INTRODUCTION & DATA DESCRIPTION

The data set is used by the consideration of importing "circle_id"s instead of "tramo_id"s. The daily, weekly and monthly routes are founded by using the same algorithm with a small change for "circle_id". (The only case we could not use "circle_id"s is "First and Last Stops Analysis"). The dimension of routes' data:

- Daily Routes: 1048575 routes are detected. (file= **daily_routes.xlsx**)
- Weekly Routes: 509925 routes are detected. (file= **weekly_routes.xlsx**)
- Monthly Routes: 195563 routes are detected. (file= **monthly_routes.xlsx**)

In the data set, two more columns are created; one for "week" and one for "month". For the week, 29 weeks are created for the data from 1st of January, 2016 to 15th of July, 2016. Encoding is done such a way that first week takes "Week1" label, whereas the last week takes "Week29". The "month" column is created using "FHSTART" column, where the operation is automatically done.

These files consist of 3 columns. The first column is for "user_id", the second column is for "day" or "week" "month (in number format)" and the third column is for the "circle_id"s which are visited by users in a day, a week or a month.

The new version of check-ins data used is **checkin_data_modified.xlsx**, and it includes both week and month columns.

DATA CLEANSING

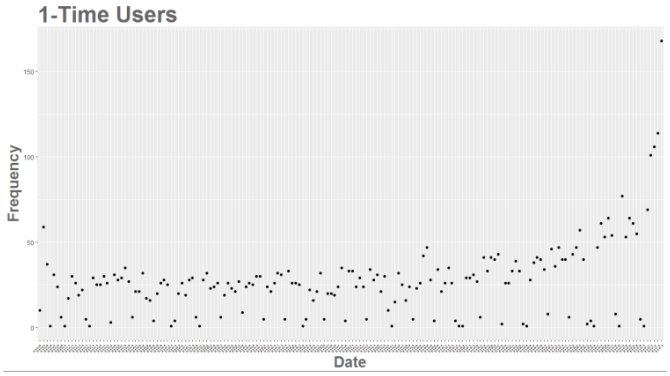


Figure 1

The plot on the left represents the increase of 1-time users from January 1, 2016 to July 15, 2016. The interpretation has to be done carefully. The 1-time users on the right side of the plot can be misleading. These are the points needed to be considered as 1-time users for now. In the future, these users can have more check-ins. For now, **3754 out of 49172** users **(7.634 %)** are labelled as 1-time users.

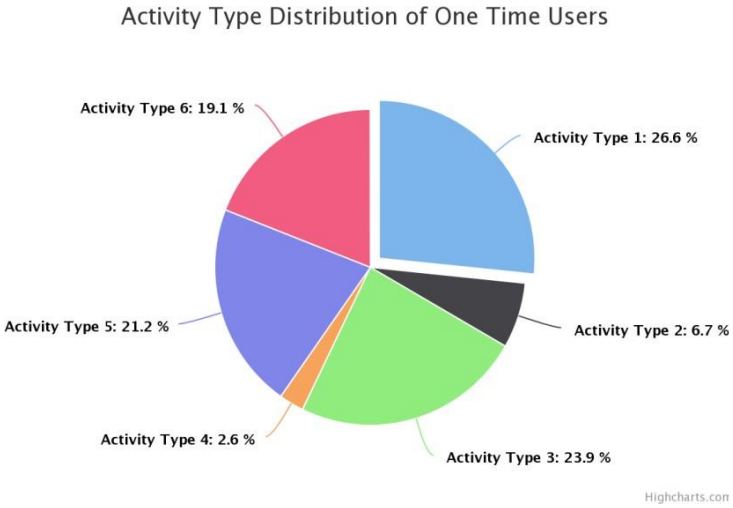


Figure 2.

The plot on the left represents the activity type distribution of 1-time users.

(File= **act_dist_one_time_users.xlsx**)

Activity Type 1, 3 and 5 are in the lead for having one time users. The activity type distribution in 1-time users is so similar the one we got for the activity type distributions in disallowed repeated check-ins in general.

DAILY ROUTES PLOTS

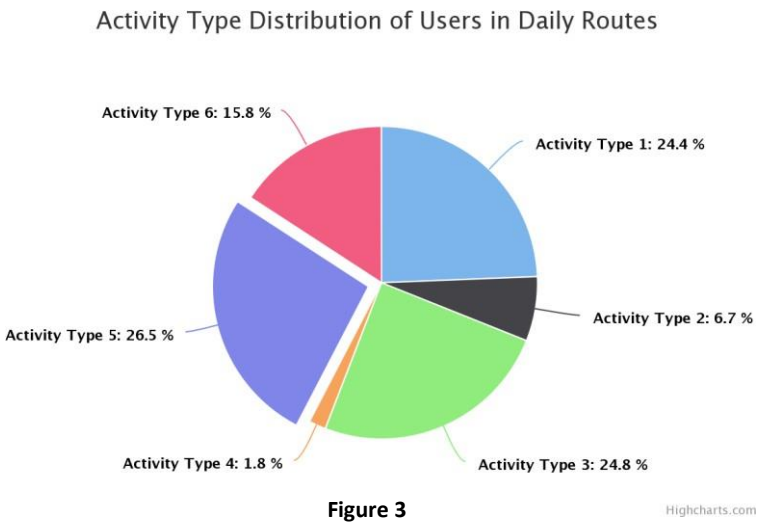


Figure 3

The plot on the left is to illustrate the activity type distribution of users in daily routes. Since the data we have is after cleansing does not contain any 1-time users, we check the activity type distribution in the rest of users we have in modified data. To interpret, e.g. 26.5 % of users who appeared in daily routes belongs to "Activity Type 5". The activity type distribution here is similar to the one we got for activity type distribution in general check-ins data.

File = **act_dist_user_daily_route.xlsx**

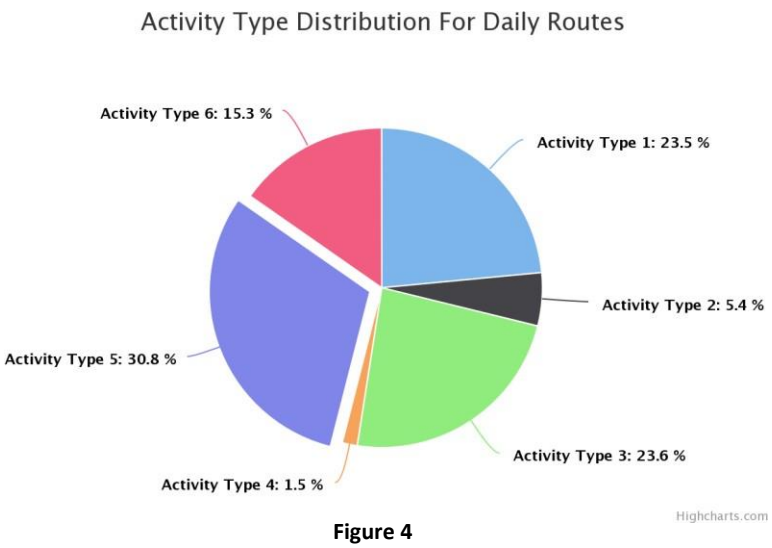


Figure 4

The plot on the left is to illustrate the activity type distribution of routes in daily data. One user can have several check-ins in a day, and also several days in general. If a user goes for a delivery in 5 different days, then we have 5 daily routes from these users. When we associate this user to his/her corresponding activity type, then his/her activity type will occur 5 times in the frequency table. This is the way we create the frequency table for activity types in daily routes. Unsurprisingly, this distribution is similar the one we got for activity distribution in general checkins data.

File=**act_dist_general_daily_routes.xlsx**

Additional Interpretation of Figure 3 and Figure 4

- Activity Type 1 users have less daily routes percentage (Figure 4) than the expectation comes from their user proportion (Figure 3). This type has 24.4 % of users in daily routes, whereas they have 23.5 % of daily routes.
- Activity Type 2 users have less daily routes percentage than the expectation comes from their user proportion. This type has 6.7 % of users in daily routes, whereas they have 5.4 % of daily routes.
- Activity Type 3 users have less daily routes percentage than the expectation comes from their user proportion. This type has 24.8 % of users in daily routes, whereas they have 23.6 % of daily routes.
- Activity Type 4 users have less daily routes percentage than the expectation comes from their user proportion. This type has 1.8 % of users in daily routes, whereas they have 1.5 % of daily routes.
- **Activity Type 5 users have more daily routes percentage than the expectation comes from their user proportion. This type has 26.5 % of users in daily routes, whereas they have 30.8 % of daily routes**.
- Activity Type 6 users have less daily routes percentage than the expectation comes from their user proportion. This type has 15.8 % of users in daily routes, whereas they have 15.3 % of daily routes.

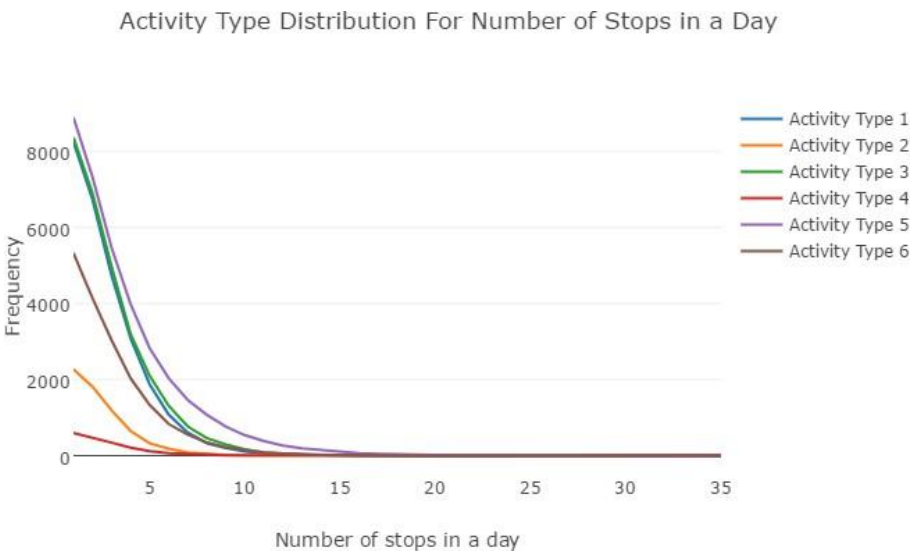


Figure 5

File= **act_dist_numberofstops.xlsx**

In Figure 5, x-axis represents the number of stops (circle ids) visited in a day by users, whereas the y-axis represents the counted numbers of check-ins come from specific number of stops (circle ids).

In the plot, it can be seemed that users tend to have only 1 circle id for their daily route.

Let's focus on "5" on x-axis. Here, "5" is the number of stops (circle ids) in daily routes. It is basically after the daily route computation, we subset the data by the number of stops (circle ids) in daily routes. For instance,

(This is an example, not related to the real data)

INDEX	USER ID	DATE	NUMBER OF STOPS
1	U1	01/01/2016	10
2	U1	02/01/2016	5
3	U1	03/01/2016	6
4	U1	04/01/2016	5
5	U2	02/01/2016	3
6	U2	03/01/2016	5
7	U2	07/01/2016	5

Subset process is done by this column

The range for the number of stops (circle ids) in daily routes is from 1 to 30 without skipping any number ([1, 30]), and then jumps from 30 to 33. From 33 to 35, it continues without skipping 34. In a nutshell,

$$\mathbf{1, 30} \cup \mathbf{33, 35}$$

For each number in this range, we create a subset. Let's go back to the starting point in our example, the number of stops (circle ids) in daily routes, which is "5". In here, we can subset the data for "5".

INDEX	USER ID	DATE	NUMBER OF STOPS
2	U1	02/01/2016	5
4	U1	04/01/2016	5
6	U2	03/01/2016	5
7	U2	07/01/2016	5

Here we take the number of rows in order to have the frequency of "5". (In our example, the frequency is 4). If we look at Figure 5 again, for "5" on x-axis," Activity Type 5 has the greatest number. In other words, Activity Type 5 is the type which has more daily routes consist of only 5 stops (circle ids).

The difference between delivery area stops and circle id data:

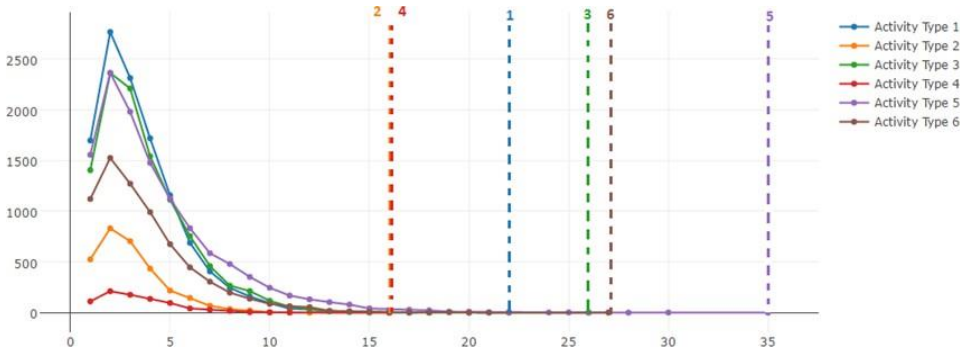


Figure 6

The plot on the left is for the number of stops in daily routes without the circle id consideration. **Each stop is a "tramo".** Maximum number of stops in daily routes for each activity type is as follows:

- Activity Type 1: 22 max
- Activity Type 2: 16 max
- Activity Type 3: 26 max
- Activity Type 4: 16 max
- Activity Type 5: 35 max
- Activity Type 6: 27 max

Conclusion: Each types' users tend to have 2 tramos in their daily delivery visits.

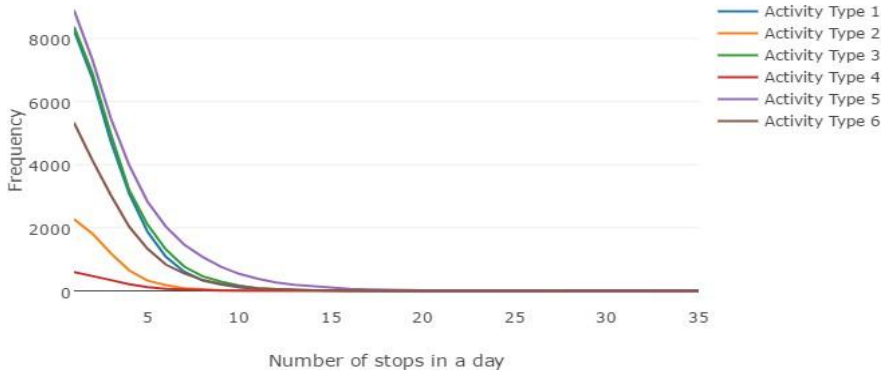


Figure 7

The plot on the left is for number of stops in daily routes with the circle id consideration. **Each stop is a "circle_id".** Maximum number of stops in daily routes for each activity type is as follows:

- Activity Type 1: 22 max
- Activity Type 2: 16 max
- Activity Type 3: 20 max
- Activity Type 4: 16 max
- Activity Type 5: 35 max
- Activity Type 6: 21 max

Conclusion: Each types' users tend to 1 circle in their daily delivery visits. The possible reason for the decrease in the percentage of Activity Type 3 and 6 can be the fact that the users from these types make stops (tramos) in the same circles separately.

Clusters	Activity Types	Repetition	Real Number	Ratio
1	1	7580	0	0
1	2	2076	0	0
1	3	7783	0	0
1	4	532	0	0
1	5	8165	0	0
1	6	4816	0	0
1	1	837	1	1
1	2	227	1	1
1	3	787	1	1
1	4	79	1	1
1	5	993	1	1
1	6	649	1	1
2	1	6786	0	0
2	2	1902	0	0
2	3	7063	0	0
2	4	498	0	0
2	5	7469	0	0
2	6	4459	0	0
2	1	410	1	0.5
2	2	83	1	0.5
2	3	251	1	0.5
2	4	19	1	0.5
2	5	467	1	0.5
2	6	273	1	0.5
2	1	1221	2	1
2	2	318	2	1
2	3	1256	2	1
2	4	94	2	1
2	5	1222	2	1
2	6	733	2	1

Figure 8

File= **act_dist_for_each_cluster_number_of_stops_daily. xlsx**

It is an instance from the data set which holds the information about the activity type distribution in clusters (each cluster represents the number of stops (circle ids) exactly repeated everyday). In the data table;

- **"Clusters" column**: Represents the cluster number from 1 to 35. These numbers are the stop (circle id) numbers which are exactly repeated everyday a user had a delivery.
- **"Activity Types" column**
- **"Real Number" column**: The number of stops (circle ids) repeated exactly. Where the "Real Number" is 0, then the "Repetition" values for each activity type represent the total number.
- **"Ratio" column:** This column holds the ratio between the cluster number and real number. For instance, in the first row, the ratio is 0. It is because of (Real number /Clusters value) 0/1=0. In a nutshell, the ratio column is the representative of repeated stops (circle ids) in daily routes for each cluster and their corresponding "Activity Types".

Let's focus on Cluster "2". It has different blocks based on its corresponding "Real Number"s. In this cluster, "2" means that daily routes which consist of only 2 stops(circle ids). For this reason, there are 3 possibilities for "Real Number" as follows:

- **Real Number 0** → The users do not repeat any of the stops (circle ids) in different days. For example;
  (This is an example, not related to the real data)

USER ID	DATE	CIRCLE IDS
U1	01/01/2016	1, 2
U1	03/01/2016	4, 7
U1	08/01/2016	5, 8

The users do not repeat any stops (circle ids) exactly every day.

- **Real Number 1** → The users repeat only 1 stop (circle id), and it is exactly repeated every day the user had delivery(s).
  (This is an example, not related to the real data)

USER ID	DATE	CIRCLE IDS
U2	01/02/2016	2, 3
U2	05/02/2016	2, 4
U2	09/02/2016	7, 2

The circle id "2" is repeated every day when U2 had a delivery.

- **Real Number 2** → The users repeat 2 out of 2 (all stops), every day they had delivery.

  (This is an example, not related to the real data)

USER ID	DATE	CIRCLE IDS
U3	04/05/2016	10, 11
U3	07/05/2016	10, 11
U3	11/07/2016	10, 11

They always go to the same stops (circle ids) without any changes.

Let' move to the column "Ratio". We use column "Clusters" and "Real Number" to calculate the ratio such as:

$$\frac{Real\ Number}{Clusters}$$

- **For Real Number 0:** The ratio is $\frac{0}{2} = 0$
- **For Real Number 1:** The ratio is $\frac{1}{2} = 0.5$
- **For Real Number 2:** The ratio is $\frac{2}{2} = 1$

These ratios are the representatives for the ratio of repeated sub-routes.

APPLICATION USAGE

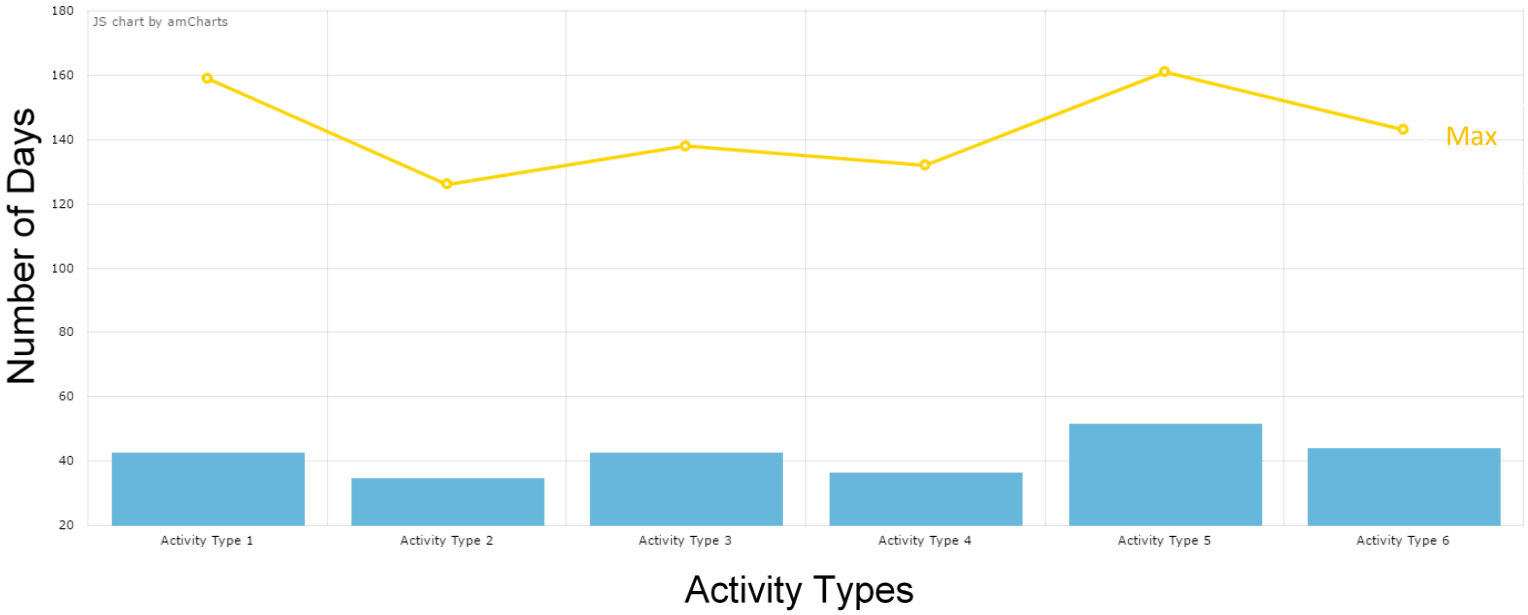


Figure 9

File= **act_dist_average_number_of_days.xlsx**

The plot above is for the average and maximum number of days for each user from their corresponding activity type in daily routes. Blue bars are for the average number of days that activity types had routes, whereas the yellow line stands for the maximum number of days occurred from each activity type. The average values are calculated by the consideration of each user in his/her corresponding activity type, and calculation of total number of days he/she had a delivery. For instance,

(This is an example, not related to the real data)

USERID	ACTIVITY TYPE	NUMBER OF DAYS (IN TOTAL)
U1	1	10
U2	3	20
U3	4	30
U4	1	40
U5	4	20
U6	3	10

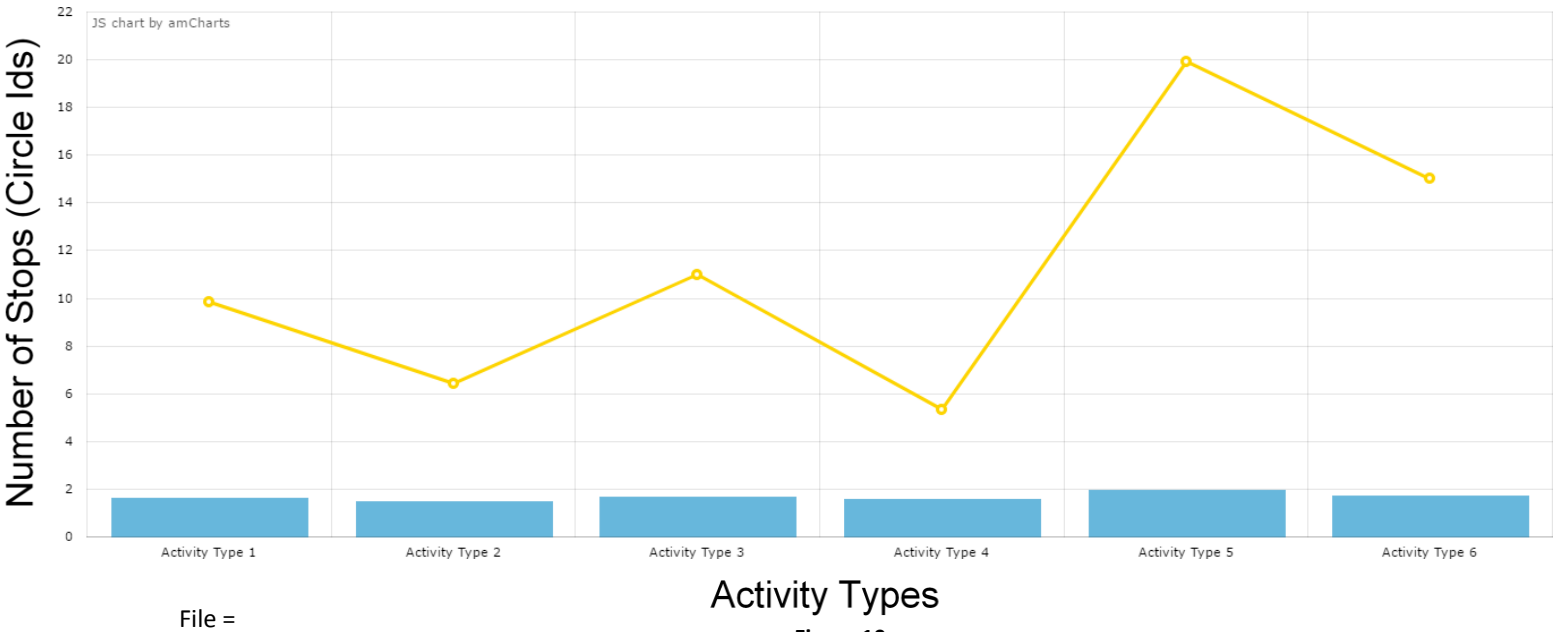
We use grouping approach for activity types and aggregation by “mean” method

$$Activity\ Type\ X\ Average\ Number\ of\ Days = \frac{sum\ of\ days}{counted\ number\ of\ rows}$$

$$Activity\ Type\ 1\ Average\ Number\ of\ Days = \frac{10 + 40}{2} = 25$$

$$Activity\ Type\ 3\ Average\ Number\ of\ Days = \frac{20 + 10}{2} = 15$$

$$Activity\ Type\ 4\ Average\ Number\ of\ Days = \frac{30 + 20}{2} = 25$$



File =

Figure 10

avg_max_circles_daily_routes.xlsx

The plot above is for the average and maximum number of stops (circle ids) for each activity type in daily routes. Activity Type 5 users have more stops (circles) in daily routes than other types. Even its maximum number of stops (circles) is much higher than the other types' maximum value.

Interpretation of Figure 9 and Figure 10: Activity Type 5 users have delivery more often than the other types, and in those days they had deliveries, they visited more stops (circle ids) than the other types' users. This fact makes Activity Type 5 is the one which is most frequently.

The Data Set for the Check-ins Frequency of Users Based On Activity Types and Hour Slots

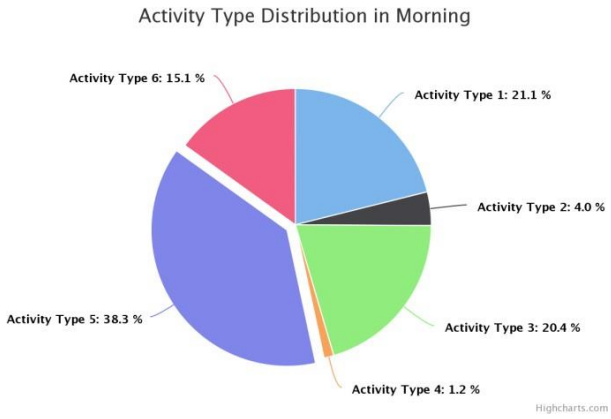
Userid	Activity Type	Hour	Frequency

The basic table above is for each hour slot to find the frequency of users and their corresponding activity types. For reading the data set in a right way, "Hour" column consist of one number. For instance, if it is "8", then it represents 08:00-08:59.

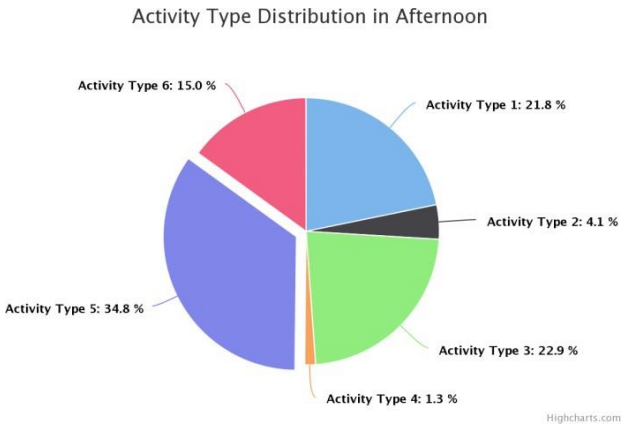
File= **act_user_hour.csv**

Using this data, we also check the activity type distribution for daytime. We created 4 daytimes based on hours as follows:

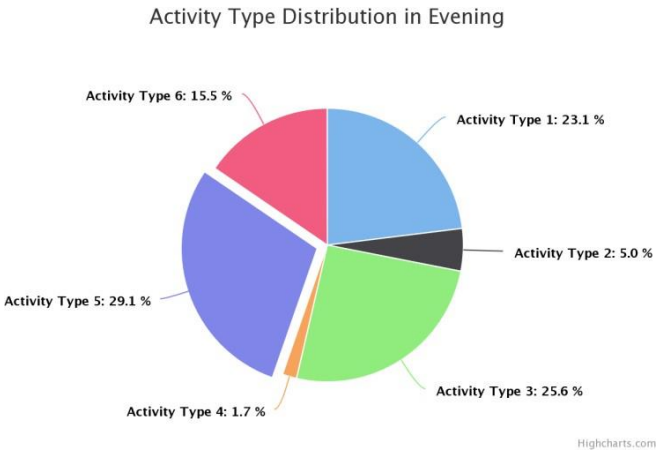
- **Morning:** From 08:00 to 12:00
- **Afternoon:** From 12:00 to 16:00
- **Evening:** From 16:00 to 20:00
- **Night:** From 20:00 to 08:00
-



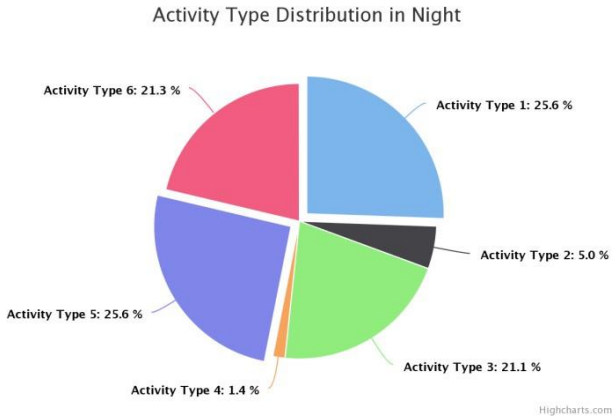
File= **act_user_hour_morning.xlsx**



File= **act_user_hour_afternoon.xlsx**



File= **act_user_hour_evening.xlsx**



File= **act_user_hour_night.xlsx**

Figure 11

Interpretation: If we order the daytimes such as “morning, afternoon, evening, night”, we can get the following observations:

- Activity Type 1, Activity Type 2 and Activity Type 6 percentages always increase from morning to night.
- Activity Type 3 and Activity Type 4 percentages increase from morning to evening, but night.
- Activity Type 5 percentage always decreases from morning to night.

The Data Set for the Check-ins Frequency of Users Based On Activity Types and Days

Userid	Activity Type	Day	Frequency

The basic table above is to illustrate our data set. We check the regular frequency for each user, each activity type and each day (date dd/mm/yyyy)

File= **act_user_day.csv**

The Data Set for the Check-ins Frequency of Users Based On Activity Types and Weeks

Userid	Activity Type	Week	Frequency

The basic table above is to illustrate our data set. We check the regular frequency for each user, each activity type and each week (From Week1 to Week29).

File= **act_user_week.csv**

The Data Set for the Check-ins Frequency of Users Based On Activity Types and Months

Userid	Activity Type	Month	Frequency

The basic table above is to illustrate our data set. We check the regular frequency for each user, each activity type and each month (From 1 to 7).

File= **act_user_month.csv**

The Data Set for the Check-ins Frequency of Circles Based On Activity Types and Hours

CircleID	Hour	Activity Type	Frequency

The basic table above is to illustrate our data set. We check the regular frequency for each circle, at each hour and each activity type.

File= **circle_hour.csv**

The Data Set for the Check-ins Frequency of Circles Based On Activity Types and Days

CircleID	Date	Activity Type	Frequency

The basic table above is to illustrate our data set. We check the regular frequency for each circle, on each day and each activity type.

File= **circle_day.csv**

The Data Set for the Check-ins Frequency of Circles Based On Activity Types and Weeks

CircleID	Week	Activity Type	Frequency

The basic table above is to illustrate our data set. We check the regular frequency for each circle, in each week and each activity type.

File= **circle_week.csv**

The Data Set for the Check-ins Frequency of Circles Based On Activity Types and Weeks

CircleID	Month	Activity Type	Frequency

The basic table above is to illustrate our data set. We check the regular frequency for each circle, in each month and each activity type.

File= **circle_month.csv**

FIRST AND LAST STOP ANALYSIS

Since we are not able to have coordinates for circle ids yet, we use the regular tramos for this part of analysis.

First stop data = **first_origin.csv**

Last stop data = **last_arrival.csv**

Combined version= **first_last.csv** (including DayTime)

FIRST STOP

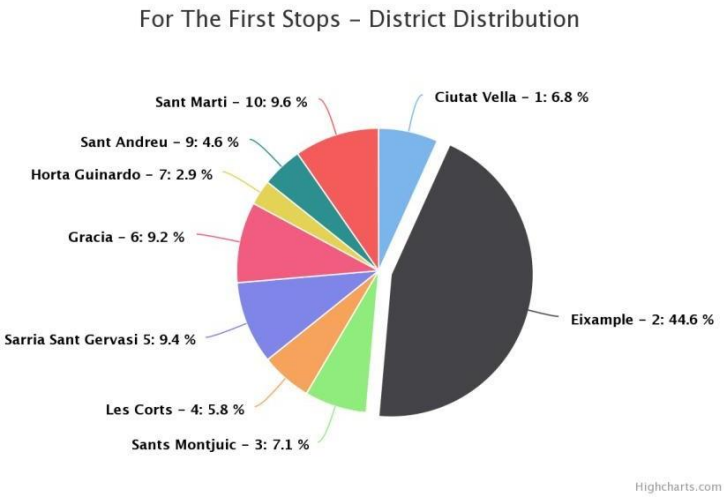
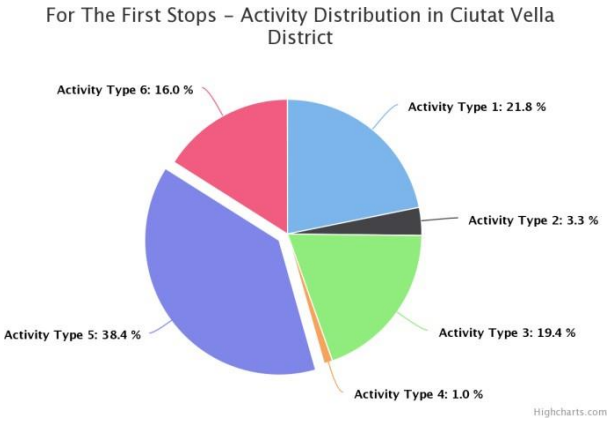


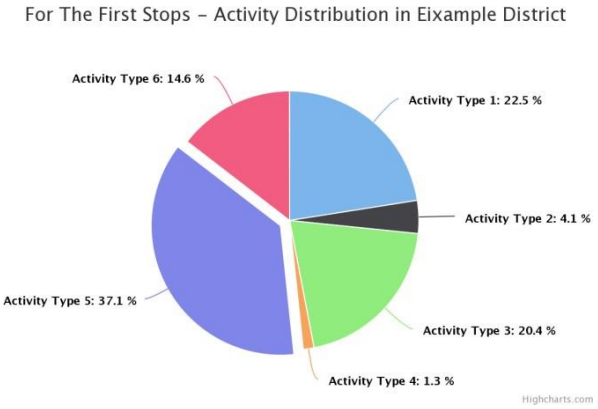
Figure 12

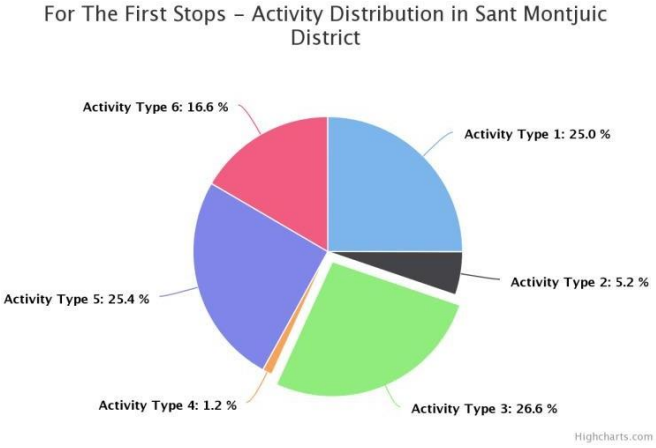
The plot on the left is to illustrate district distribution for the first stop check-ins of daily routes. Unsurprisingly, Eixample has the largest piece in the pie chart.

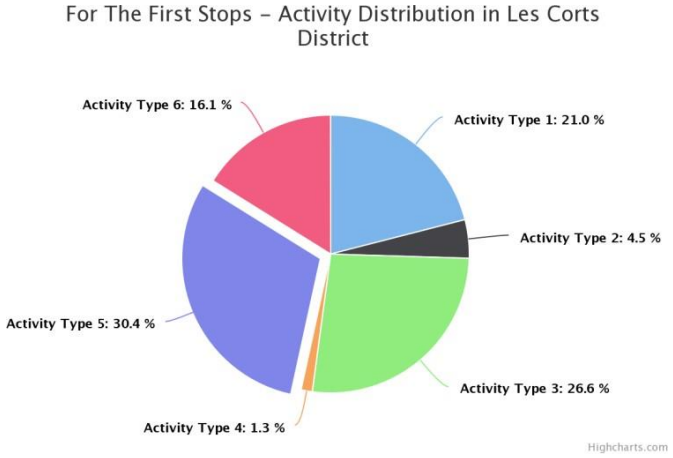
File= **firststop_district_dist.xlsx**

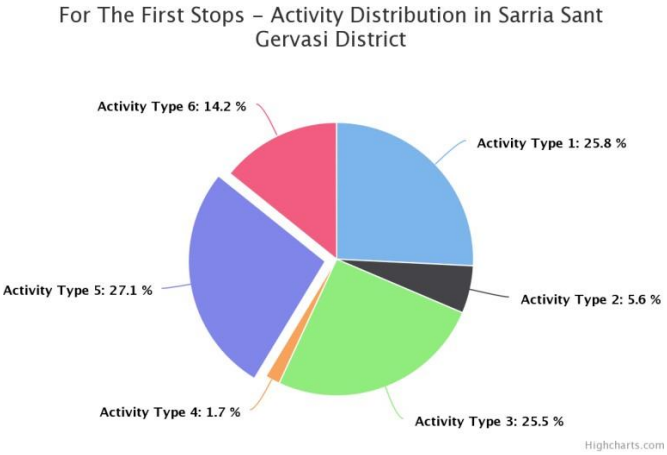
Activity Type Distribution for First Stops in Daily Routes in Each District

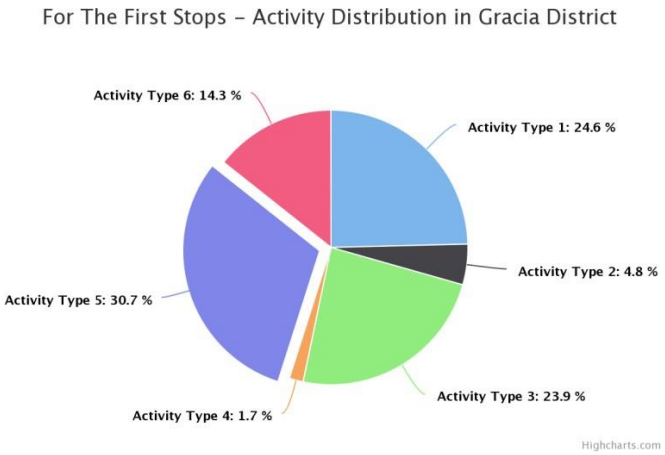


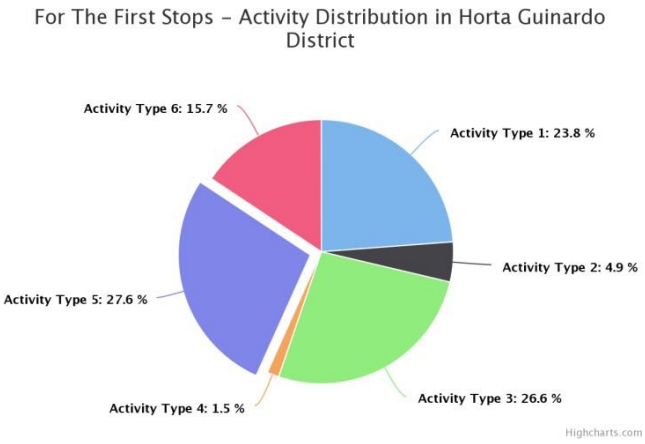


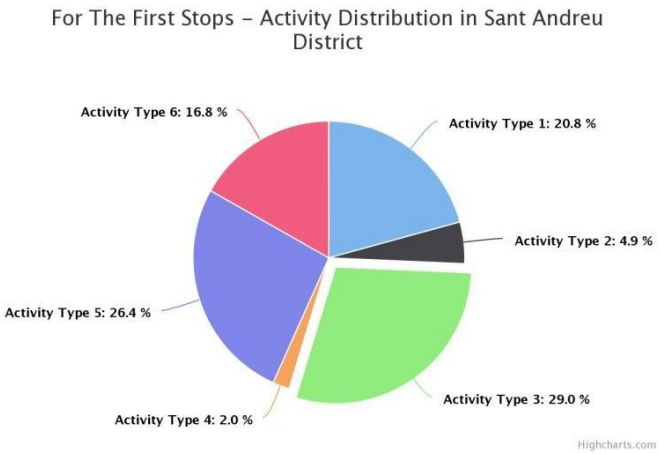












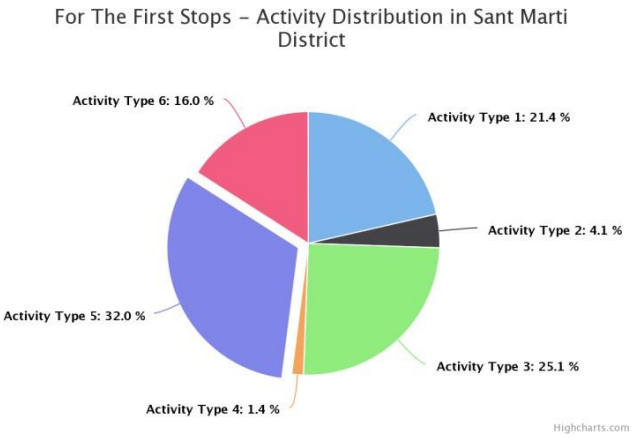


Figure 13

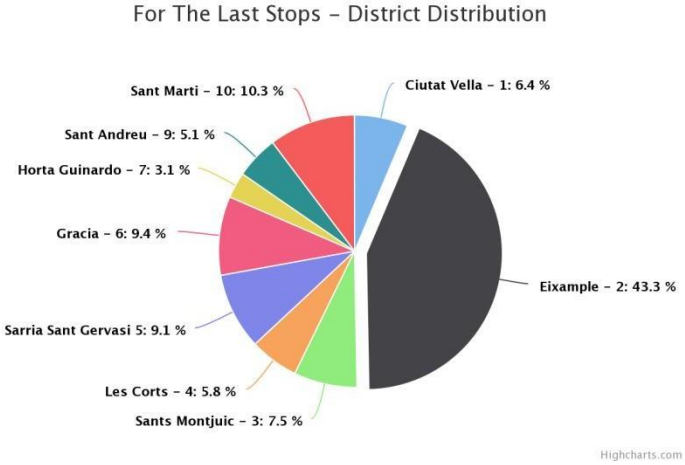
File=**firststop_district_act_dist.csv**

First Stop	Hour Slot	Frequency

The basic table above is to illustrate the frequency for each first stop and each hour slot.

File= **first_hour_freq.csv**

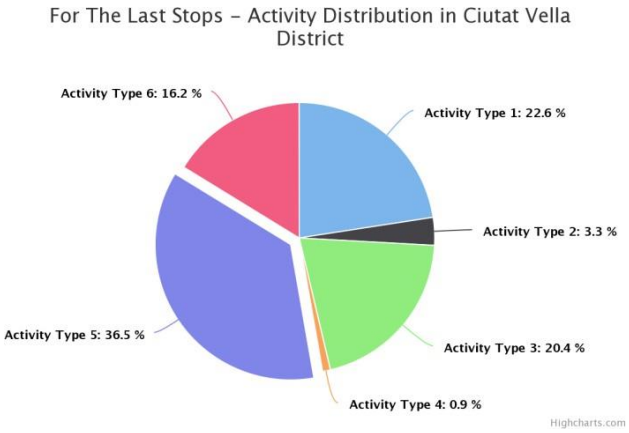
LAST STOP

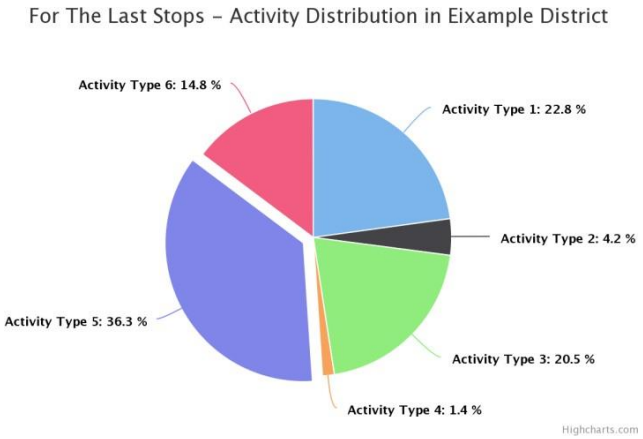


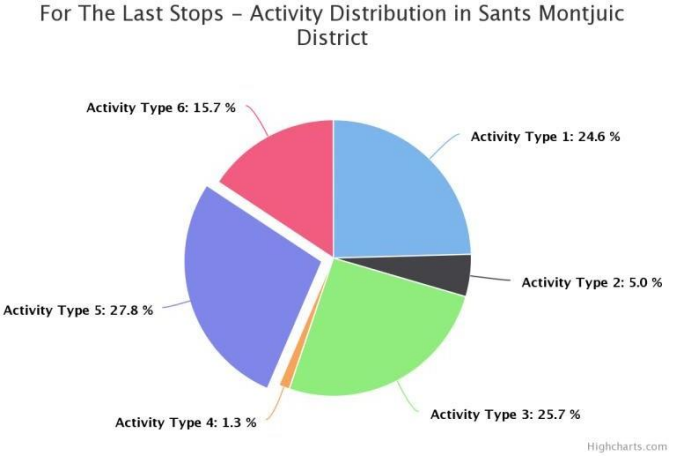
File= **laststop_district_dist.xlsx**

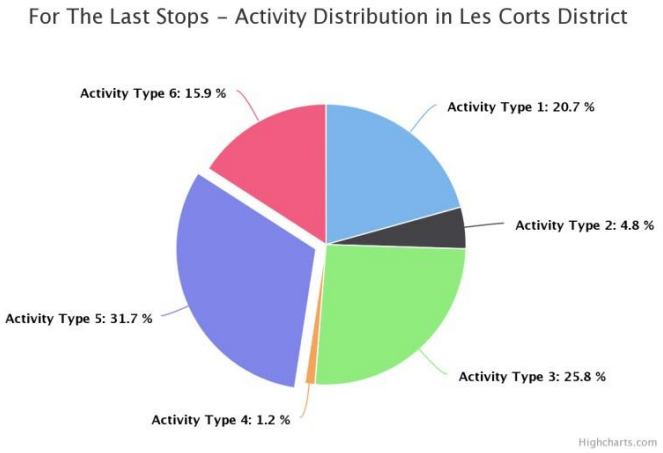
Figure 14

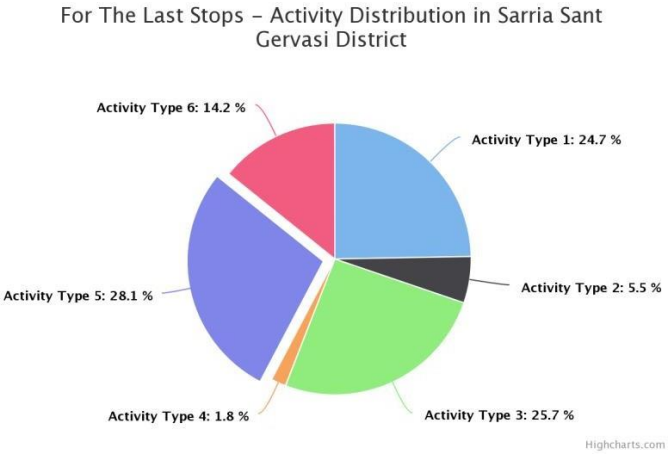
Activity Type Distribution for Last Stops in Daily Routes in Each District

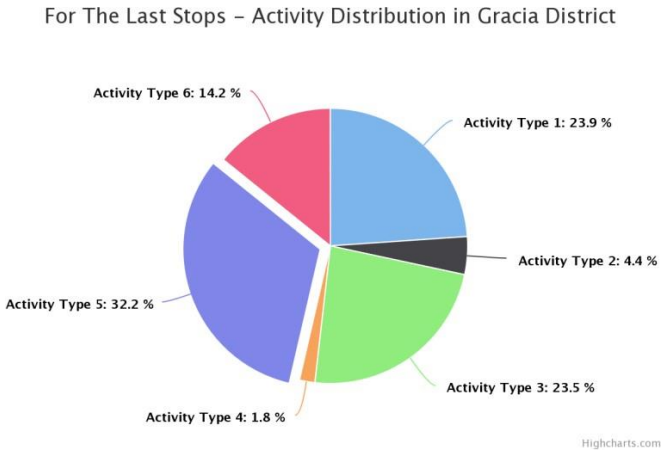


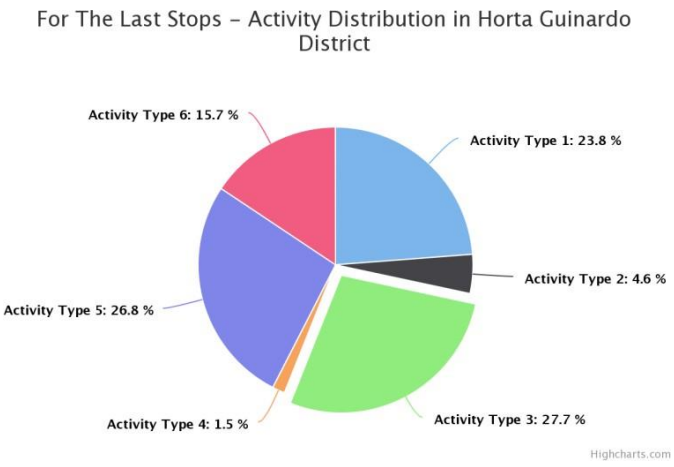


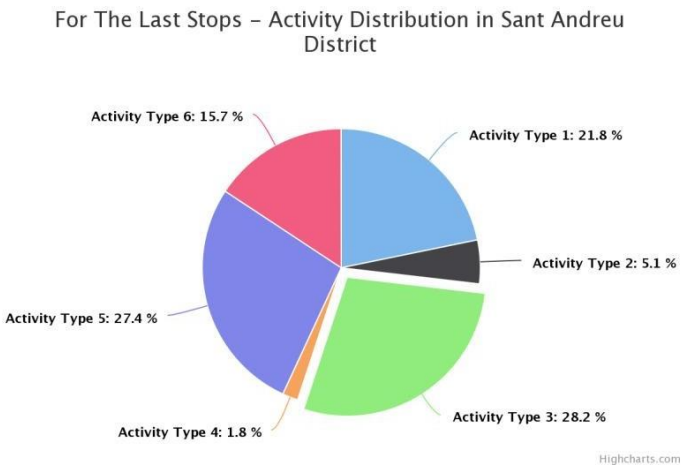












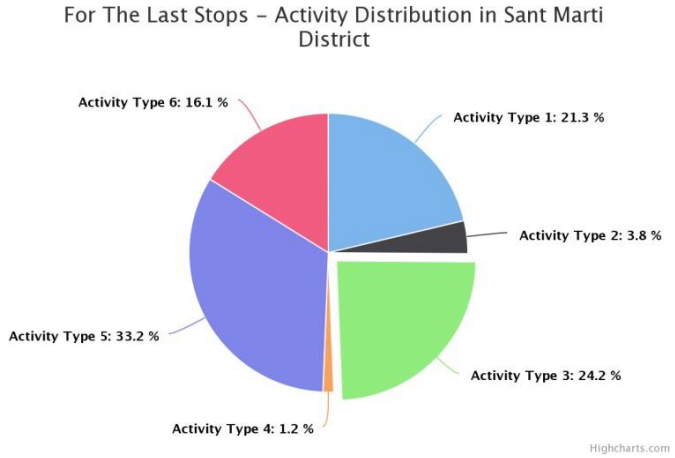


Figure 15

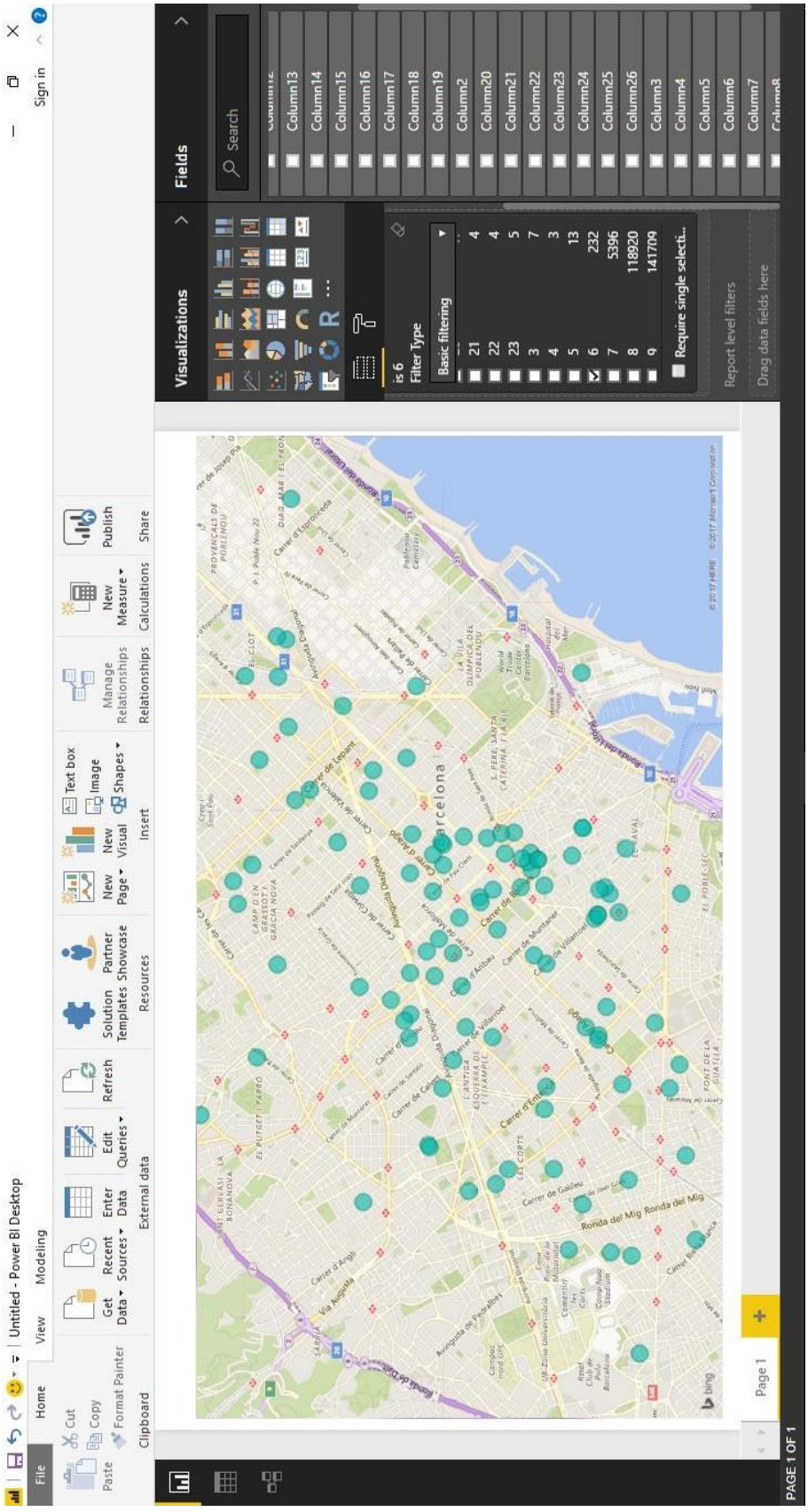
File=**laststop_district_act_dist.csv**

Last Stop	Hour Slot	Frequency

The basic table above is to illustrate the frequency for each last stop and each hour slot.

File= **last_hour_freq.csv**

Example Plotting Using PowerBI – First Tramos at 6am



TIME DIFFERENCE BETWEEN THE FIRST AND THE LAST TRAMOS

After the time difference calculation between the first and the last stop of daily routes, time groups are created.

Time Group	Time Difference Interval
1	[0, 1]
2	[1, 10]
3	[10, 20]
4	[20, 30]
5	[30, 40]
6	[40, 50]
7	[50, 60]
8	[60, 70]
9	[70, 80]
10	[80, 90]
11	[90, 100]
12	[100, 110]
13	[110, 120]
14	[120, 130]
15	[130, 140]
16	[140, 150]
17	[150, 160]
18	[160, 170]
19	[170, 180]
20	[180, 190]
21	[190, 200]
22	[200, 210]
23	[210, 220]
24	[220, 230]
25	[230, 240]
26	[240, 300]
27	[300, 360]
28	[360, 420]
29	[420, 480]
30	[480, 540]
31	[540, 600]
32	[600, 660]
33	[660, 720]
34	[720, 780]
35	[780, 840]
36	[840, 900]
37	[900, 960]
38	[960, 1020]
39	[1020, 1080]
40	[1080, 1140]
41	[1140, 1200]
42	[1200, 1260]

"Time Difference Interval" column values are formatted into minutes. Time Group 1 is created for mistaken repeated check-ins (double click etc). From Group 2 to Group 25 the intervals are for 10 minutes. The rest is designed hourly.

Time Groups	Activity Types	Frequency

File= **act_dist_timegroups.csv**

The data table above is created to check the activity type distribution for each time groups.

Since it is not feasible to plot for each time group, we also grouped the time groups.

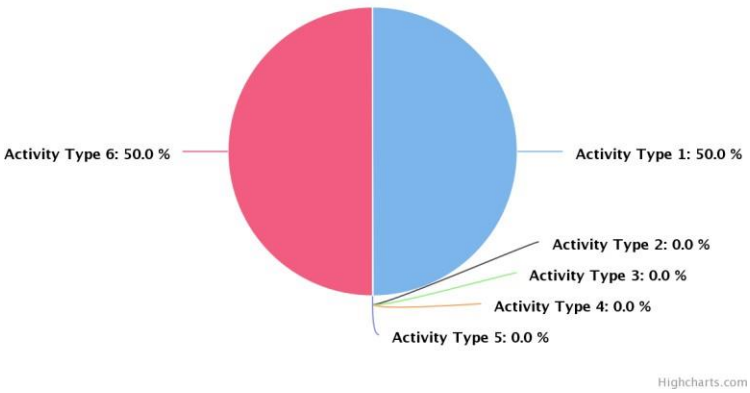
- Grouped Time Group 1 = Time Group 1
- Grouped Time Group 2 = Time Group 2, 3, 4, 5, 6, 7
- Grouped Time Group 3 = Time Group 8, 9, 10, 11, 12, 13
- Grouped Time Group 4 = Time Group 14, 15, 16, 17, 18, 19
- Grouped Time Group 5 = Time Group 20, 21, 22, 23, 24, 25
- Grouped Time Group 6 = Time Group 26, 27, 28, 29, 30, 31
- Grouped Time Group 7 = Time Group 32, 33, 34, 35, 36
- Grouped Time Group 8 = Time Group 37, 38, 39, 40, 41, 42

Grouped Time Groups	Activity Types	Frequency

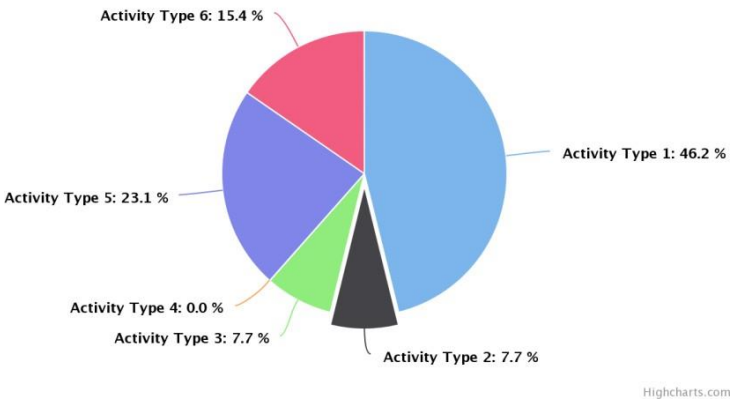
File= **act_dist_grouped_timegroups.csv**

The data table above is created to be able to have piecharts.

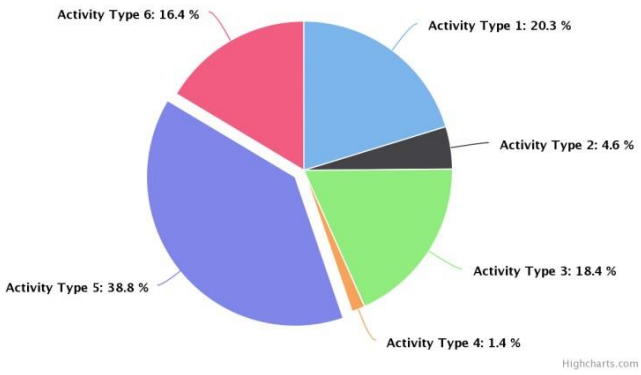
Activity Type Distribution in Grouped TimeGroupes 0



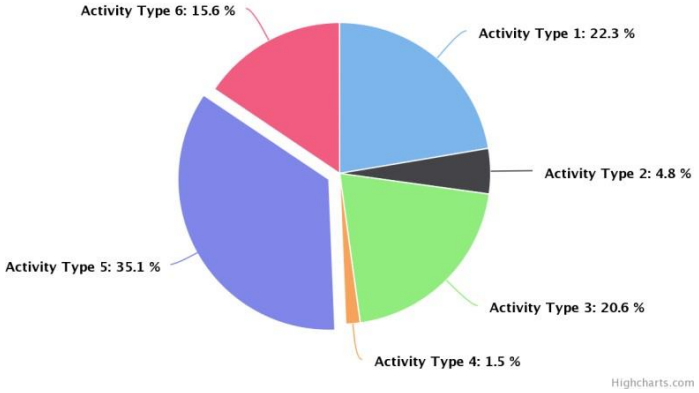
Activity Type Distribution in Grouped TimeGroupes 1



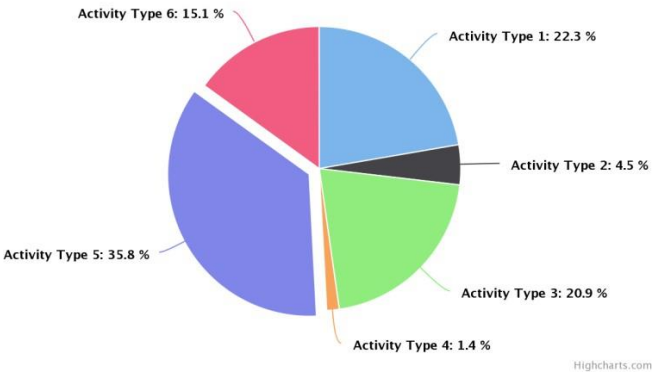
Activity Type Distribution in Grouped TimeGroupes 2



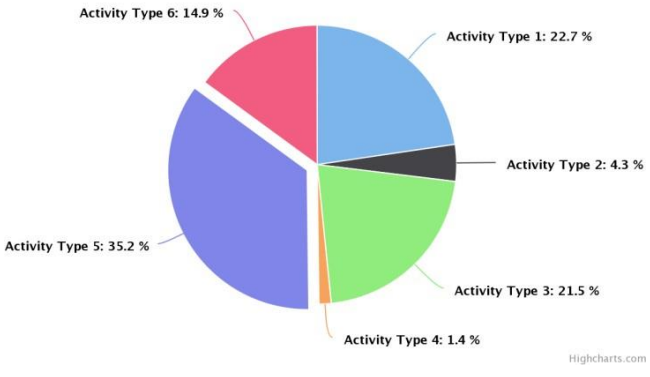
Activity Type Distribution in Grouped TimeGroupes 3

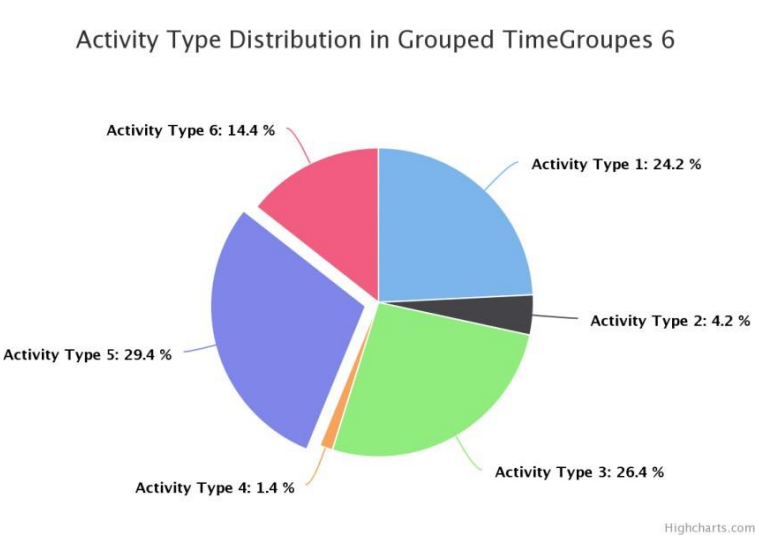


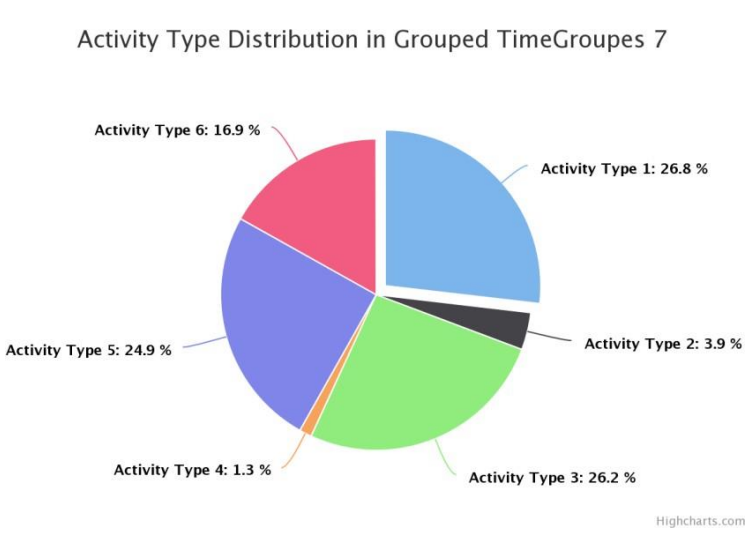
Activity Type Distribution in Grouped TimeGroupes 4



Activity Type Distribution in Grouped TimeGroupes 5







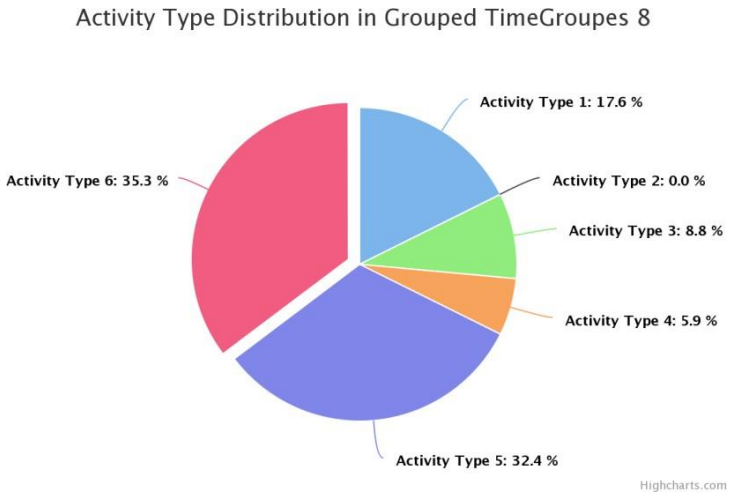
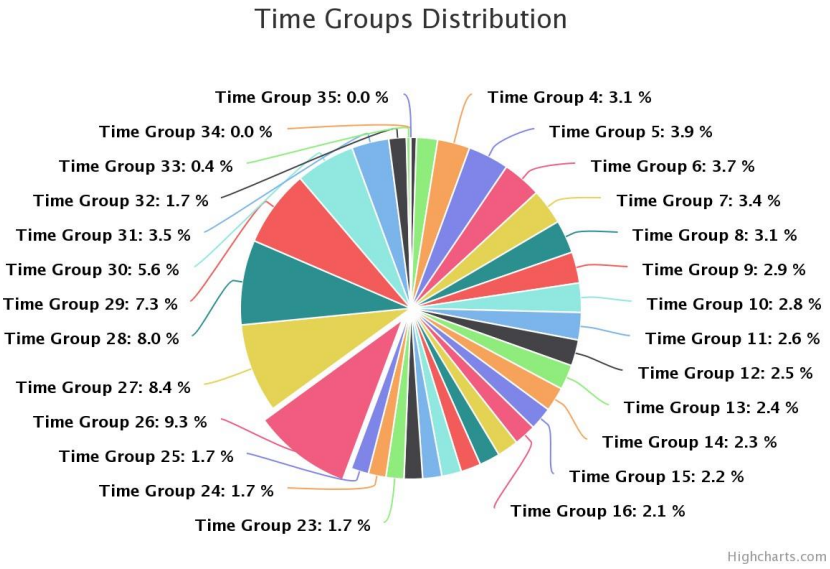


Figure 16

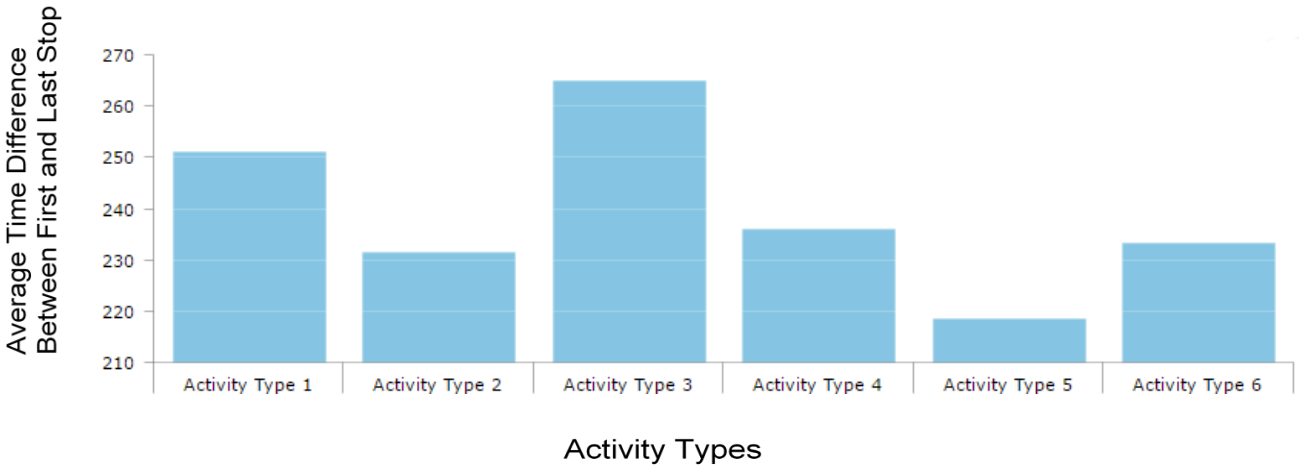
Interpretation:

Group 0 and Group 1 activity distributions are different from the others. From Group 2 to Group 5, the activity type distributions are similar. Group 6, Group 7 and Group 8 are also different by the distributions.



The plot on the left is to illustrate the time groups' distribution. Time Group 26 has the largest percentage, with the interval of [240, 300] which is from 4 hours to 5 hours.

File= **time_groups_dist.csv**



The plot above is to illustrate the average time difference for activity types.

File= **act_dist_average_time_diff.xlsx**

DISTANCE BETWEEN THE FIRST AND THE LAST TRAMOS

Distance Groups	Distance Interval
0	0
1	[0, 100]
2	[100, 250]
3	[250, 500]
4	[500, 750]
5	[750, 1000]
6	[1000, 1250]
7	[1250, 1500]
8	[1500, 1750]
9	[1750, 2000]
10	[2000, 2250]
11	[2250, 2500]
12	[2500, 2750]
13	[2750, 3000]
14	[3000, 3250]
15	[3250, 3500]
16	[3500, 3750]
17	[3750, 4000]
18	[4000, 4250]
19	[4250, 4500]
20	[4500, 4750]

21	[4750, 5000]
22	[5000, 5250]
23	[5250, 5500]
24	[5500, 5750]
25	[5750, 6000]
26	[6000, 6250]
27	[6250, 6500]
28	[6500, 6750]
29	[6750, 7000]
30	[7000, 7250]
31	[7250, 7500]
32	[7500, 7750]
33	[7750, 8000]
34	[8000, 8250]
35	[8250, 8500]
36	[8500, 8750]
37	[8750, 9000]
38	[9000, 9250]
39	[9250, 9500]
40	[9500, 9750]

"Distance Interval" column values are formatted into meters. Time Group 0 is created for the users who come to the starting point at the end of their daily route. Group 1 is created for 100 meters distance, whereas the others are for 250 meters.

Distance Groups	Activity Types	Frequency

File= **act_dist_distancegroups.csv**

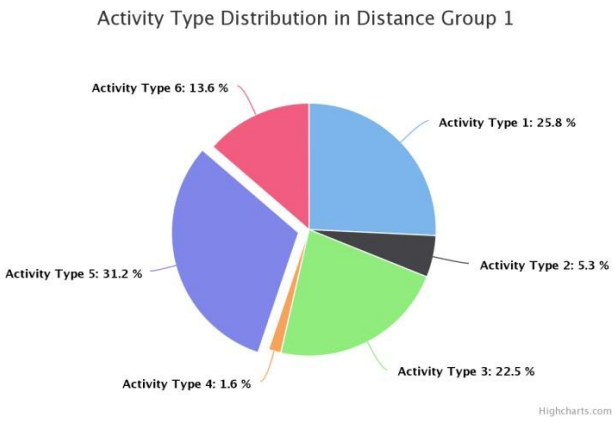
The data table above is created to illustrate the activity type distributions for distance groups. However, it is not feasible to plot everything, and we created grouped distance groups.

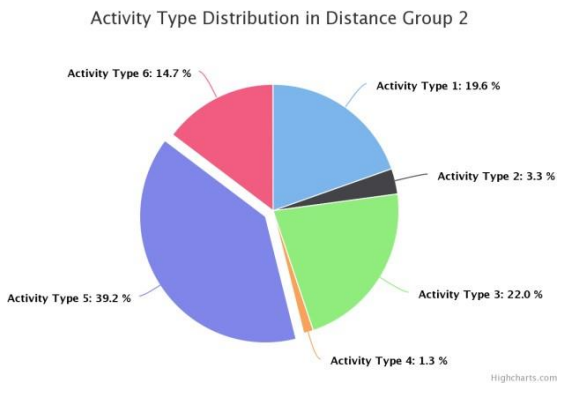
- Grouped Distance Group 0 = Distance Group 0
- Grouped Distance Group 1 = Distance Group 1, 2, 3
- Grouped Distance Group 2 = Distance Group 4, 5
- Grouped Distance Group 3 = Distance Group 6, 7, 8, 9
- Grouped Distance Group 4 = Distance Group 10, 11, 12, 13
- Grouped Distance Group 5 = Distance Group 14, 15, 16, 17
- Grouped Distance Group 6 = Distance Group 18, 19, 20, 21
- Grouped Distance Group 7 = Distance Group 22, 23, 24, 25
- Grouped Distance Group 8 = Distance Group 26, 27, 28, 29
- Grouped Distance Group 9 = Distance Group 30, 31, 32, 33
- Grouped Distance Group 10 = Distance Group 34, 35, 36, 37
- Grouped Distance Group 11 = Distance Group 38, 39, 40

Grouped Distance Groups	Activity Types	Frequency

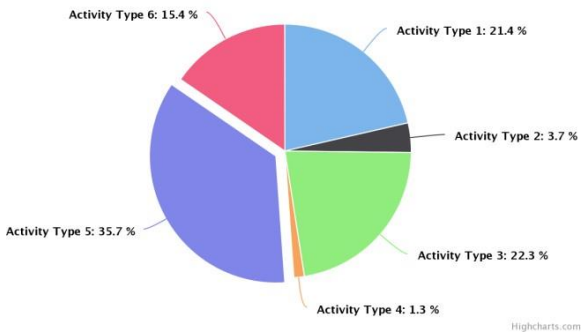
File= **act_dist_grouped_distancegroups.csv**

The data table above is created to be able to have piecharts.

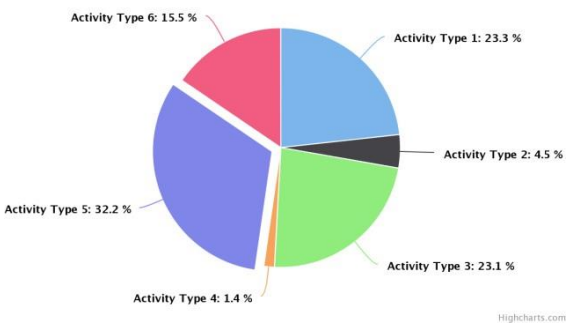




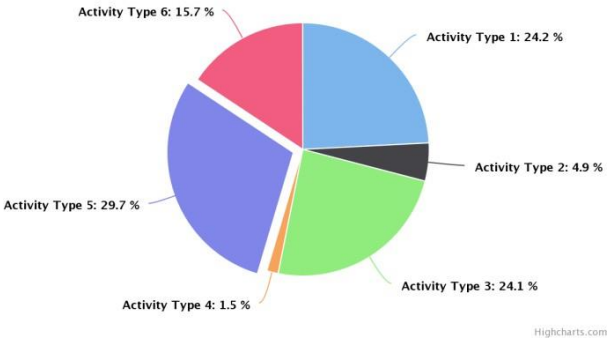
Activity Type Distribution in Distance Group 3



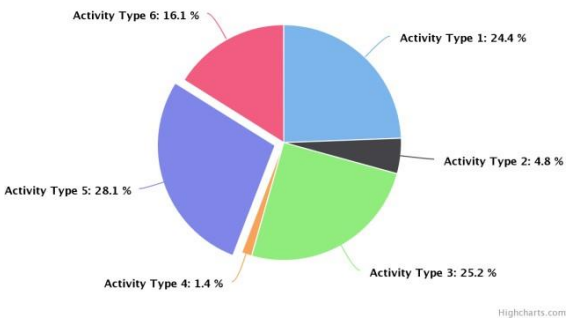
Activity Type Distribution in Distance Group 4



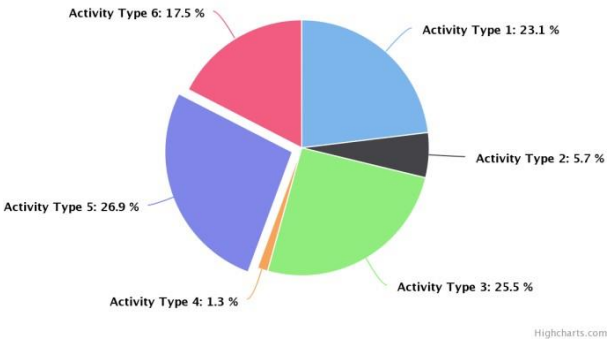
Activity Type Distribution in Distance Group 5



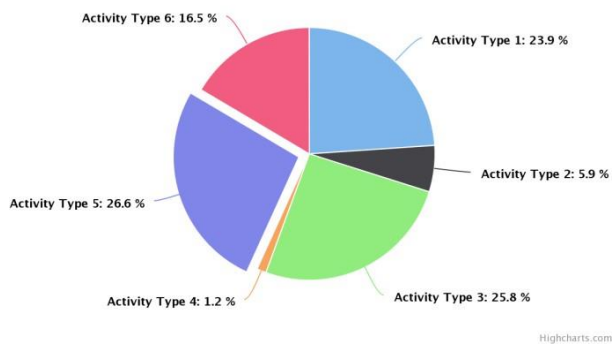
Activity Type Distribution in Distance Group 6



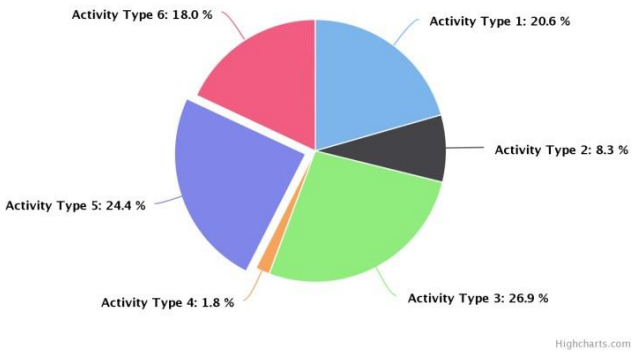
Activity Type Distribution in Distance Group 7



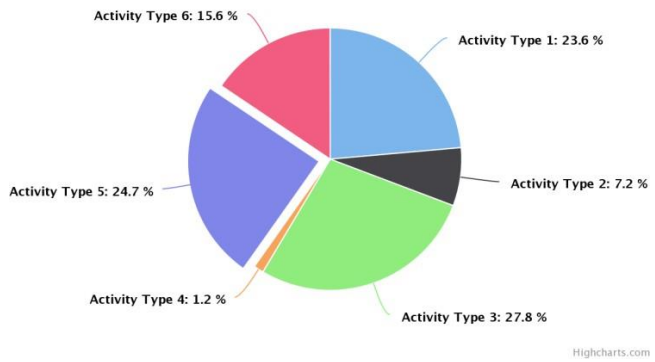
Activity Type Distribution in Distance Group 8

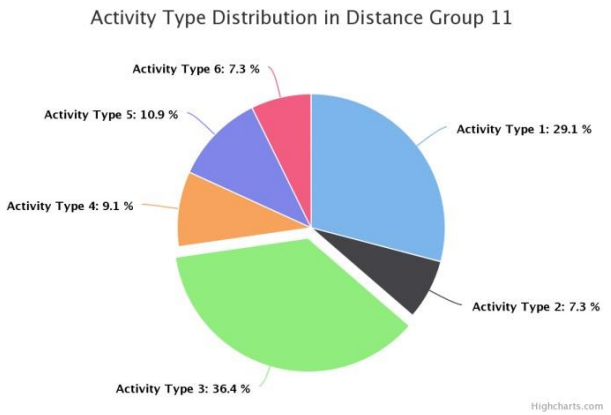


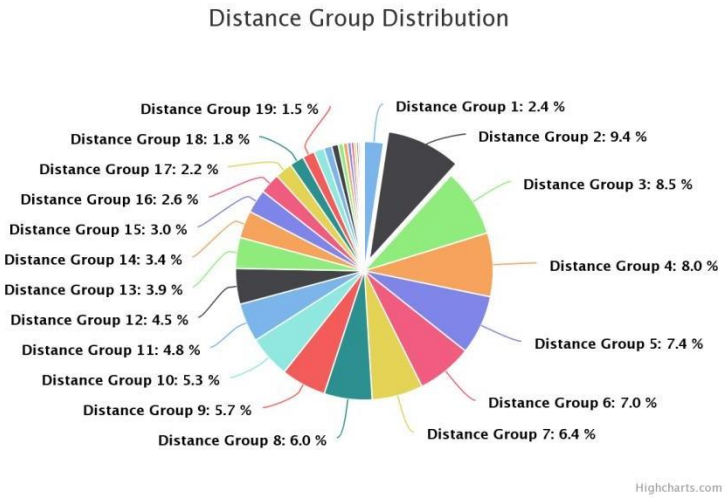
Activity Type Distribution in Distance Group 9



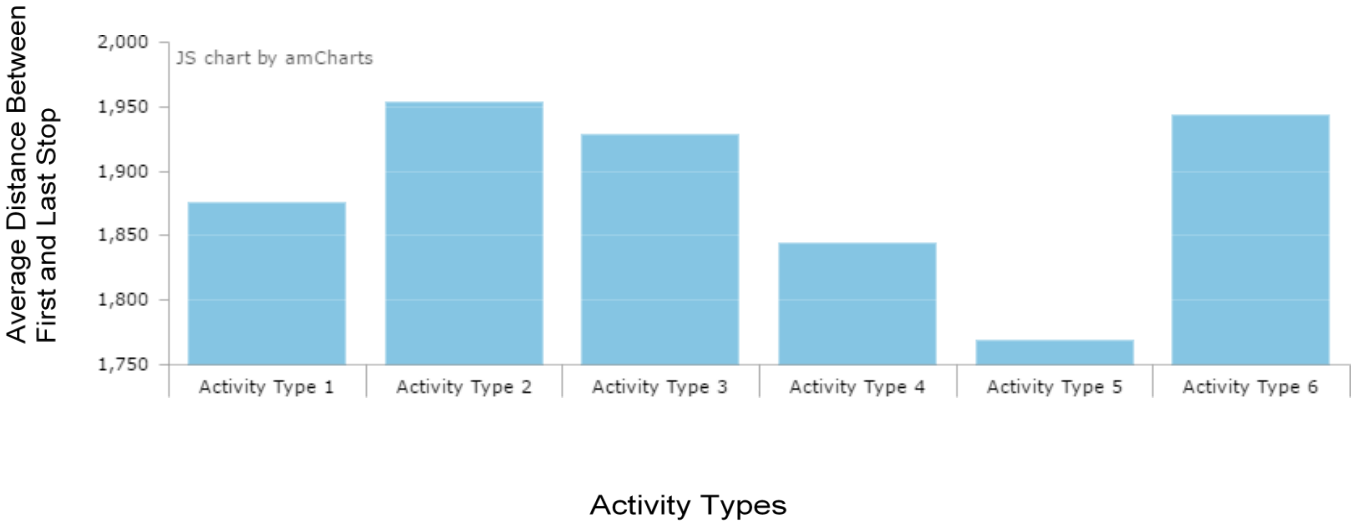
Activity Type Distribution in Distance Group 10





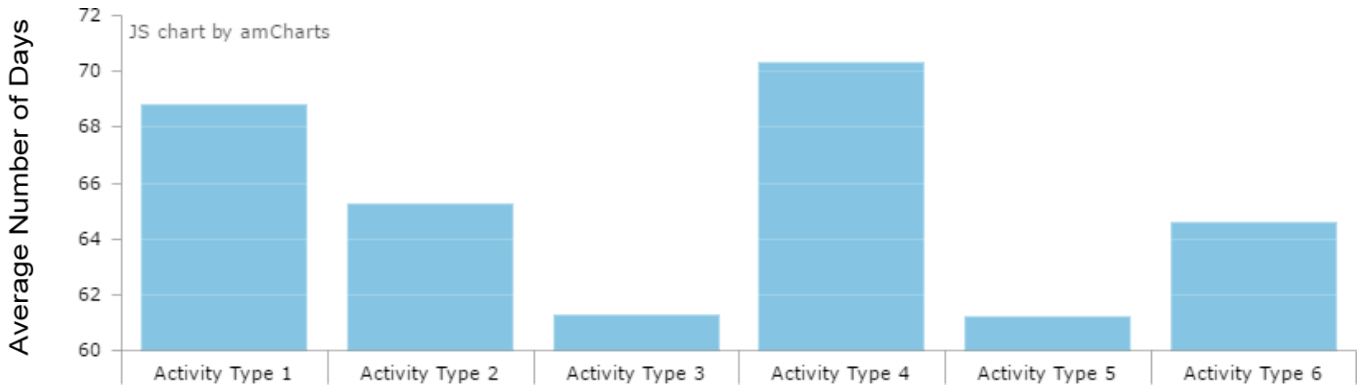


File= **act_dist_distancegroups.csv**



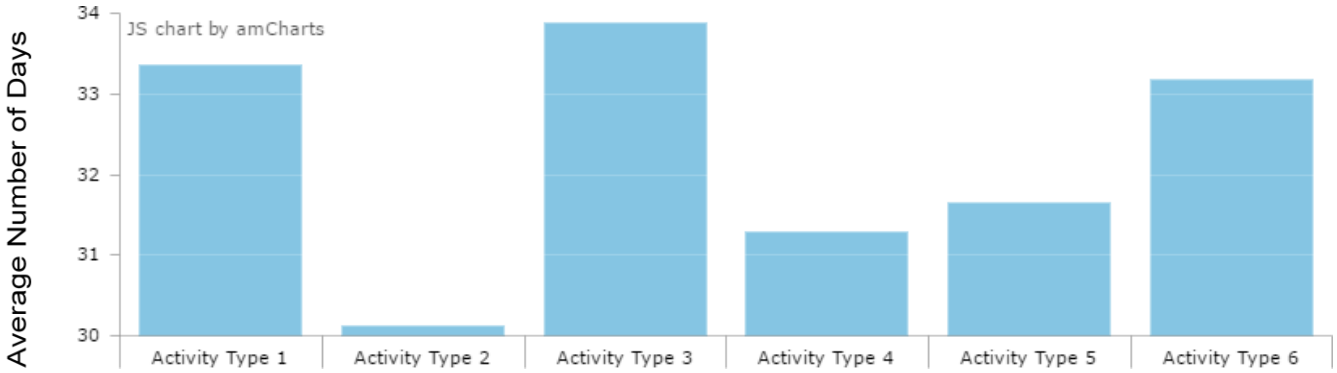
File= **act_dist_average_distance.xlsx**

Addition to the Application Usage



The plot above shows the average time difference between the first usage of the application and 01/01/2016.

File= **average_days_after_january_usage.xlsx**



The plot above shows the average time difference between the first usage of the application and 01/01/2016.

File= **average_days_before_july_usage.xlsx**