ANOVA + Tukey Test In Python

Using Statistical Testing Methods In Python To Develop An Online Advertising Strategy (With Code).

Alexander Cheng Jul 9 · 12 min read ★

Scenario

Our client is a startup clothing company that specializes in “athleisure” clothing in the US. Their marketing team wants to launch an ad campaign to increase online traffic on its website, which hopefully leads to more revenue. To best allocate time and effort for the launch of the ad campaign and maximize the audience to see their ads, they want to understand 3 things.

> ***Which keywords related to athleisure are consumers searching for most?***
>
> ***Which month are consumers searching for athleisure clothing most?***
>
> ***Which platform are consumers using most for their searches?***

. . .

Goal

> Provide recommendations to the athleisure startup company to determine the best keywords, best timing, and best platform to run their ads.

. . .

Data

To collect data for this case study, we will use Wordtracker — a paid database service for Search Engine Optimization (SEO). Search Engine Optimization is essentially just another way of saying:

> *"I want to figure out how to make my website the top result for a given search."*

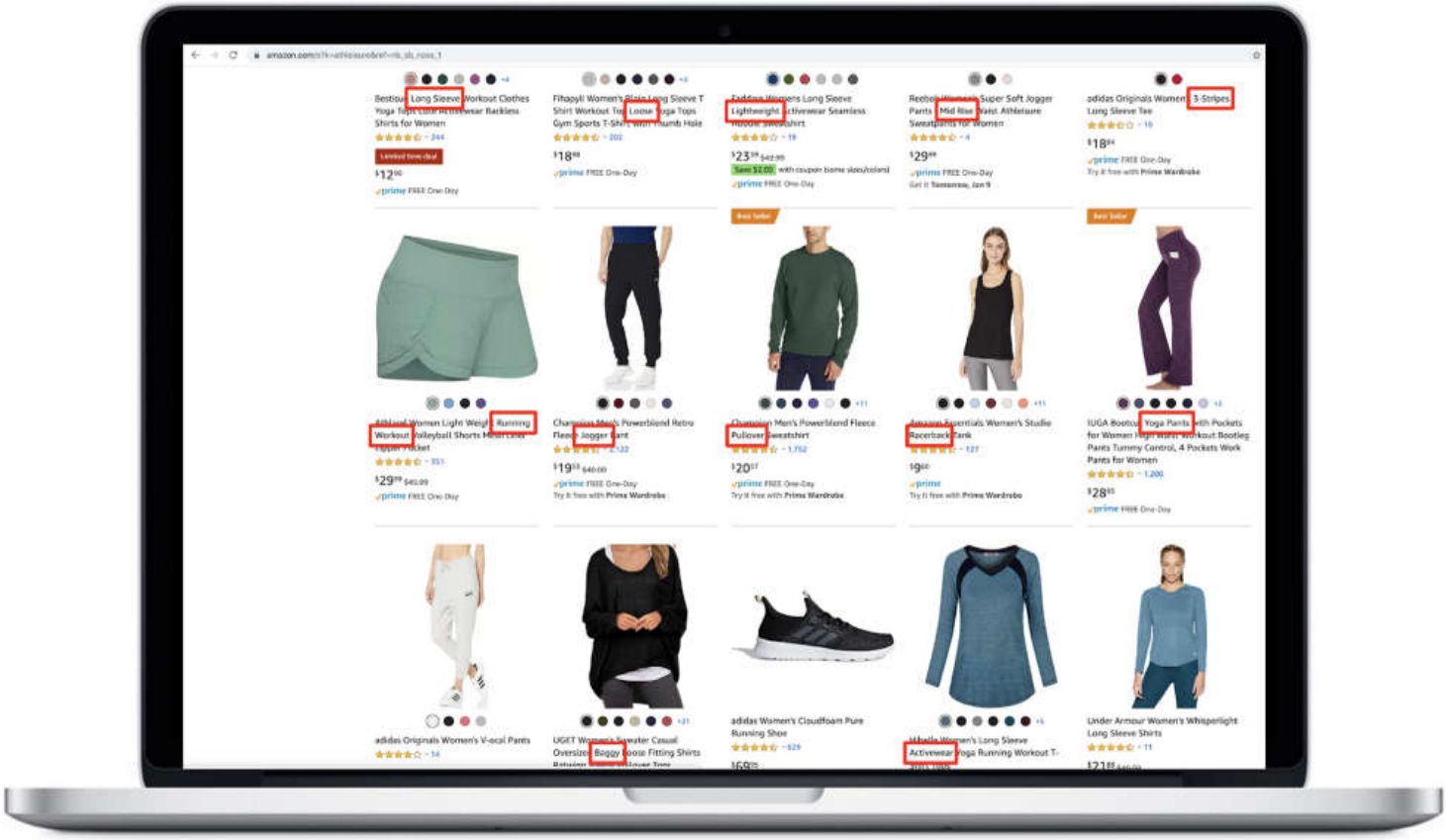
Wordtracker helps clients to get more traffic to their website or better understand what consumers are searching for. Wordtracker is similar to Google Keywords Planner service but allows access to search data on platforms beyond Google. At the time the data was extracted, Wordtracker offered a 1-year representative sample of search data from June 2018-May 2019 on Google, Youtube, Amazon, and eBay. It offered over 2 billion unique keywords from 18 million global panelists, across 106 countries. Below is a synopsis of all data in the Wordtracker database related to search volume in the United States.

Wordtracker Database Stats **(US)**

Max Volume For Any One Search	Volume Of Unique Searches	Total Volume Of Searches	Search Platform	Country
1365152	114231420	594234483	google	US
Max Volume For Any One Search	Volume Of Unique Searches	Total Volume Of Searches	Search Platform	Country
66073	9166488	60508218	youtube	US
Max Volume For Any One Search	Volume Of Unique Searches	Total Volume Of Searches	Search Platform	Country
15771	1395186	11653715	amazon	US

Source: Alex Cheng + Justin Fleury via GitHub

But how do we decide on which terms related to "athleisure" to select for our search volume queries on Wordtracker? There are many methods, but what better way than to shop for athleisure buzzwords online? We can search the term **"athleisure"** on Amazon, and find all of the most frequently occurring terms in the results.



Source: Alex Cheng + Justin Fleury via GitHub

So for our study, we can pull data from Wordtracker with the following constraints:

- **70+ terms related to athleisure, using top Amazon keyword results when searching for "athleisure".**
- **Search volumes only in the United States.**
- **Search volume data pulled from Google, YouTube, and Amazon.**

According to Search Engine Journal — Google, YouTube, and Amazon are the three most popular search engines worldwide. According to Bluelist in 2019, there are about 2 trillion Google searches per year. Wordtracker provides nearly 2 billion Google searches within a 1-year timeframe. We assume that Wordtracker provides a representative sample of roughly 1/1000th of the entire Google search database worldwide.

The cleaned dataset can be downloaded **here** as a .csv file. For the full code on how the data was called from Wordtracker's API — refer to this Jupyter Notebook **here**. For

additional information on how the data was cleaned, refer to the data cleaning Jupyter Notebook **here**.

. . .

Exploratory Data Analysis (EDA)

Tableau Dashboard

Once the data has been extracted and cleaned, we can perform **exploratory data analysis**, or EDA for short. Creating a Tableau dashboard is a great way to get a comprehensive understanding of the various relationships between the dimensions and measures in our data, all in one place. A static preview of this dashboard is displayed below. The fully interactive Tableau dashboard can be explored **here**.

Source: Alex Cheng + Justin Fleury via GitHub

Word Clouds

A collection of word clouds provide a non-technical, graphic sense of scale to understand the most searched athleisure terms in each search engine. The larger the keyword appears in the word cloud, then the more that keyword was searched compared to the others.

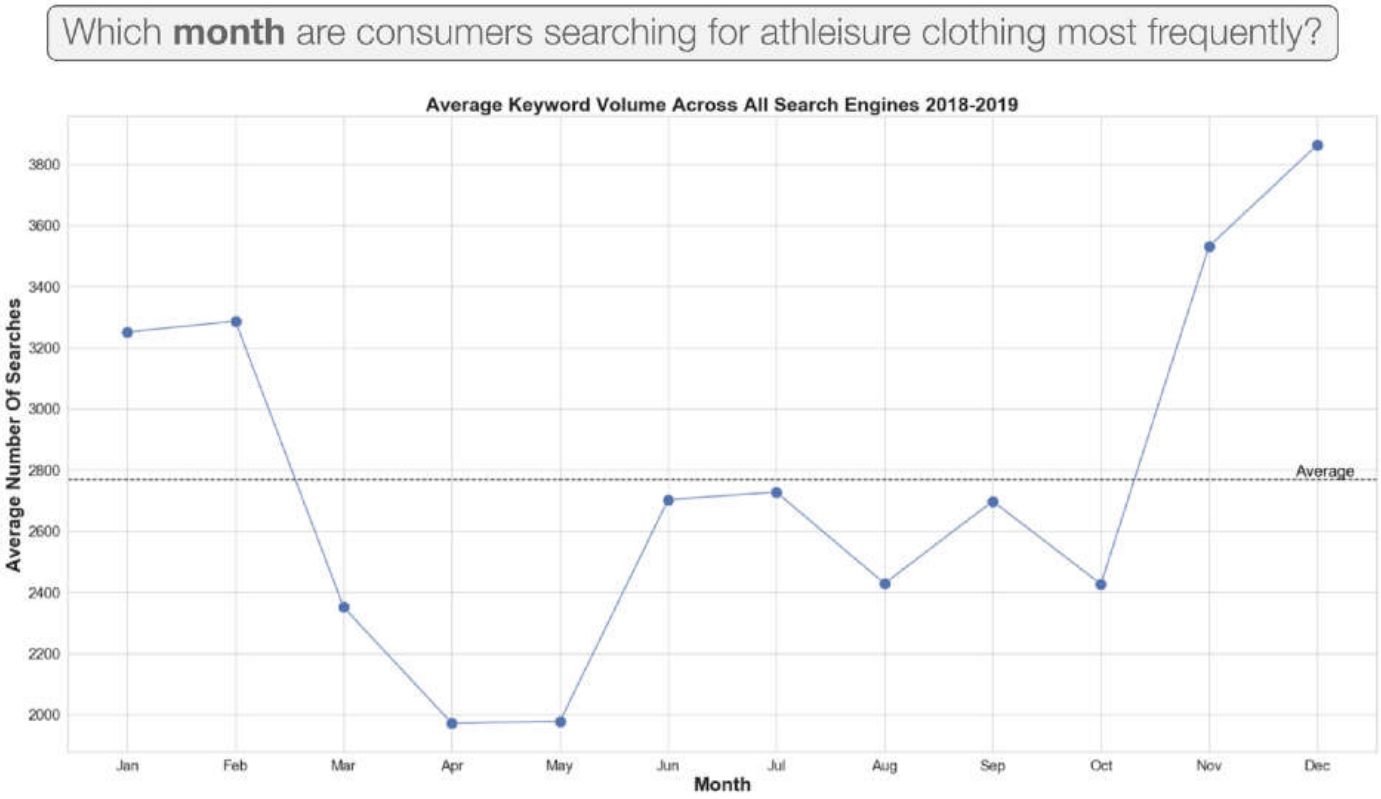
Google

YouTube

amazon

Average Keyword Search Volume Per Month — Line Plot

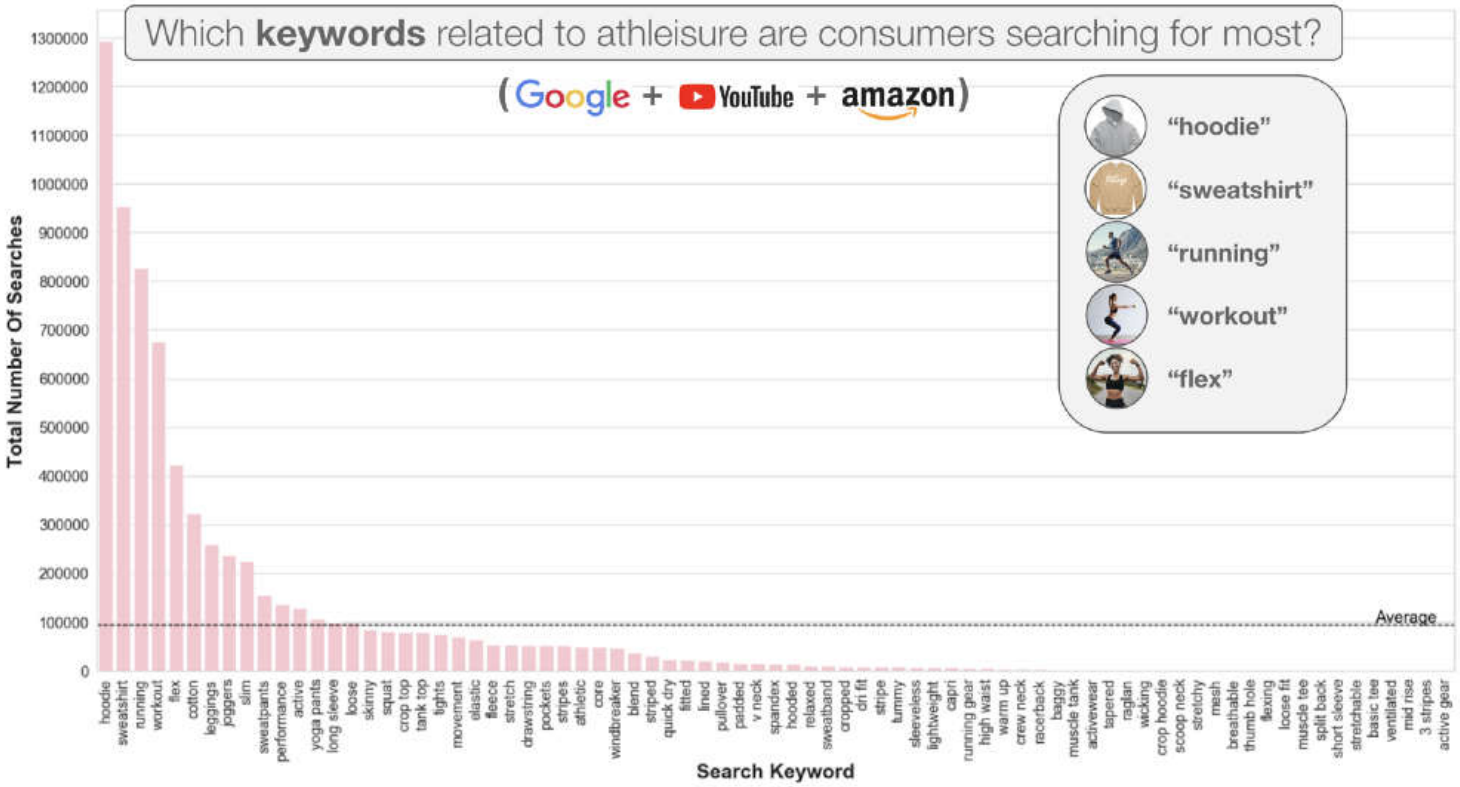
This line plot shows how often athleisure keywords are searched on average over each month of the year. There seems to be a higher number of searches in the colder months than in the warmer months, with the peak being in December.



Source: Alex Cheng + Justin Fleury via GitHub

Total Search Volume Per Keyword — Bar Plot

When we aggregated search volume for all of our athleisure-related keywords across all three search engines (Google + YouTube + Amazon) we noticed that the keyword “hoodie” was the most searched term, with 1,300,000 searches. This is followed by “sweatshirt”, “running”, “workout”, and “flex”, all with several hundred thousand searches. Across all 77 keywords that were considered, the average number of searches is about 100,000.

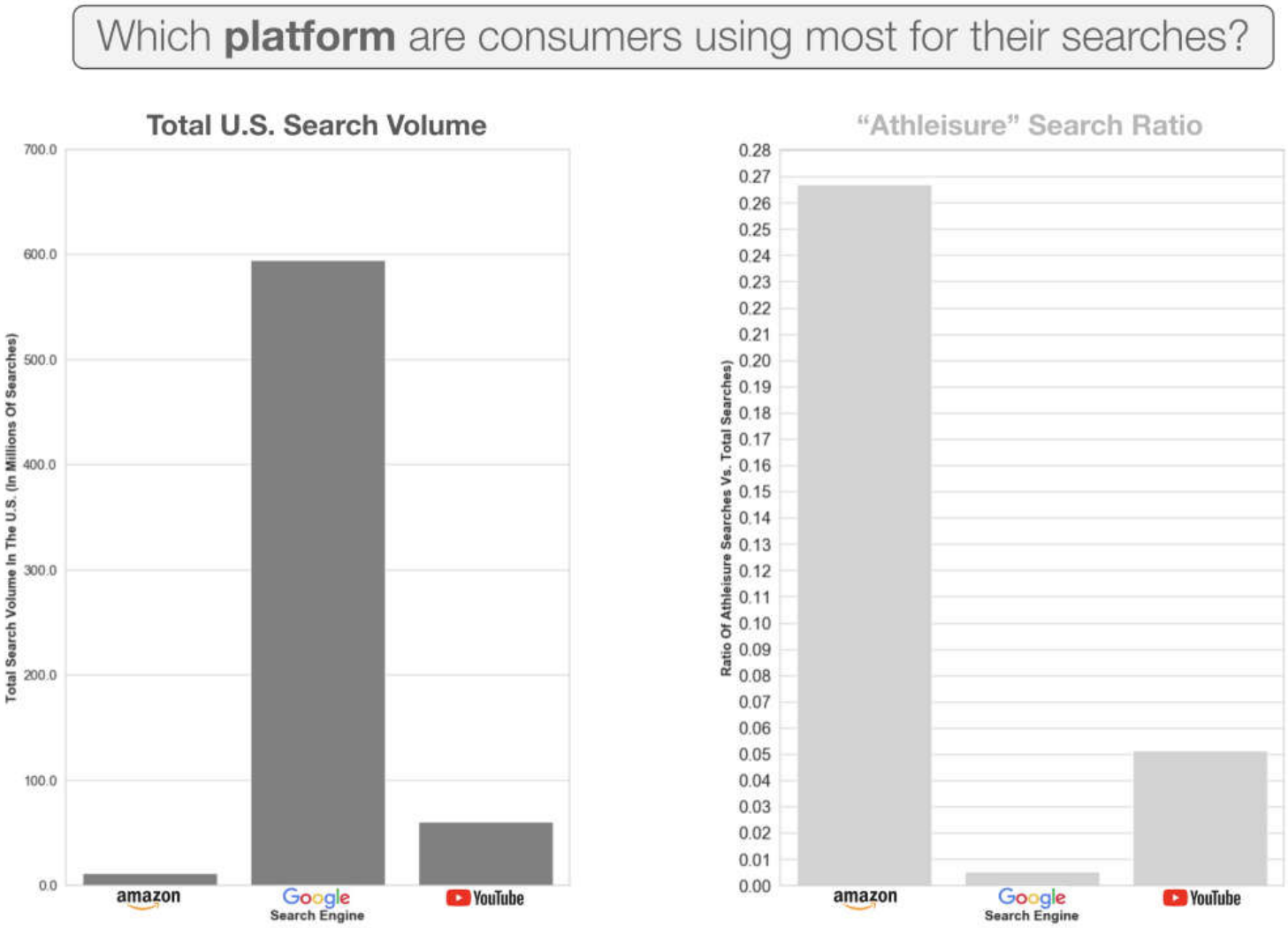


Source: Alex Cheng + Justin Fleury via GitHub

Total Volume Per Engine & “Athleisure” Search Ratio — Bar Plot

In observing metrics on each search engine, we found that Amazon and YouTube have a very low total search volume compared to Google. But even though Amazon has the lowest overall search volume, it has by far the highest ratio for athleisure-related terms. Over 25% of all searches on Amazon are related to athleisure keywords! In comparison, 5% of searches on YouTube are related to athleisure keywords. And less than 1% of searches on Google are related to athleisure keywords. These findings suggest that Amazon and Youtube may be better suited to run ads rather than Google since people are clearly searching athleisure keywords more often on those platforms.

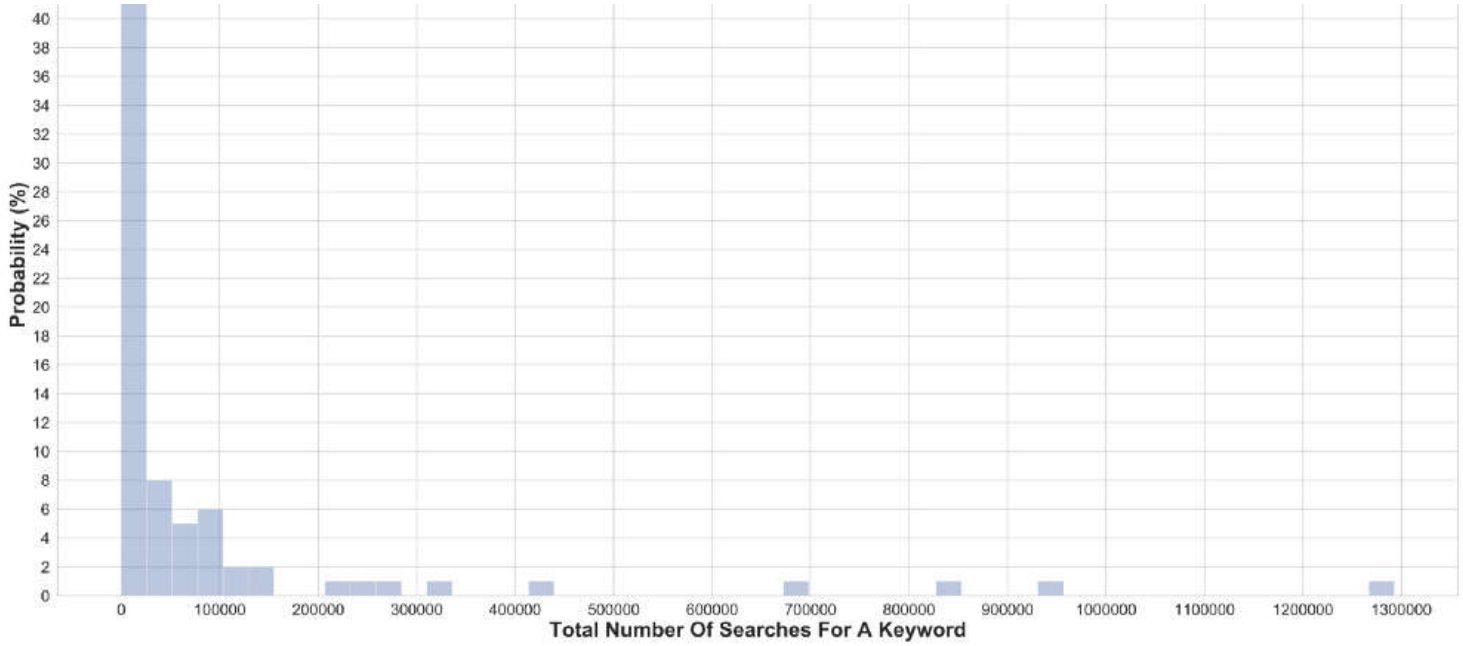
In the barplot on the left, we can see that Google has nearly 600,000,000 total searches, while Amazon and YouTube combined are less than 100,000,000. However, in the barplot on the right, we can see the search ratio for athleisure-related terms, compared to all searches on each search engine.



Source: Alex Cheng + Justin Fleury via GitHub

Probability Density Function (PDF)

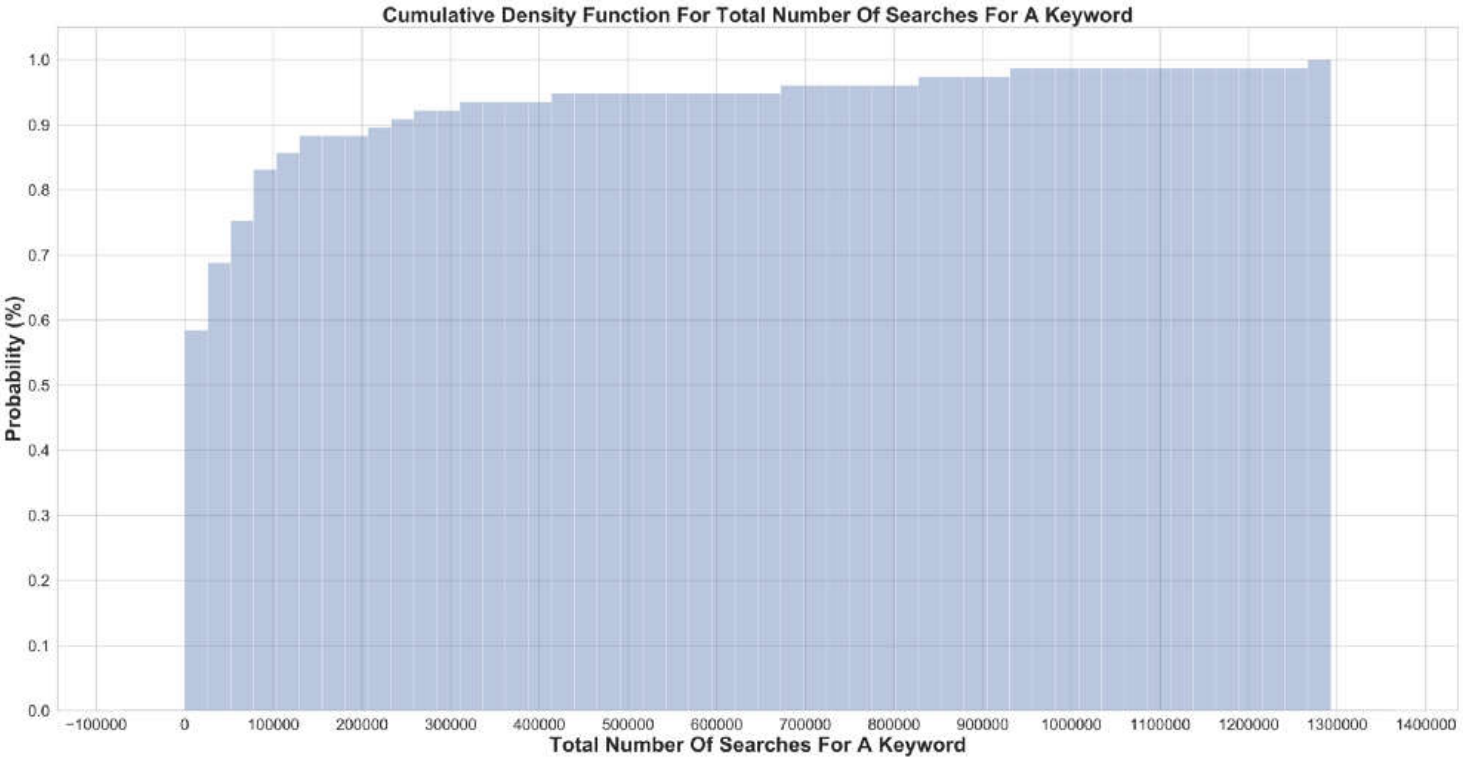
The Probability Density Function plot below shows that a pretty good chunk of our keywords have a small search volume — close to zero compared to the rest of our keywords. Clearly, some keywords are searched much higher than others, and highly searched keywords are rare.



Source: Alex Cheng + Justin Fleury via GitHub

Cumulative Density Function (CDF)

The Cumulative Density Function plot below shows that 90% of our keywords have a search volume that is less than 200,000. Once again, this obviates that a large number of our keywords have a low search volume and that there are only a handful of highly searched keywords. This CDF plot seems to be logarithmic in nature.



Source: Alex Cheng + Justin Fleury via GitHub

. . .

Statistical Testing

Before we get into the statistical testing of our variables, below is a brief overview of hypothesis testing and the concepts that we will use to determine statistical significance.

Hypothesis Testing

Put very simply — the **Null Hypothesis (H0)** is a hypothesis that claims that there is no statistically significant relationship between variables, while the **Alternative Hypothesis (HA)** claims that there is a statistically significant relationship between variables.

Alpha Value

The **alpha value** is the probability of rejecting the null hypothesis when the null hypothesis is true. This is also known as a false-positive, or Type I error. We choose the alpha value in every hypothesis test.

A **higher** alpha value means we are okay with a higher probability of error, while a **lower** alpha value means that we are okay with a very low probability of error. Higher alpha values might be okay when there is no big consequence if there is an error, while low alpha values should be used when there are dire consequences if there is an error (like medical diagnoses, criminal sentencing, or any endangering of a life).

Alpha values typically range between less than 0.01, and 0.1. For our case study, we can use an alpha value is not too lenient, nor too strict. So, We will use an alpha value of 0.05.

P-Value

The **p-value** is the "evidence" against a null hypothesis. The smaller the p-value, the stronger the evidence that we can reject the null hypothesis. We compare the p-value to the alpha value that we set for our statistical test.

- If the p-value is < 0.05, then we reject the null hypothesis.
- If the p-value is $>= 0.05$, then we fail to reject the null hypothesis.

ANOVA

One-Way Analysis Of Variance or **ANOVA** is our statistical test of choice since we are dealing with multiple groups. We will also use **Two-Way ANOVA** to determine combinations of factors that may be statistically significant.

Tukey Test

The problem with ANOVA is that it only compares the means between groups and determines if any of those means are statistically significantly different from each other. **In short: ANOVA tells us if our results or significant or not, but does not tell us where the results are significant.**

But, the interpretability of statistical significance is crucial to figure out in order to guide our athleisure advertising strategy. We have to be able to explain which keywords perform best, which search engine is best, or which month is best to run ads!

So a **Tukey Test** allows us to interpret the statistical significance of our ANOVA test and find out which specific groups' means (compared with each other) are different. **So, after performing each round of ANOVA, we should use a Tukey Test to find out where the statistical significance is occurring in our data.**

One-Way ANOVA + Tukey Test

Hypothesis Test 1: Keywords

Question: Are there any differences between athleisure-related keywords when considering search volume?

> ***Null Hypothesis (H0)*** *— All athleisure-related keywords are equal in terms of average search volume.*
>
> ***Alternative Hypothesis (HA)*** *— Some athleisure-related keywords have greater average search volumes than others.*

One-Way ANOVA Result:

P-Value = 1.3293563590514185e-119 < 0.05 (This is nearly zero.)

We reject the null hypothesis that mean search volume is equal across all athleisure-related keywords. Keyword on its own, does indeed constitute a difference in average search volume for athleisure-related items.

Tukey Test Result:

The top 5 terms that are the “most” statistically different than the rest are:

- “hoodie”
- “running”
- “sweatshirt”
- “workout”
- “flex”

	reject1	reject2	total_sum
hoodie	50.0	25.0	75.0
running	28.0	46.0	74.0
sweatshirt	11.0	62.0	73.0
workout	1.0	70.0	71.0
flex	33.0	13.0	46.0
3 stripes	5.0	0.0	5.0
spandex	2.0	3.0	5.0
muscle tank	3.0	2.0	5.0
muscle tee	3.0	2.0	5.0
pullover	3.0	2.0	5.0

Hypothesis Test 2: Months

Question: Are there any differences between months when considering search volume?

> ***Null Hypothesis (H0)*** *— People will be equally likely to search for activewear-related terms in any given month.*
>
> ***Alternative Hypothesis (HA)*** *— People will be more likely to search for activewear-related terms depending on the month.*

One-Way ANOVA Result:

- P-Value = 0.8831258135517717 > 0.05

- We fail to reject the null hypothesis that mean search volume is equal across all months.
- The month on its own does not constitute a difference in search volumes for athleisure-related items.

Tukey Test Result:

- No need to run Tukey multiple comparisons test since we failed to reject the null hypothesis here.

Hypothesis Test 3: Search Engine

Are there any differences among search engines when considering search volumes?

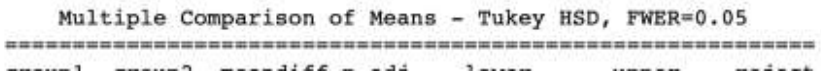
> ***Null Hypothesis (H0)*** *— There will be an equal search volume for activewear-related terms on any platform.*
>
> ***Alternative Hypothesis (HA)*** *— There will be a greater search volume for activewear-related terms on one particular platform.*

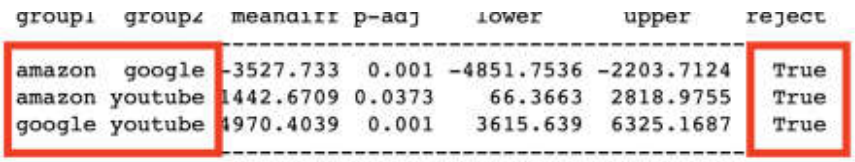
One-Way ANOVA Result:

- P-Value = 7.19196465389629e-18 < 0.05 (This is nearly zero.)
- We reject the null hypothesis that mean search volume is equal across all search engines.
- Search engine on its own, does indeed constitute a difference in average search volume for athleisure-related items.

Tukey Test Result:

- In all cases, reject the null hypothesis that search engine 1 is equal to search engine 2 in terms of average search volume.
- Search volumes are unique to each platform.





group1	group2	meandiff	p-adj	lower	upper	reject
amazon	google	-3527.733	0.001	-4851.7536	-2203.7124	True
amazon	youtube	1442.6709	0.0373	66.3663	2818.9755	True
google	youtube	4970.4039	0.001	3615.639	6325.1687	True

Code

Below are code snippet examples for how to perform the One-Way ANOVA and Tukey Test in Python.

For the One-Way ANOVA, we are using the SciPy library (note: this can also be done using the Statsmodels library). We coerce the data to a dictionary and feed the keys to the **scipy.stats.f_oneway()** function, which returns the F-statistic and p-value (which is what we're after).

```
1   import pandas as pd
2   from scipy import stats
3
4   athleisure_df = pd.read_csv('athleisure.csv')
5   athleisure_df.drop(['Unnamed: 0'], axis =1, inplace = True)
6
7   keys = list(athleisure_df.engine.unique())
8
9   values = []
10  for engine in list(athleisure_df.engine.unique()):
11      values.append(list(athleisure_df.loc[athleisure_df['engine'] == engine, 'volume']))
12
13  data = dict(zip(keys, values))
14
15  # stats f_oneway functions takes the groups as input and returns F and P-value
16  fvalue, pvalue = stats.f_oneway(data['google'],
17                                  data['youtube'],
18                                  data['amazon'])
19
20  print(f"Results of ANOVA test:\n The F-statistic is: {fvalue}\n The p-value is: {pvalue}")
21
22  # Results of ANOVA test:
23  # The F-statistic is: 40.08443136373594
24  # The p-value is: 7.19196465389629e-18
```

Search Engine ANOVA.txt hosted with ♡ by GitHub view raw

For the Tukey test, we assign a variable to the **pairwise_tukeyhsd()** function, where we provide our response variable (search volume), the group we are testing (search engine, in this case), and our alpha value (0.05). Then, we simply print the result.

```
from statsmodels.stats.multicomp import pairwise_tukeyhsd

# perform multiple pairwise comparison (Tukey HSD)
m_comp = pairwise_tukeyhsd(endog=athleisure_df['volume'], groups=athleisure_df['engine'], alpha=0
print(m_comp)
```

Search Engine Tukey Test.txt hosted with ♡ by GitHub view raw

. . .

Multiple ANOVA + Tukey Test

Two-factor ANOVA between all three factors will help us see if any two combinations of these factors are statistically significant. We essentially wanted to answer this question:

> *"Can we determine which specific 2-factor combinations of keyword/month/search engine generate the highest search volume?"*

Combo 1: Keyword + Engine

> ***Null Hypothesis (H0)*** *— All keyword/engine combinations are equal in terms of mean search volume.*
>
> ***Alternative Hypothesis (HA)*** *— Some keyword/engine combinations have greater mean search volume.*

Two-Way ANOVA Result:

- P-Value = 1.008919e-151 < 0.05 (This is nearly zero.)
- Reject the null hypothesis that the mean search volume is equal among all Keyword/Engine combinations. Tukey Test needed.

Tukey Test Result:

- There were 10 Keyword/Engine combinations that were significantly different in search volume.

	reject1	reject2	total_sum
running / youtube	85.0	128.0	213.0
hoodie / youtube	142.0	71.0	213.0
sweatshirt / youtube	35.0	177.0	212.0
workout / amazon	4.0	206.0	210.0
flex / youtube	148.0	60.0	208.0
hoodie / amazon	141.0	67.0	208.0
leggings / amazon	127.0	67.0	194.0
workout / youtube	2.0	191.0	193.0
joggers / amazon	126.0	63.0	189.0
cotton / youtube	155.0	28.0	183.0
3 stripes / google	10.0	0.0	10.0
skinny / amazon	3.0	7.0	10.0
scoop neck / youtube	3.0	7.0	10.0
short sleeve / google	3.0	7.0	10.0
short sleeve / youtube	3.0	7.0	10.0
sleeveless / amazon	3.0	7.0	10.0
skinny / google	3.0	7.0	10.0
scoop neck / amazon	3.0	7.0	10.0
sleeveless / google	3.0	7.0	10.0
slim / google	3.0	7.0	10.0

Combo 2: Keyword + Month

> ***Null Hypothesis (H0)***— *All keyword/month combinations are equal in terms of mean search volume.*
>
> ***Alternative Hypothesis (HA)*** — *Some keyword/month combinations have greater mean search volume.*

Two-Way ANOVA Result:

- P-Value = 7.896266e-01 > 0.05
- Fail to reject the null hypothesis that the mean search volume is equal among Keyword/Month combinations. No Tukey Test.

Combo 3: Engine + Month

> ***Null Hypothesis (H0)*** *— All engine/month combinations are equal in terms of mean search volume.*
>
> ***Alternative Hypothesis (HA)*** *— Some engine/month combinations have greater mean search volume.*

Two-Way ANOVA Result:

- P-Value = 7.789742e-01 > 0.05
- Fail to reject the null hypothesis that the mean search volume is equal among Engine/Month combinations. No Tukey Test.

	sum_sq	df	F	PR(>F)
C(keyword)	1.021546e+11	76.0	19.630897	9.219712e-149
C(engine)	1.072592e+10	2.0	78.324956	4.462158e-33
C(month_abbr)	8.591816e+08	11.0	1.140743	3.249134e-01
C(keyword):C(engine)	1.156474e+11	152.0	11.111892	1.008919e-151
C(keyword):C(month_abbr)	5.446891e+10	836.0	0.951564	7.896266e-01
C(engine):C(month_abbr)	1.143128e+09	22.0	0.758871	7.789742e-01
Residual	1.024321e+11	1496.0	NaN	NaN

Code

Below are code snippet examples for how to perform the Two-Way ANOVA and Tukey Test in Python.

For the Two-Way ANOVA, we provide a string "formula" to define our groups as required by the **statsmodels.formula.api.ols()** function, assign this fitted ols() function to a variable, then feed that variable into the **sm.stats.anova_lm()** function. In the output, we look to the PR(>F) column which provides our p-values (which is what we're after).

For the Tukey Test after Two-Way ANOVA, we assign a variable to the **pairwise_tukeyhsd()** function, where we provide our response variable (search volume), the groups we are testing (combo of keyword + search engine, in this case), and our alpha value (0.05). Then, we coerce the output to a dataframe for easier analysis and filtering, and add a "total_sum" column to add up all of the (True) null rejections for each observation in each group. There are a very large number of combinations, so we only show the top 20 results.

. . .

Recommendations

Keywords

There were 5 keywords that outperform any other athleisure-related keyword tested across all platforms and months. We might recommend that the ad campaign use these 5 buzzwords:

- **“hoodie”**
- **“running”**
- **“sweatshirt”**
- **“workout”**
- **“flex”**

Engine

Ads should not be launched on Google, because it has the lowest search volume for athleisure-related keywords. If the search volume is most important, then we would recommend YouTube. If market share is most important, then we would recommend Amazon.

Month

The month by itself is not statistically significant enough of a factor to provide a confident recommendation. The month should only be considered as a factor when combined with a particular platform and set of keywords.

Keyword/Engine

Below are the top 10 keyword/engine combinations that we would recommend to our athleisure clothing startup company to help them guide their advertising efforts online. Note again that Google is not recommended as a platform to run athleisure advertisements.

Improvements + Future Work

Improvements

- We might ensure all word types being tested as athleisure keywords are similar. For example, use all nouns, or all adjectives, etc…
- We might limit results to explicitly clothing-related searches. For example, consider pairing an “athleisure” related adjective to an article of clothing — ex: “breathable hoodie”, “ventilated shorts”, “striped joggers”, etc…
- We might ensure all platforms being compared are the same in the service that they provide for better accuracy. For example, compare Google to Bing, or YouTube and Vimeo, or Amazon and eBay.

Future Work

- We could explore search volume demographic statistics of each engine, for example: age, gender, or income.
- We could investigate conversion rates — meaning who actually buys the product after viewing the ad.

- We could consider the costs of running an ad on a particular platform. For example, what is the cost of an ad on Google versus Amazon?

. . .

Thanks!

Thanks for reading this blog! I hope it was useful and brought some clarity to what statistical testing is capable of in a contemporary, real-world use case. I'm open to hearing your thoughts and feedback! All code and data can be found at my ***GitHub repository****. Feel free to connect with me on* ***LinkedIn****.*

Statistics | Data Science | Python | Advertising | Hypothesis Testing