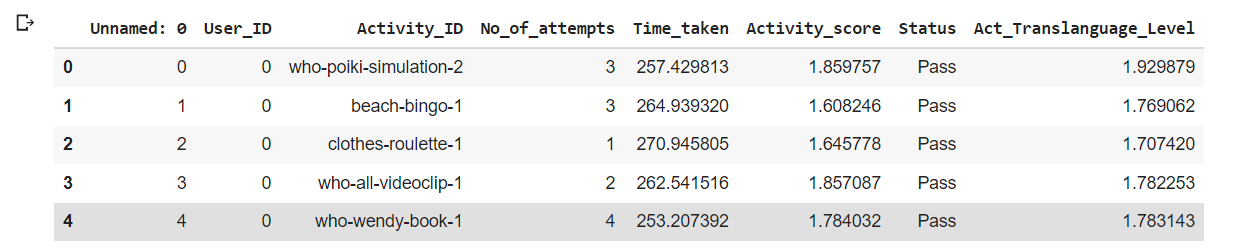
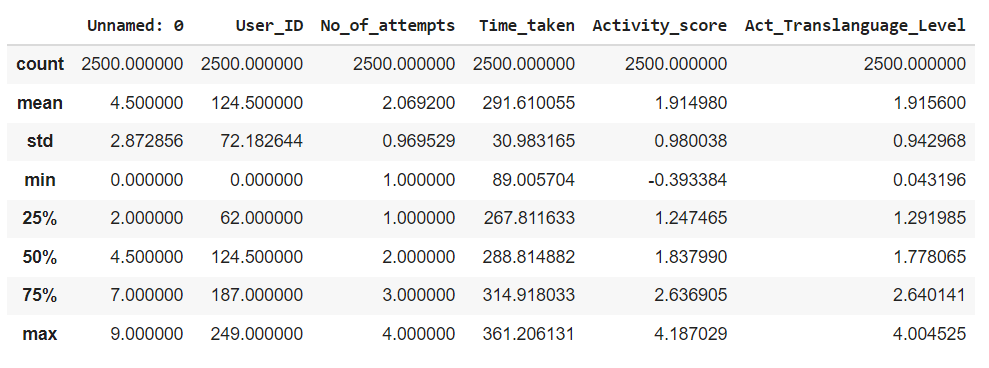
EDA for Activities data

\* Special thanks to Task6 for generating such an insightful data manually

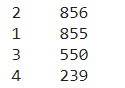
Here is the link for the [notebook](https://colab.research.google.com/drive/176L2Cj1nCXG4u2h3NOAl64ODiAyBfdpP#scrollTo=14QCMMSml4RJ)

* The data has 2500 different entries pertaining to 250 users. Each user has an entry of 10 activities.
* The columns that we have are as follows
  + User\_ID
  + Activity\_ID
  + No\_of\_attempts
  + Time\_taken
  + Activity\_score
  + Status
  + Trans\_Language\_level
* None of the entries are ‘null’ valued



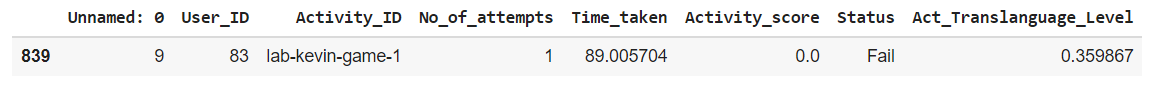


* Average number of attempts to clear the activities are 2.



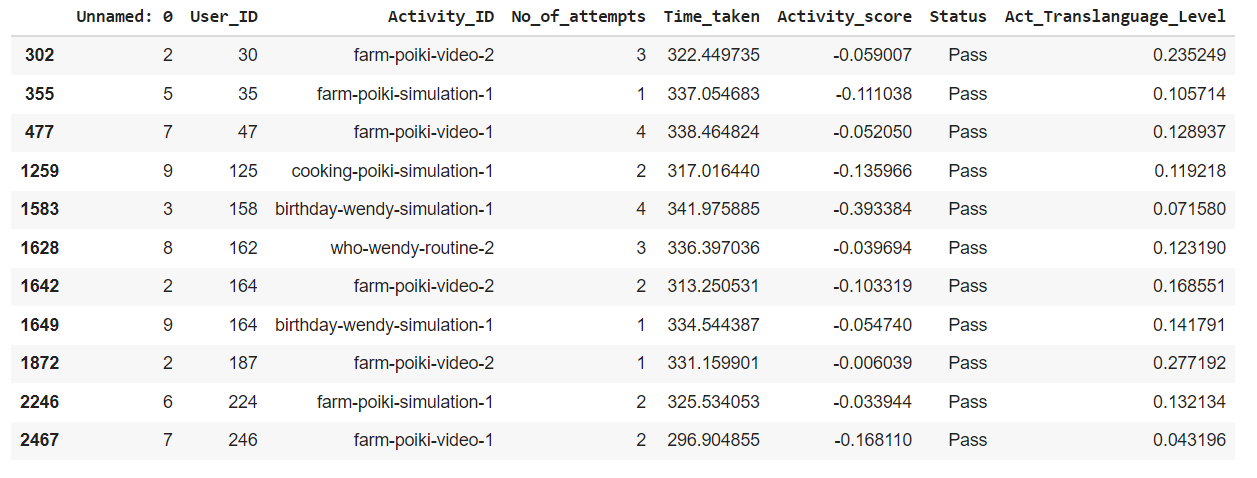
* The average time taken for completion of activity is 290 seconds with a standard deviation of 30 seconds.
* The pass/fail frequency is as follows





The one entry that failed the activity

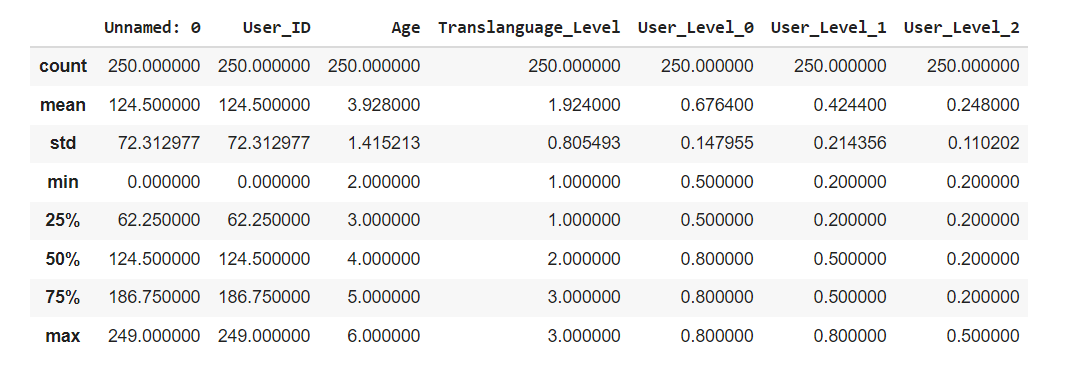
* 60 out of possible 146 activities are used in the dataset.
* There are 11 entries with a score less than zero.



EDA for Users data

Here is the link for the [notebook](https://colab.research.google.com/drive/1MYF1XJ6tZYGasGwcJtQmuy_JQTnDTIeP#scrollTo=uHAVyQUVpJR7)

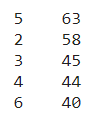
* We have 250 different users in our database.
* The columns are as follows
  + User\_ID
  + Age
  + Translanguage\_Level
  + Literate
  + Learning\_Rate
  + Learning\_Profile
  + Country\_of\_Birth
  + Country\_of\_Residence
  + Cultural\_Heritage
  + Main\_Language
  + Learning\_Language
  + User\_Level\_0
  + User\_Level\_1
  + User\_Level\_2
  + Gender
  + User\_Name



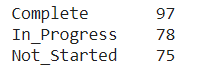
* We have the following gender based demography.



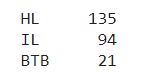
* We have the following age based demography.



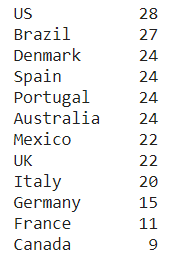
* Here is also the distribution for type of progress in our dataset



* Learning profiles are as follows



* We have users from the following geographical demography.



* We have 26 users where country of birth is equal to country of residence
* **Our average length for cultural heritage is 2. But the standard deviation is quite high which suggests the cultural heritage may be an equal split among 1,2 and 3 values.**
* We have 4 users whose learning language is the same as their main language.
* Ids of those users are 39,65,96,244.

**Final thoughts**

* The activity data only employs 60 activities out of a total of 146, we can build a test set around the remaining activities.
* Since the number of unseen activities are more than seen, the recommendation system should (especially the collaborative filtration ) adapt efficiently to suggest unseen activities too.
* The split between pass/ fail is of the ratio of 2499:1. This skewness might affect the logic where the algorithm punishes users for not performing well on an activity.
* The ‘users’ data is a fairly balanced dataset.
* The activities for users[around 40] having cultural activities originating from Italy, France and Germany are quite few. Again the CF team might have trouble suggesting recommendations as every user will have to be given every activity to fulfill the quota beforehand.
* There are 4 users whose learning language and main language is the same [ won’t really matter, but i felt worth mentioning as this might not be the case when the real data comes in]
* Finally, there is a degree of randomness to the data as it is generated manually so our models might overfit and not work well when real data comes in. [ Just something to keep in mind]
* Next points pertain to the **activties\_dataset** given by stakeholders.
* Imbalanced set of activities for age groups. [ Since we have indifferent frequencies, one algo might not work across all groups]
* Age group 6 has less than 30 activities (27 activities). [ Poikolingo clarified last time that activities used for lower age groups can be extended to upper age groups but something to keep in mind.]
* Not all age groups have the same or similar activities.
* Cultures [French, German, Italian] are attributed to hardly any activities ( <10 activities).
* SEL activities are also few in number when compared against standard and cultural activities.