Homework 2
(10 points)

Use the Indiegogo dataset (https://webrobots.io/indiegogo-dataset/) and download <u>five files</u> of data, preferable in different years.

1. For each of these categories* in the category of JSON element, check whether all keywords has a gaussian distribution. You should count the appearance of the keyword per month and then assign keyword month. e.g. "Education", "Jan", "2020", "32"

use qq plot Then, plot their distributions based on the number of year (use density plot). It means you should download the data for five years and then compare their frequency each one separately.

2. Compare following two categories: "Health & Fitness", "Fashion & Wearables" on year basis (2018, 2019, 2020).
        a. With three statistics tests, one parametric, two non-parametric tests and report results.
        b. Use the effect size test, to quantify the magnitude of differences.

3. Use three correlation coefficient tests (Pearson, Spearman, KendallTau) and report whether following two keywords have correlations: "Fashion & Wearables", "Health & Fitness".

You need to prepare a report on your tasks and findings along with a video file describing what you have done. You can copy paste your codes, its results and your description into a Word document, Python Notebook or you can use R notebook.

Your deadline for delivering this home work is written on the blackboard online. Please feel free to ask your question and prepare it for presentation for the next session.

* "Education", "Energy & Green Tech", "Health & Fitness", "Fashion & Wearables", "Wellness"