are as follows:

- Cluster 1: 1, 2, 3
- Cluster 2: 4, 5, 6, 7
- Cluster 3: 8, 9, 10, 11, 12, 13, 14
- Cluster 4: 15, 16, 17, 18, 19, 20
- Cluster 5: 21, 22, 23, 24, 25, 26, 27
- Cluster 6: 28, 29, 30, 33, 34, 35

The reason of this modification is the fact that we had put the numbers of 1, 2 and 3 into a different cluster because of insufficient dimension.

4.1.2 The Number of Loading/Unloading Areas Which Are Repeated Every Day A User Had Delivery

In this section, we only take the delivery areas which are visited **every** day a user had delivery. We do not check the sub-patterns that are repeated some days. This is the section we use our specific algorithm MLCS (Check Section 3.1).

	1,2,3	4,5,6,7	8,9,10,11,12,13,14	15,16,17,18,19,20	21,22,23,24,25,26,27	28,29,30,31,32,33,34,35
0	43514	11841	1767	64	5	1
1	1477	1427	337	22	2	1
2	122	362	142	7	2	0
3	124	99	68	7	2	0
4		3981	36	10	1	1
5		1040	13	6	0	0
6		309	8	2	3	0
7		96	2	4	0	0
8			918	2	0	0
9			317	2	0	0
10			103	0	0	0
11			32	1	0	0
12			20	0	0	0
13			9	0	0	0
14			5	0	0	0
15				55	0	0
16				24	0	0
17				14	0	0
18				8	0	0
19				3	0	0
20				1	0	0
21					3	0
22					7	0
23					0	0
24					0	0
25					1	0
26					1	0
27					0	0
28						1
29						0
30						0
31						0
32						0
33						0
34						0
35						0

Figure 6: The number of delivery areas repeated every day a user had delivery

In Figure 6, the columns represent the number of delivery areas visited in a day of users, whereas the rows represent the number of delivery areas which are repeated exactly each day when a user had delivery. The cell values represent the number of users. For the interpretation, let's take the first column as an example:

- 43514 is the number of users who never repeat a delivery area after they visited it once. Here is the restriction is the Cluster 1 with values of 1, 2 and 3.

- 1427 is the number of users who repeat only one delivery area each day they had delivery. Here is the restriction is the Cluster 1 with values of 1, 2 and 3.
- 122 is the number of users who repeat two delivery areas each day they had delivery. Here is the restriction is the Cluster 1 with values of 1, 2 and 3.
- 124 is the number of users who repeat 3 delivery areas each day they had delivery. Here is the restriction is the Cluster 1 with values of 1, 2 and 3.
- 124 is the number which tells us that there are 124 users who have fixed routes for all of the days they had delivery. They do not have extra stop or less stop for other days. Here is the restriction is the Cluster 1 with values of 1, 2 and 3.

4.1.3 The Ratio of Unrepeated Loading/Unloading Areas in Clusters

In this section, we find the probability of unrepeated loading/unloading for each user. We calculate the fraction of unrepeated areas by the total number of loading/unloading areas visited that day by the user.

The idea here is to find the probabilities of having a unique loading/unloading area in the trip plan of each user. In the previous section (Section 4.1.2), we analyzed the loading/unloading areas which are visited everyday in a fixed route by a user. Here, we do the complementary analysis which stands for the other loading/unloading areas that does not have daily repetition. These loading/unloading areas can be the unique one, or a part of a sub-pattern that is repeated some days.

The boxplot method is used to show the distribution of the data in each cluster since it is exploratory graphic. Interpretation of boxplot is as follows:

- Cluster 5 and 6 have the largest IQR values which are related to the central tendency and spread. There is more variation in Cluster 5 and 6 than the other clusters. On the other hand, Cluster 3 has the lowest IQR value.
- Range of Cluster 1 and 2 is exactly between 0 and 1. From Cluster 3 to Cluster 6, the range decreases respectively, whereas Cluster 6 has

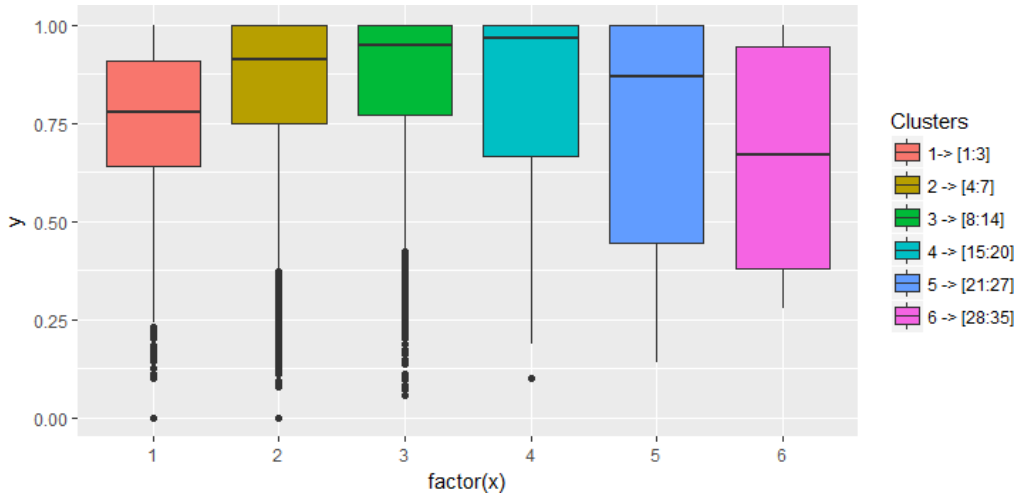


Figure 7: The Boxplot of Clustered Data

the lowest range value.

- The highest median value comes from Cluster 4, whereas the lowest one comes from Cluster 6.
- For Cluster 1 and 6, it seems that the observations are evenly split at the median (Symmetry).
- In Cluster 2, 3, 4 and 5, most of the observations are concentrated on the up end of the scale (Skew-left).

Highlights:
In general, the ratio of unrepeated delivery areas is close to 1. It indicates the fact that the users have unique delivery areas in their delivery history. Once they visit a deliver area, they are more likely not willing to revisit it in other days.

	[1:3]	[4:7]	[8:14]	[15:20]	[21:27]	[28:35]
Min:	0	0	0.5556	0.1	0.14	0.2784
Quartile 1:	0.6393	0.75	0.7692	0.6667	0.4450	0.379
Median:	0.7759	0.9130	0.9474	0.9661	0.8696	0.6692
Mean:	0.7492	0.8525	0.8542	0.8131	0.7371	0.6542
Quartile 3:	0.9091	1	1	1	1	0.9444
Max:	1	1	1	1	1	1

4.1.4 The Number of Visited Loading/Unloading Areas by Each Activity Type

In this section, we analyze the number of stops which are visited by different activity types.

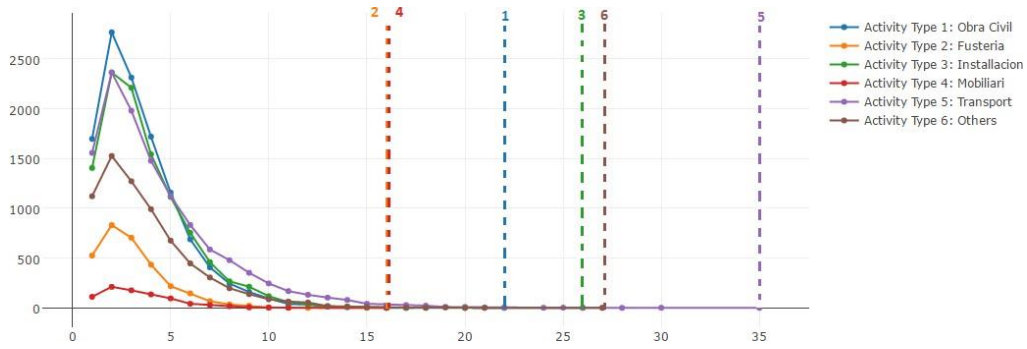


Figure 8: The number of visited delivery areas by different activity types

In Figure 8 each color represents a different activity type, which you can see in the legend of the plot. The dot lines are to highlight the point each activity types has the last value.

Highlights:

- 2 is the number that all activity types have the greatest number of visited delivery areas in a day.
- Up to number 16 we see all activity types.
- The greatest value as the representative of daily visited delivery area number pertain to *Transport (Activity Type 5)*.

If we normalized the frequencies by their corresponding data length, it gets easier to see which types have the lead by the percentage.

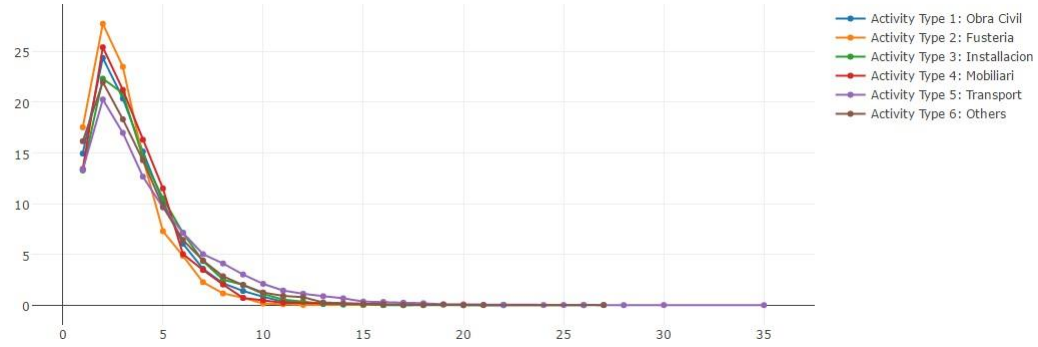


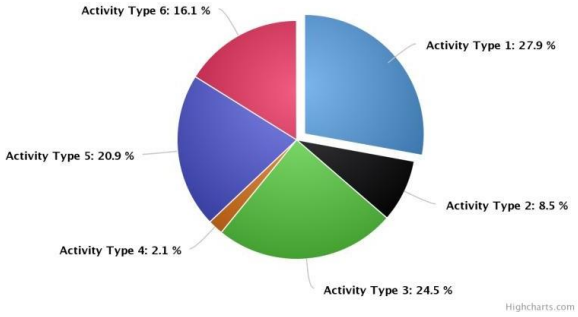
Figure 9: The normalized number of visited delivery areas by different activity types

As the next step, we illustrated the distribution of activity types in each cluster we have detected before. If you check the piecharts in the next page, you can see that *Transport (Activity Type 5)* increases the size of the slice in each cluster from Cluster 1 to Cluster 6. The observation here is that *Transport (Activity Type 5)* is the one that has more loading/unloading areas in its corresponding users' routes than the other types.

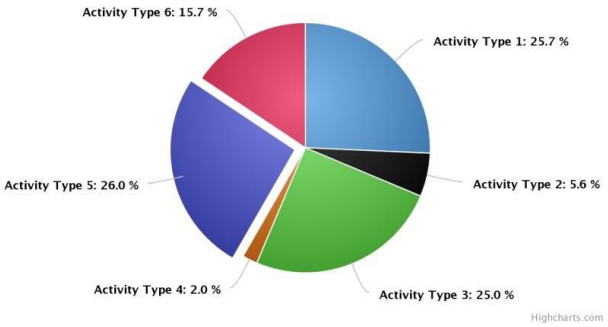
The only cluster that *Transport (Activity Type 5)* does not have the largest slice in is Cluster 1 where *Obra Civil (Activity Type 1)* has the largest slice.

In Figure 9, it can be seemed that by the in-type-percentage, *Fusteria (Activity Type 2)* has the largest percentage for the number of 2.

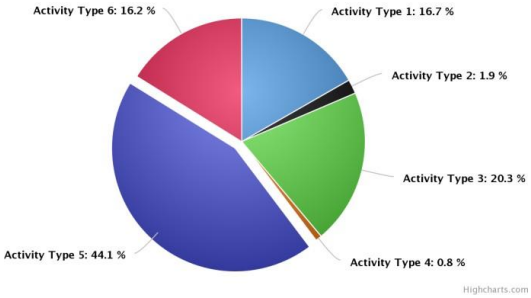
The Activity Types of The User Patterns Consist of Maximum 1, 2 or 3 Visited Delivery Areas in A Day



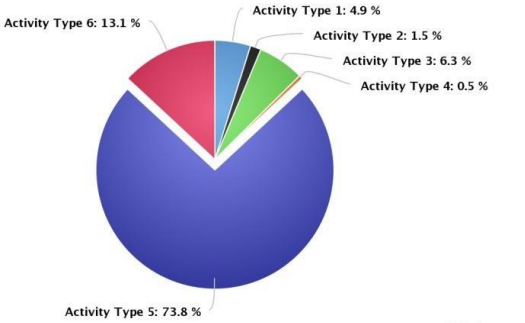
The Activity Types of The User Patterns Consist of Maximum 4, 5, 6 or 7 Visited Delivery Areas in A Day



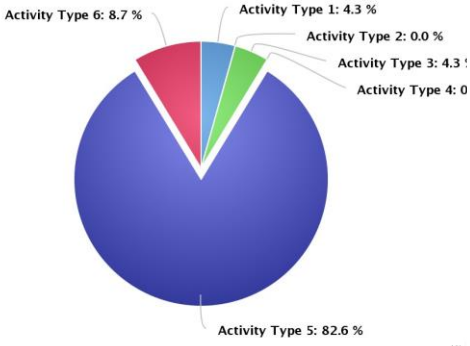
The Activity Types of The User Patterns Consist of Maximum 8, 9, 10, 11, 12, 13 or 14 Visited Delivery Areas in A Day



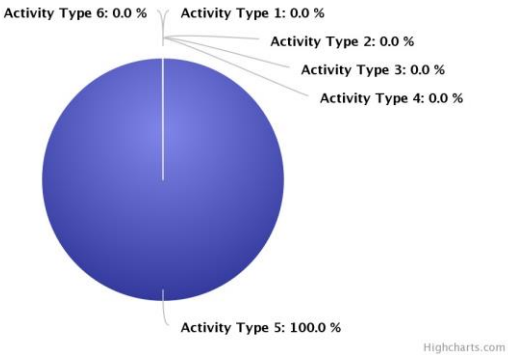
The Activity Types of The User Patterns Consist of Maximum 15, 16, 17, 18, 19 or 20 Visited Delivery Areas in A Day



The Activity Types of The User Patterns Consist of Maxim 21, 22, 23, 24, 25, 26 or 27 Visited Delivery Areas in A D



The Activity Types of The User Patterns Consist of Maximum 28, 29, 30, 31, 32, 33, 34 or 35 Visited Delivery Areas in A Day



4.1.5 The Haversine Distance Between The First and The Last Stop of Daily Routes

From the dataset, the first and the last stops for each user's daily route are extracted. Haversine is used method for the great circle distance (The reason of this choice has explained in the previous reports.). The main goal is to see if the users' starting and ending point for a day are the different points far away or same spots.

First of all, we grouped the distances into bins. Each bin represents 100 meters. We have an additional bin as the representative of 0 (zero) meter. Which basically says that the deliverer ended his journey at the same spot he started.The groups are created as follows:

- Group 0: 0 meters,
- Group 1: 0-100 meters,
- Group 2: 100-200 meters,

  :

- Group 97: 9600-9700 meters.

In Figure 10, there are 603405 daily routes which start and end at the same delivery area.

After the frequencies are calculated, we wanted to see which type of activity have which kind of distance between the first and the last stops in their daily route. The general distribution of activity types in daily routes without considering any distance as follows:

- Activity Type 1: 24.5 %
- Activity Type 2: 6 %
- Activity Type 3: 23.4 %
- Activity Type 4: 1.4 %
- Activity Type 5: 29.8 %
- Activity Type 6: 14.9 %

Category	x				
0	603405	35	6931	70	462
1	14842	36	6641	71	408
2	48428	37	6327	72	356
3	22860	38	5932	73	347
4	19035	39	5486	74	295
5	20487	40	5036	75	263
6	20889	41	4775	76	211
7	19598	42	4345	77	193
8	17888	43	4057	78	206
9	18774	44	3727	79	127
10	18269	45	3647	80	94
11	17328	46	3238	81	96
12	16801	47	3138	82	79
13	17000	48	2804	83	86
14	16512	49	2601	84	61
15	15342	50	2370	85	51
16	15164	51	2182	86	57
17	14913	52	2032	87	38
18	14889	53	1834	88	46
19	14024	54	1589	89	34
20	13466	55	1490	90	24
21	13633	56	1371	91	15
22	12981	57	1313	92	12
23	12301	58	1251	93	9
24	12026	59	1092	94	11
25	11866	60	1008	95	4
26	11475	61	990	96	3
27	10920	62	878	97	1
28	10278	63	852		
29	9678	64	751		
30	9428	65	709		
31	8834	66	688		
32	8163	67	628		
33	8011	68	609		
34	7645	69	542		

Figure 10: The frequency distribution of the distances between the first stops and the last stops (in meters)

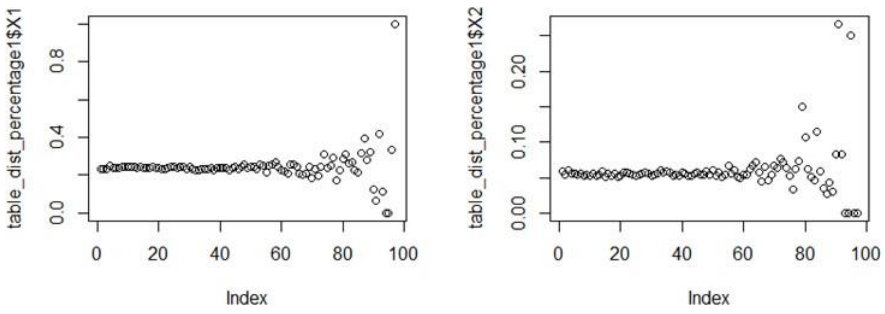


Figure 11: Activity Type 1 and Activity Type 2 - Clustered distance between the first and the last stop of daily routes

In Figure 11,

- Activity Type 1 has 24.5% of the daily routes. It follows a straight trend from *Group 0* (0 meter) to *Group 60* (5900-6000 meters). It means that this activity type has the same percentage for the daily routes' distance from 0 to 6000 meters. After *Group 60* it starts to become varied. There is one group consists of only Activity Type 1.
- Activity Type 2 has 6% of the daily routes. This activity type has the same percentage for the daily routes' distance from 0 to 6000 meters. After *Group 60* it starts to become varied. There are some groups that do not include any Activity Type 2.

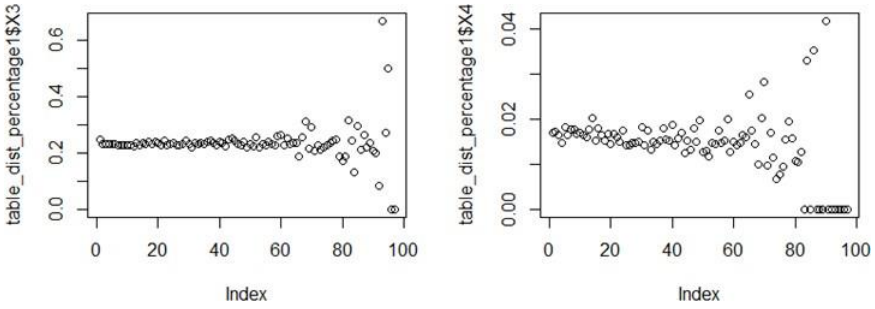


Figure 12: Activity Type 3 and Activity Type 4 - Clustered distance between the first and the last stop of daily routes

In Figure 12,

- Activity Type 3 has 23.4% of the daily routes. This activity type has the same percentage for the daily routes' distance from 0 to 6000 meters. After *Group 60* it starts to become varied. There are some groups that do not include any Activity Type 3.
- Activity Type 4 has 1.4% of the daily routes. This activity type is the one which varies more than other types. However, it is because of the low dimension of it.

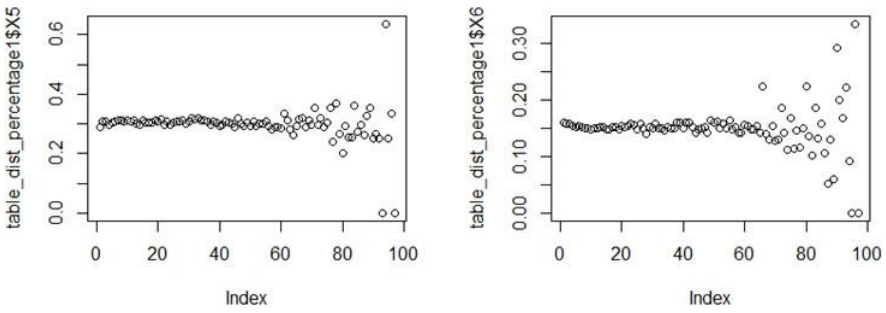


Figure 13: Activity Type 5 and Activity Type 6 - Clustered distance between the first and the last stop of daily routes

In Figure 13,

- Activity Type 5 has 29.8% of the daily routes. This activity type has the same percentage for the daily routes' distance from 0 to 6000 meters. After *Group 60* it starts to become varied. There are a few groups without this type.
- Activity Type 6 has 14.9% of the daily routes. This activity type has the same percentage for the daily routes' distance from 0 to 6000 meters. After *Group 60* it starts to become varied. There are some groups that do not include any Activity Type 6.

4.1.6 The Barrio Change Between First and Last Check-in in Daily Routes

First of all, we focused on the *Group 0* (0 meter) in order to see the barrios' frequency for unchanged destination and arrival stops.

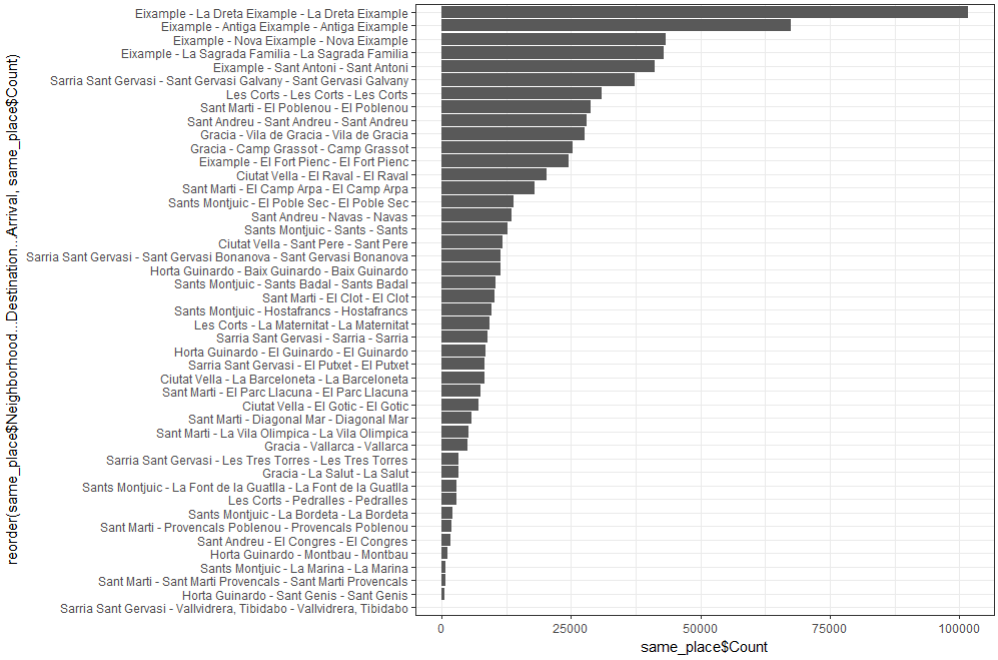


Figure 14: The Barrios' Frequency for being used for both destination and arrival point in daily routes

Figure 14 shows us that the daily routes started and ended in the same neighborhood happens mostly in La Dreta de'l Eixample and Antiga de'l Eixample neighborhoods. The reason can be that users here in these neighborhoods have a specific point that they need to start and end their journey. For instance, it can be a company they need to report, it can be a local store the user had to come back and etc.

The neighborhoods at the bottom of Figure 14 are the neighborhoods with the lowest frequency of check-ins. Thus, there is no unexpected result there.

In Figure 15, it can be seemed that the routes from one neighborhood of Eixample ends in another neighborhood in Eixample more frequently than the case it ends in another district's neighborhood.

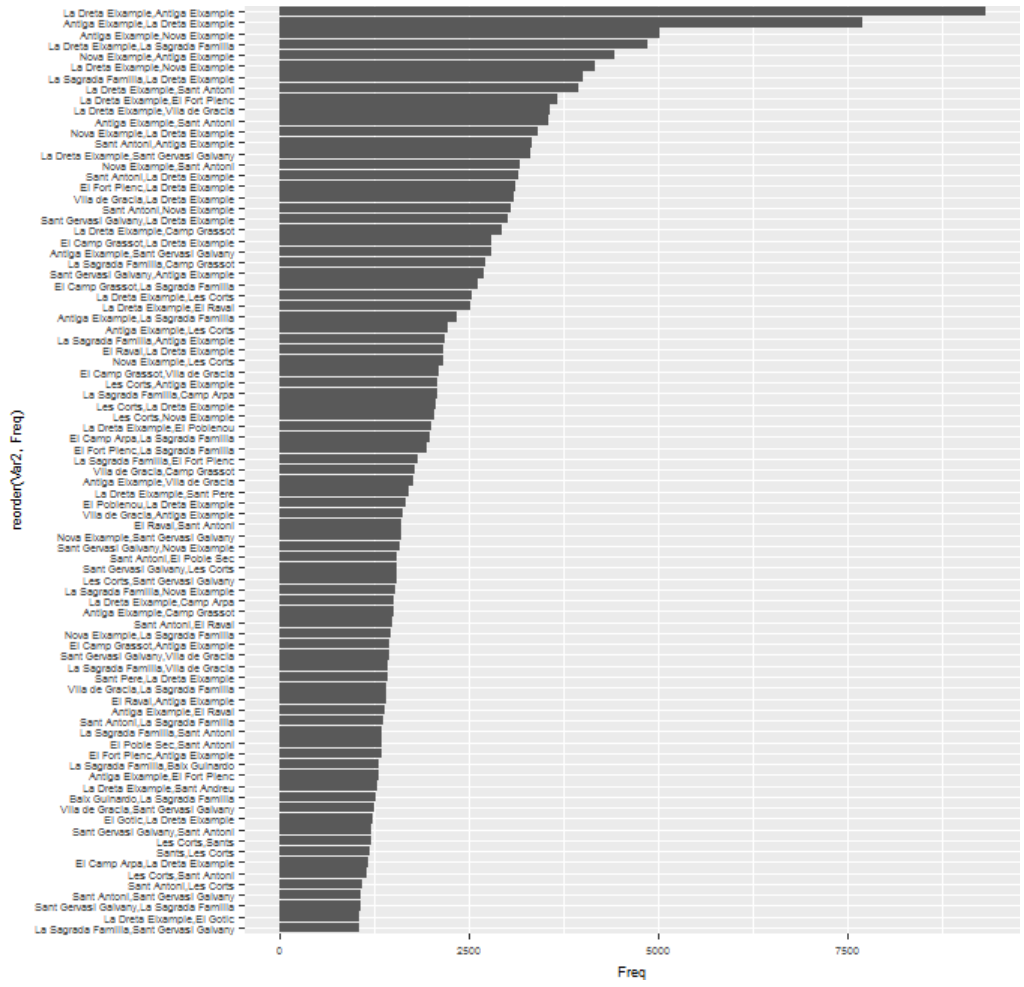


Figure 15: The Barrio pairs which have more routes than 1000 starts from one and ends in another

4.1.7 The Time Difference Between The First and The Last heck-ins in Daily Routes

```
        Min.          1st Qu.          Median            Mean         3rd Qu.            Max.
0.0000000000  0.0000000000  15.1833333333  121.4134194340  210.1166666670  1257.6833333300
```

Figure 16: Summary output of time differences (in minutes) between the first and the last stop in daily routes

In Figure 17, it can be seemed that the average time spent between the first and the last stops of daily routes is 15.18 minutes. However, there is a significant difference between mean and median. This leads us to the fact that there are outlier points in time difference data.

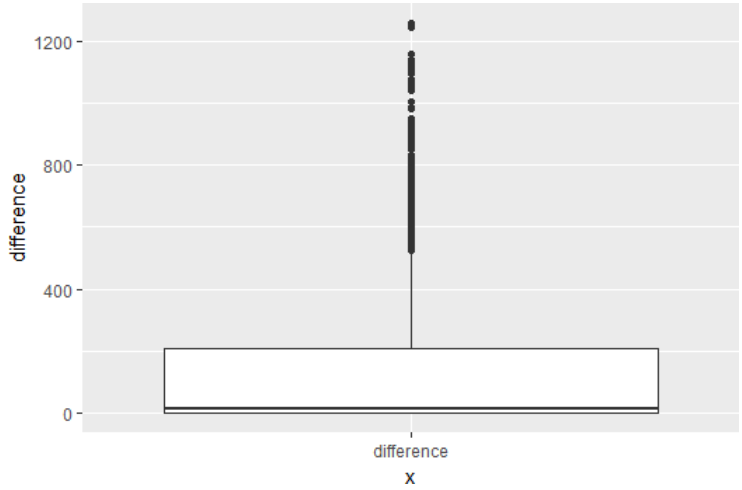


Figure 17: Boxplot representation of time differences between the first and the last stops of daily routes

The most important observation from Figure 16 and Figure 17 is the fact that the minimum value of time difference is zero. The reason is the data includes the daily routes with only one stop. We eliminate these corresponding rows from data in order to have a clean sight for the time spent between the first and the last stops of daily routes.

```
          Min.          1st Qu.           Median             Mean          3rd Qu.             Max.
0.283333333333  88.533333333300  206.216666667000  239.538776139000  371.400000000000 1257.683333330000
```

Figure 18: Summary output of time differences (in minutes) between the first and the last stop in daily routes after rows with value of zero removed

Figure 18 is the output after the zero time difference elimination. As we can see the results make much more sense. The mean value reach to 239.54 minutes which is around 4 hours.

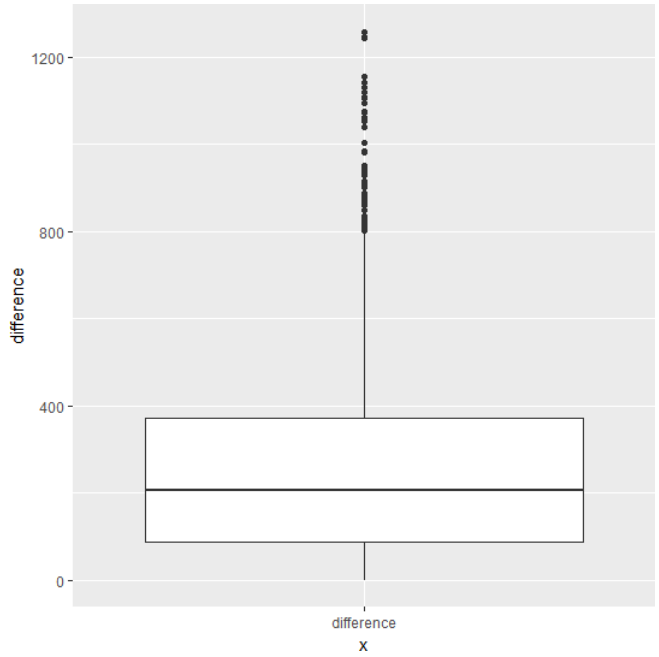


Figure 19: Boxplot representation of time differences between the first and the last stops of daily routes after the zero time difference elimination

If we compare Figure 17 and Figure 19 we can see that the range between first and third quartile is increased, and the total number of outliers is decreased.

Since we have float format numbers, it is not easy to have an interpretation easily. For this reason, we grouped the time differences between the firsts and the last stops of daily routes. We created 26 groups starts with *Group 0* and ends with *Group 25*. The groups are as follows:

- Group 0: (0, 1] (in minutes)
- Group 1: (1, 5] (in minutes)
- Group 2: (5, 15] (in minutes)
- Group 3: (15, 30] (in minutes)

- Group 4: (30, 45] (in minutes)
- Group 5: (45, 60] (in minutes)
- Group 6: (60, 120] (in minutes)
- Group 7: (120, 180] (in minutes)
- Group 8: (180, 240] (in minutes)

  .

- Group 25: (1200, 1260] (in minutes)

Group 0 is created for the time difference less than or equal to 1 minute. This group basically represents the mistaken check-ins. The user probably wrote the wrong delivery area id while making a check-in, and the user had only one delivery area in his list in that day. From *Group 1* to *Group 5*, the bins are created for each 15 minutes. After that, from *Group 6* to *Group 25*, all the bins are created hourly. The bin creation started with the minimum value and ended with the maximum value.

The next step is to check the activity type distribution for each group.

	1	2	3	4	5	6
0	6	1	1	0	3	2
1	276	49	269	20	306	165
2	1065	220	881	64	3735	1111
3	4831	1046	4140	318	11879	4358
4	7663	1749	6963	536	12871	6070
5	6969	1661	6657	545	11042	5163
6	22552	4886	20824	1553	35391	15703
7	17178	3491	16096	1099	27538	11595
8	14717	2759	13923	922	22862	9698
9	13412	2356	13369	743	19295	8618
10	12413	2209	13205	700	16184	7630
11	12126	2173	13426	678	14213	7141
12	11026	1977	12644	586	12518	6426
13	8900	1384	10193	504	9211	4734
14	5546	896	6360	350	5628	3164
15	2773	388	2749	131	2584	1653
16	654	104	568	31	583	481
17	7	6	26	2	17	19
18	3	0	5	1	10	8
19	0	0	0	0	5	3
20	2	0	3	1	2	1
21	1	0	0	0	1	1
22	2	0	0	0	2	2
23	1	0	0	1	1	3
24	0	0	0	0	0	2
25	2	0	0	0	0	2

Figure 20: The distribution of activity types in time difference groups

When we look at Figure 20, we can see that there is no different trend among activity type except their frequency. *Group 6* is the one with most

frequently appeared for all activity types. In order to ensure we plot the frequency table.

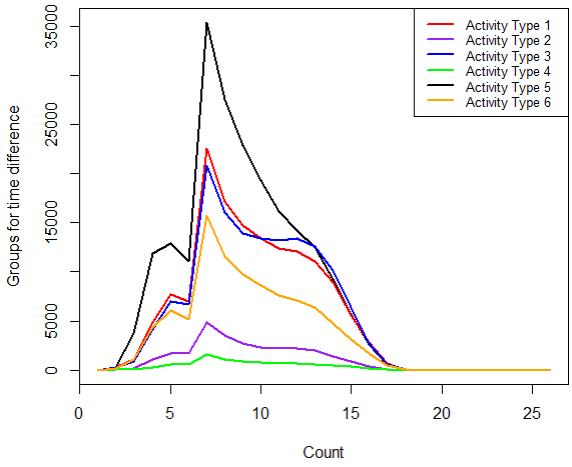


Figure 21: The distribution of activity types in time difference groups

In Figure 21, all the activity types' distribution based on groups have the peaks at same points with different number of frequencies. Each of them has the biggest peak at the *Group 6* which is the representative of (60, 120]. It is the time difference between the first and the last stop, which varies from one hour to two hours.

4.2 Based on Weekly Delivery

In this section, we analyze the routes of users in order to understand how the users' routes differentiate weekly. The aggregation is done by combining daily routes into weeks.

```
          Min.        1st Qu.         Median           Mean        3rd Qu.           Max.
0.000000000000 0.634146341463 0.780487804878 0.756894764635 0.967741935484 1.000000000000
```

Figure 22: The summary output of unrepeated loading/unloading areas for weekly routes

In Figure 22, the mean value is around 76%. It means that 76% of the loading/unloading areas are not repeated in weekly routes. It is a quite high percentage than our expectation.

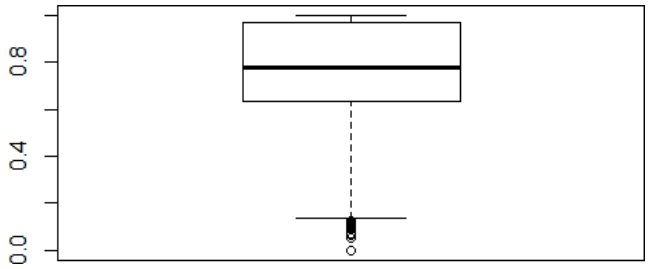


Figure 23: The ratio of unrepeated loading/unloading areas in weekly routes

Figure 23 also proves that most of the data points are located in the upper scale close to 1. All of the outlier points are located close to zero. These points represent the loading/unloading areas which are repeated in the weekly routes.

```
same_list_week
    0     1     2     3     4     5     6     7     8     9    10    11    12    13    14    15    16    17
39426  7087  1455   546   229   131    76    61    44    30    30    13    13    22    11    12     9     8
   18    19    20    21    22    23    24    26    27    30    31    33    34    35
    6     5     5     1     3     5     2     1     3     1     1     2     1     1
```

Figure 24: The frequency table of the common subsequence for each week

As we can see in Figure 24, most of the loading/unloading areas are not repeated every week.

4.3 Based on Monthly Delivery

In this section, we analyze the routes of users in order to understand how the users' routes differentiate monthly. The aggregation is done by combining daily routes into months.

```
          Min.        1st Qu.         Median           Mean        3rd Qu.           Max.
0.000000000000 0.631578947368 0.777777777778 0.753824088569 0.941176470588 1.000000000000
```

Figure 25: The summary output of unrepeated loading/unloading areas for monthly routes

In Figure 25, the mean value is around 75%. It means that 75% of the loading/unloading areas are not repeated in weekly routes.

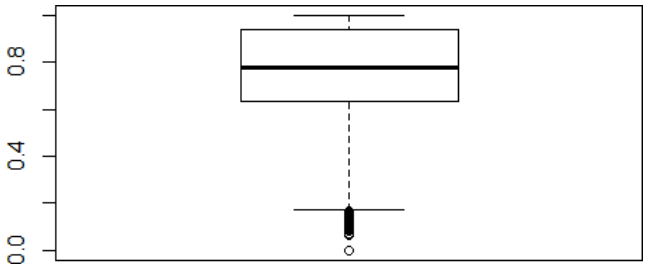


Figure 26: The ratio of unrepeated loading/unloading areas in monthly routes

Figure 26 also proves that most of the data points are located in the upper scale close to 1. All of the outlier points are located close to zero. These points represent the loading/unloading areas which are repeated in monthly routes.

same_list_month

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
27078	11928	4314	2136	1280	854	695	582	445	428	388	296	263	251	226	189	188	160
18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
108	97	83	73	62	51	46	41	35	34	27	34	33	29	20	25	20	21
36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53
27	17	21	26	18	22	16	12	22	15	23	13	21	20	15	12	11	10
54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	71	72
15	17	19	9	14	9	5	5	7	7	5	6	7	6	6	2	2	3
73	75	76	77	78	80												
2	4	1	2	2	1												

Figure 27: The frequency table of the common subsequences for each month

In Figure 27, there are more numbers in the frequency table and it is expected since we combined the daily routes into monthly bins. However, there are still a lot of unrepeated common subsequences.

5 Conclusion

In daily, weekly and monthly routes, the users generally tend to not repeat the delivery areas they already visited. It is totally unexpected since we assumed that users would have fixed routes that they were assigned. We cannot see any significant hallmark for activity types in any kind of analysis, but the percentage of the check-ins divided into clusters in daily route analysis. Activity types have similar trends with different frequency numbers. This fact shows us that different activity types do not create their own characteristics different from each other.

Els continguts d'aquesta publicació estan subjectes a una llicència de **Reconeixement – No comercial – Sense Obra Derivada (by-nc-nd)** amb finalitat no comercial i sense obra derivada. Es permet copiar i redistribuir el material en qualsevol mitjà i format, sempre que no tingui finalitats comercials i no es distribueixi cap obra derivada.
La llicència completa es pot consultar a http://creativecommons.org/licenses/by-nc-nd/4.0/deed.ca