Ryan Canfield, Pat Norcross, Alex Butera

Professor St. Hilaire

DSE4900 Data Science Capstone

14 December 2023

Final Report Draft

**Problem Statement:**

In recent years, there has been a stronger push to prioritize mental health. Mental health can manifest in many unexpected ways. A place where you might not expect it would be in the gym because exercise and fitness are usually synonymous with being healthy. The areas that we aim to target are Body Dysmorphic Disorders, specifically, muscle dysmorphia. The overall objective of this research is to determine patterns or connections between the training behaviors and habits of people with muscle dysmorphia, their mental health status, and their substance or supplement use. We are also looking at what other mental health disorders could increase within patients that have muscle dysmorphia.

**Introduction and Background:**

Muscle dysmorphia, commonly shortened to MD, is the psychological perception of being “too small” or “not muscular enough”. When they have a normal or muscular build, and in many cases, an objectively strong physique. With this information in mind, our overall goal is to understand the characteristics of the individuals with MD and to gain a better understanding of what traits might play a factor in their diagnosis. We will specifically focus on the relationship between bodybuilding fitness metrics and the individual's mental health. Along with bodybuilding fitness metrics, this research project will examine the correlation between those who use supplements or substances and the potential mental health diagnoses that adversely come with it.

Prior to completing this project, we had researched and read other studies on this topic. In one study, *The emergence of Exercise Addiction, Body Dysmorphic Disorder, and other image-related psychopathological correlates in fitness settings: A cross sectional study,* which was found on PubMed, they looked at patients from multiple fitness settings and graded them with different surveys to see if the individuals had body dysmorphic disorder. The surveys or inventories they used were Exercise Addiction Inventory (EAI), Appearance Anxiety Inventory (AAI), and Rosenburg’s Self-Esteem Scale (RSE). In their conclusions, they noted 39% of the 1,700 people they measured had body dysmorphic disorder and 40% used unprescribed supplements or substances (Corazza, 2019). No statistics or other tools were used besides filling out the different surveys. In another study relating to muscle dysmorphia, they focused on revealing a deeper understanding of MD’s implications and causes. The signs and symptoms were given, showing how it could be recognized. In their conclusion, they discussed that muscle dysmorphia can create changes in the individual’s substance use and training behaviors. It is unclear what methods they used to come to this conclusion. The paper did not mention any statistical methods or surveys done to prove their thesis. It seemed as if it was a summary paper based on a culmination of other articles. In Dr. SantaBarbara’s article, he recorded and measured twenty-one men over the course of a few weeks. This included getting their information, recording the amount of weight that they would lift in various exercises, and keeping record of their baseline mental health. He then looked at the means and standard deviations of his patients using a t-test to analyze the changes in body image and perceived muscle size between conditions (i.e., intensities) over time (SantaBarbara, 2020). From all of these articles, we can deduce that these studies largely focus on informing and surveying the population. Therefore, our objectives will use statistics to determine the relationship of the patient's training behaviors, as well as their substance or supplement use, to compare them to mental health beyond what the previous research papers accomplished. This study will also use different mental health inventories, along with a new survey sample, making it possible that it could lead to different results.

With the previous research done and their goals attained, our client has generated three major objectives for our project. Our first objective is to **characterize resistance training behavior and mental health in a sample of men with symptoms of muscle dysmorphia.** This will give us a general idea of the patients in our dataset, specifically their mental health status, their workout behaviors and other useful demographics about them. By learning these things, it will give us more context and understanding moving forward of who the people are in our project. The second objective will **explore whether a worse mental health, including symptoms of muscle dysmorphia, are associated with more frequent and intense resistance training behaviors.** This goal is to see if there are any meaningful patterns in patients with MD that will allow us to identify why the patient's training intensity is changing based on what potential mental health conditions they have been diagnosed with. Our last objective strays away from the individual’s training habits and focuses on their use of supplements and substances. This includes things such as using protein, creatine, and pre-workout, or using drugs like cigarettes, marijuana or other illicit substances. This objective is to **explore any associations between mental health and the use of substances or supplements.** This will allow us to see any potential relationships between increased supplement or substance use and how it can affect these individuals' mental health. These goals will be accomplished using the two datasets gathered during the survey and interview process.

**Data and Methods:**

The dataset we are using for this project was given to us by our client Dr. SantaBarbara. It contains information about people with muscle dysmorphia. This includes demographic data like age, race, and living situation, as well as muscle measurements and questions regarding mental health. The mental health data was collected by asking patients questions in a survey-like manner. It was collected in two parts by our client. The first dataset was collected when he did his master’s dissertation and the second half was collected during his doctoral dissertation. Both two datasets were created and recorded by Dr. SantaBarbara. Both datasets were handed to us but were combined to use for our project. To combine them, a lot of preprocessing had to be accomplished. First, both datasets had to be aligned. This meant that each variable had to have the same name, data type, and range of values. This made it so that each variable stood for the same thing. The first step of preprocessing was to remove all underscores and spaces, and then change all characters to be strictly uppercase. This allowed variables like “su\_9” in one dataset to be paired with “SU9” in the other. Then we matched other columns that could not be automated. For example, in one dataset there was “height” and “weight” and in the other dataset, we had to rename “lbs” to “height” and “ht20” to “weight”. The third step included dropping columns that did not match. The dataset used for the doctoral dissertation was more complex than the other, so we took out multiple columns from that dataset. At this point, after accounting for hundreds of columns, the two datasets were combined. Even after the dataset was combined, we still had to change some of the values so that it portrayed the exact same information. An example of this is the Income variable. In one dataset, it was on a scale and in the other, it was continuous. So when it was combined, we had to change all the values to be equal. Finally, we had a dataset that included 54 patients with and without MD. A subset was later created which removed all rows where participants did not have muscle dysmorphia, as they were not our target, and had large portions of crucial missing data. That subset had a total of 35 rows and 159 columns. Each row represents a patient with muscle dysmorphia, and each column represents a question from the survey.

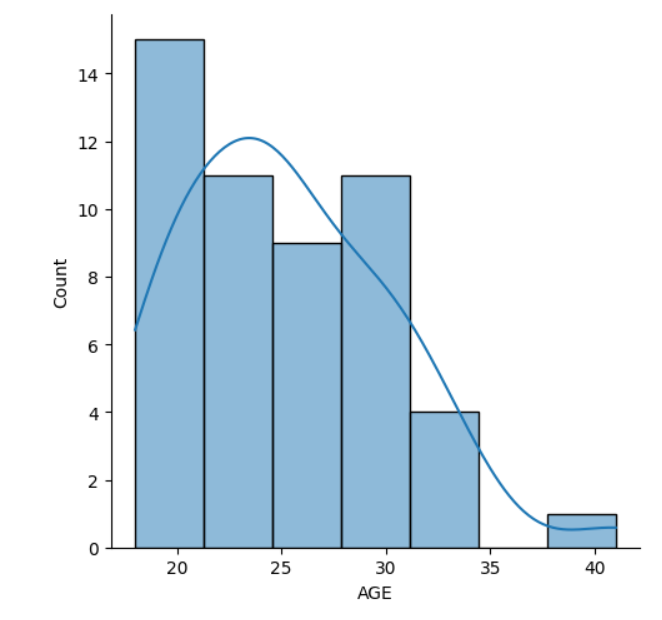
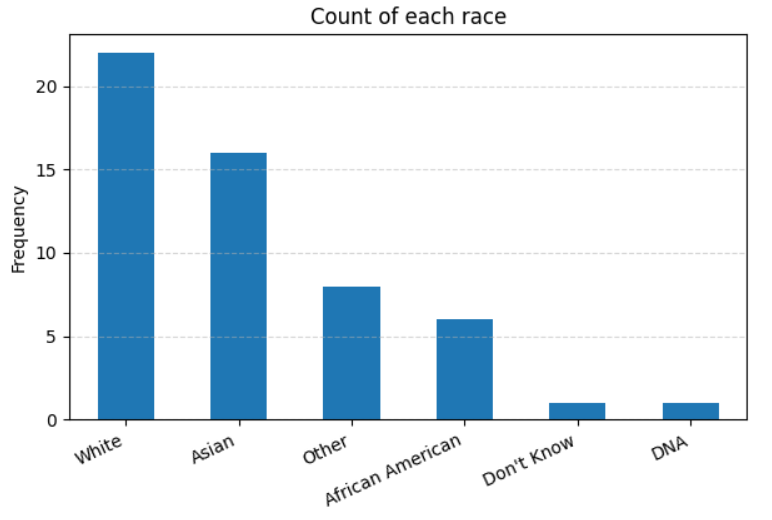
Each of our aims had different methods to answer them. We used both R and Python to complete this because the two languages have strong statistical methods built into them. R is built on the foundations of statistical computing and Python has packages like pandas and sklearn, which deal with wrangling data and creating models. For our first aim, we characterized resistance training behavior by using df.describe, which creates a statistical summary table for each variable. Along with the summary table, we also created many different bar graph visualizations to get a better understanding of the people that are included in our data set. To characterize the mental health issues, we met with our client and used the scores that were compiled at the end of the questionnaire. The surveys for anxiety and depression were the only ones that had a clinical threshold that was over 40 and 10 respectively. So if a patient scores over those numbers, they could be classified as clinically anxious or depressed. The other questionnaires did not have thresholds, but in general, the higher the scores, the worse mental health they had.

For our second aim, we did further preprocessing, dropping all the columns, except the ones that involve training (“days per week”, “intensity”, and “duration”), as well as the mental health questions. PCA was then performed on each of the seven surveys: MDI (Muscle Dysmorphia), RS (Nutrition), BDS (Bodybuilding Dependance Scale or Body Image), SPAS (Social Physique Anxiety Scale), CESD (Depression), STAI (Anxiety), and AUDIT (Alcohol). After the dimensions were reduced, they were added to the training variables and a linear regression model was run using intensity as the response variable and the results from each PCA were used as the explanatory variables. A forward stepwise regression was run on all the dimensions to get the best outcome with the data. Finally, a regression tree was used to get a different look at the data in a nonlinear fashion. The reason why these predictive models were used to answer this aim, is to see if there is direct correlation between the training and mental health variables.

For the third aim, ANOVA, and logistic regression were both used to explore any associations between mental health and substance or supplement use. These models were used due to the nature of how the substance and supplement columns were recorded. Since everything was on a binary scale, it was easier to score them and analyze their variance. Using the chi square function from the scipy package and a correlation matrix, we were able to see if there was any correlation between using marijuana, creatine, or protein and being clinically depressed. After we used ANOVA on all the supplements and drugs, we then printed a summary of the results to look at the mean squared errors and f values. The surveys with clinical cut offs (CESD, and STAI) were also converted to a binary target column which made logistic regression possible. This used the CESD score, or the depression survey score, as a target variable and the different supplements and substances as the explanatory variables. We used the function from the sklearn package to run the logistic regression.

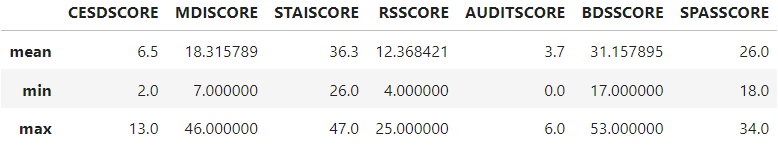
**Results:**

For our first aim, when utilizing the df.describe() command, the resulting output furnishes a statistical summary of our dataset’s numerical columns. It presents essential statistics like count, mean, standard deviation, minimum, maximum, and quartile values. Analyzing the describe() output helps in comprehending the central tendencies, spread, and distribution of the data. It has broken up so it will be easier to comprehend. The first two figures are bar graphs that show us the distribution of patients in our data set. From these two graphs we can see that we have males that are 20 - 30 years old and a petty even split between white and non-white races.

**Figure 1 (above)**: Race distribution graph.

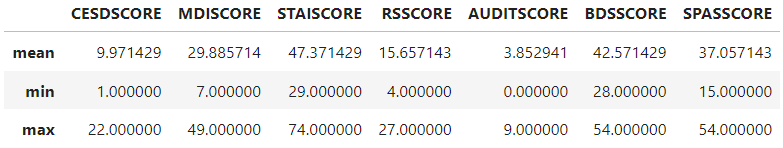
**Figure 2 (left)**: Age distribution graph.

The other tables from aim one that describe our data set show the two groups, both of patients with and without MD. Figure 3 and 4 show the mental scores on each survey, and Figures 5 and 6 show the training behaviors of the two sets.



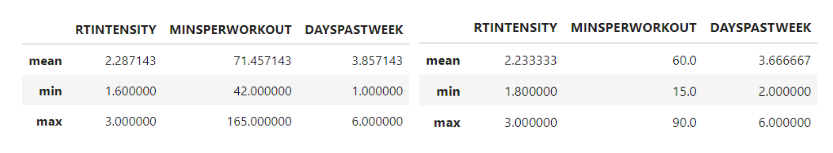
**Figure 3 (above)**: Table of mental health scores from patients without MD.

**Figure 4 (below)**: Table of mental health scores from patients with MD.



Note how every score on the table that shows patients with MD is higher than the scores in the table that shows patients without MD. From this we can say that the worse the mental health is of

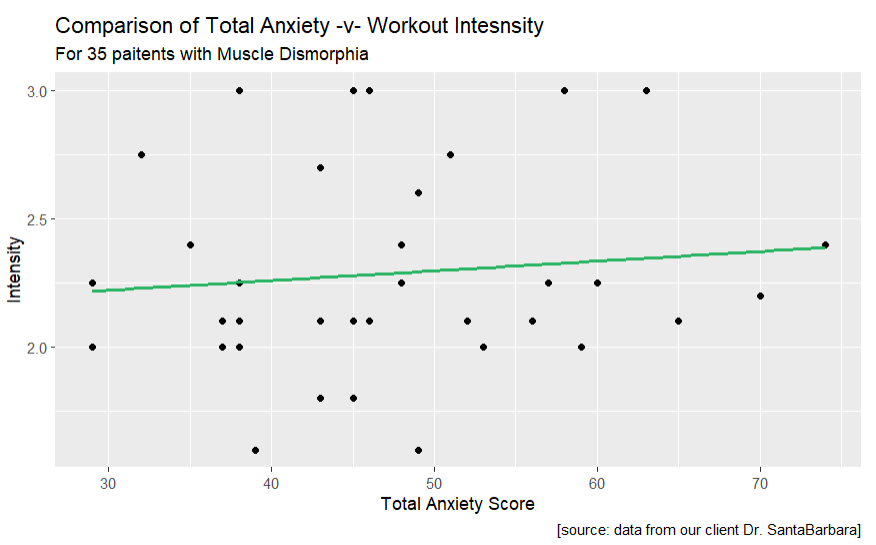
a patient the higher chance they are to have muscle dysmorphia. Next is a table that compares intensity, how long they worked out for and how many days per week they workout for.



**Figure 5 (above)**: Tables showing training behaviors between patients with MD (left) and patients without (right).

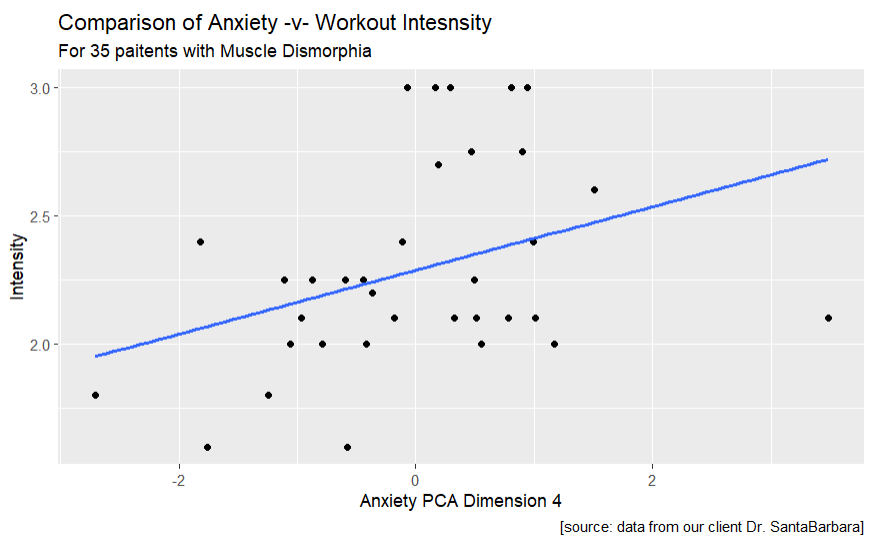
People with muscle dysmorphia have similar training habits to people without muscle dysmorphia, excluding workout length. The typical patient with muscle dysmorphia works out about 3.8 days a week, for 71.45 minutes per session, at an intensity of roughly 2.28.

To see whether worse mental health, including symptoms of muscle dysmorphia, are associated with more frequent and intense resistance training behaviors, we decided to use a linear regression model on the total scores of each survey. This produced the graph from figure 6. From this graph not much correlation is shown with the line of best fit. To get a more accurate result, we wanted to change the scores to variables that held data for each individual question, but there were too many variables to use in our model.



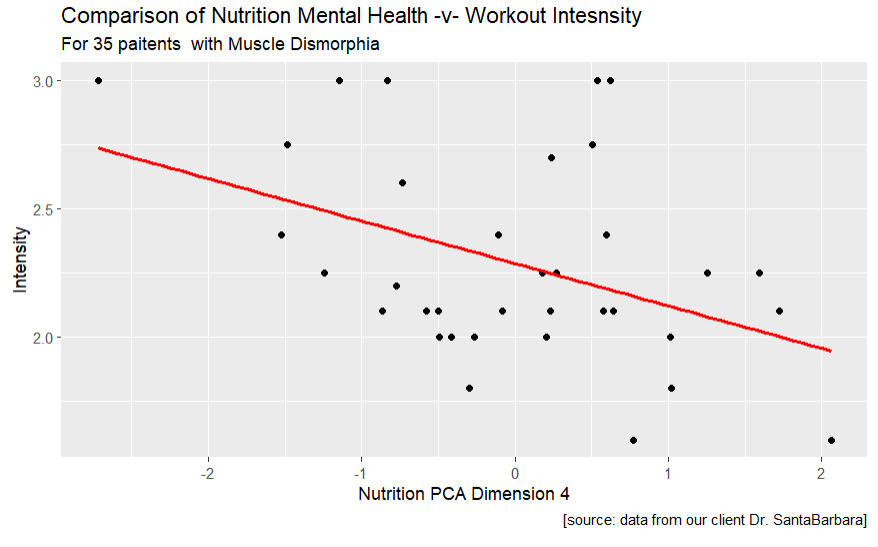
**Figure 6 (left)**: Linear regression model that shows intensity and the anxiety score as variables.

To solve this problem, we decided to use principal component analysis. Principal component analysis or PCA reduces dimensions based on variability. This way we can keep the most information and have the least number of variables. PCA was performed on each survey and the most statistically significant variables were taken to make the second graph. The next graph was made from the most significant variables created by PCA.

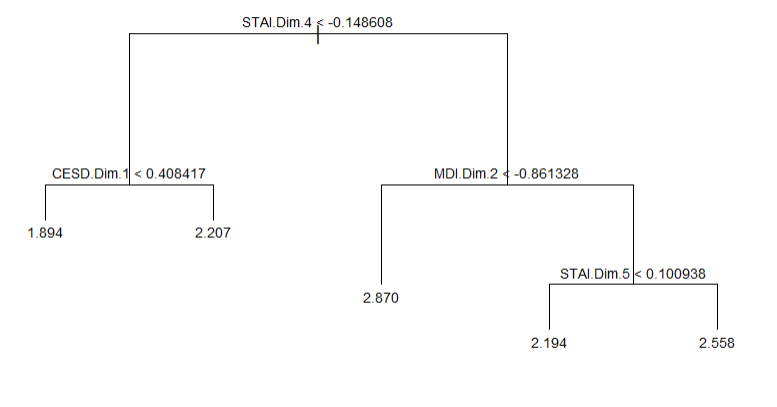


**Figure 7 (left)**: Linear regression model that shows intensity and the fourth dimension for anxiety as the variables.

While this is better, and we start to see a correlation it is still not as accurate as we wanted. To see if there was anything else we could do to make a stronger model we wanted to try one more thing. The last thing we did was to make a forward stepwise regression model using intensity and trying out combinations between all the PCA dimensions.

**Figure 8 (left):** Forward linear regression model with the best combination of variables.

While this is still not very accurate in its ability to predict intensity, it is the best linear model we could make. In another approach to see if we could answer this aim in a non linear way, we created a regression tree. The regression tree below is the best tree model that we were able to produce. At each split is the number the model decided to split the data on and in the root of each branch shows the perceived intensity the patient will work out at.

**Figure 8 (left):** Regression tree that predicts intensity.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Linear on Scores | Linear on PCA | Forward Stepwise | Regression Tree |
| R^2 | 0.1363 | 0.2141 | 0.4712 | NA |
| Mean^2 Error | 0.3748 | 0.3534 | 0.2899 | 0.0543 |

**Figure 10 (above):** Mean squared error and proportion of variance table.

From all these models we have taken out their MSE and R^2 which are measures that show how well a model can predict the data it is given. For R^2 we want the numbers to be as close to one as possible and for MSE we want the numbers to be as close to zero as possible. In both models the variables for the fourth dimension of nutrition and anxiety occur many times. The questions in those dimensions that have the most variance is “I feel that difficulties are piling up and I cannot overcome them?” and “How many pounds over your desired weight were you at your max?”. While these models are too weak to draw a definite conclusion it is something interesting to take note of. Overall, these results show that the intensity of the workout is not associated with worse mental health or having symptoms of muscle dysmorphia.

Throughout aim 3, we ran various Anova models to determine if any substance and supplement use within the month would contribute to a decrease in mental health. Through our data, we had various mental health indexes, including CESD, STAI, MDI, and others.

**Figure 11 (above):** Result table of R-Squared values for all models and AIC for ANOVA.

All of these results became inconclusive due to the small sample size and simply not having a strong correlation, returning R^2 values less than 0.35 for all mental health indexes. Many of these do not define a medical cutoff; however, CESD and STAI did. CESD’s medical cutoff of displaying symptoms of depression is 10 and greater, for STAI, the cutoff for displaying anxiety is greater than 40. It is important to note that these do not follow the same scaling. I checked for individuals above and below this value. This left me with a binary variable that I could use to run logistic regression to determine which group they were placed in, this resulted in R-Squared scores around 0.1 for both of these. Overall, we could not pull out any significant results within our dataset when trying to compare substance and supplement use to mental health. This does not imply that there are no connections between the two, but potentially that something in our data did not allow for anything to be conclusive, whether that be a small sample size, or a survey population that did not represent the whole population well.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ANOVA | | | | | | | Logistic Regression | |
| MH  Survey | CESD | STAI | MDI | RS | BDS | SPAS | AUDIT | CESD - Status | STAI - Status |
| R^2 | 0.180 | 0.208 | 0.099 | 0.163 | 0.294 | 0.189 | 0.326 | 0.082 | 0.121 |
| AIC | 219 | 270 | 266 | 228 | 237 | 262 | 154 | N/A | N/A |

**Conclusion and Discussion:**

Body dysmorphia represents a complex and distressing psychological condition, characterized by a persistent preoccupation with perceived flaws in physical appearance. This disorder not only affects one's mental health but also impacts daily functioning, relationships, and overall well-being. Addressing body dysmorphia requires a multifaceted approach involving therapeutic interventions, such as therapy, and potentially medication, alongside a supportive environment to foster body positivity. Throughout our objectives, we have seen that results have typically shown not to be statistically significant. This does not mean that in a larger population of people these connections between training behaviors, substance use, and mental health do not exist. These results tell us that our sample did not have meaningful connections because statistical significance can be very difficult to find in datasets of small numbers. Some limitations of this dataset were due to the small sample size, and the way the variables were collected. An ideal dataset would have only a few columns and thousands of rows. In the dataset used we had the opposite problem. The hundreds of columns made the variance to spread out, which is why we tried to use dimension reduction techniques. In the future, a larger and more diverse sample size may have come in handy to come to more statistically significant conclusions. We could have had better models due to there being more data to work with. If we had a good split of men with and without MD, then we could use more advanced supervised learning techniques to see if it would have yielded better results. Some better variables would be to have all the scores with clinical thresholds, as well as more other exercise metrics, like an observed intensity rather than perceived intensity. Most of the variables seemed like they would be useful; they just needed more data in them. If all of these things were in a new data set moving forward we are confident that we would produce better results. This is important because understanding the intricacies of this disorder is crucial in developing tailored strategies to help individuals navigate and overcome its detrimental effects on their lives.

**Resources:**

***References***

Corazza O, Simonato P, Demetrovics Z, Mooney R, van de Ven K, Roman-Urrestarazu A, Rácmolnár L, De Luca I, Cinosi E, Santacroce R, Marini M, Wellsted D, Sullivan K, Bersani G, Martinotti G. The emergence of Exercise Addiction, Body Dysmorphic Disorder, and other image-related psychopathological correlates in fitness settings: A cross sectional study. PLoS One. 2019 Apr 3;14(4):e0213060. doi: 10.1371/journal.pone.0213060. PMID: 30943200; PMCID: PMC6447162.

Leone JE, Sedory EJ, Gray KA. Recognition and treatment of muscle dysmorphia and related body image disorders. J Athl Train. 2005 Oct-Dec;40(4):352-9. PMID: 16404458; PMCID: PMC1323298.

SantaBarbara, N. J., Nosrat, S., Whitworth, J. W., & Ciccolo, J. T. (2020). Acute psychological effects of resistance exercise in men with symptoms of muscle dysmorphia: A pilot study. *Journal of Strength and Conditioning Research*, *37*(2), 277–283. https://doi.org/10.1519/jsc.0000000000003615

**Code Appendix:**

**Aim #1:**

*#Load in Cleaned Data*

*#Models*

*import pandas as pd*

*import numpy as np*

*import seaborn as sns*

*import matplotlib.pyplot as plt*

*df\_all = pd.read\_csv('combined\_dataset\_ALL.csv')*

*df = pd.read\_csv('combined\_dataset.csv')*

*df\_all.rename(columns={'AGE2':'AGE'}, inplace=True)*

*df\_all[['AGE','NUMBEROFROOMMATES']][df\_all['GroupFinal'] == 1].describe().iloc[1:,:]*

*df\_all[['AGE','NUMBEROFROOMMATES']][df\_all['GroupFinal'] == 0].describe().iloc[1:,:]*

*df\_all[['NUMBEROFROOMMATES','AGE']].describe().iloc[1:,:]*

*df\_all[['CESDSCORE','MDISCORE',"STAISCORE","RSSCORE","AUDITSCORE","BDSSCORE","SPASSCORE"]][df\_all['GroupFinal'] == 1].describe().iloc[[1,3,7],:]*

*df\_all[['CESDSCORE','MDISCORE',"STAISCORE","RSSCORE","AUDITSCORE","BDSSCORE","SPASSCORE"]][df\_all['GroupFinal'] == 0].describe().iloc[[1,3,7],:]*

*df\_all.loc[df\_all['RACE'] == 1, 'RACE'] = "Native American"*

*df\_all.loc[df\_all['RACE'] == 2, 'RACE'] = "Asian"*

*df\_all.loc[df\_all['RACE'] == 3, 'RACE'] = "Pacific Islander"*

*df\_all.loc[df\_all['RACE'] == 4, 'RACE'] = "African American"*

*df\_all.loc[df\_all['RACE'] == 5, 'RACE'] = "White"*

*df\_all.loc[df\_all['RACE'] == 6, 'RACE'] = "Other"*

*df\_all.loc[df\_all['RACE'] == 7, 'RACE'] = "Don't Know"*

*df\_all.loc[df\_all['RACE'] == 999, 'RACE'] = "DNA"*

*df\_all['RACE'].value\_counts()*

*ax = df\_all['RACE'].value\_counts().plot(kind='bar', figsize=(7,4), title="Count of each race")*

*ax.set\_xlabel("")*

*ax.set\_ylabel("Frequency")*

*plt.xticks(rotation = 25, ha = "right")*

*plt.grid(visible=True, which='major',axis='y', alpha = 0.5, linestyle = '--')*

*plt.show()*

*sns.displot(data=df\_all, x = 'AGE', kde = True)*

*df\_all.loc[df\_all['SEXUALORIENTATION'] == 1, 'SEXUALORIENTATION'] = "Heterosexual"*

*df\_all.loc[df\_all['SEXUALORIENTATION'] == 2, 'SEXUALORIENTATION'] = "Homosexual"*

*df\_all.loc[df\_all['SEXUALORIENTATION'] == 3, 'SEXUALORIENTATION'] = "Bisexual"*

*df\_all[["RTINTENSITY","MINSPERWORKOUT","DAYSPASTWEEK"]][df\_all['GroupFinal'] == 1].describe().iloc[[1,3,7],:]*

*df\_all[["RTINTENSITY","MINSPERWORKOUT","DAYSPASTWEEK"]][df\_all['GroupFinal'] == 0].describe().iloc[[1,3,7],:]*

***Aim 2***

***Ryan Canfield***

***2023-11-19***

***This is the code for the preliminary results for completing Aim 2.***

***Other code will be added later which will include the final model with the highest accuracy.***

***By Ryan Canfield***

***DSE4900 Data Science Capstone***

***Dr. Santa Barabara Group (Ryan Canfield, Pat Norcross, and Alex Buterra)***

***11/14/23***

*# Libraries and Packages*

*library(factoextra) # Used for PCA*

*library(FactoMineR)*

*# Reading in the data set and previewing the data*

*df <- read.csv("combined\_dataset.csv")*

*## Preprocessing*

*# Dropping columns*

*df2 = subset(df, select = -c(X, index, GROUP, ELIGIBILITYRESCREENQ1, ELIGIBILITYRESCREENQ2, ELIGIBILITYRESCREENQ3, ELIGIBILITYRESCREENQ4, GENDER2, RACEOTHER, LIVING, NUMBEROFROOMMATES, INCOME, EMPLOYMENT, HEALTHHIX23, DRUGADDICTION, ALCOHOLADDICTION, SU15))*

*## THIS WAS DONE BUT AFTER DOING WORK THIS IS NOT NECCSARY BECAUSE THESE COLUMNS ARE NOT USED.*

*# Binary column race into 1 for white all other for 0*

*# living 0 alone 1 other people*

*df2$RACE <- ifelse(df2$RACE != 5, 0, 1)*

*df2$EDUCATION <- ifelse(df2$EDUCATION <= 5, 0, 1)*

*# Fixes the Na's found in the age column*

*df2$AGE2[is.na(df2$AGE2)] <- round(median(df2$AGE2, na.rm = TRUE))*

*# Keep HISPANICLATINO column(if they are part or not )*

*# Preliminary model to see if we can use total scores to predict intensity*

*Starting.Intensity <- lm(RTINTENSITY ~ MDISCORE + RSSCORE + BDSSCORE + SPASSCORE + CESDSCORE + STAISCORE + AUDITSCORE, data = df2)*

*summary.lm(Starting.Intensity)*

*# Plotting the forward stepwise linear regression model based off of the most significant variable*

*g4 <- ggplot(data = df2, aes(x = STAISCORE, y = RTINTENSITY)) +*

*geom\_point () +*

*labs(title = "Comparison of Total Anxiety -v- Workout Intesnsity",*

*subtitle ="For 35 paitents with Muscle Dismorphia",*

*x = "Total Anxiety Score",*

*y = "Intensity",*

*caption = "[source: data from our client Dr. SantaBarbara]") +*

*geom\_smooth(method = "lm", se = 0, colour = "#28B463")*

*g4*

*## PCA for MDI*

*# Just getting the MDI variables MDI1 - MDIScore*

*subsetMDI <- df2[c(22:38)]*

*set.seed(310) #I use this so that I get the same answer each time*

*# Preforming PCA*

*res.pca.MDI <- prcomp(subsetMDI[c(1:16)], scale = TRUE) # leave off the last variable, which is the response variable "quality"*

*# Eigenvalues*

*eig.val.MDI <- get\_eigenvalue(res.pca.MDI)*

*eig.val.MDI*

*# Scree plot to look at which dimensions to keep*

*fviz\_eig(res.pca.MDI)*

*# Looking at those dimensions to extract variables with high variance*

*res.var.MDI <- get\_pca\_ind(res.pca.MDI)*

*stored <- res.var.MDI$coord[, 1:3]*

*# ^^^ Store this into a data frame*

*MDIvari <- data.frame(stored)*

*MDIvari*

*# MDI6 MDI8 MDI9 MDI10*

*# MDI2 MDI11 MDI13*

*# MDI4*

*# This lets you look at what variables are used or questions ask*

*which.var.MDI <- get\_pca\_var(res.pca.MDI)*

*which.var.MDI$coord*

*### This is the same code as abov just replaced with different survey data*

*## Not the different subsets*

*## PCA for RS*

*subsetRS <- df2[c(39:49)]*

*set.seed(310)*

*res.pca.RS <- prcomp(subsetRS[c(1:10)], scale = TRUE) # leave off the last variable, which is the response variable "quality"*

*eig.val.RS <- get\_eigenvalue(res.pca.RS)*

*eig.val.RS*

*fviz\_eig(res.pca.RS)*

*res.var.RS <- get\_pca\_ind(res.pca.RS)*

*RSvari <- data.frame(res.var.RS$coord[, 1:4])*

*# RS7 RS8*

*# RS2 RS4*

*# RS1*

*# This lets you look at what variables are used or questions ask*

*which.var.RS <- get\_pca\_var(res.pca.RS)*

*which.var.RS$coord*

*## PCA for BDS*

*subsetBDS <- df2[c(50:59)]*

*set.seed(310)*

*res.pca.BDS <- prcomp(subsetBDS[c(1:9)], scale = TRUE) # leave off the last variable, which is the response variable "quality"*

*eig.val.BDS <- get\_eigenvalue(res.pca.BDS)*

*eig.val.BDS*

*fviz\_eig(res.pca.BDS)*

*res.var.BDS <- get\_pca\_ind(res.pca.BDS)*

*BDSvari <- data.frame(res.var.BDS$coord[, 1:4])*

*# BDS4 BDS5 BDS7 BDS3*

*# Results for Variables - Prints component matrix*

*# This lets you look at what variables are used or questions ask*

*which.var.BDS <- get\_pca\_var(res.pca.BDS)*

*which.var.BDS$coord*

*# PCA for SPAS*

*subsetSPAS <- df2[c(60:72)]*

*set.seed(310)*

*res.pca.SPAS <- prcomp(subsetSPAS[c(1:12)], scale = TRUE) # leave off the last variable, which is the response variable "quality"*

*eig.val.SPAS <- get\_eigenvalue(res.pca.SPAS)*

*eig.val.SPAS*

*fviz\_eig(res.pca.SPAS)*

*res.var.SPAS <- get\_pca\_ind(res.pca.SPAS)*

*SPASvari <- data.frame(res.var.SPAS$coord[, 1:4])*

*# SPAS1 SPAS2 SPAS4 SPAS6 SPAS8*

*# This lets you look at what variables are used or questions ask*

*which.var.SPAS <- get\_pca\_var(res.pca.SPAS)*

*which.var.SPAS$coord*

*# PCA for CESD*

*subsetCESD <- df2[c(73:83)]*

*set.seed(310)*

*res.pca.CESD <- prcomp(subsetCESD[c(1:10)], scale = TRUE) # leave off the last variable, which is the response variable "quality"*

*eig.val.CESD <- get\_eigenvalue(res.pca.CESD)*

*eig.val.CESD*

*fviz\_eig(res.pca.CESD)*

*res.var.CESD <- get\_pca\_ind(res.pca.CESD)*

*CESDvari <- data.frame(res.var.CESD$coord[, 1:3])*

*# CESD3 CESD9 CESD6*

*# This lets you look at what variables are used or questions ask*

*which.var.CESD <- get\_pca\_var(res.pca.CESD)*

*which.var.CESD$coord*

*# PCA for STAI*

*subsetSTAI <- df2[c(84:104)]*

*set.seed(310)*

*res.pca.STAI <- prcomp(subsetSTAI[c(1:20)], scale = TRUE) # leave off the last variable, which is the response variable "quality"*

*eig.val.STAI <- get\_eigenvalue(res.pca.STAI)*

*eig.val.STAI*

*fviz\_eig(res.pca.STAI)*

*res.var.STAI <- get\_pca\_ind(res.pca.STAI)*

*STAIvari <- data.frame(res.var.STAI$coord[, 1:5])*

*# STAI1 STAI4 STAI9 STAI10 STAI12 STAI16 STAI19*

*# This lets you look at what variables are used or questions ask*

*which.var.STAI <- get\_pca\_var(res.pca.STAI)*

*which.var.STAI$coord*

*set.seed(310)*

*# PCA for AUDIT*

*subsetAUDIT <- df2[c(115:118)]*

*subsetAUDIT <- replace(subsetAUDIT, is.na(subsetAUDIT), 0)*

*res.pca.AUDIT <- prcomp(subsetAUDIT[c(1:3)], scale = TRUE) # leave off the last variable, which is the response variable "quality"*

*eig.val.AUDIT <- get\_eigenvalue(res.pca.AUDIT)*

*eig.val.AUDIT*

*fviz\_eig(res.pca.AUDIT)*

*res.var.AUDIT <- get\_pca\_ind(res.pca.AUDIT)*

*AUDITvari <- data.frame(res.var.AUDIT$coord[, 1:1])*

*# AUDIT1 AUDIT2 AUDIT3*

*# This lets you look at what variables are used or questions ask*

*which.var.AUDIT <- get\_pca\_var(res.pca.AUDIT)*

*which.var.AUDIT$coord*

*library(dplyr)*

*# Renaming the columns to tell the dimensions apart.*

*MDIvari <- MDIvari %>% rename("MDI.Dim.1" = "Dim.1", "MDI.Dim.2" = "Dim.2", "MDI.Dim.3" = "Dim.3")*

*RSvari <- RSvari %>% rename("RS.Dim.1" = "Dim.1", "RS.Dim.2" = "Dim.2", "RS.Dim.3" = "Dim.3", "RS.Dim.4" = "Dim.4")*

*BDSvari <- BDSvari %>% rename("BDS.Dim.1" = "Dim.1", "BDS.Dim.2" = "Dim.2", "BDS.Dim.3" = "Dim.3", "BDS.Dim.4" = "Dim.4")*

*SPASvari <- SPASvari %>% rename("SPAS.Dim.1" = "Dim.1", "SPAS.Dim.2" = "Dim.2", "SPAS.Dim.3" = "Dim.3", "SPAS.Dim.4" = "Dim.4",)*

*CESDvari <- CESDvari %>% rename("CESD.Dim.1" = "Dim.1", "CESD.Dim.2" = "Dim.2", "CESD.Dim.3" = "Dim.3")*

*STAIvari <- STAIvari %>% rename("STAI.Dim.1" = "Dim.1", "STAI.Dim.2" = "Dim.2", "STAI.Dim.3" = "Dim.3", "STAI.Dim.4" = "Dim.4", "STAI.Dim.5" = "Dim.5")*

*AUDITvari <- AUDITvari %>% rename("AUDIT.Dim.1" = "res.var.AUDIT.coord...1.1.")*

*# Forming a new dataset with the new PCA dimensions and the workout variables.*

*questionaire.Variables <- cbind(MDIvari, RSvari, BDSvari, SPASvari, CESDvari, STAIvari, AUDITvari)*

*df2.1 <- df2[ , 19:21]*

*# Here are the scores if they are needed.*

*#questionaire.Scores <- cbind(subsetMDI[c(17)], subsetRS[c(11)], subsetBDS[c(10)], subsetSPAS[c(13)], subsetCESD[c(11)], subsetSTAI[c(21)], subsetAUDIT[c(4)])*

*#questionaire.Scores*

*# This is the dataset I am using for the analysis.*

*df3 <- cbind(df2.1, questionaire.Variables)*

*head(df3)*

*Linear Regression*

*# Preforming linear regression on all the surveys to see if any dimensions are significant.*

*MDI.Intensity <- lm(RTINTENSITY ~ MDI.Dim.1 + MDI.Dim.2 + MDI.Dim.3, data = df3)*

*summary.lm(MDI.Intensity)*

*RS.Intensity <- lm(RTINTENSITY ~ RS.Dim.1 + RS.Dim.2 + RS.Dim.3 + RS.Dim.4, data = df3)*

*summary.lm(RS.Intensity)*

*BDS.Intensity <- lm(RTINTENSITY ~ BDS.Dim.1 + BDS.Dim.2 + BDS.Dim.3 + BDS.Dim.4, data = df3)*

*summary.lm(BDS.Intensity)*

*SPAS.Intensity <- lm(RTINTENSITY ~ SPAS.Dim.1 + SPAS.Dim.2 + SPAS.Dim.3 + SPAS.Dim.4, data = df3)*

*summary.lm(SPAS.Intensity)*

*CESD.Intensity <- lm(RTINTENSITY ~ CESD.Dim.1 + CESD.Dim.2 + CESD.Dim.3, data = df3)*

*summary.lm(CESD.Intensity)*

*STAI.Intensity <- lm(RTINTENSITY ~ STAI.Dim.1 + STAI.Dim.2 + STAI.Dim.3 + STAI.Dim.4 + STAI.Dim.5, data = df3)*

*summary.lm(STAI.Intensity)*

*AUDIT.Intensity <- lm(RTINTENSITY ~ AUDIT.Dim.1, data = df3)*

*summary.lm(AUDIT.Intensity)*

*# Significant variables*

*# MDI 2*

*# RS 2 4*

*# STAI 4*

*# Final linear regression model taken from the significant variables above*

*FINAL.Intensity <- lm(RTINTENSITY ~ MDI.Dim.2 + RS.Dim.4 + STAI.Dim.4, data = df3)*

*summary.lm(FINAL.Intensity)*

*## Do scores first an see if anything is significant*

*## Then go into PCA*

*## TALK ABOUT WHAT QUESTIONS GO INTO THE THREE*

*g1 <- ggplot(data = df3, aes(x = STAI.Dim.4, y = RTINTENSITY)) + # create cty-v-hwy graph from data set mpg*

*geom\_point () + # crate scatterplot with points jittered*

*labs(title = "Comparison of Anxiety -v- Workout Intesnsity", # create title above graph*

*subtitle ="For 35 paitents with Muscle Dismorphia", # create subtitle below title*

*x = "Anxiety PCA Dimension 4", # label x-axis*

*y = "Intensity", # label y-axis*

*caption = "[source: data from our client Dr. SantaBarbara]") # insert captionbelow graph # display graph*

*g2 <- g1 + # create graph g2 starting with graph g1*

*geom\_smooth(method = "lm", se = 0) # add linear regression line with standard error envelope*

*g2 # display graph*

*#To run stepwise regrression:*

*#First, let us define the null (intercept-only) model. We need this to build our forward stepwise regression:*

*intercept\_only1 <- lm(DAYSPASTWEEK ~ 1, data = df3)*

*summary.lm(intercept\_only1)*

*#Next, let us define the model with all explanatory variables included. We need this both for forward and backward stepwise regression:*

*all1 <- lm(RTINTENSITY ~ ., data = df3)*

*summary.lm(all1)*

*#If we want to see the output of the forward stepwise regression, we can use this command:*

*forward$anova*

*# Note that forward$coefficients does not give us the full summary like summary.lm does. If we want that, we need to fit the model with these predictors and print out the output:*

*bestforward <- lm(RTINTENSITY ~ RS.Dim.4 + MINSPERWORKOUT + SPAS.Dim.3 + STAI.Dim.5 + STAI.Dim.4 + STAI.Dim.2, data = df3)*

*summary.lm(bestforward)*

*# Plotting the forward stepwise linear regression model based off of the most significant variable*

*g3 <- ggplot(data = df3, aes(x = RS.Dim.4, y = RTINTENSITY)) +*

*geom\_point () +*

*labs(title = "Comparison of Nutrition Mental Health -v- Workout Intesnsity",*

*subtitle ="For 35 paitents with Muscle Dismorphia",*

*x = "Nutrition PCA Dimension 4",*

*y = "Intensity",*

*caption = "[source: data from our client Dr. SantaBarbara]") +*

*geom\_smooth(method = "lm", se = 0, colour = "red")*

*g3*

*## `geom\_smooth()` using formula 'y ~ x'*

*Make regression trees*

*library(gbm)*

*library(tree)*

*library(randomForest)*

*library(tidyverse)*

*head(df3)*

*# Bottom of the node = the level of intensity*

*tree\_model <- tree(RTINTENSITY ~ ., df3)*

*plot(tree\_model)*

*text(tree\_model, pretty = 0, cex = 0.7)*

*summary(tree\_model)*

*pred <- predict(tree\_model, df3)*

*mean((pred - df3$RTINTENSITY)^2)*

*#calculate residual standard error*

*sqrt(deviance(tree\_model)/df.residual(tree\_model))*

*set.seed(310)*

*cv\_tree\_model <- cv.tree(tree\_model, K = 10)*

*data.frame(n\_leaves = cv\_tree\_model$size,*

*CV\_RSS = cv\_tree\_model$dev) %>%*

*mutate(min\_CV\_RSS = as.numeric(min(CV\_RSS) == CV\_RSS)) %>%*

*ggplot(aes(x = n\_leaves, y = CV\_RSS)) +*

*geom\_line(col = "grey55") +*

*geom\_point(size = 2, aes(col = factor(min\_CV\_RSS))) +*

*scale\_x\_continuous(breaks = seq(1, 17, 2)) +*

*scale\_y\_continuous(labels = scales::comma\_format()) +*

*scale\_color\_manual(values = c("deepskyblue3", "green")) +*

*theme(legend.position = "none") +*

*labs(title = "Muscle Dismorphia Dataset - Regression Tree",*

*subtitle = "Selecting the complexity parameter with cross-validation",*

*x = "Terminal Nodes",*

*y = "CV RSS")*

*# Seeing in pruning the tree helps.*

*pruned\_tree\_model <- prune.tree(tree\_model, best = 2)*

*cv.pred <- predict(pruned\_tree\_model, df3)*

*mean((cv.pred - df3$RTINTENSITY)^2)*

***Aim #3***

*#Load in Data*

*import pandas as pd*

*import numpy as np*

*df\_all = pd.read\_csv('combined\_dataset\_ALL.csv')*

*df = pd.read\_csv('combined\_dataset.csv')*

*#Aim 3 Models*

*df.describe()*

*from sklearn.linear\_model import LinearRegression*

*#Create SU SCORE*

*#ANOVA*

*X = df[[col for col in df if col.startswith('SU')]]*

*X.drop(['SU7','SU15'],axis=1,inplace=True)*

*X = X.to\_numpy()*

*y = df[[col for col in df if col.startswith('MDISCORE')]]*

*y = y.to\_numpy().reshape(len(y),)*

*lm = LinearRegression()*

*lm.fit(X,y)*

*y\_pred = lm.predict(X)*

*lm.score(X,y)*

*Lm.coef\_*

*import matplotlib.pyplot as plt*

*plt.scatter(y, y\_pred)*

*plt.axline([0,0],[1,1],c='r')*

*plt.xlabel("Actual MDI Score")*

*plt.ylabel("Predicted MDI Score")*

*from sklearn.metrics import confusion\_matrix, ConfusionMatrixDisplay*

*plt.hist(data=df, x='CESDSCORE')*

*X = df[[col for col in df if col.startswith('SU')]]*

*X.drop(['SU7','SU15'],axis=1,inplace=True)*

*X = X.to\_numpy()*

*y = df[[col for col in df if col.startswith('CESDSCORE')]]*

*y = y.to\_numpy().reshape(len(y),)*

*lm = LinearRegression()*

*lm.fit(X,y)*

*y\_pred = lm.predict(X)*

*lm.score(X,y)*

*plt.scatter(y, y\_pred)*

*plt.axline([0,0],[1,1],c='r')*

*plt.xlabel("Actual CESD Score")*

*plt.ylabel("Predicted CESD Score")*

*#Create a Binary feature*

*df.STAISCORE.describe()*

*#CESD - above 10*

*#STAI - above 40 - anxiety*

*#MDI - RS SPAS - BDS - No cut offs*

*df['MHSCORE'] = df['MDISCORE'] + df['CESDSCORE'] + df['STAISCORE'] + df['SPASSCORE']*

*import statsmodels.api as sm*

*from statsmodels.tools import add\_constant*

*X = df[[col for col in df if col.startswith('SU')]]*

*X.drop(['SU7','SU15'],axis=1,inplace=True)*

*X = X.to\_numpy()*

*y = df[[col for col in df if col.startswith('MHSCORE')]]*

*y = y.to\_numpy().reshape(len(y),)*

*lm = LinearRegression()*

*lm.fit(X,y)*

*y\_pred = lm.predict(X)*

*lm.score(X,y)*

*X = add\_constant(X)*

*model\_statsmodels = sm.OLS(y,X).fit()*

*print(model\_statsmodels.summary())*

*##changing columns*

*#df['Protein\_overmonth'] = df['SU111'] + df['SU112'] + df['SU113']*

*df['Protein\_lastmonth'] = df['SU114'] + df['SU115'] + df['SU116']*

*#df['Creatine\_overmonth'] = df['SU121'] + df['SU122'] + df['SU123']*

*df['Creatine\_lastmonth'] = df['SU124'] + df['SU125'] + df['SU126']*

*#df['Diaretics\_overmonth'] = df['SU131'] + df['SU132'] + df['SU133']*

*df['Diaretics\_lastmonth'] = df['SU134'] + df['SU135'] + df['SU136']*

*#df['Ephedra\_overmonth'] = df['SU141'] + df['SU142'] + df['SU143']*

*df['Ephedra\_lastmonth'] = df['SU144'] + df['SU145'] + df['SU146']*

*df.iloc[:,-4:] = df.iloc[:,-4:]*

*df.drop('SU7',axis=1,inplace=True)*

*list = ['Marijuana','Stimulants','Inhalents','Halucogens','Opiotes','Steroids','Morphine','Xanax','Cigaretes']*

*vars = ['SU1', "SU2", "SU3", "SU4", "SU5", "SU6", "SU8", "SU9", "SU10"]*

*#good way to do it*

*for x, v in zip(list,vars):*

*df[x+"\_undermonth"] = 0*

*df.loc[df[v] >= 2, x+'\_undermonth'] = 1*

*df.MHSCORE.median()*

*df['MHSTATUS'] = np.where(df['MHSCORE'] > 125, 1, 0)*

*df['CESDSTATUS'] = np.where(df['CESDSCORE'] >= 10, 1, 0)*

*df['STAISTATUS'] = np.where(df['CESDSCORE'] >= 40, 1, 0)*

*from scipy.stats import chisquare*

*from scipy.stats import chi2\_contingency*

*x = confusion\_matrix(y\_pred=df['Marijuana\_undermonth'].to\_numpy(), y\_true=df['CESDSTATUS'].to\_numpy())*

*chi2\_contingency(x)*

*ConfusionMatrixDisplay(x).plot()*

*x = confusion\_matrix(y\_pred=df['Protein\_lastmonth'].to\_numpy(), y\_true=df['CESDSTATUS'].to\_numpy())*

*chi2\_contingency(x)*

*chi2\_contingency(confusion\_matrix(y\_pred=df['Protein\_lastmonth'].to\_numpy(), y\_true=df['CESDSTATUS'].to\_numpy()))*

*def chi(pred,true):*

*return chi2\_contingency(confusion\_matrix(y\_pred=df[pred].to\_numpy(), y\_true=df[true].to\_numpy()))*

*chi("Creatine\_lastmonth","CESDSTATUS")*

*df.iloc[:,-16:-3].columns*

*df.iloc[:,-16:-3].sum(axis=1).sort\_values(ascending=False)*

*df.iloc[:,-16:-3].sum(axis=0).sort\_values(ascending=False)*

*df.drop("Steroids\_undermonth",axis=1,inplace=True)*

*from bioinfokit.analys import stat*

*import statsmodels.api as sm*

*string = ""*

*for x in df.iloc[:,-15:-3][['Protein\_lastmonth',"Marijuana\_undermonth","Creatine\_lastmonth","Cigaretes\_undermonth","Stimulants\_undermonth"]]:*

*string = string + x + " + "*

*string = string[:-3]*

*print(string)*

*df.iloc[:,-16:] = df.iloc[:,-16:].astype("category")*

*#Anova model using MHSCORE*

*import statsmodels.api as sm*

*from statsmodels.formula.api import ols*

*# Ordinary Least Squares (OLS) model*

*model = ols(f'CESDSCORE ~ {string}', data=df).fit()*

*anova\_table = sm.stats.anova\_lm(model, type=2)*

*round(anova\_table,4)*

*model.summary()*

*model = ols(f'MDISCORE ~ {string}', data=df).fit()*

*anova\_table = sm.stats.anova\_lm(model, type=2)*

*round(anova\_table,4)*

*model.summary()*

*model = ols(f'BDSSCORE ~ {string}', data=df).fit()*

*anova\_table = sm.stats.anova\_lm(model, type=2)*

*round(anova\_table,4)*

*model.summary()*

*model = ols(f'SPASSCORE ~ {string}', data=df).fit()*

*anova\_table = sm.stats.anova\_lm(model, type=2)*

*round(anova\_table,4)*

*model.summary()*

*model = ols(f'RSSCORE ~ {string}', data=df).fit()*

*anova\_table = sm.stats.anova\_lm(model, type=2)*

*round(anova\_table,4)*

*model.summary()*

*model = ols(f'STAISCORE ~ {string}', data=df).fit()*

*anova\_table = sm.stats.anova\_lm(model, type=2)*

*round(anova\_table,4)*

*model.summary()*

*model = ols(f'AUDITSCORE ~ {string}', data=df).fit()*

*anova\_table = sm.stats.anova\_lm(model, type=2)*

*round(anova\_table,4)*

*model.summary()*

*#LogReg model using MHSTATUS*

*from sklearn.linear\_model import LogisticRegression*

*X = df.iloc[:,-15:-3][['Protein\_lastmonth',"Marijuana\_undermonth","Creatine\_lastmonth","Cigaretes\_undermonth","Stimulants\_undermonth"]].to\_numpy()*

*y = df.iloc[:,-2:-1].to\_numpy()*

*y = y.reshape((len(y),))*

*Lg = LogisticRegression()*

*Lg.fit(X,y)*

*Lg.predict(X)*

*Lg.score(X,y)*

*X = np.array(X, dtype=float)*

*import statsmodels.api as sm*

*log\_reg = sm.Logit(y, X).fit()*

*print(log\_reg.summary())*

*print(log\_reg.summary2())*

*yhat = log\_reg.predict(X)*

*prediction = [yhat.round()]*

*# comparing original and predicted values of y*

*print('Actual values', y)*

*print('Predictions :', np.array(prediction[0], dtype=int) )*

*Lg.coef\_*

*coefs = []*

*i = 0*

*for a in df.iloc[:,-15:-3][['Protein\_lastmonth',"Marijuana\_undermonth","Creatine\_lastmonth","Cigaretes\_undermonth","Stimulants\_undermonth"]].columns:*

*coefs.append([a, round(Lg.coef\_[:,i][0],4)])*

*i += 1*

*coefs*

*df[df.CESDSTATUS == 1]["CESDSTATUS"].count()*

*import seaborn as sns*

*df.iloc[:,-16:] = df.iloc[:,-16:].astype("float64")*

*dfcor = df[['Protein\_lastmonth',"Marijuana\_undermonth","Creatine\_lastmonth","Cigaretes\_undermonth","Stimulants\_undermonth","CESDSCORE"]].corr()*

*sns.heatmap(dfcor, annot=True, linewidths=0.5)*

*plt.xticks(rotation = 25, ha = "right")*

*#Kmodes MHSTATUS*

*from kmodes.kmodes import KModes*

*km = KModes(n\_clusters = 4)*

*km.fit\_predict(X)*

**Documentation:**

***Data Science Capstone***

***Identifying Muscle Dysmorphia***

***This is the repository for the Fall 2023 SantaBarbara Capstone project***

**Members**

***Alex Butera -*** [***buteraa@merrimack.edu***](mailto:buteraa@merrimack.edu)

***Ryan Canfield -*** [***canfieldr@merrimack.edu***](mailto:canfieldr@merrimack.edu)

***Patrick Norcross -*** [***norcrossp@merrimack.edu***](mailto:norcrossp@merrimack.edu)

**Requirements**

Software:

Python IDE, Jupyter notebooks, R, RStudio

Packages:

Python: Pandas, Numpy, Scipy, Matplotlib, scikit-learn, statsmodels, bioinfokit

R: tidyverse, factoextra, FactoMineR, gbm, tree, and randomForest

**How to run**

If you copy this repository to your local computer everything should run after the other requirements have been met. All .csv files are part of the repository and will load into their respective jupyter file. Please note that relative paths were not used in the .jpynb files so some adjustments should be made to load the .csv files correctly

**Files**

**The three important files to run are..**

*Aim1\_code.jpynb - Code for aim 1: Characterize resistance training behavior, and mental health, in a sample of men with symptoms of muscle dysmorphia.*

*Aim #2Rmd - Code for aim 2: Explore whether a worse mental health, including symptoms of muscle dysmorphia, are associated with more frequent and intense resistance training behaviors.*

Aim3\_code.jpynb - Code for aim 3: Explore any associations between mental health and substance and supplement use.

Other jupyter notebook files

*Preprocessing MD2\_Dataset.ipynb - Focused on specific data cleaning necessary for the MD2 original file.*

*DataCleaning.ipynb - Focused on specific data cleaning necessary for the MD\_Acute original file, as well as any final cleaning necessary once the 2 data sets were combined.*

Demographic data and information.ipynb - This was for general information and demographic data on the individuals in our dataset.

**.CSV files:**

***Clean, final data sets***

*combined\_dataset\_ALL.csv - Contains individuals with and without muscle dysmorphia*

combined\_dataset.csv - Only contains individuals with muscle dysmorphia

***Other .CSV files***

*MD2\_Data\_set.csv - The original MD2 dataset, no preprocessing*

*MDACUTE\_Database.csv - The original MDACUTE dataset, no preprocessing*

*PREPROCESSED\_MD2\_Database\_2.1.csv - Early preprocessing that was completed in Excel instead of python.*