The Effects of COVID-19 on US Exercise Behavior

Michael O'Donnell

City University of New York, School of Professional Studies

DATA698, Analytics Master's Research Project

Dr. Paul Bailo

November 17, 2020

Abstract

The new coronavirus disease that emerged in 2019, COVID-19, modified many day-to-day behaviors in the United States (Tynan & Howard, 2020). Certain behaviors changed naturally, such as Netflix viewing habits, while others changed in accordance with government orders, such as restriction of indoor restaurant dining (Gostin & Wiley, 2020). One aspect of day-to-day behavior in the US that seemingly changed both naturally and as a result of government orders was exercise (Ramirez-Campillo & Souza, 2020). Exercise behaviors are a significant component of overall American health, and this project will analyze apparent changes in these behaviors since the onset of COVID-19. More specifically, this project will use public data to measure changes in seven different types of US exercise routine. The goal of this project is to increase knowledge pertaining to US behavioral changes 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 11, 2020 (WHO, 2020). In response, the United States declared a national emergency on March 13, 2020 to slow the spread of the virus. Regarding guidelines for exercise, the national emergency declaration implemented commercial gym closures, discouraged the sharing of fitness equipment, and ordered the cessation of group exercise when ten or more people were in attendance (Ramirez-Campillo & Souza, 2020). Although these government orders did not halt exercise in the US, they altered it in numerous ways:

- People habitually 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., those using weights at a commercial gym).
- People exercising in groups had to exercise in smaller groups or alone (i.e., those attending exercise 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 three exercise alterations listed above are critical. The resulting alterations to US exercise behavior have significantly altered the fitness industry as a whole. To gain insight into these changes, this project focuses 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 new coronavirus disease that emerged in 2019, COVID-19, was declared a global pandemic by the World Health Organization on March 11, 2020 (WHO, 2020). In response, the United States declared a national emergency on March 13, 2020 to slow the spread of the virus. The national emergency imposed school closures, nonessential businesses closures, cancellation of large public gatherings, cancellation of 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 the sharing of fitness equipment, and demanded the cessation of group exercise when ten or more people were in attendance (Ramirez-Campillo & Souza, 2020).

COVID-19 US Attitude Changes and the New Normal

As a result of the global pandemic and the US national emergency, the behaviors and attitudes of members of the US public changed. Surveys from May 5-12, 2020 revealed that US citizens began to avoid gathering in groups of ten or more persons and complied 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 the months of March, April, and May of 2020 (Herlihy & Tynan, 2020). These changes are continually shaping a “new normal” in the United States (Roberts & Tehrani, 2020).

Some early and obvious differences inherent to this “new normal” have already been well covered, such as the increased number of people working from home (Ahmad, 2020) and the rise in Netflix subscriptions (Dias, 2020). Thus far, however, there has not been extensive research examining the “new normal” regarding US exercise habits and behaviors. Therefore, this project

will examine the changes that have taken place in this important aspect of American behavior relevant to health (Fletcher et al., 1992)

Relationship between Podcast Data and US Exercise Behavior

As humans, our interests are complicated and varied, with differing degrees of magnitude and purpose. For this project, it is important to consider that we are interested in topics that are relevant to ourselves (Tobarra et al., 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 dramatically gained in popularity since 2014 (Durrani, Gotkin, & Laughlin, 2015) and podcasts with a consistent number of episodes over many months indicate an interested audience (Mcclung & Johnson, 2010). Thus, this project leverages podcast data to measure seven 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.

Table 1

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

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 the US	Fletcher
Podcast Data	Relationship between interest and relevance	Tobarra & Caminero
Podcast Data	Podcast popularity	Durrani, Gotkin, & Laughlin
Podcast Data	Podcast audience	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**: According to 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 used to answer this question was US-based, exercise-related podcasts.
    - To acquire podcast data, this project connected to Spotify's API with python. To store data acquired from Spotify, this project leveraged a cloud-based data warehouse created using Oracle Database 12c.

2. **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.
3. **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 the average number of episodes per week. Finally, podcasts without enough episodes both pre-COVID-19 and post-COVID-19 were removed from the dataset.
4. **Natural Language Processing**: Each podcast episode description was run through a python natural language processor (library: NLTK) to determine the number of exercise-related terms used. Then, the number of exercise terms per podcast description was calculated.
5. **Two-Sample T-Test for Difference of Means**: With the dataset properly labelled and analyzed, seven separate two-sample t-tests were set up to test changes in seven different types of exercise. Each test used the same podcast dataset and each measured changes between pre-COVID-19 and post-COVID-19. For all t-tests, the null hypothesis was that the popularity of this type of exercise did not change after COVID-19. The alternate hypothesis was that 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 pertaining to the data collection are found in the chart below:

Table 2

Episodes Collected by Podcast

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

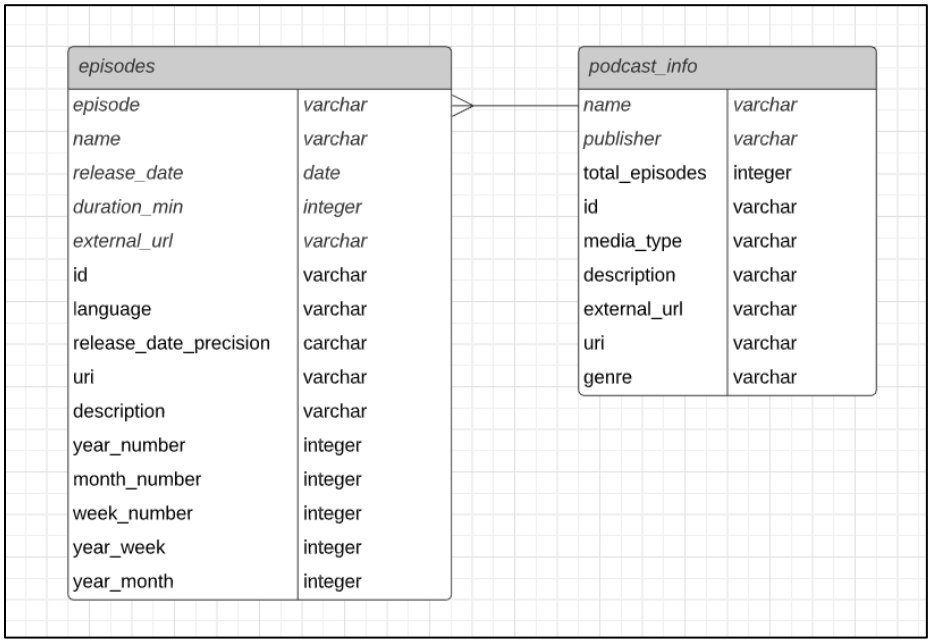
Mind Pump: Raw Fitness Truth	1450
Misfit Podcast	150
Muscle For Life with Mike Matthews	625
No Meat Athlete Radio	350
Not Real Runners	125
On Purpose with Jay Shetty	175
Pace the Nation	300
Run Selfie Repeat	50
Run to the Top Podcast | The Ultimate Guide to Running	400
RunBuzz Running Podcast	125
Runners 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:

Figure 1

Entity Relationship Diagram of Collected Data



In the "episodes" table, there are 11,447 rows (episodes) and 15 columns. In the "podcast info" table there are 47 rows (podcasts) and nine columns.

Data Exploration

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

Table 3

Episodes Collected by Year

Year	Episodes
2020	2435
2019	2266
2018	1824
2017	1440

2016	1173
2015	1005
2014	540
2013	385
2012	291
2011	79

Table 4

Episodes Collected pre-COVID-19 and post-COVID-19

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):

Figure 2

Histogram of Episode Duration in Minutes

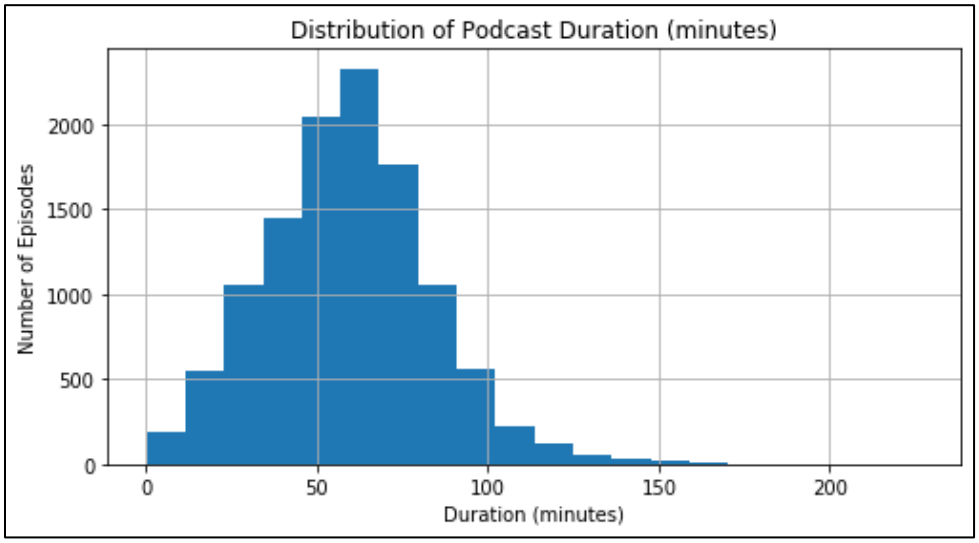
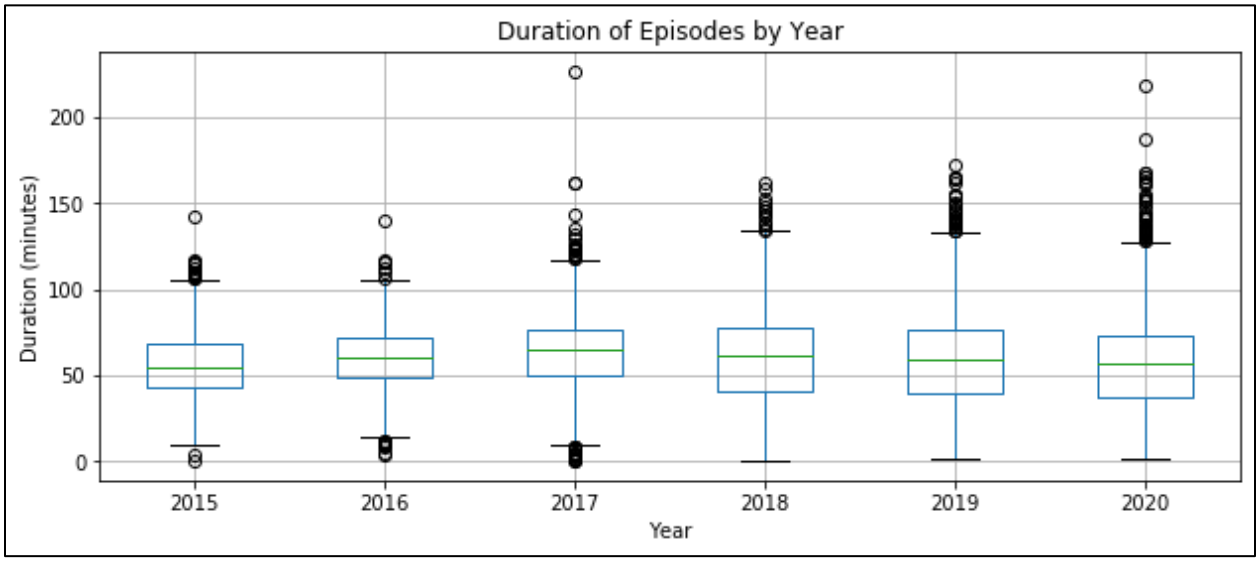


Figure 3

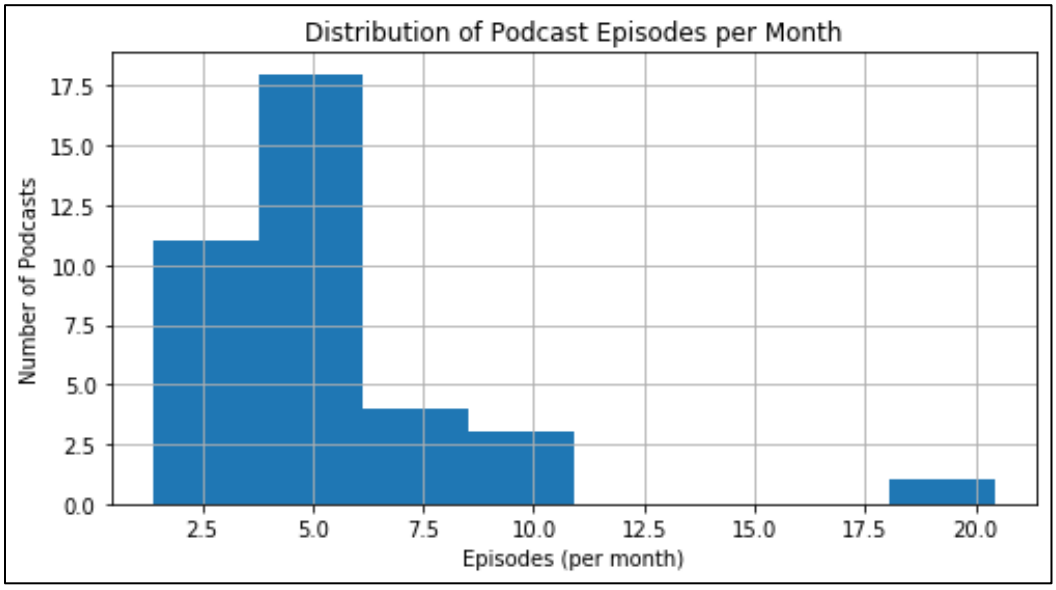
Boxplot of Episode Duration in Minutes by Year



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

Figure 4

Histogram of Episodes per Month by Podcast



Data Preparation and Natural Language Processing

To prepare the data, podcasts were removed that did not have a sufficient number of episodes both pre-COVID-19 and post-COVID-19. The threshold for a sufficient number of 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 exercise-related words are detailed below:

Table 5

Frequency Table of Exercise Related Words

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 set up for each of the seven 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: $\mu_{\text{pre-COVID exercise-specific terms}} = \mu_{\text{post-COVID exercise-specific terms}}$

Alternate Hypothesis: $\mu_{\text{pre-COVID exercise-specific terms}} \neq \mu_{\text{post-COVID exercise-specific terms}}$

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

The results of the seven t-tests are presented below:

Table 6

Hypothesis Test Results for each type of Exercise

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 began with a question: Which types of exercise are most popular in the US post-COVID-19? To measure exercise popularity, a connection was made between podcast episode content and relevancy to podcast listeners. Therefore, 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 episodes. 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 between 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 seven types of exercise: running, cycling, swimming, walking, weightlifting, Crossfit, and yoga.

Using the natural language processing results, seven 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 seven types of exercise. From the two-sample t-tests, it was clear that running and weightlifting became more popular post-COVID-19:

Table 7

Research Conclusions

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

References

Ahmad, T. (2020, April 5). Corona virus (COVID-19) pandemic and work from home: Challenges of cybercrimes and cybersecurity. http://dx.doi.org/10.2139/ssrn.3568830

Bird, S., Loper, E., & Klein, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc.

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

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

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

Czeisler, M. É., Tynan M. A., Howard, M.E. (2020). 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. *Morbidity and Mortality Weekly Report 2020*(69):751–758.
doi: http://dx.doi.org/10.15585/mmwr.mm6924e1external icon

Dias, M. (2020). Netflix: From Apollo 13 to the coronavirus pandemic. *Global Scientific Journal, 8*(8): doi:10.11216/gsj.2020.08.42678

Durrani, M., Gotkin, K., & Laughlin, C. (2015). Serial, seriality, and the possibilities for the podcast format. *American Anthropologist*

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

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

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.2020.5460Markman, 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

Mcclung, S., & Johnson, K. (2010). Examining the motives of podcast users. *Journal of Radio & Audio Media, 17*(1), 82–95. doi:10.1080/19376521003719391

Moreland, A., Herlihy, C., Tynan, M. A. 2020). Timing of state and territorial COVID-19 stay-at-home orders and changes in population movement: United States. *Morbidity and Mortality Weekly Report 2020*(69):1198–1203. doi: http://dx.doi.org/10.15585/mmwr.mm6935a2external icon.

Ngoc, L. (2014). Behavior pattern of individual investors in stock market. *International Journal of Business and Management, 9*(1). doi:10.5539/ijbm.v9n1p1

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

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

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

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

Appendix with Code

Data Acquisition

```
# get podcast episodes from a show
def get_all_podcast_episodes(self, showid, market = 'US'):

  headers = self.get_resource_header()

  podcast_name = self.get_podcast_info_by_id(showid)['name']
  num_episodes = self.get_podcast_info_by_id(showid)['total_episodes']

  limit = 25
  offset = 0
  num_runs = num_episodes//limit

  episode_df = pd.DataFrame(columns = ['podcast', 'name','release_date','duration_min',
              'external_urls','id', 'language',
              'release_date_precision', 'uri','description'])

  for i in range(num_runs):

    endpoint = f"https://api.spotify.com/v1/shows/{showid}/episodes?offset={offset}&limit={limit}&market=US"
    lookup_url = f"{endpoint}"

    r = requests.get(lookup_url, headers = headers)
    if r.status_code not in range(200,299):
      return "somethings wrong..."

    raw_json = r.json()



    for i in range(limit):
      # create a dict with the data
      temp_dict = {'podcast': podcast_name,
            'name': raw_json['items'][i]['name'],
            'release_date': raw_json['items'][i]['release_date'],
            'duration_min': round((raw_json['items'][i]['duration_ms'])/60000,2),
            'external_urls': raw_json['items'][i]['external_urls'],
            'id': raw_json['items'][i]['id'],
            'language': raw_json['items'][i]['language'],
            'release_date_precision': raw_json['items'][i]['release_date_precision'],
            'uri': raw_json['items'][i]['uri'],
            'description': raw_json['items'][i]['description']}

      df = pd.DataFrame(temp_dict, columns = ['podcast','name','release_date','duration_min',
                  'external_urls','id', 'language',
                  'release_date_precision', 'uri','description'])
      episode_df = episode_df.append(df)

    offset = offset + limit

  episode_df = episode_df.reset_index(drop=True)
  return episode_df
```

Data Exploration

```
# create histogram of episode duration
plt.figure(figsize=(8,4))

plt.hist(df['duration_min'], bins = 20)
plt.xlabel("Duration (minutes)")
plt.ylabel("Number of Episodes")
plt.title("Distribution of Podcast Duration (minutes)")
plt.grid(True)

plt.show()

# create boxplot of episode duration by year
df[df['year_number']>2014].boxplot(by='year_number',
      column=['duration_min'],
      grid=True,
      figsize = (10,4))
plt.title("Duration of Episodes by Year")
plt.suptitle("")
plt.ylabel("Duration (minutes)")
plt.xlabel("Year")
plt.show()
```

Data Preparation

```
# add year, month, and week numbers to dataframe
df['year_number'] = pd.to_datetime(df['release_date']).dt.year
df['month_number'] = pd.to_datetime(df['release_date']).dt.month
df['week_number'] = pd.to_datetime(df['release_date']).dt.week
df['year_week'] = df['year_number'].map(str) + df['week_number'].map(str)
df['year_month'] = df['year_number'].map(str) + df['month_number'].map(str)

for i in df['podcast'].unique():
  temp_df = df[df['podcast']==i]
  print(i, "|", round(len(temp_df)/len(temp_df['year_month'].unique()),2))
```

Natural Language Processing and Two-Sample T-Test

```
# function to setup hypothesis test for podcast descriptions pre vs post COVID
def podcast_description_hypothesis_test(csv, list_of_words, alpha, title = "hypothesis test"):

  # first, read in the csv
  # 'data/final_datasets/relevant_episode_data_v2.csv'
  episode_desc_df = read_episode_descriptions(csv)

preCOVID_episode_desc_df = episode_desc_df[pd.to_datetime(episode_desc_df['release_date']) < '2020-03-13']

postCOVID_episode_desc_df = episode_desc_df[pd.to_datetime(episode_desc_df['release_date']) >= '2020-03-13']

  # format description column for tokenization
```

```
  desc_df = create_episode_df(episode_desc_df)
  preCOVID_df = create_episode_df(preCOVID_episode_desc_df)
  postCOVID_df = create_episode_df(postCOVID_episode_desc_df)

  # tokenize all 3 dataframes
  episode_tokens = tokenize_descriptions(desc_df)
  preCOVID_tokens = tokenize_descriptions(preCOVID_df)
  postCOVID_tokens = tokenize_descriptions(postCOVID_df)

  # create final dataframes for analyses
  episode_dataset = pd.DataFrame(episode_tokens[1])
  preCOVID_dataset = pd.DataFrame(preCOVID_tokens[1])
  postCOVID_dataset = pd.DataFrame(postCOVID_tokens[1])

  # add number of exercise terms to each row in above datasets
  episodes_terms = []
  for i in range(len(episode_dataset)):
episodes_terms.append(count_terms(episode_dataset['description'][i], list_of_words))
  episode_dataset['number_of_words'] = episodes_terms

  preCOVID_terms = []
  for i in range(len(preCOVID_dataset)):
preCOVID_terms.append(count_terms(preCOVID_dataset['description'][i], list_of_words))
  preCOVID_dataset['number_of_words'] = preCOVID_terms

  postCOVID_terms = []
  for i in range(len(postCOVID_dataset)):
postCOVID_terms.append(count_terms(postCOVID_dataset['description'][i], list_of_words))
  postCOVID_dataset['number_of_words'] = postCOVID_terms

  # find mean, standard deviation, and count of pre-COVID-19 exercise terms
  preCOVID_mean = preCOVID_dataset['number_of_words'].mean()
  preCOVID_sd = preCOVID_dataset['number_of_words'].std()
  preCOVID_episodes = len(preCOVID_dataset['number_of_words'])

  # find mean, standard deviation, and count of post-COVID-19 exercise terms
  postCOVID_mean = postCOVID_dataset['number_of_words'].mean()
  postCOVID_sd = postCOVID_dataset['number_of_words'].std()
  postCOVID_episodes = len(postCOVID_dataset['number_of_words'])

  # use scipy to get test statistic and p-value
hyp_test = stats.ttest_ind(postCOVID_dataset['number_of_words'],preCOVID_dataset['numbe
r_of_words'], equal_var=False)
  t_statistic = hyp_test[0]
  p_value = hyp_test[1]

  # print findings
  print("words tested:", list_of_words)
  print("variance of postCOVID:", round(postCOVID_dataset['number_of_words'].var(), 2))
  print("variance of preCOVID:", round(preCOVID_dataset['number_of_words'].var(), 2))
  print("difference in means = ", round((postCOVID_mean-preCOVID_mean), 3))
  print("t = ", round(t_statistic, 3))
  print("p = ", round(p_value, 3))
  print("alpha = ", str(alpha))

if stats.ttest_ind(postCOVID_dataset['number_of_words'],preCOVID_dataset['number_of_wor
ds'], equal_var=False)[1] < alpha:
```

```
    print("Hypothesis result: REJECT the null hypothesis")
  else:
    print("Hypothesis result: cannot reject the null hypothesis")

  # export dataset
  postCOVID_dataset['post_COVID'] = 1
  preCOVID_dataset['post_COVID'] = 0
  df = preCOVID_dataset
  df = preCOVID_dataset.append(postCOVID_dataset)
  df.to_csv(f"{title}_data.csv")
```