The Effects of COVID-19 on US Exercise Behavior

University: City University of New York, School of Professional Studies

Course: DATA698, Analytics Master’s Research Project

Professor: Dr Paul Bailo

Student: Michael O’Donnell

Date: November 17th, 2020

# Abstract

The Coronavirus disease 2019, COVID-19, changed many day-to-day behaviors in the United States (Tynan & Howard, 2020). Some behaviors changed naturally like Netflix watching, while others changed in accordance with government orders like indoor dining (Gostin & Wiley, 2020). One day-to-day behavior in the US that seemingly changed both naturally and according to government orders was exercise (Ramirez-Campillo & Souza, 2020). Exercise is a large part of American health, and this project will analyze its apparent change since the onset of COVID-19. More specifically, this project will use public data to measure the change in 7 different types of US exercise routines. The goal of this project is to increase the knowledge about US behavioral change since the pandemic began with a healthy focus on exercise.

# Problem Statement

COVID-19 was declared a global pandemic by the World Health Organization on March 11th, 2020 (WHO, 2020). In response, the United States declared a national emergency on March 13th, 2020 to slow the spread of COVID-19. Regarding exercise, the national emergency implemented commercial gym closures, discouraged sharing fitness equipment, and ceased group exercise of 10 or more people (Ramirez-Campillo & Souza, 2020). Although these government order did not halt exercise in the US, they altered it in numerous ways:

* People exercising at commercial gyms had to exercise at home or elsewhere.
* People exercising with public equipment had to find their own equipment or change the equipment they exercised with (i.e. weights at a commercial gym).
* People exercising in groups had to exercise in smaller groups or alone (i.e. spin classes).

For commercial gyms, fitness equipment manufacturers, fitness apparel brands, nutritional supplement brands, fitness instructors/coaches, and many more participants in the fitness industry, the 3 exercise alterations listed above are critical. They may have changed US exercise behavior, which would change the fitness industry. To gain insight, this project focusses on the following question:

* **Which types of exercise are more popular in the US post-COVID-19?**

# Literature Review

## COVID-19 Timeline and Government-imposed Business Closures

The Coronavirus disease 2019, COVID-19, was declared a global pandemic by the World Health Organization on March 11th, 2020 (WHO, 2020). In response, the United States declared a national emergency on March 13th, 2020 to slow the spread of COVID-19. The national emergency imposed school closures, nonessential businesses closures, cancellation of large public gatherings, cancellation of as sporting and entertainment events (Chowell, 2020), travel restrictions, quarantines for travelers, and stay-at-home orders implemented by governors and mayors (Gostin & Wiley, 2020). In relation to exercise, these government orders closed commercial gyms, discouraged sharing fitness equipment, and ceased group exercise of 10 or more people (Ramirez-Campillo & Souza, 2020).

## COVID-19 US Attitude Changes and New Normal

Due to the global pandemic and US national emergency, the behavior and attitude of US public changed. Surveys from May 5-12, 2020 showed US citizens avoided groups of 10 or more persons and agreed with rules that prohibited inside dining (Tynan & Howard, 2020). Another study by the CDC showed a drastic decrease in US population movement from state-to-state during March, April, and May 2020 (Herlihy & Tynan, 2020). All this change is continually shaping a “new normal” in the United States (Roberts & Tehrani, 2020).

Some early and obvious differences are in this “new normal” are well-covered, like more working from home (Ahmad, 2020) and Netflix subscriptions (Dias, 2020). But there is not extensive research about the “new normal” regarding US exercise. Thus, this project will dig into the changes of this large component of American health (Fletcher et al., 1992)

## Relationship between Podcast Data and US Exercise Behavior

As humans, our interests are complicated. We are interested in many topics, some with greater magnitude and some for different reasons. For this project, it is important to consider that we are interested in topics that are relevant to ourselves (Tobarra, Robles-Gómez, Ros, Hernández, & Caminero, 2014). This link between interest and relevancy is important, although not perfect, because it allows us to use data regarding interest as a proxy for relevance.

One source of data regarding interest that is prevalent and available to us is podcast data. Podcasts have drastically gained popularity since 2014 (Durrani, Gotkin, & Laughlin, 2015) and podcasts with consistent episodes over many months indicate an interested audience (Mcclung & Johnson, 2010). Thus, this project leverages podcast data to measure 7 types of exercise in the US. The podcasts used in this project were chosen because they are the top exercise podcasts in the US according to Spotify.

# Summary of Literature Review Resources by Category

|  |  |  |
| --- | --- | --- |
| Category | Theoretical Construct | Source |
| COVID-19 Information | COVID-19 timeline | World Health Organization |
| COVID-19 Information | COVID-19 government orders | Chowell |
| COVID-19 Information | COVID-19 government orders | Gostin & Wiley |
| US Behavior | US commercial gym closures | Ramirez-Campillo & Souza |
| US Behavior | US attitudes regarding COVID-19 | Tynan & Howard, 2020 |
| US Behavior | US population movement | Center for Disease Control |
| US Behavior | A New Normal | Roberts & Tehrani |
| US Behavior | Working from home increase | Ahmad |
| US Behavior | Netflix subscription increase | Dias |
| US Exercise | Importance of exercise in US | Fletcher |
| Podcast data | Relationship between interest and relevance | Tobarra & Caminero |
| Podcast data | Podcast popularity | Durrani, Gotkin, & Laughlin |
| Podcast data | Podcast audiance | Mcclung & Johnson |

# Methodology

As stated above, this project’s research focused on answering the following question:

Which types of exercise are more popular in the US post-COVID-19?

To answer this question, a five-step methodology was used:

1. **Data Acquisition and Storage**: From the literary review conducted for this research project, podcast listening corresponds to interest in a subject and interest corresponds to relevancy. Thus, the primary data source to answer this question was data from US-based, exercise-related podcasts.

To acquire podcast data, this project connected to Spotify’s python API. To store data acquired from Spotify, this project leveraged a cloud-based Data Warehouse created with Oracle Database 12c.

1. **Data Exploration**: Each variable in the podcast dataset was explored for data type and distribution. Then, the three most important variables were closely explored: number of episodes, episode description, and episodes per podcast per week.
2. **Data Preparation**: All episodes with missing data were removed. Then, each episode record was labelled with a week, month, and pre-vs-post COVID-19 distinction and each podcast record was labelled with a genre and average episodes per week. Last, podcasts without enough episodes both pre-COVID-19 and post-COVID-19 were removed from the dataset.
3. **Natural Language Processing**: Each podcast episode description was run through a python natural language processor (library: NLTK) to determine the number of endurance-exercise-related terms used. Then, the number of endurance terms per podcast description were calculated.
4. **Two-Sample T-Test for Difference of Means**: With the dataset properly labelled and analyzed, 7 separate two-sample t-tests were setup to test changes in 7 different types of exercise. Each test used the same podcast dataset, and each measured change between pre-COVID-19 and post-COVID-19. For all t-tests, the null hypothesis was: the popularity of this type of exercise did not change after COVID-19. The alternate hypothesis was: this type of exercise did change post-COVID-19.

# Experimentation and Results

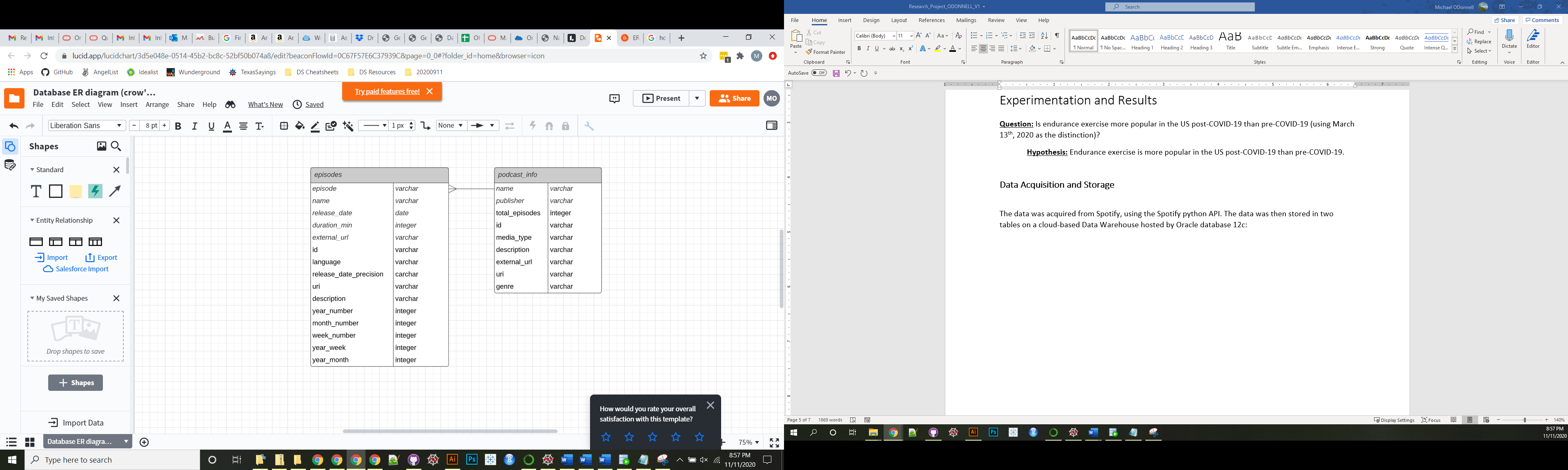
Question: Which types of exercise are more popular in the US post-COVID-19?

## **Data Acquisition and Storage**

The podcast data was acquired from Spotify using Spotify’s python API. The data was collected from 47 of the largest US exercise podcasts and totaled 11,447 episodes. More details about the data collection are below:

|  |  |
| --- | --- |
| Podcast | Episodes Collected |
| 9 to 5 Fitness | 75 |
| Align Podcast | 300 |
| AMRAP Mentality with Jason Khalipa | 50 |
| Another Mother Runner | 500 |
| BarBend Podcast | 125 |
| Ben Greenfield Fitness | 972 |
| Bulletproof Radio | 750 |
| Chasing Excellence | 125 |
| CITIUS MAG Podcast with Chris Chavez | 175 |
| Corpus Animus Podcast | 25 |
| Cultra Trail Running | 100 |
| Froning and Friends | 125 |
| IST CrossFit Podcast | 25 |
| Joe DeFranco's Industrial Strength Show | 250 |
| Kyle Kingsbury Podcast | 175 |
| Mind Pump: Raw Fitness Truth | 1450 |
| Misfit Podcast | 150 |
| Muscle For Life with Mike Matthews | 625 |
| No Meat Athlete Radio | 350 |
| Not Real Runners | 125 |
| On Purpose with Jay Shetty | 175 |
| Pace the Nation | 300 |
| Run Selfie Repeat | 50 |
| Run to the Top Podcast | The Ultimate Guide to Running | 400 |
| RunBuzz Running Podcast | 125 |
| Runners of NYC | 50 |
| Running Lean | 25 |
| Running Things Considered | 25 |
| SHRED CrossFit Podcast | 25 |
| The BibRave Podcast | 225 |
| The Brute Strength Podcast | 25 |
| The Jillian Michaels Show | 500 |
| The Mind Muscle Project | 750 |
| The Not Your Average Runner Podcast | 150 |
| The Refined Savage | 175 |
| The Runner's World Show | 75 |
| The Runners Zone | 25 |
| The Running Pod | 25 |
| The Running Public | 75 |
| The Strength Running Podcast | 150 |
| The WAG Podcast | 75 |
| The WODcast Podcast | 425 |
| Trail Runner Nation | 475 |
| Trail Running Women | 100 |
| TRAINED | 50 |
| Ultrarunnerpodcast.com | 425 |
| WHOOP Podcast | 75 |
| Total | 11447 |

As the data was collected from Spotify, it was stored in two tables on a cloud-based Data Warehouse hosted by Oracle database 12c. The structure of the tables is shown below:



In the “episodes” table, there are 11,447 rows (episodes) and 15 columns. In the “podcast\_info” table there are 47 rows (podcasts) and 9 columns.

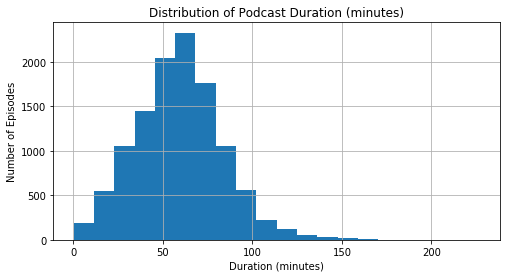
## **Data Exploration**

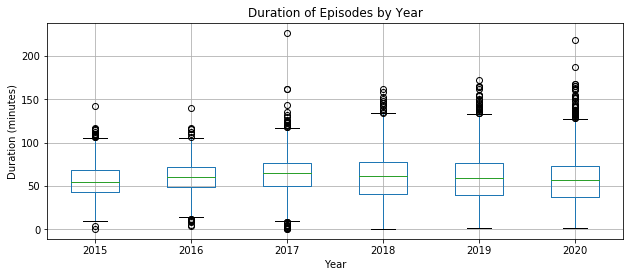
The episodes table was first explored at a high-level. Since the hypothesis relies on time (pre and post March 13th, 2020), the **number of episodes** were broken down by year and episodes pre-COVID-19 and post-COVID-19:

|  |  |
| --- | --- |
| Year | Episodes |
| 2020 | 2435 |
| 2019 | 2266 |
| 2018 | 1824 |
| 2017 | 1440 |
| 2016 | 1173 |
| 2015 | 1005 |
| 2014 | 540 |
| 2013 | 385 |
| 2012 | 291 |
| 2011 | 79 |

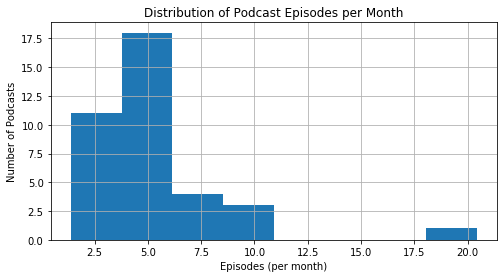
|  |  |
| --- | --- |
| Dates | Episodes |
| pre-COVID-19 | 9533 |
| post-COVID-19 | 1914 |

Next, the **duration of podcast episodes** was analyzed. First for overall distribution, then by year (including only years with >= 1000 episodes):





Last, the number of **episodes per month by podcast** was analyzed for distribution:



## **Data Preparation and Natural Language Processing**

To prepare the data, podcasts were removed that did not have enough episodes both pre-COVID-19 and post-COVID-19. The threshold for enough episodes was at least one episode per month from January 2020 to October 2020. This threshold filtered out 10 podcasts totaling 975 episodes. This left 37 podcasts totaling 10,472 episodes.

With the final set of 37 podcasts, all 10,472 episodes were run through a natural language processer in python (library: NLTK). The natural language processor counted the number of exercise-related words in each episode for 7 different types of exercise. The types of exercise and related words are detailed below:

|  |  |  |
| --- | --- | --- |
| Type of Exercise | Related Word | Occurrences |
| Running | running | 404 |
| run | 150 |
| jogging | 355 |
| jog | 147 |
| Cycling | cycling | 513 |
| cycle | 468 |
| biking | 305 |
| bike | 260 |
| Swimming | swimming | 488 |
| swim | 334 |
| freestyle | 422 |
| breaststroke | 421 |
| butterfly | 342 |
| backstroke | 487 |
| Walking | walking | 556 |
| walk | 402 |
| Weightlifting | weightlifting | 761 |
| lifting | 223 |
| lift | 69 |
| barbell | 425 |
| kettlebell | 571 |
| dumbbell | 423 |
| squat | 227 |
| press | 150 |
| bench | 528 |
| deadlift | 178 |
| Crossfit | WOD | 334 |
| box | 208 |
| amrap | 41 |
| afap | 53 |
| Yoga | yoga | 54 |
| streching | 494 |
| meditation | 209 |

## **Two-Sample T-Test for Difference of Means**

After Natural Language Processing was run on all 10,472 episode descriptions, the data was prepared to answer the research question: Which types of exercise are more popular in the US post-COVID-19?

To scientifically answer this question, a two-sample t-test was setup for each of the 7 types of exercise included in the project: running, cycling, swimming, walking, weightlifting, crossfit, and yoga.

Each t-test measured the mean of exercise-specific terms used per podcast episode description. Using weightlifting as an example, the mean of weightlifting-related terms per podcast episode description was compared between pre-COVID-19 and post-COVID-19.

Each t-test had the same null and alternative hypothesis:

Null Hypothesis: μpre-COVID exercise-specific terms= μpost-COVID exercise-specific terms

Alternate Hypothesis: μpre-COVID exercise-specific terms≠ μpost-COVID exercise-specific terms

The significance level (α) was 0.05 for all t-tests.

The results of the 7 t-tests are below:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Type of Exercise | Mean difference (post-pre) | t-statistic | p-value | significance level (α) | Hypothesis Result | Conclusion |
| Running | 0.142 | 2.043 | 0.041 | 0.05 | Reject the null | More popular |
| Cycling | -0.026 | -3.292 | 0.001 | 0.05 | Reject the null | Less popular |
| Swimming | -0.015 | -2.722 | 0.007 | 0.05 | Reject the null | Less popular |
| Walking | -0.01 | -0.068 | 0.945 | 0.05 | Cannot reject the null | No change |
| Weightlifting | 0.079 | 2.697 | 0.007 | 0.05 | Reject the null | More popular |
| Crossfit | -0.031 | -4.79 | 0.001 | 0.05 | Reject the null | Less popular |
| Yoga | -0.023 | -1.098 | 0.272 | 0.05 | Cannot reject the null | No change |

# Conclusions

This project started with a question, which types of exercise are more popular in the US post-COVID-19?

To measure exercise popularity, a connection was made between podcast episode content and relevancy to podcast listeners.

Hence, 11,447 podcast episode descriptions from 47 major US, exercise-related podcasts were acquired.

During the data exploration phase, podcast episodes were divided between pre-COVID-19 and post-COVID-19. In total, there were 9,522 pre-COVID-19 episodes and 1,914 post-COVID-19 episodes.

During the data preparation phase, 975 podcast episodes were removed because their corresponding podcasts could not be compared pre-COVID-19 and post-COVID-19.

After the data was explored and prepared, all podcast episode descriptions were run through a natural language processor in python to count the number of words related to each of the following 7 types of exercise: running, cycling, swimming, walking, weightlifting, crossfit, and yoga.

With the natural language processing results, 7 two-sample t-tests were run to compare the means of exercise-related terms used pre-COVID-19 and post-COVID-19 for each of the 7 types of exercise. From the two-sample t-tests, it was clear that running and weightlifting became more popular post-COVID-19:

|  |  |
| --- | --- |
| Type of Exercise | Conclusion |
| Running | More popular |
| Cycling | Less popular |
| Swimming | Less popular |
| Walking | No change |
| Weightlifting | More popular |
| Crossfit | Less popular |
| Yoga | No change |

# Resources

WHO. (2020, June 29). Listings of WHO's response to COVID-19. Retrieved October 20, 2020, from <https://www.who.int/news/item/29-06-2020-covidtimeline>

Chowell, G. (2020). The COVID-19 pandemic in the USA: What might we expect? *The Lancet,* *395*(10230), 1093-1094. doi:https://doi.org/10.1016/S0140-6736(20)30743-1

Gostin, L. O., & Wiley, L. F. (2020). Governmental Public Health Powers During the COVID-19 Pandemic. *Jama,* *323*(21), 2137-2138. doi:10.1001/jama.2020.5460

Czeisler MÉ, Tynan MA, Howard ME, et al. Public Attitudes, Behaviors, and Beliefs Related to COVID-19, Stay-at-Home Orders, Nonessential Business Closures, and Public Health Guidance — United States, New York City, and Los Angeles, May 5–12, 2020. *MMWR Morb Mortal Wkly Rep* 2020;69:751–758. DOI: [http://dx.doi.org/10.15585/mmwr.mm6924e1external icon](http://dx.doi.org/10.15585/mmwr.mm6924e1)

Moreland A, Herlihy C, Tynan MA, et al. Timing of State and Territorial COVID-19 Stay-at-Home Orders and Changes in Population Movement — United States, March 1–May 31, 2020. *MMWR Morb Mortal Wkly Rep* 2020;69:1198–1203. DOI: [http://dx.doi.org/10.15585/mmwr.mm6935a2external icon](http://dx.doi.org/10.15585/mmwr.mm6935a2).

Roberts, J. D., & Tehrani, S. O. (2020). Environments, Behaviors, and Inequalities: Reflecting on the Impacts of the Influenza and Coronavirus Pandemics in the United States. *International Journal of Environmental Research and Public Health,* *17*(12). doi:https://doi.org/10.3390/ijerph17124484

Gentil, P., Ramirez-Campillo, R., & Souza, D. (2020). Resistance Training in Face of the Coronavirus Outbreak: Time to Think Outside the Box. *Frontiers in physiology*, 11, 859. https://doi.org/10.3389/fphys.2020.00859

Ahmad, Tabrez, Corona Virus (COVID-19) Pandemic and Work from Home: Challenges of Cybercrimes and Cybersecurity (April 5, 2020). [http://dx.doi.org/10.2139/ssrn.3568830](https://dx.doi.org/10.2139/ssrn.3568830)

Dias, M. (2020). NETFLIX: FROM APOLLO 13 TO THE CORONAVIRUS PANDEMIC. *Global Scientific Journal,* *8*(8). doi:10.11216/gsj.2020.08.42678

Fletcher, G. F., Blair, S. N., Blumenthal, J., Caspersen, C., Chaitman, B., Epstein, S., . . . Pina, I. L. (1992). Statement on exercise. Benefits and recommendations for physical activity programs for all Americans. A statement for health professionals by the Committee on Exercise and Cardiac Rehabilitation of the Council on Clinical Cardiology, American Heart association. *Circulation,* *86*(1), 340-344. doi:10.1161/01.cir.86.1.340

Bollen, J., & Mao, H. (2011). Twitter Mood as a Stock Market Predictor. *Computer,* *44*(10), 91-94. doi:https://doi.org/10.1016/j.jocs.2010.12.007

Chu, S., Chen, H., & Sung, Y. (2015). Following brands on Twitter: An extension of theory of planned behavior. *International Journal of Advertising,* *35*(3), 421-437. doi:10.1080/02650487.2015.1037708

Ngoc, L. (2014). Behavior Pattern of Individual Investors in Stock Market. *International Journal of Business and Management,* *9*(1). doi:10.5539/ijbm.v9n1p1

Zhang, H., Dantu, R., & Cangussu, J. W. (2011). Socioscope: Human Relationship and Behavior Analysis in Social Networks. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans,* *41*(6), 1122-1143. doi:10.1109/tsmca.2011.2113335

Markman, K. M., & Sawyer, C. E. (2014). Why Pod? Further Explorations of the Motivations for Independent Podcasting. *Journal of Radio & Audio Media,* *21*(1), 20-35. doi:10.1080/19376529.2014.891211

Mchugh, S. (2020). Wide angle: Podcasts: Radio reinvented. *The UNESCO Courier,* *2020*(1), 6-9. doi:10.18356/4c5699b4-en

Bird, Steven, Edward Loper and Ewan Klein (2009), *Natural Language Processing with Python*. O’Reilly Media Inc.

Tobarra, L., Robles-Gómez, A., Ros, S., Hernández, R., & Caminero, A. C. (2014). Analyzing the students’ behavior and relevant topics in virtual learning communities. *Computers in Human Behavior,* *31*, 659-669. doi:10.1016/j.chb.2013.10.001

Durrani, M., Gotkin, K., & Laughlin, C. (2015). Serial, Seriality, and the Possibilities for the Podcast Format. *AMERICAN ANTHROPOLOGIST,* *000*(0). Retrieved November 17, 2020

Mcclung, S., & Johnson, K. (2010). Examining the Motives of Podcast Users. *Journal of Radio & Audio Media,* *17*(1), 82-95. doi:10.1080/19376521003719391

# Appendix with Code

## Data Acquisition

1. # get podcast episodes from a show
2. **def** get\_all\_podcast\_episodes(self, showid, market = 'US'):
3. headers = self.get\_resource\_header()
5. podcast\_name = self.get\_podcast\_info\_by\_id(showid)['name']
6. num\_episodes = self.get\_podcast\_info\_by\_id(showid)['total\_episodes']
8. limit = 25
9. offset = 0
10. num\_runs = num\_episodes//limit
12. episode\_df = pd.DataFrame(columns = ['podcast', 'name','release\_date','duration\_min',
13. 'external\_urls','id', 'language',
14. 'release\_date\_precision', 'uri','description'])
16. **for** i **in** range(num\_runs):
18. endpoint = f"https://api.spotify.com/v1/shows/{showid}/episodes?offset={offset}&limit={limit}&market=US"
19. lookup\_url = f"{endpoint}"
21. r = requests.get(lookup\_url, headers = headers)
22. **if** r.status\_code **not** **in** range(200,299):
23. **return** "somethings wrong..."
25. raw\_json = r.json()


29. **for** i **in** range(limit):
30. # create a dict with the data
31. temp\_dict = {'podcast': podcast\_name,
32. 'name': raw\_json['items'][i]['name'],
33. 'release\_date': raw\_json['items'][i]['release\_date'],
34. 'duration\_min': round((raw\_json['items'][i]['duration\_ms'])/60000,2),
35. 'external\_urls': raw\_json['items'][i]['external\_urls'],
36. 'id': raw\_json['items'][i]['id'],
37. 'language': raw\_json['items'][i]['language'],
38. 'release\_date\_precision': raw\_json['items'][i]['release\_date\_precision'],
39. 'uri': raw\_json['items'][i]['uri'],
40. 'description': raw\_json['items'][i]['description']}
42. df = pd.DataFrame(temp\_dict, columns = ['podcast','name','release\_date','duration\_min',
43. 'external\_urls','id', 'language',
44. 'release\_date\_precision', 'uri','description'])
45. episode\_df = episode\_df.append(df)
47. offset = offset + limit
49. episode\_df = episode\_df.reset\_index(drop=True)
50. **return** episode\_df

## Data Exploration

1. # create histogram of episode duration
2. plt.figure(figsize=(8,4))
4. plt.hist(df['duration\_min'], bins = 20)
5. plt.xlabel("Duration (minutes)")
6. plt.ylabel("Number of Episodes")
7. plt.title("Distribution of Podcast Duration (minutes)")
8. plt.grid(True)
10. plt.show()
12. # create boxplot of episode duration by year
13. df[df['year\_number']>2014].boxplot(by='year\_number',
14. column=['duration\_min'],
15. grid=True,
16. figsize = (10,4))
17. plt.title("Duration of Episodes by Year")
18. plt.suptitle("")
19. plt.ylabel("Duration (minutes)")
20. plt.xlabel("Year")
21. plt.show()

## Data Preparation

1. # add year, month, and week numbers to dataframe
2. df['year\_number'] = pd.to\_datetime(df['release\_date']).dt.year
3. df['month\_number'] = pd.to\_datetime(df['release\_date']).dt.month
4. df['week\_number'] = pd.to\_datetime(df['release\_date']).dt.week
5. df['year\_week'] = df['year\_number'].map(str) + df['week\_number'].map(str)
6. df['year\_month'] = df['year\_number'].map(str) + df['month\_number'].map(str)
8. **for** i **in** df['podcast'].unique():
9. temp\_df = df[df['podcast']==i]
10. **print**(i, "|", round(len(temp\_df)/len(temp\_df['year\_month'].unique()),2))

## Natural Language Processing and Two-Sample T-Test

1. # function to setup hypothesis test for podcast descriptions pre vs post COVID
2. **def** podcast\_description\_hypothesis\_test(csv, list\_of\_words, alpha, title = "hypothesis test"):
4. # first, read in the csv
5. # 'data/final\_datasets/relevant\_episode\_data\_v2.csv'
6. episode\_desc\_df = read\_episode\_descriptions(csv)
7. preCOVID\_episode\_desc\_df = episode\_desc\_df[pd.to\_datetime(episode\_desc\_df['release\_date']) < '2020-03-13']
8. postCOVID\_episode\_desc\_df = episode\_desc\_df[pd.to\_datetime(episode\_desc\_df['release\_date']) >= '2020-03-13']
10. # format description column for tokenization
11. desc\_df = create\_episode\_df(episode\_desc\_df)
12. preCOVID\_df = create\_episode\_df(preCOVID\_episode\_desc\_df)
13. postCOVID\_df = create\_episode\_df(postCOVID\_episode\_desc\_df)
15. # tokenize all 3 dataframes
16. episode\_tokens = tokenize\_descriptions(desc\_df)
17. preCOVID\_tokens = tokenize\_descriptions(preCOVID\_df)
18. postCOVID\_tokens = tokenize\_descriptions(postCOVID\_df)
20. # create final dataframes for analyses
21. episode\_dataset = pd.DataFrame(episode\_tokens[1])
22. preCOVID\_dataset = pd.DataFrame(preCOVID\_tokens[1])
23. postCOVID\_dataset = pd.DataFrame(postCOVID\_tokens[1])
25. # add number of exercise terms to each row in above datasets
26. episodes\_terms = []
27. **for** i **in** range(len(episode\_dataset)):
28. episodes\_terms.append(count\_terms(episode\_dataset['description'][i], list\_of\_words))
29. episode\_dataset['number\_of\_words'] = episodes\_terms
31. preCOVID\_terms = []
32. **for** i **in** range(len(preCOVID\_dataset)):
33. preCOVID\_terms.append(count\_terms(preCOVID\_dataset['description'][i], list\_of\_words))
34. preCOVID\_dataset['number\_of\_words'] = preCOVID\_terms
36. postCOVID\_terms = []
37. **for** i **in** range(len(postCOVID\_dataset)):
38. postCOVID\_terms.append(count\_terms(postCOVID\_dataset['description'][i], list\_of\_words))
39. postCOVID\_dataset['number\_of\_words'] = postCOVID\_terms
41. # find mean, standard deviation, and count of pre-COVID-19 exercise terms
42. preCOVID\_mean =  preCOVID\_dataset['number\_of\_words'].mean()
43. preCOVID\_sd = preCOVID\_dataset['number\_of\_words'].std()
44. preCOVID\_episodes = len(preCOVID\_dataset['number\_of\_words'])
46. # find mean, standard deviation, and count of post-COVID-19 exercise terms
47. postCOVID\_mean =  postCOVID\_dataset['number\_of\_words'].mean()
48. postCOVID\_sd = postCOVID\_dataset['number\_of\_words'].std()
49. postCOVID\_episodes = len(postCOVID\_dataset['number\_of\_words'])
51. # use scipy to get test statistic and p-value
52. hyp\_test = stats.ttest\_ind(postCOVID\_dataset['number\_of\_words'],preCOVID\_dataset['number\_of\_words'], equal\_var=False)
53. t\_statistic = hyp\_test[0]
54. p\_value = hyp\_test[1]
56. # print findings
57. **print**("words tested:", list\_of\_words)
58. **print**("variance of postCOVID:", round(postCOVID\_dataset['number\_of\_words'].var(), 2))
59. **print**("variance of preCOVID:", round(preCOVID\_dataset['number\_of\_words'].var(), 2))
60. **print**("difference in means = ", round((postCOVID\_mean-preCOVID\_mean), 3))
61. **print**("t = ", round(t\_statistic, 3))
62. **print**("p = ", round(p\_value, 3))
63. **print**("alpha = ", str(alpha))
64. **if** stats.ttest\_ind(postCOVID\_dataset['number\_of\_words'],preCOVID\_dataset['number\_of\_words'], equal\_var=False)[1] < alpha:
65. **print**("Hypothesis result: REJECT the null hypothesis")
66. **else**:
67. **print**("Hypothesis result: cannot reject the null hypothesis")
69. # export dataset
70. postCOVID\_dataset['post\_COVID'] = 1
71. preCOVID\_dataset['post\_COVID'] = 0
72. df = preCOVID\_dataset
73. df = preCOVID\_dataset.append(postCOVID\_dataset)
74. df.to\_csv(f"{title}\_data.csv")