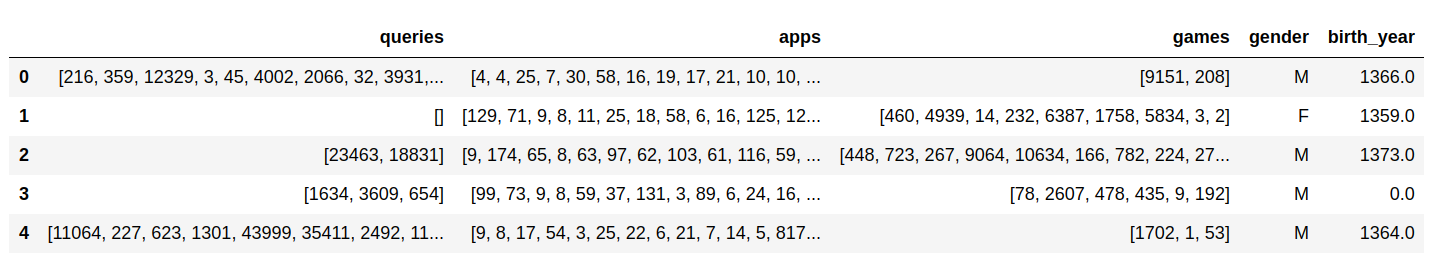
**تشخیص سن و جنسیت:**

در این پروژه قصد داریم تا سن و جنسیت کاربران را براساس **لیست برنامه‌ها، لیست بازی‌ها و لیست سرچ کوئری‌ها** مدل‌سازی کنیم و بتوانیم از مدل ساخته شده برای پیش‌بینی مابقی کاربران استفاده کنیم. بعضی از کاربران در صفحه‌ی پروفایل خود سن و جنسیت را وارد کردند. در شکل زیر نمونه‌ داده‌ی جمع‌آوری شده آورده شده است.



نکاتی در رابطه با داده مطرح است که در زیر به آنها اشاره شده:

* برنامه، بازی و کوئری با یک id جایگزین شدند. بخاطر بعضی از حساسیت‌ها در سرچ کوئری ها در شروع از خود متن کوئری صرف نظر کردم که احتمالا در ادامه دسترسی به این مورد مقدور میگردد.
* دیتاست تهیه شده نمونه‌ای تصادفی از کل داده به اندازه‌ی ۴۰۰۰۰ است هر چند که کل داده‌ای که کاربران پر کردند در مقیاس کاربران بازار زیاد نبوده و در حدود ۱۰ درصد از کل کاربران داده‌ی پروفایل خود را تکمیل کردند.
* تایپ فایل پارکت(parquet) بوده و ستون‌های aps,games و queries از نوع لیست هستند.
* توزیع داده‌ی تکمیل شده از سمت کاربران اریب به سمت کاربران فعال بازار بوده و با توزیع داده‌ای که قرار است مدل بتواند برای مابقی پیش‌بینی کند فرق می‌کند که این مورد احتمالا در مدل ‌سازی لحاظ باید گردد.
* دو ستون دیگر در داده وجود دارد که چون اکثر خالی بودند از آوردنشون خودداری کردم. یکی لیست ویدیوها و دیگری لیست سریالها که در ادامه پروژه این دوستون به داده‌ی اصلی اضافه خواهند شد.

نحوه‌ی ارزیابی فعلی balanced accuracy به صورت زیر است که این مورد ممکن است با بررسی به معیار بهتری تغییر پیدا کند.