**Assignment 2**

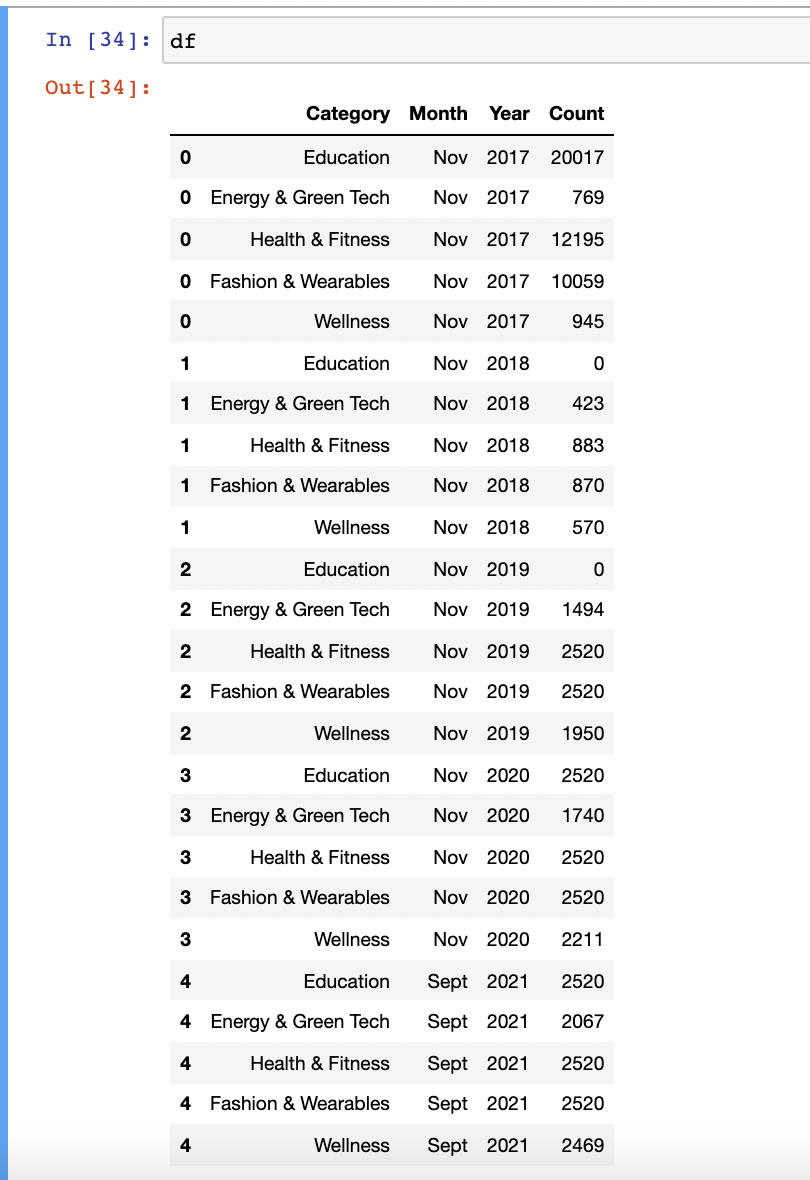
**Web Analytics and Mining**

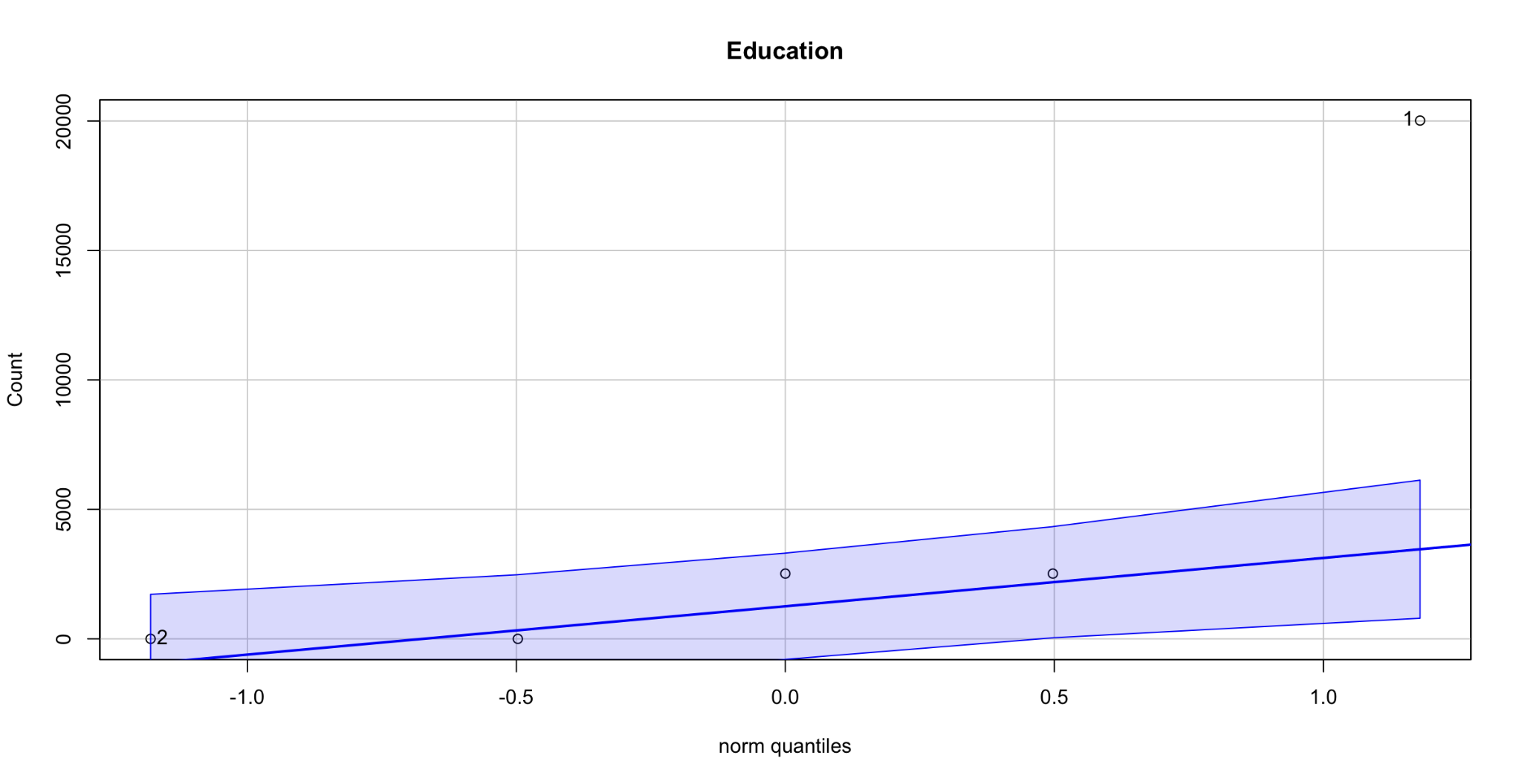
MET CS 688

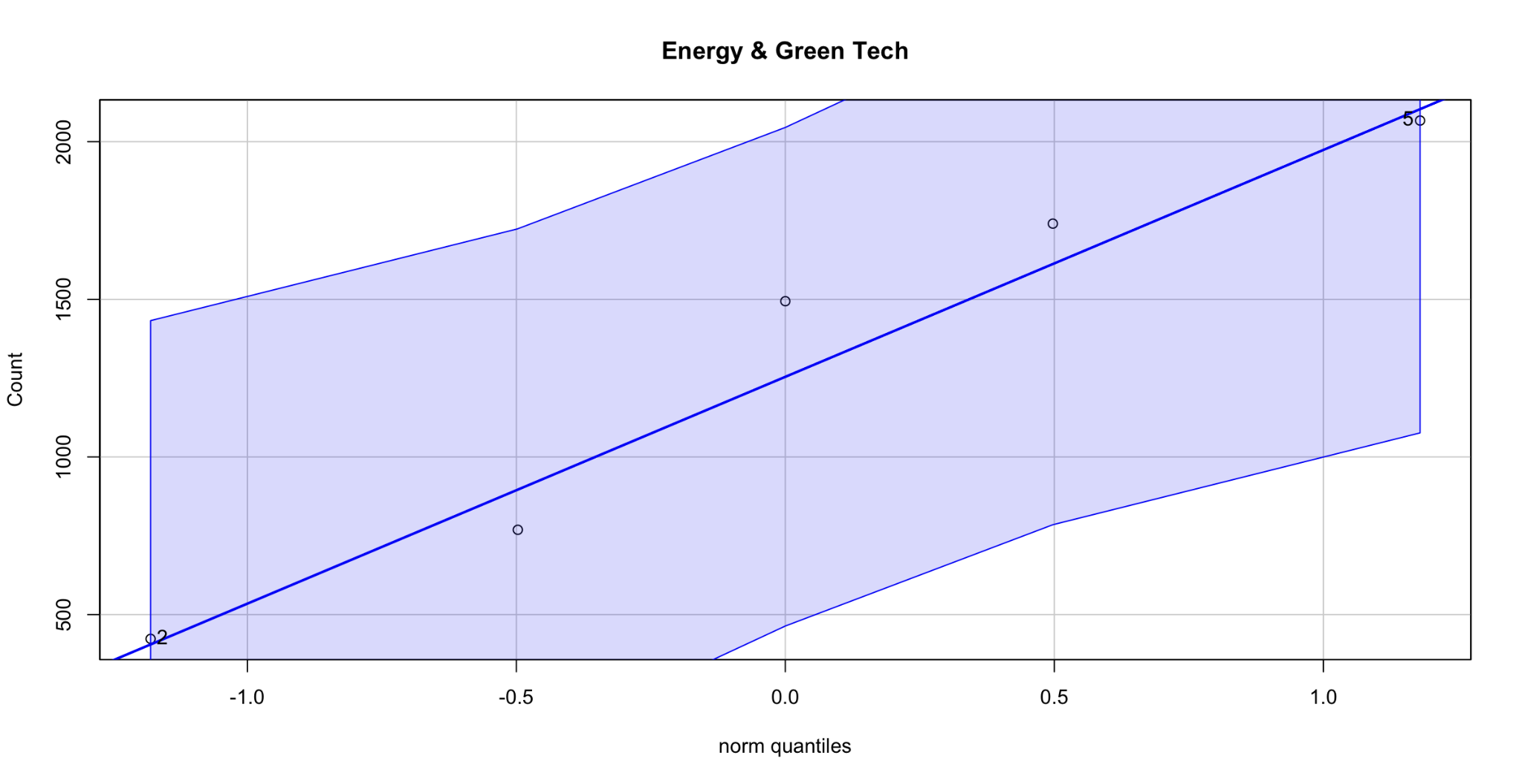
# Use the Indiegogo dataset (https://webrobots.io/indiegogo-dataset/) and download five files of data, preferable in different years.

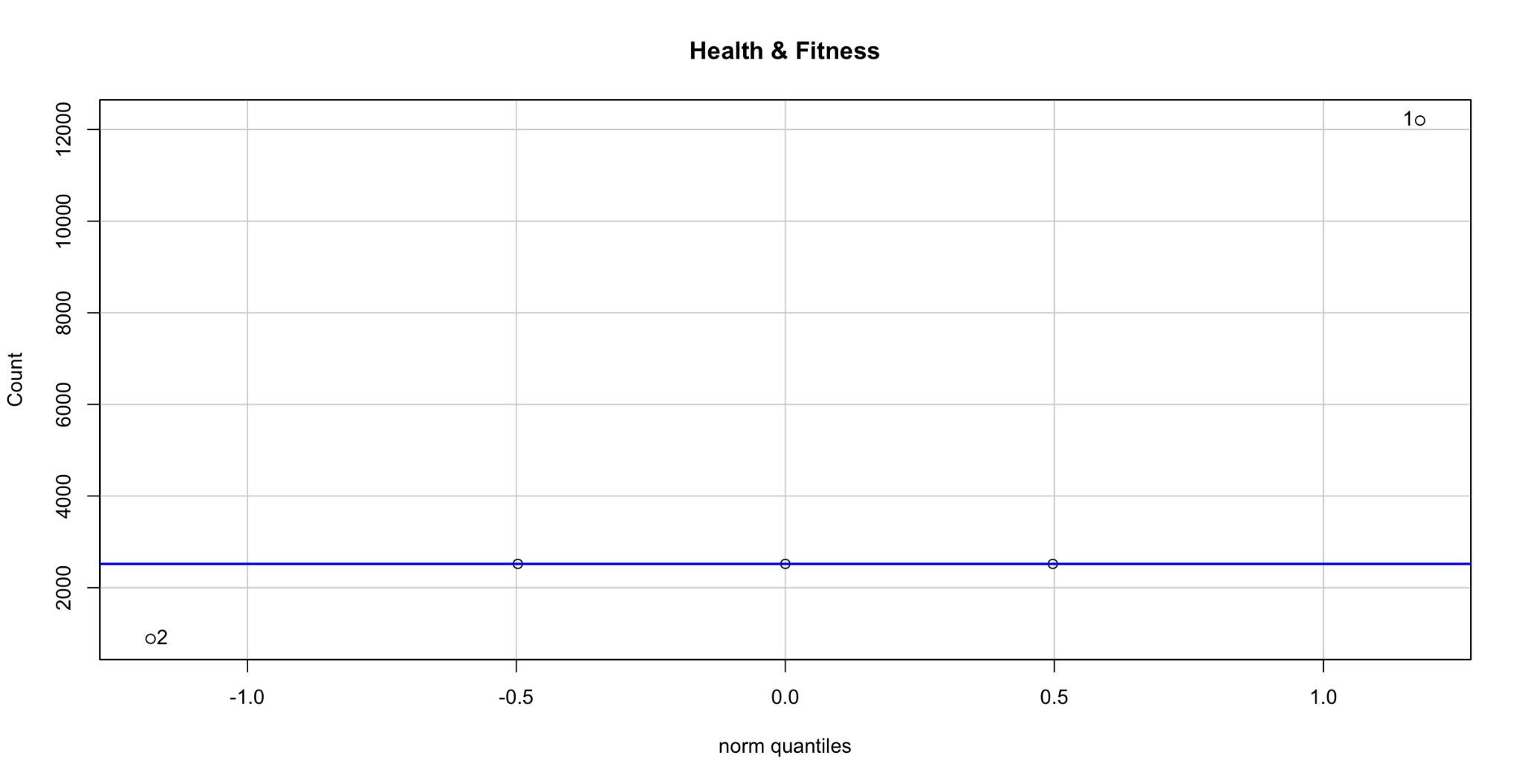
# **1.** For each of these categories\* in the category of JSON element, check whether all keywords have a Gaussian distribution. You should count the appearance of the keyword per month and then assign the keyword month. e.g., “Education,” “Jan,” “2020”, “32” Then, plot their distributions based on the number of years (use density plot). It means you should download the data for five years and then compare their frequency separately.

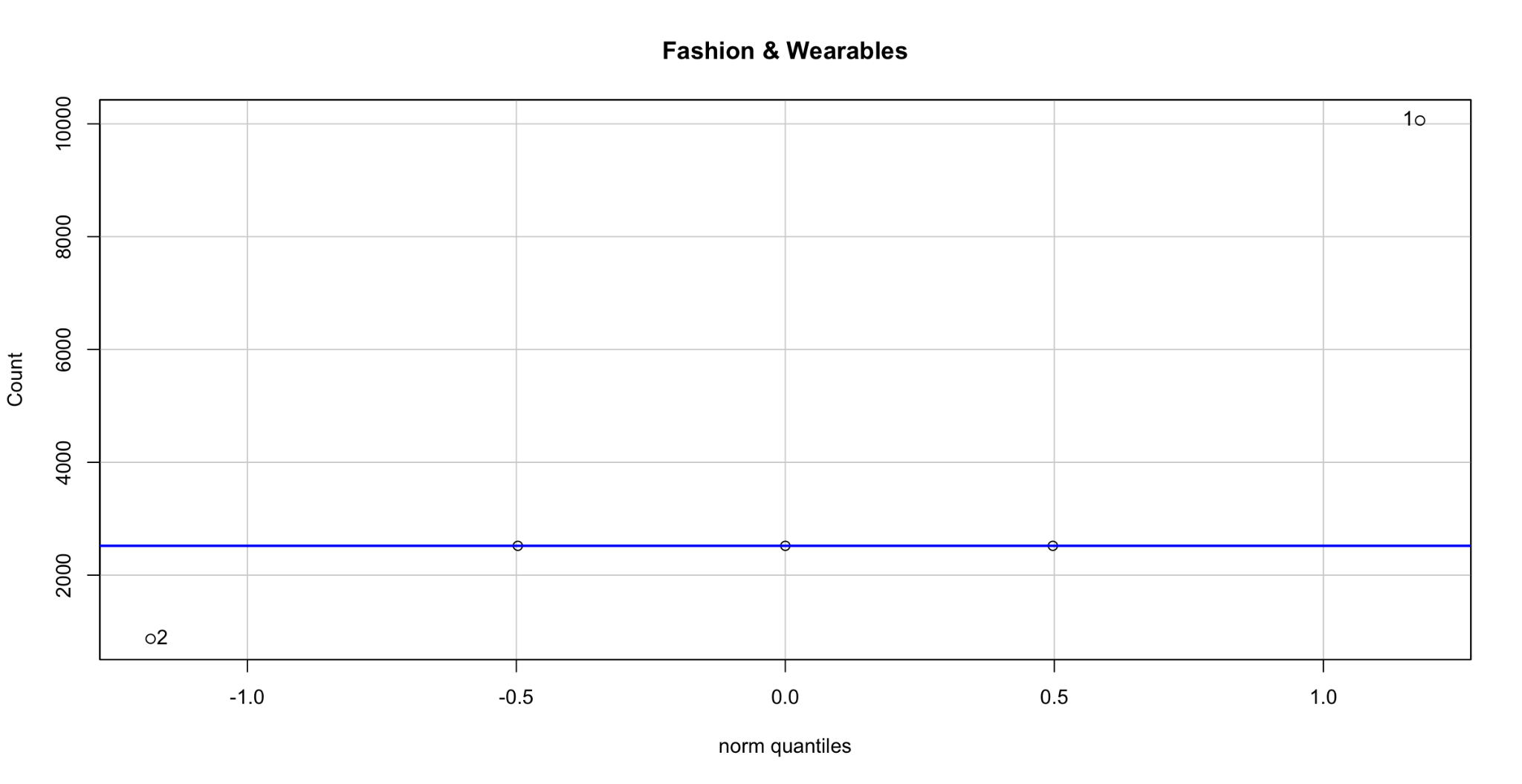
### Solution:

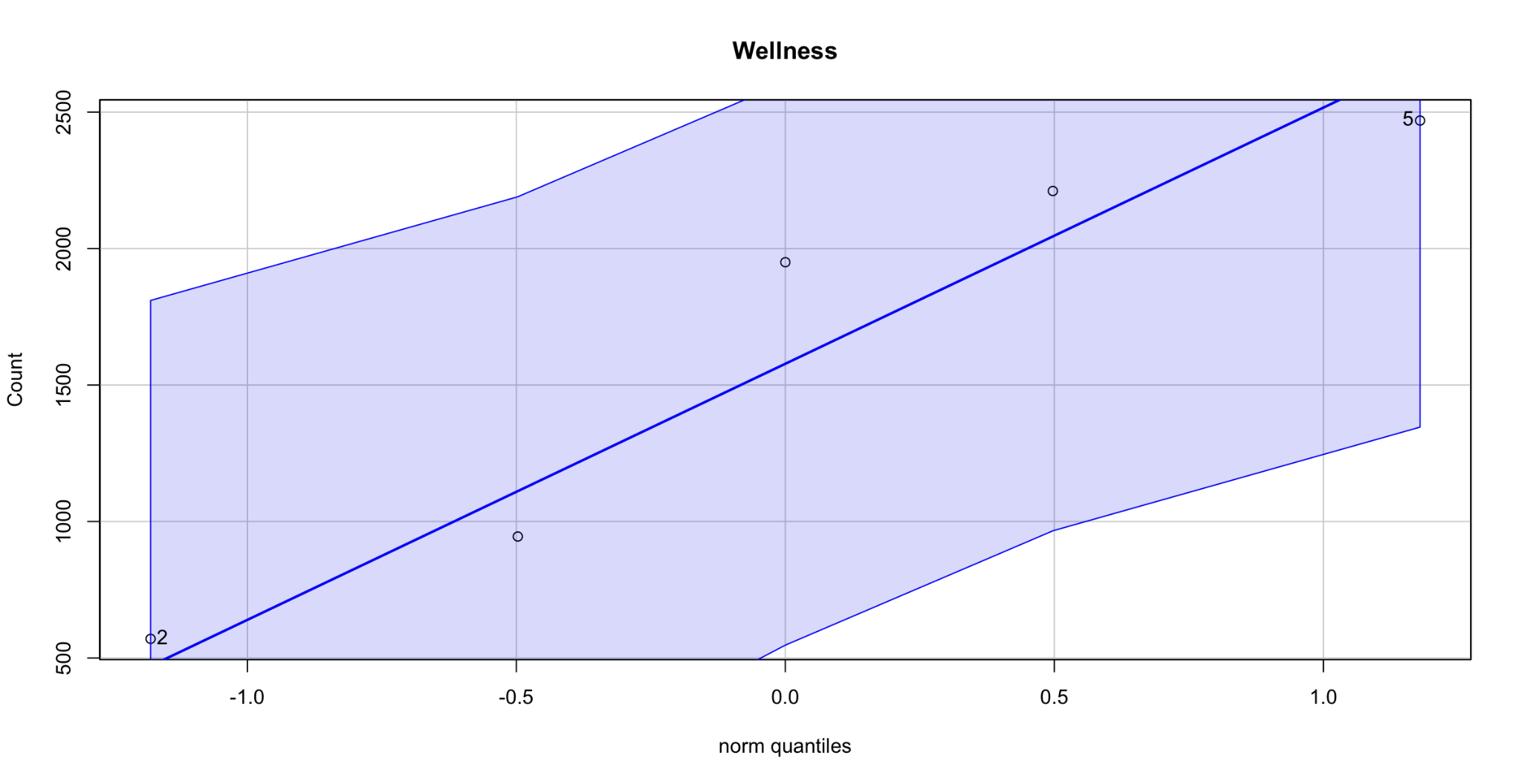




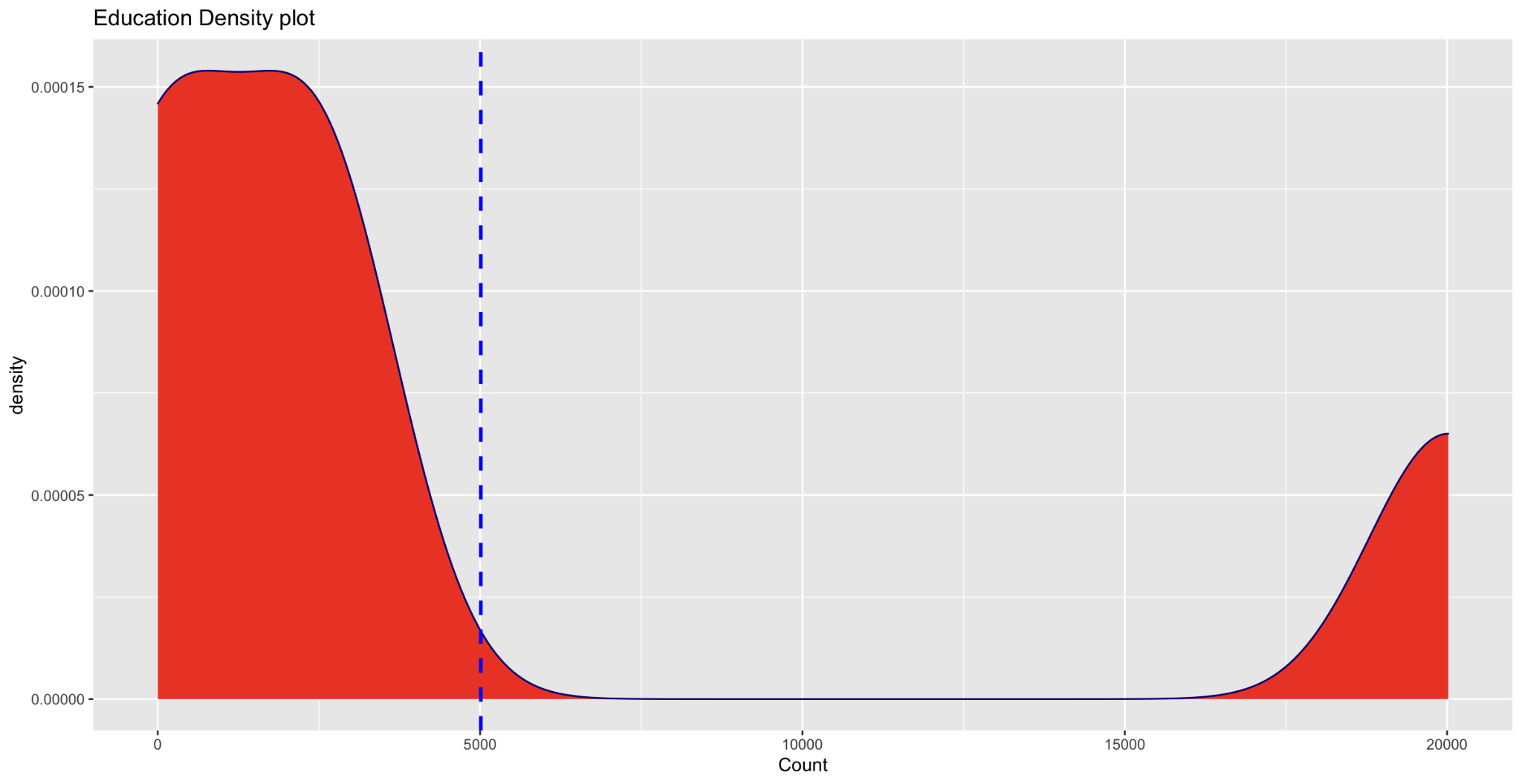


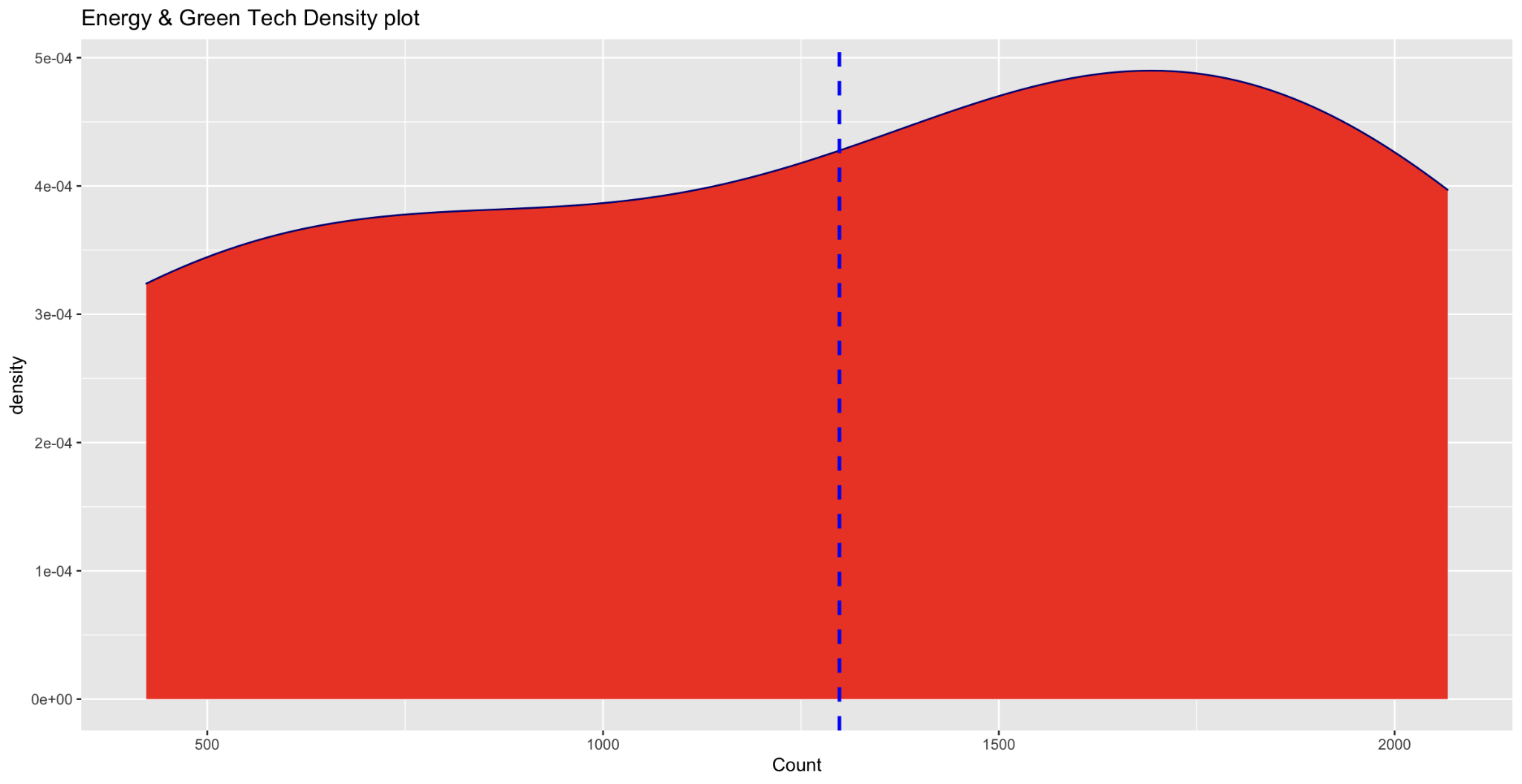


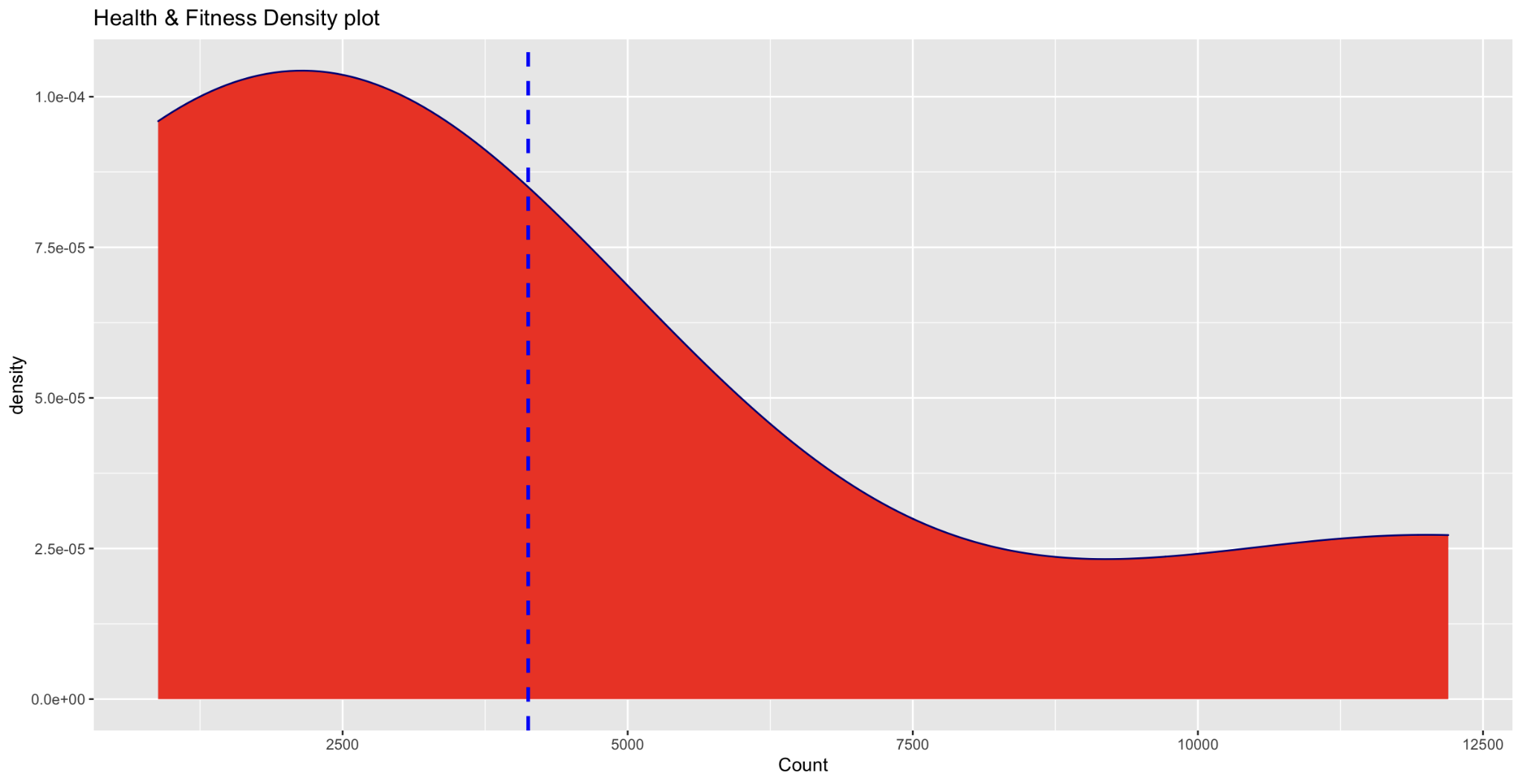


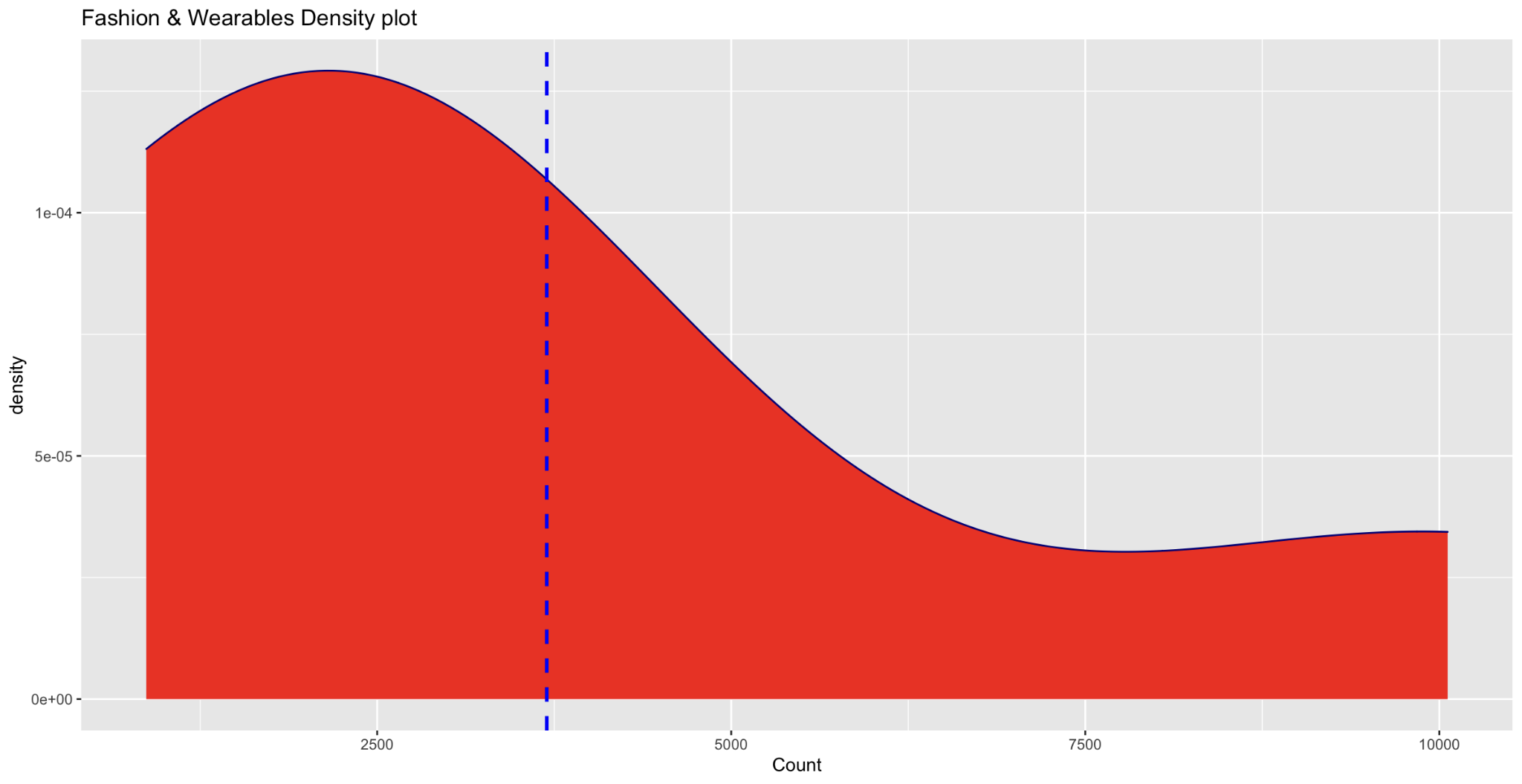


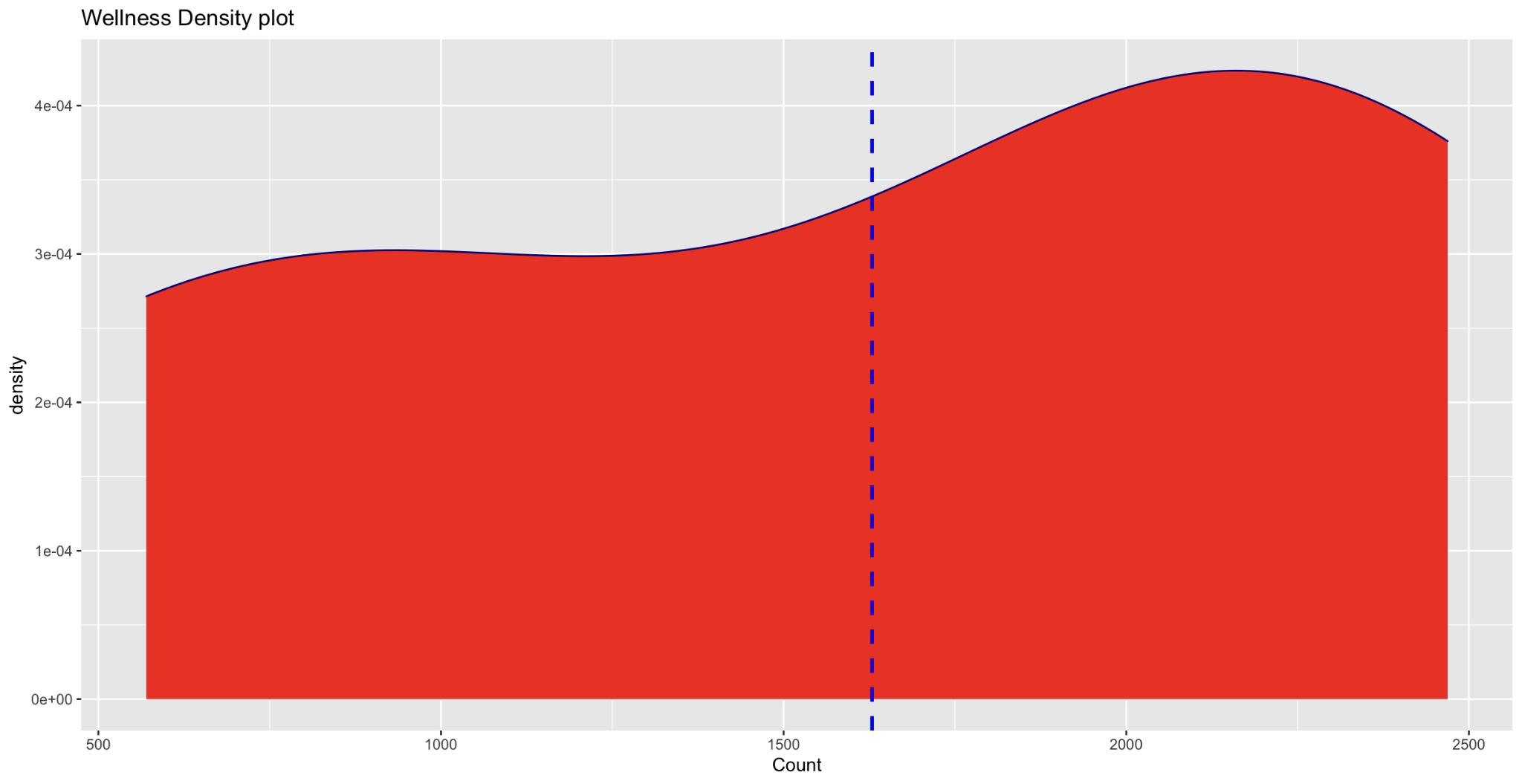
Education, Energy & Green Tech, Health & Fitness, Fashion & Wearables, and Wellness **follow Gaussian distribution.** Although, the distribution is skewed.











# **2.** Compare the following two categories: “Health & Fitness,” “Fashion & Wearables” on a year basis (2018, 2019, 2020).

## **a.** With three statistics tests, one parametric, two non-parametric tests, and report results.

### Solution:

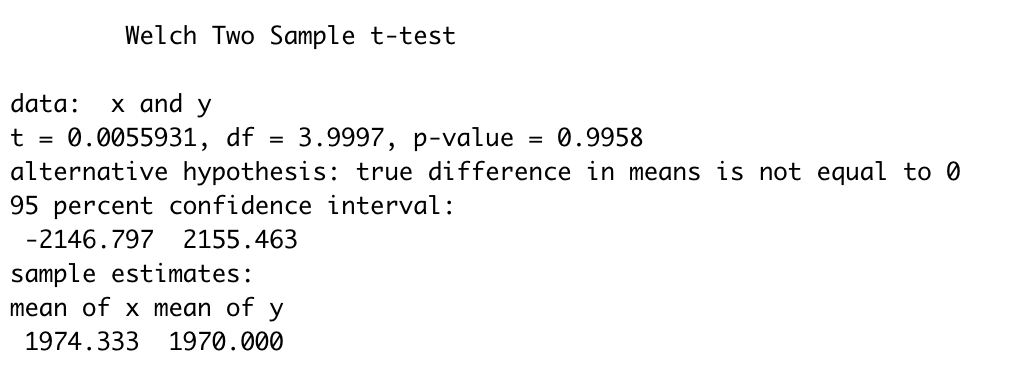
if (p-value < α) —> H0 is rejected

if (p-value >= α) —> H1 is rejected

H0 = μ1 = μ2 (means of both dataset are equal.)

H1 != μ1 != μ2 (Means are not all equal)

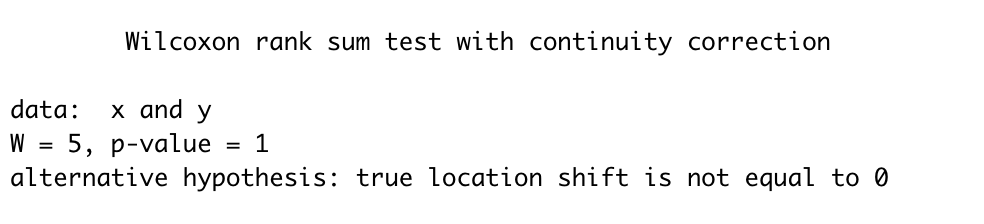
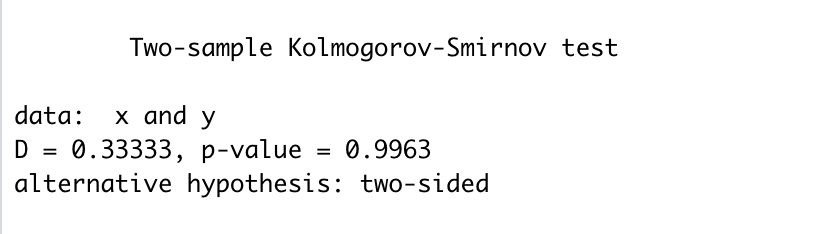
**## T-test parametric test**

****

**Since, p-value > 0.05. We reject H1, i.e., we fail to reject the null hypothesis (H0)**

**## Non-parametric tests**

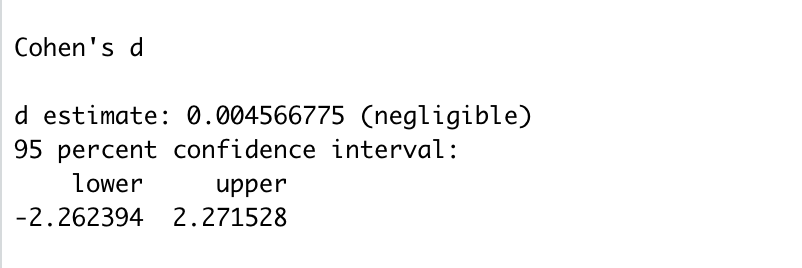
* KS-Test
* Mann-Whitney-U Test

****

## **b.** Use the effect size test to quantify the magnitude of differences.

### Solution:

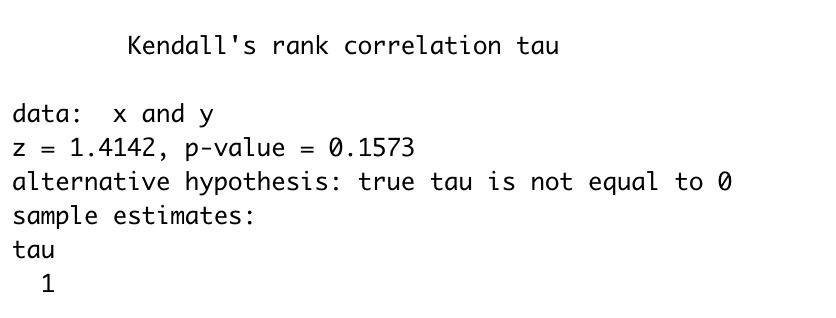
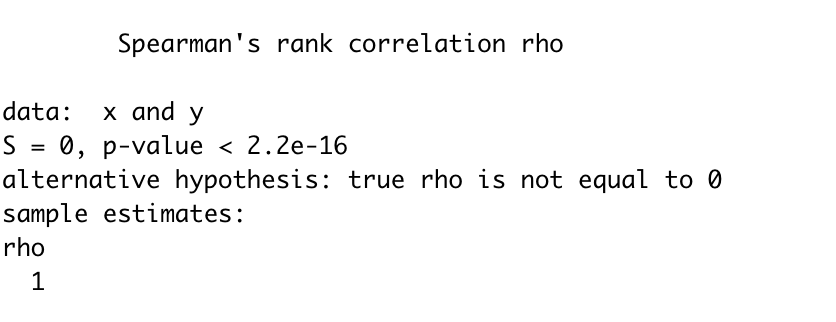
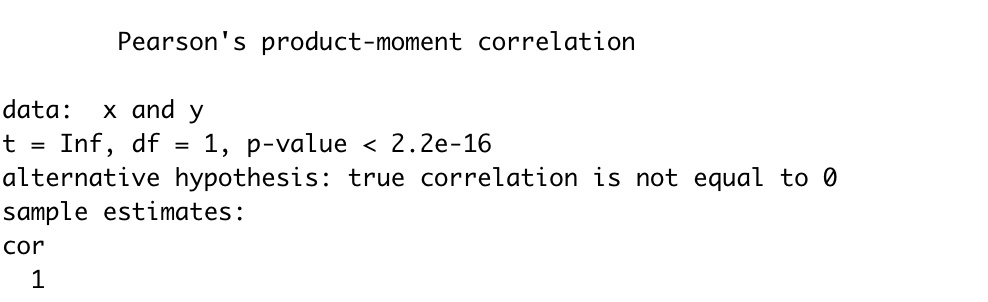
Since data belongs to a normal distribution, we will use Cohens’d test, which is parametric.

This means that the difference between two groups' means is less than **0.2 standard deviations**, the **difference is negligible**.

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

### Solution:

Correlation test is performed on Health & Fitness”, “Fashion & Wearables” on years (2018, 2019, 2020).



For all three of these, test r is 1. Positive r: means increasing one variable results in increasing the other variable.