We deal with the topic of #homeworkout and #gym tweets.

According to our limitations:

* Our samples: 1000 tweets each for both hashtags
* We limited our results to only tweets written in English language
* We included re-tweets
* Time and date for the request of the tweets was 03/12/2020 @ 13:59

**Interpretation of the top locations and map:**

According to the bar chart we can spot that most tweets were originated from the USA, and UK, which is not a surprise since we restricted our search to only tweets in English. Furthermore, we notice higher volume of tweets for #homeworkout for the top 5 locations.

Moving to the world map, we notice that #homeworkout is mostly used in the UK and USA with some appearances in north western India and in Philippines, while #gym appears more scattered, appearing as well in central and eastern Europe, Malaysia and Melbourne Australia.

It must be stated that #homeworkout has 896 NA location values, while # gym 460.

**Interpretation of unique words**

Based on the graph we can easily notice that the **fitness** word in #homeworkout tweets soars up to about 350, while for #gym the **fitness** escalates up to 200 unique counts. **Exercise** has about 60 more appearances in **#homeworkout** compared to **#gym**, while other words do not pass the 80 appearances for both cases. The word **App** appears frequently only in #homeworkout which probably indicates that users use an application for their workout in home. Which can be exploited for…

**Sentiment Analysis (MOST COMMON POS AND NEG, #HOME, #GYM)**

From the bar-charts we can observe that there is more volume of words with positive sentiment, with the word free have almost 70 appearances in #homework. In the other hand we got the word deprive soaring at about 60 for negative sentiment in #homeworkout related tweets, which might be related to users who are expressing their annoyance of the closure of gyms. Also, we can spot common words for both different tweets, like healthy, love for positive score and loss/lose, hard for negative.

Finally, it should be mentioned that some off those words within the text must have a different meaning. For example, users typed the word **free** might want to express that they want to work-out freely outside or in a gym, while the words of fat, hard, lose, and grind might be related to “hard workouts” and lose of fat/weight, which has a positive meaning.

**Sentiment Analysis (distribution of sentiment scores)**

Based on the visual observation of the density plots, we can assume that **#gym** sentiment scores follow a unimodal distribution (1 peak) and also the mean=median=1 , while for the **#homeworkout** follow a bimodal distribution with picks at -1 and 1. Also we can easily notice that the **#gym** mean sentiment is higher about 0.7. Such difference might had been resulted because we took our sample the date when lockdown in UK was lifted and gyms has opened.

However, in order to infer such difference to the population we must perform statistical testing. In this case a t-test could had been in use.

First, we should go through the assumptions of the t-test:

1. **scale of measurement, # yes (ordinal ?)**
2. **simple random sampling # not sure not all users are the same expressly (some of them do not post their opinion), and also not all the gym/homeworkout users use twitter, also we included only English tweets and gathered our sample only once time, rather than taking multiple samples in different days.**
3. **Adequacy of sample size # we got 1000 for each topic**
4. **Equality of variance # we got about 1 point of variance difference**
5. **Normality # according to the visual testing only #gym might follow a normal distribution while #homeworkout follows a bimodal.**

From the output, the p-value < 0.05 implying that both distributions of the data are significantly different from normal distribution. In other words, we cannot assume normality for none of them.

In this case we can use the Mann-Whitney U test for 2 independent samples. Based on their density plots we can assume that they have different shapes, to this end we can make a hypothesis about their mean ranks.

H0: the distribution of scores for the two groups are equal

the alternative hypothesis is as follows:

HA: the mean ranks of the two groups are not equal

The p-value = 3.6 x 10^-6 < 0.5. Therefore, at the 5% significance level, we reject the null hypothesis, and we conclude that the sentiment scores between tweets with #gym and #homeworkout are significant different.

<https://www.statsandr.com/blog/wilcoxon-test-in-r-how-to-compare-2-groups-under-the-non-normality-assumption/>

<https://statistics.laerd.com/statistical-guides/mann-whitney-u-test-assumptions.php>