Greedy Game Assignment

Dataframe Size: 171721 rows

Data Discrepancy : 4793 redundant Data rows (Server Received the same information at different time)

Data Loss: 13426 missing Data rows

1. 9378 data rows with GGSTOP event
2. 4048 data rows with GGSTART event

Outlier removal:The corresponding data for these rows are extracted from the data frame and stored in different frames.

Calculations on new data frame :-

Name: test

Data Frame Size: 153502 rows

Total Sessions: 22424

Total Valid Sessions: 14042

Average valid session time: 1546.323 sec

(Total valid session time / total valid Sessions)

User behaviour:

Total users in test: 14678

Most played game: Game with game id “55107008”

User : 9708

(timespan: 1980 to 2016)

**Calculation**:

Language Used : R

Library used : plyr, dplyr

1. First redundant data rows from the server have been removed by removing timestamp and applying “unique”. As these are datas with same information received at different time at the server.
2. Then missing data rows have been searched and extracted by searching for unpaired rows of ggstart or ggstop and data frame is ordered.
3. For some users, only ggstart or ggstop rows are present.
4. For some users uneven numbers of ggstart or ggstop are present. In this case, out of the repeated ggstart or ggstop, only last one for ggstart and first one for ggstop are taken others are extracted.
5. The new data frame test is of size 153502 rows with 76751 rows with ggstart event and 76751 rows with ggstopp event.
6. For consecutive rows,time difference has been calculated and shown in a new column as ”tdiff” where even columns give the time of a session and odd columns give the time between the end of a session and start of the next session .
7. Three Functions as Valid\_count(), Session\_count(),Valid\_sessionlength()

are used to find out the number of sessions, number of valid sessions and valid session length.

1. In each function one inner loop compares the time difference between different sessions (odd rows) to check whether it is > or < 30 secs. The outer loop act correspondingly and keeps the count and session lengths in variable “counter” and “total\_sum”
2. These functions are applied with ddply on id ai5 to find session and valid sessions count for each group and counts are summed up to find total sessions and total valid sessions as well as total valid session time. For average valid time, total time is divided by total valid session count.

**Minimization of data loss**: Data rows of users only with ggstart or ggstop event can be examined and number of sessions (not Valid) can be found out. If two ggstart events for one user take place between 30 secs then it can be assumed that those two events are part of one session and correspondingly an assumption can be made about the least number of sessions. For these purpose test\_start with 9378 data rows of “ggstart” of different users and test\_stop with 4048 data rows of “ggstop” are created. One function as session\_loss\_count is also provided for finding out relation between data rows with missing counterpart.