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.

![1-Time Users](image1)

**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.

![Activity Type Distribution of 1-time Users](image2)
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](image-url)
In Figure 2, the letters are the representatives of delivery areas. \( S_1 \) and \( S_2 \) 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 \( N - \text{dimensional} \) 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:

- \( \mu_k \): mean vector
- \( \Sigma_k \): 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 $\mu_k$, with increased density for points near the mean. Geometric features (shape, volume, orientation) of each cluster are determined by the covariance matrix $\Sigma_k$. Different possible parameterizations of $\Sigma_k$ 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 clustering 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 sub-routes 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
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:**

<table>
<thead>
<tr>
<th>log.likelihood</th>
<th>n</th>
<th>df</th>
<th>BIC</th>
<th>ICL</th>
</tr>
</thead>
<tbody>
<tr>
<td>-287.9585</td>
<td>33</td>
<td>14</td>
<td>-624.8631</td>
<td>-629.807</td>
</tr>
</tbody>
</table>

**Clustering table:**

<table>
<thead>
<tr>
<th>1 2 3 4 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 7 6 6 7</td>
</tr>
</tbody>
</table>

---

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

Figure 4: Summary of Mclust Model
Model based clustering technique created 5 clusters as follows:

![Clustered Number of Delivery Areas Visited in One Day](image)

- **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).

![Figure 6: The number of delivery areas repeated every day a user had delivery](image)

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 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 delivery area, they are more likely not willing to revisit it in other days.
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.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Min:</td>
<td>0</td>
<td>0</td>
<td>0.5556</td>
<td>0.1</td>
<td>0.14</td>
<td>0.2784</td>
</tr>
<tr>
<td>Quartile 1:</td>
<td>0.6393</td>
<td>0.75</td>
<td>0.7692</td>
<td>0.6667</td>
<td>0.4450</td>
<td>0.379</td>
</tr>
<tr>
<td>Median:</td>
<td>0.7759</td>
<td>0.9130</td>
<td>0.9474</td>
<td>0.9661</td>
<td>0.8696</td>
<td>0.6692</td>
</tr>
<tr>
<td>Mean:</td>
<td>0.7492</td>
<td>0.8525</td>
<td>0.8542</td>
<td>0.8131</td>
<td>0.7371</td>
<td>0.6542</td>
</tr>
<tr>
<td>Quartile 3:</td>
<td>0.9091</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.9444</td>
</tr>
<tr>
<td>Max:</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

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

![Figure 10: Distance Distribution](image)

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

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.

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.

![Boxplot representation of time differences between the first and the last stops of daily routes after the zero time difference elimination](image)

**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.

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>276</td>
<td>49</td>
<td>269</td>
<td>20</td>
<td>306</td>
</tr>
<tr>
<td>2</td>
<td>1065</td>
<td>259</td>
<td>881</td>
<td>64</td>
<td>3735</td>
</tr>
<tr>
<td>3</td>
<td>453</td>
<td>1046</td>
<td>4140</td>
<td>318</td>
<td>12579</td>
</tr>
<tr>
<td>4</td>
<td>7629</td>
<td>1741</td>
<td>6643</td>
<td>536</td>
<td>12172</td>
</tr>
<tr>
<td>5</td>
<td>6469</td>
<td>1461</td>
<td>8557</td>
<td>586</td>
<td>11542</td>
</tr>
<tr>
<td>6</td>
<td>5552</td>
<td>4556</td>
<td>20024</td>
<td>1550</td>
<td>30891</td>
</tr>
<tr>
<td>7</td>
<td>17178</td>
<td>2461</td>
<td>9029</td>
<td>1099</td>
<td>25209</td>
</tr>
<tr>
<td>8</td>
<td>14717</td>
<td>2789</td>
<td>19928</td>
<td>932</td>
<td>22862</td>
</tr>
<tr>
<td>9</td>
<td>13412</td>
<td>2256</td>
<td>13369</td>
<td>743</td>
<td>19295</td>
</tr>
<tr>
<td>10</td>
<td>12413</td>
<td>2209</td>
<td>13205</td>
<td>700</td>
<td>16184</td>
</tr>
<tr>
<td>11</td>
<td>12124</td>
<td>2173</td>
<td>13082</td>
<td>670</td>
<td>16215</td>
</tr>
<tr>
<td>12</td>
<td>11052</td>
<td>1877</td>
<td>12644</td>
<td>556</td>
<td>12731</td>
</tr>
<tr>
<td>13</td>
<td>8900</td>
<td>1389</td>
<td>10193</td>
<td>594</td>
<td>9211</td>
</tr>
<tr>
<td>14</td>
<td>5546</td>
<td>956</td>
<td>6560</td>
<td>550</td>
<td>5628</td>
</tr>
<tr>
<td>15</td>
<td>2773</td>
<td>389</td>
<td>2749</td>
<td>131</td>
<td>2594</td>
</tr>
<tr>
<td>16</td>
<td>654</td>
<td>104</td>
<td>568</td>
<td>31</td>
<td>583</td>
</tr>
<tr>
<td>17</td>
<td>7</td>
<td>6</td>
<td>26</td>
<td>2</td>
<td>17</td>
</tr>
<tr>
<td>18</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>19</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>21</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>22</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>23</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>25</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

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.

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.

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.

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.

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.