На гистограммах в all\_raw\_separate.pdf в логарифмической шкале отражены все данные, включая повторения (также, предлоги и числа считаются за отдельные слова). All\_raw\_comparison – то же самое, но с наложением гистограмм.

В файле comparison\_regions.pdf видно, что в некоторых регионах (значение параметра lr) преобладает одна группа, а в каких-то все группы представлены примерно одинаково. Также есть два региона, в которых сосредоточено преобладающее число пользователей из групп.

Я также посмотрел, как варьируется параметр msid – оказалось, что в выборке присутствует 718 уникальных параметров, притом, что всего их 986. Поэтому решил msid в данном задании не рассматривать.

На гистограмме comparison\_family отражено, включён ли параметра family=запросе. При этом на гистограмме представлены 4 столбца, значения которых следубщие: [' ', '2', 'active', 'yes']. Если предполжить, что «2»= «active»= «yes», то можно видеть, что в группе Detroit довольно много запросов «безопасного поиска» (~2.5%) по сравнению с группами Ufa и Andorra (~0.2%).

Comparison\_usersactivity – количество разных запросов (разделённых по времени) от одного и того же пользователя.

Из 4216 пользователей 307 (7.3%) входят одновременно в две и более группы (число пользователей получено в скрипте , hist\_users\_shared - для наглядности)