The Effects of COVID-19 on US Exercise Behavior: Data Collection and Analysis

Michael O’Donnell

City University of New York, School of Professional Studies

DATA698, Analytics Master’s Research Project

Dr. Paul Bailo

Debember 4, 2020

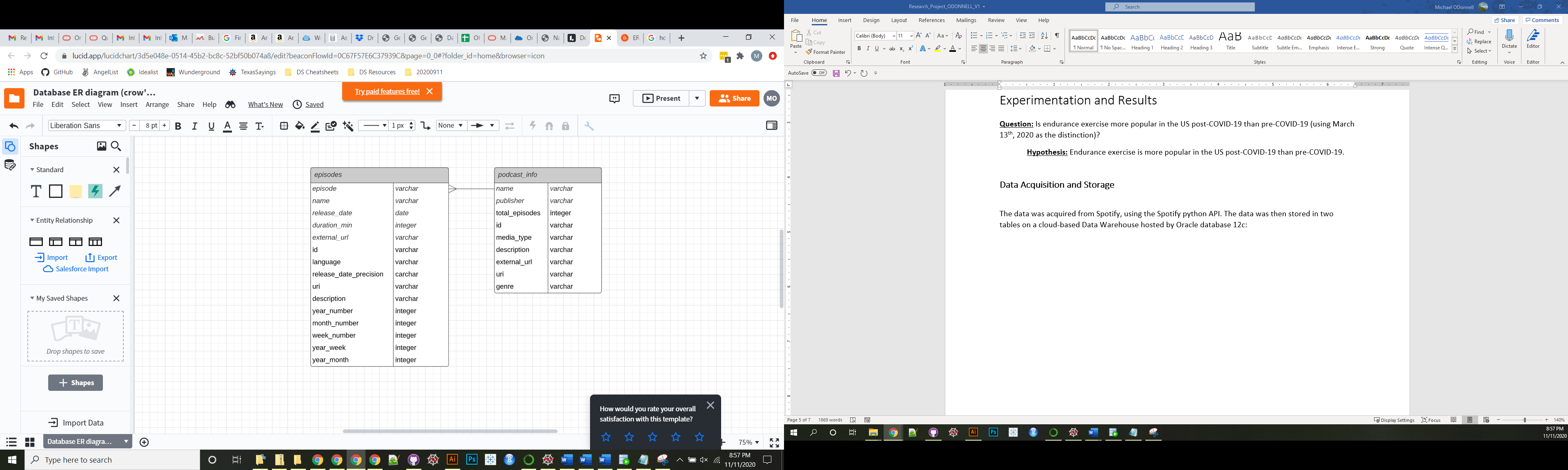
# Data Collection

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

For the research, I collected as much publicly available, US exercise data as I could find. Then, I read dozens of research papers to correctly tie the data to my research question. As a result, I only used Spotify podcast data for my project; I was able to tie podcast data to US exercise behavior. However, all the data I collected was important to the research process and it is therefore all listed below.

## Spotify Podcast Data

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. The following fields were collected from each podcast and episode:



*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 |

*Episodes Collected pre-COVID-19 versus post-COVID-19*

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

## Yahoo Finance Data

To assess public fitness companies’ stock prices pre-COVID-19 against post-COVID-19, Yahoo Finance data was scraped for 12 companies. The stock prices were compared for multiple durations before and after the pandemic was declared a US national emergency:



## Twitter

Like Spotify podcast data, Twitter data was social, public, and could be categorized. Thus, tweets from commercial gyms were analyzed pre-COVID-19 and post-COVID-19. The purpose of this was to analyze the change in engagement of tweets from large US commercial gyms. The results were quite interesting, but I was unable to gather enough tweets for statistically significant results. An example of the data is below.



## Garmin

Garmin specializes in tracking exercise data. So this data would have been a home run for the project. Unfortunately, although I was granted access to Garmin’s API I was not approved to view more than my own Garmin data. Hence, this data was not usable for the research.

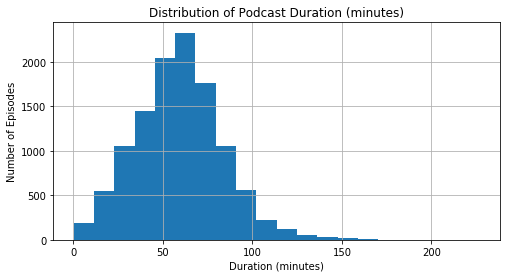
## Strava

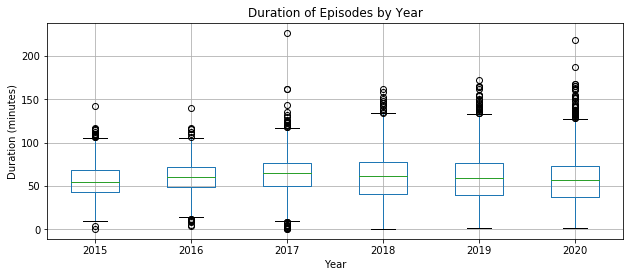
Like Garmin, Strava’s forte is tracking exercise data. But again, I was not granted access to more than my personal data. However, I explored joining multiple large Strava clubs (running clubs on the app, etc.) to use club data. In the end, this would not have produced a random sample of the US exercise community.

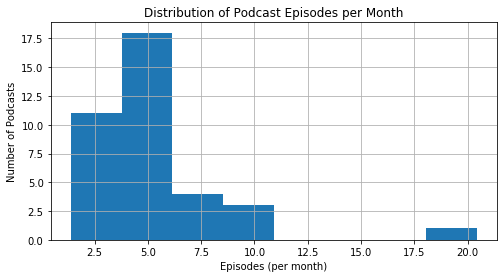
# Exploratory Data Analysis

## Pre-NLP

The exploratory data analysis done before applying Natural Language Processing to the Spotify Podcast Episode descriptions was interesting and insightful. The primary function of this analysis was to eliminate episodes that did not belong in the research. To do this, I focused on episode duration in minutes and number of episodes per month for each podcast. Some simple findings are below.







## Post-NLP

After the Natural Language Processing occurred, more exploration was necessary to ensure enough data was present for each type of exercise included in the problem statement. Fortunately, there were enough keywords for each type of exercise so I did not have to alter the problem statement.

# Progress on Model/Analysis

## Data Preparation

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.

## Final Analysis: Hypothesis Testing

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

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 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 |