# Customer Journey

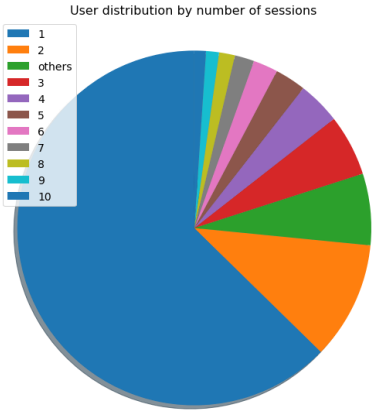
## 1. Background

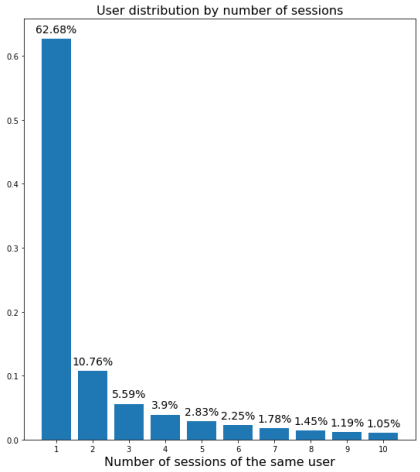
This document analyses the user sessions during the month of September 2020. The analysis aims to identify which sessions are more important to understand how familiar users use applications.

To achieve the most representative data possible, different filters have been applied to eliminate sessions that show non-human behavior.

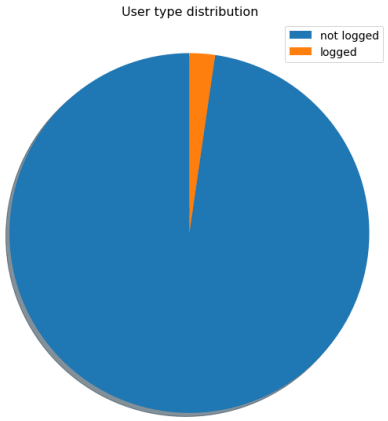
## 2. Initial Data Analysis

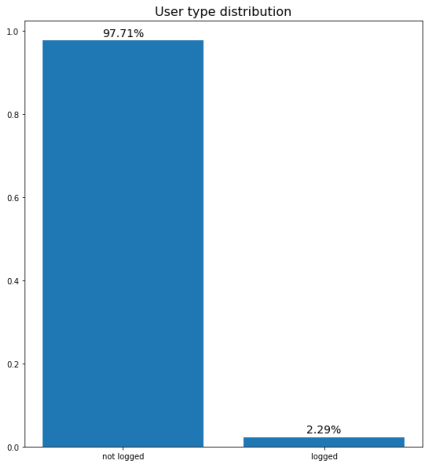
In this study, we work with user sessions from September 2020.
The dataset contains 488.032 sessions from 355.902 different users and only 17% of users have registered more than four visits in the month of September.



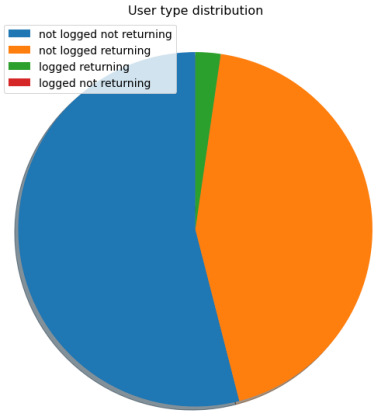

As we can see below, most of the users are not logged users, almost reaching the 98% of the users.

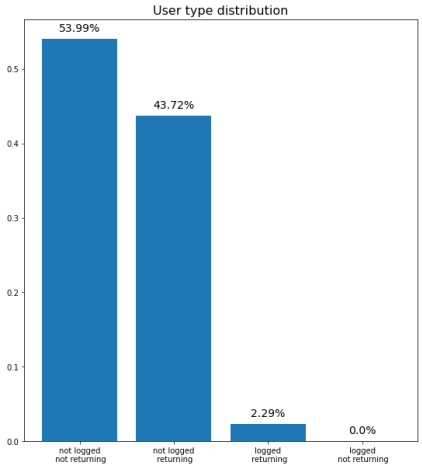


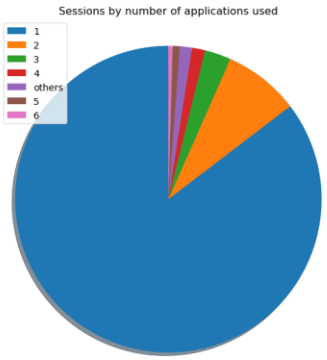
# IP Portal

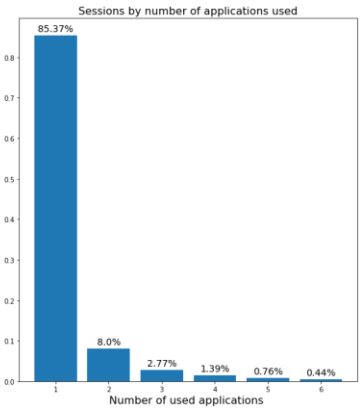
It could be expected that the non-logged in users do not have enough knowledge to use the applications correctly, but it has been observed that, as shown below, there is a large number of non-logged in users who visit the portal on multiple occasions. These users can be interesting because they can show an expert behaviour but different from that of registered users since the actions they can perform is reduced.

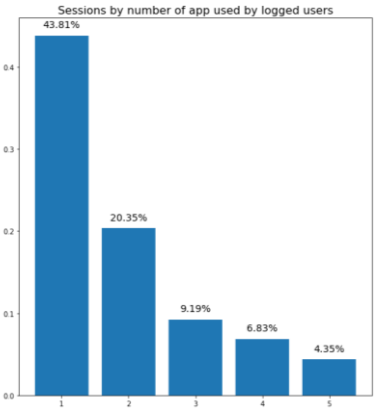


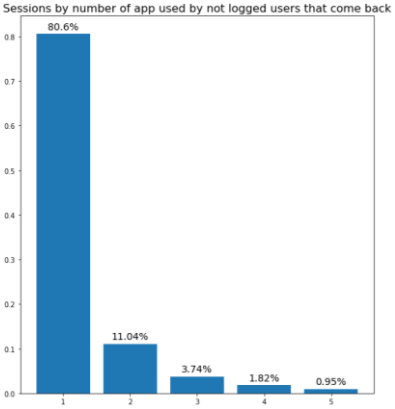



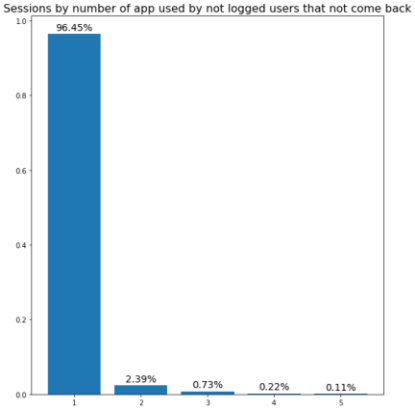

Another interesting aspect to bear in mind is the number of applications used during a session and their duration because the sessions are more informative when using multiple applications as it can be inferred what the user is doing or trying to do.



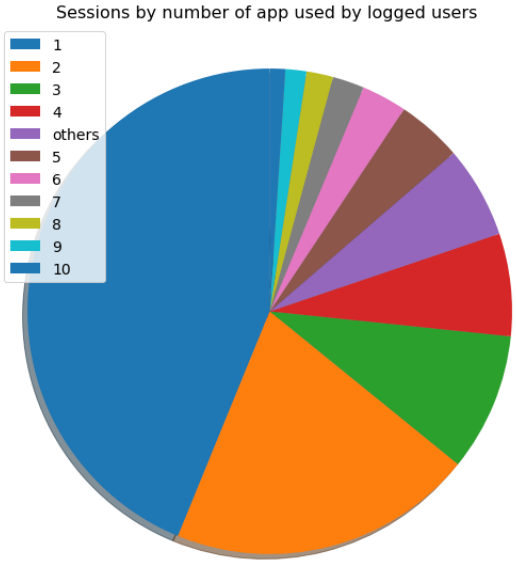



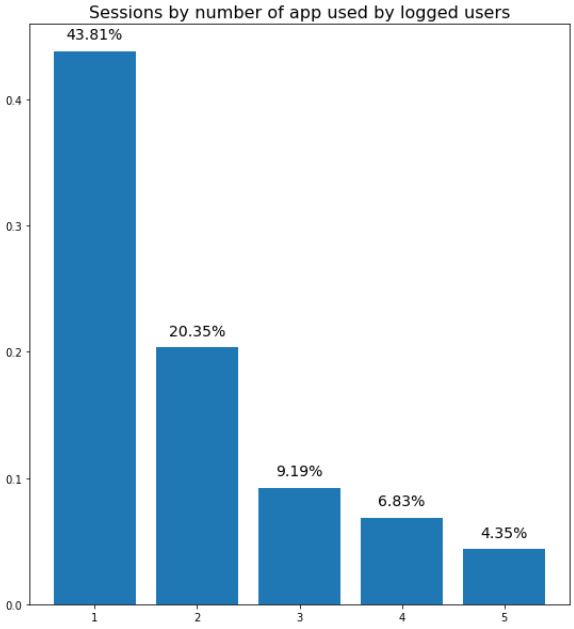



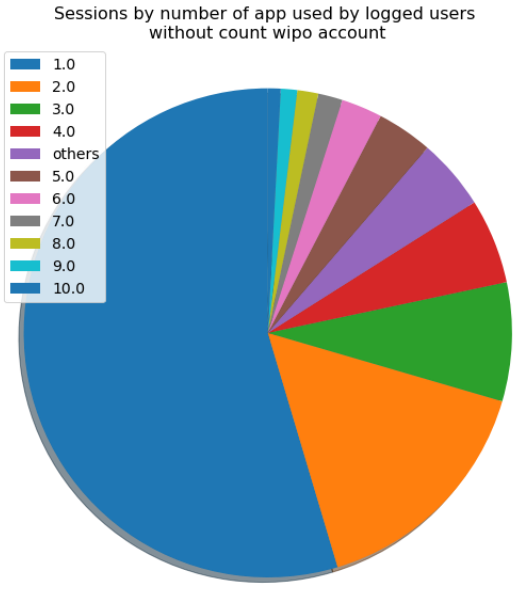
# IP Portal

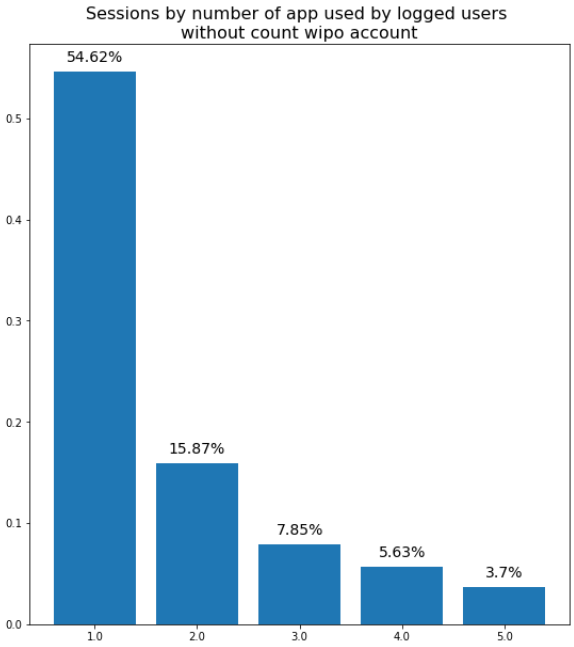
An important aspect to bear in mind when we talking about the number of applications used during a session by a logged user, is that exists an application only to perform the login.
For this reason, it is interesting (and important) to evaluate the number of applications of logged users taking into account all the applications except this one (Wipo Account).

As we can see below, with this detail, these users remain in general using mora applications than the rest.



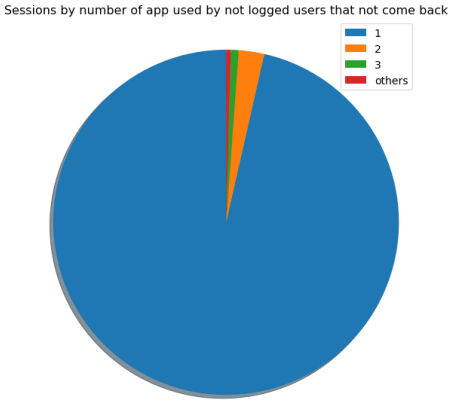



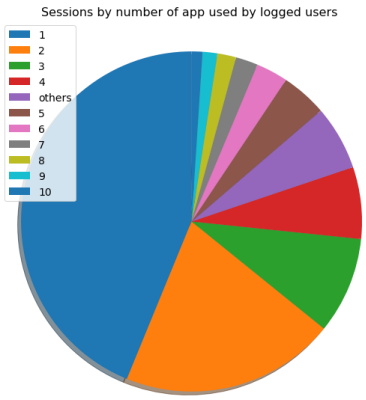



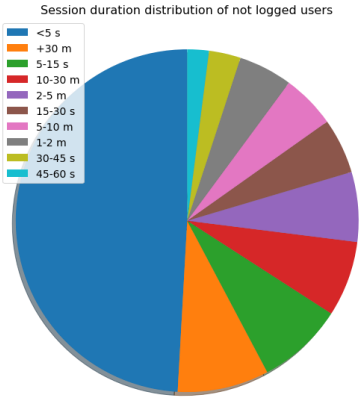
# IP Portal

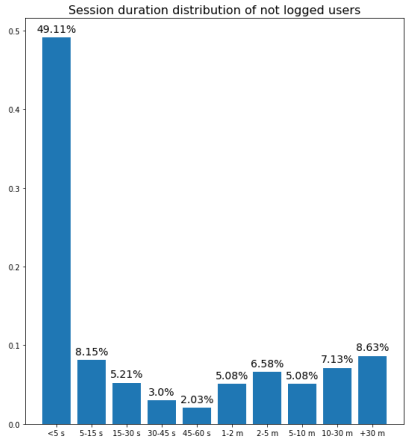
We can see that expert users tend to use more applications. If we compare the non-logged in users who do not return with logged in users, the percentage of sessions that only use one application drops from 96% to 44%.

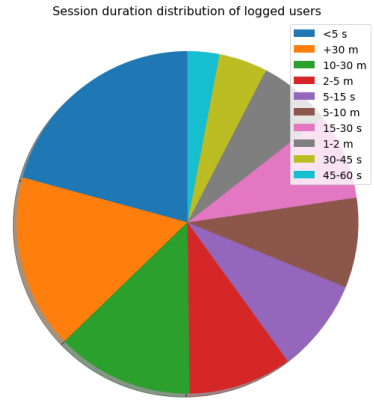


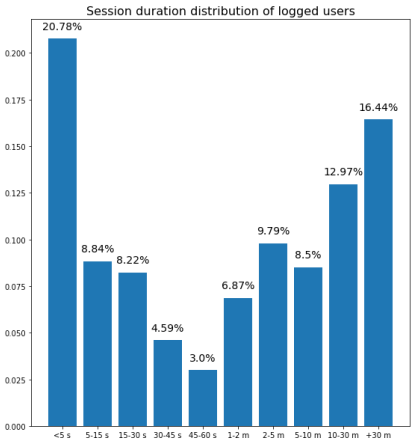




In order to analyse the duration of the sessions, it was necessary to calculate the time spent on each application based on the server time of their events.
As we can see below, there is large number of sessions with very short or very long durations.

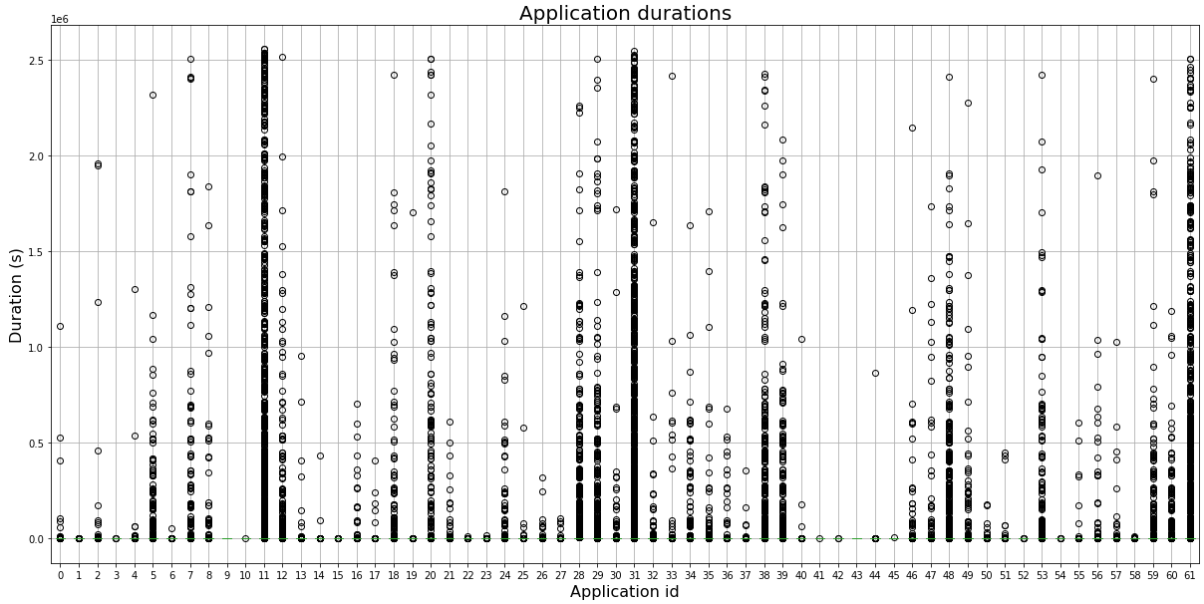


Below is an inconspicuous box plot. This graph could seem be wrong a priori, but gives us useful information such as that the vast majority of durations are very small and close to zero.
We also observe that there are many applications with durations close to even 28 days.
We can therefore affirm that there are sessions that do not follow human behaviour.
The next step of this analysis is try to identify, explain and filter these sessions.




## 3. Data Cleaning

For each application, we have the events that a user has performed in each of them, such as clicking or loading a new page. Knowing this, we can assume that a common user will perform events with short durations (few seconds or minutes) and we should expect to have sessions with a limited duration (less than a day).
By definition, a session ends after 30 minutes of inactivity so an event should not be able to reach that duration.

Based on this, the first rule created to remove undesired sessions is delete sessions with events lasting more than 25 minutes.

### 3.1. Ignore sessions with events lasting more than 25 minutes.

With this rule, the sessions that contain an event on any application with at least 25 minutes of duration are marked as outliers and ignored.
This rule affects to **17.740/488.032** sessions but even with this constraint, there are still many sessions with several days duration.

# IP Portal

3.2. Ignore sessions with at least 6 consecutive events of 10 minutes or more.

With this rule, the sessions that contain six or more consecutives events with a duration of 600 seconds are marked as outliers and ignored.
This rule affects to **648/488.032** sessions but even with this constraint, there are still many sessions with several days duration.

3.3. Ignore sessions with a first/last event server time different than first/last action time.

With this rule, the sessions where the first event server time or the last event server time differs in at least 60 seconds with the first action or the last action are marked as outlier.
This rule affects to **8.476/488.032** sessions but even with this constraint, there are still many sessions with several days duration.

| | server_time | session_hk_c | visitor_id_c | visit_first_action_time | visit_last_action_time | step_session | |
|---|---|---|---|---|---|---|---|
| 0 | 2020-09-21 12:25:57 | 0 | 77997 | 2020-09-21 07:10:47 | 2020-09-21 13:13:05 | 0 | ← 5 Hours of difference |
| 1 | 2020-09-21 12:26:07 | 0 | 77997 | 2020-09-21 07:10:47 | 2020-09-21 13:13:05 | 1 | |
| 2 | 2020-09-21 12:26:09 | 0 | 77997 | 2020-09-21 07:10:47 | 2020-09-21 13:13:05 | 2 | |
| 3 | 2020-09-21 12:27:23 | 0 | 77997 | 2020-09-21 07:10:47 | 2020-09-21 13:13:05 | 3 | |
| ⋮ | | | | | | | |
| 76 | 2020-09-21 13:10:41 | 0 | 77997 | 2020-09-21 07:10:47 | 2020-09-21 13:13:05 | 76 | |
| 77 | 2020-09-21 13:11:10 | 0 | 77997 | 2020-09-21 07:10:47 | 2020-09-21 13:13:05 | 77 | |
| 78 | 2020-09-21 13:11:34 | 0 | 77997 | 2020-09-21 13:11:34 | 2020-09-21 13:11:34 | 78 | |
| 79 | 2020-09-21 13:11:36 | 0 | 77997 | 2020-09-21 07:10:47 | 2020-09-21 13:13:05 | 79 | |
| 80 | 2020-09-21 13:13:05 | 0 | 77997 | 2020-09-21 07:10:47 | 2020-09-21 13:13:05 | 80 | ← Correct end Should be the same session? |

| | server_time | session_hk_c | visitor_id_c | visit_first_action_time | visit_last_action_time | step_session | |
|---|---|---|---|---|---|---|---|
| 5149 | 2020-09-29 14:17:50 | 741 | 28215 | 2020-09-29 14:17:50 | 2020-09-29 14:55:35 | 0 | ← Correct |
| 5150 | 2020-09-29 14:18:05 | 741 | 28215 | 2020-09-29 14:17:50 | 2020-09-29 14:55:35 | 1 | |
| 5151 | 2020-09-29 14:18:38 | 741 | 28215 | 2020-09-29 14:17:50 | 2020-09-29 14:55:35 | 2 | |
| 5152 | 2020-09-29 14:18:39 | 741 | 28215 | 2020-09-29 14:17:50 | 2020-09-29 14:55:35 | 3 | |
| ⋮ | | | | | | | |
| 5160 | 2020-09-29 14:23:42 | 741 | 28215 | 2020-09-29 14:17:50 | 2020-09-29 14:55:35 | 11 | |
| 5161 | 2020-09-29 14:23:44 | 741 | 28215 | 2020-09-29 14:17:50 | 2020-09-29 14:55:35 | 12 | |
| 5162 | 2020-09-29 14:23:57 | 741 | 28215 | 2020-09-29 14:17:50 | 2020-09-29 14:55:35 | 13 | |
| 5163 | 2020-09-29 14:24:07 | 741 | 28215 | 2020-09-29 14:17:50 | 2020-09-29 14:55:35 | 14 | ← Last action after more than 30 min? |

### 3.4. Ignore sessions with cyclic events with the same duration.

After apply the previous rule, another non-human behaviour was identified.
It was observed that sessions with long duration have a cyclical and periodic behaviour in the duration of their events.
This could be produced by a java script error or some periodical event that keeps alive the session.

The second rule to mark sessions as not valid was find a sequence of pairs of events with similar duration (10 seconds of difference at most) with durations of more than 10 seconds.

This rule affects **2.607/488.032** of the sessions.

Below is an example of this type of session where we can see how the user performs some relatively short events and some changes of application and after the eighth event, all the events follow the sequence. This pattern lasts around 18 hours.

| | session_hk_c | visitor_id_c | appname_int | server_time | event_duration | |
|---|---|---|---|---|---|---|
| 1077003 | 147161 | 285563 | 61 | 2020-09-22 11:48:36 | 19.0 | |
| 1077004 | 147161 | 285563 | 28 | 2020-09-22 11:48:55 | 151.0 | |
| 1077005 | 147161 | 285563 | 11 | 2020-09-22 11:51:26 | 9.0 | |
| 1077006 | 147161 | 285563 | 31 | 2020-09-22 11:51:35 | 1.0 | |
| 1077007 | 147161 | 285563 | 61 | 2020-09-22 11:51:36 | 557.0 | |
| 1077008 | 147161 | 285563 | 61 | 2020-09-22 12:00:53 | 7.0 | |
| 1077009 | 147161 | 285563 | 61 | 2020-09-22 12:01:00 | 9.0 | |
| 1077010 | 147161 | 285563 | 28 | 2020-09-22 12:01:09 | 1215.0 | ~ 10 min |
| 1077011 | 147161 | 285563 | 28 | 2020-09-22 12:21:24 | 611.0 | ~ 20 min |
| 1077012 | 147161 | 285563 | 28 | 2020-09-22 12:31:35 | 1191.0 | ~ 10 min |
| 1077013 | 147161 | 285563 | 28 | 2020-09-22 12:51:26 | 611.0 | ~ 20 min |
| 1077014 | 147161 | 285563 | 28 | 2020-09-22 13:01:37 | 1191.0 | |
| 1077015 | 147161 | 285563 | 28 | 2020-09-22 13:21:28 | 610.0 | |
| 1077016 | 147161 | 285563 | 28 | 2020-09-22 13:31:38 | 1191.0 | |
| 1077017 | 147161 | 285563 | 28 | 2020-09-22 13:51:29 | 610.0 | |
| 1077018 | 147161 | 285563 | 28 | 2020-09-22 14:01:39 | 1192.0 | |
| 1077019 | 147161 | 285563 | 28 | 2020-09-22 14:21:31 | 609.0 | |
| ... | | | | | | 40 h (same pattern) |
| 1077182 | 147161 | 285563 | 28 | 2020-09-24 07:03:16 | 1197.0 | |
| 1077183 | 147161 | 285563 | 28 | 2020-09-24 07:23:13 | 607.0 | |
| 1077184 | 147161 | 285563 | 28 | 2020-09-24 07:33:20 | 308.0 | Possible new session |
| 1077185 | 147161 | 285563 | 31 | 2020-09-24 07:38:28 | 1.0 | |
| 1077186 | 147161 | 285563 | 31 | 2020-09-24 07:38:29 | 4.0 | |
| 1077187 | 147161 | 285563 | 31 | 2020-09-24 07:38:33 | 4.0 | |

### 3.5. Ignore sessions with only one event with a duration less than a second

With this rule, the sessions that contain only one event with duration less than one second are marked as outliers and removed.
This rule affects to **53.738/488.032** sessions but even with this constraint, there are still many sessions with several days duration.

### 3.6. Ignore sessions that only use Global Brand Database and could be scrappers

After apply the previous rules, there are still sessions lasting several days. Commonly these sessions only use the Global Brand Database application.
In sessions with only this application another unusual behaviour has been detected.
There are multiple sessions with durations of days but with constant events of a duration of few seconds.

In this rule, sessions that only use Global Brand Database application, have more than 20 events and the mean event time is lower than 10 seconds are marked as outliers and ignored.

This rule affects to **1.205/488.032** sessions.

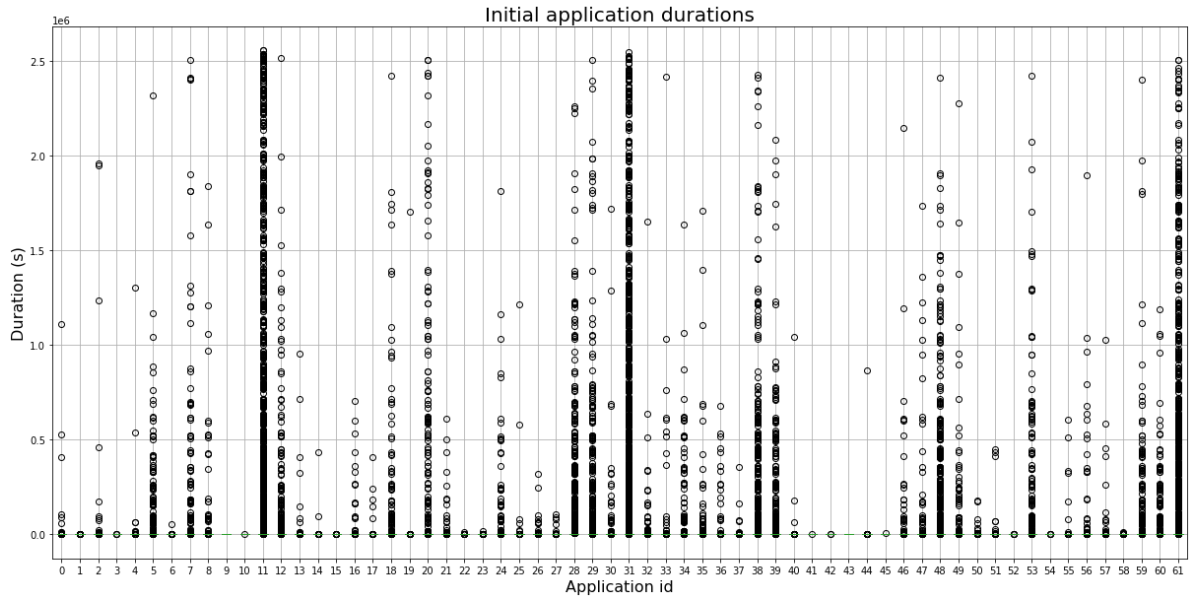
Below, we can see an example of possible scrapper of Global Brand Database, this user perform an event each 4-5 seconds during around 13 hours.

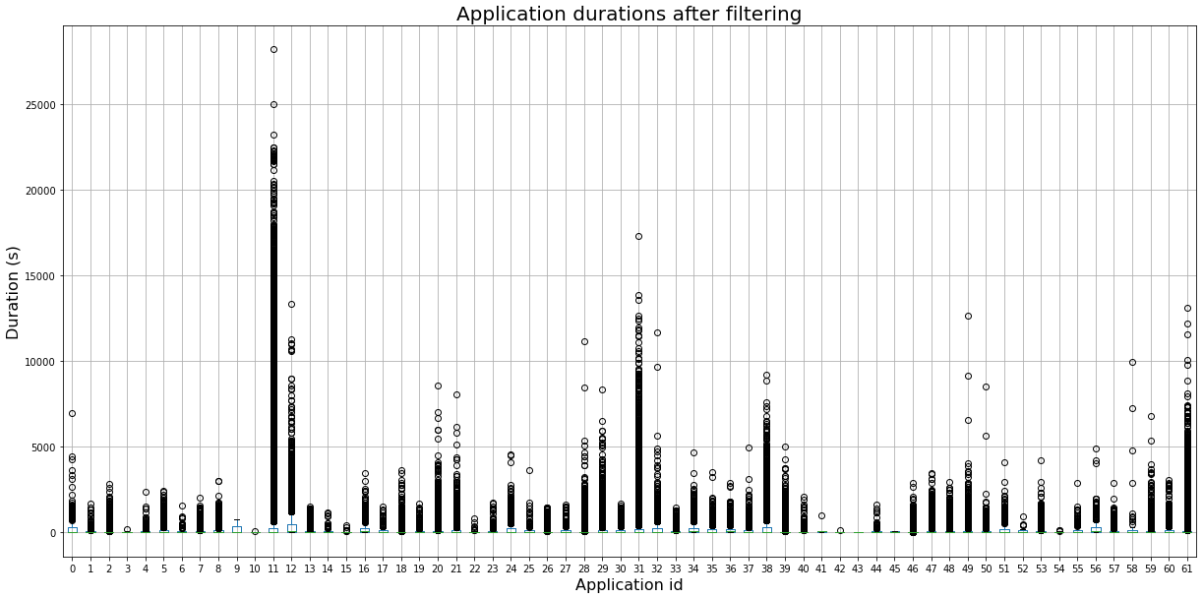
| | session_hk_c | visitor_id_c | appname_int | server_time | event_duration |
|---|---|---|---|---|---|
| 1304222 | 177795 | 220583 | 11 | 2020-09-07 10:43:04 | 6.0 |
| 1304223 | 177795 | 220583 | 11 | 2020-09-07 10:43:10 | 3.0 |
| 1304224 | 177795 | 220583 | 11 | 2020-09-07 10:43:13 | 3.0 |
| 1304225 | 177795 | 220583 | 11 | 2020-09-07 10:43:16 | 3.0 |
| 1304226 | 177795 | 220583 | 11 | 2020-09-07 10:43:19 | 3.0 |
| 1304227 | 177795 | 220583 | 11 | 2020-09-07 10:43:22 | 2.0 |
| 1304228 | 177795 | 220583 | 11 | 2020-09-07 10:43:24 | 3.0 |
| 1304229 | 177795 | 220583 | 11 | 2020-09-07 10:43:27 | 3.0 |
| 1304230 | 177795 | 220583 | 11 | 2020-09-07 10:43:30 | 3.0 |
| 1304231 | 177795 | 220583 | 11 | 2020-09-07 10:43:33 | 2.0 |
| 1304232 | 177795 | 220583 | 11 | 2020-09-07 10:43:35 | 3.0 |
| 1304233 | 177795 | 220583 | 11 | 2020-09-07 10:43:38 | 3.0 |
| 1304234 | 177795 | 220583 | 11 | 2020-09-07 10:43:41 | 3.0 |
| 1304235 | 177795 | 220583 | 11 | 2020-09-07 10:43:44 | 3.0 |
| ⋮ | | | | | |
| 1322473 | 177795 | 220583 | 11 | 2020-09-08 00:26:30 | 2.0 |
| 1322474 | 177795 | 220583 | 11 | 2020-09-08 00:26:32 | 3.0 |
| 1322475 | 177795 | 220583 | 11 | 2020-09-08 00:26:35 | 4.0 |
| 1322476 | 177795 | 220583 | 11 | 2020-09-08 00:26:39 | 2.0 |

13 h
(18256 consecutive events
with duration < 10 segs)

# IP Portal

After apply all the rules (some sessions are marked by multiple rules), the number of sessions has been reduced by **17% from 488.032 to 406.104** and, as we can see, the sessions have durations between a few seconds and seven hours.

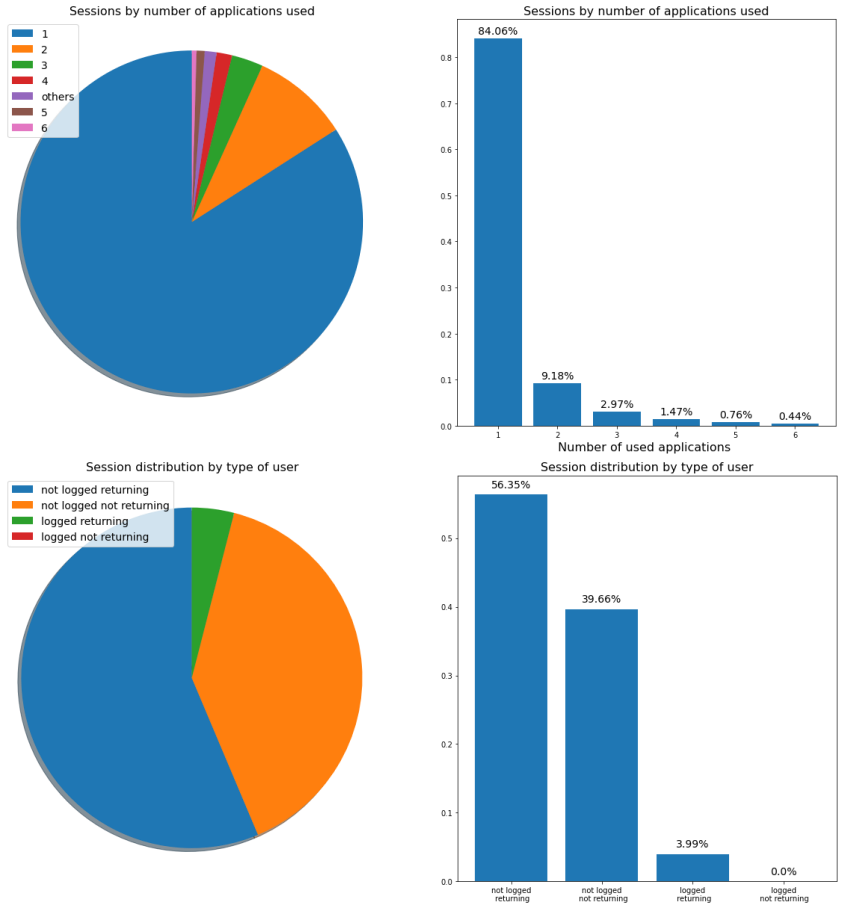


## 4. Sessions with one application analysis

After remove sessions with non-human behaviour, we can try to select what kind of sessions are useful to try to understand which sessions are more important to understand how familiar users use applications.

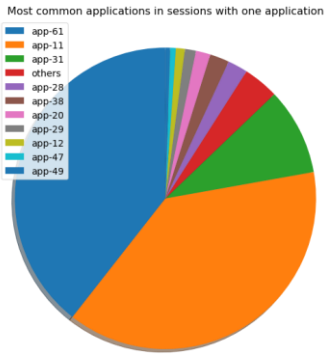
The distribution of the sessions based on their number of applications and their type of user remains quite similar to the original data.
We can see that most of sessions have a single application and the 56.35% of users are not experimented users.

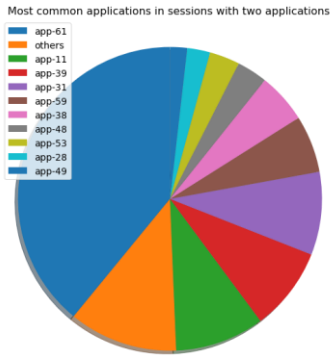


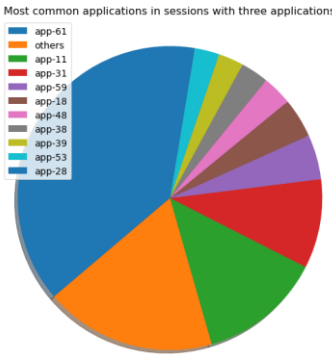

We are interested in having the largest number of sessions that use different applications, but it is also important that these sessions are from users with some experience.

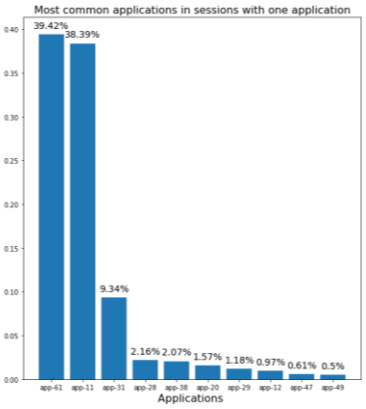
# IP Portal

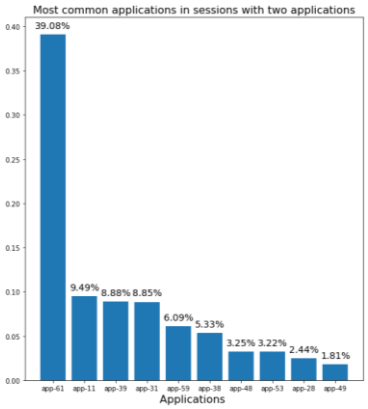
If we take a look to the sessions depending of their number of applications, we can see that in the case of sessions with a single app, the 80% of the sessions are from 2 of the 62 available applications (wipo.int and Global Brand Database). So the sessions with a single app will not be very informative.

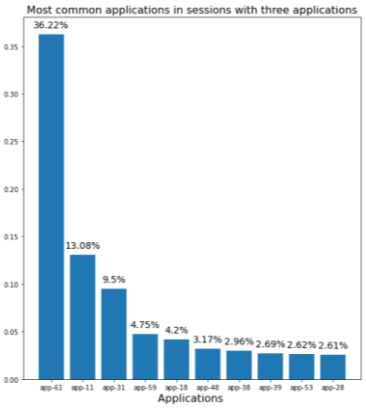


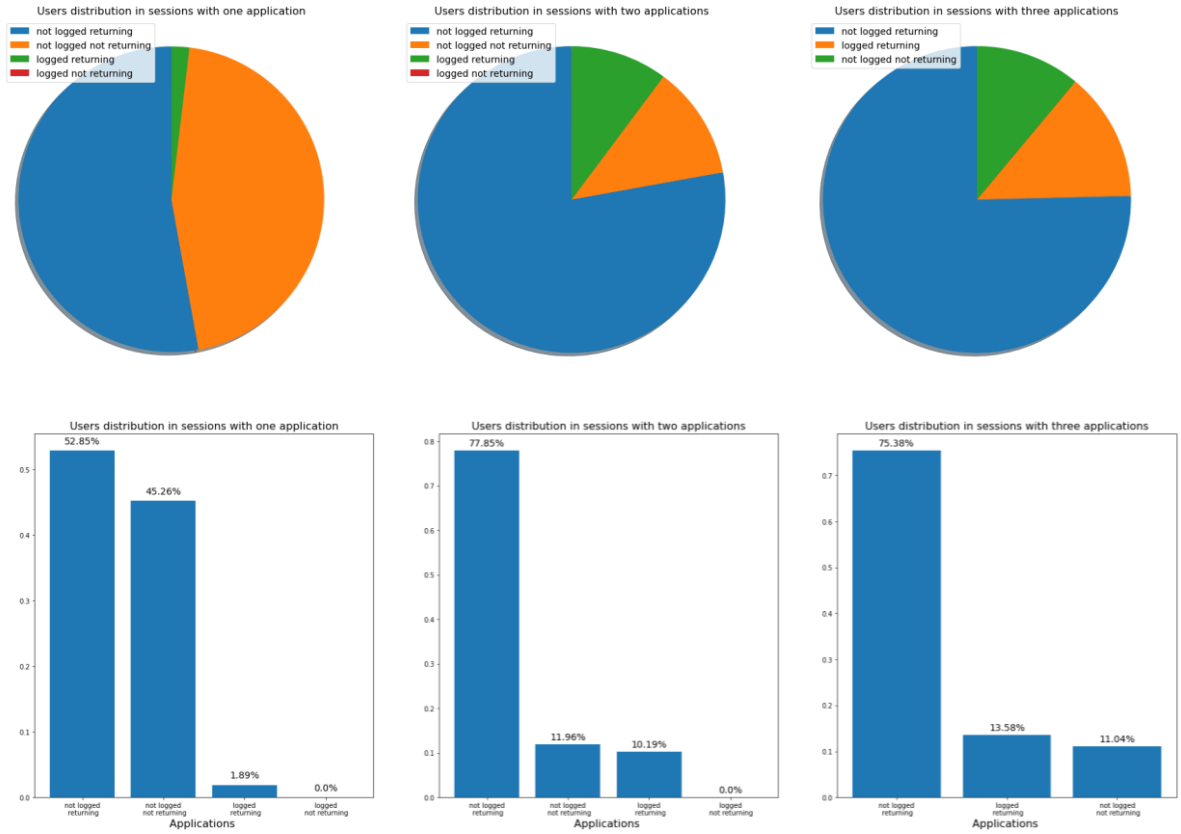



# IP Portal

If we take a look to the sessions depending of their number of experimented users, we can see that in the case of sessions with a single app, the 45% of the users are not logged users that not come back. In sessions with more than one application, these users represent 10% or less.




Due to the limited variety of applications, the large number of inexperienced users and the difficulty of knowing what the user is trying to do, we recommend filtering sessions with a single application.

## 5. Data analysis comparison

In this comparison will be shown the differences between the original dataset, the dataset after apply the filtering rules and the dataset after apply the filtering rules and remove the sessions with a single application.

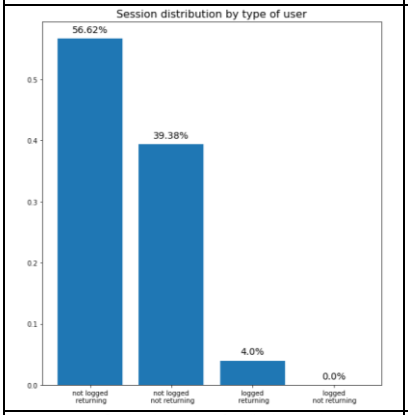
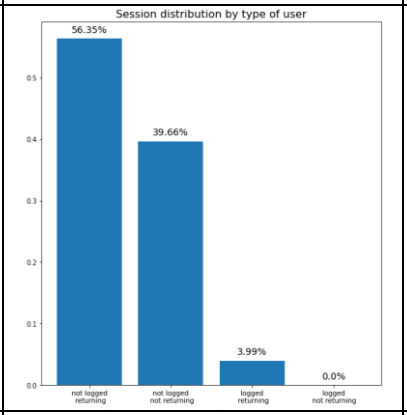
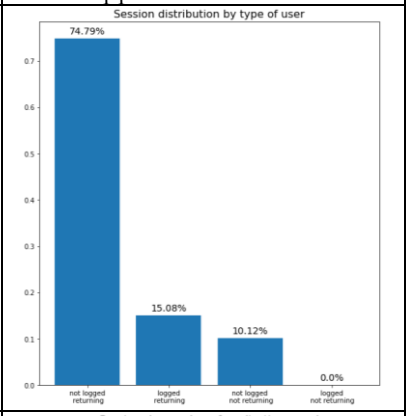
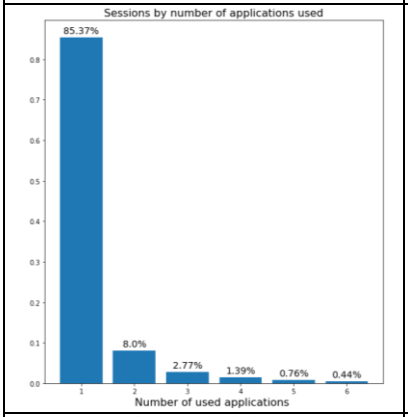
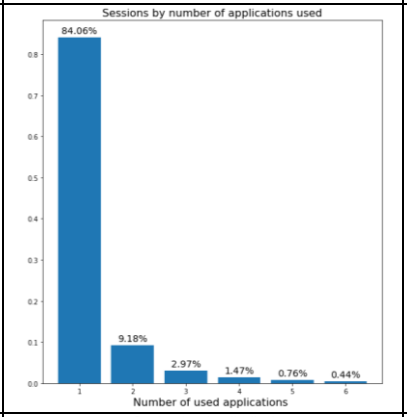
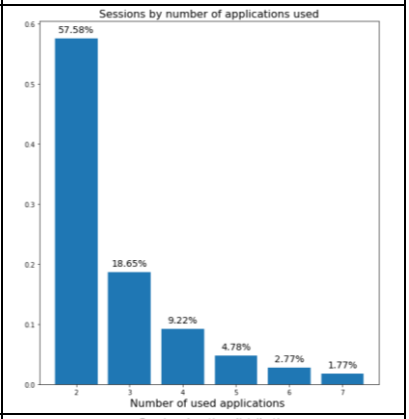
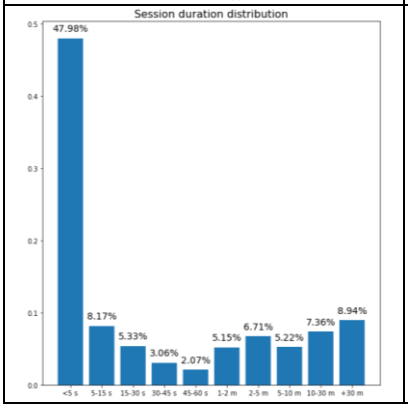
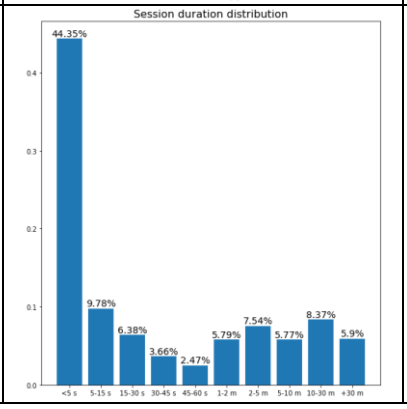
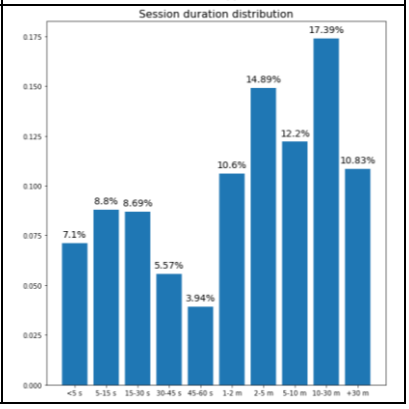
| | Logged users that come back | Logged users that not come back | Not logged users that come back | Not logged users that not come back |
|---|---|---|---|---|
| **Original dataset** | 19.515 (8.141 unique) | 7 | 276.345 (155.589 unique) | 192.165 |
| **Filtered dataset** | 16.211 (7.599 unique) | 6 | 228.832 (137.053 unique) | 161.055 |
| **Filtered dataset without single application sessions** | 9.758 (6.234 unique) | 3 | 48.403 (39.468 unique) | 6.552 |

*Table 1: Users distribution*

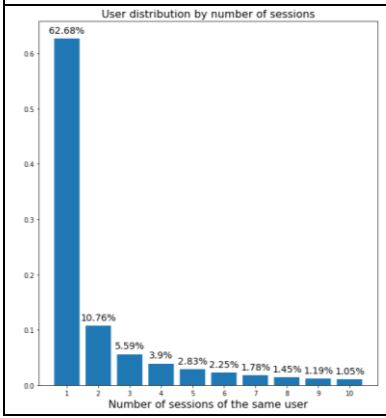
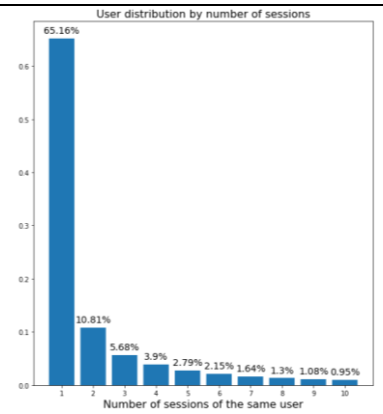
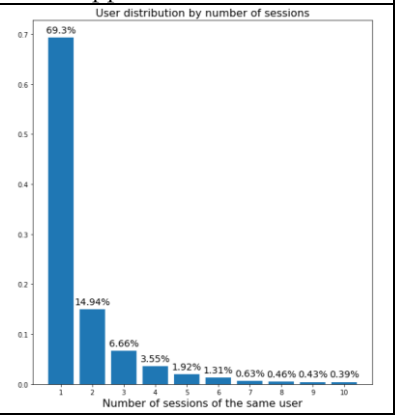
| | Sessions with 1 application | Sessions with 2 applications | Sessions with 3 applications | Sessions with more than 3 applications |
|---|---|---|---|---|
| **Original dataset** | 416.621 | 39.065 | 13.530 | 18.816 |
| **Filtered dataset** | 341.388 | 37.264 | 12.068 | 15.384 |
| **Filtered dataset without single application sessions** | 0 | 37.264 | 12.068 | 15.384 |

*Table 2: Sessions distribution by number of applications*

# IP Portal

| Original dataset | Filtered dataset | Filtered dataset without single application sessions |
|---|---|---|
|  |  |  |
|  |  |  |
|  |  |  |

# IP Portal

| Original dataset | Filtered dataset | Filtered dataset without single application sessions |
|---|---|---|
|  |  |  |