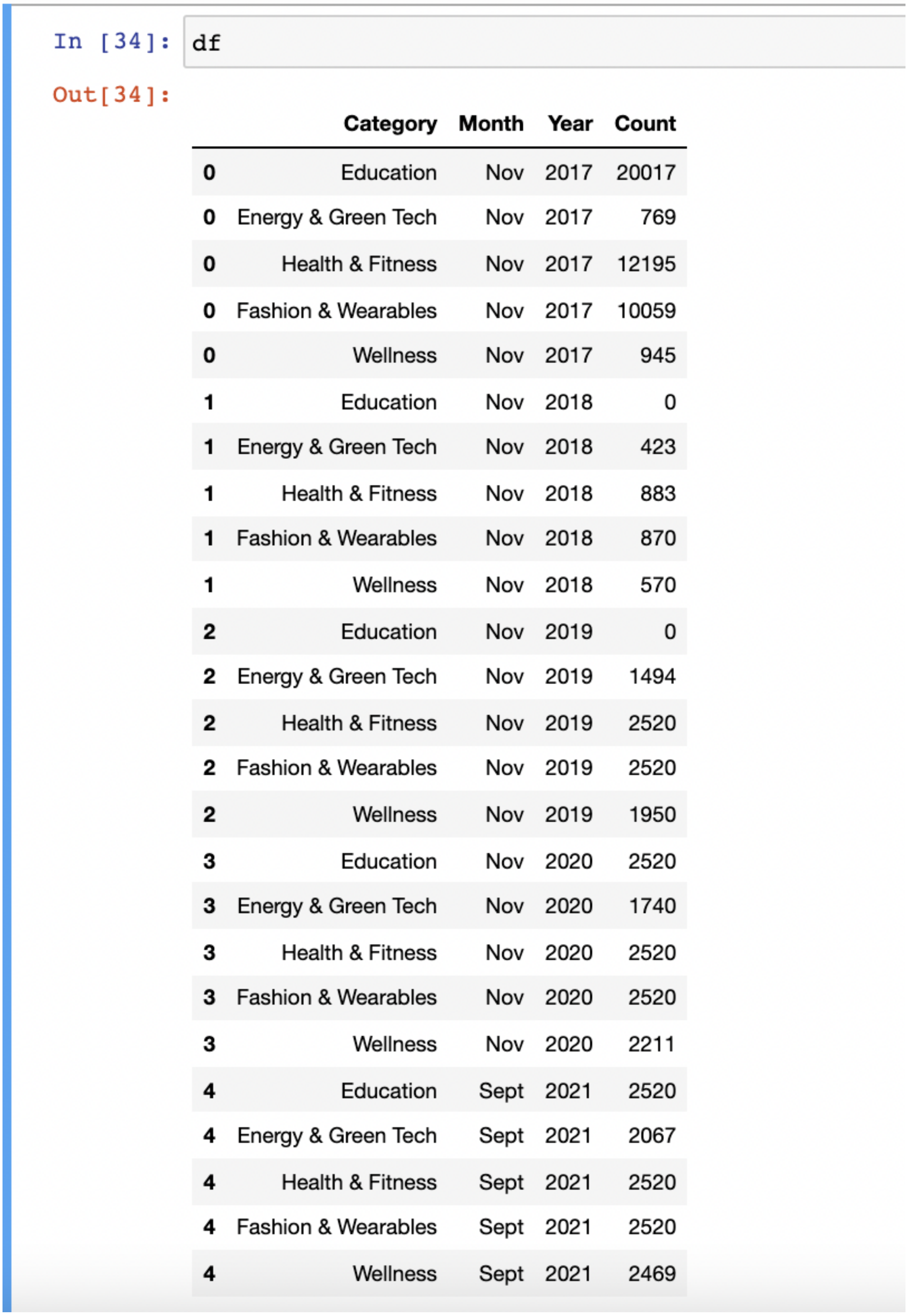
Sahil Khanna
U86445364

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:

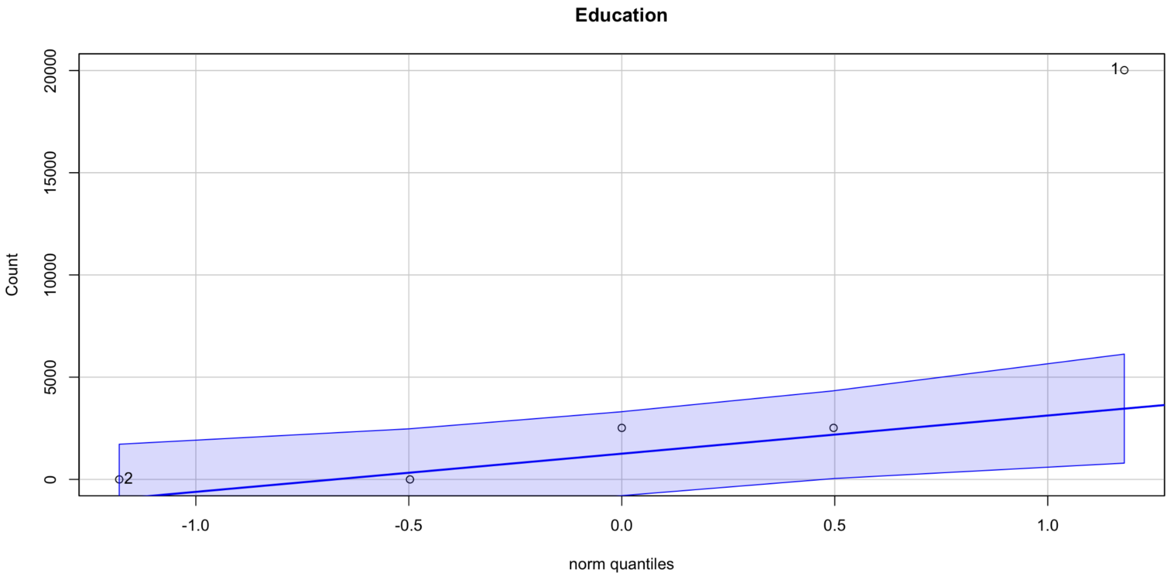


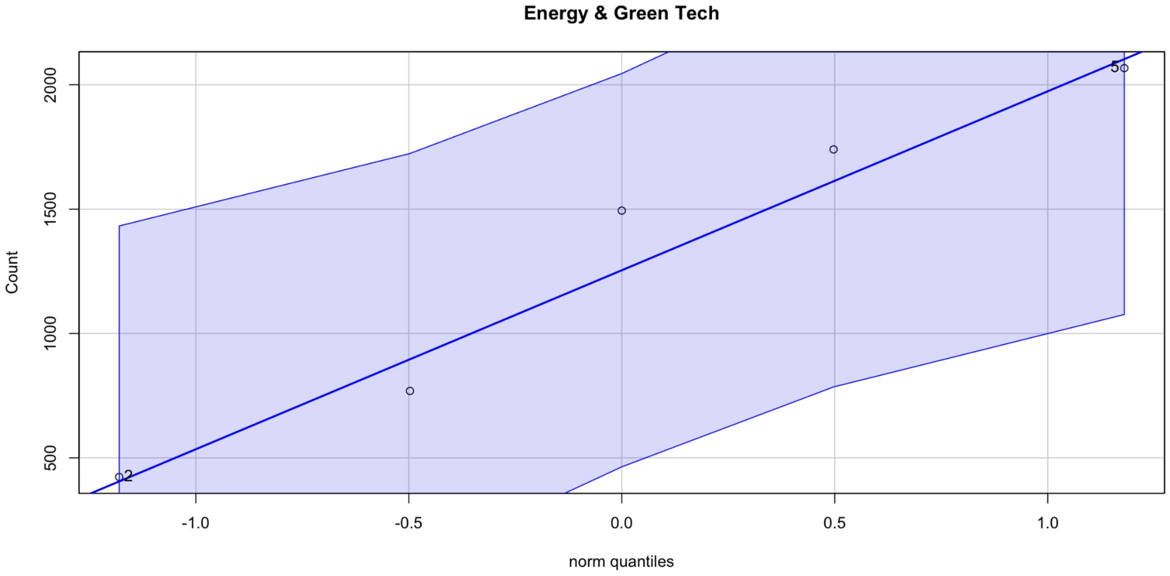
```
In [34]: df
```

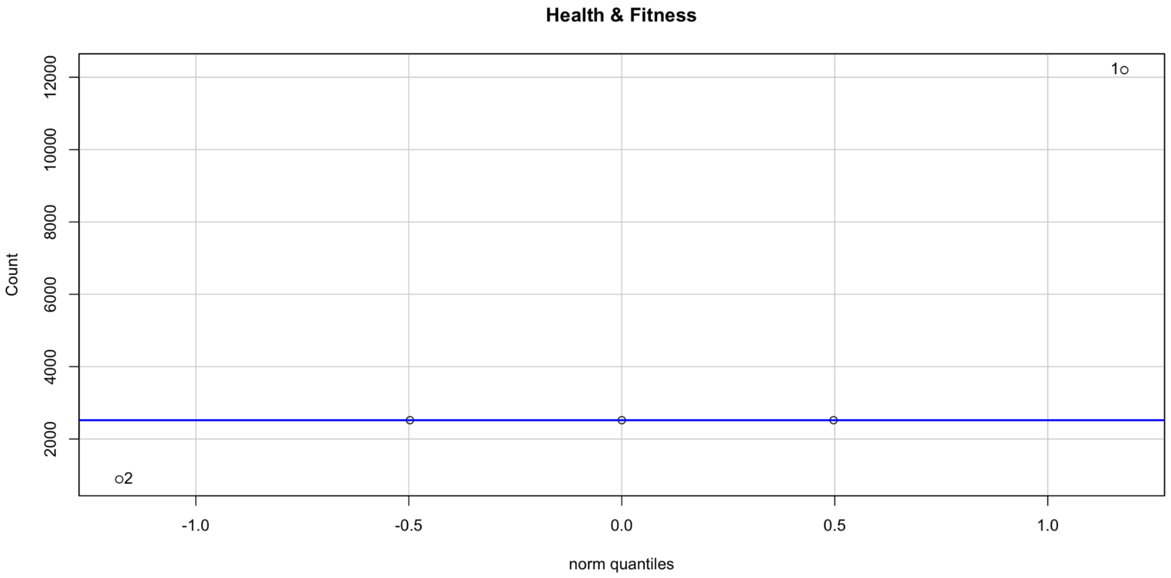
```
Out[34]:
```

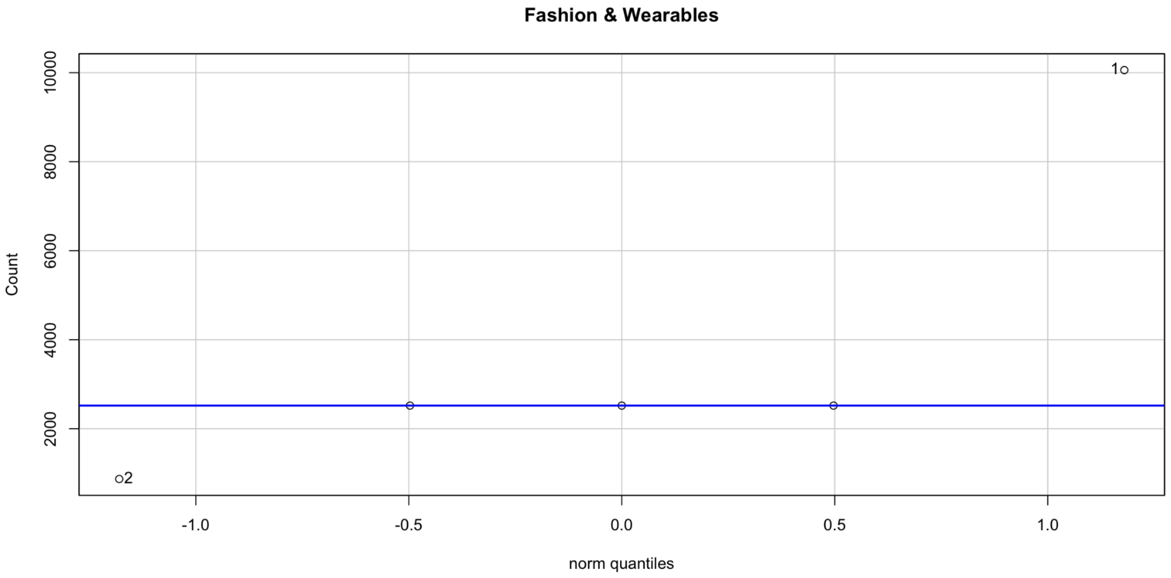
	Category	Month	Year	Count
0	Education	Nov	2017	20017
0	Energy & Green Tech	Nov	2017	769
0	Health & Fitness	Nov	2017	12195
0	Fashion & Wearables	Nov	2017	10059
0	Wellness	Nov	2017	945
1	Education	Nov	2018	0
1	Energy & Green Tech	Nov	2018	423
1	Health & Fitness	Nov	2018	883
1	Fashion & Wearables	Nov	2018	870
1	Wellness	Nov	2018	570
2	Education	Nov	2019	0
2	Energy & Green Tech	Nov	2019	1494
2	Health & Fitness	Nov	2019	2520
2	Fashion & Wearables	Nov	2019	2520
2	Wellness	Nov	2019	1950
3	Education	Nov	2020	2520
3	Energy & Green Tech	Nov	2020	1740
3	Health & Fitness	Nov	2020	2520
3	Fashion & Wearables	Nov	2020	2520
3	Wellness	Nov	2020	2211
4	Education	Sept	2021	2520
4	Energy & Green Tech	Sept	2021	2067
4	Health & Fitness	Sept	2021	2520
4	Fashion & Wearables	Sept	2021	2520
4	Wellness	Sept	2021	2469

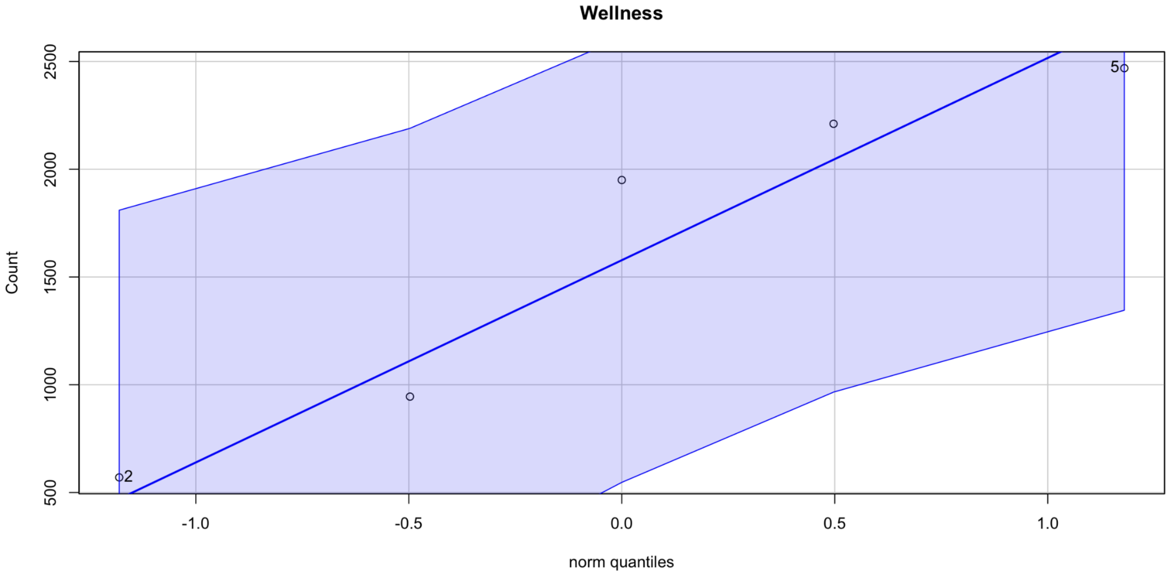
Sahil Khanna
U86445364



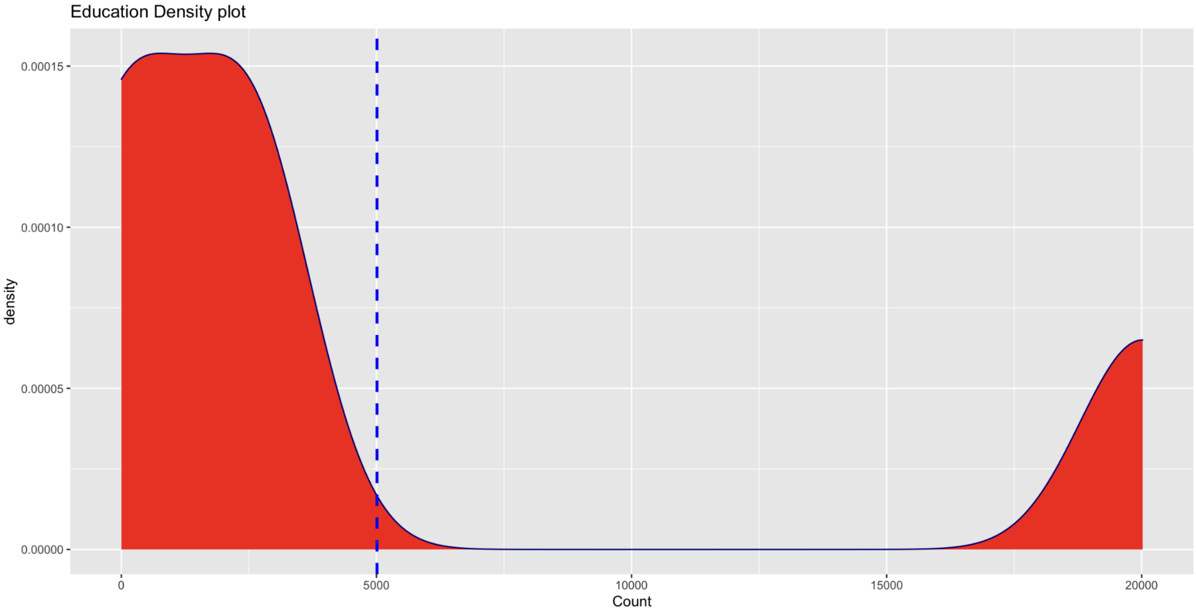



Health & Fitness
Count
12000
10000
8000
6000
4000
2000
1
2
-1.0
-0.5
0.0
0.5
1.0
norm quantiles


Fashion & Wearables
Count
10000
8000
6000
4000
2000
1
2
-1.0
-0.5
0.0
0.5
1.0
norm quantiles

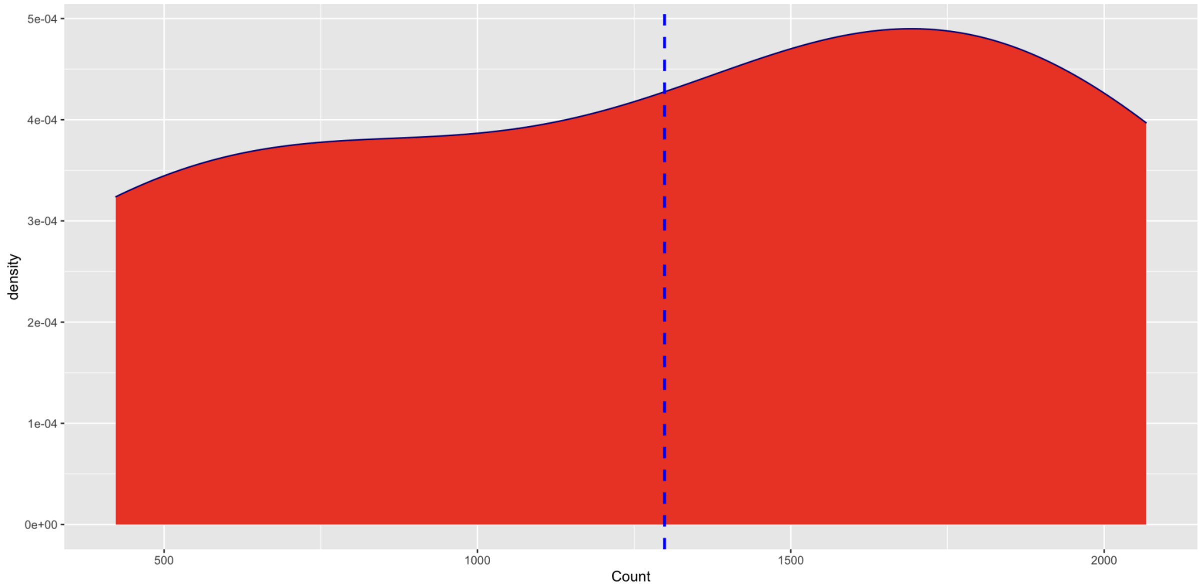


Education, Energy & Green Tech, Health & Fitness, Fashion & Wearables, and Wellness **follow Gaussian distribution.** Although, the distribution is skewed.

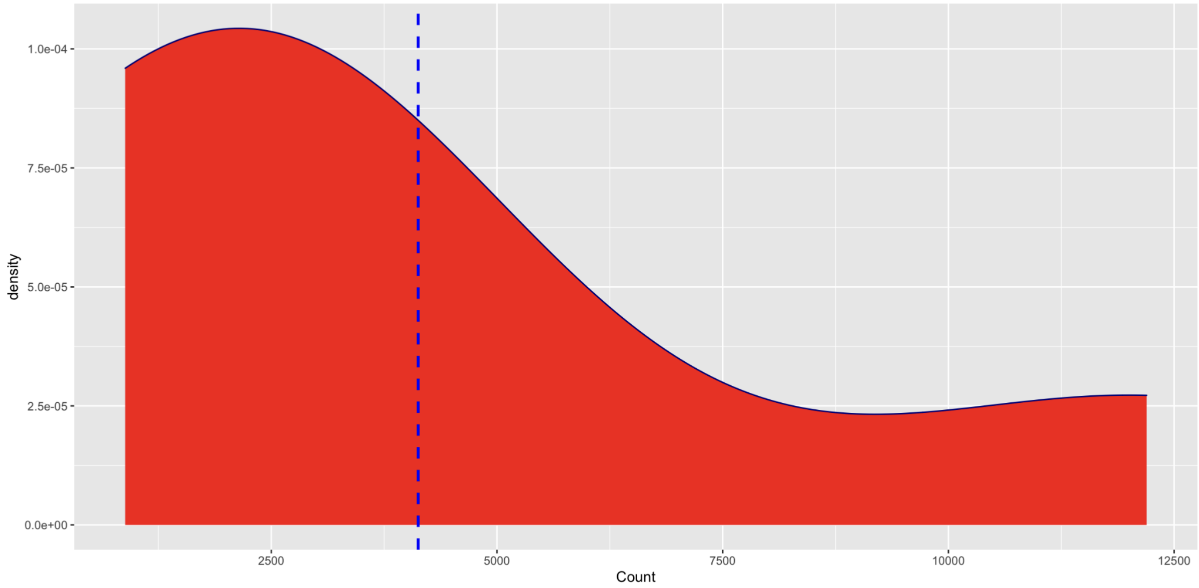


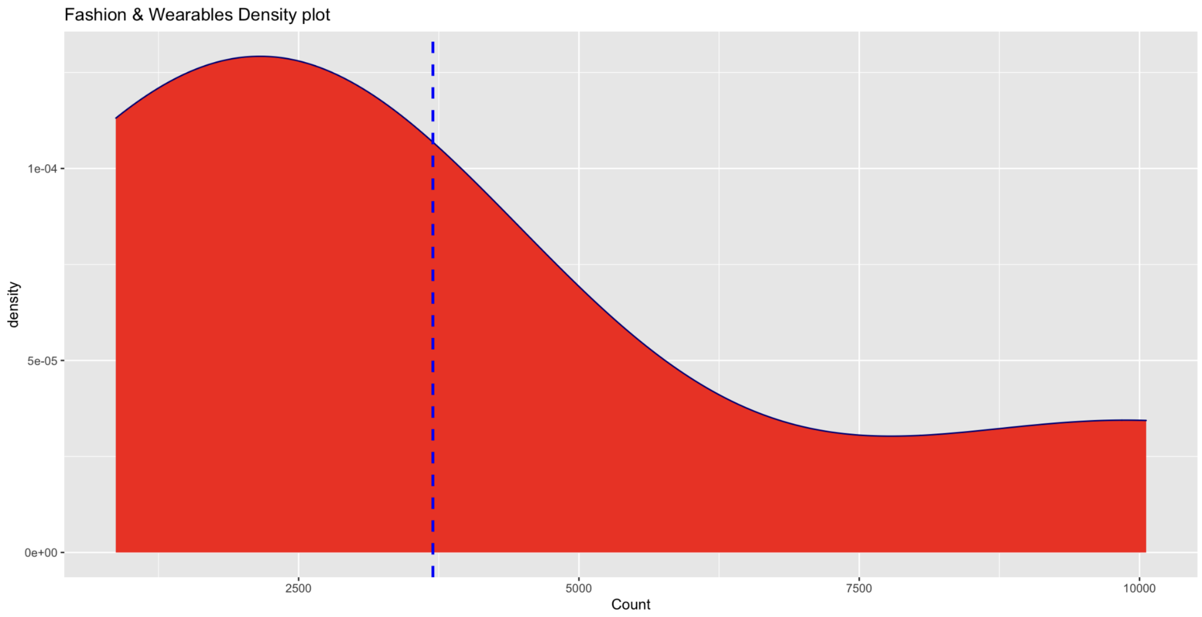
Sahil Khanna
U86445364

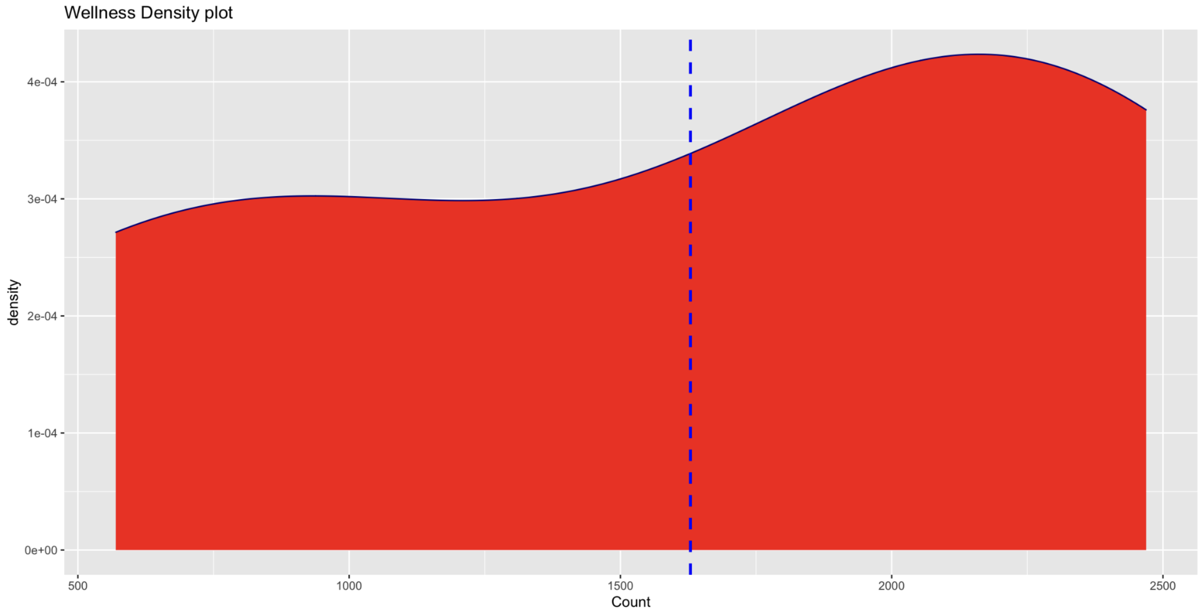
Energy & Green Tech Density plot



Health & Fitness Density plot







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:

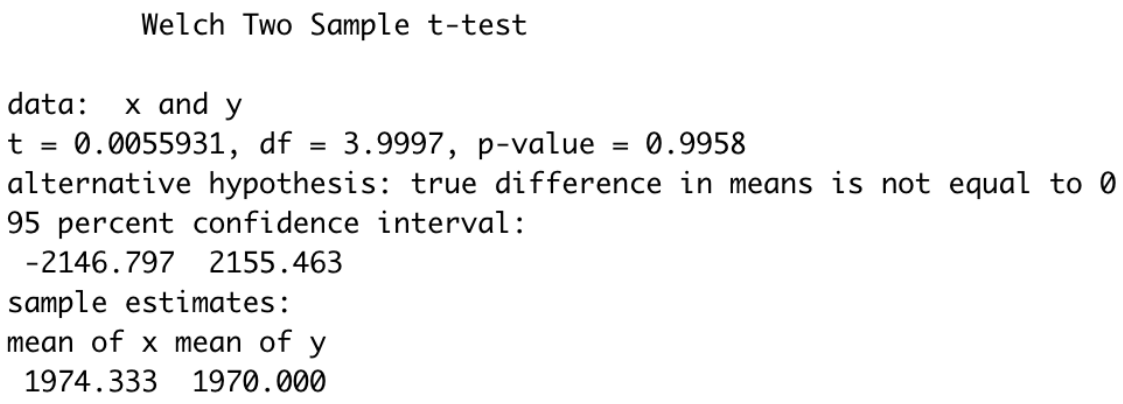
if (p-value < α) —> H0 is rejected

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

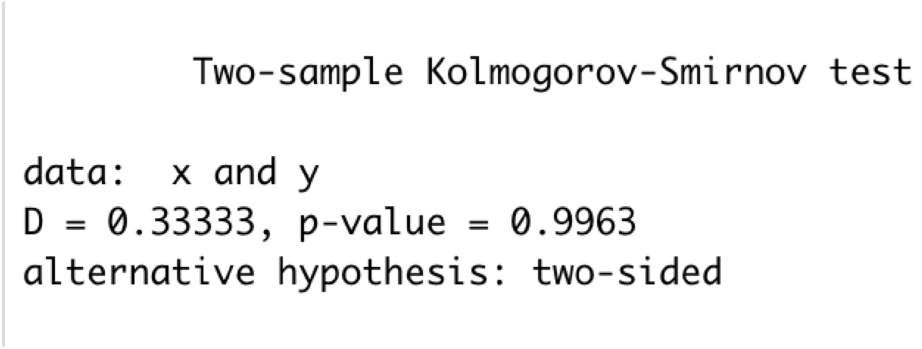


```
        Welch Two Sample t-test

data:  x and y
t = 0.0055931, df = 3.9997, p-value = 0.9958
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2146.797  2155.463
sample estimates:
mean of x mean of y
 1974.333  1970.000
```

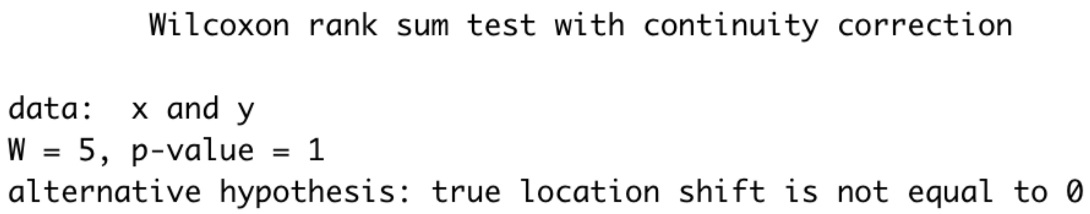
Since, p-value > 0.05. We reject H1, i.e., we fail to reject the null hypothesis (H0)

## Non-parametric tests

- KS-Test
- Mann-Whitney-U Test



```
        Two-sample Kolmogorov-Smirnov test

data:  x and y
D = 0.33333, p-value = 0.9963
alternative hypothesis: two-sided
```

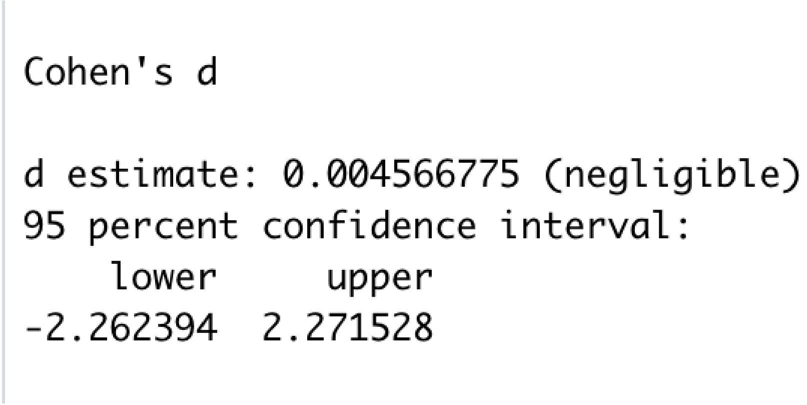


```
        Wilcoxon rank sum test with continuity correction

data:  x and y
W = 5, p-value = 1
alternative hypothesis: true location shift is not equal to 0
```

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.



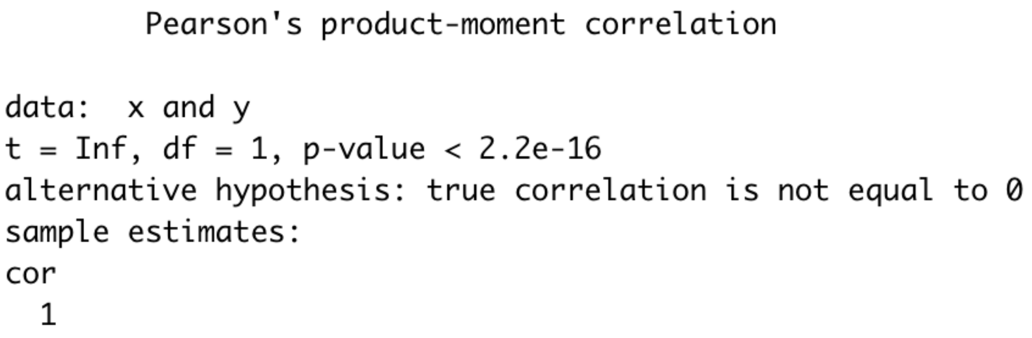
```
Cohen's d

d estimate: 0.004566775 (negligible)
95 percent confidence interval:
    lower     upper
-2.262394  2.271528
```

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

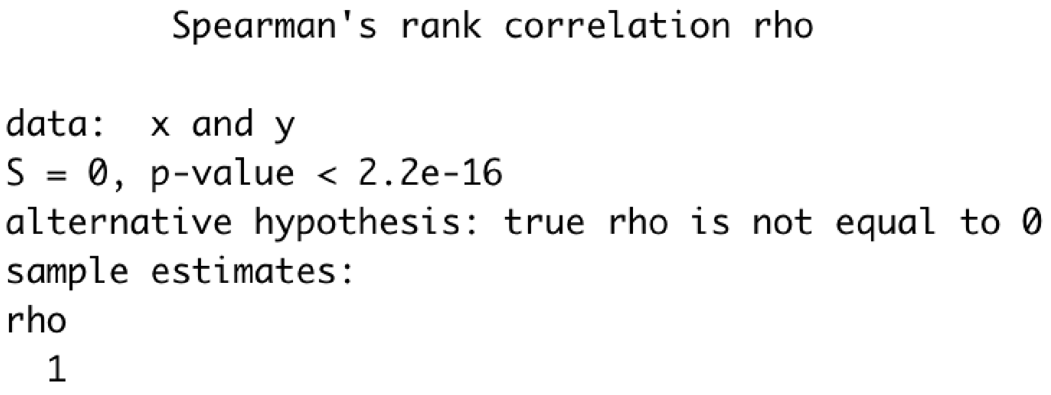
3. Use three correlation coefficient tests (Pearson, Spearman, KendallTau) and report whether the following two keywords have correlations: “Fashion & Wearables,” “Health & Fitness.”

Solution:

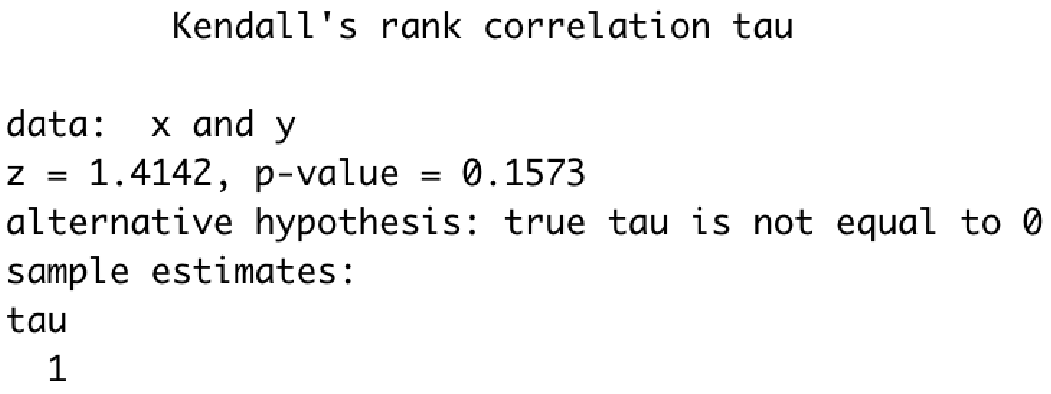
Correlation test is performed on Health & Fitness”, “Fashion & Wearables” on years (2018, 2019, 2020).



```
        Pearson's product-moment correlation

data:  x and y
t = Inf, df = 1, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
sample estimates:
cor
  1
```



```
        Spearman's rank correlation rho

data:  x and y
S = 0, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
  1
```



```
        Kendall's rank correlation tau

data:  x and y
z = 1.4142, p-value = 0.1573
alternative hypothesis: true tau is not equal to 0
sample estimates:
tau
  1
```

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